15-110: Principles of Computing

PROJECT

Due: 3rd December, 2020 at 23:59

Data analysis notebook

In this project you need to develop a python notebook (using Jupyter) containing an analysis derived from a dataset. 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 discussion of the data, some hypothesis about what you expect to find in the data, code to analyze the data, code to graph the data, and the graphs themselves.

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

Graphs

You must generate **at least 4 graphs** from the dataset of choice. A good approach is to come up with a hypothesis for the data, and plot a graph to check if it 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.

Documentation

The documentation must contain:

- Title
- Author (i.e. your name)
- Description of the dataset (What is it about? From where was it obtained? What data does it contain? 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 conclusion you have reached, if it confirms your expectations, why or why not.

Your explanations must be clear, objective and concise, and this may take longer than you think, so don't 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.

Code

The dataset file should be read once and stored in some kind of a data structure in Python. You should *not* open and read the file each time you are creating a new graph.

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 **it must follow the style guidelines**. 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.

Data

The datasets below are already approved by course staff for this project. The first link contains a brief description of the dataset and the second link is for downloading the file.

- COVID-19 Complete Dataset
  `https://www.kaggle.com/imdevskp/corona-virus-report`
  `https://web2.qatar.cmu.edu/~mhhammou/15110-f20/datasets/covid_19_clean_complete.csv`
- Avocado Prices
  `https://www.kaggle.com/neuromusic/avocado-prices`
  `https://web2.qatar.cmu.edu/~mhhammou/15110-f20/datasets/avocado.csv`
- International football results from 1872 to 2020
  `https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017`
  `https://web2.qatar.cmu.edu/~mhhammou/15110-f20/datasets/football-results.csv`
- Human Resources Data Set
  `https://www.kaggle.com/rhuebner/human-resources-data-set`
  `https://web2.qatar.cmu.edu/~mhhammou/15110-f20/datasets/HRDataset_v14.csv`
- Human Freedom Index
  `https://www.kaggle.com/gsutters/the-human-freedom-index`
  `https://web2.qatar.cmu.edu/~mhhammou/15110-f20/datasets/Human-Freedom-Index.zip`
- World Happiness Report
  `https://www.kaggle.com/unsdsn/world-happiness`
  `https://web2.qatar.cmu.edu/~mhhammou/15110-f20/datasets/World-Happiness-Report.zip`
- Hospital filing records
  `http://vincentarelbundock.github.io/Rdatasets/doc/COUNT/medpar.html`
  `https://web2.qatar.cmu.edu/~mhhammou/15110-f20/datasets/medpar.csv`
- Dengue by regions
  `http://vincentarelbundock.github.io/Rdatasets/doc/DAAG/dengue.html`
  `https://web2.qatar.cmu.edu/~mhhammou/15110-f20/datasets/dengue.csv`
- Car accidents
  `https://vincentarelbundock.github.io/Rdatasets/doc/DAAG/nassCDS.html`
  `https://web2.qatar.cmu.edu/~mhhammou/15110-f20/datasets/nassCDS.csv`
- Treatment of migraine
  `https://vincentarelbundock.github.io/Rdatasets/doc/carData/KosteckiDillon.html`
  `https://web2.qatar.cmu.edu/~mhhammou/15110-f20/datasets/KosteckiDillon.csv`
- Occurrence of pneumonia in children
  `https://vincentarelbundock.github.io/Rdatasets/doc/KMsurv/pneumon.html`
  `https://web2.qatar.cmu.edu/~mhhammou/15110-f20/datasets/pneumon.csv`

Alternatively, you can choose your own dataset but it must have at least 1000 rows and 5 columns. Additionally, **you must get the approval of the instructors for whatever dataset you choose by November 26th**.

Here are some suggested good places to find datasets:

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

Examples

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

Submission

Your notebook and dataset used must be zipped in one file and uploaded to Autolab before the deadline.

Distribution of points

This project is graded out of **100 points** (the same as roughly two homeworks), distributed as follows:

- Graphs: 40 points
- Code: 20 points
- Documentation: 40 points