can code communities

Python for Data Analytics

Final Project

Your final project is designed to demonstrate your understanding of working in Python and analyzing a dataset(s) of your choosing.

Find and utilize an interesting dataset and then use [Google Colab](#) to write Python code (with descriptive comments) and perform the following tasks:

1. Load the data source (either from a local file or from the web)

2. Explore and describe the data
   - a. Show how the data is distributed.
   - b. Determine frequency of values in columns for categorical variables.
   - c. Create a histogram for all continuous variables.
   - d. Show any outliers in your data.
     - i. Describe how the outliers impact your analysis and how you might need to deal with this.
   - e. Discuss any missing data points including why the data is missing and how it might affect your analysis. Describe your strategy for dealing with the missing data.

3. Research your topic
   - a. Find other analyses on the same or similar data.
   - b. Describe the conclusions others have created.
     Even if you have been given a task to answer with data, doing a quick search of what others have done can help you plan for how you're going to do your own analysis.
   - c. Define the target audience for your analysis.

4. Build your analysis plan
   - a. What data do you need to solve your problem? Review your data exploration for that data.
   - b. Do you need to clean or enhance the data to address outlier/missing values? List any new columns that need to be created.
   - c. Describe any techniques you use to answer the question.
     Start simple! You can always iterate and make improvements.
   - d. Describe why you consider your answer successful.
   - e. Describe the expectations you are trying to prove or disprove.

5. Turn your plan into code
   - a. Put your plan to action and start looking at your data.
     Remember you can always make adjustments to your analysis plan as you look into your data.

Example Datasets:

Listed below are some sample dataset that you can use for inspiration or pick from for your project:

1. https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset
2. https://www.kaggle.com/datasets/michaelbryantds/university-enrollments-dataset
3. https://catalog.data.gov/dataset/purchase-card-pcard-fiscal-year-2014
4. https://catalog.data.gov/dataset/percent-change-in-consumer-spending-january-2020-through-the-present
5. https://catalog.data.gov/dataset/dohmh-community-health-survey-2010-2016
6. https://data.world/education/ap-cs-pass-rate-by-racegender
7. https://www.kaggle.com/datasets/gregorut/videogamesales
8. https://www.kaggle.com/datasets/nikhilbhathi/data-scientist-salary-us-glassdoor

Part 1 Dataset Choice:

For this part you will submit the dataset that you wish to work with for the final project. You may change your mind on the dataset, but you need to let your instructor know. You will submit this bny uploading the file to teams or linking to it in the correct spot in teams. Due on [Fill in Date].

Part 2 Draft/Data Processing:

For this part you must load your data that you chose to work with into a pandas dataframe and have done the following. Due on [Fill in Date].

1. Make sure all columns are the appropriate data type.
2. Calculate the mean, median, and mode for each of the columns. You may skip any measures that do not fit the data.
3. Filter your data into any subsets that you wish to work with. Such as country, or year.

Part 3 Requirements:

The following list outlines the requirements for the two parts of your final project. Due on [Fill in Date]

1. Requirements for Python Code:
   a. Present all code in a Google Colab shared with [Instructor Email], and a public github link to your .ipynb file and writeup.
   b. Import the data into the Colab as a pandas dataframe
   c. Clean the data (removing duplicates, bad entries, extra information, adding in missing values, or anything else needed).
   d. Create either histograms, box plots/box and whisker plots, or frequency graphs to show the distribution of data, for any relevant groups of data.
   e. Calculate the mean, median, mode, and any other relevant statistical measures for your data.
   f. Create at least 7 graphs to either determine/demonstrate/show any meaningful correlations in your data.
   g. Customize the graphs to be presentable (change color, scale, labels, or whatever else is needed to make them accurate and meaningful).
   h. Remove any outliers or duplicates, if needed, to provide more meaningful analysis/figures/graphs.
   i. Place comments in your code to describe what and why you are doing in each section.
   j. Run your code fully without errors.

2. Requirements for Summary Paper (1-2 pages length [not including the code]).
   a. Write an introduction specifying what your goal is in analyzing your data.
   b. Write a summary of the dataset describing its source, why you chose that dataset(s), and how the data was gathered.
   c. Write a summary of other people's findings on analyzing your dataset(s) or a similar dataset(s).
   d. Cite any sources used in your project.
   e. Write a summary of the process you followed in writing your code.

f. Discuss your choices of the specific graphs you are displaying.

g. Conclude with an analysis of why you feel your analysis was successful (or not).

Project Checklist and Evaluation

Requirements for Summary Paper

- □ Report is 1-2 pages length
- □ Includes link to shared Google Colab
- □ An introduction specifying what your goal is in analyzing your data.
- □ A summary of the dataset describing its source, why you chose that dataset(s), and how the data was gathered.
- □ A summary of other people's findings on analyzing your dataset(s) or a similar dataset(s).
- □ All sources used in your project are cited
- □ A summary of the process you followed in writing your code.
- □ Your choices of the specific graphs you are displaying are discussed.
- □ Conclusion with an analysis of why you feel your analysis was successful (or not).

Requirements for Python Code:

- □ All code is presented in a Google Colab
- □ It is shared with Bri?
- □ Dataset is Imported into the Colab as a pandas dataframe
- □ Data is clean
  - ○ Duplicates removed
  - ○ bad entries removed
  - ○ No missing values.
- □ Distribution is shown as either histograms, box plots/box and whisker plots, or frequency graphs.
- □ The mean, median, mode, and any other relevant statistical measures for your data have been calculated.
- □ At least 5 graphs that determine/demonstrate/show any meaningful correlations in your data are included.
- □ All included graphs are presentable, accurate,and meaningful. Note the use of color, scale, labels, etc.
- □ All outliers or duplicates are removed from the data
- □ The code has sufficient comments to describe why and what you are doing in each section.
- □ The code runs fully without errors.