Final Project Phase I

Guidelines for Phase I Submission: For open ended questions, 1-3 bullet points should suffice for most answers. You do not need essay-length answers; however, there needs to be enough information, so that we can understand your topic and confirm that you have a cohesive and feasible topic. Make sure your answers are brief, but cohesive and answer all of the questions.

NOTE: Most of the points lost in this phase are due to not reading the instructions. Please make sure to ***read each question in its entirety****.*

options:

Q1. Topic - 15 points

Please provide an overview of what your topic is going to be.

Q1.1 - 5 points

What topic have you chosen for your Final Project?

> **Answer:** The relationship between median income per county and heart disease mortality rate.

Q1.2 - 5 points

Why did you choose this specific topic and what are you looking to learn from the analysis?

> **Answer:** Heart disease is a leading cause of death in the US that is caused by many socioeconomic issues. We want to figure out whether income plays a significant role in mortality rates due to heart disease.

Q1.3 - 5 points

Explain some of the concrete insights you expect to gather from your data and/or hypothesis you expect to answer.

> **Answer:** From the data, we expect to find insights into how income might affect the person's decisions before they were to suffer from heart disease. We hypothesize that lower-income counties will have less funding and therefore less access to preventative measures for heart disease leading to overall higher mortality rates by heart disease.

Q2. Downloaded Dataset - 15 points

Please provide a brief overview of your downloaded dataset. This should demonstrate that you understand the data contained within the dataset.

Q2.1 - 2 points

Provide the link (url) to your downloaded dataset.

Answer:
https://catalog.data.gov/dataset/heart-disease-mortality-data-among-us-adults-35-by-state-territory-and-county

Q2.2 - 3 points

Provide the dimensions of your downloaded dataset in terms of rows x columns and file size.

Ex. 50,000 rows x 20 columns and 5.4mb.

If your file is a .json file, state the file size .

Ex. 4.2mb.

Answer: 78,793 rows x 21 column and a file size of 23 mb

Q2.3 - 5 points

Discuss the structure of your dataset. For .csv or table type datasets, identify significant column titles and give examples of data contained within.

Ex. 'state_name', 'population' → 'GA', '10.9 million'.

For .json data, identify significant dictionary keys you will be referencing and give examples of data contained within.

Ex. 'id', 'score' → '1632', '9.5'

Answer: 'LocationAbbr', 'LocationDesc', 'Data_Value', → 'AK', 'Kenai Peninsula', '165.1'

Q2.4 - 5 points

Explain why you chose this specific dataset. How will this data be used in your analysis? Can insights be drawn from this data alone, or will it be combined with other data?

Answer: We chose this dataset since it contains data on all mortalities caused by heart disease and identifies information such as location and race. This will help with our analysis of whether income plays a role in heart disease mortalities by cross checking the amount of deaths with median income and figuring out which counties suffer the most and what their average income is.

Q3. Web Requirement #1 (Web-scrape or HTML) - 15 points

Please provide a brief overview of your downloaded dataset. This should demonstrate that you understand the data contained within the dataset.

Q3.1 - 2 points

Provide the link (url) to your downloaded dataset.

Answer: https://nccd.cdc.gov/DHDSPAtlas/Reports.aspx

Q3.2 - 3 points

Explain briefly how you plan to retrieve the data from this source.

Answer: We plan to retrieve the data in the table by web scraping the HTML web page using BeautifulSoup.

Q3.3 - 5 points

Discuss the structure of your dataset. For .csv or table type datasets, identify significant column titles and give examples of data contained within.

Ex. 'state_name', 'population' → 'GA', '10.9 million'.

For .json data, identify significant dictionary keys you will be referencing and give examples of data contained within.

Ex. 'id', 'score' → '1632', '9.5'

Answer: The dataset is a table with county statistics of social, economic, environmental data filtered by social environment, specifically median household income. The table has column tags "County", "State", "Value" Category Range" → "McDowell", "WV", "29000", "$29,000 - $51,000 (710)"

Q3.4 - 5 points

Explain why you chose this specific dataset. How will this data be used in your analysis? Can insights be drawn from this data alone, or will it be combined with other data?

Answer: We house this data because it has the median household income range of each county. We plan to combine this data with our downloaded data, "Heart Disease Mortality Data…", to find the value of heart disease along with median household income for each county.

Q4. Web Requirement #2 (API or JSON) - 15 points

Please provide a brief overview of your downloaded dataset. This should demonstrate that you understand the data contained within the dataset.

Q4.1 - 2 points

Provide the link (url) to your downloaded dataset.

> **Answer:**
> https://apps.bea.gov/api/data/?UserID=80BEF102-7B1D-4714-BA8D-7D4D45AC07A9&method=GetData&datasetname=Regional&TableName=CAGDP2&LineCode=1&Year=2020&GeoFips=COUNTY&ResultFormat=json

Q4.2 - 3 points

Explain briefly how you plan to retrieve the data from this source.

> **Answer:** We plan on using requests module to pull the data from the api and then assign it as a variable to be used to find the data of the gdp of every county and then planning on using sorted function as well as displaying the highest and lowest gdp county's

Q4.3 - 5 points

Discuss the structure of your dataset. For .csv or table type datasets, identify significant column titles and give examples of data contained within.

Ex. 'state_name', 'population' → 'GA', '10.9 million'.

For .json data, identify significant dictionary keys you will be referencing and give examples of data contained within.

Ex. 'id', 'score' → '1632', '9.5'

> **Answer:** 'GeoName', 'CL_UNIT', 'DataValue' → 'Baldwin, AL' , 'Thousands of dollars', '8762106'

Q4.4 - 5 points

Explain why you chose this specific dataset. How will this data be used in your analysis? Can insights be drawn from this data alone, or will it be combined with other data?

> **Answer:** We chose this specific dataset because it contains the GDP of all counties in the United States. We can use this information to match up with our median income web scrape to get a determination on what area has low access to funds compared to other counties. Then we combine this data with mortality rates by heart disease and figure out whether income and revenue play a role in mortality rates.

Q5. Additional Datasets - 10 points

If you have found any datasets beyond the three required, please describe them below: (If you do not plan to use any additional datasets please simply write **N/A**)

Q5.1 - 5 points

Provide the links for any additional datasets you might use

Answer:
https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_income

Q5.2 - 5 points

Explain how you will retrieve data from these sources, and how this data is going to be used for your analysis

Answer: We will use the data from the table to fill in missing data from Web Requirement #1 about median household income by parsing through the table rows and matching them with our dataset using BeautifulSoup

Q6. Inconsistencies - 15 points

Please list at least 3 valid inconsistencies (according to the Inconsistencies Description document) you have found in your datasets, and how you plan to address each of them.

Answer:

1) Web Requirement #1 has an inconsistency of additional characters in a string: at the end of each category range, a (###) is attached at the end. We plan to address them by splicing and removing the (###) from the category range to have just the income value.
2) The downloaded data has a missing data inconsistency, where some of the county data_value values are missing. We plan on addressing this inconsistency by using the average state data_value values, which is included in the downloaded dataset, to fill in the missing data.
3) Web Requirement #1 has a missing data inconsistency as well, some of the counties are missing their value and category range values. For the value, we will replace the "Insufficient Data" with "N/A" since it is not relevant data to our project. For the category range values, we will use Wikipedia to replace "Insufficient Data" with the average state median income instead.

Q7. About Your Analysis - 10 points

Provide a BRIEF list of steps of how you plan on performing your analysis and the way you will gather/present your findings. (Non-technical, high-level overview)

Answer: 1) download all datasets, 2) web scrape all necessary data using BeautifulSoup, 3) Cross reference median household income with average gdp of counties to figure out areas of low income relative to others, 4) Create a new columns in the downloaded dataset to add median household income and average gdp of county, 5) analyze and compare high mortality values to income to see if there is a correlation, 6) determine if the hypothesis is true or not, 7) create insight analysis, data visualizations, and one pager, 8) create a video to present.

Q8. About You - 5 points

Q8.1 - 2.5 points

List the names of each of the members of the group working on this project. If you are working alone, there should be one name listed.

Team Member 1: Andrew Jiang

Team Member 2(If Applicable): Kelly Zhang

Q8.2 - 2.5 points

Each member of the group should initial below to indicate that you acknowledge this statement:

I affirm that all of the work in this project will be done by me/my team and is not duplicated from any other source. In addition, any references that I use or code that I choose to model after will be appropriately credited and referenced in my project.

Team Member 1 Initials: AJ

Team Member 2 Initials (If Applicable): KZ

Total - 100 points