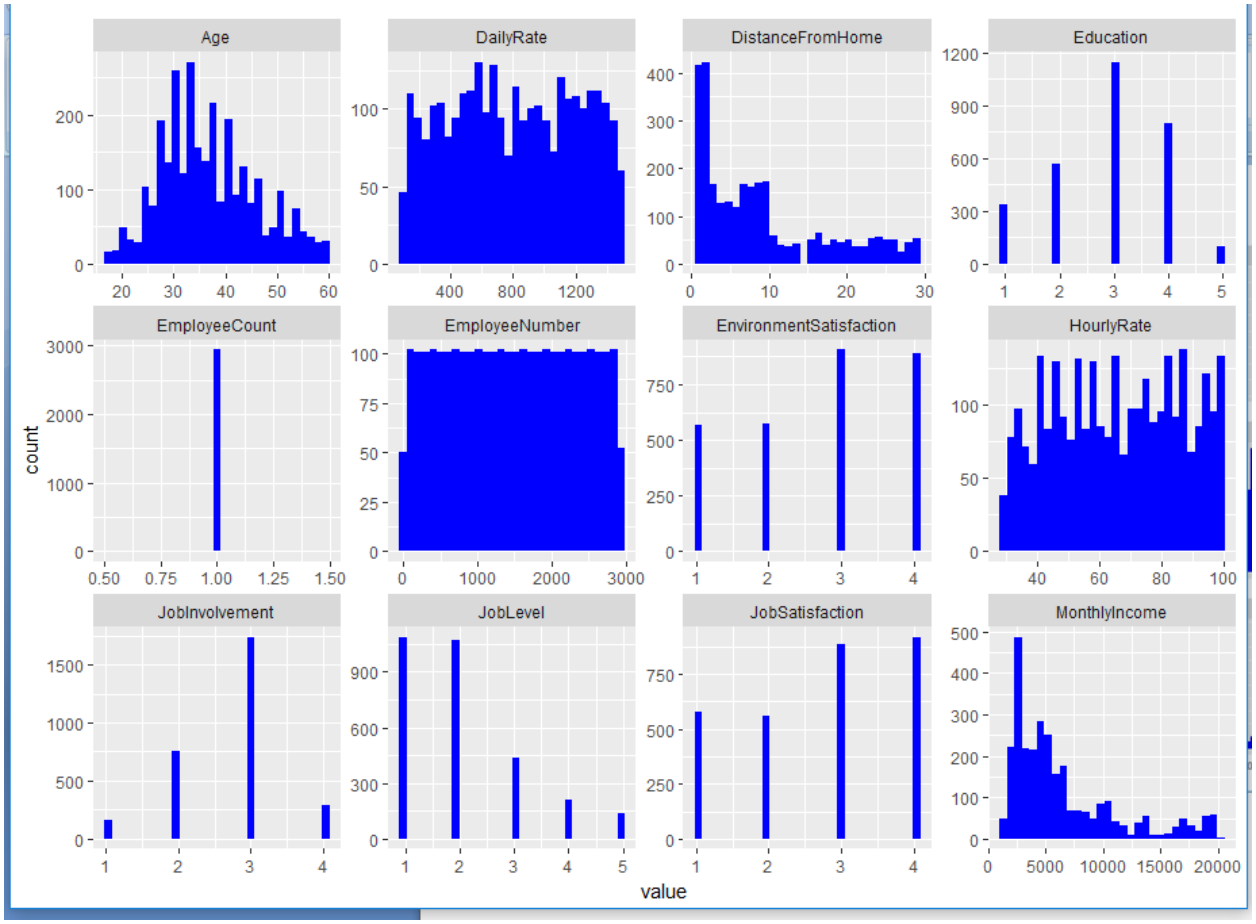
#####PREDICTING EMPLOYEE ATTRITION#################

EDA:

It helps us in separating powerful predictor variable from weak predictor variables.
Informs about variables which are highly correlated with each other, that means they are not independent and impact each other either positively or negatively.

1.UNIVARIATE ANALYSIS : **One variable at a time -** Histograms

Visualizing Each Numerical data

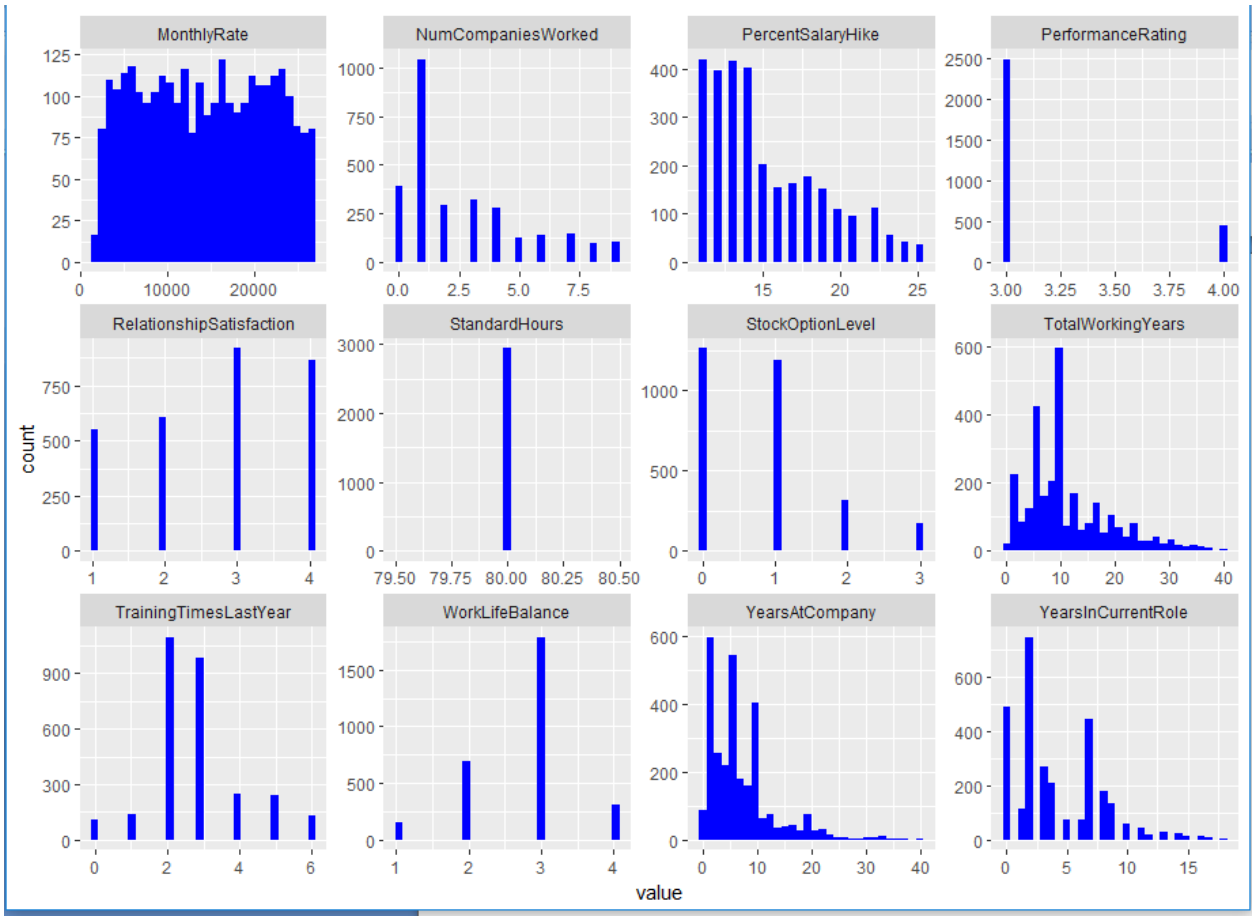


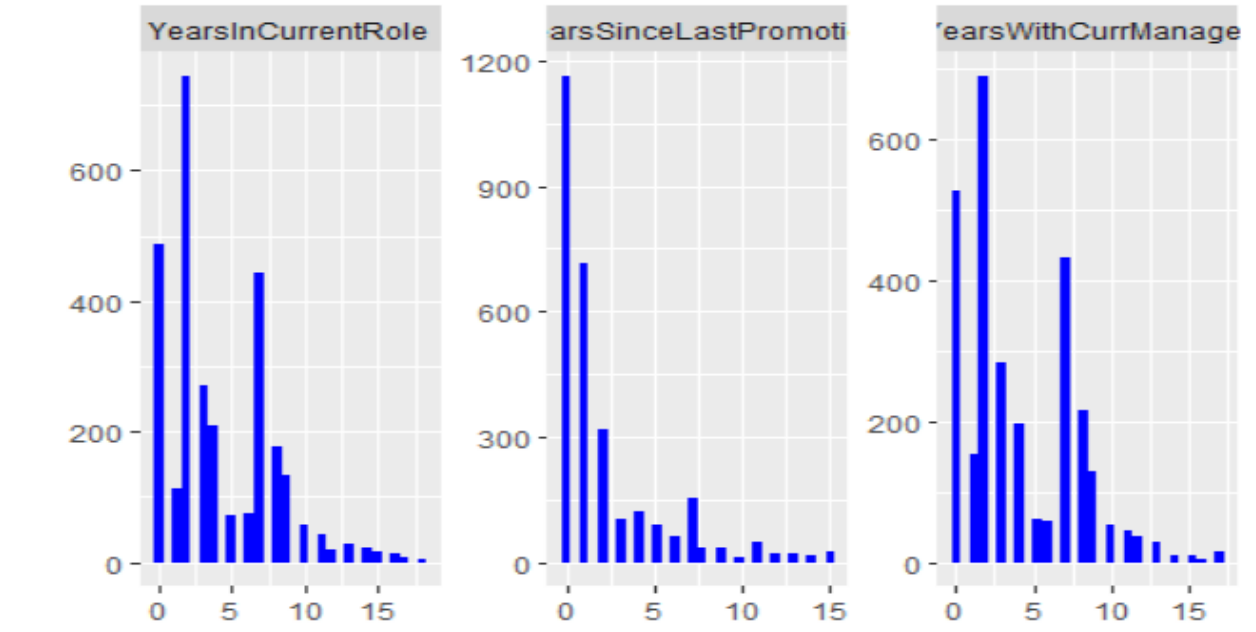
1.For all the Rows, Employee count is equal to 1,Standard hours is equal to 80,Over18 is Y which means all are above 18. Removing 4 variables including Employee Number.

2.Many of the Employees Education is 3 and Job Level is 1 and 2.Monthly Income of many of the Employees is below 2500.

3.For Many of Employees ,this is the Second company (Number of Companies Worked is more for 1) and total working years is less than 10years for most of the Employees and Years at this company is less than 5 years.

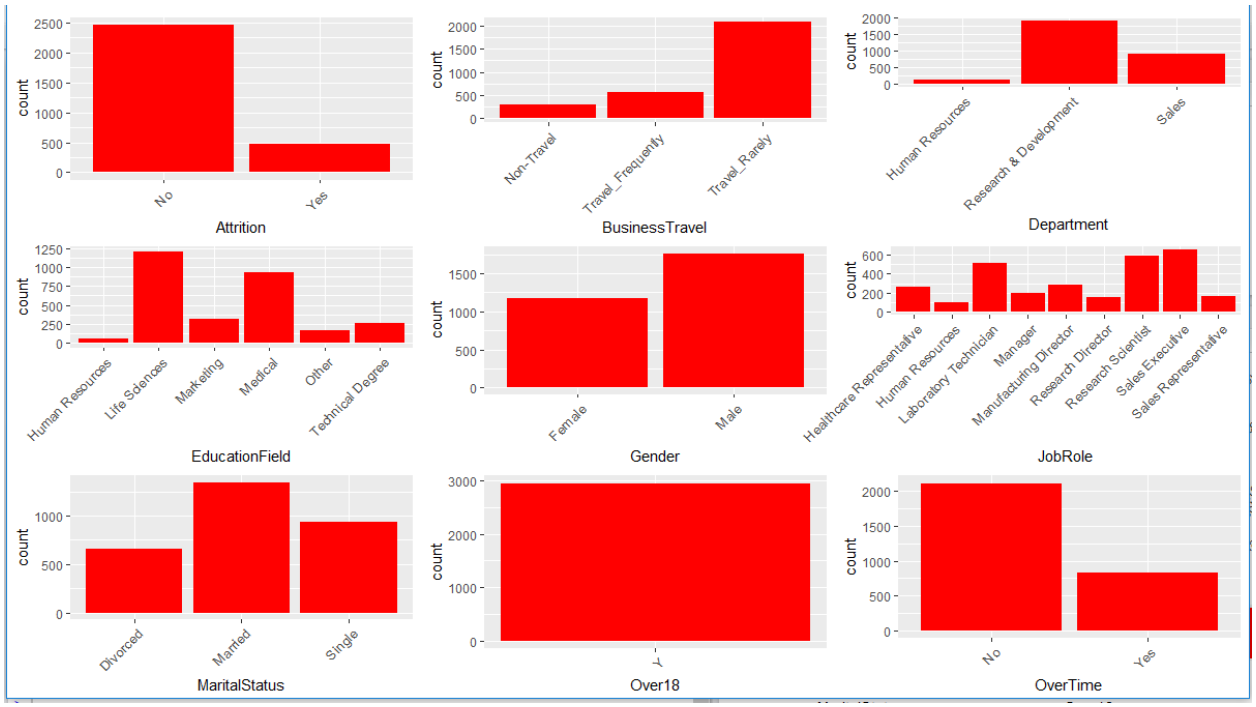
4.Most of the Employees Performance Rating is 3 and many of Employee's Percent Salary Hike is less than 15%.


MonthlyRate
NumCompaniesWorked
PercentSalaryHike
PerformanceRating
RelationshipSatisfaction
StandardHours
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole
count
value


YearsInCurrentRole
arsSinceLastPromoti
earsWithCurrManage

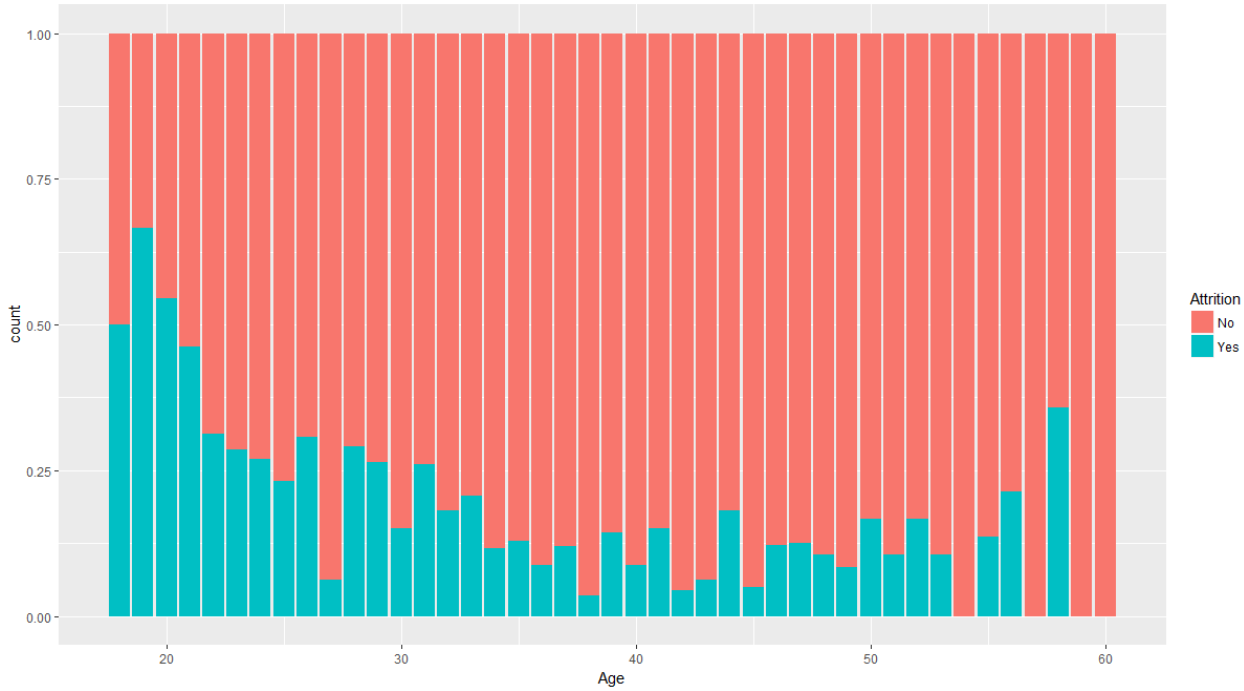
Visualizing Each Categorical Variable:

1. Many Employees Travel rarely to On-site.Attrition Rate is 20%.
2. Many of the Employees have come from LifeSciences, Medical and Marketing field and most of them are into R&D Department and Job roles are many employees is Research Scientist,Research Director,Lab Technician and Sales Executive.



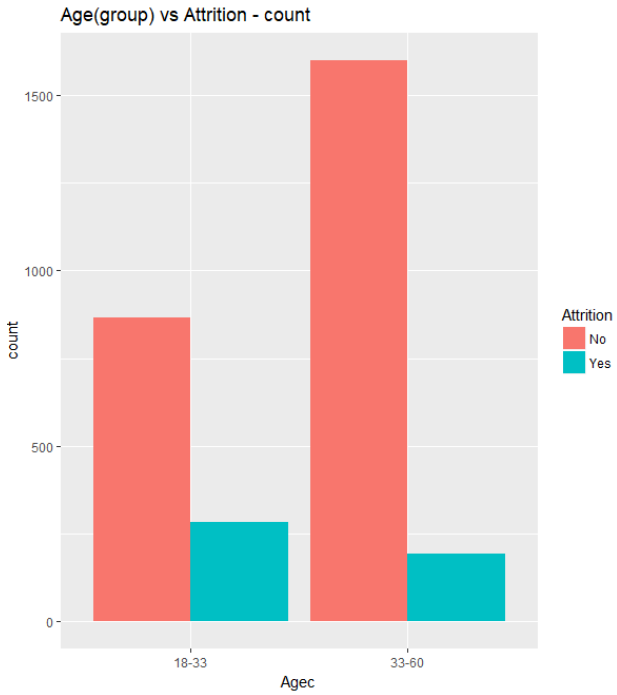
Bivariate Analysis: **Two variables at a time**

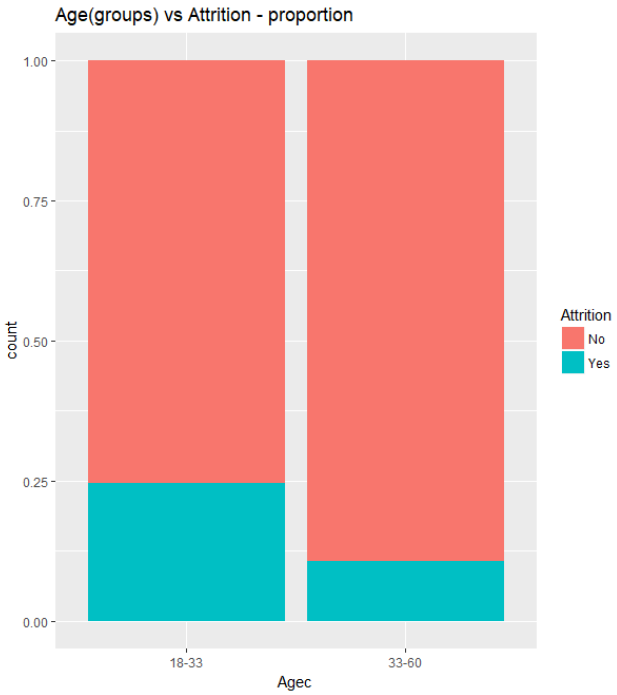
Age vs Attrition:



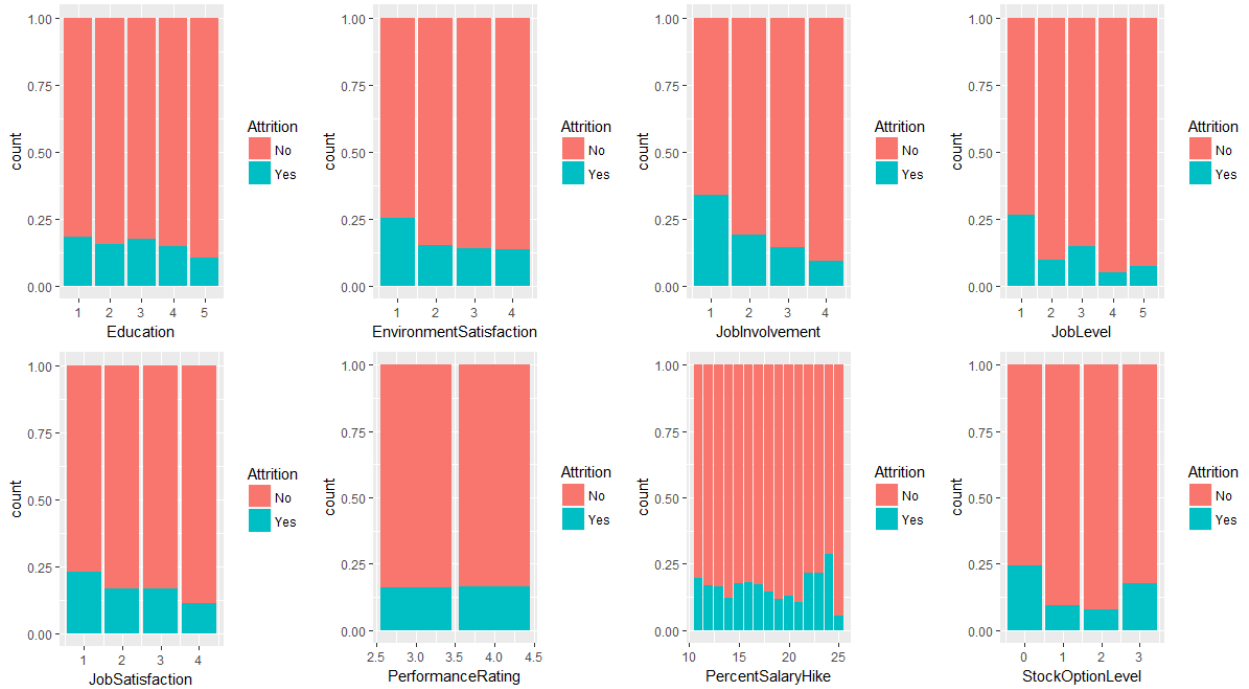
We can infer that Attrition is more between Age 18 and 33 when compared to employees whose ageis more than 33.There is increase in attrition whose Age is greater than 50 as employees usually retire at age of 55.

Grouping Employees with Age between 18 and 33 and employees Age greater than 33.This has decreased the AIC value which is a good model

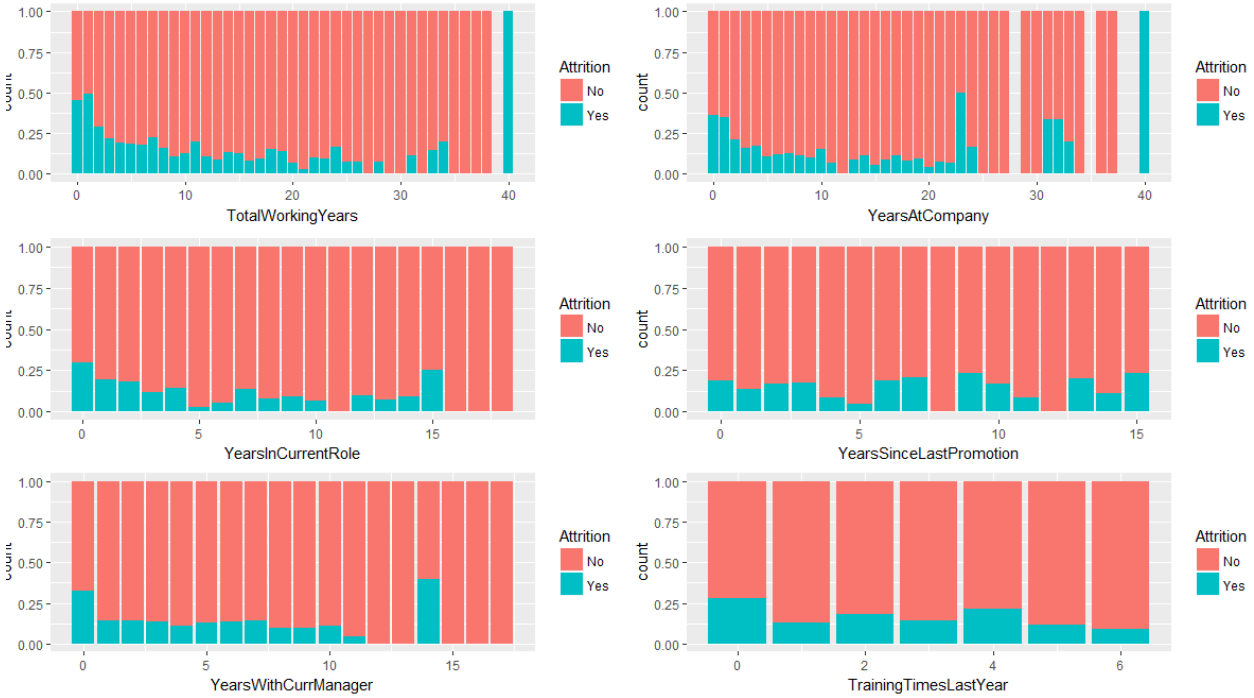




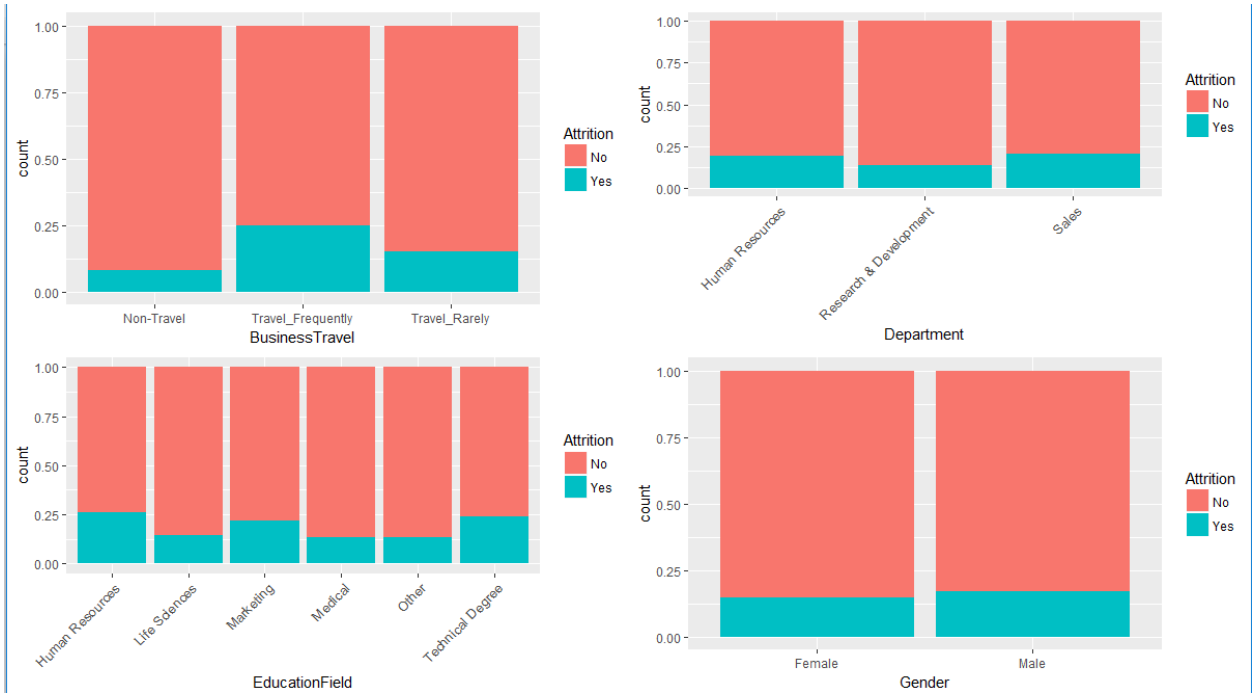
From this I can Infer that Education and Performance Rating doesnot impact on Employee Attrition



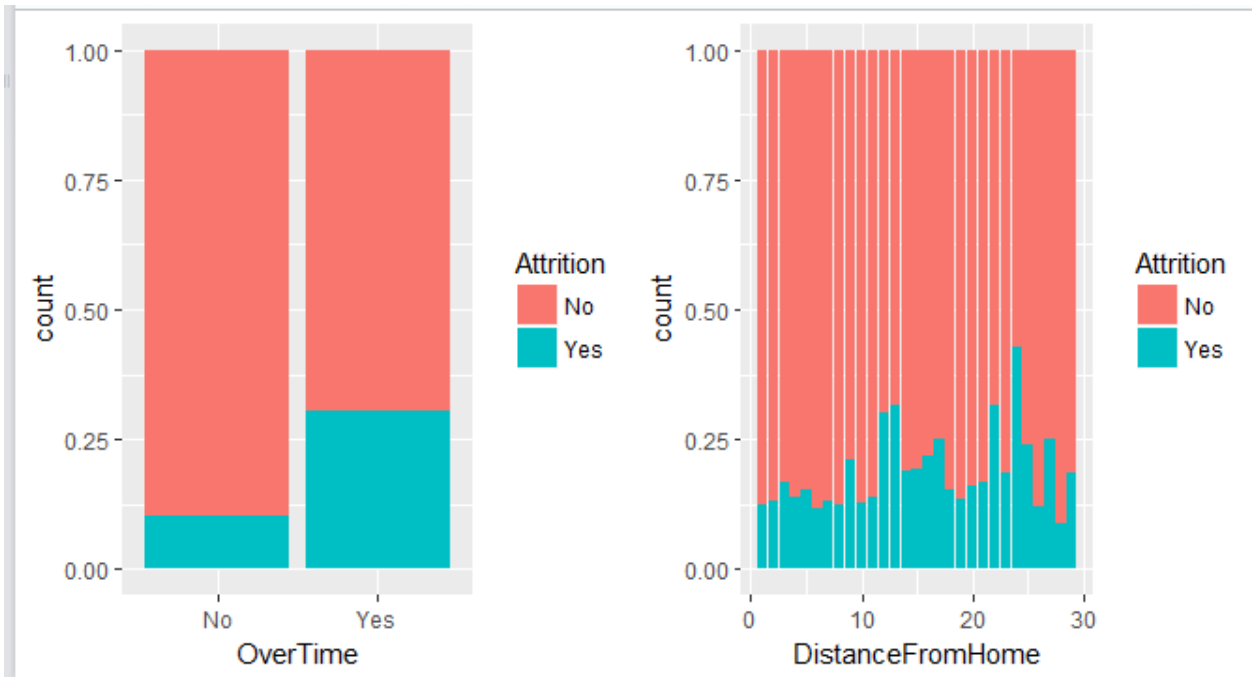
From this, We can infer that Training times since last year does not affect on attrition and rest of the variables are important.



From this ,We can infer that employees who travel frequently will leave company when company when compared to Non-Travellers.



More than 25% of Employees who work Overtime leave the company. More Employees leave the company if distance from home is greater than 12kms.



So, Education Field, Gender, Department ,Trainingtimessincelastyear, Performance rating and Education Field are not strong predictors and I will not be including these variables.

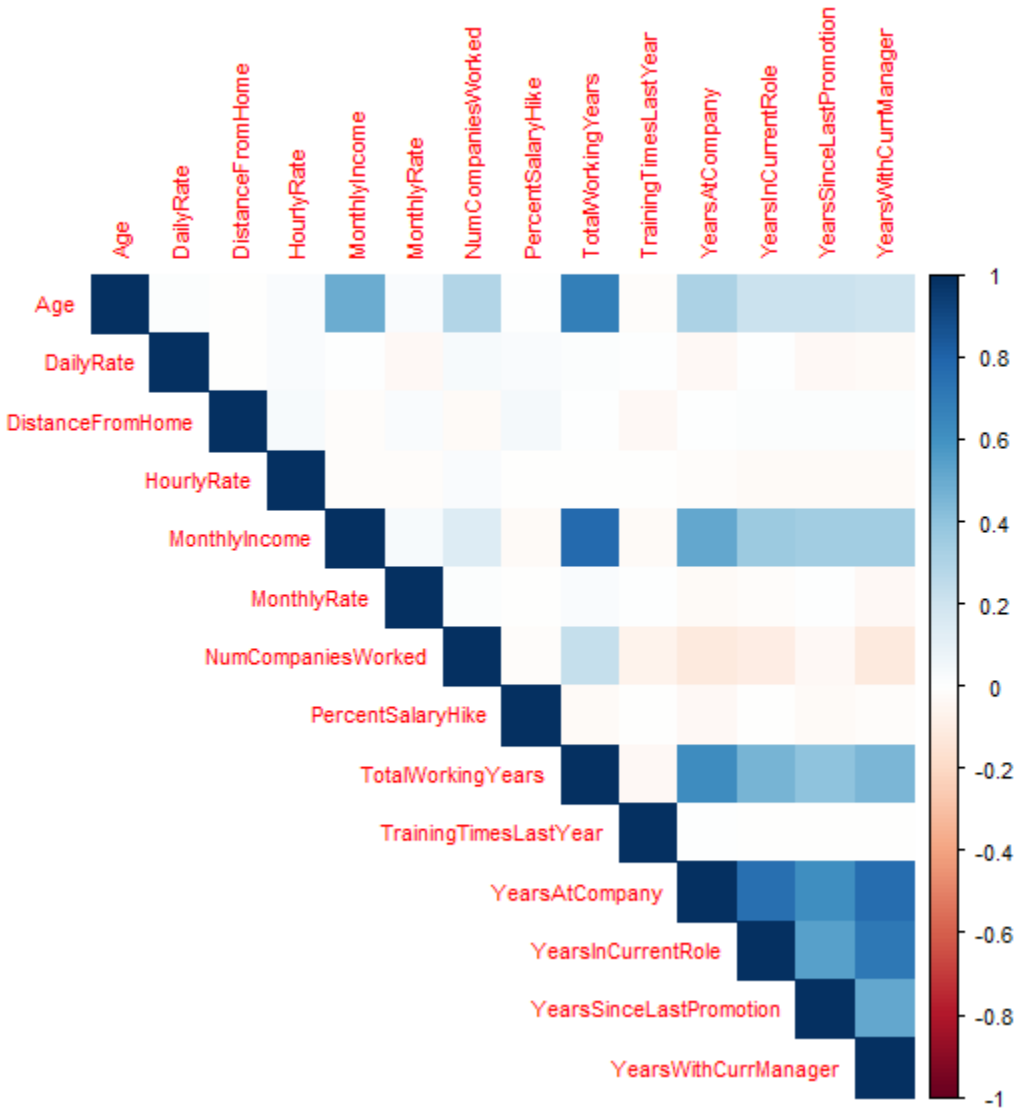
Checking if there is Multi-Co linearity - High Correlation between ndependent variables:

We can see that Age and TotalWorkingYears are highly correlated.(We know that Total working years is dependent on Age).
We can see that Monthly Income and TotalWorkingYears are highly correlated.

As Age Increases, Total working years increases and As Total working years keeps increasing, Salary will Increase. So ,I will not be considering these variables.
YearsAtCompany is correlated with YearsInCurrentRole, YearsWithCurrManager and TotalWorkingYears.So,I will not be considering YearsAtcompany.
I will not be considering YearsWithCurrManager as there is correlation with YearsInCurrentRole.

EDA has helped me to separate powerful predictor variable from weak predictor variables.

"Agec","BusinessTravel","OverTime","JobInvolvement","JobLevel","Attrition","JobSatisfaction","StockOptionLevel" "NumCompaniesWorked","StockOptionLevel","YearsAtCompany","JobRole", "DistanceFromHome","PercentSalaryHike", "RelationshipSatisfaction","WorkLifeBalance"



Built a Logistic Regression model and Accuracy is 0.90.