Business Report on Exploratory Data Analysis of COVID-19 Datasets

Introduction

This report presents the findings of an exploratory data analysis performed on two COVID-19 dataset. We aim to investigate various relationships and trends related to COVID-19 hospitalizations, deaths, and the socioeconomic impact on different demographics and job fields. The steps undertaken during this analysis include but are not limited to data importation, cleaning, visualization, and statistical analysis.

Methodology

The analysis followed a structured methodology comprising data import, cleaning, post-processing, visualization, and statistical testing if required. Here is a detailed account of each step:

Data Import

The data was imported from 2 CSVs file using the pandas library. Specific columns relevant to the analysis were selected to ensure focus and clarity.
The columns were imported as “category” data type in pandas to ensure an efficient and quick analysis. The columns used in different parts of the analysis include:

- **age_group**: Categorized age groups of individuals.
- **underlying_conditions_yn**: Presence of underlying medical conditions (Yes/No).
- **icu_yn**: ICU admission status (Yes/No).
- **hosp_yn**: Hospitalization status (Yes/No).
- **death_yn**: Death status (Yes/No).
- **res_state**: State of residence.
- **case_month**: Month and year of the reported case.
- **sex**: Gender of individuals.
- **race**: Race of individuals.
- **kindwork**: Type of employment.
- **expctloss**: Expected employment loss (Yes/No).
- **income**: Household income categories.
- **delay**: Delayed medical treatment (Yes/No).
- **notget**: Unobtained medical treatment (Yes/No).
- **tbirth_year**: Year of birth.

Data Cleaning

Data cleaning involved several steps to ensure accuracy and reliability, as well as improve the memory usage and processing speeds as we were dealing with large datasets (>10,000,000 rows and >20 cols).

- Filtering: Rows with missing, unknown, or irrelevant data were removed.
- Category Management: Unused categories were removed from categorical variables.
- Date Conversion: The `case_month` column was converted to datetime format to facilitate time-series analysis.
- Labeling: Categorical variables were appropriately labeled to enhance readability and interpretation.

Data Visualization

Visualization was a key component of the analysis, enabling the identification of trends and patterns. It was the first step to recognize any relationships in the data. The seaborn and matplotlib libraries were used to create various plots, including:

- Bar (histogram) Plots: To compare percentages of hospitalizations, deaths, and employment loss across different categories.
- Line Plots: To visualize trends over time for hospitalizations and deaths.
- KDE Plots: To show the density distribution of hospitalizations and deaths over time.

Statistical Analysis

Statistical methods, such as the chi-square test, were employed to test hypotheses and assess the significance of observed relationships. The chi-square test of independence was used to evaluate the association between categorical variables. This however was not done for all the questions we were trying to answer as some of them were answered clearly during the visualization stage.

Our Questions:

- Section 1:

1. The total number of hospitalizations versus deaths from COVID-19 over the entire US per month-year timestamp.
2. The average rates of COVID-related deaths relative to patient demographics.
3. The rates of COVID-related hospitalization and death with age (across age groups).
4. Average rate of COVID-related hospitalization and death per state over the entire study period.
5. The relationship between age, pre-existing medical conditions and/or risk behaviors, and rate of admittance to the ICU.
6. The rate of expected employment loss due to COVID-19 and sector of employment.
7. The rate of expected employment loss due to COVID-19 relative to responders demographics.
8. The rate of expected employment loss due to COVID-19 for the top 10 states with the highest rate of COVID hospitalization.
9. The relationship between household income and the rate of delayed/ OR unobtained medical treatment (Due to COVID or otherwise).
10. The relationship between COVID-19 symptom manifestation and age group.

- Section 2:

1. Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19?
2. Who are the people (the demographic segment) that appear to be most at risk of death due to COVID-19? Who is the least at risk?
3. What percent of patients who have reported exposure to any kind of travel /or congregation within the 14 days prior to illness onset end up hospitalized? What percent of those go on to be hospitalized?
4. Are Asymptomatic COVID patients less likely to be hospitalized? Are they less likely to die from their illness?
5. Which state is associated with the highest percentage of Economic Impact (stimulus) payments among survey respondents?
6. What is the relationship between age group and ICU admission rates?
7. How does symptom status impact the rate of ICU admittance?

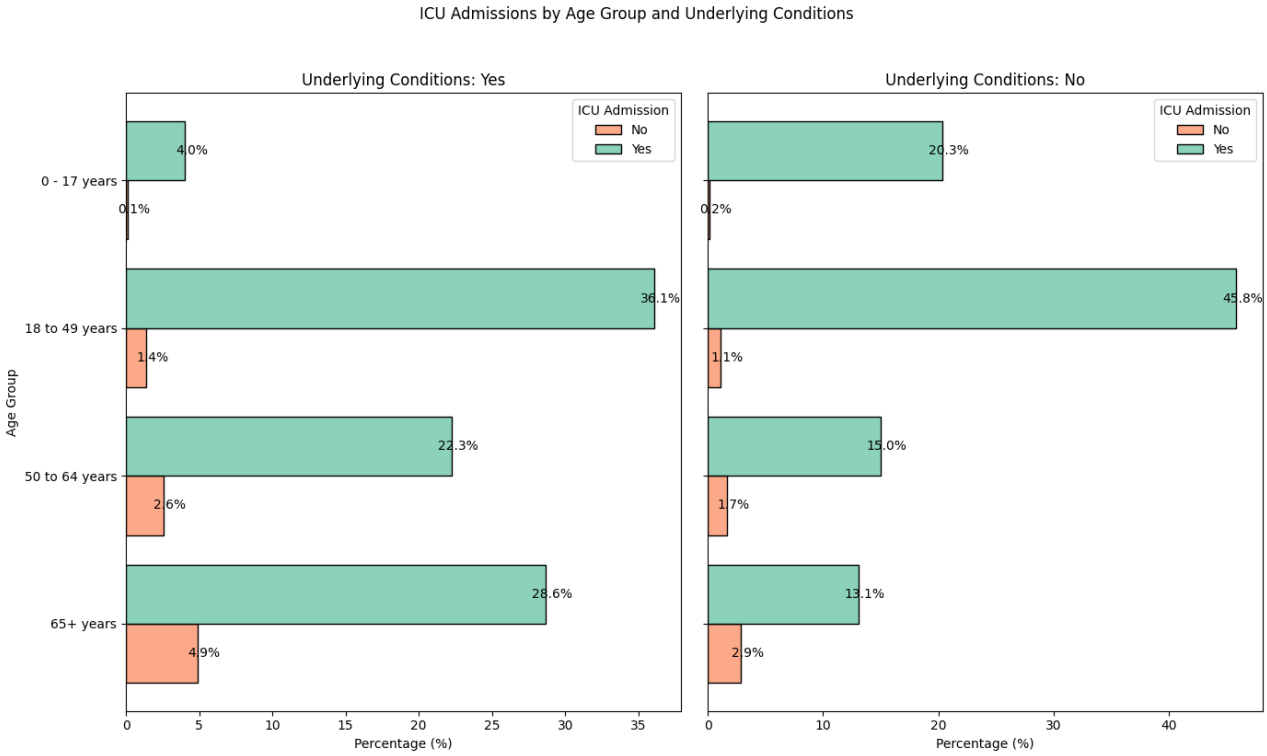
8. How does the length of time between symptom onset and first positive test correlate with hospitalization rates?
9. What is the rate of patients who are reported dead after ICU admission?
10. What is the impact of symptom onset interval on death rates?

Analysis and Significant Findings

1. Relationship between Age, Pre-existing Medical Conditions, and ICU Admissions

Steps:

- Data on age, underlying conditions, and ICU admissions were filtered and cleaned.
- Visualizations were created to compare ICU admissions across age groups with and without underlying conditions.
- A chi-square test was performed to assess the statistical significance.



Findings:

- There is a significant association between ICU admissions and the presence of underlying conditions in different age groups.
- Older age groups with underlying conditions are more likely to be admitted to the ICU.

2.

```
Contingency Table:
icu_yn                                    No    Yes
age_group      underlying_conditions_yn
0 - 17 years   No                      85618    670
               Yes                      7091    195
18 to 49 years No                     193049   4827
               Yes                     63179   2437
50 to 64 years No                      63081   6996
               Yes                     38987   4485
65+ years      No                      55042  12269
               Yes                     50161   8577

Chi-square statistic: 33628.01056573591
p-value: 0.0
Degrees of freedom: 7

Expected frequencies:
[[ 80437.35821836   5850.64178164]
 [  6791.98256975    494.01743025]
 [184459.28396551  13416.71603449]
 [ 61166.99537428   4449.00462572]
 [ 65325.52326938   4751.47673062]
 [ 40524.43951034   2947.56048966]
 [ 62747.06817908   4563.93182092]
 [ 54755.34891329   3982.65108671]]

Reject the null hypothesis: There is a significant association between ICU admission and the presence of
 underlying conditions in different age groups.
```

Hospitalizations and Deaths Over Time by State

Steps:

- Data on hospitalizations, deaths, and states were analyzed.
- Time-series plots were created to visualize hospitalizations and deaths over time for each state.

Findings:

- Hospitalization and death rates vary significantly between states.
- Trends over time indicate varying peaks and durations of COVID-19 impact across states.
- The plots will be included in the appendix as they are too large

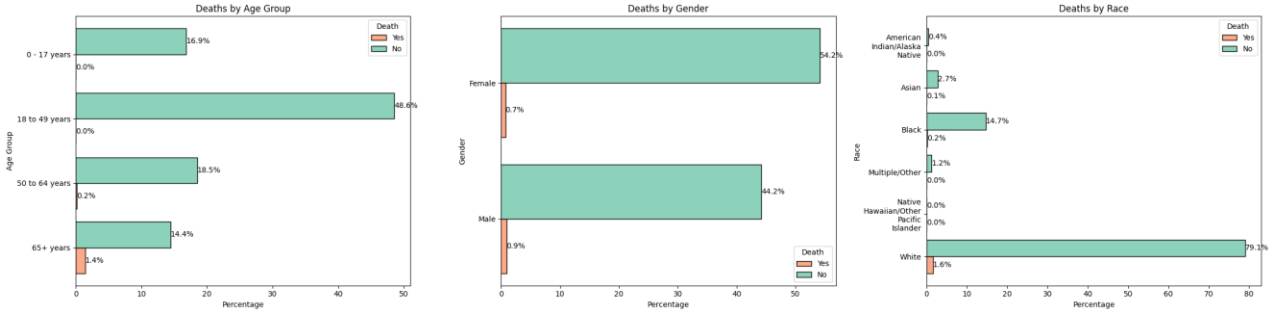
3. COVID-19 Related Deaths by Demographics (Age, Gender, Race)

Steps:

- Data on deaths, age, gender, and race were analyzed.
- Bar plots were created to show the percentage of deaths across different demographics.

Findings:

- Higher death rates are observed in older age groups, males, and certain racial groups.
- These disparities highlight the unequal impact of COVID-19 across different demographic segments.



4. Expected Employment Loss Due to COVID-19 by Sector

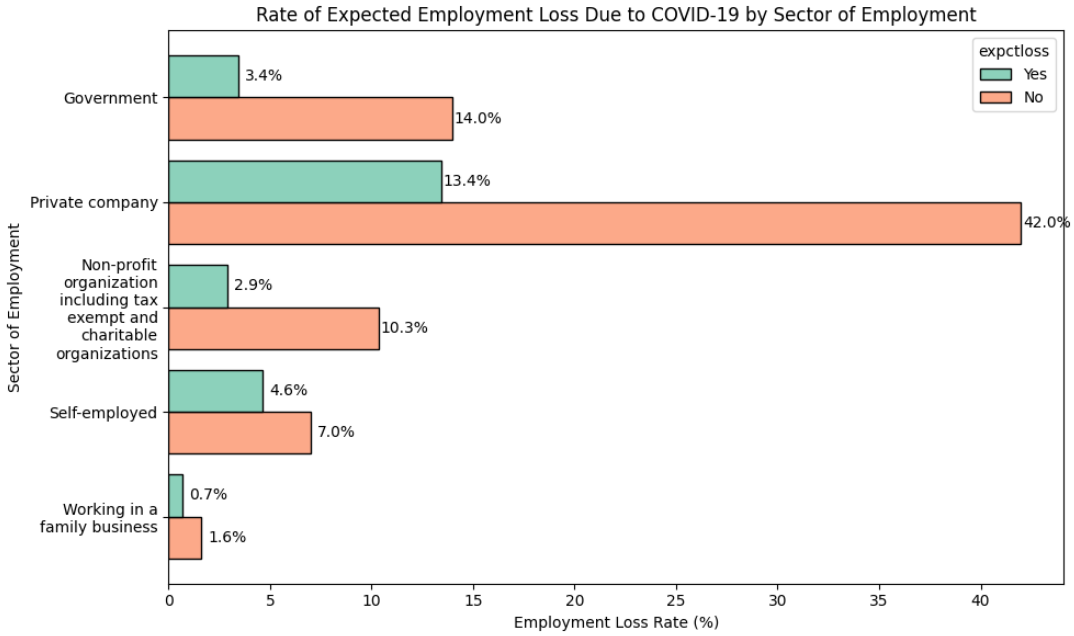
Steps:

- Data on employment sectors and expected job loss were analyzed.
- Bar plots were created to show the rate of expected job loss across different sectors.

Findings:

- Certain sectors, such as private companies and self-employed individuals, report higher expected job losses.
- Government jobs appear to be more secure during the pandemic.

5.



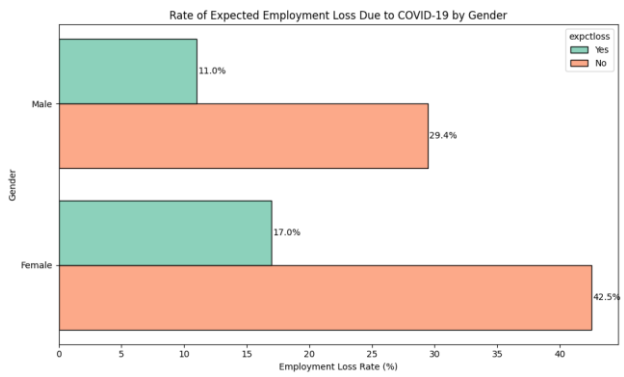
Employment Loss Relative to Demographics (Gender, Race, Age)

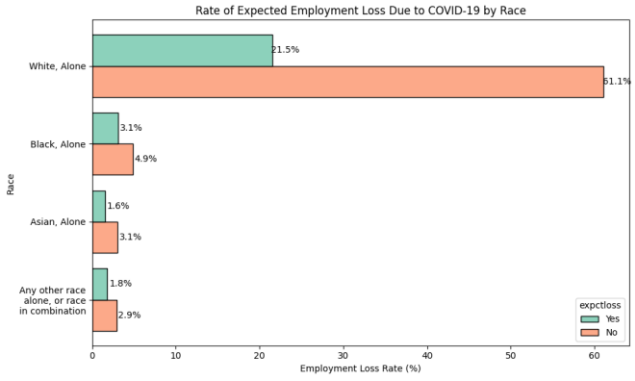
Steps:

- Data on expected job loss by gender, race, and birth year were analyzed.
- Bar plots were created to visualize the expected job loss across these demographics.

Findings:

- Expected job loss rates vary across gender and race, with some groups feeling more at risk.
- Younger age groups report higher expected job losses.





6. Delayed or Unobtained Medical Treatment by Household Income

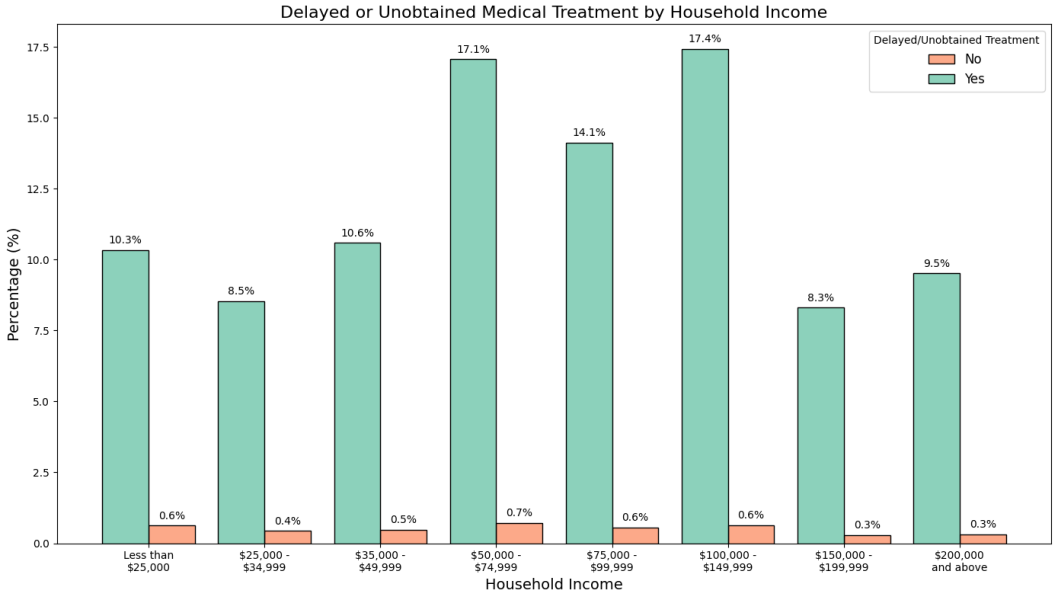
Steps:

- Data on income and delayed/unobtained medical treatment were analyzed.
- Bar plots and a chi-square test were conducted to assess the relationship.

Findings:

- Lower-income households report higher rates of delayed or unobtained medical treatment.

- There is a significant relationship between household income and access to medical treatment during the pandemic.



7.

Relationship between COVID-19 Symptom Manifestation and Age Group

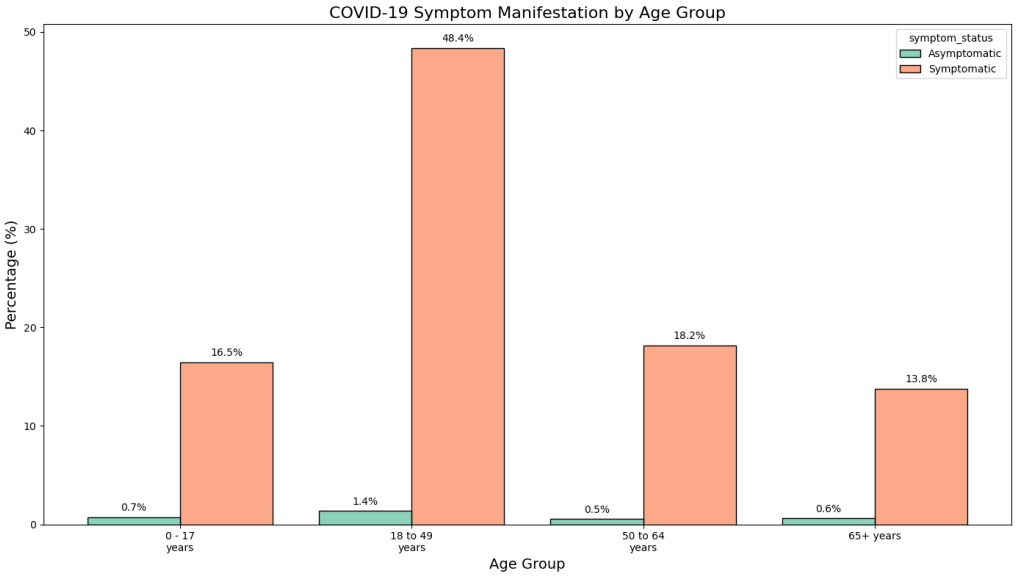
Steps:

- Data on age groups and symptom status were analyzed.
- Bar plots were created to show the percentage of symptom manifestation across different age groups.
- A chi-square test was performed to evaluate the statistical significance.

Findings:

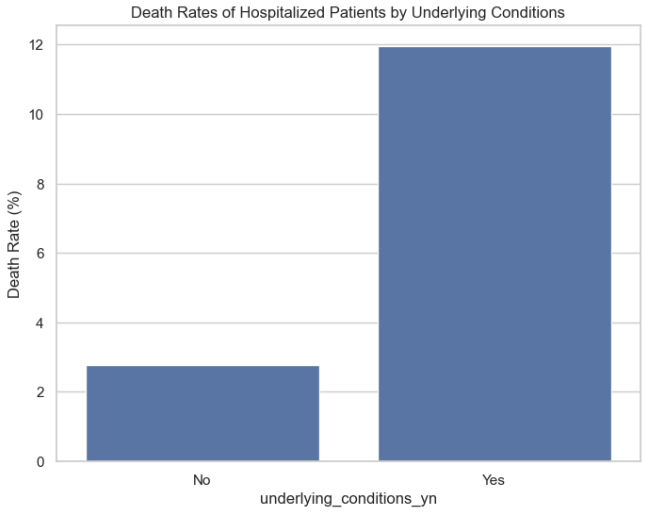
- There is a significant relationship between COVID-19 symptom manifestation and age group.
- Older age groups are more likely to manifest symptoms compared to younger age groups.

Part 2 Analysis

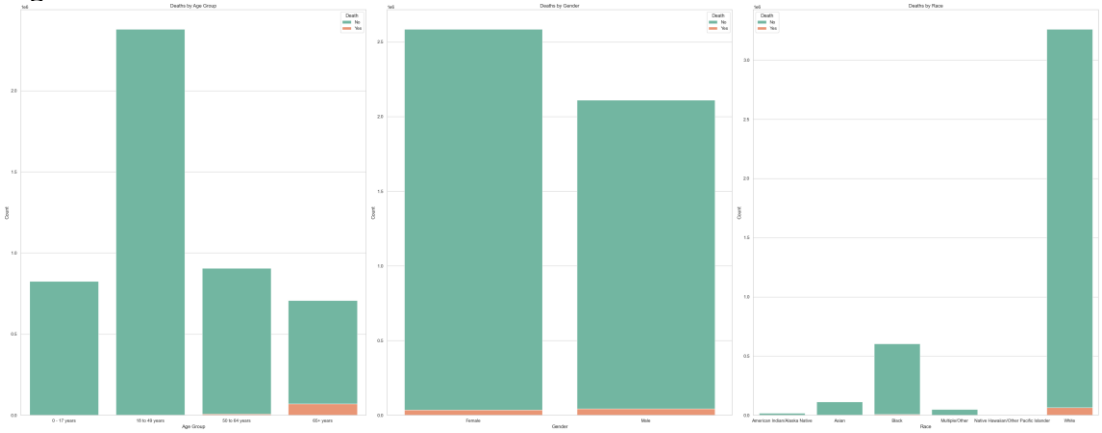


and Findings:

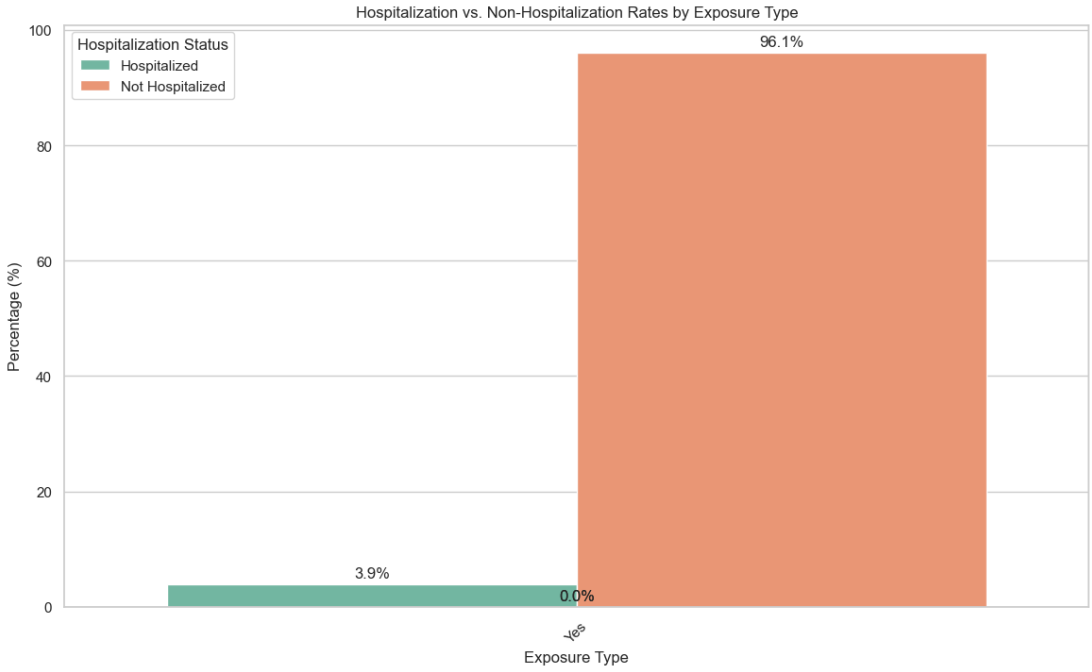
1. **Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19?**
    a. Steps:
        i. Data on underlying conditions, hospitalization status, and death status were filtered and cleaned.
        ii. Cross-tabulations and visualizations were created to compare death rates among hospitalized patients with and without underlying conditions.



    b. Findings: Patients with underlying conditions are more likely to die from COVID-19 compared to those without such conditions.
2. **Who are the people (demographic segment) most at risk of death due to COVID-19? Who is the least at risk?**
    a. Steps
        i. Data on age, gender, race, and death status were analyzed.
        ii. Bar plots were created to show death rates across different demographic segments.
        iii.

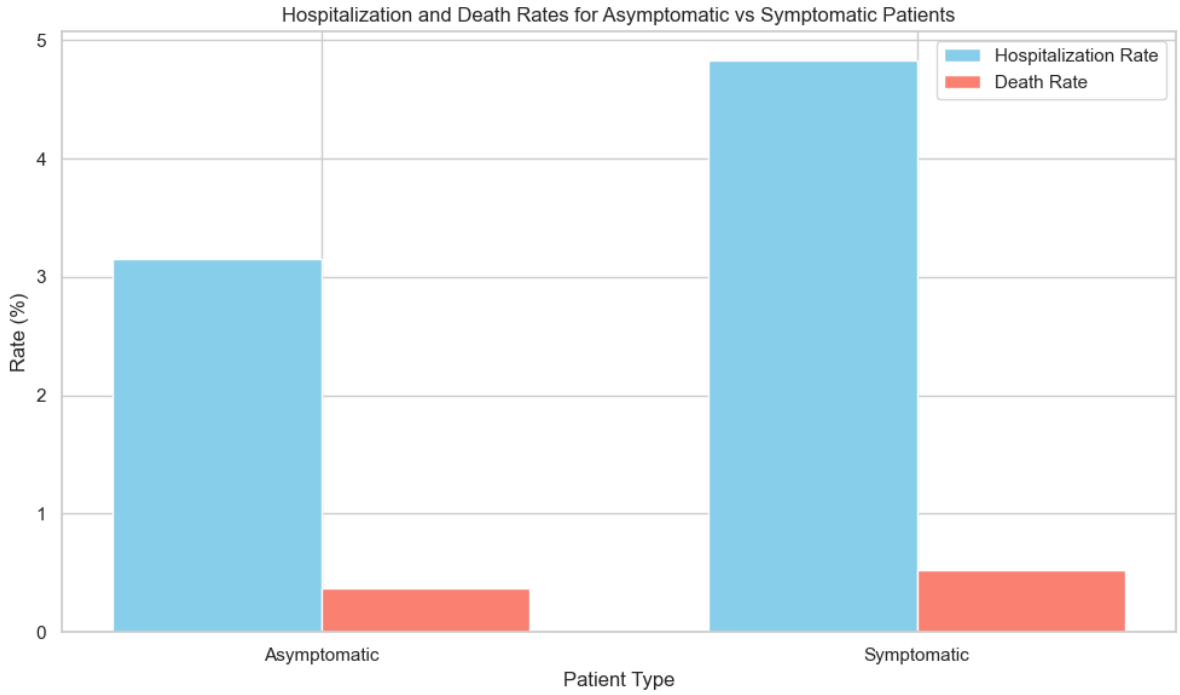


    b. Findings:
        i. Older individuals, males, and certain racial groups(White)exhibit higher death rates.

3. **What percent of patients who reported exposure to travel 14 days prior to illness onset end up hospitalized? What percent of those go on to be hospitalized?**
    a. **Steps:**
        i. Data on travel exposure, and hospitalization status were filtered and analyzed.
        ii. Percentages of hospitalization among those exposed to travel were calculated.
        iii. 

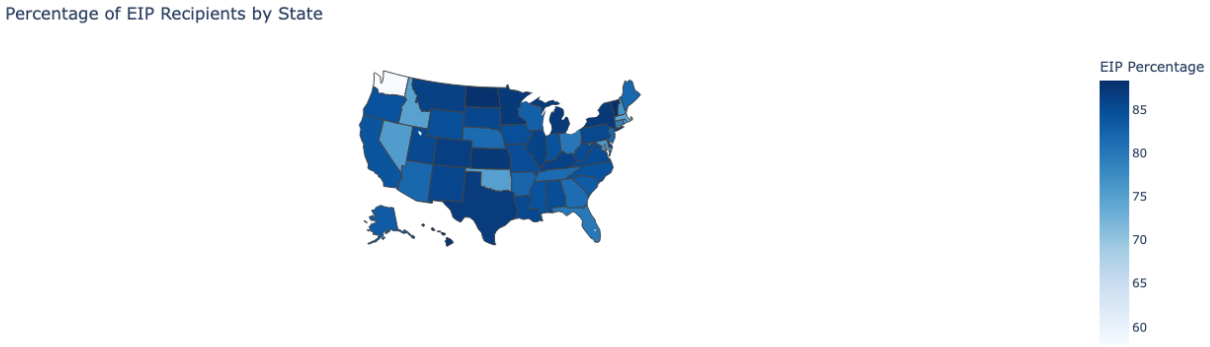
    b. Findings:
        i. Not many patients reporting exposure to travel within 14 days before illness onset end up hospitalized.

4. **Are asymptomatic COVID patients less likely to be hospitalized? Are they less likely to die from their illness?**



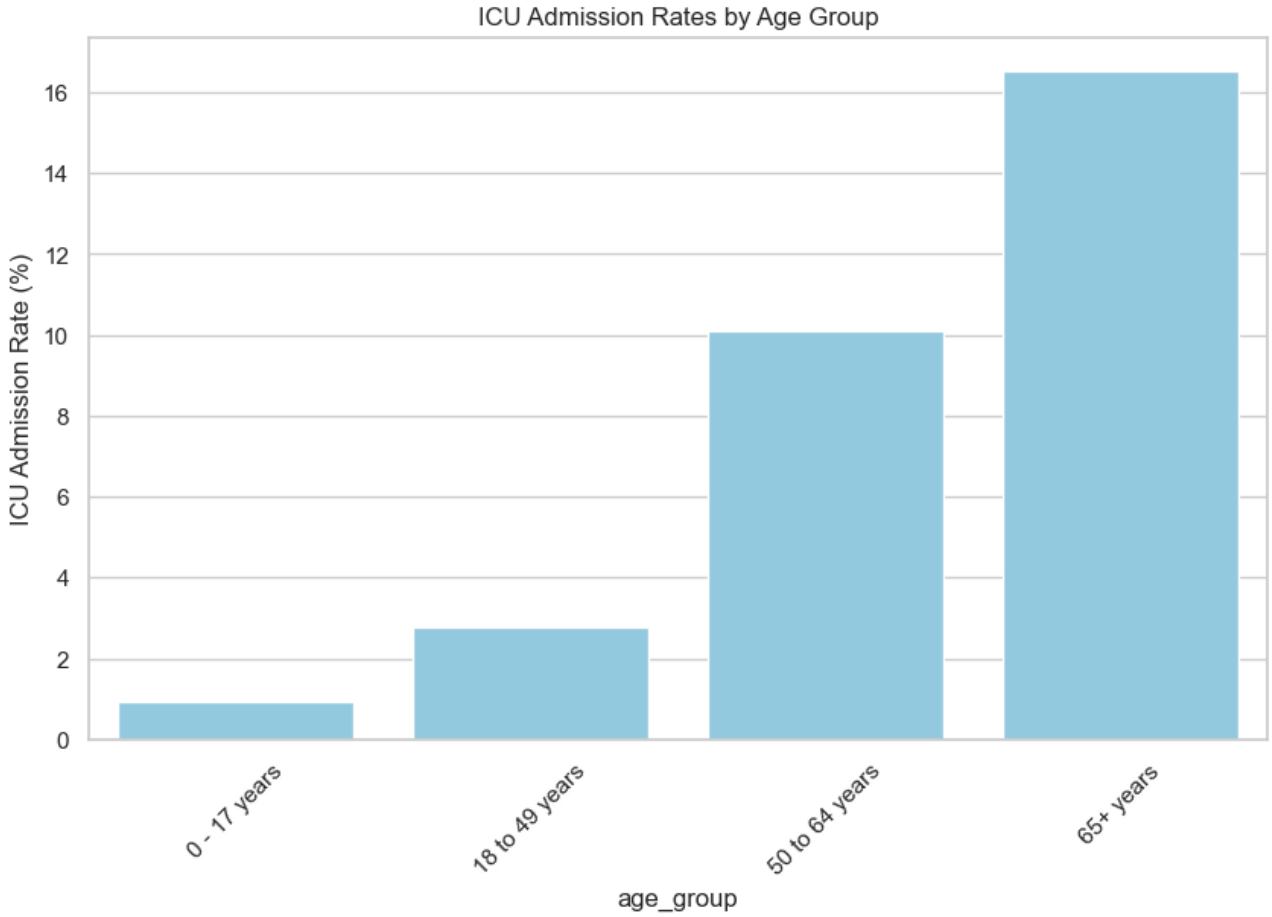
a. From analyzing the given data after cleaning, we can confirm that he hospitalization and death rates for asymptomatic patients are lower than those for symptomatic patients. This suggests that asymptomatic patients may have a milder form of the disease compared to symptomatic patients.

5. **Which state is associated with the highest percentage of Economic Impact (stimulus) payments among survey respondents?**



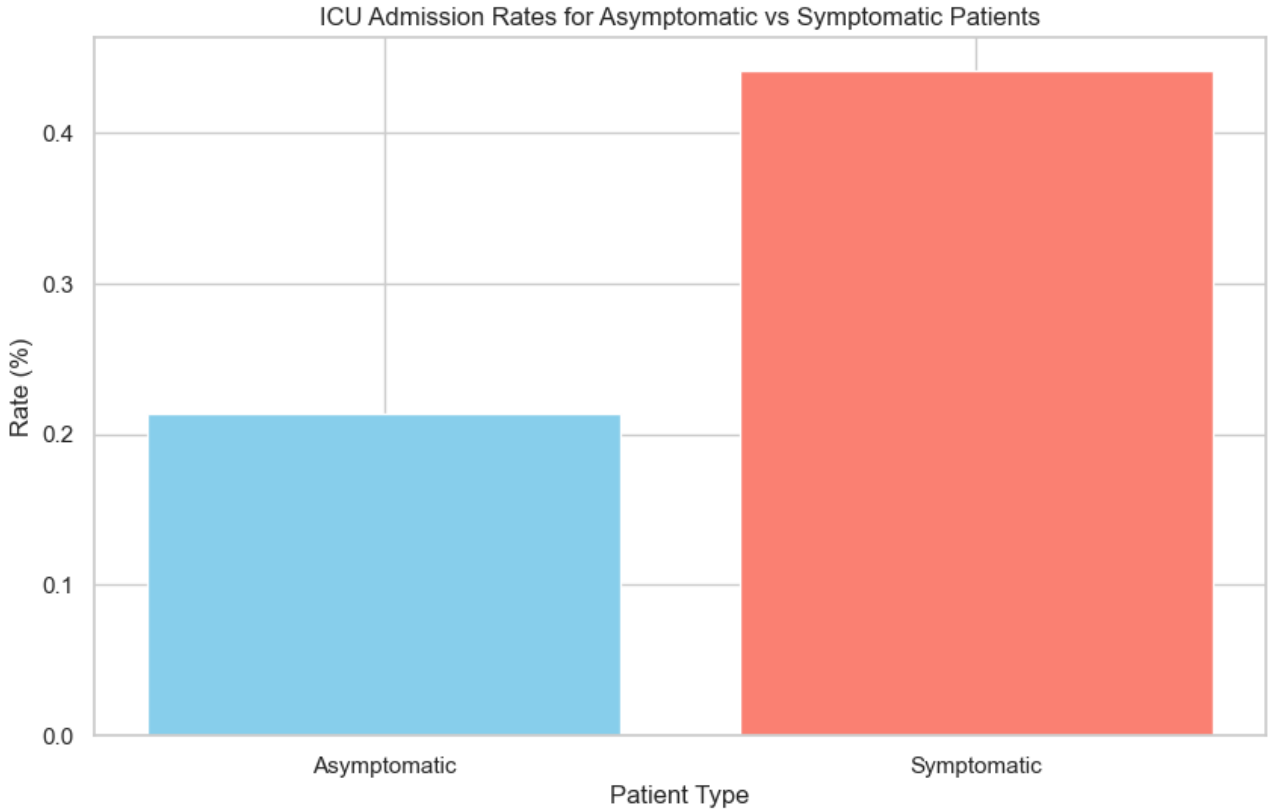
We calculated the percentages of EIP recipients in each state
Used Plotly library for heat-geographical representations
and found Virginia to be the state with the highest percentage.

6. **What is the relationship between age group and ICU admission rates?**



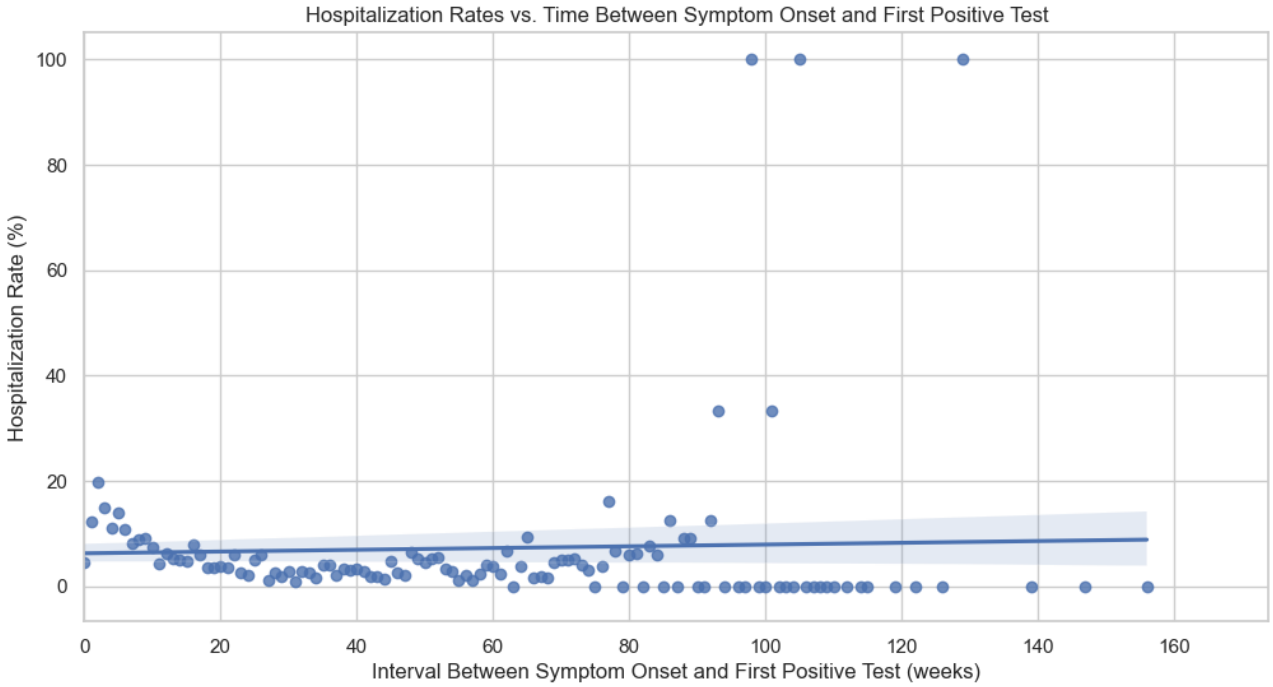
After Plotting age groups to icu admission rates distribution, we noted that as we go forward with an older age group, there is higher possibility they enter ICU.

7. **How does symptom status impact the rate of icu admmitance**



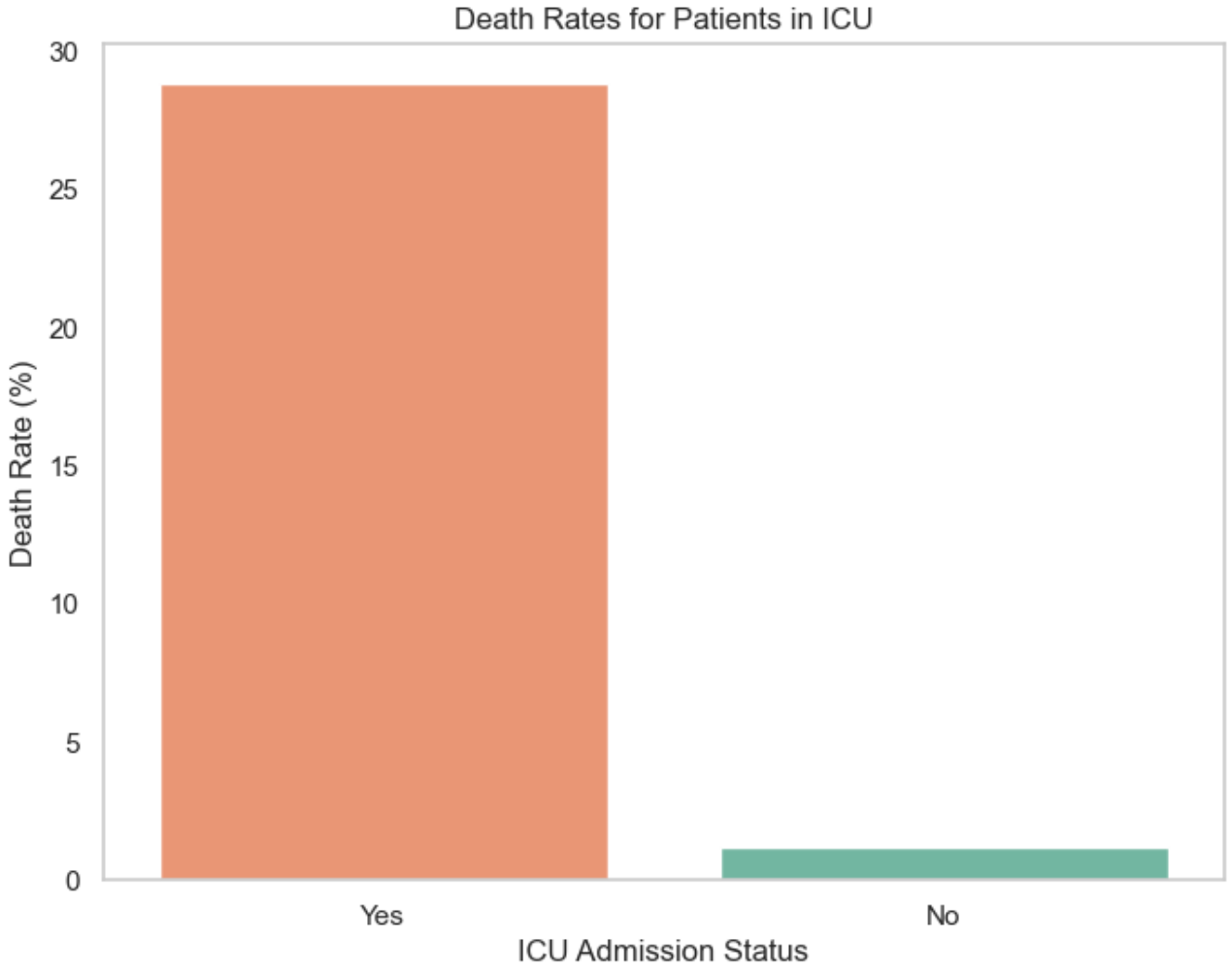
We can note that Symptomatic patients are twice as likely to be admitted to the ICU which means that they may suffer from extreme symptoms.

8. **How does the length of time between symptom onset and first positive test correlate with hospitalization rates?**

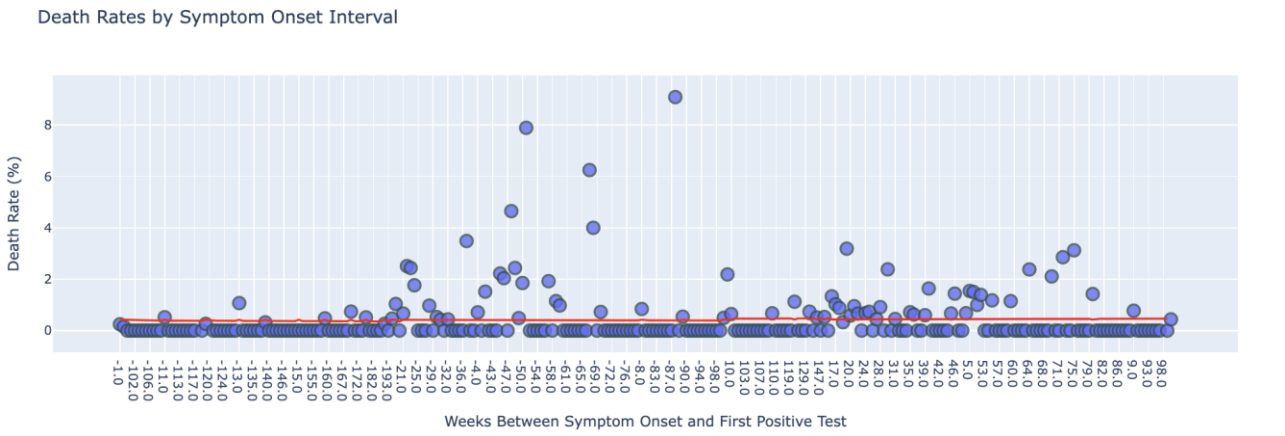


We could note a slight positive slope in the trendline, which may tell that the time between Symptoms and first positive test may affect how quick we can treat the symptoms

9. **What is the rate of patients who are dead after icu admmission**



10. **What is the impact of symptom onset interval on death rates?**



Here we calculated the death rate for groups of patients and the time of them being positive tested and having appearing symptoms. We could note that the trend line shows that the time between symptoms onset and the first positive test does not necessarily mean that this will affect the risk of death, although there a slight slope in the trendline

Limitation

The datasets used in this analysis contain missing, unknown, or inconsistent entries. Although efforts were made to clean the data, some inaccuracies may remain. Missing data, especially in crucial columns like age, underlying conditions, and hospitalization status, can lead to biased results. The amount of missing data is more than half of the dataset itself, we have seen rows from from 10M to <5M as we clean certain columns. Furthermore, there could be bias in the data due to the fact it is self reported.

Conclusion

Our exploratory data analysis reveals significant inconsistencies in the impact of COVID-19 across different demographics and socioeconomic groups. Our findings show the higher risk of ICU admissions for older individuals with underlying conditions, varying hospitalization and death rates across states, and significant employment and medical treatment challenges faced by lower-income households and certain racial groups. These insights can help us be more prepared and aware in other future challenges.

Part 3 – Hypothesis Testing:

3.1

Claim:

“There is a strong association between probability of death due to COVID-19 and patient demographics”

Test Type:

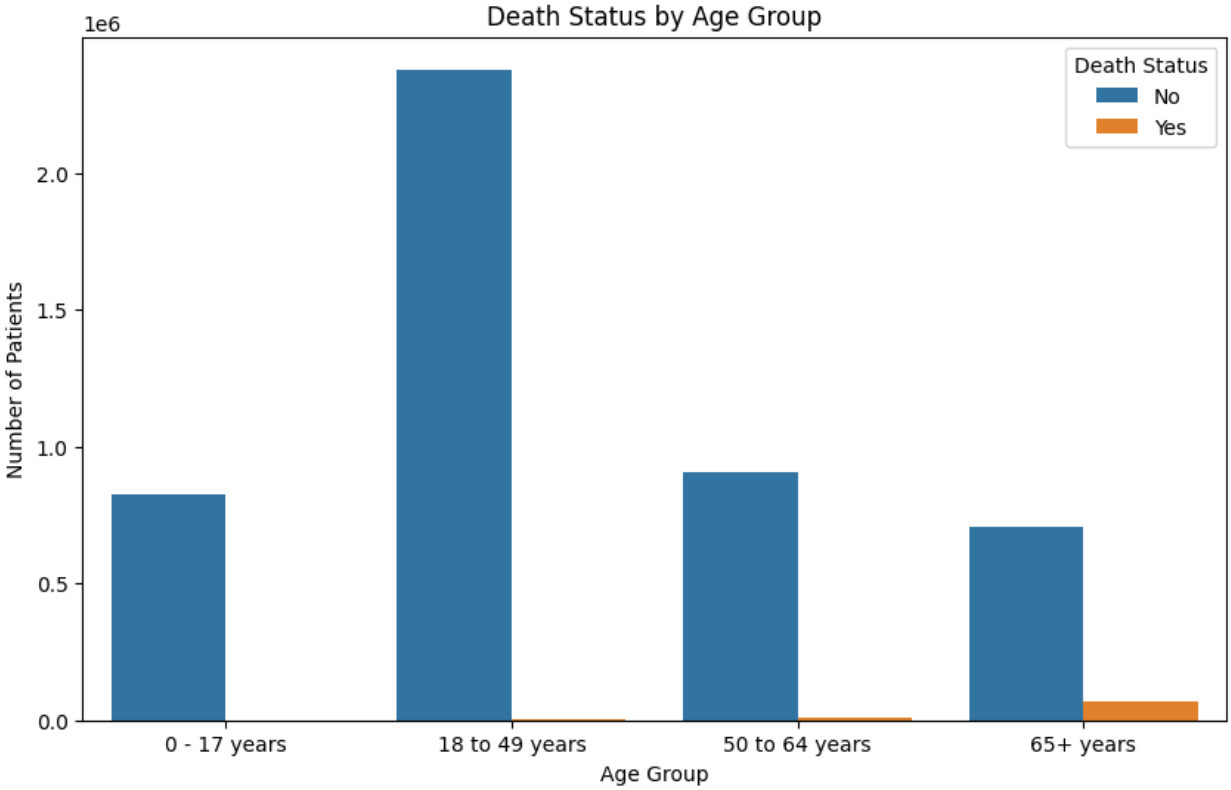
Since we are interested in investigating an association between categorical features, it is best to use the Chi-square test.

Age and Death from COVID-19:

Formulating the Hypothesis

Null Hypothesis Ho: there is no association between age and death rate for COVID-19 patients.

Alternative Hypothesis Ha: there is an association between age and death rate for COVID-19 patients.



Contingency Table

death_yn	No	Yes
age_group		
0 - 17 years	825619	0
18 to 49 years	2379165	1388
50 to 64 years	906248	7428
65+ years	707193	70562

Chi-Square Test Result

```
Chi2: 324646.1892193214
p-value: 0.0
Degrees of freedom: 3

Expected frequencies:
[[ 812237.76330482   13381.23669518]
 [2341970.13894858   38582.86105142]
 [ 898867.57769056   14808.42230944]
 [ 765149.52005604   12605.47994396]]
```

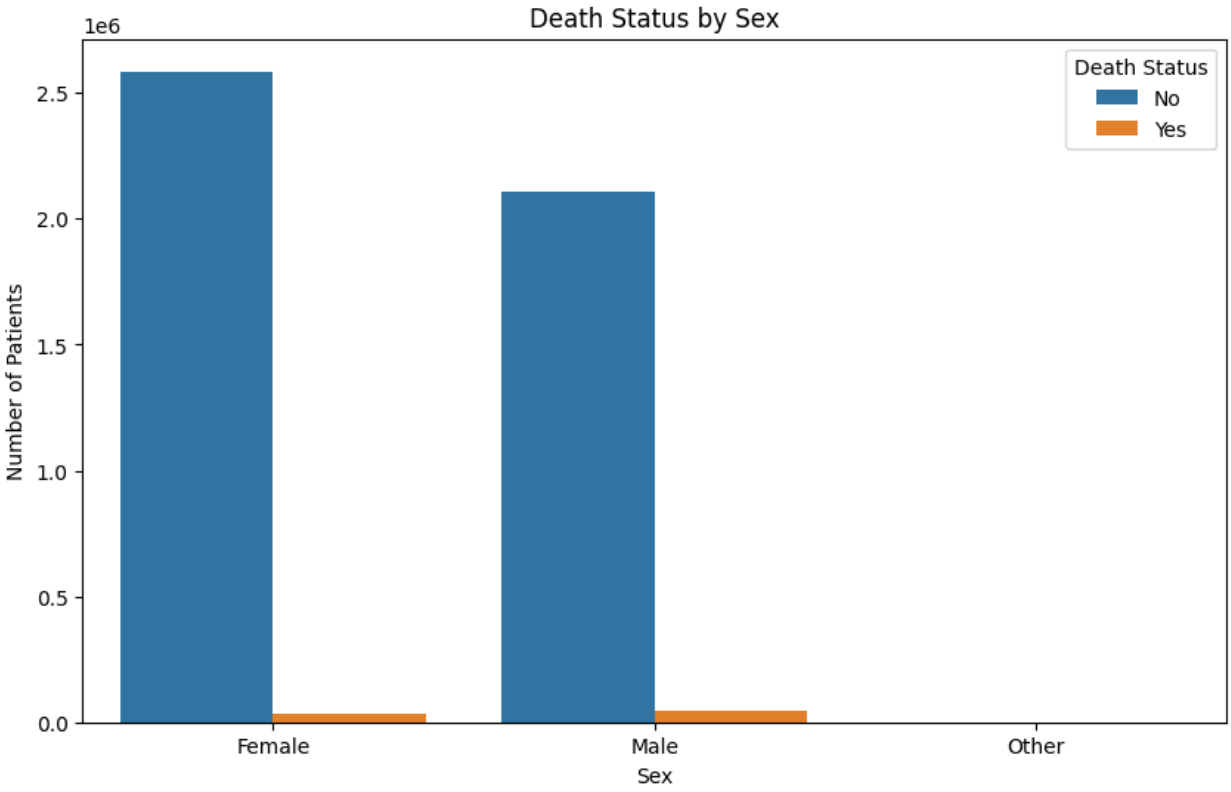
From the test result, it can be said there is enough evidence to say reject the null hypothesis, meaning that there is a significant association between age and death probability from COVID-19.

Sex and Death from COVID-19

Formulating the Hypothesis

Null Hypothesis Ho: there is no association between sex and death rate for COVID-19 patients.

Alternative Hypothesis Ha: there is an association between sex and death rate for COVID-19 patients.



Contingency Table

death_yn	No	Yes
sex		
Female	2585496	35226
Male	2109900	43912
Other	14	0

Chi-Square Test Result

```
Chi2: 3500.18092650455
p-value: 0.0
Degrees of freedom: 2

Expected frequencies:
[[2.57728361e+06 4.34383941e+04]
 [2.11811263e+06 3.56993739e+04]
 [1.37679504e+01 2.32049610e-01]]
```

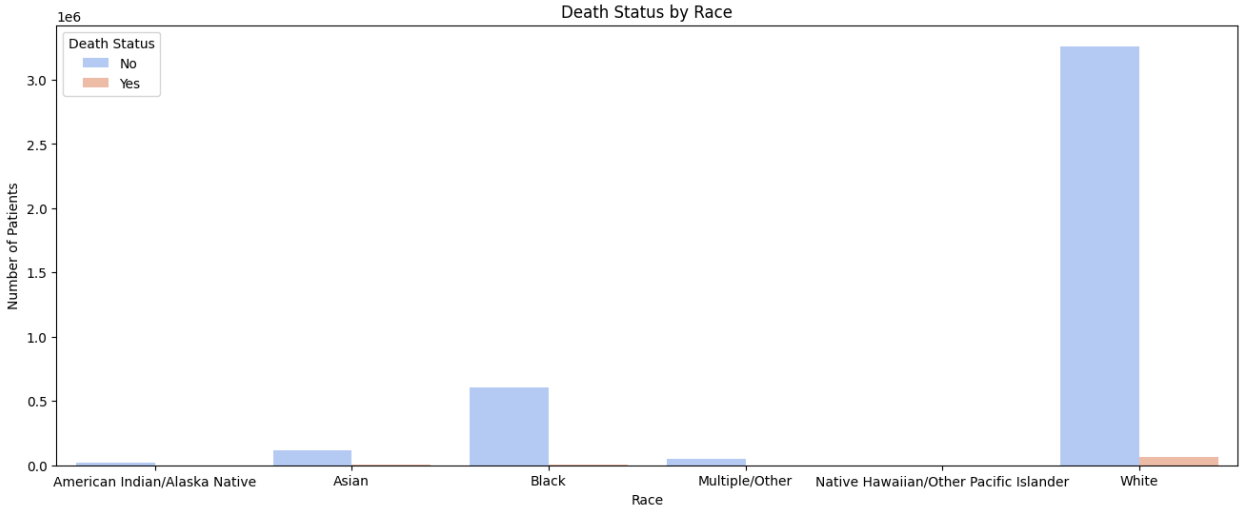
From the p-value in the test result, it can be said there is enough evidence to say reject the null hypothesis, meaning that there is a significant association between sex and death probability from COVID-19.

Race and Death from COVID-19

Formulating the Hypothesis

Null Hypothesis Ho: there is no association between race and death rate for COVID-19 patients.

Alternative Hypothesis Ha: there is an association between race and death rate for COVID-19 patients.



Contingency Table

death_yn	No	Yes
race		
American Indian/Alaska Native	17912	118
Asian	113337	2386
Black	605111	8583
Multiple/Other	50185	108
Native Hawaiian/Other Pacific Islander	915	0
White	3260997	64289

Chi-Square Test Result

```
Chi2: 1753.859070376362
p-value: 0.0
Degrees of freedom: 5

Expected frequencies:
[[1.76999816e+04 3.30018427e+02]
 [1.13604823e+05 2.11817650e+03]
 [6.02461037e+05 1.12329633e+04]
 [4.93724444e+04 9.20555559e+02]
 [8.98251977e+02 1.67480233e+01]
 [3.26442046e+06 6.08655382e+04]]
```

From the p-value in the test result, it can be said there is enough evidence to say reject the null hypothesis, meaning that there is a significant association between race and death probability from COVID-19.

3.2

My claim:

There is an association between patients having asymptomatic or symptomatic symptoms and the probability of death from COVID-19.

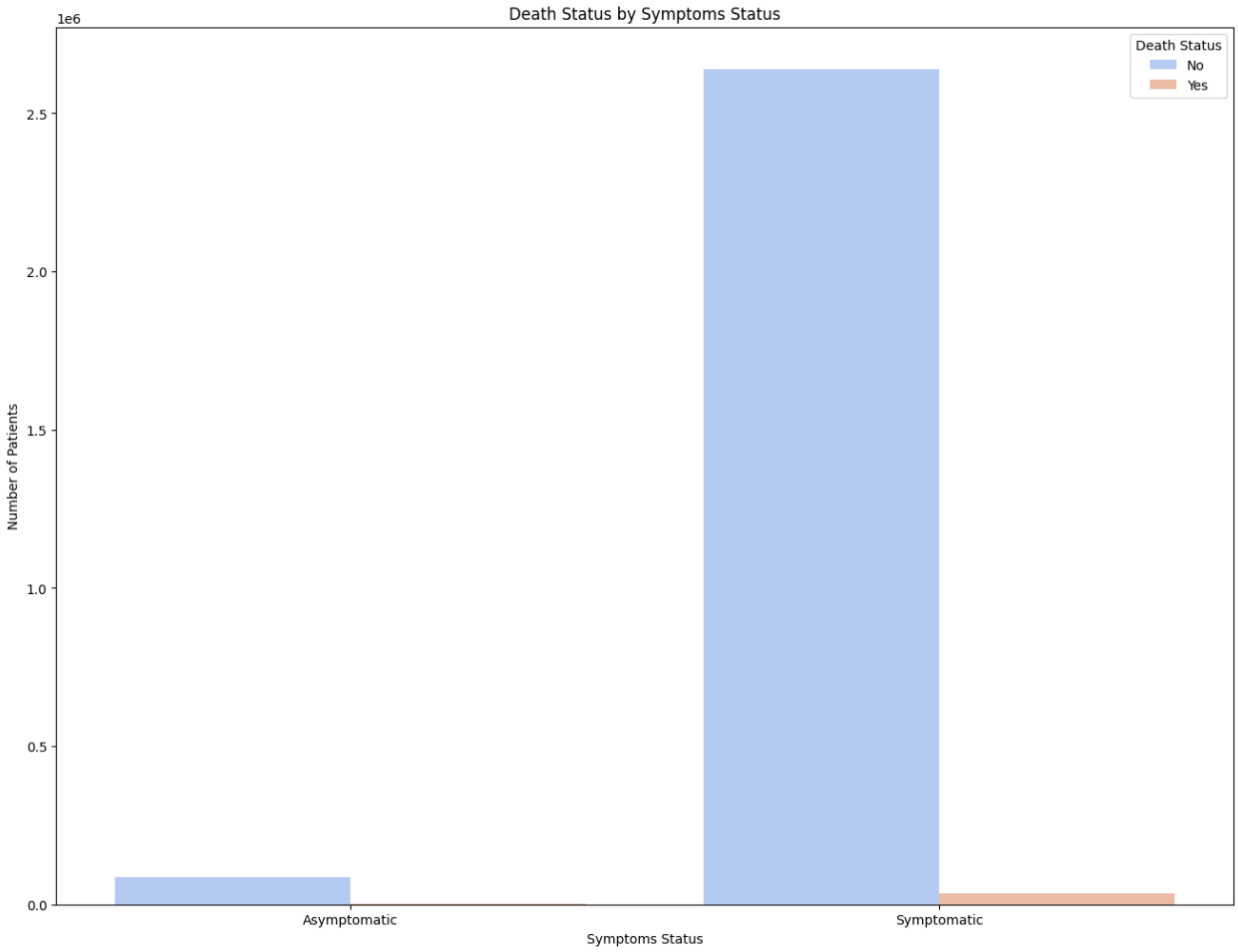
Test Type:

Since we are interested in comparing the proportion of deaths in two populations, it is best to use the Z-test to investigate the significance of the association.

Formulating the Hypothesis

Null Hypothesis Ho: the death probability from COVID-19 patients is equal for patients with asymptotic and asymptotic symptom status.

Alternative Hypothesis Ha: the death probability from COVID-19 patients is not the same for patients with asymptotic and asymptotic symptom status.



Contingency Table

death_yn	No	Yes
symptom_status		
Asymptomatic	85306	834
Symptomatic	2640259	34238

Z-Test Result

```
Z-statistic: -8.047090626655558
P-value: 8.478539931811714e-16
```

Based on the resulting p-value, it can be said that there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis, that is, the death probability from COVID-19 is not the same for patients with symptomatic and asymptomatic symptom status.

Part 4 Regression Analysis:

The COVID-19 pandemic has had a profound impact on public health, prompting the need for comprehensive analysis to understand the factors influencing mortality rates. This business report presents a regression analysis aimed at predicting the total percentage of deaths among COVID-19 cases in a given month. Utilizing the COVID Case Surveillance dataset, the analysis focuses on gender and age distribution, ICU admissions, and hospitalizations as key predictors. The goal is to identify significant variables that contribute to the variability in death rates, thereby informing public health policies and interventions.

Analysis Overview

Data Preparation

The dataset used in this analysis includes the following columns:

- **case_month**
- **age_group**
- **sex**
- **hosp_yn**
- **icu_yn**
- **death_yn**

The data was filtered to remove entries with missing or unknown values in the relevant columns, and then grouped by **case_month** to calculate the proportions of gender, age groups, ICU admissions, hospitalizations, and deaths.

Regression Model

The regression model was designed to predict the **death_proportion** based on the following predictors:

- **female_proportion**
- **age_0_17_proportion**
- **age_18_49_proportion**
- **age_50_64_proportion**
- **age_65_proportion**
- **icu_proportion**
- **hospitalized_proportion**

The model's coefficients and p-values were calculated to identify significant predictors.

Model Summary

The regression results are summarized below:

Predictor	**Coefficient**	**p-value**

Intercept	0.2141	0.002
female_proportion	-0.3489	0.015
age_0_17_proportion	-0.0413	0.184
age_18_49_proportion	-0.1005	0.035
age_50_64_proportion	0.4010	0.000
age_65_proportion	-0.0451	0.420
icu_proportion	0.5399	0.008
hospitalized_proportion	-0.1719	0.071

The model has an R-squared value of 0.692, indicating that it explains approximately 69.2% of the variability in the death proportion.

Significant Predictors

Based on the p-values, the following predictors are considered significant (p-value < 0.05):

- **female_proportion**
- **age_18_49_proportion**
- **age_50_64_proportion**
- **icu_proportion**

Predictors with higher p-values, such as **age_0_17_proportion**, **age_65_proportion**, and **hospitalized_proportion**, are not considered significant.

Multicollinearity Assessment

A pairplot and a heatmap were used to visualize relationships between predictors, revealing potential multicollinearity issues. Variance Inflation Factor (VIF) calculations indicated high multicollinearity among age group proportions.

Model Improvement Attempts

Various strategies were employed to improve the model:

- Removing the intercept: This did not significantly impact the R-squared value.
- Introducing higher-order terms and interaction terms: These attempts resulted in a marginal increase in the R-squared value to 0.882.
- Removing outliers: Identified outliers were removed to refine the model further.

Conclusion

The regression analysis identified several significant predictors of COVID-19 death proportion, including gender distribution, age distribution, and ICU admissions. However, multicollinearity among age group proportions poses a challenge. The model's R-squared value of 0.692 suggests a moderate level of explanatory power, indicating that there are other factors influencing death rates that were not captured in this analysis. Additionally, the high multicollinearity among age groups suggests that these variables may not independently contribute to the prediction, complicating the interpretation of their individual effects.

Significance:

- **Policy Implications**: The findings can help tailor public health interventions to specific demographic groups and clinical conditions, potentially reducing COVID-19 mortality rates.
- **Resource Allocation**: Understanding significant predictors can guide resource allocation, such as ICU beds and hospital staff, to areas with higher risk profiles.

Limitations:

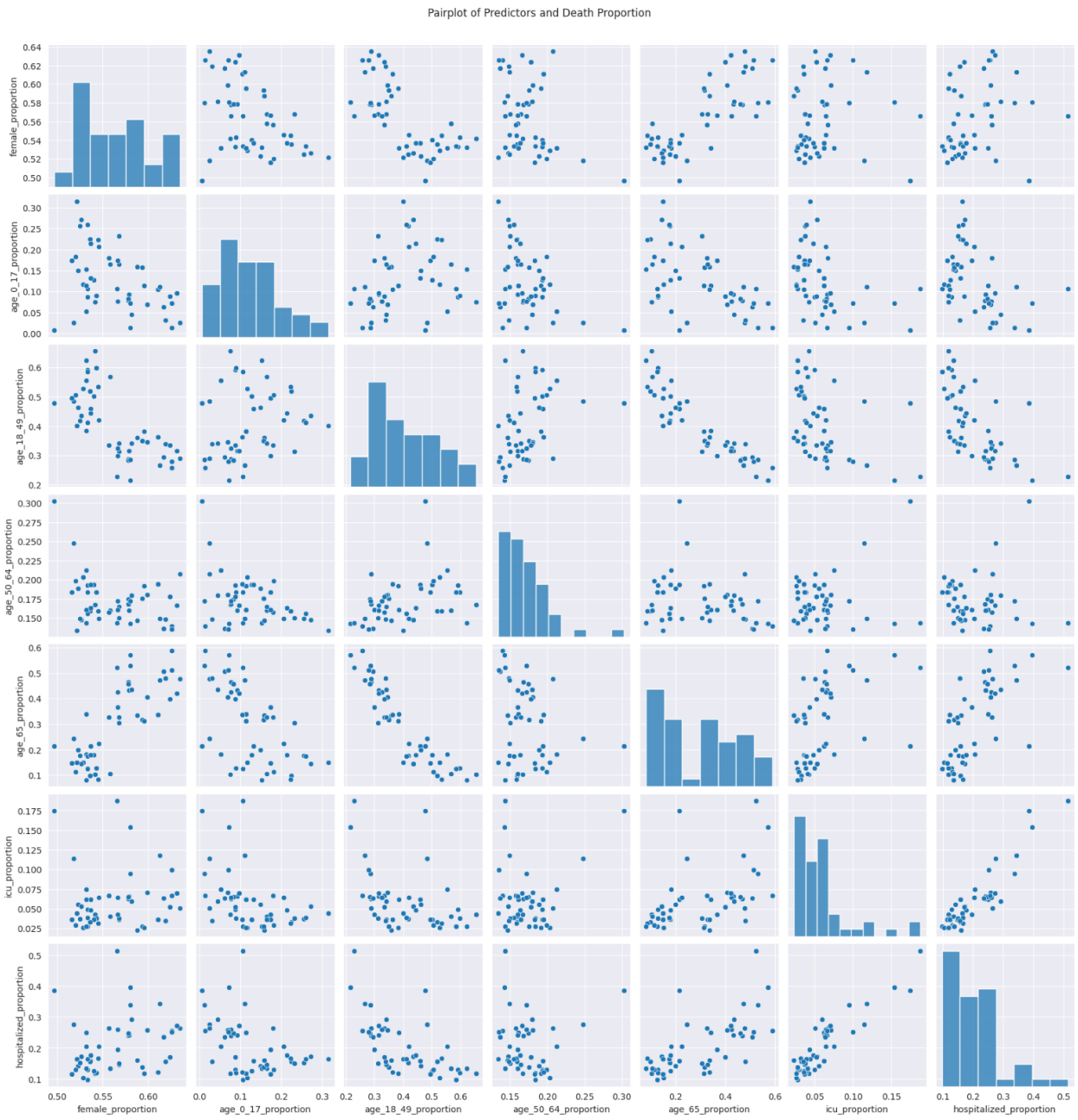
- **Data Quality**: The analysis is limited by the quality and completeness of the available data. Missing or misclassified data could bias the results.
- **Multicollinearity**: High multicollinearity among age groups complicates the interpretation of individual predictor effects.
- **Unobserved Variables**: Other important factors, such as comorbidities and vaccination status, were not included in the model, limiting its comprehensiveness.

Future research should aim to incorporate a broader range of predictors and employ advanced techniques to address multicollinearity, thereby enhancing the robustness and interpretability of the model.

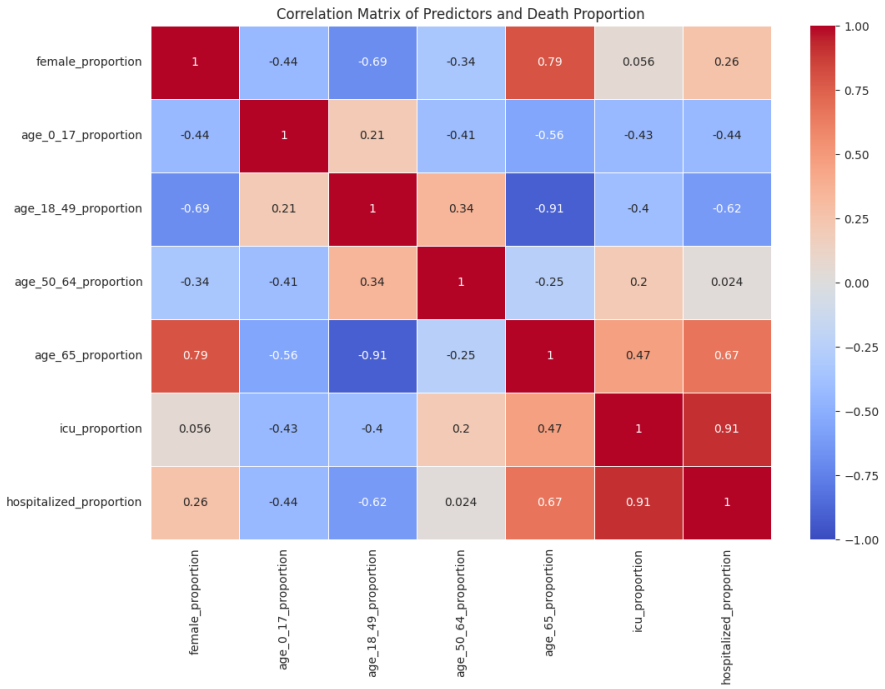
Visualizations

Below are some visualizations that support the findings of this analysis:

Pairplot of Predictors:



Heatmap of Predictor Correlations:



These visualizations help illustrate the relationships and potential multicollinearity among the predictors.

Recommendations

1. **Model Refinement**: Address multicollinearity by considering techniques like principal component analysis (PCA) to reduce dimensionality.
2. **Data Enhancement**: Incorporate additional relevant predictors, such as comorbidities and vaccination rates, to improve model accuracy.
3. **Policy Implications**: Utilize significant predictors to inform public health strategies, such as targeted interventions for specific age groups or genders.

Appendix: