Data Handling And Visualization

-Vansh. Gandhi

Table Of Contents

[Abstract 1](#_Toc80666522)

[ETL/Cleaning of Data 1](#_Toc80666523)

[Extract: 1](#_Toc80666524)

[Transform: 2](#_Toc80666525)

[Load: 2](#_Toc80666526)

[Creating the Visualizations: 2](#_Toc80666527)

[1. A Line Chart of Deaths due to Covid -19 vs Deaths due to Pneumonia 3](#_Toc80666528)

[2. A Pie Chart of Deaths due to Covid – 19 vs Deaths due to Other Reasons 4](#_Toc80666529)

[3. A Waterfall Chart of Total deaths due to Covid 19 in 2020 and 2021 4](#_Toc80666530)

[4. Filled Map of State wise distribution of Covid - 19 deaths in the U.S. 5](#_Toc80666531)

[5. A Stacked Column chart depicting Covid – 19 deaths vs Different Age Groups 6](#_Toc80666532)

[6. A stacked bar chart depicting Covid – 19 deaths in Males vs Females 7](#_Toc80666533)

[Summary: 8](#_Toc80666534)

# Abstract

This project report is the part of the Data Visualization Project carried out in our Data Handling and Visualization course taken by Prof. Jacob Shortt. This report summarizes the Visualization and ETL process done on the CDC Data Set of Provisional COVID-19 Deaths by Sex and Age. For cleaning and Analyzing the Data, MS Excel was used. Power BI and Tableau were used in making the Visualizations. In this project we were asked to create 6 Visualizations from the given data set. To showcase the visualizations, MS PowerPoint has been used.

Please Note: Data dated from 1st January 2020 to 18th August 2021 has been used in this project

# What Makes This Data Set Stand Out?

Before starting with the analyzation of a data set, we are supposed to take care of the 5V’s of Big Data. These are

* Value – As the data set consists of information about an ongoing pandemic, its insights are significant
* Veracity – As the source is CDC, A government agency itself, the data could be trusted.
* Volume – As the data set has more than 63K rows, it is enough for us to create an accurate analysis
* Variety – As the data set contains a good number of different parameters, it makes it easier for us to find relations inside the data.
* Velocity – As the source is a government website it updates on a regular basis thus we could se a significant change in the span of a few months

# ETL/Cleaning of Data

To make the analysis simpler, raw data has to go through an ETL (Extract, Transform & Load) Process thus to remove inconsistencies in the provided data set

## Extract:

The data was downloaded from the official website of the Centers for Disease Control and Prevention (CDC, Hyperlink- <https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku> ) in .csv format. The data was later exported into .xlsx format thus making it easier to store and format the data.

## Transform:

The following steps were taken for the transformation of the data:

1. Filters were applied to all the columns to get started
2. Rows containing null data were filtered out and were removed easily without disturbing surrounding data
3. Data from ‘New York City’ was deleted as it is not a state and it’s data was already considered in the collective data of the ‘New York State’. Although District of Columbia (DC) and Puerto Rico are not considered as states, they are still territories of the United States of America and thus their data was not excluded from my analysis.
4. Unnecessary columns like ‘Date As Of’, ‘Influenza Deaths’, ‘Pneumonia, Influenza, or COVID-19 Deaths’, ‘Footnote’ were removed to make the analysis easier.
5. All the Data was segregated into 3 sheets, namely ‘By Total’, ‘By Month’ and ‘By Year’ to make it easier to classify the data sets while visualizing. Year and Month in the By Total sheet were removed to further make the data analysis easier.
6. The Primary sheet was left untouched thus to be able to refer it later on.
7. To make the visualization process easier, the Month and Year columns were concatenated using an IF statement. The resultant column (named Y/M) was in the format yyyy/mm.
8. The Data was now sorted in by the alphabetical order of the states.
9. A pivot table was also used to be able to thoroughly gaze through the data and have a primitive analysis.

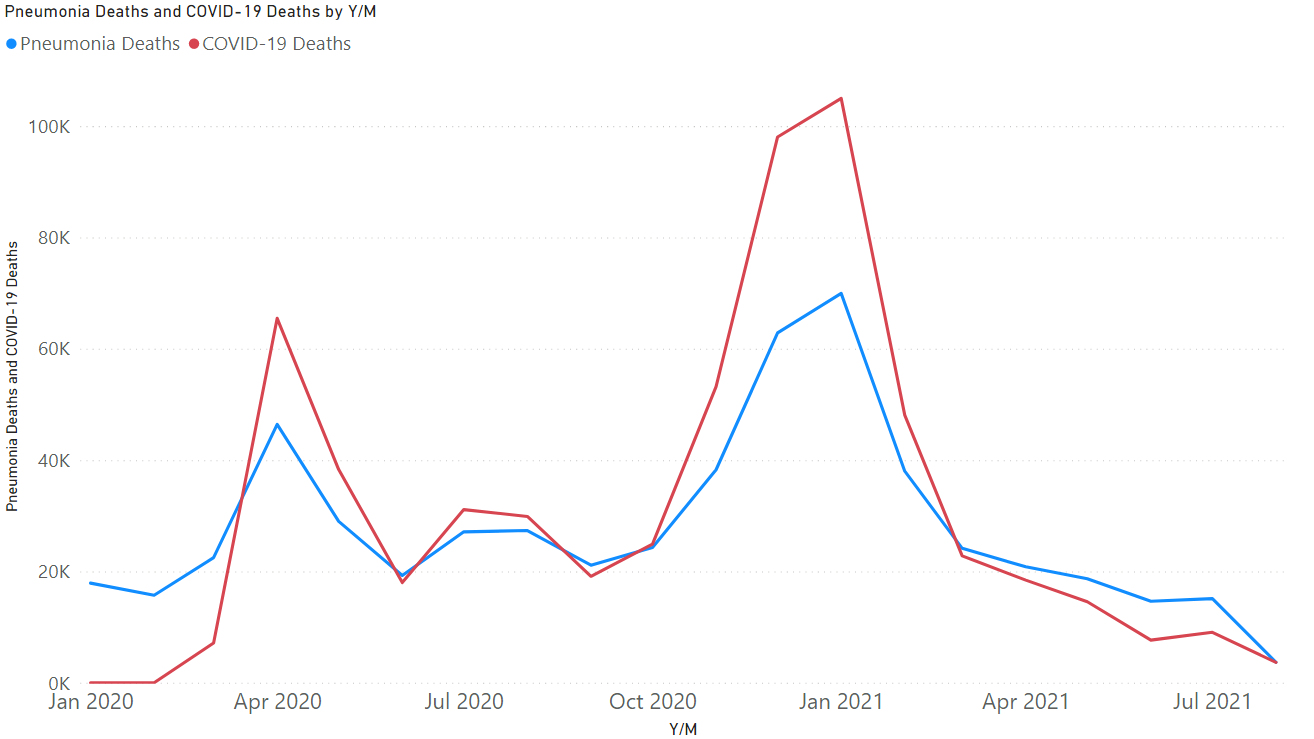
## Load:

After cleaning the data up, it was time for visualizing it. The excel file was loaded onto Tableau where relations between all the sheets were chosen manually. Later the data was loaded on Power BI where the software automatically creates relations between the sheets which is one of the advantages of Power BI. Most of the Visualizations were made using the By Month sheet.

# Creating the Visualizations:

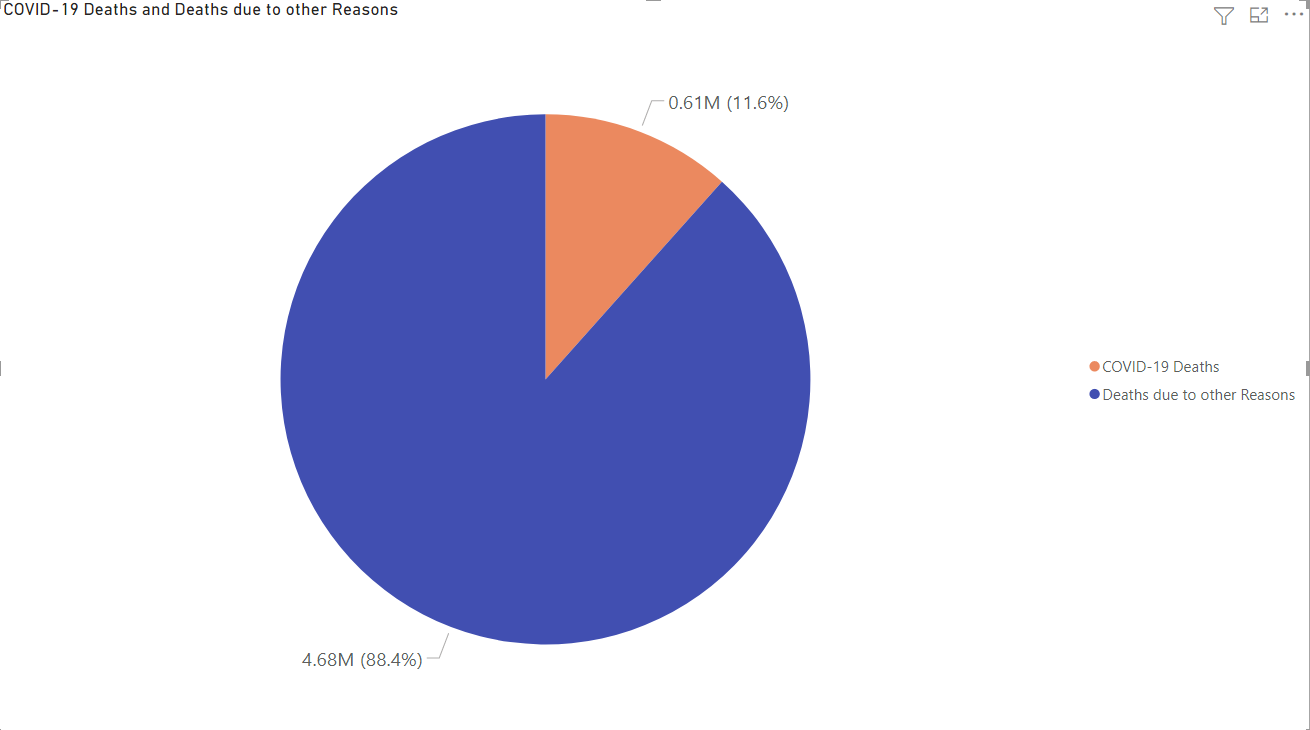
Six Visualizations of different forms have been created using both Power BI and Tableau. Images of the Visualizations have been attached to the PowerPoint. Following are the insights that I have gained from each Visualization:

## A Line Chart of Deaths due to Covid -19 vs Deaths due to Pneumonia



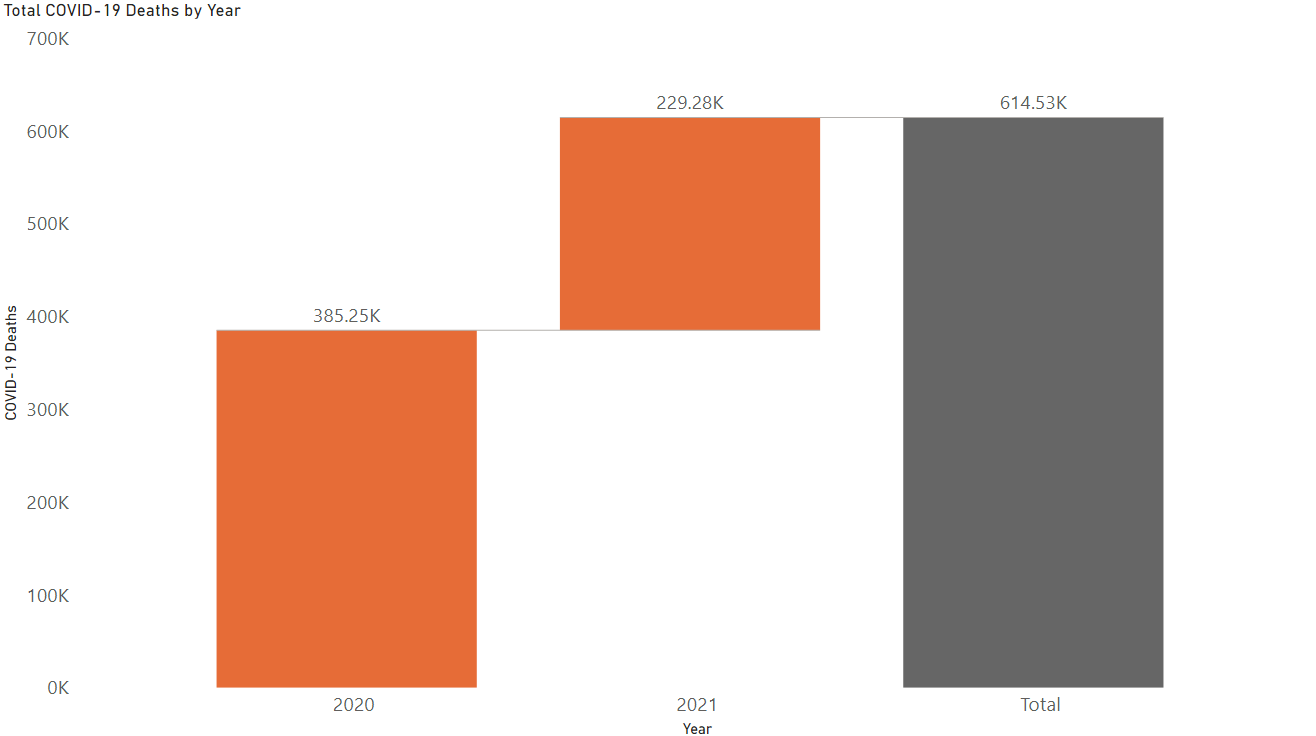
* What does the chart say: This chart shows us how the deaths by Covid – 19 surpassed the deaths by pneumonia near April 2020 and were close to each other until around December 2020 where Covid – 19 rose to a very high point until falling back below Pneumonia deaths in March 2021.
* Insights Gained: Although being almost similarly deadly for long time, Covid-19 has had its highs during the last 1.5 years proving a lot more fatal than pneumonia in these times when the medical knowledge of pneumonia has advanced a lot since its primary outbreak.
* Software Used: Power BI
* Sheet Used: By Month

## A Pie Chart of Deaths due to Covid – 19 vs Deaths due to Other Reasons



* What does the Chart Say: The chart shows us the difference between the Number of deaths due to Covid – 19 and the Number of Deaths due to Reasons other than covid – 19 since January 2020, which we can see is a very high number.
* Insights Gained: We can easily make out that although almost 4.7M people dies due to reasons other than covid – 19, covid – 19 suffices for more than 11% of the total deaths in U.S. since January 2020 which is a pretty big number for any country.
* Software Used: Power BI
* Sheet Used: By Month

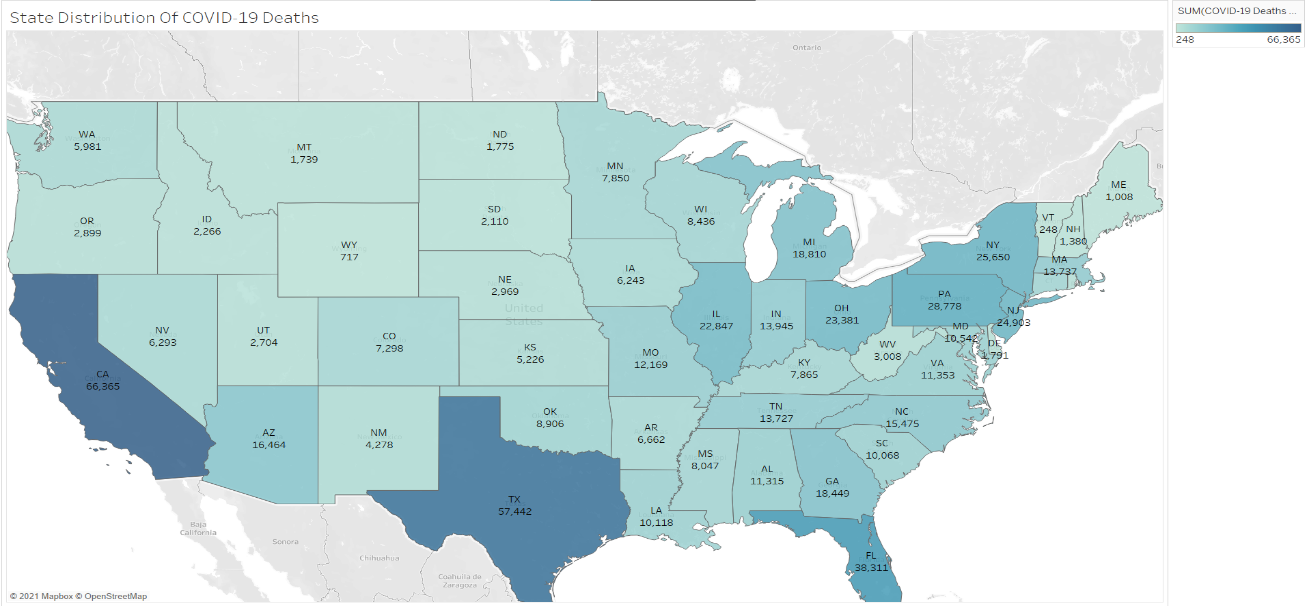
## A Waterfall Chart of Total deaths due to Covid 19 in 2020 and 2021



* What does the chart Say: The chart compares the Deaths due to Covid – 19 in 2020 to the Deaths due to Covid-19 in 2021 vs the Total covid-19 deaths. A waterfall chart has been used to easily compare the deaths while giving us an idea of the timeline.
* Insights Gained: We learnt that the number of deaths in 2021 are almost 66% of the number of deaths in 2020. Although we must keep in mind that this data is of only 8 months of 2021 thus the final data could be different owing to the Delta Variant of Covid-19.
* Software Used: Power BI
* Sheet Used: By Month

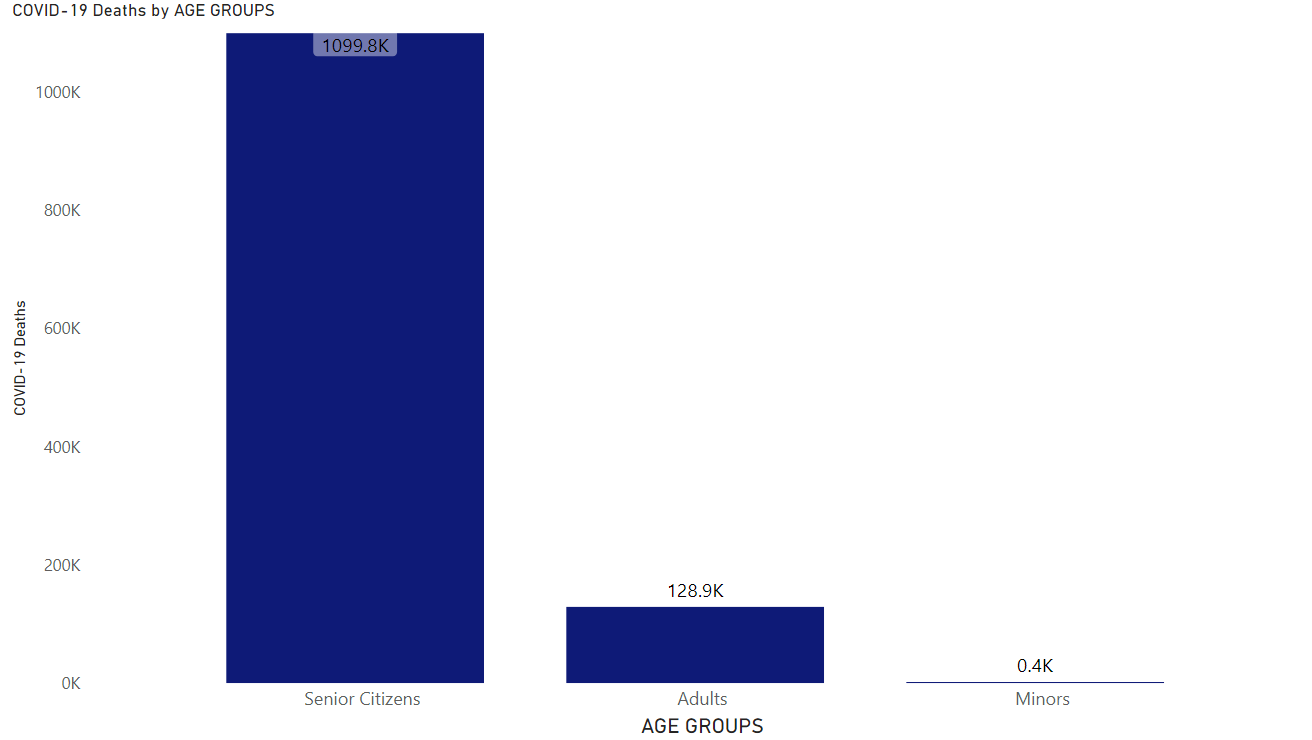
## Filled Map of State wise distribution of Covid - 19 deaths in the U.S.

* What does the Chart Say: The chart shows the distribution of Covid – 19 deaths state-wise. The chart shows the states with most deaths in darker shade of blue whereas it gets lighter as the numbers of deaths decrease. The legend on the top right corner depicts the same. I have not put in the numbers as it would have formed a clutter in the visualization. If I would’ve tried to avoid the clutter and put in the numbers, I would have to cut out Alaska, Hawaii and Puerto Rico from the chart. A chart with those changes has been inserted below.



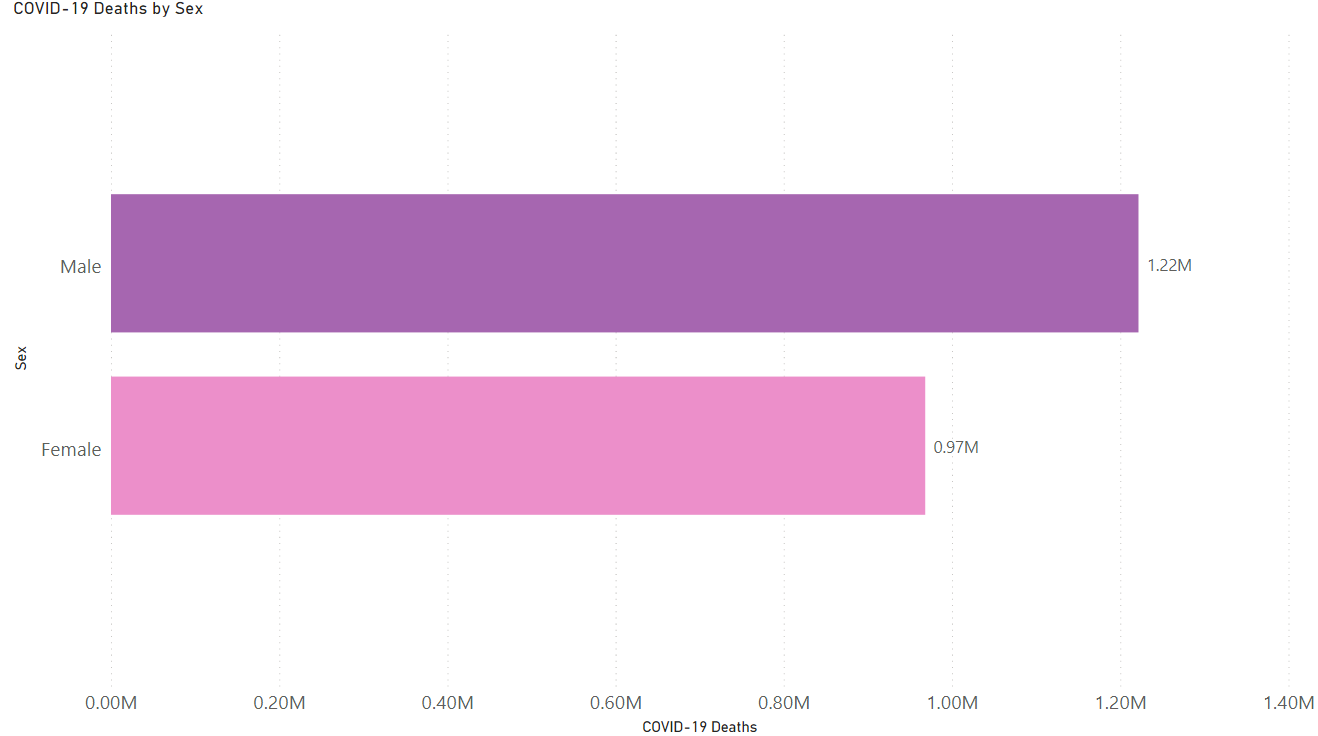
* Insights Gained: Thus, we got to know that most deaths have occurred on the East Coast except California. We also learnt that California has the highest number of Covid – 19 fatalities which is around 66K.
* Software Used: Tableau
* Sheet Used: By Total

## A Stacked Column chart depicting Covid – 19 deaths vs Different Age Groups



* What does the chart say: The chart shows us 3 different age groups Minors (Age 0 to 17), Adults (Age 18 to 64) and Senior Citizens (Age 65 and above). As the number of deaths of Minors and Adults was a relatively tiny number than that of Senior Citizens, we had to ditch the ‘Millions’ unit and had to go ahead with ‘thousands’.
* Insights Gained: We got to know that the number of fatalities in Senior Citizens (People aged above 65 years) was very high, around 10 times the deaths of adults and 2750 times the deaths of minors, which was very heartbreaking. What we learnt is Senior Citizens are most prone to an unfortunate outcome after getting infected with covid – 19 and should be taking the highest sorts of precautions.
* Software Used: Power BI
* Sheet Used: By Total

## A stacked bar chart depicting Covid – 19 deaths in Males vs Females



* What does the Chart Say: By looking at the chart we can figure out that it’s a direct comparison between the deaths caused due to Covid – 19 in Males and Females. The units here have been taken in millions as the values of both were comparably close.
* Insights Gained: We got to know that since January 2020, the deaths of males due to covid – 19 has been a significantly larger number than that of females. Although the number of female deaths is nowhere close to be called less, the trend shows a higher number of male deaths than female deaths.
* Software Used: Power BI
* Sheet Used: By Total

# Summary:

Thus, we got to know that pneumonia has gone hand in hand with covid-19 to cause not as many as covid – 19 but still a significant number of deaths. Covid – 19 has caused more than 610K deaths in the United States since January 2020. We also concluded that California has the highest number of Covid – 19 deaths rounding to around 66K. We also concluded that Covid – 19 is relatively more fatal for senior citizens than other age groups and has shown insane numbers in Deaths compared to that of other age groups. We also found out that the mortality rate of males is higher than that of females. While analyzing this data set, I realized that the horrors of Covid – 19 has left no household alone. Whether it is U.S. or India, the claws of this deadly virus have caused irreparable damage to all of us and the only way we can fight it is by taking necessary precautions until it is completely eradicated and to take the necessary vaccinations and medicines.

This analysis got me thinking as a real-world analyst for the first time since I’ve started college and has given me an idea of how it’d be outside the walls of my university. I’d like to thank Prof. Jacob Shortt for how amazingly he has taught the course and helped us grow and stand out. Looking forward to attending future classes with him.

THANK YOU.