HR ANALYTICS ANALYSIS CASE STUDY

SUBMISSION

Group Name:

1. Ashwin Rangarajan - DDA1730137
2. Prachi Prakash – DDA1730058
3. Shikha Chaturvedi – DDA1730288
4. Sumanth Sundaramurthy - DDA1730138

Business Objective – Recognise employee attrition factors for XYZ company and provide insights to minimize the same

XYZ company has Employee attrition rate of around 15% every year. It has an employee strength of around 4000 at any given point in time; this level of attrition is bad for the company because:

- The former employees' project gets delayed which makes it difficult to meet timelines
- A sizeable department has to be maintained, for the purpose of recruiting new talent
- New employees have to be trained for the job and/or given time to acclimatize themselves to the company

Business Strategy:

The company wants to analyse the factors that are responsible for employee attrition so remedial steps can be taken to retain employees and curb the attrition rate

The strategy of the analysis is to perform:

- **Univariate and Bivariate analysis:** Visually identifying the trends across different variables in the employee dataset
- **Building and evaluating the model:** Build a logistic model using the training employee dataset, test the model using the test employee dataset and evaluate the model
- **Provide insights on steps to decrease employee attrition**

Problem Solving Methodology

The approach taken here is the **CRISP-DM framework**. The Steps involved in CRISP-DM framework are –

Business Understanding

- XYZ Company has 15% average attrition and needs to be replaced with new Talent Pool. This attrition has a negative impact on the company.
- XYZ Company wants to understand what are the factors that affect Attrition rate and what changes can be done at the workplace to make employees stay and reduce attrition rate.

Data Understanding

- To perform the analysis the analysis we have the required data across 5 files.
- employee_survey_data.csv, general_data.csv, manager_survey_data.csv, in_time.csv, out_time.csv
- The data set contains employee survey information about workplace, their employee information like job role, experience, salary etc. The data set has their swipe in and out times for a period of 1 year and manager survey about the employee performance.
- Data Dictionary provided helps us understand which variables in the data set are ordinal, nominal & Continuous.

Data Preparation

- Verified for deduplication of data and if the given data set contains the information of all employees. Renamed the column names in IN & OUT time file.
- Checked for NA across dataset and handled them appropriately either by imputing or removing them.
- Outliers were treated for the continuous variables.
- Date time fields were converted to the correct date formats.
- Appropriate conversion of variables to factors.
- All data sets were merged to
- Univariate, Bivariate and Correlation Analysis performed to visually identify trends on variables of interest and gain useful insights.

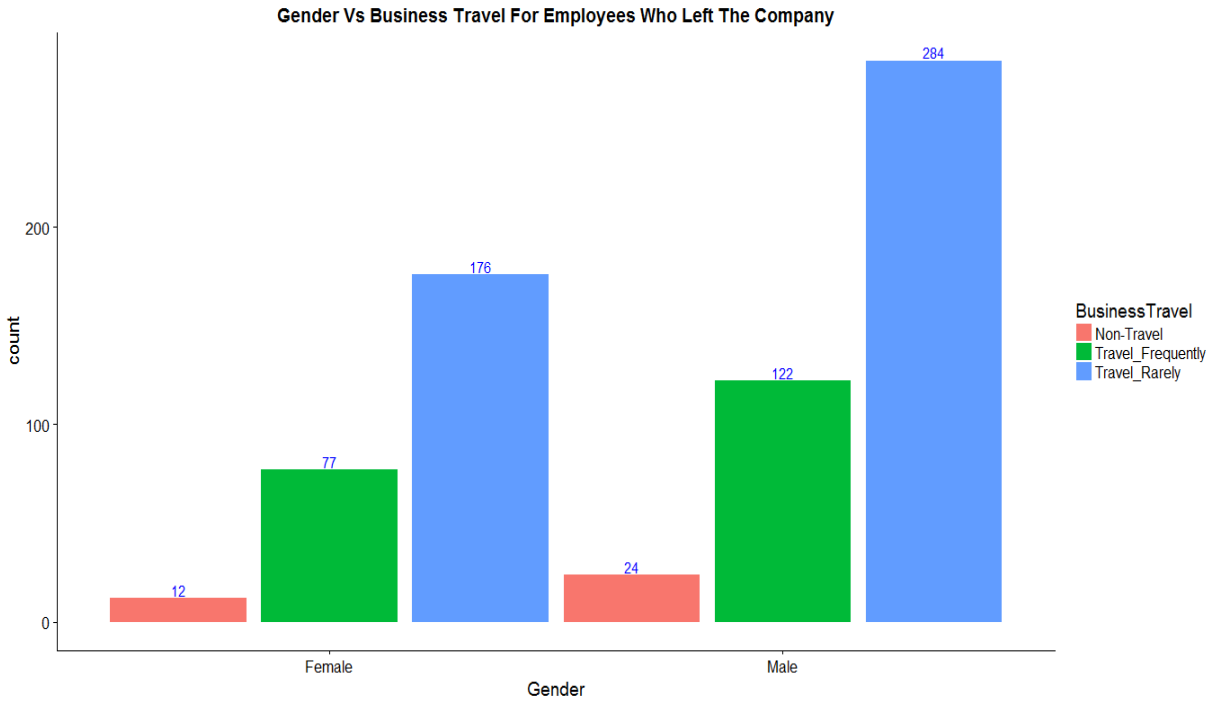
Data Modelling

- Modelling is the heart of data analytics. A model as a which takes relevant data as input and gives an output one is interested in.
- Among the available models we have used Logistic Regression models has been used to build the model.
- Merged Dataset was divided into Test (70%) and Training (30%) data set.
- Factor variables were converted to dummy variables and continuous variables were scaled.
- Model was run iteratively to analyze the effect of independent variables on the model.

Model Evaluation

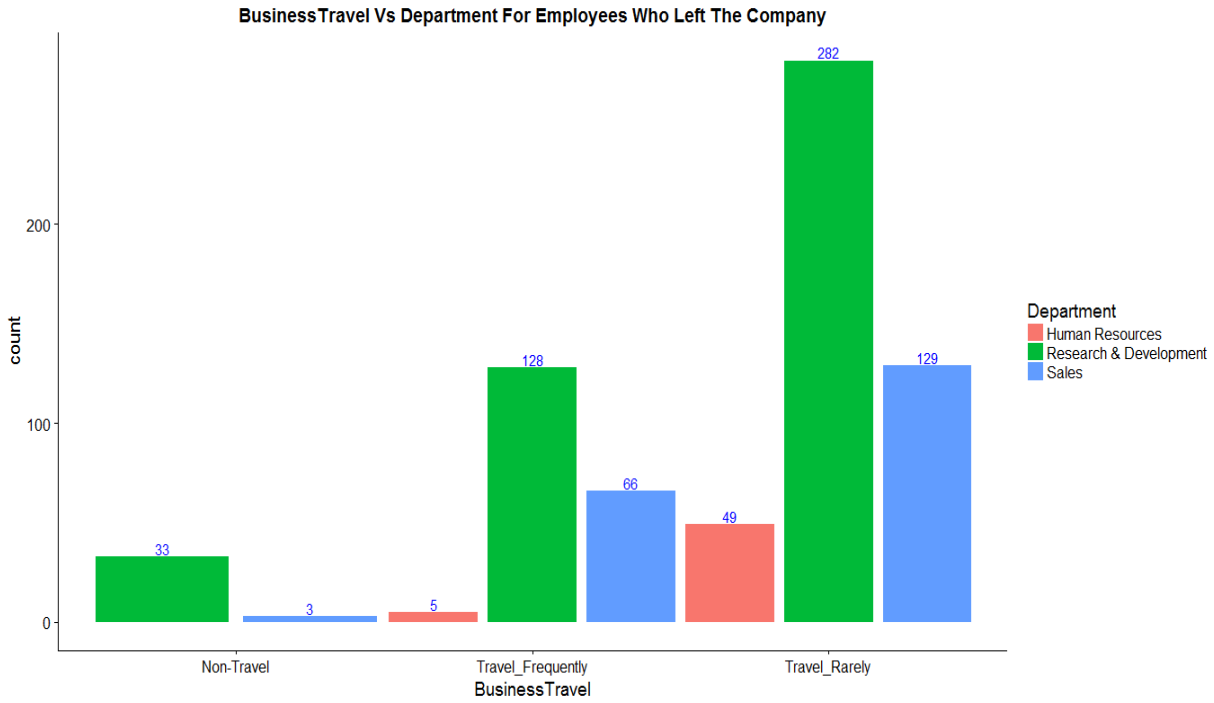
- Model Evaluation is a step where one does the litmus test.
- With the iterative approach, the final optimized model was identified and was run on the Test Data to see if the results are satisfactory.
- Model evaluation methods such as confusion matrix, KS Statistics , lift and gain chart were used to confirm the model accuracy.

Multivariate Analysis



Observation:

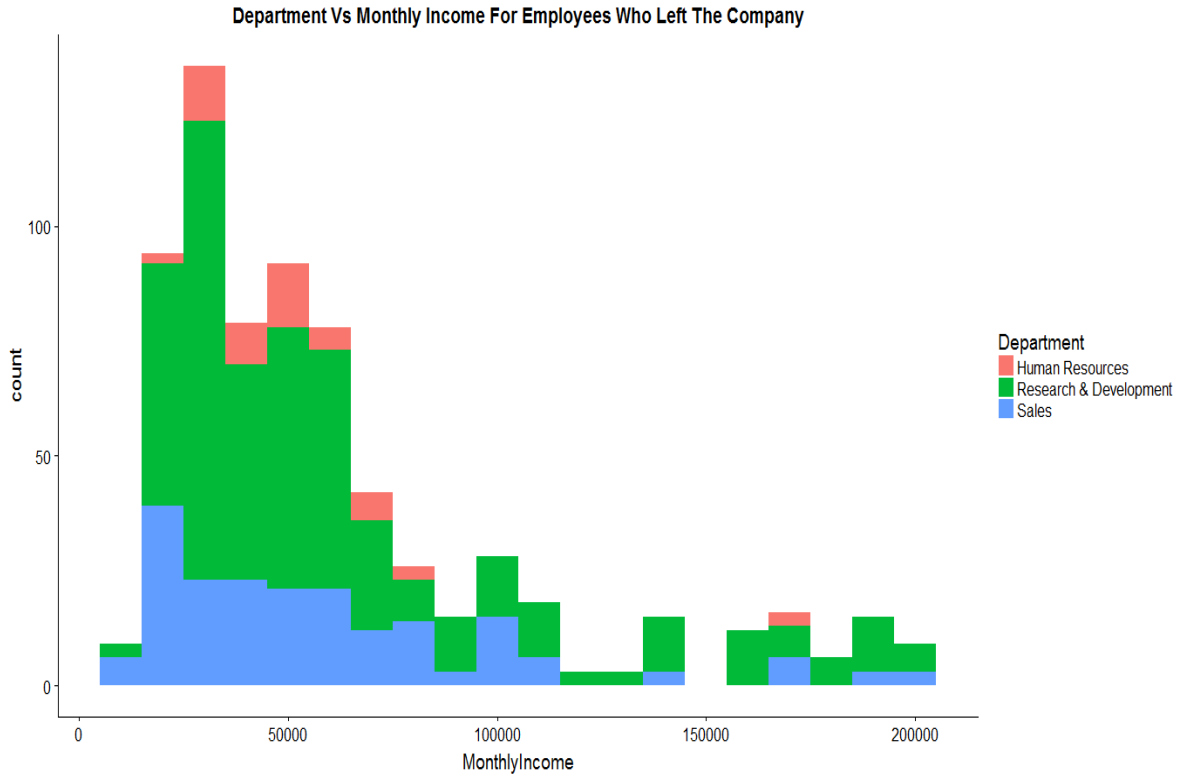
41% of Males and 25% of Females who Travel Rarely tend to leave the company than others.



Observation:

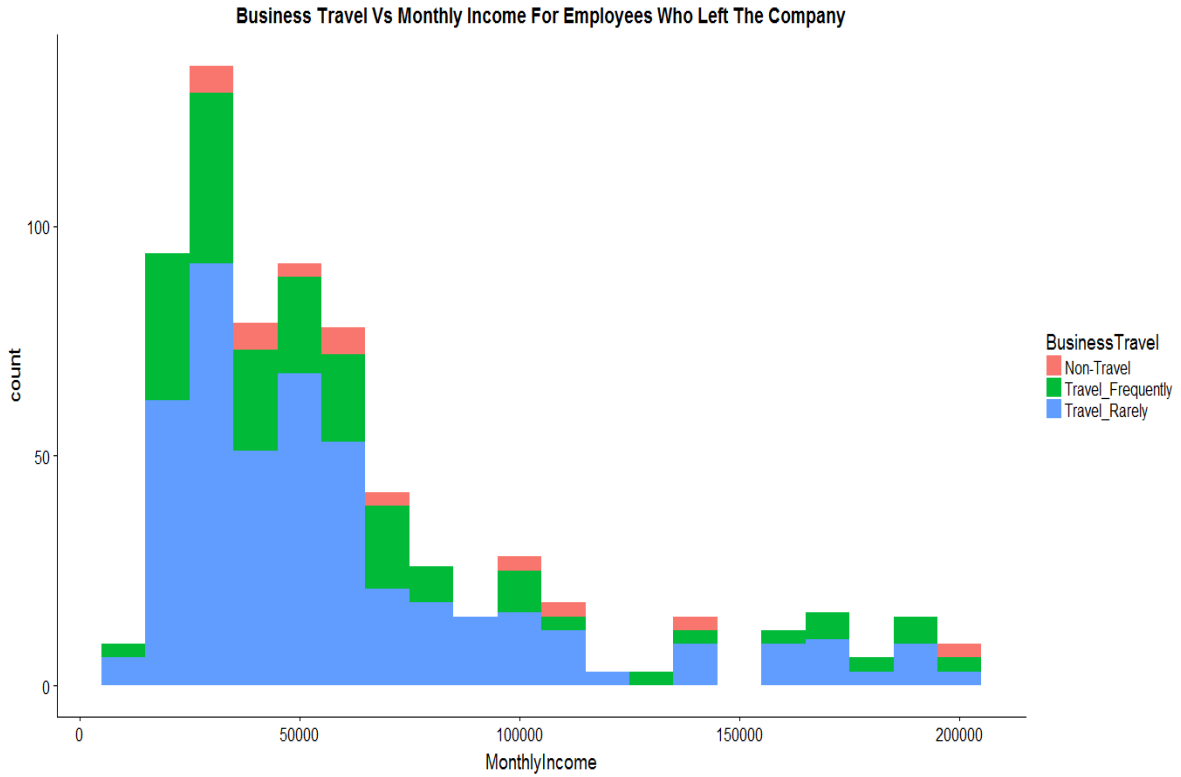
Employees who are in R&D and travel rarely (about 41%) tend to leave the company

Multivariate Analysis



Observation:

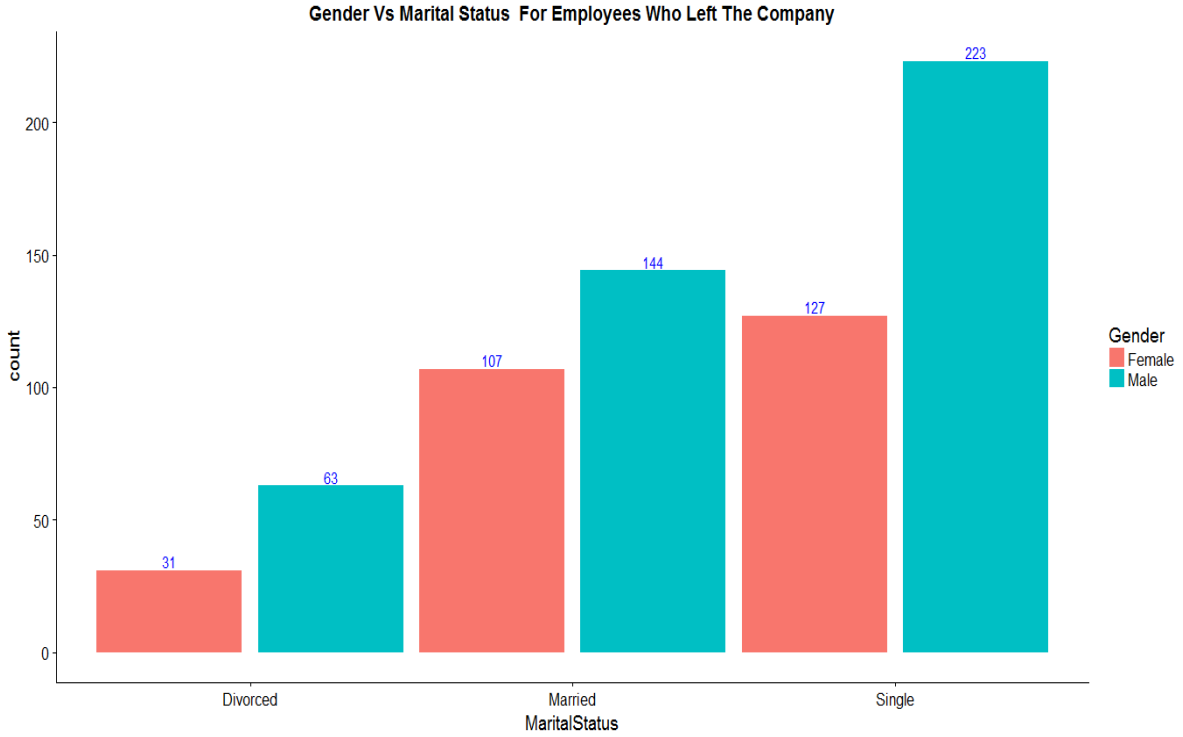
Employees below the income range of Rs. 70,000 from Research & Development have higher attrition rate.



Observation:

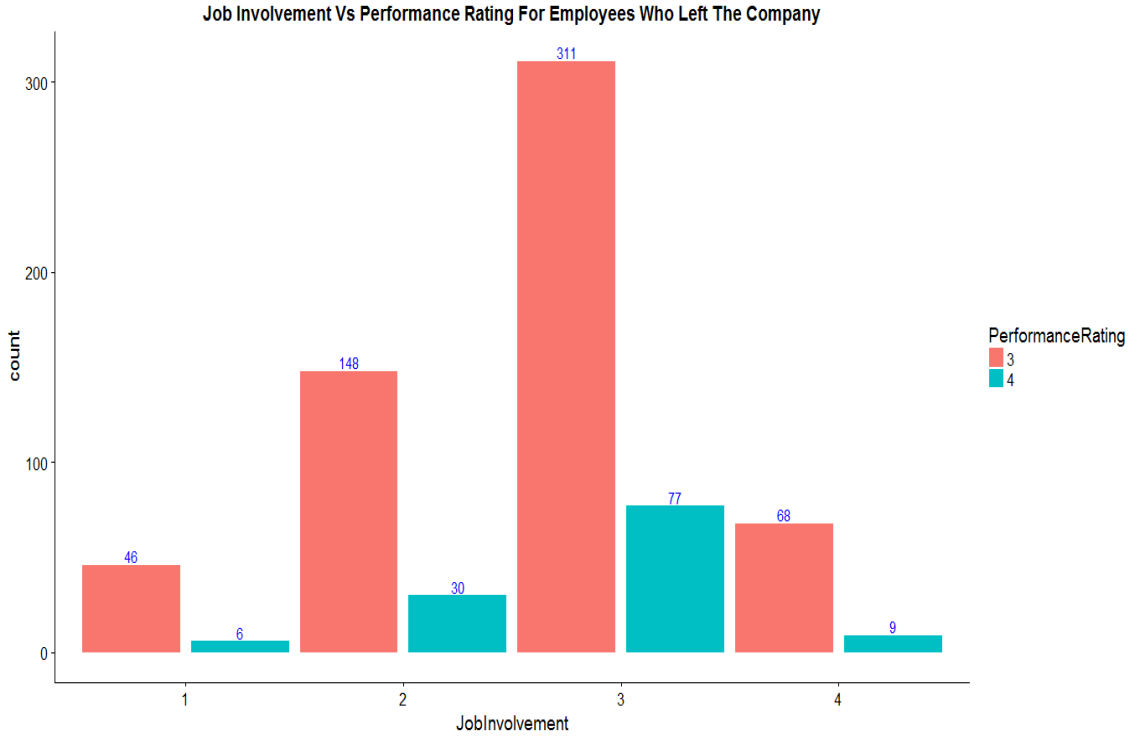
Employees below the income range Rs. 70,000 and who Travel Rarely have higher attrition rate.

Multivariate Analysis



Observation:

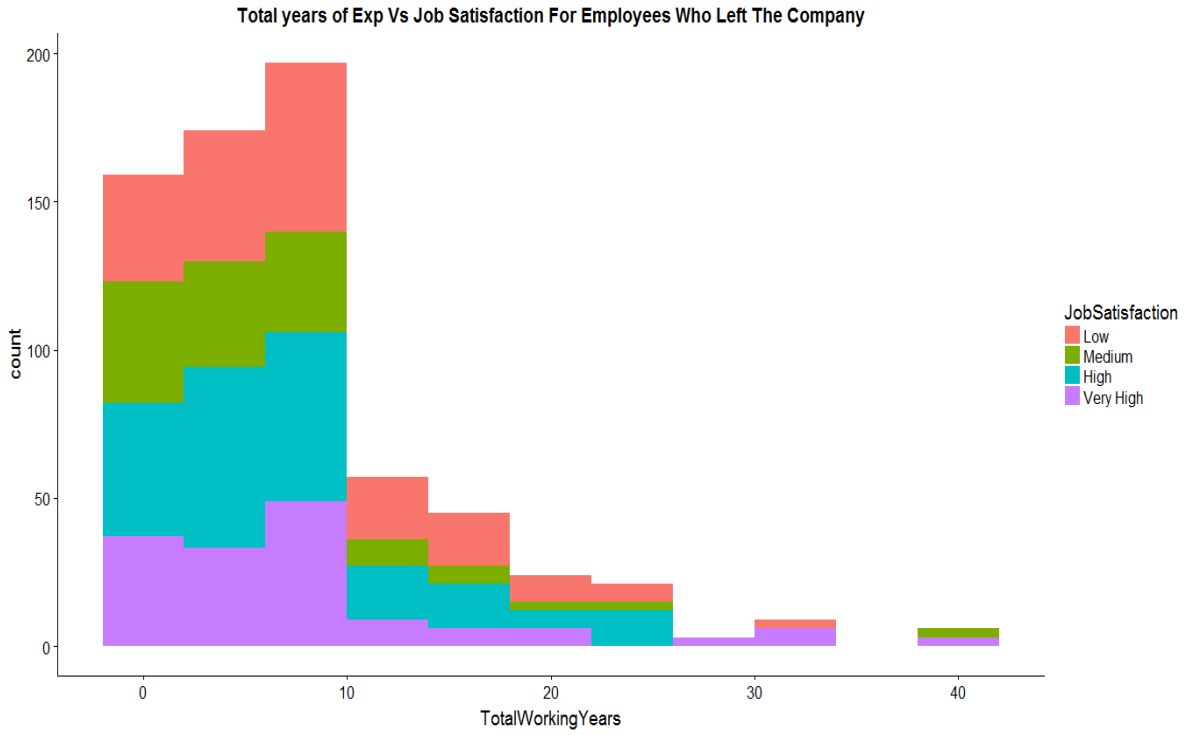
Attrition is higher among Male employees irrespective of the Marital Status. Additionally, 32% of single male employees tend to leave the company



Observation:

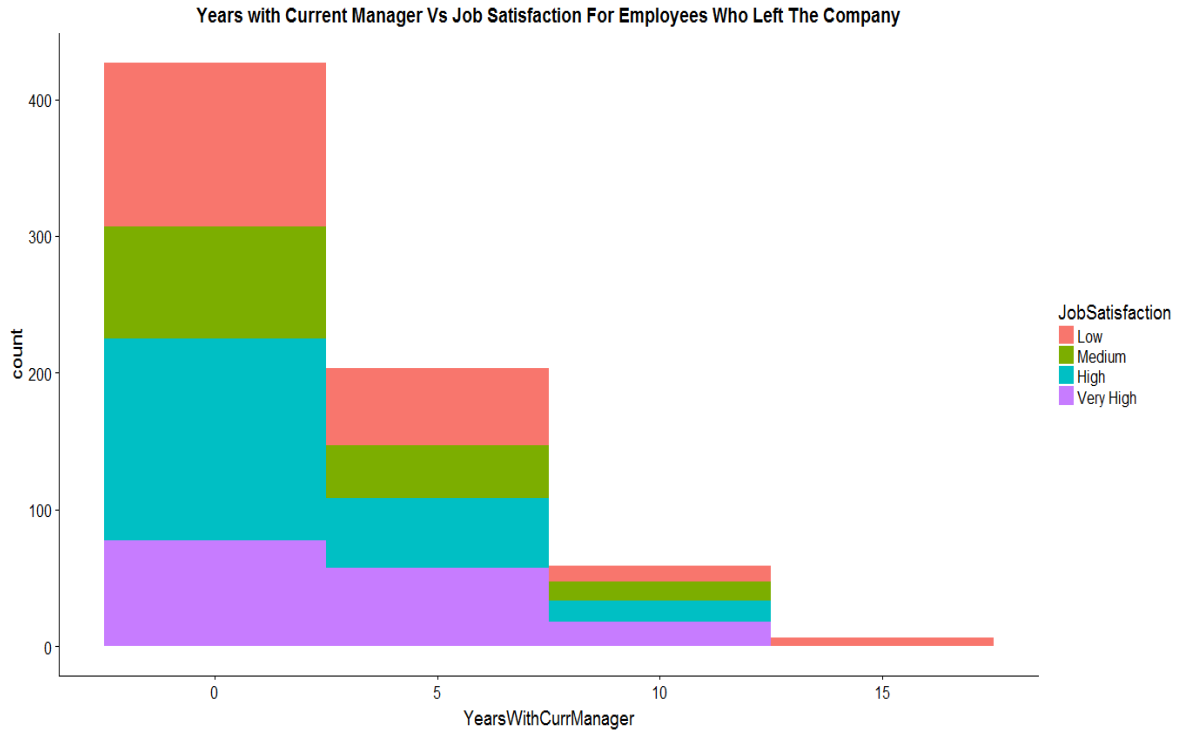
Employees with high job involvement & performance rating of 3 (about 44.7%) tend to leave the company

Multivariate Analysis



Observation:

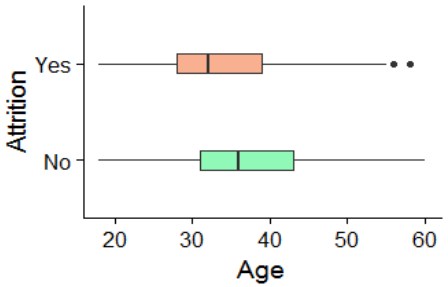
Employees less than 10 years of experience have a higher attrition even though job satisfaction is high.

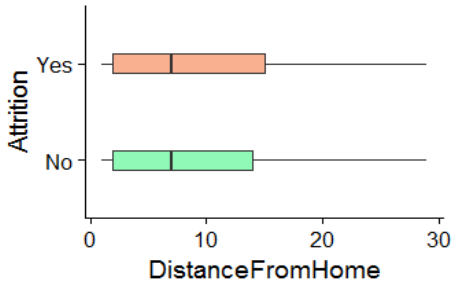


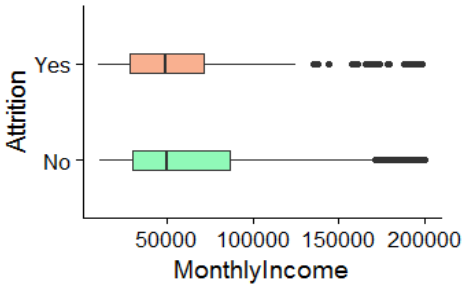
Observation:

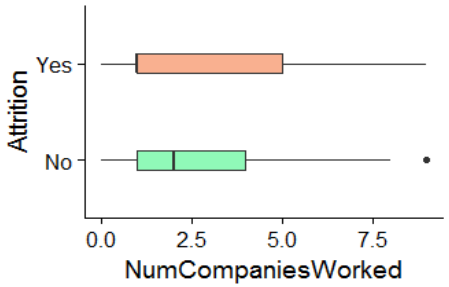
Employees who have been with their current manager for < 2.5 years have a higher attrition despite having high job satisfaction

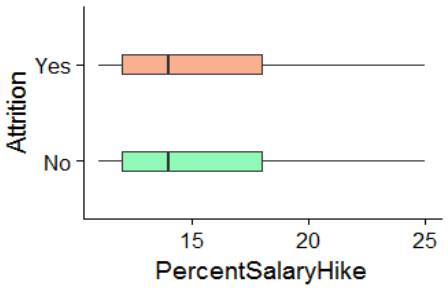
Multivariate Analysis

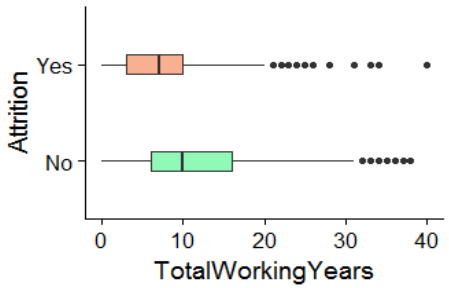


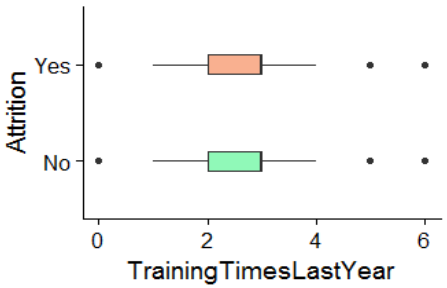


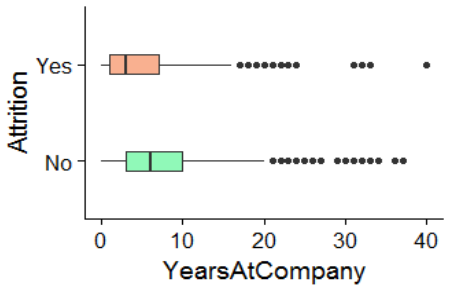


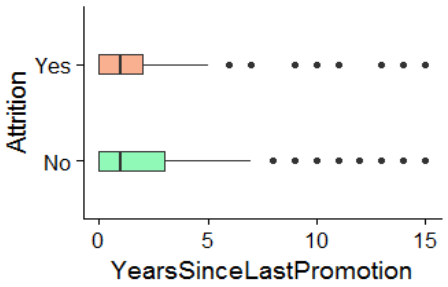


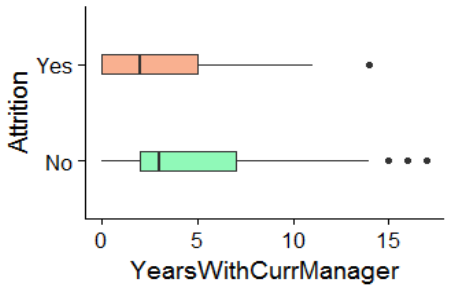


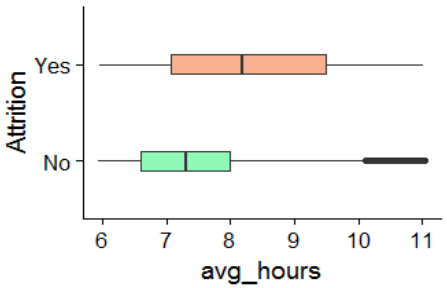








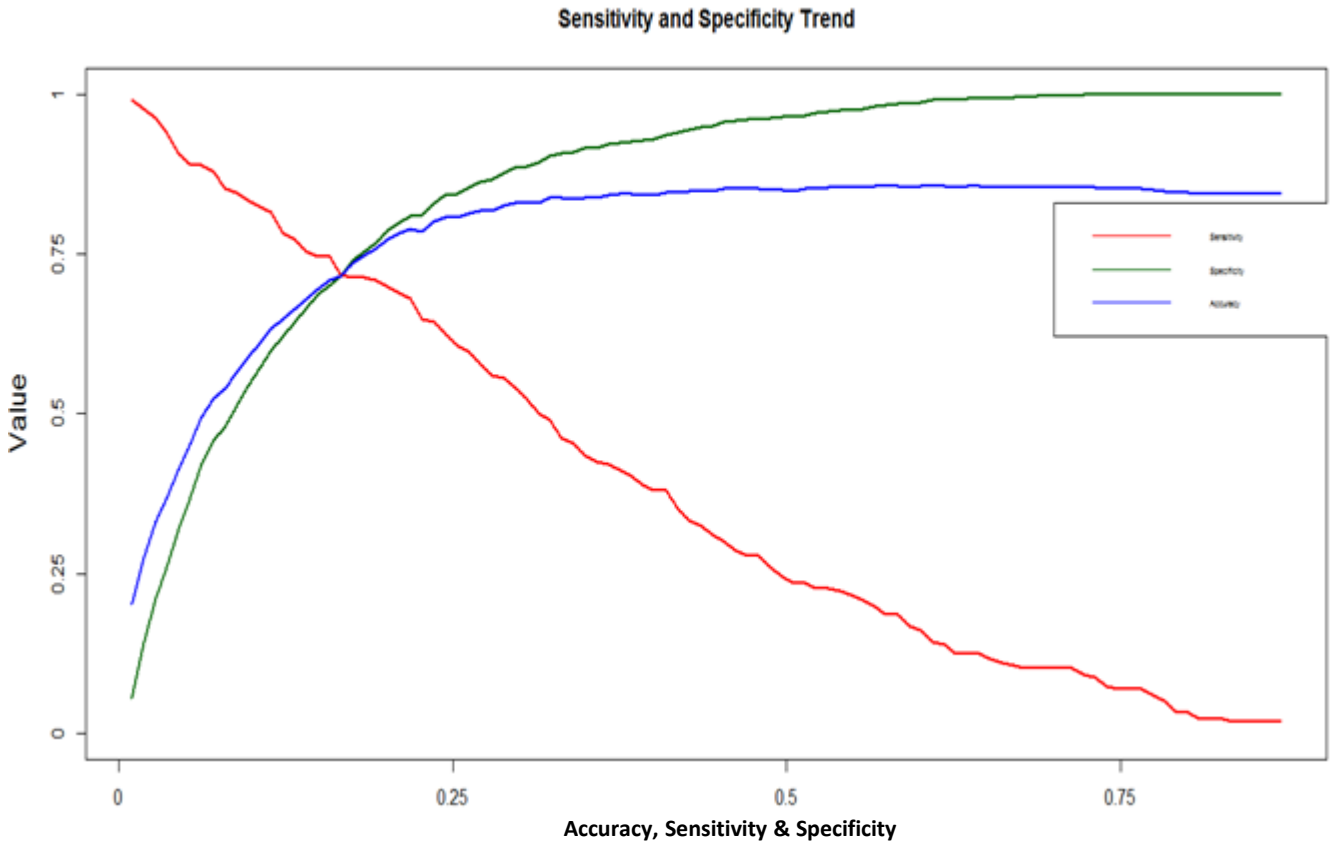




Key factors affecting Employee Attrition -

- Age < 30
- Income < 70,000
- Number of Companies Worked > 4
- Total Working Years < 10
- Average Hours > 8.5

Model Evaluation

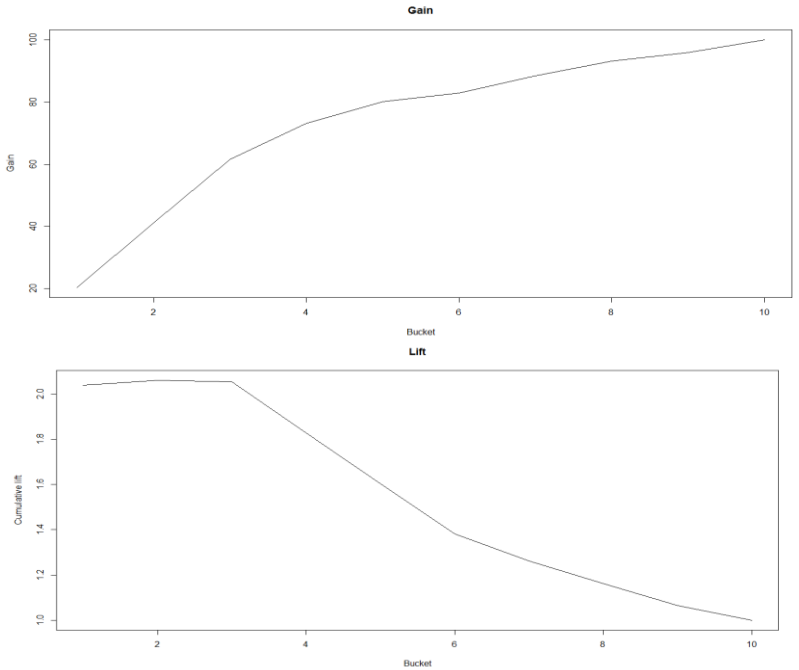


- Methodology – **Accuracy, Sensitivity & Specificity**
  - The prediction gives the probability of attrition of each employee. The best cut off for the probability is the point where **Accuracy & Sensitivity, Specificity** meet.
  - We move ahead to identify the cut off value and plot the graph where the **Sensitivity, Specificity & Accuracy** meet.
  - The cutoff Identified = **0.1663636**
  - Using the cutoff value we will calculate the cut off attrition rate using the test data set and create the Confusion Matrix.
  - Confusion Matrix gives the values of Sensitivity, Specificity & Accuracy –
  - Accuracy – 0.7154
  - Sensitivity – 0.7175
  - Specificity – 0.7150

Model Evaluation

bucket	total	totalresp	Cum-Churn	%Cum - Churn	Non Churn	Cum-Non Churn	%Cum-Non Churn	KS
1	136	44	44	20.37%	92	92	8.04%	12.33%
2	136	45	89	41.20%	91	183	16.00%	25.21%
3	136	44	133	61.57%	92	275	24.04%	37.54%
4	136	25	158	73.15%	111	386	33.74%	39.41%
5	136	15	173	80.09%	121	507	44.32%	35.77%
6	136	6	179	82.87%	130	637	55.68%	27.19%
7	136	12	191	88.43%	124	761	66.52%	21.90%
8	136	10	201	93.06%	126	887	77.53%	15.52%
9	136	6	207	95.83%	130	1017	88.90%	6.93%
10	136	9	216	100.00%	127	1144	100.00%	0.00%
	1360	216			1144			

Sl. No.	Total	Total resp.	Cumm resp.	Gain	Cumlift
1	136	44	44	20.4	2.04
2	136	45	89	41.2	2.06
3	136	44	133	61.6	2.05
4	136	25	158	73.1	1.83
5	136	15	173	80.1	1.6
6	136	6	179	82.9	1.38
7	136	12	191	88.4	1.26
8	136	10	201	93.1	1.16
9	136	6	207	95.8	1.06
10	136	9	216	100	1



- Methodology – KS Statistic
  - It is another indicator of how well the model is performing. It is an indicator how well your model discriminates between the two classes.
  - For the model we have built the KS Statistic Value = 0.4326 or 43.36%
- Methodology – Gain & Lift Chart
  - To show how well our model is performing

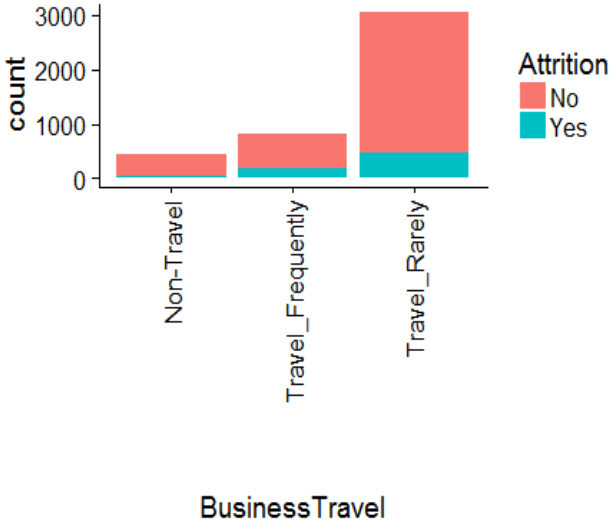
Model Evaluation

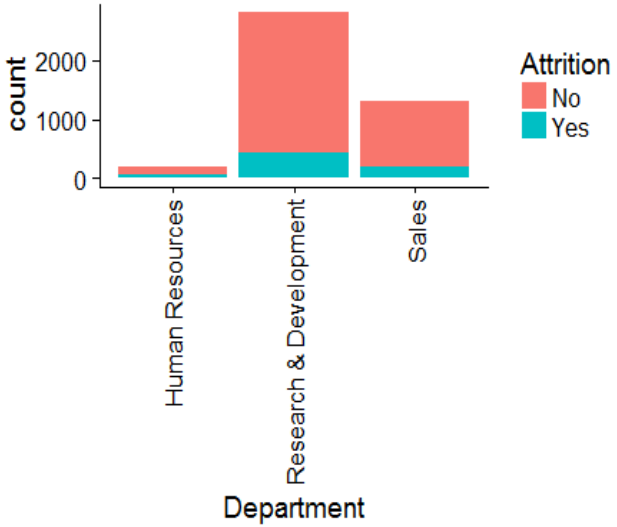
- Interpretation of the Data Collated So far –
  - Interpretation of Accuracy, sensitivity and specificity – Model has high accuracy of 71%, along with balanced specificity and sensitivity thus reflecting the model captures 71% of attrition data correctly
  - KS statistic interpretation – High KS statistic value represents that the model has all the attrition with the 1st to 4th decile
  - Gain – when you look at the gain chart you can see that 73% of events are covered in top 40% of the model. i.e. within the first 40% we can reach to about 73% of employees and stop them from leaving the company.
  - Lift – For the top two decile, cumulative lift is 2, which implies that we can cover 2 times the number of employees leaving the company by selecting only 20% of the employees based on the model as compared to selecting 20% randomly.

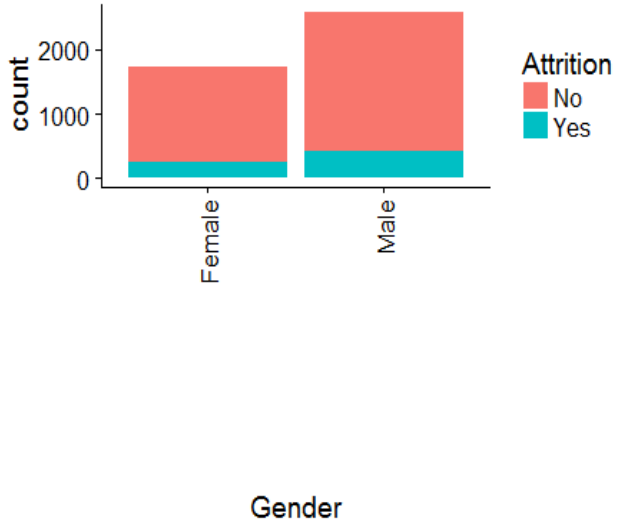
- Key Factors Affecting XYZ Company Attrition per the final model –
  - Age
  - Number of Companies Worked
  - Total Working Years
  - Years Since Last Promotion
  - Years With Current Manager
  - Average Working Hours
  - Business Travel – Employees who Travel Frequently and Travel Rarely.
  - Department – Research & Development, Sales
  - Marita Status – Single
  - Job Satisfaction – Very High
  - Work Life Balance – Good, Better & Best
  - Environmental Satisfaction – Medium, High & Very High

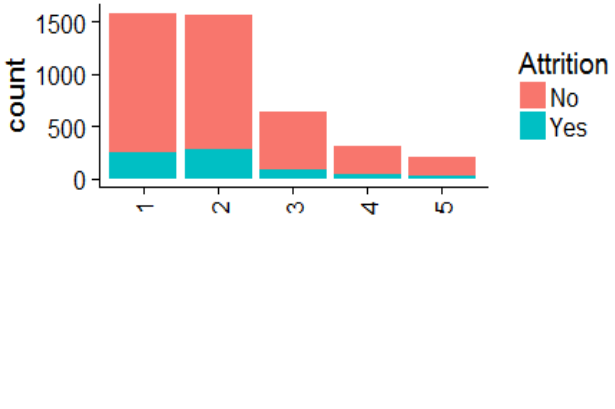
Additional Analysis

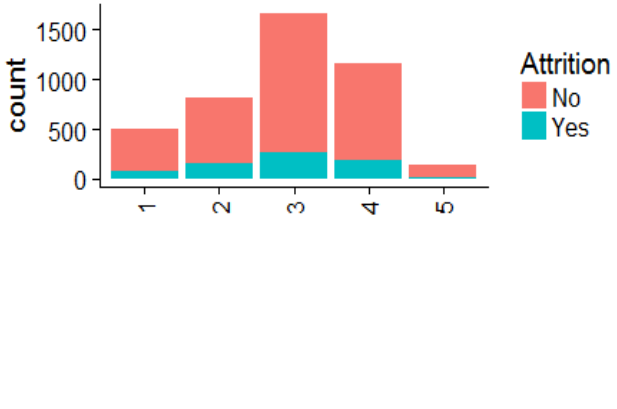
Univariate Analysis – Across Categorical Variables

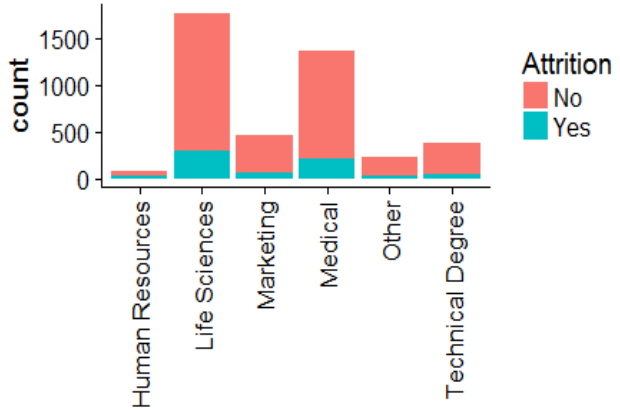








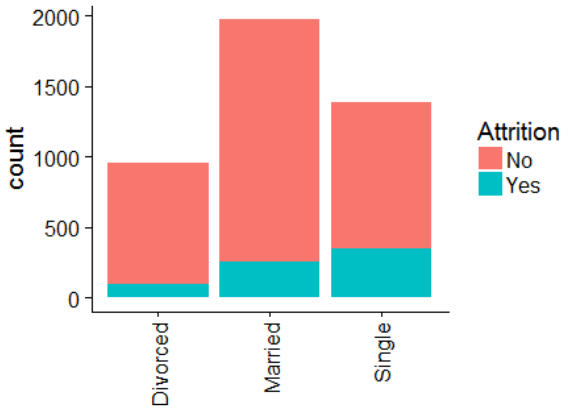


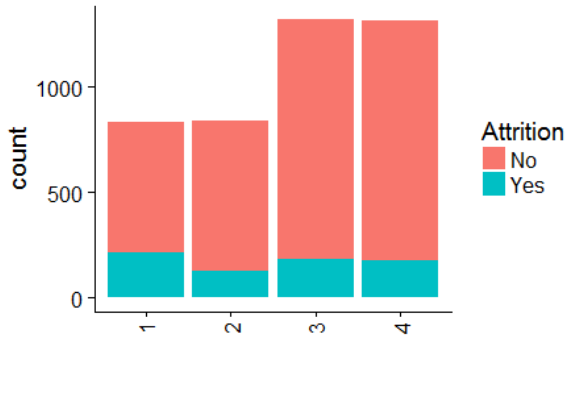


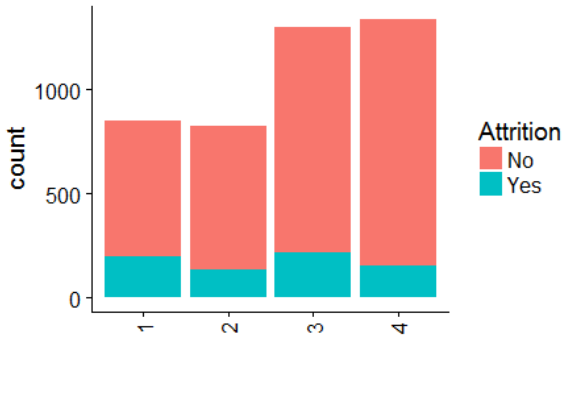
Observations –

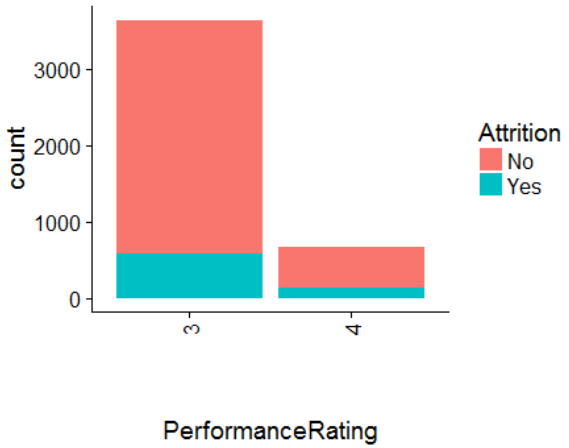
- Employees who travel rarely tend to leave
- Attrition rate is higher among employees in R&D department
- Employees with Job satisfaction rating of 3 tend to have a slightly higher chance of leaving
- Employees with Work life balance, Job Involvement and Performance rating of **3** tend to have a higher chance of leaving

Univariate Analysis – Across Categorical Variables





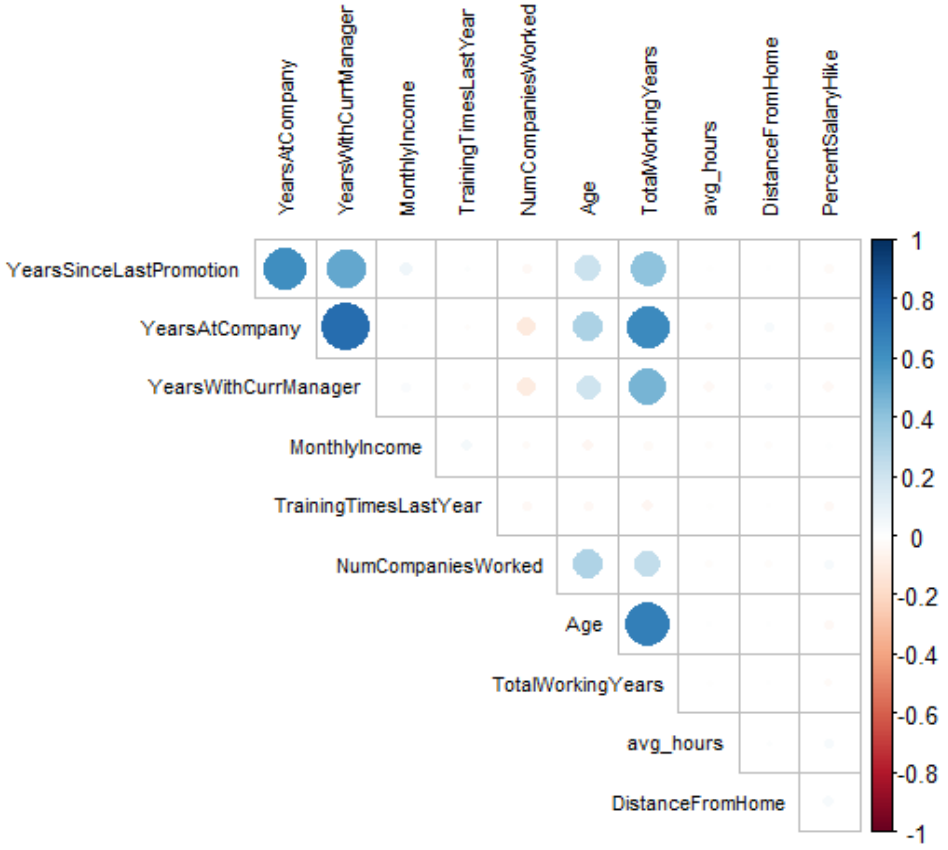


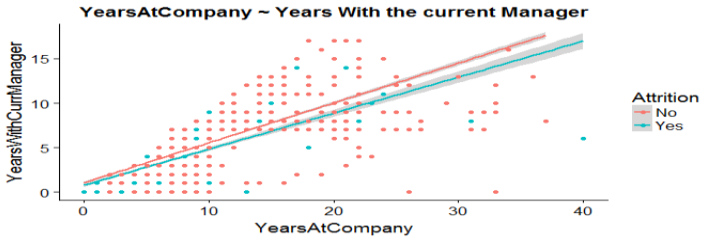


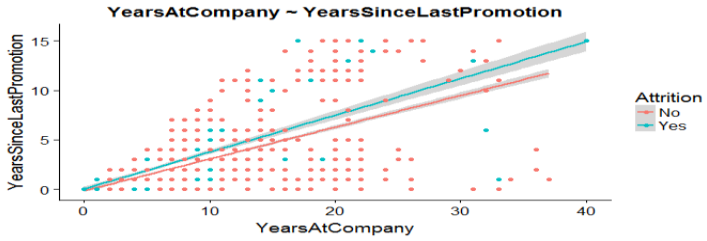
Observations –

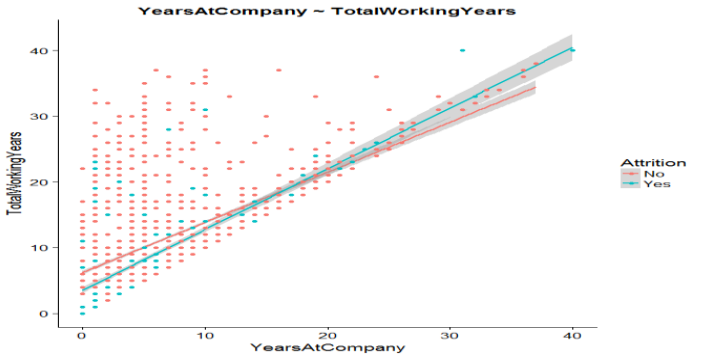
- Employees who are single tend to leave
- Employees with an environment satisfaction of 2 are less likely to leave
- Employees with Job satisfaction rating of 3 tend to have a slightly higher chance of leaving
- Employees with Work life balance, Job Involvement and Performance rating of **3** tend to have a higher chance of leaving

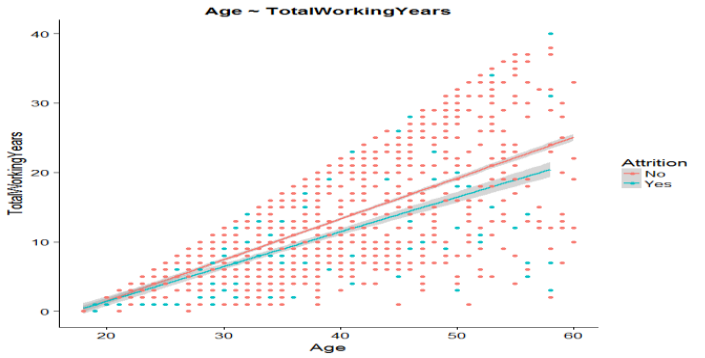
Correlation Matrix for numeric variables











Observations –

- YearsSinceLastPromotion and YearsAtCompany are positively correlated with correlation of 0.551
- YearsAtCompany and YearsWithCurrManager are positively correlated with correlation of 0.76
- YearsAtCompany and TotalWorkingYears are positively correlated with correlation of 0.62

Understanding Data In Hand

For the given problem in hand we have been provided with five data sets –

- ❑ employee_survey_data.csv
- ❑ general_data.csv
- ❑ in_time.csv
- ❑ out_time.csv
- ❑ manager_survey_data.csv

- The employee entry and exit time for 1 year have been provided in the in_time & out_time files.
- Employee related information such as age, department, education, gender etc. are part of the general_data file.
- Survey information filled by managers about their employees based on work performance is available in manager_survey_data file.
- Similarly, survey done by employees about the workplace is available in employee_survey_data file.
- Overall the data in hand is for 4410 employees.
- EmployeeID is unique identifier based on which we have confirmed there are no duplicate records and also used for merging the data sets.
- Looking at the dataset there is about 16% attrition rate for the XYZ company.

Data Cleaning

1. Percentage of NA values in the merged data set –

Attribute	NA Percentage
NumCompaniesWorked	0.43%
TotalWorkingYears	0.20%
EnvironmentSatisfaction	0.57%
JobSatisfaction	0.45%
WorkLifeBalance	0.86%

2. Categorical variables such as "Attrition", "BusinessTravel", "Department", "EducationField", "Gender", "JobLevel", "JobRole", "MaritalStatus", "Over18", "JobSatisfaction", "JobInvolvement", "PerformanceRating", "WorkLifeBalance", "EnvironmentSatisfaction", "Education", "StockOptionLevel" are converted to factor
3. Outliers for continuous variables are identified using boxplot and are treated by capping high and low end values
4. Column names for in_time and out_time data set has been changed to "EmployeeID"

Data Modelling

So far we have seen the EDA analysis and their observations on the variables of merged data set. Now we get to the modelling stage where we will identify which are the variables that affect the attrition rate. Steps that will be followed –

- Feature Scaling – if the continuous variables are not of different scale they do affect the model outcome. So using the **scale()** function the continuous variable will be brought to same order of magnitude.
- Dummy Variable Creation – The categorical variables are converted to dummy variables either by using **model.matrix()** function where the number of levels are more than 2 OR by converting the factors to 1 or 0 where the level is 1 or 2. The Dummy variables are merged with the data set in order for us to have a final data set prior to building the model.
- Test & Train Data Set - Prior to building the model set we need to split the existing data set into Test and Training data set. So that when we have the final model we can run the Training data set on the model and see how well the model predicts.
- Model Building – Building a model is iterative process till we reach the final optimized model. Steps followed –
  - Using glm() we run the whole test data set to create our first model.
  - stepAiC() is run on the first model which help us eliminating all the insignificant variables through multiple iterations. The output model of stepAIC is fed to glm() to create the next model.
  - By comparing the VIF and p-value of the model, iteratively we remove the insignificant variables till the final optimized model is arrived at.
  - Final Model which was arrived at contained the following variables and all of them are significant –
    - Age, NumCompaniesWorked, TotalWorkingYears, YearsSinceLastPromotion, YearsWithCurrManager, avg_hours, BusinessTravel.xTravel_Frequently, BusinessTravel.xTravel_Rarely, Department.xResearch...Development, Department.xSales, MaritalStatus.xSingle, JobSatisfaction.x4, WorkLifeBalance.x2, WorkLifeBalance.x3, WorkLifeBalance.x4, EnvironmentSatisfaction.x2, EnvironmentSatisfaction.x3,EnvironmentSatisfaction.x4,
  - With the final model obtained we will run the prediction using **predict()** function to see how well the model is doing.