PHOENIX GLOBAL EMPLOYEE ATTRITION

RESEARCH PROJECT BY – ALI AHMAD

CONTENTS

- Objective
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Statistical Testing
- Model Building
- Observations
- Different Classifier Models
- Insights & Business Recommendations

OBJECTIVE

- **Business Problem** : Phoenix Global company has been facing the issue of constant employee **attrition.** Close to 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.

- **Goal** : To understand what factors they should focus on, in order to curb attrition in the company. 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.

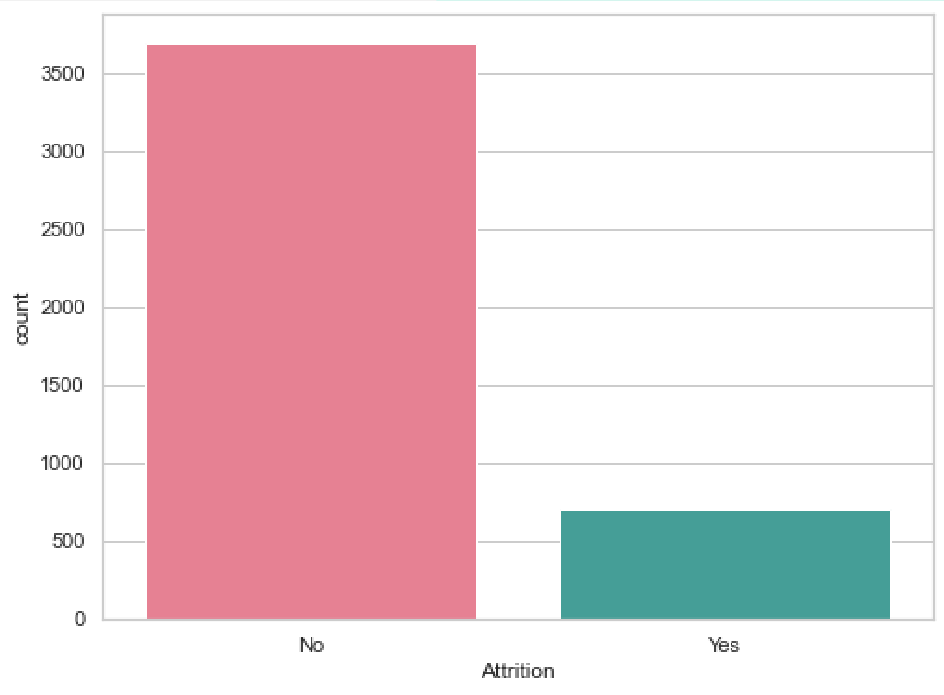
- We are required to model the probability of attrition using a **Logistic Regression**. The results obtained will be used by the management to understand what changes they should make to their workplace. We will also model the probability of attrition using different classifiers like Random Forest, SVM, etc for better results.

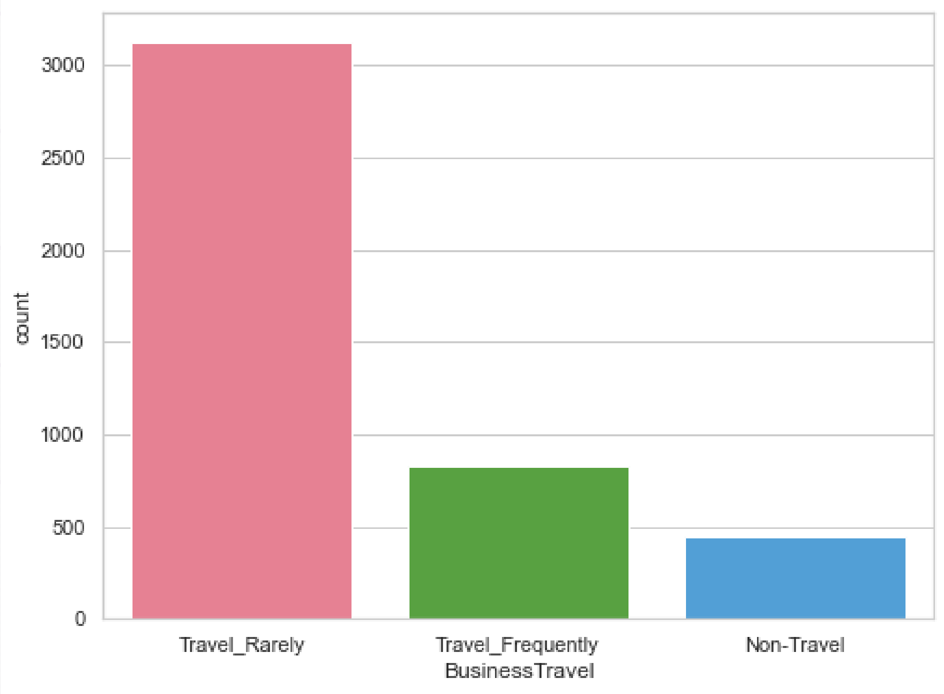
VARIABLES IN DATASET

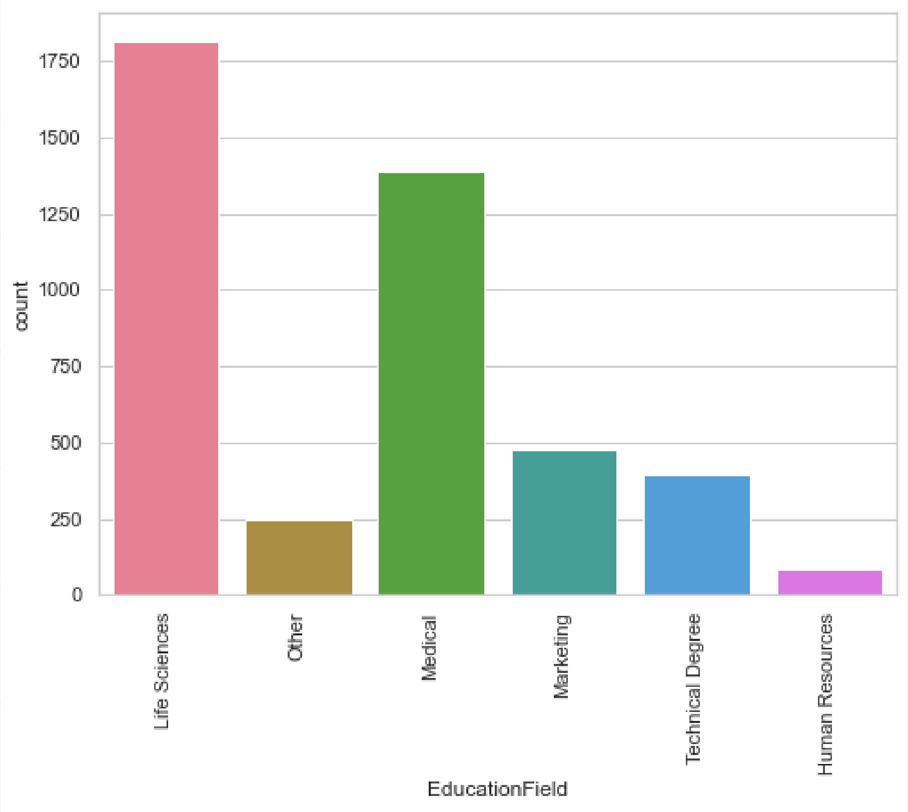
```
'EmployeeID', 'hrs', 'JobInvolvement', 'PerformanceRating',
'EnvironmentSatisfaction', 'JobSatisfaction’,
 'WorkLifeBalance', 'Age', 'Attrition', 'BusinessTravel',
'Department', 'DistanceFromHome’, 'Education',
'EducationField', 'EmployeeCount', 'Gender', 'JobLevel',
'JobRole', 'MaritalStatus’, 'MonthlyIncome',
'NumCompaniesWorked', 'Over18', 'PercentSalaryHike',
'StandardHours', 'StockOptionLevel’,
'TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany', 'YearsSinceLastPromotion',
'YearsWithCurrManager’
```

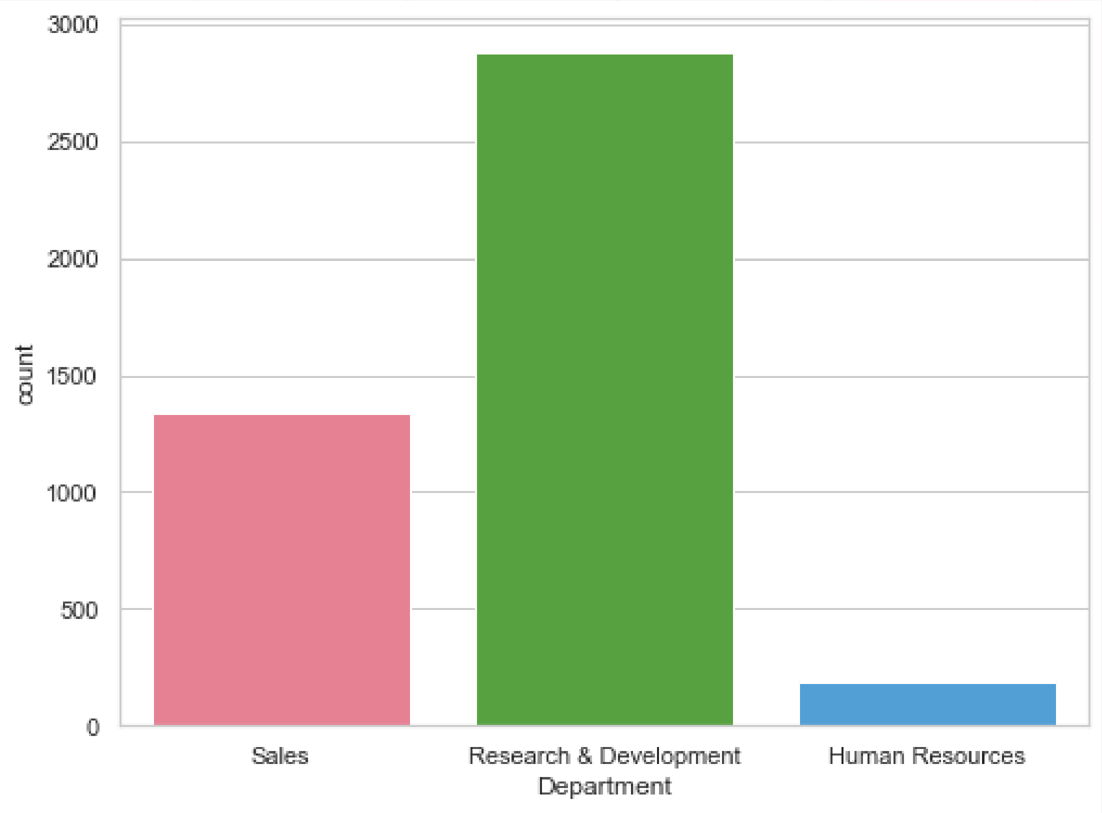
UNIVARIATE ANALYSIS

- Over here we see that maximum employees stayed at the company (No) and less employees have left the company
- After checking the plot, it implies that this Phoenix Global data is imbalanced. We will use SMOTE technique to balance the data.
- We can see that employees who left the company are mostly travelling rarely followed by frequent travel employees



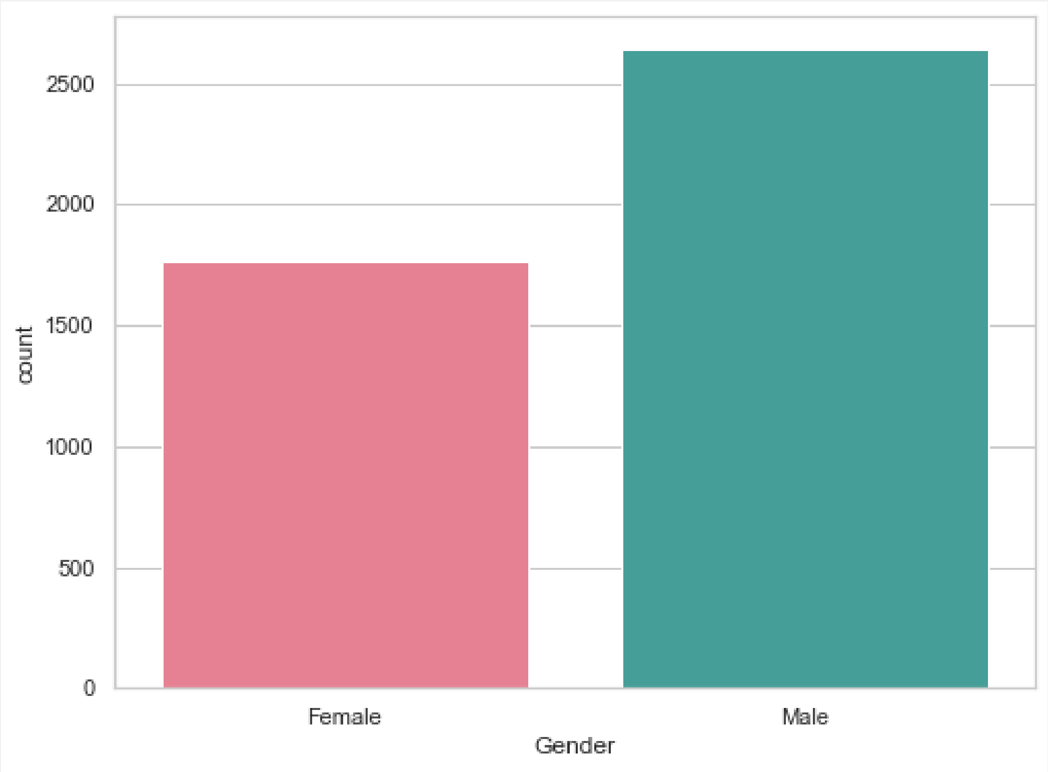


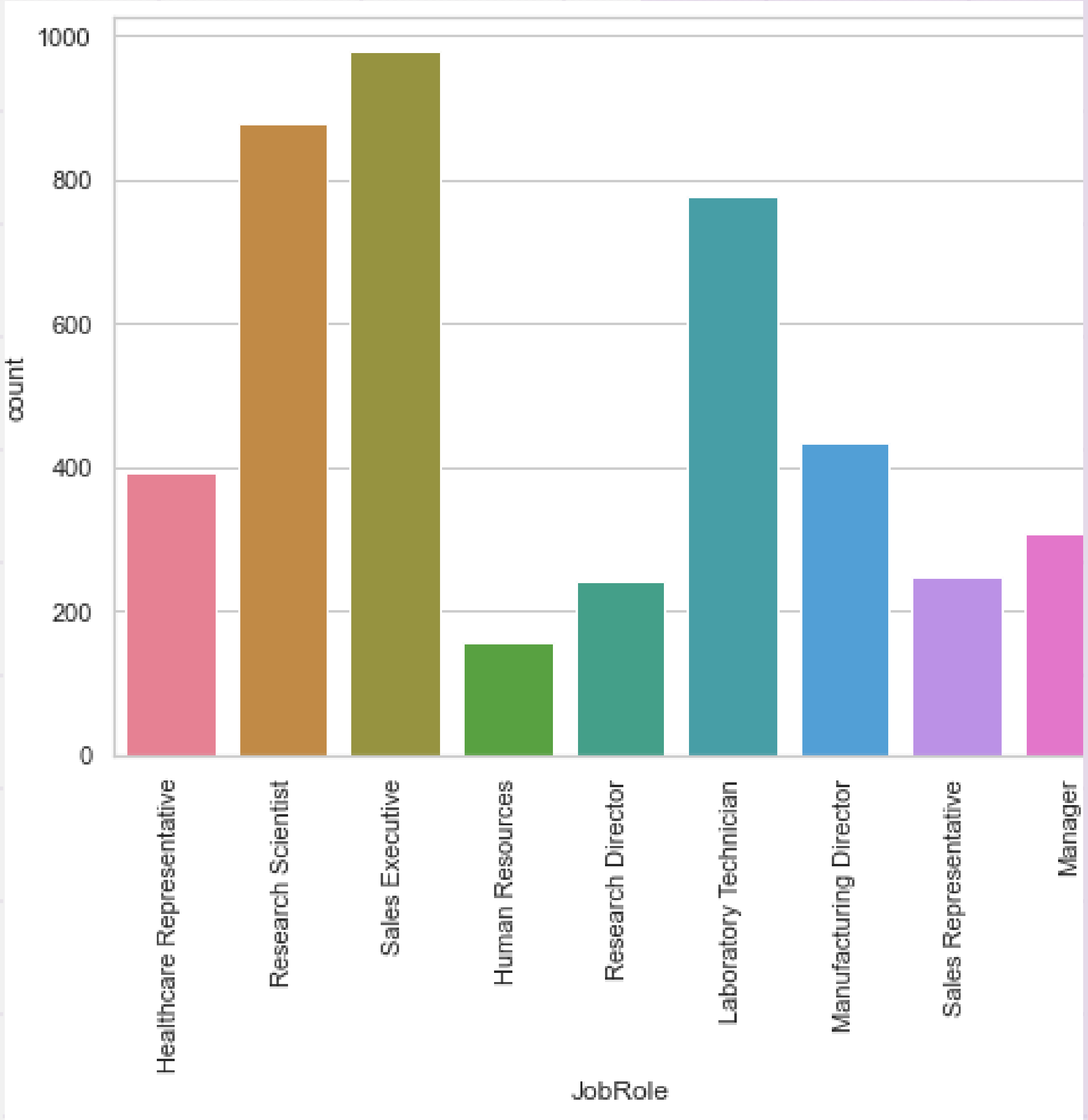




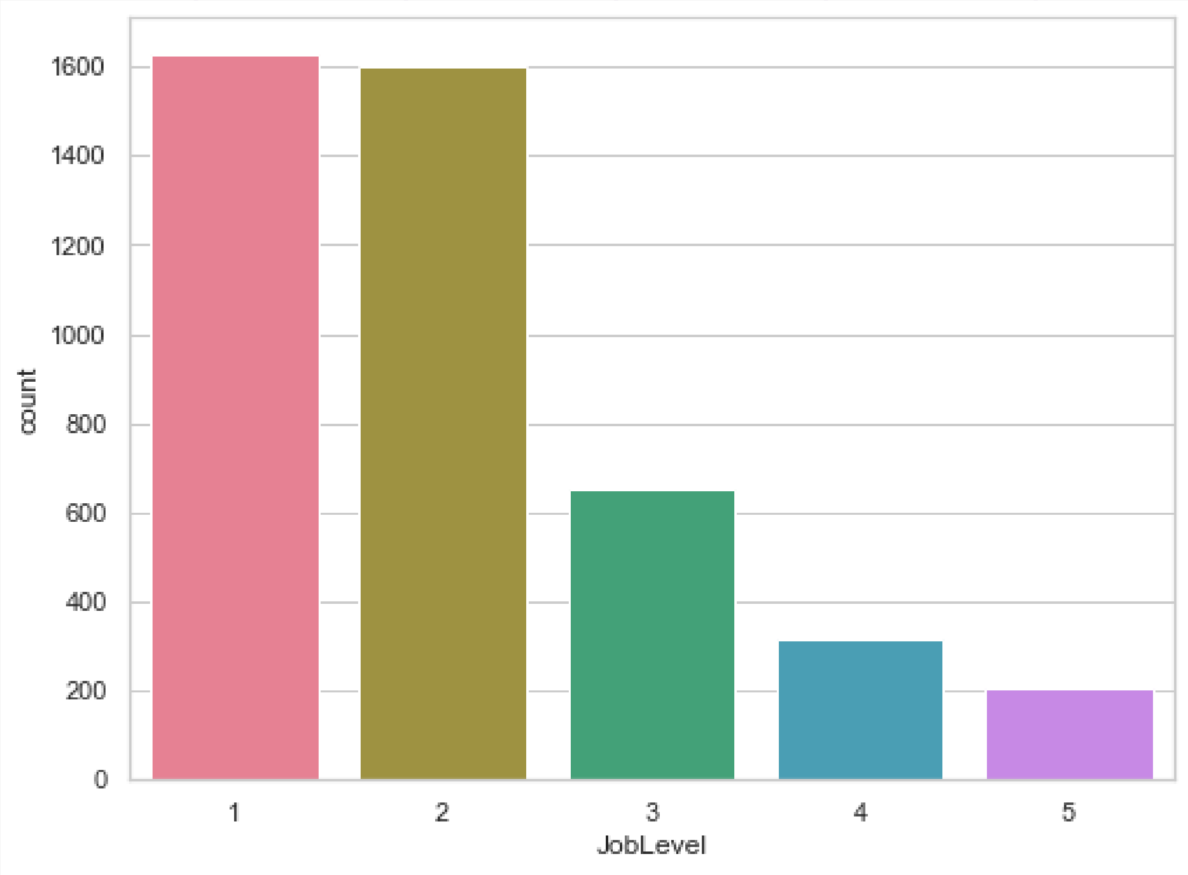
- The maximum attritions is seen where the employee's have the education field in Life Sciences followed by Medical.

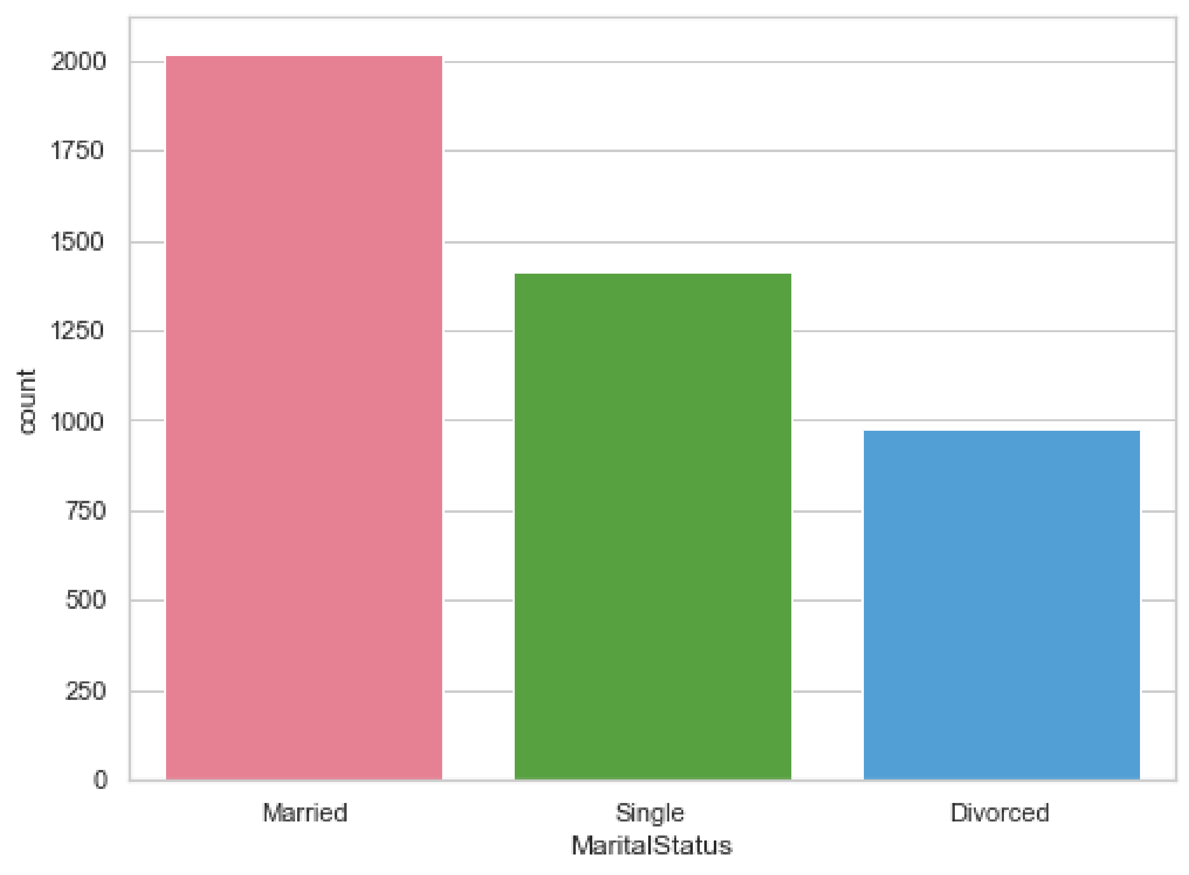
- Over here the employees that worked in the Research & Development Department have most attritions and very less attritions are seen from the Human Resources department





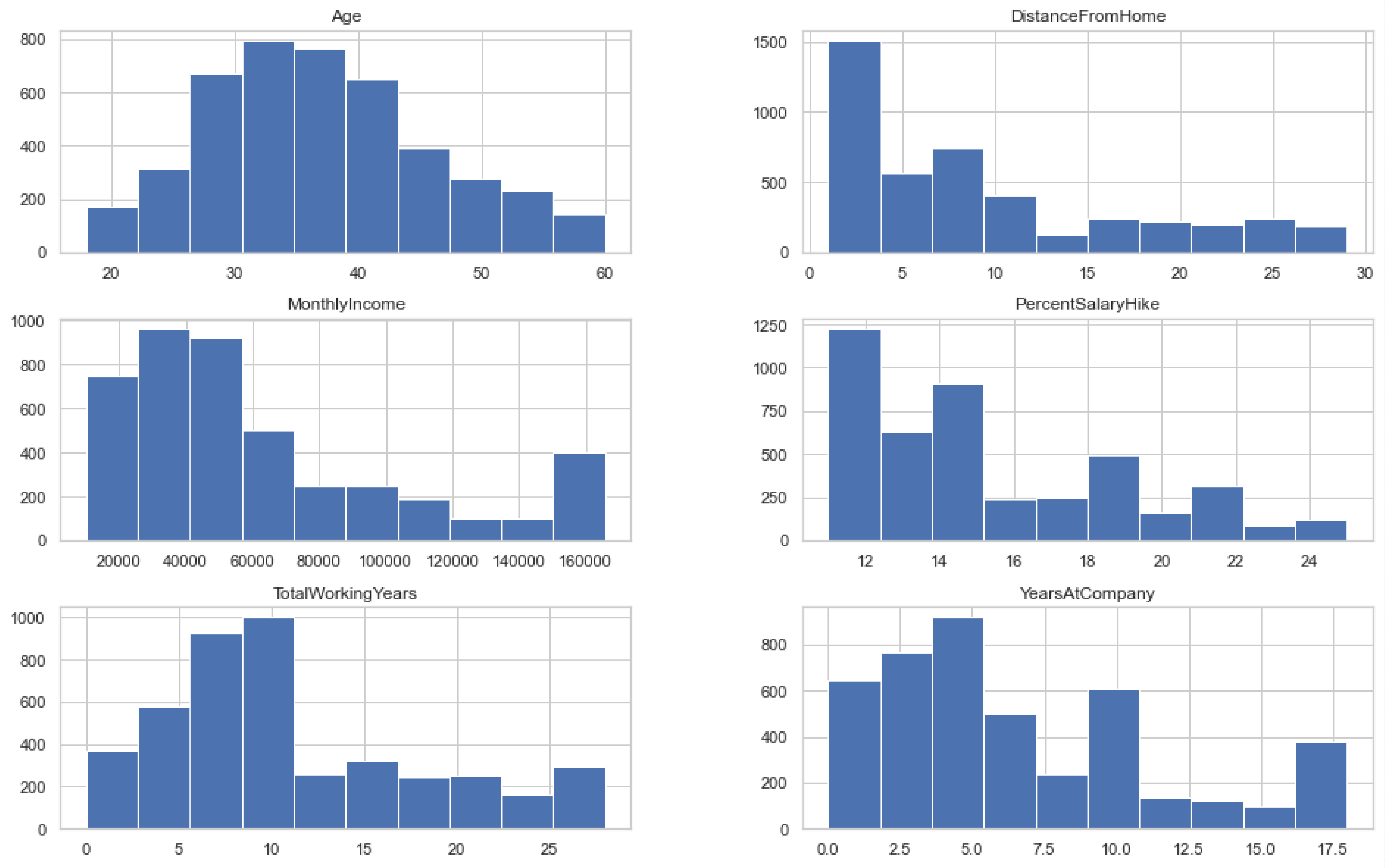
- We can see that most employees' roles who left is Sales Executive followed by Research Scientist and Laboratory Technician
- Most employees who left the company is male





- The most attritions of the employees can be seen in Job Level 1 followed by job level 2.
- The Married employees are most likely to be seen leaving the company.

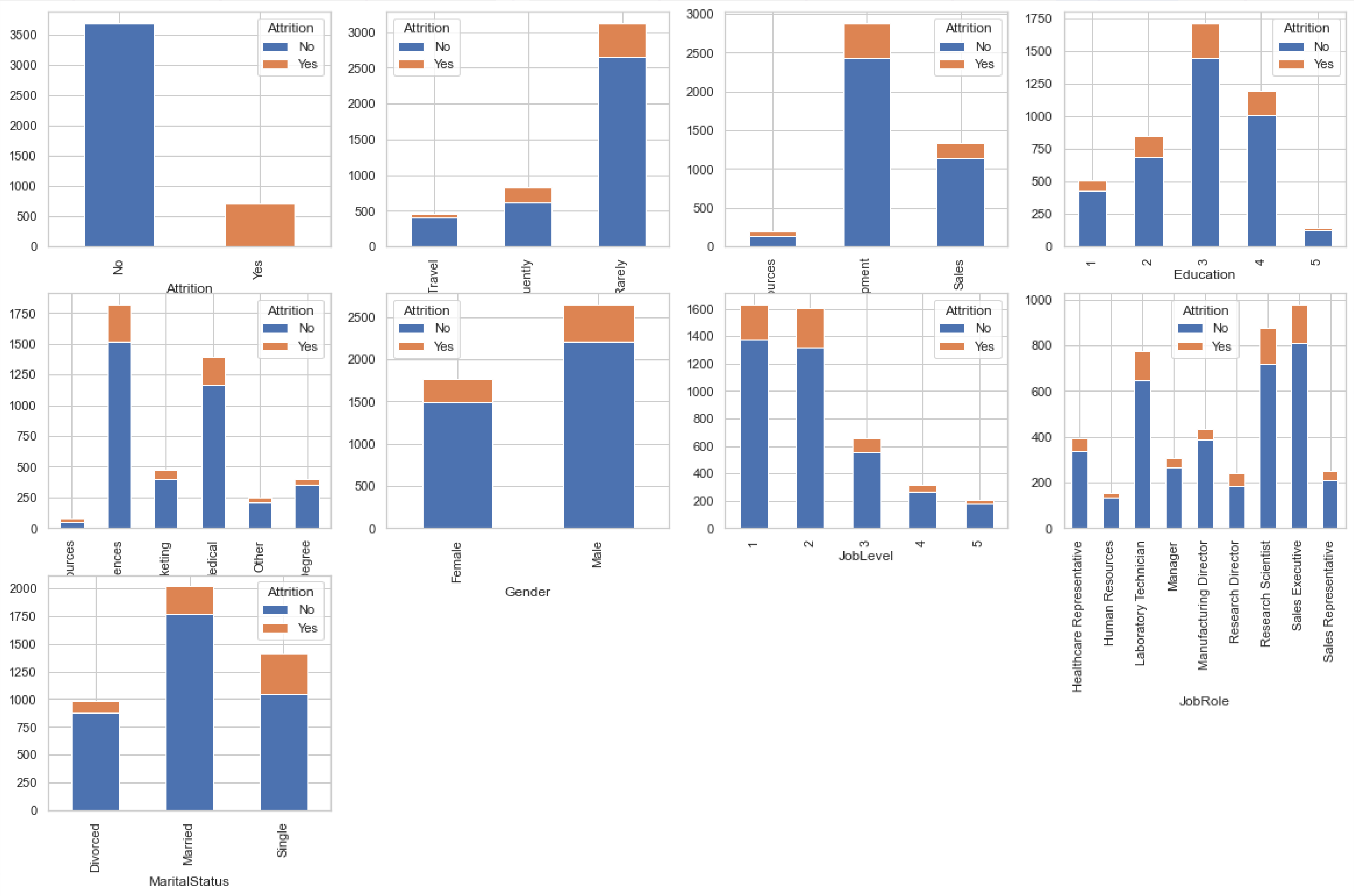
UNIVARIATE ANALYSIS



- Age Feature Distribution is almost Normal Distribution.

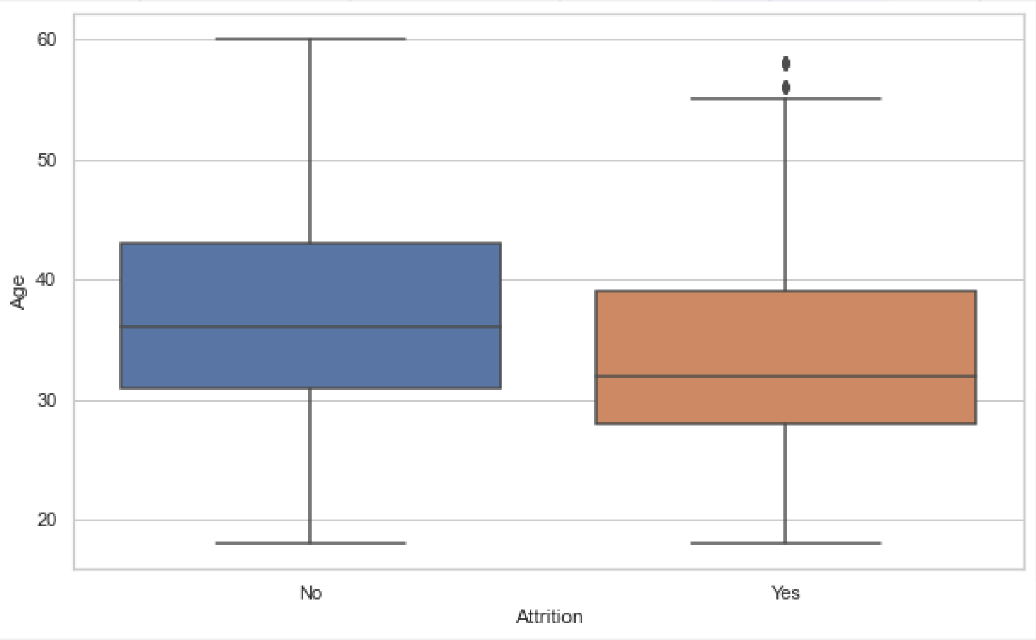
- Most of the Columns are in Skew Distribution form except Age.

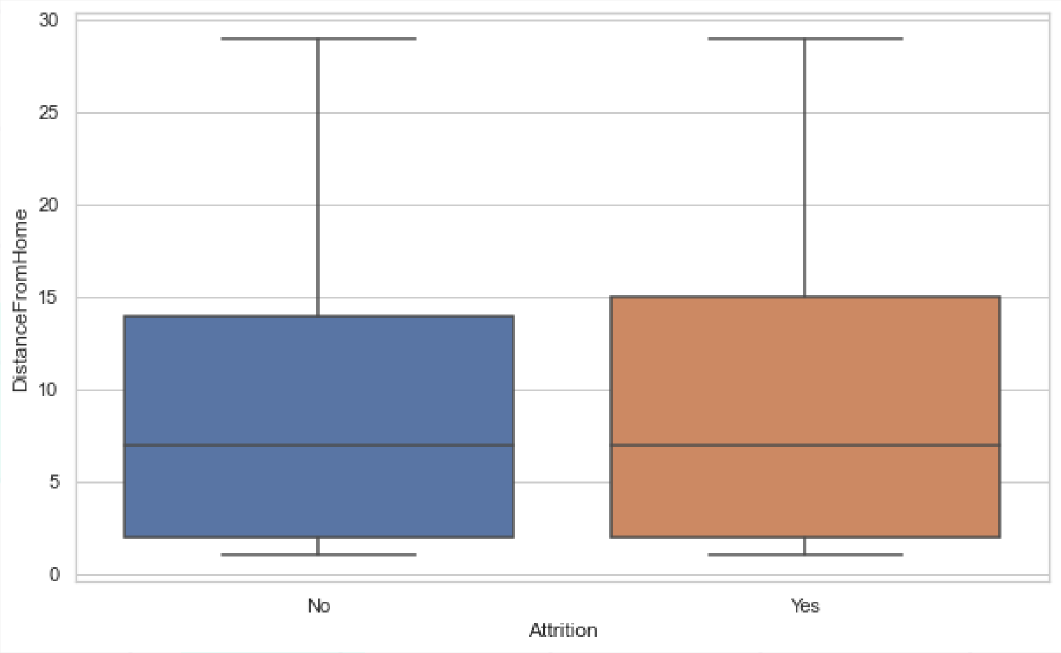
BIVARIATE ANALYSIS

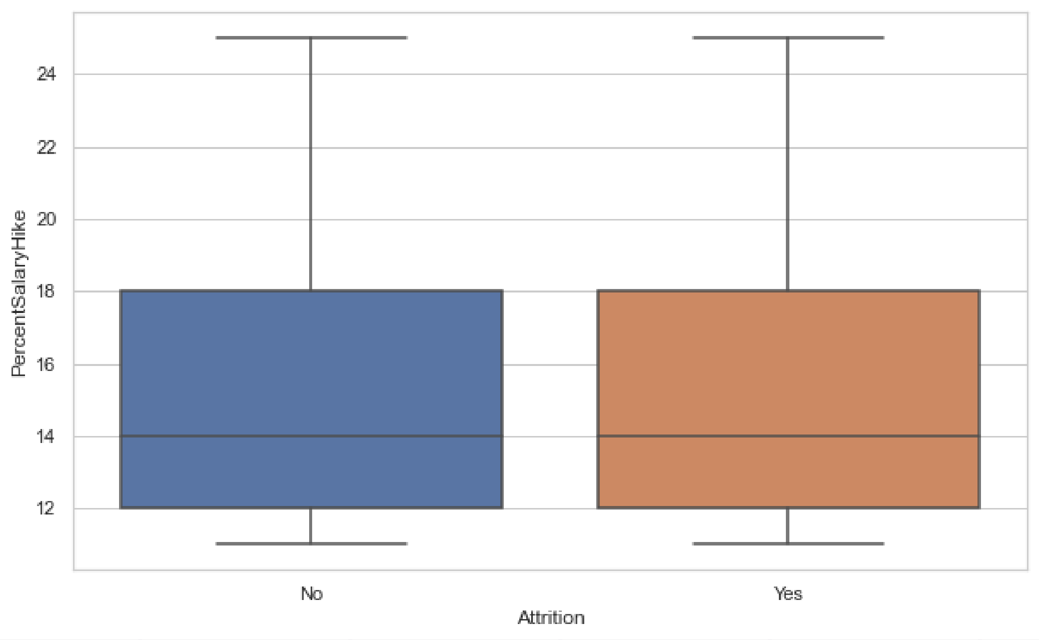


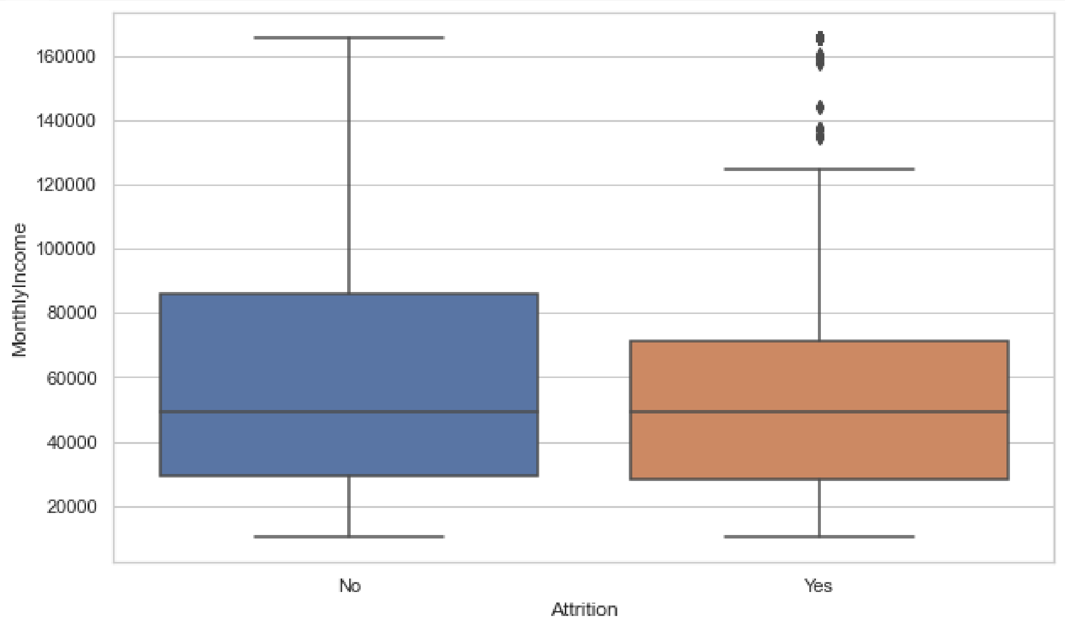
BIVARIATE ANALYSIS : using Boxplots

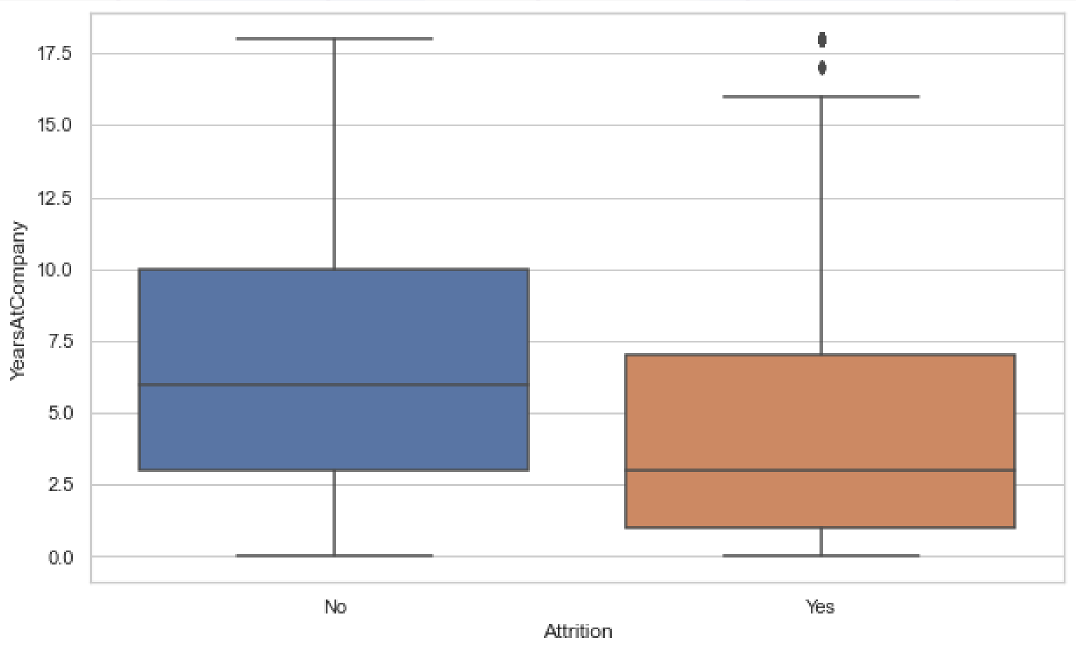
Boxplot shows the distribution of quantitative data that facilitates comparisons between variables or across levels of a categorical variable.

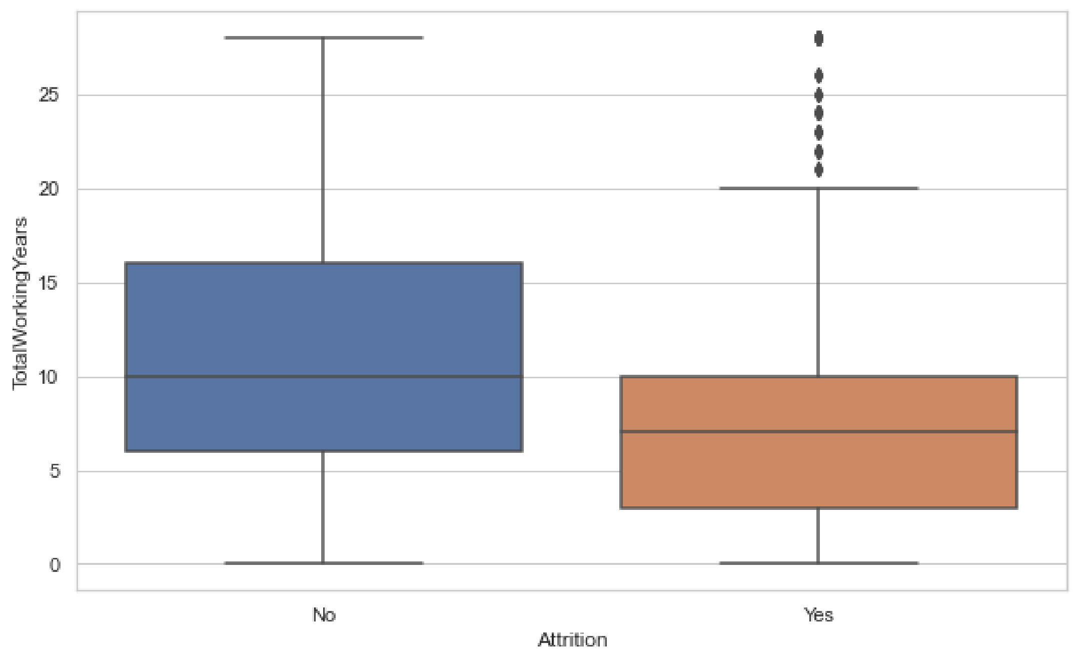


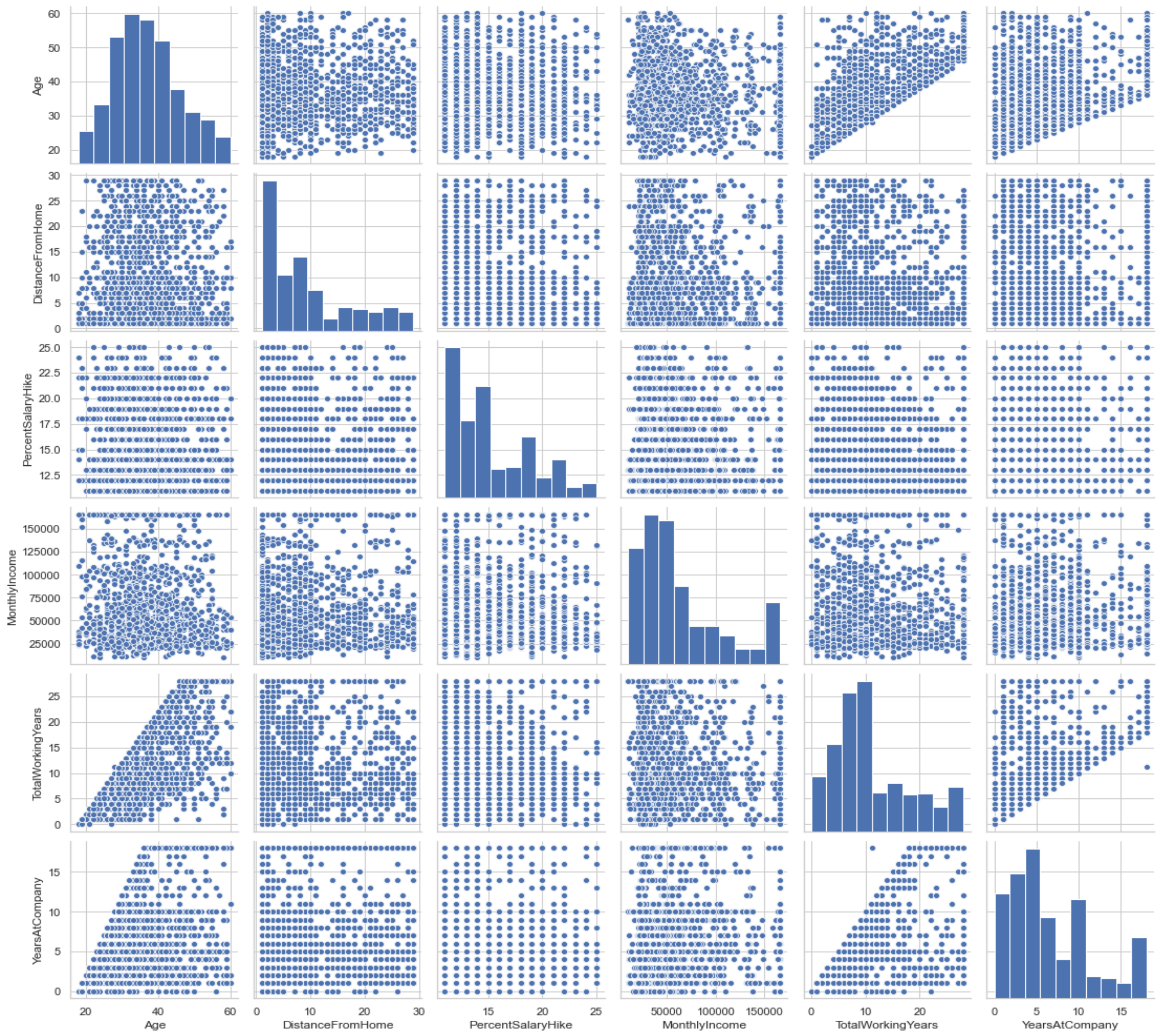





MonthlyIncome
160000
140000
120000
100000
80000
60000
40000
20000
No
Yes
Attrition

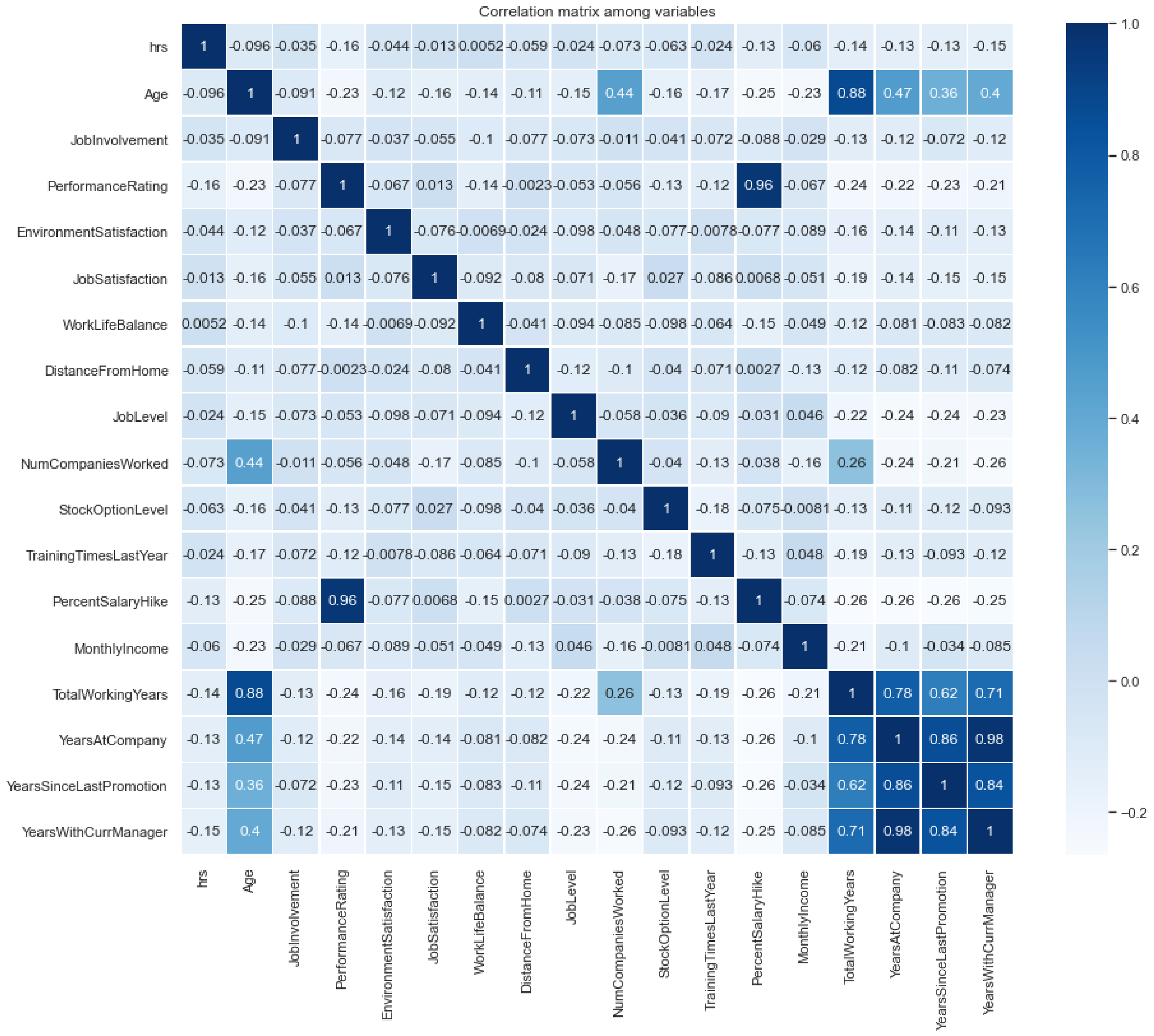

YearsAtCompany
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
No
Yes
Attrition


TotalWorkingYears
25
20
15
10
5
0
No
Yes
Attrition



Multivariate Analysis:
Using Pair plots for
numerical variables

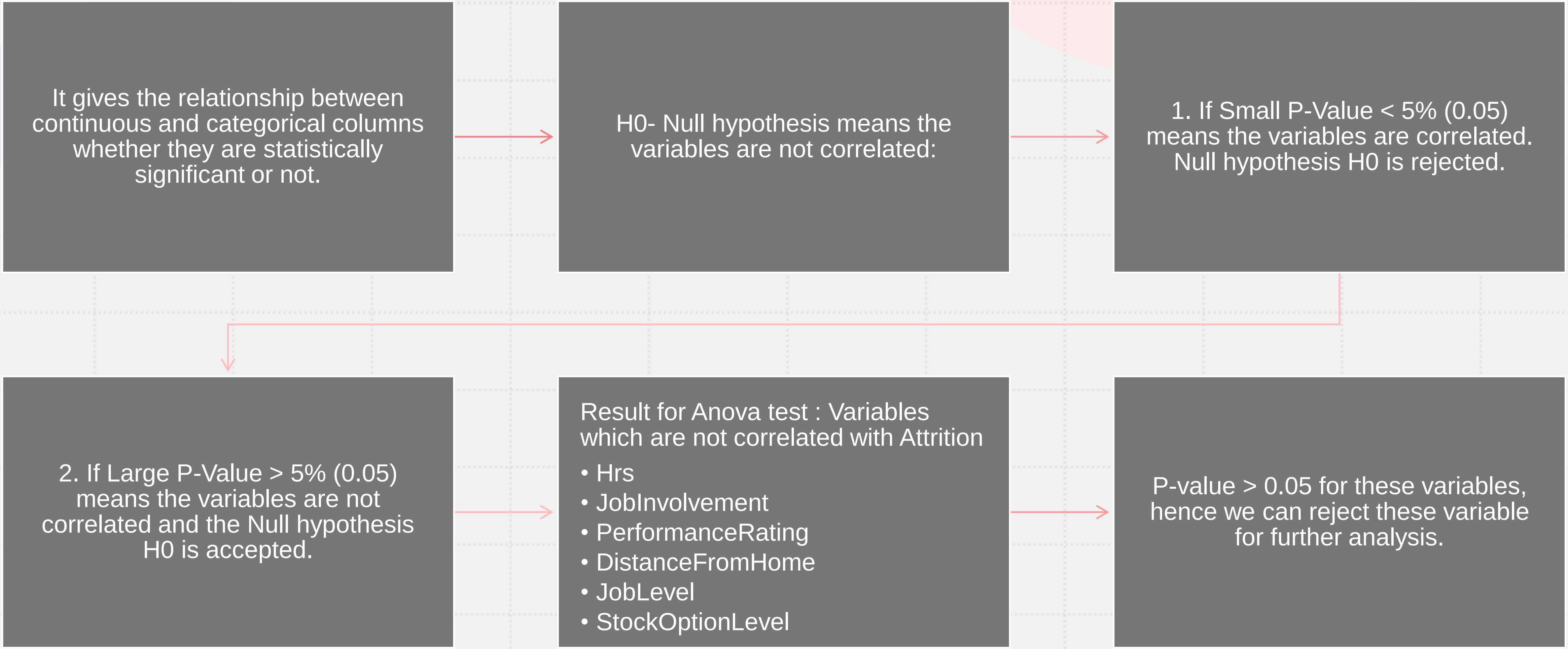
Correlation: Heatmap



Correlation matrix among variables

	hrs	Age	JobInvolvement	PerformanceRating	EnvironmentSatisfaction	JobSatisfaction	WorkLifeBalance	DistanceFromHome	JobLevel	NumCompaniesWorked	StockOptionLevel	TrainingTimesLastYear	PercentSalaryHike	MonthlyIncome	TotalWorkingYears	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager
hrs	1	-0.096	-0.035	-0.16	-0.044	-0.013	0.0052	-0.059	-0.024	-0.073	-0.063	-0.024	-0.13	-0.06	-0.14	-0.13	-0.13	-0.15
Age	-0.096	1	-0.091	-0.23	-0.12	-0.16	-0.14	-0.11	-0.15	0.44	-0.16	-0.17	-0.25	-0.23	0.88	0.47	0.36	0.4
JobInvolvement	-0.035	-0.091	1	-0.077	-0.037	-0.055	-0.1	-0.077	-0.073	-0.011	-0.041	-0.072	-0.088	-0.029	-0.13	-0.12	-0.072	-0.12
PerformanceRating	-0.16	-0.23	-0.077	1	-0.067	0.013	-0.14	-0.0023	-0.053	-0.056	-0.13	-0.12	0.96	-0.067	-0.24	-0.22	-0.23	-0.21
EnvironmentSatisfaction	-0.044	-0.12	-0.037	-0.067	1	-0.076	-0.0069	-0.024	-0.098	-0.048	-0.077	-0.0078	-0.077	-0.089	-0.16	-0.14	-0.11	-0.13
JobSatisfaction	-0.013	-0.16	-0.055	0.013	-0.076	1	-0.092	-0.08	-0.071	-0.17	0.027	-0.086	0.0068	-0.051	-0.19	-0.14	-0.15	-0.15
WorkLifeBalance	0.0052	-0.14	-0.1	-0.14	-0.0069	-0.092	1	-0.041	-0.094	-0.085	-0.098	-0.064	-0.15	-0.049	-0.12	-0.081	-0.083	-0.082
DistanceFromHome	-0.059	-0.11	-0.077	-0.0023	-0.024	-0.08	-0.041	1	-0.12	-0.1	-0.04	-0.071	0.0027	-0.13	-0.12	-0.082	-0.11	-0.074
JobLevel	-0.024	-0.15	-0.073	-0.053	-0.098	-0.071	-0.094	-0.12	1	-0.058	-0.036	-0.09	-0.031	0.046	-0.22	-0.24	-0.24	-0.23
NumCompaniesWorked	-0.073	0.44	-0.011	-0.056	-0.048	-0.17	-0.085	-0.1	-0.058	1	-0.04	-0.13	-0.038	-0.16	0.26	-0.24	-0.21	-0.26
StockOptionLevel	-0.063	-0.16	-0.041	-0.13	-0.077	0.027	-0.098	-0.04	-0.036	-0.04	1	-0.18	-0.075	-0.0081	-0.13	-0.11	-0.12	-0.093
TrainingTimesLastYear	-0.024	-0.17	-0.072	-0.12	-0.0078	-0.086	-0.064	-0.071	-0.09	-0.13	-0.18	1	-0.13	0.048	-0.19	-0.13	-0.093	-0.12
PercentSalaryHike	-0.13	-0.25	-0.088	0.96	-0.077	0.0068	-0.15	0.0027	-0.031	-0.038	-0.075	-0.13	1	-0.074	-0.26	-0.26	-0.26	-0.25
MonthlyIncome	-0.06	-0.23	-0.029	-0.067	-0.089	-0.051	-0.049	-0.13	0.046	-0.16	-0.0081	0.048	-0.074	1	-0.21	-0.1	-0.034	-0.085
TotalWorkingYears	-0.14	0.88	-0.13	-0.24	-0.16	-0.19	-0.12	-0.12	-0.22	0.26	-0.13	-0.19	-0.26	-0.21	1	0.78	0.62	0.71
YearsAtCompany	-0.13	0.47	-0.12	-0.22	-0.14	-0.14	-0.081	-0.082	-0.24	-0.24	-0.11	-0.13	-0.26	-0.1	0.78	1	0.86	0.98
YearsSinceLastPromotion	-0.13	0.36	-0.072	-0.23	-0.11	-0.15	-0.083	-0.11	-0.24	-0.21	-0.12	-0.093	-0.26	-0.034	0.62	0.86	1	0.84
YearsWithCurrManager	-0.15	0.4	-0.12	-0.21	-0.13	-0.15	-0.082	-0.074	-0.23	-0.26	-0.093	-0.12	-0.25	-0.085	0.71	0.98	0.84	1

ANOVA TEST (Analysis of Variance)



It gives the relationship between continuous and categorical columns whether they are statistically significant or not.

H0- Null hypothesis means the variables are not correlated:

1. If Small P-Value < 5% (0.05) means the variables are correlated. Null hypothesis H0 is rejected.

2. If Large P-Value > 5% (0.05) means the variables are not correlated and the Null hypothesis H0 is accepted.

Result for Anova test : Variables which are not correlated with Attrition

- Hrs
- JobInvolvement
- PerformanceRating
- DistanceFromHome
- JobLevel
- StockOptionLevel

P-value > 0.05 for these variables, hence we can reject these variable for further analysis.

Chi-Square Test

Chi-square is a quantitative measure used to determine whether relationship exists between two categorical variables.
H0- Null hypothesis means the variables are not correlated:
1. If Small P-Value < 5% (0.05) means the variables are correlated. Null hypothesis H0 is rejected
2. If Large P-Value > 5% (0.05) means the variables are not correlated and the Null hypothesis H0 is accepted

The variables (p-value < 0.05) that are found significant from Chi-square test:

```
['BusinessTravel_Non-Travel', 'BusinessTravel_Travel_Frequently',
'BusinessTravel_Travel_Rarely', 'Department_Human Resources’,
 'EducationField_Human Resources', 'EducationField_Technical Degree',
'JobRole_Manufacturing Director', 'JobRole_Research Director’,
 'MaritalStatus_Divorced', 'MaritalStatus_Married', 'MaritalStatus_Single']
```

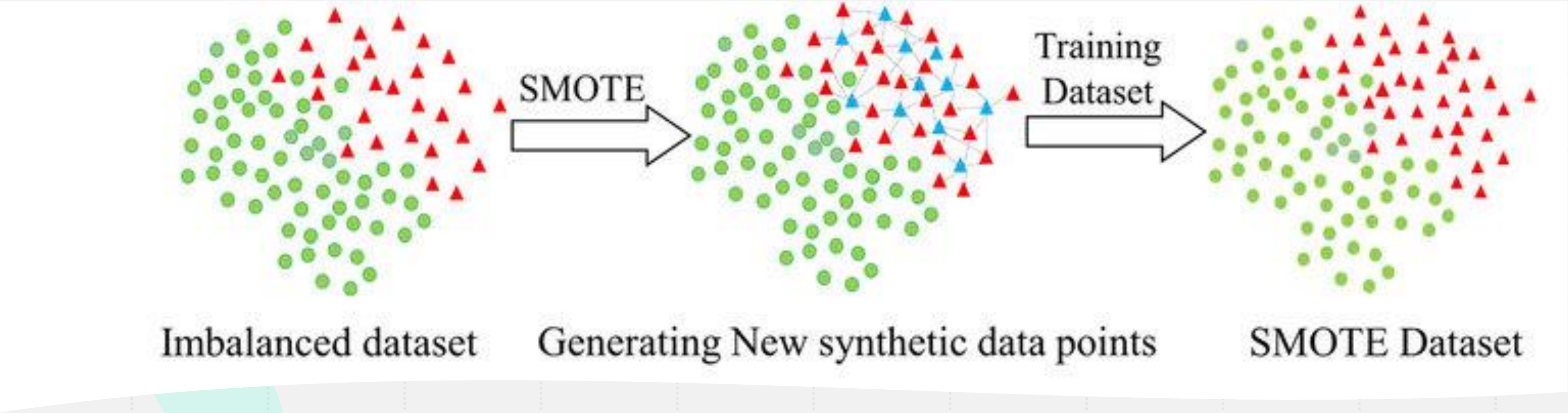
The variables that were found with large p-value i.e. p-value > 0.05. So, we can reject these variables.

```
['JobRole_Sales Representative', 'JobRole_Sales Executive', 'JobRole_Research
Scientist', 'JobRole_Manager’,
 'JobRole_Laboratory Technician', 'JobRole_Human Resources', 'JobRole_Healthcare
Representative', 'Gender_Male', 'Gender_Female’,
'EducationField_Other', 'EducationField_Medical', 'EducationField_Marketing’,
 'EducationField_Life Sciences', 'Department_Sales', 'Department_Research &
Development']
```

LOGISTIC REGRESSION MODEL BUILDING

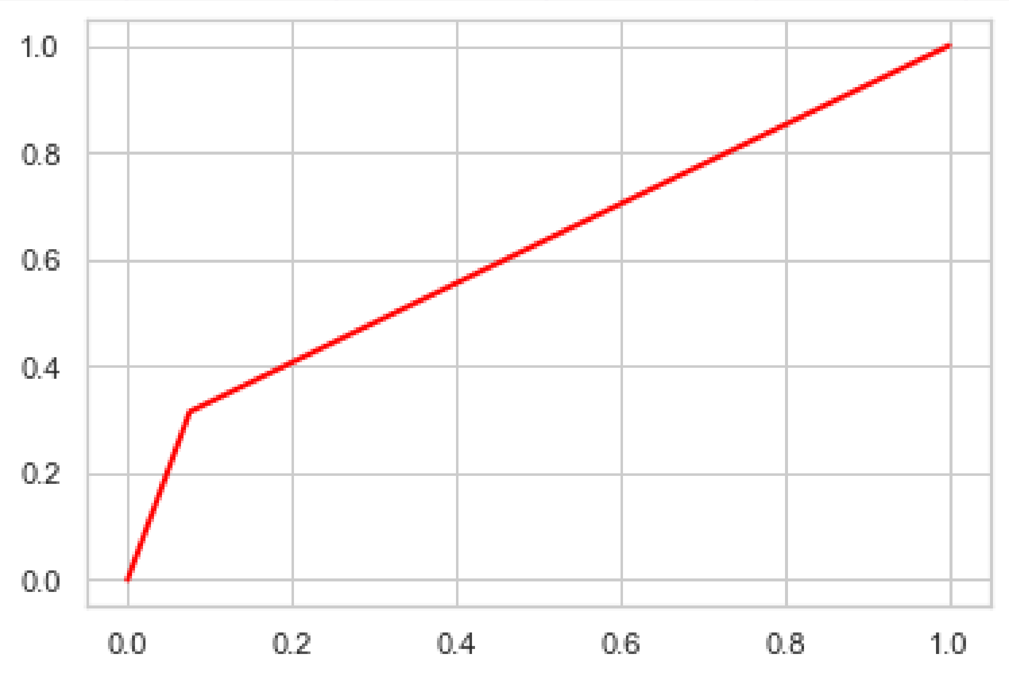
For model building, Scikit-Learn model is used to predict the attrition rate of the employees.

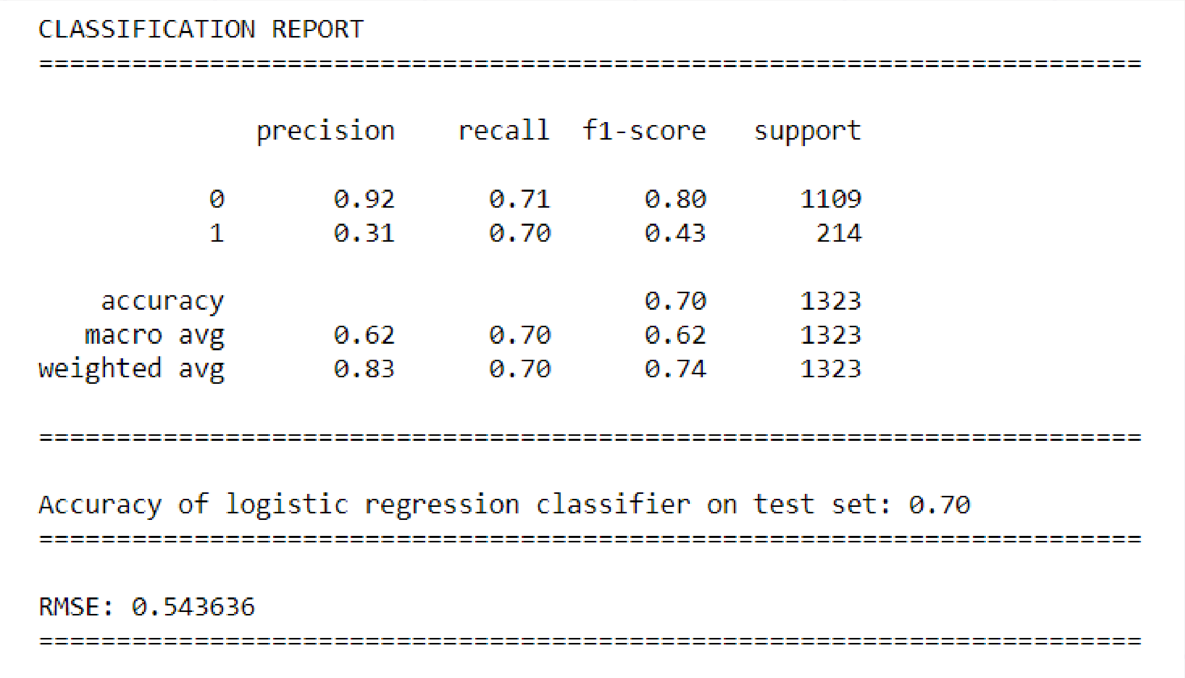
SMOTE technique with oversampling is used to balance the dataset.



OBSERVATIONS

- The accuracy of logistic regression classifier is coming around 70%.
- Root Mean Square Error (RMSE) of logistic regression is 0.54.
- The ROC AUC score for logistic regression model is approximately 62%



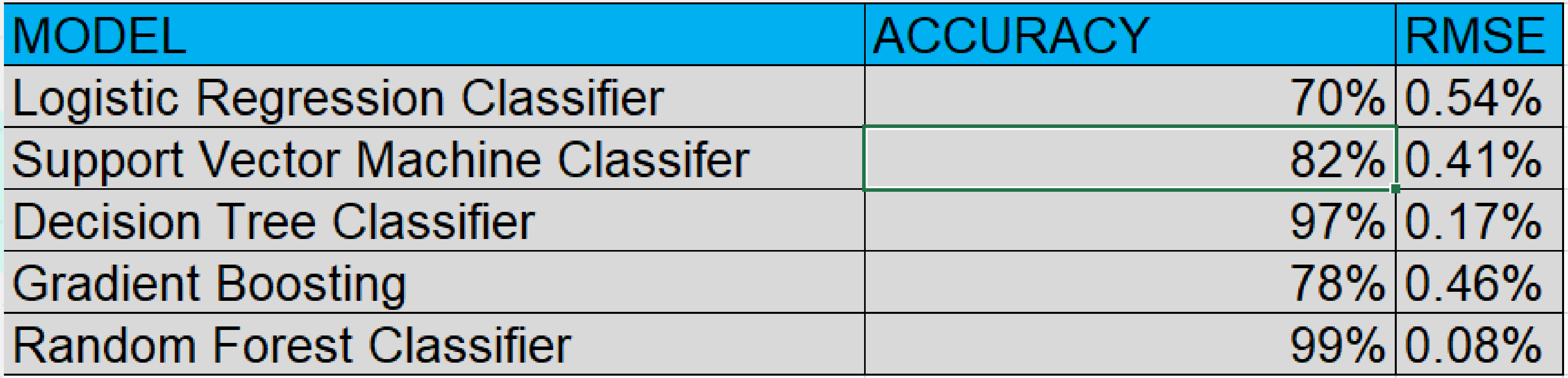


```
CLASSIFICATION REPORT
=======================================================================

              precision    recall  f1-score   support

           0       0.92      0.71      0.80      1109
           1       0.31      0.70      0.43       214

    accuracy                           0.70      1323
   macro avg       0.62      0.70      0.62      1323
weighted avg       0.83      0.70      0.74      1323

=======================================================================

Accuracy of logistic regression classifier on test set: 0.70
=======================================================================

RMSE: 0.543636
=======================================================================
```

CLASSIFIER MODELS TO ACHIEVE BETTER ACCURACY

- *Support Vector Machine Classifier (SVM)* : SVC is a linear model and creates a line or a hyperplane which separates the data into classes.
- *Decision Tree Classifier* : DTs are a non-parametric supervised learning method used for classification and regression.
- *Gradient Boosting* : Gradient boosting is another CART ensemble algorithm which relies on the assumption that the best possible next model.
- *Random Forest Classifier* : The random forest is an ensemble learning algorithm which builds upon the Classification And Regression Tree (CART) paradigm. It uses bagging and feature randomness

Below are the results from different classification models to achieve a better accuracy and rmse :-



MODEL	ACCURACY	RMSE
Logistic Regression Classifier	70%	0.54%
Support Vector Machine Classifer	82%	0.41%
Decision Tree Classifier	97%	0.17%
Gradient Boosting	78%	0.46%
Random Forest Classifier	99%	0.08%

Insights and Business Recommendations

INSIGHTS :

- From feature importance, we find that Age appears to be the most important feature followed by the employee's Total Working Years, Monthly Income and Years At Company.
- Digging further we found also that the employees who left the company worked in the Research & Development department and Sales department and more specifically those who had Sales executive, Research scientist and Laboratory technicians as their Job role.
- We observed that most of the employees who left the company has normal Work Life Balance and employees with education level 3 and 4 had most attritions and the least attritions were seen with employees having education level 5.
- Most attritions is seen where the employee is travelling rarely.
- If we go by Random forest or Decision tree approach for model building, we will yield a nice accuracy and less root mean square error. Here we are getting a 99% accuracy and less than 10 % Rmse from Random forest model approach.

RECOMMENDATIONS :

- So my recommendation is to check the job satisfaction and environment satisfaction to lower the risk of attrition.
- Employees who didn't get promoted since last 5 years have higher risk of attrition.
- More years an employee spends with the manager lesser the risk of attrition.
- If the employee travels frequently then it can lead to a higher risk of attrition.

THANK YOU

ivy™

Professional School

ISO 9001 : 2000 Organisation