COVID-19 Death Cases Analysis by Date, State, Sex and Age Groups

Chen Chen 6381370662

2022-12-07

Introduction

Dataset Background

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. As we know, most people infected with COVID-19 will experience mild to moderate respiratory illness and recover without requiring special treatment. However, some will become seriously ill or die at any age [1]. The data set of this project is from CDC (Centers for Disease Control and Prevention) and describes deaths involving COVID-19, pneumonia, and influenza reported to the National Center for Health Statistics by sex, age group, and jurisdiction of occurrence [2]. The provisional counts for COVID-19 deaths are based on a current flow of mortality data in the National Vital Statistics System. National provisional counts include deaths occurring within the 50 states and the District of Columbia that have been received and coded as of the date specified [3].

Dataset Description

The dataset has 107, 406 observations and 16 variables. Each row represents COVID-19 deaths by sex, age, state, year, and month. For columns, the dataset has 8 character variables and 8 integer variables. Their name and description are listed below.

Variable_Name	Description
Data As Of	Date of analysis
Start Date	First date of data period
End Date	Last date of data period
Group	Indicator of whether data measured by Month, by Year, or Total
Year	Year in which death occurred
Month	Month in which death occurred
State	Jurisdiction of occurrence
Sex	Sex
Age Group	Age group
COVID-19 Deaths	Deaths involving COVID-19
Total Deaths	Deaths from all causes of death
Pneumonia Deaths	Pneumonia Deaths
Pneumonia and COVID-19 Deaths	Deaths with Pneumonia and COVID-19
Influenza Deaths	Influenza Deaths

Variable_Name	Description
Pneumonia, Influenza, or COVID-19 Deaths	Deaths with Pneumonia, Influenza, or COVID-19
Footnote	Suppressed counts (1-9)

The key variables being used in this project are End Date, Year, Month, State, Sex, and Age Group. The key output variable in the analysis is COVID-19 Deaths. This dataset tells the case number of the United States Covid-19 deaths in each state, gender, age, etc. from January 1, 2020, to November 19, 2022. We will conduct more specific analyses based on them below.

Project Objective

It is known that Covid-19 began to spread widely in 2019 and three years later, with the widespread vaccination, it has entered into a more stable state. Both vaccination and virus mutation can significantly reduce the death rate of this virus. As the date information is provided by the dataset, the first question this project going to explore is: do COVID-19 death cases decrease by date, including months and years?

In addition, we are still curious about whether the Covid-19 death toll is related to state, sex, and age groups. By exploring distributions, we can formulate more targeted epidemic prevention policies according to different states, sex, and people of different age groups. Therefore, the second objective of this project is: Do COVID-19 death cases vary by state, sex, and age group?

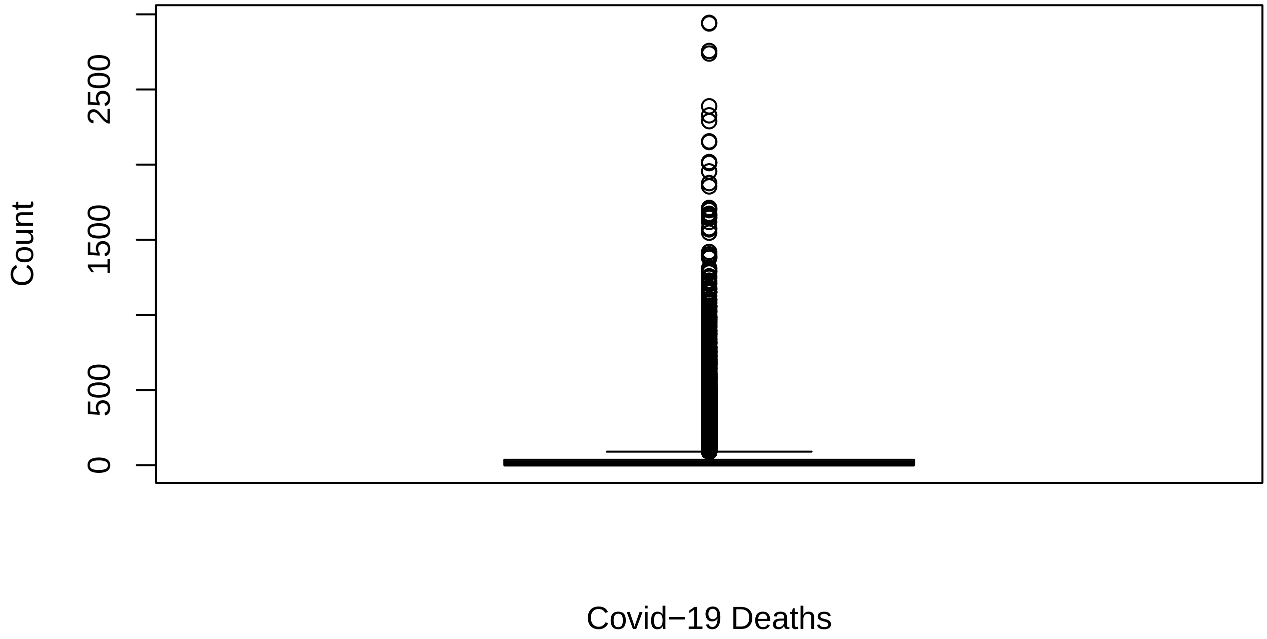
These two questions will be explored and explained in the later part of this report.

Methods

Data Cleaning and Preparation

The first step of data cleaning is to have an overview of the whole dataset. We can notice that each variable in this dataset has a "summary observation" for summarizing the variable. For example, for a column of "State", the table contains data for each different state and also has some observations called "United States", which summarizes the total number of all states stratified by other variables. Similarly, for the variable of "Sex", it has "All Sexes" as the summary and for "Age Groups", it is "All Ages". Therefore, prior to future analysis, we need to select key variables we are going to use first and filter out those summary observations to avoid redundancy.

After filtering out those "summary rows" and dropping NAs, the next step is to check if there are any irregular values and drop them off. Do the summary of the column of "COVID-19 Deaths" in the new dataset. We find that there is no negative value but the maximum value is quite large to 2943. We perform the box plot and find that there are several extremely large values above 2500. Try to filter them out and the results are shown below.



End Date	Group	Year	Month	State	Sex	Age Group	COVID-19 Deaths	Total Deaths
01/31/2021	By Month	2021	1	California	Male	50-64 years	2739	5700
01/31/2021	By Month	2021	1	California	Male	65-74 years	2939	6026
01/31/2021	By Month	2021	1	California	Male	75-84 years	2943	6297
01/31/2021	By Month	2021	1	California	Female	85 years and over	2756	8248

From the table, it looks like those big values are from January 2021 in California. They have meaning so we cannot see them as irregular values to drop them off. In addition to the above steps, we also converted the format of the variable representing the date. Converting to the date format can make later analysis and visualization easier.

Data Analysis and Exploration

Now we get a clean dataset. It is time to carry out data analysis. For our first question, which is to explore the relationship between the number of deaths and dates, we first make a summary table for different years.

Year	Total_Death_Cases	Avg_Death_Cases	Min_Death_Cases	Max_Death_Cases
2020	445344	39	0	2388

Year	Total_Death_Cases	Avg_Death_Cases	Min_Death_Cases	Max_Death_Cases
2021	588675	56	0	2943
2022	253859	30	0	952

The table above summarizes the total death cases, average Covid-19 death cases as well as the min and max for each year. From the table, it can be found that 2021 is the year with the most total death data, including the average and maximum. Instead, all values for 2022 are the smallest. In order to explore the distribution by months, we do a similar summary for all 3 years by month and arrange the result in descending order. However, the results of 3 years are quite different, and hard to find out the regulations. So it can be concluded that the number of deaths has little relationship with the month in these three years. Results for 2021 are presented as examples in the table below. Interactive tables for all three years can be referenced on the project website.

Month	Total_Death_Cases	Avg_Death_Cases	Min_Death_Cases	Max_Death_Cases
1	124471	130	0	2943
9	88992	90	0	1645
8	67957	74	0	1664
12	59040	62	0	587
2	57703	66	0	1282
10	57362	59	0	808
11	41353	45	0	424
3	27854	33	0	456
4	23025	28	0	266
5	18297	24	0	219
7	13693	18	0	338
6	8928	12	0	139

For the second question, we can do a similar summary stratified by state and age groups. The results are shown below.

State	Total_Death_Cases	Avg_Death_Cases
Texas	131617	160
California	130495	154
Florida	94636	119
Pennsylvania	57740	82
Ohio	54983	76
New York	45201	69
New York City	44274	68
Georgia	43743	62
Illinois	43563	62
Michigan	41548	61

For states, we only show the top 10 results. The complete interactive result can be found on the website Appendix. From the table, Texas, California, and Florida have the top 3 highest levels of both total death cases and average. Vermont, Alaska, and Hawaii are the three states with the fewest cases.

Age Group	Total_Death_Cases	Avg_Death_Cases
85 years and over	281032	94
75-84 years	275956	94

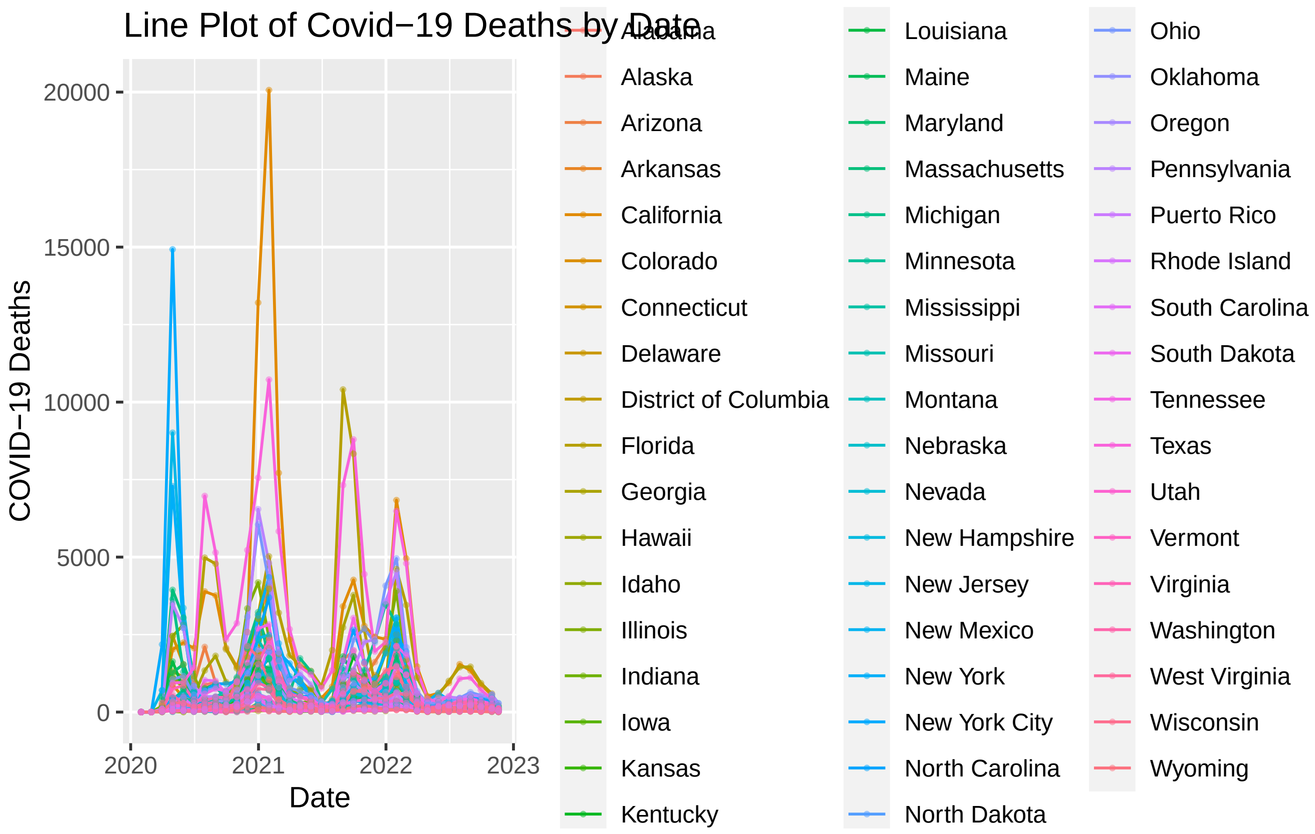
Age Group	Total_Death_Cases	Avg_Death_Cases
65-74 years	240511	85
50-64 years	192244	73
55-64 years	149402	60
45-54 years	63394	32
40-49 years	38645	23
35-44 years	22837	15
30-39 years	13366	9
25-34 years	7025	5
18-29 years	2865	2
15-24 years	547	0
0-17 years	54	0
1-4 years	0	0
5-14 years	0	0
Under 1 year	0	0

The table above summarizes the death cases by age group. It can be seen that the number of death cases shows an apparent increasing trend with the increasing age.

Results

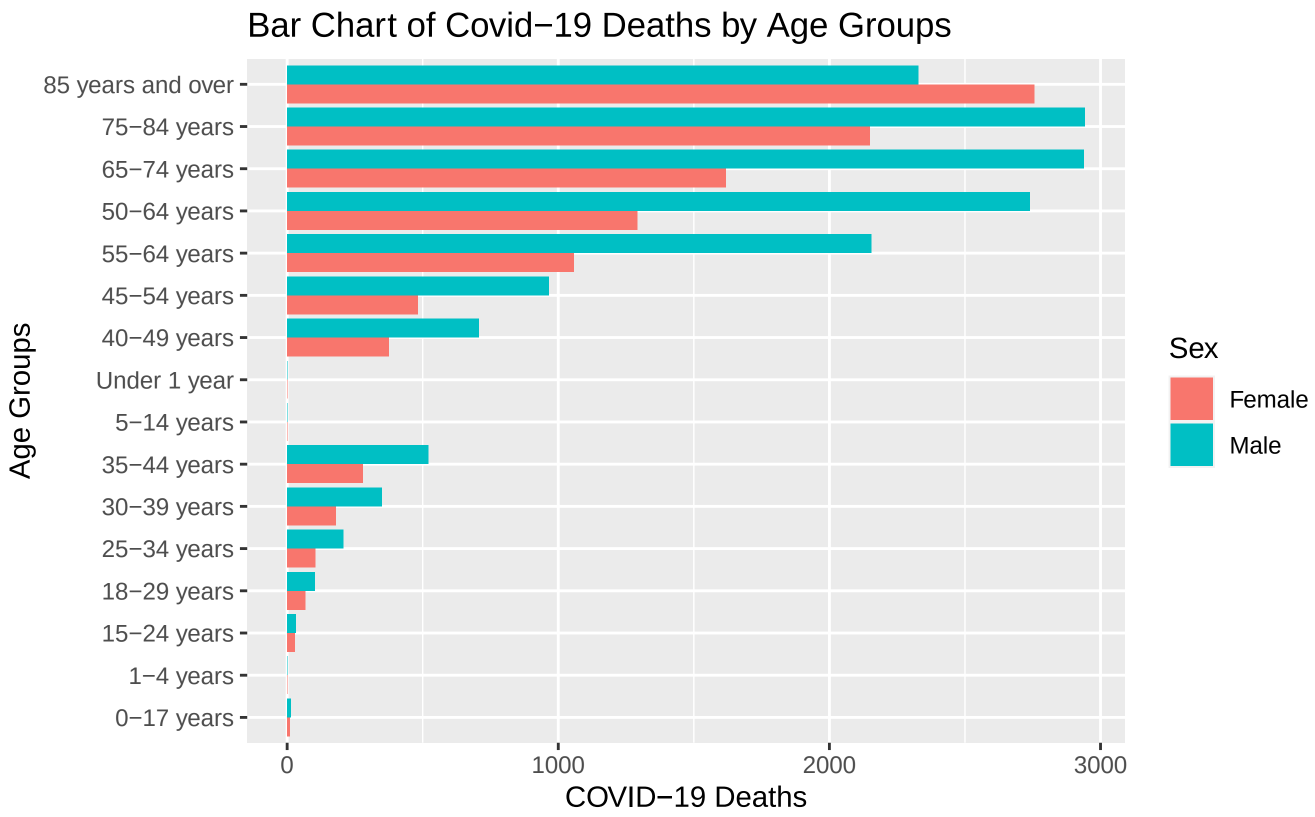
After a series of cleaning, summarizing, and exploring the dataset, we obtain several visualization results for the questions. They will be presented one by one later. And the interactive version of these plots will also be presented on the project website.

The first chart is the line plot of Covid-19 death cases by date. The range of the date is from January 2020 to November 2022. The different color here in the plot represents the different state. By this, we can find out distribution by date among states as well.



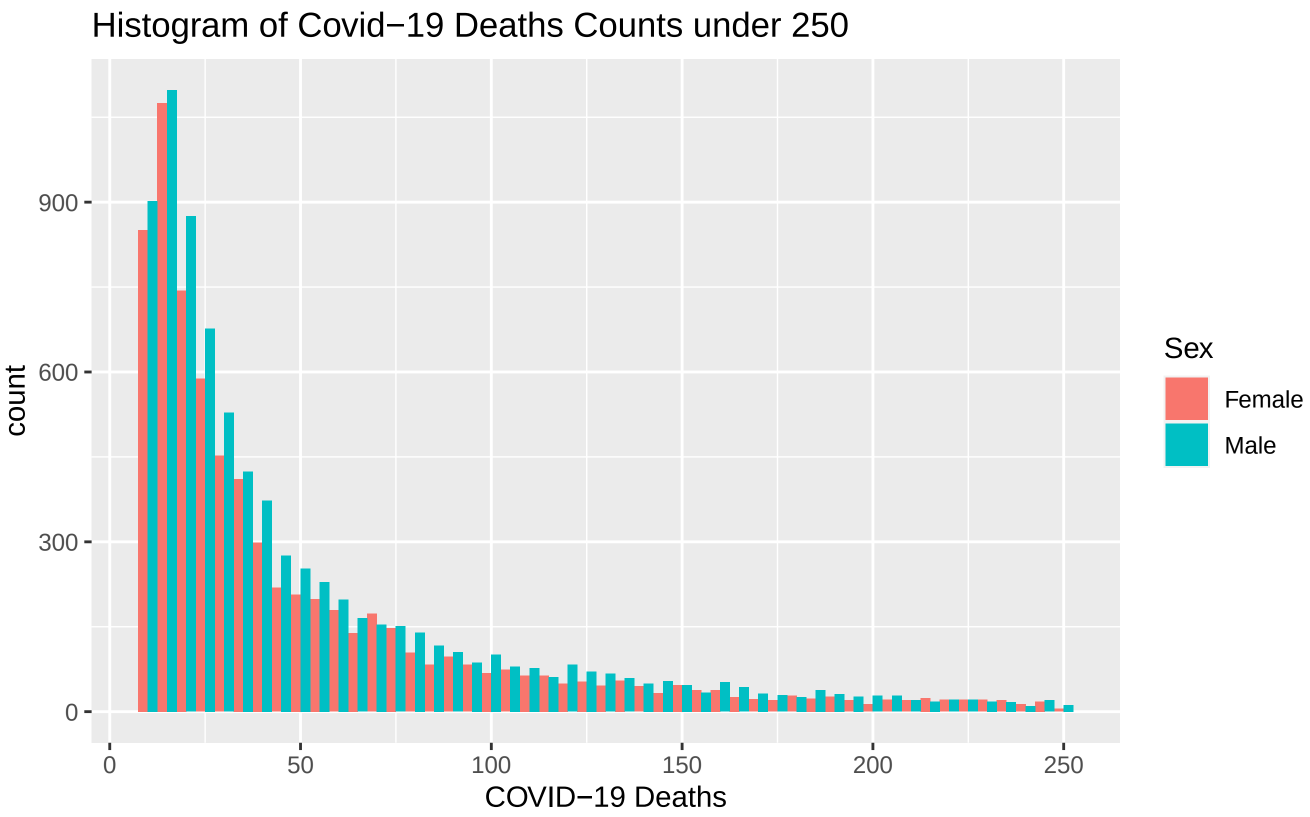
From the chart, We can find that from 2020 to 2022, the overall number of deaths shows a trend of increasing first and then decreasing. The data peaks in early 2021, and the recent death data (end of 2022) is much smaller than the data when the virus just come out to spread (2020). It indicates that the death rate from Covid-19 has been greatly reduced. The data distribution in this image matches our initial expectations. From the perspective of states, the area with the largest number of deaths in 2020 in New York City. It is noticed that blue lines are at higher levels, referring to the vicinity of New York State. After entering 2021, California's data began to rise rapidly and reached the highest peak. At the end of 2021, there is a decline in CA, and the peaks become Florida and Texas. This distribution is consistent with the results presented in our previous summary table.

The following is the visualization presentation for the second question, because Age Group is a categorical variable with relatively more classifications, we choose a bar chart to display the death cases data for different age groups.



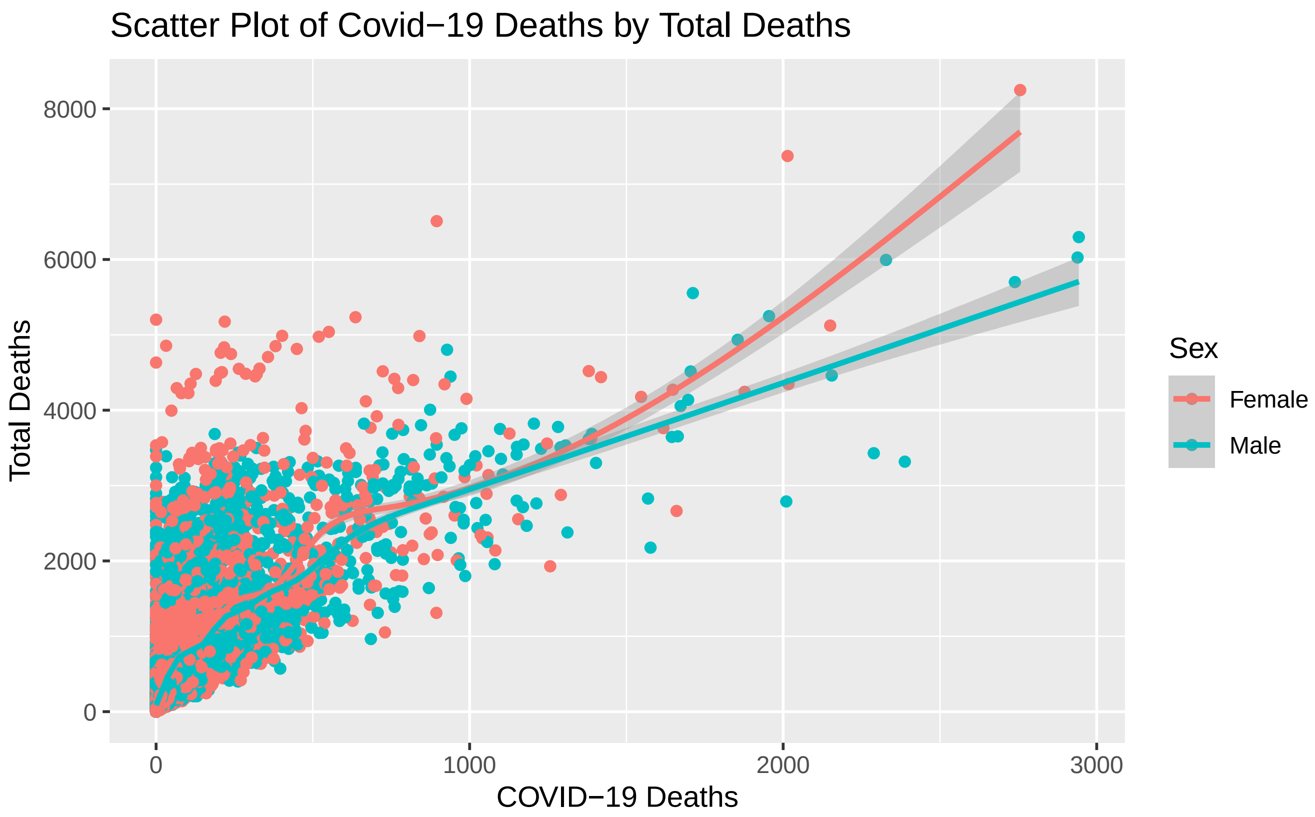
Combined with the summary table in the previous section, we can get the same distribution result: the number of death cases shows an apparent increasing trend with increasing age. It is worth mentioning that the bars of different colors in the figure refer to different genders: blue represents males, and red represents females. From the chart, except for the two age groups 55-64 and over 85, most of the other groups have more male deaths than females. This is a gender-related result that can be reflected in this graph.

For sex groups analysis, the next figure is one histogram of death cases colored by gender. The figure only shows the data with the count of death cases below 250. Because after observing the dataset, most of the observations are below 250. In this case, the distribution of gender classification can be indicated more intuitively.



It can be clearly seen from the histogram that most of the observations for death cases under 250 have more males than females. Combining the above visualization results, we can conclude that in the three-year Covid-19 death data from 2020 to 2022, there are more males than females.

The last graph is a scatter plot of Covid-19 Deaths by total death cases with regression lines and points colored by gender.



This plot shows the relationship between the number of Covid-19 deaths and the total number of deaths. It is not difficult to find from the regression line that these two variables have a certain degree of correlation. With the increase in the total deaths, number of deaths from Covid-19 is close to linear growth. From the perspective of gender, at the same level of total deaths, males have a higher proportion of covid-19 deaths (the blue line is on the right side of the red line), which also confirms our previous conclusions.

Conclusion and Summary

This project performs exploration into the objective problems through the observation, processing, and analysis of the 2020-2022 US Covid-19 death data and gets the following results. During these three years, the number of Covid-19 deaths has fluctuated, showing an increasing trend until the beginning of 2021 and starting to decline significantly thereafter; For the data of each state, California, Texas, and Florida have the top three most total Covid-19 death data. New York City, California, and Florida have all reached the highest level among states at different times; The Covid-19 death cases increase by age group, with higher age groups having more deaths; In terms of gender, most of the total number and proportion of Covid-19 deaths in the male group is higher the female group.

Reference List

[1]World Health Organizations, https://www.who.int/health-topics/coronavirus#tab=tab_1.

[2]Centers for Disease Control and Prevention, https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku.

[3]Centers for Disease Control and Prevention, https://www.cdc.gov/nchs/covid19/covid-19-mortality-data-files.htm