



Zewail City of Science, Technology, and Innovation
University of Science and Technology
Communications and Information Engineering
Analysis of U.S. COVID-19 Data Impact

Statistical Inference and Data Analysis

Prepared By

Mariam Elseedawy	201901281
Rodina Mohamed	201900642
Shrouk Shata	201902199
Omar Ayman	202100443

Supervised By

Dr. Mahmoud Abdelaziz
Eng. Asmaa Mostafa

Spring 2024

Contents

1	Introduction	2
1.1	Background	2
1.2	Purpose of the Report	2
1.3	Objectives	2
1.4	Scope	2
1.5	Significance	3
2	Methodolody and Results	3
2.1	Exploratory Analysis	3
2.2	Questions Analysis	13
2.3	Hypothesis Tests	19
2.3.1	First Claim	19
2.3.2	Second Claim	23
3	Regression Analysis	24
3.1	Preprocessing and Proportion Calculation	25
3.1.1	Preprocessing Steps	25
3.1.2	Proportion Calculation	25
3.2	Trial 1 Summary	26
3.2.1	Results	26
3.2.2	Interpretation	26
3.3	Trial 2 Summary	27
3.3.1	Results	27
3.3.2	Interpretation	27
3.4	Comparison Between the Two Trials	27
4	Conclusion	28
4.1	Key Insights	28
4.1.1	Exploratory Analysis	28
4.1.2	Specific Questions	28
4.1.3	Hypothesis Testing	28
4.1.4	Regression Analysis	28
4.2	Future Research	28
4.2.1	Deeper Analysis	28
4.2.2	Longitudinal Studies	28

1 Introduction

1.1 Background

The COVID-19 pandemic has emerged as one of the most significant global health crises in over a century, affecting millions of lives and reshaping societies and economies worldwide. Since its outbreak in late 2019, the virus has spread rapidly, necessitating urgent and extensive public health responses. The United States has been particularly hard-hit, with millions of confirmed cases and a substantial death toll, prompting widespread economic and social disruptions.

1.2 Purpose of the Report

This report aims to analyze the extensive data collected during the pandemic to uncover critical trends and insights into the spread and impact of the virus across different demographics and regions within the United States. By leveraging the COVID-19 Case Surveillance Public Use Data and the Household Pulse Survey, this analysis seeks to understand the factors influencing disease outcomes and to provide evidence-based recommendations for policymakers and health practitioners.

1.3 Objectives

- **To analyze COVID-19's impact across different demographic groups**, including age, sex, and race, to identify which populations are most at risk.
- **To study the trends in hospitalizations and deaths** over time and across various states to evaluate the effectiveness of public health interventions.
- **To explore the economic implications** of the pandemic, focusing on employment impacts and healthcare access disparities.
- **To develop predictive models** using statistical and machine learning techniques to forecast future trends in COVID-19 cases and outcomes.

1.4 Scope

The report will cover data up to the current date and will include detailed exploratory data analysis, hypothesis testing, regression analysis, and predictive modeling. The findings will be presented

with comprehensive visualizations to aid in the interpretation of the data and to support the conclusions and recommendations provided.

1.5 Significance

The insights generated from this report are intended to aid in decision-making and strategy formulation for handling ongoing and future public health challenges posed by COVID-19. By identifying key trends and demographic disparities, the report aims to contribute to more targeted and effective public health responses and policies.

2 Methodology and Results

2.1 Exploratory Analysis

In this section, we show our EDA to answer 10 main questions, commenting on them as well.

1. The total number of hospitalizations versus deaths from COVID-19 over the entire US per month-year timestamp.

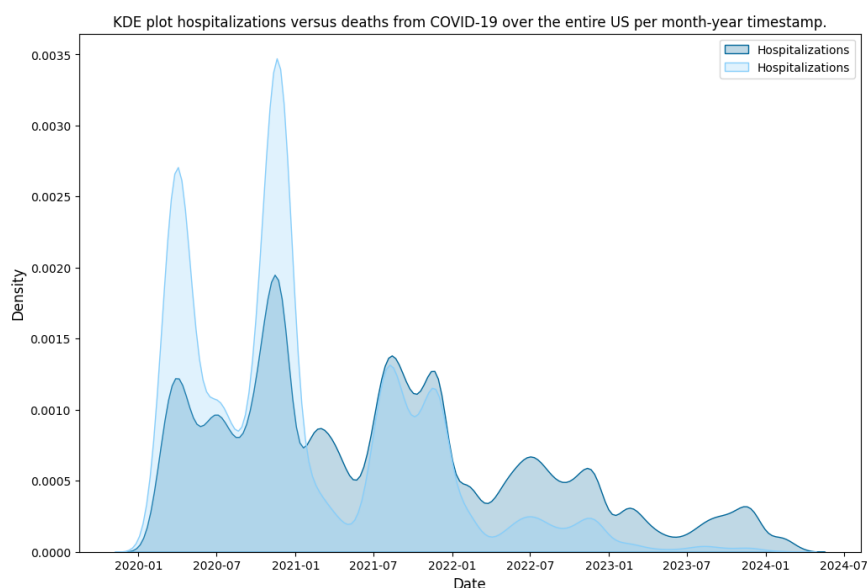


Figure 1: Hospitalizations versus deaths from COVID-19 over the entire US.

Comments: The number of deaths and hospitalizations reached their peak in the first year of the pandemic. However, as time passes, both numbers decrease as the number of infected people decreases.

2. The average rates of COVID-related deaths relative to patient demographics (age, sex, race)

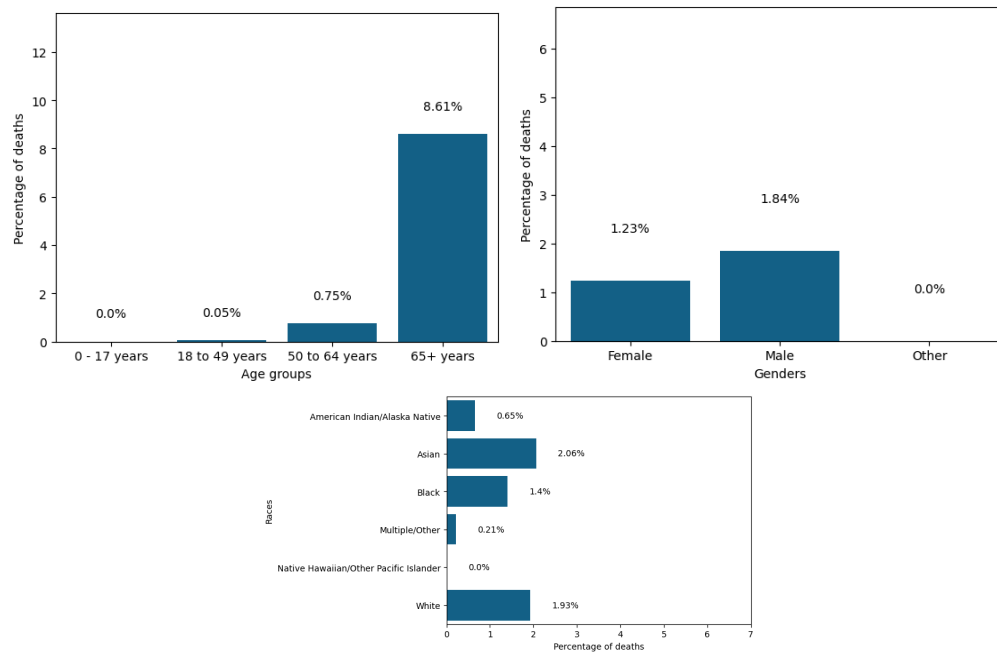


Figure 2: Percentages of deaths with different demographics

Comments: People aged 65+ years are exposed to deaths than any other age group. This is mainly because of the age factor where the elderly are more prone to diseases. Concerning gender, Men and women have an equal likelihood of dying from COVID-19. For different races, most of the deaths from the collected data are Asians living in the US.

3. The rates of COVID-related hospitalization and death with age (across age groups).

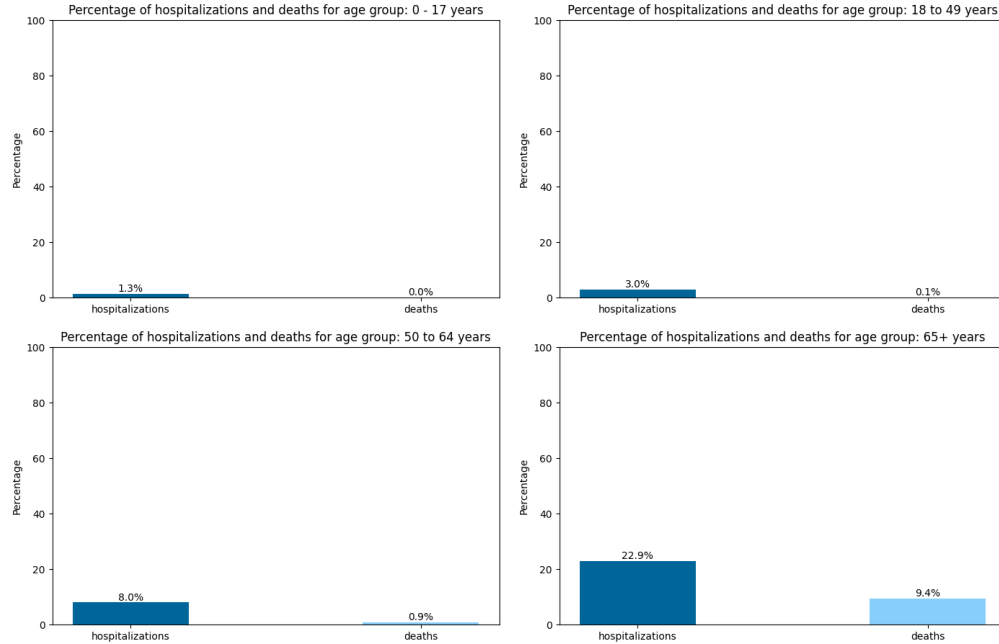


Figure 3: Hospitalization and death rates with different age groups

Comments: We can see that the most affected age group is those aged 65+ years. In contrast, the least affected age group is the group aged from 0 to 17 years old, which makes sense as the elderly are more prone to getting infected.

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

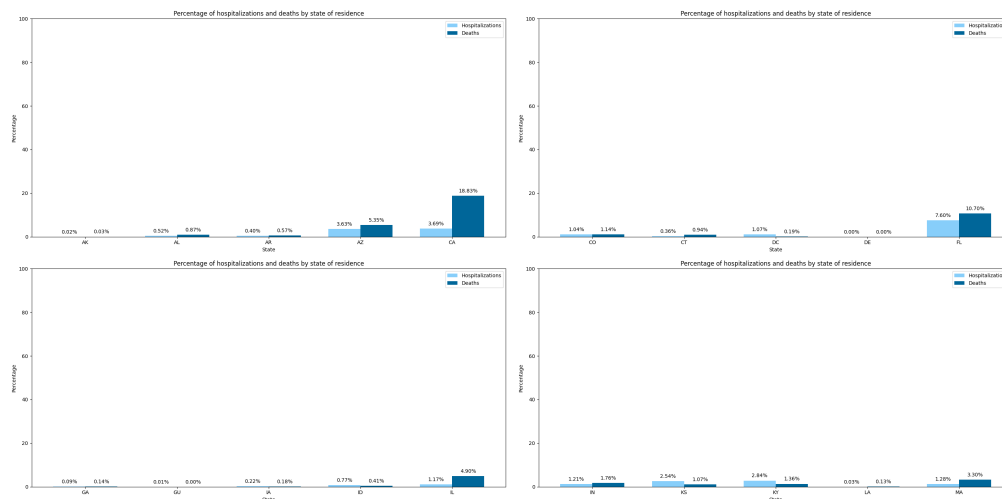


Figure 4: Rate of COVID-related hospitalization and death per some U.S. states (1)

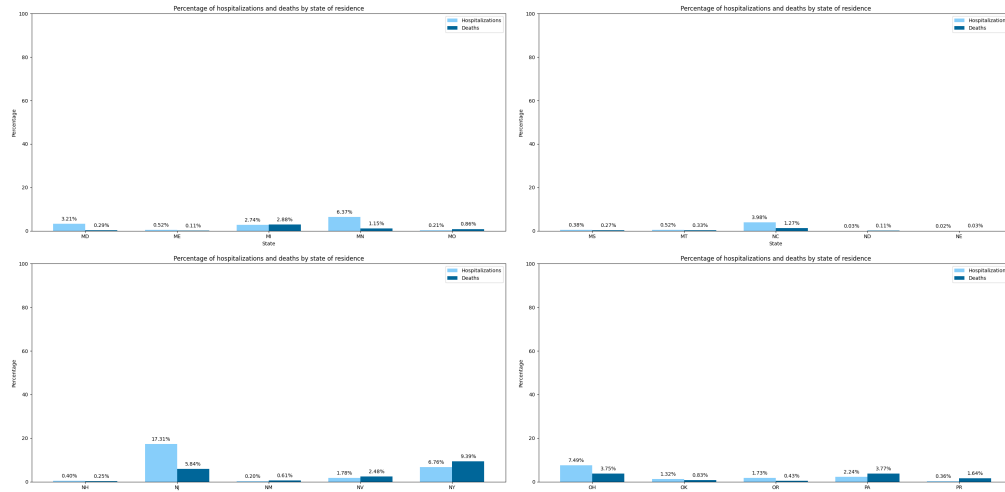


Figure 5: Rate of COVID-related hospitalization and death per some U.S. states (2)

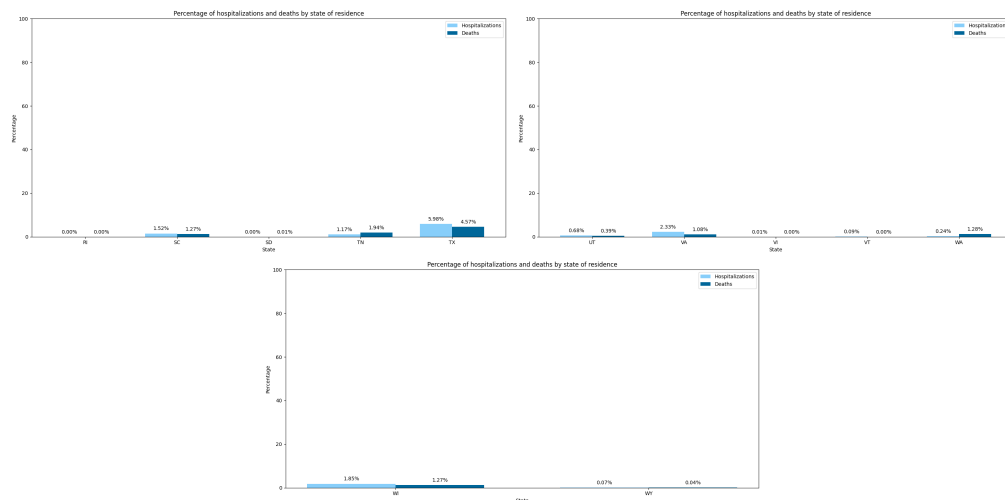


Figure 6: Rate of COVID-related hospitalization and death per some U.S. states (3)

Comments: Most affected states are California and New York. This can be explained as they are from the most populated states in the US.

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

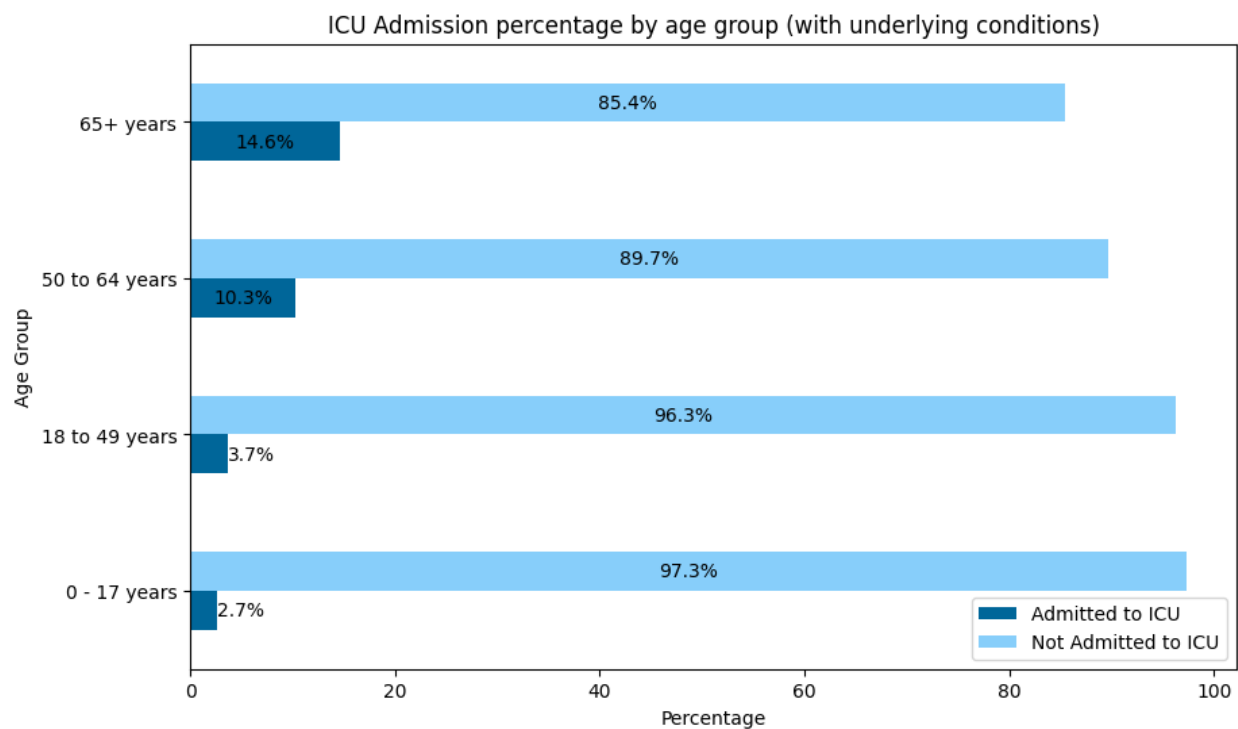


Figure 7: ICU Admission percentage by age group (with underlying conditions)

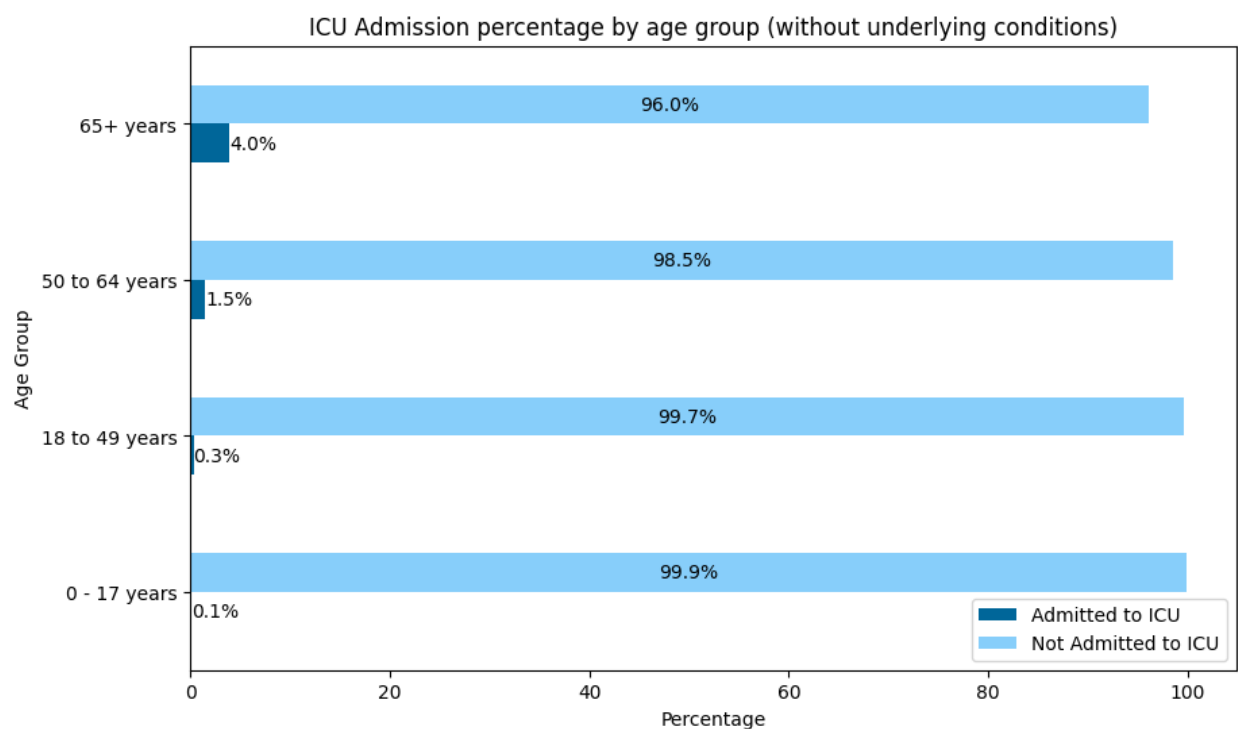


Figure 8: ICU Admission percentage by age group (without underlying conditions)

Comments: People with underlying conditions are more likely to be admitted to ICU. Also, The older, the more likely to be admitted to the ICU.

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

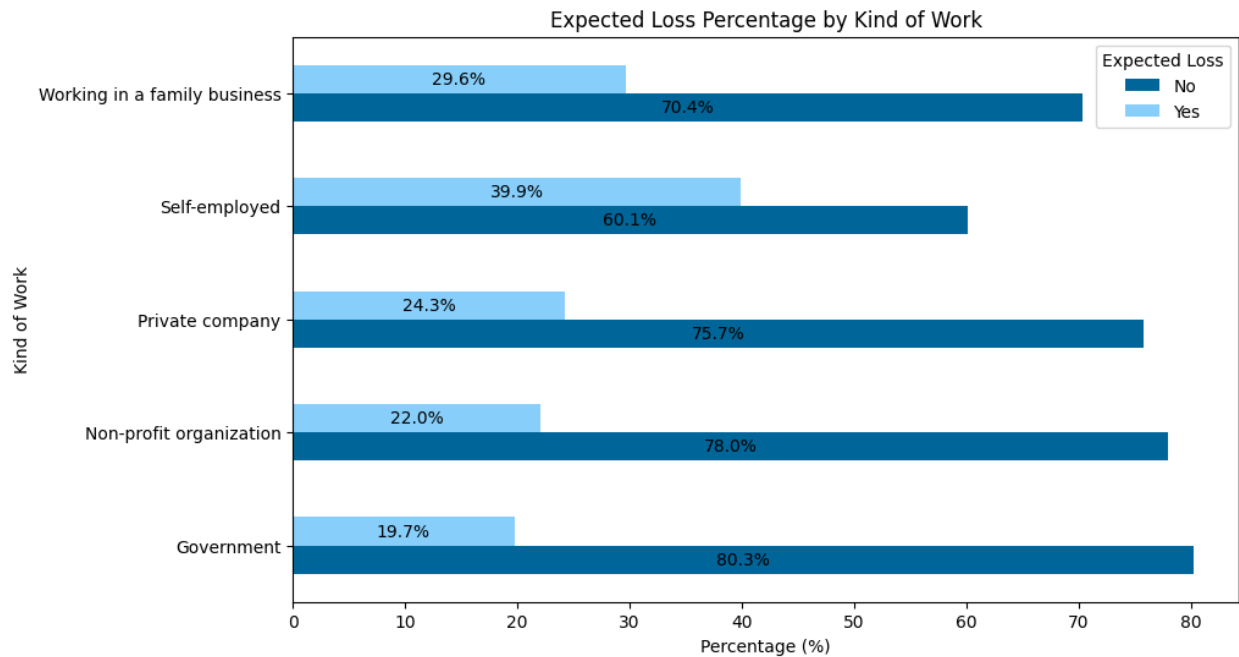


Figure 9: Expected employment loss percentage by kind of work

Comments: People who had their own business are the most affected by the pandemic because of the lockup that caused a big global recession (A significant, widespread, and prolonged downturn in economic activity.)

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

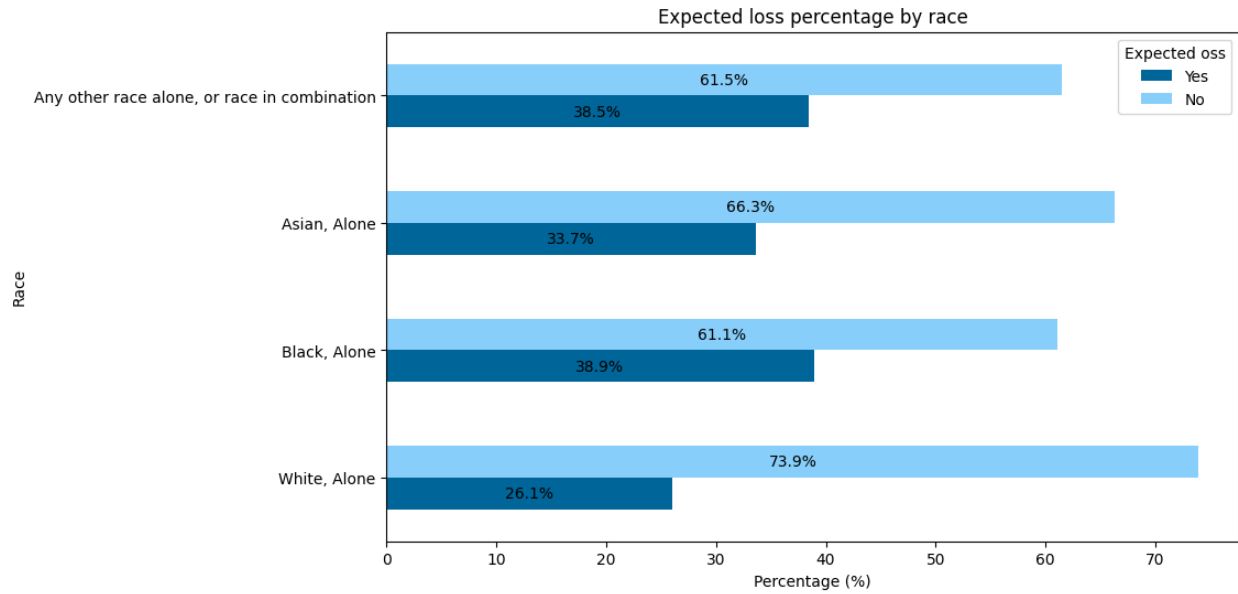


Figure 10: Expected employment loss percentage by race

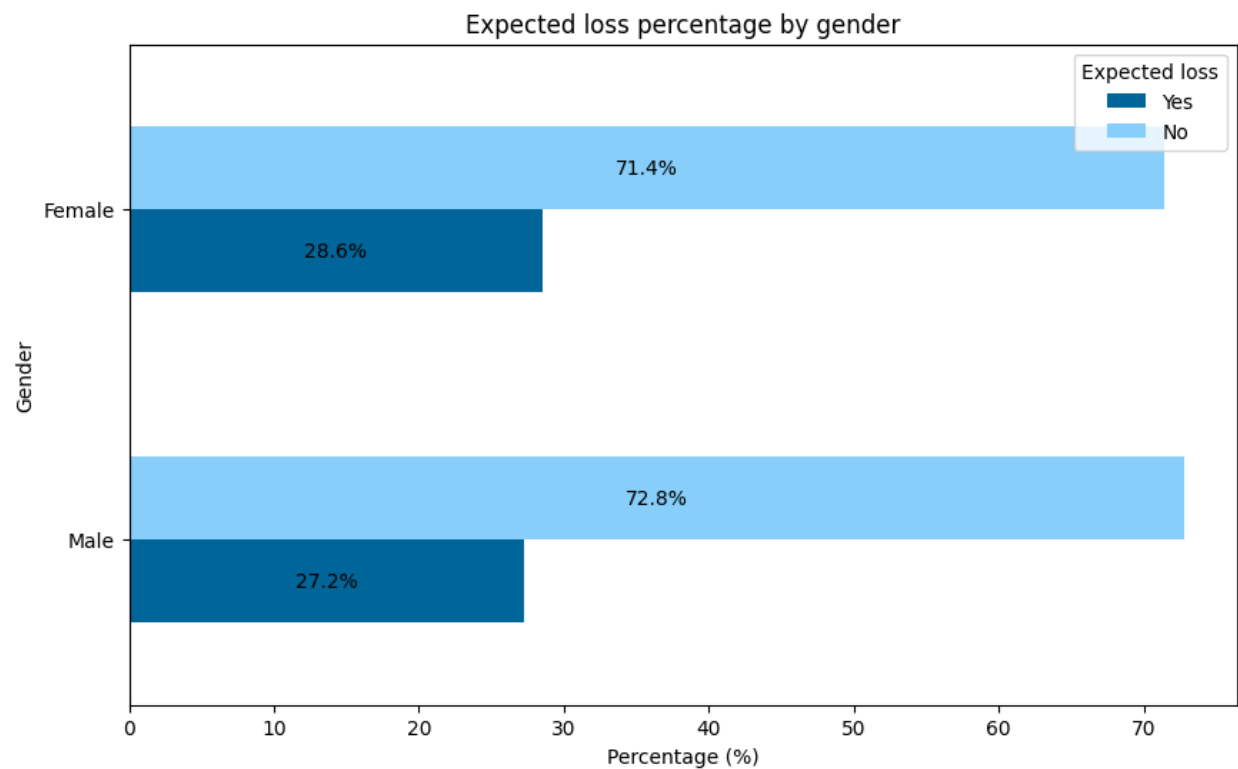


Figure 11: Expected employment loss percentage by gender

Comments: Men and women have an equal likelihood of losing their jobs. However, white

individuals are less likely to lose their jobs compared to black individuals.

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

Expected employment loss due to COVID-19 for the top 10 states with the highest rate of COVID hospitalization.

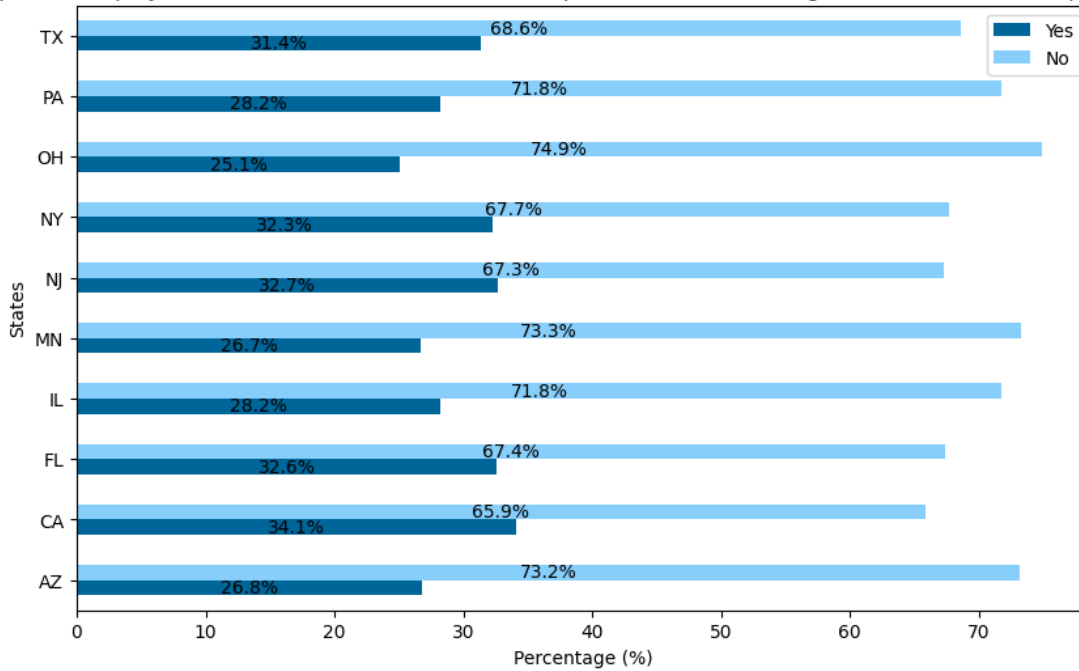


Figure 12: Expected employment loss for the top 10 states with the highest rate of COVID hospitalization.

Comments: Among the top 10 states with the highest rate of COVID-19 hospitalization, California has the highest rate of expected employment loss with **34.1%**.

9. The relationship between household income and the rate of delayed/ OR unobtained medical treatment.

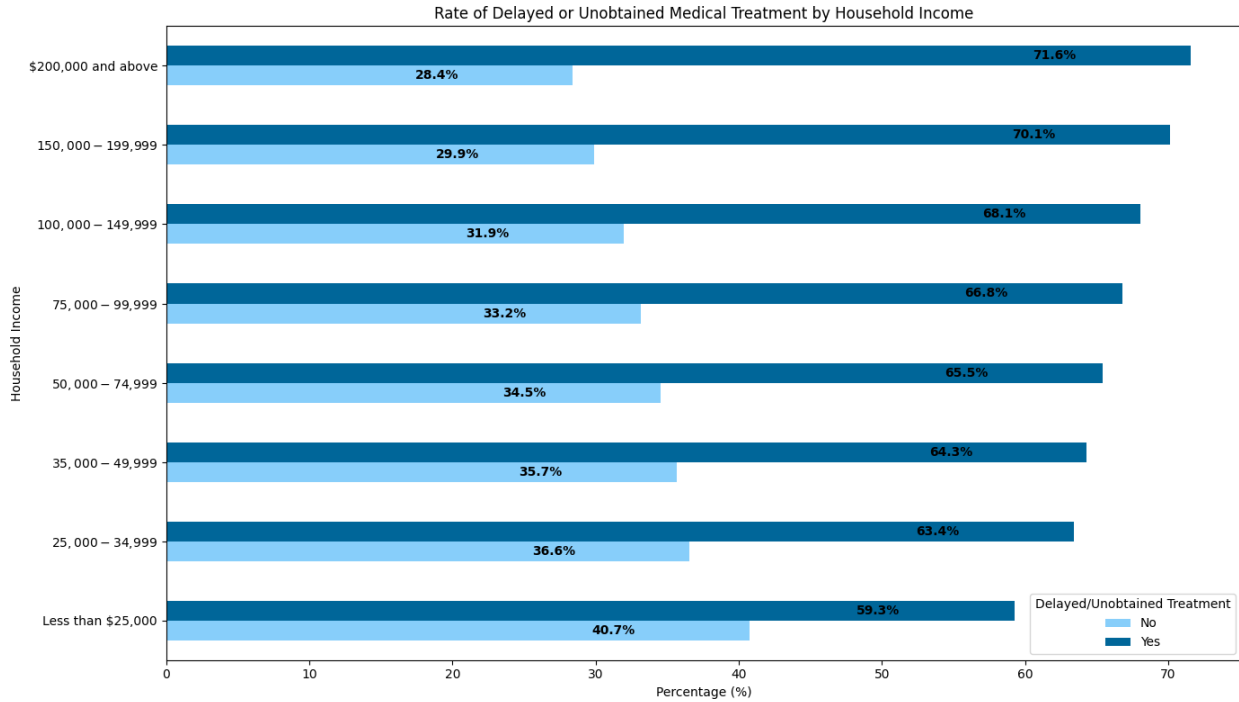


Figure 13: Rate of delayed or unobtained medical treatment by household income

Comments: Individuals with less income tend to delay the treatment or decide not to take it. This may be a result of the belief of anti-vaxxers about vaccines in general.

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

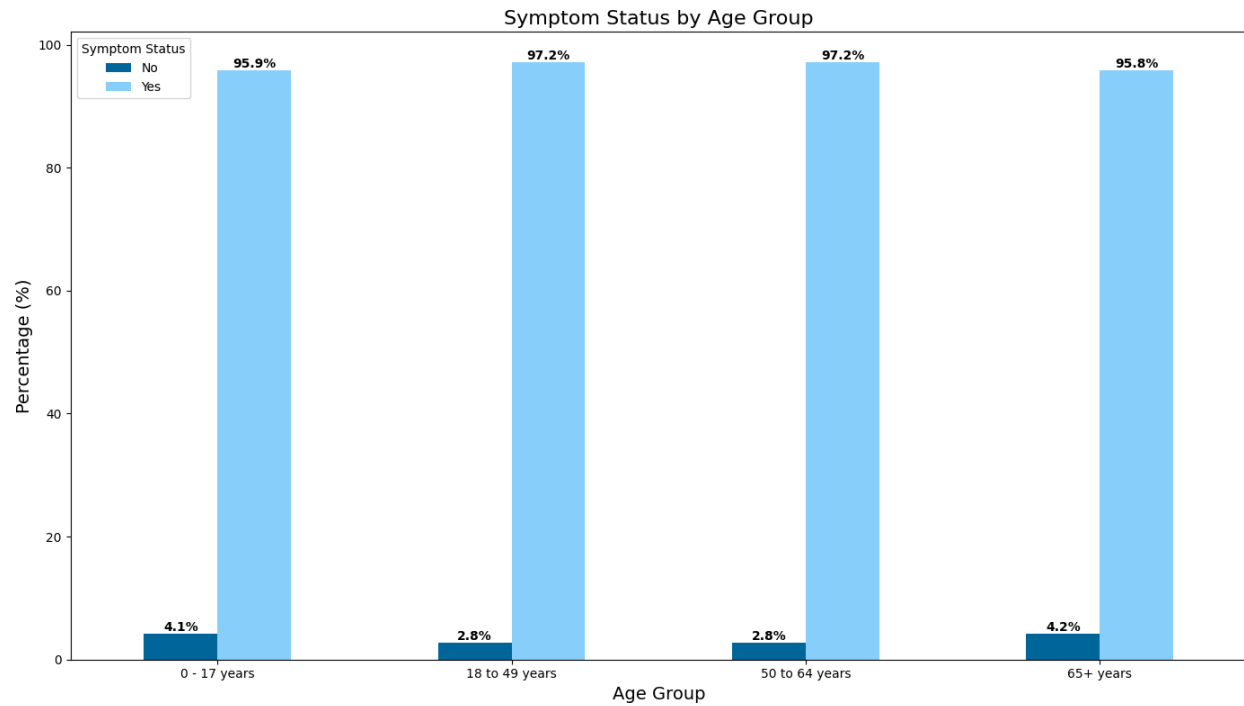


Figure 14: Rate of delayed or unobtained medical treatment by household income

Comments: Old individuals are more likely to experience symptom manifestation, which can lead to their hospitalization and maybe their death.

2.2 Questions Analysis

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

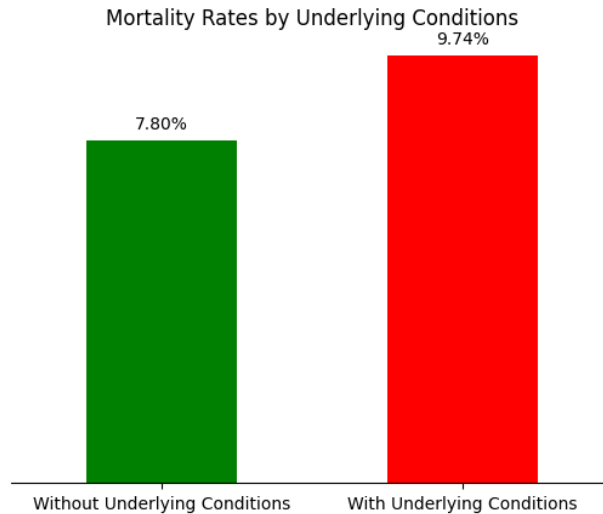


Figure 15: Mortality rates by underlying conditions

- The mortality rate for individuals **without** underlying conditions is **7.80%** while the mortality rate for individuals **with** underlying conditions is **9.74%**.
- Hospitalized patients with underlying medical conditions and/or risk behaviors are 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?

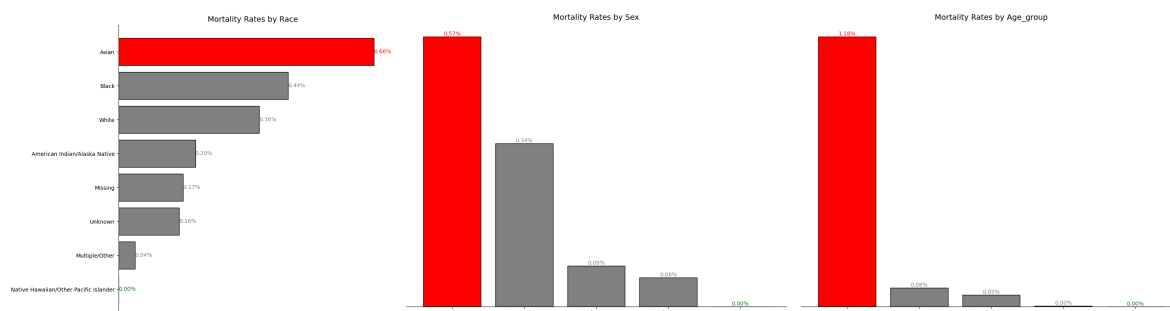


Figure 16: Demographic segment analysis

Most at Risk:

- (a) Older Adults (65+ years)
- (b) Males
- (c) Asian Racial Group

Least at Risk:

- (a) Children and Teenagers (0 - 17 years)
- (b) Individuals Identified as "Other" in Sex
- (c) Native Hawaiian/Other Pacific Islander Racial Group

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?**

Percentage of Exposed Patients Hospitalized



Figure 17: Percentage of Exposed Patients Hospitalized

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

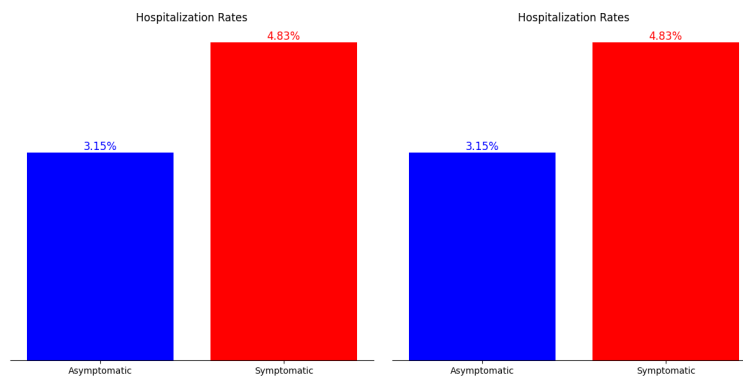


Figure 18: Symptoms, hospitalization, and Death Analysis

- Yes, asymptomatic COVID-19 patients are less likely to be hospitalized.
- Yes, asymptomatic COVID-19 patients are less likely to die from their illness.

5. How does household size affect the receipt and use of the Economic Impact Payment (EIP)?

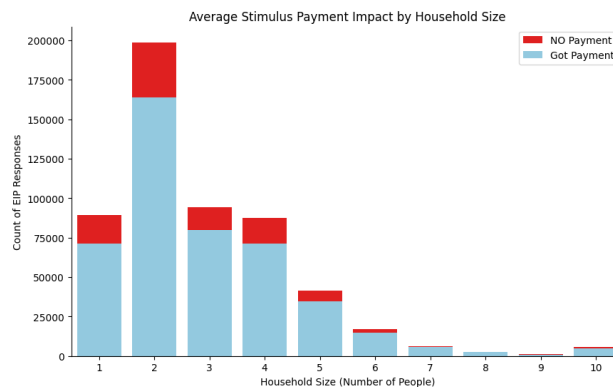


Figure 19: Average Stimulus Payment Impact by Household Size

- Those with 1 to 4 members, have a higher count of responses both for receiving and not receiving payments. This suggests that smaller households are more prevalent in the dataset or more likely to respond to the survey.
- The highest overall response is observed in households with 3 members, indicating that this household size is either more common or more engaged with the survey process.

- There is a significant drop in the number of responses as household size increases. Households with more than 5 members show markedly fewer responses, which might reflect their relative rarity or lower response rates from these groups.
- Overall, count of people who got payment is less than count of people who did not get payment.

6. Which state is associated with the highest percentage of Economic Impact (stimulus) payments among survey respondents? I focused on "Mostly spend it (1)" Option

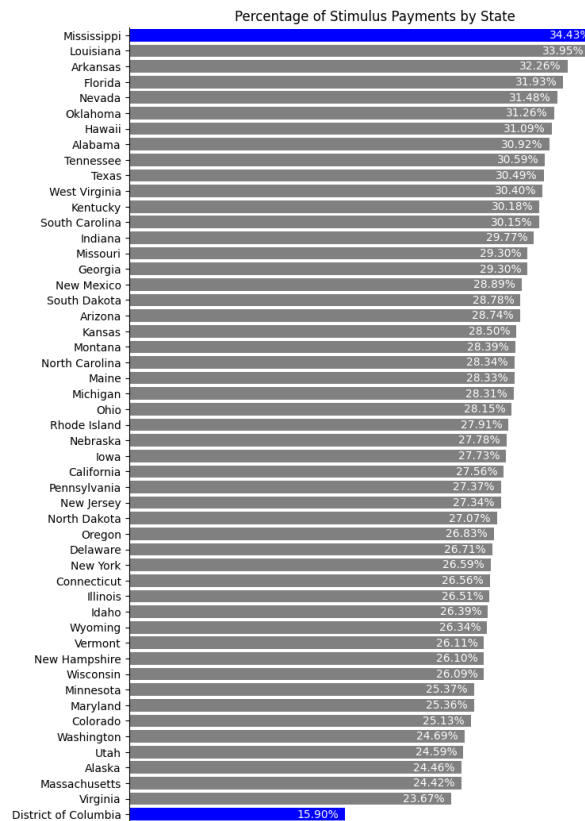


Figure 20: Percentage of Stimulus Payments by State

As shown, the **highest** percentage of Economic Impact (stimulus) payments among survey respondents is **Mississippi**.

7. How does household size correlate with the mental health impact of COVID-19 lockdown measures?

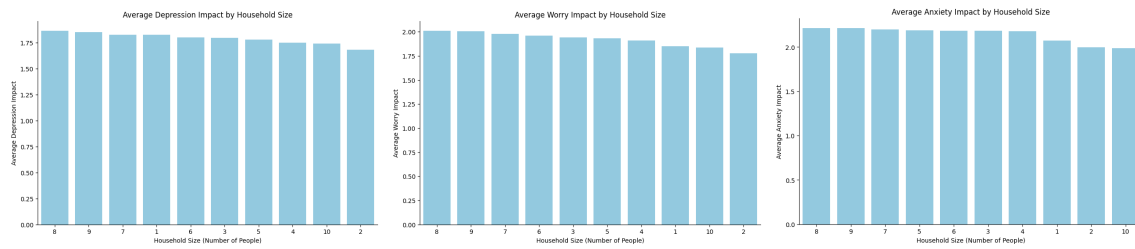


Figure 21: Mental Health correlation with household size

The depression, worry and anxiety impact is relatively **similar** across different household sizes, with a slight decrease in smaller households.

8. Is there a correlation between ICU admissions and the presence of underlying conditions in COVID-19 patients?

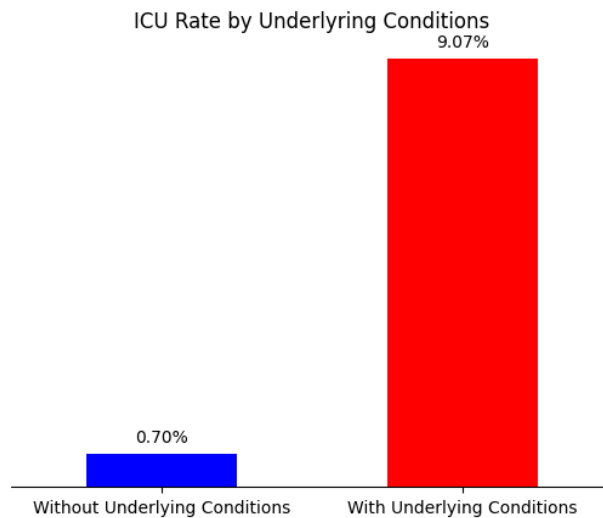


Figure 22: ICU Rate by Underlying Conditions

People with underlying conditions are the most people need ICU.

9. Is there a correlation between age and the time from earliest identification to symptom onset?

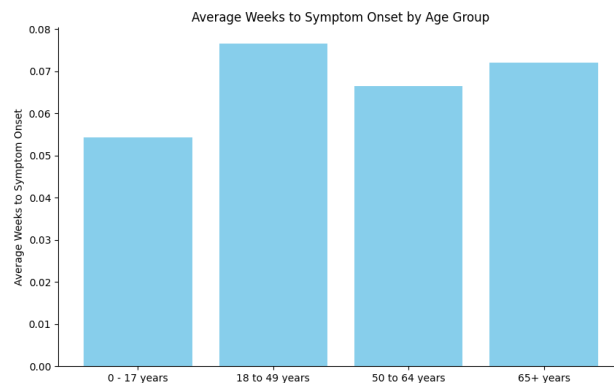


Figure 23: Average Weeks to Symptom Onset by Age Group

- 18 to 49 years old people experience a delayed onset of symptoms compared to others.
- 0-17 years old children experience onset of symptoms fastly compared to others.

10. How do the number of cases in different states compare over the same months?

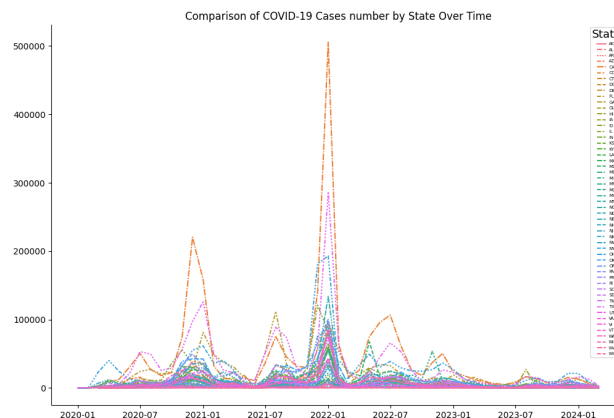


Figure 24: Comparison of COVID-19 Cases number by State Over Time

- The graph shows synchronized peaks across most states, indicating nationwide COVID-19 waves, particularly prominent in early 2021 and late 2021.
- By mid-2022, there's a visible decline in cases across all states, which may reflect the effectiveness of interventions like vaccinations.
- There's significant variation in case numbers between states, suggesting differing regional impacts and responses to the pandemic.

-
- Most of time, California had the most cases numbers. This is because of bias of California biggest number of people. Normalization is needed.

2.3 Hypothesis Tests

In this section, we explore a fundamental aspect of hypothesis testing. Hypothesis testing is a cornerstone of statistical analysis, providing a framework for making inferences about populations based on sample data. By establishing hypotheses, conducting tests, and interpreting results, researchers can draw conclusions and make informed decisions. This discussion serves to elucidate the key principles and methodologies underlying hypothesis testing, offering insights into its application and significance in statistical analysis. In the following subsections, we will explore two claims and subject them to various tests for validation.

2.3.1 First Claim

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

1. Stating the Test and Justifying the Choice:

We utilized the chi-square test for independence, which is appropriate for examining the relationship between two categorical variables: patient demographics (a combination of age group, sex, and race) and the outcome (death due to COVID-19).

2. Stating the Hypotheses:

- Null Hypothesis (H_0): There is no association between patient demographics (age group, sex, race) and death due to COVID-19.
- Alternative Hypothesis (H_1): There is an association between patient demographics (age group, sex, race) and death due to COVID-19.

The following figures represent the observed and expected frequencies of the combined patient demographics (age_group, sex, and race) versus the outcome of COVID-19 (death_yn) before removing the missing and unknown from the data.

observed Frequencies:						
	death_yn	Missing	No	Unknown	Yes	Total
combination						
0 - 17 years, Female, American Indian/Alaska Native		6134	1855	717	0	8706
0 - 17 years, Female, Asian		31348	9396	2416	0	43160
0 - 17 years, Female, Black		82361	55568	18021	0	155950
0 - 17 years, Female, Missing		100619	18536	1116	0	120271
0 - 17 years, Female, Multiple/Other		16843	5543	2246	0	24632
0 - 17 years, Female, Native Hawaiian/Other Pacific Islander		1262	48	24	0	1334
0 - 17 years, Female, Unknown		119005	42913	29244	0	191162
0 - 17 years, Female, White		307197	251254	102341	0	660792
0 - 17 years, Male, American Indian/Alaska Native		9031	1996	950	0	11977
0 - 17 years, Male, Asian		33591	10265	2629	0	46485
0 - 17 years, Male, Black		82011	55336	17615	0	154962
0 - 17 years, Male, Missing		103184	19208	1163	0	123555
0 - 17 years, Male, Multiple/Other		20525	7274	3069	0	30868
0 - 17 years, Male, Native Hawaiian/Other Pacific Islander		1525	69	87	0	1681
0 - 17 years, Male, Unknown		120995	45674	29853	0	196522
0 - 17 years, Male, White		313187	255197	103108	0	671492
0 - 17 years, Missing, American Indian/Alaska Native		21	0	0	0	21
0 - 17 years, Missing, Asian		278	1	0	0	279

Figure 25: Observed Frequencies

Expected Frequencies with Totals:						
	death_yn	Missing	No	Unknown	Yes	Total
combination						
0 - 17 years, Female, American Indian/Alaska Native		4.706120e+03	2.708715e+03	1.246272e+03	44.892511	17412.0
0 - 17 years, Female, Asian		2.333059e+04	1.342846e+04	6.178395e+03	222.554649	86320.0
0 - 17 years, Female, Black		8.430042e+04	4.852104e+04	2.232439e+04	804.156570	311900.0
0 - 17 years, Female, Missing		6.501376e+04	3.742016e+04	1.721691e+04	620.177716	240542.0
0 - 17 years, Female, Multiple/Other		1.331509e+04	7.663804e+03	3.526094e+03	127.014970	49264.0
0 - 17 years, Female, Native Hawaiian/Other Pacific Islander		7.211078e+02	4.150501e+02	1.909634e+02	6.878774	2668.0
0 - 17 years, Female, Unknown		1.033346e+05	5.947062e+04	2.736502e+04	985.727337	382324.0
0 - 17 years, Female, White		3.571981e+05	2.055935e+05	9.459299e+04	3407.375621	1321584.0
0 - 17 years, Male, American Indian/Alaska Native		6.474294e+03	3.726428e+03	1.714519e+03	61.759431	23954.0
0 - 17 years, Male, Asian		2.512796e+04	1.446297e+04	6.654371e+03	239.700020	92970.0
0 - 17 years, Male, Black		8.376634e+04	4.821364e+04	2.218296e+04	799.061945	309924.0
0 - 17 years, Male, Missing		6.678896e+04	3.844192e+04	1.768701e+04	637.111670	247110.0
0 - 17 years, Male, Multiple/Other		1.668602e+04	9.604023e+03	4.418783e+03	159.170920	61736.0
0 - 17 years, Male, Native Hawaiian/Other Pacific Islander		9.086823e+02	5.230129e+02	2.406367e+02	8.668081	3362.0
0 - 17 years, Male, Unknown		1.062320e+05	6.114429e+04	2.813231e+04	1013.366191	393044.0
0 - 17 years, Male, White		3.629821e+05	2.089227e+05	9.612471e+04	3462.550199	1342984.0
0 - 17 years, Missing, American Indian/Alaska Native		1.135177e+01	6.533772e+00	3.006170e+00	0.108287	42.0

Figure 26: Expected Frequencies

3. Conducting the Test and Reporting the Result:

We conducted the chi-square test for independence using the provided data. Here are the detailed results:

- Chi-squared value: 321949.4142723772
- P-value: 0.0
- Degrees of Freedom: 100

These results were obtained by analyzing the observed frequencies of the combination of age group, sex, and race against the death outcome due to COVID-19.

The following figures represent the observed and expected frequencies of the combined patient demographics (age_group, sex, and race) versus the outcome of COVID-19 (death_yn) after removing the missing and unknown from the data.

Observed Frequencies with Totals:

	death_yn	No	Yes	Total
combination				
0 - 17 years, Female, American Indian/Alaska Native		1855	0	1855
0 - 17 years, Female, Asian		9396	0	9396
0 - 17 years, Female, Black		55568	0	55568
0 - 17 years, Female, Multiple/Other		5543	0	5543
0 - 17 years, Female, Native Hawaiian/Other Pacific Islander		48	0	48
0 - 17 years, Female, White		251254	0	251254
0 - 17 years, Male, American Indian/Alaska Native		1996	0	1996
0 - 17 years, Male, Asian		10265	0	10265
0 - 17 years, Male, Black		55336	0	55336
0 - 17 years, Male, Multiple/Other		7274	0	7274
0 - 17 years, Male, Native Hawaiian/Other Pacific Islander		69	0	69
0 - 17 years, Male, White		255197	0	255197
18 to 49 years, Female, American Indian/Alaska Native		6131	0	6131
18 to 49 years, Female, Asian		39557	1	39558
18 to 49 years, Female, Black		206928	131	207059
18 to 49 years, Female, Multiple/Other		15557	0	15557
18 to 49 years, Female, Native Hawaiian/Other Pacific Islander		407	0	407
18 to 49 years, Female, White		890300	225	890525
18 to 49 years, Male, American Indian/Alaska Native		5491	6	5497
18 to 49 years, Male, Asian		32843	11	32854
18 to 49 years, Male, Black		130265	144	130409
18 to 49 years, Male, Multiple/Other		14417	0	14417

Figure 27: Observed Frequencies without "missing" & "unknown" data

Expected Frequencies with Totals:			
	death_yn	No	Yes
combination			Total
0 - 17 years, Female, American Indian/Alaska Native	1.821046e+03	33.953643	3710.0
0 - 17 years, Female, Asian	9.224017e+03	171.982980	18792.0
0 - 17 years, Female, Black	5.455089e+04	1017.108371	111136.0
0 - 17 years, Female, Multiple/Other	5.441542e+03	101.458244	11086.0
0 - 17 years, Female, Native Hawaiian/Other Pacific Islander	4.712142e+01	0.878585	96.0
0 - 17 years, Female, White	2.466551e+05	4598.915682	502508.0
0 - 17 years, Male, American Indian/Alaska Native	1.959466e+03	36.534486	3992.0
0 - 17 years, Male, Asian	1.007711e+04	187.889027	20530.0
0 - 17 years, Male, Black	5.432314e+04	1012.861878	110672.0
0 - 17 years, Male, Multiple/Other	7.140858e+03	133.142209	14548.0
0 - 17 years, Male, Native Hawaiian/Other Pacific Islander	6.773703e+01	1.262966	138.0
0 - 17 years, Male, White	2.505259e+05	4671.087765	510394.0
18 to 49 years, Female, American Indian/Alaska Native	6.018779e+03	112.220908	12262.0
18 to 49 years, Female, Asian	3.883394e+04	724.063723	79116.0
18 to 49 years, Female, Black	2.032690e+05	3789.977004	414118.0
18 to 49 years, Female, Multiple/Other	1.527225e+04	284.753004	31114.0
18 to 49 years, Female, Native Hawaiian/Other Pacific Islander	3.995503e+02	7.449667	814.0
18 to 49 years, Female, White	8.742250e+05	16300.036567	1781050.0
18 to 49 years, Male, American Indian/Alaska Native	5.396384e+03	100.616267	10994.0
18 to 49 years, Male, Asian	3.225265e+04	601.354708	65708.0
18 to 49 years, Male, Black	1.280220e+05	2386.986855	260818.0

Figure 28: Expected Frequencies without "missing" & "unknown" data

4. Making a Conclusion as to the Validity of the Claim:

- **Conclusion:** Since the p-value is 0.0, significantly less than the significance level of 0.05, we reject the null hypothesis.
- **Interpretation:** The null hypothesis stated that there is no association between patient demographics (age group, sex, race) and death due to COVID-19. By rejecting the null hypothesis, we conclude that there is a significant association between patient demographics (age group, sex, race) and the probability of death due to COVID-19. This supports the claim of a strong association between these variables. Thus, based on the given data and the chi-square test results, we find the claim to be valid.

```
Chi-squared Test Results:
Chi-squared value: 321949.4142723772
P-value: 0.0
Degrees of Freedom: 100

Conclusion:
There is a significant association between the combination of age_group, sex, race and death_yn (reject the null hypothesis).
```

Figure 29: Chi-Square Output

2.3.2 Second Claim

Claim: "There is a significant difference in the mean age between COVID-19 patients who were admitted to the ICU and those who were not admitted to the ICU."

1. Stating the Test and Justifying the Choice:

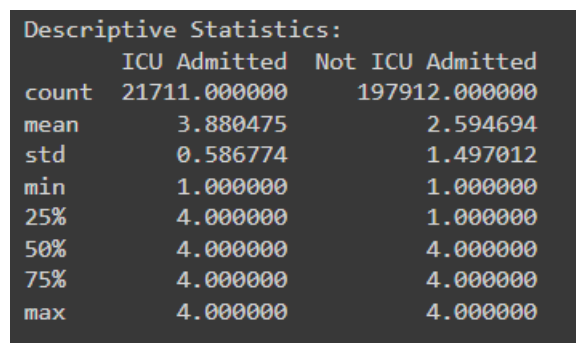
We employed the independent two-sample t-test, as it is suitable for comparing the means of two independent groups (COVID-19 patients admitted to the ICU and those not admitted to the ICU) to determine if there is a significant difference in their mean ages.

2. Stating the Hypotheses:

- Null Hypothesis (H_0): There is no significant difference in the mean age between COVID-19 patients who were admitted to the ICU and those who were not admitted to the ICU.
- Alternative Hypothesis (H_1): There is a significant difference in the mean age between COVID-19 patients who were admitted to the ICU and those who were not admitted to the ICU.

3. Conducting the Test and Reporting the Result:

The descriptive statistics table reveal that patients admitted to the ICU tend to be older (mean age group value of 3.880) compared to those not admitted to the ICU (mean age group value of 2.595). The data indicates a clear trend that older patients have a higher likelihood of ICU admission, which aligns with the claim that there is a significant difference in the mean age between these two groups.

A screenshot of a statistical software output window titled "Descriptive Statistics:". It contains a table with two columns: "ICU Admitted" and "Not ICU Admitted". The rows represent various statistical measures: count, mean, std, min, 25%, 50%, 75%, and max. The "ICU Admitted" column shows values for a group of 21,711 patients, while the "Not ICU Admitted" column shows values for a group of 19,791 patients. The mean age for the ICU group is 3.880475, and for the non-ICU group, it is 2.594694.

	ICU Admitted	Not ICU Admitted
count	21711.000000	197912.000000
mean	3.880475	2.594694
std	0.586774	1.497012
min	1.000000	1.000000
25%	4.000000	1.000000
50%	4.000000	4.000000
75%	4.000000	4.000000
max	4.000000	4.000000

Figure 30: Descriptive Statistics Output from the t-test

4. Making a Conclusion as to the Validity of the Claim:


```
Custom t-test results:  
T-statistic: 246.619, P-value: 0.000  
  
Scipy t-test results:  
T-statistic: 125.502, P-value: 0.000  
  
Conclusion:  
There is a significant difference in the mean age of patients who were admitted to the ICU and those who were not admitted to the ICU (reject the null hypothesis).
```

Figure 31: t-test Output

- **Interpretation:** Both the custom and SciPy t-tests yielded extremely low p-values (0.000), which are far below the significance level of 0.05. This indicates that the observed difference in mean ages between the two groups is highly significant.
- **Conclusion:** Since the p-value is less than the significance level of 0.05, we reject the null hypothesis. There is a significant difference in the mean age of COVID-19 patients who were admitted to the ICU and those who were not admitted to the ICU. This supports the claim that the mean age significantly differs between these two groups.

3 Regression Analysis

In this section, we undertake a comprehensive regression analysis using the COVID Case Surveillance dataset to predict the total percent (or proportion) of deaths out of all COVID cases in a given month. Our analysis aims to understand the impact of various demographic and clinical factors on the mortality rate of COVID-19. Specifically, we focus on the following predictors:

- **Gender Distribution:** The proportion or percentage of female and male cases over the month.
- **Age Distribution:** The proportion or percentage of each age group among the cases over the month.
- **ICU Admissions:** The proportion or percentage of cases that end up in the ICU over the month.
- **Hospitalizations:** The proportion or percentage of cases that result in hospitalization over the month.

By fitting a regression model to these predictors, we aim to:

1. **Estimate Model Coefficients and Significance:** Report the coefficients and p-values to assess the statistical significance of each predictor.

-
2. **Evaluate Predictor Importance:** Determine which variables are good predictors of variability in the target (the proportion of deaths) and which are not.
 3. **Assess Predictor Correlation:** Identify any correlations between the predictors that might affect the model's performance.
 4. **Enhance Model Fit and Interpretability:** Experiment with various techniques such as adding or removing the intercept, introducing higher-order terms, and removing outliers to improve the model.

3.1 Preprocessing and Proportion Calculation

In this section, we outline the preprocessing steps and proportion calculation performed on the COVID Case Surveillance dataset for regression analysis. The objective is to prepare the data for further analysis by extracting key insights regarding mortality rates and their relationship with demographic and clinical factors.

3.1.1 Preprocessing Steps

1. **Data Selection:** We selected relevant columns from the dataset including `case_month`, `hosp_yn`, `icu_yn`, `age_group`, `sex`, and `death_yn` for analysis.
2. **Data Cleaning:** We removed unknown, NaN, and missing values from the selected columns to ensure data integrity and accuracy in subsequent analysis.

3.1.2 Proportion Calculation

We calculated proportions for each selected column within the context of `case_month` to understand the distribution of demographic and clinical factors over time. Specifically, we calculated the proportion of:

- **Age Groups:** The percentage distribution of age groups among COVID cases.
- **Gender Distribution:** The percentage distribution of males and females among COVID cases.
- **Hospitalization Status:** The percentage of cases resulting in hospitalization.
- **ICU Admission Status:** The percentage of cases admitted to the ICU.
- **Mortality Status:** The percentage of cases resulting in death.

These proportions provide valuable insights into the demographic and clinical factors associated with COVID-19 mortality rates. They serve as foundational elements for subsequent regression analysis, enabling us to identify predictors and assess their impact on mortality rates.

The calculated proportions are aggregated into a single DataFrame for further analysis and modeling. This consolidated dataset facilitates comprehensive regression analysis, helping us derive actionable insights to inform strategic decision-making and public health interventions.

3.2 Trial 1 Summary

In this trial, we compared 4 cases for the input columns to the regression model: with intercept, without intercept, with introducing higher order features, and with removing outliers.

3.2.1 Results

Case	MAE	MSE	R-squared
With Intercept	2.0979	6.0857	0.975
Higher Order Features	2.6485	10.9350	0.977
Without Intercept	2.0979	6.0857	0.975
Removing Outliers	1.3695	3.3171	0.716

Table 1: Regression Model Performance Metrics

3.2.2 Interpretation

With and Without Intercept:

The models with and without intercepts exhibit similar performance, as indicated by identical MAE, MSE, and R-squared values. This suggests that the intercept term does not significantly impact the model's predictive capability in this case.

Higher Order Features:

Although the R-squared value slightly increases, indicating a better fit to the training data, the increase in both MAE and MSE suggests potential overfitting.

Removing Outliers:

After removing outliers, there is a noticeable improvement in model performance. The MAE and MSE decrease significantly, indicating a better fit to the data. However, the R-squared value decreases, suggesting that the model explains less of the variance in the target variable after removing outliers.

3.3 Trial 2 Summary

In this trial, we compared the same 4 cases as the previous trial but excluded males from the analysis, as they were dependent on females. Additionally, we removed the age group variable encompassing ages from 0 to 17 to mitigate potential overfitting issues.

3.3.1 Results

Case	MAE	MSE	R-squared
With Intercept	1.9647	5.9400	0.974
Higher Order Features	2.9555	15.7943	0.968
Without Intercept	2.0634	6.3002	0.977
Removing Outliers	1.3588	3.2345	0.715

Table 2: Regression Model Performance Metrics (Trial 2)

3.3.2 Interpretation

With and Without Intercept:

The models with and without intercepts exhibit similar performance, as indicated by identical MAE, MSE, and R-squared values. This suggests that the intercept term does not significantly impact the model's predictive capability in this case.

Higher Order Features:

Although the R-squared value slightly increases, indicating a better fit to the training data, the increase in both MAE and MSE suggests potential overfitting.

Removing Outliers:

After removing outliers, there is a noticeable improvement in model performance. The MAE and MSE decrease significantly, indicating a better fit to the data. However, the R-squared value decreases, suggesting that the model explains less of the variance in the target variable after removing outliers.

3.4 Comparison Between the Two Trials

The current trial generally outperforms the previous trial in terms of MAE and MSE across most model configurations, indicating improved accuracy and precision in predicting the target variable. However, the previous trial achieved higher R-squared values in the "With Intercept Case" and "Higher Order Features" configurations, suggesting better overall fit to the data in those cases.

4 Conclusion

4.1 Key Insights

4.1.1 Exploratory Analysis

We conducted a comprehensive exploratory analysis of COVID-19 data across the U.S., identifying trends in hospitalizations, deaths, and demographic impacts. This analysis provided valuable insights into the spread and impact of the virus.

4.1.2 Specific Questions

Our investigation addressed specific questions regarding risk factors, affected demographics, and economic impacts of COVID-19. Through data analysis, we gained insights that can inform public health policies and interventions.

4.1.3 Hypothesis Testing

We confirmed strong associations between patient demographics and COVID-19 outcomes through hypothesis testing. These findings underscore the importance of considering demographic factors in pandemic response strategies.

4.1.4 Regression Analysis

Using regression analysis, we developed a predictive model for COVID-19 death rates based on age, gender, and ICU admissions. This model serves as a valuable tool for understanding mortality risk factors and informing healthcare resource allocation.

4.2 Future Research

4.2.1 Deeper Analysis

Future research should delve deeper into sub-demographics to uncover nuanced insights into COVID-19's impact on vulnerable populations.

4.2.2 Longitudinal Studies

Longitudinal studies are essential to track the long-term effects of the pandemic on health outcomes, socioeconomic factors, and healthcare systems.

4.2.3 Data Integration

Integrating COVID-19 data with other health datasets can provide comprehensive insights into disease patterns, comorbidities, and healthcare disparities.