Project Description

15-110 - Principles of Computing

Due: 9th November, 2022, 6:00 pm

Checkpoint: 3rd November, 2022, 10:00 pm

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 the end, 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 helpful, 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 includes various information about food consumption across the world, as a first hypothesis you might want to propose that in rich countries, there is a more significant 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 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 with the code and the plot(s). Based on the results, you will move on to 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, explain the conclusions you have reached: if they confirm your expectations, and why or why not.
- Citations (see below).
- Additional help received (see below).

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 you cannot make sense of it the next day.

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). 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 data structure in Python. You must *not* open and read the file each time you create a new graph (this will be considered 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.

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, with 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 making a good choice of the dataset that you are going to analyze. Below you will find several datasets selected for you as possible choices.

You are not limited to choosing 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

[1] https://web2.qatar.cmu.edu/~mhhammou/15110-f22/resources/style.pdf

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 relevant features to test your hypotheses.

IMPORTANT: Each student must choose a different dataset. 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/1SVhznpOEZ2LtcLm0WYZ6vMrAvtRSKTwi1AFqWVrH5VA/edit?usp=sharing
The selection is **first-come-first-serve**.

1. COVID-19 World Vaccination Progress
   Description
   Dataset
2. COVID-19 Variants Worldwide Evolution
   Description
   Dataset
3. COVID-19 Healthy Diet Dataset
   Description
   Dataset
4. Heart Failure Prediction Dataset
   Description
   Dataset
5. Tesla Daily Stocks Prices
   Description
   Dataset
6. IBM Real Time Stock Analysis
   Description
   Dataset
7. Bitcoin Data
   Description
   Dataset
8. Credit Card Approval Prediction
   Description
   Dataset
9. World Happiness Report up to 2022
   Description
   Dataset
10. World Sustainability Dataset
    Description
    Dataset
11. Global Human Trafficking
    Description
    Dataset
12. Brazilian Amazon Rainforest Degradation 1999-2019
    Description
    Dataset
13. Global Seawater Oxygen-18 Levels
    Description
    Dataset

14. India Sub-division Rainfall 1901-2015
    Description
    Dataset

15. Weather Conditions in Seattle
    Description
    Dataset

16. 9000+ Movies Dataset
    Description
    Dataset

17. Oscar Best Picture Movies
    Description
    Dataset

18. Indian Premier League 2008-2019
    Description
    Dataset

19. FIFA Football World Cup Dataset
    Description
    Dataset

20. Dog Adoption
    Description
    Dataset

21. Flight Delays
    Description
    Dataset

22. World Energy Consumption
    Description
    Dataset

23. COVID-19 Education Impact Survey
    Description
    Dataset

24. Avocado Prices
    Description
    Dataset

25. International football results from 1872 to 2020
    Description
    Dataset

26. Human Resources Data Set
    Description
    Dataset

27. Human Freedom Index
    Description
    Dataset

28. Hospital filing records
    Description
    Dataset

29. Dengue by regions
    Description
    Dataset

30. Car accidents
    Description
    Dataset

31. Treatment of migraine
    Description
    Dataset

32. Occurrence of pneumonia in children
    Description
    Dataset

33. University instructor evaluations
    Description
    Dataset

34. Students Adaptability Level in Online Education
    Description
    Dataset

35. DataCo Supply Chain Data
    Description
    Dataset

As pointed out above, you are not limited to choosing 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 various topics.

4 Python Modules

You are not allowed to use external Python modules apart from those demonstrated in class and in the lecture notes.

5 Online Resources

You may consult any materials from any sources you may discover online. However, you must very clearly cite each such use, so it is very clear what is yours and what is not, and in the latter case where the materials came from. We will grade you only on your original contributions, and we will penalize use of external materials without citation. Note that you must also clearly cite code that comes from the course notes! Non-cited external code will receive a penalty on the checkpoint, a severe penalty on the final submission, and will be investigated as potential cheating cases.

6 Additional Help

This project is solo. You may not collaborate or discuss it with anyone outside of the course, and your options for discussing with other students currently taking the course are limited. See the academic honesty policy for more details.

If you receive additional help, you should mention it on the notebook (e.g., in the documentation); you must include the names of the students and staff that you consulted on this project.

7 Checkpoint

The checkpoint submission consists of the following:

- the **notebook (`.ipynb` file)**
- **the dataset (`.csv` file(s))** used

These files must be zipped in one file and uploaded to gradescope before the corresponding deadline.

In the checkpoint, you should demonstrate that you can read the dataset files and store the data in appropriate data structures. You should also describe a concrete plan for your hypotheses and the graphs you intend to make. This plan should be explained in text using appropriate markdown cells.

7.1 Checkpoint meeting

You must attend a project checkpoint meeting with the CAs before the checkpoint deadline. These can take place on Wednesday and Thursday. The goal is for your to present your progress to the CA, receive comments on how you can improve your notebook and your plan, and solve any issues you may be having before submitting the checkpoint.

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

8 Submission

The final submission consists of the following:

- the **notebook (`.ipynb` file)**
- **the dataset (`.csv` file(s))** used

These files must be zipped in one file and uploaded to gradescope before the corresponding deadline.

9 Distribution of points

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

- Checkpoint: 10%
- Graphs: 30%

- Code: 20%
- Documentation: 30%
- Style and aesthetics: 10%