BA222 Project 2 - Applied Multi-Variate Data Analysis in Python

Due: Tu, Nov 14 at 12PM

Fall 2023

Description

In this project, you will be identifying an interesting data set from the real world and using techniques from BA222 to analyze it. Like Project 1, this is an open-ended project. Please carefully read the guidelines below, as they are much more intricate than project 1.

Groups

Please work in **Groups of 4**. Your partners can be from any of Jetson's sections.

You must **change some of your partners** from Project 1. You may have at most 1 overlapping partner of your three partners from project 1.

All participants should put their name on the report and on the Jupyter notebook. All participants should upload and submit the report and spreadsheet to Questrom Tools; the course staff will only assign 1 grade but that will help us keep track.

Please plan to identify your partners no later than Oct 29 and submit them in this form:

`https://forms.gle/uoUUnU4HrX5sz3Lo6`

If you are unable to find a partner, please email the TA Nicholas Antonelli (nantonel@bu.edu) and he will help you find a partner. Sooner the better, and no later than Oct 29.

Data Sources

Data can come from any source you choose. Be wary of "messy" data, due to Python errors. You may need to clean the data manually. That is okay, but please describe the steps in your report. Good writing style dictates that

small nuances like data cleaning should be described later in the report, or in a footnote or appendix.

Please make sure your dataset has at least 2,000 observations and 10 variables. More is better, but please keep observations under a few million. In order to submit the data to the course staff, please make sure you can compress it to a 100MB file or find a way to transfer the dataset to the course staff (e.g. via Dropbox). Please do not use simulated data.

As with project 1, use data you are passionate about, and convey that passion in your analysis.

Expectations

You will conduct an analysis in a Jupyter notebook spreadsheet and produce a **4-6 page written report**. Submit both the notebook and the report. The report should be a standalone, readable document with embedded figures. Analyses left in Jupyter will not be counted. Please submit the report as a PDF.

Here are some guidelines for what this should entail. Reports are expected to be well written and detailed. Please plan to exceed these guidelines; this is just a roadmap of how a basic analysis can be conducted.

- Write an introduction for the reader about why this data set and topic are interesting and what open questions there are.
- Describe where you got your data set, as well as the features of the data.
- Describe the variables in the data set.
- How is the data quality? What cleaning processes did you undertake if any?
- Produce and describe summary statistics of the dataset, as well as one or more single-variable analyses with visualizations. Use what you learned in and since Project 1. Graphs should follow best-practice graphing as described in class. Graphs should be well-labeled and standalone. The text should reference the figure by number and describe it.
- **This project must include a multi-variable analysis component** based on the following few questions
- How do the data relate to each other? Are there any interesting correlations? Describe it in complete sentences to the reader.
- Create one or more tables using Groupby, and make plot(s) of these data. What do we learn from this table? Describe it in complete sentences to the reader.
- Run one or more regressions on the data. Write down the model. Interpret your regressions using full sentences, and explain their meaning to the

reader.

- What do your regressions tell you? Is there a causal relationship between variables?
- Create one or more multi-variate plots. Label the plots well, and describe them to the reader.
- Please include a paragraph at the end of the report (not counted toward page length) describing how each team member participated.