**Huang, Shiwei Final Project**

**Data Analytics**

1. **Abstract and Introduction (2%) Describe your motivation, initial hypothesis/ idea that you wanted to investigate, and if applicable any prior work, interest in the topic (like an intro for a paper, with references), Min. 1/2 page.**

**Introduction:**

The main goal of my final project is to compare two sets of data: COVID19 up-to-date data within the United States and COVID19 up-to-date data in China. In the following part of my data analysis, I will use the mortality rate, confirmed case, suspected case to try to understand why the COVID19 mortality rate is different from different countries. I will use my data model to compare with the scholar paper and hopefully get a conclusion. The second goal of my analyses is to understand whether the testing rate has anything to do with America’s geographic location, that being said, I want to understand whether the higher latitude has the higher testing rate, or vice versa.

Since mid-march, COVID-19 has hit America hard. I wanted to use the latest COVID-19 data as an excel file and perform various analyses on the COVID data. Mainly, my main interest is to analyze the COVID-19 data points within the United States across different states. In my data analyses, I will perform basic reports such as the historical number of max death in the state and number of minimum death within each state; number of confirmed cases maximum, mean and minimum within each state; number of maximum, minimum and average recovered people within each state.

According to the summary of the COVID data, the min confirmed case across the state is 0; the maximum confirmed case across the state is 295106 historically until 04/28/2020; the average confirmed case from January through April is 4575; and the mean for confirmed case is 17162. In terms of total number of deaths, the minimum number of deaths is 0; the maximum number of deaths within a day is 22912; the average number of deaths across the state is 192 and the mean number of deaths is 989. In terms of recovered rate, the minimum number of recovered cases within a day is 0; the average recovered cases historically is 1146; and the maximum recovered cases is 115936.

A close up of a receipt

Description automatically generated

*Data summary on the US data*

In order to get precise table summaries, I listed *covid042820* command and R demonstrated a clear readable way of my COVID-19 latest data:

A close up of text on a white background

Description automatically generated

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

A close up of text on a white background

Description automatically generated

*Above is the summary on the China data*

In addition to the U.S. data, I also obtained a China data set. I performed the summary of my China dataset. Up till April 28th, the total death toll is 9276 people. Total confirmed case is: 167223 people. Total cured case is: 156270.

**2. Data Description (3%) 1NOTE: 6000-level students must develop at least two different types of models, not just change the number of variables for a given model type. Describe how you determined which datasets you used in this project, the criteria, source, data and information- types in detail, associated documentation and any other supporting materials. Min. 1/2 page text (+graphics if applicable).**

One important question I want to estimate is to calculate the mortality rate within the United States. I used the linear model to predict the mortality rate based on the number of deaths, number of recovered people and number of people who are still being hospitalized together and ran my linear regression model.

My summary of the mortality rate within the United State is 0.2; meaning if there are 100 people who are tested for COVID-19, there will be around 2 people to encounter death within the United States.

A screenshot of text

Description automatically generated

However, the mortality rate in China is only 0.011. My data shows me that for the same amount of people who have gotten COVID-19; people in the U.S are more likely to face death compared with the number in China.

I further analyze the COVID-19 data with a decision tree model. Decision tree is a supervised model for continuous input and output models. I want to use decision tree to give me a visual analyzation to see whether the testing rate within the United States has anything to do with America’s geographic location.

**3. Analysis (5%) Explore the statistical aspects of your datasets. Perform any transformations, interpolations, smoothing, cleaning, etc. required on the data, to begin to explore your hypothesis/ questions. Analyze the distributions; provide summaries of the relevant statistics and plots of any fits you made. Discuss and specify or estimate possible sources of error, uncertainty or bias in the data you used (or did not use). Min. 2 pages text + graphics.**

I performed some basic analyses on the covid 19 data in the United States up to Apr 28th. For the smoothing and cleaning process, I got rid of useless values from the data rows. I performed linear regression on both U.S data and on the Chinese data.

Below is the Chinese dataset on a histogram on confirmed cases cumulatively in a day:

A screenshot of a cell phone

Description automatically generated

**4. Model Development and Application of model(s) (12%) What types of models you used to describe the data (regression, classification, clustering, etc.), patterns/ trends you found, visual approaches that helped you choose models, and or variables (type/ number) in the model, other parameter choices or settings for the models (e.g. distance metrics, kernels, etc.). Apply the models to assess model performance (i.e. predict). Discuss the confidence in your results including any statistic measures. Discuss how you validated your models and performed any optimization (give details). Min. 6 pages text + graphics.**

This is the summary of the linear regression mortality rate for the Chinese data. I realized after reading my summary that my variable’s coefficients are very small. So my assumption is that a linear regression model might not be the best model to analyze the predicted mortality rate using my variables. Another suspicion that I have is that the linear regression model probably needs more variables than my current variables such as confirmed cases and suspected cases.

A screenshot of text

Description automatically generated

*Linear regression model on China data.*

For the Chinese data, my linear regression result showed below:

A close up of a piece of paper

Description automatically generated

The second model I used is the decision tree model. I want to understand whether the high latitude has anything to do with the testing rate. My decision tree graph demonstrated that the higher latitude is, the higher the testing rate is. For example, for latitude of around 40; the people who are tested is 88%; however, for latitude of around 38, only 66% of people who have been tested; for latitude of around 37, only 54% people who have tested.

A screenshot of a cell phone

Description automatically generated

**5. Conclusions and Discussion (3%) Describe your conclusions; interpret the results, predictions you made, the models and their characteristics, and a give summary of what changed as you went through the project (data, analysis, model choices, etc.), what you would do next, or do differently in a subsequent exploration. Min. 1 page text + graphics (optional). References – websites, papers, packages, data refs, etc. should be included at the end. Include your R scripts! (e.g. in a zip file) and also include the Github URL that contains the code. There is no specific citation format, just be consistent.**

My data told me two important things with COVID-19 in between the United States and China. First, the mortality rate of COVID-19 in China is 50% lower than the mortality rate within the United States. Second, within the United States, places with a higher latitude has a higher percentage to get tested.

Reference:

<https://www.datacamp.com/community/tutorials/decision-trees-R>