## IST 652 - Group 3 Final Project Proposal

### Deliverables [5 points]

For this proposal, we’ll need submit a paper (.docx) which consists of the following

information:

#### 1. Team as Assigned

* Daniel Jeski - dljeski
* Mark Gopani - mgopani
* Warren Justin Fernandes - wjfernan

#### 2. Pick a topic of investigation. [1 point]

The topic we have chosen to review is COVID-19 data found on the Coronavirus Pandemic Statistics and Research page (<https://ourworldindata.org/coronavirus>). We’ll be looking to compare and contrast some of the key variables between some selected major countries. We will be picking China and India to keep the scope of the project manageable. We anticipate this to include data regarding cases, deaths, test, and safety/mask policy.

#### 3. Pick the data sets you plan on using. [1 point]

* United States, China, and India COVID datasets snapshot from the start of the pandemic to the date of date we pull the data.
* Datasets will include separated CSV for each country for each of:
  + Cases
  + Deaths
  + Tests
  + Mask Policy
* The data will be downloaded directly from <https://ourworldindata.org/coronavirus>

#### 4. Pick several (2-3) possible methods of data acquisition and analysis. [1.5 points]

Potential Questions:

* How do both the case and death rates of the United States compare to both China and India? Are there any significant notations found in the comparison?
* Was the requirement of masks effective given the death rate in the United States versus both China and India?
* How has the United States performed in vaccinating the population compared to testing? How has this compared to China and India?

For actual data processing, we’ll be simply downloading the data CSV from the source website - <https://ourworldindata.org/coronavirus>

Then, once read into Jupyter, we anticipate using the following packages:

* Numpy - Used to create arrays given numerical data used to then compare statistical information provided by the data
* Pandas - Used to iterate through and create dataframes and series in order to appropriately compare and contrast data between data sets.
* CSV - We’ll need this as the anticipated date will be downloaded in the form of CSV files. We’ll need this to import it into our Jupyter Notebook for analysis, and then to export to new CSV files for sharing.

We will then likely have the results displayed in readable tables with clear indication of the results accompanied by written analysis.

#### 5. Identify potential development tasks (2-3) and whether you believe you’ll need additional guidance in achieving results. [1.5 points]

* Identifying exactly which data sets we’ll need. The data source has a significant amount of data. Identifying which sets of data we’ll need will be vital to our success.
* Import our data directly into Juptyer to being analysis using CSV.
* Concatenate data as needed for proper analysis.
* Perform data inspection tasks such as identifying NaN/Null values to determine if action is needed to proceed. Clean the data as needed.
* Iterate and manipulate the data as needed given the questions at hand. This will include various means of counting, comparing, and statistical analysis.

## Original Instructions

Final Project Proposal

IST652 - Scripting for Data Analysis

M002: Spring 2019

Due: Saturday, March 9 by midnight

Purpose

For this assignment, you are designing an initial plan for your final project.

In the final project assignment (due at the end of the semester), you will demonstrate

your ability to write Python scripts to access and amass data from fields in one or more

of the three types of data studied in the course and prepare the data to produce

summaries, lists and other structures for analysis.

Deliverable [5 points]

For this proposal, you’ll need submit a paper (.docx) which consists of the following

information:

1. Choose whether to work individually or to work in a team of 2-3 people. If you wish

to work in a team, specify the people that you have talked with to form a team.

2. Pick a topic of investigation. [1 point]

3. Pick the data sets you plan on using. [1 point]

• The topic could focus on one main data set, but also have supporting data.

• The data may come from any source: those that you have found online,

collected from social media or obtained through other means.

• The provenance of the data must however be known and cited in your final

report.

4. Pick several (2-3) possible methods of data acquisition and analysis. [1.5 points]

• These analyses will allow you to answer the types of questions that you have

worked on for the lab and mini-project assignments.

• NOTE: Since we are not focused on visualization in this course, the results of

your analysis can be reported as structured tables, with a unit of analysis and

collected, summarized or computed values for those units.

5. Identify potential development tasks (2-3) and whether you believe you’ll need

additional guidance in achieving results. [1.5 points]

This can range from:

• Big things (e.g. I want to get information from Yelp reviews), to

• Small things (e.g. I’d like a program that helps me to get dates from the

documents in my collection and all me to compare them over time).

Assignment Result

The scale of the final project must be larger in scope than the mini-project assignments

in the following dimensions:

• Incorporating multiple datasets, possibly combining structured, semi-structured

or un-structured text data.

• Conducting additional analytical enquiry, beyond what we have done in class:

o This includes additional types of analysis or collecting data (e.g. using

another API or using Social Network Analysis)

Ideas for Data or Projects

Many web sites where people have done analysis also give the sources of their

information. For example:

• Nate Silver’s 538 web site has many examples of analysis. One is this article by

Rob Arthur and Jeff Asher on gun violence in Chicago. 538 indicates that they

retrieved their crime data from the City of Chicago open data portal.

o All 538 articles have their analyses and data posted to Github here:

fivethirtyeight/data

• Data Journalist Yue Qiu has a web site with several projects reporting data on

workers, trains carrying crude oil, and other statistics from various government

web sites.

Examples of data sets used by students in the past:

• Baseball hall of fame data

• Airbnb data from Kaggle

• Used car test data from the EPA web site

• Somerville surveys for sense of safety

• Victim crime data from the Bureau of Justice Statistics

• Data sets from the UPI web site: faculty use of Wikipedia data, forest fires, red

wine quality, bike rentals

Examples of Previous Student Final Projects

1.

• Data Sources: EPA car review data, Edmund’s Car Reviews and Dealership

Reviews.

• Questions: What are the car dealerships located in the vicinity of Syracuse and

how far away are they? How would you rate American made automobiles

according to mileage, horsepower, fuel efficiency and cost?

2.

• Data Source: City of Chicago crimes.

• Questions: Which types of crimes are most and least frequent? How many

crimes occurred in each year? In the past year, which crimes increased or

decreased over the previous year?

3.

• Data Sources: Twitter collection from the bands Dave Matthews Band and

Phish, including the users and the last 2000 tweets from their user timelines.

• Question: Compare the popularity of the two bands by comparing follower and

favorite counts from each profile, average numbers of retweets, and retweets

and favorites per followers.

4.

• Data Source: Tweets collected April 18, 2016 around #parisattacks OR #bataclan

with 32K tweets.

• Questions: What are the demographics of the tweets? Who are the most

influential users (using SNA and retweets)? What are the demographics of the

information (looking at the URLs)? What are the sentiments expressed in the

tweets?

5.

• Data Source: Tweets about Boston Red Sox and NY Yankees, and Facebook posts

and comments from the two teams Facebook fan pages.

• Questions: Updated analysis of many questions based on earlier article by Bialik

in Five-thirty-eight for 2014.

6.

• Data sources: Airbnb data set, collected tweets about airbnb

• Questions: What factors influence the customer review scores? How much

money can each host make in a particular time period? Can we use tweets

about airbnb to discover recent popular travel trends?

7.

• Data Sources: MovieLens dataset with 100K reviews and selected movie reviews

downloaded from IMDb in HTML.

• Questions: Do movie ratings differ according to gender and genre? Do movie

reviews differ by gender for movies with male or female protagonists?