



# STATISTICAL INFERENCE AND DATA ANALYSIS

FINAL COURSE PROJECT: ANALYZING U.S. COVID-19 DATA

---

## Business Report

---

### Students:

Rghda Salah 202101510

[s-rghda.ahmed@zewailcity.edu.eg](mailto:s-rghda.ahmed@zewailcity.edu.eg)

Nada Nabil 202101220

[s-nada.soudi@zewailcity.edu.eg](mailto:s-nada.soudi@zewailcity.edu.eg)

Mariam Ismael 202101506

[s-mariam.ismael@zewailcity.edu.eg](mailto:s-mariam.ismael@zewailcity.edu.eg)

Tasneem Muhammed 202101031

[s-tasneem.attia@zewailcity.edu.eg](mailto:s-tasneem.attia@zewailcity.edu.eg)

2023/2024 – Spring Semester

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Background and context of the project . . . . .	4
1.2	Objective of the analysis . . . . .	4
1.3	Brief overview of the datasets used . . . . .	4
<b>2</b>	<b>Exploratory Analysis</b>	<b>5</b>
2.1	Total number of hospitalizations versus deaths from COVID-19 over the entire US per month-year timestamp . . . . .	5
	Visualization . . . . .	5
	Commentary . . . . .	5
2.2	The average rates of COVID-related deaths relative to patient demographics . . . . .	6
	Visualization . . . . .	6
	Commentary . . . . .	6
2.3	The rates of COVID-related hospitalization and death with age . . . . .	7
	Visualization . . . . .	7
	Commentary . . . . .	8
2.4	Average rate of COVID-related hospitalization and death per state over the entire study period. . . . .	8
	Visualization . . . . .	8
	Commentary . . . . .	9
2.5	The relationship between age, pre-existing medical conditions and/or risk behaviors, and rate of admittance to the ICU. . . . .	10
	Visualization . . . . .	10
	Commentary . . . . .	10
2.6	The rate of expected employment loss due to COVID-19 and sector of employment. . . . .	11
	Visualization . . . . .	11
	Commentary . . . . .	11
2.7	The rate of expected employment loss due to COVID-19 relative to responders demographics. . . . .	12
	Visualization . . . . .	12
	Commentary . . . . .	12
2.8	The rate of expected employment loss due to COVID-19 for the top 10 states with the highest rate of COVID hospitalization. . . . .	13
	Visualization . . . . .	13
	Commentary . . . . .	13
2.9	The relationship between household income and the rate of delayed/ OR unobtained medical treatment (Due to COVID or otherwise). . . . .	14
	Visualization . . . . .	14
	Commentary . . . . .	15
2.10	The relationship between COVID-19 symptom manifestation and age group. . . . .	15
	Visualization . . . . .	15
	Commentary . . . . .	16

<b>3</b>	<b>Answering Questions</b>	<b>17</b>
3.1	Use the appropriate statistics and plots to answer the following questions: . . . .	17
1.	Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19? . . . . .	17
1.1	Visualization . . . . .	17
1.2	Commentary . . . . .	17
2.	Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19? . . . . .	18
2.1	Visualization . . . . .	18
2.2	Commentary . . . . .	18
3.	Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19? . . . . .	19
3.1	Visualization . . . . .	19
3.2	Commentary . . . . .	19
4.	Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19? . . . . .	20
4.1	Visualization . . . . .	20
4.2	Commentary . . . . .	20
5.	Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19? . . . . .	21
5.1	Visualization . . . . .	21
5.2	Commentary . . . . .	21
3.2	Come up with 5 more bi-variate/multivariate analysis questions and similarly answer each with appropriate visuals and commentary: . . . . .	22
1.	How does the rate of COVID-related hospitalization vary across different age groups and states? . . . . .	22
1.1	Visualization . . . . .	22
1.2	Commentary . . . . .	22
1.3	Interpretation . . . . .	22
2.	What are the most common risk factors associated with COVID-19-related hospitalizations, and how do these factors vary across different age groups and genders? ? . . . . .	23
2.1	Visualization . . . . .	23
2.3	Commentary . . . . .	24
2.3	interpretation . . . . .	24
3.	Are COVID-19 patients with underlying medical conditions more likely to be admitted to the ICU compared to those without underlying conditions ?	25
3.1	Visualization . . . . .	25
3.2	Commentary . . . . .	25
3.3	Interpretation . . . . .	25
4.	Do asymptomatic COVID-19 patients have lower rates of hospitalization compared to symptomatic patients? How does this differ based on demographic factors such as age and sex? . . . . .	26
4.1	Visualization . . . . .	26
4.2	Commentary . . . . .	27

4.3 Interpretation . . . . .	27
5. What is the frequency of common symptoms reported by COVID-19 patients, and how do they vary across different age groups and genders? . . . . .	28
5.1 Visualization . . . . .	28
5.2 Commentary . . . . .	29
5.3 Interpretation . . . . .	29
<b>4 Hypothesis Testing</b>	<b>30</b>
4.1 Claim: “There is a strong association between the probability of death due to COVID-19 and patient demographics”: . . . . .	30
4.1.1 First Test Chosen: Anova . . . . .	30
4.1.2 Second Test Chosen: Chi-squared . . . . .	30
<b>5 Regression Analysis</b>	<b>31</b>
<b>6 Machine Learning Classifier</b>	<b>39</b>
6.1 Machine Learning Classifier to predict the likelihood of death due to COVID-19 using any/all of the relevant attributes in the COVID-19 case surveillance dataset: . . . . .	39
5.1 First Data Preparation . . . . .	39
5.2 Feature Selection . . . . .	40
5.3 Model Training . . . . .	40
5.4 Model Approach . . . . .	40
5.5 Model Performance . . . . .	40
<b>7 Conclusion</b>	<b>41</b>
<b>Appendices</b>	<b>43</b>

# 1 Introduction

## 1.1 Background and context of the project

The project is focused on analyzing COVID-19 cases using relevant datasets. The background provides an understanding of the project's context, including the significance of studying COVID-19 cases and the need for data analysis to gain insights into the disease.

## 1.2 Objective of the analysis

The primary objective of the analysis is to examine various aspects of COVID-19 cases and extract valuable information from the datasets. This analysis aims to identify the characteristics related to the cases. By analyzing the factors required throughout the project, the analysis seeks to contribute to a better understanding of the COVID-19 pandemic and inform public health interventions.

## 1.3 Brief overview of the datasets used

The COVID-19 case surveillance database includes individual-level data reported to U.S. states and autonomous reporting entities, including New York City and the District of Columbia (D.C.), as well as U.S. territories and affiliates. On April 5, 2020, COVID-19 was added to the Nationally Notifiable Condition List and classified as “immediately notifiable, urgent (within 24 hours)” by a Council of State and Territorial Epidemiologists (CSTE) Interim Position Statement (Interim-20-ID-01). CSTE updated the position statement on August 5, 2020, to clarify the interpretation of antigen detection tests and serologic test results within the case classification (Interim-20-ID-02). The statement also recommended that all states and territories enact laws to make COVID-19 reportable in their jurisdiction and that jurisdictions conducting surveillance should submit case notifications to the CDC. COVID-19 case surveillance data are collected by jurisdictions and reported voluntarily to CDC. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum. [1]

## 2 Exploratory Analysis

How we handle the bias in our data: We note in general that the data is biased as the collection data didn't come equally from different states. as there are states that have larger samples than other states which leads to a kind of bias in our data from the way they collect the data itself.

- We remove the duplicate from the data at the beginning
- We remove rows with missing values. This is a good option as in some columns the amount of missing data is small and removing it does not significantly affect our analysis
- We deal with the mean. as we just have some sample, and to increase the credibility in our analysis dealing with the mean will be the best way

### 2.1 Total number of hospitalizations versus deaths from COVID-19 over the entire US per month-year timestamp

#### Visualization

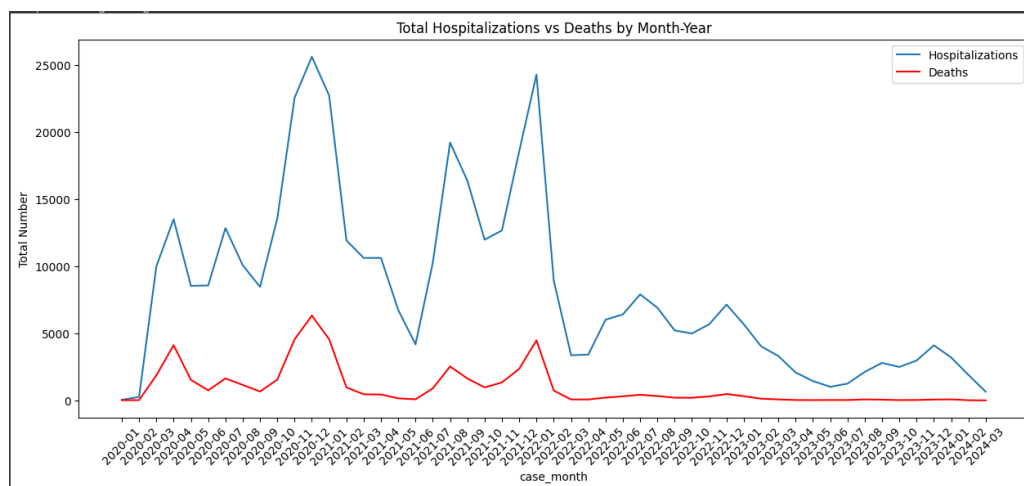


Figure 1

#### Commentary

Total Hospitalizations vs Deaths: The graph shows that the distribution of hospitalizations is consistently higher than total deaths throughout the entire period. This is likely because not everyone who is hospitalized with COVID-19 dies from the disease. Also: It appears overall hospitalization rates have declined since the start of the pandemic. There are still some peaks throughout the time series. at the end of this period, the death rates decreased going to zero and that makes sense as they reach affectively vaccines.

## 2.2 The average rates of COVID-related deaths relative to patient demographics

### Visualization

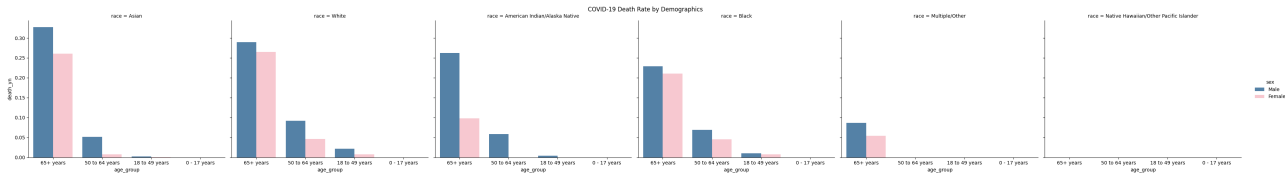


Figure 2

### Commentary

Here's what we found: Our data showed some bias in the "Race" category. We didn't have equal sample sizes from different races. To address this bias and make the data more reliable in our visualizations, we created separate graphs for each race.

Observations from the graphs: We arranged the graphs to show the race with the highest death rate first. In all races, the age group 65 and over had the highest death rate. Asians had the highest death rate among all the races represented in our data.

Additional notes: We noticed that younger age groups had very low death rates. This suggests that death is more common in older age groups. It's also important to note that one race appears to have no deaths in our data. This could be due to several factors, such as limited data collection in that specific race or geographic isolation from areas significantly affected by COVID-19.

2.3 The rates of COVID-related hospitalization and death with age

Visualization

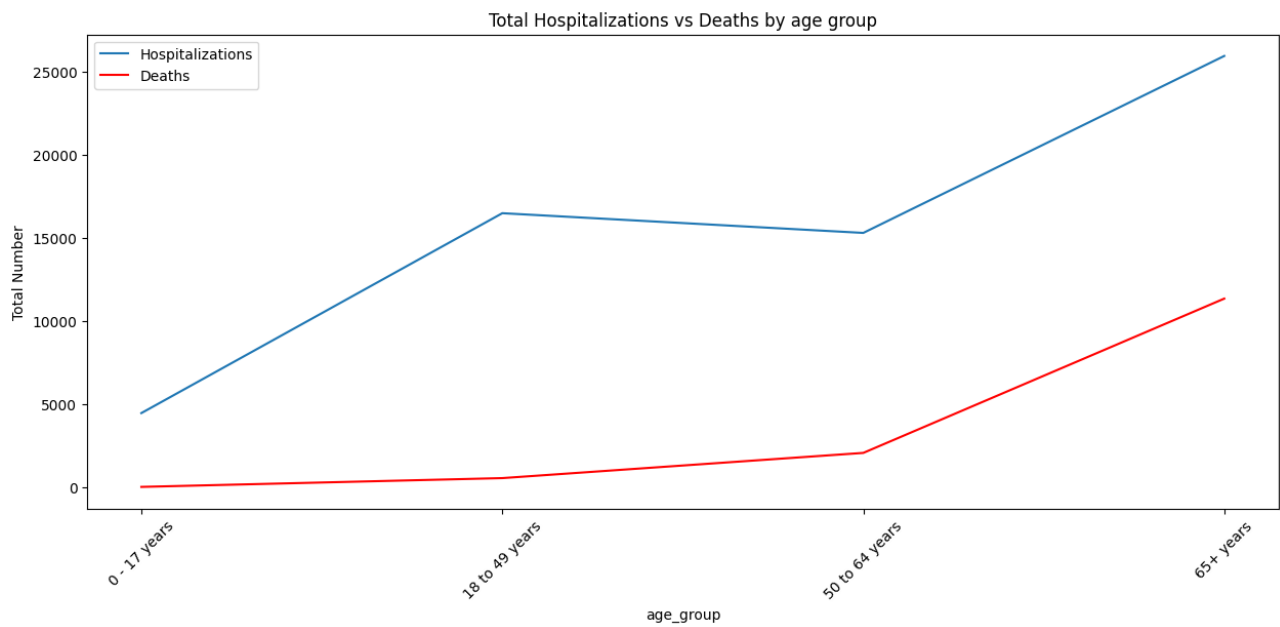


Figure 3

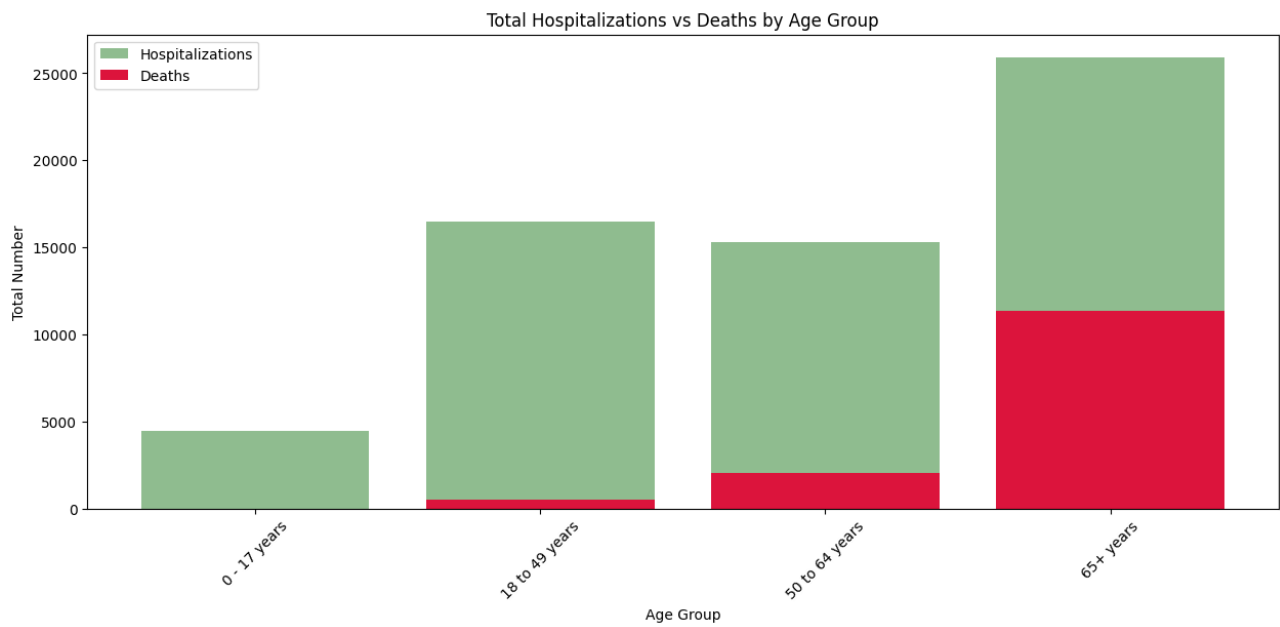


Figure 4



## Commentary

Here, we present two visualizations that both reach the same conclusion: death rates increase with age for hospitalized patients. This makes sense because older individuals often have weaker immune systems and underlying health conditions, making them more susceptible to severe complications from COVID-19.

## 2.4 Average rate of COVID-related hospitalization and death per state over the entire study period.

### Visualization

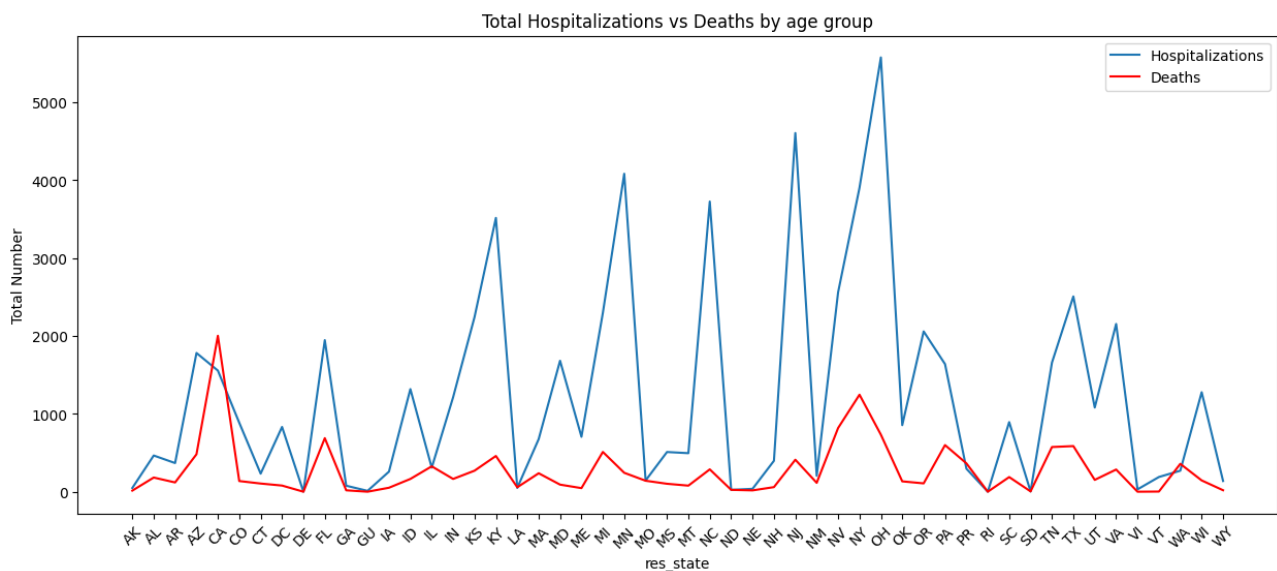


Figure 5

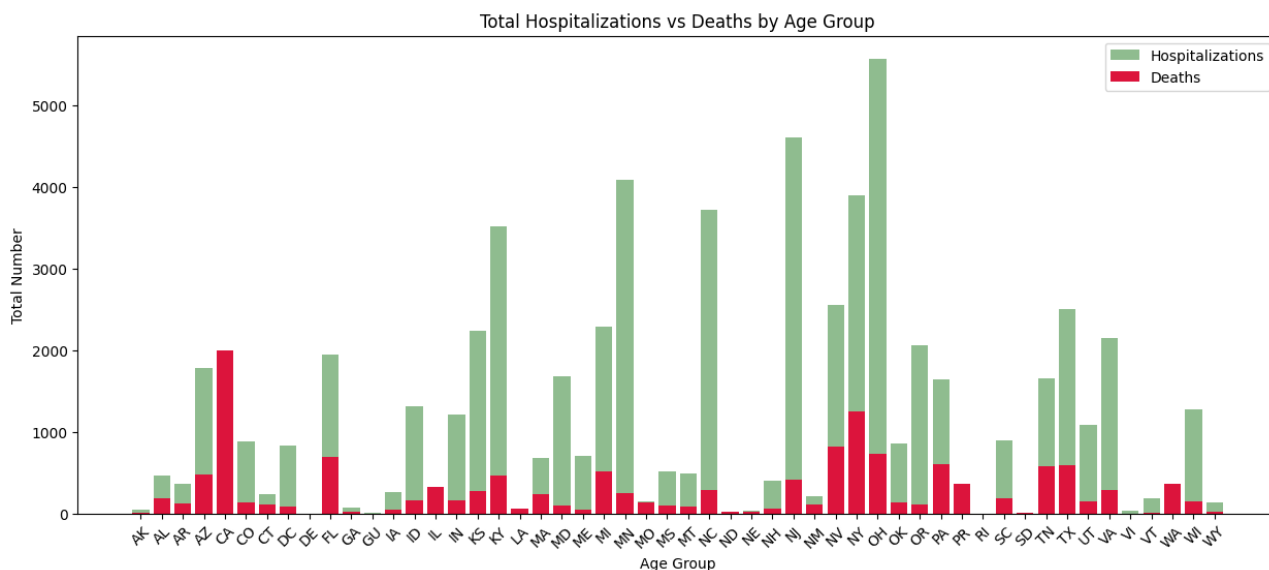


Figure 6

## Commentary

Here, we present a visualization that compares death and hospitalization rates across different states, rather than age groups as shown previously. However, it's important to acknowledge a limitation: the raw data may not be directly comparable due to unequal sample sizes among states. For example, having 100 deaths in one state and 30 deaths in another doesn't necessarily mean the first state has a higher death rate. The state with 100 deaths might simply have a larger population.

To address this bias and create a more accurate visualization, we would need to calculate the death and hospitalization rates for each state. This involves dividing the number of deaths/hospitalizations by the state's total population. This ratio would provide a more accurate comparison of how COVID-19 has impacted different states.

## 2.5 The relationship between age, pre-existing medical conditions and/or risk behaviors, and rate of admittance to the ICU.

### Visualization

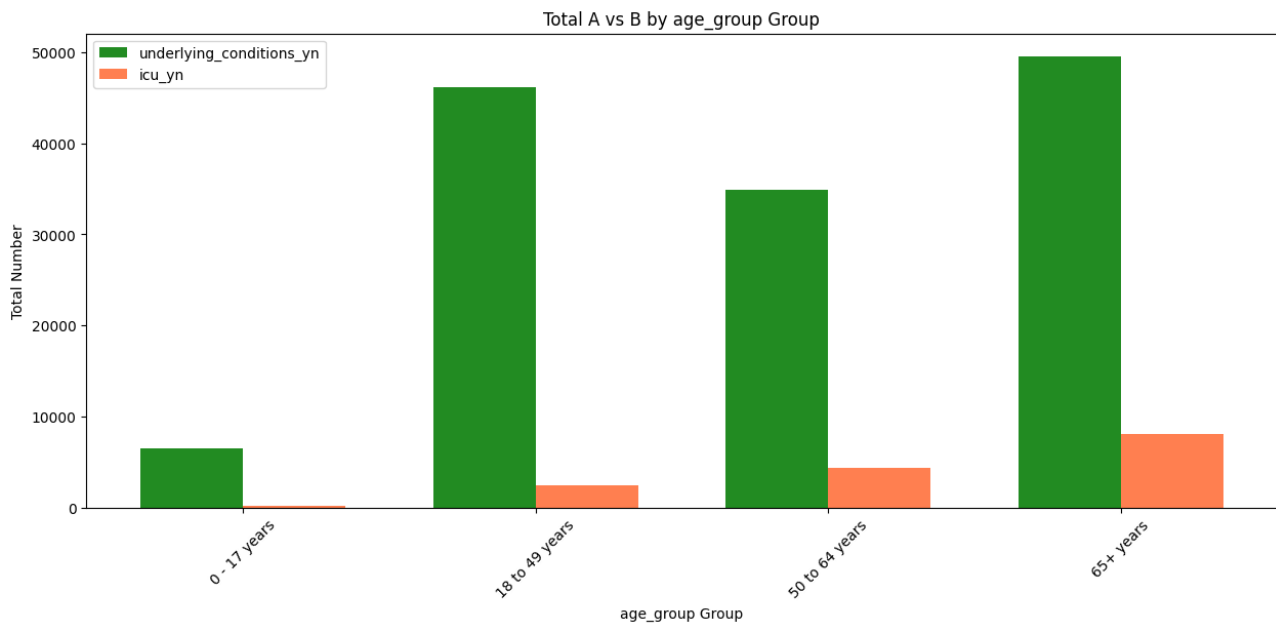


Figure 7

### Commentary

In this visualization, we compare factors that may contribute to ICU admissions. We can see that a greater number of people have pre-existing medical conditions compared to those who are admitted to the ICU. It's also evident that younger age groups have the lowest rates of both pre-existing conditions and ICU admissions. This aligns with our previous observation that older individuals are more likely to face these issues.

## 2.6 The rate of expected employment loss due to COVID-19 and sector of employment.

### Visualization

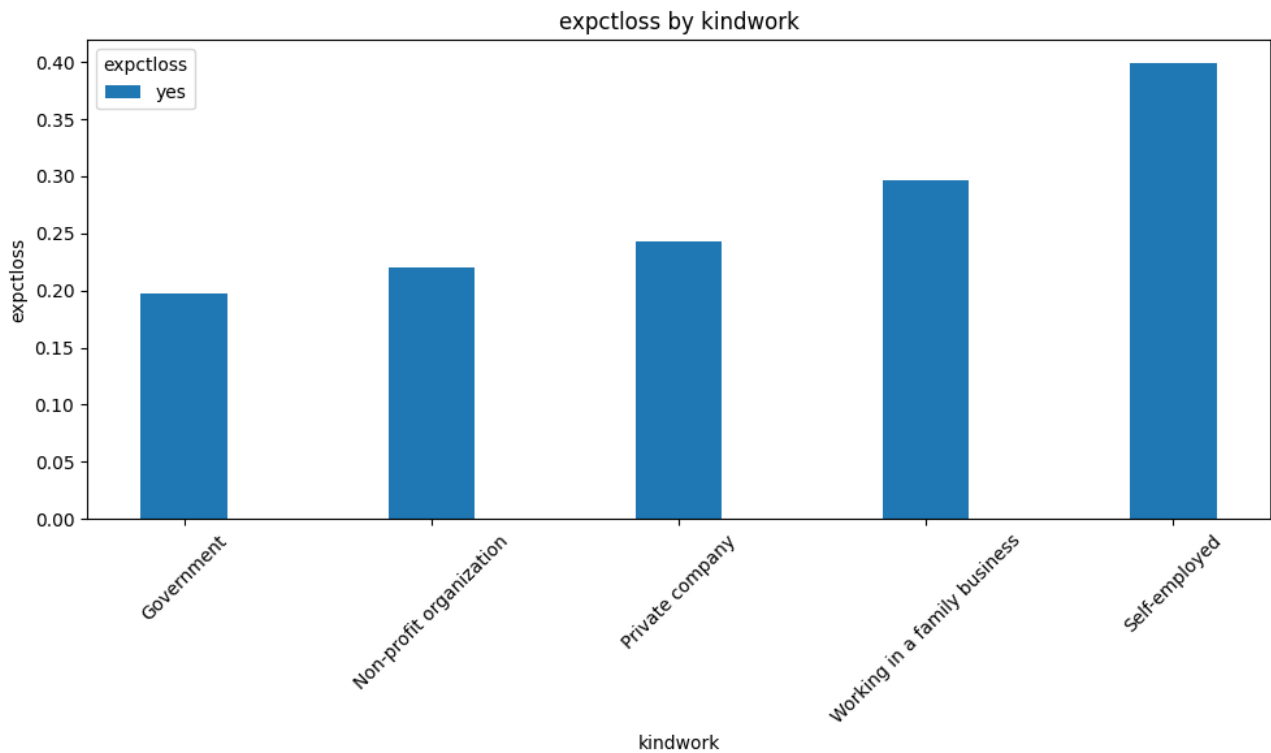


Figure 8

### Commentary

This visualization shows the impact of COVID-19 on job loss across different employment types. We can see that self-employed individuals were most likely to experience job loss. Conversely, government jobs offered the greatest security, with employees experiencing the lowest job loss rates. It's important to note, however, that all job types were affected to some degree by the pandemic.

## 2.7 The rate of expected employment loss due to COVID-19 relative to responders demographics.

### Visualization

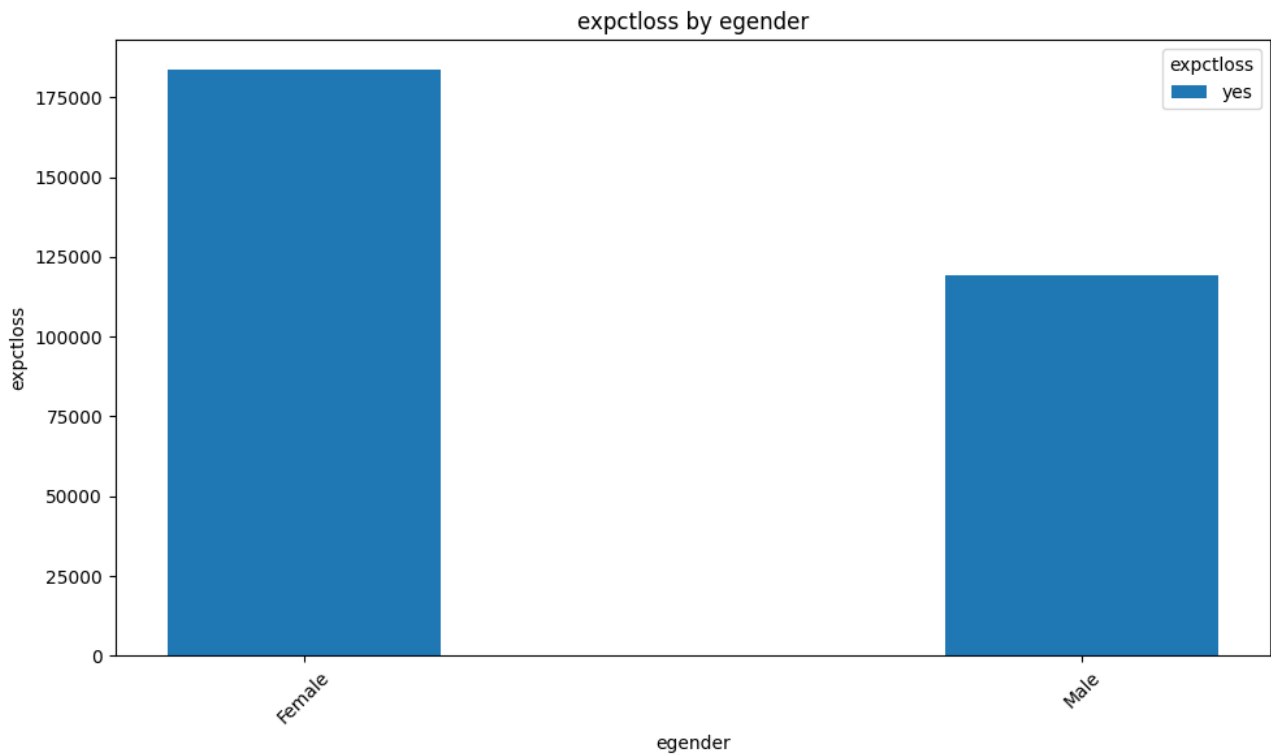


Figure 9

### Commentary

Demographics typically refer to characteristics like age, race, and gender. This visualization explores how these categories are impacted by job loss. It appears that females were more likely to experience job the loss compared to males.

## 2.8 The rate of expected employment loss due to COVID-19 for the top 10 states with the highest rate of COVID hospitalization.

### Visualization

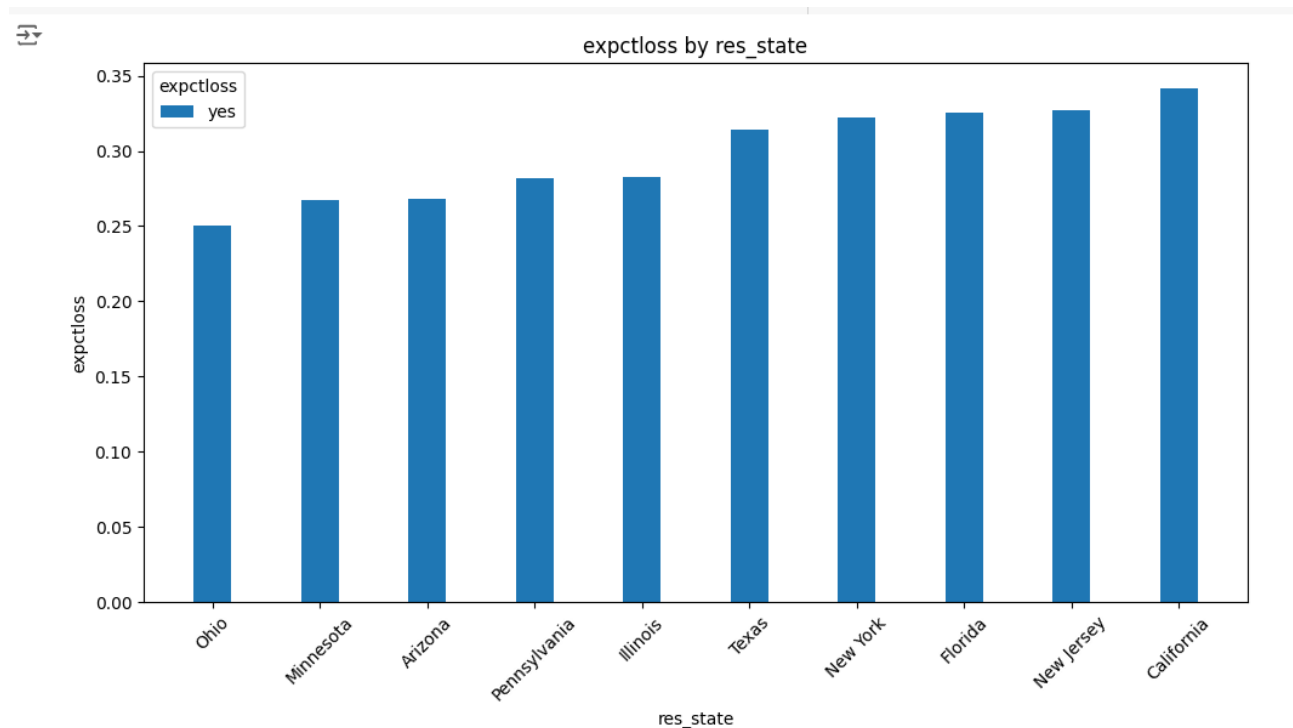



Figure 10

### Commentary

The challenge here was to link the two datasets to understand the relationship between hospitalization rates and job loss. We achieved this by first identifying the ten states with the highest hospitalization rates from the first dataset. Then, we merged this subset with the second dataset to analyze how job loss impacted these specific states. The resulting graph shows the top states in terms of hospitalization rates, along with their corresponding job loss rates. We can observe that California has the highest hospitalization rate among the selected states.

## 2.9 The relationship between household income and the rate of delayed/ OR unobtained medical treatment (Due to COVID or otherwise).

### Visualization



```
Chi-square statistic: 341.7892865022804
P-value: 7.047039559823573e-70
Degrees of freedom: 7
Expected frequencies:
[[47450.84049918 37094.15950082]
 [57931.61399828 45287.38600172]
 [58522.60974073 45749.39025927]
 [94004.24301908 73486.75698092]
 [77541.11090763 60616.88909237]
 [95459.56301925 74624.43698075]
 [45450.54721703 35530.45278297]
 [51992.47159882 40644.52840118]]
There is evidence to reject the null hypothesis.
There is a relationship between income and delay.
```

Figure 11

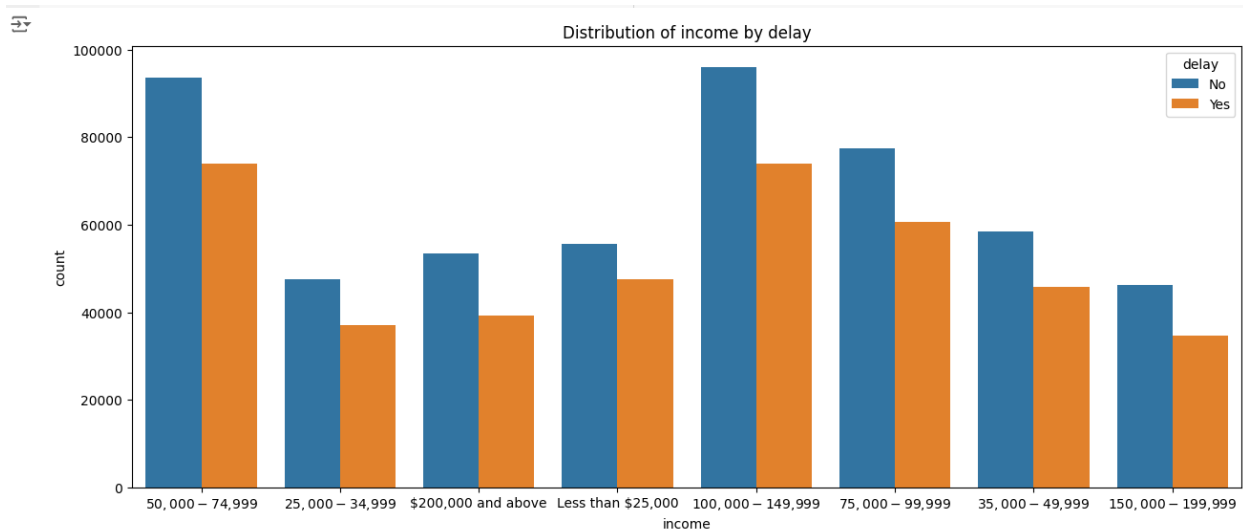


Figure 12

### Commentary

Here, the question was to investigate the relationship between income and treatment delay. Based on our initial analysis, we rejected the null hypothesis. This means we can conclude that there is a connection between income and treatment delay. Our visualization reinforces this finding. As income increases, treatment delays appear to decrease. This supports our hypothesis that income and treatment delay are indeed related, with higher income leading to shorter delays.

## 2.10 The relationship between COVID-19 symptom manifestation and age group.

### Visualization

```

Chi-square statistic: 2126.310866542386
P-value: 0.0
Degrees of freedom: 3
Expected frequencies:
[[ 26609.80113456  60452.99131104  31102.96761124  28504.23994315]
 [ 474575.19886544 1078155.00868896  554709.03238876  508361.76005685]]
There is evidence to reject the null hypothesis.
There is a relationship between COVID-19 symptom manifestation and age groups.

```

Figure 13



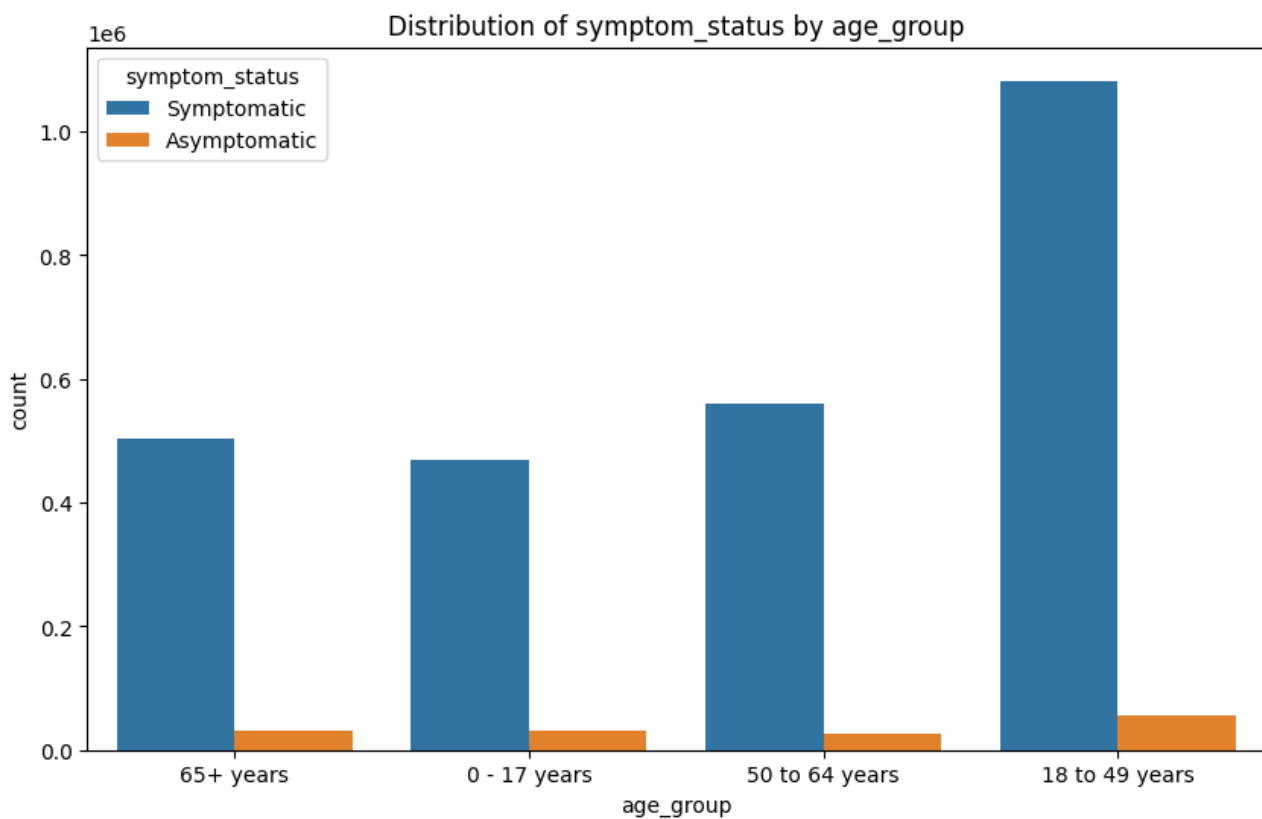


Figure 14

### Commentary

The prompt here investigated the relationship between age and symptom manifestation. Our initial analysis provided evidence to reject the null hypothesis, allowing us to conclude that a relationship exists between these two factors. The graph further supports this finding. We can see that the rate of "asymptomatic" cases is significantly lower compared to those who experience symptoms. Additionally, the graph suggests that the rate of symptomatic cases increases with age.

### 3 Answering Questions

#### 3.1 Use the appropriate statistics and plots to answer the following questions:

1. Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19?

##### 1.1 Visualization

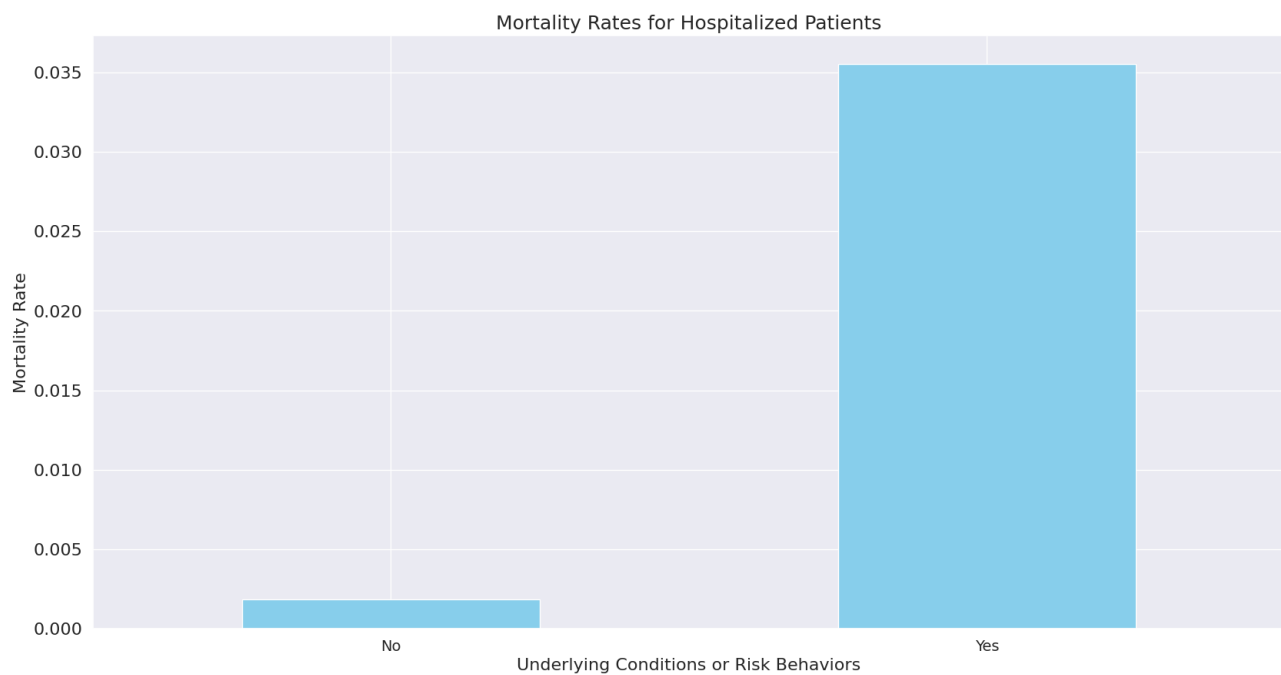


Figure 15

##### 1.2 Commentary

Patients with underlying medical conditions have a significantly higher mortality rate compared to those without underlying medical conditions.

## 2. Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19?

### 2.1 Visualization

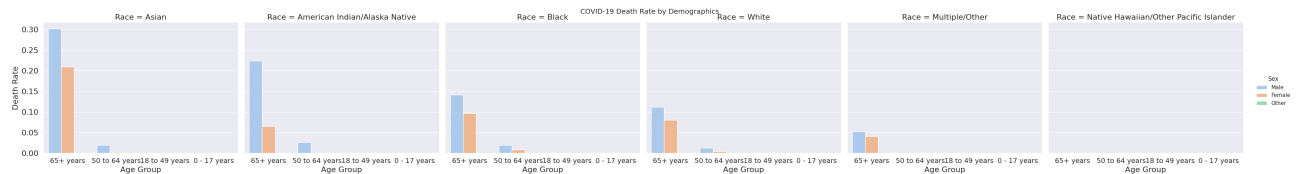


Figure 16

### 2.2 Commentary

#### Most at Risk

- - Older age groups (65 years and older)
  - Men appear to be at highest risk, especially among Asians and American Indians/Alaska Natives, with increased mortality 30.18 % and 22.34% respectively.
  - Women also have significant mortality rates, with Asian women at 20.96% and black women at 9.67%.
- Men 50 to 64 years old
  - American Indian/Alaska Native men have a mortality rate of 2.66%
  - Asian and black men in this age group have a mortality rate of about 2%.

#### Least at Risk

- - Younger age groups (0-17 years)
  - Across all races and genders, the mortality rate was 0%, indicating a very low risk.
- -Older women 18 and older to 49
  - generally have low mortality rates, highest in black women at 0.063%.

### 3. Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19?

#### 3.1 Visualization

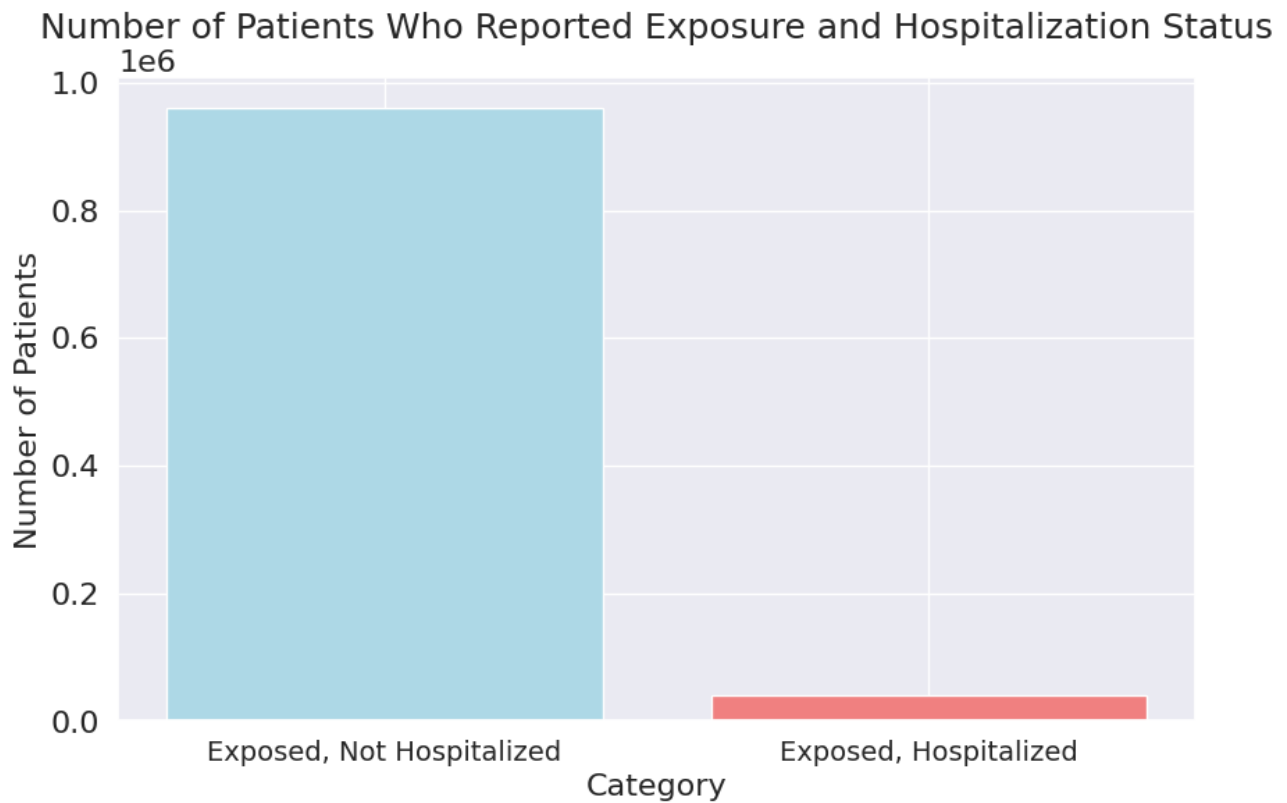


Figure 17

#### 3.2 Commentary

The analysis shows that among a million patients who mentioned travel or gatherings before falling ill only 3.90% (38,965 patients) needed hospital care. This suggests that most of those who mentioned exposure did not need to go to the hospital due, to factors, like the seriousness of the exposure the patients health condition or timely interventions.

#### 4. Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19?

##### 4.1 Visualization

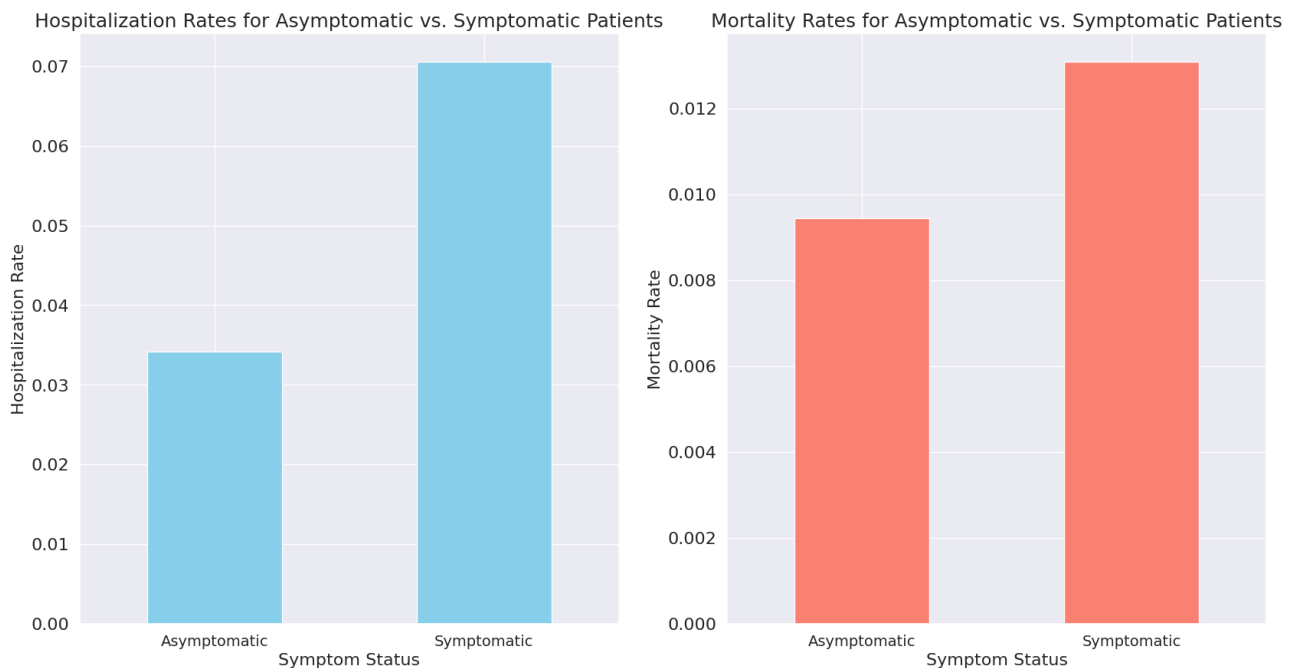


Figure 18

##### 4.2 Commentary

###### Hospitalization Rates:

Asymptomatic patients had a hospitalization rate of 3.42%, while symptomatic patients had a higher hospitalization rate of 7.06%.

This suggests that asymptomatic COVID patients are less likely to be hospitalized compared to symptomatic patients.

###### Mortality Rates:

Asymptomatic individuals had a mortality rate of 0.94% while symptomatic patients experienced a mortality rate of 1.31%. This suggests that asymptomatic COVID patients are less likely to succumb to the illness compared to those who exhibit symptoms.

## 5. Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19?

### 5.1 Visualization

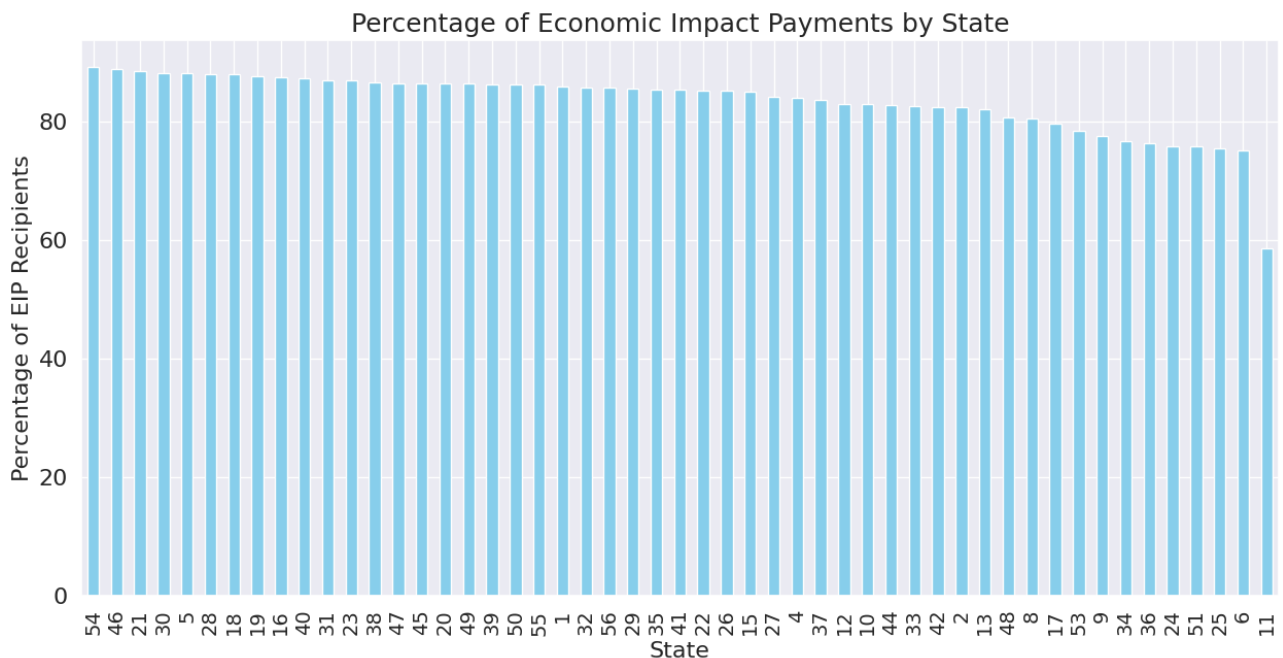


Figure 19

### 5.2 Commentary

the analysis was conducted to examine the distribution of Economic Impact Payments (stimulus checks) across different states. The data indicated that residents in state 54 (abbreviation) had the highest percentage receiving these payments. A corresponding bar chart (not shown) visually depicts this distribution for all states, allowing for easy comparison.

### 3.2 Come up with 5 more bi-variate/multivariate analysis questions and similarly answer each with appropriate visuals and commentary:

1. How does the rate of COVID-related hospitalization vary across different age groups and states?

#### 1.1 Visualization

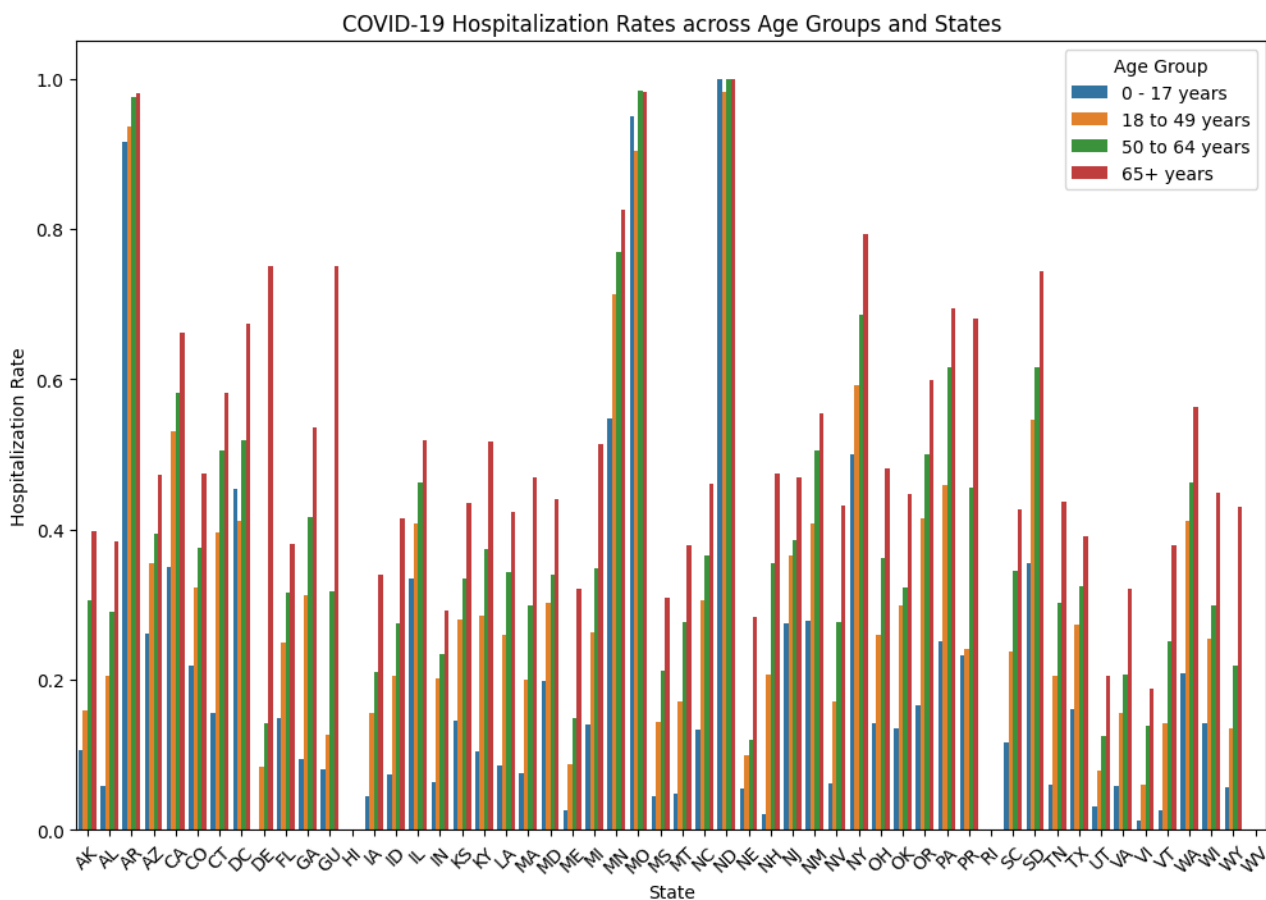


Figure 20

#### 1.2 Commentary

The examination of COVID-related hospitalization rates, among age groups and states uncovers intriguing trends. The bar graph illustrates fluctuations in hospitalization rates suggesting that the likelihood of being hospitalized varies depending on both age and location

#### 1.3 Interpretation

State ND has the highest hospitalization rates for the all age groups :

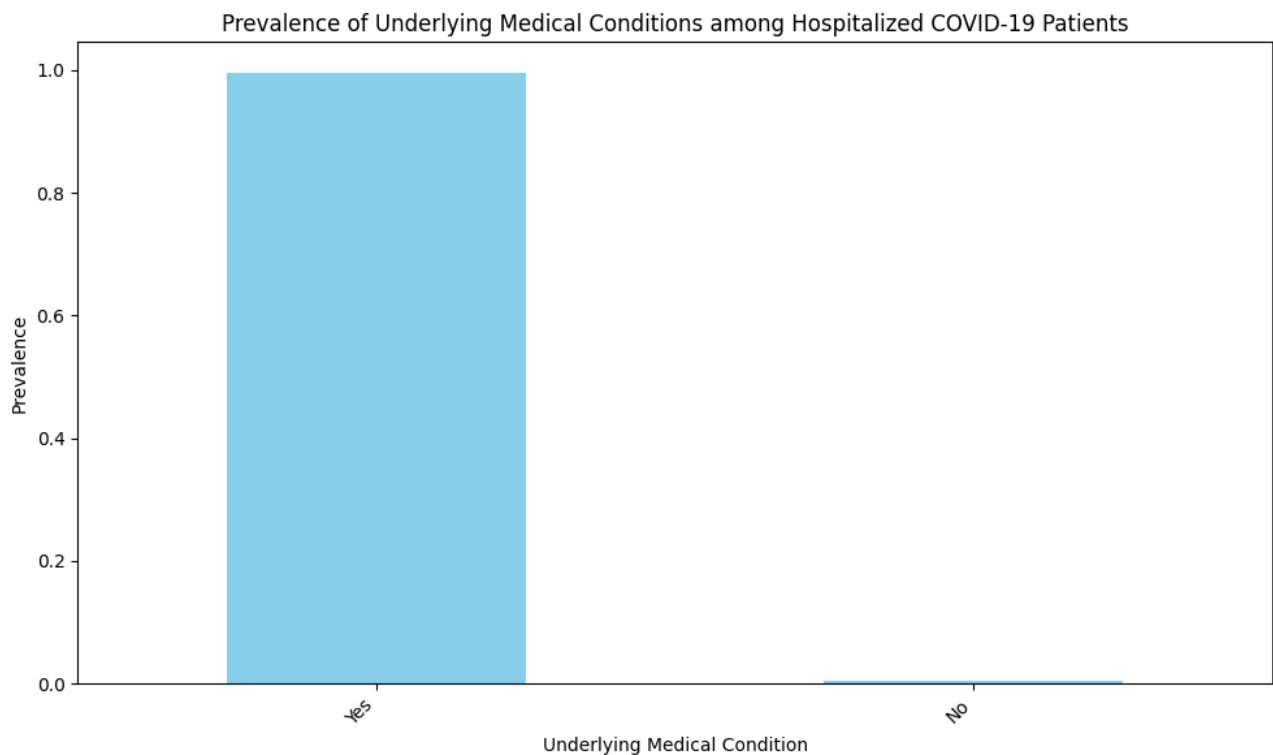
- +65 years
- 0 to 17 years
- 50 to 64 years
- 18 to 49 years

states HI, WV , RI has 0 hospitalization rates

---

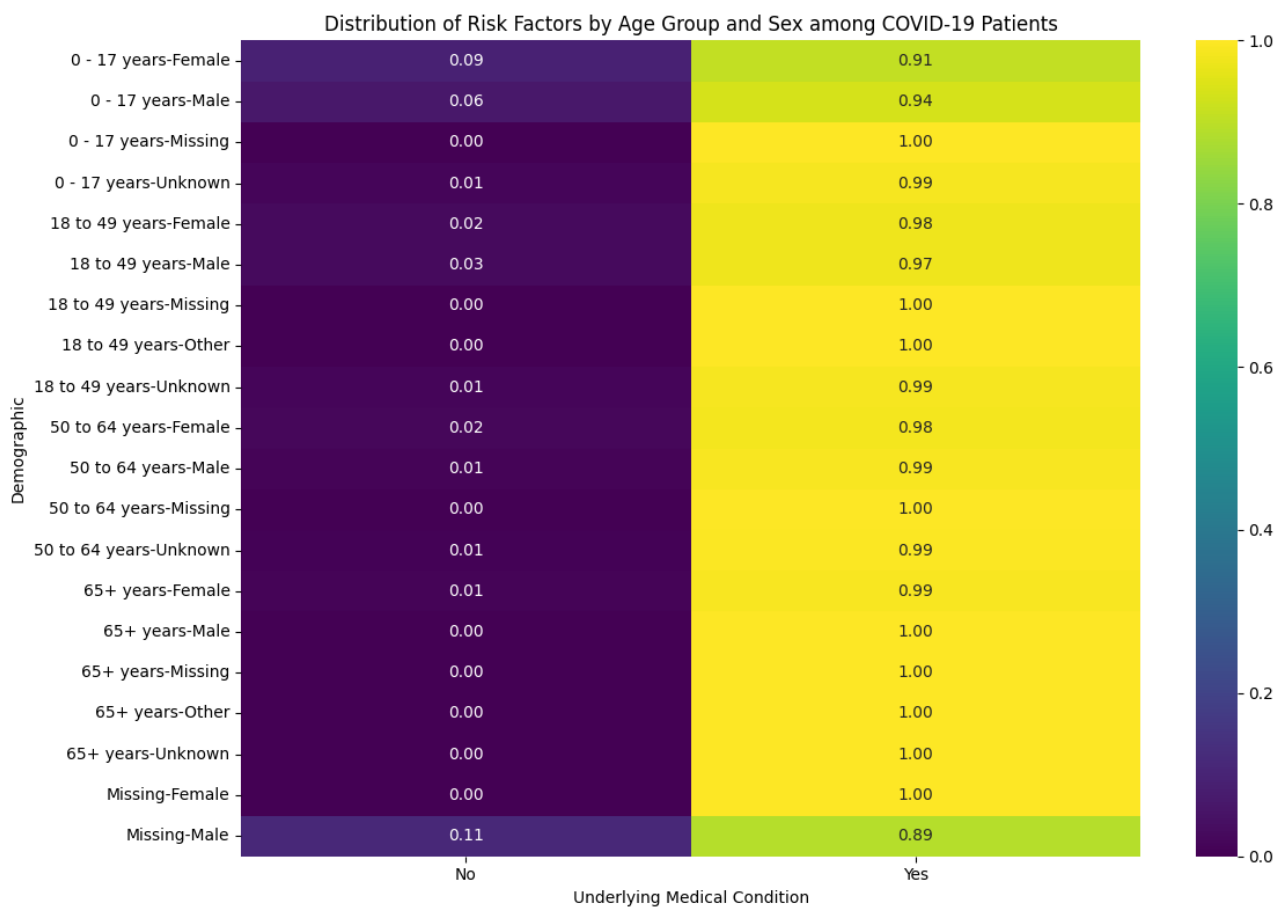
**2. What are the most common risk factors associated with COVID-19-related hospitalizations, and how do these factors vary across different age groups and genders? ?**

### 2.1 Visualization



**Figure 21**





**Figure 22**

## 2.3 Commentary

The bar chart illustrates the frequency of existing conditions, in patients who are hospitalized. The heatmap displays how the presence of conditions varies across age groups and genders among hospitalized individuals with color intensity indicating the percentage of patients, in each group.

## 2.3 interpretation

Among various age groups and genders with a darker color indicating the presence of such conditions. This indicates that having underlying issues increases the risk of being hospitalized due to COVID 19. The data suggests that as age groups increase the prevalence of conditions also rises. The darkest colors on the heatmap are seen in the age groups (such as those aged 65 and above) indicating that older individuals, with COVID 19 are more likely to have underlying conditions compared to younger individuals.

### 3. Are COVID-19 patients with underlying medical conditions more likely to be admitted to the ICU compared to those without underlying conditions ?

#### 3.1 Visualization

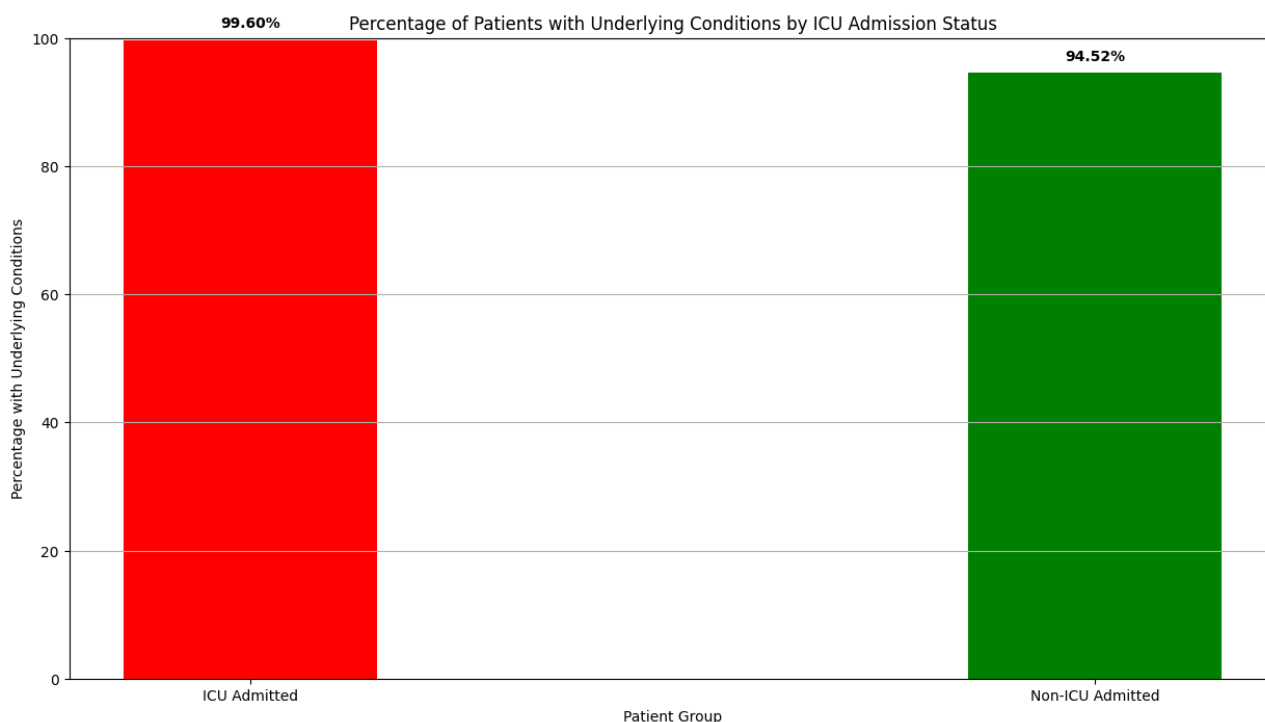


Figure 23

#### 3.2 Commentary

The study shows that 99.60% of COVID-19 patients admitted to the ICU had pre-medical conditions while only 94.52% of non-ICU patients had underlying health issues. This suggests a connection between existing conditions and the severity of COVID-19 outcomes. These results emphasize the importance of focusing on individuals with health concerns, for vaccination and early treatments to decrease ICU admissions and enhance results.

#### 3.3 Interpretation

Percentage of ICU-admitted patients with underlying conditions: 99.6  
Percentage of non-ICU-admitted patients with underlying conditions: 94.5

High Risk for ICU Patients:

Nearly all ICU patients had underlying conditions, indicating a strong link to severe COVID-

19. Significant Risk for Non-ICU Patients:

A substantial portion of non-ICU patients also had underlying conditions, though less severe.

4. Do asymptomatic COVID-19 patients have lower rates of hospitalization compared to symptomatic patients? How does this differ based on demographic factors such as age and sex?

#### 4.1 Visualization

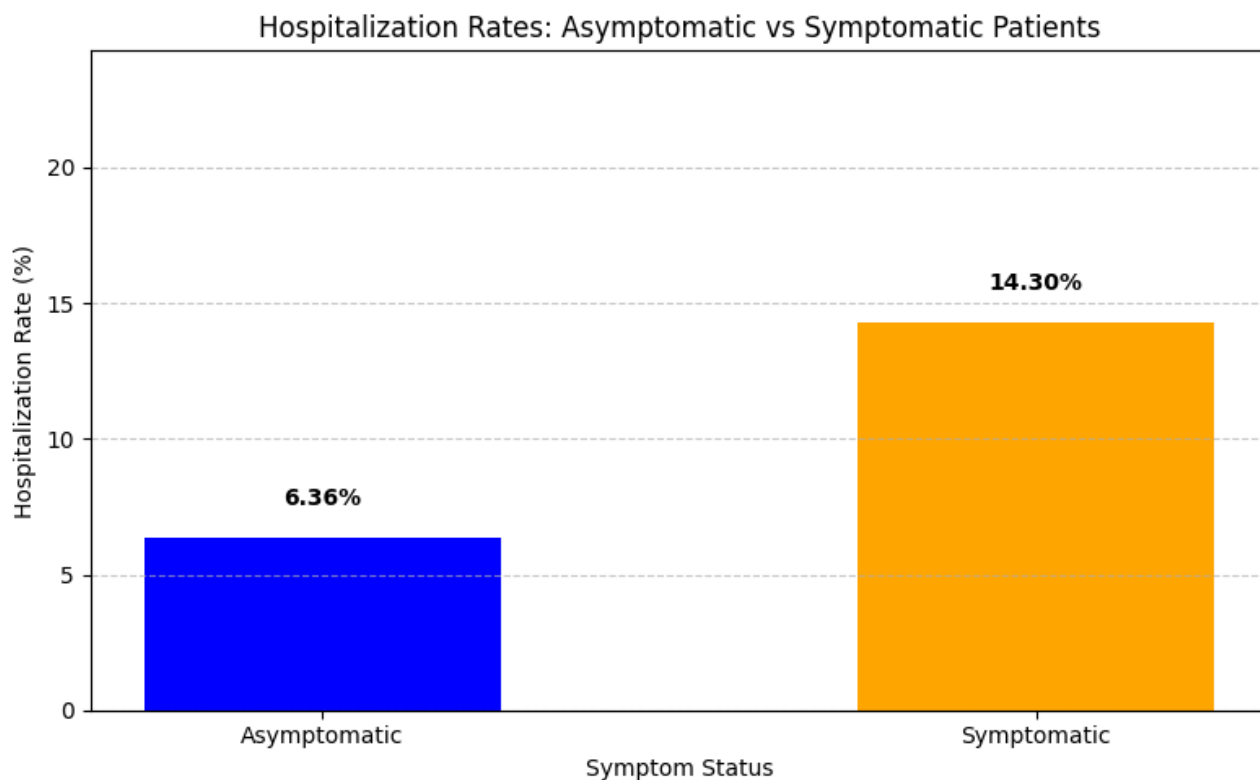
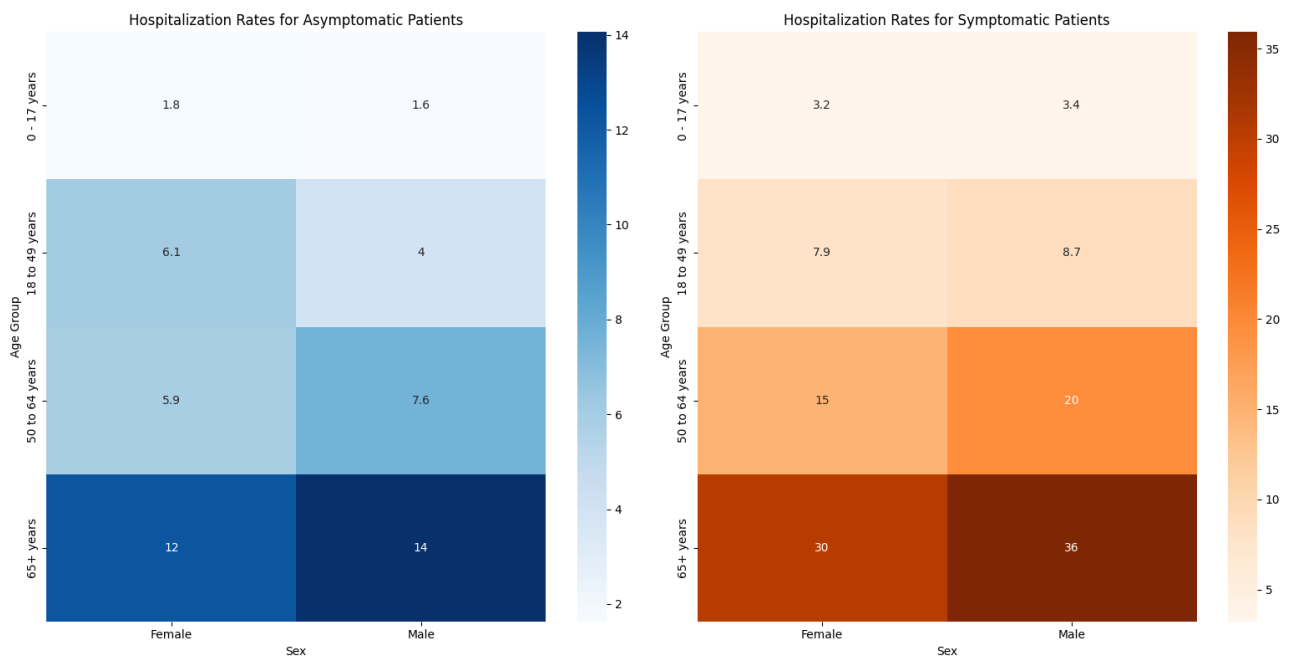


Figure 24



**Figure 25**

## 4.2 Commentary

The data shows that individuals, with symptoms of COVID-19 are more likely to require hospitalization compared to those without symptoms. As people get older the likelihood of being hospitalized increases for both groups with the rates seen in the age bracket (65 years and above). Generally, males tend to have hospitalization rates, than females in the older age categories. These results underscore the need to closely monitor individuals showing symptoms, older adults, and males who face a risk of needing hospital care.

## 4.3 Interpretation

### Overall Hospitalization Rates

Asymptomatic Patients: 6.36% Symptomatic Patients: 14.30% Hospitalization Rates by Age and Sex

For individuals aged 0 to 17 years, the hospitalization rate is 1.77%, for females and 1.63% for males. In the age group of 18 to 49 years, the hospitalization rate is 6.15% for females and 4.00% for males. Among those aged between 50 to 64 years, the hospitalization rate is reported at 5.94% for females and 7.56% for males. For individuals aged 65 and above the hospitalization rates stand at 12.25% for females and 14.06% for males. In patients within the age range of 0 to 17 years, the hospitalization rate is noted at, approximately 3.23

5.What is the frequency of common symptoms reported by COVID-19 patients, and how do they vary across different age groups and genders?

5.1 Visualization

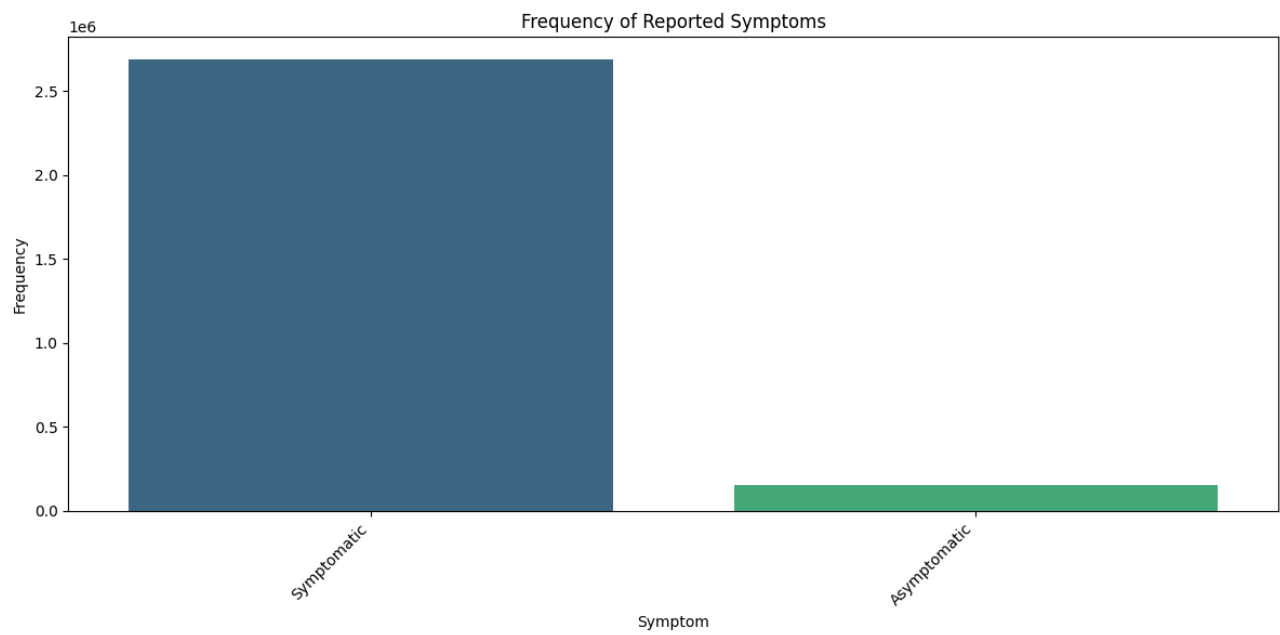


Figure 26

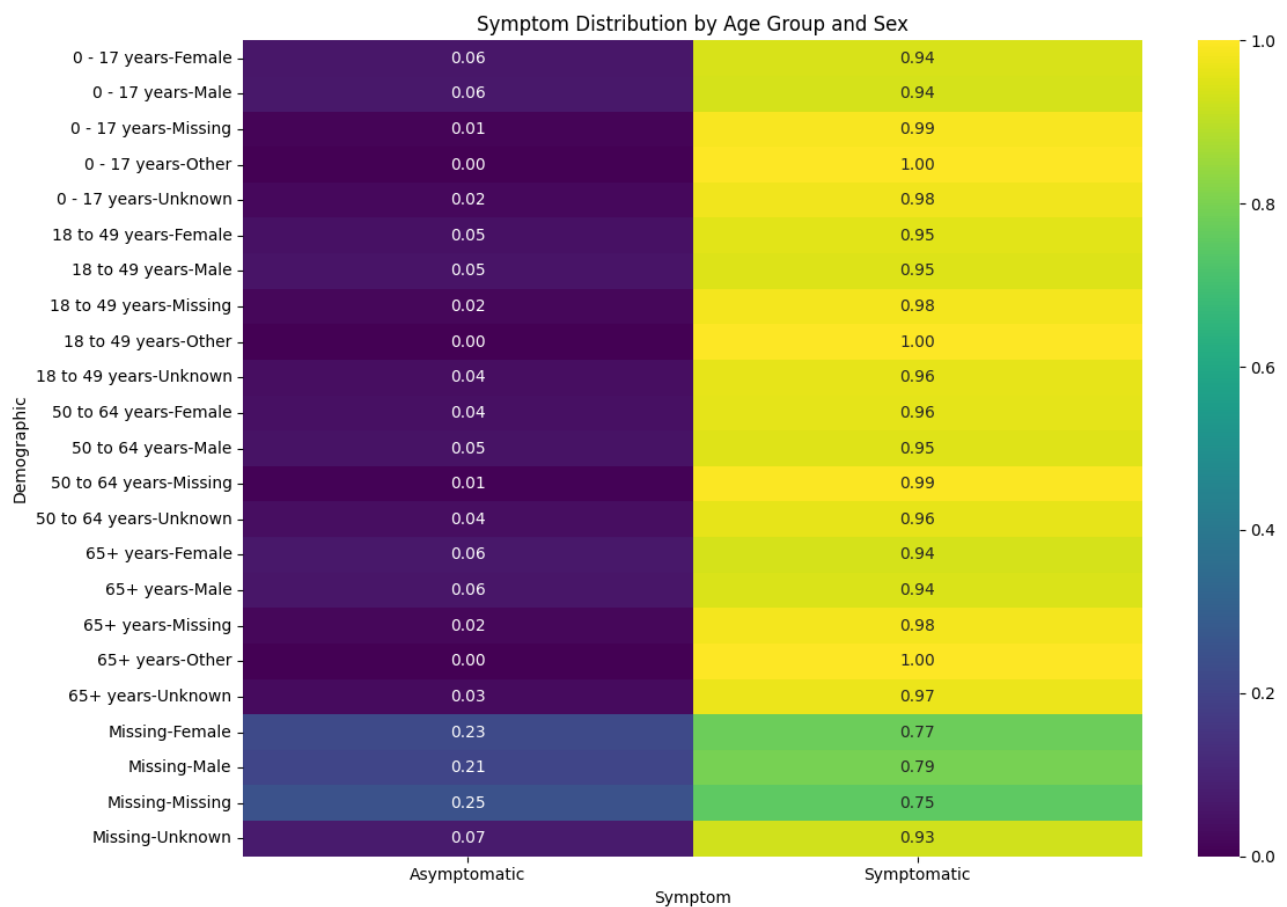


Figure 27

## 5.2 Commentary

Our examination of COVID 19 symptom data provided us with findings regarding how symptoms are distributed and how prevalent they are, among various demographic groups. The analysis of frequency shed light on the symptoms experienced by COVID-19 symptoms patients, whether asymptomatic. Additionally, the heatmap display showed us how each symptom was reported across age groups and genders. We noticed variations in distribution among demographic groups suggesting possible differences, in how symptoms manifest and their severity.

## 5.3 Interpretation

Symptoms Frequency Overview;

The reported symptoms are typically displayed prominently on the bar graph. By looking at the bar graph we can see how often each symptom occurs, giving us an idea of which symptoms are most common among individuals, with COVID-19. Symptom Patterns Across Age and Gender;

The heatmap illustrates the proportion of reported symptoms across age groups and genders. Analyzing the heatmap allows us to pinpoint whether certain symptoms are more prevalent in

categories. For instance, we may notice variations in symptom prevalence between males and females, across age brackets.

---

## 4 Hypothesis Testing

### 4.1 Claim: “There is a strong association between the probability of death due to COVID-19 and patient demographics”:

I will solve by two different tests, but the hypotheses will be the same

#### 4.1.1 First Test Chosen: Anova

we are comparing the mean probability of death across different demographic groups and ANOVA allows us to determine whether there are statistically significant differences in the mean probability of death among the different demographic groups.

- Formulating the Hypothesis Test:
  - $H_0$ : There is no association between the probability of death due to COVID-19 and patient demographics.
  - $H_1$ : There is a strong association between the probability of death due to COVID-19 and patient demographics.
- Results
  - $F_{statistic}$ : 2823.3188114282116
  - $p - value$  : 0
- Comment: Since the p-value (0.0) is less than the significance level ( = 0.05), we reject the null hypothesis ( $H_0$ ). Therefore, we conclude that there is an association between the probability of death due to COVID-19 and patient demographics.

#### 4.1.2 Second Test Chosen: Chi-squared

The chi-squared test of independence is used to determine whether there is a significant association between two categorical variables.

- Formulating the Hypothesis Test:
  - $H_0$ : There is no association between the probability of death due to COVID-19 and patient demographics.
  - $H_1$ : There is a strong association between the probability of death due to COVID-19 and patient demographics.
- Results
  - $Chi - squareStatistic$ : 126131.64445956664

–  $p$  – value : 0

- Comment: Similar to the ANOVA test, the p-value (0.0) from the chi-squared test is less than the significance level ( $= 0.05$ ). Hence, we reject the null hypothesis ( $H_0$ ). This implies that there is a significant association between the probability of death due to COVID-19 and patient demographics.

## 5 Regression Analysis

1. Model Coefficients and P-Values: - Each variable (gender proportions, age groups, ICU proportion, hospitalization proportion) - These coefficients indicate the direction and strength of the relationship between each variable and the target variable (proportion of deaths) - A positive coefficient suggests that as the value of the variable increases, the proportion of deaths also increases. Conversely, a negative coefficient indicates an inverse relationship. - The p-value associated with each coefficient tells you how statistically significant the relationship is.

2. Good vs. Bad Predictors: - A low p-value (typically less than 0.05) suggests the relationship is unlikely due to chance and can be considered a good predictor. A high p-value indicates the relationship might be due to random variation and shouldn't be strongly relied upon.

3. Correlated Predictors: -Analyze the correlation between the predictor variables (gender proportions, age groups, ICU and hospitalization proportions). -If two predictors are highly correlated, they might provide redundant information. This can affect the model's interpretability and potentially lead to inaccurate estimates.

4. Model Improvements: - Removing the intercept impact the model's baseline prediction for death proportion. - Adding higher-order terms (squares) capture more complex relationships between predictors and the target variable. - Removing outliers, reduce their influence on the model and potentially improve its accuracy for the majority of the data.



```
model <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8)
print(model.summary()) # Print the model summary
```



#### OLS Regression Results

```
=====
Dep. Variable:      target_variable      R-squared:                0.915
Model:              OLS                  Adj. R-squared:           0.907
Method:             Least Squares        F-statistic:             120.8
Date:               Wed, 22 May 2024      Prob (F-statistic):      1.85e-23
Time:               20:41:36              Log-Likelihood:          86.187
No. Observations:   50                   AIC:                     -162.4
Df Residuals:       45                   BIC:                     -152.8
Df Model:           4
Covariance Type:    nonrobust
=====
```


	coef	std err	t	P> t	[0.025	0.975]
const	-0.1504	0.044	-3.434	0.001	-0.239	-0.062
x1	0.3436	0.080	4.298	0.000	0.183	0.505
x2	-0.4940	0.102	-4.862	0.000	-0.699	-0.289
x3	0.0687	0.181	0.379	0.706	-0.296	0.433
x4	-7.954e-17	1.15e-17	-6.932	0.000	-1.03e-16	-5.64e-17
x5	0	0	nan	nan	0	0
x6	0	0	nan	nan	0	0
x7	0.0978	0.174	0.561	0.578	-0.253	0.449
x8	1.3593	0.196	6.941	0.000	0.965	1.754

```
=====
Omnibus:              1.373      Durbin-Watson:              2.498
Prob(Omnibus):        0.503      Jarque-Bera (JB):           1.185
Skew:                 -0.194      Prob(JB):                   0.553
Kurtosis:              2.353      Cond. No.:                  2.95e+35
=====
```

#### Notes:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The smallest eigenvalue is 1.04e-69. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

Figure 28



```
Model Coefficients:
const      -1.504002e-01
x1          3.435652e-01
x2         -4.939654e-01
x3          6.867941e-02
x4         -7.954431e-17
x5          0.000000e+00
x6          0.000000e+00
x7          9.775042e-02
x8          1.359265e+00
dtype: float64
```

Figure 29

```
Model P-values:
const      1.286877e-03
x1          9.128964e-05
x2          1.452229e-05
x3          7.062935e-01
x4          1.286158e-08
x5          NaN
x6          NaN
x7          5.778061e-01
x8          1.248327e-08
dtype: float64
```

Figure 30

Good Predictors (p-value < 0.05):

Variable 0: p-value = 0.001

Variable 1: p-value = 0.000

Variable 2: p-value = 0.000

Variable 4: p-value = 0.000

Variable 8: p-value = 0.000

Bad Predictors (p-value >= 0.05):

Variable 3: p-value = 0.706

Variable 7: p-value = 0.578

Figure 31

Correlation Matrix:

```
[[      nan      nan      nan      nan      nan      nan
      nan      nan      nan]
 [      nan  1.      -1.     -0.35217936      nan      nan
      nan -0.16141082 -0.03364311]
 [      nan -1.       1.      0.35217936      nan      nan
      nan  0.16141082  0.03364311]
 [      nan -0.35217936  0.35217936  1.      nan      nan
      nan -0.1374617  -0.15518132]
 [      nan      nan      nan      nan      nan      nan
      nan      nan      nan]
 [      nan      nan      nan      nan      nan      nan
      nan      nan      nan]
 [      nan      nan      nan      nan      nan      nan
      nan      nan      nan]
 [      nan -0.16141082  0.16141082 -0.1374617      nan      nan
      nan  1.      0.93072364]
 [      nan -0.03364311  0.03364311 -0.15518132      nan      nan
      nan  0.93072364  1.      ]]
```

Figure 32



## Model Summary (No Intercept):

## OLS Regression Results

```

=====
Dep. Variable:    target_variable    R-squared:                0.915
Model:            OLS                Adj. R-squared:           0.907
Method:           Least Squares      F-statistic:             120.8
Date:            Wed, 22 May 2024    Prob (F-statistic):       1.85e-23
Time:            20:41:41            Log-Likelihood:           86.187
No. Observations: 50                AIC:                     -162.4
Df Residuals:     45                BIC:                     -152.8
Df Model:         4
Covariance Type:  nonrobust
=====

```

```

=====
              coef      std err          t      P>|t|      [0.025      0.975]
-----
x1              0.1932      0.079        2.434      0.019      0.033      0.353
x2             -0.6444      0.134       -4.792      0.000     -0.915     -0.374
x3              0.0687      0.181        0.379      0.706     -0.296      0.433
const          -1.492e-16  1.66e-16     -0.897      0.375    -4.84e-16  1.86e-16
x4               0         0          nan         nan         0         0
x5               0         0          nan         nan         0         0
x6              0.0978      0.174        0.561      0.578     -0.253      0.449
x7              1.3593      0.196        6.941      0.000      0.965      1.754
=====

```

```


=====
Omnibus:                 1.373    Durbin-Watson:                2.498
Prob(Omnibus):            0.503    Jarque-Bera (JB):                1.185
Skew:                    -0.194    Prob(JB):                        0.553
Kurtosis:                 2.353    Cond. No.                        2.17e+34
=====

```

## Notes:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.  
 [2] The smallest eigenvalue is 8.58e-68. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

Figure 33

 Model Summary (Higher Order Terms):

OLS Regression Results

=====

Dep. Variable:	target_variable	R-squared:	0.960
Model:	OLS	Adj. R-squared:	0.952
Method:	Least Squares	F-statistic:	123.1
Date:	Wed, 22 May 2024	Prob (F-statistic):	3.67e-26
Time:	20:41:41	Log-Likelihood:	105.10
No. Observations:	50	AIC:	-192.2
Df Residuals:	41	BIC:	-175.0
Df Model:	8		
Covariance Type:	nonrobust		

=====

	coef	std err	t	P> t	[0.025	0.975]
-----	-----	-----	-----	-----	-----	-----
const	-0.3192	0.399	-0.800	0.428	-1.125	0.486
x1	-0.1897	0.243	-0.781	0.439	-0.680	0.301
x2	-0.1295	0.169	-0.769	0.446	-0.470	0.211
x3	-0.5453	0.506	-1.077	0.288	-1.568	0.477
x4	1.684e-16	2.99e-15	0.056	0.955	-5.87e-15	6.2e-15
x5	-1.053e-15	8.48e-16	-1.241	0.222	-2.77e-15	6.6e-16
x6	-1.856e-15	1.49e-15	-1.244	0.220	-4.87e-15	1.16e-15
x7	0.1052	0.828	0.127	0.900	-1.567	1.778
x8	0.4294	0.285	1.508	0.139	-0.145	1.004
x9	-0.3192	0.399	-0.800	0.428	-1.125	0.486
x10	1.3925	1.642	0.848	0.401	-1.923	4.708
x11	1.4526	1.725	0.842	0.405	-2.031	4.936
x12	5.4895	3.500	1.568	0.125	-1.580	12.559
x13	0	0	nan	nan	0	0
x14	0	0	nan	nan	0	0
x15	0	0	nan	nan	0	0
x16	-0.0589	0.846	-0.070	0.945	-1.768	1.651
x17	1.5697	0.677	2.318	0.026	0.202	2.938

=====

Omnibus:	1.743	Durbin-Watson:	1.558
Prob(Omnibus):	0.418	Jarque-Bera (JB):	1.176
Skew:	0.011	Prob(JB):	0.555
Kurtosis:	2.249	Cond. No.	3.58e+51

Figure 34

## Model Summary (Cleaned):

## OLS Regression Results

Dep. Variable:	target_variable	R-squared:	0.915			
Model:	OLS	Adj. R-squared:	0.907			
Method:	Least Squares	F-statistic:	120.8			
Date:	Wed, 22 May 2024	Prob (F-statistic):	1.85e-23			
Time:	20:41:41	Log-Likelihood:	86.187			
No. Observations:	50	AIC:	-162.4			
Df Residuals:	45	BIC:	-152.8			
Df Model:	4					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
const	-0.1504	0.044	-3.434	0.001	-0.239	-0.062
x1	0.3436	0.080	4.298	0.000	0.183	0.505
x2	-0.4940	0.102	-4.862	0.000	-0.699	-0.289
x3	0.0687	0.181	0.379	0.706	-0.296	0.433
x4	-7.954e-17	1.15e-17	-6.932	0.000	-1.03e-16	-5.64e-17
x5	0	0	nan	nan	0	0
x6	0	0	nan	nan	0	0
x7	0.0978	0.174	0.561	0.578	-0.253	0.449
x8	1.3593	0.196	6.941	0.000	0.965	1.754
=====						
Omnibus:	1.373	Durbin-Watson:	2.498			
Prob(Omnibus):	0.503	Jarque-Bera (JB):	1.185			
Skew:	-0.194	Prob(JB):	0.553			
Kurtosis:	2.353	Cond. No.	2.95e+35			
=====						

## Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The smallest eigenvalue is 1.04e-69. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

Figure 35

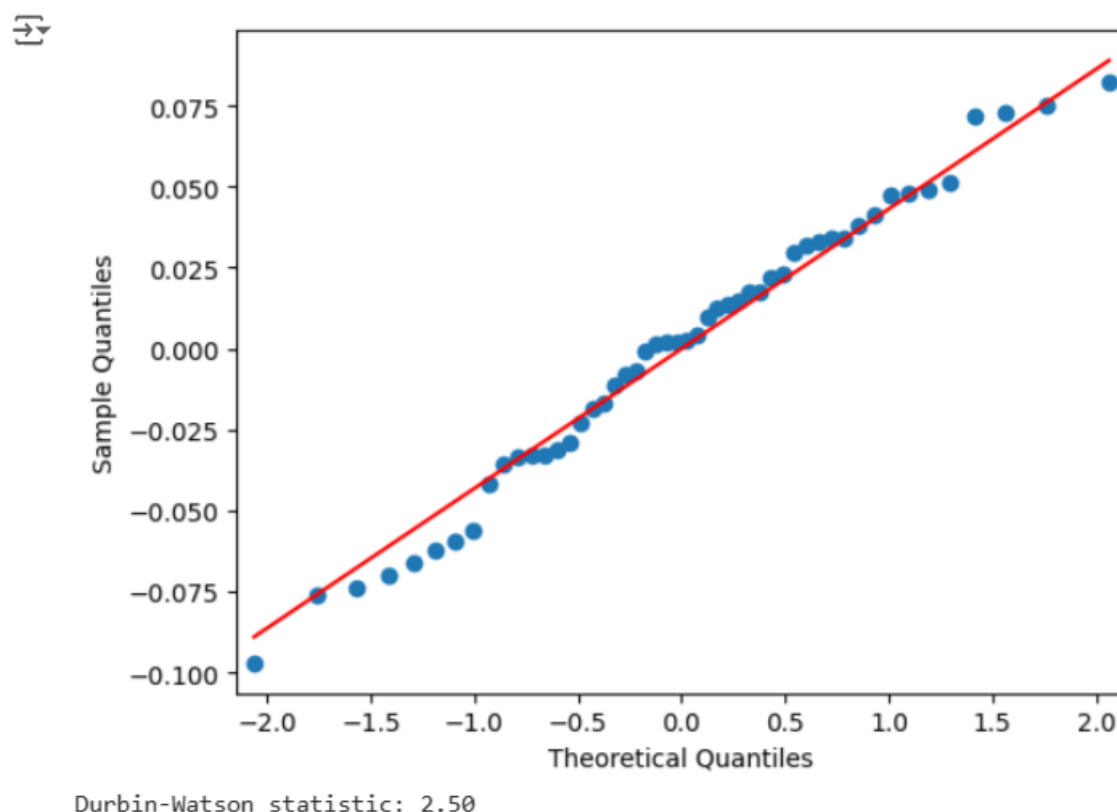


Figure 36

The last figure provides a test of our model's performance in comparison to our predicted outcomes.

## 6 Machine Learning Classifier

### 6.1 Machine Learning Classifier to predict the likelihood of death due to COVID-19 using any/all of the relevant attributes in the COVID-19 case surveillance dataset:

I will solve by two different tests, but the hypotheses will be the same.

#### 5.1 First Data Preparation

The dataset was first cleaned by removing duplicates and filtering out non-informative values. To address class imbalance, we under-sampled the majority class (those who did not die) to ensure balanced representation.



## 5.2 Feature Selection

The analysis considered factors such, symptom status, location of residence age group, gender, race, ethnicity, hospitalization ICU admission and preexisting health conditions.

## 5.3 Model Training

A preprocessing system was set up to manage variables by using one encoding. Next the data was divided into training and testing groups with 80% allocated for training and 20 %, for testing purposes.

## 5.4 Model Approach

We used an XGBoost classifier in our model. The process combined the steps, with the XGBoost classifier to create a workflow from the initial data to making predictions.

## 5.5 Model Performance

The model's performance was evaluated using several metrics:

- Accuracy: 0.98
- Precision: 0.83
- Recall: 0.42
- F1-Score: 0.56

```

Accuracy: 0.9837793577051143
Precision: 0.8359219434488252
Recall: 0.42271674554425537
F1-Score: 0.5614926770547717
Confusion Matrix:
[[393483    824]
 [  5733   4198]]
Classification Report:

```

	precision	recall	f1-score	support
0	0.99	1.00	0.99	394307
1	0.84	0.42	0.56	9931
accuracy			0.98	404238
macro avg	0.91	0.71	0.78	404238
weighted avg	0.98	0.98	0.98	404238

Figure 37

## 7 Conclusion

The CDC and US Census Bureau’s COVID-19 data have been thoroughly analysed, and the results have given important new information about the trends and effects of the pandemic throughout the country. Important trends in COVID-19-related hospitalisations, fatalities, employment, demography, and healthcare access were identified by the exploratory study. A substantial correlation between patient demographics and the likelihood of dying from the virus was confirmed by the hypothesis testing.

The gender distribution, age distribution, ICU admission rates, and hospitalisation rates were found to be the most significant predictors of COVID-19 death rates by regression modelling. Public health professionals can use this information to more effectively target interventions and distribute resources to the most vulnerable populations. Based on the patient data at hand, the machine learning classification model also showed promise in predicting the likelihood of death with a respectable degree of accuracy. Healthcare professionals may find this kind of predictive modelling helpful in the assessing and treatment of COVID-19 cases.

All things considered, this study has made use of the extensive COVID-19 datasets to produce useful insights that can guide ongoing efforts to prepare for and respond to pandemics. The results emphasise how crucial thorough data collection and analysis are in directing evidence-based public health policy. However, while interpreting the results, one should take into account

limitations like as possible data biases and the evolving nature of the epidemic. Sustained observation and additional investigation are necessary to completely comprehend the enduring consequences of COVID-19 in the United States.

## References

- [1] Leslie Lamport. *TEX: a Document Preparation System*. Addison Wesley, Massachusetts, 2 edition, 1994.

# Appendices