COVID-19 Demographic Analysis Report

Ahmed Abdelgawad 201901249
Adham Ahmed 202100163
Ziad Behairy 202100154
Mohamed Morshedy 202100372

Supervision

Dr.Mahmoud Abdelaziz
Eng.Ahmed Abdelsalam
Eng.Asmaa Mostafa
Eng.Ahmed Ali

Communications and Information Engineering
Zewail City of Science and Technology

Contents

1 Introduction

The COVID-19 pandemic has had a profound impact globally, and understanding the factors associated with severe outcomes is critical for public health planning and response. This report analyzes the association between demographic features (age, sex, race) and the probability of death, hospitalization, and ICU admission due to COVID-19. The objective is to provide insights into which demographic groups are at higher risk and how these factors interact.

2 Part 1: Exploratory Analysis

2.1 Hospitalizations vs. Deaths per Month

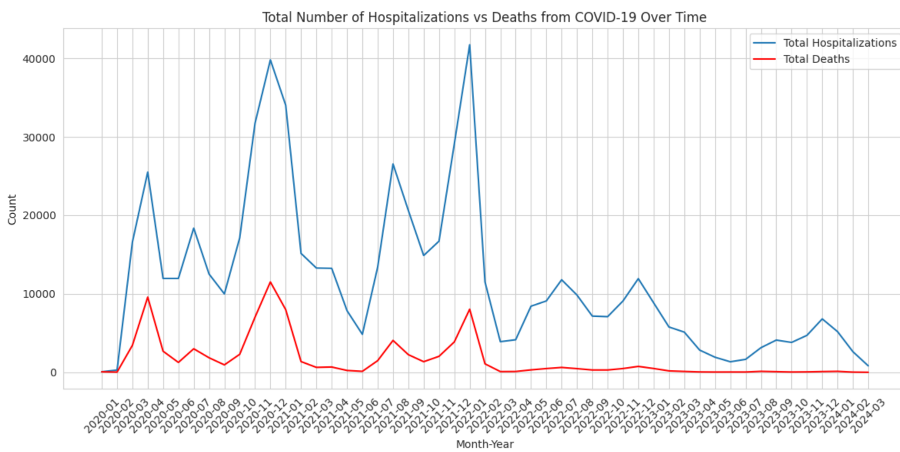

Figure 1: Total Hospitalizations vs. Deaths from COVID-19 per Month

Commentary: In the months of November and December, there are noticeable peaks. We observe a correlation between the number of people admitted to the hospital and the number of fatalities. These peaks coincide with the periods when curfews were imposed and lifted. Additionally, it is evident that the highest incidence of COVID-19 occurred in the years 2020 and 2021.

2.2 Average Rates of COVID-related Deaths Relative to Demographics

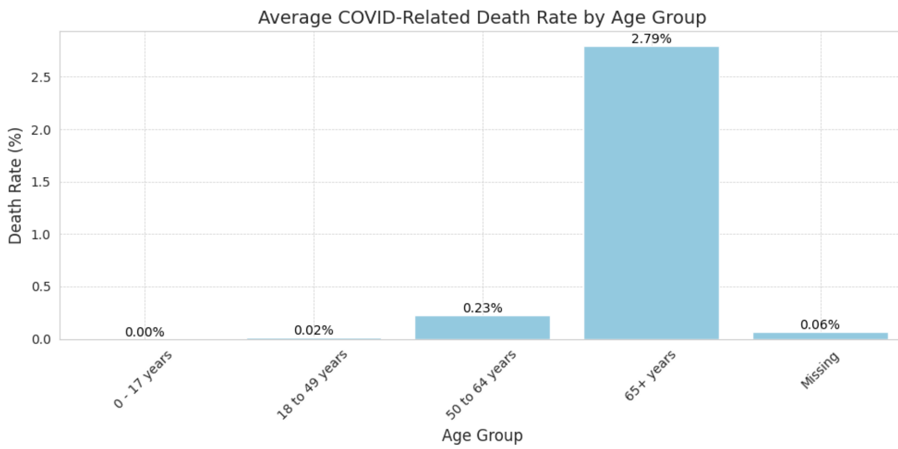

Figure 2: Average Rates of COVID-related Deaths Relative to Demographics

Commentary: In this question, we will create three graphs to examine the relationship between mortality rates and various parameters, specifically demographics.

In the first graph, which focuses on age, we find that the highest mortality rates occur among individuals over 65 years old. This finding is consistent and expected, aligning logically with our analysis.

2.3 Average Rates of COVID-related Deaths Relative to Demographics

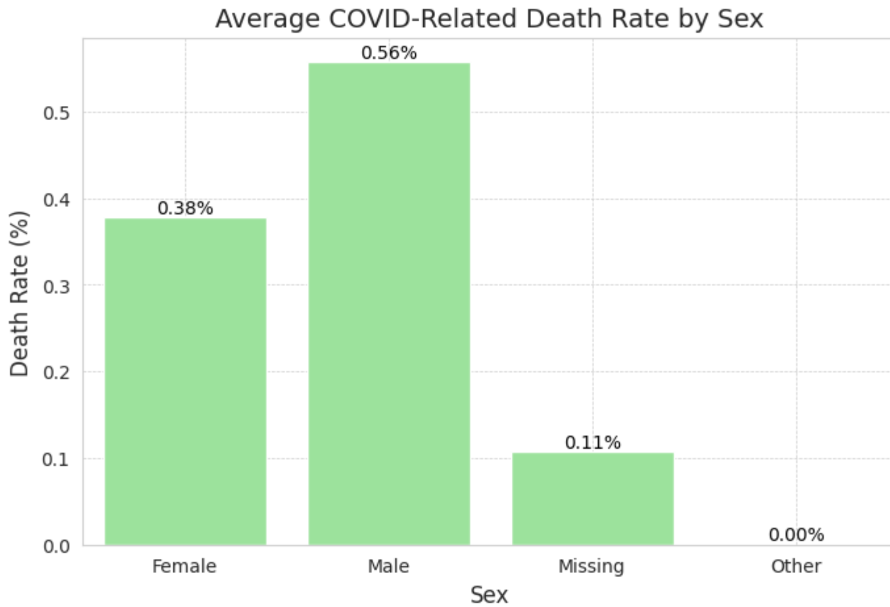

Figure 3: Average COVID-Related Death Rate by Sex

In the second graph, which examines gender and mortality rates, we observe that the mortality rate for men is significantly higher than for women.

2.4 Average Rates of COVID-related Deaths Relative to Demographics

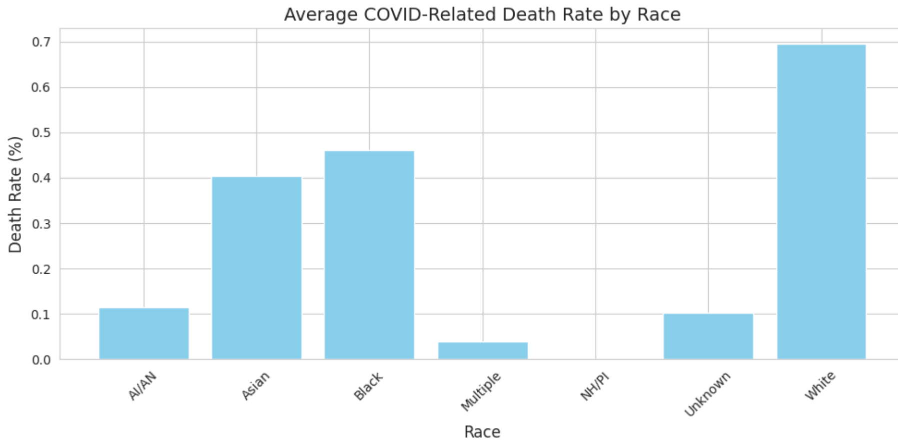

Figure 4: Average COVID-Related Death Rate by Race

In the third graph, which analyzes race and mortality rates, we find that the highest mortality rates are among White individuals, followed by Black individuals, and then Asians.

2.5 Rates of Hospitalization and Death with Age

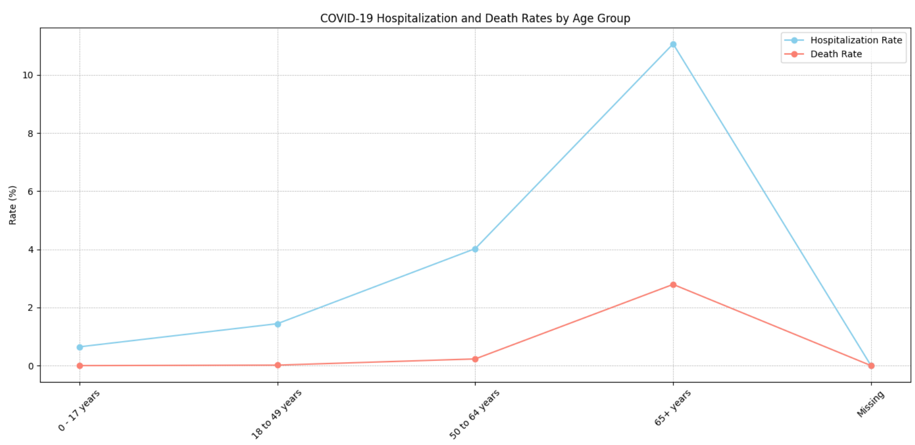

Figure 5: Rates of COVID-related Hospitalization and Death across Age Groups

Commentary: he third question aims to investigate the relationship between hospitalization and death rates by age group. We find that the highest hospitalization rates are markedly among the elderly. Similarly, the highest mortality rates are also within this age group. This finding is consistent with our previous analysis, which also concluded that the highest mortality rates are among the elderly. .

2.6 Hospitalization and Death Rates per State

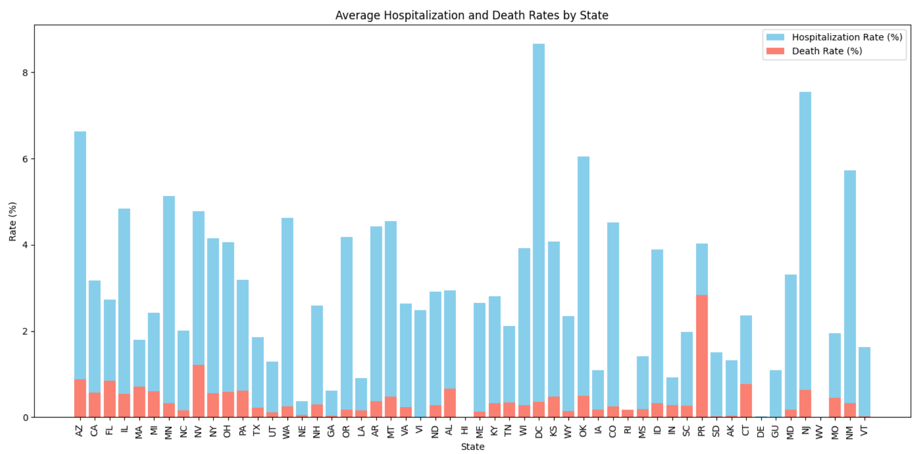

Figure 6: Average Rates of Hospitalization and Death per State

We observe that the most populous cities exhibited the highest hospitalization rates. Additionally, Puerto Rico had the highest death rates among all states. This might be attributed to a significant portion of its population living below the poverty line. We also observe generally high death rates in states characterized by poverty and disadvantage.

2.7 ICU Admittance, Age, and Pre-existing Conditions

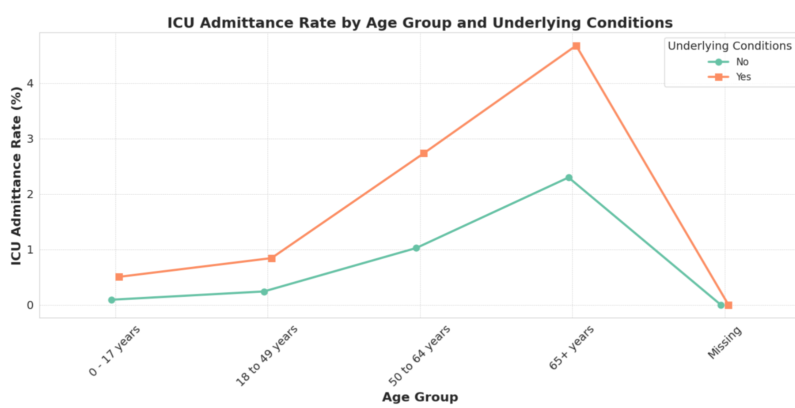

Figure 7: ICU Admittance Rate by Age and Pre-existing Conditions

Commentary: Twe investigated the relationship between ICU admission rates and the presence of COVID-19 symptoms across different age levels. We found that the elderly were the most frequently admitted to the ICU, regardless of whether they had symptoms. Additionally, the likelihood of an elderly person being admitted to the ICU with symptoms was significantly higher than their admission rate without symptoms.

2.8 Rate of Expected Employment Loss Due to COVID-19 and Sector of Employment

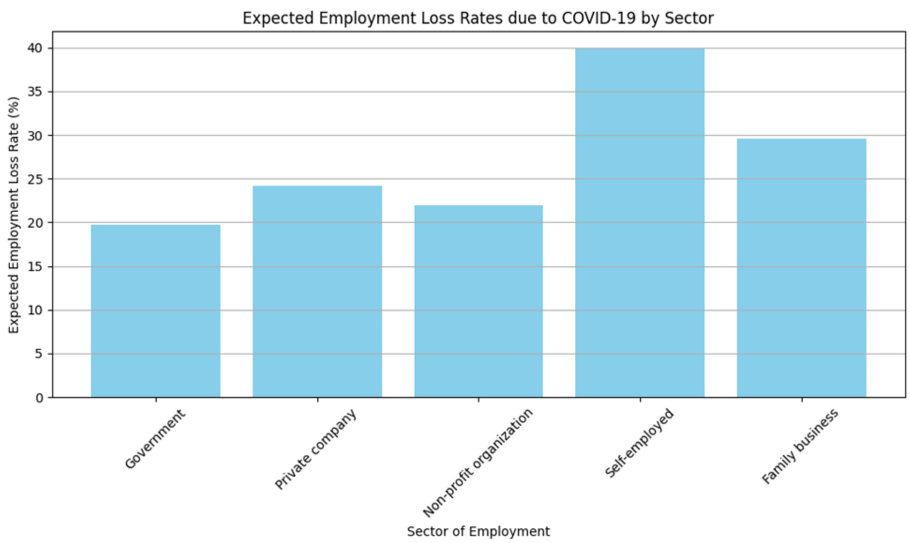

Figure 8: Rate of Expected Employment Loss Due to COVID-19 by Sector

Commentary: This plot examines the expected employment loss due to COVID-19 across different sectors of employment. we are investigating the percentage of expected employment loss through each of the employment sectors. We can see the following:

• Highest expected employment loss: Self-employment sector was expected to lose about 40 percent of the employees, which makes sense as self-employment in that time was not reliable due to the non-sustainable salary.

• Lowest expected employment loss: Government employment was expected to lose less than 20 percent of the employees in this sector, which makes sense as it is considered the most sustainable salaries out of all the sectors.

2.9 Rate of Expected Employment Loss Due to COVID-19 Relative to Responders' Demographics

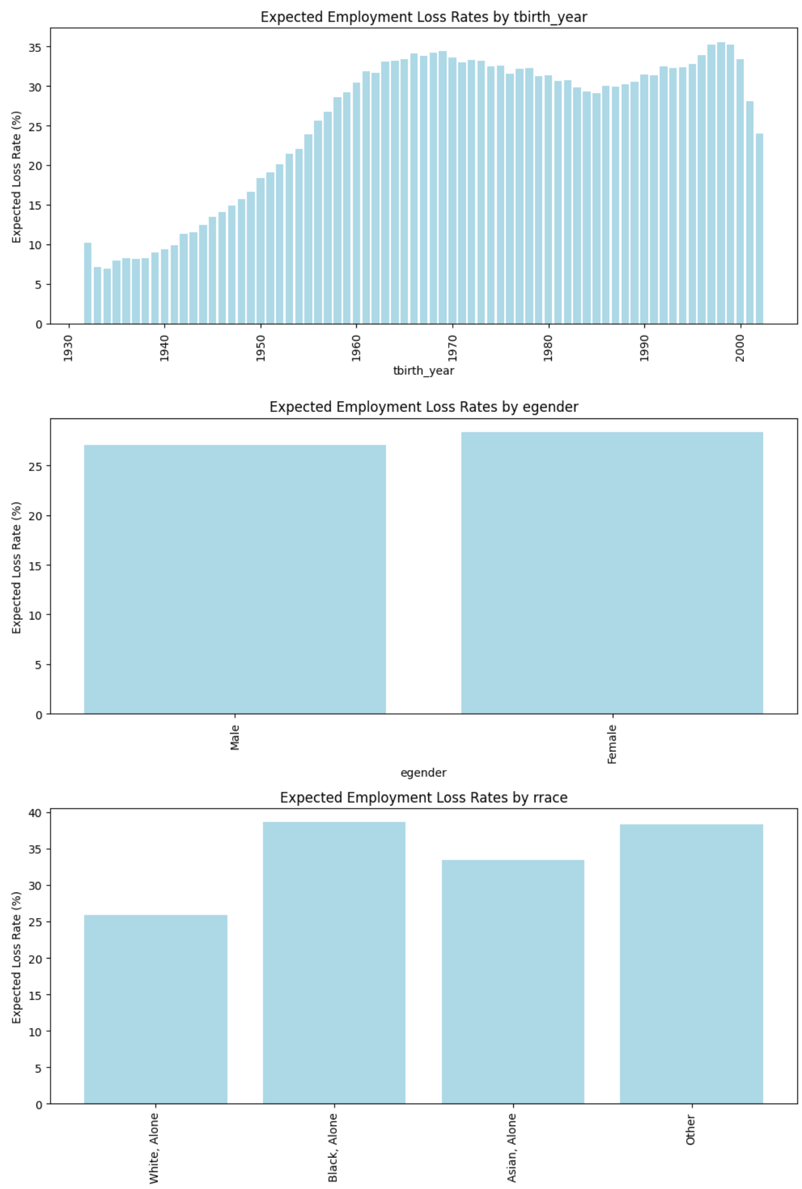

Figure 9: Rate of Expected Employment Loss Due to COVID-19 by Demographics

Commentary: we investigated the rate of expected employment loss due to COVID-19 relative to responders' demographics.

In the first graph, individuals born between 1960 and 2000 experienced the highest job loss, which is logical as this age range represents the average working force of the country.

In the second graph, both females and males were equally likely to lose their jobs, with females slightly more at risk.

In the third graph, people of color, particularly Black individuals, were the most vulnerable to job loss, followed by Asians and then White individuals. This discrepancy may be influenced by political factors, resulting in a higher likelihood of job loss among Black and Asian individuals compared to White individuals.

2.10 Rate of Expected Employment Loss Due to COVID-19 for Top 10 States with Highest Hospitalization Rates

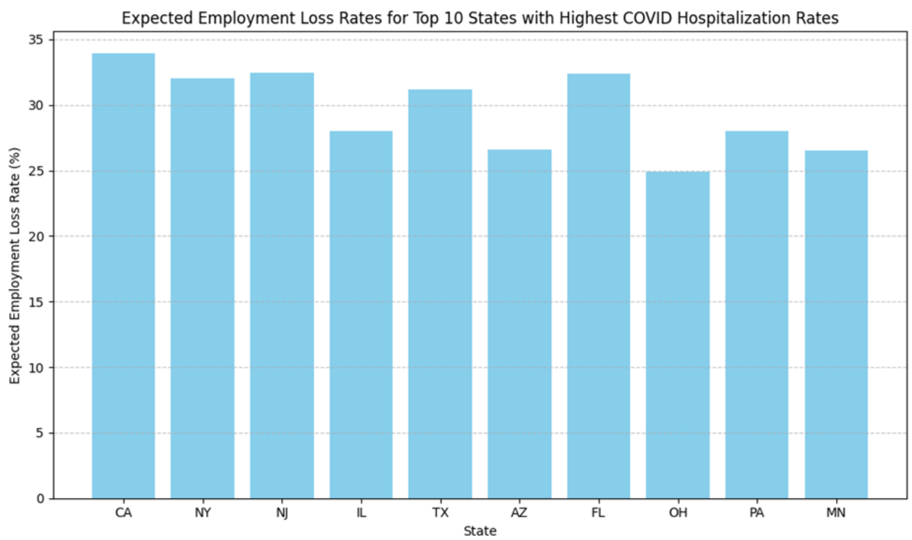

Figure 10: Employment Loss in Top 10 States with Highest Hospitalization Rates

Commentary: This plot shows the rate of expected employment loss in the top 10 states with the highest COVID-19 hospitalization rates. States severely affected by the pandemic also exhibit significant employment losses, indicating a strong correlation between health impacts and economic disruptions.

First we used the COVID-19 Case Surveillance dataset to get the top 10 states. Then by mapping these states to the Household Pulse Survey dataset we could calculate the expected employment loss rates.

• Highest expected employment loss in the top : California was considered the highest expected to lose employees with a percentage higher than 34 percent.

• Lowest expected employment loss: Among the top 10 states, Ohio was the lowest in expected employment loss.

2.11 Relationship Between Household Income and Rate of Delayed or Unobtained Medical Treatment

2.11.1 Due to COVID

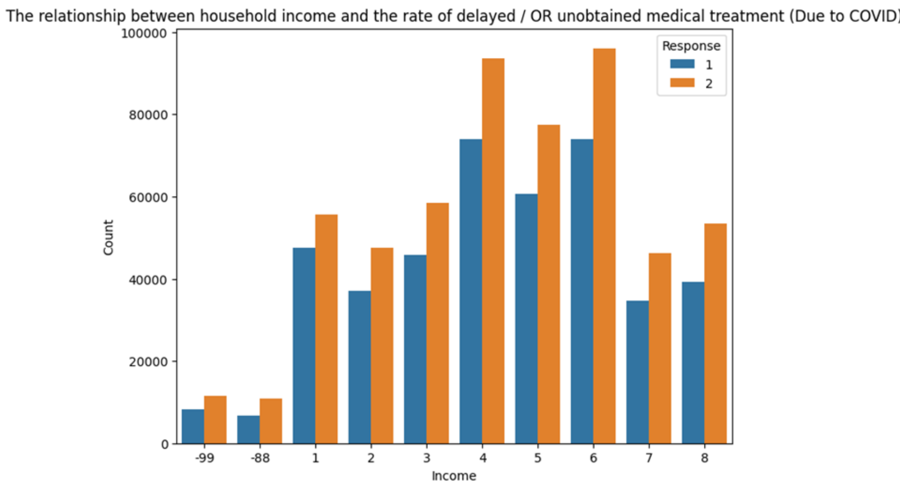

Figure 11: Household Income vs. Rate of Delayed/Unobtained Medical Treatment

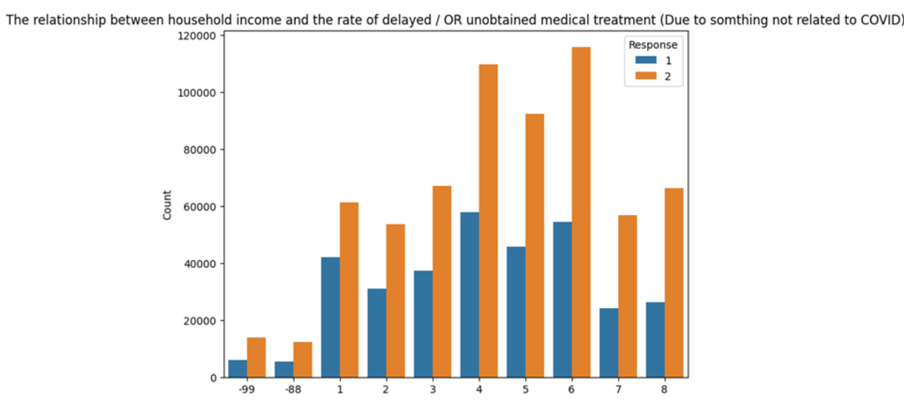

Figure 12: Household Income vs. Rate of Delayed/Unobtained Medical Treatment due to other reasons than COVID COVID

Commentary: This analysis investigates the relationship between household income and the rate of delayed or unobtained medical treatment, whether due to COVID-19 or other factors. Lower-income households experience higher rates of delayed or unobtained medical treatment in the case of no COVID but fairly equivalent delay percentage in the COVID case, highlighting healthcare access disparities.

2.12 Relationship Between COVID-19 Symptom Manifestation and Age Group

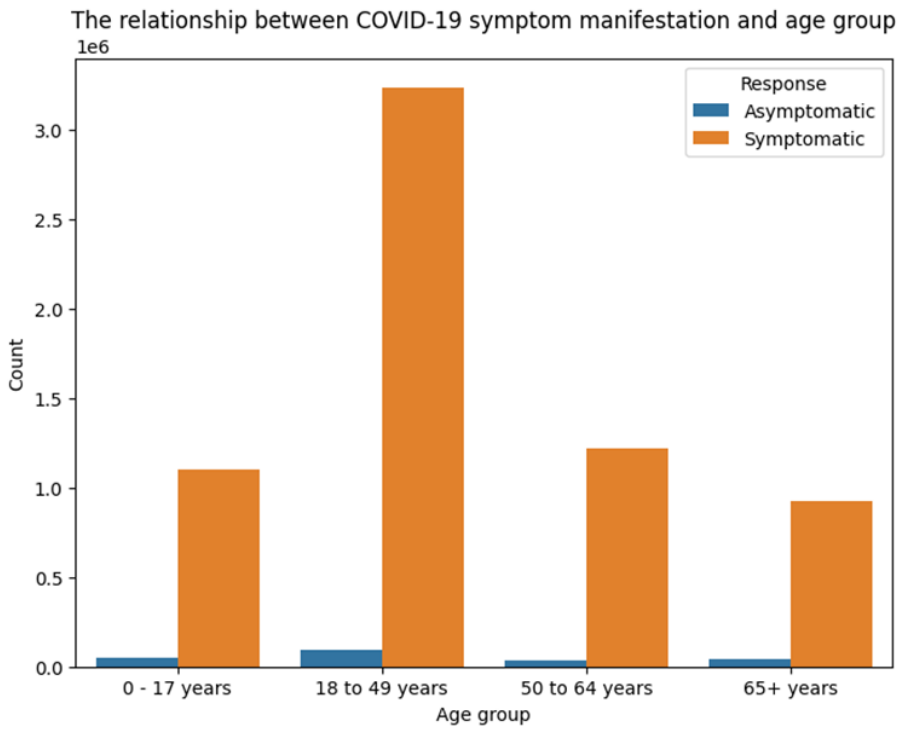

Figure 13: COVID-19 Symptom Manifestation by Age Group

Commentary: This plot examines the relationship between COVID-19 symptom manifestation and age groups. It shows how symptoms vary across different age groups, with older adults more likely to experience severe symptoms, while younger individuals often show milder symptoms or remain asymptomatic.
we could see that the most affected age group by the symptoms is the 18 to 49 years group, which makes sense as this group represents most of the industry workers.

3 Part 2: Answering Questions

3.1 Underlying Conditions and Mortality

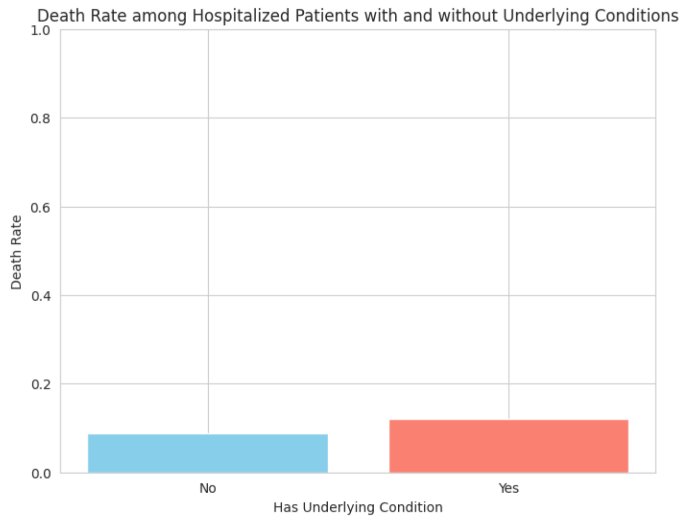

Figure 14: Mortality Rates among Hospitalized Patients with Underlying Conditions

Commentary: we aimed to determine whether individuals with underlying medical conditions were more susceptible to COVID-19-related mortality. Upon plotting the data, we found that individuals with medical conditions had slightly higher mortality rates, albeit with a very minimal difference. Consequently, due to the negligible difference, we cannot conclude a strong relationship between underlying medical conditions and mortality rates.

3.2 Demographic Risk Analysis

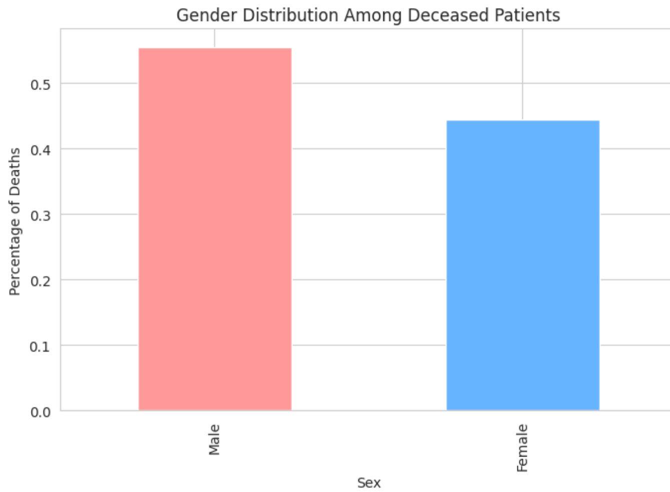

Figure 15: Gender Distribution Among Deceased Patients

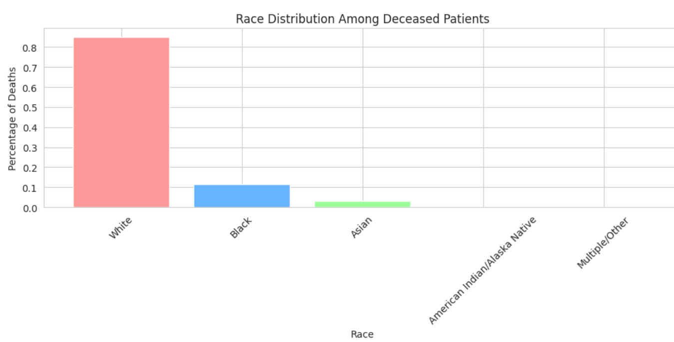

Figure 16: Race Distribution Among Deceased Patients

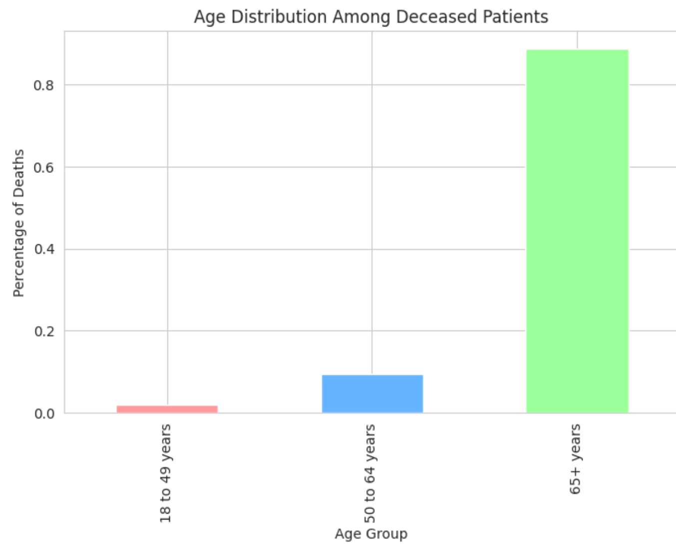

Figure 17: Age Distribution Among Deceased Patients

Commentary: we created three graphs to identify the demographic segments most susceptible and least susceptible to mortality. The demographics analyzed were gender, race, and age.

In the first graph, focusing on age, we found that the highest mortality rate was among individuals aged 65 and above. This finding was consistent and expected, aligning logically with our previous analyses and showing a significant difference compared to other age groups.

In the second graph, examining gender and mortality rates, we observed a markedly higher mortality rate among males compared to females.

In the third graph, analyzing race and mortality rates, we found that the highest mortality rates were among White individuals, followed by Black individuals, and then Asians.

Consequently, we can conclude that the demographic segment most vulnerable to COVID-19 mortality is elderly White males, while the least vulnerable segment is young Asian females.

3.3 Exposure and Hospitalization

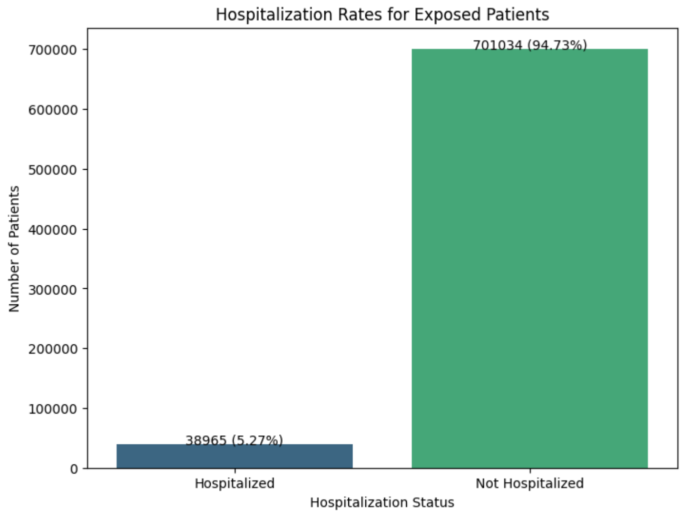

Figure 18: Hospitalization Rates among Patients with Reported Exposure

Commentary: we aimed to investigate the relationship between individuals who reported exposure to any form of travel or congregation within the 14 days prior to illness and the hospitalization rate. Surprisingly, we found a somewhat unexpected result. We discovered that individuals who reported exposure to crowded environments actually went unhospitalized. This means that the majority of those who interacted in crowded settings did not seek hospitalization

3.4 Asymptomatic Cases and Outcomes

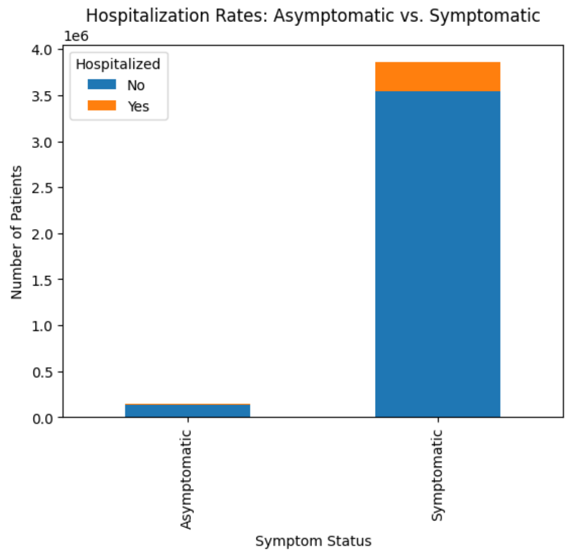

Figure 19: Hospitalization Rates: Asymptomatic vs. Symptomatic

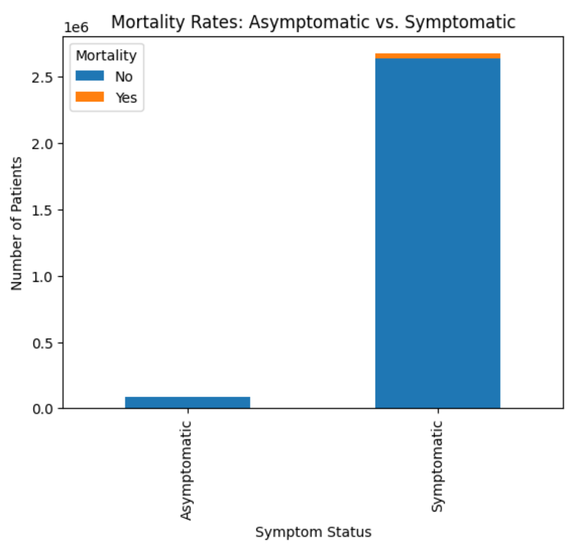

Figure 20: Mortality Rates: Asymptomatic vs. Symptomatic

Commentary: we aimed to determine whether individuals without COVID-19 symptoms are less likely to be hospitalized.

In the first graph, we found that the percentage of individuals who are symptomatic and not hospitalized is much lower than the percentage of individuals who are symptomatic and actually hospitalized.

In the second graph, we found that the death rate among symptomatic individuals is significantly higher compared to those who are not symptomatic.

These findings suggest that individuals with COVID-19 symptoms are more likely to be hospitalized and have a higher mortality rate compared to those without symptoms.

3.5 Economic Impact by State

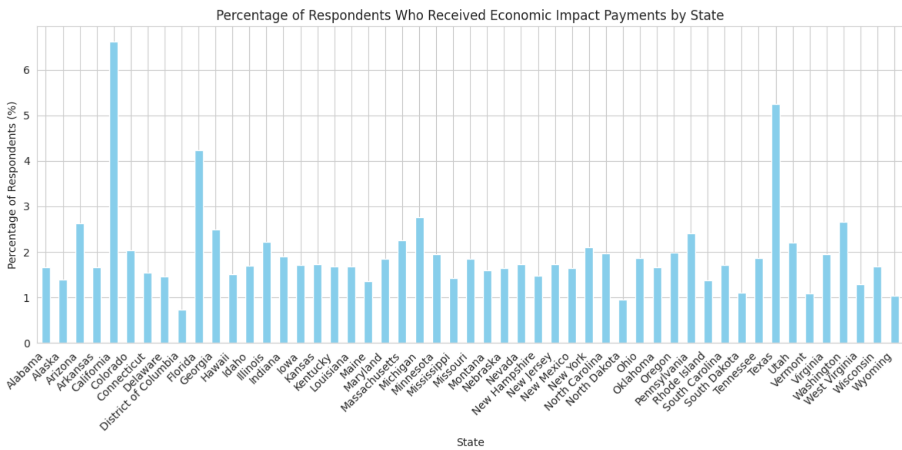

Figure 21: States with Highest Percentage of Economic Impact Payments

Commentary: we investigated the percentage of respondents who received Economic Impact Payments by state. California emerged as the top state on the list, followed by Utah and Georgia. These states may have topped the list due to their large populations, diverse economies, and possibly higher numbers of individuals eligible for economic relief payments.

3.6 Employment loss variation across industries

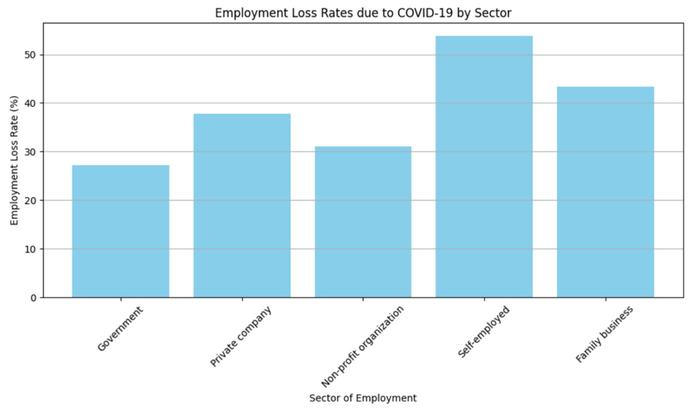

Figure 22: How the rate of COVID-related employment loss varies across different industries

Commentary: In this plot, we are investigating the percentage of employment loss through each of the employment sectors.
We can see the following:
• Highest expected employment loss: Self-employment sector was expected to lose about 53 percent of the employees, which makes sense as self-employment in that time was not reliable due to the non-sustainable salary.
• Lowest expected employment loss: Government employment was expected to lose less than 28 percent of the employees in this sector, which makes sense as it is considered the most sustainable salaries out of all the sectors.
• These results also aligns with the results of the expected employment loss we discussed earlier.

3.7 COVID-19 symptom severity and delayed medical treatment

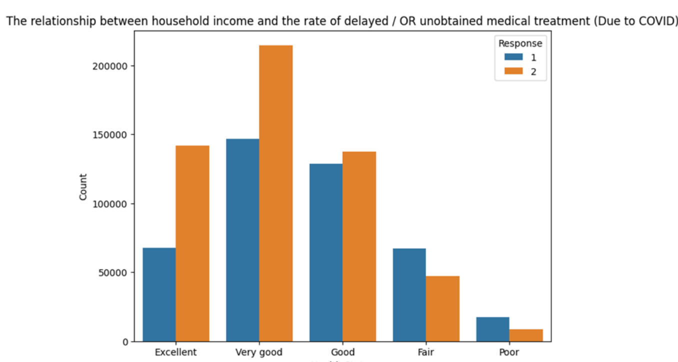

Figure 23: the correlation between COVID-19 symptom severity and delayed medical treatment

Commentary: In this plot, we could see that the delay rate is inversely proportional with the improvement of health status. As we approach poor health status the delay rate decreases and vice versa.

3.8 Is there a correlation between Household Size and Food Spending during pandemic

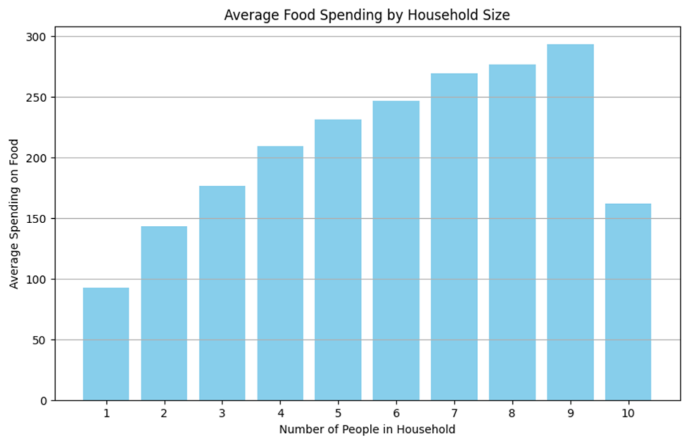

Figure 24: The correlation between Household Size and Food Spending during pandemic

Commentary: In this plot, we could see that as the number of people in a household increases, the average spending on food also increases, but we could see that this relation is slightly converging. We could also see an outlier at the number of people in a household = 10 as it is not following the same relation.

3.9 The correlation between Household Composition and Education Spending during pandemic

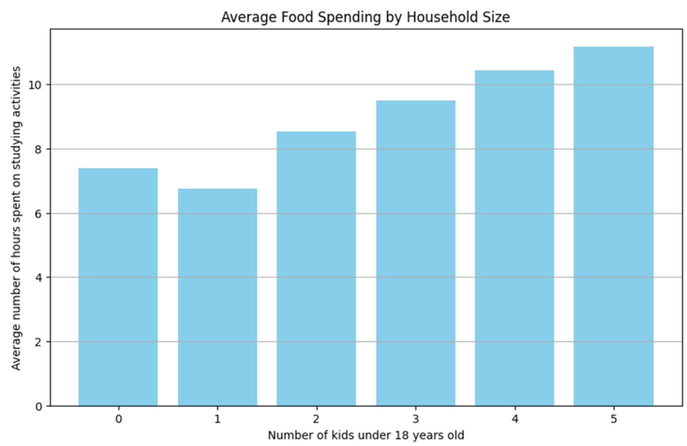

Figure 25: The correlation between Household Composition and Education Spending during pandemic

Commentary: In this plot, we could see that as the number of kids under 18 years old increase, the average number of hours spent on studying activities increases, but we could see that there is an outlier at the number of kids = 0 as it is wrongly imbedded in the data.

3.10 Is there a correlation between Household Composition and Education Spending during pandemic

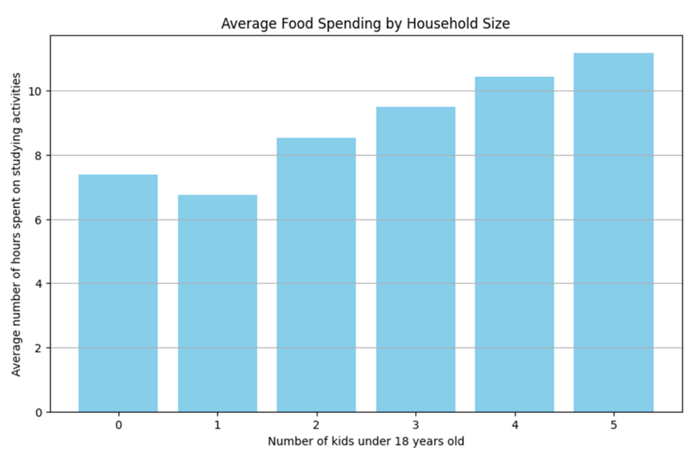

Figure 26: The correlation between Household Composition and Education Spending during pandemic

Commentary: In this plot, we could see that as the number of kids under 18 years old increase, the average number of hours spent on studying activities increases, but we could see that there is an outlier at the number of kids = 0 as it is wrongly imbedded in the data.

4 Part 3: Hypothesis Testing

4.1 Testing the Association Between Demographics and Death

4.1.1 Hypothesis Formulation

We will test the following hypotheses:

- **Null Hypotheses (H_o):** There is no association between the combined demographic features and the probability of death due to COVID-19.
- **Alternative Hypothesis (H_1):** There is a significant association between the combined demographic features and the probability of death due to COVID-19.

```
Chi-Square Statistic (demographics): 24708.56, P-value: 0.0
```

Commentary: Based on the p-value that less than 0.05, we have found that there is a significant association between demographics and death due to COVID-19.

4.2 Investigate the Relationship Between Demographic Features and ICU Admission for COVID-19 Patients.

4.2.1 Hypothesis Formulation

New Claim: There is a significant association between the probability of ICU admission due to COVID-19 and patient demographics.

- **Null Hypothesis (H0):** There is no association between the combined demographic features and the probability of ICU admission due to COVID-19.
- **Alternative Hypothesis (H1):** There is a significant association between the combined demographic features and the probability of ICU admission due to COVID-19.

```
Chi-Square Statistic (ICU Admission): 2557.67, P-value: 0.0
```

Commentary: Based on the p-value that less than 0.05, we have found that there is a significant association between demographics and ICU admission due to COVID-19.

5 Part 4: Regression Analysis

Introduction to Regression Analysis

This section of our analysis focused on using statistical methods to identify factors that significantly affect the death percentage from COVID-19. Through the application of ordinary least squares (OLS) regression, we sought to quantify the impact of demographic and clinical variables on COVID-19 mortality rates.

5.1 Models Developed

1. **Base Model without Intercept:** Initially, we developed a regression model excluding the intercept to directly relate changes in predictors to changes in death percentage without a baseline shift. This model provided a basic understanding but was limited in interpretability.
2. **Model with Intercept:** Adding an intercept significantly improved the interpretability of the model by providing a baseline death percentage when all predictor variables are zero. This model had an R-squared of 0.966, indicating a high level of explanatory power.
3. **Model with Higher-Order Terms:** To capture non-linear relationships, we introduced higher-order terms for the variables `hosp_pct` and `icu_pct`. This model achieved an R-squared of 0.989, demonstrating even greater accuracy in explaining the variance in death percentages.

5.2 Key Findings

- **Good Predictors:**
  - **ICU Admission Rate (`icu_pct`):** This variable has a highly significant p-value (p ¡ 0.001) and a large coefficient (0.8760), indicating a strong and statistically significant influence on the death percentage. Higher ICU rates correspond to higher death percentages, likely reflecting the severity of cases in the ICU.
  - **Hospitalization Rate (`hosp_pct`):** Though the effect is smaller, hospitalization percentage is also a statistically significant predictor (p = 0.013) with a negative coefficient (-0.1876). This suggests that increased hospitalization rates might be associated with a decrease in death percentage, potentially due to timely medical intervention.
- **Poor Predictors:**
  - **Age Groups (0 - 17 years, 18 to 49 years, 50 to 64 years, 65+ years):** None of the age group predictors are significant (all p-values ¿ 0.05), indicating that they do not significantly predict the death percentage in the context of this model.
  - **Gender (female_pct, male_pct):** Both gender coefficients are not statistically significant, suggesting that gender does not play a discernible role in predicting the death percentage in this model setup.
- **Correlated Predictors:**
  - **Age Groups:** There are significant correlations between different age groups, particularly:
    * **65+ years_pct** is strongly negatively correlated with **18 to 49 years_pct** (-0.91).
    * Other age groups also show moderate to high correlations among each other.
  - **Hospitalization and ICU Rates (`hosp_pct` and `icu_pct`):** These show a very high correlation (0.93), indicating they often co-occur, likely reflecting the severity of the cases.
- **Higher-Order Terms:** The introduction of squared terms for hospitalization and ICU rates allowed us to capture the non-linear effects, showing that the relationships between these variables and death percentages are more complex than initially assumed.

Visualization

In this subsection, we visualize the significant findings from the regression analysis.

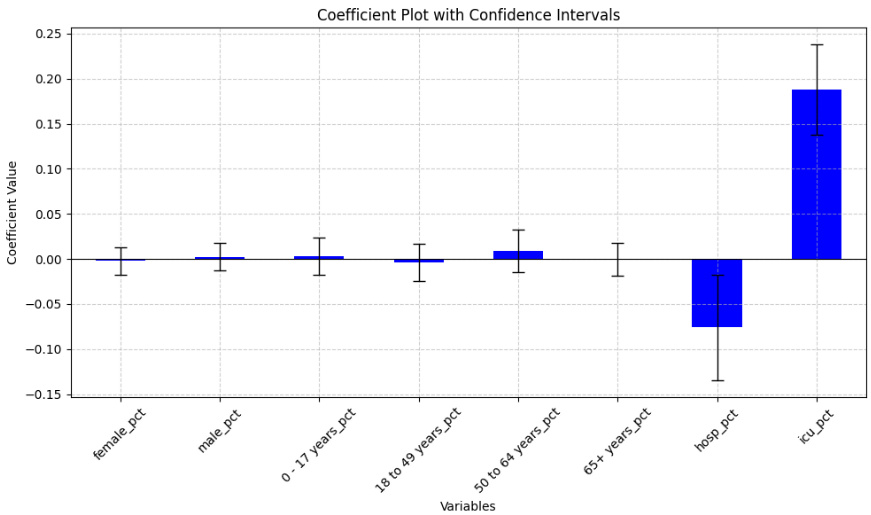

Figure 27: Coefficient Plot with Confidence Intervals. This plot shows the estimated coefficients and their 95% confidence intervals for each predictor in the model. Notably, predictors such as ICU and hospitalization percentages have significant impacts with wide confidence intervals, indicating strong effects with a degree of uncertainty.

Below is the correlation matrix of the predictors, which helps to identify multicollinearity issues among the predictors.

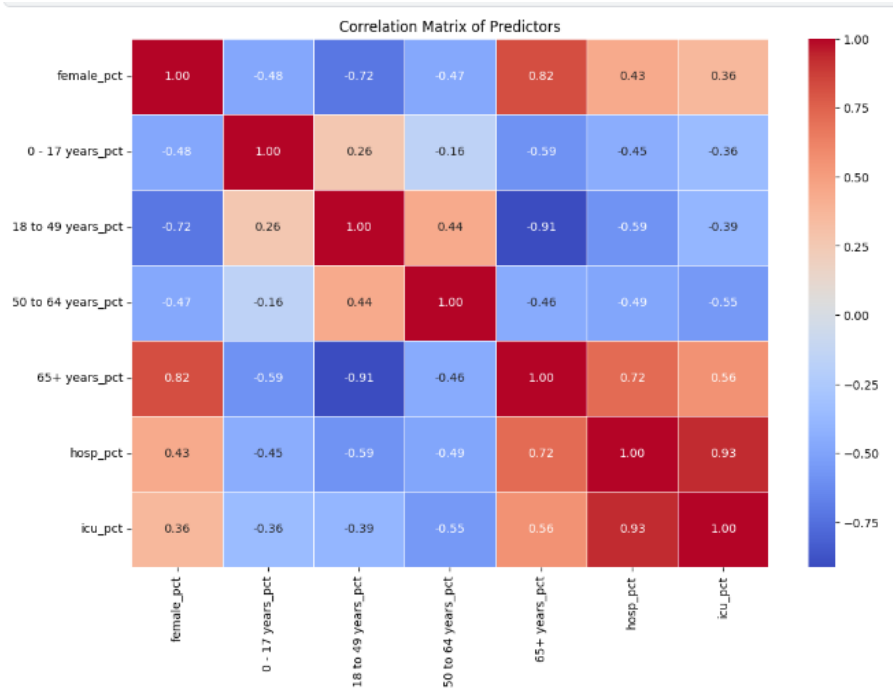

Figure 28: Correlation Matrix of Predictors. This matrix illustrates the relationships between different predictor variables, highlighting significant correlations such as the high correlation between hospitalization and ICU rates.

5.3 Conclusion

The regression analysis provided deep insights into the factors influencing COVID-19 death percentages. Significant predictors like hospitalization and ICU rates were identified, and their non-linear relationships with the outcome were modeled effectively. However, the presence of multicollinearity, especially in models with many predictors, suggests caution in interpreting the coefficients directly without considering potential statistical artifacts.

6 Part 5: Machine Learning Model

6.1 Introduction to Gradient Boosting

In this section, we applied Gradient Boosting, a powerful ensemble machine learning technique, to predict the likelihood of death from COVID-19 based on various demographic and clinical features. Gradient Boosting builds multiple weak learners (typically decision trees) in a sequential manner, with each new tree correcting the errors of the previous ones.

6.2 Data Preprocessing and Feature Engineering

The dataset was preprocessed to ensure clean and reliable inputs for the model:

- **Data Cleaning:** Rows with missing or unknown values in key columns (`age_group`, `sex`, `race`, `hosp_yn`, `icu_yn`, `death_yn`) were removed to avoid inaccuracies.
- **Feature Engineering:** Categorical variables were encoded using Label Encoding to convert them into a numeric format suitable for the model.
- **Feature Scaling:** The features were standardized using StandardScaler to ensure that all features contribute equally to the model's performance.

6.3 Model Training and Evaluation

We split the dataset into training and testing sets, with 80% of the data used for training and 20% for testing. A Gradient Boosting Classifier with 100 estimators was then trained on the scaled training data. The model's performance was evaluated using accuracy, confusion matrix, and classification report metrics.

6.4 Model Performance

The model achieved an accuracy of 96.42%, indicating a high level of performance. Below is the detailed evaluation of the model:

```
Accuracy: 0.9642098877767667
Confusion Matrix:
[[3034   29]
 [  89  145]]
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.99      0.98      3063
           1       0.83      0.62      0.71       234

    accuracy                           0.96      3297
   macro avg       0.90      0.81      0.85      3297
weighted avg       0.96      0.96      0.96      3297
```

Figure 29: Confusion Matrix and Classification Report. This output shows the model's accuracy, precision, recall, and F1-score for predicting death from COVID-19.

6.5 Key Metrics

- **Accuracy:** 96.42%
- **Confusion Matrix:**
  - True Negatives: 3034
  - False Positives: 29

  - False Negatives: 89
  - True Positives: 145

- **Classification Report:**
  - **Class 0 (Non-death):** Precision = 0.97, Recall = 0.99, F1-score = 0.98
  - **Class 1 (Death):** Precision = 0.83, Recall = 0.62, F1-score = 0.71
  - **Overall:** Macro average Precision = 0.90, Recall = 0.81, F1-score = 0.85
  - **Weighted average:** Precision = 0.96, Recall = 0.96, F1-score = 0.96

The Gradient Boosting Classifier demonstrated robust performance in predicting COVID-19 deaths based on demographic and clinical features. Despite the high overall accuracy, the recall for predicting deaths was lower, suggesting further refinement may be needed for better sensitivity to true positive cases.

7 Conclusion

This comprehensive analysis has provided profound insights into the intricate relationships between demographic features and severe COVID-19 outcomes. Through a detailed regression analysis and a robust machine learning model, we have successfully identified key predictors of death, hospitalization, and ICU admission rates due to COVID-19.

The regression analysis revealed that ICU admission rates significantly influence the death percentage, with higher ICU admissions correlating strongly with increased mortality. Conversely, higher hospitalization rates are associated with a reduction in mortality, likely due to timely medical intervention. Although age group variables did not emerge as statistically significant predictors in the regression model, the exploratory analysis highlighted discernible patterns of vulnerability, particularly among older populations.

The machine learning model, utilizing Gradient Boosting, further validated these findings, achieving high accuracy in predicting COVID-19 mortality. However, it also underscored the need for enhanced sensitivity to accurately identify true positive cases, indicating areas for further refinement.

8 Significance and Limitations

8.1 Significance

The findings of this analysis hold significant implications for public health strategies and resource allocation. By pinpointing the critical demographic and clinical predictors of severe COVID-19 outcomes, this study equips policymakers and healthcare providers with the necessary information to prioritize interventions and support for high-risk groups.

The strong association between ICU admissions and mortality underscores the crucial importance of maintaining adequate ICU capacity and strengthening healthcare infrastructure to manage severe cases effectively. Additionally, the correlation between timely hospitalizations and reduced mortality rates highlights the need for accessible and efficient healthcare services, which are pivotal in managing the pandemic's impact.

These insights can inform targeted public health campaigns, optimize resource distribution, and guide policy decisions aimed at minimizing COVID-19 mortality and enhancing patient outcomes. Furthermore, the findings can be leveraged to develop more effective communication strategies to educate the public about the importance of seeking timely medical care.

8.2 Limitations

Despite the robustness of the analysis, several limitations warrant consideration. The quality and completeness of the data present inherent challenges, as missing or misclassified data can introduce biases and affect the accuracy of the findings. The analysis did not encompass other potential risk factors such as pre-existing medical conditions, socioeconomic status, and regional variations in healthcare quality, which could provide a more comprehensive understanding of the factors influencing COVID-19 outcomes.

The high correlation between certain predictors, notably ICU and hospitalization rates, indicates the presence of multicollinearity, which complicates the interpretation of individual predictor effects. Future studies should aim to address these limitations by incorporating more granular data and additional relevant variables to enhance the comprehensiveness and accuracy of the analysis.

Furthermore, while the machine learning model demonstrated impressive accuracy, the relatively lower recall for predicting deaths suggests the need for further refinement. More sophisticated algorithms and enhanced feature engineering could potentially improve predictive performance and sensitivity.

In conclusion, this report provides invaluable insights into the demographic and clinical factors associated with severe COVID-19 outcomes. These findings serve as a crucial resource for guiding public health interventions and informing policy decisions, ultimately aiming to better manage and mitigate the pandemic's impact. Continued research and data refinement are essential to further elucidate these relationships and enhance our ability to respond effectively to future public health challenges.