**M.S. in Applied Data Science Project Portfolio**

**Milestone Requirement**

**Ian Ustanik**

**Social Media Activity and the Coronavirus**

**Due Date: 4/20/2021 Date Submitted: 4/20/2021**

**File(s):** Final Presentation.pdf, Final Project Report.pdf, Submission\_pkg.final.zip

**Tool(s) Required For Viewing:** Any text editor for .json files, Adobe Reader for .pdf files, Anaconda – Jupyter Notebook for .ipynb files, Microsoft Excel for .csv files,

**Further Details:** To investigate the relationship between social media activity and the coronavirus, this project required the use of two separate data sources. The first of these two data sources were coronavirus-related tweets. The Tweepy package was utilized as a wrapper to make a GET request to the Twitter API. To see if there was a difference in results depending on specific tweet content, there was a search for tweets with keywords #covid19, #maskup, and #travel specifically. For the second data source, the project required Covid data that included new case counts for each day by state. The required numbers were found in the Cases and Deaths by State portion of the CDC website. Following data exploration and data cleaning, three methods of analysis were used to answer the following three questions: What is the current state of Covid-19? What is the sentiment of coronavirus-related tweets? Does a correlation exist between social media sentiment and Covid-19 in the United States? Methods of analysis included location analysis, sentiment analysis, in addition to various regression models.

**(FURTHER DETAIL FOR .ZIP FILE)**

**Program Description**

Explanation of workflow and order of script execution. The following list contains the script names with their corresponding input/outputs:

1. Get Tweets (“1\_Twitter\_API.ipynb”)

a. Input:

i. Query parameters

ii. File name and save location

b. Output:

i. Tweepy API response to csv

ii. Csv for each day (five days) with a query for each category or bin (3 bins)

iii. Example: “covid1913.csv”, “mask14.csv”, “travel19.csv”, …

2. Combine Tweets (“2\_final\_project\_script”)

a. Input:

i. Directory name containing all csv files output from the “1\_Twitter\_API.ipynb”

script

b. Output:

i. One csv file containing all twitter data with sentiment scores (“all\_scores.csv”)

3. Get Covid Data (“3\_CovidData.ipynb”)

a. Input:

i. csv from CDC website (“US\_COVID\_FINAL.csv”)

b. Output:

i. “covidData.csv”

4. Extract locations from tweets then insert (“4\_StatesLocationFull.ipynb”)

a. Input:

i. “all\_scores.csv”

b. Output:

i. “all\_states.csv”

5. Map out data for analysis (“5\_Maps.ipynb”)

a. Input:

i. “covidData.csv”, “center.csv”, “allStatesR.csv”, “2019Pops.csv”,

“WMC3-us-states.json”

b. Output:

i. Map plots

6. Regression Analysis (“6\_AnalysisFinal.ipynb”)

c. Input:

i. “allStatesR.csv”, “2019Pops.csv”

d. Output:

i. Various regression models

**Output Description**

The following list corresponds with each element of the previous list, further explaining the output of each script file.

1. Get Tweets

a. The files that were returned from the Tweepy requests. Each file represents the return of one single day for one specific category.

2. Combine All Tweets

a. This csv file represents all the Twitter data collected, along with their corresponding sentiment analysis computations.

3. Get Covid Data

a. This file represents a cleaned, Pandas friendly version of the data collected from the CDC website.

4. Extract Locations

a. This file is the same shape as the output of “combine all Tweets” output, but with more accurate location information inserted into their respective locations.

5. Map Data Analysis

a. These outputs are plots that display the processed data geographically.

6. Regression Analysis

a. These outputs are various regression models fit to investigate the direct correlation between social media activity and the coronavirus.