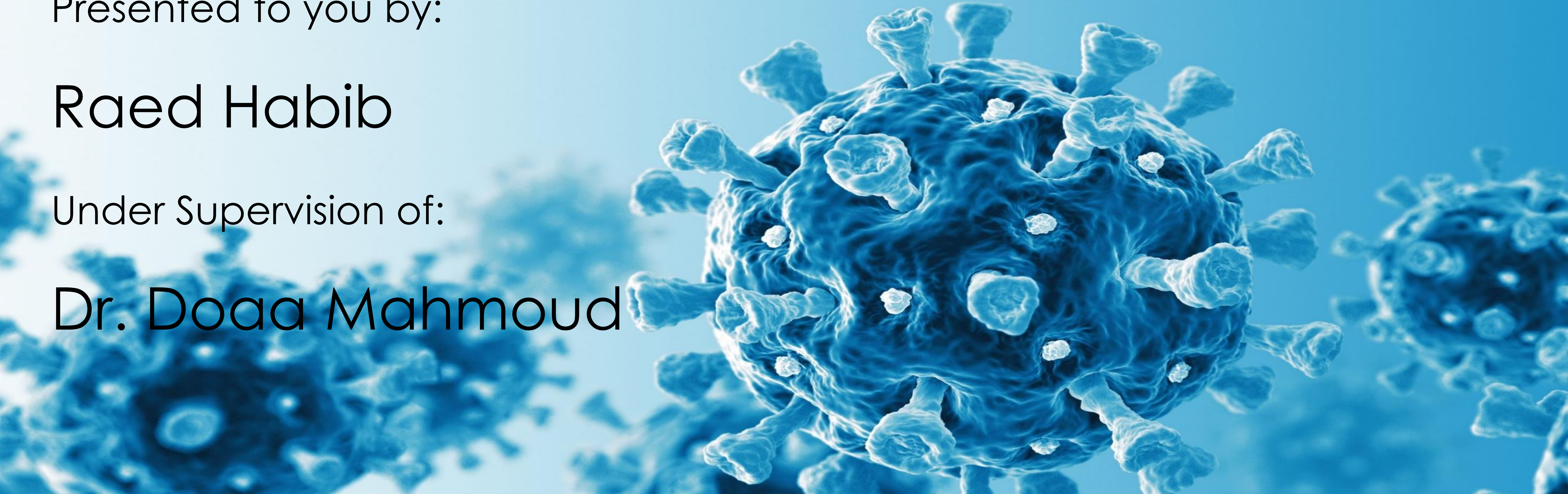
Covid 19 Analysis and Risk Prediction

Presented to you by:

Raed Habib

Under Supervision of:

Dr. Doaa Mahmoud



Agenda

- Introduction
- Dataset Overview
- Initial Questions to be Answered
- Exploratory Data Analysis
- Data Pre-processing and Problems
- Modeling and Results
- Conclusion

Introduction

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people infected with COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment. Older people, and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness.

Introduction

During the entire course of the pandemic, one of the main problems that healthcare providers have faced is the shortage of medical resources and a proper plan to efficiently distribute them. In these tough times, being able to predict what kind of resource an individual might require at the time of being tested positive or even before that will be of immense help to the authorities as they would be able to procure and arrange for the resources necessary to save the life of that patient.

Project goal

The main goal of this project is to build a machine learning model that, given a Covid-19 patient's current symptom, status, and medical history, will predict whether the patient is at high risk or not.

Dataset Overview

The dataset is obtained from Kaggle at the following link:

https://www.kaggle.com/datasets/meirnizri/covid19-dataset, and was provided by the Mexican government, which can be found at this link:

https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico

This dataset contains an enormous number of anonymized patient-related information including pre-conditions. The raw dataset consists of 21 unique features and 1,048,576 unique patients. In the Boolean features, **1 means "yes" and 2 means "no". values as 97 and 99 are missing data**.

Dataset Overview

sex: 1 for female and 2 for male.

age: of the patient.

classification: Covid test findings. Values 1-3 mean that the patient has Covid in different degrees. 4 or higher means that the patient is not a carrier of Covid or that the test is inconclusive.

- **sex**: 1 for female and 2 for male

patient type: type of care the patient received in the unit. 1 for returned home and 2 for hospitalization.

pneumonia: whether the patient already have air sacs inflammation or not.

pregnancy: whether the patient is pregnant or not.

diabetes: whether the patient has diabetes or not.

Dataset Overview

copd: Indicates whether the patient has Chronic obstructive pulmonary disease or not.

asthma: whether the patient has asthma or not.

inmsupr: whether the patient is immunosuppressed or not.

- **sex**: 1 for female and 0 for male

hypertension: whether the patient has hypertension or not.

cardiovascular: whether the patient has heart or blood vessels related disease.

renal chronic: whether the patient has chronic renal disease or not.

other disease: whether the patient has other disease or not.

Dataset Overview

obesity: whether the patient is obese or not.

tobacco: whether the patient is a tobacco user.

firusmr: Indicates whether the patient treated medical units of the st, second or third level.

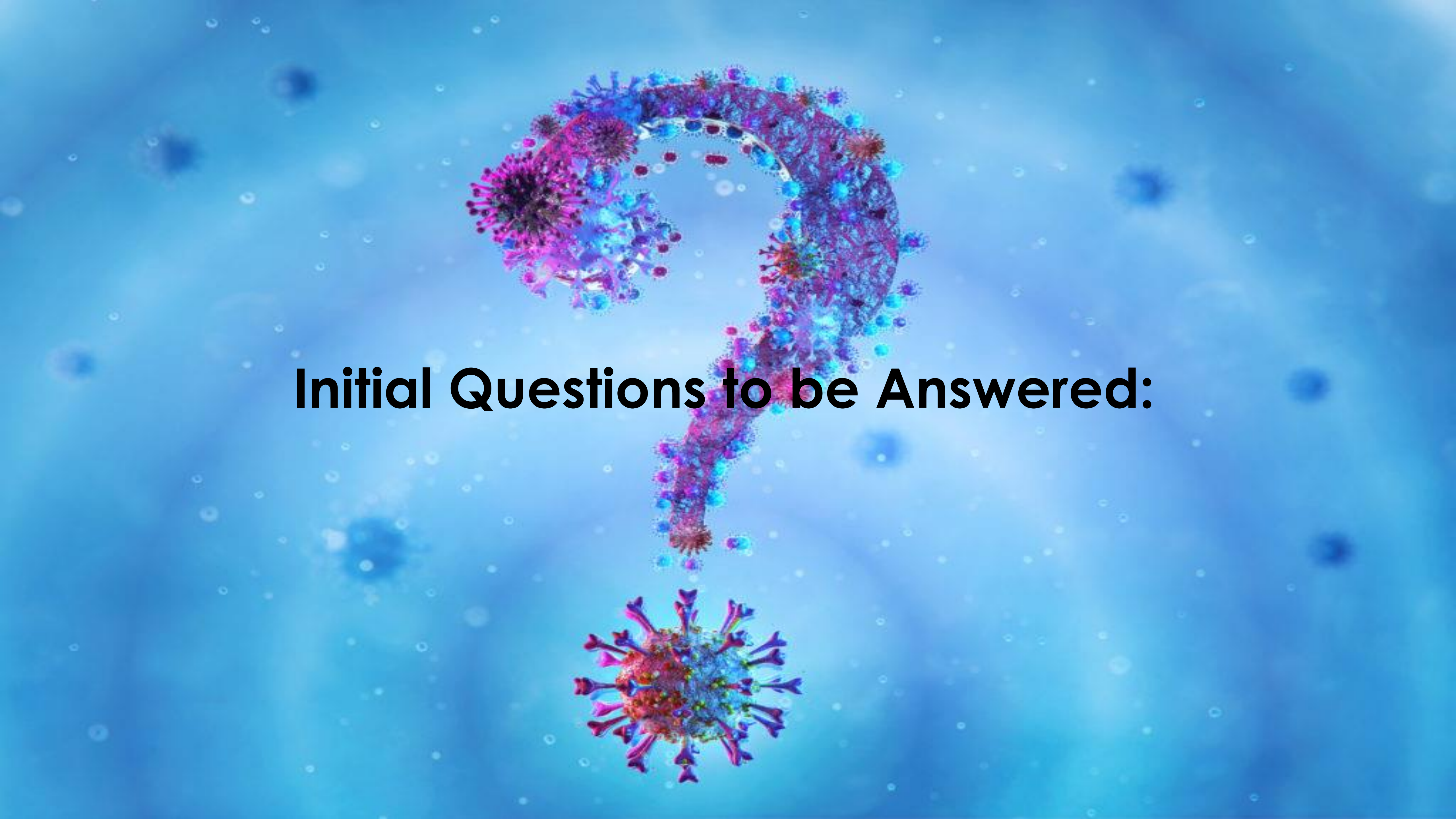
- **sex**: 1 for female and 0 for male

medical unit: type of institution of the National Health System that provided the care.

Intubed:whether the patient was connected to the ventilator.

icu: Indicates whether the patient had been admitted to an Intensive Care Unit.

date died: If the patient died indicate the date of death, and 9999-99-99 otherwise.

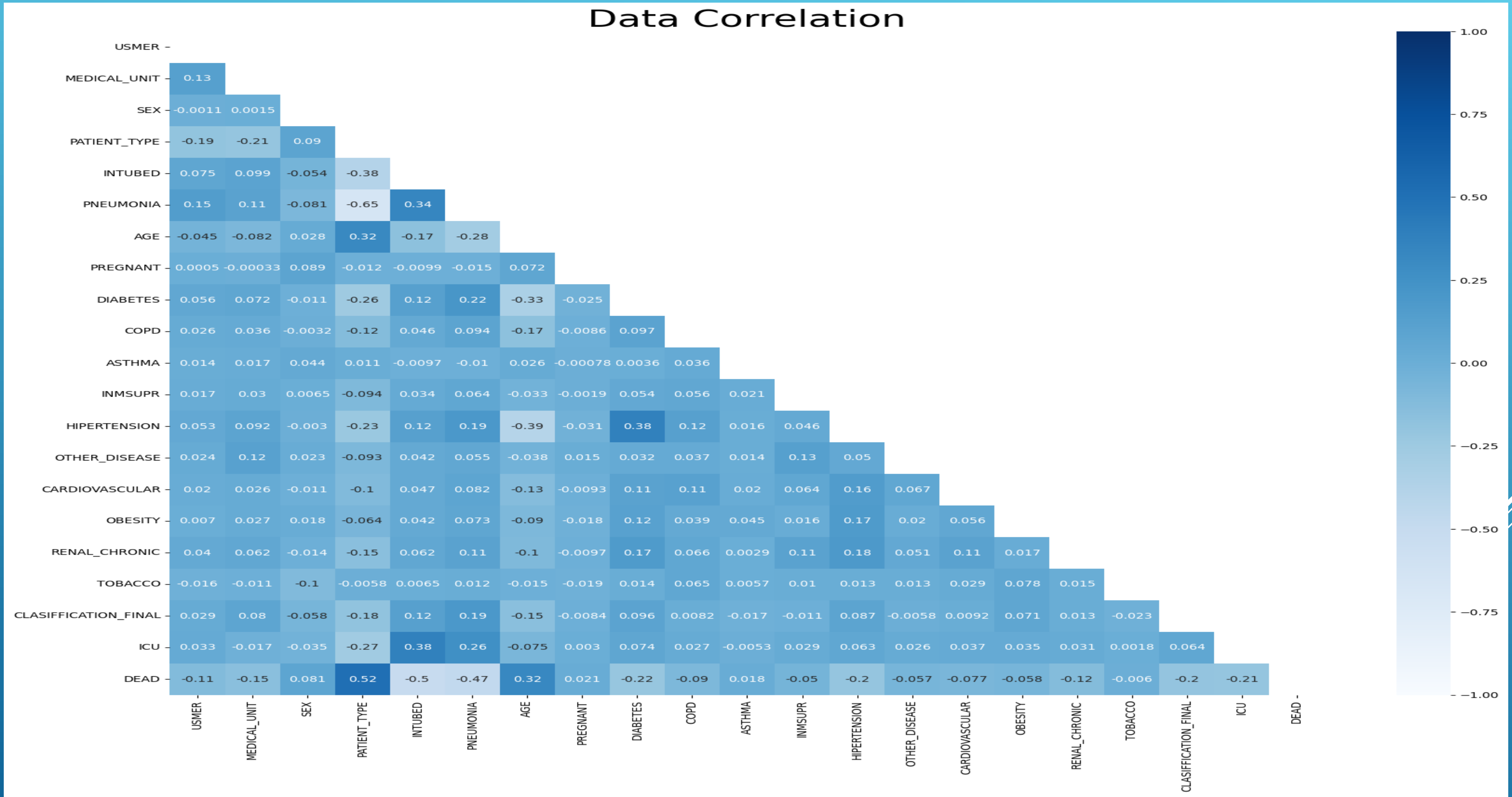


Initial Questions to be Answered:

Initial Questions to be Answered:

1- How many people have died?
2- And, did they all have the same results?
3- Is it true that age has some impact?
4- Does Obesity have any impact
5- Does gender have any impact?
6- How do other diseases affect the patient classification?
7- How many patients were diagnosed with Covid 19 of 1st, 2nd, or 3rd degrees?
8- How many of them have died?
9- How many patients were hospitalized?
10- How many of them have died?
11- How many patients had been admitted to an Intensive Care Unit?
12- How many of them have died?

Exploratory Data Analysis

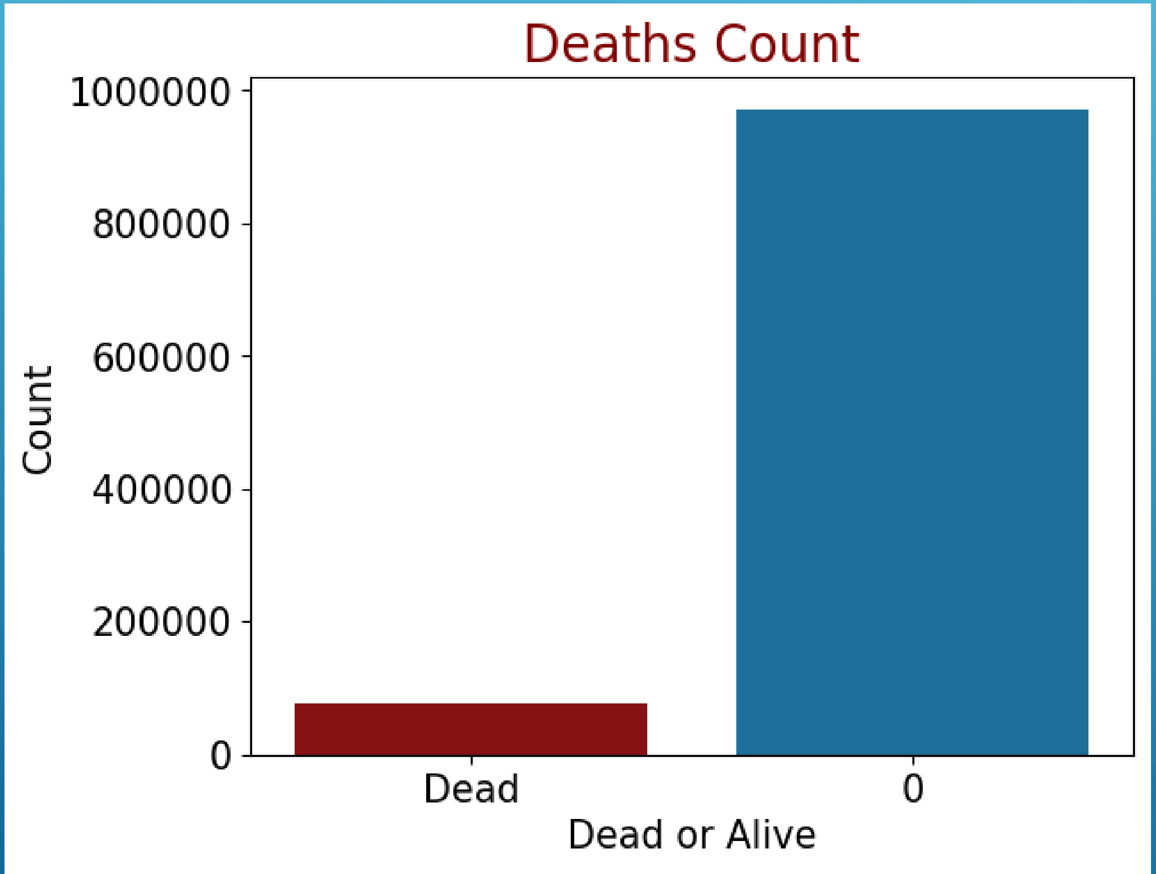


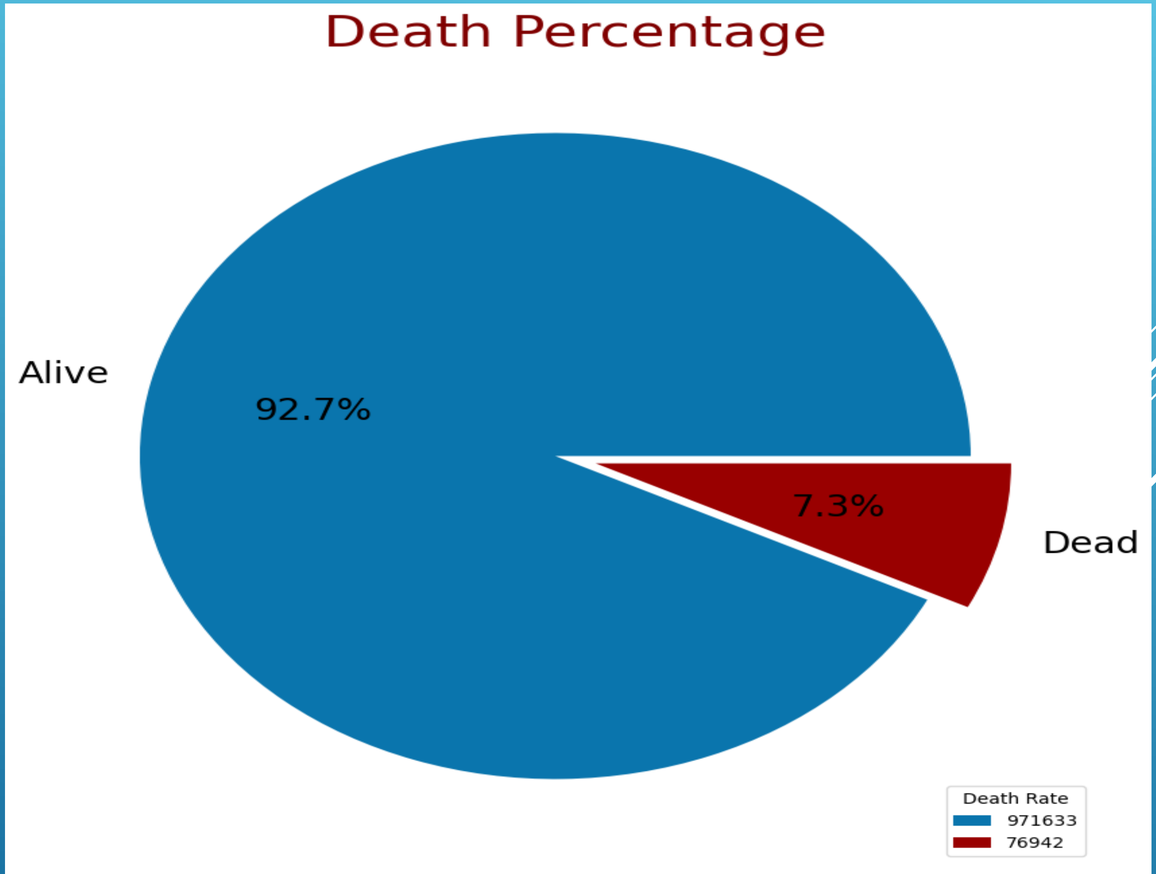
Data Correlation

	USMER	MEDICAL_UNIT	SEX	PATIENT_TYPE	INTUBED	PNEUMONIA	AGE	PREGNANT	DIABETES	COPD	ASTHMA	INMSUPR	HIPERTENSION	OTHER_DISEASE	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO	CLASIFFICATION_FINAL	ICU	DEAD
USMER																					
MEDICAL_UNIT	0.13																				
SEX	-0.0011	0.0015																			
PATIENT_TYPE	-0.19	-0.21	0.09																		
INTUBED	0.075	0.099	-0.054	-0.38																	
PNEUMONIA	0.15	0.11	-0.081	-0.65	0.34																
AGE	-0.045	-0.082	0.028	0.32	-0.17	-0.28															
PREGNANT	0.0005	-0.00033	0.089	-0.012	-0.0099	-0.015	0.072														
DIABETES	0.056	0.072	-0.011	-0.26	0.12	0.22	-0.33	-0.025													
COPD	0.026	0.036	-0.0032	-0.12	0.046	0.094	-0.17	-0.0086	0.097												
ASTHMA	0.014	0.017	0.044	0.011	-0.0097	-0.01	0.026	-0.00078	0.0036	0.036											
INMSUPR	0.017	0.03	0.0065	-0.094	0.034	0.064	-0.033	-0.0019	0.054	0.056	0.021										
HIPERTENSION	0.053	0.092	-0.003	-0.23	0.12	0.19	-0.39	-0.031	0.38	0.12	0.016	0.046									
OTHER_DISEASE	0.024	0.12	0.023	-0.093	0.042	0.055	-0.038	0.015	0.032	0.037	0.014	0.13	0.05								
CARDIOVASCULAR	0.02	0.026	-0.011	-0.1	0.047	0.082	-0.13	-0.0093	0.11	0.11	0.02	0.064	0.16	0.067							
OBESITY	0.007	0.027	0.018	-0.064	0.042	0.073	-0.09	-0.018	0.12	0.039	0.045	0.016	0.17	0.02	0.056						
RENAL_CHRONIC	0.04	0.062	-0.014	-0.15	0.062	0.11	-0.1	-0.0097	0.17	0.066	0.0029	0.11	0.18	0.051	0.11	0.017					
TOBACCO	-0.016	-0.011	-0.1	-0.0058	0.0065	0.012	-0.015	-0.019	0.014	0.065	0.0057	0.01	0.013	0.013	0.029	0.078	0.015				
CLASIFFICATION_FINAL	0.029	0.08	-0.058	-0.18	0.12	0.19	-0.15	-0.0084	0.096	0.0082	-0.017	-0.011	0.087	-0.0058	0.0092	0.071	0.013	-0.023			
ICU	0.033	-0.017	-0.035	-0.27	0.38	0.26	-0.075	0.003	0.074	0.027	-0.0053	0.029	0.063	0.026	0.037	0.035	0.031	0.0018	0.064		
DEAD	-0.11	-0.15	0.081	0.52	-0.5	-0.47	0.32	0.021	-0.22	-0.09	0.018	-0.05	-0.2	-0.057	-0.077	-0.058	-0.12	-0.006	-0.2	-0.21	

Exploratory Data Analysis

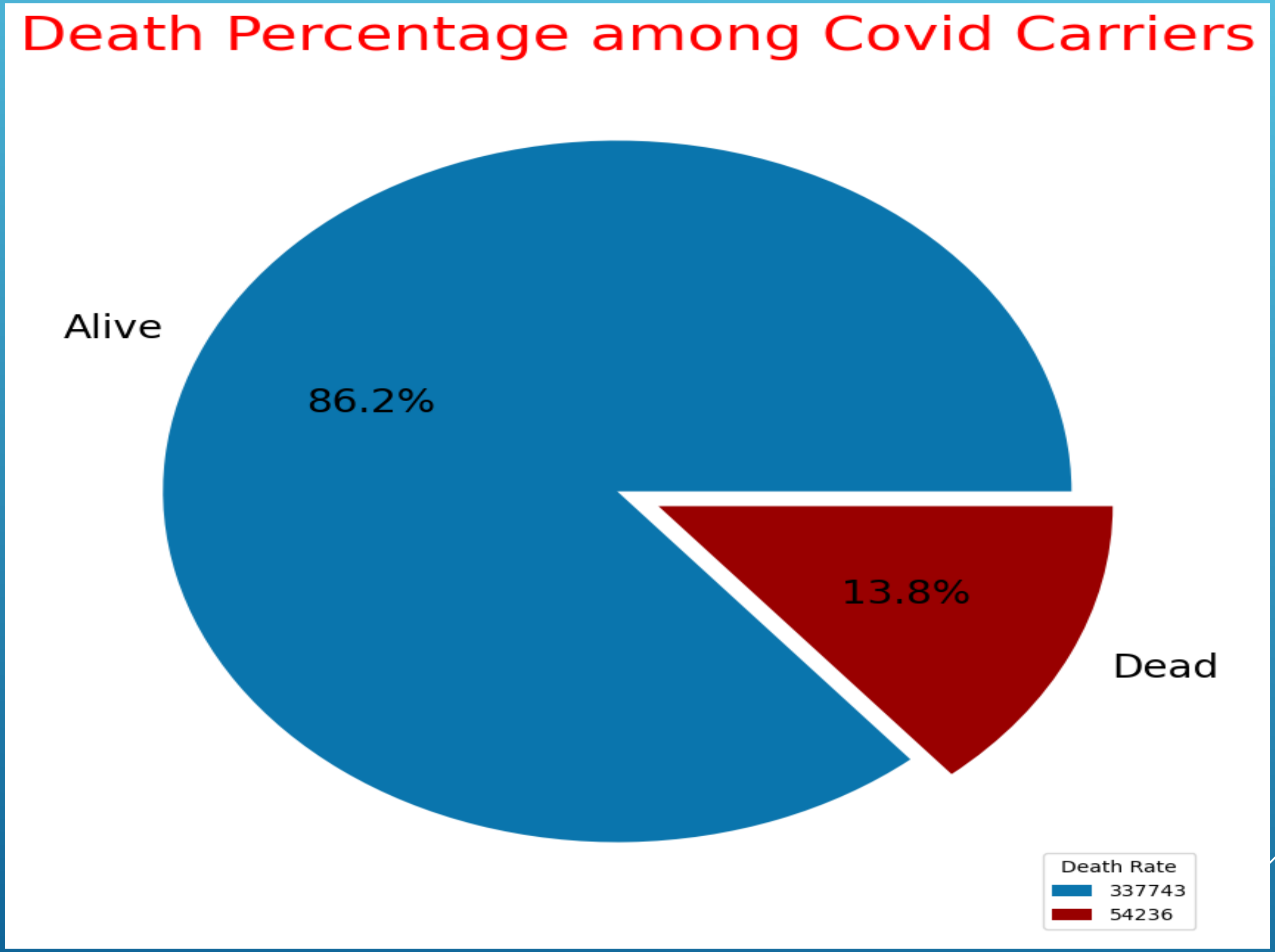
Death statistics





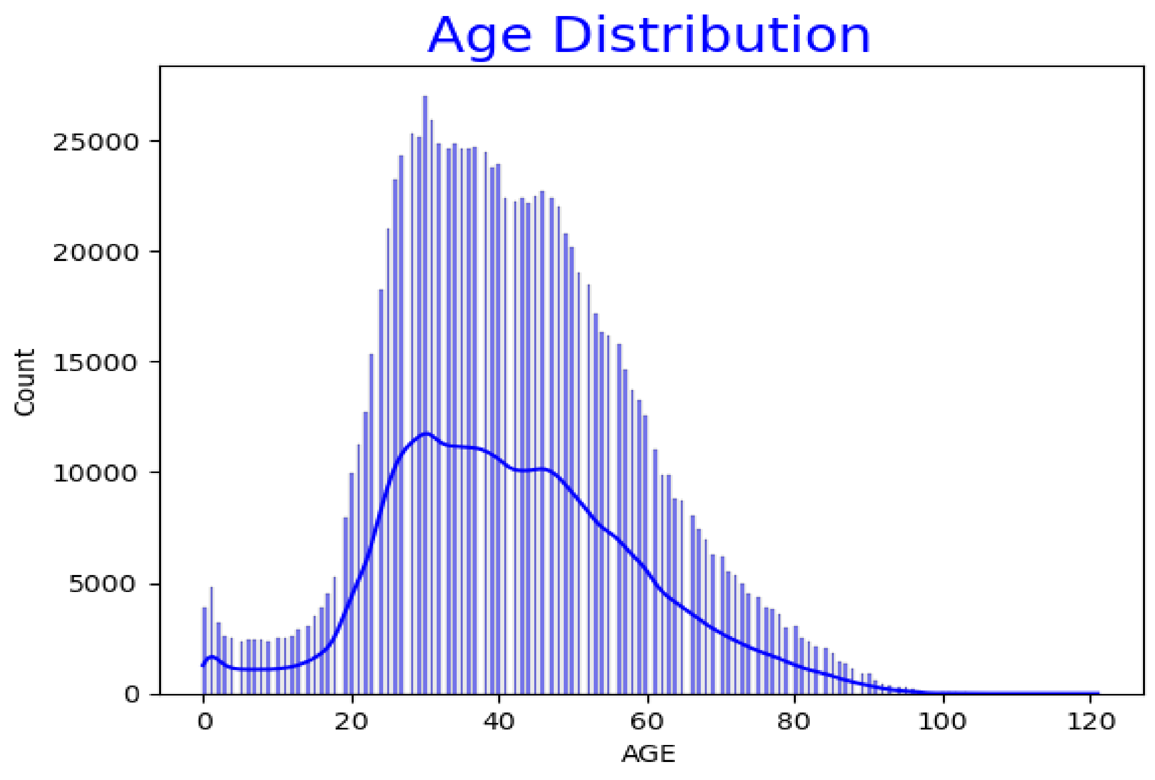
Exploratory Data Analysis

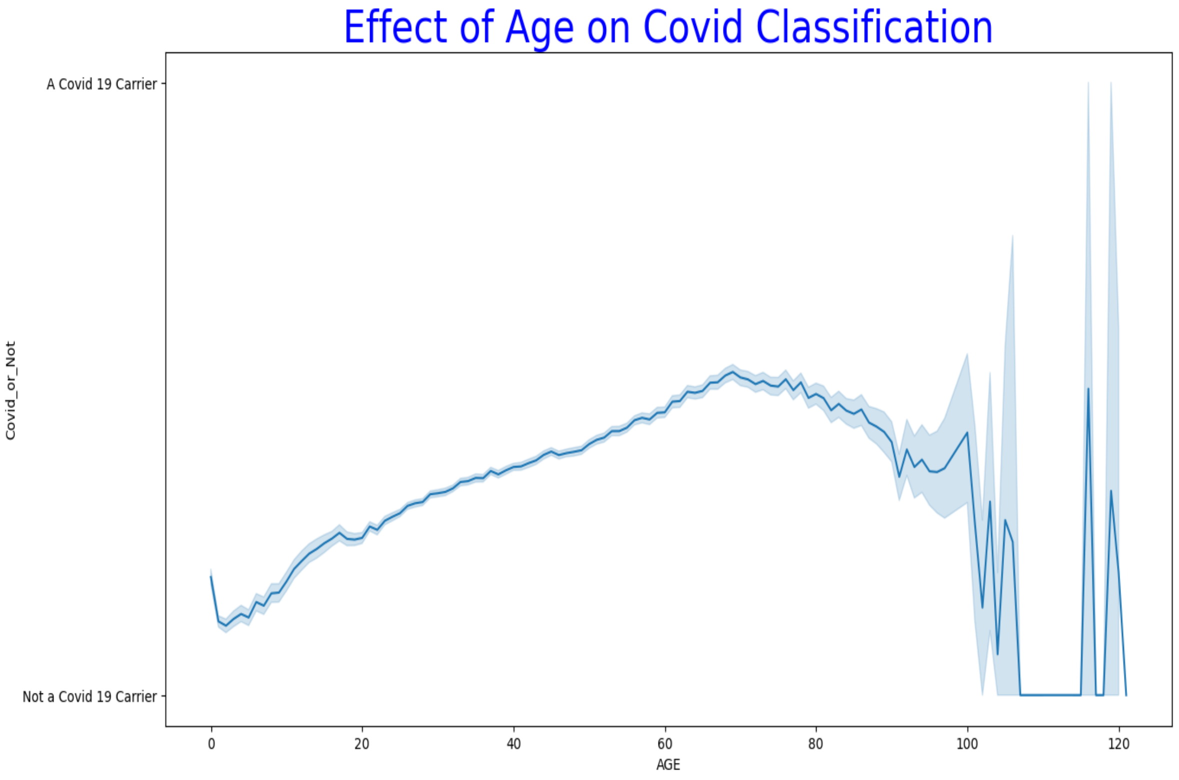
Death statistics



Exploratory Data Analysis

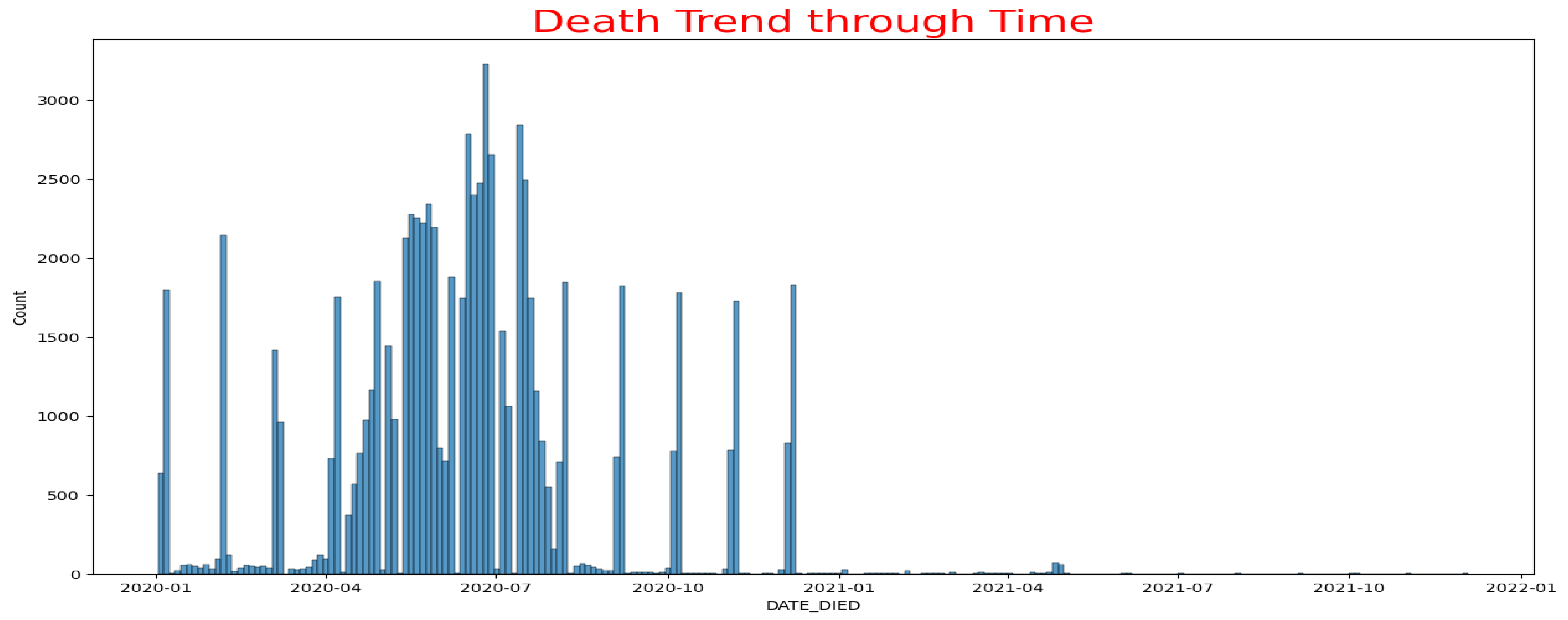
Death statistics





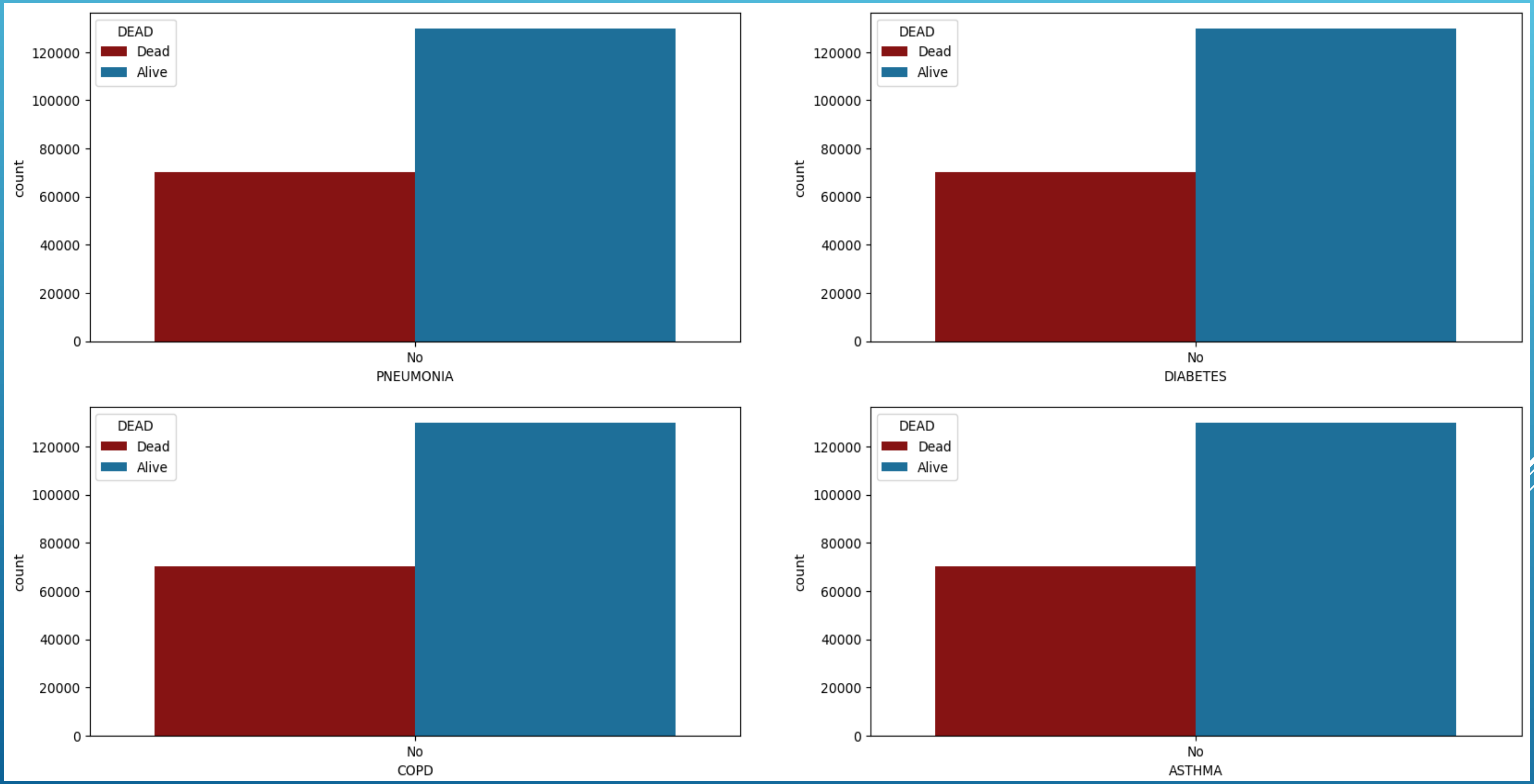
Exploratory Data Analysis

Death statistics



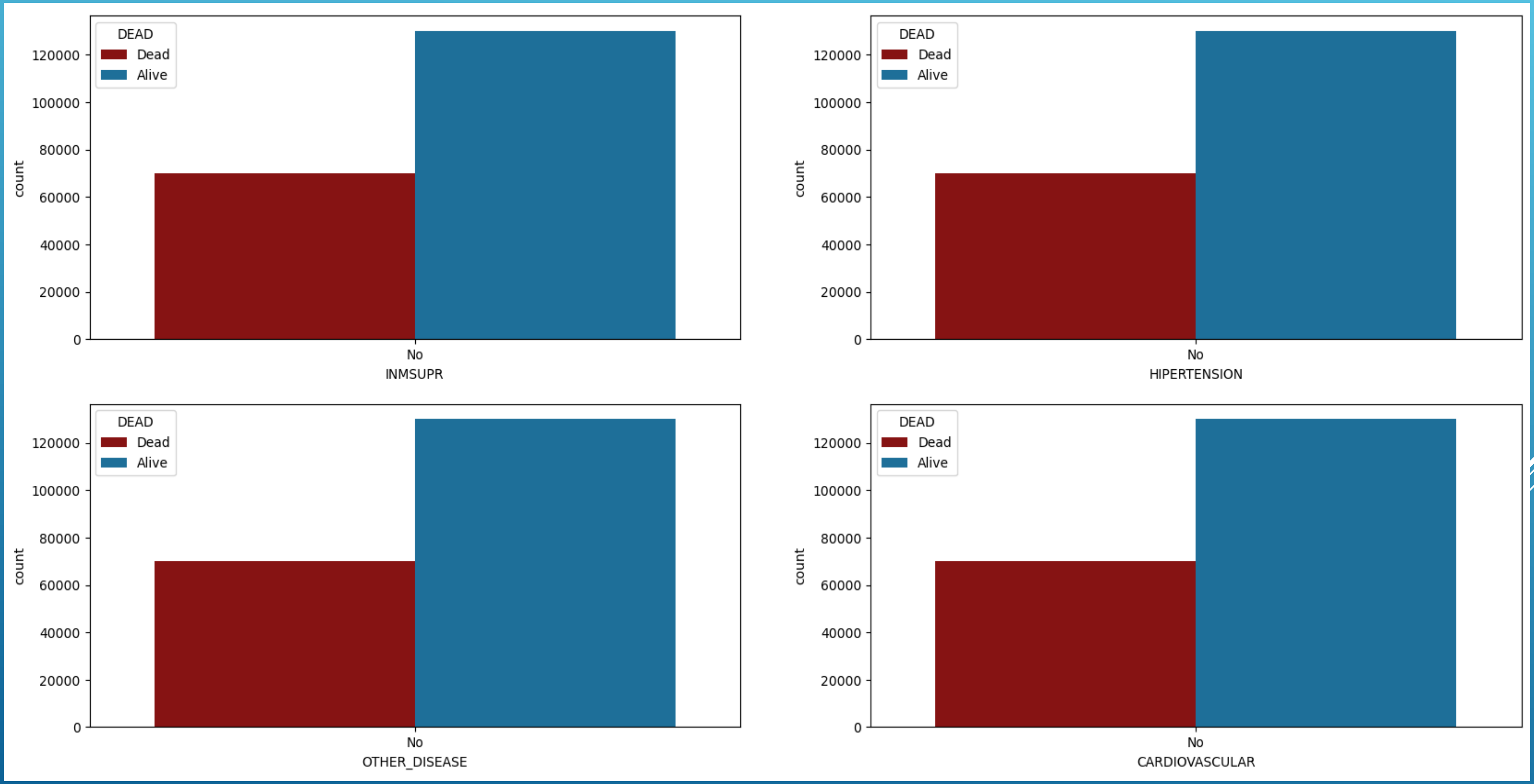
Exploratory Data Analysis

Effect of diseases on Death



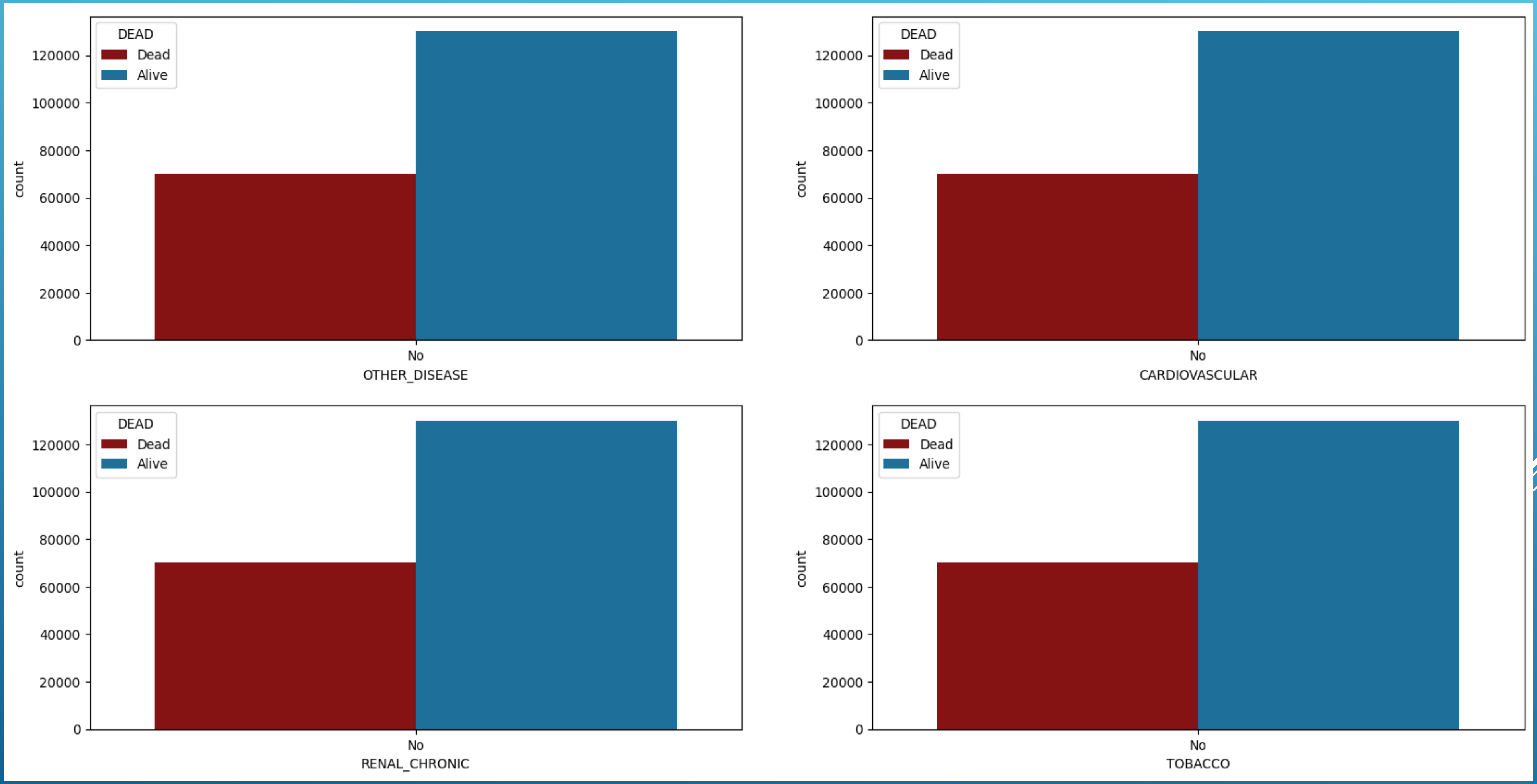
Exploratory Data Analysis

Effect of diseases on Death



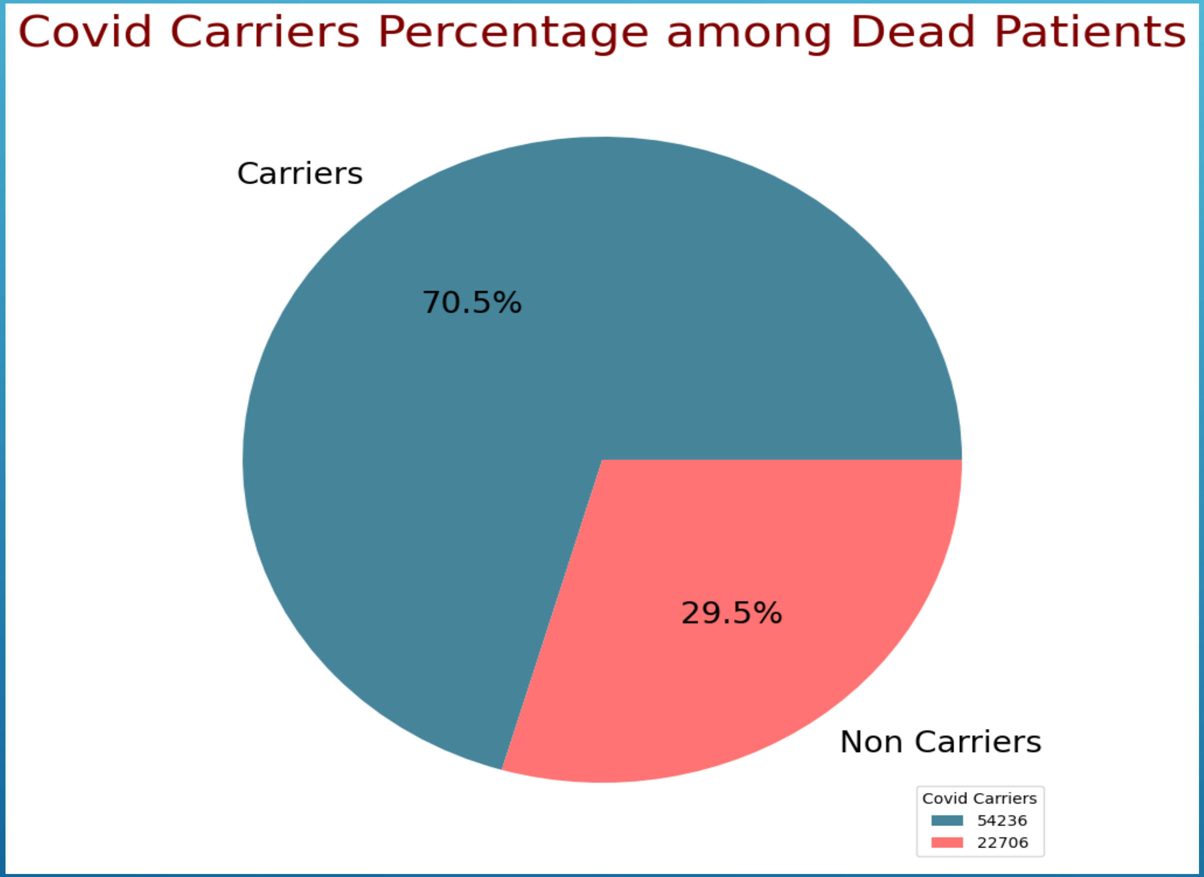
Exploratory Data Analysis

Effect of diseases on Death

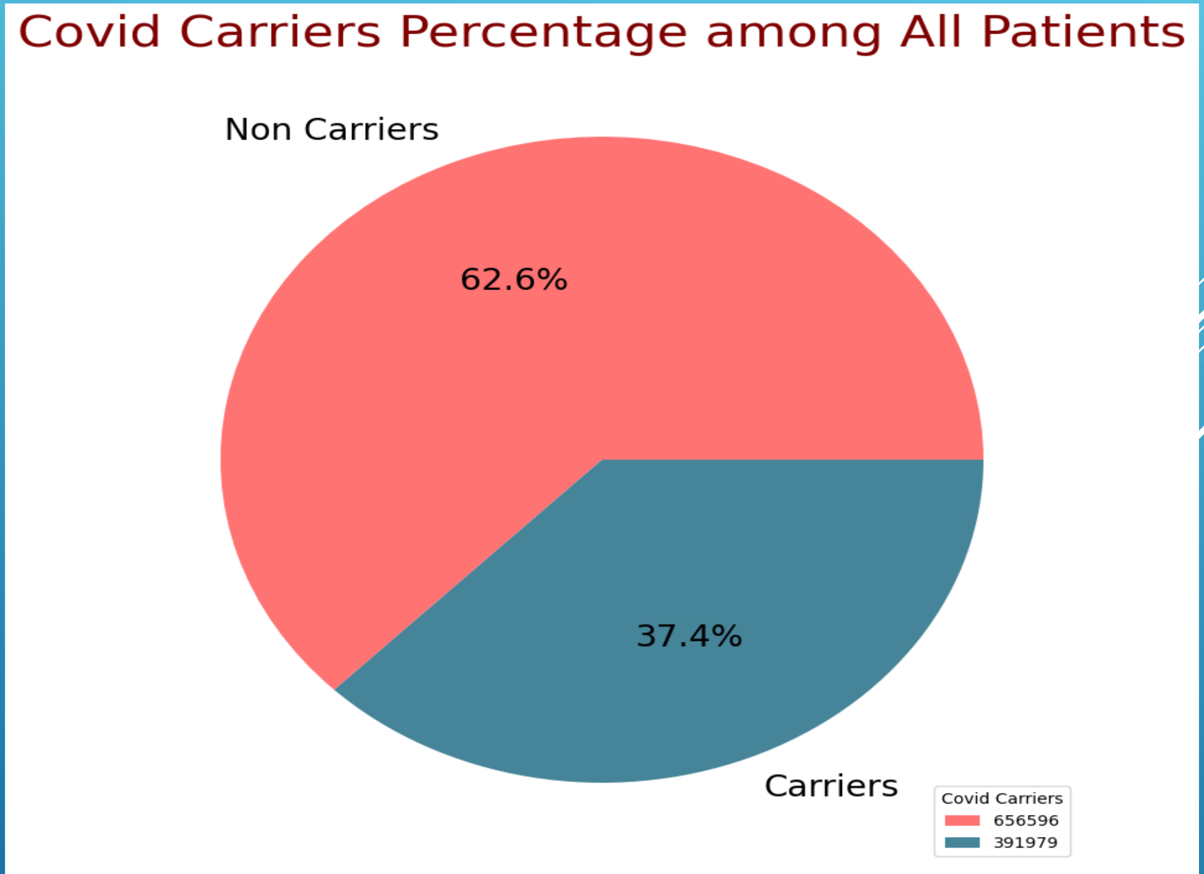


Exploratory Data Analysis

Covid statistics



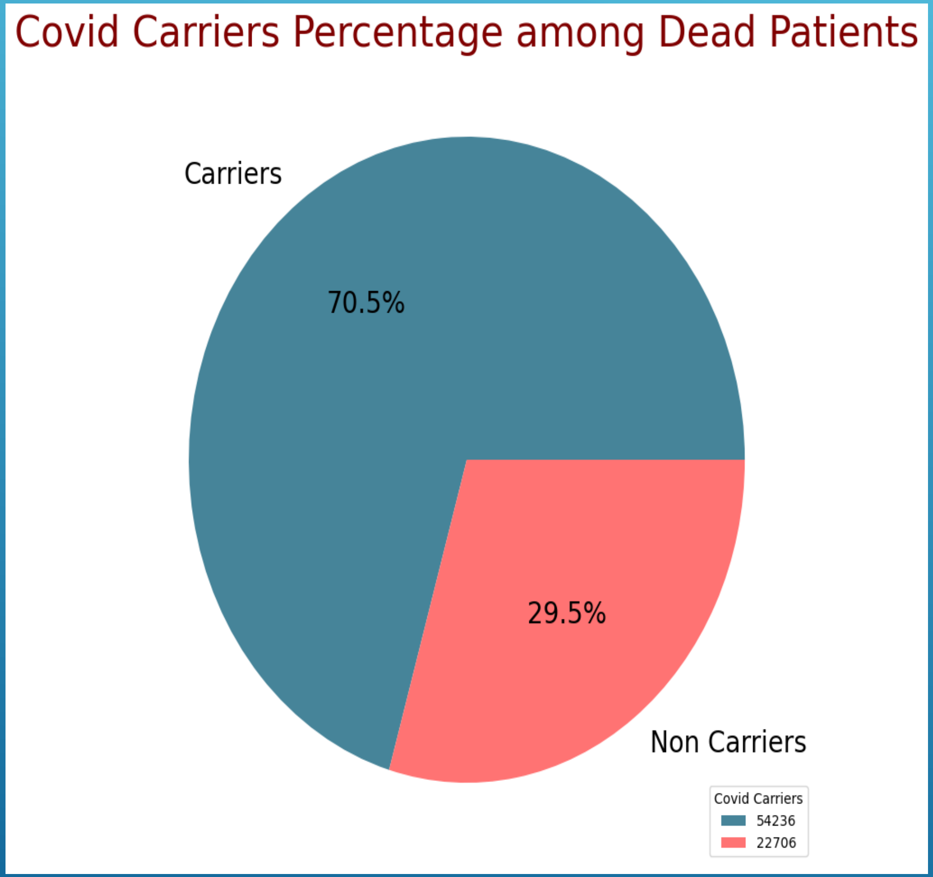
Covid Carriers Percentage among Dead Patients

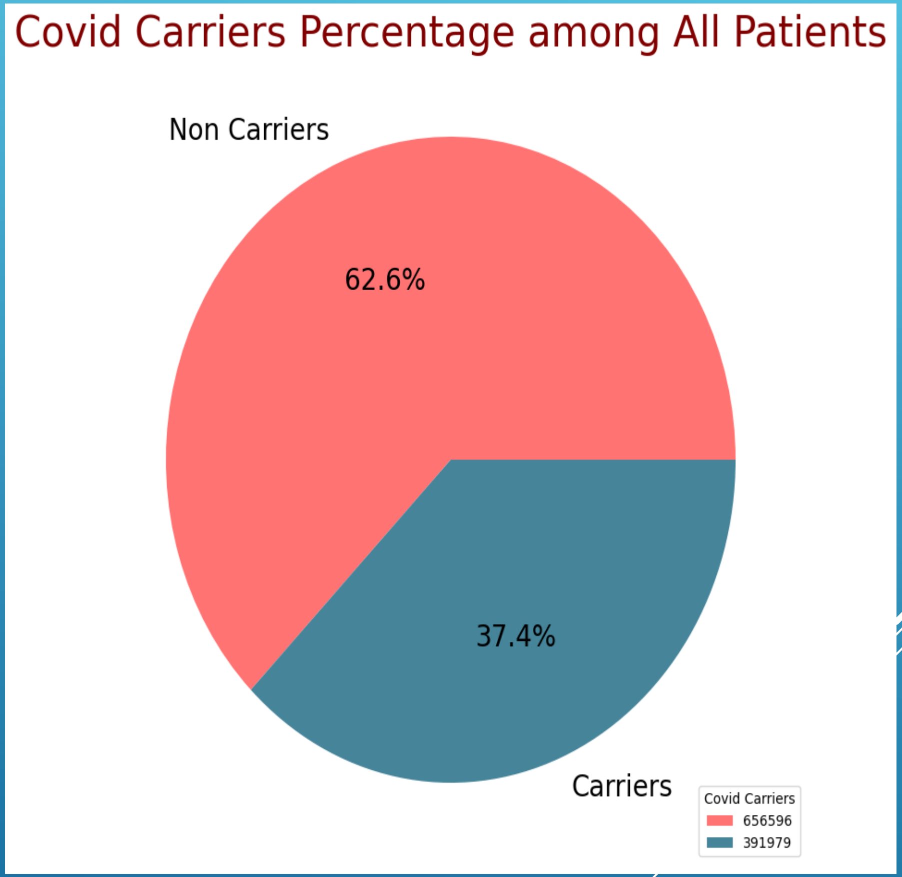


Covid Carriers Percentage among All Patients

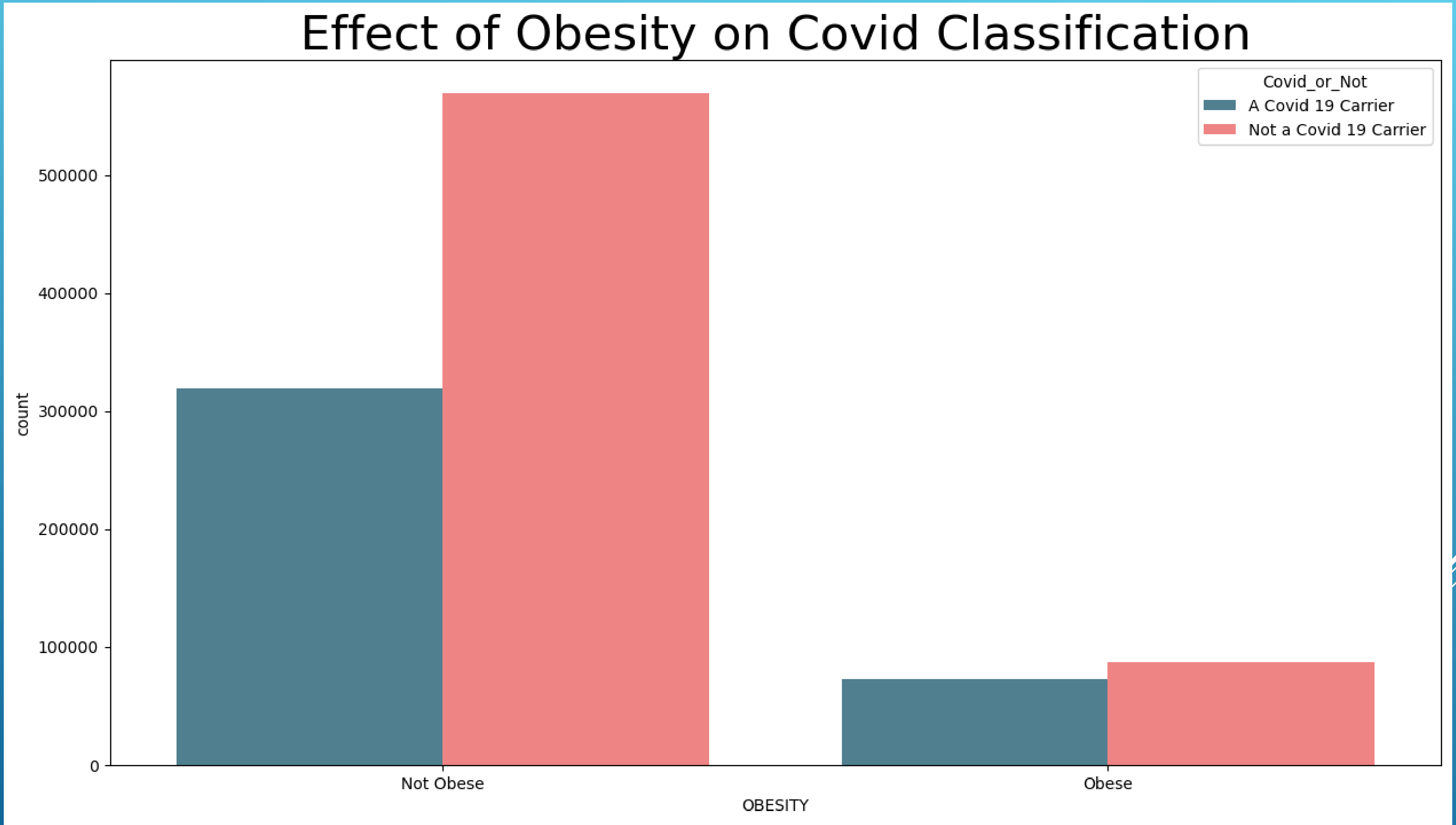
Exploratory Data Analysis

Covid statistics

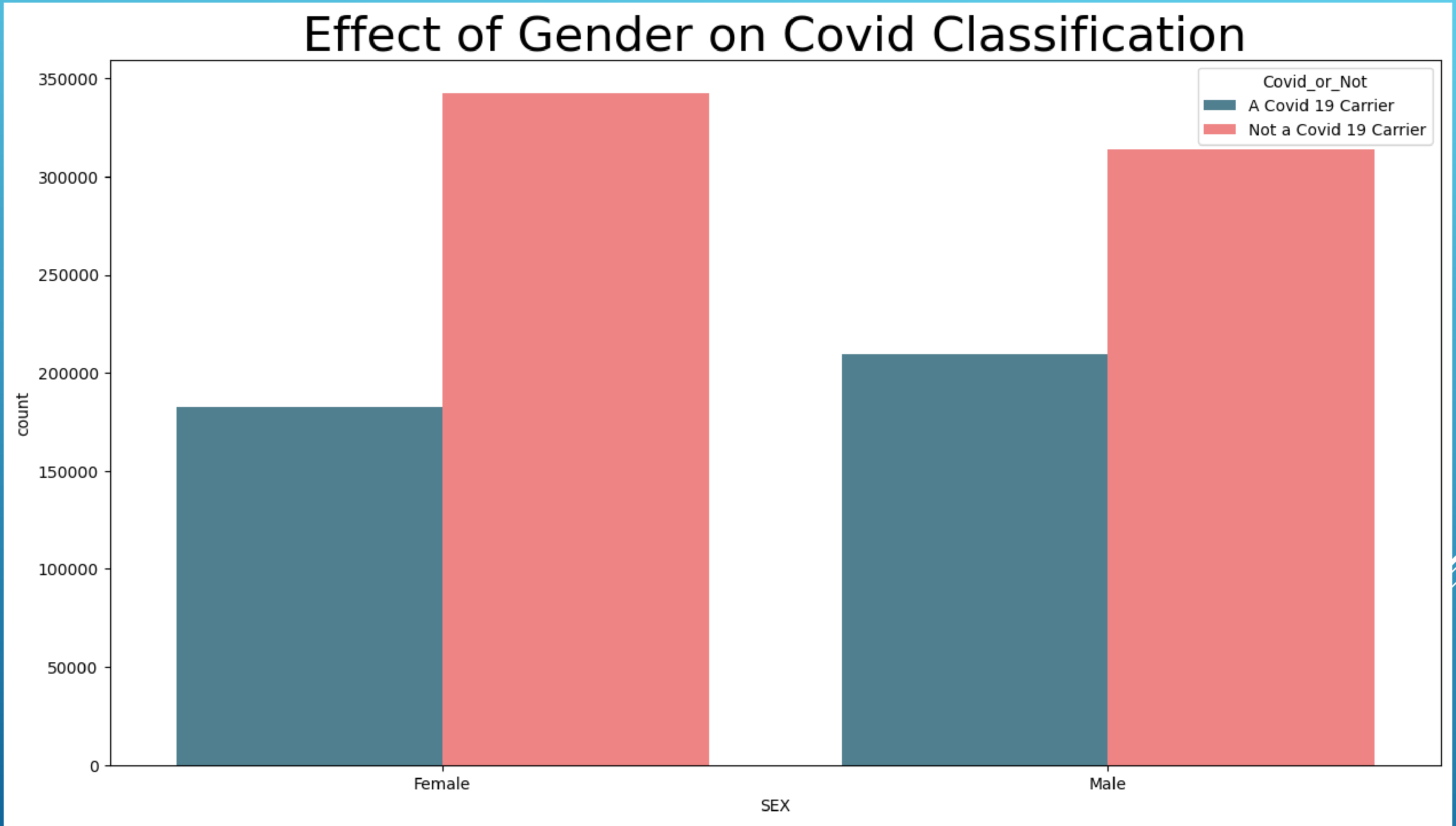




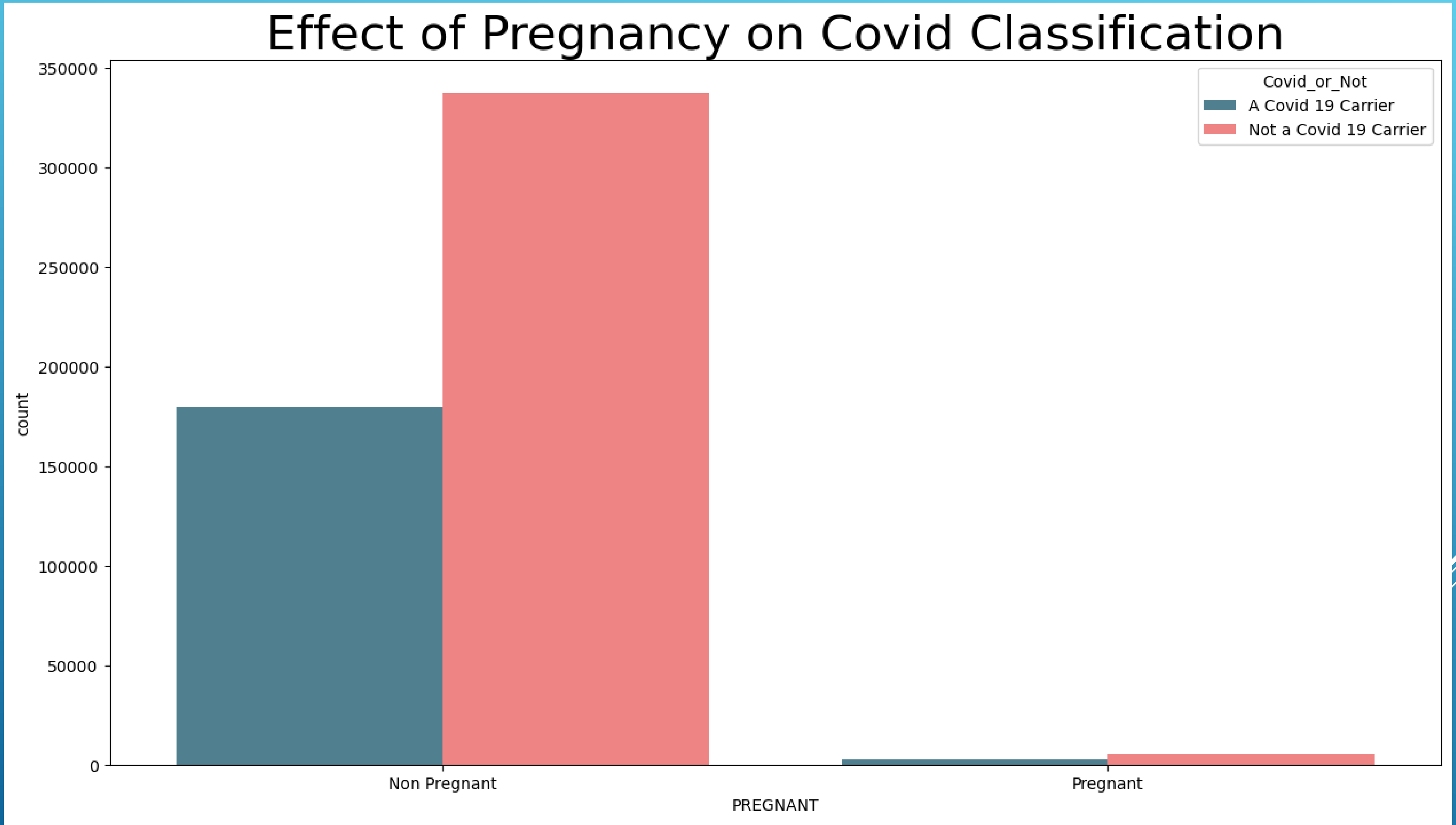
Exploratory Data Analysis



Exploratory Data Analysis

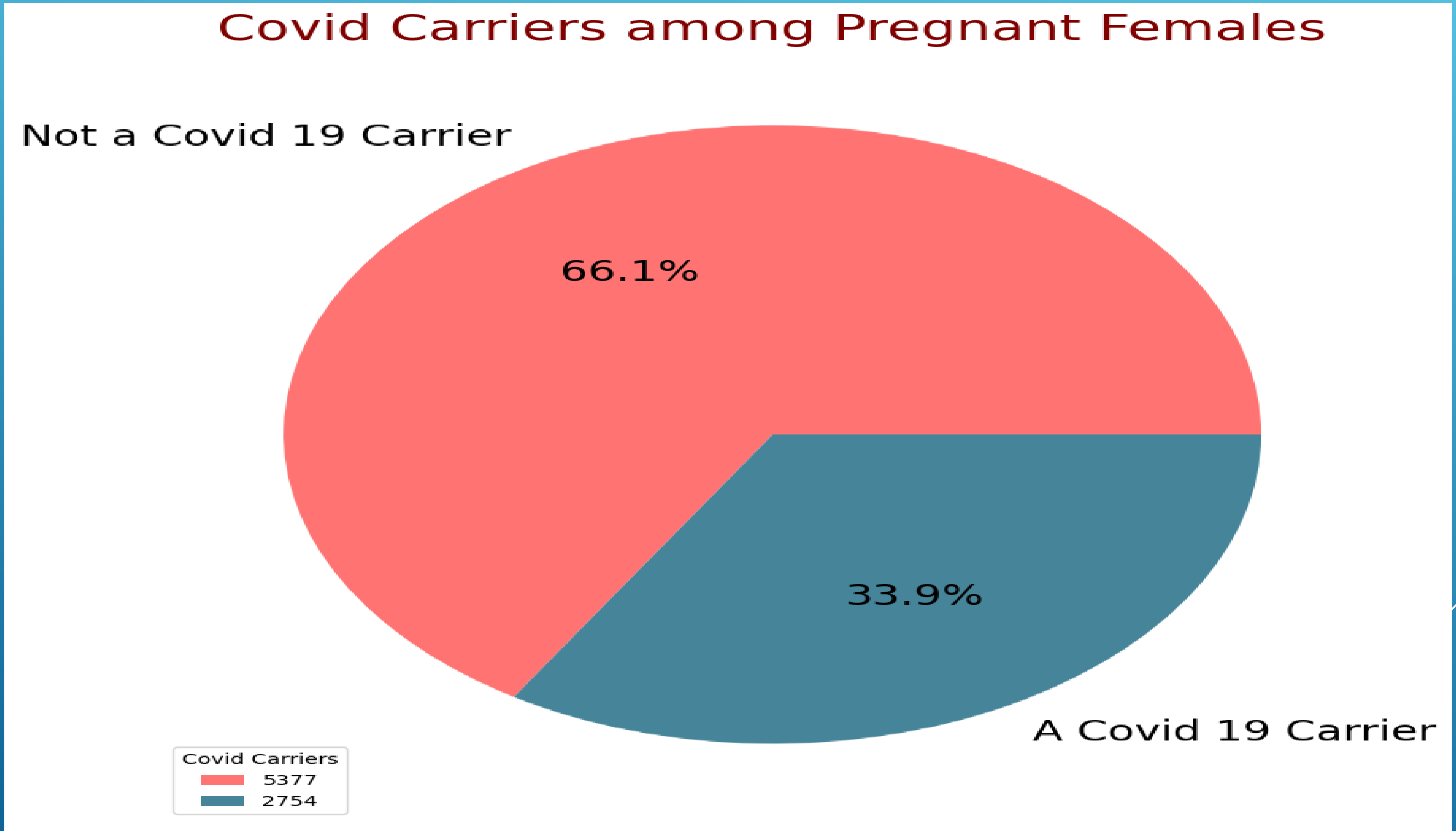


Exploratory Data Analysis



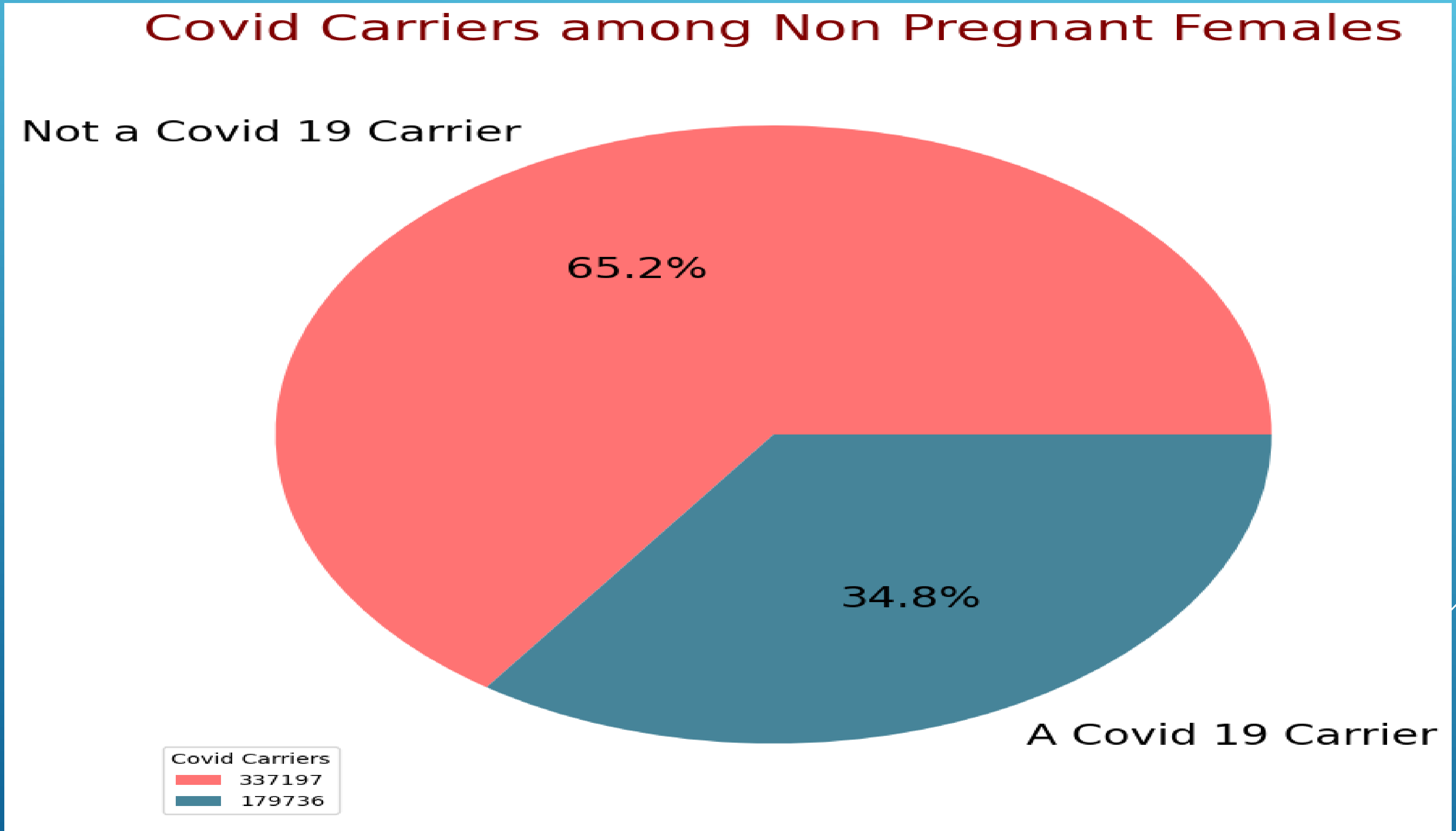
Exploratory Data Analysis

Covid statistics



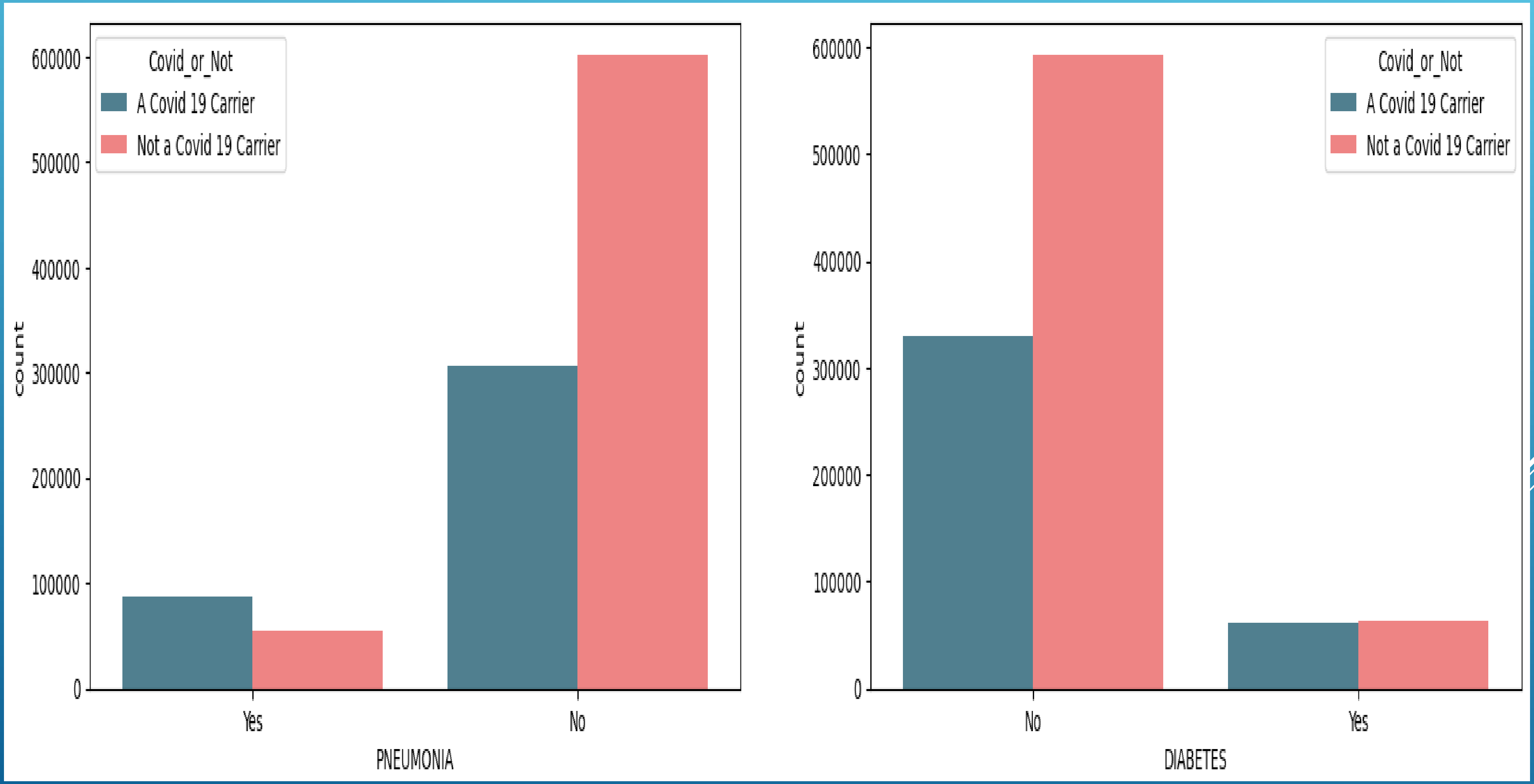
Exploratory Data Analysis

Covid statistics



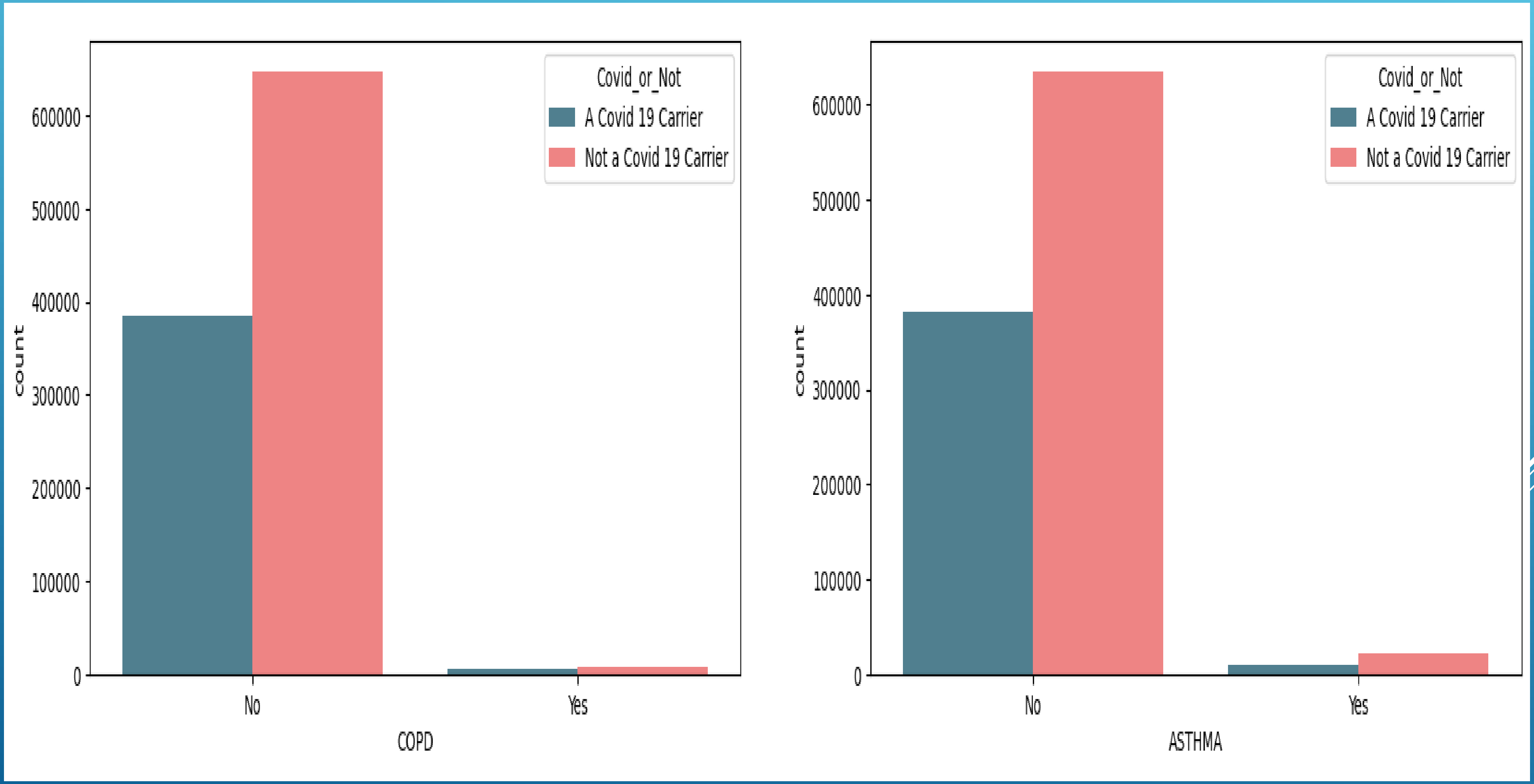
Exploratory Data Analysis

Effect of diseases on Covid



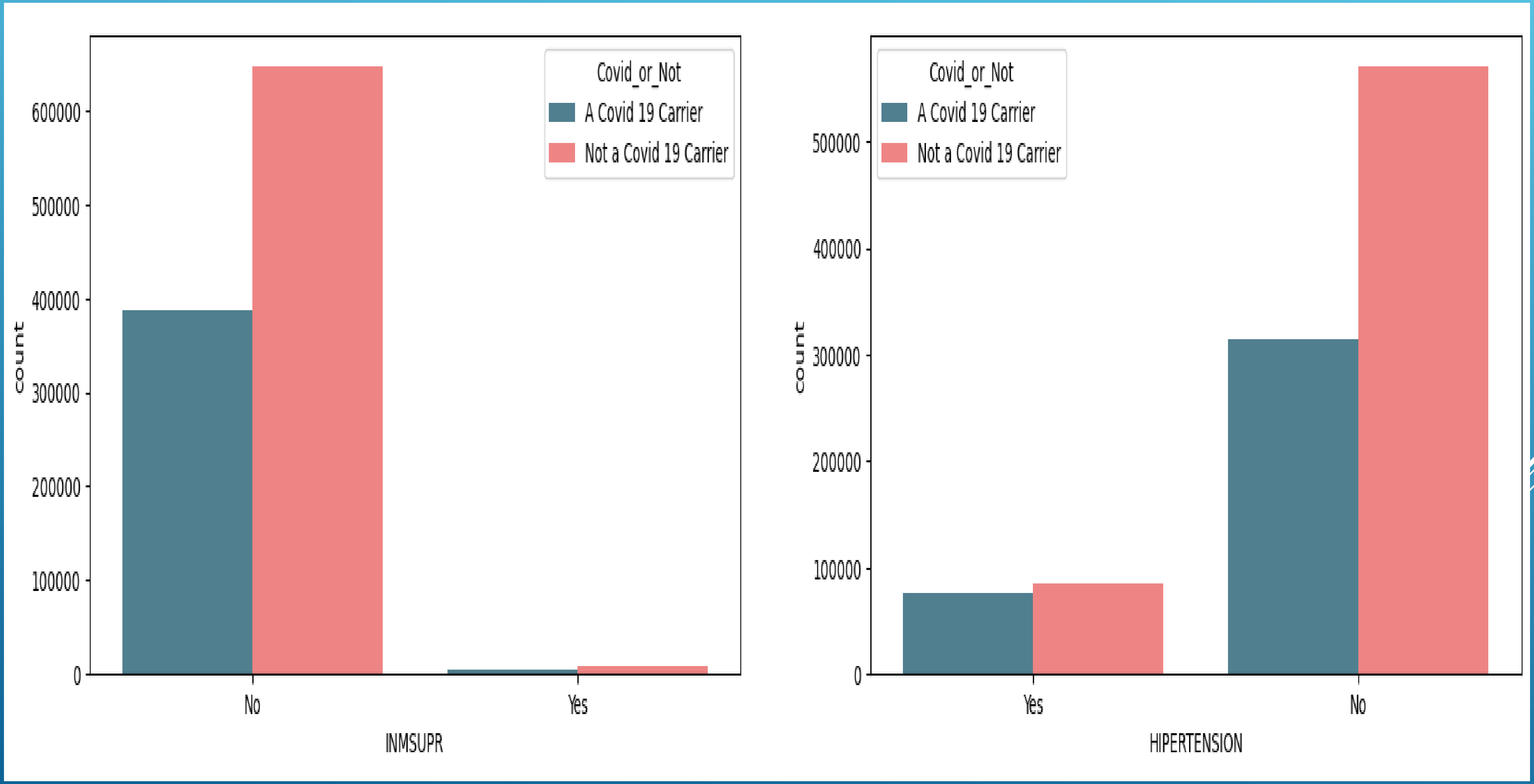
Exploratory Data Analysis

Effect of diseases on Covid



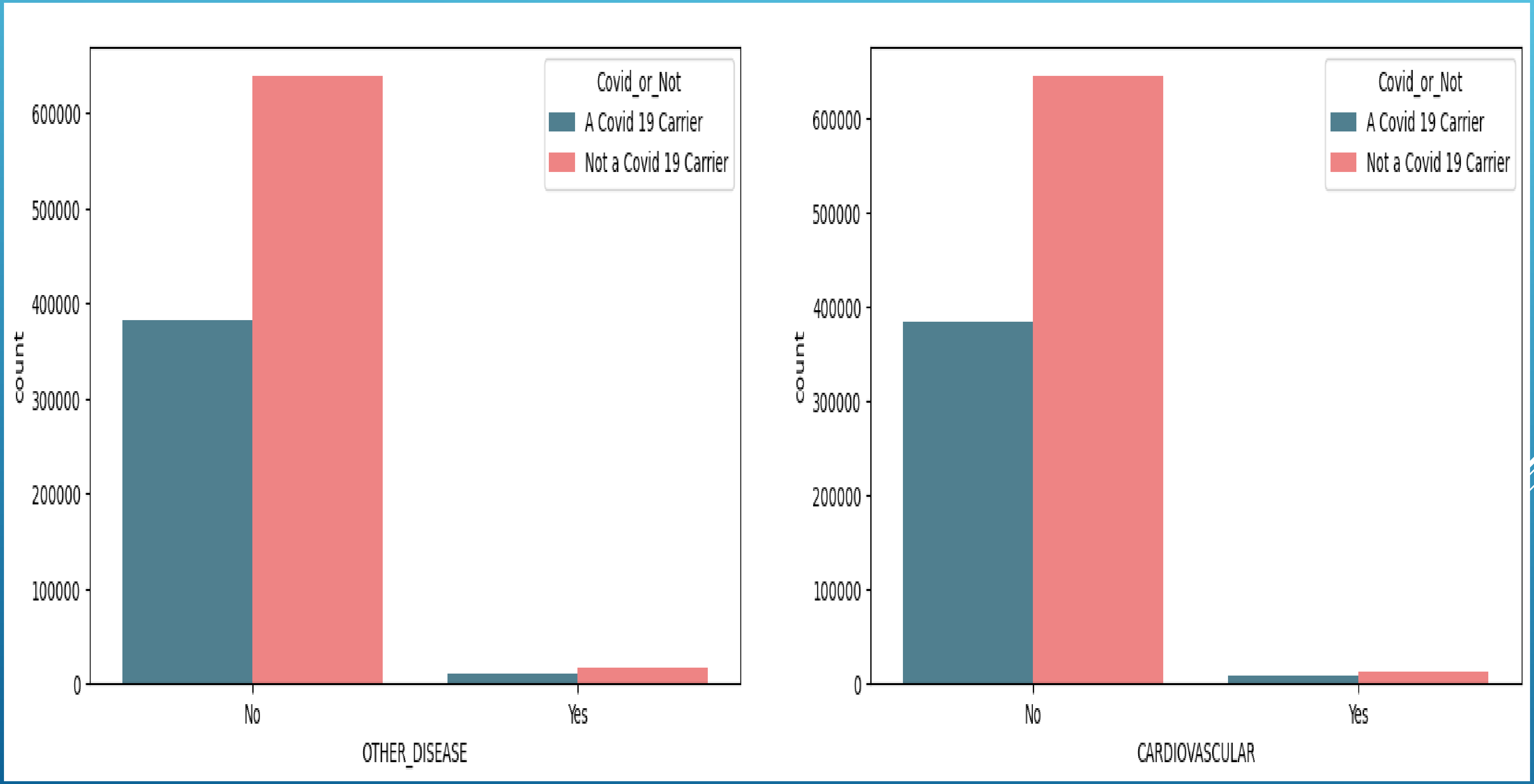
Exploratory Data Analysis

Effect of diseases on Covid



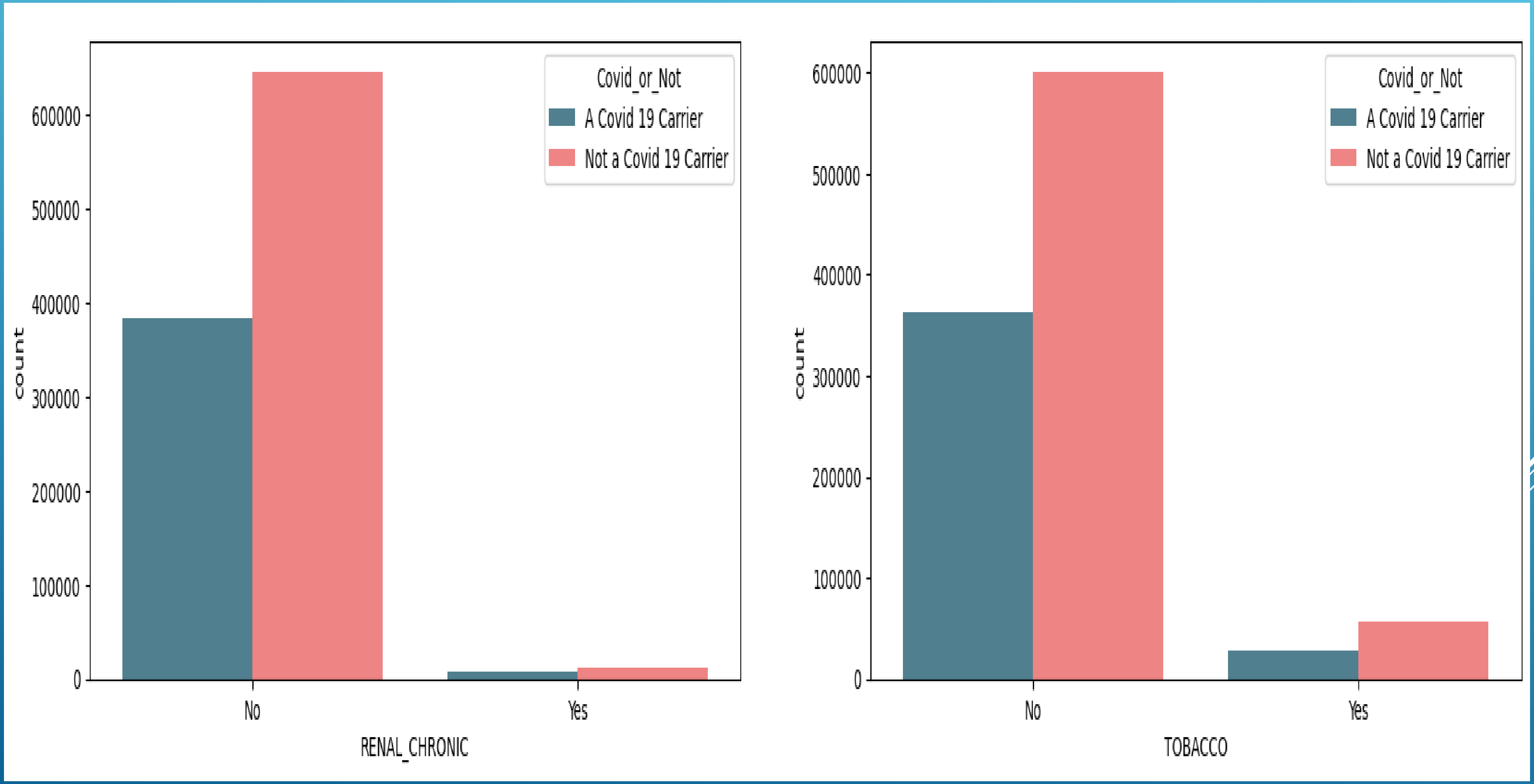
Exploratory Data Analysis

Effect of diseases on Covid



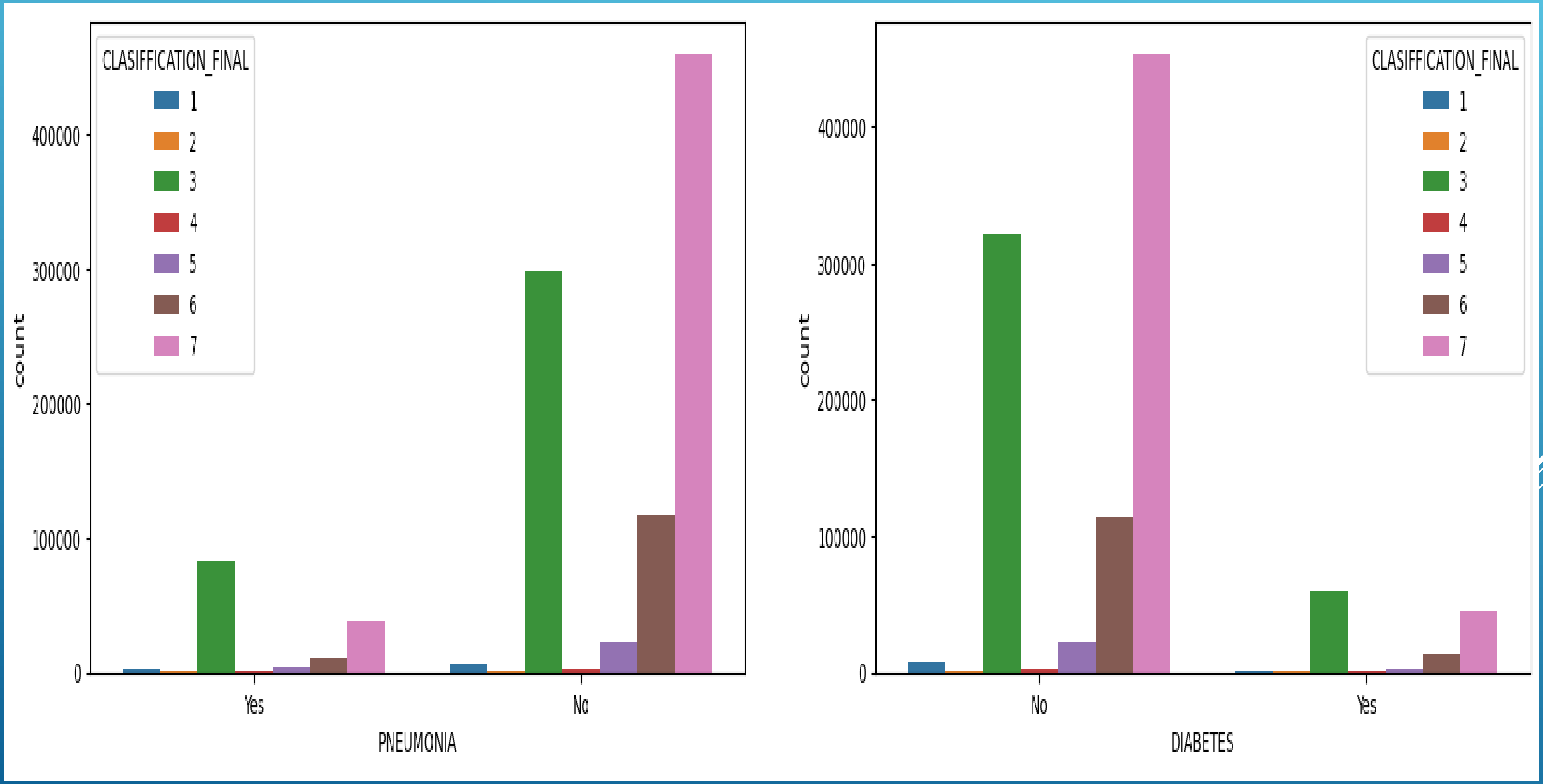
Exploratory Data Analysis

Effect of diseases on Covid



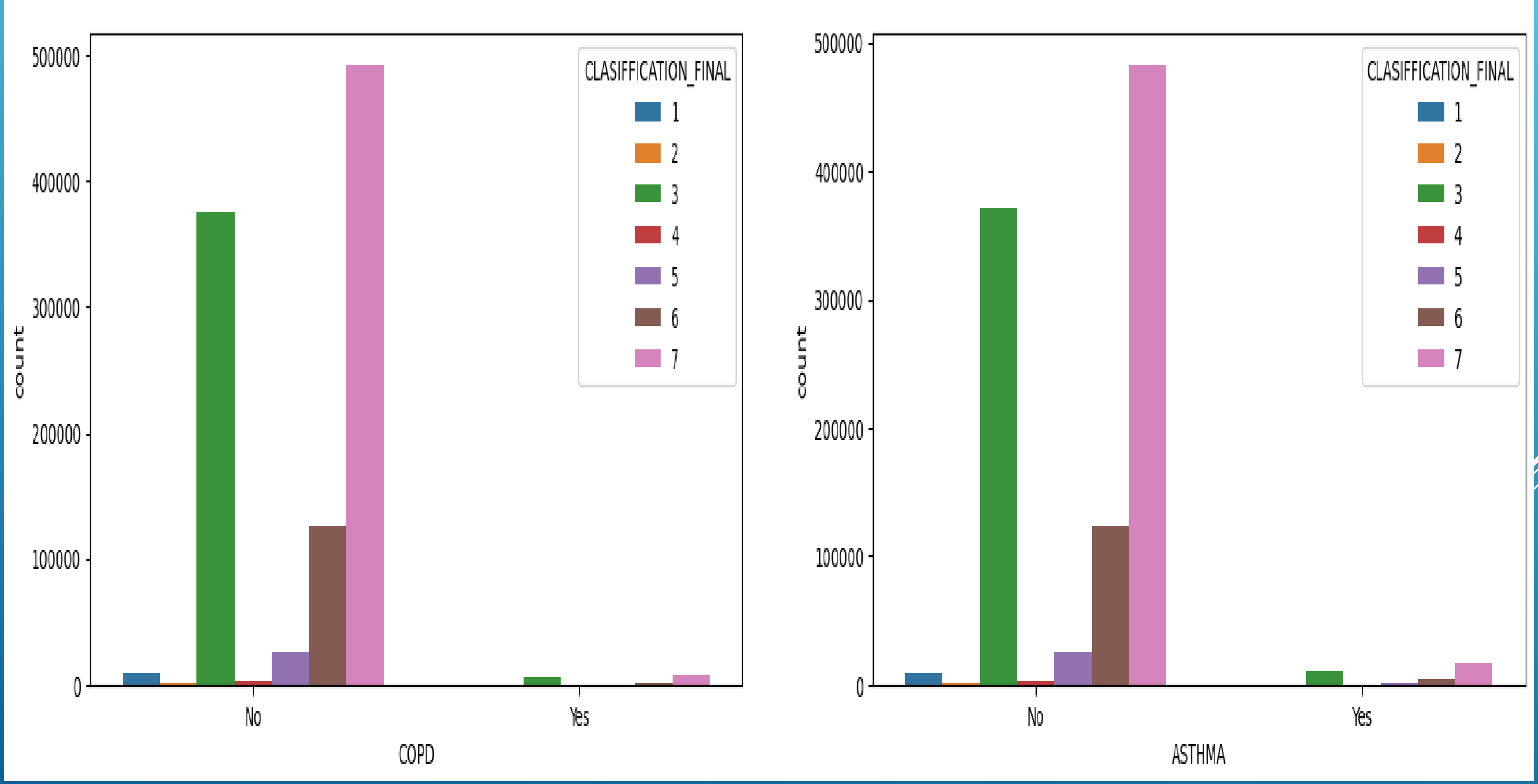
Exploratory Data Analysis

Effect of diseases on Covid Classification



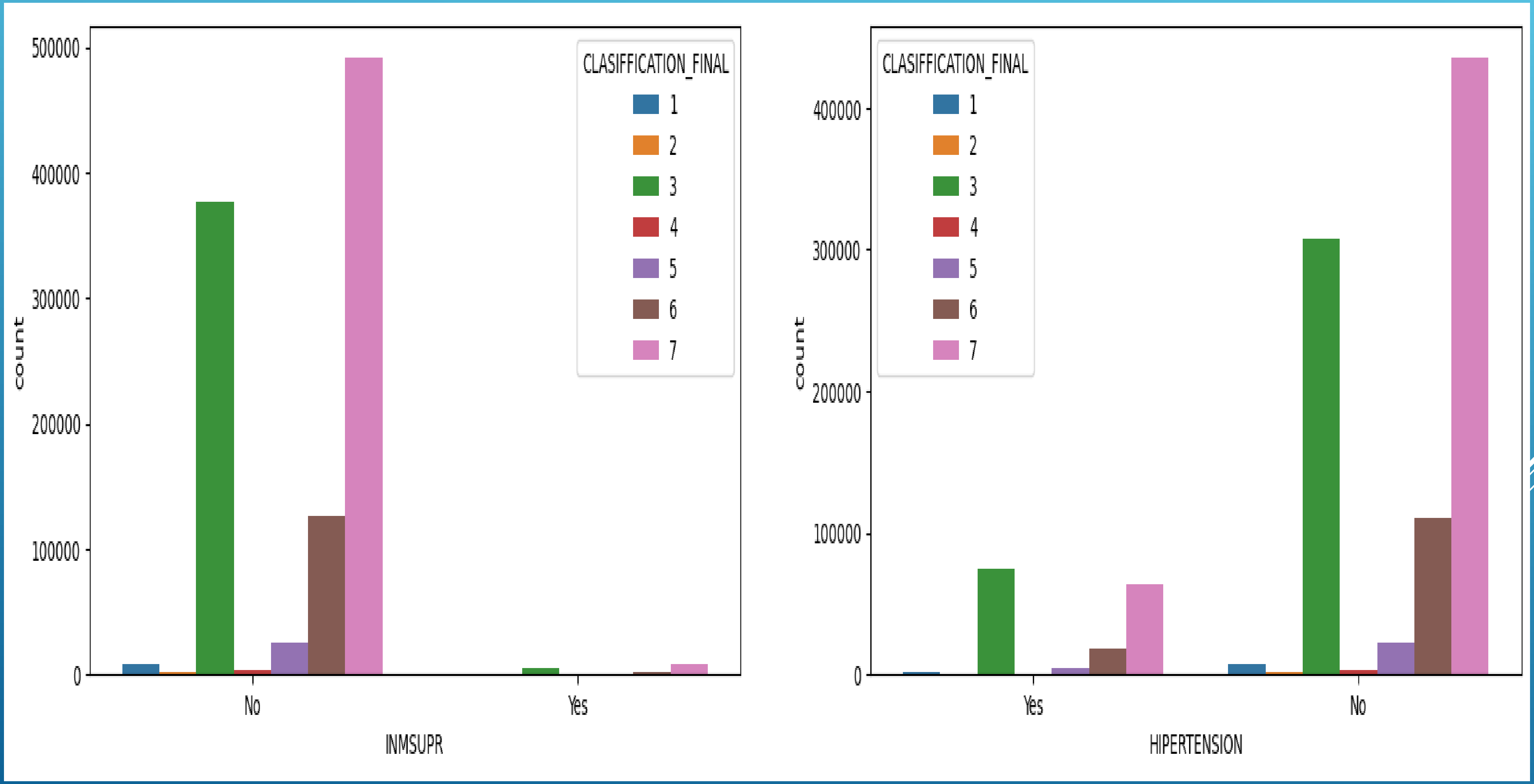
Exploratory Data Analysis

Effect of diseases on Covid Classification



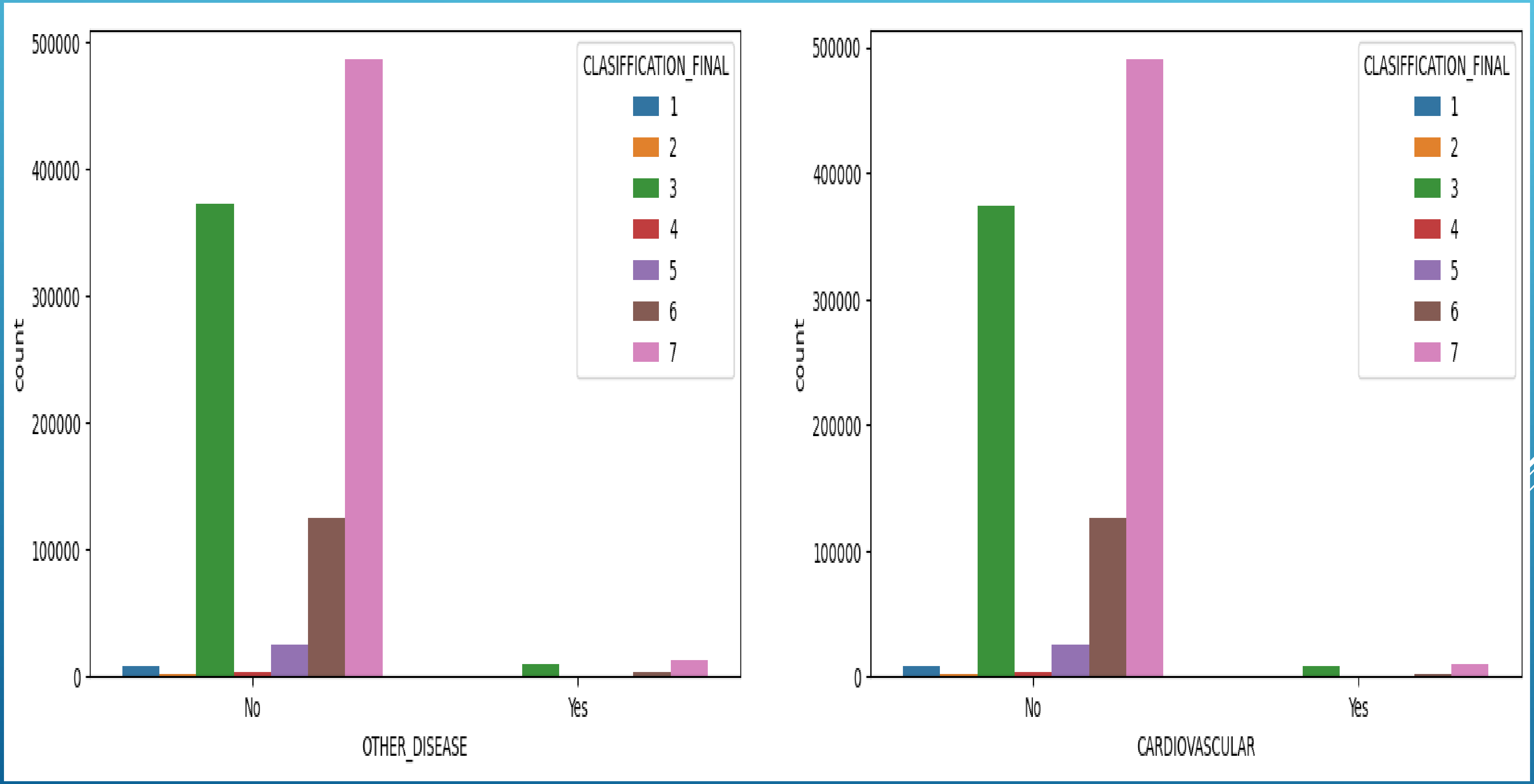
Exploratory Data Analysis

Effect of diseases on Covid Classification



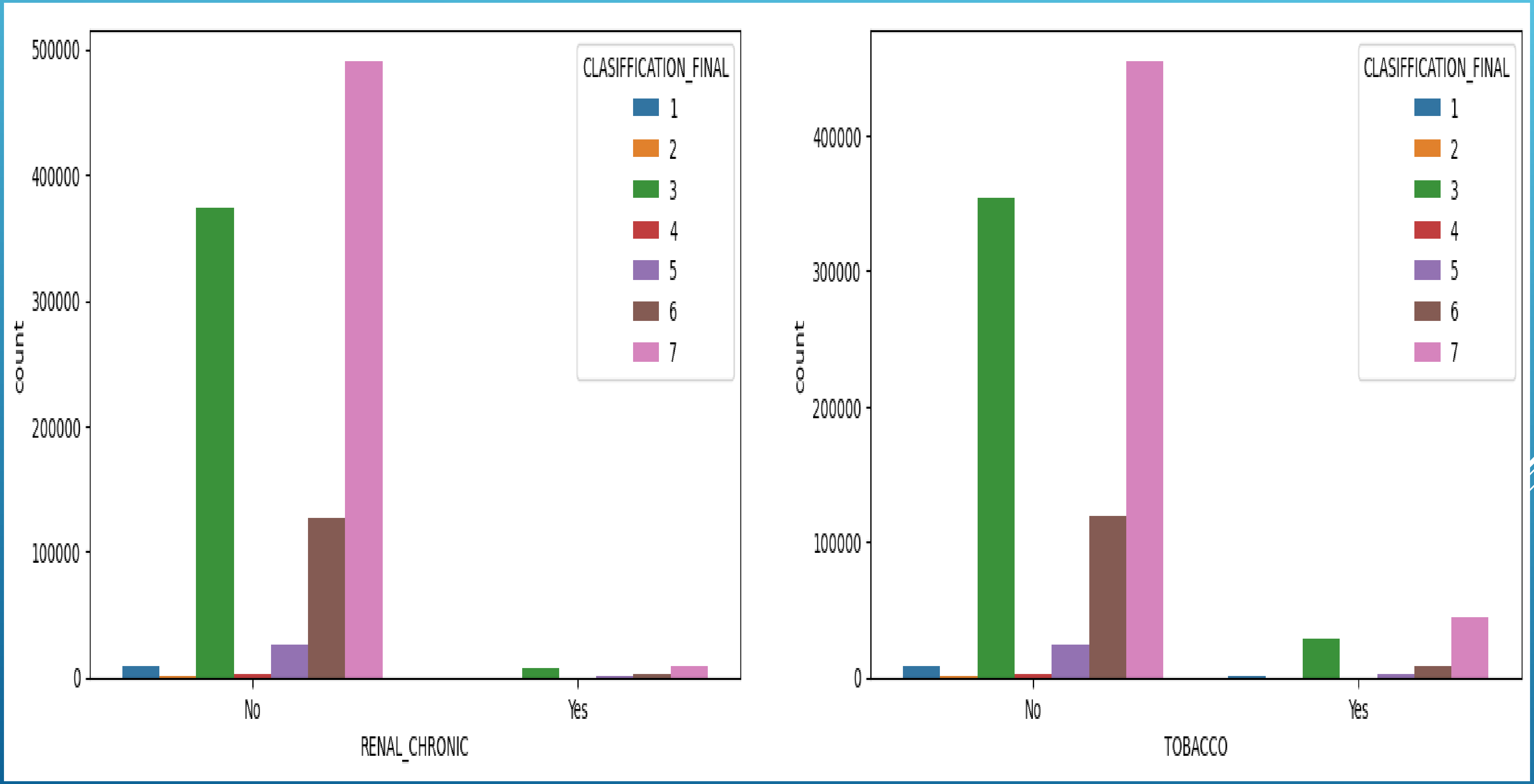
Exploratory Data Analysis

Effect of diseases on Covid Classification



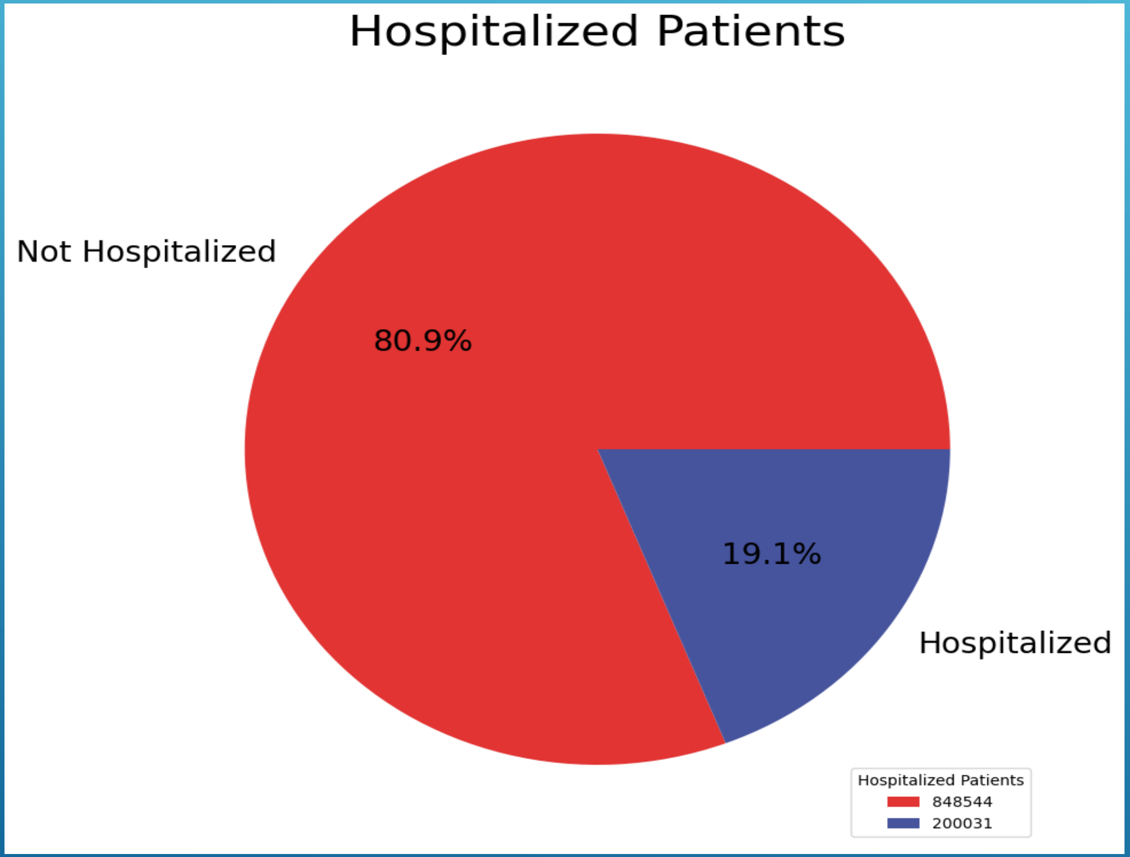
Exploratory Data Analysis

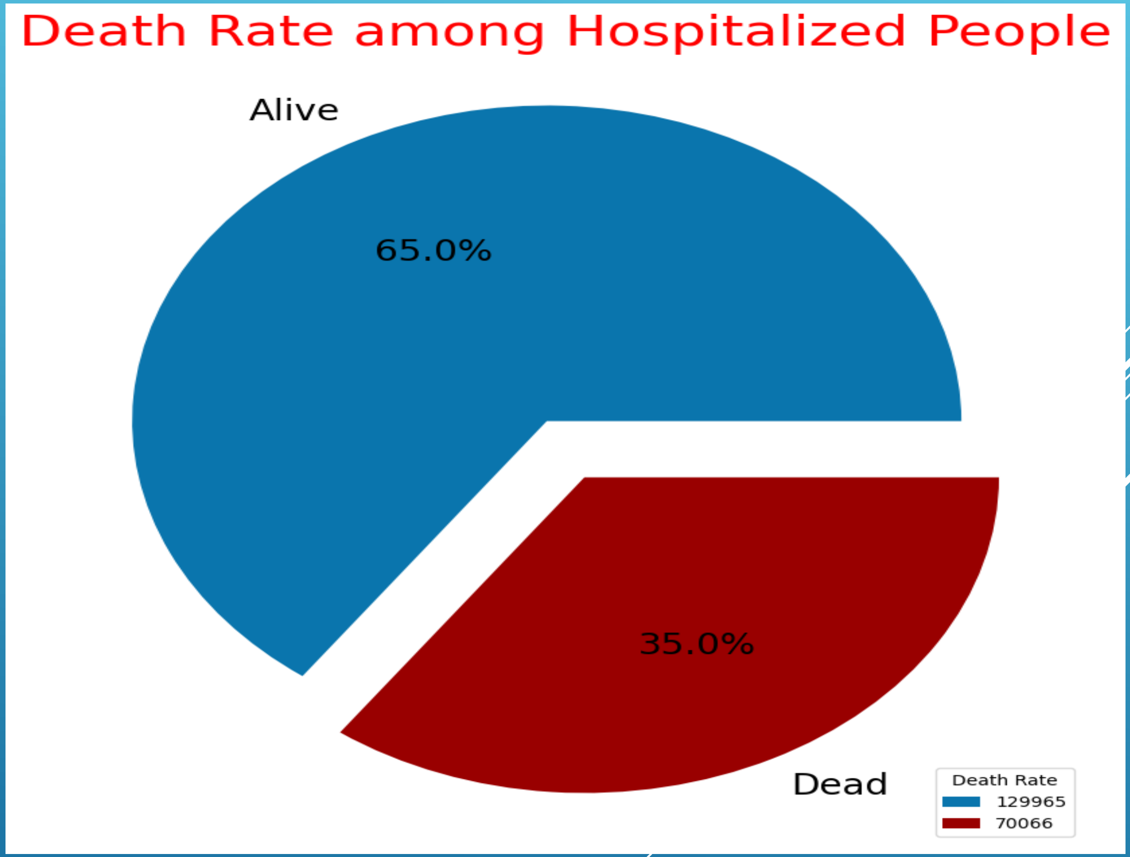
Effect of diseases on Covid Classification



Exploratory Data Analysis

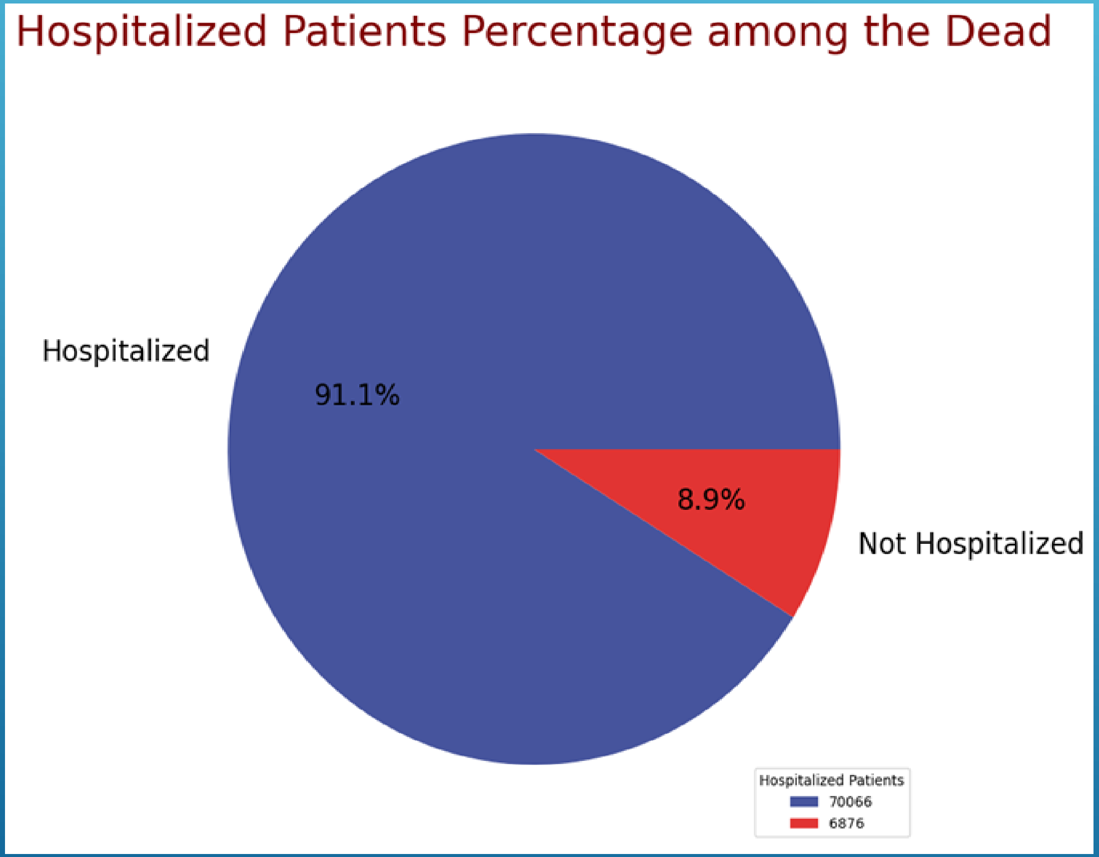
Hospital statistics

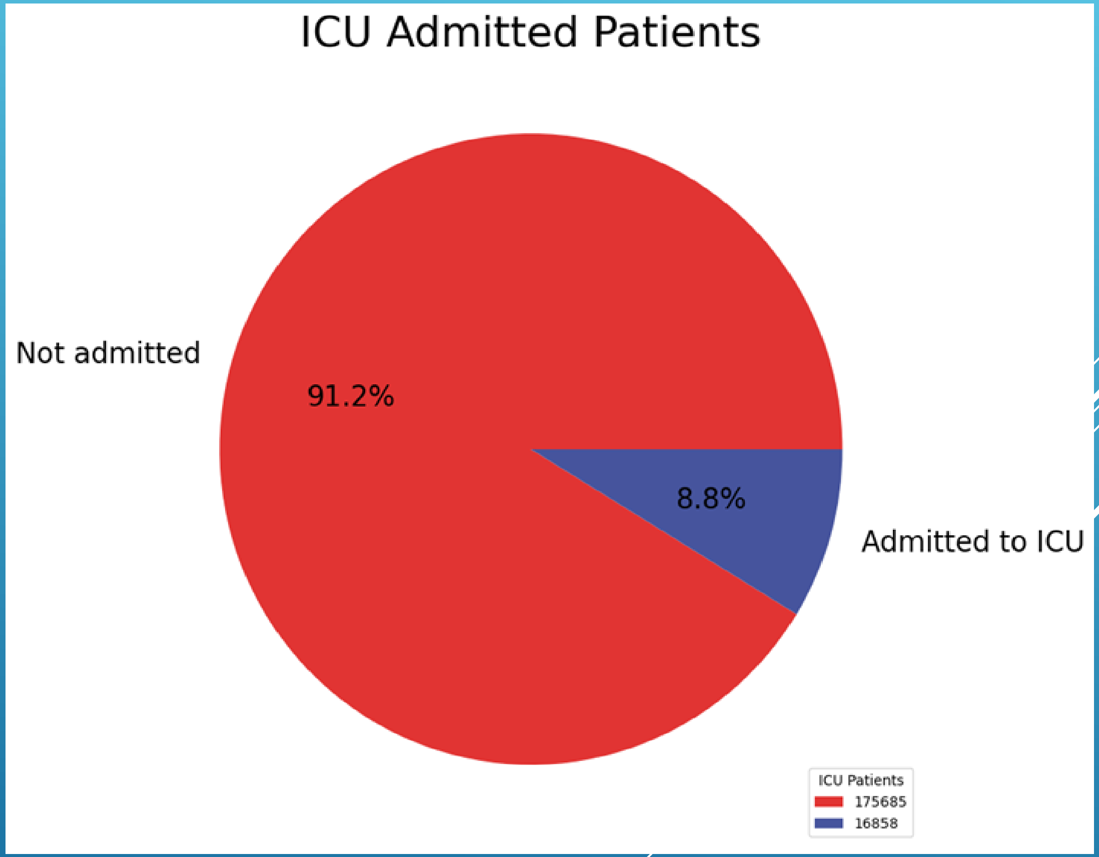




Exploratory Data Analysis

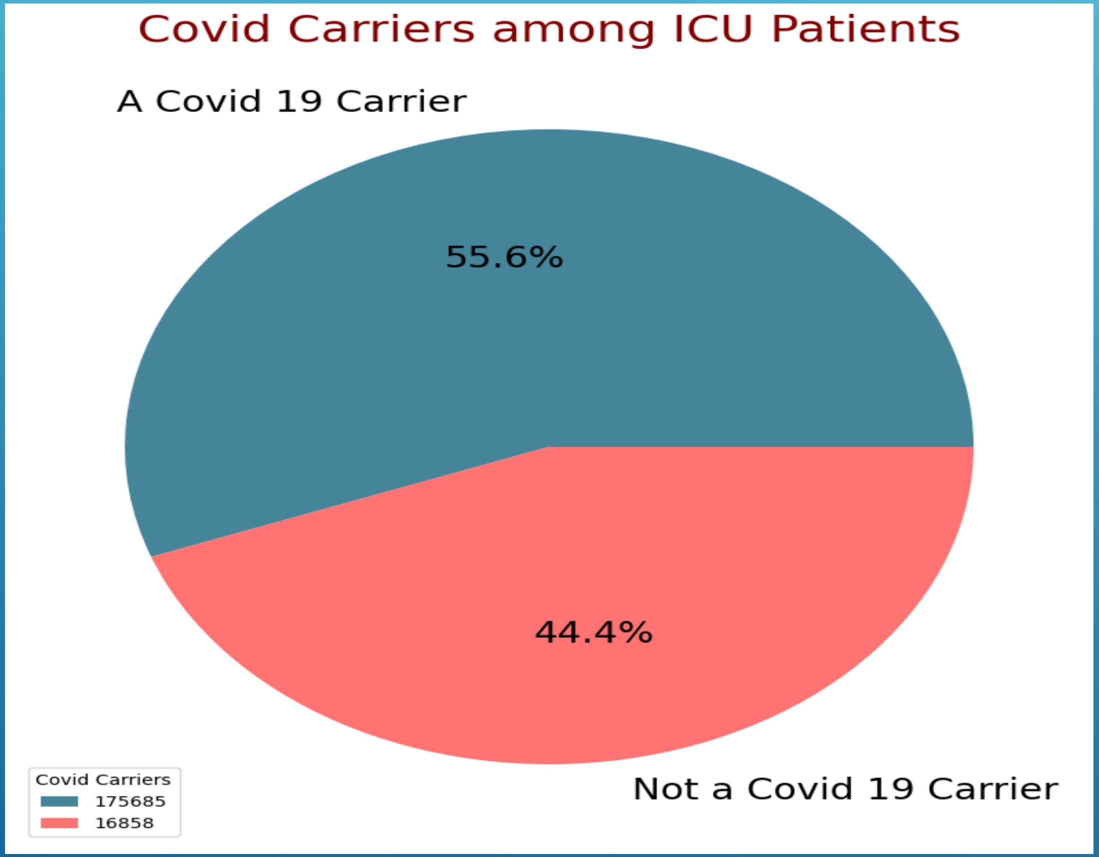
Hospital statistics

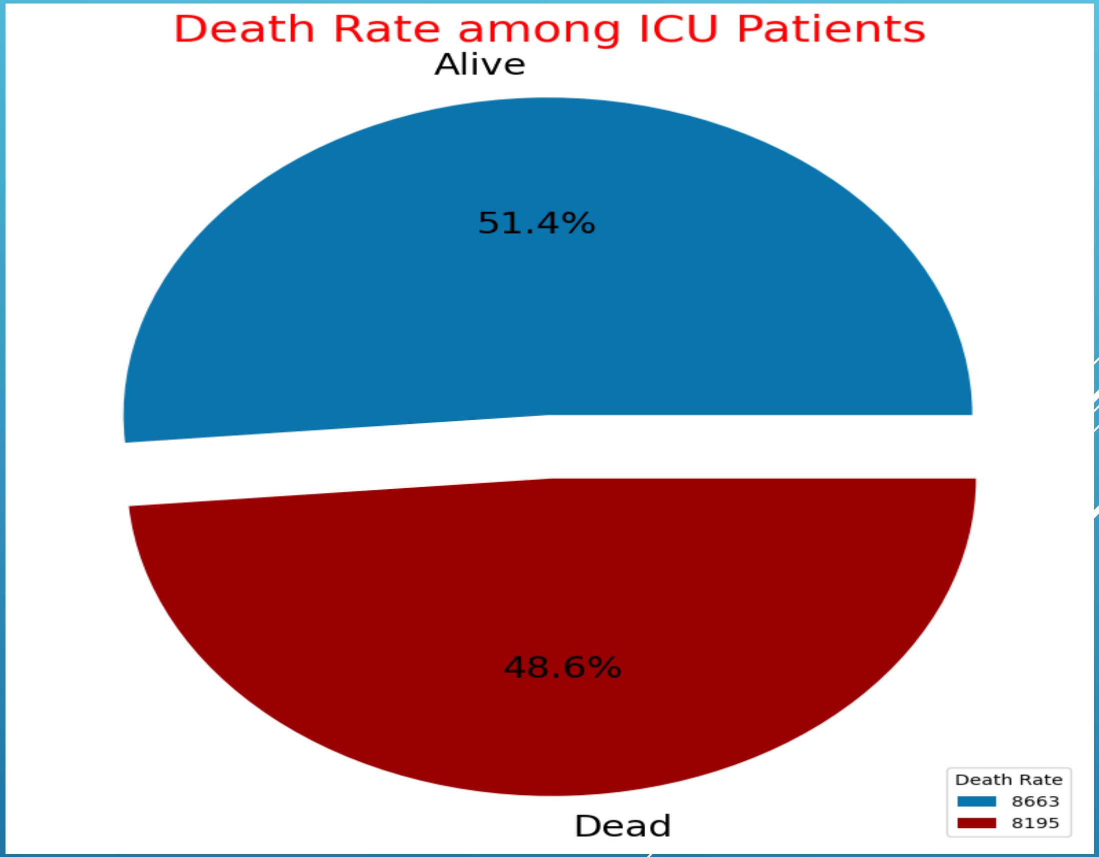




Exploratory Data Analysis

Hospital statistics

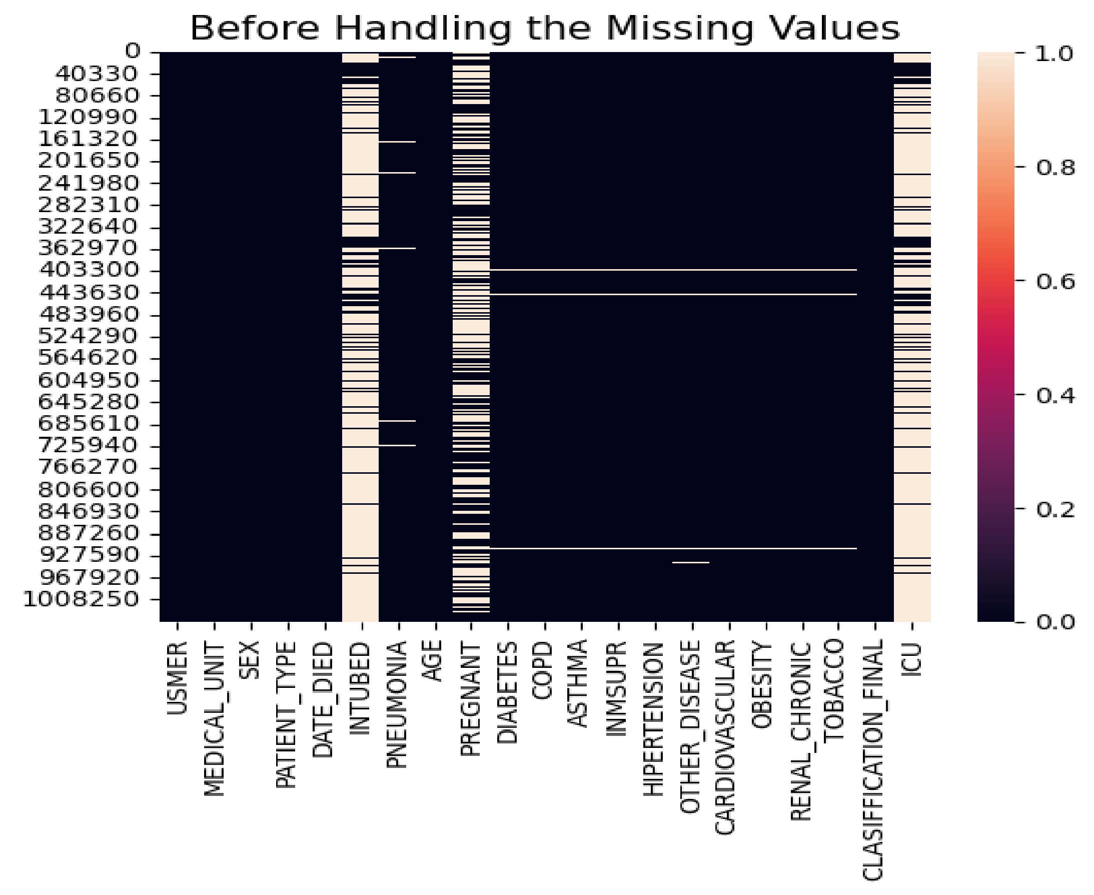


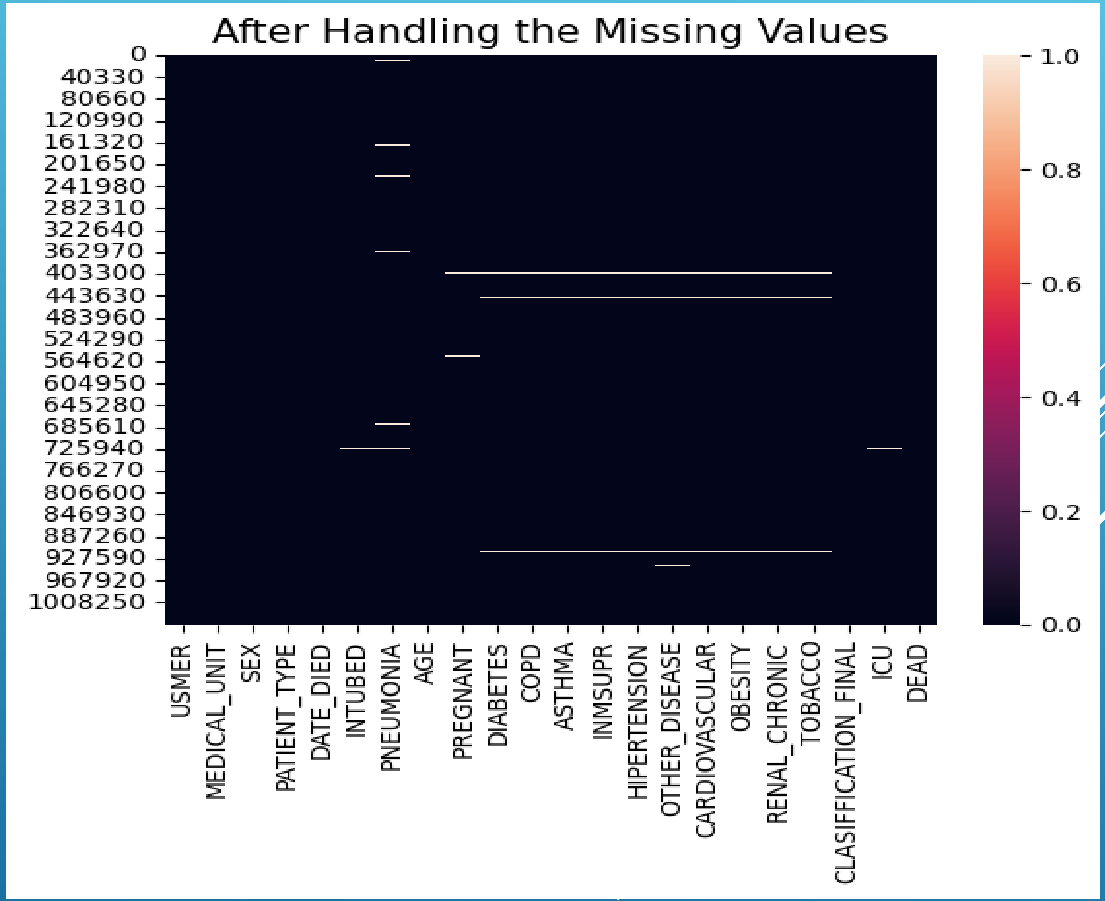


Data Pre-processing and Problems

Data Pre-processing and Problems

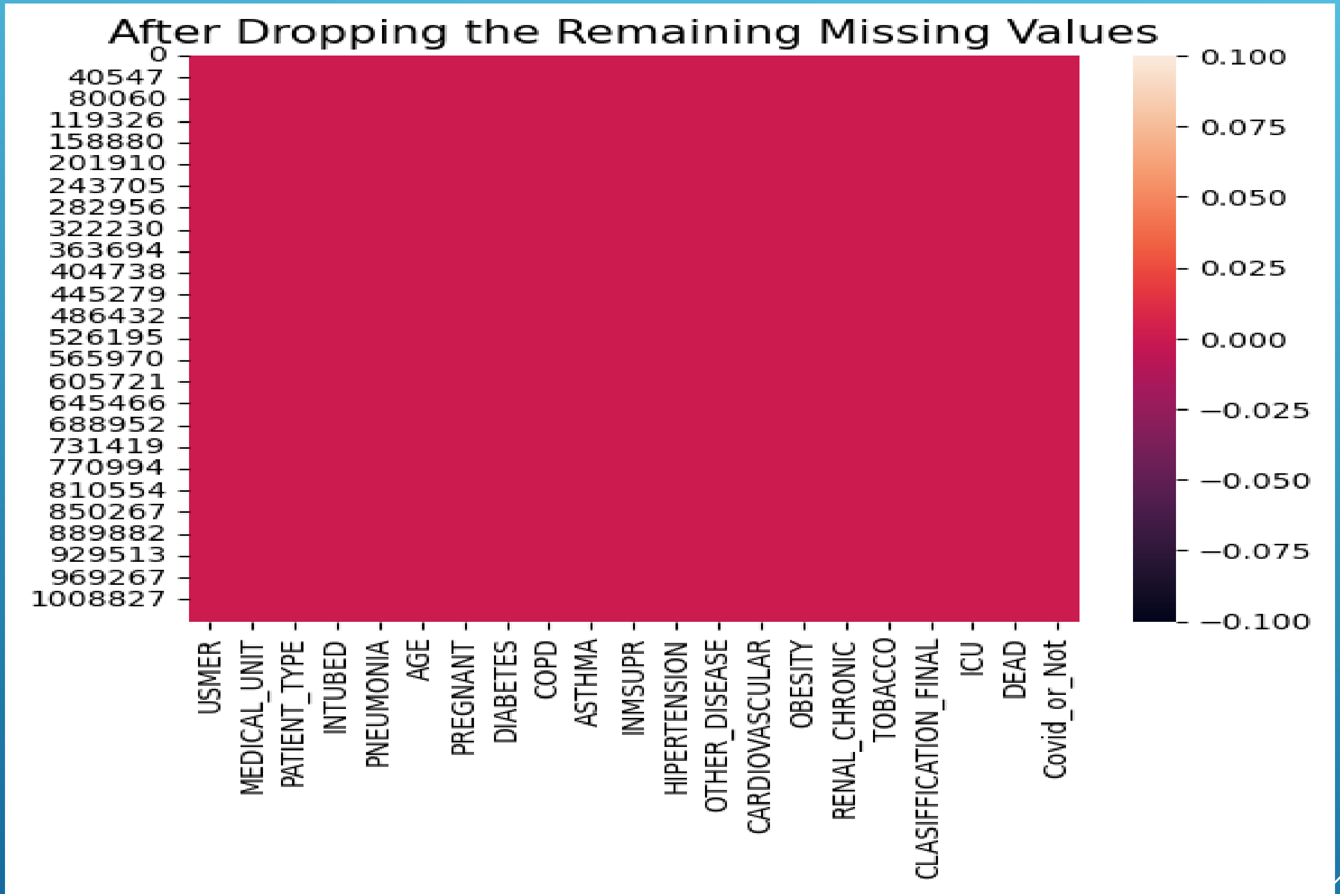
Handling Missing Values





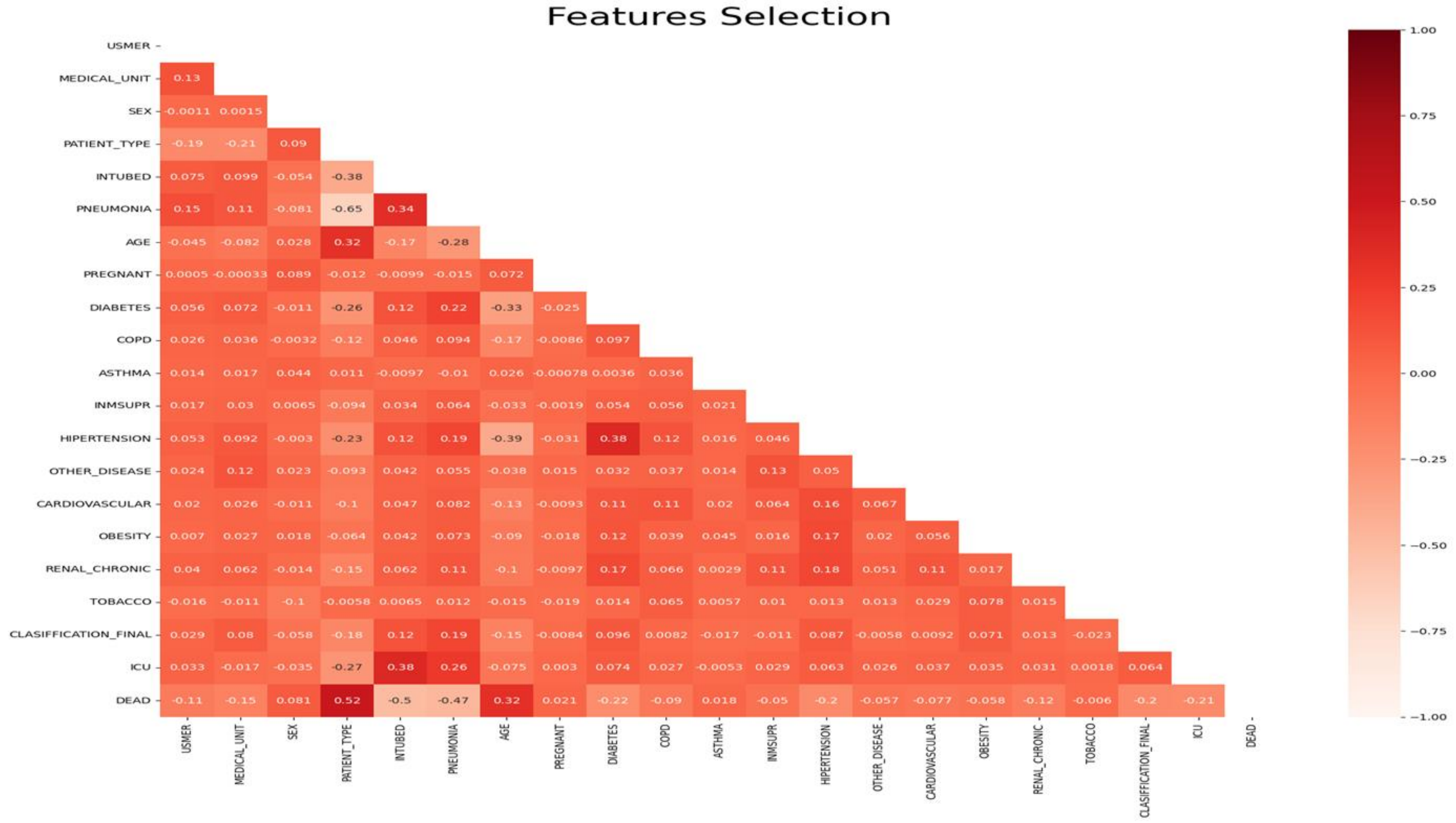
Data Pre-processing and Problems

Handling Missing Values



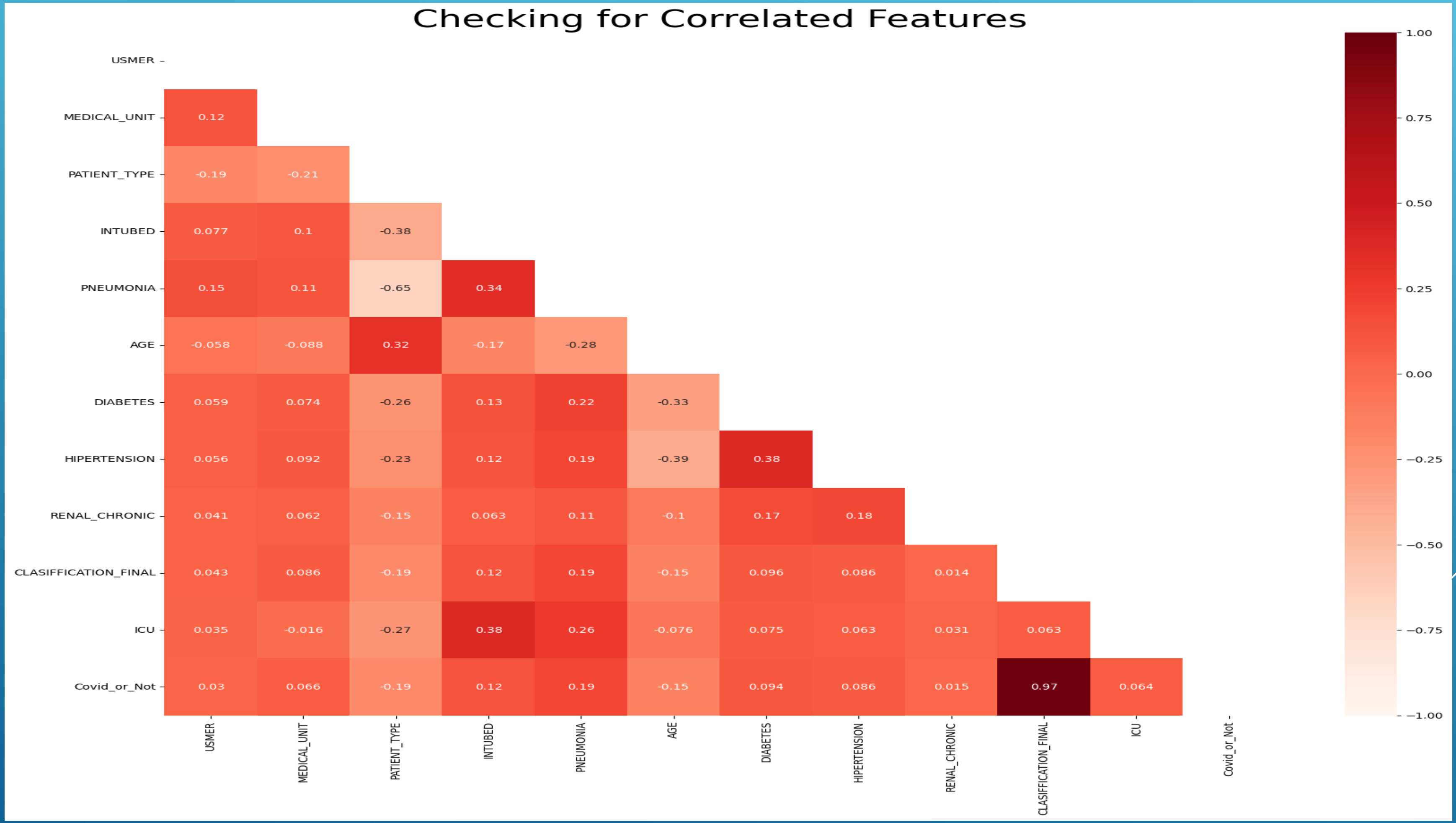
Data Pre-processing and Problems

Features Engineering



Data Pre-processing and Problems

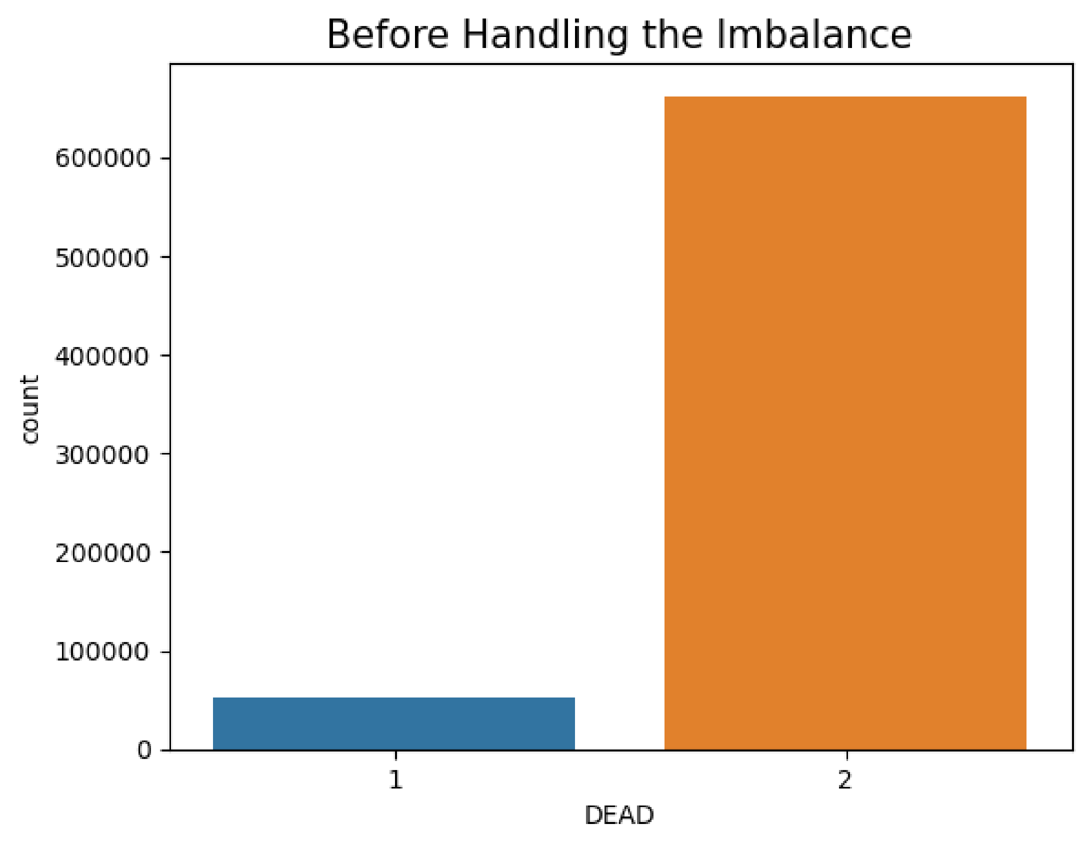
Features Engineering

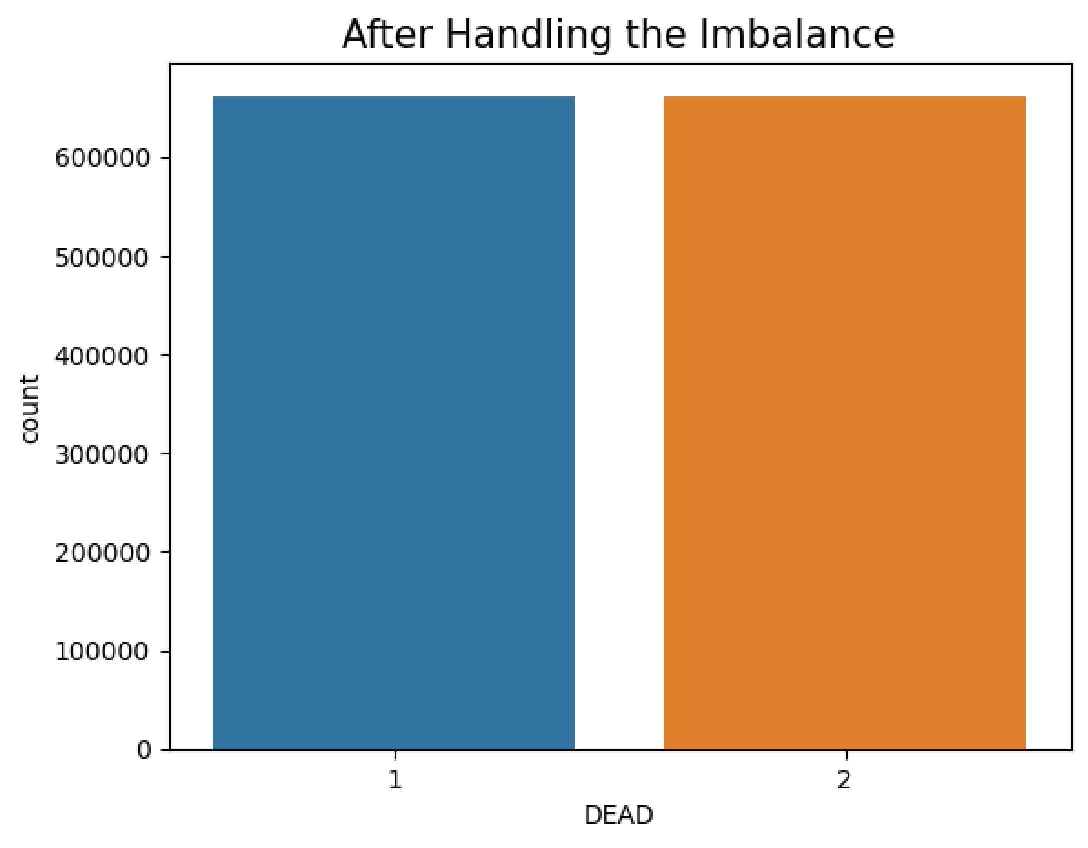


Checking for Correlated Features

Data Pre-processing and Problems

Handling Imbalance





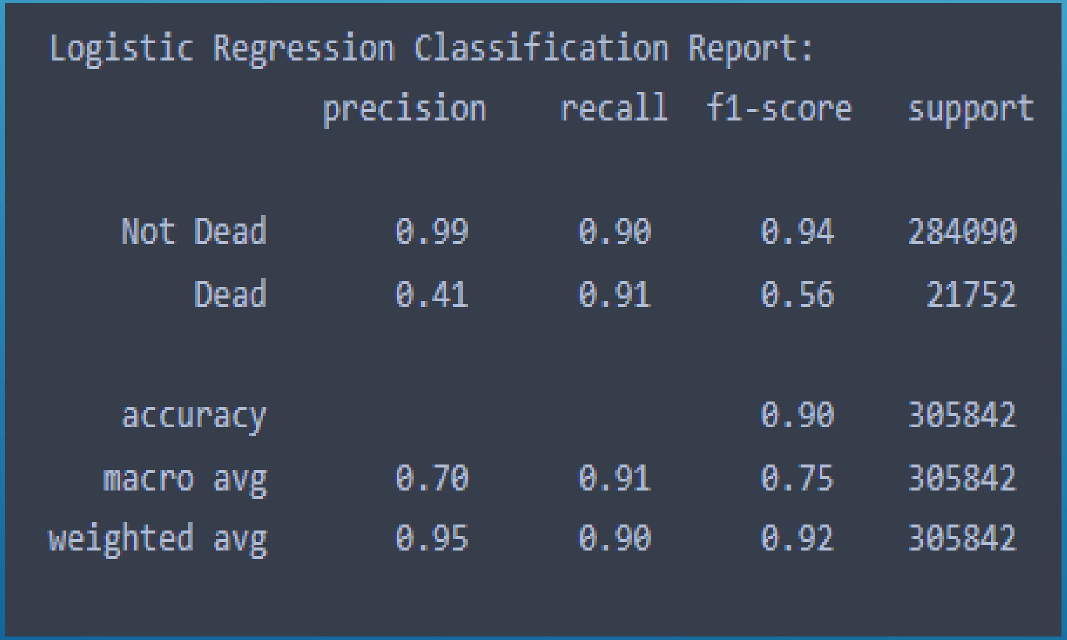
Modeling and Results

Algorithm Used (Model)	Accuracy	Recall (Lowest)
Logistic Regression	90%	90%
Decision Tree	92.3%	66%
Random Forest	91.5%	78%
Naive Bayes	88.1%	88%
XGBoost	92%	83%

Modeling and Results

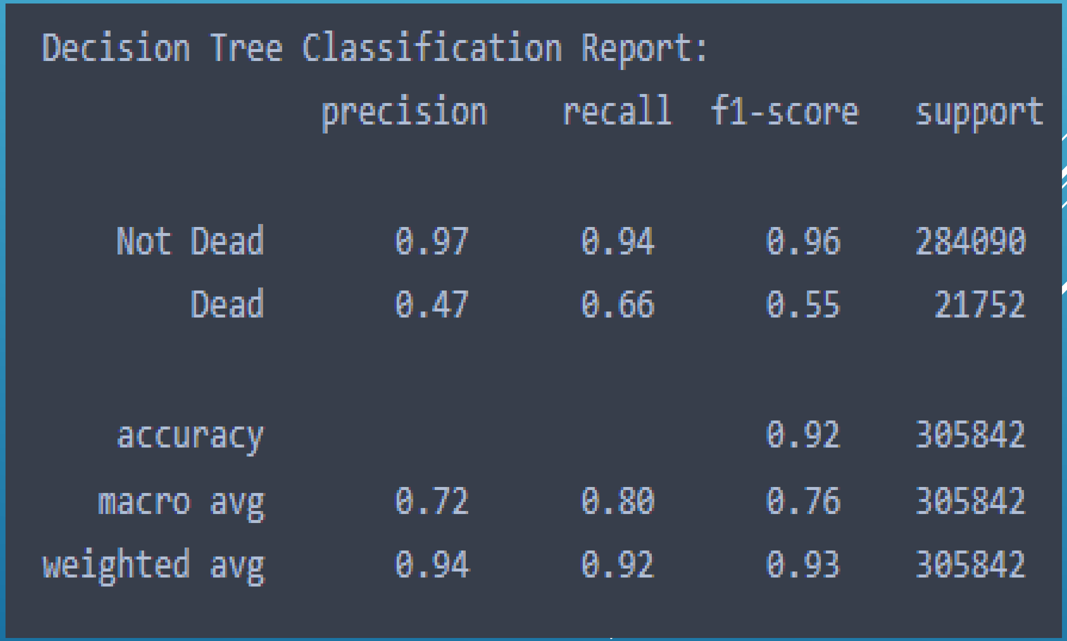
For more details, here are the classification reports for these models:

Logistic Regression



```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

    Not Dead       0.99      0.90      0.94    284090
        Dead       0.41      0.91      0.56     21752

    accuracy                           0.90    305842
   macro avg       0.70      0.91      0.75    305842
weighted avg       0.95      0.90      0.92    305842
```

Decision Tree

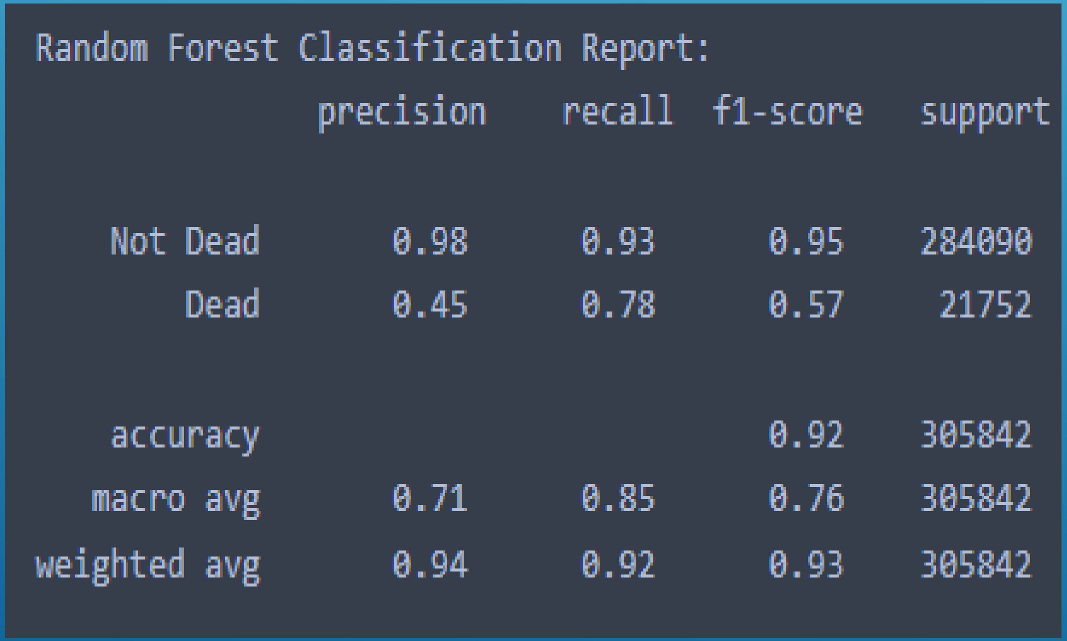


```
Decision Tree Classification Report:
              precision    recall  f1-score   support

    Not Dead       0.97      0.94      0.96    284090
        Dead       0.47      0.66      0.55     21752

    accuracy                           0.92    305842
   macro avg       0.72      0.80      0.76    305842
weighted avg       0.94      0.92      0.93    305842
```

Modeling and Results

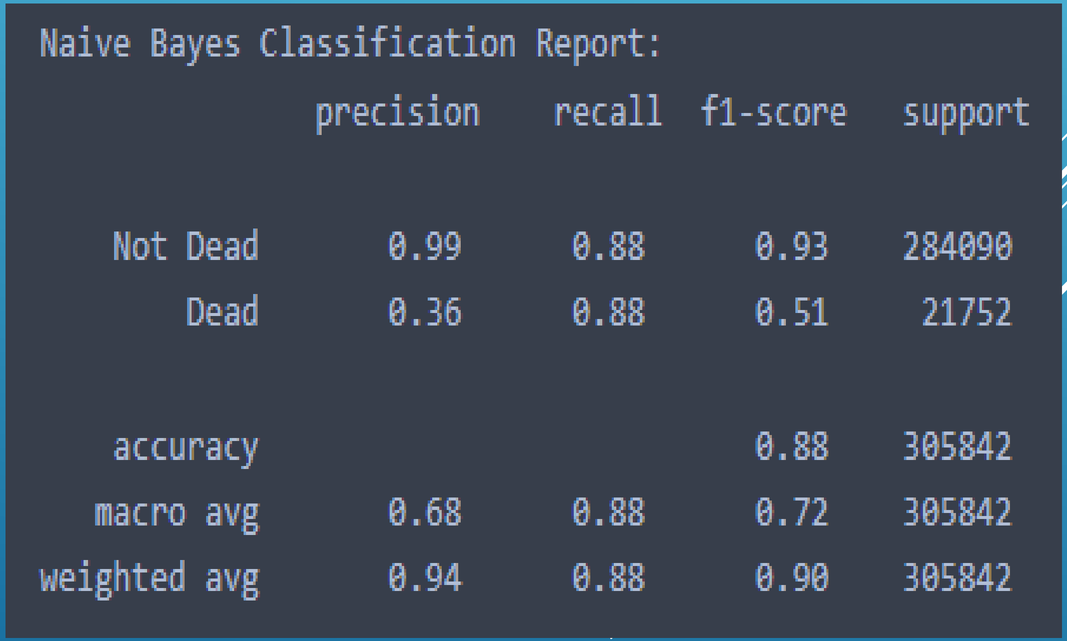
For more details, here are the classification reports for these models:

Random Forrest



```
Random Forest Classification Report:
              precision    recall  f1-score   support

    Not Dead       0.98      0.93      0.95    284090
        Dead       0.45      0.78      0.57     21752

    accuracy                           0.92    305842
   macro avg       0.71      0.85      0.76    305842
weighted avg       0.94      0.92      0.93    305842
```

Naive Bayes

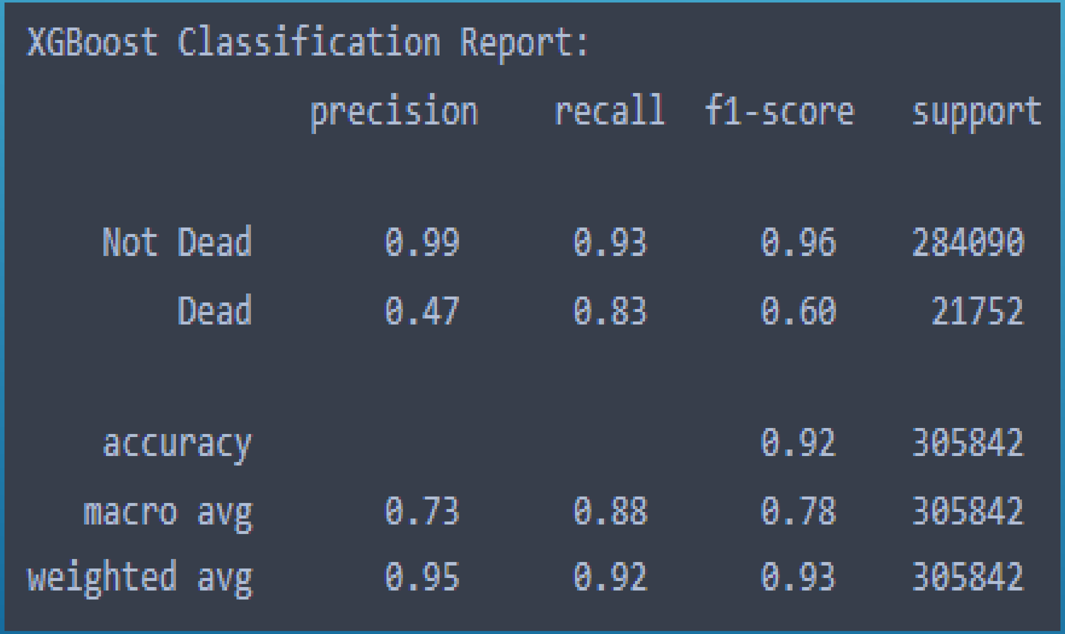


```
Naive Bayes Classification Report:
              precision    recall  f1-score   support

    Not Dead       0.99      0.88      0.93    284090
        Dead       0.36      0.88      0.51     21752

    accuracy                           0.88    305842
   macro avg       0.68      0.88      0.72    305842
weighted avg       0.94      0.88      0.90    305842
```

Modeling and Results

For more details, here are the classification reports for these models:

XGBoost



```
XGBoost Classification Report:
              precision    recall  f1-score   support

    Not Dead       0.99      0.93      0.96    284090
        Dead       0.47      0.83      0.60     21752

    accuracy                           0.92    305842
   macro avg       0.73      0.88      0.78    305842
weighted avg       0.95      0.92      0.93    305842
```

Conclusion

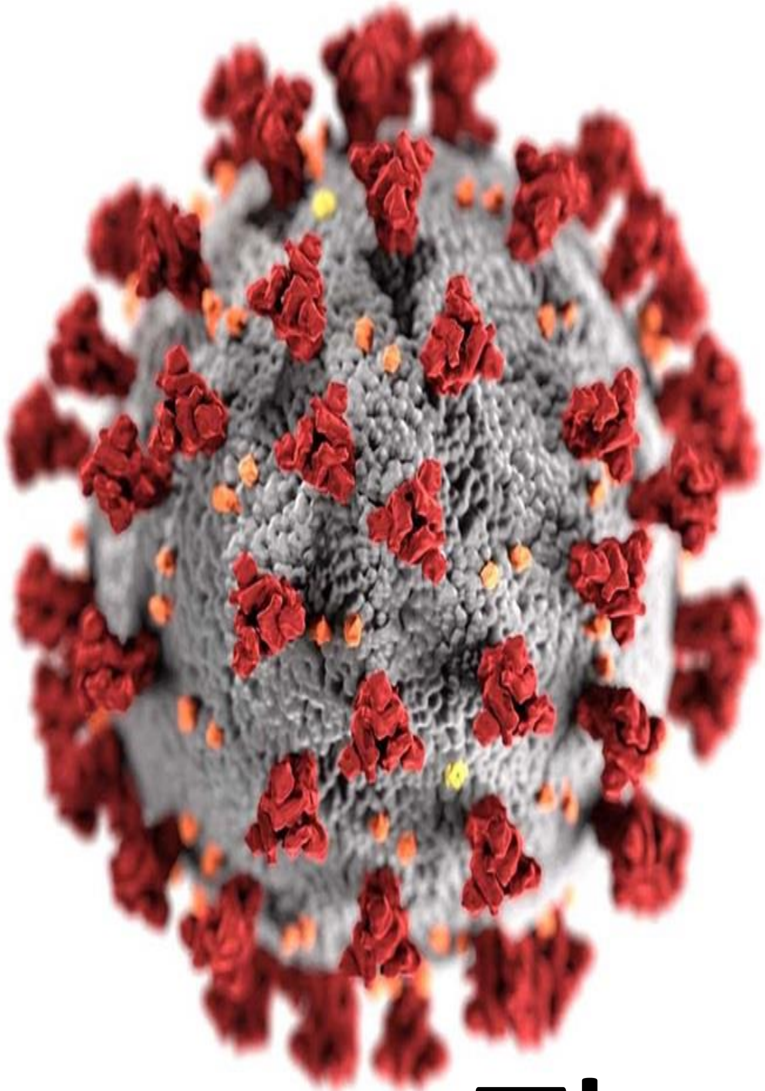
- Unfortunately 7.3% of the total patients have died, with about 70.5% of them were Covid Carriers
- As for the total carriers they were about 37.5% out of the total patients.
- About 14% of those carriers have died.
- We found that age has a significant impact; as it increases chances of getting the virus increases.
- We also found that people who are suffering from obesity are more likely to carry the virus.
- As for pregnancy, we couldn't find any impact on Covid classification.
- We noticed that patients with "Pneumonia", "Hypertension", "Diabetes" and tobacco users have a great chance of getting the virus with "Pneumonia" patients being the most.

Conclusion

- We also noticed that there's a positive correlation between having "Hypertension" and "Diabetes" diseases; as most patients with one of those two diseases are subjected to get the other.
- We saw that among all the patients of these diseases the patients classified with 3rd degree of Covid are the highest by far
- About 19% of the total patients were hospitalized, with a death percentage of 35%.
- The "Pneumonia" disease has the greatest impact on that percentage (35%).
- About 91% of the dead patients were hospitalized
- About 9% out of the hospitalized patients were admitted to the ICU, with about 56% of them being classified as Covid carriers, and with a great percentage of death of about 49%.

Conclusion

- The death was very trending during mid 2020 starting from April up till August.
- As for our modeling, we got the highest accuracy of **92**% using both "**Random Forest**" and "**Decision Tree**" algorithms, but when checking the recall we found that those accuracies are misleading. So, we settled with the "**Logistic Regression**" that achieved great scores regarding both accuracy and recall with about **90**%.



Thank You!