Report for Data Science Project

As a student who is currently a STEM major, I was really curious as to what aspects are involved for students to land a good job. In addition to that, I wanted to find a way such that I would be able to predict the amount of salary an employee makes or will make. Apart from satisfying my inquisitiveness regarding the issue, I would also be able to understand more about this side of entering the workforce. Moreover, as I started working on the project, I realized that it had far more applications than I imagined it would. For example, it could be used by companies to calculate fair wages, it can be used as a source of information to promote equality in the work sector, and many more uses that can be facilitated by making a few tweaks. As it stands, it is an analysis of specifically STEM-based jobs, a yearly compensation prediction function, and a thorough analysis of some factors.

GOALS:

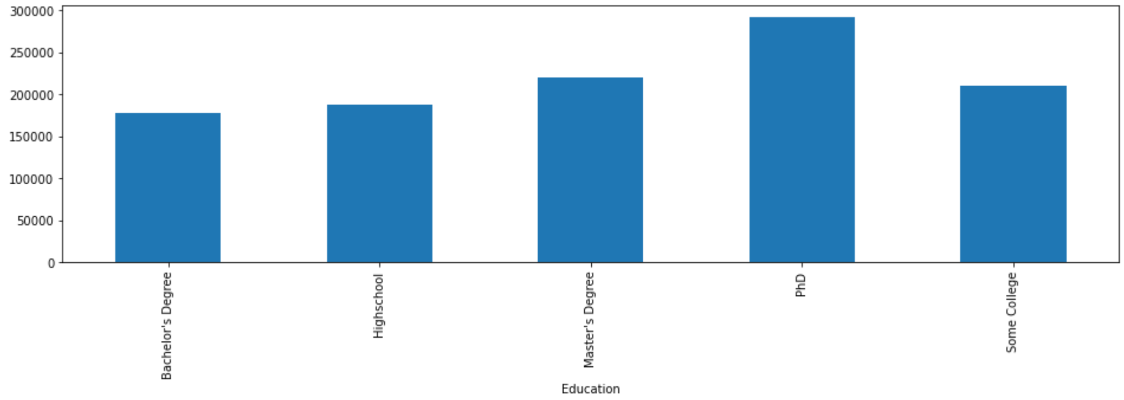
- Analysis of a Dataset that contains information about STEM Based salaries (Source: Kaggle).
- Creating and Implementing a Machine Learning Model to Predict the salary of an employee.
- Analyzing the data and drawing conclusions.

OVERVIEW of the PROJECT:

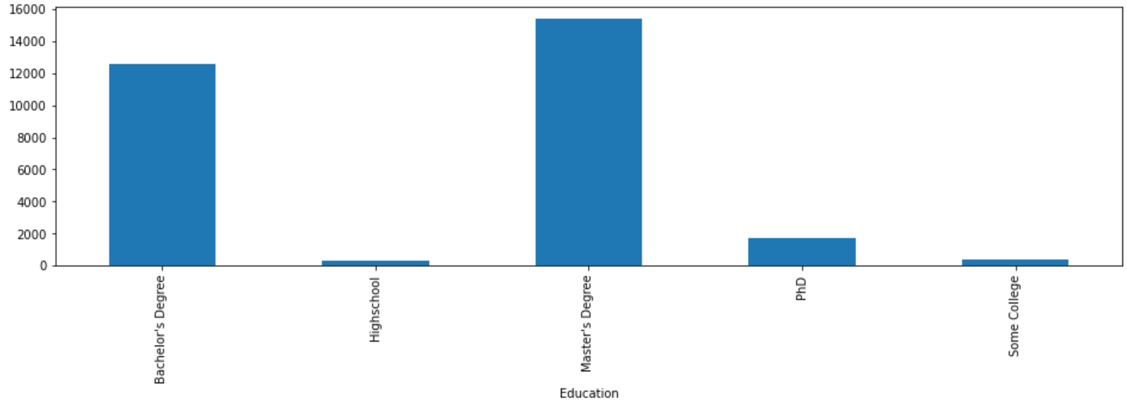
- Understanding the data and getting a rough idea of how factors relate to each other
- Some analysis to gain an understanding of certain aspects like gender, education, job title etc. interplay.
- To further understand the data, I used data visualization like bar graphs, scatterplots, heat maps etc.
- After that the next step was to clean the data, this included removing NAN values and unnecessary columns. Following this, I proceeded to perform label-encoding so we can facilitate classification algorithms.
- Finally, I implemented a Machine Learning Algorithm using a DecisionTreeRegressor
- Next Steps:
    - There may be cases when certain data points may not be known, hence need to make a function to account for that (eg. when the user is trying to make a prediction he may not know the bonus value or something else).
    - Need to create a more accurate clustering algorithm, so we can factor in data like education level etc. to improve the prediction.
    - Implement a function that integrates clustering and regression algorithms.

OBSERVATIONS and CONCLUSIONS:

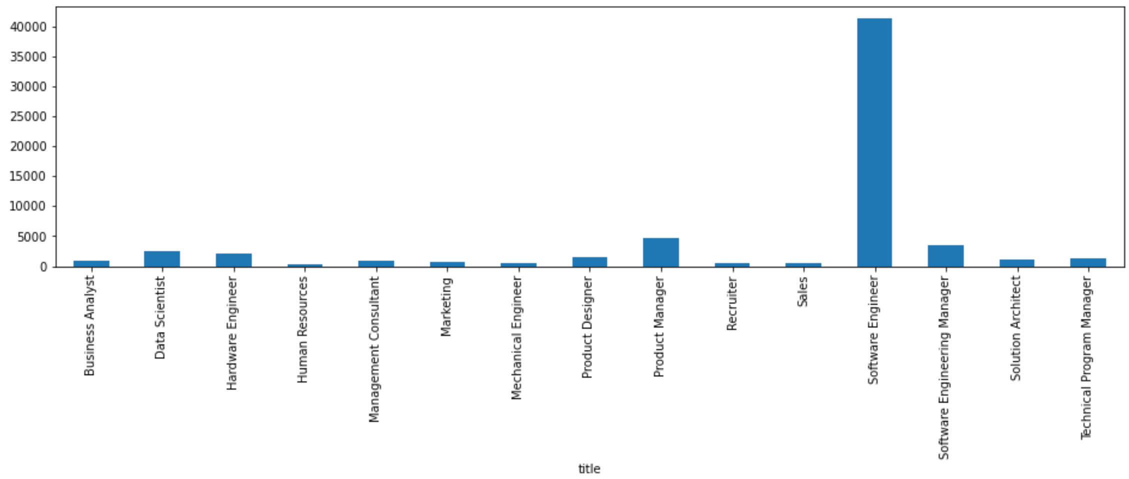
- Conclusions Drawn from Data Exploration and Basic Analysis and Notes
    - Education
        - Ph.D. Students get the largest compensation on average compared to employees with other education levels.



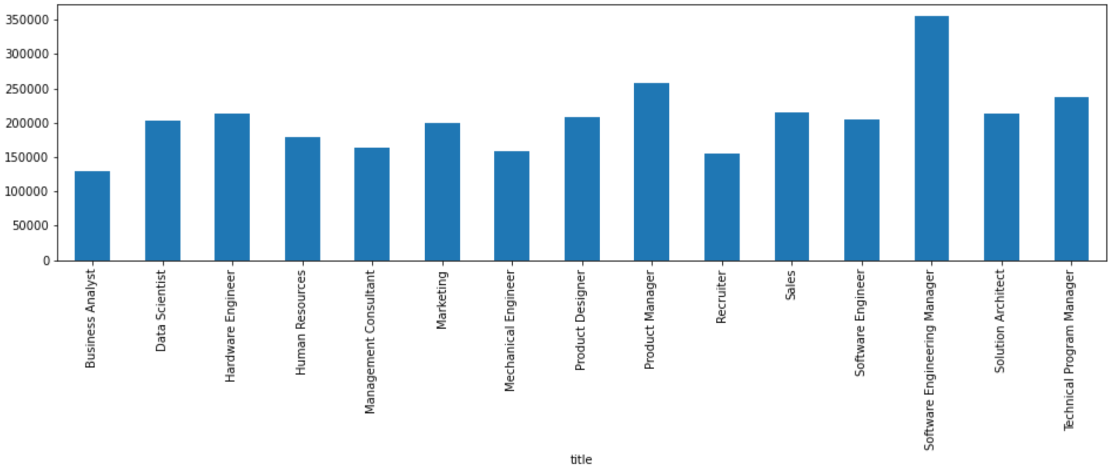
        -
        - At the same time, it is important to note that the number of students pursuing a Ph.D. is far less, in fact in upper education the trends are as follows:



        -
        - I also analyzed the pay gap between different educations levels and was surprised to find that HighSchool and Undergraduate students have a similar pay gap, but we must also understand that there may be some confounders present as there is a huge difference in the number of students.
    - Jobs
        - The most common job role was that of a software engineer.

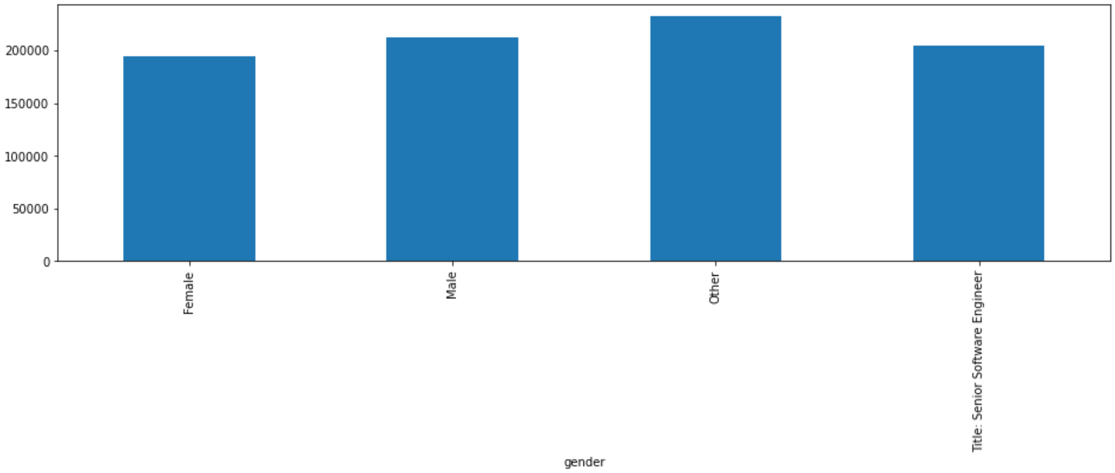


    -
    - The highest paying job role was that of a software engineering manager.



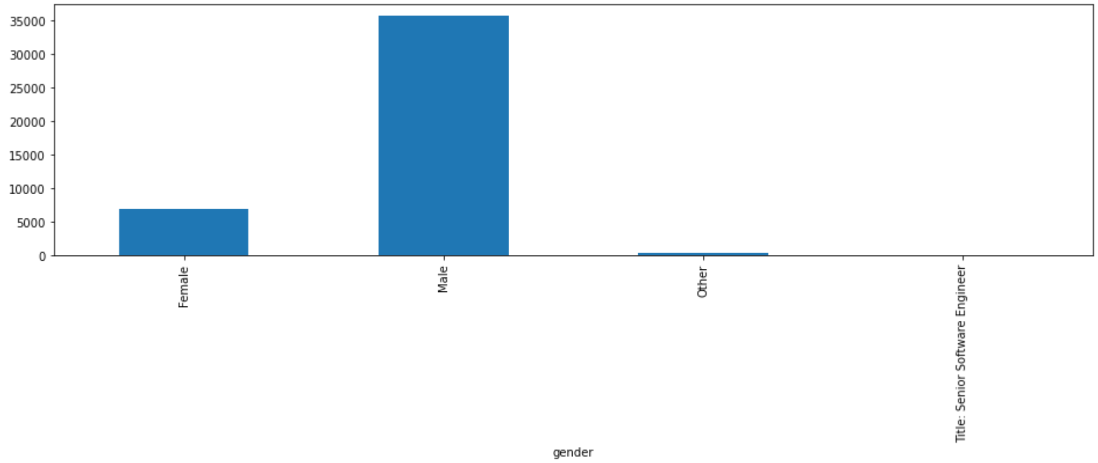
    -

  - Checking for Inequality in Gender
    - An important issue that we must address, this part of my analysis, helped me understand a social aspect of the job roles.
    - In terms of total yearly compensation, we can see a slight difference:



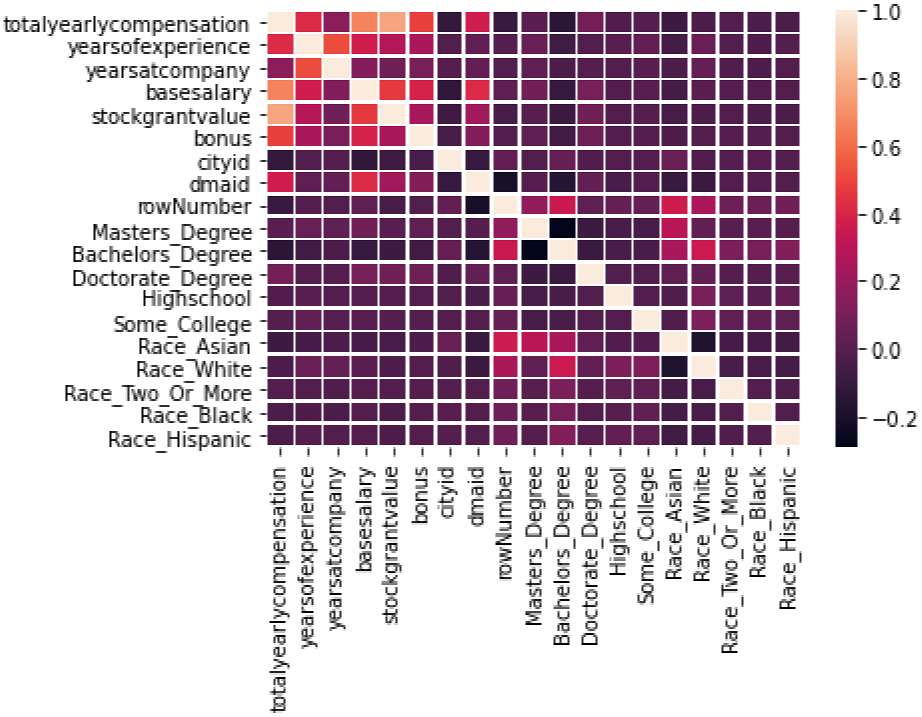
    -

- In terms of the number of people, we can see a huge disparity:



-
- NOTE: there is a small error in plotting as one part of 'Title' has been included.

- Conclusions Drawn from the Machine Learning Algorithm and Notes:
    - Used a Correlation Matrix which gave me a good idea of what factors contribute to the data and what does not. I used a heatmap to get a good visualization.



    -
    - 'Totalyearlycompensation' has a strong correlation with 'stockgrantvalue' and 'bonus'
    - In addition to that 'yearsofexperience' and 'basesalary' has a good correlation.
    - Surprisingly 'yearsatcompany' did not matter as much.

- Something that I took into account while implementing the ML model is that when a person wants to predict his salary he may not know the stock grant value, as that too is usually part of the pay, and hence while implementing the model in my function I provided for a case where one could choose whether or not that data should be included.
- A good accuracy model was expected due to the large size of the data.
- As far as Regression was concerned, I tried a few models based on research and finally decided on the final one (DecisionTree) based on its accuracy score.
- I could not implement an accurate model with categorical data and hence had to default to using the regression model only.
- The Highest Accuracy model which I implemented was a DecisionTreeRegressor with a 97.5 % accuracy (when data for all columns are available i.e columns like stock grant value was known).
- In cases where it was not known, the accuracy was around 94.1 %
- The FINAL FUNCTION implemented takes the data based on which we need to make a prediction and a boolean value to know if ‘stockgrantvalue’ is provided. Based on that the function returns
    - The Predicted Value
    - Accuracy Score
    - Residuals

***