مدينة زويل للعلوم والتكنولوجيا
Zewail City of Science and Technology

Statistical inference Final Project

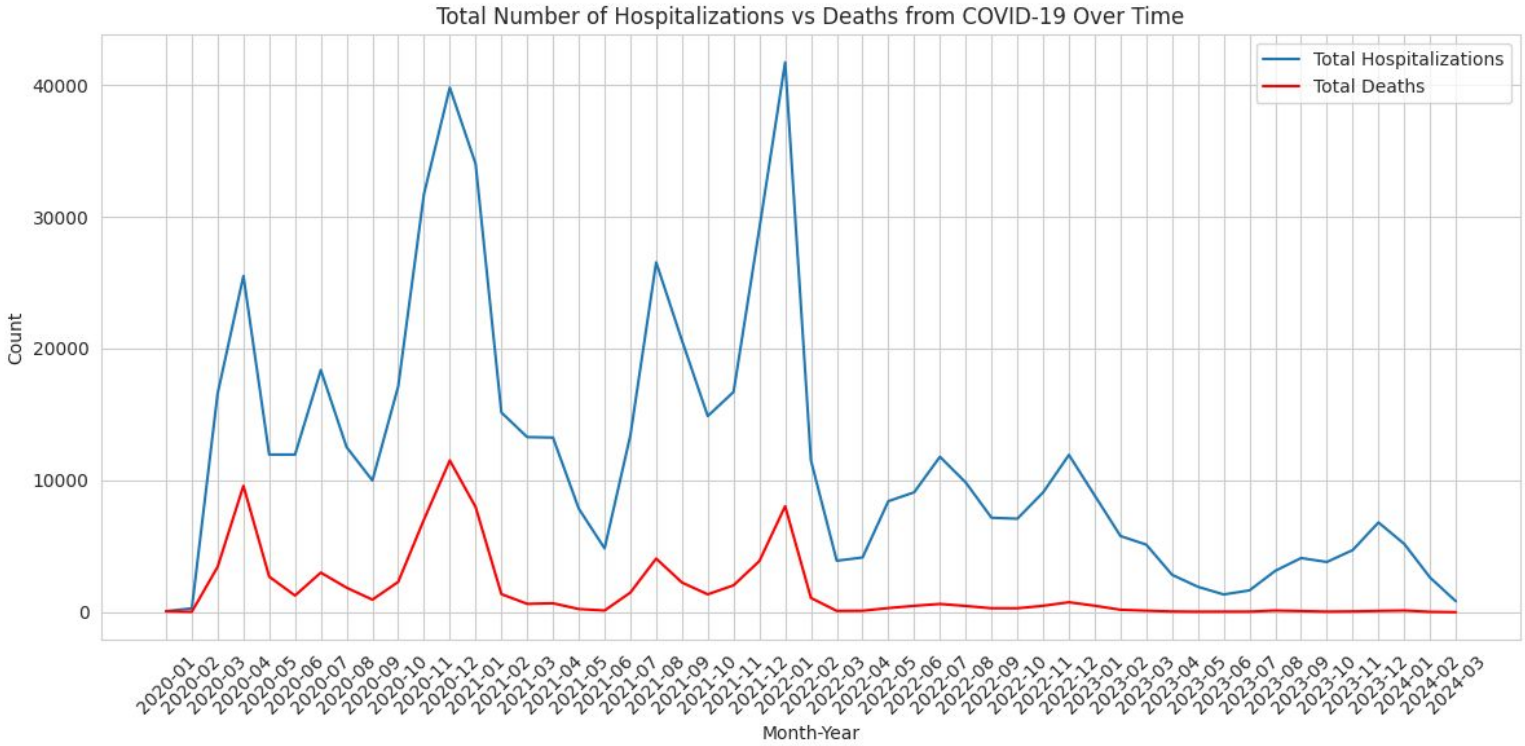
CIE 457 - Spring 24

By: Ahmed Abdelgawad (201901249) - Ziad Mohammed(202100154) –
Adham Ahmed(202100163) – Mohamed Morshedy(202100372)

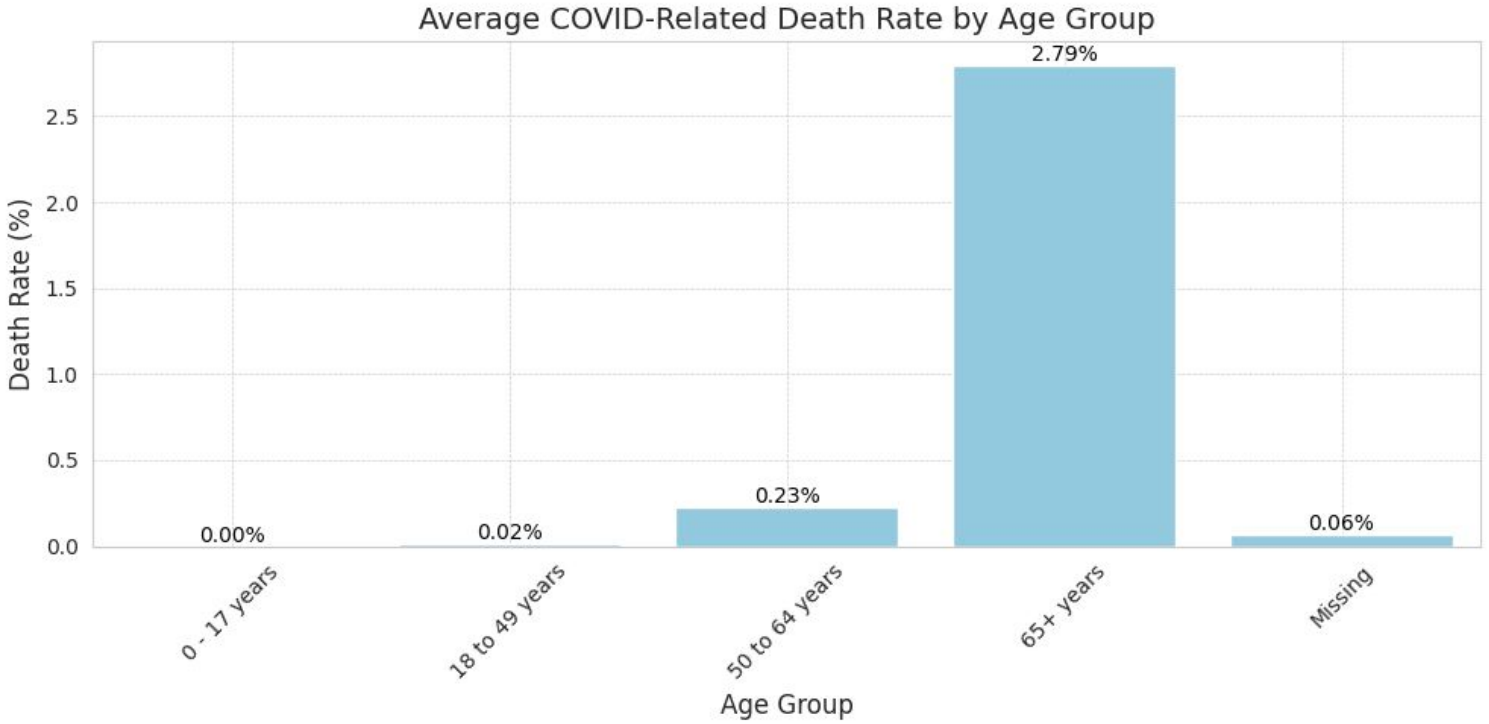
Supervised By: Dr. Mahmoud Abdelaziz Abdelaziz

PART 1: Exploratory Analysis:

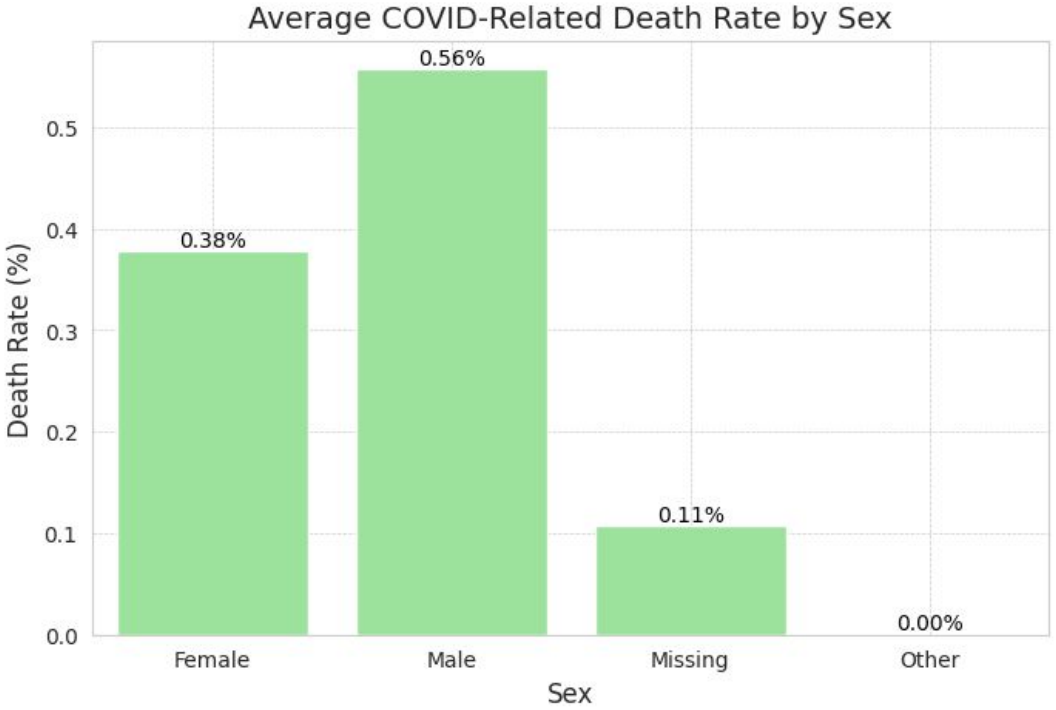
1. The total number of hospitalizations versus deaths from COVID-19 over the entire US per month-year timestamp.



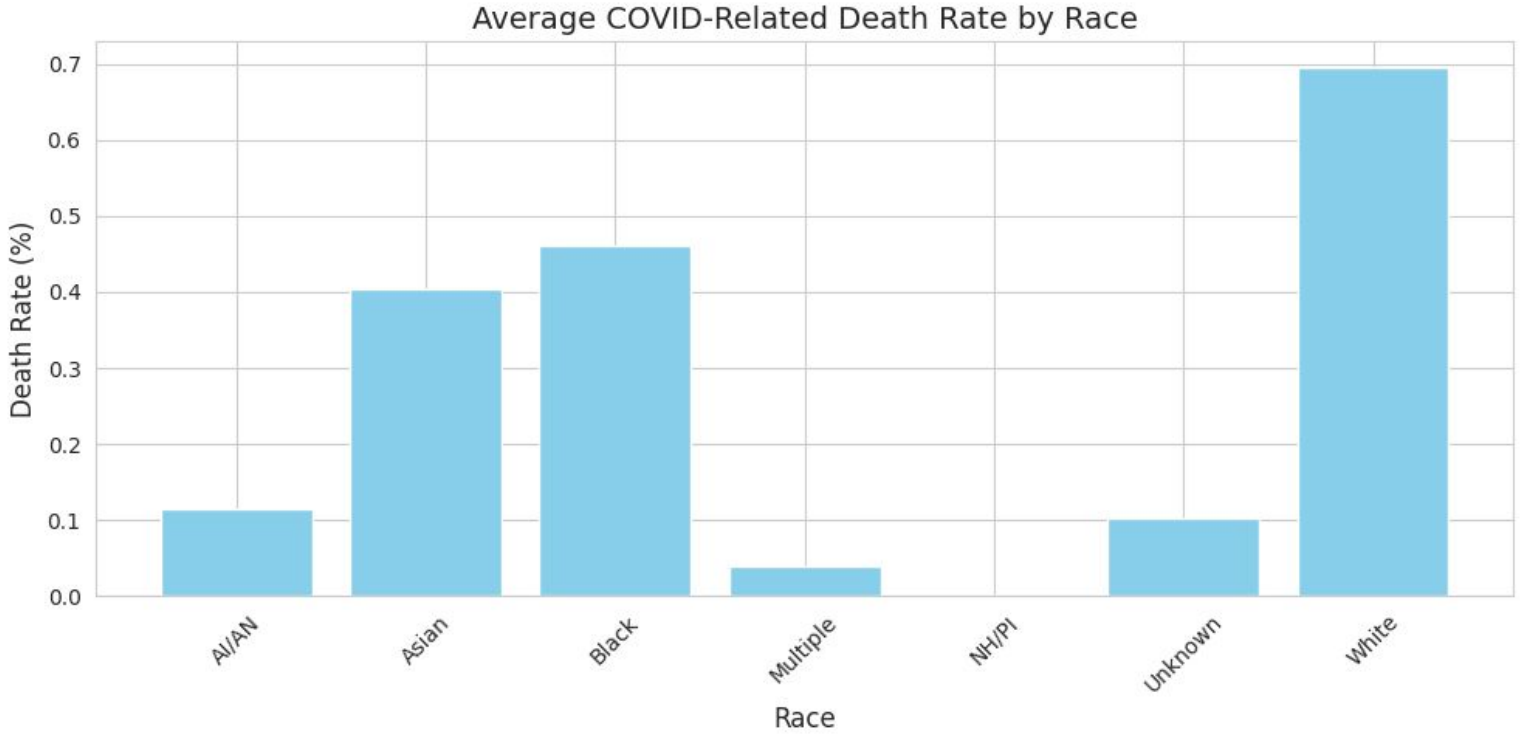
2. The average rates of COVID-related deaths relative to patient demographics (Age Group)



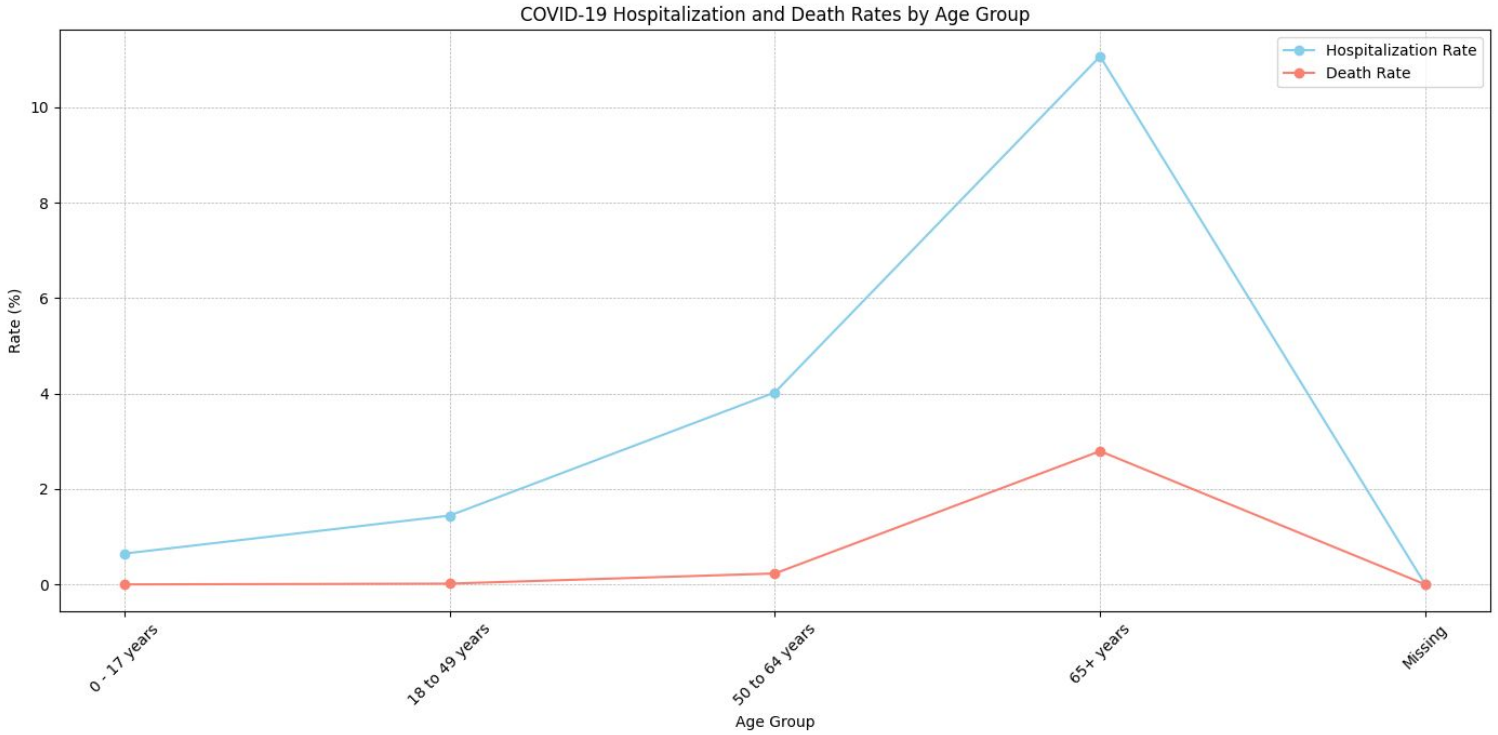
2. The average rates of COVID-related deaths relative to patient demographics (Sex)



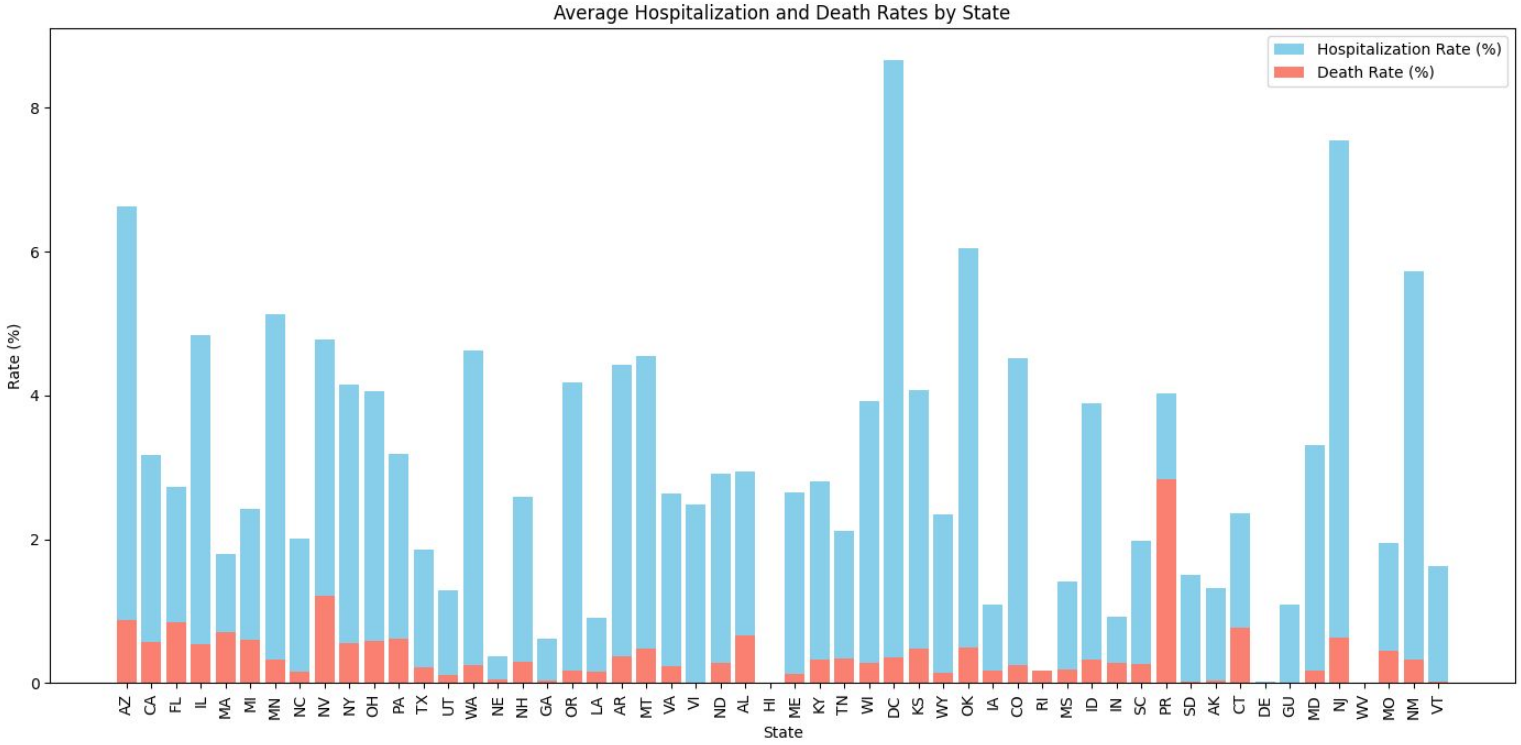
2. The average rates of COVID-related deaths relative to patient demographics (Race)



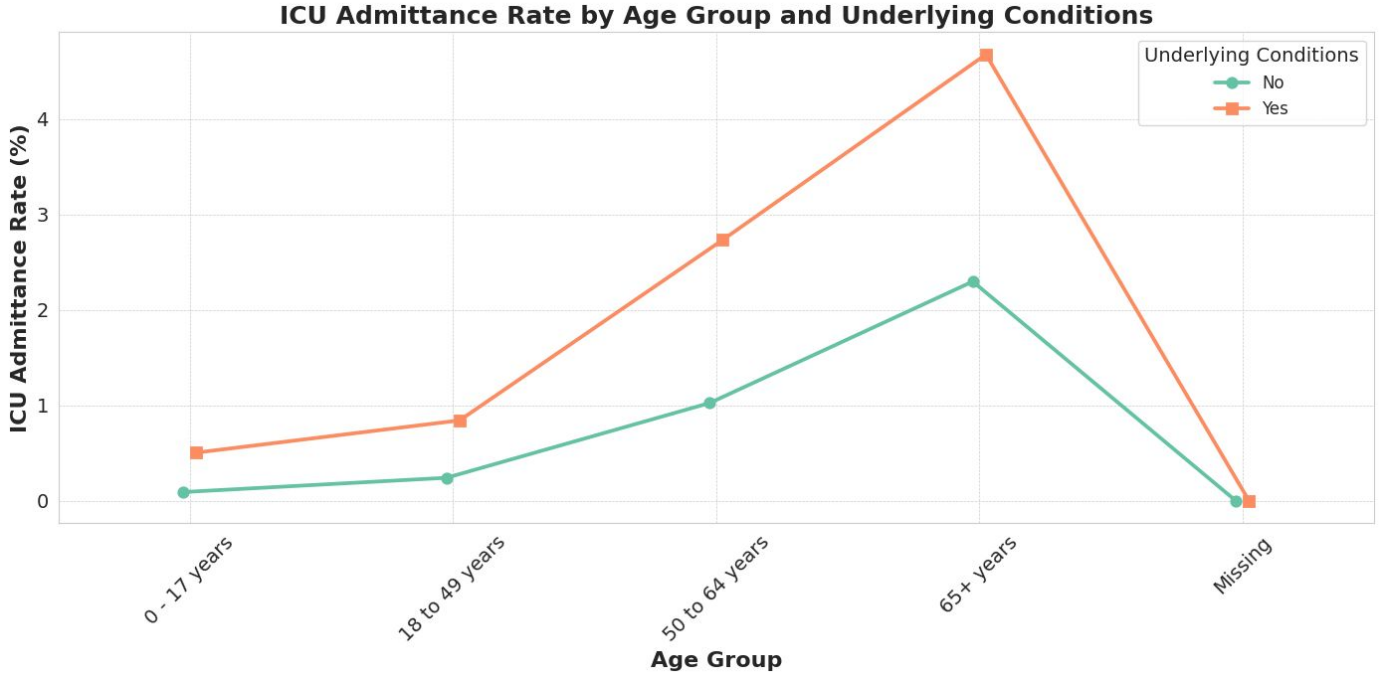
3. The rates of COVID-related hospitalization and death with age (across age groups).



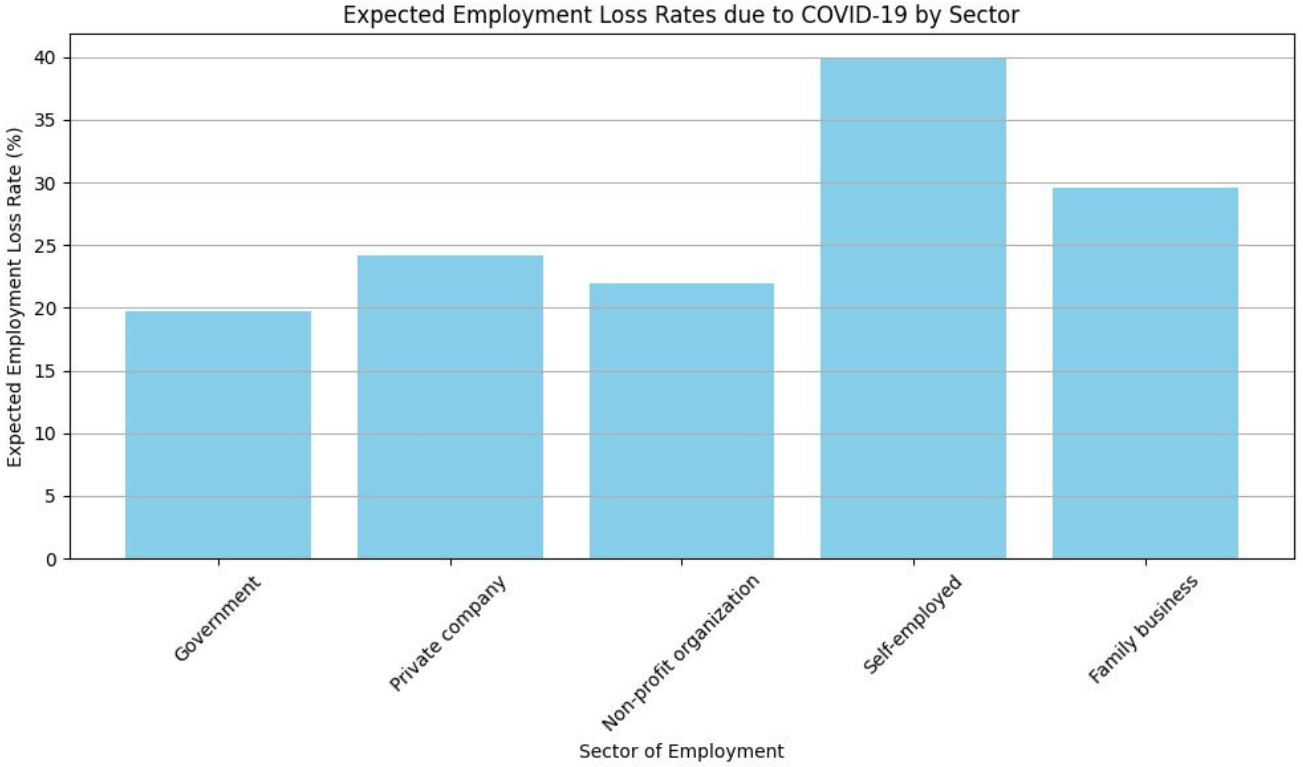
4. The Average rate of COVID-related hospitalization and death per state over the entire study period.



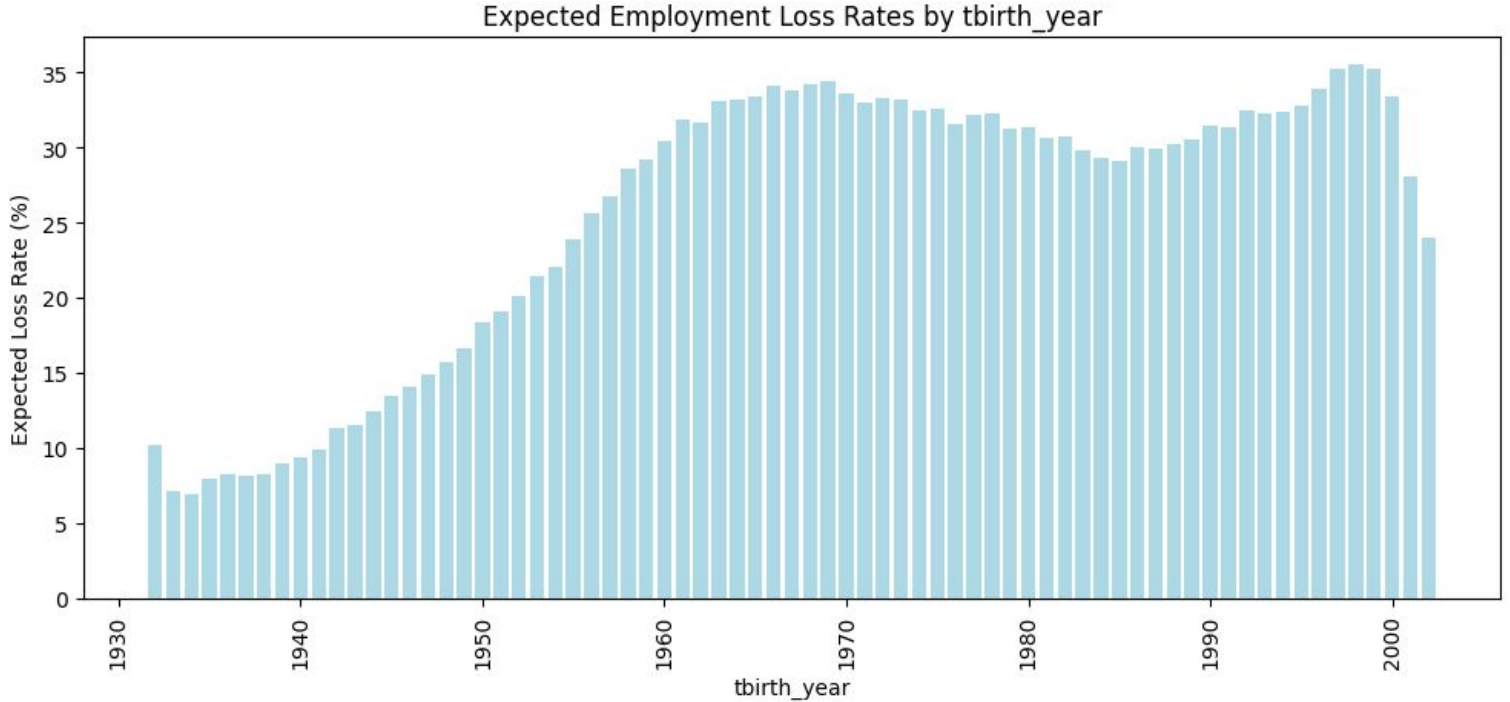
5. The relationship between age, pre-existing medical conditions and/or risk behaviors, and rate of admittance to the ICU.



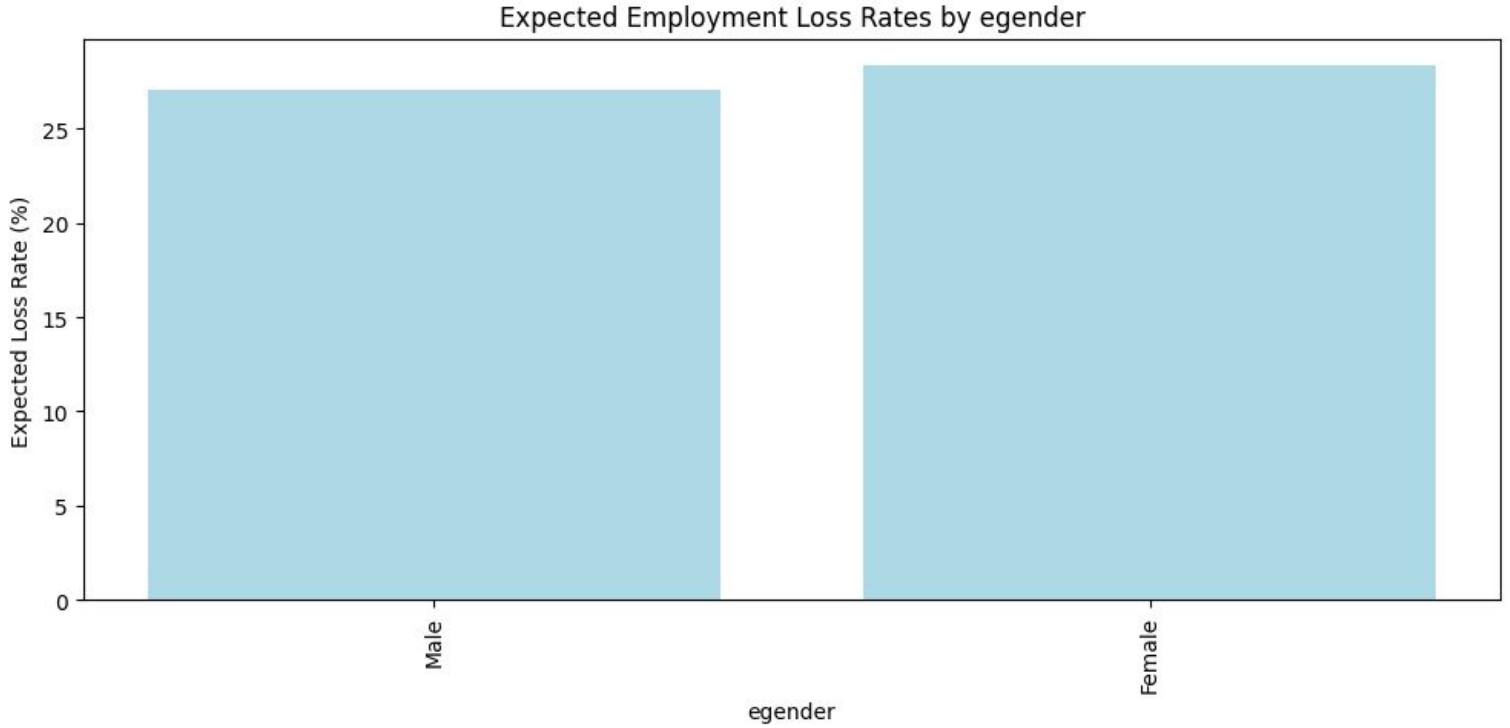
6. The rate of expected employment loss due to COVID-19 and sector of employment.



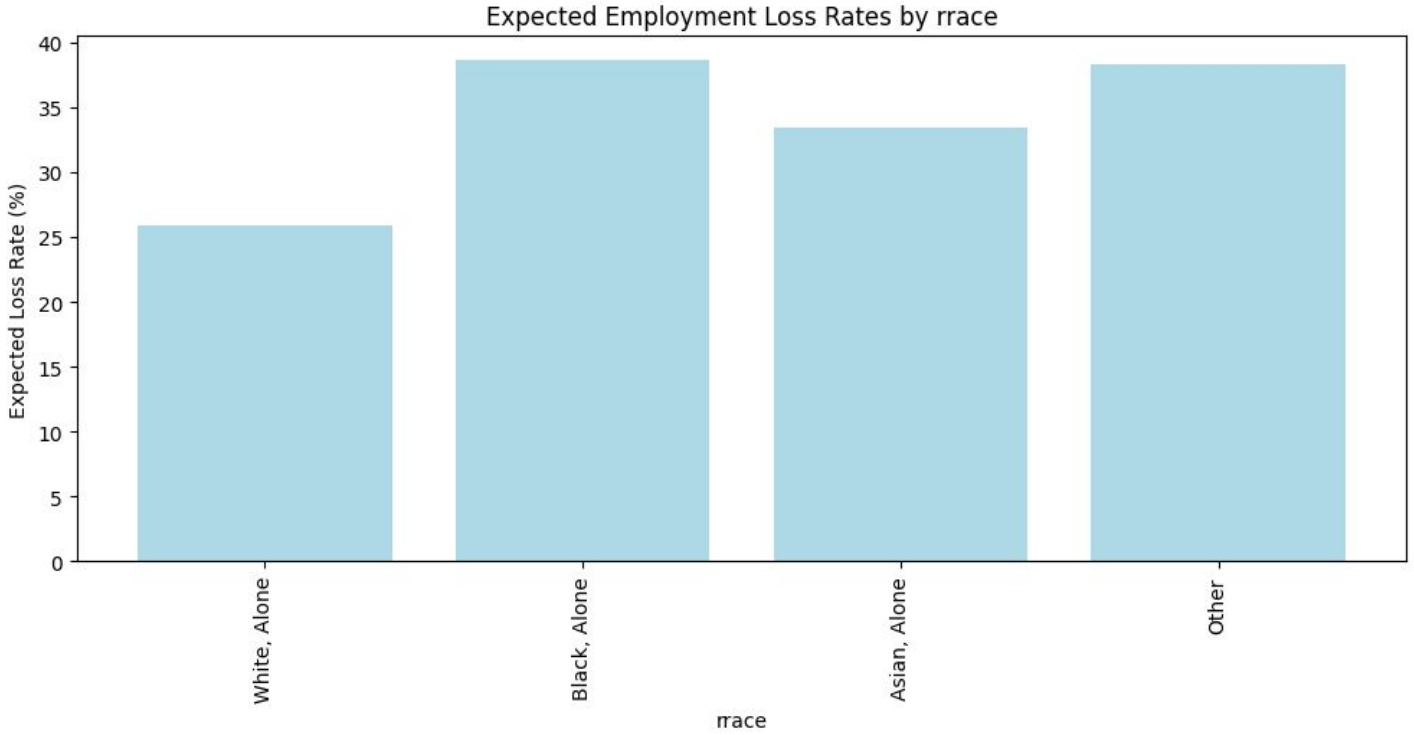
7. The rate of expected employment loss due to COVID-19 relative to responders demographics. (Year of birth)



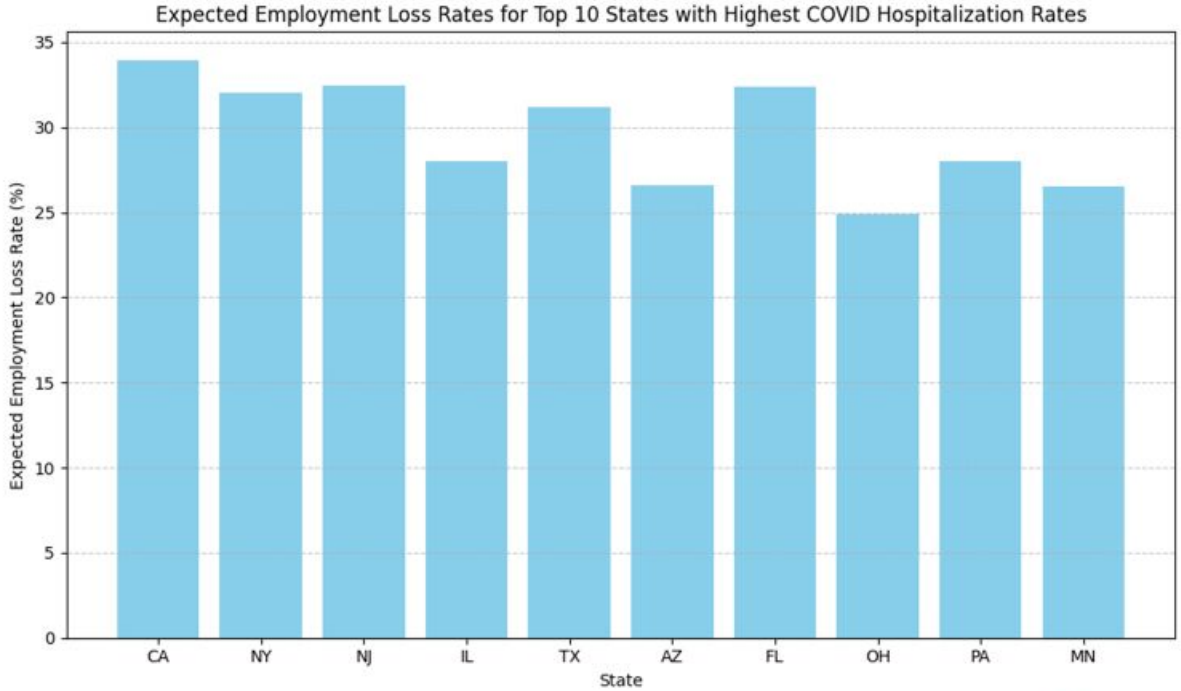
7. The rate of expected employment loss due to COVID-19 relative to responders demographics. (Gender)



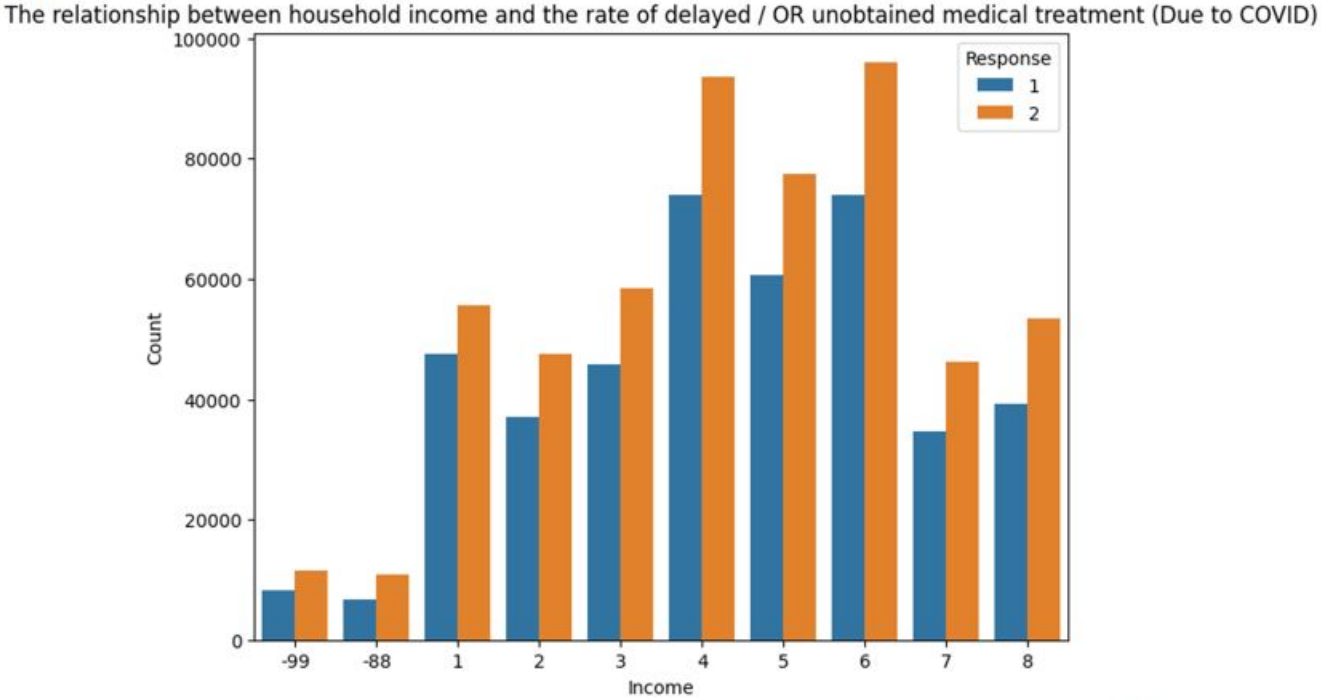
7. The rate of expected employment loss due to COVID-19 relative to responders demographics. (Race)



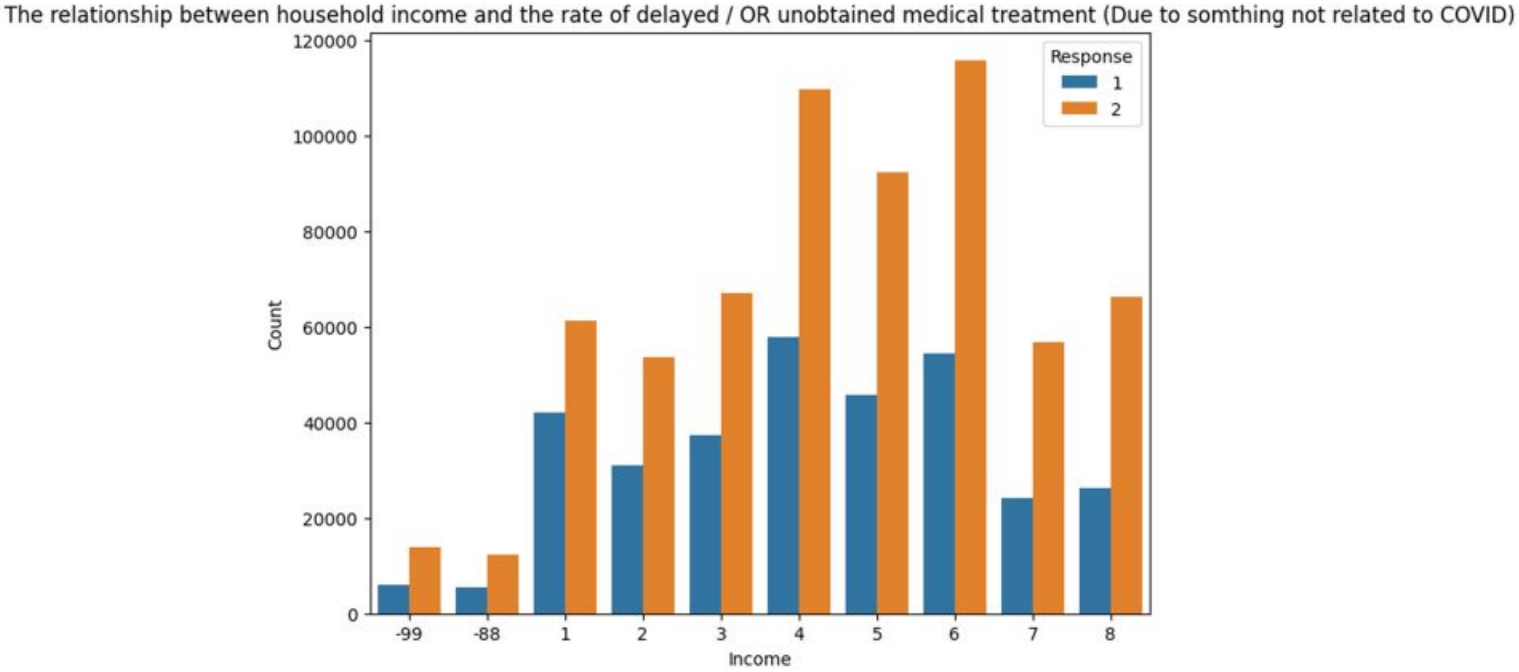
8. The rate of expected employment loss due to COVID-19 for the top 10 states with the highest rate of COVID hospitalization.



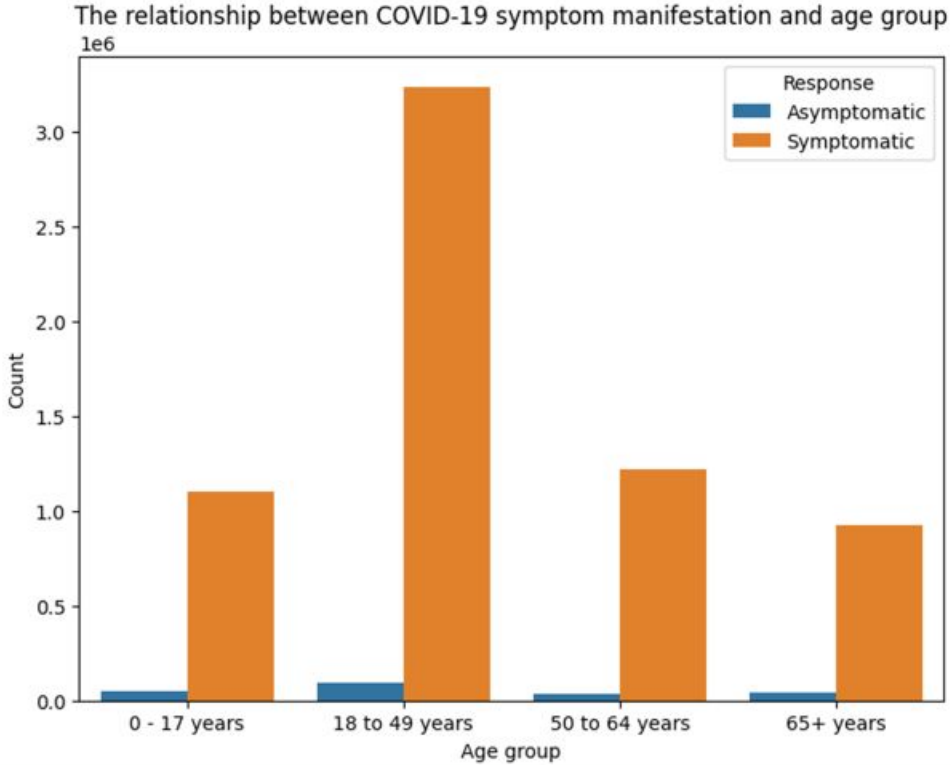
9. The relationship between household income and the rate of delayed/ OR unobtained medical treatment (Due to COVID).



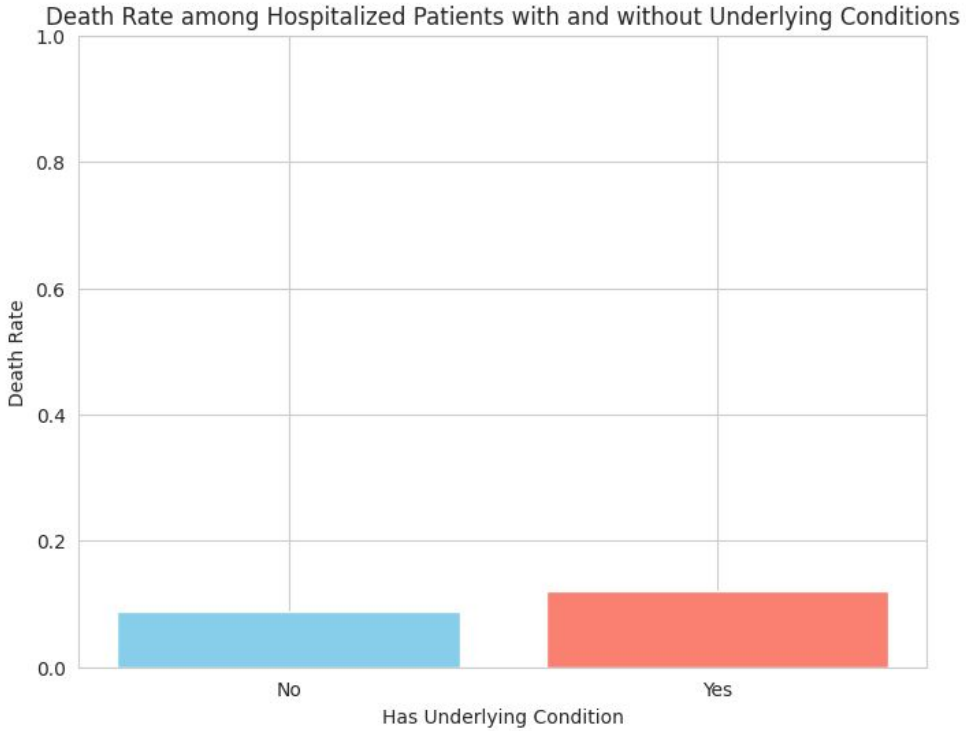
9. The relationship between household income and the rate of delayed/ OR unobtained medical treatment (Due to other reasons than COVID).



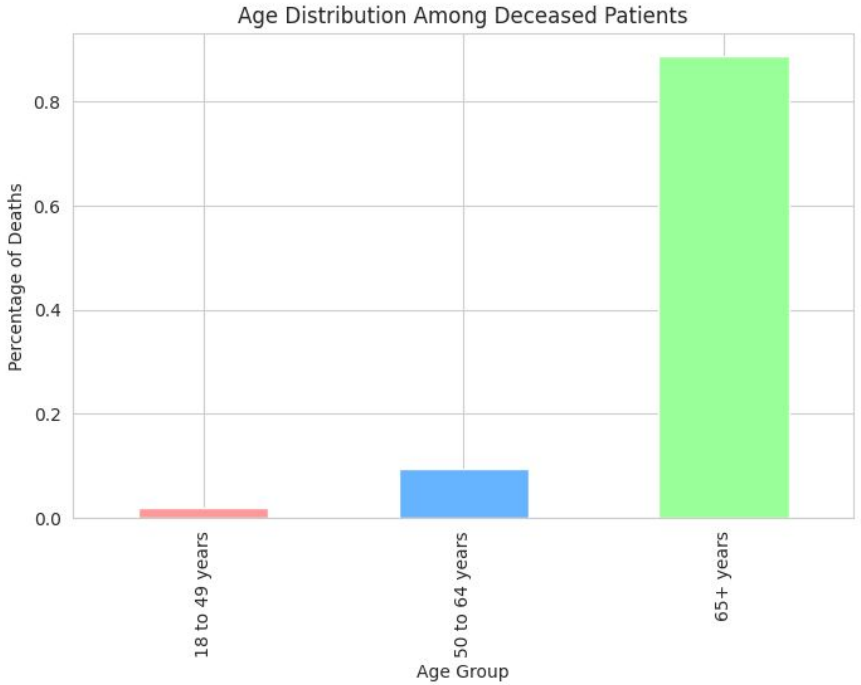
10. The relationship between COVID-19 symptom manifestation and age group.



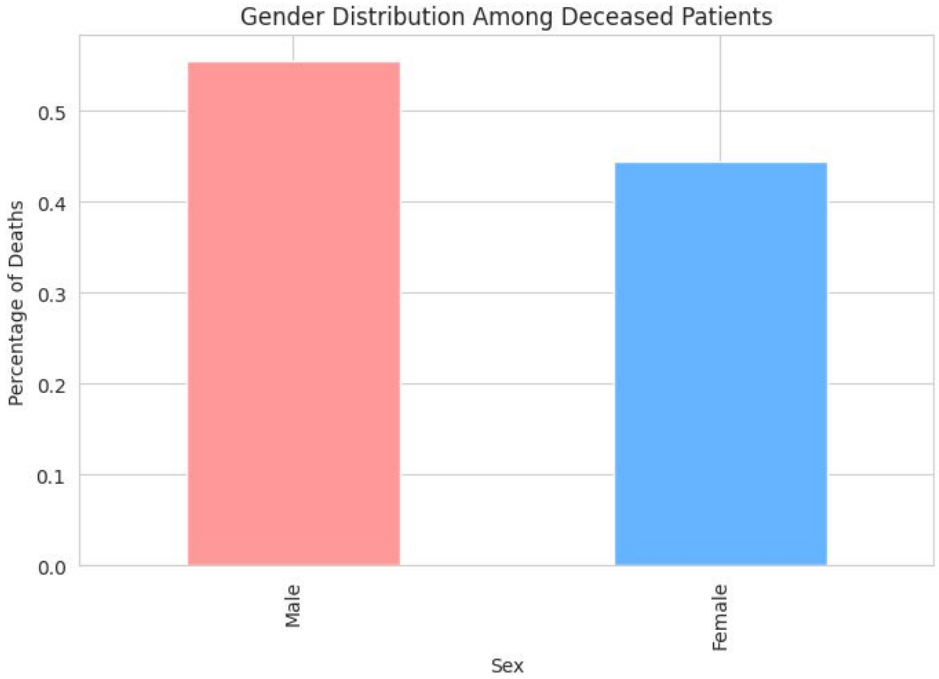
1. Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19?



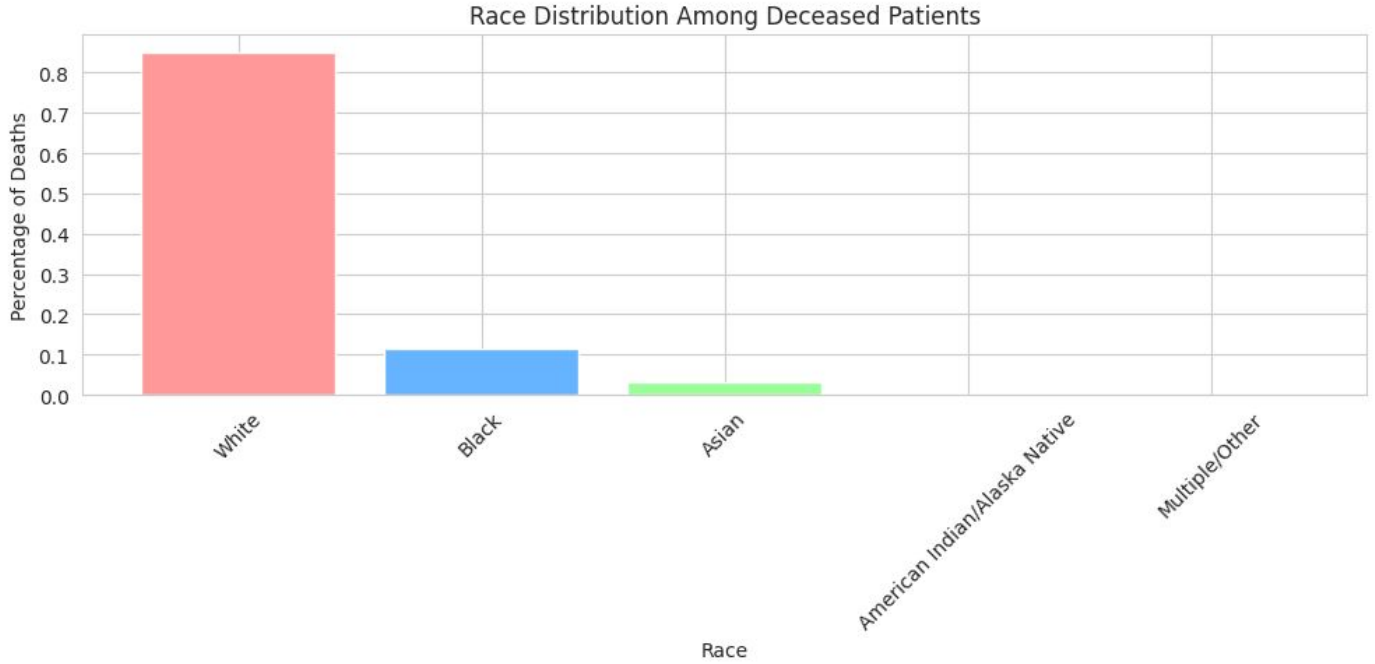
2. Are Who are the people (the demographic segment) that appear to be most at risk of death due to COVID-19? Who is the least at risk?



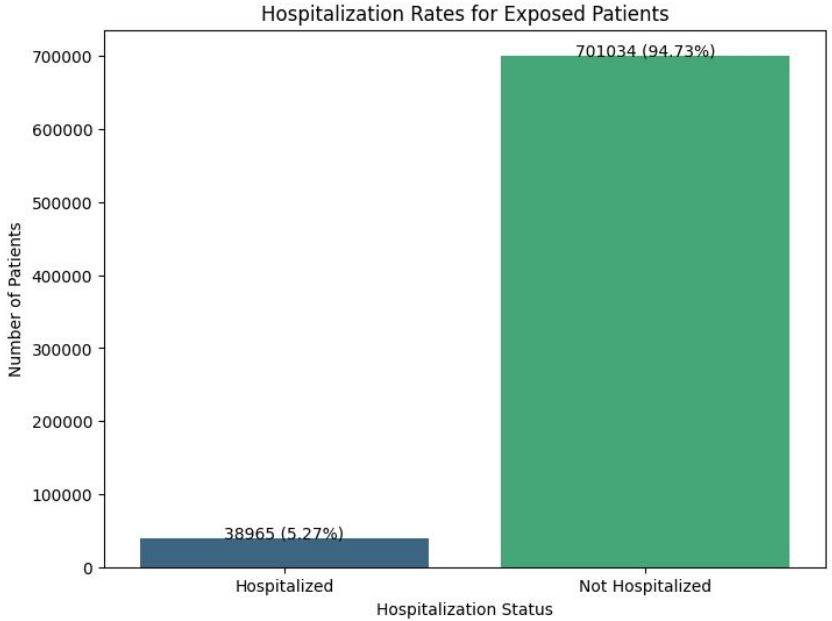
2. Are Who are the people (the demographic segment) that appear to be most at risk of death due to COVID-19? Who is the least at risk?



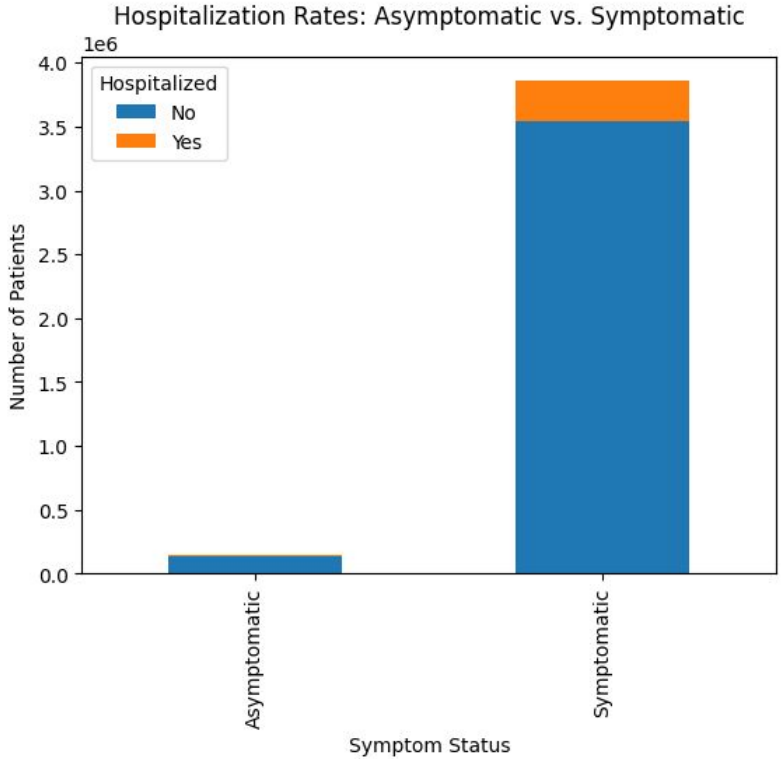
2. Are Who are the people (the demographic segment) that appear to be most at risk of death due to COVID-19? Who is the least at risk?



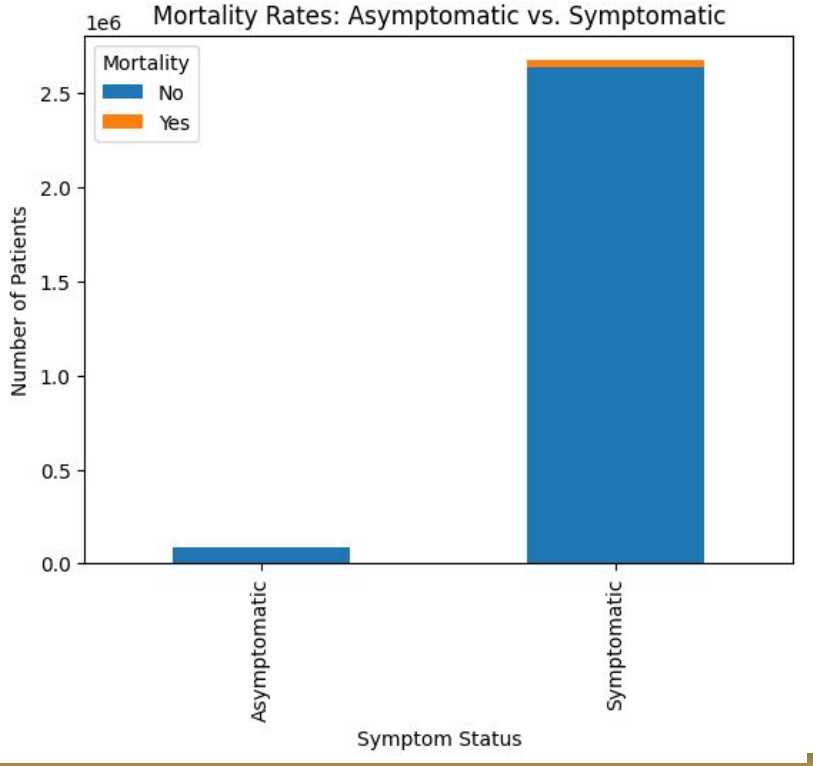
3. What percent of patients who have reported exposure to any kind of travel / or congregation within the 14 days prior to illness onset end up hospitalized? What percent of those go on to be hospitalized?



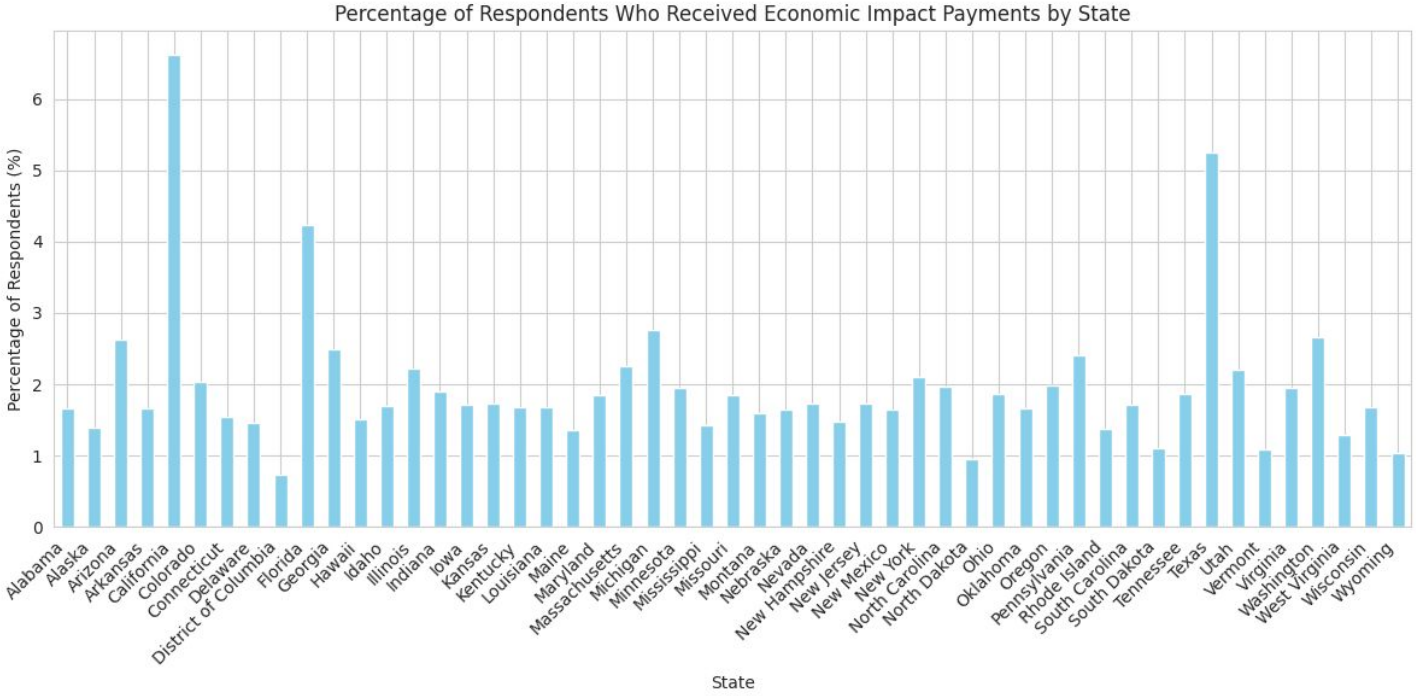
4. Are Asymptomatic COVID patients less likely to be hospitalized? Are they less likely to die from their illness?



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

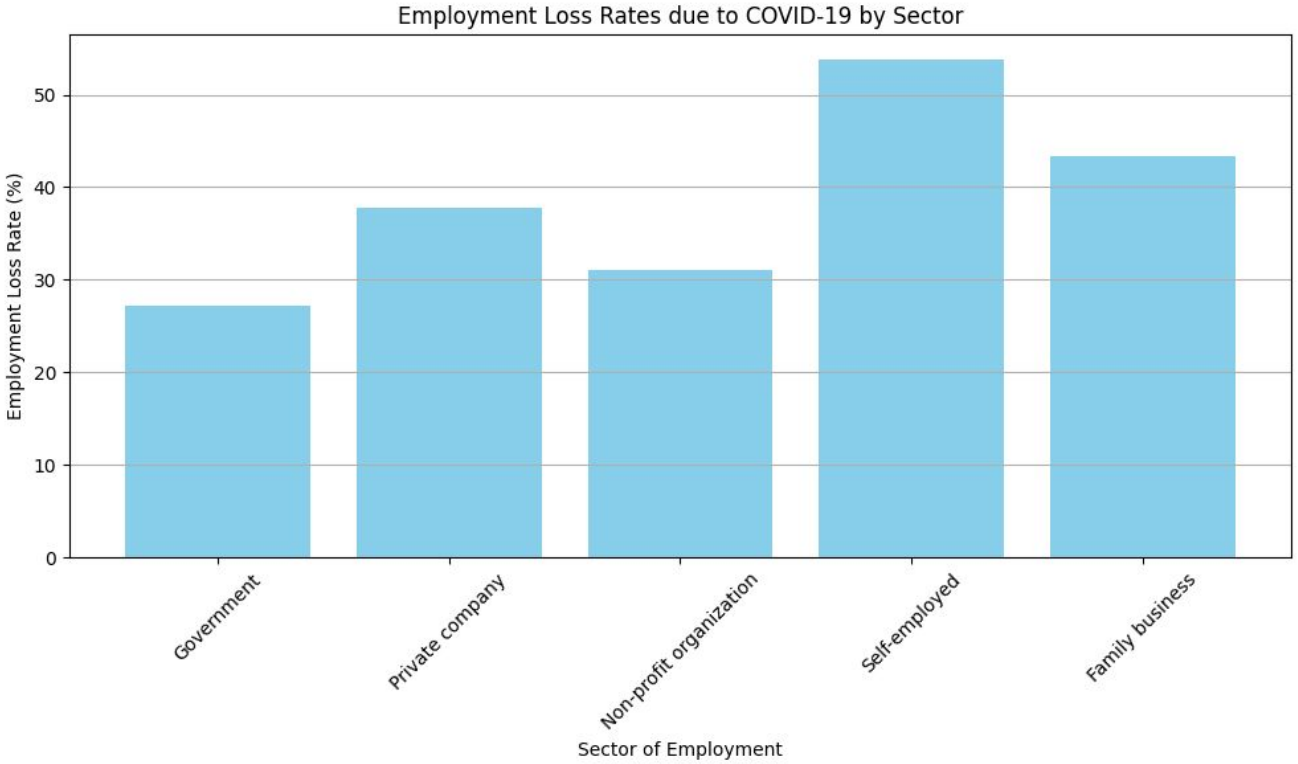


5. Which state is associated with the highest percentage of Economic Impact (stimulus) payments among survey respondents?

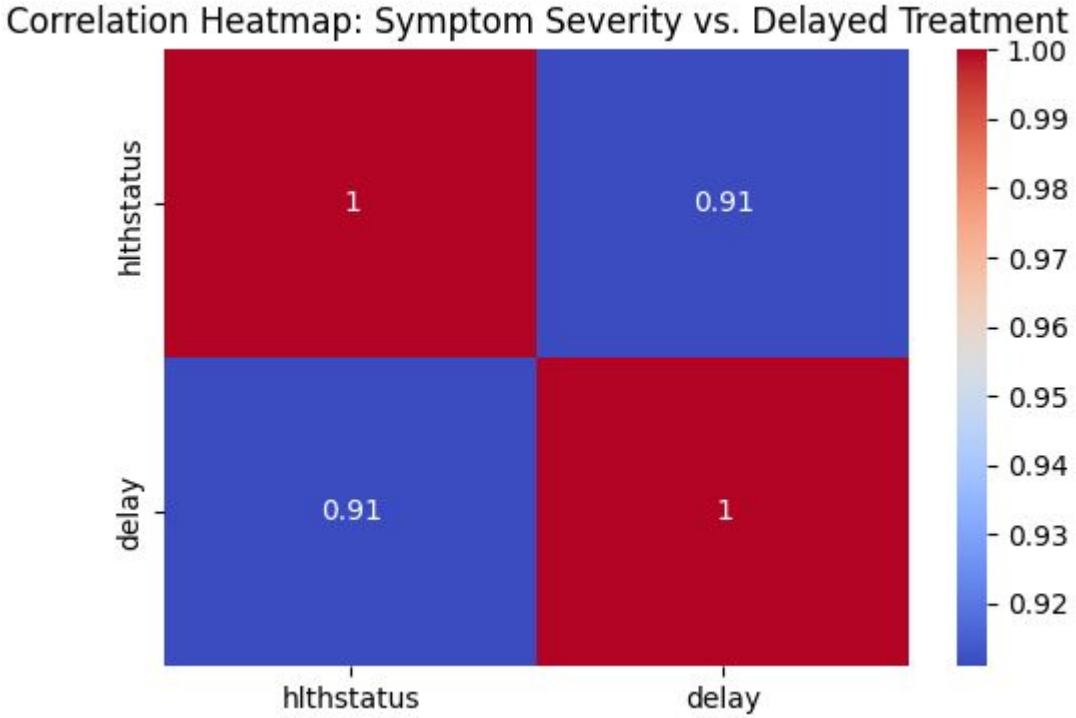


2.2 Come up with 5 more bivariate/multivariate analysis questions and similarly answer each with appropriate visuals and commentary.

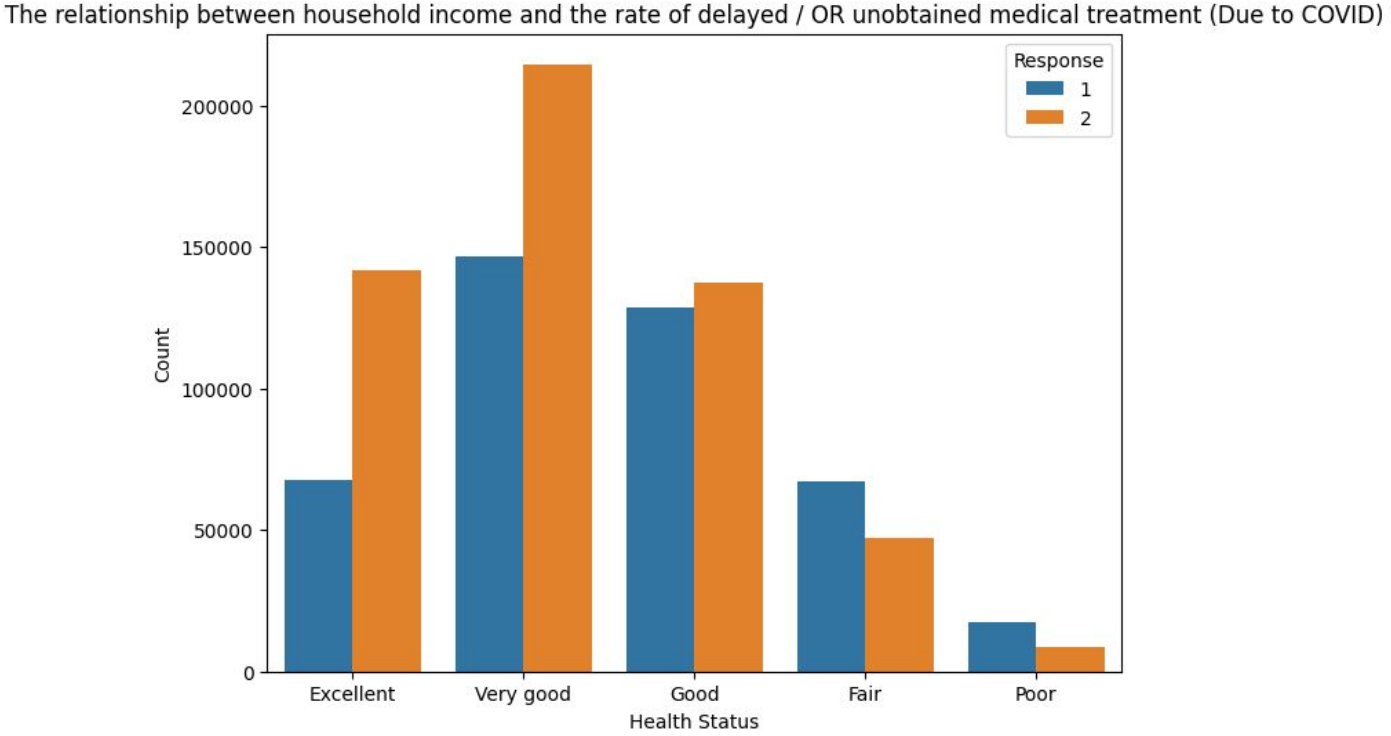
1. How does the rate of COVID-related employment loss vary across different industries?



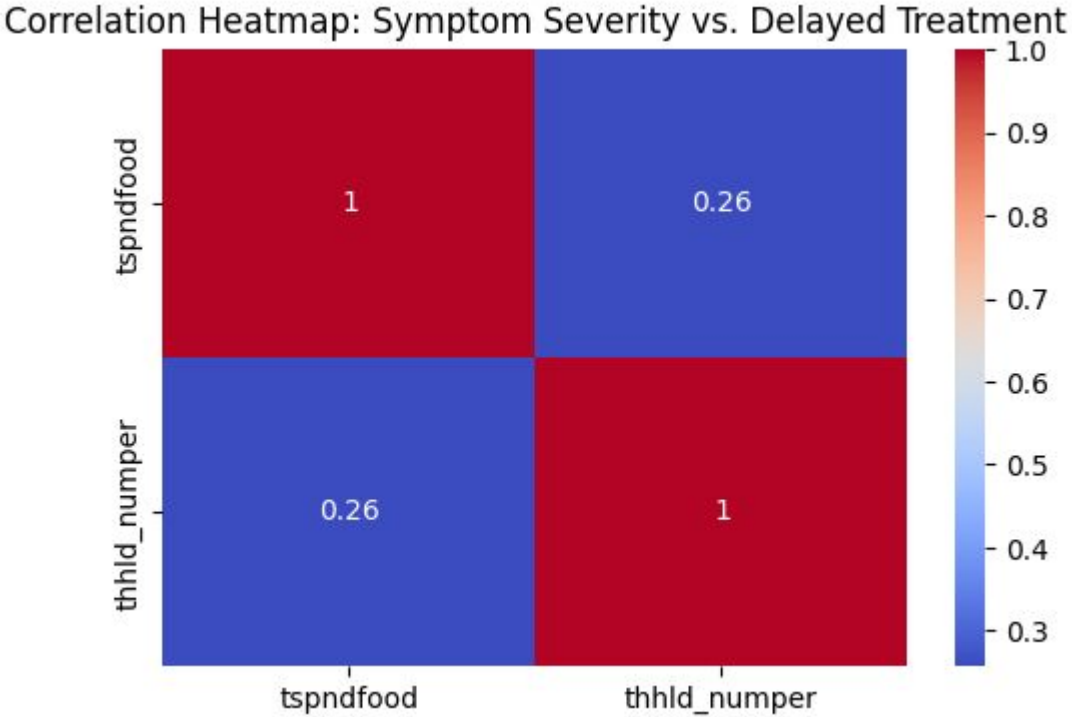
2. Is there a correlation between COVID-19 symptom severity and delayed medical treatment?



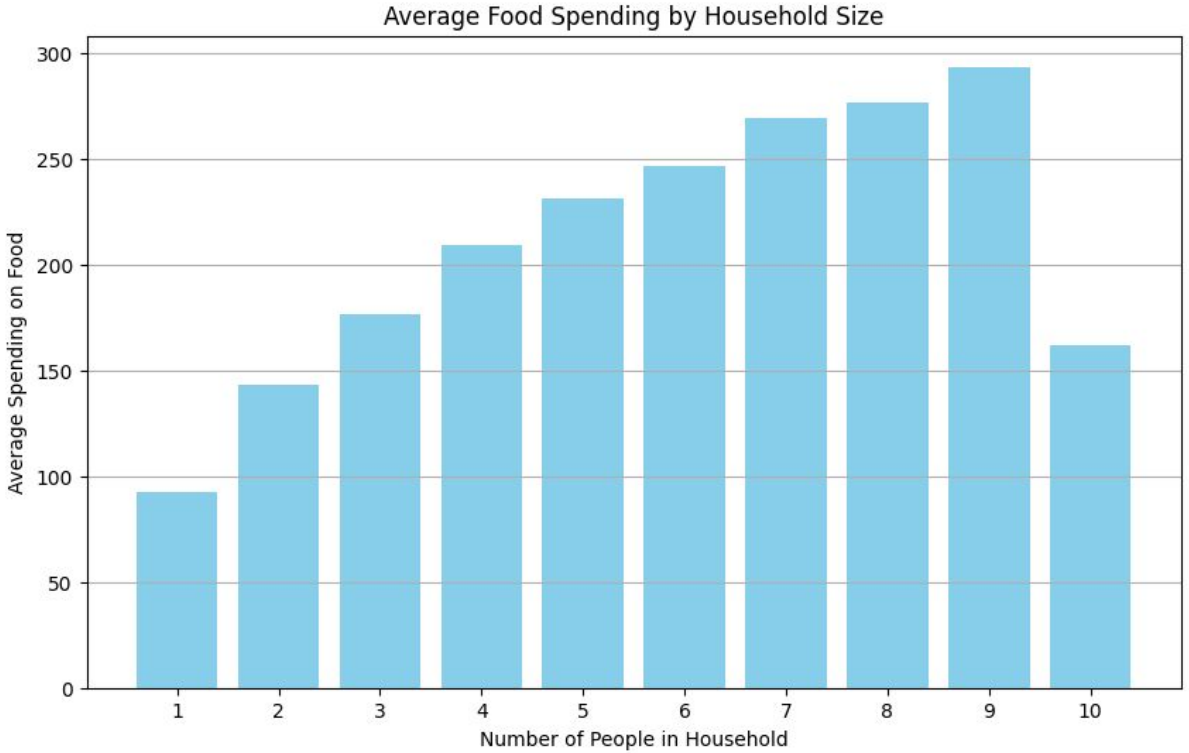
2. Is there a correlation between COVID-19 symptom severity and delayed medical treatment?



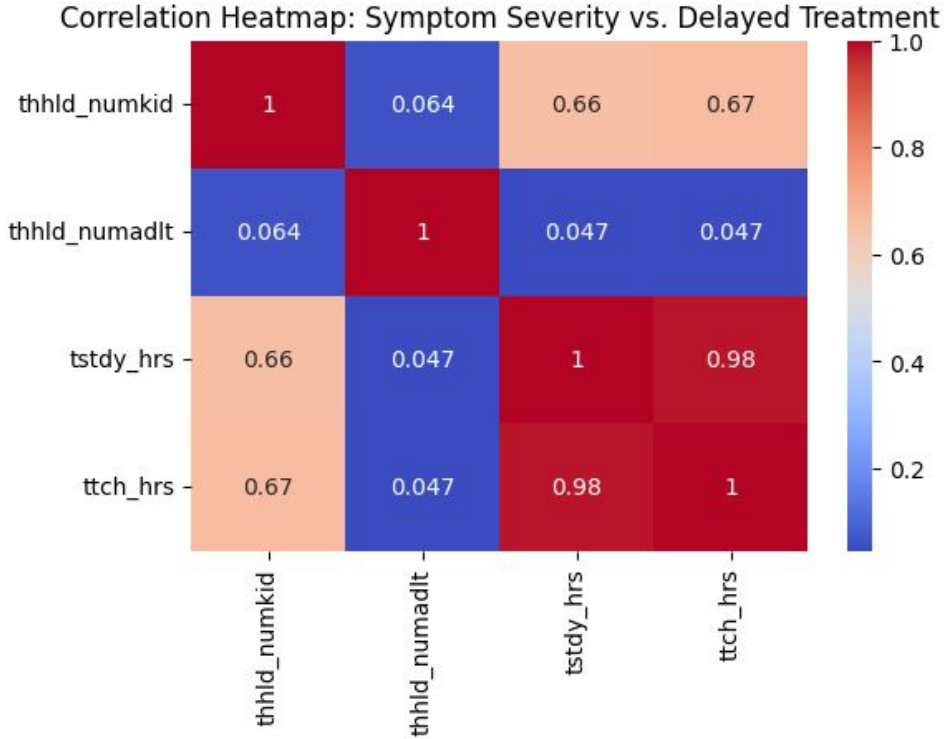
3. Is there a correlation between Household Size and Food Spending during pandemic?



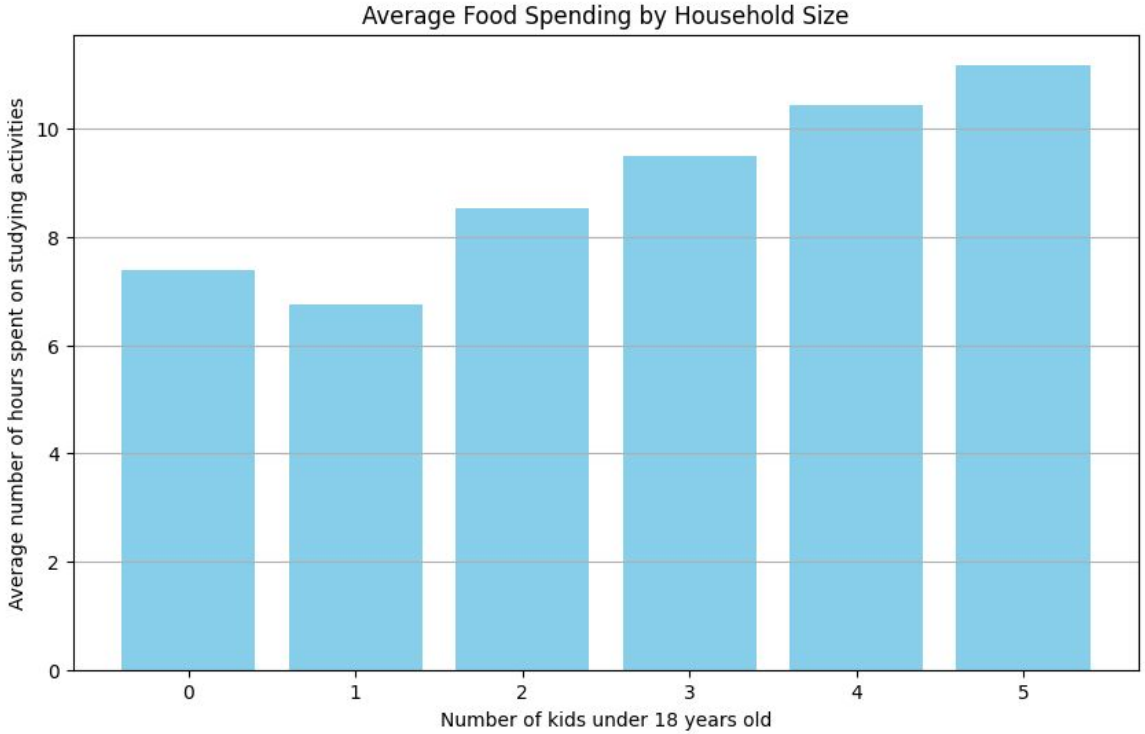
3. Is there a correlation between Household Size and Food Spending during pandemic?



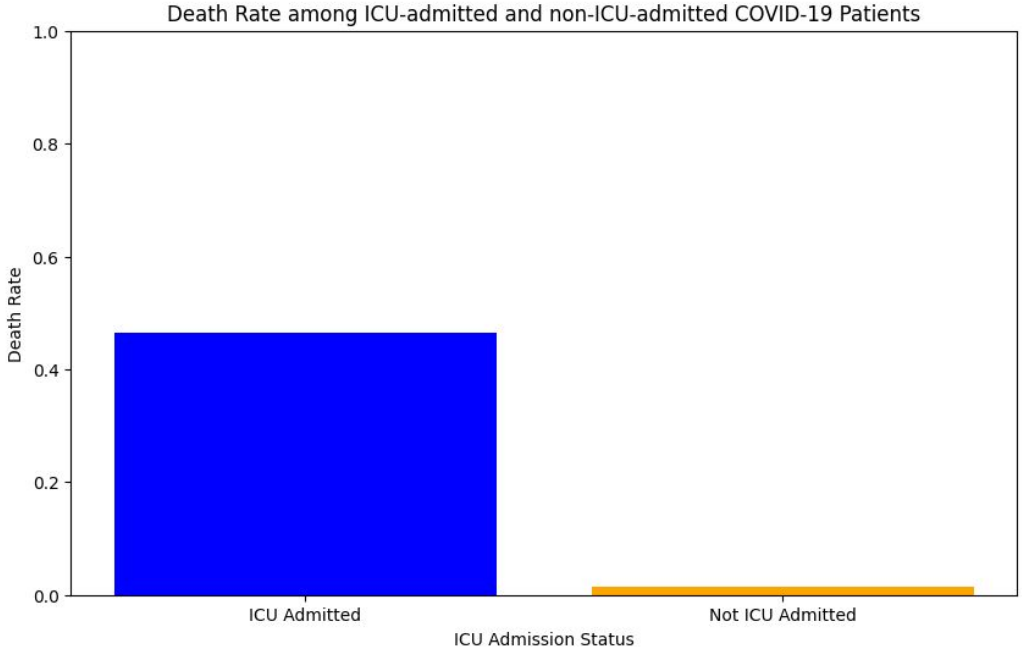
4. Is there a correlation between Household Composition and Education Spending during pandemic?



4. Is there a correlation between Household Composition and Education Spending during pandemic?



5. Is there a correlation between admission to the intensive care unit (ICU) and the death rate among COVID-19 patients?



PART 3: Hypothesis Testing:

3.1

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

Hypothesis Formulation

- Null Hypothesis (H0): There is no association between the demographic feature and the probability of death due to COVID-19.

- Alternative Hypothesis (H1): There is a significant association between the demographic feature and the probability of death due to COVID-19.

3.1

Clean the Data

```
def clean_and_preprocess_data(df):
    # Create a copy of the relevant columns
    df_filtered = df[['age_group', 'sex', 'race', 'death_yn']].copy()

    # Remove rows with 'Missing' or 'Unknown' values
    df_filtered = df_filtered[~df_filtered['age_group'].isin(['Missing', 'Unknown'])]
    df_filtered = df_filtered[~df_filtered['sex'].isin(['Missing', 'Unknown'])]
    df_filtered = df_filtered[~df_filtered['race'].isin(['Missing', 'Unknown'])]
    df_filtered = df_filtered[~df_filtered['death_yn'].isin(['Missing', 'Unknown'])]

    # Drop rows with any NaN values
    df_filtered.dropna(inplace=True)

    # Combine 'age_group', 'sex', and 'race' into a single 'demographics' column
    df_filtered['demographics'] = df_filtered[['age_group', 'sex', 'race']].astype(str).agg('_'.join, axis=1)
    # Display the first few rows of the cleaned DataFrame
    df_filtered.head()
    return df_filtered

df_filtered= clean_and_preprocess_data(df)
```

	age_group	sex	race	death_yn	demographics
4	65+ years	Female	White	No	65+ years_Female_White
10	18 to 49 years	Female	Black	No	18 to 49 years_Female_Black
12	18 to 49 years	Female	White	No	18 to 49 years_Female_White
13	50 to 64 years	Female	White	No	50 to 64 years_Female_White
16	0 - 17 years	Male	White	No	0 - 17 years_Male_White

3.1

Test Choice and Justification

Chi-Square Test of Independence: This test is suitable for examining the association between two categorical variables. It is appropriate here because both the demographic features and the outcome (death due to COVID-19) are categorical. The Chi-Square test is non-parametric and does not assume any specific distribution of the data.

```
Chi-Square Test for demographics and death_yn
Chi2 Statistic: 24708.562478559325, p-value: 0.0
Reject the null hypothesis - There is a significant association between the demographic feature and the probability of death due to COVID-19.
```

3.2

Our Claim: “There is a significant association between the combined demographic features and the probability of entering ICU admission due to COVID-19”

Hypothesis Formulation

- 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.

3.2

Conduct the test and report the result.

```
Chi-Square Statistic (ICU Admission): 2557.6799468278 35
P-value (ICU Admission): 0.0
```

There is a significant association between the combined demographic features and the probability of ICU admission due to COVID-19 ($p < 0.05$)

Advantages of Testing on Combined Demographics:

1. Reduced Multiple Testing: By combining multiple demographic features into a single variable, as we reduce the number of statistical tests performed. This can help mitigate the risk of Type I errors (false positives) associated with multiple comparisons.

2. Capturing Interactions: Sometimes, different demographic factors can affect each other in unexpected ways. By combining them, we can capture these interactions or how they work together, giving us a better overall picture of their connection to the outcome we're interested in, like whether someone dies from COVID-19.

3. Simplicity: Instead of dealing with many separate factors, testing on just one combined variable makes things simpler. It's easier to analyze and understand the results, especially when we're looking at several demographic features at once..

Disadvantages of Testing on Combined Demographics:

1. Loss of Specificity: Combining demographic features into a single variable may result in loss of specificity, as individual effects of each demographic variable may be obscured.

2. Assumption of Independence: Chi-Square test assumes independence between variables. Combining demographic features may violate this assumption if there are dependencies among them.

PART 4: Regression Analysis

1-Report your model's coefficients and p-values.

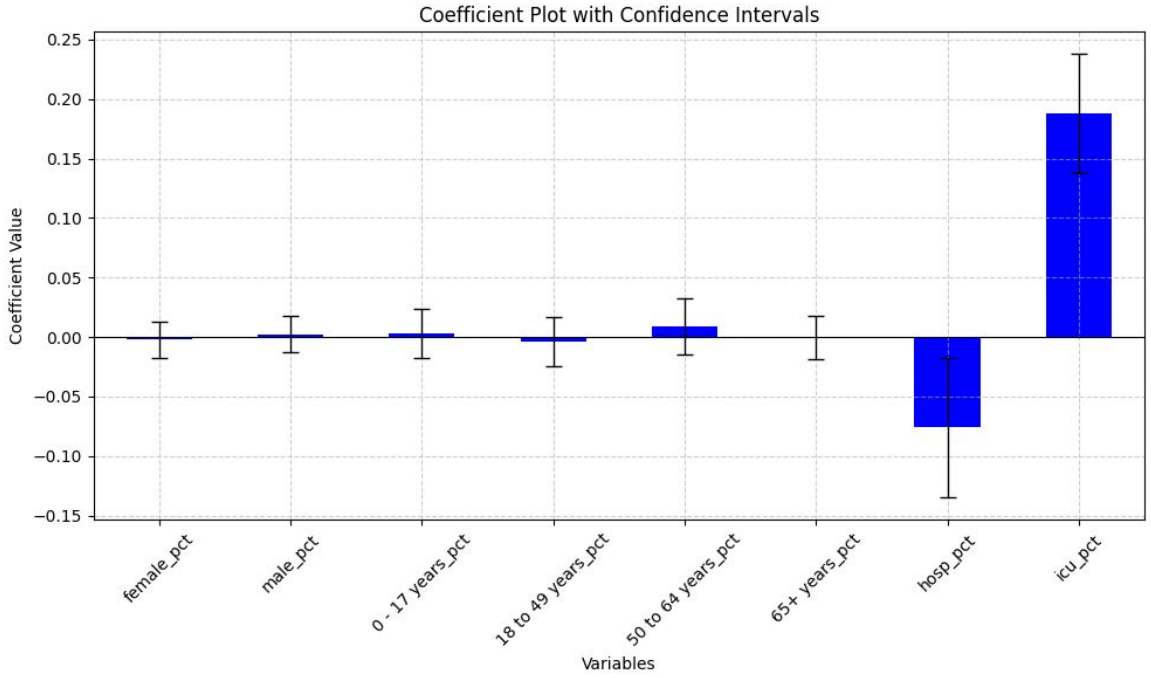
OLS Regression Results

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:	Thu, 23 May 2024	Prob (F-statistic):	2.48e-17
Time:	05:48:00	Log-Likelihood:	83.151
No. Observations:	50	AIC:	-154.3
Df Residuals:	44	BIC:	-142.8
Df Model:	6		
Covariance Type:	nonrobust		

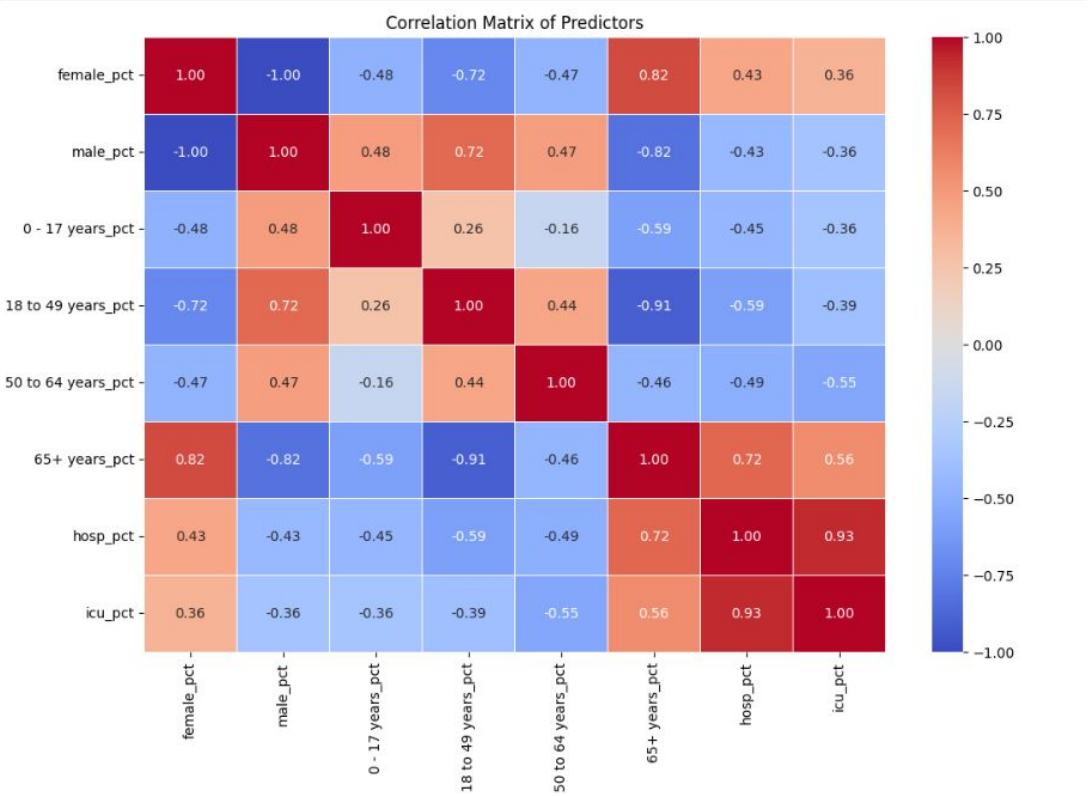
	coef	std err	t	P>|t|	[0.025	0.975]
female_pct	-0.0024	0.008	-0.308	0.760	-0.018	0.013
male_pct	0.0024	0.008	0.308	0.760	-0.013	0.018
0 - 17 years_pct	0.0031	0.010	0.312	0.756	-0.017	0.023
18 to 49 years_pct	-0.0039	0.010	-0.379	0.706	-0.025	0.017
50 to 64 years_pct	0.0088	0.012	0.746	0.459	-0.015	0.032
65+ years_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

Omnibus:	10.474	Durbin-Watson:	0.152
Prob(Omnibus):	0.005	Jarque-Bera (JB):	10.355
Skew:	-0.905	Prob(JB):	0.00564
Kurtosis:	4.301	Cond. No.	1.12e+16

2-Which of these variables are good predictors of the variabilities in the target? Which are bad ones?



3. Are any of these predictors correlated with each other?



4. Experiment with different ways to improve the fit and interpretability of the model.

OLS Regression Results

Dep. Variable:	death_pct	R-squared:	0.966
Model:	OLS	Adj. R-squared:	0.962
Method:	Least Squares	F-statistic:	206.6
Date:	Thu, 23 May 2024	Prob (F-statistic):	4.87e-30
Time:	06:31:47	Log-Likelihood:	121.39
No. Observations:	50	AIC:	-228.8
Df Residuals:	43	BIC:	-215.4
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0406	0.003	12.468	0.000	0.034	0.047
female_pct	-0.0024	0.004	-0.654	0.517	-0.010	0.005
male_pct	0.0024	0.004	0.654	0.517	-0.005	0.010
0 - 17 years_pct	0.0031	0.005	0.664	0.511	-0.006	0.013
18 to 49 years_pct	-0.0039	0.005	-0.806	0.425	-0.014	0.006
50 to 64 years_pct	0.0088	0.006	1.585	0.120	-0.002	0.020
65+ years_pct	-0.0007	0.004	-0.158	0.876	-0.009	0.008
hosp_pct	-0.0760	0.014	-5.530	0.000	-0.104	-0.048
icu_pct	0.1876	0.012	16.011	0.000	0.164	0.211

Omnibus:	10.474	Durbin-Watson:	0.701
Prob(Omnibus):	0.005	Jarque-Bera (JB):	10.355
Skew:	-0.905	Prob(JB):	0.00564
Kurtosis:	4.301	Cond. No.	2.26e+17

4. Experiment with different ways to improve the fit and interpretability of the model.

OLS Regression Results

Dep. Variable:	death_pct	R-squared:	0.989
Model:	OLS	Adj. R-squared:	0.987
Method:	Least Squares	F-statistic:	480.9
Date:	Thu, 23 May 2024	Prob (F-statistic):	5.44e-38
Time:	06:32:31	Log-Likelihood:	150.30
No. Observations:	50	AIC:	-282.6
Df Residuals:	41	BIC:	-265.4
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0398	0.003	11.689	0.000	0.033	0.047
female_pct	-0.0063	0.002	-2.939	0.005	-0.011	-0.002
male_pct	0.0063	0.002	2.939	0.005	0.002	0.011
0 - 17 years_pct	0.0008	0.003	0.278	0.783	-0.005	0.006
18 to 49 years_pct	-0.0029	0.003	-1.015	0.316	-0.009	0.003
50 to 64 years_pct	0.0130	0.003	3.861	0.000	0.006	0.020
65+ years_pct	-0.0013	0.002	-0.517	0.608	-0.006	0.004
hosp_pct	-0.0222	0.010	-2.157	0.037	-0.043	-0.001
icu_pct	0.0924	0.019	4.990	0.000	0.055	0.130
hosp_pct_squared	-0.0264	0.004	-6.159	0.000	-0.035	-0.018
icu_pct_squared	0.0273	0.003	9.355	0.000	0.021	0.033

Omnibus:	1.442	Durbin-Watson:	1.683
Prob(Omnibus):	0.486	Jarque-Bera (JB):	0.673
Skew:	0.160	Prob(JB):	0.714
Kurtosis:	3.469	Cond. No.	4.13e+16

PART 5: Bonus

Model output

```
Accuracy: 0.96420988777667667
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
```

THANK YOU