HR Analytics Case Study

Group Name: Datalogists

1. Indira Kumari Saraswatula
2. Samadunnisha Sheik
3. Prudvi Bilakanti
4. Sampath Kanduri

Problem statement | Objective | Goal

Problem Statement:

XYZ company is seeing attrition of 15% every year and backfilling the positions to meet the head count of 4000. effecting there project deliverables due 15 % of its employees leave the company every year. The attrition is impacting the project timelines, increased recruitment and training costs of acquired resources

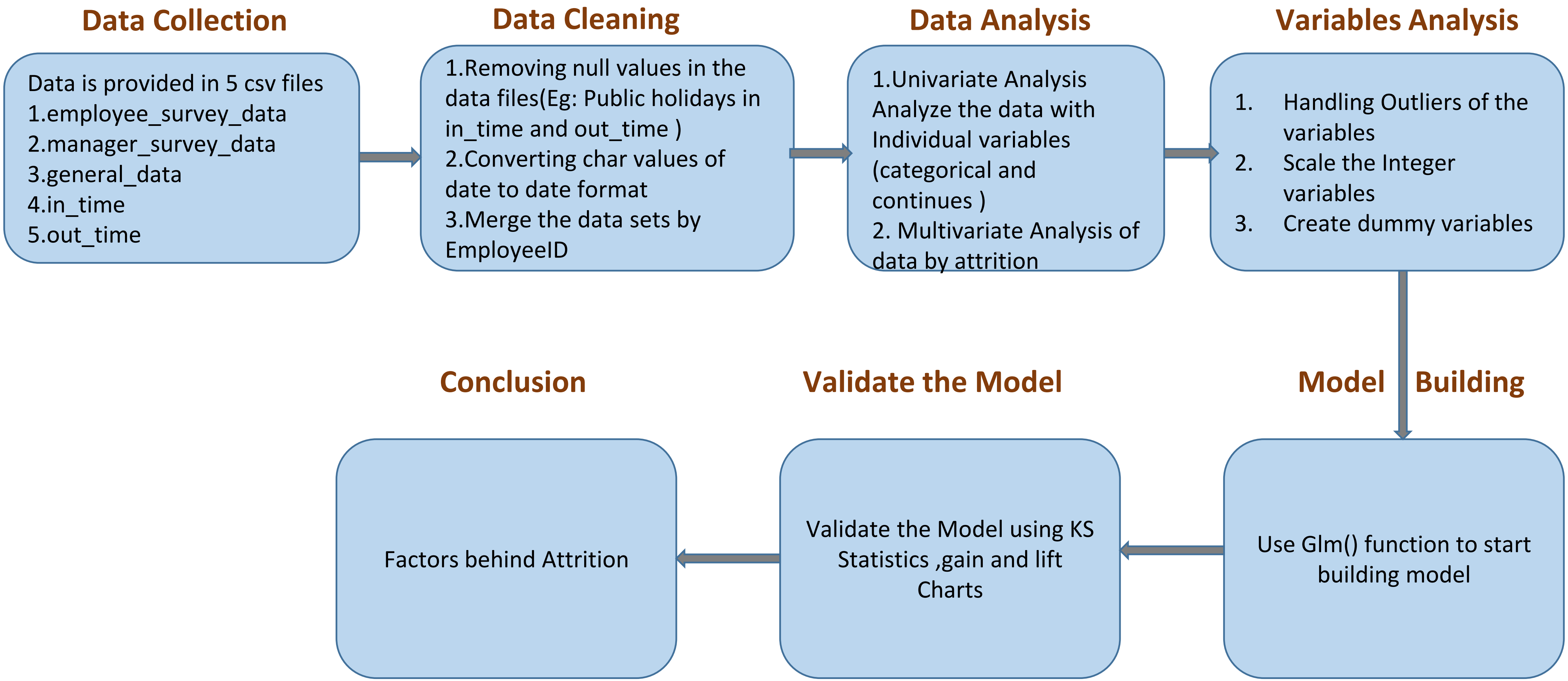
Objective:

Identify the factors contributing to the attrition and suggest the changes XYZ to bring to their workplace to curb the attrition

Goal:

Model the **probability of attrition** using a **Logistic Regression** which helps the management to make changes and retain employees

Problem solving Methodology



Data Collection

Data is provided in 5 csv files
1.employee_survey_data
2.manager_survey_data
3.general_data
4.in_time
5.out_time

Data Cleaning

1.Removing null values in the data files(Eg: Public holidays in in_time and out_time)
2.Converting char values of date to date format
3.Merge the data sets by EmployeeID

Data Analysis

1.Univariate Analysis
Analyze the data with Individual variables (categorical and continues)
2. Multivariate Analysis of data by attrition

Variables Analysis

1. Handling Outliers of the variables
2. Scale the Integer variables
3. Create dummy variables

Model Building

Use Glm() function to start building model

Validate the Model

Validate the Model using KS Statistics ,gain and lift Charts

Conclusion

Factors behind Attrition

Data Cleaning and Preparation

Data Cleaning:

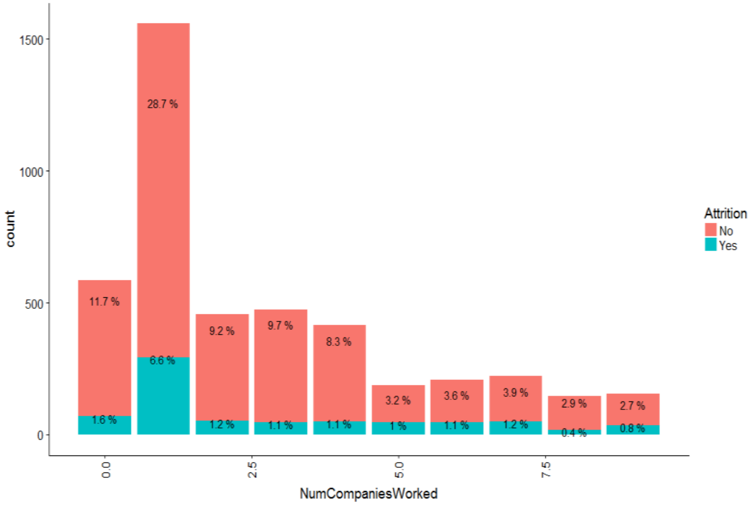
- Dropped columns not contributing to attrition/standard across – Standard hours, Employee Count, Over18
- Add missing column names – Employee ID in ïn-time”and “out-time”
- Remove ‘X’ prefixed to date column names
- Remove holiday columns from “in-time and out-time”
- Remove “ËmployeeID”

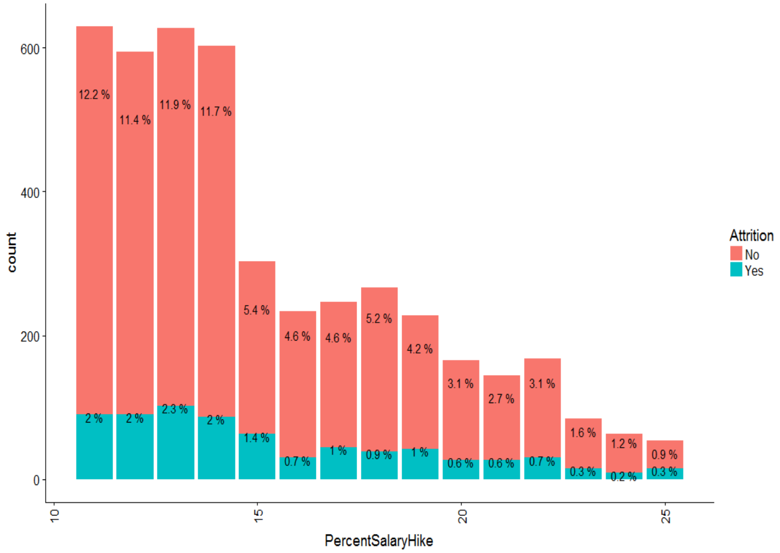
Data Preparation:

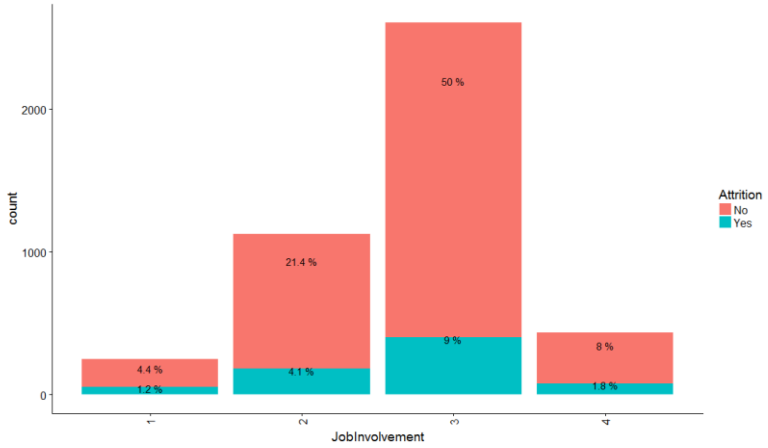
- Convert character variables to Factors to simplify analysis
- Count the no. of leaves and calculate the no. of hours worked
- Merge the files to make one master file

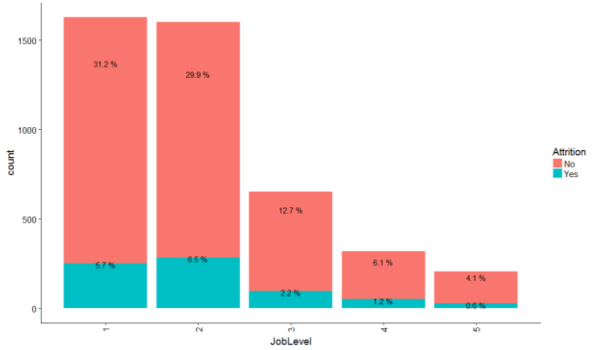
Univariate Analysis of Employee Data

- Analyze the data with categorical variable :- Below are the Plots of all categorical variables by Attrition

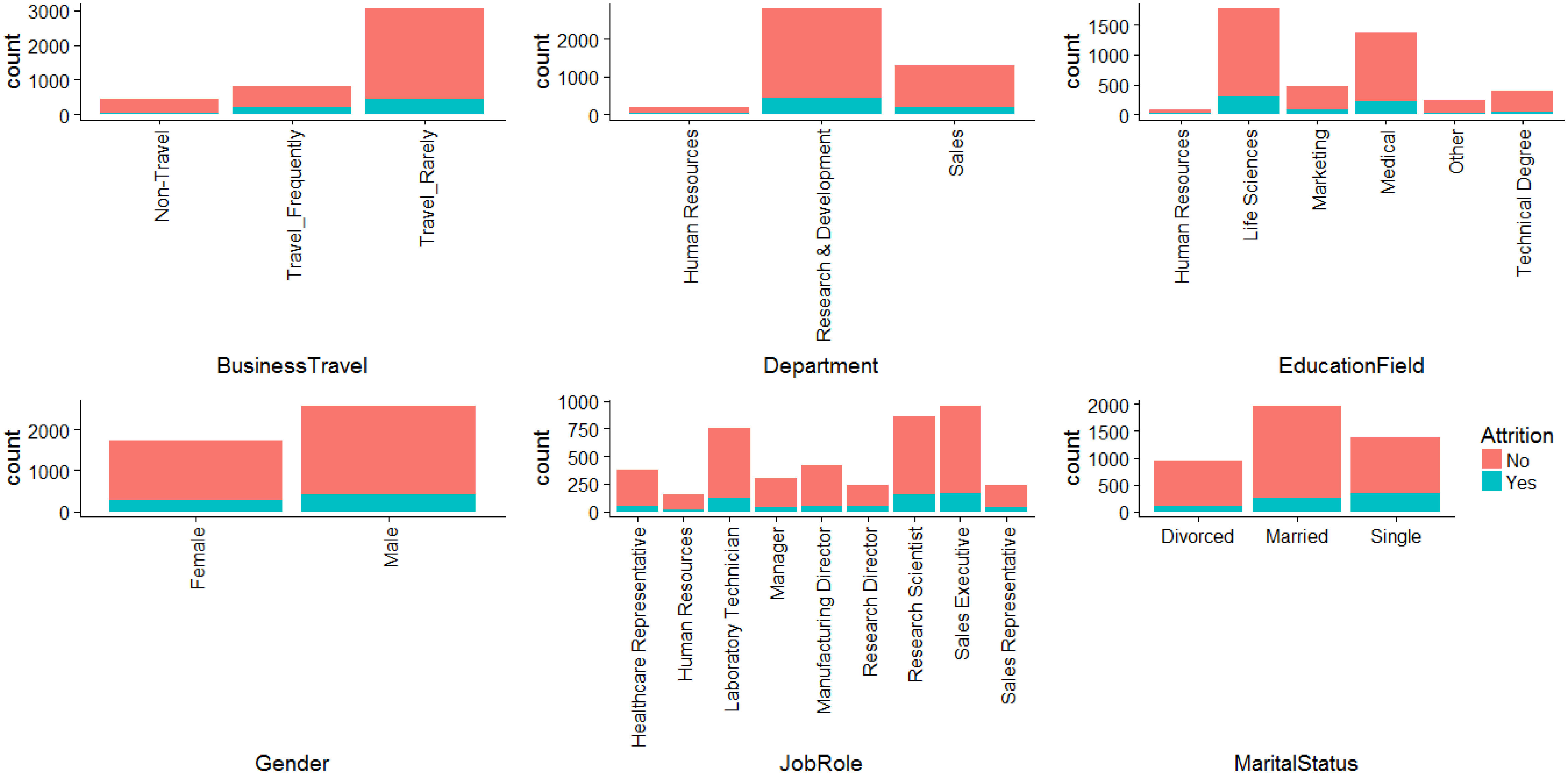




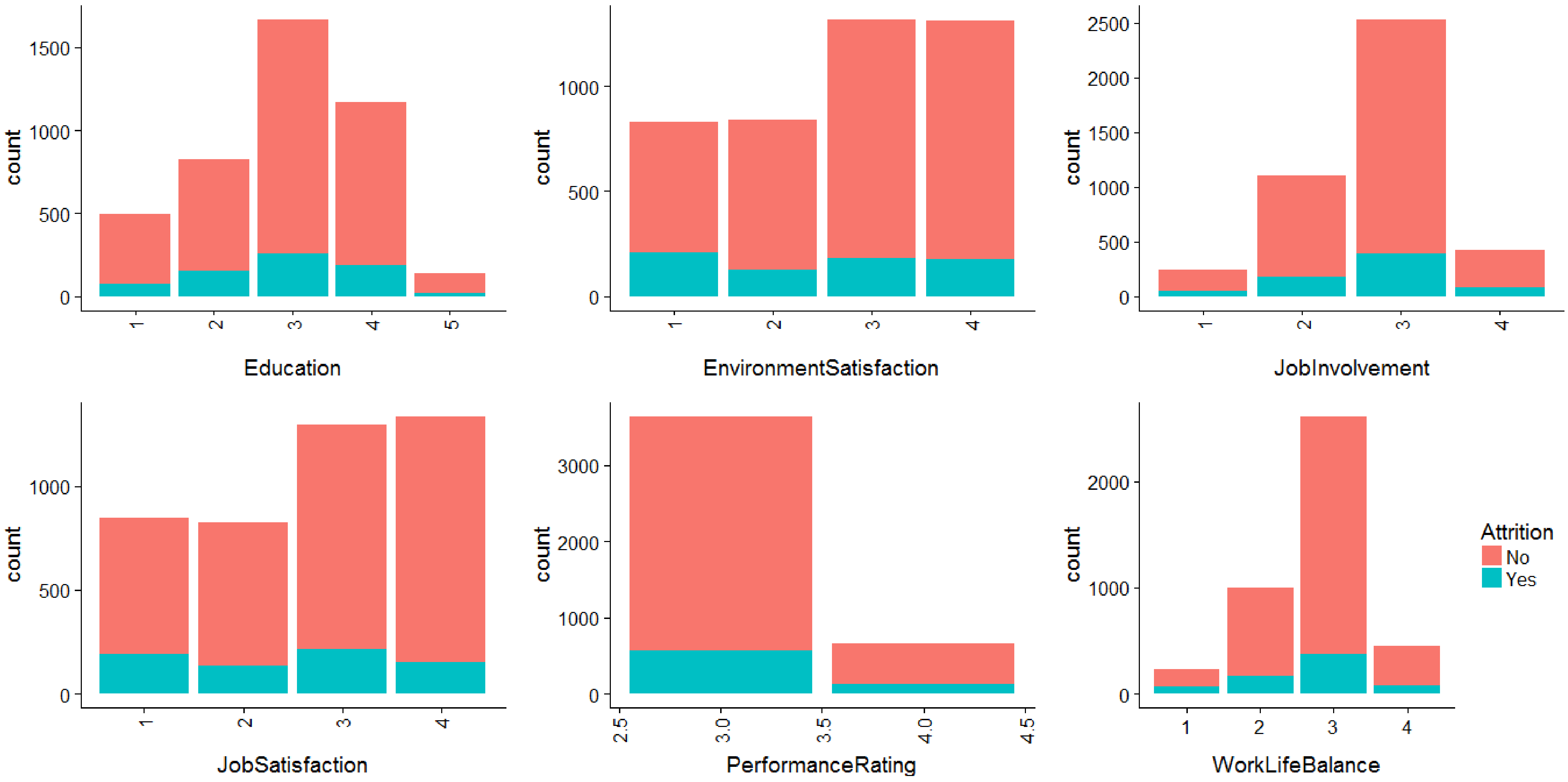




Univariate Analysis – Categorical Variables

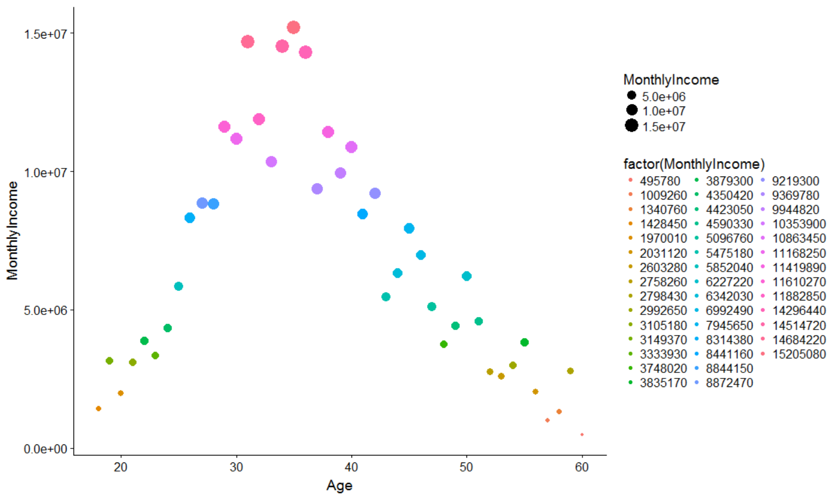


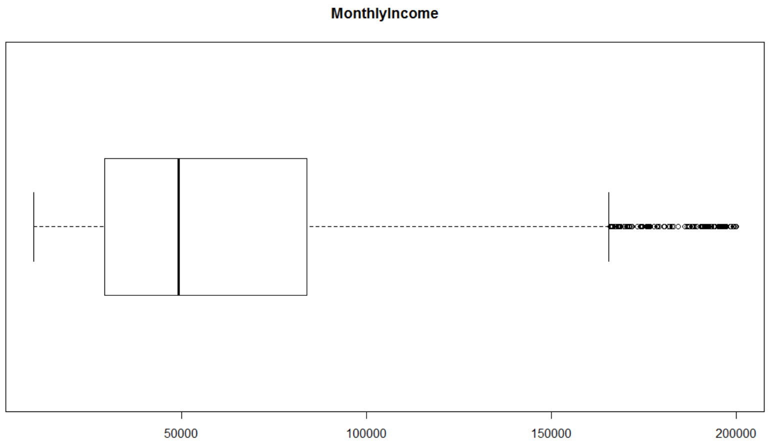
Univariate Analysis – Ordinal Variables



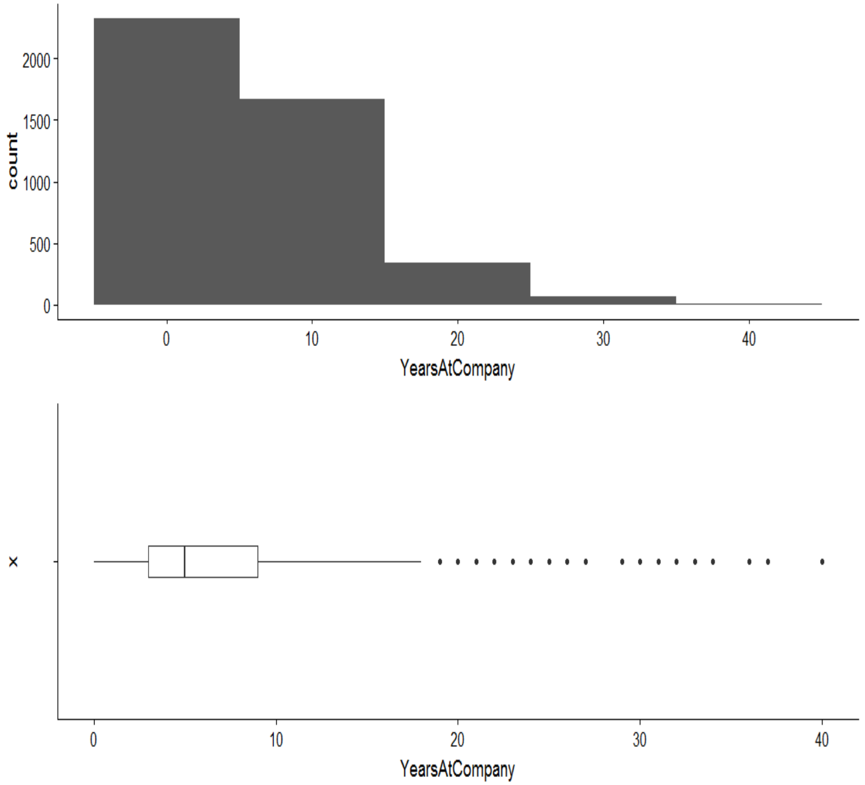
Univariate Analysis – Continues variable

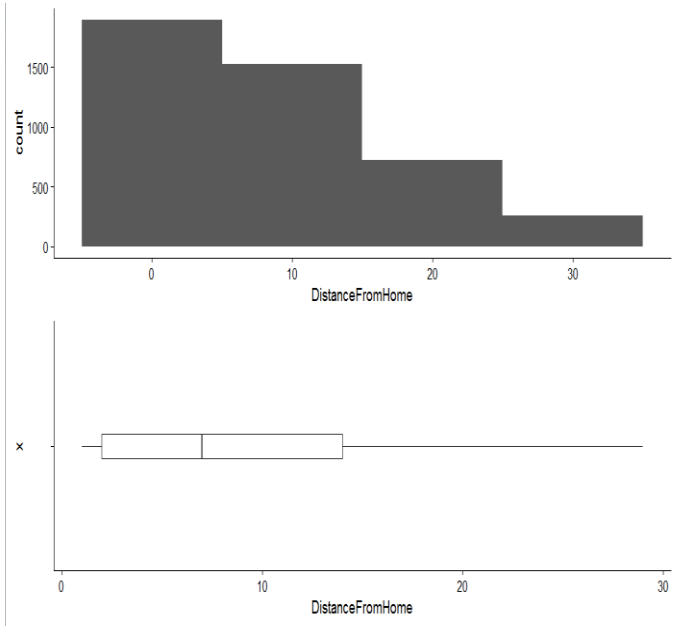
Used Histogram and Box plot for clear understanding of continuous variables like Monthly Income, YearsAtcompany





Highly skewed towards the right and having outliners

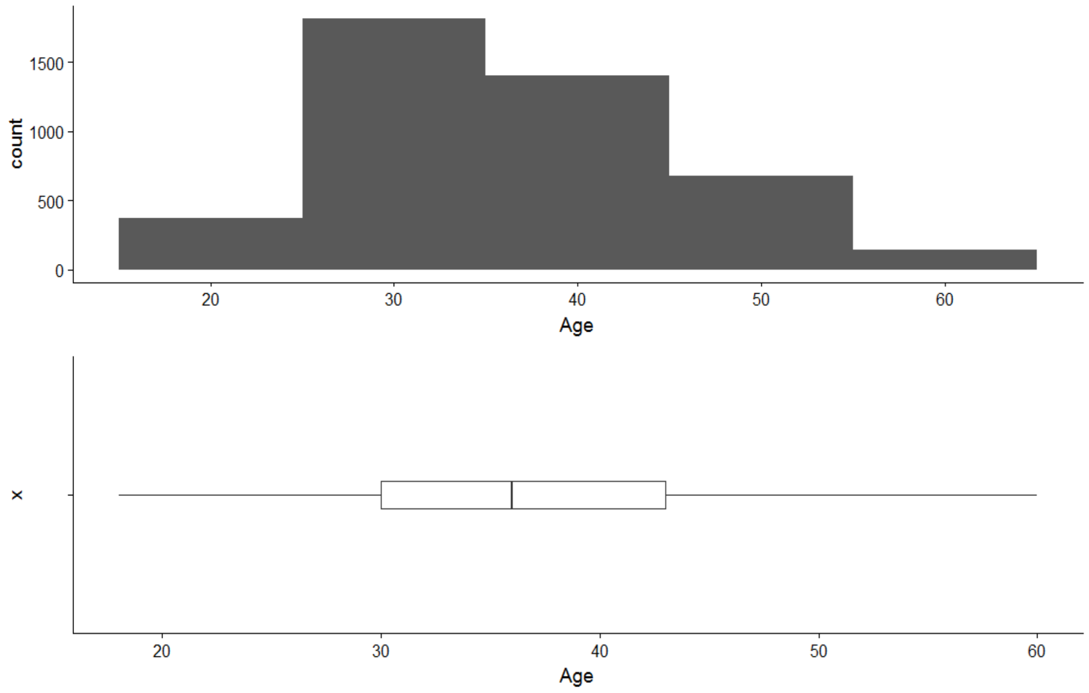


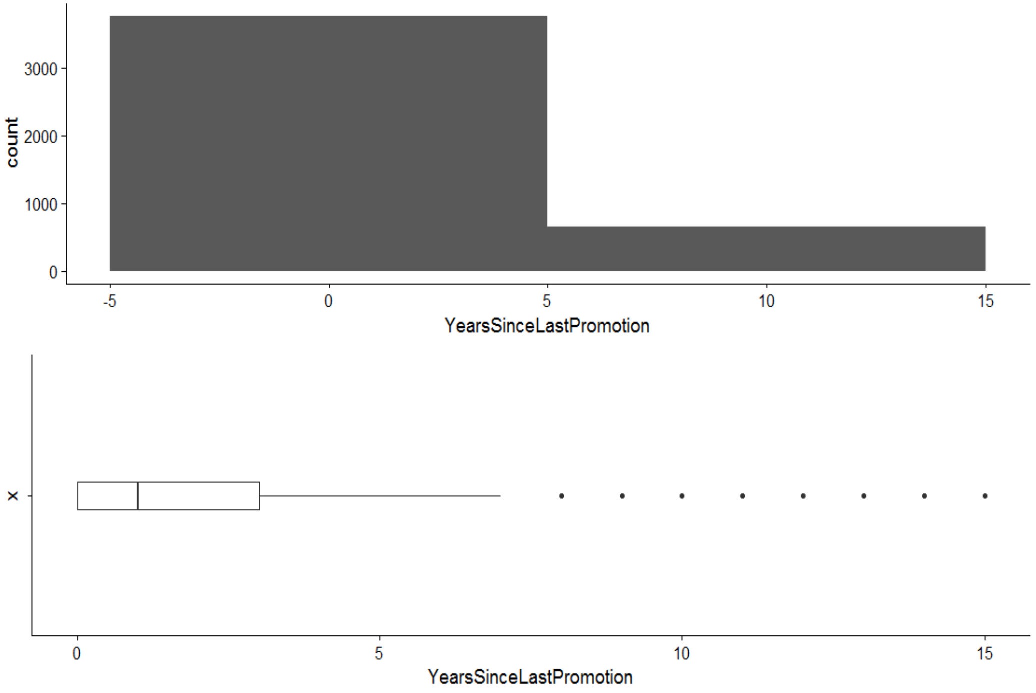


Highly skewed towards the right and having outliners

Univariate Analysis – Continues variable

Adding few more continues variables visualization





Highly skewed towards the right and having outliners

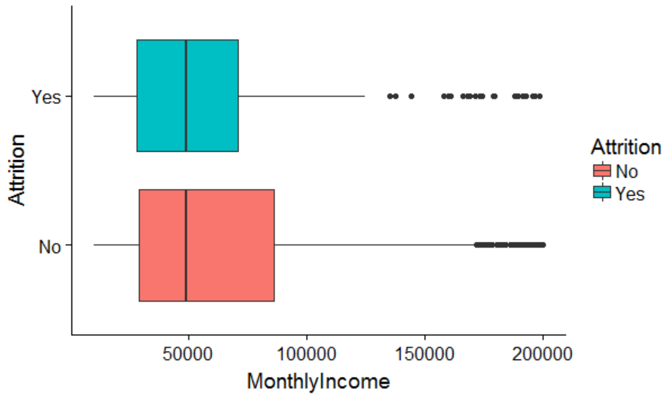
Highly skewed towards the right and having outliners

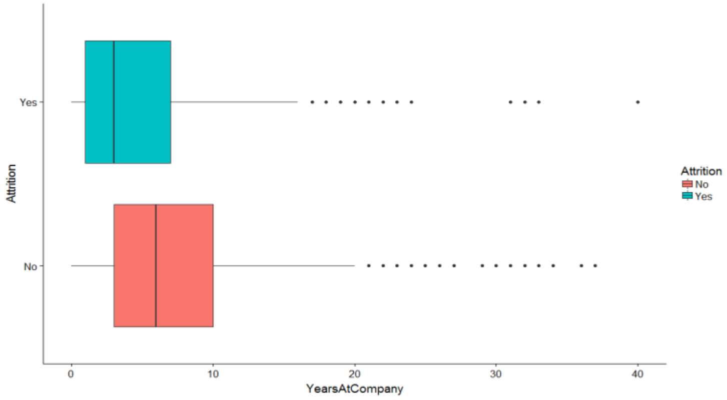
Observations from Univariate Analysis

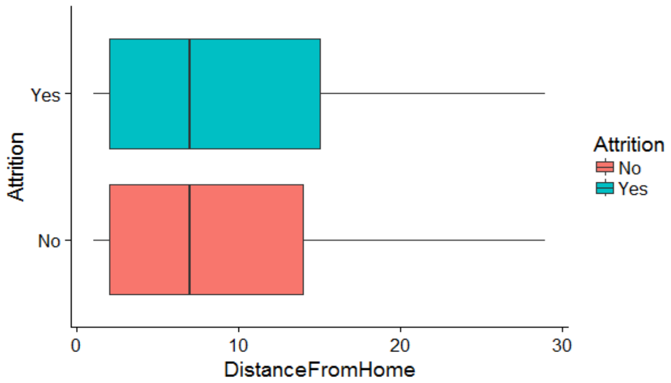
- Employees travelled rarely are tend to leave the company
- Attrition in HR looks low and R&D look high
- Life Sciences department shows more attrition and HR at low
- Female associates show less attrition than Male associates
- Employees from HR stay longer than the Sales executives
- Divorced employees are less likely to leave the company
- Associates holding bachelor degrees are more prone to leave and Doctors seem to be committed to the company
- Low work environment is impacting the attrition
- People with low performance rating are prone to leave the company
- The other factors contributing to attrition are – the low performance rating, less monthly salary, less hikes,

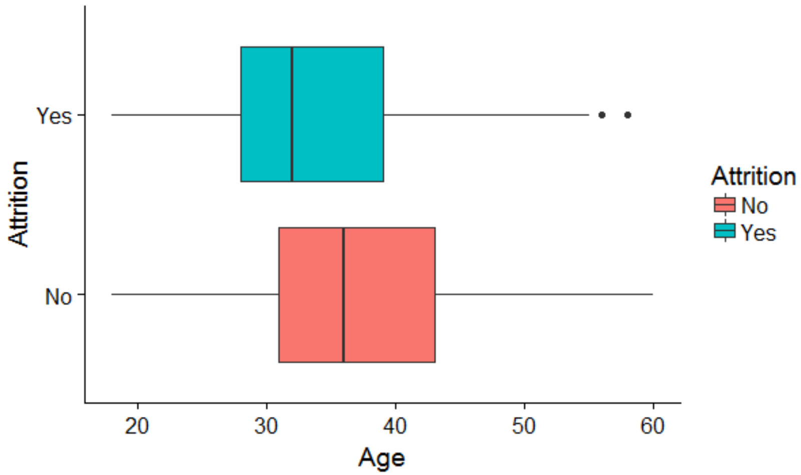
  no trainings , working with same manager for long period

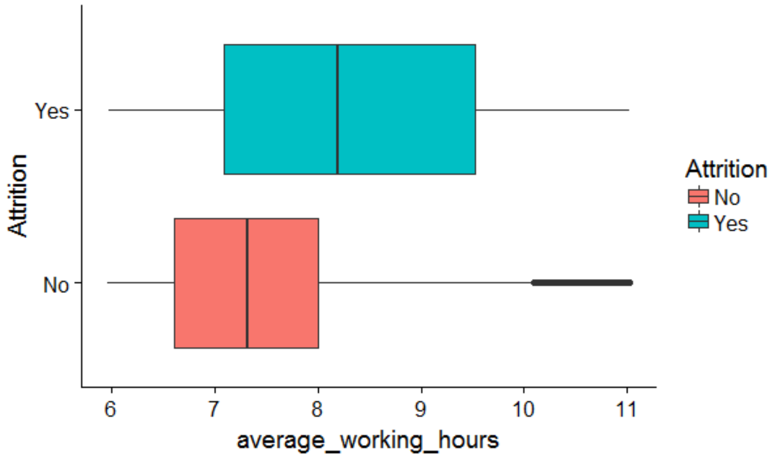
Multivariate Analysis of Data









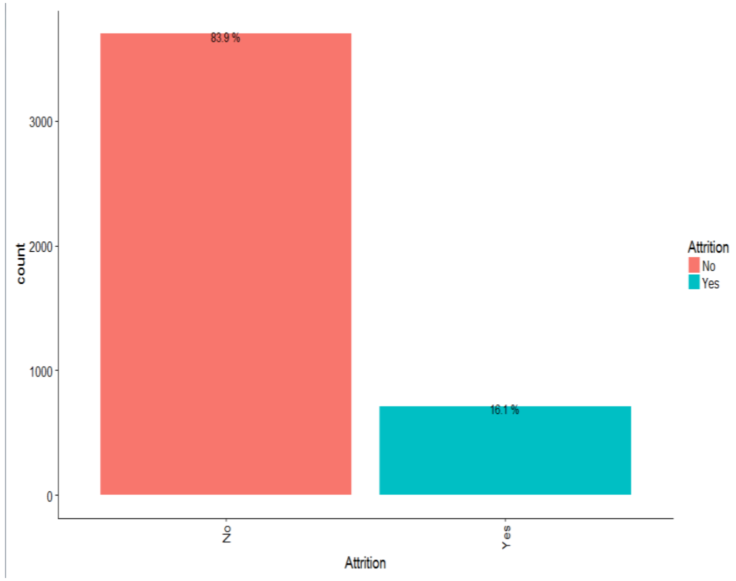


There is significant difference in **YearAtCompany** and **Average working hours**, based on these 2 attrition this will taking appropriate action during model

Model Building

Pre-Model Stage

- Prepare the Input variables for Model.
- Normalize the continuous variable and scale them using R functions.
- Used the Logistic regression on target variable ‘Attrition’ because it is a categorical variable.
- Given data has Attrition as 16.1%.



- Created a dummy variables for categorical variables as a part of variable reduction technique.
- Split the data into 7:3 as train and test data.

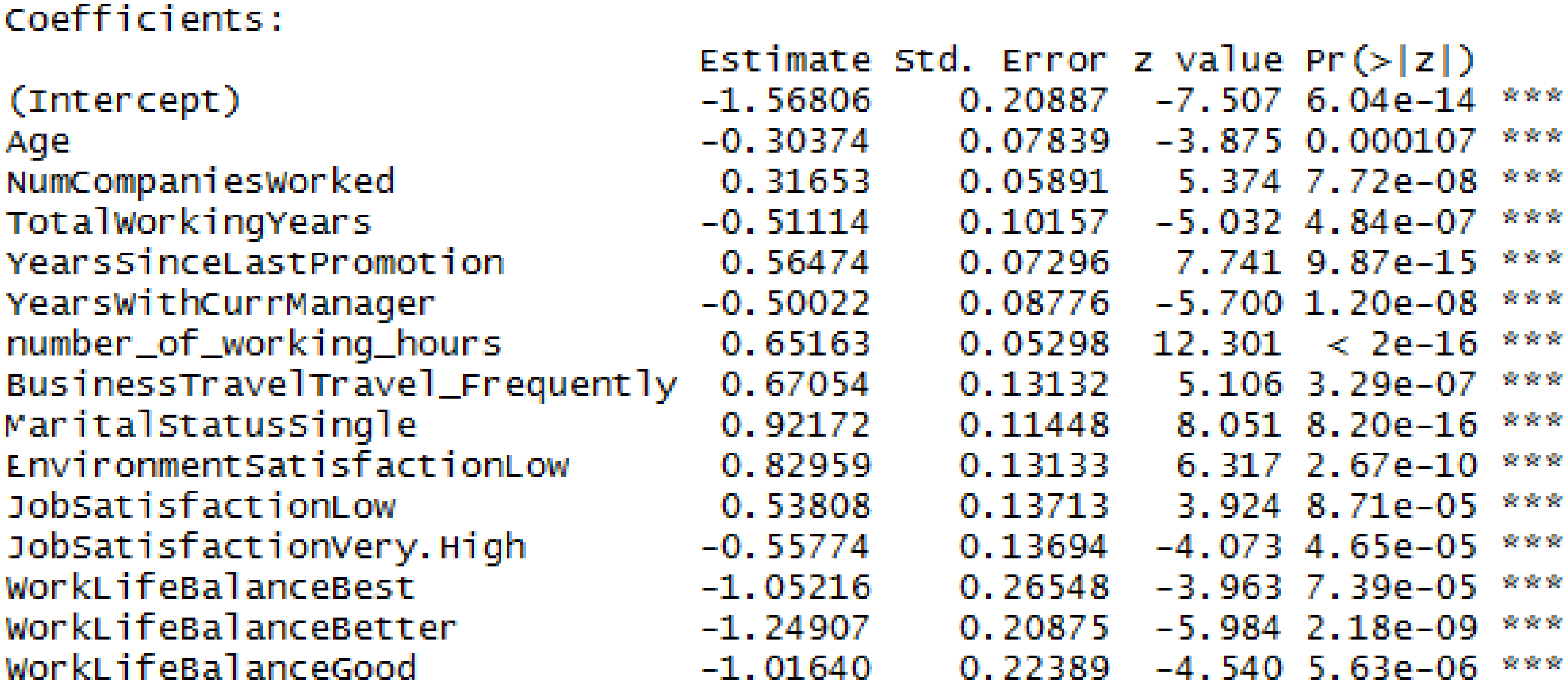
Model Building contd..

Steps followed in Modeling Stage

- Used Generalized linear regression to predict Attrition.
- Performs stepwise model selection by using AIC. stepAIC selects the model based on Akaike Information Criteria, not p-values. The goal is to find the model with the smallest AIC by removing or adding variables in your scope. Since each variable in the model is penalized with a factor of 2, this leads to LR tests with p-values < 0.1573.
- Checking for multicollinearity among your model predictors using VIF analysis.
- Removed variables whose P- value is high.

Model Building - Final Model

- There are 14 significant variables in the final model that with positive and negative coefficients. The Positive coefficients represent the chances of attrition whereas the negative coefficients impact it negatively



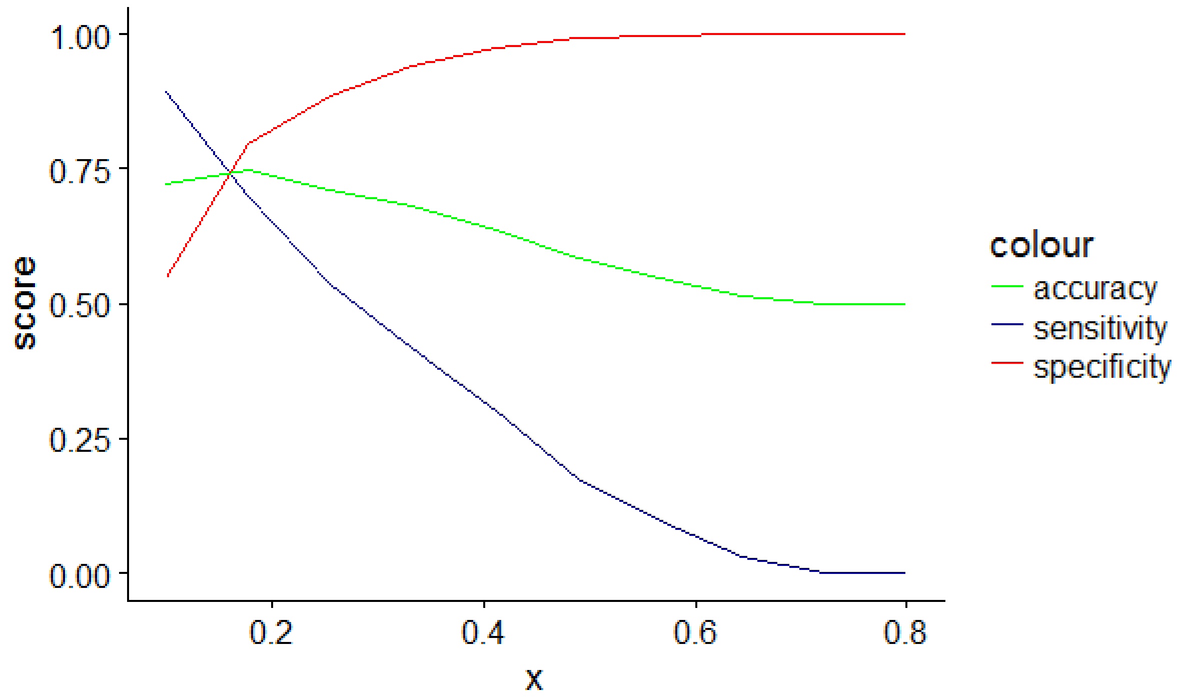
```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                     -1.56806    0.20887  -7.507 6.04e-14 ***
Age                             -0.30374    0.07839  -3.875 0.000107 ***
NumCompaniesWorked               0.31653    0.05891   5.374 7.72e-08 ***
TotalWorkingYears               -0.51114    0.10157  -5.032 4.84e-07 ***
YearsSinceLastPromotion          0.56474    0.07296   7.741 9.87e-15 ***
YearsWithCurrManager            -0.50022    0.08776  -5.700 1.20e-08 ***
number_of_working_hours          0.65163    0.05298  12.301  < 2e-16 ***
BusinessTravelTravel_Frequently  0.67054    0.13132   5.106 3.29e-07 ***
MaritalStatusSingle              0.92172    0.11448   8.051 8.20e-16 ***
EnvironmentSatisfactionLow       0.82959    0.13133   6.317 2.67e-10 ***
JobSatisfactionLow               0.53808    0.13713   3.924 8.71e-05 ***
JobSatisfactionVery.High        -0.55774    0.13694  -4.073 4.65e-05 ***
WorkLifeBalanceBest             -1.05216    0.26548  -3.963 7.39e-05 ***
WorkLifeBalanceBetter           -1.24907    0.20875  -5.984 2.18e-09 ***
WorkLifeBalanceGood             -1.01640    0.22389  -4.540 5.63e-06 ***
```

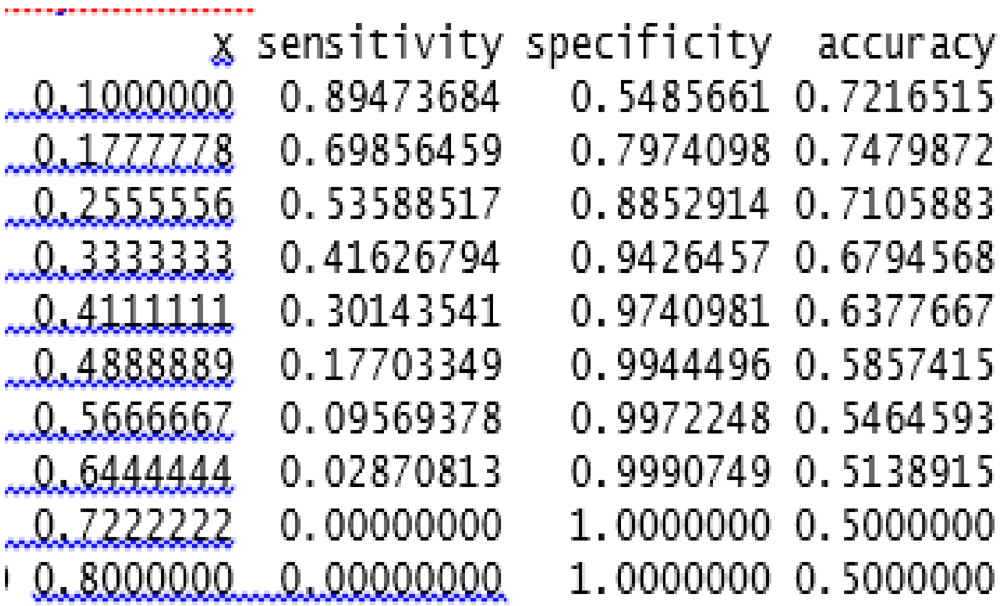
Model Validation

Now predict and check performance of model on test data.

1. **Confusion Matrix :** Used **0.5** as initial cutoff to check accuracy and other scores of the model

- Accuracy : 0.8651
- Sensitivity : 0.30144
- Specificity : 0.97410

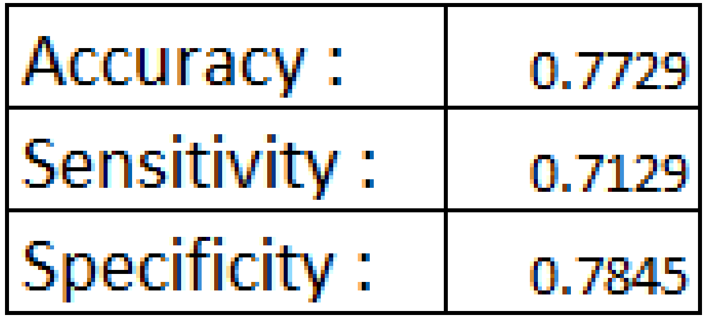




```
          x sensitivity specificity  accuracy
  0.1000000  0.89473684   0.5485661 0.7216515
  0.1777778  0.69856459   0.7974098 0.7479872
  0.2555556  0.53588517   0.8852914 0.7105883
  0.3333333  0.41626794   0.9426457 0.6794568
  0.4111111  0.30143541   0.9740981 0.6377667
  0.4888889  0.17703349   0.9944496 0.5857415
  0.5666667  0.09569378   0.9972248 0.5464593
  0.6444444  0.02870813   0.9990749 0.5138915
  0.7222222  0.00000000   1.0000000 0.5000000
  0.8000000  0.00000000   1.0000000 0.5000000
```

Model Validation contd..

- The graph in slide **15** shows that specificity increases as we increase cutoff and vise versa for specificity and accuracy reaches down to **0.50**
- To have a balance between specificity and sensitivity we can get value from the intersection ,which is around **0.19** Run the confusion matrix



Accuracy :	0.7729
Sensitivity :	0.7129
Specificity :	0.7845

With the cut off **0.19**, **71 %** predictions of the employees leaving the company and **78%** predictions of not leaving the company are accurate

Model Validation contd..

2. KS Statistics calculation

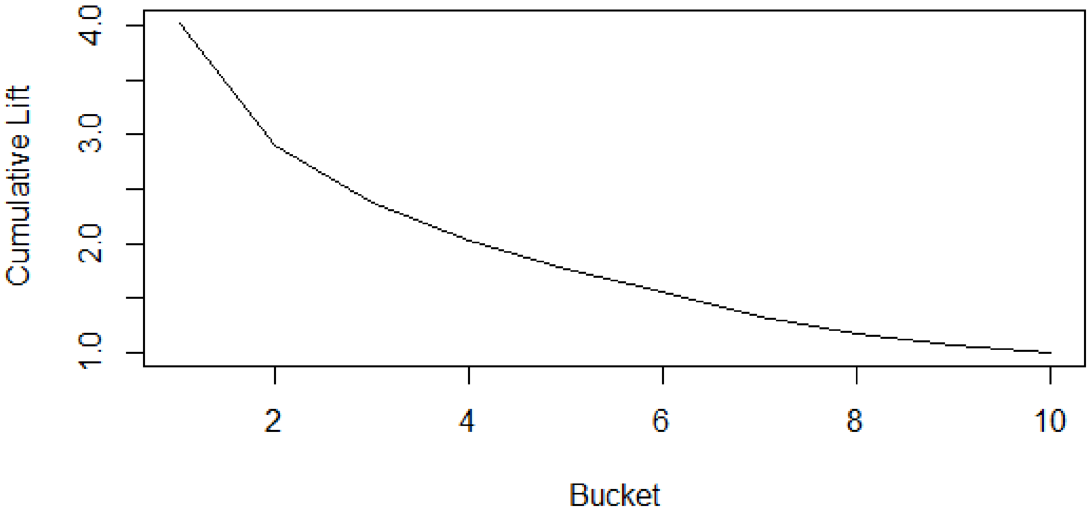
KS Test is a non-parametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution.

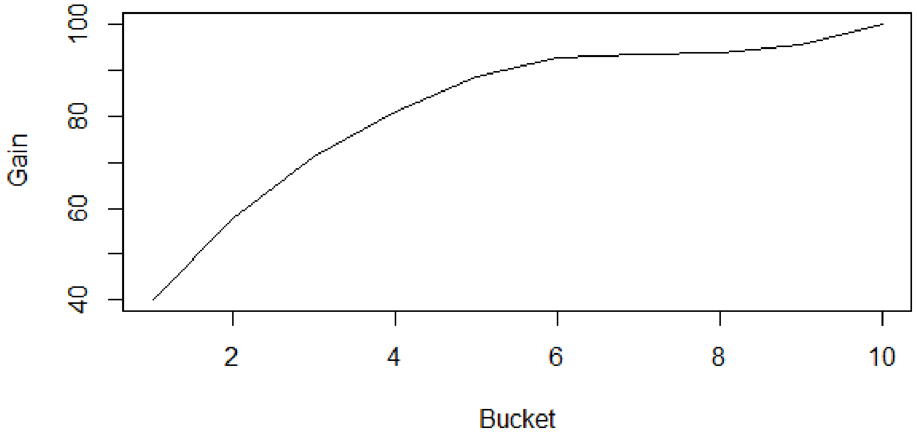
KS Statistics for our model is : **0.4973775** , It a good Statistics as its between 40- 60

3.Gain and Lift Charts

Gain or lift is a measure of the effectiveness of a classification model calculated as the ratio between the results obtained with and without the model.

Gain and lift charts are visual aids for evaluating performance of classification models.





Recommendations to Curb Attrition

- XYZ needs to look at the no. of companies a candidate has worked before hiring him. This can indicate the commitment of the candidate
- Job role, work life balance, Job satisfaction are linked. Company to increase the engagement levels of the associates thereby reducing the attrition
- XYZ to track the working hours to ensure no employee is over working or under working. Over working continuously may result in attrition
- If an employee works with the same manager for a longer period of time the lesser are the chances that employee will leave the company.
- Hire skilful and expert candidates with more experience as they are less likely to leave the company. However, if the person has worked in many companies then the chances that he/she will leave the company increases.
- Employees who are unmarried are prone to leaving the company.