CS 457 – Data Science Final Project

Please submit on LMS

Select one dataset from the list below and write 8-10 page Final Analytics project report.

1. Consumer Complaint Database (Data contains complaints received about financial products and services such as bank account, credit card, loans etc. Good for NLP as well.
https://catalog.data.gov/dataset/consumer-complaint-database

2. Department of the Treasury Income Tax Return (This study provides detailed tabulations of individual income tax return data at the state and ZIP code level. You can use the most recent year data only for analysis, or you can compare two consecutive years in your analysis)
https://catalog.data.gov/dataset/zip-code-data

3. College Scorecard Data (College Scorecards make it easier for students to search for a college that is a good fit for them. They can use the College Scorecard to find out more about a college's affordability and value so they can make more informed decisions about which college to attend)
https://catalog.data.gov/dataset/college-scorecard

4. Immunization Data by School This data is based on counts of students by immunization status in all grades (kindergartner through 12th grade) in public and private schools for 2016-2017. The student immunization status is based on parent reports to schools and may not be verified by a healthcare provider.
https://catalog.data.gov/dataset/all-students-kindergarten-through-12th-grade-immunization-data-by-school-2016-2017

5. NCDC Storm Database (Contains chronological listing, by state, of hurricanes, tornadoes, thunderstorms, hail, floods, drought conditions, lightning, high winds, snow, temperature extremes, statistics on personal injuries and damage estimates. Storm Data covers the United States of America)
https://catalog.data.gov/dataset/ncdc-storm-events-database

Please follow the same structure below in your report and make sure to include all the sections with detailed briefing on each of them.

Introduction (about overall project)

- Briefly explain the project and its scope

Who (company, agency, organization) collected the data?

- Who they are, what do they do?
- What is their role/purpose?

Need

- Why did they collect this data?
- What potential questions could be answered by studying this data?
    - List some specific questions, and be sure to answer them in your analysis
- Is there any privacy, quality, or other issues with this data?

Requirements and Resources needed

- What software and hardware resources you have used in this project?

Dataset Description

- Briefly describe the dataset
- Prepare and describe relevant metadata (types of attributes/variables in the dataset)
- Include schema for SQL if you are using any database. Schema should contain data types and sample values of columns/attributes. If you are using dataset in csv, xls or txt file, then you need to describe data types and sample values of all the columns/attributes in the data file. That will be considered as your dataset schema.

Results/Findings

- Explore the dataset using relevant tools discussed in the course such as R, SQL, Python, Tableau etc. Prepare relevant descriptive statistics and visualizations for selected data
    - **You do not need to analyze all the columns in the dataset**

- Include your results for each of the following analysis. You can add more than one analysis for each of them but at least one analysis on each of them is required
    - Use different set of variables/columns for each analysis. Do not use same set of variables in those analysis.

    - Scatterplot
    - Boxplot
    - Hypothesis test
    - Regression/Correlation analysis
    - Unsupervised/Supervised Learning method

- Interpret the results
    - **What conclusions can be supported for each analysis?**
    - This should reflect answers to the specific questions specified in the "Need" section.

- Graphics must follow good visualization practices discussed in course lectures.

Explain/define terms

- Include explanation of any technical terms relevant to the project

References

- Provide appropriate citations and references
    - Include citation for the dataset

Submission

- **<u>Submit the report and code files only. Do not submit the dataset.</u>**
- Provide the link to download the dataset used in the report in case instructor needs to download it. Do not submit the dataset (due to large size) with report on blackboard.

If you have a specific area of interest (security, social media, system performance, commerce, etc.) and want to work on dataset in your area of interest then send an email to Instructor Zeehasham Rasheed (zeehasham.rasheed@sse.habib.edu.pk) indicating the subject of your Final Project and Dataset for approval.