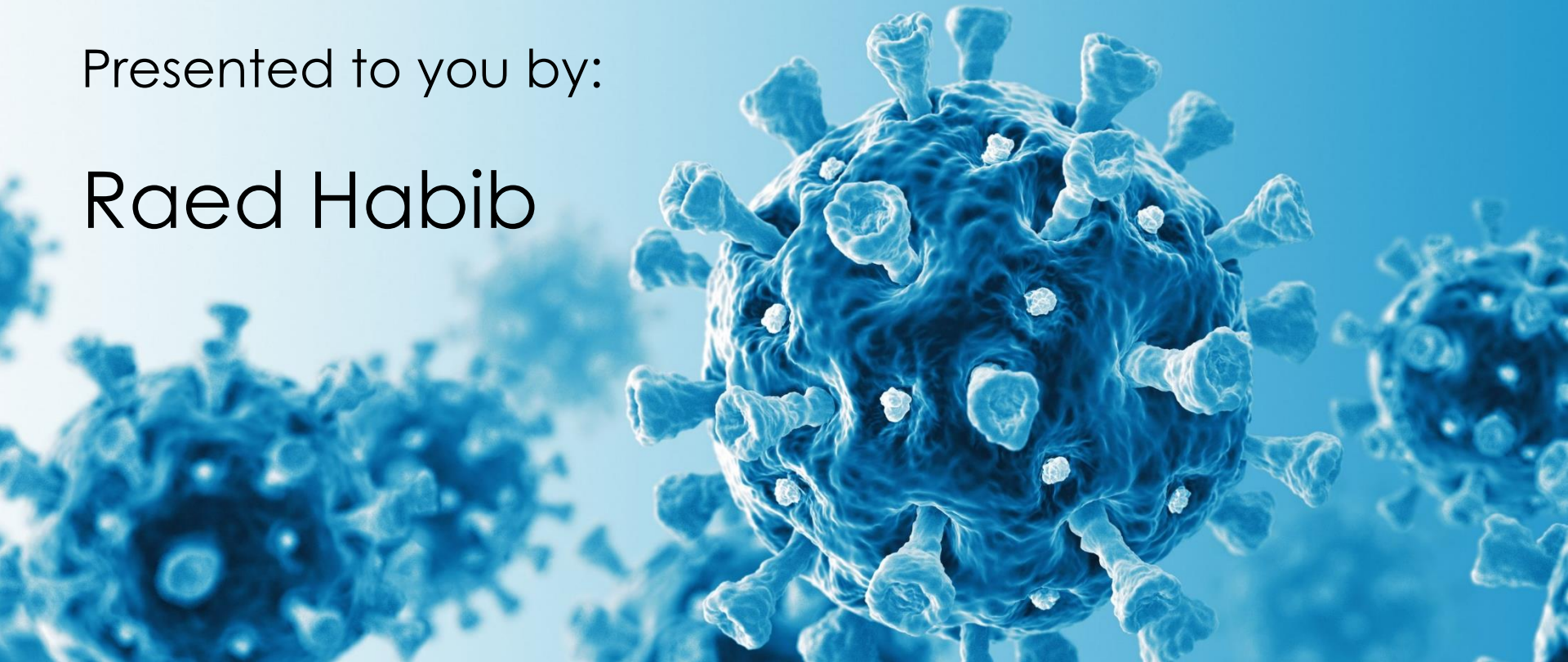


# Covid 19 Analysis and Risk Prediction

Presented to you by:

Raed Habib



# 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/meirnazri/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\*\***.

Several white lines of varying lengths and angles are drawn in the bottom right corner of the slide, creating a modern, abstract graphic element.

# 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.

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.

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.

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:**

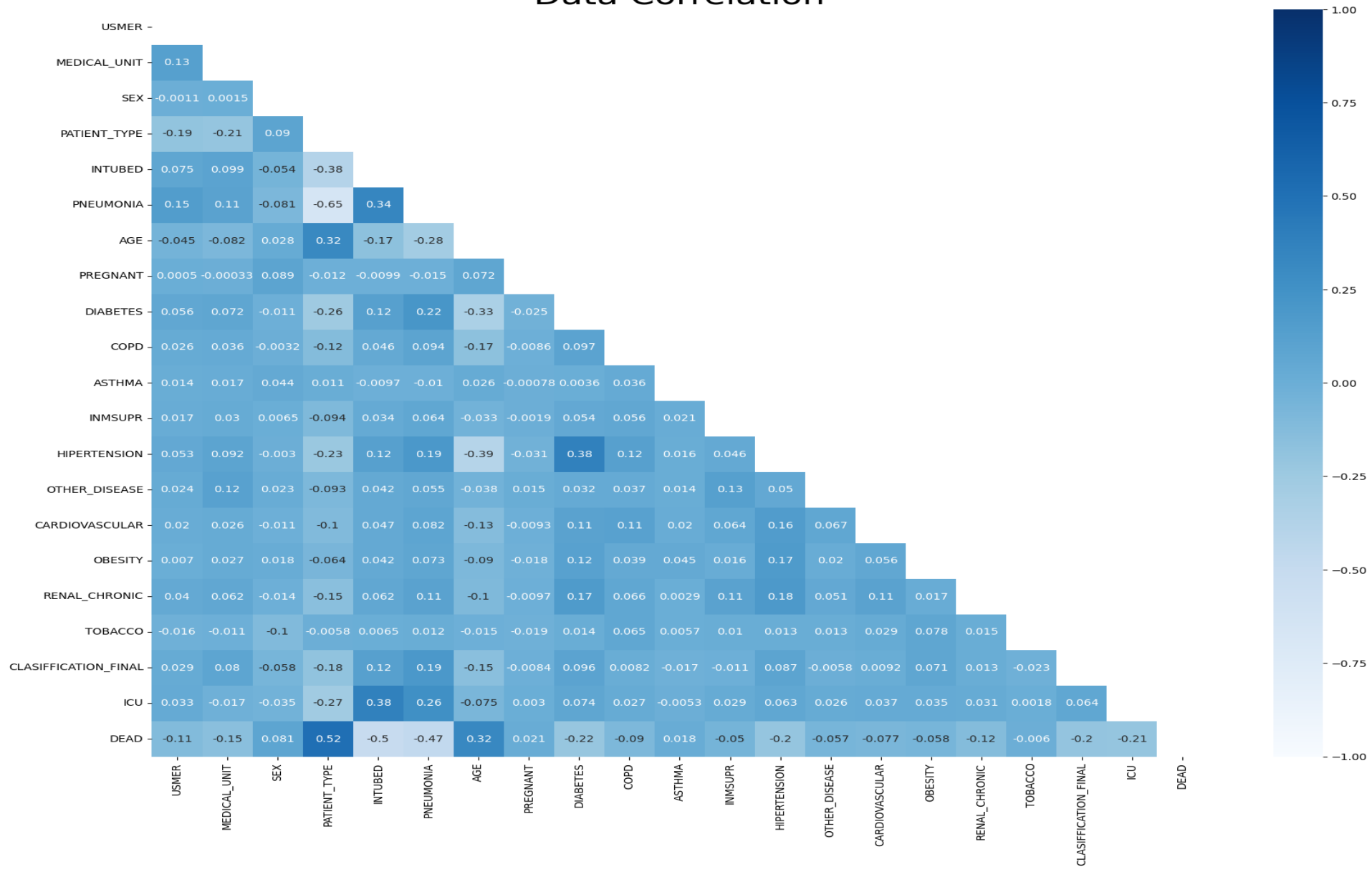
# Initial Questions to be Answered:

- 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

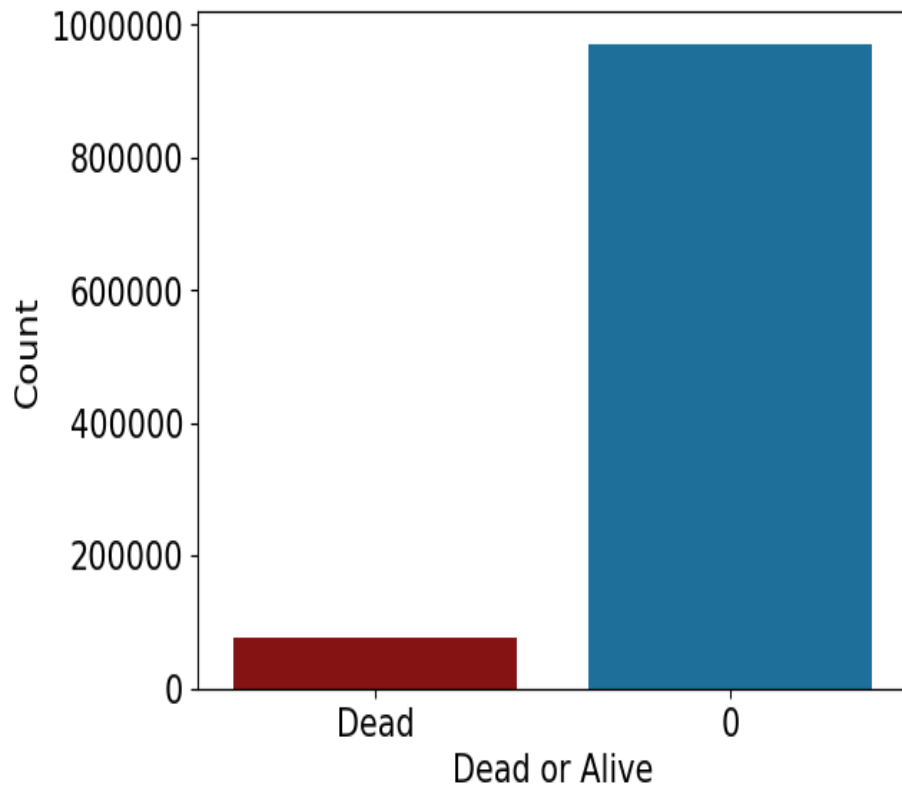
## Data Correlation



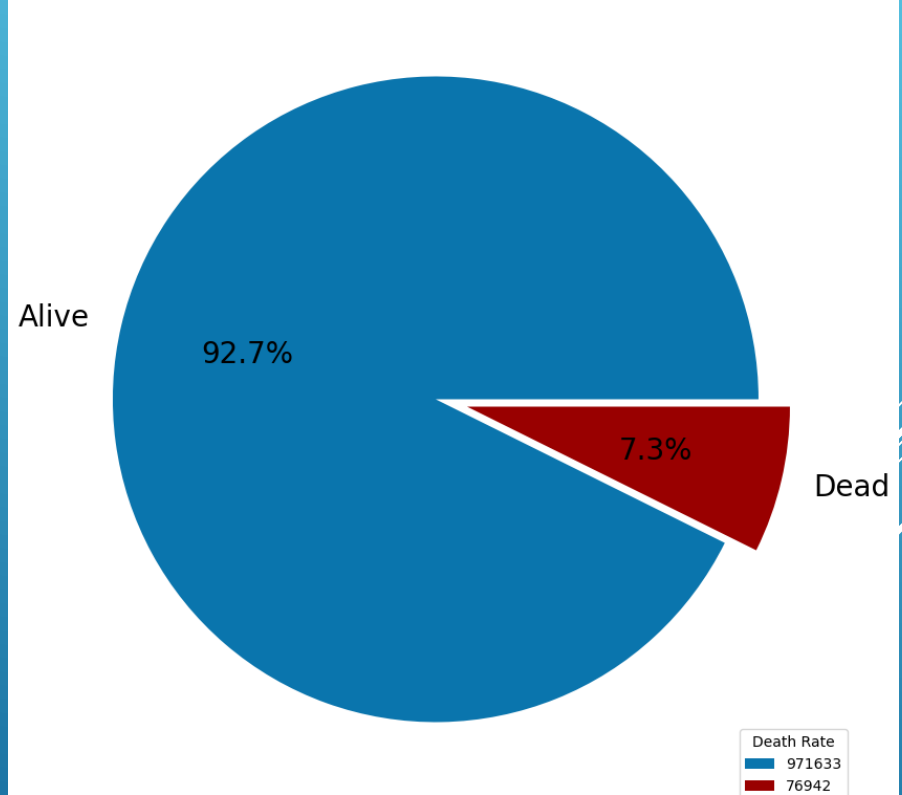
# Exploratory Data Analysis

## Death statistics

Deaths Count



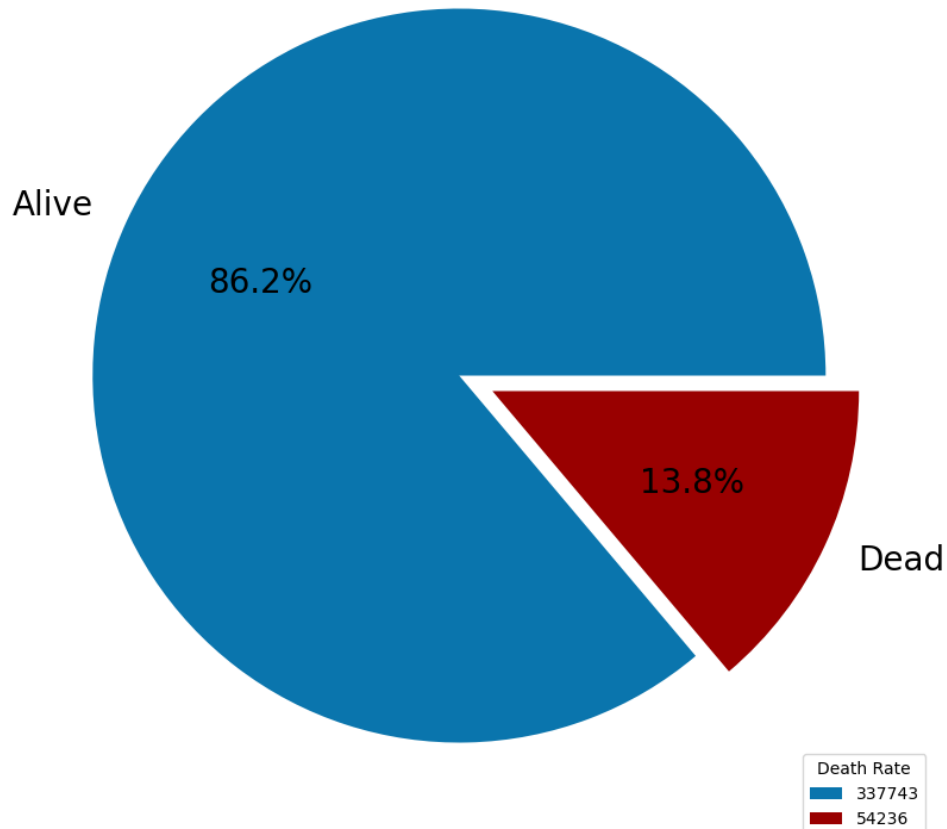
Death Percentage



# Exploratory Data Analysis

## Death statistics

### Death Percentage among Covid Carriers

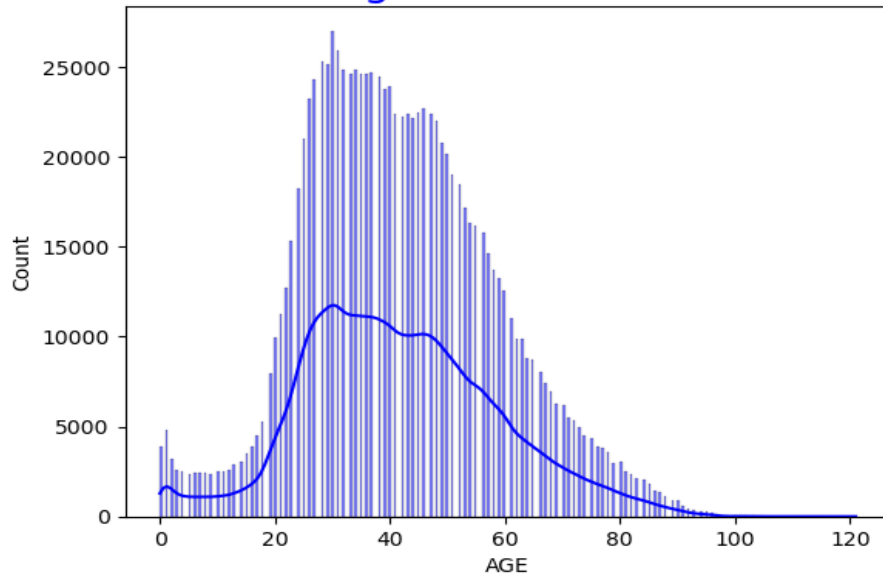




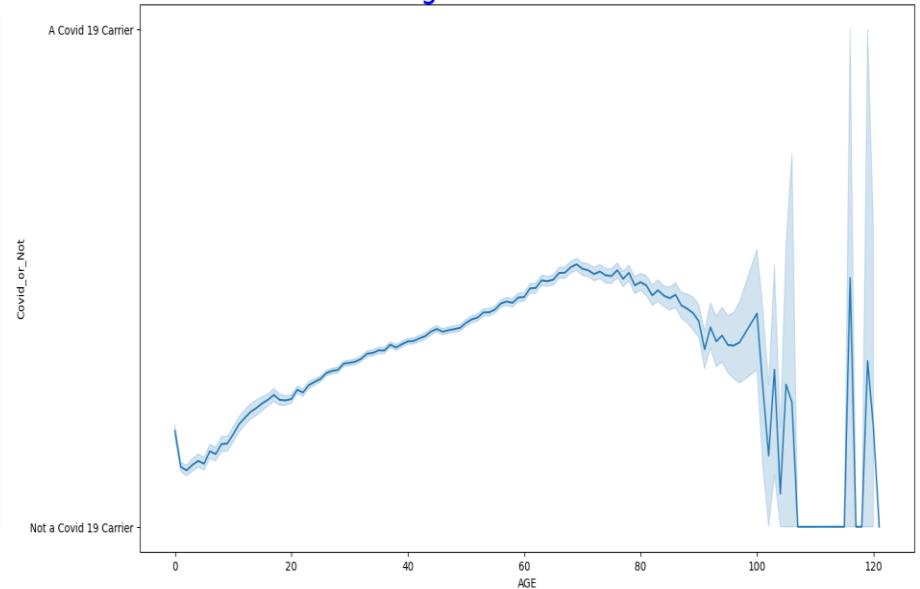
# Exploratory Data Analysis

## Death statistics

Age Distribution

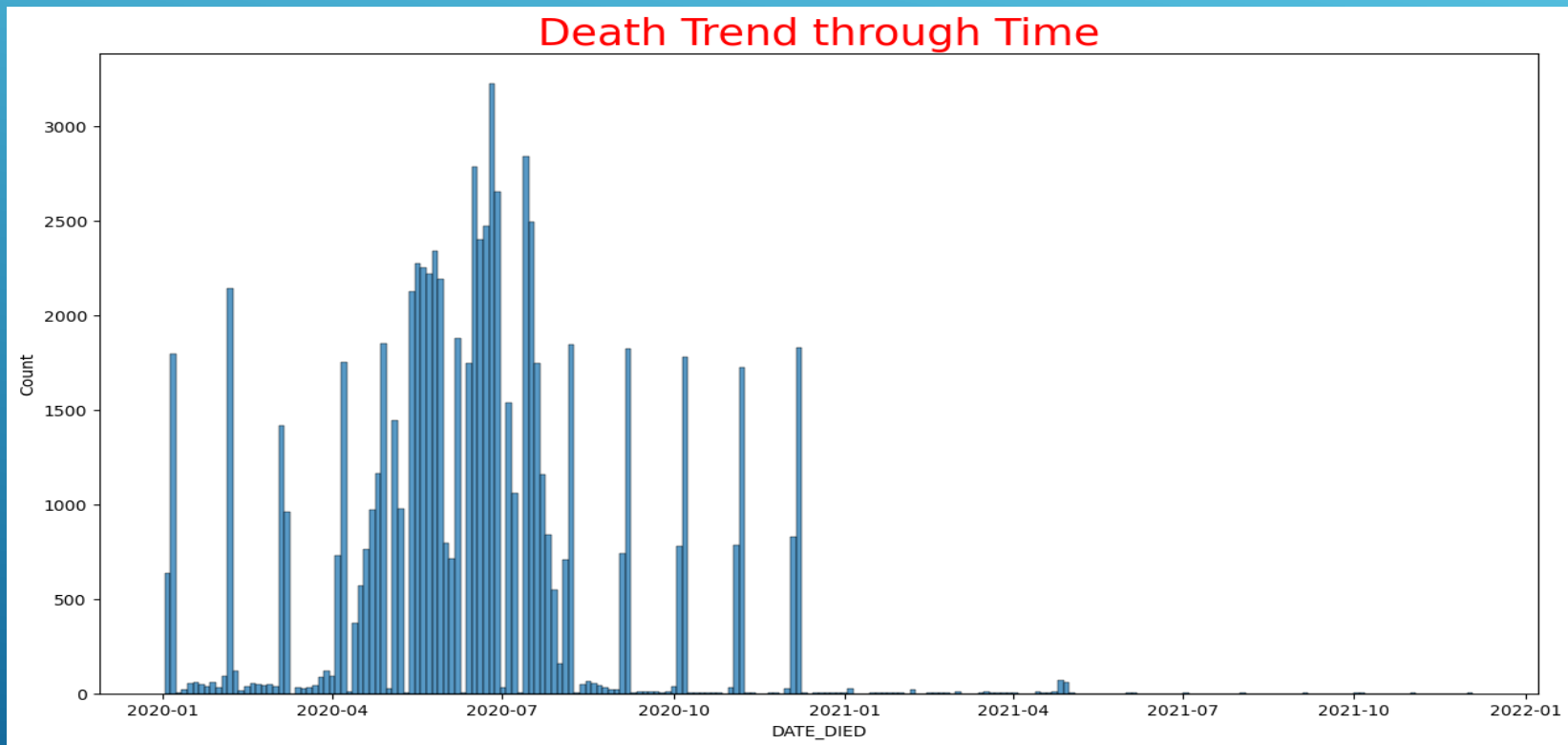


Effect of Age on Covid Classification



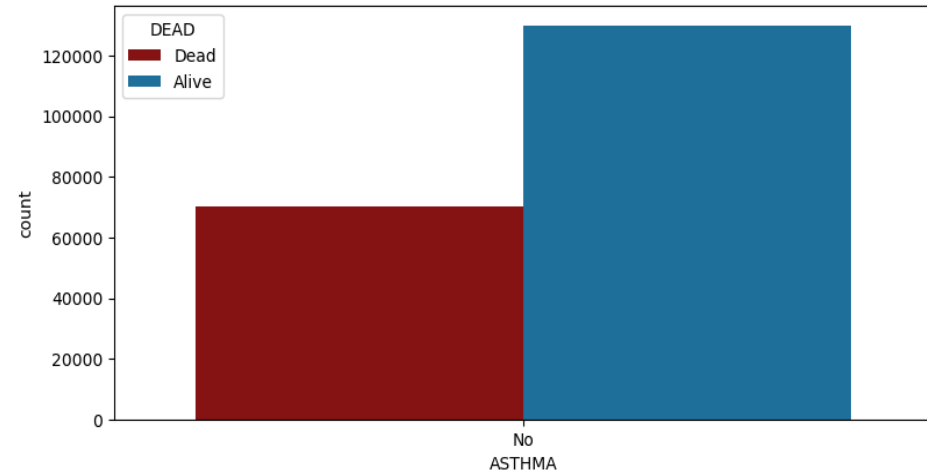
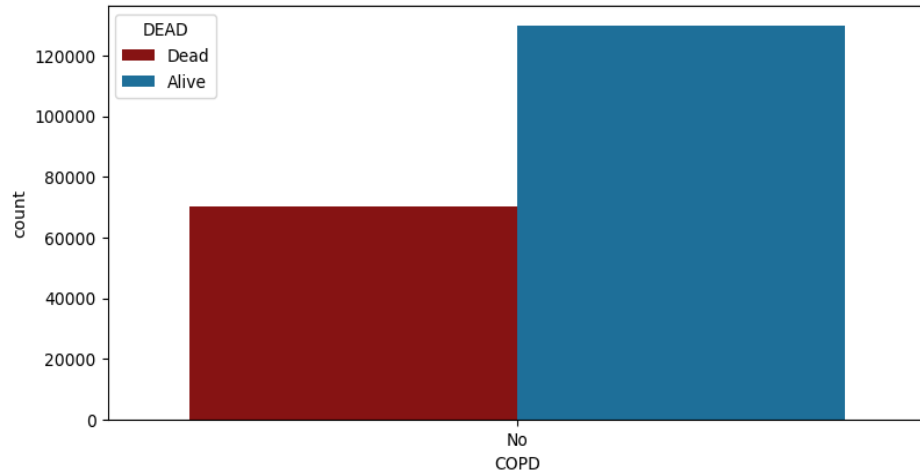
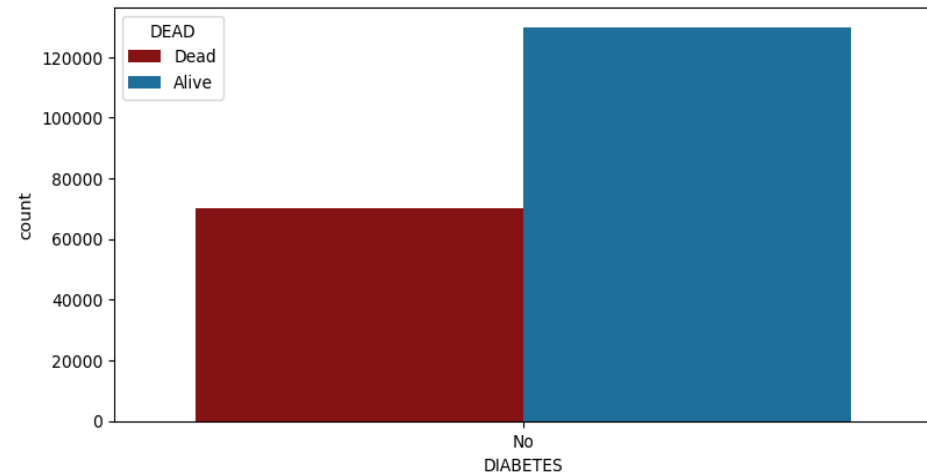
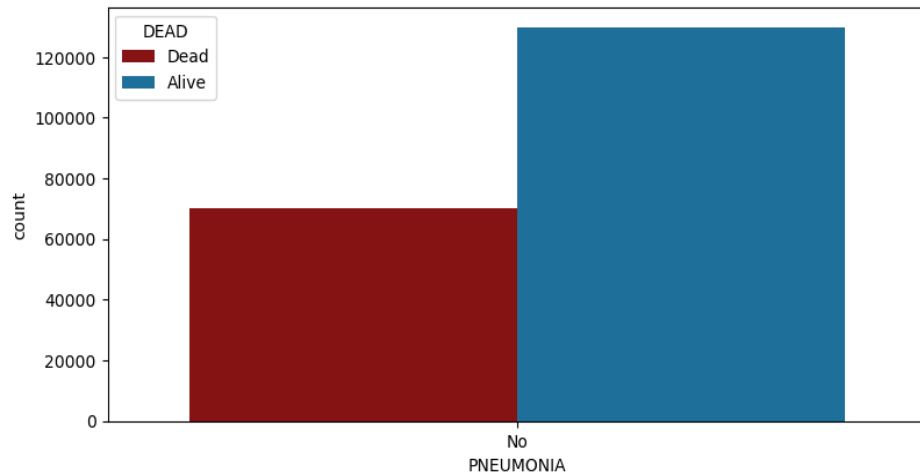
# Exploratory Data Analysis

## Death statistics



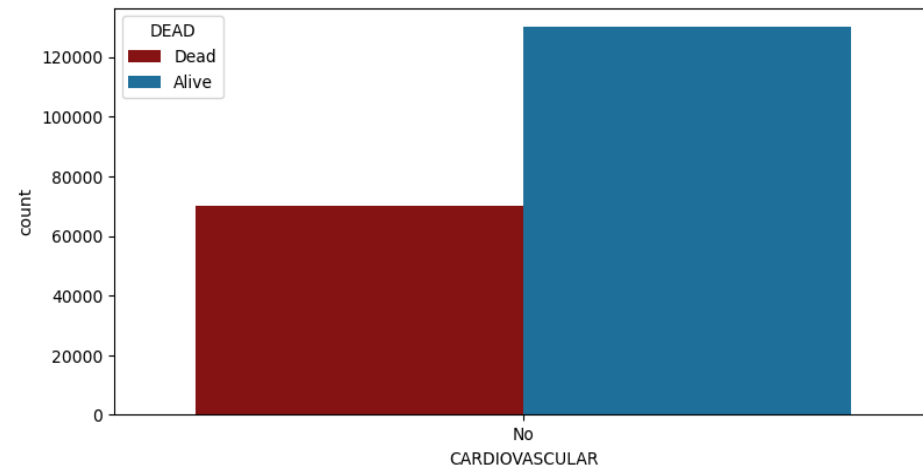
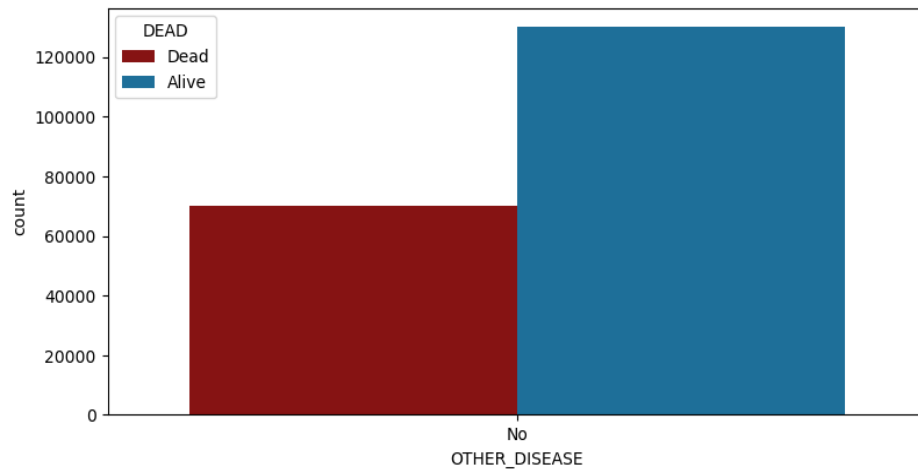
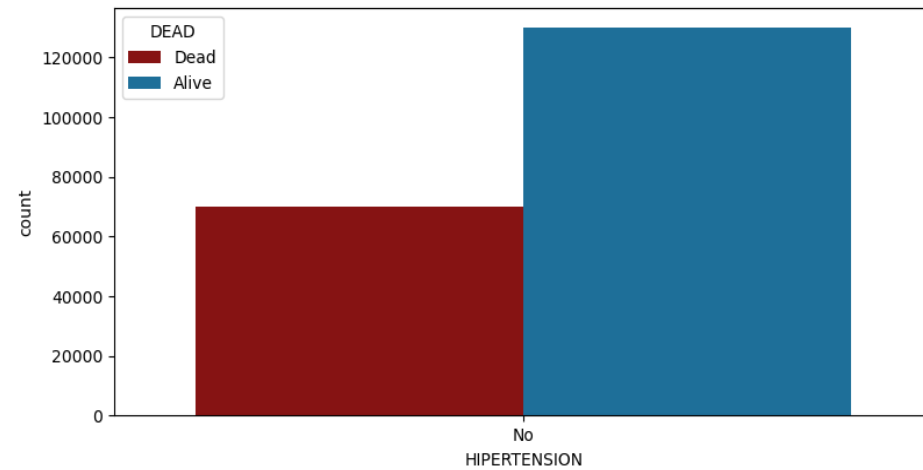
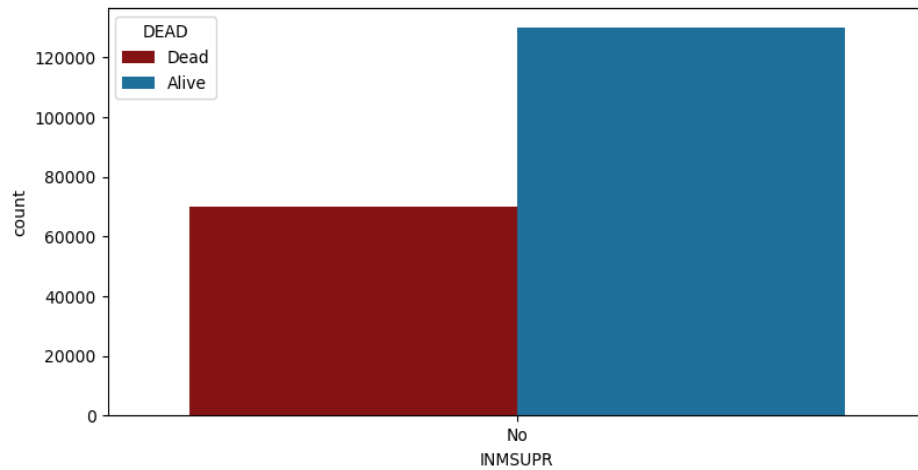
# Exploratory Data Analysis

## Effect of diseases on Death



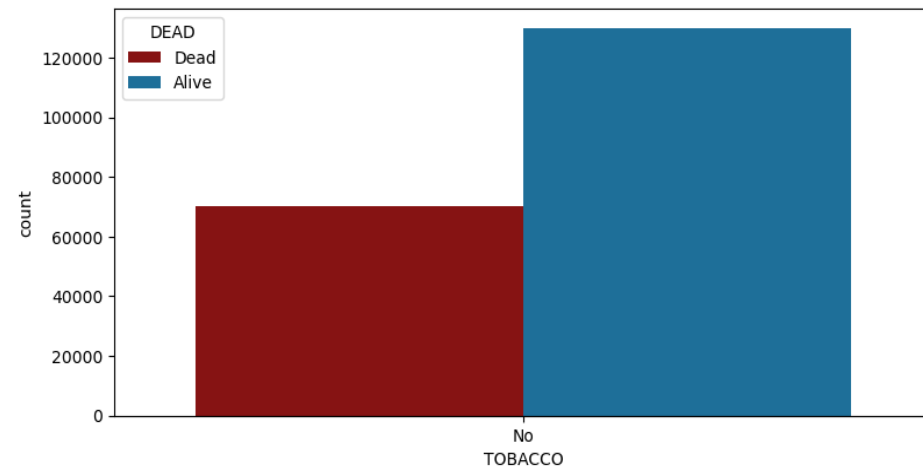
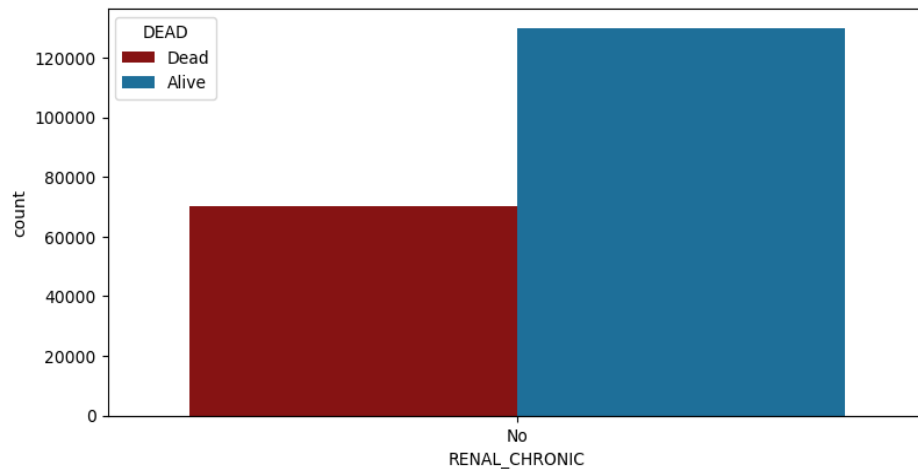
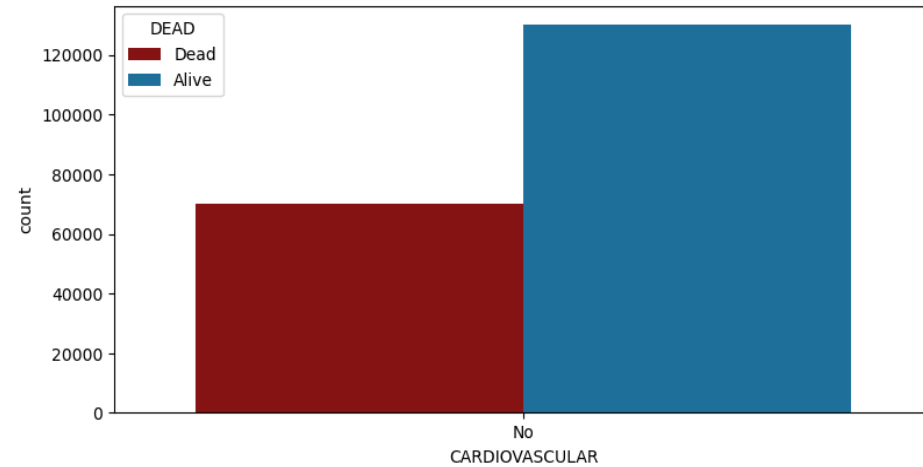
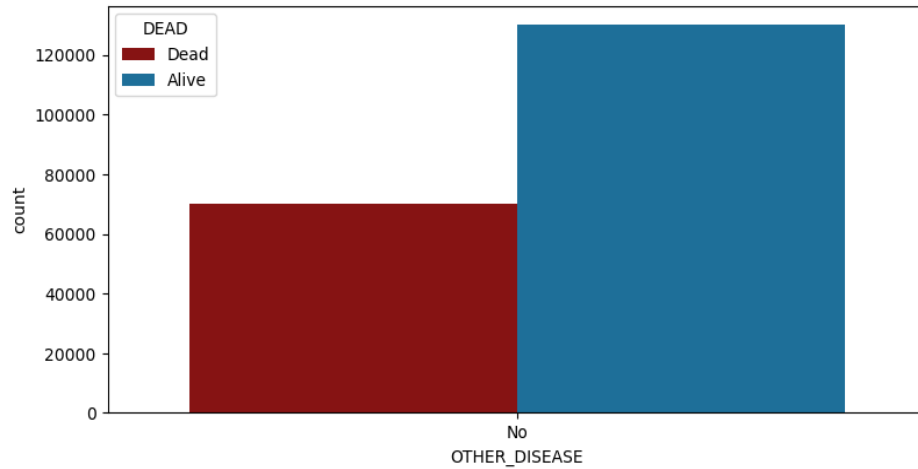
# Exploratory Data Analysis

## Effect of diseases on Death



# Exploratory Data Analysis

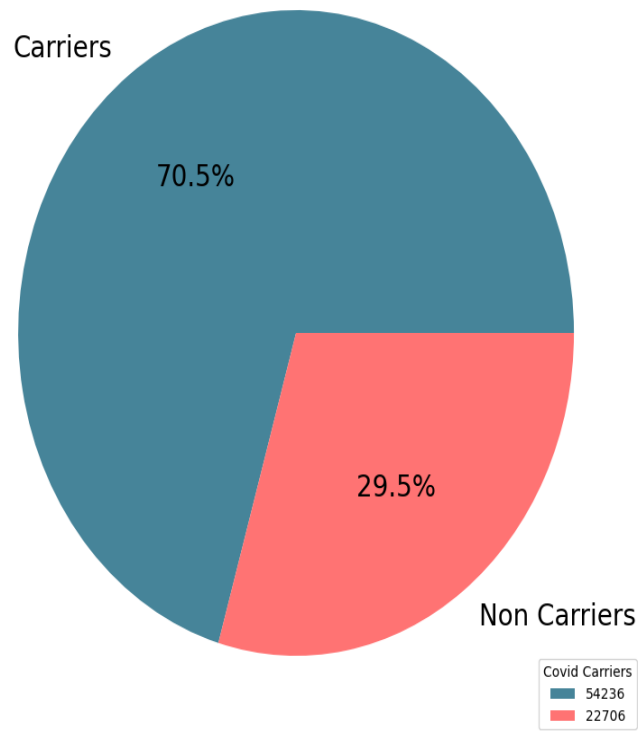
## Effect of diseases on Death



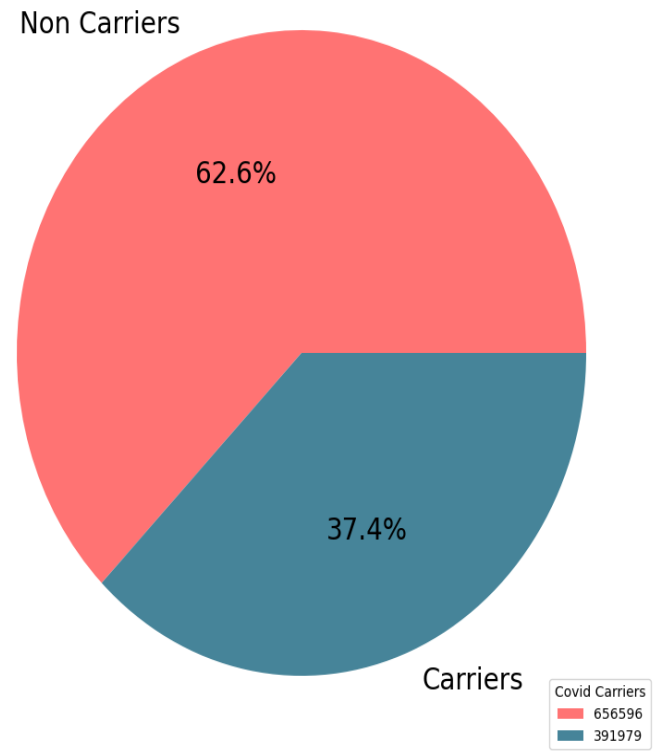
# Exploratory Data Analysis

## Covid statistics

Covid Carriers Percentage among Dead Patients



Covid Carriers Percentage among All Patients

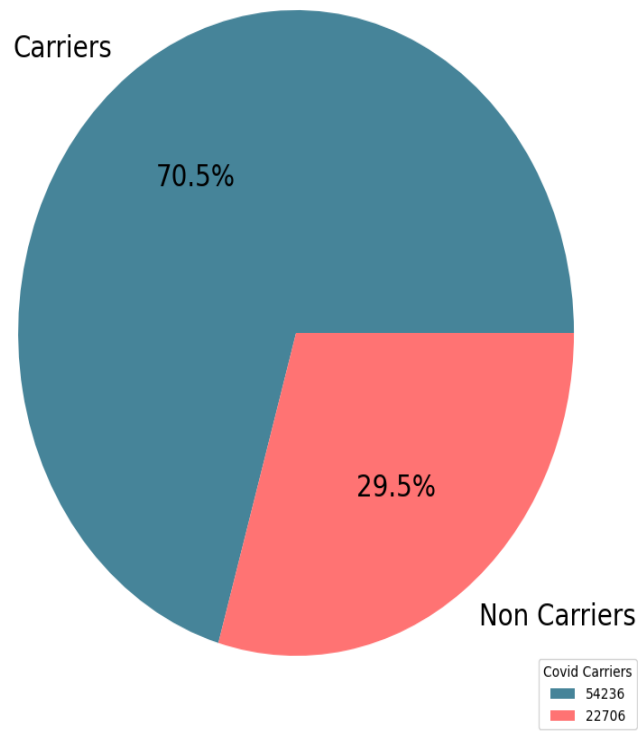




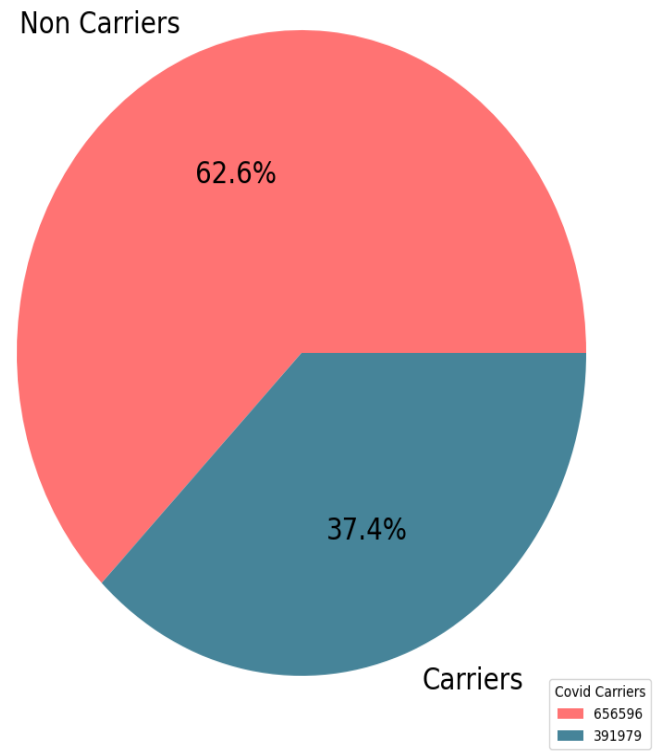
# Exploratory Data Analysis

## Covid statistics

Covid Carriers Percentage among Dead Patients



Covid Carriers Percentage among All Patients

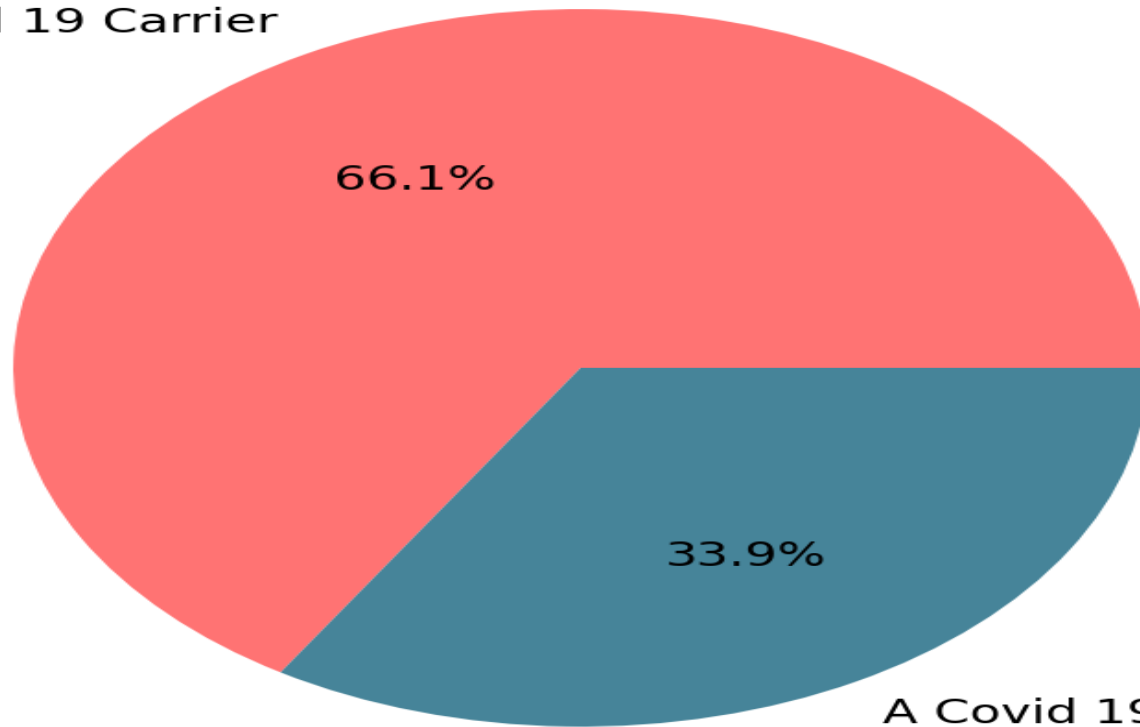


# Exploratory Data Analysis

## Covid statistics

### Covid Carriers among Pregnant Females

Not a Covid 19 Carrier



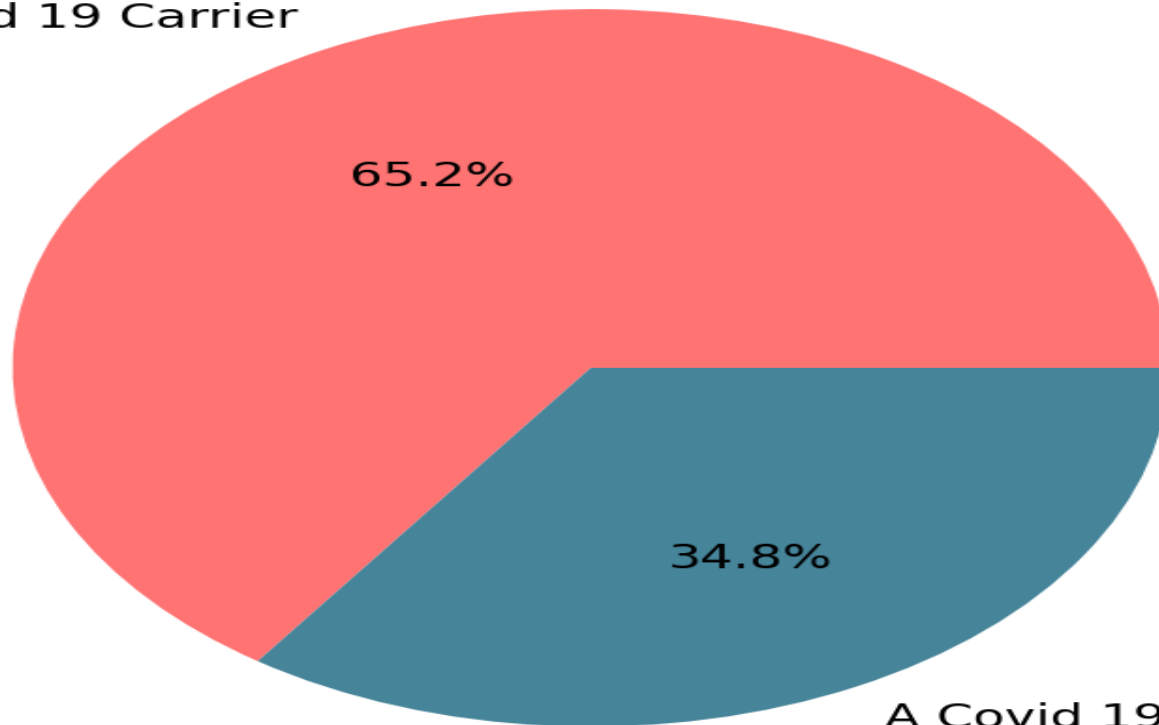
Covid Carriers	
	5377
	2754

# Exploratory Data Analysis

## Covid statistics

### Covid Carriers among Non Pregnant Females

Not a Covid 19 Carrier

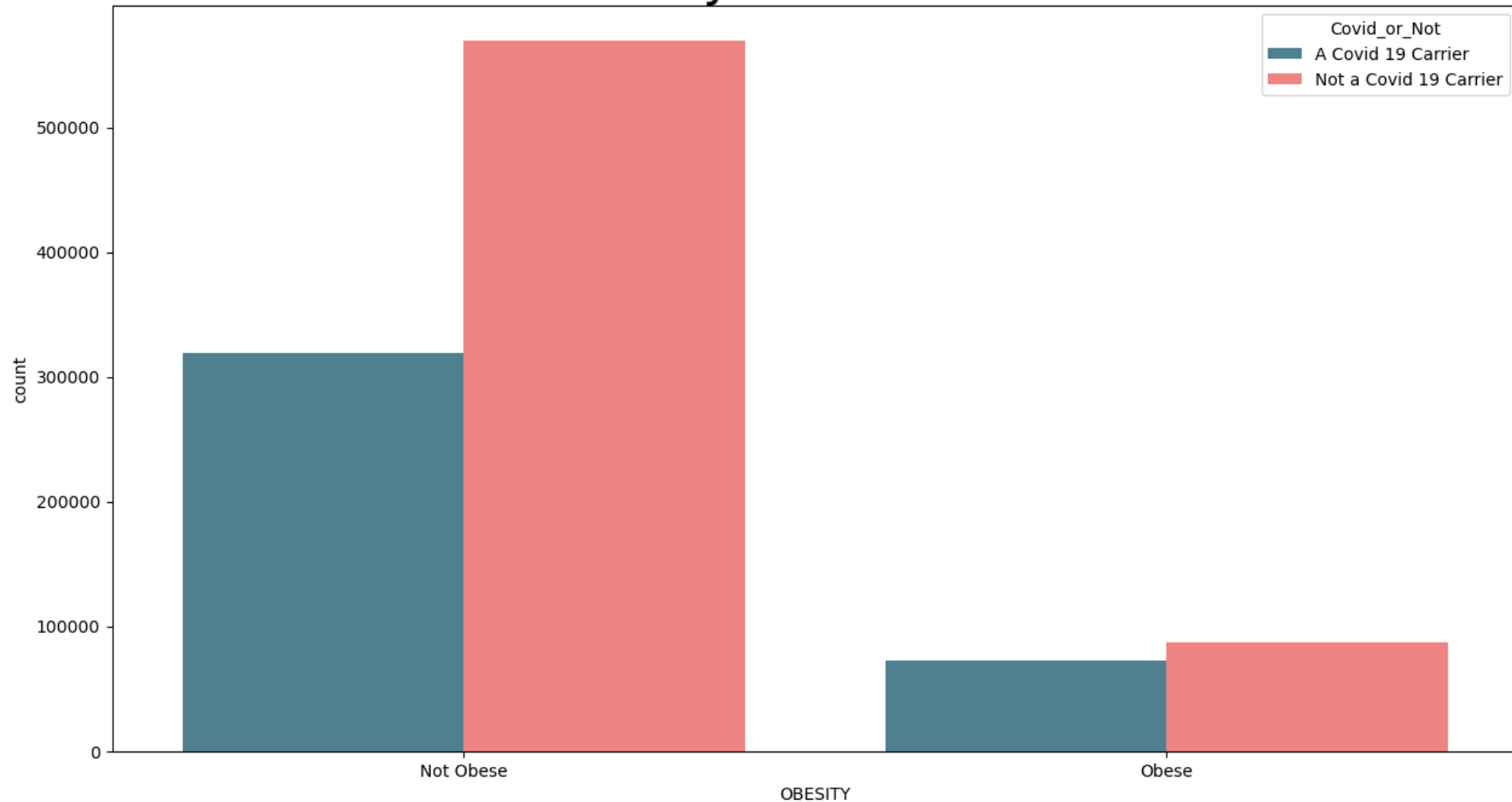


Covid Carriers	
	337197
	179736

A Covid 19 Carrier

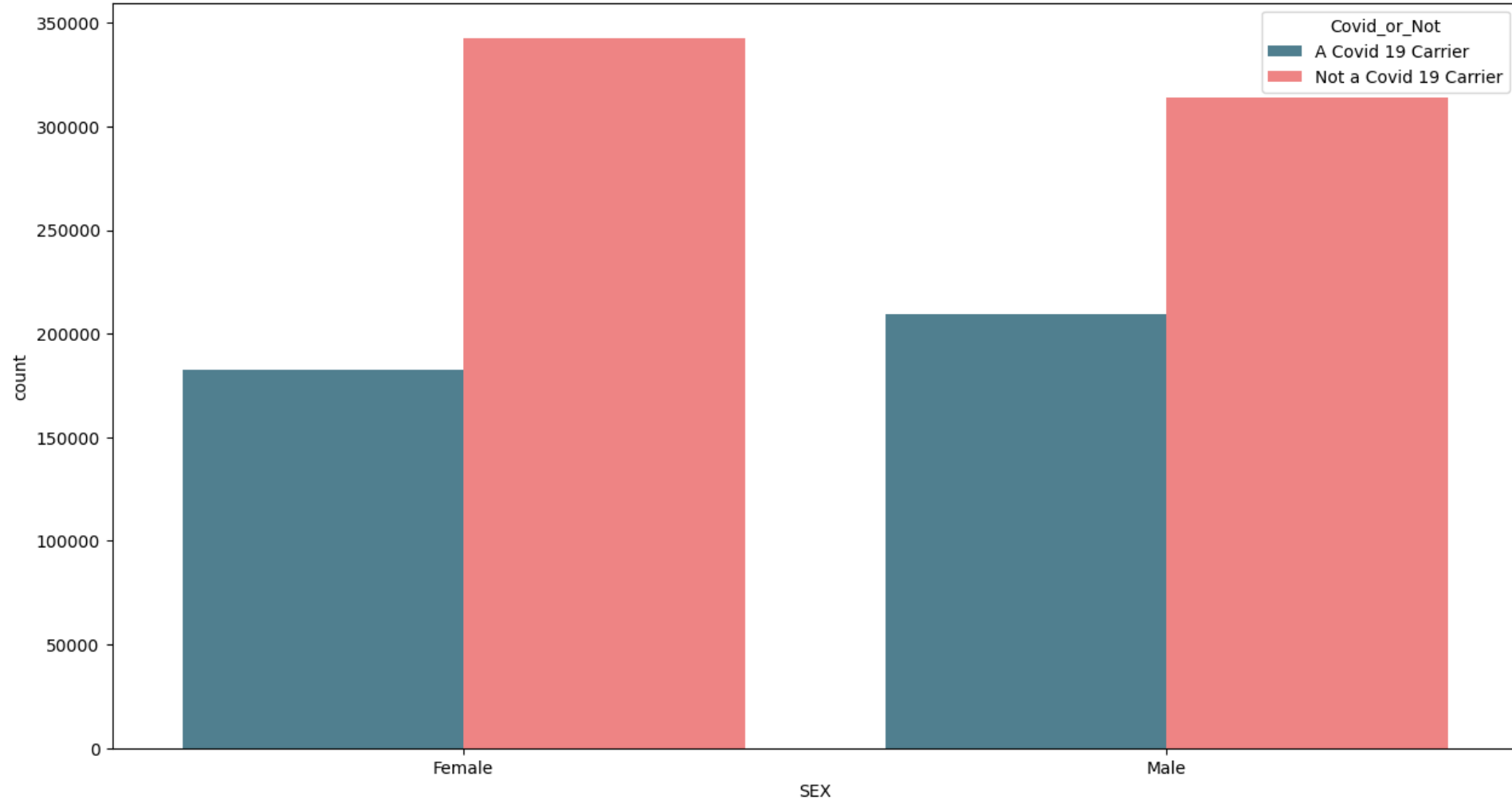
# Exploratory Data Analysis

## Effect of Obesity on Covid Classification



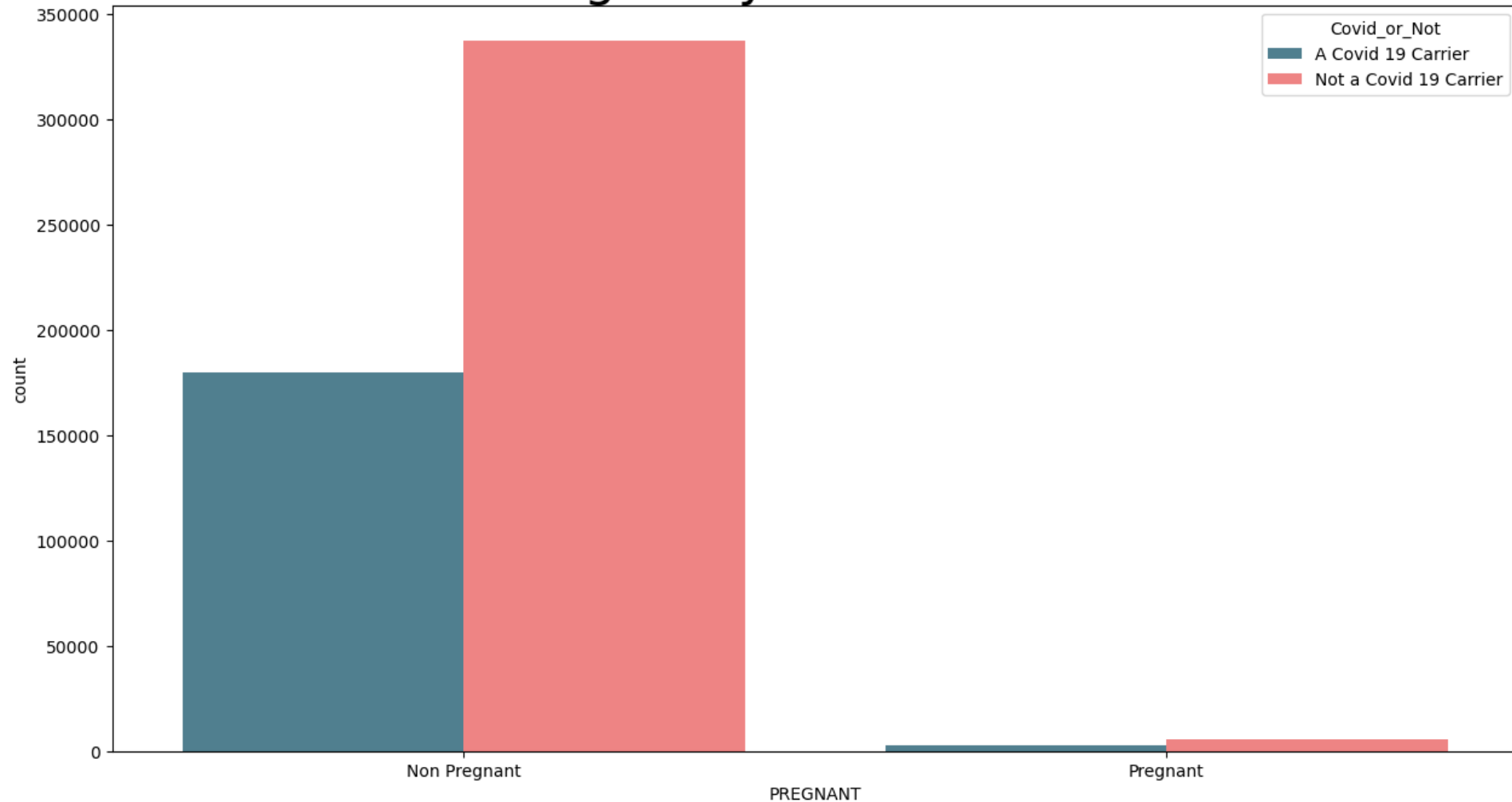
# Exploratory Data Analysis

## Effect of Gender on Covid Classification



# Exploratory Data Analysis

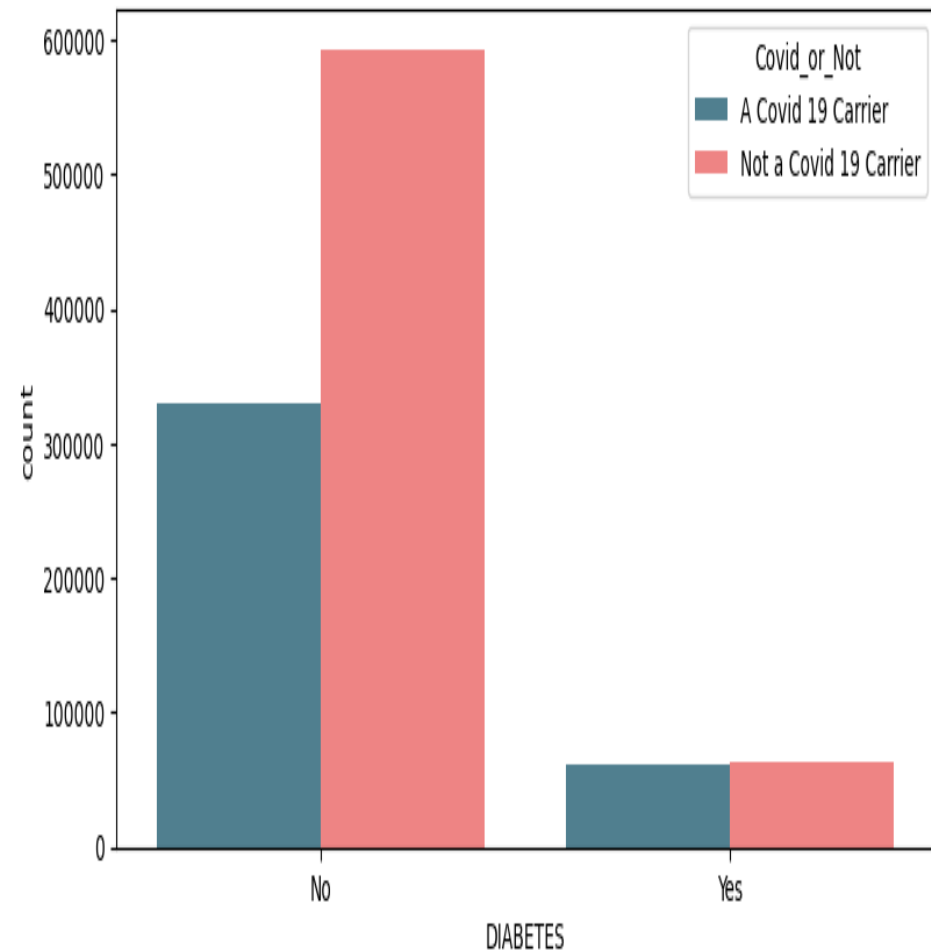
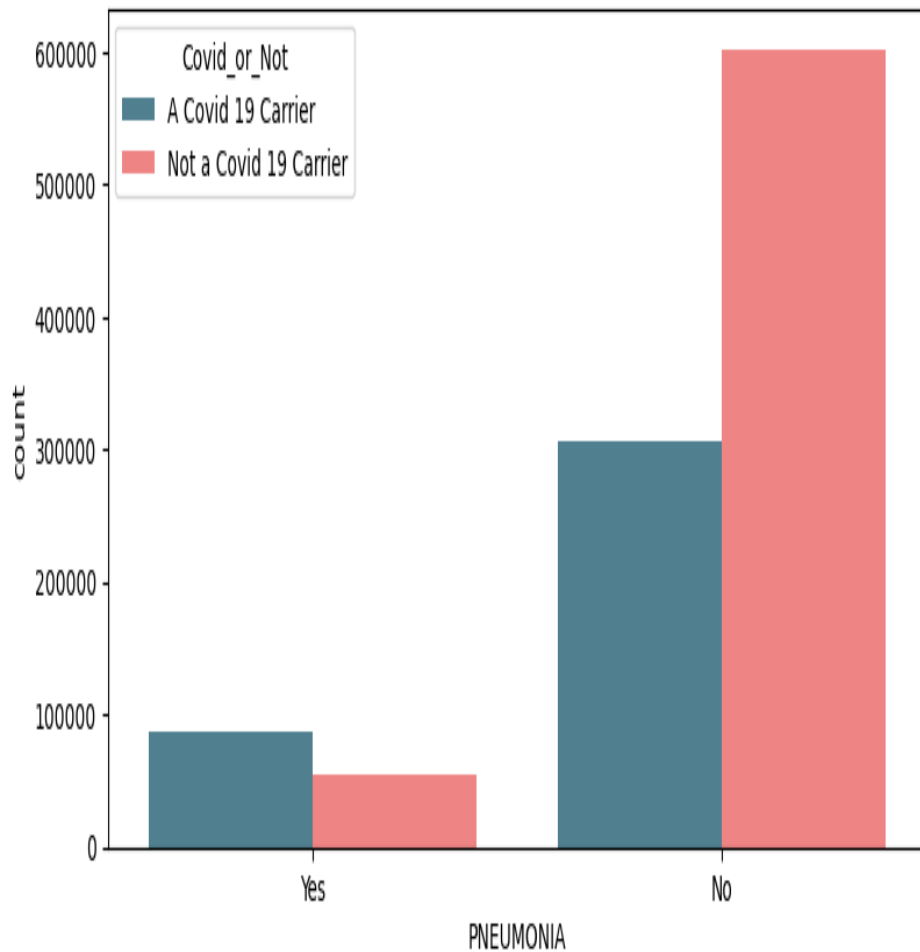
## Effect of Pregnancy on Covid Classification





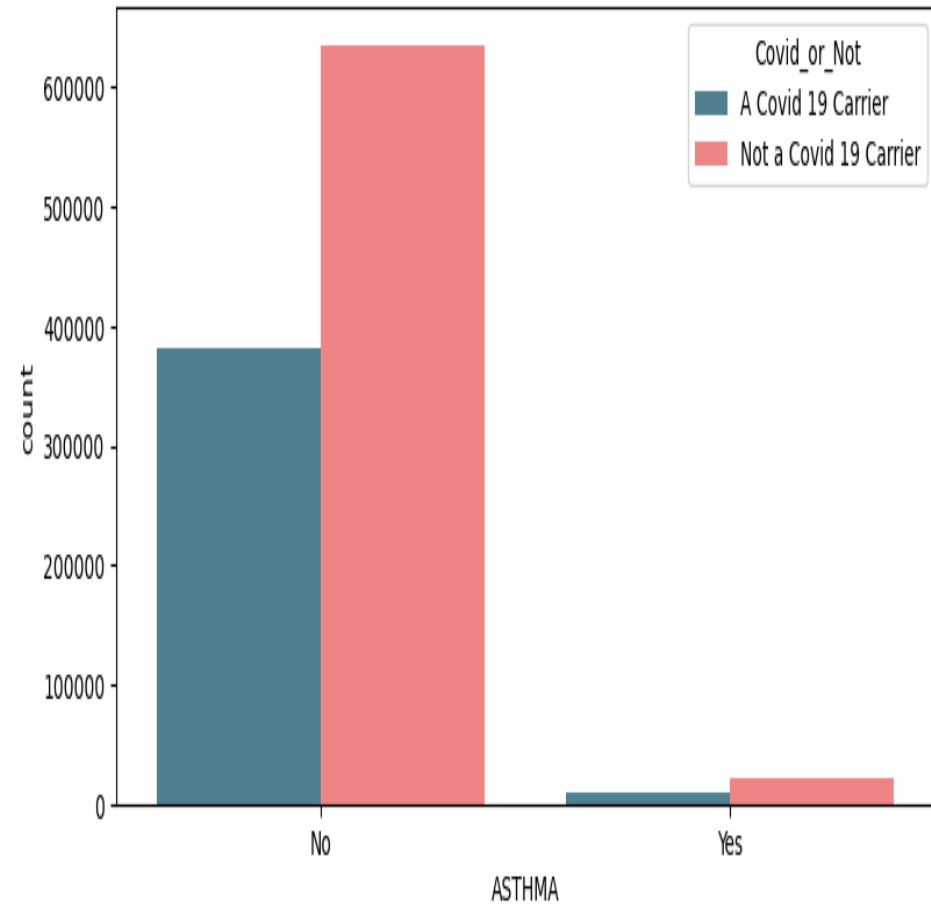
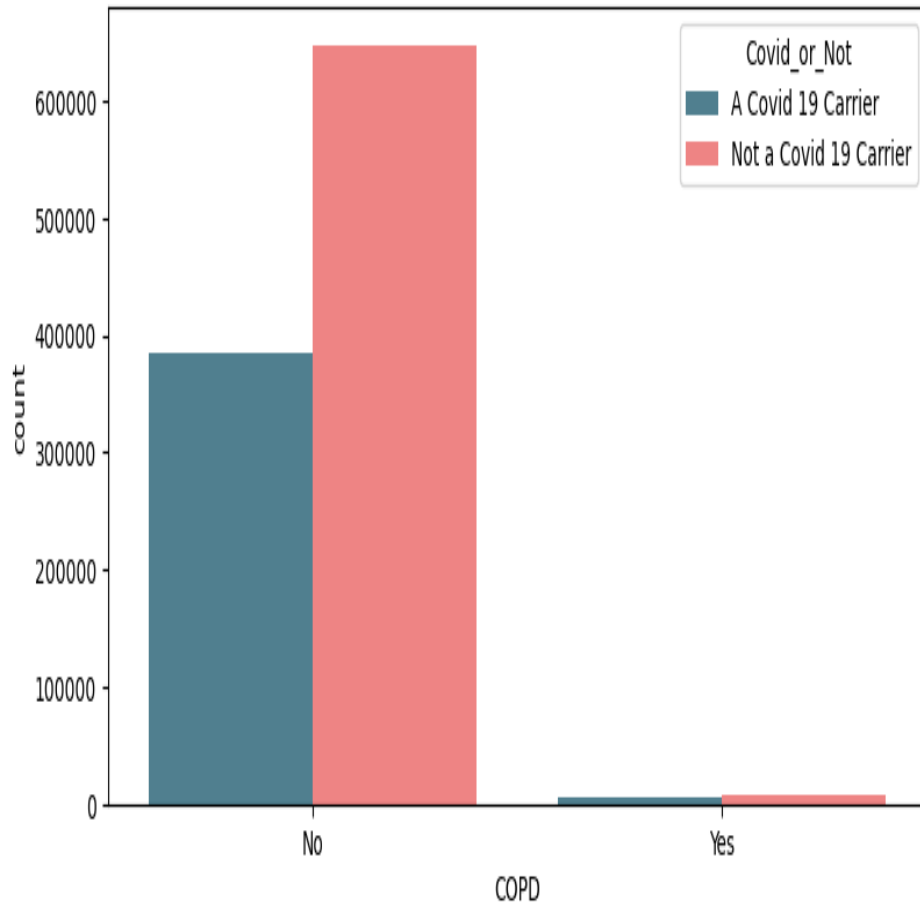
# Exploratory Data Analysis

## Effect of diseases on Covid



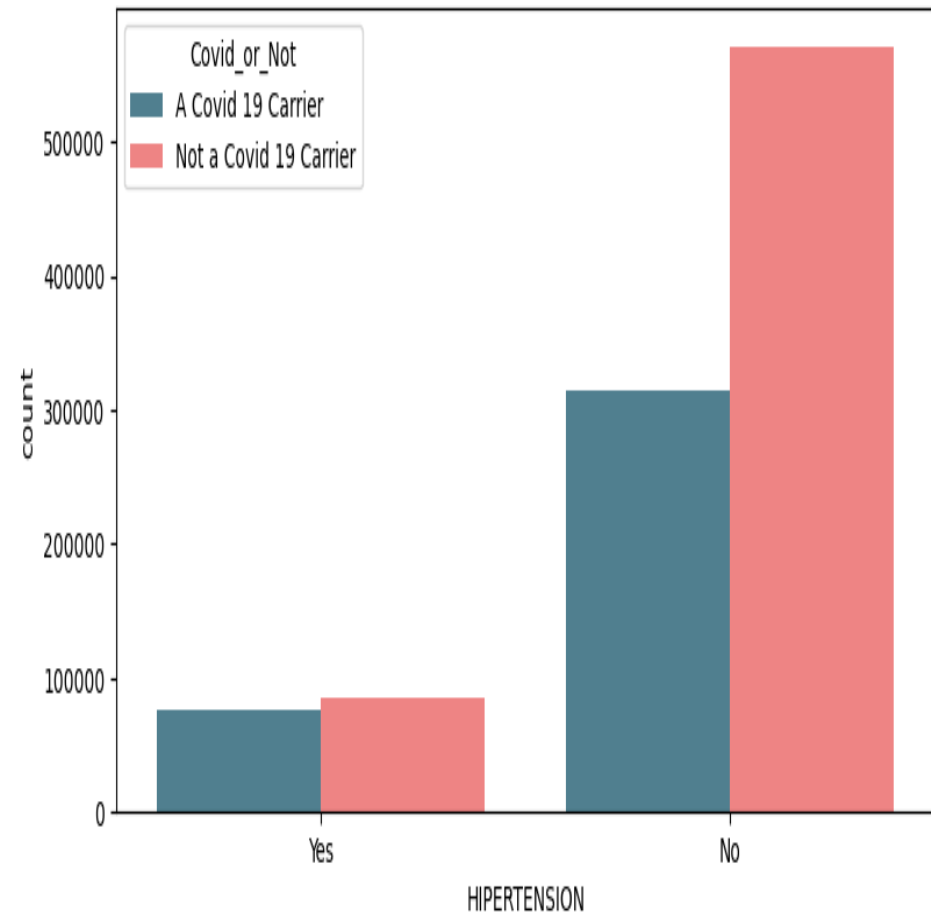
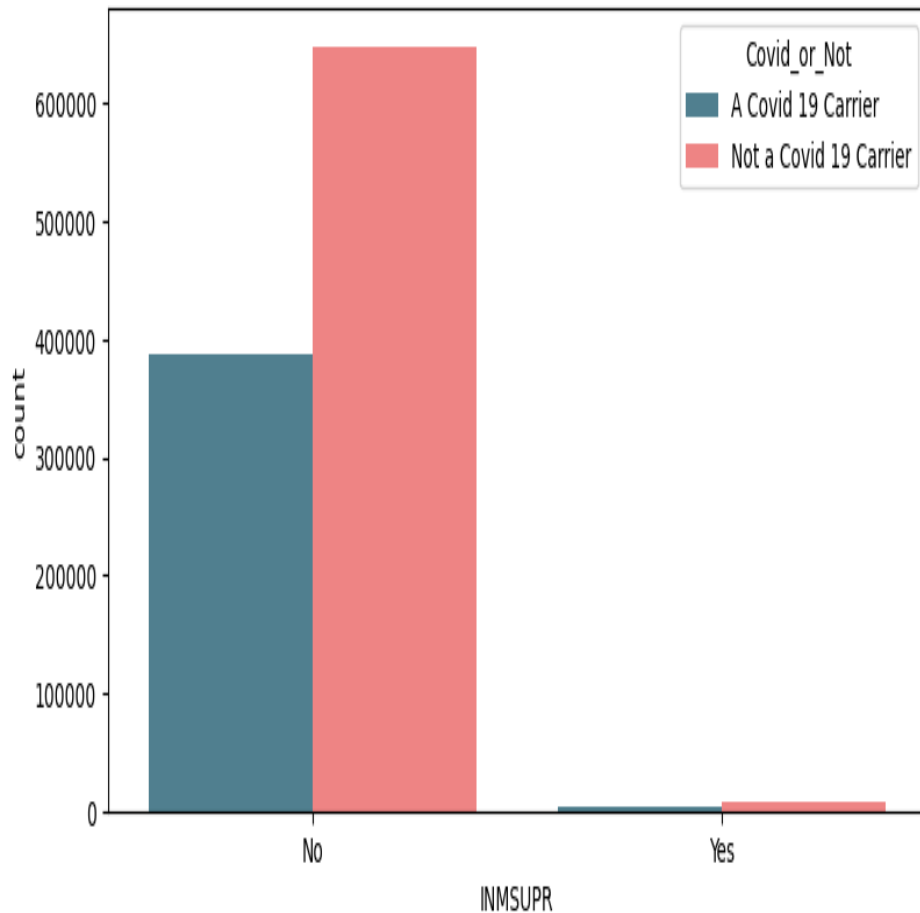
# Exploratory Data Analysis

## Effect of diseases on Covid



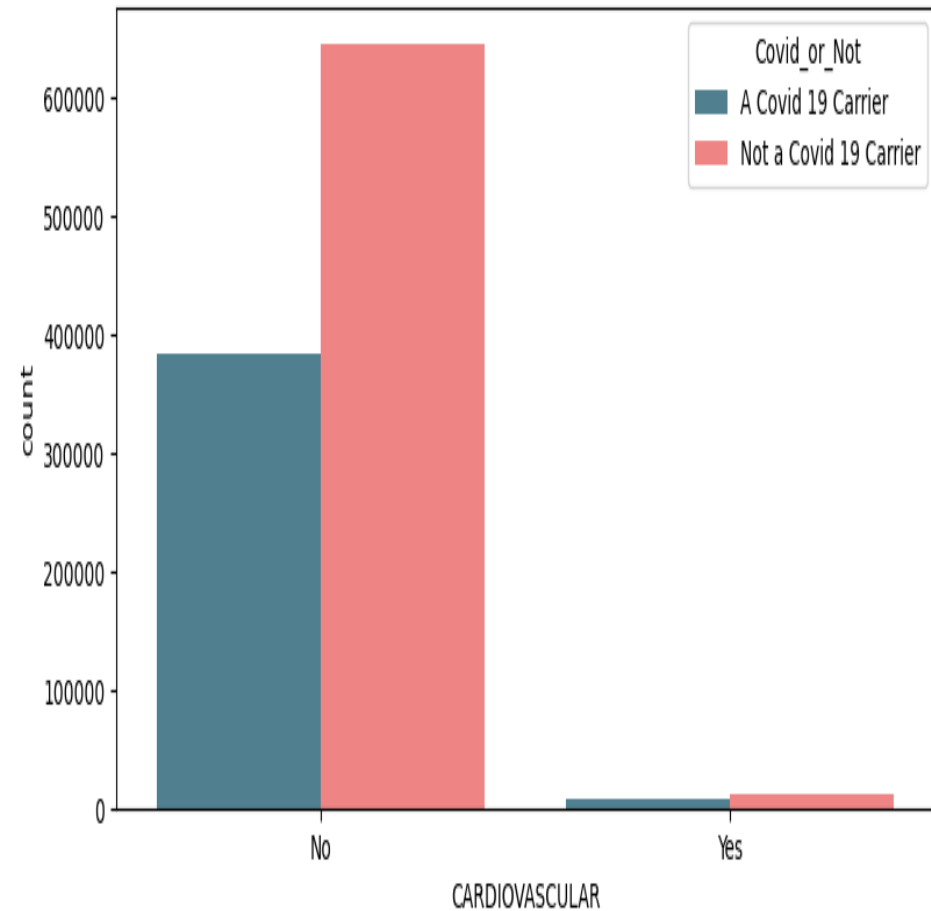
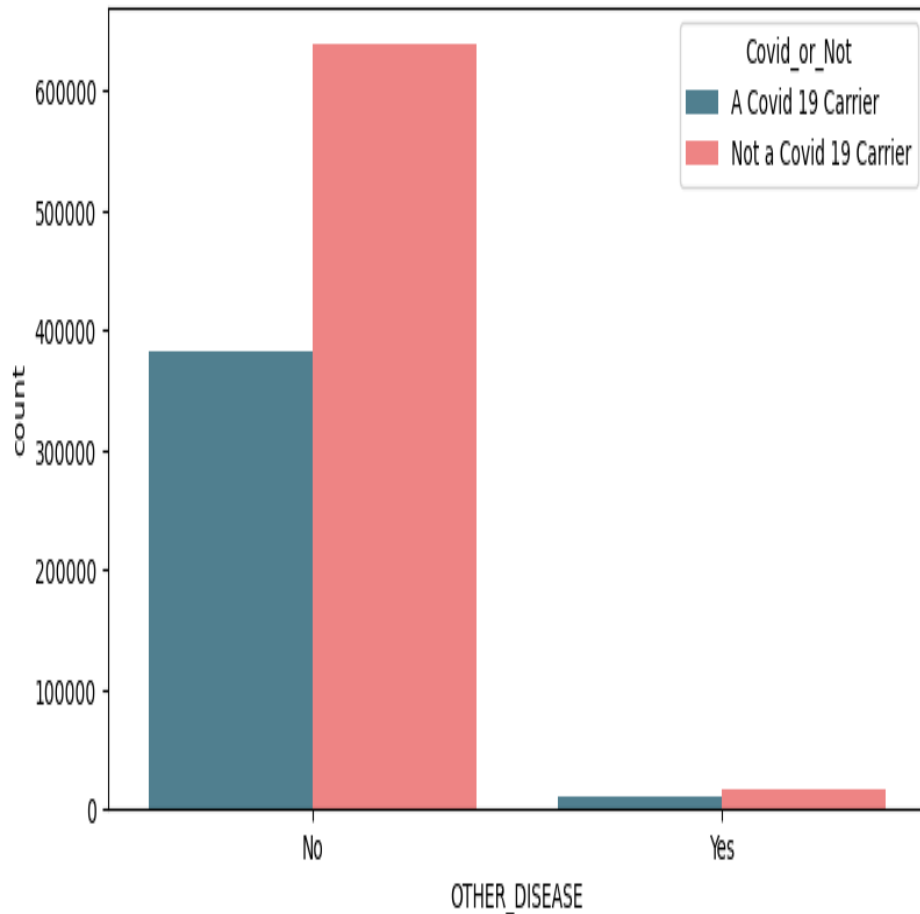
# Exploratory Data Analysis

## Effect of diseases on Covid



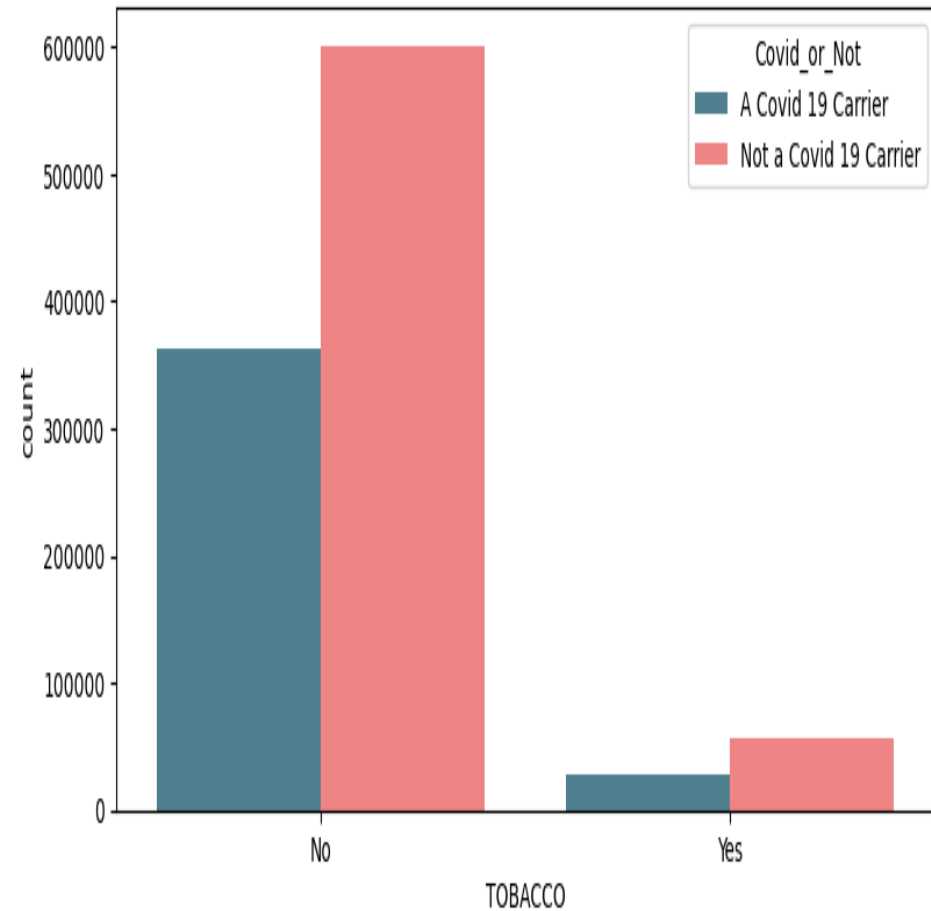
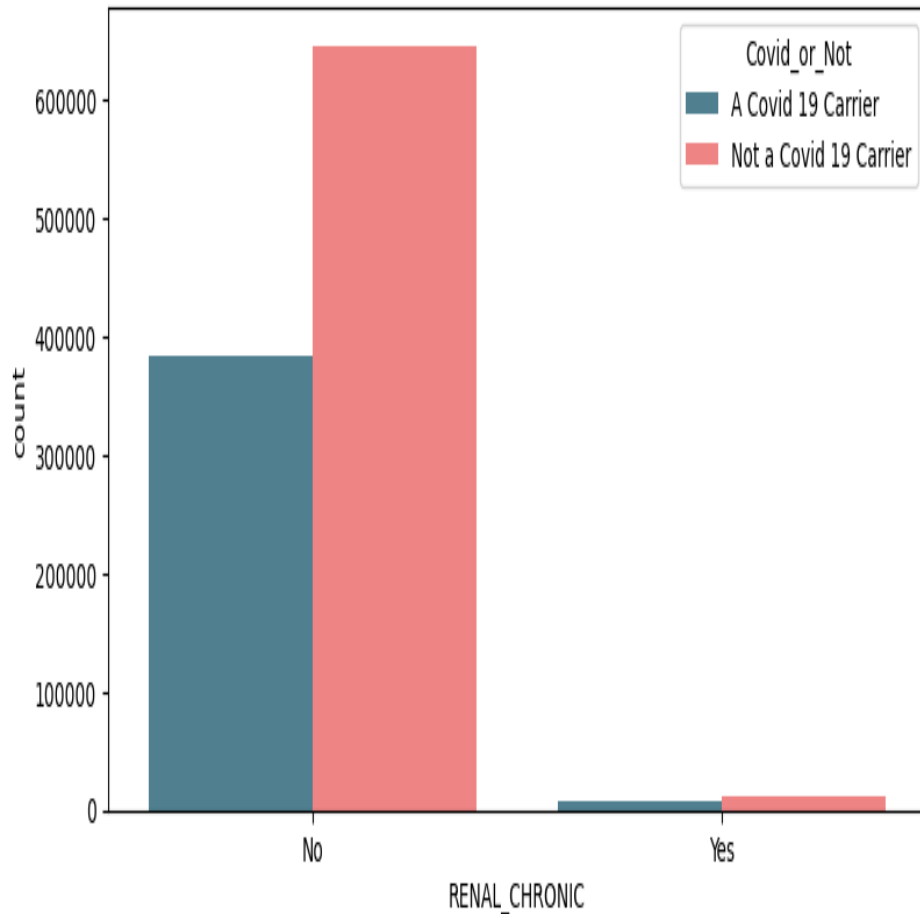
# Exploratory Data Analysis

## Effect of diseases on Covid



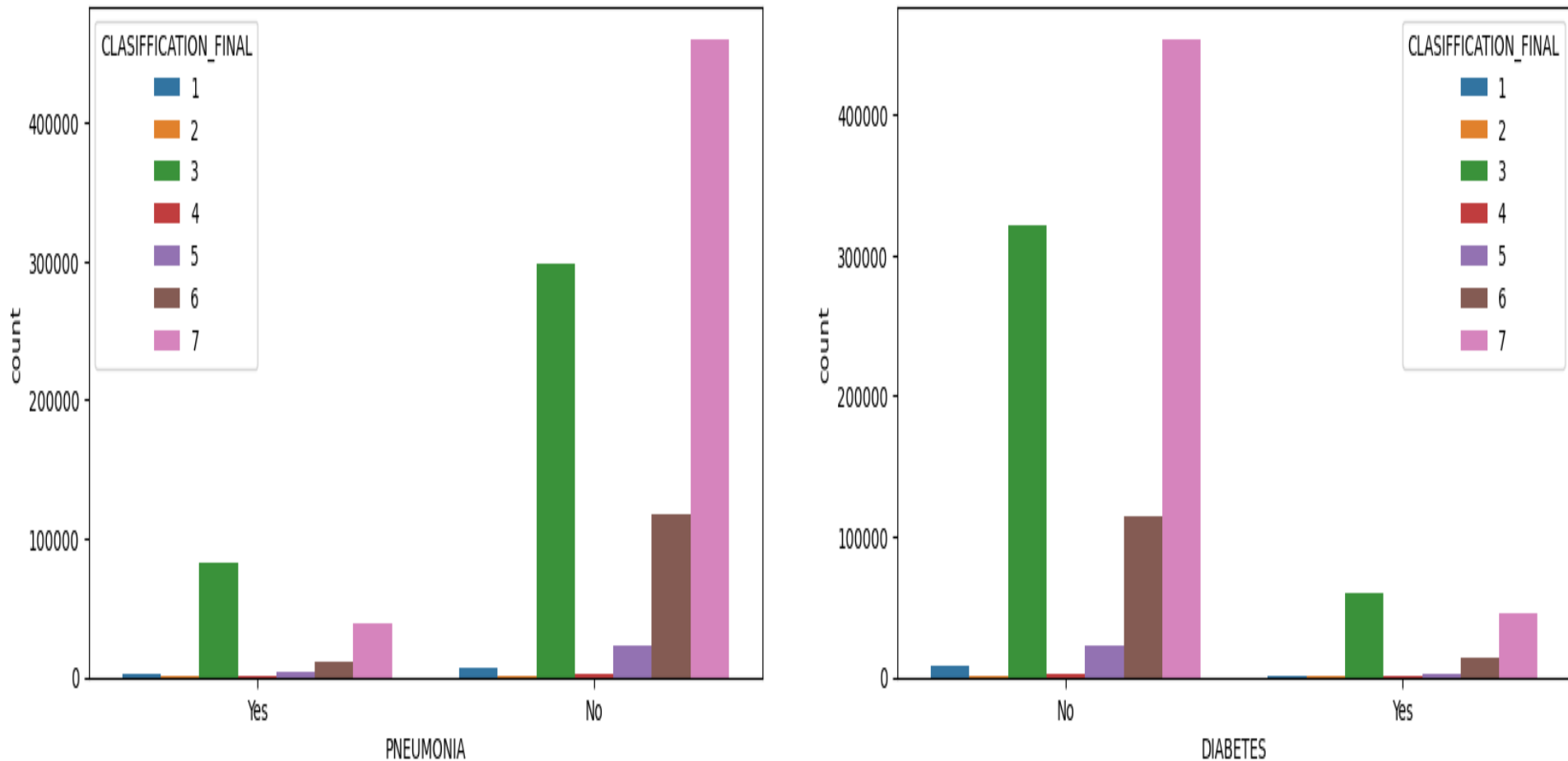
# Exploratory Data Analysis

## Effect of diseases on Covid



# Exploratory Data Analysis

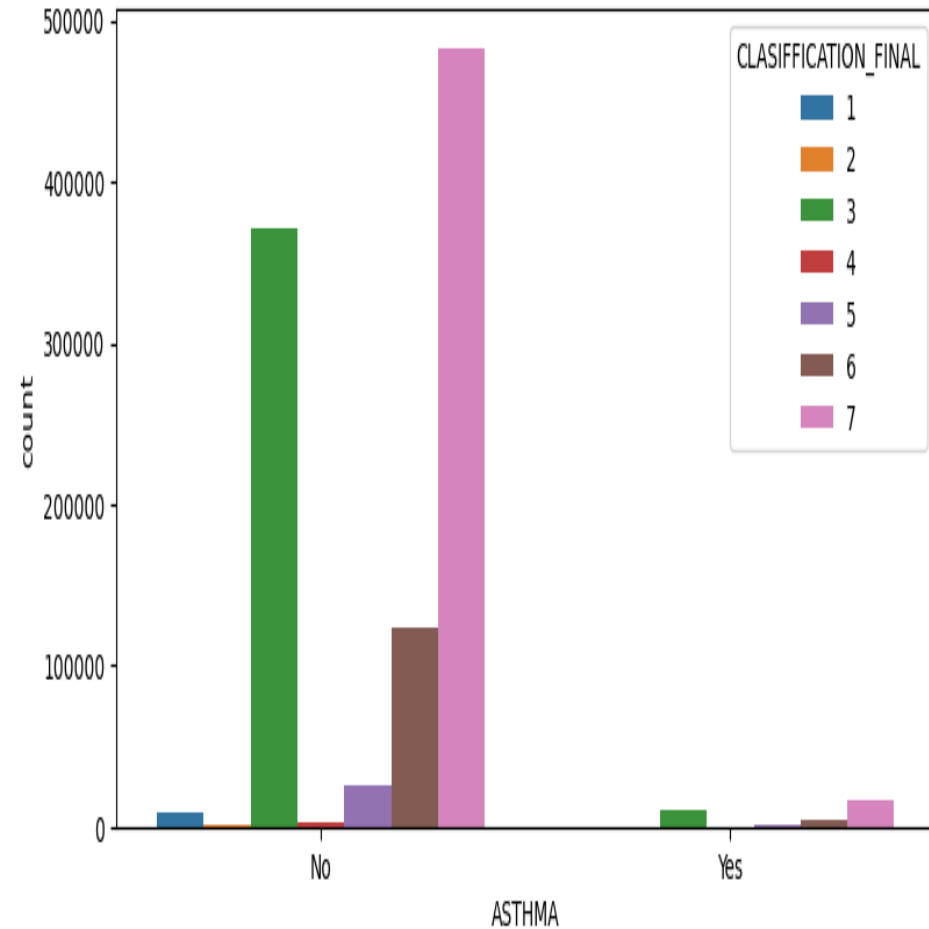
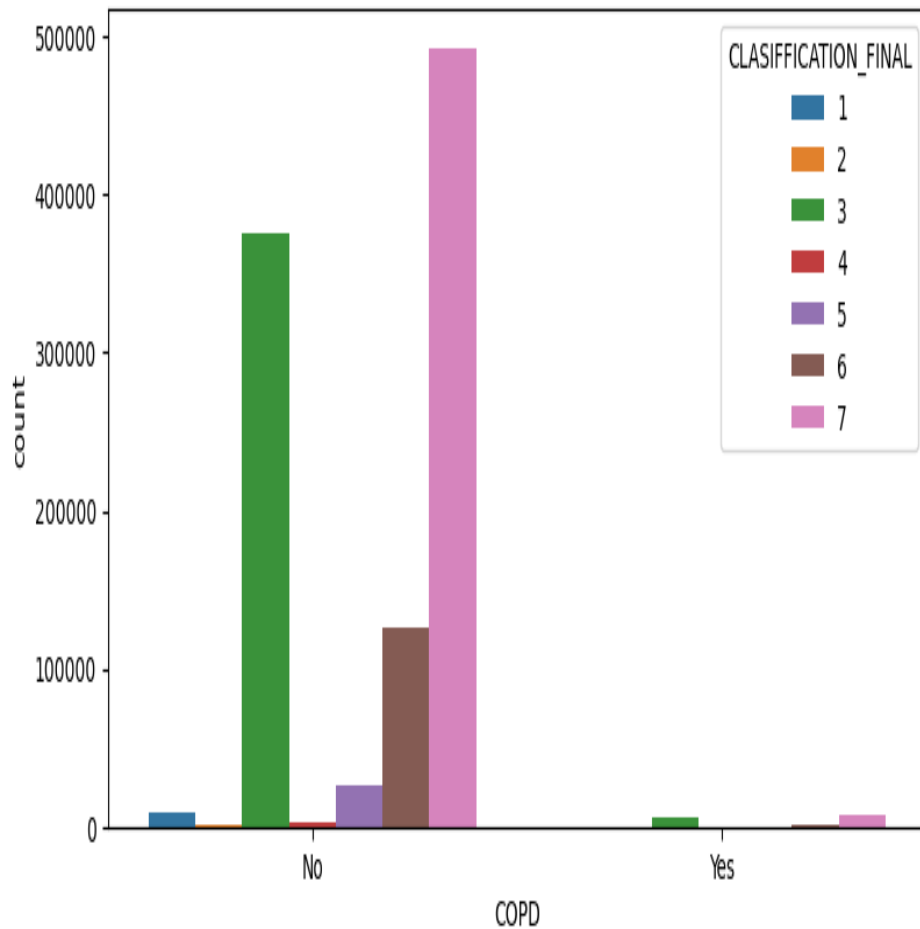
## Effect of diseases on Covid Classification





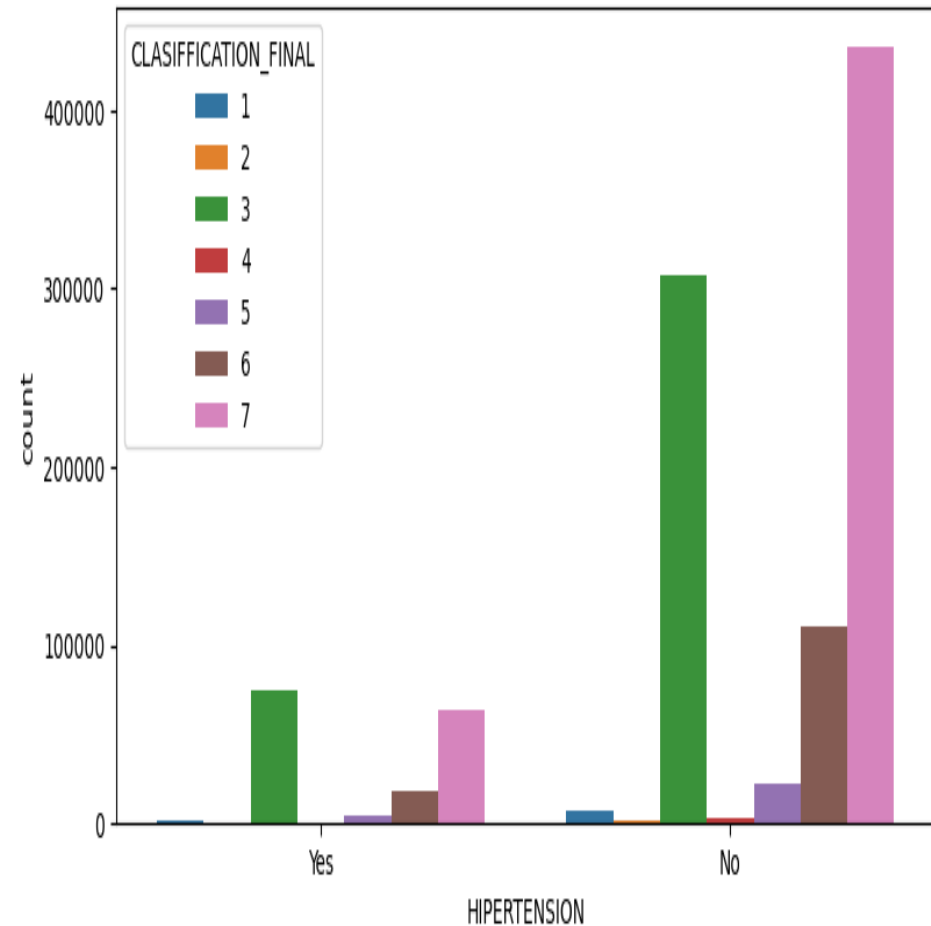
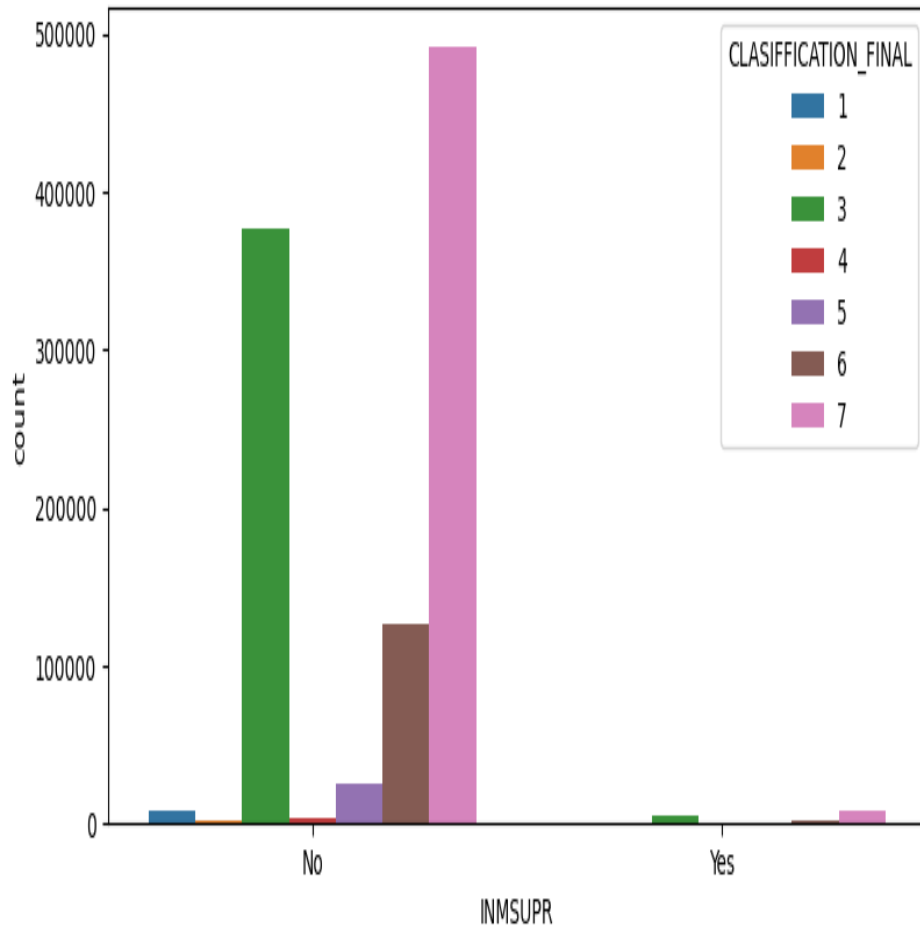
# Exploratory Data Analysis

## Effect of diseases on Covid Classification



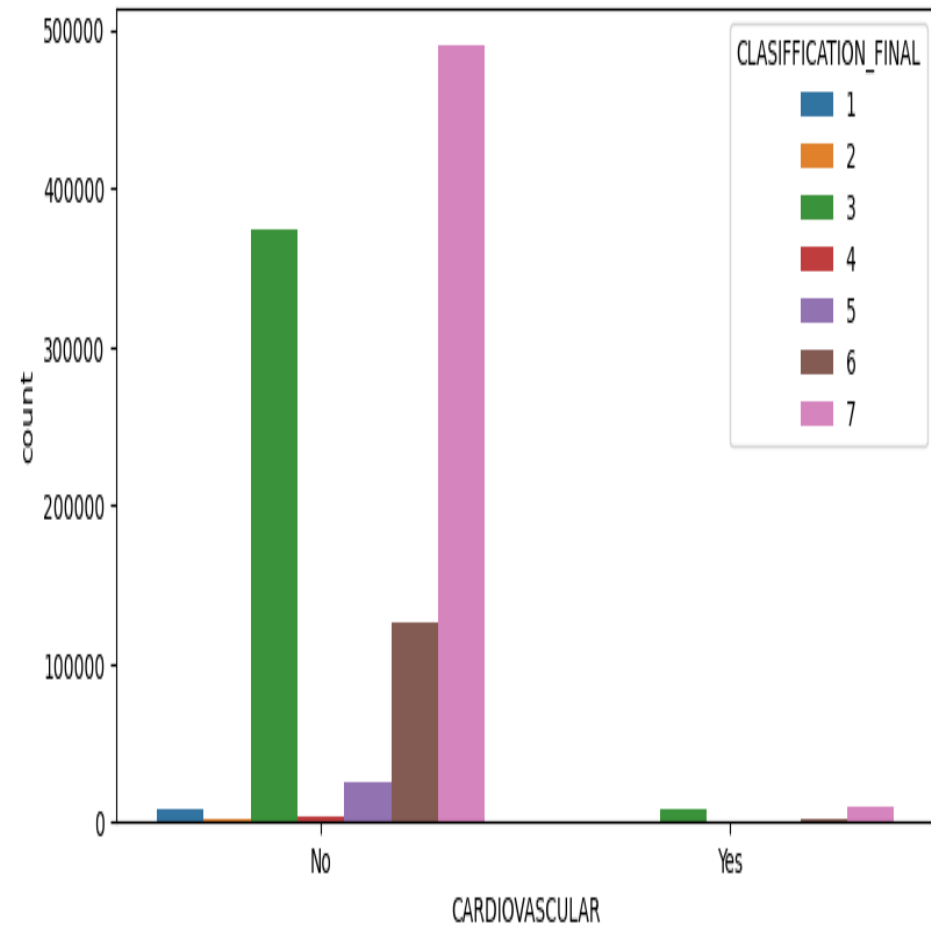
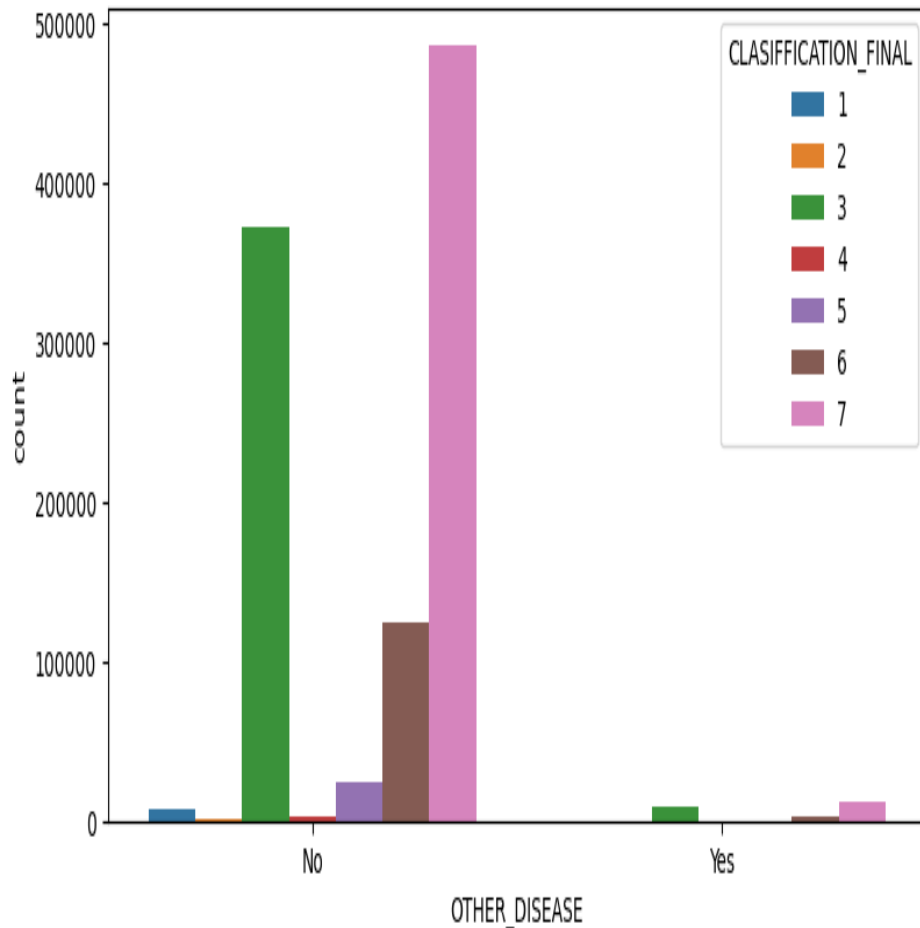
# Exploratory Data Analysis

## Effect of diseases on Covid Classification



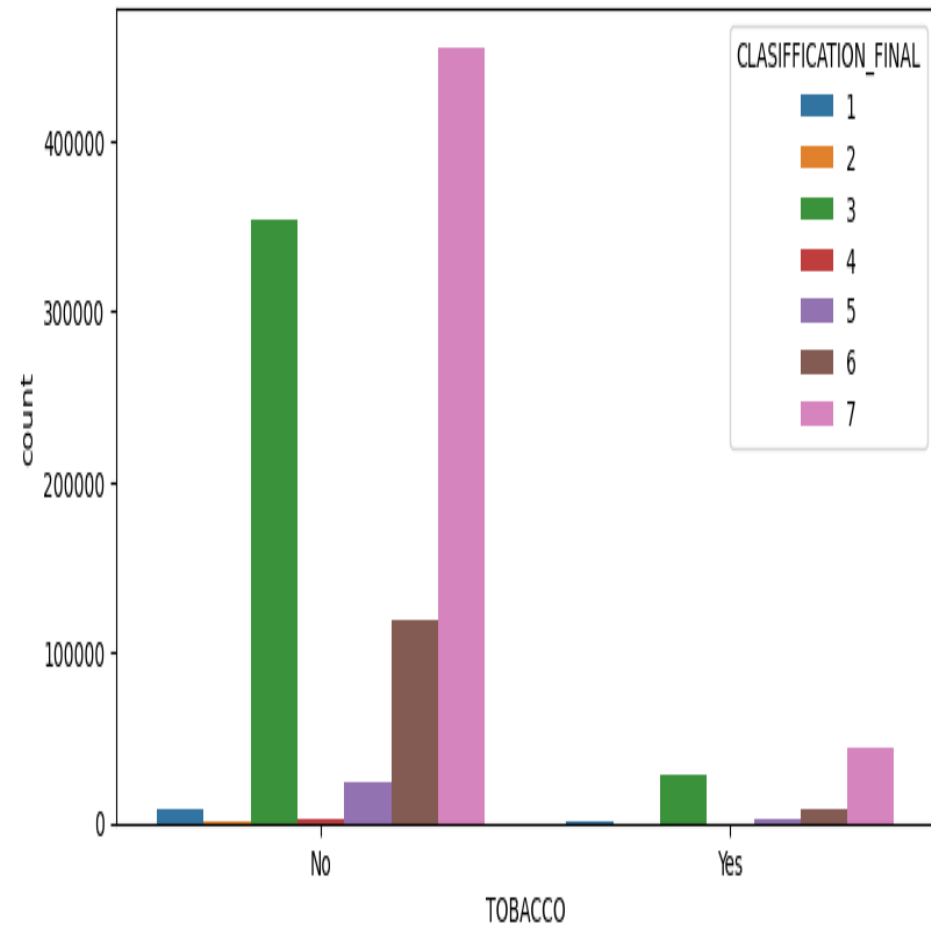
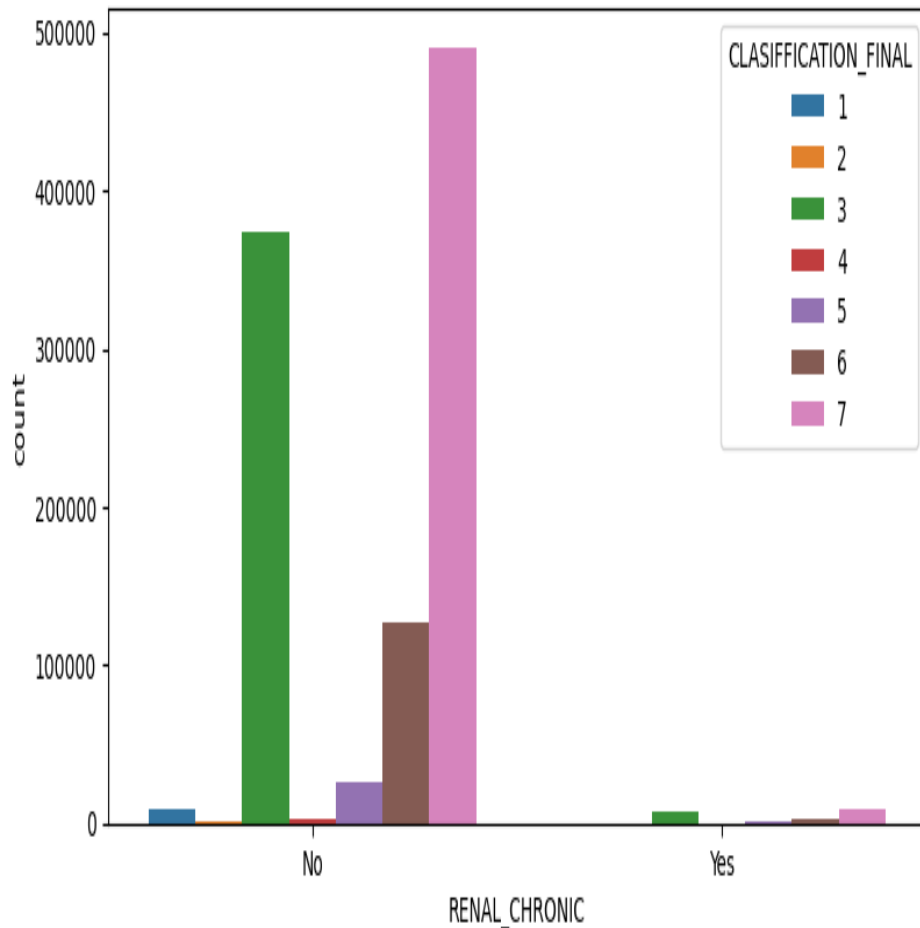
# Exploratory Data Analysis

## Effect of diseases on Covid Classification



# Exploratory Data Analysis

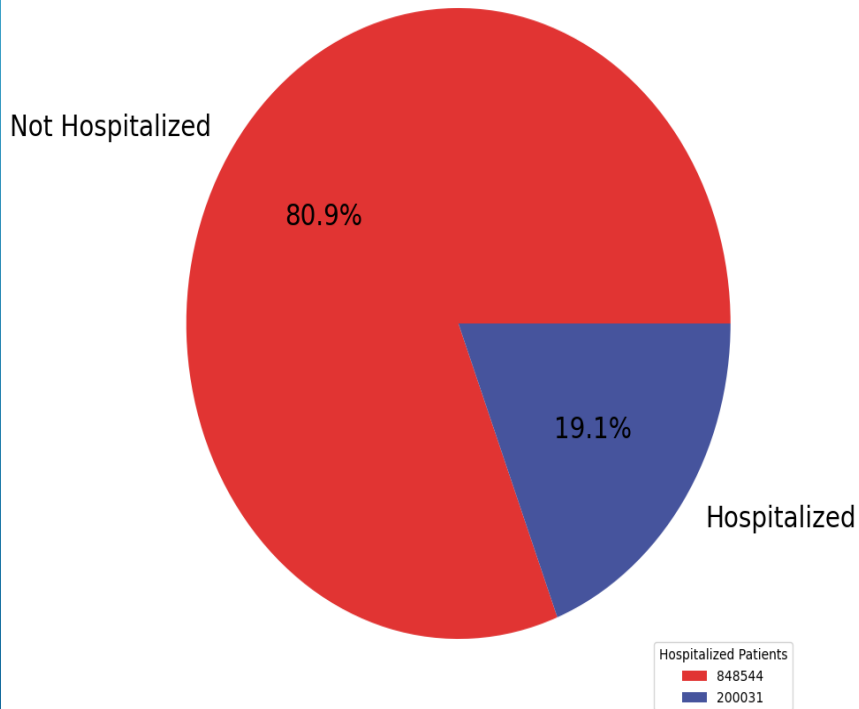
## Effect of diseases on Covid Classification



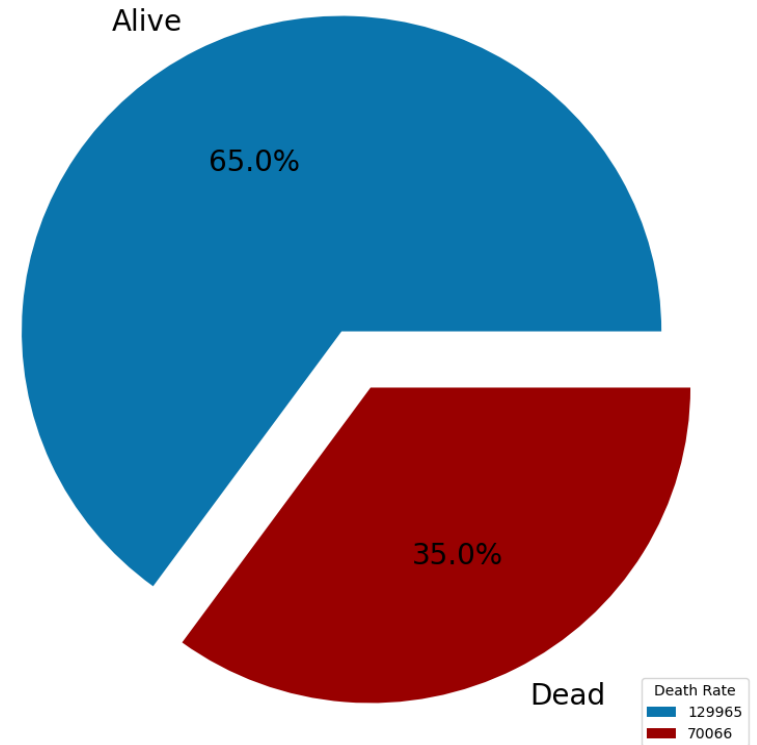
# Exploratory Data Analysis

## Hospital statistics

Hospitalized Patients



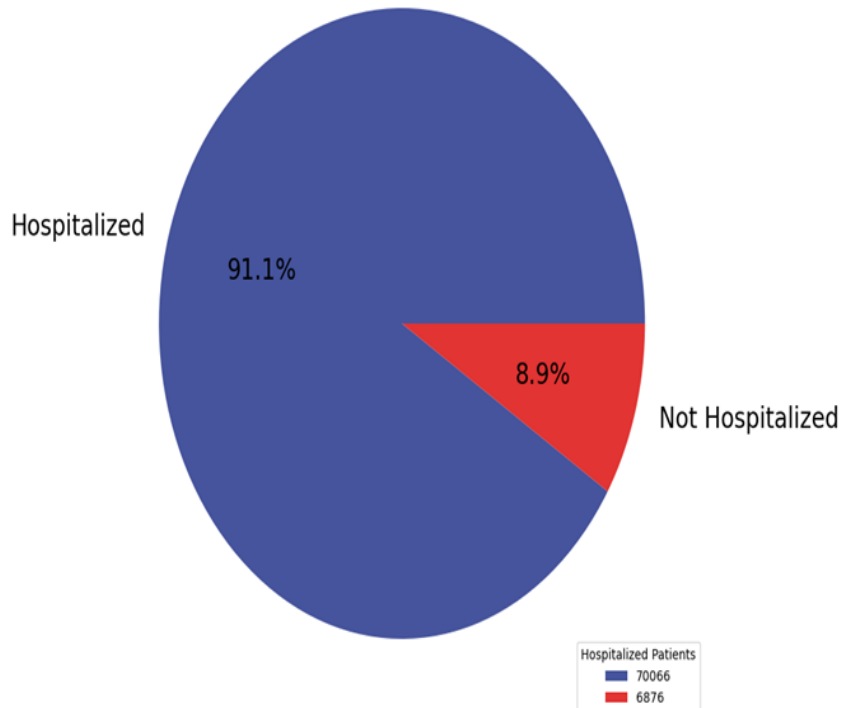
Death Rate among Hospitalized People



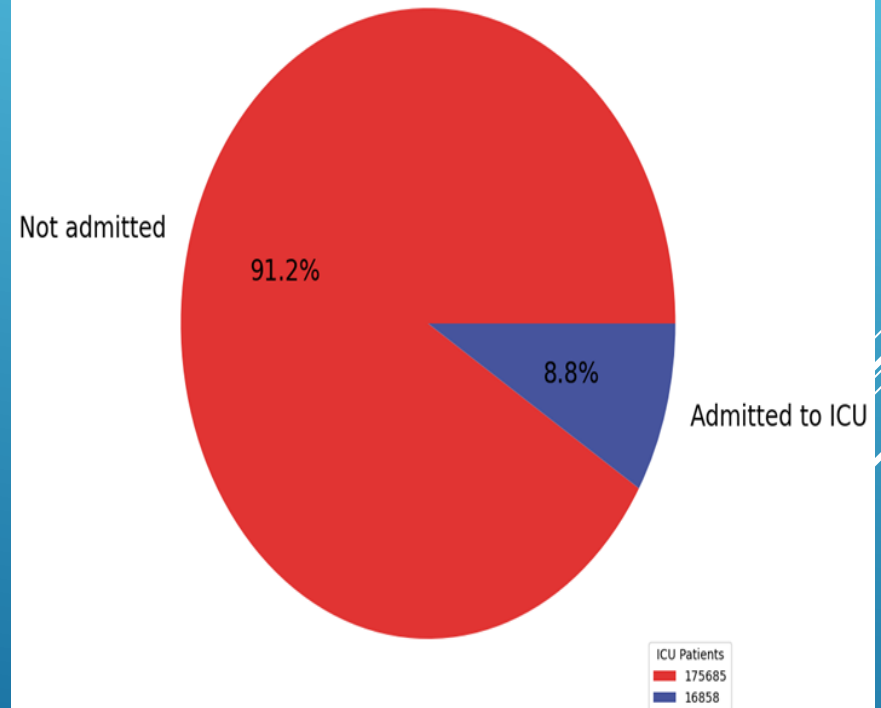
# Exploratory Data Analysis

## Hospital statistics

Hospitalized Patients Percentage among the Dead



ICU Admitted Patients

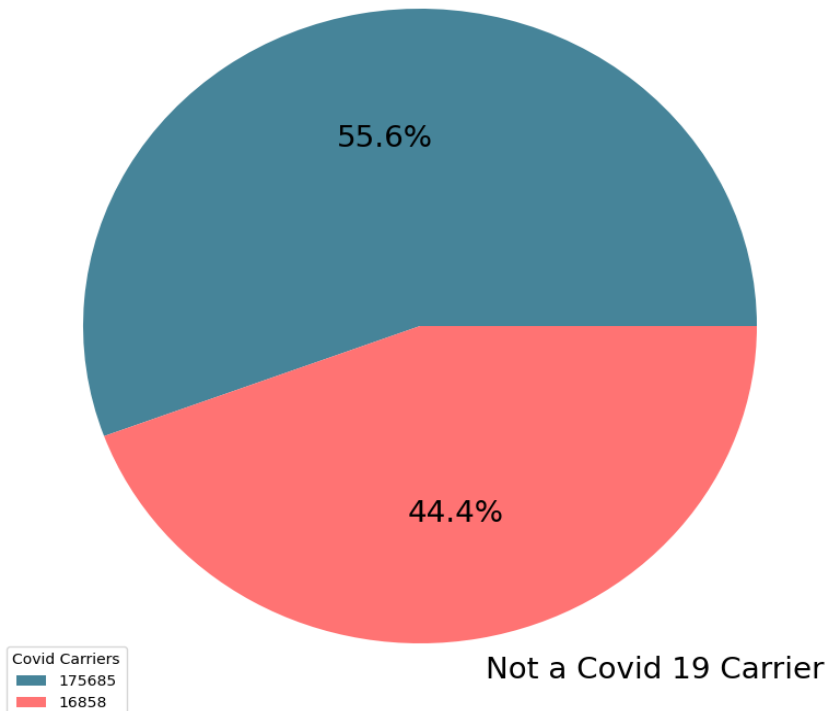


# Exploratory Data Analysis

## Hospital statistics

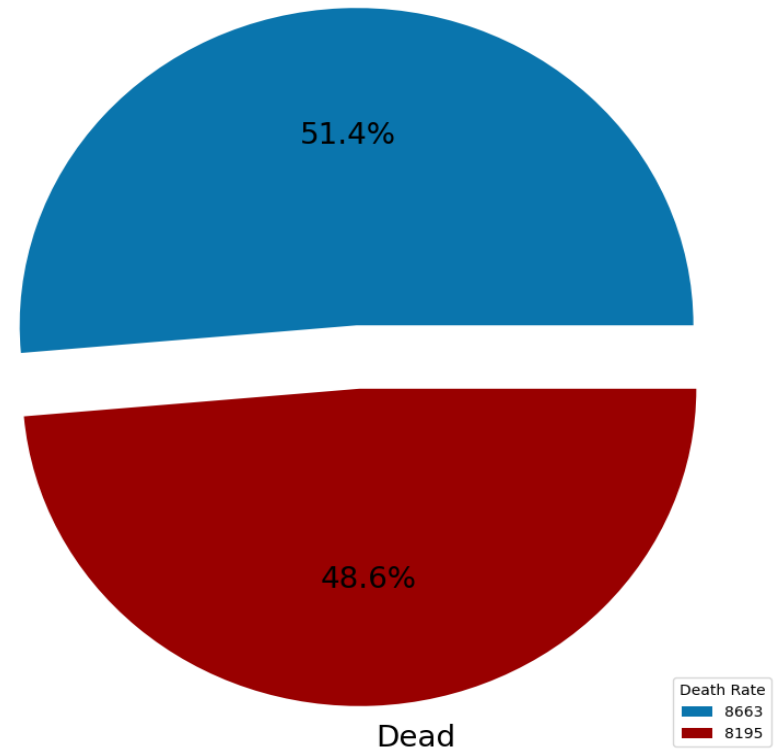
Covid Carriers among ICU Patients

A Covid 19 Carrier



Death Rate among ICU Patients

Alive



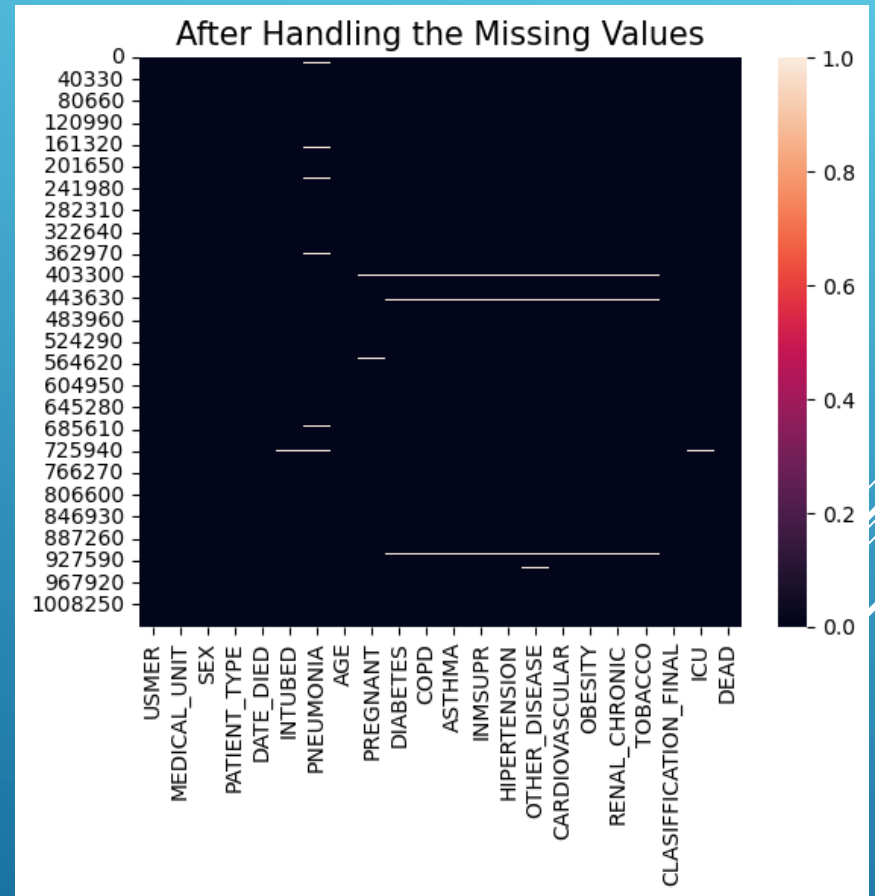
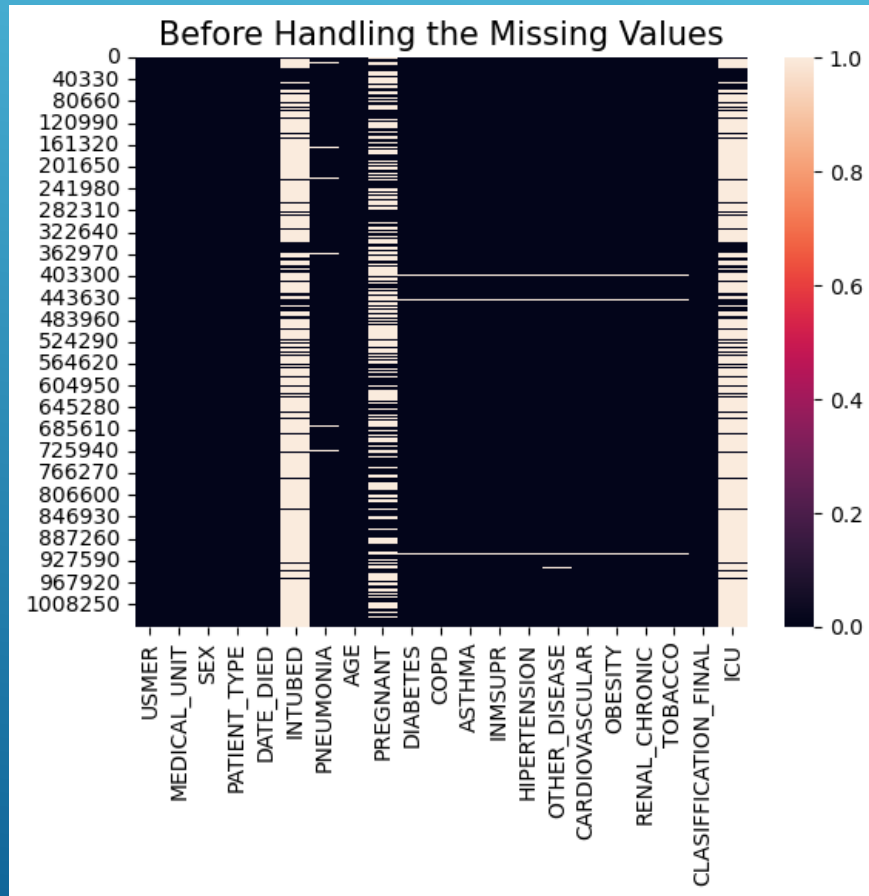
# Data Pre-processing and Problems





# Data Pre-processing and Problems

## Handling Outliers



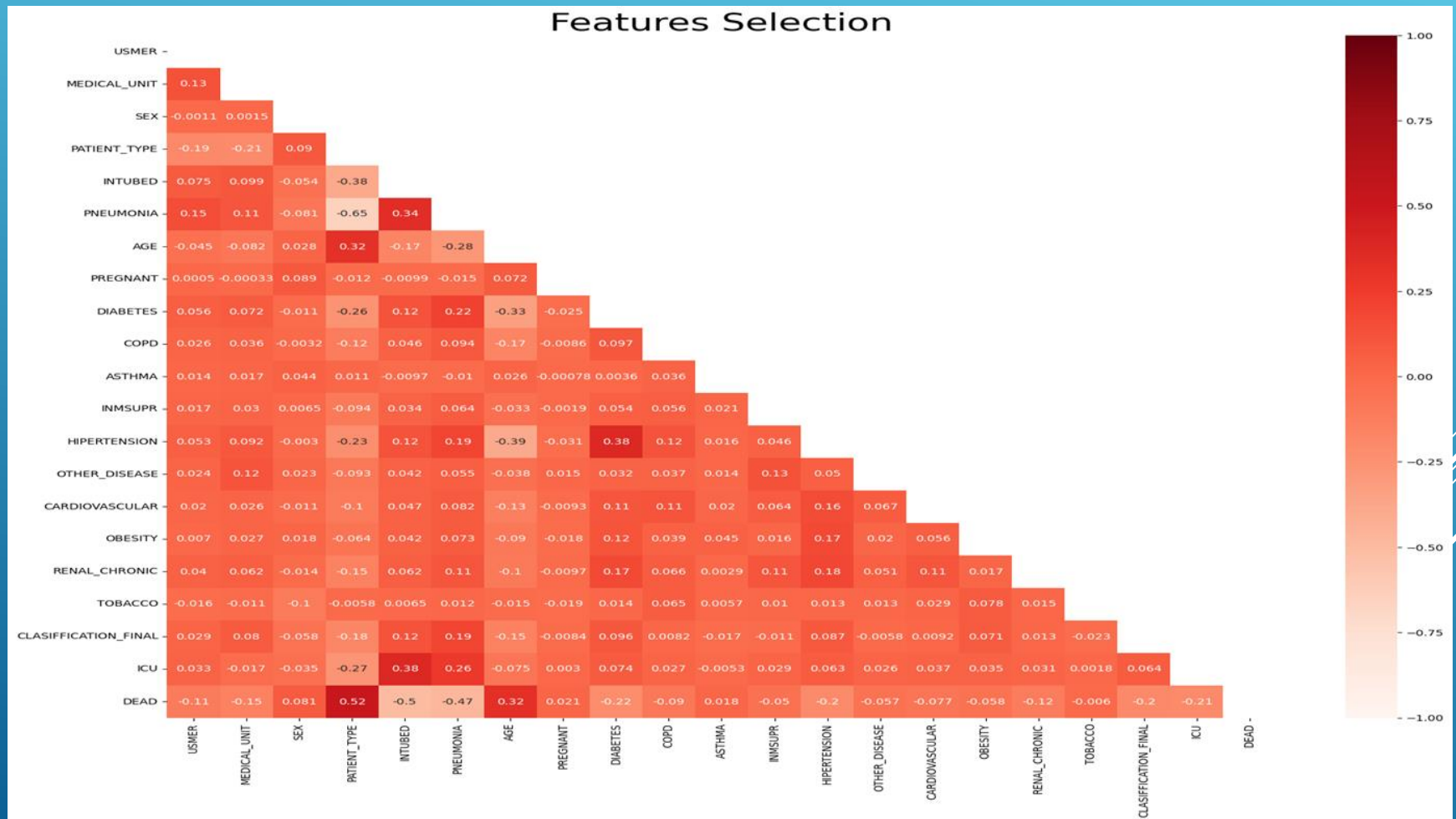
### After Dropping the Remaining Missing Values

Heatmap showing the correlation matrix of 20 variables after dropping missing values. The variables are listed on the y-axis and x-axis. The color scale ranges from -0.100 (dark purple) to 0.100 (light orange). The diagonal is dark purple (0.000). Most correlations are near zero, indicated by the uniform dark purple color.

Variable	USMER	MEDICAL_UNIT	PATIENT_TYPE	INTUBED	PNEUMONIA	AGE	PREGNANT	DIABETES	COPD	ASTHMA	INMSUPR	HIPERTENSION	OTHER_DISEASE	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO	CLASIFFICATION_FINAL	ICU	DEAD	Covid_or_Not		
USMER	0.000																						
MEDICAL_UNIT		0.000																					
PATIENT_TYPE			0.000																				
INTUBED				0.000																			
PNEUMONIA					0.000																		
AGE						0.000																	
PREGNANT							0.000																
DIABETES								0.000															
COPD									0.000														
ASTHMA										0.000													
INMSUPR											0.000												
HIPERTENSION												0.000											
OTHER_DISEASE													0.000										
CARDIOVASCULAR														0.000									
OBESITY															0.000								
RENAL_CHRONIC																0.000							
TOBACCO																	0.000						
CLASIFFICATION_FINAL																		0.000					
ICU																			0.000				
DEAD																				0.000			
Covid_or_Not																					0.000		

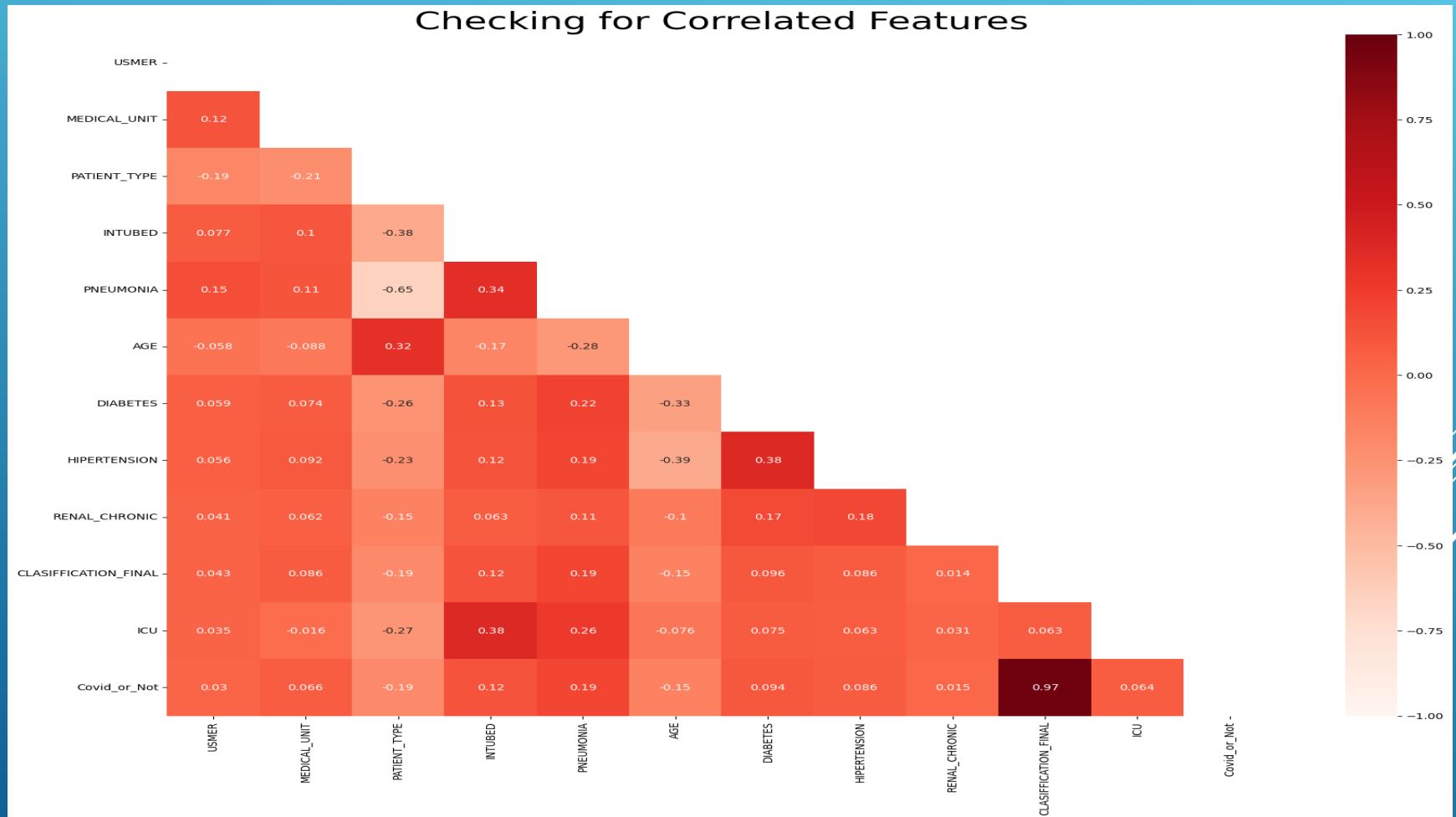
# Data Pre-processing and Problems

## Features Engineering



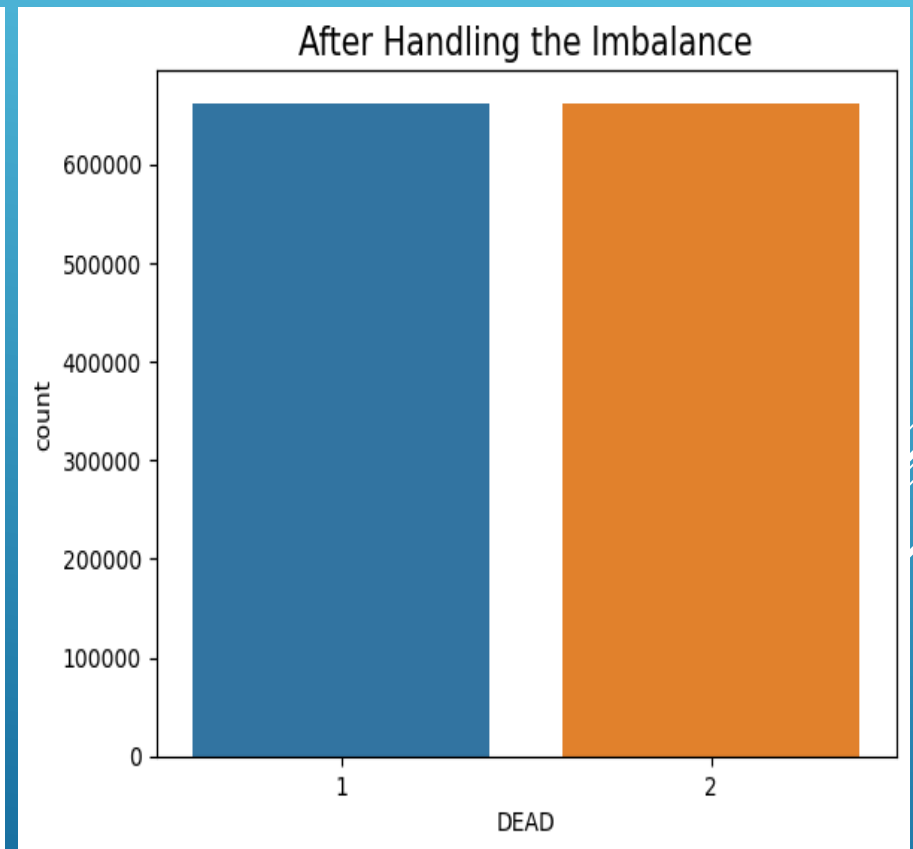
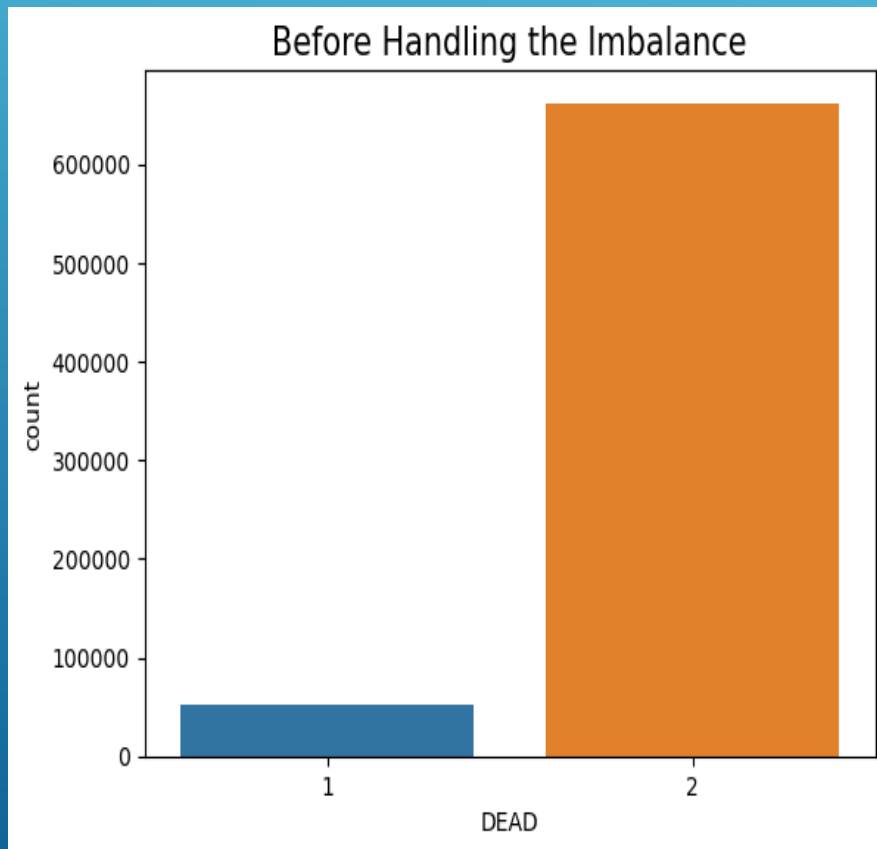
# Data Pre-processing and Problems

## Features Engineering



# 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%

# 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

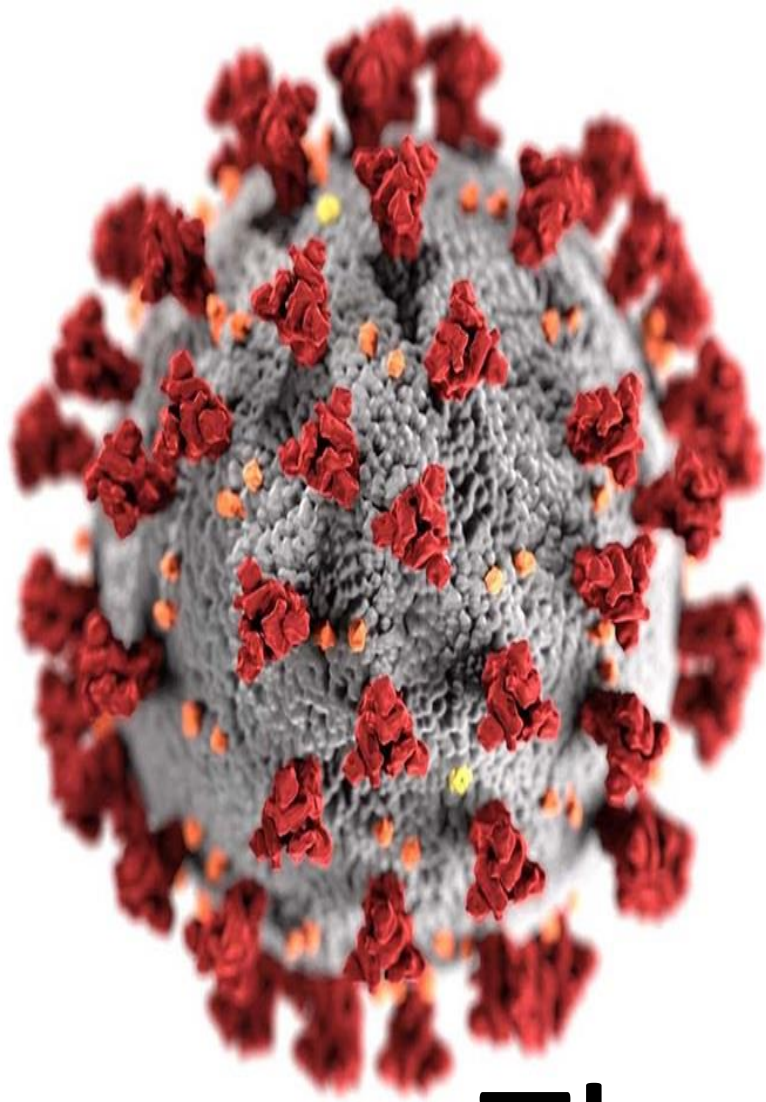
- We also noticed that there's a positive correlation between having "Hypertension" and "Diabetes" diseases; as most patients with one of those two diseases are subjected to get the other.
- We saw that among all the patients of these diseases the patients classified with 3rd degree of Covid are the 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!