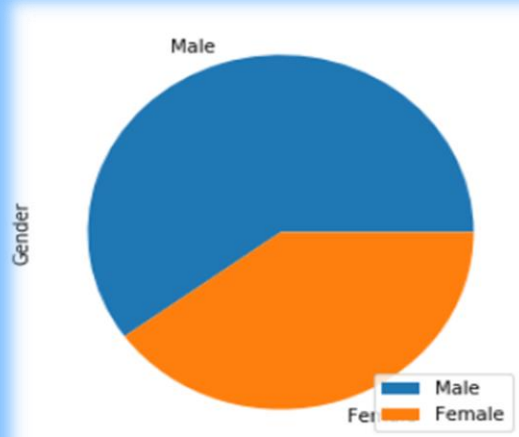
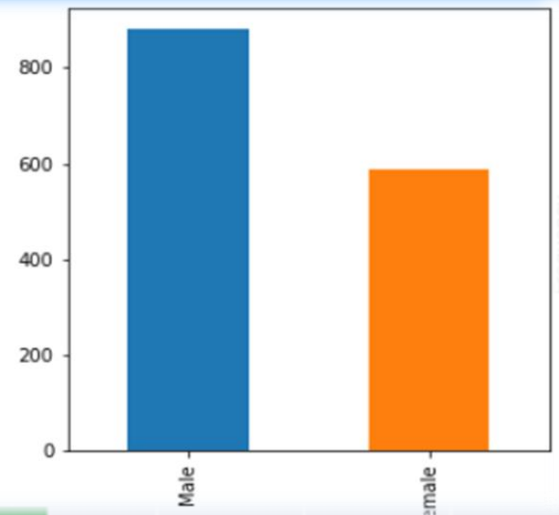
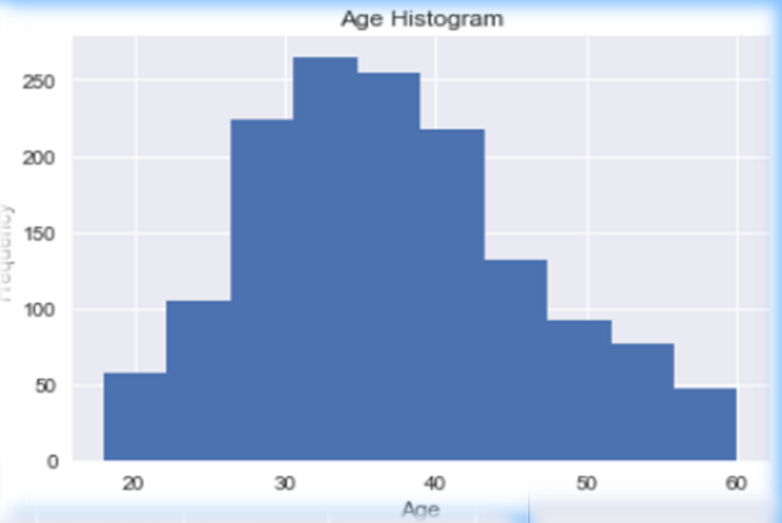
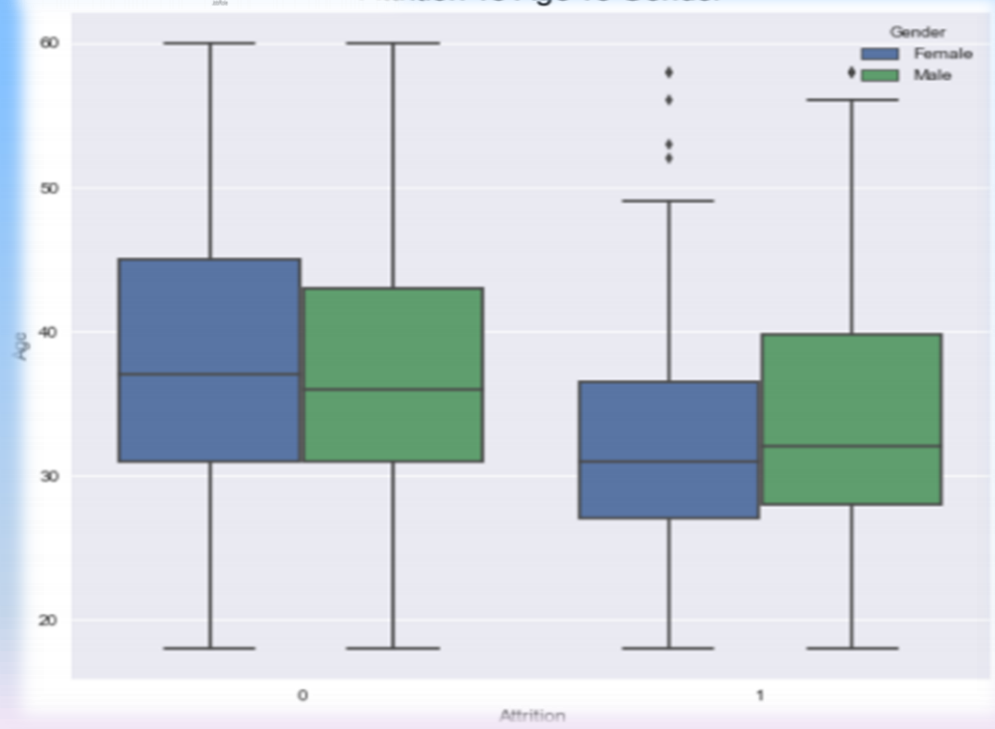
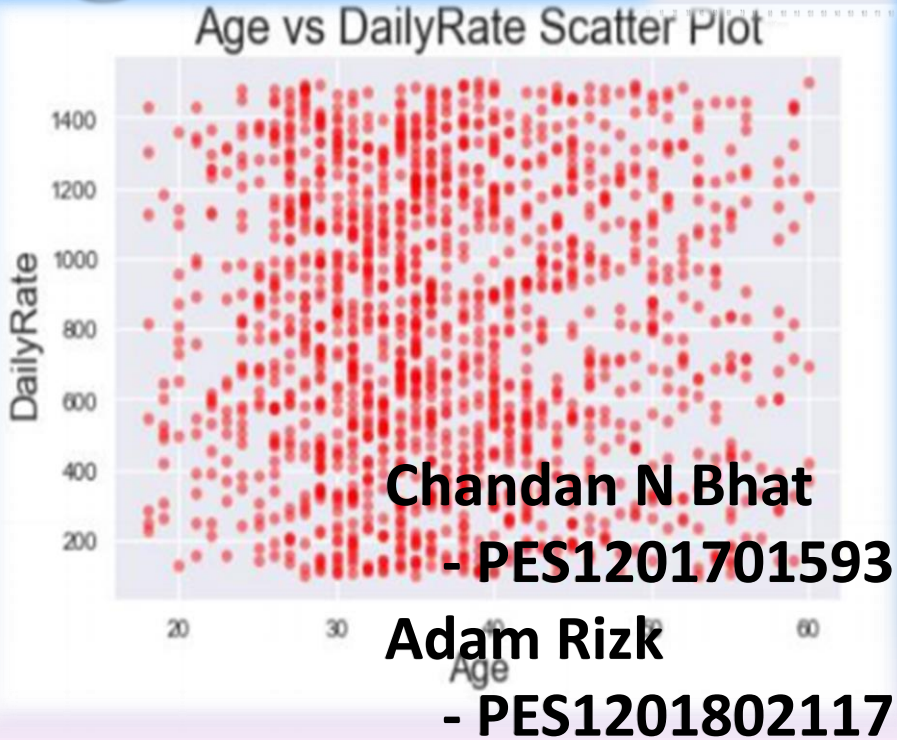






IDS Assignment





Chandan N Bhat
- PES1201701593
Adam Rizk
- PES1201802117

Aim of this Assignment

- We are studying a data set which includes detailed information about current and former employees of the IBM corporation in spreadsheet form.
- Our aim in this assignment is to find insights that could potentially lead to improved functioning in IBM's workforce.
- We have analyzed IBM employees with the aim of gaining further information about two key issues effecting IBM: attrition and employee productivity.

- In order to carry out our goals, we relied on various data visualization techniques such that we could discern the relationships different employee characteristics have with attrition and productivity rates.
- For instance, we used box plots to find that job requirements such as business travel and overtime significantly increase the risk of attrition, and that women are more sensitive than men to these requirements, leading to an increased risk of attrition.
- By the end we try obtain suggestions on how to reduce these friction points among the workplace to decrease attrition and increase productivity.

About Our Dataset

- Our data set was collected from Kaggle, and it is information about IBM employees collected and compiled from its human resources department.
- It is a fairly large dataset, 1471 rows and 31 columns.
- Each row corresponds with one IBM employee, and each column in our spreadsheet characterizes the given employee in some way.
- Importantly, this our Data Set includes both current employees, and employees who have already undergone attrition.

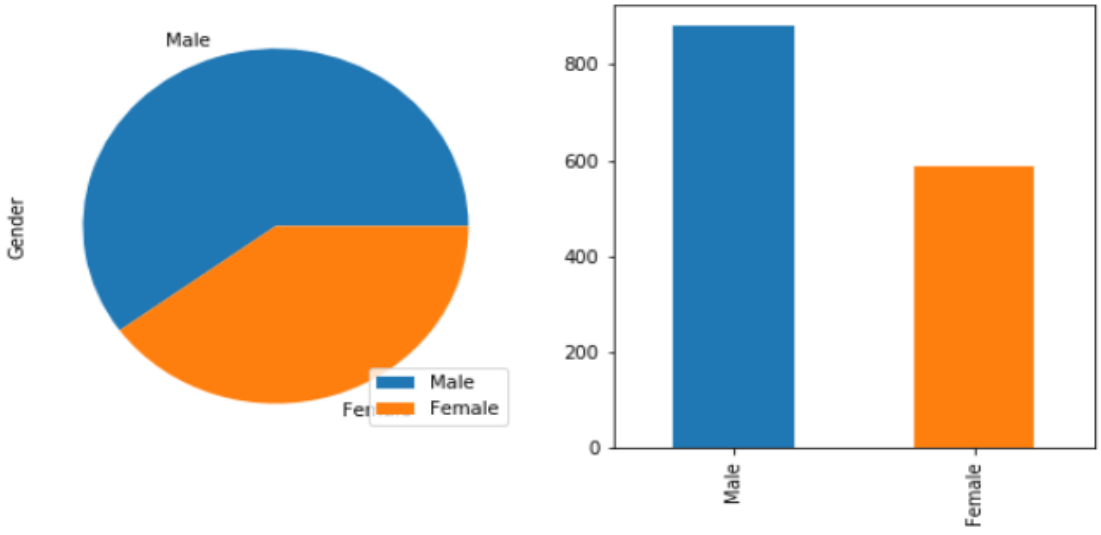
Processing(Data Cleaning)

- Our data is already relatively clean, so our data cleaning only consists of removing extraneous columns, relabeling the values of certain columns, and removing outliers.
- A cursory look at our dataframe reveals that all the cells in column Over18 have the value ‘Y’, all the cells in column EmployeeCount have a value of ‘1’, all the cells in columns StandardHours have a value of ‘80’, and that the column EmployeeNumber contains arbitrary numbers in ascending order.

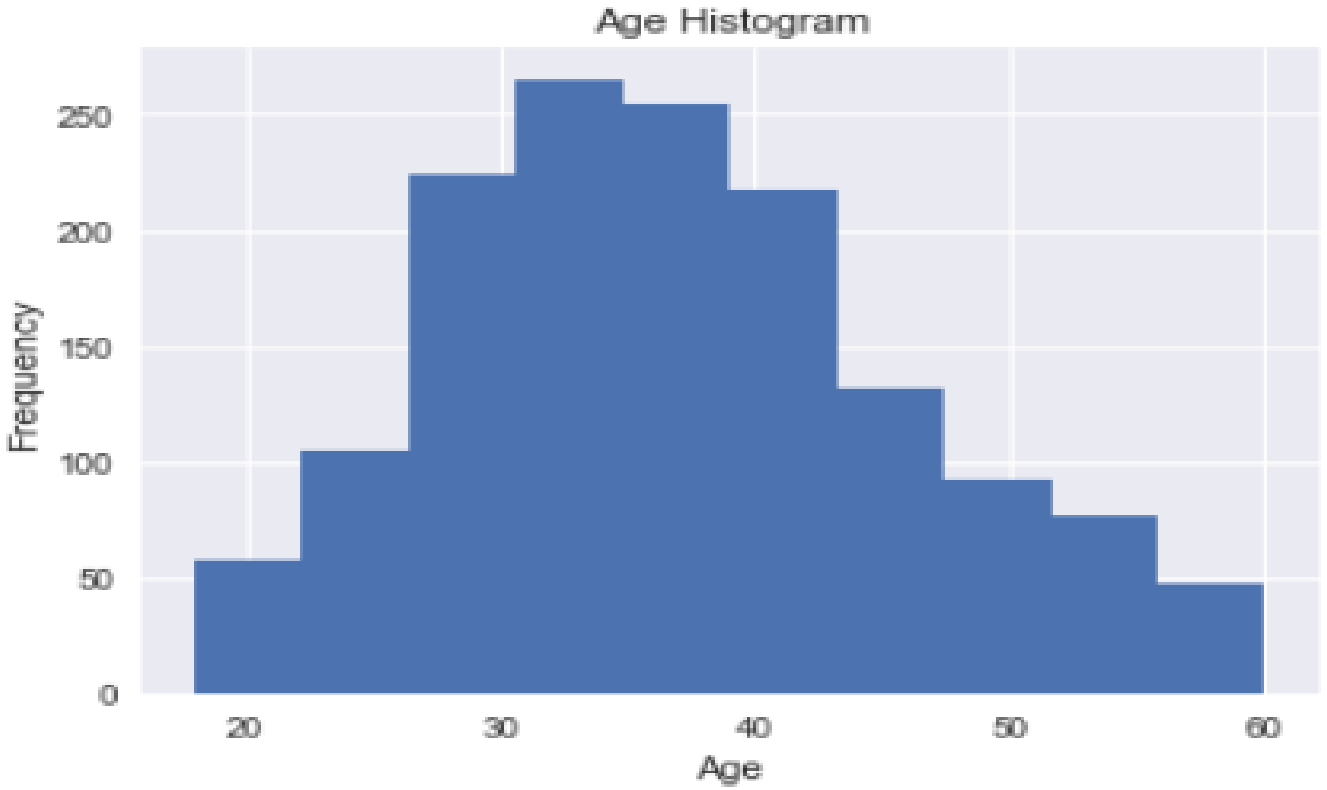
- Our next step is to relabel the values for certain columns.
- Some columns, such as Education are labeled numerically. This is not ideal as it makes the dataframe unreadable.
- We relabel these numerical values with High School, Diploma, Bachelors, Masters, and Doctorate. We obtained these label values from Kaggle.
- Finally we remove outliers from our dataframe. One common definition of outlier is a value which is 1.5 times the inter quartile range above or below the first and third quarter of the data
- We used this definition of outlier and a python script to remove outliers from out dataframe.
- We replaced each outlier with the median of the corresponding column.

Exploratory Analysis

- Here, we will use data visualization techniques to form insights about IBM employees with regard to demographics, attrition, and productivity.

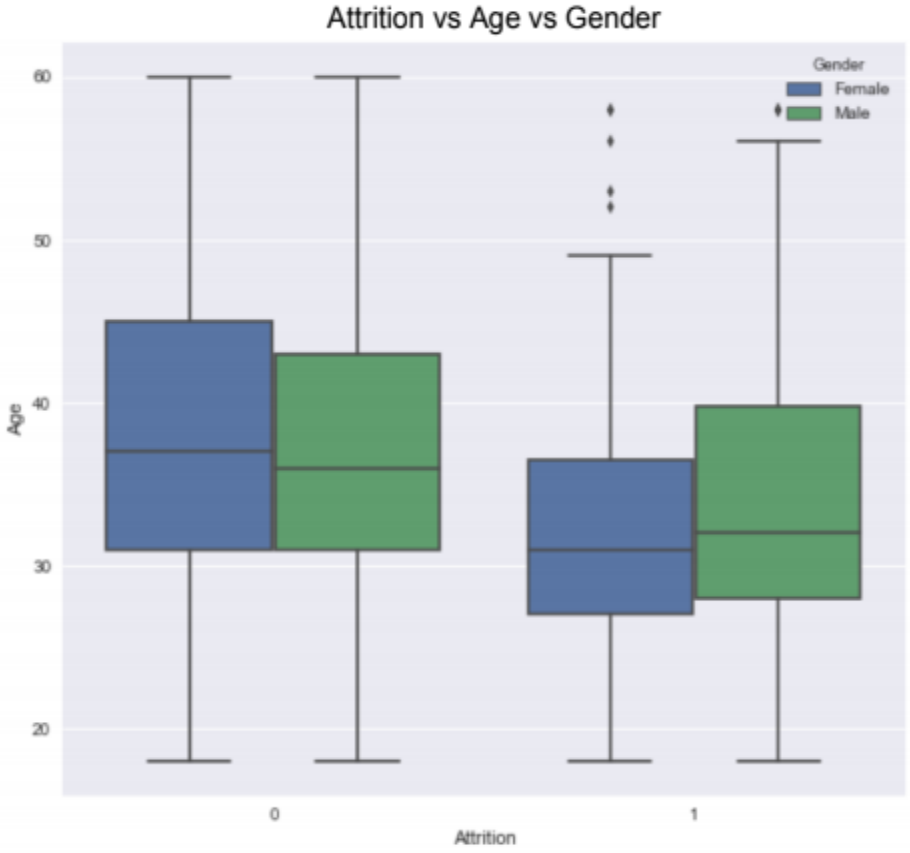


- This graphs shows us the count of Male and Female Employees



- The above histogram comes out to be almost a unimodal Normal curve, although it is slightly skewed to the right.
- We see that a large number of employees fall under the age group of 30 to 45.

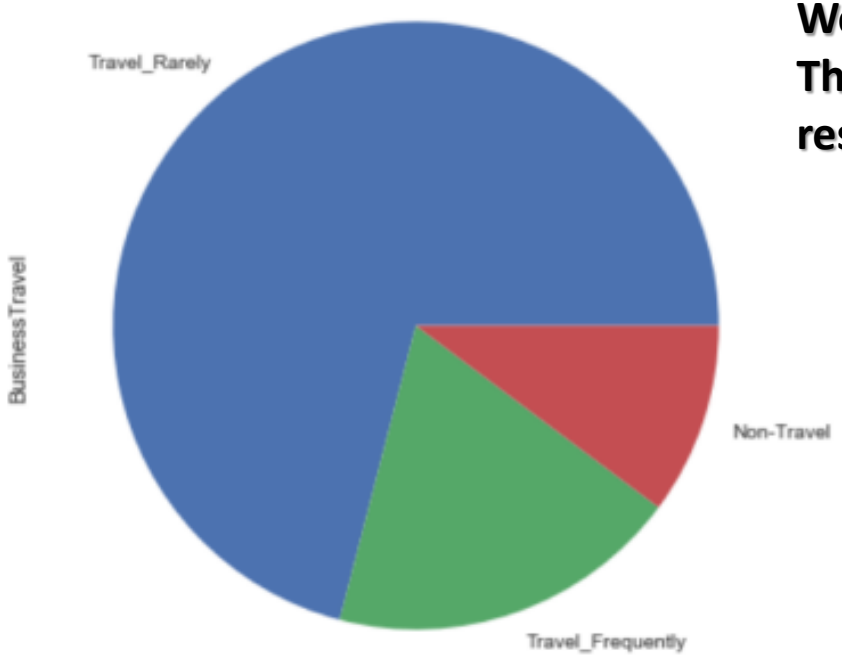
- Next, we will dive into attrition, discovering what causes it and how it can be prevented.Even though the number of employees with Attrition is relatively small compared to those without attrition, 16.12% of employees cannot be neglected as employee attrition is a loss to company, in both monetary and non-monetary terms.



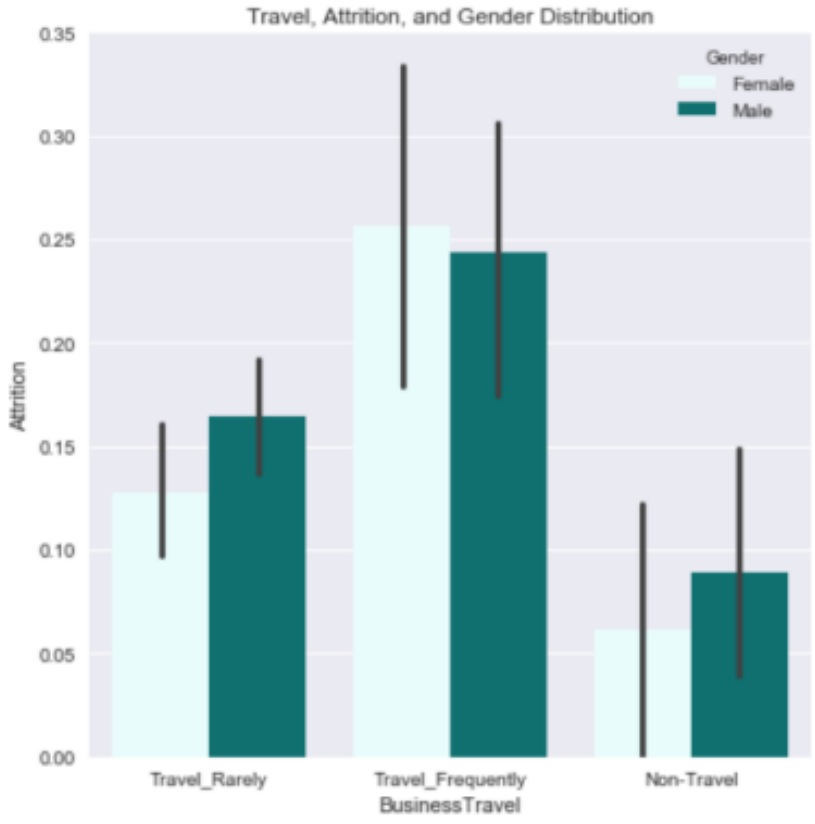
- A clear insight from the this plot is that **attrited employees are slightly younger than non-attrited employee**

- There does not appear to be significant **correlation between gender and attrition based on Age.**

Let's get a sense of how prevalent traveling is among IBM employees.



Women are more sensitive to travel than men. This may be due to reasons such as personal responsibilities.



First of all, we can clearly observe that there is a direct correlation between the frequency of an employee's business travel, and his/her chances of attrition.

Frequent business travel has a strong correlation with attrition.

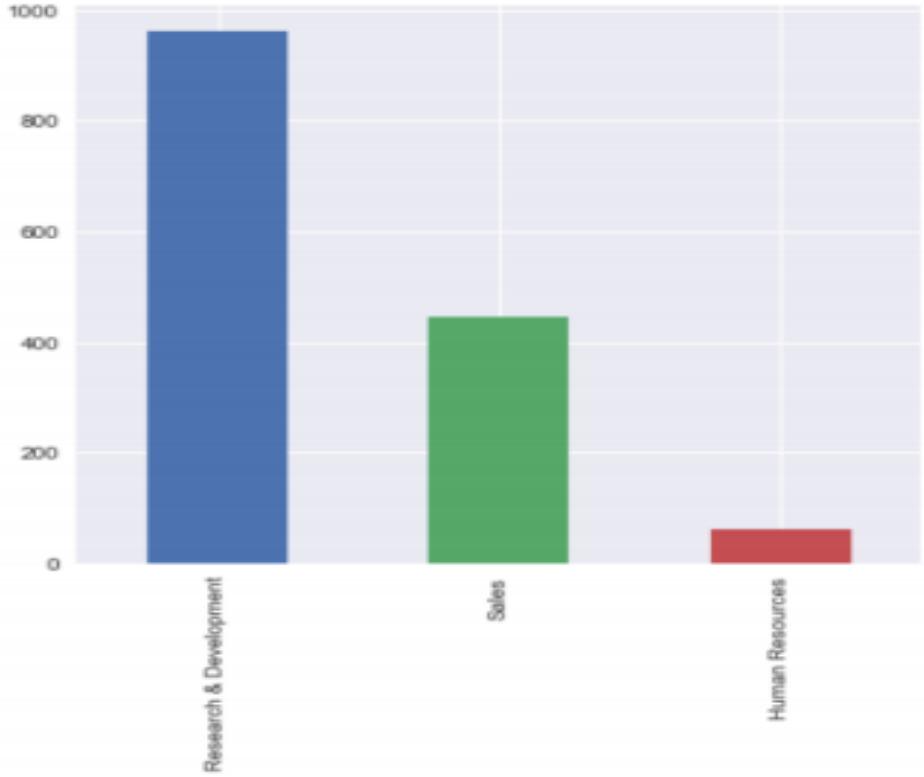
- Next, we compare how monthly income correlates with attrition in the case of men and women.

This is a significant correlation between income and attrition.

This is not surprising, as unsatisfactory pay is one of the main reasons one would expect to cause attrition among employees.

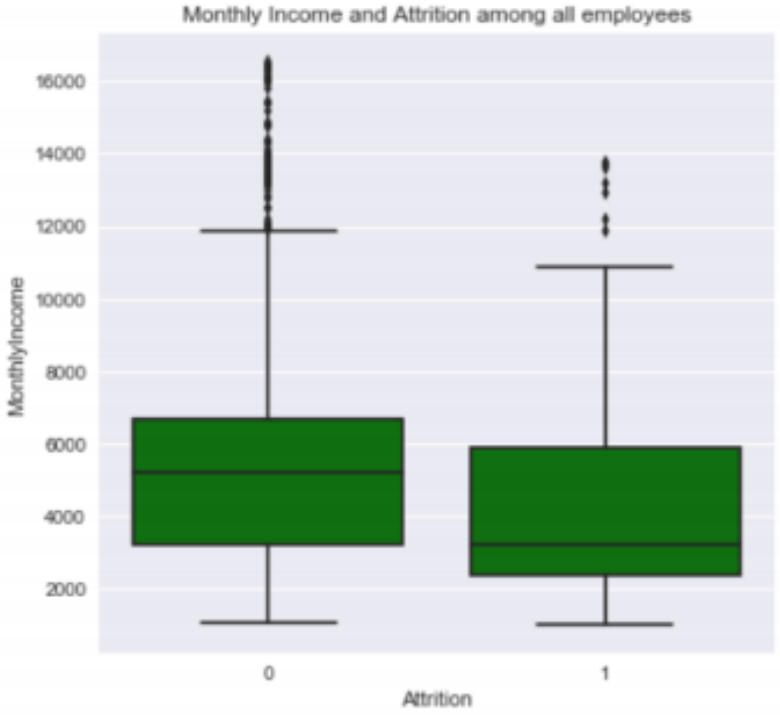
Men are potentially more sensitive to their monthly income than male employees.

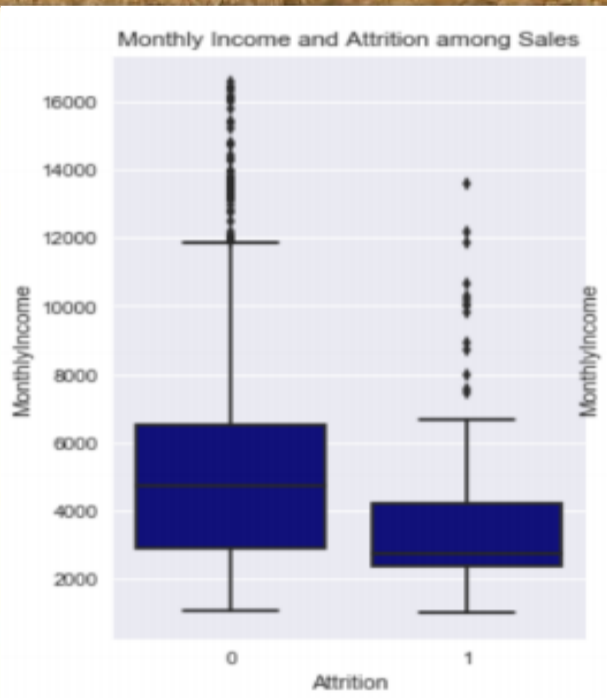
- Next, we will analyze the three different departments at IBM and gain some insights on their respective employees' attrition.

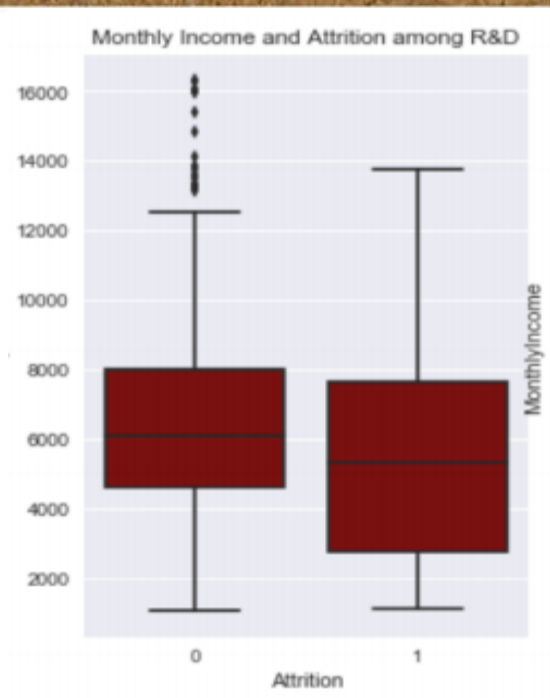


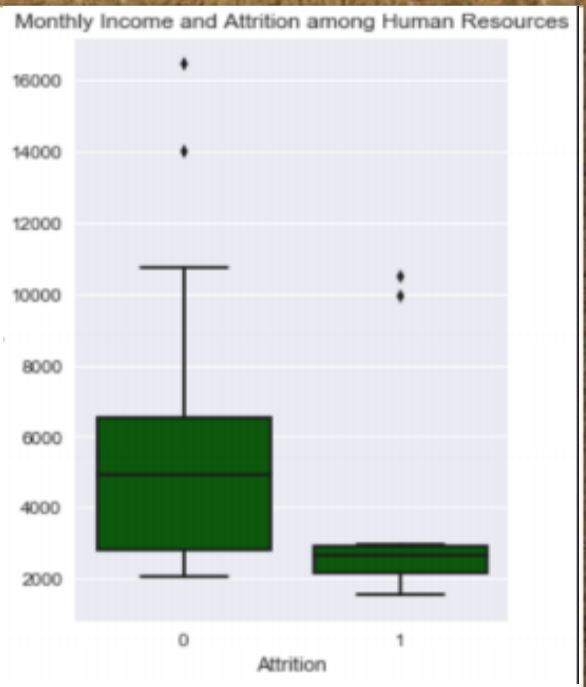
As we can see in the above bar chart, R&D is by far the largest department, with well over half the company in it.

The box plot shows that employees with attrition earn significantly less money than those who stay with the company.



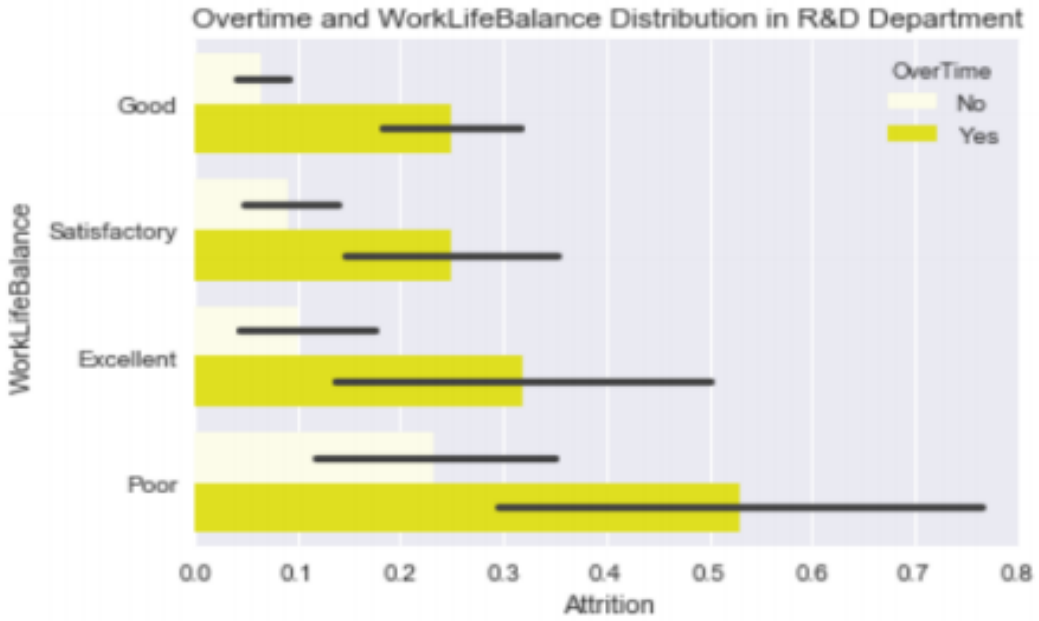






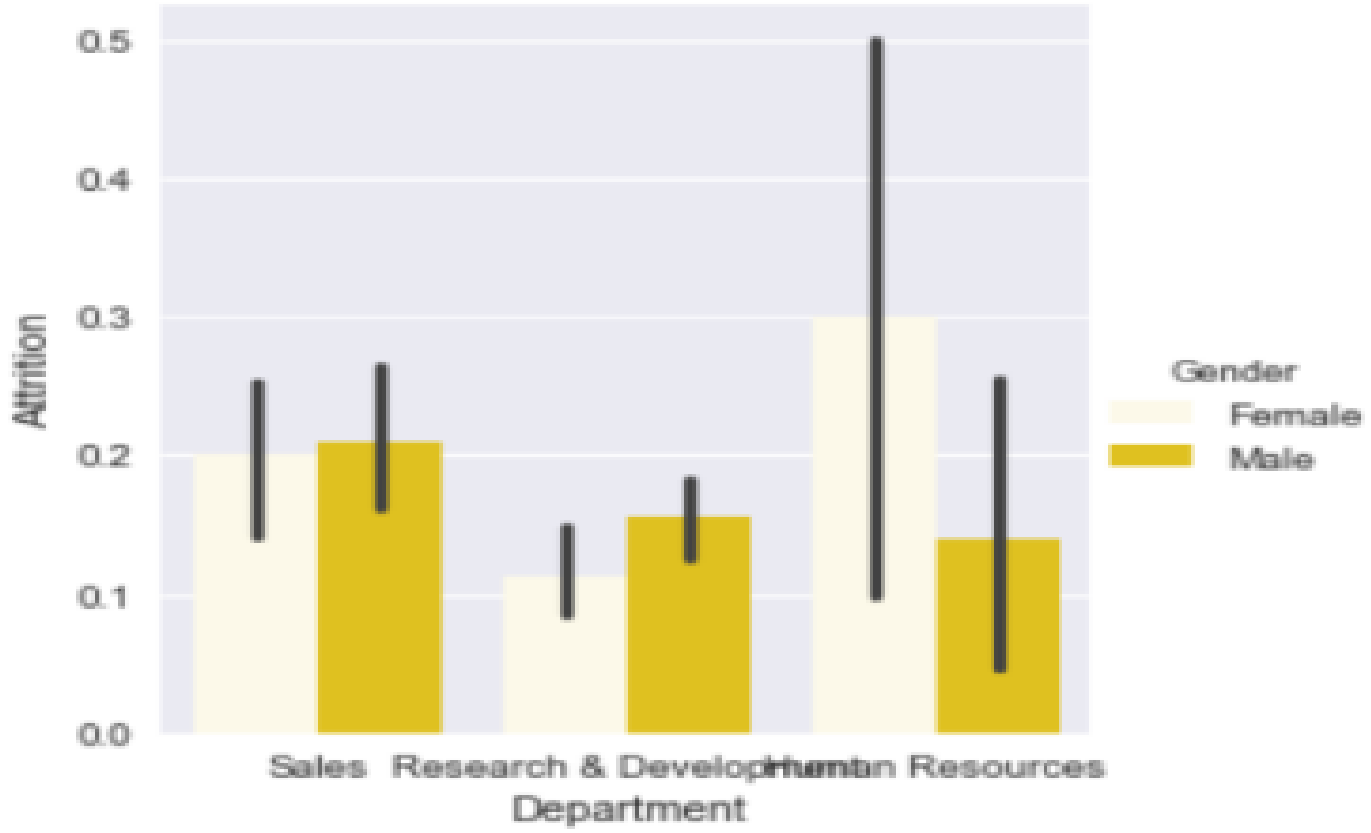
- **These three box plots again confirm the correlation between income and attrition.**
- **Human resources employees are especially sensitive to monthly income and could underpaid.**
- **This means that other factors, not the size of income, push R&D employees to attrition. This is important, as it means that the salary of R&D employees is adequate but other factors in their work environment may need improvement prevent their attrition.**

Let's dig deeper into the R&D department to see what causes its employees to have attrition, since income doesn't appear to be the main reason.



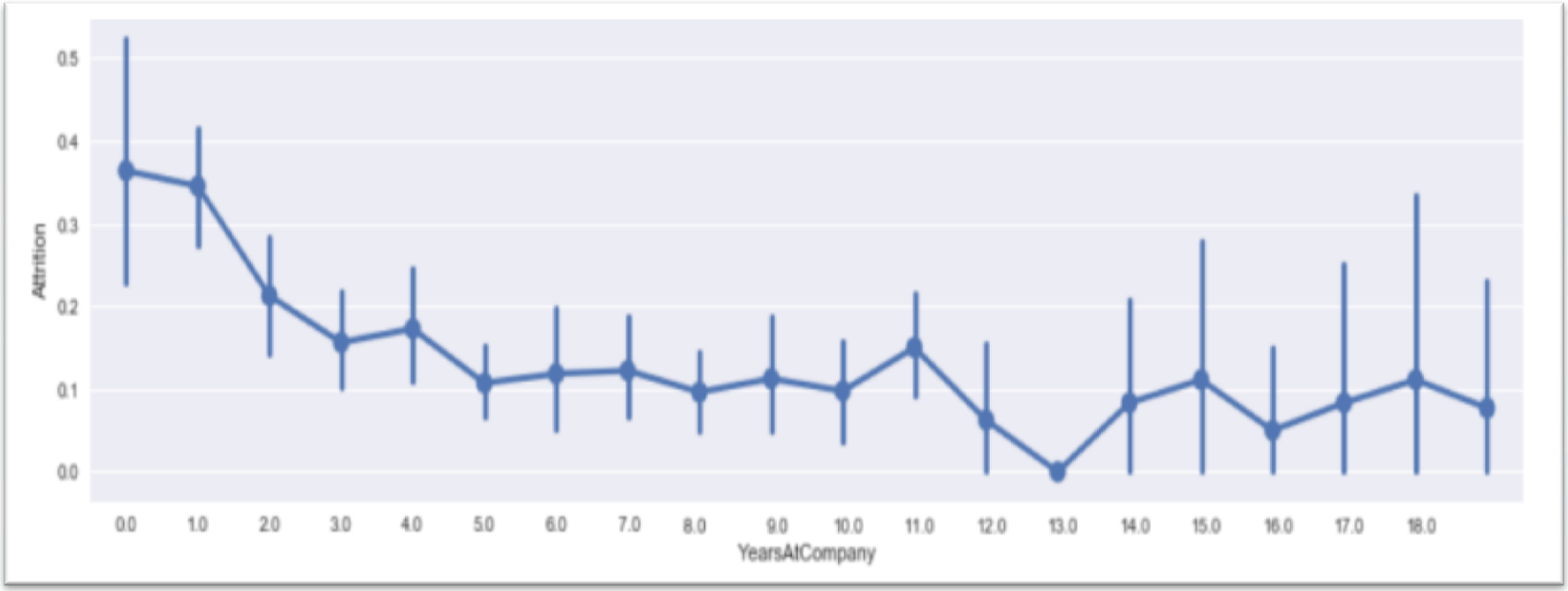
- **IBM should reduce the workload of R&D Department employees significantly to improve their work life balance.**
- **Overtime should only be given when absolutely necessary, as it is corelates very strongly with attrition.**
- **The pay of R&D employees is sufficient as seen above, but they may be overworked.**

Now, let's see how the attrition in each department corelates with each gender.

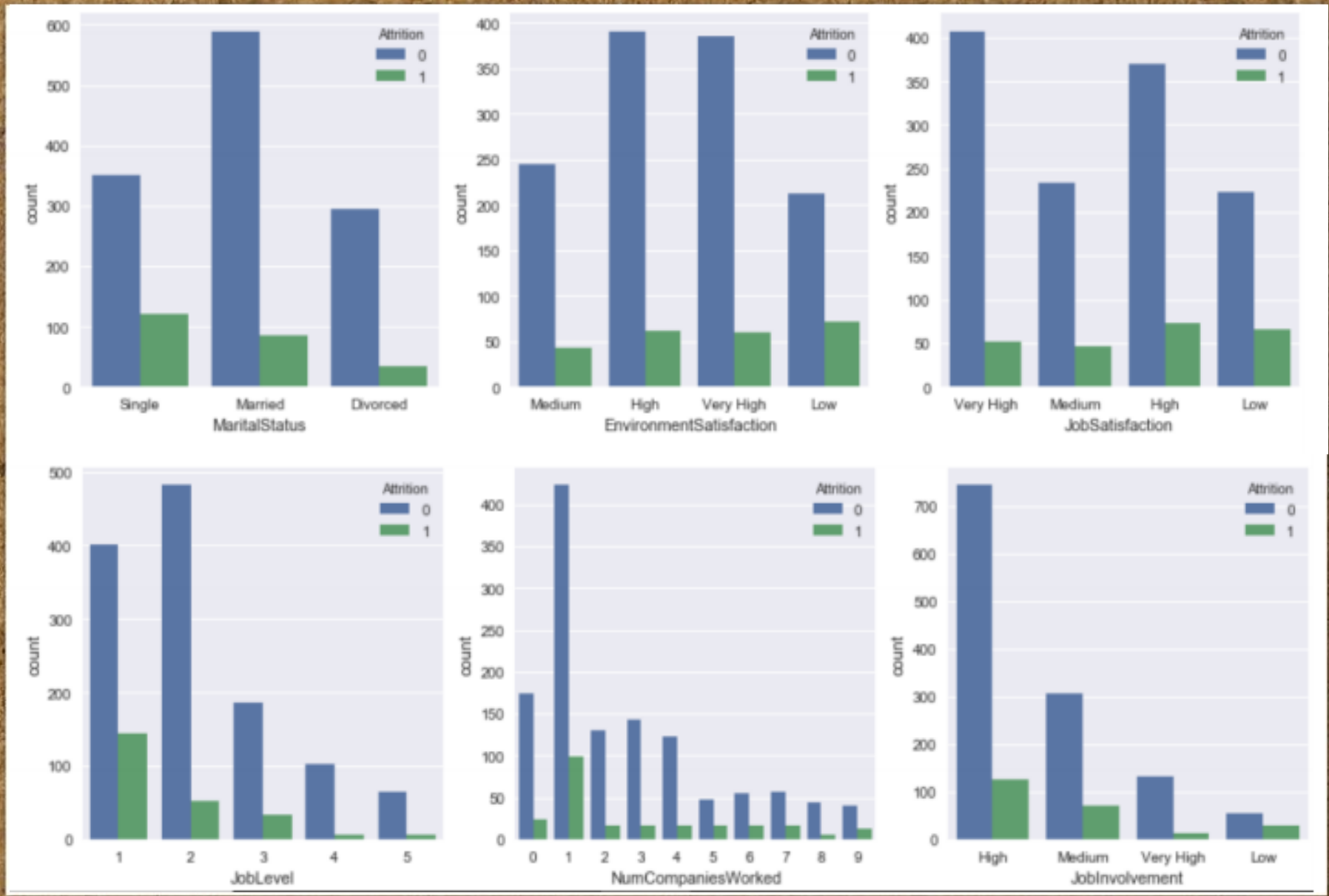


As you can see, men and women have comparable rate of attrition in Sales and R&D, but women are much more likely than men to undergo attrition in the Human Resources department.

There is also a considerable inverse correlation between the number of years an employee has worked for IBM and the chances of him/her having attrition.



It is likely that employees who have been with IBM for many years are well established in their position and are not interested in attrition as much as newer employees.


Attrition
0
1
count
MaritalStatus
Single
Married
Divorced
EnvironmentSatisfaction
Medium
High
Very High
Low
JobSatisfaction
Very High
Medium
High
Low
JobLevel
NumCompaniesWorked
JobInvolvement
High
Medium
Very High
Low

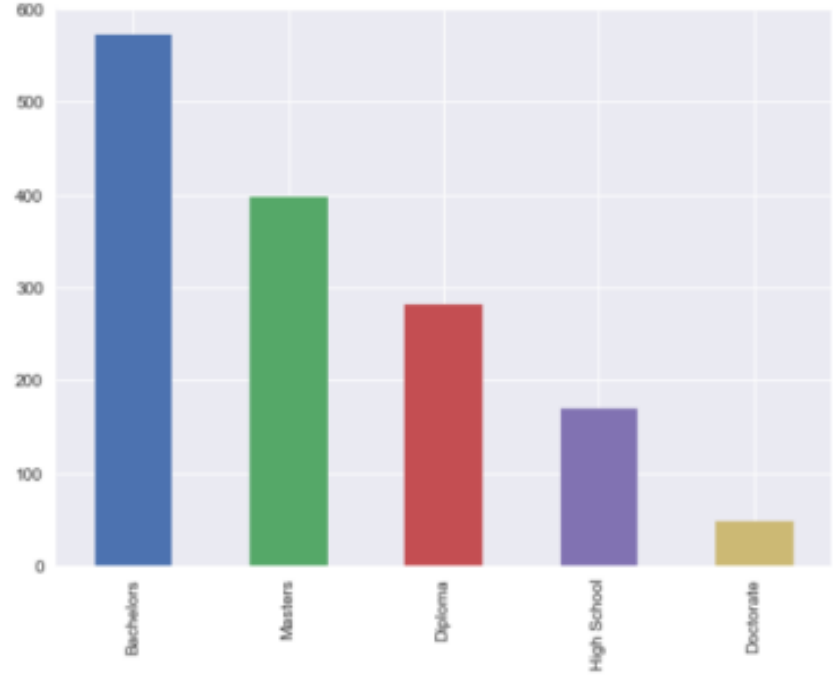
A few insights from the previous graphs :

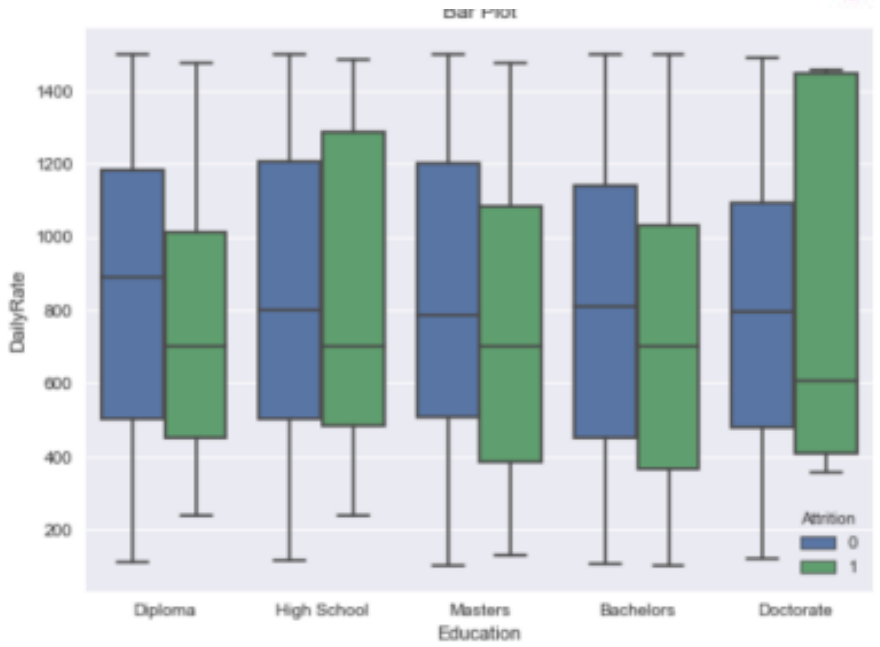
- **There is a strong negative correlation between job level and attrition. Employees with a higher job level tend to have low attrition.**
- **Single employees have higher attrition than married employees. This is likely due to the fact that single people have fewer commitments in their personal lives and can afford to take risks.**
- **IBM should improve their employees' environmental satisfaction to prevent attrition.**

There is a significant negative correlation between Environmental Satisfaction and Attrition.

- **Interestingly, employees who have worked for one company in the past have a higher attrition rate. This is likely due to inexperience and naivety among such employees, and due to the fact that they are probably younger and less likely to be married.**

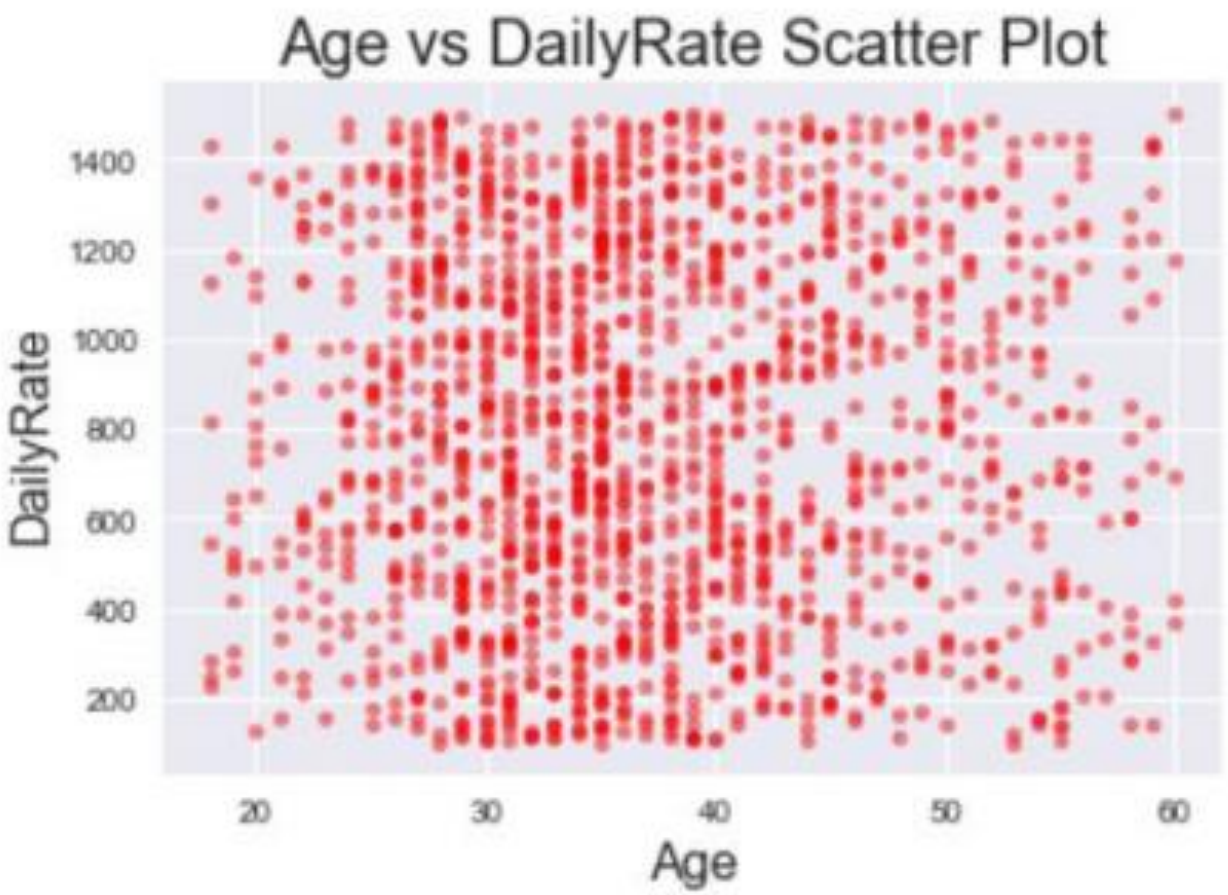
Next, we will investigate what factors correlate with employee productivity:





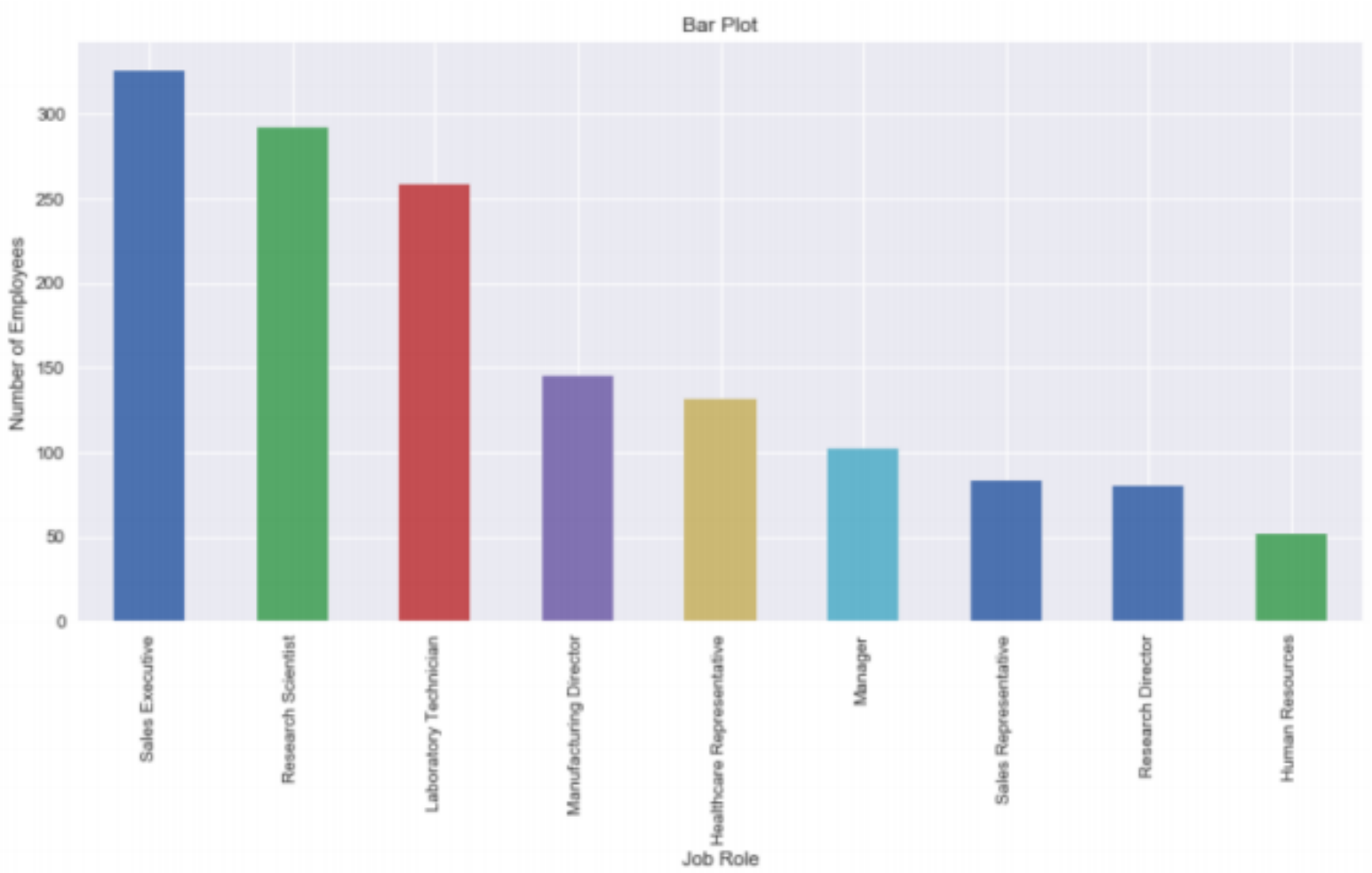
Above, we explore the relationship between education level and Daily Rate. Observing the bar chart, it is clear that Bachelors degrees are the most common education level among IBM employees, followed by Masters and Diploma.

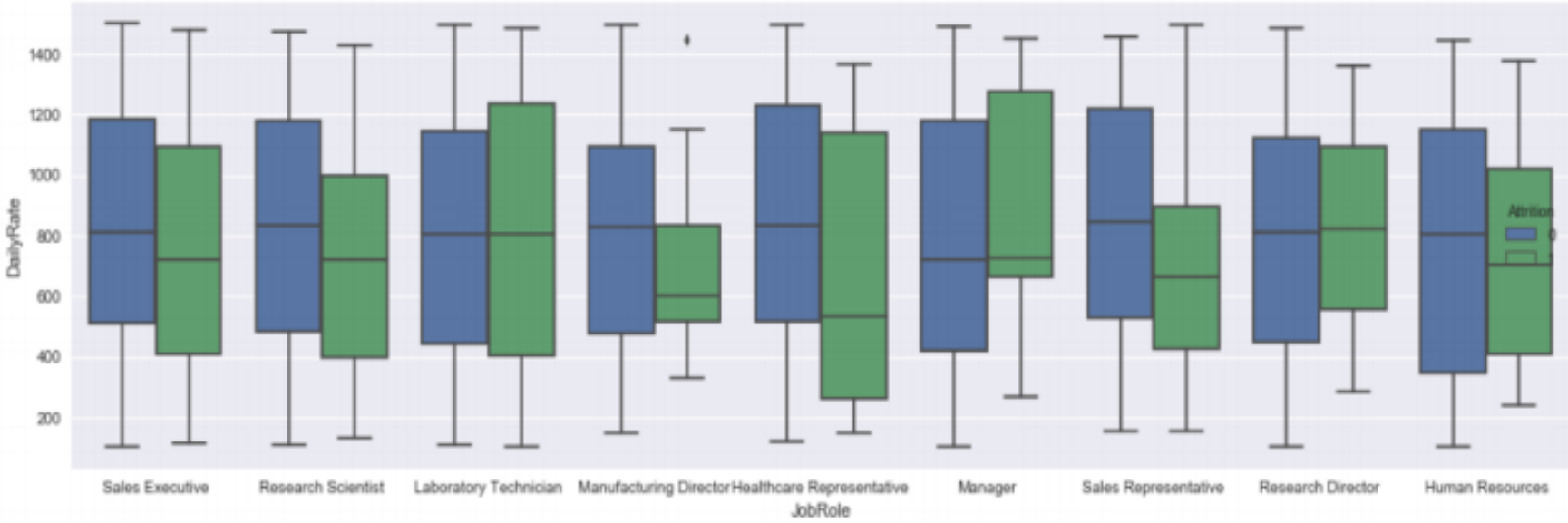
There is no significant relation between Daily Rate and Education Level. The box plots are comparable at each education level and are approximately symmetric. Therefore, hiring more educated employees is not an effective way for IBM to increase the Daily Rate of its employees. In addition, workers who have lower Daily Rate are more likely to undergo attrition.



- **We can see that there doesn't appear to be a significant correlation between Age and Daily Rate.**
- **The scatterplot appears denser around age thirty.**

Lets see whether there is correlation between Daily Rate and one's job title:



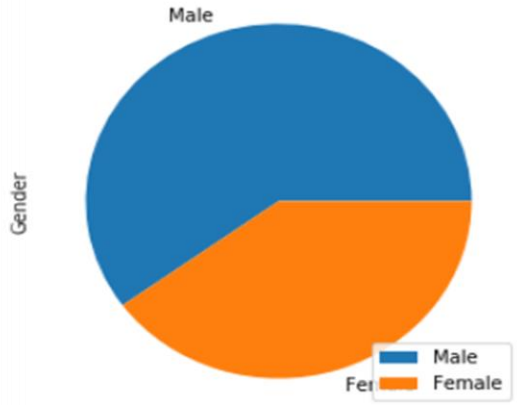


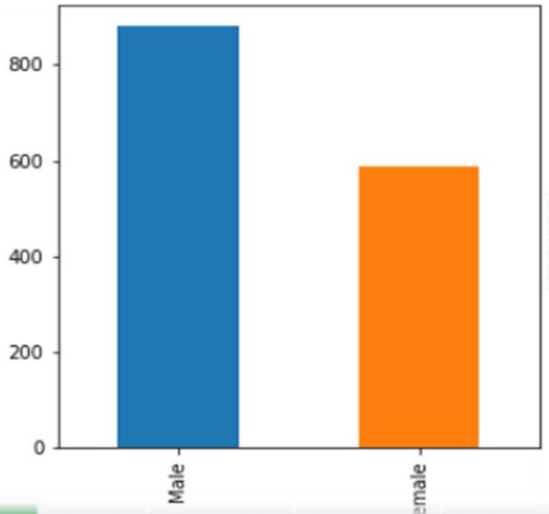
Here, we can see that there is not a significant correlation between one's Job Role and one's daily rate (among employees without attrition).

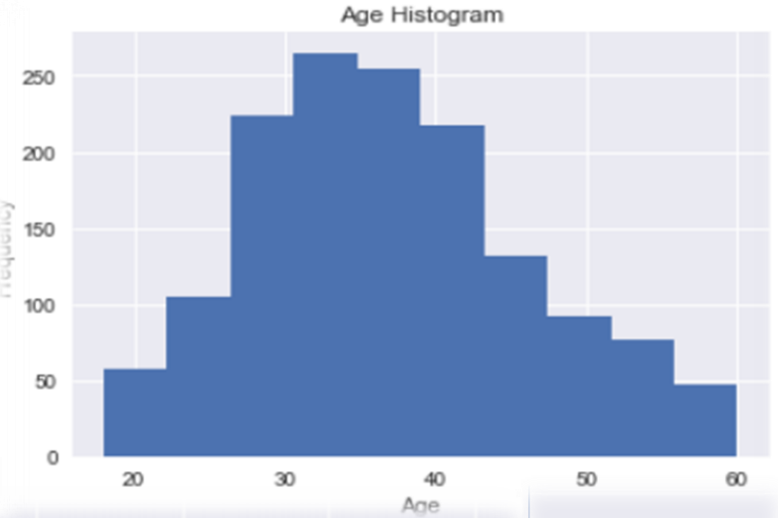
Our Takeaway From this Project

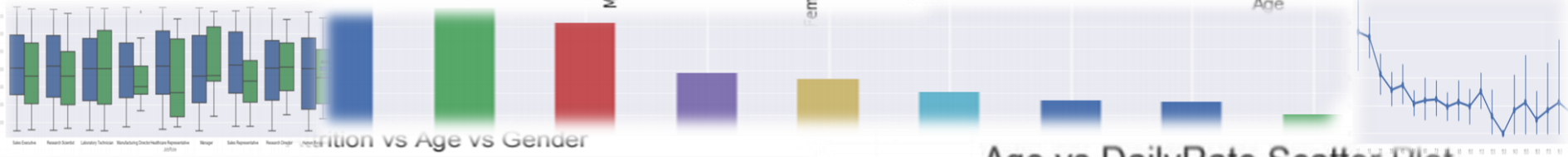
Our takeaway from this project that an organization like IBM can have numerous underlying issues which may not become evident until they are scrutinized statistically; this task is made much easier by the use of visual visualization techniques.

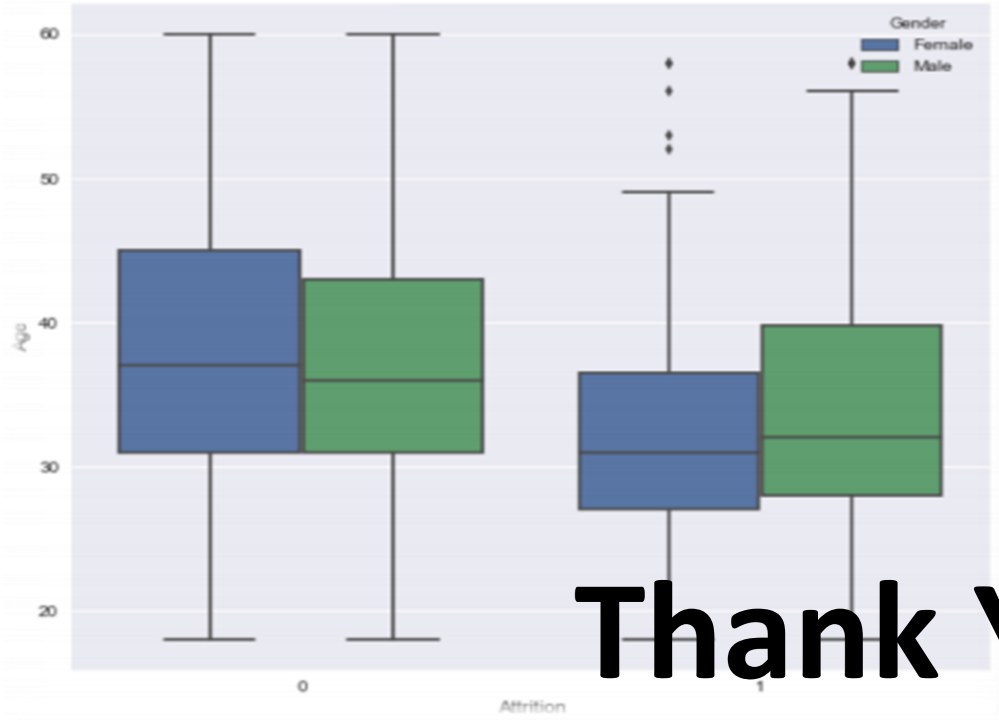
- **Even very subtle relationships and problems in the dataset become evident when visualized using the appropriate chart. For example, by using a bar chart, we discovered that women employees are potentially facing marginalized by IBM's HR department.**
- **We discovered several factors on the part of the employer that are contributing to a higher attrition rate among its employees.**
- **Using these insights, one can discover how these issues can be mitigated. The practice of adept statistical analysis with the help of computational data science techniques can make any task or organization run more efficiently.**
- **Through this assignment, we practiced our python skills along with the Pandas, Seaborn and matplotlib library, and most of all; we gained a newfound appreciation for the field of data science.**

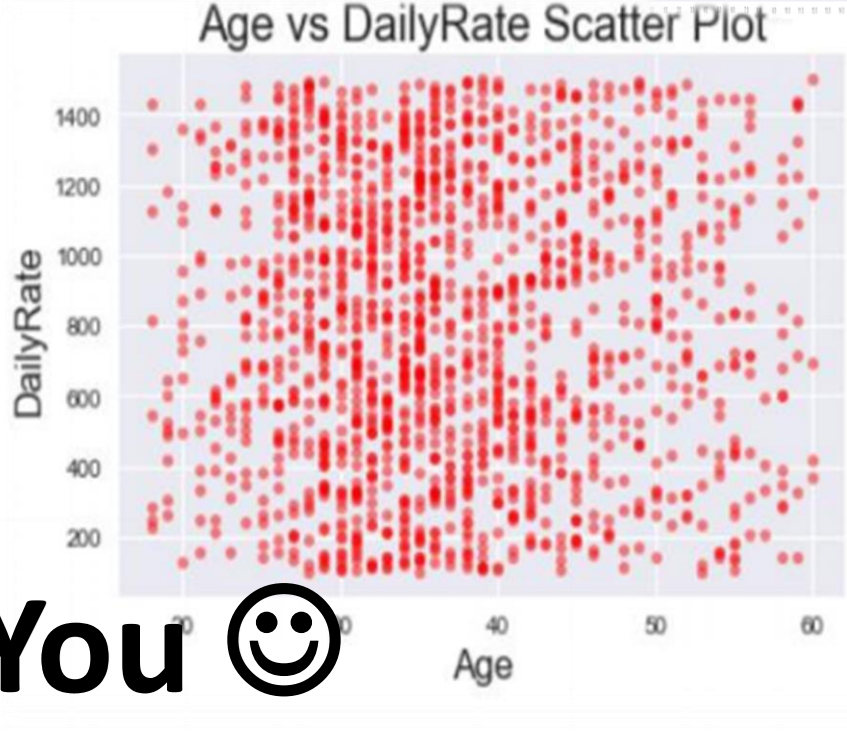

Male
Gender
Male
Female


800
600
400
200
0
Male


Age Histogram
250
200
150
100
50
0
20
30
40
50
60
Age




Gender
Female
Male
60
50
40
30
20
0
1
Attrition


Age vs DailyRate Scatter Plot
DailyRate
1400
1200
1000
800
600
400
200
20
40
50
60
Age

Thank You ☺