Dissertation

Data Science Profession Survey

Overall Aim of the project

The aim of the project is to create a data analysis visualisation on the datasets which was collected from internet jobs sites about the skills requirements in the field of data and analytics. We are using three datasets namely, Glassdoor Data, Profession Survey Data, Salary Data.

- Glassdoor Data
  This dataset consists of all the recruitments posted on Glassdoor till the year 2020. The job postings are related to all the positions which are relevant to data and analytics field such as machine learning engineer, data analyst, business analyst.

- Profession Survey Data
  This dataset consists of question panel which asks the applying candidate about his/her education qualification. We are analysing this dataset because this gives us the highest and lowest degree required to be employed in the sector of data and analytics. The dataset has around 40 questions distributed among the 18-45 age groups.

- Salary Data
  This dataset would mainly be used for the gender differences in the field of data analysis. This dataset will help us in determining the salary differentiation with respect to gender. This dataset is similar to the above dataset, “Profession Survey” because that has a gender question.

Approach

1. **Data Cleaning**
   The cleaning of the dataset will be initial step to dive in the project. As the dataset is collected from the internet, it will contain lot of misinterpreted data. Though the dataset was available on Kaggle the cleaning makes us sure to approach the project in a right way. The cleaning of the data contained some of tasks mentioned below:
   a) Company name text
   The raw data consists of the company name and the ratings provided by the Glassdoor website. The cleaning required here was to extract the text name of company as we have the ratings column already.
   After going through the dataset, the Salary Estimate column has unnecessary filling of "Glassdoor est.". This was removed too.
   After clearing the salary field, we have the salary range as lower and higher. Creation of maximum and minimum salary was done.

   b) State field
   The need of state name and job location is required to be cleaned. We can see in the location column that city and state name are provided. The use of "replace" command is used to filter out the state names.

   c) Age of company
   The age of company is calculated by the year founded. But prior to that we have to make sure the state name is same as the headquarter name. We extracted the state name above and now we compare it with the headquarter name with the help of headquarters column. We created a new column name, "same_state".

A new column was created for determining the age of the company. As the current year is 2022, and the data is from the year 2020, I have subtracted the founding year with the year 2020.

d) Parsing of job description (python, etc)

After going through the dataset, the number of skills required are python, R, Spark, AWS, Excel. I have filtered out every skill listed on the dataset to know which skill is the most common in all the positions.

Separate column was introduced with respect to every profession.

Dropping of "Unamed" column.

2. EDA on Glassdoor Dataset

This process can be done after the dataset is cleaned and saved. Now as the dataset is cleaned, we will now identify the job title given in the dataset and the seniority of the position. The next step will be to analyse the job description length according to the job position posted. And at last, will be the analysis of competitors, this will also include the hourly wage.

After all these analyses we will plot histogram to visualise the distribution and for better clarification we can use box-plot between salary and avg. salary, etc.

We can also plot a correlation matrix and with the help of that we can plot other visualisation graphs which are in relation to other features.

3. Profession Survey Dataset

This analysis is done in the second dataset. This dataset can be use to analyse each question and number of choices made for that question. We can use "plotly express" and "plotly graph" to demonstrate the EDA. As the skill specification question is also asked in the dataset, we can make a histogram to analyse the distribution of skills most of the candidate have.

As the skill distribution is plotted, same can be done for education qualification. After the graphs are plotted successfully, I will try to compare different education levels with different career paths. This may a bit tricky as I wanted to explore the features by using a dropdown menu. I have read that it can be done with the help of "plotly", but need to refer it from the documentation.

4. Gender Difference Analysis

This analysis is a short analysis related to gender in the sector of data & analysis. The initial steps to explore the data will be similar as the above one, but in this exploration, we are concerned about the question which asks the candidate about it gender. We can calculate the sum of the question and plot a histogram between Man and Woman. After the distribution is clear we can merge this distribution to the earlier analysis which required education qualification, female distribution, etc. A normalisation graph can also be created for this analysis.

5. Model Building (if possible)

At the final stage, we will try to create a model which gives the accuracy value by using Regression techniques, GridSearchCV. And at last, we will try to create a web application which will act as a job portal for data field enthusiasts (I am not familiar with web application, will need help in this. I am looking to use Streamlit library of python and will try to dump the pickle file in that).