Final Project – Data Analytics

The purpose of this project is to learn to create analyses and visualizations that are useful in real life.

You are required to choose **one** dataset from four provided datasets. The datasets will be available for download on Apr 15. Complete all requirements outlined below for the dataset you have chosen.

Requirements

Your analysis should include:

1. At least two statistical summary (mean, sum, count, median etc).
2. At least two analyses of patterns, relationship etc. found in the data.
3. At least two visualizations - at least one for statistical summary, at least one for analysis.
4. A written summary of no less than 200 words describing the analysis and the results.

Using the ramen ratings dataset as an example,

1. A statistical summary of the average, min, and max ratings, the number of brands, styles, and countries in the datasets. The country that has the greatest number of ramen brands, etc.
2. Relationship between location of countries and the average ratings, relationship between brand and style and ratings, etc
3. A bar chart indicating the number of brands of each country (or region such as East Asia, Southeast Asia, Europe, etc), and an overlaying line chart that shows the average ratings of each country, etc
4. What can be concluded based on the analysis? Are ramens from a certain country / region more popular in general? Are brands doing better with a certain style in terms of ratings?

For those of you who are interested in text analysis, you might want to do some additional reading on some relevant tools and libraries. Recommended text analytics resources:

https://www.nltk.org/book/
https://www.youtube.com/watch?v=3lEsR57xDNY

Notes

1. No manual editing or modification of the original dataset (e.g. via Excel or other app) is allowed. All data manipulation needs to be done in the notebook with Python, pandas, matplotlib, or any other libraries that suit your needs.
2. You are not supposed to create and work on your own csv files. You should only use and work on the original csv datasets.
3. There is no resubmission policy for both the draft and the final project. All submissions are final.
5. May 8 is the last day I will accept any late submission of the final project. 3 points will be deducted per day the submission is late. Any submission after May 8 will not be graded and receive a zero.

Due dates

Apr 29 – Submit the draft of your project. The draft submission is optional. Feedback will be given within 1 to 3 days after submission.
May 6 – Submit the final version of your project.

Grading

1. Completion 40%

   Are all requirements satisfied?
2. Accuracy 30%

   Any errors running the program?

   Does the program accurately summarize and analyze the data?

   Is your analysis correctly represented by the visualizations?
3. Content 25%

   Are you able to find meaningful relationships and patterns from the datasets?
4. Style 5%

   Does the program follow best practice? (e.g. no redundant logic, easy to read and follow, sufficient comments etc.)