Assignment 1

Submission information

Due date: Monday, November 27th, 2023. Submission closes at 11:59pm.
Format: 1x Jupyter notebook (.ipynb)

Description

In this assignment, you will examine a data science problem of your choice. What data set and analyses you choose to do are up to you. Your project must satisfy the following requirements:

Requirements

1. Incorporate the key aspects of the data science workflow.
2. Explore one or more questions.
3. Be communicated in the form of a Jupyter Notebook.

Requirement 1

Your project must include the essential aspects of the data science workflow. These include:

- Locating data (see below for some data repositories)
- Importing data (into Python)
- Cleaning data (optional/if needed)
- Exploratory data analysis to explore relationships. Ideally, this will include a summary table.
- Communicate your process and findings in the form of written text, that includes at least 1 generated image/plot.

We will discuss the data science workflow (see Figure 1) in class.

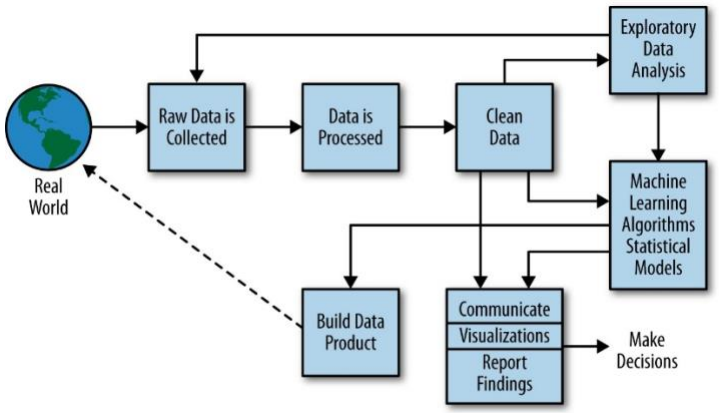

Figure 1. Data science workflow. From Doing Data Science, 1st Edition by C. O'Neil & R. Schutt O'Reilly Media, 2014.

Requirement 2

Your project should explore at least one question or issue that is of interest to you, from your chosen dataset. For example, you download a dataset from the waterways department, and after looking at the data, you become interested in the relationship between rainfall levels and pollution measures in local waterways. A question often means exploring a relationship between two or more variables (e.g., a person's height, and time).

Requirement 3

Your submission must be in the form of a single Jupyter workbook (.ipynb) file. Use Markdown to communicate to the reader. It should make effective use of Markdown's formatting syntax to delineate each section of the notebook (i.e. use headings, bold etc).
It must be structured to contain the following elements:

1. **Title** – Give your assignment a title. Include your name and ID.
2. **Introduction** – Provide the reader with some background on your question and data. Where did you get your data? What question looked interesting given the data you chose? This section should be between 150-500 words.
3. **Analyses** – This is where you perform your Python coding. Use Markdown to explain the major steps (e.g., data loading, cleaning, manipulation, generated plot).
4. **Discussion** – Summarize the numbers, graphs, and tables from the analyses section. What does it all mean? For example, you may have found a relationship between heat waves and violent crime frequency. What might this mean?
5. **References** – Must contain a link to where you go the data, and any additional packages you used (if any).

Source data

Your assignment must use an external data file in .CSV format. When grading, teaching staff will "run" your Jupyter notebook. For this to happen, you must provide the data for the notebook to work. You may do this in one of two ways:

1. Upload your CSV, along with your .ipynb file, to Canvas. Your notebook must be written to load the CSV file from the same local directory as your notebook file. Do not link to some other folder on your personal computer, as it it will break on a different machine.
2. Load the CSV file via a publicly-accessible URL

Example 1:

```
# Load data from local folder
d = pd.read_csv('TitanicSurvival.csv')
```

Example 2

```
# Load data from public URL (no account or login required)
d = pd.read_csv('https://somefreedata.github.io/data/TitanicSurvival.csv')
```

Data repositories

The following is a short list of websites that offer open access datasets for you to download and explore:

- Kaggle [url]
- UC Irvine ML repo [url]
- Canada Open Data [url]
- Data.gov [url]
- Microsoft Open Data [url]
- Open AWS [url]

When looking for a data, keep it simple. The file shouldn't be too big (i.e. < 5 Mb), and should contain a mixture of categorical (e.g., red, black, green), and continuous variables (e.g., -234.34, 84). It should have column headers, and a description of what the data is so you can understand what you're looking at. If it looks complicated or problematic, find another dataset.

Projects to avoid

Your project should have some original thought. The following projects are not permitted:

- Those using the Titanic dataset or any other dataset used in class.
- Minor extensions of something you found elsewhere (e.g., on Kaggle).
- Minor extensions of projects done in class.

Use of LLMs

You may use Large Language Models (e.g., ChatGPT) in your assignment. As noted in L01, you must include the prompts used, and the context in which you used it (e.g., To generate the introductory text, or the general problem).

Academic Integrity

Follow Ontario Tech's policy on academic integrity, and be aware of the reporting policy. We covered academic integrity in Lecture 01 - your submission should not involve collusion or plagiarism. Cite your sources. Using code from other sources with citation is acceptable if it's a few lines (<=5 lines), but wholesale copying of large code chunks is not permitted (> 5 lines). Email Dr Livingstone or Dr Lee if you're unsure about usage.

Grading rubric

Elements	Points
Markdown use and effectiveness	15
Data importing & cleaning	15
Exploratory analysis	15
Modelling	20
Graphic	15
Communication	20
Total	**100**