Statistical Analysis of

COVID-19 DATA

Derek Wong 998761202

Ninh Nguyen 917287958

Diana Rueda 912172783

Youngchan Park 914955115

Dr. Fushing Hsieh STA 160

June 9th, 2021

Abstract

After the devastating impact of the COVID-19, governments are distributing vaccines and reopening establishments. With the resources available from CDC open dataset, this report will be able to inform you of an overview of the ongoing crisis and our take on the decision of reopening.

Focusing on cases in the US, we try to find the relationship between COVID-19 cases, deaths, state, sex, age groups, and the capacity of hospitals.

I. *Introduction*

II. *Description of data*

III.  *Materials and Methods*

1. *Exploratory Data Analysis*
2. *Checking of model assumptions*
3. *Model Diagnostics*

IV. *Conclusion and Discussion*

V. Appendix

1. *Data Visualization*

VI. References

1. ***Introduction***

The 2019 Novel CoronaVirus has swept every individual in the United States and the globe. Unlike SARS and EBOLA viruses, the COVID-19 is highly contagious with a high mortality rate. Unfortunately, this virus reveals many social problems and vulnerable parts in the community. In the short period of time, many agencies including government, university and laboratories have collected many COVID-19 data to fight back the pandemic and to prevent spreading of the virus. Unfortunately, the data is complicated and needs good insight to analyze. To handle this problem, it is necessary to have a deep understanding of data, statistical techniques and understanding of how to implement the statistical model.

Analyzing features of COVID-19 data would help to prevent the further pandemic situation. Our team decided to analyze the unique features or patterns of COVID-19 about hospitalization, deaths by sex and age group, and how politcial perspective affects on vaccination distribution. For each subject, we would like to use any statistical model including linear regression model. Throughout our analysis, we will be using a multitude of tools and packages in the programming languages of R and Python.

**II. *Description of data***

**Hospitals**

After exploring Confirmed Covid cases, we ask if the capacity of hospitals affects Deaths cases.To find the relationship between COVID cases and hospital capacity in the US, we combined "covid19.analytics" R package with dataset population in state in R (data(usmap::statepop)) and dataset about hospitals at <https://opendata.arcgis.com/datasets/6ac5e325468c4cb9b905f1728d6fbf0f_0.csv>.

| **Variable Name** | **Description** |
| --- | --- |
| abbr | Name of state in the US |
| pop2015 | Population in the US 2015 |
| STATES\_BEDS | Number of beds in hospitals of states in theUS |
| HOSPITAL | Number of hospitals of states in the US |
| MISSING BEDS | Number of hospitals don’t fill out the the number of beds of states |
| Confirmed | Confirmed Covid cases of states in the US |
| Deaths | Confirmed Covid cases of states in the US |

**--------------------------------------------------------------------------------------------------**

**Deaths**

We would use CDC COVID-19 data “Provisional\_COVID-19\_Deaths\_by\_Sex\_and\_Age” by NCHS which is approved by CDC and there are three data types in the dataset: date, numerical and categorical (string).

| **Variable Name** | **Description** |
| --- | --- |
| Data As Of | Date of analysis  (Date & Time) |
| Start Date | First date of data period  (Date & Time) |
| End Data | Last date of data period  (Date & Time) |
| Group | Indicator of whether data measured by Month, by Year, or Total  (numerical) |
| Year | Year in which death occurred  (numerical) |
| Month | Smaller more precise area within the neighbourhoods in neighbourhood\_group in which the posting is located  (Month in which death occurred) |
| State | Jurisdiction of occurrence  (categorical) |
| Sex | Sex  (categorical) |
| Age Group | Age group  (categorical) |
| COVID-19 Deaths | Deaths involving COVID-19  (numerical) |
| Total Deaths | Deaths from all causes of death  (numerical) |
| Pneumonia Deaths | Pneumonia Deaths  (numerical) |
| Pneumonia and COVID-19 Deaths | Deaths with Pneumonia and COVID-19  (numerical) |
| Influenza Deaths | Influenza Deaths  (numerical) |
| Pneumonia, Influenza, or COVID-19 Deaths | Deaths with Pneumonia, Influenza, or COVID-19  (numerical) |

**--------------------------------------------------------------------------------------------------**

**Vaccination**

The data used to explore COVID-19 vaccination was an integration and transformation of datasets obtained from the CDC data center, the Github repository of Our World in Data, USAFacts.org and NYDatabases.com. This data was structured into 4 different data frames which served as the base for our analysis. The first three data frames contain data in COVID-19 vaccine distribution by State and manufacturers. The columns in each are as follows:

* Johnson & Johnson/Jenssen COVID-19 Vaccine Allocations

| **Column Name** | **Description** |
| --- | --- |
| Jurisdiction | Object: US State name which received the vaccines |
| Week of Allocations | Date: Week in which the vaccines were distributed |
| 1st Dose Allocations | Numeric: number of vaccines distributed for said week |

* Moderna COVID-19 Vaccine Allocations

| **Column Name** | **Description** |
| --- | --- |
| Jurisdiction | Object: US State name which received the vaccines |
| Week of Allocations | Date: Week in which the vaccines were distributed |
| 1st Dose Allocations | Numeric: number of vaccines distributed for said week |
| 2nd Dose Allocations | Numeric: number of vaccines distributed as a second dose |

* Pfizer COVID-19 Vaccine Allocations

| **Column Name** | **Description** |
| --- | --- |
| Jurisdiction | Object: US State name which received the vaccines |
| Week of Allocations | Date: Week in which the vaccines were distributed |
| 1st Dose Allocations | Numeric: number of vaccines distributed for said week |
| 2nd Dose Allocations | Numeric: number of vaccines distributed intended for a second dose |

The remaining data frame contains percentages on COVID-19 vaccine hesitation, cases, deaths and political attitudes during our last Presidential Election. The columns are as follow:

* COVID-19 Vaccine

| **Column Name** | **Description** |
| --- | --- |
| FIPS\_code | Numeric: FIPS identification code unique to each US county for purpose of plotting |
| County | Object: Names of counties and States abbreviations |
| strongly\_hesitant | Numeric: Estimated percentage of people in the county who feel strongly hesitant about the COVID-19 vaccine |
| fully\_vac\_adults | Numeric: percentage of adults in the county who are fully vaccinated against COVID-19 |
| total\_cases | Numeric: percentage of people in the county who have been infected with COVID-19 at any point during the pandemic |
| total\_deaths | Numeric: percentage of accumulated deaths in the county per number of total cases |
| 2020\_Trump\_votes | Numeric: percentage of votes in favor of Donald Trump per county during the 2020 Election |

The vaccine distribution data frames were used to explore the trends of distribution throughout States with visualizations while the last data frame described was used to explore linear models that might provide insight into linear correlation between some of these variables.

**---------------------------------------------------------------------------------------------------**

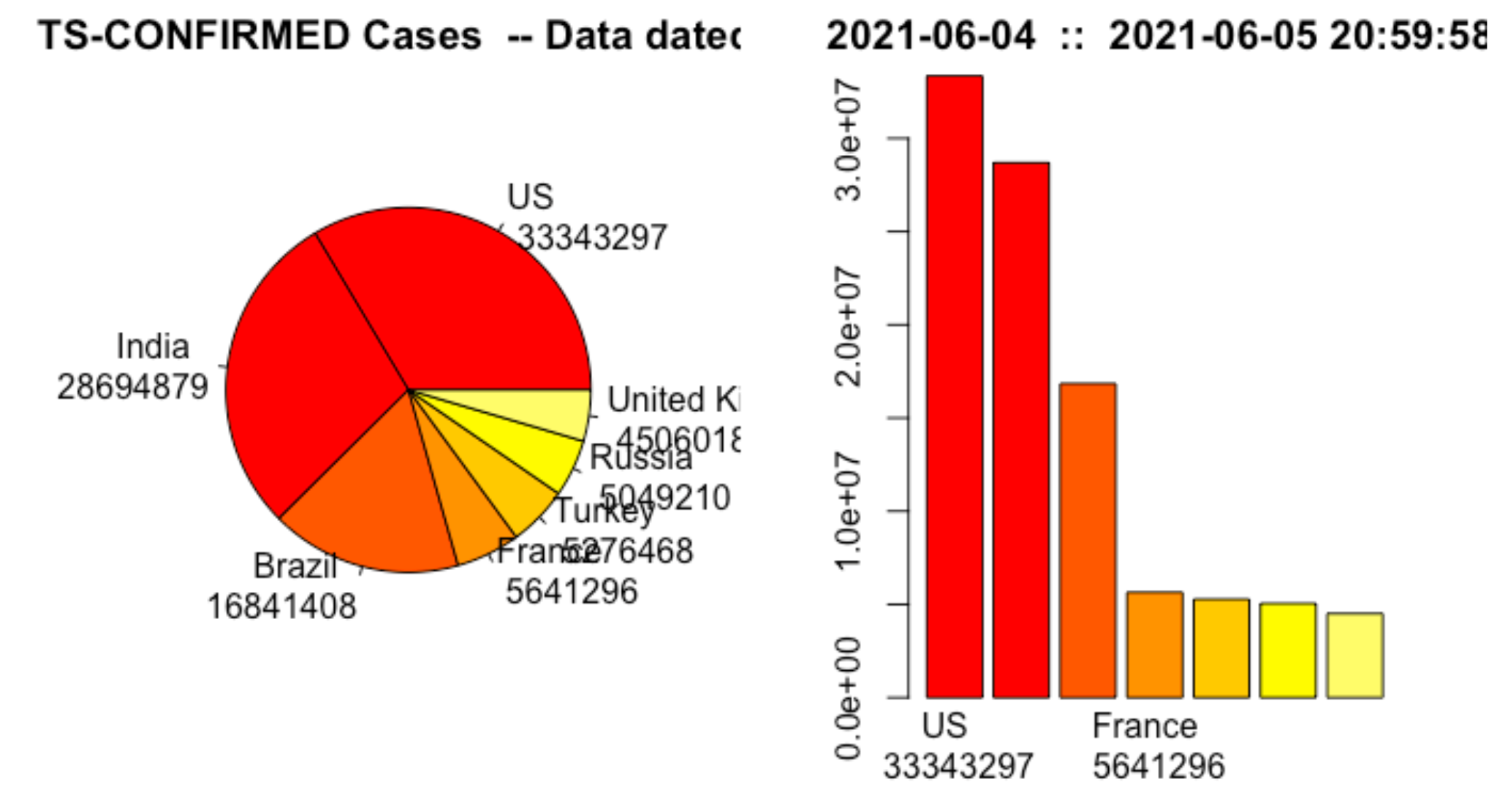
**General**

The "covid19.analytics" R package allows users to obtain live worldwide data from CoronaVirus Disease originally reported in 2019, CoViD-19, as published by the John Hopkins University CCSE repository. This package also provides basic analysis tools and functions to investigate these datasets.

Below is an overview of the top 7 regions with confirmed cases by using time series analysis:







It is clear that despite the distribution of vaccines, positive Covid-19 cases increased at a rate of 2000 cases in two days.

**III. *Materials and Methods***

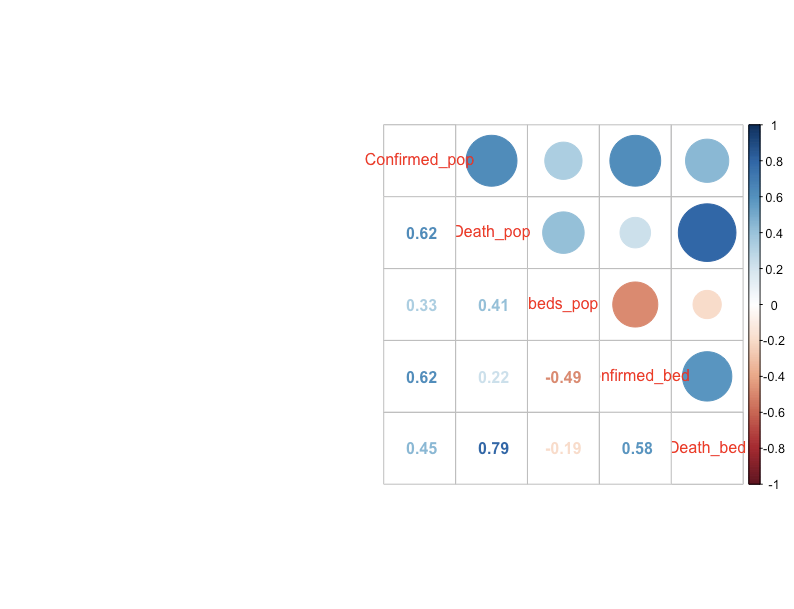
### ***A.*** ***Exploratory Data Analysis***

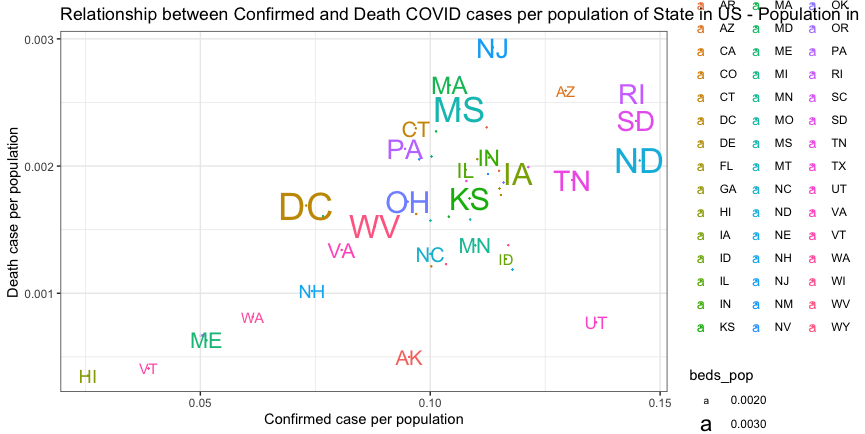
**Hospitals**

After cleaning data, we standardize the data by calculating the ratio between case/per population, case/bed in hospitals and bed/populations, and we just use data from the active hospitals.

The correlation between the ratios. We can see that there are negative relationships between the ratio beds/population and ratio death Covid case/beds, the ratio beds/population and ratio Confirmed Covid case/beds.

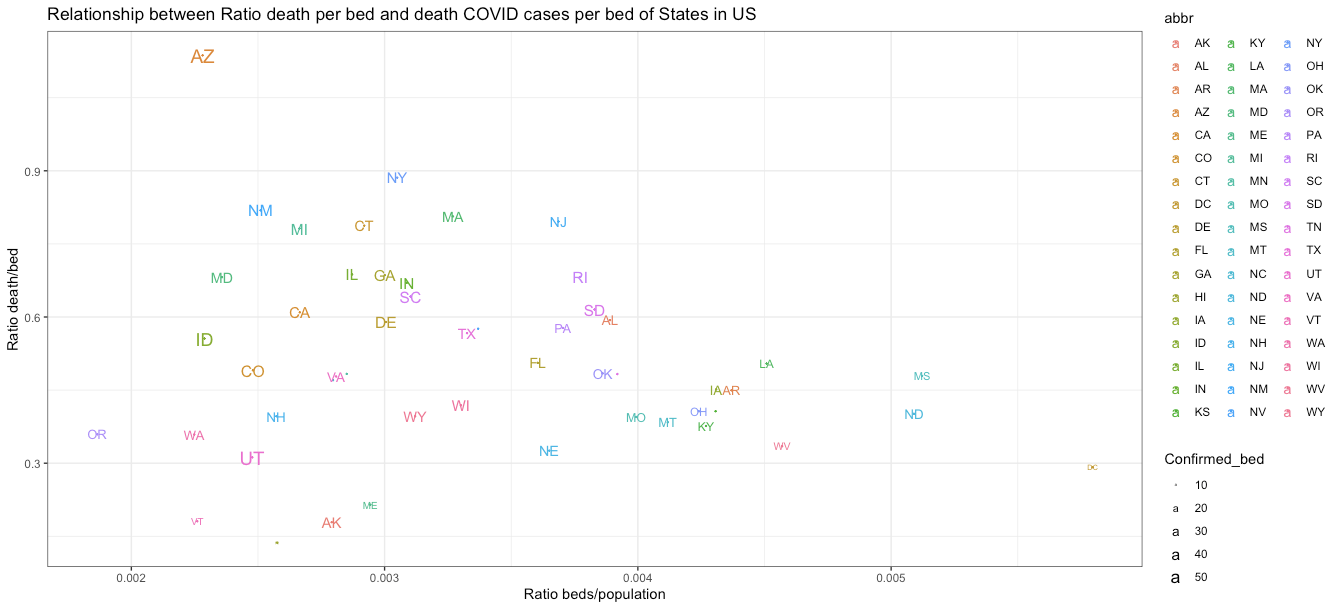
**Visualizations**

****



Font size: ratio beds/population in the states.

There is a positive correlation between Confirmed and Death case per population. The size of the name is the ratio between beds/population.



Font size: ratio Confirmed case/beds in the states.

As we can see, there are negative relationships between the ratio beds/populations and the ratio between death COVID cases/beds of hospitals in each state in the US.

In DC, beds/population is high so the death case per bed is low, and the confirmed COVID cases are low too.

In AZ, beds/population is low so the death case per bed is high.

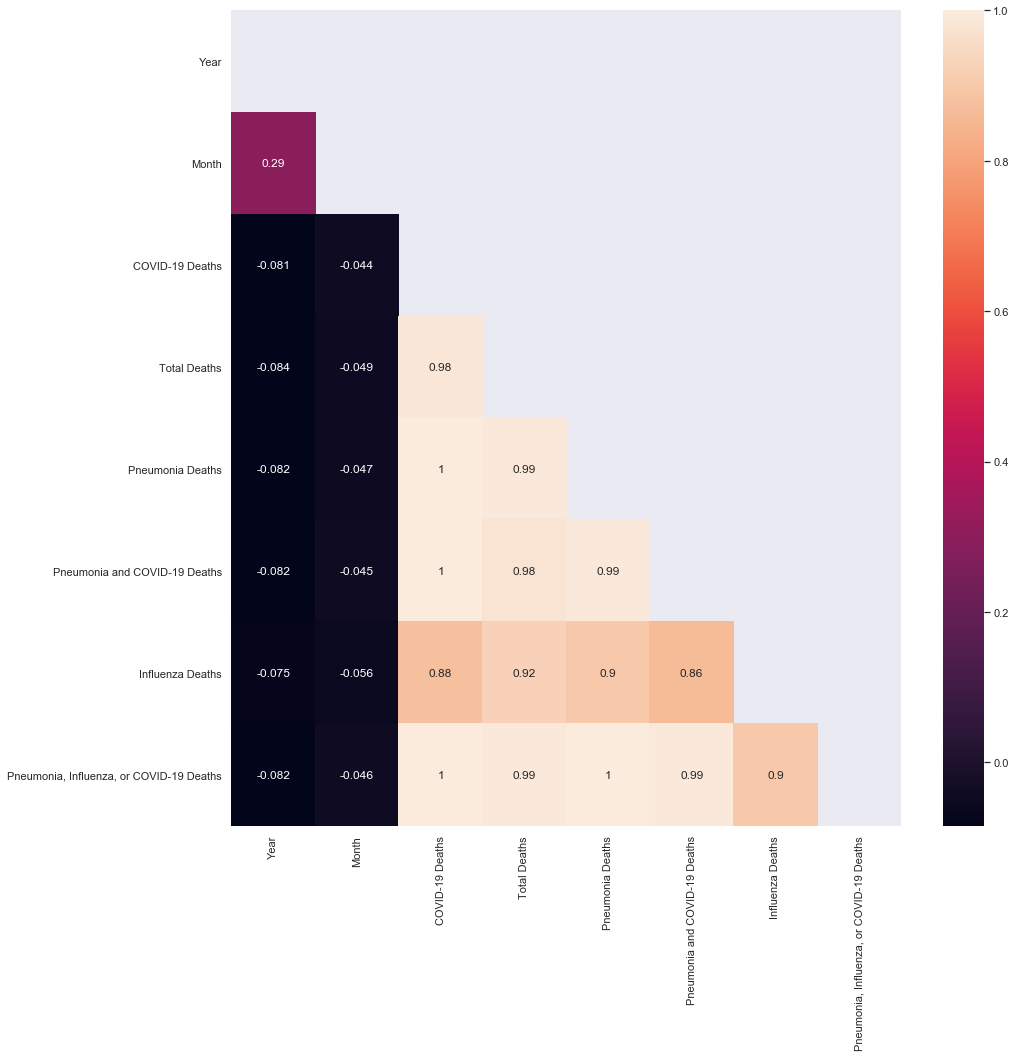
In UT, the confirmed cases/beds are high but the death case is low even though beds/ population is normal because the population in UT is not high.

In NY, confirmed cases are the same with a lot of other states like AK, WY, GA, DE... but the death cases per bed is high in normal capacity of hospitals.

Zoom in the Confirmed case/bed is larger than 50, we have 3 states ( AZ, ID, UT) with the ratio of death cases/beds so different, even though the ratio beds/population is not much different.

**--------------------------------------------------------------------------------------------------**

**Deaths by sex and age group**

******

**Figure 1. Correlation matrix between variables**

From the above plot, the number of COVID-19 and total Pneumonia deaths gives a correlation as mostly 1 which shows perfectly related with each variable. Furthermore, influenza deaths are correlated with other deaths with mostly .9. So I regard all deaths in the dataset as mostly related to each other's deaths. Furthermore, in the matter of analyzing data about COVID-19, COVID-19 influences the number of total deaths in the US.

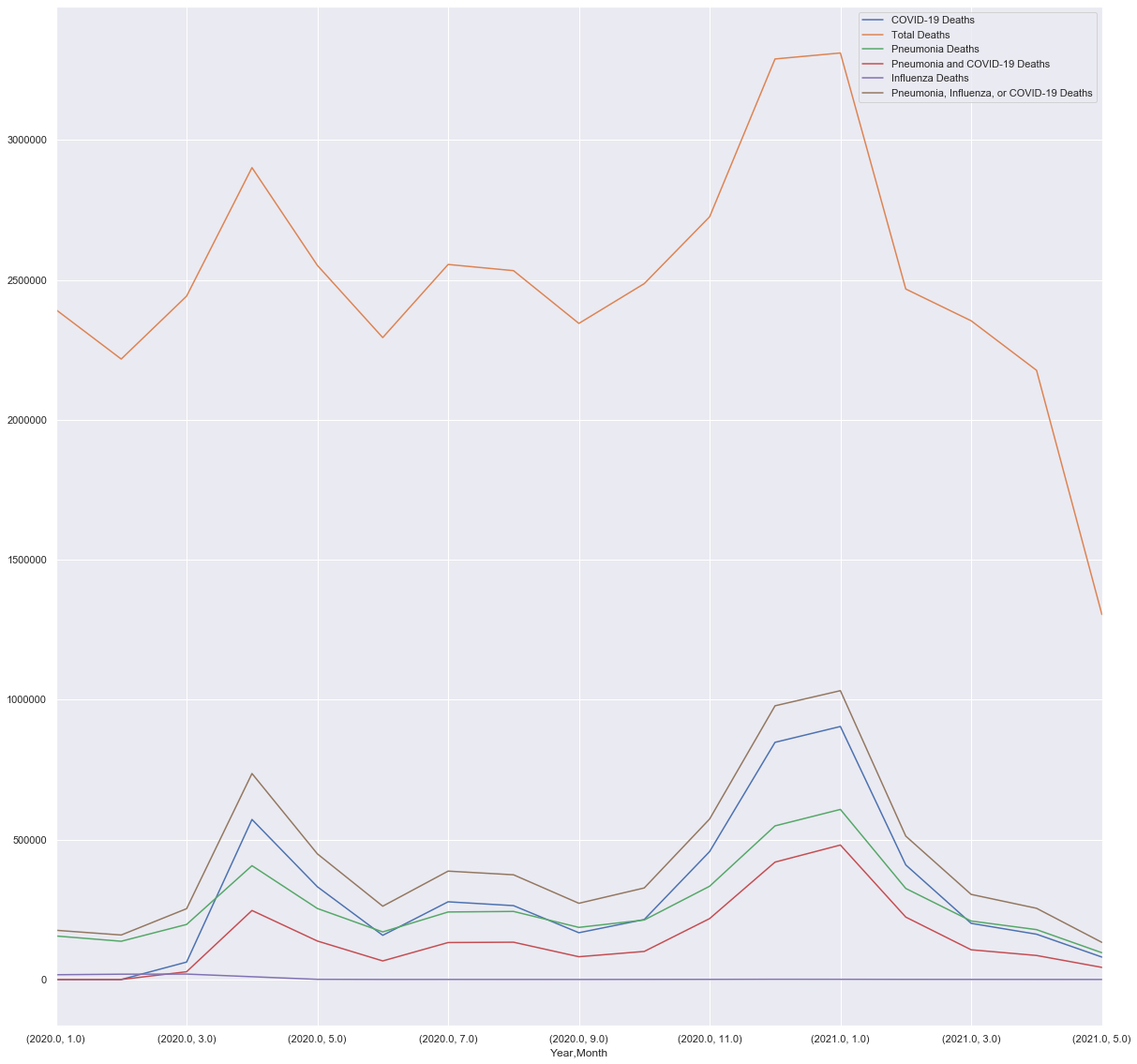


Figure 2. The number of deaths for each month

From the above plot, all causes of death including the number of COVID-19, total deaths, pneumonia deaths show a very similar trend similar to the correlation matrix. Furthermore, the number of COVID-19 deaths are similar to the number of pneumonia deaths. But in April and the winter season, the number of COVI-19 deaths greatly outnumber the number of pneumonia deaths. Similar to the correlation matrix of different causes of death, the number of COVID-19 and total deaths show the same trends with dramatic increases in April and the winter season. This implies that COVID-19 deaths are more seasonal than pneumonia deaths.

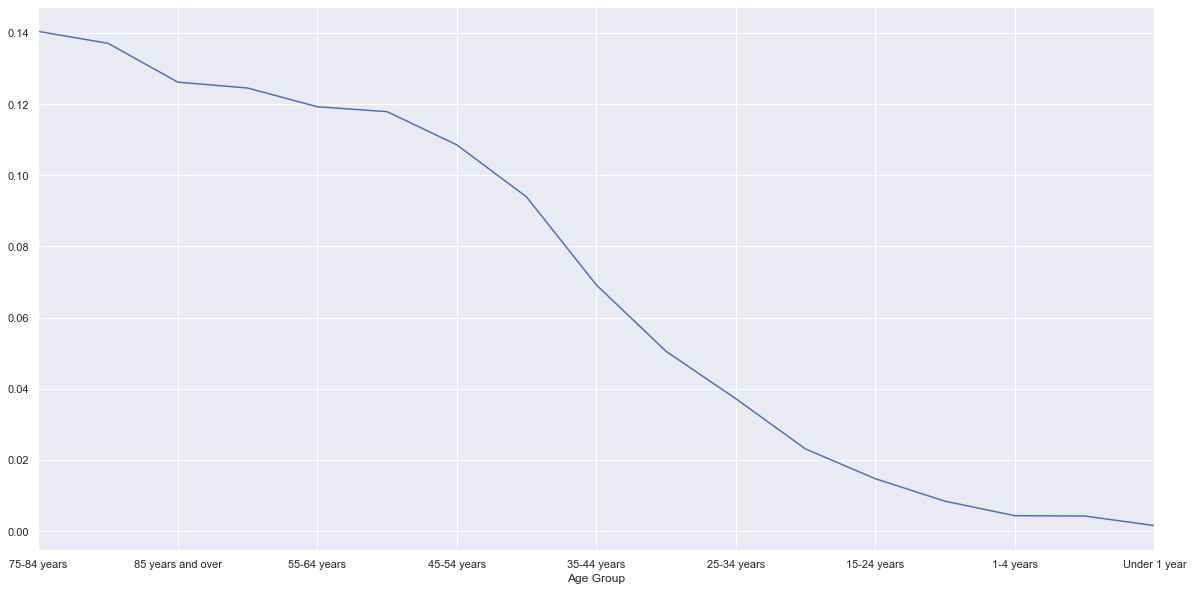


Figure 3. The ratio of COVID-19 deaths by the total deaths for each age group

From the above plot, the ratio of the COVID-19 deaths by total deaths are increased by the older age groups. Interestingly, the deaths of the age group of 75-84 years are more affected by COVI-19 than the age group of 85 years and over. The graph definitely shows that older age groups are more affected by COVID-19 spread than younger age groups. This shows that early vaccination distribution from the older generation was reasonable to prevent the possible deaths caused by COVID-19.

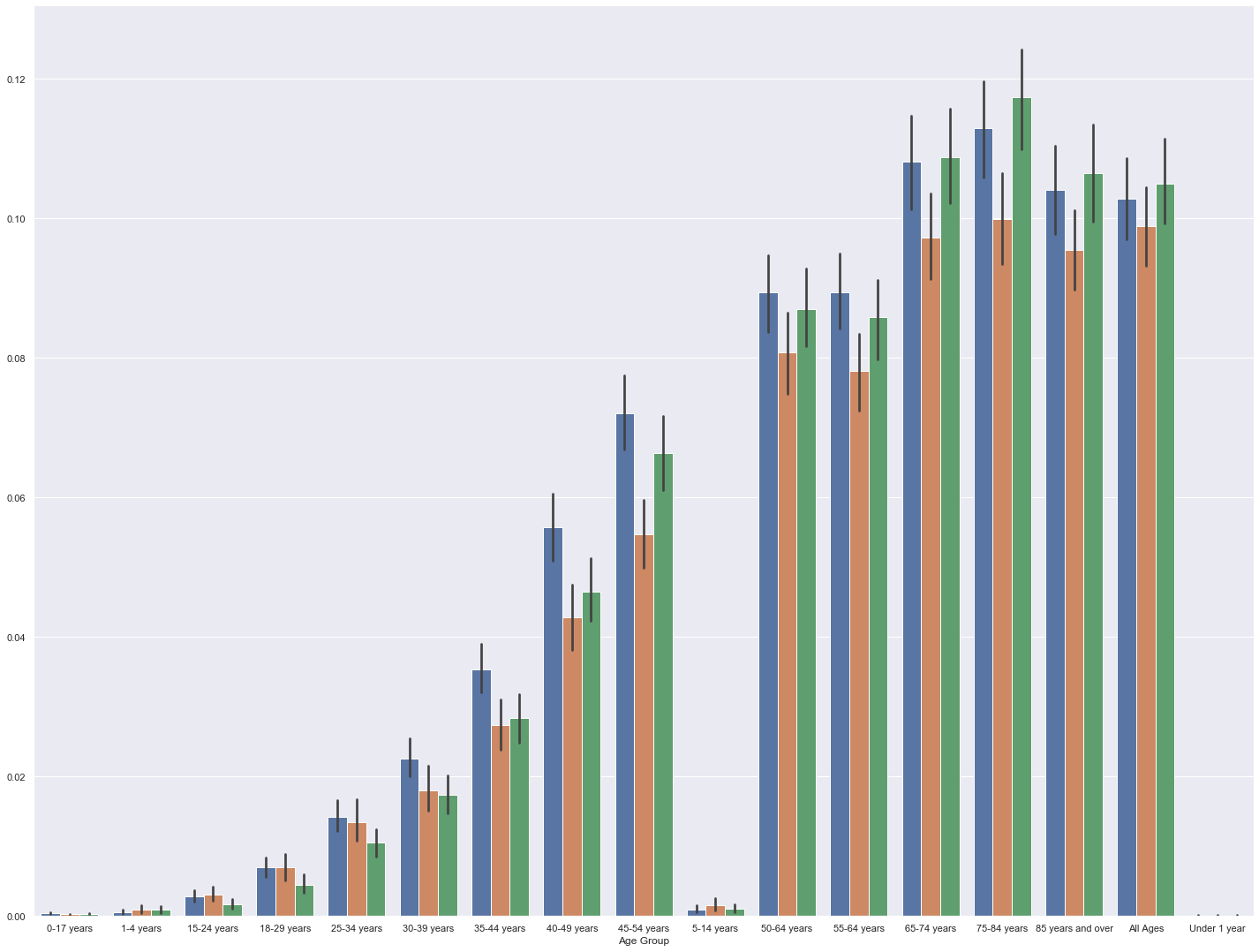


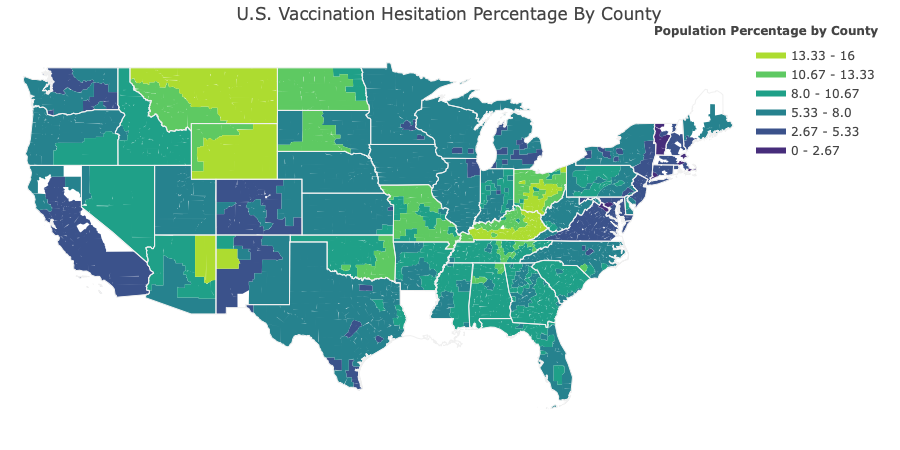
Figure 4. The ratio of COVID-19 deaths by the total deaths for each sex

From the above plot, the ratio of the COVID-19 deaths by total deaths are different by each gender. The COVID-19 deaths ratio of women is higher than men’s COVID-19 deaths ratio until age 39. But after age 40, th COVID-19 deaths ratio of women is lower than men’s COVID-19 deaths ratio and its gap increases by the age getting older. This histogram shows that over 40 year old men are more vulnerable to COVID-19 spread than women.

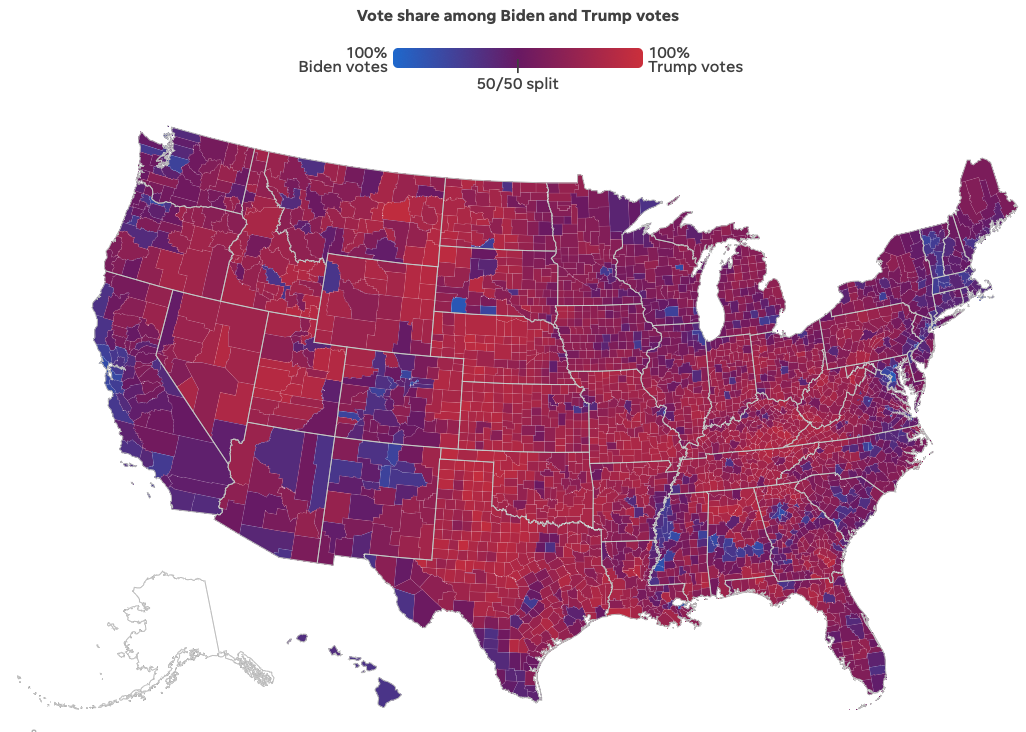
**--------------------------------------------------------------------------------------------------**

**Vaccination**

As it is perceived by the media there is a preconception about what we would expect to find if we look into the relationship between those areas with higher concentration of individuals who feel strongly hesitant about the COVID-19 vaccine and the areas corresponding to recent support for Donald Trump. We thought it was worth it to take a look at the data and see if it can tell us more.

******

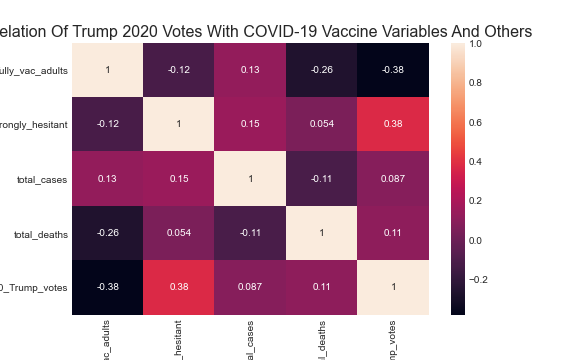
U.S. Map outlining each county per State color coded for percentage of strongly hesitant attitudes against the COVID-19 vaccine. The light green areas show counties with the highest concentration of individuals who reported feeling strongly hesitant.

******

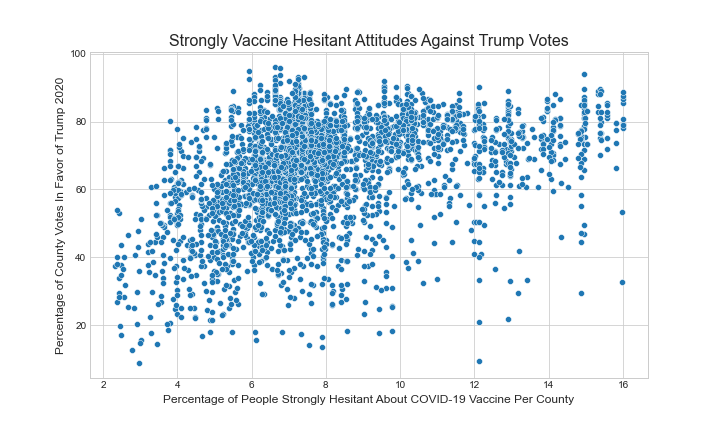
U.S. Map\* outlining each county per State color coded for percentage of vote sharing between Trump and Baiden during the 2020 Presidential Elections. The brightest red areas mark the counties where the vote percentage was dominated by Trump supporters.

\* from usatoday.com

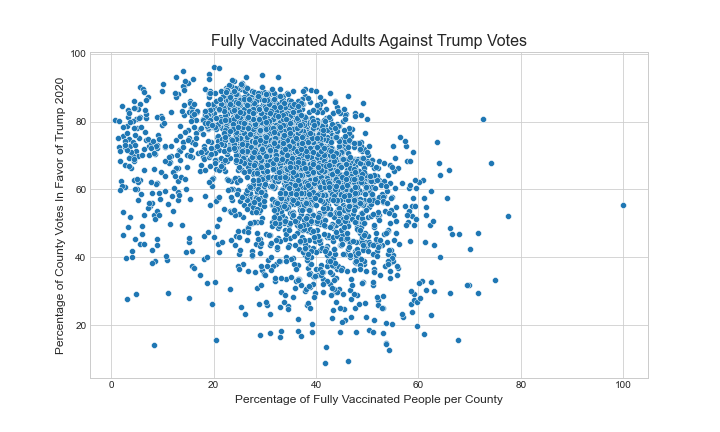
The illustrations above help us have a more complete image of the relationship we can hope to find in the data. It seems that the areas with high percentages of strong hesitation for the vaccine coincide to some degree with the areas with the highest percentage of support for Trump during our past election. In order to observe how strong these correlation is we make use of a correlation heatmap and scatter plots.

******

Heatmap showing the correlation between the variables of the COVID-19 vaccine data. The variables with the highest correlation are fully\_vac\_adults and strongly\_hesitant with 2020\_Trump\_votes.

******

Data distribution of the percentage of people with strongly hesitant attitudes and the percentage of people who voted for Trump in the most recent Presidential Election. The plot confirms that there is a positive linear relationship between the two variables.

******

Data distribution of the percentage of adults fully vaccinated and the percentage of people who voted for Trump in the most recent Presidential Election. The plot confirms that there is a weak negative linear relationship between the two variables.

### ***B.*** ***Checking Model Assumptions***

* **Hospitals**

Test the relationship between ratio death cases/beds and the ratio beds/pop. We failed to reject the F-test(p-value 0.18). The test is not fully reliable at this stage. We check the assumptions of normality and equal variances for the residuals.

Call:

aov(formula = ratio$Death\_bed ~ ratio$beds\_pop)

Terms:

ratio$beds\_pop Residuals

Sum of Squares 0.0701271 1.8605524

Deg. of Freedom 1 49

Residual standard error: 0.1948601

Estimated effects may be unbalanced

Df Sum Sq Mean Sq F value Pr(>F)

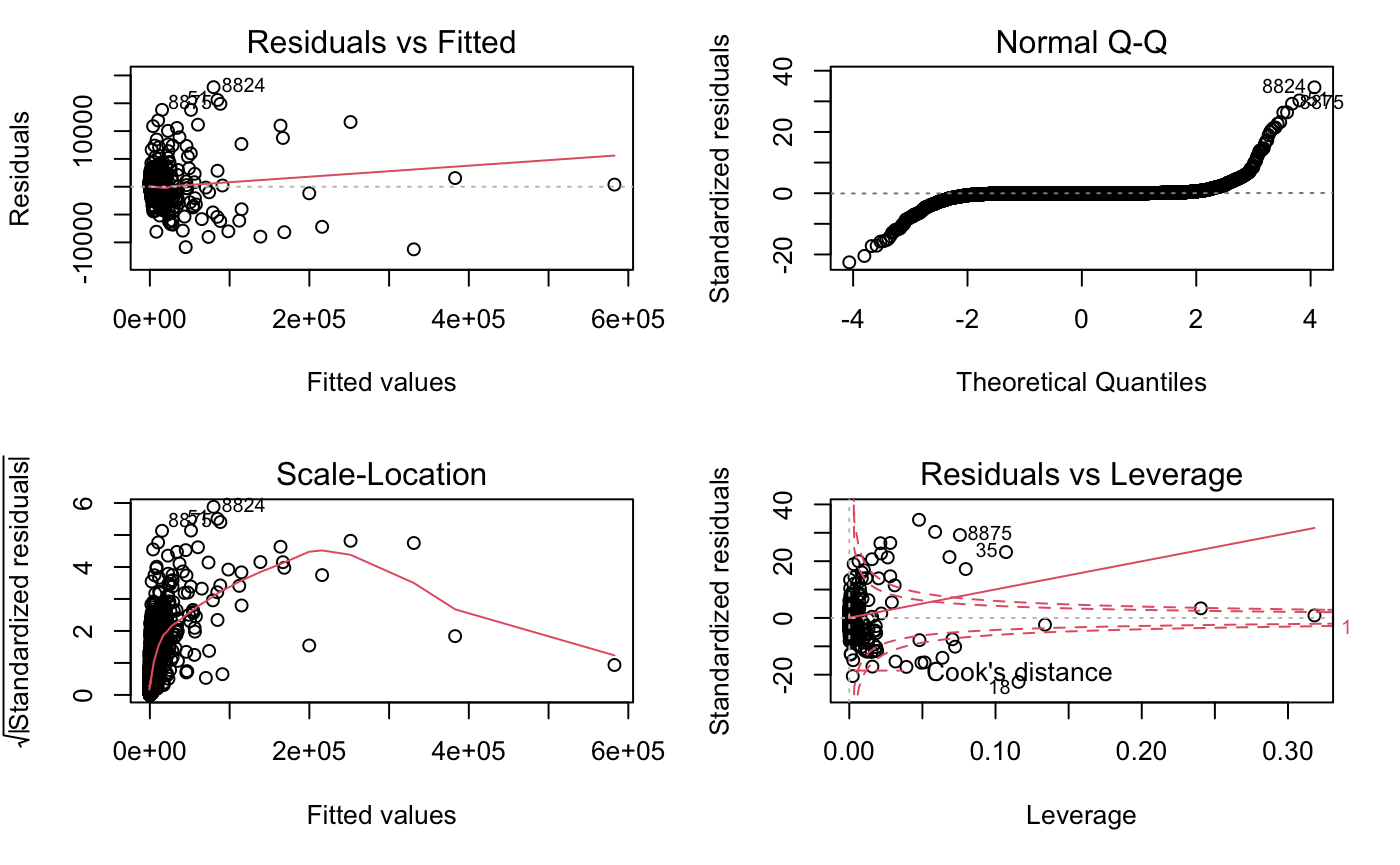
ratio$beds\_pop 1 0.0701 0.07013 1.847 0.18

Residuals 49 1.8606 0.03797

**--------------------------------------------------------------------------------------------------**

* **Deaths by sex and age group**

Since we are interested in the relationship between COVID-19 deaths data and other death causes including Influenza Deaths, Pneumonia, and total deaths. So the linear model was set up with dependent variables as COVID-19 deaths and independent variables as Influenza Deaths, Pneumonia, and total deaths.

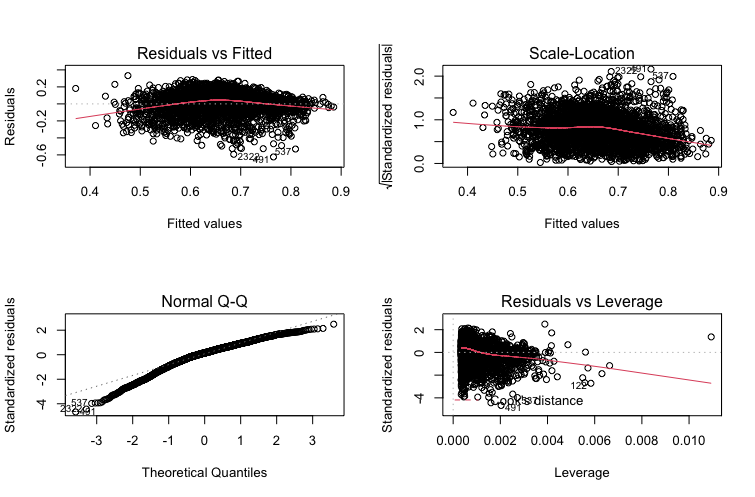
****

From the Residual vs Fitted plot we can see the red line is tilted a bit but not much and gives the implication that residuals and fitted are mostly linear relationships. In the QQ-plot, the most of the data is along the horizontal line and regarded as normal distribution. In the Scale-Location plot, the horizontal line is not straight at all and shows uneven distribution of values among standardized residuals. Residuals vs Leverage shows tilted horizontal lines. From the diagnostics, the model has homoscedasticity issues.

**--------------------------------------------------------------------------------------------------**

* **Vaccination**

Since we are interested in exploring a model for a linear relationship of these three variables there are certain assumptions we need to check first. In order to do this we build a linear regression model with 2020\_Trump\_votes as the dependent variable and the two variables with the highest correlation to this as the independent variables namely fully\_vac\_adults and strongly\_hesitant. The following are the plots obtained from this regression on the raw data.

******

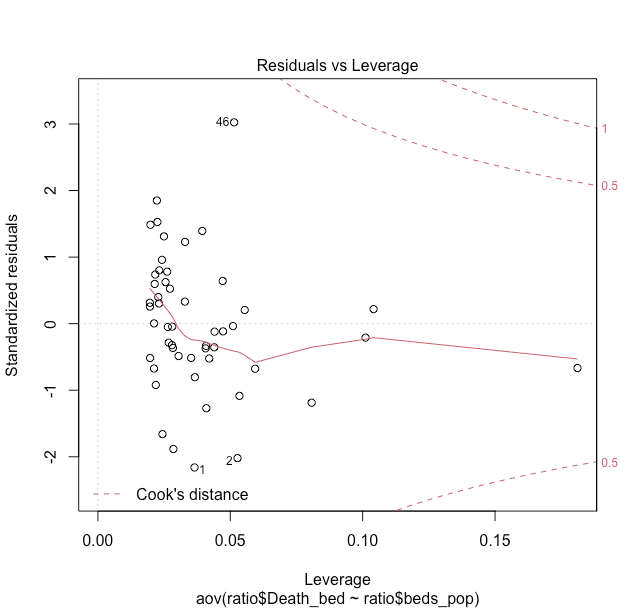
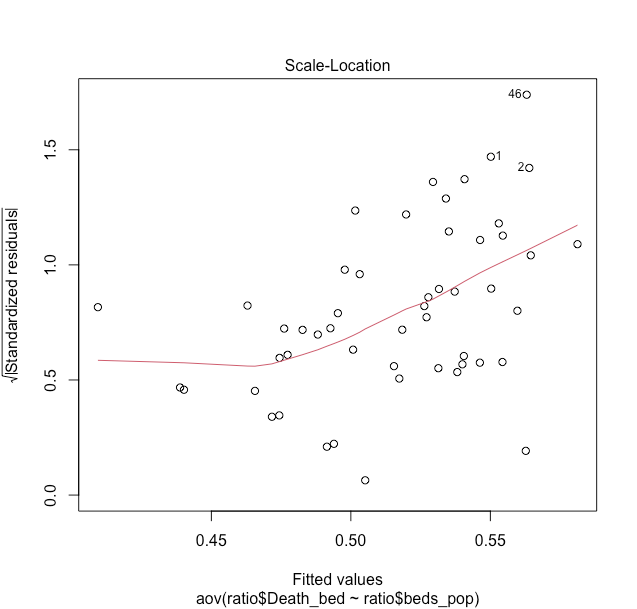
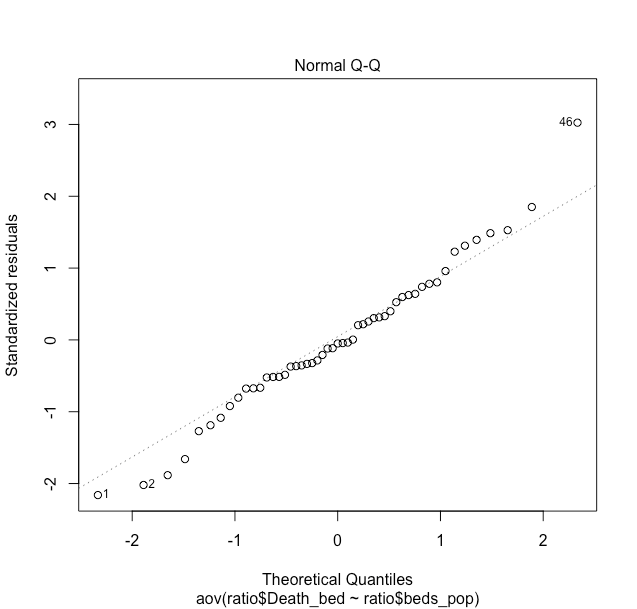
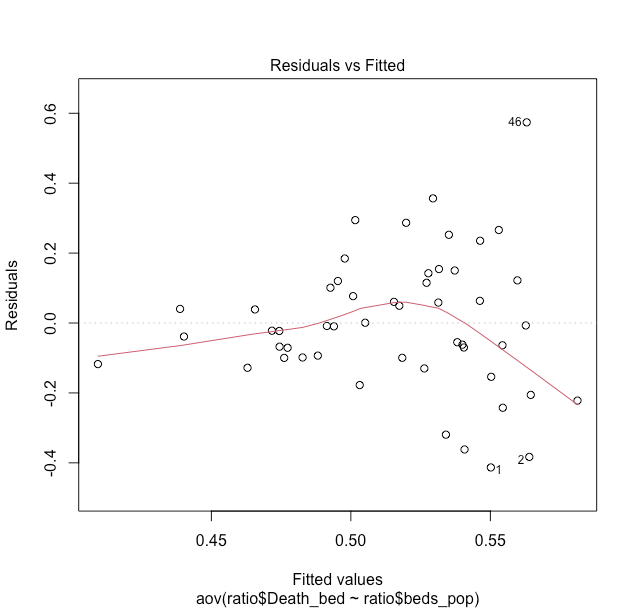
From the Residual vs Fitted plot we can see the red line remains horizontal and uniform around zero with a very small amount of distortion. This is an indication of a linear relationship. In the Normal Q-Q plot we have that the residuals follow a normal distribution to a reasonable extent. Perhaps after transformation and normalization the tail behavior of the residuals will improve. In the Scale-Location plot we can see a horizontal line reasonably straight with an even distribution of values around it which indicates a homogenous variance of the residuals. Based on this information we can continue to improve this model since the assumptions for linearity and homoscedasticity stand.

--------------------------------------------------------------------------------------------------

### ***C.*** ***Model Diagnostics***

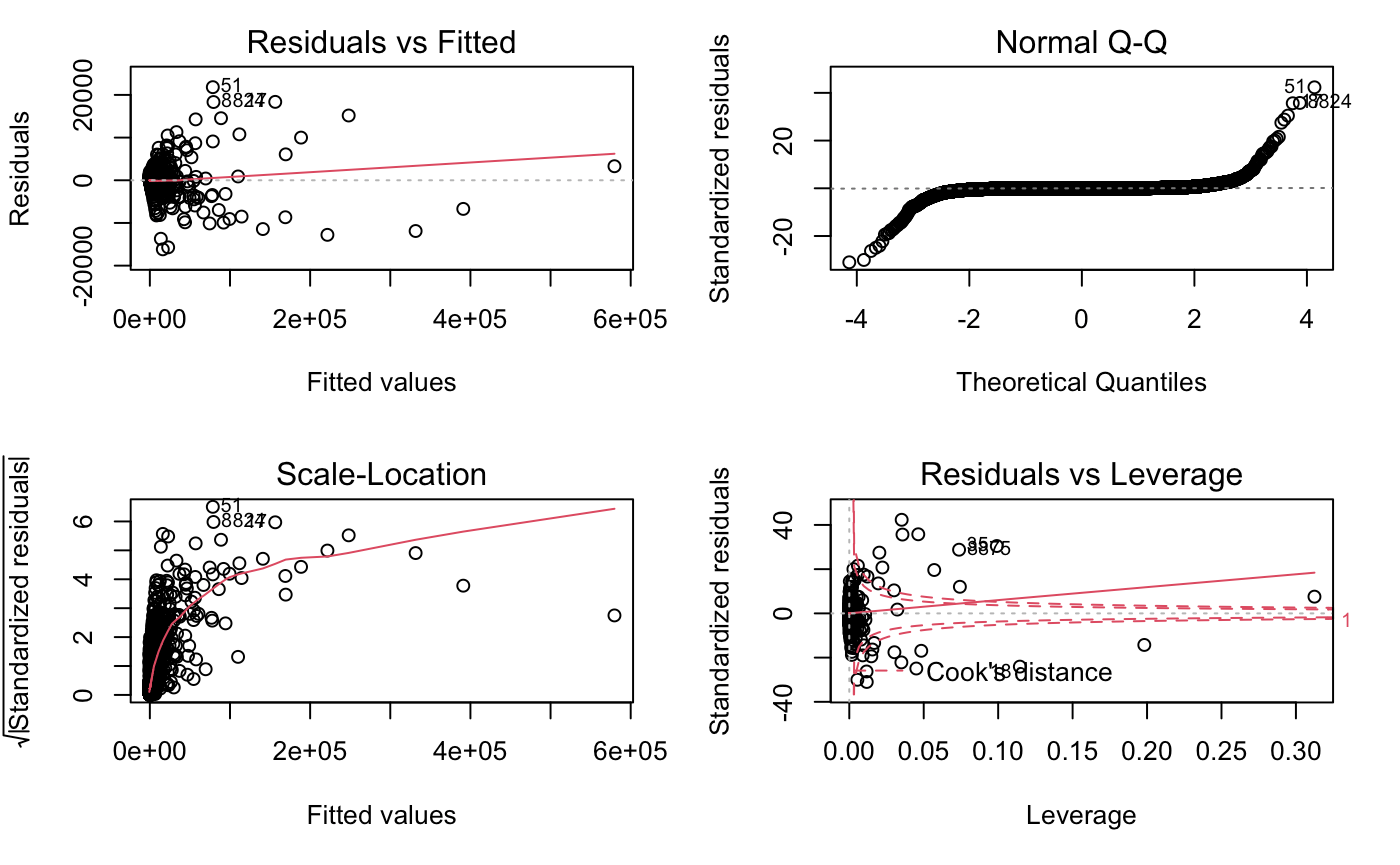
* **Hospitals**

After checking the assumptions, we can see that the relationship between ratio death cases/beds and the ratio beds/pop is not clear. However, some states with low capacity hospitals /population but they have a low ratio of death/beds. We can consider more the conditions of the state to figure out which states are vulnerable in controlling pandemics in the future or help the sceneries to the capacity of hospitals.

****

**--------------------------------------------------------------------------------------------------**

* **Deaths by sex and age group**

****

From this model we obtained significant estimates for beta1, beta2 and beta3 of -.012, 1.359 and -6.13 respectively.

From t-test for beta1 we obtained a t-statistic of -18.363 and a 2e-16 critical value. Therefore we can reject the null hypothesis that the estimate beta1 is zero at level alpha = 0.05.

From t-test for beta2 we obtained a t-statistic of 254.725 and a 2e-16 critical value. Therefore we can reject the null hypothesis that the estimate beta1 is zero at level alpha = 0.05.

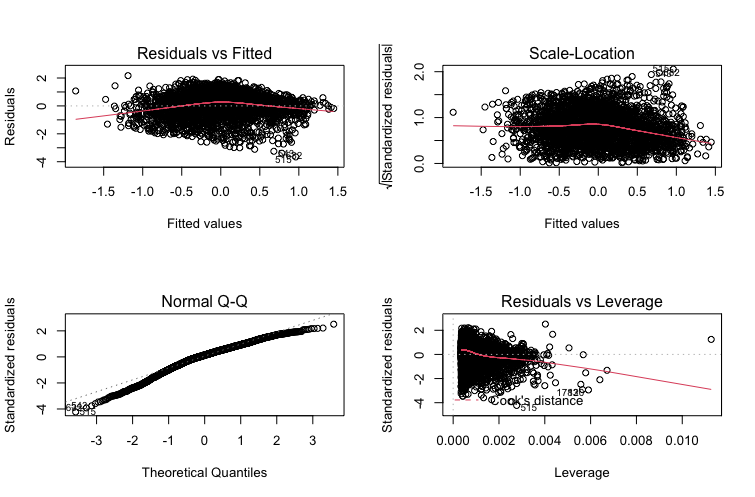
From t-test for beta3 we obtained a t-statistic of -76.984 and a 2e-16 critical value. Therefore we can reject the null hypothesis that the estimate beta1 is zero at level alpha = 0.05.

In addition, with a level alpha = 0.05 we find that the f-statistic is 1.32e+06 while the critical value is 2.2e-16. Therefore we can reject the null hypothesis that the estimates beta1 = beta2 = 0 at level alpha = 0.05 .

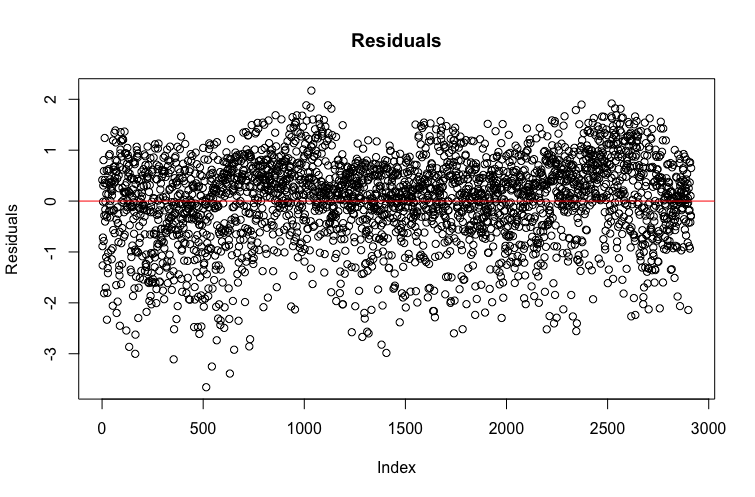
**--------------------------------------------------------------------------------------------------**

* **Vaccination**

Based on our preliminary analysis we proceeded to standardize the data and eliminate the outliers before implementing model diagnostics once more. The variables that provided the most significant and adequate linear model results were the initial model with clean data. The model consists of 2020\_Trump\_votes as the dependent variable and the two variables with the highest correlation to this as the independent variables are fully\_vac\_adults and strongly\_hesitant. The diagnostic plots are as follows:

****

We can notice small improvements to the Residual vs Fitted and scale-location plots as the residuals maintain an evenly distributed cluster. However the most significant improvement occurs in the Normal Q-Q plot where we can see an improvement in the distribution of the residuals. In order to complement these plots below is the plot of residuals and as you can see they are considerably well distributed around zero with a small skewness at the bottom.

****

From these model we obtained significant estimates for beta1 and beta2 of 0.3363 and -0.3383 respectively. We also have an estimated intercept of 2.874e-16 with no significance and a R-squared value of 0.2499 meaning that only 0.2499 of the variance can be explained with this model.

From t-test for beta1 we obtained a t-statistic of 20.85 and a 1.96 critical value. Therefore we can reject the null hypothesis that the estimate beta1 is zero at level alpha = 0.05.

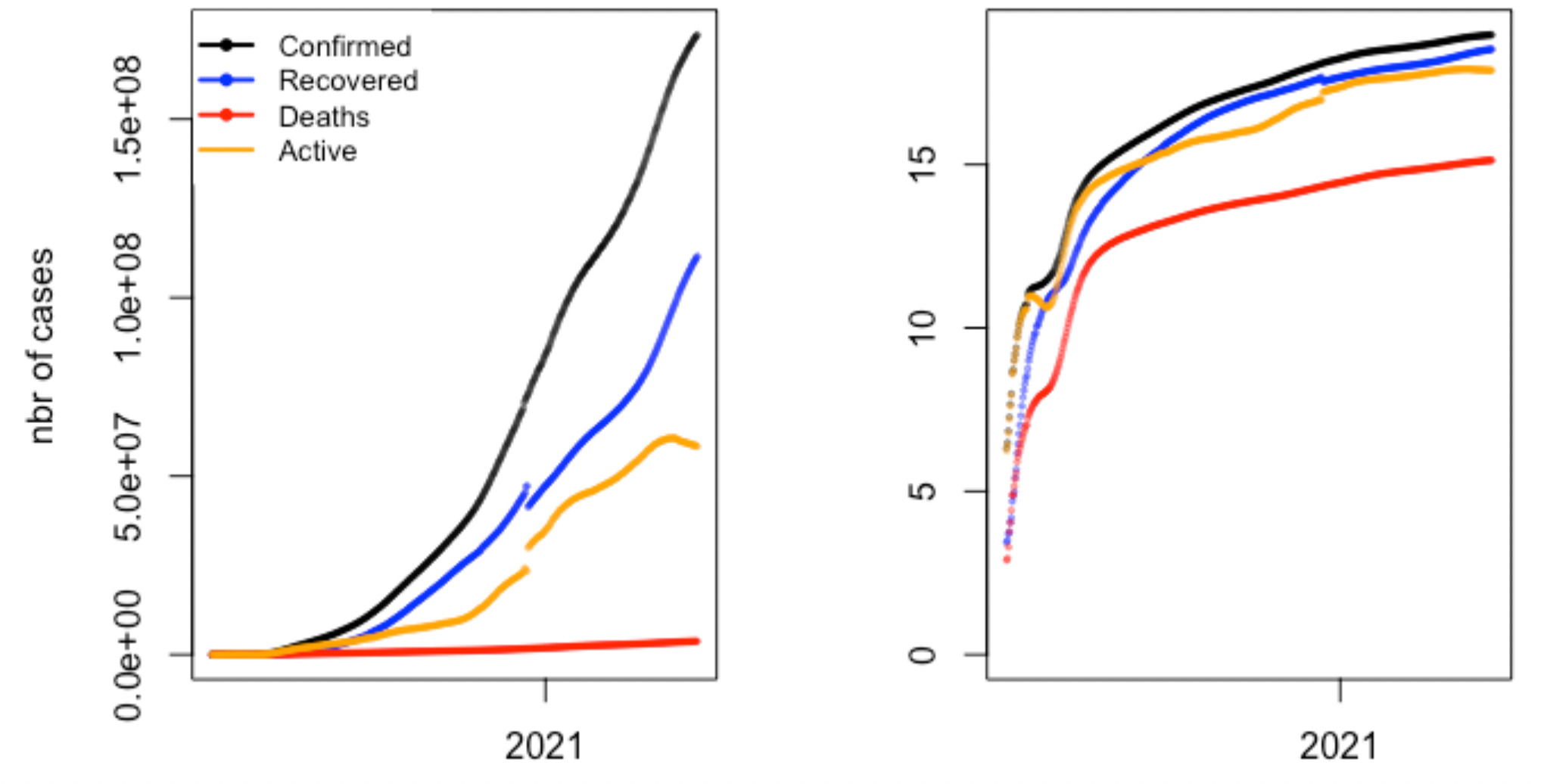
From t-test for beta2 we obtained a t-statistic of -20.97 and a 1.96 critical value. Therefore we can reject the null hypothesis that the estimate beta1 is zero at level alpha = 0.05.

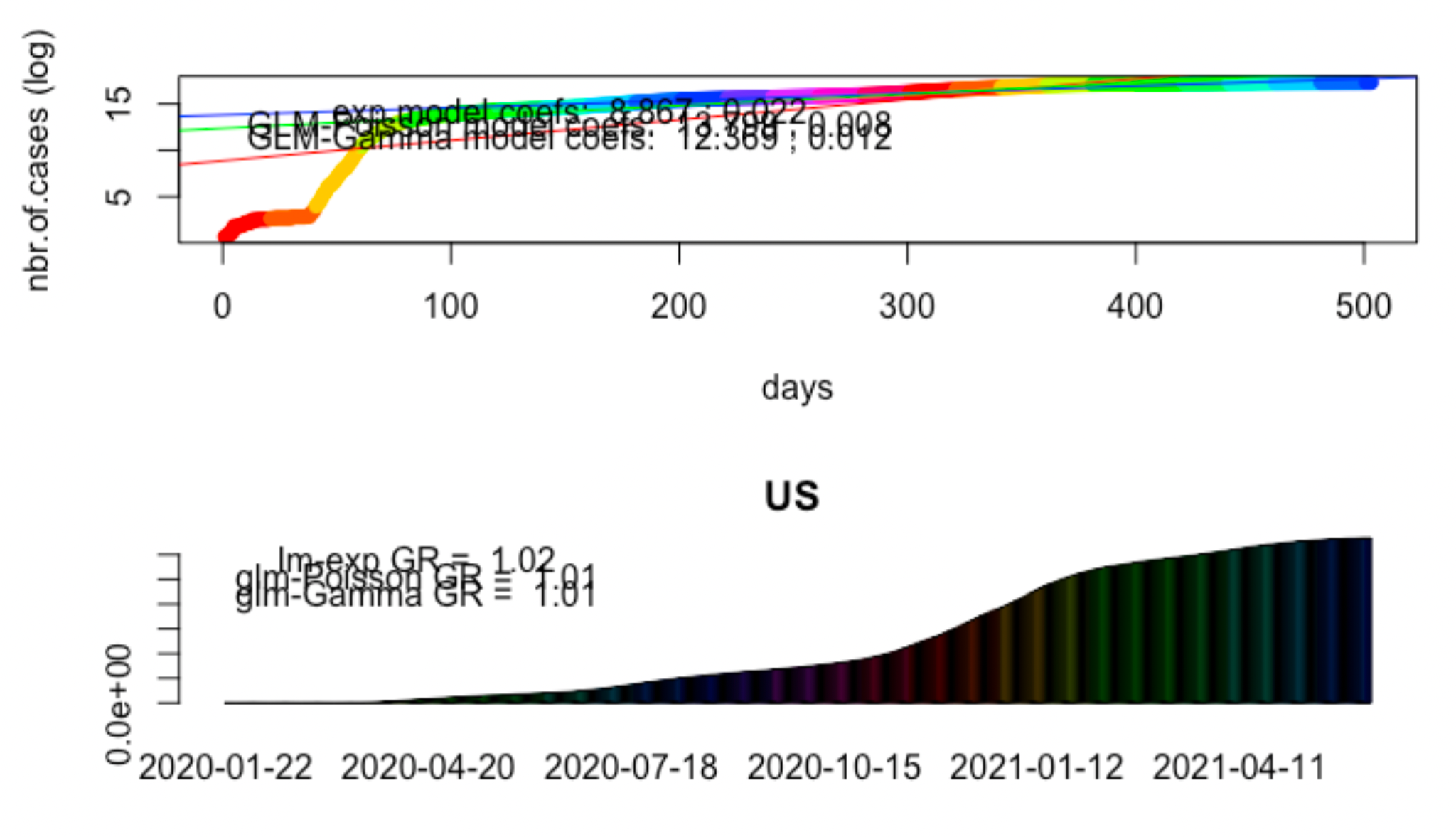
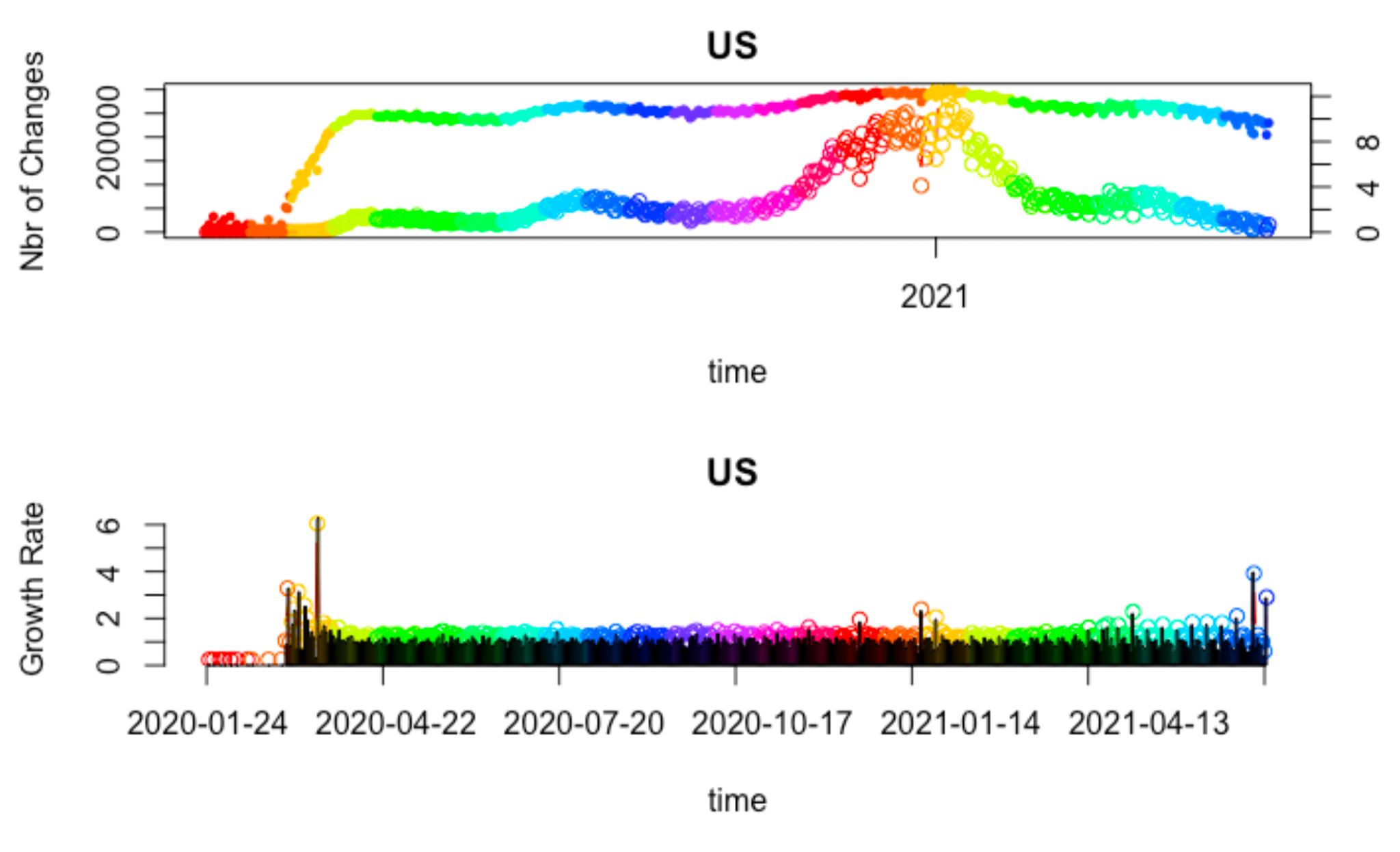
In addition, with a level alpha = 0.05 we find that the f-statistic is 727.86 while the critical value is 4.50. Therefore we can reject the null hypothesis that the estimates beta1 = beta2 = 0 at level alpha = 0.05 .

From the 95% confidence intervals for our estimates we have that only the interval for our intercept contains zero while the interval for beta1 = (0.3047, 0.3679) and beta2 = (-0.3699, -0.3067)

**--------------------------------------------------------------------------------------------------**

**General**

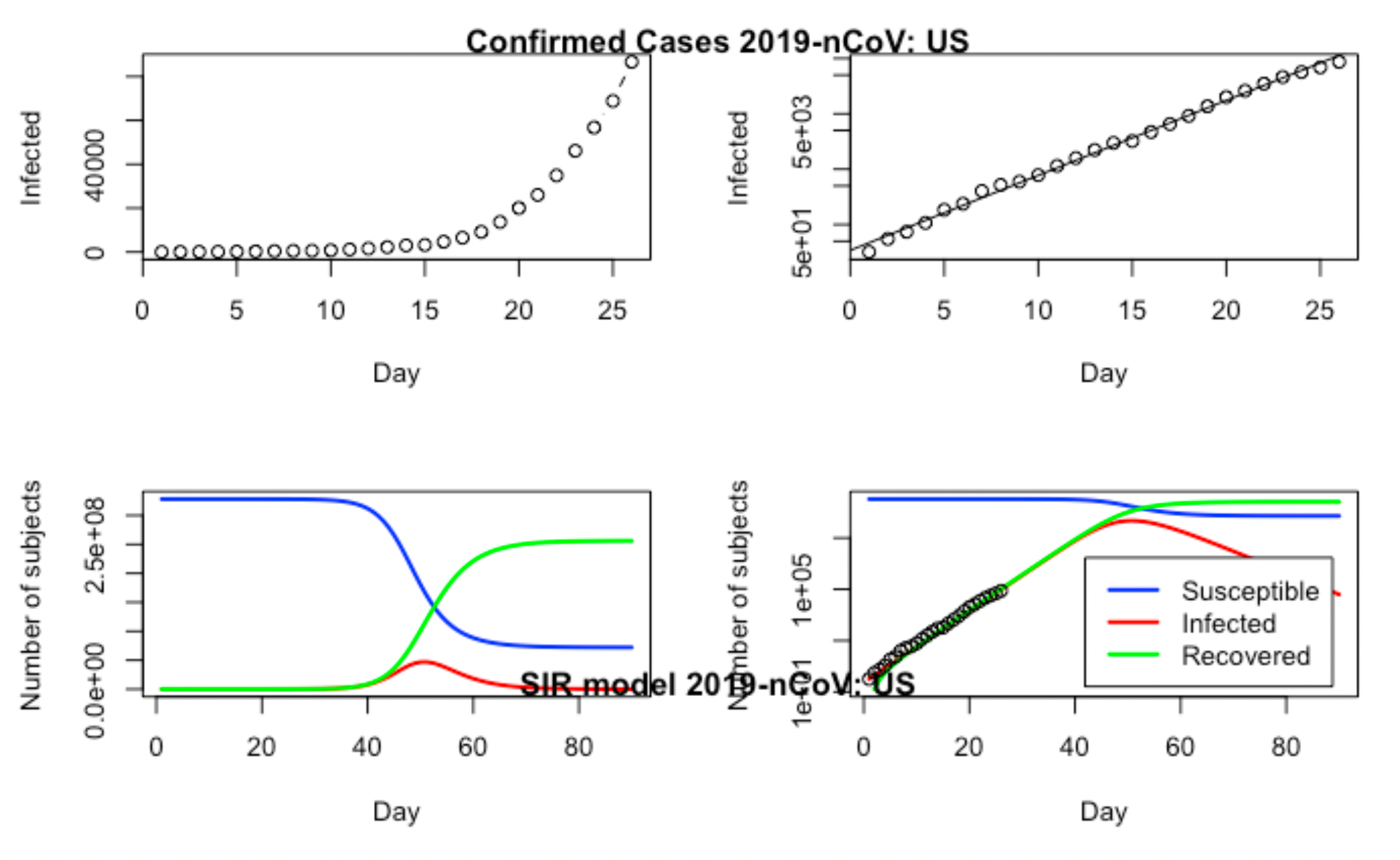




Using also R package “Shiny”

Simulating the Virus spread in the US on the basis of time series data for confirmed cases.

The time series model’ assumption predicts cases will continue to rise in the US for the remaining 2021.



Model Diagnostics show the same result.

**IV. *Conclusion & Discussion***

Through our analysis we were able to find interesting things regarding COVID-19 and its deaths and political perspective. And also, our exploration of the data in question revealed relationships between COVID-19 and other variables which suggested the existence of an overall picture to be explored.

Throughout confirmed cases, death, hospitalization by each state, there is a positive correlation between Confirmed and Death case per population. Also, there is a linear negative relationship between the ratio beds/populations and the ratio between death COVID cases/beds of hospitals in each state in the US.

The ratio of COVID-19 deaths by the total deaths reveals that the early vaccination allocation from elderly people and elderly men are the most vulnerable to COVI-19. From the plot, correlation and linear regression modeling shows Pneumonia, influenza and COVID-19 share mostly the same features which shows seasonal in spring and winter season. COVID-19 is more seasonal than Pneumonia, influenza. The perfect correlation of total deaths and COVID-19 deaths implies that recent rapid increases of the number of total deaths are the cause of COVID-19 .

Even though we saw a relationship between vaccine hesitation and political inclination the data was not significant enough to draw more precise conclusions. We know there is a correlation between these two variables through the examination of the data and through what we perceive from the media. However we would need more information and examination to be able to have a concrete image of the relationship.

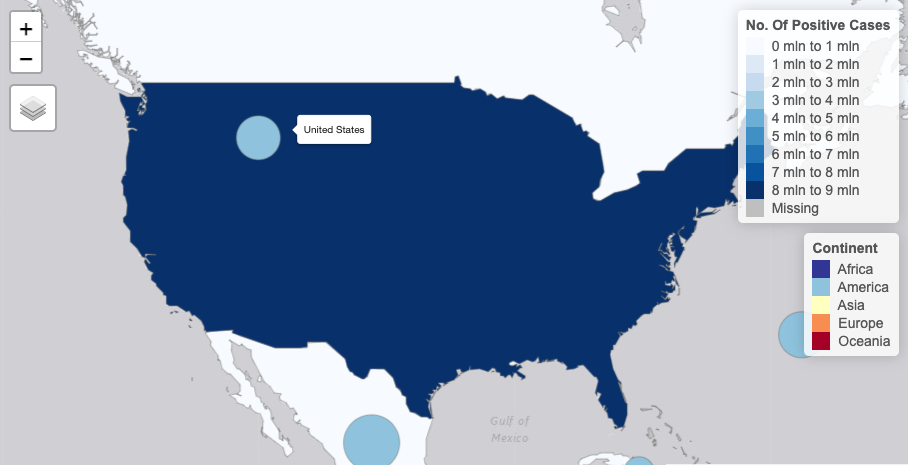
In conclusion, COVID-19 greatly affects the increase of the total number of deaths. In such that, the more beds per population and vaccination allocation from elderly people are great measures to prevent possible deaths caused by COVID-19. To promote vaccination efforts, the relationship between vaccine hesitation and political inclination is not certain. So, we suggest vaccination advantages such as lottery.

**V. Appendix**

[STA 160 - Relationship between COVID cases and Capacity of hospitals of States in US.pdf](https://drive.google.com/file/d/1TjYABeSlwsBmM0Ngv5MjRpXvMk8Gnoen/view?usp=sharing)

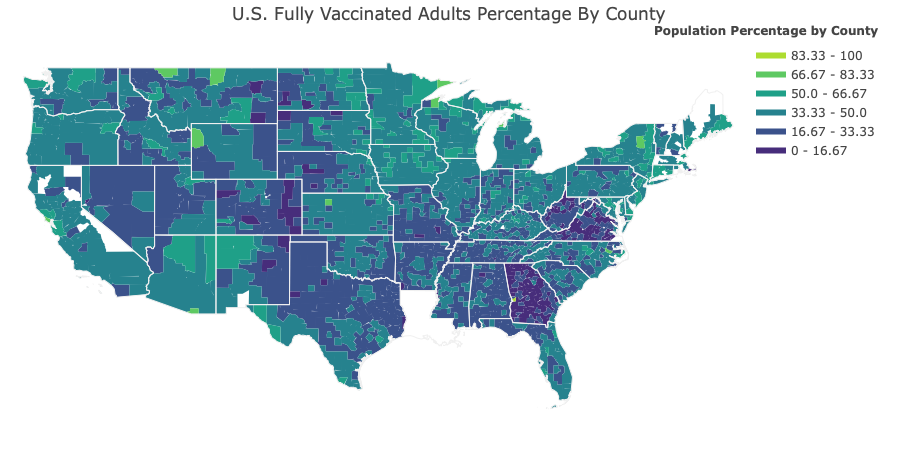
[Vaccine\_visualizations.pdf](https://drive.google.com/file/d/1ad-eYpE4JIMkDw-mPvzzM2Pxs-T36TPC/view?usp=sharing)

[STA160\_Final\_deaths.pdf](https://drive.google.com/file/d/1PlSVqpvzkSe4CNovSVVzXtoYTHBsTHFx/view?usp=sharing)



Interactive world map of real time positive cases

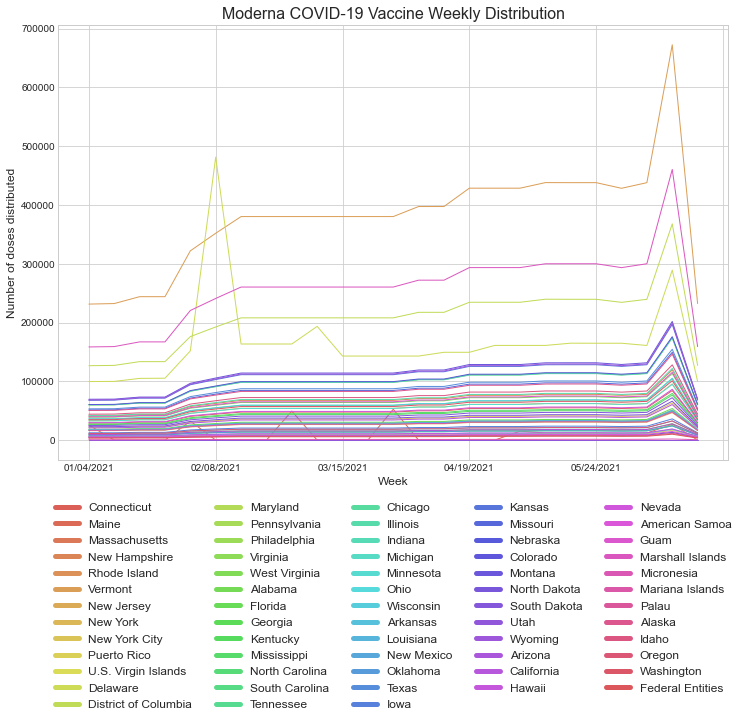
**--------------------------------------------------------------------------------------------------**

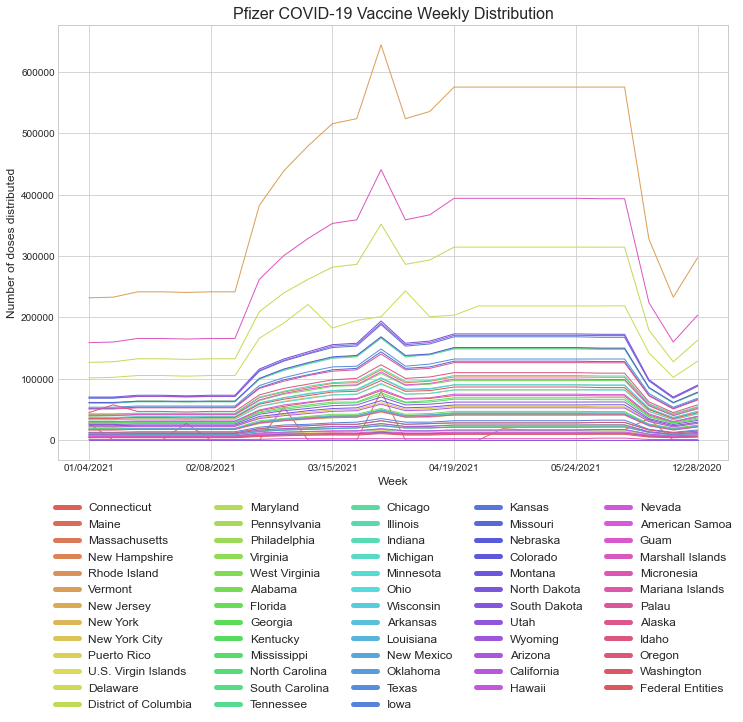


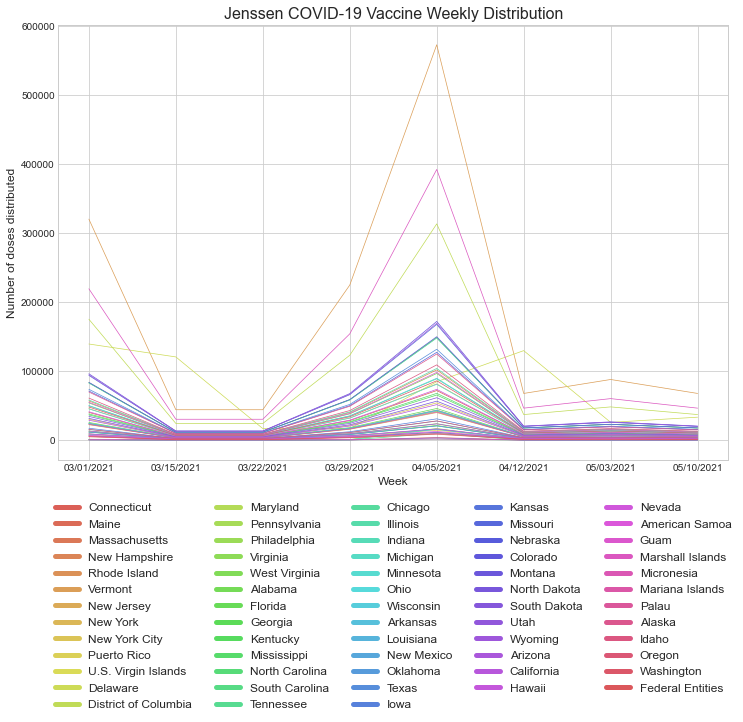
U.S. Map outlining each county per State color coded for percentage of fully vaccinated people per county population. The light green areas show counties with the lowest concentration of individuals who have been vaccinated.

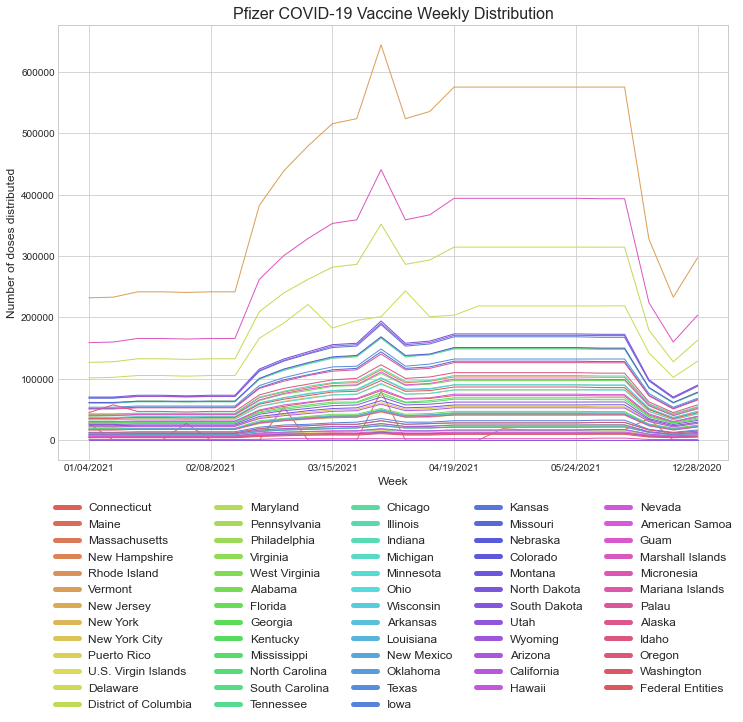
**--------------------------------------------------------------------------------------------------**

The following three plots illustrate the even distribution of each COVID-19 vaccine across States. The legend is provided after the third plot. The takeaway from these plots is that the allocation of vaccines is proportional to the state size









**--------------------------------------------------------------------------------------------------**

**VI. References (Bibliographic)**

* <https://rpubs.com/benthecoder/covid-19-slidy>
* <https://rviews.rstudio.com/2020/08/10/us-covid-19-forecasts/>
* <https://mine-cetinkaya-rundel.github.io/covid19-r/>
* <https://github.com/joachim-gassen/tidycovid19>
* <https://cran.r-project.org/web/packages/COVID19/index.html>
* <https://youtu.be/CYwA8lvYQx8https://youtu.be/CYwA8lvYQx8>
* Vaccination and case data: <https://github.com/owid/covid-19-data>
* Vaccination hesitancy data: <https://data.cdc.gov/Vaccinations/COVID-19-County-Hesitancy/c4bi-8ytd>
* Jenssen COVID-19 vaccine allocation data: <https://data.cdc.gov/Vaccinations/Fully-Vaccinated-Adults/jm79-dz78>
* <https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/saz5-9hgg>
* Moderna COVID-19 vaccine allocation data: <https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/b7pe-5nws>
* Pfizer COVID-19 vaccine allocation data: <https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/w9zu-fywh>
* Vaccine distribution update: <https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-authorizes-pfizer-biontech-covid-19-vaccine-emergency-use>
* Article as a response to COVID-19 vaccine hesitancy: <https://www.healthaffairs.org/do/10.1377/hblog20200814.996612/full/>
* Safety related subtopics which is worth studying: <https://www.cdc.gov/coronavirus/2019-ncov/vaccines/recommendations/pregnancy.html>
* On effectiveness of COVID-19 vaccine: <https://www.cdc.gov/coronavirus/2019-ncov/vaccines/effectiveness/why-measure-effectiveness/breakthrough-cases.html>
* On taking steps back: <https://www.nytimes.com/live/2021/06/04/world/covid-vaccine-coronavirus-mask>