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.

```
> summary(covid042820)
  Province_State Country_Region     Last_Update         Lat              Long_            Confirmed
 Alabama     : 1   US:59      2020-04-29 02:32:33:59   Min.   :-14.27   Min.   :-170.13   Min.   :      0
 Alaska      : 1                                       1st Qu.: 34.59   1st Qu.:-101.17   1st Qu.:   1248
 American Samoa: 1                                     Median : 39.06   Median : -87.94   Median :   4575
 Arizona     : 1                                       Mean   : 36.84   Mean   : -85.21   Mean   :  17162
 Arkansas    : 1                                       3rd Qu.: 42.36   3rd Qu.: -76.97   3rd Qu.:  15464
 California   : 1                                      Max.   : 61.37   Max.   : 145.67   Max.   : 295106
 (Other)     :53                                       NA's   :3        NA's   :3
     Deaths          Recovered          Active            FIPS          Incident_Rate     People_Tested
 Min.   :    0.0   Min.   :     0   Min.   :-115936   Min.   :   1.00   Min.   :   0.00   Min.   :     3
 1st Qu.:   44.0   1st Qu.:   466   1st Qu.:   1458   1st Qu.:  18.25   1st Qu.:  97.66   1st Qu.: 20921
 Median :  192.0   Median :  1146   Median :   4520   Median :  32.50   Median : 150.71   Median : 59251
 Mean   :  989.1   Mean   :  6898   Mean   :  14453   Mean   :3288.09   Mean   : 258.19   Mean   :103495
 3rd Qu.:  761.0   3rd Qu.:  2260   3rd Qu.:  15227   3rd Qu.:  47.75   3rd Qu.: 252.84   3rd Qu.:115240
 Max.   :22912.0   Max.   :115936   Max.   : 272194   Max.   :99999.00  Max.   :1750.23   Max.   :844994
                   NA's   :22       NA's   :1         NA's   :1         NA's   :3         NA's   :3
 People_Hospitalized Mortality_Rate      UID            ISO3       Testing_Rate      Hospitalization_Rate
 Min.   :   56.0     Min.   : 0.000   Min.   :      16   ASM: 1   Min.   :   5.392   Min.   : 6.788
 1st Qu.:  277.0     1st Qu.: 2.995   1st Qu.:84000012   GUM: 1   1st Qu.:1271.876   1st Qu.:10.627
 Median :  902.5     Median : 3.954   Median :84000028   MNP: 1   Median :1559.246   Median :15.888
 Mean   : 3542.2     Mean   : 4.249   Mean   :76885809   PRI: 1   Mean   :1902.382   Mean   :15.725
 3rd Qu.: 2227.8     3rd Qu.: 5.143   3rd Qu.:84000043   USA:54   3rd Qu.:2401.672   3rd Qu.:19.767
 Max.   :64275.0     Max.   :14.286   Max.   :84099999   VIR: 1   Max.   :5446.019   Max.   :29.280
 NA's   :29          NA's   :2                                    NA's   :3          NA's   :29
```

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:

```
> covid042820
          Province_State Country_Region         Last_Update      Lat       Long_  Confirmed Deaths Recovered
1                Alabama             US 2020-04-29 02:32:33  32.3182  -86.9023      6750    242        NA
2                 Alaska             US 2020-04-29 02:32:33  61.3707 -152.4044       351      9       228
3         American Samoa             US 2020-04-29 02:32:33 -14.2710 -170.1320         0      0        NA
4                Arizona             US 2020-04-29 02:32:33  33.7298 -111.4312      6955    275      1450
5               Arkansas             US 2020-04-29 02:32:33  34.9697  -92.3731      3127     57      1146
6             California             US 2020-04-29 02:32:33  36.1162 -119.6816     46164   1864        NA
7               Colorado             US 2020-04-29 02:32:33  39.0598 -105.3111     14316    736      2275
8            Connecticut             US 2020-04-29 02:32:33  41.5978  -72.7554     26312   2087        NA
9               Delaware             US 2020-04-29 02:32:33  39.3185  -75.5071      4575    137      1096
10       Diamond Princess             US 2020-04-29 02:32:33       NA       NA        49      0         0
11   District of Columbia           US 2020-04-29 02:32:33  38.8974  -77.0268      3994    190       660
12                Florida             US 2020-04-29 02:32:33  27.7663  -81.6868     32848   1171        NA
13                Georgia             US 2020-04-29 02:32:33  33.0406  -83.6431     24922   1036        NA
14         Grand Princess           US 2020-04-29 02:32:33       NA       NA       103      3         0
15                   Guam           US 2020-04-29 02:32:33  13.4443  144.7937       141      5       129
16                 Hawaii           US 2020-04-29 02:32:33  21.0943 -157.4983       609     16       493
17                  Idaho           US 2020-04-29 02:32:33  44.2405 -114.4788      1952     60      1039
18               Illinois           US 2020-04-29 02:32:33  40.3495  -88.9861     48102   2125        NA
19                Indiana           US 2020-04-29 02:32:33  39.8494  -86.2583     16588    901        NA
```

```
20              Iowa    US 2020-04-29 02:32:33  42.0115  -93.2105    6376    136    2164
21            Kansas    US 2020-04-29 02:32:33  38.5266  -96.7265    3652    127      NA
22          Kentucky    US 2020-04-29 02:32:33  37.6681  -84.6701    4375    225    1521
23         Louisiana    US 2020-04-29 02:32:33  31.1695  -91.8678   27286   1801   17303
24             Maine    US 2020-04-29 02:32:33  44.6939  -69.3819    1040     51     585
25          Maryland    US 2020-04-29 02:32:33  39.0639  -76.8021   20113   1016    1295
26     Massachusetts    US 2020-04-29 02:32:33  42.2302  -71.5301   58302   3153      NA
27          Michigan    US 2020-04-29 02:32:33  43.3266  -84.5361   39262   3568    8342
28         Minnesota    US 2020-04-29 02:32:33  45.6945  -93.9002    4181    301    1912
29       Mississippi    US 2020-04-29 02:32:33  32.7416  -89.6787    6342    239      NA
30          Missouri    US 2020-04-29 02:32:33  38.4561  -92.2884    7450    330      NA
31           Montana    US 2020-04-29 02:32:33  46.9219 -110.4544     451     15     356
32          Nebraska    US 2020-04-29 02:32:33  41.1254  -98.2681    3517     56      NA
33            Nevada    US 2020-04-29 02:32:33  38.3135 -117.0554    4821    219      NA
34     New Hampshire    US 2020-04-29 02:32:33  43.4525  -71.5639    2010     60     798
35        New Jersey    US 2020-04-29 02:32:33  40.2989  -74.5210  113856   6442   15642
36        New Mexico    US 2020-04-29 02:32:33  34.8405 -106.2485    2974    105     666
37          New York    US 2020-04-29 02:32:33  42.1657  -74.9481  295106  22912   51630
38    North Carolina    US 2020-04-29 02:32:33  35.6301  -79.8064    9755    363      NA
39      North Dakota    US 2020-04-29 02:32:33  47.5289  -99.7840     991     19     409
40 Northern Mariana Islands  US 2020-04-29 02:32:33  15.0979  145.6739      14      2      12
41              Ohio    US 2020-04-29 02:32:33  40.3888  -82.7649   16769    799      NA
42          Oklahoma    US 2020-04-29 02:32:33  35.5653  -96.9289    3410    207    2260
43            Oregon    US 2020-04-29 02:32:33  44.5720 -122.0709    2385     99      NA
44      Pennsylvania    US 2020-04-29 02:32:33  40.5908  -77.2098   45137   2046      NA
45       Puerto Rico    US 2020-04-29 02:32:33  18.2208  -66.5901    1400     86      NA
46      Rhode Island    US 2020-04-29 02:32:33  41.6809  -71.5118    7927    239     466
47    South Carolina    US 2020-04-29 02:32:33  33.8569  -80.9450    5735    192    2830
48      South Dakota    US 2020-04-29 02:32:33  44.2998  -99.4388    2313     11    1392
49         Tennessee    US 2020-04-29 02:32:33  35.7478  -86.6923   10052    188    4921
50             Texas    US 2020-04-29 02:32:33  31.0545  -97.5635   26357    719   11786
51              Utah    US 2020-04-29 02:32:33  40.1500 -111.8624    4345     41    1704
52           Vermont    US 2020-04-29 02:32:33  44.0459  -72.7107     862     47      NA
53     Virgin Islands    US 2020-04-29 02:32:33  18.3358  -64.8963      57      4      51
54          Virginia    US 2020-04-29 02:32:33  37.7693  -78.1700   14339    492    1914
55        Washington    US 2020-04-29 02:32:33  47.4009 -121.4905   13842    786      NA
```

```
> summary(covidata)
        provinceName provinceShortName                   cityName   confirmedCount     suspectedCount
Chongqing      : 41   Chongqing: 41   境外输入          : 19   Min.   :    0.00   Min.   :  0.000
Guangdong Province: 22   Guangdong: 22   Luzhou            :  6   1st Qu.:    7.00   1st Qu.:  0.000
Sichuan Province  : 22   Sichuan  : 22   Area to be identified:  3   Median :   18.00   Median :  0.000
Henan Province    : 20   Henan    : 20   Fuzhou            :  2   Mean   :  349.84   Mean   :  3.226
Beijing        : 19   Beijing  : 19   Suzhou            :  2   3rd Qu.:   52.75   3rd Qu.:  0.000
Shanghai       : 19   Shanghai : 19   (Other)           :412   Max.   :68128.00   Max.   :384.000
(Other)        :335   (Other)  :335   NA's              : 34
  curedCount         deadCount
Min.   :    0.0   Min.   :    0.00
1st Qu.:    6.0   1st Qu.:    0.00
Median :   16.0   Median :    0.00
Mean   :  326.9   Mean   :   19.41
3rd Qu.:   48.0   3rd Qu.:    0.00
Max.   :63616.0   Max.   : 4512.00
```

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.

```
Call:
lm(formula = data$Mortality_Rate ~ data$Deaths + data$Recovered +
    data$People_Hospitalized)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6200 -0.8327  0.2027  0.9321  2.0522

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.3378824  0.5558322   4.206  0.00103 **
data$Deaths                0.0213010  0.0085718   2.485  0.02735 *
data$Recovered            -0.0003597  0.0003585  -1.003  0.33400
data$People_Hospitalized  -0.0022215  0.0023340  -0.952  0.35857
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.347 on 13 degrees of freedom
Multiple R-squared:  0.5417,    Adjusted R-squared:  0.4359
F-statistic: 5.121 on 3 and 13 DF,  p-value: 0.01479
```

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:

Histogram of B
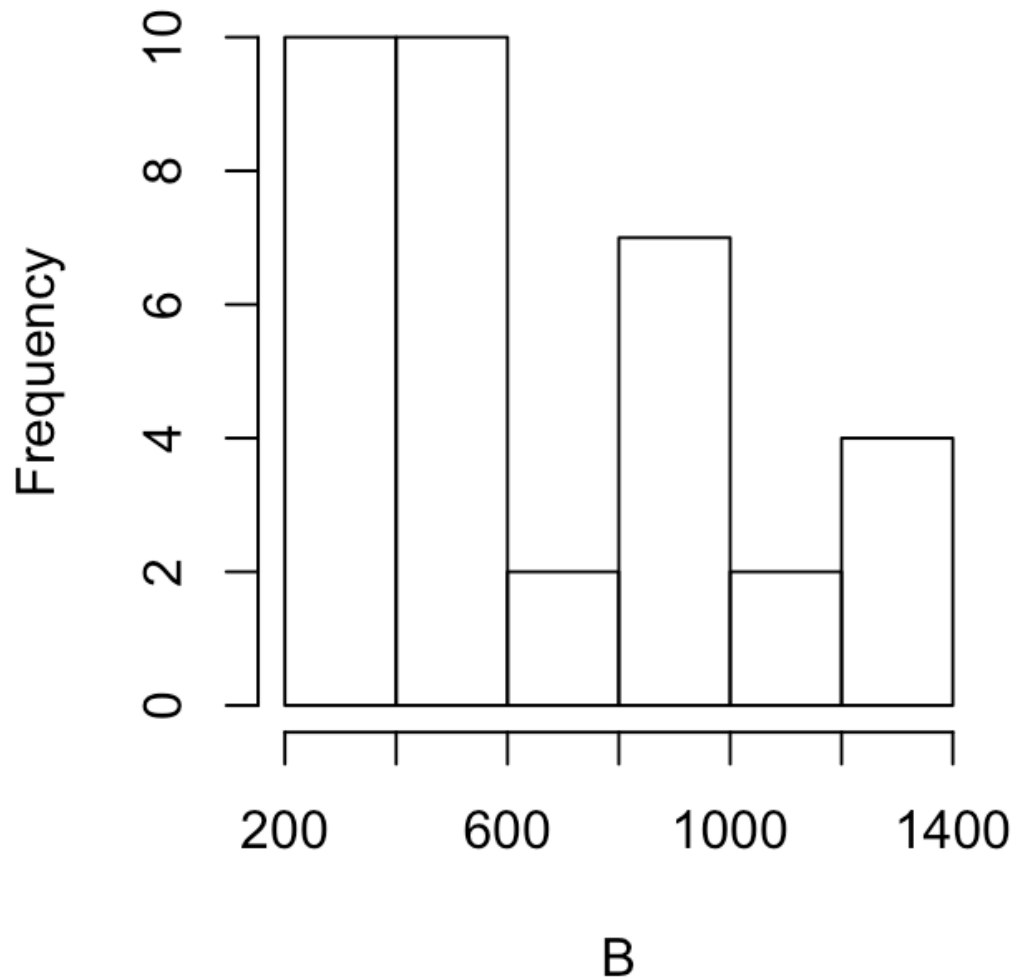
4. Model Development and Application of model(s) (12%) What types of models you used to describe the data (regression, classification, clustering, etc.), patterns/ trends you found, visual approaches that helped you choose models, and or variables (type/ number) in the model, other parameter choices 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.

```
> summary(lmMortalityRate)

Call:
lm(formula = data$Morality ~ data$Confirmed + data$Suspected +
    data$Current)

Residuals:
    Min      1Q   Median      3Q      Max
-0.01648 -0.01105 -0.01103 -0.01093  0.32231

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.102e-02  1.678e-03   6.570 1.35e-10 ***
data$Confirmed -5.241e-06  5.422e-05  -0.097    0.923
data$Suspected -1.020e-06  7.495e-05  -0.014    0.989
data$Current    6.740e-06  5.829e-05   0.116    0.908
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03588 on 465 degrees of freedom
Multiple R-squared:  0.01252,   Adjusted R-squared:  0.006148
F-statistic: 1.965 on 3 and 465 DF,  p-value: 0.1184
```

Linear regression model on China data.

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

```
~/ ↪
> na.omit(data)
      Confirmed Suspected Current      Morality
1           939       384      586 0.0138445154
2           386        34      116 0.0000000000
3           263         8      195 0.0152091255
4            20         0       15 0.0000000000
5            27         0       25 0.0370370370
6            52         0       49 0.0576923077
7            47         0       43 0.0851063830
8            46         0       46 0.0000000000
9            43         0       42 0.0232558140
10           17         0       17 0.0000000000
11           15         0       15 0.0000000000
12           14         0       14 0.0000000000
13            5         0        5 0.0000000000
14            3         0        3 0.0000000000
15            1         0        1 0.0000000000
16         1037        47      811 0.0038572806
18          644       300      584 0.0108695652
```

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.
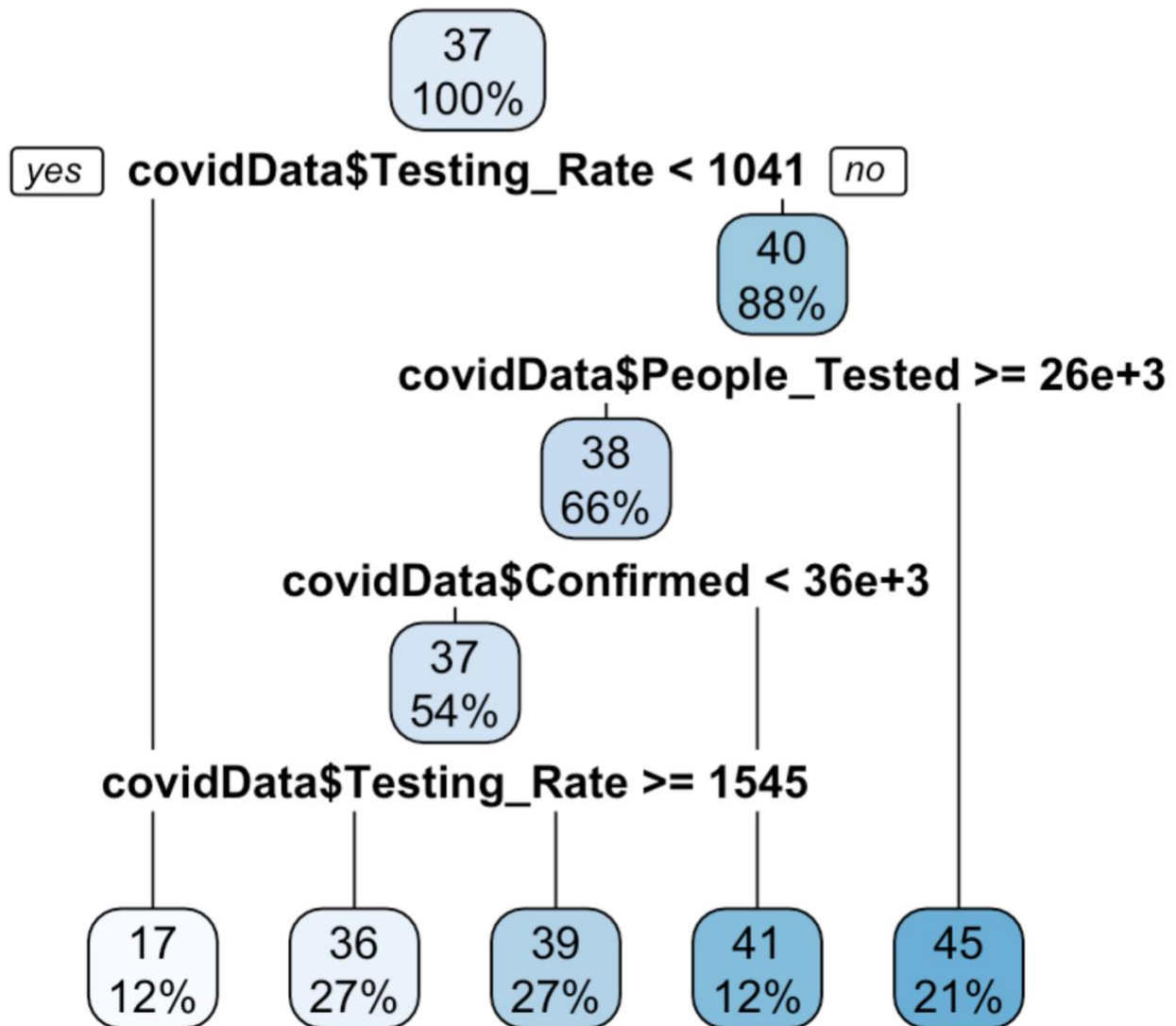
5. Conclusions and Discussion (3%) Describe your conclusions; interpret the results, predictions you made, the models and their characteristics, and a give summary of what changed as you went through the project (data, analysis, model choices, etc.), what you would do next, or do differently in a subsequent 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