HR Analytics

Predict if an employee is likely to leave the company

Authors:
Andy Huang
Arun Gopal
Shweta Gujrathi
Sahana Nataraja

DePaul University, Chicago, IL

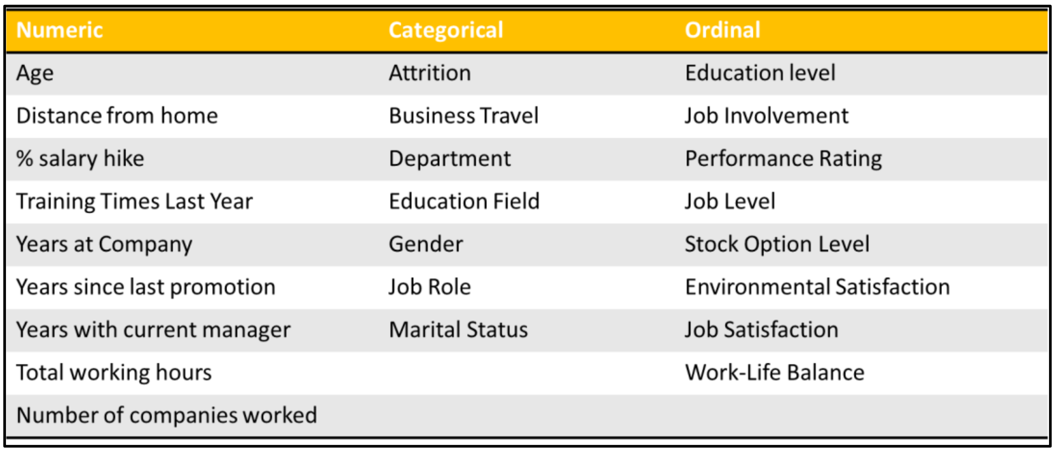
Table of Contents

Executive Summary

“ Employee Benefits News reported in 2017 that turnover can cost employers 33 percent of an employee’s annual salary. The culprit? The hiring of a replacement. To put a dollar amount on it, if the employee earned a median salary of $45,000 a year, this would cost the company $15,000 per person — on top of the annual $45,000. Considering that a survey from Willis Tower Watson found that one in three hires will leave a company within two years, you see how quickly this can add up.” (Forbes).

Even though the Forbes report discusses the cost to company in dollars for companies in the USA, the same principal can be applied to any other company in any part of the world. XYZ company is based in India (and therefore the currency used in Rupees) has more than 4000 employees and around 15% of its employees leave each year Employee Attrition is a causing the company to rethink their relationships with the employees. High Attrition is not only costing the company in terms of money spent on replacement and training, client projects are getting delayed as well as the image of the company amongst prospective employees is getting affected too (Kaggle). The HR Analytics project analyzes significant features that is causing the employees to leave the company. The project also evaluates employee’s monthly income to analyze the factors that are influencing an employee’s monthly income and to make sure that everyone is compensated equally and fairly.

Several factors were studied for the purpose of finding the reason for employee attrition and monthly income–



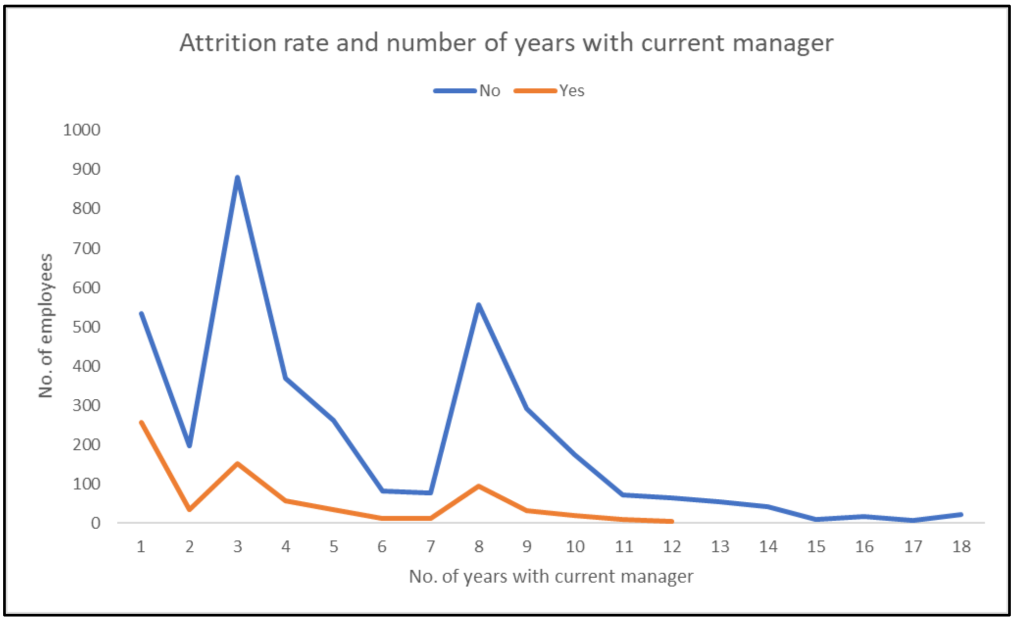
Numeric	Categorical	Ordinal
Age	Attrition	Education level
Distance from home	Business Travel	Job Involvement
% salary hike	Department	Performance Rating
Training Times Last Year	Education Field	Job Level
Years at Company	Gender	Stock Option Level
Years since last promotion	Job Role	Environmental Satisfaction
Years with current manager	Marital Status	Job Satisfaction
Total working hours		Work-Life Balance
Number of companies worked		

Employees were surveyed for the features such as Environmental Satisfaction, Job Satisfaction, Work-Life Balance on a scale of 1-4 with 4 being the highest. Managers were surveyed for the features such as Job Involvement and Performance Rating on a scale of 1-4 with 4 being the highest.

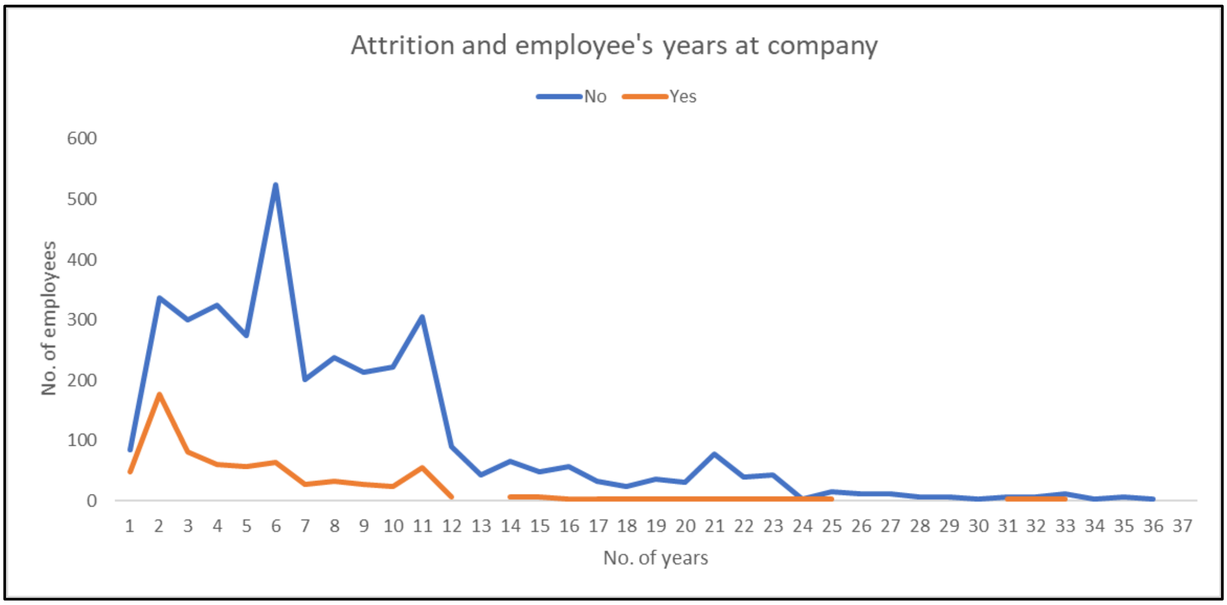
Over the course of the analysis, it was decided that the features Years at company, Years since last promotion and Years with current manager can be grouped together under ‘Experience with company’. Similarly, age, total working years and number of companies worked can be grouped together under ‘Overall Experience’. Distance from home, Job Level and monthly income can be grouped together under ‘Job Satisfaction’.

The results show that ‘Experience with company’ plays the most important role in whether an employee leaves the company or not. Since this feature consists of data about employee’s years at company, years since they last got promoted and years with current manager, these are the factors that the company can work on to reduce employee attrition.

The company can actively track if their employees are duly getting promoted and recognized. They can also periodically assess the manager – employee relationship because if an employee is not happy working with current manager, it's unlikely they would spend too many years working under them.



As it can be seen from the above graph, the highest employee attrition has happened within the first 2 years of an employee working with their manager. If employees are happy working with managers and share a good relationship with them, it will not give them a cause to leave the company.



As per the above graph, a high number of employees left within the first 2-3 years at the company. This aligns with the previous analysis as well. If the employees are unhappy and not being recognized, they will move early on.

The employee attrition numbers show that a high number of employees are leaving within the first few years at the company. The company is spending significant money in training and onboarding the employees and if they leave early on, the company isn't getting value from their investment.

In the monthly income analysis, it was found that experience with company positively influences monthly income whereas overall experience negatively affects monthly income. The negative impact on monthly income is not what is expected, and it may be because the model for monthly income is not statistically significant and therefore, it is advised that it may not be used for prediction purposes. Any findings for the monthly income variable are directional.

To conclude, the company can make significant changes in the employee attrition rate by improving employee – manager relationship and duly recognizing employees through promotions. This will not only help save costs but also help in creating better value to the clients.

Introduction

This HR Analytics data is a case study project got from Kaggle. It is always in the best interests of the company to know what changes to make in the workplace so that the employee does not leave the company and curb attrition rate. Also, it is very important to predict other questions like monthly income. This dataset gives us the means to analyze all the above discussed parameters and implement logistic and linear regression and use PCA as there are more the 20 features (to help avoid overfitting) and other advanced analytics to analyze data, find insights and finally interpret the results.

To get the final dataset, we have combined data from 3 different data source files. The general employee data was combined with manager and employee survey. The survey asked the manager to rank employees on a scale of 1-4 on Performance Rating and Job Involvement. The other survey asked the employees to rank features such as Environment Satisfaction, Work Life Balance, Job Satisfaction on a scale of 1-4.

Goals

The goal of the project is as follows:

- Logistic Regression
  - Predict probability of attrition
- Linear Regression
  - Predict monthly income

Dataset Details

a. **Dataset Name:** HR Analytics Case Study

b. **Number of dependent variables:**

- Attrition – Binary variable
- Monthly Income – Numeric variable

c. **Target (independent variables) and its type:**
There are various metric/numeric variables and categorical/binary variables within our dataset.

Based on other 25 predictors, we are interested in predicting the attrition of the company as well as the employees' monthly income. There are 27 variables been kept for now, including the employee id, in case there is further interest to add more data. Within the 27 variables, 25 of the predictors will be used to predict the likelihood of the employees' attrition and the employees' monthly income.

Our predictors are as follow:

- Employee ID
- Numeric/metric variables:
  - **Age** – Age of the employee
  - **Distance from Home** – Distance from home in kms
  - **Environment Satisfaction** – 1"Low"/1"Medium/3"High"/4"Very High/NA's
  - **Job Satisfaction** – 1"Low"/1"Medium/3"High"/4"Very High"/NA's
  - **Work Life Balance** – 1"Bad"/2"Good"/3"Better"/4"Best"/NA's
  - **Job Involvement** – 1"Low"/1"Medium/3"High"/4"Very High"
  - **Performance Rating** – 1"Low"/2"Good"/3"Excellent"/4"Outstanding"
  - **Job Level** – Job level at company on a scale of 1 to 5
  - **Monthly Income –** (in Rupees) Predictor
  - **NumCompaniesWorked** – Total number of companies the employee has worked for

- **PercentSalaryHike** – percent salary hike for last year
- **Standard Hours** – Standard hours of work for the employee
- **StockOptionLevel** – Stock option level of the employee
- **TotalWorkingYears -** Total number of years the employee has worked so far
- **TrainingTimesLastYear -** Number of times training was conducted for this employee last year
- **YearsAtCompany -** Total number of years spent at the company by the employee
- **YearsSinceLastPromotion -** Number of years since last promotion
- **YearsWithCurrManager -** Number of years under current manager

- Categorcal/Binary variables:
  - **Attrition** – Whether the employee left in the previous year or not (Yes / No)
  - **Business Travel** – Non-Travel/Travel rarely/ Travel Frequently
  - **Department** – Human Resources/ Research & Development / Sales
  - **Education** – 1"Below College"/2"College"/3"Bachelor"/4"Master"/5"Doctor"
  - **Education Field** – Human Resources/Life Sciences/Marketing/Medical/Technical Degree/Other
  - **Gender** – Male/Female
  - **Job Role** – Healthcare Representative/Human Resources/Laboratory Technician/Manager/Manufacturing Director/Research Director/Research Scientist/Sales Executive/Sales Representative
  - **Marital Status** – Married/Single/Divorced

d. **Dependent Variable and Independant Variable:**

- **Attrition:** 0 or 1
- **Monthly income:** numeric continuous variable

e. **Dataset URL**: https://www.kaggle.com/vjchoudhary7/hr-analytics-case-study#general_data.csv

f. **Missing data/observations :** There is not a lot of missing data. We do have some observations which are 'NA's. Those are as follows -
Environment Satisfaction – 25 NA's
Job Satisfaction – 20 NA's
Work Life Balance – 38 NA's
Number of Companies Worked – 19 NA's
Total Working years – 9 NA's

Methodology

The basic overview of the analysis methodology used here is:

- Exploratory data analysis
- Initial Model Building
- PCA on continuous variables
- Ordinal Factor Analysis
- PFA on numeric variables
- Correspondence Analysis
- Advanced Model building

Technical Summary

1. Exploratory Data Analysis

In the exploratory data analysis, the quantitative variables were analyzed using histograms and the categorical variables were analyzed using the frequency tables. Most of the employee fall in the age range of 30 to 40, and most of the employees in the company live approximately less than 5 km or 10 km from the company. Fewer employees need to commute more than 10 km to the company. Most employees have a Bachelor or master's degree and have a high to medium level of the job involvement. Most employees have a monthly income of 30000 rupees, but fewer employees have a monthly income ranging from 80000 to 200000 rupees. The employees generally 11 to 12 percent of the salary hike for last year and the stock option level of the employees are usually 0 or 1.

Last year, most of the employees at the company had 2 to 3 times of the training time. As for the length of the years the employees stayed in the company, our data gathered the length of years from less than 1 year to more than 30 years, with most of the employees stay in the company for less than 10 years. Finally, most employees have received a promotion about 1 year ago at the point our data gathered. In addition, most employees have the same manager for less than or about 2 years. Many of the employees have also have the same manager for 7 years.

Frequency Table

For variables that are not suitable for plotting histogram or scatter plot, a frequency table was generated for each of the predictor to better understand the spread of the data. For the parameter of interest, Attrition, the dataset contains 711 employees that already left the company in the previous year. On the other hand, there are 3699 employees that are still with the company, which account for approximately 83.88%. For the performance rating, only about 15.37% of the total employees got" outstanding," the remaining 84.63% of the employees received" excellent." As for some demographic information, there are 60% of the total employees within our dataset are male while 40% of the employees are female. There are 2883 employees work in the R&D department, which accounted for roughly 65.37% of the total employees. Only 189 employees work in Human Resources which only account for 4.29% of the total employees. 70.95% of the employees indicated that they rarely travel. However, there are 18.84% of the employees indicated that they travel frequently. There are also 10.2% of the employees do not travel at all. As for employees' marital status, 45.78% of the employees are married but a total of about 54.21% of the employees are single or divorced. Most employees were majored in Life Sciences (41.22%) and Medical (31.56%). Only 81 employees within our dataset (1.84%) have majored in Human Resources. There are employees who also majored in Marketing (10.82%), owns a technical degree (8.98%), or have other education field (5.58%). The job roles for most employees are Sales Executives (22.18%), Research Scientist (19.86%), or Laboratory Technician (17.62%). Together these three job roles have accounted for approximately 60% of the total employees.

Correlation among the continuous variables:

The dataset has a limited number of multi-collinearities among the predictor variables. Among the 15 numeric variables, only the following variables have high correlation between them. The variables with high correlation are:

PercentSalaryHike – PerformanceRating – 77 percent correlation
YearsAtCompany – YearsSinceLastPromotion – 62 percent correlation
YearsAtCompany – YearsWithCurrManager – 72 percent correlation
YearsSinceLastPromotion – YearswithCurrManager – 51 percent correlation

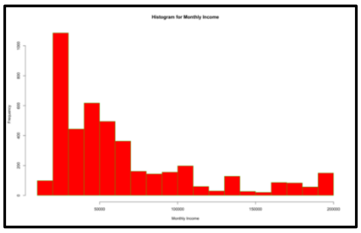
The correlation between the variables makes total sense in understanding of the data. Performance rating and hike in salary shows a positive correlation which is as expected. One would have expected a negative correlation between YearsAtCompany and YearsSinceLastPromotion, instead it had positive correlation. The variables overall do not have much correlation between them. The multi-collinearity needs more analysis where the correlation among the variables can be better understood.

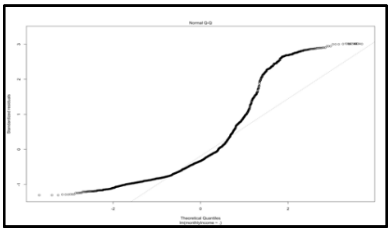
2. Initial Model Building

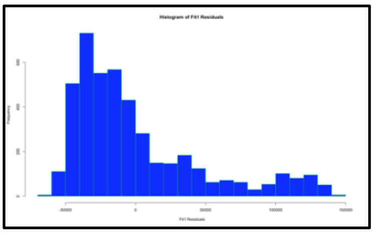
Ordinary Least Squares

The original dataset came with 4410 observations and a total of 27 independent variables. Each type of variables should be investigated, after separating the numeric variables, the ordinal variables, and the categorical variables. Since there is basically no correlation among ordinal variables, the ordinal variables have been treated as numeric variables, and a new data frame has been created which contains both numeric and ordinal variables for ease of use of the analysis. As a result, both ordinal and categorical variables are being used to run the Ordinary Least Square with the assumption that there are some relationships worth investigating. Furthermore, Forward, Backward, and Stepwise Model Selection have been performed to allow us to gain a better picture as which variables have relatively stronger effects on the dependent variable, *Monthly Income*. However, a log transformation is needed for Ordinary Least Square, so log transformation has been applied on the linear regression model. Even after log transformation, there are still too much variance within Monthly Income have not been explained. It is a both a conclusion and a limitation that Ordinary Least Square does not fit the nature of the chosen dataset, even after log transformation. Moreover, the Forward, Backward, and Stepwise Selection have all given the same result, which represent that the variables chosen by these three methods are potentially relatively important predictors for *Monthly Income*.

As a first step of Exploratory data analysis, many independent variables were plotted as histogram to get a first understanding of the data. Most of the predictors shows a right skewed on their respective histograms. One of the parameters of interest, *Monthly Income*, also shows a right skewed on the histogram, as shown below.

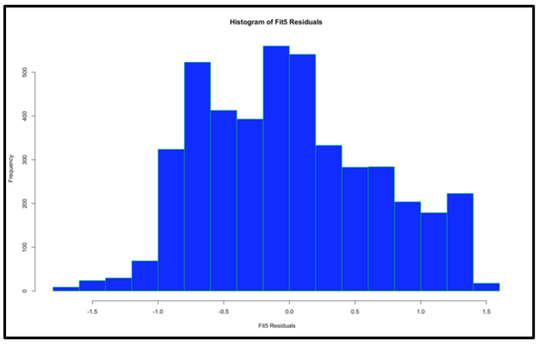


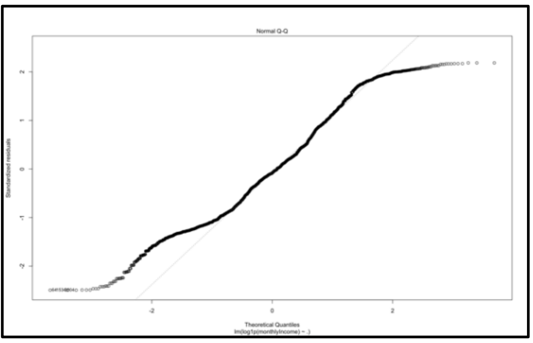




As for the other parameter of interest, Attrition, a frequency table has been generated, which revealed that there are about 83.88% of the total employees are still with the company, while 16.12% percent of the total employees have left the company in the previous year. After exploratory analysis, several linear regression models were fitted to predict Monthly Income. However, it had become obvious that a log transformation is needed, as the R-square are very low (about 1%), the Normal QQ plot does not show close to a straight line, and the histogram of the residuals was not close to normally distributed, as shown above.

Going forward, log transformation has been applied on Monthly Income, with the predictors being both numeric and ordinal variables – a total of 17 variables are used, 9 are numeric variables and 8 are ordinal variables.



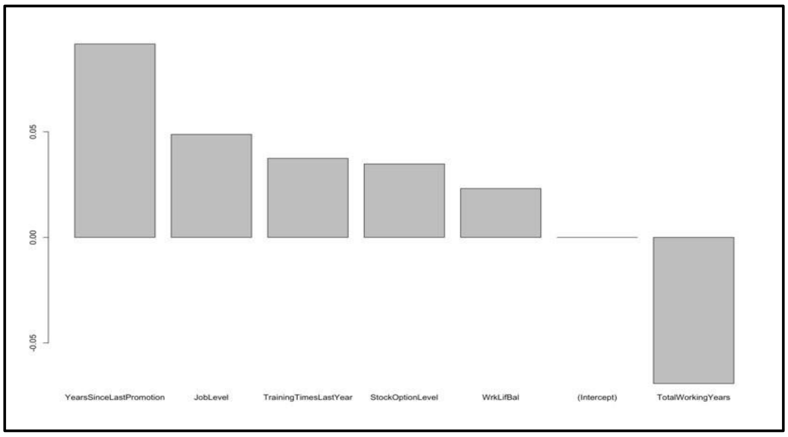


The R-square value still shows only 1.55% after transformation, which is really low. 4 Variables – *Training Time Last Year*, *Years Since Last Promotion*, *Job Level*, and *Stock Option Level* –

shows significant at the level of alpha = 0.05. The normal Q-Q plot does show an improvement, as presented above, but there are still many outliers that did not been captured. As for the histogram of the residuals, similarly, it has improved to somehow closer to normally distributed, but it still does not show normal distribution. OLS does not really benefit our analysis much, as there is still too much variance that did not been captured and explained by OLS.

Automate Model Selection

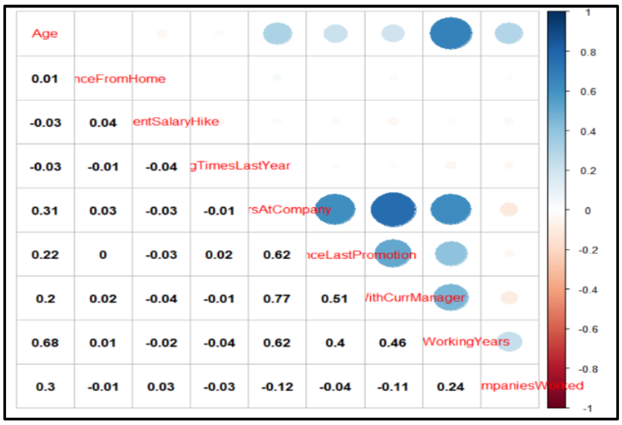
Forward Selection, Reverse Elimination, and Stepwise Selection have also been performed to compare the results. All three model selection methods return the same results, as shown below. Forward, Backward, and Stepwise selections show that *Training Time Last Year*, *Years Since Last Promotion*, *Job Level*, *Stock Option Level*, *Work Life Balance*, and *Total Working Years* are significant predictors for *Monthly Income*, with *Total Working Years* being the only negative predictor. In addition, *Years Since Last Promotion* has the strongest positive effect on *Monthly Income*.

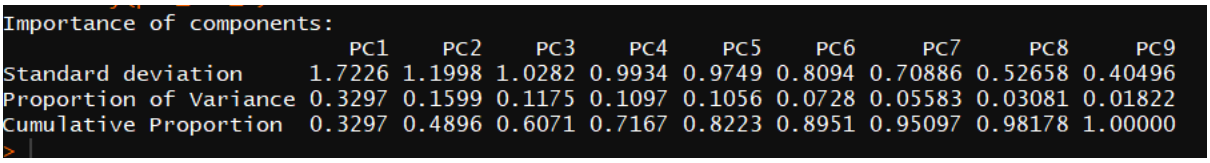


Training Time Last Year, *Years Since Last Promotion*, *Job Level*, *Stock Option Level*, *Work Life Balance*, and *Total Working Years* seem to have relatively stronger influence for *Monthly Income*, as these variables are all been selected by Forward, Backward, and Stepwise Selection. Automate model selection also revealed that *Total Working Years* seems to have negative effect on *Monthly Income*, if the number of total working years increase, the monthly income tends to decrease. On the other hand, these model selections also suggest that if the years since last promotion increase, the monthly income tends to increase as well. As for Ordinary Least Square, a very low percentage of the variance within *Monthly Income* has successfully been captured by the model, even after logistic transformation. As a limitation, Ordinary Least Square does not benefit the analysis much – still too much variance that did not been captured. However, the results from automate model selection could be used as comparison to the results from other techniques used in this project.

3. Principal Component Analysis on the continuous variables

PCA was performed on the 9 numeric variables in the dataset. Initial PCA produced the following results. It took 7 components to capture 90% variance in the data.

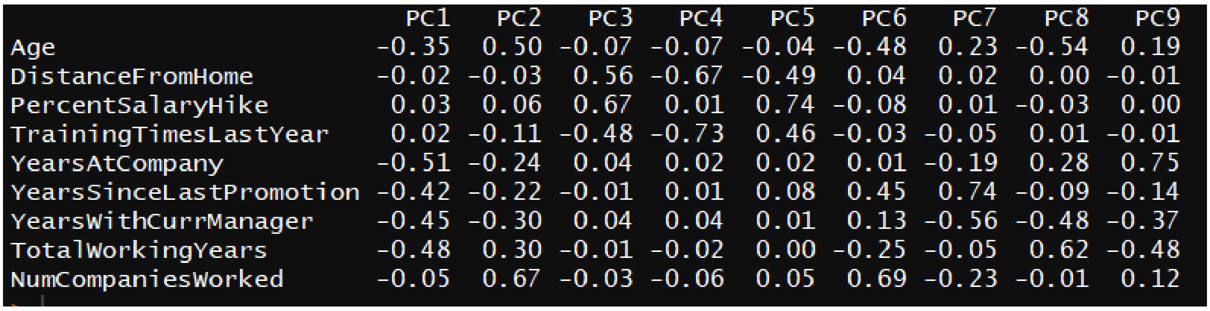




```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6     PC7     PC8     PC9
Standard deviation     1.7226 1.1998 1.0282 0.9934 0.9749 0.8094 0.70886 0.52658 0.40496
Proportion of Variance 0.3297 0.1599 0.1175 0.1097 0.1056 0.0728 0.05583 0.03081 0.01822
Cumulative Proportion  0.3297 0.4896 0.6071 0.7167 0.8223 0.8951 0.95097 0.98178 1.00000
>
```

Since we had only 9 numeric variables, we analyzed the correlation between the numeric variables. The correlation plot explained the correlation between the numeric variables and furthered our understanding between the numeric variables. There was very limited correlation between these variables. The correlation plot complements the results of the PCA analysis. Since there were limited correlation or covariance between the variable, data point rotation was not able to reduce the dimensions in the data.

The component loadings are shown in the table below. Component 1 explains the years of work experience spent in a company with contributions from age, YearsAtCompany, YearsSinceLastPromotion, YearsWithCurrManager, and TotalWorkingYears which all in a way explains the number of years spent working or the working experience of the professional. It is be noted that YearsAtCompany and YearsWithCurr Manager are also strong contributions of the component. Also, NumCompaniesWorked isn't a strong contribution of the component. Also, NumCompaniesWorked isn't a strong contributor. Hence it can be said that component 1 explains the experience of a professional in one company. PC2 gets high contribution from age and NumCompaniesWorked. Years at company have opposite contribution. Hence, we could say that component 2 explains the overall experience of the professional. Though other components have significant contributions from a few variables, the components are difficult to interpret, and the underlying meaning of the components are hard to find. The components start to get high contribution from single variables which again explains the low correlation between these variables.



```
                          PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9
Age                     -0.35  0.50 -0.07 -0.07 -0.04 -0.48  0.23 -0.54  0.19
DistanceFromHome        -0.02 -0.03  0.56 -0.67 -0.49  0.04  0.02  0.00 -0.01
PercentSalaryHike        0.03  0.06  0.67  0.01  0.74 -0.08  0.01 -0.03  0.00
TrainingTimesLastYear    0.02 -0.11 -0.48 -0.73  0.46 -0.03 -0.05  0.01 -0.01
YearsAtCompany          -0.51 -0.24  0.04  0.02  0.02  0.01 -0.19  0.28  0.75
YearsSinceLastPromotion -0.42 -0.22 -0.01  0.01  0.08  0.45  0.74 -0.09 -0.14
YearsWithCurrManager    -0.45 -0.30  0.04  0.04  0.01  0.13 -0.56 -0.48 -0.37
TotalWorkingYears       -0.48  0.30 -0.01 -0.02  0.00 -0.25 -0.05  0.62 -0.48
NumCompaniesWorked      -0.05  0.67 -0.03 -0.06  0.05  0.69 -0.23 -0.01  0.12
```

It was clear from the two analysis that it is difficult to analyze the parameters of interest with just the numeric variables. It is also clear that the ordinal variables in the data should be dealt differently and the categorical variable must be dealt as well. Hence, the next steps was in analyzing the ordinal variables using the Spearman, Kendall and Pearson correlation techniques and correspondence analysis on the categorical variables.

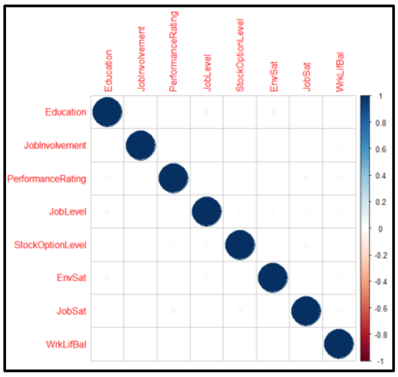
4. Ordinal Factor Analysis

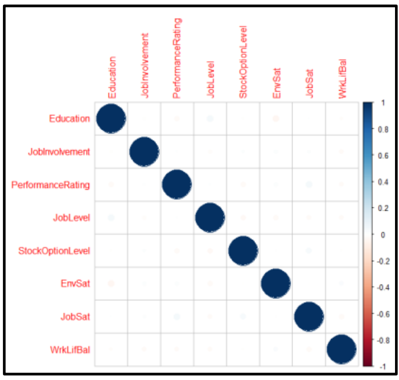
The HR Analytics dataset consisted on 8 ordinal features. This was a huge chuck of ordinal data out of a total 27 features. The goals for the ordinal data analysis are:

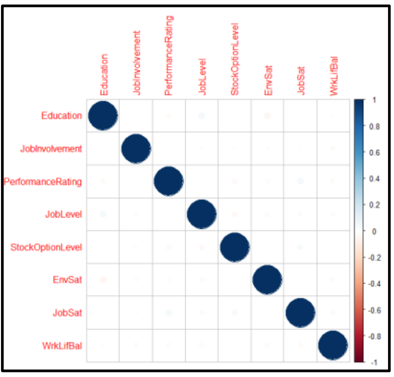
- Find if there are any correlations between the 8 ordinal features using Pearson, Spearman and Kendall methods
- Perform factor analysis on the ordinal data to see if any meaningful groupings are identified
- Combine ordinal and numeric features to do PCA. Use the factor data from PCA to do OLS and logistic regression
- Use "Hetcor" to find correlations between all kinds of features: numeric, categorical and ordinal

Initial step involved in handling missing/NA values in 3 of the 8 ordinal features which were encoded as character data in the dataset. Those character features were converted to numeric features and Missing/NA values were replaced by "mode". Totally 8 ordinal features were available for analysis: *Education*, *JobInvolvement*, *PerformanceRating*, *JobLevel*, *StockOptionLevel*, *Environment Satisfaction*, *JobSatisfaction*, *WorkLifeBalance*.

The correlations of the ordinal features between each other using Pearson, Spearman and Kendall was conducted and all the 3 gave same results showing absolutely no correlations between them:

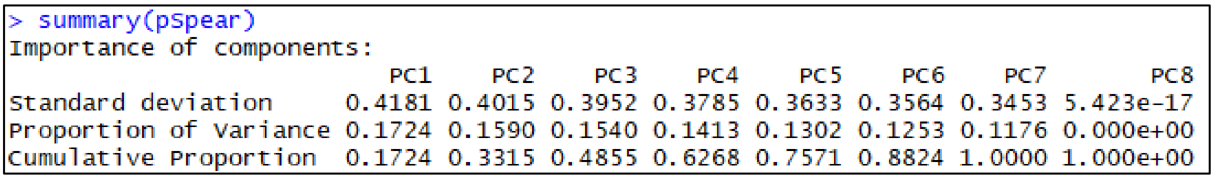




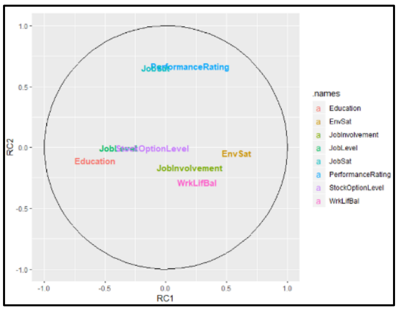


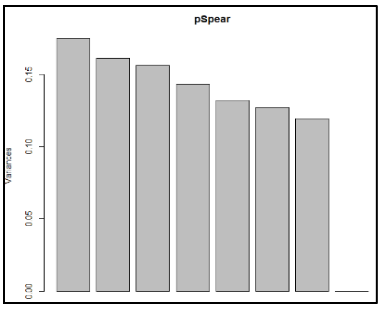
At this point, performing PCA on uncorrelated ordinal data wouldn't make any sense. However, out of curiosity when factor analysis on the uncorrelated data were performed, the Spearman method gave some interesting groupings. This was surprising to see.

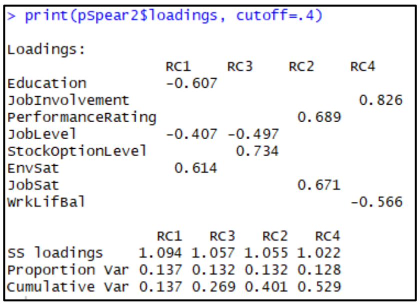
The results of the PCA psych plot and the scree plot and the summary to choose the number of factors is shown below. The result of the Spearman was not impressive, but it was better than the other methods with groupings. 7 features were required to account for a 90% variance in the data (graphs shown below).


```
> summary(pSpear)
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6    PC7       PC8
Standard deviation     0.4181 0.4015 0.3952 0.3785 0.3633 0.3564 0.3453 5.423e-17
Proportion of Variance 0.1724 0.1590 0.1540 0.1413 0.1302 0.1253 0.1176 0.000e+00
Cumulative Proportion  0.1724 0.3315 0.4855 0.6268 0.7571 0.8824 1.0000 1.000e+00
```

4 components were used to do principal factor analysis which accounted to 63% total variance in data. There was no clear knee pattern from the screen plot indicating large number of components to be used in order to get 90% variance in data.



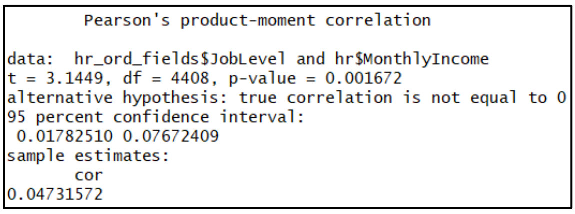



```
> print(pSpear2$loadings, cutoff=.4)

Loadings:
                  RC1    RC3    RC2    RC4   
Education         -0.607                     
JobInvolvement                          0.826
PerformanceRating               0.689        
JobLevel          -0.407 -0.497              
StockOptionLevel          0.734              
EnvSat             0.614                     
JobSat                          0.671        
WrkLifBal                              -0.566

                 RC1   RC3   RC2   RC4
SS loadings    1.094 1.057 1.055 1.022
Proportion Var 0.137 0.132 0.132 0.128
Cumulative Var 0.137 0.269 0.401 0.529
```

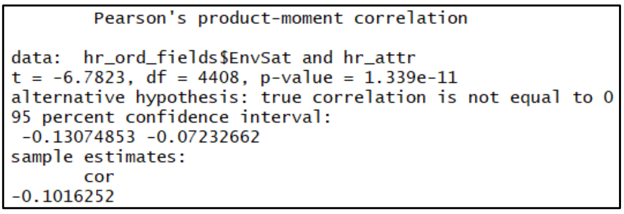
The psych plot with rotated components gave some interesting groupings. RC1 is having positive grouping of Environment Satisfaction and negative for Job level and education. RC2 shows clear positive grouping of Job Satisfaction and Performance Rating and low negative for Job Involvement and Work life balance.

Since there were not much of insightful information was got from ordinal factor analysis, as per feedback from professor and research, correlation analysis of ordinal features with dependent variables were conducted. There were only 4 features that came significant with dependent variables. Decision was made to consider just those important ordinal features as numeric features and performed PCA on it.

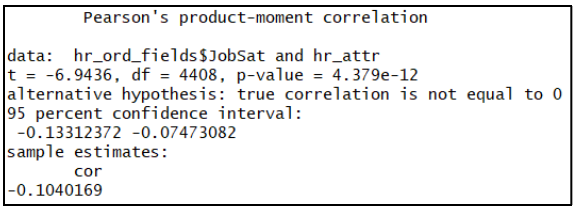
- **EnvironmentSatisfaction, JobSatisfaction** and **WorkLifeBalance** are highly significant with Attrition
- **JobLevel** is highly significant with Monthly Income



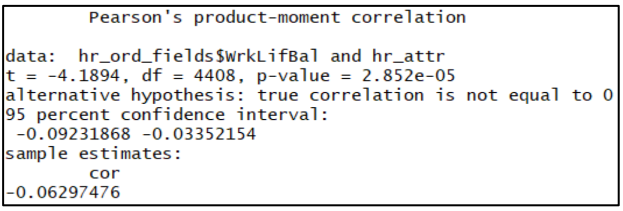
```
        Pearson's product-moment correlation

data:  hr_ord_fields$JobLevel and hr$MonthlyIncome
t = 3.1449, df = 4408, p-value = 0.001672
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01782510 0.07672409
sample estimates:
       cor 
0.04731572
```



```
        Pearson's product-moment correlation

data:  hr_ord_fields$EnvSat and hr_attr
t = -6.7823, df = 4408, p-value = 1.339e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.13074853 -0.07232662
sample estimates:
       cor 
-0.1016252
```



```
        Pearson's product-moment correlation

data:  hr_ord_fields$JobSat and hr_attr
t = -6.9436, df = 4408, p-value = 4.379e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.13312372 -0.07473082
sample estimates:
       cor 
-0.1040169
```



```
        Pearson's product-moment correlation

data:  hr_ord_fields$WrkLifBal and hr_attr
t = -4.1894, df = 4408, p-value = 2.852e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.09231868 -0.03352154
sample estimates:
        cor 
-0.06297476
```

The above correlation coefficients and test indicates that we can reject the null hypothesis and accept the alternate hypothesis as the P-value was very low < 0.05 at 95% confidence interval. However, the correlation percentages are 10% or less. Ideal correlation percentage it to be 40% or higher to be considered. But these features were still considered to be used in PCA to see if it produced any interesting results.

The ordinal features were considered as numeric features. PCA was performed again with numeric and ordinal features to see if it produced better results compared to PCA with just numeric continuous features. The next section explains how it was impleted.

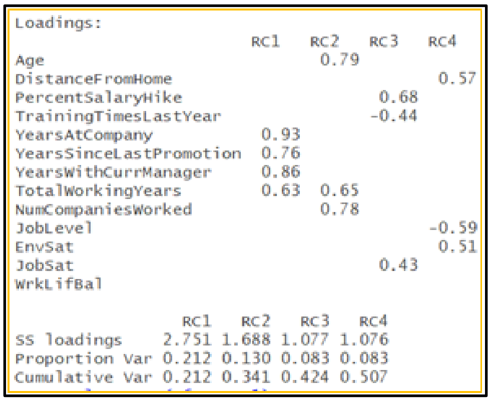
5. Principal Factor Analysis with Continuous and Ordinal Variables

Principal Factor Analysis (PFA) was used to analyze the underlying factors/ components in the numeric variables present in the data. In earlier analyses, Pearson correlation, Spearman correlation and Kendall correlation were calculated, and these techniques provided evidence that the ordinal variables had no correlation between them and hence these variables were treated as continuous variables in our analysis. All the continuous variables and only the ordinal variables significantly correlated with the response variable were included in the analysis. It was found using PFA that 4 factors could capture 50% of the variance in the data and these factors gave highly relevant and interpretable factors. The factors were named as “Experience with the Company”, “Overall work experience”, “Job Satisfaction” and “Environmental Satisfaction”. These factors were able to reduce the correlation between the independent variables, reduce the number of dimensions and helped in the interpretation of the variable contribution. Moreover, the scores from these 4 factors were used in further analysis such as logistic regression.

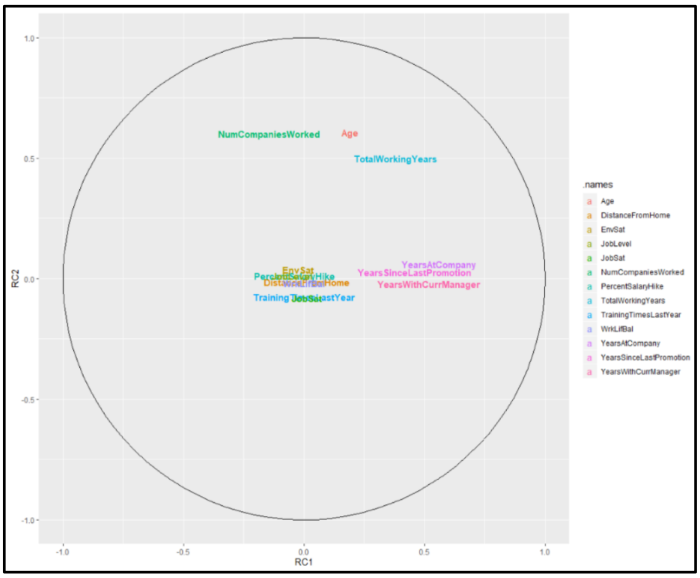
There were two separate principal factor analysis performed. Since the project has two goals - one to predict the *Monthly Income* and another to classify the *Attrition*, the PFA had two different approach. In one approach, the variable *MonthlyIncome* was not included in PFA and in another, the variable was included in the analysis. The results of the PFA without including the *MonthlyIncome* were further used in the prediction model of *MonthlyIncome*. The results of the PFA including *MonthlyIncome* were further used in the classification of the attrition.

Only the significant ordinal variables were included in this analysis. Variables such as *JobSatisfaction , EnvironmentSatisfaction, WorkLifeBalance* and *JobLevel* were included in the analysis

PFA with continuous and ordinal features without the inclusion of *MonthlyIncome* gave the below results:



```
Loadings:
                         RC1   RC2   RC3   RC4
Age                            0.79
DistanceFromHome                           0.57
PercentSalaryHike                    0.68
TrainingTimesLastYear               -0.44
YearsAtCompany           0.93
YearsSinceLastPromotion  0.76
YearsWithCurrManager     0.86
TotalWorkingYears        0.63  0.65
NumCompaniesWorked             0.78
JobLevel                                  -0.59
EnvSat                                     0.51
JobSat                               0.43
WrkLifBal

                 RC1   RC2   RC3   RC4
SS loadings    2.751 1.688 1.077 1.076
Proportion Var 0.212 0.130 0.083 0.083
Cumulative Var 0.212 0.341 0.424 0.507
```



As it can be seen in the above psych plot, the groupings are much more significant with more meaningful information. Four factors were able to capture more than 50% variance in the data.

- Factor 1 = 0.93 * YearsAtCompany + 0.76 * YearsSinceLastPromotion + 0.86 * YearswithCurrManager + 0.63 * TotalWorkingYears
- Factor 2 = 0.79 * Age + 0.65 * TotalWorkingYears + 0.78 * NumCompaniesWorked
- Factor 3 = 0.68 * PercentSalaryHike – 0.44 * TrainingTimeLastYear + 0.43 * WrkLifBal
- Factor 4 = 0.57 * DistanceFromHome – 0.59 * JobLevel + 0.51 * EnvSat

The factors can be named using the loadings. The factors can be named as below.

- Factor 1 – Experience with Same Company
- Factor 2 – Overall. Experience
- Factor 3 – Job. Satisfaction
- Factor 4 – Environmental. Satisfaction

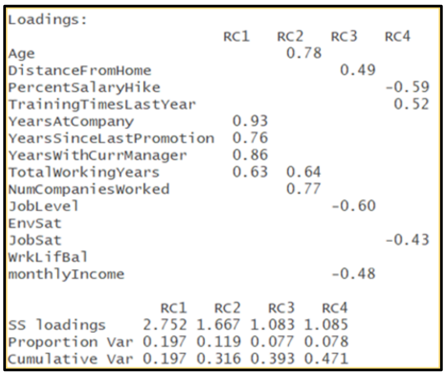
PFA with continuous and ordinal features with the inclusion of *MonthlyIncome* gave the below results. As it can be seen from the loadings, again 4 factors were able to capture 50% variance in the data. Only one of the factor loadings changed. The factor loadings changed and the formulas for each factor is given as follows.

- Factor 1 = 0.93 * YearsAtCompany + 0.76 * YearsSinceLastPromotion + 0.86 * YearswithCurrManager + 0.63 * TotalWorkingYears
- Factor 2 = 0.78 * Age + 0.64 * TotalWorkingYears + 0.77 * NumCompaniesWorked
- Factor 3 = 0.49 * DistanceFromHome – 0.60 * JobLevel – 0.48 * MonthlyIncome
- Factor 4 = -0.59 * PercentSalaryHike + 0.52 * TrainingTimeLastYear - 0.43 * WrkLifBal

Although the loadings changed minimally, the definition of one of the factors changed. The new factors can be named as follows.

- Factor 1 – Experience with Same Company
- Factor 2 – Overall. Experience
- Factor 3 – Job. Satisfaction
- Factor 4 – Job.Level

These factors were included in furthering model building like logistic regression and partial least squares regression method.



```
Loadings:
                         RC1   RC2   RC3   RC4
Age                            0.78
DistanceFromHome                     0.49
PercentSalaryHike                         -0.59
TrainingTimesLastYear                      0.52
YearsAtCompany           0.93
YearsSinceLastPromotion  0.76
YearsWithCurrManager     0.86
TotalWorkingYears        0.63  0.64
NumCompaniesWorked             0.77
JobLevel                            -0.60
EnvSat
JobSat                                    -0.43
WrkLifBal
monthlyIncome                       -0.48

                 RC1   RC2   RC3   RC4
SS loadings    2.752 1.667 1.083 1.085
Proportion Var 0.197 0.119 0.077 0.078
Cumulative Var 0.197 0.316 0.393 0.471
```

6. Correspondence Analysis on Categorical Variables

Correspondence Analysis is useful to understand the relationship of our categorical data, and it could also be plotted like PCA for better visualization. As a result, under the assumption that there is relationship among the categorical variables, Correspondence Analysis has been performed to identify the correlation between the categorical variables. Each categorical variable was paired with one of the parameters of interest, *Attrition*. *Business Travel*, *Department*, *Gender*, *Job Role*, *Marital Status* and *Education Field* in relation with *Attrition* were examined. Moreover, another pair of categorical variables that also been examined is *Business Travel* vs *Department* and *Education Field vs Job Role*. The results indicate that, compare to other departments, Human Resources does have higher attrition rate. In addition, it has been revealed that the employee who travel frequently tend to have a higher attrition rate.

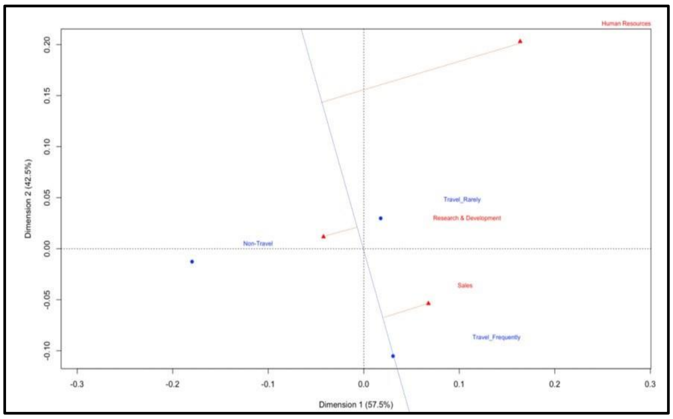
From Correspondence Analysis, Human Resources in this company came out as an interesting department. Comparing to R&D and Sales departments, HR shows a higher employee attrition rate. From the education field perspective, 40.74% of the employees who hold an HR degree have left the company in the previous year, which is significantly higher than the employees who hold other areas of degree.

To further explore the reason of the relatively higher attrition rate of HR, *Department* and *Business Travel* have been paired. Over 80% of the employees who work in HR rarely travel, and over 70% of the employees who work in both R&D and Sales departments rarely travel as well.

The Chi-squared test of independent was performed for each pair of the categorical variables to test the independence. The p-values of 7 pairs of the categorical variables are all significantly less than the .05 significance level, except for *Gender* versus *Attrition*. In the analysis of gender versus attrition, it was found through chi-squared test that there were no association between *Gender* and *Attrition*. As a result, the null hypothesis that the 7 pairs of the categorical variables are independent has been rejected. In other words, the 7 pairs of the categorical variables are not independent.

From the contingency table, it seems like the frequency of business travel, however, does not have direct influence on the employee attrition rate. However, the mosaic indicates that the frequency of the employees who travel frequently and are still with the company is less than we expected. In other words, the employees who travel frequently tend to leave the company.

The mosaic plots of *Attrition vs Department* and *Attrition vs Education Field* indicate the same findings as mentioned earlier about the higher attrition rate of HR. The mosaic plot of *Attrition vs Department* indicates that the frequency of the employees who work in HR department and are still with the company is less than we expected. The mosaic plot of *Attrition vs Education Field* indicates that the frequency of the employees who hold an HR degree and have left the company in the previous year is more than we expected.



Drawing a line from Travel Frequently through origin, it became obvious that Sales department corresponds most to travel frequently. HR department, however, corresponds the least to travel frequently but correspond the most to travel Rarely. These findings have aligned with the output presented above.

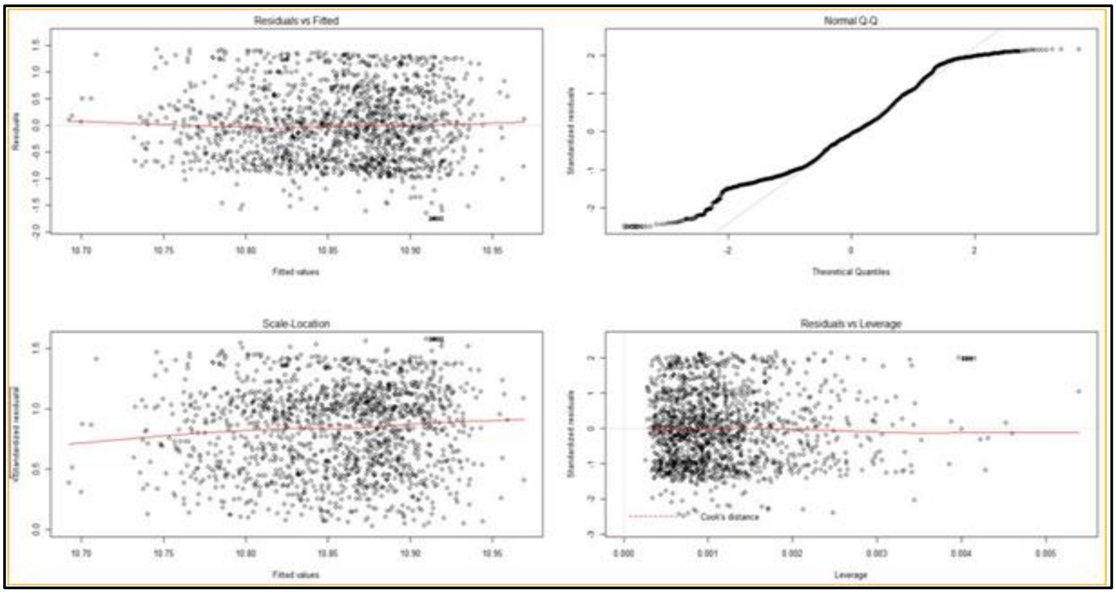
In the analysis of marital status versus attrition, it was found that the likelihood of singles leaving the company was very high. Also, the likelihood of married and divorced workforce leaving the company was very low. Chi-squared test, mosaic plot and contingency table also proves the same. Similarly, in the analysis of job role versus attrition, it was found that the research directors have a very high attrition rate. As opposed the research directors, the manufacturing directors have a very low attrition rate and the likelihood of them leaving the company was very low.

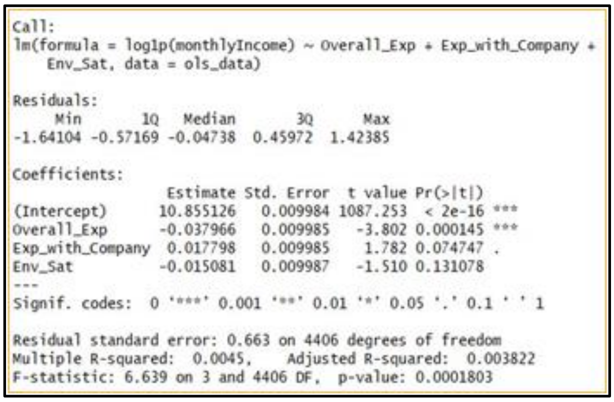
In the analysis of education field versus job role, it was found that the likelihood of sales executives in the company having a technical degree is very high. This explains the nature of the company and that the client facing sales executives are required a technical degree rather than the research scientists. These were the associations between the categorical variables found in the data set. Further graphs and plots are attached in the appendix.

7. Advanced Model building:

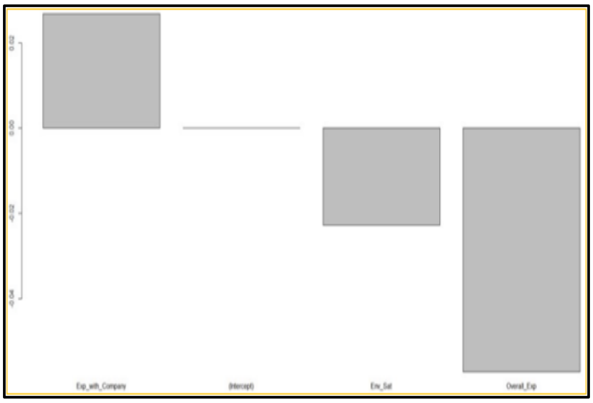
OLS with continuous numeric and ordinal features

OLS using the above factor data produced below results. The R^2 lower with 0.45% but the model overall was significant with p-value < 0.05 at 95% confidence interval. The residuals did not show any pattern and the distribution looked almost normal. However, the straight line seen in the residual plot is due to inclusion of ordinal feature in OLS.





```
Call:
lm(formula = log1p(monthlyIncome) ~ Overall_Exp + Exp_with_Company +
    Env_Sat, data = ols_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.64104 -0.57169 -0.04738  0.45972  1.42385

Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)      10.855126   0.009984 1087.253  < 2e-16 ***
Overall_Exp      -0.037966   0.009985   -3.802 0.000145 ***
Exp_with_Company  0.017798   0.009985    1.782 0.074747 .
Env_Sat          -0.015081   0.009987   -1.510 0.131078
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.663 on 4406 degrees of freedom
Multiple R-squared:  0.0045,    Adjusted R-squared:  0.003822
F-statistic: 6.639 on 3 and 4406 DF,  p-value: 0.0001803
```



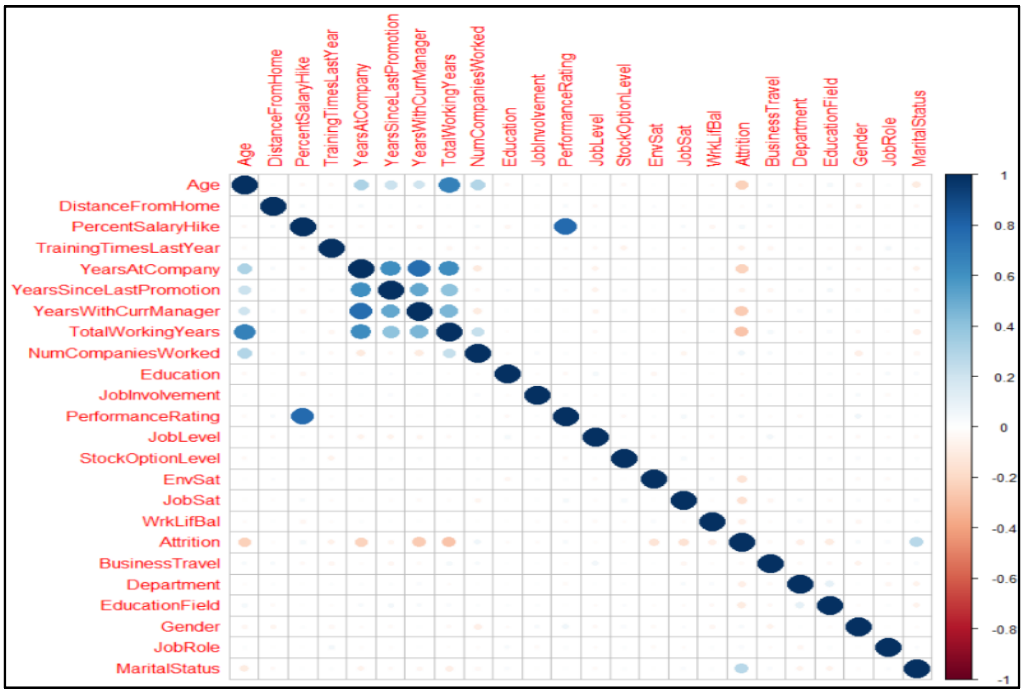
The significant features selected by the feature selection methods were:

- **Experience with company** positively influences Monthly Income
- **Environment Satisfaction** and **Overall Experience** negatively influences Monthly Income
- **Overall Experience** is the strongest influencer

8. Correlation analysis using Hetcor:

“Hetcor” correlation was used to check correlation of all 3 types of features in our dataset: continuous, ordinal and categorical. This analysis was important as it would give the correlation using all the features. The hetcor correlation resulted with the grouping came as the “Experience with Company”.

YearsAtCompany,YearsSinceLastPromotion, YearsWithCurrManager, TotalWorkingYears shows stronger correlations. These features seem to be more important for our analysis than the rest.



This analysis gave the confidence that the features used for final analysis were justified and it was in the right direction.

Logistic Regression

One of the parameters of interest is attrition which is a binary variable and logistic regression was used to understand the predictors which have an impact on attrition. Logistic regression was performed initially only on the numeric variables and then again on the PCA factor data for numeric and ordinal variables.

Logistic Regression was performed in 2 steps –

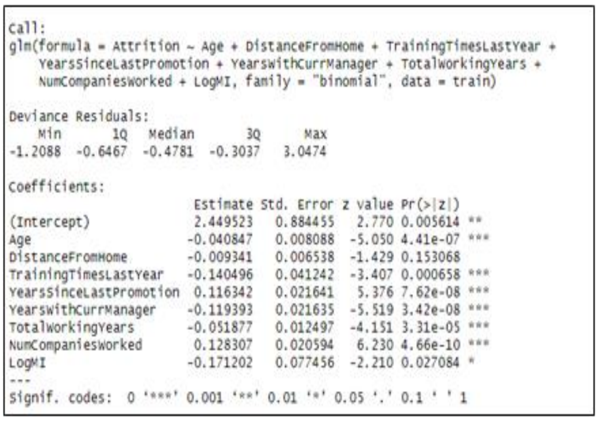
1. Initially Logistic Regression was performed on the 10 numeric variables as the technique considers ordinal and categorical variables as dummy variables.
2. Factor Analysis was performed on ordinal data and since pearson, kendall and spearman correlation were not much different, ordinal predictors were considered numeric. PCA was performed on the numeric and ordinal predictors. Logistic Regression was performed on these factors where they were treated as predictors.

The numeric variables are – Income, Age, Distance from home, % Salary Hike, Training Times Last Year, Years at Company, Years since last promotion, Years with current manager, Total working years, Number of companies worked.

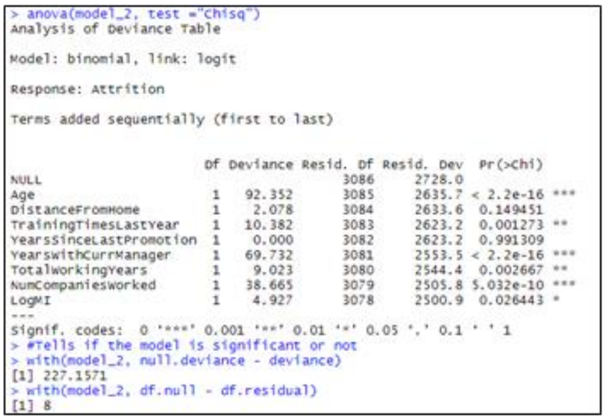
The ordinal variables are – Education Level, Job Involvement, Performance Rating, Job Level, Stock Option Level, Environmental Satisfaction, Job Satisfaction, Work-Life Balance.

Initial PCA:

Initial PCA was performed on the 10 numeric variables.

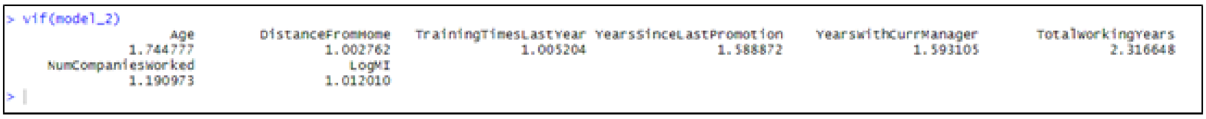


```
Call:
glm(formula = Attrition ~ Age + DistanceFromHome + TrainingTimesLastYear +
    YearsSinceLastPromotion + YearsWithCurrManager + TotalWorkingYears +
    NumCompaniesWorked + LogMI, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2088  -0.6467  -0.4781  -0.3037   3.0474

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              2.449523   0.884455   2.770 0.005614 **
Age                     -0.040847   0.008088  -5.050 4.41e-07 ***
DistanceFromHome        -0.009341   0.006538  -1.429 0.153068
TrainingTimesLastYear   -0.140496   0.041242  -3.407 0.000658 ***
YearsSinceLastPromotion  0.116342   0.021641   5.376 7.62e-08 ***
YearsWithCurrManager    -0.119393   0.021635  -5.519 3.42e-08 ***
TotalWorkingYears       -0.051877   0.012497  -4.151 3.31e-05 ***
NumCompaniesWorked       0.128307   0.020594   6.230 4.66e-10 ***
LogMI                   -0.171202   0.077456  -2.210 0.027084 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
> anova(model_2, test ="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Attrition

Terms added sequentially (first to last)


                        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                     3086     2728.0
Age                      1   92.352      3085     2635.7 < 2.2e-16 ***
DistanceFromHome         1    2.078      3084     2633.6  0.149451
TrainingTimesLastYear    1   10.382      3083     2623.2  0.001273 **
YearsSinceLastPromotion  1    0.000      3082     2623.2  0.991309
YearsWithCurrManager     1   69.732      3081     2553.5 < 2.2e-16 ***
TotalWorkingYears        1    9.023      3080     2544.4  0.002667 **
NumCompaniesWorked       1   38.665      3079     2505.8 5.032e-10 ***
LogMI                    1    4.927      3078     2500.9  0.026443 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #Tells if the model is significant or not
> with(model_2, null.deviance - deviance)
[1] 227.1571
> with(model_2, df.null - df.residual)
[1] 8
```

Log on the variable income was considered for this analysis. The data was split into training and test sets. Logistic Regression Model was created on the training set and its accuracy was checked on the test set. Stepwise feature selection technique with the AIC criterion was used to build the model with all the significant variables.



```
> vif(model_2)
                    Age        DistanceFromHome   TrainingTimesLastYear YearsSinceLastPromotion    YearsWithCurrManager       TotalWorkingYears
               1.744777                1.002762                1.005204                1.588872                1.593105                2.316648
     NumCompaniesWorked                   LogMI
               1.190973                1.012010
> |
```

None of the explanatory variables show any correlation with each other all of them have VIF less than 3.

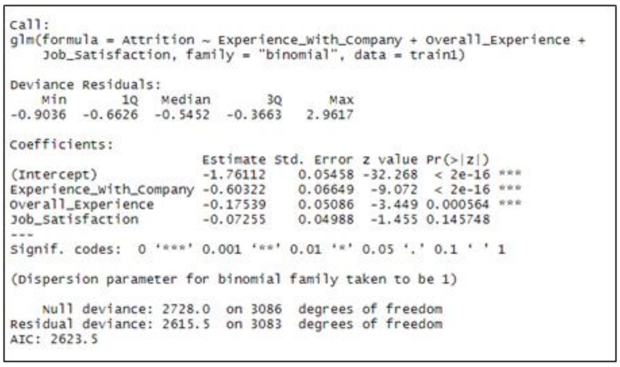
The full model contained 10 explanatory variables whereas the model created by stepwise feature selection created a model with 8 explanatory variables.

The model created by stepwise feature selection using only numeric variables has a chi-square of 227 with 8 degrees of freedom and p=value of less than 0.05. This is an indicator that the model is significant and fits better than a null model.

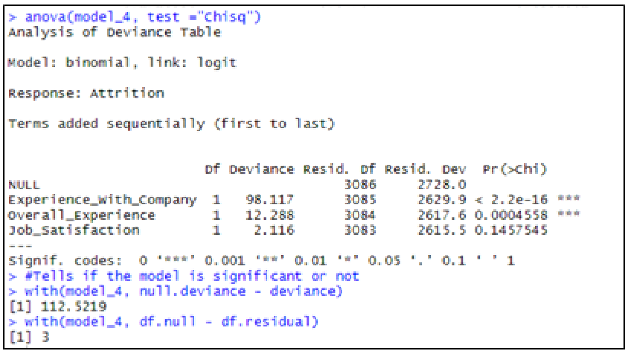
The model has an accuracy of 83.82% of the test set.

However, the above model only used numeric predictors whereas the HR analytics dataset has ordinal and categorical variables as well. Ordinal factor analysis was performed on the ordinal data and it implied that ordinal variables can be used as numeric variables. PCA was performed on this data and the resulting factors were used to perform logistic regression.

The new explanatory variables (factors) were – Experience with company, Overall Experience, Job Level, Job Satisfaction. The data was split into train and test set.

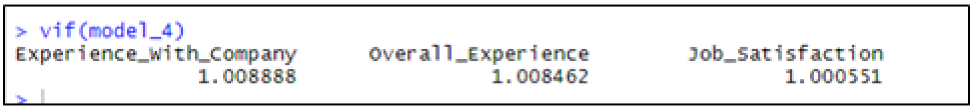


```
Call:
glm(formula = Attrition ~ Experience_With_Company + Overall_Experience +
    Job_Satisfaction, family = "binomial", data = train1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9036  -0.6626  -0.5452  -0.3663   2.9617

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.76112    0.05458 -32.268  < 2e-16 ***
Experience_With_Company -0.60322    0.06649  -9.072  < 2e-16 ***
Overall_Experience      -0.17539    0.05086  -3.449 0.000564 ***
Job_Satisfaction        -0.07255    0.04988  -1.455 0.145748
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2728.0  on 3086  degrees of freedom
Residual deviance: 2615.5  on 3083  degrees of freedom
AIC: 2623.5
```



```
> anova(model_4, test ="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Attrition

Terms added sequentially (first to last)


                        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                     3086     2728.0
Experience_With_Company  1   98.117      3085     2629.9 < 2.2e-16 ***
Overall_Experience       1   12.288      3084     2617.6 0.0004558 ***
Job_Satisfaction         1    2.116      3083     2615.5 0.1457545
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #Tells if the model is significant or not
> with(model_4, null.deviance - deviance)
[1] 112.5219
> with(model_4, df.null - df.residual)
[1] 3
```

Although the full model has 4 explanatory variables (factors), the model created by stepwise feature selection technique has 3 variables.



```
> vif(model_4)
Experience_With_Company      Overall_Experience        Job_Satisfaction
               1.008888                1.008462                1.000551
>
```

There is no multicollinearity in the data as the VIF is very low.

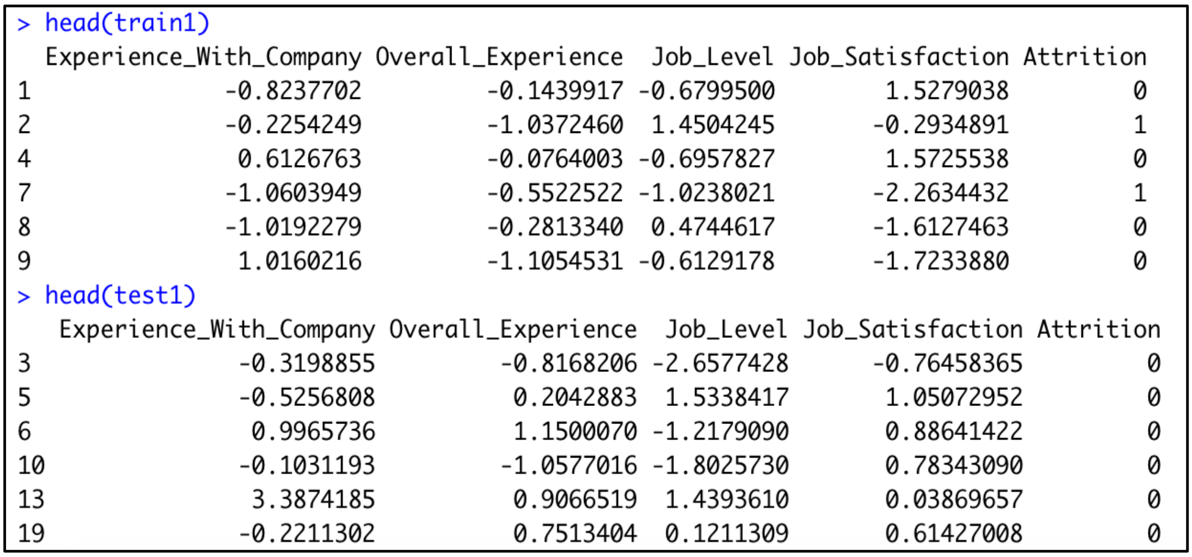
The model created by stepwise feature selection using numeric + ordinal factors from PCA has a chi-square of 112 with 3 degrees of freedom and p=value of less than 0.05. This is an indictor that the model is significant and fits better than a null model. The accuracy of the model on the test set was 83.9%.

The data is more suited for logistic regression than for linear regression. Some important conclusions can be drawn from the logistic regression – Experience with company (includes years at company, years since last promotion, years with current manager) plays an important role in attrition of employees at a company. HR executives can monitor employee-manager relationships as well as periodically check if all the employees are getting due promotion/recognition to make sure they are not leaving the company because of these reasons.

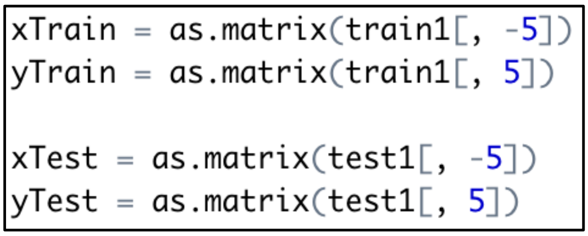
Lasso Regression for logistic with factor data

Lasso logistic model was performed since Lasso is capable of providing feature selection. Under the assumption that the nature of the data is suitable for performing Lasso logistic model, Lasso logistic model was performed on new variables: Experience_With_Company, Overall_Experience, Job_Level, and Job_Satisfaction. Lasso has only selected Experience_With_Company, which contains Years at Company, Years since last promotion, Years with current manager, and Total working years. This has aligned with our previous finding that these four variables are correlated with each other, thus they also been selected by Lasso as strong predictor for Attrition.

“glmnet” package in R is a hybrid between LASSO regression and Ridge regression. By setting α=1, a pure Lasso model was performed on the new variables – Experience_With_Company, Overall_Experience, Job_Level, and Job_Satisfaction, with parameter of interest being *Attrition*.

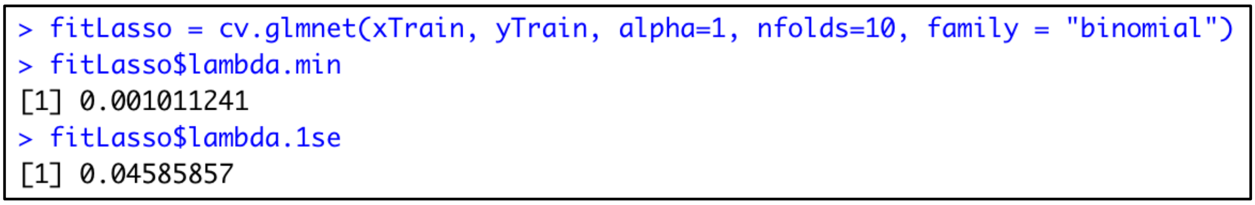


```
> head(train1)
  Experience_With_Company Overall_Experience  Job_Level Job_Satisfaction Attrition
1              -0.8237702         -0.1439917 -0.6799500        1.5279038         0
2              -0.2254249         -1.0372460  1.4504245       -0.2934891         1
4               0.6126763         -0.0764003 -0.6957827        1.5725538         0
7              -1.0603949         -0.5522522 -1.0238021       -2.2634432         1
8              -1.0192279         -0.2813340  0.4744617       -1.6127463         0
9               1.0160216         -1.1054531 -0.6129178       -1.7233880         0
> head(test1)
   Experience_With_Company Overall_Experience  Job_Level Job_Satisfaction Attrition
3               -0.3198855         -0.8168206 -2.6577428      -0.76458365         0
5               -0.5256808          0.2042883  1.5338417       1.05072952         0
6                0.9965736          1.1500070 -1.2179090       0.88641422         0
10              -0.1031193         -1.0577016 -1.8025730       0.78343090         0
13               3.3874185          0.9066519  1.4393610       0.03869657         0
19              -0.2211302          0.7513404  0.1211309       0.61427008         0
```



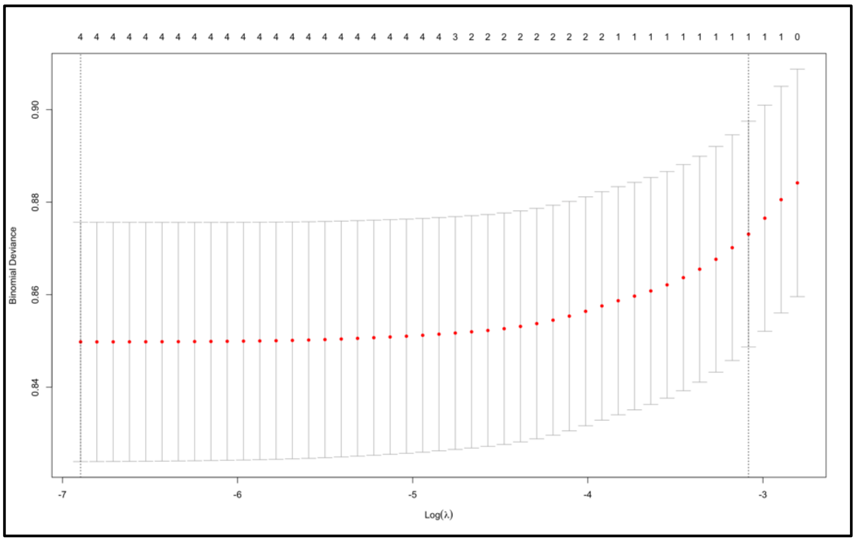
```
xTrain = as.matrix(train1[, -5])
yTrain = as.matrix(train1[, 5])

xTest = as.matrix(test1[, -5])
yTest = as.matrix(test1[, 5])
```

Going forward, a separation of the X's and Y's for training and test set as matrices are performed, in order to validate the result later on.



```
> fitLasso = cv.glmnet(xTrain, yTrain, alpha=1, nfolds=10, family = "binomial")
> fitLasso$lambda.min
[1] 0.001011241
> fitLasso$lambda.1se
[1] 0.04585857
```

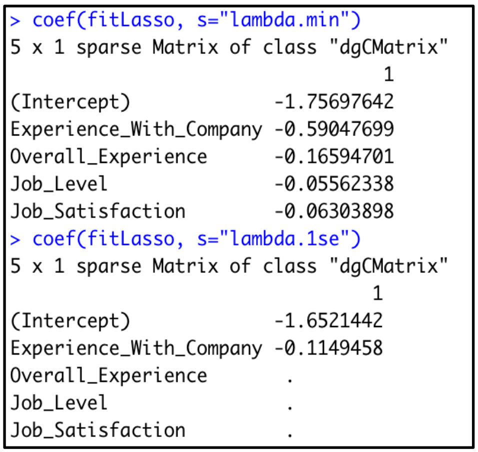
Since the interest of ours is in logistic regression, “binomial” has been set, with the number of cross-validation being 10. The lambda.min is computed as 0.001011241 while the lambda.1se is computed as 0.04585857.



In addition, as shown at the left, we can visualize the plot of Binomial Deviance versus Log(lambda). The minimum of the Log(lambda) and the Log(lambda) within 1 standard error are both denoted at the plot. If we transformed the values back from Log, we would get lambda.min and lambda.1se, which is 0.001011241 and 0.04585857, respectively. Using *lambda.min* as our lambda does not provide us any model selection since all of the predictors are included in the model, as shown at the right.

However, using *lambda.1se* as our lambda does provide us a selected model. The predictor, *Experience_With_Company*, has been selected by Lasso logistic model. Note that Experience_With_Company is a combination of otiginal variables: Years at company, Years since last promotion, Years with current manager, and Total working years.

The RMSE values we compute after running the R commands above are as follow:



```
> coef(fitLasso, s="lambda.min")
5 x 1 sparse Matrix of class "dgCMatrix"
                                  1
(Intercept)             -1.75697642
Experience_With_Company -0.59047699
Overall_Experience      -0.16594701
Job_Level               -0.05562338
Job_Satisfaction        -0.06303898
> coef(fitLasso, s="lambda.1se")
5 x 1 sparse Matrix of class "dgCMatrix"
                                 1
(Intercept)             -1.6521442
Experience_With_Company -0.1149458
Overall_Experience       .
Job_Level                .
Job_Satisfaction         .
```

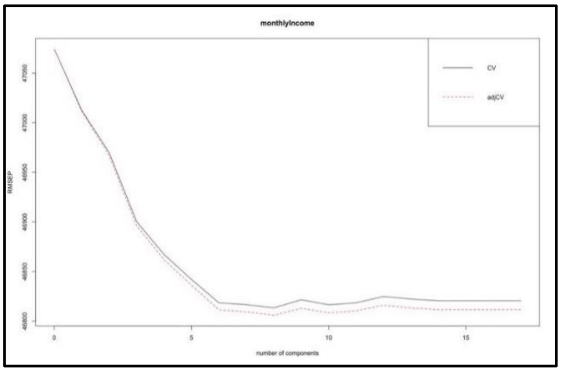
	Lasso Logistic (Lambda.min)	Lasso Logistic (Lambda.1se)
RMSE for Training Set	2.032248	1.850809
RMSE for Test Set	2.024971	1.849363

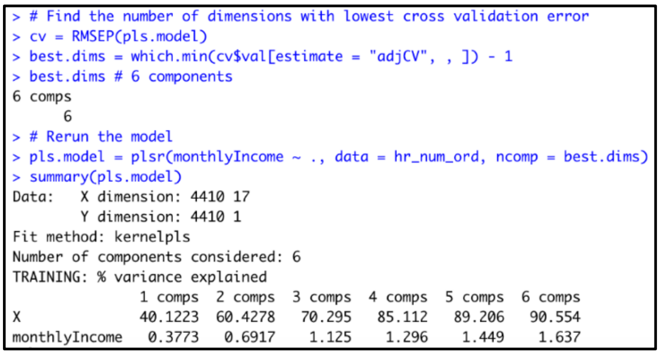
Overall, Lasso logistic model with lambda being *lambda.1se* has provided us a better performed model, the RMSE for the test set is significantly lower (1.85). In addition, using *lambda.1se* has provided us a more parsimonious model, which is desired. As a conclusion for Lasso Logistic model, *Experience_With_Company* (Years at company, Years since last promotion, Years with current manager, and Total working years) has relatively stronger effect on *Attrition*.

With the largest value of lambda such that error is within 1 standard error of the minimum (lambda.1se), Experience_With_Company has been selected by Lasso logistic model as the strongest predictor among *Experience_With_Company, Overall_Experience, Job_Level, and Job_Satisfaction*. Experience_With_Company contains Years at Company, Years since last promotion, Years with current manager, and Total working years, which proves that these four variables are having stronger correlations as well as stronger effect on the parameter of interest, *Attrition*. The results of Lasso logistic provide us a different approach to evaluate the feature importance.

Partial Least Squares

Similar with Principal Component Analysis, Partial Least Square is another dimension reduction technique. However, PCA does not allow us to predefine an outcome variable, while Partial Least Square allows us to do so. As a result, consider the nature of our data, under the assumption that Partial Least Square would potentially be more beneficial than Ordinary Least Square and Principal Component Analysis, *Monthly Income* was chosen to be our parameter of interest in Partial Least Square. Same with Ordinary Least Square, ordinal variables were treated as numeric variables and have been used in Partial Least Square, along with numeric variables. As a result, Partial Least Square does not seem to be more suitable for the data more than Ordinary Least Square does. Nevertheless, three positive predictors and two negative predictors were selected by partial least square, and the result was used in comparison with the result from automate model selection. *Training Time Last Year*, *Years Since Last Promotion*, and *Job Level* were selected by both techniques, which signaled that these three predictors are likely to have relatively strong influence on *Monthly Income*.

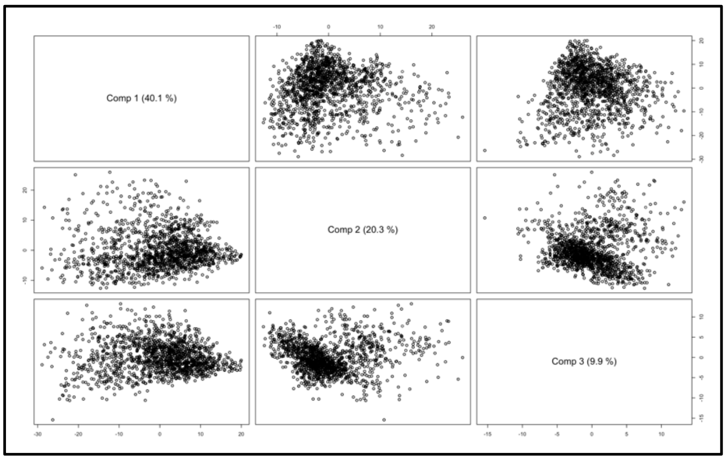




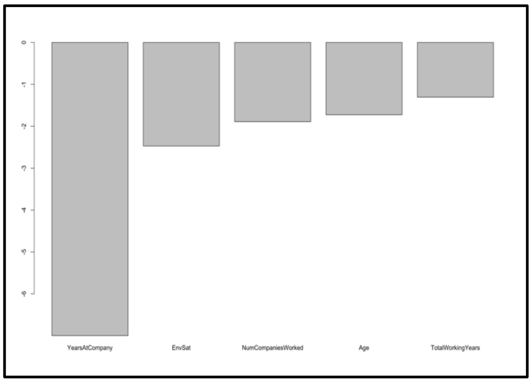
```
> # Find the number of dimensions with lowest cross validation error
> cv = RMSEP(pls.model)
> best.dims = which.min(cv$val[estimate = "adjCV", , ]) - 1
> best.dims # 6 components
6 comps
      6
> # Rerun the model
> pls.model = plsr(monthlyIncome ~ ., data = hr_num_ord, ncomp = best.dims)
> summary(pls.model)
Data: 	X dimension: 4410 17
	Y dimension: 4410 1
Fit method: kernelpls
Number of components considered: 6
TRAINING: % variance explained
               1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X              40.1223  60.4278   70.295   85.112   89.206   90.554
monthlyIncome   0.3773   0.6917    1.125    1.296    1.449    1.637
```

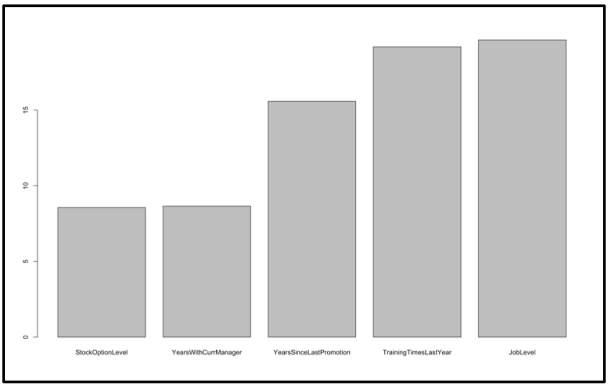
For Partial Least Square, cross-validation has been used with the purpose of finding the optimal number of dimensions. Optimally, the number of dimensions should be found with the lowest cross-validation error. However, the nature of our chosen dataset does not seem to fit Partial Least Square, since each time the algorithm returns different number of dimensions to us, which approximately range from 6 to 8 dimensions. However, according to the plot of Root Mean Square Error of Prediction (RMSEA) above, the number of components seems could be 6.

Similar with PCA, the first 2 dimensions capture the most variance. Our output from Partial Least Square shows that component 1 has captured 40.1% of the variance, while component 2 has captured 20.3% of the variance. Together, 2 components have captured 60.4% of the variance.



To extract the meaningful information from Partial Least Square, the regression coefficients have been sorted and produced to find out the relative importance to *Monthly Income*.





Partial Least Square shows that *Job Level*, *Training Times Last Year*, *Year Since Last Promotion* are the top 3 positive predictors to *Monthly Income*. On the other hand, *Years At Company* and *Environmental Satisfaction* are the top 2 negative predictors to *Monthly Income*.

Automate Model Selection versus Partial Least Square

The following table present the comparison between the variables that selected by Forward, Backward, and Stepwise versus the relative important variables that suggested by Partial Least Square.

Technique	Automate Model Selection	Partial Least Square
Variables Selected	*Training Time Last Year* *Years Since Last Promotion Job Level* *Stock Option Level* *Work Life Balance* *Total Working Years*	*Job Level* *Training Times Last Year* *Year Since Last Promotion* *Years At Company Environmental Satisfaction*

Training Time Last Year, *Years Since Last Promotion*, and *Job Level* have been selected in both techniques. These 3 variables seem to be having the strongest positive influence on *Monthly Income*.

For Partial Least Square, cross-validation was used to find the ideal number of dimensions. However, each time different number of the dimension were selected, which represent that the data under analysis seems to be not suitable for Partial Least Square. Nevertheless, the number of dimensions were selected based on the plot of Root Mean Square Error of Prediction (RMSEA) – 6 components were selected. Partial Least Square indicates that the employees' *Job Level*, *Training Times Last Year*, *Year Since Last Promotion* are having positive relationship with their monthly income. On the other hand, if employees' years at company or their environmental satisfaction increase, their monthly income somehow tend to decrease. Furthermore, the variables that selected by both PLS and automate model selections are identified, which benefit the overall project significantly. Employees' training time last year, years since last promotion, and their job level in the company are identified by both techniques as three most strongly positive variables, in relationship with their monthly income.

Conclusion

From the initial model selection methods, *Training Time Last Year*, *Years Since Last Promotion*, *Job Level*, *Stock Option Level*, *Work Life Balance*, and *Total Working Years* seem to have relatively stronger influence for *Monthly Income*, as these variables are all been selected by Forward, Backward, and Stepwise Selection. Automate model selection also revealed that *Total Working Years* seems to have negative effect on *Monthly Income*, if the number of total working years increase, the monthly income tends to decrease. On the other hand, these feature selection methods also suggest that if the years since last promotion increase, the monthly income tends to increase as well. As for Ordinary Least Square, a very low percentage of the variance within *Monthly Income* has successfully been captured by the model, even after log transformation. As a limitation, Ordinary Least Square does not benefit the analysis much – still too much variance that has not been captured. However, the results from automate model selection can be used to compare the results from other techniques used in this project.

It was clear from the principal component analysis using continuous variables that it is difficult to analyze the parameters of interest with just the continuous variables. It was also clear that the ordinal variables in the data should be dealt differently and the categorical variable must be dealt as well. Hence, the next steps were to analyze the ordinal variables using the Spearman, Kendall and Pearson correlation techniques and correspondence analysis on the categorical variables. The ordinal factor analysis was performed primarily with the following goals: 1.Find if there are any correlations between the 8 ordinal features using Pearson, Spearman and Kendall methods 2. Perform factor analysis on the ordinal data to see if any meaningful groupings are identified 3.Combine ordinal and numeric features to do PCA. Use the factor data from PCA to do OLS and logistic regression 4. Use "Hetcor" to find correlations between all kinds of features: numeric, categorical and ordinal.

Correlation analysis using Pearson, Spearman and Kendall methods gave the same results. There were absolutely no correlations between the ordinal features. Using these uncorrelated ordinal features to perform PCA/factor analysis doesn't make any sense and it features the core meaning of PCA. However, when factor analysis was performed on ordinal features using Spearman method, it gave surprisingly some meaningful groupings. Keeping this information aside for future use, correlation of ordinal features with dependent features Attrition and Monthly Income was studied. There were only 4 features that were significant having p-value <0.05. Significant feature information shown below:

- EnvSat, JobSat and WrkLifBal were significantly correlated with Attrition
- JobLevel was significantly correlated with Monthly Income

The above significant features were treated as numeric features to perform PCA. The factor data from the PCA were used for OLS and Logistic regression to see if it produced any different results. Cluster analysis on the ordinal data did not produce any meaningful results as the nature of the data was not suited for cluster analysis.

On the OLS regression model (built using PCA factor data), feature selection was conducted using forward, backward and stepwise methods. All the 3 selected same feature and significance. The results were meaningful knowing the business.

If a company HR would want to understand what influences the monthly income of an employee, they can investigate the below feature groupings:

- **Experience with company**: Years at company, years since last promotion, year with current manager and total working years. All are positive with years at company being highest. Longer experience with company or current manager is important have a higher monthly income

- **Overall Experience**: Age, total working years and number of companies worked. All are positive age and number of companies worked being highest. Higher overall experience in their career, lesser chance of that employee having higher monthly income. This indicates, these employees are not just income driven and other factors matters to them to have a healthy lifestyle.

- **Environment Satisfaction**: Distance from home and environment satisfaction being positive and job level being negative.

- **Job Satisfaction**: Higher monthly income doesn't necessarily mean the employees are satisfied; employees may be over stressed delivering the results and meeting the deadlines to have good performance rating.

These factors were later used in logistic regression and it was found that the factor *Experience with company* (includes *years at company*, *years since last promotion*, *years with current manager*) play an important role in attrition of employees at a company. HR executives can monitor employee-manager relationships as well as periodically check if all the employees are getting due promotion/recognition to make sure they are not leaving the company because of these reasons. It was also found that the data was suitable for the logistic regression problem than a linear regression on predicting the *MonthlyIncome.*

Further Partial Least Squares method was used in the analysis. Similar to Principal Component Analysis, Partial Least Square is another dimension reduction technique. However, PCA does not allow to predefine an outcome variable, while Partial Least Square allows to do so. As a result, considering the nature of data, under the assumption that Partial Least Square would potentially be more beneficial than Ordinary Least Square and Principal Component Analysis, *Monthly Income* was chosen to be the parameter of interest in Partial Least Square. Same with Ordinary Least Square, ordinal variables were treated as numeric variables and have been used in Partial Least Square, along with numeric variables. As a result, Partial Least Square does not seem to be more suitable for the data more than Ordinary Least Square does. Nevertheless, three positive predictors and two negative predictors were selected by partial least square, and the result was used in comparison with the result from automate model selection. *Training Time Last Year, Years Since Last Promotion,* and *Job Level* were selected by both techniques, which signaled that these three predictors are likely to have relatively strong influence on *Monthly Income*.

Overall, it was found that the data was tailored to predict the attrition rate of the company rather than predicting the Monthly Income of the employee. This was proven during various stages if the analysis. Further analysis revealed that the four factors could be used to capture 50% variance and could simplify the interpretation of the variables. Hence these factors were used in the model building process. Logistic regression model build using the factors produced an accuracy of 83.9% and it was also found that one of the factor *Experience with company* (includes *years at company*, *years since last promotion*, *years with current manager*) play an important role in attrition of employees at a company.

Future Work

In the analysis, the categorical variables were never used in the model building process. This is the limitation of the project. Since there were 7 categorical variables in the dataset and given the high association of these variables on the parameter of interest, the inclusion of these variables would have boosted the model performance. This would be one limitation of the work. In future work, the categorical variables combined with the factors would be used in the predictive model.

One other future work would be canonical correlation. Since the data set had two parameters of interest – Monthly Income and Attrition Rate, canonical correlation would be one way to analyze multiple response variables with the independent variables simultaneously. These would be the interesting future work in the project.

References

Forbes. 9 May 2019. <https://www.forbes.com/sites/johnhall/2019/05/09/the-cost-of-turnover-can-kill-your-business-and-make-things-less-fun/#ec0915b79437>.

Kaggle. n.d. <https://www.kaggle.com/vjchoudhary7/hr-analytics-case-study#general_data.csv>.

Appendices

Appendix 1 – Final Individual Report

Andy Huang

At the stage of initial analysis, I have used both ordinal and categorical variables to perform Ordinary Least Square. Since there is not much correlation between the ordinal variables, so these ordinal variables are treated as numeric variables. However, based on the results of the R-square (very low, about 1%), the Normal Q-Q Plot, and the histogram of the residuals, a log transformation is identified as needed for Ordinary Least Square. As a result, I also applied log transformation on the dependent variable, *Monthly Income*.

In addition, Forward, Backward, and Stepwise Model Selection have also been performed to allow our group to gain a better understanding as which variables have relatively stronger effects on the dependent variable, *Monthly Income*. Forward, Backward, and Stepwise selections all provide a unified result that *Training Time Last Year*, *Years Since Last Promotion*, *Job Level*, *Stock Option Level*, *Work Life Balance*, and *Total Working Years* are significant predictors for *Monthly Income*, with *Total Working Years* being the only negative predictor. In addition, *Years Since Last Promotion* has the strongest positive effect on *Monthly Income*.

For exploring relationships among categorical variables, Arun and I decided to analyze the categorical variables of our dataset by performing correspondence analysis, with the hope to reveal insights hidden between each categorical variable and one of our dependent variables, *Attrition*, which is also categorical. Arun and I each took 3 categorical variables and pair each categorical variable with *Attrition* and performed correspondence analysis. Moreover, Arun and I also further explored the categorical variables by replacing *Attrition* with other categorical variables that we are interested in exploring. Correspondence analysis revealed interesting findings such as HR department has a higher attrition rate, and employees with HR degree tend to leave the company.

Furthermore, I also performed Lasso logistic model since Lasso is capable of providing feature selection. Lasso logistic model was performed on our new variables: *Experience_With_Company*, *Overall_Experience*, *Job_Level*, and *Job_Satisfaction*. The finding from Lasso logistic has aligned with our previous finding that *Years at Company*, *Years since last promotion*, *Years with current manager*, and *Total working years* are significant predictors *for Attrition*.

Lastly, I performed Partial Least Square Regression with the hope to gain interesting finding for comparison with the results from previous techniques. Partial Least Square does not seem to be more suitable for the data more than Ordinary Least Square does, since each time the algorithm returns a different number of dimensions to us, which approximately range from 6 to 8 dimensions. Nevertheless, according to the plot of Root Mean Square Error of Prediction (RMSEA), 6 components were identified. Partial Least Square shows that *Job Level*, *Training Times Last Year*, *Year Since Last Promotion* are the top 3 positive predictors to *Monthly Income*. Moreover, *Years at Company* and *Environmental Satisfaction* are the top 2 negative predictors to *Monthly Income*. This findings were used in comparison with the result from automate model selection, and *Training Time Last Year*, *Years Since Last Promotion,* and *Job Level* were selected by both techniques, which signaled that these three predictors are likely to have relatively strong influence on *Monthly Income*.

Arun Gopal

We, as a team worked on the HR analytics data set found from Kaggle. As mentioned in the milestone 2, our data set consists of a mixture of variables – 8 categorical, 8 ordinal and 9 numeric variables. We explored principal component analysis, ordinary least squares, and logistic regression on the numeric analysis, ordinal analysis on the ordinal variables, and correspondence analysis on the categorical variables. Further, we explored methods such as canonical correlation and partial least squares.

The team worked together on different modules. Everyone had a chance of working on all the aspects covered in our analysis at least once. However, my significant contribution was on the principal component analysis (PCA) on the continuous variables, then building on the analysis to a principal factor analysis (PFA). Later on in the project, the PFA was performed on combined set of continuous and ordinal variables. Factor analysis was a huge part of my contribution towards the goal of the project. Similarly, I helped the team with correspondence analysis to understand the association between the categorical variables in the data.

Apart from PCA, PFA and correspondence analysis, I was involved in several exploratory analysis and initial model building. Initial model building includes simple ordinary least squares method, stepwise linear regression models, logistic regression model and linear discriminant model.

Detailed analysis on the correspondence analysis was submitted during one of the milestones. The report included a 7-page detailed analysis with mosaic plots, chi-square tests and correspondence analysis using the library “ca”. Some of the highlights of the analysis can be seen below.

Correspondence Analysis:

We have 7 categorical variables in the dataset including the parameter of interest. Andy and I decided to further divide the task into two. The categorical variables that Andy and I analyzed are as follows:

- Andy - Attrition versus Business Travel (3 levels: Non-Travel/Travel rarely/Travel Frequently)
- Andy - Attrition versus Department (3 levels: Human Resources/Research & Development/ Sales)
- Andy - Attrition versus Education Field (6 levels: Human Resources/Life Sciences/Marketing/Medical/Technical Degree/Other)
- Arun – Attrition versus Gender (2 levels: Male/Female)
- Arun – Attrition versus Job Role (9 levels: Healthcare Representative/Human Resources/Laboratory Technician/Manager/Manufacturing Director/Research Director/Research Scientist/Sales Executive/Sales Representative)
- Arun – Attrition versus Marital Status (3 levels: Married/Single/Divorced)

In addition to our analysis of the categorical variables against the parameter of interest, we also analyzed the following:

- Andy – Business Travel versus Department
- Arun – Education Field versus Job Role

Of the 3 variables that I analyzed against attrition; gender was concluded to have no association/ correlation with the response variable attrition. In our analyzes, we found that the unmarried and single employees had a high attrition rate as opposed to the married and divorced workforce. Similarly, in our analysis of job role against attrition, we found that the research directors had a high attrition rate as opposed to manufacturing director who tend to have low attrition rate. Using the correspondence analysis, various interesting association between the categorical variables were found which is included in the detailed report.

Principal Factor Analysis:

The continuous variables in the data set were initially analyzed using PCA techniques discussed in class. The initial PCA resulted in 4 components capturing 70% of variance in the data. The factors had contribution from continuous variables only. Since we had ordinal variables in the dataset, these ordinal variables were analyzed separately by Shweta and Sahana and found that the ordinal variables had no correlation between them. Hence, these ordinal variables were further used in the principal factor analyzes. Overall, we had 17 numeric variables in our dataset. With PFA, 4 factors were able to capture 50% of the variance in the data. These factors were named based on the contributions of the variables. Detailed report is attached to the final report. The conclusions from the factors are as follows.

Based on the loadings of these factors, the factors were named as follows.

- Factor 1 - Experience. With. Same. Company
- Factor 2 - Overall. Experience
- Factor 3 - Job. Satisfaction
- Factor 4 - Environmental. Satisfaction

In conclusion, these factors were intended to be used in further analysis such as logistic regression, linear discriminant analysis and partial least square methods.

Shweta Gujrathi

For the data exploration, I specifically worked on understanding data spread for the 2 response variables – monthly salary (numeric) and attrition (binary). The data for monthly income is right skewed and we have almost 5 times more observations where the response for the attrition variable is 'yes' than the response 'no. Following which I also checked box plots for monthly income and salary hike which implied that attrition was higher in employees who fall have low monthly salaries but what was surprising was that attrition is also higher in employees who have higher % of salary hike. Additionally, out of the employees who left the company, 47% were in the 25-34 age bracket and 65% held lower job levels 1 & 2. The data exploration also showed that environment satisfaction and job satisfaction do not have much effect on attrition rate. After performing similar analysis for ordinal and categorical variables, I checked the correlation between the 9 numeric variables, out of which only 4 showed any correlation with each other – Years with current manager, Years since last promotion, Years at company, Percent Salary Hike.

We started data analysis on numeric variables with linear and logistic regression out of which I performed Logistic regression. We did not include the categorical and ordinal variables at this point since the above two regression techniques code categorical and ordinal variables as dummy variables. In logistic regression, I took a log of monthly income since its range was much higher than the rest of the numeric data. The chi-square of the logistic model showed that the model is significant. I also divided the data into train and test to check the prediction accuracy and the accuracy is 83%. The model has a chi-square of 227 with 8 df and p-value less than 0.05, proving it is a better fit than an empty model.

I also worked on the ordinal factor analysis with Sahana and the results showed that there is no correlation between the ordinal variables – Education level, Job Involvement, Performance rating, Job Level, Stock Option Level, Environment Satisfaction, Job Satisfaction and Work Life Balance. Performed Pearson, Spearman and Kendall correlation techniques for the ordinal variables. Since none of these techniques showed any correlation, the group decided that we will be treating the ordinal variables as numeric variables and will include in principal component analysis.

After PCA was performed numeric and ordinal variables, I again did logistic regression on the new factors – Experience with Company, Overall Experience, Job Satisfaction & Job Level. Out of which with feature selection technique, 3 factors remained in the final model – Experience with company, Overall Experience and Job Satisfaction.[6] The model had a p-value less than 0.05 thus proving its significance and it got an accuracy of 83% on the test set.

To understand the impact of experience with company which involves the variables, years with company, years with current manager and years after last promotion, I graphed the trends for these variables and found that out of the employees that left the company, the highest number left within 2 years of working with their current manager and within 2-3 years of joining the company.

In conclusion, experience with company, overall experience and job satisfaction plays an important role in whether an employee will leave the company or not. The company can use this insight to decrease attrition rate by monitoring promotions, employee relationship with their managers

Key Takeaways and Learning : There was a lot of learning involved in this project, not just from data analysis point of view but also skills such as collaborating with a group from different background, handling conflicting schedules and deadlines. From data analysis perspective, it was interesting to perform different analysis for different kinds of variables such as numeric, ordinal and categorical and then try to make sense of it all together.

Sahana Natraja

Summary:

I took up the task for doing Ordinal Factor Analysis, OLS with Numeric and Factor data and Hetcor correlation in this project. Goals, steps and processes followed for project analysis is described as follows. Our data consisted of 8 ordinal features; it was a big chunk of features. The goal behind the ordinal data analysis are:

- Find if there are any correlations between the 8 ordinal features using Pearson, Spearman and Kendall methods
- Perform factor analysis on the ordinal data to see if any meaningful groupings are identified
- Combine ordinal and numeric features to do PCA. Use the factor data from PCA to do OLS and logistic regression
- Use "Hetcor" to find correlations between all kinds of features: numeric, categorical and ordinal.

Initial step I took was to handle missing/NA values. 3 of the 8 ordinal features which were encoded as character data in the dataset. Those character features were converted to numeric features and Missing/NA values were replaced by "mode". Totally 8 ordinal features were available for analysis: **Education, JobInvolvement, PerformanceRating, JobLevel, StockOptionLevel , EnvironmentSatisfaction, JobSatisfaction , WorkLifeBalance**.

Correlation analysis using Pearson, Spearman and Kendall methods gave the same results. There were absolutely no correlations between the ordinal features. Using these uncorrelated ordinal features to perform PCA/factor analysis doesn't make any sense and it features the core meaning of PCA. However, when factor analysis was performed on ordinal features using Spearman method gave surprisingly some meaningful groupings. Keeping this information aside for future use, correlation of ordinal features with dependent features Attrition and Monthly Income was studied. There were only 4 features that were significant having p-value <0.05. Significant feature information shown below:

- **EnvironmentSatisfaction, JobSatisfaction** and **WorkLifeBalance** are highly significant with Attrition
- **JobLevel** is highly significant with Monthly Income

The above significant features were treated as numeric features to perform PCA. The factor data from the PCA were used for OLS and Logistic regression to see if it produced any different results. Cluster analysis on the ordinal data did not produce any meaningful results as the nature of the data was not suited for cluster analysis.

On OLS regression model (built using PCA factor data), feature selection was conducted using forward, backward and stepwise methods. All the 3 selected same feature and significance. The results were meaningful knowing the business.

"Hetcor" correlation analysis was conducted on all kinds of features: numeric, categorical and ordinal data. Features popped up in the correlation were same as the features shown in the PCA first grouping "Experience with Company". This was the interesting insight to know. This method gave more confidence in the analysis indicating it is the right direction for analysis.

If a company HR would want to understand what influences the monthly income of an employee, they can look into the below feature groupings:

- **Experience with company**: Years at company, years since last promotion, year with current manager and total working years. All are positive with years at company being highest.

- **Overall Experience**: Age, total working years and number of companies worked. All are positive age and number of companies worked being highest.
- **Environment Satisfaction**: Distance from home and environment satisfaction being positive and job level being negative.

Conclusion:

This project gave a good understanding on data analysis is not all about prediction and not all data are suited for prediction. This is one such case. There were a lot of scope for analysis in this project which consisted of numeric, categorical and ordinal features. Various advanced analytical techniques were used to analyze the dataset such as: OLS regression, Logistic regression, PCA, Ordinal Factor analysis, Hetcor correlations, Correspondence analysis, Partial Least Square analysis. The goal of the project was to use all the above techniques and it was successfully implemented in the analysis.

Interesting conclusion from the project for Monthly Income are:

- Longer experience with company or current manager is important have a higher monthly income
- Higher monthly income doesn't necessarily mean the employees are satisfied; employees may be over stressed delivering the results and meeting the deadlines to have good performance rating
- Higher overall experience in their career, lesser chance of that employee having higher monthly income. This indicates, these employees are not just income driven and other factors matters to them to have a healthy lifestyle.

HR in a company might want to investigate the above aspects to learn about what influence monthly income and how to keep the employees satisfied.

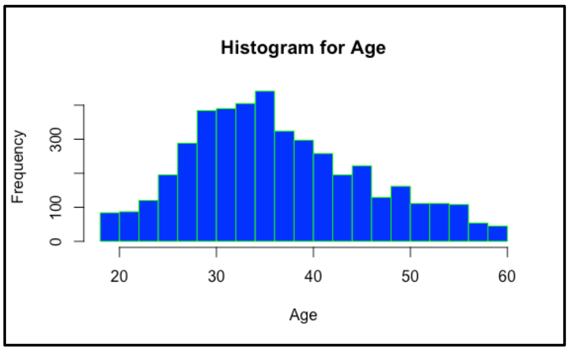
Project Key Takeway:

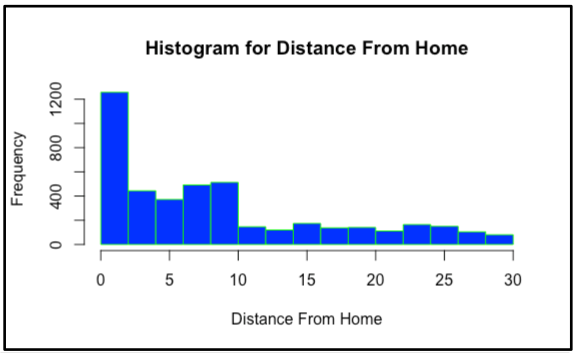
- Years at company Vs. Avg Monthly Income.
- Average Salary hike over the years in the company.
- Experience within company plays an important role in whether an employee will leave the company or not. The company can use this insight to decrease attrition rate by monitoring promotions, employee relationship with their managers.

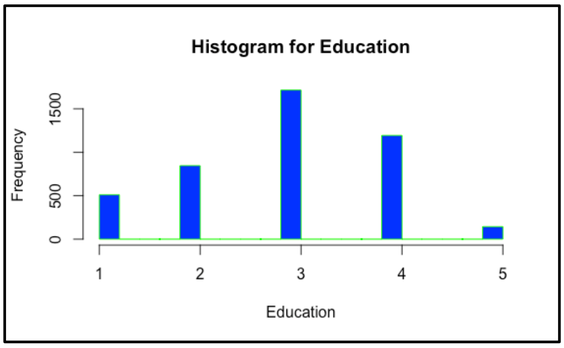
Learning:

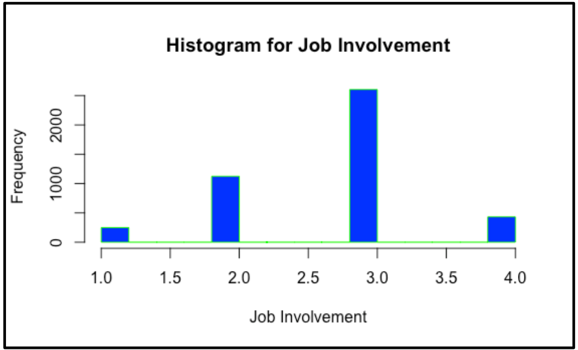
I learnt many advanced analytical techniques to analyze numeric, categorical and ordinal data. How to perform PCA and find the best components best explaining the data. This project also gave us an understanding that not all data is suited for prediction and how to focus more on data analysis part to find interesting insights. We are used Partial Least Square analysis on the data which was challenging to learn and implement. Overall, this class was very helpful in helping me analyze the data where prediction is not always the solution.

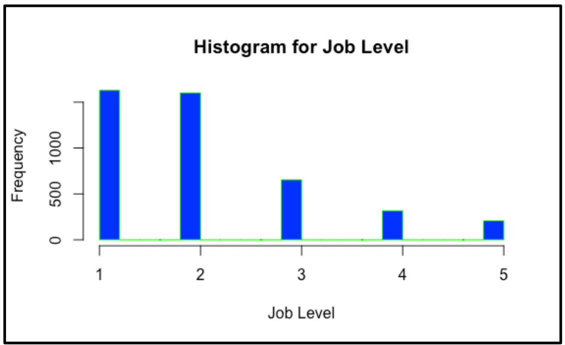
Appendix 2 – Supplementary Graphs

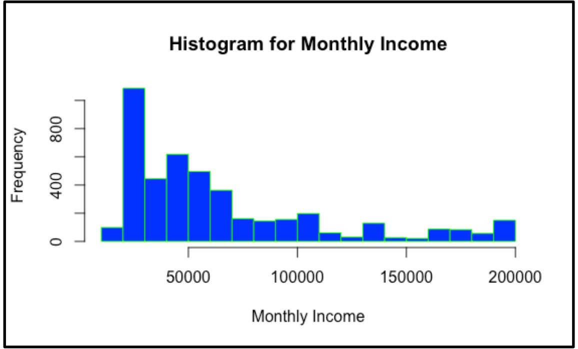


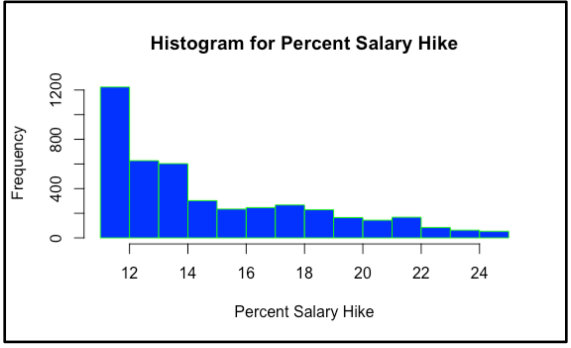


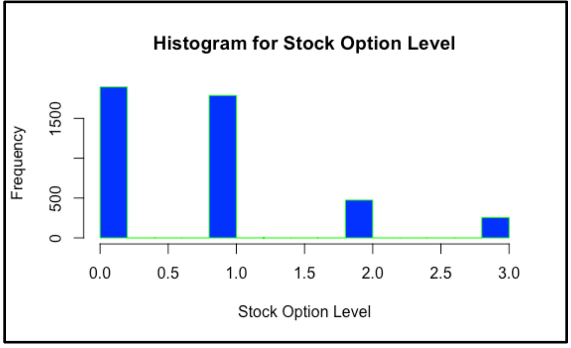


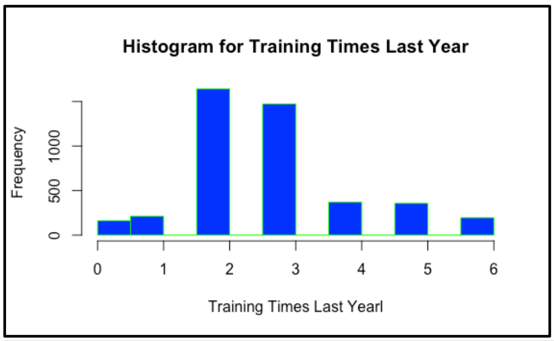


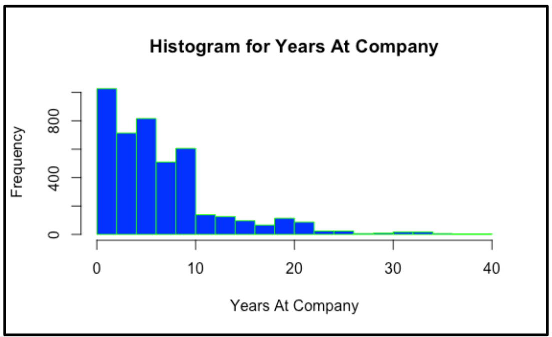


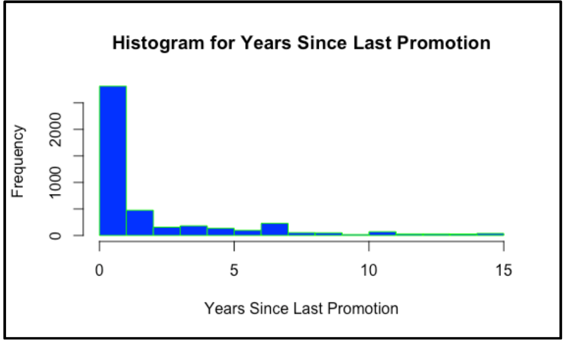


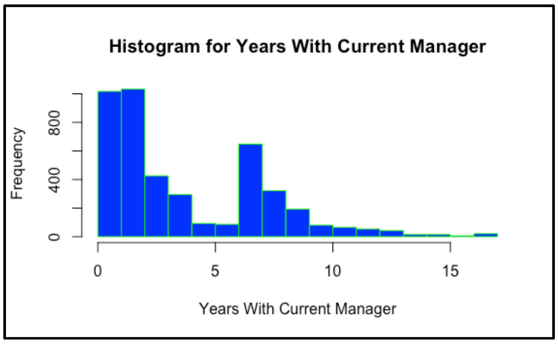


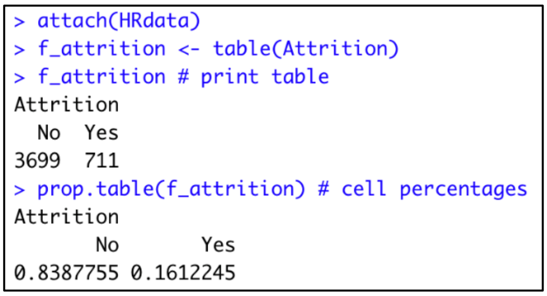




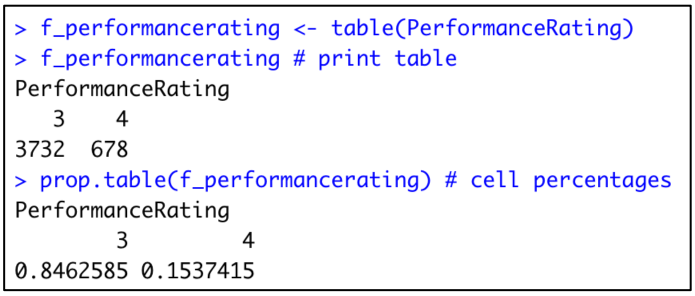




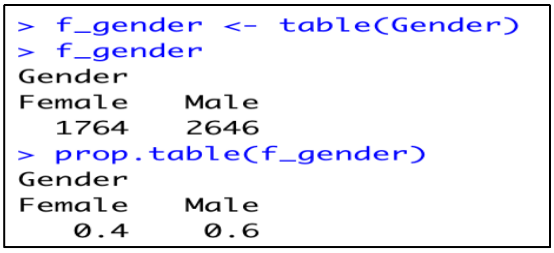




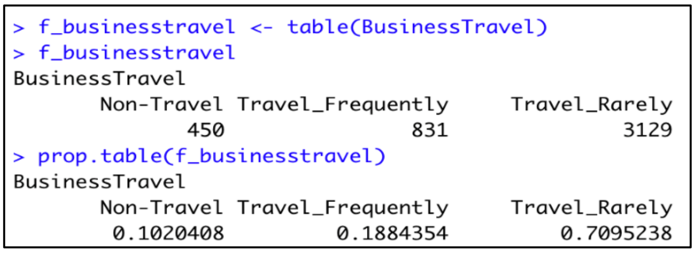
```
> attach(HRdata)
> f_attrition <- table(Attrition)
> f_attrition # print table
Attrition
  No  Yes
3699  711
> prop.table(f_attrition) # cell percentages
Attrition
       No       Yes
0.8387755 0.1612245
```



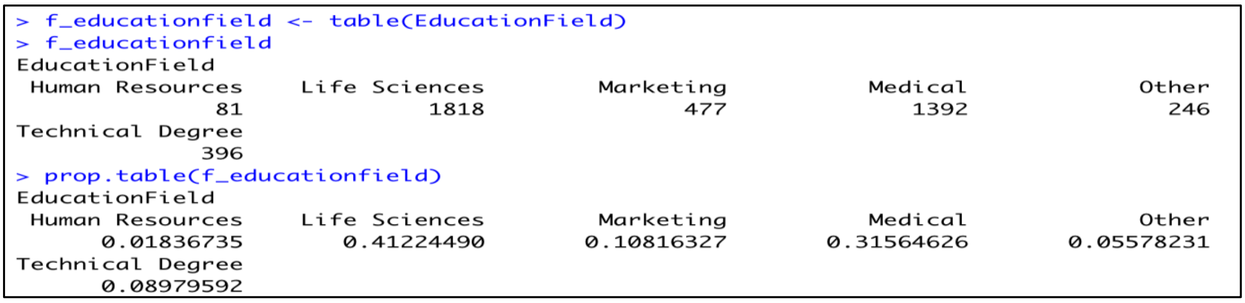
```
> f_performancerating <- table(PerformanceRating)
> f_performancerating # print table
PerformanceRating
   3    4
3732  678
> prop.table(f_performancerating) # cell percentages
PerformanceRating
        3         4
0.8462585 0.1537415
```



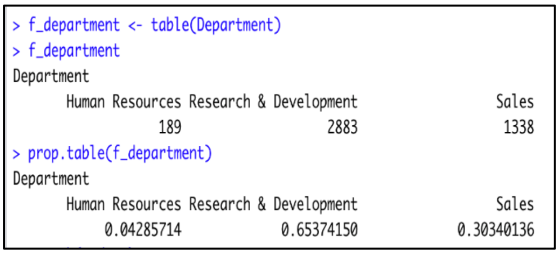
```
> f_gender <- table(Gender)
> f_gender
Gender
Female   Male
  1764   2646
> prop.table(f_gender)
Gender
Female   Male
   0.4    0.6
```



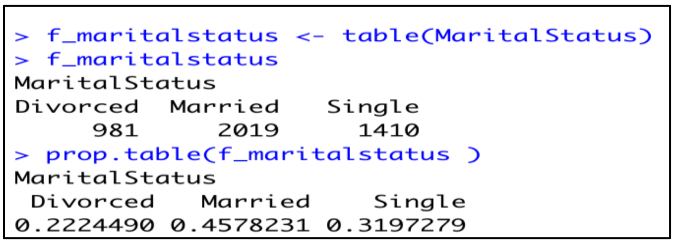
```
> f_businesstravel <- table(BusinessTravel)
> f_businesstravel
BusinessTravel
       Non-Travel Travel_Frequently     Travel_Rarely
              450               831              3129
> prop.table(f_businesstravel)
BusinessTravel
       Non-Travel Travel_Frequently     Travel_Rarely
        0.1020408         0.1884354         0.7095238
```



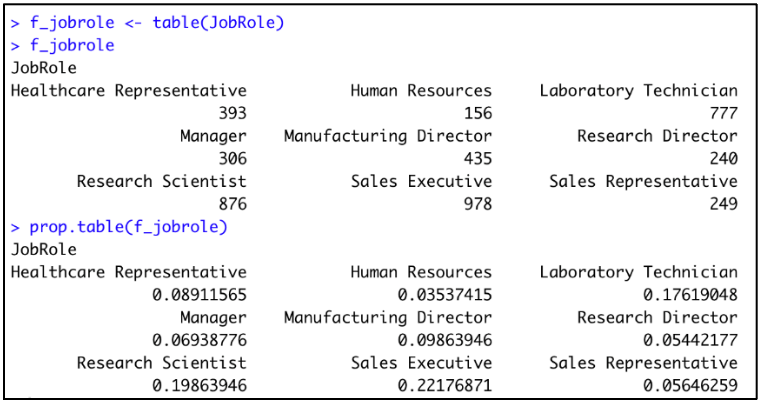
```
> f_educationfield <- table(EducationField)
> f_educationfield
EducationField
 Human Resources     Life Sciences         Marketing           Medical             Other
              81              1818               477              1392               246
Technical Degree
             396
> prop.table(f_educationfield)
EducationField
 Human Resources     Life Sciences         Marketing           Medical             Other
      0.01836735        0.41224490        0.10816327        0.31564626        0.05578231
Technical Degree
      0.08979592
```



```
> f_department <- table(Department)
> f_department
Department
       Human Resources Research & Development                  Sales
                   189                   2883                   1338
> prop.table(f_department)
Department
       Human Resources Research & Development                  Sales
            0.04285714             0.65374150             0.30340136
```

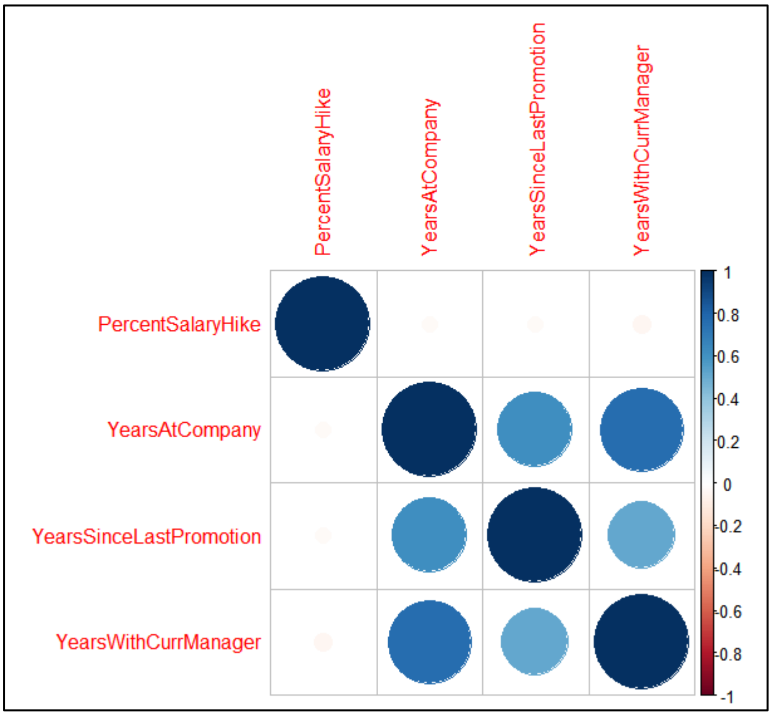


```
> f_maritalstatus <- table(MaritalStatus)
> f_maritalstatus
MaritalStatus
Divorced  Married   Single
     981     2019     1410
> prop.table(f_maritalstatus )
MaritalStatus
 Divorced   Married    Single
0.2224490 0.4578231 0.3197279
```

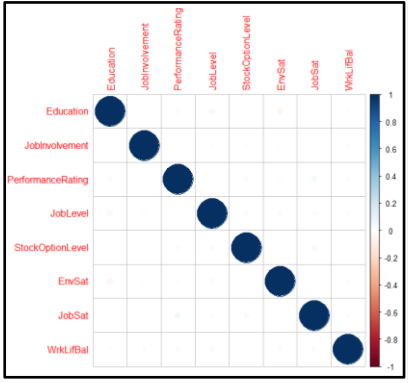


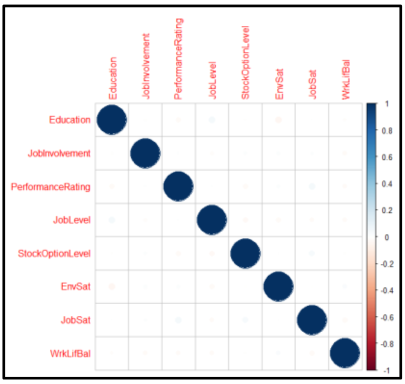
```
> f_jobrole <- table(JobRole)
> f_jobrole
JobRole
Healthcare Representative           Human Resources     Laboratory Technician
                      393                       156                       777
                  Manager    Manufacturing Director         Research Director
                      306                       435                       240
       Research Scientist           Sales Executive      Sales Representative
                      876                       978                       249
> prop.table(f_jobrole)
JobRole
Healthcare Representative           Human Resources     Laboratory Technician
               0.08911565                0.03537415                0.17619048
                  Manager    Manufacturing Director         Research Director
               0.06938776                0.09863946                0.05442177
       Research Scientist           Sales Executive      Sales Representative
               0.19863946                0.22176871                0.05646259
```

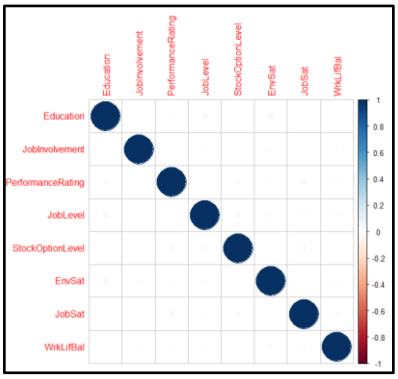
Correlation



Ordinal Factor Analysis

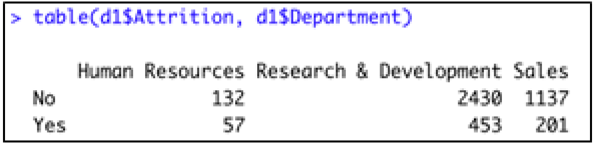




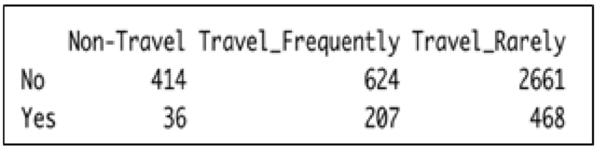


Correspondance Analysis

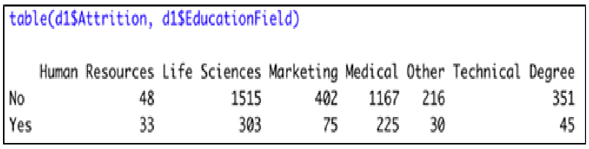
Contingency Table



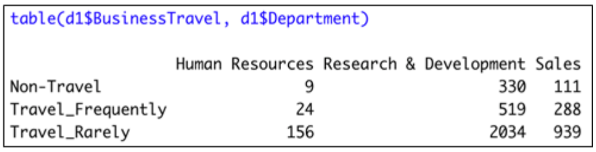
```
> table(d1$Attrition, d1$Department)

      Human Resources Research & Development Sales
  No              132                   2430  1137
  Yes              57                    453   201
```



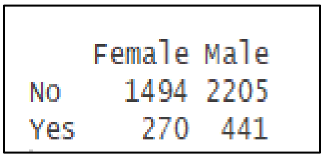
```
     Non-Travel Travel_Frequently Travel_Rarely
No          414               624          2661
Yes          36               207           468
```



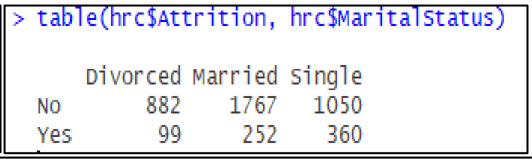
```
table(d1$Attrition, d1$EducationField)

      Human Resources Life Sciences Marketing Medical Other Technical Degree
No                 48          1515       402    1167   216              351
Yes                33           303        75     225    30               45
```



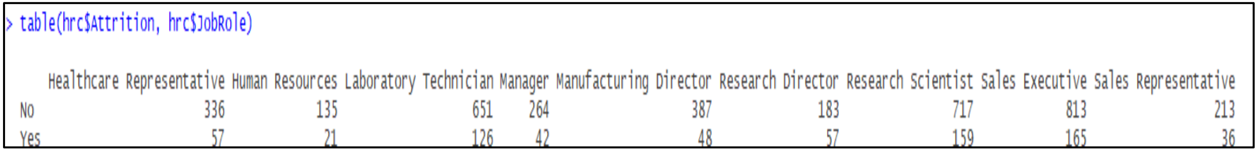
```
table(d1$BusinessTravel, d1$Department)

                  Human Resources Research & Development Sales
Non-Travel                      9                    330   111
Travel_Frequently              24                    519   288
Travel_Rarely                 156                   2034   939
```



```
     Female Male
 No    1494 2205
 Yes    270  441
```

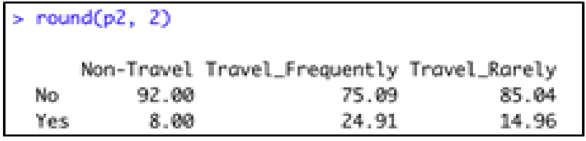


```
> table(hrc$Attrition, hrc$MaritalStatus)

      Divorced Married Single
  No       882    1767   1050
  Yes       99     252    360
```

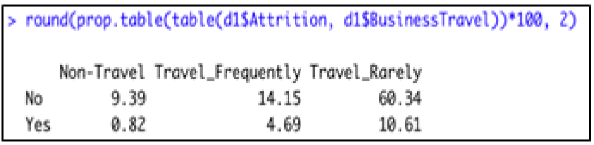


```
> table(hrc$Attrition, hrc$JobRole)

     Healthcare Representative Human Resources Laboratory Technician Manager Manufacturing Director Research Director Research Scientist Sales Executive Sales Representative
  No                       336             135                   651     264                    387               183                717             813                  213
  Yes                       57              21                   126      42                     48                57                159             165                   36
```

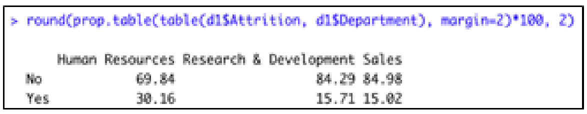
Get Percentages from Contingency Table



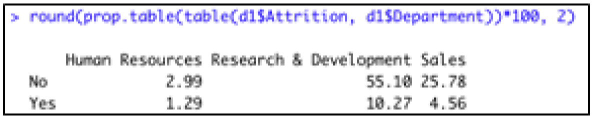
```
> round(p2, 2)

      Non-Travel Travel_Frequently Travel_Rarely
  No       92.00             75.09         85.04
  Yes       8.00             24.91         14.96
```



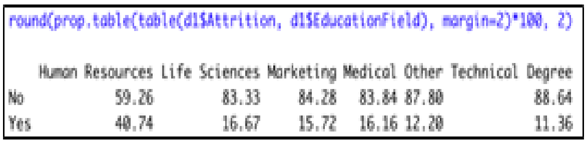
```
> round(prop.table(table(d1$Attrition, d1$BusinessTravel))*100, 2)

      Non-Travel Travel_Frequently Travel_Rarely
  No        9.39             14.15         60.34
  Yes       0.82              4.69         10.61
```



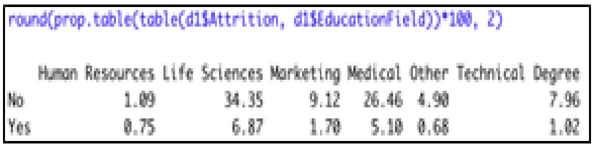
```
> round(prop.table(table(d1$Attrition, d1$Department), margin=2)*100, 2)

      Human Resources Research & Development Sales
  No            69.84                  84.29 84.98
  Yes           30.16                  15.71 15.02
```



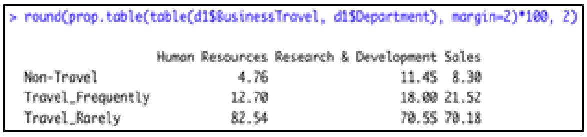
```
> round(prop.table(table(d1$Attrition, d1$Department))*100, 2)

      Human Resources Research & Development Sales
  No             2.99                  55.10 25.78
  Yes            1.29                  10.27  4.56
```



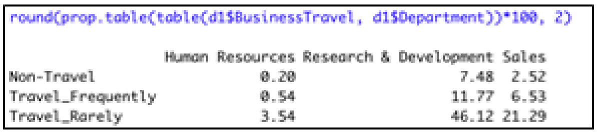
```
round(prop.table(table(d1$Attrition, d1$EducationField), margin=2)*100, 2)

      Human Resources Life Sciences Marketing Medical Other Technical Degree
No              59.26         83.33     84.28   83.84 87.80            88.64
Yes             40.74         16.67     15.72   16.16 12.20            11.36
```



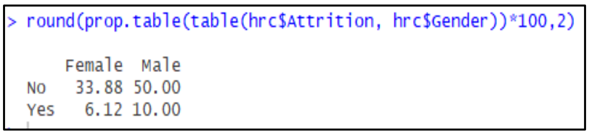
```
round(prop.table(table(d1$Attrition, d1$EducationField))*100, 2)

      Human Resources Life Sciences Marketing Medical Other Technical Degree
No               1.09         34.35      9.12   26.46  4.90             7.96
Yes              0.75          6.87      1.70    5.10  0.68             1.02
```



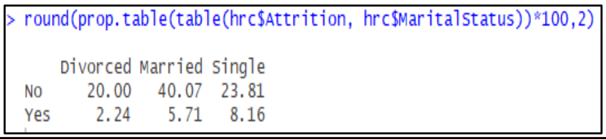
```
> round(prop.table(table(d1$BusinessTravel, d1$Department), margin=2)*100, 2)

                    Human Resources Research & Development Sales
  Non-Travel                   4.76                  11.45  8.30
  Travel_Frequently           12.70                  18.00 21.52
  Travel_Rarely               82.54                  70.55 70.18
```



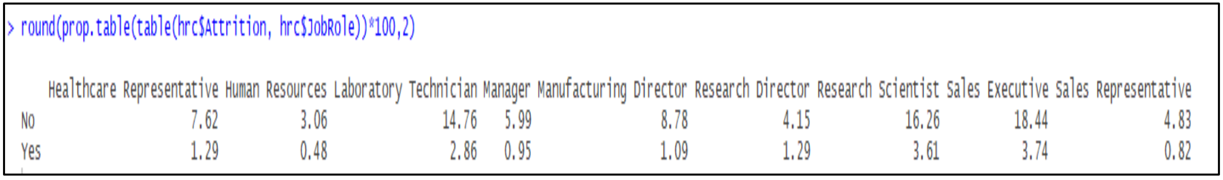
```
round(prop.table(table(d1$BusinessTravel, d1$Department))*100, 2)

                  Human Resources Research & Development Sales
Non-Travel                   0.20                   7.48  2.52
Travel_Frequently            0.54                  11.77  6.53
Travel_Rarely                3.54                  46.12 21.29
```



```
> round(prop.table(table(hrc$Attrition, hrc$Gender))*100,2)

      Female  Male
  No   33.88 50.00
  Yes   6.12 10.00
```

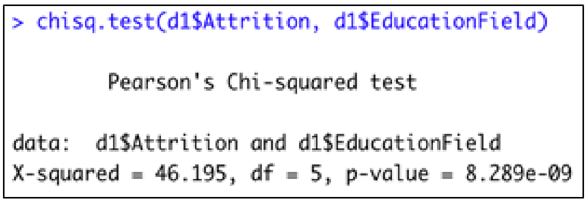


```
> round(prop.table(table(hrc$Attrition, hrc$MaritalStatus))*100,2)

      Divorced Married Single
  No     20.00   40.07  23.81
  Yes     2.24    5.71   8.16
```

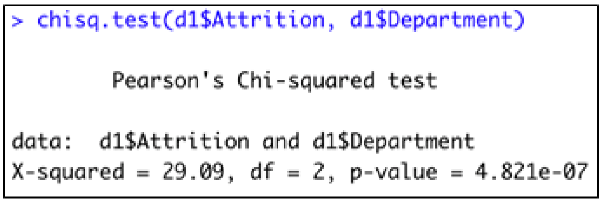


```
> round(prop.table(table(hrc$Attrition, hrc$JobRole))*100,2)

      Healthcare Representative Human Resources Laboratory Technician Manager Manufacturing Director Research Director Research Scientist Sales Executive Sales Representative
  No                       7.62            3.06                 14.76    5.99                   8.78              4.15              16.26           18.44                 4.83
  Yes                      1.29            0.48                  2.86    0.95                   1.09              1.29               3.61            3.74                 0.82
```

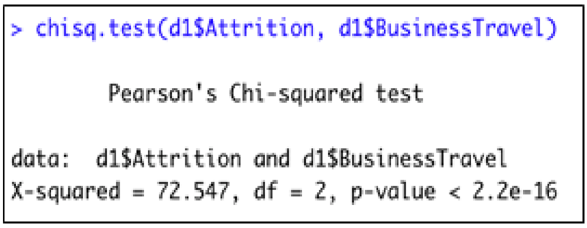
Chi-Squared Test



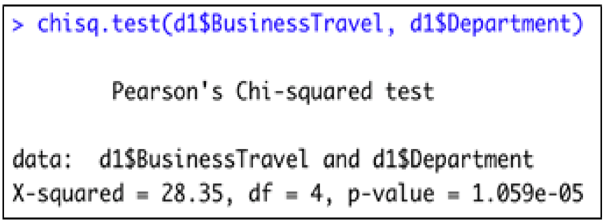
```
> chisq.test(d1$Attrition, d1$EducationField)

        Pearson's Chi-squared test

data:  d1$Attrition and d1$EducationField
X-squared = 46.195, df = 5, p-value = 8.289e-09
```



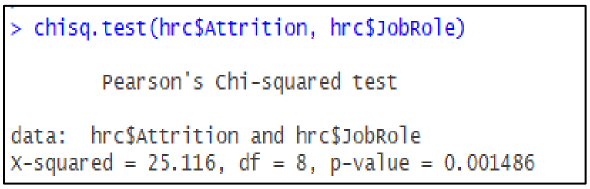
```
> chisq.test(d1$Attrition, d1$Department)

        Pearson's Chi-squared test

data:  d1$Attrition and d1$Department
X-squared = 29.09, df = 2, p-value = 4.821e-07
```



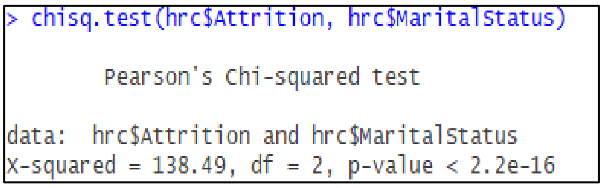
```
> chisq.test(d1$Attrition, d1$BusinessTravel)

        Pearson's Chi-squared test

data:  d1$Attrition and d1$BusinessTravel
X-squared = 72.547, df = 2, p-value < 2.2e-16
```



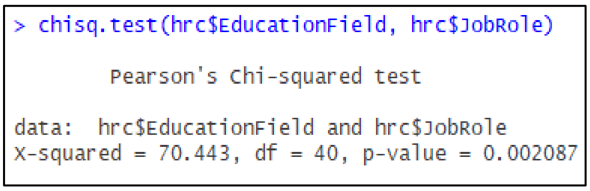
```
> chisq.test(d1$BusinessTravel, d1$Department)

        Pearson's Chi-squared test

data:  d1$BusinessTravel and d1$Department
X-squared = 28.35, df = 4, p-value = 1.059e-05
```



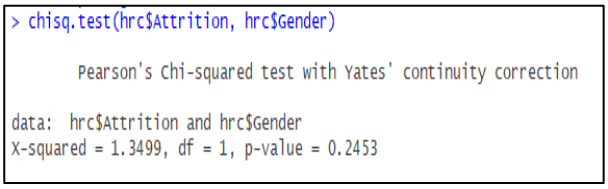
```
> chisq.test(hrc$Attrition, hrc$JobRole)

        Pearson's Chi-squared test

data:  hrc$Attrition and hrc$JobRole
X-squared = 25.116, df = 8, p-value = 0.001486
```



```
> chisq.test(hrc$Attrition, hrc$MaritalStatus)

        Pearson's Chi-squared test

data:  hrc$Attrition and hrc$MaritalStatus
X-squared = 138.49, df = 2, p-value < 2.2e-16
```

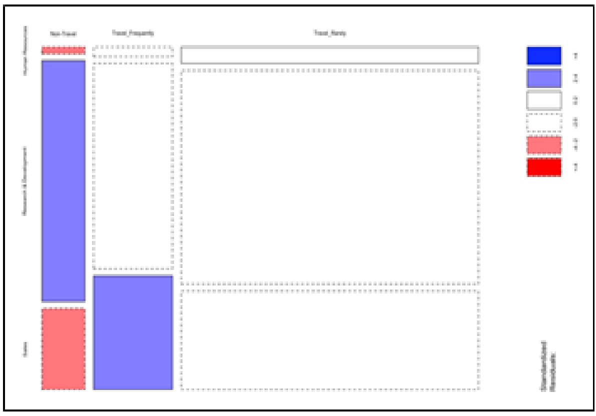


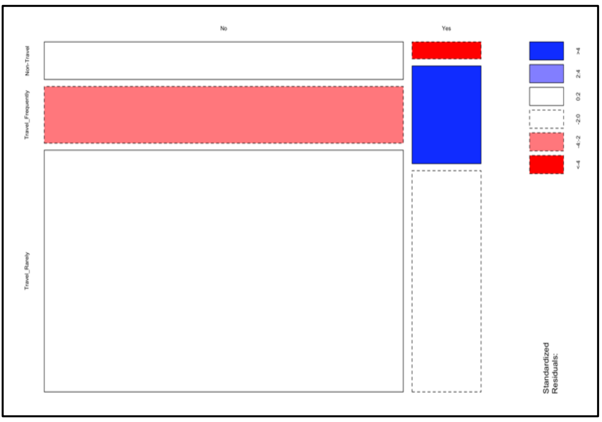
```
> chisq.test(hrc$EducationField, hrc$JobRole)

        Pearson's Chi-squared test

data:  hrc$EducationField and hrc$JobRole
X-squared = 70.443, df = 40, p-value = 0.002087
```

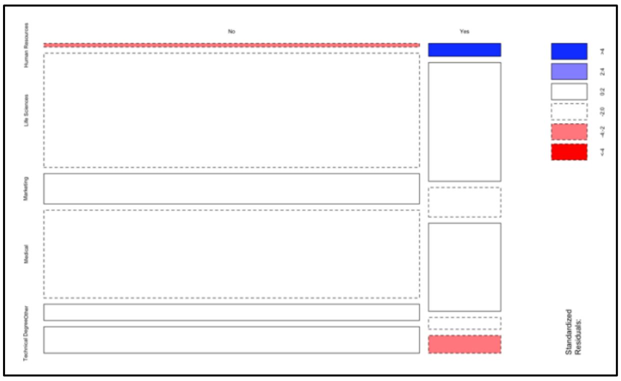


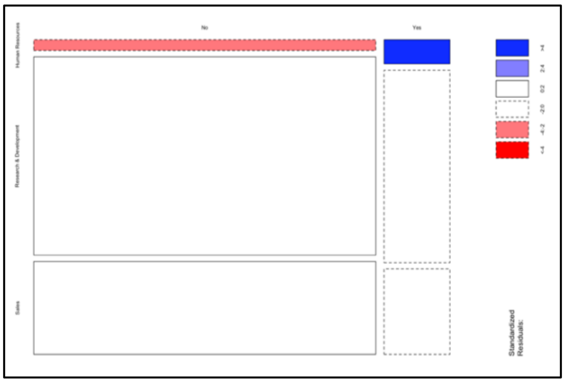
```
> chisq.test(hrc$Attrition, hrc$Gender)

        Pearson's Chi-squared test with Yates' continuity correction

data:  hrc$Attrition and hrc$Gender
X-squared = 1.3499, df = 1, p-value = 0.2453
```

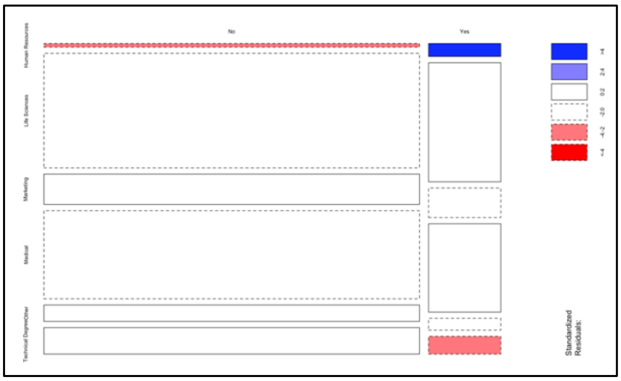
Mosaic Plot

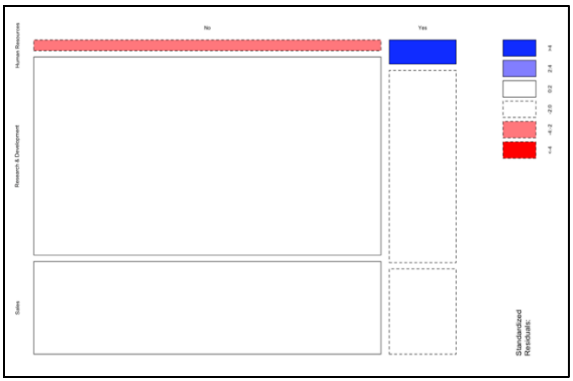


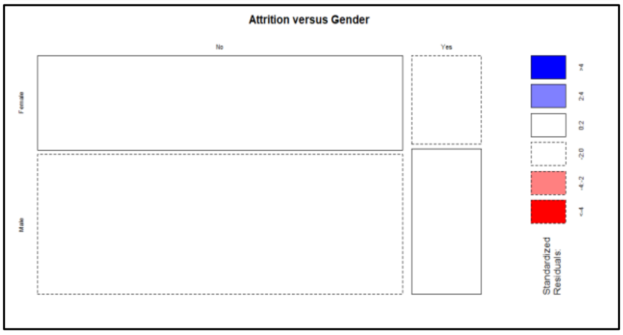


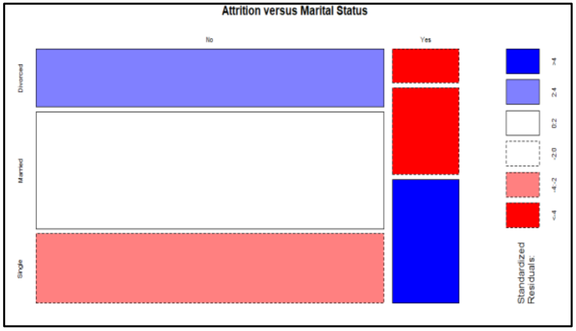

No
Yes
Human Resources
Life Sciences
Marketing
Medical
Technical DegreeOther
>4
2:4
0:2
-2:0
-4:-2
<-4
Standardized Residuals:

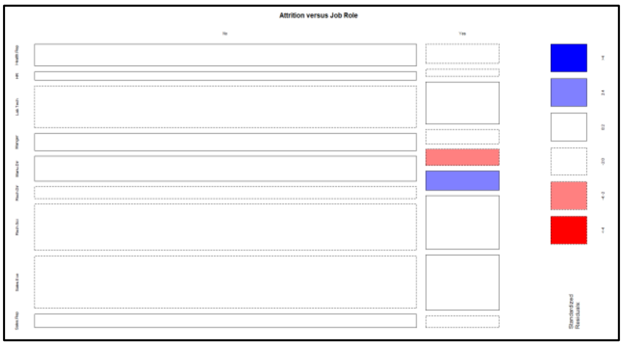

No
Yes
Human Resources
Research & Development
Sales
>4
2:4
0:2
-2:0
-4:-2
<-4
Standardized Residuals:

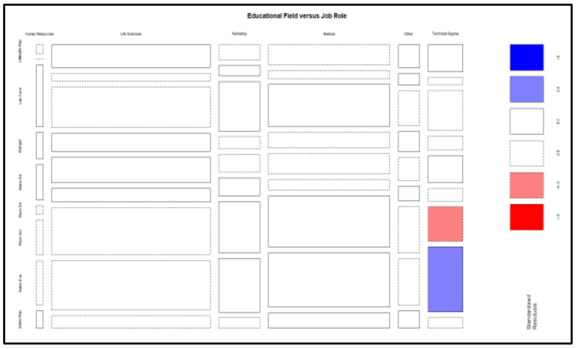

No
Yes
Human Resources
Life Sciences
Marketing
Medical
Technical DegreeOther
>4
2:4
0:2
-2:0
-4:-2
<-4
Standardized Residuals:

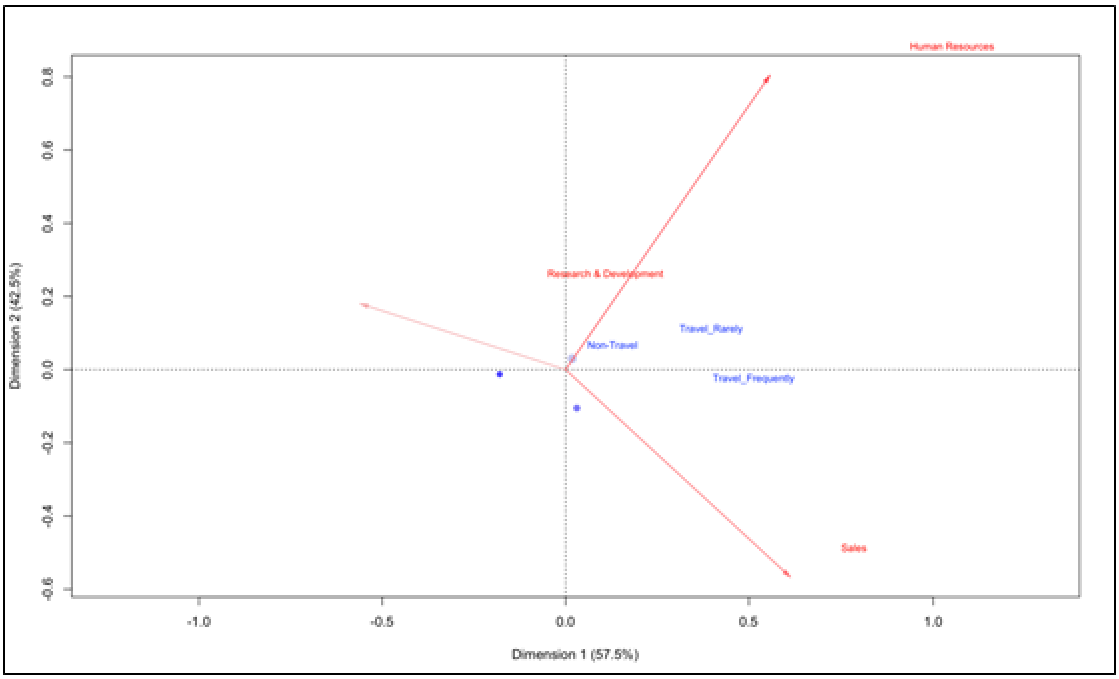

No
Yes
Human Resources
Research & Development
Sales
>4
2:4
0:2
-2:0
-4:-2
<-4
Standardized Residuals:

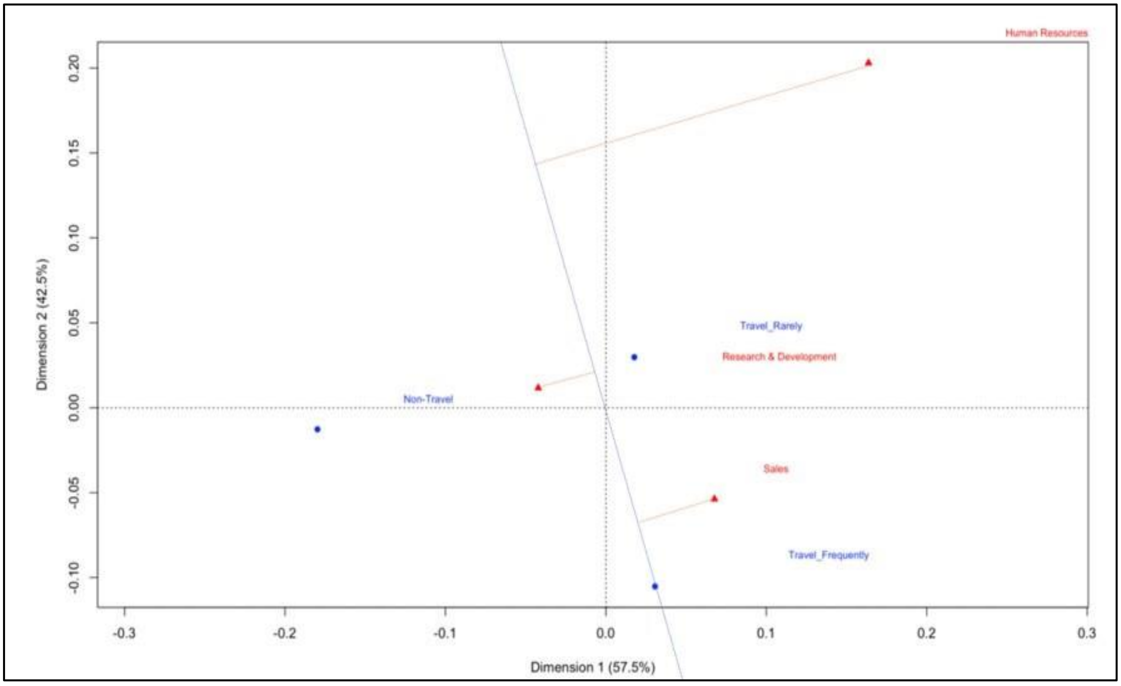

Attrition versus Gender
No
Yes
Female
Male
>4
2:4
0:2
-2:0
-4:-2
<-4
Standardized Residuals:


Attrition versus Marital Status
No
Yes
Divorced
Married
Single
>4
2:4
0:2
-2:0
-4:-2
<-4
Standardized Residuals:

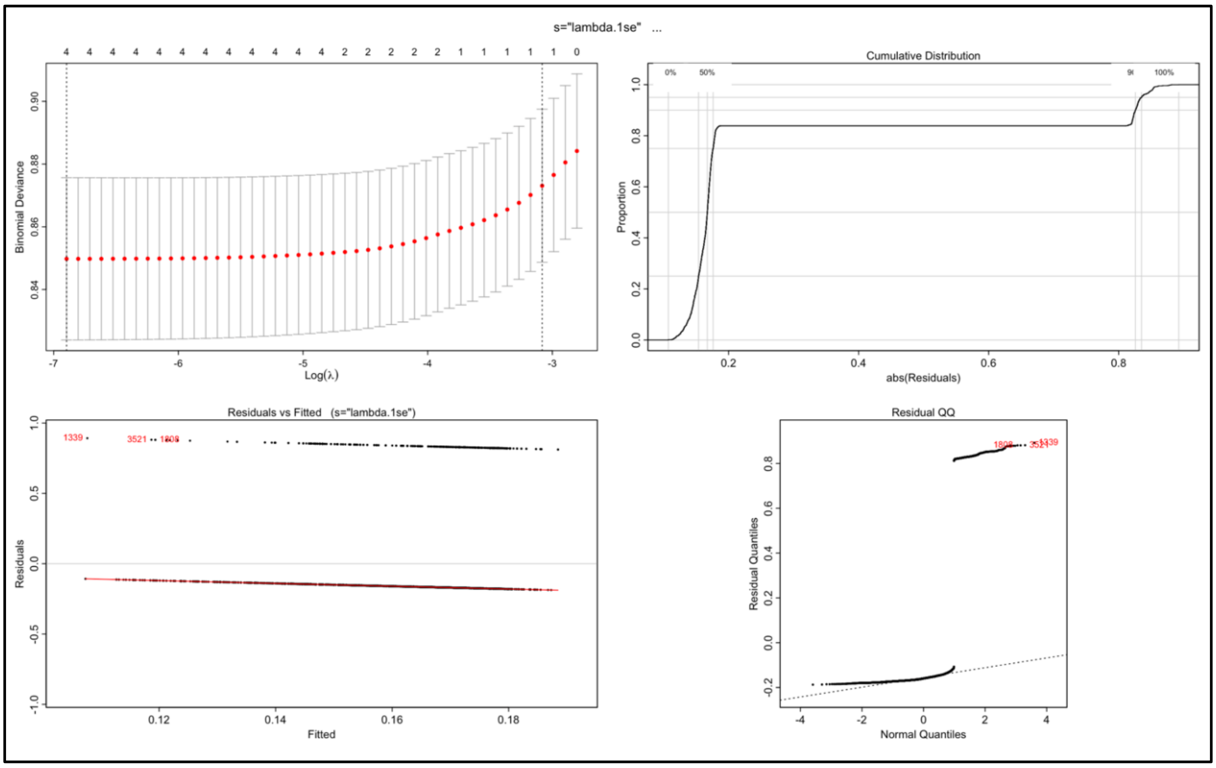

Attrition versus Job Role
No
Yes
Health Rep
HR
Lab Tech
Manger
Manu Dir
Rsch Dir
Rsch Sci
Sales Exe
Sales Rep
>4
2:4
0:2
-2:0
-4:-2
<-4
Standardized Residuals:


Educational Field versus Job Role
Human Resources
Life Sciences
Marketing
Medical
Other
Technical Degree
HHealth Rep
Lab Tech
Manger
Manu Dir
Rsch Dir
Rsch Sci
Sales Exe
Sales Rep
>4
2:4
0:2
-2:0
-4:-2
<-4
Standardized Residuals:

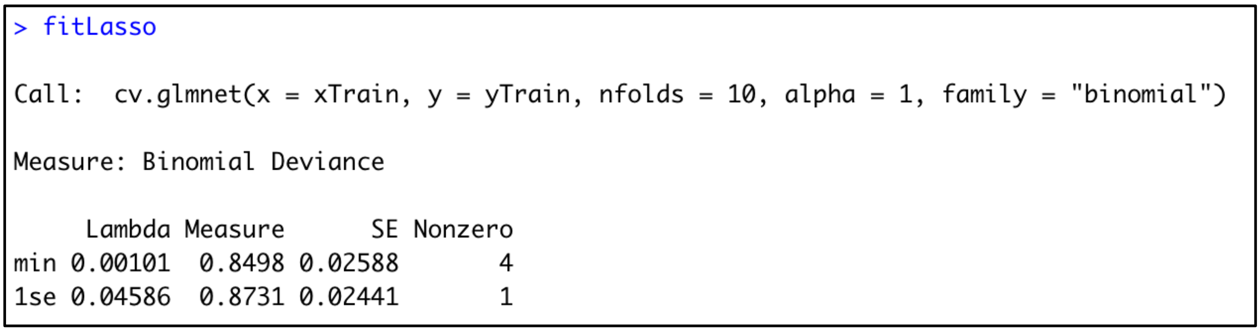

Human Resources
Research & Development
Travel_Rarely
Non-Travel
Travel_Frequently
Sales
Dimension 2 (42.5%)
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-1.0
-0.5
0.0
0.5
1.0
Dimension 1 (57.5%)


Human Resources
Travel_Rarely
Research & Development
Non-Travel
Sales
Travel_Frequently
Dimension 2 (42.5%)
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
Dimension 1 (57.5%)

Lasso Logistic Result



Lasso Logistic Model Overview



```
> fitLasso

Call:  cv.glmnet(x = xTrain, y = yTrain, nfolds = 10, alpha = 1, family = "binomial")

Measure: Binomial Deviance

     Lambda Measure      SE Nonzero
min 0.00101  0.8498 0.02588       4
1se 0.04586  0.8731 0.02441       1
```

Appendix 3 - R Code

```
#DSC424 - Advanced Data Analysis
#Project Name: HR Analytics
#Prediction:
#a. Employee Attrition: Binary
#b. Employee Monthly Salary: Continuous numeric

#TEAM Members
#Andy, Arun, Sahana, Shweta

#Data Source: Kaggle
#################### Install Packages - START #########################
install.packages("corrplot")
install.packages("QuantPsyc")
install.packages("car")
install.packages("leaps")
install.packages("lm.beta")
install.packages("readxl")
install.packages("tibble")
install.packages("polycor")
install.packages("ca")
install.packages("FactoMineR")
install.packages("factoextra")
install.packages("pls") #Partial Leased Squared
install.packages("dplyr")
install.packages ("forecast")
install.packages ("caret")
install.packages("caTools")
install.packages("GGally")
install.packages("glmnet")
install.packages("plotmo")

#################### Install Packages - END #########################
#################### Load Libraries - START ########################
library(psych)     # Has a much better scatterplot matrix function
library(corrplot)  # A nice correlation matrix visualization
library(car)       # Misc statistical methods
library(QuantPsyc) # Misc statistical methods
library(leaps)     # Gives forward, backward and stepwise
library(lm.beta)   # Gives us standardized coefficients
library(dplyr)
library(ggplot2)
library(readxl)
library(polycor)
library(ca)
library(FactoMineR)
library(factoextra)
library(pls)
library(hetcor)
library(caTools)
library(forecast)
library(caret)
library(lmtest)
library(GGally)
library(glmnet)
library(plotmo)
#################### Load Libraries - START ########################

#Set working directory
setwd("C:/Users/sahan/School Materials/DSC424-Advanced Data Analysis/Final
Project/Proposal")

#Load Psych plot for visualization
source("PCA_Plot.R")

## Reading/Loading the data
hr = read_excel("HRdataset_combined.xlsx",sheet=1)
head(hr)
dim(hr)
str(hr)

#Checking numeric variables
hr_num_fields = select_if(hr, is.numeric)
head(hr_num_fields)
dim(hr_num_fields)
str(hr_num_fields)

#Checking character variables
hr_chr_fields = select_if(hr, is.character)
head(hr_chr_fields)
dim(hr_chr_fields)
str(hr_chr_fields)

#Dependent variable - separating it out
monthlyIncome = hr$MonthlyIncome

############################# Building numeric data - START ####################

#Removing non-numeric field and dependent variable MonthlyIncome for PCA analysis
hr_num_fields = subset(hr_num_fields, select = -
c(EmployeeID,MonthlyIncome,StandardHours,JobInvolvement,PerformanceRating,JobLevel,
StockOptionLevel,Education))
dim(hr_num_fields)
```

```
#Converting TotalWorkingYears and NumCompaniesWorked to numeric and replacing NA's with 0 after analyzing the data
twh = suppressWarnings(as.numeric(hr$TotalWorkingYears))
length(twh)
sum(is.na(twh))
twh[is.na(twh)] <- 0
hr_num_fields$TotalWorkingYears = twh

comp = suppressWarnings(as.numeric(hr$NumCompaniesWorked))
length(comp)
unique(comp)
sum(is.na(comp))
comp[is.na(comp)] <- 0
hr_num_fields$NumCompaniesWorked = comp

head(hr_num_fields)

############################## Building numeric data - END #####################
############################## Building Ordinal Data: START ######################

hr_ord_fields = hr$Education  ## Considering education as an ordinal variable
hr_ord_fields = subset(hr, select=c(Education,EnvironmentSatisfaction,JobSatisfaction,WorkLifeBalance,JobInvolvement,PerformanceRating,JobLevel,StockOptionLevel))
head(hr_ord_fields)
dim(hr_ord_fields)
str(hr_ord_fields)

#Converting EnvironmentSatisfaction, JobSatisfaction and WorkLifeBalance from char type to numeric
envSat = suppressWarnings(as.numeric(hr_ord_fields$EnvironmentSatisfaction))
length(envSat)
sum(is.na(envSat))
val <- unique(envSat[!is.na(envSat)])
modeEnvSat = val[which.max(tabulate(match(envSat, val)))] # Mode of envSat
envSat[is.na(envSat)] <- modeEnvSat
hr_ord_fields$EnvSat = envSat

jobSat = suppressWarnings(as.numeric(hr_ord_fields$JobSatisfaction))
length(jobSat)
sum(is.na(jobSat))
val <- unique(jobSat[!is.na(jobSat)])
modeJobSat = val[which.max(tabulate(match(jobSat, val)))] # Mode of envSat
jobSat[is.na(jobSat)] <- modeJobSat
hr_ord_fields$JobSat = jobSat
```

```
wrkLifBal = suppressWarnings(as.numeric(hr_ord_fields$WorkLifeBalance))
length(wrkLifBal)
sum(is.na(wrkLifBal))
val <- unique(wrkLifBal[!is.na(wrkLifBal)])
modeWrkLifBal = val[which.max(tabulate(match(wrkLifBal, val)))] # Mode of envSat
wrkLifBal[is.na(wrkLifBal)] <- modeWrkLifBal
hr_ord_fields$WrkLifBal = wrkLifBal

#check the ordinal data after imputing for NAs
str(hr_ord_fields)
hr_ord_fields = subset(hr_ord_fields, select = -
c(EnvironmentSatisfaction,JobSatisfaction,WorkLifeBalance))
str(hr_ord_fields)

hist(hr_ord_fields$Education, main="Education Levels",
    xlab="Education Levels", col="blue")

######################## Building Ordinal Data: END #######################
############################# Building Categorical Data: START ################

# 6 Categorical variables are used to explore the relationship with Attrition in
Correspondence Analysis
str(hr)
str(hr_chr_fields)

hr_cate_fields = subset(hr, select=c(Attrition, BusinessTravel, Department, EducationField,
Gender, JobRole, MaritalStatus))
head(hr_cate_fields)
dim(hr_cate_fields)
str(hr_cate_fields)

############################# Building Categorical Data: END ####################
#Going forward use the above final numeric, categorical and ordinal data for analysis
############################# Data Explorations and Graphs code - START ##########

head(hr_num_fields)
str(hr_num_fields)

# Plotting basic histogram of single variable
hist(hr_num_fields$Age,
    main="Histogram for Age",
    xlab="Age",
    border="green",
    col="blue",
    breaks=15)

hist(hr_num_fields$DistanceFromHome,
    main="Histogram for Distance From Home",
    xlab="Distance From Home", border="green",
    col="blue",
    breaks=15)

hist(hr_num_fields$PercentSalaryHike,
    main="Histogram for Percent Salary Hike",
    xlab="PercentSalaryHike", border="green",
    col="blue",
    breaks=15)

hist(hr_num_fields$TrainingTimesLastYear,
    main="Histogram for Training Times Last Year",
    xlab="TrainingTimesLastYear", border="green",
    col="blue")

hist(hr_num_fields$YearsAtCompany,
    main="Histogram for Years At Company",
    xlab="Years At Company", border="green",
    col="blue")

hist(hr_num_fields$YearsSinceLastPromotion,
    main="Histogram for Years Since Last Promotion",
    xlab="YearsSinceLastPromotion", border="green",
    col="blue",
    breaks=15)

hist(hr_num_fields$YearsWithCurrManager,
    main="Histogram for Years With Current Manager",
    xlab="Years With Current Manager", border="green",
    col="blue",
    breaks=15)

hist(hr_num_fields$TotalWorkingYears,
    main="Histogram for Total Working Years",
    xlab="Total Working Years", border="green",
    col="blue")

hist(hr_num_fields$NumCompaniesWorked,
    main="Histogram for Number of Companies Worked",
    xlab="Number of Companies Worked", border="green",
    col="blue")

# What about "Monthly Income"?
hist(monthlyIncome,
    main="Histogram for Monthly Income",
    xlab="Monthly Income", border="green",
    col="red")

# What about "Attrition"?
attach(hr)
f_attrition <- table(Attrition)
f_attrition # print table
prop.table(f_attrition) # cell percentages


# All numeric variables without log tranformation on Monthly Income
Fit1 = lm(monthlyIncome ~ ., data=hr_num_fields)
summary(Fit1) # R-square = 1.3%
lm.beta(Fit1)
plot(Fit1)
hist(Fit1$residuals,
    main="Histogram of Fit1 Residuals",
    xlab=" Fit1 Residuals", border="green",
    col="blue",
    breaks=15)

# Numeric variables with log tranformation on Monthly Income
Fit2 = lm(log1p(monthlyIncome) ~ ., data=hr_num_fields)
summary(Fit2) # R-square = 1.09%
lm.beta(Fit2)
plot(Fit2)
hist(Fit2$residuals,
    main="Histogram of Fit2 Residuals",
    xlab=" Fit2 Residuals", border="green",
    col="blue",
    breaks=15)

# All ordinal variables without log tranformation on Monthly Income
str(hr_ord_fields)
Fit3 = lm(monthlyIncome ~ ., data=hr_ord_fields)
Anova(Fit3)
summary(Fit3) # R-square = 0.39%
plot(Fit3)
hist(Fit3$residuals,
    main="Histogram of Fit3 Residuals",
    xlab=" Fit3 Residuals", border="green",
    col="blue",
    breaks=15)

# All ordinal variables with log tranformation on Monthly Income
Fit4 = lm(log1p(monthlyIncome) ~ ., data=hr_ord_fields)
Anova(Fit4)
summary(Fit4) # R-square = 0.43%
```

```
plot(Fit4)
hist(Fit4$residual,
    main="Histogram of Fit4 Residuals",
    xlab=" Fit4 Residuals", border="green",
    col="blue",
    breaks=15)

####################### Combine numeric and ordinal in a data frame #############

str(hr_ord_fields)
Education = hr_ord_fields$Education
JobInvolvement = hr_ord_fields$JobInvolvement
PerformanceRating = hr_ord_fields$PerformanceRating
JobLevel = hr_ord_fields$JobLevel
StockOptionLevel = hr_ord_fields$StockOptionLevel
EnvSat = hr_ord_fields$EnvSat
JobSat = hr_ord_fields$JobSat
WrkLifBal = hr_ord_fields$WrkLifBal

hr_num_ord = cbind(hr_num_fields, Education)
hr_num_ord_2 = cbind(hr_num_ord, JobInvolvement)
hr_num_ord_3 = cbind(hr_num_ord_2, PerformanceRating)
hr_num_ord_4 = cbind(hr_num_ord_3, JobLevel)
hr_num_ord_5 = cbind(hr_num_ord_4, StockOptionLevel)
hr_num_ord_6 = cbind(hr_num_ord_5, EnvSat)
hr_num_ord_7 = cbind(hr_num_ord_6, JobSat)
hr_num_ord_8 = cbind(hr_num_ord_7, WrkLifBal) # This is the data frame which contain all
Final numeric and oridinal variables

hr_num_ord = hr_num_ord_8 # Assign it back to the name "hr_num_ord"
str(hr_num_ord)

#########################################################################
###################################################################

# All numeric and ordinal variables with log tranformation on Monthly Income
Fit5 = lm(log1p(monthlyIncome) ~ ., data=hr_num_ord)
Anova(Fit5)
summary(Fit5) #R-square = 1.55%
plot(Fit5)
hist(Fit5$residual,
    main="Histogram of Fit5 Residuals",
    xlab=" Fit5 Residuals", border="green",
    col="blue",
    breaks=15)

################################ Forward Selection #########################
```

```
# Feed the two "bounding" models
null = lm(log1p(monthlyIncome) ~ 1, data=hr_num_ord)
null
full = lm(log1p(monthlyIncome) ~ ., data=hr_num_ord)
summary(full)

# forward search
log_monthly_Forward = step(null, scope = list(lower=null, upper=full),
                    direction="forward", trace=F)
summary(log_monthly_Forward)

# The lm.beta gives "standardized betas" which better tell how large
# an effect a variable has on the parameter of interest than the raw
# beta does.
lm.beta(log_monthly_Forward)

# Look at the standardized coefficients to see which influence the
# parameter of interest to a greater degree.
stdCoef = coef(lm.beta(log_monthly_Forward))    # Grab the standardized coefficients
barplot(sort(stdCoef))
barplot(rev(sort(stdCoef)))            # Graph the coefficients in order of importance
stdCoef

print(log_monthly_Forward)             # Model Equation

############################ Backward Selection ########################

log_monthly_Backward = step(full, scope=list(lower=null, upper=full),
direction="backward", trace=F)
log_monthly_Backward = step(full, direction="backward", trace=F)
summary(log_monthly_Backward)

stdCoef = coef(lm.beta(log_monthly_Backward))    # Grab the standardized coefficients
barplot(rev(sort(stdCoef)))            # Graph the coefficients in order of importance
stdCoef


###################### Stepwise Selection ######################
log_monthly_Step = step(null, scope=list(lower=null, upper=full), direction="both", trace=F)
stdCoef = coef(lm.beta(log_monthly_Step))    # Grab the standardized coefficients
barplot(rev(sort(stdCoef)))            # Graph the coefficients in order of importance
stdCoef

summary(log_monthly_Step)
summary(log_monthly_Forward)
summary(log_monthly_Backward)
```

```
anova(log_monthly_Step, log_monthly_Forward)    # Is there any difference in predictive
power? - NO
anova(log_monthly_Step, log_monthly_Backward)   # Is there any difference in predictive
power? - NO


################################# Dependent Variable = Attrition ################
################Perform Logistic regression for Attrition, dependent variable#########

str(hr_cate_fields)
Attrition = hr_cate_fields$Attrition

hr_num_fields_logistic = cbind(hr_num_fields, Attrition)
str(hr_num_fields_logistic)

LogMI = log(monthlyIncome)
hist(LogMI)

hr_num_fields_logistic = cbind(hr_num_fields_logistic, LogMI)
str(hr_num_fields_logistic)

hr_num_fields_logistic$Attrition<- ifelse(hr_num_fields_logistic$Attrition=="Yes",1,0)
str(hr_num_fields_logistic)

set.seed(100)
indices = sample.split(hr_num_fields_logistic$Attrition, SplitRatio = 0.7)
train = hr_num_fields_logistic[indices,]
test = hr_num_fields_logistic[!(indices),]

model_1 = glm(Attrition ~ Age+ DistanceFromHome + PercentSalaryHike +
TrainingTimesLastYear + YearsAtCompany + YearsSinceLastPromotion +
YearsWithCurrManager + TotalWorkingYears + NumCompaniesWorked + LogMI, data =
train, family = "binomial")
summary(model_1)
confint(model_1)
exp(coef(model_1))
anova(model_1, test ="Chisq")

model_2<- stepAIC(model_1, direction="both")
summary(model_2)
vif(model_2)

confint(model_2)
exp(coef(model_2))
anova(model_2, test ="Chisq")

#Tells if the model is significant or not
```

```
with(model_2, null.deviance - deviance)
with(model_2, df.null - df.residual)
with(model_2, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))

# Variable importance
varImp(model_2)
lrtest(model_1, model_2)

# Predict on test data

test_pred = predict(model_2, type = "response", newdata = test)
test_pred <- ifelse(test_pred > 0.5,1,0)

#Accuracy
misClasificError <- mean(test_pred != test$Attrition)
accuracy(test$Attrition,test_pred)
print(paste('Accuracy',1-misClasificError)) #Accuracy is 0.83 is a good result.

####### End Logistic Regression on Numeric Data#######
################ Recode Attrition (Yes/No)######################################

hr_2 = hr %>% mutate(Attrition=recode(Attrition,
                          `Yes`="1",
                          `No`="2"))

num_attrition = as.numeric(hr_2$Attrition)
str(hr_2)
num_attrition

cbind(hr_2, num_attrition)
head(hr_2, 3)
hr_2$EmployeeID = NULL # Get rid of Employee ID
str(hr_2)
############################## Data Explorations and Graphs code - END ###########
############################################# PCA - START ####################

## Data transformation

## Exploratory Data Analysis
max(monthlyIncome)
boxplot(monthlyIncome)

## Principal Component Analaysis
# Get an initial plot
plot(hr_num_fields, pch=16, col="red") #there is correlations between features
pca_num_1 = prcomp(hr_num_fields, scale = T)
plot(pca_num_1)
```

```
abline(1,0)
print(pca_num_1)
summary(pca_num_1)
names(pca_num_1)
round(pca_num_1$x,2)

PCA_Plot(pca_num_1)
round(pca_num_1$rotation,2)

#choosing 6 factors which gives 90% variance in data
#Factor Analysis - with rotation

pfa_num_1 = principal(hr_num_fields, nf = 4)
print(round(pfa_num_1$loadings,2), cutoff = 0.4)
PCA_Plot_Psyc(pfa_num_1)

#Factor Analysis - without rotation
pfa_num_2 = principal(hr_num_fields, nf = 4, rotate = 'none')
print(pfa_num_2)
print(round(pfa_num_2$loadings,2), cutoff = 0.4)
PCA_Plot(pfa_num_2)

#################################### PCA - END ##############################
################################### Ordinal Factor Analysis – START #############

#Before doing FA, we will check the correlations of the Ordinal data with our dependent
variables Attrition and Monthly Income
hr_attr <- ifelse(hr$Attrition=="Yes", 1, 0)
str(hr_attr)
length(hr_attr)

#Checking correlation of each ordinal data with Attrition/Monthly Income dependent
variable
cor(hr_ord_fields$Education,hr_attr,  method = c("pearson", "kendall", "spearman"))
cor.test(hr_ord_fields$Education, hr_attr, method=c("pearson", "kendall", "spearman"))

cor(hr_ord_fields$Education,hr$MonthlyIncome,  method = c("pearson", "kendall",
"spearman"))
cor.test(hr_ord_fields$Education, hr$MonthlyIncome, method=c("pearson", "kendall",
"spearman"))
#Education is not having any significance to Attrition/Monthly Incoem - has very high p-
value of 0.31/0.67 respectively

cor(hr_ord_fields$JobInvolvement,hr_attr,  method = c("pearson", "kendall", "spearman"))
cor.test(hr_ord_fields$JobInvolvement, hr_attr, method=c("pearson", "kendall",
"spearman"))
```

```
cor(hr_ord_fields$JobInvolvement,hr$MonthlyIncome,  method = c("pearson", "kendall", "spearman"))
cor.test(hr_ord_fields$JobInvolvement, hr$MonthlyIncome, method=c("pearson", "kendall", "spearman"))
#Job Involvement is not having any significance to Attrition/Monthly Incoem - has very high p-value of 0.3/0.12 respectively

cor(hr_ord_fields$PerformanceRating,hr_attr,  method = c("pearson", "kendall", "spearman"))
cor.test(hr_ord_fields$PerformanceRating, hr_attr, method=c("pearson", "kendall", "spearman"))

cor(hr_ord_fields$PerformanceRating,hr$MonthlyIncome,  method = c("pearson", "kendall", "spearman"))
cor.test(hr_ord_fields$PerformanceRating, hr$MonthlyIncome, method=c("pearson", "kendall", "spearman"))
#Perfomance Rating is not having any significance to Attrition/Monthly Incoem - has very high p-value of 0.12/0.28 respectively

cor(hr_ord_fields$JobLevel,hr_attr,  method = c("pearson", "kendall", "spearman"))
cor.test(hr_ord_fields$JobLevel, hr_attr, method=c("pearson", "kendall", "spearman"))
#Job Level is not having any significance to Attrition - has very high p-value of 0.49

cor(hr_ord_fields$JobLevel,hr$MonthlyIncome,  method = c("pearson", "kendall", "spearman"))
cor.test(hr_ord_fields$JobLevel, hr$MonthlyIncome, method=c("pearson", "kendall", "spearman"))
#Job Level shows good significance with p-value low as 0.0016 and corr coefficient is showing 4.7% correlation with Monthly Income

cor(hr_ord_fields$StockOptionLevel,hr_attr,  method = c("pearson", "kendall", "spearman"))
cor.test(hr_ord_fields$StockOptionLevel, hr_attr, method=c("pearson", "kendall", "spearman"))

cor(hr_ord_fields$StockOptionLevel,hr$MonthlyIncome,  method = c("pearson", "kendall", "spearman"))
cor.test(hr_ord_fields$StockOptionLevel, hr$MonthlyIncome, method=c("pearson", "kendall", "spearman"))
#StockOptionLevel is not having any significance to Attrition/Monthly Incoem - has very high p-value of 0.64/0.07 respectively

cor(hr_ord_fields$EnvSat,hr_attr,  method = c("pearson", "kendall", "spearman"))
cor.test(hr_ord_fields$EnvSat, hr_attr, method=c("pearson", "kendall", "spearman"))
#Environment satisfaction is having p-value 0, meaning highly significant to Attrition and the corr coefficient is -10.2%.
```

```
cor(hr_ord_fields$EnvSat,hr$MonthlyIncome,  method = c("pearson", "kendall",
"spearman"))
cor.test(hr_ord_fields$EnvSat, hr$MonthlyIncome, method=c("pearson", "kendall",
"spearman"))
#Environment Satisfacton is not having any significance Monthly Incoem - has very high p-
value of0.70

cor(hr_ord_fields$JobSat,hr_attr,  method = c("pearson", "kendall", "spearman"))
cor.test(hr_ord_fields$JobSat, hr_attr, method=c("pearson", "kendall", "spearman"))
#Job satisfaction is having p-value 0, meaning highly significant to Attrition and the corr
coefficient is -10.4%.

cor(hr_ord_fields$JobSat,hr$MonthlyIncome,  method = c("pearson", "kendall",
"spearman"))
cor.test(hr_ord_fields$JobSat, hr$MonthlyIncome, method=c("pearson", "kendall",
"spearman"))
#JobSat is not having any significance toMonthly Incoem - has very high p-value of 0.81
respectively

cor(hr_ord_fields$WrkLifBal,hr_attr,  method = c("pearson", "kendall", "spearman"))
cor.test(hr_ord_fields$WrkLifBal, hr_attr, method=c("pearson", "kendall", "spearman"))
#WrkLifBal is significance with Attrition - has very low p-value of 0 and corr coefficients -
6.3%

cor(hr_ord_fields$WrkLifBal,hr$MonthlyIncome,  method = c("pearson", "kendall",
"spearman"))
cor.test(hr_ord_fields$WrkLifBal, hr$MonthlyIncome, method=c("pearson", "kendall",
"spearman"))
#WrkLifBal is not having any significance to Monthly Incoem - has very high p-value of 0.8
respectively

#From the above correlation analysis of ordinal data with dependent variables, there were
only couple of variables that was significant
#EnvSat,JobSat and WrkLifBal are highly significant with Attrition
#JobLevel is highly signifiant with Monthly Income

#Checking correlations or ordinal data
corrOrd = cor(hr_ord_fields)
corrplot(corrOrd, method="ellipse")
corrplot(cor(hr_ord_fields))

#We do not see much correlations between the ordinal variables
# corrplot was made with the Pearson correlation, so let's try spearman
corrOrdS = cor(hr_ord_fields, method="spearman")
corrplot(corrOrdS, method="ellipse")
corrplot(corrOrdS)
```

```
#correlations for kendal
corrOrdK = cor(hr_ord_fields, method="kendal")
corrplot(corrOrdK, method="ellipse")
corrplot(corrOrdK)

# how different would the factor analysis
max(corrOrdS - corrOrd)
min(corrOrdS - corrOrd)

max(corrOrdK - corrOrd)
min(corrOrdK - corrOrd)

range(corrOrdS)
range(corrOrdK)
range(corrOrd)

# Pearson is not very helpful! We need minimum 5 components to get 80% variance in the data
pPearson = prcomp(hr_ord_fields)
summary(pPearson)
plot(pPearson)
abline(1, 0, col="red")
PCA_Plot(p)

pPearson2 = principal(cor(hr_ord_fields), nfactors=4)
summary(pPearson2)
print(pPearson2$loadings, cutoff=.4)

# We do a bit better with the spearman, it looks like about 4 components.
pSpear = prcomp(cor(hr_ord_fields, method="spearman"))
summary(pSpear)
plot(pSpear)
abline(1, 0, col="red")
PCA_Plot(pSpear)

pSpear2 = principal(cor(hr_ord_fields, method="spearman"), nfactors=4)
summary(pSpear2)
print(pSpear2$loadings, cutoff=.4)

scores = as.data.frame(pSpear2$scores)
head(scores)

#prcomp for method = kendal
pKendal = prcomp(cor(hr_ord_fields, method="kendal"))
summary(pKendal)
plot(pKendal)
abline(1, 0, col="red")
```

```
PCA_Plot(pKendal)

pKendal2 = principal(cor(hr_ord_fields, method="kendal"), nfactors=4)
summary(pKendal2)
print(pKendal2$loadings, cutoff=.4)

#Polychoric
R = hector(hr_ord_fields)
P = principal(R)

# Let's now check with common factor analysis
f = factanal(covmat=corrOrdS, factors = 2)
print(f$loadings, cutoff=.4)

## Polychoric correlation
poly_cor = polychoric(hr_ord_fields)
rho = poly_cor$rho

### Thresholds/Scaling results
poly_cor$tau

cor.plot(poly_cor$rho, numbers=T, upper=FALSE, main = "Polychoric Correlation",
show.legend = FALSE)

# Scree plot
fa.parallel(rho, fm="pa", fa="fa", main = "Scree Plot")

# Polychoric factor analysis
poly_model = fa(hr_ord_fields, nfactor=3, cor="poly", fm="mle", rotate = "none")
poly_model$loadings

# Cluster analysis plot
fa.diagram(poly_model)

pc <- hetcor(hr_ord_fields, ML=TRUE)   # polychoric corr matrix

faPC <- fa(r=pc$correlations, nfactors=2, rotate="varimax")
faPC$loadings

################################### Ordinal Factor Analysis - END
#################################
################################### Hetcor - correlation analysis - START
##########################
#hr_num_fields - numeric
#hr_ord_fields - ordinal
#hr_cate_fields - categorical
```

```
data = data.frame(hr_num_fields, hr_ord_fields, hr_cate_fields)
p = hetcor(data)$cor
corrplot(p)

################################### Hetcor - correlation analysis – END ###########
################################## Cluster Analysis for Ordinal data - START ########

#get_dist: for computing a distance matrix between the rows of a data matrix.
#The default distance computed is the Euclidean
#fviz_dist: for visualizing a distance matrix
distance = get_dist(hr_ord_fields)
fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))

# Now, compute a k-means clustering with the three clusters we see in the data
ordHRClust = kmeans(hr_ord_fields, 3)
plot(hr_ord_fields$EnvSat, hr_ord_fields$JobSat, col=ordHRClust$cluster)

## Reading/Loading the data
fact_data = read.csv("factor_data.csv")
fact_data = round(fact_data,2)
head(fact_data)
dim(fact_data)
str(fact_data)
names(fact_data)

cor(fact_data,  method = c("spearman"))
corrplot(cor(fact_data,  method = c("spearman")))

fdClust = kmeans(fact_data, 2)
plot(fact_data$Exp_with_Company, fact_data$Overall_Exp, col=fdClust$cluster)

################################## Cluster Analysis for Ordinal data - END #########
################################## LDA - START ############################

#Performing OLS with factor data
ols_data = cbind(fact_data, monthlyIncome)

# Compute the correlation matrix and visualize it
cor_ols_data = cor(ols_data)
corrplot(cor_ols_data)

fit1 = lm(log1p(monthlyIncome) ~ Exp_with_Company + Overall_Exp + Overall_Satisfaction +
Env_Sat, data = ols_data)
summary(fit1)

# If you want to plot all four at the same time, this will do it
par(mfrow=c(2, 2))    # This will set up a 2x2 grid of plots
```

```
plot(fit1)          # Plot all four
par(mfrow=c(1, 1))    # Return the plot window to one plot

#Forward selection
null = lm(log1p(monthlyIncome) ~ 1, data = ols_data) # using only constant and no variable
in it
null
full = lm(log1p(monthlyIncome) ~ ., data = ols_data)
summary(full)

olsForward = step(null, scope = list(lower=null, upper=full),
             direction="forward", trace=T) # from leaps package - stepwise regression
#step - give starting point, give scope lower and upper point, give which direction
summary(olsForward)

# The lm.beta gives "standardized betas" which better tell how large
# an effect a variable has on the parameter of interest than the raw
# beta does.
lm.beta(olsForward)  # Sales force image has the biggest impact # standardize the betas

# Look at the standardized coefficients to see which influence the
# parameter of interest to a greater degree.
stdCoef = coef(lm.beta(olsForward))    # Grab the standardized coefficients
barplot(sort(stdCoef))
barplot(rev(sort(stdCoef)))           # Graph the coefficients in order of importance
stdCoef

#Backward selection
olsBackward = step(full, scope=list(lower=null, upper=full), direction="backward", trace=F)
olsBackward = step(full, direction="backward", trace=F)
summary(olsBackward)

stdCoef = coef(lm.beta(olsBackward))    # Grab the standardized coefficients
barplot(rev(sort(stdCoef)))           # Graph the coefficients in order of importance
stdCoef

# Note that Delivery Speed and Complaint Resolution have been replaced
# by Product Line and Price Flexibility!

# Finally we do a "stepwise" search combining the two
olsStep = step(null, scope=list(lower=null, upper=full), direction="both", trace=F)
summary(olsStep)
stdCoef = coef(lm.beta(olsStep))    # Grab the standardized coefficients
barplot(rev(sort(stdCoef)))           # Graph the coefficients in order of importance
stdCoef

anova(olsStep, olsForward)    # Finding difference in the predictive power
anova(olsStep, olsBackward)   # Finding difference in the predictive power
anova(olsForward, olsBackward)

Attrition = hr$Attrition
lda_data = cbind(fact_data, Attrition)
fit2 = lda(Attrition ~ Exp_with_Company + Overall_Exp + Overall_Satisfaction + Env_Sat,
data = lda_data)
summary(fit2)
print(fit2)
pred = predict(fit2, newdata = lda_data)$class
table(lda_data$Attrition, pred)
length(pred)

################################## LDA - END ################################

################################ Ordinal and Numeric Factor Analysis - Start #######

hr_num_ord = cbind(hr_num_fields,hr_ord_fields)
names(hr_ord_fields)
## remove uncorrelated ordinal variables from combined data
hr_num_ord$Education = hr_num_ord$JobInvolvement = hr_num_ord$PerformanceRating
= hr_num_ord$StockOptionLevel = NULL
names(hr_num_ord)

print("Initial Principal Component Analysis - Ordinal and Numeric Combined")
pca_num_1 = prcomp(hr_num_ord, scale = T)
plot(pca_num_1)
abline(1,0)
print(pca_num_1)
summary(pca_num_1)

pfa_num_1 = principal(hr_num_ord, nf = 4)
print(round(pfa_num_1$loadings,2), cutoff = 0.4)
PCA_Plot_Psyc(pfa_num_1)

names(pfa_num_1)
factor_data = pfa_num_1$scores
#colnames(factor_data) =
c("Experience.With.Company","Overall.Experience","Satisfaction","Environmental")
print(factor_data)

write.csv(factor_data,"factor_data.csv",row.names=FALSE)


########################## Ordinal and Numeric Factor Analysis - End ##############
############################### Correspondence Analysis - START ################
```

```
str(hr_cate_fields)

# Contingency Table
table(hr_cate_fields$Attrition, hr_cate_fields$BusinessTravel)
table(hr_cate_fields$Attrition, hr_cate_fields$Department)
table(hr_cate_fields$Attrition, hr_cate_fields$EducationField)
table(hr_cate_fields$BusinessTravel, hr_cate_fields$Department)

#Conversion to percents (multiply by 100)
#This is a joint probability distribution
round(prop.table(table(hr_cate_fields$Attrition, hr_cate_fields$BusinessTravel))*100, 2)
round(prop.table(table(hr_cate_fields$Attrition, hr_cate_fields$Department))*100, 2)
round(prop.table(table(hr_cate_fields$Attrition, hr_cate_fields$EducationField))*100, 2)
round(prop.table(table(hr_cate_fields$BusinessTravel, hr_cate_fields$Department))*100, 2)

# More often, we are interested in the distribution of one variable within groups created by
another
round(prop.table(table(hr_cate_fields$Attrition, hr_cate_fields$BusinessTravel),
margin=2)*100, 2)
round(prop.table(table(hr_cate_fields$Attrition, hr_cate_fields$Department),
margin=2)*100, 2)
round(prop.table(table(hr_cate_fields$Attrition, hr_cate_fields$EducationField),
margin=2)*100, 2)
round(prop.table(table(hr_cate_fields$BusinessTravel, hr_cate_fields$Department),
margin=2)*100, 2)

# Chi-square test of independence
chisq.test(hr_cate_fields$Attrition, hr_cate_fields$BusinessTravel)
chisq.test(hr_cate_fields$Attrition, hr_cate_fields$Department)
chisq.test(hr_cate_fields$Attrition, hr_cate_fields$EducationField)
chisq.test(hr_cate_fields$BusinessTravel, hr_cate_fields$Department)

# Mosaic plot
mosaicplot(table(hr_cate_fields$Attrition, hr_cate_fields$BusinessTravel), shade=T,
main="")
mosaicplot(table(hr_cate_fields$Attrition, hr_cate_fields$Department), shade=T, main="")
mosaicplot(table(hr_cate_fields$Attrition, hr_cate_fields$EducationField), shade=T,
main="")
mosaicplot(table(hr_cate_fields$BusinessTravel, hr_cate_fields$Department), shade=T,
main="")

# Plot like PCA
# Can only do this on Business Travel v.s. DePartment because we need 2 dimensions to
plot, other pairs only generate 1 dimension
# The ca library has a nice correspondence analysis function

c = ca(table(hr_cate_fields$BusinessTravel, hr_cate_fields$Department))
```

```
c$N
c$rowcoord  # 2 dimensions
summary(c)
plot(c)

# The following plot puts arrows to the letters so that we can compare their relative
frequencies to the texts
plot(c, mass=T, contrib="absolute",
    map="rowgreen", arrows=c(F, T))


##There are two different functions - CA and ca
##The below libraries are for the purpose of correspondance analysis (CA - UpperCase)

## reading the data
job = table(hr_cate_fields$Attrition, hr_cate_fields$JobRole)
marital = table(hr_cate_fields$Attrition, hr_cate_fields$MaritalStatus)
gender = table(hr_cate_fields$Attrition, hr_cate_fields$Gender)
edu = table(hr_cate_fields$EducationField, hr_cate_fields$JobRole)

colnames(job) =
c("Health.Rep","HR","Lab.Tech","Manger","Manu.Dir","Rsch.Dir","Rsch.Sci","Sales.Exe","Sales.Rep")
colnames(edu) =
c("Health.Rep","HR","Lab.Tech","Manger","Manu.Dir","Rsch.Dir","Rsch.Sci","Sales.Exe","Sales.Rep")

round(prop.table(table(hr_cate_fields$Attrition, hr_cate_fields$JobRole))*100,2)
round(prop.table(table(hr_cate_fields$Attrition, hr_cate_fields$MaritalStatus))*100,2)
round(prop.table(table(hr_cate_fields$Attrition, hr_cate_fields$Gender))*100,2)
round(prop.table(table(hr_cate_fields$EducationField, hr_cate_fields$JobRole))*100,2)


chisq.test(hr_cate_fields$Attrition, hr_cate_fields$Gender)
chisq.test(hr_cate_fields$Attrition, hr_cate_fields$MaritalStatus)
chisq.test(hr_cate_fields$Attrition, hr_cate_fields$JobRole)
chisq.test(hr_cate_fields$EducationField, hr_cate_fields$JobRole)


mosaicplot(job, shade =T, main ="Attrition versus Job Role")
mosaicplot(gender, shade =T, main ="Attrition versus Gender")
mosaicplot(marital, shade =T, main ="Attrition versus Marital Status")
mosaicplot(edu, shade =T, main ="Educational Field versus Job Role")

## correspondance analysis
corres = ca(edu)
summary(corres)
```

```
## $N gives our original data set
corres$N

## computes the eigen vectors for the rows
corres$rowcoord
rowC = corres$rowcoord[, 1:2]
rowC[order(rowC[,1]), ]
rowC[order(rowC[,2]), ]

## plots the correspondance for rows only
plot(corres, what=c("all","none"))

## computes the eigen vectors for the colums
corres$colcoord
colC = corres$colcoord[, 1:2]
colC[order(colC[,1]), ]
colC[order(colC[,2]), ]

## plots the correspondance for columns only
plot(corres, what=c("none","all"))

## plot the correspondance for rows and columns
plot(corres)

## plot the arrows
plot(corres, mass=T, contrib="absolute", map="rowgreen", arrows=c(F, T))


################################## Correspondence Analysis - END ##############


###########Perform Logistic regression for Attrition as dependent variable with Factor
Data ###############

ds = read.csv("factor_data_with_MI.csv")

head(ds)

str(hr_cate_fields)
Attrition = hr_cate_fields$Attrition

ds_factor_logistic = cbind(ds, Attrition)
str(ds_factor_logistic)

ds_factor_logistic$Attrition<- ifelse(ds_factor_logistic$Attrition=="Yes",1,0)
str(ds_factor_logistic)
```

```
set.seed(123)
indices1 = sample.split(ds_factor_logistic$Attrition, SplitRatio = 0.7)
train1 = ds_factor_logistic[indices1,]
test1 = ds_factor_logistic[!(indices1),]

model_3 = glm(Attrition ~ ., data = train1, family = "binomial")
summary(model_3)
confint(model_3)
exp(coef(model_3))
anova(model_3, test ="Chisq")

model_4<- stepAIC(model_3, direction="both")
summary(model_4)
vif(model_4)

confint(model_4)
exp(coef(model_4))
anova(model_4, test ="Chisq")

#Tells if the model is significant or not
with(model_4, null.deviance - deviance)
with(model_4, df.null - df.residual)
with(model_4, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))

varImp(model_4)

lrtest(model_3, model_4)

# Predict on test data

test_pred1 = predict(model_4, type = "response", newdata = test1)
test_pred1 <- ifelse(test_pred1 > 0.5,1,0)

misClasificError1 <- mean(test_pred1 != test1$Attrition)
accuracy(test1$Attrition,test_pred1)
print(paste('Accuracy',1-misClasificError1)) #Accuracy is 0.84 is a good result.

######################## Logistic Regression with factor data END#################
############ Performing Lasso for logistic regression with factor data ################
set.seed(470)

#Make a matrix of plots with train1
ggpairs(train1)

# lasso is obtained by setting alpha = 1 in library(glmnet)
# Separate the X's and Y's as matrices
```

```
head(train1)
head(test1)
xTrain = as.matrix(train1[, -5])   # Take out "Attrition", column 5
yTrain = as.matrix(train1[, 5])    # Take only "Attrition", column 5

xTest = as.matrix(test1[, -5])   # Take out "Attrition", column 5
yTest = as.matrix(test1[, 5])    # Take only "Attrition", column 5

lRange = seq(0, 5, .1)
fitLasso = glmnet(xTrain, yTrain, alpha=1, lambda=lRange, family = "binomial")

plot(fitLasso, xvar="lambda")

fitLasso
fitLasso = cv.glmnet(xTrain, yTrain, alpha=1, nfolds=10, family = "binomial")
fitLasso$lambda.min
fitLasso$lambda.1se

plot(fitLasso)
coef(fitLasso, s="lambda.min") # Everything got selected
coef(fitLasso, s="lambda.1se") # Only Experience_With_Company is selected

lassoPred = predict(fitLasso, xTrain, s="lambda.min")
rmseLasso_Train = sqrt(mean((lassoPred - yTrain)^2))
rmseLasso_Train

# To predict with this model
lassoPred_2 = predict(fitLasso, xTest, s="lambda.min")
rmseLasso_Test = sqrt(mean((lassoPred_2 - yTest)^2))
rmseLasso_Test

# Compute rmse for training set - with lambda = lambda.1se
lassoPred_3 = predict(fitLasso, xTrain, s="lambda.1se")
rmseLasso_Train2 = sqrt(mean((lassoPred_3 - yTrain)^2))
rmseLasso_Train2

rmseLasso_Train

# To predict with this model - with lambda = lambda.1se
lassoPred_4 = predict(fitLasso, xTest, s="lambda.1se")
rmseLasso_Test2 = sqrt(mean((lassoPred_4 - yTest)^2))
rmseLasso_Test2

rmseLasso_Test

#install.packages("plotmo")
#library(plotmo)
```

```
plotres(fitLasso)
summary(fitLasso)
fitLasso

################################## (E.C) Partial Least Square Regression #########

# Cross validation is used to find the optimal number of retained dimensions.
# Then the model is rebuilt with this optimal number of dimensions.
pls.model = plsr(monthlyIncome ~ ., data = hr_num_ord, validation = "CV")
summary(pls.model)
# Visualize cross-validated RMSEP curves
plot(RMSEP(pls.model), legendpos = "topright") # Judge the RMSEP # 7 components

# Find the number of dimensions with lowest cross validation error
cv = RMSEP(pls.model)
best.dims = which.min(cv$val[estimate = "adjCV", , ]) - 1
best.dims # 6 components

# Rerun the model
pls.model = plsr(monthlyIncome ~ ., data = hr_num_ord, ncomp = best.dims)
summary(pls.model)

# Once the number of components has been chosen, we can inpect different aspects of the fit by plotting
# predictions, scores, loadings, etc.
plot(pls.model, ncomp = 6, asp = 1, line = TRUE) # prediction plot
plot(pls.model, plottype = "scores", comps = 1:3) # a pairwise plot of the score values for the first three components

# extract the explained variances explicitly
explvar(pls.model)

# Print the loadings for interpretation purposes
# plot(pls.model, "loadings", comps = 1:2, legendpos = "topleft", labels = "numbers", xlab = "nm") # doesnt work
# abline(h=0)

loading.weights(pls.model)
pls.model$loadings

# predict the monthly income
predict(pls.model, ncomp = 2, data = hr_num_ord)

# Extract the useful information and format the output
# The regression coefficients are normalized so their absolute sum is 100 and the result is sorted
coefficients = coef(pls.model)
```

```
sum.coef = sum(sapply(coefficients, abs))
coefficients = coefficients * 100 / sum.coef
coefficients = sort(coefficients[, 1 , 1])
barplot(tail(coefficients, 5)) # Job Level, Training Time Last Year, and Years Since Last
Promotion are positive predictors of Monthly Income
barplot(head(coefficients, 5)) # to see that at the other end of the scale what are negative
predictors # Years at Company, Environmental Satisfication


####################Dependent Variable: Attrition ##########################
############ Partial Least Square Regression seems not that useful to Attrition #######

# Cross validation is used to find the optimal number of retained dimensions.
# Then the model is rebuilt with this optimal number of dimensions.
pls.model2 = plsr(num_attrition ~ ., data = hr_num_ord, validation = "CV")
summary(pls.model2)
# Visualize cross-validated RMSEP curves
plot(RMSEP(pls.model2), legendpos = "topright") # Judge the RMSEP # seems 6
components?

# Find the number of dimensions with lowest cross validation error
cv = RMSEP(pls.model2)
best.dims = which.min(cv$val[estimate = "adjCV", , ]) - 1
best.dims # 10 or 11 components # A LOT

# Rerun the model
pls.model2 = plsr(num_attrition ~ ., data = hr_num_ord, ncomp = best.dims)
summary(pls.model2)

# Once the number of components has been chosen, we can inpect different aspects of the
fit by plotting
# predictions, scores, loadings, etc.
plot(pls.model2, ncomp = 11, asp = 1, line = TRUE) # prediction plot # not useful to Attrition
plot(pls.model2, plottype = "scores", comps = 1:3)
# extract the explained variances explicitly
explvar(pls.model2)
# Print the loadings for interpretation purposes
# plot(pls.model2, "loadings", comps = 1:2, legendpos = "topleft", labels = "numbers", xlab =
"nm") # doesnt work
# abline(h=0)

# predict Attrition # Not useful to Attrition
# predict(pls.model2, ncomp = 11, data = hr_num_ord)

# Extract the useful information and format the output
# The regression coefficients are normalized so their absolute sum is 100 and the result is
sorted
```

```
coefficients = coef(pls.model2)
sum.coef = sum(sapply(coefficients, abs))
coefficients = coefficients * 100 / sum.coef
coefficients = sort(coefficients[, 1 , 1])
barplot(tail(coefficients, 5)) # Job Satisfaction, Environment Satisfaction, and Work Life Balance are positive predictors of Attrition
barplot(head(coefficients, 5)) # to see that at the other end of the scale what are negative predictors # Number Companies Worked, Years Since Last Promotion
```