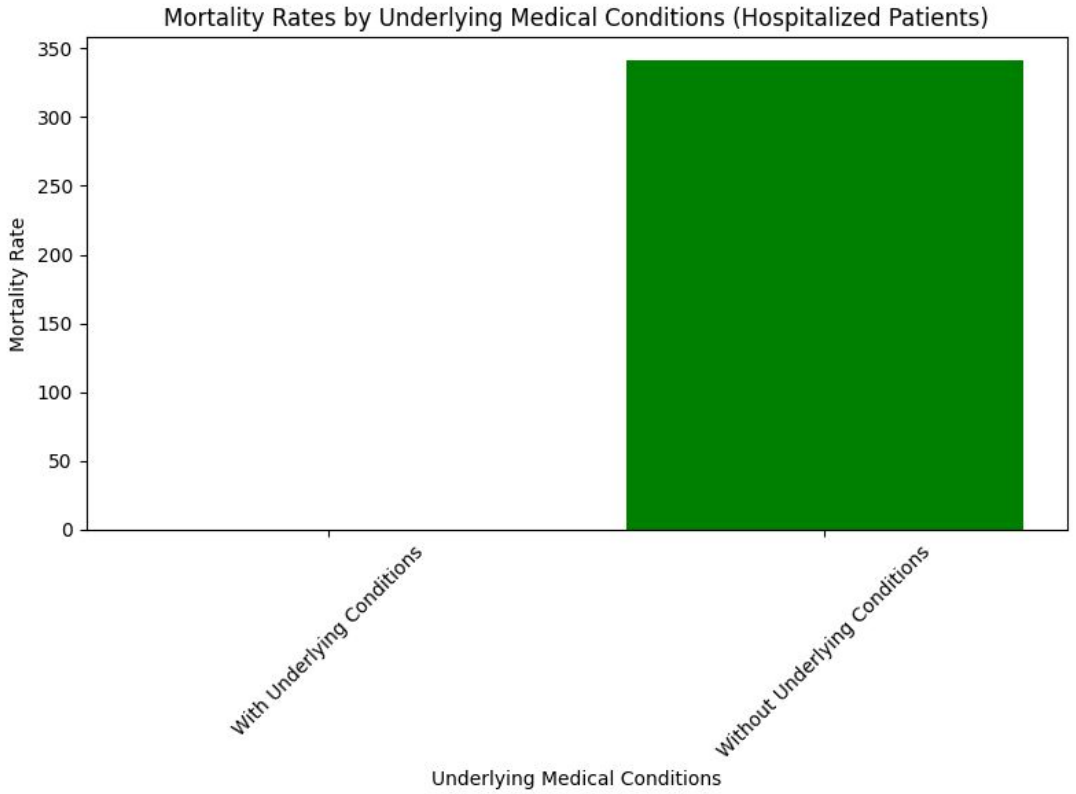
Analyzing U.S.COVID-19 Data

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

Explanatory Analysis

Explanatory Analysis

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)

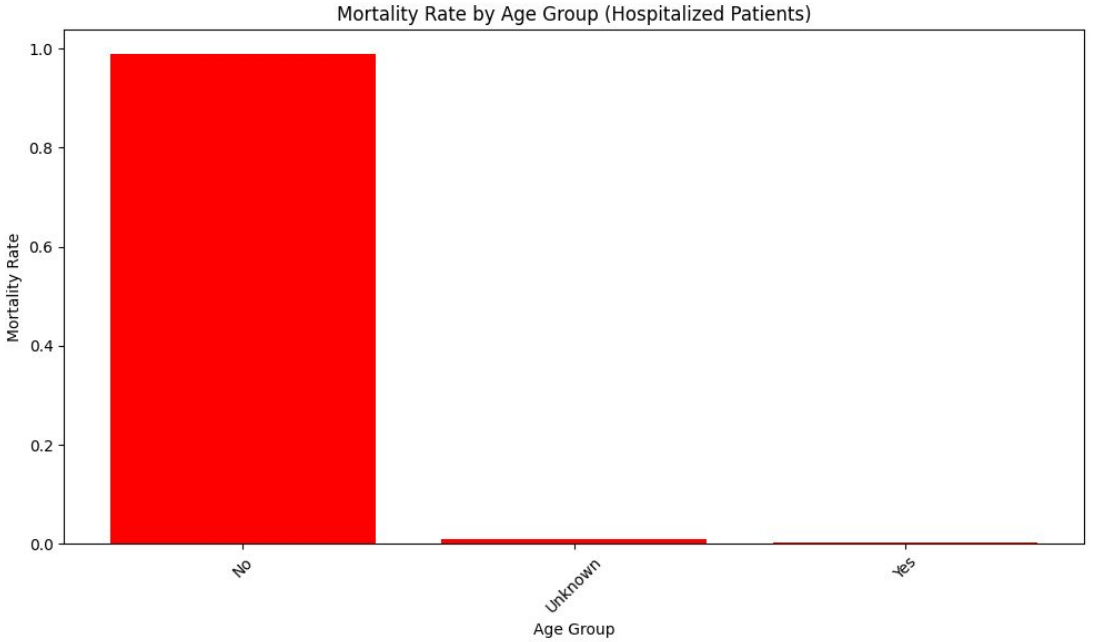


Explanatory Analysis

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.

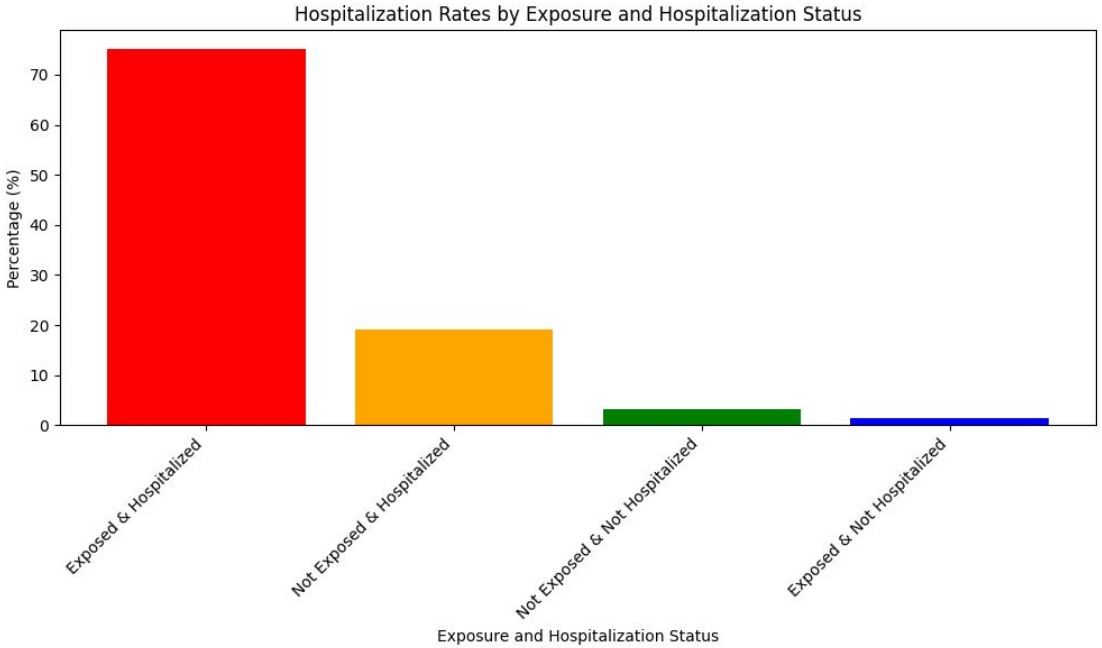
Increased Risk with Age: Mortality rates increase considerably with older age groups (65+ years) compared to younger groups (0-17 years).

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

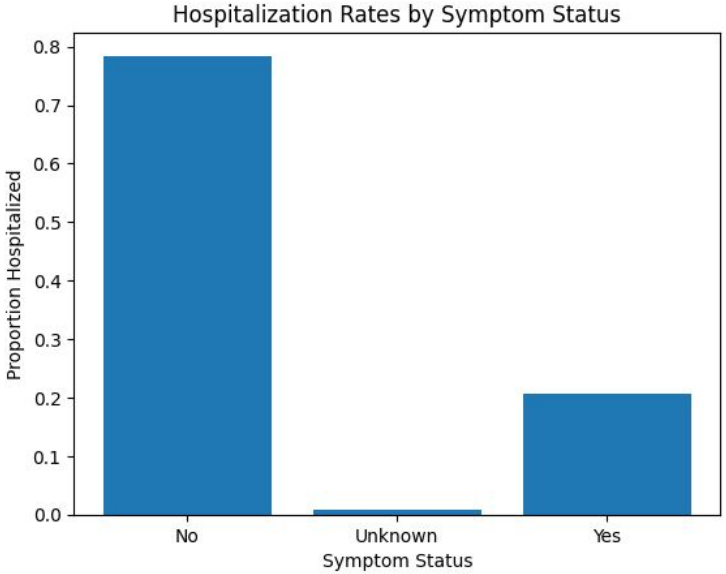


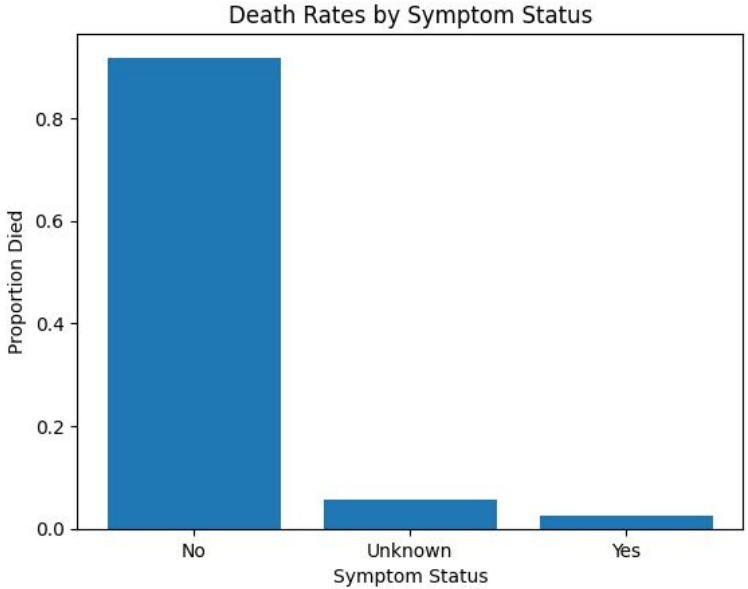
Explanatory Analysis

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



Explanatory Analysis



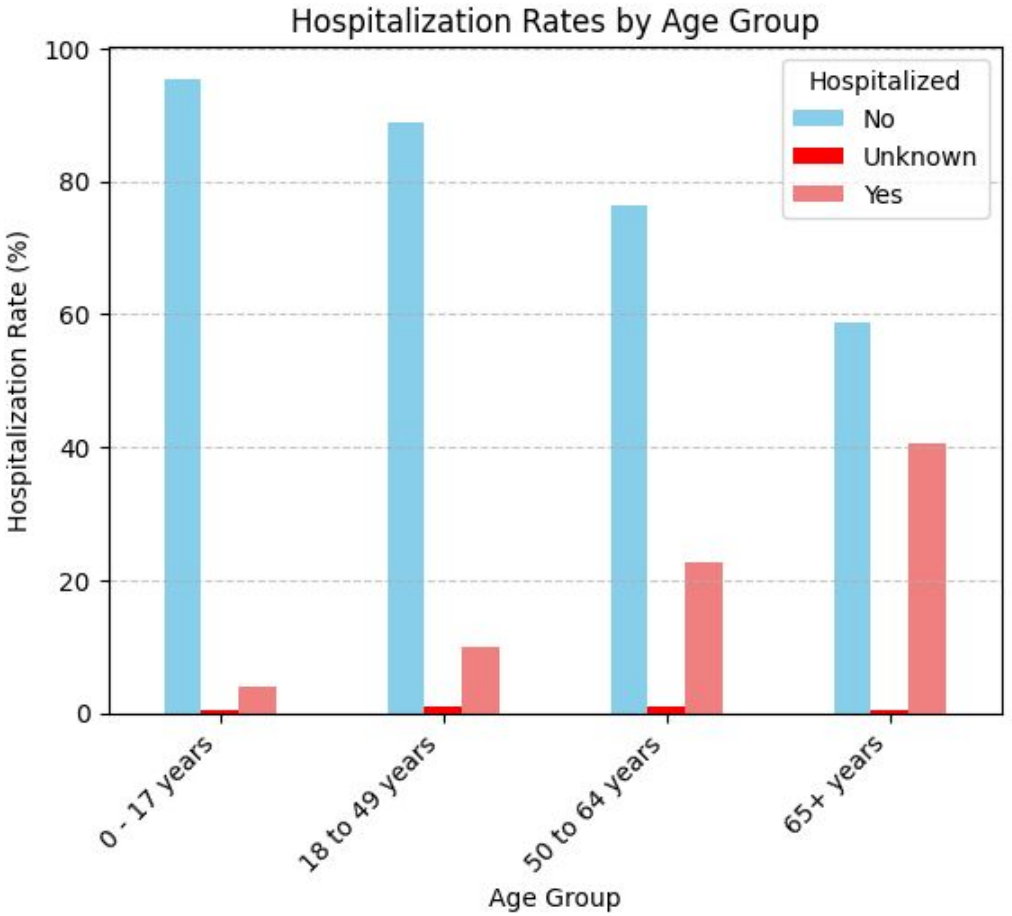


Asymptomatic COVID patients are less likely to be hospitalized although they are also less likely to die from their illness

Explanatory Analysis

Is there a correlation between age and hospitalization rates?

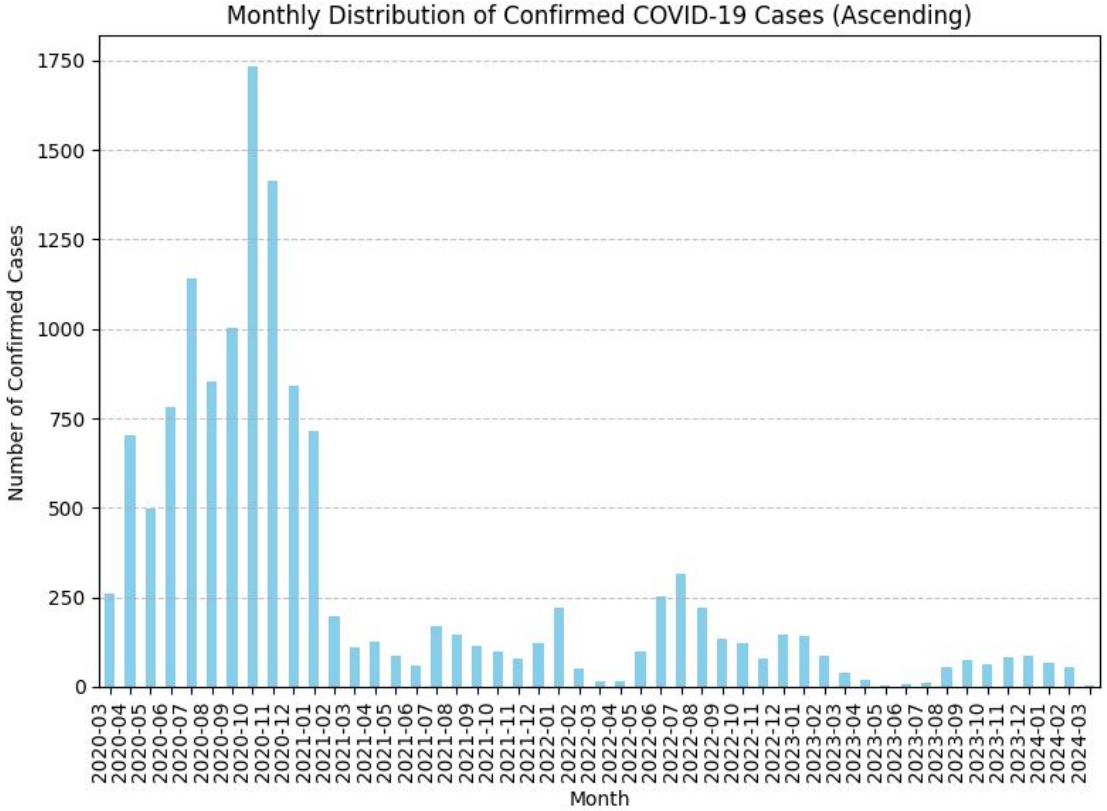
its shown that +50 years old cases are most likely to be hospitalized than the different age groups



Explanatory Analysis

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

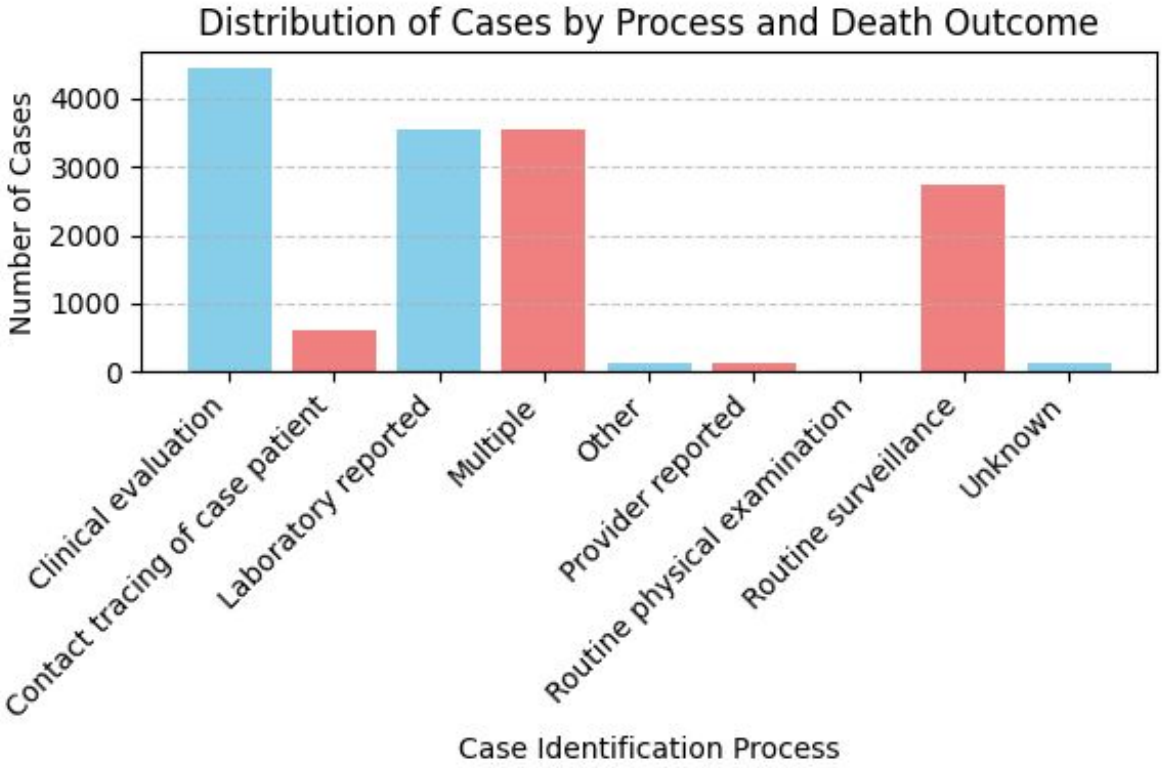
it's shown that most confirmed cases lied in the range between march 2020 and jan 2020



Explanatory Analysis

Is there a relation between Process an death outcome?

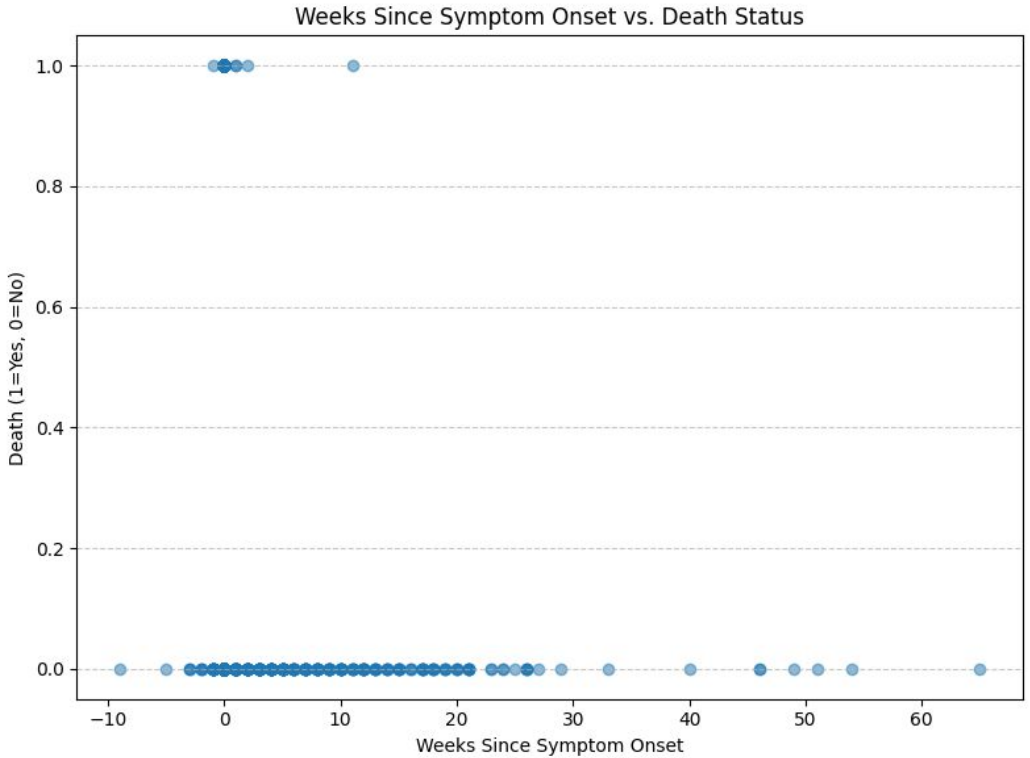
There is a statistically significant association between case identification process and death (p-value = 5.933999696316748e-48).



Explanatory Analysis

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

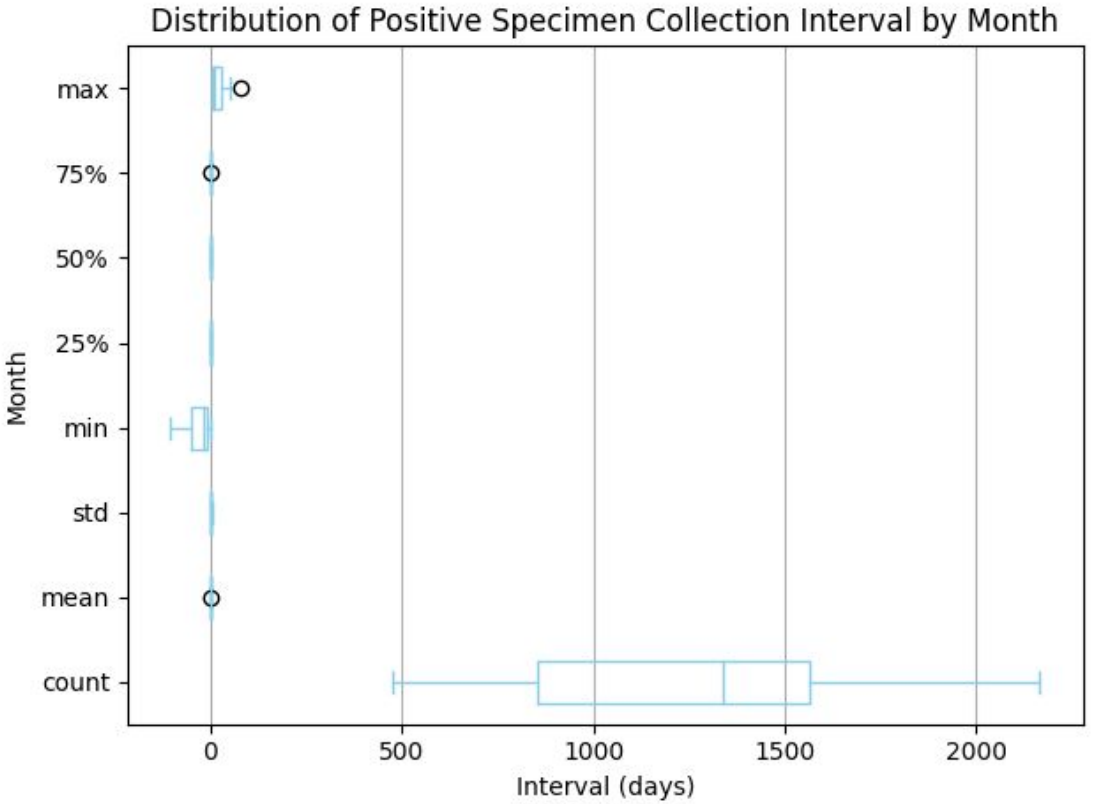
There is no significant correlation between weeks since symptom onset and death (correlation = nan , p-value = nan).



Explanatory Analysis

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?

No significant change shown



Answering Questions

Hypothesis Testing

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

For this claim the null hypothesis was **partially valid** as most of the columns was higher than the significance level (Alpha = 0.05)

```
P-values:
 const                                          9.330255e-01
age_group_18 to 49 years                        9.501193e-01
age_group_50 to 64 years                        9.421819e-01
age_group_65+ years                             9.348131e-01
sex_Male                                        0.000000e+00
race_Asian                                      4.988423e-10
race_Black                                      3.494668e-02
race_Multiple/Other                             1.288777e-29
race_Native Hawaiian/Other Pacific Islander     9.609803e-01
race_White                                      1.000775e-10
ethnicity_Non-Hispanic/Latino                   0.000000e+00
dtype: float64
```

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

For this claim, It was rejected as the p value was less that the significance value (Alpha = 0.05) So the Null Hypothesis was **rejected.**

```
p-value: 9.865158227128456e-72
```

Regression Analysis

Model Coefficients & 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

Coefficients & 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 & 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

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

Introducing Higher-Orders terms

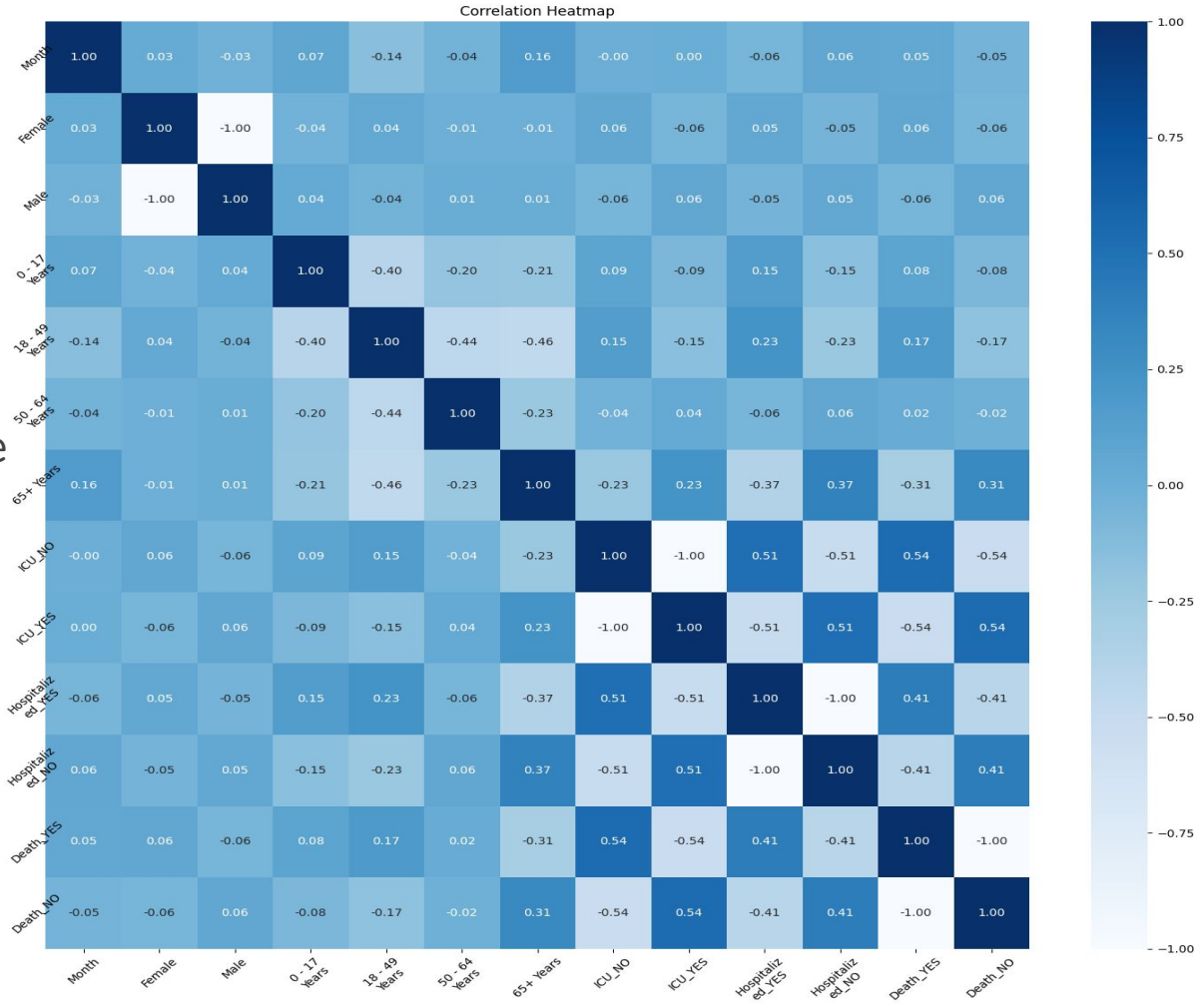
OLS Regression Results

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

- The Opposite Heatmap shows some correlation between predictors.
- Correlation value ranges from [-1 , 1]



Machine Learning

Results of Training & 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

Thanks !