Project: Final

Python

Overview

Each student will choose a dataset related to a topic they are interested in. I have attached links to various options below. You will turn in a Jupyter notebook (.ipynb) that answers a data question using the dataset you choose.

There will be multiple checkpoints when you will turn in a portion of the final project. Each checkpoint will receive its own grade even though you are expected to include it in an expanded capacity with the final project.

There will be four (4) such checkpoints:

1. Dataset & Question (end Week 3)
   a) Choose a dataset.
2. Calculate Basic Statistics (end Week 7)
3. Data Visualizations (end week 10)
4. t-test and ANOVA test (after week 13)
5. Final due Date (Week 16)
6. Process Paper (Week 16)
   a) This paper should be written in Markdown and should explain your data problem and solution briefly. Then it should spend most of the time discussing

> **i Note**
>
> The Final Project is completed through connection with the GitHub classroom. Once you have created a GitHub account (with your Queens email), you will be able to join the classroom and access the assignments for the Final Project.
> See the instructions for connecting to the GitHub Classroom

Criteria

The final project must include the following to receive full credit

- Markdown explaining your problem, solution, and the process you went through to get the answer.
- At least 3 visualizations (of different types)
- At least 1 t-test
- At least 1 ANOVA test
- A written explanation of the results of the t-test and ANOVA test
  - The explanation should include the null hypothesis, the alternative hypothesis, the p-value, and the conclusion.
  - Explain why you chose the variable you did for the t-test and ANOVA test.

Possible Datasets By Area

You are not limited to these datasets. You can choose any dataset you would like. These are just some suggestions to get you started. Also, you have to look around and click around. The sites do not all look the same. Some are more user-friendly than others. Some have more data than others. You might need to combine data or clean data. Data analysis is about solving problems. Some problems will be related to aligning the data correctly.

Good places to search for datasets are:

- Kaggle Datasets

Business and Finance

- Yahoo Finance API
  - Yahoo Finance has an expansive API that allows you to pull in stock data with extensive amounts of information. This is actually a Python library that allows you to interact the the API more easily. It is available by default in Colab.
- World Bank Open Data
  - The World Bank has a variety of data on countries around the world. You can search for data by country, topic, and more.
  - The World Bank data has a significant amount of data.
-

Sports

- NFL Play-by-Play Data
  - This package allows you to pull in NFL play-by-play data. This data is very detailed and can be used to answer a variety of questions about the NFL. This is also a Python library that allows you to interact with the API more easily. It is not available by default in Colab. You can add it by running `!pip install nfl_data_py` in a code cell then import it.
- NFL Data
  - This dataset contains a variety of information about NFL games, including scores, betting lines, and more.
- World Cup 2018 Event Data
  - This data is available on my GitHub account. It was collected by Luca Pappalardo and Emanuele Massucco and is available on the figshare platform. There is also a GitHub repo explaining how to work with the data (here)[https://github.com/Friends-of-Tracking-Data-FoTD/mapping-match-events-in-Python]. I have cleaned the data to make it easier to work with. It has every event in every game of the world cup and where on the field it took place: passes, shots, challenges, etc. It comes in JSON format. You can find the raw data here. You will want to use the url for the raw data when you request the data in your Python script. This is super cool data of a similar granular quality as the NFL play-by-play data.

Climate and Environment

- NOAAA Climate Data Online
  - The National Oceanic and Atmospheric Administration has a variety of climate data that you can use for your project. You can search for data by location, date, and more.

Search for Other Data

- Kaggle Datasets
  - Kaggle has a variety of datasets that you can use for your project. You can search for datasets by topic or popularity.