PES UNIVERSITY

PES University, Bangalore

UE17CS23

(Established under Karnataka Act No. 16 of 2013)

B.Tech, Sem III
Session : Aug-Dec, 2018

UE17CS203 – INTRODUCTION TO DATA SCIENCE

REPORT

EXPLORATORY ANALYSIS ON
IBM HR Analytics On Employee Attrition & Performance

DATA SET LINK :	**https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset**
TEAM MEMBERS	**NAME : Adam Rizk** **SRN : PES120802117** **EMAIL ID : adamrizk9@gmail.com** **CONTACT NO. : +91 9108274202**
	NAME : Chandan N Bhat **SRN : PES1201701593** **EMAIL ID : chandanbhat9799@gmail.com** **CONTACT NO. : +91 9632916454**

ABSTRACT

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. Attrition is the phenomenon of employees leaving the company due to being dissatisfied with their jobs; it is a significant issue for any company, and we have studied what factors corelate with attrition and what can be done to mitigate it. Similarly, we have also analyzed 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. We discovered that employees with different personal and professional characteristics respond differently to their job requirements with regards to attrition. 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. Similarly, we discovered what factors cause attrition at various departments at IBM, and which departments are more likely to be affected by certain factors. For example, we found that Sales and Human Resources employees were much more sensitive to their incomes in determining attrition compared to the Research and Development department. We then used bar charts to discover that R&D employees undergo attrition due to poor work-life balance, and not due to an unsatisfactory income like the two aforementioned departments. This report contains detailed information and visualizations of numerous other such insights; we also offer suggestions on how to reduce these friction points among the workplace to decrease attrition and increase productivity.

Data Set

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. 237 out of the 1470 employees (16.12%) in our data set have left IBM through attrition.

The two columns we will be focusing on the most in our analysis are *Attrition* and *DailyRate*. *Attrition* is a binary variable, as an employee can either have attrition or not. DailyRate is a measure of productivity for an employee on a day-by-day basis. According to Kaggle, *DailyRate* is measured as the dollar value of productivity that the employee contributes to IBM in a given day. We have been unable to determine the exact methodology IBM uses to value the productivity of its employees in dollars, however, in this analysis will assume that *DailyRate* is a sufficiently accurate measure of employees' productivity. We will rely on this column as we characterize the various factors which effect employees' productivity.

Other columns in our data set contain demographic information such as age, marital status, and gender. These are all self-explanatory. There are numerous columns that provide us with information about the employee's professional life. The data set has columns indicating how often they must travel for business; their hourly, daily, and monthly rate; how many years of experience they have, how many years they have worked for IBM, their monthly income, the number of years they have been with their current manager, education level, etc. These are all fairly self-explanatory variables and have numerical values.

There are also columns which are qualitative: such as environmental satisfaction and job satisfaction. These variables have numerical values, however, we transform them into strings during our data cleaning process. It is our opinion that numerical values are not ideal for recording variables such as one's education level and job satisfaction, and that they must be labeled more clearly. For example, the column *JobSatisfaction* has numerical values 1-4. We transform these into strings such as *low, medium*, etc. Similarly, we transform the *Education* column, which has numerical values 1-5, into strings such as *High School, Diploma*, etc. We got this information from Kaggle and incorporated these transformations to make our dataset more readable. Thus, our data set is fairly comprehensive and provides us with the ability to form numerous insights into attrition and productivity at IBM.

Introduction

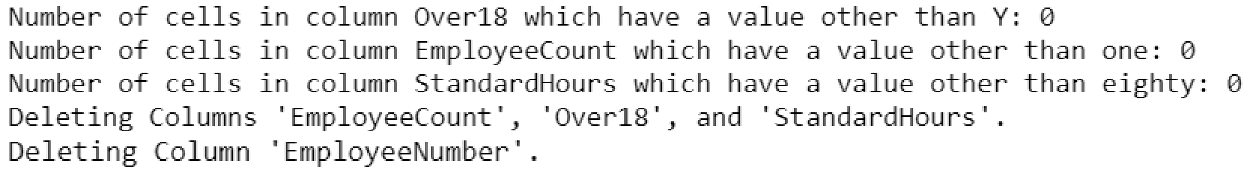
We have analyzed IBM employees with the aim of gaining further information about two key issues: attrition and employee productivity. First, we investigate how IBM can reduce attrition among its employees, and what employee characteristics corelate with attrition. Attrition is the phenomenon of employees resigning from their company for a multitude of possible reasons: unsatisfactory income, poor work-life balance, overtime, toxic work environment, and other personal and professional reasons. Minimizing attrition is essential to the overall health of a company. Attrition costs IBM a significant amount of time and money and reduces employee morale. Leaving attrition unchecked can also increase employee turnover and reduce overall productivity. For these reasons we have studied various factors that relate to attrition and offer recommendations which could potentially minimize this problem.

In addition to this, we have also analyzed the productivity of employees, and gathered information on what characteristics corelate with poor productivity, and what action IBM can take to mitigate this. We have also studied issues with IBM causes low productivity among its employees and offer suggestions to potentially mitigate this issue.

Our first step in analyzing these issues was to get a sense of the demographics of IBM employees. We used pie charts, bar plots, and histograms to find the distribution of age and gender in our sample. Our next step in order to choose which variables we should explore was to find the mean value of each column for employees with attrition, and those without attrition. We accomplished that by a Python script which grouped our dataset by attrition, and gave us the means of each column in tabular form. From there, we observed which columns have different values for attrition and non-attrition, and implemented those columns in various charts to discover various insights regarding attrition, productivity, income, etc. Using this technique, we can answer questions such as what characteristics of employees correspond to attrition, and what characteristics by the employer correspond to attrition. We also have the ability to explore which combinations of employee characteristics and work place requirements lead to the best or worse productivity.

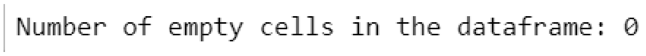
Processing (Data Cleaning)

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. We run a python script to confirm our suspicions and then delete these extraneous columns.



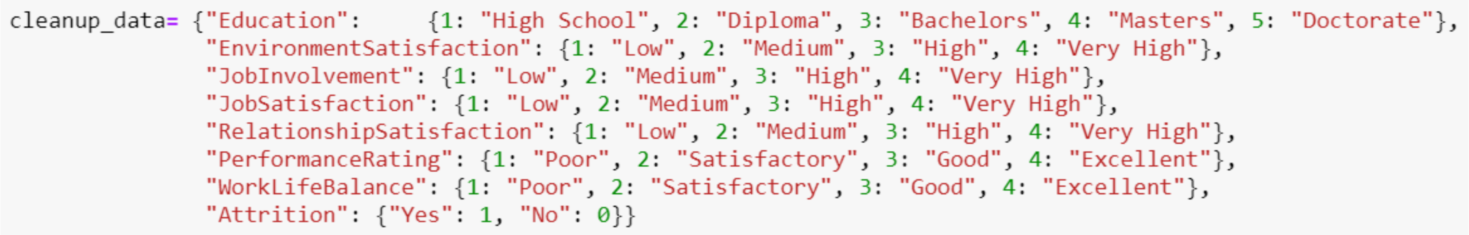
```
Number of cells in column Over18 which have a value other than Y: 0
Number of cells in column EmployeeCount which have a value other than one: 0
Number of cells in column StandardHours which have a value other than eighty: 0
Deleting Columns 'EmployeeCount', 'Over18', and 'StandardHours'.
Deleting Column 'EmployeeNumber'.
```

We then run a python script to check whether there are any empty cells in our dataframe.



```
Number of empty cells in the dataframe: 0
```

We determine that there are no empty cells, so no further action is necessary in this regard. 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.



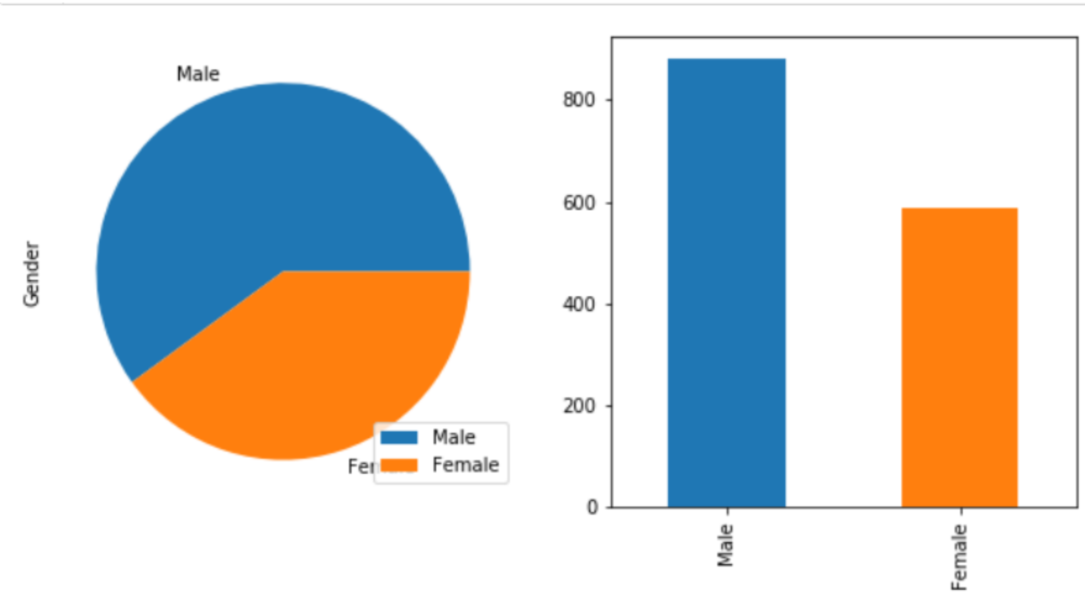
```
cleanup_data= {"Education":     {1: "High School", 2: "Diploma", 3: "Bachelors", 4: "Masters", 5: "Doctorate"},
               "EnvironmentSatisfaction": {1: "Low", 2: "Medium", 3: "High", 4: "Very High"},
               "JobInvolvement": {1: "Low", 2: "Medium", 3: "High", 4: "Very High"},
               "JobSatisfaction": {1: "Low", 2: "Medium", 3: "High", 4: "Very High"},
               "RelationshipSatisfaction": {1: "Low", 2: "Medium", 3: "High", 4: "Very High"},
               "PerformanceRating": {1: "Poor", 2: "Satisfactory", 3: "Good", 4: "Excellent"},
               "WorkLifeBalance": {1: "Poor", 2: "Satisfactory", 3: "Good", 4: "Excellent"},
               "Attrition": {"Yes": 1, "No": 0}}
```

Now that we have relabeled some of our columns as indicated in the code shown above, our dataframe is significantly easier to read and allows us to label our charts appropriately. Note that we have relabeled *Attrition* from values *Yes* and *No* to 1 and 0. We do this because some of our charts require numerical variables. Since attrition is the most important variable in our analysis, we have made it numerical.

The only thing left in our data cleaning is to remove outliers from our dataframe. One common definition of outlier is a value which is 1.5 time the interquartile range above or below the first and third quarter of the data. We used the describe() command in Pandas to obtain statistical values of each column in tabular form. We use this to gain a preliminary idea of what the distribution of each column is, and to determine which columns have outliers. 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.

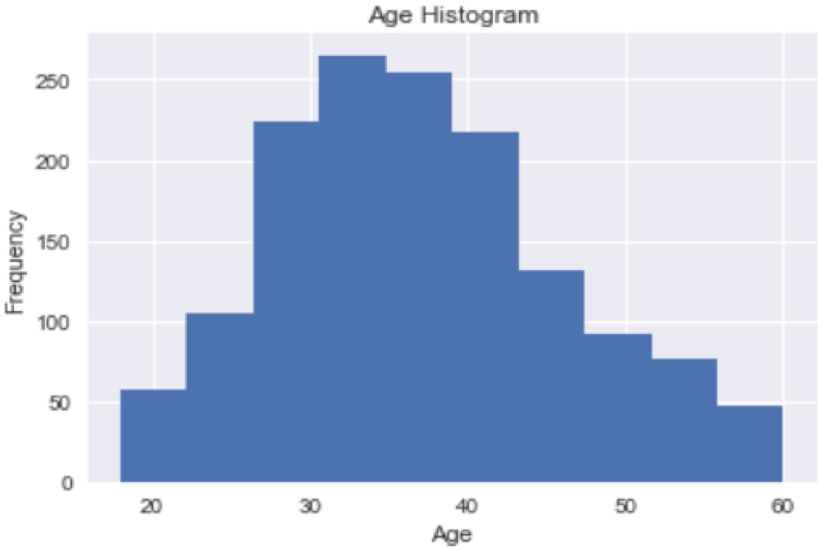
EXPLORATORY ANALYSIS

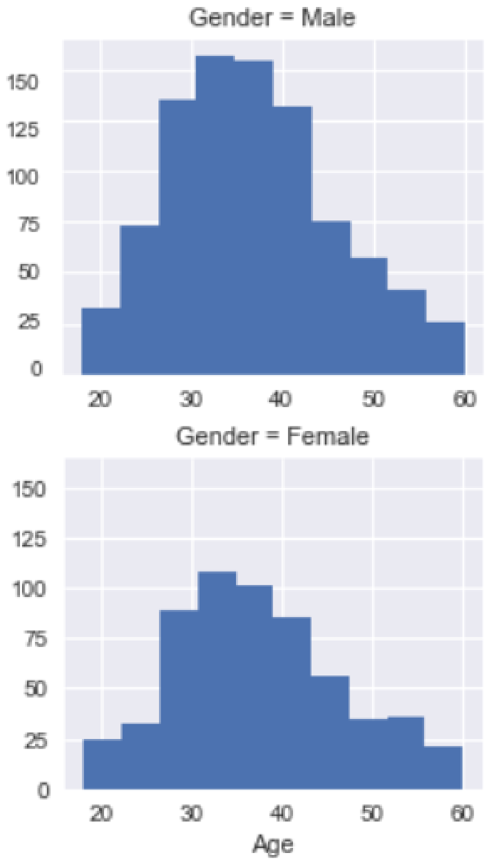
Here, we will use data visualization techniques to form insights about IBM employees with regard to demographics, attrition, and productivity. We also offer suggestions on how to mitigate any problems. Our first step is to get a sense of the demographics of our dataset, as we should know the distribution of demographic characteristics such as age and gender before we relate these characteristics to attrition.



Here, we can see that most of the employees at IBM are male. We can only speculate as to the cause of this distribution, but factors such as familial responsibilities are likely the reason for this disparity.

Next, we will observe the distribution by age IBM employees as a whole, and when separated by gender.

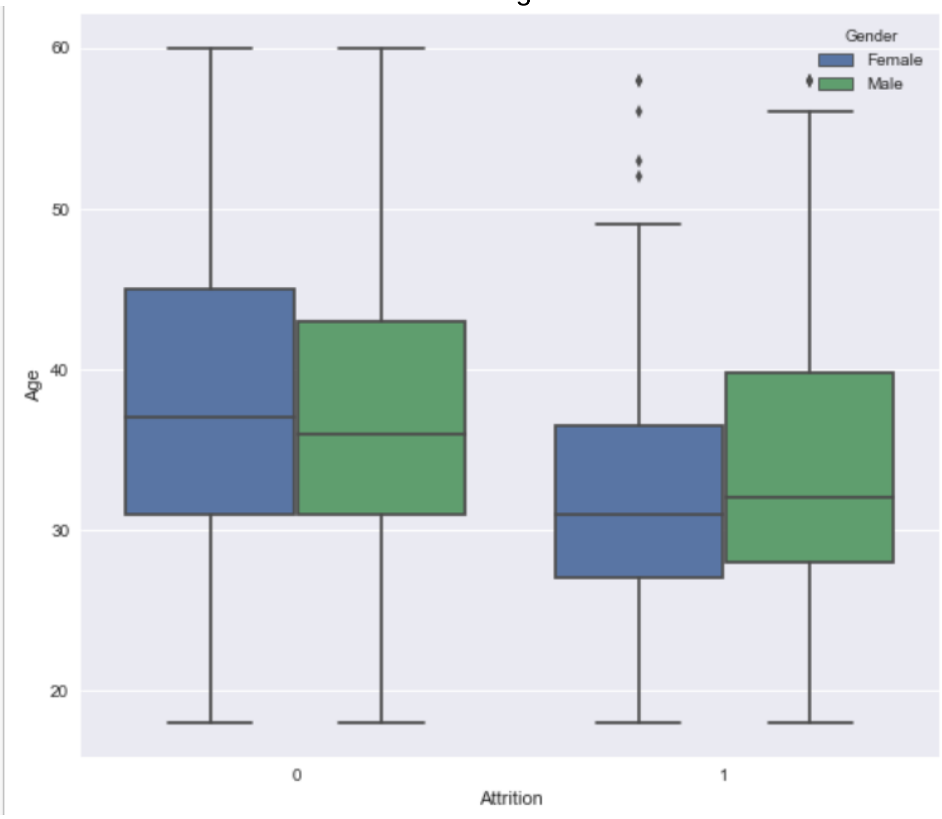




The above histogram comes out to be almost a unimodal Normal curve, although it is slightly skewed to the right. We see that a large number of employees fall under the age group of 30 to 45 which isn't surprising as companies often retain employees of this age group due to their experience. The graph to the right shows us the Age distribution in Males and Females separately. Their distributions have approximately the same shape. Thus, **there is no significant correlation between age and gender** in our sample of IBM employees.

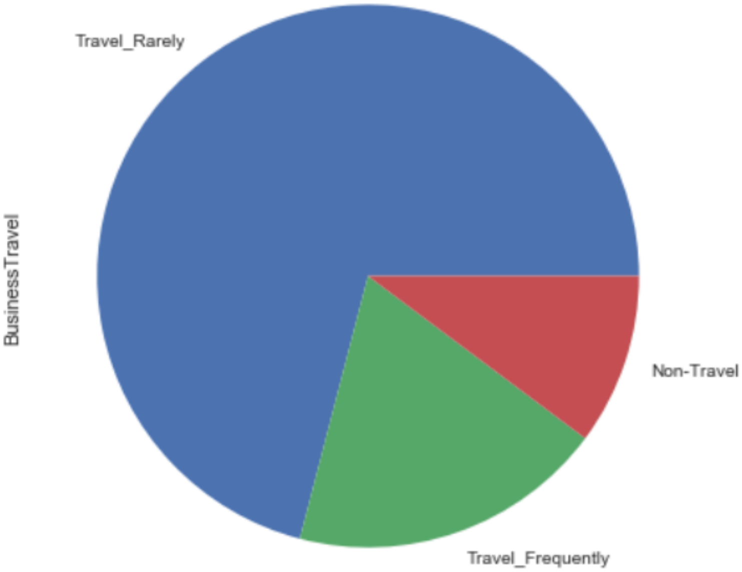
Next, we will dive into attrition, discovering what causes it and how it can be prevented. Even though the number of employees with Attrition is relatively small compared to those without attrition, 237 out of 1470 (16.12%) of employees cannot be neglected as employee attrition is a loss to company, in both monetary and non-monetary terms. In our python code, we have made a table which gives statistical data about each column for employees with attrition and those without attrition separately. We will study this table in order to determine factors, which corelate to employee Attrition. These factors may be of the employee, the IBM corporation, or a combination of both employee and employer. The first factor we will explore is the age of the employees and its correlation with attrition. We accomplish this with the help of a box plot.

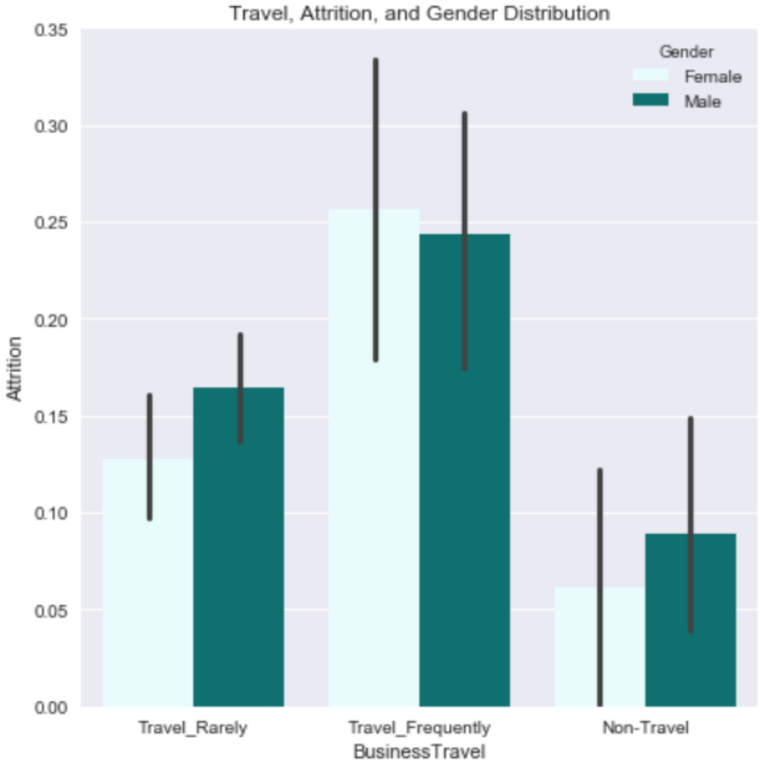
Attrition vs Age vs Gender



One insight from the above plot is that **employees with attrition are slightly younger than employees without attrition, although this correlation isn't very strong.** We can only speculate as to the cause; perhaps younger employees are more willing to change jobs due to lesser familial responsibility and less risk aversity. **There does not appear to be significant correlation between gender and attrition.**

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



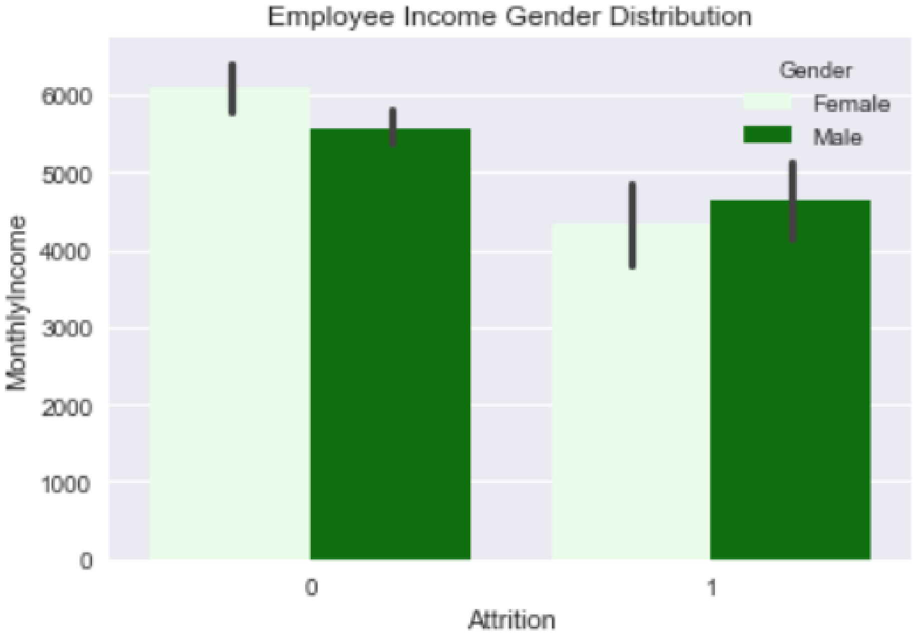


We can see from our pie chart below that the overwhelming majority of employees never travel, or only seldom travel. We can make a few interesting conclusions from looking at the graphs above.

- 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. Non-travel has the lower attrition rates, followed by Travel Rarely, followed by Travel Frequently. This correlation exists regardless of gender. **Therefore, IBM should we wary of having employees travel too much. This is because frequent business travel has a strong correlation with attrition.**
- Another insight we can draw is that women tend to have lower attrition rates than men when they undergo business travel rarely or never, but they have a higher attrition rate than men when asked to travel frequently. In addition, women consistently have longer confidence intervals, meaning they're more volatile when confronted with business travel. **Therefore, Women are more sensitive to travel than men. This may be due to reasons such as personal responsibilities.**

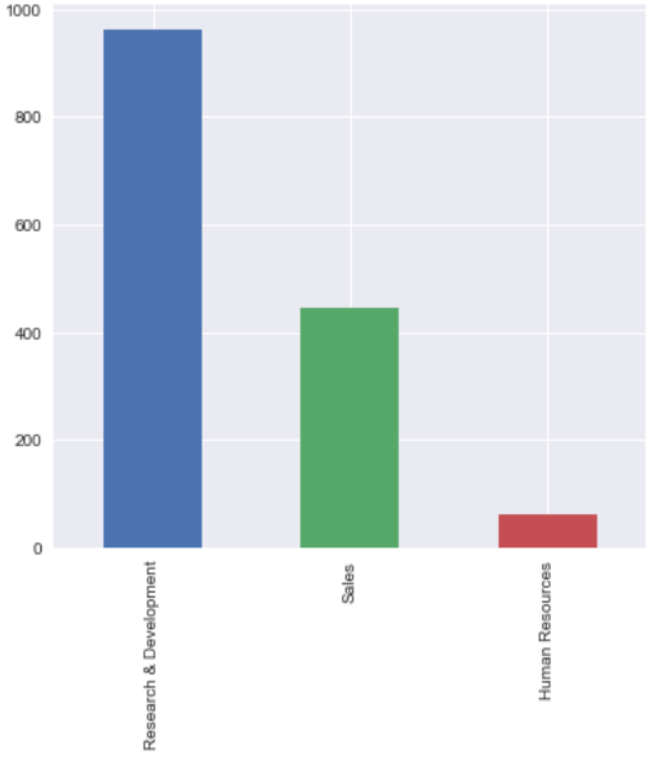
IBM should try to reduce business travel among its employees to reduce attrition. This applies especially strongly in the case of female employees.

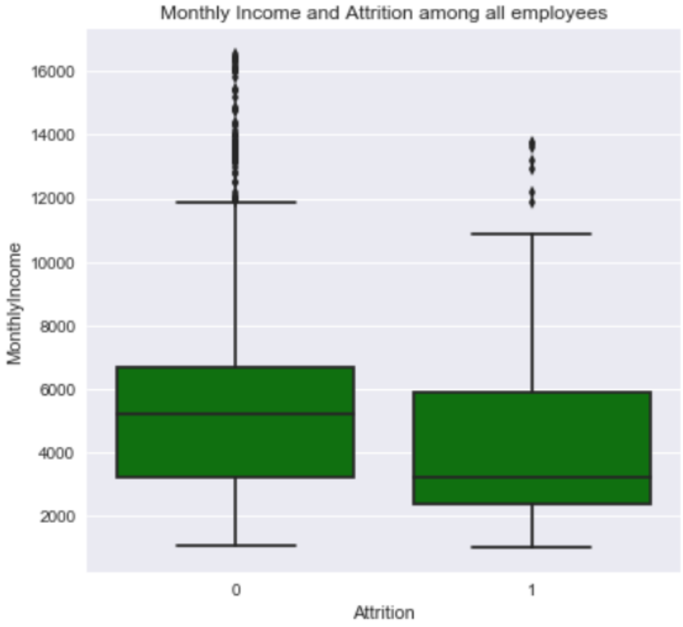
Next, we compare how monthly income correlates with attrition in the case of men and women.



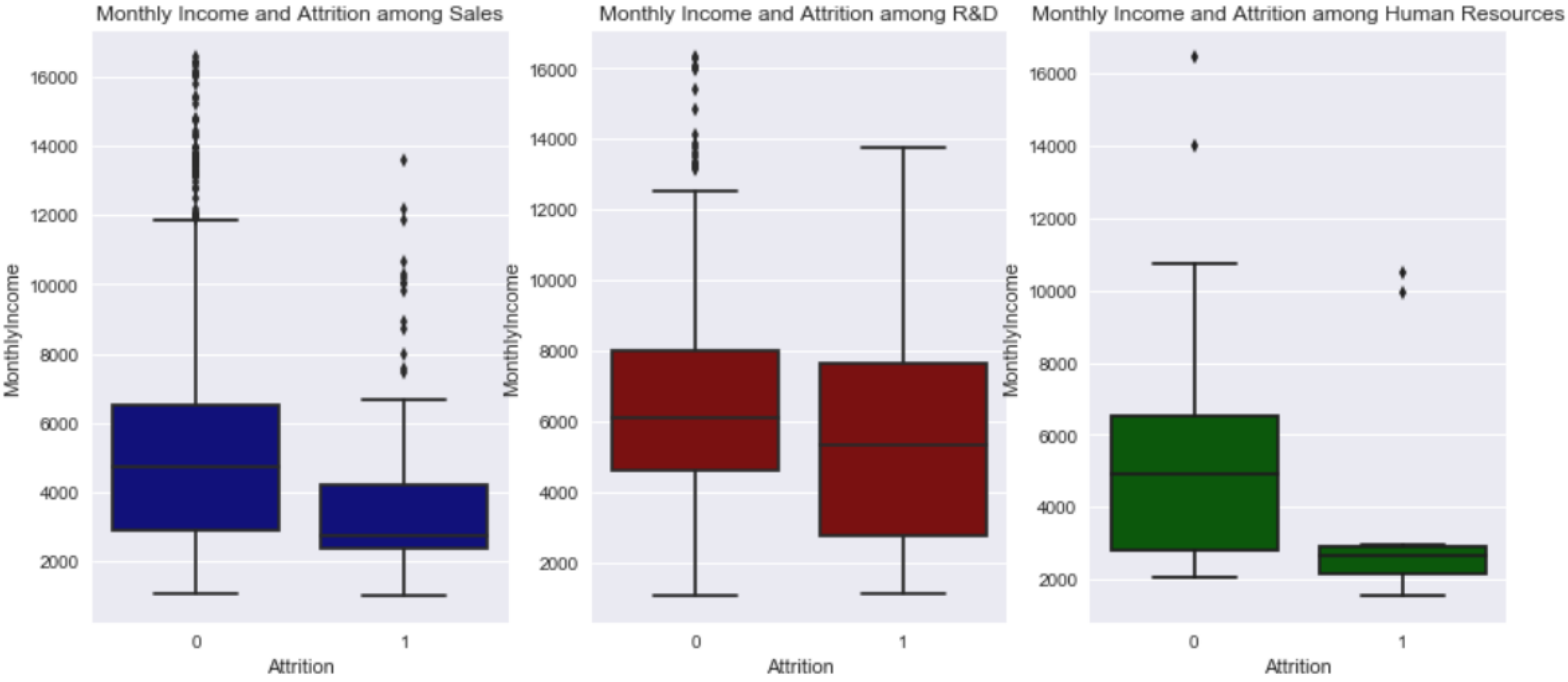
This is a significant correlation between income and attrition. This is not surprising, as unsatisfactory pay is one of the main reasons one would expect to cause attrition among employees. Observing the graph, we arrive at an interesting conclusion: **Men are potentially more sensitive to their monthly income than male employees.** We arrived at this conclusion by observing that male employees with attrition had a higher monthly income than female employees with attrition (i.e. Female employees tolerate a lower income than males as far as attrition is concerned). Interestingly, among current employees women tend to have higher incomes than men.

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 to its right shows that employees with attrition earn significantly less money than those who stay with the company. Also note that the box plot of employees with attrition is right skewed, further reinforcing the correlation between income and attrition. **IBM should consider raising the pay of its employees to decrease the chances of attrition**

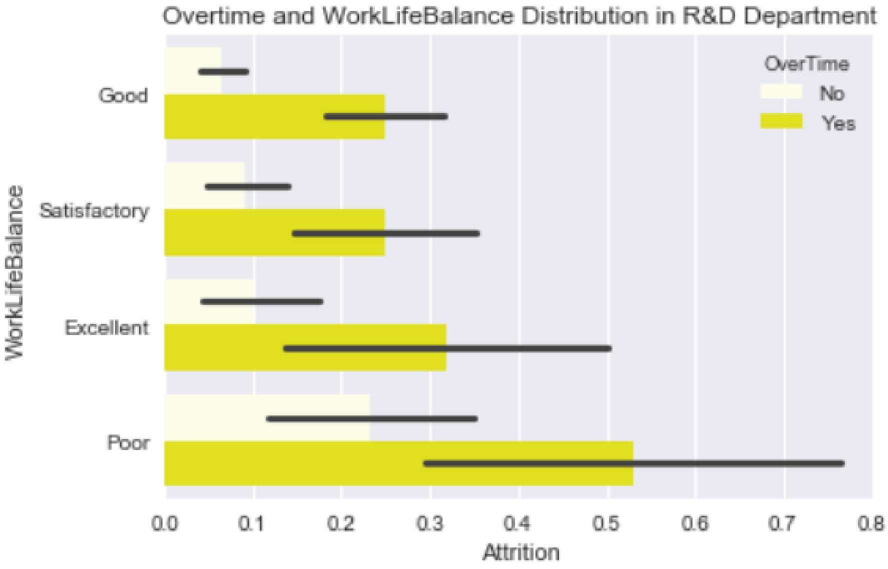


These three box plots again confirm the correlation between income and attrition. Note, however, that the disparity in income is especially significant in the Human Resources Department. It appears that a significant number of employees in HR might be being underpaid, and hence the large pay gap between those with and without attrition.

Human resources employees are especially sensitive to monthly income and could underpaid. IBM should consider increasing the monthly salary of HR employees to prevent attrition.

Similarly, we can see that the difference in pay among those who have attrition in R&D is not as large. **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 to prevent their attrition.**

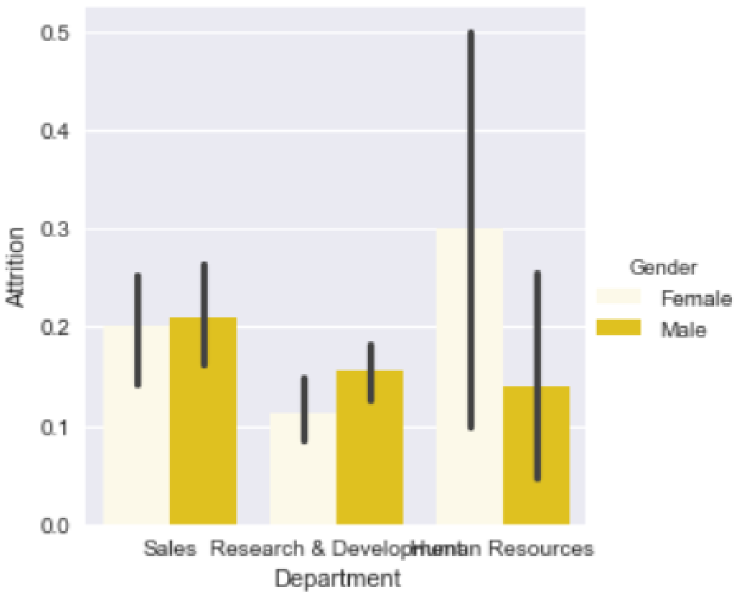
Let's dig deeper into the R&D department to see what causes its employees to have attrition, since income doesn't appear to be the main reason.



As you can see, two key factors in the attrition of R&D Employees is the work life balance and overtime, both of which corelate positively with attrition. This is a very stark correlation and leads to the conclusion that R&D employees, while they are paid well, are perhaps overworked. The inclusion of overtime more than doubles the chances of attrition regardless of work life balance, so overtime should be restricted significantly to reduce attrition in the R&D Department. In fact, an employee who has poor work life balance and is also given overtime has a >50% chance of attrition!

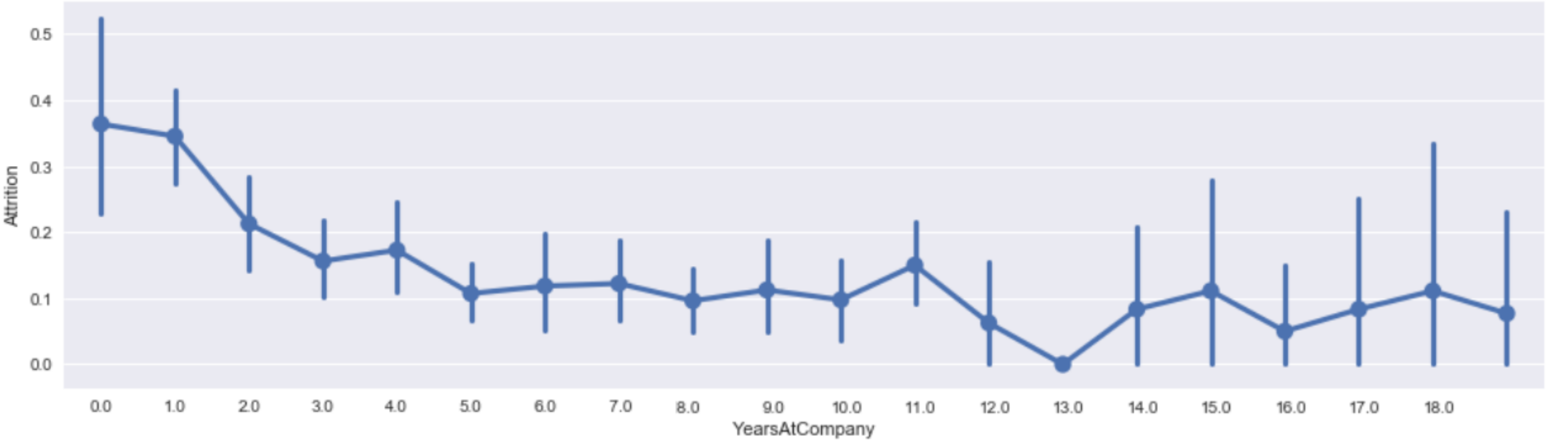
- **IBM should reduce the workload of R&D Department employees significantly to improve their work life balance.**
- **Overtime should only be given when absolutely necessary, as it is corelates very strongly with attrition.**
- **The pay of R&D employees is sufficient as seen above, but they may be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

Now, let's see how the attrition in each department corelates with each gender. As you can see, attrition rates are fairly comparable for male employees across all three departments, although Sales is slightly higher than the other two.

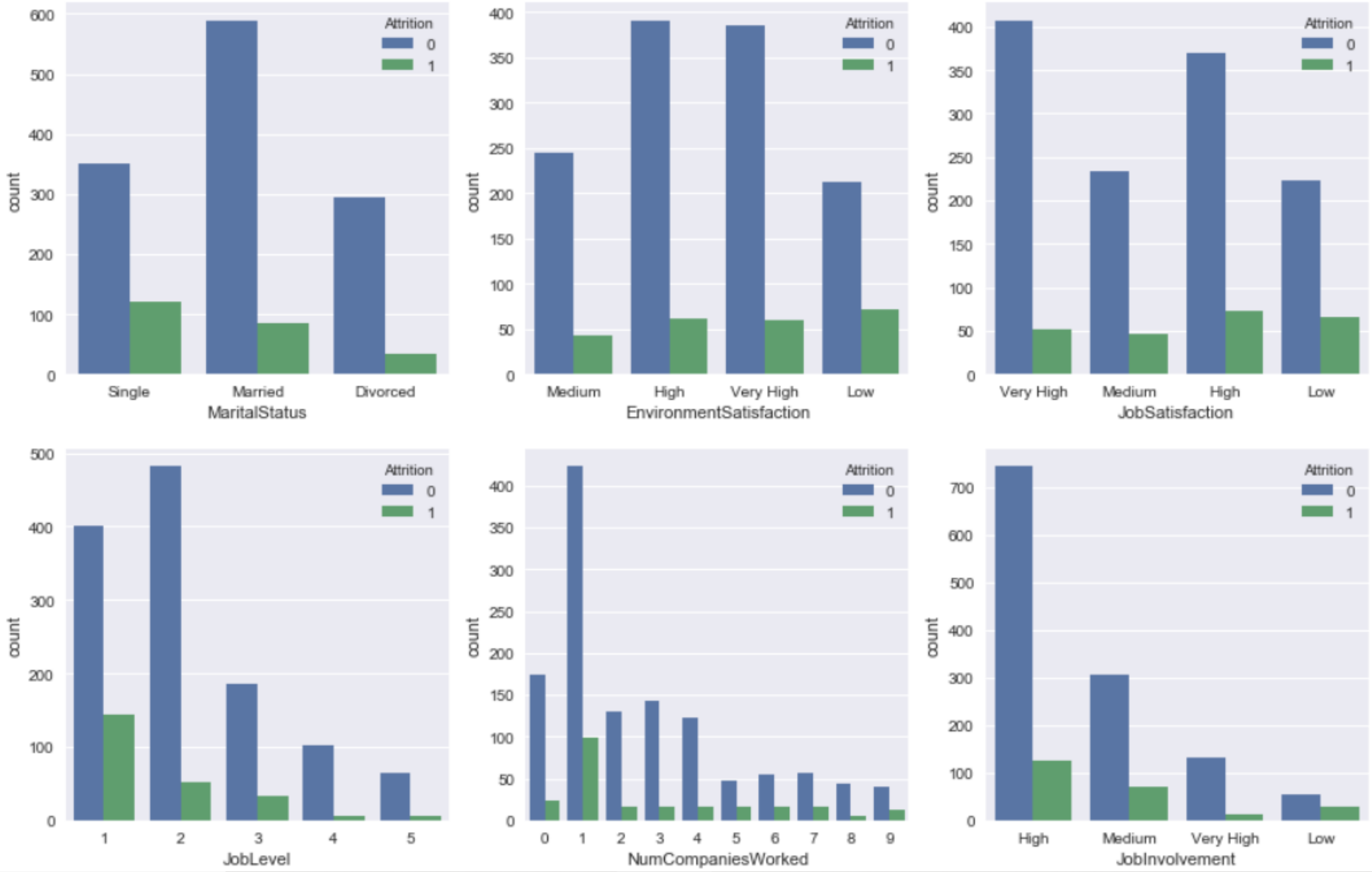


As you can see, men and women have comparable rate of attrition in Sales and R&D, **but women are much more likely than men to undergo attrition in the Human Resources department.** We can only speculate why this is, perhaps due to poor work culture or stress experienced by women in the HR department.

Therefore, **IBM should be cautious in hiring women employees for HR positions and investigate whether women are mistreated in the HR department. There is an unknown issue with the HR department that drives female employees to attrition, and this should be investigated and rectified.**



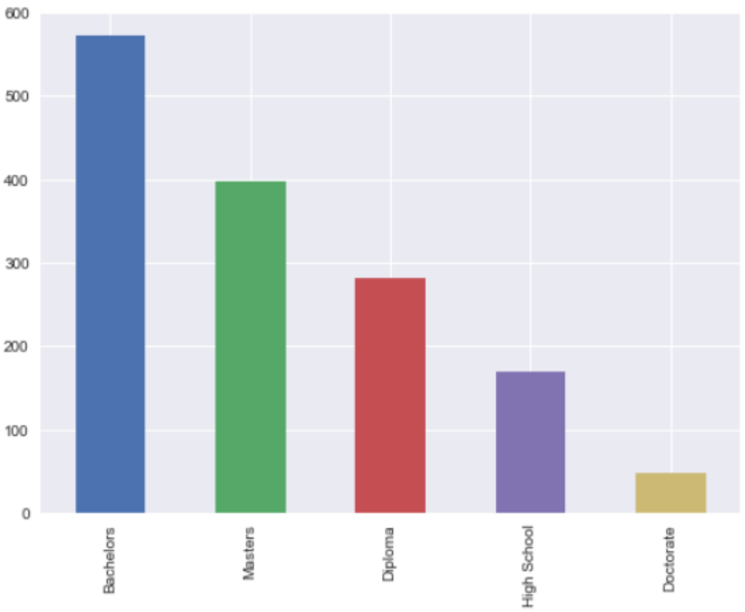
There is also a considerable inverse correlation between the number of years an employee has worked for IBM and the chances of him/her having attrition. It is likely that employees who have been with IBM for many years are well established in their position and are not interested in attrition as much as newer employees.

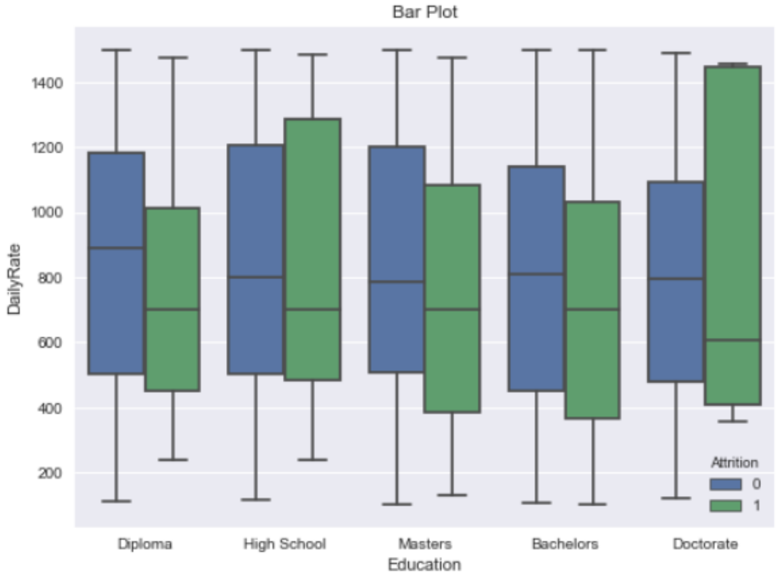


A few insights from the graphs above:

- **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:

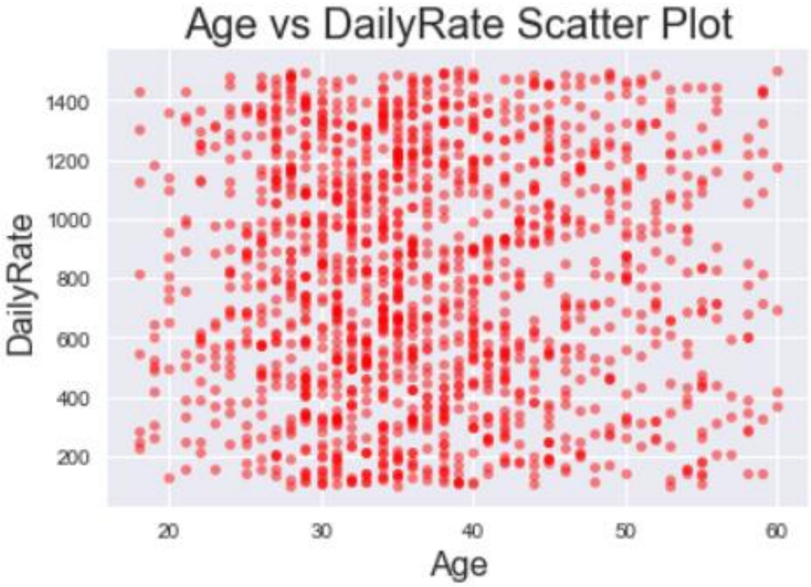




Above, we explore the relationship between education level and Daily Rate. Observing the bar chart, it is clear that Bachelors degrees are the most common education level among IBM employees, followed by Masters and Diploma.

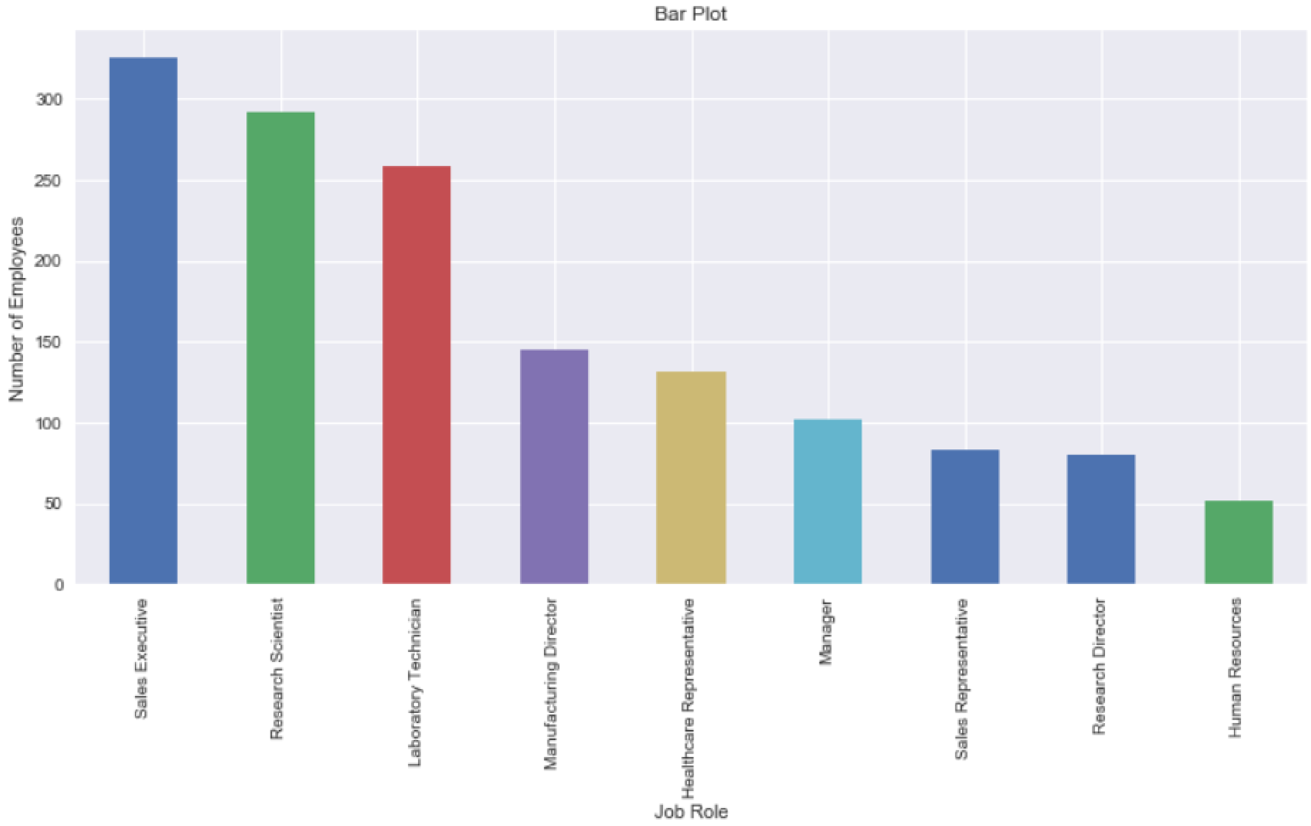
Looking at the box plot, we see that:

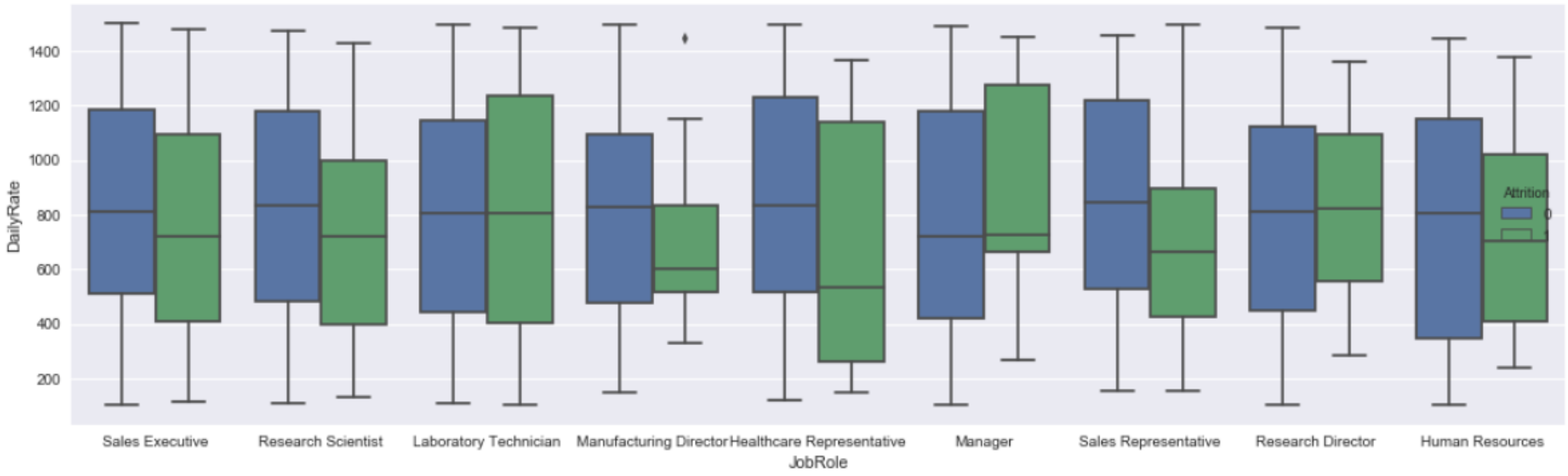
- **There is no significant relation between Daily Rate and Education Level. The box plots are comparable at each education level and are approximately symmetric.**
- **Therefore, hiring more educated employees is not an effective way for IBM to increase the Daily Rate of its employees.**
- **In addition, 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, but that's only because the most employees are of that age group and does not reflect on the Daily Rate. Hiring younger or older employees is not an effective way for IBM to increase the Daily Rate of its employees.

Below, we use box plots to investigate 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).

Interestingly, we note that Laboratory Technicians and Managers with attrition have similar daily rates to those who have not had attrition. Daily Rate is not a significant factor that leads Managers and Technicians to have attrition. There are likely other causes that lead employees in these roles to have attrition.

CONCLUSION

Through our analysis, we discovered various insights regarding employee attrition and performance at IBM:

1. Employees with attrition tend to be younger than employees without attrition.

2. IBM should avoid asking employees to travel very frequently. Frequent business travel has a strong correlation with attrition.

3. The correlation between frequent travel and attrition is especially prevalent among female employees. IBM should try to reduce business travel especially for women as they more sensitive to travel compared to men.

4. There is a significant inverse correlation between income and attrition among IBM employees. IBM should offer increased salaries to reduce attrition significantly.

5. The relationship between income and attrition is starker in the case of male employees. Male employees appear more insistent on a higher income and this reflects in the case of attrition of male employees with lower monthly incomes, relative to women. This can also be mitigated by raising salaries.

6. HR employees are more sensitive to monthly income than other departments, this is evident by the stark difference in the income levels of HR employees with attrition, and those without attrition This points to HR employees being underpaid relative to other departments. Income is the main factor that determines whether an HR employee has attrition.

7. Conversely, R&D employees are much less sensitive to monthly income than other departments, indicating that R&D employees are receiving satisfactory salaries. This also means that other factors in the workplace, not income, are responsible for causing attrition in the R&D department.

8. Instead of income, the main reason R&D employees have attrition is because they are overworked. There is a *very strong* correlation between work-life balance and

overtime with attrition among R&D employees. In fact, an R&D employee with poor-work-life balance and overtime has a greater than 50% chance of attrition! Adding overtime to an already poor work-life balance multiplies the chances of attrition. To mitigate this, IBM should reduce the workload of R&D employees, giving overtime only when absolutely necessary.

9. Female employees have a lower attrition rate than men in all departments except Human Resources. Women are twice as likely as men to have attrition in the HR department. This is a very good indication that the HR department is experiencing a problem that affects women. The HR department should be investigated to uncover what issues exist among female HR employees.

10. There is an inverse correlation between the number of years an employee has worked for IBM and the chances of him/her having attrition. IBM should be aware that experienced employees are unlikely to have attrition and delegate responsibilities accordingly.

11. There is an inverse correlation between job level and attrition. Employees who have a senior position in the company are less likely to have attrition. This is not surprising as only employees who establish themselves well at IBM are likely to have a senior position in the company.

12. 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.

13. Interestingly, employees who have worked for one company in the past have a higher attrition rate, than those for whom IBM is their first company. This is likely due to fresher employees being risk averse and compromising with regards to the workplace.

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 library, and most of all; we gained a newfound appreciation for the field of data science.