Final Project Summary

****

College of Science and Technology, Bellevue University

DSC540-T301: Data Preparation (2233-1)

Sashidhar Bezawada

March 02, 2023

Final Project Summary (Milestone 5)

As a part of my project deliverables, I have considered three datasets regarding post-graduate earnings for the students , university acceptance rate and College Scorecard. The three datasets are linked based on the university name.

Dataset and Sources:

CSV file: Data source as flat file

https://www.kaggle.com/datasets/wsj/college-salaries?select=salaries-by-region.csv
https://www.kaggle.com/datasets/wsj/college-salaries?select=salaries-by-college-type.csv

Description: The two datasets were obtained from the Wall Street Journal based on data from Payscale, Inc. They represent universities by region & college type and the post-graduate earnings for the students at these various universities.

To complete Milestone 2,

- The exercises for weeks 3 and 4 helped me a lot on how to extract data from CSV files and perform data cleansing techniques. I have completed the below steps to finish up the task.
- I have renamed few column names to make it more meaningful to understand and deduped the complete dataset.
- I have subjected the data set to pandas, profiling to identify the correlation between variables missing values and outliers in the dataset.

Website : Data source as website

https://oedb.org/rankings/acceptance-rate/

Description: The tabular layout in this website represents various numerical factors for universities in the United States. These factors include: student-to-faculty ratio, graduation rate, retention rate, acceptance rate, enrollment rate, school aid rate and default rate.

To complete Milestone 3, I have learned and performed the following:

- The exercises for weeks 4 and 5 helped me a lot on how to extract data from HTML sources and perform data parsing techniques. I have completed the following steps to finish up the task.
- I have imported the HTML data and get it into a more readable format.

- I have removed other redundant data which is not required for my use case.
- To match the file more manageable, I will only keep the columns needed, renamed columns , Filling Missing values.

API : Data source as website

https://collegescorecard.ed.gov/data/documentation/

Description: College Scorecard provides data at the institution-level and by field of study. It is a resource for prospective students to utilize for searching many degrees of information on colleges/universities and assessing their fit.

To complete Milestone 4, I will conduct the following:

- I have converted the returned data into JSON format for ease of access.
- I have renamed and dropped a few of the columns which are not required for my use case.
- I have organized the retrieved data into a data frame.

Milestone 5 :

I have loaded each of the above dataset into sqlite database and relationship between all the datasets is based on **University name.**

The CSV file has a 1:1 relationship with the API, and the website also has a 1:1 relationship with the API, as they contain one record per university and the API does as well.

CSV	: School Name
Website	: School Name
API	: Institution Name

I have created the below visualizations:

- Scatter Plot Between Graduation Rate and Starting Median Salary
- Median Graduation Rates Per Region Bar Plot
- Average Overall SAT Scores per Region per School Type Bar Plot
- Northeastern Universities' Boxplots of Starting & Mid-Career Salaries
- Histogram of 4-Year Retention Rates
- Pie Chart of School Ownership Categories

Ethical Implication of data from the Website data

Accountability of the College rankings is not considered as part of the data collection via the Website data sources used in this dataset.