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? If the scope of your project is too narrow, you are subject to lose points.

Answer: The topic we are choosing to investigate is regarding the salaries of individuals post higher education and how they may correlate with the average standardized test scores of students attending that university.

Q1.2 - 5 points

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

Answer: This topic is a major source of debate as colleges look to keep or remove standardized testing requirements during applications. Furthermore, we look to see whether there is a correlation between the average testing scores of a college and how it may relate to post education salaries. We are also able to see the background information of an average student at the school allowing us to consider various factors such has average household income.

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: Through this, we aim to gather insight about post college salaries, and how they differ amongst different colleges. We will also be able gain insight on not only the salary, but also major credit levels and debt a college student can possibly accumulate after graduation. Our essential hypothesis is that students who go to colleges with higher standardized test scores are more likely to earn a higher salary post college.

Q2. Downloaded Dataset - 15 points

Please provided a brief overview of you 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://collegescorecard.ed.gov/data
In particular we will be using the file FieldOfStudyData1718_1819_PP file

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: Our downloaded dataset is a .csv file with 224818 rows x 122 columns and is approximately 13.796mb in size.

Q2.3 - 5 points

Briefly discuss the structure of your dataset. For .csv or table type datasets list out the column titles you anticipate using and give examples of the data contained within. This is not binding. For json data map out the dictionary and give examples of the data contained within.

Answer: The data contains the college name, code, control (whether it is public or private) as well as a lot more such as major, class (undergrad, grad, phd) debt, and much more. In particular we aim to look into college information whether they are public or private, as well as look into any financial data that is provided to help us determine correlations amongst this and the other two sources being used.

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 by itself could provide insights on details about each college and its statistics of student's post-college performance, but will be combined with other data, such as acceptance rate and SAT range needed for acceptance to produce meaningful information regarding pre-college as well.

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

Please provided a brief overview of you 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.reachhighscholars.org/scores_and_acceptance.html

Q3.2 - 3 points

Briefly explain how you plan to retrieve the data from this source, including the necessary Python libraries/modules.

Answer: The dataset consists of a table in a website. Thus, BeautifulSoup will be used to parse through the webpage and codify the data so that we can manipulate it and use it with the data from other resources.

Q3.3 - 5 points

Briefly discuss the structure of your dataset. For .csv or table type datasets list out the column titles you anticipate using and give examples of the data contained within. This is not binding. For json data map out the dictionary and give examples of the data contained within.

Answer: The data that we intend to use comes from the table on the webpage. More specifically we aim to use the columns with titles name of the college, SAT 25-75th percentile, and the % of acceptance.

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 has been chosen as it provides information on a wide variety of college's SAT range and acceptance rate. This data by itself could provide insight into how difficult it is to get into the college but will provide more significant information once combined with other data regarding post-college performance of students.

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

Please provided a brief overview of you downloaded dataset. This should demonstrated that you understand the data contained within the dataset.

Q4.1 - 2 points

Provide the link (url) to your downloaded dataset.

Answer:
https://api.data.gov/ed/collegescorecard/v1/schools.json?&api_key=6QnJqX2bDvlVKsS4xPyPkMZG28mHTKRJ5rLgNxeW
The link provides more information as it is queried differently.

The directions on how to do this are provided here:

https://github.com/RTICWDT/open-data-maker/blob/master/API.md#api-access-for-the-college-scorecard-data

Q4.2 - 3 points

Briefly explain how you plan to retrieve the data from this source, including the necessary Python libraries/modules.

Answer: We aim to use the requests module in python to obtain data from the source. We will then clean it and obtain relevant data and attempt to link with data from the other sources being used.

Q4.3 - 5 points

Briefly discuss the structure of your dataset. For .csv or table type datasets list out the column titles you anticipate using and give examples of the data contained within. This is not binding. For json data map out the dictionary and give examples of the data contained within.

Answer: This api is in json format with many useful fields. Some in particular pertain to the endowment.


```
{'zip': '30332-0530', 'city': 'Atlanta', 'name': 'Georgia Institute of
Technology-Main Campus', 'alias': 'Georgia Tech', 'state': 'GA', 'locale':
11, 'dolflag': 1, 'branches': 1, 'men_only': 0, 'operating': 1,
'ownership': 1, 'region_id': 5, 'accreditor': 'Southern Association of
Colleges and Schools Commission on Colleges',…}.
```


Some sections in particular that we aim to utilize are the endowment, demographics, as well as financial information such as median household income as this may give us insight on factors that are play a role in the relationship between standardized test scores, colleges, and post college salaries.

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: This dataset has been chosen as it provides deeper insight into college life of students, such as completion rate and ethnicity. This, combined with other data will allow us to have a full set of insights regarding pre-college, during college, and post-college life of students which will all eventually impact the salaries of graduates, which is essentially the question we aim to answer through this project.

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: N/A

Q5.2 - 5 points

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

Answer: N/A

Q6. Inconsistencies - 15 points

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

Answer:

- Colleges are repeated multiple times in the dataset, making it difficult to obtain the data for the college in one tuple.
  - We plan to address this by taking the sum, average, etc (depending on the field) and making one tuple using the data from all the different rows.
- Some columns have missing or incorrect characters such as 'Bachelorâ€™s' Degree.
  - We plan to address this when cleaning our data by replacing such elements with what they are supposed to be such as Bachelor Degree.
- Many fields in the table have the value 'PrivacySuppressed' while others use NULL
  - We are going to have to uniformize these fields when working with them as some colleges have values while some don't this will have to be kept in mind when putting all the data together for a single college.
- Instead of True or False the data set uses 1 or 0
  - This is something that we will be translating over to Boolean values when working with the dataset in python
- Some columns have operators in their fields in addition to integers.
  - Remove the operators and have it be the integer with a positive or negative number depending on the direction of the sign

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:

- **Clean the downloaded dataset by removing inconsistencies.**
- **Analyze the downloaded dataset using the csv module.**
- **Web scrape the HTML file using BeautifulSoup to collect only the data needed from the columns.**
- **API can be queried to obtain information for most relevant colleges, we then plan to analyze the data independently to obtain insights on the portions and factors that we would like to use in the larger comparison.**
- **After cleaning and analyzing each source's individually we aim to combine information from all three sources and analyzing the colleges holistically using all the data rather than just a small subset.**

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. Failure to list your teammate and group them in Canvas may result in working individually on the project.

Team Member 1: Katniss Min

Team Member 2: Prit Patel

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 my 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: KM

Team Member 2 Initials (If Applicable): PP

Total - 100 points