**Predicting Employee Attrition**

**Fatima Soytemiz**

**April 1, 2021**

**Contents**

1. Introduction …………………………………………………………………………………………………………………..... 3
2. Data Acquisition ………………………………………………………………………………………………………………. 3
3. Data Cleaning ………………………………………………………………………………………………………………….. 3
4. Exploratory Analysis ………………………………………………………………………………………………………... 3
5. Data Pre-Processing ………………..………………………………………………………………………………………. 15
6. Modeling ………………………………………………………………………………………………………………………... 16
7. Conclusion ……………………………………………………………………………………………………………..………. 16
8. **Introduction**

Attrition is a terminology in human resources which refers to the employees leaving the company. Attrition is measured by the number of employees are going out of the company either by voluntary resigning or laid off by the company. In general, high attrition is problematic for companies that causes the huge loss. There are many reasons for which the employees leave the company, such as; salary dissatisfaction, no career growth etc. The loss is not only in terms of the money but also the company sometimes loses the skilled employees who are the most valuable assets to the company (Morrison, 2014). If the company can predict the employee attrition (employees which are going to leave the company) in near future, they can also work on retention beforehand and avoid the loss of valuable employee. The prediction of attrition and retention is the part of the HR Analytics. how to retain talent and avoid attrition in the organizations.

In this project, I analyze the IBM Employee Attrition dataset to find the main factors why employees choose to leave and to help the company to predict whether or not a certain employee will leave the company by utilizing machine learning models.

1. **Data Acquisition**

I use IBM HR Analytics Employee Attrition & Performance dataset from Kaggle. The data is a fictional data set created by IBM data scientists. The data can be reached from this [link](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset). The data has 35 columns and 1470 observations and contains numeric and categorical data types. Each row has various attributes from each unique employee such as age, gender, job role, educational field, marital status, monthly income etc.

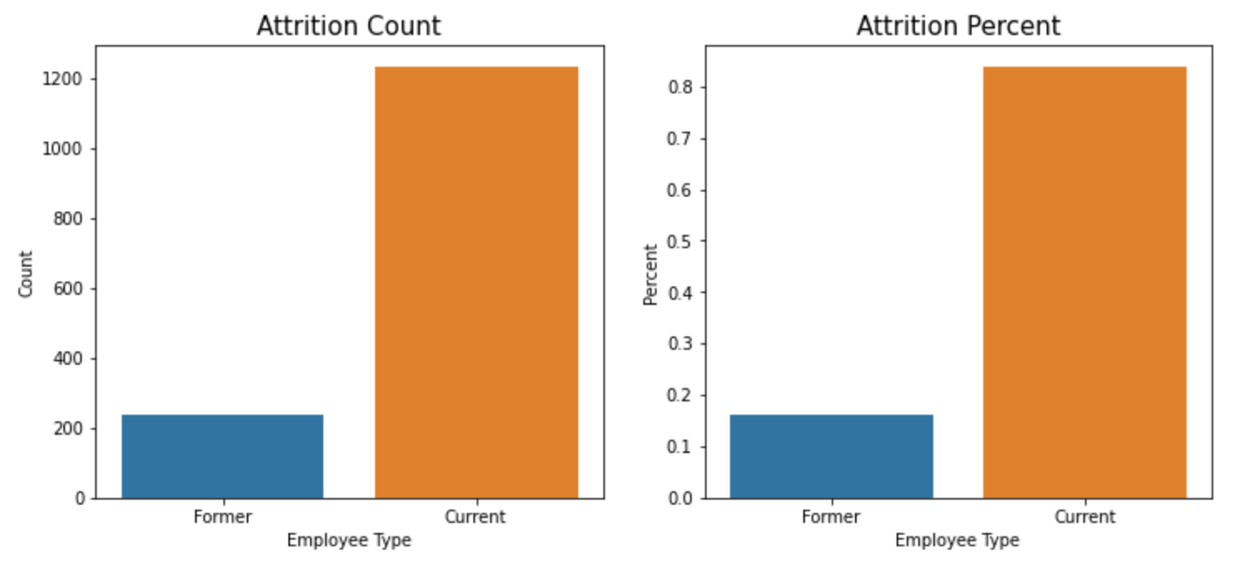
1. **Data Cleaning**

There are no missing values. Also, there are no duplicated values in the dataset. “Over18”, “EmployeeCount” and “StandardHours” variables are deleted since they contain only one value and there is no variation in these features. “EmployeeNumber” is also deleted because it has all unique numbers like ID (unique indentifier) and it doesn’t generate any value for the model. I count the number of employees who left the company vs who stayed in the company.

1. **Exploratory Data Analysis**

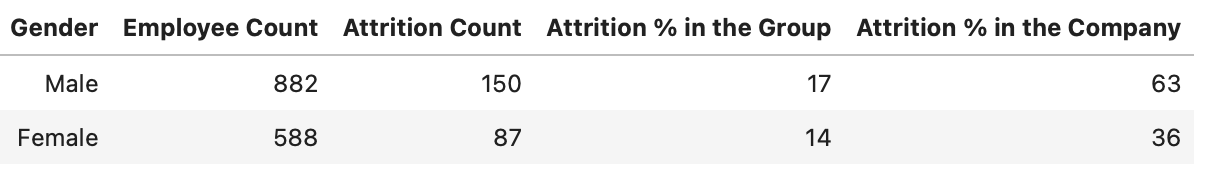
**4.1 Target Variable (Attrition):**

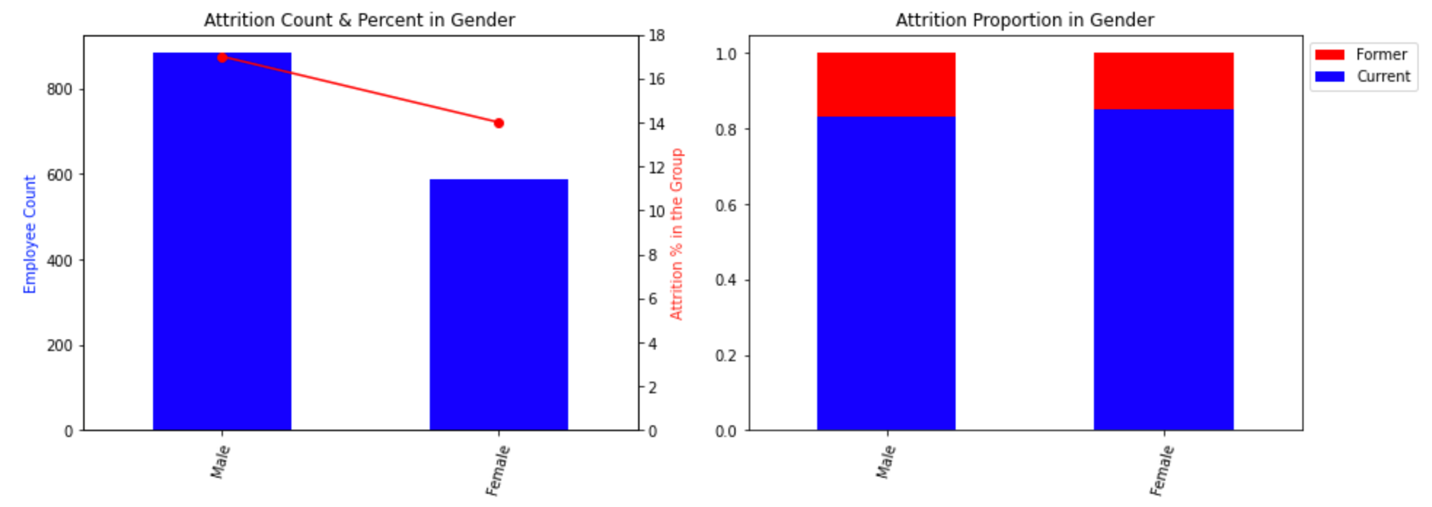
There are 1233 employees who stayed and 237 employees who left the company. As it is shown in the chart below, 16% of the employees left the company which tells us there is quite a large skew in target. Therefore, we have a quite big imbalance in our target variable. We will use an oversampling method known as SMOTE to treat the imbalance in the data.



**4.2 Categorical Features**

**4.2.1 Gender:**





According to the table and charts above, we see that the number of male employees is higher than females. Also, the attrition rate in males is higher than females.

**4.2.1.1 Hypothesis Testing:**

Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data.  Hypothesis Testing is basically an assumption that we make about the population parameter.

**4.2.1.1.1 Testing Employee Counts in Gender**:

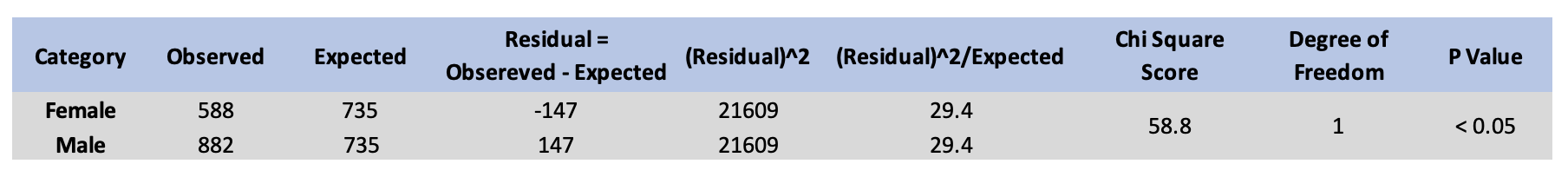
There are 558 females and 882 males are in the company. There is a difference between the employee counts. Is the difference significant? Is it too big or not really big? I can find the answer by using Chi Square test. The Chi Square statistics is commonly used for testing relationships between categorical variables. I would like to see the difference in numbers between female and male employee categories.

*Null Hypothesis:*

H0: There is no difference in the number of males and females in the company.

*Alternative Hypothesis:*

HA: There is difference in the number of males and females in the company.



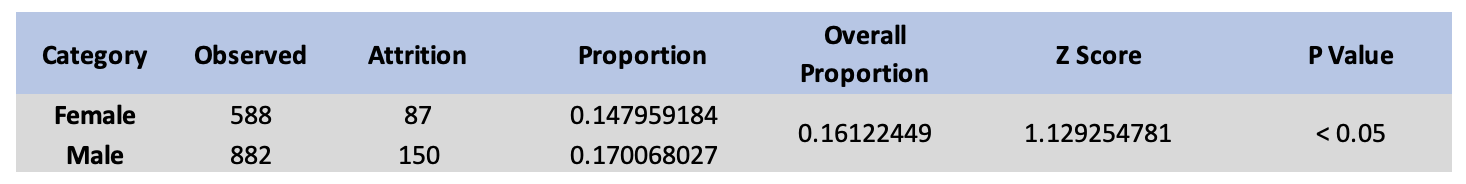
I use 95% significance level. As in the statistical calculation table above, the p value is less than 0.05. Therefore, we fail reject null hypothesis. I am confident that I can accept that there is no difference between males and females.

**4.2.1.1.2 Testing Attrition Rates in Gender:**

The attrition rate in male is 17% and 14% in female. It is obvious that male attrition is more than female attrition. Is the difference significant? Since I compare the attrition rate and the sample size is greater than 30 it is appropriate to use a two-proportion z-test.

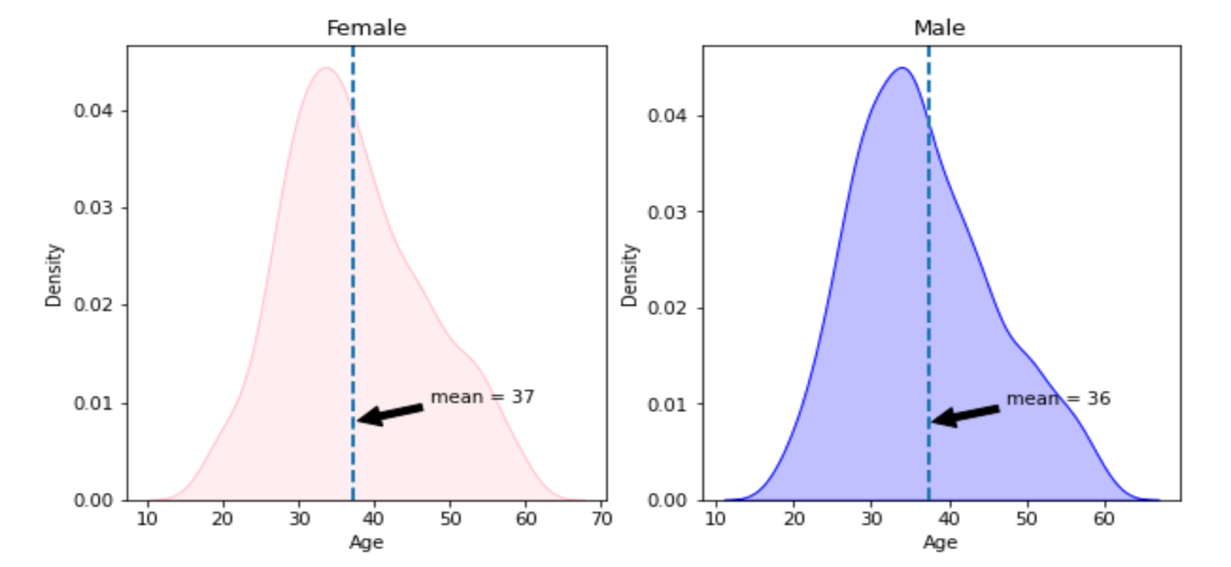
H0: There is no difference in the attrition rate for male and female employees in the company

HA: There is a significant difference in the attrition rate for males and female employees in the company.



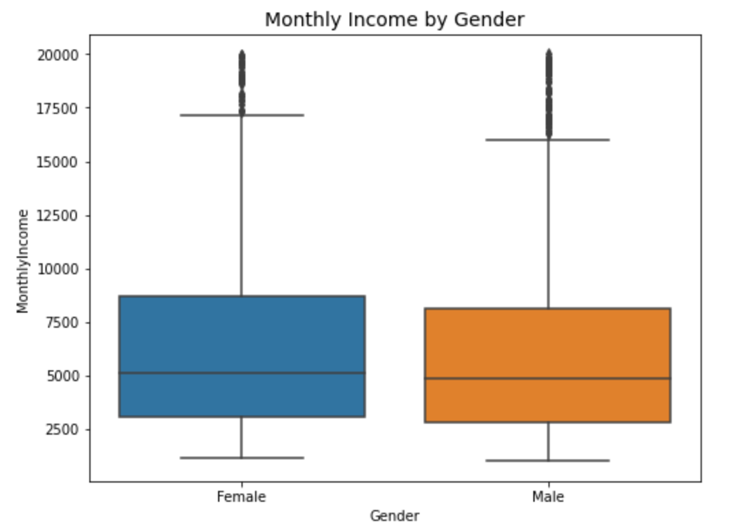
The p-value in the table above is below the significance level (0.05). Therefore, I fail to reject the null hypothesis. I can conclude that there is no significant difference in the attrition rates between female and male employees in the company.

**4.2.1.2 Age Distribution by Gender:**



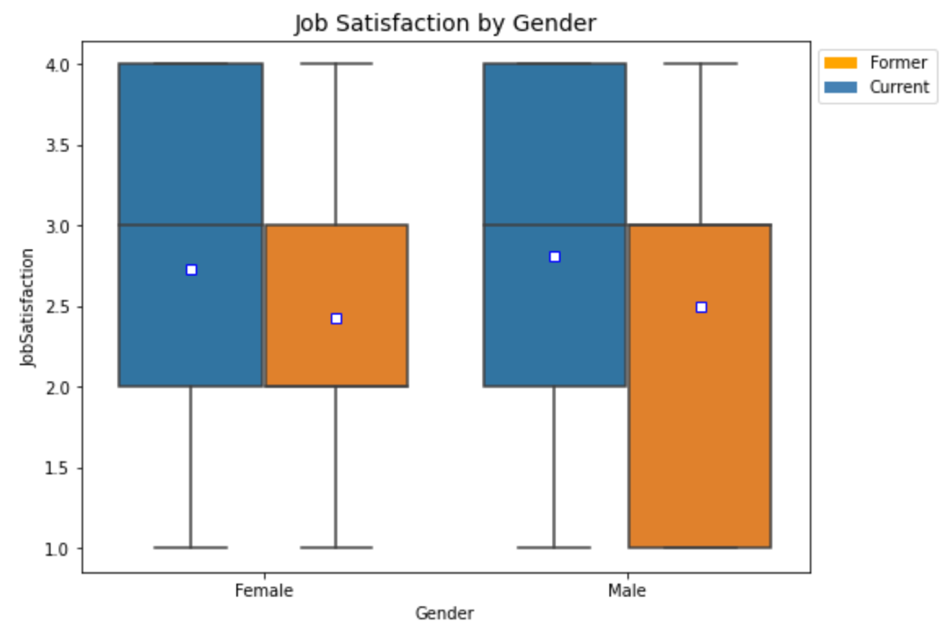
What is the age distribution between males and females? They both have the same distribution as below. The average age of females is 37 and for males is 36.

**4.2.1.3 Monthly Income by Gender:**



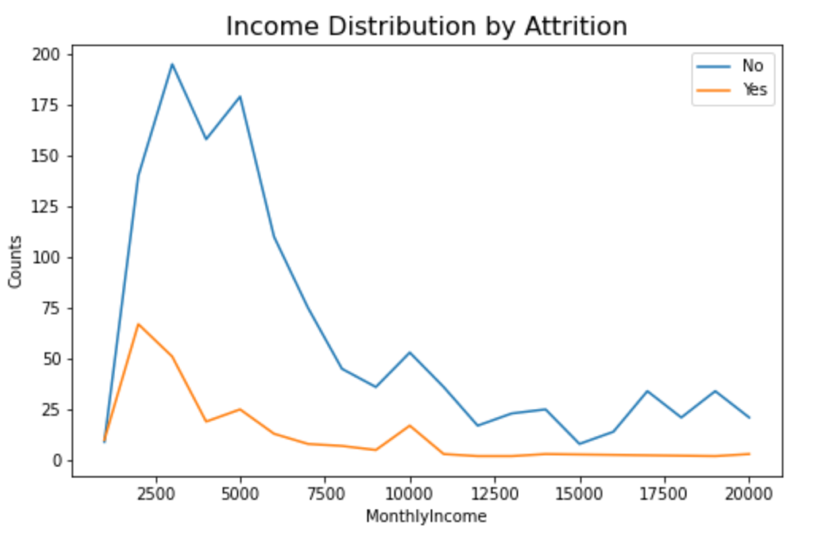
Is there any gender disparity in income? Based on the chart above, it seems that males and females have equal payments on average.

**4.2.1.4 Job Satisfaction by Gender:**

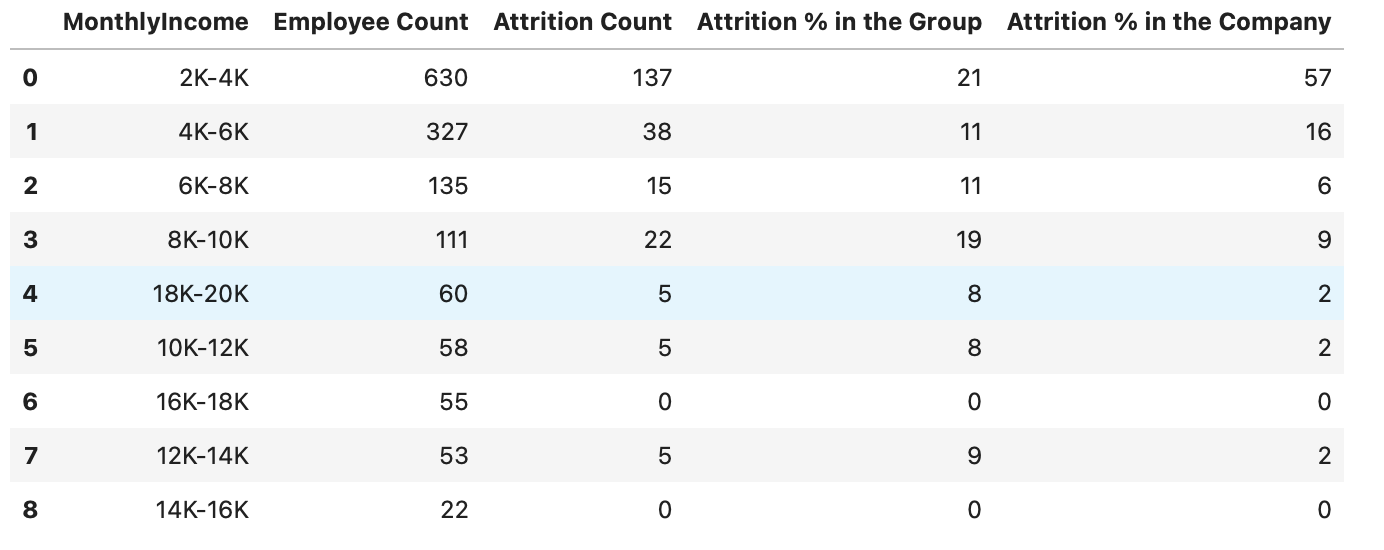


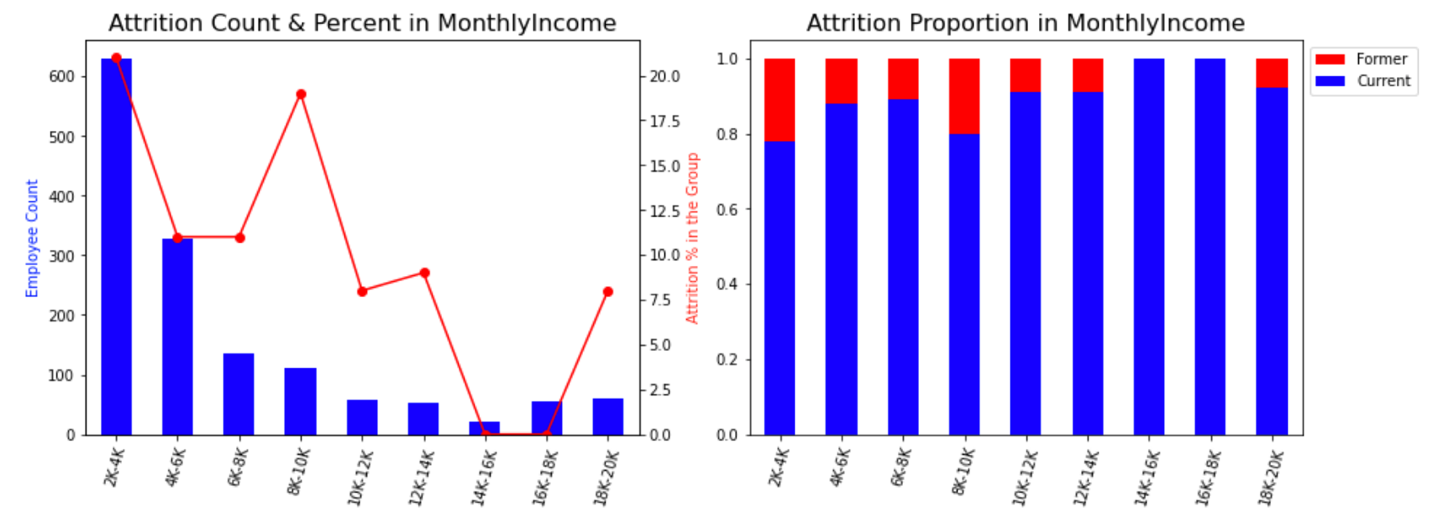
Based on the chart above, it is obvious that job satisfaction is low in males and females who left the company.

**4.2.2 Monthly Income:**

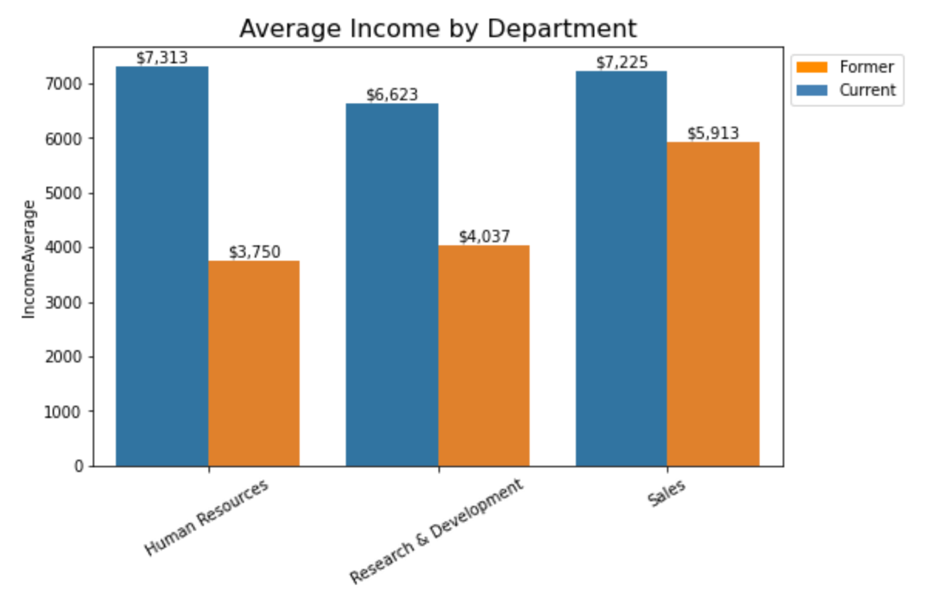


Is income the main factor toward employee attrition? As seen in the chart above, the attrition rate is obviously high at low-income levels, especially for the income levels less than $4,000. The attrition rate decreases after $4,000 but there is a minor spike at around $10,000. Employees at mid-income level tend to go to different company to upgrade salary or to shift toward a better standard of living. When the income is pretty decent, the chances of an employee leaving the company is low as it is seen by the flat line.



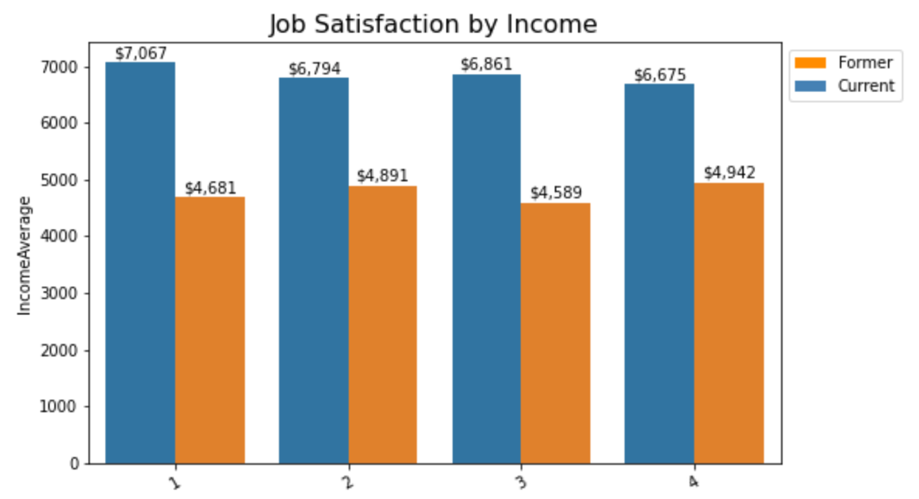


**4.2.2.1 Monthly Income by Department:**



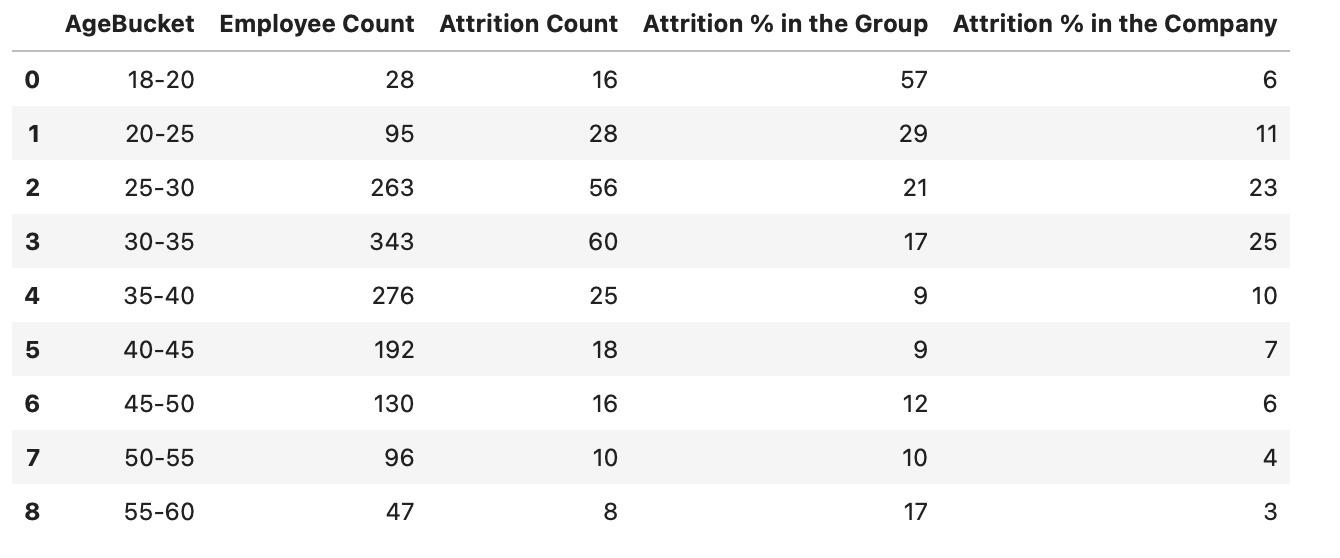
The attrition rate is high in Sales department despite the high average income compared to other departments.

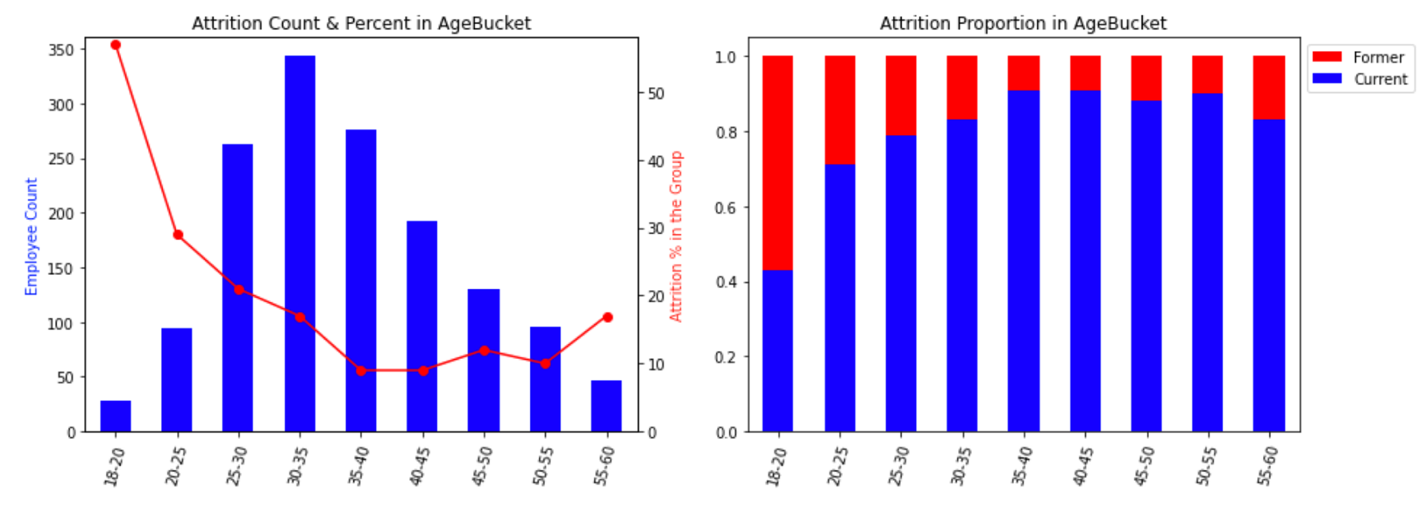
**4.2.2.2 Job Satisfaction by Monthly Income:**



As it is clearly seen in the chart above, employees who left the company have low income in all satisfaction levels.

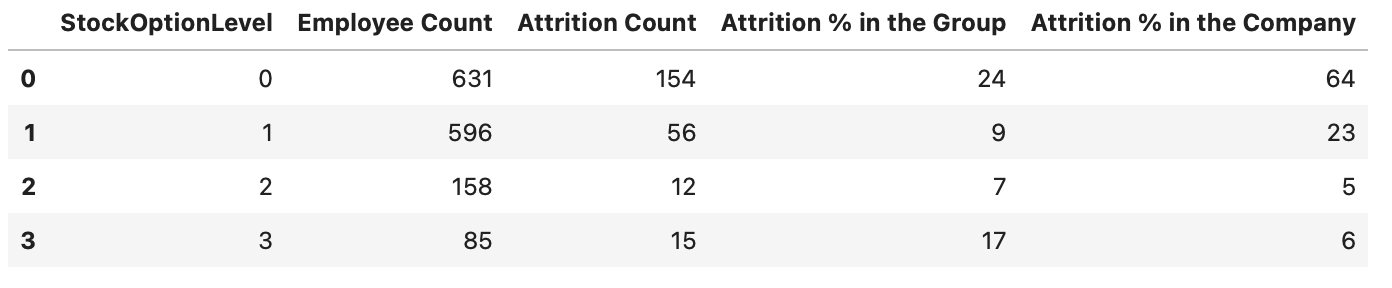
* + 1. **Age:**

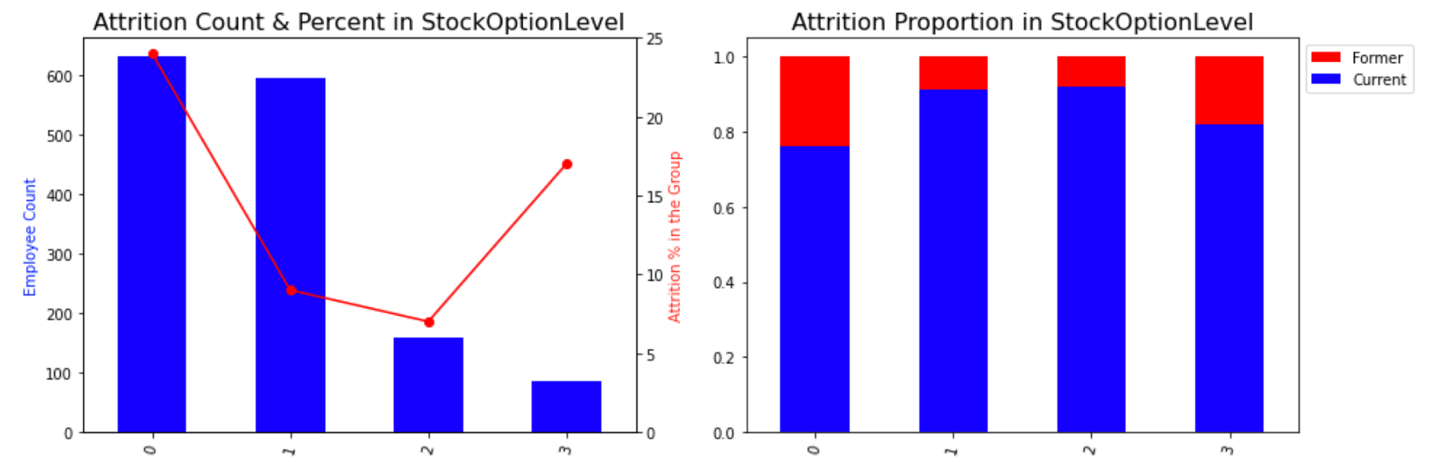




As it is seen in the chart above, the attrition very high for young people who are less than 25. The attrition keeps on failing with increasing age. After around age 35, people prefer stability in their jobs.

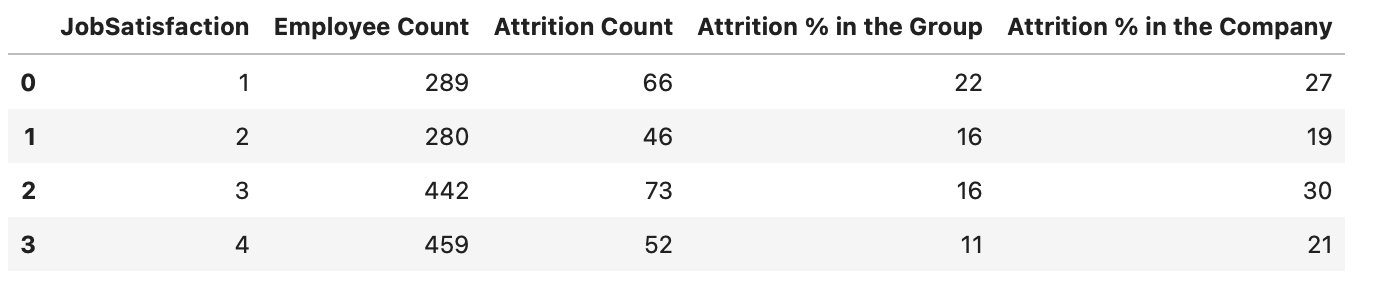
* + 1. **Stock Option:**

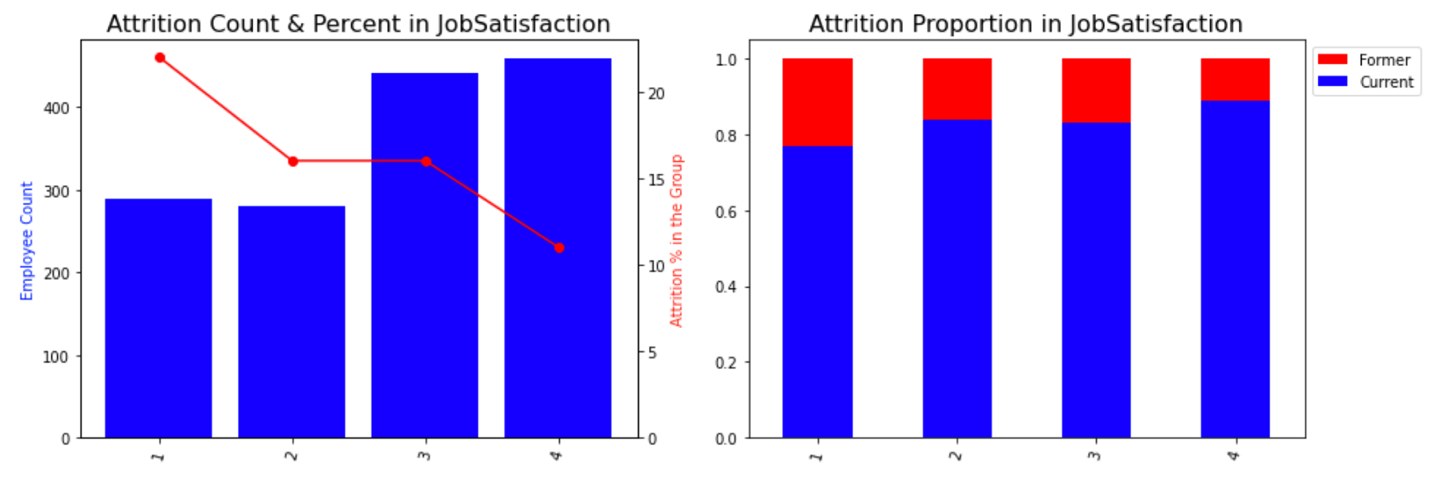




What is the attrition rate depending on the stock payments? The tendency of employees to leave the company is high when there is low or no stock options. Since the stocks contributes a huge amount of money, people do not want to lose that opportunity. People tend to leave the company if they are not happy with stock options.

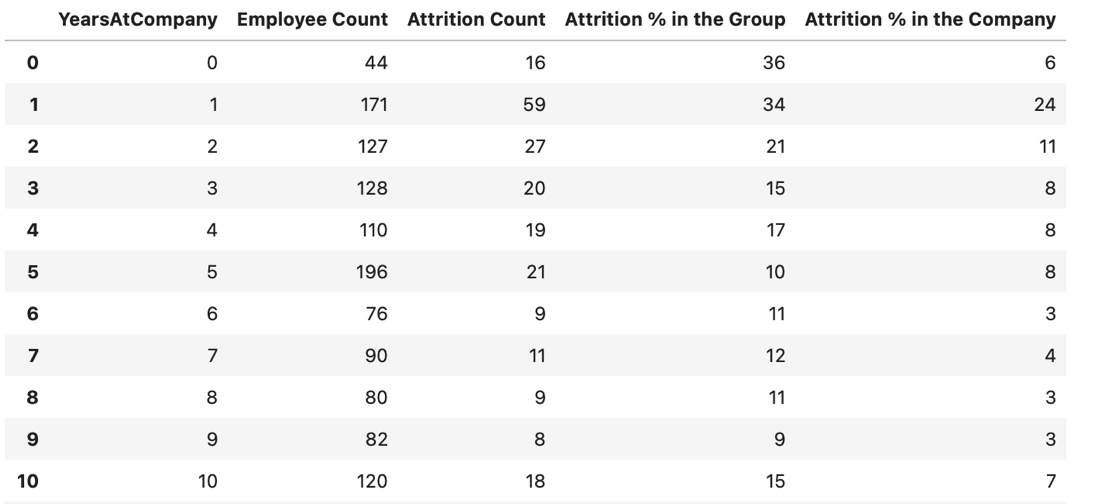
* + 1. **Job Satisfaction:**

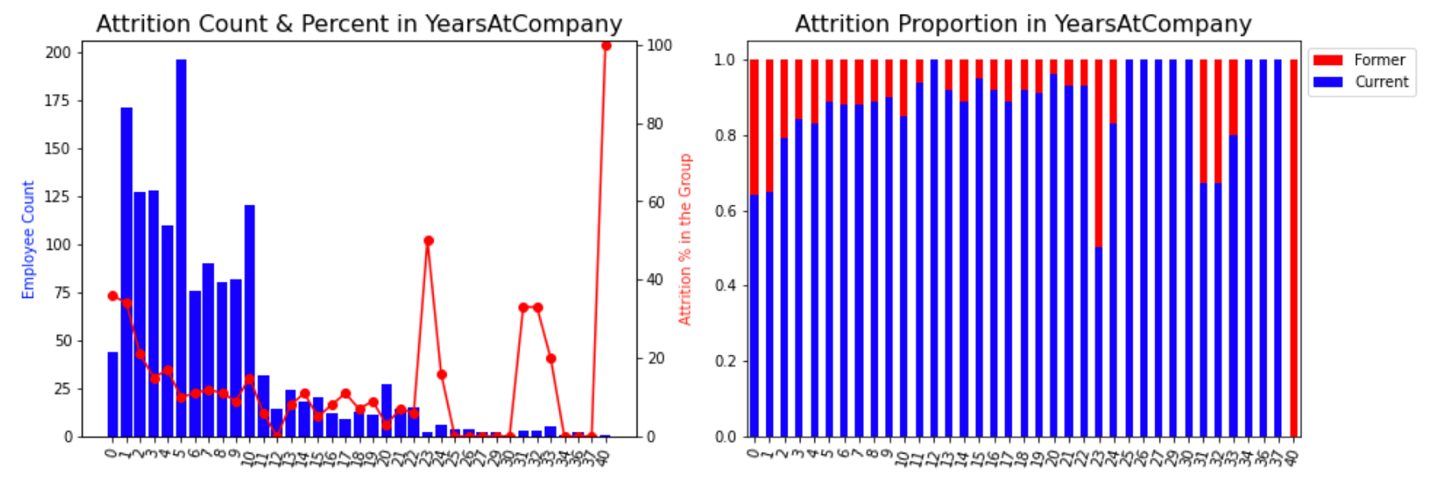




There are four job satisfaction levels in the data. 1 indicates low satisfaction and 4 indicates high satisfaction. As it is seen in the chart above, employees who have low job satisfaction most likely to leave the company more than employees who have high job satisfaction. The attrition rate in low job satisfaction is 27% and the attrition rate in high job satisfaction is 21%. With an increasing job satisfaction, the attrition rates decrease.

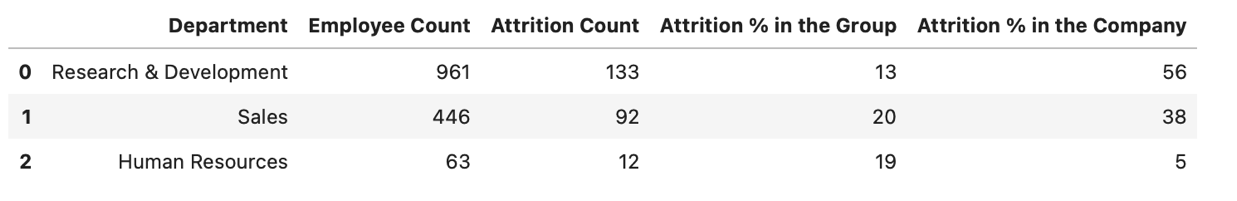
* + 1. **Years at Company:**

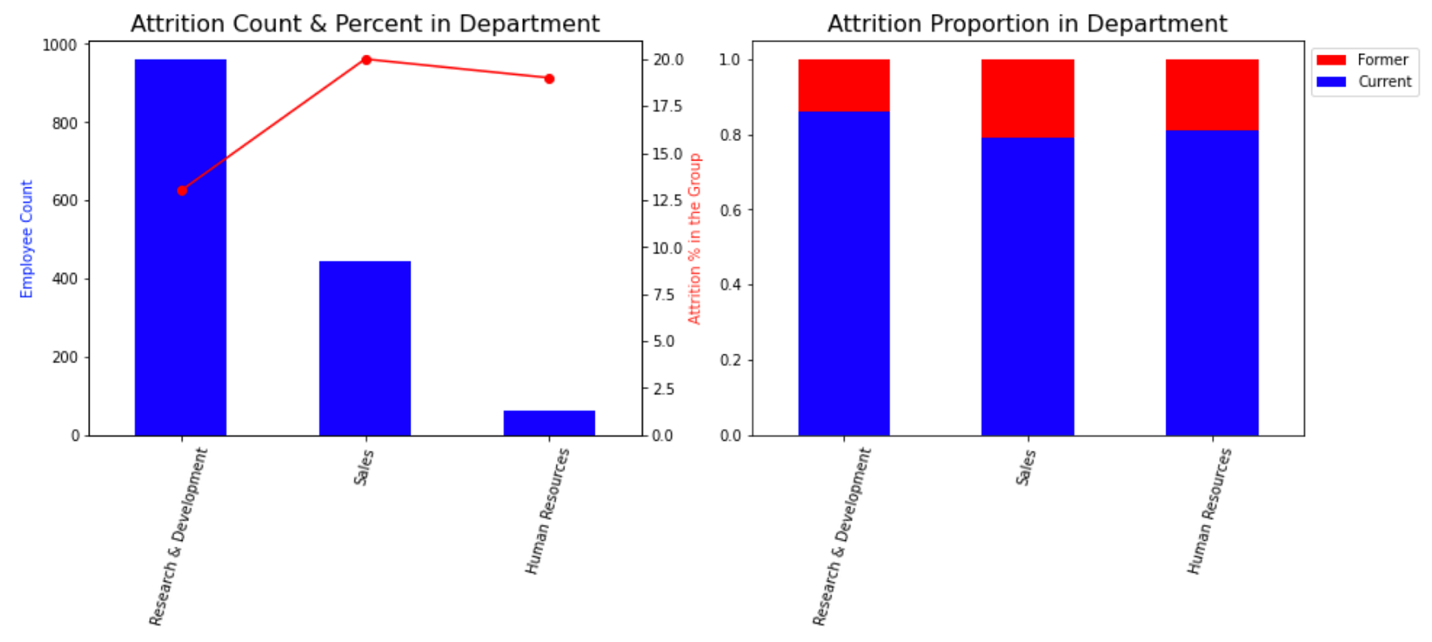




As it is seen in the chart and table above, employees who have two years or less working experience in the company has the highest percentage. It composes 41% of the total attrition in the company. Employees who are in their initial years have a higher chanced of leaving the company. People who have gained working experience tend to stay in the company.

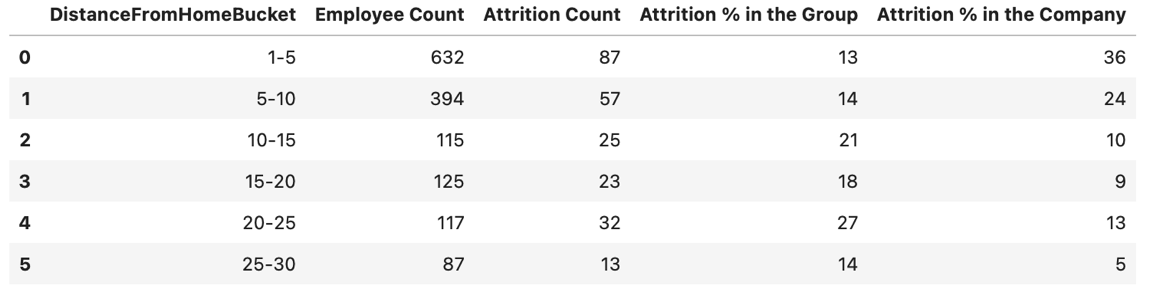
* + 1. **Department:**

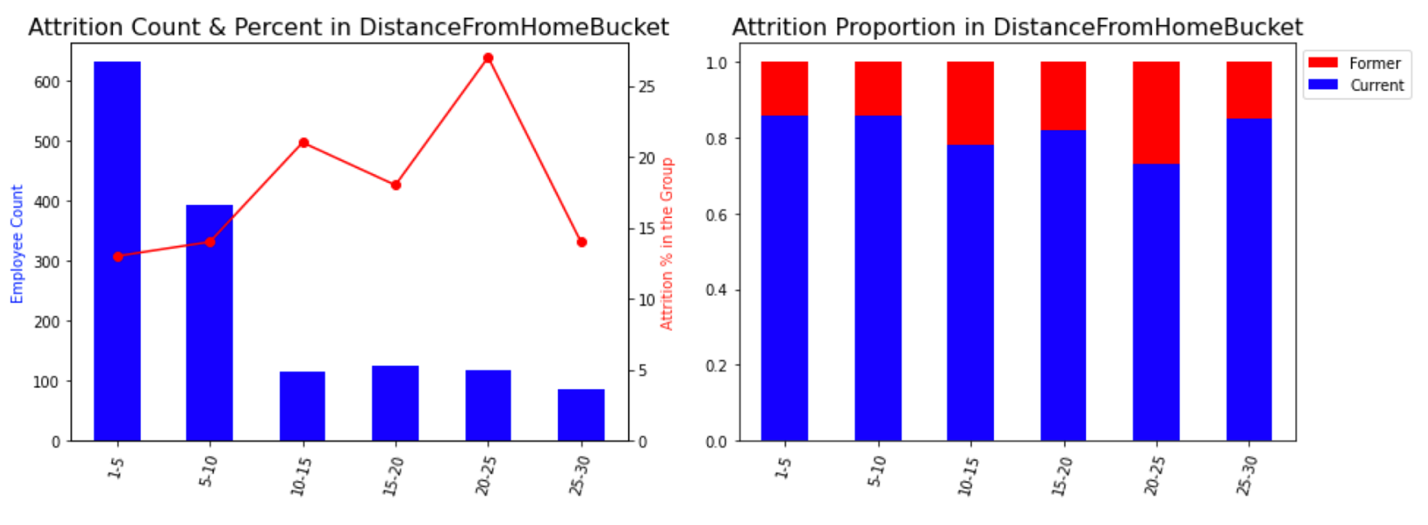




There are three departments in the company based on the data. The Sales Department has the highest attrition rates 20% and it is followed by the Human Resource Department which has 19%. Research and Development has the least attrition rates which shows employees prefer stability.

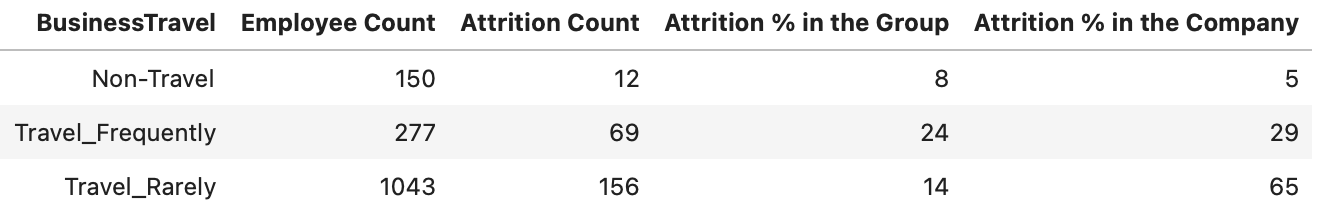
* + 1. **Distance from Home:**

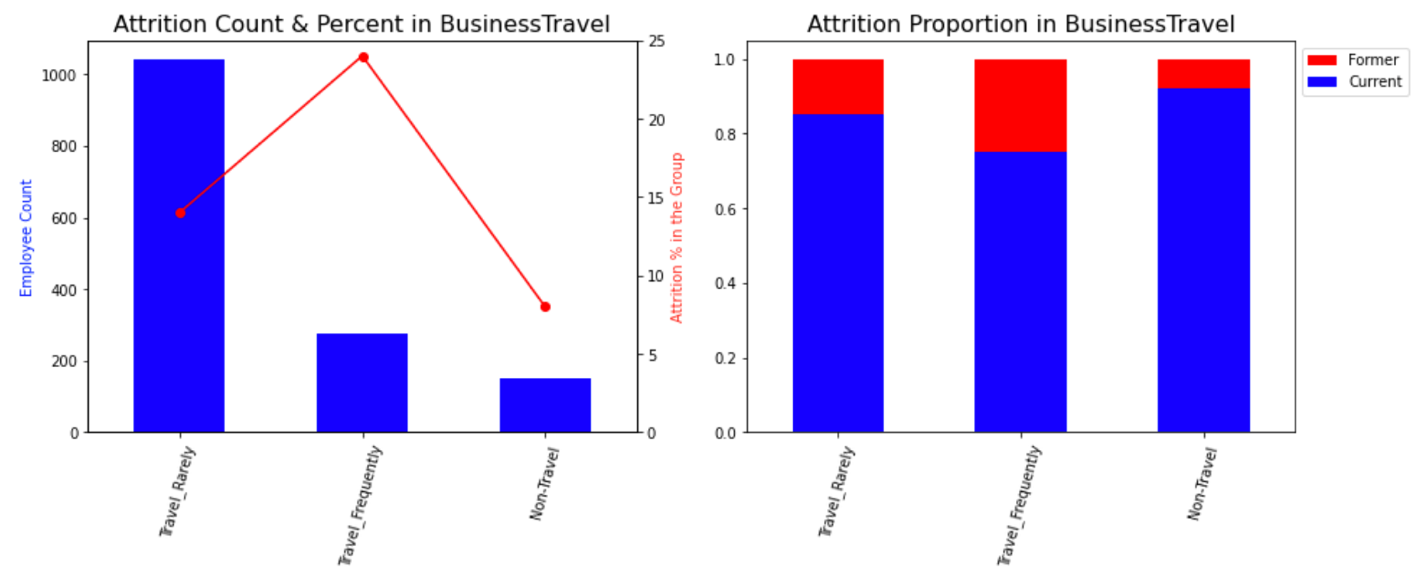




How does the distance from home impact attrition? It is clearly shown in the chart that people who live more than 10 miles away from the company are more likely to leave the company. Employees who live more than 10 miles away from the company compose 1/3 of the whole company attrition.

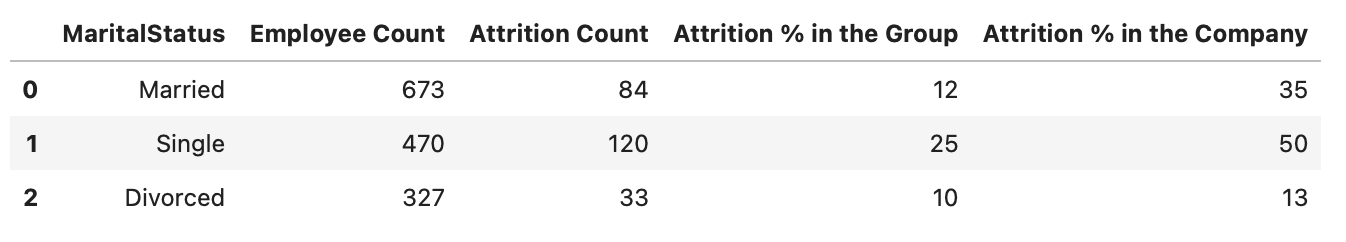
* + 1. **Business Travel:**

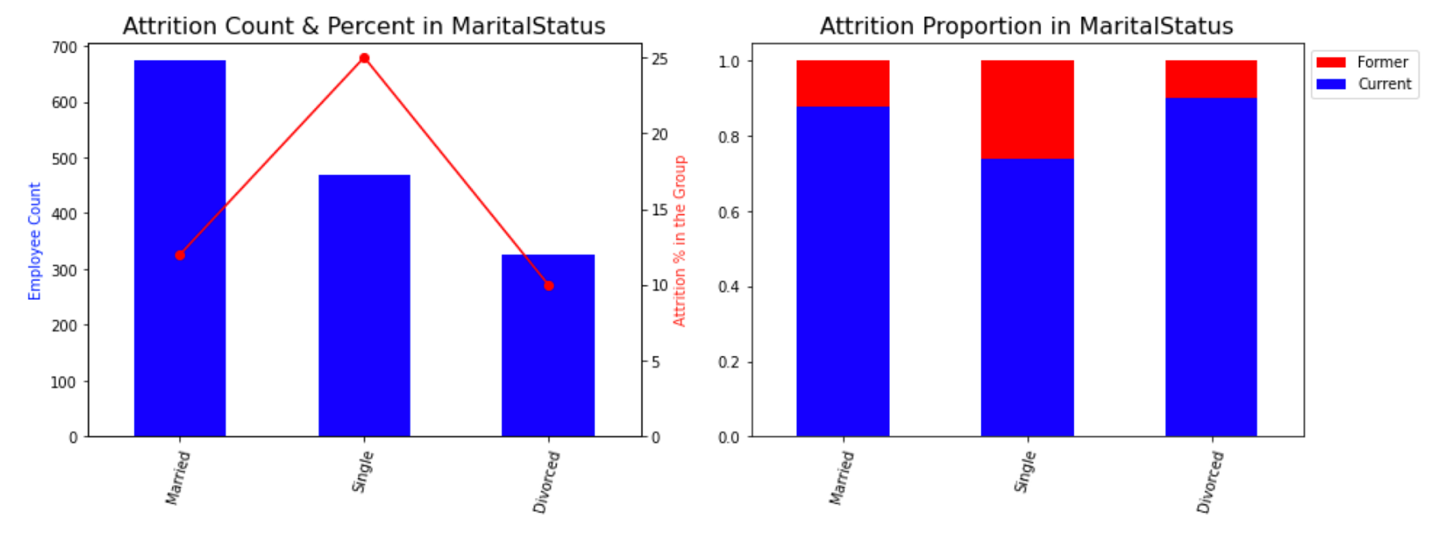




What is the impact of business travel? Employees who travel frequently have the highest attrition rate 24%. Employees who don’t travel has the lowest attrition rate in the company 8%. People do not prefer to travel in general.

* + 1. **Marital Status:**





Single employees are more likely to leave the company. They have the highest attrition rate which makes up 50% of all attrition in the company.

1. **Data Pre-Processing**

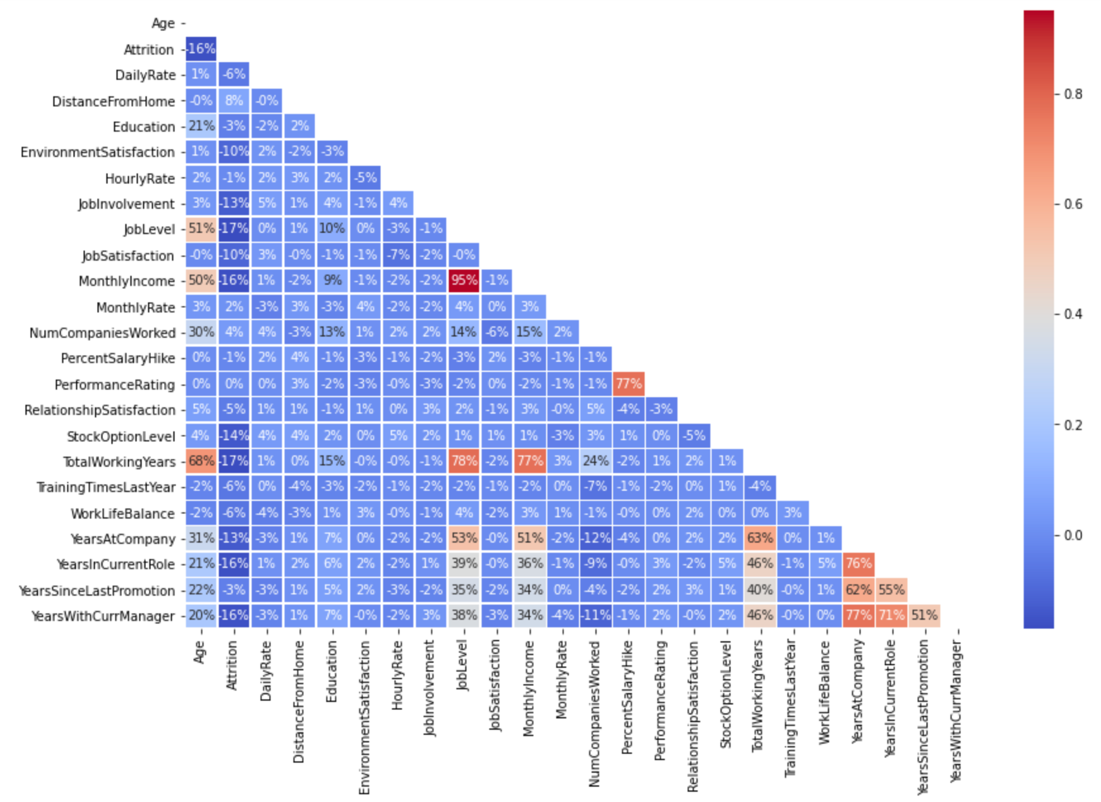
Before running the machine learning algorithms, I will do some pre-processing steps to make the dataset ready for model building.

**5.1 Feature Selection**

Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.  Adding redundant variables reduces the generalization capability of the model and may also reduce the overall accuracy of a classifier. Furthermore, adding more and more variables to a model increases the overall complexity of the model.

Some features just have one data level that do not contribute anything to our model. EmployeeCount, Over18 and StandardHours, and employee number does not have meaning in analyzing resulting so we also deleted these features.

We built correlation matrix which is a table showing correlation coefficients between variables as in the heat map below. MonthlyIncome variable and JobLevel variables have strong correlation (95%). Therefore, we delete JobLevel and keep MontlyIncome in the model. The rest of the variables have less correlation in general, we will keep all others for now.



**5.2 Dummy Variables**

We need to convert all the categorical data into numerical data for the Machine Learning model to work. We have used one hot encoding to create dummy variables. The basic strategy is to convert each category value into a new column and assign a 1 or 0 (True/False) value to the column. Dummy coding is a commonly used method for converting a categorical variable into continuous variable.

**5.3 Train Test Split**

First, we separated features and response variable as X and y. Then, we divided the dataset into the training and test sets. We have used 25% of the data for testing and 75% of the dataset for the training. Splitting our dataset is essential for an unbiased evaluation of prediction performance. We use the training set to build and train the model. Once the model is ready, we will test it on the testing set for accuracy and how well it performs. The objective is to have the model perform on any data with the highest accuracy.

**5.4 Over Sampling Imbalance Data**

An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed. Most of the machine learning algorithms used for classification that were designed around the assumption of an equal number of examples for each class. Imbalanced classifications significantly affect the model performance. This will result poor predictive performance, specifically for the minority class. As it is mentioned in the earlier sections, we have imbalanced data with 16% minority class. I have used oversampling method known as the SMOTE (Synthetic Minority Oversampling Technique) which increase the number of observations for the minority class. The SMOTE method randomly creates synthetic instances of the minority class so that the net observations of both the class get balanced out. We balanced only the training dataset and didn’t touch the test dataset.

**5.5 Feature Scaling**

Each feature can have different magnitude and different units. Variables that are measured at different scales do not contribute equally to the model. Some machine learning models are sensitive for scaling and some are not. Since most of the machine learning algorithms use Euclidean distance between two data points in their computation, this is a problem. We used StandardScaler to scale the data. With this scaling method, features will have mean of 0 and a standard deviation of 1. We fit the scaler on the training data and then used it to transform the testing data. This would avoid any data leakage during the model testing process.

1. **Modeling**
2. **Conclusion**