*Provisional COVID-19 Deaths by Sex and Age Group:*

*A Data Storytelling Report*

By Sauravi Lalge

8/23/2021

Provisional COVID-19 Deaths by Sex and Age Group:

A Data Storytelling Report

Contents

[Abstract: 2](#_Toc80671261)

[Software’s Used: 2](#_Toc80671262)

[About the Dataset: 3](#_Toc80671263)

[Provisional COVID-19 Deaths by Sex and Age Group: 3](#_Toc80671264)

[ETL Process: 3](#_Toc80671265)

[Visualizations: 4](#_Toc80671266)

[ Stacked Column Chart: 4](#_Toc80671267)

[ Funnel Chart: 5](#_Toc80671268)

[ Tree Map: 6](#_Toc80671269)

[ Clustered Bar Chart: 7](#_Toc80671270)

[ Line Chart: 7](#_Toc80671271)

[ Donut Chart: 8](#_Toc80671272)

[Conclusion: 10](#_Toc80671273)

[About the data: 10](#_Toc80671274)

[About the project: 10](#_Toc80671275)

# Abstract:

This is a project report for Data Handling and Visualization course. This report sums up the process of transforming data-driven analyses into a widely-accessible visual format to influence a decision, strategy and turn it into actionable insights and apply methods learnt from the course. It focuses on visualizing and analyzing data provided by the Centers for Disease Control (CDC) website on Provisional COVID-19 Deaths by Sex and Age.

# Software’s Used:

1. MS Excel:

Microsoft Excel is one of the top tools for data analysis and the built-in pivot tables are arguably the most popular analytic tool.

1. Power BI:

Power BI provides rapid insight and is used to extract and visualize data. Power BI brings together data from multiple sources to give you a comprehensive view of information assets.

# About the Dataset:

## Provisional COVID-19 Deaths by Sex and Age Group:

I chose ‘Provisional COVID-19 Deaths by Sex and Age Group’ as in my judgement this data set encloses all the 5 V’s i.e., Variety, Velocity, Value, Veracity, and Volume, that an ideal dataset should consist of. Another justification considering this dataset is that given the present situation of a pandemic, which area of the USA populace was influenced and to what extent was appealing to me. This particular dataset gives us an overview of Deaths involving coronavirus disease 2019 (COVID-19), pneumonia, and influenza by sex, age group, and jurisdiction of occurrence.

# ETL Process:

The ETL process being the pivotal step, encompasses data **E**xtraction, **T**ransformation, and **L**oading of data.

The data was downloaded directly from CDC website as a CSV file and was exported to MS Excel. This CSV file was then transformed to a .xlsx file to avoid any loss of data. This dataset is genuinely direct, there wasn't a lot to do when it came to cleaning the information. In the dataset there was a column that contained cumulative data for the United States. This had to me removed as the we have only taken individual states into consideration.

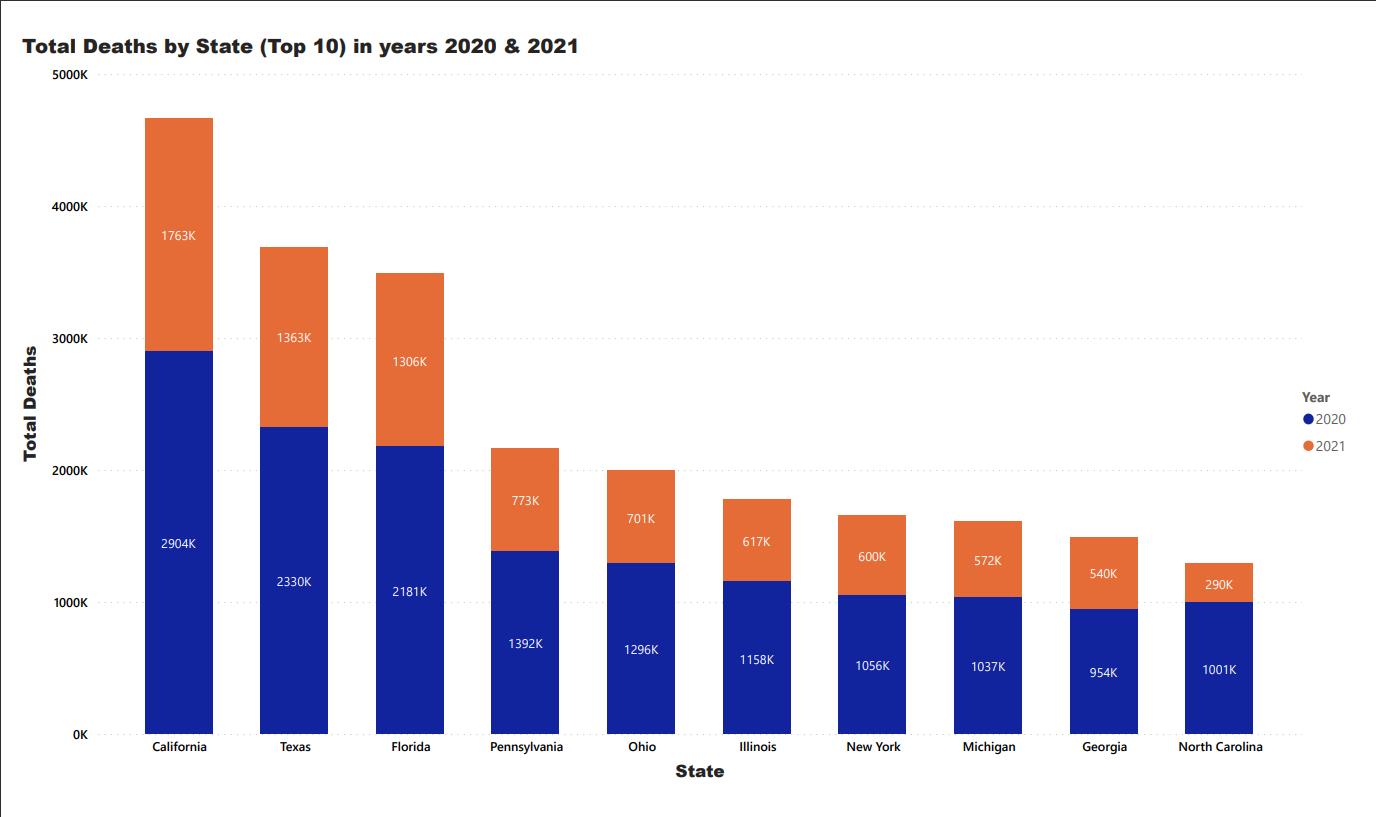
The data was then loaded straight to Power BI.

# Visualizations:

The data in all visualizations is from 2020 & 2021 unless said otherwise.

## Stacked Column Chart:

Stacked column chart is a very common and easy way to analyze data as the data points are clearer when they are on top of each other. Here, a comparison between the total deaths in the states in years 2020 & 2021 is displayed. Post filtering, the visualization shows top 10 states with the greatest number of deaths in the USA.

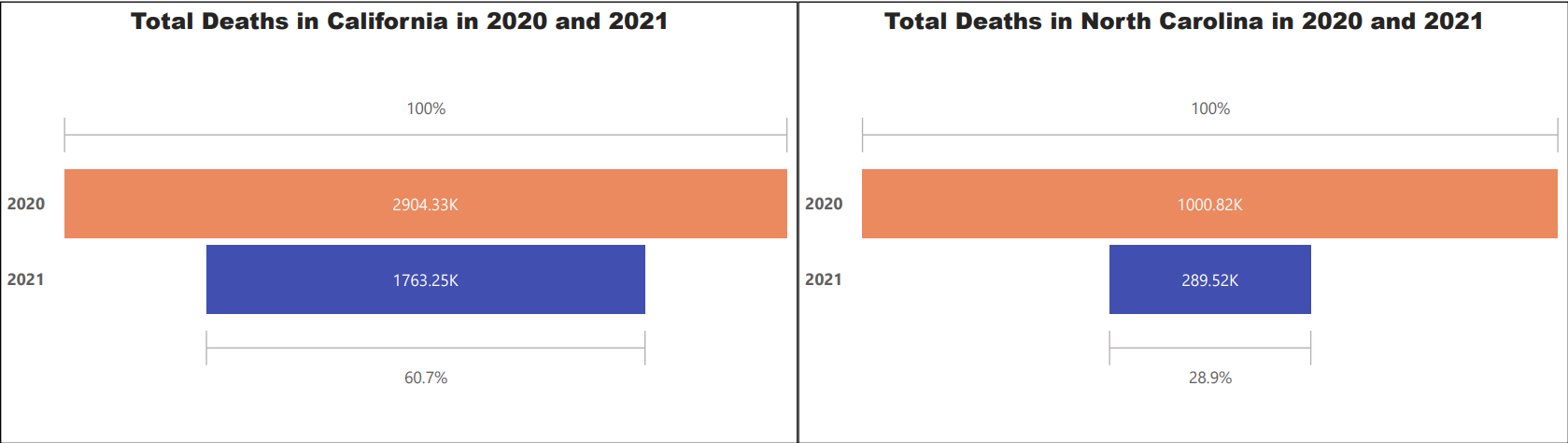


Insights:

* California has the maximum deaths amongst all states in the past two years.
* The number of deaths in Texas have almost halved in 2021 as compared to 2020.
* North Carolina has the lowest ratio for total deaths in the two consecutive years in this visualization.

## Funnel Chart:

A funnel chart helps you visualize a linear process that has sequential connected stages. Since California and North Carolina had the maximum and minimum deaths within the above visualization respectively, here is a comparison in percentage. The subsequent chart shows the percentage decrease in deaths within the particular state by taking data from 2020 and 2021 into consideration.

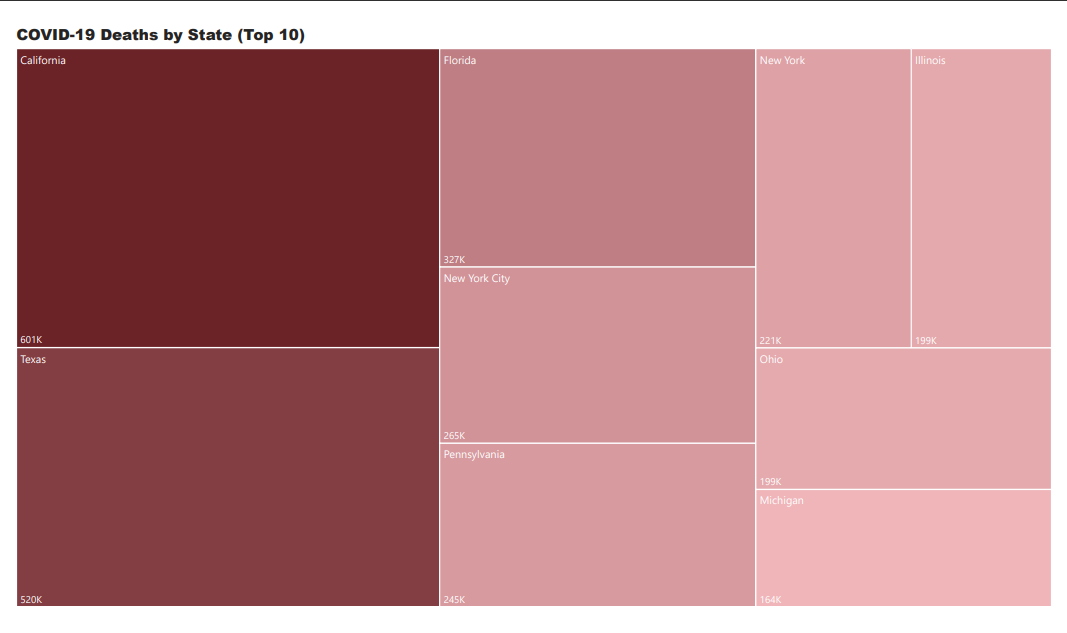


Insights:

* The total number of deaths in California reduced by 40%.
* The total number of deaths in North Carolina reduced by 70% approximately.
* The difference between the number of deaths in California within the years 2020 and 2021 continues to be greater than the overall deaths in North Carolina within the year 2020 alone.

## Tree Map:

Treemaps display data in nested rectangles. You use dimensions to define the structure of the treemap, and measures to define the size or color of the individual rectangles. Treemaps are a relatively simple data visualization that can provide insight in a visually attractive format. This treemap shows COVID-19 deaths in the United States. To avoid clutter, visualization was filtered to show the top 10 states.

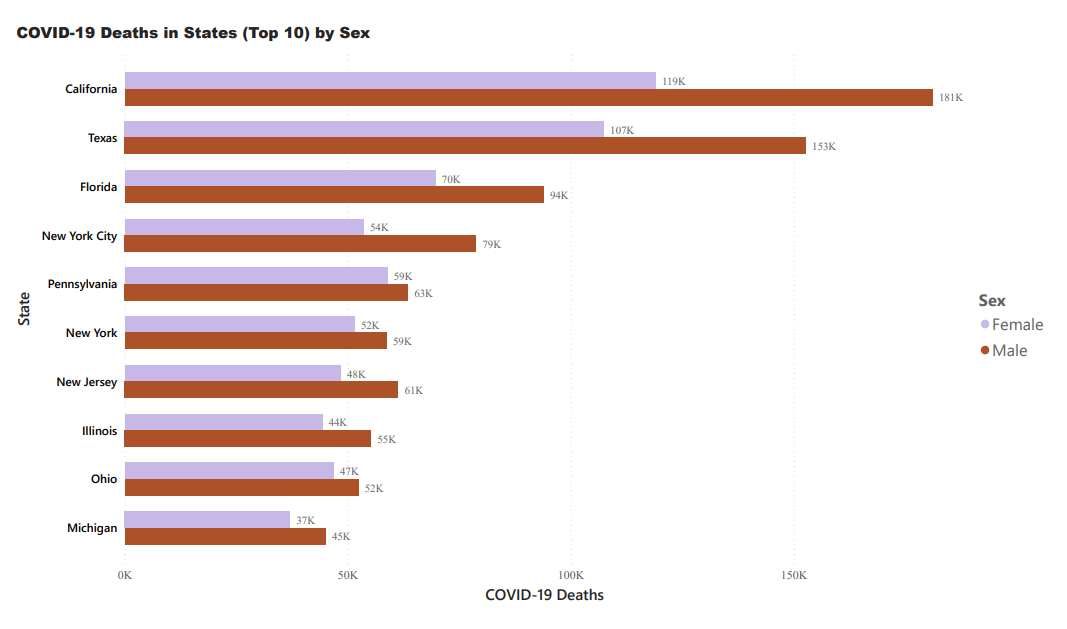


Insights:

* California has the most COVID-19 deaths over the two years.
* Unlike the total number of deaths, Florida ranks behind Texas, and the trend has changed.

## Clustered Bar Chart:

A clustered bar chart extends the bar chart, plotting numeric values for levels of two categorical variables instead of one. The visualization below shows the number of COVID19 deaths among men and women in different states. This chart was then filtered to show top 10 states so as to avoid clutter.

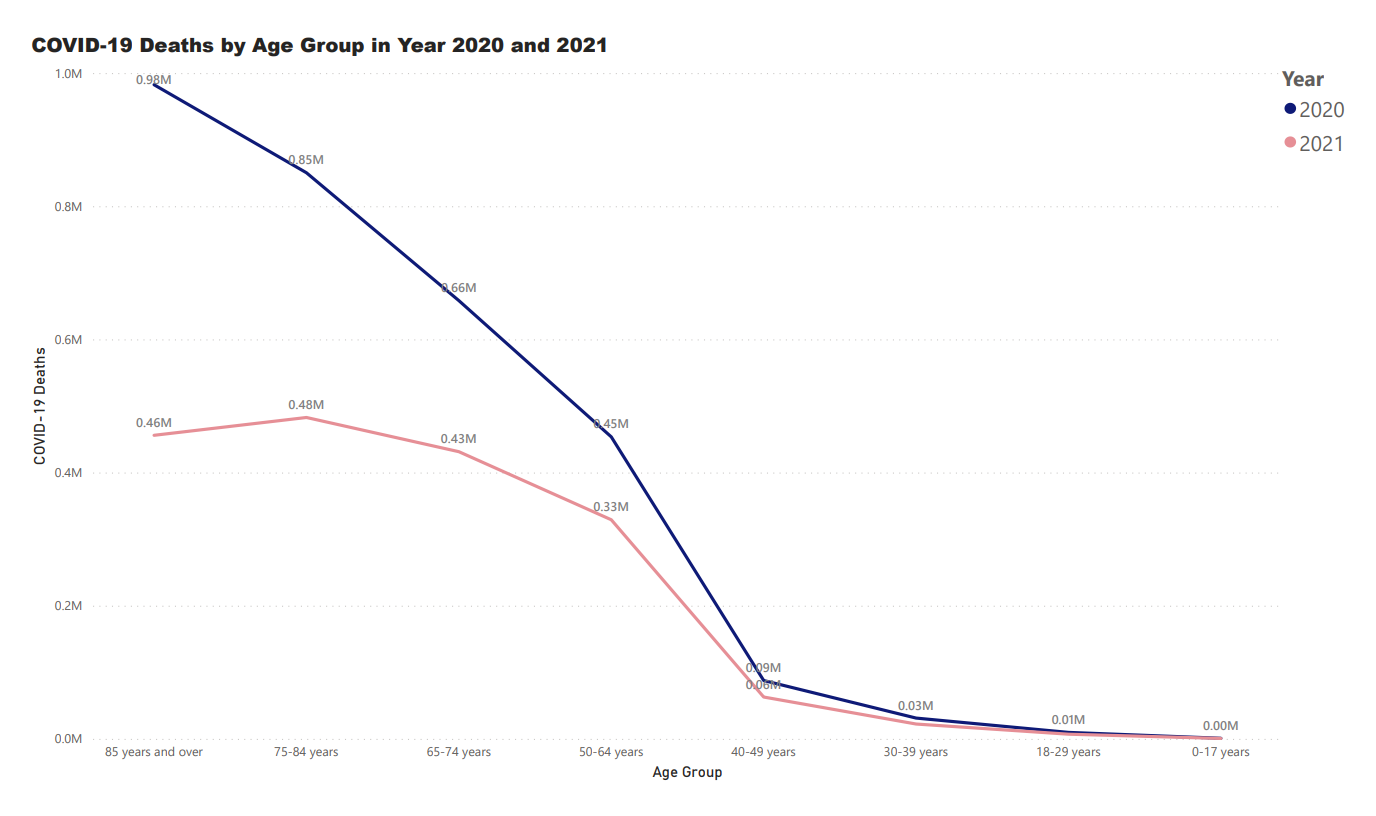


Insights:

* Deaths in females are lesser than deaths in males for all the states.

## Line Chart:

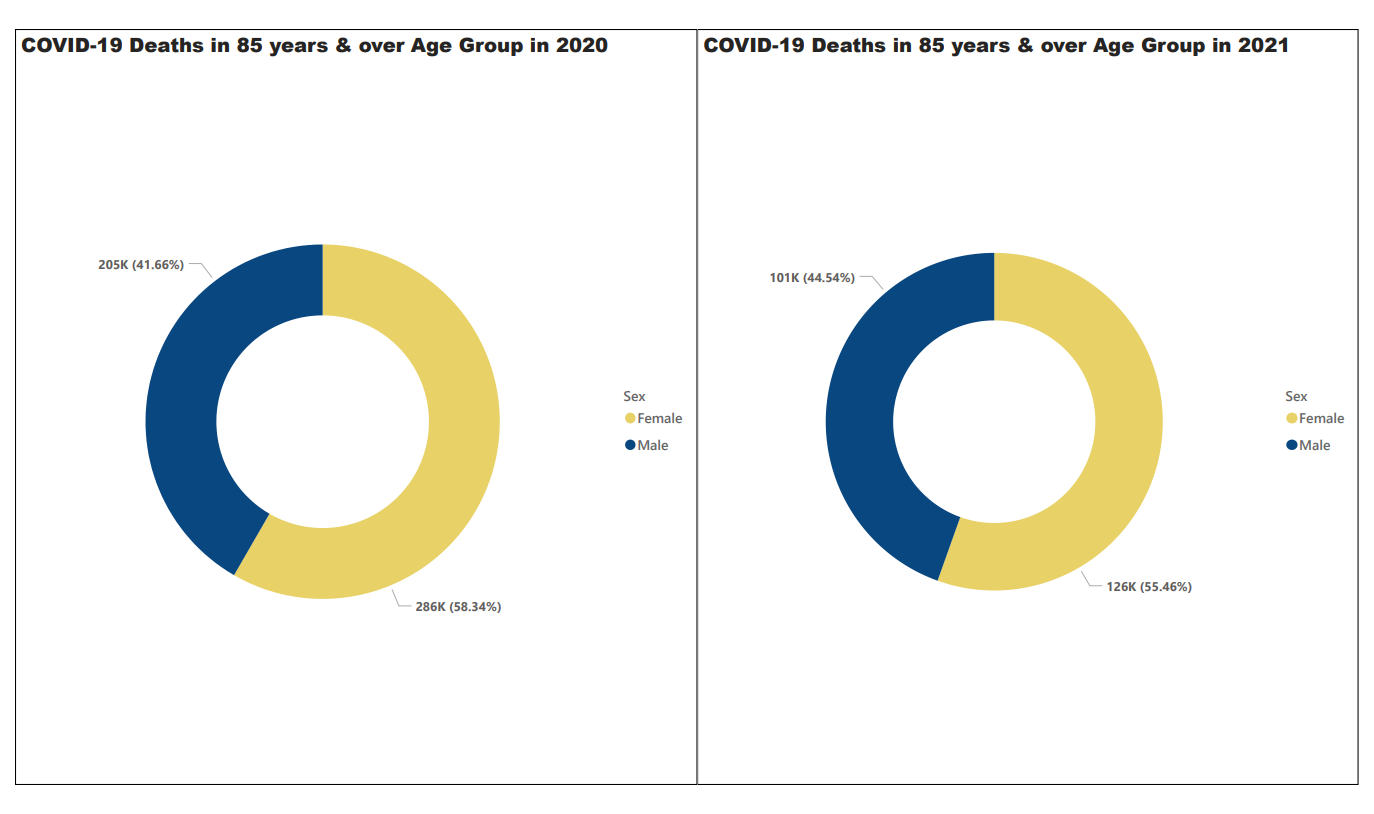
A line chart is a graphical representation of an asset's historical action that connects a series of data points with a continuous line. The data presented here is the number of deaths from COVID-19 in different age groups. Line graphs for separate years is best suited for this analysis.



Insights:

* As seen, the most affected age group was the elderly.
* The number of COVID deaths rapidly went up post 50 years and above age group.
* The deaths in 2021 have significantly reduced as compared to 2020.

## Donut Chart:

Donut charts are used to show the proportions of categorical data, with the size of each piece representing the proportion of each category. Using a donut chart helped as here we are making an analysis by comparing two different data points- by gender and by year in the elderly age group.

Insights:

* Male to Female death ratio has barely changed in 2021.
* Although, the number of deaths in females has reduced more than half of that in 2020. It has also remarkably reduced for males.

# Conclusion:

## About the data:

COVID-19 has been devastating to health and to society more broadly. We observed higher mortality rate in elderly age groups but it has gradually improved over time. A more speculative possibility of why mortality has decreased is the possibility that larger infective doses of COVID-19. There is reason to be proud of how much medical progress, including upcoming vaccines and monoclonal antibodies, as well as increased masking and social distancing, has been made to combat a virus that was unknown until the end of 2019.

## About the project:

I can confidently state that this was an excellent learning opportunity. This activity required me to start from the beginning and work my way through it using various software’s. To the best of my ability, I have attempted to apply all of the discussed topics. The chosen dataset was pretty straightforward, with only a few of data points to concentrate on. This significantly aided in the development of a solid foundation for future projects.