Portfolio Project #1

COVID-19 Deaths by Medical Condition

Data Source:

https://healthdata.gov/dataset/Conditions-Contributing-to-COVID-19-Deaths-by-Stat/uvkj-kpue/about_data

Questions:

- Which age group has the most deaths caused by COVID-19?
- Which condition has the most deaths? Condition group?
- Which condition leads to the most deaths by age group?

Key Variables:

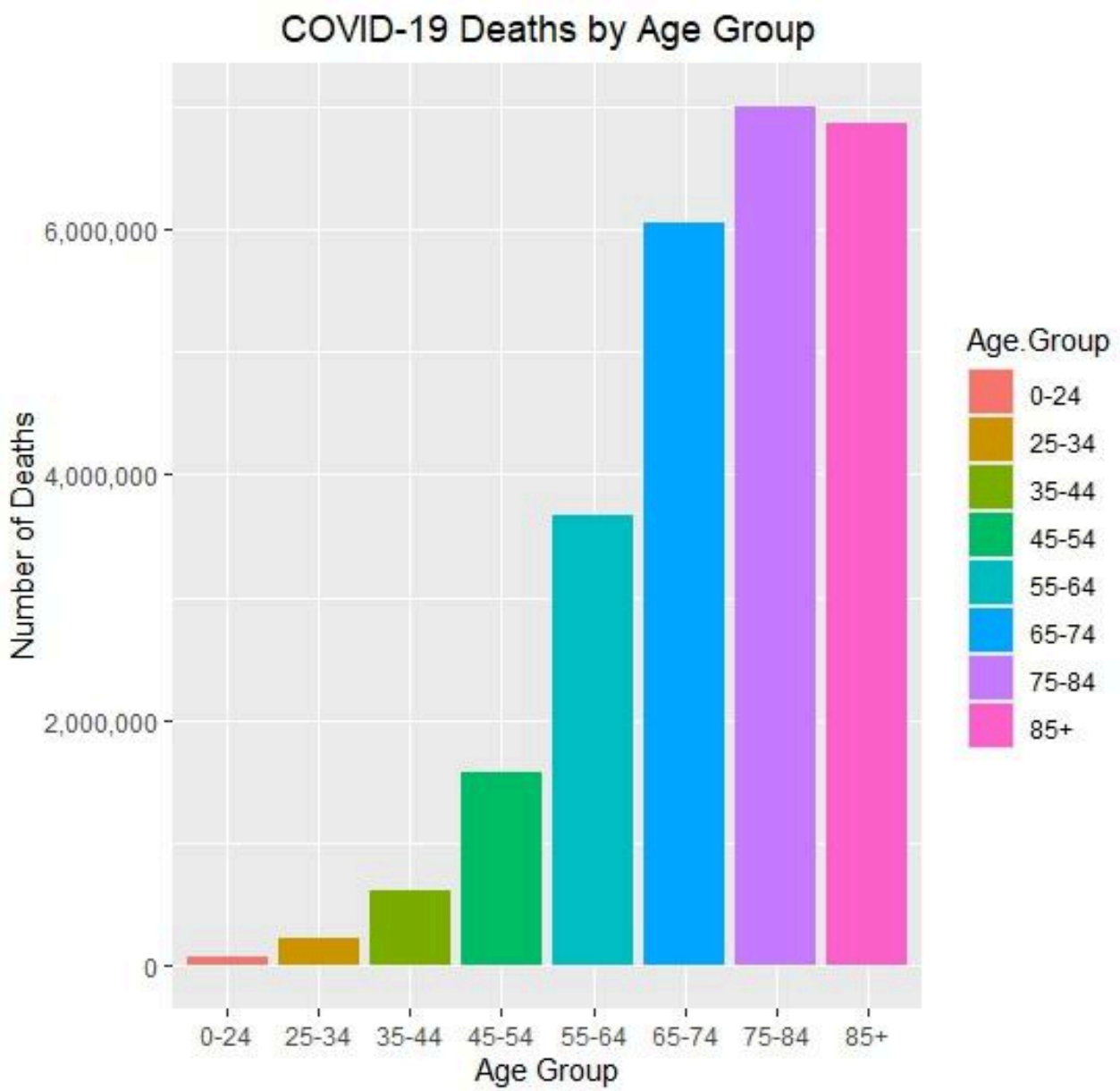
- Condition.group
- Condition
- Age.Group
- COVID.19.Deaths

Which age group has the most deaths caused by COVID-19?

1. First filter the data so that only Age.Group and COVID.19.Deaths are left in dataset
2. Combine all Age groups by various conditions together into total count
3. Plot bar graph representing each age group on x-axis and number of deaths on y-axis

CODE:

```
#Installing tidyverse package
install.packages("tidyverse")
library(tidyverse)

#Reading CSV
Covid19_data <-
read.csv("D://Conditions_Contributing_to_COVID-19_Deaths__by_State_and_Age__Provisional_2020-2023.csv")

#Filtering data to only have Age.Group and COVID.19.Death Variables
Question_1 <- Covid19_data %>% select(Age.Group, COVID.19.Deaths)

#Combining number of deaths by age group
Grouped_Question_1 <- Question_1 %>% filter(!is.na(COVID.19.Deaths)) %>% group_by(Age.Group) %>%
summarise(sumDeaths = sum(COVID.19.Deaths))

#Cleaning data to exclude ‘All Ages’ and ‘Not Stated’ as they are not relevant
Clean_Grouped_Question_1 <- Grouped_Question_1[-c(9,10), ]

#Plotting data on a bar chart
#Adding Color, Title (Centered), X-axis Label, Y-axis Label
#Changing Y-axis scale to display comma number instead of exponent
ggplot(Clean_Grouped_Question_1, aes(x=Age.Group, y=sumDeaths, fill=Age.Group)) + geom_bar(stat="identity") +
scale_y_continuous(labels = scales::comma) + labs(title="COVID-19 Deaths by Age Group", x="Age Group",
y="Number of Deaths") + theme(plot.title = element_text(hjust=0.5))
```


COVID-19 Deaths by Age Group
Number of Deaths
6,000,000
4,000,000
2,000,000
0
0-24
25-34
35-44
45-54
55-64
65-74
75-84
85+
Age Group
Age.Group
0-24
25-34
35-44
45-54
55-64
65-74
75-84
85+

Which condition has the most deaths? Condition group?

1. Group together the number of deaths in each condition (combining all age groups)
2. Plot bar graph representing each condition on x-axis and number of deaths on y-axis
3. Repeat with condition group

CODE: Condition

#Installing tidyverse package

```
install.packages("tidyverse")
library(tidyverse)
```

#Reading CSV

```
Covid19_data <-
read.csv("D://Conditions_Contributing_to_COVID-19_Deaths__by_State_and_Age__Provisional_2020-2023.csv")
```

#Filtering data to only have Condition and COVID.19.Death Variables

```
Question_2 <- Covid19_data %>% select(Condition, COVID.19.Deaths)
```

#Combining number of deaths by Condition

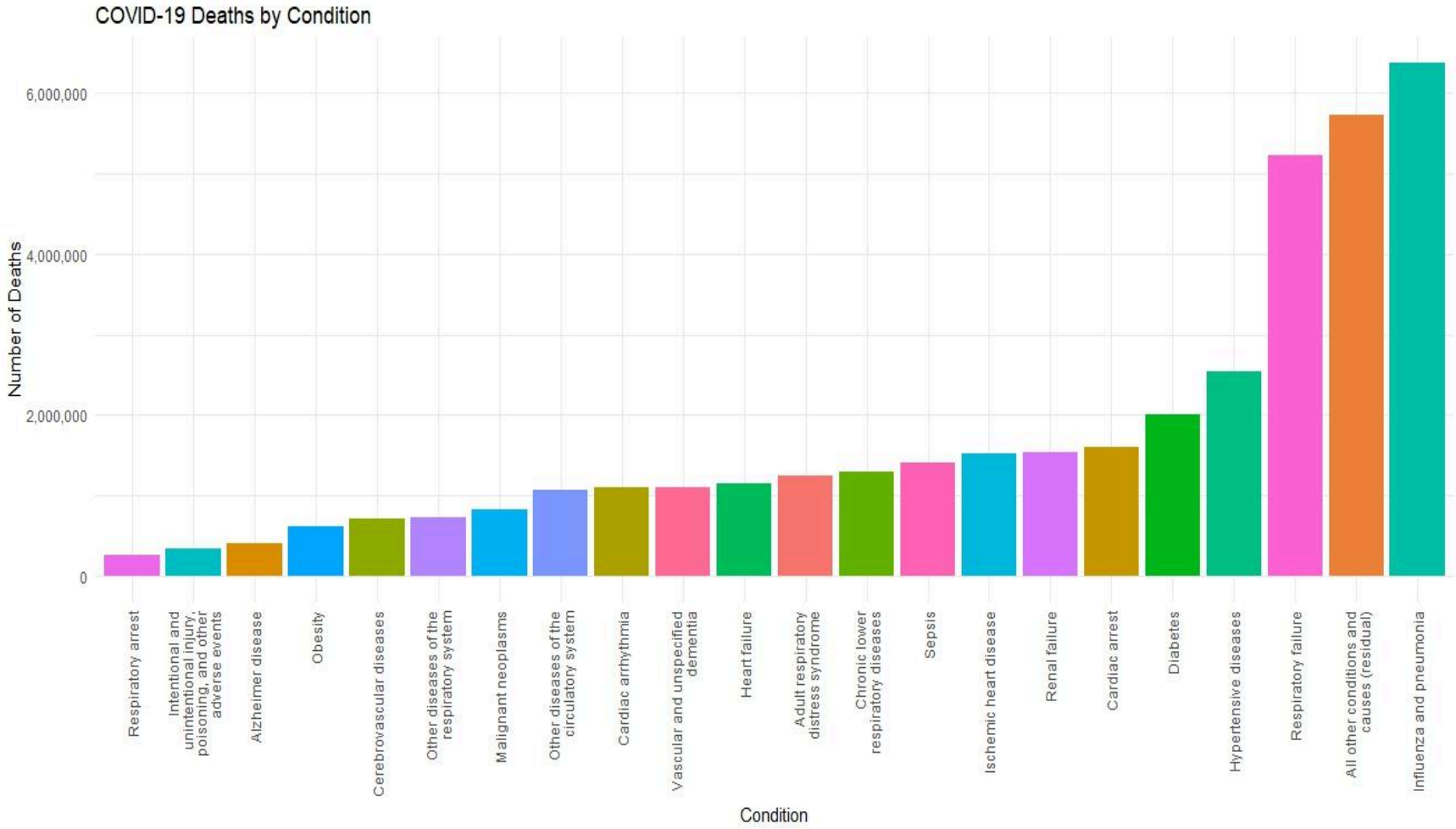
```
Grouped_Question_2 <- Question_2 %>% filter(!is.na(COVID.19.Deaths)) %>% group_by(Condition) %>%
summarise(sumDeaths = sum(COVID.19.Deaths))
```

#Cleaning data to exclude COVID-19 as it is not relevant

```
Clean_Grouped_Question_2 <- Grouped_Question_2[-c(4), ]
```

#Plotting the data
#Adding Color, Title, X-axis Label, Y-axis Label
#Changing Y-axis scale to display comma number instead of exponent
#Wrapping x-axis labels at 25 characters

```
ggplot(Clean_Grouped_Question_2, aes(x=reorder(Condition,sumDeaths), y=sumDeaths, fill=Condition)) +
geom_bar(stat="identity") + scale_y_continuous(labels = scales::comma) + labs(title="COVID-19 Deaths by
Condition", x="Condition", y="Number of Deaths") + theme_minimal() + theme(legend.position = "none") +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + scale_x_discrete(labels =
label_wrap(25))
```



CODE: Condition Group

#Installing tidyverse package

```
install.packages("tidyverse")
library(tidyverse)
```

#Reading CSV

```
Covid19_data <-
read.csv("D://Conditions_Contributing_to_COVID-19_Deaths__by_State_and_Age__Provisional_2020-2023.csv")
```

#Filtering data to only have Condition Group and COVID.19.Death Variables

```
Question_2_Group <- Covid19_data %>% select(Condition.Group, COVID.19.Deaths)
```

#Combining number of deaths by Condition

```
Grouped_Question_2_Group <- Question_2_Group %>% filter(!is.na(COVID.19.Deaths)) %>% group_by(Condition.Group)
%>% summarise(sumDeaths = sum(COVID.19.Deaths))
```

#Cleaning data to exclude COVID-19 as it is not relevant

```
Clean_Grouped_Question_2_Group <- Grouped_Question_2_Group[-c(3), ]
```

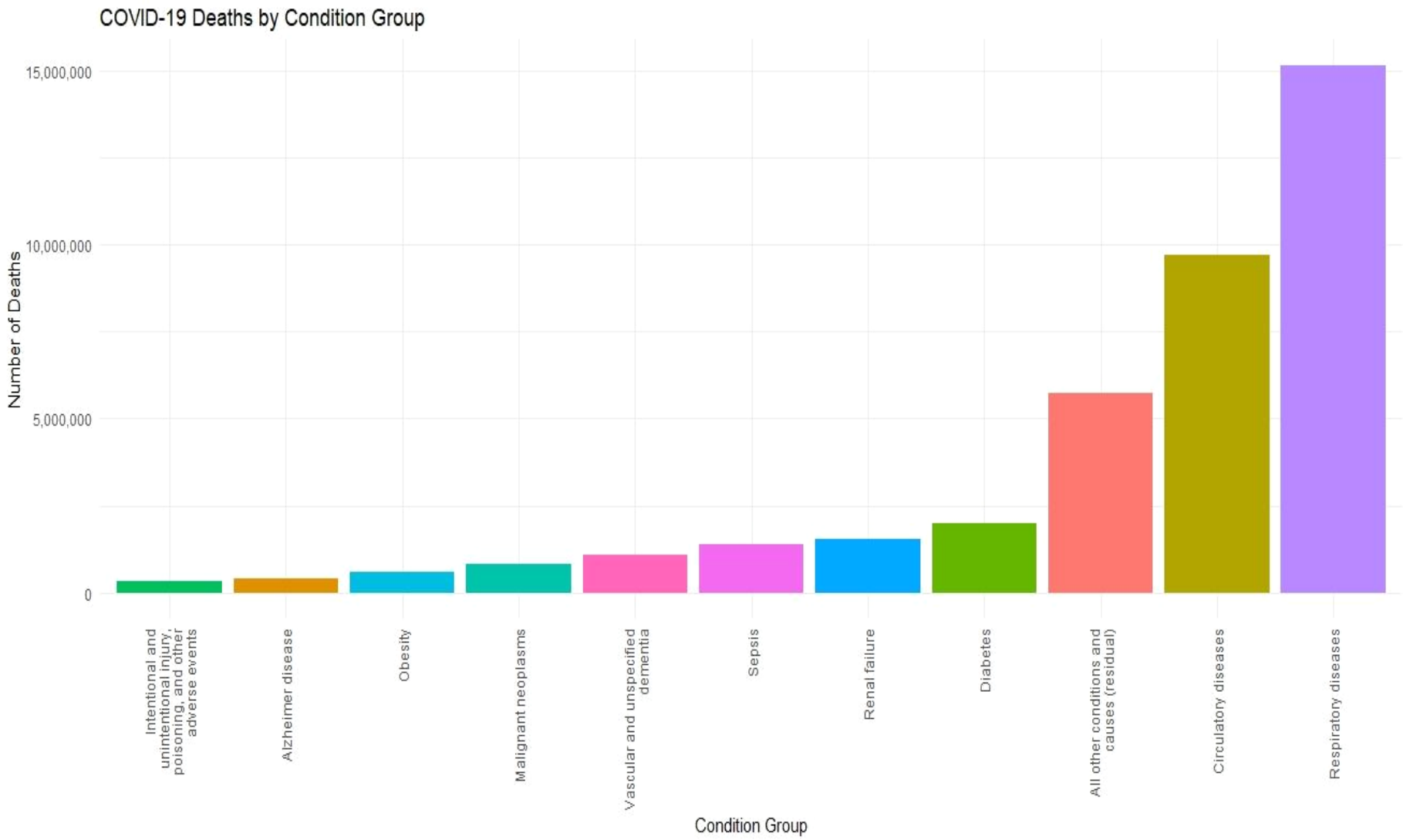
#Plotting the data

#Adding Color, Title, X-axis Label, Y-axis Label

#Changing Y-axis scale to display comma number instead of exponent

#Wrapping x-axis labels at 25 characters

```
ggplot(Clean_Grouped_Question_2_Group, aes(x=reorder(Condition.Group,sumDeaths), y=sumDeaths,
fill=Condition.Group)) + geom_bar(stat="identity") + scale_y_continuous(labels = scales::comma) +
labs(title="COVID-19 Deaths by Condition Group", x="Condition Group", y="Number of Deaths") +
theme_minimal() + theme(legend.position = "none") + theme(axis.text.x = element_text(angle = 90, vjust =
0.5, hjust=1)) + scale_x_discrete(labels = label_wrap(25))
```


COVID-19 Deaths by Condition Group
Number of Deaths
15,000,000
10,000,000
5,000,000
0
Intentional and unintentional injury, poisoning, and other adverse events
Alzheimer disease
Obesity
Malignant neoplasms
Vascular and unspecified dementia
Sepsis
Renal failure
Diabetes
All other conditions and causes (residual)
Circulatory diseases
Respiratory diseases
Condition Group

Which condition leads to the most deaths by age group?

1. Filter dataset to include each age_group one by one
2. Filter that data further in ascending order to show which conditions lead to the most deaths
3. Repeat for all age groups

CODE:

#Reading CSV

```
Covid19_data <-
read.csv("D://Conditions_Contributing_to_COVID-19_Deaths__by_State_and_Age__Provisional_2020-2023.csv")
```

#Filtering data to only have Condition, COVID.19 Deaths, and Age.Group Variables

```
Question_3 <- Covid19_data %>% select(Age.Group, COVID.19.Deaths, Condition)
```

#Grouping Age Group and Condition by Number of Deaths

```
Grouped_Question_3 <- Question_3 %>% filter(!is.na(COVID.19.Deaths)) %>% group_by(Age.Group, Condition) %>%
summarise(sumDeaths = sum(COVID.19.Deaths))
```

#Cleaning data to exclude COVID-19 as it is not relevant

```
Clean_Grouped_Question_3 <- Grouped_Question_3[-c(2,4,25,27,48,50,71,73,94,96,117,119,140,142,163,165), ]
```

#Creating Age 0-24 group dataset

```
Age_Group_0_24 <- Clean_Grouped_Question_3 %>%
filter(Age.Group == "0-24")
```

#Arrange to show conditions leading to most deaths in Age 0-24 group

```
arrange(Age_Group_0_24, desc(sumDeaths))
```

	Age.Group <chr>	Condition <chr>	sumDeaths <int>
1	0-24	Influenza and pneumonia	7755
2	0-24	Respiratory failure	4868
3	0-24	Obesity	3236

4	0-24	Adult respiratory distress syndrome	2042
5	0-24	Cardiac arrest	1990
6	0-24	Other diseases of the circulatory system	1936
7	0-24	Intentional and unintentional injury, poisoning, and other adverse events	1701
8	0-24	Sepsis	1685
9	0-24	Other diseases of the respiratory system	1546
10	0-24	Renal failure	1057

#Creating Age 25-34 group dataset

```
Age_Group_25_34 <- Clean_Grouped_Question_3 %>%
filter(Age.Group == "25-34")
```

#Arrange to show conditions leading to most deaths in Age 25-34 group

```
arrange(Age_Group_25_34, desc(sumDeaths))
```

	Age.Group	Condition	sumDeaths
	<chr>	<chr>	<int>
1	25-34	Influenza and pneumonia	32298
2	25-34	Respiratory failure	21975
3	25-34	Obesity	15077
4	25-34	Adult respiratory distress syndrome	9117
5	25-34	Cardiac arrest	8218
6	25-34	Sepsis	6852
7	25-34	Diabetes	6013
8	25-34	Renal failure	5843
9	25-34	Other diseases of the circulatory system	5536
10	25-34	Hypertensive diseases	4124

#Creating Age 35-44 group dataset

```
Age_Group_35_44 <- Clean_Grouped_Question_3 %>%
filter(Age.Group == "35-44")
```

#Arrange to show conditions leading to most deaths in Age 35-44 group

```
arrange(Age_Group_35_44, desc(sumDeaths))
```

	Age.Group	Condition	sumDeaths
	<chr>	<chr>	<int>
1	35-44	Influenza and pneumonia	87411
2	35-44	Respiratory failure	62326
3	35-44	Obesity	32734
4	35-44	Adult respiratory distress syndrome	25886
5	35-44	Diabetes	23279
6	35-44	Cardiac arrest	22339
7	35-44	Sepsis	20828
8	35-44	Hypertensive diseases	19437
9	35-44	Renal failure	18062
10	35-44	Other diseases of the circulatory system	13577

#Creating Age 45-54 group dataset

```
Age_Group_45_54 <- Clean_Grouped_Question_3 %>%
filter(Age.Group == "45-54")
```

#Arrange to show conditions leading to most deaths in Age 45-54 group

```
arrange(Age_Group_45_54, desc(sumDeaths))
```

	Age.Group	Condition	sumDeaths
	<chr>	<chr>	<int>
1	45-54	Influenza and pneumonia	221872
2	45-54	Respiratory failure	166170
3	45-54	Diabetes	70350
4	45-54	Adult respiratory distress syndrome	65874
5	45-54	Hypertensive diseases	60533
6	45-54	Cardiac arrest	57151
7	45-54	Obesity	56158

```
 8 45-54     Sepsis                                     56068
 9 45-54     Renal failure                              51977
10 45-54     Other diseases of the circulatory system   33223
```

#Creating Age 55-64 group dataset

```
Age_Group_55_64 <- Clean_Grouped_Question_3 %>%
filter(Age.Group == "55-64")
```

```
  Age.Group Condition                          sumDeaths
   <chr>     <chr>                                  <int>
 1 55-64     Influenza and pneumonia               494167
 2 55-64     Respiratory failure                   393528
 3 55-64     Diabetes                              168843
 4 55-64     Hypertensive diseases                 159316
 5 55-64     Adult respiratory distress syndrome   134393
 6 55-64     Sepsis                                129194
 7 55-64     Cardiac arrest                        125665
 8 55-64     Renal failure                         120300
 9 55-64     Obesity                                78892
10 55-64     Ischemic heart disease                 75709
```

#Arrange to show conditions leading to most deaths in Age 55-64 group

```
arrange(Age_Group_55_64, desc(sumDeaths))
```

#Creating Age 65-74 group dataset

```
Age_Group_65_74 <- Clean_Grouped_Question_3 %>%
filter(Age.Group == "65-74")
```

#Arrange to show conditions leading to most deaths in Age 65-74 group

```
arrange(Age_Group_65_74, desc(sumDeaths))
```

	Age.Group	Condition	sumDeaths
	<chr>	<chr>	<int>
1	65-74	Influenza and pneumonia	773279
2	65-74	Respiratory failure	646850
3	65-74	Hypertensive diseases	279765
4	65-74	Diabetes	271990
5	65-74	Renal failure	193808
6	65-74	Sepsis	192547
7	65-74	Cardiac arrest	189865
8	65-74	Adult respiratory distress syndrome	177502
9	65-74	Ischemic heart disease	162939
10	65-74	Chronic lower respiratory diseases	162936

#Creating Age 75-84 group dataset

```
Age_Group_75_84 <- Clean_Grouped_Question_3 %>%
filter(Age.Group == "75-84")
```

#Arrange to show conditions leading to most deaths in Age 75-84 group

```
arrange(Age_Group_75_84, desc(sumDeaths))
```

	Age.Group	Condition	sumDeaths
	<chr>	<chr>	<int>
1	75-84	Influenza and pneumonia	831481
2	75-84	Respiratory failure	712050
3	75-84	Hypertensive diseases	345036
4	75-84	Diabetes	268955
5	75-84	Ischemic heart disease	232827
6	75-84	Chronic lower respiratory diseases	211851
7	75-84	Renal failure	200007
8	75-84	Cardiac arrest	199868
9	75-84	Sepsis	173745
10	75-84	Heart failure	168222

#Creating Age 85+ group dataset

```
Age_Group_85 <- Clean_Grouped_Question_3 %>%
filter(Age.Group == "85+")
```

#Arrange to show conditions leading to most deaths in Age 85+ group

```
arrange(Age_Group_85, desc(sumDeaths))
```

	Age.Group	Condition	sumDeaths
	<chr>	<chr>	<int>
1	85+	Influenza and pneumonia	727237
2	85+	Respiratory failure	592044
3	85+	Hypertensive diseases	392193
4	85+	Vascular and unspecified dementia	330541
5	85+	Ischemic heart disease	247990
6	85+	Heart failure	234030
7	85+	Cardiac arrhythmia	208430
8	85+	Cardiac arrest	185247
9	85+	Diabetes	179043
10	85+	Renal failure	164459

Key Takeaways

- Age is a significant factor in COVID-19 mortality: The 85+ age group experienced the most deaths due to COVID-19, closely followed by the 75-84 age group, indicating a strong correlation between increasing age and higher COVID-19 death tolls.
- Respiratory conditions are the primary contributors to COVID-19 deaths: “Influenza and pneumonia” was the most common co-condition leading to death across almost all age groups, with “Respiratory failure” consistently ranking as the second most frequent.
- Common pre-existing conditions play a role in leading to deaths by COVID-19. Beyond respiratory issues there are other significant contributing conditions across various age groups including but not limited to “Obesity,” “Diabetes,” “Hypertensive diseases,” “Cardiac arrest,” “Sepsis,” and “Renal failure”.