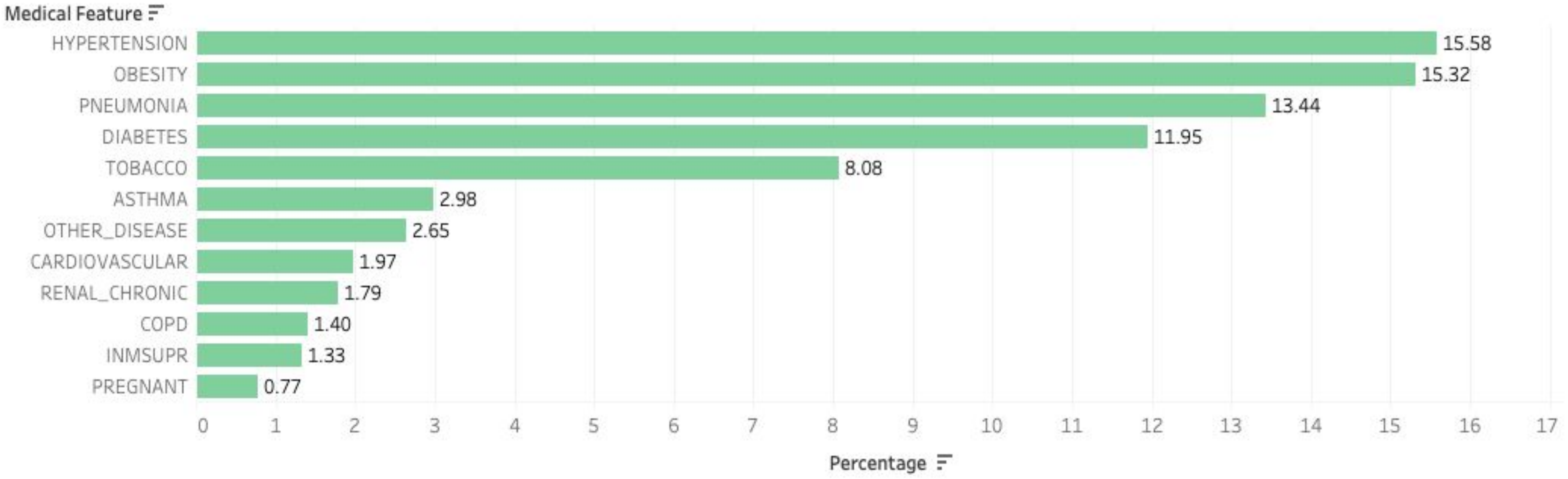
COVID-19

Overview

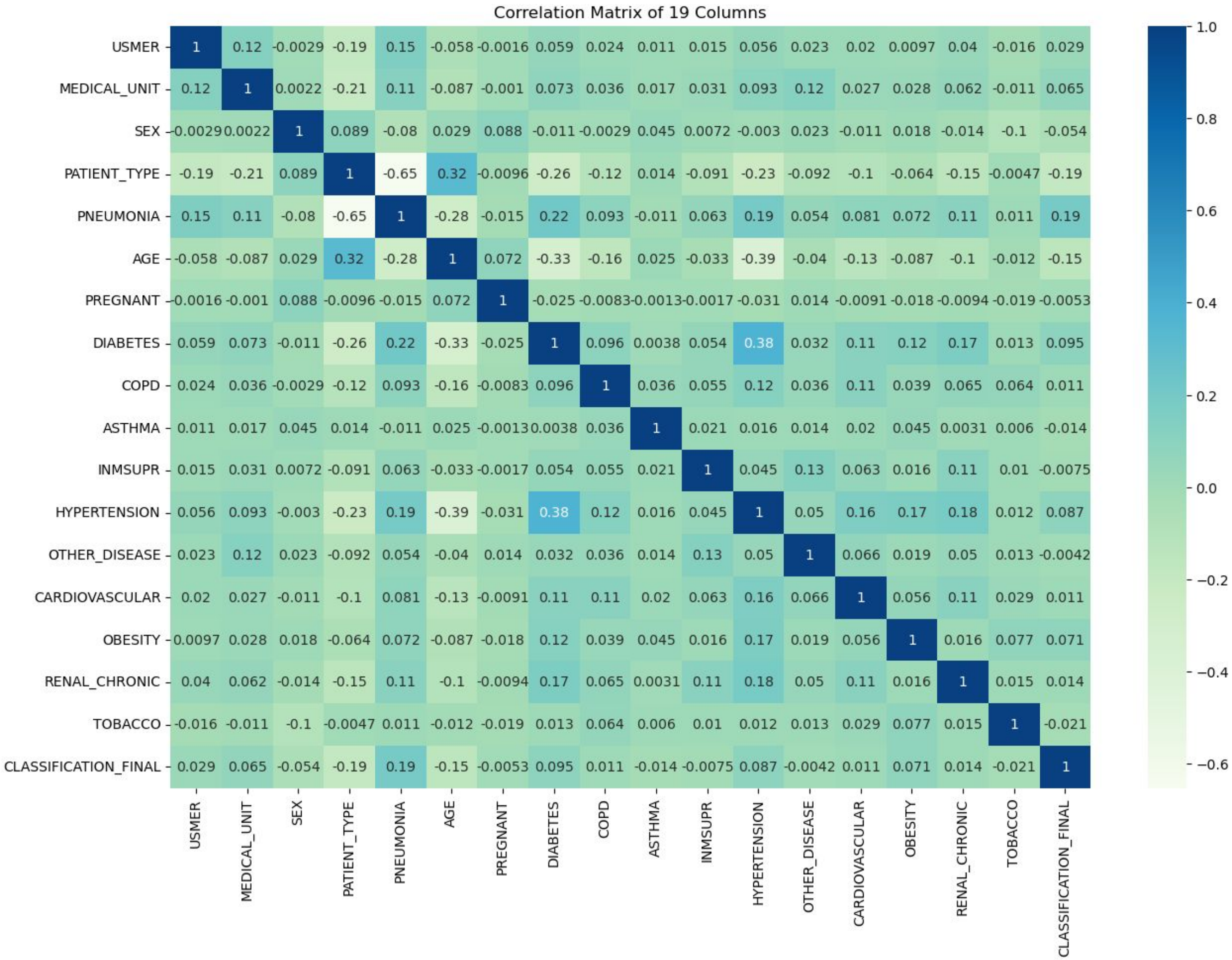
- Exploration of different medical features that affect survival rate of COVID including:
    - Hypertension
    - Obesity
    - Pneumonia
    - Diabetes
- Model prediction of survival rate

Medical Features



Correlation of features

- Pneumonia and patient type are negatively correlated at -0.65, followed by hypertension and age at -0.39 then hypertension and diabetes at 0.38.

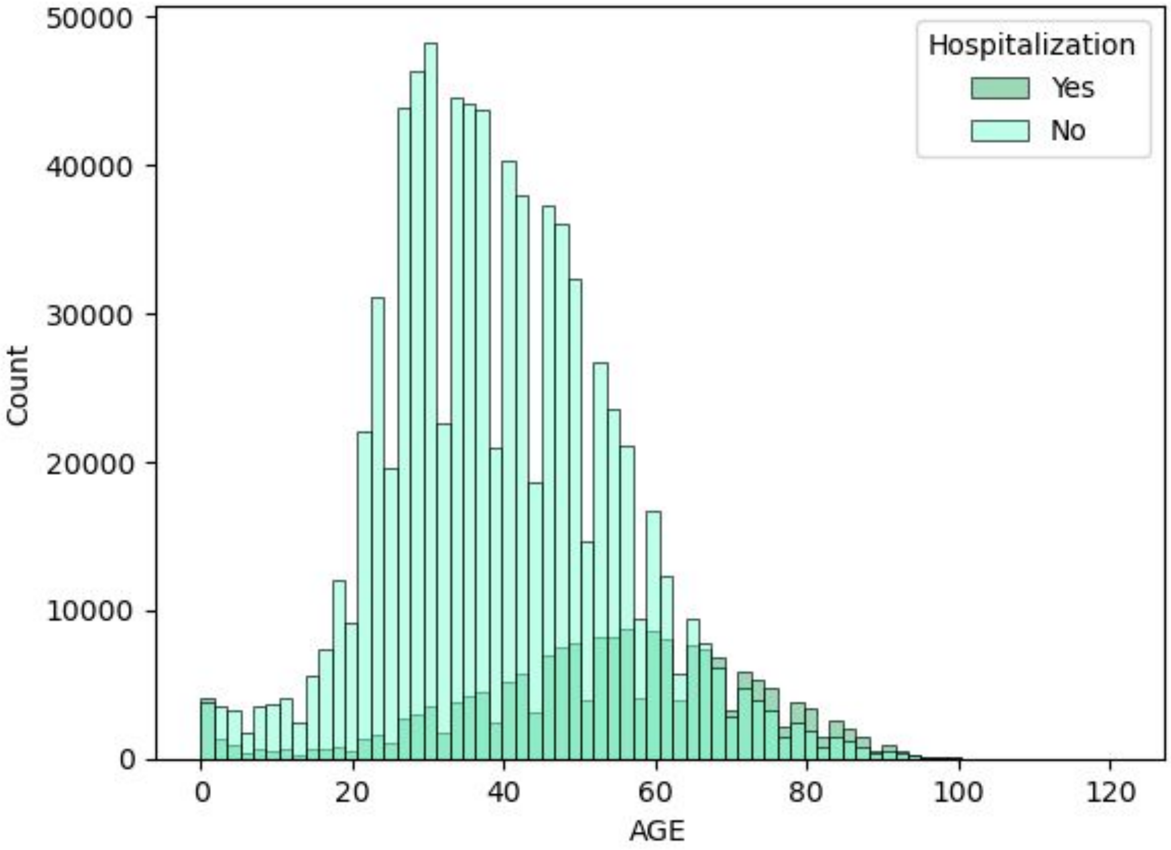


Correlation Matrix of 19 Columns

	USMER	MEDICAL_UNIT	SEX	PATIENT_TYPE	PNEUMONIA	AGE	PREGNANT	DIABETES	COPD	ASTHMA	INMSUPR	HYPERTENSION	OTHER_DISEASE	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO	CLASSIFICATION_FINAL
USMER	1	0.12	-0.0029	-0.19	0.15	-0.058	-0.0016	0.059	0.024	0.011	0.015	0.056	0.023	0.02	0.0097	0.04	-0.016	0.029
MEDICAL_UNIT	0.12	1	0.0022	-0.21	0.11	-0.087	-0.001	0.073	0.036	0.017	0.031	0.093	0.12	0.027	0.028	0.062	-0.011	0.065
SEX	-0.0029	0.0022	1	0.089	-0.08	0.029	0.088	-0.011	-0.0029	0.045	0.0072	-0.003	0.023	-0.011	0.018	-0.014	-0.1	-0.054
PATIENT_TYPE	-0.19	-0.21	0.089	1	-0.65	0.32	-0.0096	-0.26	-0.12	0.014	-0.091	-0.23	-0.092	-0.1	-0.064	-0.15	-0.0047	-0.19
PNEUMONIA	0.15	0.11	-0.08	-0.65	1	-0.28	-0.015	0.22	0.093	-0.011	0.063	0.19	0.054	0.081	0.072	0.11	0.011	0.19
AGE	-0.058	-0.087	0.029	0.32	-0.28	1	0.072	-0.33	-0.16	0.025	-0.033	-0.39	-0.04	-0.13	-0.087	-0.1	-0.012	-0.15
PREGNANT	-0.0016	-0.001	0.088	-0.0096	-0.015	0.072	1	-0.025	-0.0083	-0.0013	-0.0017	-0.031	0.014	-0.0091	-0.018	-0.0094	-0.019	-0.0053
DIABETES	0.059	0.073	-0.011	-0.26	0.22	-0.33	-0.025	1	0.096	0.0038	0.054	0.38	0.032	0.11	0.12	0.17	0.013	0.095
COPD	0.024	0.036	-0.0029	-0.12	0.093	-0.16	-0.0083	0.096	1	0.036	0.055	0.12	0.036	0.11	0.039	0.065	0.064	0.011
ASTHMA	0.011	0.017	0.045	0.014	-0.011	0.025	-0.0013	0.0038	0.036	1	0.021	0.016	0.014	0.02	0.045	0.0031	0.006	-0.014
INMSUPR	0.015	0.031	0.0072	-0.091	0.063	-0.033	-0.0017	0.054	0.055	0.021	1	0.045	0.13	0.063	0.016	0.11	0.01	-0.0075
HYPERTENSION	0.056	0.093	-0.003	-0.23	0.19	-0.39	-0.031	0.38	0.12	0.016	0.045	1	0.05	0.16	0.17	0.18	0.012	0.087
OTHER_DISEASE	0.023	0.12	0.023	-0.092	0.054	-0.04	0.014	0.032	0.036	0.014	0.13	0.05	1	0.066	0.019	0.05	0.013	-0.0042
CARDIOVASCULAR	0.02	0.027	-0.011	-0.1	0.081	-0.13	-0.0091	0.11	0.11	0.02	0.063	0.16	0.066	1	0.056	0.11	0.029	0.011
OBESITY	0.0097	0.028	0.018	-0.064	0.072	-0.087	-0.018	0.12	0.039	0.045	0.016	0.17	0.019	0.056	1	0.016	0.077	0.071
RENAL_CHRONIC	0.04	0.062	-0.014	-0.15	0.11	-0.1	-0.0094	0.17	0.065	0.0031	0.11	0.18	0.05	0.11	0.016	1	0.015	0.014
TOBACCO	-0.016	-0.011	-0.1	-0.0047	0.011	-0.012	-0.019	0.013	0.064	0.006	0.01	0.012	0.013	0.029	0.077	0.015	1	-0.021
CLASSIFICATION_FINAL	0.029	0.065	-0.054	-0.19	0.19	-0.15	-0.0053	0.095	0.011	-0.014	-0.0075	0.087	-0.0042	0.011	0.071	0.014	-0.021	1

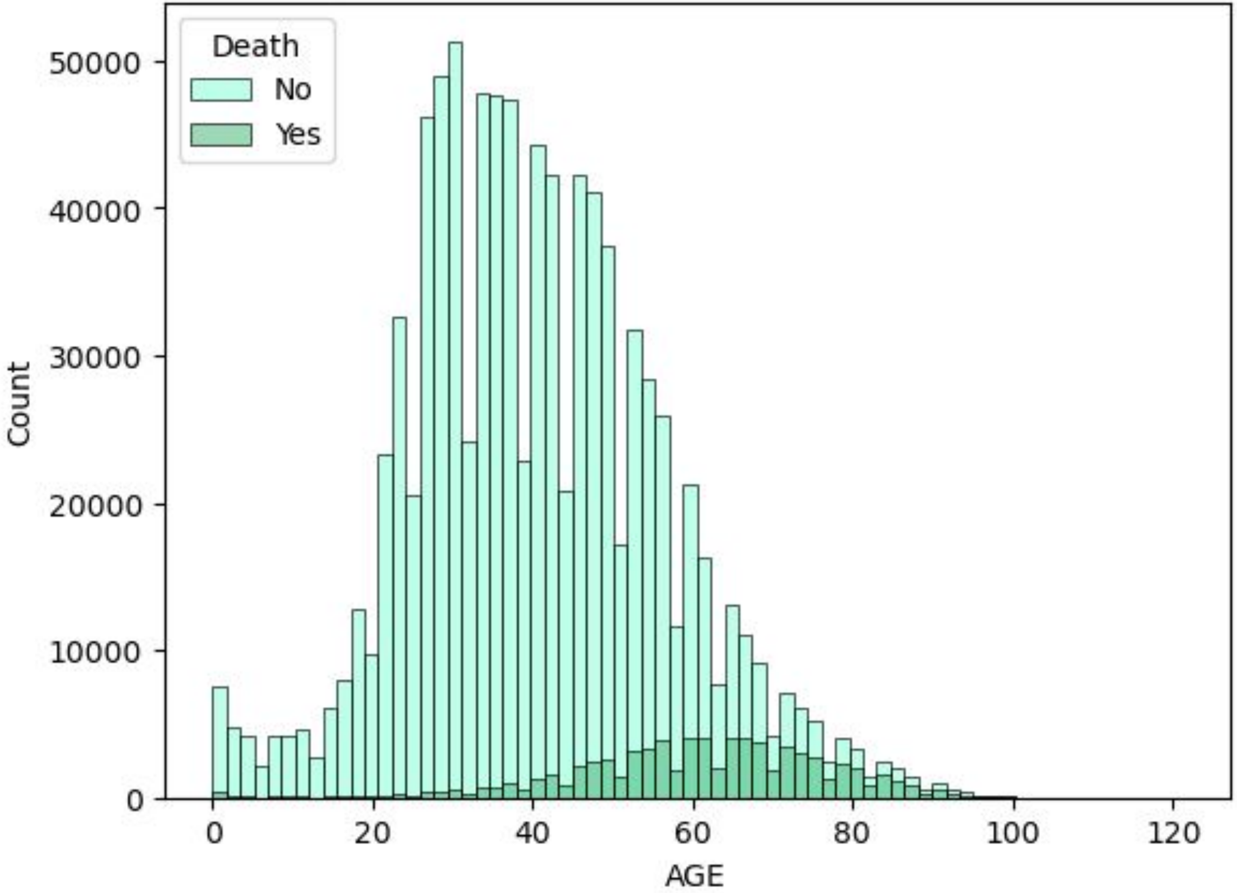
Age vs Hospitalization

- Hospitalization occurs more frequently between aged 50-80 and also in young infants.



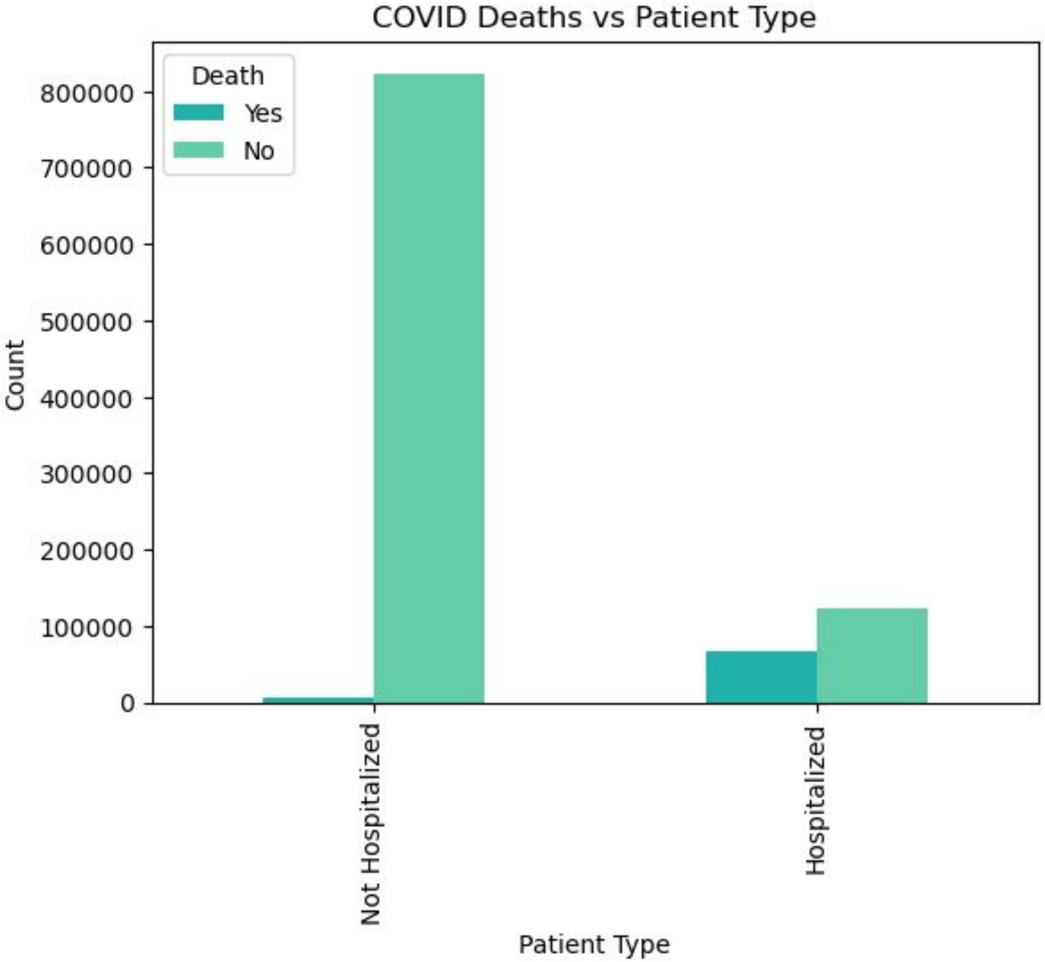
Age vs Covid death

- Death occurs most frequently between 50-80 inline with hospitalization.



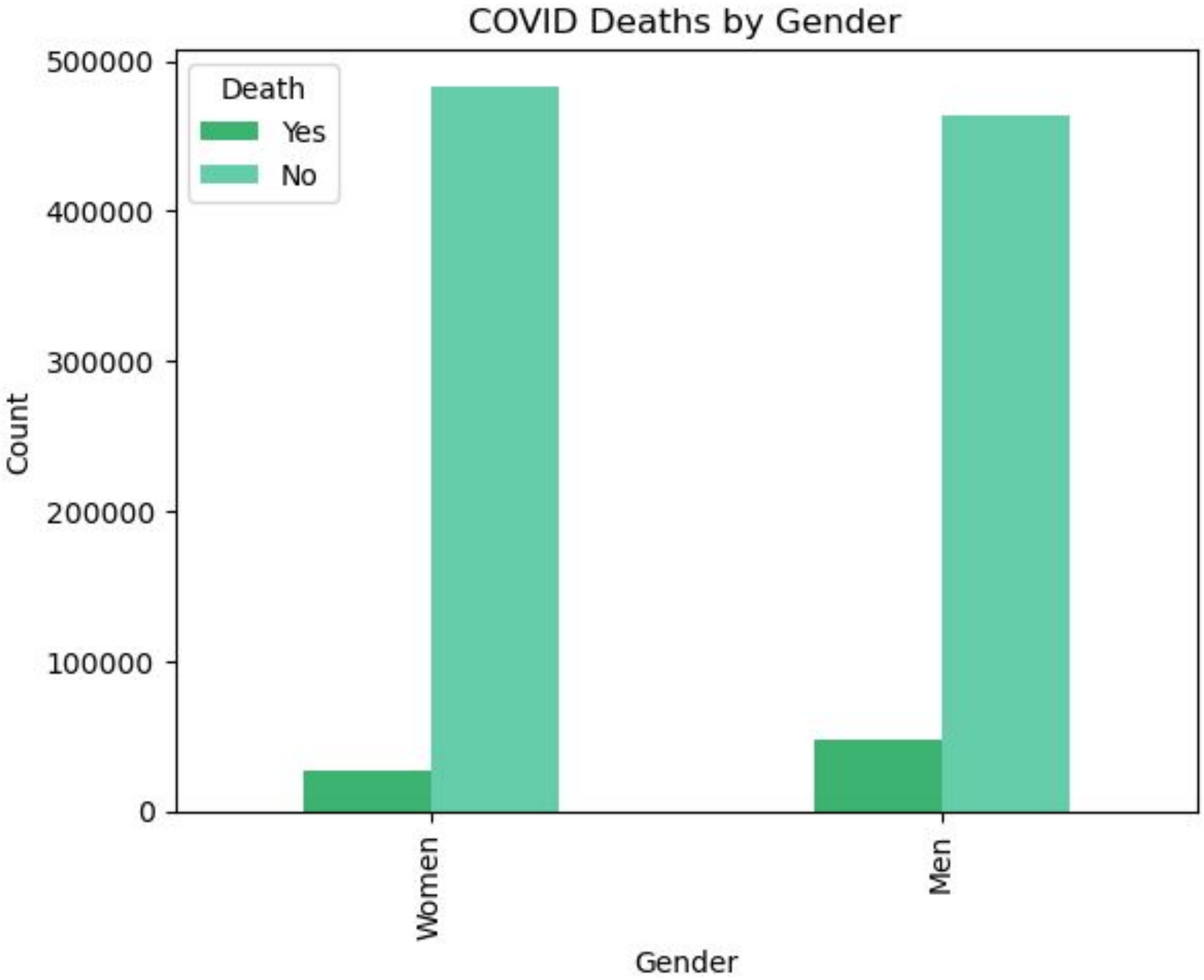
Hospitalization vs Covid Death

- Majority of patients are not hospitalized and of those only 0.8% have died.
- Of those who have been hospitalized death rate is much higher at 35%.



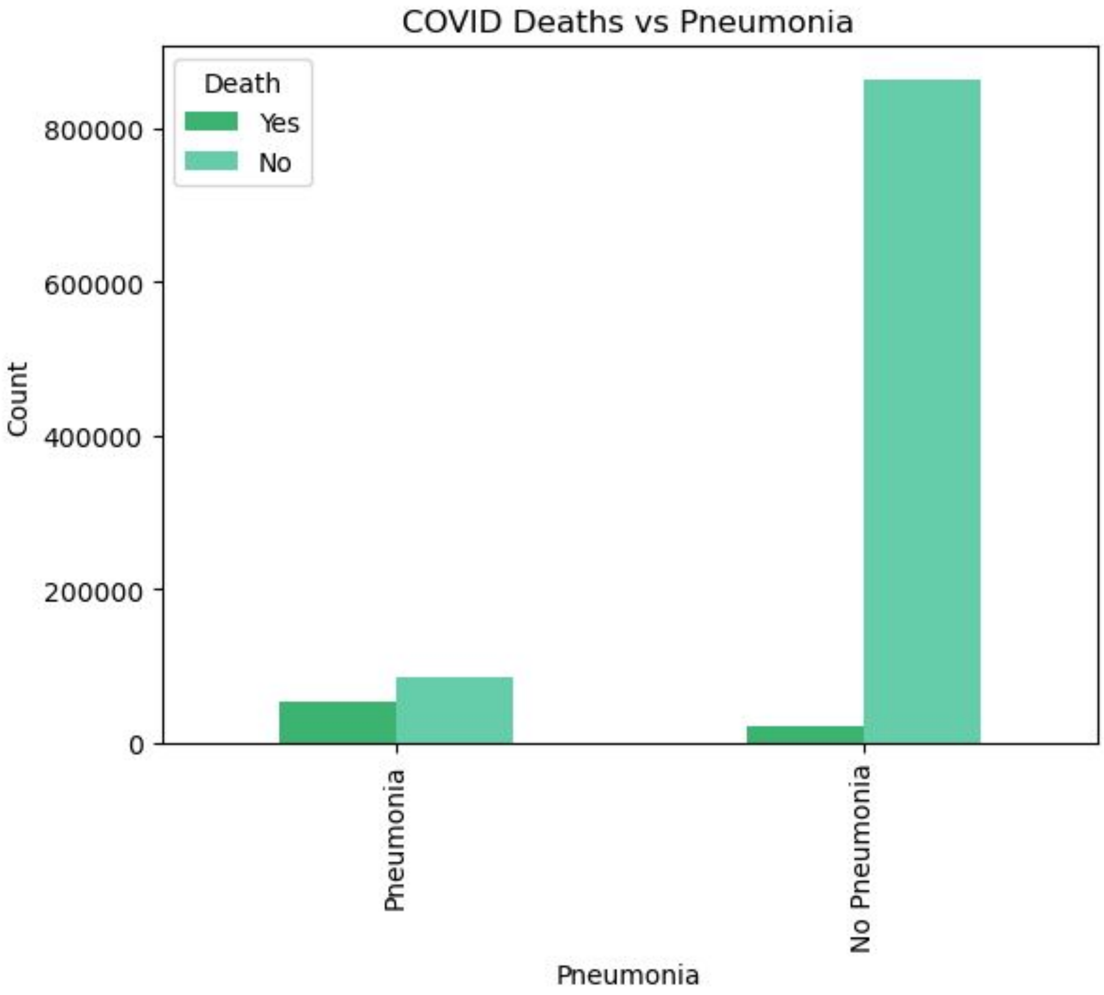
Gender vs Covid Death

- Percentage of death in Female patients: 5.21 %
- Percentage of death in Male patients: 9.39 %



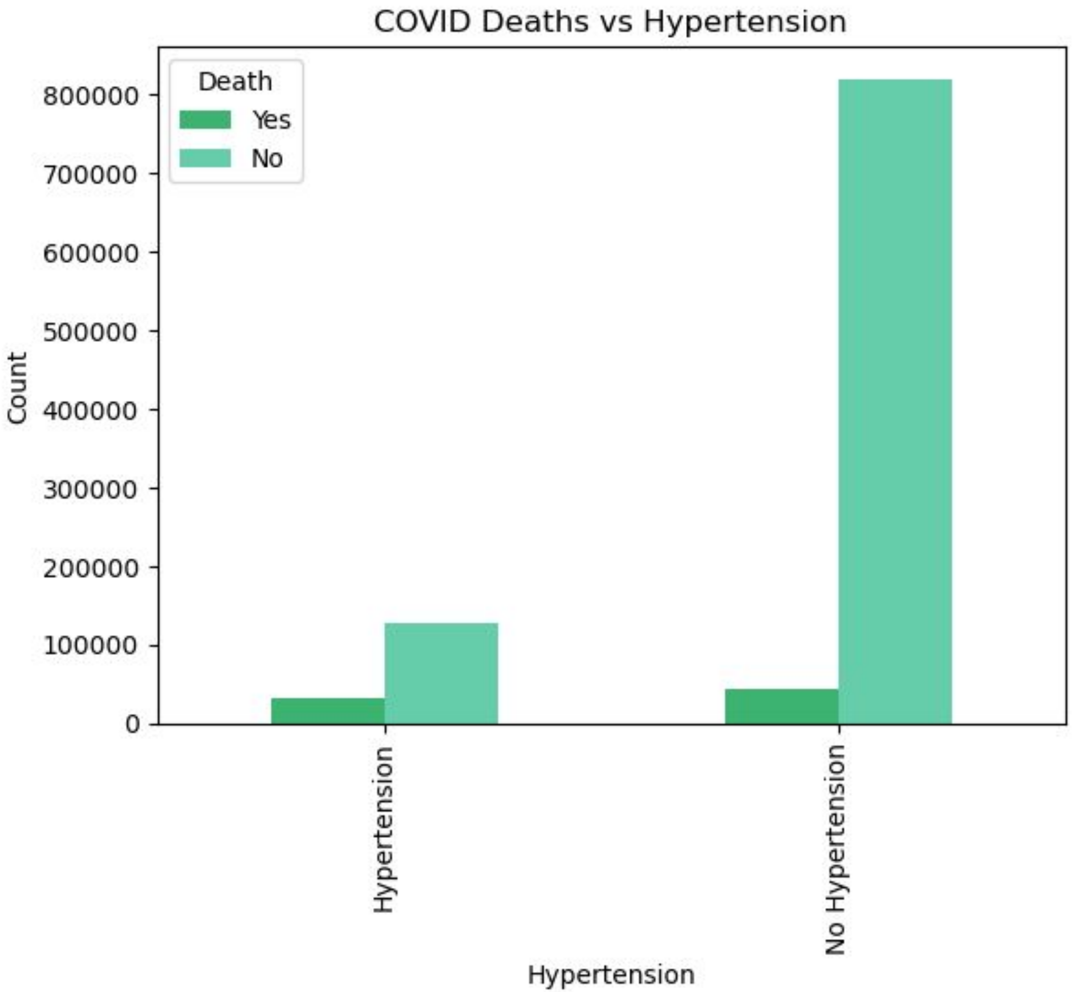
Pneumonia vs Covid Death

- Percentage of death in pneumonia patients: 38.4 %
- Percentage of death in non-pneumonia patients: 2.48 %



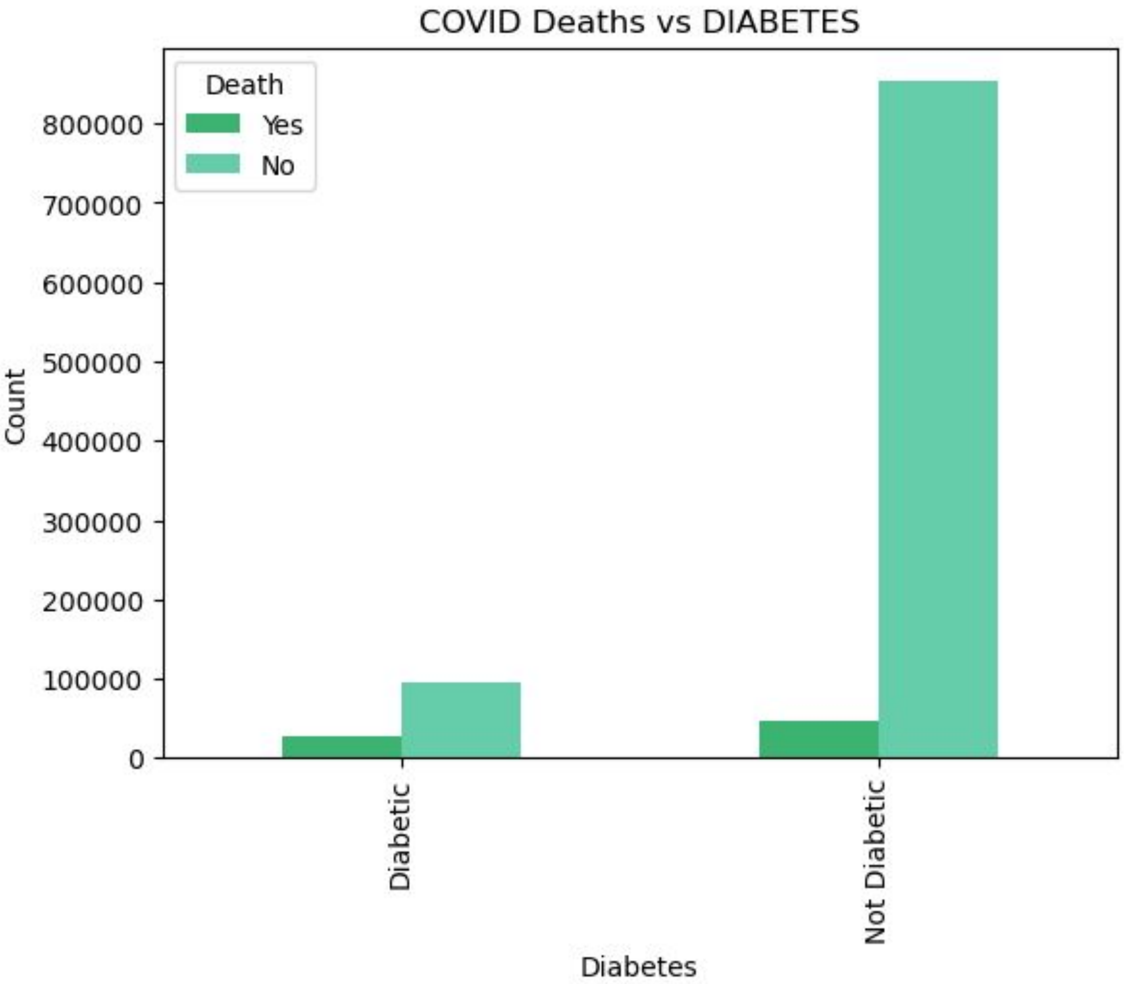
Hypertension vs Covid Death

- Percentage of death in patients with Hypertension: 19.73 %
- Percentage of death in patients without Hypertension: 5.01 %



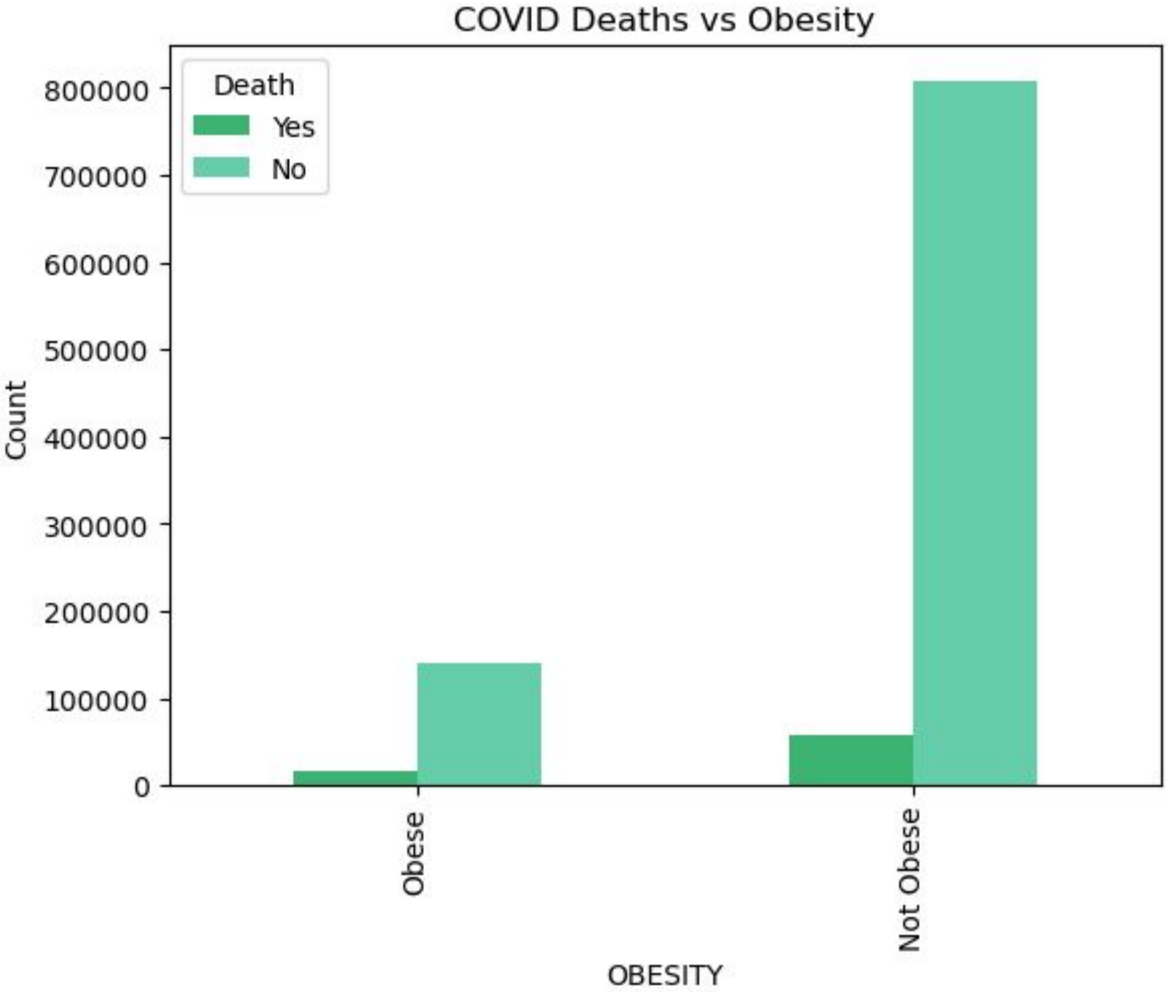
Diabetes vs Covid Death

- Percentage of death in diabetic patients: 22.64 %
- Percentage of death in non-diabetic patients: 5.22 %



Obesity vs Covid Death

- Percentage of death in obese patients: 10.76 %
- Percentage of death in non-obese patients: 6.68 %



Modelling the data

- The aim of the model is to predict the survival of a patient based on certain medical features they may have.
- Dataset is heavily imbalanced in the target variable therefore different sampling techniques were tested in the models and evaluated.
- Models that were tested: Logistic regression, random forests, random forest with gridsearch cross validation, over/under sampling, SMOTE, variance threshold and RFE.

Model Evaluations

Model	Class Imbalance	Under Sampling	Over Sampling	SMOTE	Variance Threshold	RFE
Logistic Regression	0.47	0.51	0.51	0.51	0.41	0.37
Random Forest	0.32	0.49	0.49	0.48	-	-
Random Forest GS	0.47	0.83	**0.88**	-	-	-

- Random forest with grid search cross validation and oversampling is the best model with a Kappa score of 0.88.
- This indicates a high level of agreement between the model predictions and the actual outcomes with high accuracy.

Final Model	Kappa	Precision	Recall	F1-Score
Random Forest GS	0.88	0.70	0.89	0.75