HR ANALYTICS CASE STUDY

Group Name:

1. Snigdha Prakash
2. Amisha
3. Rahul Doshi

Business Objective

The HR Analytics case study is being carried for the company who wants to act upon the cause of attrition in the company. Attrition in any company, be it because of the employees leaving voluntarily or being fired harms the companies reputation and the issues must be addressed in a right way.

Here, the objective is to find the factor that affects the attrition rate and the ways to minimize their cause.

Business Constraints:

- Only 1-year data of the employees has been provided

Goals:

- Find the factors affecting the attrition rate
- Perform univariate and multivariate analysis of variables of interest
- Express the findings in terms of neat visualizations
- Suggest methods of minimizing the cause of attrition

Problem solving methodology

Data Collection

Data has been provided in 5 categories -

Employee survey data - which provide information on employees emotional well being

Manager survey data – which tells about the performance of the employee under a manager

In and out time – which gives the time recorded for employee reporting at office and time at which the employee leaves the office

General data – This is the master data which contains other details of he employee

Data Cleaning

It involves -

- Merging the data from different sources into a common dataframe for analysis
- Handling the missing values
- Checking outliers in data and treating them
- Formatting date and time properly for proper analysis
- Giving a proper column name to in-time and out-time dataset

Univariate and Multivariate Analysis

This includes –

- Analysis the behavior of each variables and also finding the relationship amongst them
- We also find the correlation between the variables and analyze their significance
- We also plot them to visually see the effects

Variable Treatment

In order to feed correct inputs to our model, we treat the categorical and continuous variables –

- We create dummies of categorical variables
- We scale the continuous variable to avoid misinterpretation of results due to large deviations in values

Model building

- Analyzing the nature of predicted variable.
- Here, it is Attrition which is categorical in nature
- So we use glm function available in R to build logistic regression model

Model testing

We test the model against the test dataset using –

- Confusion Matrix(determining accuracy, specificity and sensitivity)
- KS statistics
- Gain and lift charts

If test results are not acceptable

Restructuring data and rebuilding model

Data Preparation and Cleaning

- **There are 5 datasets given –**

Employee survey data	Manager survey data	
General data	In time of the employees	Out time of the employees

- **There is 1-year worth of data in in-time and out-time dataset**
  - In whole year, there are 12 days when it was a holiday
  - We verified this by finding the index column in both the dataset when all 4410 rows are having NA values
  - We found the time spend by each employee each data by subtracting in-time with the out time
  - There are around 5% of NA values in the resultant dataset. Since the row index of NA in in-time matches with that in out-time dataset, we can safely assume that the employee was absent
  - We replaced those values with zero
  - We are focusing on average time that an employee spends in office each day
- **Merging data**
  - Since the number of rows of all datasets are same and “EmployeeID” is the key field, we can merge them safely
  - There are few variables like – EmployeeCount, Standard Hours, Over18 which are categorical variables having only 1 value. Since they will not be having any significant impact, it is safe to remove them

Data Preparation and Cleaning

- **Other variables having NA –**

Variable Name	Number of NA	% of NA
NumCompaniesWorked	19	0.4%
TotalWorkingYears	9	0.2%
EnvironmentSatisfaction	25	0.5%
JobSatisfaction	20	0.4%
WorkLifeBalance	38	0.8%

Since the % of NA value are very less, we have omitted them from the data

- **Categorization of Age variable**

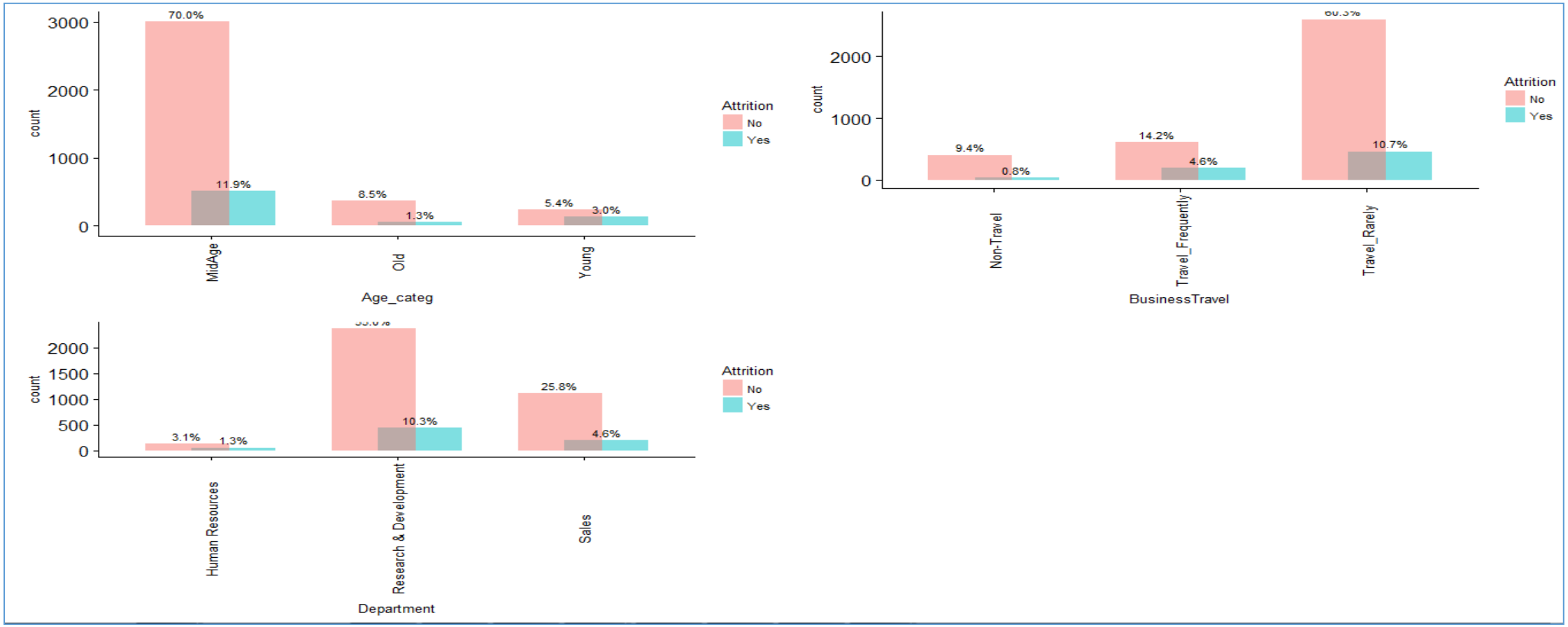
Age	Group
<=25	Young
25 < Age <= 50	MidAge
>50	Old

- **All categorical variables were converted to factor type**

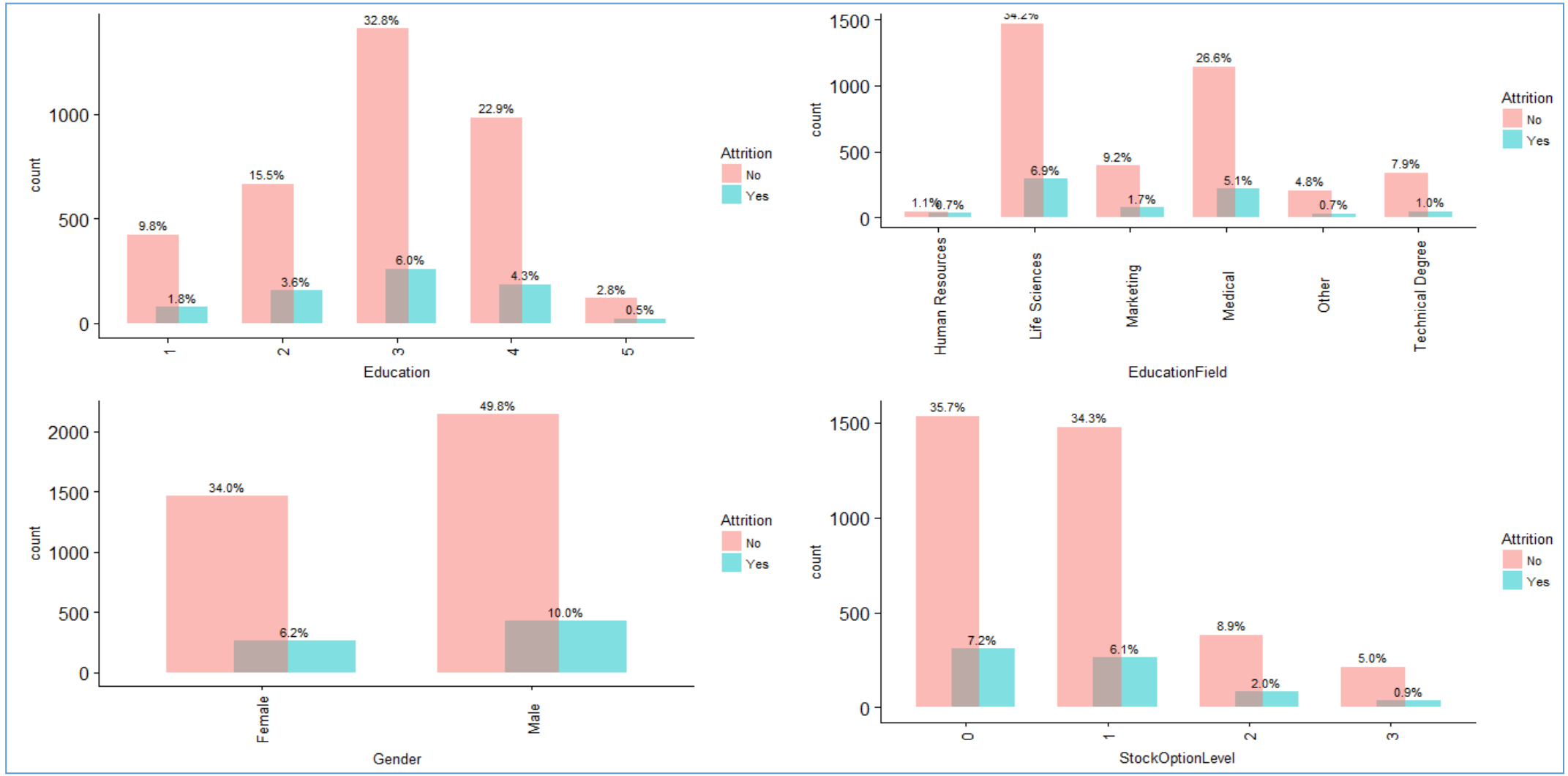
Univariate Analysis of Data

1. Analyzing categorical variables -

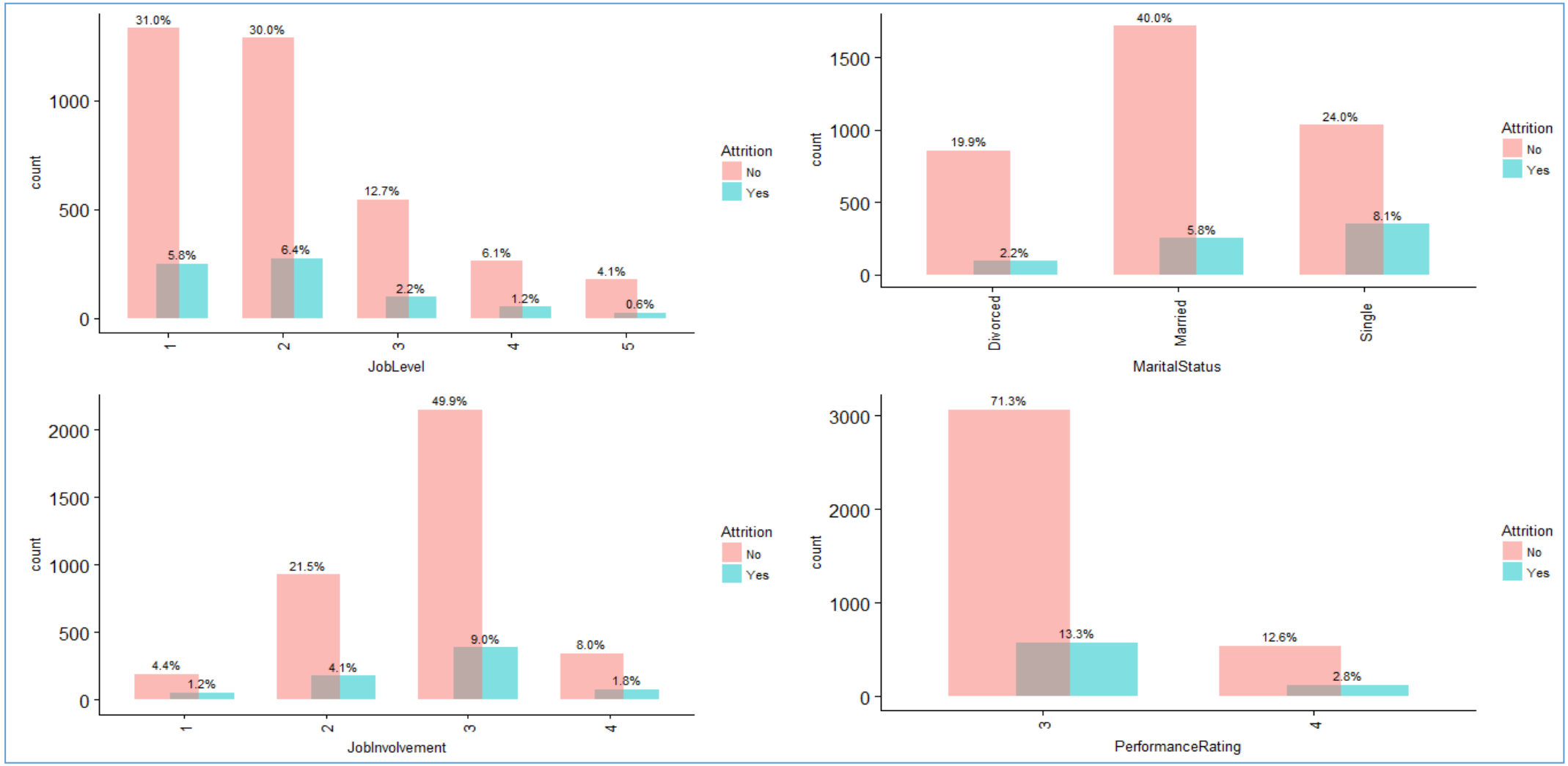
We found the proportion of employees who have left the company against those who are still in the company for each category. Here is the visualization of all categorical variables against attrition.



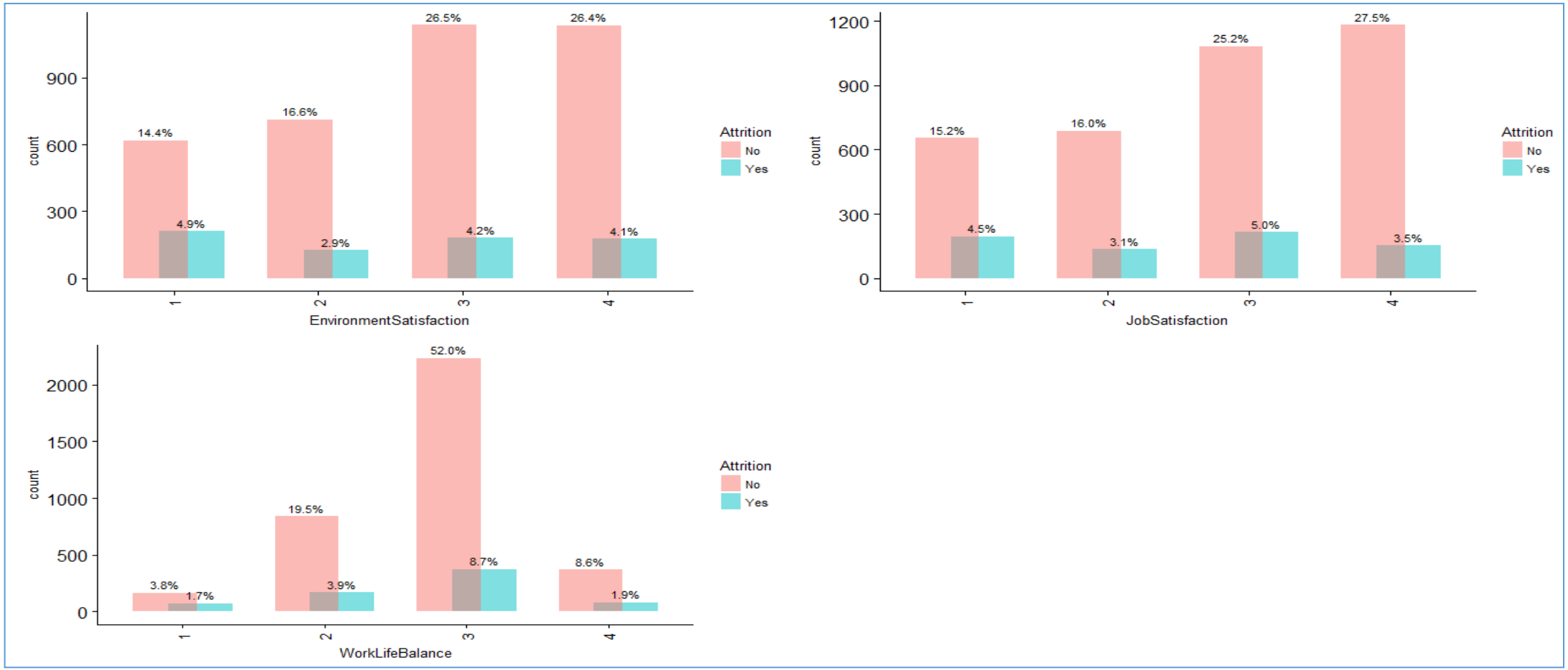
Similarly, we analyze effect of other categorical variable on Attrition rate visually



Similarly, we analyze effect of other categorical variable on Attrition rate visually

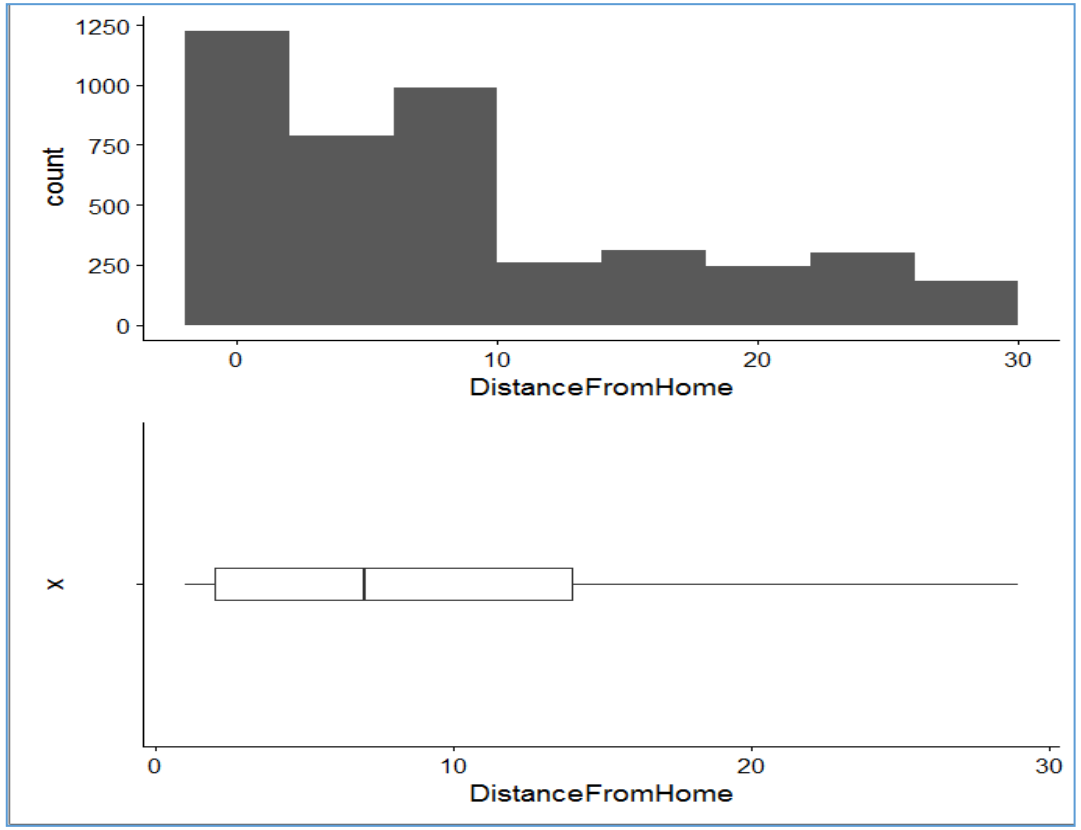


Similarly, we analyze effect of other categorical variable on Attrition rate visually

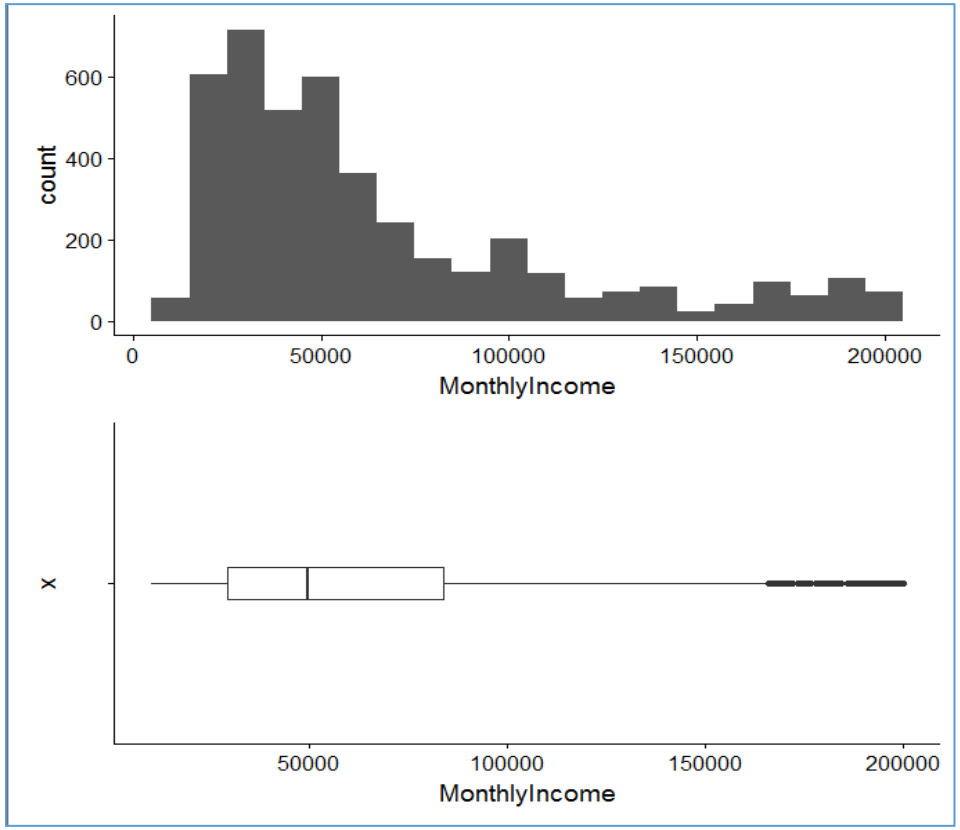


2. Analyzing Continuous variables -

- We plotted the histogram and box plot to get a clear picture of the outliers in the data.
- Since the outliers in the data can be misleading, we have further calculated the quantile values and treated the outlier data with the nearest quantile value

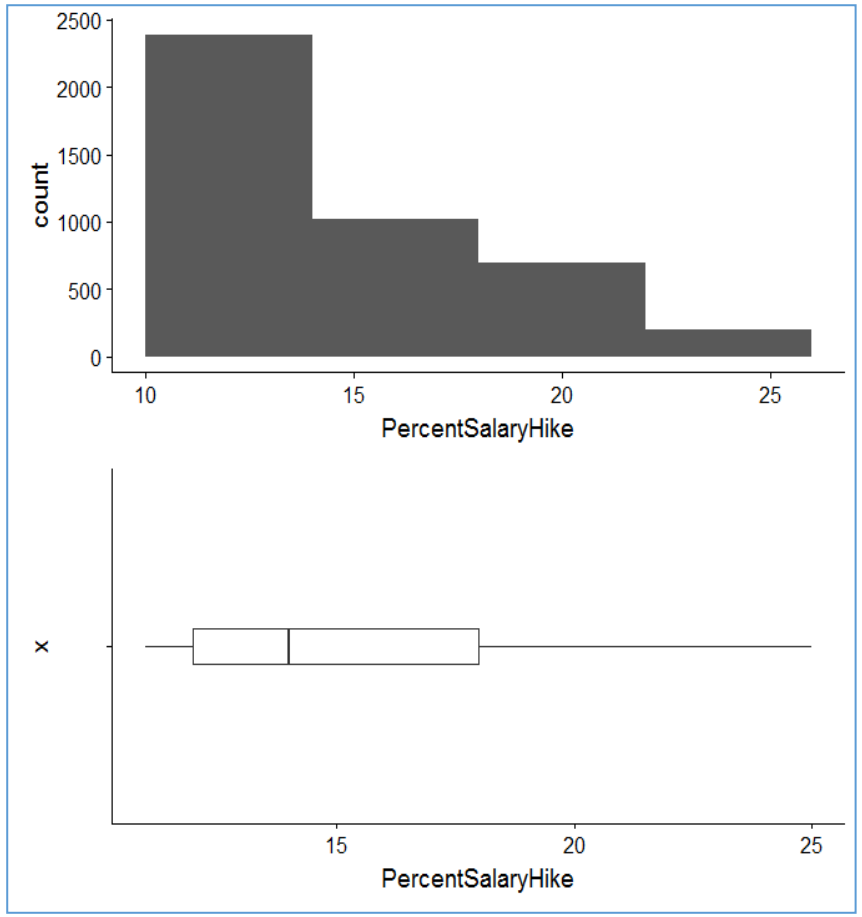


Skewness : 0.95 [Skewed towards the right]

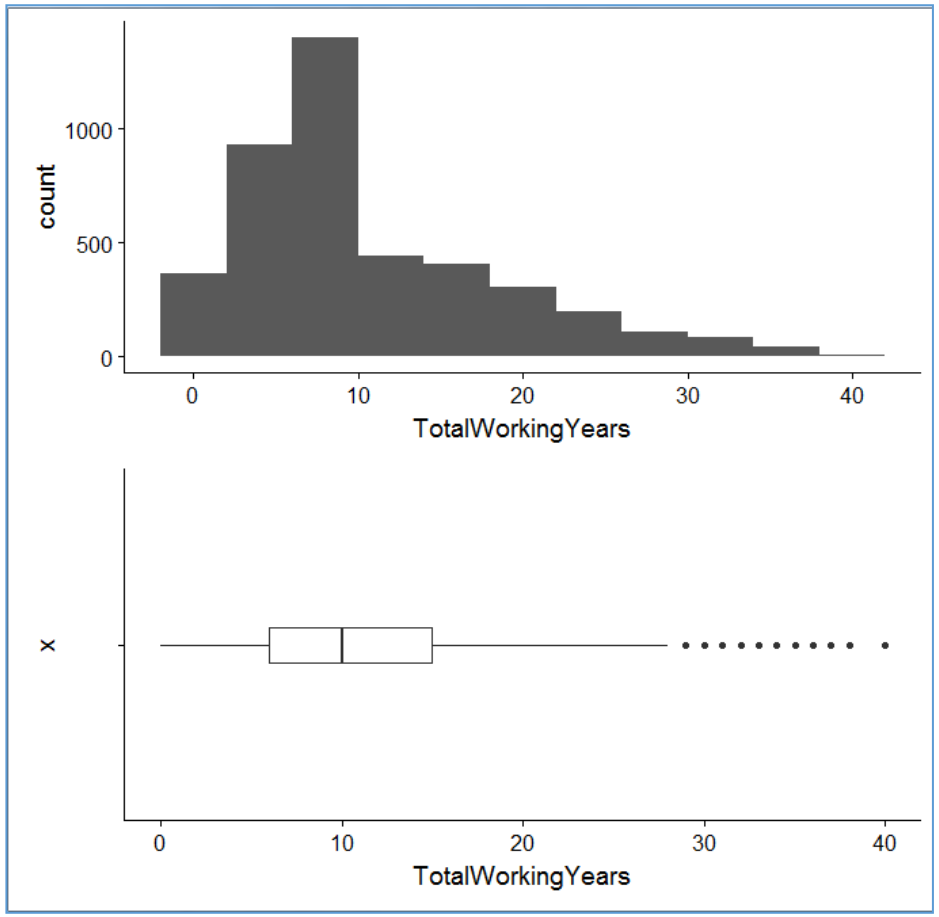


Skewness : 1.36 [Highly Skewed towards the right and having outliers]

Similarly we plot for other continuous variables

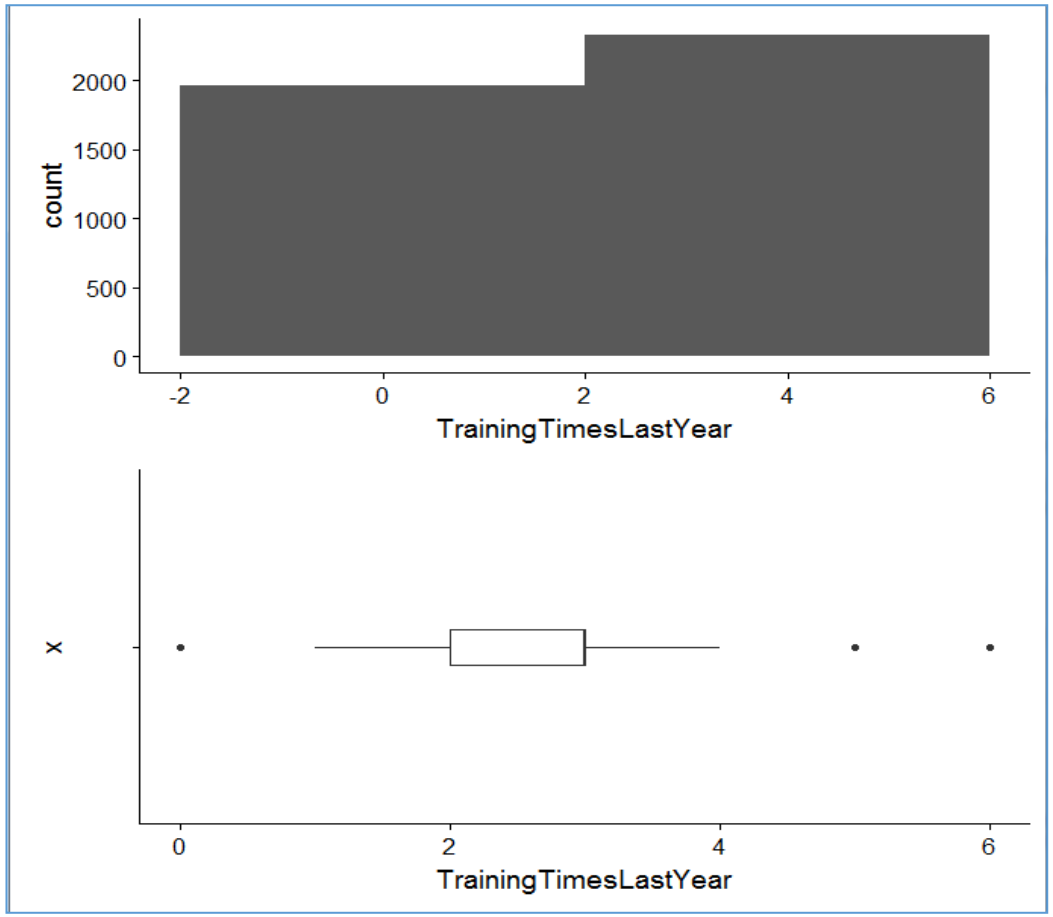


Skewness : 0.82 [Skewed towards the right]

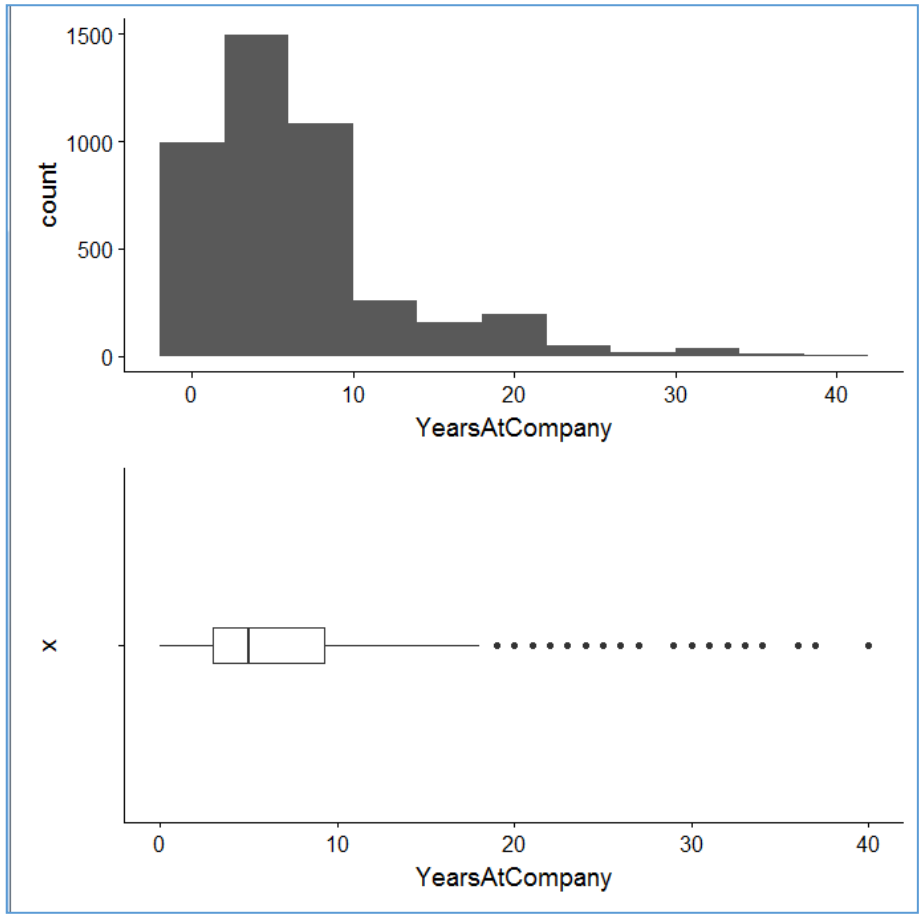


Skewness : 1.36 [Highly Skewed towards the right and having outliers]

Similarly we plot for other continuous variables

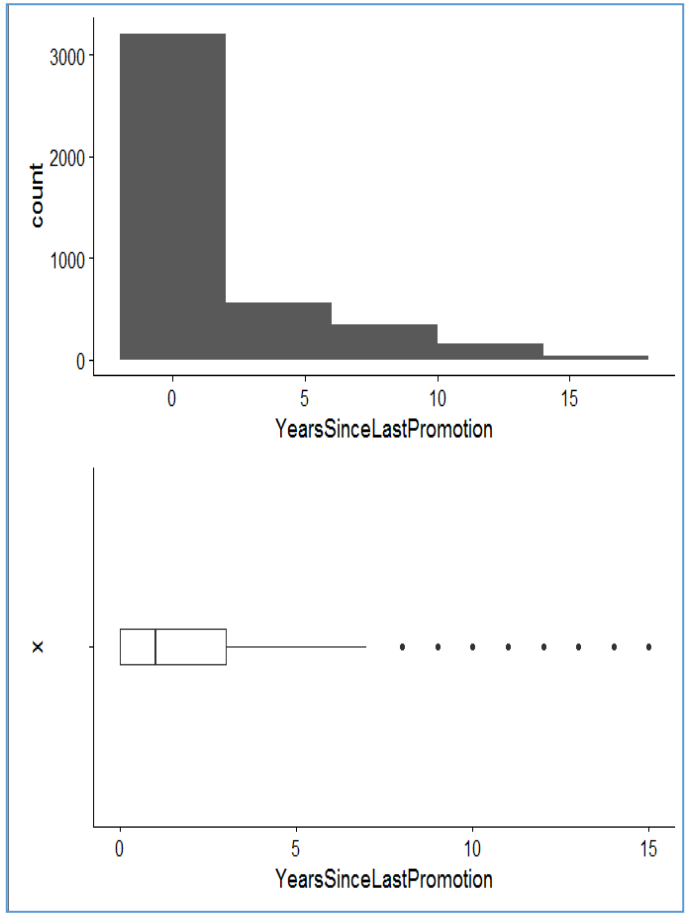


Skewness : 0.54 [not a significant skew. However, it has outliers on both the sides]

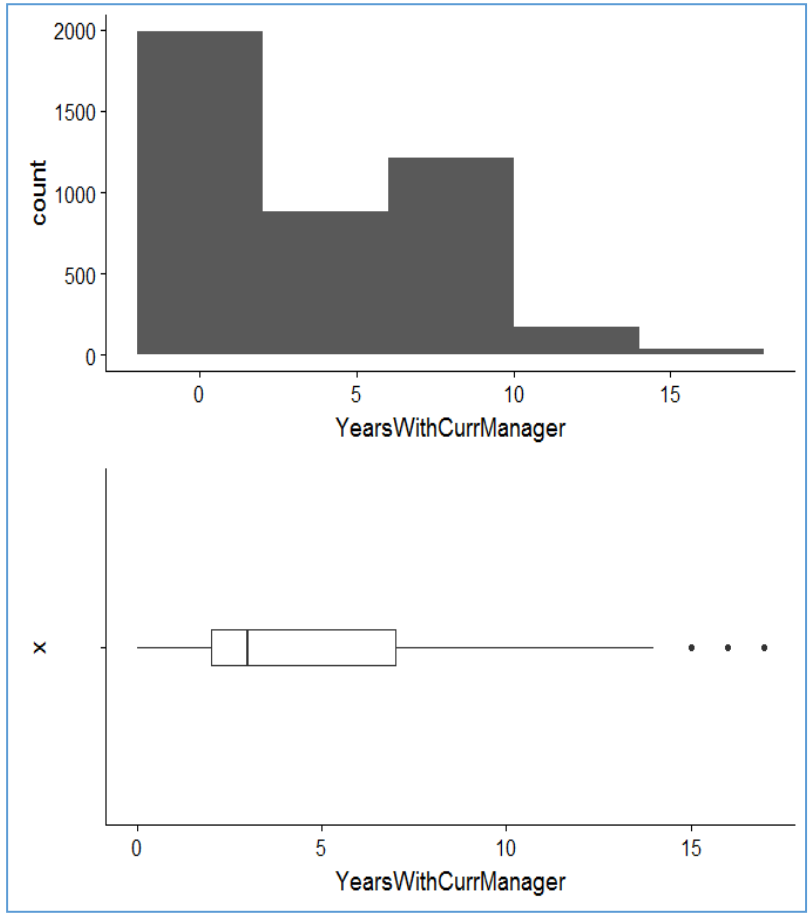


Skewness : 1.76 [Highly Skewed towards the right and having outliers on both the sides]

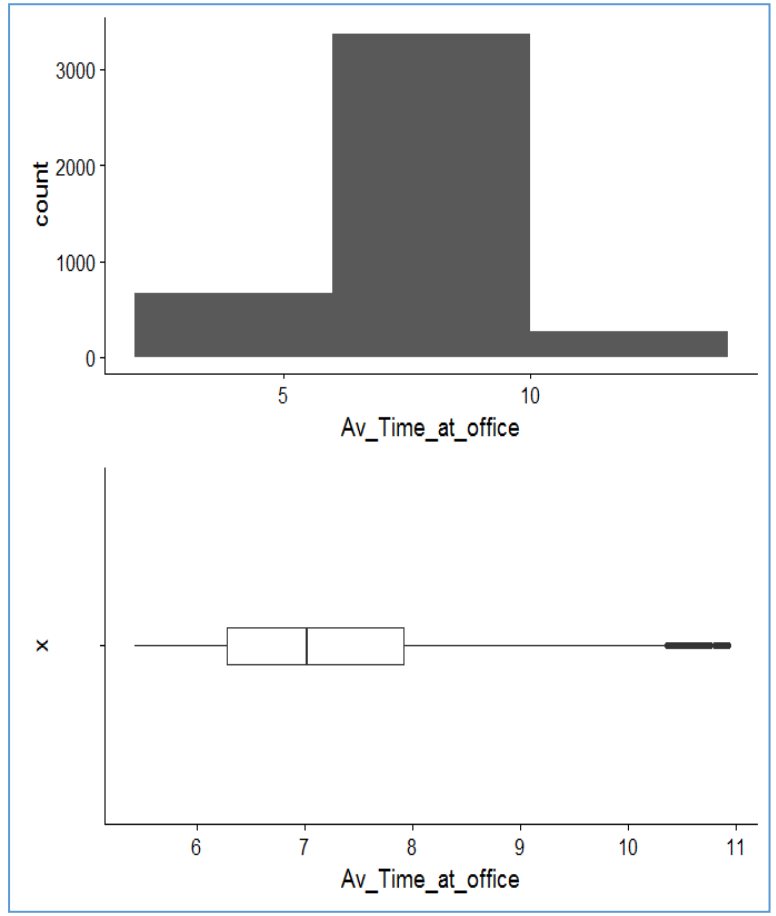
Similarly we plot for other continuous variables



Skewness : 1.98 [Highly skewed towards the right and have outliers]



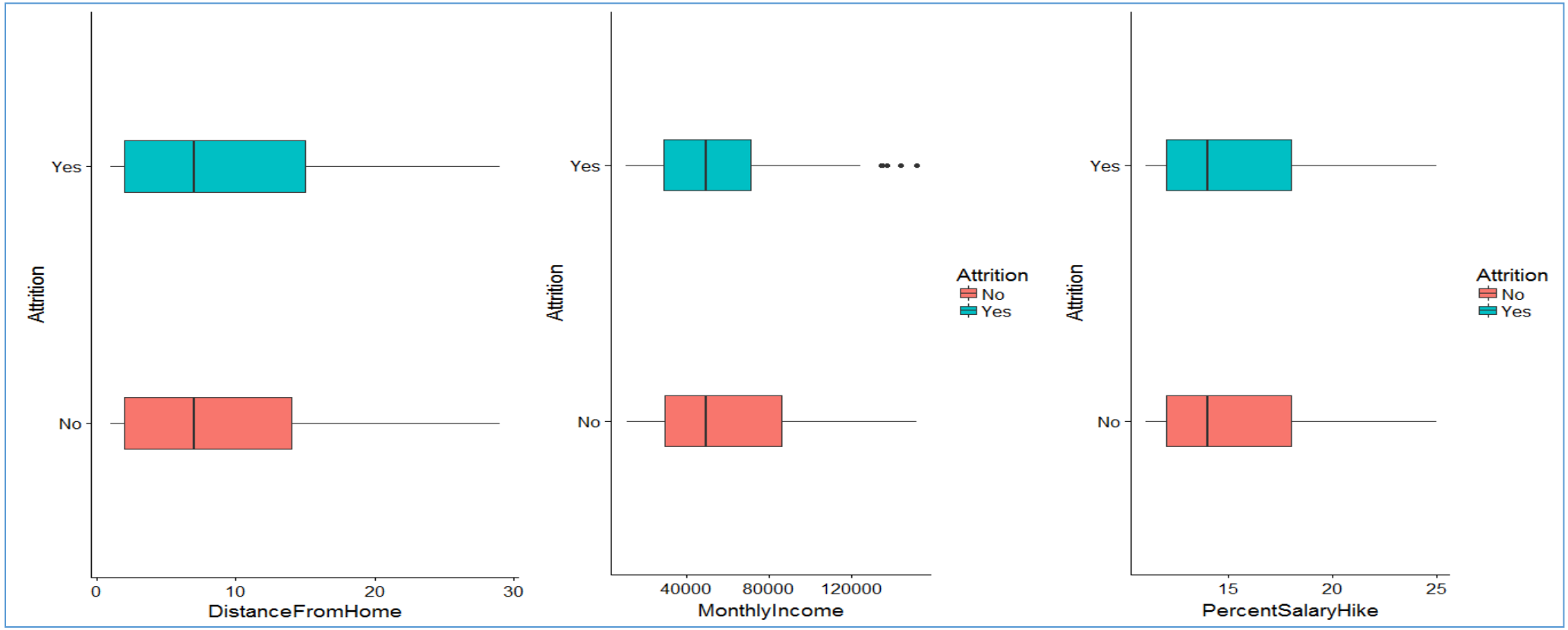
Skewness :0.83[Skewed towards the right and having outliers]



Skewness : 0.86 Skewed towards the right and having outliers

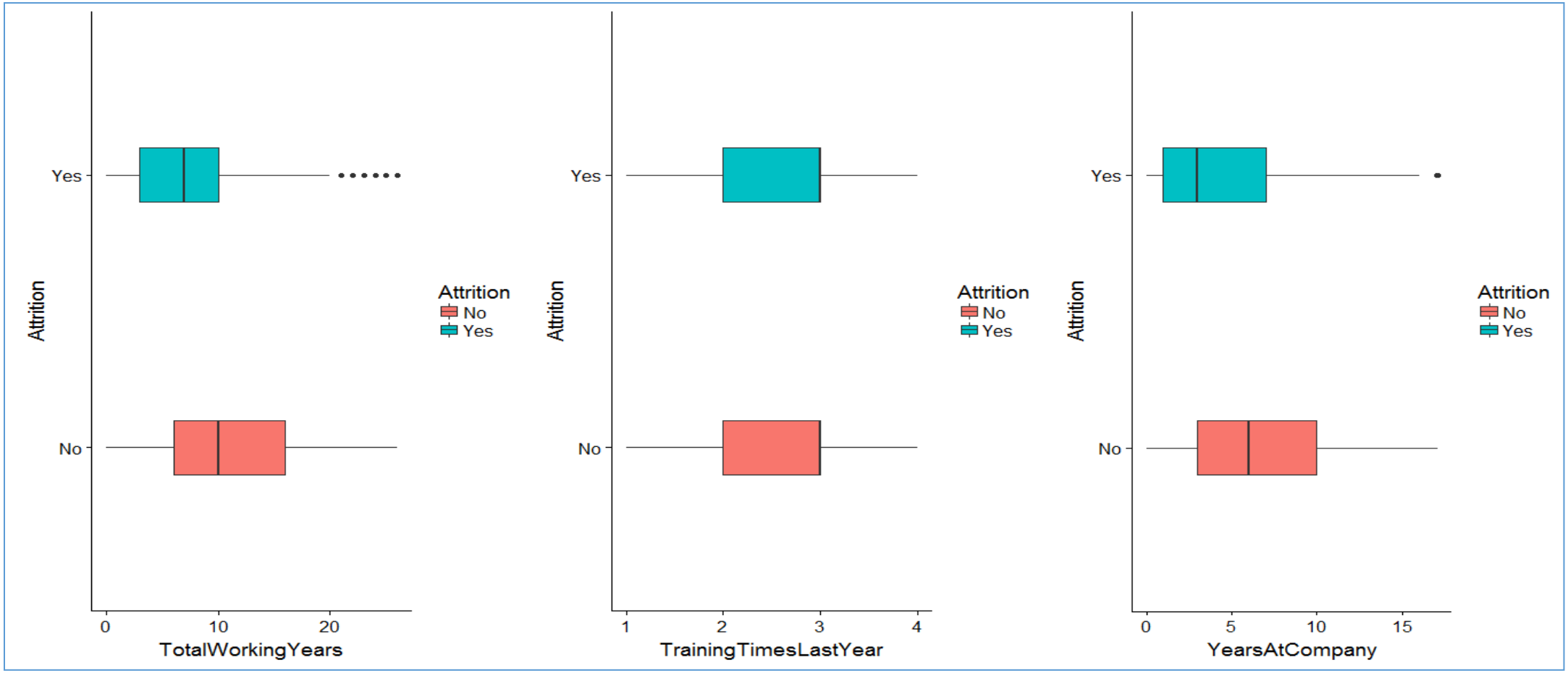
Multivariate Analysis of Data

1. Analyzing the distribution of continuous variables against Attrition and observing how they behave against different levels of attrition



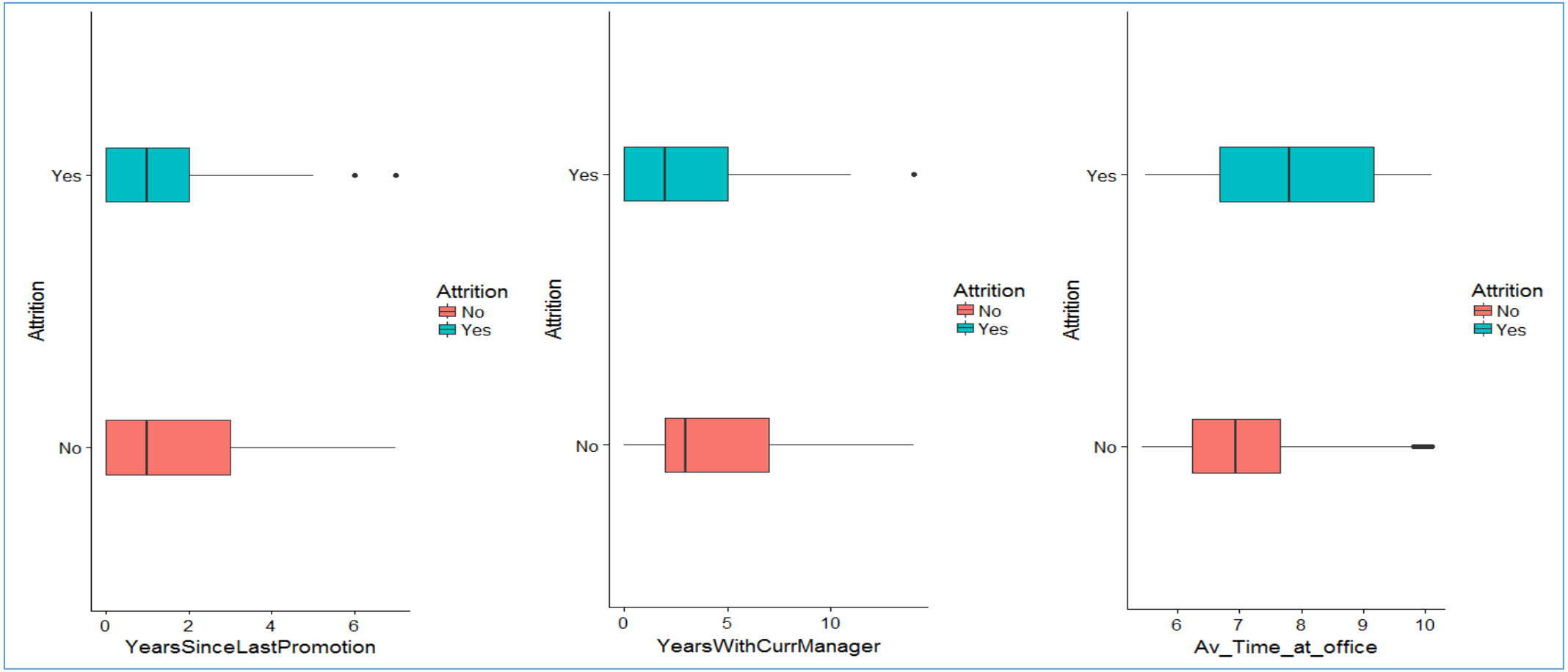
Here, we do not see appreciable difference in distribution for the 2 categories. During modelling, we checked for the significance of the variables and handled them accordingly

Analyzing the distribution of other continuous variables against Attrition and observing how they behave against different levels of attrition



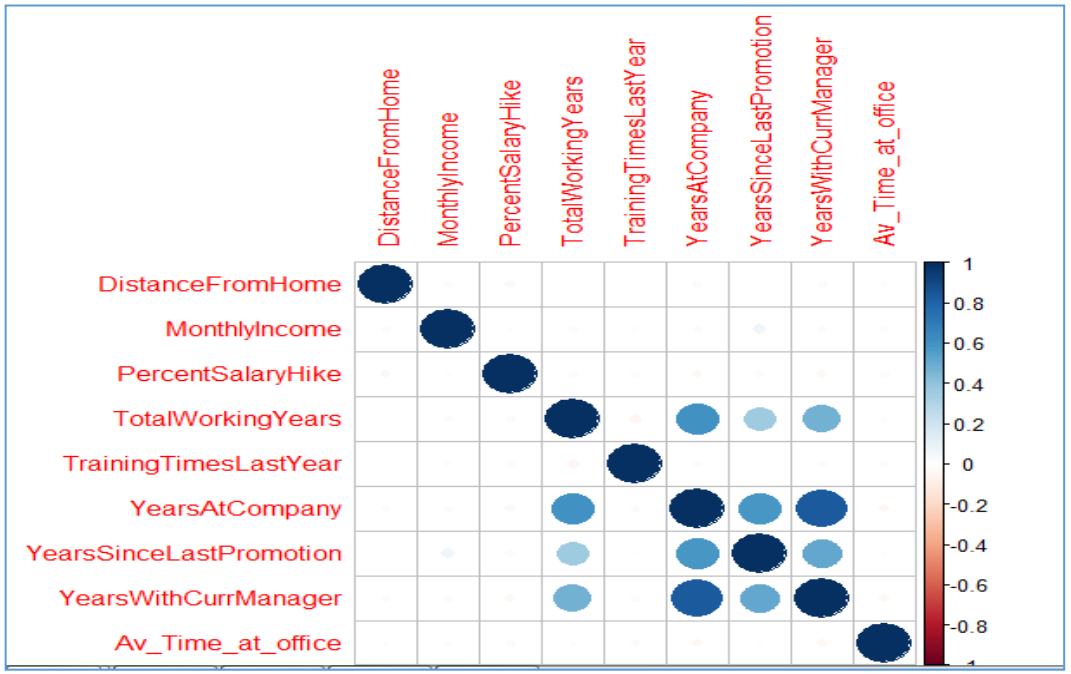
Here, we see that the variables like Total Working Years and Years at Company shows a significant difference for the 2 attrition levels. Based on the significance of each variable we will be taking appropriate actions during model building stage

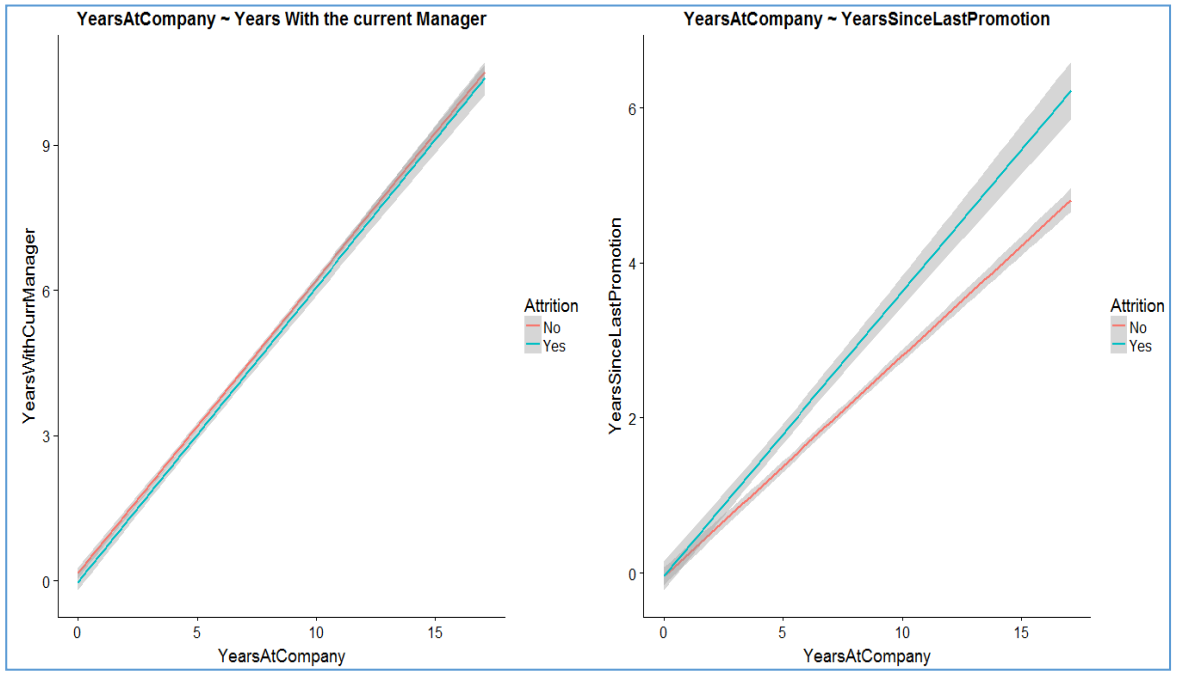
Analyzing the distribution of other continuous variables against Attrition and observing how they behave against different levels of attrition



Here, we see that the variables like Years with the Current Manager and Average Time at office shows a significant difference for the 2 attrition levels. Based on the significance of each variable we will be taking appropriate actions during model building stage

2. Correlation between the continuous variables - We observed some high correlation between the variables





Correlation Chart	
YearsAtCompany - YearsWithCurrManager	0.84
YearsAtCompany - YearsSinceLastPromotion	0.58

We will be taking actions on these correlated variables while checking Variable Inflation factor during model building stage

Model Building

1. Pre-Modelling Stage

- In this stage, we prepared the input variables which were to be given to the model
- We normalized the continuous variables. We scaled them using scale function in R
- We found the nature of the target variable to decide the model to be used –
  - Here our target variable is Attrition which is categorical in nature and hence we used logistic regression model
  - Attrition rate found in our dataset is 16% only
- We created dummies for categorical variable as part of variable reduction technique
- We split the data into train and test data in ratio 7:3

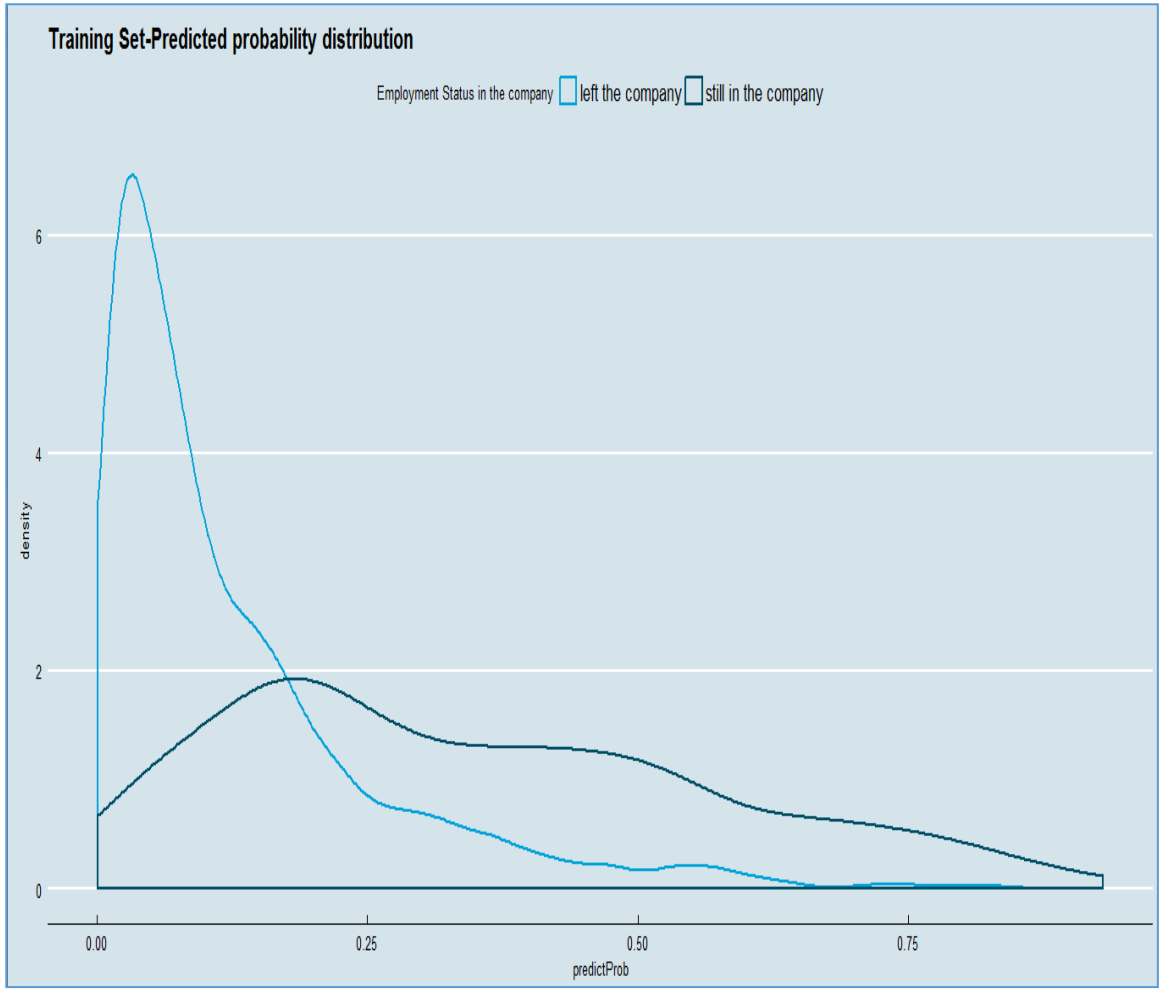
2. Modelling Stage

- We used generalized linear model to build our model to predict the attrition
- We used step AIC to get the starting model. Here we are predicting how Attrition behaves with respect to other variables. Since step AIC suggests the optimal model based on Akaike Information Criteria (model having lowest AIC value is preferred), there are some insignificant variables included in the model
- We used vif (Variable Inflation Factor) to check the correlation between the variables
- Finally we used p-values to remove insignificant variables

2. Modelling Stage [cont..]

Final Model

Predicted Variable	Attrition			
Coefficients:	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	0.63971	0.32783	1.951	0.051017
TotalWorkingYears	-0.55599	0.08071	-6.889	5.63E-12
YearsSinceLastPromotion	0.59261	0.07356	8.056	7.89E-16
YearsWithCurrManager	-0.50988	0.08785	-5.804	6.49E-09
Av_Time_at_office	0.68564	0.05459	12.559	<2.00E-16
BusinessTravel.xTravel_Frequently	0.71584	0.13419	5.335	9.57E-08
Department.xResearch...Development	-0.97034	0.23149	-4.192	2.77E-05
Department.xSales	-0.91566	0.24308	-3.767	0.000165
JobRole.xManufacturing.Director	-0.90704	0.21763	-4.168	3.07E-05
MaritalStatus.xSingle	0.92981	0.11677	7.963	1.68E-15
NumCompaniesWorked.x5	1.17446	0.23566	4.984	6.24E-07
NumCompaniesWorked.x7	0.95704	0.23452	4.081	4.49E-05
EnvironmentSatisfaction.x2	-0.62848	0.16984	-3.7	0.000215
EnvironmentSatisfaction.x3	-0.82446	0.15719	-5.245	1.56E-07
EnvironmentSatisfaction.x4	-0.97057	0.15916	-6.098	1.07E-09
JobSatisfaction.x2	-0.58016	0.17359	-3.342	0.000832
JobSatisfaction.x3	-0.51824	0.1513	-3.425	0.000614
JobSatisfaction.x4	-1.12396	0.16329	-6.883	5.85E-12
WorkLifeBalance.x2	-1.09395	0.22766	-4.805	1.55E-06
WorkLifeBalance.x3	-1.30052	0.2118	-6.14	8.23E-10
WorkLifeBalance.x4	-1.04621	0.26751	-3.911	9.19E-05
Age_categ.xYoung	0.73425	0.17646	4.161	3.17E-05

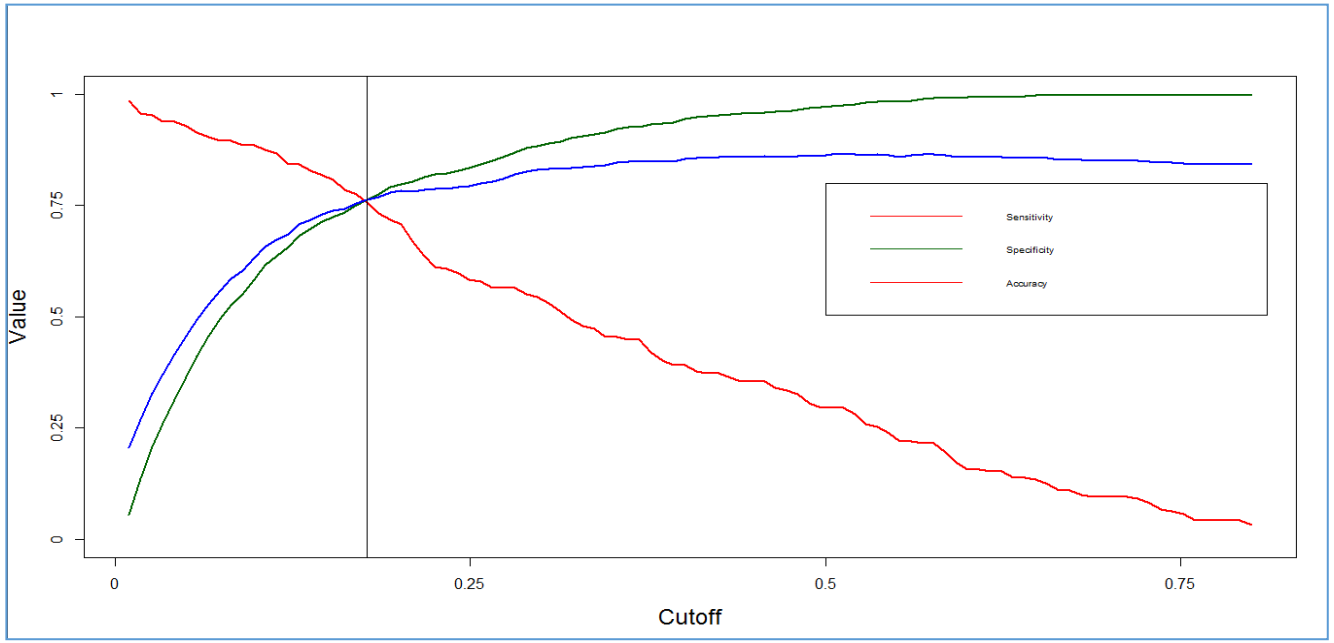


We checked the model against Train dataset to see how well it has predicted the known values. Here we see that there is no clear demarcation between the Yes and Nos. This is because our dataset had only 16% attrition rate. We will now run the model against the test dataset to see what probability of attrition we can correctly predict.

3. Post-Modelling Stage

- We ran the model against the test dataset
- We performed the following **model validations** -

1. **<u>Finding Accuracy, Specificity and Sensitivity through Confusion Matrix</u>**
   - In order to find a suitable probability cut-off, we checked the Accuracy, sensitivity and specificity for 1% to 80% probability values
   - The optimum cut-off probability is the one where the value of specificity and sensitivity are close to each other. Here we have taken a safe range of 0.01. Cut-off probability was found to be ~0.18



At Cut-off	0.18
Specificity	0.76
Sensitivity	0.75
Accuracy	0.76

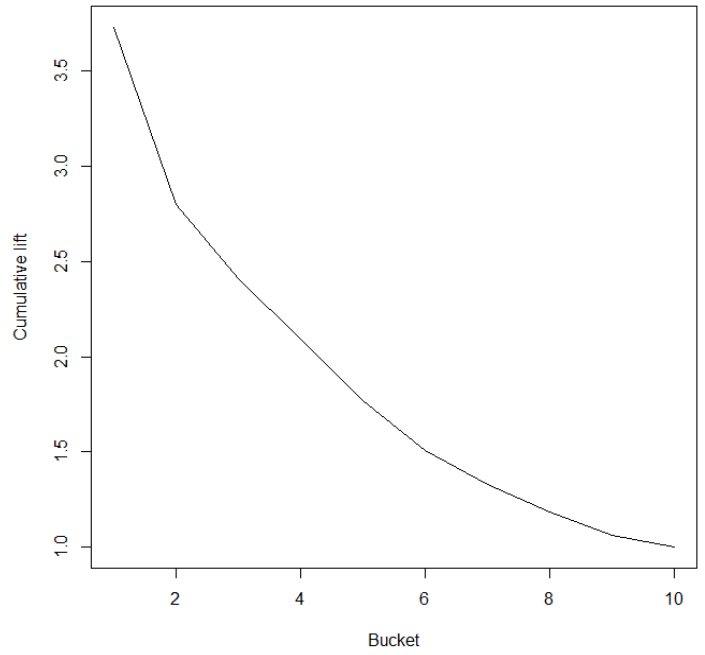
Our model can correctly identify 75% of time likelihood of the employees leaving the company and 76% of time when the employees are likely to stay in the company.
Our model can, in 76% of the case, predict the correct results

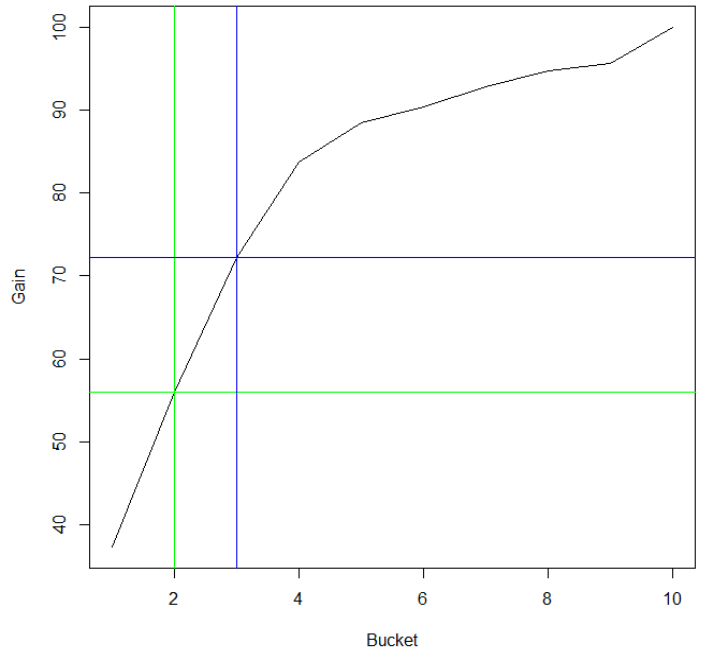
2. Calculating KS Statistics

- KS statistics measures the degree of separation between positive and negative distribution
- The optimal value of KS statistics for a good model should lie between 40-60 and should be within first 3 deciles
- KS-Statistics for our model is 0.513

3. Gain and Lift Charts

- It helps to measure the effectiveness of the model by calculating the percentage of events captured in each decile





Here, we can see that even by targeting top 20% of the employees, we can predict approx 55% of time if they are going to leave and if you go to even 30%, this number can go as high as 72%

Conclusion

Variables having positive impact on attrition		
Factors	**Inferences**	**Suggestions**
Number of companies worked	If an employee has **switched companies 5 to 7 times**, they are more likely to leave the current company	Company should focus more on the individual's aspirations and check if they are inline with the companies aspirations before hiring such employees. They should also ask for a reason for switch from these employees to prevent attrition of them in the current company
Marital Status	Those who are **single** are more likely to leave the company	
Age	The employees having **age below 25** are more likely to leave	Company should provide suitable opportunities to these employees and reward the critical ones to retain them for long
Business Travel	Those who **travel frequently** are likely to quit the company. This may be because they do not like to travel and they are forced to or they have found better opportunities in other places	Company should keep a continuous check on employee's aspirations here and see if they are happy with travelling. They should also keep a tap on what other opportunities employee have in other places and keep pace with the changing workforce dynamics
Average time spent at office each day	Those who spend **more time** at office each day are more likely to quit. The reason could be work pressure and frustration	Company should ensure that work in the team are divided uniformly and there should not be pressure of a single employee. Knowledge sharing should be increased within the team so that there is no pressure on team members during absence of others

Conclusion

Variables having negative impact on attrition

Factors	Inferences	Suggestions
Work Life Balance	Employees having a **good , better and best work life balance** are likely to stay	Company should ensure that the employee enjoys personal and family time apart from work. They can roll out some holiday offers to the employees once in a year or two. They should be a bit lenient on leaves given to the employees
Job Satisfaction	Those who are having a **medium, high and very high job satisfaction** are likely to stay with company	Company should have a mechanism in place where the employees can share the feedback of the work they are doing. Based on this they should work on the negative ones and promote the positive ones
Environment Satisfaction	Those with **medium, high and very high satisfaction** are likely to continue long with the company	Company should ensure conducive atmosphere to work where the employee are motivated. Employees should be given rewards in terms of money or coupons from time to time. Company should also ensure that the workspace is clean and safe for the employees to enjoy working
Department	Those **working on research and development and sales** are more likely to stay	Company should ensure that whenever there are vacancies in these department, they should check on movement of employees internally as opposed to hiring someone outside the company
Job Role	Those who are **Manufacturing directors** are more likely to stay	
Total working years	**More is the working year**, lesser are the chances to quit the company	Company should look to utilize the experience of such professionals
Years with current Manager	The more the employee **spends time on working with the current manager**, more likely are they to stay	Company should ensure a healthy relation between the employees and the manager. The manager should easily be approachable and should be considerate on the employees goals. This will help the company retain the employees