Homework 1

Due: 9/20/2024

Content Covered

Data science on small data set

1. General Homework Requirements

Work Environment: This homework must be written in **Python**.

Programming: Relevant tutorials and resources are linked below to aid in the programming portion of this homework.

Academic Integrity: You will get an automatic F for the course if you violate the academic integrity policy.

Note: This homework is an **individual** assignment. You are not permitted to work with anyone else on this assignment. All work submitted must be yours and yours alone.

2. Overview

The objective is to walk through the process of data cleaning, EDA and some elementary analysis using Jupyter Notebook Python. You will be working with **LAPD's crime data since 2020 to present from USA's data.gov** catalog to understand how to preprocess data and derive meaningful insights.

3. Install a Python notebook IDE

There are several IDEs that support .ipynb format. The popular ones are Jupyter Notebook, JupyterLab and VSCode:

Jupyter Notebook: https://github.com/jupyterlab/jupyterlab-desktop

JupyterLab: https://github.com/jupyterlab/jupyterlab-desktop

VSCode: https://code.visualstudio.com/docs/datascience/jupyter-notebooks

Installing any of the above IDEs is ok. There can be other options as well. Please test your installation before moving onto the next part. **Make sure that your application is able to export a PDF file from a notebook file.**

4. Getting familiar with pandas, numpy and matplotlib

Basic data science packages include pandas, numpy and matplotlib. Please check out the official documentation of these packages

pandas: https://pandas.pydata.org/docs/

numpy: https://numpy.org/doc/stable/

matplotlib: https://matplotlib.org/stable/index.html

5. Tasks

Below is the URL address of the LAPD's crime data:

https://data.lacity.org/api/views/2nrs-mtv8/rows.csv?accessType=DOWNLOAD

For a deeper understanding of the features in the dataset, refer to the following link:

https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data

Create a Jupyter notebook to code and respond to the following tasks: (100 pts total)

a. **(5 pt)** Load the data into the DataFrame using the URL. (You loose points if reading from local folder)
b. Initial Exploratory and visualization:
    i. **(5 pt)** Print the metadata of column information.
    ii. **(10 pts)** What is the total number of crimes committed according to the description of the crime code? Make a visualization using just one graph that shows a distribution of several crimes.
    iii. **(10pts)** Make a visualization to suggest highest crime prone areas. You may plot multiple graphs.
    iv. **(10pts)** Make a visualization to warn general public about the trend crimes according to the time of crime occurence, sex and age of victim and the area in which it can occur. You may plot multiple graphs.
c. Investigating Patterns of Vehicle Thefts in Los Angeles:
    i. **(10pt)** Apply conditions to make it a valid problem statement. Also provide features which you think are important according to your problem statement.
    ii. **(10 pts)** Explain your approach to your problem statement.
    iii. **(10 pts)** Perform data cleaning to get the pure data for this problem. Explain your data cleaning steps.(At least 3 cleaning steps)
    iv. **(10 pts)** Implement your approach to this problem and justify your hypothesis.
d. Exploring cases in Identity Theft Cases in Los Angeles:
    i. **(10 pts)** Explain your approach to this problem. Also provide features which you think are important according to your problem statement.

ii. **(10 pts)** Perform data cleaning to get the pure data for this problem. Explain your data cleaning steps.

6. Submission

The submission includes two files: **the .ipynb file and the .pdf file exported from the snapshot of running all sections of your notebook file.**

Name the files as: **hw1_ubpersonnumber.ipynb** and **hw1_ubpersonnumber.pdf**.

For example, hw1_57333333.ipynb.

Submit the files to UBLearns.