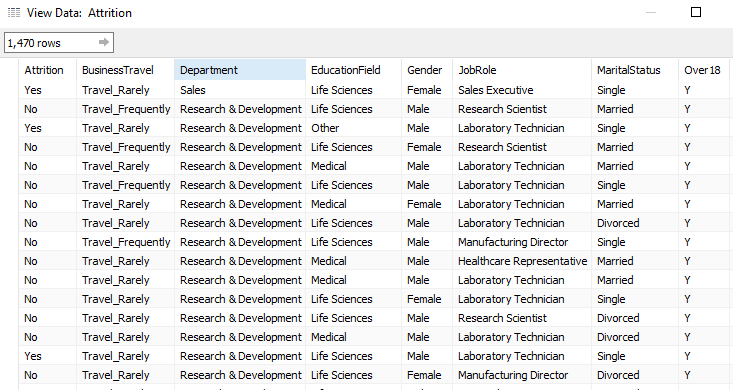
**Data Science Group Project**

# Data set description

* The ‘Attrition’ data set contains 1470 observations and 35 variables

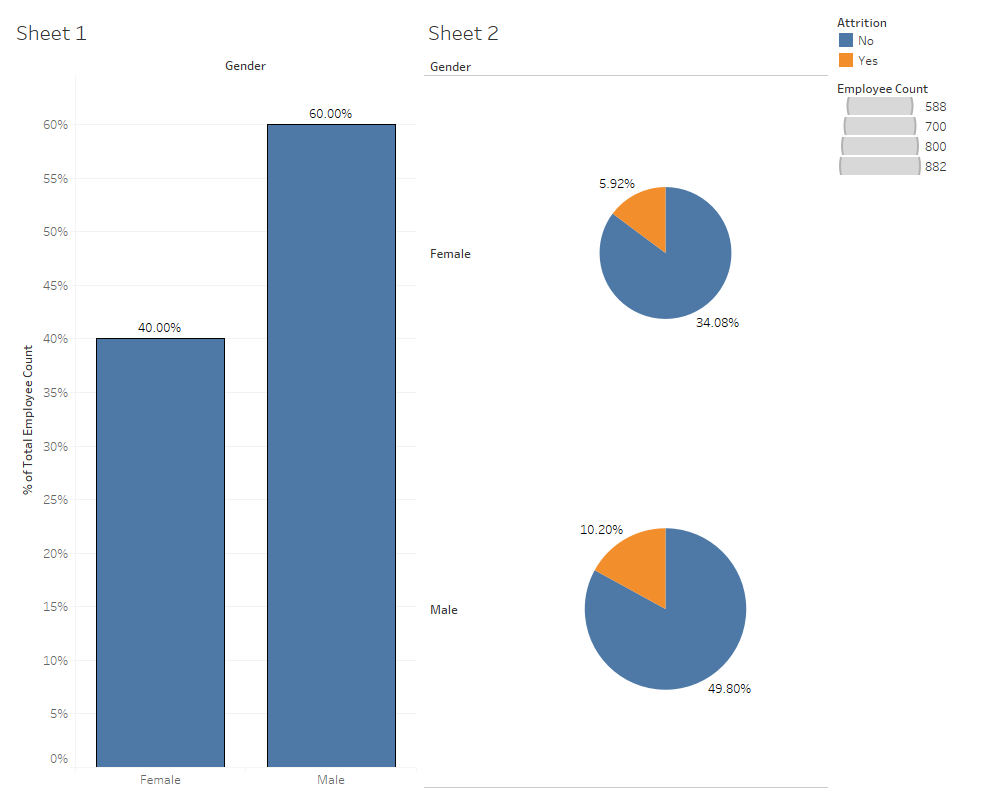


# Objective

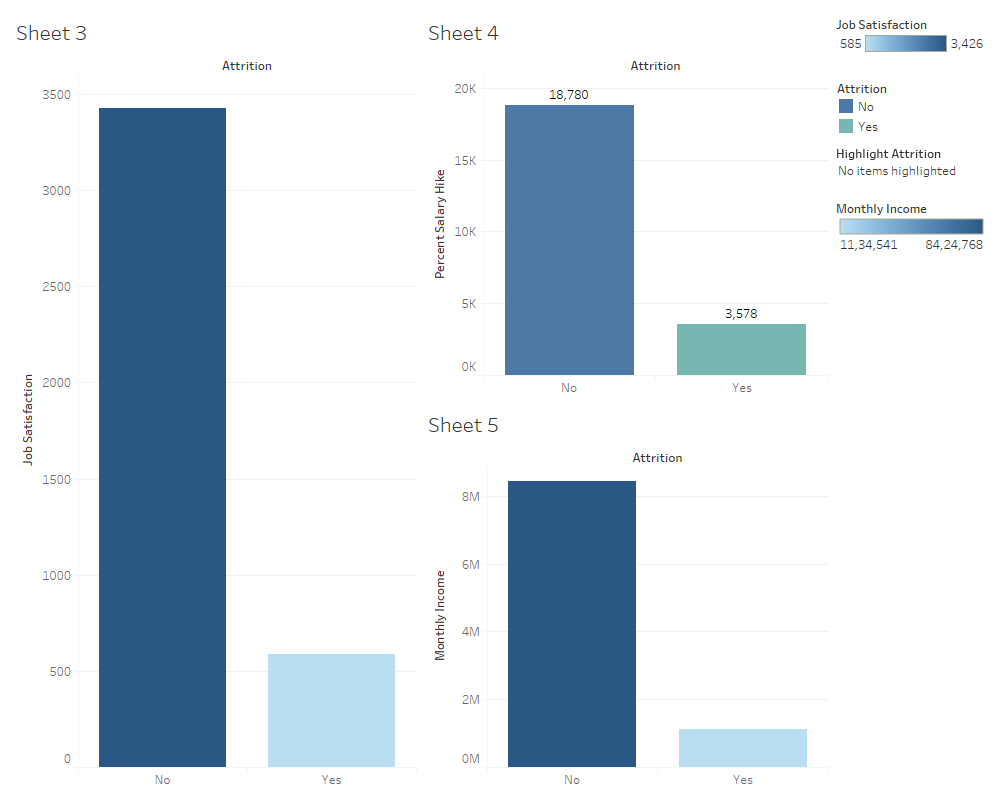
* To predict which employee is likely to churn



* We see that 60% of employees are male

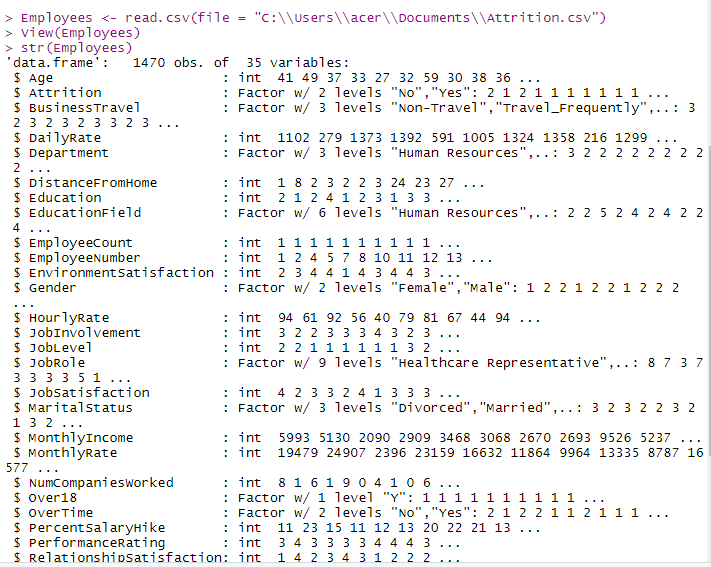


* Below we see the impact of different variables like ‘Job Satisfaction’ , ‘monthly income’ and ‘Salary hike’ on Attrition.



# Analysis in R

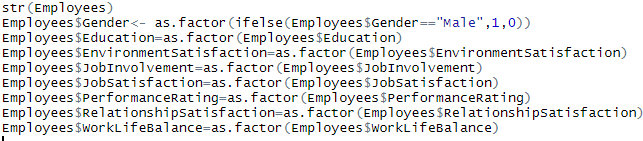
* Importing and structure of data



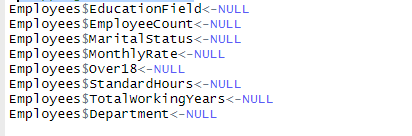
# Converting ‘Y’ variable into Factor

* Employees$Attrition<- as.factor(ifelse(Employees$Attrition=="Yes",1,0))

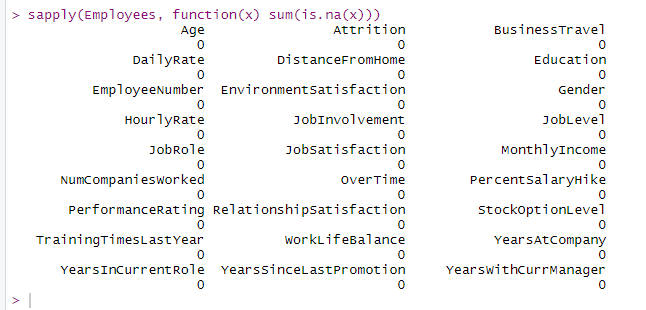
# Converting ‘X’ variables from integer to categorical or Factor variable



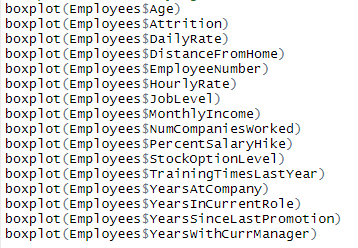
# Dropping the unwanted variables



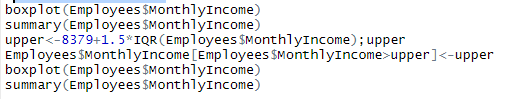
# Checking for missing values (no missing values found)



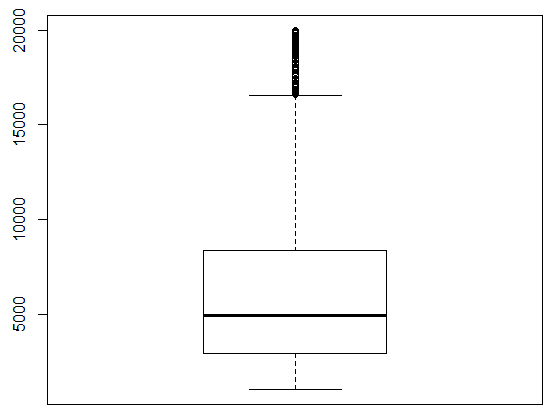
# Boxplot for identifying outlier



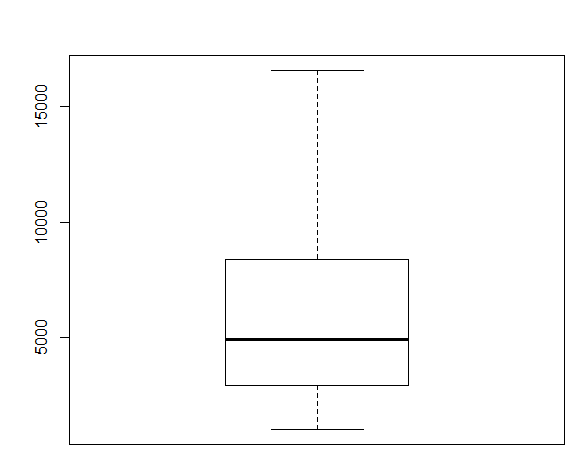
# Treatment of outliers



# Before Treatment

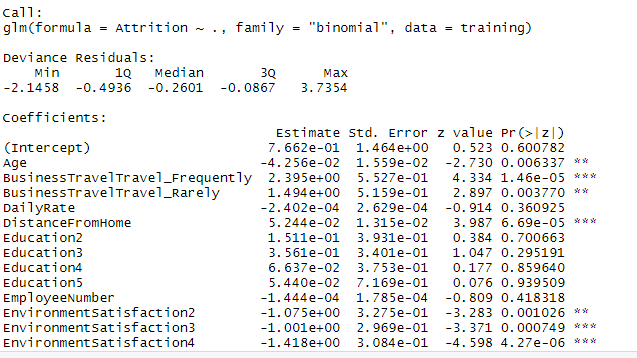


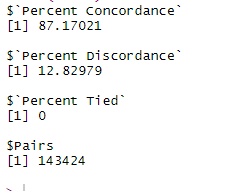
# After Treatment



Multiple Logistics regression (Logistic regression is a regression model where the dependent variable is Categorical)

* Start with data partition, create Training and testing data.
* Build model
* Check summary and look for significant variables
* Check model accuracy using ‘Acc’ function (87.17021 % Concordance in our Case)
* Run model on test data, accuracy - 86.75396 %
* check multicollinearity
* VIF values are <5 thus no multicolinearity
* Finding odds ratio
* Check confusion matrix
* Check AUC





**Confusion Matrix and Statistics**

Reference

Prediction 0 1

0 367 2

1 54 17

