**Title**: Study on the Surge of Covid-19 in All State Data with Data Visualization Methodologies

**Group Members**

# Name 1: Rengalakshmi Balasubramanian

# Name 2: Quang Dao Tran (Jimmy Chan)

# Name 3: Huu Lam Nguyen (Hubert Nguyen)

**I. Introduction:**

Coronavirus pandemics have affected all the countries around the globe and the US has emerged as a global hotspot. According to the Niko Kommenda (2020), the US has recorded the highest confirmed Covid-19 cases and deaths globally. Currently, the number of new Covid-19 cases has continued to increase in the recent months. A detailed study on the cases, death and its trend are discussed in this report.

**II. Background and Motivation**

- The Covid-19 disease has been one of the top concerns in the U.S this year. It is essential to get updated about the latest details regarding the pandemic, such as the number of cases and deaths as well as the increase in these figures.

- Many people might be not aware of the effects of the virus in their locations. The project aims to clarify which states are witnessing the most rapid spread of the virus, which states seem to achieve a way to deal with it.

- In the winter of 2020, many experts predict that there will be a surge in the number of positive cases. The project intends to evaluate this opinion.

**III. Research Questions**

***First question:*** ***How have the U.S gone through the pandemic in 2020?***

*Solution:* Achieve the monthly Covid-19 data related to the U.S (the whole country) in 2020 (up to present). Use charts or graphs to illustrate the modification and trend in the cumulative number of cases and deaths, and calculate the percentage of people who have recovered from Covid-19 as well as the percentage of deaths in the total number of cases.

***Second question:*** ***Where are the centers of the pandemic in the U.S at present?***

*Solution:* Use charts to illustrate the numbers of cases and death in each jurisdiction of the U.S at present to conclude with the 5 most hard-hit states by Covid-19. Also attempt to demonstrate the way these states have been dealing with the pandemic by including data related to the number of people who have been tested for Covid-19 in each of the states.

***Third question: Is there going to be a surge in the number of positive cases in U.S states?***

*Solution:* Applying the time series forecasting techniques to forecast the number of positive cases and deaths of the U.S at the end of 2020.

**IV. Dataset**

The dataset is obtained from The COVID Tracking Project website, which is a volunteer organization launched from The Atlantic and dedicated to collecting and publishing the data required to understand the COVID-19 outbreak in the United States. In this dataset, the COVID19 situation in the 56 jurisdictions are tracked. The URL to the website is listed below: <https://covidtracking.com/data>. The data of each state which is updated daily is extracted from the website and combined altogether to make a whole dataset with the help of data extracting techniques in R studio. The dataset used in the report has a range of time, starting from April 1st to December 14th, 2020. A summary of the dataset is provided below. The dataset has a total number of observations of 14448 with 15 variables.

|  |  |  |
| --- | --- | --- |
| **Variables** | **Data type** | **Definition** |
| States | Factor | Name of the state or territory |
| Date | Date | Date of report |
| Totalcases | Number | Cumulative number of confirmed plus probable cases of COVID-19 |
| Newcases | Number | Number of new cases reported in the date |
| Confirmedcases | Number | Cumulative number of confirmed cases of COVID-19 |
| Probablecases | Number | Cumulative number of probable cases of COVID-19 |
| TotalPCR | Number | Cumulative number of people tested via PCR testing |
| TotalABT\_p | Number | Cumulative number of completed antibody tests |
| TotalABT\_s | Number | Cumulative number of unique people who have been tested at least once via antibody testing |
| Totaltests | Number | Cumulative number of TotalPCR, TotalABT\_p and TotalABT\_s |
| Recovered | Number | Cumulative number of people that are identified as recovered from COVID-19. |
| Totaldeath | Number | Cumulative fatalities with confirmed or probable COVID-19 case diagnosis |
| Newdeath | Number | Number of new deaths reported in the date |
| Probabledeath | Number | Cumulative fatalities with probable COVID-19 case diagnosis |
| Confirmeddeath | Number | Cumulative fatalities with confirmed COVID-19 case diagnosis |

**V. ETL**

**1. Extracting the data**

The link to the data of each jurisdiction is obtained by adding the name of the state or territory at the end of a base URL: <https://covidtracking.com/data/state/>. This data is divided into seven main categories: Cases, Viral (PCR) tests, Antibody tests, Hospitalization, Hospitalization (HHS data), Outcomes and Long-term care (LTC). To achieve the data that is exclusive to a single category, the name of the category is also required. For example, the cumulative number of confirmed plus probable cases of COVID-19, which belongs to the Cases category, of Alaska can be extracted from the URL: <https://covidtracking.com/data/state/alaska/cases>. Afterwards, the data extracted from each category is combined altogether to make the data of a state or territory. The report only uses the data extracted from three categories: Cases, Viral (PCR) tests, Antibody tests and Outcomes.

Apart from the variables which are obtained from the data source, there are two new variables created manually in the used dataset. The first one is States and the second one is Totaltests. The reason for creating the second one is that there is a difference in the method of reporting the tested people among jurisdictions. Some states decided to report using the TotalPCR only while others chose TotalABT\_p and TotalABT\_s. This may be caused by the inadequate understanding of the related metrics. Therefore, a new variable is added to make it less challenging to compare among states and territories in terms of the number of people who are tested for Covid-19. In the final step, data from each jurisdiction is linked to complete the dataset for the whole country.

**2. Transforming the data**

After completing the whole dataset, it is important to understand about the structure of the dataset and make necessary adjustments to the variables. The data types for all variables are correct, apart from the States, which has a character data type originally. It is a categorical variable; therefore, it is wise to change its data type into factor.

Another important aspect is how to deal with null values. There are a lot of missing values in the dataset. However, if they are removed, a lot of data points would be deleted as well. In addition, in the data source, null values are considered as 0 and the report also applies this method.

**3. Loading the data**

In the end, the technically correct and consistent data is saved as a csv file, which is used later for the visualization using Tableau.

**VI. Design of visualization and implementation**

The project uses three dashboards to answer the research questions, respectively.

**First dashboard:** **COVID in US 2020: The Whole Country**

There are three graphs presented in the dashboard. The first bar chart is used to illustrate the cumulative number of positive cases, deaths, and recoveries in the U.S through months of 2020, while the second one projects the number of new cases and deaths for each month of 2020. For both these charts, the length of the bars demonstrates the number of cases while the color informs about the number of deaths. The only line chart in the dashboard helps to illustrate the trend in the percentage of cumulative number of recoveries and deaths based on the cumulative number of cases through months. There are two lines in the chart, the red one represents the percentage of deaths while the other represents the figure for recoveries.

**Implementation:** Tooltips are used in the two bar charts to bring out the data related to a single month. Moreover, both bar graphs are used as filters to link all the graphs in the dashboard and help viewers only focus on one specific month by clicking on one bar.

**Second dashboard:** **COVID Map of USA: State by State Report:**

The map in the dashboard provides information about the cumulative cases, deaths, and the population of each jurisdiction in the U.S at present. The area of the circle at the center of each jurisdiction illustrates the number of cases within the territory while the color of it presents the number of deaths. Also, the 2018 population of the jurisdiction is shown by the color of the territory area. The bubble graph illustrates the number of new cases and new deaths in the 5 most hard-hit jurisdictions of the U.S in December only (from December 1 to December 14), including: California, New York, Texas, Illinois, and Florida. The area of the circle related to each state represents its number of new cases in December, and the number of new deaths is also added as an extra detail. The only bar chart in the dashboard illustrates the cumulative number of people who have been tested for Covid-19 in these five mentioned states through months. Color is applied to represent the states and the length of bars is used to demonstrate the number of tested people.

**Implementation:** Tooltips are used for all the graphs to bring out the data related to a specific jurisdiction quickly. Furthermore, the first two figures are used as filters to link all the graphs in the dashboard and help viewers only focus on one specific state or territory by clicking on the appropriate circle or bubble.

**Third dashboard:** **Trends of Surging COVID-19**:

The only bar chart in this dashboard provides information about the number of new cases and new deaths in each state or territory of the U.S on only December 14, 2020. The length of bars represents the number of new cases, while the color of them demonstrates the new death. Data of the whole U.S is also annotated in the area of the chart. Then, the two-line charts illustrate the trend in the number of cases and deaths for the whole U.S through days and forecast this trend in the remaining weeks of December. Colors are used to differentiate between these two categories: cases and deaths. Several data points are labelled in the last line chart to emphasize the new high peaks in the number of cases and deaths due to the forecast.

**Implementation:** Tooltips are used for all the charts in this dashboard to bring out the data related to a specific state or date quickly and responsively.

**\* Applied design principles:**

- Appropriate chart types are chosen for all the charts. Bar charts and line charts are used only when there is information about time included. Also, the map is only applied to the states’ data with geographical features (population).

- Useful design elements are added to help compare among states and features. There are only three main mark types applied in the visualization: length, size, and color.

- Incorporate helpful elements to increase both clarity and aesthetics. Well-placed and aligned titles are added for each of the graphs. Details in font selection are also paid attention to. For some graphs, order is applied to make these graphs more beautiful. Some labels and annotations are also created to make the graph more informative.

**VII. Results and Discussion**

The report has succeeded in answering all the three research questions. Regarding the first question, it is apparent that the number of COVID-19 cases has been increasing rapidly in the U.S through months of 2020, especially for the last two months November and December, the number of new cases arising has considerably outnumbered the figures of previous months. However, the U.S has witnessed a rise in the percentage of recoveries in the total number of people who have been infected while the percentage of deaths tends to fall, which is a good signal.

In terms of the second question, California, New York, Texas, Florida, and Illinois are the 5 most hard-hit jurisdictions in the U.S. California is the top-most state that is reporting the biggest number of positive cases while most of the deaths related to the pandemic in the country are confirmed in New York. These five states are also the ones which have the most populations in the U.S and the governments of states have been struggling to prevent the spread of COVID-19. One of their efforts is illustrated by the surge in the number of people receiving tests (both Antibody and PCR) through months.

Finally, there is a high probability that the number of positive cases and deaths in the U.S will continue to grow in the remaining days of 2020. The report provides two-line graphs, one for forecasting the number of new cases and deaths for each day in the next two weeks (until December 28) and the other for forecasting the cumulative number of positive cases and deaths. As on December 14, the U.S has reported more than 16.3 million people who have been infected and 292 thousand deaths. By December 28, the cumulative COVID-19 cases figure is likely to go up to 19.6 million and the number for deaths is more than 329.5 thousand.

**VIII. Conclusion**

Overall, the project provides thorough and meaningful answers for all the research questions. As clearly demonstrated in the project, COVID-19 is still the one of the urgent concerns in the U.S at present, and in the following days when the figures of cases and deaths will reach new high peaks (Adeline et al., 2020). One remaining shortcoming is that there is no insight about the hospitalization regarding COVID-19 in all jurisdictions included (Buerhaus et al., 2020). The reason for this drawback is lack of relevant states’ data, most states such as California have not reported this feature.

**Links for the public visualization:**

Rengalakshmi Balasubramanian: <https://public.tableau.com/views/realtimecovid/COVID-19intheU_S2020?:language=en&:display_count=y&publish=yes&:origin=viz_share_link>

**References:**

1. Niko Kommenda, Pablo Gutiérrez and Juweek Adolphe. 2020. Coronavirus map of the US: latest cases state by state. Available on https://www.theguardian.com/world/ng-interactive/2020/sep/21/coronavirus-covid-19-map-us-cases-deaths-state-by-state
2. Coronavirus disease (COVID-19). Available on https://apps.who.int/iris/bitstream/handle/10665/333905/nCoV-weekly-sitrep1-eng.pdf
3. Alkuzweny M, Raj A, Mehta S. Preparing for a COVID-19 surge. Eclinicalmedicine, 04 Jun 2020, 23:100400. DOI: 10.1016/j.eclinm.2020.100400
4. Peter I. Buerhaus; David I. Auerbach; Douglas O. Staiger. Older Clinicians and the Surge in Novel Coronavirus Disease 2019 (COVID-19). JAMA. 2020;323(18):1777-1778. doi:10.1001/jama.2020.4978
5. Stephanie Adeline, Connie Hanzhang Jin, Alyson Hurt, Thomas Wilburn, Daniel Wood, Ruth Talbot. 2020. Coronavirus Is Surging: How Severe Is Your State's Outbreak? Available on https://www.npr.org/sections/health-shots/2020/09/01/816707182/map-tracking-the-spread-of-the-coronavirus-in-the-u-s