Project Description

15-110 - Principles of Computing

Due: 18th April, 2022

Checkpoint: 13th April, 2022

1 Data analysis notebook

In this project you need to develop a python notebook (using Jupyter) containing an analysis derived from a dataset (in a `.csv` format). That is, once you select a dataset, you have to design a set of questions (or hypotheses) that that data allows you to investigate. Then you will use python to extract and process the relevant data, and create informative graphs and summaries. The compiled information will allow you to reach some conclusions.

Your notebook must contain the documentation of the project, code, and any graphs you have used to reach your conclusions. In a very real sense you are using Jupyter to write a report that includes:

- a discussion of the data;
- hypotheses about what you expect to find out;
- code to analyze the data;
- code to graph the data;
- the graphs themselves;
- and your conclusions.

Read carefully the description of each of these elements below and make sure your notebook contains all the required information before submitting.

1.1 Graphs

You must generate (at least) **3 different types of graphs** from the dataset of choice. Typically each graph shows the data used to test a different hypothesis. If you feel it will be useful, you may plot multiple graphs for one hypothesis.

A good approach is to come up with a first hypothesis for the data, extract or generate the relevant information from the dataset, and plot a graph to check if the hypothesis is confirmed or not. If it is, you can come up with another hypothesis and/or refine the graph for extracting more information. If it is not, you can generate other graphs to find out why, and so on. For instance, if your dataset is about food and inlcuded various information about food consumption across the world, as a first hypothesis you might want to propose that in rich countries there is a larger consumption of meat compared to poorer countries. To test your hypothesis you will have to extract the relevant data from the dataset and use it to to make a plot representing the distribution of meat consumption versus GDP. You will then analyze the data and draw some conclusions that you will document in the notebook together with code and the plot(s). Based on the results, you will move on making another hypothesis about food consumption, and so on, until you have (at least) 3 graphs of different types, each used for testing a different hypothesis.

Note that a graph can be in the form of a **plot (with points or a line, histogram, bar plot, or pie chart**. You are welcome to make more than 3 graphs, of course :)

1.2 Documentation

The documentation must contain:

- Title
- Author (i.e., your name)
- Description of the dataset (what it is about, where it was obtained, the data it contains, etc.)
- Explanation of how the data is represented in python (do you use lists, dictionaries, tuples? What are keys and values? etc.)
- For each graph generated, a description of what you are trying to find out and how the information needed for the graph can be obtained from the data.
- For each graph generated, an explanation of the conclusions you have reached, if they confirm your expectations, why or why not.

Your explanations must be clear, objective, and concise, and this may take longer than you think, so do not leave it to the last day! Document as you go and refine your text if on the next day you cannot make sense of it anymore.

Note that the text of the documentation *must be nicely formatted using Markdown* (e.g., use of lists, use of italic and boldface characters to emphasize text when necessary, appropriate use of space between paragraphs, use of sectioning). The use of a sloppy, unformatted presentation text will be penalized at grading time. Last but not least, make sure to check your spelling!

It is strongly suggested to *organize the notebook in sections*, where each hypothesis is discussed in a separate section.

1.3 Code

In your code, the dataset file must be read once and stored in some kind of data structure in Python. You must *not* open and read the file each time you are creating a new graph (this will be considered as a major error).

All the code submitted should work properly and generate the graphs in the notebook. In particular, if you click on `Cell` → `Run all` no errors should appear and all output should be generated on the fly.

It goes without saying that the code must be written following a good style (i.e., meaningful variable and function names, good spacing in the statements, use of inline comments, use of functions to make the code as modular as possible and avoid code repetitions). Please revisit the style document[1] on the course website.

Think of a good way to split your code into cells, so there is a nice balance of text, code and graphs. Use comments when necessary to explain what each part does.

2 Data

The dataset is your source of information. Spend some time to make a good choice of the dataset that you are going to analyze. Below you will find a number of datasets that have been selected for you as possible choices.

You are not limited to choose among the datasets below. Indeed, you can choose your own dataset, but this needs to be approved by the course instructor in advance.

In the list below, the first link contains a brief description of the dataset and the second link is for downloading the file. Note that Kaggle requires a login to download the files. Also note that many datasets include an extensive number of columns and files. You do not have to use all of them, but you should select the subset of features that are relevant to test your hypotheses.

[1] `https://web2.qatar.cmu.edu/cs/15110/resources/style.pdf`

IMPORTANT: Since there will be presentations, each student must choose a different dataset. In order to ensure this, you must add your name as a comment to the cell corresponding to your choice in the following spreadsheet:
https://docs.google.com/spreadsheets/d/1xU7t76QOIWbBsnUyWu4gYGf9sbTCdc98gRSHG9-mAOs/edit?usp=sharing
The selection is first-come-first-serve.

1. COVID-19 World Vaccination Progress
   https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress
   https://web2.qatar.cmu.edu/cs/15110/datasets/covid-19-world-vaccination-progress.zip
2. COVID-19 Variants Worldwide Evolution
   https://www.kaggle.com/datasets/gpreda/covid19-variants
   https://web2.qatar.cmu.edu/cs/15110/datasets/covid-19-variants-worldwide-evolution.zip
3. COVID-19 Healthy Diet Dataset
   https://www.kaggle.com/datasets/mariaren/covid19-healthy-diet-dataset
   https://web2.qatar.cmu.edu/cs/15110/datasets/covid-19-healthy-diet-dataset.zip
4. Heart Failure Prediction Dataset
   https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
   https://web2.qatar.cmu.edu/cs/15110/datasets/heart-failure-prediction.zip
5. Tesla Daily Stocks Prices
   https://www.kaggle.com/datasets/timmofeyy/-tesla-daily-stocks-prices
   https://web2.qatar.cmu.edu/cs/15110/datasets/tesla-daily-stocks-prices.zip
6. IBM Real Time Stock Analysis
   https://www.kaggle.com/datasets/bhanuprasanna527/stock-market-prediction
   https://web2.qatar.cmu.edu/cs/15110/datasets/ibm-real-time-stock-analysis.zip
7. Bitcoin Data
   https://www.kaggle.com/datasets/varpit94/bitcoin-data-updated-till-26jun2021
   https://web2.qatar.cmu.edu/cs/15110/datasets/bitcoin-data.zip
8. Credit Card Approval Prediction
   https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction
   https://web2.qatar.cmu.edu/cs/15110/datasets/credit-card-approval-prediction.zip
9. World Happiness Report up to 2022
   https://www.kaggle.com/datasets/mathurinache/world-happiness-report
   https://web2.qatar.cmu.edu/cs/15110/datasets/world-happiness-report-up-to-2022.zip
10. World Sustainability Dataset
    https://www.kaggle.com/datasets/truecue/worldsustainabilitydataset
    https://web2.qatar.cmu.edu/cs/15110/datasets/world-sustainability-dataset.zip
11. Human Freedom Index
    https://www.kaggle.com/gsutters/the-human-freedom-index
    https://web2.qatar.cmu.edu/cs/15110/datasets/the-human-freedom-index.zip
12. Global Human Trafficking
    https://www.kaggle.com/datasets/andrewmvd/global-human-trafficking
    https://web2.qatar.cmu.edu/cs/15110/datasets/global-human-trafficking.zip

13. Brazilian Amazon Rainforest Degradation 1999-2019
https://www.kaggle.com/datasets/mbogernetto/brazilian-amazon-rainforest-degradation
https://web2.qatar.cmu.edu/cs/15110/datasets/brazilian-amazon-rainforest-degradation-1999-2
zip

14. Global Seawater Oxygen-18 Levels
https://www.kaggle.com/datasets/tjkyner/global-seawater-oxygen18-levels
https://web2.qatar.cmu.edu/cs/15110/datasets/global-seawater-oxygen-18-levels.zip

15. India Sub-division Rainfall 1901-2015
https://www.kaggle.com/datasets/kkhandekar/statewise-rainfall-1901-to-2015
https://web2.qatar.cmu.edu/cs/15110/datasets/indian-sub-division-rainfall-1901-2015.zip

16. Weather Conditions in Seattle
https://www.kaggle.com/datasets/ananthr1/weather-prediction
https://web2.qatar.cmu.edu/cs/15110/datasets/weather-conditions-in-seattle.zip

17. 9000+ Movies Dataset
https://www.kaggle.com/datasets/disham993/9000-movies-dataset
https://web2.qatar.cmu.edu/cs/15110/datasets/9000-movies-dataset.zip

18. Oscar Best Picture Movies
https://www.kaggle.com/datasets/martinmraz07/oscar-movies
https://web2.qatar.cmu.edu/cs/15110/datasets/oscar-best-picture-movies.zip

19. Indian Premier League 2008-2019
https://www.kaggle.com/datasets/nowke9/ipldata
https://web2.qatar.cmu.edu/cs/15110/datasets/indian-premier-league-2008-2019.zip

20. FIFA Football World Cup Dataset
https://www.kaggle.com/datasets/iamsouravbanerjee/fifa-football-world-cup-dataset
https://web2.qatar.cmu.edu/cs/15110/datasets/fifa-football-world-cup-dataset.zip

As pointed out above, you are not limited to choose among the above datasets. Indeed, you can choose your own dataset. Good places to find datasets include:

- https://www.kaggle.com/datasets
- http://vincentarelbundock.github.io/Rdatasets/datasets.html
- https://archive.ics.uci.edu/ml/index.php

If you choose your own dataset, note that it must have at least 1000 rows and 5 columns, and you will also need to get the **the instructor's approval** before you start working on it.

3 Examples

You may be inspired by looking through some Jupyter notebooks created by others. Visit https://github.com/jupyter/jupyter/wiki for an entire page of links to Jupyter notebooks on a variety of topics.

4 Submission

Your **notebook (.ipynb file) and dataset (.csv file(s))** used must be zipped in one file and uploaded to gradescope before the deadline.

5 Checkpoint

One week before the final submission, you must attend a project checkpoint meeting with the CAs. These will take place on a Wednesday, at the same time as your homework interviews. The goal is for your to present your progress to the CA, receive comments on how you can improve your notebook, and solve any issues you may be having.

This checkpoint meeting counts for 10% of the project grade.

6 Presentation

After submitting your project, you need to present it to the class. Your will have *5 minutes* to show off your notebook, explaining which dataset you have chosen, and your most interesting or surprising hypothesis and conclusion.

The presentation counts for 10% of the project grade.

7 Distribution of points

This project is graded out of **100 points** distributed as follows:

- Checkpoint: 10%
- Graphs: 30%
- Code: 20%
- Documentation: 30%
- Presentation: 10%