|  |  |
| --- | --- |
|  |  |
| Employee Attrition  Project |  |
|  |  |
|  |  |

|  |
| --- |
| Data Management and Visualization |
| Basmah Alnasair |

Table of Contents

**Acknowledgment 3**

**Objective 4**

**Background of the dataset………………………………………………………………………………..……………….5**

1. **Introduction and Motivated…………………………………………………………………..……………6**
2. **Related study……………………………………………………………………………………………..……..6**
3. **Processed Methods…………………………………………………………………………………………….7**

**A.Descriptive of the dataset……………………………………..…………………………..7**

**B.Dealing with missing data…………………………………………………………………9**

**C.Interesting statistic and data Visualization………………………………………..11**

1. **Discussion………………………………………………………………………………………………………25**
2. **References……………………………………………………………………………………………………….29**
3. **Appendix………………………………………………………………………………………………………29**

**Acknowledgment**

*We would like to express our special thanks of gratitude to our prof. Hanan Hosni and prof. Zuhaira Muhd Zain or delivering this valuable course, which is really important in our career as a data analyst. As well as giving us an opportunity to do this project (IBM employee attrition), which helped us in doing a lot of research and learn more.*

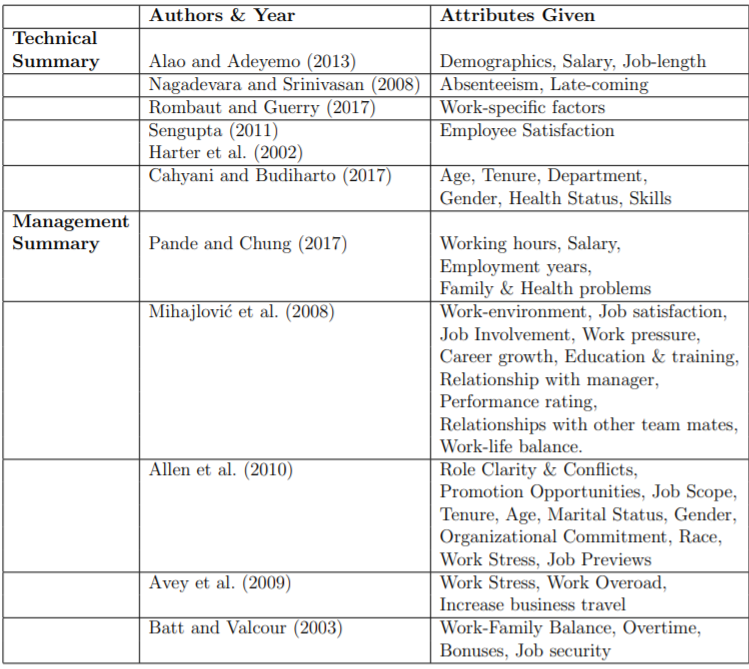
*Secondly, we would also like to thank our family and friends who helped us a lot in finalizing this project.*

**Objective**

The main objective of this project is to implement what we have learned throughout the course successfully. As long as using the right tools to the right problem. Furthermore, we aimed to make the dataset more readable for any organization to used in a future study. Visualization and figuring out the interesting statistic about some features of the *IBM employee attrition* dataset. In addition to that, we want to know whether the features that we have been selected impact the attrition state of employees.

**Background of the dataset**

For this project, an IBM employee, attrition dataset, has been picked, which is available on the IBM website1. Their objective is to use the supplied information for each employee to predict the attrition outcome. Therefore, employees with a high likelihood of leaving the company can be identified. This dataset was gathered throughout the years from 2003 to 2017 and ended up with 35 features. Table 1 shows how the researchers produced new features every year by conducting researches that used techniques of machine learning and deep learning to find out what causes the employees to attrite. This process was done by the HR analytics team in different countries until the dataset was finalized and published in 2017. This process, interesting enough, made it very difficult for the data to be used in a real-time analysis or studies. An additional difficulty was the collection of data, as due to the establishment of the data governance and privacy policies, it was very important to IBM to ensure the maintenance of the privacy of their candidates, customers, or even their employee's data [1].



*Table 1: a copy of Attributes responsible for Employee Attrition [1].*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

<https://developer.ibm.com/patterns/data-science-life-cycle-in-action-to-solve-employee-attrition-problem/#description>

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | I. Introduction and Motivated What does employee attrition mean? Employee attrition is known as losing an organization's employees by reasons related to the employee. For example, not offering a rewarding salary, working endless overtime, demanding roles that can't be handled, not working in a motivating environment or any personal reasons. Losing employees causes organizations bossily both time and money in replacement hiring and training new employees. As a result, employee attrition has always been a vital matter of concern for any workplace [2].  Because the attrition of skilled employees is a problem faced by companies worldwide, we were motivated to select the "IBM HR employee attrition" dataset as IBM is a well-known corporation. IBM published this dataset which listed information for 1470 employees with an indication of attrition outcome as a separate binary attribute. This project aims to present a view of employee attrition attributes and we have selected Python as our analytics tools.  In a beginning, we started by questioning one of the significant factors to see whether the employee got attrition because they were not getting paid well or not. We have answered this question by plotting a bar chart for the employees hourly rate. The answer found is not. In this project, we carefully selected the top effective features based on [2]. So we will be concentrating on the attributes (Overtime, Total Working Years, Monthly Income, Marital Status) and few other demographic variables. II. Related study The literature review shows that many factors can impact employee attrition. For Instance, In [2], which is tackled the class of imbalanced data. Th authors handled the imbalanced data by using an adaptive synthetic (ADSYN) approach. By doing that he achieved an F1-score of 0.909 using 12 features selection and random forest.  The remainder of this report is organized as follows. Section 2 proceeds the methods of exploring the data set and show the sample of the data. By first, explore the dataset. Then, it introduces descriptive statistics and data distribution for numeric features in the dataset. After that, we visualized decretive statistics and correlation relationships between different columns in the dataset. Section 3 discusses the findings of exploratory analysis. | |  |

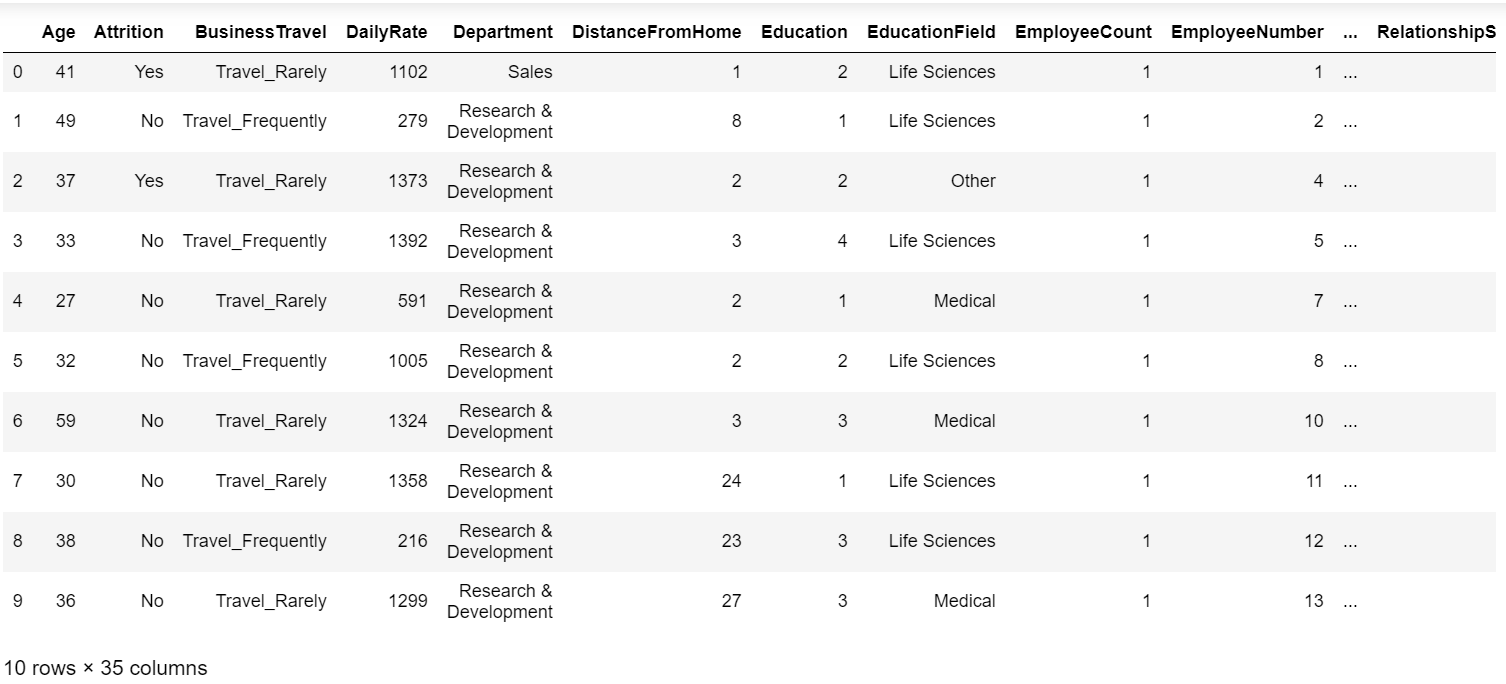
# III. Processed Methods

In this report, we explored the dataset using three main methodologies. First, we started to explore the dataset by using NumPy and pandas to inspect the meta-data and value range visually in the description of the dataset part. Second, we have dealt with the missing values part. We were using the same libraries as the first part. Third, we indicate the dataset statistics and dataset distribution as well as data visualization using in the addiction matlplotlip, seaborn, and plotly python libraries. In data visualization, we visualized the overall employee attrition rate and taking that into consideration with the top effective features. The features have been selected carefully, according to [2]. Also, we presented the relationship between attrition and four attributes ( two numeric, two categorical ), followed by the correlation between all features among each other.

The proposed methodologies are presented below with more details.

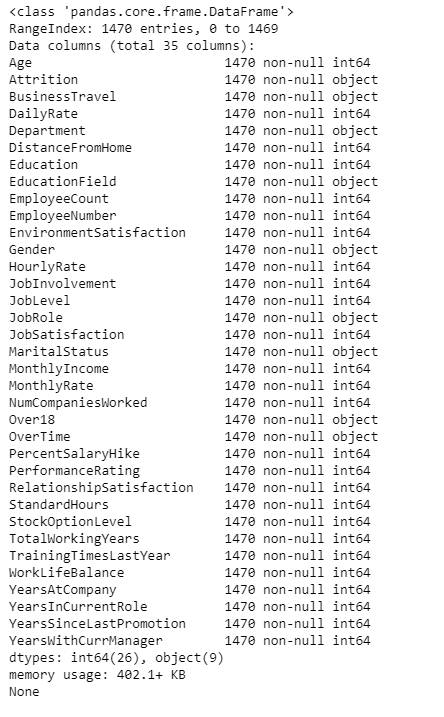
## A. Description of the dataset

We start by exploring the first ten rows from IBM employee attrition dataset to visually inspect the meta-data and value ranges. Figure 1 shows the result of selecting the first ten rows from the dataset. It can be observed that there are 35 columns in total for this dataset.

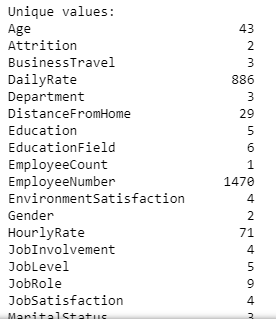


*Figure 1:* Head of the IBM employee, attrition dataset.

To inspect the meta-data of the employee attrition dataset. Figure 2 lists the available columns in the dataset with their data types. And in order to check the Noise data, Figure 3 shows the unique values for each column. It can be observed that we have **26** of integer columns and **9** of string columns with a total of **35** columns in this dataset.



*Figure 2:* list of features in the employee attrition dataset along with data type and number of rows for each of them



*Figure 3:* list the unique value of the employee attrition dataset.

## B. Dealing with missing value

In order to inspect for missing values in each column, we count the total number of rows with missing values and list the outcome in Figure 4. It can be observed that there are no missing values in this dataset. However, in case we had another scenario we would have used useful techniques. For example, dropping the columns or rows or fail in with the replacement of the average value.

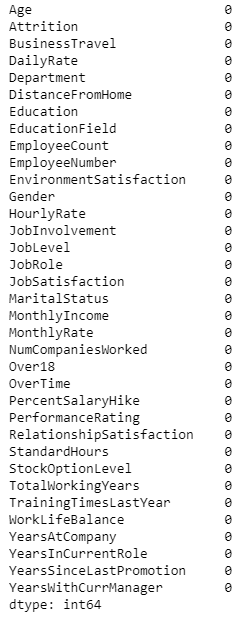


Figure 4: A list of columns with a total number of missing values.

As a result, from the above figures, we observed that the columns “Over18”, “Standard Hours,” and “Employee Count” contain the same value for each observation, which we dropped them as they do not support us in visualizing the dataset.

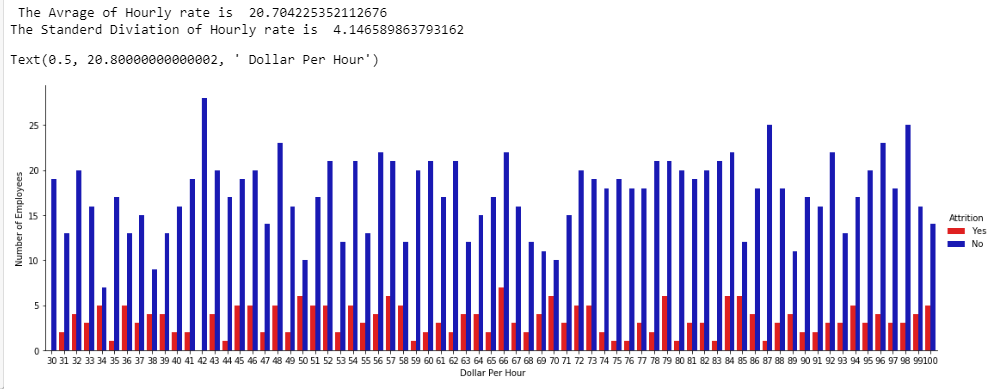
So, to summarize our findings, Table 2 illustrates each column with its corresponding data type. And how many unique values in each column does have.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Variable Name | Variable Type | Unique Values |
|  | Age | Quantitative – [Discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo)- Ratio | 43 |
|  | BusinessTravel | Qualitative – Nominal | 3 |
|  | DailyRate | Quantitative – [discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo)- Ratio | 886 |
|  | Department | Qualitative – Nominal | 3 |
|  | DistanceFromHome | Quantitative – [Discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo) - Ratio | 29 |
|  | Education | Qualitative – Ordinal | 5 |
|  | EducationField | Qualitative – Nominal | 6 |
|  | EmployeeCount | Quantitative – [Discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo) | 1 |
|  | EmployeeNumber | Qualitative – Nominal | 1470 |
|  | EnvironmentSatisfaction | Qualitative – Ordinal | 4 |
|  | Gender | Qualitative – Nominal | 2 |
|  | HourlyRate | Quantitative – [discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo)- Ratio | 71 |
|  | JobInvolvement | Qualitative – Ordinal | 4 |
|  | JobLevel | Qualitative – Ordinal | 5 |
|  | JobRole | Qualitative – Nominal | 9 |
|  | JobSatisfaction | Qualitative – Ordinal | 4 |
|  | MaritalStatus | Qualitative – Nominal | 3 |
|  | MonthlyIncome | Quantitative – [discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo)- Ratio | 1349 |
|  | MonthlyRate | Quantitative – [discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo)- Ratio | 1427 |
|  | NumCompaniesWorked | Qualitative – Nominal | 10 |
|  | Over18 | Quantitative – [discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo) | 1 |
|  | OverTime | Qualitative – Nominal | 2 |
|  | PercentSalaryHike | Quantitative – [discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo)- Ratio | 15 |
|  | PerformanceRating | Qualitative – Ordinal | 2 |
|  | RelationshipSatisfaction | Qualitative – Ordinal | 4 |
|  | StandardHours | Quantitative – [discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo) | 1 |
|  | StockOptionLevel | Qualitative – Ordinal | 4 |
|  | TotalWorkingYears | Quantitative – [discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo)- Ratio | 40 |
|  | TrainingTimesLastYear | Quantitative – [discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo)- Ratio | 7 |
|  | WorkLifeBalance | Qualitative – Ordinal | 4 |
|  | YearsAtCompany | Quantitative – [discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo)- Ratio | 37 |
|  | YearsInCurrentRole | Quantitative – [discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo)- Ratio | 19 |
|  | YearsSinceLastPromotion | Quantitative – [discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo)- Ratio | 16 |
|  | YearsWithCurrManager | Quantitative – [discrete](https://www.google.com/search?safe=strict&q=continuous+discrete&spell=1&sa=X&ved=2ahUKEwiZltPap7_oAhUDXRoKHU9HCGsQkeECKAB6BAgRECo)- Ratio | 18 |
|  | Attrition | Qualitative – Nominal | 2 |

*Table 2 :* data type and unique values of the IBM employee attrition dataset.

## C. Interesting statistics and data visualization

In the beginning, we started by examining whether the employee got attrition because they didn't get a rewarding salary? Figure bellowed gives us a clear answer, and this highlights the importance of data visualization. They were getting fast and correct answer from a big dataset.



*Hourly rate Vs. attrition state*

We see from the graph that the probability of attrition has occurred over a different amount of hourly rate. The average hourly rate was 20.70$ per hour, while the standard deviation was 4.14 .

In this part, we introduce the first-order statistics describing the distribution of numeric columns in the dataset. These statistics include meaning, standard deviation, max, min, the 25 percentile, the 50 percentile (the median), and the 75 percentile. Figure 5 lists the descriptive statistics for the first 9 numeric columns.



Figure 5: List of descriptive statistics for the first 9 numeric columns in the employee attrition dataset.

It can be observed that there is a significant difference in the value ranges across the numeric columns (see the highlight mean row in Figure 5). For example, the mean of the “*DailyRate*” column is “802.4”, while the mean of the “*JobInvolvement*” column is “2.7”. This difference demands the use of normalization techniques to neutralize its effect on the modeling method.

*Figure 6* depicts a boxplot for the value ranges of the numeric columns in the dataset. A significant difference in ranges can be spotted for the “*DailyRate*”, “*EmployeeNumber*”, “*MonthlyIncome*”, and “*MonltyRate*” columns in comparison to the remaining columns. Figure 7 we can see all the value ranges after doing a normalization.

Figure 7, we do correction for having different values range across all features by using normalization.

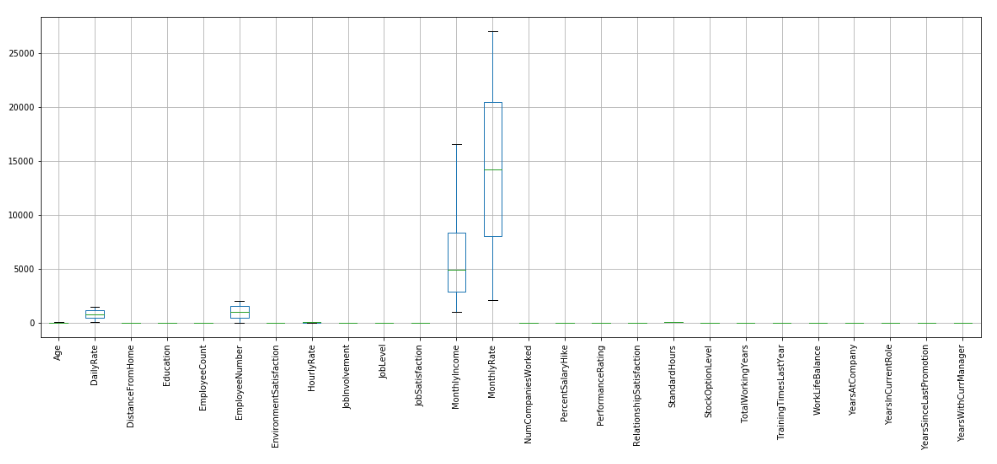


Figure 6: Boxplots for value ranges of numeric columns in the attrition dataset.

# A close up of a piece of paper Description automatically generated

Figure 7: Boxplots for value ranges of numeric columns in the attrition dataset using normalization.

Also, in this part, we show histogram visualizations for the statistical distributions of numeric columns in the employee attrition dataset. Figure 8 shows the histogram graphs for the numeric columns. It can be observed that almost 8 columns contribute to the majority of variance in the numeric columns’ values.

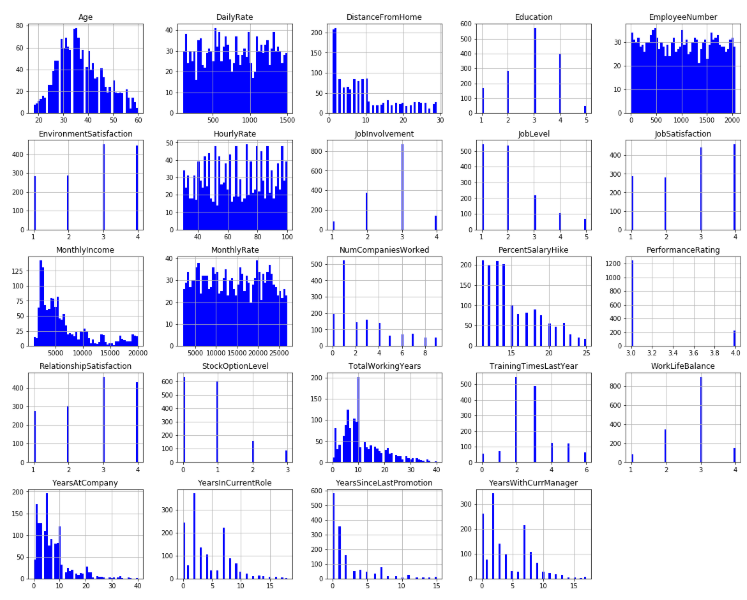
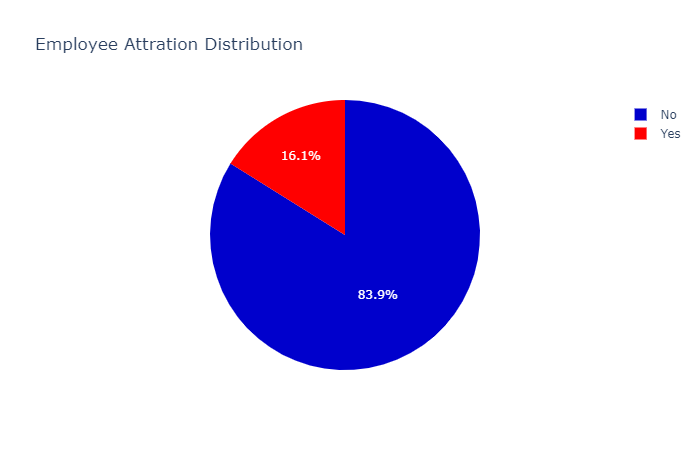
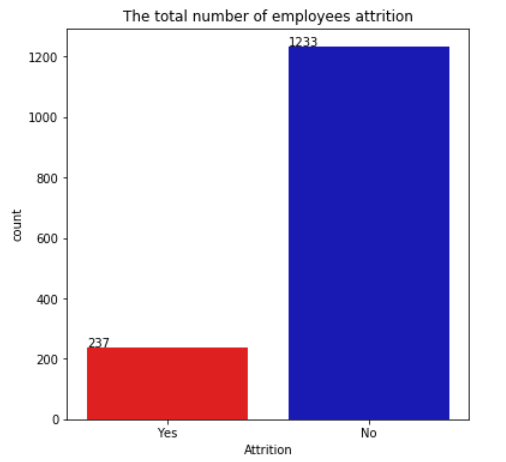


Figure 8: Histograms for the 24 numeric columns in the employee attrition dataset.

Figure 9 shows a pie chart of the overall attrition rate. First, we can see that the percentage of the total number of employees who got attrition represented only 16.1% of the total number of employees. Second, the employee not attrition rate reached nearly three times higher than the employee who really suffers from attrition. Lastly, 83.9% of employees not got attrition, which is a significant difference. This finding indicates a class label imbalance between the attrition classes (i.e., “Yes,” “No”). This imbalance will demand the use of suitable sampling techniques (e.g., stratified sampling) during the training procedure of a machine learning model to guarantee good generalization on future data.



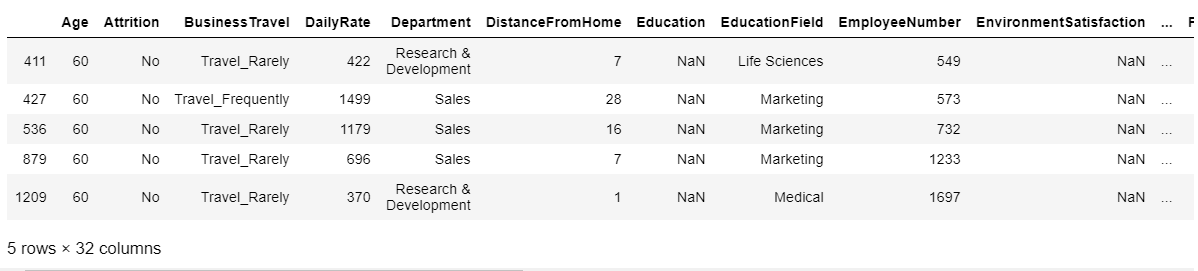
*Figure 9: A pie chart showing Employee Attrition Percentage*

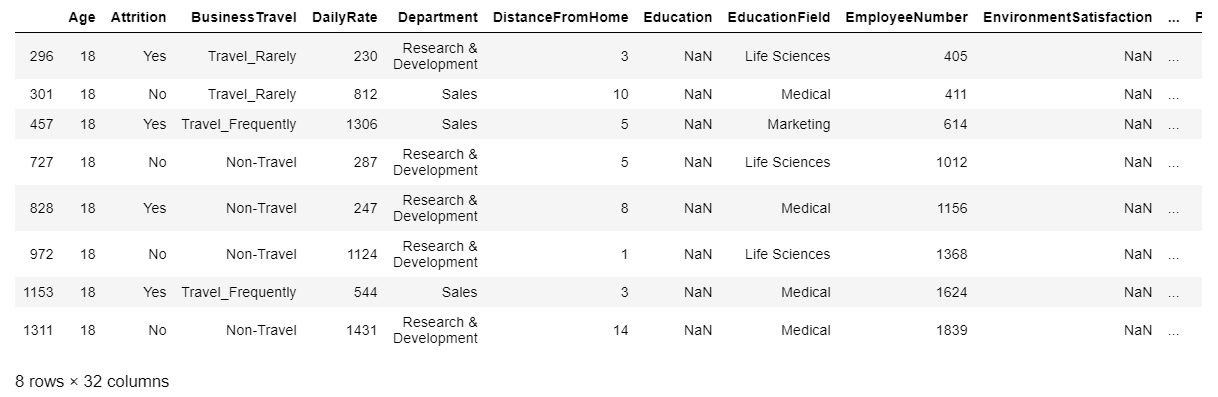


*Figure 10 : A bar chart showing the total number of samples for each attrition possible value, including “Yes” and “No.”*

Figure 10 shows a bar chart of the total number of records for each attrition value. As we can see, the numbers, 237 employees got attrition, and 1233 employees have continued in the company. We conclude the number of NO groups is largely higher than the YES group.

Now, we want to understand the employees' age. We cheeked about employees' age to find how old the older employee was the same, the younger age. We found, according to the result in Figure 11 and Figure 12, the oldest employee was sixty years old, while the youngest employee was eighteen years old. There are five employees had 60 years old, and eight employees had 18 years old.



*Figure 11: A list of the Oldest employees*

*Figure 12: A list of the Youngest employees*

As a result, older employees show more patient and responsible than the younger once. All the employees in their 60 years old were not got attrition. However, 50% of the employee in their 18 were suffered from attrition, and the other half of them were not.

In order to have a clear perception of the dataset, we wanted to know about the company's department and which one has the highest attrition rate. Figure 12 has answered that.

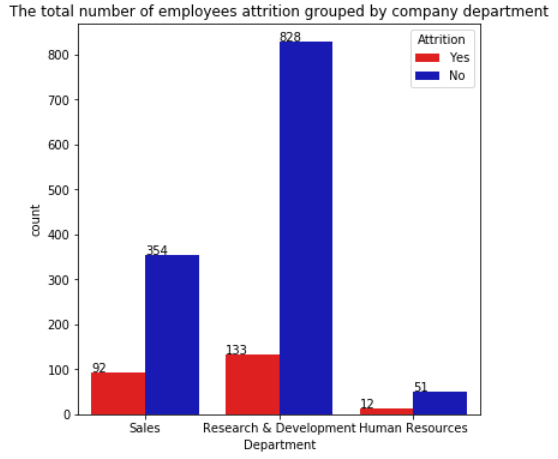
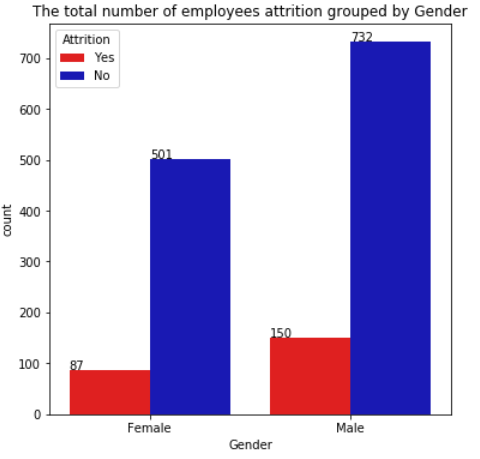


Figure 12: A bar chart showing the record count for the “Department” column across attrition column values.

Figure 12 shows that the R&D department had the highest number of employees who got attrition, followed by the sales department. In contrast, the human resources department had a minority of attritive employees. This indicates that demanding jobs such as researchers or salespeople have a higher likelihood of got attrition than less demanding ones such as recruitment specialists.

One more important question to be asked is, "which gender affected the most from attrition*?"*. Figure 13 answers this question.



*Figure 13: A bar chart showing the record count for the “Gender” column across attrition column values.*

Figure 13, shows that overall the number of male workers is higher than females. Concerning the attrition of males, employees got more attrition than female employees. As what numbers shown in the figure, 150 male employees got attrition. Female employees, meanwhile, are less likely to do so.

Due to the mentioned imbalance across attrition value, we wanted to focus on the records with attrition (16.1% of the total dataset records) to explore the distribution of their values. Figure 14 focuses only on the distribution of the numeric columns with records with attrition (i.e., attrition=” Yes”). We highlighted interesting findings from these distributions by the red rectangles, and we are going to discuss them in the next section.

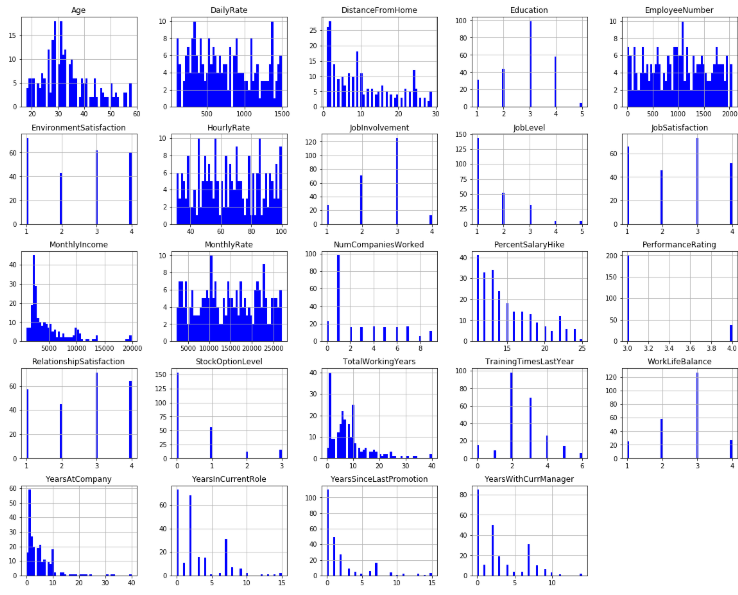


Figure 14: Histograms for the numeric columns in the employee attrition dataset with attrition (attrition = “Yes”).

Before we get in-depth with visualizing the dataset, we thought of mapping the categorical values with its corresponding meaning instead of having dummy numbers. We did that based on the published metadata. Of course, this is not a useful step to use in modeling, but it does with visualization. Figure 15 describes the meaning of numbers of categorical values.

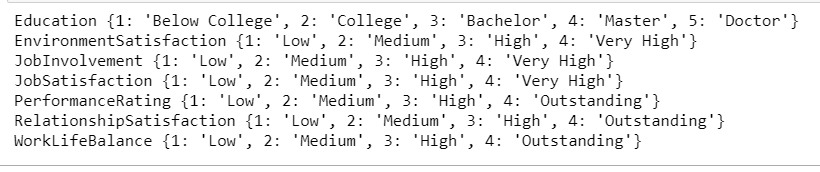
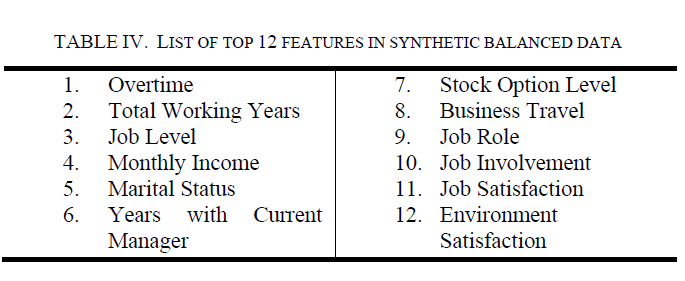


Figure 15, mapping the categorical values.

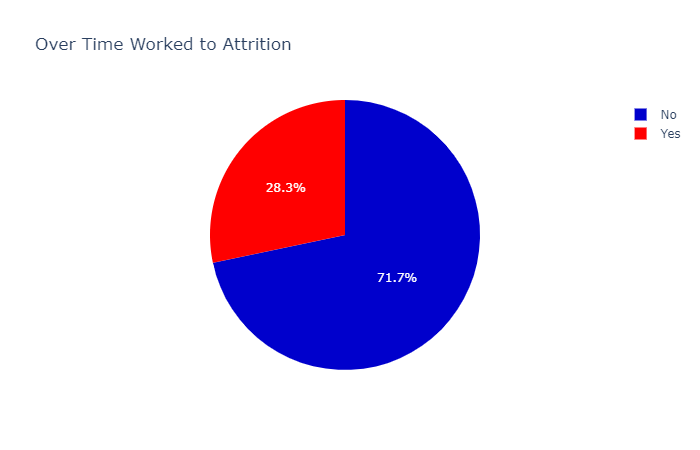
As well as the previous research [2] has tackled this problem in the same dataset and came up with the top 12 features used in training. They reached a high result of 0.909 F1 scores. Table 16 below shows a copy form their results:



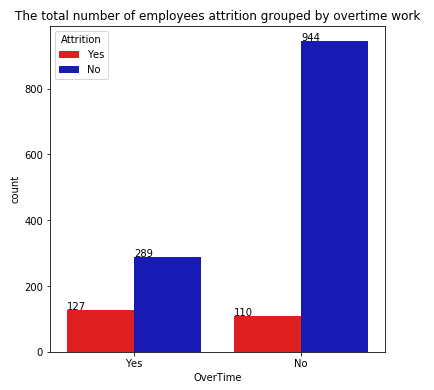
*Figure 16: a copy of 12 top features* of IBM employee attrition dataset [2]

On account of this result, we will be concentrating on this project on four features to study the relationship between them across by the attrition feature. In this part, we illustrate the results and discuss further information in the next section.

1. *Relationship of Over Time Worked to Attrition*

**

*Figure 17 : A pie chart showing Over Time Percentage*



*Figure 18: A bar chart showing the record count for the “Over Time” column across attrition column values*

1. *Relationship of Total Working Years to Attrition*

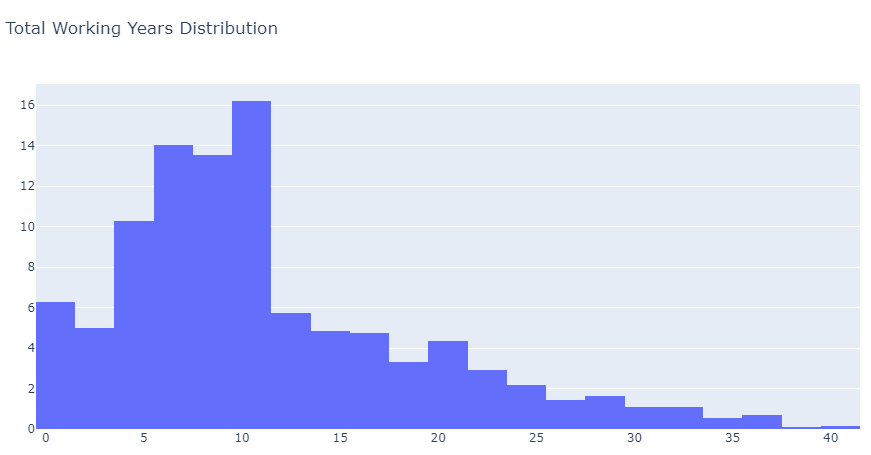


Figure 19: histogram of total working years of employee attrition dataset.

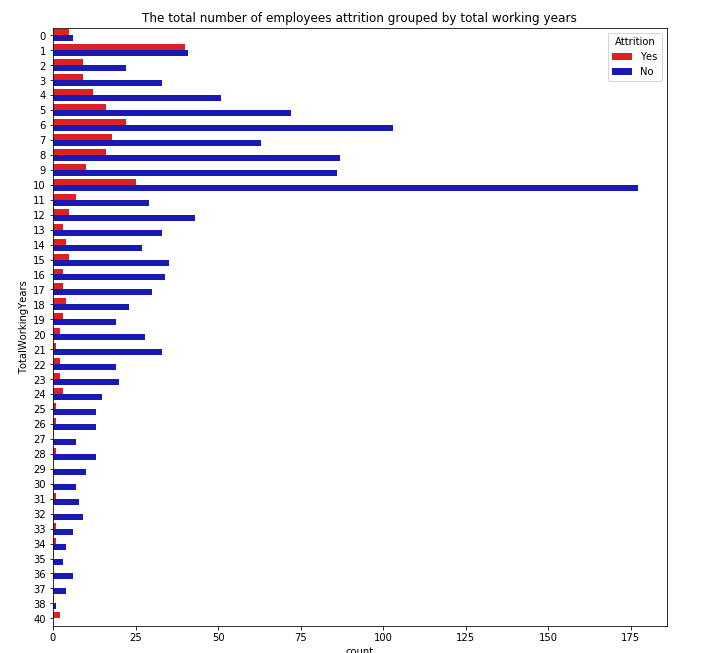


Figure 20: a bar chart of total working years of employee attrition dataset.

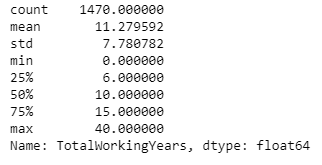
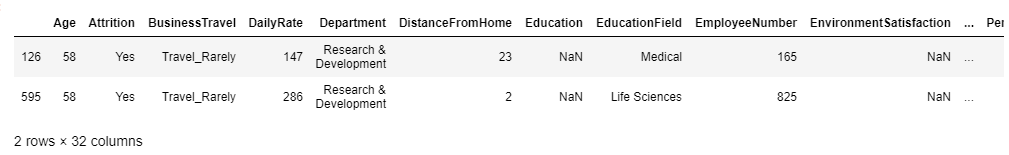
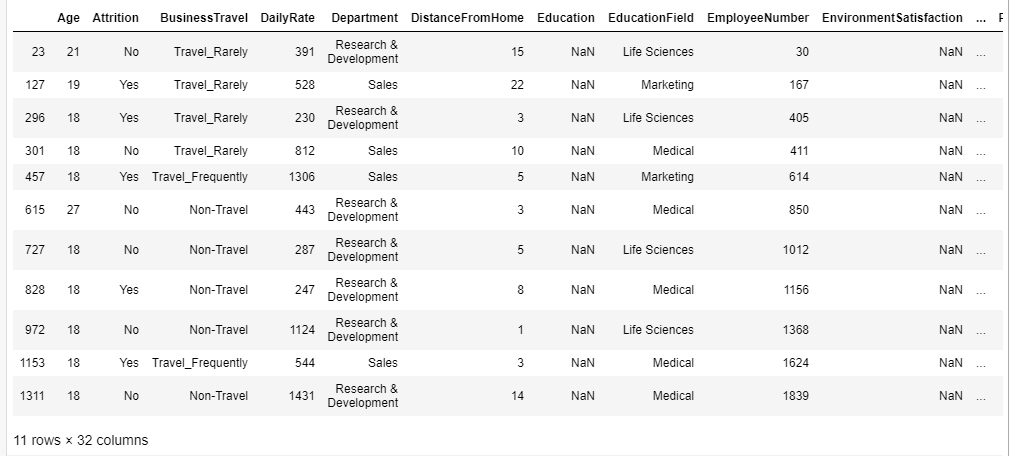


Figure 21: List of descriptive statistics for the total working years of the employee attrition dataset.



*Figure 22: A list of highest total working years employees*



*Figure 23: A list of lowest total working years employees*

1. *Relationship of Monthly Income to Attrition*

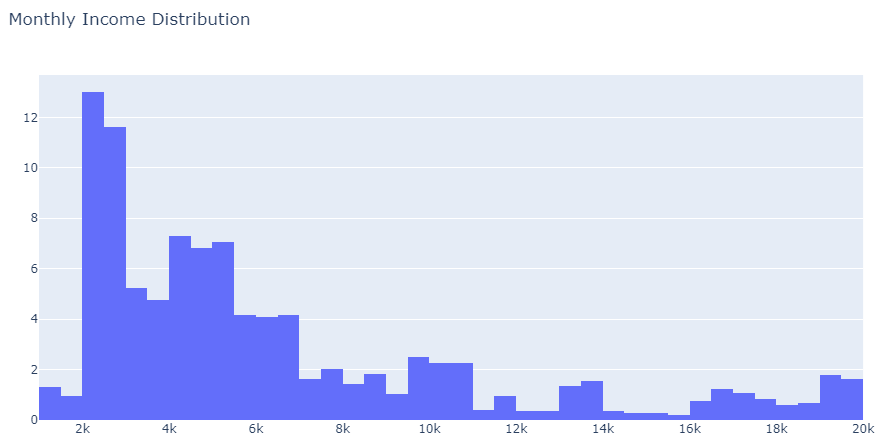


Figure 24: histogram of the monthly income of employee attrition dataset.

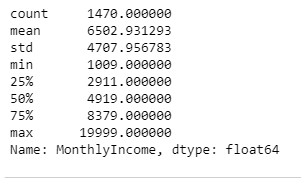
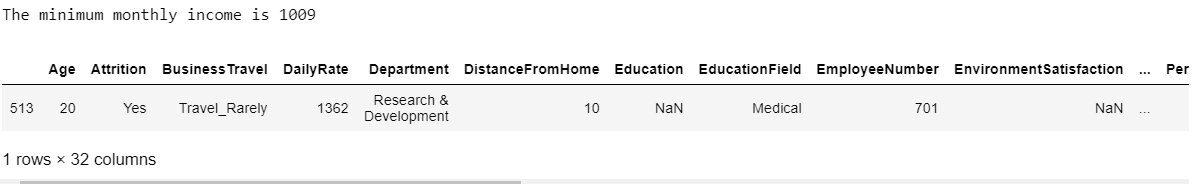
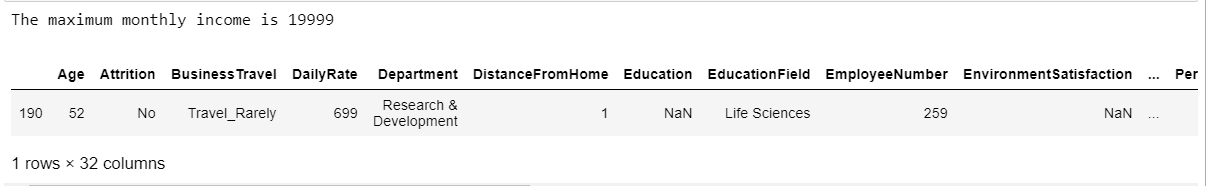
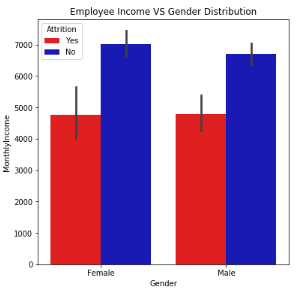


Figure 25: List of descriptive statistics for the monthly income of the employee attrition dataset.



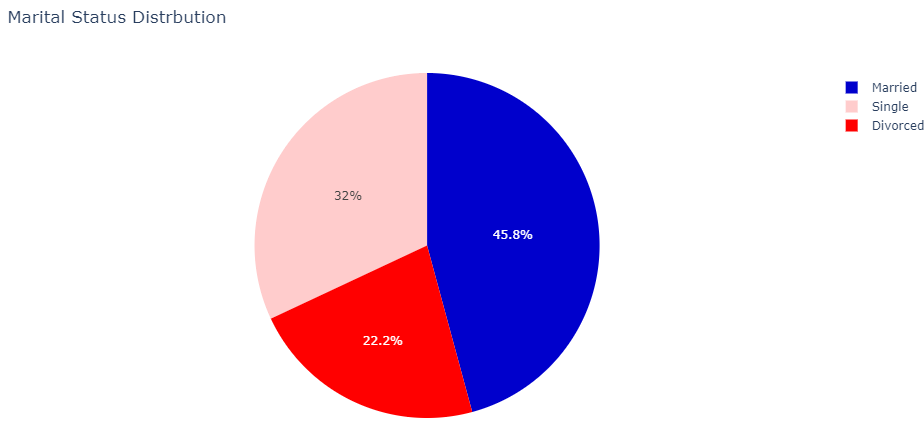
*Figure 26: A list of lowest earn a monthly income of total employees* 

*Figure 27: A list of the highest earn a monthly income of total employees*

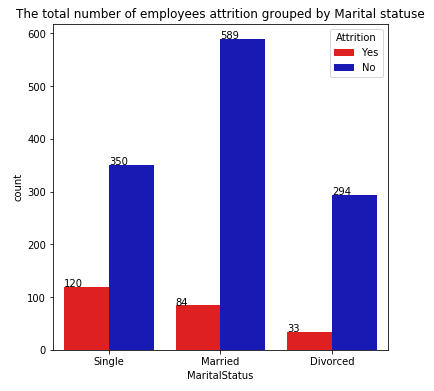


*Figure 28: A bar chart showing Monthly Income and Gender Distribution across attrition column values*

1. *Relationship of Marital Statues to Attrition*



*Figure 29 : A pie chart showing the Marital Status Percentage*



*Figure 30: A bar chart showing the record count for the "Marital Status” column across attrition column values.*

Moreover, we explored the correlation between the numeric columns except the “*EmployeeCount*” and “*StandardHours*” columns as they had constant values across all records, as observed from Figure.

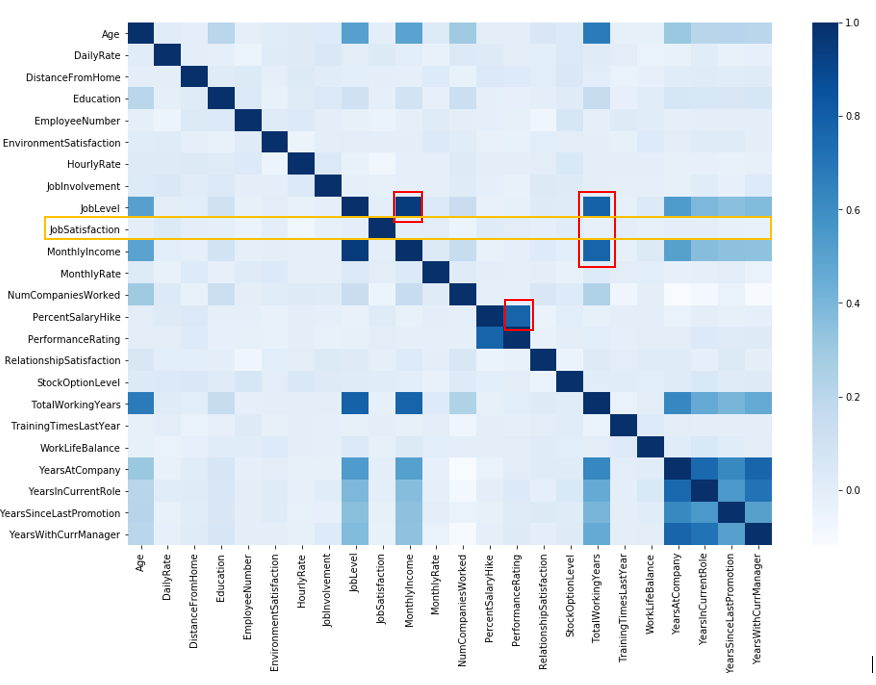


Figure 31 : A heatmap visualizing the correlation value between each numeric column in the employee attrition dataset.

# IV. Discussion

Based on the observations from Figures 14,' there are four findings. First, most of the employees with attrition falls in the age range [25-35] years old, which suggests that there is a high likelihood of attrition for younger employees in comparison to older ones. Second, a distance from home, which less than or equal 10 was covering many of the employees with attrition. Therefore, good separation distance from the work location helps to reduce attrition. Third, almost 150 employees with attrition were in job level 1 with a decreasing pattern for this number with the increase of the job level, which indicates that junior employees tend to be more prone to attrition than senior ones. Forth, most of the employees with attrition had been paid less than $5000 per month, which indicates that the more rewarding the salary, the more the chance to overcome attrition. Finally, many of the attrition cases occurred to employees with less than 10 years with the company, which reflects how the culture of the company over time is important in fighting symptoms of attrition.

For the findings from the relationship of (Over Time, Total working years, Monthly income, Marital statuses) columns, we discuss as follows. From Figure 17, the pie chart shows that the percentage of employees who worked for an overtime rate. These workers represent 71.7% of the total number, whereas only 28.3% didn't experience overtime worked. Furthermore, as what numbers are shown in Figure 18, 127 of employees who worked for overtime and got attrition. Witch is slightly higher than who didn't work for extra time and got attrition. So, we can conclude that most attrition is in the non-overtime group in terms of the proportion of employees who leave the company. However, regarding overall numbers, the highest attrition is from the overtime worked for the group. And despite the fact that overtime worked is the main reason for suffering from attrition, these people who are working for overtime usually stressed out, and that affects them in the whole aspect of their life. We didn't come with a significant result due to the issue with the majority of not attrition group.

From figure 19: the histogram distribution is right-skewed. And it shows the percentage of total working years in comparison with all the employees. Table [3] explain the percentage distribution of years sorted by largest to smallest percent in detail. Next, we visualized the total working years across the attrition state to see how many years most people try to leave the company. From Figure 20: we can see the attrition probably is highest for lower total working years. After that, regarding the statistics, Figure 21 gives us a clear glance at the interesting statistics. First, as shown, the average number of total working years is 11 years and 2 months. Second, the standers deviation is nearly 7 years and 8 months. Third, the minimum years are less than a year, and the greatest years are 40 years. Finally, we can also indicate that 25percentile of employees had 6 total experience years, while 75% of the employee had 15 total experience years. In addition, we wanted to see based on the maximum and minimum of total working years for employees whether they got attrition or not. Figure 22 and Figure 23 have answered our queries. As a result of Figure 22, which lists the highest total working years of the employee, we as clearly see that had only 2 employees who worked 40 years, what's interesting, that he did not get attrition the same as the oldest employees. Compared to the minimum of total working years in Figure 23, Over 50% have less than a year of total working years suffer from attrition, which is, more than the percentage of those who didn't suffer from attrition on the same working year and the represented 11 employees of total numbers of all employees.

|  |  |
| --- | --- |
| Total Working Years Percentage | |
| years | Percentage |
| 10-11 | 16.19% |
| 6-7 | 14% |
| 8-9 | 13.5% |
| 4-5 | 10.2% |
| 0-1 | 6.2% |
| 12-13 | 5.7% |
| 2-3 | 4.9% |
| 14-15 | 4.8% |
| 16-17 | 4.7% |
| 20-21 | 4.35% |
| 18-19 | 3.3% |
| 22-23 | 2.92% |
| 24-25 | 2.17% |
| 28-29 | 1.63% |
| 26-27 | 1.42% |
| 30-31 | 1.08% |
| 32-33 | 1.08% |
| 36-37 | 0.68% |
| 34-35 | 0.54% |
| 40-41 | 0.136% |
| 38-39 | 0.068% |

*Table 3: Percentage of total working years* of the IBM employee attrition dataset.

From figure 24: the histogram distribution of monthly income is right-skewed. Also, it shows the percentage of the total monthly income in comparison with all the employees. Table [4] explains the percentage distribution of years sorted by largest to smallest percent in detail. Next, Figure 25 gives us a clear view of the interesting statistics. First, as shown, the average number of monthly income is 6502$. Second, the standard deviation is 4707$. Third, the smallest monthly income is 1009, and the most is 1999$. Finally, we can illustrate that 25% of employees earned 2911$ per month, while 75% of employees earned 8379$ per month. Also, we wanted to find out whether the employees with the highest monthly income got attrition or not—the same as with the employee with the lowest monthly income. As a result, we found that the employee got attrition with the lowest salary, and he did not with the highest salary. So, we conclude the attrition sate decreased with an increase in the monthly income.

From Figure 28, according to monthly income, we have tried to examine the question of whether the average monthly income before attrition is different for males and females. For this query, we plot a bar chart that shows the average salary for the employee who suffers from attrition were the same for both male and female employees. The average salary of a female was slightly higher than the male employees. So, we conclude the attrition state decreased with an increase in the monthly income.

|  |  |
| --- | --- |
| Total Monthly Income Percentage | |
| income | Percentage |
| 1000$-1499$ | 1.29% |
| 1500$-1999$ | .95% |
| 2000$-2499$ | 12.99% |
| 2500$-2999$ | 11.63% |
| 3000$-3499$ | 5.23% |
| 3500$-3999$ | 4.76% |
| 4000$-4499$ | 7.27% |
| 4500$-4999$ | 6.80% |
| 5000$-5499$ | 7.07% |
| 5500$-5999$ | 4.14% |
| 6000$-6499$ | 4.08% |
| 6500$-6999$ | 4.14% |
| 7000$-7499$ | 1.63% |
| 7500$-7999$ | 2.04% |
| 8000$-8499$ | 1.42% |
| 8500$-8999$ | 1.83% |
| 9000$-9499$ | 1.02% |
| 9500$-9.999k$ | 2.51 % |
| 10k$-10.499k$ | 2.244% |
| 10.5k$-10.999k$ | 2.24% |
| 11k$-11.499k$ | .408% |
| 11.5k$-11.99k$ | .95% |
| 12k$-12.499k$ | .34% |
| 13k$-13.499k$ | 1.36% |
| 13.5k$-13.999k$ | 1.5% |
| 14k$-14.499k$ | .34% |
| 14.5k$-14.99k$ | .27% |
| 15k$-15.499k$ | .27% |
| 15.5k$-15.999k$ | .20% |
| 16k$-16.499k$ | .74% |
| 16.5k$-16.999k$ | 1.22% |
| 17k$-17.499k$ | 1.08% |
| 17.5k$-17.999k$ | .81% |
| 18k$-18.499k$ | .61% |
| 18.5k$-18.999k$ | .68 % |
| 19k$-19.499k$ | 1.76% |
| 19.5k$-19.999k$ | 1.63% |

*Table 4: Percentage of total working years* of the IBM employee attrition dataset.

Figure 29: the pie chart shows the percentage of the Marital state. It is clearly shown that nearly half of employees were married while the lowest rate of employees was divorced. The single employees represented almost 33% of the total number of employees. Also, from Figure 30: we can see, it shows that the married employees have formed the majority of attrition cases, while divorced ones formed the minority, with the singles as the median number of cases. This is quite interesting finding as divorced employees showed the highest tolerance to attrition, maybe this is due to favoring the work side over the emotional side in life after recovering from the divorce experience.

The correlation analysis in Figure 31 : shows interesting findings. First, there is a high positive correlation between the “TotalWorkingYears” column and the “JobLevel” and “MonthlyIncome”, which reflects a sort of fairness in promoting and paying people in the company based on their experience level. Second, there was a high positive correlation between “PerformanceRating” and “PercentSalaryHike” columns, which again confirms that the increase in salary is based on the increase in the performance level. Third, the “JobSatisfaction” column does not have any correlation with the remainder of the numeric columns, which is somehow unexpected as it would be reasonable to have it increased with the increase in “MonthlyIncome” or “JobLevel” columns. The latest finding suggests either an issue in estimating employee satisfaction (i.e., the way it was calculated) or missing important aspects that might reflect the employee’s satisfaction in the dataset.

# References

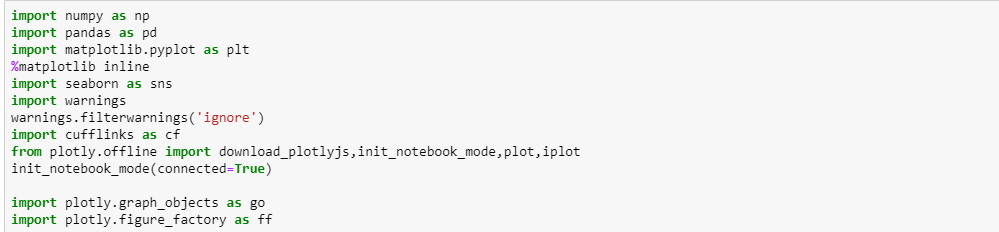
[1] T. Attri, "Why an Employee Leaves: Predicting using Data Mining Techniques," Dublin, National College of Ireland, 2018.

[2] S. S. Alduayj and K. Rajpoot, "Predicting Employee Attrition using Machine Learning," in *2018 International Conference on Innovations in Information Technology (IIT)*, 2018, pp. 93-98.

# Appendix

**Appendix A**

Importing the required dataset and reading the dataset

**

*Appendix A : Import required packages*

*Appendix A.1 : Reading the dataset*

**Appendix B**

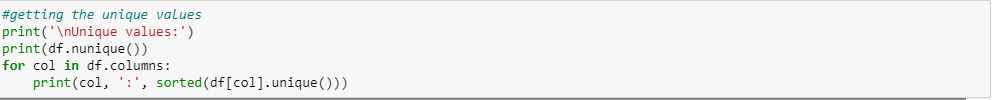
Exploring the dataset



*Appendix B.1 : Reading the first 10 rows*



*Appendix B.2 : Shape of the dataset*

 *Appendix B.3 : list the columns names*

*Appendix B.4 : getting the unique values*



*Appendix B.5: getting the columns*

**Appendix C**

Missing values



*Appendix c.1 : checking the missing values*



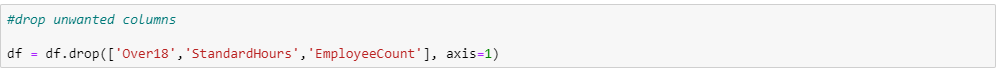
*Appendix c.2 : checking the total number of values in each columns*



*Appendix c.3 : getting a Boolean result of not having a result*

**Appendix D**

Dropping the columns



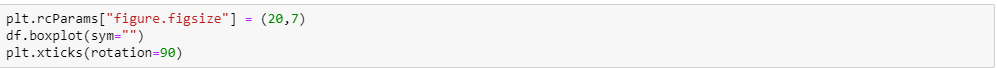
*Appendix D : Dropping columns*

**Appendix E**

Descriptive statistics



*Appendix E.1 : getting descriptive statistics of all numeric features*



*Appendix E. 2: Visualizes the statistics of all numeric features*

**Appendix F**

Data Visualization



*Appendix F. 1: Visualizes the distribution of all numeric features*

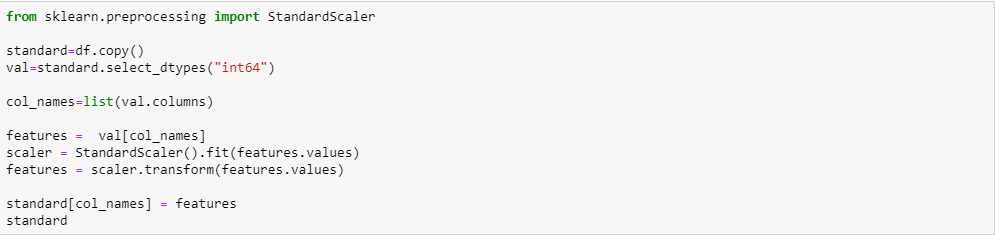


*Appendix F. 2: Visualizes the pie chart of the attrition*

Mapping the dataset



Normalize the dataset



|  |  |  |
| --- | --- | --- |
|  |  |  |
| ***“To win in the marketplace, you must first win in the workplace.”***  ***-Doug Conant*** |
|  |