Spencer Mui - sm3061

Data Science Salary Prediction

Objective
Forecast the growth of Data Science careers by using historical data from recent years to interpret potential salaries and remote work opportunities

Importance
This dataset is an encapsulation of what the Introduction to Data Management class is meant for. The data set that will be used pertains to jobs that are based around the concepts and strategies we have learned in the class.

Expected Outcomes
From the project, the goal is to create an accurate model to use as a predictor for prospecting data science students or people looking to switch careers. If the salaries are strong and the job outlook is good, such as having a high availability of jobs, potential users can reference the model and estimated predictions to see if moving into data science is a good career path. The ideal prediction model shows year over year job and salary growth.

Notes
The dataset used is (https://www.kaggle.com/datasets/yusufdelikkaya/datascience-salaries-2024). Scraping job boards goes against their terms of service so it will not actually be included in the dataset, however I will add it as a part of steps for data collection and handling to to simulate what the process would look like but will not implement it in the actual project. Some manual entry of data from these boards will be used to add more usable data.

* - *Denotes pieces of the project that will not be implemented but explained*

1. Data Collection
- Extract data from Kaggle data set
- Inject random unrelated variables/missing variables into dataset (for data cleaning)
- Gather data from job boards *
    - Including job board listings would add more validity to the data set and add more data to interpret

2. Data Handling

- Error handle potential defects from csv file
- Load data into Pandas dataframe
- Remove extraneous data
- Handle missing values in Pandas by removing data that cannot be interpreted
- Delete duplicate data from job boards*
    - Job boards have many incorrect values and variables, many of which are not important for interpretation or completely invalidate the dataset. For example, job postings for internships that are actually full-time jobs and vice-versa can be incorrectly listed, leading to incorrect data and models.

3. Databases

- Design SQL database to perpetually store collected data
    - Create Job_Posting to store data of job market, contains data such as posted date and job type * (Will be entered into the SQL database but will not contain any data inside)
    - Create Historic_Jobs to store data of disclosed salaries and relevant jobs information
- Create structure to store collected data into SQL
- Load existing data into the new database
- Verify correctness of loaded data
    - Checks to make sure all the data entered into SQL is accurate, to test this, query SQL and check the first 3 lines are entered correctly

3. Modeling

- Calculate mean and variance for data
- Remove data outliers
- Create Z-Score normalization of data
- Split dataset into testing and prediction sets
- Use linear regression to calculate prediction model
- Visualize current data
    - Show the job markets with data science

4. Analysis

- Use model to predict salaries over the next 5 years
- Visualize prediction showing both historic and new data points
    - Create graphs to visualize the growth of jobs and salaries in the industry
    - Show the countries with a high demand for data science employees
- Produce R2 Scores
- Calculate the Mean Squared Error
- Compare the historical data to the current prediction model