Homework 1

Due: 9/19/2024 23:59 p.m.

Content Covered

Data science on small data

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.

Teams: 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 a COVID-19 dataset 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 owid covid data:

https://covid.ourworldindata.org/data/owid-covid-data.csv

Create a Jupyter notebook to code and respond to the following tasks: (30 pts total)

a. **(1 pt)** Load the data into the DataFrame structure using the URL. (Lose points if reading from local folder)
b. Initial Exploratory and visualization
    i. **(1 pt)** Print the metadata of column information.
    ii. **(2 pts)** What is the total number of infection and death cases for each country? Make a visualization using just one graph that shows the top 10 death cases accompanied by infection cases.
    iii. **(2 pts)** Make a visualization to suggest vaccination to old people. You may plot multiple graphs.
    iv. **(2 pts)** Make a visualization to warn your neighborhood about the trend of covid. You may plot multiple graphs.
c. How effective is the vaccination?
    i. **(1 pt)** Apply conditions to make it a valid problem statement.
    ii. **(2 pts)** Explain your approach to your problem statement.
    iii. **(2 pts)** Perform data cleaning to get the pure data for this problem. Explain your data cleaning steps.
    iv. **(3 pts)** Implement your approach to this problem and justify your hypothesis.
d. How long does the virus across all variations take on average to kill a person if it does kill a person after infection?
    i. **(2 pts)** Explain your approach to this problem.
    ii. **(3 pts)** Perform data cleaning to get the pure data for this problem. Explain your data cleaning steps.
    iii. **(3 pts)** Implement your approach and compute the estimate.
e. **(6 pts)** What are your other findings from the dataset? Name one of your findings. State the problem then analyze it using the data.

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_[your_ubnumber].ipynb** and **hw1_[your_ubnumber].pdf**.

For example, hw1_57333333.ipynb.

Submit the files to UBLearns.