



Data Analysis and Statistical Inference

“COVID-19 OUTBREAK ANALYSIS IN THE USA”

Contributors: Supreet Bhatia, Mullica, Emilio Sagre, Steven Whang

Contents

Overview	2
Purpose	2
Introduction	2
Dataset	2
Variables	2
Statistical Software	3
Cleaning and Data Transformation	3
Descriptive Analysis	4
Overview of Dataset	4
Timeseries Analysis for Confirmed and Deaths	6
Confirmed Cases Analysis	9
Death Cases Analysis	12
Inferential Analysis	14
Overview	15
General Analysis	15
Confirmed Cases Analysis	16
Death Cases Analysis	19
Conclusion	22
Attachments	23
Appendix A	24
Removed pre-identified variable using MS Excel Office 365	24
Appendix B	26
Variable Types and Descriptions	26
Appendix C	30
Inferential Analysis R code results and graphs	30
Appendix D	36
Cleaning and Transformation Codes	36
Appendix E	42
Descriptive Analysis Codes	42
Appendix F	50
Inferential Analysis Codes	50

Overview

Purpose

The purpose of this study is to investigate the COVID-19 outbreak in the United States. This paper will use different statistical techniques in analysis. The details of various features are described in the article written by Haratian et al. (2021).

References:

Haratian, A., Fazelinia, H., Maleki, Z., Ramazi, P., Wang, H., Lewis, M.A., Greiner, R., et al. (2021), "Dataset of COVID-19 outbreak and potential predictive features in the USA", Data in Brief, Data Brief, Vol. 38, available at: <https://doi.org/10.1016/J.DIB.2021.107360>.

Introduction

The COVID 19 pandemic, which started in 2019, has changed the way the world works. Due to global deaths, nations around the world have taken precautions that have affected their respective economies. Because of this, it is important to study what has already transpired in order to brace for what is about to come. From the past couple of years that the pandemic has taken place, there have been heaps of real data that can be studied. This dataset provides these data. Although it only includes data regarding the United States, it can give us a clearer picture of what we are actually dealing with, so we can predict and plan accordingly for the future.

This dataset includes numerous variables that may explain the pandemic's dynamics. Most importantly, it also includes the number of daily confirmed cases and deaths of all 50 states and their respective counties across 9 months in 2021. From this, we can investigate which variables contribute to cases and deaths the most, and properly predict these all-important variables as well. By doing so, we can slow down the deaths globally since we will find out the contributing factors to both cases and deaths.

Note:

Several pre-identified variables were removed from the dataset using Microsoft Excel. See appendix A for details.

Dataset

- Covid Dataset.csv (original dataset)
- Covid_cleanish1.csv (initially cleaned dataset where several columns were dropped using MS Excel based on the note above. We will use this dataset moving forward)

Variables

There are 36 variables. These are listed below. As for the descriptions and type, please refer to Appendix B.

- | | |
|----------------------------|---------------------------------|
| • date | • covid_19_deaths |
| • county_fips | • social_distancing_total_grade |
| • count_name | • daily_state_test |
| • state_fips | • precipitation |
| • state_name | • temperature |
| • covid_19_confirmed_cases | • virus_pressure |
| • total_population | • political_party |

- female_percent
- area
- population_density
- hospital_beds_ratio
- ventilator_capacity_ratio
- icu_beds_ratio
- houses_density
- total_college_population
- percent_smokers
- percent_diabetes
- religious_congregation_ratio
- airport_distance
- passenger_load_ratio
- meat_plants
- median_household_income
- percent_insured
- deaths_per_100000
- gdp_per_capita
- Age_0_19
- Age_20_59
- Age_60
- immigrant_student_ratio

Statistical Software

- Data Transformation and Inferential Analysis

R language using RStudio 2021.09.2+382 “Ghost Orchid” Release (fc9e217980ee9320126e33cdf334d4f4e105dc4f, 2022-01-04) for Windows Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) QtWebEngine/5.12.8 Chrome/69.0.3497.128 Safari/537.36

- Visualization

Tableau Desktop Professional Edition 2021.4.4 (20214.22.0213.1102) 64-bit

Cleaning and Data Transformation

The partially cleaned dataset has a total of 562,128 observations with 36 variables. The team tried loading the whole dataset into R Studio but came across an issue when running certain inferential analysis like linear regression. Our current machine can't handle large dataset computation. This leads us into reducing the number of observations with the following steps:

1. We created additional columns based on week count as reference to the date variable.
2. We converted the categorical variable 'social_distancing_total_grade' to numerical variable by converting them to a factor and then assigning them as numeric (A as 1, A- as 2 to F as 12).
3. Dropped the following variables: 'date', 'county_name', 'state_name'.
4. Grouping of the datasets will be based on aggregating the observations by 'week' and 'county_fips' since the data were recorded based on 'date' and 'county_fips'.
5. Created 2 groups of datasets for joining later since a different aggregation process will be applied.
 - a. Dataset 1: 'week', 'county_fips', 'covid_19_confirmed_cases', 'covid_19_deaths' will be grouped by getting the sum of 'covid_19_confirmed_cases', 'covid_19_deaths' values.
 - b. Dataset 2: all other variables except 'covid_19_confirmed_cases', 'covid_19_deaths' will be grouped by getting the mean values.
6. 2 datasets were merged using row names as reference

The new dataset has a total of 82,320 from 562,128 observations. The new dataset was reduced to 85.36% without affecting the behaviour and nature of the data.

Based on the data brief, the following age variables should add up to 100%. However, upon checking the data there were several observations with a sum of around 103%. To solve the proportion issue, we applied a simple mathematical formula to resolve the issue. We divided the variable with a certain age bracket with the total age multiplied by 100 to get the new age proportion. See below:

$$\text{Age_0_19}/(\text{Age_0_19} + \text{Age_20_59} + \text{Age_60.}) * 100$$

$$\text{Age_20_59}/(\text{Age_0_19} + \text{Age_20_59} + \text{Age_60.}) * 100$$

$$\text{Age_60.}/(\text{Age_0_19} + \text{Age_20_59} + \text{Age_60.}) * 100$$

Note: For the codes, see Appendix B.

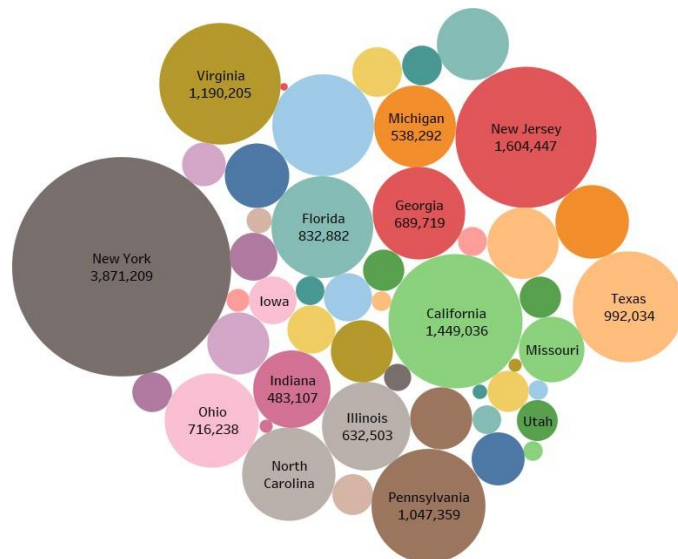
Descriptive Analysis

Overview of Dataset

We will look at different obvious factors in our dataset that may impact the spread of COVID-19 and if these factors also impact the number of deaths caused by the virus. We will look at the performance by state and see how each differs from each other.

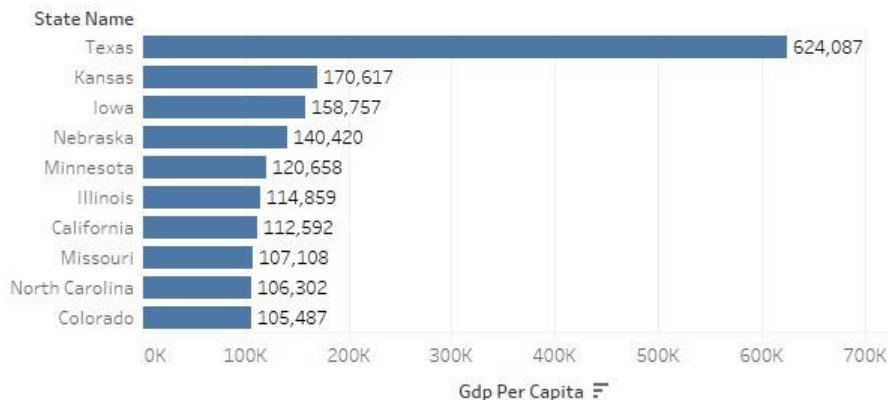
A. Population Density

The USA has several big cities across the country. This massive growth was a result of the industrial revolution which attracted more jobs to different people from the rural areas and immigrants across the world. The top dominating dense states are New York at 3.87 million per square mile followed by New Jersey at 1.6 million per square mile, California at 1.49 million per square mile, Virginia at 1.19 million per square mile, and Pennsylvania at 1.04 per square mile. This massive urbanization can increase the density of the state, hence may impact the spread of COVID-19.



B. Gross Domestic Product (GDP) per capita

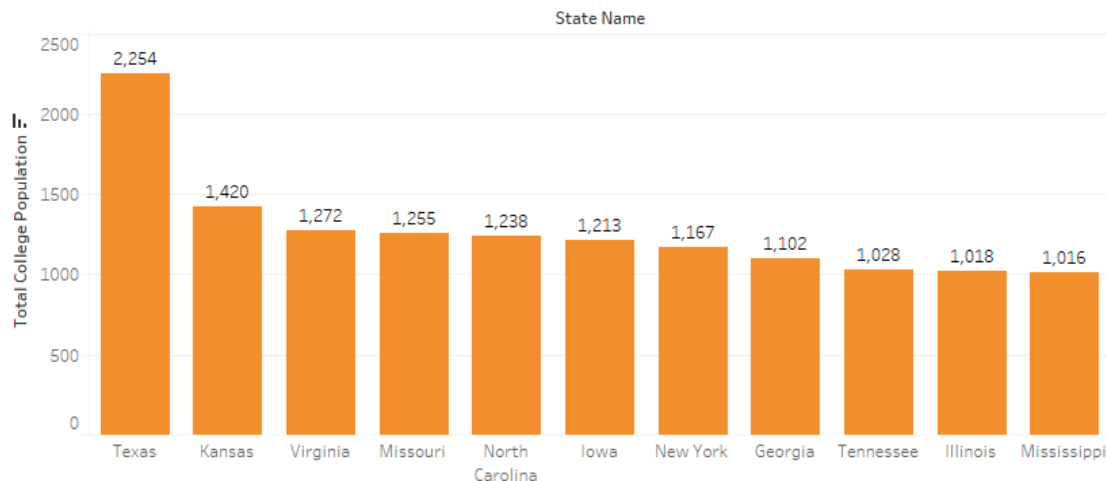
GPD per Capita_high



Looking at the GDP per Capita per state, Texas dominated the list at 624,087 followed by Kansas at 170,617, Iowa at 158,757, Nebraska at 140,420 and Minnesota at 120,658. The top 5 most dense state did not make it to the top 10 except for California at rank 7. This showed population density is not a good indicator that the state is performing well in terms of standard of living across different demographics. We will try to see later if having high GDP per capita can help curb the fight against the spread of COVID-19.

C. Total College Population

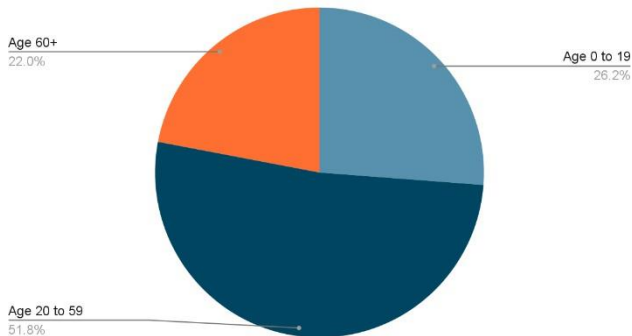
College Population



We are in the age of information. There are several sources we can find and consume our information. The way we consume it is based on how our critical thinking was molded. Education level plays a big role in this thinking process development. The data showed that Texas has the most educated demographics at 2,254 per total population. This is followed by Kansas at 1,420 per total population and Virginia at 1,272 per total population. This tells us that having high standard of living could be attributed to higher education as well. Later, we will see if this has impact on the spread of COVID-19 and deaths as during the peak of the pandemic there were several measures like stay-at-home orders, social distancing and wearing masks were introduced to help curve the spread. The enforcement of this however rely on the individuals who believes the true impact of this virus to the community.

D. Age Distribution

Age Distribution

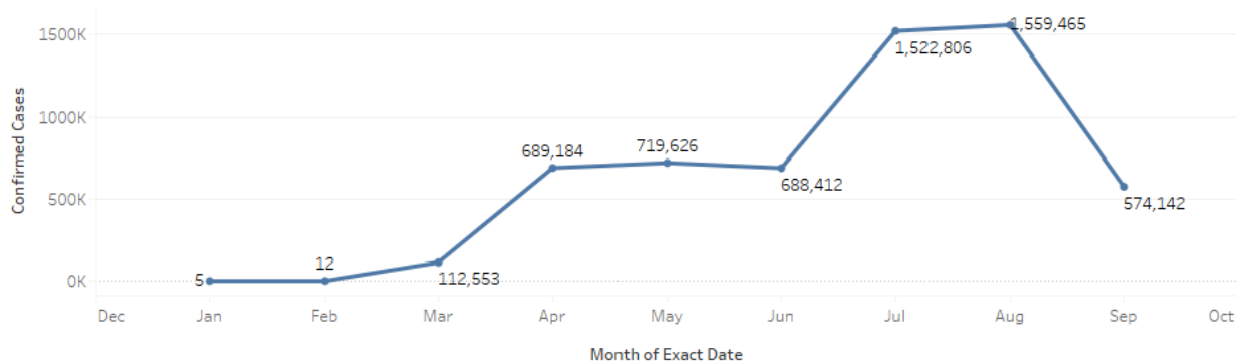


The last factor we are looking at that can impact the spread of COVID-19 and deaths is age. 51% of the Americans are of ages 20 to 59. This is followed by 26 at ages 0 to 19 and age 60 plus at 22%. We know that COVID-19 virus is more lethal to ages 60 plus and above and this has no to little impact to ages 19 and below. However, since 73% of the population is ages 20 and above, we can tell the a lot of the population will be impacted by this virus.

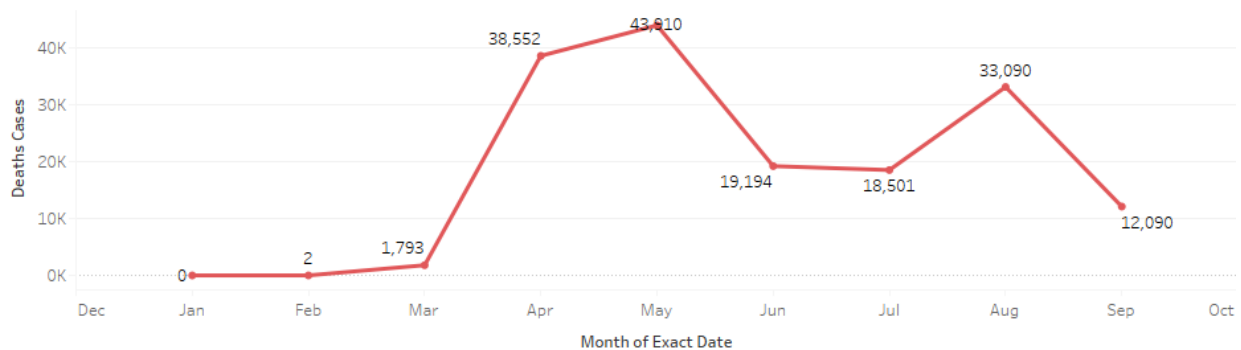
We will look at the impact of the COVID-19 virus in different states and what factors could drive the increase or decrease in cases or deaths.

Timeseries Analysis for Confirmed and Deaths

Timeseries_case



Timeseries_death



From the timeseries above, we can infer that there were 2 surges of the covid19 virus across 9 months. Evidently, the first surge was more severe as cases increased from 12 in February to 112k in March. Proportionally, the death count has a delayed response with confirmed cases as it shoots up from 1,793 in March to 38k in April. The death count continually increases all the way to almost 44k in May, which is the peak of the pandemic (in terms of death) for the US in this dataset. Confirmed cases hovers around the same number until the month of June but increases once again thereafter, which is when the second surge of covid takes place. In July, the death toll increased from 18k to 33k in August, which is the peak of the pandemic (in terms of confirmed cases) with a count of 1.5 million.

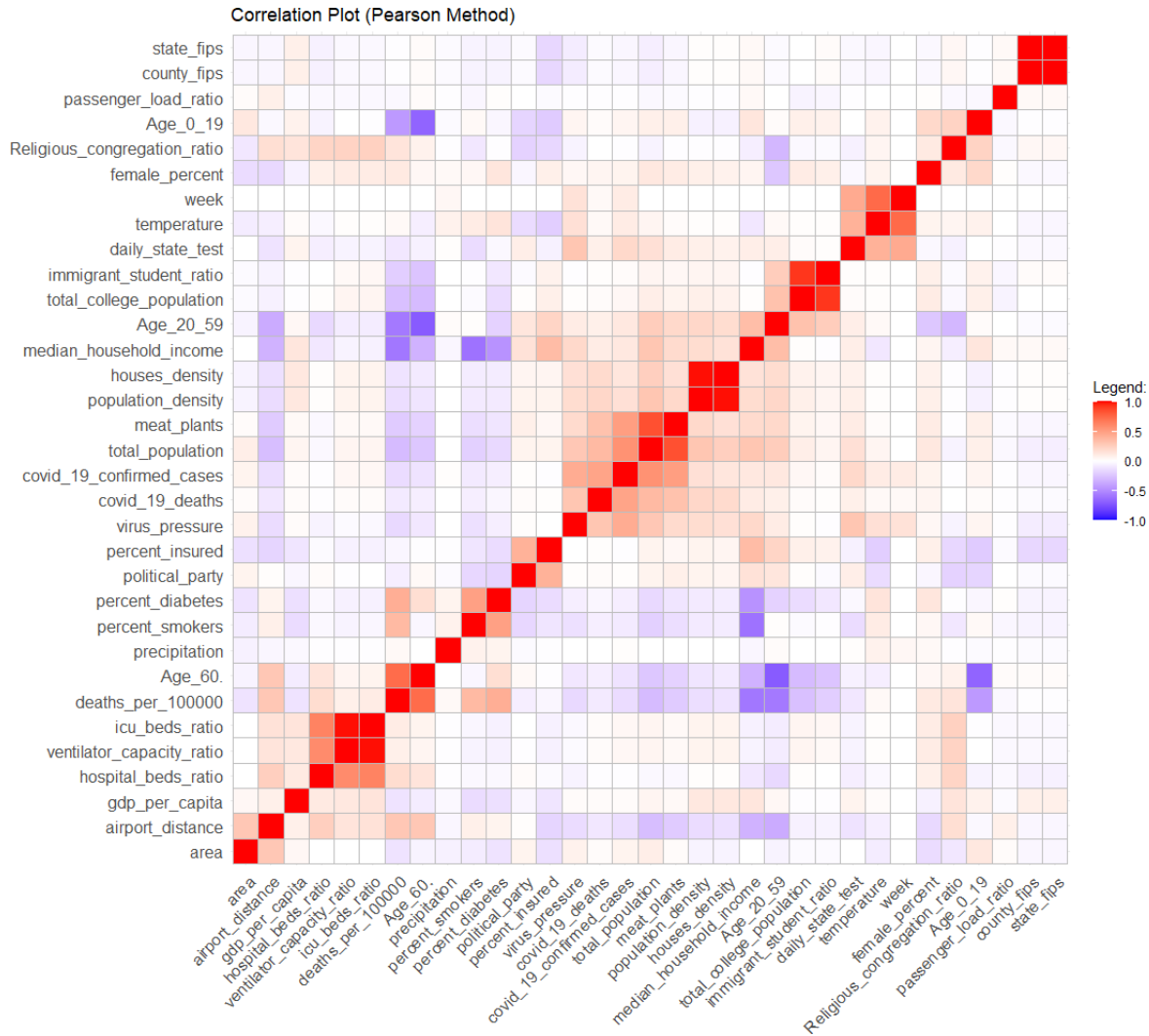
The Centers for Disease Control and Prevention (CDC 2022) reports that in the month of March, President Trump declared a state of emergency and required mass testing nationwide as cases exponentially increase. On March 19 California, one of the number 1 state in terms of confirmed cases and deaths, issues a “stay-at-home” order as confirmed cases steadily increased from the month of March onwards. In addition to this, social distancing measures were tightened starting from April, which is another possible explanation of covid-19 confirmed cases slowed down and plateaued. In June, the World Health Organization declared that the virus can be airborne (AJMC 2021). From this, we see the 2nd surge of covid-19 and the confirmed cases and deaths increase exponentially again.

References:

Centers for Disease Control and Prevention, “CDC Museum COVID-19 Timeline,” Smithsonian Institution January 5 2022, <https://www.cdc.gov/museum/timeline/covid19.html>

AJMC Staff, “A Timeline of COVID-19 Developments in 2020,” The Center for Biosimilars January 1 2021, <https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020>

A. Correlation Plot of All Variables



According to the correlation plot and coefficients, the variable covid_19_confirmed_cases have the highest correlation coefficient with covid19_deaths with a value of 0.64. The proceeding variable that has a high correlation coefficient with confirmed cases is total_population with a coefficient of 0.60. In addition to these, we can infer from the correlation plot that the relationship between these variables is positive. This means that if covid_19_deaths or total_population is high, then covid_19_confirmed_cases will likely increase as well. However, it is important to note that this does not immediately imply causality between the two variables. This merely shows that they have a relationship with one another.

For the succeeding analysis, we will use the result of the correlation plot as our basis which variables we will dig deeper. This way, we can narrow down the impact on cases and deaths. Based on common domain knowledge, we will check which factors we think can affect the cases and deaths. This means that the other variables may not have good correlation value but are still worthy of the analysis.

Confirmed Cases Analysis

A. Correlation Values

	[,1]
week.x	0.100660092
county_fips.x	-0.048698548
covid_19_confirmed_cases	1.000000000
covid_19_deaths	0.641141351
state_fips	-0.048634042
social_distancing_total_grade	0.084377505
daily_state_test	0.214811983
precipitation	0.012948648
temperature	0.117847563
virus_pressure	0.474153200
total_population	0.605453079
female_percent	0.067056324
area	0.060751863
population_density	0.170690338
hospital_beds_ratio	-0.011051323
ventilator_capacity_ratio	0.039574184
icu_beds_ratio	0.038195423
houses_density	0.142151770
total_college_population	0.046455330
percent_smokers	-0.094027015
percent_diabetes	-0.074695840
Religious_congregation_ratio	-0.008801956
political_party	0.005980894
airport_distance	-0.149413142
passenger_load_ratio	-0.016092317
meat_plants	0.548638489
median_household_income	0.126032298
percent_insured	-0.012344338
deaths_per_100000	-0.158198359
gdp_per_capita	0.026470138
Age_0_19	0.059133151
Age_20_59	0.131870985
Age_60.	-0.134433313
immigrant_student_ratio	0.025098413

Figure 1: Correlation values versus confirmed cases

B. Top 5 per Location

On figure 2, the top 5 states in terms of number of covid-19 cases are California, Texas, Florida, New York, and Illinois. Arizona comes in at number 6. When compared to confirmed cases, these states are also part of the top 5 of those variables, which makes sense. In theory, if there are more cases, there will most likely be more deaths as well.

In the following section (Figure 1), we will be analyzing the dataset according to covid19 confirmed cases and deaths separately. We will examine the specific variables that may or may not have an effect on these 2 variables based on theory. For example, we can say that total_college_population has an effect on covid19 confirmed cases because education possibly has an important role on people's views on getting vaccinated, and thus would lead to less deaths. The same thing can be said for meat plants and GDP per capita. In terms of meat plants, we can theoretically say that diet has an effect on a person's ability to fight or get covid. In terms of GDP per capita, we can infer that people's incomes dictate how much medicine they can afford as well. Therefore, in this next section, we will explore if these variables truly contribute to confirmed cases.

Location_case

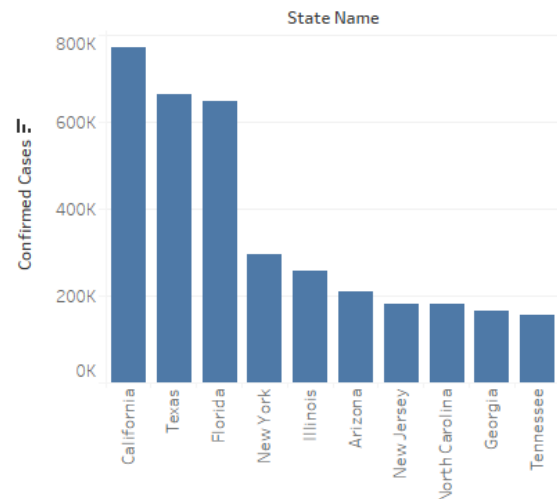


Figure 2: Confirmed cases per state

Top 5 states in terms of gdp_per_capita		Top 5 states in terms of meat plants	
Texas	624,087	California	23,835
Kansas	170,617	Texas	16,485
Iowa	158,757	Illinois	14,455
Nebraska	140,420	Pennsylvania	14,105
Minnesota	120,658	Florida	11,970

Top 5 states in terms of total_college_population	
Texas	2,254
Kansas	1,420
Virginia	1,272
Missouri	1,255
North Carolina	1,238

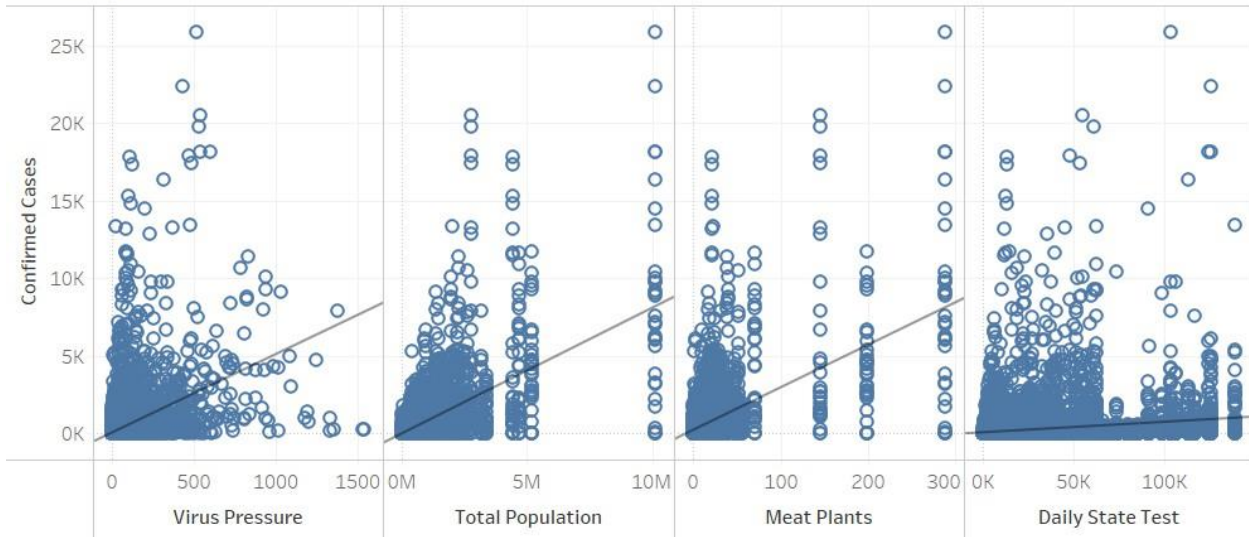
Based on the lists above, we can get a clearer picture of which variables have an effect on confirmed cases. If the top 5 states of these variables contribute to confirmed cases, then they should also be near the right side of our bar graph. After further analysis, we can say that among these 3 variables, meat_plants would have to be the variable that contributes the most to confirmed cases. We can see that among the top 5 states that have the greatest number of meat plants, the majority of them have really high confirmed cases. For example, California, Texas, Illinois, and Florida are in the top 5 in terms of confirmed cases, while at the same time in the top of 5 in terms of meat plants. This gives us reason that meat plants will possibly have a strong contributing factor to confirmed cases. The same cannot be said for gdp_per_capita and total_college_population. The top 5 states in both of these variables have really low confirmed cases relative to the other states. Therefore, we can infer that these 2 variables do not contribute to confirmed cases as much as meat plants.

With further examination, looking at the correlation coefficients of these variables tells the same story. The correlation coefficient of gdp per capita and total college population are 0.026 and 0.046 respectively, which are relatively low when compared to the correlation coefficient of meat plants, which is 0.548. Thus, this confirms our theory that diet may have an important impact on confirmed cases. This can be confirmed through other highly correlated variables with covid 19 confirmed cases. Both virus pressure and daily state test have relatively high correlation coefficients with values of 0.474 and 0.214, respectively. As seen, the top 5 states in these 2 variables are also part of the top 10 states in terms of confirmed cases which means that they have more of an impact on covid 19 confirmed cases than other variables.

Top 5 states in terms of virus pressure		Top 5 states in terms of daily_state_test	
California	111,277	Texas	153,050,644
Texas	89,580	California	107,312,343
Florida	83,631	New York	75,702,551
Arizona	37,871	Illinois	52,506,145
New York	34,634	Florida	41,236,536

C. Selected Scatter Plots versus Confirmed Cases

Scatterplot_case



These 4 variables have the highest correlation coefficients to covid-19 confirmed cases. As seen in the correlation coefficients table, virus pressure, total population, meat plants, and daily state test have correlation coefficients of 0.474, 0.60, 0.54, and 0.21, respectively. The scatterplots shown above give us a visual representation of these individual relationships. The slopes of the scatterplots show that the relationship is increasing and positive. This means that from the data, if virus pressure, total population, meat plants, and daily state test individually increase, then confirmed cases will likely increase as well. In addition to this, the slope of the scatterplot is steeper if the correlation coefficient is higher, which means that confirmed cases increase at a higher rate with the variables of higher coefficients.

Death Cases Analysis

Correlation Plot versus deaths cases

	[,1]
week.x	0.0383753424
county_fips.x	-0.0278140972
covid_19_confirmed_cases	0.6411413514
covid_19_deaths	1.0000000000
state_fips	-0.0276728379
social_distancing_total_grade	0.0307510810
daily_state_test	0.1054815463
precipitation	0.0091544272
temperature	0.0457891628
virus_pressure	0.4011166551
total_population	0.4576261827
female_percent	0.0656503880
area	0.0217860453
population_density	0.2727449059
hospital_beds_ratio	-0.0077721036
ventilator_capacity_ratio	0.0292181277
icu_beds_ratio	0.0288622243
houses_density	0.2380982452
total_college_population	0.0312815124
percent_smokers	-0.0836150216
percent_diabetes	-0.0575498945
Religious_congregation_ratio	-0.0001965108
political_party	0.0308086655
airport_distance	-0.1260215367
passenger_load_ratio	-0.0101387059
meat_plants	0.4131992069
median_household_income	0.1229562343
percent_insured	0.0172492832
deaths_per_100000	-0.1148515848
gdp_per_capita	0.0301774686
Age_0_19	0.0244814998
Age_20_59	0.1048739340
Age_60.	-0.0897724387
immigrant_student_ratio	0.0156103948

Figure 3: Correlation values vs death

On figure 3, from the correlation coefficients of all variables against covid 19 deaths, the most correlated variable is covid 19 confirmed cases. The following most correlated variable has a value of 0.457, which is also total population similar to confirmed cases. Further, the relationship of these variables are significantly positive. In addition to this, we can utilize the same process to verify the theoretical relationships to covid19 deaths, and investigate which variables truly have a relationship with covid19 deaths. In theory, being a smoker and/or having diabetes would mean you are less healthy than the individuals who don't smoke or don't have diabetes. This would also mean that you are more susceptible to death due to covid 19:

Top 5 per location

On figure 4, the top 5 states in terms of covid-19 deaths are California, Texas, Florida, New York, and New Jersey. Illinois comes in at number 6. When compared to confirmed cases, these states are also part of the top 5 of that variable, which makes sense. In theory, if there are more cases, there will most likely be more deaths as well.

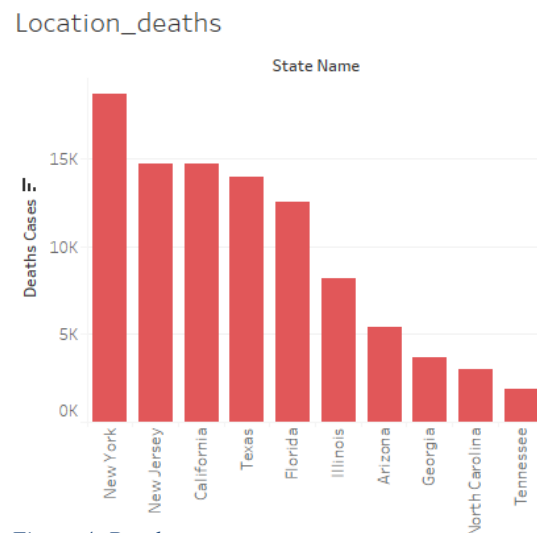


Figure 4: Death cases per state

Top 5 states in terms of percent smoker		Top 5 states in terms of percent diabetes	
Texas	96,017	Texas	73,619
Missouri	70,424	Missouri	44,447
Tennessee	58,168	North Carolina	39,718
Kansas	54,518	Kansas	37,660
North Carolina	53,284	Tennessee	37,387

Interestingly, the top 5 states of percent smoker and diabetes are identical just in different order. Therefore, we can interpret them together in terms of covid 19 deaths. In theory, these variables may have a significant effect on covid 19 deaths. However, we can see from the tables above that only the state of Texas is in the top 5 in terms of covid 19 deaths, while the other 4 states in the list have relatively lower death counts. This gives us reason that in reality, percent smoker and percent diabetes do not have a contributing factor to covid-19 death.

Top 5 states in terms of hospital beds ratio		Top 5 states in terms of ventilator capacity ratio	
Kansas	20	Kansas	2
Texas	12	Texas	2
Minnesota	10	Nebraska	1
Iowa	9	Iowa	1
Nebraska	9	North Carolina	1

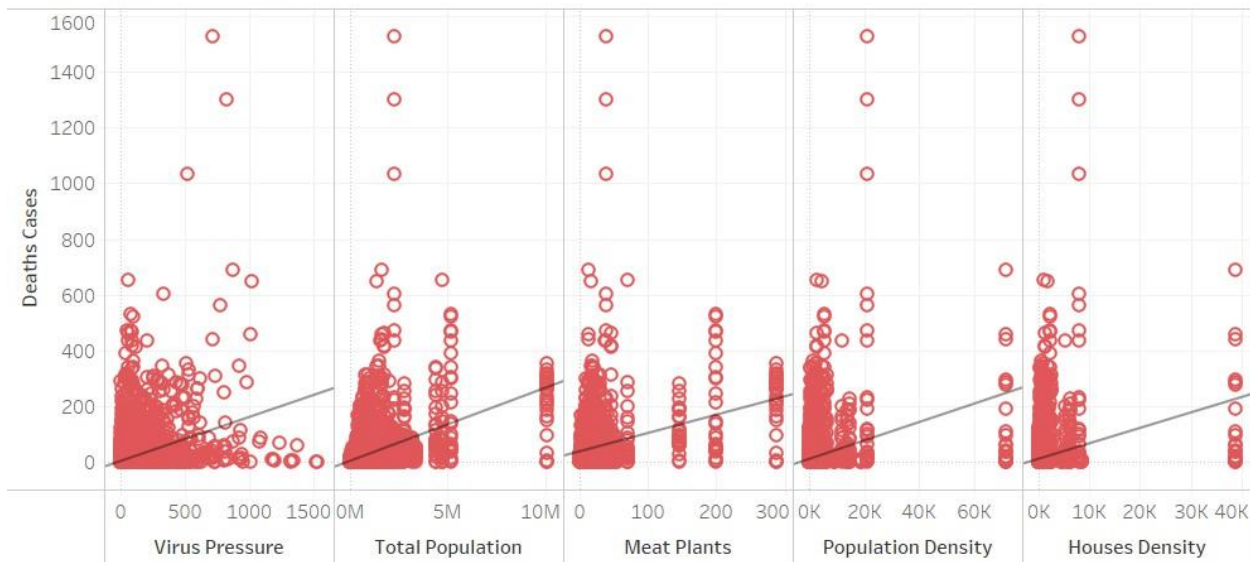
Top 5 states in terms of icu beds ratio	
Kansas	2
Texas	2
Nebraska	1
Iowa	1
North Carolina	1

Similar to the variable pairing before this, the top 5 states for hospital beds ratio, ventilator capacity ratio, and icu beds ratio are all alike with each other. Again, only Texas is the only state in the top 5 in terms of covid-19 deaths. In theory, hospital beds ratio, ventilator capacity ratio, and icu beds ratio have a significant relationship to covid-19 deaths because if these services are not available during the pandemic then the covid-19 death count will increase. However in reality, from the tables above, only the state of Texas is in the top 5 of covid-19 deaths and the other 4 states are significantly low. Therefore, they do not contribute heavily to covid-19 deaths.

These hypotheses can be confirmed from the correlation coefficients. Percent smokers and percent diabetes have correlation coefficients of -0.08 and -0.05 respectively. Hospital beds ratio, ventilator capacity ratio, and icu beds ratio have correlation coefficients of -0.007, 0.029, and 0.028, respectively. With their correlation coefficients being significantly low, there is no surprise that the top 5 states in these respective variables do not appear in the top 5 states in terms of covid-19 deaths.

- **Scatter plot other vs deaths**

Scatterplot_death



The variables chosen from the above scatterplots are of the highest correlation coefficients in relation to covid-19 deaths. Virus pressure, total population, meat plants, population density, and houses density have correlation coefficients of 0.40, 0.457, 0.41, 0.27, 0.23, respectively. The scatterplots above show the visual relationship between these variables individually to covid-19 deaths. In addition, the slope of these scatterplots are also positive as they have an upward slope. This means that covid-19 death counts are likely to increase if these variables increase individually. Furthermore, the slopes of these variables are steeper if the individual correlation coefficients are higher, which is similar to the situation in confirmed cases.

Inferential Analysis

From the descriptive analysis, we know that confirmed cases and deaths have very high correlation at 64% followed by total population at 60%. We found that confirmed cases are highly correlated with the following variables: virus pressure, total population, meat plants and daily state tests. For death cases the following variables are highly correlated as well: virus pressure, total population, population density and houses density. However, this analysis only showed that they have relationship. We will now look at if the following variables are statistically significant. Do they really impact the increase in confirmed and death cases?

On the 1st part of the analysis, we know that living in high population density states like California, Texas, Florida, and New York can increase your chance of getting the virus. While, living in less population density states like Alaska, Wyoming and North Dakota have lesser chance of getting the virus. From this, we will look at what factors are driving the increase in cases for both groups (higher versus lower population density states). We then check if a relationship exists between these groups.

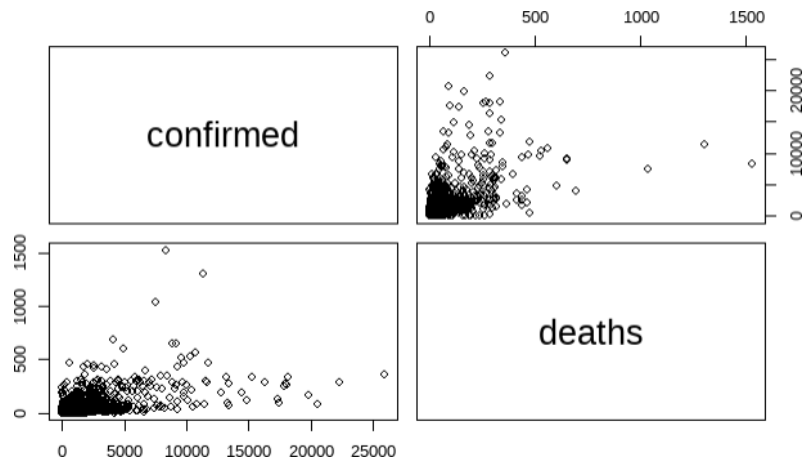
Overview

Before we proceed with the inferential analysis of the data, we checked the distribution of our response variables: confirmed and death cases. The boxplot (see figure A in appendix C) showed several points outside the whiskers for both confirmed and death cases. The results showed that the data is not normally distributed. When checking the distribution of the data, it is not normally distributed, same is true for most of the remaining variables (see figure B in appendix C). For confirmed and death cases histogram are highly skewed to the right. The QQ plot (see figure C in appendix C) and Kolmogorov-Smirnov Test (see figure D in appendix C) tells the same.

We checked what was affecting the distribution, we observed that 36.68% (or 30,200) of the data in the confirmed cases are equal to zero and 77.20% (or 63,551) of the death cases are equal to zero. When the two variables are combined, there are a total of 36.4% (or 30,021) are equal to zero and either of them will result to 77.4% (or 63,730) is equal to zero. This means that the data has a lot of zero recorded cases or deaths per country per week from the start of the observations in January to September of 2020.

We tried implementing a cut off start point on March 11, 2020, where WHO officially announced the spread of COVID-19 as a pandemic. The result showed that 21% (or 13,842) of the data is equal to zero for confirmed cases and 77.19%

(or 47,097) of death cases are equal to zero. This means that even during the start of the pandemic, there were zero recorded cases or deaths per country per week. Zero cases were not only recorded before the WHO announcement but also after the announcement. You can see from the graph on the left the scatter plot distribution of confirmed and death cases. All the observations are concentrated on the lower left portion.



With this, we tried to cut off some of the extreme points from the data to create a bit reliable model later. We used the Mahalanobis Distance technique. From this, it suggested to remove 1,271 observations outside the 95% confidence interval. The new dataset has now 81,049 observations from 82,320.

General Analysis

From the full model with **confirmed cases** as the response variable (see figure E from appendix C), we can see that all variables are statistically significant but when checked for multicollinearity it, several variables have VIF value of more than 5 like population density at 94.63, ventilator at 53.52, icu bed ratio at 58.71, houses density at 90.73, age 0 to 19 at 99.3, age 20 to 59 at 114.29, age 60+ at 212.71 and immigration student at 7.15.

After removing certain variables that have multicollinearity such as houses density, ventilator, age 0 to 19 and immigration student, the number of significant variables were reduced. The variables hospital bed and college population became less significant in the model. Comparing the model from the descriptive part, the variable meat plants resulted in negative beta which is opposite. This means that, as number of meat plants decreases the number of confirmed cases increase.

From the full model with **deaths cases** as the response variable (see figure G from appendix C), almost all were statistically significant except the following variables: hospital beds, college population, percent smokers, religious congregation ratio, passenger load, income, percent insured, age 0 to 19, age 20 to 59, age 60 plus and immigration student. When running VIF for multicollinearity, the result showed almost the same with the confirmed cases full model.

After removing certain variables that have multicollinearity such as houses density, ventilator, age 0 to 19 and immigration student, the number of significant variables were reduced. The variables temperature, hospital bed, college population percent smokers, religious congregation ratio, passenger load, percent insured, and age 60+ became less significant in the model. Similar to findings in confirmed case, when comparing the model from the descriptive part, the variable meat plants resulted in negative beta which is opposite. This means that, as number of meat plants decreases the number of death cases increase.

Confirmed Cases Analysis

Top 5 states:

1. California
2. Texas
3. Florida
4. New York
5. Illinois

We filtered the dataset based on the performance of top 5 states. We were able to get 7,953 observations. We run linear regression on full and reduced model based on stepwise selection and after removal of multicollinearity. Below is the result.

Full model	Reduced model
<pre>Call: lm(formula = confirmed ~ ., data = county_fips - state_fips - mahal - badmahal, data = confirmedT5) Residuals: Min 1Q Median 3Q Max -645.66 -39.07 -6.63 21.50 793.68 Coefficients: (Intercept) -2.414e+02 3.428e+02 -0.704 0.481428 deaths 1.735e+01 2.635e-01 65.840 < 2e-16 *** social_dist 1.027e+00 5.315e-01 1.931 0.053460 daily_state_test 1.163e-03 5.223e-05 22.260 < 2e-16 *** precipitation 3.629e-02 2.293e-02 1.583 0.113556 temperature -4.305e-01 1.949e-01 -2.209 0.027187 * virus_pressure 5.738e-01 1.871e-02 30.659 < 2e-16 *** total_population 3.907e-05 4.886e-06 7.996 1.47e-15 *** female_percent -9.612e+01 8.030e+01 -1.197 0.231325 area 4.832e-05 9.217e-04 0.052 0.958195 population_density 2.601e-02 3.352e-03 7.759 9.59e-15 *** hosp_beds -3.130e+03 1.008e+03 -3.105 0.001909 ** ventilator -4.286e+03 4.055e+04 -0.106 0.915810 icu_beds_ratio 6.314e+04 4.664e+04 1.354 0.175842 houses_density -4.969e-02 6.336e-03 -7.843 4.97e-15 *** college_pop 4.046e+01 6.777e+00 5.971 2.46e-09 *** percent_smokers 4.303e-01 8.473e-01 0.508 0.611567 percent_diabetes 2.193e-01 4.256e-01 0.515 0.606351 religious_congregation_ratio 3.796e-01 9.935e-02 3.821 0.000134 *** political_party -7.029e+01 8.295e+00 -8.474 < 2e-16 *** airport_distance -2.745e-02 3.034e-02 -0.905 0.365534 pass_load 1.088e+00 3.092e+00 0.352 0.724829 meat_plants -8.861e-01 1.625e-01 -5.453 5.09e-08 *** income 4.944e-04 1.466e-04 3.372 0.000750 *** percent_insured 2.038e+00 6.938e-01 2.937 0.003322 *** deaths_per_100000 1.202e-02 1.110e-02 1.083 0.278954 gdp_per_capita 2.796e-01 7.525e-02 3.716 0.000204 *** Age_0_19 4.936e+00 3.475e+00 1.420 0.155522 Age_20_59 -1.158e+00 3.500e+00 -0.331 0.740722 Age_60 -5.926e-01 3.444e+00 -0.172 0.863392 immig_student -7.544e+02 1.632e+02 -4.623 3.84e-06 *** Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 94.56 on 7922 degrees of freedom Multiple R-squared: 0.6478, Adjusted R-squared: 0.6464 F-statistic: 485.7 on 30 and 7922 DF, p-value: < 2.2e-16</pre>	<pre>Call: lm(formula = confirmed ~ +virus_pressure + daily_state_test + total_population + airport_distance + ventilator + Age_0_19 + income + meat_plants + college_pop + Religious_congregation_ratio + percent_insured + social_dist + pass_load + hosp_beds + gdp_per_capita, data = ConfirmedT5) Residuals: Min 1Q Median 3Q Max -1063.84 -52.97 -12.63 24.85 840.80 Coefficients: (Intercept) 2.120e+02 3.138e+01 6.755 1.53e-11 *** virus_pressure 8.130e-01 2.279e-02 35.680 < 2e-16 *** daily_state_test 1.340e-03 4.532e-05 29.577 < 2e-16 *** total_population 1.149e-04 5.313e-06 21.622 < 2e-16 *** airport_distance -2.906e-01 3.130e-02 -9.285 < 2e-16 *** ventilator 8.506e+04 9.177e+03 9.268 < 2e-16 *** Age_0_19 4.848e+00 4.365e-01 11.107 < 2e-16 *** income 5.513e-04 1.273e-04 4.331 1.50e-05 *** meat_plants -2.746e+00 1.933e-01 -14.204 < 2e-16 *** college_pop 1.306e+01 2.704e+00 4.830 1.39e-06 *** Religious_congregation_ratio 4.831e-01 1.203e-01 4.017 5.95e-05 *** percent_insured -4.461e+00 3.763e-01 -11.857 < 2e-16 *** social_dist 1.850e+00 6.417e-01 2.883 0.00395 ** pass_load 3.560e+00 3.611e+00 0.986 0.32414 hosp_beds -2.207e+03 1.017e+03 -2.169 0.03010 * gdp_per_capita 1.452e-03 6.484e-02 0.022 0.98213 --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 119.5 on 7937 degrees of freedom Multiple R-squared: 0.4361, Adjusted R-squared: 0.435 F-statistic: 409.2 on 15 and 7937 DF, p-value: < 2.2e-16</pre>

Regression Model	R-squared	Adjusted R-squared	RMSE	MAE
Full model	0.64	0.64	94.56	53.43
Reduced model	0.43	0.43	119.5	68.98

From the initial inspection of the results, the statistical measures indicate that the full model performs much better than the reduced model. The adjusted R-square value of the full model is much almost 20% higher than the adjusted R-square value of the reduced model. Further, a lower RMSE and MAE is desirable. In the case of our models, both the RMSE and MAE of the full model are lower than that of the reduced model. Although the statistical measures look better for the full model, it is observable that there are numerous statistically insignificant variables in the full model. On the other hand, the reduced model has only 2 variables that are not statistically significant. The reason why the R-squared and adjusted R-squared values are higher for the full model is because the full model includes more variables. As we know, the R-squared values will increase when there are more observations.

Bottom 5 states:

1. Alaska
2. Wyoming
3. New Hampshire
4. West Virginia
5. North Dakota

We filtered the dataset based on the performance of bottom 5 states. We were able to get 3,010 observations. We run linear regression on full and reduced model based on stepwise selection and after removal of multicollinearity. Below is the result.

Full model

Call:

lm(formula = confirmed ~ . - date - county_fips - state_fips - mahal - badmahal, data = confirmedB5)

Residuals:

	Min	1Q	Median	3Q	Max
	-122.46	-8.01	-0.46	5.65	384.30

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.975e+01	1.360e+02	0.513	0.608000
deaths	1.439e+01	6.236e-01	23.075	< 2e-16 ***
social_dist	4.846e-01	1.934e-01	2.506	0.012255 *
daily_state_test	4.587e-03	4.488e-04	10.220	< 2e-16 ***
precipitation	-3.569e-02	1.858e-02	-1.921	0.054808 .
temperature	-8.498e-02	7.430e-02	-1.144	0.252838
virus_pressure	1.308e+00	2.560e-01	5.107	3.48e-07 ***
total_population	2.495e-04	2.378e-05	10.490	< 2e-16 ***
female_percent	-3.431e+01	4.008e+01	-0.856	0.391993
area	-1.208e-04	1.704e-04	-0.709	0.478422
population_density	8.727e-02	7.551e-02	1.156	0.247916
hosp_beds	-7.298e+02	1.884e+02	-3.873	0.000110 ***
ventilator	2.963e+04	8.352e+03	3.548	0.000394 ***
icu_beds_ratio	-2.185e+04	9.265e+03	-2.359	0.018396 *
houses_density	-2.164e-01	1.603e-01	-1.350	0.176997
college_pop	-4.165e+00	3.523e+00	-1.182	0.237217
percent_smokers	-5.013e-01	2.228e-01	-2.250	0.024547 *
percent_diabetes	-2.190e-01	2.058e-01	-1.064	0.287294
Religious_congregation_ratio	3.239e-02	2.932e-02	1.105	0.269376
political_party	NA	NA	NA	NA
airport_distance	5.816e-03	8.913e-03	0.653	0.514083
pass_load	9.274e-03	4.552e-02	0.204	0.838577
meat_plants	1.330e+00	6.104e-01	2.179	0.029434 *
income	9.549e-06	9.473e-05	0.101	0.919715
percent_insured	-1.066e-01	2.015e-01	-0.529	0.596807
deaths_per_100000	-4.977e-03	3.671e-03	-1.356	0.175304
gdp_per_capita	-4.066e-02	1.672e-02	-2.431	0.015119 *
Age_0_19	-6.308e-01	1.327e+00	-0.476	0.634448
Age_20_59	-1.280e-01	1.297e+00	-0.099	0.921397
Age_60	-6.627e-01	1.310e+00	-0.506	0.613009
immig_student	8.914e+01	7.635e+01	1.167	0.243109

Reduced model

Call:

lm(formula = confirmed ~ +deaths + total_population + daily_state_test + percent_diabetes + virus_pressure + houses_density + deaths_per_100000 + ventilator + hosp_beds + percent_smokers + gdp_per_capita + social_dist + female_percent + meat_plants + precipitation, data = confirmedB5)

Residuals:

	Min	1Q	Median	3Q	Max
	-124.94	-8.02	-0.73	5.81	383.97

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.171e+01	1.507e+01	3.432	0.000608 ***
deaths	1.450e+01	6.185e-01	23.445	< 2e-16 ***
total_population	2.600e-04	2.055e-05	12.652	< 2e-16 ***
daily_state_test	4.321e-03	2.954e-04	14.629	< 2e-16 ***
percent_diabetes	-2.746e-01	1.867e-01	-1.471	0.141513
virus_pressure	1.193e+00	2.456e-01	4.857	1.26e-06 ***
houses_density	-4.215e-02	1.575e-02	-2.677	0.007477 **
deaths_per_100000	-8.548e-03	2.206e-03	-3.875	0.000109 ***
ventilator	1.005e+04	1.578e+03	6.371	2.17e-10 ***
hosp_beds	-7.041e+02	1.773e+02	-3.972	7.29e-05 ***
percent_smokers	-4.758e-01	1.354e-01	-3.513	0.000449 ***
gdp_per_capita	-4.689e-02	1.378e-02	-3.402	0.000677 ***
social_dist	5.979e-01	1.834e-01	3.261	0.001122 **
female_percent	-8.366e+01	3.122e+01	-2.680	0.007411 **
meat_plants	1.343e+00	5.531e-01	2.428	0.015235 *
precipitation	-3.491e-02	1.800e-02	-1.939	0.052602 .

Regression Model	R-squared	Adjusted R-squared	RMSE	MAE
Full model	0.456	0.451	24.11	11.34
Reduced model	0.453	0.451	24.11	11.29

The same conclusions can be drawn for the bottom 5 states for confirmed cases. However, in this case, the 4 statistical measures do not decrease as much compared to the case of top 5 states in confirmed cases as previously shown. This confirms our assumptions that the reduced model is superior because even if we filter out statistically insignificant variables, the reduced model still performs as well as the full model. In addition to this, the results, and coefficients from both the top 5 and bottom 5 states make sense in theory. For example, the coefficient of daily_state_test is positive in both models. This means that as the number of daily_state_tests increase, so will the count for confirmed cases.

Another interesting variable is social_distancing. For the bottom 5 states, social_distancing has a much higher positive coefficient compared to that from the top 5 states. This can imply that the top 5 states should learn from the bottom 5 states and improve their social distancing measures as this would be a helpful way to alleviate covid19 confirmed cases.

However, there are also variables that have an opposite effect when comparing the top 5 and bottom 5 states. An example of this would be the variable meat_plants. In the case of the top 5 states, the coefficient for meat_plants is negative. However, in the case of the bottom 5 states, the coefficient for

meat_plants is positive. Variables such as these that have contradicting coefficients may need some further investigation. A possible explanation for this is that the top 5 states have more meat plants.

With this, we checked if there is a relationship between the top and bottom 5 states by using one-way ANOVA. We first created a new variable for each data frame and label top and bottom based on their category. After this, we run one-way ANOVA for the rank (top and bottom 5 states categories) with the number of confirmed cases.

The result showed a p-value that is less than 0.05. This means we will need to reject the null hypothesis that both groups are equal. This means, that there is no relationship between top and bottom 5 states in terms confirmed cases.

```
# anova
summary(aov(confirmed~rank, data = Confirmed_rank))

      Df    Sum Sq Mean Sq F value Pr(>F)
rank      1    8519678  8519678   457.1 <2e-16 ***
Residuals 10961 204317940   18640
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

p value is less than 0.05, so we reject null hypothesis that both groups are equal.
```

Death Cases Analysis

Top 5 states

1. New York
2. New Jersey
3. California
4. Texas
5. Florida

We filtered the dataset based on the performance of top 5 states. We were able to get 7,953 observations. We run linear regression on full and reduced model based on stepwise selection and after removal of multicollinearity. Below is the result.

Full model

Call:

lm(Formula = deaths ~ . - date - county_fips - state_fips - mahal - badmahal, data = Deathst5)

Residuals:

Min	1Q	Median	3Q	Max
-16.2420	-0.8893	-0.3200	0.3339	29.9135

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.509e+00	8.403e+00	-0.418	0.676205
confirmed	2.007e-02	2.582e-04	77.758	< 2e-16 ***
social_dist	-2.024e-02	1.394e-02	-1.452	0.146508
daily_state_test	-2.865e-05	1.684e-06	-17.010	< 2e-16 ***
precipitation	2.433e-03	6.737e-04	3.611	0.000306 ***
temperature	9.194e-02	5.927e-03	15.511	< 2e-16 ***
virus_pressure	-2.051e-03	5.888e-04	-3.483	0.000498 ***
total_population	7.319e-07	1.359e-07	5.385	7.39e-08 ***
female_percent	5.261e+00	1.762e+00	2.985	0.002838 **
area	1.959e-05	2.877e-05	0.681	0.495930
population_density	-2.753e-04	9.495e-05	-2.899	0.003750 **
hosp_beds	6.847e+01	2.588e+01	2.645	0.008171 **
ventilator	1.683e+02	6.214e+02	0.271	0.786583
icu_beds_ratio	-4.040e+02	7.327e+02	-0.551	0.581356
houses_density	5.133e-04	1.781e-04	2.882	0.003957 **
college_pop	-3.656e-01	1.340e-01	-2.728	0.006380 **
percent_smokers	5.332e-02	1.699e-02	3.138	0.001707 **
percent_diabetes	-2.685e-03	8.740e-03	-0.307	0.758661
Religious_congregation_ratio	6.973e-04	2.379e-03	0.293	0.769464
political_party	2.236e+00	1.705e-01	13.113	< 2e-16 ***
airport_distance	-2.793e-03	6.828e-04	-4.090	4.34e-05 ***
pass_load	3.137e-03	1.012e-02	0.310	0.756616
meat_plants	-1.109e-02	5.370e-03	-2.065	0.038953 *
income	3.519e-06	3.434e-06	1.025	0.305541
percent_insured	-5.153e-02	1.131e-02	-4.558	5.22e-06 ***
deaths_per_100000	-1.041e-05	2.341e-04	-0.044	0.964552
gdp_per_capita	1.283e-04	2.159e-04	0.594	0.552425
Age_0_19	2.068e-02	8.434e-02	0.245	0.806339
Age_20_59	2.805e-02	8.433e-02	0.333	0.739420
Age_60	5.292e-02	8.360e-02	0.633	0.526705
immig_student	6.336e+00	2.580e+00	2.456	0.014066 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.201 on 12176 degrees of freedom

Multiple R-squared: 0.4761, Adjusted R-squared: 0.4748

F-statistic: 368.8 on 30 and 12176 DF, p-value: < 2.2e-16

Reduced model

Call:

lm(Formula = deaths ~ +confirmed + total_population + daily_state_test + temperature + Age_20_59 + precipitation + airport_distance + virus_pressure + hosp_beds + percent_smokers + percent_insured + Age_60 + meat_plants, data = Deathst5)

Residuals:

Min	1Q	Median	3Q	Max
-16.3600	-0.8093	-0.3268	0.2339	30.4127

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.209e+00	7.213e-01	-3.062	0.00220 **
confirmed	1.975e-02	2.559e-04	77.168	< 2e-16 ***
total_population	5.474e-07	1.236e-07	4.430	9.51e-06 ***
daily_state_test	-1.923e-05	1.508e-06	-12.746	< 2e-16 ***
temperature	5.586e-02	5.234e-03	10.671	< 2e-16 ***
Age_20_59	-6.321e-02	1.130e-02	-5.593	2.28e-08 ***
precipitation	2.038e-03	6.758e-04	3.016	0.00257 **
airport_distance	-1.683e-03	5.513e-04	-3.052	0.00228 **
virus_pressure	-1.621e-03	5.785e-04	-2.803	0.00508 **
hosp_beds	4.539e+01	1.582e+01	2.870	0.00411 **
percent_smokers	-9.267e-03	1.230e-02	-0.753	0.45125
percent_insured	6.324e-02	6.113e-03	10.346	< 2e-16 ***
Age_60	8.429e-04	7.643e-03	0.110	0.91218
meat_plants	1.550e-03	5.120e-03	0.303	0.76206

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.228 on 12193 degrees of freedom

Multiple R-squared: 0.4667, Adjusted R-squared: 0.4661

F-statistic: 820.7 on 13 and 12193 DF, p-value: < 2.2e-16

Regression Model	R-squared	Adjusted R-squared	RMSE	MAE
Full model	0.4761	0.478	3.201	1.546
Reduced model	0.4667	0.4661	3.228	1.511

The statistical measures indicate that the full model for deaths performed better than the reduced model except for MAE. Although the R-squared, Adjusted R-squared, and RMSE are all more favorable in the full model, it cannot be automatically assumed that it is better than the reduced model. After further investigation there are numerous statistically insignificant variables included in the full model. After filtering out some of these variables using the stepwise selection method, we are left with only 3 insignificant variables in the reduced model. In addition to this, the same conclusions can be drawn from the confirmed cases. Both the R-squared and Adjusted R-squared values are higher in the reduced model simply because there are more variables and observations in the full model compared to the reduced model. This will inevitably increase the R-squared and adjusted R-squared values.

Bottom 5 states

1. Alaska
2. Wyoming
3. Vermont
4. Hawaii
5. Maine

We filtered the dataset based on the performance of bottom 5 states. We were able to get 1,885 observations. We run linear regression on full and reduced model based on stepwise selection and after removal of multicollinearity. Below is the result.

Full model	Reduced model
<pre> call: lm(formula = deaths ~ ., data = county_fips - state_fips - mahal - badmahal, data = DeathsB5) Residuals: Min 1Q Median 3Q Max -7.5316 -0.1506 -0.0363 0.0671 7.1162 Coefficients: (Intercept) 7.160e-02 5.488e+00 0.013 0.9896 confirmed 1.411e-02 5.678e-04 24.860 < 2e-16 *** social_dist -3.647e-02 5.602e-03 -6.510 9.63e-11 *** daily_state_test -2.445e-05 1.858e-05 -1.316 0.1883 precipitation -1.200e-04 6.297e-04 -0.191 0.8489 temperature -2.763e-03 1.939e-03 -1.425 0.1544 virus_pressure -1.536e-02 1.021e-02 -1.504 0.1328 total_population 3.436e-07 5.195e-07 0.661 0.5085 female_percent 2.027e+00 2.612e+00 0.776 0.4378 area 3.810e-06 8.836e-06 0.431 0.6664 population_density -5.183e-03 1.069e-03 -4.850 1.34e-06 *** hosp_beds -7.299e+00 1.183e+01 -0.617 0.5374 ventilator 5.520e+02 5.701e+02 0.968 0.3331 icu_beds_ratio -3.627e+02 6.455e+02 -0.562 0.5743 houses_density 1.450e-02 2.520e-03 5.756 1.01e-08 *** college_pop -1.849e-01 1.669e-01 -1.108 0.2682 percent_smokers -2.572e-02 1.469e-02 -1.750 0.0802 . percent_diabetes 3.043e-02 1.212e-02 2.512 0.0121 * Religious_congregation_ratio political_party 9.845e-03 5.644e-02 0.174 0.8616 airport_distance -2.109e-04 4.677e-04 -0.451 0.6521 pass_load 2.907e-03 2.270e-02 0.128 0.8981 meat_plants -1.096e-02 1.415e-02 -0.775 0.4386 income -1.579e-06 3.385e-06 -0.466 0.6410 percent_insured -5.927e-03 8.702e-03 -0.681 0.4959 deaths_per_100000 -2.872e-05 1.677e-04 -0.171 0.8641 gdp_per_capita 7.760e-05 1.273e-03 0.061 0.9514 Age_0_19 4.421e-03 6.138e-02 0.072 0.9426 Age_20_59 4.401e-03 5.346e-02 0.082 0.9344 Age_60 -1.691e-02 5.740e-02 -0.295 0.7684 immig_student 5.249e+00 3.990e+00 1.316 0.1885 --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 0.5975 on 1854 degrees of freedom Multiple R-squared: 0.3831, Adjusted R-squared: 0.3731 F-statistic: 38.38 on 30 and 1854 DF, p-value: < 2.2e-16 </pre>	<pre> call: lm(formula = deaths ~ +confirmed + female_percent + social_dist + Age_60 + temperature + airport_distance + gdp_per_capita, data = DeathsB5) Residuals: Min 1Q Median 3Q Max -8.0221 -0.1443 -0.0514 0.0289 7.2352 Coefficients: (Intercept) -2.7280087 0.6216608 -4.388 1.21e-05 *** confirmed 0.0148675 0.0005038 29.512 < 2e-16 *** female_percent 6.4735686 1.2755469 5.075 4.26e-07 *** social_dist -0.0285662 0.0052994 -5.390 7.92e-08 *** Age_60 -0.0082490 0.0039346 -2.097 0.036168 * temperature -0.0055754 0.0014574 -3.826 0.000135 *** airport_distance -0.0007928 0.0002335 -3.395 0.000701 *** gdp_per_capita 0.0016603 0.0007314 2.270 0.023318 * --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 0.6083 on 1877 degrees of freedom Multiple R-squared: 0.3528, Adjusted R-squared: 0.3503 F-statistic: 146.1 on 7 and 1877 DF, p-value: < 2.2e-16 </pre>

Regression Model	R-squared	Adjusted R-squared	RMSE	MAE
Full model	0.383	0.373	0.597	0.239
Reduced model	0.352	0.350	0.608	0.229

Similar to the previous results from the top 5 in terms of death counts, the statistical models indicate that the full model performs better than the reduced model except for MAE. However, as mentioned, the R-squared and adjusted R-squared values are higher for the full model simply because there are more observations in the full model. In addition, all the variables in the reduced model for the bottom 5 states in terms of death counts are statistically significant.

With this, we checked the if there is a relationship between the top and bottom 5 states by using one-way ANOVA. We first created a new variable for each data frame and label top and bottom based on their category. After this, we run one-way ANOVA for the rank (top and bottom 5 states categories) with the number of death cases. The result showed a p-value that is less than 0.05. This means we will need to reject the null hypothesis that both groups are equal. This means, that

```

# anova
summary(aov(deaths~rank, data = Deaths_rank))
***

              Df Sum Sq Mean Sq F value Pr(>F)
rank           1   3426    3426   201.7 <2e-16 ***
Residuals    14090  239257      17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

p value is less than 0.05, so we reject null hypothesis that both groups are equal.

```


there is no relationship between top and bottom 5 states in terms death cases.

Conclusion

In conclusion, the reduced model is almost always better than the full model even though the statistical measures indicate otherwise. After running the stepwise selection method for the top 5 and bottom 5 states for both confirmed cases and death counts, the reduced models contain statistically significant variables only. This is the main reason why R-squared and adjusted R-squared decrease for the reduced model because observations were removed. Although they are lower for the reduced model, the differences are minimal and can almost be neglected especially after filtering out statistically insignificant variables. In addition to this, the RMSE and MAE of the reduced model (in some cases) is higher than the full model, which confirms our findings. Furthermore, the coefficients from the regression output can be interpreted in numerous ways depending on how one looks at it. There are a lot of moving parts (such as the total population, social distancing measures, etc.) when it comes to the relationships between these variables and thus, they may sometimes contradict each other when talking about the top 5 and bottom 5 states for both confirmed cases and deaths alike.

This confirmed it when we run one-way ANOVA test between top and bottom 5 states, it showed that there is no relationship between 2 groups. Thus, we can say that both groups performed different in terms of distribution of confirmed and death cases.

Attachments

- weekfinal1.xlsx
- Deaths_rank.xlsx
- Confirmed_rank.xlsx
- Week_finalT.xlsx

Appendix A

Removed pre-identified variable using MS Excel Office 365

Parameters	Affected Variables
Components of 'social_distancing_total_grade'	<ul style="list-style-type: none"> * social_distancing_encounters_grade * social_distancing_travel_distance_grade
Components of 'total_college_population'	<ul style="list-style-type: none"> * less_than_high_school_diploma * high_school_diploma_only * some_college_or_higher
Unnecessary variables	<ul style="list-style-type: none"> * latitude * longitude
Combined in 'Age_0_19'	<ul style="list-style-type: none"> * age_0_4 * age_5_9 * age_10_14 * age_15_19
Combined in 'Age_20_59'	<ul style="list-style-type: none"> * age_20_24 * age_25_29 * age_30_34 * age_35_39 * age_40_44 * age_45_49 * age_50_54

	* age_55_59
Combined in 'Age_60+'	* age_60_64 * age_65_69 * age_70_74 * age_75_74 * age_80_84 * age_85_or_higher

Appendix B

Variable Types and Descriptions

Variables	Type	Scale of Measurement	Description
date	Numerical: Discrete	Interval	date formatted as day-month-year
county_fips	Categorical	Nominal	code for each unique county
count_name	Categorical	Nominal	name of county per state
state_fips	Categorical	Nominal	code for each unique state
state_name	Categorical	Nominal	name of each state
covid_19_confirmed_cases	Numerical: Discrete	Ratio	number of daily confirmed COVID cases
covid_19_deaths	Numerical: Discrete	Ratio	number of daily COVID deaths
social_distancing_total_grade	Categorical	Ordinal	average numerical score of unnecessary activities
daily_state_test	Numerical: Continuous	Ratio	number of tests performed daily in each county
precipitation	Numerical: Continuous	Ratio	daily precipitation
temperature	Numerical: Continuous	Interval	daily average temperature

Variables	Type	Scale of Measurement	Description
virus_pressure	Numerical: Continuous	Ratio	measures virus transmission from neighboring counties based on their covid cases
total_population	Numerical: Discrete	Ratio	population of each county
female_percent	Numerical: Continuous	Ratio	total percentage of females over whole population
area	Numerical: Continuous	Ratio	area in square miles per county
population_density	Numerical: Continuous	Ratio	population per square mile per county
hospital_beds_ratio	Numerical: Continuous	Ratio	number of hospital beds over total population
ventilator_capacity_ratio	Numerical: Continuous	Ratio	number of total ventilators divided by total population
icu_beds_ratio	Numerical: Continuous	Ratio	number of ICU beds divided by total population
houses_density	Numerical: Continuous	Ratio	number of housing units per square mile
total_college_population	Numerical: Discrete	Ratio	number of college students over total population
percent_smokers	Numerical: Continuous	Ratio	percentage of adult smokers

Variables	Type	Scale of Measurement	Description
percent_diabetes	Numerical: Continuous	Ratio	percentage of diabetic adults
religious_congregation_ratio	Numerical: Continuous	Ratio	number of active members of active religious congregations over total population
political_party	Categorical	Nominal	political party of the state's governor (0 - Republican, 1 - Democratic)
airport_distance	Numerical: Continuous	Ratio	distance to the nearest international airport (daily passenger load > 10)
passenger_load_ratio	Numerical: Continuous	Ratio	average daily passenger load of nearest international airport over total population
meat_plants	Numerical: Discrete	Ratio	number of meat processing plants
median_household_income	Numerical: Discrete	Ratio	average household income
percent_insured	Numerical: Continuous	Ratio	percentage of health insured residents
deaths_per_100000	Numerical: Continuous	Ratio	deaths per 100,000 residents
gdp_per_capita	Numerical: Continuous	Ratio	gross domestic product per capita
Age_0_19	Numerical: Discrete	Ratio	age group of children and young adults
Age_20_59	Numerical: Discrete	Ratio	age group of adults

Variables	Type	Scale of Measurement	Description
Age_60	Numerical: Discrete	Ratio	age group of elderly
immigrant_student_ratio	Numerical: Continuous	Ratio	total number of students who study in the county but come from another state over total population

Appendix C

Inferential Analysis R code results and graphs

Figure A

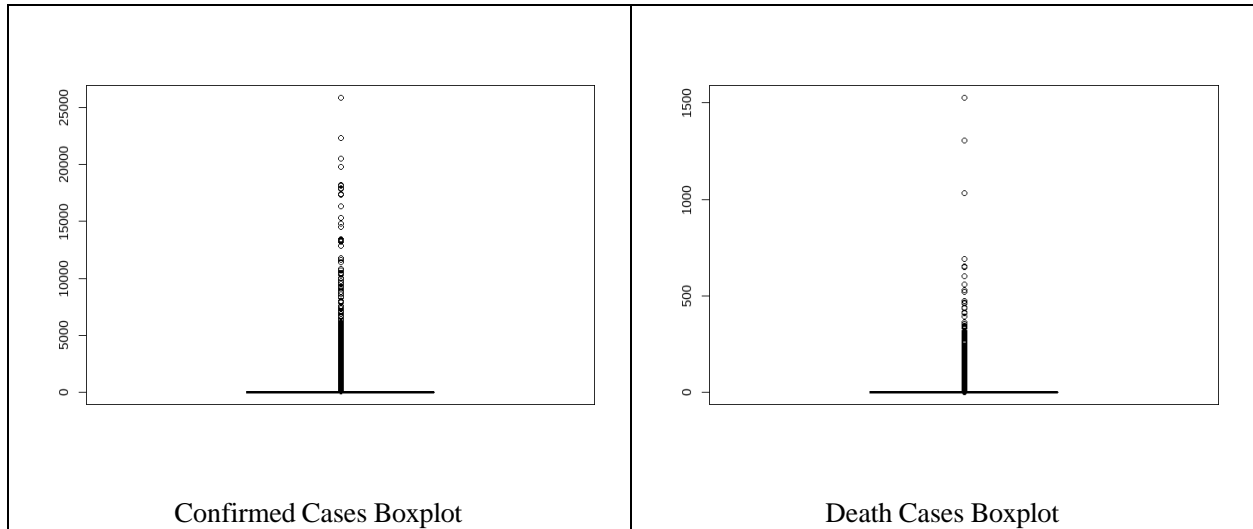
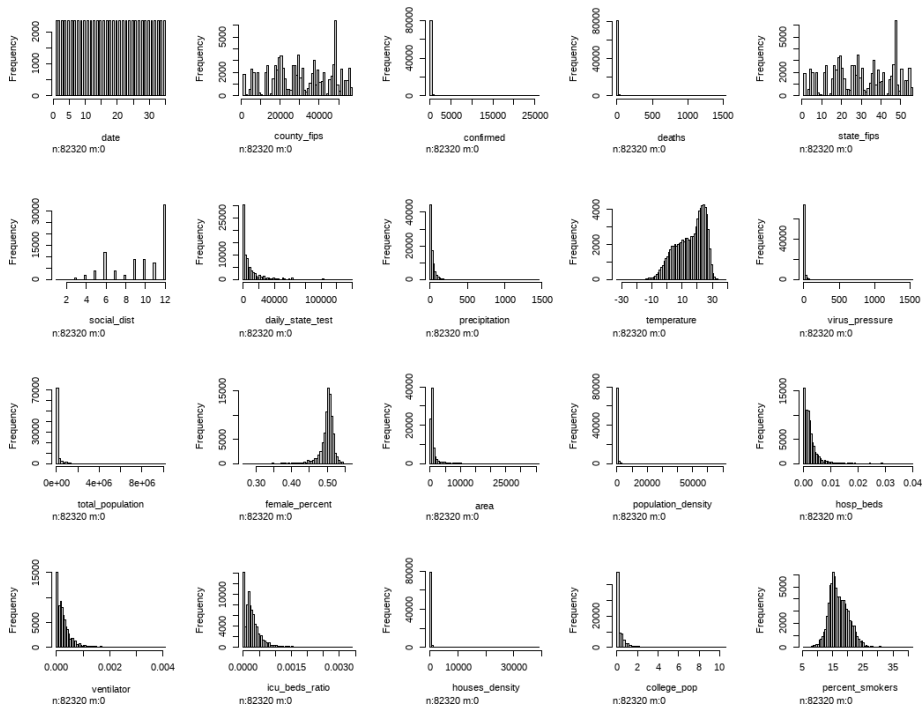


Figure B



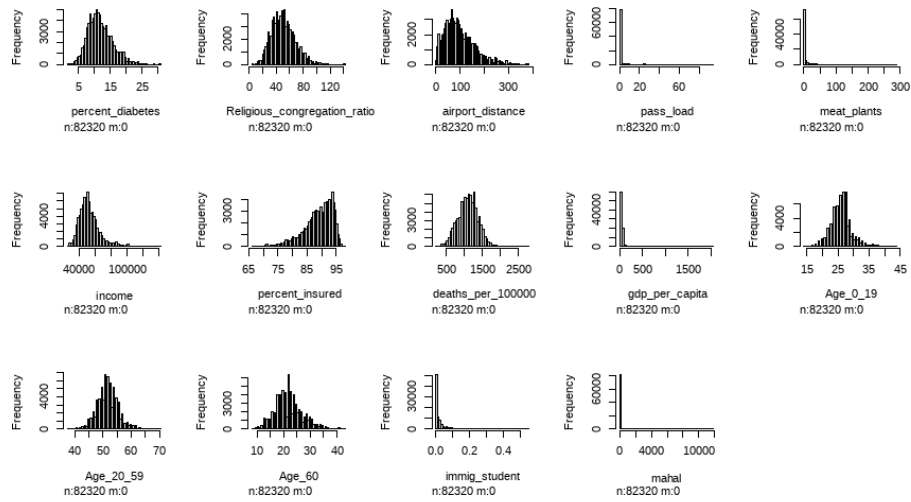


Figure C

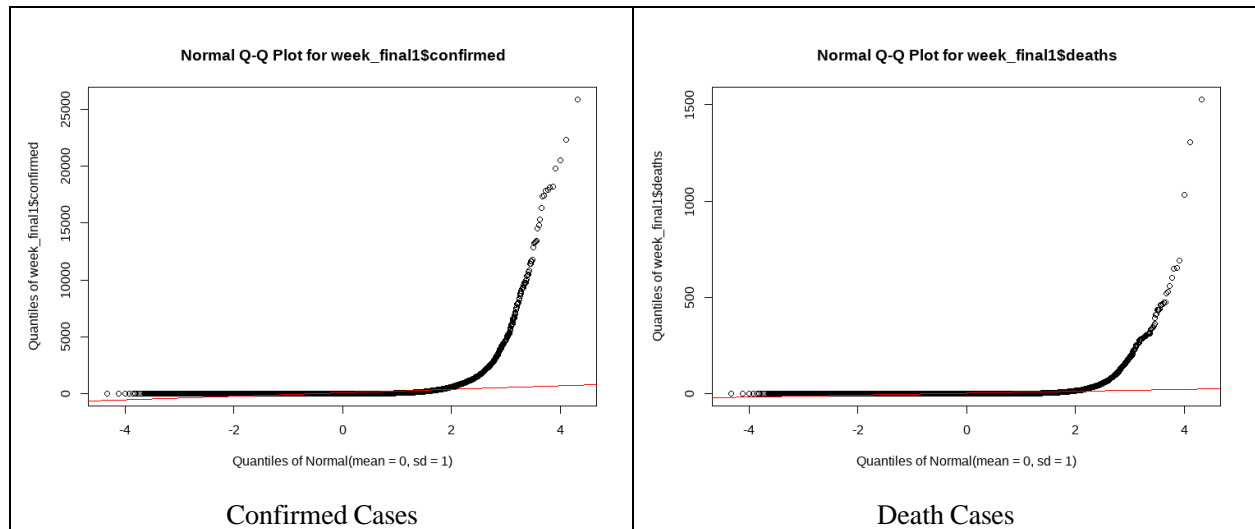


Figure D

Test for normality using Kolmogorov-Smirnov test. Null hypothesis = sample is normal distribution. The result for both showed p-value less than 0.05, therefore we reject the null hypothesis.

<pre>> ks.test(week_final1\$confirmed, "pnorm") warning in ks.test(week_final1\$confirmed, "p default ks.test() cannot compute correct p see help page for one-sample kolmogorov tes One-sample kolmogorov-smirnov test data: week_final1\$confirmed D = 0.54462, p-value < 2.2e-16 alternative hypothesis: two-sided</pre> <p>Confirmed Cases</p>	<pre>> ks.test(week_final1\$deaths, "pnorm") warning in ks.test(week_final1\$deaths, "pnorm default ks.test() cannot compute correct p- see help page for one-sample kolmogorov test One-sample kolmogorov-smirnov test data: week_final1\$deaths D = 0.5, p-value < 2.2e-16 alternative hypothesis: two-sided</pre> <p>Death Cases</p>
--	---

Figure E

```
Call:
lm(formula = confirmed ~ . - deaths - date - county_fips - state_fips -
    political_party - mahal - badmahal, data = week_finalT)

Residuals:
    Min       1Q   Median       3Q      Max
-1449.59  -28.78    -8.46   12.91   950.08

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.342e+01  8.359e+01   0.639  0.522757
social_dist   1.981e+00  1.269e-01  15.610 < 2e-16 ***
daily_state_test 5.587e-04  2.179e-05  25.645 < 2e-16 ***
precipitation  2.928e-02  7.903e-03   3.705  0.000211 ***
temperature   1.732e+00  3.655e-02  47.386 < 2e-16 ***
virus_pressure 1.070e+00  1.122e-02  95.384 < 2e-16 ***
total_population 1.285e-04  1.990e-06  64.589 < 2e-16 ***
female_percent 3.153e+02  1.694e+01  18.610 < 2e-16 ***
area          1.421e-03  1.965e-04   7.230  4.85e-13 ***
population_density 3.326e-02  1.983e-03  16.776 < 2e-16 ***
hosp_beds     -4.681e+02  1.277e+02  -3.664  0.000248 ***
ventilator    -2.442e+04  5.779e+03  -4.225  2.39e-05 ***
icu_beds_ratio  5.614e+04  6.704e+03   8.375 < 2e-16 ***
houses_density -6.182e-02  3.806e-03 -16.242 < 2e-16 ***
college_pop    4.635e+00  1.062e+00   4.362  1.29e-05 ***
percent_smokers  3.983e-01  1.323e-01   3.009  0.002618 **
percent_diabetes -7.741e-01  9.431e-02  -8.208  2.27e-16 ***
Religious_congregation_ratio 2.015e-01  1.999e-02  10.077 < 2e-16 ***
airport_distance -8.271e-02  6.056e-03 -13.657 < 2e-16 ***
pass_load     -2.175e-01  5.688e-02  -3.823  0.000132 ***
meat_plants   -1.028e+00  7.847e-02 -13.103 < 2e-16 ***
income        -3.069e-04  3.792e-05  -8.093  5.92e-16 ***
percent_insured  6.948e-01  7.110e-02   9.772 < 2e-16 ***
deaths_per_100000 -2.937e-02  2.065e-03 -14.225 < 2e-16 ***
gdp_per_capita -1.740e-02  4.600e-03  -3.784  0.000155 ***
Age_0_19      -3.105e+00  8.362e-01  -3.713  0.000205 ***
Age_20_59     -1.991e+00  8.341e-01  -2.387  0.016974 *
Age_60        -3.685e+00  8.294e-01  -4.443  8.88e-06 ***
immig_student -1.324e+02  2.339e+01  -5.660  1.52e-08 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 81.87 on 81020 degrees of freedom
Multiple R-squared:  0.3904,    Adjusted R-squared:  0.3901
F-statistic: 1853 on 28 and 81020 DF, p-value: < 2.2e-16
```

Figure F

```

call:
lm(formula = confirmed ~ . - deaths - date - county_fips - state_fips -
    political_party - mahal - badmahal - houses_density - ventilator -
    Age_0_19 - immig_student, data = week_finalT)

Residuals:
    Min       1Q   Median       3Q      Max
-1460.83  -28.92    -8.61    12.78   949.05

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.714e+02  1.503e+01 -18.060 < 2e-16 ***
social_dist     1.944e+00  1.271e-01  15.298 < 2e-16 ***
daily_state_test 5.647e-04  2.182e-05  25.875 < 2e-16 ***
precipitation   2.862e-02  7.917e-03   3.615 0.000300 ***
temperature     1.721e+00  3.662e-02  47.013 < 2e-16 ***
virus_pressure   1.076e+00  1.123e-02  95.841 < 2e-16 ***
total_population 1.351e-04  1.958e-06  68.990 < 2e-16 ***
female_percent   3.266e+02  1.691e+01  19.321 < 2e-16 ***
area            1.254e-03  1.964e-04   6.388 1.69e-10 ***
population_density 1.253e-03  2.232e-04   5.611 2.02e-08 ***
hosp_beds      -3.046e+02  1.217e+02  -2.502 0.012362 *
icu_beds_ratio   2.853e+04  1.163e+03  24.538 < 2e-16 ***
college_pop     -5.364e-01  4.368e-01  -1.228 0.219414
percent_smokers   4.137e-01  1.325e-01   3.122 0.001797 **
percent_diabetes -7.880e-01  9.428e-02  -8.358 < 2e-16 ***
Religious_congregation_ratio 2.186e-01  1.999e-02  10.938 < 2e-16 ***
airport_distance -8.152e-02  6.056e-03 -13.461 < 2e-16 ***
pass_load       -2.133e-01  5.698e-02  -3.744 0.000181 ***
meat_plants     -9.441e-01  7.839e-02 -12.044 < 2e-16 ***
income          -2.703e-04  3.766e-05  -7.176 7.23e-13 ***
percent_insured   6.676e-01  7.107e-02   9.393 < 2e-16 ***
deaths_per_100000 -2.909e-02  2.060e-03 -14.127 < 2e-16 ***
gdp_per_capita  -1.678e-02  4.561e-03  -3.680 0.000234 ***
Age_20_59        1.291e+00  1.538e-01   8.391 < 2e-16 ***
Age_60          -5.985e-01  1.202e-01  -4.980 6.36e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 82.03 on 81024 degrees of freedom
Multiple R-squared:  0.3879,    Adjusted R-squared:  0.3877
F-statistic: 2140 on 24 and 81024 DF, p-value: < 2.2e-16

```

Figure G

```
Call:
lm(formula = deaths ~ ., data = week_finalT)

Residuals:
    Min       1Q   Median       3Q      Max
-15.9575  -0.4110  -0.1799   0.0620  30.9525

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.311e+00  2.362e+00  -1.825  0.068025 .
confirmed    1.745e-02  9.929e-05 175.760 < 2e-16 ***
social_dist  -3.282e-02  3.591e-03  -9.138 < 2e-16 ***
daily_state_test -4.682e-06  6.183e-07  -7.574  3.67e-14 ***
precipitation  7.722e-04  2.234e-04   3.457  0.000546 ***
temperature   2.116e-03  1.047e-03   2.021  0.043265 *
virus_pressure  4.228e-03  3.343e-04 12.647 < 2e-16 ***
total_population 1.659e-06  5.768e-08 28.762 < 2e-16 ***
female_percent  5.130e+00  4.798e-01 10.691 < 2e-16 ***
area          4.674e-05  5.555e-06   8.415 < 2e-16 ***
population_density 2.047e-04  5.613e-05   3.647  0.000265 ***
hosp_beds     -5.656e+00  3.611e+00  -1.566  0.117256
ventilator    -3.883e+02  1.633e+02  -2.377  0.017439 *
icu_beds_ratio  5.396e+02  1.895e+02   2.847  0.004418 **
houses_density -3.491e-04  1.077e-04  -3.239  0.001198 **
college_pop   -6.583e-03  3.003e-02  -0.219  0.826485
percent_smokers  7.148e-03  3.741e-03   1.911  0.056017 .
percent_diabetes  6.965e-03  2.666e-03   2.612  0.008998 **
Religious_congregation_ratio 1.233e-04  5.654e-04   0.218  0.827325
airport_distance -2.065e-03  1.714e-04 -12.047 < 2e-16 ***
pass_load     1.299e-03  1.608e-03   0.808  0.419248
meat_plants   -1.758e-02  2.220e-03  -7.917  2.46e-15 ***
income        2.133e-06  1.072e-06   1.989  0.046662 *
percent_insured  8.638e-04  2.011e-03   0.430  0.667481
deaths_per_100000 2.711e-04  5.843e-05  4.640  3.49e-06 ***
gdp_per_capita  4.164e-04  1.300e-04   3.203  0.001362 **
Age_0_19      7.865e-03  2.363e-02   0.333  0.739314
Age_20_59     2.001e-02  2.357e-02   0.849  0.395913
Age_60        1.322e-02  2.344e-02   0.564  0.572908
immig_student -2.655e-01  6.611e-01  -0.402  0.687942
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.314 on 81019 degrees of freedom
Multiple R-squared:  0.4511, Adjusted R-squared:  0.4509
F-statistic: 2296 on 29 and 81019 DF, p-value: < 2.2e-16
```

Figure H.1

```
Call:
lm(formula = confirmed ~ . - date - county_fips - state_fips -
    mahal - badmahal, data = confirmedT5)

Residuals:
    Min       1Q   Median       3Q      Max
-645.66  -39.07   -6.63   21.50  793.68

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.414e+02  3.428e+02  -0.704  0.481428
deaths       1.735e+01  2.635e-01  65.840 < 2e-16 ***
social_dist  1.027e+00  5.315e-01   1.931  0.053460 .
daily_state_test 1.163e-03  5.223e-05  22.260 < 2e-16 ***
precipitation 3.629e-02  2.293e-02   1.583  0.113556
temperature -4.305e-01  1.949e-01  -2.209  0.027187 *
virus_pressure 5.738e-01  1.871e-02  30.659 < 2e-16 ***
total_population 3.907e-05  4.886e-06   7.996  1.47e-15 ***
female_percent -9.612e+01  8.030e+01  -1.197  0.231325
area         4.832e-05  9.217e-04   0.052  0.958195
population_density 2.601e-02  3.352e-03   7.759  9.59e-15 ***
hosp_beds    -3.130e+03  1.008e+03  -3.105  0.001909 **
ventilator   -4.286e+03  4.055e+04  -0.106  0.915810
icu_beds_ratio 6.314e+04  4.664e+04   1.354  0.175842
houses_density -4.969e-02  6.336e-03  -7.843  4.97e-15 ***
college_pop   4.046e+01  6.777e+00   5.971  2.46e-09 ***
percent_smokers 4.303e-01  8.473e-01   0.508  0.611567
percent_diabetes 2.193e-01  4.256e-01   0.515  0.606351
Religious_congregation_ratio 3.796e-01  9.935e-02   3.821  0.000134 ***
political_party -7.029e+01  8.295e+00  -8.474 < 2e-16 ***
airport_distance -2.745e-02  3.034e-02  -0.905  0.365534
pass_load     1.088e+00  3.092e+00   0.352  0.724829
meat_plants   -8.861e-01  1.625e-01  -5.453  5.09e-08 ***
income        4.944e-04  1.466e-04   3.372  0.000750 ***
percent_insured 2.038e+00  6.938e-01   2.937  0.003322 **
deaths_per_100000 1.202e-02  1.110e-02   1.083  0.278954
gdp_per_capita 2.796e-01  7.525e-02   3.716  0.000204 ***
Age_0_19     4.936e+00  3.475e+00   1.420  0.155522
Age_20_59    -1.158e+00  3.500e+00  -0.331  0.740722
Age_60       -5.926e-01  3.444e+00  -0.172  0.863392
immig_student -7.544e+02  1.632e+02  -4.623  3.84e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 94.56 on 7922 degrees of freedom
Multiple R-squared:  0.6478,    Adjusted R-squared:  0.6464
F-statistic: 485.7 on 30 and 7922 DF,  p-value: < 2.2e-16
```


Appendix D

Cleaning and Transformation Codes

Library used

```
library(tidyverse) # for data manipulation
library(olsrr) # for residuals plot
library(equationomatic) # to generate lm equation using TeX code
library(ggcorrplot) # correlation plot library(caret) # for
RMSE and MAE library(plotly) # for interactive graph
library(hrbrthemes) # theme components for ggplot2
library(dlookr) # for diagnose outlier library(lattice) #
multiple columns boxplots library(leaps) # for variable
selection
library(trackdown) # collaborate Rmarkdown document through Google Drive

# set working directory
setwd("C:/Users/emili/OneDrive - Langara College/02 DANA 4810 -- Quantitative/Project")
```

Load Dataset

```
# load clean revised dataset
covid_c <- read.csv("C:/Users/emili/OneDrive - Langara College/02 DANA 4810 -
- Quantitative/Project/Covid cleanish1.csv")
```

Investigate Dataset

```
# row and column count
cat("Total number of rows:", nrow(covid_c), "\n") ## Total number
of rows: 562128

cat("Total number of variables/columns:", ncol(covid_c)) ## Total number
of variables/columns: 36
```

```
##      i..date county_fips      county_name state_fips state_name
## 1 27-04-20      56013  Fremont County      56      Wyoming
## 2 25-04-20      56013  Fremont County      56      Wyoming
## 3 05-07-20      56005  Campbell County      56      Wyoming
## 4 16-03-20      56013  Fremont County      56      Wyoming
## 5 26-04-20      56013  Fremont County      56      Wyoming
## 6 10-05-20      56013  Fremont County      56      Wyoming
## covid_19_confirmed_cases covid_19_deaths social_distancing      _total_grade
```

##	1		15	0	C
##	2		10	0	C
##	3		7	0	D+
##	4		7	0	D+
##	5		7	0	C
##	6		7	0	D+
##	daily_state_test precipitation temperature virus_pressure				
##	total_population				
##	1	8	1.0	7.18	0
##	2	101	0.0	4.68	0
##	3	28	0.0	21.30	0
##	4	0	0.0	-1.46	0
##	5	13	4.6	6.78	0
##	6	17	0.0	4.90	0
##	female_percent area population_density hospital_beds_ratio				
##	1	0.4984696	9183.81	4.304423	0.002352584
##	2	0.4984696	9183.81	4.304423	0.002352584
##	3	0.4846337	4802.71	9.607076	0.004659731
##	4	0.4984696	9183.81	4.304423	0.002352584
##	5	0.4984696	9183.81	4.304423	0.002352584
##	6	0.4984696	9183.81	4.304423	0.002352584
##	ventilator_capacity_ratio icu_beds_ratio houses_density				
##	1	0.001340720	0.001113051	1.9625	
##	2	0.001340720	0.001113051	1.9625	
##	3	0.000368444	0.000303424	4.2303	
##	4	0.001340720	0.001113051	1.9625	
##	5	0.001340720	0.001113051	1.9625	
##	6	0.001340720	0.001113051	1.9625	
##	total_college_population percent_smokers percent_diabetes 0.4240217				
##	1		19.39994	9.9	
##	2	0.4240217	19.39994	9.9	
##	3	0.0000000	18.48914	7.5	
##	4	0.4240217	19.39994	9.9	
##	5	0.4240217	19.39994	9.9	
##	6	0.4240217	19.39994	9.9	
##	Religious_congregation_ratio political_party airport_distance				
##	1	33	0	176.8667	
##	2	33	0	176.8667	
##	3	35	0	166.5740	
##	4	33	0	176.8667	
##	5	33	0	176.8667	
##	6	33	0	176.8667	
##	passenger_load_ratio meat_plants median_household_income percent_insured				
##	1	0.000455339	0	51204	79.88766

```
## 2      0.000455339      0      51204      79.88766
## 3      0.000498483      0      78112      87.15840
## 4      0.000455339      0      51204      79.88766
## 5      0.000455339      0      51204      79.88766
## 6      0.000455339      0      51204      79.88766
##  deaths_per_100000  gdp_per_capita  Age_0_19  Age_20_59  Age_60.
## 1      1130.8      39.82      29      52      22
## 2      1130.8      39.82      29      52      22
## 3      587.3      130.06      31      60      9
## 4      1130.8      39.82      29      52      22
## 5      1130.8      39.82      29      52      22
## 6      1130.8      39.82      29      52      22
##  immigrant_student_ratio
## 1      0.01495029
## 2      0.01495029
## 3      0.00000000
## 4      0.01495029
## 5      0.01495029
## 6      0.01495029
```

Cleaning and Transformation

rename date column

```
covid_c <- covid_c %>% rename(date =
  ï..date)
```

format date variable as.date format

```
covid_c$date <- as.Date(covid_c$date, "%d-%m-%y")
```

check data format and column name

```
head(covid_c)
```

```
##      date county_fips county_name state_fips state_name
## 1 2020-04-27    56013  Fremont County      56    Wyoming
## 2 2020-04-25    56013  Fremont County      56    Wyoming
## 3 2020-07-05    56005  Campbell County      56    Wyoming
## 4 2020-03-16    56013  Fremont County      56    Wyoming
## 5 2020-04-26    56013  Fremont County      56    Wyoming
## 6 2020-05-10    56013  Fremont County      56    Wyoming
## covid_19_confirmed_cases covid_19_deaths social_distancing_total_grade
## 1      15      0      C
## 2      10      0      C
## 3      7      0      D+
## 4      7      0      D+
## 5      7      0      C
## 6      7      0      D+
##  .... daily_state_test precipitation temperature virus_pressure
total_population
## 1      8      1.0      7.18      0
39531
```

## 2	101	0.0	4.68	0	
39531					
## 3	28	0.0	21.30	0	
46140					
## 4	0	0.0	-1.46	0	
39531					
## 5	13	4.6	6.78	0	
39531					
## 6	17	0.0	4.90	0	
39531					
##	female_percent	area	population_density	hospital_beds_ratio	
## 1	0.4984696	9183.81	4.304423	0.002352584	
## 2	0.4984696	9183.81	4.304423	0.002352584	
## 3	0.4846337	4802.71	9.607076	0.004659731	
## 4	0.4984696	9183.81	4.304423	0.002352584	
## 5	0.4984696	9183.81	4.304423	0.002352584	
## 6	0.4984696	9183.81	4.304423	0.002352584	
##	ventilator_capacity_ratio	icu_beds_ratio	houses_density		
## 1	0.001340720	0.001113051	1.9625		
## 2	0.001340720	0.001113051	1.9625		
## 3	0.000368444	0.000303424	4.2303		
## 4	0.001340720	0.001113051	1.9625		
## 5	0.001340720	0.001113051	1.9625		
## 6	0.001340720	0.001113051	1.9625		
##	total_college_population	percent_smokers	percent_diabetes	0.4240217	
## 1		19.39994	9.9		
## 2	0.4240217	19.39994	9.9		
## 3	0.0000000	18.48914	7.5		
## 4	0.4240217	19.39994	9.9		
## 5	0.4240217	19.39994	9.9		
## 6	0.4240217	19.39994	9.9		
##	Religious_congregation_ratio	political_party	airport_distance		
## 1		33	0	176.8667	
## 2		33	0	176.8667	
## 3		35	0	166.5740	
## 4		33	0	176.8667	
## 5		33	0	176.8667	
## 6		33	0	176.8667	
##	passenger_load_ratio	meat_plants	median_household_income	percent_insured	
## 1	0.000455339	0	51204	79.88766	
## 2	0.000455339	0	51204	79.88766	
## 3	0.000498483	0	78112	87.15840	
## 4	0.000455339	0	51204	79.88766	
## 5	0.000455339	0	51204	79.88766	
## 6	0.000455339	0	51204	79.88766	
##	deaths_per_100000	gdp_per_capita	Age_0_19	Age_20_59	Age_60.
## 1	1130.8	39.82	29	52	22
## 2	1130.8	39.82	29	52	22
## 3	587.3	130.06	31	60	9
## 4	1130.8	39.82	29	52	22

```
## 5      1130.8      39.82      29      52      22
## 6      1130.8      39.82      29      52      22
##      immigrant_student_ratio
## 1      0.01495029
## 2      0.01495029
## 3      0.00000000
## 4      0.01495029
## 5      0.01495029
## 6      0.01495029

# group data per week per county
library(dplyr) covid_c <-
covid_c %>%
  mutate(week = cut.Date(date, breaks = "1 week", labels = FALSE)) %>% arrange(date)

unique(covid_c$week) # get unique numbers in week

## [1] 1 2 3 4 5      6 7 8 9 10 11 12 13 14      15 16 17 18 19 20 21 22 23
24 25
## [26] 26 27 28 29 30      31 32 33 34 35

# change categorical variable to numeric equivalent
unique(covid_c$social_distancing_total_grade) # check unique variables

## [1] "F" "D" "C" "D+" "D-" "C-" "B-" "B" "A-" "C+" "A" "B+"

covid_c1 <- covid_c %>% mutate(social_distancing_total_grade =
as.numeric(as.factor(covid_c$social_distancing_total_grade)))
unique(covid_c1$social_distancing_total_grade) # check unique variables as numeric

## [1] 12 9 6 11 10 7 4 3 2 8 1 5
```

No A+ grading was recorded.

```
# drop columns in dataset: date, county_name, state_name
covid_c1 <- subset(covid_c1, select = -c(date, county_name, state_name))

# Group the observations together by week by county based on mean but
separate covid_19_confirmed_cases and covid_19_deaths from the res
# dataframe 1
week_meancounty_cd <- subset(covid_c1, select = c(week, county_fips, covid_19_confirmed_cases,
covid_19_deaths))
week_meancounty_cd <- aggregate(~ week + county_fips, data = week_meancounty_cd, sum)
# dataframe 2
week_meancounty <- subset(covid_c1, select = -c(covid_19_confirmed_cases, covid_19_deaths))
week_meancounty <- aggregate(~ week + county_fips, data = week_meancounty, mean)
```

Check if sum of original dataframe and aggregated are the same

```
sum(week_meancounty_cd$covid_19_confirmed_cases) ## [1]
```

5866205

```
sum(covid_c1$covid_19_confirmed_cases) ## [1]
```

5866205

Merge 2 dataframe 1 and 2 by row names

```
week_final <- merge(week_meancounty_cd, week_meancounty, by = 0)
```

drop duplicate variable names and rename final dataset for analysis week_final

```
<- subset(week_final, select = -c(Row.names, week.y, county_fips.y))
```

Total observations are 82,320

round to whole digit social_distancing_total_grade so you can covert later easily to categorical

```
week_final$social_distancing_total_grade <-  
round(week_final$social_distancing_total_grade, digits = 0)
```

NEXT TO DO: fix proportion of age group to equal to 100%

fix age distribution per age group by getting the ratio of the total of total percent

```
week_final <- week_final %>%  
  mutate(Age_0_19 = Age_0_19/(Age_0_19 + Age_20_59 + Age_60.)*100) %>% mutate(Age_20_59 =  
    Age_20_59/(Age_0_19 + Age_20_59 + Age_60.)*100) %>% mutate(Age_60. = Age_60./(Age_0_19 +  
    Age_20_59 + Age_60.)*100)
```

rename variables by removing '.x' and '.'

```
week_final <- week_final %>% rename(date =  
  week.x) %>%  
  rename(county_fips = county_fips.x) %>% rename(Age_60 =  
    Age_60.)
```

remove splitted dataframe

Appendix E

Descriptive Analysis Codes

Library used

```
library(tidyverse) # for data manipulation
library(olsrr) # for residuals plot
library(equatiomatic) # to generate lm equation using TeX code
library(ggcorrplot) # correlation plot library(caret) # for
RMSE and MAE library(plotly) # for interactive graph
library(hrbrthemes) # theme components for ggplot2
library(dlookr) # for diagnose outlier library(lattice) #
multiple columns boxplots library(leaps) # for variable
selection
library(trackdown) # collaborate Rmarkdown document through Google Drive

# set working directory
setwd("C:/Users/emili/OneDrive - Langara College/02 DANA 4810 -- Quantitativ e/Project")
```

Load Aggregated Dataset from Github

```
week_final1 <- read.csv("https://raw.githubusercontent.com/emiliosagre/COVID19-
US/main/week_final1.csv")
```

Investigate Dataset

```
# row and column count
cat("Total number of rows:", nrow(week_final1), "\n") ## Total number
of rows: 82320

cat("Total number of variables/columns:", ncol(week_final1))
## Total number of variables/columns: 24
```

Check for Outliers

```
summary(week_final1)
```

##	date	county_fips	covid_19_confirmed_cases	Min.	covid_19_deaths
##	Min. : 1	Min. : 1003	:	0.00	Min. : 0.00
##	1st Qu.: 9	1st Qu.: 19041	1st Qu.:	0.00	1st Qu.: 0.00
##	Median : 18	Median : 29162	Median :	3.00	Median : 0.00
##	Mean : 18	Mean : 30309	Mean :	71.26	Mean : 2.03
##	3rd Qu.: 27	3rd Qu.: 45046	3rd Qu.:	26.00	3rd Qu.: 0.00
##	Max. : 35	Max. : 56039	Max. :	25882.00	Max. : 1528.00
##	state_fips	social_distancing_total_grade	daily_state_test		

##	Min. : 1.00	Min. : 1.000	Min. : 0.0	
##	1st Qu.:19.00	1st Qu.: 7.000	1st Qu.: 332.6	
##	Median :29.00	Median :10.000	Median : 4132.7	
##	Mean :30.22	Mean : 9.552	Mean : 9986.6	
##	3rd Qu.:45.00	3rd Qu.:12.000	3rd Qu.: 11659.6	
##	Max. :56.00	Max. :12.000	Max. :138859.4	
##	precipitation	temperature	virus_pressure	total_population
##	Min. : 0.000	Min. : -32.314	Min. : 0.0000	Min. : 1227
##	1st Qu.: 4.286	1st Qu.: 8.174	1st Qu.: 0.0357	1st Qu.: 13119
##	Median : 17.200	Median : 17.343	Median : 1.3333	Median : 32398
##	Mean : 28.991	Mean : 15.446	Mean : 9.9183	Mean : 124755
##	3rd Qu.: 39.629	3rd Qu.: 23.449	3rd Qu.: 6.6735	3rd Qu.: 87010
##	Max. :1446.000	Max. : 38.074	Max. :1535.5000	Max. :10105518
##	female_percent	area	population_density	hospital_beds_ratio
##	Min. :0.2684	Min. : 2.5	Min. : 0.22	Min. :0.0000000
##	1st Qu.:0.4946	1st Qu.: 471.7	1st Qu.: 18.39	1st Qu.:0.0008365
##	Median :0.5031	Median : 651.7	Median : 48.13	Median :0.0017630
##	Mean :0.4994	Mean : 1086.4	Mean : 257.72	Mean :0.0025211
##	3rd Qu.:0.5100	3rd Qu.: 974.7	3rd Qu.: 136.96	3rd Qu.:0.0030786
##	Max. :0.5687	Max. :35572.6	Max. :71340.39	Max. :0.0399348
##	ventilator_capacity_ratio	icu_beds_ratio	houses_density	
##	Min. :0.0000000	Min. :0.0000000	Min. : 0.08	
##	1st Qu.:0.0001048	1st Qu.:0.0001188	1st Qu.: 9.28	
##	Median :0.0002189	Median :0.0002285	Median : 23.22	
##	Mean :0.0003105	Mean :0.0003046	Mean : 113.82	
##	3rd Qu.:0.0003991	3rd Qu.:0.0003918	3rd Qu.: 60.83	
##	Max. :0.0040732	Max. :0.0033943	Max. :38819.49	
##	total_college_population	percent_smokers	percent_diabetes	
##	Min. : 0.000000	Min. : 5.909	Min. : 1.80	
##	1st Qu.: 0.000000	1st Qu.:14.801	1st Qu.: 9.10	
##	Median : 0.005462	Median :16.673	Median :11.40	
##	Mean : 0.387971	Mean :17.169	Mean :11.87	
##	3rd Qu.: 0.506748	3rd Qu.:19.341	3rd Qu.:14.10	
##	Max. :10.586403	Max. :41.491	Max. :31.00	
##	Religious_congregation_ratio	political_party	airport_distance	
##	Min. : 5.00	Min. :0.0000	Min. : 2.675	
##	1st Qu.: 39.00	1st Qu.:0.0000	1st Qu.: 53.906	
##	Median : 50.00	Median :0.0000	Median : 87.143	
##	Mean : 51.14	Mean :0.4575	Mean : 98.660	
##	3rd Qu.: 62.00	3rd Qu.:1.0000	3rd Qu.:133.886	
##	Max. :141.00	Max. :1.0000	Max. :383.144	
##	passenger_load_ratio	sured	meat_plants	median_household_income
##	Min. : 0.00002	Min. : 0.000	Min. : 26278	percent_in
.25				
##	1st Qu.: 0.00157	1st Qu.: 0.000	1st Qu.: 44565	1st Qu.:86
.24				
##	Median : 0.00610	Median : 1.000	Median : 51121	Median :89
.78				
##	Mean : 0.80283	Mean : 2.963	Mean : 53410	Mean 88


```
.94
## 3rd Qu.: 0.04684      3rd Qu.: 3.000      3rd Qu.: 59243      3rd Qu.:92
.83
## Max. :93.58695      Max. :288.000      Max. :140382      Max. :97
.74
## deaths_per_100000 gdp_per_capita      Age_0_19      Age_20_59
## Min. : 235.4      Min. : 10.61      Min. :14.14      Min. :37.70
## 1st Qu.: 919.1      1st Qu.: 29.31      1st Qu.:24.00      1st Qu.:49.61
## Median :1109.2      Median : 39.12      Median :26.00      Median :51.63
## Mean :1103.0      Mean : 47.70      Mean :26.20      Mean :51.78
## 3rd Qu.:1287.8      3rd Qu.: 52.15      3rd Qu.:28.00      3rd Qu.:54.00
## Max. :2790.7      Max. :2027.95      Max. :43.56      Max. :70.41
## Age_60      immigrant_student_ratio
## Min. : 8.00      Min. :0.0000000
## 1st Qu.:18.88      1st Qu.:0.0000000
## Median :21.92      Median :0.0002023
## Mean :22.02      Mean :0.0157857
## 3rd Qu.:25.00      3rd Qu.:0.0201942
## Max. :43.00      Max. :0.5400094

diagnose(week_final1)

## # A tibble: 34 x 6
## variables      types missing_count missing_percent unique_count unique_rate
## <chr>      <chr>      <int>      <dbl>      <int>
<dbl>
## 1 date      inte~      0      0      35      0.
000425
## 2 county_fips inte~      0      0      2352      0.
0286
## 3 covid_19_confir~ inte~      0      0      1737      0.
0211
## 4 covid_19_deaths inte~      0      0      261      0.
00317
## 5 state_fips inte~      0      0      50      0.
000607
## 6 social_distanci~ inte~      0      0      12      0.
000146
## 7 daily_state_test nume~      0      0      1515      0.
0184
## 8 precipitation nume~      0      0      5002      0.
0608
## 9 temperature nume~      0      0      13737      0.
167
## 10 virus_pressure nume~      0      0      16790      0.
204
## # ... with 24 more rows
```

No missing values in the data.

```
# library dlookr
# convert to dataframe
covid_outlier <- as.data.frame(diagnose_outlier(week_final1))

# get difference between outlier with mean and without mean
covid_outlier$difference <- (covid_outlier$with_mean - covid_outlier$without_mean)

# sort difference in descending order
covid_outlier %>% arrange(desc(covid_outlier$difference))
```

##		variables	outliers_cnt	outliers_ratio	outliers_mean
## 1		virus_pressure	10155	12.33600583	6.152980e+01
## 2		ventilator_capacity_ratio	5250	6.37755102	1.346640e-03
## 3		total_population	10745	13.05272109	6.792133e+05
## 4		total_college_population	6160	7.48299320	2.440349e+00
## 5		temperature	28	0.03401361	-1.918184e+01
## 6		state_fips	0	0.00000000	NaN
## 7		social_distancing_total_grade	0	0.00000000	NaN
## 8		Religious_congregation_ratio	700	0.85034014	1.063500e+02
## 9		precipitation	4807	5.83940719	1.396803e+02
## 10		population_density	11095	13.47789116	1.508432e+03
## 11		political_party	0	0.00000000	NaN
## 12		percent_smokers	735	0.89285714	2.881225e+01
## 13		percent_insured	1575	1.91326531	7.339556e+01
## 14		percent_diabetes	1085	1.31802721	2.376129e+01
## 15		passenger_load_ratio	16380	19.89795918	3.988672e+00
## 16		median_household_income	3780	4.59183673	9.459195e+04
## 17		meat_plants	6895	8.37585034	1.920305e+01
## 18		immigrant_student_ratio	7035	8.54591837	9.519714e-02
## 19		icu_beds_ratio	5320	6.46258503	1.229842e-03
## 20		houses_density	10850	13.18027211	6.701818e+02
## 21		hospital_beds_ratio	5495	6.67517007	1.129817e-02
## 22		gdp_per_capita	3885	4.71938776	1.956587e+02
## 23		female_percent	6335	7.69557823	4.544444e-01
## 24		deaths_per_100000	700	0.85034014	1.474400e+03
## 25		date	0	0.00000000	NaN
## 26		daily_state_test	7084	8.60544218	5.346615e+04
## 27		covid_19_deaths	18769	22.80004859	8.904683e+00
## 28		covid_19_confirmed_cases	12218	14.84207969	4.314966e+02
## 29		county_fips	0	0.00000000	NaN
## 30		area	10185	12.37244898	4.072637e+03
## 31		airport_distance	1960	2.38095238	2.872448e+02
## 32		Age_60	1365	1.65816327	3.508528e+01
## 33		Age_20_59	2205	2.67857143	5.561758e+01
## 34		Age_0_19	2695	3.27380952	3.148534e+01
##		with_mean without_mean difference			
## 1		9.918332e+00 2.655607e+00 7.262724e+00			
## 2		3.105403e-04 2.399614e-04 7.057897e-05			
## 3		1.247548e+05 4.151818e+04 8.323657e+04			

```
## 4 3.879712e-01 2.219701e-01 1.660011e-01
## 5 1.544597e+01 1.545775e+01 -1.178217e-02
## 6 3.021556e+01 3.021556e+01 0.000000e+00
## 7 9.552259e+00 9.552259e+00 0.000000e+00
## 8 5.114201e+01 5.066852e+01 4.734819e-01
## 9 2.899104e+01 2.212660e+01 6.864439e+00
## 10 2.577202e+02 6.289195e+01 1.948283e+02
## 11 4.574830e-01 4.574830e-01 0.000000e+00
## 12 1.716937e+01 1.706448e+01 1.048909e-01
## 13 8.894042e+01 8.924364e+01 -3.032159e-01
## 14 1.186688e+01 1.170801e+01 1.588655e-01
## 15 8.028250e-01 1.143621e-02 7.913888e-01
## 16 5.340991e+04 5.142788e+04 1.982024e+03
## 17 2.963010e+00 1.478422e+00 1.484588e+00
## 18 1.578568e-02 8.365082e-03 7.420596e-03
## 19 3.045541e-04 2.406252e-04 6.392895e-05
## 20 1.138233e+02 2.936149e+01 8.446186e+01
## 21 2.521076e-03 1.893284e-03 6.277923e-04
## 22 4.769686e+01 4.036809e+01 7.328768e+00
## 23 4.993922e-01 5.031396e-01 -3.747380e-03
## 24 1.103045e+03 1.099860e+03 3.184861e+00
## 25 1.800000e+01 1.800000e+01 0.000000e+00
## 26 9.986594e+03 5.892687e+03 4.093907e+03
## 27 2.030272e+00 0.000000e+00 2.030272e+00
## 28 7.126099e+01 8.475921e+00 6.278507e+01
## 29 3.030864e+04 3.030864e+04 0.000000e+00
## 30 1.086409e+03 6.647723e+02 4.216364e+02
## 31 9.865959e+01 9.405995e+01 4.599638e+00
## 32 2.202343e+01 2.180319e+01 2.202388e-01
## 33 5.178345e+01 5.167792e+01 1.055265e-01
## 34 2.620292e+01 2.602413e+01 1.787897e-01
```

covid_outlier

##	variables	outliers_cnt	outliers_ratio	outliers_mean
## 1	date	0	0.00000000	NaN
## 2	county_fips	0	0.00000000	NaN
## 3	covid_19_confirmed_cases	12218	14.84207969	4.314966e+02
## 4	covid_19_deaths	18769	22.80004859	8.904683e+00
## 5	state_fips	0	0.00000000	NaN
## 6	social_distancing_total_grade	0	0.00000000	NaN
## 7	daily_state_test	7084	8.60544218	5.346615e+04
## 8	precipitation	4807	5.83940719	1.396803e+02
## 9	temperature	28	0.03401361	-1.918184e+01
## 10	virus_pressure	10155	12.33600583	6.152980e+01
## 11	total_population	10745	13.05272109	6.792133e+05
## 12	female_percent	6335	7.69557823	4.544444e-01
## 13	area	10185	12.37244898	4.072637e+03
## 14	population_density	11095	13.47789116	1.508432e+03
## 15	hospital_beds_ratio	5495	6.67517007	1.129817e-02

##	16	ventilator_capacity_ratio	5250	6.37755102	1.346640e-03
##	17	icu_beds_ratio	5320	6.46258503	1.229842e-03
##	18	houses_density	10850	13.18027211	6.701818e+02
##	19	total_college_population	6160	7.48299320	2.440349e+00
##	20	percent_smokers	735	0.89285714	2.881225e+01
##	21	percent_diabetes	1085	1.31802721	2.376129e+01
##	22	Religious_congregation_ratio	700	0.85034014	1.063500e+02
##	23	political_party	0	0.00000000	NaN
##	24	airport_distance	1960	2.38095238	2.872448e+02
##	25	passenger_load_ratio	16380	19.89795918	3.988672e+00
##	26	meat_plants	6895	8.37585034	1.920305e+01
##	27	median_household_income	3780	4.59183673	9.459195e+04
##	28	percent_insured	1575	1.91326531	7.339556e+01
##	29	deaths_per_100000	700	0.85034014	1.474400e+03
##	30	gdp_per_capita	3885	4.71938776	1.956587e+02
##	31	Age_0_19	2695	3.27380952	3.148534e+01
##	32	Age_20_59	2205	2.67857143	5.561758e+01
##	33	Age_60	1365	1.65816327	3.508528e+01
##	34	immigrant_student_ratio	7035	8.54591837	9.519714e-02
##		with_mean without_mean difference			
##	1	1.800000e+01 1.800000e+01 0.000000e+00			
##	2	3.030864e+04 3.030864e+04 0.000000e+00			
##	3	7.126099e+01 8.475921e+00 6.278507e+01			
##	4	2.030272e+00 0.000000e+00 2.030272e+00			
##	5	3.021556e+01 3.021556e+01 0.000000e+00			
##	6	9.552259e+00 9.552259e+00 0.000000e+00			
##	7	9.986594e+03 5.892687e+03 4.093907e+03			
##	8	2.899104e+01 2.212660e+01 6.864439e+00			
##	9	1.544597e+01 1.545775e+01 -1.178217e-02			
##	10	9.918332e+00 2.655607e+00 7.262724e+00			
##	11	1.247548e+05 4.151818e+04 8.323657e+04			
##	12	4.993922e-01 5.031396e-01 -3.747380e-03			
##	13	1.086409e+03 6.647723e+02 4.216364e+02			
##	14	2.577202e+02 6.289195e+01 1.948283e+02			
##	15	2.521076e-03 1.893284e-03 6.277923e-04			
##	16	3.105403e-04 2.399614e-04 7.057897e-05			
##	17	3.045541e-04 2.406252e-04 6.392895e-05			
##	18	1.138233e+02 2.936149e+01 8.446186e+01			
##	19	3.879712e-01 2.219701e-01 1.660011e-01			
##	20	1.716937e+01 1.706448e+01 1.048909e-01			
##	21	1.186688e+01 1.170801e+01 1.588655e-01			
##	22	5.114201e+01 5.066852e+01 4.734819e-01			
##	23	4.574830e-01 4.574830e-01 0.000000e+00			
##	24	9.865959e+01 9.405995e+01 4.599638e+00			
##	25	8.028250e-01 1.143621e-02 7.913888e-01			
##	26	2.963010e+00 1.478422e+00 1.484588e+00			
##	27	5.340991e+04 5.142788e+04 1.982024e+03			
##	28	8.894042e+01 8.924364e+01 -3.032159e-01			
##	29	1.103045e+03 1.099860e+03 3.184861e+00			
##	30	4.769686e+01 4.036809e+01 7.328768e+00			

```
## 31 2.620292e+01 2.602413e+01 1.787897e-01 ## 32
5.178345e+01 5.167792e+01 1.055265e-01 ## 33
2.202343e+01 2.180319e+01 2.202388e-01 ## 34
1 578568e-02 8 365082e-03 7 420596e-03
```

Data showed a lot of outliers we will not remove any of them.

Correlation for covid_19_confirmed_cases and covid_19_deaths with other variables

A. Confirmed Cases

covid_19_confirmed_cases

```
cor_confirmed <- as.data.frame(cor(week_final1[, colnames(week_final1) != "covid_19_confirmed_cases"], # Calculate correlations
                                week_final1$covid_19_confirmed_cases))
```

sort difference in descending order

```
cor_confirmed <- cor_confirmed %>% arrange(desc(cor_confirmed,v1))
```

```
head(cor_confirmed, 10) # top 10 most correlated
```

```
##                                V1
## covid_19_deaths                0.6411414
## total_population                0.6054531
## meat_plants                    0.5486385
## virus_pressure                  0.4741532
## daily_state_test                0.2148120
## population_density              0.1706903
## houses_density                  0.1421518
## Age_20_59                      0.1317190
## median_household_income         0.1260323
## temperature                     0.1178476
```

```
tail(cor_confirmed, 10)
```

```
##                                V1
## hospital_beds_ratio             -0.01105132
## percent_insured                 -0.01234432
## passenger_load_ratio            -0.01609232
## state_fips                      -0.04863404
## county_fips                     -0.04869855
## percent_diabetes                 -0.07469584
## percent_smokers                  -0.09402702
## Age_60                         -0.13482410
## airport_distance                -0.14941314
## deaths_per_100000               -0.15819836
```

B. Deaths

covid_19_confirmed_cases

```
cor_deaths <- as.data.frame(cor(week_final1[, colnames(week_final1) != "covid_19_confirmed_cases"], # Calculate correlations
                                week_final1$covid_19_confirmed_cases))
```

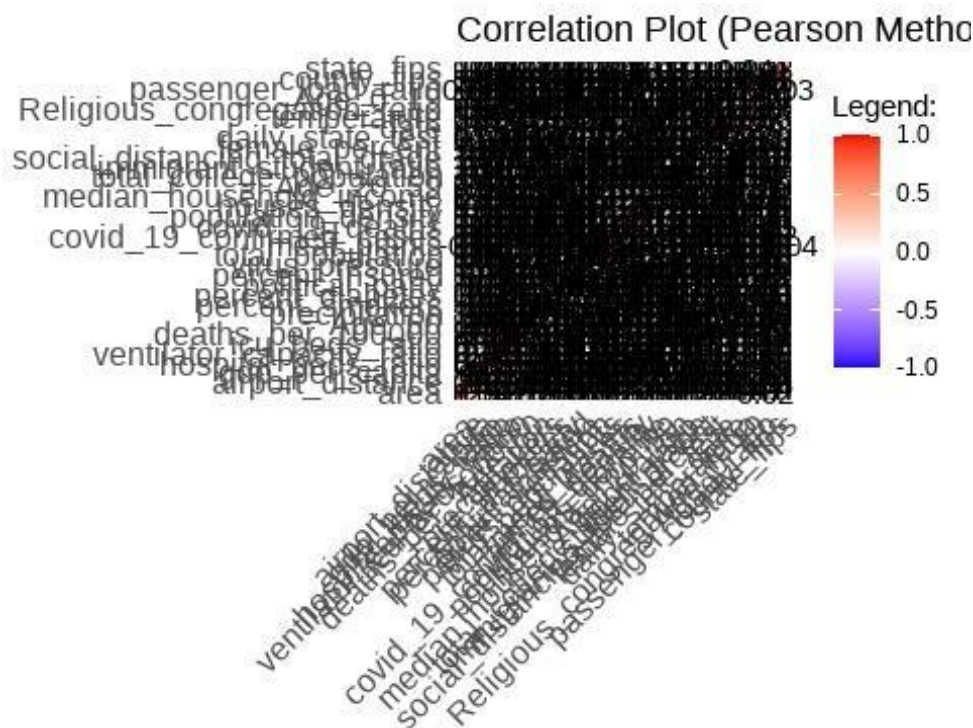
```
d_19_deaths"], # Calculate correlations
      week_final1$covid_19_deaths))

# sort difference in descending order
cor_deaths <- cor_deaths %>% arrange(desc(cor_deaths,v1))

head(cor_confirmed, 10) # top 10 most correlated

##
## covid_19_deaths ##
## total_population ##
## meat_plants
## virus_pressure ##
## population_density
## houses_density
## Age_20_59
## median_household_income
## temperature

# correlation plot of all variables using pearson method with values
ggcorrplot(cor(week_final1[,unlist(lapply(week_final1,is.numeric))], method = "pearson"), hc.order =
TRUE, insig = "blank", lab = TRUE,
  title = "Correlation Plot (Pearson Method)", legend.title = "Legend:")
```



Appendix F

Inferential Analysis Codes

set working directory

```
setwd("C:/Users/emili/OneDrive - Langara College/02 DANA 4810 -- Quantitative/Project/FINAL")
```

Library used *

```
library(tidyverse) # for data manipulation library(olsrr) #  
for residuals plot library(dplyr)  
library(equatiomatic) # to generate lm equation using TeX code  
library(ggcorrplot) # correlation plot library(caret) # for  
RMSE and MAE library(plotly) # for interactive graph  
library(hrbrthemes) # theme components for ggplot2  
library(dlookr) # for diagnose outlier library(lattice) #  
multiple columns boxplots library(leaps) # for variable  
selection  
library(trackdown) # collaborate Rmarkdown document through Google Drive  
library(corrplot)
```

Load Aggregated Dataset

```
week_final1 <- read.csv("https://raw.githubusercontent.com/emiliosagre/COVID19-  
US/main/week_final1.csv")
```

Dataset shape

```
# row and column count  
cat("Total number of rows:", nrow(week_final1), "\n") ## Total number  
of rows: 82320  
  
cat("Total number of variables/columns:", ncol(week_final1))  
## Total number of variables/columns: 24
```

General Analysis

```
# rename date column  
week_final1 <- week_final1 %>%  
  rename(confirmed = covid_19_confirmed_cases) %>% rename(deaths  
= covid_19_deaths) %>%  
  rename(social_dist = social_distancing_total_grade1) %>%
```

```
rename(hosp_beds = hospital_beds_ratio) %>%  
rename(college_pop = total_college_population) %>%  
rename(ventilator = ventilator_capacity_ratio) %>%  
rename(pass_load = passenger_load_ratio) %>% rename(income =  
median_household_income) %>% rename(immig_student =
```

Check summary of statistics

```
summary(week_final1$confirmed)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	0.00	0.00	3.00	71.26	26.00	25882.00

```
summary(week_final1$deaths)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	##	0.00
	0.00	0.00	2.03	0.00	1528.00			

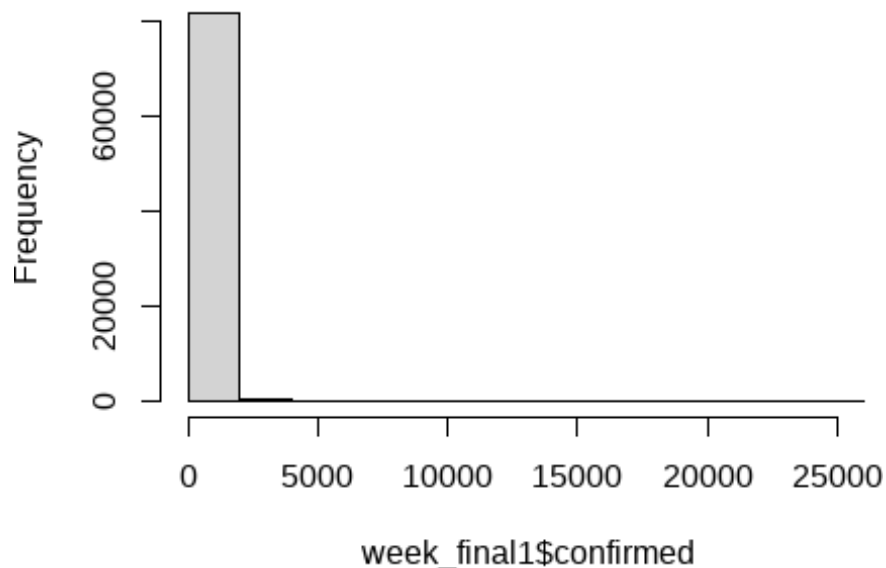
values are within limit but median is near Q1 and max value is very large and very far from the mean for both

check histogram

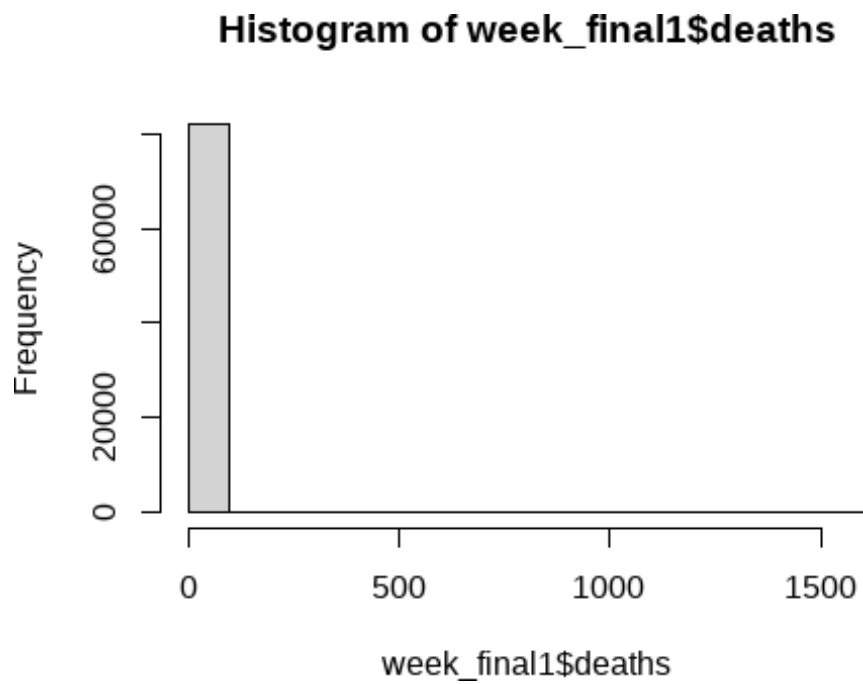
```
# test for normality using histogram
```

```
hist(week_final1$confirmed, bins = 30)
```

Histogram of week_final1\$confirmed

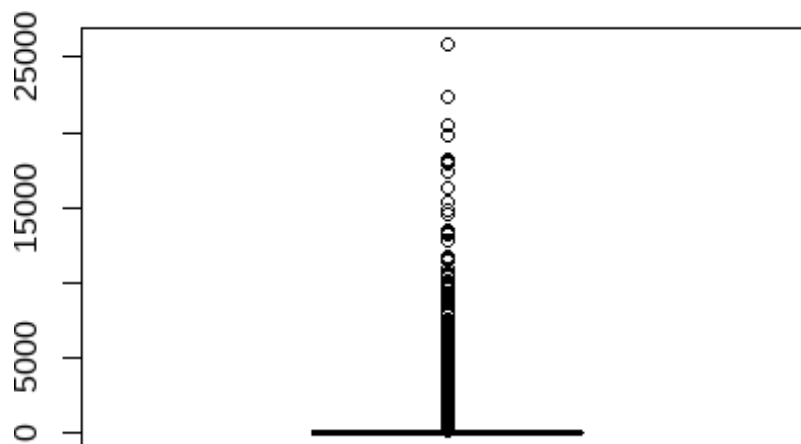


```
hist(week_final1$deaths, bins = 30)
```

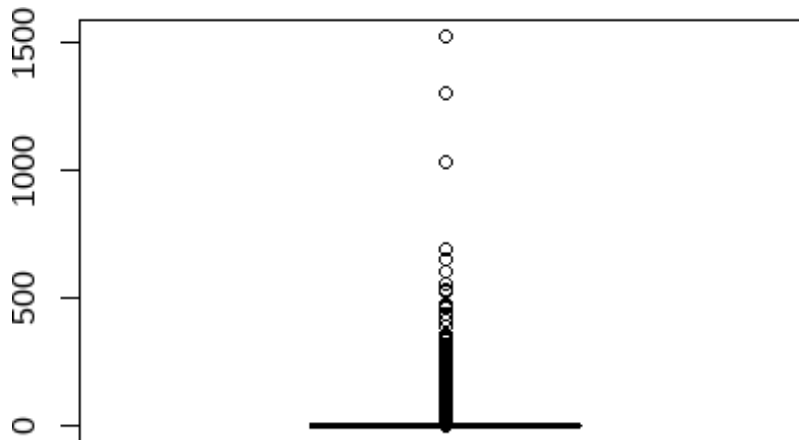



The data is heavily skewed to the right

```
boxplot(week_final1$confirmed)
```



```
boxplot(week_final1$deaths)
```



the lower value

> all points are near

```
# test for normality using QQ plot
```

```
library(EnvStats)
```

```
## Warning: package 'EnvStats' was built under R version 4.1.3 ##
```

```
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:dlookr': ##
```

```
##      kurtosis, skewness
```

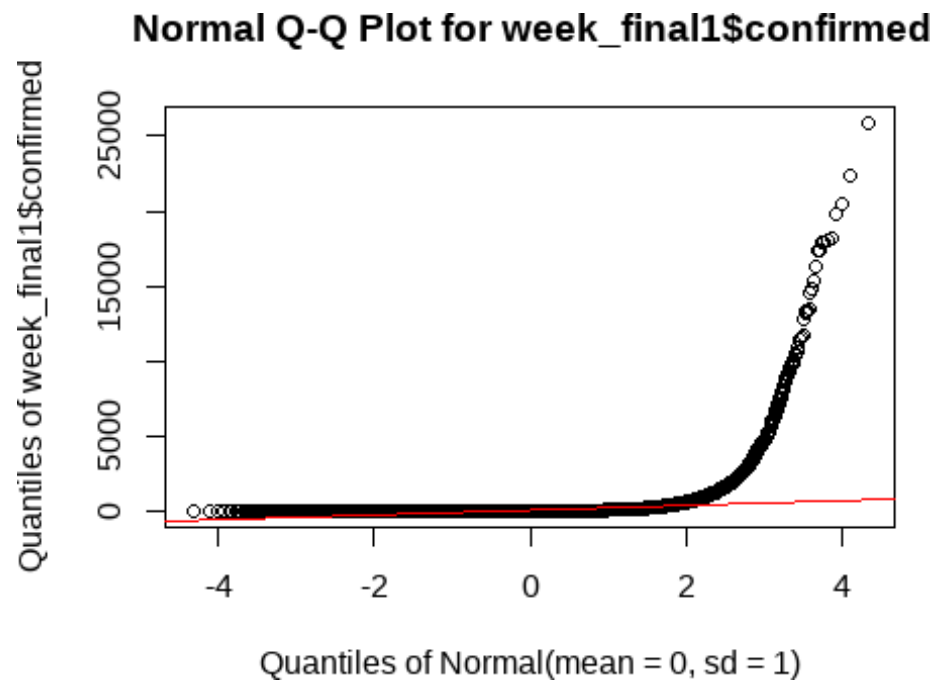
```
## The following objects are masked from 'package:stats': ##
```

```
##      predict, predict.lm
```

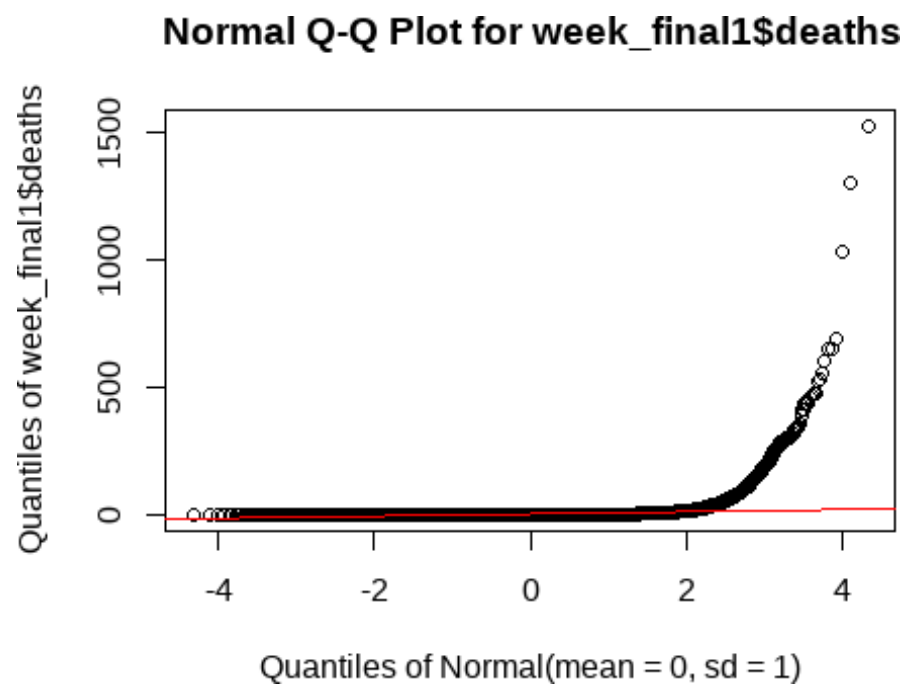
```
## The following object is masked from 'package:base': ##
```

```
##      print.default
```

```
qqPlot(week_final1$confirmed, add.line = TRUE, line.col = "red")
```



```
qqPlot(week_final1$deaths, add.line = TRUE, line.col = "red")
```



```
# test for normality using Kolmogorov-Smirnov test
# null hypothesis = sample is normal distribution

library(dgof) ##
## Attaching package: 'dgof'

## The following object is masked from 'package:stats': ##
##      ks.test ks.test(week_final1$confirmed,

"pnorm")

## Warning in ks.test(week_final1$confirmed, "pnorm"): default ks.test() cannot compute correct p-
values with ties;
## see help page for one-sample Kolmogorov test for discrete distributions.

##
## One-sample Kolmogorov-Smirnov test ##
## data: week_final1$confirmed ## D =
0.54462, p-value < 2.2e-16
## alternative hypothesis: two-sided ks.test(week_final1$deaths,

"pnorm")

## Warning in ks.test(week_final1$deaths, "pnorm"): default ks.test() cannot compute correct p-values
with ties;
## see help page for one-sample Kolmogorov test for discrete distributions.

##
## One-sample Kolmogorov-Smirnov test ##
## data: week_final1$deaths ## D =
0.5, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

both variables are not normally distributed.

check what's affecting this distribution

```
# count number of zero
week_final1 %>% count(confirmed == 0) # zeros = 30,200 / 82,320 or 36.6%

## confirmed == 0      n
## 1 FALSE 52120
## 2 TRUE 30200

week_final1 %>% count(deaths == 0) # zeros = 63,551 / 82,320 or 77.19%
```

```
## deaths == 0 n
## 1 FALSE 18769
## 2 TRUE 63551

# number of zeros combined
week_final1 %>% count(deaths == 0 & confirmed == 0) # zeros = 30,021 / 82,320 or 36.4%
```

```
## deaths == 0 & confirmed == 0 n
## 1 FALSE 52299
## 2 TRUE 30021
```

```
# number of zeros with deaths or confirmed
week_final1 %>% count(deaths == 0 | confirmed == 0) # 63,730 / 82,320 or 77.4%
```

```
## deaths == 0 | confirmed == 0 n
## 1 FALSE 18590
## 2 TRUE 63730
```

There are too many zeros in our data set that we can't just drop them. Better to analyse independently confirmed and deaths

check for outliers

```
# library(dlookr)
# diagnose_outlier(week_final1, confirmed, deaths)
outliersdf <- as.data.frame(diagnose_outlier(week_final1)) outliersdf$diff <-
(outliersdf$with_mean - outliersdf$without_mean)
```

```
# library(dplyr)
# in terms of highest difference with and without outliers mean
outliersdf %>%
```

```
  arrange(desc(diff)) %>% slice(1:15)
```

	variables	outliers_cnt	outliers_ratio	outliers_mean	with_mean
## 1	total_population	10745	13.0527211	6.792133e+05	1.247548e+05
## 2	daily_state_test	7084	8.6054422	5.346615e+04	9.986594e+03
## 3	income	3780	4.5918367	9.459195e+04	5.340991e+04
## 4	area	10185	12.3724490	4.072637e+03	1.086409e+03
## 5	population_density	11095	13.4778912	1.508432e+03	2.577202e+02
## 6	houses_density	10850	13.1802721	6.701818e+02	1.138233e+02
## 7	confirmed	12218	14.8420797	4.314966e+02	7.126099e+01

```
## 8      gdp_per_capita      3885      4.7193878  1.956587e+02  4.769686e+
01
## 9      virus_pressure      10155      12.3360058  6.152980e+01  9.918332e+
00
## 10     precipitation      4807      5.8394072   1.396803e+02  2.899104e+
01
## 11     airport_distance      1960      2.3809524   2.872448e+02  9.865959e+
01
## 12     deaths_per_100000      700      0.8503401   1.474400e+03  1.103045e+
03
## 13            deaths      18769      22.8000486   8.904683e+00  2.030272e+
00
## 14            meat_plants      6895      8.3758503   1.920305e+01  2.963010e+
00
## 15            pass_load      16380      19.8979592   3.988672e+00  8.028250e-
01
##      without_mean      diff
## 1  4.151818e+04  8.323657e+04
## 2  5.892687e+03  4.093907e+03
## 3  5.142788e+04  1.982024e+03
## 4  6.647723e+02  4.216364e+02
## 5  6.289195e+01  1.948283e+02
## 6  2.936149e+01  8.446186e+01
## 7  8.475921e+00  6.278507e+01
## 8  4.036809e+01  7.328768e+00
## 9  2.655607e+00  7.262724e+00
## 10 2.212660e+01  6.864439e+00
## 11 9.405995e+01  4.599638e+00
## 12 1.099860e+03  3.184861e+00
## 13 0.000000e+00  2.030272e+00
## 14 1.478422e+00  1.484588e+00
## 15 1.143621e-02  7.913888e-01
```

highest number of outliers

outliersdf %>%

arrange(desc(outliers_cnt)) %>% slice(1:15)

##	variables	outliers_cnt	outliers_ratio	outliers_mean	with_me
## 1	deaths	18769	22.800049	8.904683e+00	2.030272e+
## 2	pass_load	16380	19.897959	3.988672e+00	8.028250e-
## 3	confirmed	12218	14.842080	4.314966e+02	7.126099e+
## 4	population_density	11095	13.477891	1.508432e+03	2.577202e+
## 5	houses_density	10850	13.180272	6.701818e+02	1.138233e+
## 6	total_population	10745	13.052721	6.792133e+05	1.247548e+

```

05
## 7          area          10185          12.372449          4.072637e+03  1.086409e+
03
## 8      virus_pressure      10155          12.336006          6.152980e+01  9.918332e+
00
## 9      daily_state_test      7084          8.605442          5.346615e+04  9.986594e+
03
## 10     immig_student      7035          8.545918          9.519714e-02  1.578568e-
02
## 11     meat_plants      6895          8.375850          1.920305e+01  2.963010e+
00
## 12     female_percent      6335          7.695578          4.544444e-01  4.993922e-
01
## 13     college_pop      6160          7.482993          2.440349e+00  3.879712e-
01
## 14     hosp_beds      5495          6.675170          1.129817e-02  2.521076e-
03
## 15     icu_beds_ratio      5320          6.462585          1.229842e-03  3.045541e-
04
##      without_mean      diff
## 1  0.000000e+00  2.030272e+00
## 2  1.143621e-02  7.913888e-01
## 3  8.475921e+00  6.278507e+01
## 4  6.289195e+01  1.948283e+02
## 5  2.936149e+01  8.446186e+01
## 6  4.151818e+04  8.323657e+04
## 7  6.647723e+02  4.216364e+02
## 8  2.655607e+00  7.262724e+00
## 9  5.892687e+03  4.093907e+03
## 10 8.365082e-03  7.420596e-03
## 11 1.478422e+00  1.484588e+00
## 12 5.031396e-01 -3.747380e-03
## 13 2.219701e-01  1.660011e-01
## 14 1.893284e-03  6.277923e-04
## 15 2.406252e-04  6.392895e-05

```

data shows that there are too many outliers

let start our analysis during the time WHO declared COVID-19 as pandemic in Mar 11, 2020 or week 7 from 1st case

```
week_final2 <- week_final1 %>% filter(date > 7) nrow(week_final2) # 65,856 rows.
```

16,464 rows removed

```
## [1] 65856
```

```
week_final2 %>% count(confirmed == 0 | deaths == 0)
```

```
## confirmed == 0 | deaths == 0      n
```

```
## 1 FALSE 18581
```

```
## 2 TRUE 47275
```

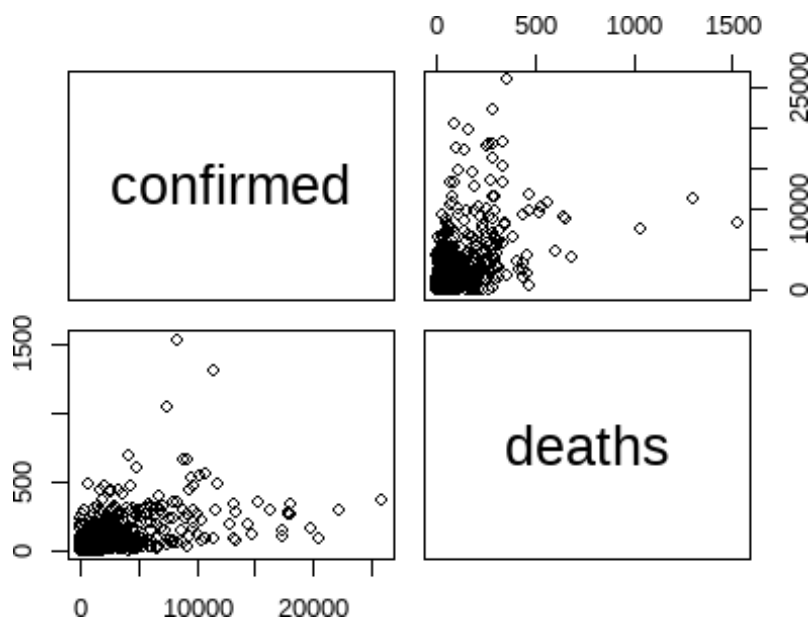
```
week_final2 %>% count(confirmed == 0) # zeros = 21.0% from 36.5%
```

```
## confirmed == 0      n
## 1      FALSE 52014
## 2      TRUE 13842
```

```
week_final2 %>% count(deaths == 0) # zeros = 71.51 from 77.19%
```

```
## deaths == 0      n
## 1      FALSE 18759
## 2      TRUE 47097
```

```
pairs(week_final1[3:4])
```



data are concentraed on the lower left

Get the Mahalanobis Distance

```
# get mahalanobis distance for response variable confirmed and deaths only as
they are skewed to the right
```

```
week_final1$mahal <- mahalanobis(week_final1[,c(3:4)],
                                colMeans(week_final1[,c(3:4)]), cov(week_final1[,c(3:4)]))
```

```
# check summary of statistics for mahalanobis distance
```

```
summary(week_final1$mahal)
```

```
##      Min.      1st Qu.      Median      Mean      3rd Qu.      Max.
## 0.000      0.023      0.028      2.000      0.028     11756.365
```


data is concentrated around mean of 2

```
# determine cut off score as reference
cutoff <- qchisq(1 - 0.05, ncol(week_final1[,3:4])) cutoff

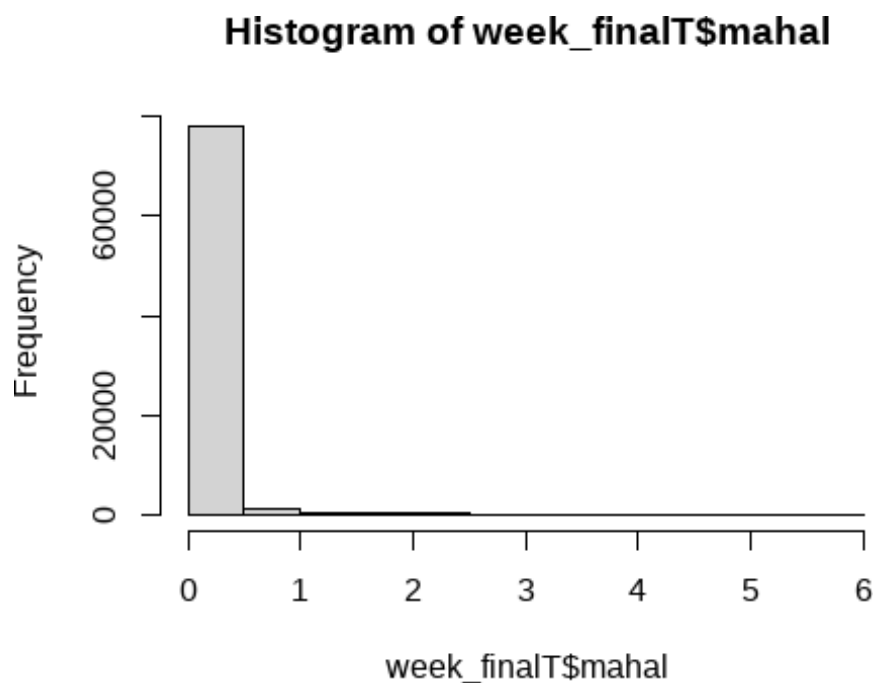
## [1] 5.991465

week_final1$badmahal <- as.numeric(week_final1$mahal > cutoff) table(week_final1$badmahal) # 1 =
yes, 0 = no

##
##      0      1
## 81049 1271
```

there are 1,271 outside the cut off score based on chi-square limit of 95%

```
# remove
week_finalT <- week_final1 %>% filter(badmahal == 0)
hist(week_finalT$mahal)
```



```
summary(week_finalT$mahal)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.000004 0.023170 0.027922 0.102839 0.027922 5.977876

summary(week_finalT$confirmed)
```

```
##      Min. 1st Qu.      Median      Mean 3rd Qu.      Max. 1122.00
##      0.00      0.00          3.00    38.29    23.00
```

summary(week_finalT\$deaths)

```
##      Min. 1st Qu.      Median      Mean 3rd Qu.      Max.
##      0.000      0.000          0.000    0.909      0.000    41.000
```

Run linear regression after removing outliers

```
model_co <- lm(formula = confirmed ~ .-deaths -date -county_fips -state_fips
-political_party -mahal - badmahal, data = week_finalT) summary(model_co)
```

```
##
## Call:
## lm(formula = confirmed ~ . - deaths - date - county_fips - state_fips - ##      political_party -
mahal - badmahal, data = week_finalT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max ## -1449.59 -
28.78  -8.46   12.91  950.08 ##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      5.342e+01    8.359e+01     0.639  0.522757
## social_dist                      1.981e+00    1.269e-01    15.610 < 2e-16 ***
## daily_state_test                  5.587e-04    2.179e-05    25.645 < 2e-16 ***
## precipitation                    2.928e-02    7.903e-03     3.705  0.000211 ***
## temperature                      1.732e+00    3.655e-02    47.386 < 2e-16 ***
## virus_pressure                   1.070e+00    1.122e-02    95.384 < 2e-16 ***
## total_population                 1.285e-04    1.990e-06    64.589 < 2e-16 ***
## female_percent                   3.153e+02    1.694e+01    18.610 < 2e-16 ***
## area                             1.421e-03    1.965e-04     7.230  4.85e-13 ***
## population_density               3.326e-02    1.983e-03    16.776 < 2e-16 ***
## hosp_beds                       -4.681e+02    1.277e+02    -3.664  0.000248 ***
## ventilator                      -2.442e+04    5.779e+03    -4.225  2.39e-05 ***
## icu_beds_ratio                   5.614e+04    6.704e+03     8.375 < 2e-16 ***
## houses_density                  -6.182e-02    3.806e-03    -16.242 < 2e-16 ***
## college_pop                     4.635e+00    1.062e+00     4.362  1.29e-05 ***
## percent_smokers                   3.983e-01    1.323e-01     3.009  0.002618 **
## percent_diabetes                -7.741e-01    9.431e-02    -8.208  2.27e-16 ***
## Religious_congregation_ratio     2.015e-01    1.999e-02    10.077 < 2e-16 ***
## airport_distance                -8.271e-02    6.056e-03    -13.657 < 2e-16 ***
## pass_load                       -2.175e-01    5.688e-02    -3.823  0.000132 ***
## meat_plants                     -1.028e+00    7.847e-02    -13.103 < 2e-16 ***
## income                          -3.069e-04    3.792e-05    -8.093  5.92e-16 ***
## percent_insured                  6.948e-01    7.110e-02     9.772 < 2e-16 ***
## deaths_per_100000               -2.937e-02    2.065e-03    -14.225 < 2e-16 ***
## gdp_per_capita                  -1.740e-02    4.600e-03    -3.784  0.000155 ***
## Age_0_19                        -3.105e+00    8.362e-01    -3.713  0.000205 ***
```

```
## Age_20_59 -1.991e+00 8.341e-01 -2.387 0.016974 *
## Age_60 -3.685e+00 8.294e-01 -4.443 8.88e-06 ***
## immig_student -1.324e+02 2.339e+01 -5.660 1.52e-08 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 81.87 on 81020 degrees of freedom
## Multiple R-squared: 0.3904, Adjusted R-squared: 0.3901
## F-statistic: 1853 on 28 and 81020 DF, p-value: < 2.2e-16

ols_vif_tol(model_co)

## Variables Tolerance VIF
## 1 social_dist 0.731335042 1.367362
## 2 daily_state_test 0.696973125 1.434776
## 3 precipitation 0.932651646 1.072212
## 4 temperature 0.686923628 1.455766
## 5 virus_pressure 0.836791625 1.195041
## 6 total_population 0.299495801 3.338945
## 7 female_percent 0.632689749 1.580554
## 8 area 0.773815666 1.292297
## 9 population_density 0.010597434 94.362466
## 10 hosp_beds 0.506585091 1.974002
## 11 ventilator 0.018682627 53.525662
## 12 icu_beds_ratio 0.017032674 58.710689
## 13 houses_density 0.011020938 90.736381
## 14 college_pop 0.127479237 7.844415
## 15 percent_smokers 0.410402836 2.436630
## 16 percent_diabetes 0.630960106 1.584886
## 17 Religious_congregation_ratio 0.710343978 1.407769
## 18 airport_distance 0.603276340 1.657615
## 19 pass_load 0.970112360 1.030808
## 20 meat_plants 0.366071424 2.731707
## 21 income 0.316793905 3.156626
## 22 percent_insured 0.671047975 1.490206
## 23 deaths_per_100000 0.249095246 4.014529
## 24 gdp_per_capita 0.875355088 1.142394
## 25 Age_0_19 0.010067602 99.328522
## 26 Age_20_59 0.008749029 114.298401
## 27 Age_60 0.004701055 212.718195
## 28 immig_student 0.139744521 7.155916

model_co1 <- lm(formula = confirmed ~ .-deaths -date -county_fips -state_fips
-political_party -mahal - badmahal -houses_density -ventilator -Age_0_19 -imm ig_student, data =
week_finalT)
summary(model_co1)

##
```

Call:

```
## lm(formula = confirmed ~ . - deaths - date - county_fips - state_fips - ## political_party -  
mahal - badmahal - houses_density - ventilator -
```

```
##      Age_0_19 - immig_student, data = week_finalT)
##
## Residuals:
##      Min        1Q      Median        3Q        Max
## -1460.83    -28.92     -8.61     12.78     949.05
##
## Coefficients:
##
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)    -2.714e+02  1.503e+01 -18.060    < 2e-16 ***
## social_dist      1.944e+00  1.271e-01  15.298    < 2e-16 ***
## daily_state_test  5.647e-04  2.182e-05  25.875    < 2e-16 ***
## precipitation    2.862e-02  7.917e-03   3.615  0.000300 ***
## temperature     1.721e+00  3.662e-02  47.013    < 2e-16 ***
## virus_pressure   1.076e+00  1.123e-02  95.841    < 2e-16 ***
## total_population 1.351e-04  1.958e-06  68.990    < 2e-16 ***
## female_percent   3.266e+02  1.691e+01  19.321    < 2e-16 ***
## area            1.254e-03  1.964e-04   6.388  1.69e-10 ***
## population_density 1.253e-03  2.232e-04   5.611  2.02e-08 ***
## hosp_beds       -3.046e+02  1.217e+02  -2.502   0.012362 *
## icu_beds_ratio   2.853e+04  1.163e+03  24.538    < 2e-16 ***
## college_pop     -5.364e-01  4.368e-01  -1.228   0.219414
## percent_smokers    4.137e-01  1.325e-01   3.122  0.001797 **
## percent_diabetes  -7.880e-01  9.428e-02  -8.358    < 2e-16 ***
## Religious_congregation_ratio 2.186e-01  1.999e-02  10.938    < 2e-16 ***
## airport_distance -8.152e-02  6.056e-03 -13.461    < 2e-16 ***
## pass_load        -2.133e-01  5.698e-02  -3.744   0.000181 ***
## meat_plants      -9.441e-01  7.839e-02 -12.044    < 2e-16 ***
## income           -2.703e-04  3.766e-05  -7.176   7.23e-13 ***
## percent_insured   6.676e-01  7.107e-02   9.393    < 2e-16 ***
## deaths_per_100000 -2.909e-02  2.060e-03 -14.127    < 2e-16 ***
## gdp_per_capita    -1.678e-02  4.561e-03  -3.680   0.000234 ***
## Age_20_59         1.291e+00  1.538e-01   8.391    < 2e-16 ***
## Age_60            -5.985e-01  1.202e-01  -4.980   6.36e-07 ' ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.03 on 81024 degrees of freedom Multiple R-
## squared:  0.3879, Adjusted R-squared:  0.3877 F-statistic: 2140 on 24 and
## 81024 DF,  p-value: < 2.2e-16
ols_vif_tol(model_co1)
##
##      Variables  Tolerance    VIF
## 1      social_dist  0.7320661  1.365997
## 2    daily_state_test  0.6973556  1.433989
## 3      precipitation  0.9329417  1.071878
## 4      temperature  0.6871159  1.455359
## 5      virus_pressure  0.8378075  1.193592
## 6    total_population  0.3105596  3.219994
## 7      female_percent  0.6378733  1.567710
```

```
## 8          area 0.7778670 1.285567
## 9    population_density 0.8393574 1.191388
## 10         hosp_beds 0.5600136 1.785671
## 11        icu_beds_ratio 0.5683951 1.759340
## 12        college_pop 0.7571691 1.320709
## 13      percent_smokers 0.4109351 2.433474
## 14      percent_diabetes 0.6338090 1.577762
## 15 Religious_congregation_ratio 0.7134481 1.401644
## 16      airport_distance 0.6057801 1.650764
## 17         pass_load 0.9708470 1.030028
## 18         meat_plants 0.3682931 2.715229
## 19         income 0.3224151 3.101592
## 20      percent_insured 0.6743367 1.482939
## 21 deaths_per_100000 0.2513216 3.978965
## 22      gdp_per_capita 0.8935999 1.119069
## 23      Age_20_59 0.2582821 3.871736
## 24      Age_60 0.2248554 4.447302
```

```
model_do <- lm(formula = deaths ~ . -date -county_fips -state_fips -political
_party -mahal - badmahal, data = week_finalT) summary(model_do)
```

```
##
```

```
## Call:
```

```
## lm(formula = deaths ~ . - date - county_fips - state_fips - political_part y -
```

```
##      mahal - badmahal, data = week_finalT) ##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -15.9575   -0.4110   -0.1799    0.0620   30.9525
```

```
##
```

```
## Coefficients:
```

```
##              Estimate      Std. Error  t value    Pr(>|t|)
## (Intercept)  -4.311e+00    2.362e+00   -1.825    0.068025 .
## confirmed    1.745e-02    9.929e-05  175.760   < 2e-16 ***
## social_dist  -3.282e-02    3.591e-03   -9.138   < 2e-16 ***
## daily_state_test -4.682e-06    6.183e-07  -7.574   3.67e-14 ***
## precipitation 7.722e-04    2.234e-04    3.457   0.000546 ***
## temperature  2.116e-03    1.047e-03    2.021   0.043265 *
## virus_pressure 4.228e-03    3.343e-04   12.647   < 2e-16 ***
## total_population 1.659e-06    5.768e-08   28.762   < 2e-16 ***
## female_percent 5.130e+00    4.798e-01   10.691   < 2e-16 ***
## area         4.674e-05    5.555e-06    8.415   < 2e-16 ***
## population_density 2.047e-04    5.613e-05    3.647   0.000265 ***
## hosp_beds    -5.656e+00    3.611e+00   -1.566   0.117256
## ventilator   -3.883e+02    1.633e+02   -2.377   0.017439 *
## icu_beds_ratio 5.396e+02    1.895e+02    2.847   0.004418 **
## houses_density -3.491e-04    1.077e-04   -3.239   0.001198 **
## college_pop  -6.583e-03    3.003e-02   -0.219   0.826485
```

```
## percent_smokers          7.148e-03   3.741e-03   1.911 0.056017 .
## percent_diabetes        6.965e-03   2.666e-03   2.612 0.008998 **
## Religious_congregation_ratio 1.233e-04   5.654e-04   0.218 0.827325
## airport_distance        -2.065e-03   1.714e-04  -12.047 < 2e-16 ***
## pass_load               1.299e-03   1.608e-03   0.808 0.419248
## meat_plants             -1.758e-02   2.220e-03  -7.917 2.46e-15 ***
## income                  2.133e-06   1.072e-06   1.989 0.046662 *
## percent_insured         8.638e-04   2.011e-03   0.430 0.667481
## deaths_per_100000       2.711e-04   5.843e-05   4.640 3.49e-06 ***
## gdp_per_capita          4.164e-04   1.300e-04   3.203 0.001362 **
## Age_0_19                7.865e-03   2.363e-02   0.333 0.739314
## Age_20_59               2.001e-02   2.357e-02   0.849 0.395913
## Age_60                  1.322e-02   2.344e-02   0.564 0.572908
## immig_student           -2.655e-01   6.611e-01  -0.402 0.687942
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.314 on 81019 degrees of freedom
## Multiple R-squared:  0.4511, Adjusted R-squared:  0.4509
## F-statistic: 2296 on 29 and 81019 DF, p-value: < 2.2e-16

ols_vif_tol(model_do)
```

##	Variables	Tolerance	VIF
## 1	confirmed	0.609642378	1.640306
## 2	social_dist	0.729142170	1.371475
## 3	daily_state_test	0.691361224	1.446422
## 4	precipitation	0.932493629	1.072393
## 5	temperature	0.668399383	1.496111
## 6	virus_pressure	0.752311151	1.329237
## 7	total_population	0.284830047	3.510866
## 8	female_percent	0.629996673	1.587310
## 9	area	0.773316665	1.293131
## 10	population_density	0.010560749	94.690257
## 11	hosp_beds	0.506501152	1.974329
## 12	ventilator	0.018678511	53.537458
## 13	icu_beds_ratio	0.017017942	58.761512
## 14	houses_density	0.010985171	91.031812
## 15	college_pop	0.127449301	7.846257
## 16	percent_smokers	0.410356965	2.436903
## 17	percent_diabetes	0.630435811	1.586204
## 18	Religious_congregation_ratio	0.709454815	1.409533
## 19	airport_distance	0.601890680	1.661431
## 20	pass_load	0.969937356	1.030994
## 21	meat_plants	0.365297346	2.737496
## 22	income	0.316538039	3.159178
## 23	percent_insured	0.670258022	1.491963
## 24	deaths_per_100000	0.248474672	4.024555

## 25	gdp_per_capita	0.875200432	1.142595
## 26	Age_0_19	0.010065889	99.345424


```
## 27                Age_20_59  0.008748413    114.306441
## 28                Age_60  0.004699910    212.770025
## 29                immig_student  0.139689286    7.158745

model_do1 <- lm(formula = deaths ~ . -date -county_fips -state_fips -politica l_party -mahal - badmahal
                -houses_density -ventilator -Age_0_19 -immig_student, data =
week_finalT) summary(model_do1)

##
## Call:
## lm(formula = deaths ~ . - date - county_fips - state_fips - political_part y -
##      mahal - badmahal - houses_density - ventilator - Age_0_19 - ##
##      immig_student, data = week_finalT)
##
## Residuals:
```

##	Min	1Q	Median	3Q	Max
##	-15.9913	-0.4096	-0.1791	0.0583	30.9688
##					

```
## Coefficients:
```

##	Estimate	Std. Error	t value	Pr(> t)	
## (Intercept)	-3.650e+00	4.248e-01	-8.593	< 2e-16	***
## confirmed	1.747e-02	9.910e-05	176.322	< 2e-16	***
## social_dist	-3.290e-02	3.590e-03	-9.165	< 2e-16	***
## daily_state_test	-4.644e-06	6.182e-07	-7.513	5.85e-14	***
## precipitation	7.608e-04	2.233e-04	3.406	0.000658	***
## temperature	2.029e-03	1.047e-03	1.938	0.052639	.
## virus_pressure	4.234e-03	3.343e-04	12.667	< 2e-16	***
## total_population	1.694e-06	5.684e-08	29.796	< 2e-16	***
## female_percent	5.201e+00	4.780e-01	10.882	< 2e-16	***
## area	4.590e-05	5.541e-06	8.285	< 2e-16	***
## population_density	2.365e-05	6.298e-06	3.755	0.000173	***
## hosp_beds	-3.069e+00	3.434e+00	-0.894	0.371587	
## icu_beds_ratio	9.610e+01	3.292e+01	2.919	0.003511	**
## college_pop	-1.588e-02	1.232e-02	-1.289	0.197489	
## percent_smokers	7.241e-03	3.739e-03	1.937	0.052760	.
## percent_diabetes	6.961e-03	2.661e-03	2.616	0.008895	**
## Religious_congregation_ratio	2.227e-04	5.643e-04	0.395	0.693072	
## airport_distance	-2.050e-03	1.710e-04	-11.989	< 2e-16	***
## pass_load	1.284e-03	1.607e-03	0.799	0.424535	
## meat_plants	-1.709e-02	2.213e-03	-7.720	1.18e-14	***
## income	2.333e-06	1.063e-06	2.195	0.028160	*
## percent_insured	7.474e-04	2.006e-03	0.373	0.709466	
## deaths_per_100000	2.779e-04	5.817e-05	4.777	1.78e-06	***
## gdp_per_capita	4.481e-04	1.287e-04	3.482	0.000498	***
## Age_20_59	1.363e-02	4.341e-03	3.139	0.001696	**

## Age_60	5.650e-03	3.390e-03	1.667	0.095602	.
## ---					

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ##
## Residual standard error: 2.314 on 81023 degrees of freedom ## Multiple R-
squared:  0.451, Adjusted R-squared:  0.4508 ## F-statistic: 2662 on 25 and
81023 DF,  p-value: < 2.2e-16

ols_vif_tol(model_do1)

##              Variables Tolerance          VIF
## 1              confirmed 0.6120822 1.633767
## 2              social_dist 0.7299578 1.369942
## 3          daily_state_test 0.6916406 1.445838
## 4              precipitation 0.9327912 1.072051
## 5              temperature 0.6688703 1.495058
## 6              virus_pressure 0.7524984 1.328907
## 7          total_population 0.2933286 3.409146
## 8              female_percent 0.6349478 1.574933
## 9                  area 0.7774754 1.286214
## 10         population_density 0.8390314 1.191851
## 11              hosp_beds 0.5599704 1.785809
## 12              icu_beds_ratio 0.5642024 1.772414
## 13              college_pop 0.7571550 1.320734
## 14              percent_smokers 0.4108857 2.433767
## 15              percent_diabetes 0.6332630 1.579123
## 16 Religious_congregation_ratio 0.7123962 1.403713
## 17              airport_distance 0.6044284 1.654456
## 18              pass_load 0.9706790 1.030207
## 19              meat_plants 0.3676349 2.720090
## 20                  income 0.3222103 3.103563
## 21              percent_insured 0.6736032 1.484554
## 22              deaths_per_100000 0.2507041 3.988765
## 23              gdp_per_capita 0.8934506 1.119256
## 24              Age_20_59 0.2580578 3.875101
## 25              Age_60 0.2247866 4.448664

bothfit.pod <- ols_step_both_p(model_do, pent = 0.05, TRUE, details =
FALSE)

## Stepwise Selection Method ##
## -----
## Candidate Terms: ##
## 1. confirmed
## 2. social_dist
## 3. daily_state_test
## 4. precipitation
## 5. temperature
## 6. virus_pressure
## 7. total_population
## 8. female_percent
```

```
## 9. area
## 10. population_density
## 11. hosp_beds
## 12. ventilator
## 13. icu_beds_ratio
## 14. houses_density
## 15. college_pop
## 16. percent_smokers
## 17. percent_diabetes
## 18. Religious_congregation_ratio
## 19. airport_distance
## 20. pass_load
## 21. meat_plants
## 22. income
## 23. percent_insured
## 24. deaths_per_100000
## 25. gdp_per_capita
## 26. Age_0_19
## 27. Age_20_59
## 28. Age_60
## 29. immig_student ##
## We are selecting variables based on p value... ##
## Variables Entered/Removed: ##
## - Age_60 added
## - female_percent added ## -
Age_60 added
## - area added ## -
income added
## - gdp_per_capita added ## -
precipitation added ## -
icu_beds_ratio added ## -
percent_diabetes added ## -
temperature added
## - college_pop added ##
## No more variables to be added/removed. ##
##
## Final Model Output ##
##
## ----- Model Summary
##
## R 0.672 RMSE 2.314
## R-Squared 0.451 Coef. Var 254.549
## Adj. R-Squared 0.451 MSE 5.354
## Pred R-Squared 0.450 MAE 0.907
```

##	-----					
##	RMSE: Root Mean Square Error					
##	MSE: Mean Square Error MAE:					
##	Mean Absolute Error					
##						
##	ANOVA					
##	-----					
##		Sum of				
##		Squares	DF	Mean Square	F	Sig.
##	-----					
##						
##	Regression	356386.041	18	19799.225	3697.84	0.0000
##	Residual ##	433856.239	81030	5.354		
Total ##		790242.281	81048			
-	-----					
##						
##						
##	Parameter Estimates					
##	-----					
##	model	Beta	Std. Error	Std. Beta	t	Sig
lower	upper					
##	-----					
##	(Intercept)	-3.176	0.339		-9.372	0.00
0	-3.840 -2.512					
##	confirmed	0.017	0.000	0.587	179.223	0.00
0	0.017 0.018					
##	social_dist	-0.034	0.004	-0.029	-9.557	0.00
0	-0.041 -0.027					
##	daily_state_test	0.000	0.000	-0.022	-7.804	0.00
0	0.000 0.000					
##	virus_pressure	0.004	0.000	0.039	13.060	0.00
0	0.004 0.005					
##	total_population	0.000	0.000	0.141	29.307	0.00
0	0.000 0.000					
##	population_density	0.000	0.000	0.096	3.812	0.00
0	0.000 0.000					
##	houses_density	0.000	0.000	-0.083	-3.387	0.00
1	-0.001 0.000					
##	airport_distance	-0.002	0.000	-0.041	-12.720	0.00
0	-0.002 -0.002					
##	meat_plants	-0.018	0.002	-0.035	-8.233	0.00
0	-0.022 -0.014					
##	deaths_per_100000	0.000	0.000	0.029	7.750	0.00
0	0.000 0.000					
##	Age_20_59	0.012	0.003	0.014	3.727	0.00

0	0.006	0.019					
##	female_percent		5.312	0.454	0.036	11.692	0.00

```
0      4.422      6.203
##           area      0.000      0.000      0.024      8.336      0.00
0      0.000      0.000
##      gdp_per_capita      0.000      0.000      0.010      3.521      0.00
0      0.000      0.001
##      precipitation      0.001      0.000      0.010      3.812      0.00
0      0.000      0.001
##      icu_beds_ratio      72.629      26.088      0.008      2.784      0.00
5      21.498      123.761
##      percent_diabetes      0.007      0.002      0.008      2.745      0.00
6      0.002      0.011
##      college_pop      -0.023      0.012      -0.006      -1.982      0.04
7      -0.047      0.000
##
```

Confirmed Cases per States

Top 5 states: California, Texas, Florida, New York, Illinois

```
ConfirmedT5 <- week_finalT %>% filter(state_fips == 6 | state_fips == 12 | state_fips == 17 | state_fips == 36 | state_fips == 17) unique(ConfirmedT5$state_fips) # checking
```

```
## [1] 12 17 6 36
```

7,953 observations

```
model_T5 <- lm(formula = confirmed ~ .-date -county_fips -state_fips -mahal - badmahal, data = ConfirmedT5)
```

```
summary(model_T5)
```

```
##
```

```
## Call:
```

```
## lm(formula = confirmed ~ . - date - county_fips - state_fips - ## mahal - badmahal, data = ConfirmedT5)
```

```
##
```

```
## Residuals:
```

```
##      Min      1Q  Median      3Q      Max ## -645.66 -39.07 -6.63 21.50 793.68
```

```
##
```

```
## Coefficients:
```

```
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.414e+02  3.428e+02  -0.704  0.481428
## deaths      1.735e+01  2.635e-01  65.840 < 2e-16 ***
## social_dist  1.027e+00  5.315e-01   1.931  0.053460 .
## daily_state_test 1.163e-03  5.223e-05  22.260 < 2e-16 ***
## precipitation 3.629e-02  2.293e-02   1.583  0.113556
## temperature -4.305e-01  1.949e-01  -2.209  0.027187 *
```

## virus_pressure	5.738e-01	1.871e-02	30.659	< 2e-16	***
## total_population	3.907e-05	4.886e-06	7.996	1.47e-15	***


```
## female_percent      -9.612e+01  8.030e+01  -1.197  0.231325
## area                4.832e-05  9.217e-04   0.052  0.958195
## population_density  2.601e-02  3.352e-03   7.759  9.59e-15 ***
## hosp_beds          -3.130e+03  1.008e+03  -3.105  0.001909 **
## ventilator         -4.286e+03  4.055e+04  -0.106  0.915810
## icu_beds_ratio      6.314e+04  4.664e+04   1.354  0.175842
## houses_density     -4.969e-02  6.336e-03  -7.843  4.97e-15 ***
## college_pop        4.046e+01  6.777e+00   5.971  2.46e-09 ***
## percent_smokers      4.303e-01  8.473e-01   0.508  0.611567
## percent_diabetes    2.193e-01  4.256e-01   0.515  0.606351
## Religious_congregation_ratio  3.796e-01  9.935e-02   3.821  0.000134 ***
## political_party    -7.029e+01  8.295e+00  -8.474  < 2e-16 ***
## airport_distance   -2.745e-02  3.034e-02  -0.905  0.365534
## pass_load          1.088e+00  3.092e+00   0.352  0.724829
## meat_plants        -8.861e-01  1.625e-01  -5.453  5.09e-08 ***
## income              4.944e-04  1.466e-04   3.372  0.000750 ***
## percent_insured     2.038e+00  6.938e-01   2.937  0.003322 **
## deaths_per_100000   1.202e-02  1.110e-02   1.083  0.278954
## gdp_per_capita      2.796e-01  7.525e-02   3.716  0.000204 ***
## Age_0_19            4.936e+00  3.475e+00   1.420  0.155522
## Age_20_59          -1.158e+00  3.500e+00  -0.331  0.740722
## Age_60              -5.926e-01  3.444e+00  -0.172  0.863392
## immig_student      -7.544e+02  1.632e+02  -4.623  3.84e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 94.56 on 7922 degrees of freedom
```

```
## F-statistic: 485.7 on 30 and 7922 DF, p-value: < 2.2e-16
```

```
bothfit.T5 <- ols_step_both_p(model_T5, pent = 0.05, prem = 0.05, progress = TRUE, details = FALSE)
```

```
## Stepwise Selection Method ##
```

```
## -----
```

```
## Candidate Terms: ##
```

- ## 1. deaths
- ## 2. social_dist
- ## 3. daily_state_test
- ## 4. precipitation
- ## 5. temperature
- ## 6. virus_pressure
- ## 7. total_population
- ## 8. female_percent
- ## 9. area
- ## 10. population_density
- ## 11. hosp_beds
- ## 12. ventilator

```
## Multiple R-squared:  0.6478, Adjusted R-squared:  0.6464
```

```
## 13. icu_beds_ratio
## 14. houses_density
## 15. college_pop
## 16. percent_smokers
## 17. percent_diabetes
## 18. Religious_congregation_ratio
## 19. political_party
## 20. airport_distance
## 21. pass_load
## 22. meat_plants
## 23. income
## 24. percent_insured
## 25. deaths_per_100000
## 26. gdp_per_capita
## 27. Age_0_19
## 28. Age_20_59
## 29. Age_60
## 30. immig_student ##
## We are selecting variables based on p value... ##
## Variables Entered/Removed: ##
## - virus_pressure added ## -
daily_state_test added ## -
total_population added ## -
airport_distance added ## -
ventilator added
## - Age_0_19 added
## - political_party added ## -
income added
## - meat_plants added ## -
college_pop added
## - Religious_congregation_ratio added ## -
airport_distance added
## - immig_student added ## -
percent_insured added ## -
social_dist added
## - pass_load added ## -
hosp_beds added
## - gdp_per_capita added ##
## No more variables to be added/removed. ##
##
## Final Model Output ##
## -----
##                                     Model Summary
## -----
```

## R		0.803	RMSE		94.961	
## R-Squared		0.644	Coef. Var		131.962	
## Adj. R-Squared		0.643	MSE		9017.575	
## Pred R-Squared		0.640	MAE		53.842	
##						
## RMSE: Root Mean Square Error						
## MSE: Mean Square Error						
## MAE: Mean Absolute Error						
##						
## ANOVA						
##						
--						
##		Sum of				
##		Squares	DF	Mean Square	F	Sig
.						
## -----						
--						
## Regression	129577716.256		17	7622218.603	845.263	0.000
0						
## Residual	71554453.661		7935	9017.575		
## Total	201132169.917		7952			
## -----						
--						
## Parameter Estimates						
## -----						
##		model	Beta	Std. Error	Std. Beta t	
##	Sig	lower	upper			
## -----						
##		(Intercept)	-340.321	48.922		-
6.956	0.000	-436.222	-244.421			
##		deaths	17.378	0.261	0.507	6
6.646	0.000	16.867	17.889			
##		virus_pressure	0.567	0.019	0.233	3
0.400	0.000	0.530	0.603			
##		daily_state_test	0.001	0.000	0.218	2
9.747	0.000	0.001	0.001			
##		total_population	0.000	0.000	0.165	1
0.568	0.000	0.000	0.000			
##		ventilator	46660.178	7485.975	0.058	
6.233	0.000	31985.699	61334.657			
##		Age_0_19	4.583	0.363	0.098	1
2.640	0.000	3.872	5.294			
##		political_party	-65.764	6.330	-0.174	-1
0.390	0.000	-78.172	-53.356			
##		income	0.001	0.000	0.050	
5.288	0.000	0.000	0.001			
##		meat_plants	-0.958	0.156	-0.086	

6.154	0.000	-1.263	-0.652		
##		college_pop	36.434	5.876	0.127
6.200	0.000	24.915	47.953		
##		Religious_congregation_ratio	0.428	0.095	0.032
4.498	0.000	0.241	0.614		
##		immig_student	-699.233	154.517	-0.091
4.525	0.000	-1002.127	-396.339		
##		percent_insured	2.028	0.581	0.058
3.492	0.000	0.890	3.166		
##		social_dist	1.531	0.515	0.022
2.976	0.003	0.523	2.540		
##		pass_load	6.105	2.935	0.015
2.080	0.038	0.351	11.859		
##		hosp_beds	-1954.238	816.198	-0.021
2.394	0.017	-3554.201	-354.276		
##		gdp_per_capita	0.115	0.053	0.020
2.199	0.028	0.013	0.218		
##					

```

-----
model_T5_step <- lm(formula = confirmed ~
                    + virus_pressure
                    + daily_state_test
                    + total_population
+ airport_distance
+ ventilator
+ Age_0_19
+ political_party
+ income
+ meat_plants
+ college_pop
+ Religious_congregation_ratio
+ immig_student
+ percent_insured
+ social_dist
+ pass_load
+ hosp_beds
+ gdp_per_capita, data =
ConfirmedT5)

summary(model_T5_step) ##
## Call:
## lm(formula = confirmed ~ +virus_pressure + daily_state_test +
##     total_population + airport_distance + ventilator + Age_0_19 +
##     political_party + income + meat_plants + college_pop + Religious_congr egation_ratio +
##     immig_student + percent_insured + social_dist + pass_load +
##     hosp_beds + gdp_per_capita, data = ConfirmedT5)

```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1072.34    -53.15   -12.64    27.52   822.71
##
## Coefficients:
##              Estimate      Std. Error  t value Pr(>|t|)
## (Intercept)   -4.508e+02    6.349e+01   -7.101 1.34e-12 ***
## virus_pressure  8.391e-01    2.268e-02   36.999 < 2e-16 ***
## daily_state_test 1.379e-03    4.499e-05   30.647 < 2e-16 ***
## total_population 1.144e-04    5.262e-06   21.743 < 2e-16 ***
## airport_distance -1.709e-01    3.264e-02   -5.235 1.69e-07 ***
## ventilator     5.893e+04    9.336e+03    6.312 2.91e-10 ***
## Age_0_19       6.453e+00    4.529e-01   14.248 < 2e-16 ***
## political_party -9.818e+01    8.260e+00  -11.887 < 2e-16 ***
## income         4.284e-04    1.265e-04    3.386 0.000712 ***
## meat_plants    -2.580e+00    1.920e-01  -13.434 < 2e-16 ***
## college_pop     4.612e+01    7.327e+00    6.294 3.25e-10 ***
## Religious_congregation_ratio 5.939e-01    1.195e-01    4.971 6.79e-07 ***
## immig_student  -9.386e+02    1.926e+02   -4.873 1.12e-06 ***
## percent_insured 3.188e+00    7.373e-01    4.324 1.55e-05 ***
## social_dist     5.983e-01    6.443e-01    0.929 0.353056
## pass_load       1.237e+01    3.662e+00    3.377 0.000735 ***
## hosp_beds      -5.006e+02    1.018e+03   -0.492 0.623035
## gdp_per_capita  1.176e-01    6.548e-02    1.796 0.072587 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 118.4 on 7935 degrees of freedom
## Multiple R-squared:  0.447, Adjusted R-squared:  0.4458 F-statistic: 377.3
## on 17 and 7935 DF, p-value: < 2.2e-16

ols_vif_tol(model_T5_step)

##              Variables  Tolerance      VIF
## 1          virus_pressure 0.8043321  1.243267
## 2        daily_state_test 0.8383056  1.192882
## 3      total_population 0.1945447  5.140207
## 4    airport_distance 0.6622137  1.510087
## 5          ventilator 0.5194865  1.924978
## 6         Age_0_19 0.7499205  1.333475
## 7      political_party 0.1452779  6.883360
## 8             income 0.4621593  2.163756
## 9         meat_plants 0.2357616  4.241574
## 10         college_pop 0.1060903  9.425935
## 11 Religious_congregation_ratio 0.8535396  1.171592
## 12         immig_student 0.1105729  9.043810
## 13      percent_insured 0.1541884  6.485571
## 14         social_dist 0.8205789  1.218652
## 15         pass_load 0.8076685  1.238132
```

```
## 16          hosp_beds 0.5999217 1.666884
## 17          gdp_per_capita 0.5646469 1.771018
```

immig_student has high VIF of 9. Will remove this and run another model

```
model_T5_step1 <- lm(formula = confirmed ~
                      + virus_pressure
                      + daily_state_test
                      + total_population
+ airport_distance
+ ventilator
+ Age_0_19
+ income
+ meat_plants
+ college_pop
+ Religious_congregation_ratio
+ percent_insured
+ social_dist
+ pass_load
+ hosp_beds
+ gdp_per_capita,
data=ConfirmedT5)# removed political party since this is categorical nominal

summary(model_T5_step1)

##
## Call:
## lm(formula = confirmed ~ +virus_pressure + daily_state_test +
##     total_population + airport_distance + ventilator + Age_0_19 +
##     income + meat_plants + college_pop + Religious_congregation_ratio + ## percent_insured
##     + social_dist + pass_load + hosp_beds + gdp_per_capita
##     ,
##     data = ConfirmedT5) ##
## Residuals:
```

##	Min	1Q	Median	3Q	Max
##	-1063.84	-52.97	-12.63	24.85	840.80
##					

```
## Coefficients:
```

##	Estimate	Std. Error	t value	Pr(> t)	
## (Intercept)	2.120e+02	3.138e+01	6.755	1.53e-11	***
## virus_pressure	8.130e-01	2.279e-02	35.680	< 2e-16	***
## daily_state_test	1.340e-03	4.532e-05	29.577	< 2e-16	***
## total_population	1.149e-04	5.313e-06	21.622	< 2e-16	***
## airport_distance	-2.906e-01	3.130e-02	-9.285	< 2e-16	***
## ventilator	8.506e+04	9.177e+03	9.268	< 2e-16	***
## Age_0_19	4.848e+00	4.365e-01	11.107	< 2e-16	***
## income	5.513e-04	1.273e-04	4.331	1.50e-05	***
## meat_plants	-2.746e+00	1.933e-01	-14.204	< 2e-16	***

## college_pop	1.306e+01	2.704e+00	4.830	1.39e-06 ***
----------------	-----------	-----------	-------	--------------

```
## Religious_congregation_ratio      4.831e-01    1.203e-01      4.017 5.95e-05 ***
## percent_insured                   -4.461e+00    3.763e-01     -11.857 < 2e-16 ***
## social_dist                       1.850e+00    6.417e-01      2.883 0.00395 **
## pass_load                         3.560e+00    3.611e+00      0.986 0.32414
## hosp_beds                        -2.207e+03    1.017e+03     -2.169 0.03010 *
## gdp_per_capita                    1.452e-03    6.484e-02      0.022 0.98213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 119.5 on 7937 degrees of freedom
## Multiple R-squared:  0.4361, Adjusted R-squared:  0.435
## F-statistic: 409.2 on 15 and 7937 DF, p-value: < 2.2e-16
```

```
ols_vif_tol(model_T5_step1)
```

	Variables	Tolerance	VIF
## 1	virus_pressure	0.8122960	1.231078
## 2	daily_state_test	0.8422542	1.187290
## 3	total_population	0.1945661	5.139642
## 4	airport_distance	0.7342436	1.361946
## 5	ventilator	0.5480946	1.824502
## 6	Age_0_19	0.8229092	1.215201
## 7	income	0.4655205	2.148133
## 8	meat_plants	0.2371772	4.216257
## 9	college_pop	0.7939595	1.259510
## 10	Religious_congregation_ratio	0.8586254	1.164652
## 11	percent_insured	0.6034627	1.657103
## 12	social_dist	0.8433392	1.185763
## 13	pass_load	0.8468587	1.180835
## 14	hosp_beds	0.6127359	1.632024
## 15	gdp_per_capita	0.5870687	1.703378

```
MAE(predict(model_T5, newdata = ConfirmedT5), ConfirmedT5$confirmed) # full model
```

```
## [1] 53.43031
```

```
MAE(predict(model_T5_step1, newdata = ConfirmedT5), ConfirmedT5$confirmed) # reduced model
```

```
## [1] 68.98505
```

```
anova(model_T5)
```

```
## Analysis of Variance Table ##
```

```
## Response: confirmed ##
```

```
## deaths
```

```
**
```

```
## social_dist
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
1	97159139	97159139	10864.8734	< 2.2e-16	*
1	1460896	1460896	163.3655	< 2.2e-16	*

**						
##	daily_state_test	1	11407073	11407073	1275.6021	< 2.2e-16 *
**						
##	precipitation	1	41253	41253	4.6131	0.0317589 *
##	temperature	1	544346	544346	60.8717	6.869e-15 *
**						
##	virus_pressure	1	10943311	10943311	1223.7416	< 2.2e-16 *
**						
##	total_population	1	3172333	3172333	354.7478	< 2.2e-16 *
**						
##	female_percent	1	111316	111316	12.4479	0.0004208 *
**						
##	area	1	50661	50661	5.6651	0.0173288 *
##	population_density	1	88547	88547	9.9019	0.0016572 *
*						
##	hosp_beds	1	17943	17943	2.0064	0.1566711
##	ventilator	1	842221	842221	94.1818	< 2.2e-16 *
**						
##	icu_beds_ratio	1	138	138	0.0155	0.9010372
##	houses_density	1	589182	589182	65.8856	5.499e-16 *
**						
##	college_pop	1	215897	215897	24.1428	9.124e-07 *
**						
##	percent_smokers	1	43161	43161	4.8265	0.0280547 *
##	percent_diabetes	1	8108	8108	0.9067	0.3410146
##	Religious_congregation_ratio	1	156995	156995	17.5561	2.820e-05 *
**						
##	political_party	1	731805	731805	81.8345	< 2.2e-16 *
**						
##	airport_distance	1	191284	191284	21.3904	3.806e-06 *
**						
##	pass_load	1	2641	2641	0.2953	0.5868664
##	meat_plants	1	437460	437460	48.9191	2.884e-12 *
**						
##	income	1	371143	371143	41.5032	1.245e-10 *
**						
##	percent_insured	1	89	89	0.0099	0.9207402
##	deaths_per_100000	1	483852	483852	54.1070	2.089e-13 *
**						
##	gdp_per_capita	1	69233	69233	7.7421	0.0054077 *
*						
##	Age_0_19	1	957881	957881	107.1156	< 2.2e-16 *
**						
##	Age_20_59	1	202	202	0.0225	0.8806594
##	Age_60	1	462	462	0.0516	0.8202537
##	immig_student	1	191109	191109	21.3709	3.845e-06 *
**						
##	Residuals	7922	70842492	8943		
##	---					
##	Signif. codes: 0 '***' 0.001 '**' 0.01 '*'			0.05 '.'	0.1 ' ' 1	

```
anova(model_T5_step1)

## Analysis of Variance      Table
##
## Response: confirmed
##
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
virus_pressure	1	48294389	48294389	3379.6803	< 2.2e-16	*
daily_state_test	1	7975593	7975593	558.1385	< 2.2e-16	*
total_population	1	17019296	17019296	1191.0241	< 2.2e-16	*
airport_distance	1	3974912	3974912	278.1675	< 2.2e-16	*
ventilator	1	1962224	1962224	137.3180	< 2.2e-16	*
Age_0_19	1	1435869	1435869	100.4833	< 2.2e-16	*
income	1	226441	226441	15.8465	6.931e-05	*
meat_plants	1	3892825	3892825	272.4230	< 2.2e-16	*
college_pop	1	49087	49087	3.4352	0.063859	.
Religious_congregation_ratio	1	99701	99701	6.9772	0.008272	*
percent_insured	1	2599366	2599366	181.9057	< 2.2e-16	*
social_dist	1	103216	103216	7.2231	0.007212	*
pass_load	1	14963	14963	1.0471	0.306205	
hosp_beds	1	67471	67471	4.7217	0.029814	*
gdp_per_capita	1	7	7	0.0005	0.982134	
Residuals	7937	113416811	14290			

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bottom 5 states: Alaska, Wyoming, New Hampshire, West Virginia, North Dakota

```
ConfirmedB5 <- week_finalT %>% filter(state_fips == 2 | state_fips == 56 | state_fips == 54 | state_fips == 38)
unique(ConfirmedB5$state_fips) # checking

## [1] 2 38 54 56

# 3,010 observations

model_B5 <- lm(formula = confirmed ~ .-date -county_fips -state_fips -mahal - badmahal, data = ConfirmedB5)
summary(model_B5)
```

```
##
## Call:
## lm(formula = confirmed ~ . - date - county_fips - state_fips -
##      mahal - badmahal, data = ConfirmedB5)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-122.46	-8.01	-0.46	5.65	384.30

```
##
## Coefficients: (1 not defined because of singularities)
##
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.975e+01	1.360e+02	0.513	0.608000	
deaths	1.439e+01	6.236e-01	23.075	< 2e-16	***
social_dist	4.846e-01	1.934e-01	2.506	0.012255	*
daily_state_test	4.587e-03	4.488e-04	10.220	< 2e-16	***
precipitation	-3.569e-02	1.858e-02	-1.921	0.054808	.
temperature	-8.498e-02	7.430e-02	-1.144	0.252838	
virus_pressure	1.308e+00	2.560e-01	5.107	3.48e-07	***
total_population	2.495e-04	2.378e-05	10.490	< 2e-16	***
female_percent	-3.431e+01	4.008e+01	-0.856	0.391993	
area	-1.208e-04	1.704e-04	-0.709	0.478422	
population_density	8.727e-02	7.551e-02	1.156	0.247916	
hosp_beds	-7.298e+02	1.884e+02	-3.873	0.000110	***
ventilator	2.963e+04	8.352e+03	3.548	0.000394	***
icu_beds_ratio	-2.185e+04	9.265e+03	-2.359	0.018396	*
houses_density	-2.164e-01	1.603e-01	-1.350	0.176997	
college_pop	-4.165e+00	3.523e+00	-1.182	0.237217	
percent_smokers	-5.013e-01	2.228e-01	-2.250	0.024547	*
percent_diabetes	-2.190e-01	2.058e-01	-1.064	0.287294	
Religious_congregation_ratio	3.239e-02	2.932e-02	1.105	0.269376	
political_party	NA	NA	NA	NA	
airport_distance	5.816e-03	8.913e-03	0.653	0.514083	
pass_load	9.274e-03	4.552e-02	0.204	0.838577	
meat_plants	1.330e+00	6.104e-01	2.179	0.029434	*
income	9.549e-06	9.473e-05	0.101	0.919715	
percent_insured	-1.066e-01	2.015e-01	-0.529	0.596807	
deaths_per_100000	-4.977e-03	3.671e-03	-1.356	0.175304	
gdp_per_capita	-4.066e-02	1.672e-02	-2.431	0.015119	*
Age_0_19	-6.308e-01	1.327e+00	-0.476	0.634448	
Age_20_59	-1.280e-01	1.297e+00	-0.099	0.921397	
Age_60	-6.627e-01	1.310e+00	-0.506	0.613009	
immig_student	8.914e+01	7.635e+01	1.167	0.243109	

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.11 on 2980 degrees of freedom
## Multiple R-squared:  0.4562, Adjusted R-squared:  0.451
## F-statistic: 86.22 on 29 and 2980 DF, p-value: < 2.2e-16
```

```
bothfit.B5 <- ols_step_both_p(model_B5, pent = 0.05, prem = 0.05, progress = TRUE, details = FALSE)

## Stepwise Selection Method ##
##
## -----
## Candidate Terms:
##
## 1. deaths
## 2. social_dist
## 3. daily_state_test
## 4. precipitation
## 5. temperature
## 6. virus_pressure
## 7. total_population
## 8. female_percent
## 9. area
## 10. population_density
## 11. hosp_beds
## 12. ventilator
## 13. icu_beds_ratio
## 14. houses_density
## 15. college_pop
## 16. percent_smokers
## 17. percent_diabetes
## 18. Religious_congregation_ratio
## 19. political_party
## 20. airport_distance
## 21. pass_load
## 22. meat_plants
## 23. income
## 24. percent_insured
## 25. deaths_per_100000
## 26. gdp_per_capita
## 27. Age_0_19
## 28. Age_20_59
## 29. Age_60
## 30. immig_student ##
## We are selecting variables based on p value... ##
## Variables Entered/Removed:
##
## Note: model has aliased coefficients
##      sums of squares computed by model comparison
##
## - deaths added
##
## Note: model has aliased coefficients
##      sums of squares computed by model comparison
```

```
## - total_population added

## Note: model has aliased coefficients
##      sums of squares computed by model comparison

## - daily_state_test added

## Note: model has aliased coefficients
##      sums of squares computed by model comparison

## - percent_diabetes added

## Note: model has aliased coefficients
##      sums of squares computed by model comparison

## - virus_pressure added

## Note: model has aliased coefficients
##      sums of squares computed by model comparison

## - houses_density added

## Note: model has aliased coefficients
##      sums of squares computed by model comparison

## - deaths_per_100000 added

## Note: model has aliased coefficients
##      sums of squares computed by model comparison

## - ventilator added

## Note: model has aliased coefficients
##      sums of squares computed by model comparison

## - hosp_beds added

## Note: model has aliased coefficients
##      sums of squares computed by model comparison

## - percent_smokers added ## -
percent_diabetes added

## Note: model has aliased coefficients
##      sums of squares computed by model comparison

## - gdp_per_capita added

## Note: model has aliased coefficients
##      sums of squares computed by model comparison
```

- social_dist added

```
## Note: model has aliased coefficients
##      sums of squares computed by model comparison

## - female_percent added

## Note: model has aliased coefficients
##      sums of squares computed by model comparison

## - meat_plants added

## Note: model has aliased coefficients
##      sums of squares computed by model comparison

## - icu_beds_ratio added

## Note: model has aliased coefficients
##      sums of squares computed by model comparison

## - precipitation added

## Note: model has aliased coefficients
##      sums of squares computed by model comparison
```

```
##
## No more variables to be added/removed. ##
##
## Final Model Output ##----
```

Model Summary

## R	0.674	RMSE	24.103
## R-Squared	0.454	Coef. Var	253.786
## Adj. R-Squared	0.451	MSE	580.951
## Pred R-Squared	0.440	MAE	11.232

```
##
## RMSE: Root Mean Square Error
## MSE: Mean Square Error Absolute
## MAE: n Error
```

```
##      Mea
```

```
##      n ANOVA
```

##	##	Sum of	##	Mean Square	F	Sig.
##		Squares	DF			
## Regression	1446404.532	15	96426.969	165.981	0.0000	
## Residual	1739365.946	2994	580.951			

##	Total	3185770.479	3009
##			
##			
##			Parameter Estimates

##						
##	model	Beta	Std. Error	Std. Beta	t	S
ig	lower	upper				
##	-----					
##	(Intercept)	44.581	15.339		2.906	0.
004	14.505	74.657				
##	deaths	14.391	0.619	0.342	23.232	0.
000	13.176	15.605				
##	total_population	0.000	0.000	0.275	12.493	0.
000	0.000	0.000				
##	daily_state_test	0.004	0.004	0.230	14.488	0.
000	0.004	1.214	0.246	0.076	4.935	0.
##	virus_pressure	1.697				
031	0.732	-0.035	0.016	-0.043	-2.154	0.
##	houses_density	-0.003				
000	-0.067	-0.010	0.002	-0.108	-5.725	0.
##	deaths_per_100000	-0.007				
001	-0.014	25145.122	7609.526	0.402	3.304	0.
##	ventilator	40065.551				
000	10224.693	-624.490	177.486	-0.087	-3.519	0.
##	hosp_beds	-276.482				
000	-972.497	-0.550	0.122	-0.075	-4.513	0.
##	percent_smokers	-0.311				
001	-0.788	-0.045	0.014	-0.055	-3.284	0.
##	gdp_per_capita	-0.018				
002	-0.072	0.561	0.183	0.050	3.059	0.
##	social_dist	0.920				
033	0.201	-67.802	31.714	-0.038	-2.138	0.
##	female_percent	-5.619				
011	-129.985	1.405	0.555	0.040	2.533	0.
##	meat_plants	2.493	8426.812	-0.245	-1.983	0.
047	0.318	-16712.519				
##	icu_beds_ratio	-189.591	0.018	-0.029	-1.975	0.
048	-33235.447	-0.036				
##	precipitation	0.000				
	-0.071					

```
model_B5_step <- lm(formula = confirmed ~  
                      + deaths  
                      + total_population  
                      + daily_state_test  
                      + percent_diabetes  
                      + virus_pressure  
                      + houses_density  
                      + deaths_per_100000  
                      + ventilator
```

```

+ hosp_beds
+ percent_smokers
+ gdp_per_capita
+ social_dist
+ female_percent
+ meat_plants
+ icu_beds_ratio
+ precipitation , data =
ConfirmedB5)

summary(model_B5_step)

##
## Call:
## lm(formula = confirmed ~ +deaths + total_population + daily_state_test + ## percent_diabetes
+ virus_pressure + houses_density + deaths_per_100000
+
## ventilator + hosp_beds + percent_smokers + gdp_per_capita + ## social_dist +
female_percent + meat_plants + icu_beds_ratio + ## precipitation, data = ConfirmedB5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max ## -124.19
-7.93 -0.70  5.89 384.36 ##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|) ## (Intercept)
              4.590e+01  1.537e+01      2.987  0.00284 **
## deaths              1.442e+01  6.197e-01  23.268 < 2e-16 ***
## total_population  2.578e-04  2.058e-05  12.530 < 2e-16 *** ## daily_state_test
4.280e-03  2.961e-04  14.455 < 2e-16 *** ## percent_diabetes -2.563e-01
1.869e-01 -1.371  0.17048
## virus_pressure      1.233e+00  2.464e-01      5.003 5.97e-07 *** ##
houses_density        -3.269e-02  1.650e-02   -1.981  0.04769 * ##
deaths_per_100000 -8.595e-03  2.205e-03   -3.897 9.94e-05 *** ## ventilator
              2.432e+04  7.632e+03      3.186  0.00146 ** ##
hosp_beds            -6.560e+02  1.789e+02   -3.666  0.00025 *** ##
percent_smokers        -4.684e-01  1.354e-01   -3.459  0.00055 *** ##
gdp_per_capita        -4.508e-02  1.381e-02   -3.265  0.00111 ** ##
social_dist           5.764e-01  1.836e-01      3.139  0.00171 ** ##
female_percent        -7.166e+01  3.183e+01   -2.251  0.02445 * ##
meat_plants           1.450e+00  5.557e-01      2.610  0.00910 ** ##
icu_beds_ratio        -1.612e+04  8.437e+03   -1.910  0.05618 . ##
precipitation         -3.439e-02  1.800e-02   -1.911  0.05614 . ## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ##
## Residual standard error: 24.1 on 2993 degrees of freedom ## Multiple R-
squared:  0.4544, Adjusted R-squared:  0.4514 ## F-statistic: 155.8 on 16 and
2993 DF, p-value: < 2.2e-16

```

```
ols_vif_tol(model_B5_step)
```

##	Variables	Tolerance	VIF
## 1	deaths	0.84008276	1.190359
## 2	total_population	0.37569305	2.661747
## 3	daily_state_test	0.72605812	1.377300
## 4	percent_diabetes	0.36674611	2.726682
## 5	virus_pressure	0.76501964	1.307156
## 6	houses_density	0.46150265	2.166835
## 7	deaths_per_100000	0.34410876	2.906058
## 8	ventilator	0.01226400	81.539491
## 9	hosp_beds	0.29359137	3.406095
## 10	percent_smokers	0.53441698	1.871198
## 11	gdp_per_capita	0.64567162	1.548775
## 12	social_dist	0.67292874	1.486041
## 13	female_percent	0.56918340	1.756903
## 14	meat_plants	0.74069429	1.350085
## 15	icu_beds_ratio	0.01191432	83.932639
## 16	precipitation	0.86247240	1.159457

will remove icu_beds with assumption that it has ventilators

```
model_B5_step1 <- lm(formula = confirmed ~  
                      + deaths
```

```
+ total_population  
+ daily_state_test  
+ percent_diabetes  
+ virus_pressure  
+ houses_density  
+ deaths_per_100000  
+ ventilator  
+ hosp_beds  
+ percent_smokers  
+ gdp_per_capita  
+ social_dist  
+ female_percent  
+ meat_plants  
+ precipitation , data =  
ConfirmedB5)
```

```
summary(model_B5_step1) ##
```

```
## Call:
```

```
## lm(formula = confirmed ~ +deaths + total population + daily state test +  
##    percent_diabetes + virus_pressure + houses_density + deaths_per_100000  
+  
##    ventilator + hosp_beds + percent_smokers + gdp_per_capita +  
##    social_dist + female_percent + meat_plants + precipitation, data =  
##    ConfirmedB5)  
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.94    -8.02    -0.73     5.81   383.97 ##

## Coefficients:
##              Estimate      Std. Error  t value Pr(>|t|)
## (Intercept)   5.171e+01   1.507e+01   3.432 0.000608 ***
## deaths        1.450e+01   6.185e-01  23.445 < 2e-16 ***
## total_population 2.600e-04  2.055e-05  12.652 < 2e-16 ***
## daily_state_test 4.321e-03  2.954e-04  14.629 < 2e-16 ***
## percent_diabetes -2.746e-01  1.867e-01  -1.471 0.141513
## virus_pressure 1.193e+00  2.456e-01   4.857 1.26e-06 ***
## houses_density -4.215e-02  1.575e-02  -2.677 0.007477 **
## deaths_per_100000 -8.548e-03  2.206e-03  -3.875 0.000109 ***
## ventilator      1.005e+04  1.578e+03   6.371 2.17e-10 ***
## hosp_beds       -7.041e+02  1.773e+02  -3.972 7.29e-05 ***
## percent_smokers  -4.758e-01  1.354e-01  -3.513 0.000449 ***
## gdp_per_capita  -4.689e-02  1.378e-02  -3.402 0.000677 ***
## social_dist      5.979e-01  1.834e-01   3.261 0.001122 **
## female_percent  -8.366e+01  3.122e+01  -2.680 0.007411 **
## meat_plants      1.343e+00  5.531e-01   2.428 0.015235 *
## precipitation    -3.491e-02  1.800e-02  -1.939 0.052602 .
## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ##

## Residual standard error: 24.11 on 2994 degrees of freedom ## Multiple R-
squared:  0.4537, Adjusted R-squared:  0.451 ## F-statistic: 165.8 on 15 and
2994 DF,  p-value: < 2.2e-16

ols_vif_tol(model_B5_step1)

##              Variables Tolerance      VIF
## 1              deaths 0.8440999 1.184694
## 2    total_population 0.3768880 2.653308
## 3    daily_state_test 0.7300036 1.369856
## 4    percent_diabetes 0.3677190 2.719468
## 5      virus_pressure 0.7705426 1.297787
## 6      houses_density 0.5071971 1.971620
## 7 deaths_per_100000 0.3441507 2.905704
## 8      ventilator 0.2872108 3.481764
## 9      hosp_beds 0.2994974 3.338927
## 10 percent_smokers 0.5348546 1.869667
## 11      gdp_per_capita 0.6487221 1.541492
## 12      social_dist 0.6754738 1.480442
## 13      female_percent 0.5922417 1.688500
## 14      meat_plants 0.7483470 1.336278
## 15      precipitation 0.8626694 1.159193

MAE(predict(model_B5, newdata = ConfirmedB5), ConfirmedB5$confirmed) # full model
```

```
## Warning in stats::predict.lm(object, ...): prediction from a rank-deficient fit
## may be misleading ##

[1] 11.34005

MAE(predict(model_B5_step1, newdata = ConfirmedB5), ConfirmedB5$confirmed) # reduced model
## [1] 11.2955
```

ANOVA

```
# create another column "rank" for top and bottom 5 states
ConfirmedT5_rank <- ConfirmedT5 %>% add_column(rank = "top")
ConfirmedB5_rank <- ConfirmedB5 %>% add_column(rank = "bottom")

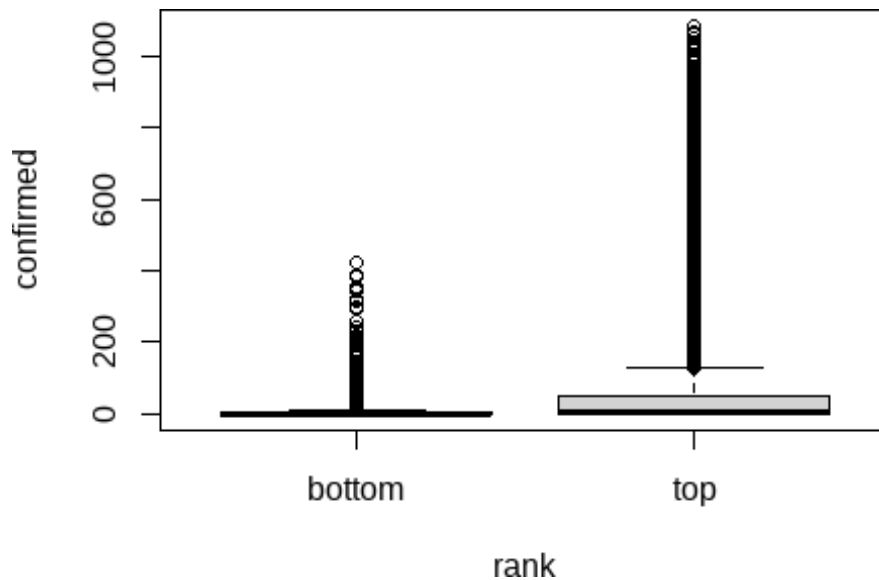
# merge 2 dataframe
Confirmed_rank <- rbind(ConfirmedT5_rank, ConfirmedB5_rank)

rm(ConfirmedT5_rank, ConfirmedB5_rank) # remove dataframes

# anova
##              Df    Sum Sq Mean Sq F value Pr(>F)
## rank          1    8519678  8519678    457.1 <2e-16 ***
## Residuals ##    10961 204317940    18640
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p value is less than 0.05, so we reject null hypothesis that both groups are equal.

```
boxplot(confirmed~rank, data = Confirmed_rank) # checking for the boxplot
```



Deaths Cases

Top 5 states: New York, New Jersey, California, Texas, Florida

```
DeathsT5 <- week_finalT %>% filter(state_fips == 36 | state_fips == 34 | state_fips == 6 | state_fips == 48 | state_fips == 12) unique(DeathsT5$state_fips) # checking
## [1] 12 34 6 36 48
# 7,953 observations

model_T5d <- lm(formula = deaths ~ .-date -county_fips -state_fips -mahal - b admahal, data = DeathsT5)
summary(model_T5d)

##
## Call:
## lm(formula = deaths ~ . - date - county_fips - state_fips - mahal - ## badmahal, data = DeathsT5)
##
## Residuals: ##
##      Min       1Q   Median       3Q      Max
## -16.2420   -0.8893  -0.3200    0.3339   29.9135
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -3.509e+00 8.403e+00 -0.418 0.676205
## confirmed 2.007e-02 2.582e-04 77.758 < 2e-16 ***
## social_dist -2.024e-02 1.394e-02 -1.452 0.146508
## daily_state_test -2.865e-05 1.684e-06 -17.010 < 2e-16 ***
## precipitation 2.433e-03 6.737e-04 3.611 0.000306 ***
## temperature 9.194e-02 5.927e-03 15.511 < 2e-16 ***
## virus_pressure -2.051e-03 5.888e-04 -3.483 0.000498 ***
## total_population 7.319e-07 1.359e-07 5.385 7.39e-08 ***
## female_percent 5.261e+00 1.762e+00 2.985 0.002838 **
## area 1.959e-05 2.877e-05 0.681 0.495930
## population_density -2.753e-04 9.495e-05 -2.899 0.003750 **
## hosp_beds 6.847e+01 2.588e+01 2.645 0.008171 **
## ventilator 1.683e+02 6.214e+02 0.271 0.786583
## icu_beds_ratio -4.040e+02 7.327e+02 -0.551 0.581356
## houses_density 5.133e-04 1.781e-04 2.882 0.003957 **
## college_pop -3.656e-01 1.340e-01 -2.728 0.006380 **
## percent_smokers 5.332e-02 1.699e-02 3.138 0.001707 **
## percent_diabetes -2.685e-03 8.740e-03 -0.307 0.758661
## Religious_congregation_ratio 6.973e-04 2.379e-03 0.293 0.769464
## political_party 2.236e+00 1.705e-01 13.113 < 2e-16 ***
## airport_distance -2.793e-03 6.828e-04 -4.090 4.34e-05 ***
## pass_load 3.137e-03 1.012e-02 0.310 0.756616
## meat_plants -1.109e-02 5.370e-03 -2.065 0.038953 *
## income 3.519e-06 3.434e-06 1.025 0.305541
## percent_insured -5.153e-02 1.131e-02 -4.558 5.22e-06 ***
## deaths_per_100000 -1.041e-05 2.341e-04 -0.044 0.964552
## gdp_per_capita 1.283e-04 2.159e-04 0.594 0.552425
## Age_0_19 2.068e-02 8.434e-02 0.245 0.806339
## Age_20_59 2.805e-02 8.433e-02 0.333 0.739420
## Age_60 5.292e-02 8.360e-02 0.633 0.526705
## immig_student 6.336e+00 2.580e+00 2.456 0.014066 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.201 on 12176 degrees of freedom
## Multiple R-squared: 0.4761, Adjusted R-squared: 0.4748
## F-statistic: 368.8 on 30 and 12176 DF, p-value: < 2.2e-16

bothfit.T5d <- ols_step_both_p(model_T5d, pent = 0.05, prem = 0.05, progress
= TRUE, details = FALSE)

## Stepwise Selection Method ##
##
## Candidate Terms: ##
## 1. confirmed
## 2. social_dist
## 3. daily_state_test
## 4. precipitation
```



```
## 5. temperature
## 6. virus_pressure
## 7. total_population
## 8. female_percent
## 9. area
## 10. population_density
## 11. hosp_beds
## 12. ventilator
## 13. icu_beds_ratio
## 14. houses_density
## 15. college_pop
## 16. percent_smokers
## 17. percent_diabetes
## 18. Religious_congregation_ratio
## 19. political_party
## 20. airport_distance
## 21. pass_load
## 22. meat_plants
## 23. income
## 24. percent_insured
## 25. deaths_per_100000
## 26. gdp_per_capita
## 27. Age_0_19
## 28. Age_20_59
## 29. Age_60
## 30. immig_student ##
## We are selecting variables based on p value... ##
## Variables Entered/Removed: ##
## - confirmed added
## - total_population added ## -
daily_state_test added ## -
temperature added
## - political_party added ## -
Age_20_59 added
## - precipitation added ## -
airport_distance added ## -
virus_pressure added ## -
hosp_beds added
## - percent_smokers added ## -
percent_insured added ## -
Age_60 added
## - meat_plants added ##
## No more variables to be added/removed. ##
##
## Final Model Output
```

```
## -----
##
##                               Model Summary
## -----
## R                               0.689          RMSE                3.203
## R-Squared                       0.475          Coef. Var          200.698
## Adj. R-Squared                   0.474          MSE                10.259
## Pred R-Squared                   0.473          MAE                1.541
##
## RMSE: Root Mean Square Error ##
MSE: Mean Square Error
## MAE: Mean Absolute Error ##
##                               ANOVA
## -----
##                               Sum of
##                               Squares          DF          Mean Square          F          Sig.
## -----
## Regression          113110.462              14          8079.319          787.564          0.0000
## Residual            125073.051             12192              10.259
## Total               238183.514             12206
## -----
##
##                               Parameter Estimates
## -----
##                               model          Beta          Std. Error          Std. Beta          t          Sig.
## lower          upper
## -----
## (Intercept)          2.464          0.792              3.112          0.002
0.912          4.016
## confirmed          0.020          0.000              0.664          78.494          0.000
0.019          0.020
## total_population          0.000          0.000              0.083          5.874          0.000
0.000          0.000
## daily_state_test          0.000          0.000             -0.191         -17.475          0.000
0.000          0.000
## temperature          0.092          0.006              0.178          15.853          0.000
0.081          0.104
## political_party          2.147          0.156              0.230          13.807          0.000
1.842          2.452
## Age_20_59          -0.038          0.011             -0.037          -3.345          0.001
-0.060         -0.016
## precipitation          0.003          0.001              0.026          3.778          0.000
0.001          0.004
## airport_distance          -0.003          0.001             -0.036          -4.790          0.000
```

-0.004	-0.002					
##	virus_pressure	-0.002	0.001	-0.023	-2.989	0.003
-0.003	-0.001					

```
##      hosp_beds      43.745      15.695      0.019      2.787      0.005
12.980      74.510
## percent_smokers      0.038      0.013      0.023      3.025      0.002
0.014      0.063
## percent_insured      -0.034      0.009      -0.055      -3.625      0.000
-0.052      -0.015
##      Age_60      0.019      0.008      0.025      2.492      0.013
0.004      0.034
##      meat_plants      -0.012      0.005      -0.030      -2.273      0.023
-0.022      -0.002
##
-----
```

```
model_T5d_step <- lm(formula = deaths ~
                      + confirmed
+ total_population
+ daily_state_test
+ temperature
+ Age_20_59
+ precipitation
+ airport_distance
+ virus_pressure
+ hosp_beds
+ percent_smokers
+ percent_insured
+ Age_60
+ meat_plants ,
data=DeathsT5) # removed political party since it's categorical nominal
```

```
##
## Call:
## lm(formula = deaths ~ +confirmed + total_population + daily_state_test +
##     temperature + Age_20_59 + precipitation + airport_distance + virus_pressure +
##     hosp_beds + percent_smokers + percent_insured + Age_60 + meat_plants, data =
##     DeathsT5)
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3600   -0.8093   -0.3268    0.2339   30.4127
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.209e+00  7.213e-01  -3.062  0.00220 **
## confirmed    1.975e-02  2.559e-04  77.168 < 2e-16 ***
## total_population  5.474e-07  1.236e-07   4.430  9.51e-06 ***
## daily_state_test -1.923e-05  1.508e-06 -12.746 < 2e-16 ***
## temperature    5.586e-02  5.234e-03  10.671 < 2e-16 ***
```

```
## precipitation      2.038e-03  6.758e-04   3.016  0.00257 **
## airport_distance   -1.683e-03  5.513e-04  -3.052  0.00228 **
## virus_pressure     -1.621e-03  5.785e-04  -2.803  0.00508 **
## hosp_beds          4.539e+01  1.582e+01   2.870  0.00411 **
## percent_smokers     -9.267e-03  1.230e-02  -0.753  0.45125
## percent_insured     6.324e-02  6.113e-03  10.346  < 2e-16 ***
## Age_60              8.429e-04  7.643e-03   0.110  0.91218
## meat_plants        1.550e-03  5.120e-03   0.303  0.76206
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ##
```

```
## Residual standard error: 3.228 on 12193 degrees of freedom ## Multiple R-
squared:  0.4667, Adjusted R-squared:  0.4661 ## F-statistic: 820.7 on 13 and
12193 DF,  p-value: < 2.2e-16
```

```
ols_vif_tol(model_T5d_step)
```

```
##          Variables Tolerance          VIF
## 1          confirmed 0.6035658 1.656820
## 2 total_population 0.2182577 4.581740
## 3 daily_state_test 0.4388161 2.278859
## 4          temperature 0.4303568 2.323653
## 5          Age_20_59 0.3684285 2.714231
## 6          precipitation 0.9342694 1.070355
## 7 airport_distance 0.7882525 1.268629
## 8          virus_pressure 0.7581458 1.319007
## 9          hosp_beds 0.9551414 1.046965
## 10 percent_smokers 0.7860575 1.272172
## 11 percent_insured 0.4354491 2.296480
## 12          Age_60 0.4374165 2.286151
## 13          meat_plants 0.2648353 3.775931
```

```
MAE(predict(model_T5d, newdata = DeathsT5), DeathsT5$deaths) # full model
```

```
## [1] 1.546916
```

```
MAE(predict(model_T5d_step, newdata = DeathsT5), DeathsT5$deaths) # reduced model
```

```
## [1] 1.511238
```

Bottom 5 states: Alaska, Wyoming, Vermont, Hawaii, Maine

```
DeathsB5 <- week_finalT %>% filter(state_fips == 2 | state_fips == 56 | state
_fips == 50 | state_fips == 15 | state_fips == 23)
unique(DeathsB5$state_fips) # checking
```

```
## [1] 15  2 23 50 56
```

```
# 1,885 observations
```

```

model_B5d <- lm(formula = deaths ~ .-date -county_fips -state_fips -mahal - b admahal, data =
DeathsB5)
summary(model_B5d)

##
## Call:
## lm(formula = deaths ~ . - date - county_fips - state_fips - mahal - ## badmahal, data =
DeathsB5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.5316 -0.1506 -0.0363  0.0671  7.1162 ##
## Coefficients:
##              Estimate      Std. Error  t value Pr(>|t|)
## (Intercept)   7.160e-02   5.488e+00   0.013   0.9896
## confirmed     1.411e-02   5.678e-04  24.860 < 2e-16 ***
## social_dist   -3.647e-02   5.602e-03  -6.510 9.63e-11 ***
## daily_state_test
## precipitation -1.200e-04   6.297e-04  -0.191  0.8489
## temperature   -2.763e-03   1.939e-03  -1.425  0.1544
## virus_pressure -1.536e-02   1.021e-02  -1.504  0.1328
## total_population
## female_percent 2.027e+00   2.612e+00   0.776  0.4378
## area          3.810e-06   8.836e-06   0.431  0.6664
## population_density
## hosp_beds      -7.299e+00   1.183e+01  -0.617  0.5374
## ventilator     5.520e+02   5.701e+02   0.968  0.3331
## icu_beds_ratio -3.627e+02   6.455e+02  -0.562  0.5743
## houses_density  1.450e-02   2.520e-03   5.756 1.01e-08 ***
## college_pop    -1.849e-01   1.669e-01  -1.108  0.2682
## percent_smokers -2.572e-02   1.469e-02  -1.750  0.0802 .
## percent_diabetes
## Religious_congregation_ratio
## political_party  9.845e-03   5.644e-02   0.174  0.8616
## airport_distance
## pass_load      2.907e-03   2.270e-02   0.128  0.8981
## meat_plants    -1.096e-02   1.415e-02  -0.775  0.4386
## income         -1.579e-06   3.385e-06  -0.466  0.6410
## percent_insured -5.927e-03   8.702e-03  -0.681  0.4959
## deaths_per_100000
## gdp_per_capita  7.760e-05   1.273e-03   0.061  0.9514
## Age_0_19       4.421e-03   6.138e-02   0.072  0.9426
## Age_20_59      4.401e-03   5.346e-02   0.082  0.9344
## Age_60        -1.691e-02   5.740e-02  -0.295  0.7684
## immig_student  5.249e+00   3.990e+00   1.316  0.1885
## ---
## Signif. codes:  0      '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5975 on 1854 degrees of freedom

```

```
## Multiple R-squared:  0.3831, Adjusted R-squared:  0.3731 ## F-statistic:
38.38 on 30 and 1854 DF,  p-value: < 2.2e-16

bothfit.B5d <- ols_step_both_p(model_B5d, pent = 0.05, prem = 0.05, progress
= TRUE, details = FALSE)

## Stepwise Selection Method ##
## -----
##
## Candidate Terms:
##
## 1. confirmed
## 2. social_dist
## 3. daily_state_test
## 4. precipitation
## 5. temperature
## 6. virus_pressure
## 7. total_population
## 8. female_percent
## 9. area
## 10. population_density
## 11. hosp_beds
## 12. ventilator
## 13. icu_beds_ratio
## 14. houses_density
## 15. college_pop
## 16. percent_smokers
## 17. percent_diabetes
## 18. Religious_congregation_ratio
## 19. political_party
## 20. airport_distance
## 21. pass_load
## 22. meat_plants
## 23. income
## 24. percent_insured
## 25. deaths_per_100000
## 26. gdp_per_capita
## 27. Age_0_19
## 28. Age_20_59
## 29. Age_60
## 30. immig_student ##
## We are selecting variables based on p value... ##
## Variables Entered/Removed: ##
## - confirmed added
## - female_percent added ## -
social_dist added ## - Age_60
added
```

```
## - temperature added
## - airport_distance added
## - gdp_per_capita added
##
## No more variables to be added/removed.
##
##
## Final Model Output
## -----
##
##                               Model Summary
##
## R                0.594          RMSE                0.608
## R-Squared        0.353          Coef.   Var          412.447
## Adj. R-Squared   0.350          MSE                0.370
## Pred R-Squared   0.307          MAE                0.229
##
## RMSE: Root Mean Square Error
## MSE: Mean Square Error
## MAE: Mean Absolute Error
##
##                               ANOVA
##
##                Sum of
##                Squares          DF      Mean Square          F          Sig.
## Regression      378.508           7          54.073        146.141      0.0000
## Residual        694.493        1877           0.370
## Total          1073.001        1884
##
##                               Parameter Estimates
##
## -----
##                model      Beta      Std. Error      Std. Beta      t      Sig
## lower      upper
## -----
##      (Intercept)      -2.728          0.622                -4.388      0.000
## -3.947      -1.509
##      confirmed         0.015          0.001                0.569      29.512      0.000
## 0.014         0.016
##      female_percent     6.474          1.276                0.135      5.075      0.000
## 3.972         8.975
##      social_dist       -0.029          0.005               -0.109     -5.390      0.000
## -0.039       -0.018
##      Age_60            -0.008          0.004               -0.053     -2.097      0.036
```


-0.016	-0.001					
##	temperature	-0.006	0.001	-0.075	-3.826	0.000
-0.008	-0.003					

```
## airport_distance      -0.001      0.000      -0.076      -3.395      0.001
-0.001      0.000
##   gdp_per_capita      0.002      0.001      0.060      2.270      0.023
0.000      0.003
## -----
```

```
model_B5d_step <- lm(formula = deaths ~
                      + confirmed
```

```
+ female_percent
+ social_dist
+ Age_60
+ temperature
+ airport_distance
+ gdp_per_capita , data =
DeathsB5)
```

```
summary(model_B5d_step)
```

```
##
```

```
## Call:
```

```
## lm(formula = deaths ~ +confirmed + female_percent + social_dist      +
##     Age_60 + temperature + airport_distance + gdp_per_capita, data = DeathsB5)
```

```
##
```

```
## Residuals:
```

```
##      Min      1Q  Median      3Q      Max
## -8.0221 -0.1443 -0.0514  0.0289  7.2352
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
```

```
##
```

```
## (Intercept)      -2.7280087  0.6216608  -4.388 1.21e-05 ***
```

```
## confirmed         0.0148675  0.0005038  29.512 < 2e-16 ***
```

```
## female_percent    6.4735686  1.2755469    5.075 4.26e-07 ***
```

```
## social_dist      -0.0285662  0.0052994  -5.390 7.92e-08 ***
```

```
## Age_60            -0.0082490  0.0039346  -2.097 0.036168 *
```

```
## temperature      -0.0055754  0.0014574   -3.826 0.000135 ***
```

```
## airport_distance  -0.0007928  0.0002335   -3.395 0.000701 ***
```

```
## gdp_per_capita    0.0016603  0.0007314    2.270 0.023318 *
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ##
```

```
## Residual standard error: 0.6083 on 1877 degrees of freedom ## Multiple R-
squared:  0.3528, Adjusted R-squared:  0.3503 ## F-statistic: 146.1 on 7 and
1877 DF,  p-value: < 2.2e-16
```

```
ols_vif_tol(model_B5d_step)
```

```
##          Variables Tolerance      VIF
```

```
## 1          confirmed 0.9263912 1.079458
```

```
## 2    female_percent 0.4908379 2.037332
```

```
## 3      social_dist 0.8510990 1.174951
## 4      Age_60     0.5459178 1.831777
## 5      temperature 0.8912085 1.122072
## 6      airport_distance 0.6806707 1.469139
## 7      gdp_per_capita 0.4993489 2.002608
```

```
MAE(predict(model_B5d, newdata = DeathsB5), DeathsB5$deaths) # full model
```

```
## [1] 0.2395637
```

```
MAE(predict(model_B5d_step, newdata = DeathsB5), DeathsB5$deaths) # reduced model
```

```
## [1] 0.2291375
```

Anova

```
# create another column "rank" for top and bottom 5 states
```

```
DeathsT5_rank <- DeathsT5 %>% add_column(rank = "top") DeathsB5_rank <-
DeathsB5 %>% add_column(rank = "bottom")
```

```
# merge 2 dataframe
```

```
Deaths_rank <- rbind(DeathsT5_rank, DeathsB5_rank)
```

```
rm(DeathsT5_rank, DeathsB5_rank) # remove dataframes #
```

```
anova
```

```
summary(aov(deaths~rank, data = Deaths_rank))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## rank          1   3426      3426   201.7 <2e-16 ***
## Residuals ##    14090 239257         17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p value is less than 0.05, so we reject null hypothesis that both groups are equal.

```
boxplot(deaths~rank, data = Deaths_rank) # checking for the boxplot
```

