


Final Team Project

Employee Attrition:

 *Fictional dataset on HR Employee attrition and performance*

GOLD TEAM 8 | MSIS 510

Introduction To Data Mining And Analytics

December 6, 2021

Team Members: Sukalpa Datta | Nisha Fotedar | Manali Deb | JF Seide |
Conlan Mcnamee | Kiran Vanam

SUMMARY

Objectives

Objective:

Workforce retention is a critical aspect for the long-term success of any organization. Losing workforce leads to delay in completion of projects and needs additional time, effort, and money to hire and train new employees. On an average, organizations lose 12-16% of the workforce every year. This is true even to sample data used in this analysis. Attrition can be attributed to many causes. For instance, employees may leave because they are looking for better opportunities, negative working environments, toxic management, work-related accidents, sickness (or death), or excessive working hours that affect the employee's ability to have a more balanced worklife.

Our goal in this project is to identify the factors that lead to employee attrition and prescribe solutions these organizations can use to retain the talent.

In this paper, we are reviewing the employee attrition rate within a fictional company from Kaggle.com to analyse the data. For our analysis, we will use Attrition as our target variable to:

- Show the parameters the company should improve upon to reduce attrition.
- Predict the most likely factors affecting attrition.
- Run the Models (Classification Tree and Random Forest) to analyse what factors could be the most important contributing factors and what factors are less likely to have an impact.
- To further analyse, if the model could be improved.

DATASET INFORMATION

Dataset Description & Understanding

As a team, we have chosen to work on the following data set: [Employee Attrition: Fictional dataset on HR Employee attrition and performance](#) from Kaggle. It contains 1,470 observations and 35 variables(as imported by R).

| Name | Type | Description |
|------------------|-------------------------|---|
| Age | integer | Age of the employee |
| BusinessTravel | Factor/string(3 levels) | Whether the employee travels and how frequently she travels |
| DailyRate | integer | Rate of an employee per day |
| Department | Factor/String(3 levels) | Department of employee |
| DistanceFromHome | integer | Distance between employee's home and office |

| | | |
|--------------------------|-------------------------|---|
| Education | Factor/String(5 levels) | Highest level of education of the employee |
| Education Field | Factor/String(6 levels) | Employee's field of education |
| EmployeeCount | integer | Count of Employee per record |
| EmployeeNumber | integer | Employee id assigned to an employee |
| EnvironmentSatisfaction | Factor/String(4 levels) | Whether the employee is satisfied with the current work environment |
| Gender | Factor/string(2 levels) | Gender of the employee |
| HourlyRate | integer | Hourly rate of an employee |
| Job role | Factor/string(9 levels) | Job Role of an employee |
| Marital Status | Factor/string(3 levels) | Marital status of an employee |
| Monthly income | integer | Monthly income of an employee |
| MonthlyRate | integer | Calculation based on the cost to the company |
| NumCompaniesWorked | integer | Total number of companies an employee worked |
| Over18 | Factor/string(1 level) | If the employee's age is above 18 years |
| PercentSalaryHike | integer | % of salary hike an employee received |
| PerformanceRating | Factor/string(2 level) | Performance rating of an employee |
| RelationshipSatisfaction | Factor/string(4 level) | Employee's relationship in the workplace |
| StandardHours | integer | Standard working hours of an employee |
| StockOptionLevel | integer | Stocks provided for the no.of years |
| TotalWorkingYears | integer | Total working years of an employee |
| TrainingTimesLastYear | integer | Training provided to an employee in hours |
| Work-life balance | Factor/string(4 levels) | Is the work-life balance of an employee bad, good, better, or best? |
| YearsAtCompany | integer | Number of years the employee spent in this company |
| YearsInCurrentRole | integer | Number of years the employee spent in this current role |
| YearsSinceLastPromotion | integer | Number of years since the last promotion |
| YearsWithCurrManager | integer | Number of years spent with the current manager |
| Attrition | Integer (0 or 1) | 1 means Attrition=YES and 0 means Attrition=NO |

Each row in this dataset represents a unique employee with the corresponding attribute value as shown above.

Data Cleaning and Preprocessing

The following table describe the status and action taken to some of the variables:

| Check name | Status | Action |
|--------------------|-----------|---------------------|
| Missing Values | No | None |
| Duplicated Records | No | None |
| Unit of Measure | Different | Scaling is required |

| | | |
|-------------------------------|---|--|
| Columns with Uniformed Values | ***Yes: "EmployeeCount", "Over18" and "StandardHours" | Drop these columns as they provide no useful information |
| Time series Data | No | None |

We will not consider the following variables in our analysis as they do not help us to predict Attrition:

1. Employee Count: counts the number of employees. It always takes the value 1.
2. Over 18: identifies if an employee is over 18 years of age. It always takes the value "Yes".
3. Standard Hours: standard working hours. It is always 80 for all employees.
4. Employee Number: It represents the employee number assigned to a particular employee.

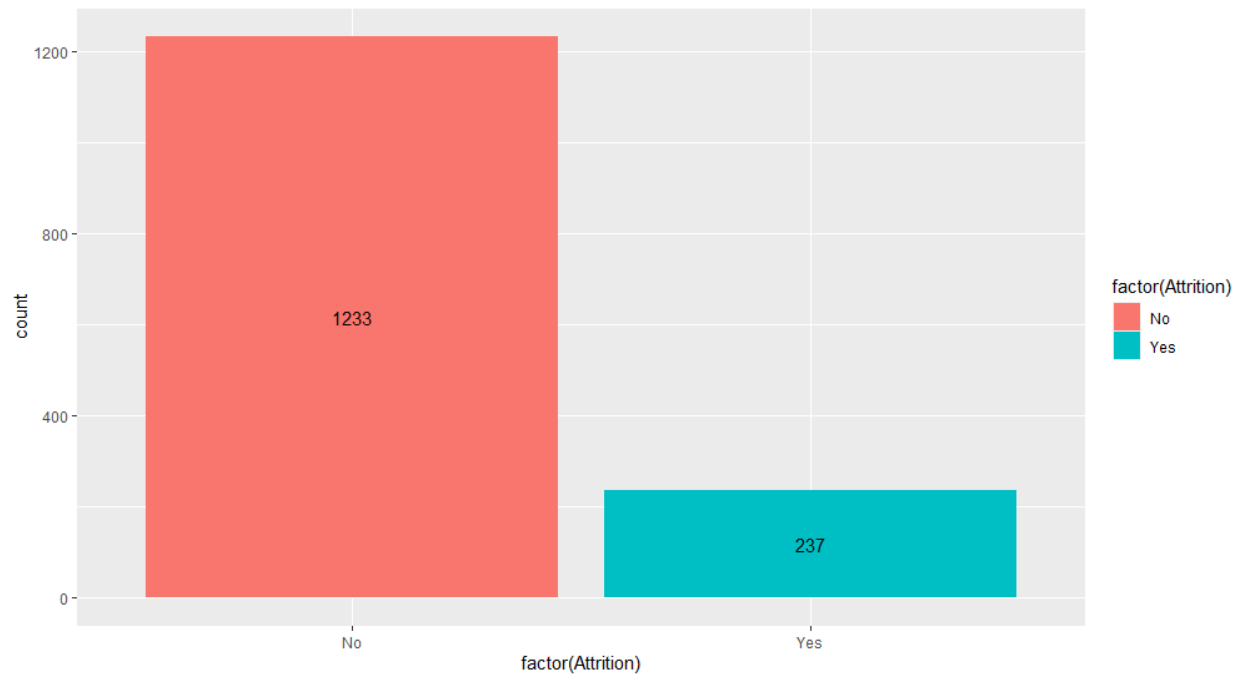
This dataset is fairly clean, as it contains no NULL values and two data types: factors and integers. Hence, no further treatment is required pertaining to the NULL values, which makes it easier to work with the dataset. A lot of the variables have a range from 1-4 or 1-5; with the lower the ordinal variable, the worse and the higher the better. For instance, Job Satisfaction 1 = "Low" while 4 = "Very High".

Attrition is the main label in our dataset; it will help us figure out why employees are leaving the organization. We provided descriptive names to the levels of some categorical variables like Education, EnvironmentSatisfaction, JobInvolvement, JobSatisfaction, PerformanceRating, RelationshipSatisfaction, and WorkLifeBalance. This helps us to better understand the levels assigned to the categorical variables.

Attrition Breakdown

This figure represents the breakup of our target variable (Attrition) for the total 1,470 observations. Based on the representation, we can conclude that the company has a 16% rate of attrition. This rate of attrition is considered high and very important for the company to resolve.

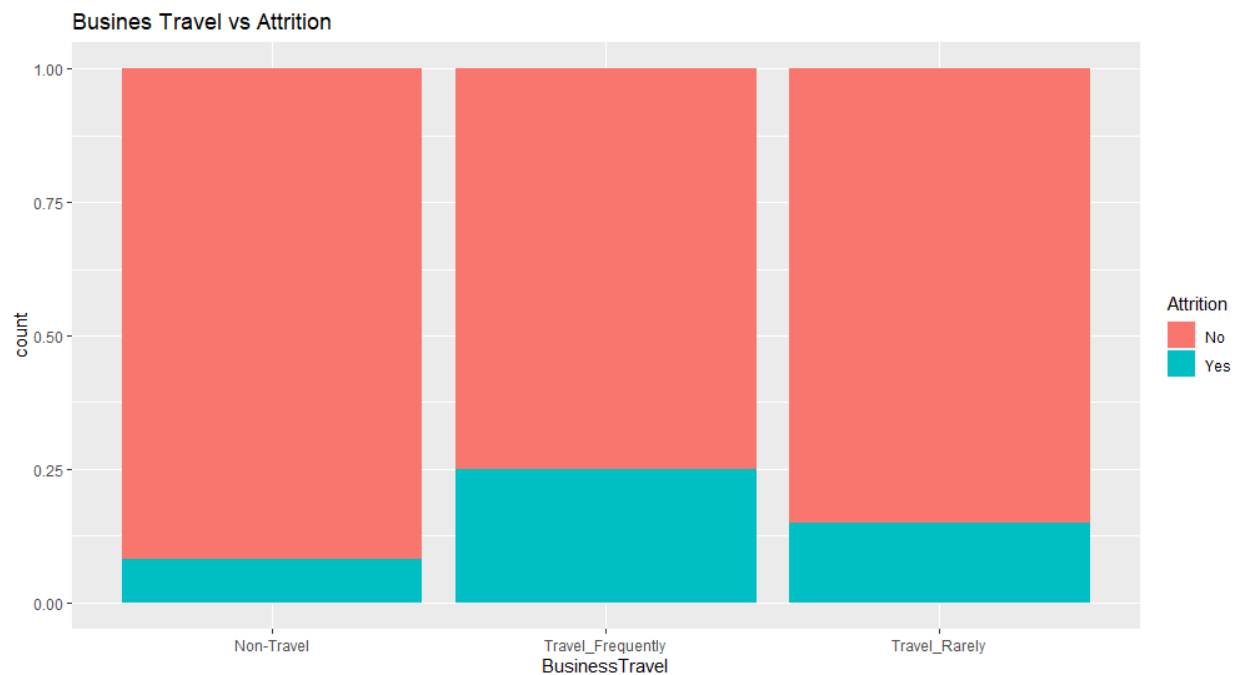
- 1,233 observations (Attrition = No)
- 237 observations (Attrition = Yes)



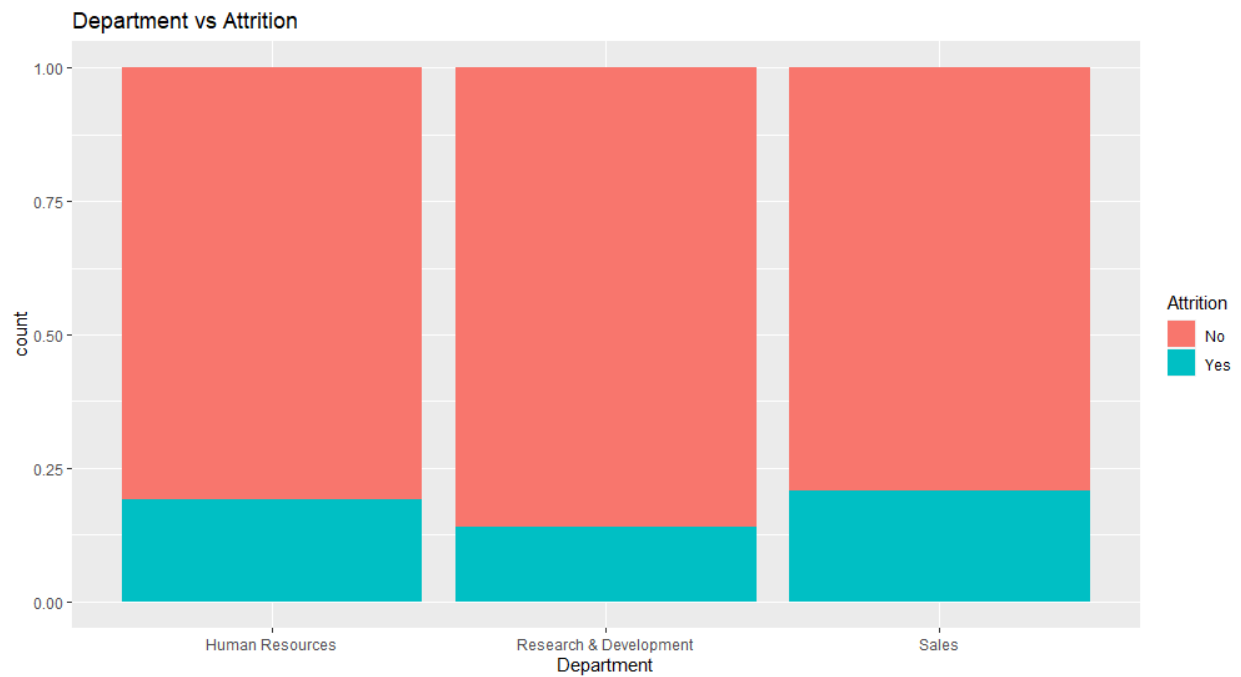
Data Visualization and Exploratory Analysis

In this section of the report, we will use our target variable (attrition) against the most obvious predictors to analyse the dataset and interpret its results.

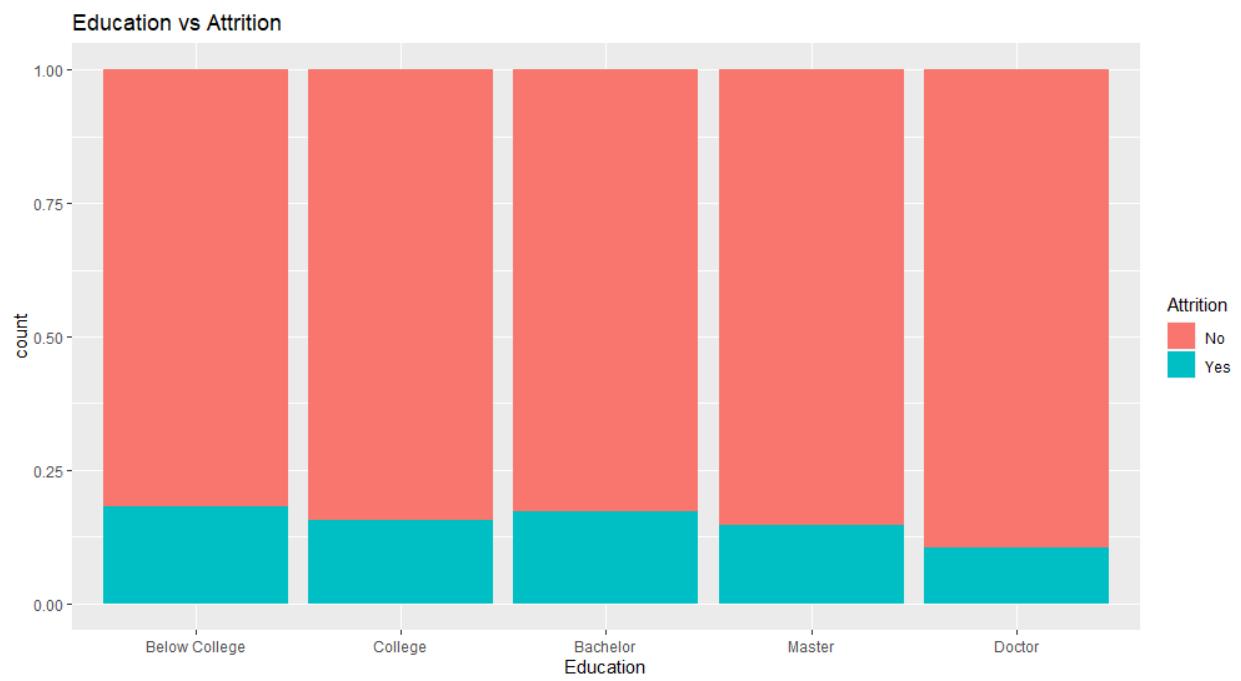
Analyzing the categorical variables vs Attrition:



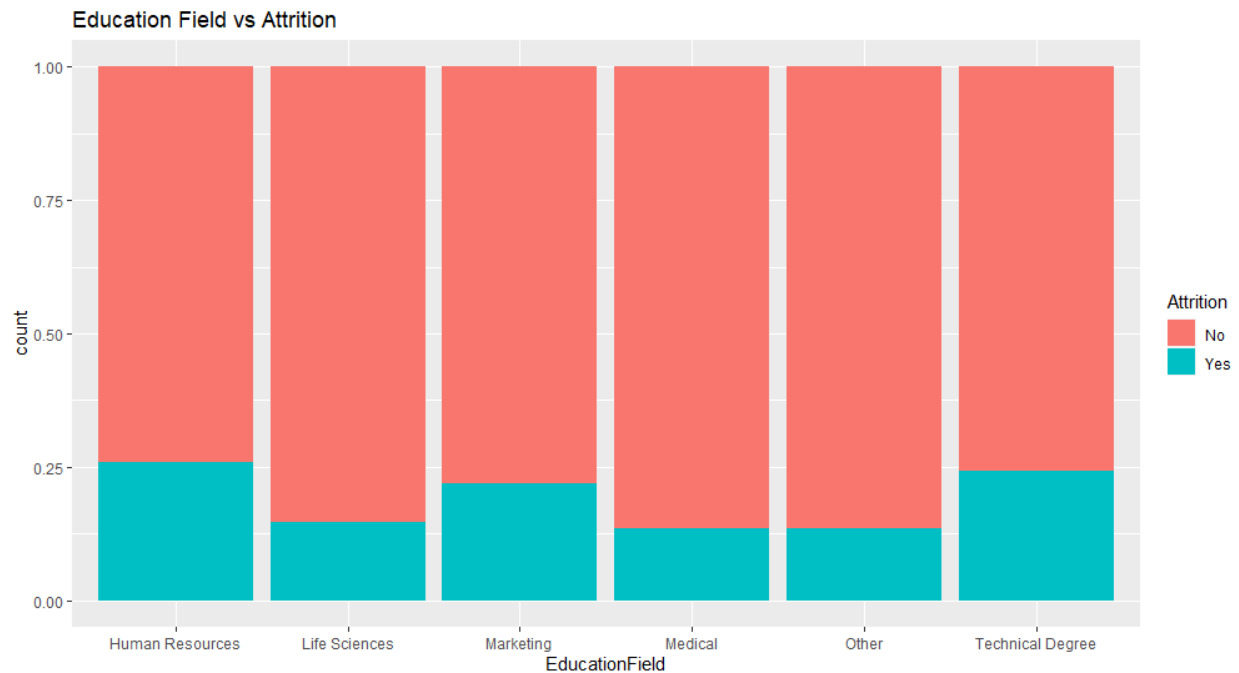
Employees who travel frequently and rarely have a higher attrition rate than employees who do not travel.



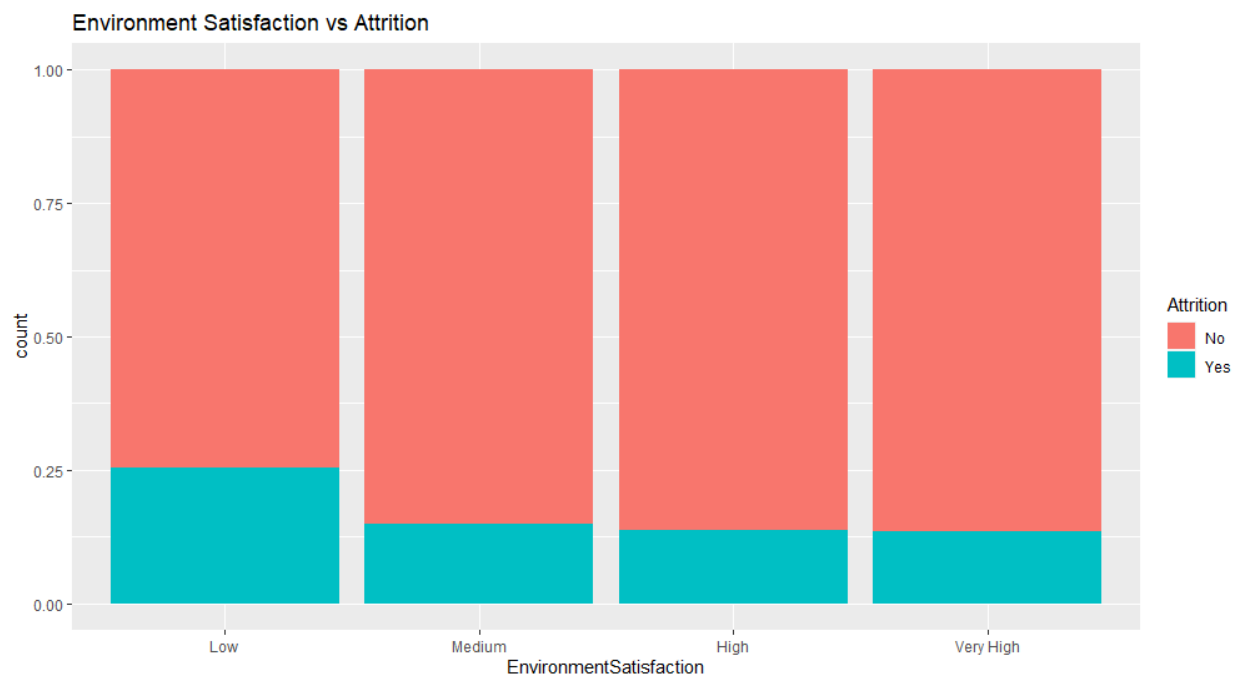
There seems to be a higher attrition rate in the Sales and Human Resource department than Research and Development.



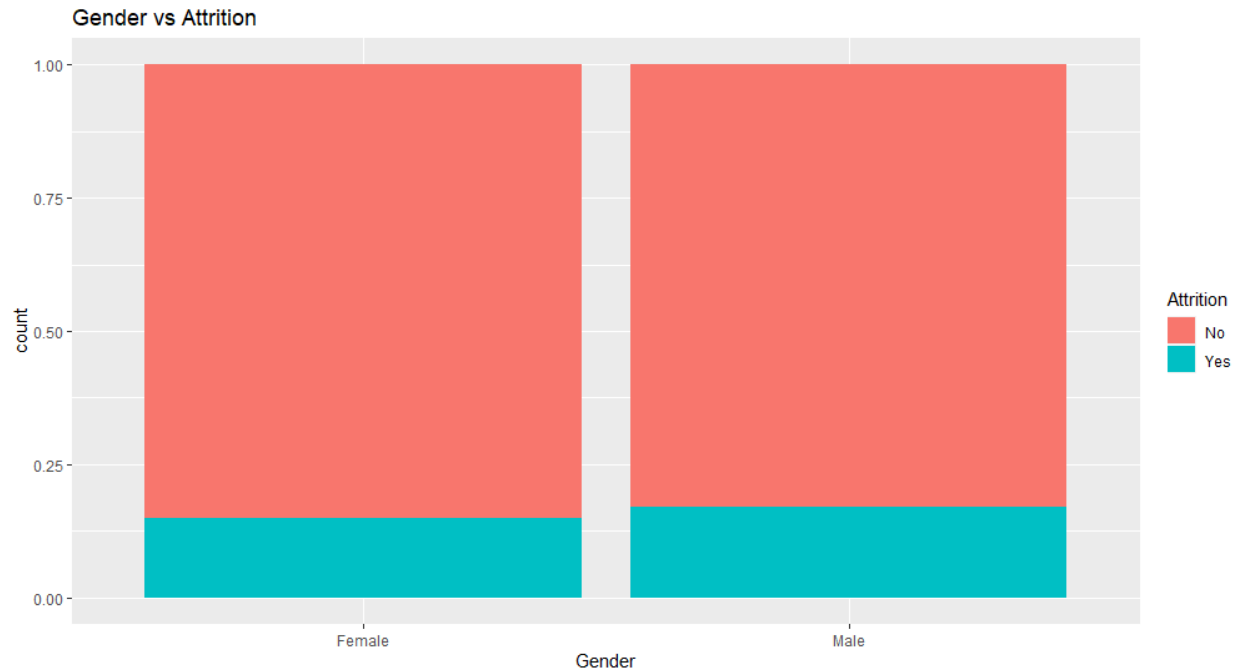
We cannot predict accurately from the above graph whether education level plays a role in Attrition. The only thing evident from the above is that the Doctor degree has the lowest Attrition and below college has the highest Attrition.



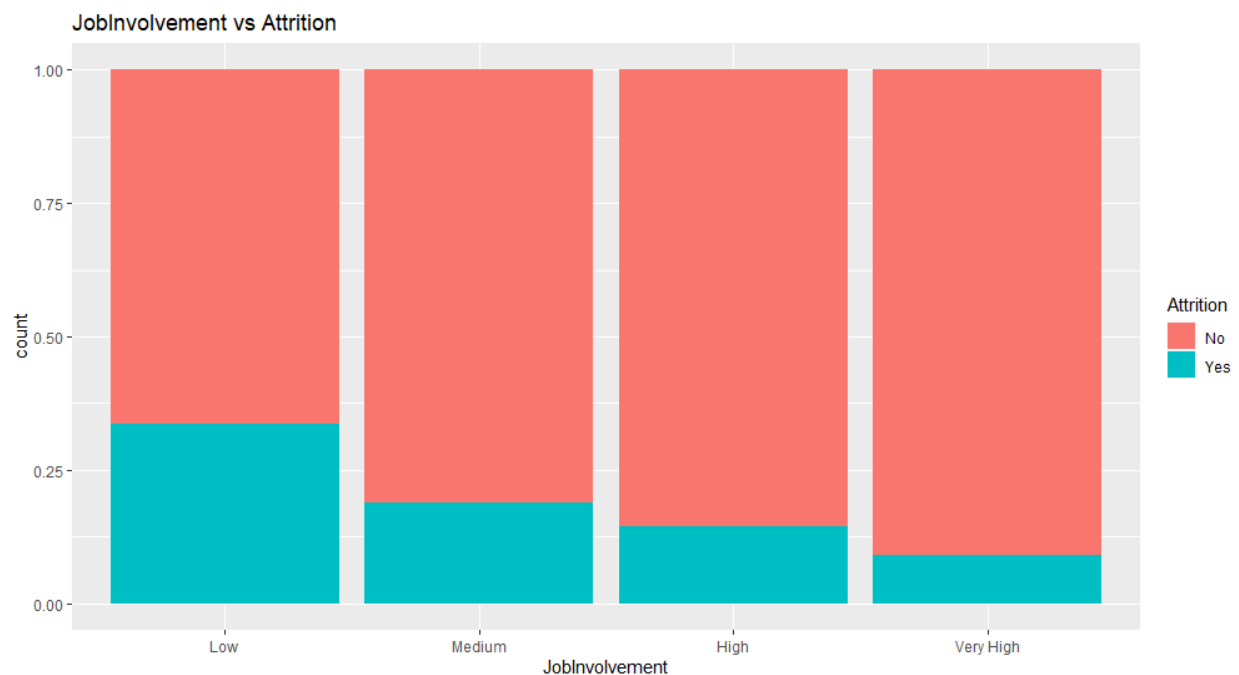
Technical Degree, Human Resource and Marketing has higher Attrition



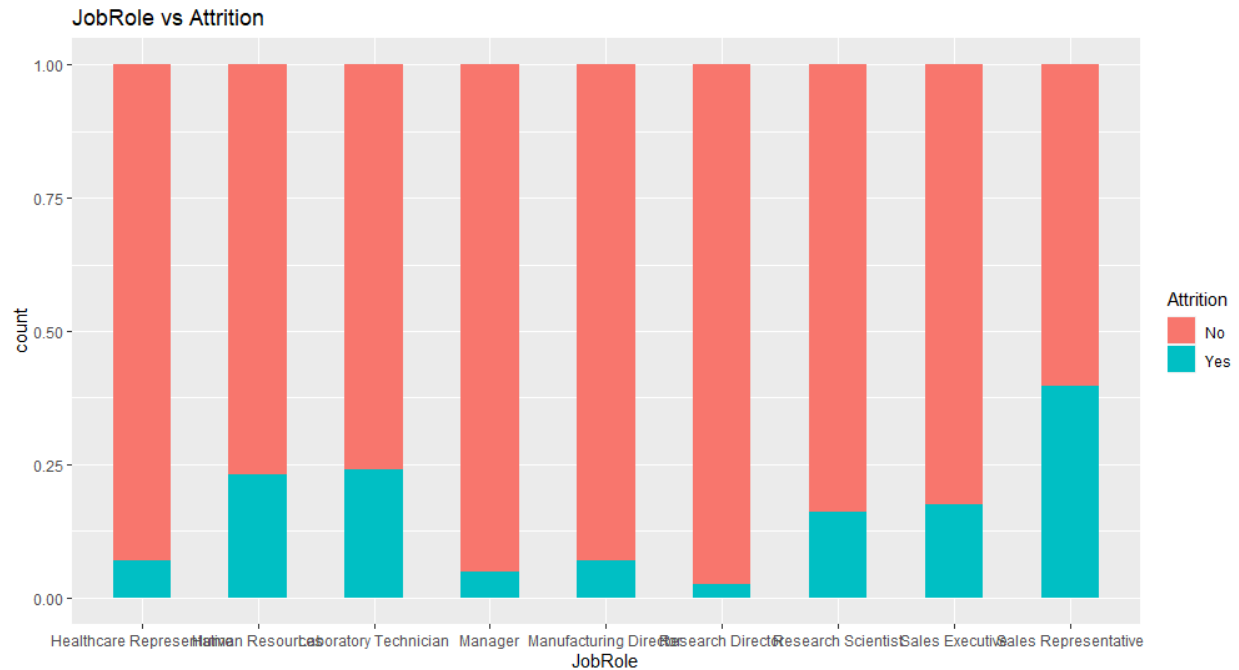
The lower an employee is satisfied with the work environment, the higher is the Attrition.



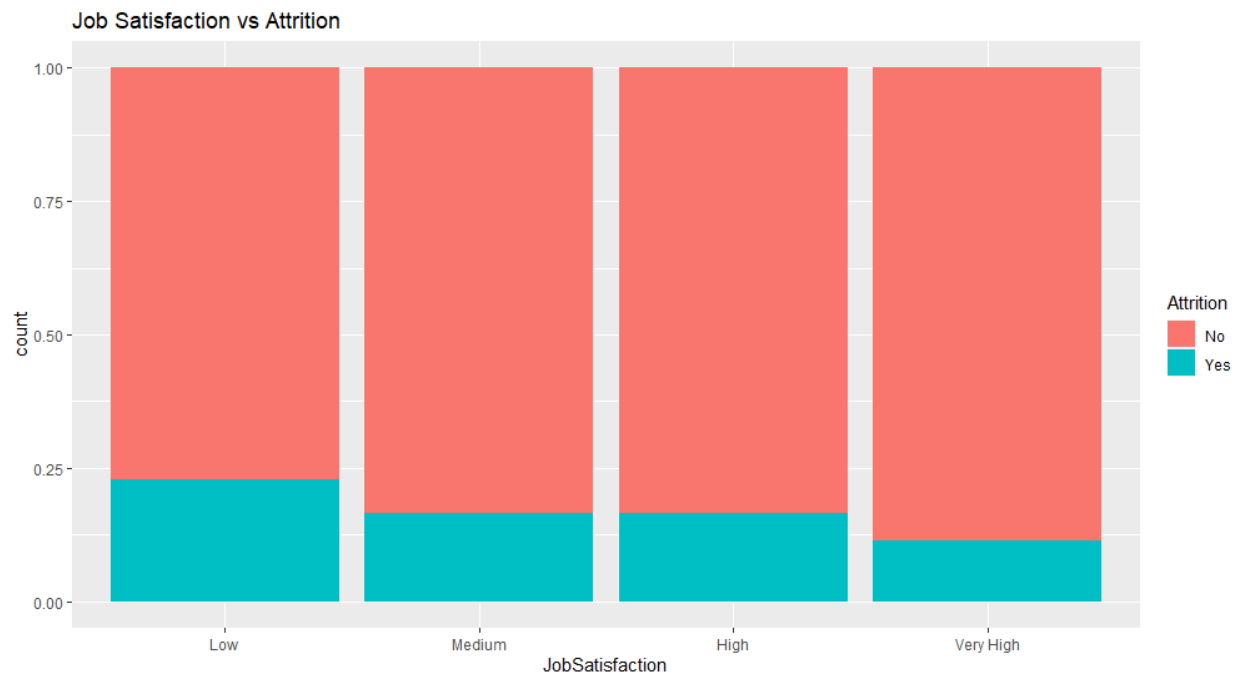
There is negligible difference in Attrition rate between the two Genders. We will not consider Gender in our Analysis



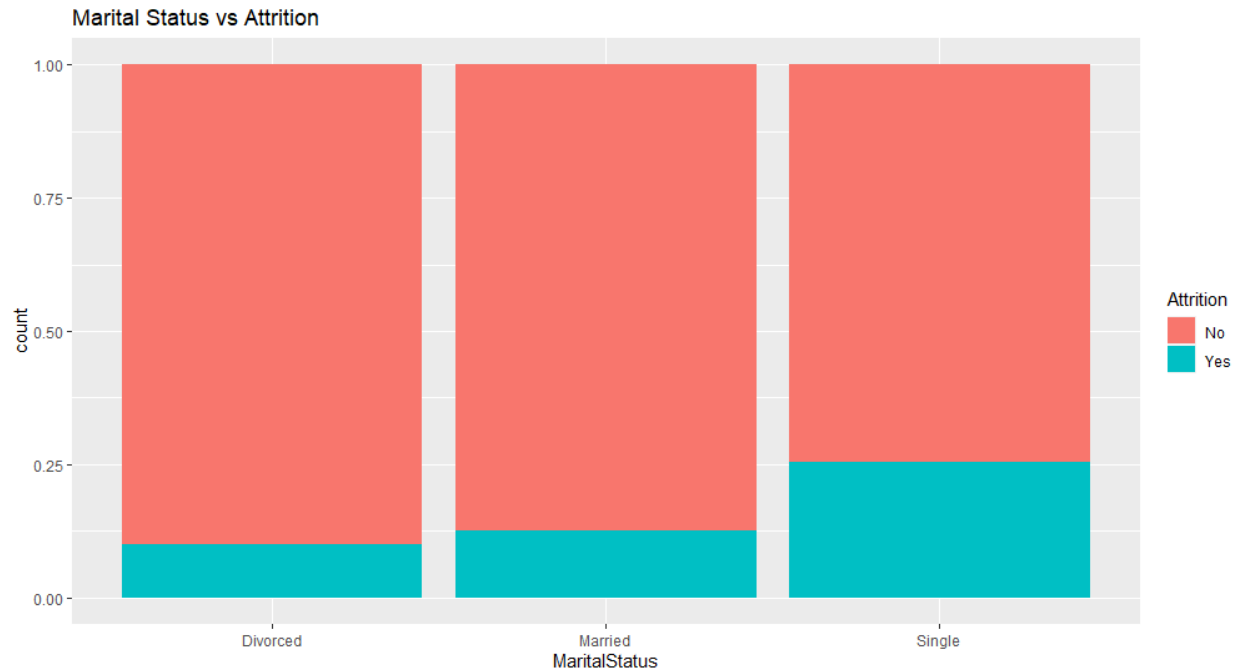
The less the employee is involved in his job, the more is the Attrition. Less involvement would probably mean the employee is not interested in the work and is probably looking for better prospects elsewhere.



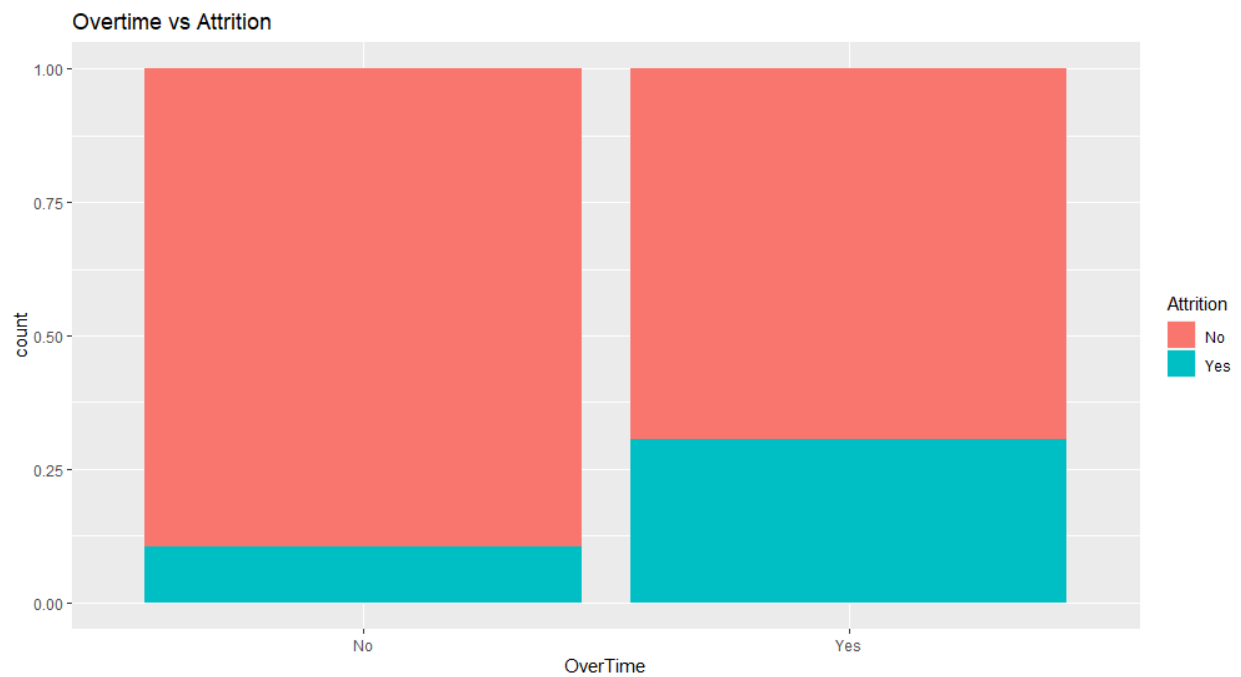
From the above graph we can conclude Human Resource, Laboratory Technician, Research Scientist, Sales Executive, Sales Representative have higher Attrition



It is clear from the above figure that lower job satisfaction has the highest Attrition.



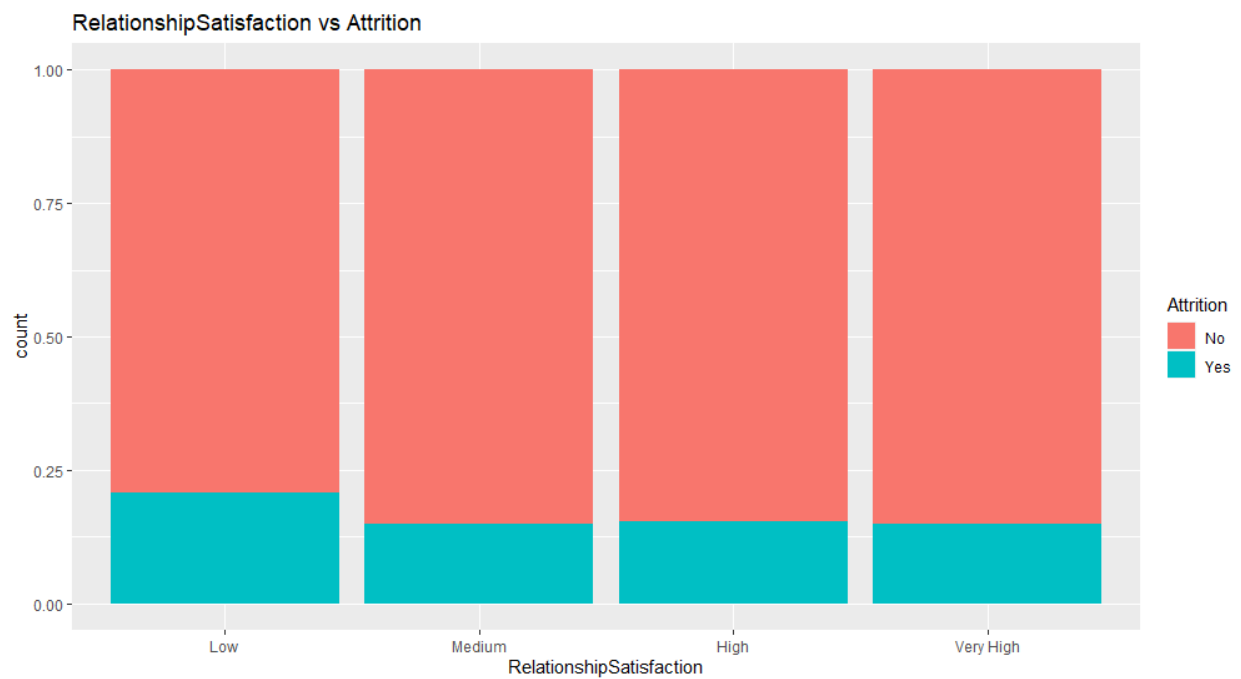
Attrition is higher among individuals who are single.



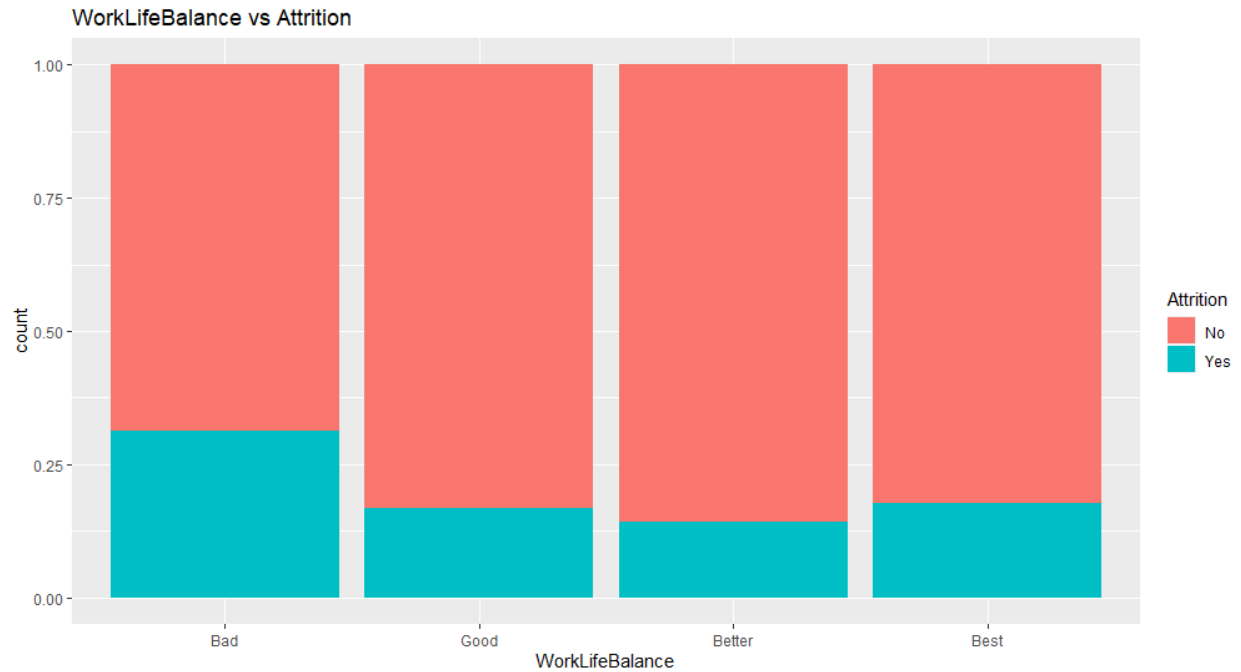
If employees are working overtime there are higher chances of Attrition



The above graph does not seem to influence Attrition. We will not consider Performance Rating in our further analysis.



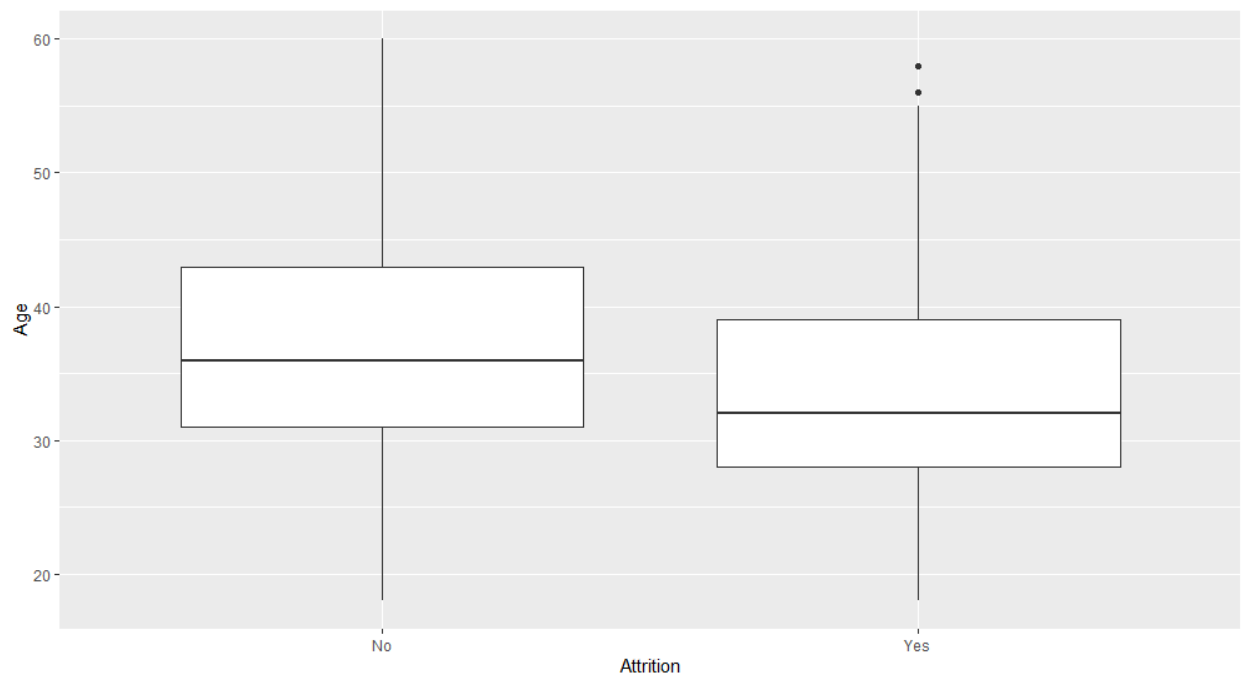
The lower the relationship Satisfaction, higher is the Attrition.



If the work life balance is bad, there is more attrition

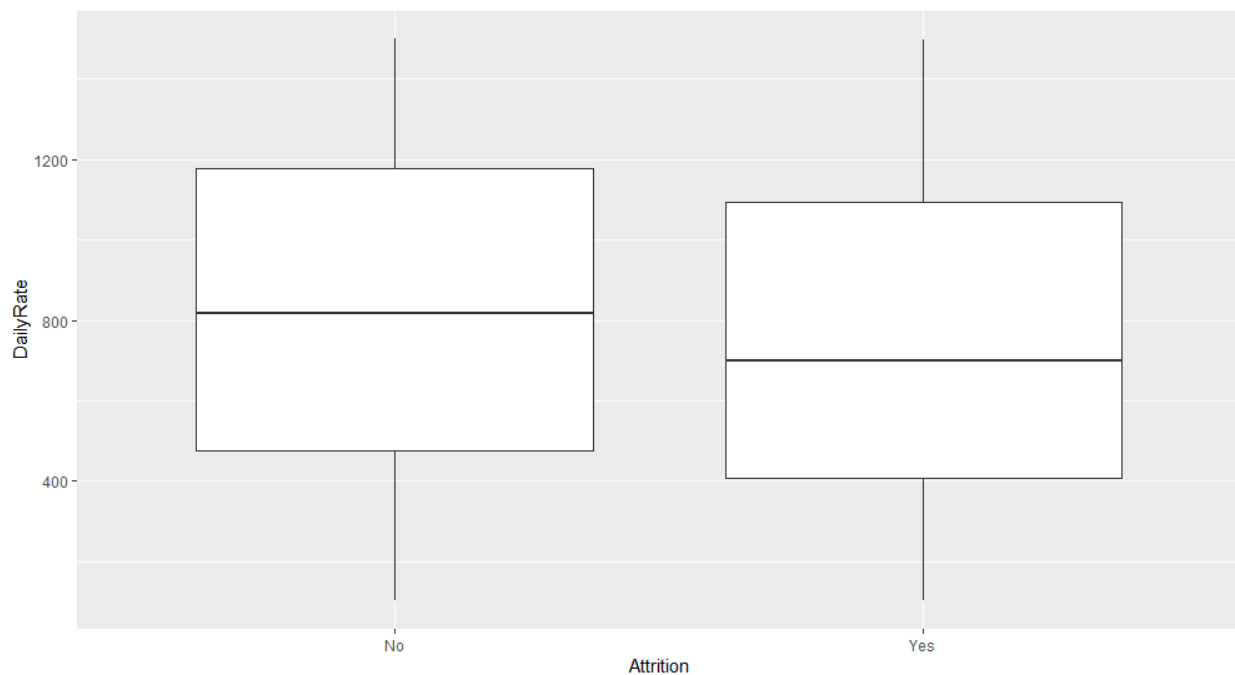
Analyzing continuous variables with respect to Attrition

Attrition vs Age



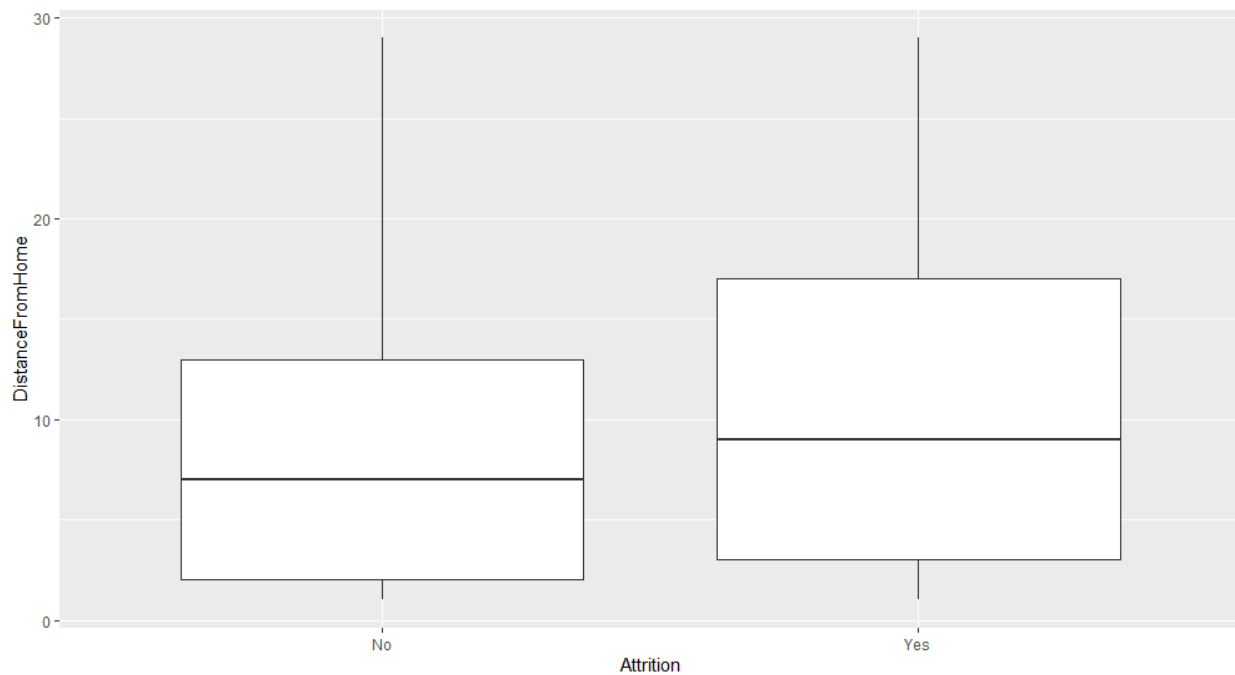
The median of attrition= “No” is higher than the median of Attrition=“Yes” which means 50% of the people below the age of 32 and above the age of 32 are prone to attrition and 50% of the people below the age of 36 and above the age of 36 is not prone to attrition. Also, the highest age for attrition(55 years) is lower than the highest age for not attrition. Hence employees above 55 years can be predicted to stay with the company.

Attrition vs Daily Rate



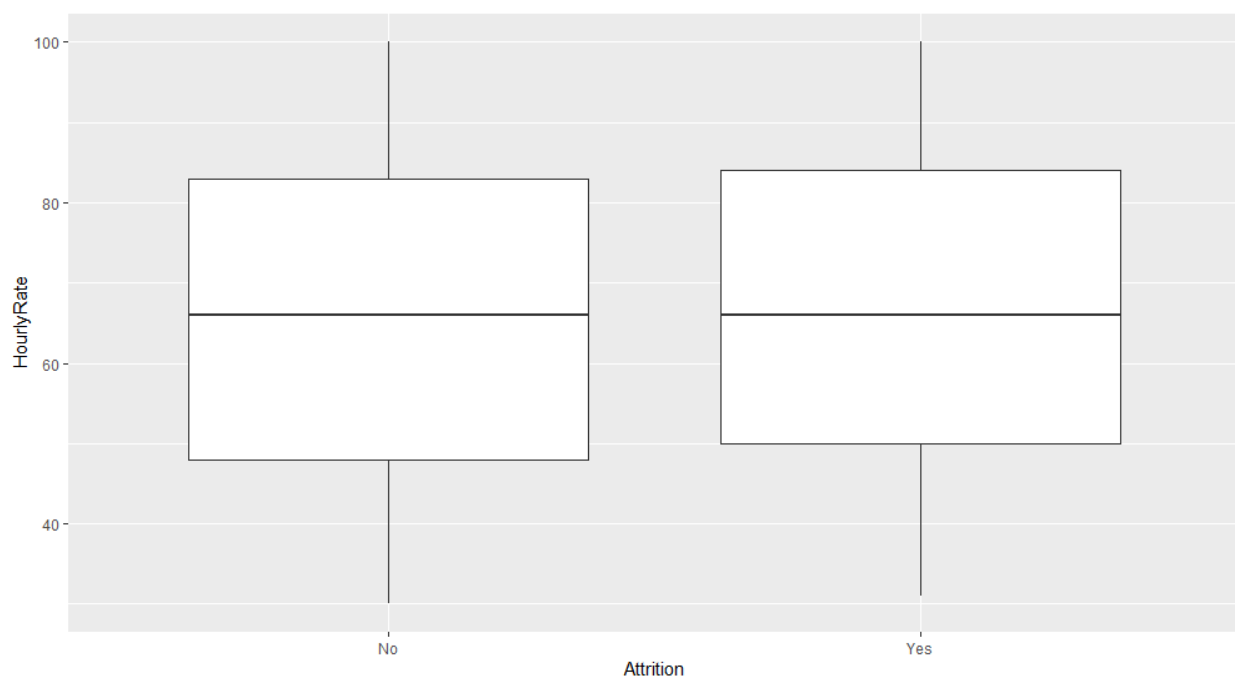
Both the box plots are almost similar. It is difficult to predict whether Daily rate influences Attrition.

Attrition vs Distance From Home



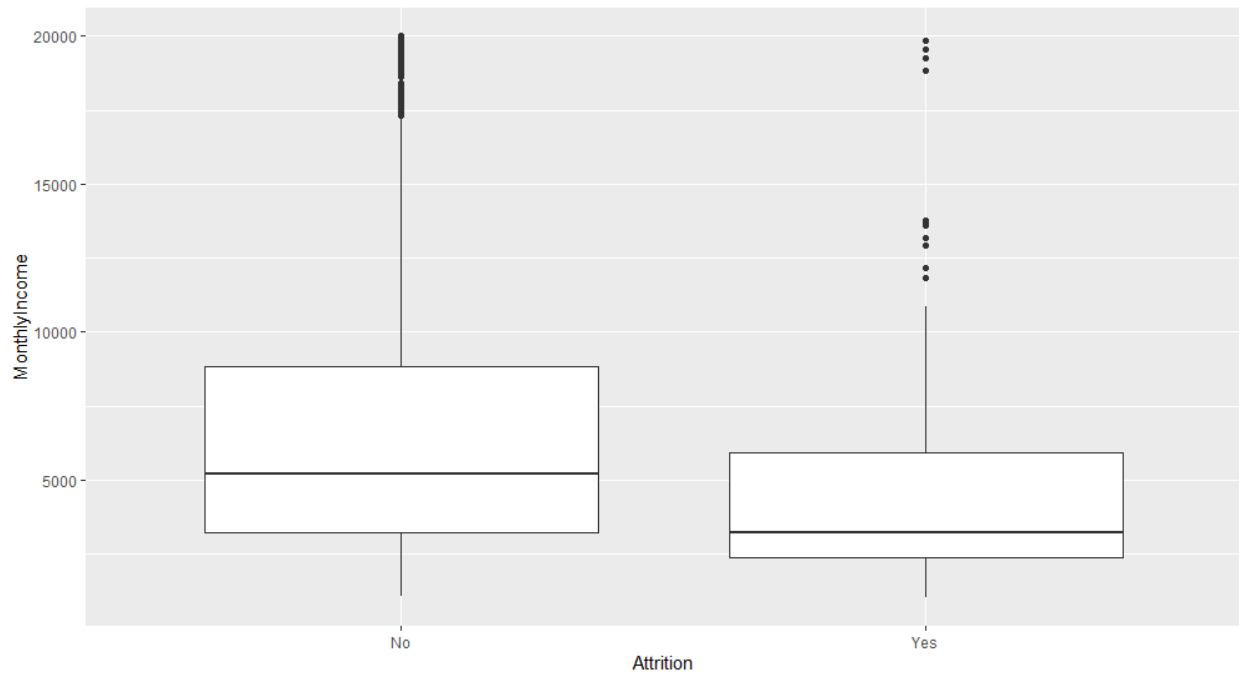
Distance from home for Attrition=Yes is more spread out across the median than distance from home for Attrition=No. There are more people around the median for Attrition=Yes. The higher the distance from home, the higher is the Attrition.

Attrition vs Hourly Rate



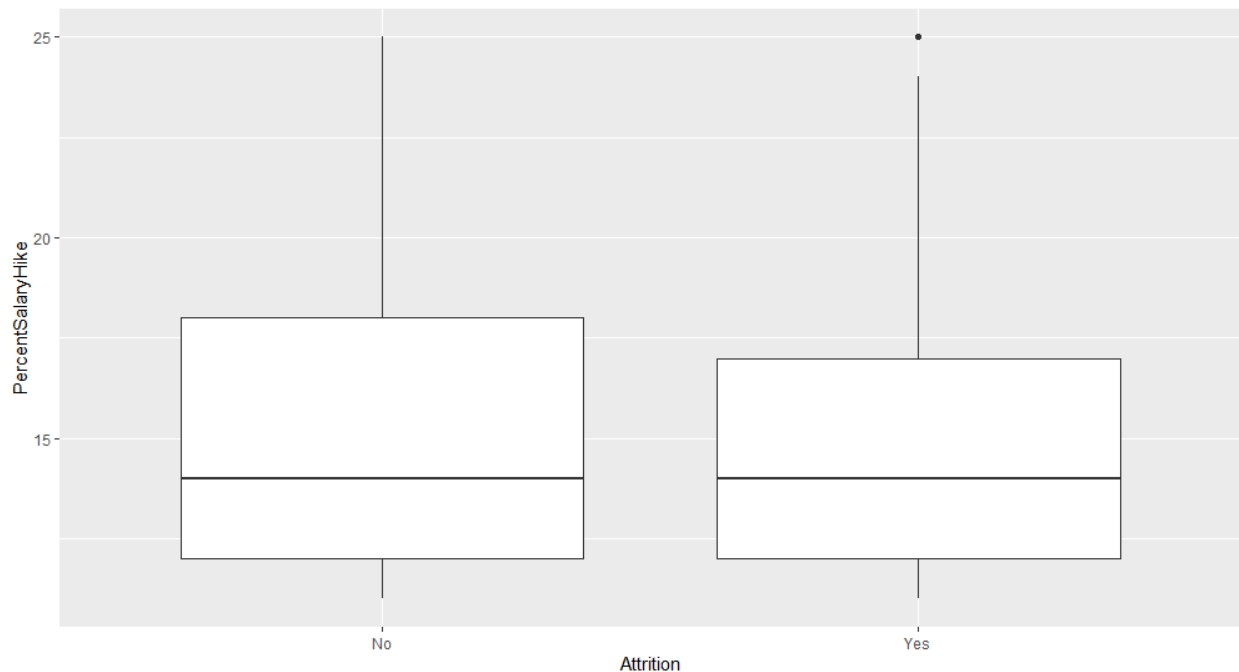
The two box plots are almost similar, we will not consider hourly rate in our analysis.

Attrition vs Monthly Income



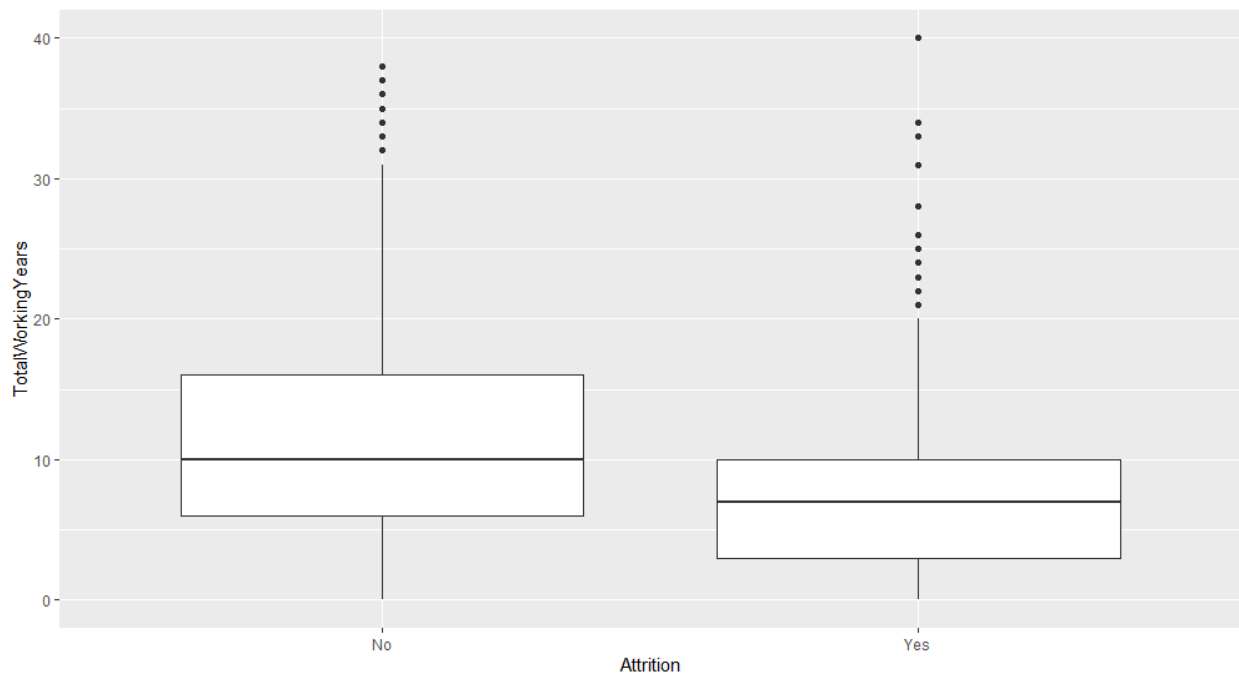
The upper 50% of the monthly income for Attrition=Yes is much lower than the upper 50% of the monthly income for Attrition=No. Also, the maximum salary for Attrition= Yes is much lower than the maximum salary for Attrition=No. Higher the monthly income, lesser is the Attrition

Attrition vs Percent Salary Hike



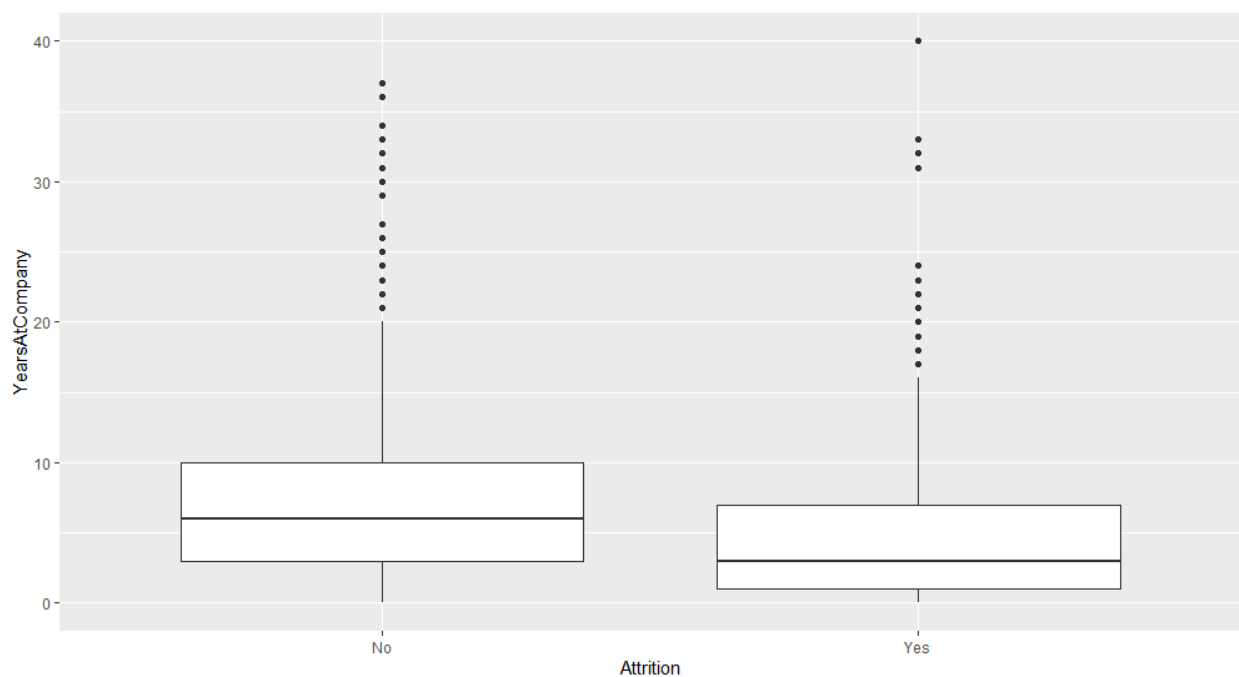
There are more employees with higher percentage salary hikes for Attrition=No

Attrition vs Total Working Years



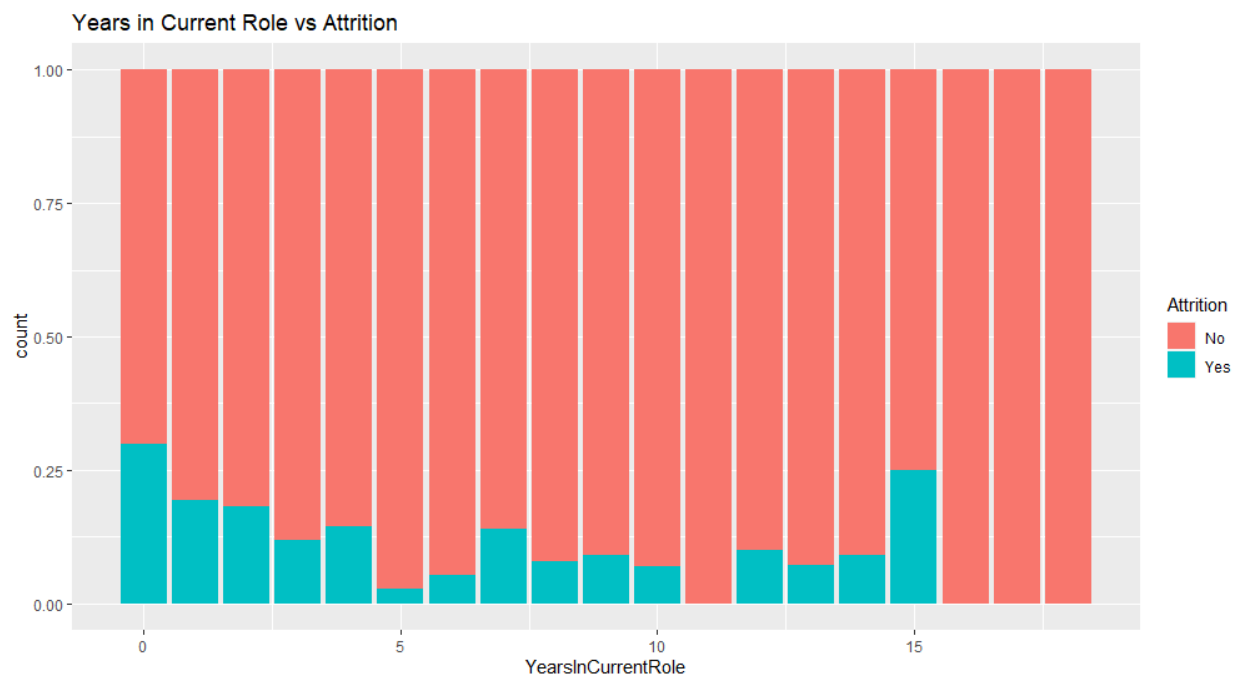
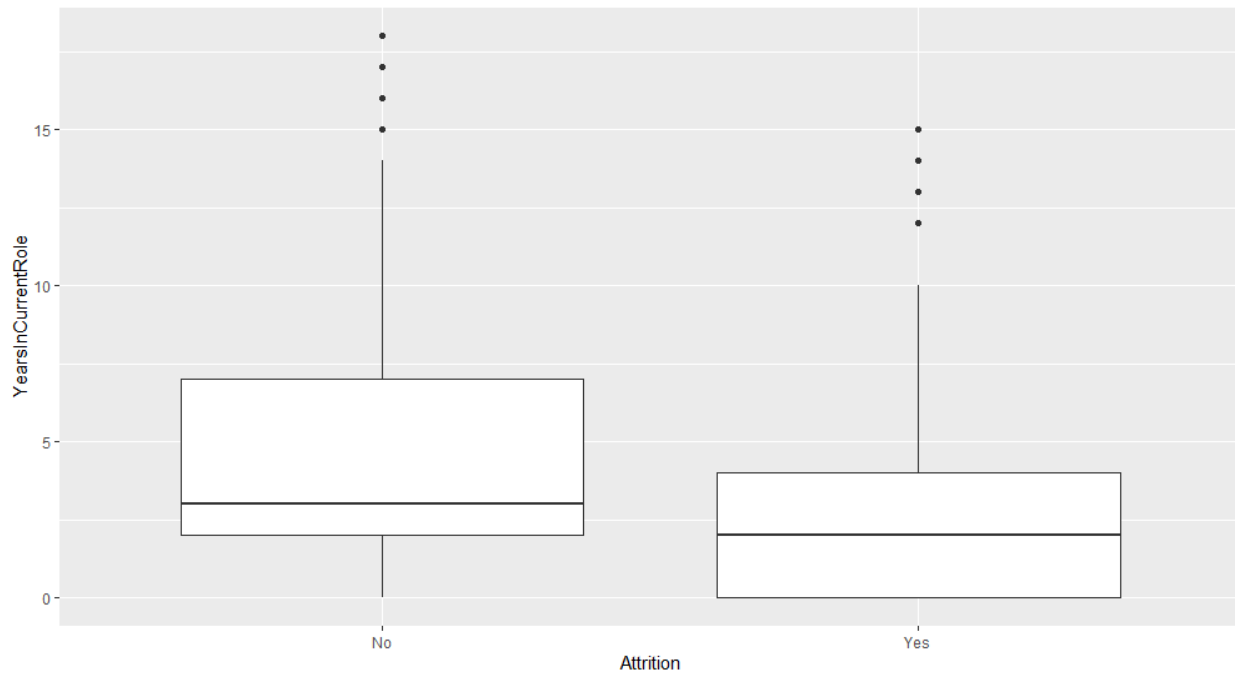
The employees with risk of Attrition are the ones with a lesser number of total working years.

Attrition vs Years At Company



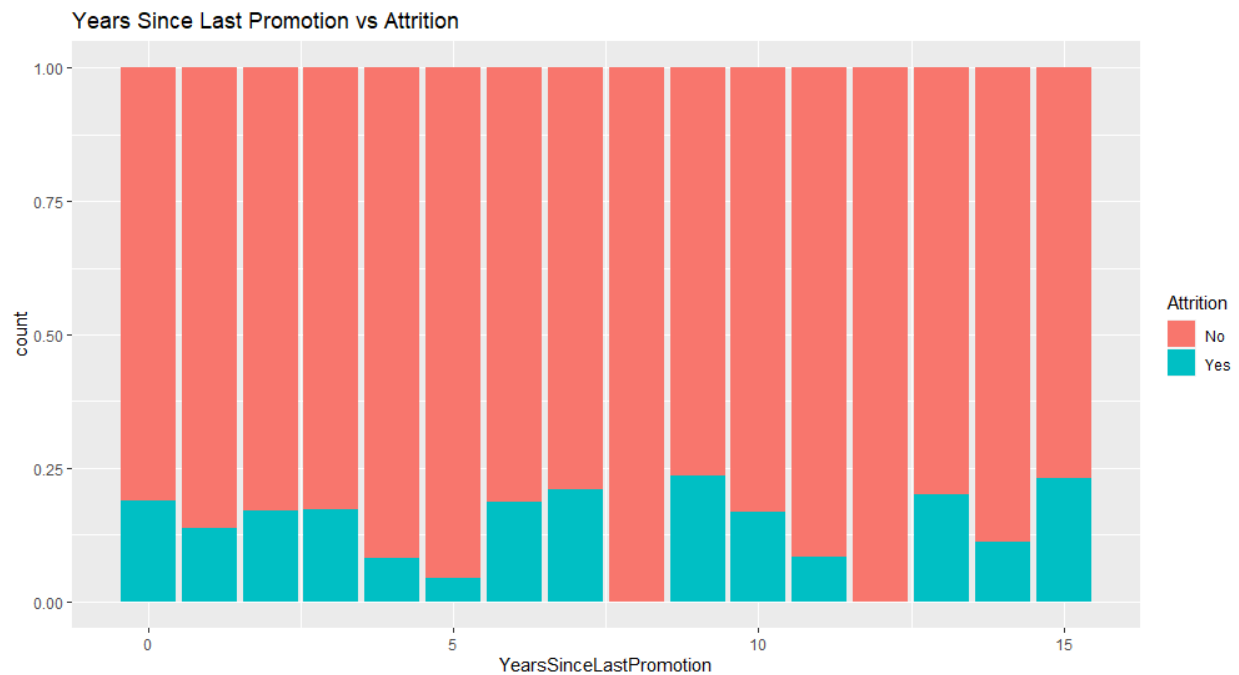
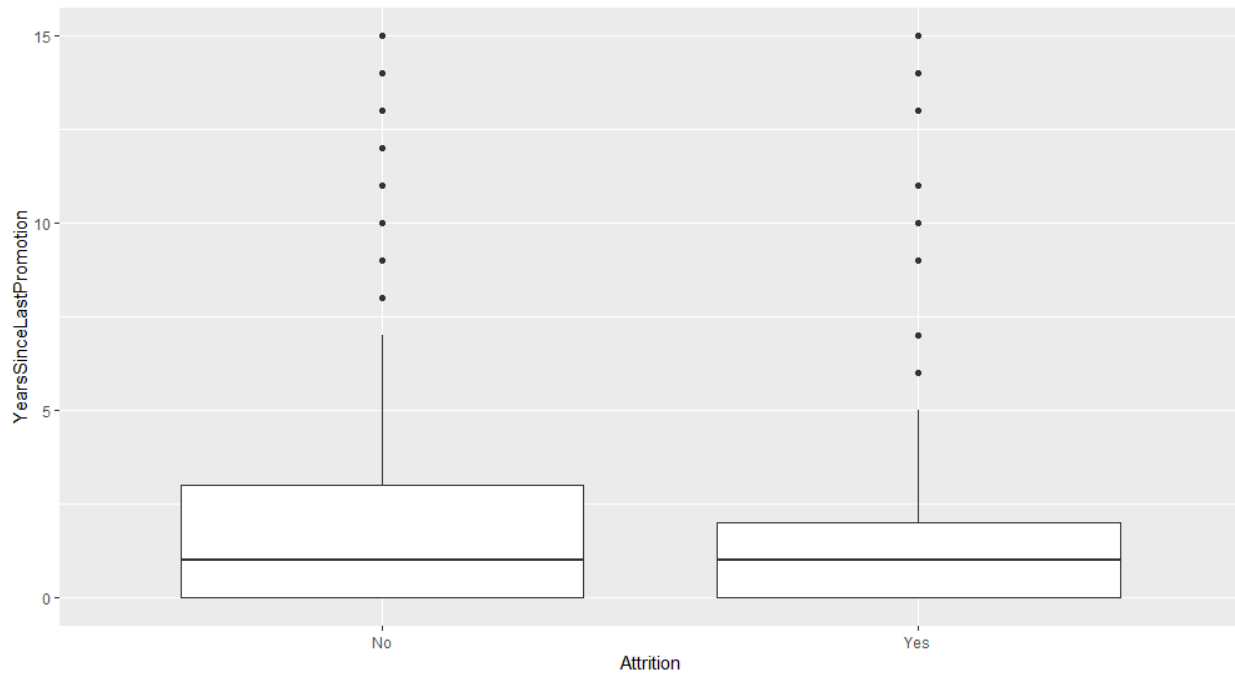
There is more Attrition among employees who have spent a lesser number of years at the company.

Attrition vs Years in Current Role



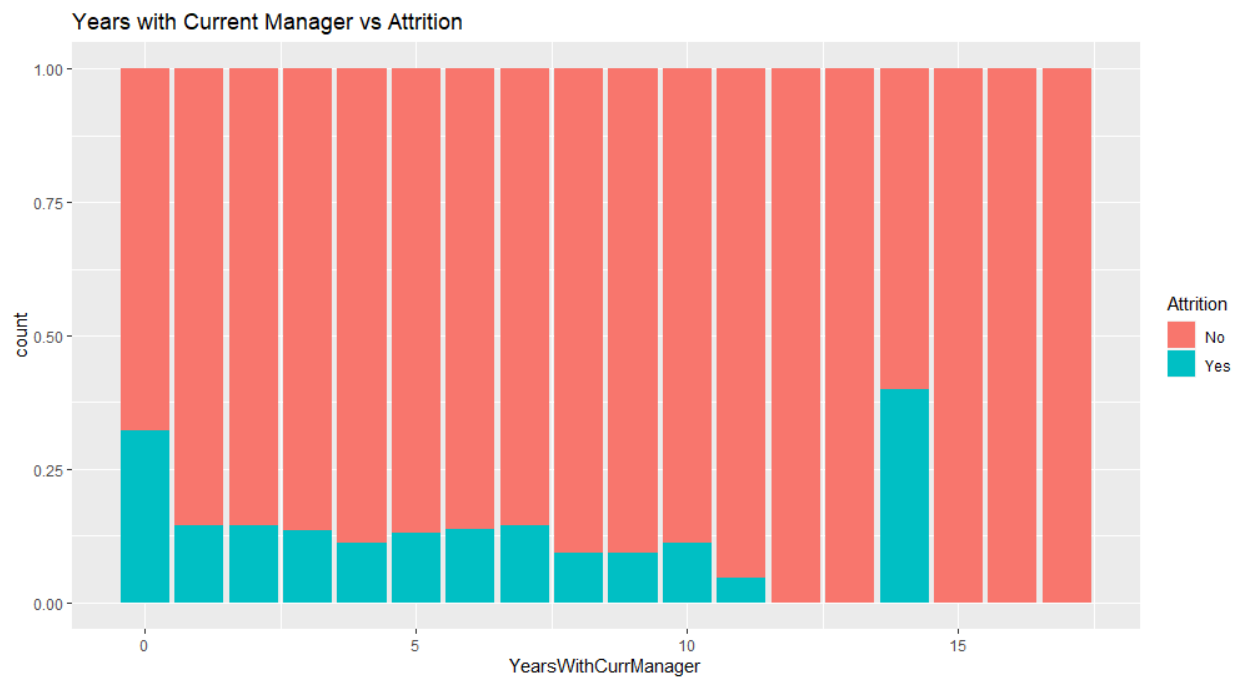
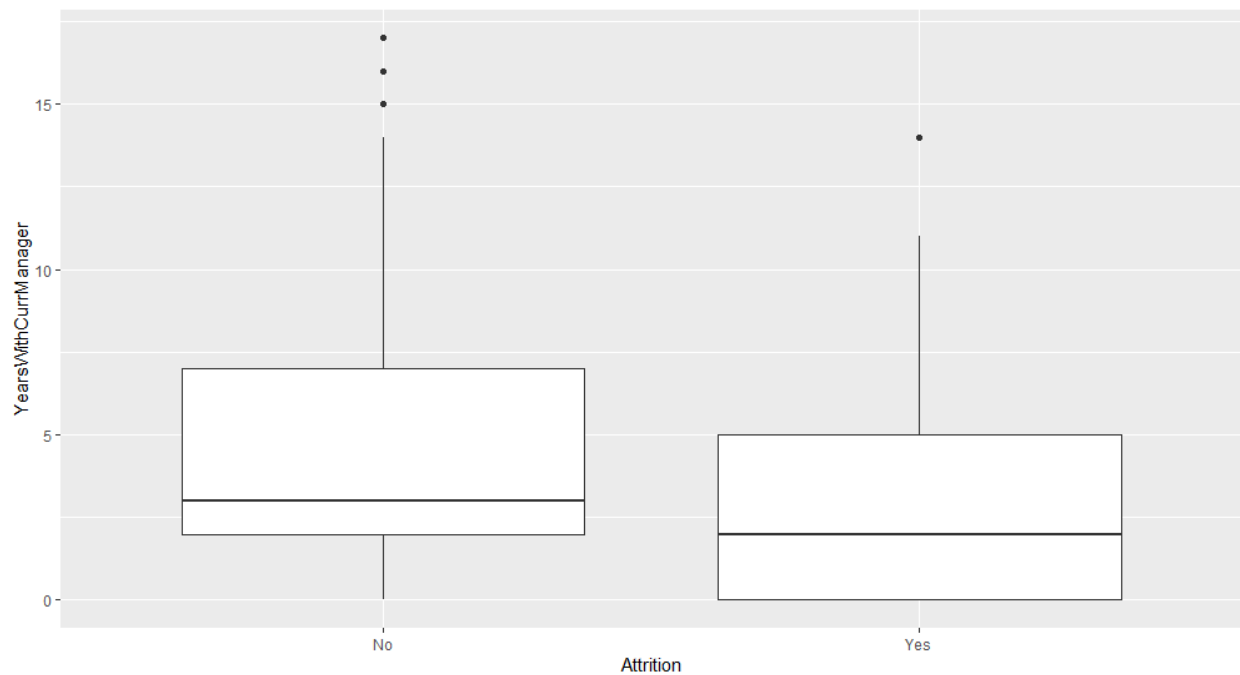
Attrition is higher for employees who have spent a lesser number of years in the current role. As per the box plot, probably the data for years=15 is an outlier and hence it deviates from the trend of higher the number of years, lower is the Attrition.

Years Since Last Promotion vs Attrition

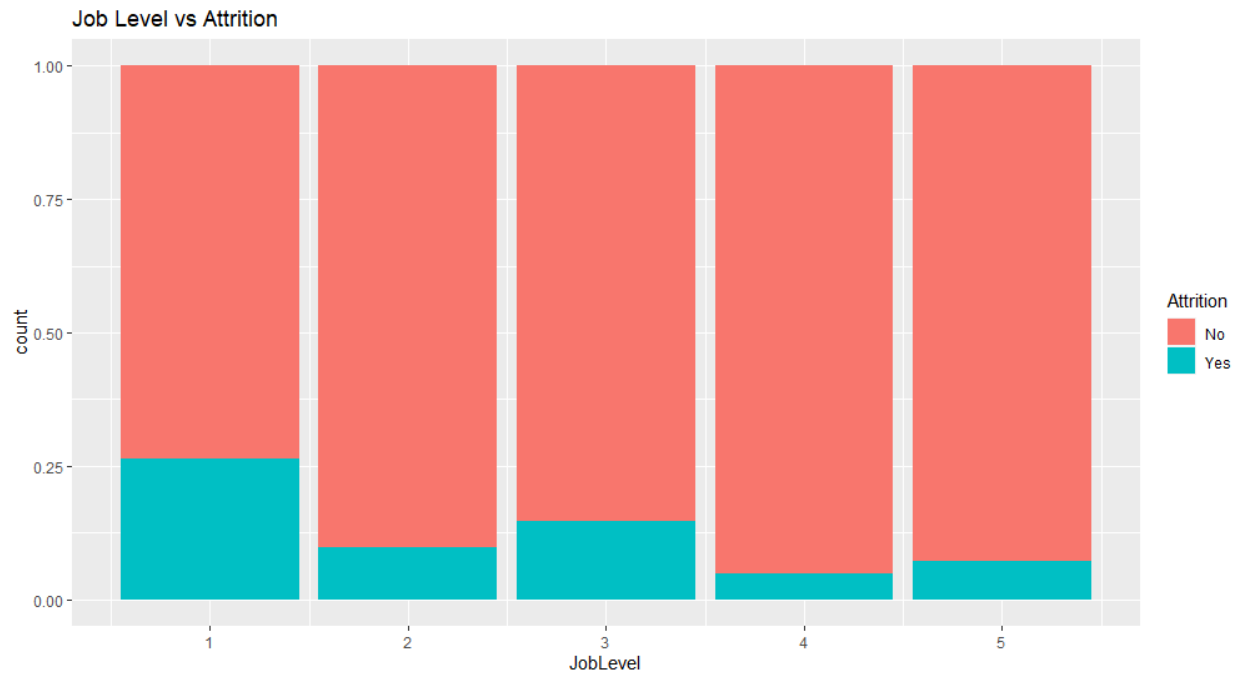


We cannot predict whether years since the last promotion affects Attrition since the graph represents varying results. We will not consider Years Since Last promotion in our analysis.

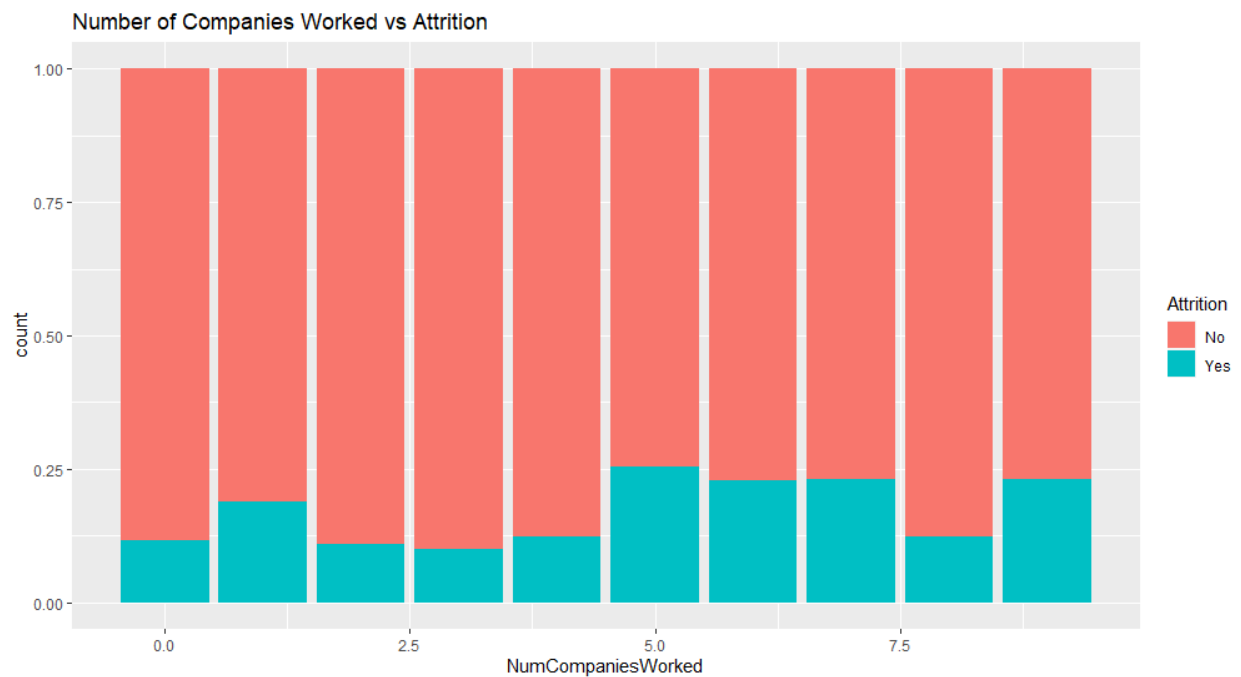
Years with current Manager vs Attrition



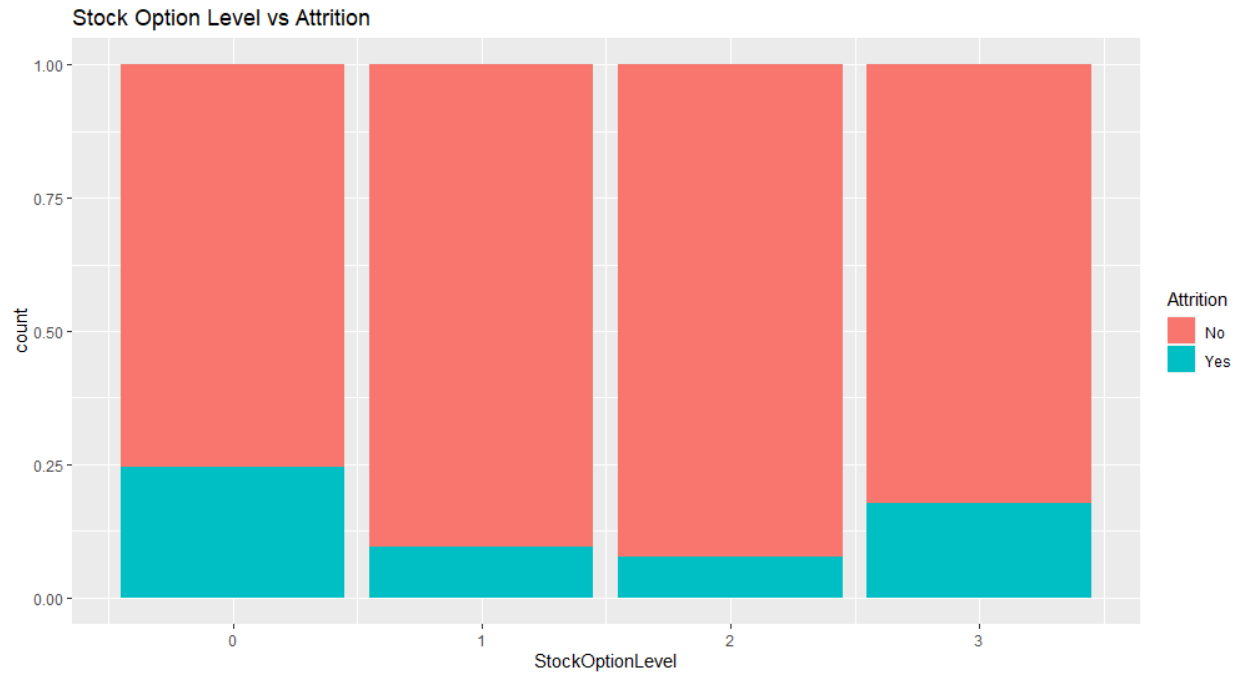
Employees with the risk of Attrition have spent only 0-5 years with their current Manager.



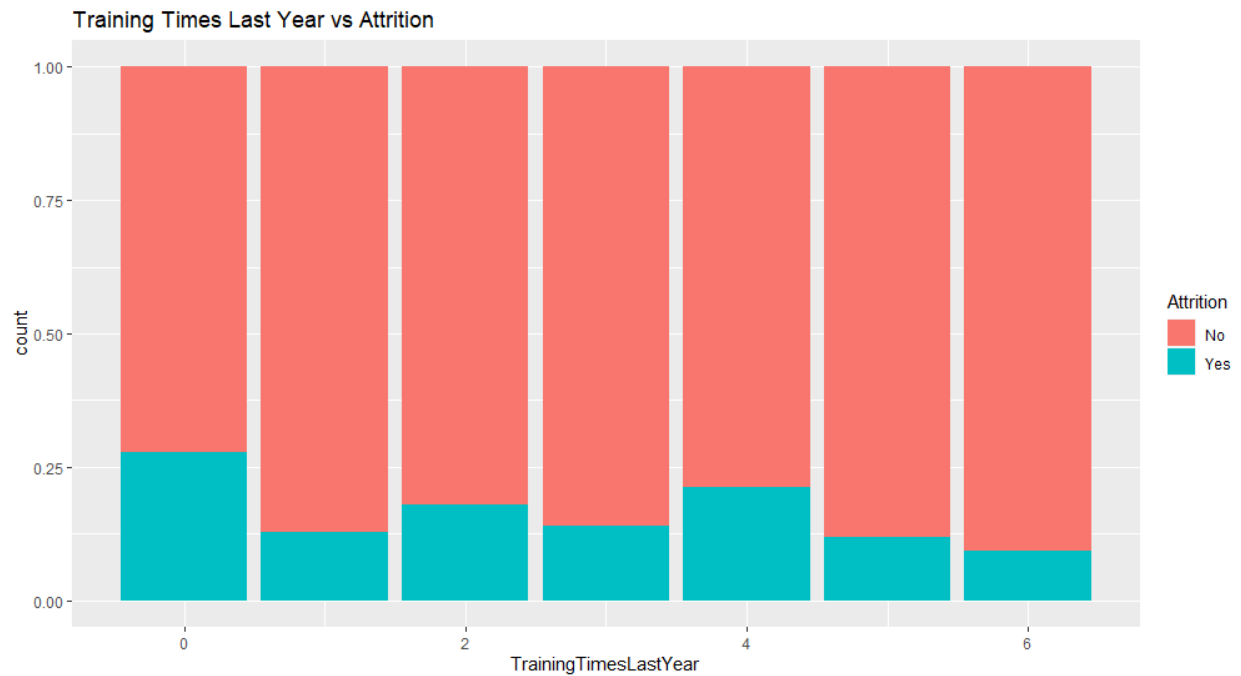
There is less Attrition for Senior level roles(4 and 5)



There is more Attrition among employees who worked for 5 or more companies.



Stock Option level of 0 and 3 have the highest Attrition



If the employee training time last year is zero then there is the highest chance of Attrition.

After visualizing the above graphs we have observed that the following variables are not important in predicting Attrition and hence discarded them for our further analysis:

1. Gender
2. Performance Rating
3. Hourly Rate
4. Years Since Last Promotion

DATA MODELING & METHODOLOGY

Methodology

Our main goal in the project is to predict whether there will be employee Attrition or not. Our target variables in this case are categorical hence it translates into a problem of classification. Our approach towards building the model was to select independent variables using data visualization that has some effect on the dependent variable(Attrition in this case). The models that we have used are Classification Tree and Random Forest

Selection of Important Independent Variables:

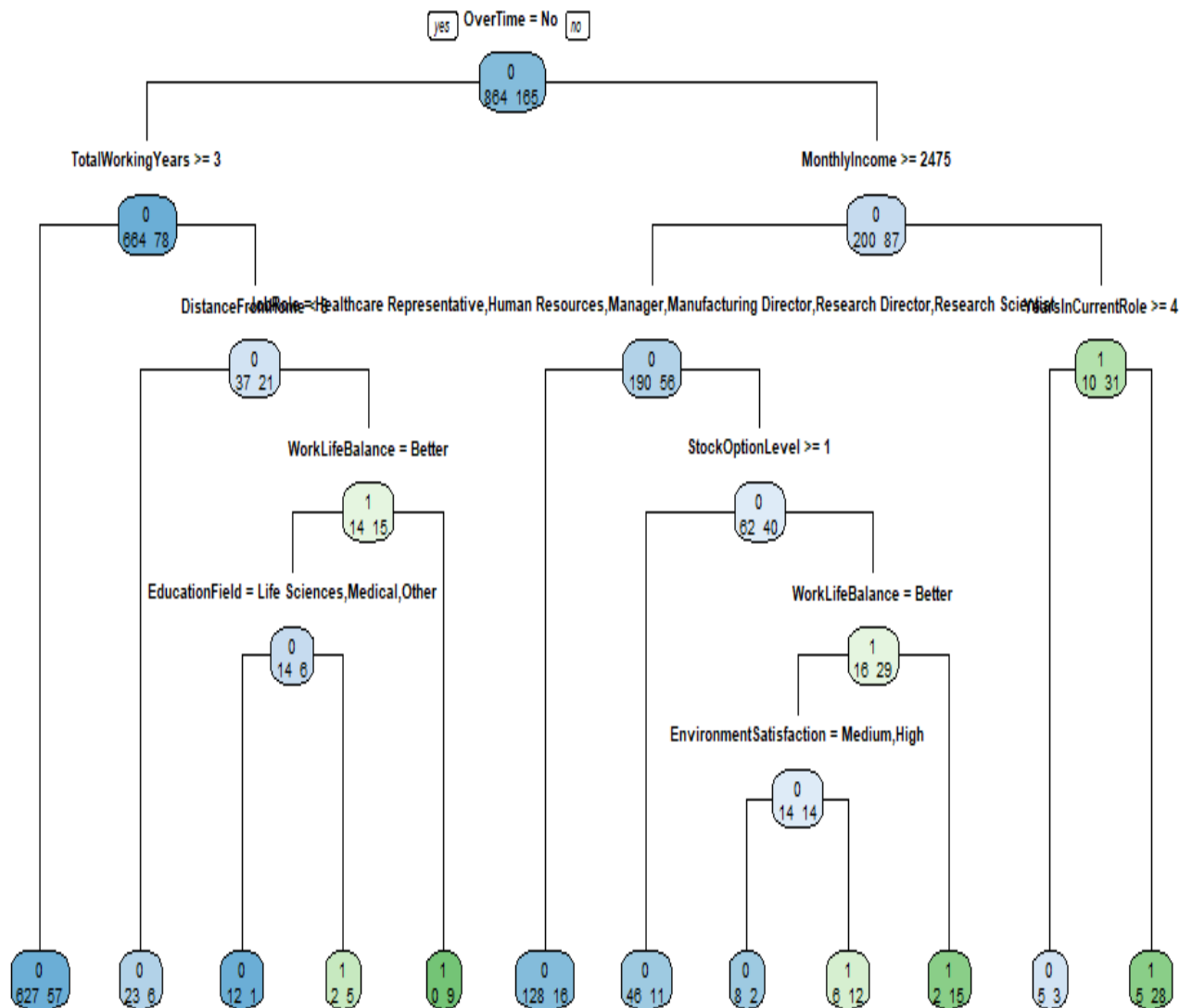
Based on our data visualization and exploratory analysis we have considered the following independent variables as important

- Age
- Business Travel
- Daily Rate
- Department
- Distance From Home
- Education
- Education Field
- Environment Satisfaction
- Job Involvement
- Job Role
- Job Level
- Job Satisfaction
- Marital Status
- MonthlyIncome.
- Number of Companies Worked
- OverTime
- Percent Salary Hike
- Relationship Satisfaction
- Stock Option Level
- Total working years

- Training Time Last Year
- Work Life Balance
- Years at Company
- Years in current role
- Years with current Manager

Data Modeling 1: Classification Tree

We selected the above, mentioned independent variables and performed a 70-30 split of the dataset. We obtained the following classification tree



Analysis:

We found that monthly income, overtime, job role, total working years, stock option level, and work life balance are the most important variables as per the model. The following values were assigned to the variables:

- **MonthlyIncome:** 23.37
- **OverTime:** 16.229
- **Job Role:** 14.03
- **TotalWorkingYears:** 11.54
- **Stock Option Level:** 10.25
- **Work Life Balance:** 9.17

Confusion Matrix:

From the confusion matrix, we found 85.49% accuracy. Our positive class in this case is Class 1. The sensitivity value we received is 0.36111 and the specificity value we received is 0.95122.

| | | Reference | |
|------------|--|-----------|----|
| Prediction | | 0 | 1 |
| 0 | | 351 | 46 |
| 1 | | 18 | 26 |

Accuracy : 0.8549

95% CI : (0.8185, 0.8864)

No Information Rate : 0.8367

P-Value [Acc > NIR] : 0.1671310

Kappa : 0.3703

Mcnemar's Test P-Value : 0.0007382

Sensitivity : 0.36111

Specificity : 0.95122

Pos Pred Value : 0.59091

Neg Pred Value : 0.88413

Prevalence : 0.16327

Detection Rate : 0.05896

Detection Prevalence : 0.09977

Balanced Accuracy : 0.65617

As evident from our classification tree, we do not have highest purity for most of the nodes. We can increase the purity of the nodes by growing the full length of the tree. But this will lead to the problem of overfitting. Overfitting might affect the accuracy of prediction in a new data set because of the introduction of errors/noise . We decided to solve the problem of overfitting using the Random Forest algorithm. Random Forest applies the technique of bootstrap aggregating and bagging to solve the overfitting issue of the classification tree.

Data Modeling 2: Random Forest

After training the model with random forest we observed that the model's accuracy of prediction was 87.3%. The sensitivity value we received is 0.29167 and the specificity value we received is 0.98645

| | Reference | |
|------------|-----------|----|
| Prediction | 0 | 1 |
| 0 | 364 | 51 |
| 1 | 5 | 21 |

Accuracy : 0.873

95% CI : (0.8383, 0.9026)

No Information Rate : 0.8367

P-Value [Acc > NIR] : 0.02039

Kappa : 0.3744

Mcnemar's Test P-Value : 1.817e-09

Sensitivity : 0.29167

Specificity : 0.98645

Pos Pred Value : 0.80769

Neg Pred Value : 0.87711

Prevalence : 0.16327

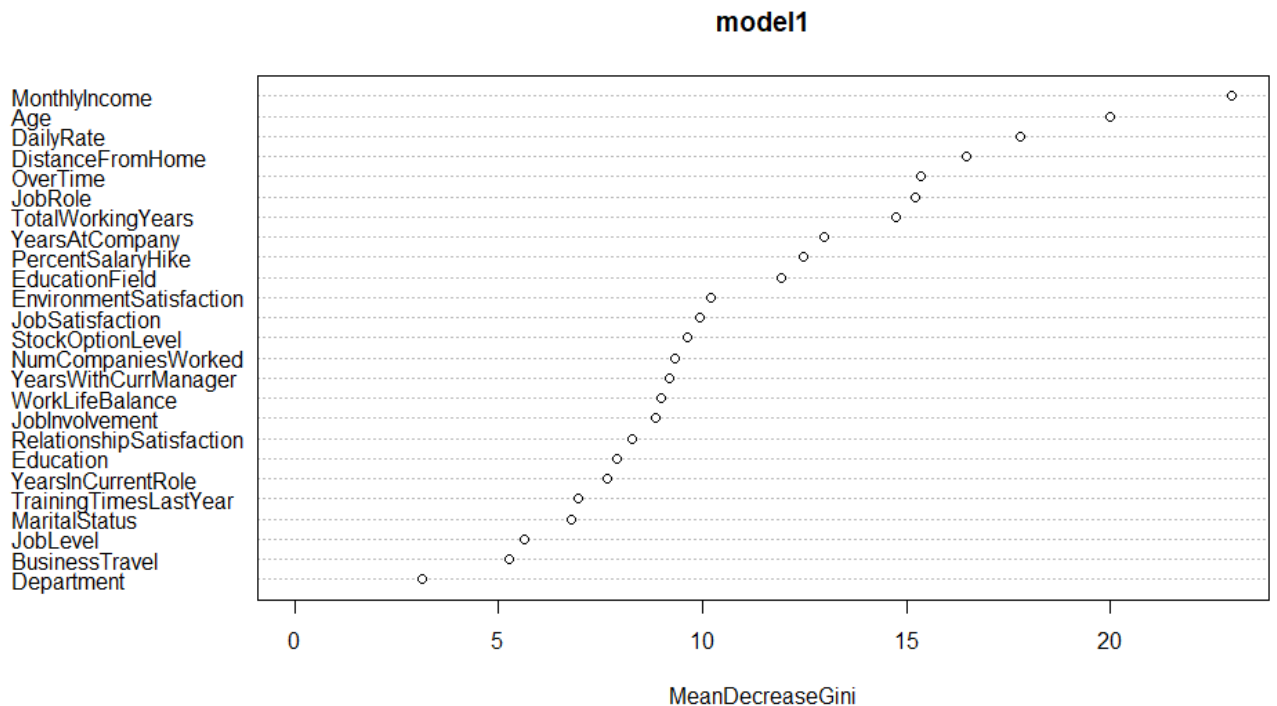
Detection Rate : 0.04762

Detection Prevalence : 0.05896

Balanced Accuracy : 0.63906

'Positive' Class : 1

The following graph was generated from the random forest model highlighting the important variables



Analysis:

As seen in the graph above, the five most important variables for predicting Attrition are monthly income, age, daily rate, distance from home, overtime.

Handling the Class Imbalance Problem

We observed that only 237 observations out of a total of 1470 observations in our data set belong to the important class which in our case is Class 1(Attrition=Yes). Our goal is to predict the possibility of Attrition and hence our model should be able to predict Class 1 more accurately than Class 0. We decided to handle this class imbalance problem by applying the technique of oversampling and undersampling.

Applying Over Sampling in Classification Tree

Confusion Matrix and Statistics

| | Reference | |
|------------|-----------|----|
| Prediction | 0 | 1 |
| 0 | 286 | 29 |

1 83 43

Accuracy : 0.746

95% CI : (0.7027, 0.786)

No Information Rate : 0.8367

P-Value [Acc > NIR] : 1

Kappa : 0.286

McNemar's Test P-Value : 5.499e-07

Sensitivity : 0.59722

Specificity : 0.77507

Pos Pred Value : 0.34127

Neg Pred Value : 0.90794

Prevalence : 0.16327

Detection Rate : 0.09751

Detection Prevalence : 0.28571

Balanced Accuracy : 0.68614

'Positive' Class : 1

We observed that the sensitivity increased to 0.59722 from the previous .36111 and the model correctly predicted 43 observations as belonging to Class 1 in comparison to the earlier 21 observations.

Applying under sampling in Classification Tree

| Reference | | |
|------------|-----|----|
| Prediction | 0 | 1 |
| 0 | 272 | 23 |
| 1 | 97 | 49 |

Accuracy : **0.7279**

95% CI : (0.6838, 0.7689)

No Information Rate : 0.8367

P-Value [Acc > NIR] : 1

Kappa : 0.2955

Mcnemar's Test P-Value : 2.666e-11

Sensitivity : **0.6806**

Specificity : **0.7371**

Pos Pred Value : 0.3356

Neg Pred Value : 0.9220

Prevalence : 0.1633

Detection Rate : 0.1111

Detection Prevalence : 0.3311

Balanced Accuracy : 0.7088

'Positive' Class : 1

The sensitivity increased to 0.6806 and the model correctly predicted 49 observations as belonging to Class 1.

Applying Over Sampling in Random Forest

Reference

Prediction 0 1

0 358 46

1 11 26

Accuracy : 0.8707

95% CI : (0.8358, 0.9006)

No Information Rate : 0.8367

P-Value [Acc > NIR] : 0.02816

Kappa : 0.4119

Mcnemar's Test P-Value : 6.687e-06

Sensitivity : 0.36111

Specificity : 0.97019

Pos Pred Value : 0.70270

Neg Pred Value : 0.88614

Prevalence : 0.16327

Detection Rate : 0.05896

Detection Prevalence : 0.08390

Balanced Accuracy : 0.66565

'Positive' Class : 1

The sensitivity increased to 0.36111 in comparison to the previous 0.29167 and the model correctly predicted 26 observations as belonging to Class1 in comparison to the previous 21 observations.

Applying under sampling in Random Forest

Reference

Prediction 0 1

0 358 48

1 11 24

Accuracy : 0.8662

95% CI : (0.8308, 0.8966)

No Information Rate : 0.8367

P-Value [Acc > NIR] : 0.05091

Kappa : 0.3827

McNemar's Test P-Value : 2.775e-06

Sensitivity : 0.33333

Specificity : 0.97019

Pos Pred Value : 0.68571

Neg Pred Value : 0.88177

Prevalence : 0.16327

Detection Rate : 0.05442

Detection Prevalence : 0.07937

Balanced Accuracy : 0.65176

'Positive' Class : 1

The sensitivity decreased to 0.33333 and the model correctly predicted 24 observations as belonging to Class1.

Comparing the two data models we can conclude that after using oversampling and undersampling, the Classification tree model is able to predict data points belonging to Class 1 more efficiently than Random Forest.

Although the accuracy of Random Forest is much higher than the Classification tree, in the case of our data set since more observations belong to Class 0 than Class 1, the class(Class 1) having lesser data, has minimal effect on the accuracy.

RESULTS & INSIGHTS

Who can benefit from the data?

Building a financially sound organization includes developing structures that keep your best employees around. The data mainly helps companies who see's high attrition rate as well as other companies who are not seeing such trends but can use this analysis to keep the rates below in future.

How would the data help them to make better decisions?

Based on our analysis of the data, we can provide the following suggestion:

1. The employees who travel frequently have a higher Attrition than those who do not travel. Travel can cause burn out, hence the company needs to provide appropriate travel allowance or some kind of compensation to these employees to make up for it.
2. The Sales and Human Resource Department has the highest Attrition. These departments should be studied closely to identify issues like job dissatisfaction, overtime etc.
3. If the work environment is good, there is a higher chance of Attrition. Some ways to achieve this would be by conducting fun team activities once a month or conducting team lunch etc.
4. The less the employee is involved in his job, the less satisfied he is and more is the Attrition. Less involvement would probably mean the employee is not interested in the work and is probably looking for better prospects elsewhere. Employees should be encouraged to approach their supervisors to speak up on

this issue. If lesser job involvement can be identified at an earlier stage, Attrition can be stopped.

5. Attrition among single individuals is higher. Efforts should be made to keep these employees motivated.
6. There is less Attrition in Senior level roles than fresher or mid level roles. It is important to identify the job prospects in these roles and determine whether these employees are seeking for more challenging projects or is it the monthly income or overtime work which is causing the Attrition.
7. The training time plays an important role in Attrition. There is more Attrition among employees who have not been trained even one time. Hence training must be provided to all employees.

What other data would be useful to have?

The dataset does not contain time series that would allow us to see the time point when the company experiences more attrition.

Recommendations

The coefficients generated through the model could be used to identify the importance of the variables. Organizations can focus on these important variables to control its attrition problem.

- **Overtime:** Implement policies and practices to limit working hours to a reasonable time.
- **Experienced Employees and Not Promoted:** Experienced employees who are not promoted might be lacking motivation to continue working for the organization. More transparency around promotions would help.
- **MonthlyIncome:** Offer competitive compensation for low wage employees.
- **Marital Status:** Singles are more likely to quit. Organizations can use this data to encourage singles to stay.
- **Frequent Business Travels:** Look to reduce unnecessary travel and distribute travel among peers on a rotation basis.

Reference:

Dataset: <https://www.kaggle.com/patelprashant/employee-attrition/version/1>

Attrition definition: <https://www.merriam-webster.com/dictionary/attrition>