Name: Gia Phat Le

Student ID: 119708238

Data Mining_Final Project

Source: https://www.kaggle.com/datasets/jainaru/covid-19-deaths-predict-conditions-contributing

Dataset description:

This dataset provides insights into the health conditions and contributing causes associated with deaths of COVID-19 in the United States from 2020 to 2023. In total, it has 13 columns and 621000 rows. The attributes include:

1. **Start Date**: The starting date of the data collection period.
2. **End Date**: The ending date of the data collection period.
3. **Group**: The categorization of the data (e.g., "By Total" for aggregated data across all subcategories).
4. **Year**: The specific year of the data record.
5. **Month**: The specific month of the data record.
6. **State**: The U.S. state where the data was collected. "United States" indicates national-level data.
7. **Condition Group**: The broader category of health conditions (e.g., "respiratory diseases").
8. **Condition**: The specific health condition mentioned in conjunction with COVID-19 deaths (e.g., "Influenza and pneumonia").
9. **ICD10_codes**: The ICD-10 codes associated with the health condition (e.g., "J09-J18" for influenza and pneumonia).
10. **Age Group**: The age range of individuals in this data record (e.g., "0-24", "25-34").

11. **COVID-19 Deaths**: The number of deaths where COVID-19 was mentioned as a cause.
12. **Number of Mentions**: The number of times the specific condition was mentioned in conjunction with COVID-19 deaths.
13. **Flag**: Highlighting if there are any special conditions or exceptions that apply to the data entry.

Business problem point out:

How can we identify trends in COVID-19 deaths to improve resource allocation and public health interventions?

R studio code:

Step 1: Dataset Description

 # Load necessary libraries

```
install.packages(c("dplyr", "ggplot2", "readr", "tidyr", "knitr"))

library(dplyr)

library(ggplot2)

library(readr)

library(tidyr)
```

First, I loaded these necessary libraries: dplyr for data manipulation, ggplot for creating plot as the visualization step to have a depth understanding of our dataset, readr for reading CSV files, tidyr for reshaping data.

 # Load the dataset

```
covid_data <- read.csv("C:/Users/giaph/OneDrive/Máy tính/Data mining project/Conditions_Contributing_to_COVID-19_Deaths__by_State_and_Age__Provisional_2020-2023.csv")
```

Code_Project.R | Code_Project.R | covid_data

Filter

	Start.Date	End.Date	Group	Year	Month	State	Condition.Group	Condition	ICD10_codes
1	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Influenza and pneumonia	J09-J18
2	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Influenza and pneumonia	J09-J18
3	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Influenza and pneumonia	J09-J18
4	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Influenza and pneumonia	J09-J18
5	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Influenza and pneumonia	J09-J18
6	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Influenza and pneumonia	J09-J18
7	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Influenza and pneumonia	J09-J18
8	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Influenza and pneumonia	J09-J18
9	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Influenza and pneumonia	J09-J18
10	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Influenza and pneumonia	J09-J18
11	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Chronic lower respiratory diseases	J40-J47
12	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Chronic lower respiratory diseases	J40-J47
13	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Chronic lower respiratory diseases	J40-J47
14	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Chronic lower respiratory diseases	J40-J47
15	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Chronic lower respiratory diseases	J40-J47
16	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Chronic lower respiratory diseases	J40-J47
17	1/1/2020	9/23/2023	By Total	NA	NA	United States	Respiratory diseases	Chronic lower respiratory diseases	J40-J47

 # View the structure of the dataset

str(covid_data)

```
> str(covid_data)
'data.frame':   621000 obs. of  13 variables:
 $ Start.Date        : chr  "1/1/2020" "1/1/2020" "1/1/2020" "1/1/2020" ...
 $ End.Date          : chr  "9/23/2023" "9/23/2023" "9/23/2023" "9/23/2023" ...
 $ Group             : chr  "By Total" "By Total" "By Total" "By Total" ...
 $ Year              : int  NA NA NA NA NA NA NA NA NA NA ...
 $ Month             : int  NA NA NA NA NA NA NA NA NA NA ...
 $ State             : chr  "United States" "United States" "United States" "United States" ...
 $ Condition.Group   : chr  "Respiratory diseases" "Respiratory diseases" "Respiratory diseases" "Respiratory dis
eases" ...
 $ Condition         : chr  "Influenza and pneumonia" "Influenza and pneumonia" "Influenza and pneumonia" "Influe
nza and pneumonia" ...
 $ ICD10_codes       : chr  "J09-J18" "J09-J18" "J09-J18" "J09-J18" ...
 $ Age.Group         : chr  "0-24" "25-34" "35-44" "45-54" ...
 $ COVID.19.Deaths   : int  1569 5804 15080 37414 82668 129005 138503 121119 12 531174 ...
 $ Number.of.Mentions: int  1647 6029 15699 38878 85708 133088 141868 123018 12 545947 ...
 $ Flag              : chr  "" "" "" "" ...
>
```

This line displays the structure of the dataset (data types and dimensions)

head(covid_data)

```
> head(covid_data)
  Start.Date  End.Date    Group Year Month         State      Condition.Group                Condition
1   1/1/2020 9/23/2023 By Total   NA    NA United States Respiratory diseases Influenza and pneumonia
2   1/1/2020 9/23/2023 By Total   NA    NA United States Respiratory diseases Influenza and pneumonia
3   1/1/2020 9/23/2023 By Total   NA    NA United States Respiratory diseases Influenza and pneumonia
4   1/1/2020 9/23/2023 By Total   NA    NA United States Respiratory diseases Influenza and pneumonia
5   1/1/2020 9/23/2023 By Total   NA    NA United States Respiratory diseases Influenza and pneumonia
6   1/1/2020 9/23/2023 By Total   NA    NA United States Respiratory diseases Influenza and pneumonia
  ICD10_codes Age.Group COVID.19.Deaths Number.of.Mentions Flag
1     J09-J18      0-24            1569               1647
2     J09-J18     25-34            5804               6029
3     J09-J18     35-44           15080              15699
4     J09-J18     45-54           37414              38878
5     J09-J18     55-64           82668              85708
6     J09-J18     65-74          129005             133088
> |
```

This line show the first 6 rows of the dataset

tail(covid_data)

```
> tail(covid_data)
       Start.Date  End.Date    Group Year Month       State Condition.Group Condition ICD10_codes Age.Group
620995   4/1/2023 4/30/2023 By Month 2023     4 Puerto Rico        COVID-19  COVID-19        U071  All Ages
620996   5/1/2023 5/31/2023 By Month 2023     5 Puerto Rico        COVID-19  COVID-19        U071  All Ages
620997   6/1/2023 6/30/2023 By Month 2023     6 Puerto Rico        COVID-19  COVID-19        U071  All Ages
620998   7/1/2023 7/31/2023 By Month 2023     7 Puerto Rico        COVID-19  COVID-19        U071  All Ages
620999   8/1/2023 8/31/2023 By Month 2023     8 Puerto Rico        COVID-19  COVID-19        U071  All Ages
621000   9/1/2023 9/23/2023 By Month 2023     9 Puerto Rico        COVID-19  COVID-19        U071  All Ages
       COVID.19.Deaths Number.of.Mentions Flag
620995              44                 44
620996              67                 67
620997             122                122
620998             114                114
620999              78                 78
621000              36                 36
> |
```

This line shows the last 6 rows of the dataset

summary(covid_data)

```
> summary(covid_data)
  Start.Date         End.Date           Group                Year          Month          State
 Length:621000      Length:621000      Length:621000      Min.   :2020   Min.   : 1.0   Length:621000
 Class :character   Class :character   Class :character   1st Qu.:2020   1st Qu.: 3.0   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median :2021   Median : 6.0   Mode  :character
                                                          Mean   :2021   Mean   : 6.2
                                                          3rd Qu.:2022   3rd Qu.: 9.0
                                                          Max.   :2023   Max.   :12.0
                                                          NA's   :12420  NA's   :62100

 Condition.Group     Condition          ICD10_codes         Age.Group          COVID.19.Deaths
 Length:621000      Length:621000      Length:621000      Length:621000      Min.   :      0.0
 Class :character   Class :character   Class :character   Class :character   1st Qu.:      0.0
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :      0.0
                                                                            Mean   :    120.1
                                                                            3rd Qu.:     18.0
                                                                            Max.   :1146242.0
                                                                            NA's   :183449

 Number.of.Mentions      Flag
 Min.   :      0.0   Length:621000
 1st Qu.:      0.0   Class :character
 Median :      0.0   Mode  :character
 Mean   :    129.3
 3rd Qu.:     19.0
 Max.   :1146242.0
 NA's   :177577
```

This line provides descriptive statistics of the numeric columns (min, max, mean, quantile) and frequency counts for the categorical columns

Step 2: Data Cleaning and Preprocessing

 # Check for missing values

missing_values <- colSums(is.na(covid_data))

print(missing_values)

```
> missing_values <- colSums(is.na(covid_data))
> print(missing_values)
       Start.Date          End.Date             Group              Year             Month
                0                 0                 0             12420             62100
            State    Condition.Group         Condition       ICD10_codes         Age.Group
                0                 0                 0                 0                 0
  COVID.19.Deaths Number.of.Mentions              Flag
           183449            177577                 0
>
```

In this part, I check for the missing values per each column

 # Remove rows with missing values

covid_data_cleaned <- na.omit(covid_data)

The missing values are going to be removed in order to keep the consistency for further steps, and then the cleaned data gonna be named as ‘covid_data_cleaned’

 # Convert Year and Month columns to integers

covid_data_cleaned$Year <- as.integer(covid_data_cleaned$Year)

covid_data_cleaned$Month <- as.integer(covid_data_cleaned$Month)

Then, for this part I converted both year and month into integer data types

 # Check for remaining missing values

colSums(is.na(covid_data_cleaned))

```
> colSums(is.na(covid_data_cleaned))
       Start.Date          End.Date             Group              Year             Month
                0                 0                 0                 0                 0
            State    Condition.Group         Condition      ICD10_codes         Age.Group
                0                 0                 0                 0                 0
  COVID.19.Deaths Number.of.Mentions              Flag
                0                 0                 0
>
```

This step mainly to ensure there is no remaining missing values in the covid_data_cleaned dataset

 # Calculate IQR for COVID-19 deaths

Q1 <- quantile(covid_data_cleaned$COVID.19.Deaths, 0.25, na.rm = TRUE)

Q3 <- quantile(covid_data_cleaned$COVID.19.Deaths, 0.75, na.rm = TRUE)

IQR_value <- Q3 - Q1

```
> IQR_value
75%
 12
>
```

Because the death column is going to the main attribute to do further investigation, so in this step I define Q1 (25th quantile) and Q3 (75th quantile) and calculate the IQR value. As we see here, IQR = 12

 # Define outlier thresholds

lower_bound <- Q1 - 1.5 * IQR_value

upper_bound <- Q3 + 1.5 * IQR_value

```
> upper_bound
75%
 30
> lower_bound
25%
-18
> |
```

Followed with that, the upper bound and lower bound are defined by this above formula. Lower bound = -18 and Upper bound = 30. It means that any values out of this range (lower than -18 or higher than 30) will be detected as outliers in the Death attribute.

 # Remove outliers

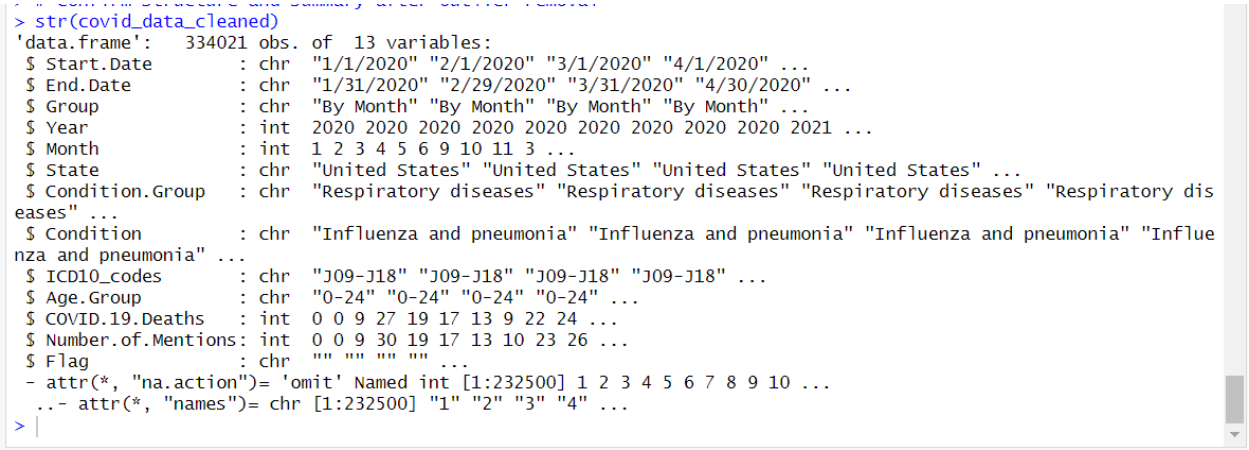
covid_data_cleaned <- covid_data_cleaned %>%

filter(COVID.19.Deaths >= lower_bound & COVID.19.Deaths <= upper_bound)

Next, I updated the covid_data_cleaned dataset by removing the outliers as I describe in the above step.

 # Confirm structure and summary after outlier removal

str(covid_data_cleaned)



```
> str(covid_data_cleaned)
'data.frame':   334021 obs. of  13 variables:
 $ Start.Date        : chr  "1/1/2020" "2/1/2020" "3/1/2020" "4/1/2020" ...
 $ End.Date          : chr  "1/31/2020" "2/29/2020" "3/31/2020" "4/30/2020" ...
 $ Group             : chr  "By Month" "By Month" "By Month" "By Month" ...
 $ Year              : int  2020 2020 2020 2020 2020 2020 2020 2020 2020 2021 ...
 $ Month             : int  1 2 3 4 5 6 9 10 11 3 ...
 $ State             : chr  "United States" "United States" "United States" "United States" ...
 $ Condition.Group   : chr  "Respiratory diseases" "Respiratory diseases" "Respiratory diseases" "Respiratory dis
eases" ...
 $ Condition         : chr  "Influenza and pneumonia" "Influenza and pneumonia" "Influenza and pneumonia" "Influe
nza and pneumonia" ...
 $ ICD10_codes       : chr  "J09-J18" "J09-J18" "J09-J18" "J09-J18" ...
 $ Age.Group         : chr  "0-24" "0-24" "0-24" "0-24" ...
 $ COVID.19.Deaths   : int  0 0 9 27 19 17 13 9 22 24 ...
 $ Number.of.Mentions: int  0 0 9 30 19 17 13 10 23 26 ...
 $ Flag              : chr  "" "" "" "" ...
 - attr(*, "na.action")= 'omit' Named int [1:232500] 1 2 3 4 5 6 7 8 9 10 ...
  ..- attr(*, "names")= chr [1:232500] "1" "2" "3" "4" ...
> |
```

summary(covid_data_cleaned)

```
> summary(covid_data_cleaned)
  Start.Date          End.Date            Group               Year          Month           State
 Length:334021      Length:334021      Length:334021      Min.   :2020   Min.   : 1.000   Length:334021
 Class :character   Class :character   Class :character   1st Qu.:2020   1st Qu.: 3.000   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median :2021   Median : 6.000   Mode  :character
                                                          Mean   :2021   Mean   : 6.032
                                                          3rd Qu.:2022   3rd Qu.: 9.000
                                                          Max.   :2023   Max.   :12.000
 Condition.Group     Condition          ICD10_codes         Age.Group          COVID.19.Deaths
 Length:334021      Length:334021      Length:334021      Length:334021      Min.   : 0.000
 Class :character   Class :character   Class :character   Class :character   1st Qu.: 0.000
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 0.000
                                                                            Mean   : 2.992
                                                                            3rd Qu.: 0.000
                                                                            Max.   :30.000

 Number.of.Mentions      Flag
 Min.   : 0.000     Length:334021
 1st Qu.: 0.000     Class :character
 Median : 0.000     Mode  :character
 Mean   : 3.296
 3rd Qu.: 0.000
 Max.   :83.000
```

In this part, I just basically recheck the structure and statistics of the cleaned dataset.

Step 3: Exploratory Data Analysis

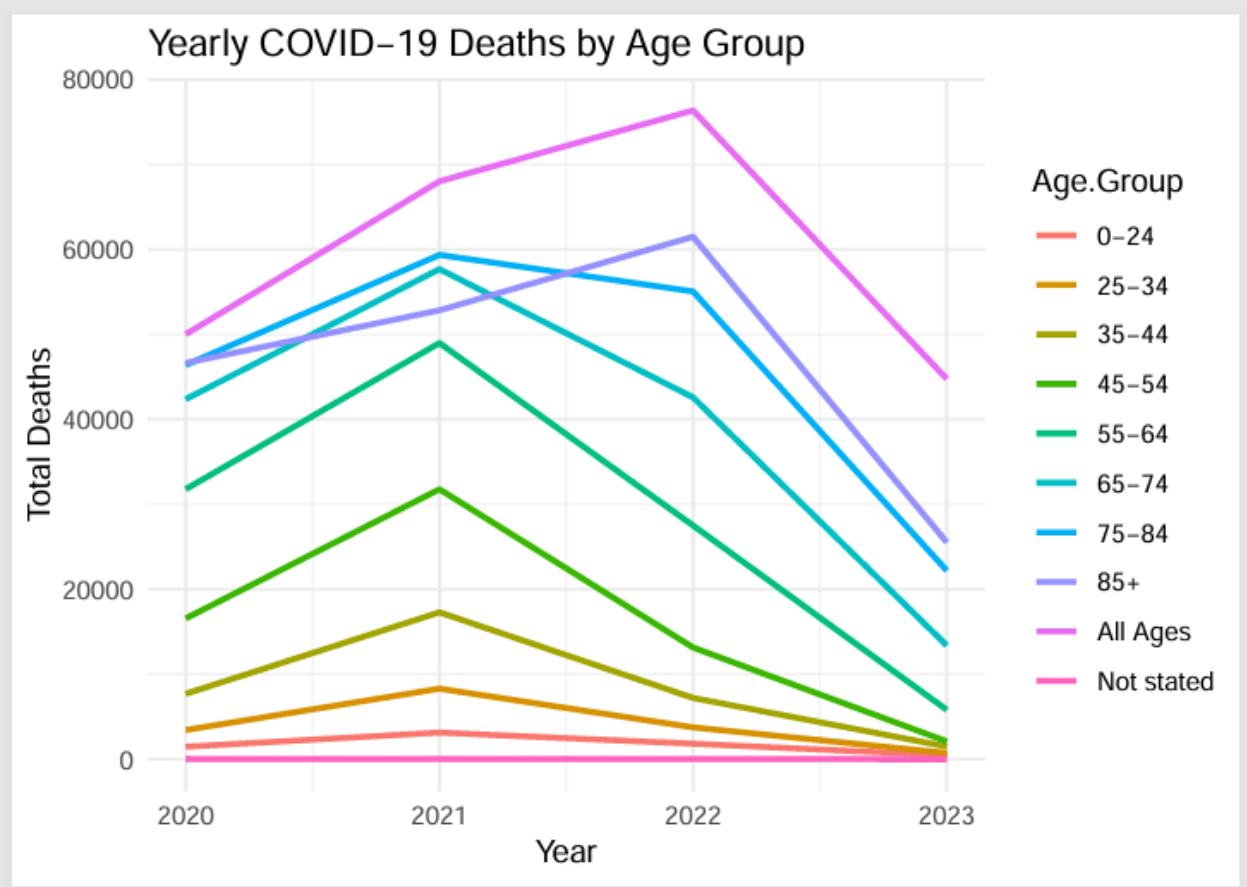
a) Yearly deaths by age group:

```
yearly_deaths_by_age <- covid_data_cleaned %>%
  group_by(Year, Age.Group) %>%
  summarise(Total_Deaths = sum(COVID.19.Deaths, na.rm = TRUE))


ggplot(yearly_deaths_by_age, aes(x = Year, y = Total_Deaths, color = Age.Group)) +
  geom_line(size = 1) +
  labs(title = "Yearly COVID-19 Deaths by Age Group",
       x = "Year",
       y = "Total Deaths") +
  theme_minimal()
```

	Year	Age.Group	Total_Deaths
1	2020	0-24	1472
2	2020	25-34	3437
3	2020	35-44	7705
4	2020	45-54	16574
5	2020	55-64	31787
6	2020	65-74	42377
7	2020	75-84	46406
8	2020	85+	46638
9	2020	All Ages	50023
10	2020	Not stated	45
11	2021	0-24	3136
12	2021	25-34	8306
13	2021	35-44	17291
14	2021	45-54	31777
15	2021	55-64	48985
16	2021	65-74	57688

	Year	Age.Group	Total_Deaths
16	2021	65-74	57688
17	2021	75-84	59357
18	2021	85+	52838
19	2021	All Ages	68019
20	2021	Not stated	27
21	2022	0-24	1827
22	2022	25-34	3752
23	2022	35-44	7216
24	2022	45-54	13150
25	2022	55-64	27510
26	2022	65-74	42589
27	2022	75-84	55057
28	2022	85+	61497
29	2022	All Ages	76363
30	2022	Not stated	19
31	2023	0-24	465
32	2023	25-34	734
33	2023	35-44	1493
34	2023	45-54	2096
35	2023	55-64	5817
36	2023	65-74	13398
37	2023	75-84	22215
38	2023	85+	25536
39	2023	All Ages	44768
40	2023	Not stated	0



The 1st method of visualization is done by grouping Year and Age.Group then calculate the total deaths for each group as the plot showing here. The x-axis represents the year (from 2020 to 2023) and the y-axis represents the total number of deaths. Each line color represents the specific age-group. The obvious trend over time we could see here is that most of the age group deaths peak in a specific year (could be in 2020 or 2021) before declining. The elder's groups show significantly higher number of total death compared to the younger generation groups. Younger groups like 0-24 and 25-34 years old show consistently lower death rates, when the number of death is less than 10000 even at their peak. This detail reflects a fact, older people have higher chance of death due to their immune system decline associated with underlying disease due to age.

b) Top 10 states for a specific condition:

```
state_condition_deaths <- covid_data_cleaned %>%
  filter(Condition.Group == "Respiratory diseases") %>%
  group_by(State) %>%
```

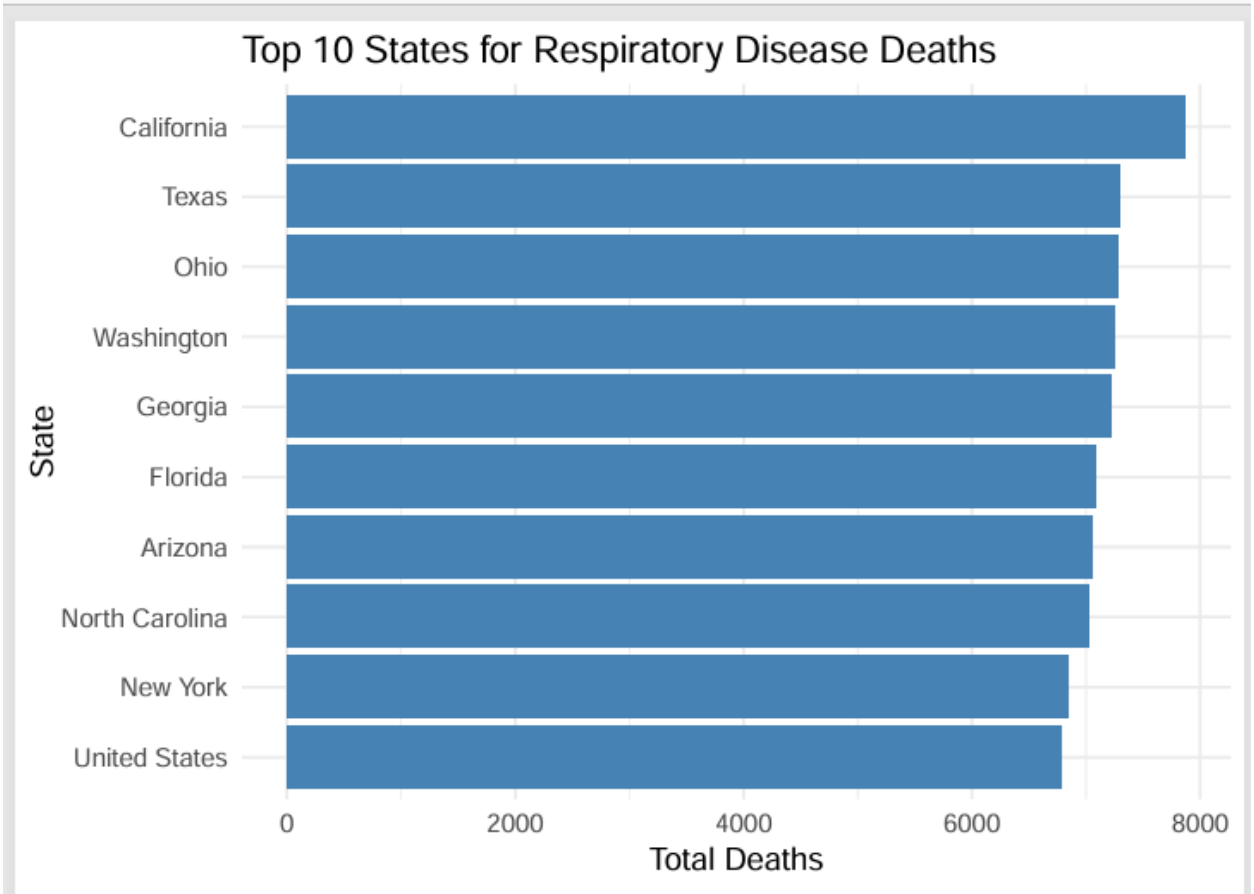
```
  summarise(Total_Deaths = sum(COVID.19.Deaths, na.rm = TRUE)) %>%
  arrange(desc(Total_Deaths))

top_states <- head(state_condition_deaths, 10)

ggplot(top_states, aes(x = reorder(State, Total_Deaths), y = Total_Deaths)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(title = "Top 10 States for Respiratory Disease Deaths",
       x = "State",
       y = "Total Deaths") +
  theme_minimal()
```

Filter

	State	Total_Deaths
1	California	7873
2	Texas	7300
3	Ohio	7279
4	Washington	7259
5	Georgia	7219
6	Florida	7090
7	Arizona	7060
8	North Carolina	7021
9	New York	6844
10	United States	6791

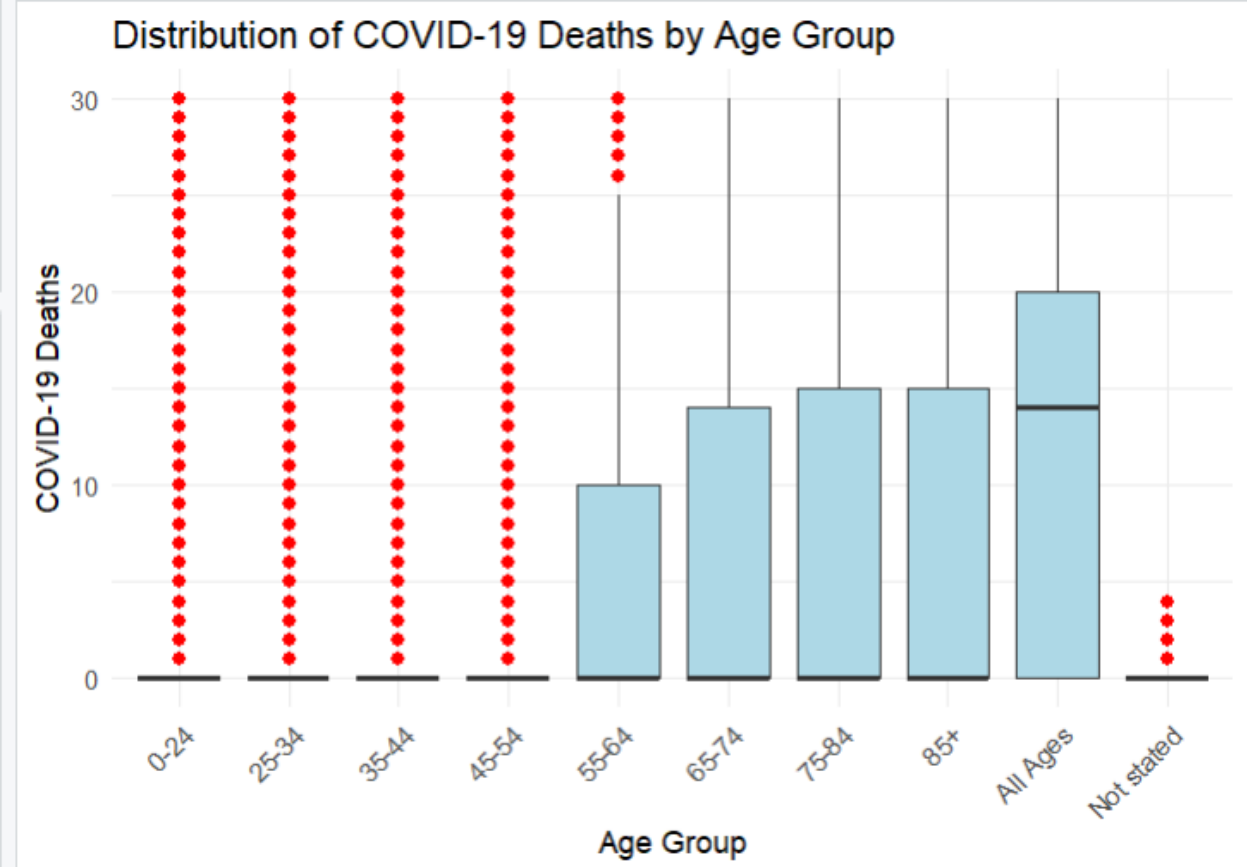


The 2nd method is a horizontal bar chart displays the top 10 states with the highest total of death due to respiratory diseases. X axis represents the total deaths while the y axis represents each state order by the number of deaths. There is not a big gap between the highest state (California) and the lowest one (New York). States like California or Texas exhibit the highest total of deaths which reflects their large populations. While New York and North Carolina show relatively fewer than the top states above, however it still a significant amount. The reason I choose this condition to compared between states because this is the main cause of COVID-19 spread at that time, that why people are encourage to wear mask and disinfect regularly by the goverment at the peak of the pandemic.

c) Distribution of Deaths Across Age Groups

```
ggplot(covid_data_cleaned, aes(x = Age.Group, y = COVID.19.Deaths)) +
  geom_boxplot(fill = "lightblue", outlier.color = "red", outlier.size = 2) +
  labs(
```

```
  title = "Distribution of COVID-19 Deaths by Age Group",
  x = "Age Group",
  y = "COVID-19 Deaths"
) +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



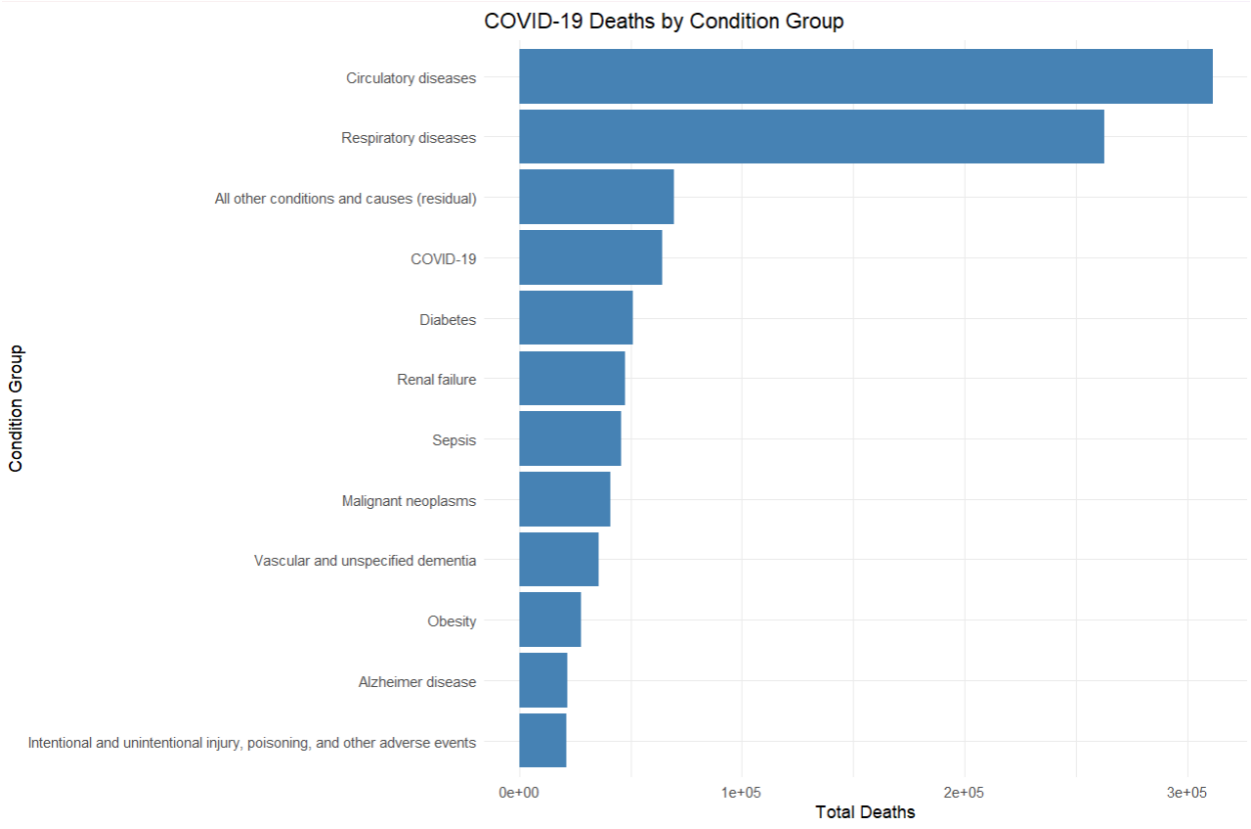
This box plot here also shows the spread of COVID-19 deaths for each age group. X-axis represents the age groups while the Y-axis represents the number of deaths. The outliers are the extreme values which are plotted as red points in this plot, the line in the middle of each box is known as the median. The median number of deaths increases with age, indicating that elderly people were likely to be affected. Also, the ages group 85+ has a wider range of deaths, might be due to variations in regional healthcare resources or other factors.

d) COVID-19 Deaths by Condition Group

```
condition_group_deaths <- covid_data_cleaned %>%
  group_by(Condition.Group) %>%
  summarise(Total_Deaths = sum(COVID.19.Deaths, na.rm = TRUE)) %>%
  arrange(desc(Total_Deaths))


ggplot(condition_group_deaths, aes(x = reorder(Condition.Group, Total_Deaths), y =
Total_Deaths)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(
    title = "COVID-19 Deaths by Condition Group",
    x = "Condition Group",
    y = "Total Deaths"
  ) +
  theme_minimal()
```

	Condition.Group	Total_Deaths
1	Circulatory diseases	311561
2	Respiratory diseases	262687
3	All other conditions and causes (residual)	69540
4	COVID-19	64384
5	Diabetes	50838
6	Renal failure	47615
7	Sepsis	45674
8	Malignant neoplasms	40966
9	Vascular and unspecified dementia	35442
10	Obesity	27612
11	Alzheimer disease	21710
12	Intentional and unintentional injury, poisoning, and other a...	21361



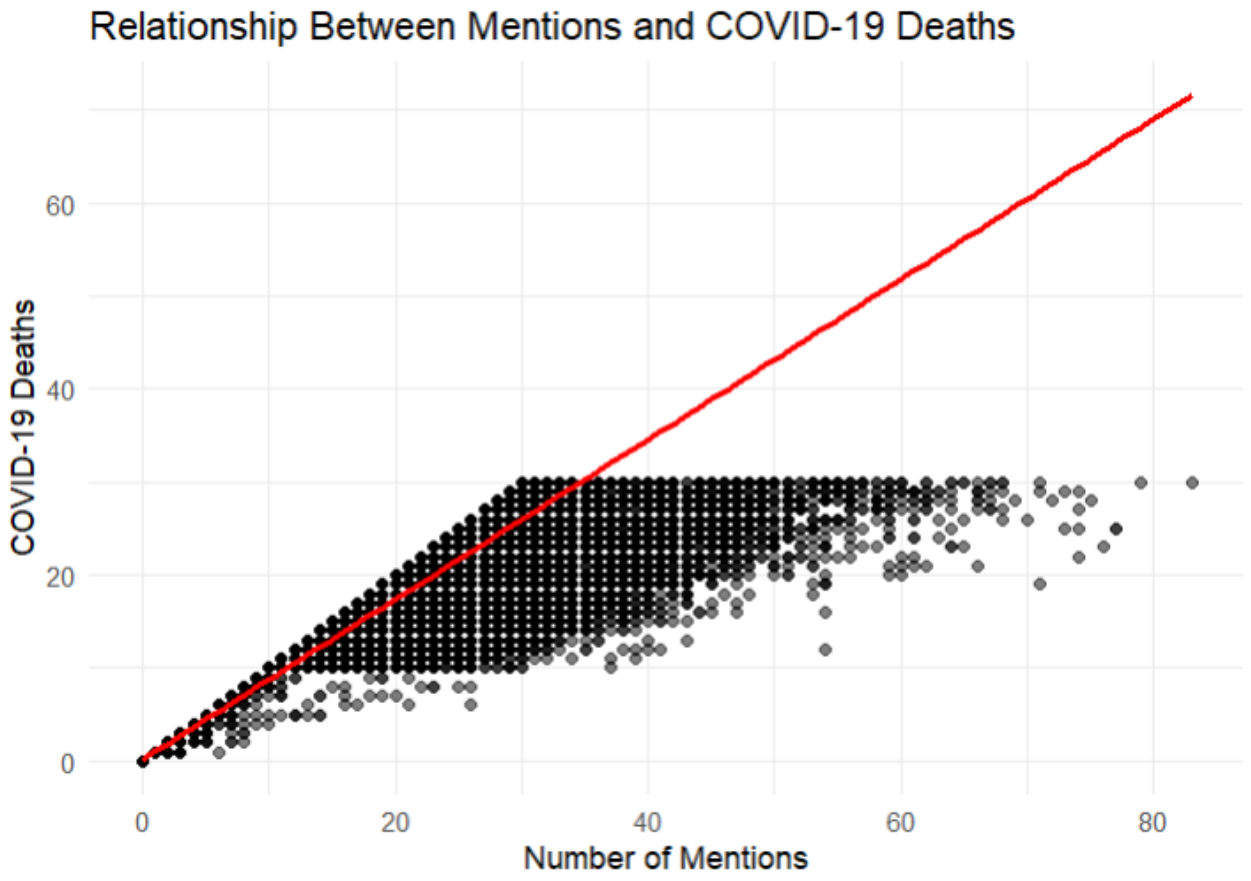
This horizontal bar chart compares the total deaths associated with different condition groups. The X axis represents the total number of deaths while the Y axis represent various condition groups. As we can see that ‘Circulatory diseases’ and ‘respiratory diseases’ are the leading contributors to COVID-19 deaths, create a big gap compares to other conditions in the groups with a significant higher total. Conditions like ‘Obesity’ and ‘Alzheimer’s disease’ show relatively fewer deaths.

Step 4: Correlation Analysis

Analyze the relationship between condition mentions and deaths.

```
ggplot(covid_data_cleaned, aes(x = Number.of.Mentions, y = COVID.19.Deaths)) +
  geom_point(alpha = 0.5) +
  labs(title = "Relationship Between Mentions and COVID-19 Deaths",
       x = "Number of Mentions",
       y = "COVID-19 Deaths") +
  theme_minimal()

cor(covid_data_cleaned$Number.of.Mentions, covid_data_cleaned$COVID.19.Deaths,
use = "complete.obs")
```




```
> cor(covid_data_cleaned$Number.of.Mentions, covid_data_cleaned$COVID.19.Deaths, use = "complete.obs")
[1] 0.9735962
```

This scatter plot right here visualizes the relationship between number of mentions (x-axis) and COVID-19 deaths (y-axis). There is probably a visible positive relationship between these two variables. As we see, when the x increases the number of y also tends to increase. Therefore, there is a possible correlation between COVID-19 deaths and Number of mentions. There are some outliers appear to be in isolated points on the top-right where both values are high, these may indicate extreme cases. The data points are clustered and form a triangle pattern as we see here, means that for higher mentions there is a wider range of death counts. This may indicate variability or inconsistency at higher mention levels.

As the correlation coefficient value = 0.97 which is close to +1 means that these two variables are highly linearly related. So, in conclusion, regions, time periods or categories with higher mentions (through news, social media) are associated with higher reported deaths from COVID-19.

Business problem possible solutions:

As we could see through the dataset and the visualization analytics, the causes of COVID-19 related mainly to these 3 factors: medical condition, age groups and geographic locations. So, there are a few solutions to improve resource allocation and public health:

1. Increase resources in regions with high prevalence of circulatory or respiratory diseases because they are contributing the most death proportion through the above graphs. Also, high elderly populations need to be priority taken care by the healthcare system associate with the government.
2. Focus on mitigating chronic diseases through awareness campaigns and healthcare accessibility.
3. Improve the citizen's awareness by encourage them to follow the trends through social media or news to predict emerging hotspots and mobilize resources immediately.
4. Adjust public health interventions to the specific needs of geographic areas based on disease prevalence and resource deficits (such as California, Ohio, Texas are the ones need more medical support proved by data).