Honors Project- Exploratory Data Analysis

Overview

Obtain 2-3 datasets online. These datasets must not be perfect (i.e., they can/should be messy in order to align with requirements). For these datasets, you must:

- Extract important variables and leave behind useless ones
- Identify outliers, missing values, or human error
- Understand the relationship(s), or lack of, between variables
- In the end, maximize your insights and
- With this, able to visualize and present your analyses in a neat format through a Notebook file.
- Describe the process and steps of their respective explorations.

The goal of this project is to extract unique insights from the respective datasets, sharpening data presentation skills, and encouraging experimentation.

Requirements

The data analyses require the following:

- A dataset that can be used to potentially demonstrate trends that translate to reality (e.g., COVID-19 data, climate change data, stock data, etc.)
- Using the *pandas* package in order to clean-up datasets, demonstrate and explain methods used for these.
- Add 2-3 columns with *pandas* to each dataset, that can add further insights to the analyses. In addition, explain how and why you created them.
- Plot 3-5 graphs e/a that provide significant insight to the respective datasets. They must all be unique.
- Finally, present everything neatly through a Notebook.
- *Optional requirement*: experiment with a data science package that could allow you to extract more findings from each dataset.

Schedule

The schedule below is designed to keep you on track. Feel free to go faster if you would like. We will meet

Fridays at 10:45 AM

- Feb 5 - Intro meeting
- Feb 19 - Students obtain 2-3 datasets they would like to use, demonstrate some exploration of each (e.g., Are they messy? Is the dataset sufficient for the project requirements? Has the student thought of additional columns for new insight?)
- Apr 2 - Students submit up to *pandas* data manipulation (new columns for each dataset)
- Apr 23 - Students present 2-3 Notebooks

Deliverables

Will be primarily graded on final product, grading parameters include: presentation neatness, good data analysis etiquette, uniqueness of insights, **experimentation**, functionality, etc.