PM 566 and 592 Final Assignment

Edward Kim

Contents

Introduction

The COVID-19 pandemic still heavily impacts the United States, with the US surpassing over 200,000 deaths since cases have first been recorded. As of November 16th, when this report was written, the number of Covid-19 cases and deaths have risen drastically throughout the United States. To shed further insight into the severity of the pandemic, the **"Provisional COVID-19 Death Count by Sex, Age, and State"** (https://data.cdc.gov/resource/9bhg-hcku.json) data was taken from the **Center for Disease Control** and analyzed. The data includes the number of COVID-19 deaths between February 2020 and August 2020 that was reported to the **National Center for Health Statistics** by sex and age group. In addition, the number of deaths due in which pneumonia, often caused by severe COVID symptoms,was diagnosed alongside with Covid-19 was also included in the data set. Data gathered by this data set is incomplete due to the length of time in which it takes for a death certificate to be completed and submitted to the NCHS after death. Furthermore, the number of Covid-19 deaths listed in this report does not accurately reflect the current state of the pandemic as the dataset only included Covid-19 data from February 2020 and August 2020.

In the **"Provisional COVID-19 Death Count by Sex, Age, and State"**, the number of Covid-19 deaths reported by state was recorded without standardization for each state. Therefore, the effect of population density of each state on the number of Covid-19 deaths from that state was also included in the analysis as a possible confounding factor or effect modifier. The data on population density was taken from the **"2010 Census: Population Density Data"** reported by the **United States Census Bureau**. The data set lists the population density by peopler per square mile for each state. However, as the data was from 2010 there may have been shifts in the population density since then.

The main purpose of the report is to:

1)Explore and analyze the relationship between **age** and **state** on the number of **Covid-19 Deaths**

2)Explore the frequency of **pneumonia** in Covid-19 Patients, and it's affect on patient mortality

Methods:

The data set, **"Provisional COVID-19 Death Count by Sex, Age, and State"** (https://data.cdc.gov/resource/9bhg-hcku.json), was accessed from the **Center for Disease Control** website through an API. Once downloaded, the desired information was extracted through regular expressions and formed into a data

table. The key independent variables that were examined in this study were **age**, **gender**, and **state** while the data of interest included number of **deaths from COVID-19**, number of **deaths from pneumonia**, and the number of deaths which both **COVID-19 and pneumonia** were involved.

Data on the population density for each state was taken from the **"2010 Census: Population Density Data"** reported by the **United States Census Bureau**. The data was download from the US Census Bureau using a CSV file for analysis. The purpose of this data set was to adjust as a potential confounding factor for the relationship between the location by state and number of deaths due to Covid-19

I. Examining Age, Gender, and State:

Age: The age variable was separated into different age groups, including ranges from 0-17, 15-24, 18-29, etc. When the age group was extracted from the raw data, there were observations that did not include data regarding the age. Therefore, those observations were removed.

In addition, there were also overlapping age-ranges in the data set. To prevent double-counting of deaths, the overlapping age-ranges were removed. The final age groups started from age 5 to age 84, broken down into increments of 10 years (5-14,15-24,26-35, etc.)

To ensure the accuracy of the data cleaning process, the number of Covid-19 deaths per age composition was added to determine if it was equal to the total number of Covid-19 deaths reported across all age groups.

Gender: The number of deaths due to COVID-19 was separated by gender. Some observation listed the gender as "All Genders." Because the goal of this project is to determine the influence of gender on COVID-19 mortality, these values were removed. In addition, removing those values would prevent double-counting of the data. Cases where the gender was unknown were also removed.

To ensure the accuracy of the data, the number of COVID-19 deaths for each gender was added to determine if it was equal to the total number of Covid-19 deaths reported throughout all genders.

State: The number of deaths due to Covid-19 was separated by state. In the original dataset, New York City was listed as a separate category than the state of New York due to the large amount of Covid-19 cases centralized in that area. The number of covid-19 cases in New York City was added to the state totals. US Territories such as Puerto Rico was also included in the original dataset. However, because the focus of this dataset is on individual states, the data for Puerto Rico was excluded.

II. Examining the COVID-19 Deaths and Pneumonia Deaths by Age:

The number of COVID-19 Deaths, Pneumonia Deaths, and deaths involving both Covid-19 and Pneumonia, were all organized by age group. The same age range was included the initial analysis of age group and Covid-19 mortality. Data in which the gender not known, as well as overlapping age categories were excluded from this analysis. The numbers from the resulting analysis may be incomplete due to missing data and lag in the reporting of deaths due to all three conditions.

All of the tables and figures were made through knitrr and ggplot2. Interactive figures (published on the website: https://eshkim1021.github.io/PM-566-Final/) were made by the Data Table package and the plotly package.

III. Building the Model

The primary research question was to determine the relationship between two demographic characteristics, age and location by state, on the number of Covid-19 Deaths in the United States from February 2020 to August 2020. In addition to the two main parameters, the gender of the individuals and the population

density of each state were included in the model as possible confounding factors. As the data includes the number of deaths due to Covid-19, a Poisson regression model will be used to determine the effects of age and state on the number of deaths. The form of each variable and its univariate effect on the number of deaths were determined.

To approximate **age** as a continuous variable for data analysis, each age group was coded as a numeric value equivalent to the midpoint of each age range. For example, the age category from 5-14 was coded as the number 10, while the age category from 15-24 was coded as the number 20. This was done not only to approximate age as a continuous variable but also to help simplify the interpretation of the results. The **gender** of those that had died was kept as a categorical variable with Female as the reference group. The **population density** for each state was kept as a continuous variable. The **location by state**, however, when examined individually, showed the relationship between each individual state and the number of deaths due to Covid-19 deaths. Rather than include all 50 states in the model, only the states with statistically significant relationships with Covid-19 deaths was included. These states include: California, New York, New Jersey, Texas, and Florida. The state.adj variable now indicatess whether the death occurred in one of those five states. Formatting the state variable in this way reduces the complexity of the model and also simplifies the interpretation.

The gender and population variables were included as possible confounding factors and effect modifiers in the Poisson regression model. Confounding factors had to significantly change the parameter estimates of age and or location by state with inclusion with statistical significance. Effect modifiers also had to significantly influence the paramter estiamte of either age or location by state with statistical significance.

The final model was reached by examining the univariate effect of each parameter estimate on the number of deaths due to Covid-19, and determining if the potential bias introduced by gender and population density is statistically significant and changes the parameter estimates of age and location by state. The overdipersion of the model was also tested to determine if the variance is larger than what would be expected under a Poisson distribution.

Results:

Visualizations for Exploratory Data Analysis:

Age The following table and figure analyze the relationship between age group and Covid-19 death:

Age_Group	Covid_Deaths
5-14 years	42
15-24 years	428
25-34 years	1812
35-44 years	4663
45-54 years	12369
55-64 years	29888
65-74 years	51666
75-84 years	64574

The table above lists the number of Covid-19 deaths for each age group in the United States from February to August 2020. The numbers range from 35 deaths, for those between 5-14 years old, to 52,617 deaths, for those in between 75-84 years old.

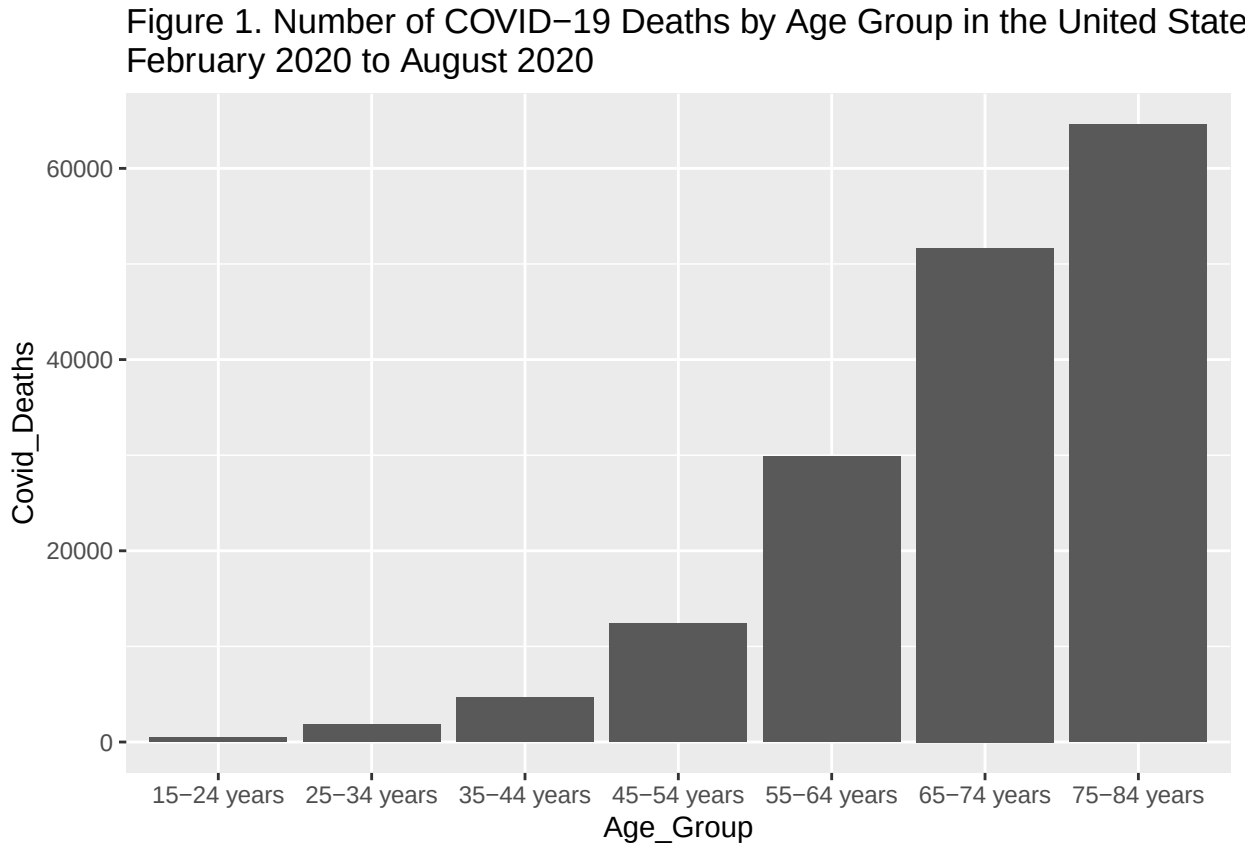

Figure 1 illustrates the number of Covid-19 deaths by age group in the United States from February to August 2020. The age group of 5-14 years was removed from the data set as the number of deaths due to Covid-19 was significantly less than the other age groups. There was 35 deaths from Covid-19 in the age group of 5-14, which comprised of <0.025% of the total Covid-19 death.

According to **Figure 1**, the number of deaths due to Covid-19 increased for every age group. The older the patient, the greater the Covid-19 mortality rate. The increase in the number of deaths was particular pronounced after the age of 55, as the number of deaths seems to increase exponentially for each increase in age group.

Gender

The following table shows the difference in Covid-19 deaths by gender in the United States from February 2020 to August 2020.

Gender	Covid Death
Female	91166
Male	137837

The number of males that have died due to Covid-19 is 114,291, while the number of females that have died due to Covid-19 is 75,203. These numbers are different from the total number of Covid-19 deaths calculated from the Covid-19 due to age distribution because different observations were omitted depending on the unknown or repetitive variables for each category.

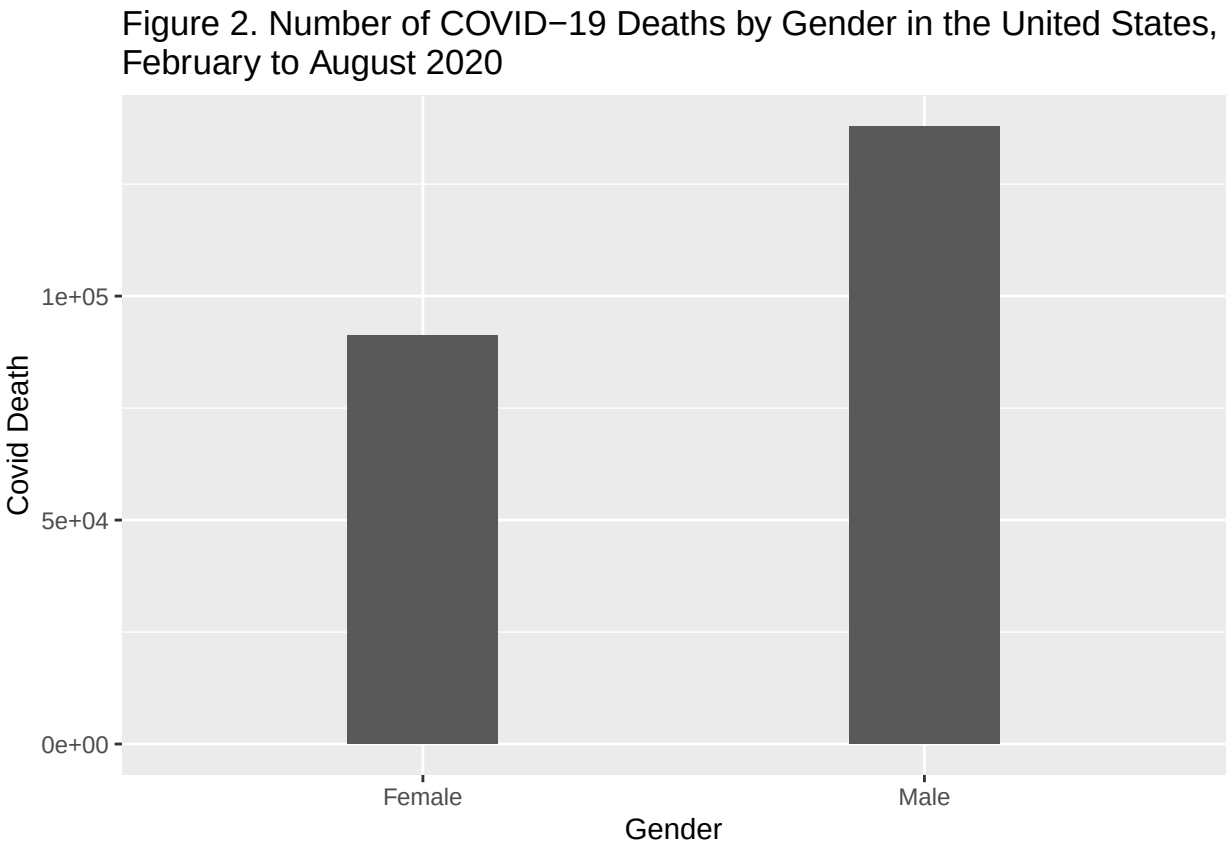

Figure 2 illustrates the difference in the number of Covid-19 deaths in the United States by gender from February to August 2020. The cases in which the gender was unknown were removed from this figure. According to the data, there have been more cases of males dying due to Covid-19 than females. The ratio of males to females that have died due to COVID-19 is 1.520, indicating that the number of males that have died from Covid-19 is 1.520 times greater than the number of females that have died.

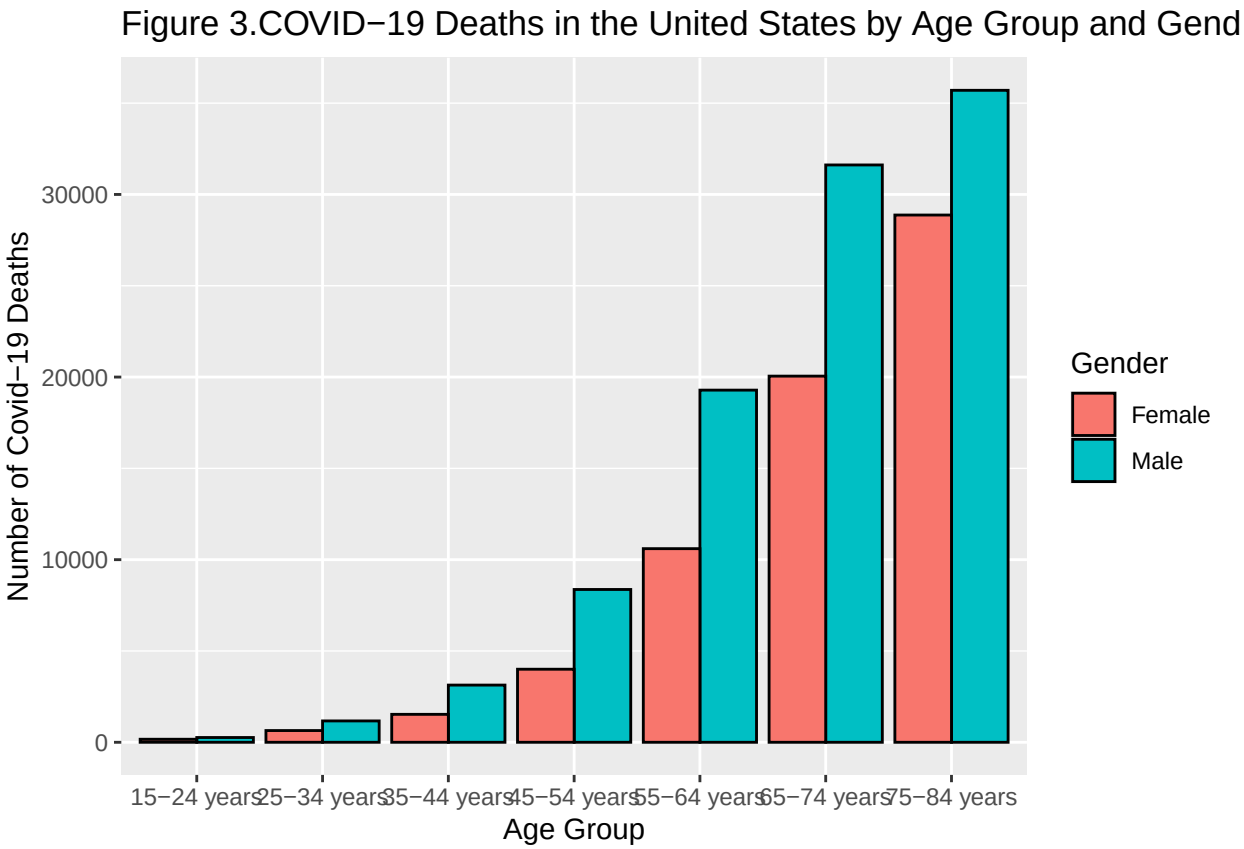

Figure 3 illustrates the number of Covid-19 deaths by age group and gender in the United States from February to August 2020. The gender distribution of the number of Covid-19 deaths for each age group mirrors that of the overall population of the United States. In each age group, the number of male deaths from Covid-19 are greater than that of females. 4

State

A table consisting of the raw data for the number of Covid-19 Deaths by state can be found on the website (https://eshkim1021.github.io/PM-566-Final/) under the "Additional Figures" tab.

An interactive map detailing the deaths due to Covid-19 by state can be found on the website: (https://eshkim1021.github.io/PM-566-Final/)

COVID-19 Deaths and Pneumonia Deaths by Age

A table consisting of the raw data for the number of Covid-19 Deaths and Pneumonia Deaths by age can be found on the website (https://eshkim1021.github.io/PM-566-Final/) under the "Additional Figures" tab.

The number of deaths for each condition increased as the individual gets older, which is expected. The number of Covid deaths recorded does not include the number of deaths where both Covid-19 and pneumonia are found.

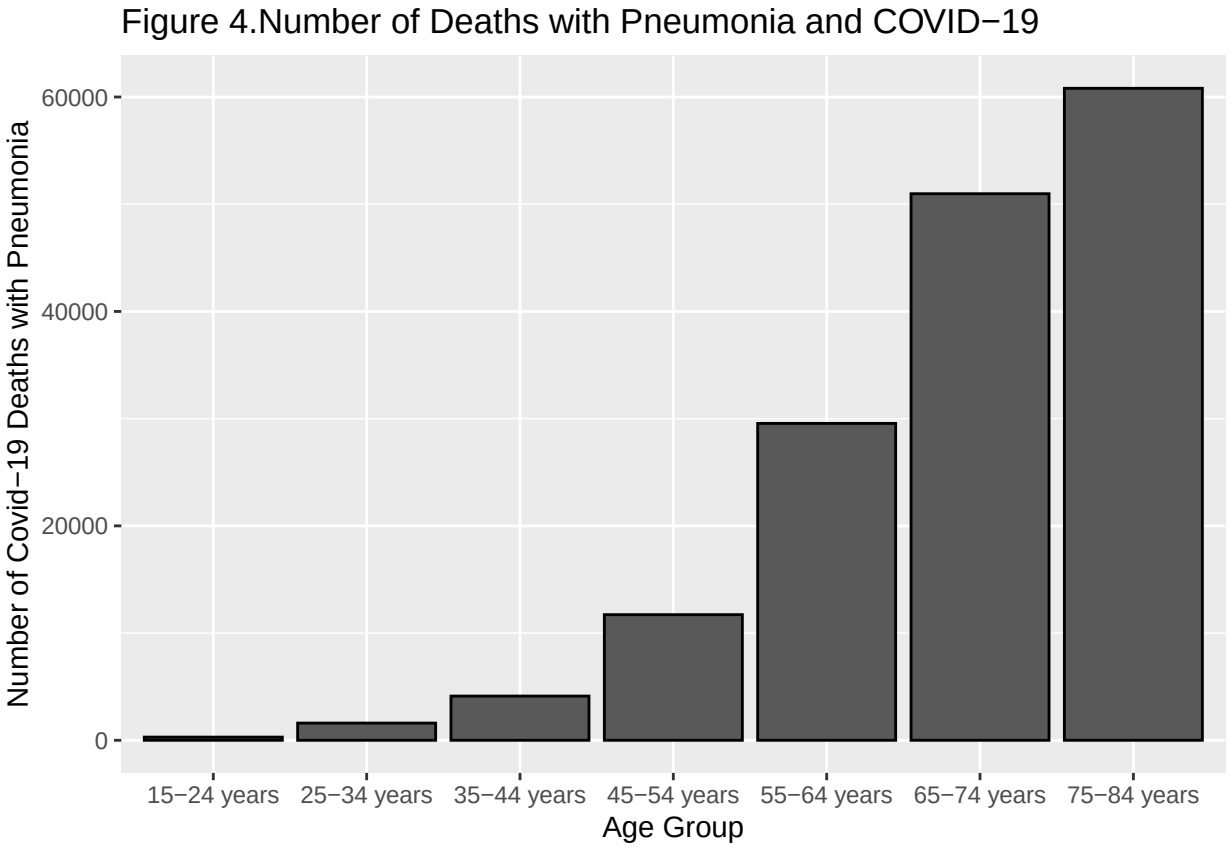

Figure 4 illustrates the number of deaths in which both Covid-19 and Pneumonia were involved. The number of cases with Covid-19 and pneumonia increase with age, and reaches the highest values at those between 75-84 years of age.

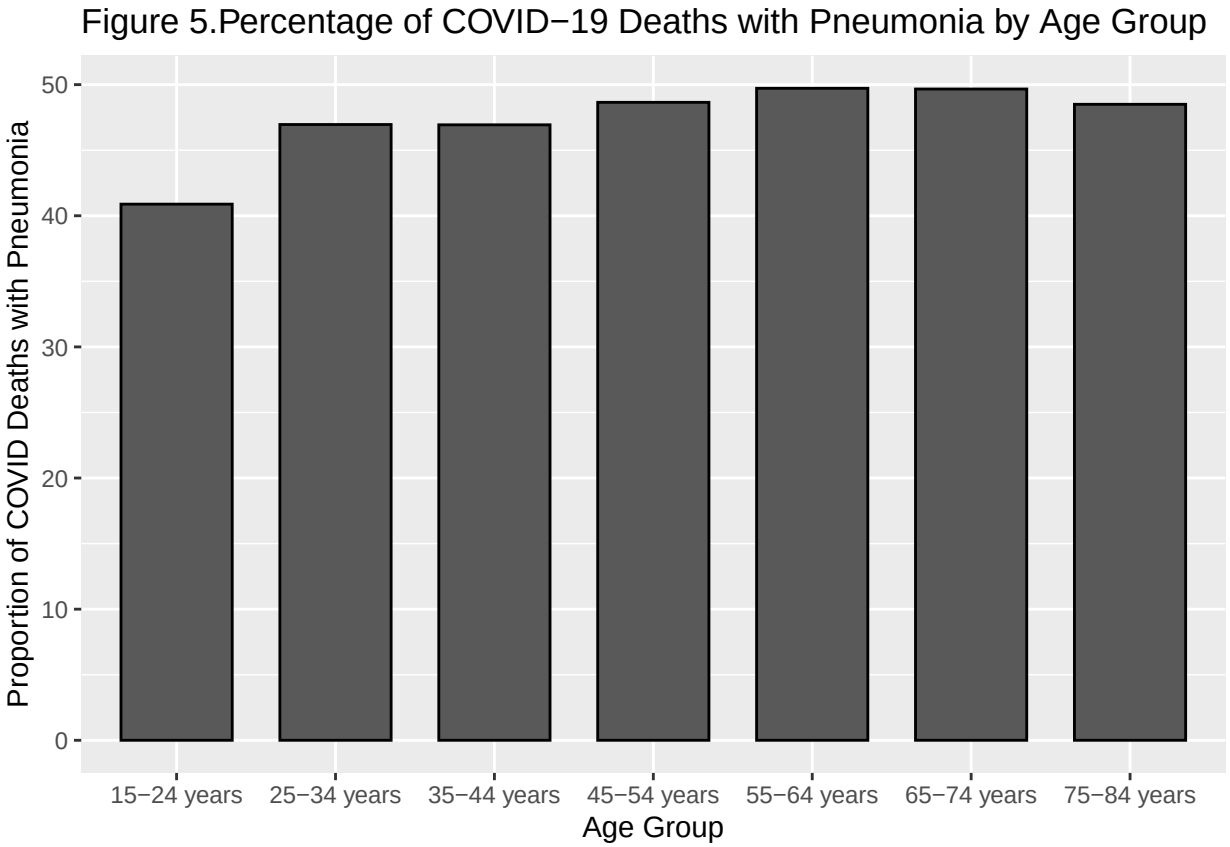

Figure 5 illustrates the percentage of deaths in which both Covid-19 and Pneumonia were present compared to the total number of deaths due to Covid-19. Throughout most age groups, the percentage of deaths in which both Covid-19 and Pneumonia are present account for around 40-50% of the total Covid-19 deaths. It stays relatively consistent throughout the age groups.

Data Analysis:

The following table illustrates the average age and population density for the study population:

Table 3: Descriptive Statics for Continuous Variables

Variable	Average	STDEV
Age	45.000	22.927
Population Density	384.404	1364.621

The results from the Poisson regression univaraite anlysis between each variable and the number of Covid-19 deaths are illustrated by the following table.

Table 4: Univaraite Analysis Between Each Variable and Number of COVID Deaths

Variable	Regression Coefficient	Standard Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound	P-Value
Age	0.055	1.271	0.054	0.055	0.000
Gender	0.311	1.858	0.299	0.322	0.000
State.Adj	1.677	1.555	1.666	1.689	0.000
Population Density	-2e-05	1.890	-2e-05	-1e-05	0.000

All the variables examined, age, gender, state, and population density, had a statistically significant p-value in its relationship to the number of deaths due to Covid-19. The p-value in the table is indicated as 0, illustrating that the p-value is too small to be represented by the decimal places in the table.

To determine if the gender and population density were confounding factors or effect modifiers, both variables were added separately to the model to determine their effect on the relationship between age and location by state with the number of Covid-19 Deaths. The inclusion of sex significantly changed the effect of age and location by state with the number of Covid-19 cases, indicating that it is a confounding factor. Although the population density did not significantly change the parameter estimates, there was a statistically significant interaction between the state and the population density and was included as an interation term.

The following table illustrates the final Poisson Regression Model including all relevant variables:

Table 5: Poisson Regression Multivariable Model for Relationship with Number of Covid Deaths

Parameter	Parameter Estimates	Rate Ratio	2.5 %	97.5 %	P-value
(Intercept)	1.10623	3.02293	2.93099	3.11776	0
state.adj	1.87535	6.52308	6.42192	6.62583	0
age_cont	0.05768	1.05938	1.05894	1.05982	0
sexMale	0.36953	1.44705	1.43027	1.46404	0
density	-0.00004	0.99996	0.99995	0.99996	0
state.adj:density	-0.00015	0.99985	0.99982	0.99987	0

However, it is important to check for overdisperion in a Poisson Regression Model. The following test examins the overdisperion of the final Poisson Model.

```
AER::dispersiontest(final.m)
```

```
##
##  Overdispersion test
##
## data:  final.m
## z = 10.556, p-value < 2.2e-16
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   117.4771
```

The test for overdipserion indicates that the dispersion is greater than what would be expected in a Poisson Model. This was corrected by changing the regression model to a negative binomial model.

Table 6: Negative Binomial Multivariable Model for Relationship with Number of Covid Deaths

Parameter	Parameter Estimates	Rate Ratio	2.5 %	97.5 %	P-value
(Intercept)	-1.16091	0.31320	0.23943	0.40970	0.00000
state.adj	2.07537	7.96751	5.29293	11.99358	0.00000
age_cont	0.09359	1.09811	1.09349	1.10276	0.00000
sexMale	0.51461	1.67298	1.39007	2.01348	0.00000
density	-0.00006	0.99994	0.99987	1.00001	0.07427
state.adj:density	-0.00037	0.99963	0.99897	1.00029	0.27620

After adjusting for overdisperion, the final model indicates that the interaction between location by state and the population density, and the population density variable itself, is not statistically significant. However, the state, age, and gender all continue to be statistically significant in its relationship to the number of Covid-19 Deaths. According to the model, individuals living in the states of California, New York, New Jersey, Texas, and Florida are 7.96 times more likely to die due to Covid-19 than individuals living elsewhere. Males are 1.67 times more likely (or 67% more likely) to die due to Covid-19 than females. For a 1 year increase in age, people are 1.09 times more likely (or 9% more likely) to die due to Covid-19.

The goodness of fit for the final negative binomial model was examined. The Pearson chi-sqaured test indicated that the model does show departure from goodness of fit. However, this could be due to the large sample size of the data. However, the Deviance Goodness of Fit test also indicates a departure from goodness of fit, even though the p-value is not as small as that of the Pearson chi-squared test.

```
pois_pearson_gof(final.adj)
```

```
## $pval
## [1] 0
##
## $df
## [1] 592
```

```
pois_dev_gof(final.adj)
```

```
## $pval
## [1] 0.002282986
##
## $df
## [1] 592
```

Conclusion:

In conclusion, the number of deaths due to Covid-19 is influenced by age and gender. It was found that the number of deaths due to Covid-19 increase with age, with there being the most deaths from individuals in between the ages of 75-84. There was a greater number of Covid-19 deaths in males than females. In addition, it was determined that a majority of the deaths due to Covid-19 were in states with large populations, such as New York and California. Also, pneumonia was prsent in around 50% of the deaths due to Covid-19, with that percentage staying relatively similar throughout all the age groups. Although pneumonia is found in around half of Covid-19 cases, there is not enough information in this data set to determine if a co-diagnosis of pneumonia lead to an increase in mortality rate from Covid-19.

Through analysis with a negative binomial regression model, it was determined that individuals living in the states of California, New York, New Jersey, Texas, and Florida are 7.96 times more likely to die due to Covid-19, while males are 1.67 times more likely (or 67% more likely) to die due to Covid-19 than females. In addition, for a 1 year increase in age, people are 1.09 times more likely (or 9% more likely) to die due to Covid-19. Although the model showed a slight departure from goodness of fit, the parameter estimates showed statistical significant with their relationship to the number of deaths. Therefore, through this analysis, we can determine the increase in risk of death due to Covid-19 based on age, location, and sex.