HR ANALYTICS
Case Study - Logistic Regression

PRESENTATION

Group Members Name:

1. Thanigachalam TamizhSelvam
2. Swati Bansal
3. Rajanish Mirajkar

Problem Statement

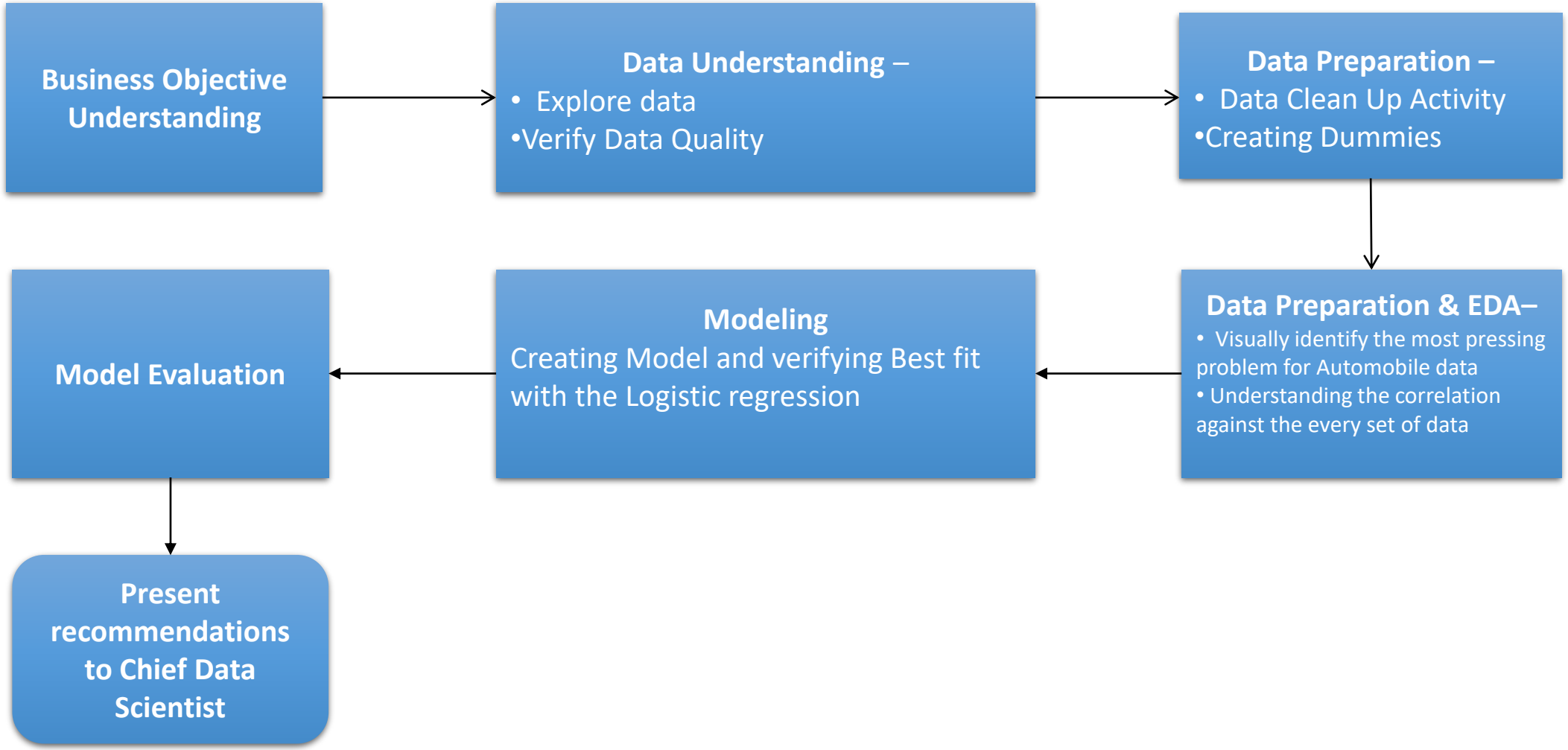
A large company named XYZ, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition(employees leaving, either on their own or because they got fired) is bad for the company, because of the following reasons –

1. The former employees' projects get delayed, which makes it difficult to meet **timelines**, resulting in a reputation loss among consumers and partners
2. A sizeable department has to be maintained, for the purposes of **recruiting** new talent
3. More often than not, the new employees have to be **trained** for the job and/or given time to acclimatize themselves to the company

Business Objective

Focus on employee retention by identifying the transformation required to drastically reduce the employee attrition rate. Spot the key parameters that contribute to it and make immediate and necessary changes to XYZ workplace and encourage their employees to continue their association with the company.

Identifying the attrition causing key parameters



Data Clean Up Activity

Performed the following data clean up :-

Source File : ‘in_time.csv’ and ‘out_time.csv’, ‘general_data.csv’

1. Converted all the character columns to upper case for avoiding any case issue.
2. Converted categorical columns has a vector type
3. In the source file, in-time and out-time datetime are in separate file we will be calculating in-time in hours by for all the employees
4. Created summary of the in-time by employee and driving following columns Total.logged.hours, Average.logged.hours, Total.leaves.taken, Total.excess.logged, Average.excess.logged
5. Validated for NA values, and found that 5 to 6 percentage of data is NA, considered this for calculating leaves taken by employee
6. Formated date column to ensure all the column date values in same formate (in_time, out_time)
7. created calculated column worked hour (intime - out time)
8. Below listed column has been removed from the analysis for in and out time since these columns only contain NA values.

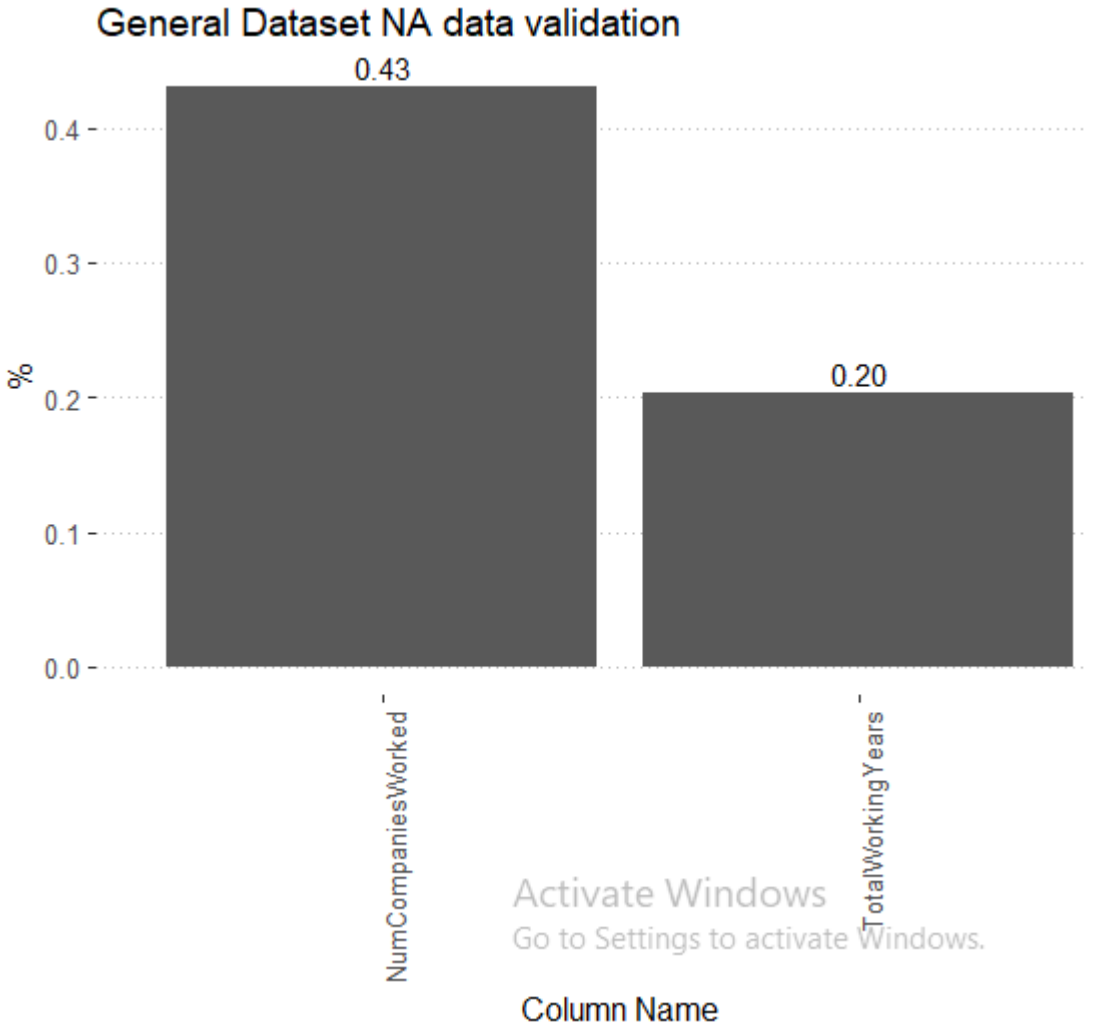
   "X2015.01.01" "X2015.01.14" "X2015.01.26" "X2015.03.05" "X2015.05.01" "X2015.07.17“
   "X2015.09.17" "X2015.10.02" "X2015.11.09" "X2015.11.10" "X2015.11.11" "X2015.12.25“

Assumptions Used For Modeling

The below listed assumptions are used to drive the inferences on NA values.

Source File :‘general_data.csv’

1. NA records of the general data set is having less than 0.43 percentage. This is a negligible value and so we are ignoring this data for analysis.

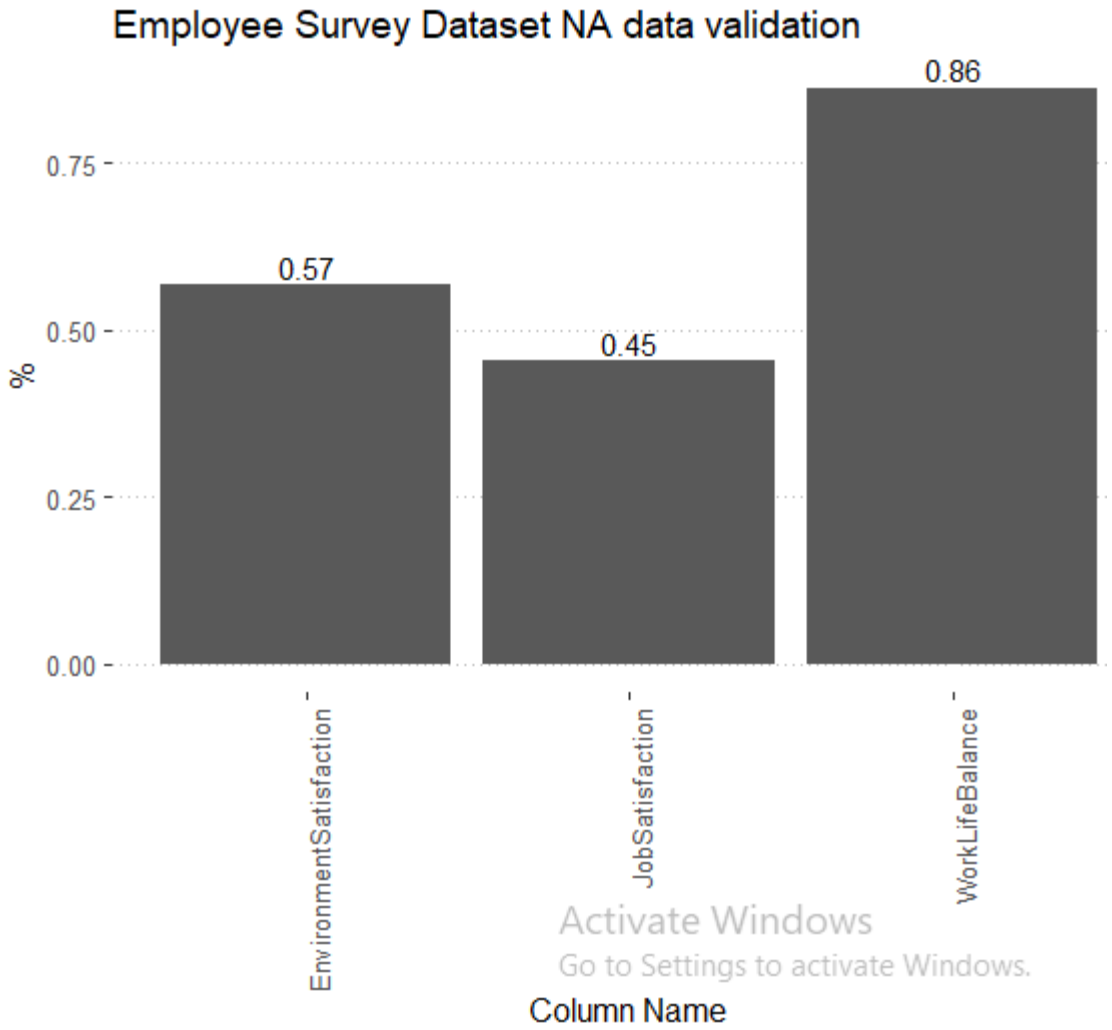


Assumptions Used For Modeling

The below listed assumptions are used to drive the inferences on NA values.

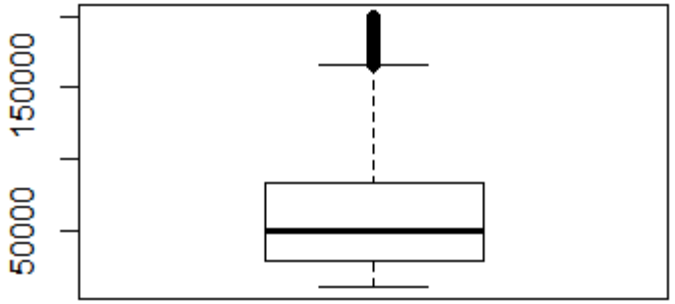
Source File :‘employee_survey_data.csv’

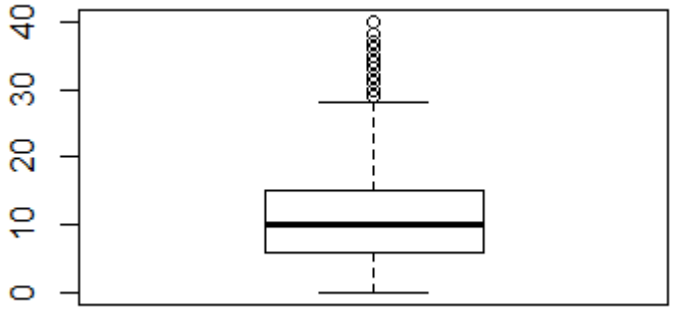
1. NA records in the employee survey data set is having less than 1 percentage, So we are ignoring this data for analysis.



Outlier Treatment

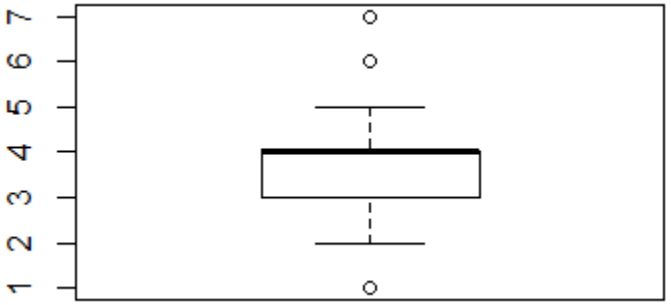
1. **Monthly Income**
    1. There are outliers on the upper end in Monthly Income
    2. modifying all Monthly incomes above 96% to 186437.6





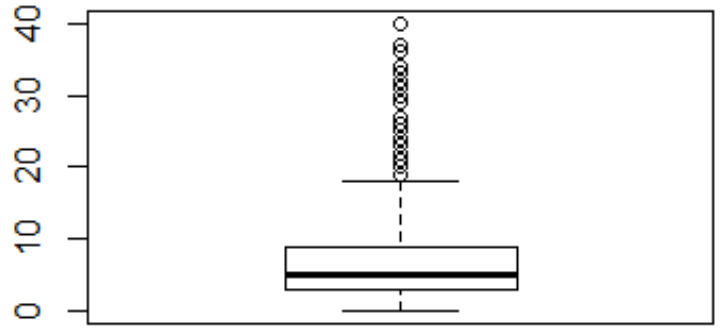
2. **Total Working Years**
    1. There are Outliers on the Higher Side
    2. modifying all Total Working Years above 98% to 32

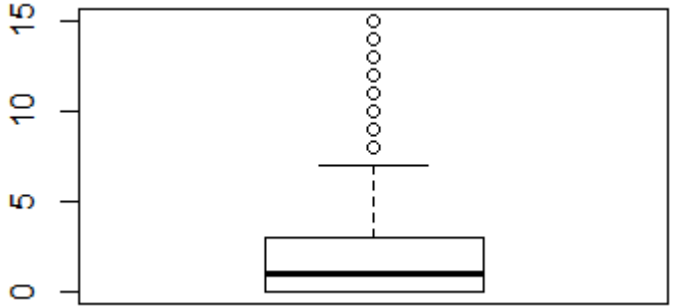
3. **Training Times Last Year**
    1. There are outliers on both upper and lower ends of the boxplot
    2. we have not found any high range so we will be avoid data clumping gut its very low so we can ignore this.



Outlier Treatment

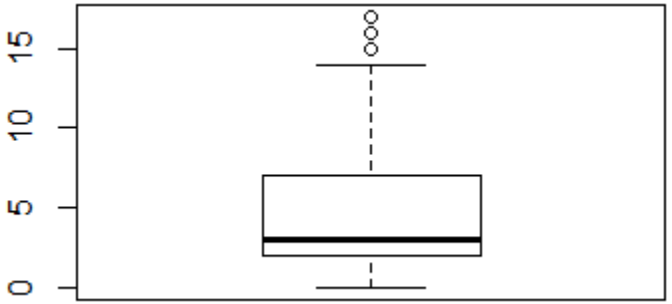
4. **Monthly Income**
    1. There are significant outliers in the higher values of Years At Company
    2. Outlier Treatment: We will cap Years At 98% will be updating to 24 for above 98%





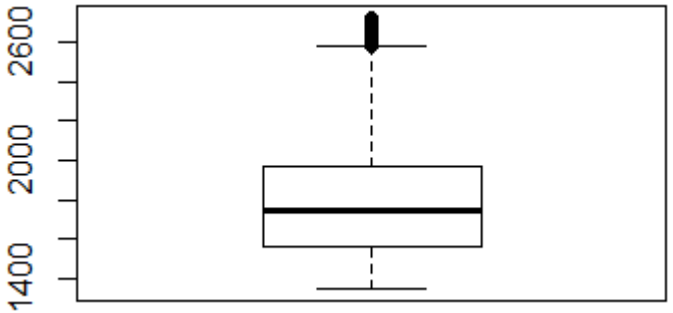
5. **Years Since Last Promotion**
    1. There are significant outliers on the upper values of Years Since Last Promotion
    2. Modifying all Years Since Last Promotion after 97% updating to 11

6. **Years With CurrManager**
    1. There are outliers on the higher values of Years With Curr Manager
    2. Updating all Years WithCurrManager value above 99% update to 14



Outlier Treatment

7. **Total.logged.hours**
    1. There are significant outliers in the upper region of Tot.logged.hours
    2. Updating all records for 98%>2589



Assumptions Used For Modeling

The below listed assumptions are used to drive the inferences and Outlier Treatments :-

1. Considered NA values between in-time and out-time as “leave” taken by the employees.
2. Ignored the columns where all the rows are NA.

Created New Derive / Dummy variables

The following new derive variables

1. Age_range, MonthlyIncome_Range, TotalWorkingYears_range, PercentSalaryHike_range, YearsAtCompany_range, YearsSinceLastPromotion_range, YearsWithCurrManager_range, Education, EnvironmentSatisfaction, JobInvolvement, JobSatisfaction, PerformanceRating, WorkLifeBalance,
2. Removed the following columns that are not considered for analysis.
    - EmployeeCount where all the column value is 1
    - Over18 all the values in the column are Y
    - StandardHouras all the values in the column are 8
3. Created Dummy variables for following columns.

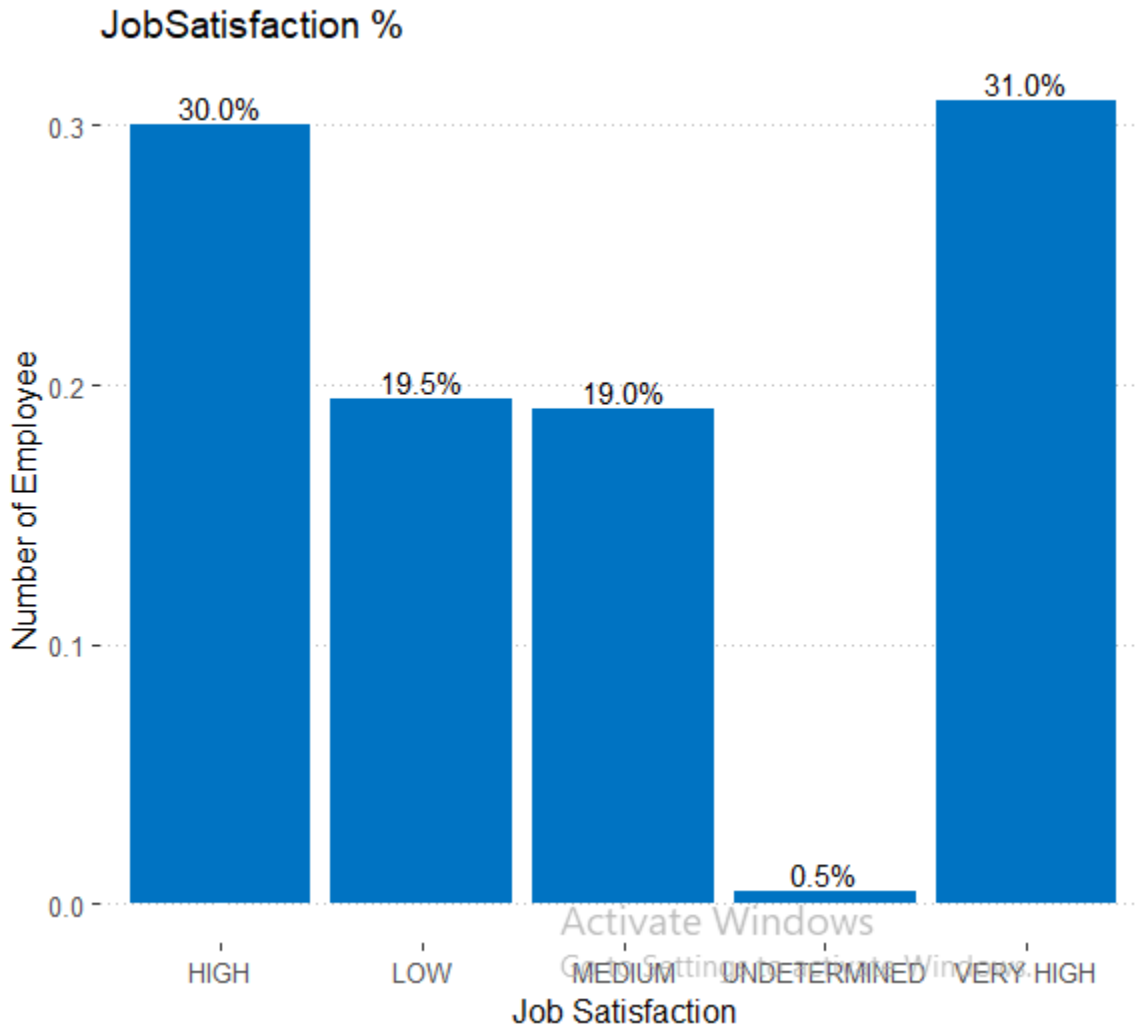
    Fueltype, aspiration, doornumber, carbody, drivewheel, enginelocation, enginetype, cylindernumber, fuelsystem, symbolling, company_name

EDA Analysis

Analysis on Job Satisfaction data

The graph depicts the percentage of employee satisfaction ratio as

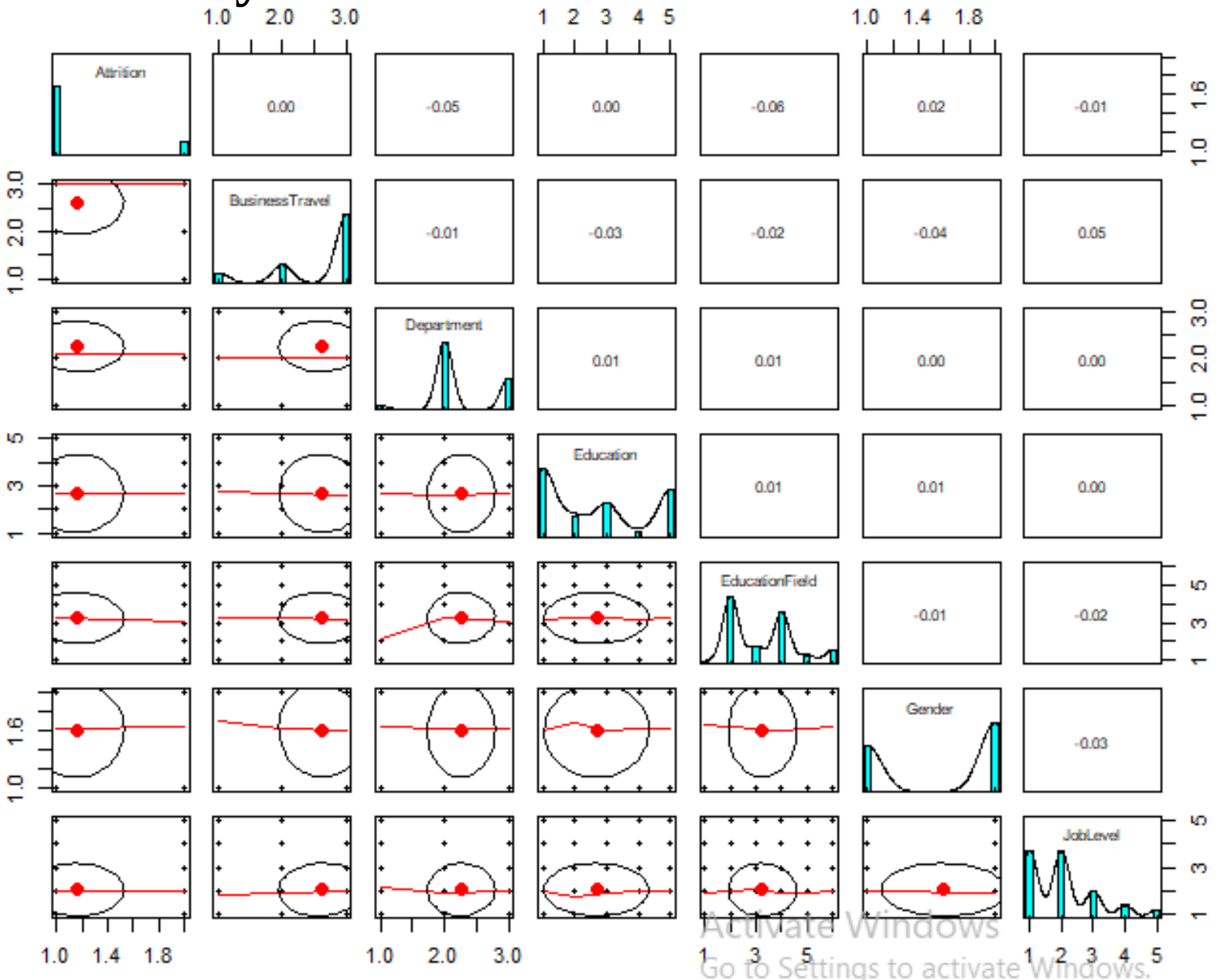
- 31% employees with **Very High** satisfaction.
- 30% employees with **High** satisfaction.
- 19% employee with **Medium** satisfaction.
- 19.5% employee with **Low** satisfaction.
- 0.5% employee are **Undetermined** satisfaction.



EDA Analysis

Correlation between Attrition, Business Travel, Department, Education, Education Field, Gender and Job Level.

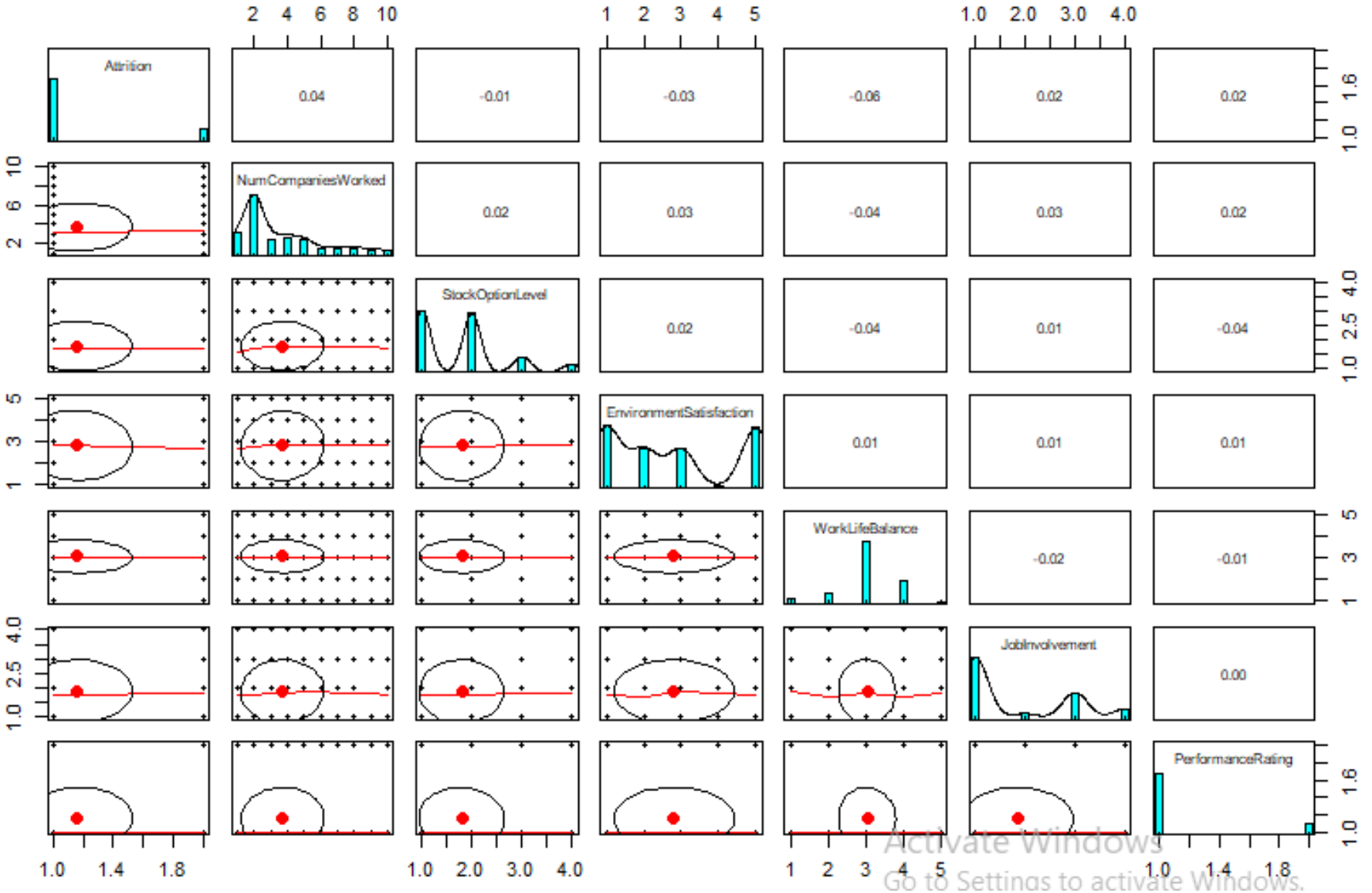
We are not able to find any **significance inference** on this.



EDA Analysis

Analysis on finding **correlation** between following variables

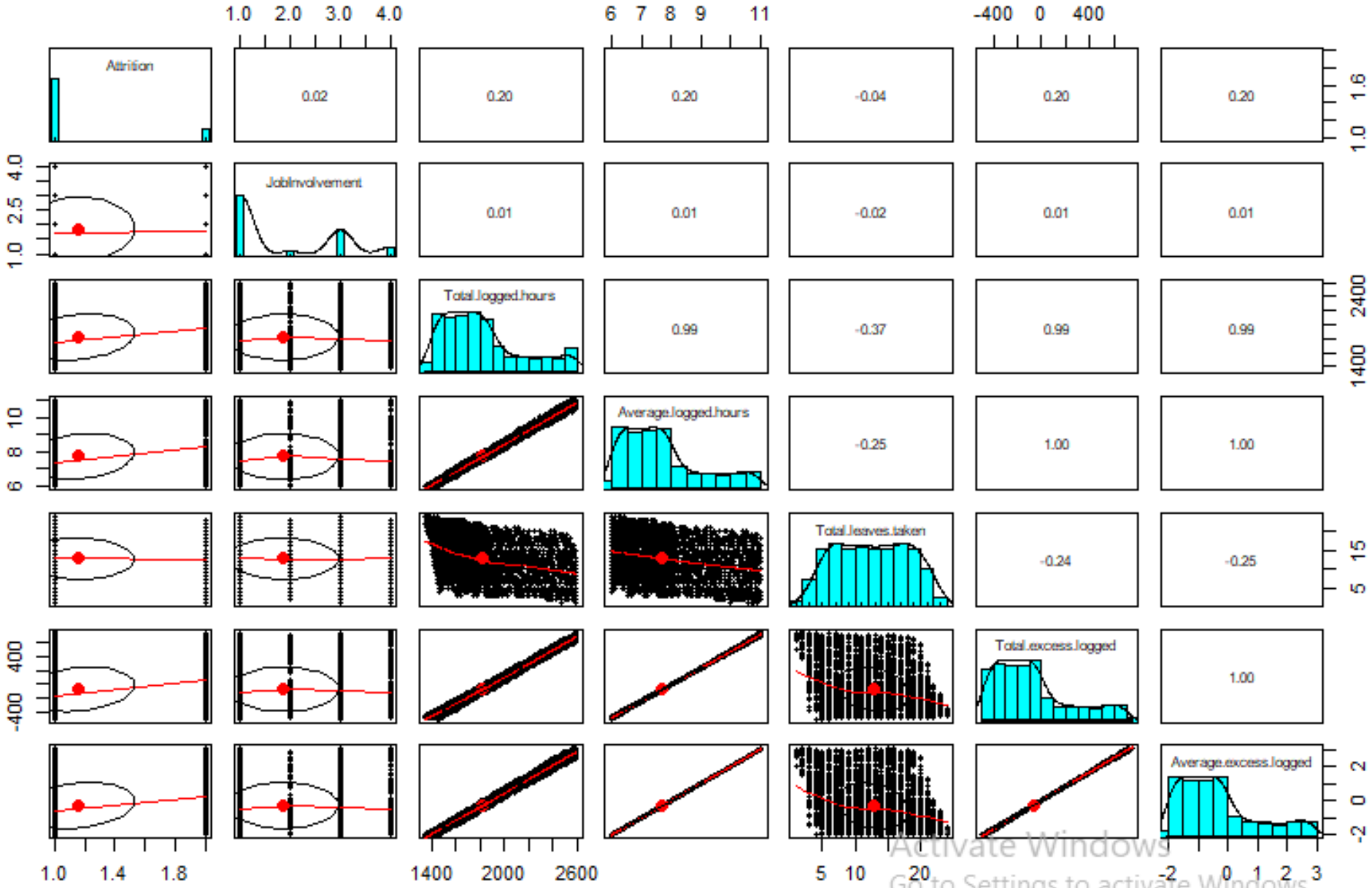
- Attrition
- NumCompaniesWorked
- StockOptionLevel
- Environment Satisfaction
- WorkLifeBalance
- JobInvolvement
- Performance Rating



EDA Analysis

Analysis on finding correlation between following variables

- Attrition
- JobInvolvement
- Total.logged.hours
- Average.logged.hours
- Total.leaves.taken
- Total.excess.logged
- Average.excess.logged

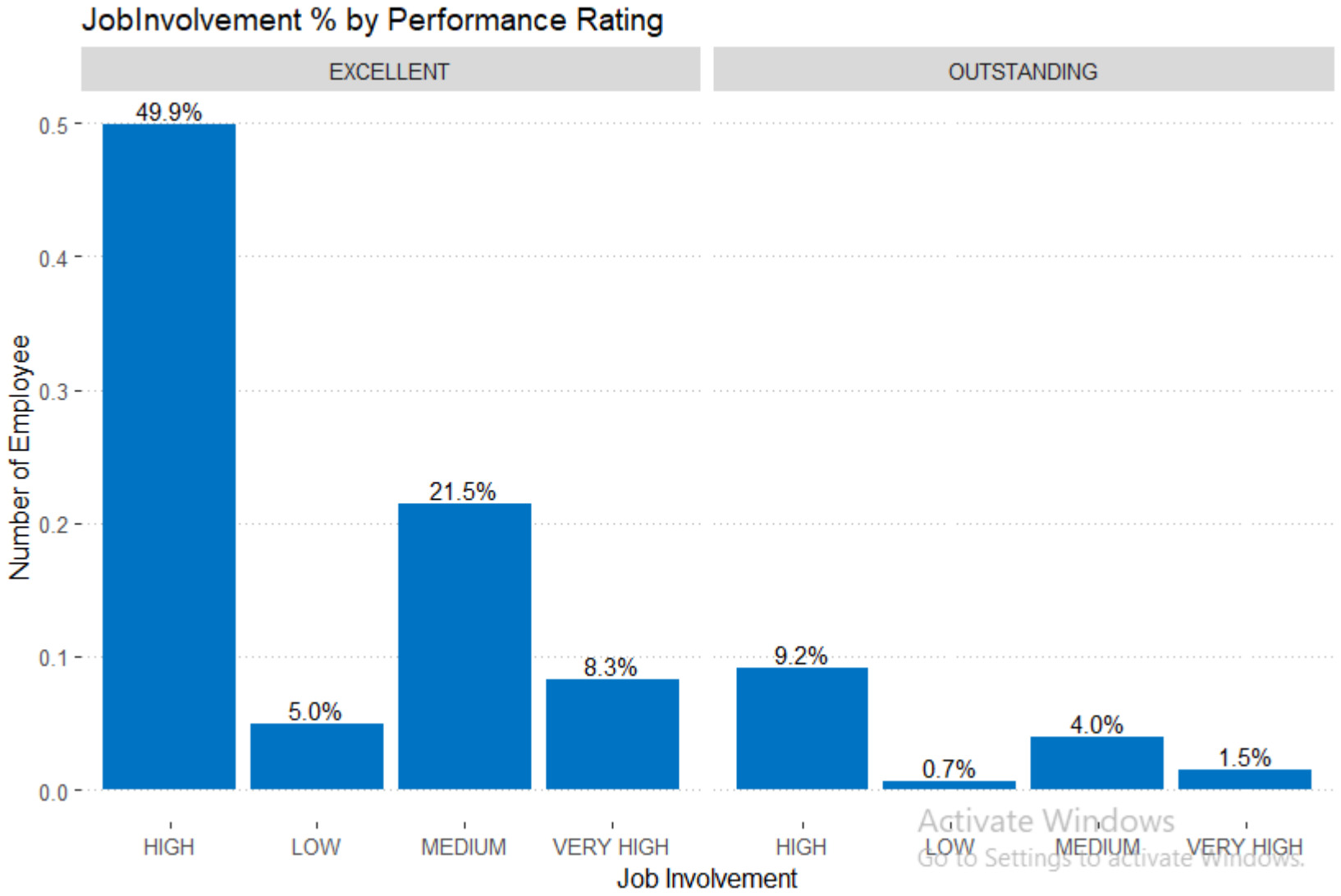


EDA Analysis

The graph depicts JobInvolvement percentage by Performance Rating

Excellent
- 8.3% Very High
- 49.9% High
- 21.5% Medium
- 5% Low

Outstanding
- 1.5% Very High
- 9.2% High
- 4% Medium
- 0.7% Low

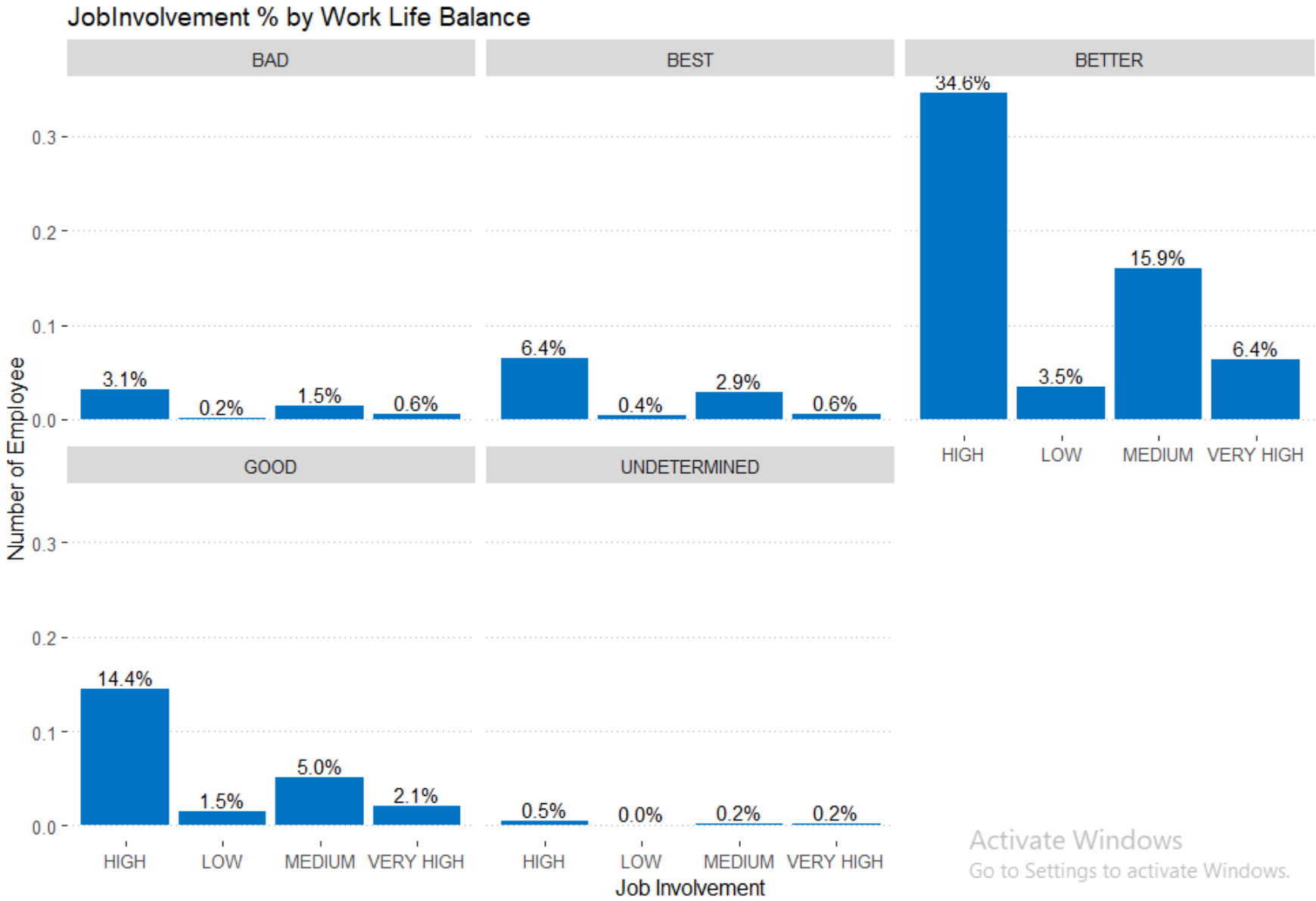


EDA Analysis

JobInvolvement percentage by Work Life Balance

With categories

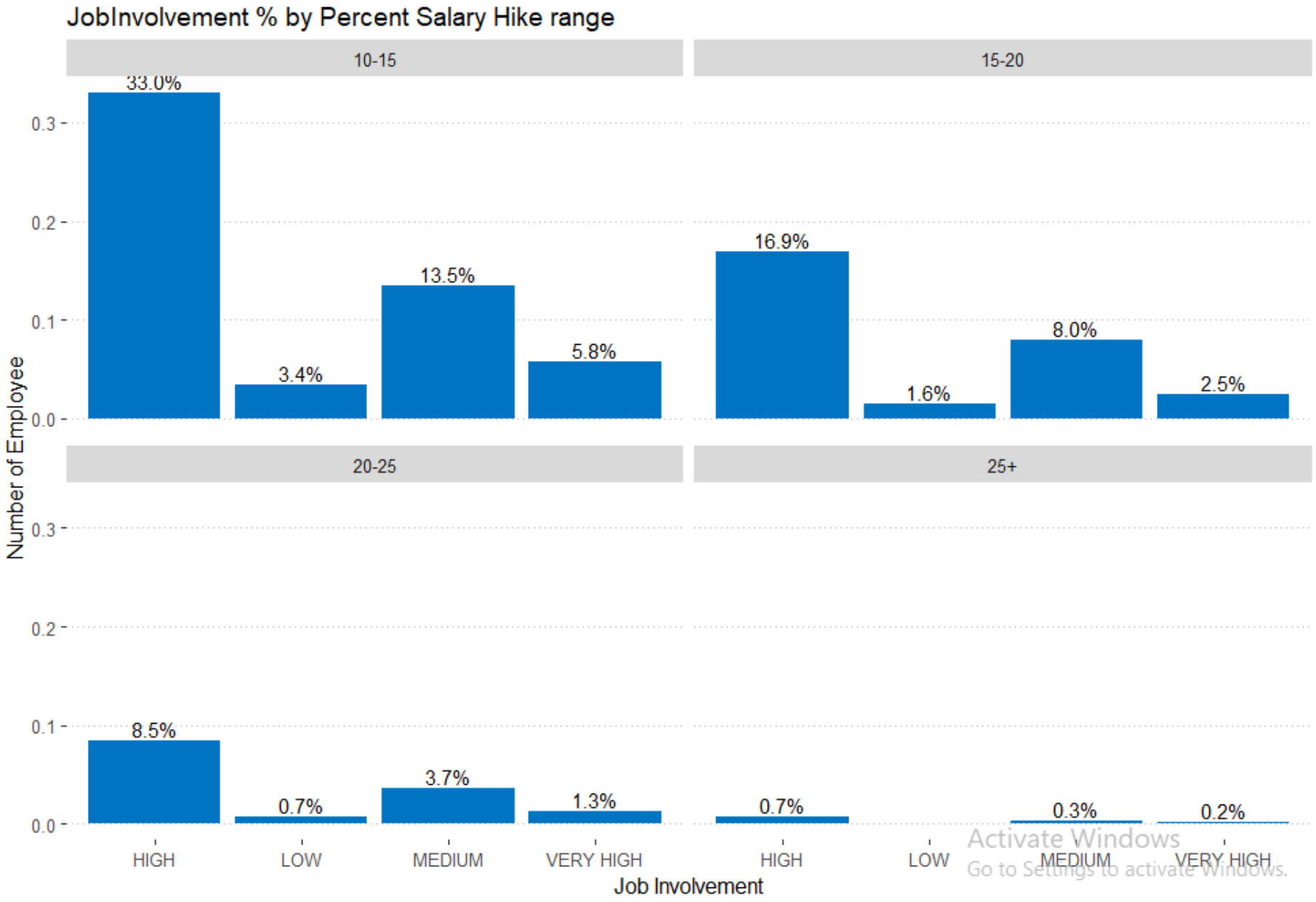
- Bad
- Good
- Better
- Best
- Undetermined



EDA Analysis

JobInvolvement % by Percent Salary Hike range

- With 10-15 % hike Job involvement shows good
- With 25% hike Job involvement shows poor

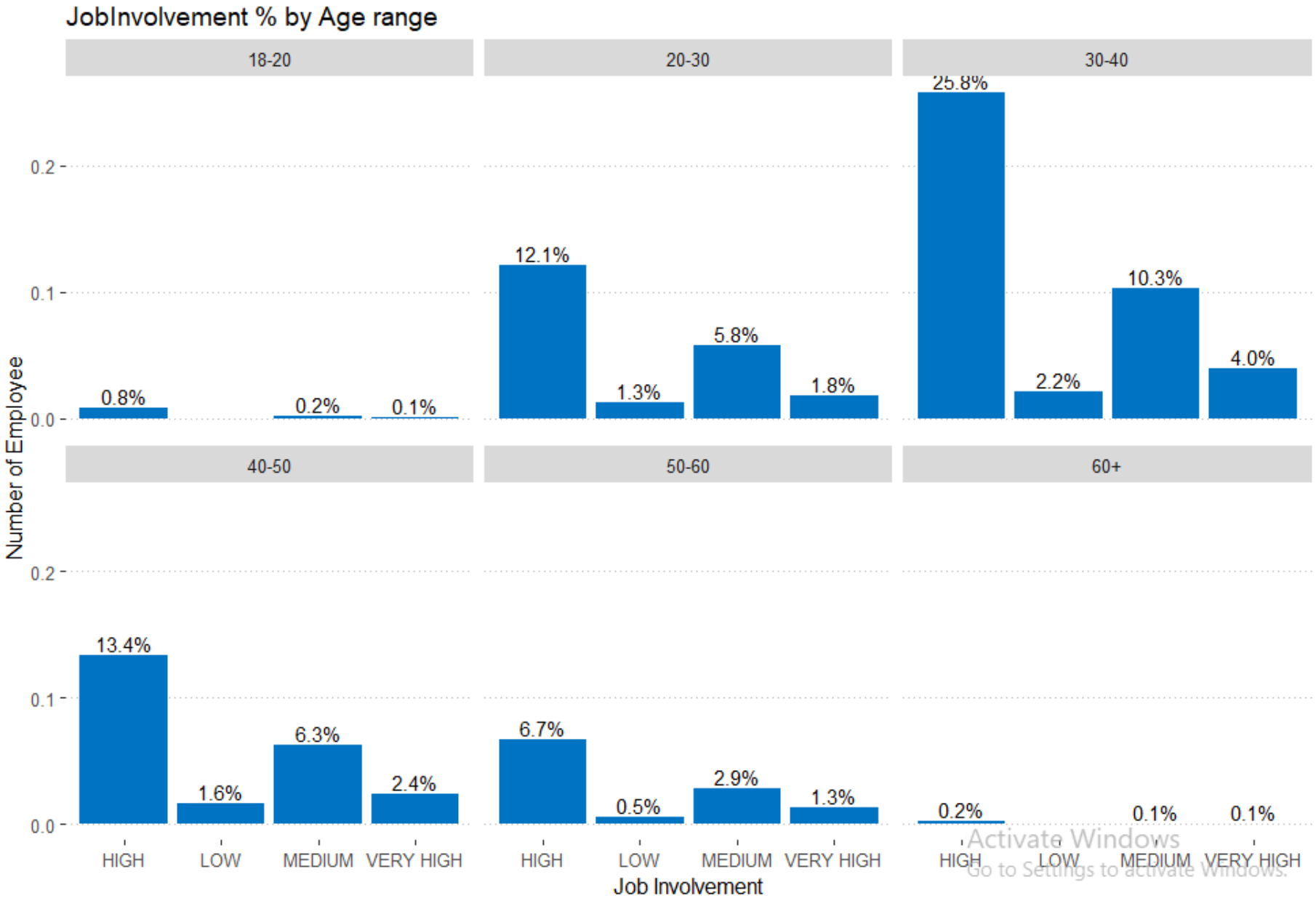


EDA Analysis

JobInvolvement % by Age range

With 30-40 age group Job involvement shows good Where as

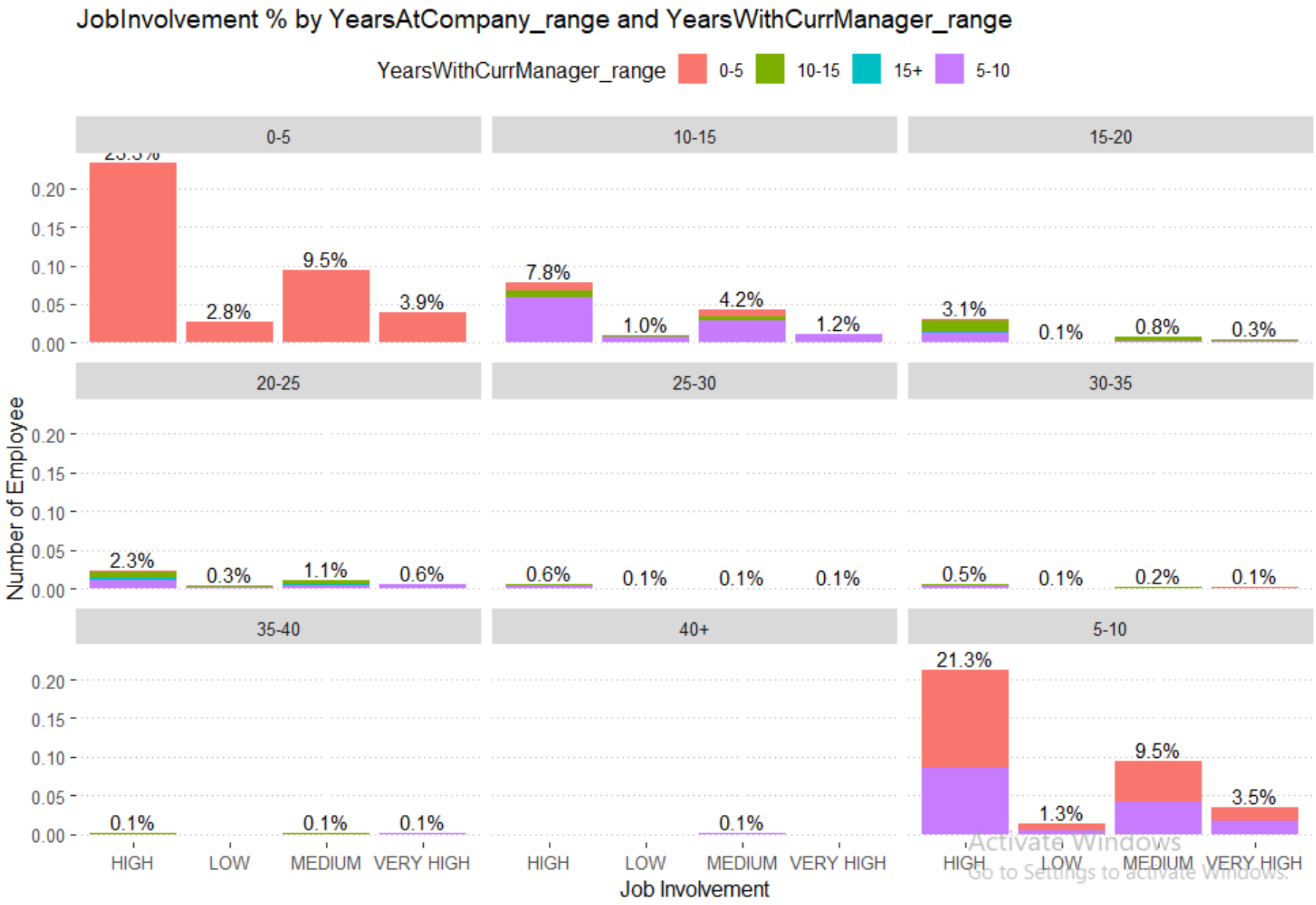
With 18-20 and 60+ age group shows poor.



EDA Analysis

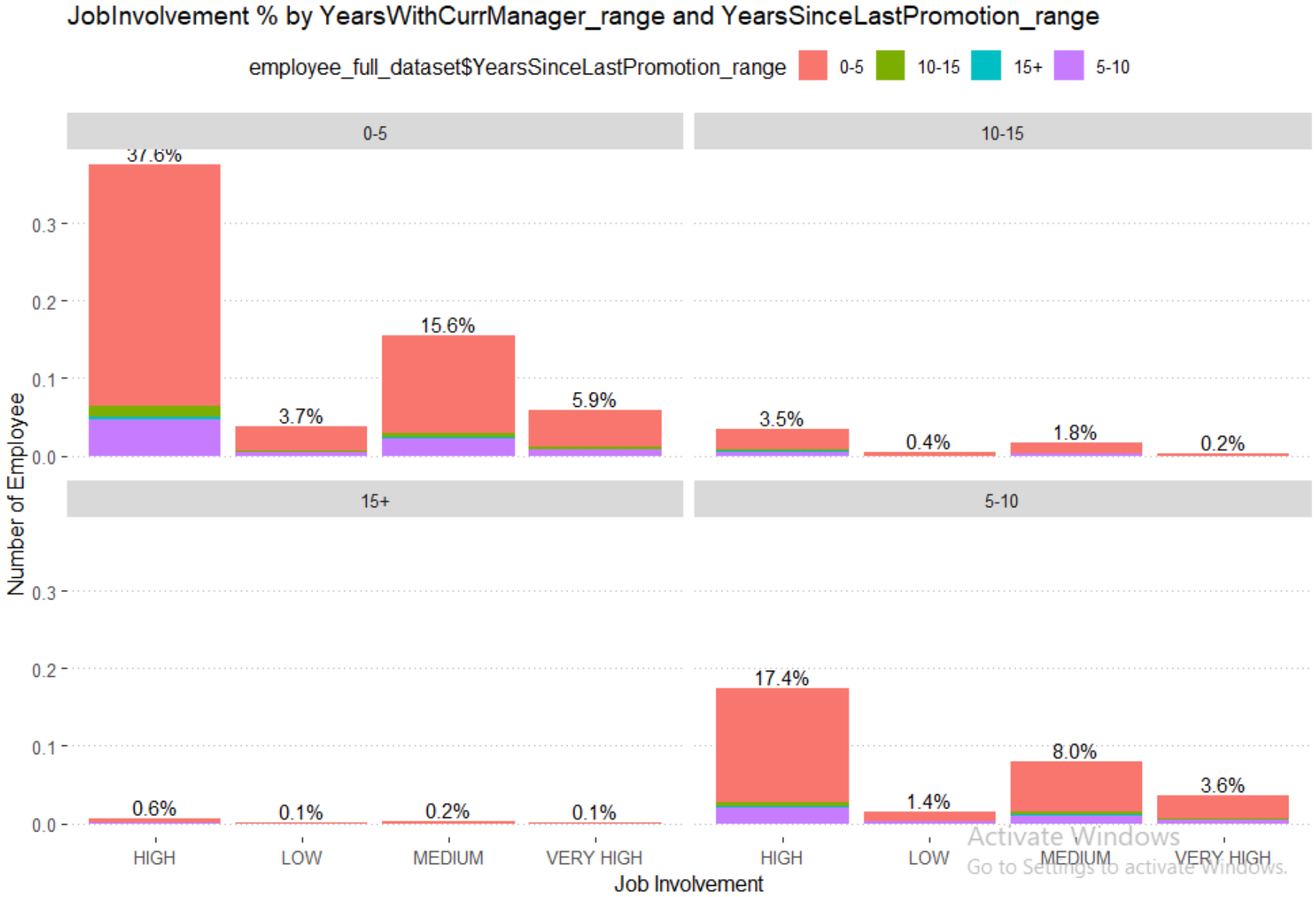
JobInvolvement % by YearsAtCompany_range and YearsWithCurrManager_range

- With 0-5 Years at company Job shows good involvement
- With 15-20 Years at company Job shows poor involvement



EDA Analysis

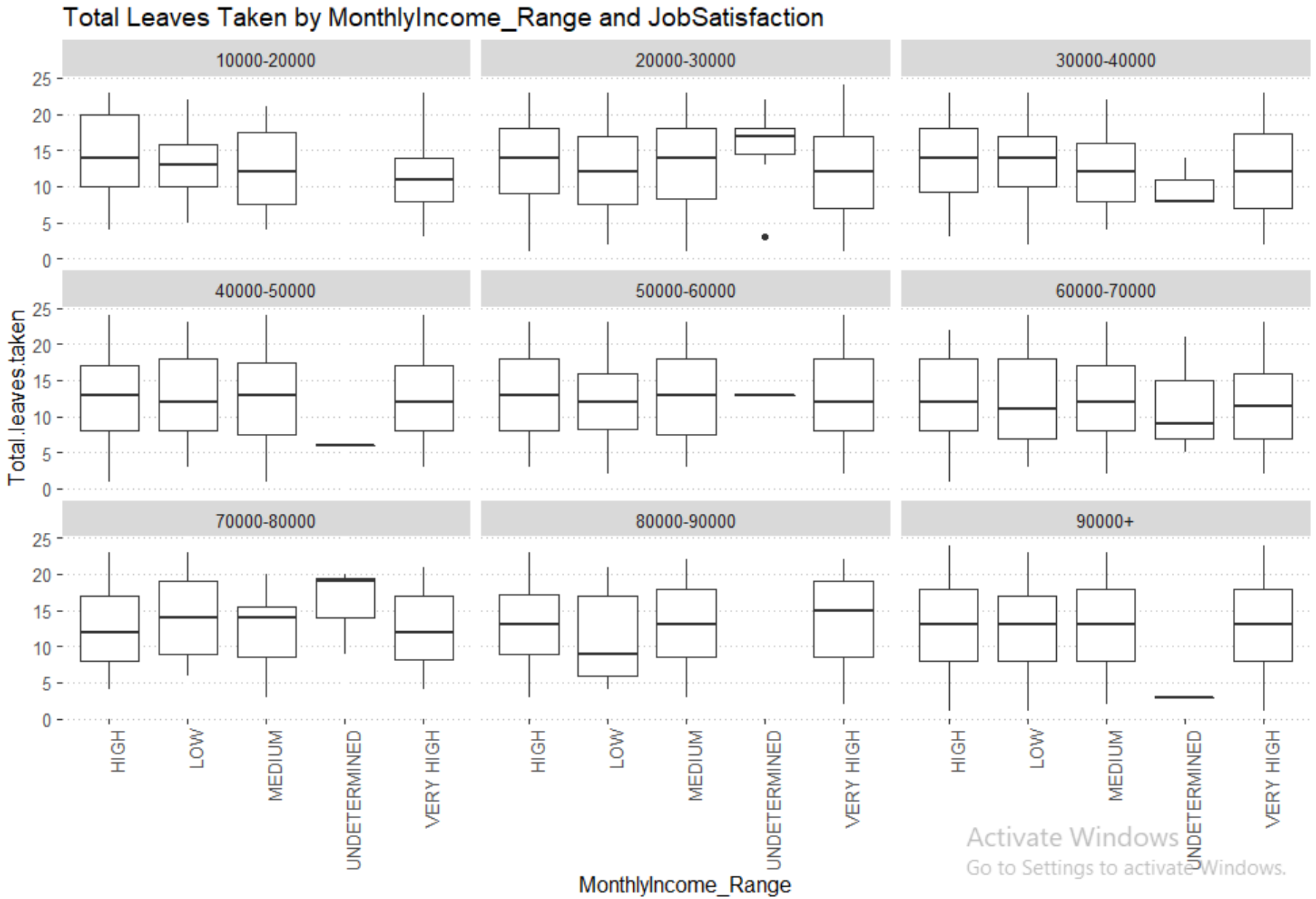
The graph shows JobInvolvement percentage by YearsWith CurrManager_range and YearsSince LastPromotion_range



EDA Analysis

Total Leaves Taken by MonthlyIncome

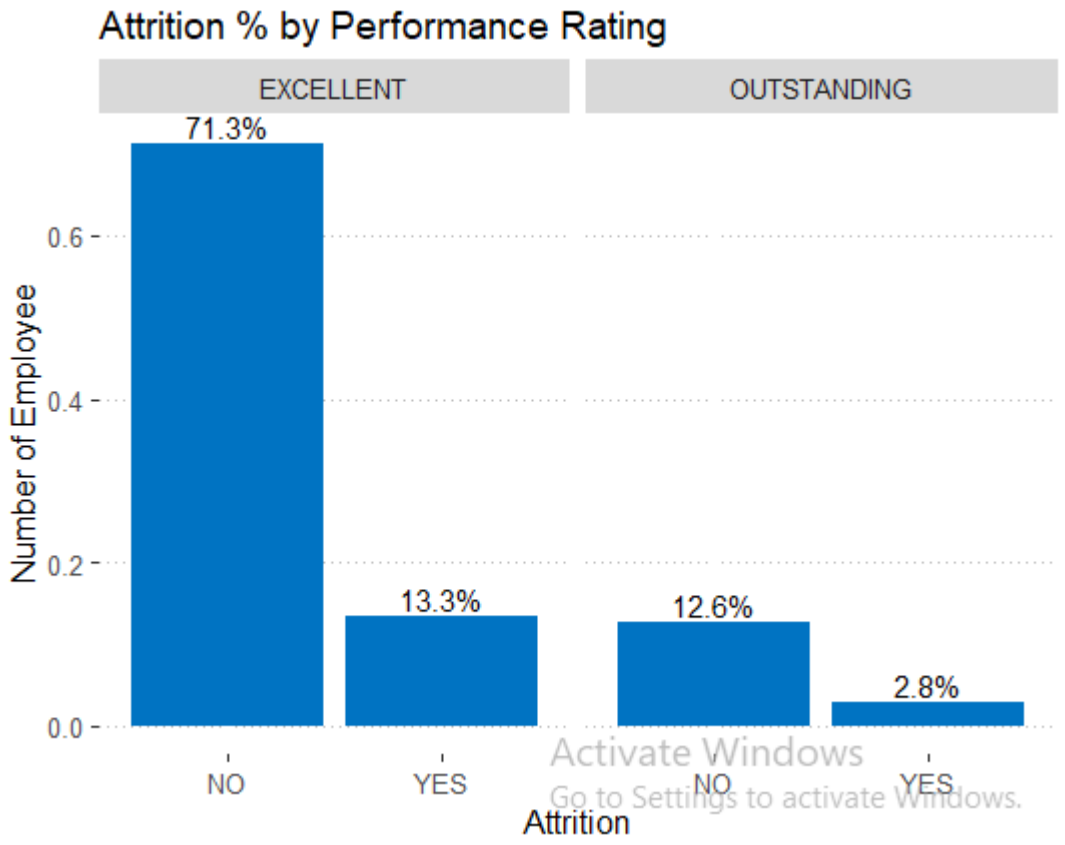
_Range and JobSatisfaction Distribution



EDA Analysis

Attrition percentage by Performance Rating

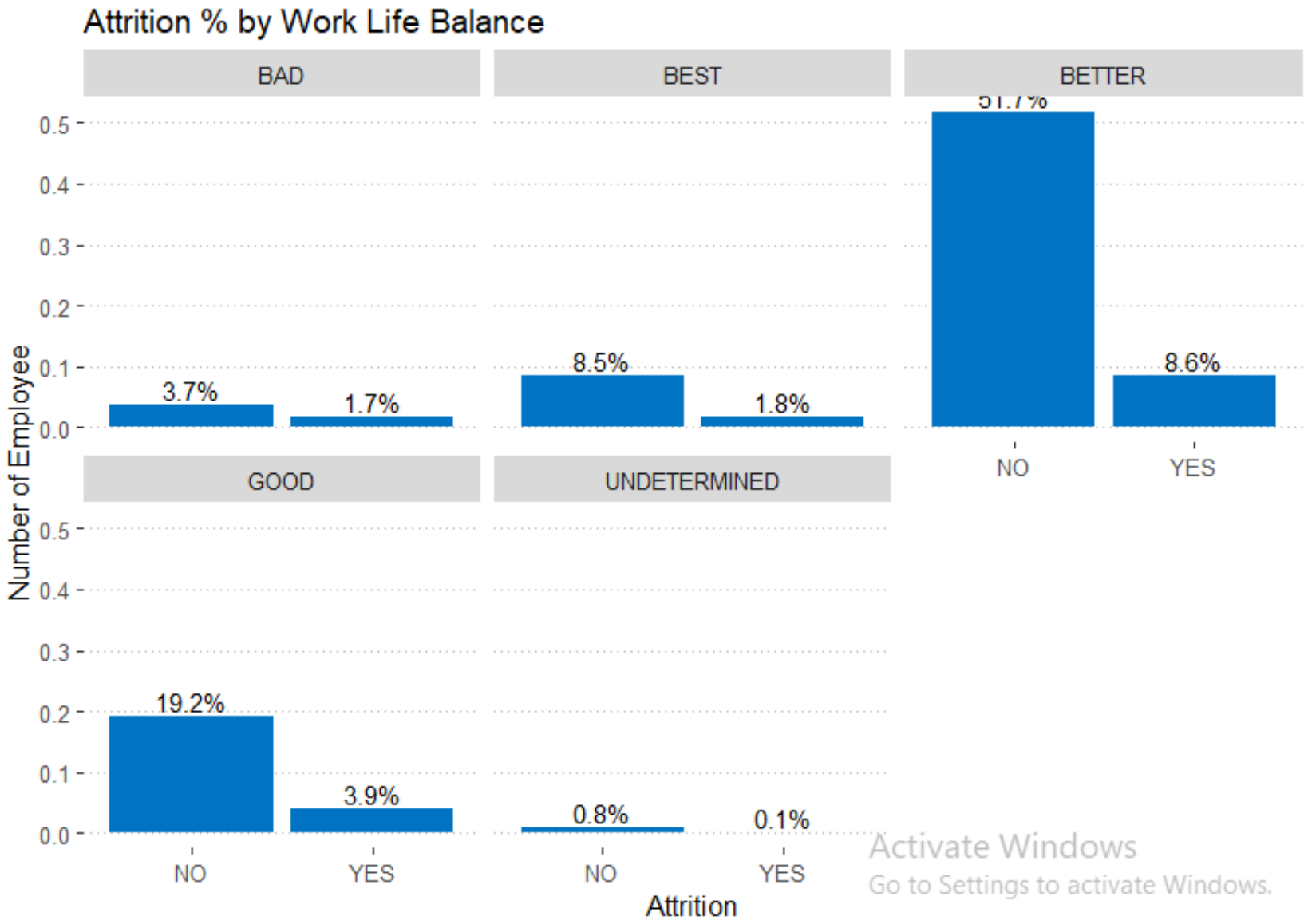
- With excellent performance rating shows 13.3% attrition
- With Outstanding performance rating shows 2.8% attrition



EDA Analysis

Attrition % by Work Life Balance

Group with BAD work life balance shows attrition rate high



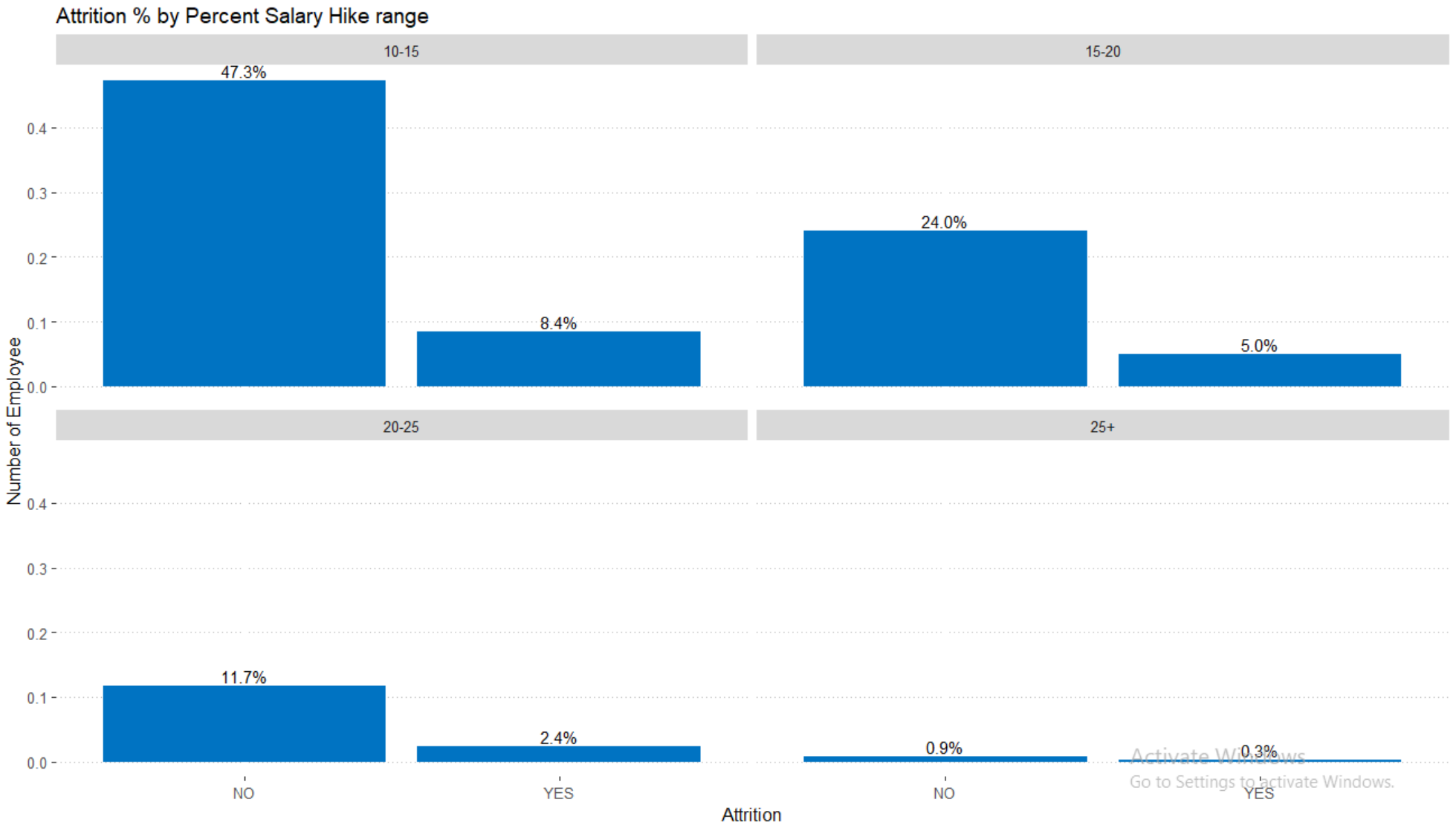
EDA Analysis

Attrition % by Percent Salary Hike range

With salary hike 10-15% shows attrition rate is high

And

With salary hike 25% shows attrition rate is low



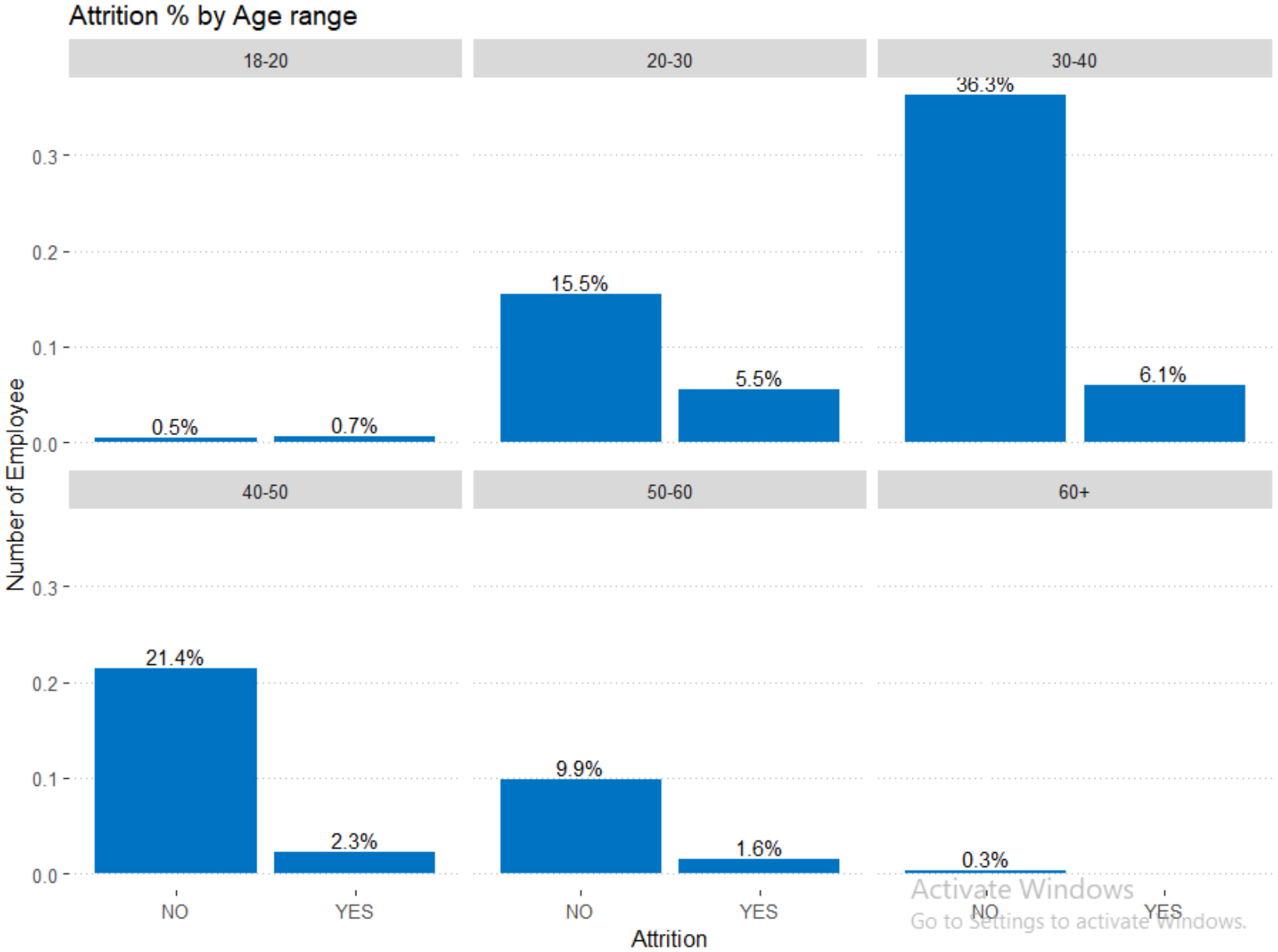
EDA Analysis

Attrition % by Age_range

Age group 20-30 shows attrition rate high

And

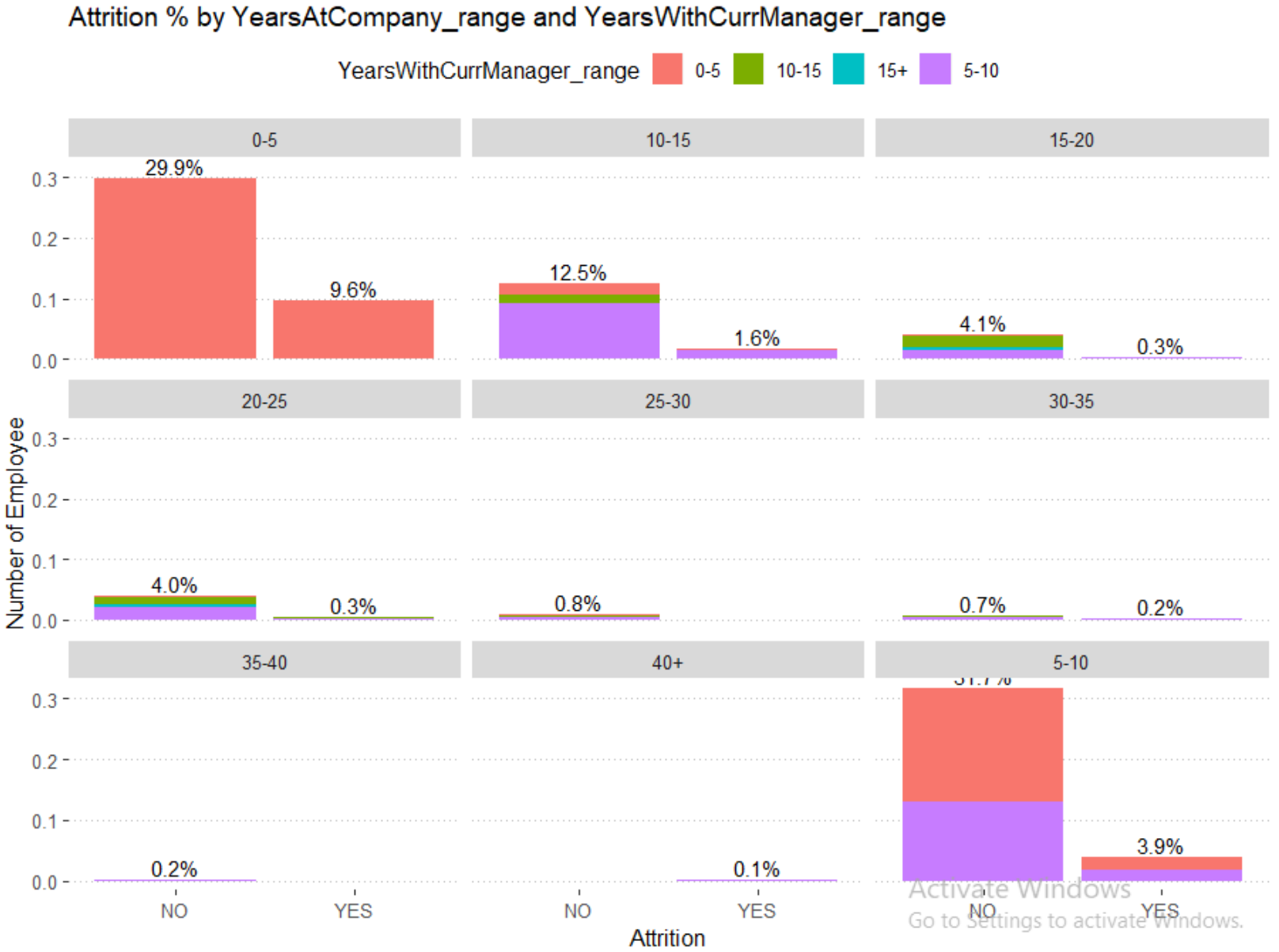
With Age group 60+ shows attrition rate



EDA Analysis

Attrition % by YearsAtCompany_range and YearsWithCurrManager_range

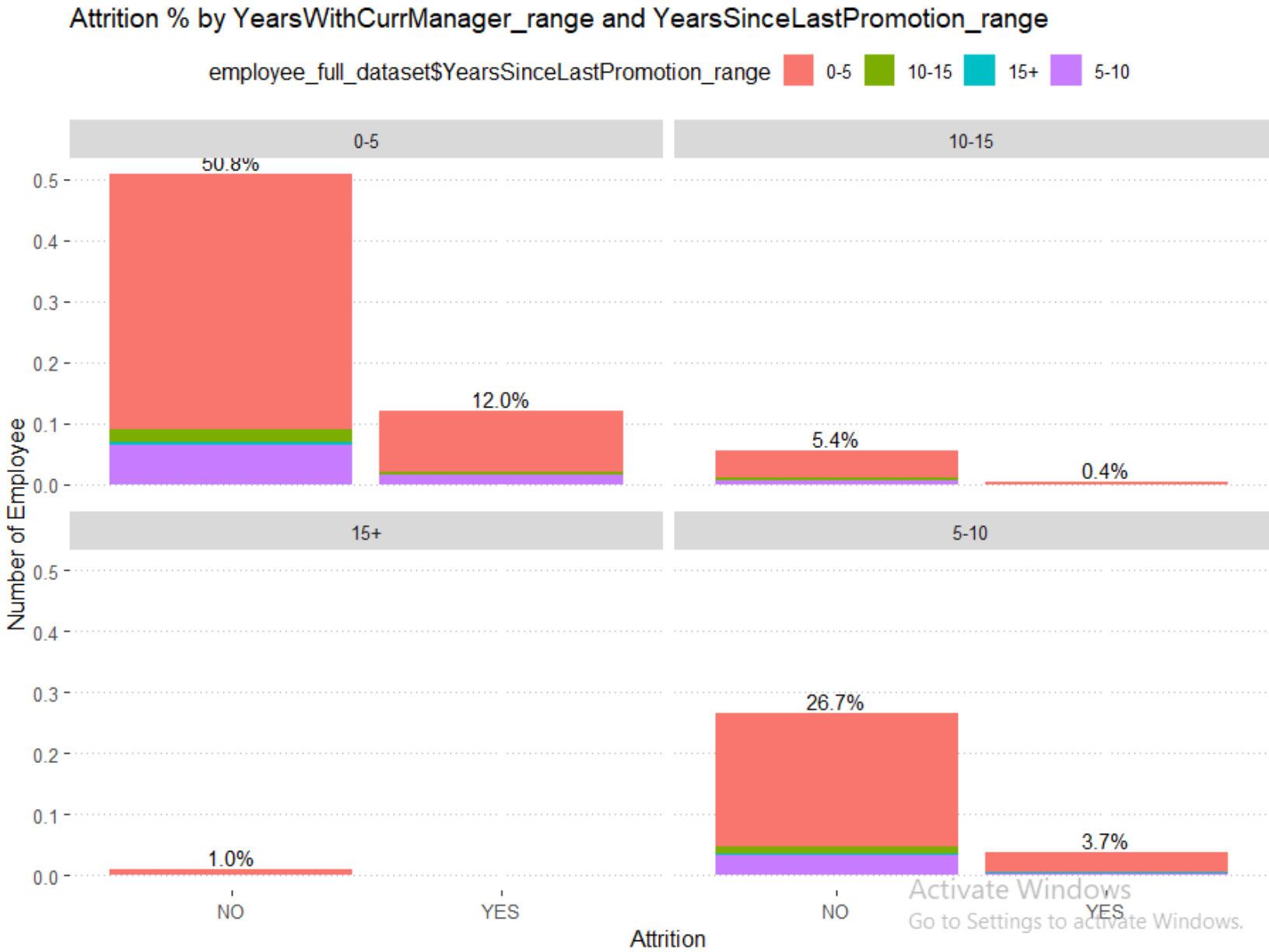
With 0-5 years group shows high attrition rate



EDA Analysis

Attrition % by YearsWithCurrManager_range and YearsSinceLastPromotion_range

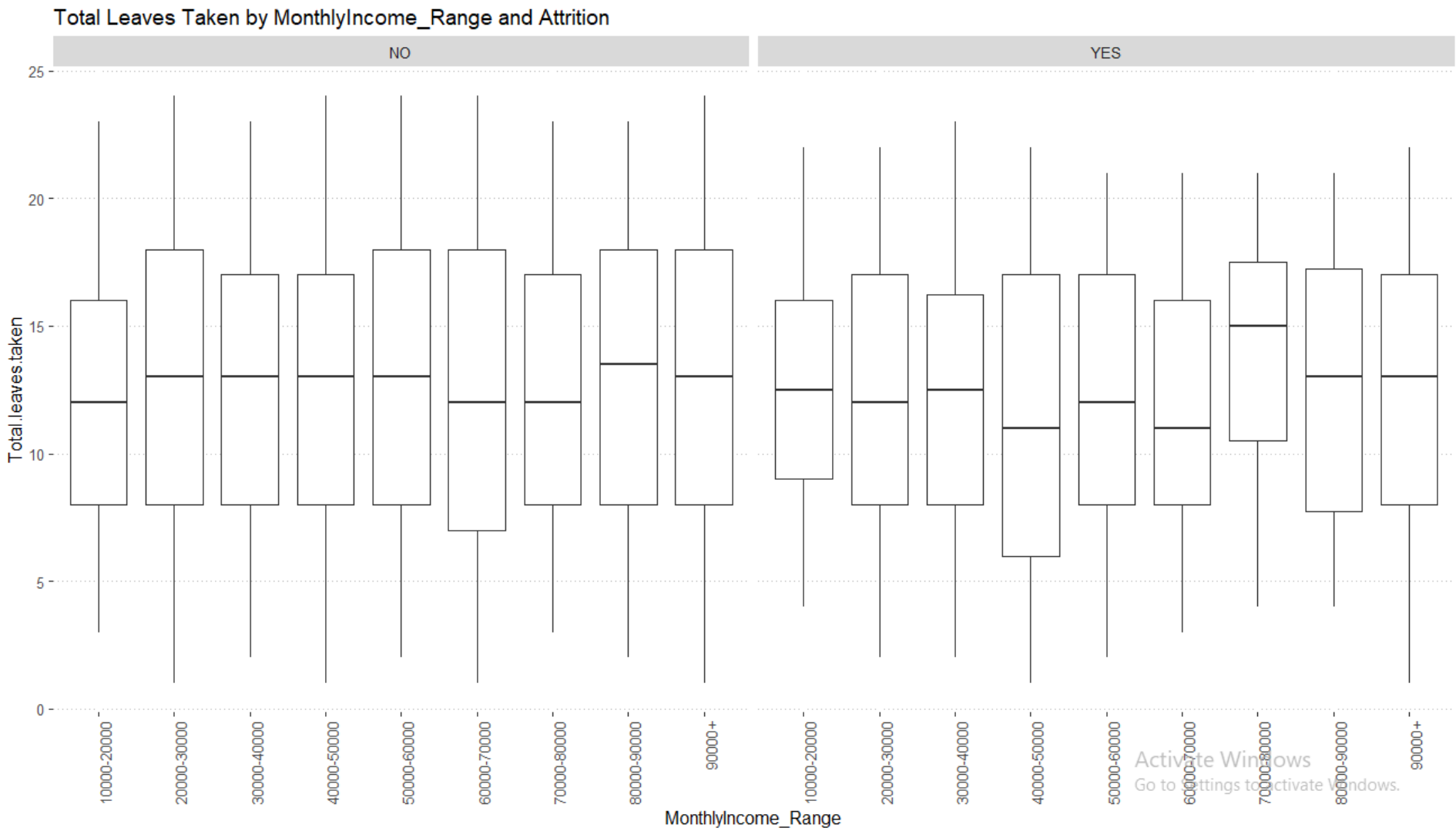
With 0-5 years group shows high attrition rate



EDA Analysis

Total Leaves Taken by MonthlyIncome_Range and Attrition

With 40000-50000 monthly income group taken more leaves.



Model Building

Information related to data used for model building

1. Total variables used for Model building is 4382 obs. Of 63 variables.
2. We used Logistic Regression model to build a model, in this we considered 70% training data and 30% as a test data.
3. Using the Family as a “Binomial”, since Attrition is 0 or 1.
4. Dependent Column “Attrition”

Operations and AIC values

	Null deviance	Residual deviance	AIC
Model 1 (62 independent + 1 dependent)	2704.5	1979.3	2105.3
StepAIC execution Number of columns suggested : 27	2704	1996	2068
Model 10 (37 independent + 1 dependent)	2704.5	1991.3	2069.3
Model 20 (12 independent + 1 dependent)	2704.5	2104.0	2132
Final Model : Model 20 (9 independent + 1 dependent)	**2704.5**	**2135.0**	**2157**

Model Building - Significant Features P-values and coefficients

Coefficient values

Feature	Coefficients	P value	Significance Stars
(Intercept)	-2.20686	2e-16	***
NumCompaniesWorked	0.25770	5.44e-06	***
TotalWorkingYears	-0.74781	2e-16	***
YearsSinceLastPromotion	0.50694	3.00e-11	***
YearsWithCurrManager	-0.49320	1.24e-08	***
Total.logged.hours	0.64084	2e-16	***
BusinessTravel.xTRAVEL_FREQUENTLY	0.33455	2.98e-11	***
MaritalStatus.xSINGLE	0.49690	2e-16	***
EnvironmentSatisfaction.xLOW	0.42791	2e-16	***
JobSatisfaction.xLOW	0.25880	1.33e-06	***
JobSatisfaction.xVERY.HIGH	-0.25696	5.64e-05	***

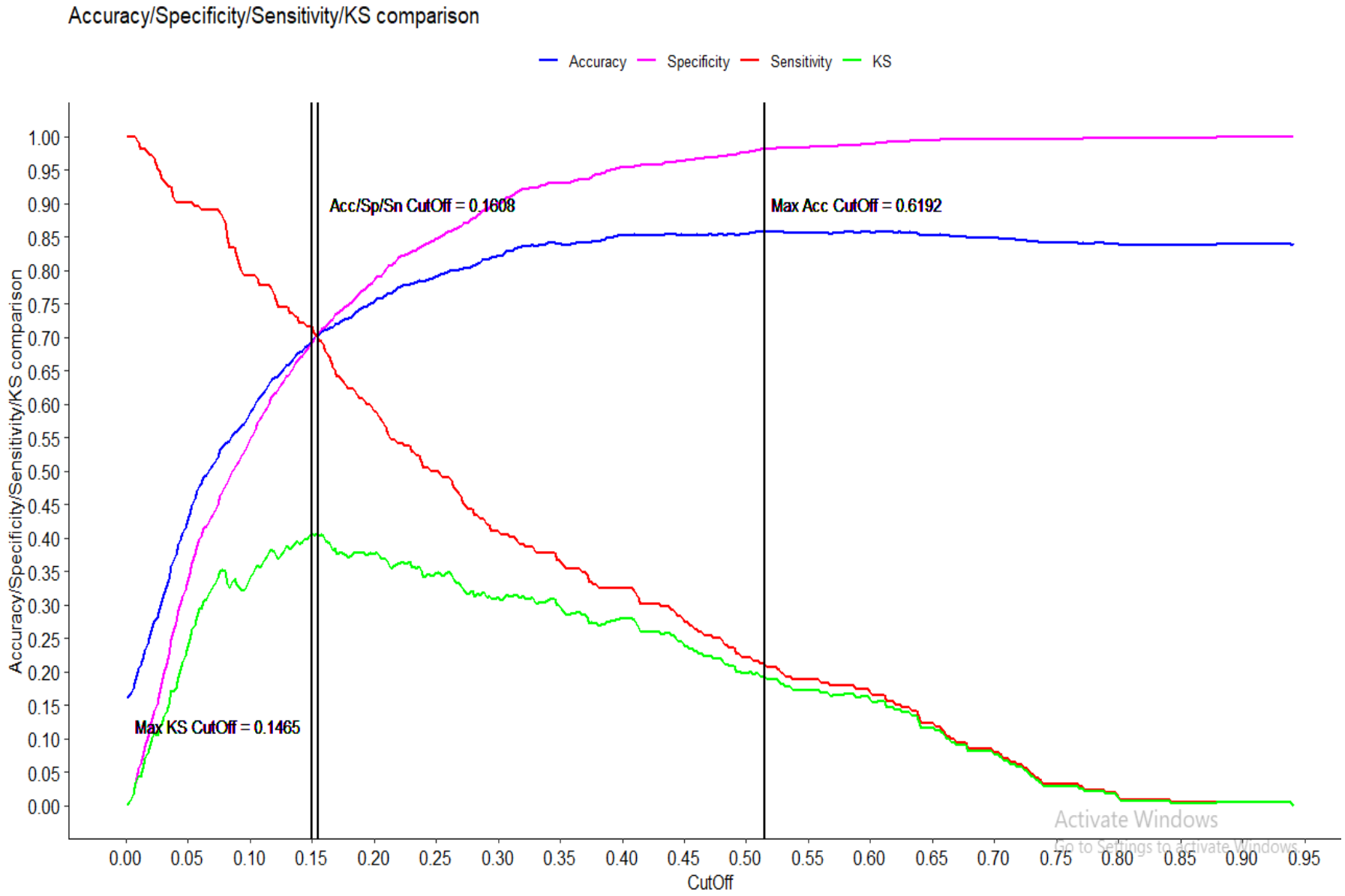
Model Evaluation – Deciding different cut offs

Step 1: Once Model is preparation is completed a cutoff values need to be decided for classifying 1 and 0 for the test data set.

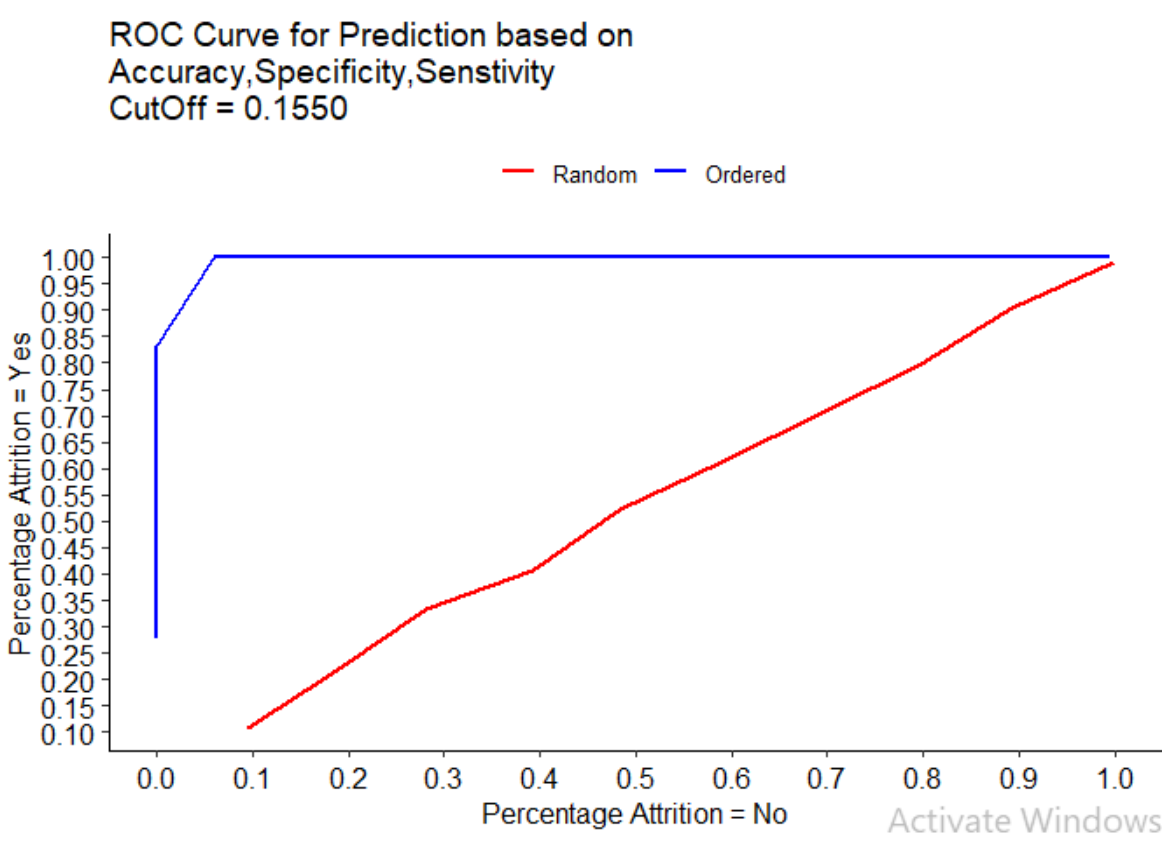
Cutoff value at Accuracy/SN/SP intersection (cutoff = 0.1608)

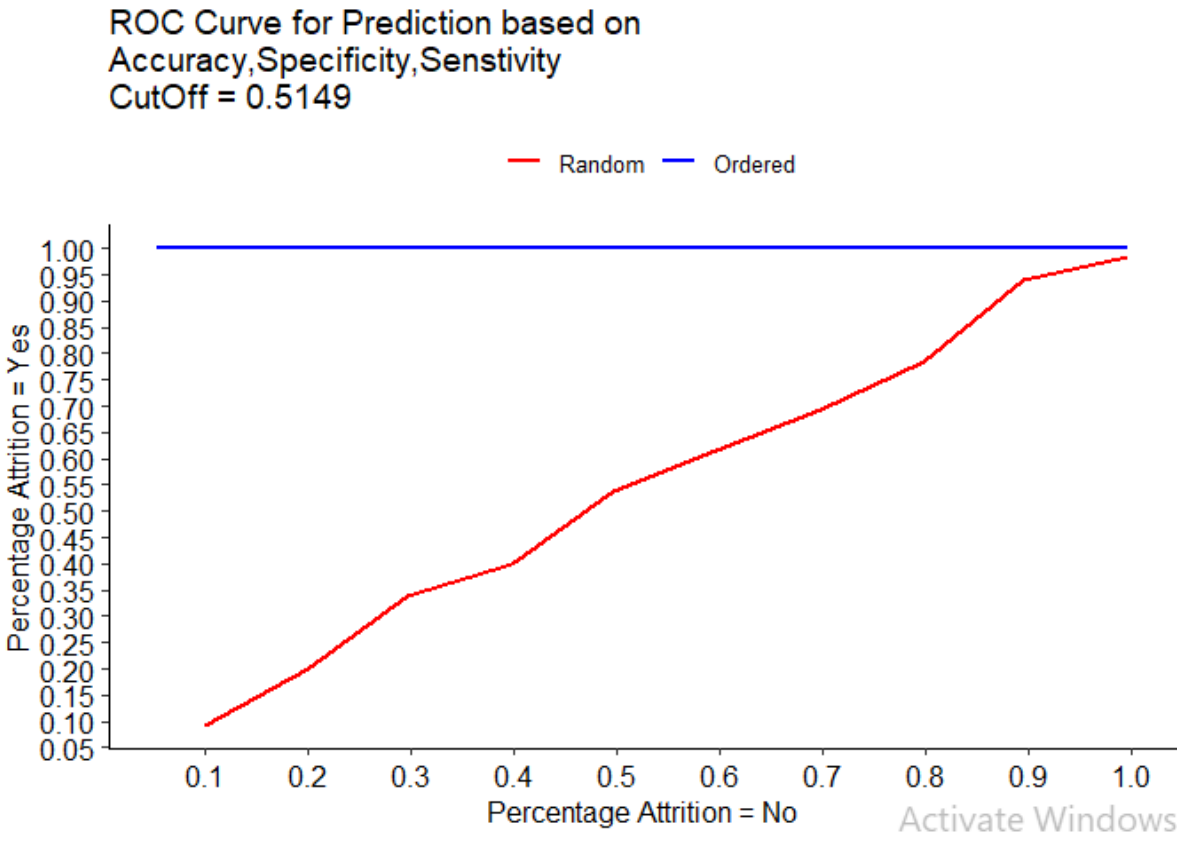
Cutoff with Minimum KS (cutoff = 0.1465)

Cutoff with maximum Accuracy (cutoff = 0.6192)



Model Evaluation Metrics- ROC curves





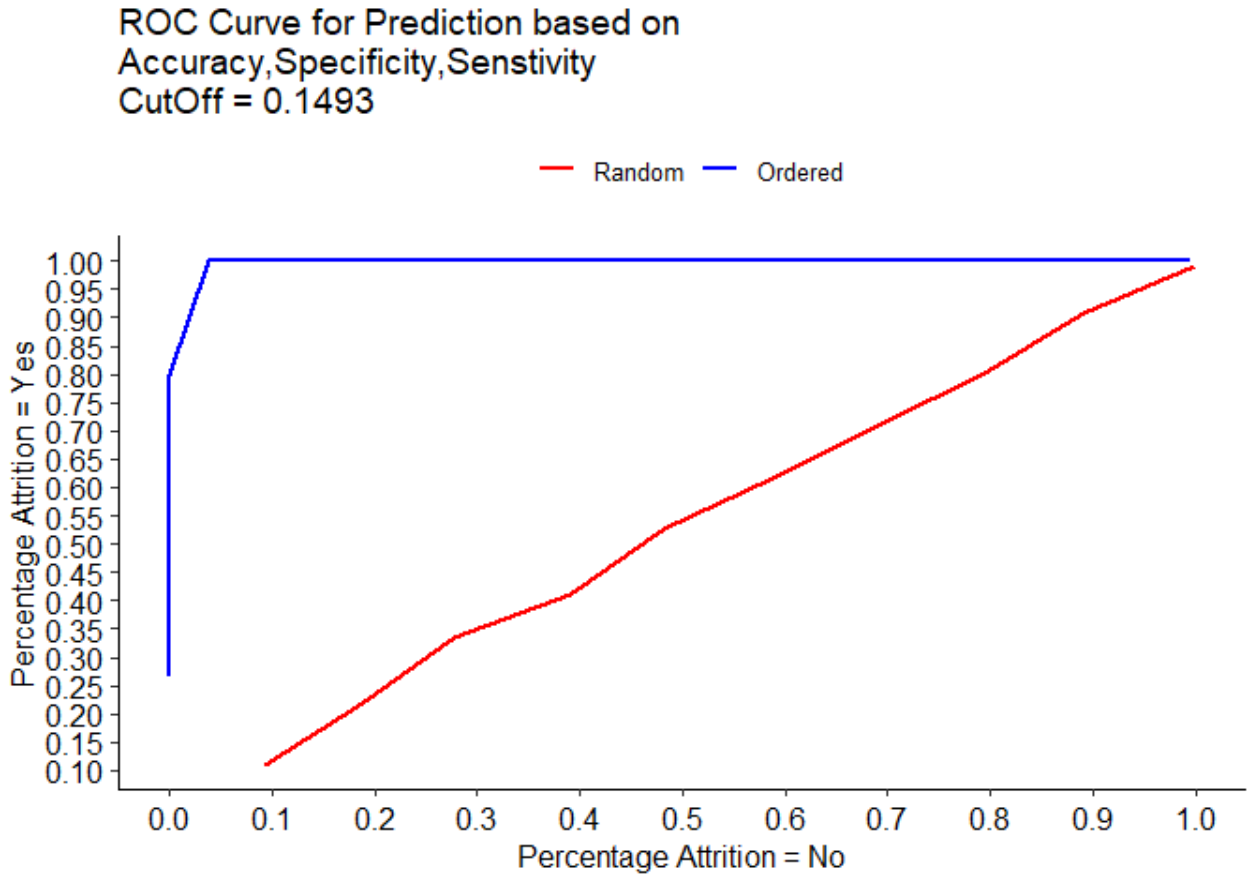
Model Evaluation Metrics- ROC curves

Inferences on ROC Curves:

- Cutoffs 1 and 3 are able to achieve a good separation between classes as can be seen from the almost rectangular ROC curves

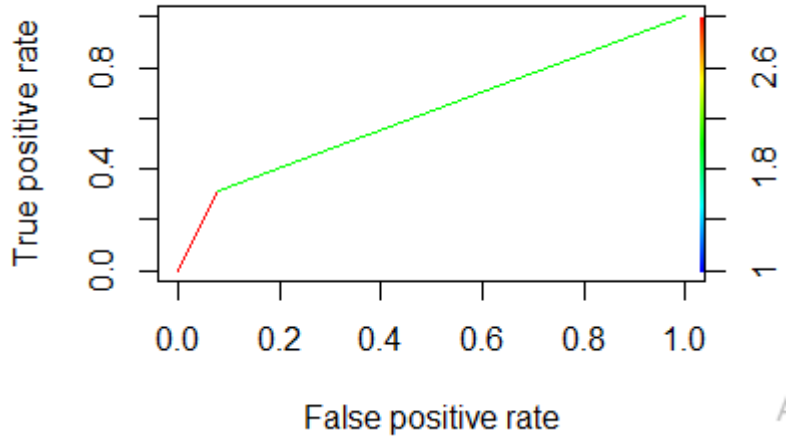
- The ROC curve for KS-Statistics cut off is slightly sharper (higher slope)

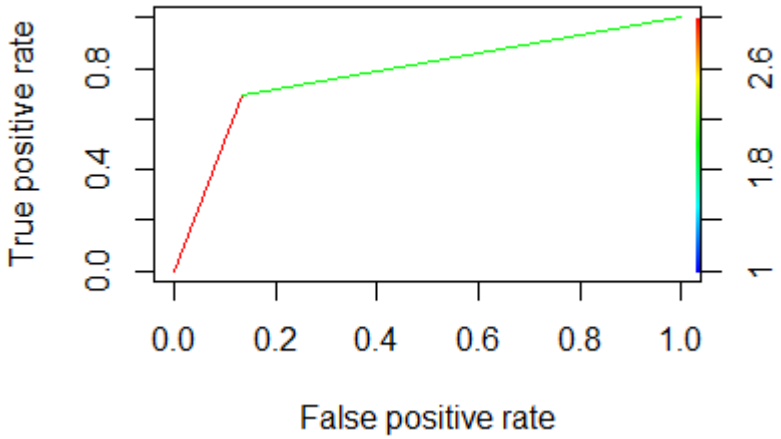
- KS should be our metric of choice as it achieves a slightly sharper curve and has greater AUC. (AUC available in next slide)



Model Evaluation Metrics - AUC

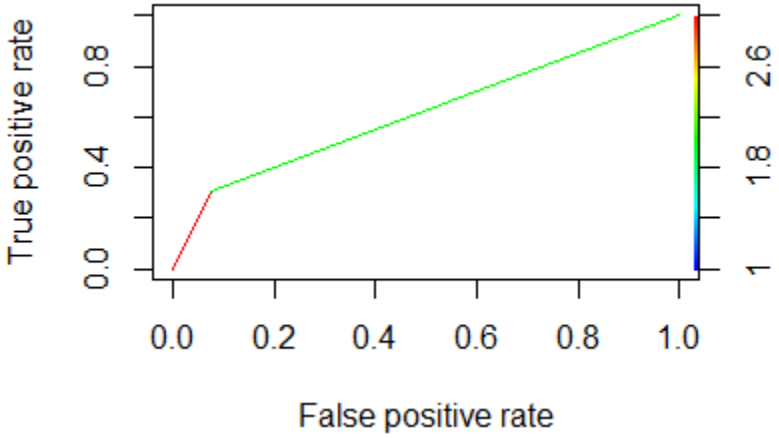
AUC value for Cut Off 1 is 0.6180195





AUC value for Cut Off 1 is 0.7793538

AUC value for Cut Off 1 is 0.6180195



From the above AUC we can conclude that "Accuracy : 0.7793538"

Inferences From Coefficients

Variable Name	Coefficients	P value	Significance
NumCompaniesWorked	0.25770	5.44e-06	Positive coefficient : Employee frequently switches job frequently = higher attrition probability
TotalWorkingYears	-0.74781	2e-16	Negative coefficient : Employee with Higher experience = lesser the attrition probability
YearsSinceLastPromotion	0.50694	3.00e-11	Positive coefficient : Lesser the promotion = higher attrition probability
YearsWithCurrManager	-0.49320	1.24e-08	Negative coefficient : frequently changing the manager = higher attrition probability
Total.logged.hours	0.64084	2e-16	Positive coefficient : Lesser the logged hours = higher attrition probability
BusinessTravel.xTRAVEL_FREQUENTLY	0.33455	2.98e-11	Positive coefficient : Higher Business Travel = higher attrition probability
MaritalStatus.xSINGLE	0.49690	2e-16	Positive coefficient : Employees in Single status = higher attrition probability
EnvironmentSatisfaction.xLOW	0.42791	2e-16	Positive coefficient : Lower the Employee Statisfaction = higher attrition probability
JobSatisfaction.xLOW	0.25880	1.33e-06	Positive coefficient : Lower the Job Statisfaction = higher attrition probability
JobSatisfaction.xVERY.HIGH	-0.25696	5.64e-05	Negative coefficient : Higher the Job Statisfaction = higher attrition probability

Conclusion and Recommendations…

One of the primary and highly influential factors is the number of hours an employee puts in XYZ Corporation. Practice should be in place to avoid late working hours of the employees. This would result from better project planning by management.

1. Recognition is always the perfect appreciation. Promotions are an important factor in influencing attrition. Promotions & appreciations should become a core value for the organization. This will build employee motivation and lower the probability of attrition.
2. Frequent changes to the management is a negative element for attrition. It is identified that employees have concern and reservation against working with new managers or different managers very frequently.
3. Job Satisfaction, Environmental Satisfaction & Work Life Balance, are obvious reasons influencing attrition. HR should work to enhance employee engagement apart from regular work by organizing events and team outings. Team building activities would also help increase employee satisfaction.
4. Business Travel has come out to be negative element for attrition. Higher management should leverage technologically enhanced tools (WebEx, Virtualization and Remote Management etc) for avoiding travel for employees.