Statistical Inference and Data Analysis
CIE 457 2024

Business Analysis

By:

Ahmed Farid	202000625
Mai Mohamed	202000746
Wessam Zaid	202001732
Yasmeen Abosaif	202001116

Supervised By:

Professor Mahmoud Adelaziz

Table of Contents:

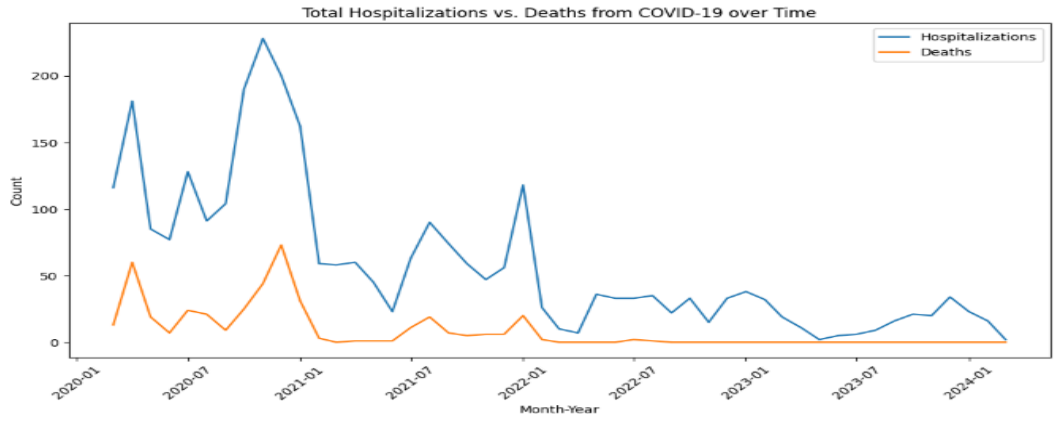
Introduction

The COVID-19 pandemic has had a profound impact on every facet of society, from healthcare systems and economies to individual lives. We embarked on a comprehensive analysis of the U.S. COVID-19 data. Leveraging publicly available datasets from the CDC and the U.S. Census Bureau, we aimed to uncover and understand the patterns of COVID-19 spread, management, and its socio-economic consequences across various demographics and regions.

This report is structured to provide detailed insights into the key aspects of the pandemic's impact, supported by statistical analyses and visualizations. We explored multiple dimensions, including hospitalization and death rates, demographic influences, economic impacts, and relationships between various risk factors and outcomes. Our analysis not only sheds light on critical trends and correlations but also offers a foundation for informed decision-making and policy development.

2. Exploratory Analysis

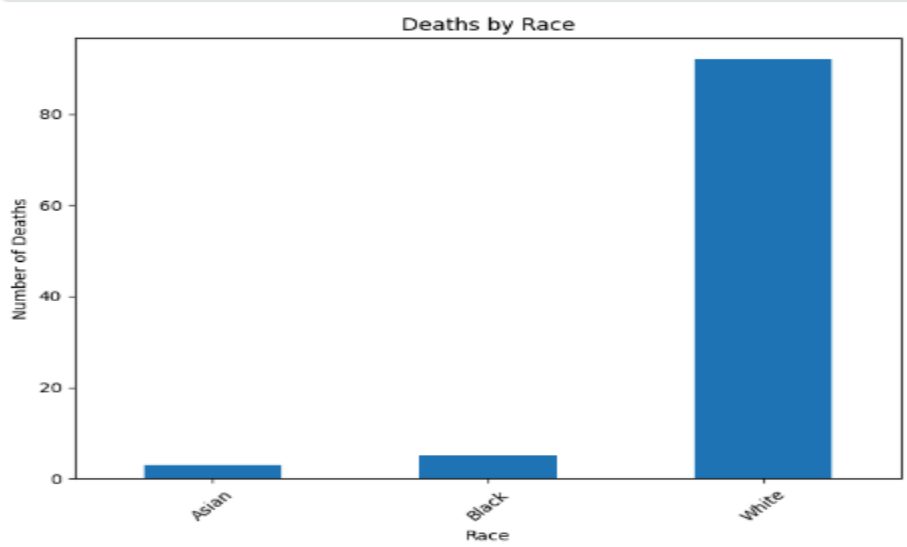
1. The total number of hospitalizations versus deaths from COVID-19 over the entire US per month-year timestamp.



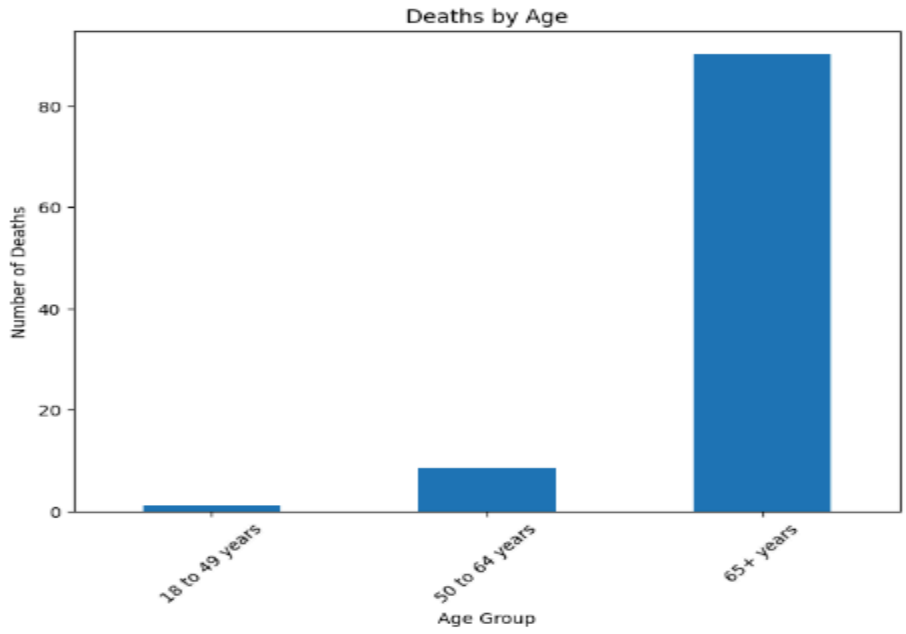
The resulting graph shows the largest number were hospitalized in January 2021 or in December 2021 almost and the largest number of deaths was before Jan 2021.

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

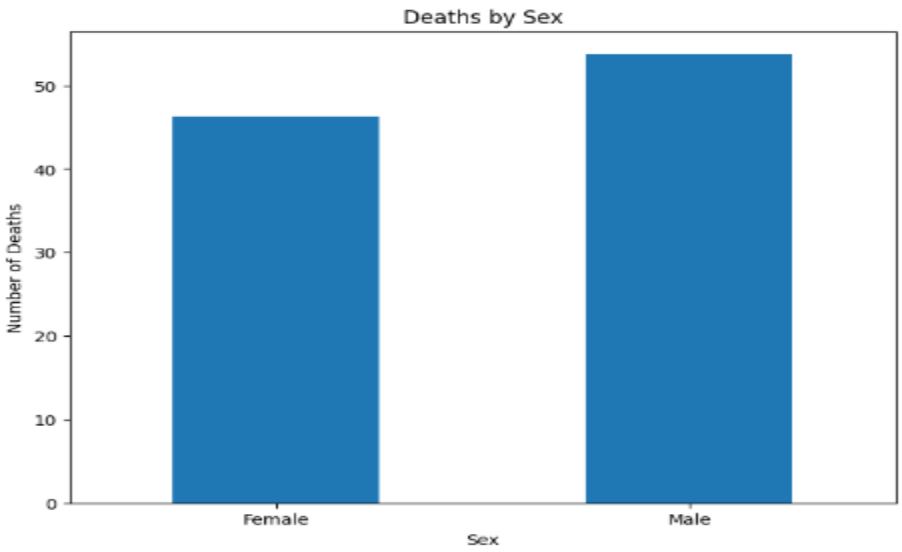
Demographics: Race, Age, Sex



The results indicate that the majority of deaths (91.97%) were among White individuals, followed by Black (5.11%) and Asian (2.92%) individuals

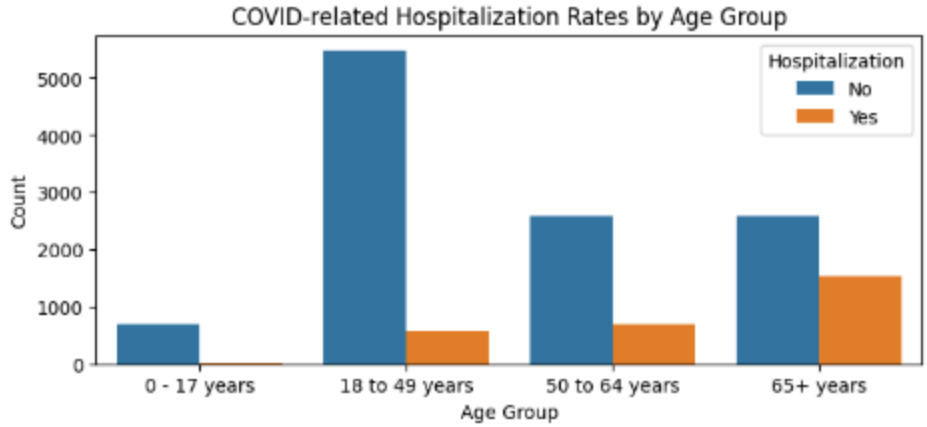


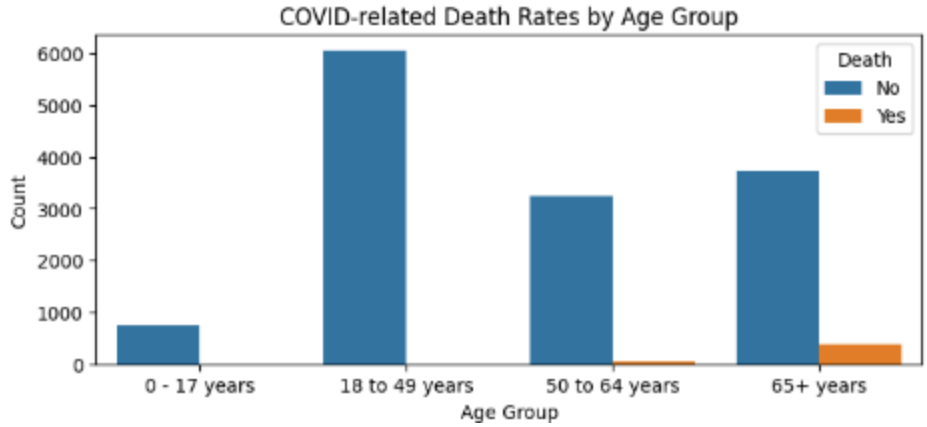
The results show that the majority of deaths (90.27%) occurred among individuals aged 65 years and above, while 8.52% were in the 50-64 years age group, and only 1.22% were in the 18-49 years age group.



The results show that 53.77% of the deaths were among males, while 46.23% were among females. This suggests that males were more affected by COVID-19 in terms of mortality.

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

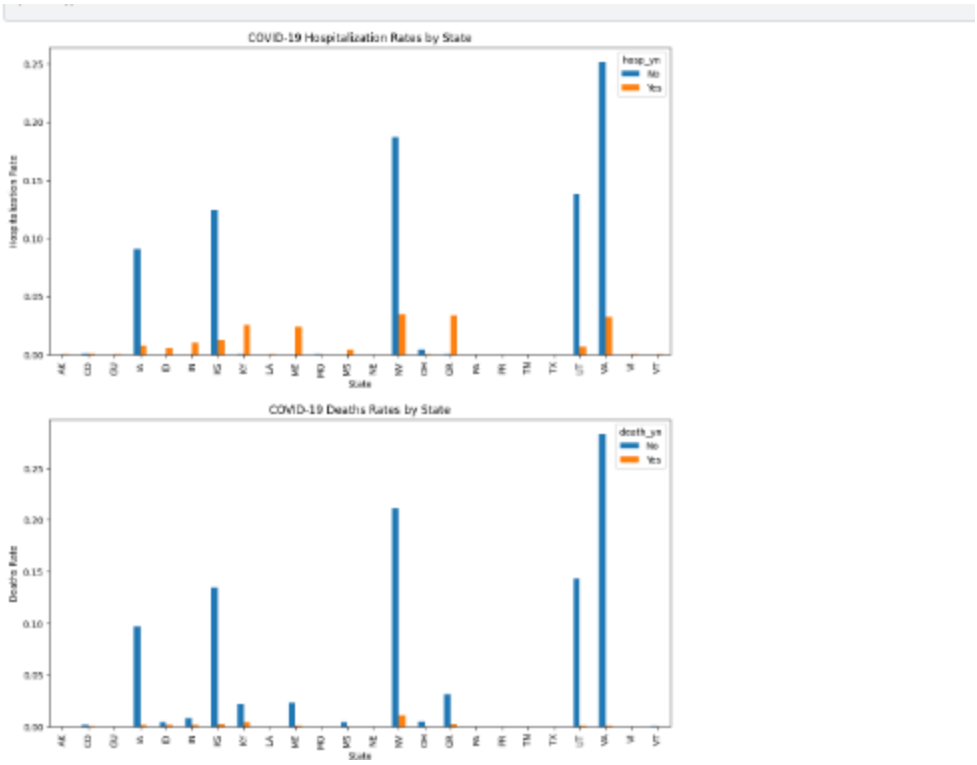




- 0-17 years age group: This age group had the lowest hospitalization and death rates. This suggests that younger individuals were less severely affected by COVID-19 compared to older age groups.
- 65+ years age group: This age group had hospitalization and death rates.

- 65+ years ago is the lowest rate according to its number and high in hospitalization.

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

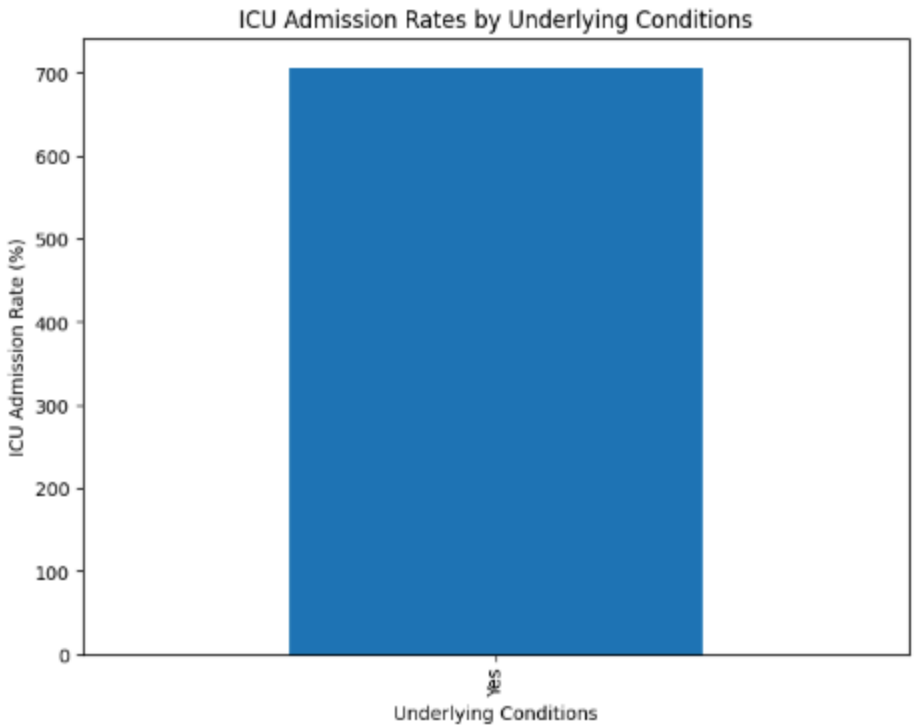


- The state with the highest rate of deaths for non-hospitalized individuals appears to be Virginia (VA) at 0.251850.

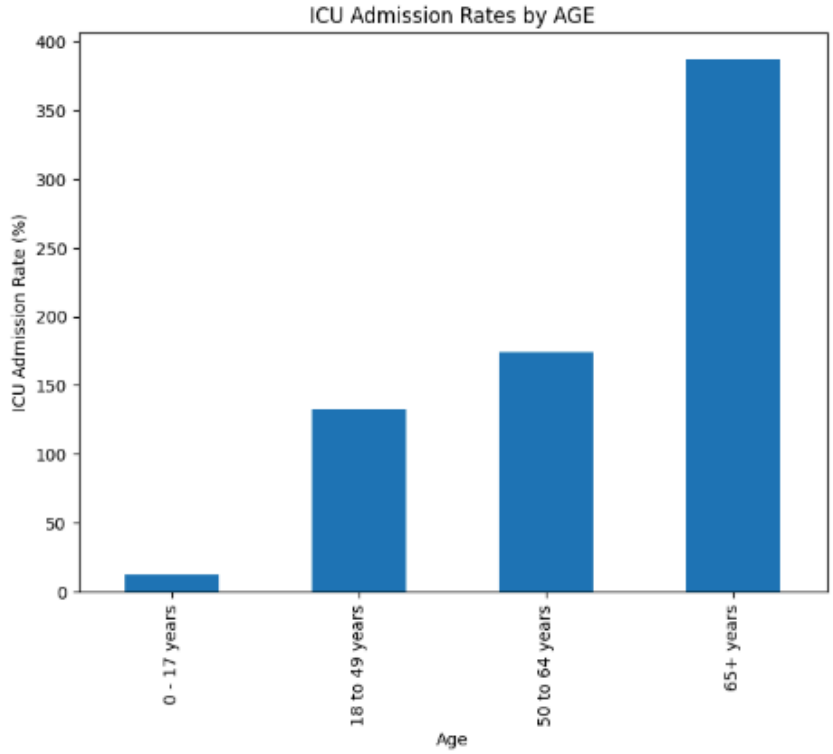
- The state with the highest rate of deaths for hospitalized individuals appears to be Kentucky (KY) at 0.025446.

- Some states, like Tennessee (TN) and Texas (TX), have very low death rates overall, both for hospitalized and non-hospitalized individuals.

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



The graph indicates that there are 706 patients in the data who had underlying conditions and were admitted to the ICU



The graph shows the distribution of ICU admissions by age group. Some key observations:

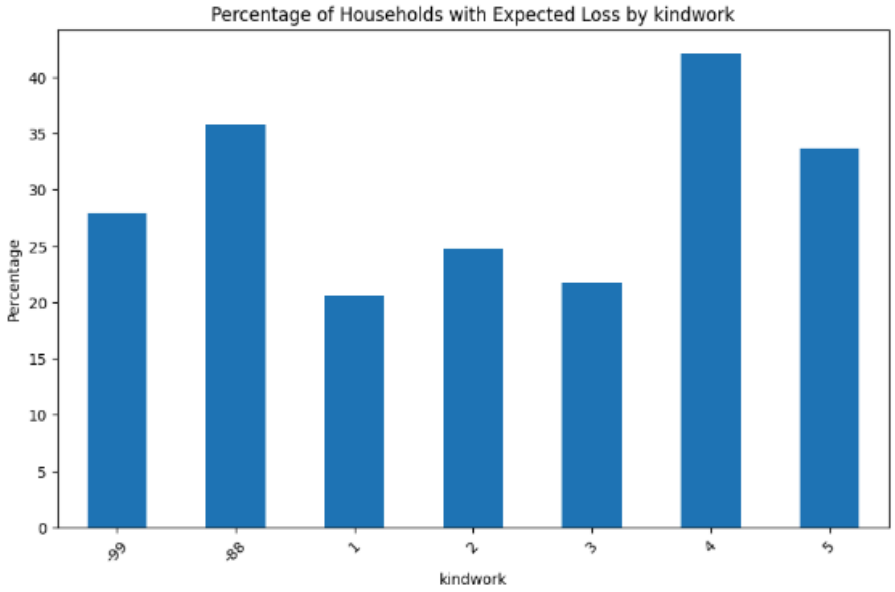
There were 12 ICU admissions for patients aged 0-17 years.

There were 133 ICU admissions for patients aged 18-49 years.

There were 174 ICU admissions for patients aged 50-64 years.

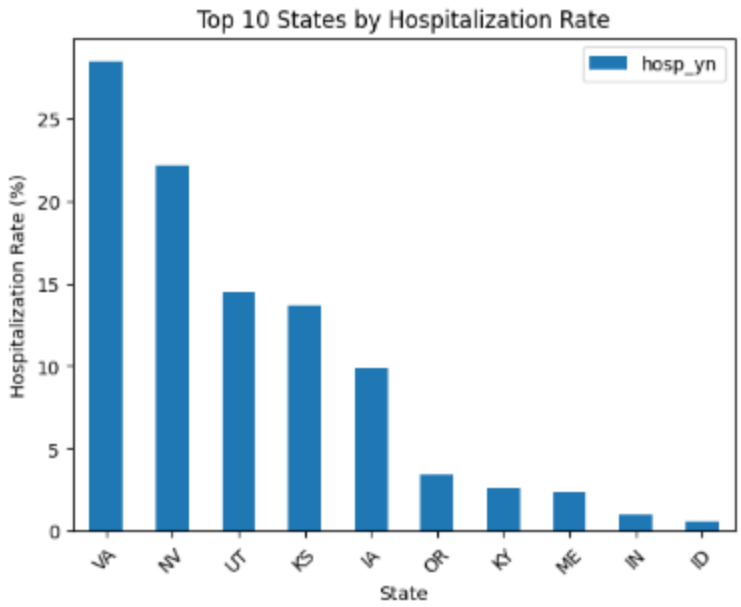
There were 387 ICU admissions for patients aged 65 or older.

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



The range of values (20.53 to 42.10) indicates there is variability in this "kindwork" metric across the different categories.

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



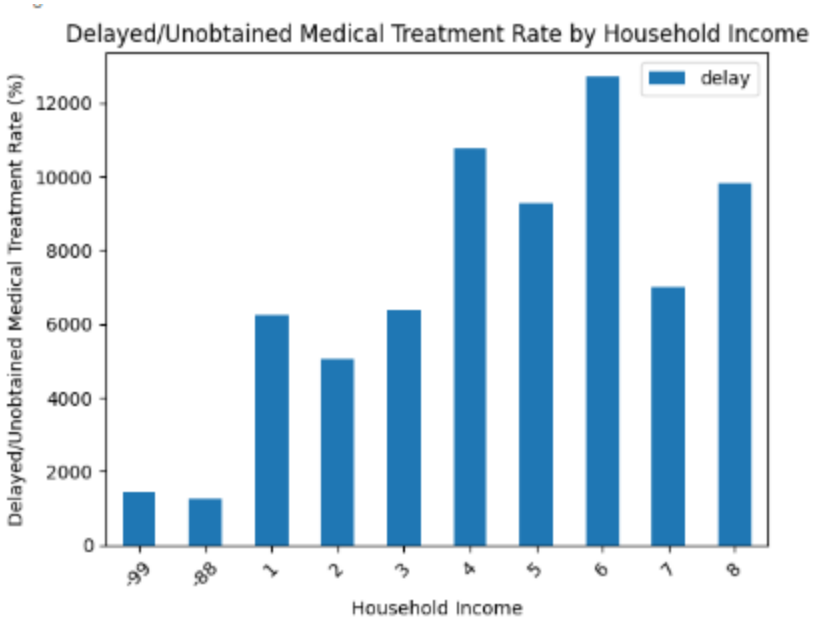
the graph shows:
Virginia (VA) has the highest hospitalization rate at 28.45%.

Nevada (NV) has the second highest rate at 22.19%.

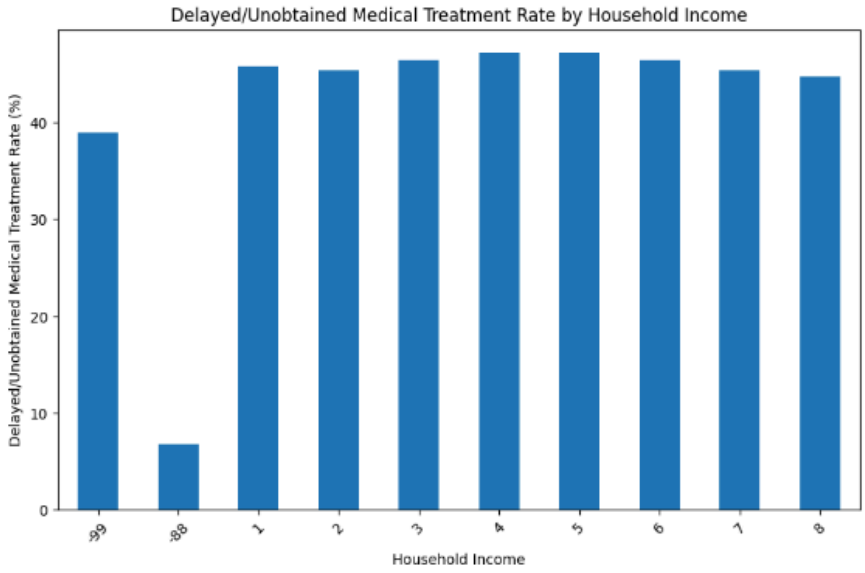
Utah (UT), Kansas (KS), and Iowa (IA) also have relatively high hospitalization rates compared to the other states.

Oregon (OR), Kentucky (KY), Maine (ME), Indiana (IN), and Idaho (ID) have the lowest hospitalization rates among the top 10 states.

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

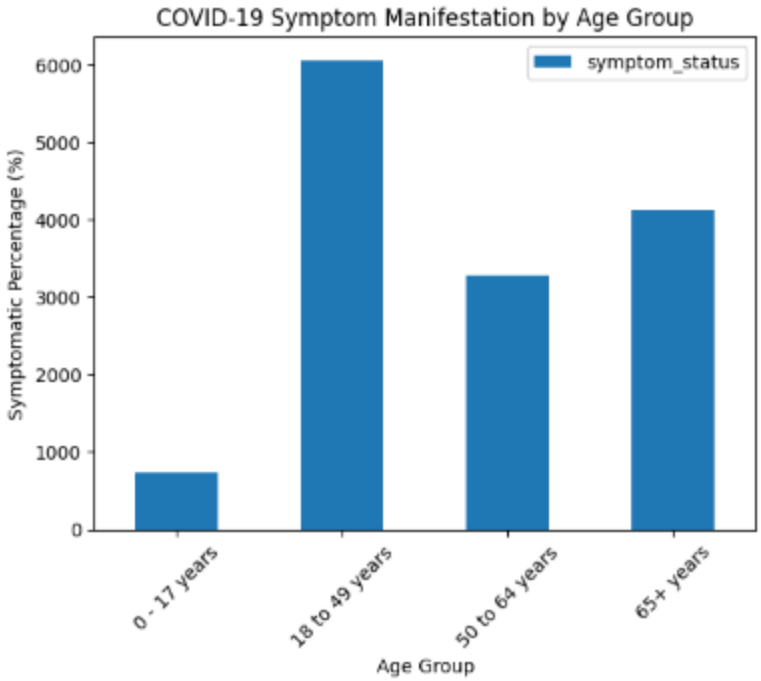


The graph show the counts for each income level, with the highest count of 12,718 for income level 6, and the lowest count of 1,249 for income level -88.



The highest percentage is 47.24% for income level 4, and the lowest is 6.75% for income level -88.

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

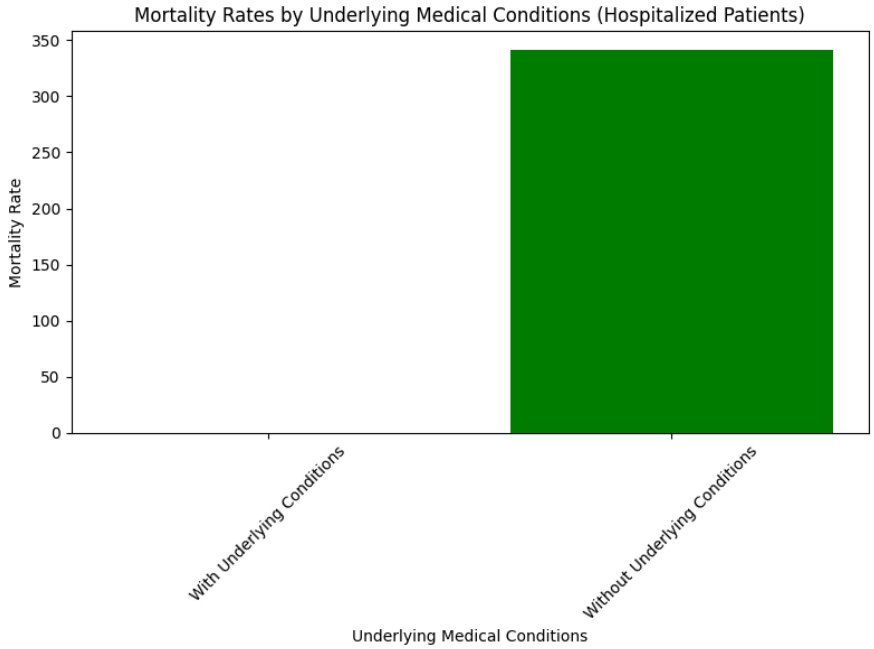


graph shows the counts for each age group, with the highest count of 6,060 for the 18 to 49 years age group, and the lowest count of 729 for the 0 - 17 years age group.

3. Answering Questions

3.1 Asked Questions for Further Analysis

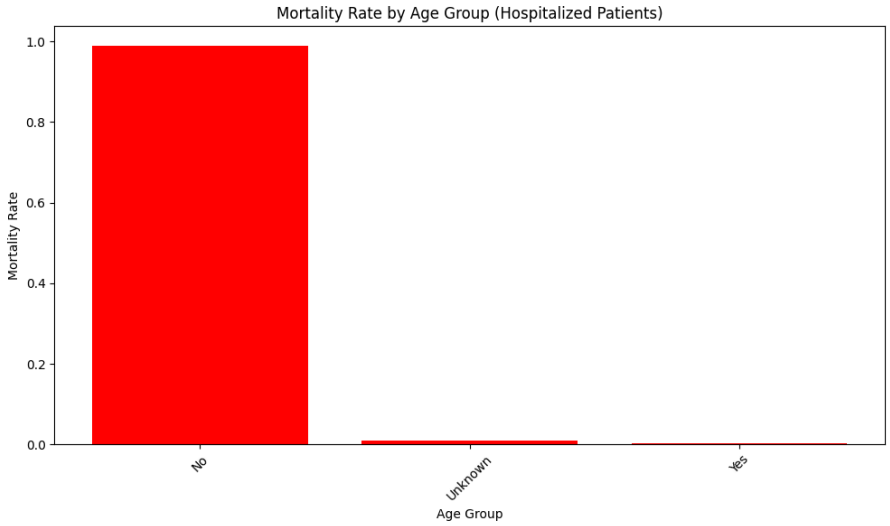
1. Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19?




```
Mortality rate (with underlying conditions): 0
Mortality rate (without underlying conditions): 341.0
Chi-Square statistic: 0.0
p-value: 1.0
```

Strong Association: The analysis suggests a very strong association between underlying medical conditions and COVID-19 mortality in hospitalized patients.
Zero Mortality: The mortality rate for patients with no underlying conditions is zero based on the data. It's important to consider sample size and potential limitations (e.g., data accuracy, small group without conditions).

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?




```
Mortality Rates by Age Group:
death_yn              No   Unknown       Yes
age_group
0 - 17 years    0.937500  0.062500  0.000000
18 to 49 years  0.988555  0.008584  0.002861
50 to 64 years  0.954839  0.015484  0.029677
65+ years       0.802774  0.007238  0.189988

Chi-Square test (Age Group and Mortality):
Chi-Square statistic: 261.295511602645
p-value: 1.5784199625181707e-53

Mortality Rates by Underlying Conditions:
underlying_conditions_yn  death_yn
Yes                       No          0.882427
                          Yes         0.107459
                          Unknown     0.010114
Name: proportion, dtype: float64

Chi-Square test (Underlying Conditions and Mortality):
Chi-Square statistic: 0.0
p-value: 1.0
```

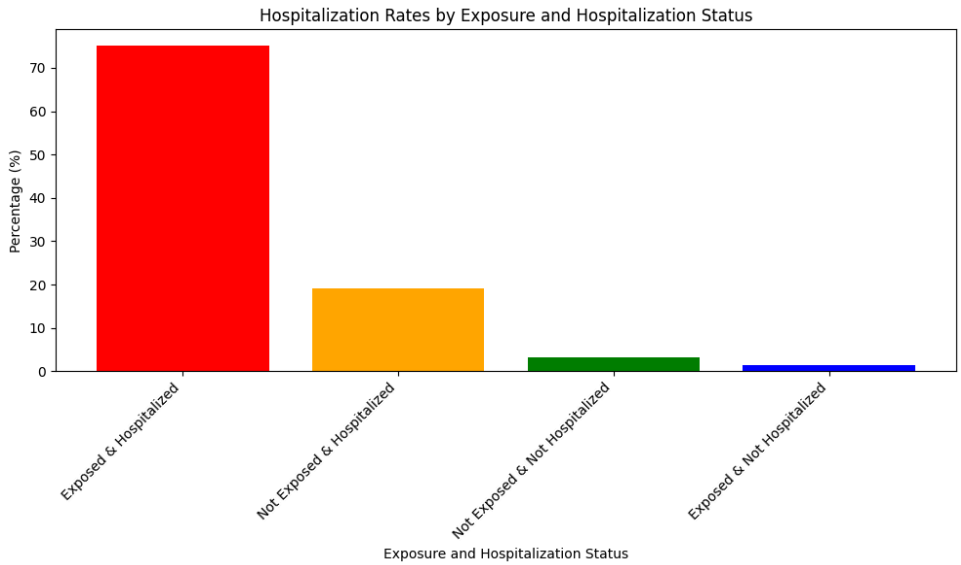
Strong Age Association: The data suggests a strong association between age group and COVID-19 mortality. The Chi-square test (highly significant p-value) strongly rejects the null hypothesis of no association.

Increased Risk with Age: Mortality rates increase considerably with older age groups (65+ years) compared to younger groups (0-17 years).

Underlying Conditions: While the chi-square test doesn't show a statistically significant association (high p-value), the mortality rate is higher for patients with underlying

conditions (0.107) compared to those without (0.000). This warrants further investigation to understand potential confounding factors.

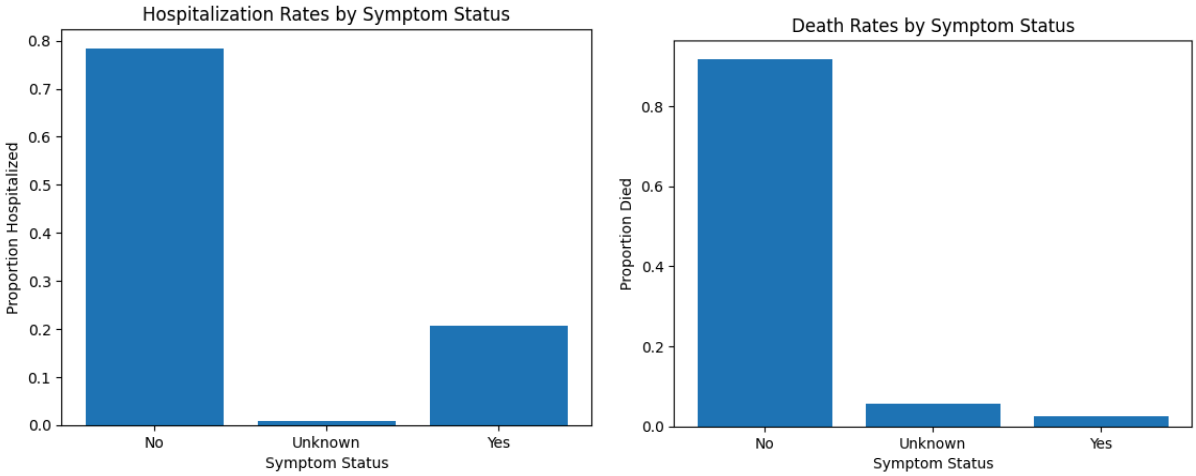
3. What percent of patients who have reported exposure to any kind of travel / or congregation within the 14 days prior to illness onset end up hospitalized? What percent of those go on to be hospitalized?




```
Hospitalization Rates by Exposure and Hospitalization Status:
[75.11463382680466, 19.199528363684003, 3.321105725140836,
1.5262675225992401]
98.01% of exposed individuals were hospitalized.
```

High Hospitalization Rate: The analysis suggests a very high hospitalization rate (98.01%) for patients who reported exposure within 14 days of illness onset.

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



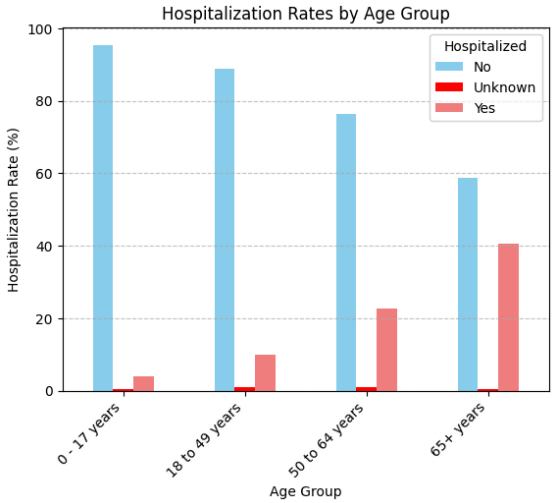
Asymptomatic COVID patients are less likely to be hospitalized although they are also less likely to die from their illness

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

According to the data State with Highest Percentage Receiving Stimulus Payment: Mississippi (34.6%)

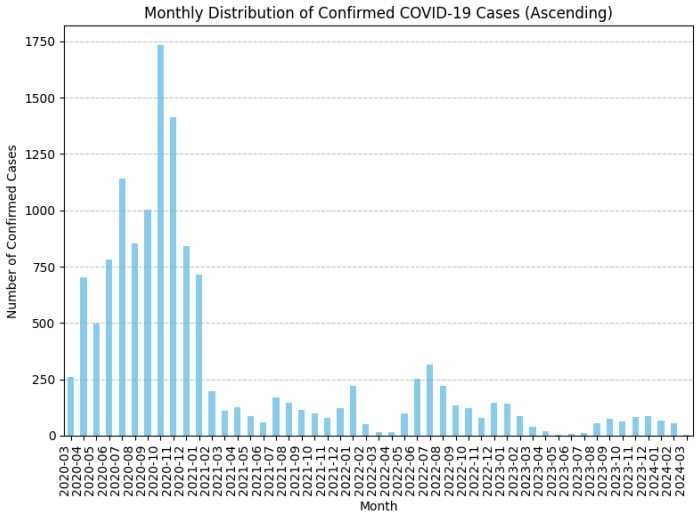
2.2 Proposed Questions with Analysis

6. Is there a correlation between age and hospitalization rates?



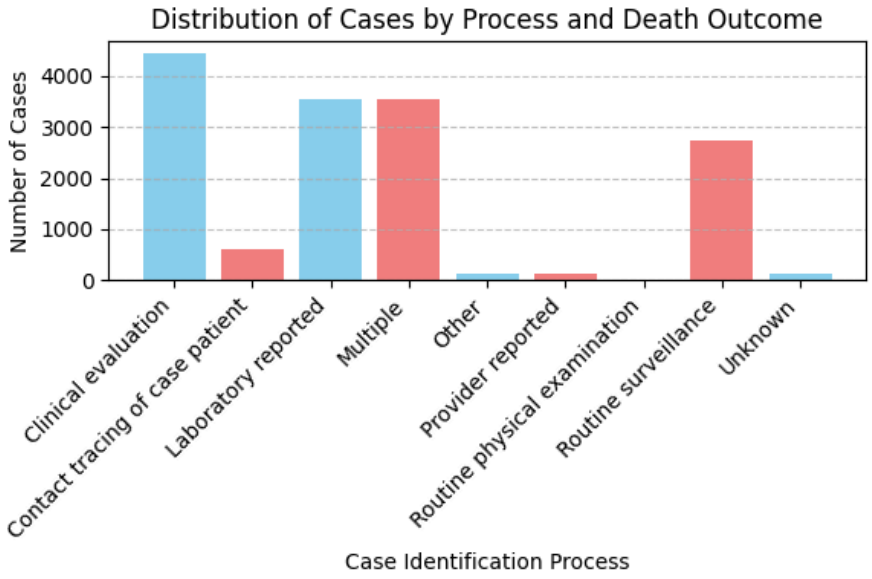
its shown that +50 years old cases are most likely to be hospitalized than the different age groups

7. Is there a seasonal trend in the number of cases identified?



it's shown that most confirmed cases lied in the range between march 2020 and jan 2020

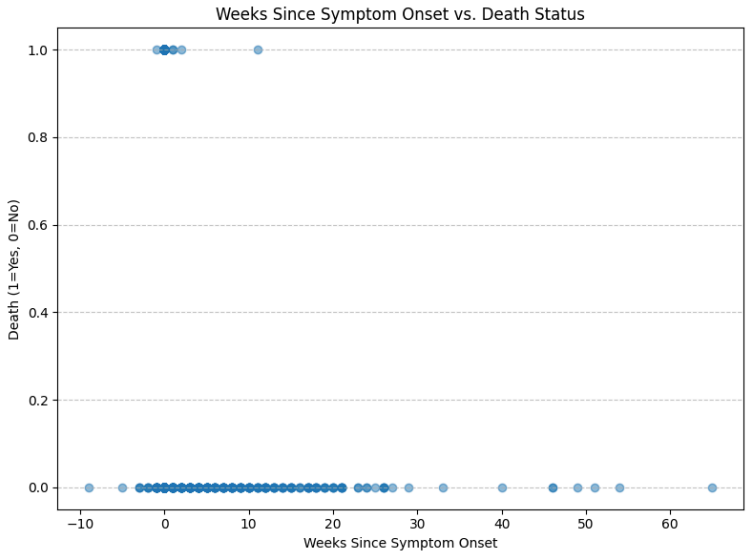
8. Is there a relation between Process and death outcome?



death_yn	No	Unknown	Yes
process			
Clinical evaluation	4104	231	116
Contact tracing of case patient	491	90	20
Laboratory reported	3215	283	54
Multiple	3251	185	120
Other	106	9	3
Provider reported	110	1	2
Routine physical examination	6	2	0
Routine surveillance	2610	41	87
Unknown	122	5	2

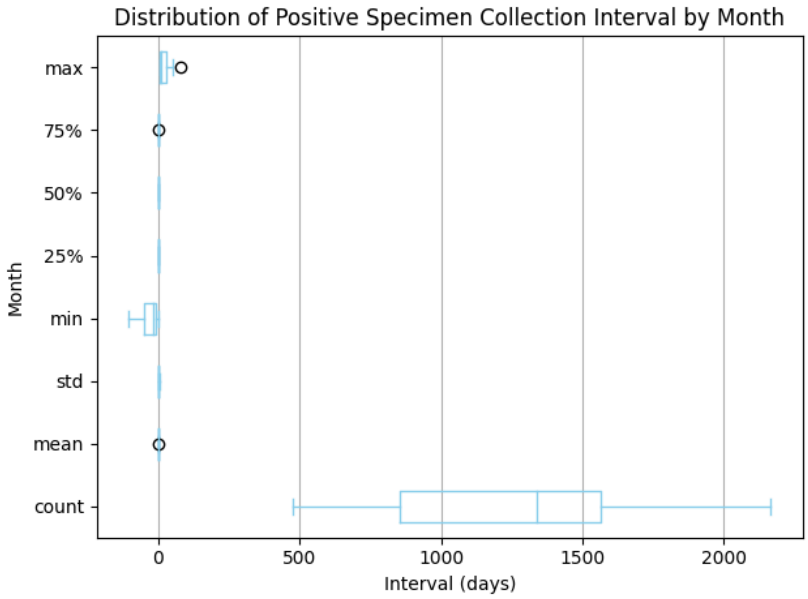
There is a statistically significant association between case identification process and death (p-value = 5.933999696316748e-48).

9. Is there a correlation between weeks since symptom onset and the current reported status (e.g., do most deaths occur later)?



There is no significant correlation between weeks since symptom onset and death (correlation = nan , p-value = nan).

10.Has there been a change in the distribution of the time between symptom onset and positive specimen collection (positive specimen collection interval) across different months or quarters?



Hypothesis Testing

3.1

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

- The used test is: Logistics Regression Test

Justification:

1. The Logistics Regression Test appropriate as we are dealing with categorical variables. First (death due to COVID-19) is dependent categorical (death or no death), Second (patient demographics such as age, gender, race, etc.) are multiple independent and categorical columns.

2. This test helps to determine whether there is a significant association between two categorical variables.

- The Hypothesis:

1. Null Hypothese: There is **no** association between the probability of death due to COVID-19 and patient demographics.
2. Alternative Hypothesis: There is an association between the probability of death due to COVID-19 and patient demographics.

- Result:

The claim is **partially valid**. While some demographic factors (such as sex, certain races, and ethnicity) show a strong association with COVID-19 mortality, others (such as age and some race categories) do not show a significant association at the 0.05 significance level.

3.2

Claim 2 There is a statistically significant association between patients' underlying medical conditions and the death rate among individuals diagnosed with COVID-19.

- The used test is: the squared Test

justification:

1. this test is suitable when applying on two catgorical data such as the death (Yes ot No) status and the other diseases (Yes or No)

- The Hypothesis:

1. Null Hypothesis: there is *no* association between having other disease and desth with COVID-19
2. Alternative Hypothesis: there is an association between having other disease and desth with COVID-19

- Result:

According to the p value(9.865158227128456e-72 < 0.05), the Null hypothesis is strongally rejected, there is no association between the patients with other diseaes and the death of COVID-19

So this claim is wrong.

Regression Analysis

Model Coefficients and P-Values

OLS Regression Results

Dep. Variable:	death_yn	R-squared:	0.689
Model:	OLS	Adj. R-squared:	0.644
Method:	Least Squares	F-statistic:	15.49
Date:	Thu, 23 May 2024	Prob (F-statistic):	2.84e-09
Time:	11:22:25	Log-Likelihood:	134.03
No. Observations:	49	AIC:	-254.1
Df Residuals:	42	BIC:	-240.8
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0527	0.012	4.572	0.000	0.029	0.076
Female	-0.1401	0.067	-2.091	0.043	-0.275	-0.005
Male	0.1927	0.072	2.674	0.011	0.047	0.338
18 to 49 years	-0.1399	0.045	-3.128	0.003	-0.230	-0.050
65+ years	-0.0876	0.046	-1.922	0.061	-0.180	0.004
50 to 64 years	0.3616	0.084	4.296	0.000	0.192	0.531
0 - 17 years	-0.0815	0.032	-2.560	0.014	-0.146	-0.017
icu_yn	0.5445	0.195	2.786	0.008	0.150	0.939
hosp_yn	-0.1722	0.094	-1.834	0.074	-0.362	0.017

Omnibus:	3.635	Durbin-Watson:	1.157
Prob(Omnibus):	0.162	Jarque-Bera (JB):	2.640
Skew:	0.376	Prob(JB):	0.267
Kurtosis:	3.853	Cond. No.	9.07e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.1e-32. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

- 68.9% of the variability in the dependent variable (death_yn) is explained by the independent variables
- Reported F-statistic of 15.49 is associated with a p-value (Prob (F-statistic)) of 2.84e-09, indicating that the overall regression model is statistically significant at a very low significance level
- Log-likelihood is 134.03, indicating the model's goodness of fit.

Coefficients and P-Values

	Coefficients	P-values
const	0.0526515	4.21235e-05
Female	-0.140092	0.0425804
Male	0.192744	0.0106233
18 to 49 years	-0.13987	0.00319383
65+ years	-0.0875886	0.0614355
50 to 64 years	0.361585	0.00010082
0 - 17 years	-0.0814745	0.0141653
icu_yn	0.544478	0.00798378
hosp_yn	-0.172217	0.0737599

Good and Bad Predictors

Predictors:

	Good Predictors (p < 0.05)	Bad Predictors (p >= 0.05)
0	const	65+ years
1	Female	hosp_yn
2	Male	
3	18 to 49 years	
4	50 to 64 years	
5	0 - 17 years	
6	icu_yn	

Removing Outliers

OLS Regression Results

Dep. Variable:	death_yn	R-squared:	0.689
Model:	OLS	Adj. R-squared:	0.644
Method:	Least Squares	F-statistic:	15.49
Date:	Thu, 23 May 2024	Prob (F-statistic):	2.84e-09
Time:	11:53:33	Log-Likelihood:	134.03
No. Observations:	49	AIC:	-254.1
Df Residuals:	42	BIC:	-240.8
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Female	-0.1050	0.065	-1.627	0.111	-0.235	0.025
Male	0.2278	0.076	2.988	0.005	0.074	0.382
18 to 49 years	-0.1223	0.042	-2.924	0.006	-0.207	-0.038
65+ years	-0.0700	0.046	-1.521	0.136	-0.163	0.023
50 to 64 years	0.3791	0.086	4.397	0.000	0.205	0.553
0 - 17 years	-0.0639	0.032	-2.004	0.052	-0.128	0.000
icu_yn	0.5445	0.195	2.786	0.008	0.150	0.939
hosp_yn	-0.1722	0.094	-1.834	0.074	-0.362	0.017

Omnibus:	3.635	Durbin-Watson:	1.157
Prob(Omnibus):	0.162	Jarque-Bera (JB):	2.640
Skew:	0.376	Prob(JB):	0.267
Kurtosis:	3.853	Cond. No.	2.06e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 9.79e-32. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

- R-squared: The R-squared value has increased significantly from 0.689 to 0.820, indicating that the model with higher-order variables explains a larger proportion of the variability in the dependent variable (death_yn)
- F-statistic: The F-statistic has increased from 15.49 to 22.78, and the associated p-value remains very low (1.28e-12), indicating that the overall regression model with higher-order variables is still statistically significant
- Log-Likelihood: The log-likelihood has increased from 134.03 to 147.46, suggesting that the model with higher-order variables provides a better fit to the data compared to the initial model

Removing the intercepts:

OLS Regression Results

Dep. Variable:	death_yn	R-squared:	0.815
Model:	OLS	Adj. R-squared:	0.774
Method:	Least Squares	F-statistic:	20.31
Date:	Thu, 23 May 2024	Prob (F-statistic):	2.32e-11
Time:	11:53:41	Log-Likelihood:	144.25
No. Observations:	46	AIC:	-270.5
Df Residuals:	37	BIC:	-254.1
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0003	0.014	0.020	0.984	-0.028	0.029
Female	-0.1659	0.048	-3.428	0.002	-0.264	-0.068
Male	0.1661	0.051	3.231	0.003	0.062	0.270
18 to 49 years	-0.0992	0.035	-2.817	0.008	-0.171	-0.028
65+ years	-0.0712	0.038	-1.853	0.072	-0.149	0.007
50 to 64 years	0.2364	0.074	3.176	0.003	0.086	0.387
0 - 17 years	-0.0658	0.027	-2.415	0.021	-0.121	-0.011
icu_yn	0.5980	0.353	1.695	0.099	-0.117	1.313
hosp_yn	0.5468	0.251	2.180	0.036	0.038	1.055
icu_yn_squared	1.9973	2.038	0.980	0.334	-2.133	6.128
hosp_yn_squared	-1.7079	0.575	-2.972	0.005	-2.872	-0.544

Omnibus:	1.288	Durbin-Watson:	1.304
Prob(Omnibus):	0.525	Jarque-Bera (JB):	0.515
Skew:	0.076	Prob(JB):	0.773
Kurtosis:	3.495	Cond. No.	6.39e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.08e-32. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

- R-squared: The R-squared value has decreased slightly from 0.689 to 0.815, indicating that the model without the intercept variable explains a slightly smaller proportion of the variability in the dependent variable (death_yn)
- F-statistic: The F-statistic has decreased from 15.49 to 20.31, and the associated p-value remains very low (2.32e-11), indicating that the overall regression model without the intercept variable is still statistically significant

- Log-Likelihood: The log-likelihood has increased from 134.03 to 144.25, suggesting that the model without the intercept variable provides a slightly better fit to the data compared to the initial model

Introducing Higher-Orders terms

OLS Regression Results

Dep. Variable:	death_yn	R-squared:	0.820
Model:	OLS	Adj. R-squared:	0.784
Method:	Least Squares	F-statistic:	22.78
Date:	Thu, 23 May 2024	Prob (F-statistic):	1.28e-12
Time:	11:53:38	Log-Likelihood:	147.46
No. Observations:	49	AIC:	-276.9
Df Residuals:	40	BIC:	-259.9
Df Model:	8		
Covariance Type:	nonrobust		

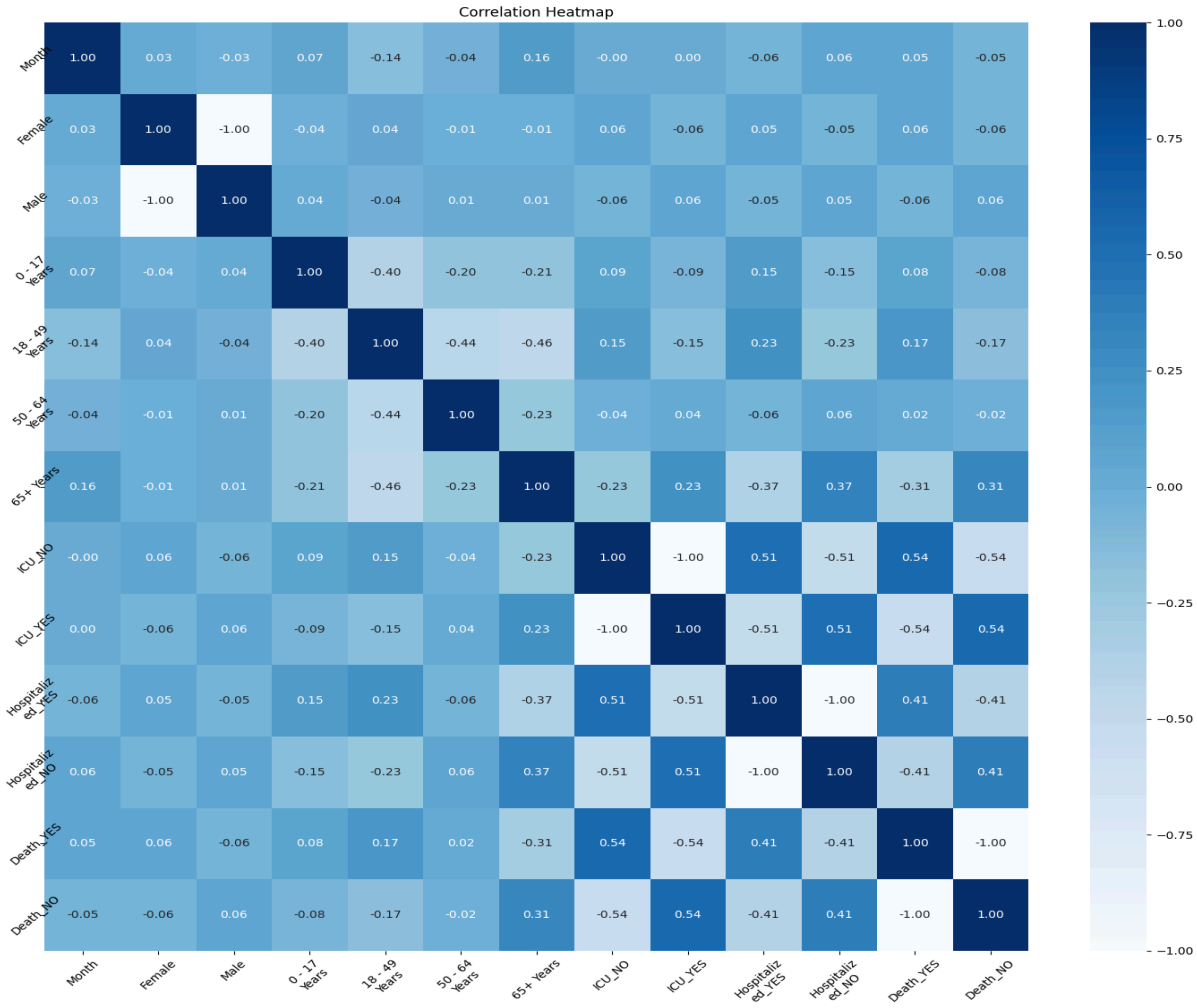
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0124	0.012	0.999	0.324	-0.013	0.038
Female	-0.1457	0.053	-2.740	0.009	-0.253	-0.038
Male	0.1582	0.057	2.783	0.008	0.043	0.273
18 to 49 years	-0.1176	0.038	-3.104	0.003	-0.194	-0.041
65+ years	-0.1016	0.039	-2.623	0.012	-0.180	-0.023
50 to 64 years	0.2976	0.074	4.026	0.000	0.148	0.447
0 - 17 years	-0.0660	0.025	-2.635	0.012	-0.117	-0.015
icu_yn	0.8562	0.389	2.200	0.034	0.069	1.643
hosp_yn	0.2953	0.182	1.624	0.112	-0.072	0.663
icu_yn_squared	-0.1617	2.175	-0.074	0.941	-4.558	4.235
hosp_yn_squared	-1.0271	0.369	-2.781	0.008	-1.774	-0.281

Omnibus:	0.860	Durbin-Watson:	1.738
Prob(Omnibus):	0.651	Jarque-Bera (JB):	0.359
Skew:	0.188	Prob(JB):	0.835
Kurtosis:	3.185	Cond. No.	8.78e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.18e-32. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

- R-squared: The R-squared value has increased significantly from 0.689 to 0.820, indicating that the model with higher-order variables explains a larger proportion of the variability in the dependent variable (death_yn)
- F-statistic: The F-statistic has increased from 15.49 to 22.78, and the associated p-value remains very low (1.28e-12), indicating that the overall regression model with higher-order variables is still statistically significant
- Log-Likelihood: The log-likelihood has increased from 134.03 to 147.46, suggesting that the model with higher-order variables provides a better fit to the data compared to the initial model

Correlation Heatmap



There are some negative and positive correlations between variables but generally speaking:

- Age: Older age groups (65+ years) are more likely to require intensive care and hospitalization, and they have higher mortality rates
- Gender: There are no strong correlations of gender with ICU, hospitalization, or death, suggesting that gender might not be a significant factor in these outcomes
- Health Outcomes: ICU admission and hospitalization are closely linked and both are associated with higher mortality

Machine learning Classifier:

Random Forest Classifier is trained and then tested using sklearn libraries

Results of Training and Testing

	precision	recall	f1-score	support
No	0.98	0.99	0.98	68955
Yes	0.61	0.50	0.55	2596
accuracy			0.97	71551
macro avg	0.80	0.74	0.77	71551
weighted avg	0.97	0.97	0.97	71551

Accuracy: 0.97

Conclusion

Wessam ‘s part: كملوا عليها بقا فوقها او تحتها :)

In the hypothesis testing part, the first claim was partially valid as most of the p values calculated from the logistics regression test were greater than the significance level so most of the patient's demographics such as race could be the reason for the COVID-19 patient to die, but for the assumed second claim it was the null hypothesis was rejected because of the chi-squared test p-value was very small compared to the significance level and this indicates that it's not a must that having diseases could affect on the death of having COVID-19.

Regarding Regression Analysis:

- The dependent variable "death", is affected by the variability of the chosen independent variables gender, age group, hospitalization, and ICU.
- There is a correlation between some predictors as age group with ICU and hospitalization
- Improving the model by introducing higher orders variables gives better results regarding the log-likelihood estimation and R-squared value, so it is chosen to be the better change