HR ANALYTICS CASE STUDY

Presented By:

1. Rutuja Mowade
2. Alay Shah
3. Rahul Rungta
4. Sabyasachi Tripathy

BUSINESS OBJECTIVE

XYZ company is a large company with 4000 employees at any point of time.

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.

Hence, the management has contracted an HR analytics firm to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay.

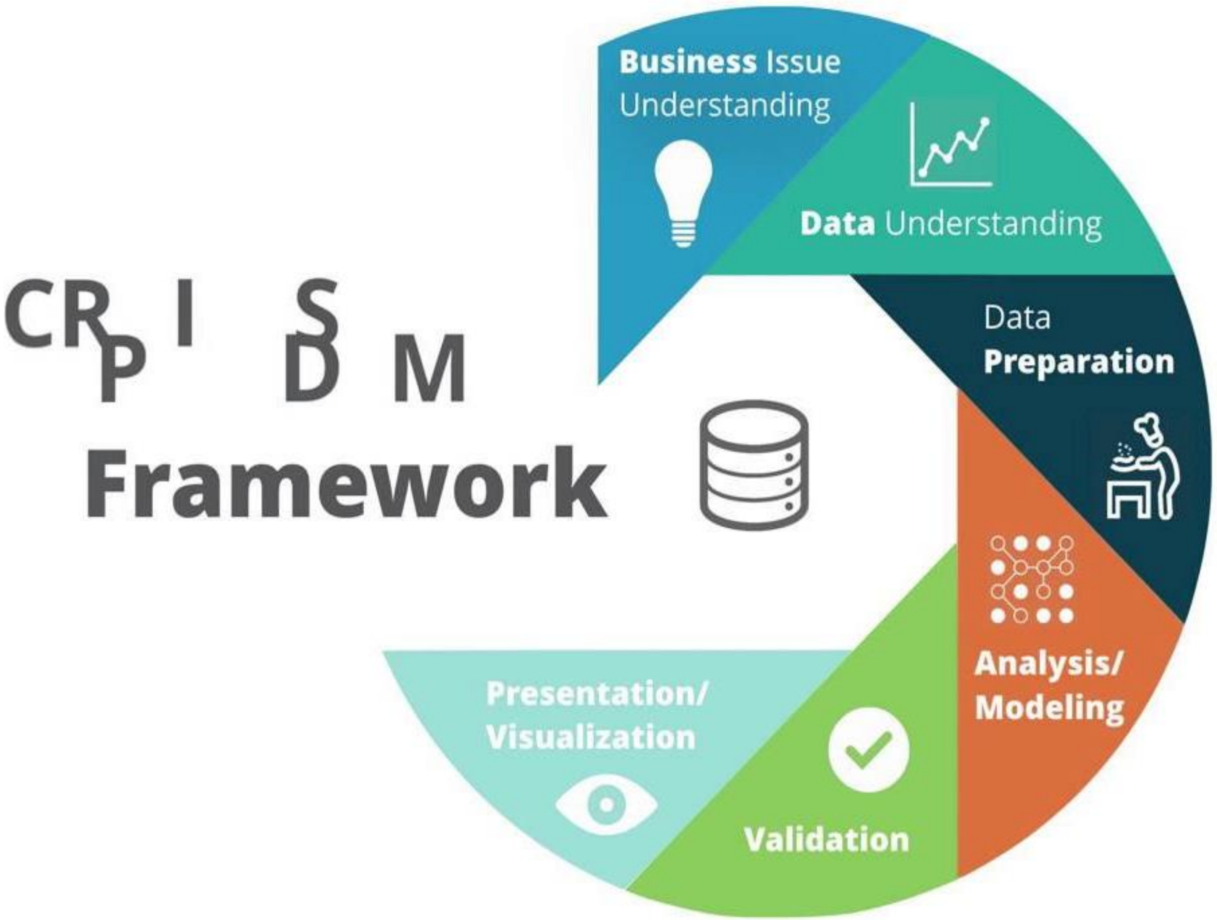
Also, they want to know which of these variables is most important and needs to be addressed right away.

GOAL OF THE CASE STUDY

The goal of this case study is to **model the probability of attrition** using a logistic regression.

The results thus obtained will be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay.

PROBLEM SOLVING METHODOLOGY



- The methodology we followed is termed as CRISP DM – Cross Industry Standard Process for Data Mining.
- As shown in figure, this methodology involves six steps:
  - Business Understanding
  - Data Understanding
  - Data Cleaning & EDA & Data Preparation
  - Model Building
  - Model Evaluation / Validation
  - Presentation/Visualisation
- The Business Objective was already explained earlier, hence we will move on to the next step i.e. Data Understanding, in the following slide.

DATA UNDERSTANDING

We had access to the following data :-

1) **GENERAL DATA :**
   - This data included all the general information about the employees like salary, distance from home, Education field, Education level, Department and 20+ such variables including the target variable – Attrition (Whether the employee left the company or not).
2) **MANAGER SURVEY DATA :**
   - As the name suggests, these data set includes the survey about the employees collected from the managers of the company. The survey was collected on the basis of two parameters, Job Involvement and Performance Rating. The manager had to rate the employee on the scale of 1 to 4, where 1 indicates Bad rating and 4 indicates Best Rating.
3) **EMPLOYEE SURVEY DATA :**
   - These data set includes the survey collected from employees to better understand their Environment Satisfaction, Job Satisfaction and their Work Life Balance. Again, the survey for the above three parameters were collected on the scale of 1 to 4, where 1 is the lowest satisfaction level and 4 is the highest possible satisfaction level.
4) **IN TIME**
   - The data set included information of the recorded date and time when employee entered the office.
5) **OUT TIME**
   - The data set included information of the recorded date and time when employee exited the office.

DATA CLEANING

The Data cleaning was carried out in the following manner :

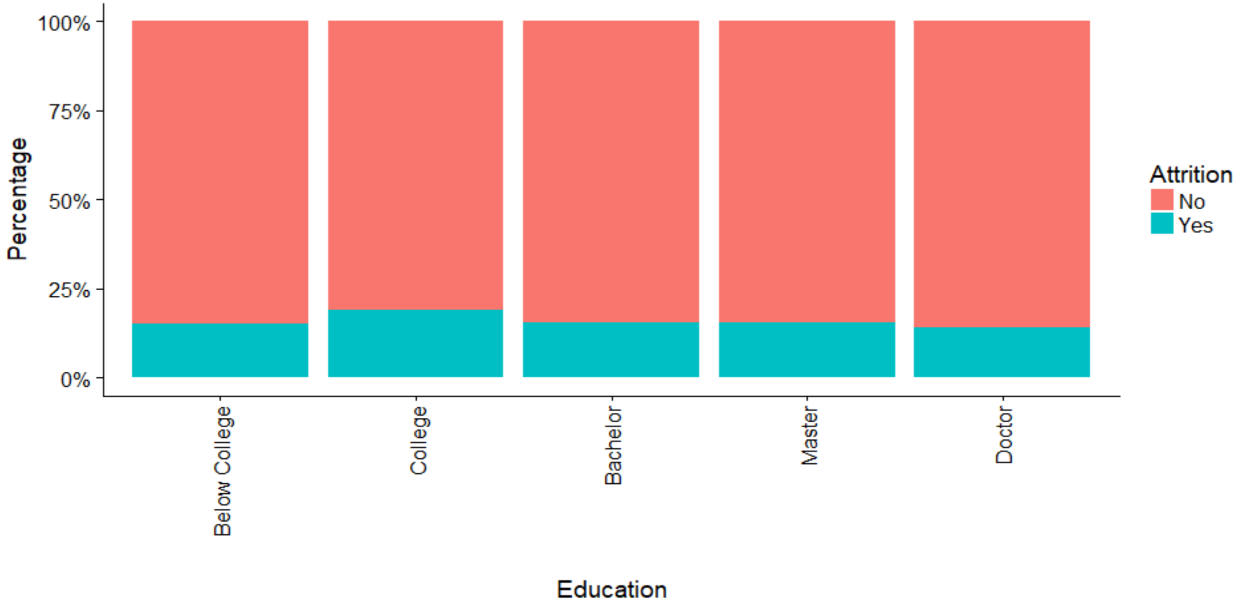
- All the five data sets were first imported in the R Studio.
- The In-time and Out-time data set were of no use individually. But, then we realized that a very useful information i.e. "Hours spent by employee at company/office" can be extracted with combination of these data sets. Hence we calculated the difference between in_time and out_time for each day, for each employee and prepared a new data frame out of it, which eventually led us to calculate average hours spent by each employee over the period. Once we had average hours, it was not difficult for us to calculate whether a person was working over time, under time or just completing his standard hours expected out of him. And all these was done with the help of another variable column "Standard Hours" which was present in the general dataset.
- Also, we found a lot of NA's in in_time and out_time data sets, which initially appeared as a data quality issue. However, lately we realized that the NA's were just an indication of holidays, or a leave taken by an employee if it was not a holiday. So we moved on to convert the NA's into leaves taken by each employee and saved that as another variable to use later.
- The next data set which required attention was employee survey data set. These data set had around 90+ NA's, and omitting the NA's would be like literally throwing away nearly 2% of the data. Hence, we kept that at last option and started searching out ways to impute these missing values. And the best possible way we found to deal with these NA's was to replace the missing values with the median. Why median? Because this was a survey data set and not a numeric entity. Hence taking mean wont work, and median would represent the crowd. That way, we did the missing value imputation.
- The last data cleaning issue we faced was to remove some NA's from general data set. Only two variables had NA values. They were : Num of companies worked (19 NA's) and Total working years (9NA's). The total NA's to be treated was less than 30 i.e. around 0.5% data. Hence, we though to remove that missing values instead of exploiting or interpreting them, as removing less than 1% of the data would not create any significant harm to the final model.
- Finally, our data was ready to be analysed. Hence we merged the data sets and moved on to exploratory data analysis.

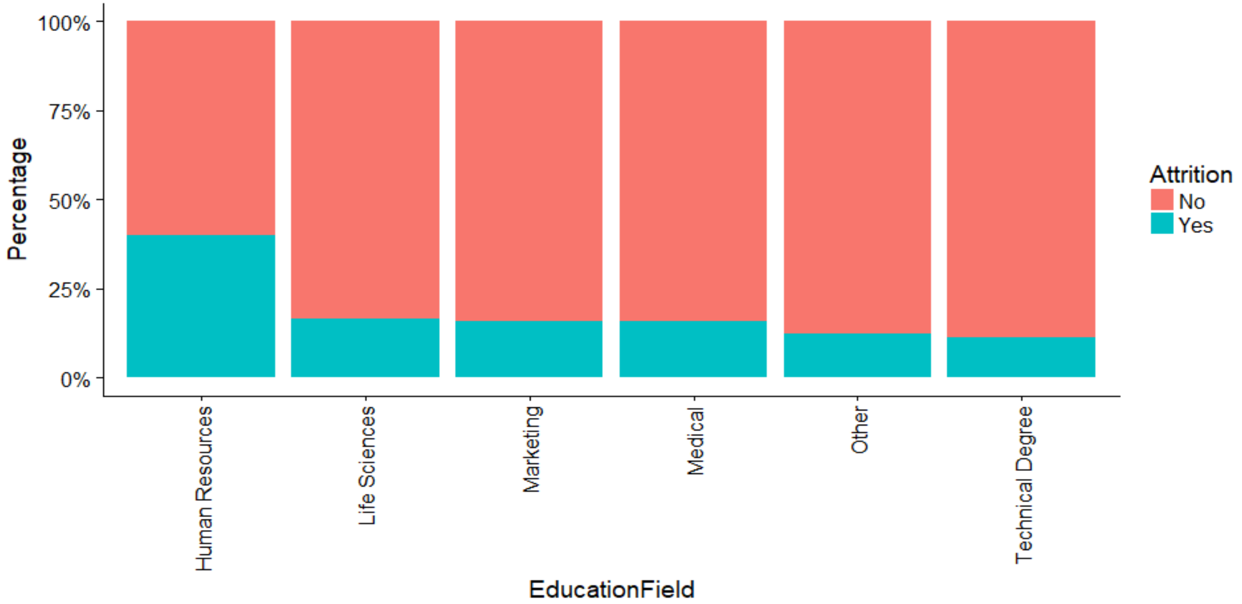
EXPLORATORY DATA ANALYSIS

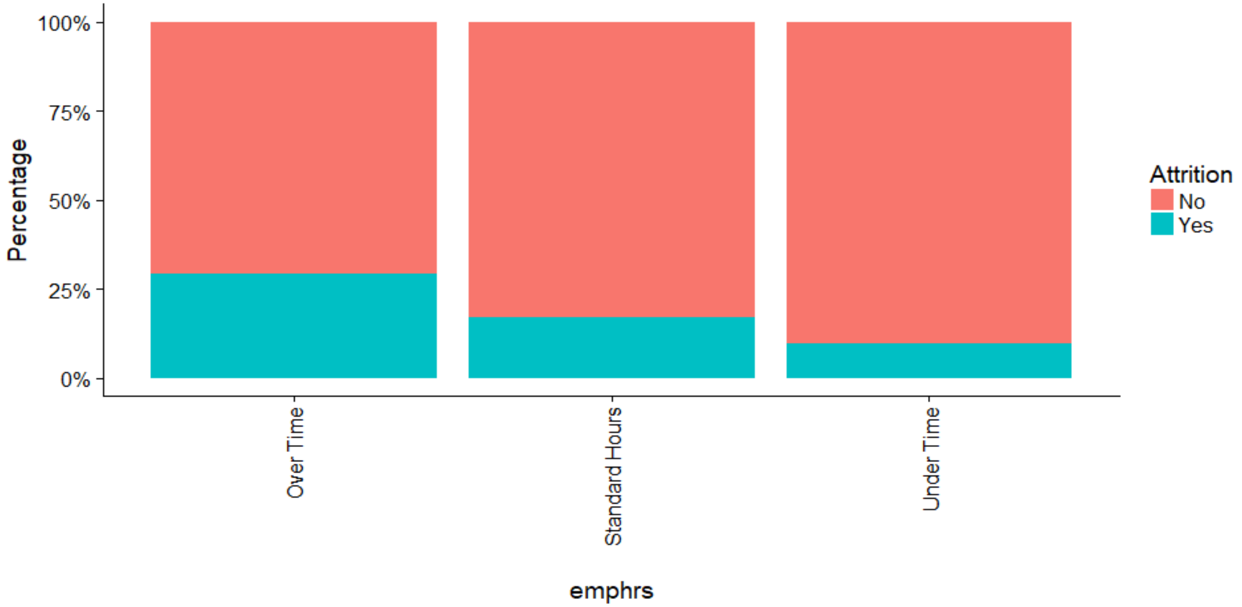
Below is the curated list of most important insights derived from exploratory data analysis. Following this, the next few slides will show all the plots made in R for all the variables.

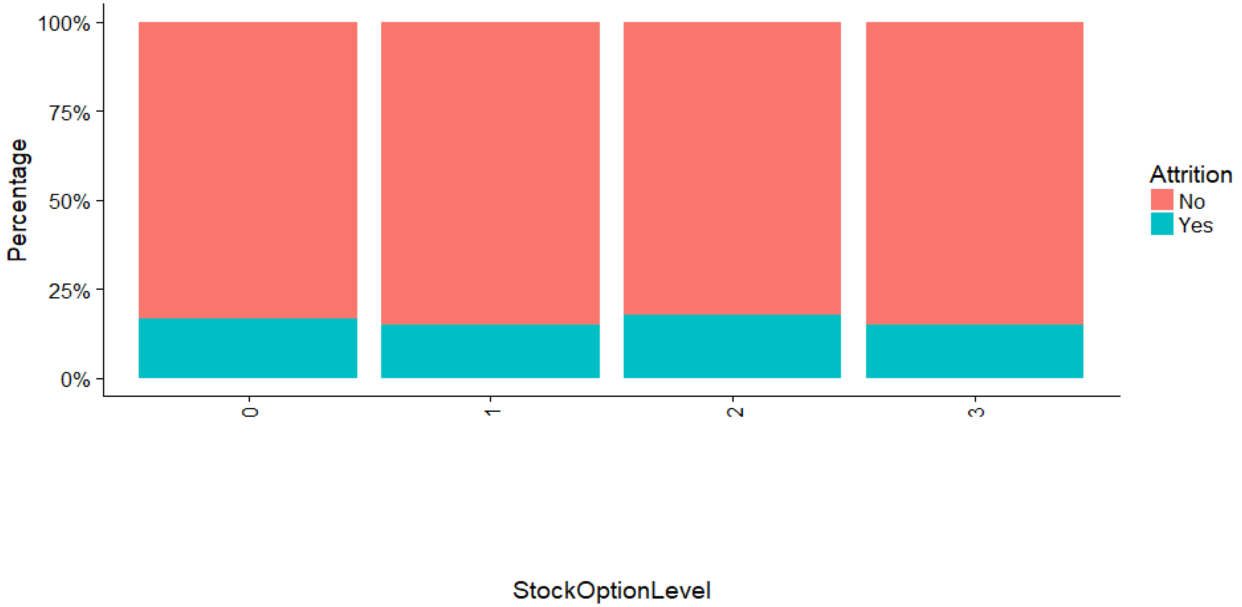
Variable Name	Important Insights
1) Education Field	1) The Human Resources Section observed the highest percentage of attrition, nearly twice as compared to any other field.
2) emphrs	2) A person who works over time is more likely to leave a job than the one who works for standard hours or under time.
3) Job Role	3) Job Role of Research Director observed the highest attrition, whereas that of Manufacturing director observed the least.
4) Travel Frequency	4) The Person who travels frequently is most likely to leave the job, as compared with the one who does not travel at all
5) Department	5) The HR department experiences the maximum percentage of attrition as compared to any other department.
6) Single	6) A person who is unmarried is twice more likely to leave the job as compared to a divorced or a married person.
7) Environment Satisfaction	7) The lesser it is, the more are the chances of person leaving the job.
8) Job Satisfaction	8) The lesser it is, the more are the chances of person leaving the job.
9) Work Life Balance	9) The lesser it is, the more are the chances of person leaving the job.
10) Age	10) A person in 20’s is nearly twice more likely to leave a job as compared to a person who is in 50’s.
11) Num Companies Worked	11) More is the number, more will be the probability of attrition.
12) Training times last year	12) More is the number, lesser will be the probability of attrition.
13) Years with Current Manager	13) More years, less probability of attrition.
14) Total Working Years	14) An experienced person is less likely to leave the company i.e. More is the number of working years, higher is the stability.

EDA PLOTS

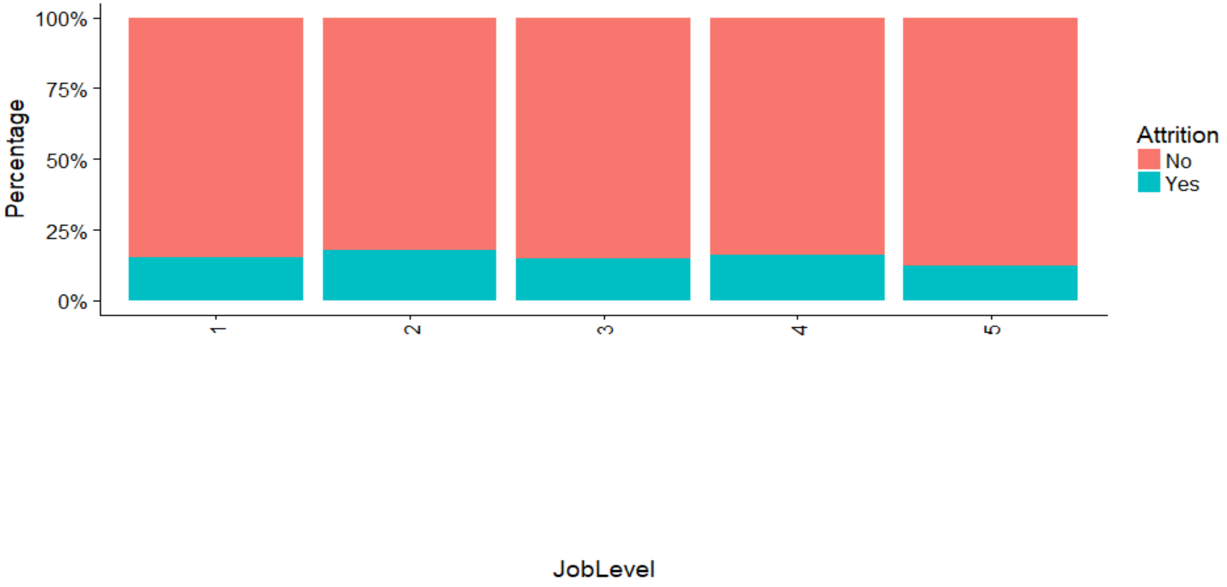

Percentage
100%
75%
50%
25%
0%
Below College
College
Bachelor
Master
Doctor
Education
Attrition
🟥 No
🟦 Yes


Percentage
100%
75%
50%
25%
0%
Human Resources
Life Sciences
Marketing
Medical
Other
Technical Degree
EducationField
Attrition
🟥 No
🟦 Yes

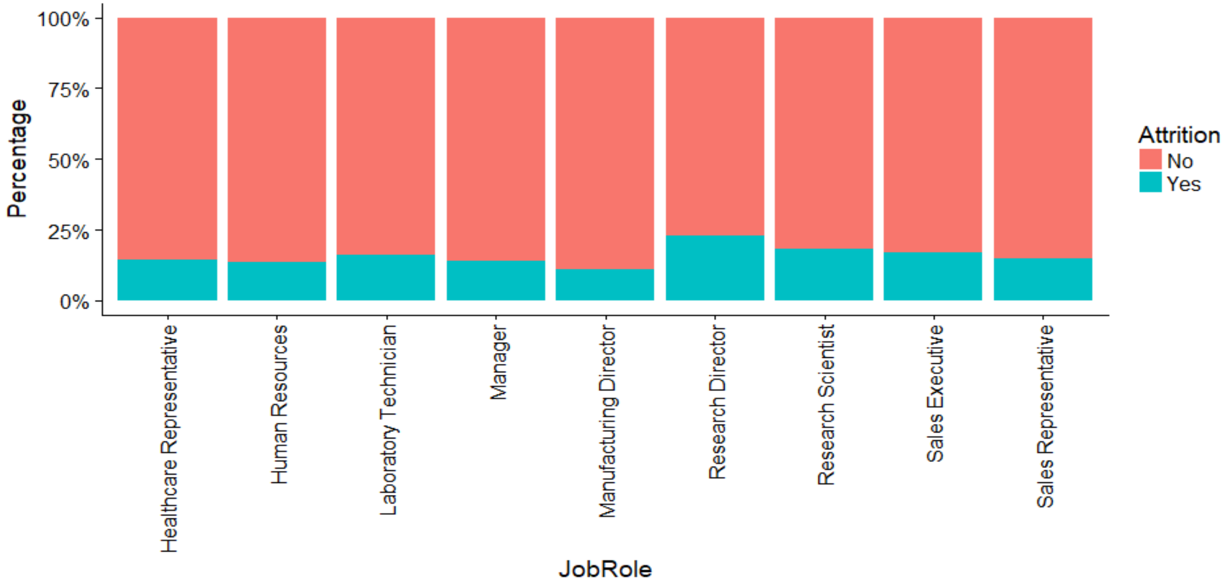

Percentage
100%
75%
50%
25%
0%
Over Time
Standard Hours
Under Time
emphrs
Attrition
🟥 No
🟦 Yes


Percentage
100%
75%
50%
25%
0%
0
1
2
3
StockOptionLevel
Attrition
🟥 No
🟦 Yes

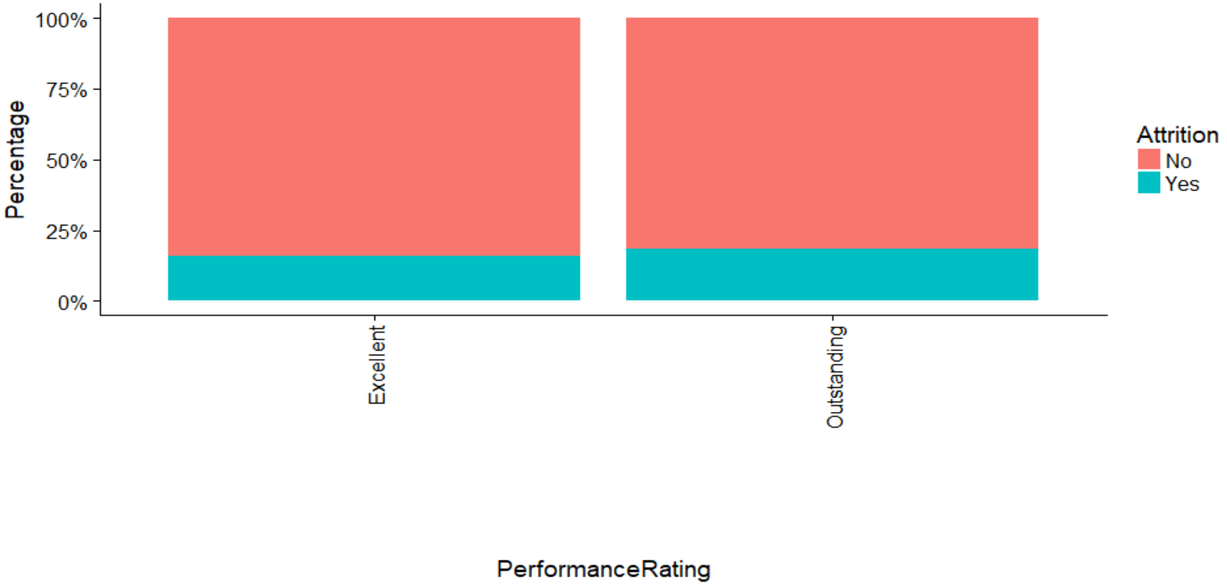
EDA PLOTS



Percentage
100%
75%
50%
25%
0%
1
2
3
4
5
JobLevel
Attrition
No
Yes

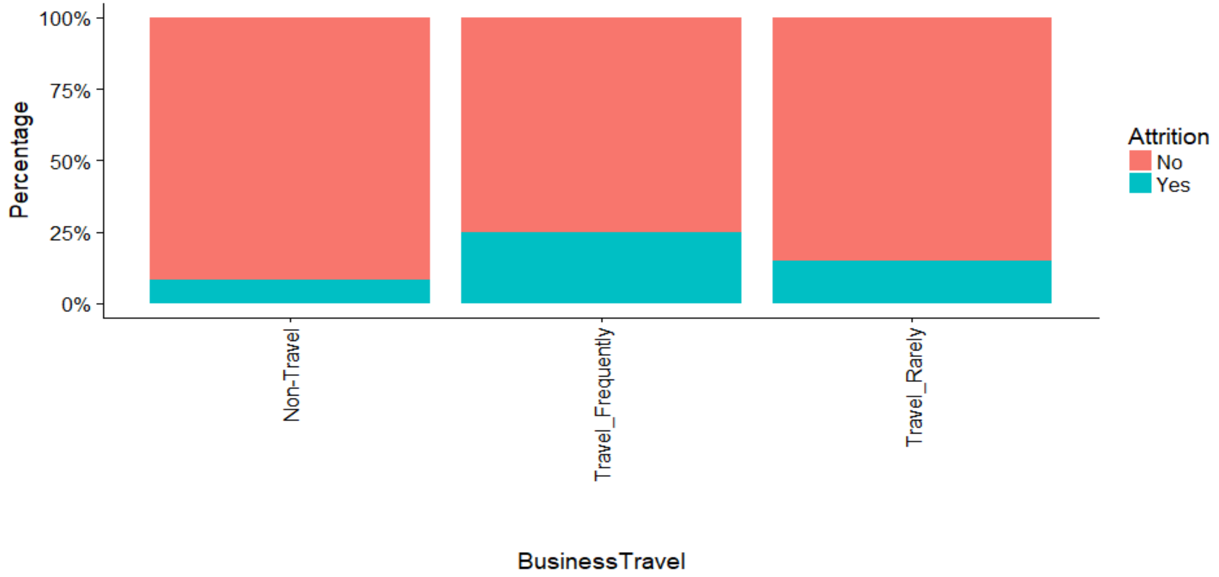


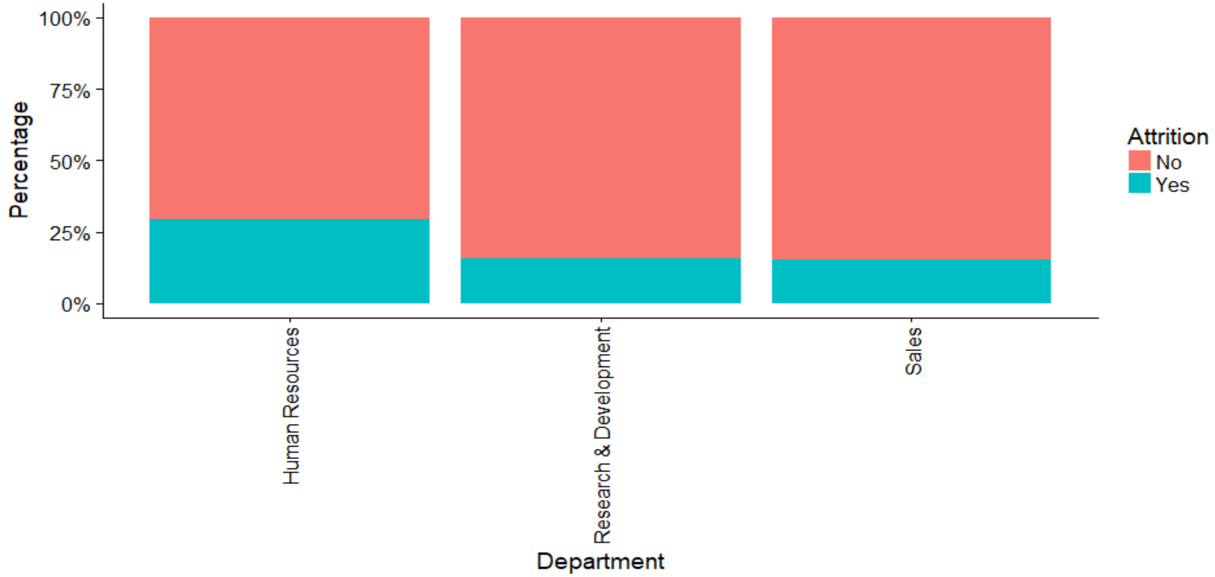
Percentage
100%
75%
50%
25%
0%
Healthcare Representative
Human Resources
Laboratory Technician
Manager
Manufacturing Director
Research Director
Research Scientist
Sales Executive
Sales Representative
JobRole
Attrition
No
Yes

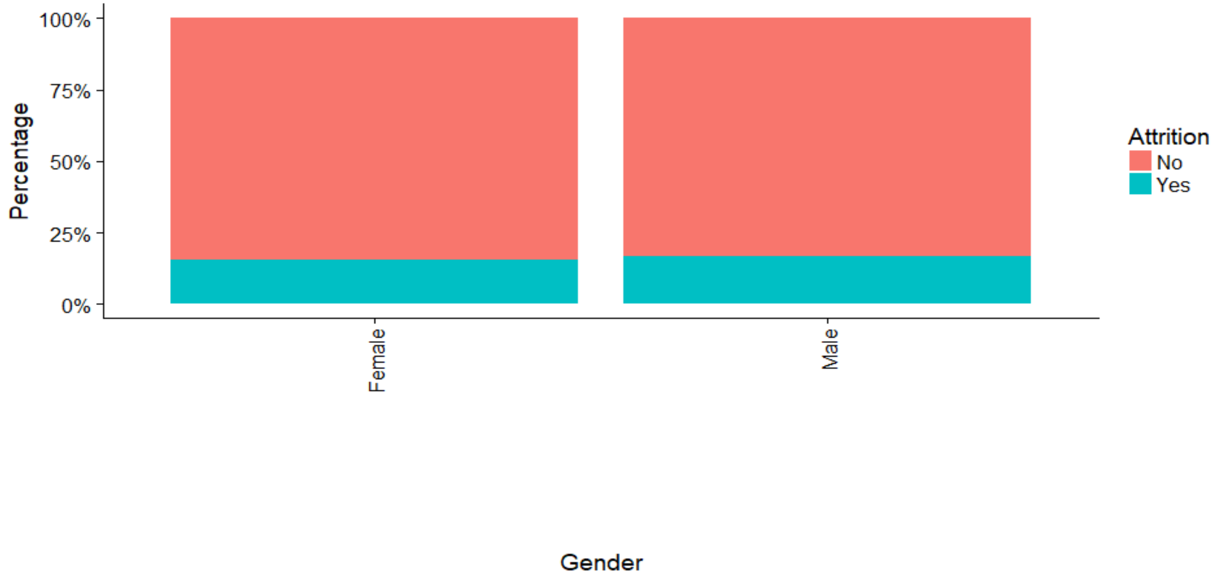


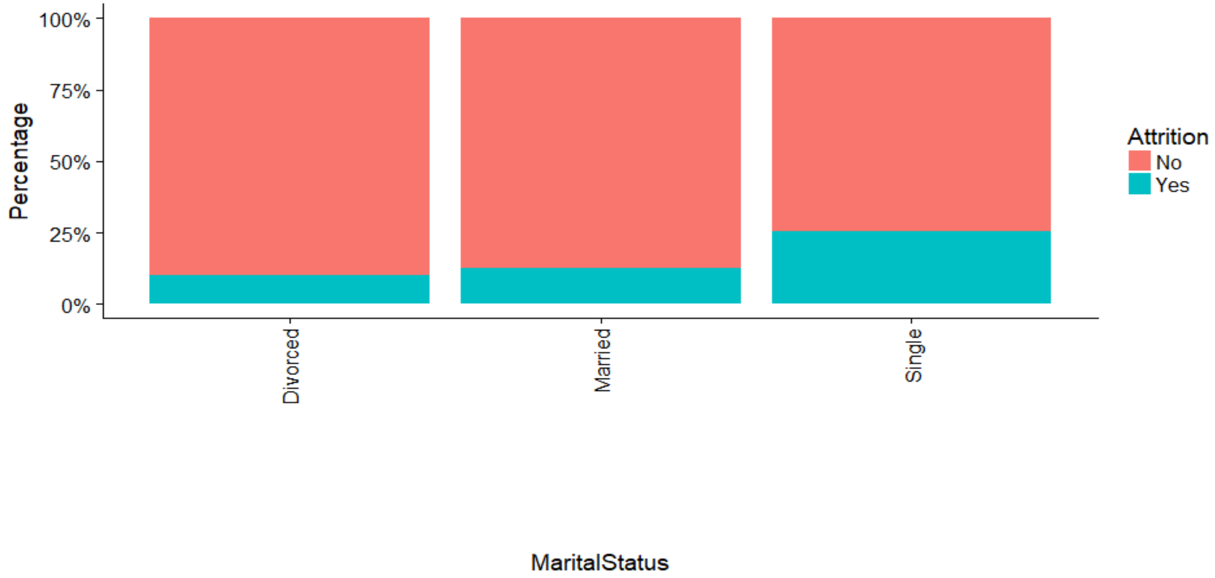
Percentage
100%
75%
50%
25%
0%
Excellent
Outstanding
PerformanceRating
Attrition
No
Yes

EDA PLOTS

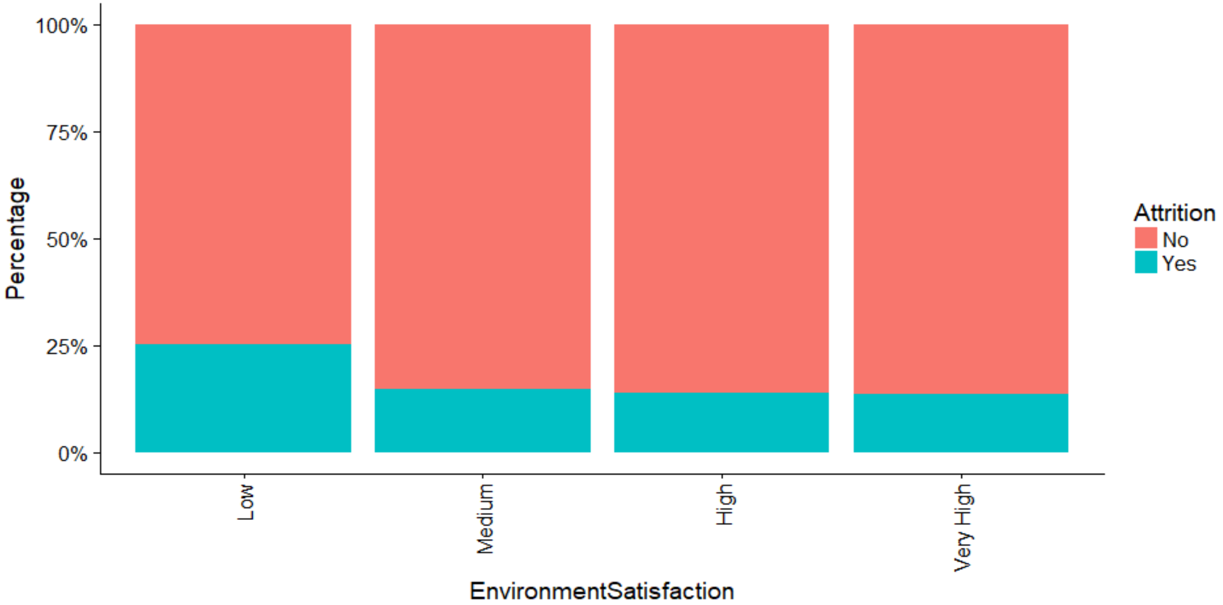

100%
75%
50%
25%
0%
Percentage
Non-Travel
Travel_Frequently
Travel_Rarely
BusinessTravel
Attrition
No
Yes

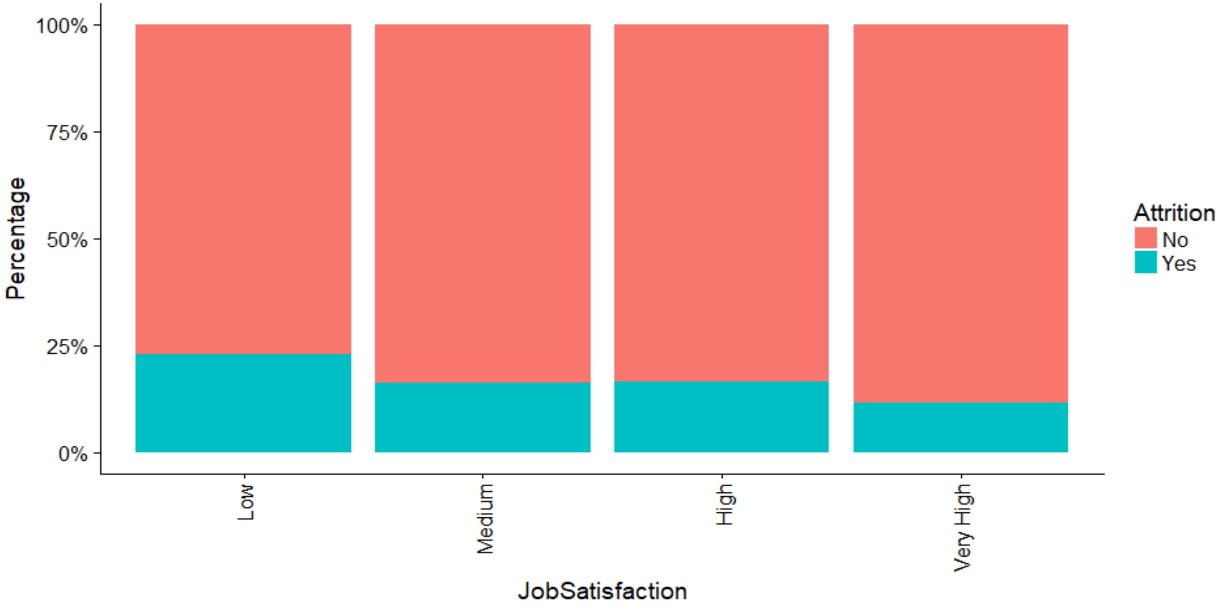

100%
75%
50%
25%
0%
Percentage
Human Resources
Research & Development
Sales
Department
Attrition
No
Yes

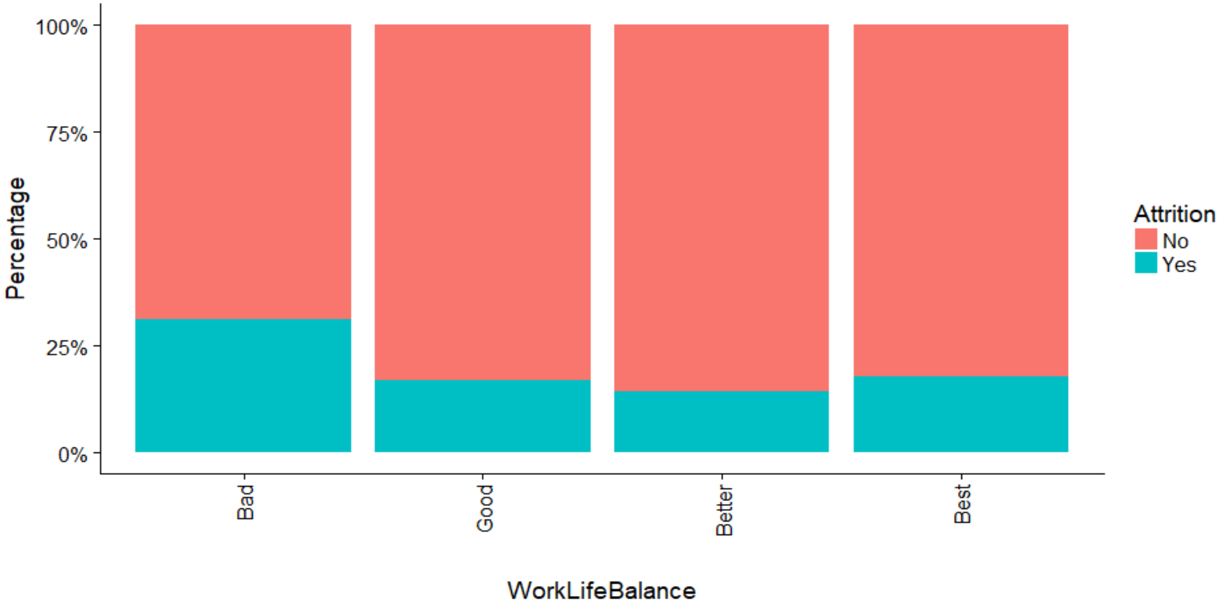

100%
75%
50%
25%
0%
Percentage
Female
Male
Gender
Attrition
No
Yes

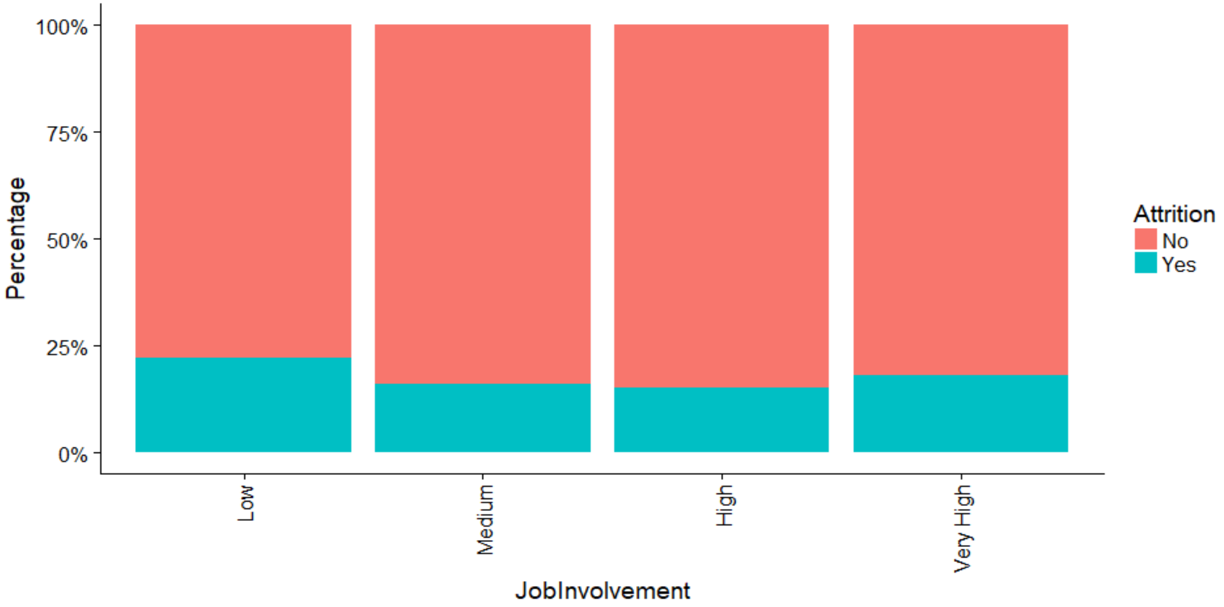

100%
75%
50%
25%
0%
Percentage
Divorced
Married
Single
MaritalStatus
Attrition
No
Yes

EDA PLOTS

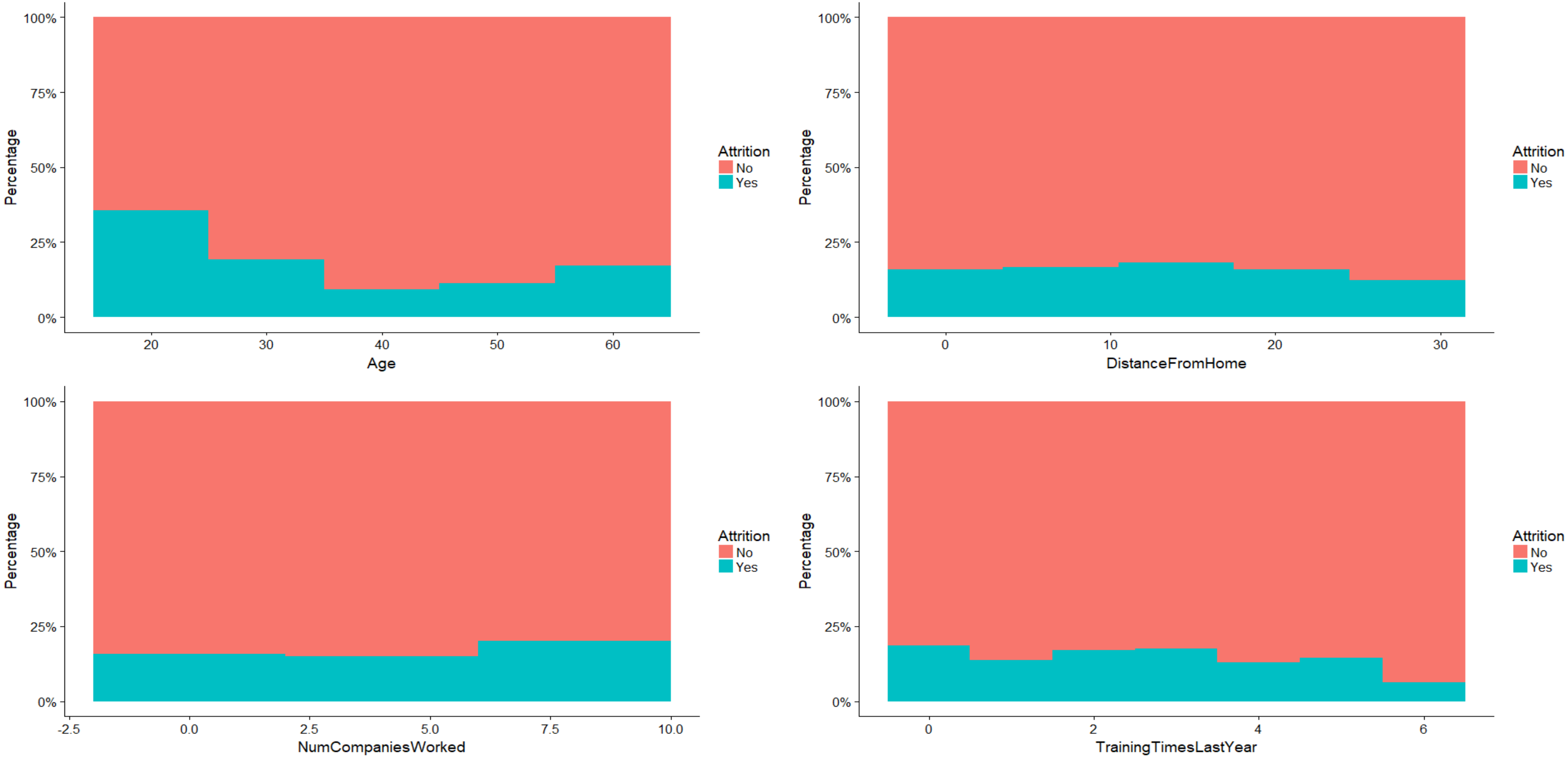

Percentage
100%
75%
50%
25%
0%
Low
Medium
High
Very High
EnvironmentSatisfaction
Attrition
No
Yes


Percentage
100%
75%
50%
25%
0%
Low
Medium
High
Very High
JobSatisfaction
Attrition
No
Yes


Percentage
100%
75%
50%
25%
0%
Bad
Good
Better
Best
WorkLifeBalance
Attrition
No
Yes

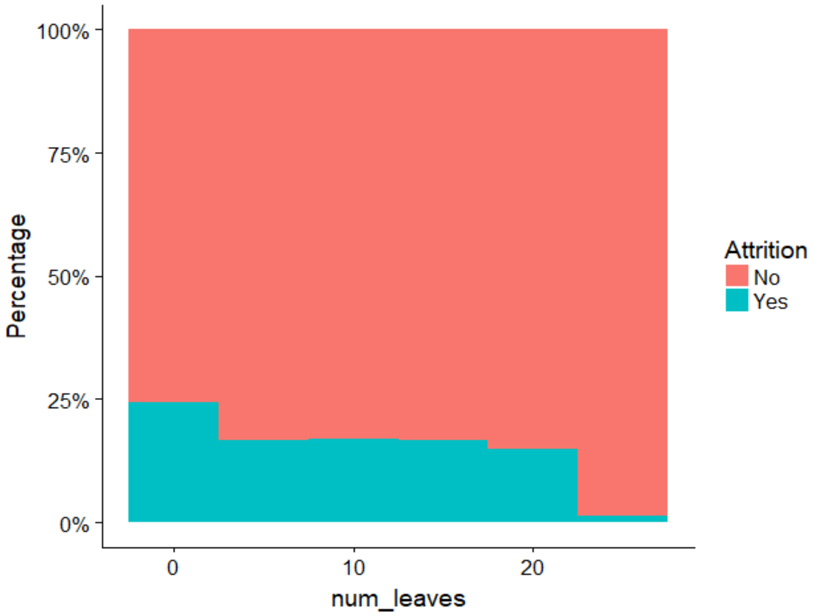

Percentage
100%
75%
50%
25%
0%
Low
Medium
High
Very High
JobInvolvement
Attrition
No
Yes

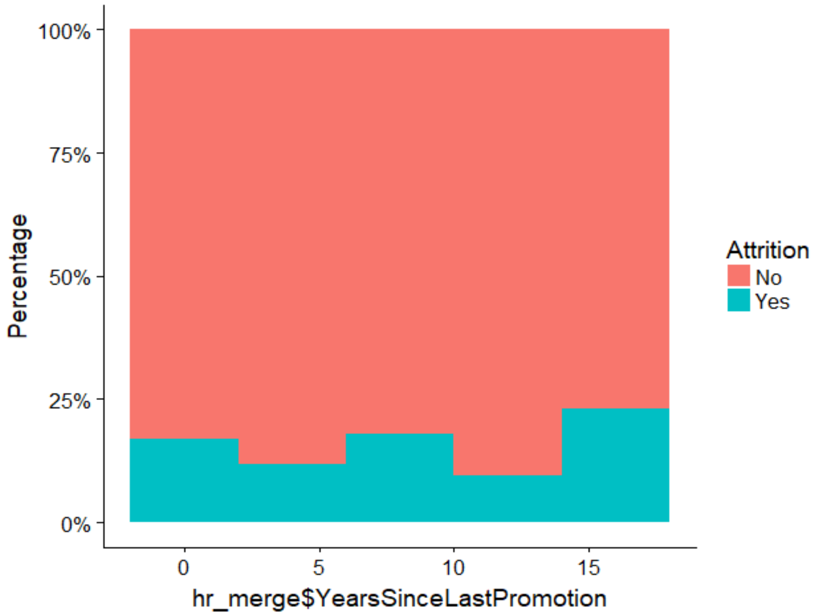
EDA PLOTS

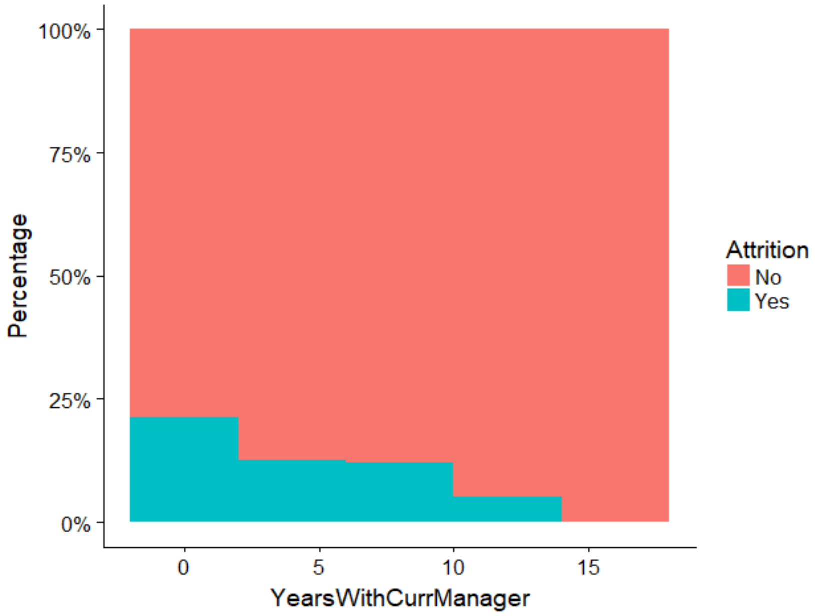


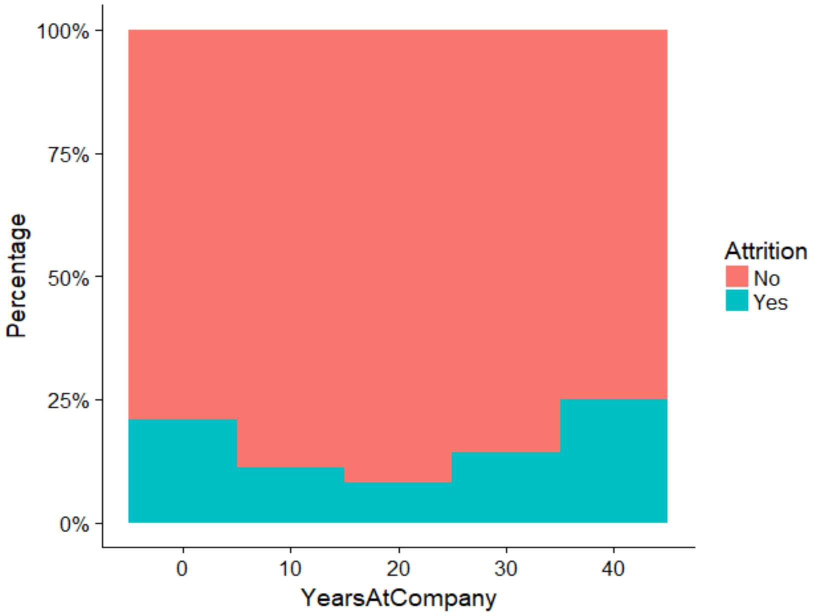
Percentage
100%
75%
50%
25%
0%
20
30
40
50
60
Age
Attrition
No
Yes
Percentage
100%
75%
50%
25%
0%
0
10
20
30
DistanceFromHome
Attrition
No
Yes
Percentage
100%
75%
50%
25%
0%
-2.5
0.0
2.5
5.0
7.5
10.0
NumCompaniesWorked
Attrition
No
Yes
Percentage
100%
75%
50%
25%
0%
0
2
4
6
TrainingTimesLastYear
Attrition
No
Yes

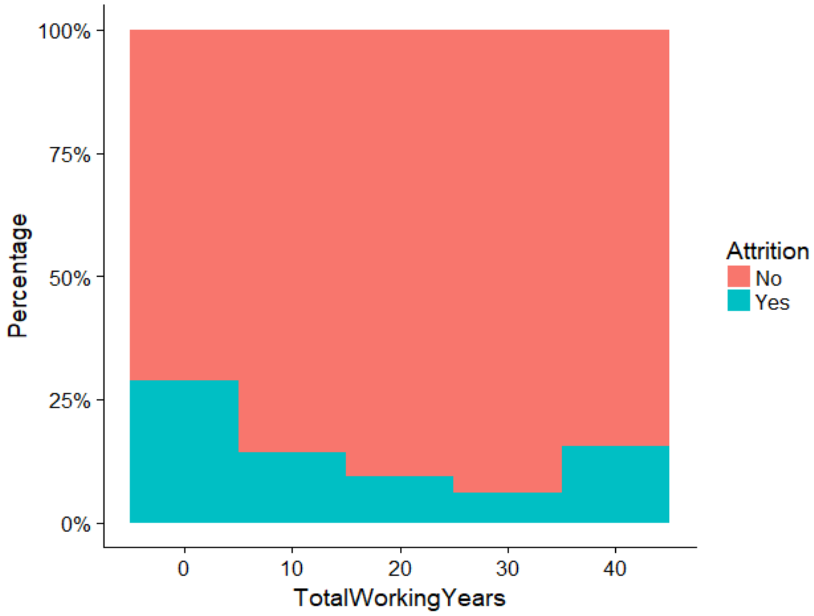
EDA PLOTS


100%
75%
50%
25%
0%
Percentage
0
10
20
num_leaves
Attrition
No
Yes

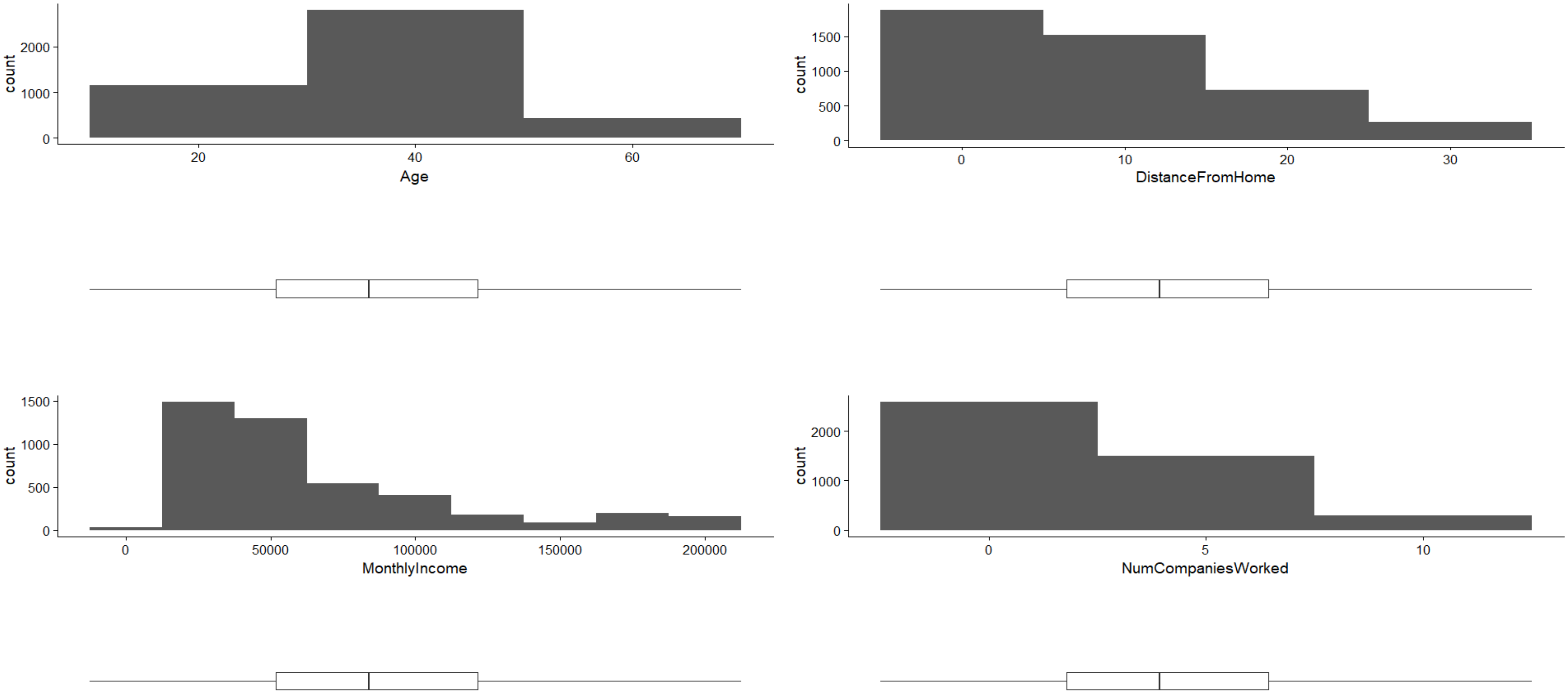

100%
75%
50%
25%
0%
Percentage
0
5
10
15
hr_merge$YearsSinceLastPromotion
Attrition
No
Yes


100%
75%
50%
25%
0%
Percentage
0
5
10
15
YearsWithCurrManager
Attrition
No
Yes

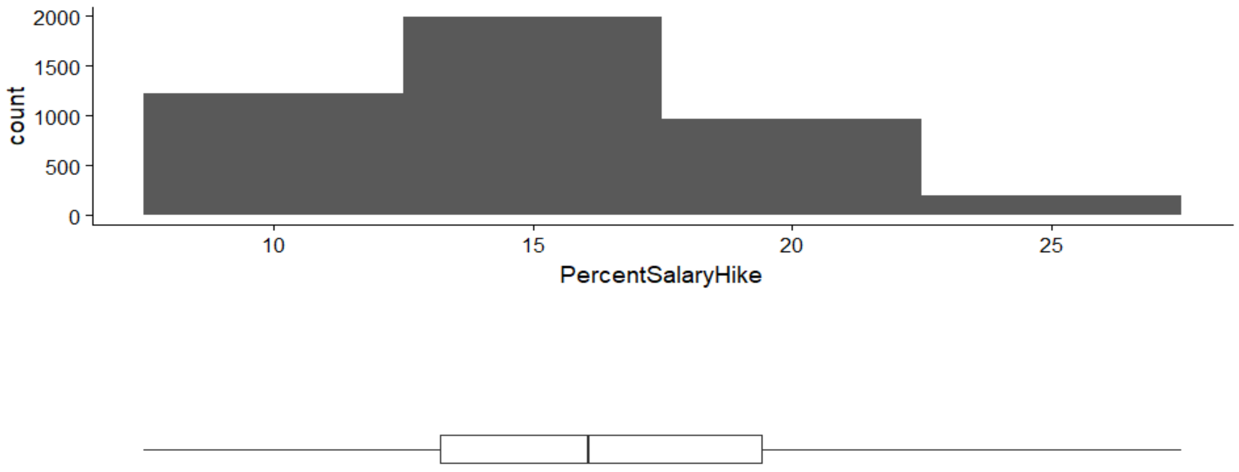

100%
75%
50%
25%
0%
Percentage
0
10
20
30
40
YearsAtCompany
Attrition
No
Yes

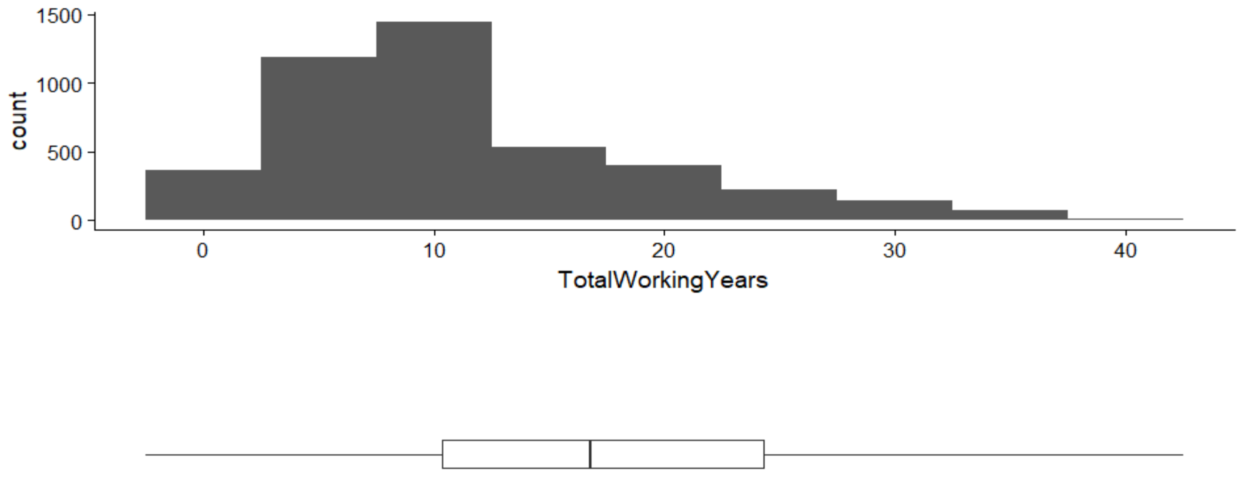

100%
75%
50%
25%
0%
Percentage
0
10
20
30
40
TotalWorkingYears
Attrition
No
Yes

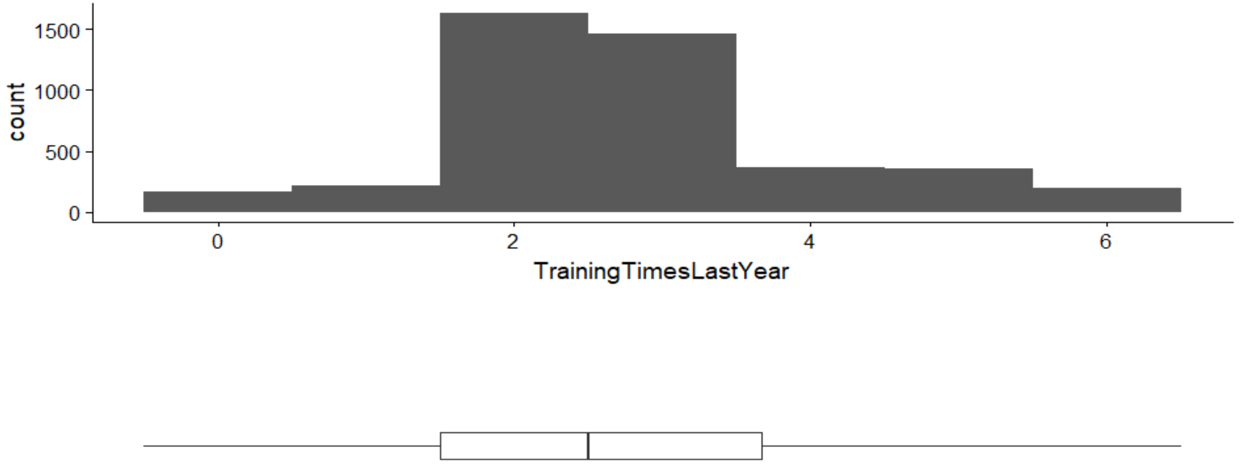
EDA PLOTS

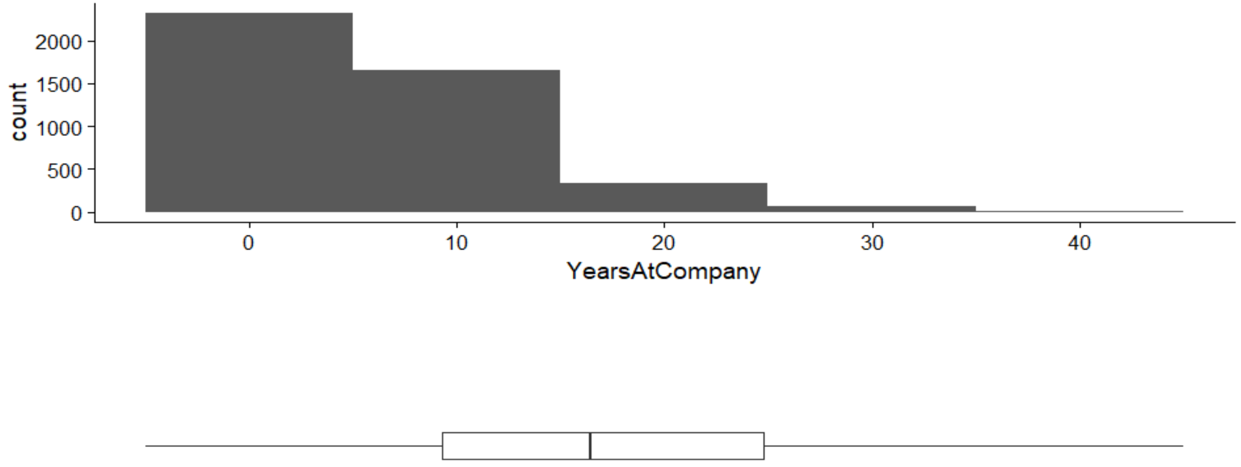

count
2000
1000
0
20
40
60
Age
count
1500
1000
500
0
0
10
20
30
DistanceFromHome
count
1500
1000
500
0
0
50000
100000
150000
200000
MonthlyIncome
count
2000
1000
0
0
5
10
NumCompaniesWorked

EDA PLOTS

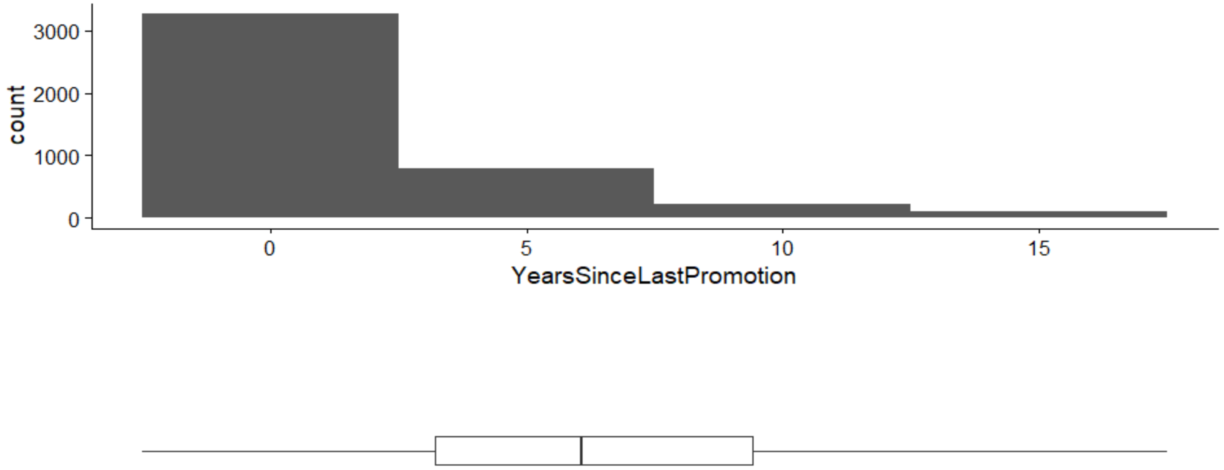

count
2000
1500
1000
500
0
10
15
20
25
PercentSalaryHike

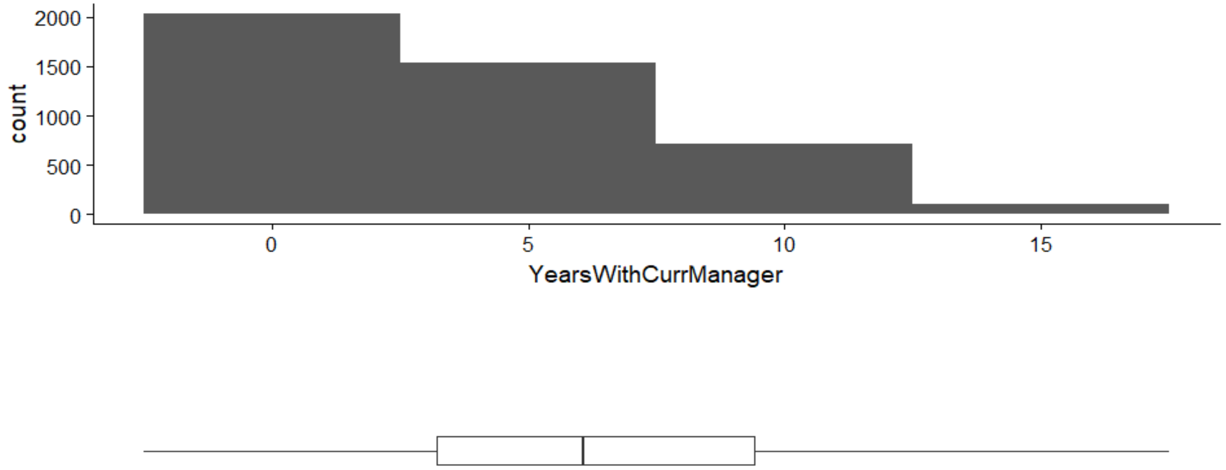

count
1500
1000
500
0
0
10
20
30
40
TotalWorkingYears

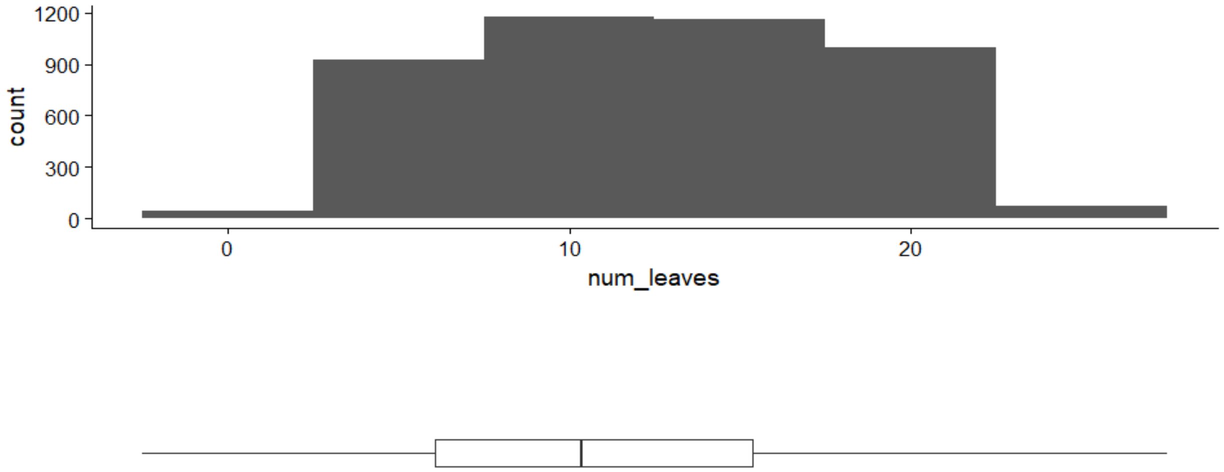

count
1500
1000
500
0
0
2
4
6
TrainingTimesLastYear


count
2000
1500
1000
500
0
0
10
20
30
40
YearsAtCompany

EDA PLOTS


count
3000
2000
1000
0
0
5
10
15
YearsSinceLastPromotion


count
2000
1500
1000
500
0
0
5
10
15
YearsWithCurrManager


count
1200
900
600
300
0
0
10
20
num_leaves

DATA PREPARATION

Once EDA was done, we had to do prepare data before building models on it. The Data preparation mainly included the following two steps:

1) FEATURE STANDARDISATION
   - Feature Standardisation is done on numeric data/ variables. This is because, the range of variables vary by a large extent. For example, the range of monthly income was from 20,000 to 2,00,000, whereas thh range of Age was 18 to 60. And some variables even had a lower range than that. Hence, the variables with higher range would create an unrequired impact on the model. To nullify that impact and to bring everything under the same scale, we need to standardise these variabels with scale function().

2) DUMMY VARUABLES CREATION.
   - The next step was creating dummy variables for all categorical columns excluding the target variable column - Attrition. Dummy variable is an artificial variable created to represent an attribute with two or more distinct categories/levels through numeric values of 1's and 0's.

Finally, the converted variables were all merged into one file as master file. The master file had 58 variables and 4382 observations including the target variable "Attrition".

Before moving on further, we converted the target variable levels from "YES", "NO" to 1,0. To be more specific, "YES = 1" & "NO = 0"

After that, the master file was divided in the 70:30 ratio, were 70% data set was taken for training and 30% data set was kept aside for testing the model on it later.

Thus, this way, we prepared the data and were ready to build models on it.

MODEL BUILDING

We began the model building process initially by taking the target variable i.e. Attrition and training it using all the variables with glm() function. That was the first model.

Next, we removed the most insignificant variables through in built stepAIC function/method. The stepAIC() function begins with a full or null model, and methods for stepwise regression can be specified in the direction argument with character values “forward”, “backward” and “both”. For our code, we set the direction argument to “both”.

The stepAIC() method provided us a standard model by removing the most insignificant variables through iterative process. But the model provided by step AIC was just a head-start to quickly get rid of most insignificant variables, however, it was not the optimal model that we can use for evaluation.

Hence, now we had to proceed with further model building by checking the p-values and VIF values. The p values provide a measure of significance, whereas the VIF are used to check multi-colinearity of the model. Basically, the final target is to get to a model which includes only highly significant variables with minimum correlation. Thus, the thumb rule is: higher VIF and higher p-values are a strong indicators for removing that variable. Following that thumb rule, we continued to remove the variables step by step until we reached to a stage where all the variables were extremely significant, and at the same time the were not much correlated.

The final model looked like this :

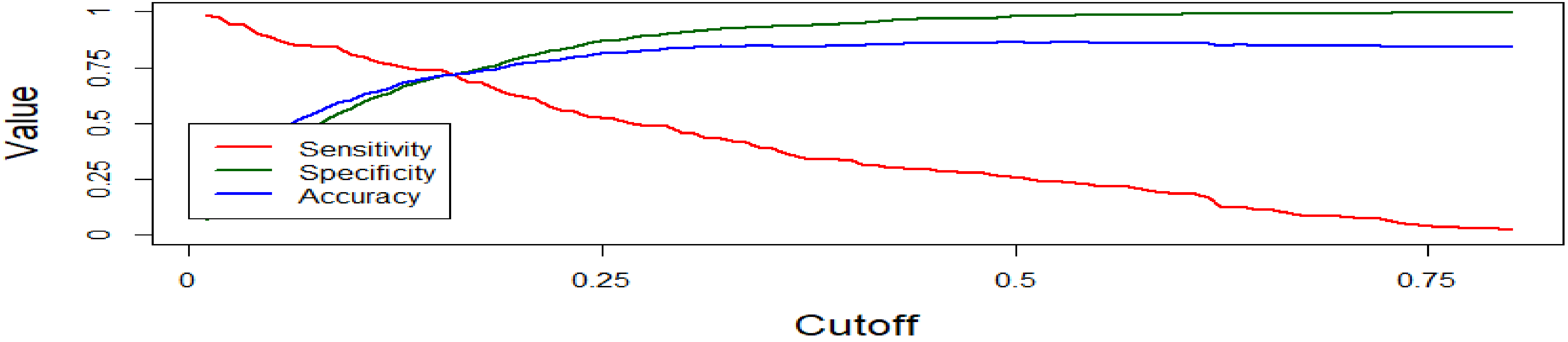
FINAL MODEL :

```
glm(formula = Attrition ~ Age + NumCompaniesWorked + TotalWorkingYears + YearsSinceLastPromotion +
YearsWithCurrManager + BusinessTravel.xTravel_Frequently + JobRole.xManufacturing.Director +
MaritalStatus.xSingle + EnvironmentSatisfaction.xMedium + EnvironmentSatisfaction.xHigh +
EnvironmentSatisfaction.xVery.High + JobSatisfaction.xMedium + JobSatisfaction.xHigh +
JobSatisfaction.xVery.High + emphrs.xStandard.Hours + emphrs.xUnder.Time, family = "binomial", data = train)
```

MODEL EVALUATION / VALIDATION

For evaluating the model, we observed the model behaviour for different cut-off values and compared its behaviour to the model with optimal cut-off value. Below is the table comparing the sensitivity, specificity and accuracy of the model for different cut-off levels including the optimal cut-off value, and the difference in models is clearly visible. Also, a graph is presented below showing the behaviour of model for different cut-off values.

CUTOFF PROBABILITY	ACCURACY	SENSITIVITY	SPECIFICITY
50%	0.8624	0.25943	0.97824
40%	0.8502	0.33491	0.94923
30%	0.8411	0.46226	0.91387
16.16% (OPTIMAL CUTOFF)	0.7217	0.7170	0.7226

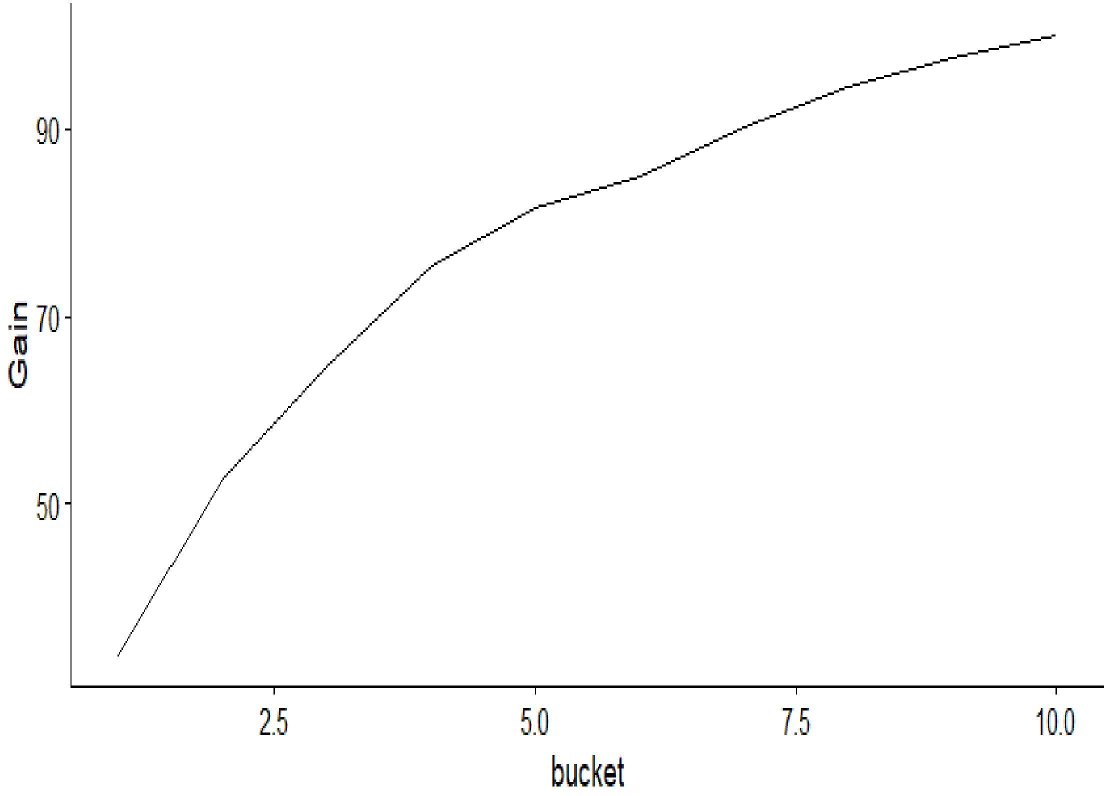


MODEL EVALUATION / VALIDATION

TABLE

Bucket	Observations	Attrition	Cum-Attrition	Gain	Lift
1	132	72	72	34%	3.4
2	131	40	112	52.8%	2.64
3	132	25	137	64.6%	2.15
4	131	23	160	75.5%	1.89
5	132	13	173	81.6%	1.63
6	131	7	180	84.9%	1.42
7	132	11	191	90.1%	1.29
8	131	9	200	94.3%	1.18
9	132	7	207	97.6%	1.08
10	131	5	212	100%	1.00

GAIN CHART

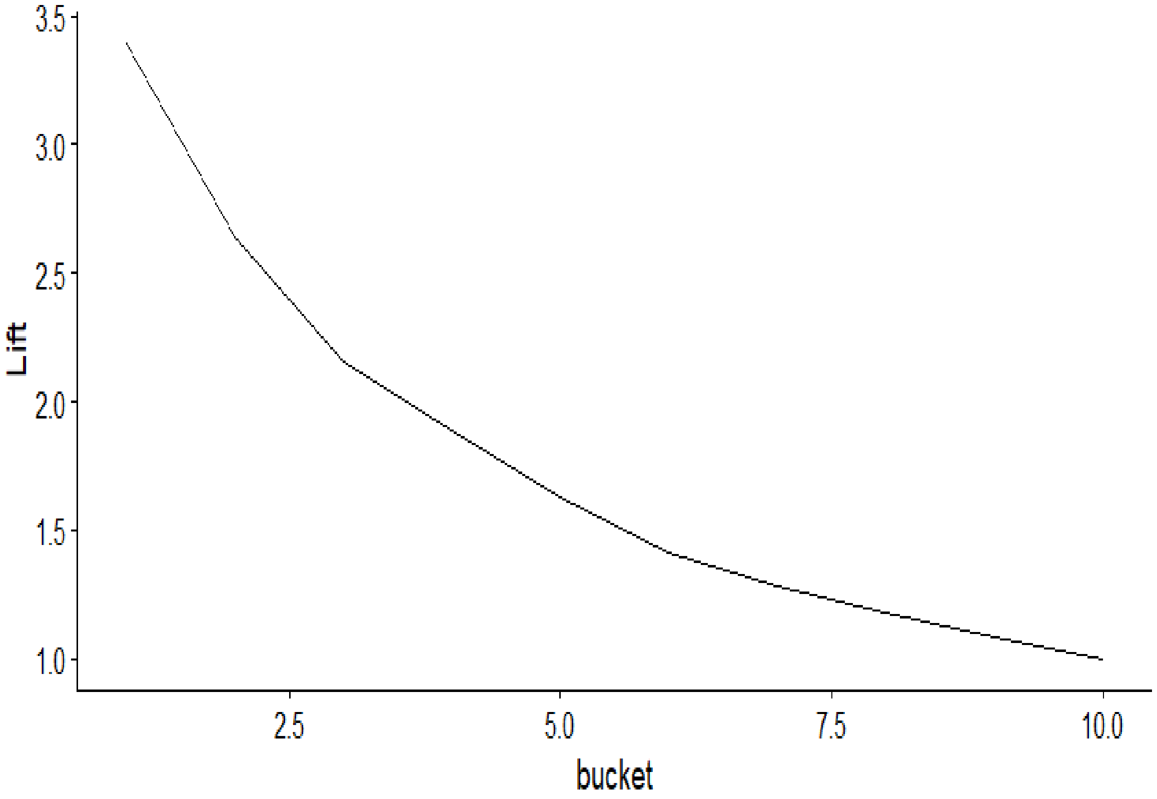


MODEL EVALUATION / VALIDATION

TABLE

Bucket	Observations	Attrition	Cum-Attrition	Gain	Lift
1	132	72	72	34%	3.4
2	131	40	112	52.8%	2.64
3	132	25	137	64.6%	2.15
4	131	23	160	75.5%	1.89
5	132	13	173	81.6%	1.63
6	131	7	180	84.9%	1.42
7	132	11	191	90.1%	1.29
8	131	9	200	94.3%	1.18
9	132	7	207	97.6%	1.08
10	131	5	212	100%	1.00

LIFT CHART



MODEL SUMMARY

- The Model has a decreasing Lift and Increasing Gain.

- For Model Prepared by taking optimal cut-off values, the Accuracy of the model turns out to be 72%, while the Sensitivity and Specificity are 71% and 72% respectively. Since all the values are above 70%, it is a pretty good model.

- The Model predicts around 75% Attrition within 4th Decile, hence it is nearly 2 times better as compared to any random model.

- The KS-Statistics turns out to be nearly 44%, which indicates that not only the model have all the attrition at the top, it also has all the non-attrition at the bottom.

- Hence considering all the above facts and figures, we can conclude that we have managed to prepare a fairly good model

SIGNIFICANT VARIABLES AND THEIR RELATION WITH THE MODEL

VARIABLE NAME	COEFFICIENTS
Age	-0.258
NumCompaniesWorked	0.316
TotalWorkingYears	-0.587
YearsSinceLastPromotion	0.571
YearsWithCurrManager	-0.552
BusinessTravel.xTravel_Frequently	0.899
JobRole.xManufacturing.Director	-0.797
MaritalStatus.xSingle	1.052
EnvironmentSatisfaction.xMedium	-1.010
EnvironmentSatisfaction.xHigh	-1.005
EnvironmentSatisfaction.xVery.High	-1.327
JobSatisfaction.xMedium	-0.573
JobSatisfaction.xHigh	-0.697
JobSatisfaction.xVery.High	-1.265
emphrs.xStandard.Hours	-1.167
emphrs.xUnder.Time	-1.736

INTERPRETING THE COEFFICIENTS

- Overall there are 16 significant variables with positive and negative coefficients.

- Positive Coefficient means that variable will increase the chances of attrition, whereas a negative coefficient will decrease the chances of attrition.

- For example, the Age variable has a negative coefficient of -0.258. This means that an older employee is less likely to leave the company as compared to the a younger employee. This way, we can interpret all the remaining variables by looking at their coefficients.

- In the last slide, we will be giving the "Recommendations to the Management" by interpreting the coefficients of the significant variables given in the above table.

RECOMMENDATIONS TO THE MANAGEMENT :

As stated in the earlier slide, an older employee is less likely to leave the company than a younger person. Also, an experienced person has lesser chances of leaving the company. Hence the management should keep that in mind while hiring a new employee in the future.

Larger the number of companies a person has worked for in the past, more likely he is to leave this company as well.

An employee who has not been promoted for many years may leave the company. So please ensure timely compliments.

More the years spent by an employee with the same/current manager, lesser are his chances of attrition. Hence avoid switching the managers of employees.

A person who travels frequently for the business work will probably resign soon. Avoid giving a lot of business trips to the same group of employees.

Out of all the Job Roles, the job role of manufacturing director observed the least attrition. Hence, the management need not worry much about these group.

An unmarried person/employee is more likely to switch the company. Hence ensure proper retention polices for them.

In General, an employee who is satisfied with environment and Job will not leave the company. And this factor also usually goes without saying. Hence management should pay serious attention to ensure an employee’s Environment Satisfaction and Job Satisfaction.

Another factor, which is also obvious and got reflected in our analysis is: An employee who works just for standard hours or works under time is less likely to leave the company as compared to the ones who work over time. Thus, management should try to manage things and allot work to the employees in such a way that no employee needs to work after office hours.

Well, that’s the end of out part. Up to the management now.

THANK YOU