

***Introduction***

Data refers to information or facts that can be processed, analyzed and used for decision making. In this project, we cleaned the dataset we have on many subjects such as monthly income, education levels, races, genders, and positions of 300 employees of the IBM company via Python, processing them, visualizing various graphs and making analyzes and inferences on these visualizations.

***Data Cleaning Part***

First, to be able to use packages the pandas and math are imported. Then the data is read. To understand the data and execute ideas for cleaning, .info() function has been used. As a result, the number of null values, column names and data types could be observed. The .head() and tail() functions have been used to see the problems more clearly. After that, the column names are fixed by using .str.title() and rename() functions. The data has no duplicates. To see the problems in variable names, the frequency table has been made. All the typos has been fixed by using .title() and .loc[] functions. With the .select\_dtypes() function the typos were checked, and no mistakes were found. The number of null values were checked with a percentage and frequency table. The categorical null values filled with the mode of the data by using .fillna() function. After that, from the description table it was observed that the maximum of some of the variables are too big. They can be called as an outlier. To eliminate this, those maximum values were equalized to zero, and then changed to mean of the variables. And finally, the data checked again with the .describe() and .isnull() functions after no problems were observed, the new clean data were saved as excel file.

***Research Questions and Interpretations***

**1.) How Much Are Different Races Paid Monthly?**

Chart

Description automatically generated

In this research question, we aim to find race monthly income for two genders, and we choose the violin plot to visualize our question.

🡺For White workers, the male's median income is higher than the female's median income. Yet, we can see that female workers’ monthly income is more spread than male female workers’ monthly income.

🡺For Black or African American workers, again males' median income is higher than females' median income. Moreover, unlike white workers, males' monthly income is more spread out than females' monthly income. Additionally, we can see that males' monthly income distribution has a bimodal distribution.

🡺Lastly for Asian workers, female workers’ median income value equals male workers’ first quartile income value. Also, some male workers, are paid more than female workers. Again, like black or African American male workers, Asian male workers’ monthly income distribution has a bimodal shape.

Overall, when we look at the violin plot, there is not much difference between races’ monthly income. We can say that there is no income disparity between races in that company.

**2.) Which Age Group Works in Different Positions?**

Chart, box and whisker chart

Description automatically generated

In this question, we aim to find which age groups work in which positions and we chose the box plot to visualize our question.  
🡺In Production Technician Department, the median male age is higher than the females median age, and age groups are between 32-36 for female workers, 33-38 for male workers. Here we have some outliers, four of them above the upper whisker, and one of them below the lower whisker in females. However, there is not any outlier for male workers. Lastly, we can say that while female workers' age distribution is skewed to the right(slightly), the male age distribution is skewed to the right(slightly).  
🡺 In IT Department, again male workers' median age is higher than female workers' median age. Also, age groups are 33-39 for female workers, 34-37 for male workers, In male workers, we have outliers two of them above the upper whisker, and two of them below the lower whisker. For female workers, there is only one outlier which is below the lower whisker, and it is the youngest person in this company. In addition, the female age distribution is skewed to the right(slightly), and the male age distribution is skewed to the left.  
🡺 Finally, in Sales Department, the male worker's median age is higher than the female workers’ median age, age groups are 32-36 for female workers, 33-37 for male workers. We have outliers for females, three of them are above the upper whisker and one of them is below the lower whisker. Also, there is only one outlier for males, and it is the oldest person in this company. Lastly, we can say that while female age distribution is skewed to the right, male workers' age distribution is bell-shaped.

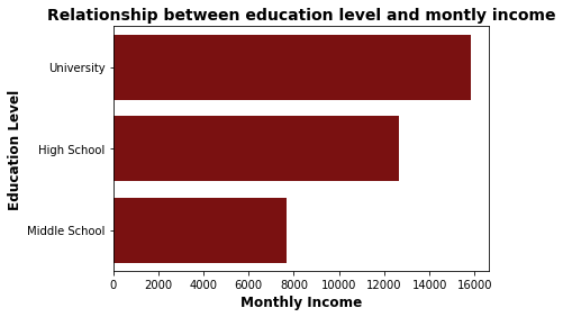
3**.) How Is the Perform Evaluation of the Genders Distributed by Positions?**

Chart, bar chart

Description automatically generated

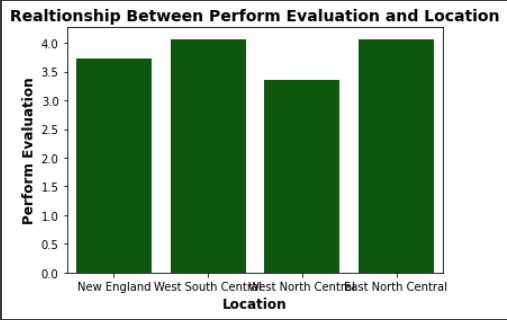
In this research question, we aim to find genders perform evaluation in different positions and we choose the clustered the bar chart to compare female and male workers. Generally, male workers perform evolution is slightly higher than female workers' perform evaluation for different positions. Therefore, we can conclude that male workers better in that company.

**4.) What Is The Relationship Between  Education Level and MontlyIncome?**



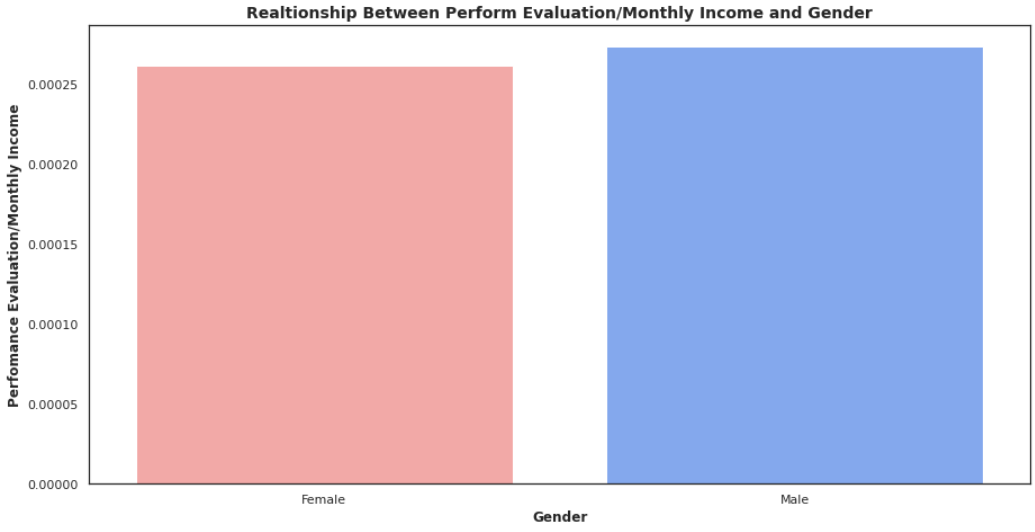
There is a positive linear relationship between education level and monthly income.

**5.) Which District Take More Monthly Income?**



West South Central and East Norh Central have the most perform evaluation value. Third New England and forth West North Central. If I choose one of the two people with equal benefits. I choose one from West South Central and East North Central.

**6.) How Effects Male Performans Their Salerys and Are They Deserve to Their Salerys?**

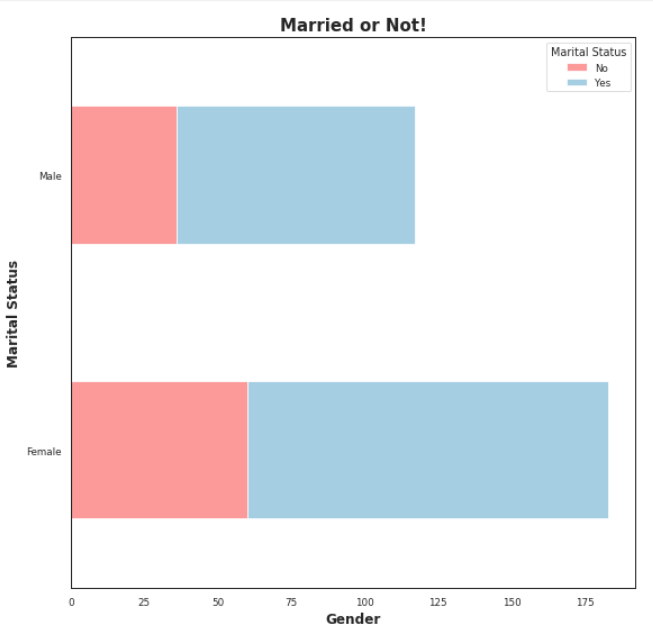


Male take more salery than female. Male's salery is not enough for their performance evaluation.Against the female, male take less salery based on their performance.

**7.) Is There Any Relationship Between Working Years and Perform Evaluation?**

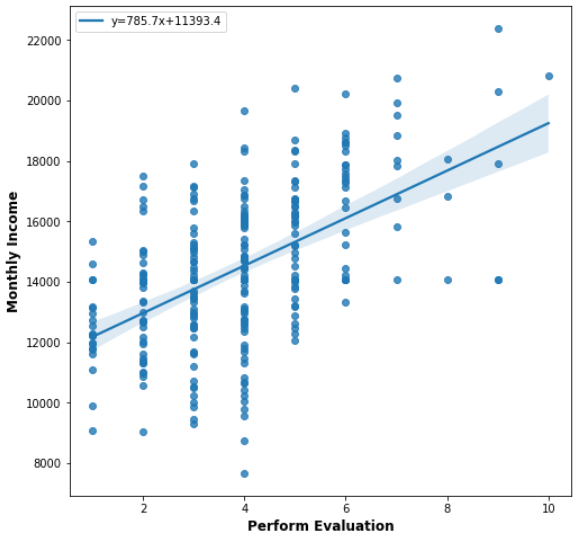
From the above chart, it can be interpreted that, as the working year increases, perform evaluation is also increases. They have a positive relationship. The data consist of people who at least work two years, and the maximum working year is five. The perform evaluation is a rating from zero to ten. While ones working year is increasing their perform evaluation is increases as well. Therefore, it can be said that people get better when they work more years and when they get experience, as a result their performance increases.

***8*.) How Is Gender Shaped According to Marital Status?**



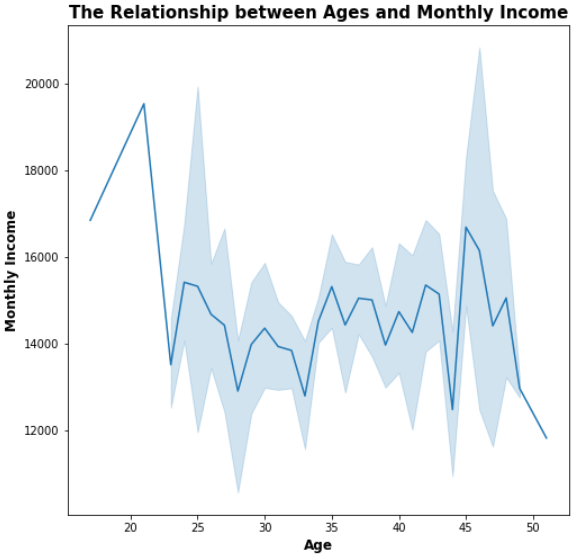
From the graph, first of all it can be seen that the number of female workers are more than male workers. There are a total of 160 female and 117 male workers. There are approximately 55 unmarried and 105 married woman workers. Also, there are approximately 30 unmarried and 87 married male workers. The majority in this workplace is married female workers. After them the most people are married male workers. Then unmarried female workers and unmarried males come respectively. Generally, most of the workers are married.

**9.) Estimate The Correlation Coefficients According to Genders and Monthly Incomes and Explain the Reasons for Your Estimations?**



Looking at the data in the chart, it can be easily observed that the monthly income values that are far from each other, but in a linear balance, increase with the performance evaluation. First of all, it can be said that there is a positive correlation in this graph, but the estimate about the correlation coefficient may not be very consistent. Because the correlation here is a weak correlation. The value will be very far from 1 but less than 0 which is (0.10-0.25).

**10.) Is There any Linear Relation Between Monthly Income and Age?**



By looking at the graph above, it can be clearly observed that there is no linear relationship between age and monthly income, and that monthly income does not vary according to the increase or decrease in age. A 20-year-old can earn more than a 25-year-old individual, less than a 50-year-old individual.

***Conclusion***

To sum up, we must first clean up this messy data, that is, fix it and fix it. If we can then visualize the clean data we have using the right visualization techniques, we can analyze and interpret the data much easier and faster. The accuracy and consistency of these analyzes depends on the ability to properly clean and visualize the data.

(We used MidJurney Al bot to draw Cover Page )