HR Analytics - Case Study Presentation

SUBMISSION

1. Shivam Kakkar (Facilitator)
2. Ashwin Suresh
3. Manohar Shanmugasundaram
4. Sundeep Gupta

HR Analytics - Case Study Analysis

Background

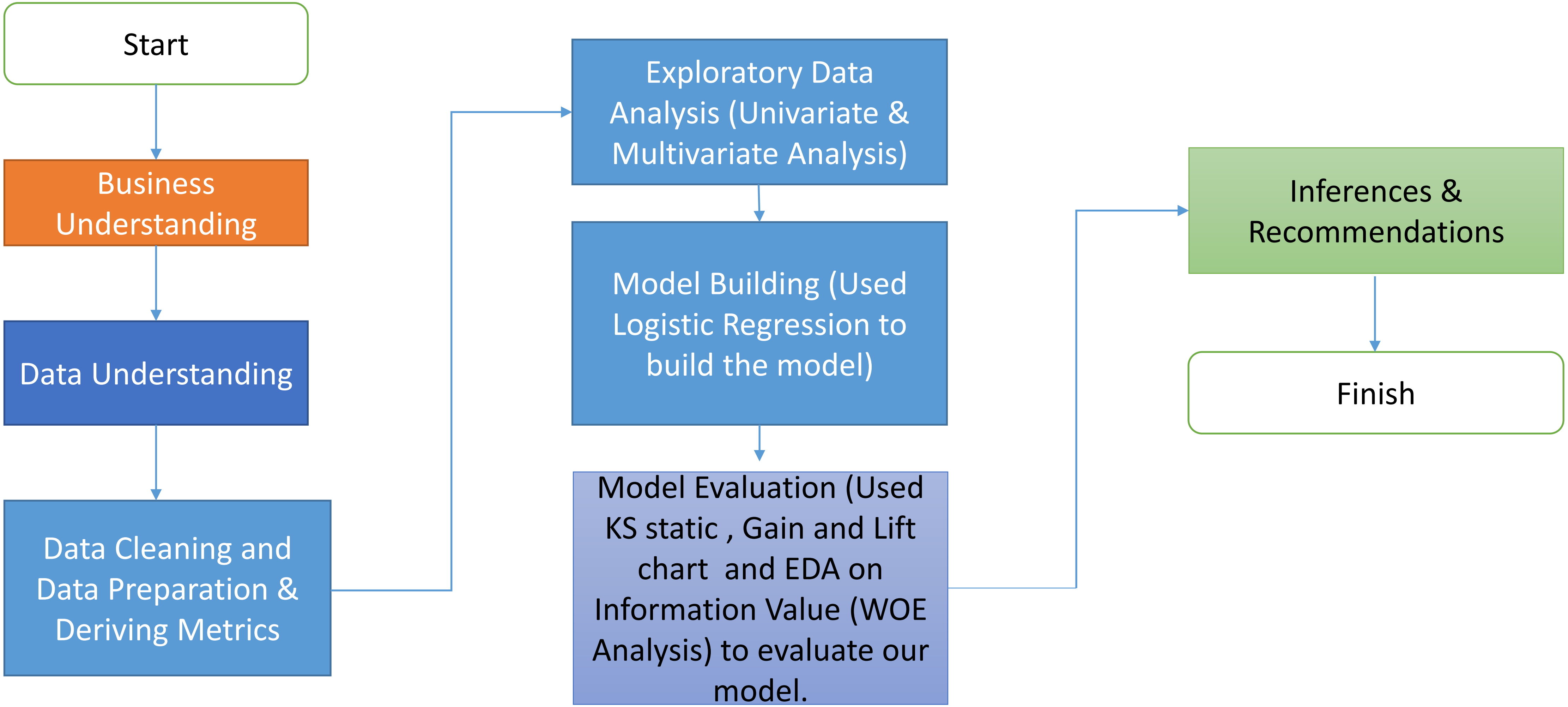
The company ‘**XYZ’** employs around 4000 employees. However, every year around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market.

Business Objectives

The management believes this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company. Hence the **HR analytics** firm has been requested by the management to understand what factors they should focus on, in order to curb attrition.

As part of this analysis, the factors that helps identifying the attrition will be identified and shared with the management.

Leveraging “CRISP-DM Framework” to solve the Problem



Start

Business Understanding

Data Understanding

Data Cleaning and Data Preparation & Deriving Metrics

Exploratory Data Analysis (Univariate & Multivariate Analysis)

Model Building (Used Logistic Regression to build the model)

Model Evaluation (Used KS static , Gain and Lift chart and EDA on Information Value (WOE Analysis) to evaluate our model.

Inferences & Recommendations

Finish

Data Cleaning & Derived Metrics

Data Preparation:

- **Source the data files :**
  - General Data
  - Employee Data
  - Manager Data
  - In time Data
  - Out time Data

Data Cleaning:

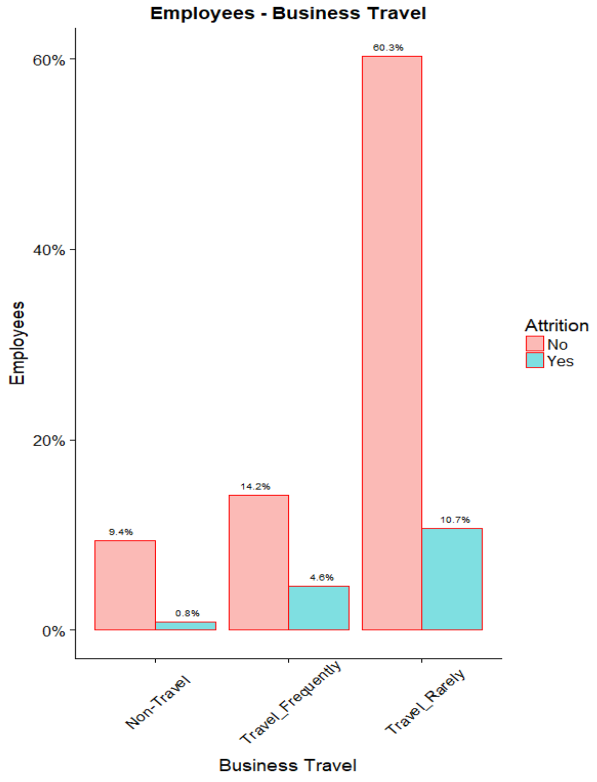
- Identify the empty fields and replaced with the 'NA'.
- Check for 'NA' and duplicated records.
- Validated all the characters are spelt correctly.
- Omitted the NA records as their percentage is very low : 0.6% in "General Data" & 1.8% in Employee Data.
- Ignored the columns which have only one unique value. For example, Employee Count it has all the records with the value as '1' .
- Done the detailed analysis on outliers along with proper rational.
- Identified and removed the outliers in the variables like monthly income (using box plots).
- Checked for any spurious records and reported.

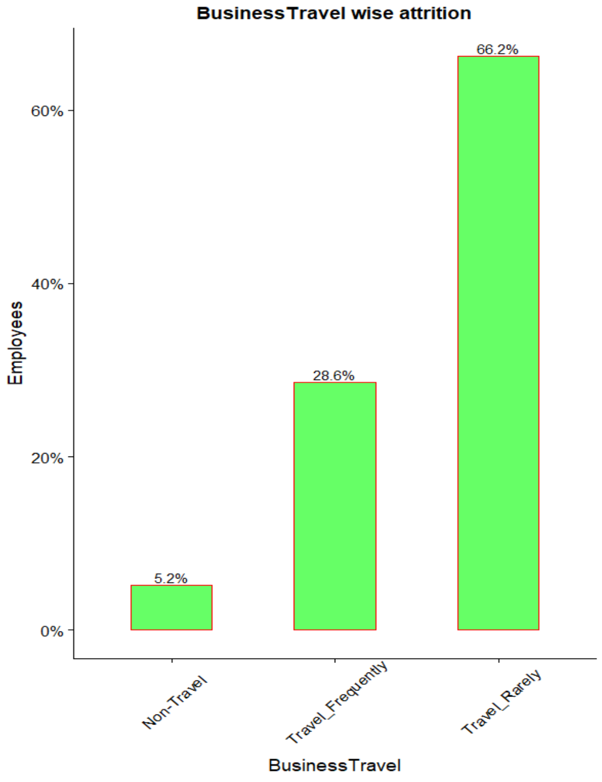
Derived Metrics:

1. **Average Working Hour Per Day (avg_working_hr_per_day)** – Created from the in and out time data sets, this field will represent the average work time.
2. **Number of leaves (no_of_leaves)** – Created from the in and out time data sets, this field will represent the number of days the employee is absence from work.

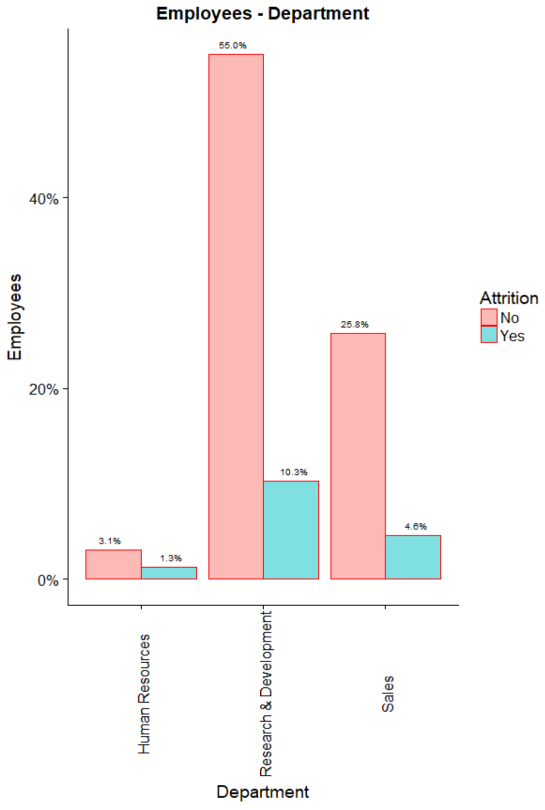
Univariate Analysis On Categorical Variables

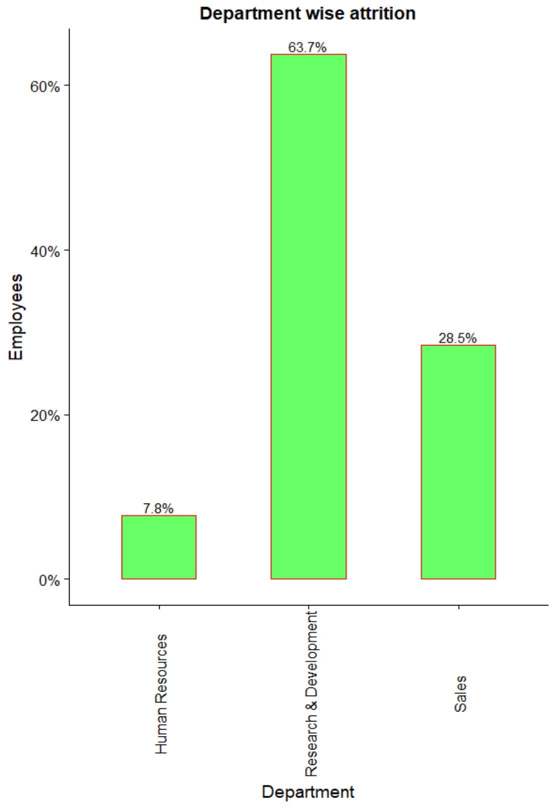
Business Travel



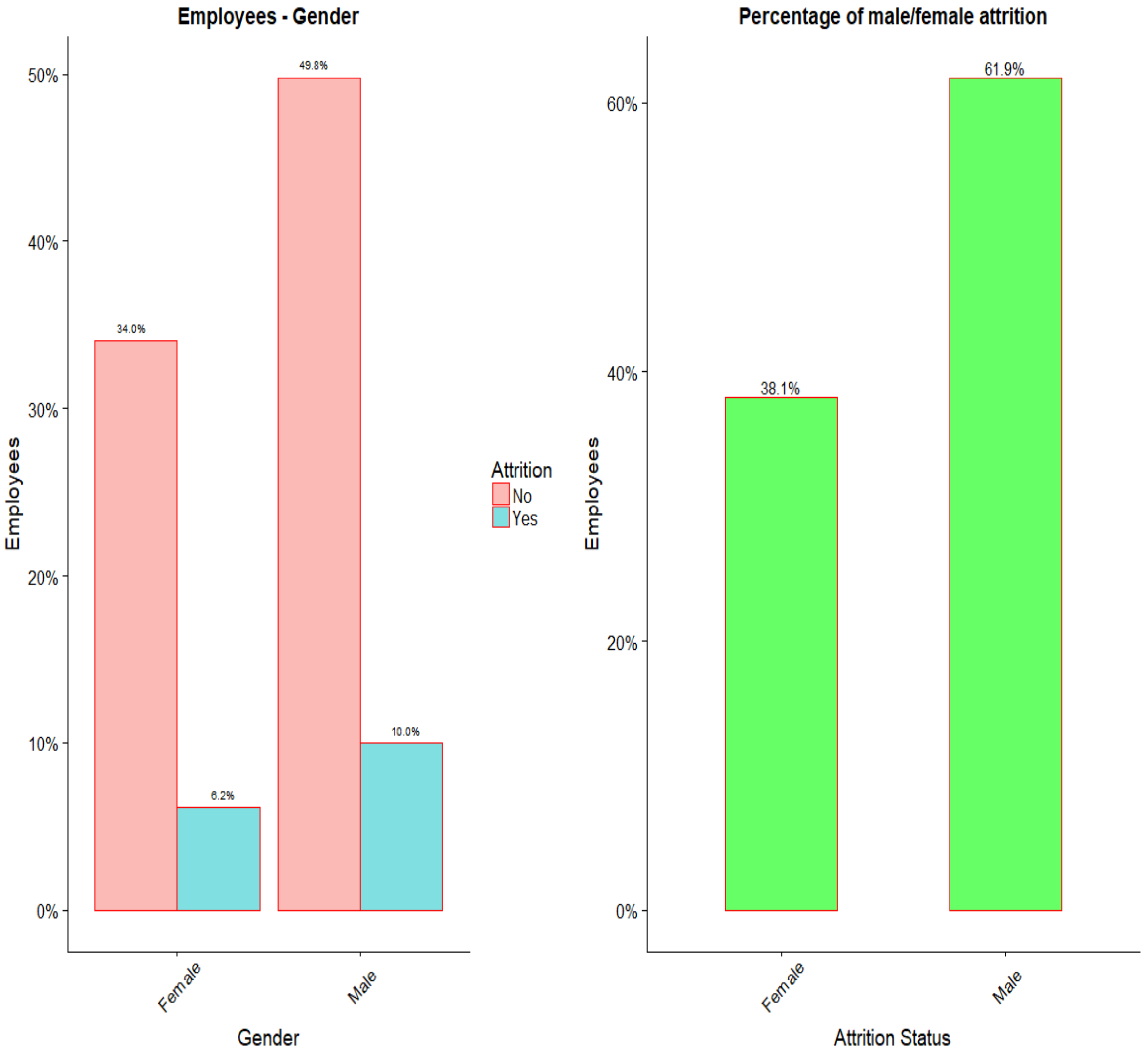


Department

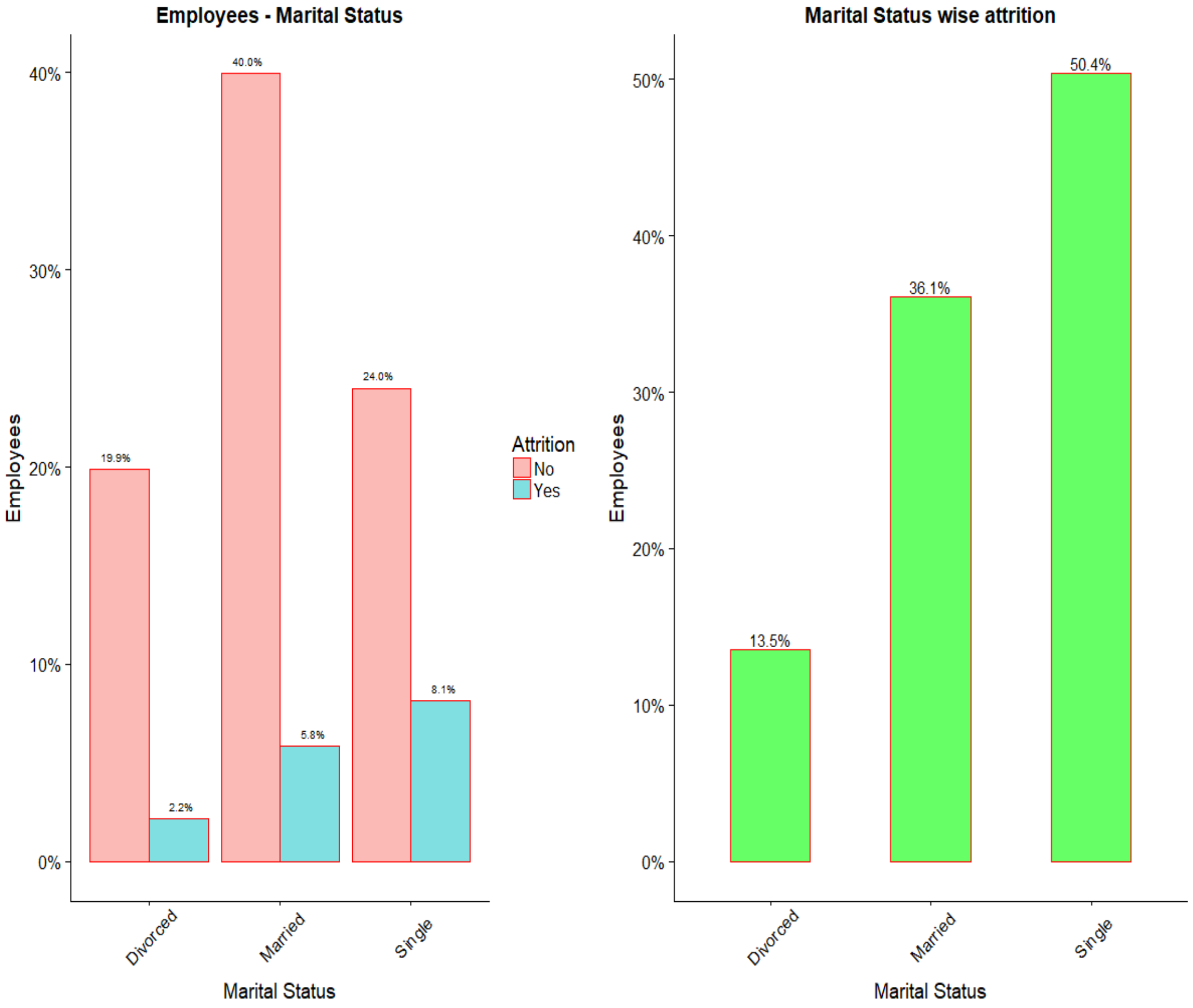




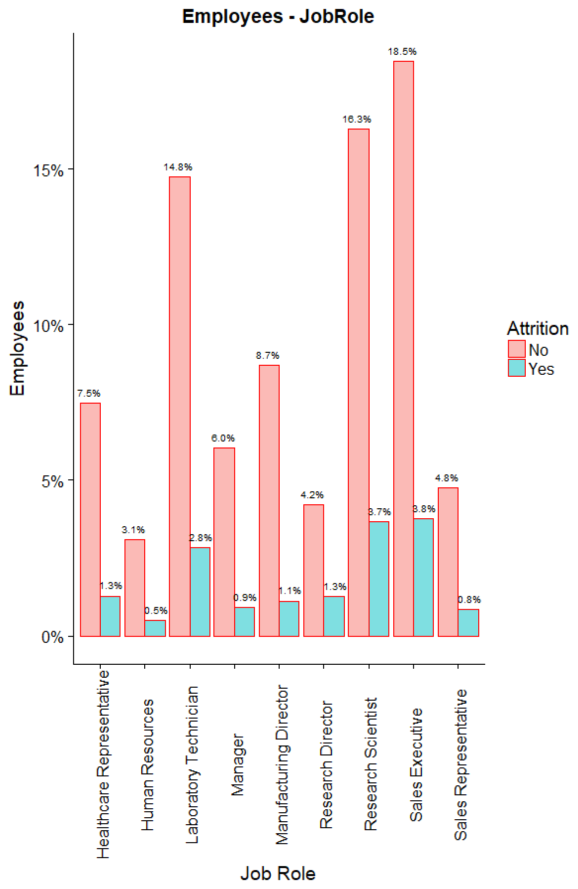
Gender

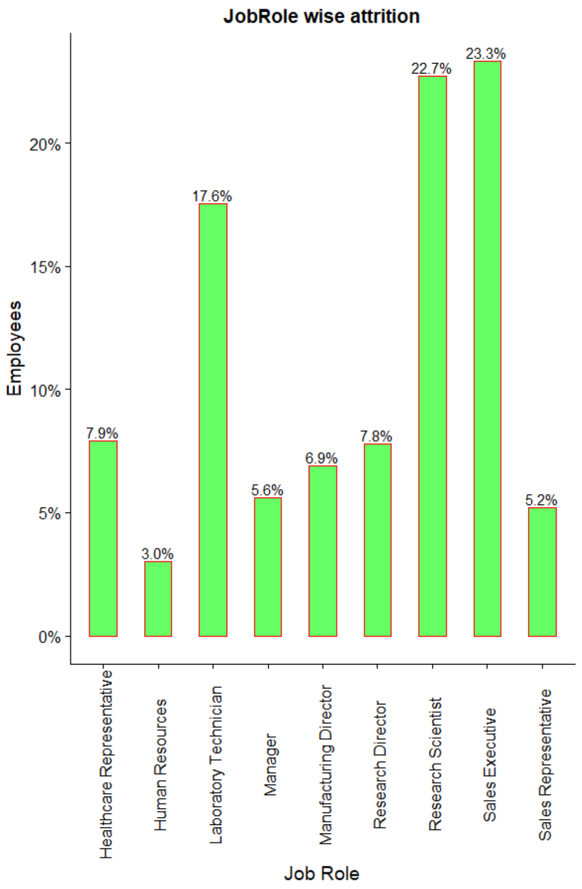


Marital Status

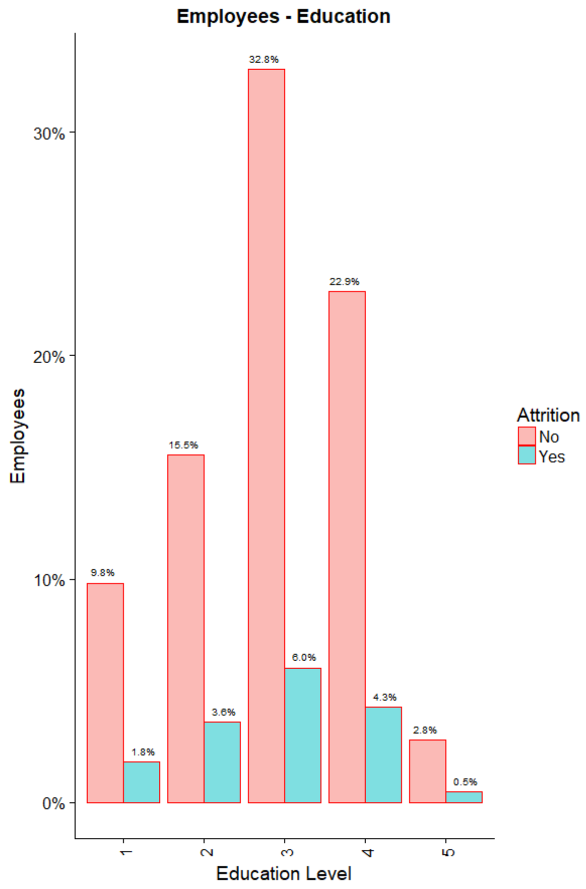


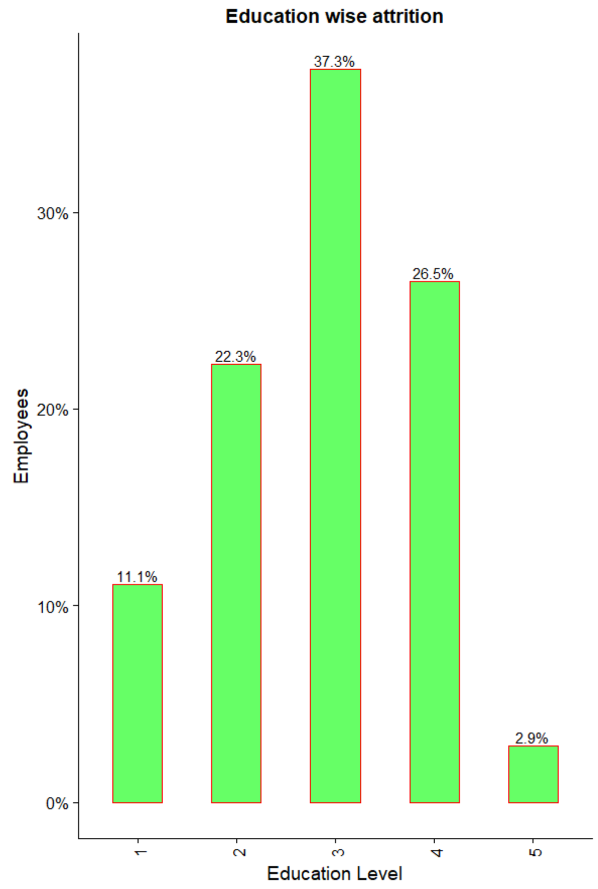
Job Role





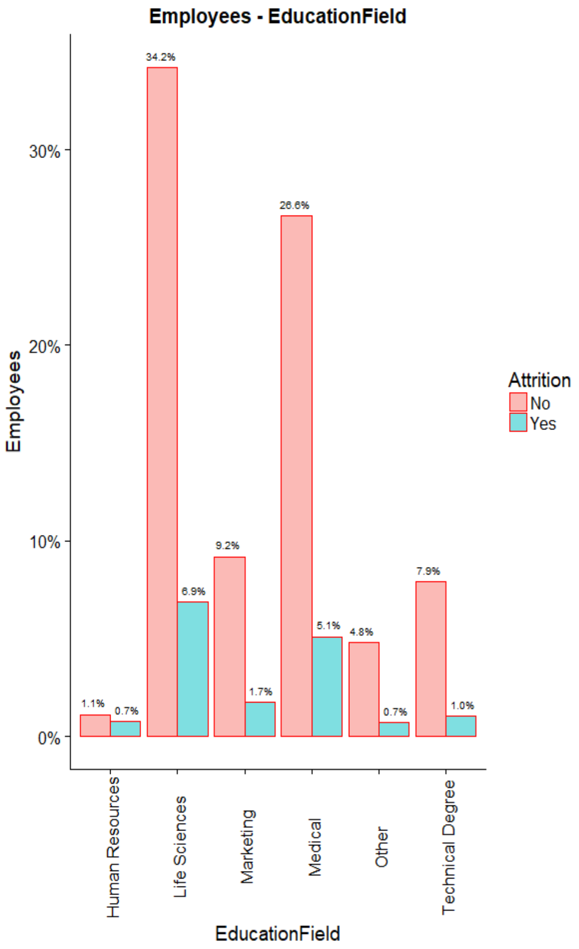
Education Level

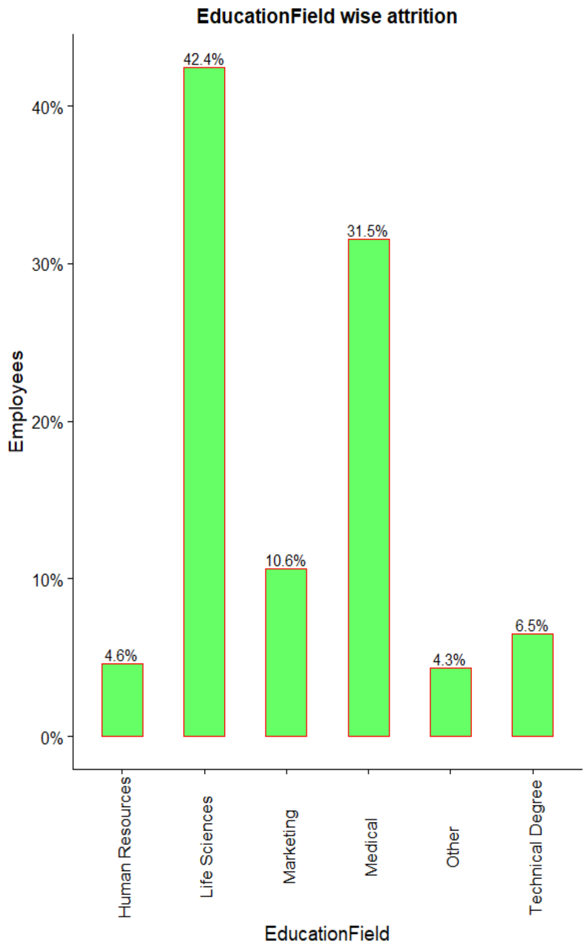


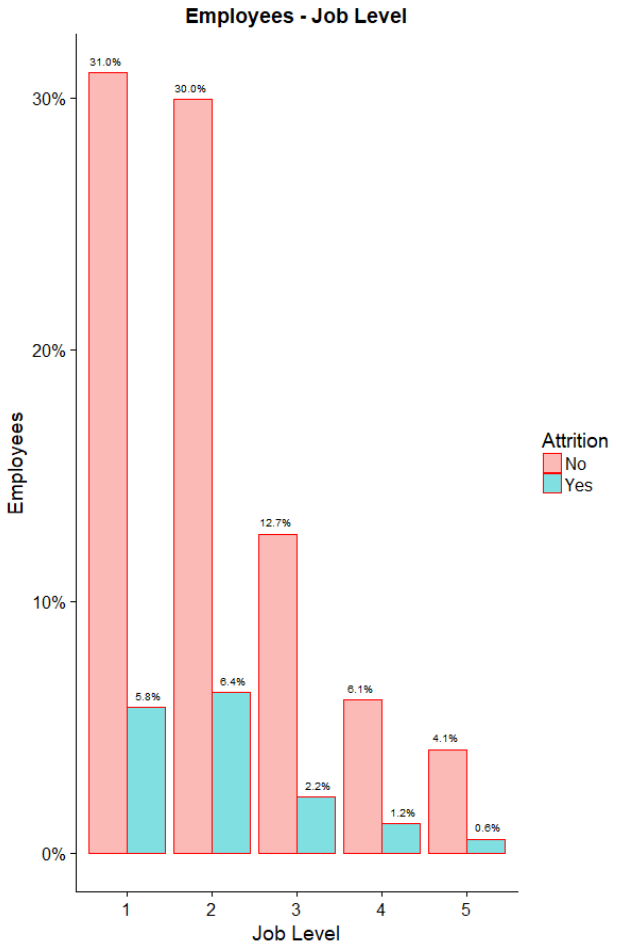


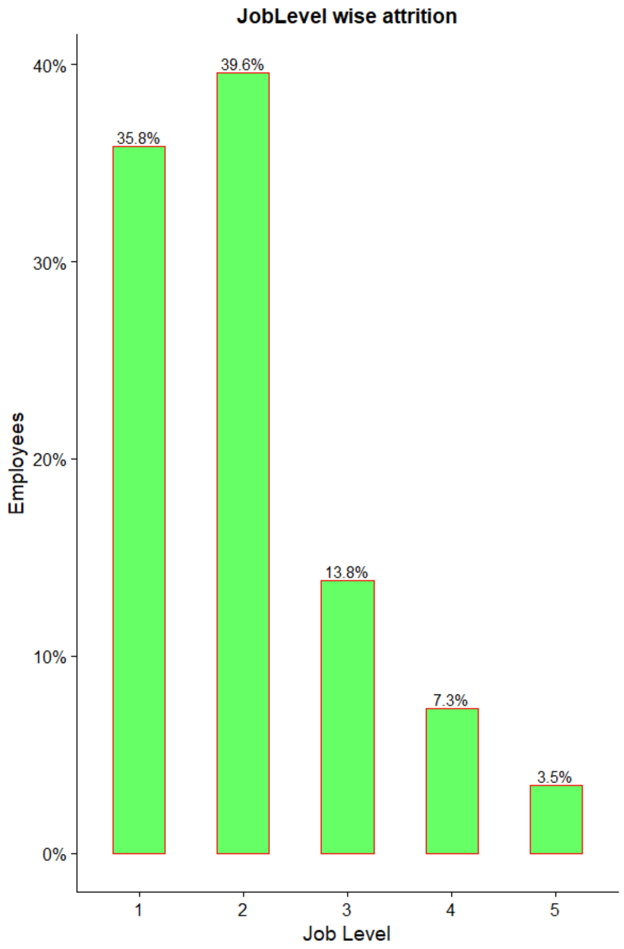
Education Field

Job Level

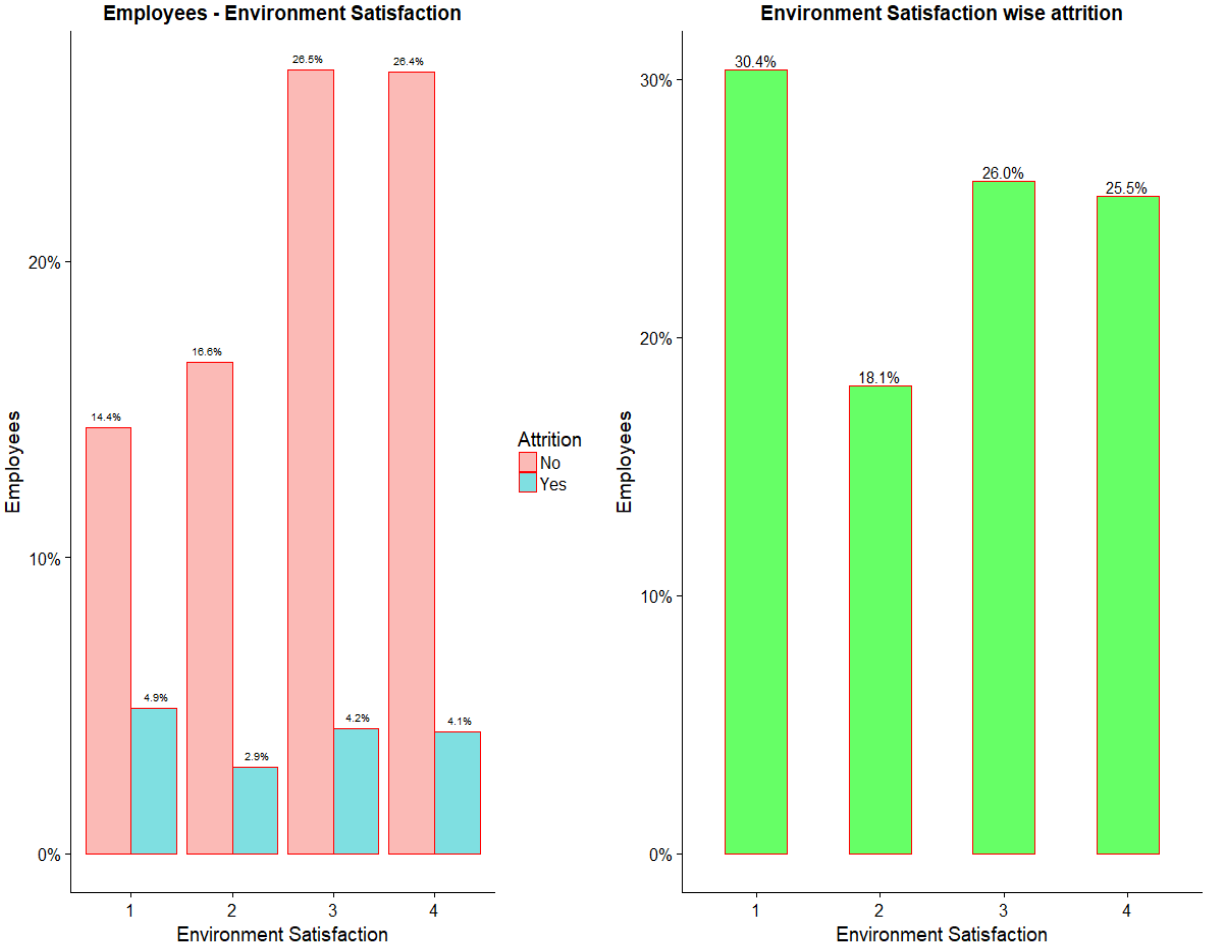




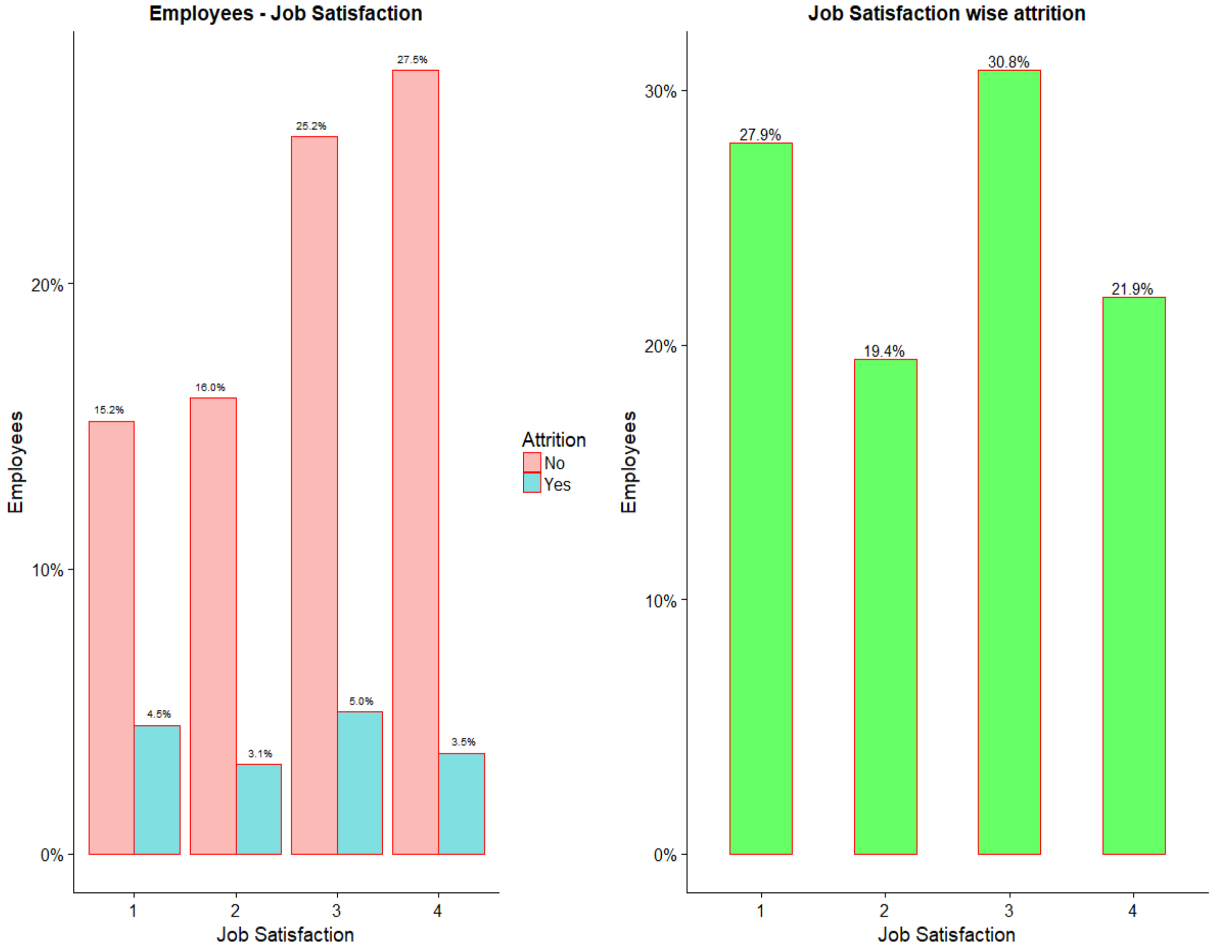




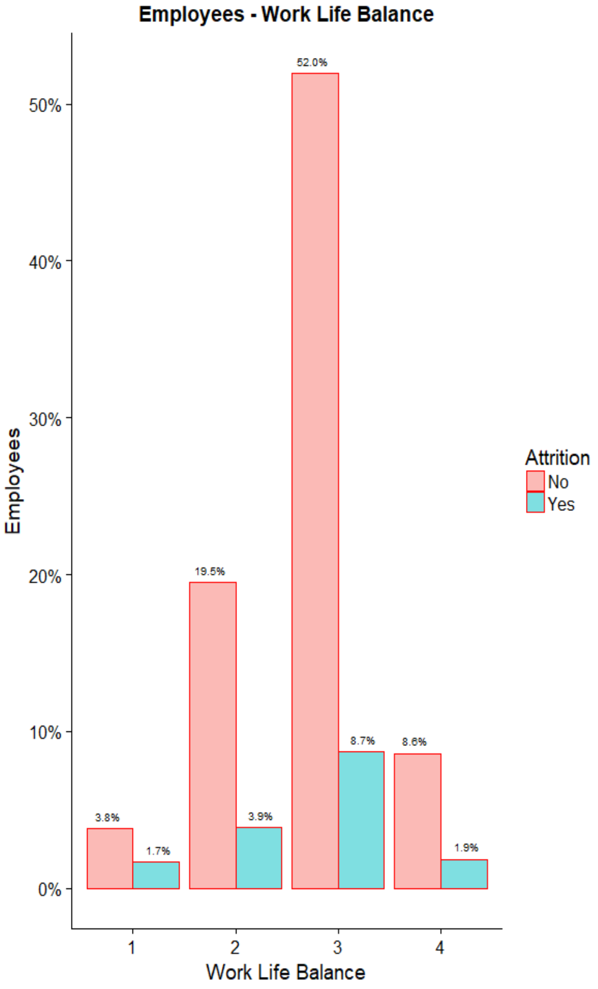
Environment Satisfaction

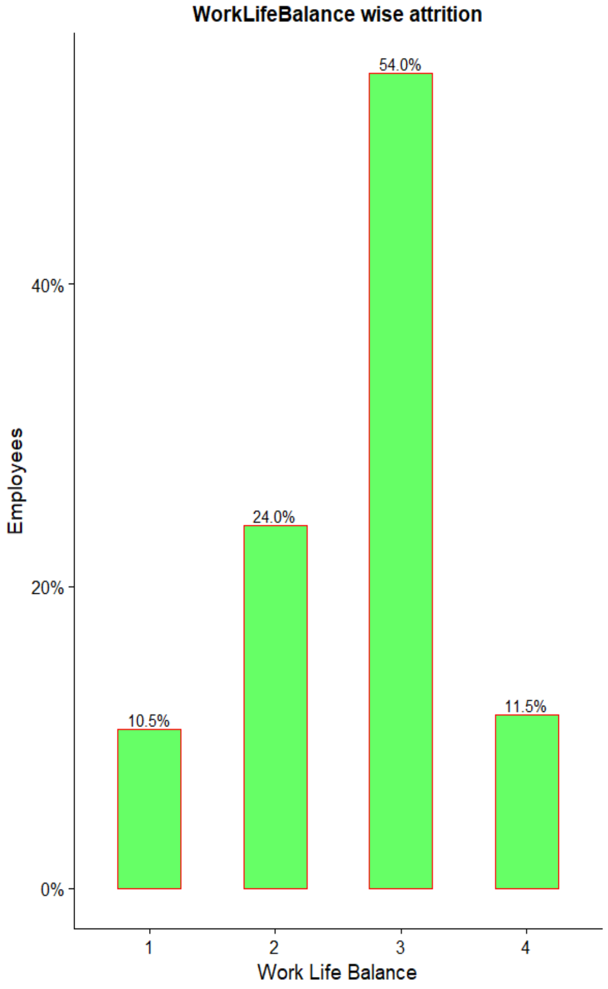


Job Satisfaction

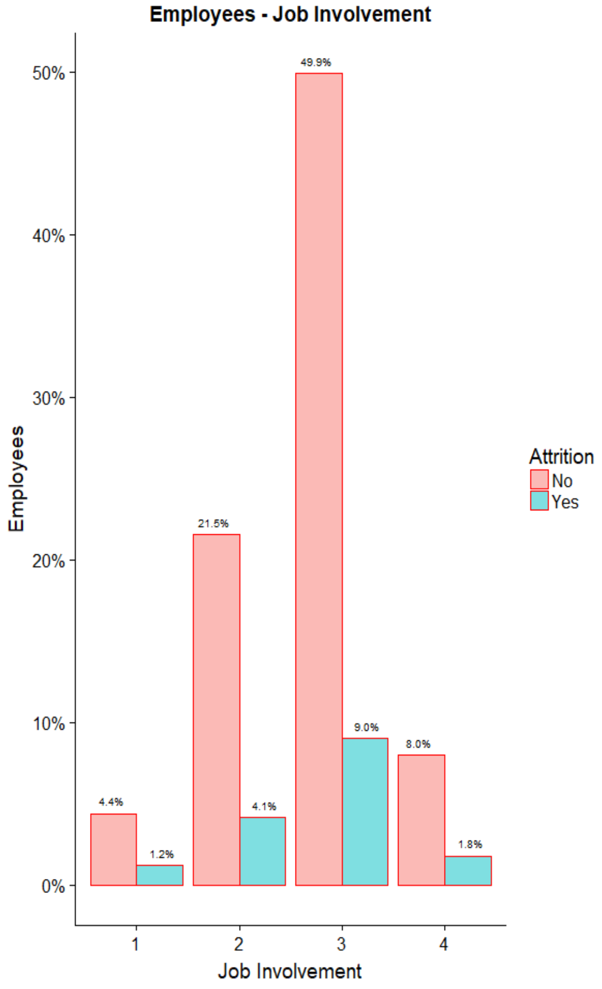


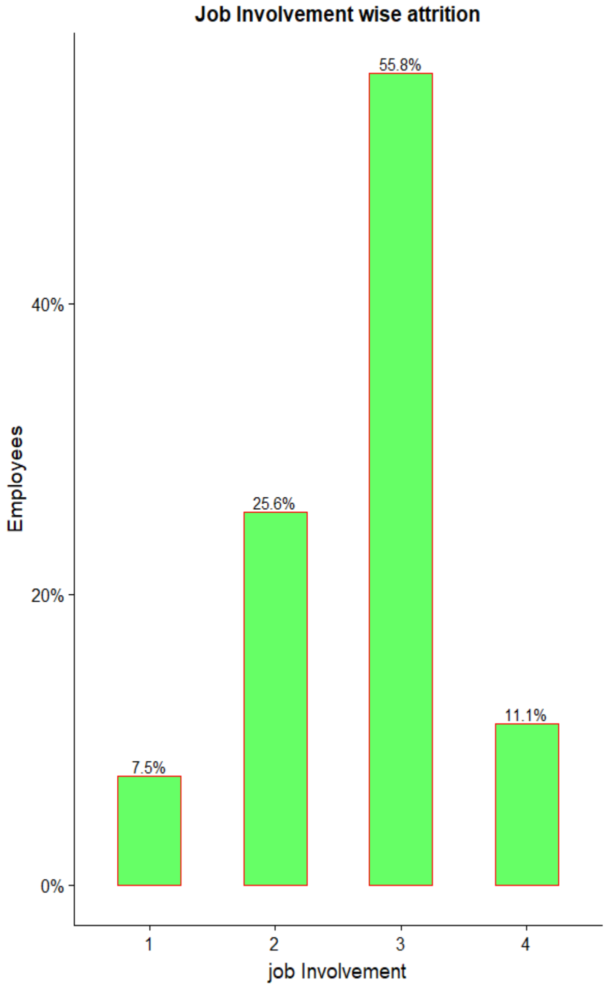
Work Life Balance



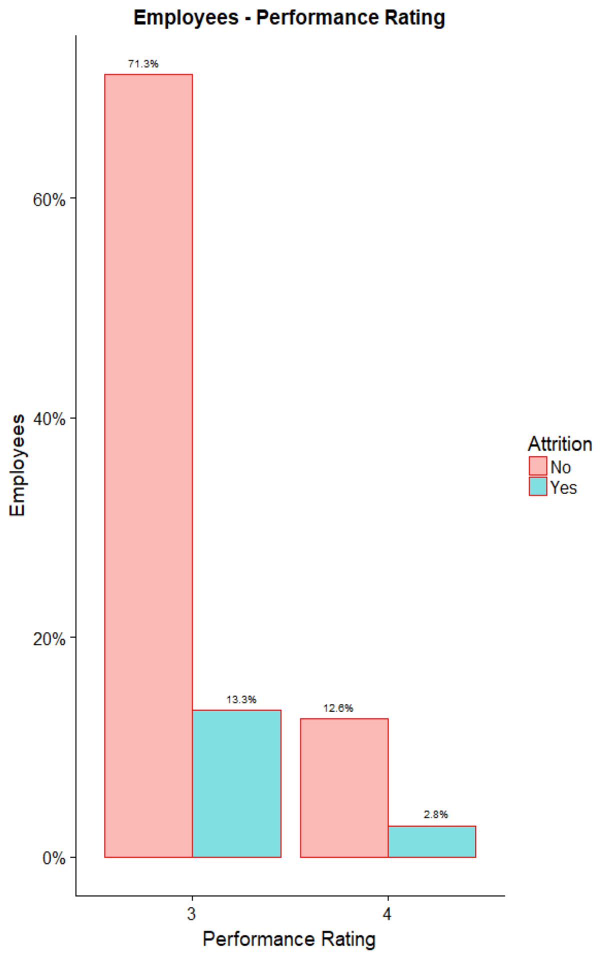


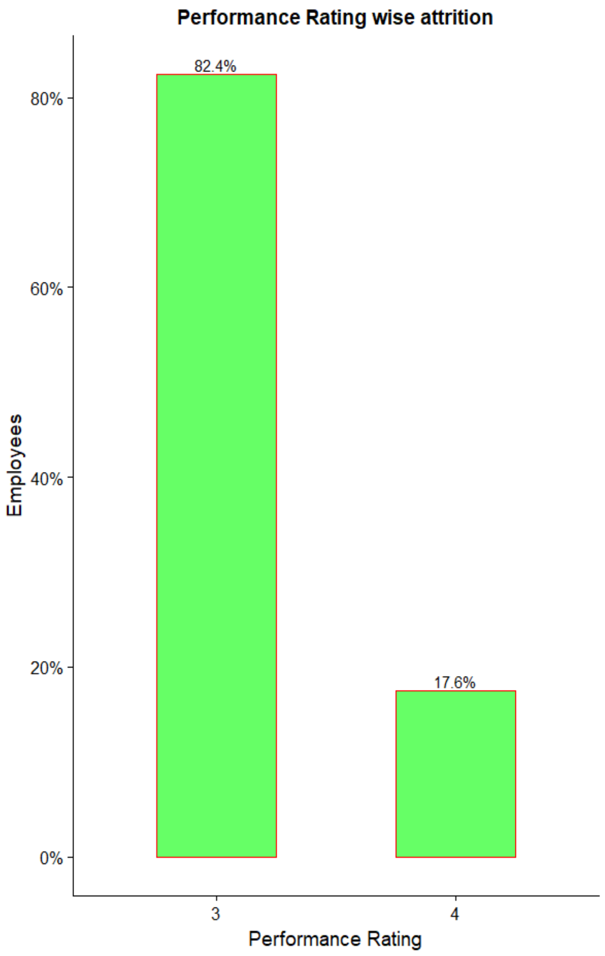
Job Involvement



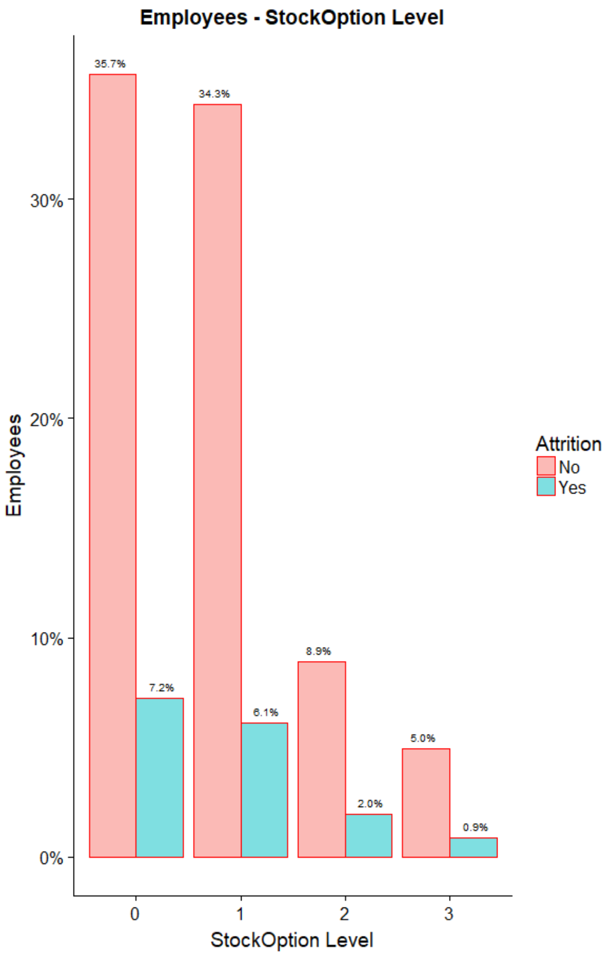


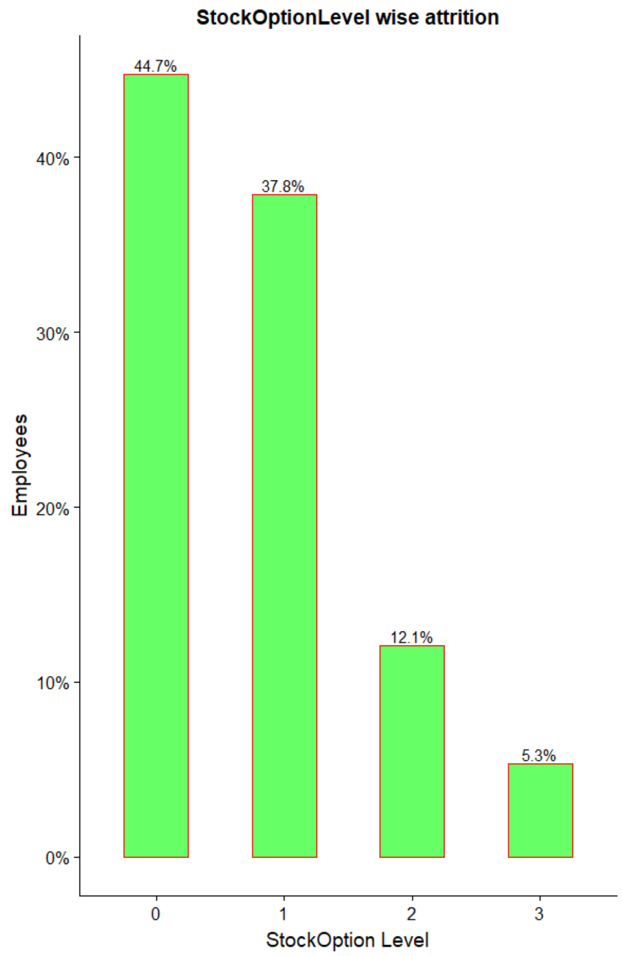
Performance Rating



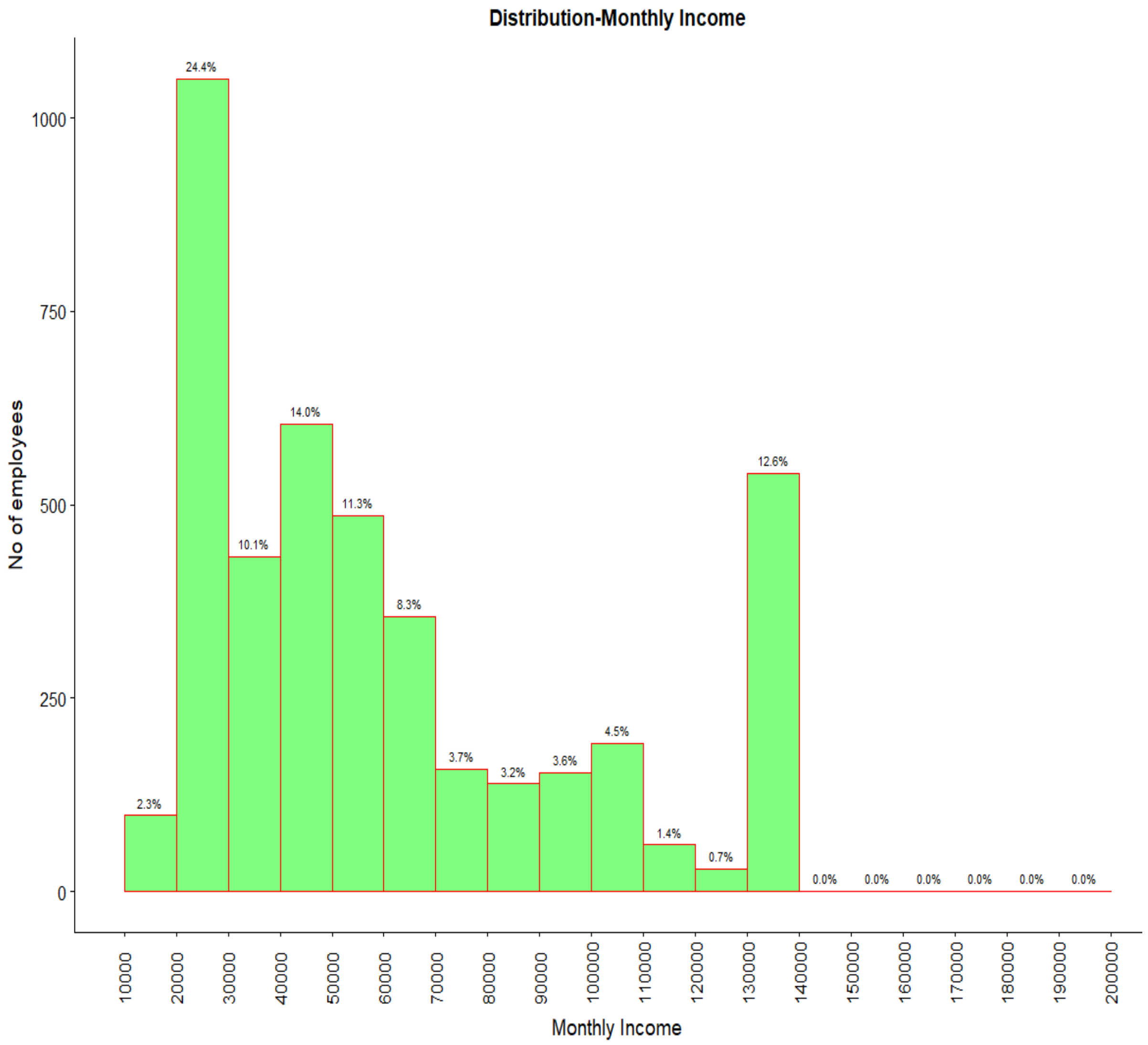


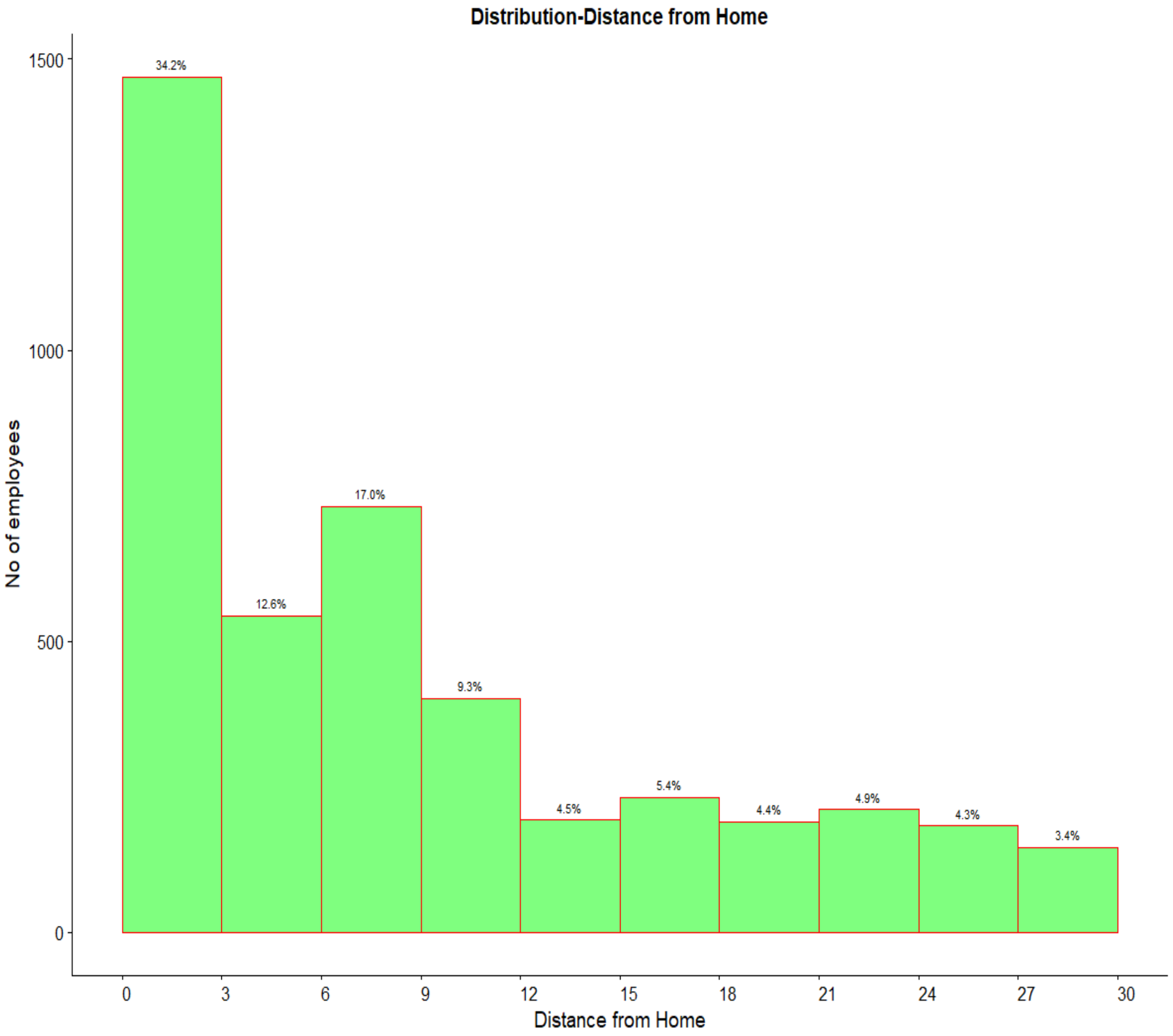
StockOptionLevel

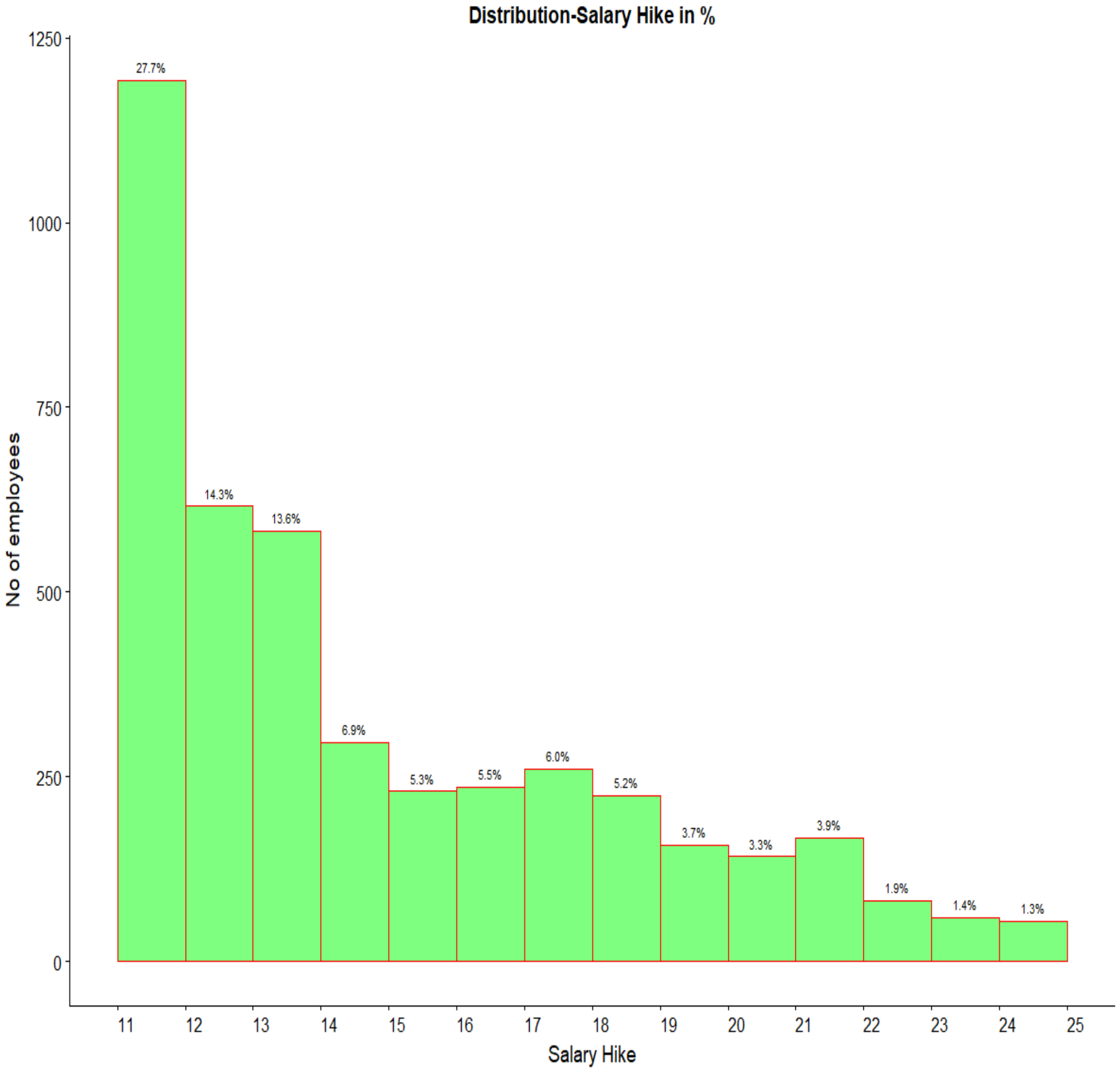




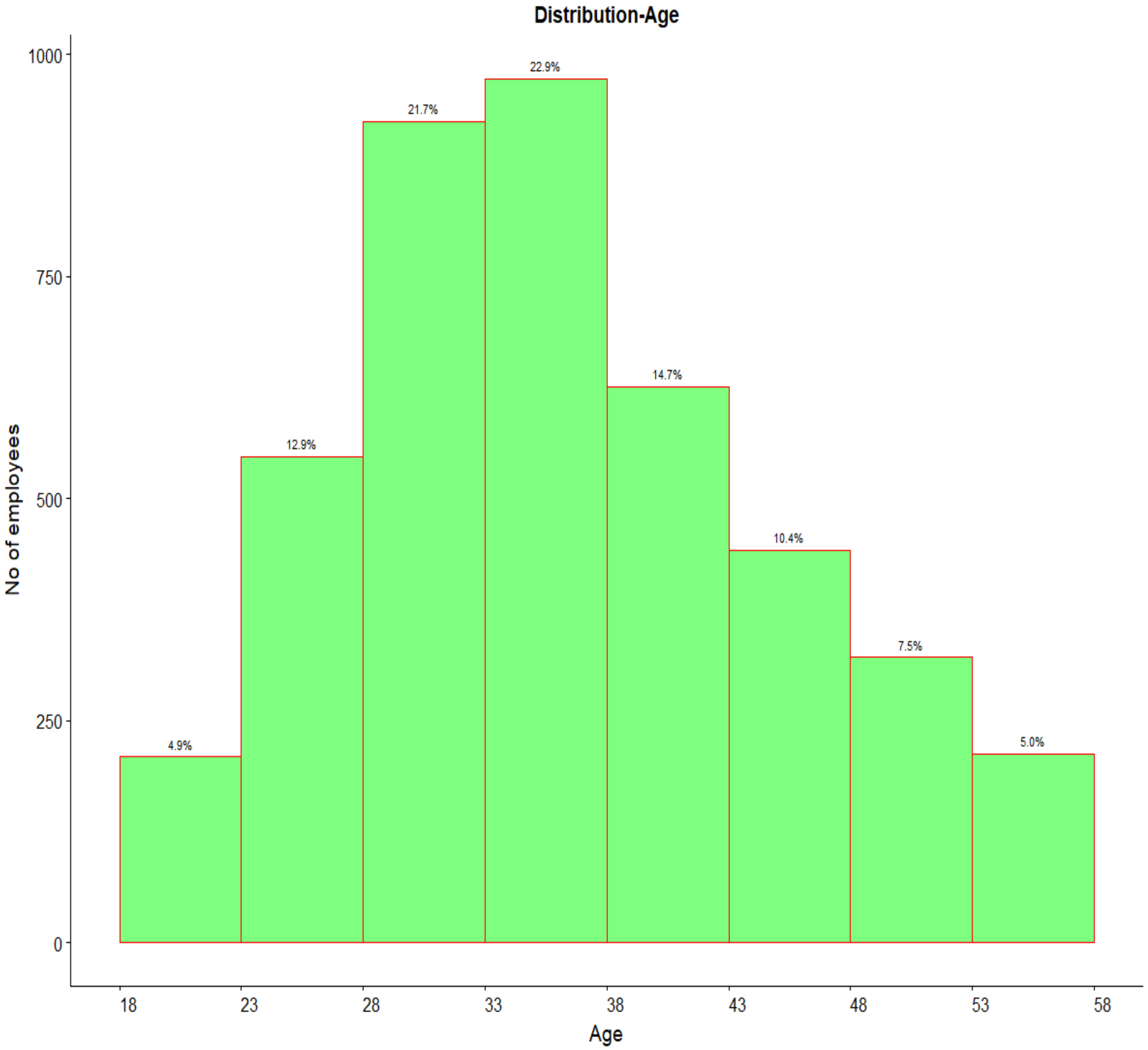
Univariate Analysis On Continuous Variables


Distribution-Monthly Income

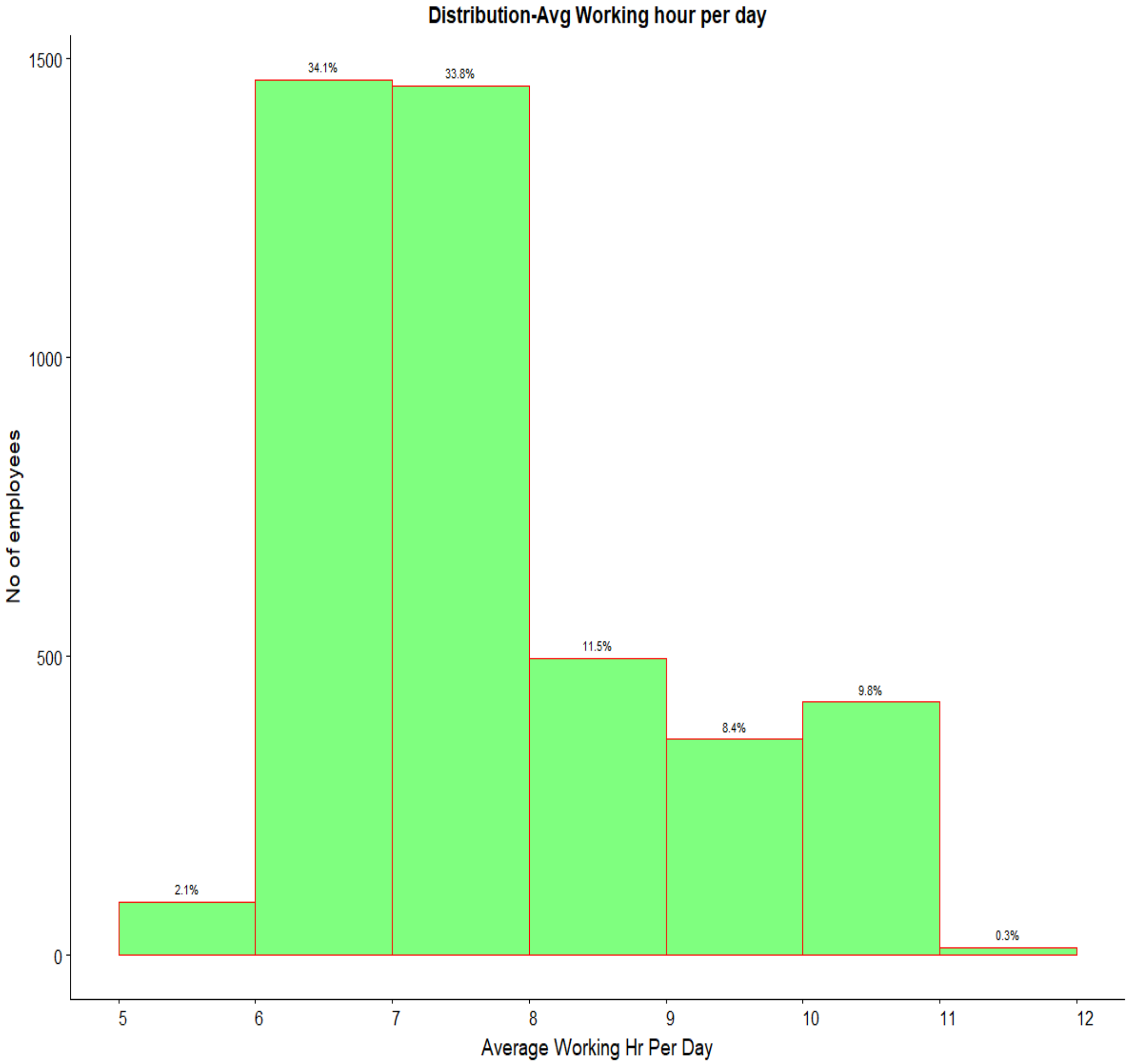

Distribution-Distance from Home

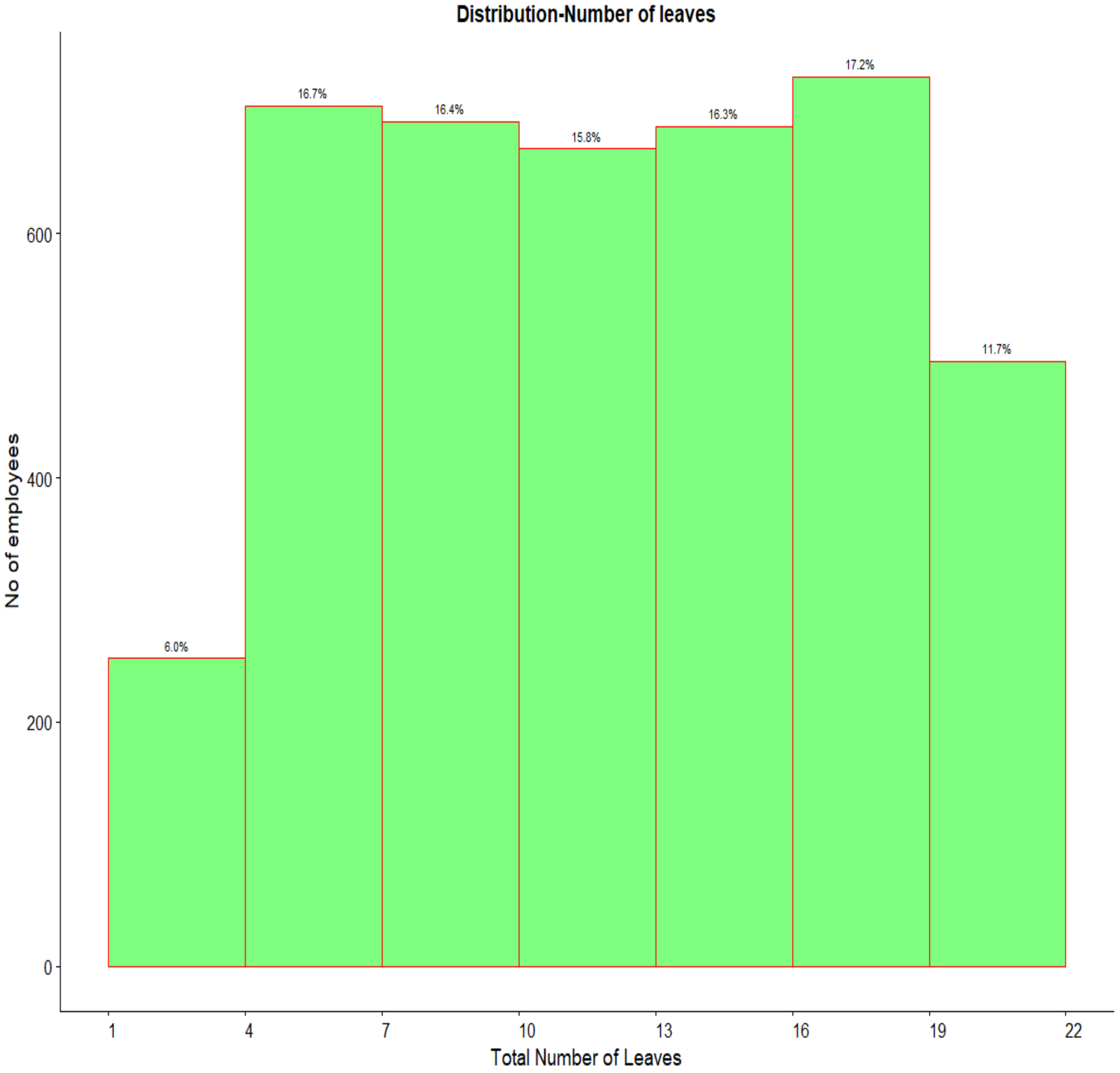


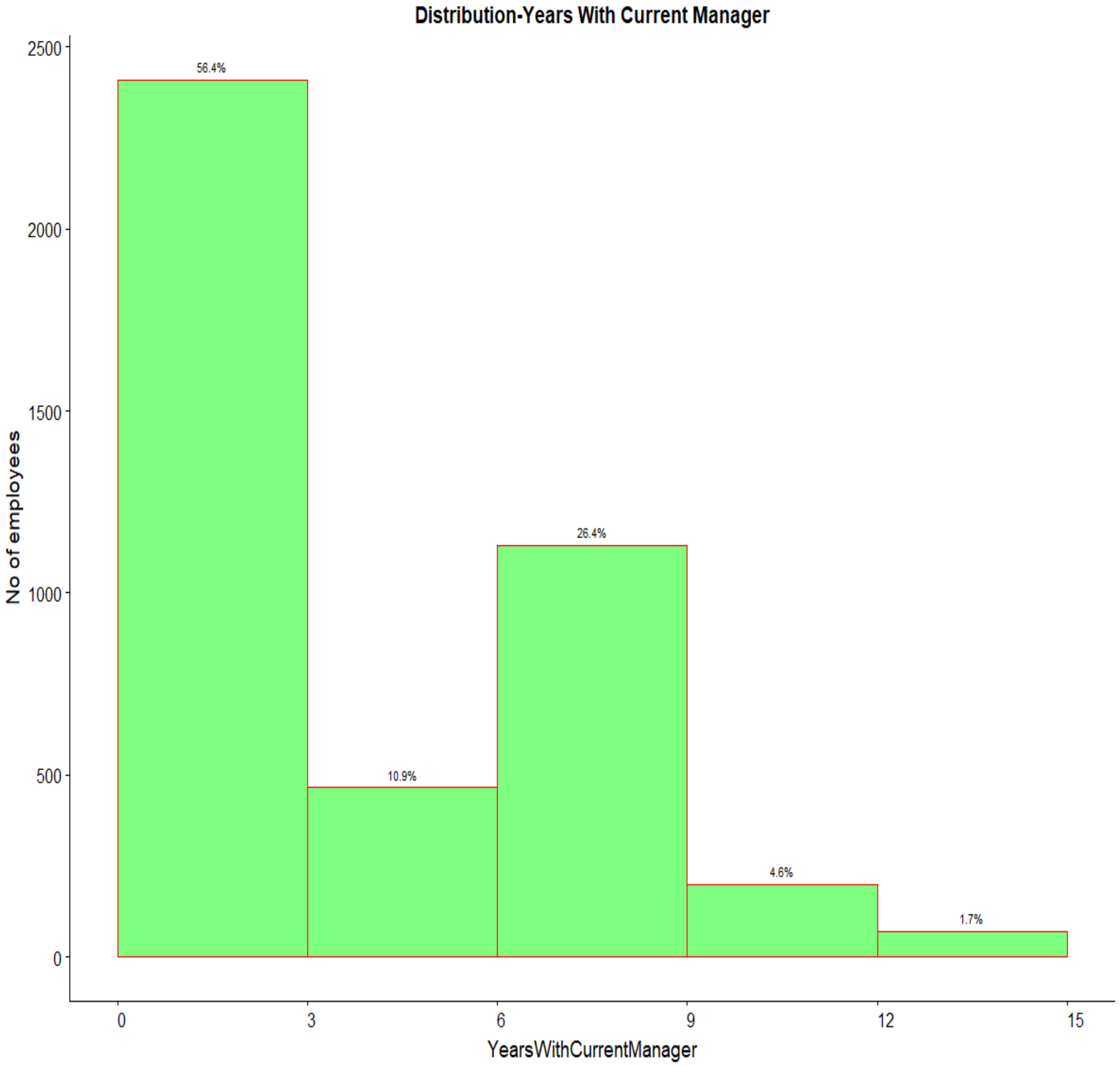
Distribution-Salary Hike in %



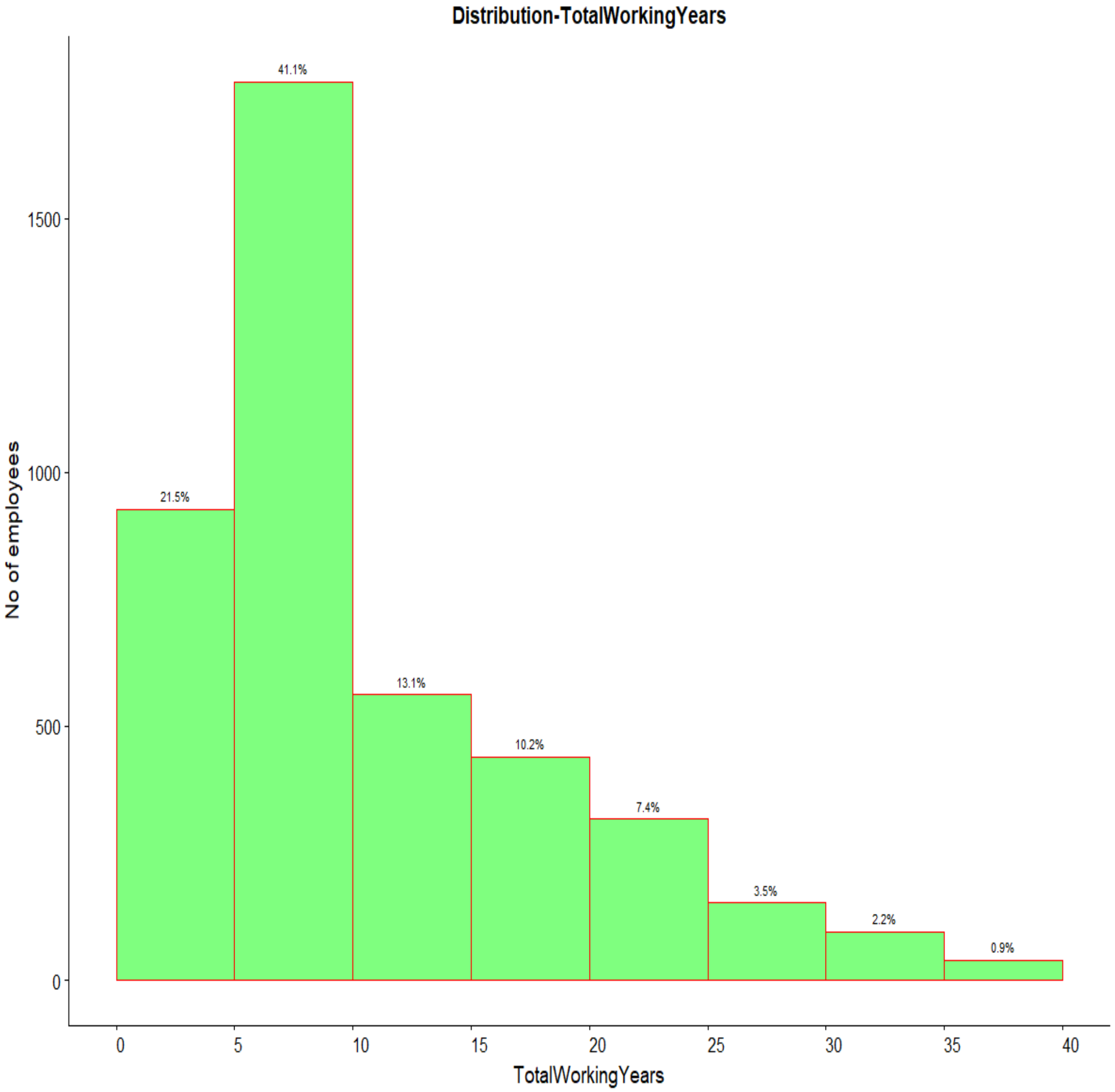
Distribution-Age


Distribution-Avg Working hour per day
No of employees
1500
1000
500
0
34.1%
33.8%
11.5%
8.4%
9.8%
2.1%
0.3%
5
6
7
8
9
10
11
12
Average Working Hr Per Day

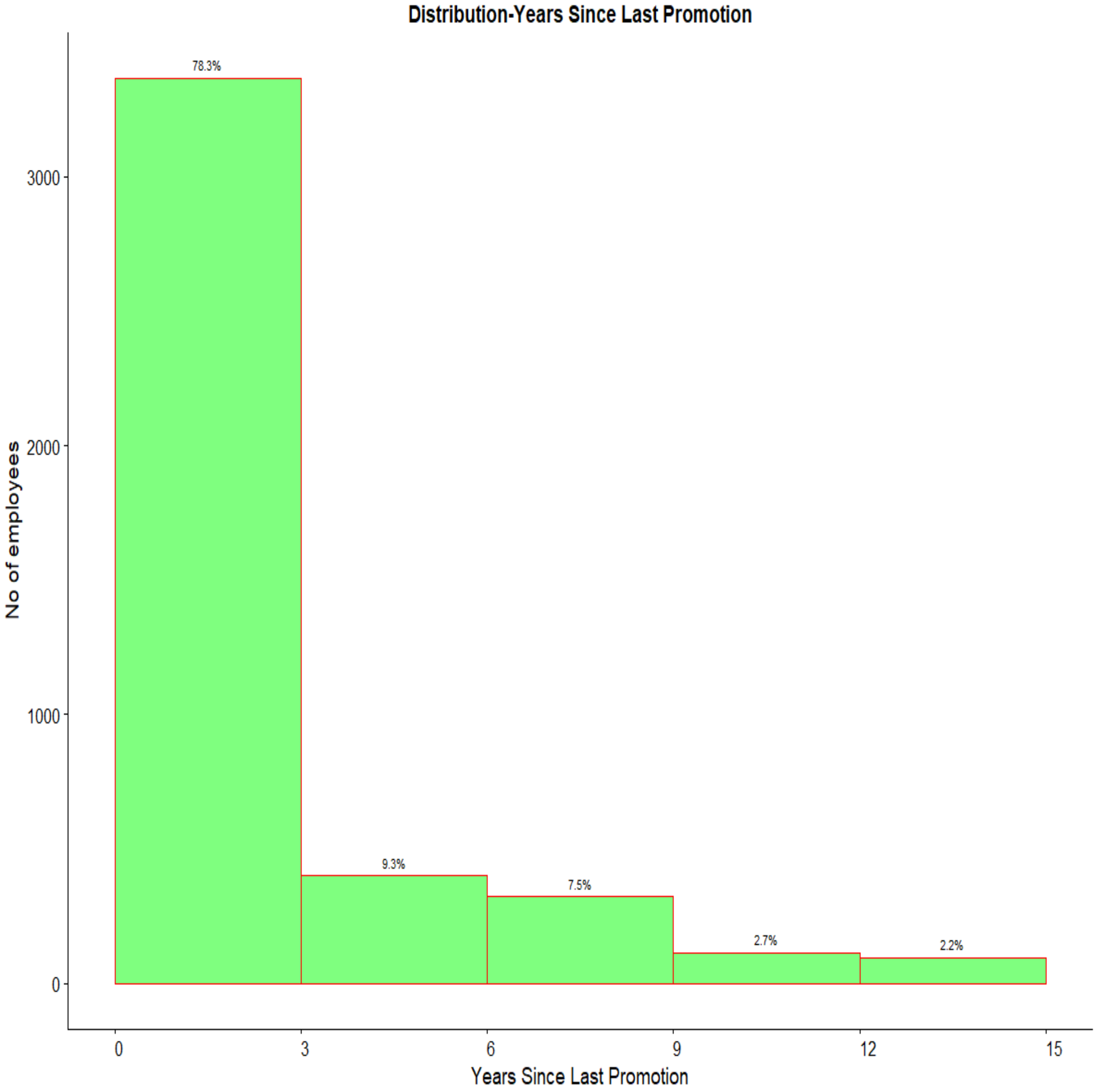

Distribution-Number of leaves
No of employees
600
400
200
0
6.0%
16.7%
16.4%
15.8%
16.3%
17.2%
11.7%
1
4
7
10
13
16
19
22
Total Number of Leaves



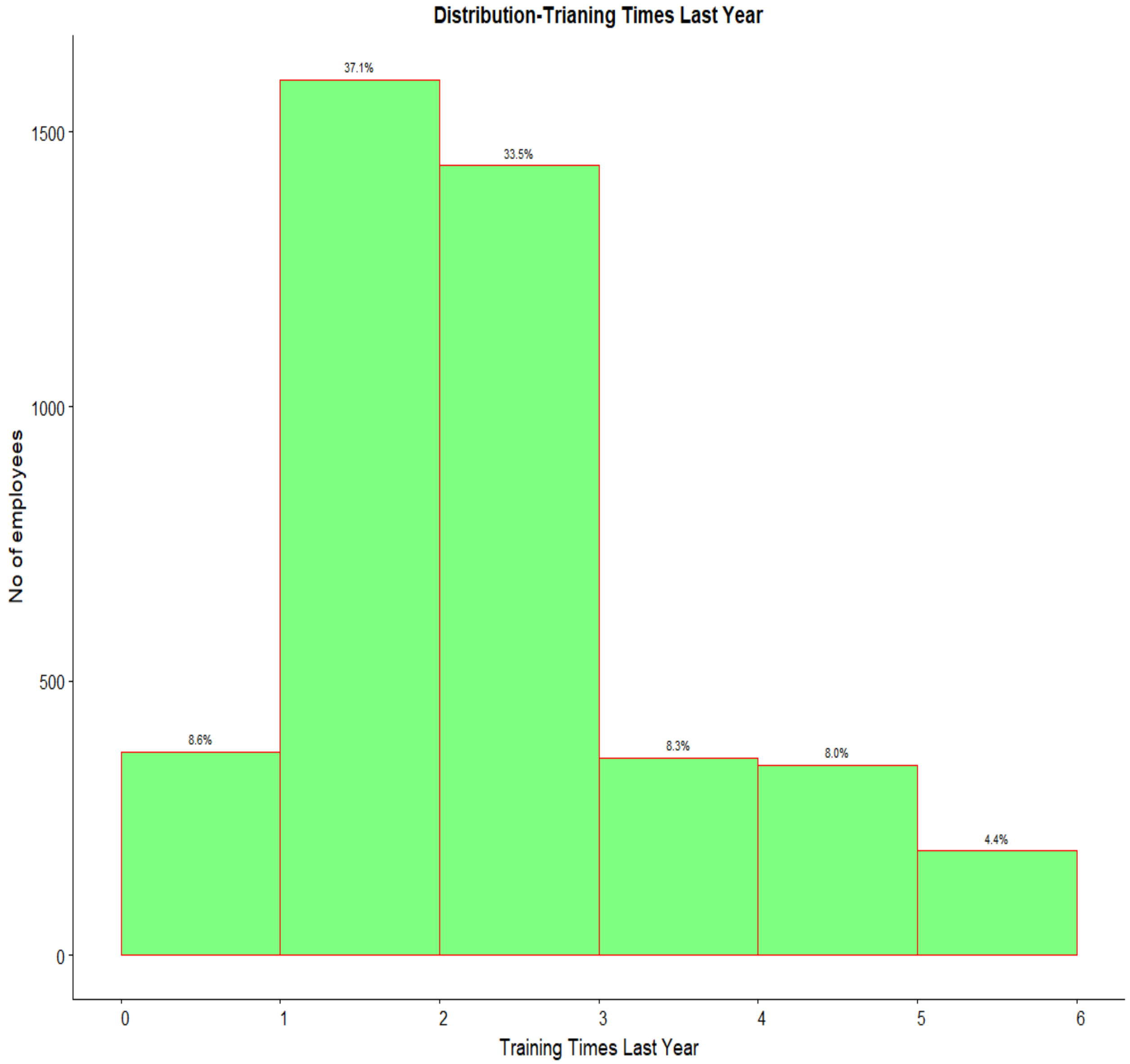
Distribution-Years With Current Manager



Distribution-TotalWorkingYears

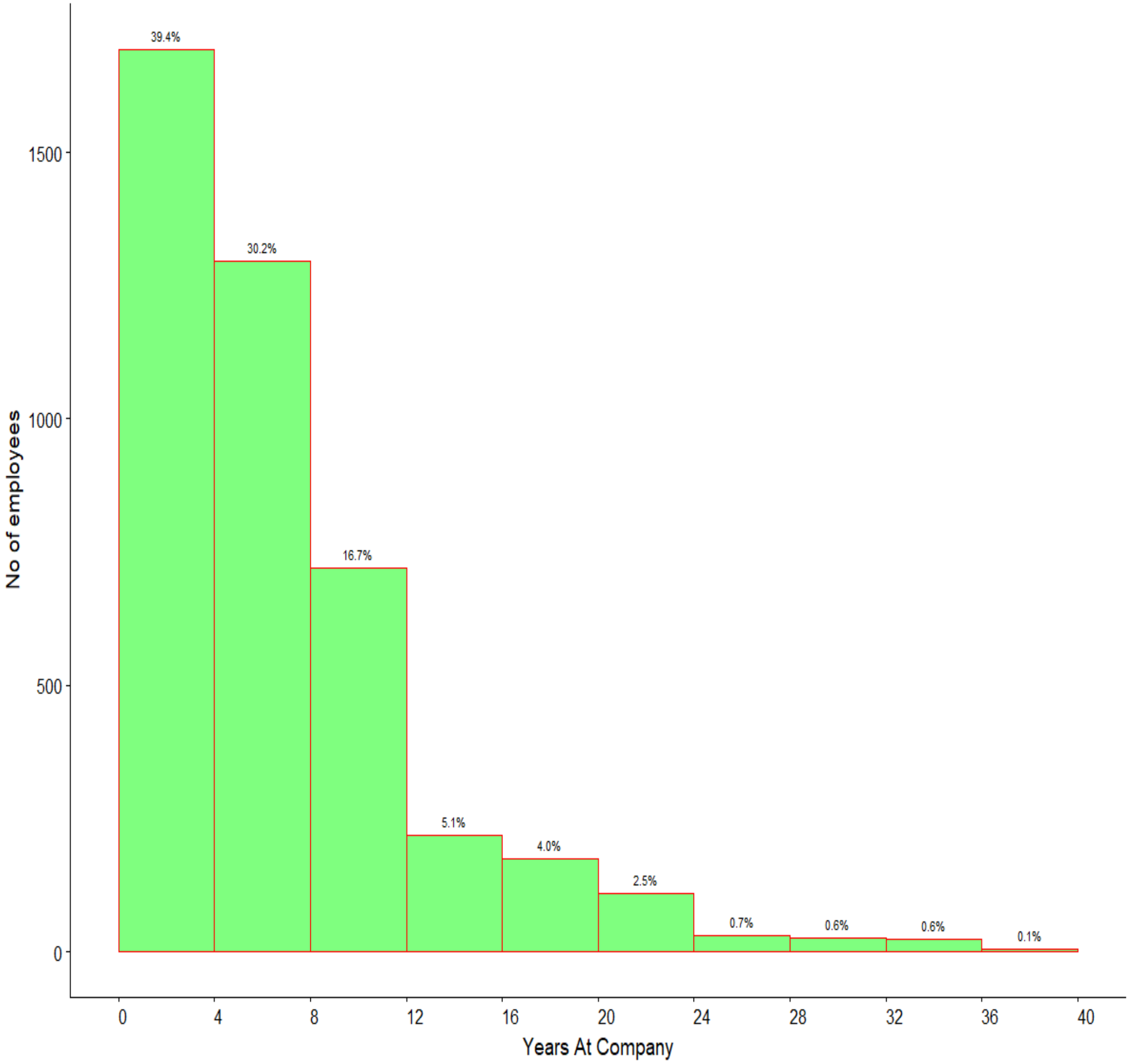


Distribution-Years Since Last Promotion

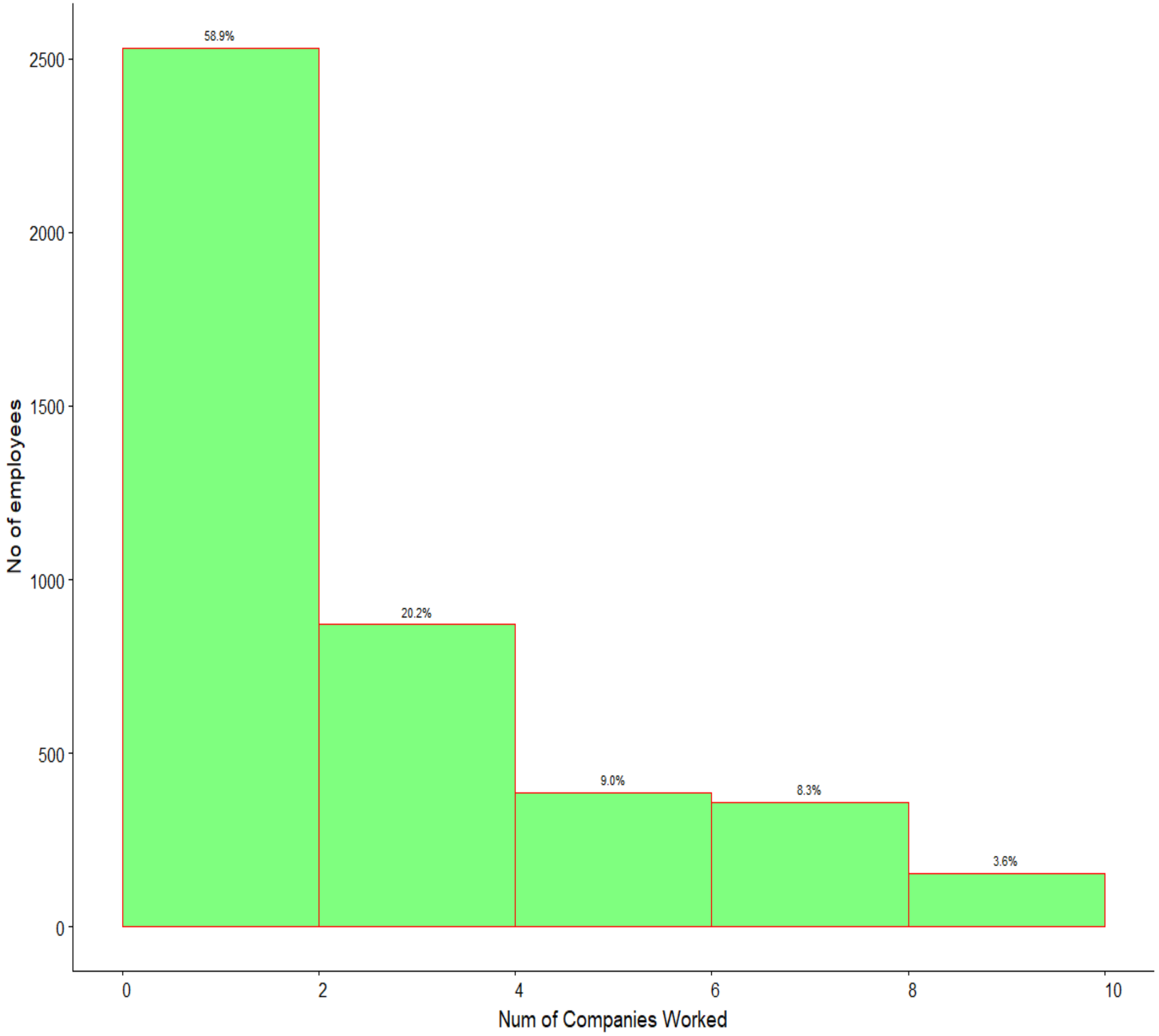


Distribution-Trianing Times Last Year

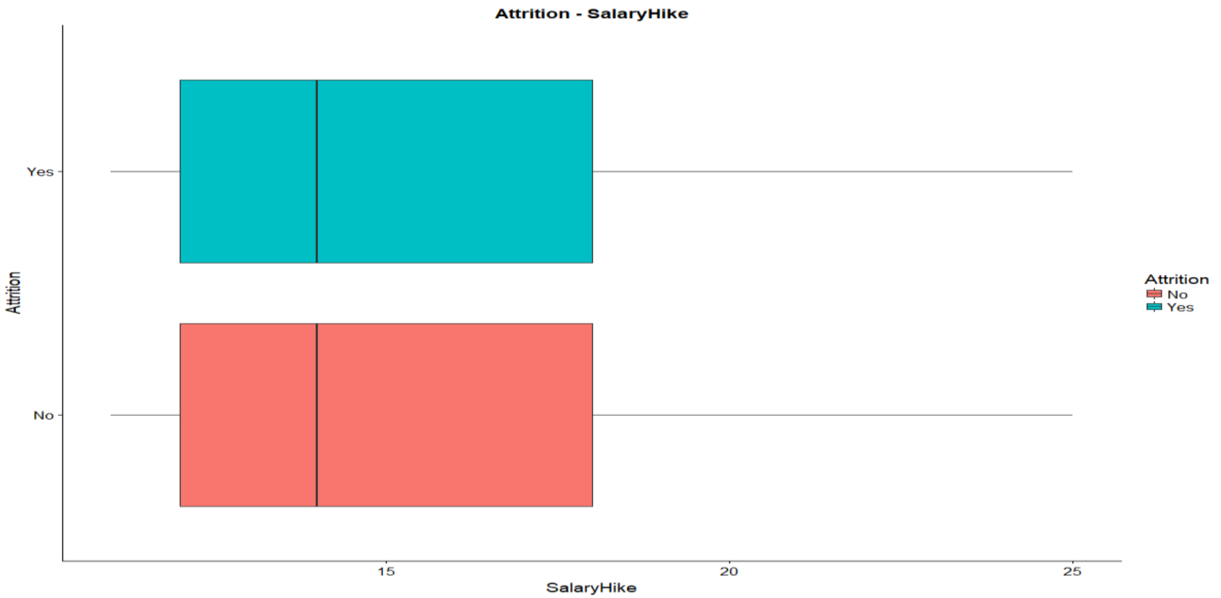
Distribution-Years At company

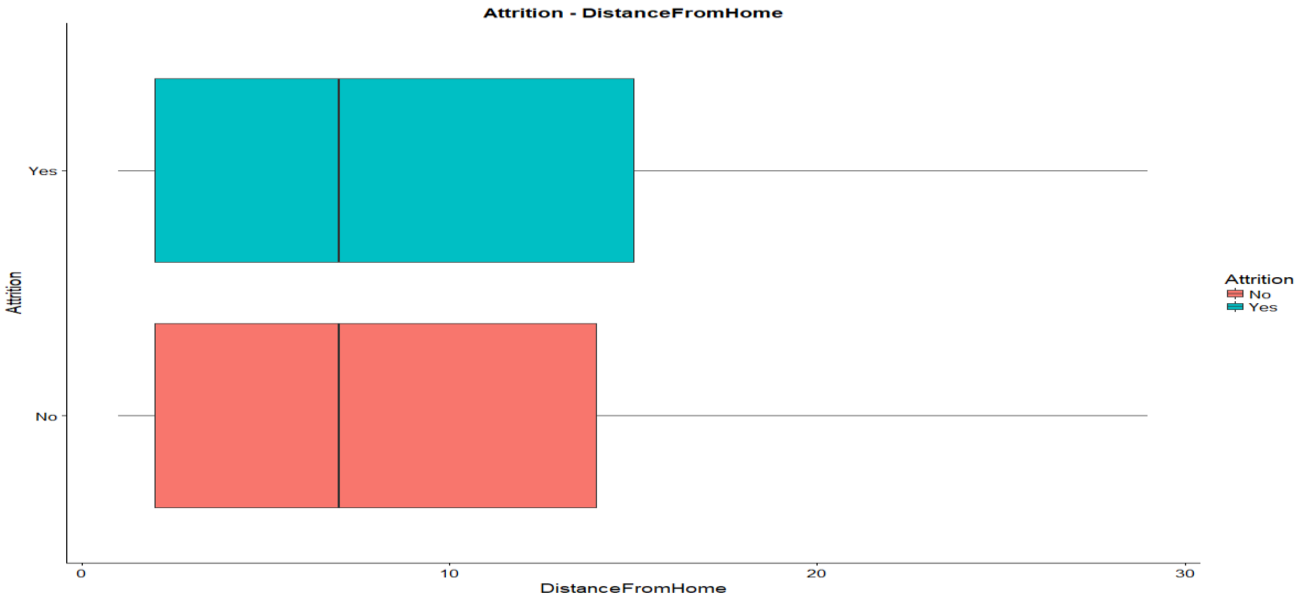


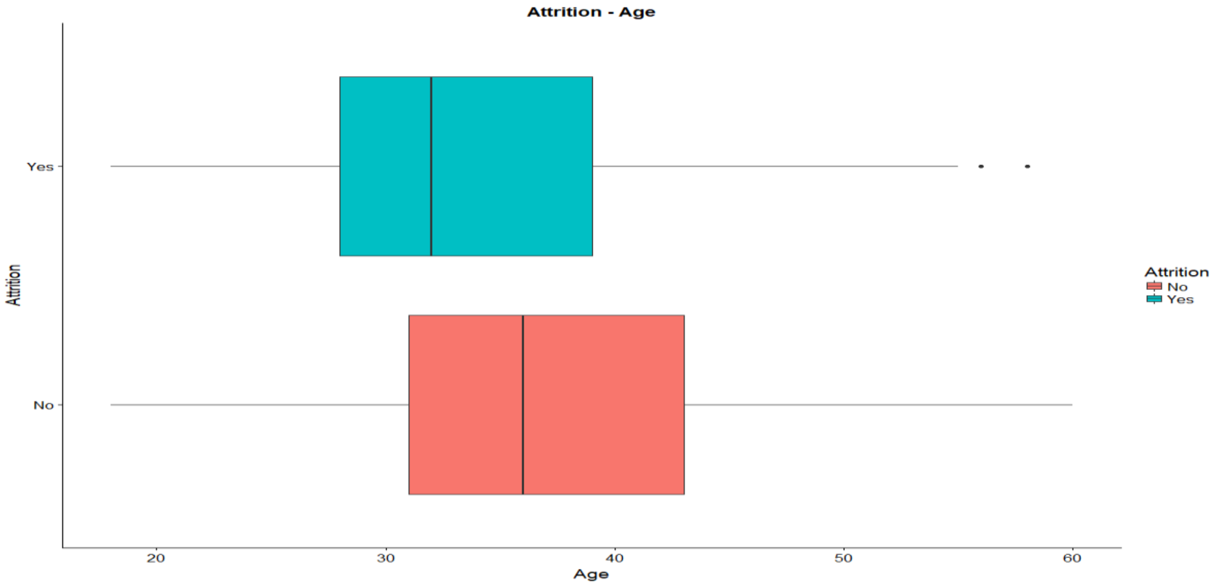
Distribution-Num Of Companies Worked

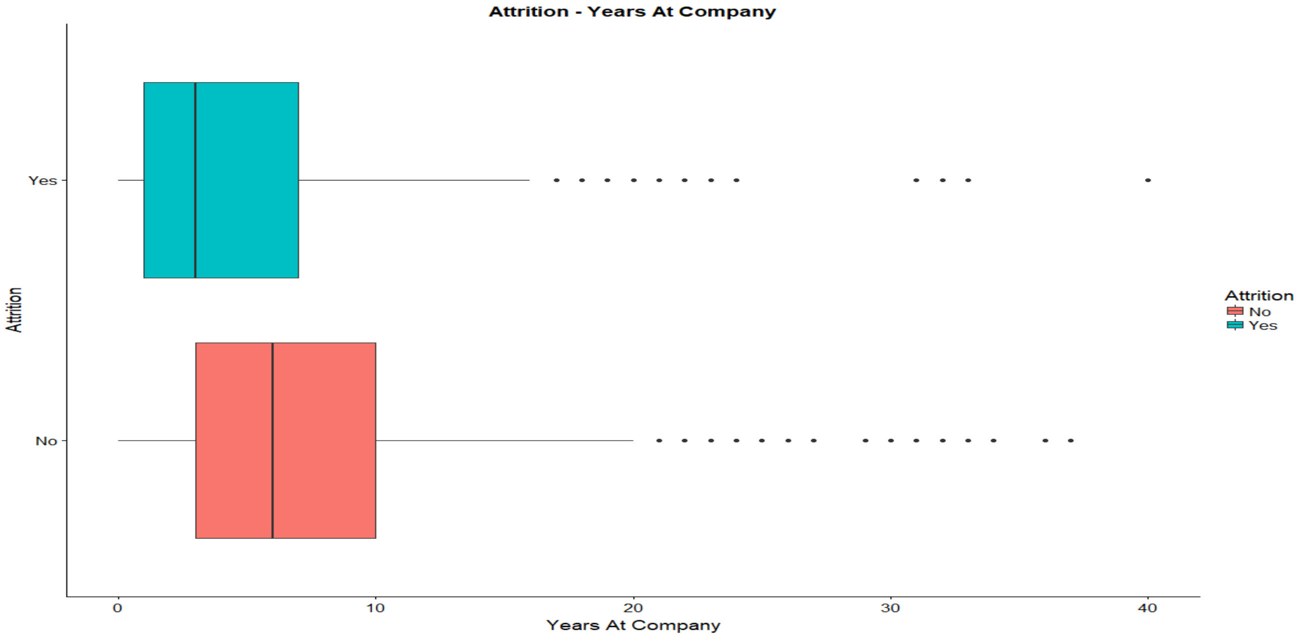


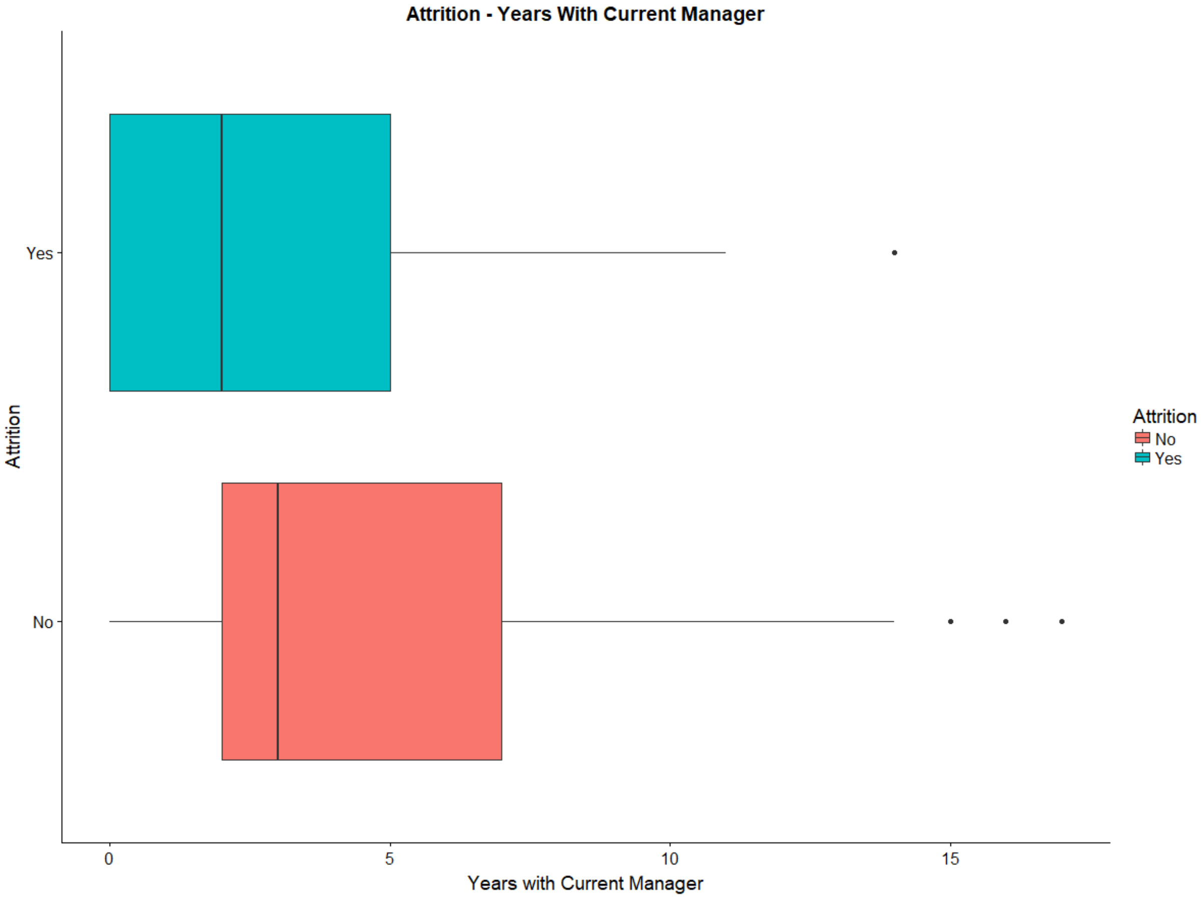
Bivariate Analysis I (Continuous Vs Categorical Variables)

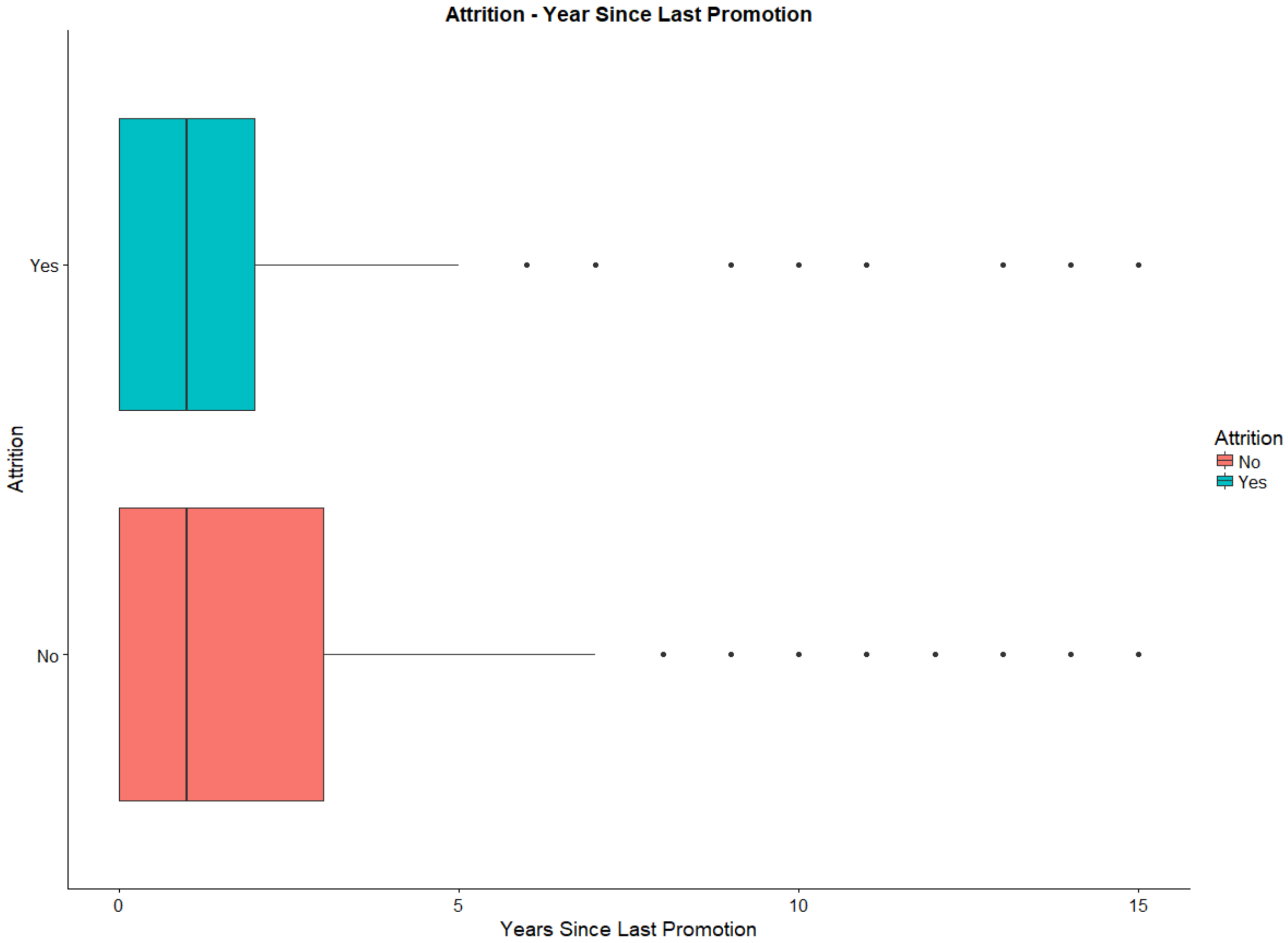










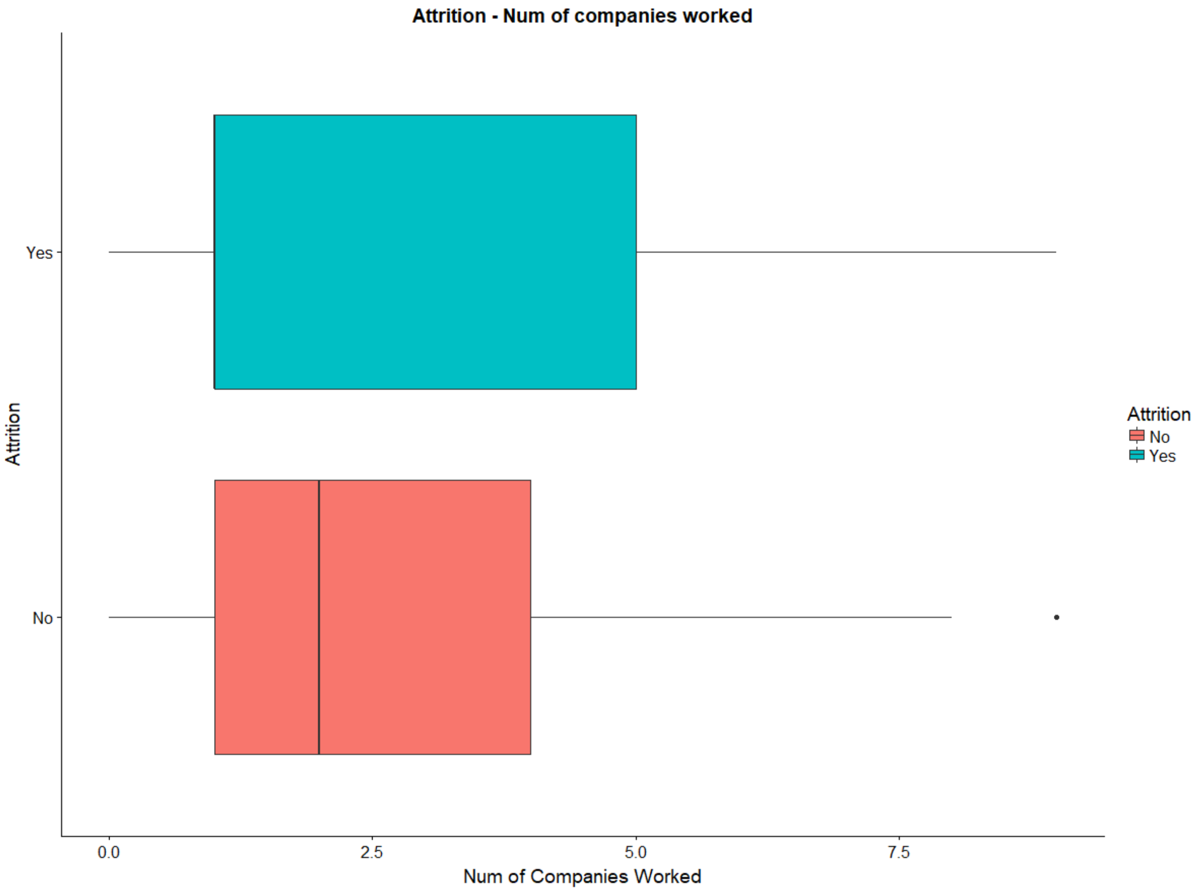


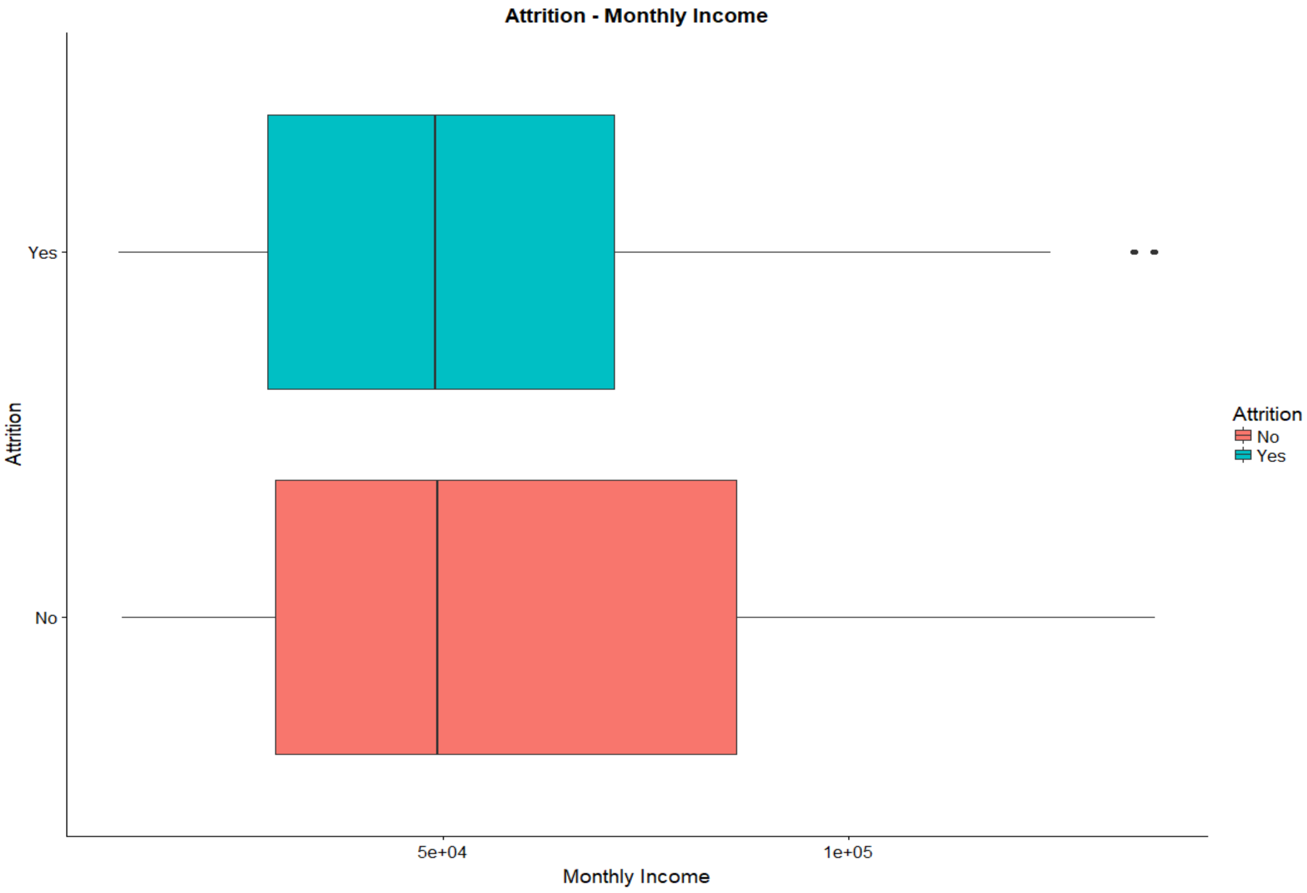
INFERENCE:

Employees who have attrited have a lesser range of years with current manager.

INFERENCE:

No major inference can be drawn.



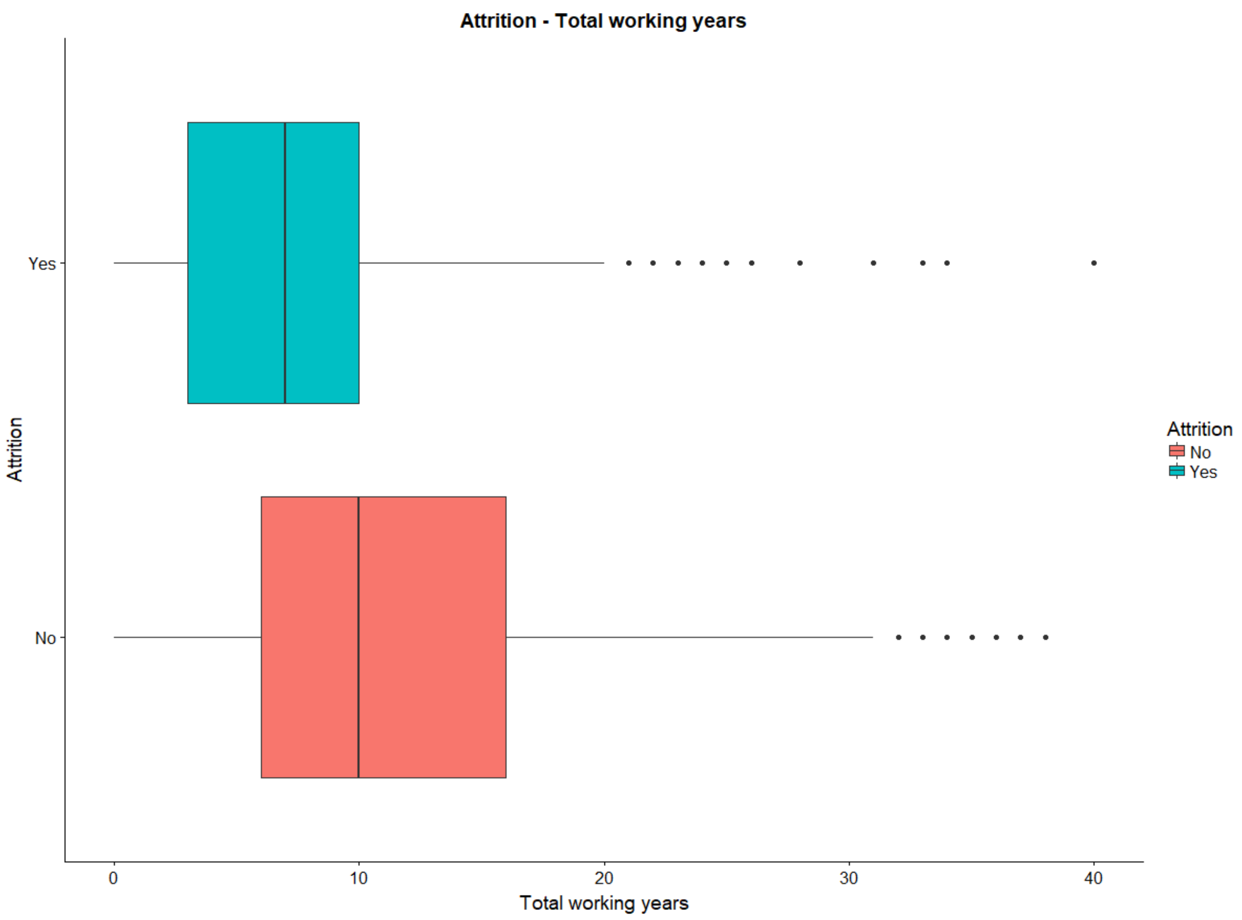


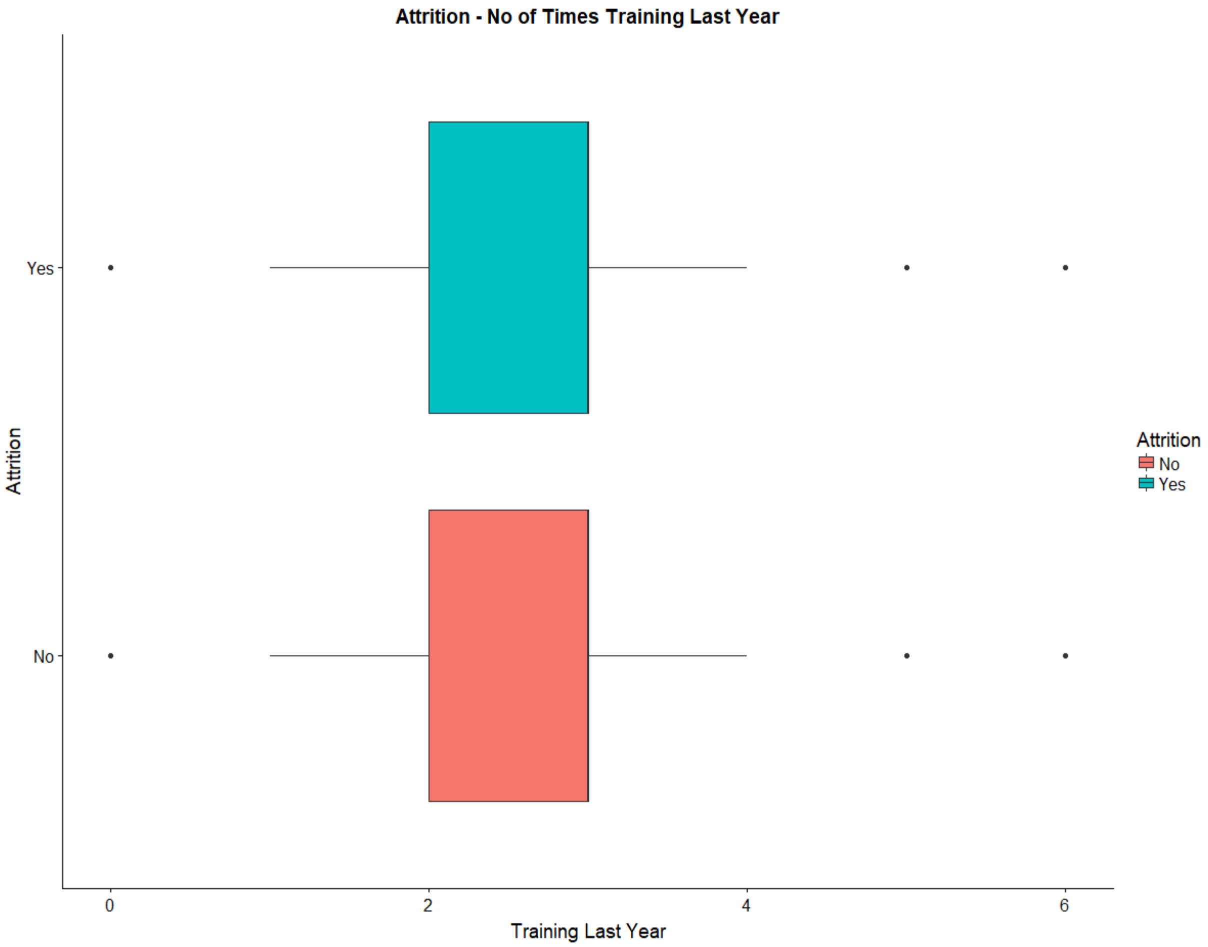
INFERENCE:

Employees who have attrited have worked for more number of companies.(range is more as we can see in box plot)

INFERENCE:

Employees who have attrited have "monthly income range" lesser than the monthly income range of employees who are still working with the company



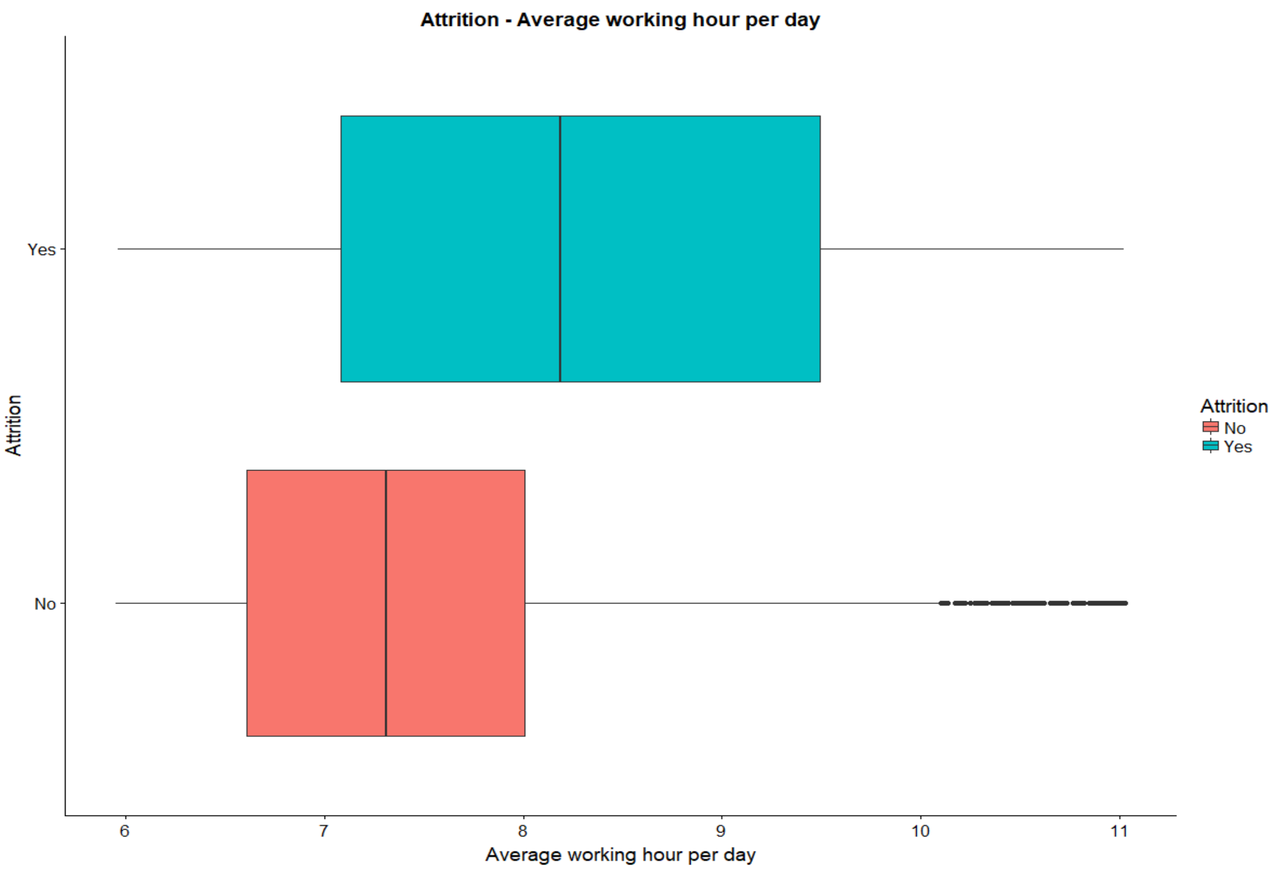


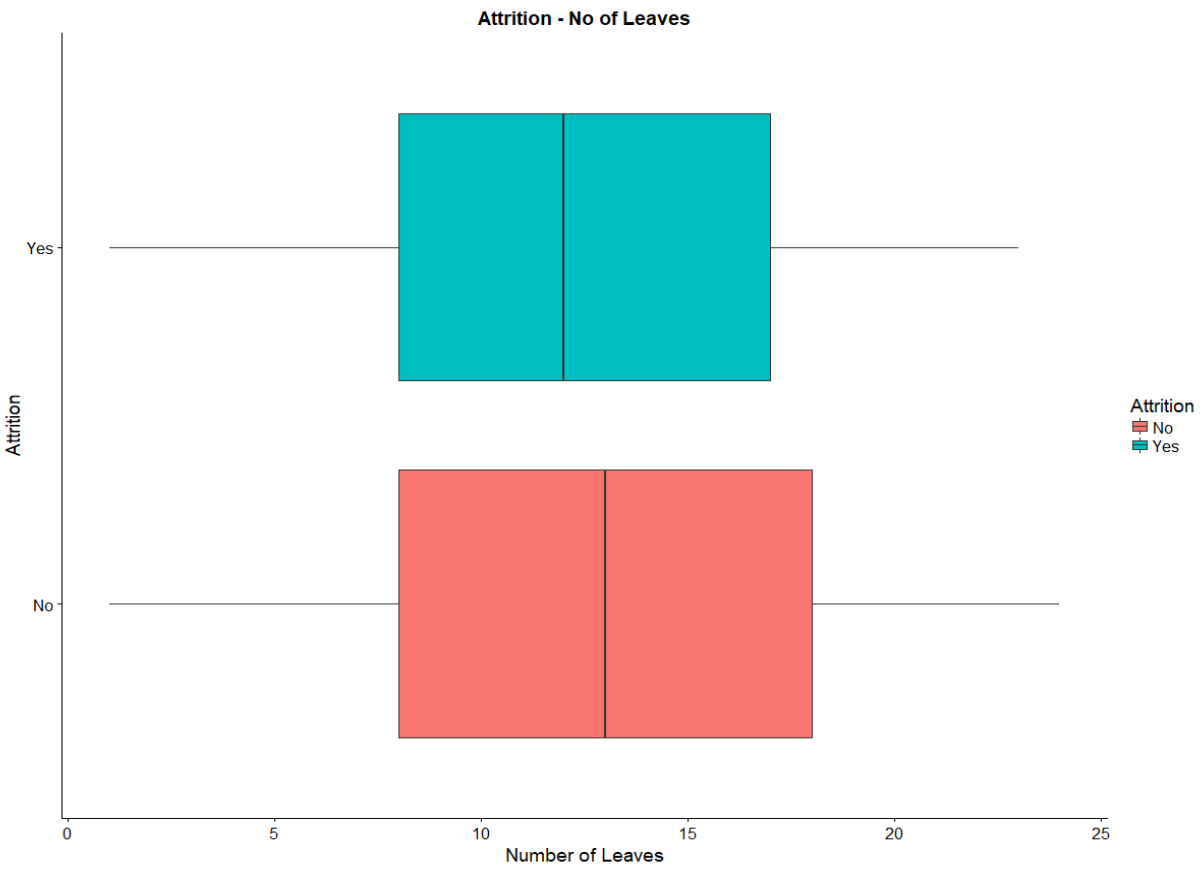
INFERENCE:

People who have attrited have lesser “TotalWorkingYears” than the people who are working with the company

INFERENCE:

No major inference can be drawn.





INFERENCE:

Employees who have attrited have an higher average of working hour per day. The range of their working hours is from 7 -9.5 which in comparison is higher than the employees who are still with the company.

INFERENCE:

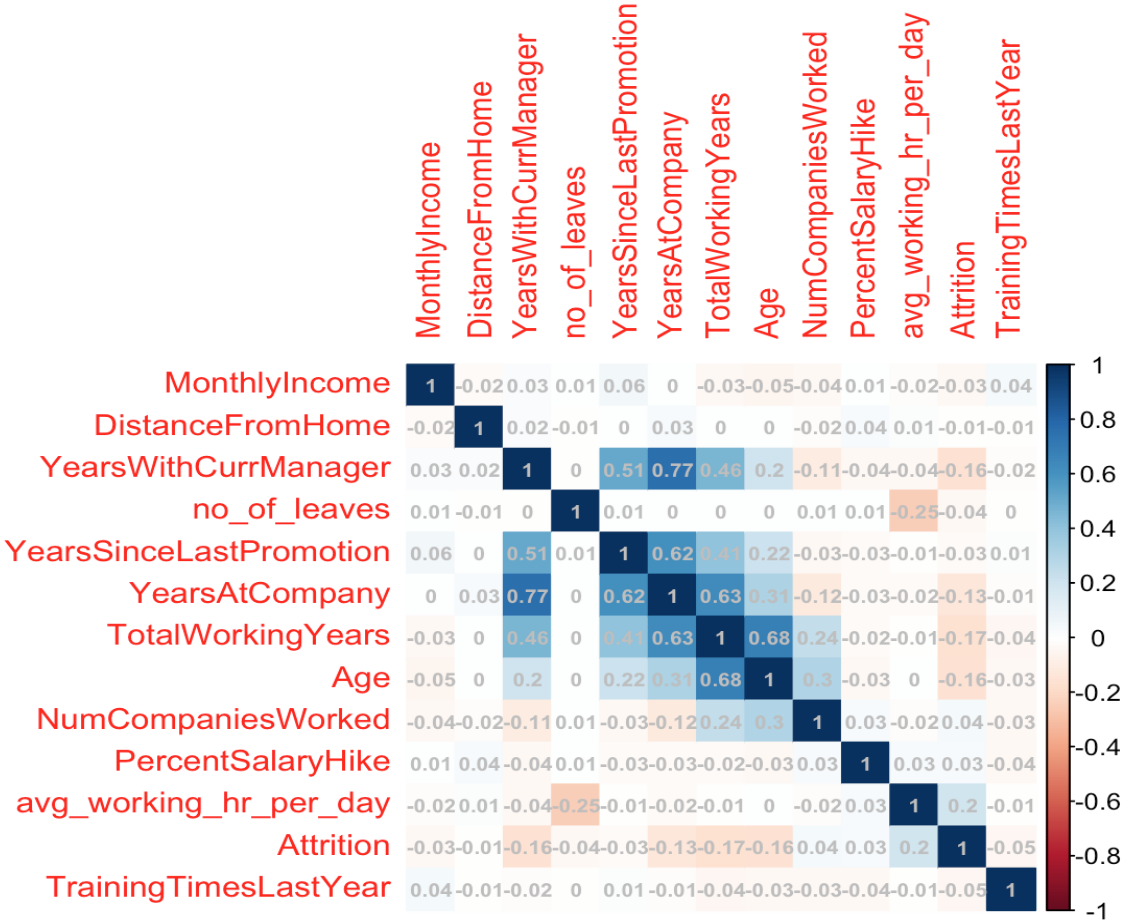
No major inference can be drawn from it.

Bivariate Analysis II (Continuous Vs Continuous Variable)

Bivariate analysis is done on the variables after univariate analysis are complete. This analysis give more insights about the variables with relation with another variable.

Bivariate analysis on the continuous variables (using correlation, etc.)

Correlation Matrix:



	MonthlyIncome	DistanceFromHome	YearsWithCurrManager	no_of_leaves	YearsSinceLastPromotion	YearsAtCompany	TotalWorkingYears	Age	NumCompaniesWorked	PercentSalaryHike	avg_working_hr_per_day	Attrition	TrainingTimesLastYear
MonthlyIncome	1	-0.02	0.03	0.01	0.06	0	-0.03	-0.05	-0.04	0.01	-0.02	-0.03	0.04
DistanceFromHome	-0.02	1	0.02	-0.01	0	0.03	0	0	-0.02	0.04	0.01	-0.01	-0.01
YearsWithCurrManager	0.03	0.02	1	0	0.51	0.77	0.46	0.2	-0.11	-0.04	-0.04	-0.16	-0.02
no_of_leaves	0.01	-0.01	0	1	0.01	0	0	0	0.01	0.01	-0.25	-0.04	0
YearsSinceLastPromotion	0.06	0	0.51	0.01	1	0.62	0.41	0.22	-0.03	-0.03	-0.01	-0.03	0.01
YearsAtCompany	0	0.03	0.77	0	0.62	1	0.63	0.31	-0.12	-0.03	-0.02	-0.13	-0.01
TotalWorkingYears	-0.03	0	0.46	0	0.41	0.63	1	0.68	0.24	-0.02	-0.01	-0.17	-0.04
Age	-0.05	0	0.2	0	0.22	0.31	0.68	1	0.3	-0.03	0	-0.16	-0.03
NumCompaniesWorked	-0.04	-0.02	-0.11	0.01	-0.03	-0.12	0.24	0.3	1	0.03	-0.02	0.04	-0.03
PercentSalaryHike	0.01	0.04	-0.04	0.01	-0.03	-0.03	-0.02	-0.03	0.03	1	0.03	0.03	-0.04
avg_working_hr_per_day	-0.02	0.01	-0.04	-0.25	-0.01	-0.02	-0.01	0	-0.02	0.03	1	0.2	-0.01
Attrition	-0.03	-0.01	-0.16	-0.04	-0.03	-0.13	-0.17	-0.16	0.04	0.03	0.2	1	-0.05
TrainingTimesLastYear	0.04	-0.01	-0.02	0	0.01	-0.01	-0.04	-0.03	-0.03	-0.04	-0.01	-0.05	1

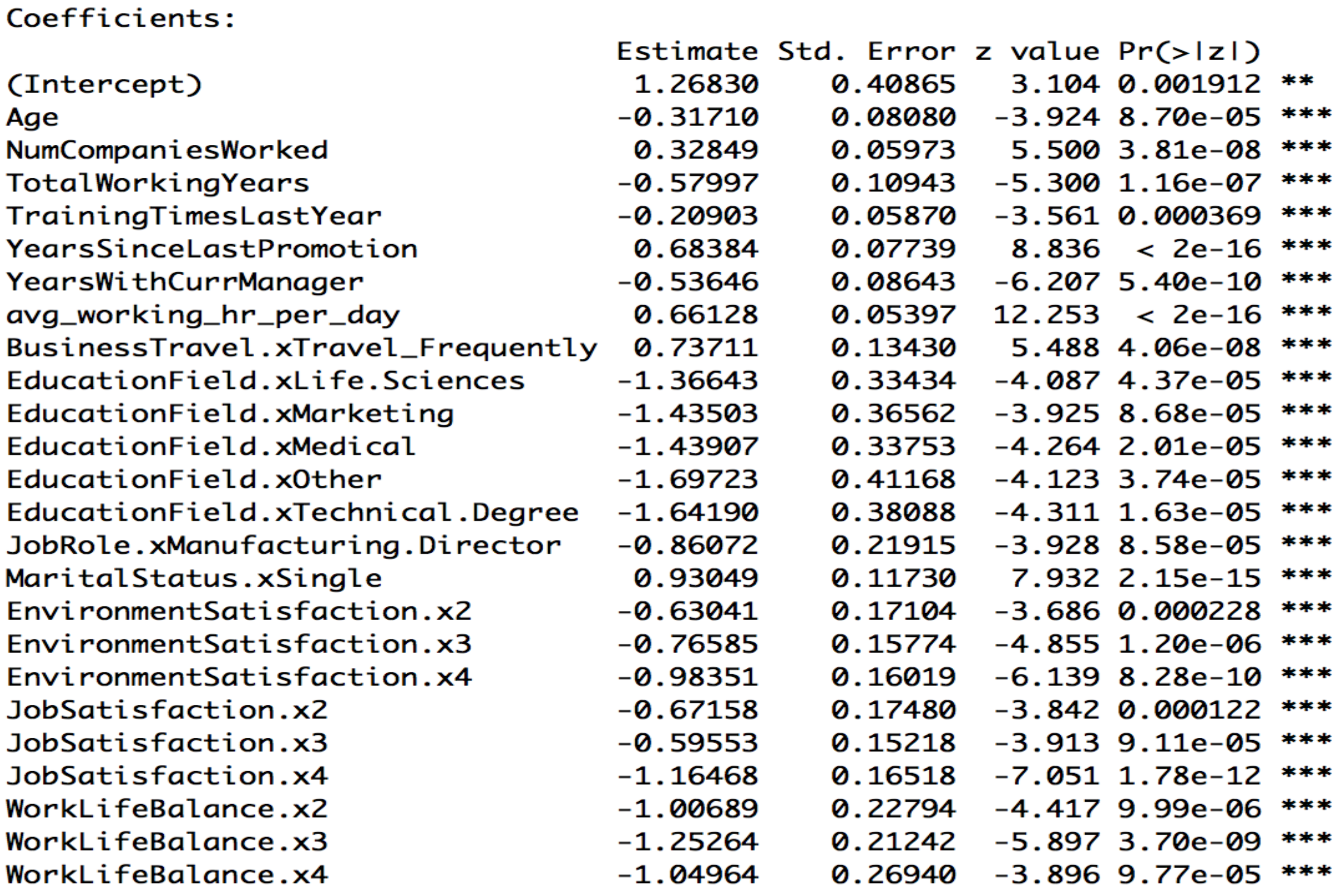
Model Building

Following is the process followed to build the models

1. Built the models using the **glm** ‘generalized linear model’ method and distribution as ‘binomial’, since the dependent variable ‘attrition’ only have ‘Y’ or ‘N’ occurrences.
2. Used the **stepAIC** function on the initial model to remove the non significant variables from the model.
3. From Model 2 onwards the insignificant variables are removed based on the VIF (variance inflation factor) and p-value.
4. For each iteration a single non-significant variable is removed, this process is repeated till all the non-significant variables are removed from the model.
5. For this analysis, we gone till model #13, to get the final model.

Final model (driver variables)

Below are the variables from the final model.

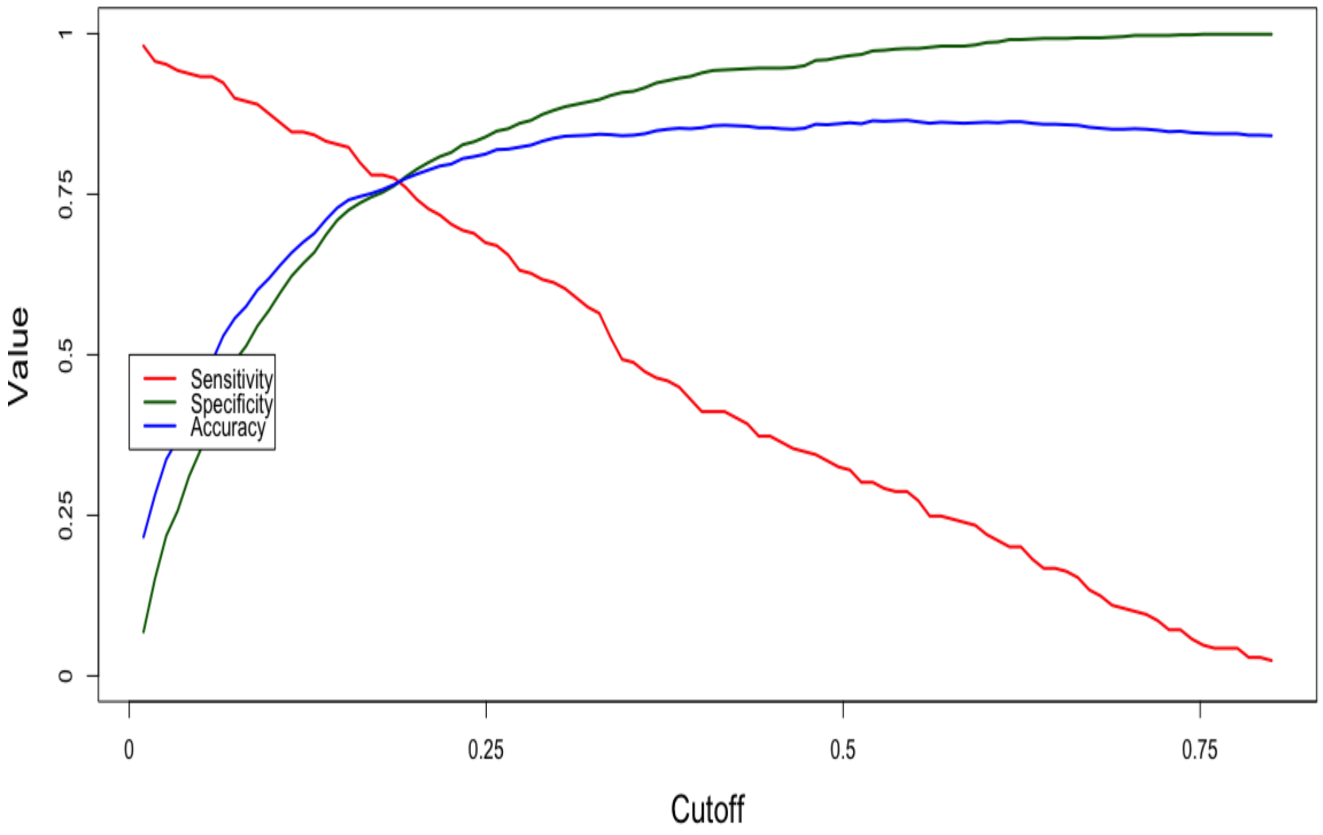


```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)                        1.26830    0.40865   3.104 0.001912 ** 
Age                               -0.31710    0.08080  -3.924 8.70e-05 ***
NumCompaniesWorked                 0.32849    0.05973   5.500 3.81e-08 ***
TotalWorkingYears                 -0.57997    0.10943  -5.300 1.16e-07 ***
TrainingTimesLastYear             -0.20903    0.05870  -3.561 0.000369 ***
YearsSinceLastPromotion            0.68384    0.07739   8.836  < 2e-16 ***
YearsWithCurrManager              -0.53646    0.08643  -6.207 5.40e-10 ***
avg_working_hr_per_day             0.66128    0.05397  12.253  < 2e-16 ***
BusinessTravel.xTravel_Frequently  0.73711    0.13430   5.488 4.06e-08 ***
EducationField.xLife.Sciences     -1.36643    0.33434  -4.087 4.37e-05 ***
EducationField.xMarketing         -1.43503    0.36562  -3.925 8.68e-05 ***
EducationField.xMedical           -1.43907    0.33753  -4.264 2.01e-05 ***
EducationField.xOther             -1.69723    0.41168  -4.123 3.74e-05 ***
EducationField.xTechnical.Degree  -1.64190    0.38088  -4.311 1.63e-05 ***
JobRole.xManufacturing.Director   -0.86072    0.21915  -3.928 8.58e-05 ***
MaritalStatus.xSingle              0.93049    0.11730   7.932 2.15e-15 ***
EnvironmentSatisfaction.x2        -0.63041    0.17104  -3.686 0.000228 ***
EnvironmentSatisfaction.x3        -0.76585    0.15774  -4.855 1.20e-06 ***
EnvironmentSatisfaction.x4        -0.98351    0.16019  -6.139 8.28e-10 ***
JobSatisfaction.x2                -0.67158    0.17480  -3.842 0.000122 ***
JobSatisfaction.x3                -0.59553    0.15218  -3.913 9.11e-05 ***
JobSatisfaction.x4                -1.16468    0.16518  -7.051 1.78e-12 ***
WorkLifeBalance.x2                -1.00689    0.22794  -4.417 9.99e-06 ***
WorkLifeBalance.x3                -1.25264    0.21242  -5.897 3.70e-09 ***
WorkLifeBalance.x4                -1.04964    0.26940  -3.896 9.77e-05 ***
```

Model Evaluation

Model - Cut-off

Used the ROC curve to identify the ideal probability cut off value for the model for an optimal Sensitivity, Specificity and Accuracy.



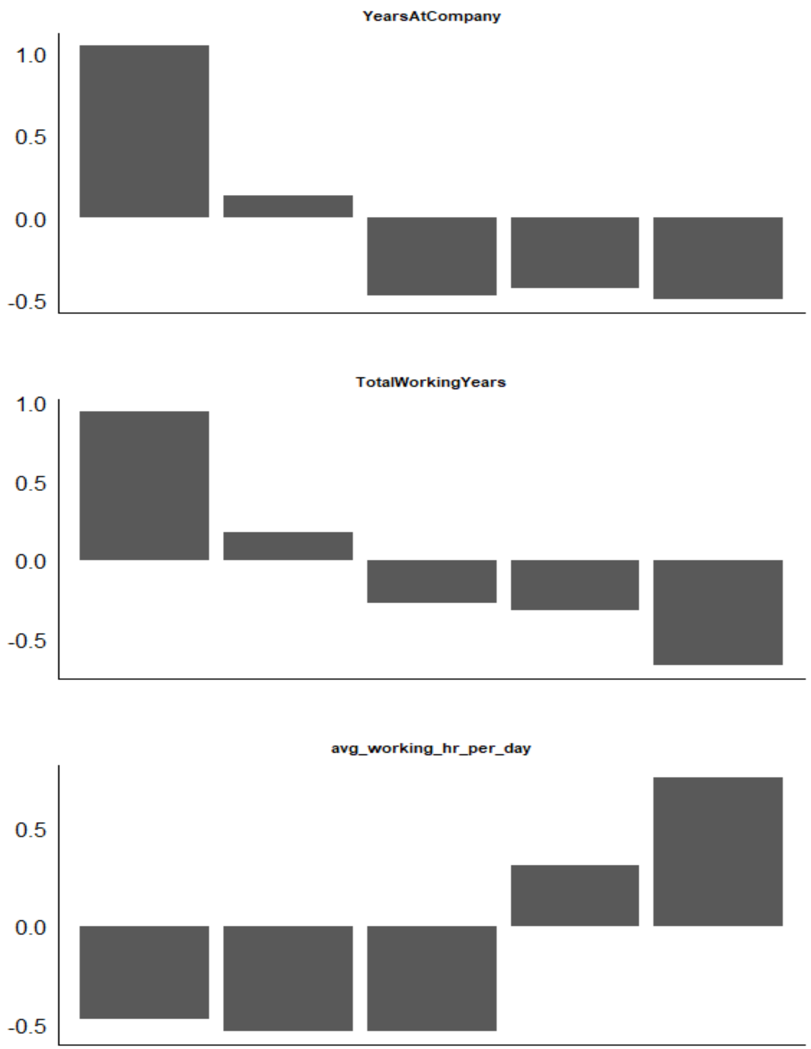
The ideal cut off identified using the ROC curve is **0.1855556** and it is used in the final model to obtain the following result:

Accuracy - 0.7651
Sensitivity - 0.7751
Specificity - 0.7632

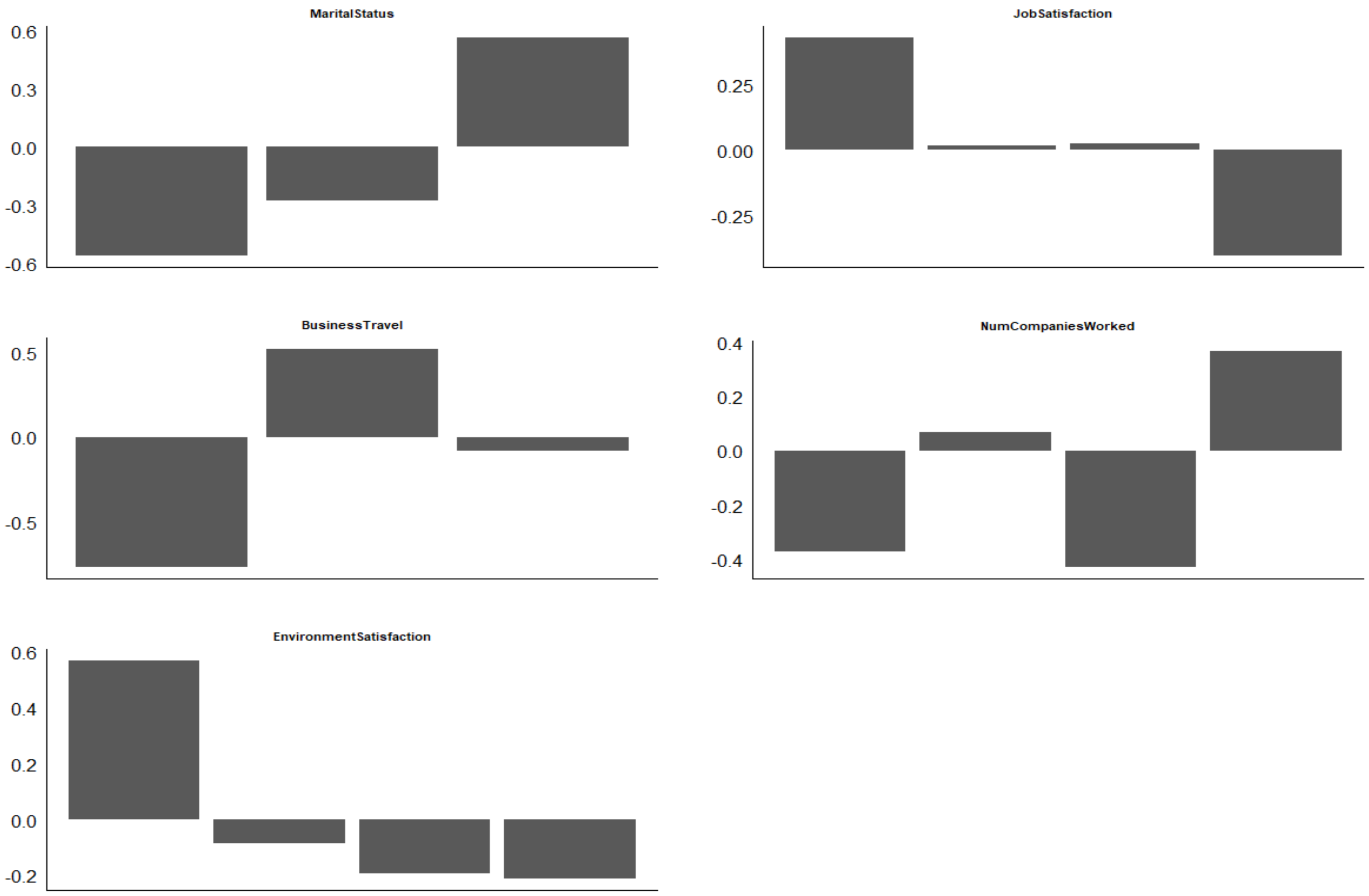
Following is the confusion matrix for the final model:

Predicted	Actual	
	No	**Yes**
No	825	47
Yes	256	162

EDA Based On Information Gain(WOE Analysis)

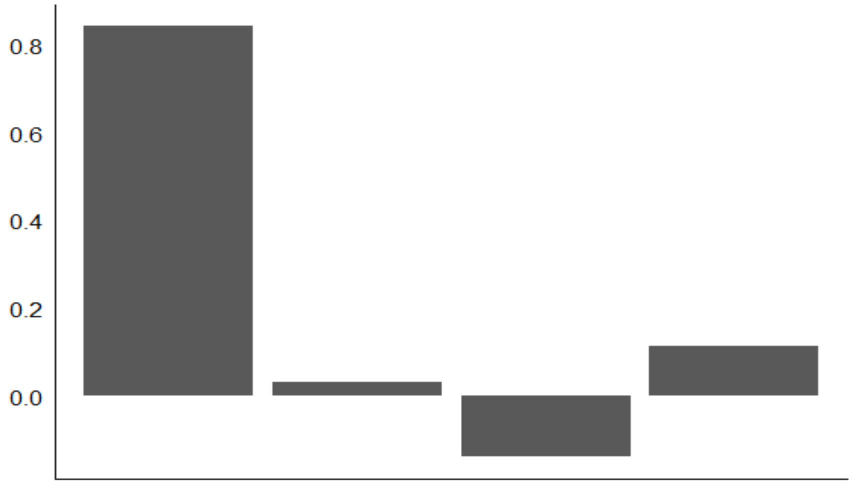


Based on their IV values all these variables variables are “strong” predictors.

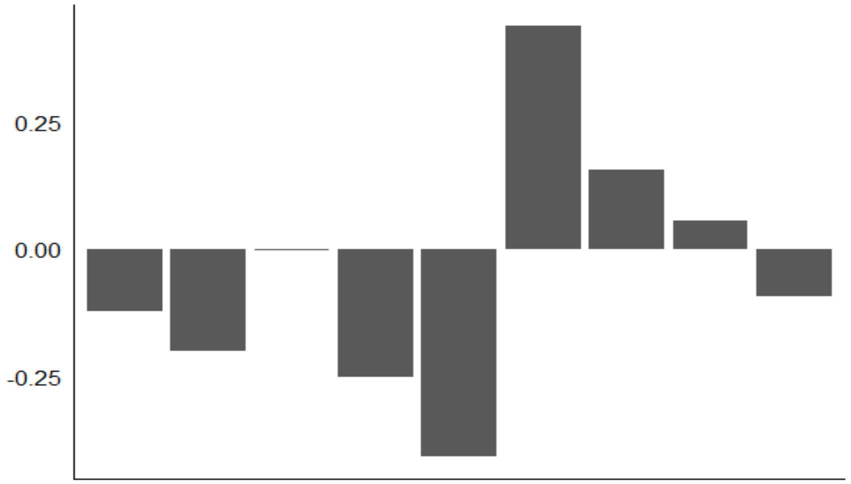


Based on their IV values all these variables variables are “moderate” predictors.

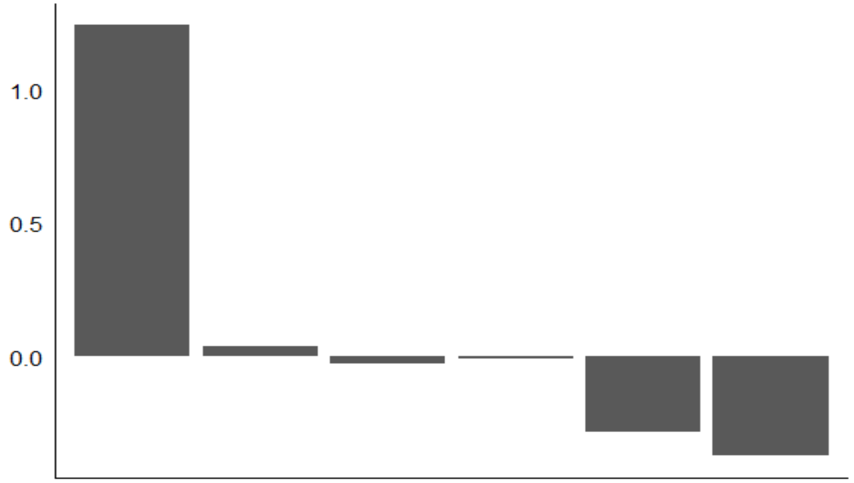
WorkLifeBalance



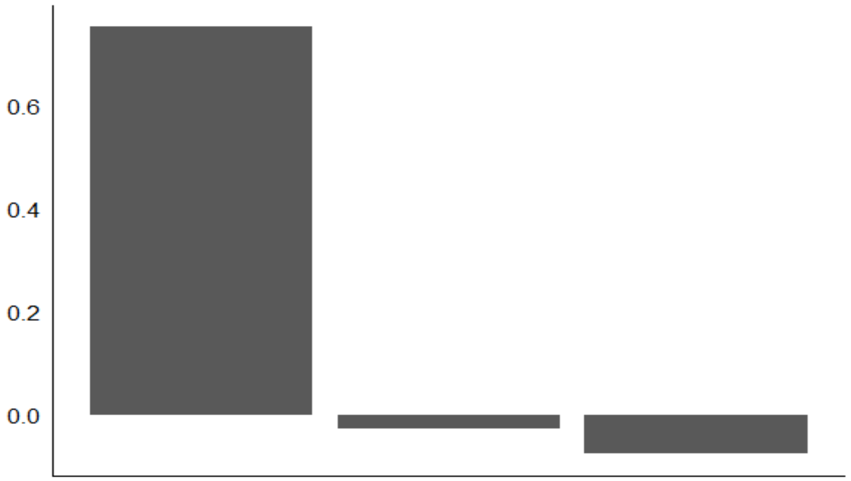
JobRole



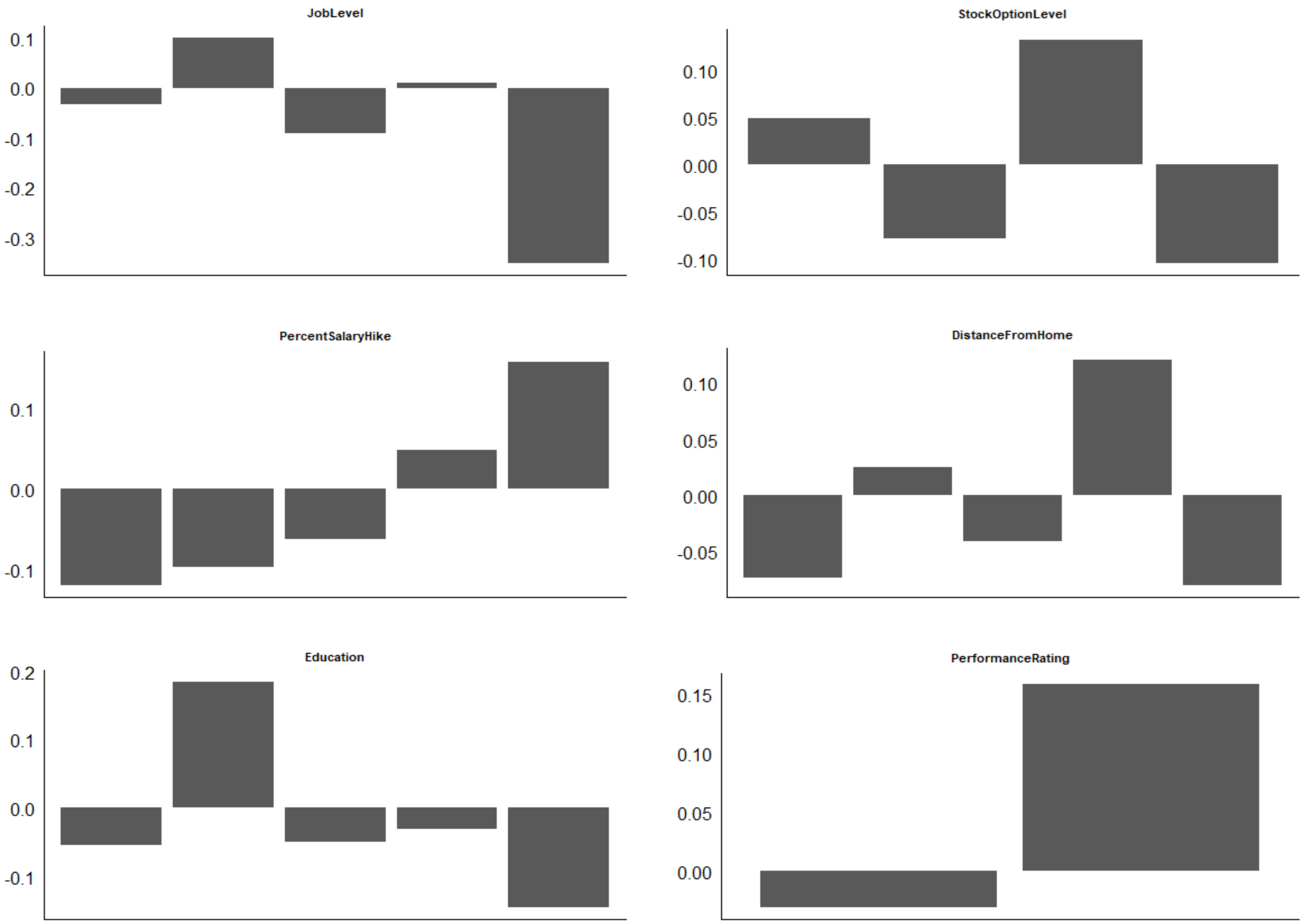
EducationField



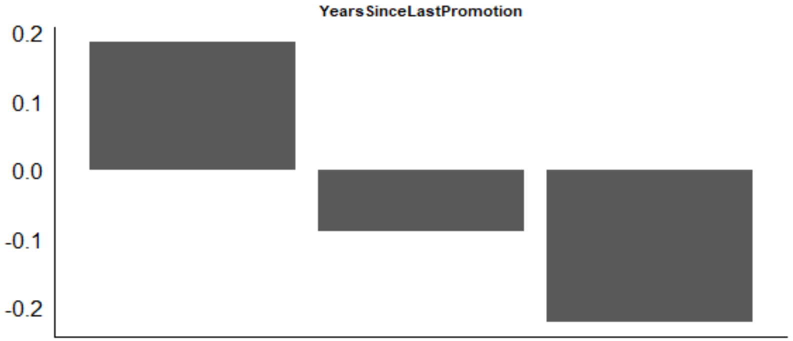
Department

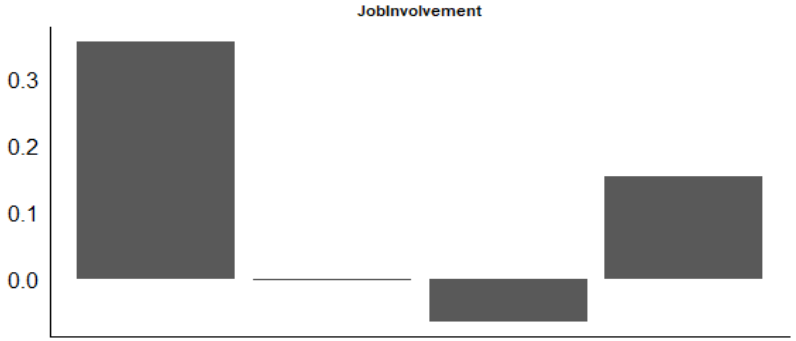


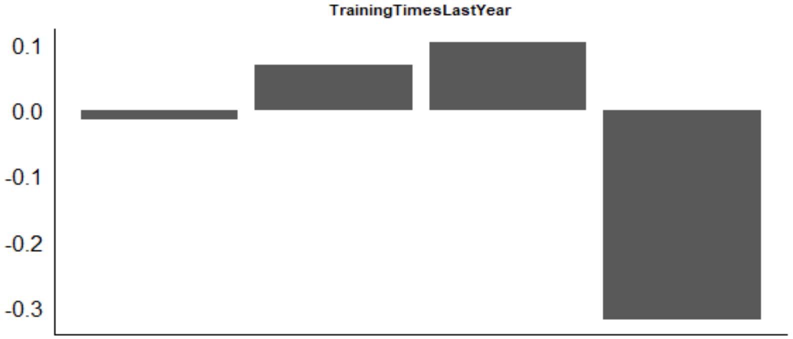
Based on their IV values all these variables variables are “moderate” predictors.

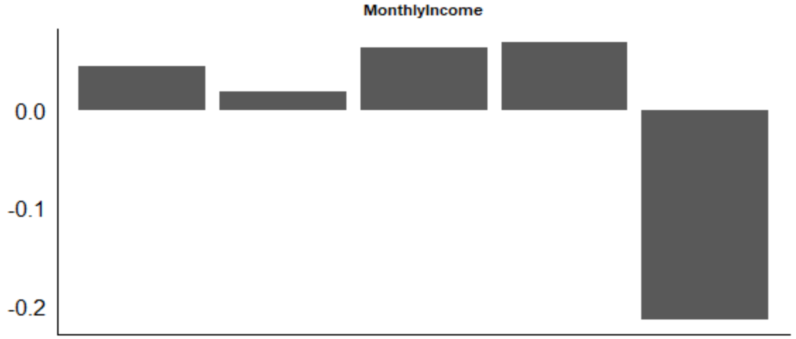


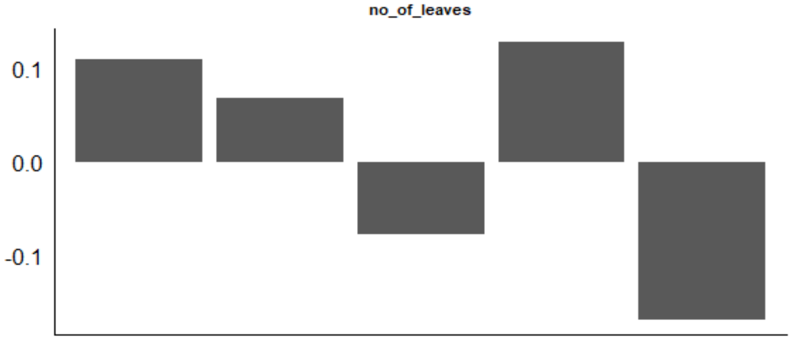
Based on their IV values all these variables variables are “weak” predictors.





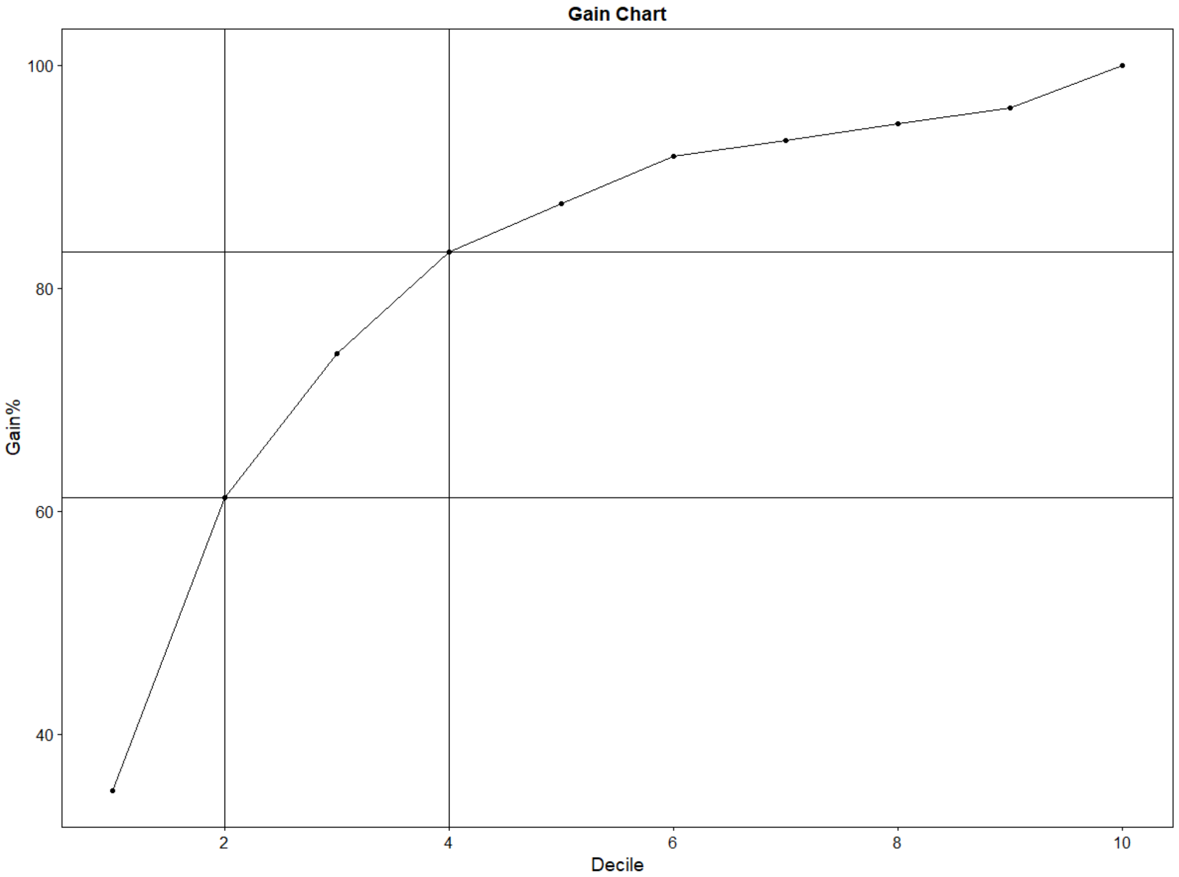


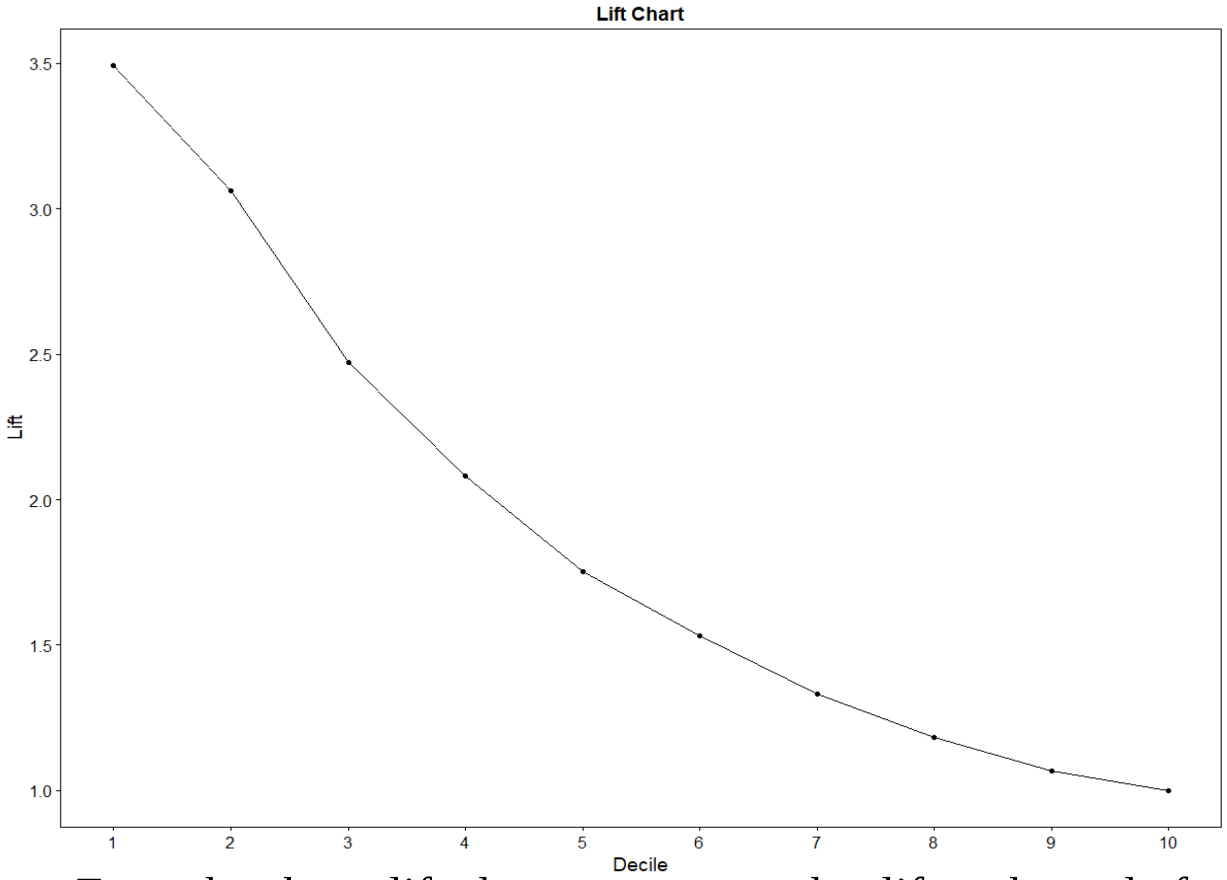




Based on their IV values all these variables variables are “weak” predictors.

Model Evaluation - ‘Gain and Lift Chart’





From the above GAIN chart, we can infer that by focusing on the top 40% of the employees (after sorting them by probabilities) .We can focus on top 83% of the employees who are likely to leave the company.

From the above lift chart, we can see that lift at the end of
1st decile is 3.49
2nd decile is 3.06
3rd decile is 2.47
4th decile is 2.08

Model Evaluation – KS Statistic

- **KS statistic** is an indicator to assess the predicting power of the models.
- It is used to give indication of quality of the model.
- A good model is one for which the KS statistic is equal to 40% or more
- **The KS statistic for our model came as 0.538019. (~ 54%)**
- **Hence our Model KS statistic is nearly 54%**

INFERENCES - RECOMMENDATIONS

VARIABLES	INFERENCES	RECOMMENDATIONS
MaritalStatus.Xsingle [Estimate : 0.93049]	Employees who are single have high probability to leave the company.	Try to have interaction with employees who are single at regular intervals to understand their concerns.
Avg_working_hr_per_day[Estimate: 0.66128]	Employees who are working for longer hours (more than the nominal working hrs. per day) are more likely to leave.	Keep track of the employees who are logging for more than the required hours. Discuss with their managers, on how their workload can be reduced.
YearsSinceLastPromotion [Estimate: 0.6834]	Employees who haven't been promoted for longer number of years are more likely to leave.	Keep track of such employees , discuss with their managers why they are not promoted ? If the employees are not willing to take promotion ,try to understand what is the reason for it ? If there are not enough perks post promotion, then this is the area "HR" should work on . If there are not enough opportunities then HR should discuss with management on how new roles can be created and can be leveraged to give promotions to eligible employees (who have not been promoted for a time).

VARIABLE	INFERENCES	RECOMMENDATIONS
BusinessTravel.xTravel_Frequently [Estimate:0.73711]	Employees those who travel frequently are **more likely** to leave	Track the employees who travel very frequently. Discuss with these employees are they ok to travel ? If not, what are the reasons ? Try to address the grievances of such employees by discussing with their reporting managers.
WorkLifeBalance	Employees who have work life balance 2,3 or 4(i.e. Good , Better, Best) are **less likely** to leave.	Have regular surveys to track employees “Work life balance” .Those who are having bad work life balance, reach out to their managers to keep them updated about this info and see what can be done to improve their work life balance.
JobSatisfaction	Employees those who have job satisfaction[2,3,4](i.e. Medium , High, VeryHigh) are **less likely** to leave.	Have regular surveys to keep rack of Job Satisfaction ratings. Be in regular touch with the reporting Managers of the employees who have lower job satisfaction rating and see how those employees can be addressed.

VARIABLE	INFERENCES	RECOMMENDATIONS
TotalWorkingYears	Higher the total Working Years, less likely employee will leave.	N/A
YearsWithCurrManager	Higher the number of years the employee works with a manager, less likely employee will leave.	Need to create a framework where the manager and his same team can work together for a longer period of time.
EnvironmentSatisfaction	If the employee is satisfied with the working environment on the rating of (2,3,4), less likely employee will leave.	Need to have regular surveys and inputs from the employee on how the working environment can be improved.
JobRole	Employees those have job role as Manufacturing Director are less likely to leave.	N/A