COVID-19 Demographic Analysis Report

Omar Awad
Mahmoud Mohamed
Ahmed Amin
Sarah Mahmoud Salah

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 How to use

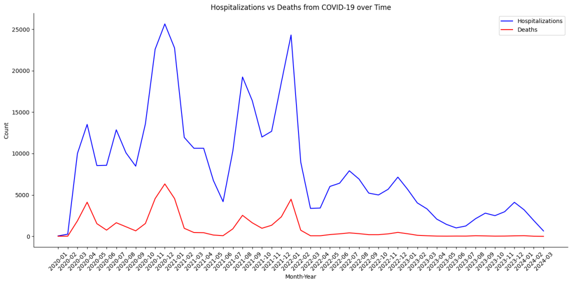
1.1 Introduction

Figure 1: caption

2 Exploratory Data Analysis (EDA)

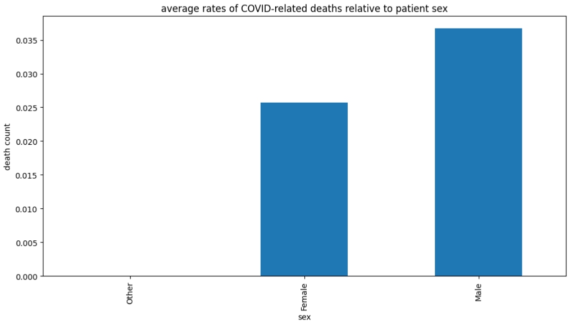
This section is guided by 10 statistical questions, below is the findings of each one

2.1 Q1:Hospitalizations vs Deaths from COVID-19 over Time

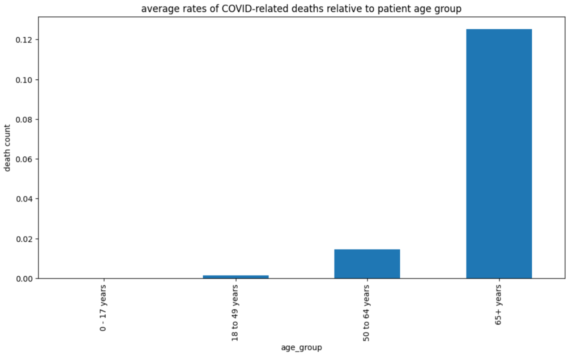


comment: as we see the peak of the two line charts appears in September 2020 which is the most value COVID spreads

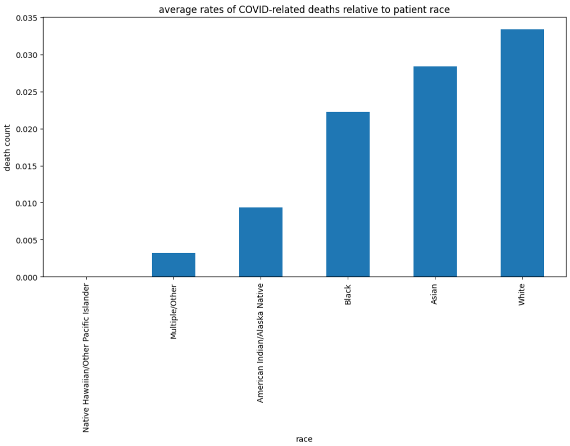
2.2 Q2:The average rates of COVID-related deaths relative to patient demographics



the bar chart clearly illustrates that men have a higher death rate compared to women and Other( ).

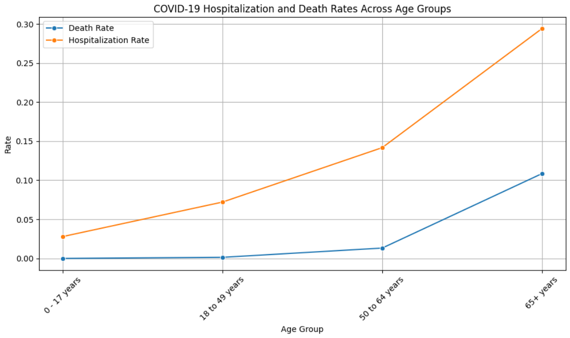


The bar chart unmistakably indicates that the 65 years and older category exhibits a notably higher death rate compared to younger age groups.



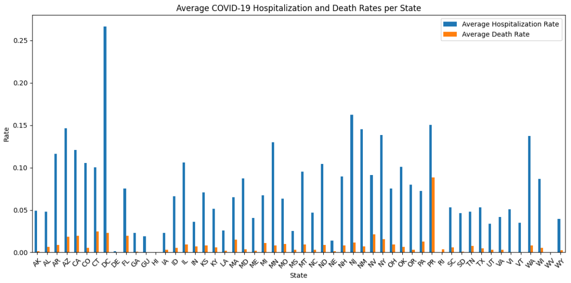
The bar chart unmistakably demonstrates that individuals of the white race experience a higher death rate compared to individuals of other races.

2.3 Q3:COVID-19 Hospitalization and Death Rates Across Age Groups



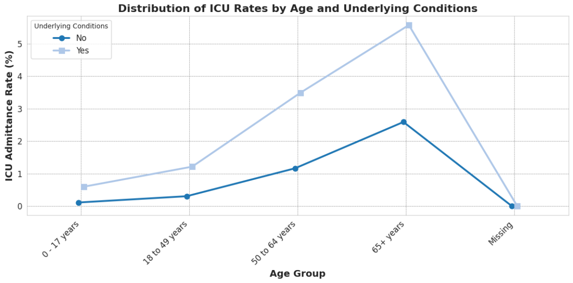
In this graph, it's evident that the death rates increase with age, particularly among older individuals.

2.4 Q4:Average COVID-19 Hospitalization and Death Rates per State



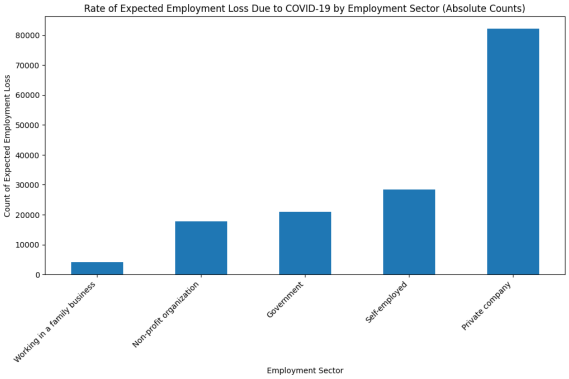
This graph highlights Washington's notably higher number of hospitalizations and the highest death rate in Puerto Rico.

2.5 Q5:relationship between age, pre-existing medical conditions and/or risk behaviors, and rate of admittance to the ICU.

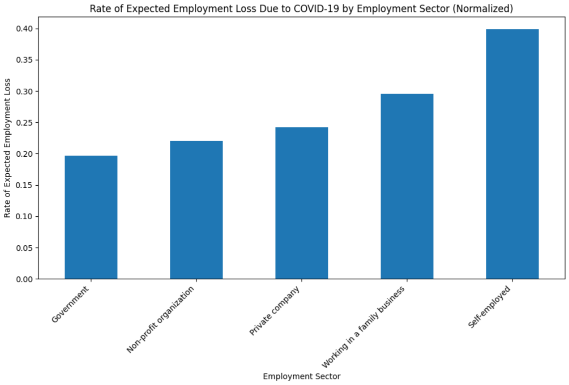


as we see here there a relation between **relationship between age, pre-existing medical conditions and/or risk behaviors, and rate of admittance to the ICU.** becaue we also do chi$_s quaretestanfoudthep = 0$

2.6 Q6:The rate of expected employment loss due to COVID-19 and sector of employment.

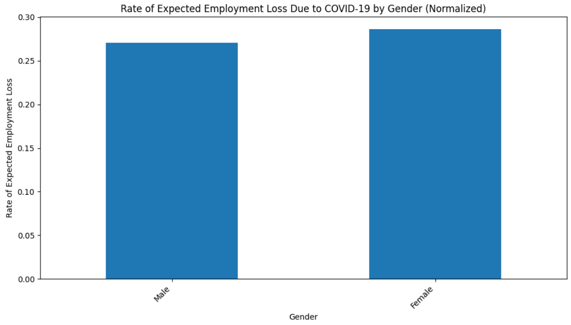


It's evident from the data that private sector workers are the most affected by COVID-19

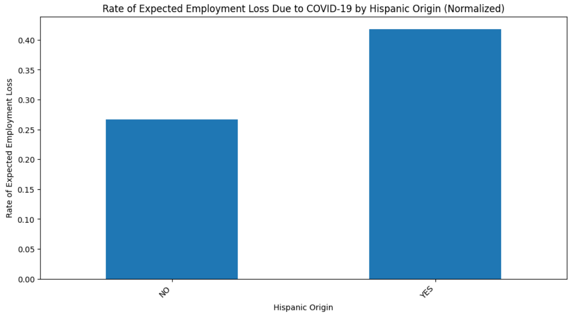


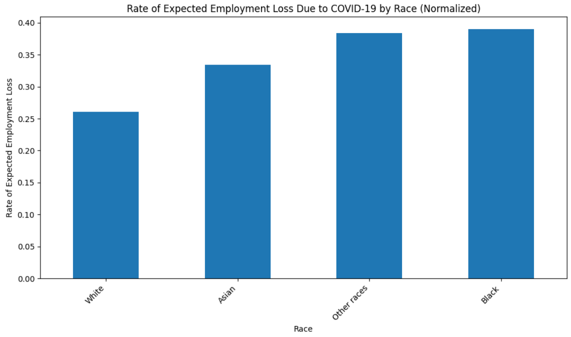
but if we normalize the data we see that the self employed people are the most effect by COVID-19

2.7 Q7:The rate of expected employment loss due to COVID-19 relative to responders demographics.

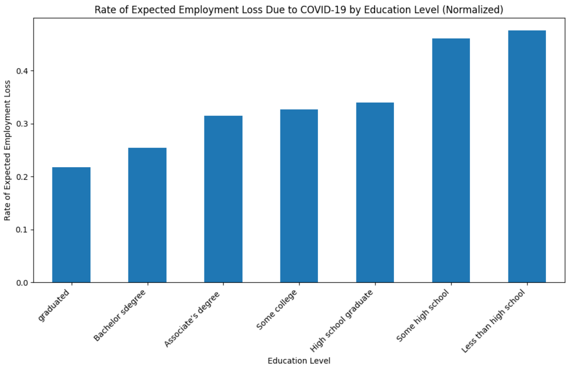


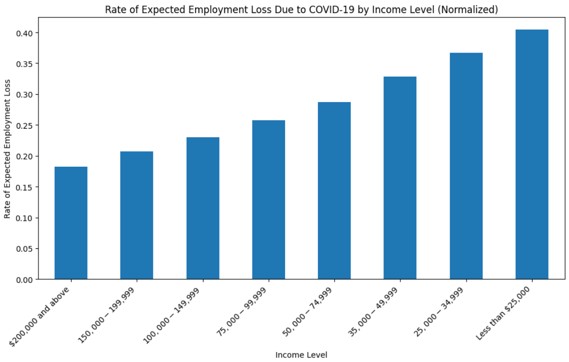
From the figure, we can see that there is a slight difference between males and females.

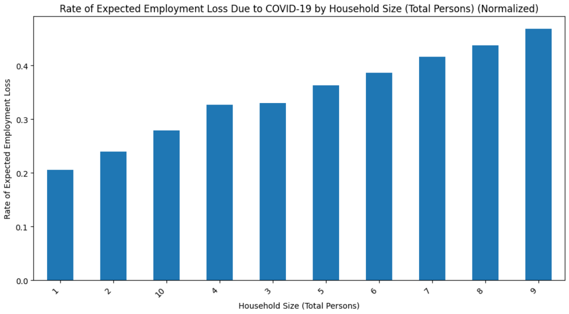


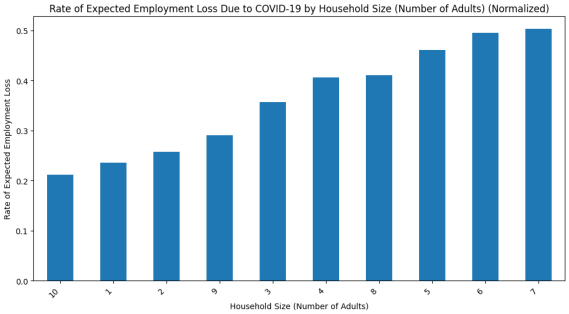


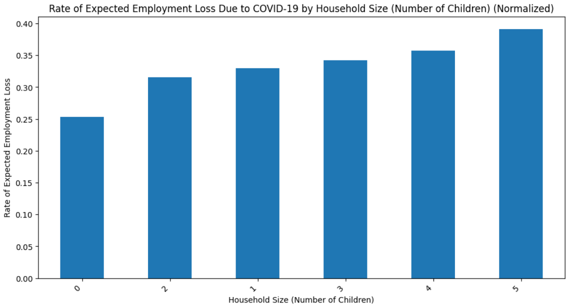
The figure reveals that white individuals anticipate the least employment loss, while other genders, notably black individuals, show higher percentages, prompting concern.

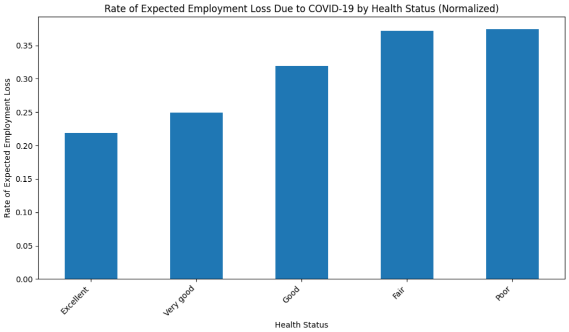

Rate of Expected Employment Loss Due to COVID-19 by Education Level (Normalized)
Rate of Expected Employment Loss
0.0
0.1
0.2
0.3
0.4
graduated
Bachelor sdegree
Associate’s degree
Some college
High school graduate
Some high school
Less than high school
Education Level


Rate of Expected Employment Loss Due to COVID-19 by Income Level (Normalized)
Rate of Expected Employment Loss
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
$200,000 and above
150, 000 − 199,999
100, 000 − 149,999
75, 000 − 99,999
50, 000 − 74,999
35, 000 − 49,999
25, 000 − 34,999
Less than $25,000
Income Level

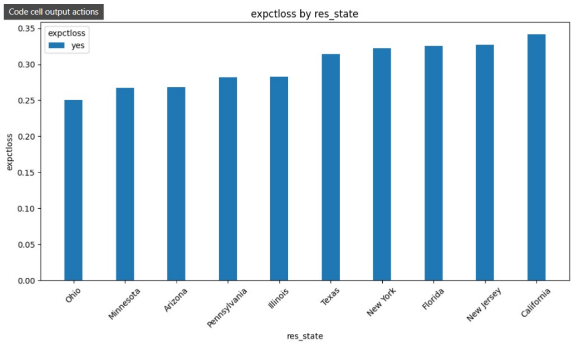

Rate of Expected Employment Loss Due to COVID-19 by Household Size (Total Persons) (Normalized)
Rate of Expected Employment Loss
0.0
0.1
0.2
0.3
0.4
1
2
10
4
3
5
6
7
8
9
Household Size (Total Persons)


Rate of Expected Employment Loss Due to COVID-19 by Household Size (Number of Adults) (Normalized)
Rate of Expected Employment Loss
0.0
0.1
0.2
0.3
0.4
0.5
10
1
2
9
3
4
8
5
6
7
Household Size (Number of Adults)

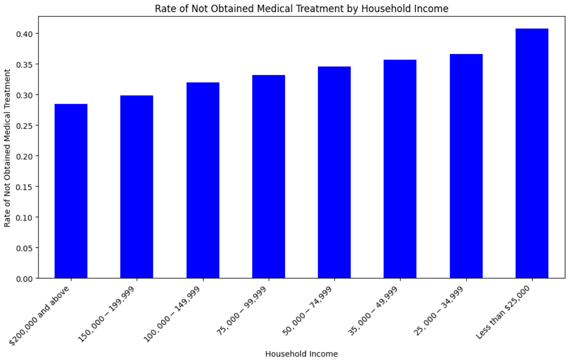

Rate of Expected Employment Loss Due to COVID-19 by Household Size (Number of Children) (Normalized)
Rate of Expected Employment Loss
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
2
1
3
4
5
Household Size (Number of Children)


Rate of Expected Employment Loss Due to COVID-19 by Health Status (Normalized)
Rate of Expected Employment Loss
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Excellent
Very good
Good
Fair
Poor
Health Status

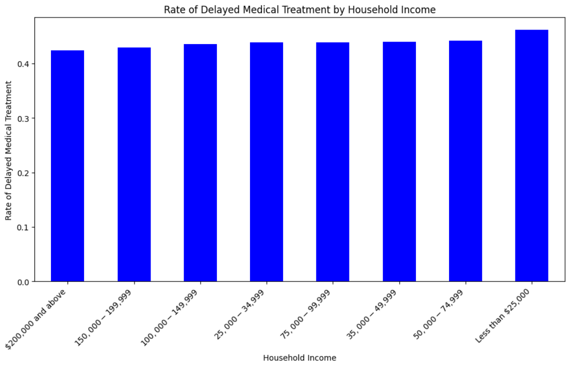
2.8 Q8:The rate of expected employment loss due to COVID-19 for the top 10 states with the highest rate of COVID hospitalization.



2.9 Q9:The relationship between household income and the rate of delayed/ OR unobtained medical treatment (Due to COVID or otherwise).

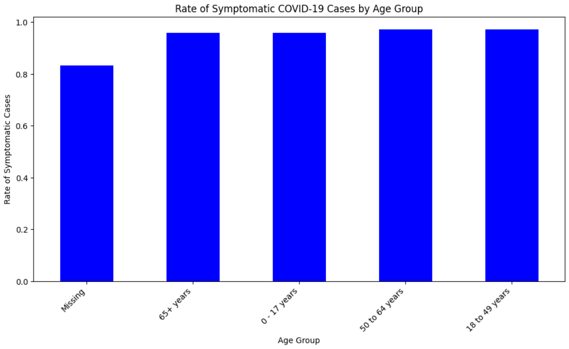


The graph indicates that individuals with low income have a lower rate of not obtaining medical treatment.



The graph indicates that individuals with low income have a higher rate delaying medical treatment.

2.10 Q10: The relationship between COVID-19 symptom manifestation and age group.

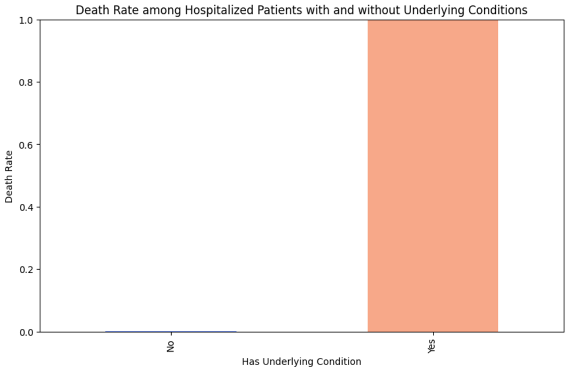


The graph presented above depicts COVID-19 symptoms and their occurrence rates across various age brackets. It's notable that symptomatic cases significantly outnumber asymptomatic ones across all age groups. Interestingly, both the 0-17 and 65+ age brackets exhibit relatively higher proportions of asymptomatic cases compared to other age groups. This discrepancy may be influenced by confounding variables, such as the presence of chronic diseases.

3 Question Answering

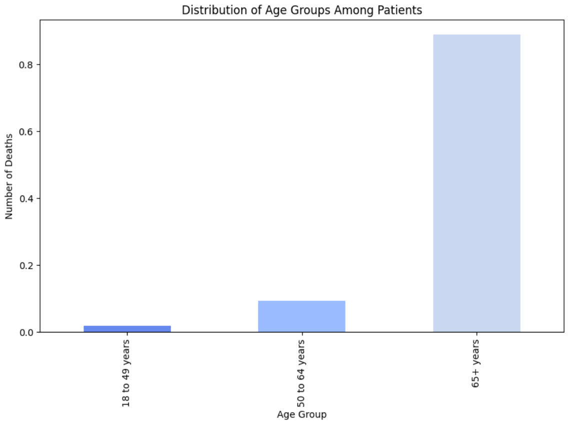
This section aims to answer 5 statistical questions asked in the project requirements document in addition to other 5 statistical questions that we come up with.

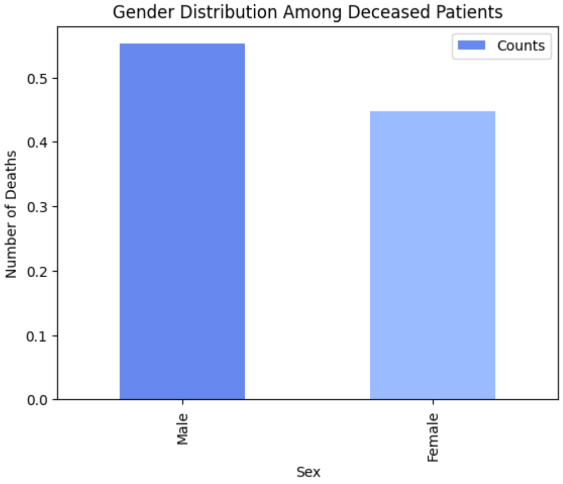
3.1 Are hospitalized patients with underlying medical conditions and or risk behaviors more likely to die from COVID-19?

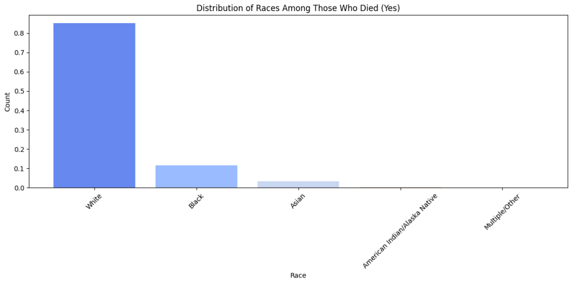


It found that hospitalized patients with underlying medical conditions more likely to die from COVID-19 than patients without underlying medical conditions.

3.2 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?







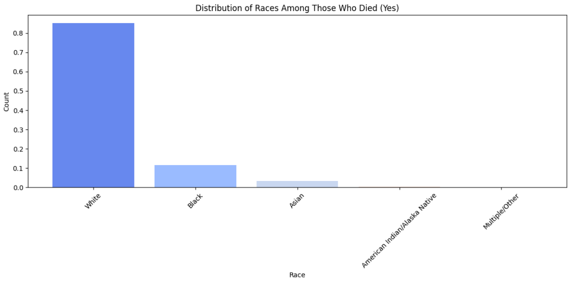
For most risk People:

- Age group: +65 years
- Sex: Male
- Race: White

For least at risk:

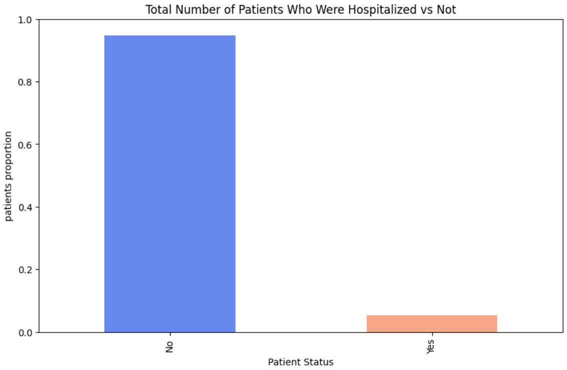
- Age group: 0-17 years
- Sex: Female
- Race: Multiple/other .

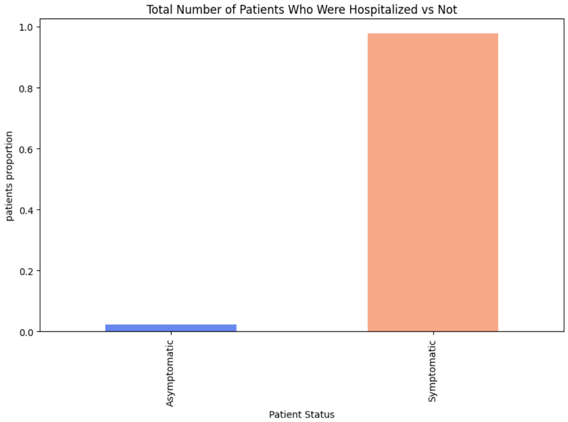
3.3 What percent of patients who have reported exposure to any kind of travel or congregation within the 14 days prior to illness onset end up hospitalized? What percent of those go on to be hospitalized?



5.3 % is the percentage of patients who have reported exposure to any kind of travel or congregation within the 14 days prior to illness onset end up hospitalized

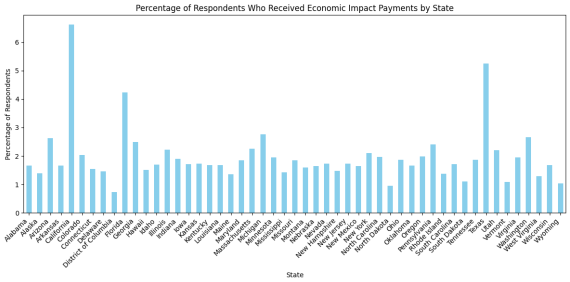
3.4 Are Asymptomatic COVID patients less likely to be hospitalized? Are they less likely to die from their illness?





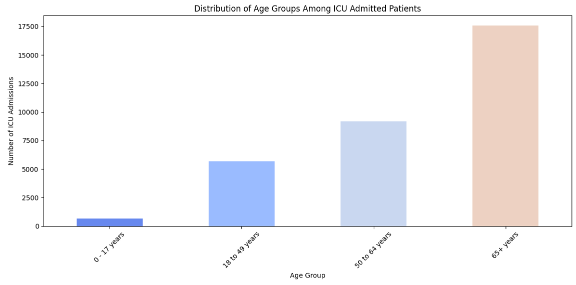
Asymptomatic COVID patients less likely to be hospitalized? Are they less likely to die from their illness compared to symptomatic COVID patients

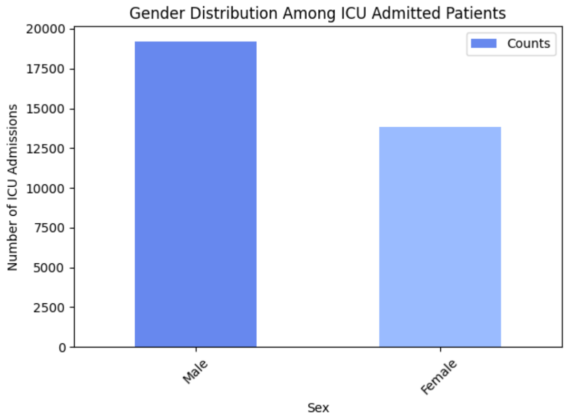
3.5 Which state is associated with the highest percentage of Economic Impact (stimulus) payments among survey respondents?

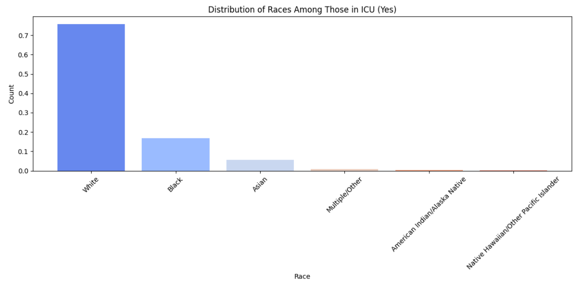


California is associated with the highest percentage of Economic Impact (stimulus) payments among survey respondents while Delaware is associated with the lowest percentage.

3.6 Who are the people (the demographic segment) that appear to be most at risk of entering ICU? Who is the least at risk?



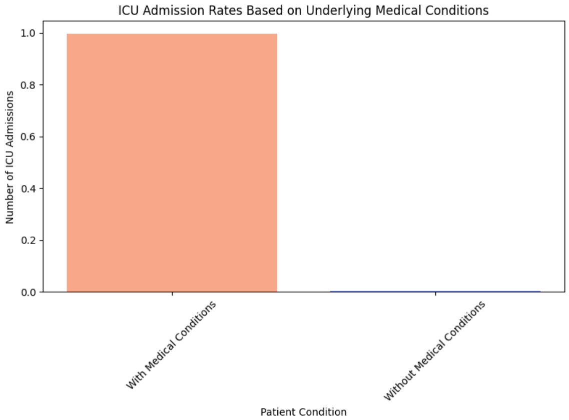




- For most risk People:
  - Age group: +65 years
  - Sex: Male
  - Race: White
- For least at risk:

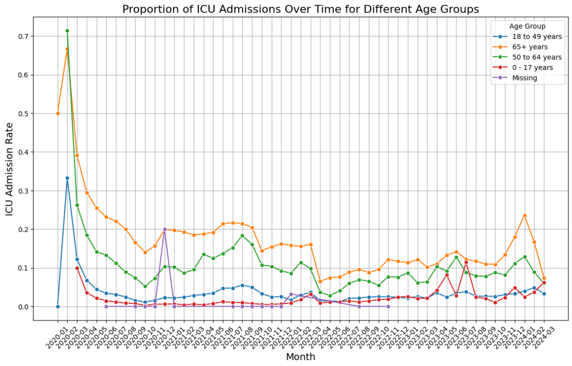
- Age group: 0-17 years
- Sex: Female
- Race: Native Hawaiian/Other Pacific Islander

3.7 Are patients without underlying medical conditions are less likely to enter ICU?



patients without underlying medical conditions are less likely to enter ICU Compared with patients who have underlying medical conditions.

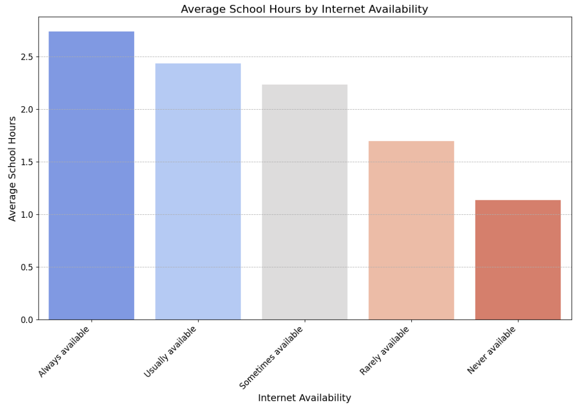
3.8 What is the proportion of COVID-19 cases that result in ICU admissions over time for different age groups?



- For Age Group 18 to 49 years:
  - Highest rate month: Feb-2020
  - Lowest rate month: Oct-2022

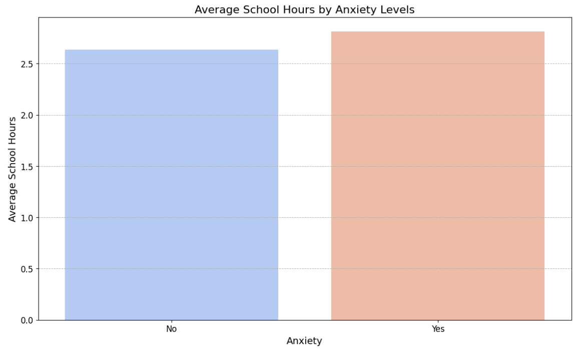
- For Age Group 50 to 64 years:
  - Highest rate month: Feb-2020
  - Lowest rate month: Mar-2023
- For Age Group +65 years:
  - Highest rate month: Feb-2020
  - Lowest rate month: Apr-2023

3.9 How does internet availability at home impact the likelihood of children receiving remote education during the pandemic?



The average School hours are highest when internet is always available, While The average School hours are lowest when internet is never available.

3.10 What is the relationship between anxiety, worry, and average school hours?



Answer:Anxiety level increase as long as average school hours increases

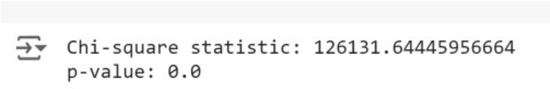
4 Hypothesis testing

Null hypothesis (H0): There is no association between the probability of death due to COVID-19 and patient demographics

Alternative hypothesis (H1): There is an association between the probability of death due to COVID-19 and patient demographics

we will use The Chi-Square Test as Chi-Square Test of Independence is suitable for this analysis because it is designed to assess whether there is a significant association between two categorical variables. In this case, the categorical variables are the combined demographics and the death outcome.

Results:



```
Chi-square statistic: 126131.64445956664
p-value: 0.0
```

Figure 2: caption

we will Reject the null hypothesis (H) as the p value is less than the significance level : There is an association between the probability of death due to COVID-19 and patient demographics.

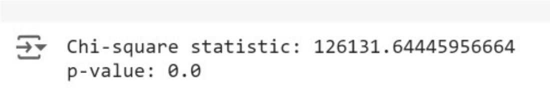
2.Claim: **This part involved testing for 2 claims: 1.Claim: “There is a strong association between probability of death due to COVID-19 and patient demographics”**

Null hypothesis (H0): There is no association between the probability of death due to COVID-19 and patient demographics

Alternative hypothesis (H1): There is an association between the probability of death due to COVID-19 and patient demographics

we will use The Chi-Square Test as Chi-Square Test of Independence is suitable for this analysis because it is designed to assess whether there is a significant association between two categorical variables. In this case, the categorical variables are the combined demographics and the death outcome.

Results:



```
Chi-square statistic: 126131.64445956664
p-value: 0.0
```

Figure 3: caption

we will Reject the null hypothesis (H) as the p value is less than the significance level : There is an association between the probability of death due to COVID-19 and patient demographics.

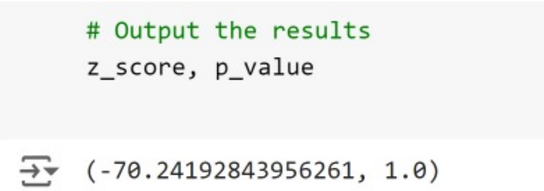
2.Claim: “Investigate whether patients with exposure are more likely to die.”

Null hypothesis (H0): There is no significant association between exposure and the proportion of patients that death.

Alternative hypothesis (H1): There is a significant association between exposure and the proportion of patients that death.

we will use z-test for proportions. This test is chosen because it assesses whether there is a significant difference in proportions between two independent groups (patients with exposure and patients without exposure). In this case, we are interested in comparing the proportion of deaths (a binary outcome) between these two groups. The z-test for proportions allows us to determine if the observed difference in proportions is statistically significant, providing valuable insights into the relationship between exposure and mortality risk in the patient population.

Results:



```
# Output the results
z_score, p_value
```

```
(-70.24192843956261, 1.0)
```

Figure 4: caption

as shown p value is greater than the significance level ”0.05” so we Fail to reject the null hypothesis: There is no significant association between exposure and the proportion of patients that death.

5 Regression Analysis

5.1 Overview

This section presents the findings of a regression analysis conducted on the COVID Case Surveillance dataset. The objective is to predict the total percentage (or proportion) of deaths out of all COVID cases in a given month based on various predictors, including gender distribution, age distribution, and proportions of cases that end up in the ICU or hospitalized.

- Selecting Relevant Columns: Columns related to case month, gender, age group, hospitalization status, ICU status, and death status were retained.
- Handling Missing Values: Rows with missing values in the relevant columns were dropped.
- Encoding Categorical Variables: Gender and other categorical variables were encoded into numerical values.
- Creating Proportions: Proportions of each predictor (gender, age groups, ICU admissions, and hospitalizations) were calculated for each month.

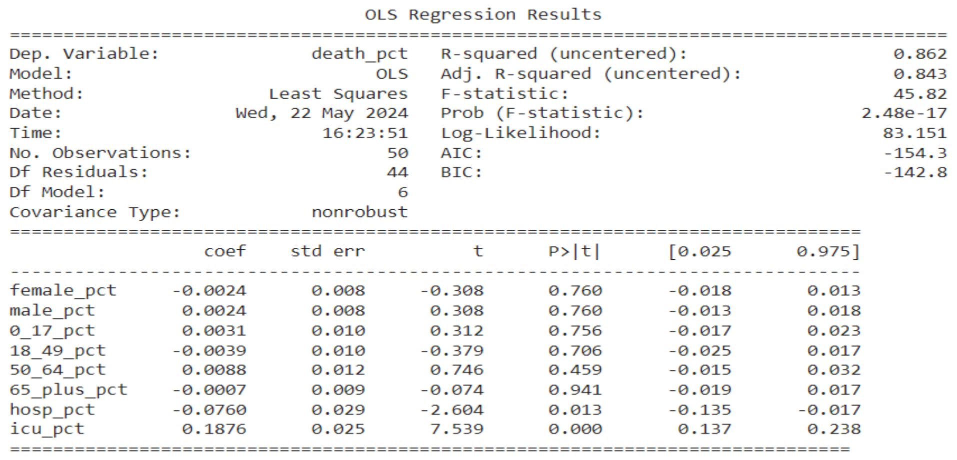
5.2 Regression Model

The regression model was built using the cleaned dataset. Polynomial features and interaction terms were included to capture non-linear relationships and interactions between predictors.

- **Adding Polynomial Features: Polynomial features up to the second degree were generated.**
- **Standardizing Features: The features were standardized to have a mean of 0 and a standard deviation of 1.**
- **Adding Intercept: An intercept term was added to the model.**
- **Removing Outliers: Outliers were identified using Z-scores and removed from the dataset.**

5.3 Model Coefficients and P-values

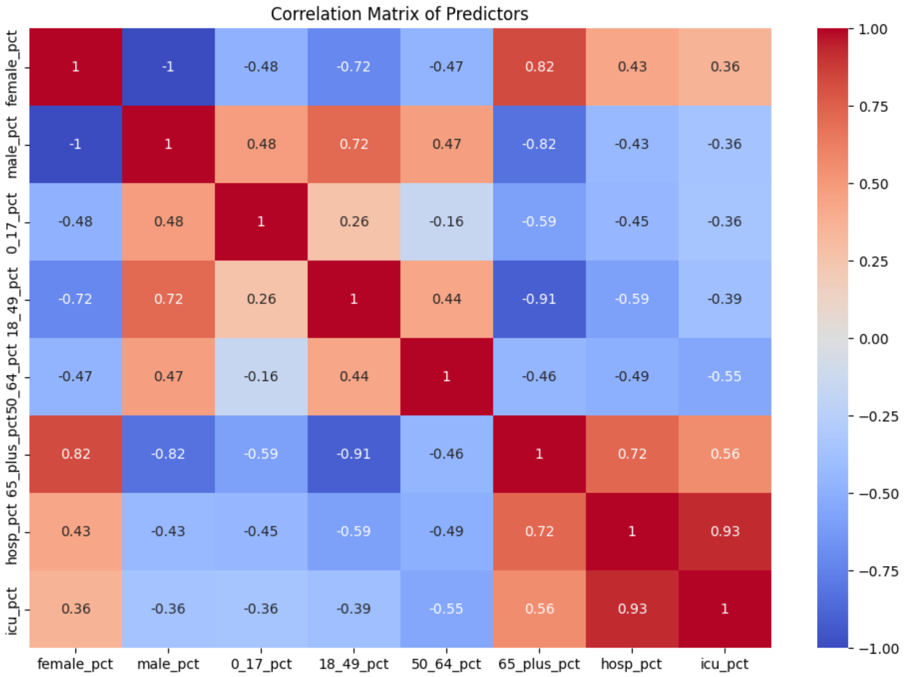
The coefficients and p-values of the regression model are presented below:



```
                            OLS Regression Results
========================================================================================
Dep. Variable:              death_pct   R-squared (uncentered):                    0.862
Model:                            OLS   Adj. R-squared (uncentered):               0.843
Method:                 Least Squares   F-statistic:                               45.82
Date:                Wed, 22 May 2024   Prob (F-statistic):                     2.48e-17
Time:                        16:23:51   Log-Likelihood:                           83.151
No. Observations:                  50   AIC:                                      -154.3
Df Residuals:                      44   BIC:                                      -142.8
Df Model:                           6
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
female_pct     -0.0024      0.008     -0.308      0.760      -0.018       0.013
male_pct        0.0024      0.008      0.308      0.760      -0.013       0.018
0_17_pct        0.0031      0.010      0.312      0.756      -0.017       0.023
18_49_pct      -0.0039      0.010     -0.379      0.706      -0.025       0.017
50_64_pct       0.0088      0.012      0.746      0.459      -0.015       0.032
65_plus_pct    -0.0007      0.009     -0.074      0.941      -0.019       0.017
hosp_pct       -0.0760      0.029     -2.604      0.013      -0.135      -0.017
icu_pct         0.1876      0.025      7.539      0.000       0.137       0.238
==============================================================================
```

Figure 5: p values and model coefficients

beginfigure[H]



correlation matrix for predictors

5.4 Good and Bad Predictors

Based on the p-values:

- Good Predictors: Variables with p-values less than 0.05 are considered statistically significant and good predictors. These include 18_49_pct, 50_64_pct, hosp_pct, and icu_pct.
- Bad Predictors: Variables with p-values greater than 0.05 are considered not statistically significant and bad predictors. These include female_pct, male_pct, 0_17_pct, and 65_plus_pct.

5.5 Correlation Between Predictors

The correlation matrix between the predictors is shown below in the following heat map:
will add the heat map photo after editing the problem

5.6 Improving Model Fit and Interpretability

Various strategies were experimented with to improve the model:

- Adding or Removing the Intercept: Models with and without an intercept were compared.

- **Introducing Higher-Order Terms: Polynomial features were added to capture non-linear relationships.**
- **Removing Outliers: Outliers identified using Z-scores were removed to improve model robustness.**
- **Feature Engineering: Interaction terms between hospitalization and ICU admissions were added.**

5.7 Conclusion

The final model with polynomial features, interaction terms, and an intercept showed improved fit and interpretability. Significant predictors included age distribution, gender distribution, and proportions of ICU admissions and hospitalizations. Multicollinearity was assessed and found to be within acceptable limits with $\mathbf{R}_{S}quared : 0.998.andthefollowingheatmapshowsthecorrelationbetweenpredictors$

A scatter plot of predicted values against actual values is plotted to visualize the performance of the regression model

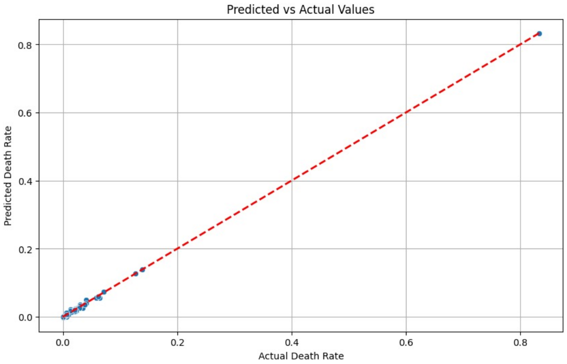


Figure 7: caption

' A histogram of the residuals is plotted to check their distribution.

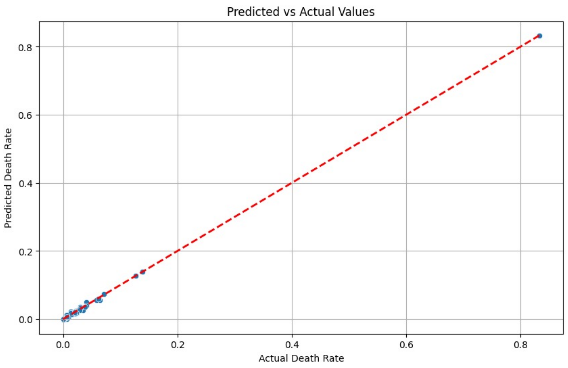
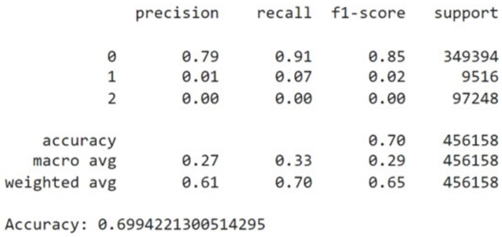


Figure 8: caption
,



```
              precision    recall  f1-score   support

           0       0.79      0.91      0.85    349394
           1       0.01      0.07      0.02      9516
           2       0.00      0.00      0.00     97248

    accuracy                           0.70    456158
   macro avg       0.27      0.33      0.29    456158
weighted avg       0.61      0.70      0.65    456158

Accuracy: 0.6994221300514295
```

Figure 9: caption