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.

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: Comparison of nutrient values from lab tested data and survey data, including % Daily Value from food labels

Q1.2 - 5 points

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

Answer: Throughout my life, I have felt that I had been given a lot of nutritional misinformation. For this project, I want to learn and verify information whilst also questioning the reliability of the information. This will help me understand which sources are most reliable and how misinformation can affect health decisions

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: I expect to find some inconsistencies in between the two methods of measuring the nutrient data such as differences in mineral or vitamin content, or how outdated nutritional information may play a factor in giving false assumptions. My hypothesis is that survey data tends to underreport or generalize nutrition information. This analysis will reveal how accurate daily value percentages really are when matched against verified data.

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://fdc.nal.usda.gov/download-datasets

FNDDS: Survey nutrional Data

Q2.2 - 3 points

What are 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 (mb, gb, etc.).

Answer: 27MB, 7 columns and about 350000 rows (Json to CSV converted, as talked with TA)

Q2.3 - 5 points

Briefly discuss the structure of your dataset. For .csv or table type datasets list out the column titles and give examples of the data contained within. For json data map out the dictionary and give examples of the data contained within.

Answer: This is a CSV file(converted file, as talked with TA about this).
Columns: fdcId,description,nutrient_name,amount,unit_name,nutrient_id,rank
example: 2710797,"Mushrooms, cooked, as ingredient",Niacin,4.65,mg,1167,6600

Q2.4 - 5 points

Please 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: This dataset provides nutrient information from national dietary surveys conducted across various facilities. I will compare this data with other lab tested data and scraped information from reputable websites in order to verify, correct and decide whether survey represent accurate information.

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://www.nutritionvalue.org/

Q3.2 - 3 points

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

Answer: I plan on extracting % Daily Value (%DV) and other nutritional information by scraping the HTML content of nutrition pages using BeautifulSoup. I can use the HTML elements containing nutrient names, and their value to then compare and analyze the data.

Q3.3 - 5 points

Briefly discuss the structure of your dataset. For .csv or table type datasets list out the column titles and give examples of the data contained within. For json data map out the dictionary and give examples of the data contained within.

Answer:

The dataset is a webpage table shows the nutrition label as you would see it in a product in store with providing the macronutrient conctents such as fats, sugars, added sugars follwing a left side(title or nutrient name) and amount on right.

Q3.4 - 5 points

Please 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: This dataset is neccesary because since it represents commonly given nutrition information to the public. It will be used alongside USDA lab tested and survey based values to find inconsistencies and help support the final conclusions.

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://fdc.nal.usda.gov/download-datasets

Q4.2 - 3 points Foundation Foods JSON

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

Answer: I plan to download the Json file and then use the json python modules(or anything related) and then parse for the necessary fields such as the name, nutrients, fcid, etc.

Q4.3 - 5 points

Briefly discuss the structure of your dataset. For .csv or table type datasets list out the column titles and give examples of the data contained within. For json data map out the dictionary and give examples of the data contained within.

Answer: The JSON dataset is structured as a list of food entries under the "FoundationFoods" key. Each entry also contains things like "fdcId", "description", "foodClass", and a nested "foodNutrients". Where within the foodnutrients array there is name, amount, rank, etc. i.e "nutrient": { "id": 1003, "name": "Protein", "unitName": "g", "rank": 600 ...}

Q4.4 - 5 points

Please 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 I chose this data due to its reliability coming from a reputable source and government website. It also has potential to contain some inconsistencies, but also can serve as a dataset to utilize for my comparisons for the accuracy of the food datasets or additional information that can be found.

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://www.myfooddata.com/
(might interchange with webrequirement 2)

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: Similar to web requirement 2 by using scraping through beautifulsoup or scrappy, I might draw some daily value data or other potential data that might comes to mind as we progress.

Q6. Inconsistencies - 15 points

Please list at least 3 inconsistencies you have found in your datasets, and how you plan to address each of them.

Answer:

1) Missing nutrient values in survey dataset. I will either flag them or draw them from foundational dataset or scraped data .

2) Inconsistent nutrition names or labeling, i.e vitamin b12 or cobalamin(same thing), but I plan to use a single source for all naming conventions.

3) Differenct measurements across the various datasets. I plan on keeping consistent with mg, since its unrealistic for nutrition to exist in grams if not supplemented.

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: I plan to take some time to learn more about this topics, gather datasets, next find and fix inconsistencies using methods above, then analyze and compare the data, test my hypothesis and and finally draw conclusions.

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: Abid Khan

Team Member 2(If Applicable): Fardin Bahar

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: A.K

Team Member 2 Initials (If Applicable): F.B

Total - 100 points