trabajo_hecho

RialPepe

2022-10-18

I. Data Pre-processing

1. Looking through the Data

Result- Over 18, EmployeeCount, EmployeeNumber & Standard working hours have same values for all employees. Summary result of rest of the data seems okay with no major discrepancies. (We have 'Att' coloum which was introduced in the data file before uploading on R to give binary values to Attrition: 0 as 'No attrition' and 1 as 'yes attrition')

Result- We have dropped four column- attrition (repeated coloum as we have new coloum Att),Over 18, EmployeeCount, EmployeeNumber & Standard working hours. The dataset now has 1470 rows and 32 variables and has been renamed to 'data'.

2. Looking for Missing values

Result- There are No missing values in the dataset.

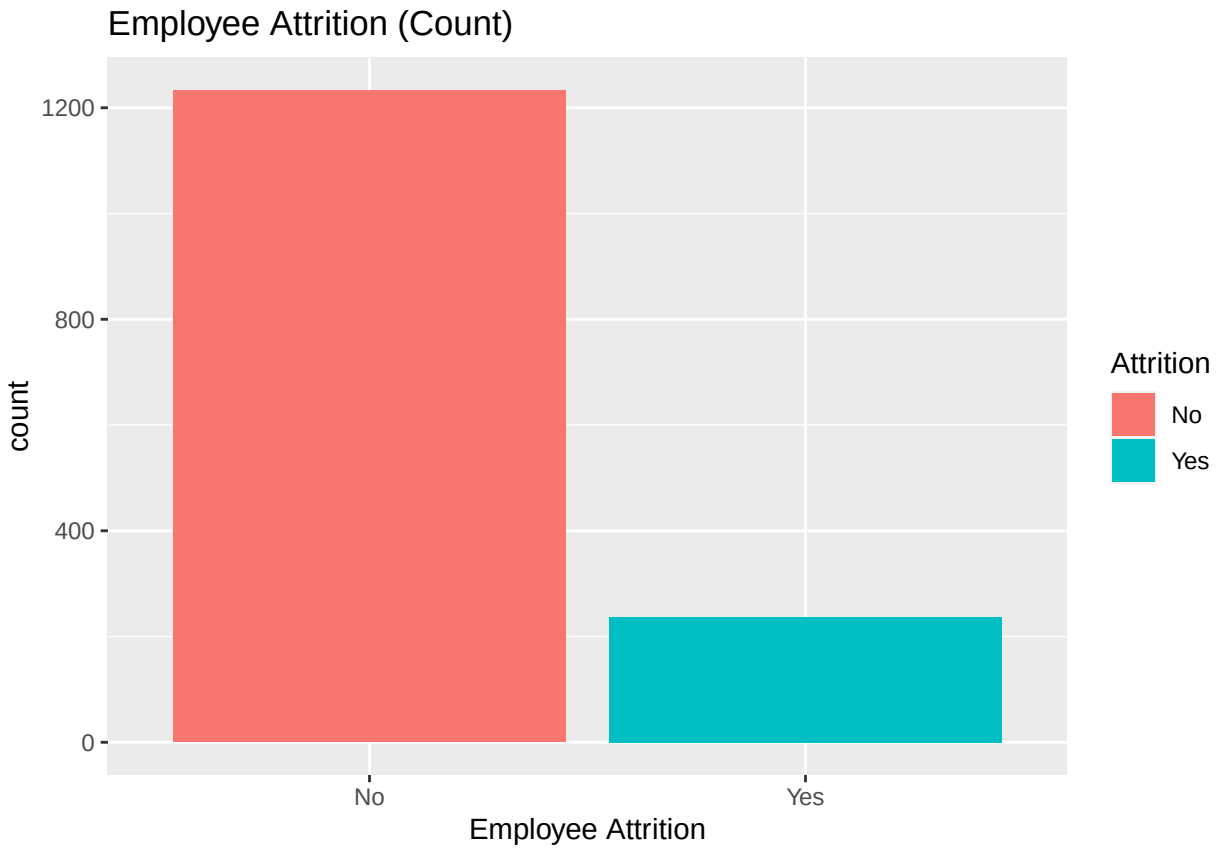
3. Structure of the Data to identify Integers and categorical values

Result- No interger/float type values, Categorical values which needs to be changed to factors- BusinessTravel", "Department", "DistanceFromHome", "Education", "EducationField", "EnvironmentSatisfaction", "Gender", "JobInvolvement", "JobLevel", "JobRole", "JobSatisfaction", "MaritalStatus", "OverTime", "PerformanceRating", "RelationshipSatisfaction", "StockOptionLevel", "WorkLifeBalance"

Result- All the categorical values changed to factors and the final dataset is ready for next step.

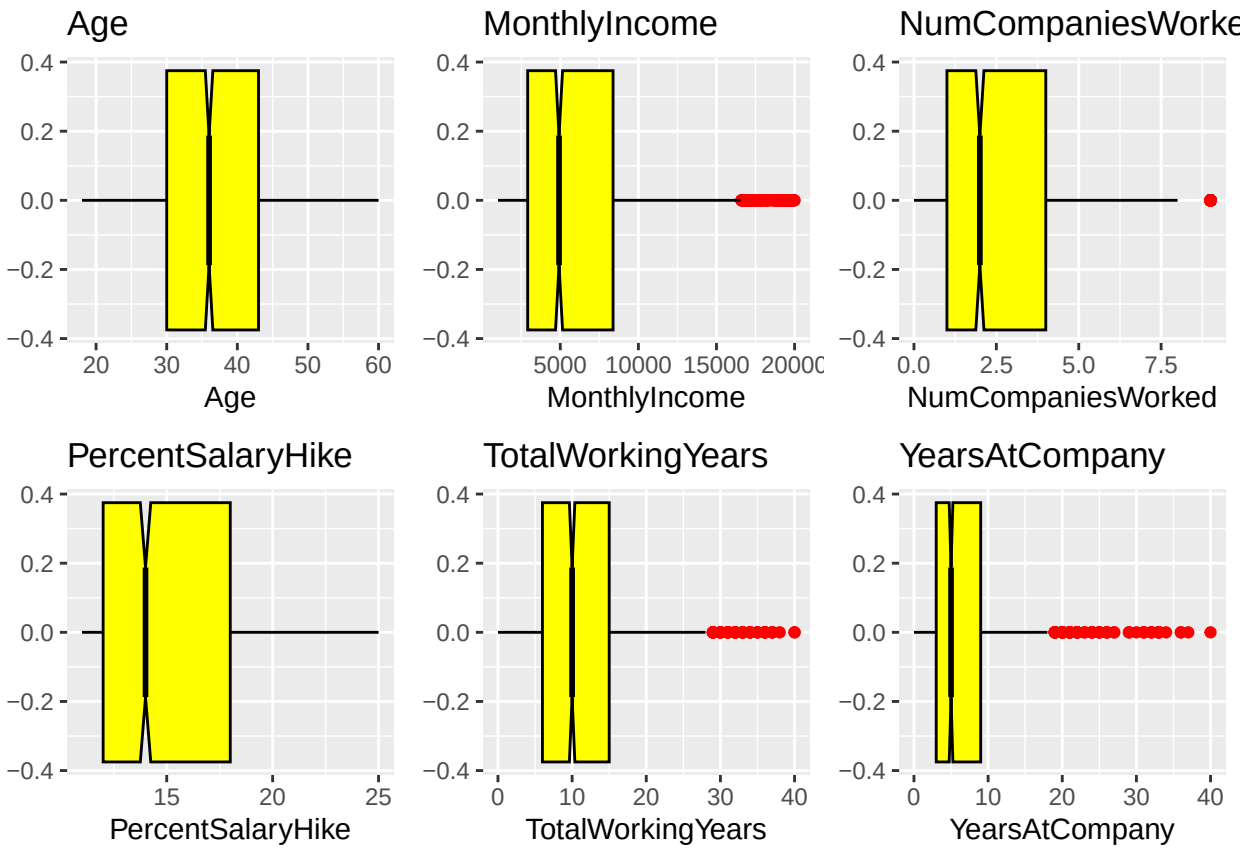
II. Exploratory Data Analysis

1. Checking for distribution of data



Result- The data sample is imbalanced as most of the instances in the dataset belong to category 0 (No Attrition). We will keep this in mind while distributing our data in 'test' and 'train' samples

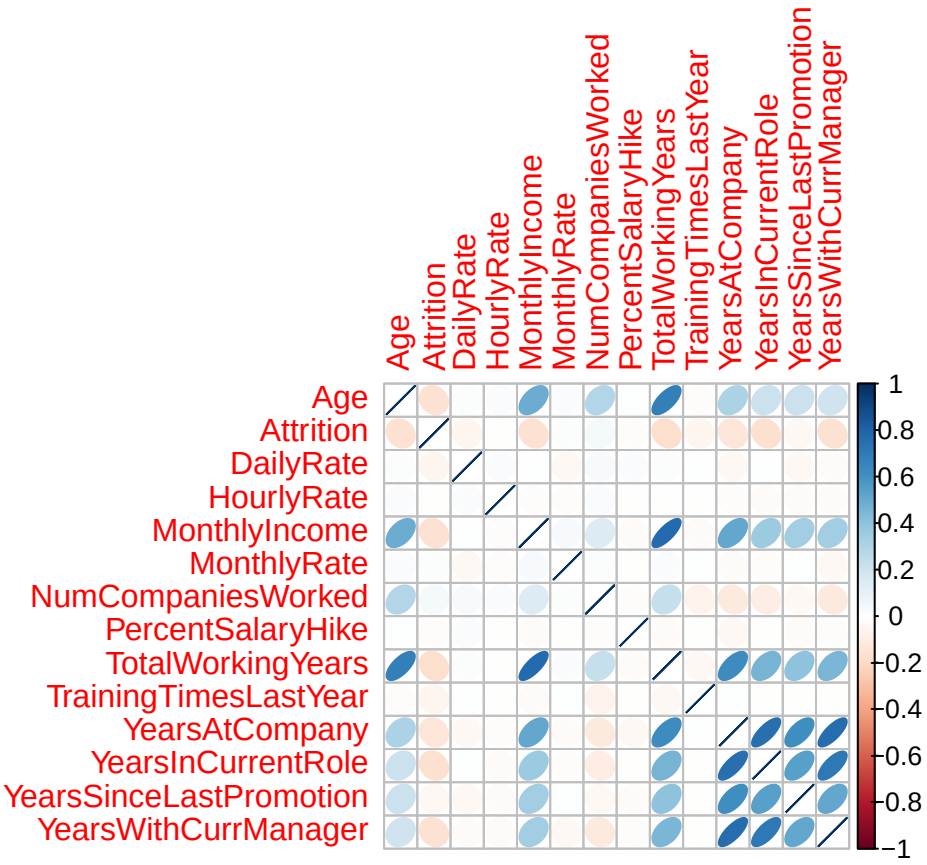
2. Box plots to check data distribution and outliers on numerical variables



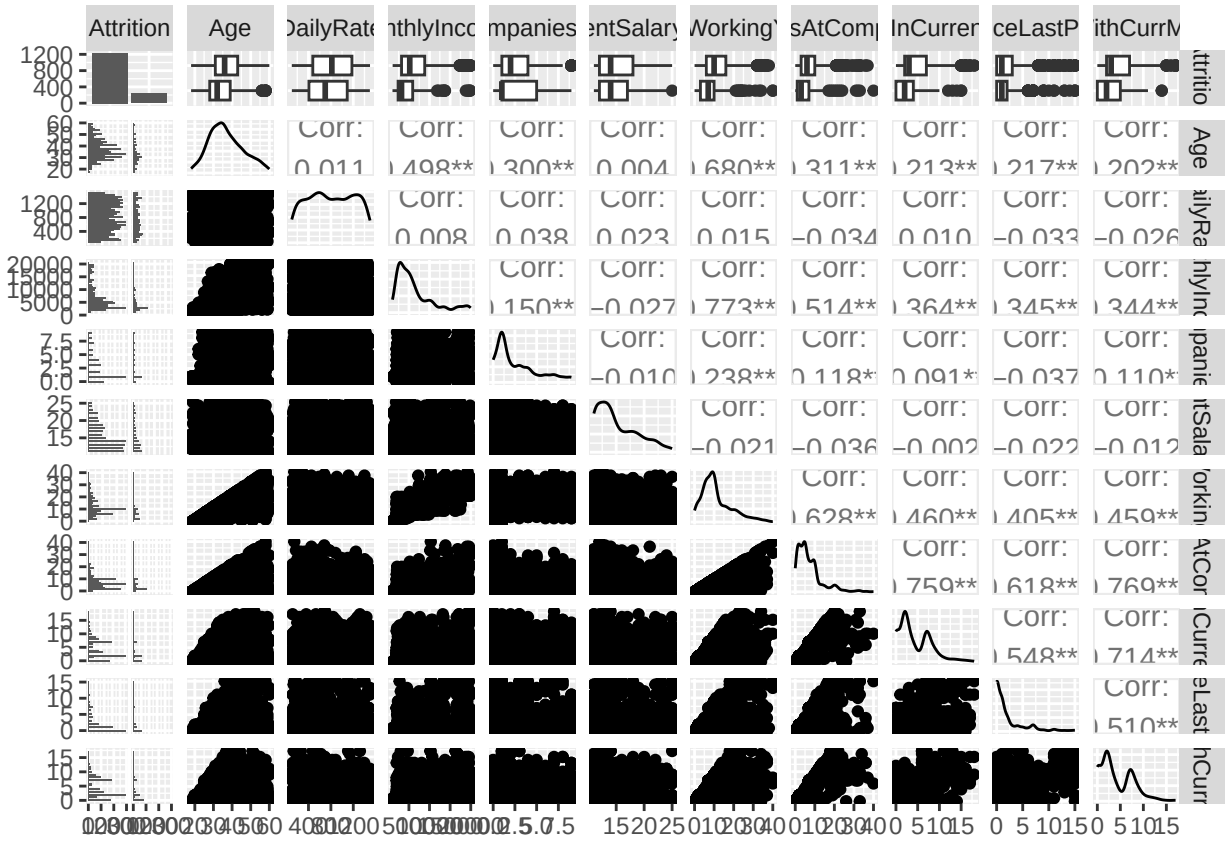
Result- Slight skewness and some outliers are detected, we will explore it further via cook distance with our first model.

3. Correlation between numerical variables

first grouping the numerical variables



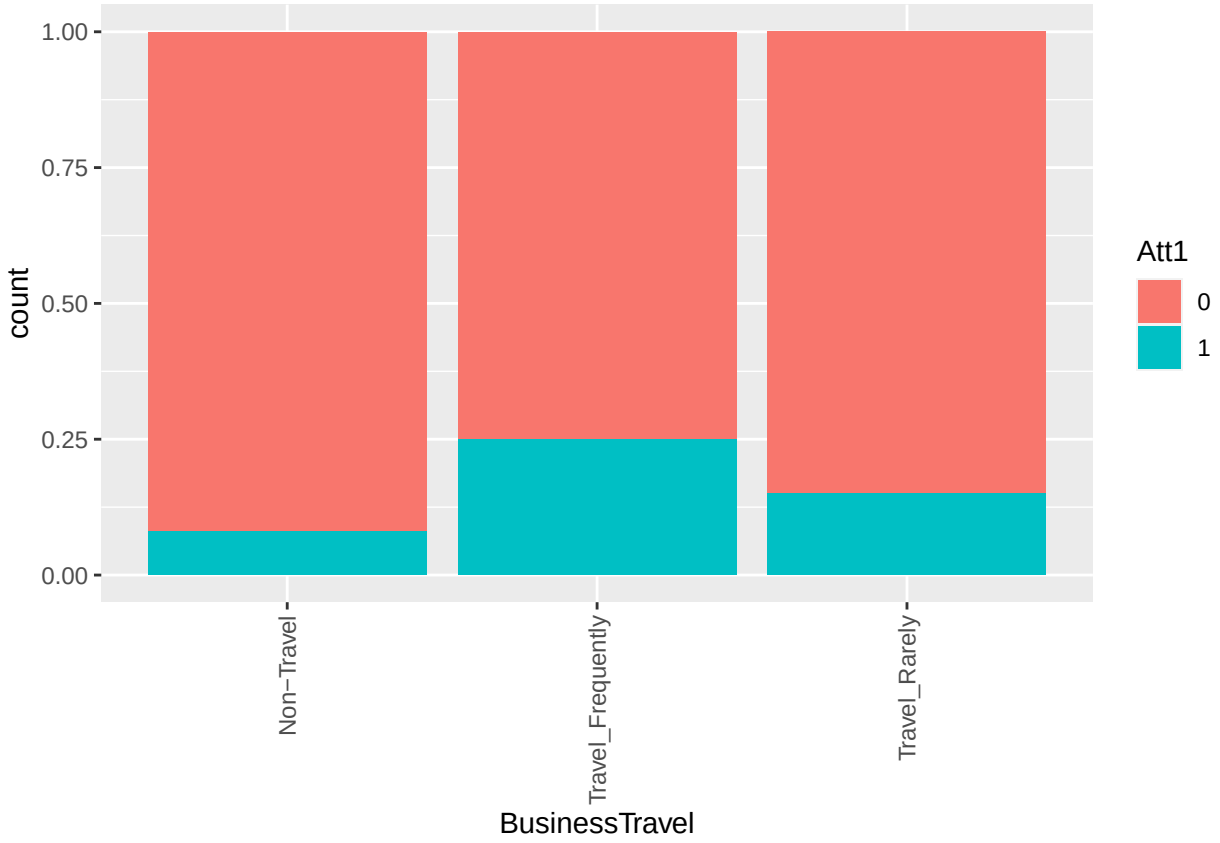
Scatterplot matrix

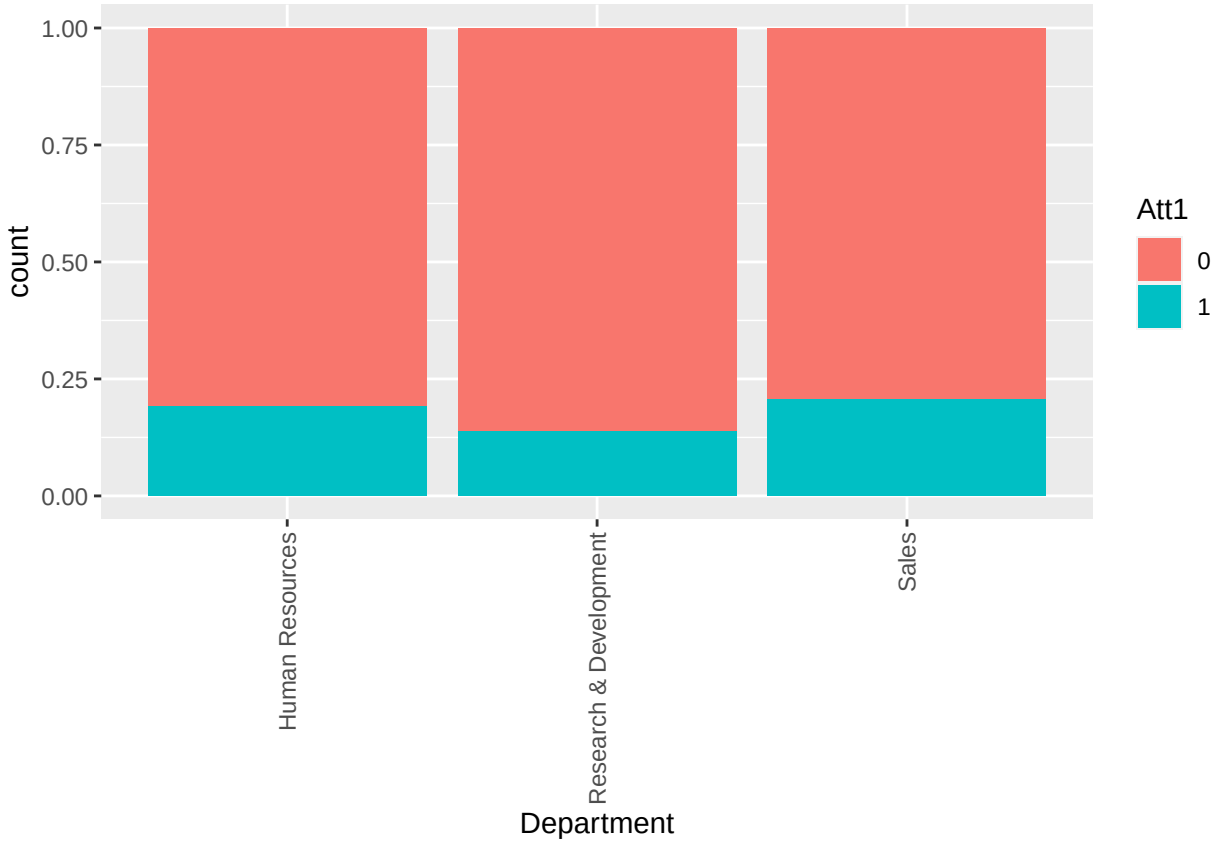


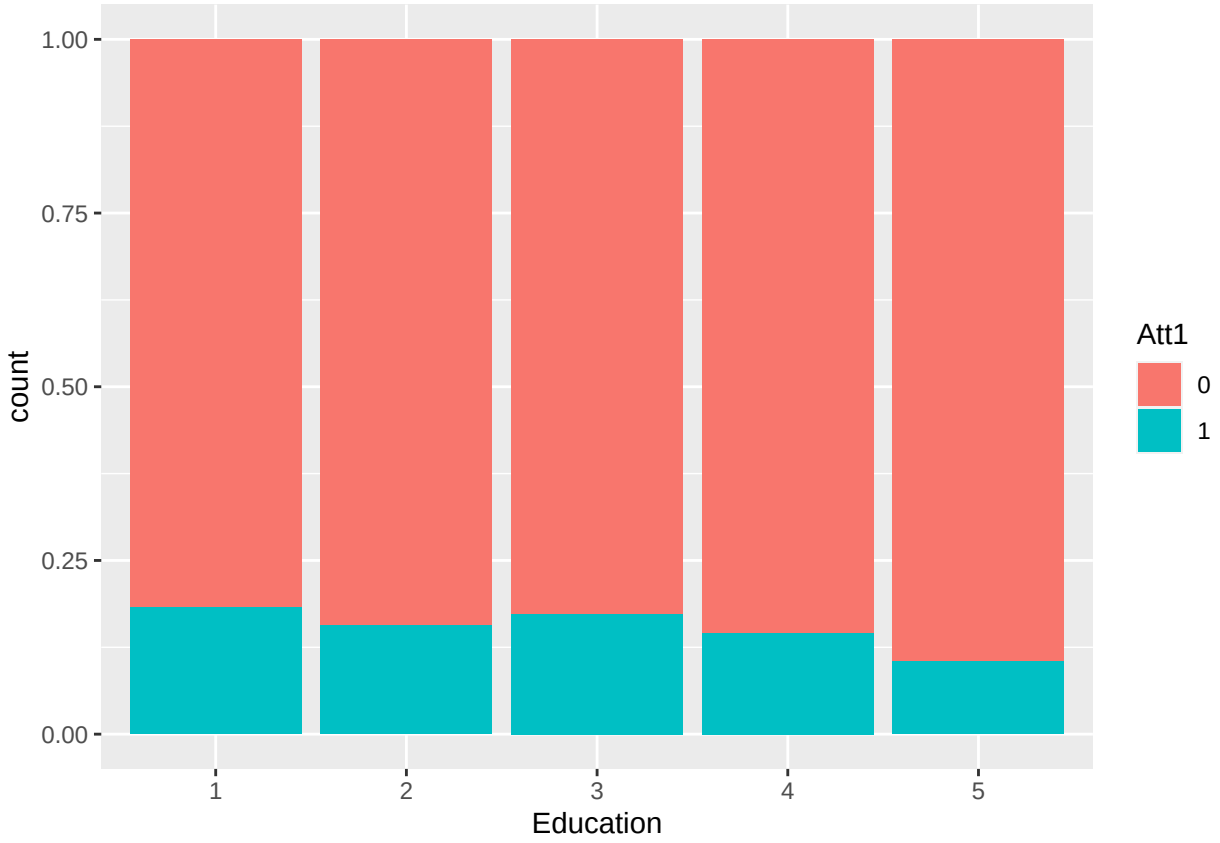
Result- High correlations were found for some variables like:

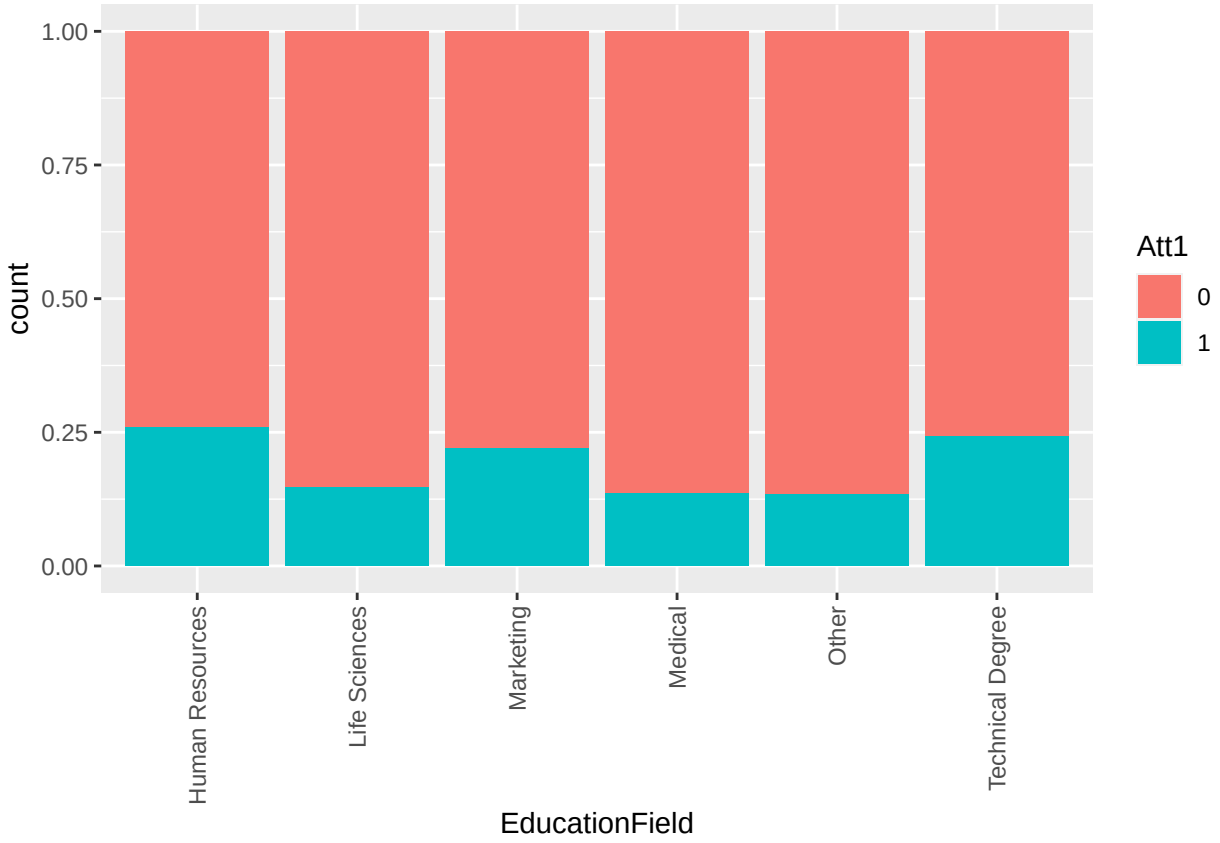
YearsAtCompany~YearsWithCurrManager 0.769 YearsAtCompany~YearsSinceLastPromotion 0.618 YearsAtCompany~YeasinCurrentRole 0.759 YearsinCurrentRole~YearsWithCurrManager 0.714

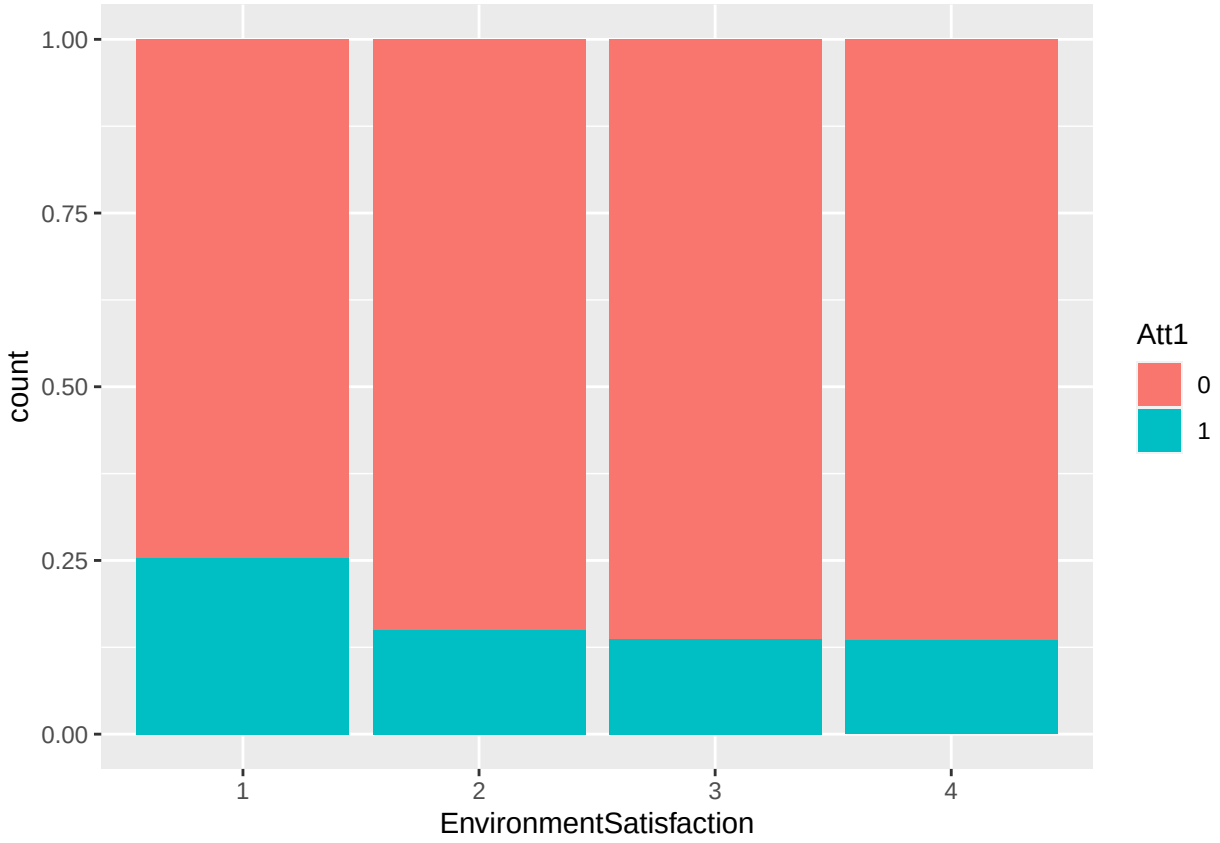
4. Checking relationship between attrition and categorical values via histograms

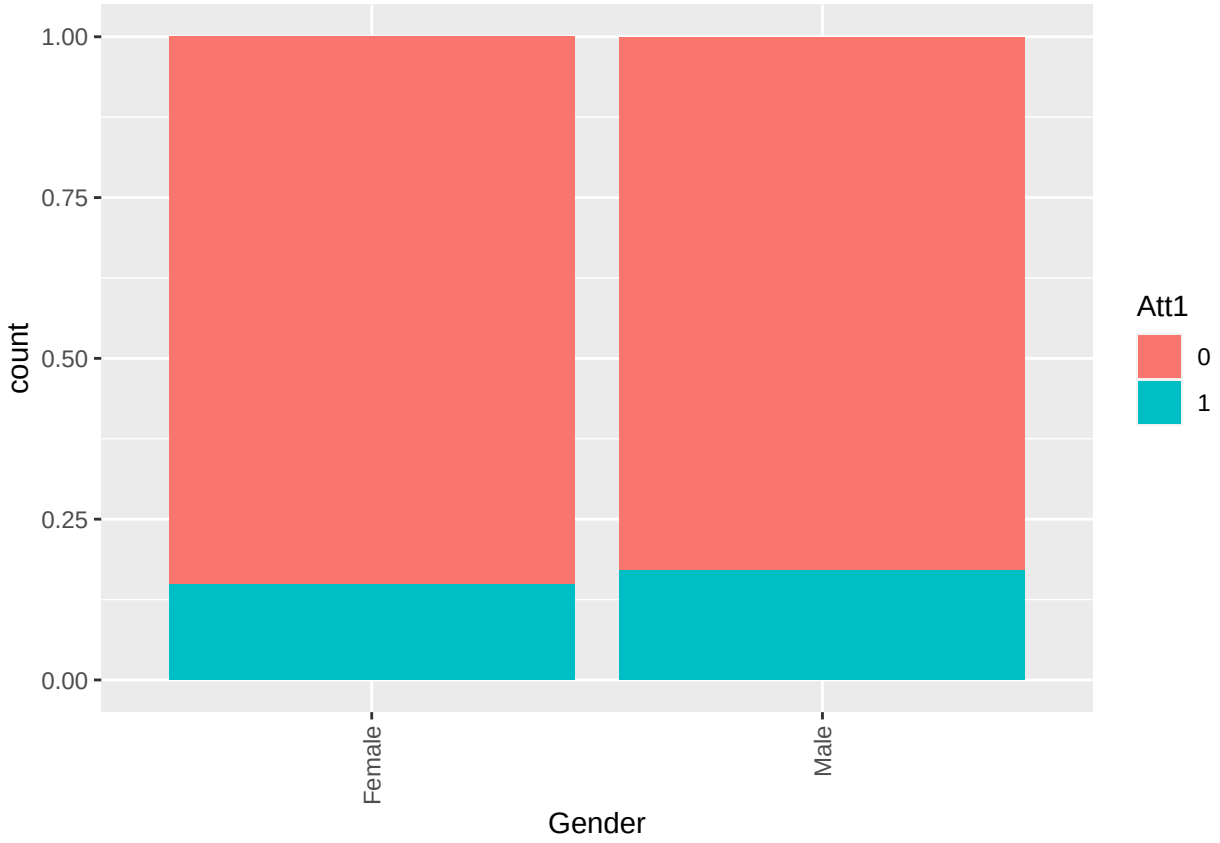

1.00
0.75
0.50
0.25
0.00
count
Non−Travel
Travel_Frequently
Travel_Rarely
BusinessTravel
Att1
0
1

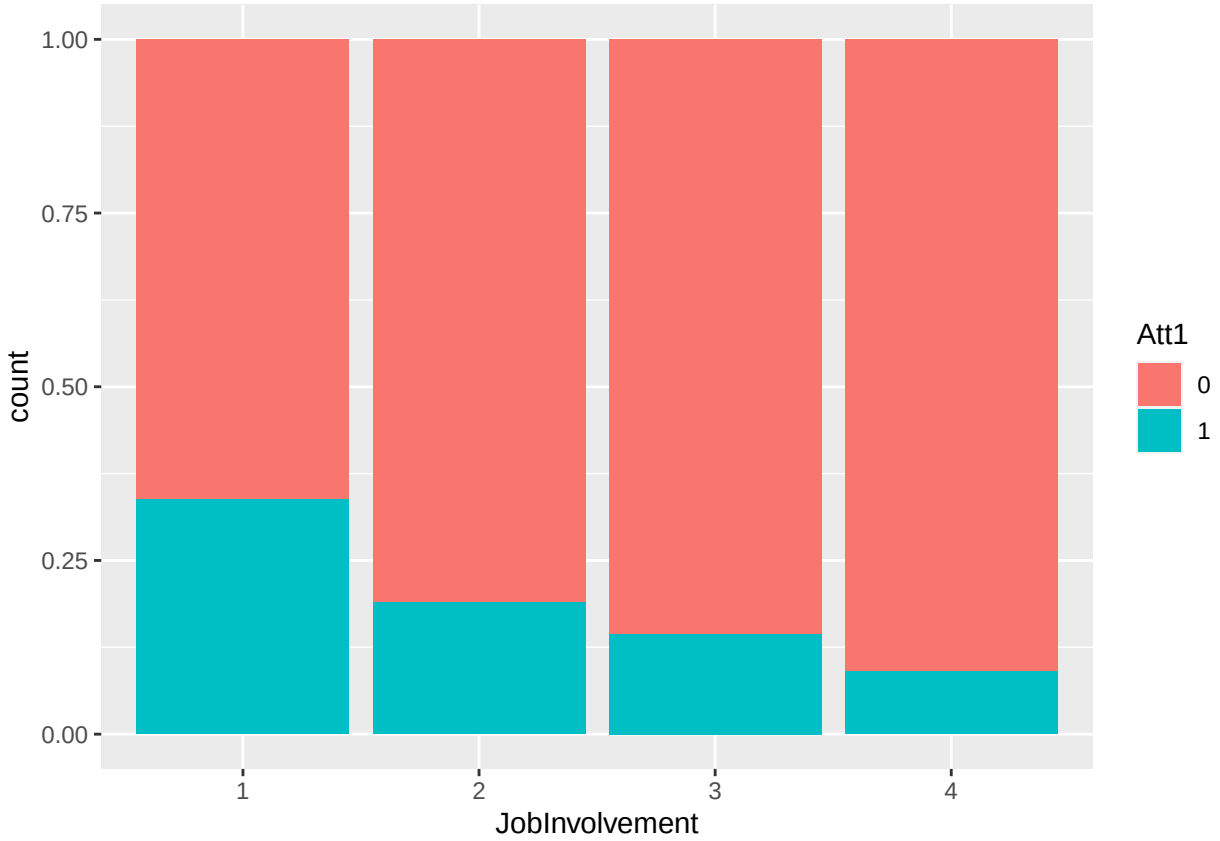

1.00
0.75
0.50
0.25
0.00
count
Human Resources
Research & Development
Sales
Department
Att1
0
1

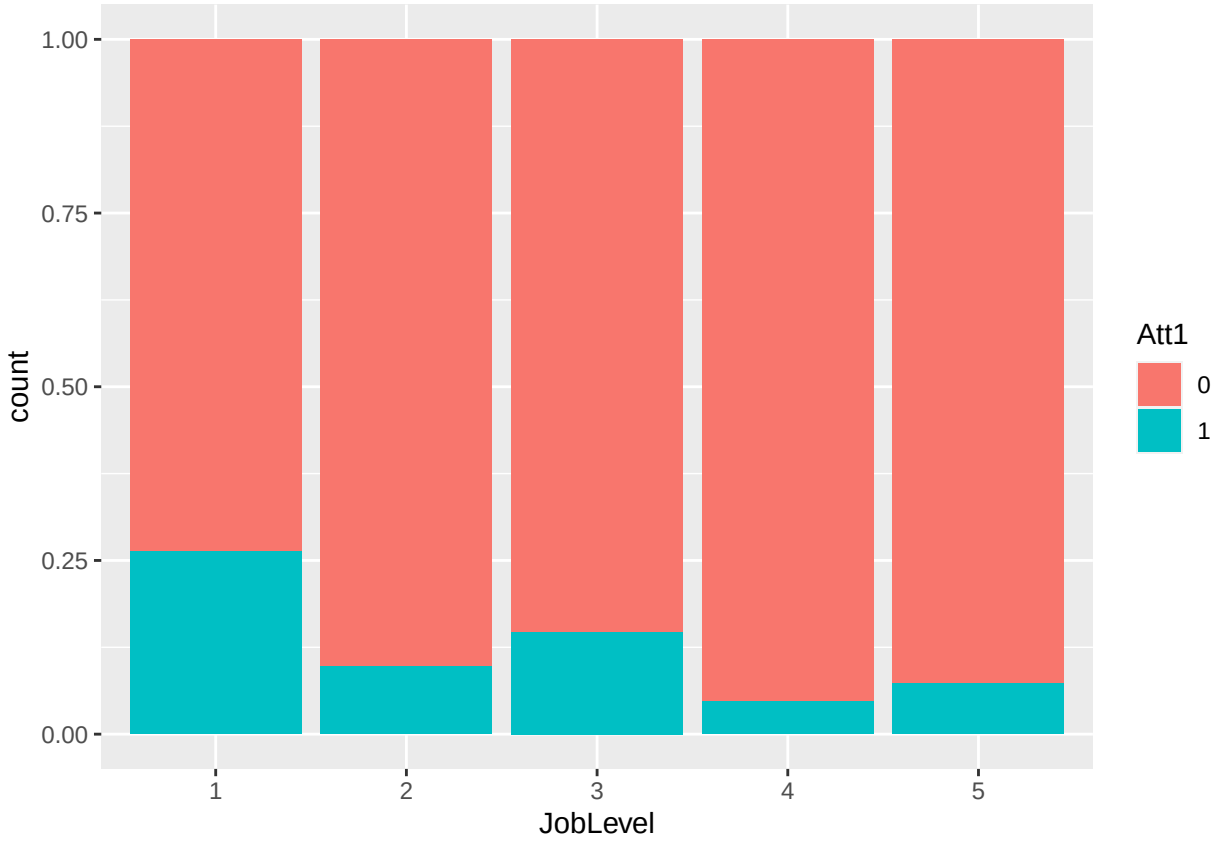

1.00
0.75
0.50
0.25
0.00
count
1
2
3
4
5
Education
Att1
0
1

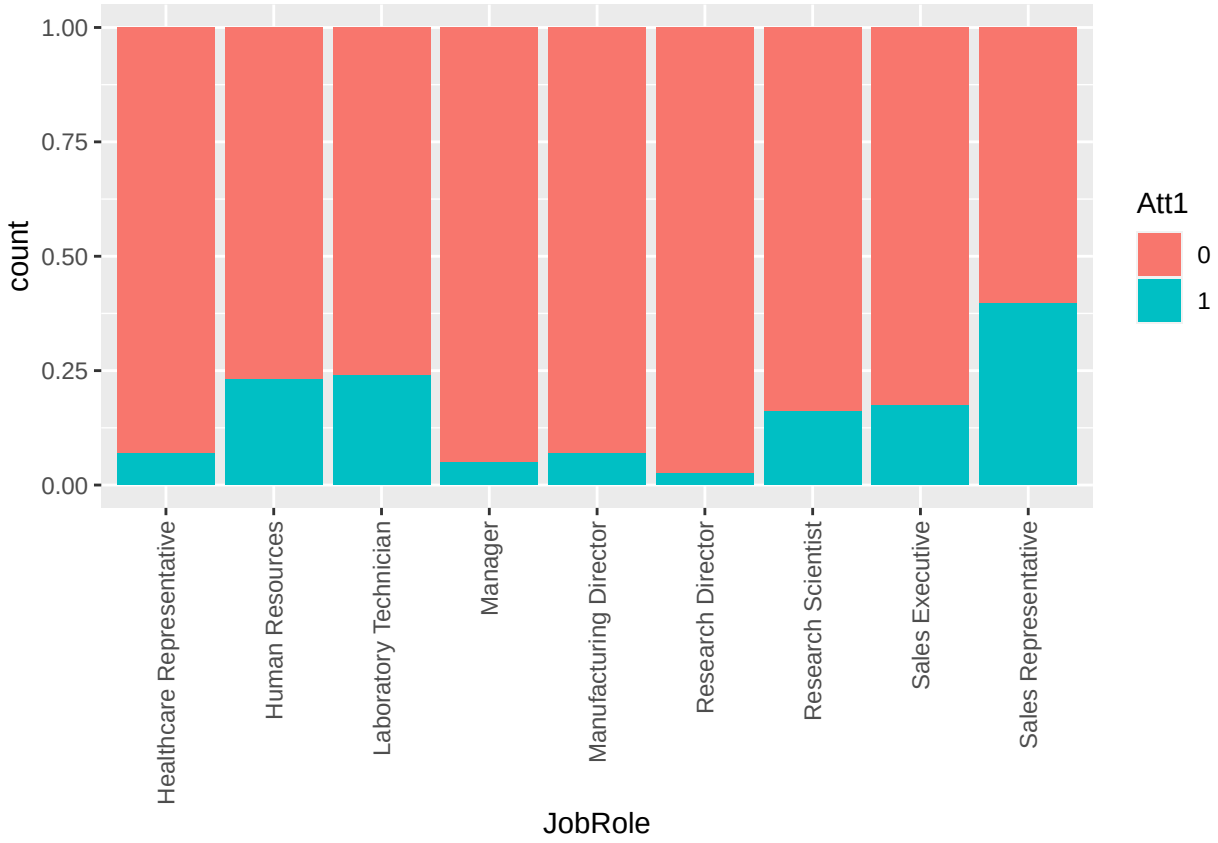

1.00
0.75
0.50
0.25
0.00
count
Human Resources
Life Sciences
Marketing
Medical
Other
Technical Degree
EducationField
Att1
0
1

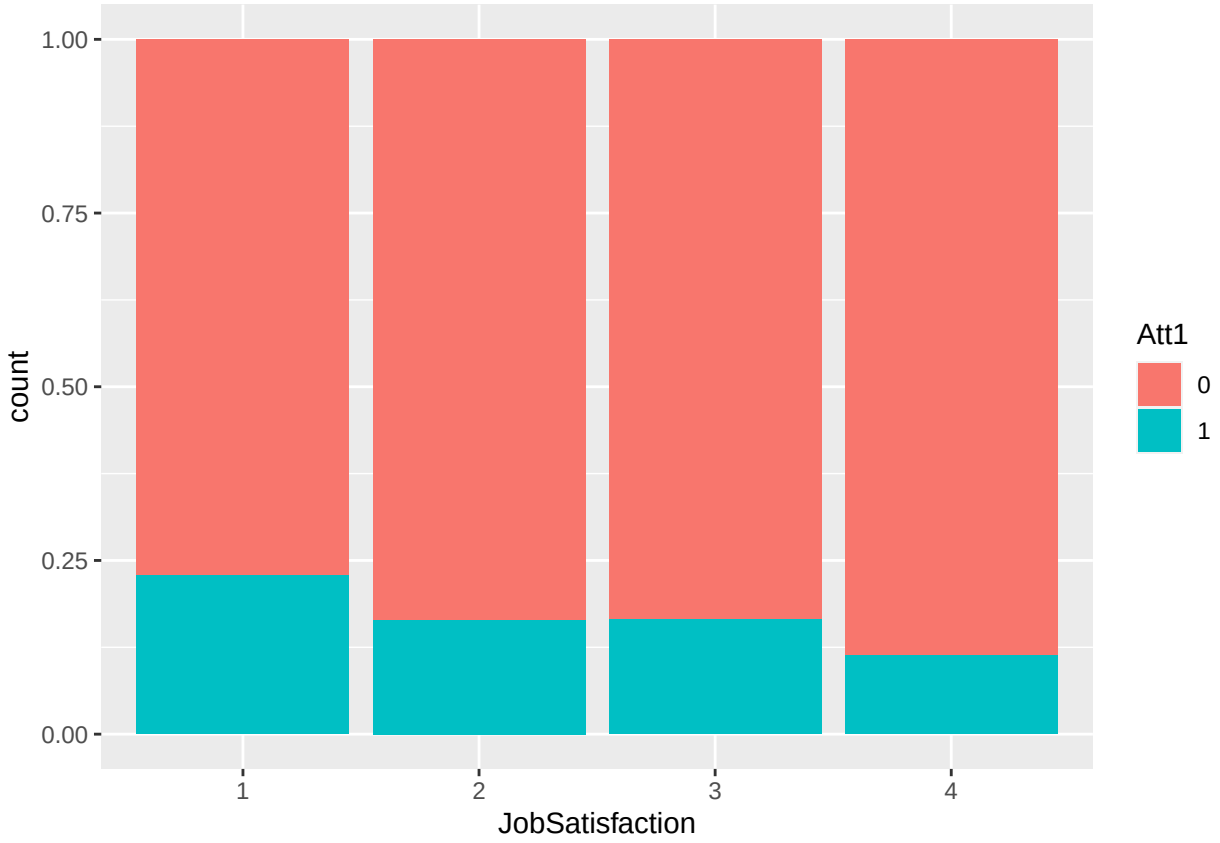

1.00
0.75
0.50
0.25
0.00
count
1
2
3
4
EnvironmentSatisfaction
Att1
0
1

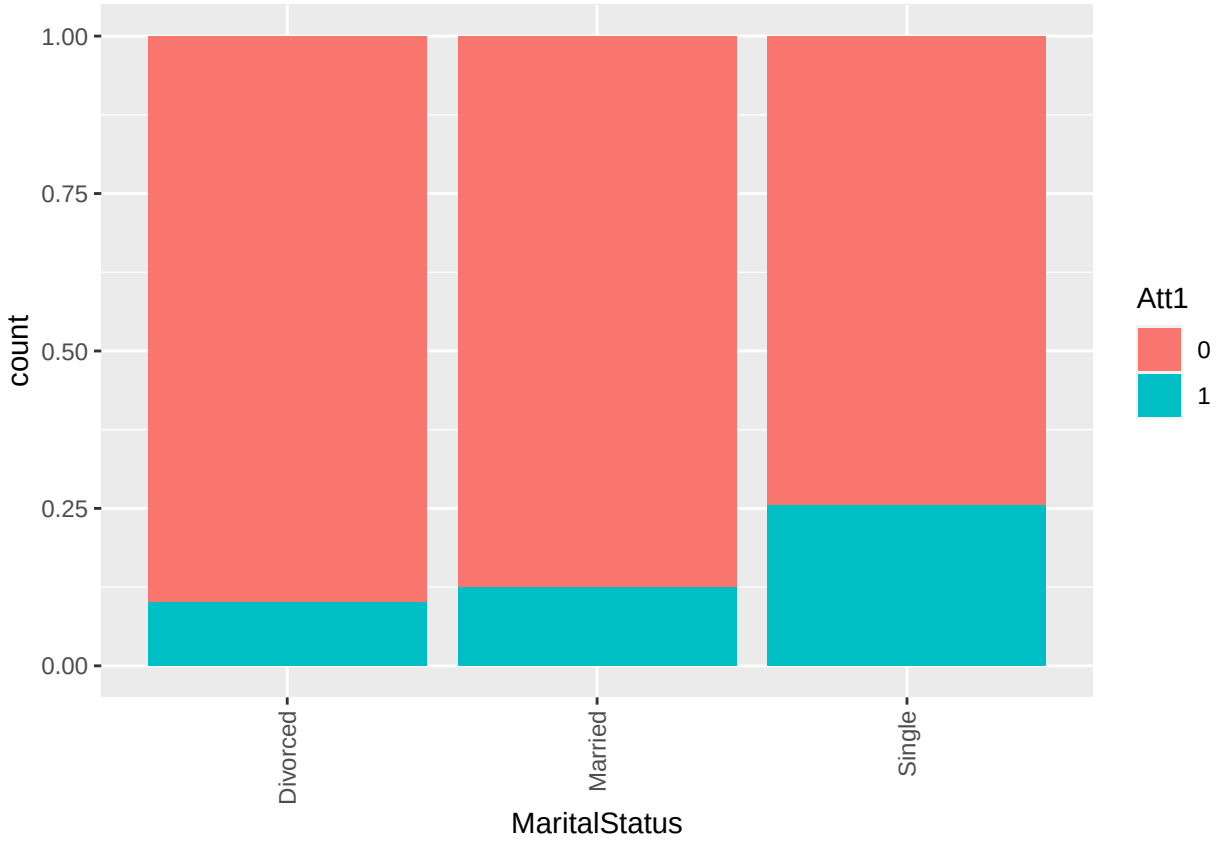

1.00
0.75
0.50
0.25
0.00
count
Female
Male
Gender
Att1
0
1

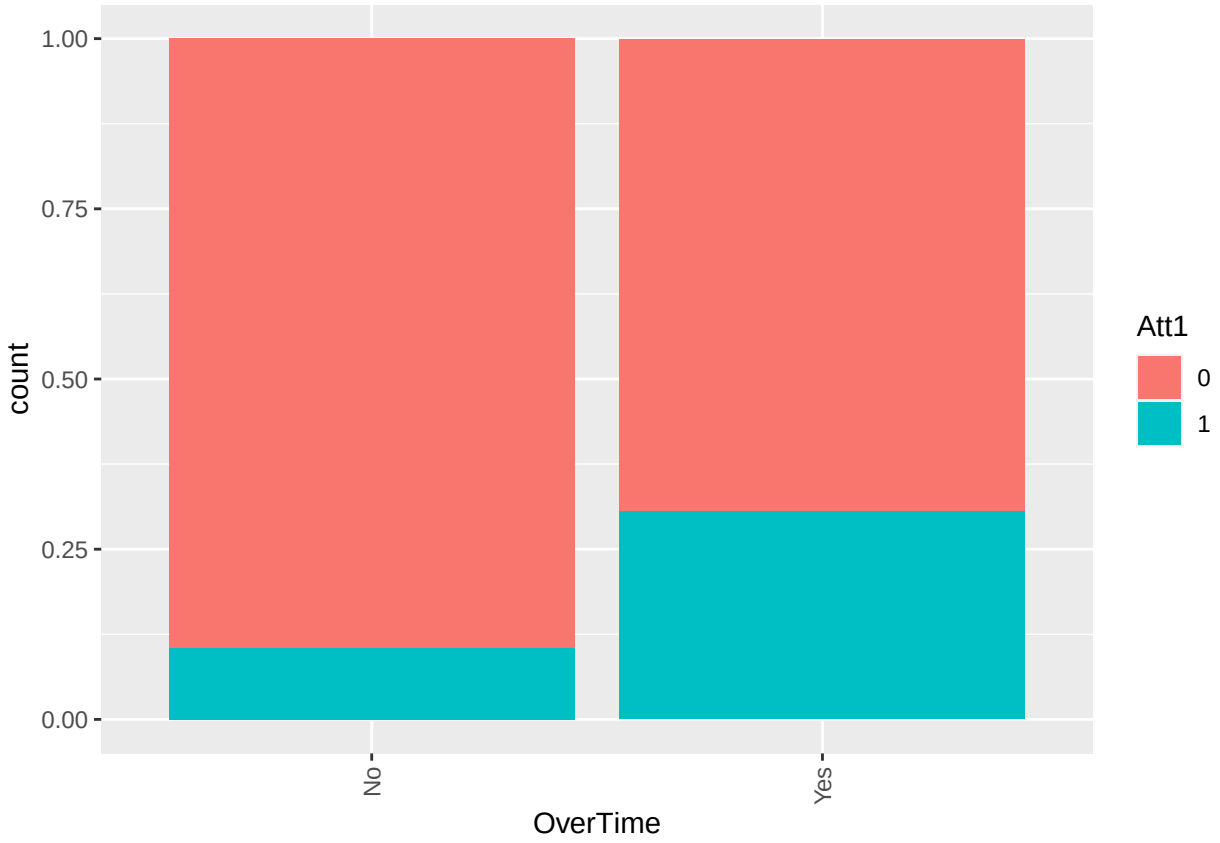

1.00
0.75
0.50
0.25
0.00
count
1
2
3
4
JobInvolvement
Att1
0
1

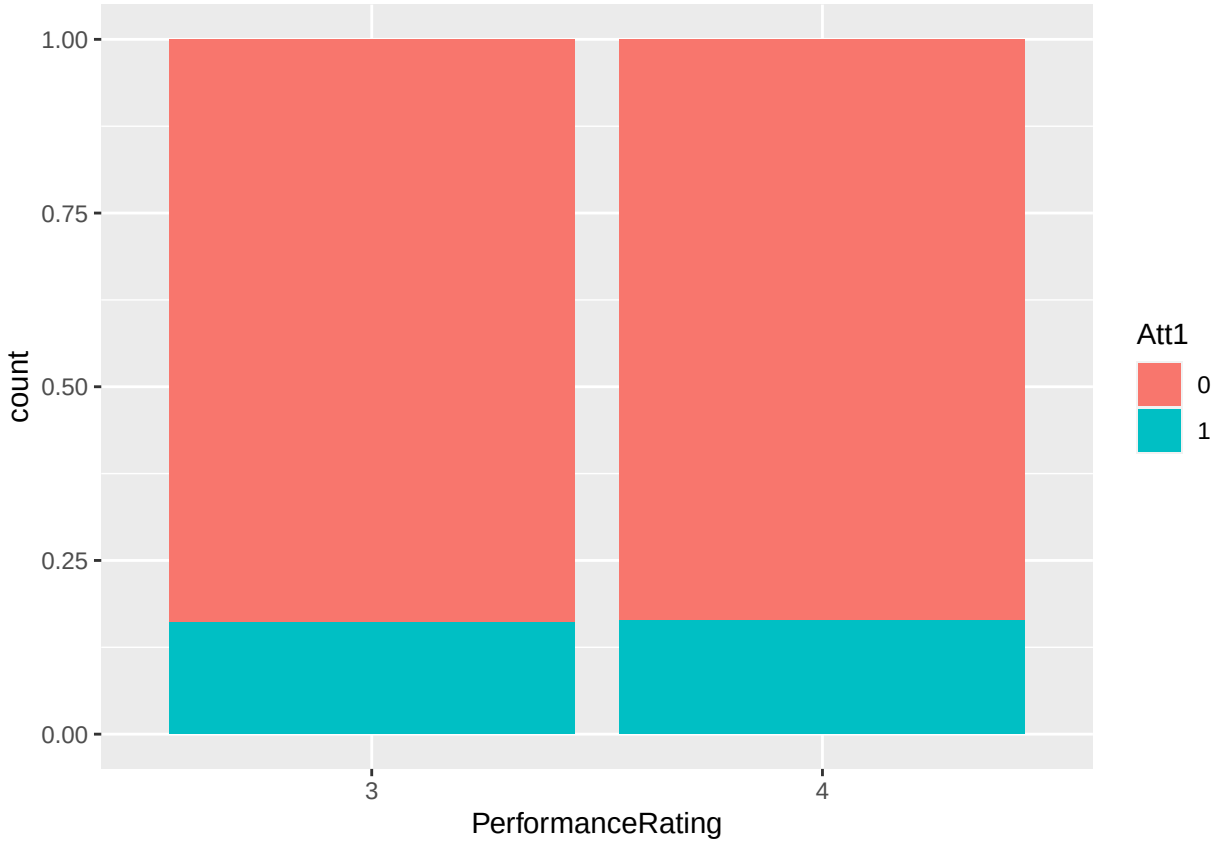

1.00
0.75
0.50
0.25
0.00
count
1
2
3
4
5
JobLevel
Att1
0
1

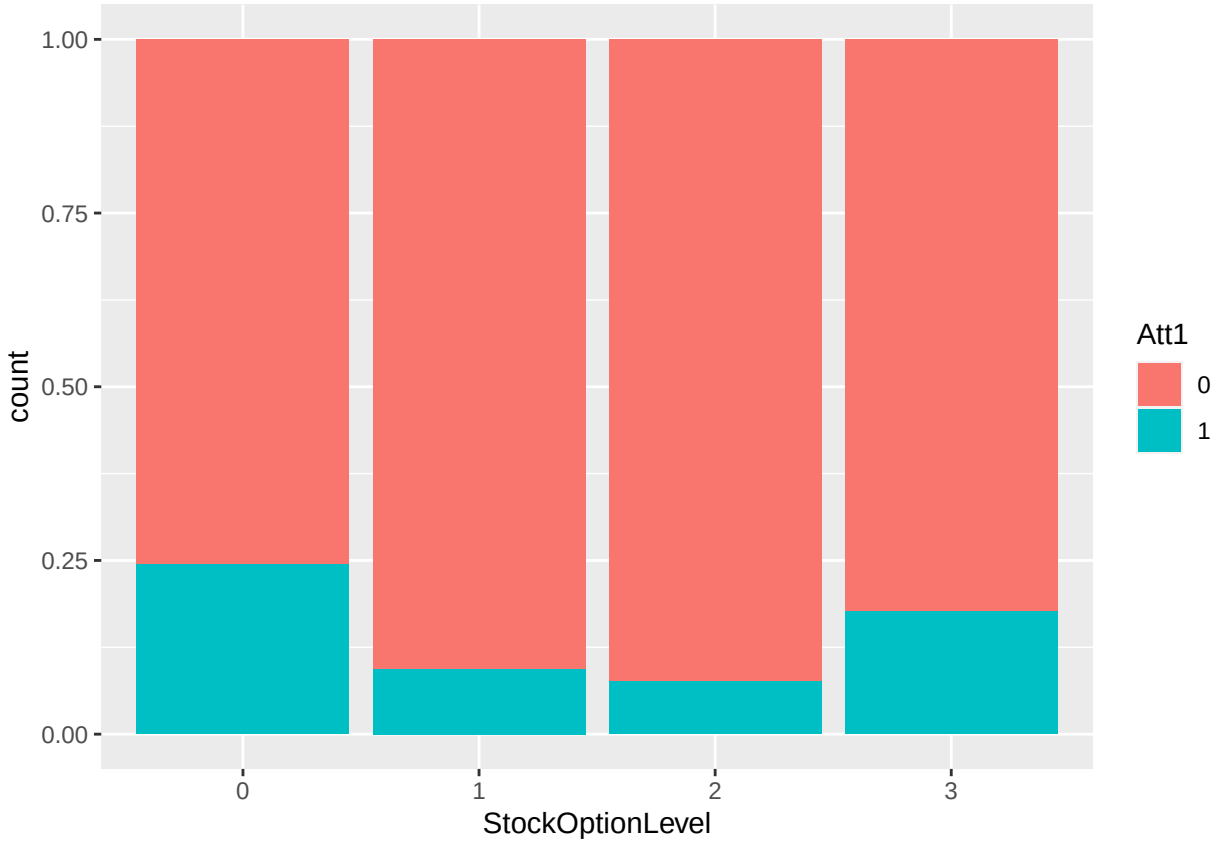

1.00
0.75
0.50
0.25
0.00
count
Healthcare Representative
Human Resources
Laboratory Technician
Manager
Manufacturing Director
Research Director
Research Scientist
Sales Executive
Sales Representative
JobRole
Att1
0
1

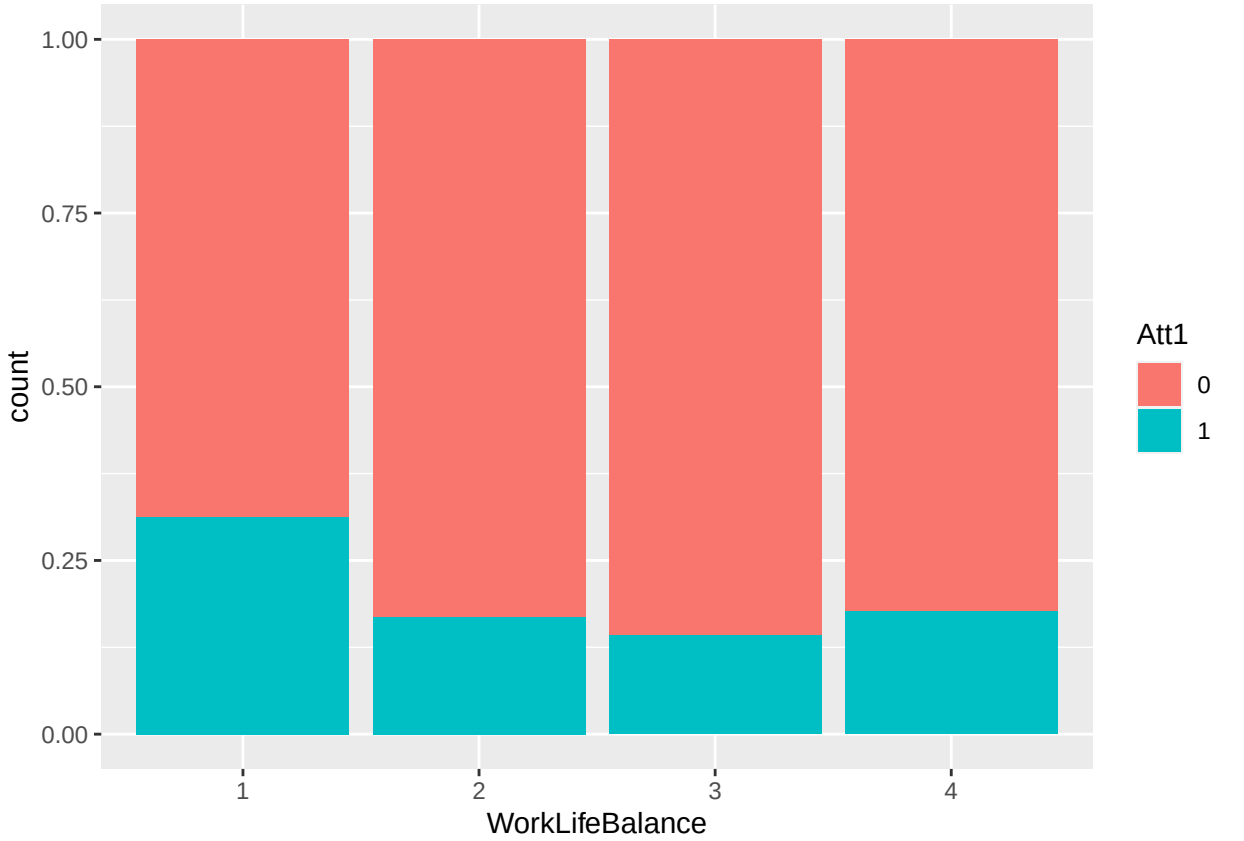

1.00
0.75
0.50
0.25
0.00
count
1
2
3
4
JobSatisfaction
Att1
0
1


1.00
0.75
0.50
0.25
0.00
count
Divorced
Married
Single
MaritalStatus
Att1
0
1


1.00
0.75
0.50
0.25
0.00
count
No
Yes
OverTime
Att1
0
1

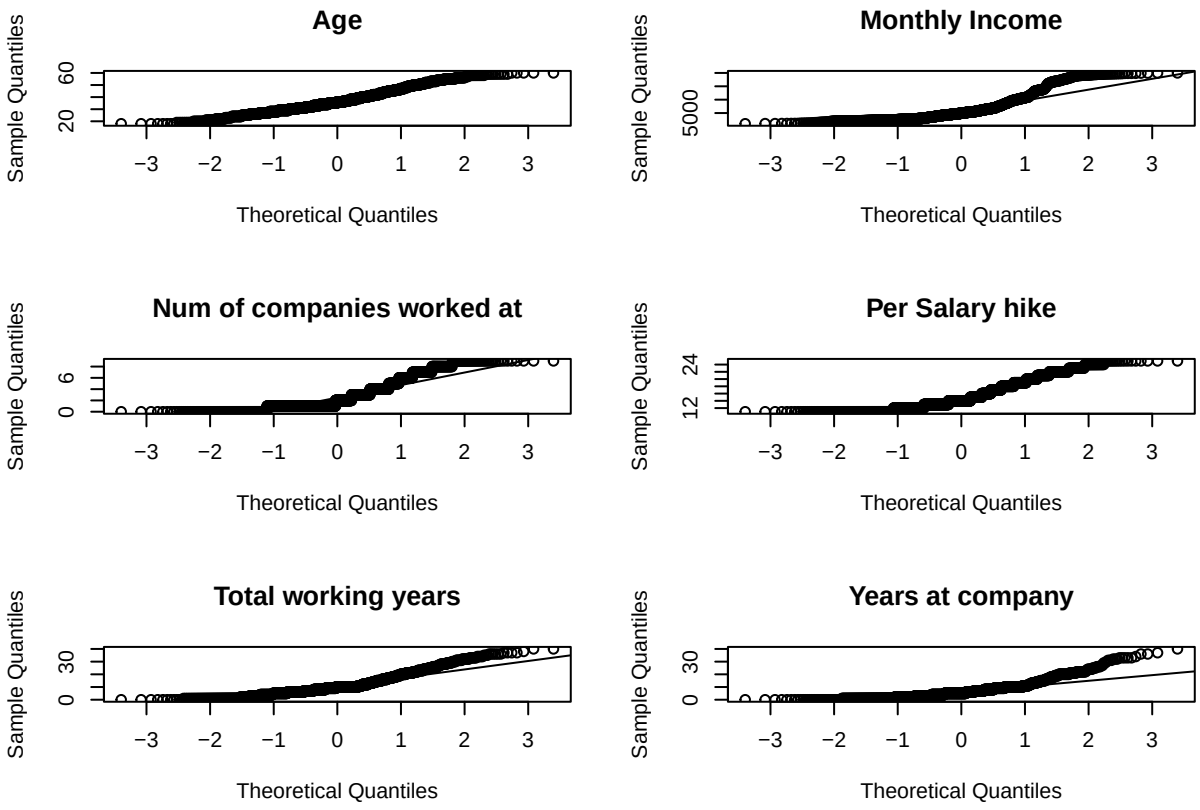

1.00
0.75
0.50
0.25
0.00
count
3
4
PerformanceRating
Att1
0
1


1.00
0.75
0.50
0.25
0.00
count
0
1
2
3
StockOptionLevel
Att1
0
1



Result- In these plots, we can see certain variables which are likely to contribute more to potential models. For eg- more frequent travel and lower levels of environment satisfaction are correlated with greater attrition, more job involvement has lower attrition, greater attrition for married people than divorced people. We had expected to see more attrition among singles, which is also clear from this plot. Overtime is also correlated with a higher attrition level.

4. Checking for Normality of the numerical values to ensure there is no skewness in the distribution of the variables can which can effect our analysis



Result- Most of the data attributes seem almost normal. There is slight variations, but nothing major that we should be concerned about

III. Model building using logistics regression

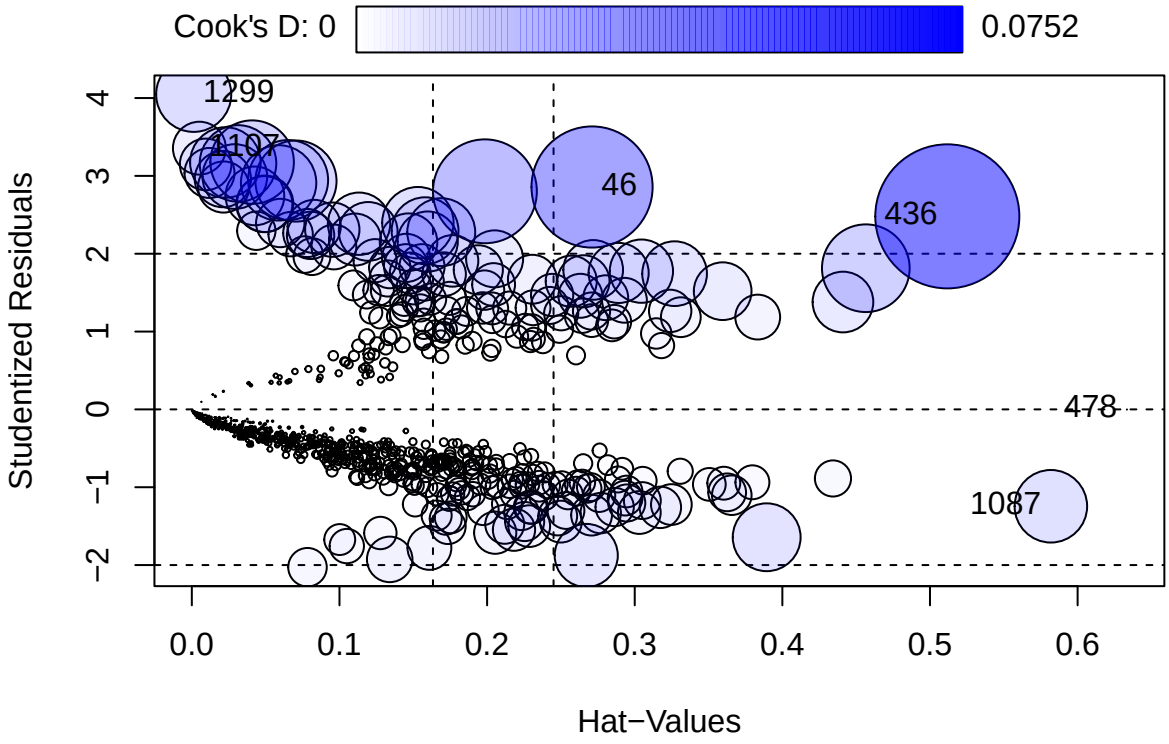
1. Splitting the data into test and train (taking into consideration the unbalanced data)

Result- the sample is distributed into train and test data with 75% as train and 25% as test

2. Full model with all variables

Result- First model with all variables, AIC- 749.37

3. Checking for outliers and leverage points in the data for remediation before we start the model selection process



Result- No high cook's d values in the table, therefore no further remediation steps needed

IV.Model selection

1. Step model of g1

Result- reduction in variables from 30 to 22, significant AIC reduction

2. Interaction model- Looking at corr matrix, we observe there are some high ccorrelations between YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager, therefore we fit an interaction model here

Result- Significant reduction in AIC value

3. we run the step function on the above model

Result- Further reduction in the AIC value

4. Anova test on g1.step and g1.step.inter.step model to select which model performs better

Result- Since p- value is small so we reject null hypothesis and consider g1.step.inter.step to be a better fit. All variables are important in the model either independently or in interaction terms

V. Further remediation and refining the model

1. Collinearity check using 'vif' on the best fit model we have till now- g1.step.inter.step

Result- we see 'YearsAtCompany' variable has highest VIF value. Therefore fitting a model without YearsAtCompany

2. Fitting model g2 without 'YearsAtCompany' and running step function
3. Compare g2.step with g1.step.inter.step model

Result- Since p-value is small , therefore we reject null hypothesis and accept that g1.step.inter.step is a better fit model

4. Looking at the individual confidential interval of the coefficients with Bonferroni confidence interval for all variables of g1.step.inter.step model

Result- Checking all the non-interaction terms , we find Age+ BusinessTravel+ EducationField+ EnvironmentSatisfaction+ JobInvolvement+ JobSatisfaction+ OverTime+ StockOptionLevel+ YearsSinceLastPromotion does not contain zero.

5. Fitting the model with variables that do not contain zero in their confidence interval and running step function

VI Cross validation on train data

1. We perform k-fold Cross Validation of following models- g1.step ,g1.step.inter.step , g2.step ,g3.step

Result- 1st value for each model is raw cross-validation estimate of prediction error and second value is bias corrected estimate of prediction error. We see all the models errors rates are very close. However we select g1.step.inter.step as the best fiiting model as it has the lowest error %

VII. Confusion Matrix and model accuracy on test data

Using test data, we predict the results via confusion matrix

Result- We get 67.7% accuracy in predicting attrition (21 attrition values were correctly predicted out of total 31 values). We see an overall accuracy of of 86.1 %