SAMSUNG

Together for Tomorrow!
Enabling People
Education for Future Generations

Samsung Innovation Campus

Artificial Intelligence Course

Who we are?

Hello, We are Hunters Team.

_Here It’s out team members:

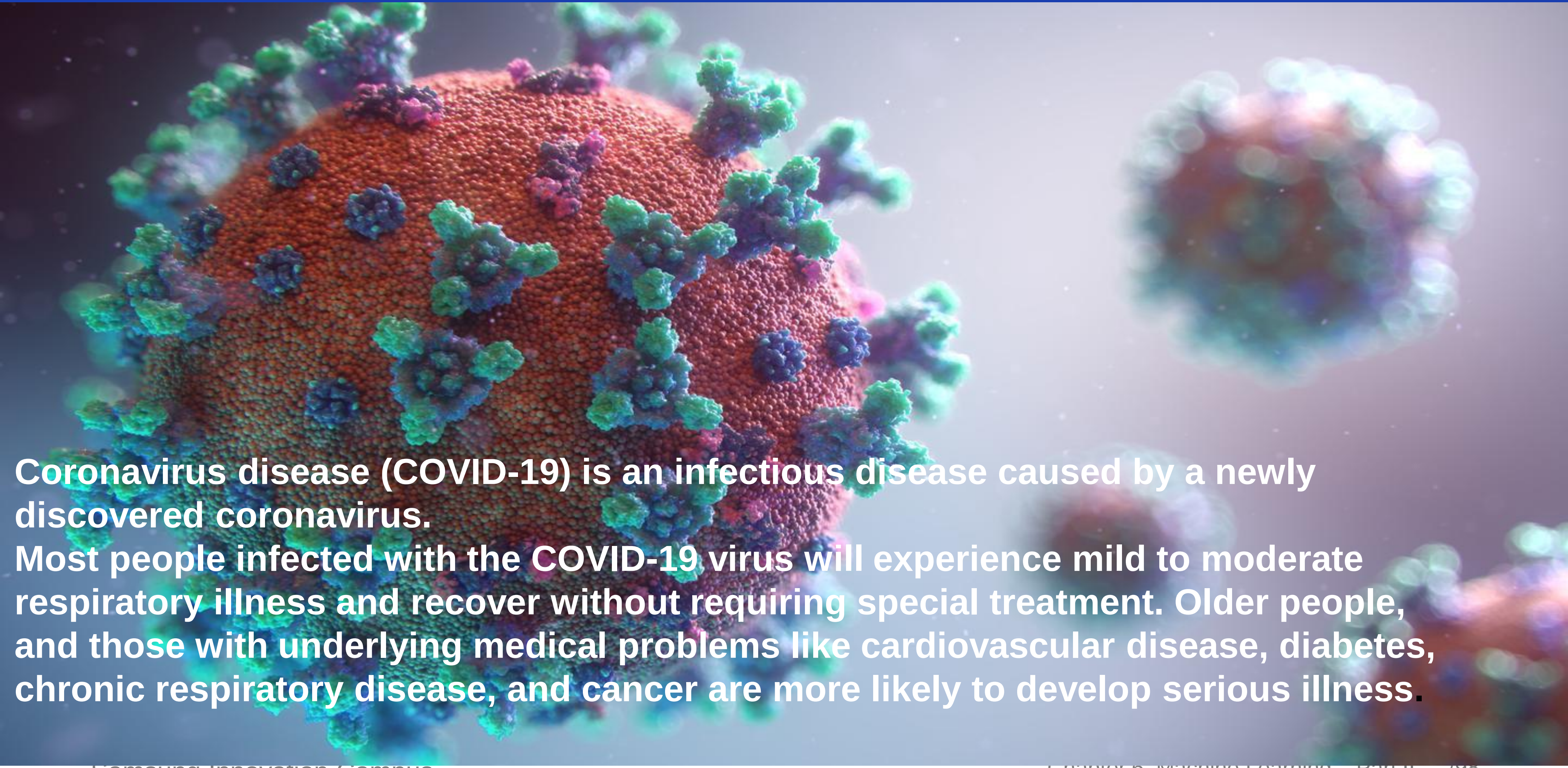
- **Mohamed Abd El-Mohsen**
- **Ahmed Gaber**
- **Ahmed Fathy**

_And Our supervisor**: Eng. Shima Osman.**

What we will inserted in ?

1) Description the data
2) Checking if there is missing values
3) Data analysis with visualization
4) Data Preprocessing
5) Applying the classification model
6) Checking the accuracy of the model
7) predict a random samble

What is COVID 19 ?



Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus.
Most people infected with the 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.

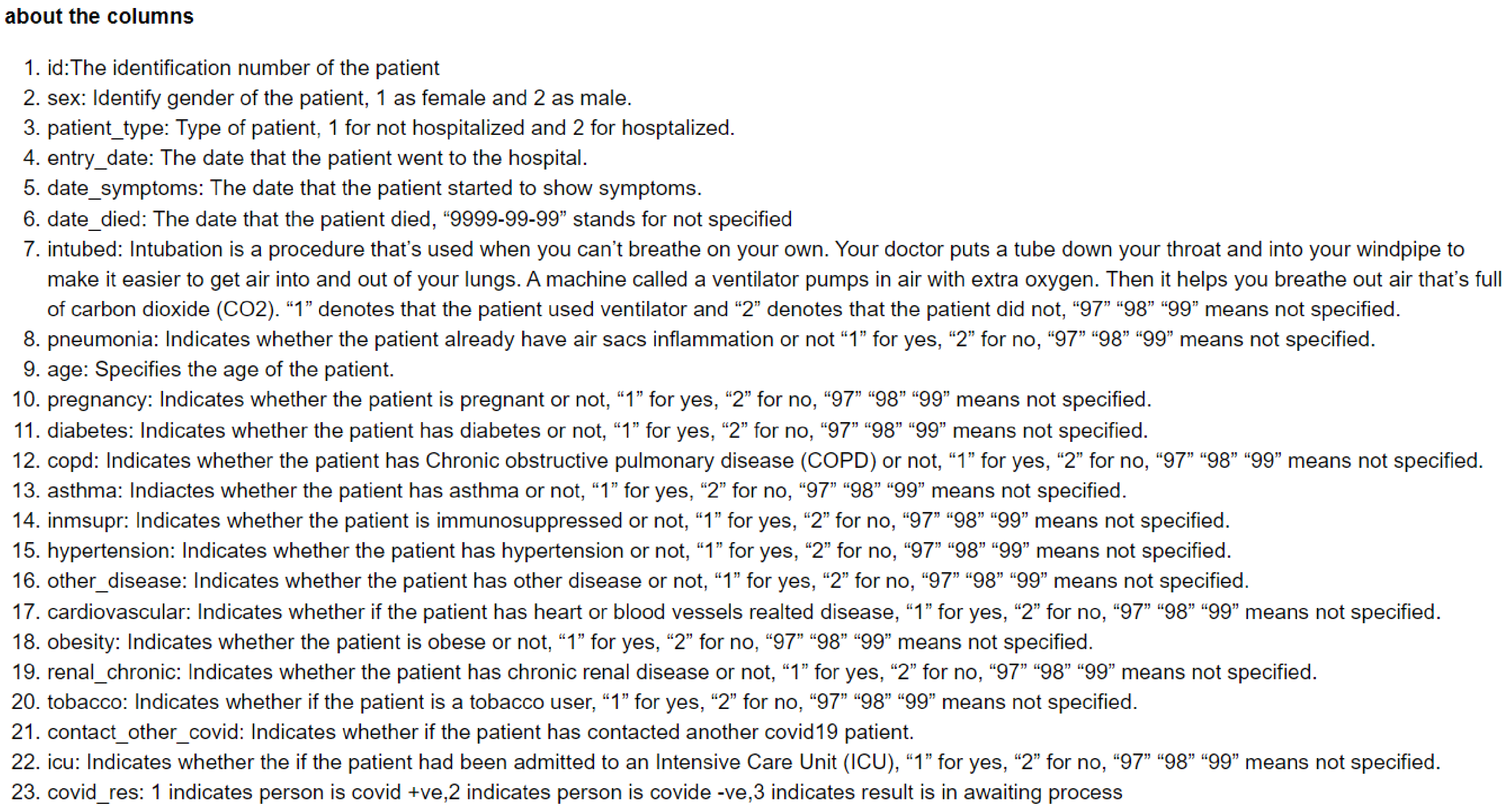
_About our data:

- Covid-19 data set
- Data obtained from Mexican government data set
- The Data Link

_The objective:

- the current COVID-9 pandemic provides us **with an opportunity to ponder and reflect over what we can better in the way we deal with healthcare to make us humans be more prepared** and enabled to combat such an event in the future.
- **getting insights** which help the Medical kits.

Get The 1st Intuition.



about the columns

1. id:The identification number of the patient
2. sex: Identify gender of the patient, 1 as female and 2 as male.
3. patient_type: Type of patient, 1 for not hospitalized and 2 for hosptalized.
4. entry_date: The date that the patient went to the hospital.
5. date_symptoms: The date that the patient started to show symptoms.
6. date_died: The date that the patient died, “9999-99-99” stands for not specified
7. intubed: Intubation is a procedure that’s used when you can’t breathe on your own. Your doctor puts a tube down your throat and into your windpipe to make it easier to get air into and out of your lungs. A machine called a ventilator pumps in air with extra oxygen. Then it helps you breathe out air that’s full of carbon dioxide (CO2). “1” denotes that the patient used ventilator and “2” denotes that the patient did not, “97” “98” “99” means not specified.
8. pneumonia: Indicates whether the patient already have air sacs inflammation or not “1” for yes, “2” for no, “97” “98” “99” means not specified.
9. age: Specifies the age of the patient.
10. pregnancy: Indicates whether the patient is pregnant or not, “1” for yes, “2” for no, “97” “98” “99” means not specified.
11. diabetes: Indicates whether the patient has diabetes or not, “1” for yes, “2” for no, “97” “98” “99” means not specified.
12. copd: Indicates whether the patient has Chronic obstructive pulmonary disease (COPD) or not, “1” for yes, “2” for no, “97” “98” “99” means not specified.
13. asthma: Indiactes whether the patient has asthma or not, “1” for yes, “2” for no, “97” “98” “99” means not specified.
14. inmsupr: Indicates whether the patient is immunosuppressed or not, “1” for yes, “2” for no, “97” “98” “99” means not specified.
15. hypertension: Indicates whether the patient has hypertension or not, “1” for yes, “2” for no, “97” “98” “99” means not specified.
16. other_disease: Indicates whether the patient has other disease or not, “1” for yes, “2” for no, “97” “98” “99” means not specified.
17. cardiovascular: Indicates whether if the patient has heart or blood vessels realted disease, “1” for yes, “2” for no, “97” “98” “99” means not specified.
18. obesity: Indicates whether the patient is obese or not, “1” for yes, “2” for no, “97” “98” “99” means not specified.
19. renal_chronic: Indicates whether the patient has chronic renal disease or not, “1” for yes, “2” for no, “97” “98” “99” means not specified.
20. tobacco: Indicates whether if the patient is a tobacco user, “1” for yes, “2” for no, “97” “98” “99” means not specified.
21. contact_other_covid: Indicates whether if the patient has contacted another covid19 patient.
22. icu: Indicates whether the if the patient had been admitted to an Intensive Care Unit (ICU), “1” for yes, “2” for no, “97” “98” “99” means not specified.
23. covid_res: 1 indicates person is covid +ve,2 indicates person is covide -ve,3 indicates result is in awaiting process

Data Wrangling.

- **The Data Shape**: 566602 rows, 23 Columns.

- **No_Null** Values**.**

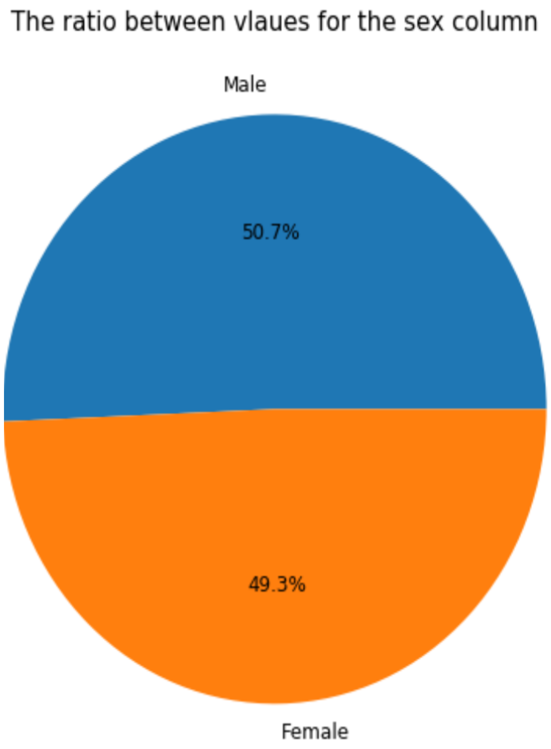
- **No_Duplicated** Rows.

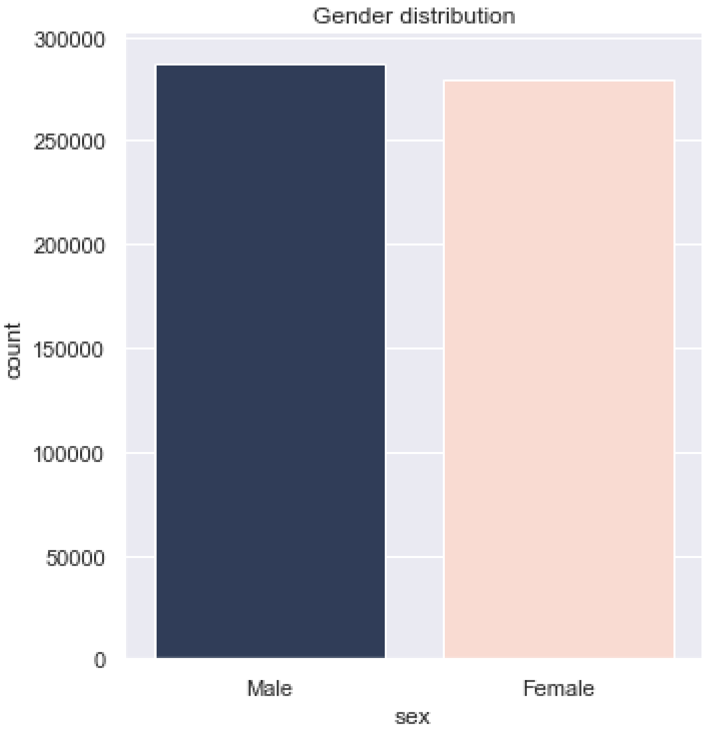
_Data To Clean:

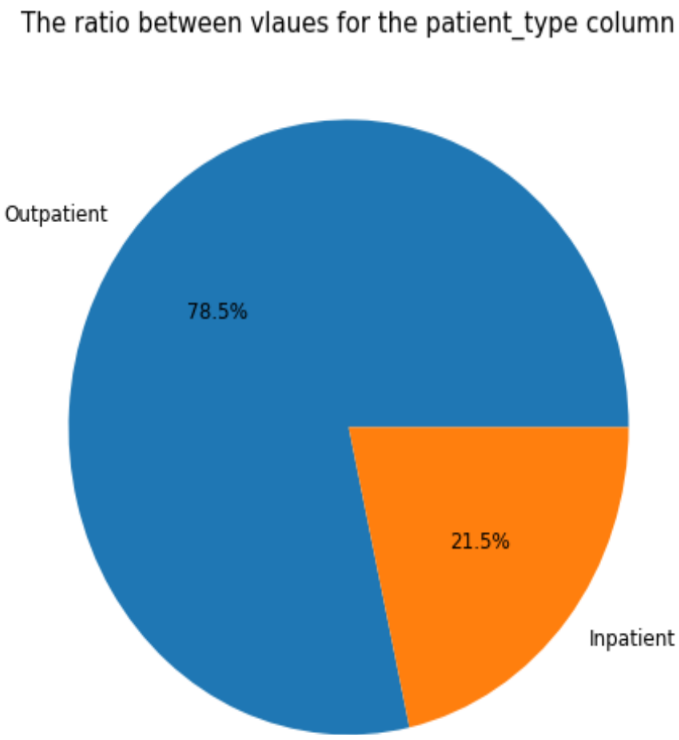
- **Drop** Un necessary column: The **id** column.
- Dealing with **Un specified** Values (date entry failure):
  {97:'Not Specified', 98:'Not Specified',99:'Not Specified'}
- **De-code** some columns. So, it becomes easy to understand in data Viz:
  (**e.g.** {1: 'Female', 2: 'Male'})
- Handling the **Date** columns **from** string **to** date-frame.

EDA: “Let’s Explore Our Data”

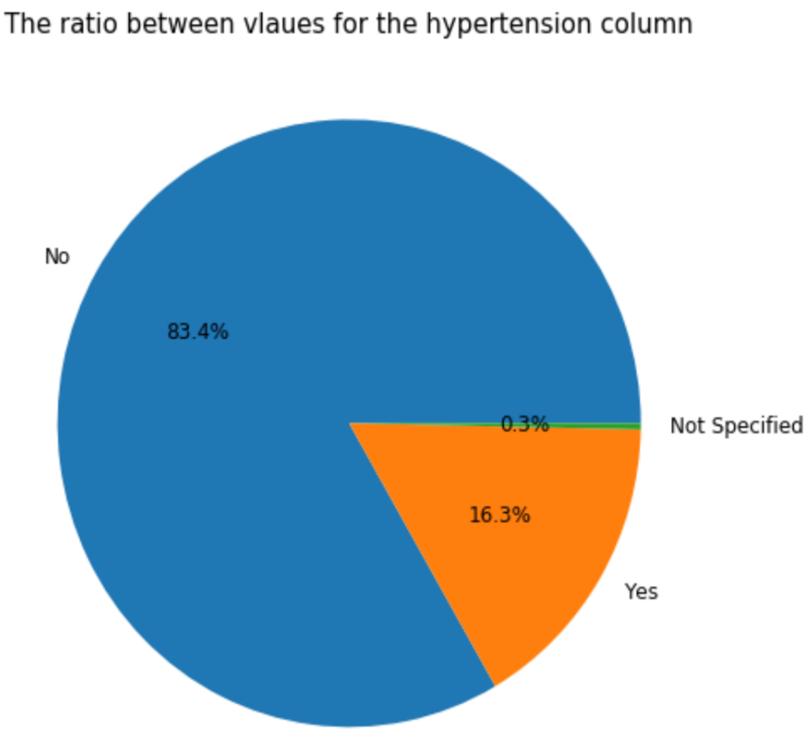
First, We try To get the Ratio of the data and this the most attentional outcomes.





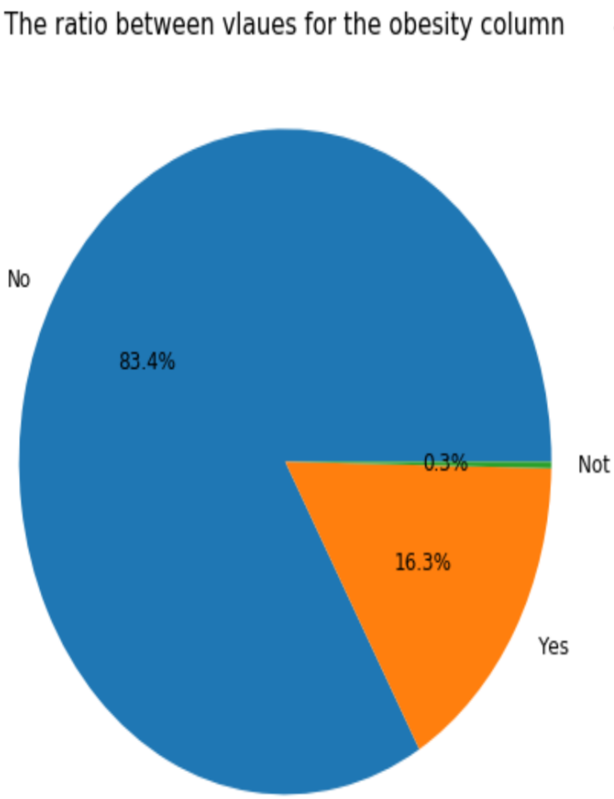


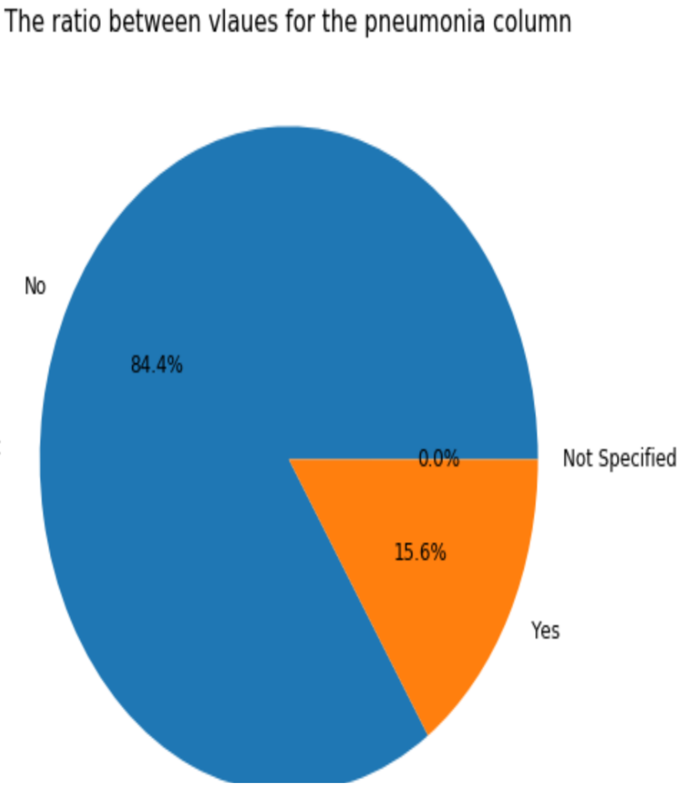
The ratio between vlaues for the patient_type column

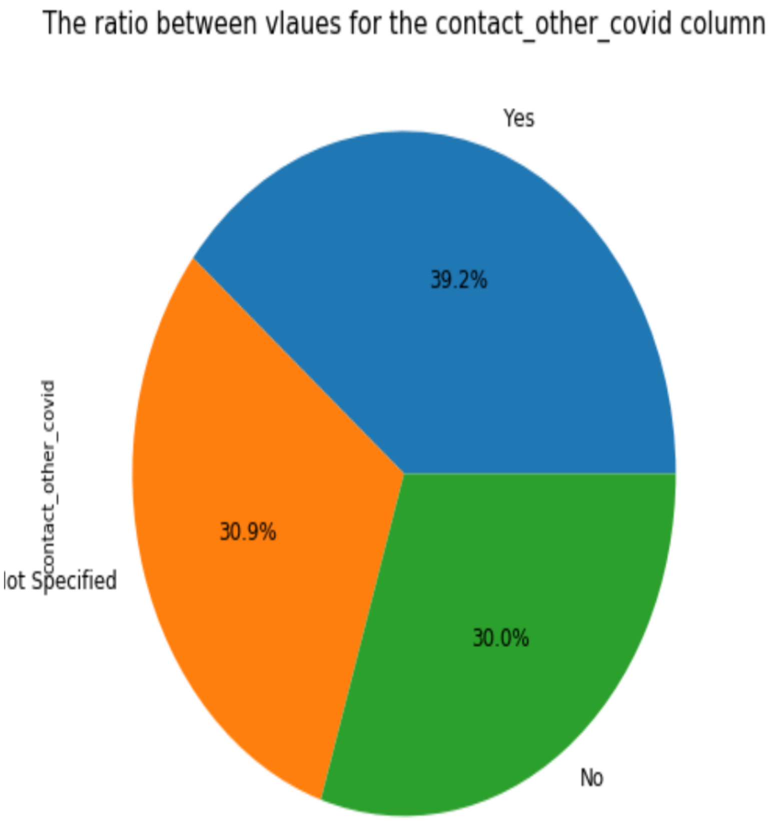


The ratio between vlaues for the hypertension column

First, We try To get the Ratio of the data and this the most attentional outcomes.

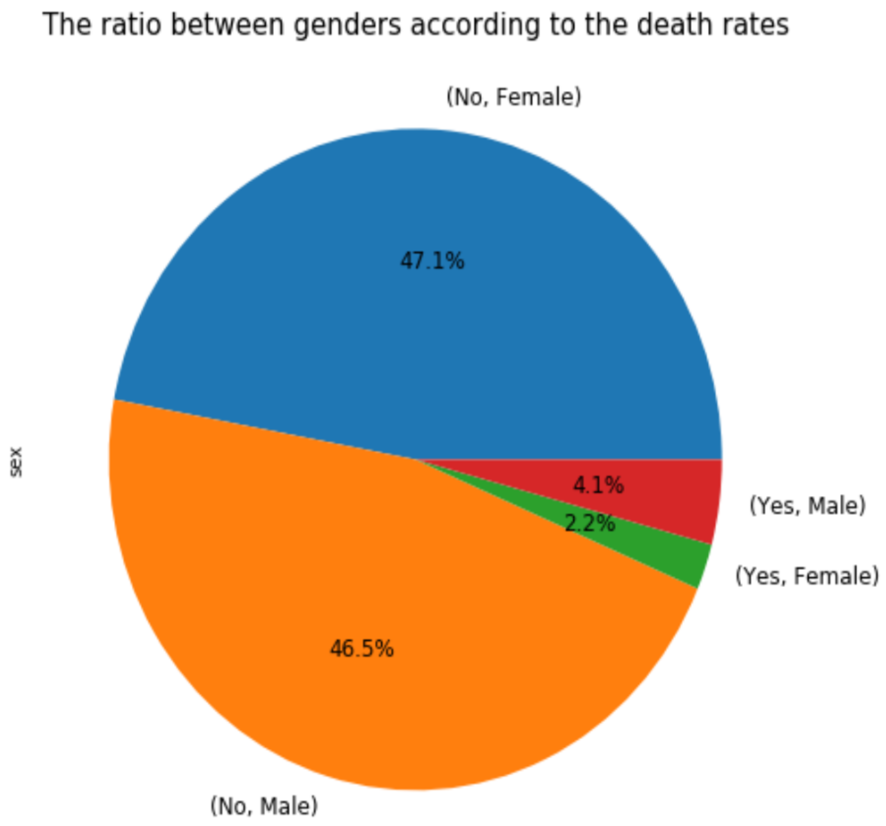

The ratio between vlaues for the obesity column

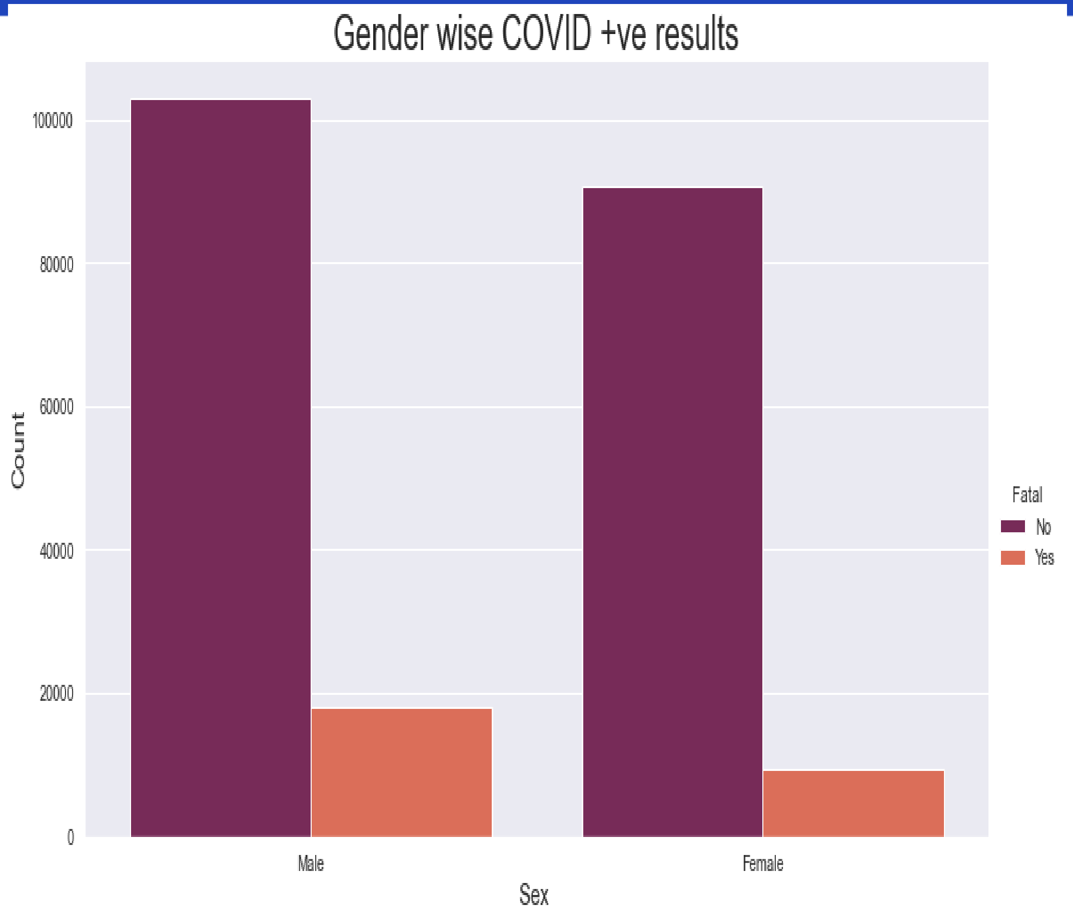

The ratio between vlaues for the pneumonia column


The ratio between vlaues for the contact_other_covid column

“Deal With the 1st outcome”

Did the **ratio** between the gender **tests** and **deaths** are the **same** too?

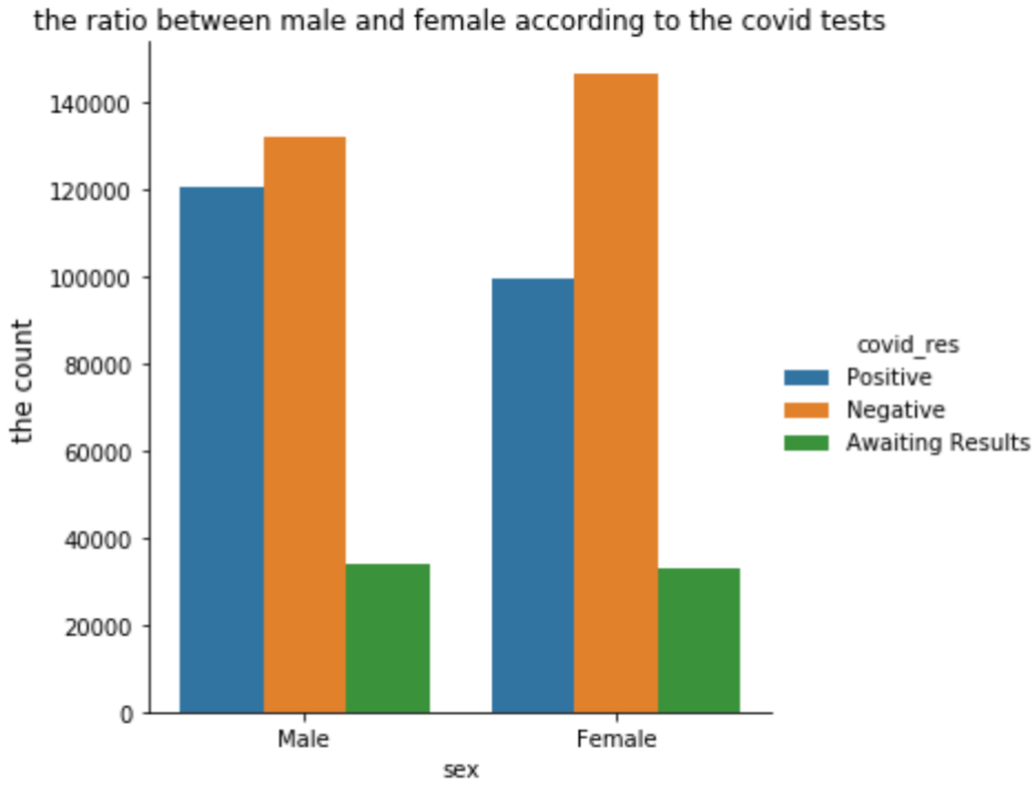




Male fatality : 14.81 %

Female fatality: 7.64 %

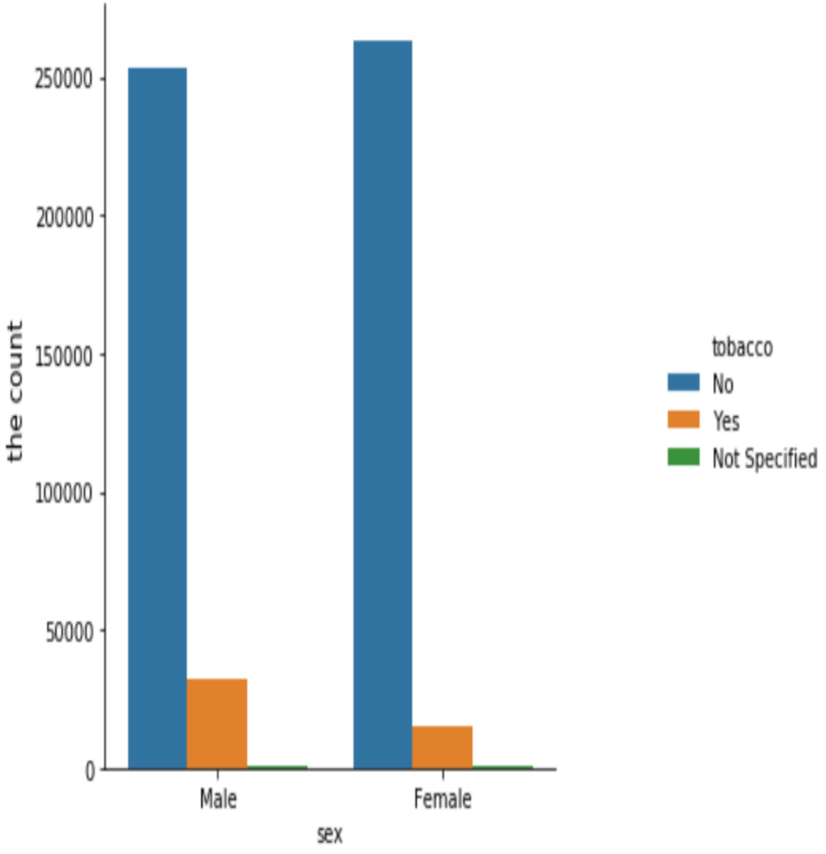
Conclusion:The ratio of people who dead of males is duobled from the feamles



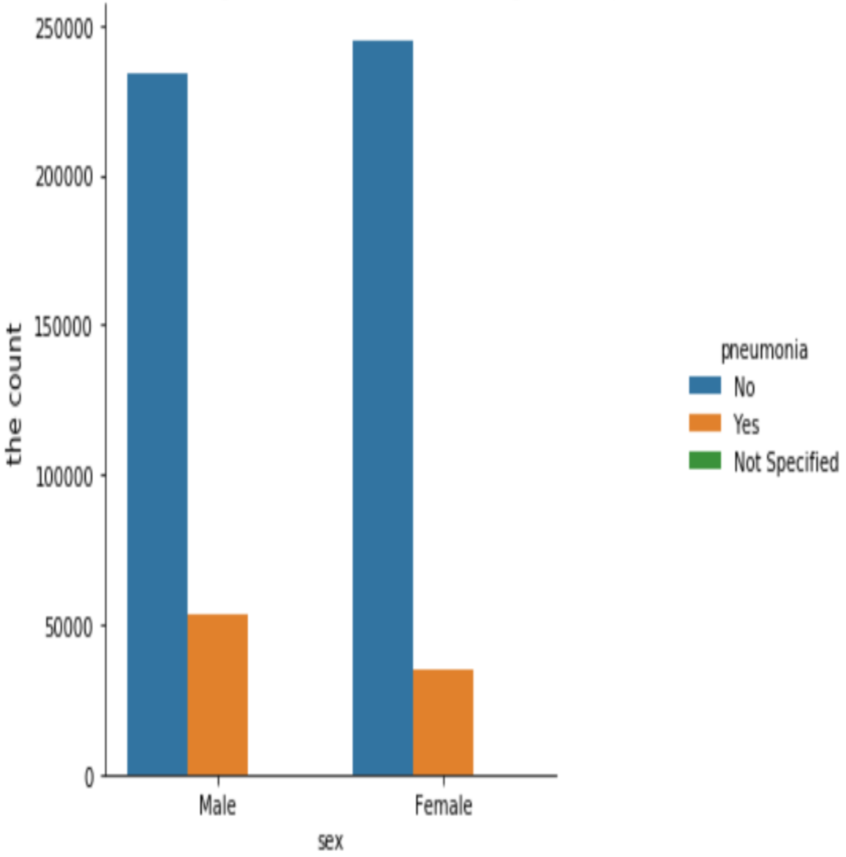
Conclusion: the most patiens who become positive are males, and the negative are females

So, Why Men?!

#Men_Life_Matters :D



The ratio between vlaues for the gender column according to the tobacco column



The ratio between vlaues for the gender column according to the pneumonia column

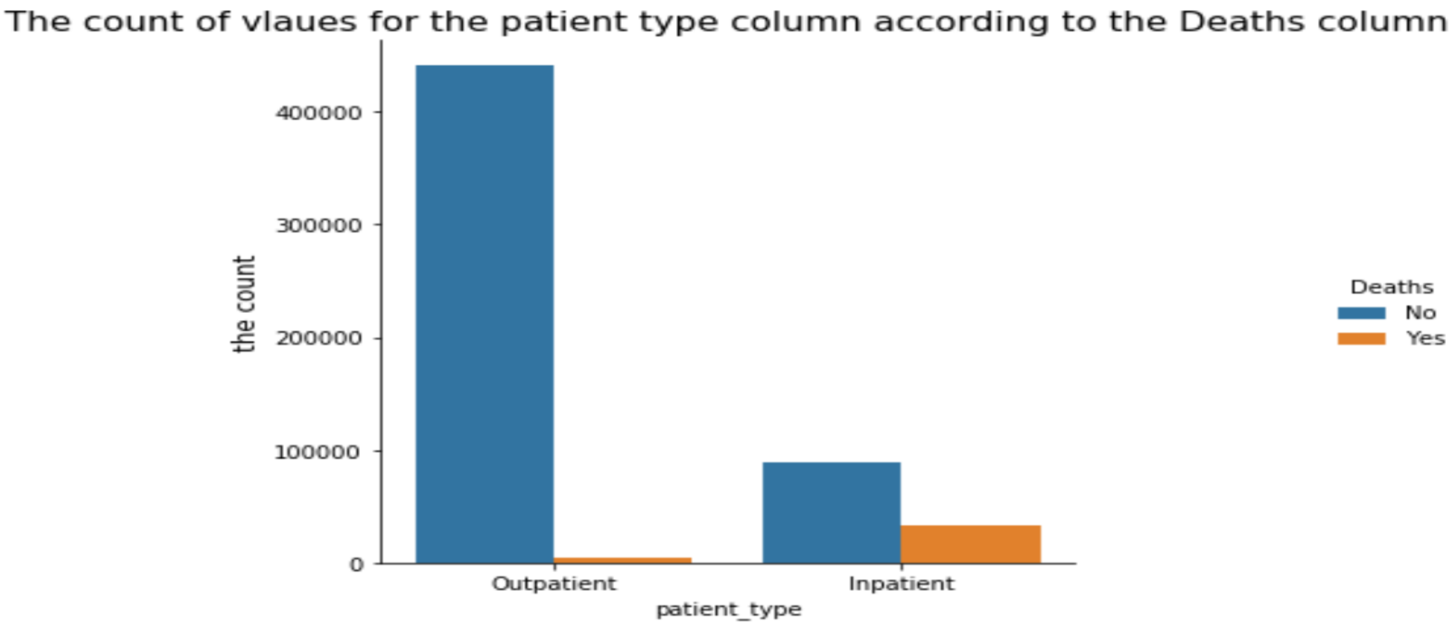
Conclusions: the big ratio of patients who got *pneumonia* and *smoke* are from males

About the pie plots second the Conclusion’s.

most of the patients is **not hospitalized**

_**Outpatient:** 444,689, **Inpatient**: 121,913.

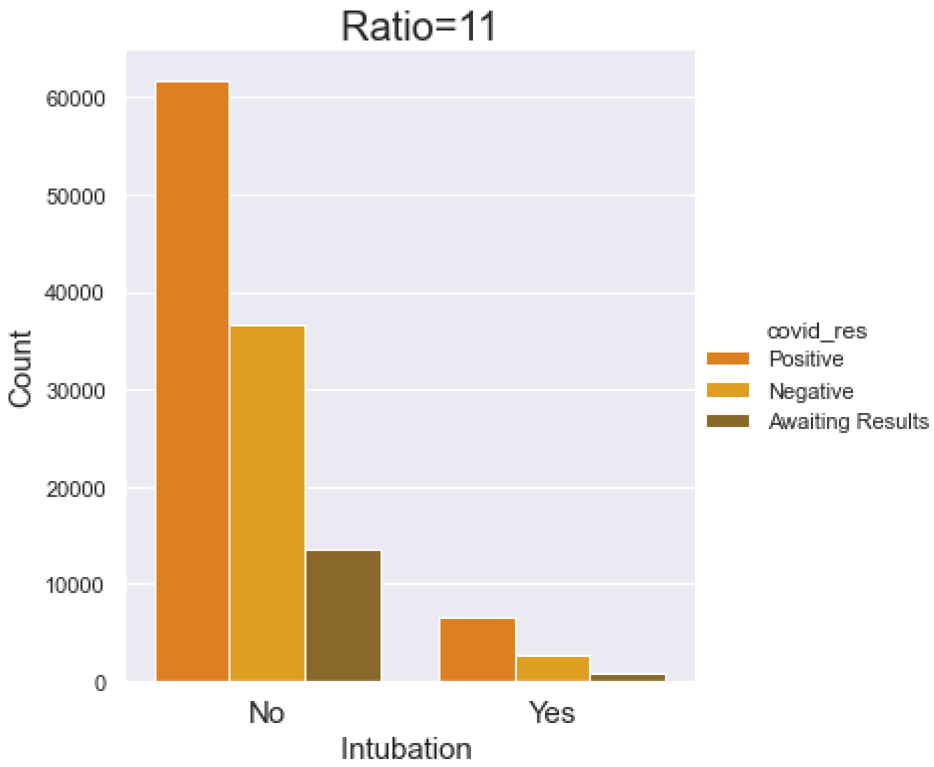
- The ratio of hospitals per thousands: ***1.38***
- This means **for each 100K there’re 138** only available hospitals.
- And the Mexico **total cases for July** only is about: **313,3192.**
- And this means **for this month** we need about **(3 times138)** hospitals.



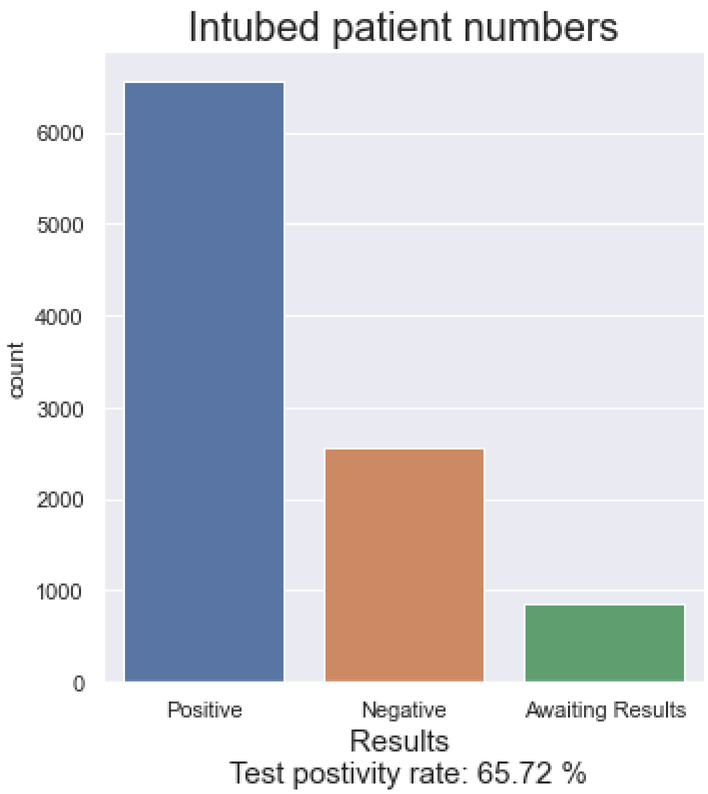
Conclusion: Most of the people who leaves survive
According to the ratio of hospitals per thousands: 1.38. So, Mexico gov gives the priority to the dangerous cases

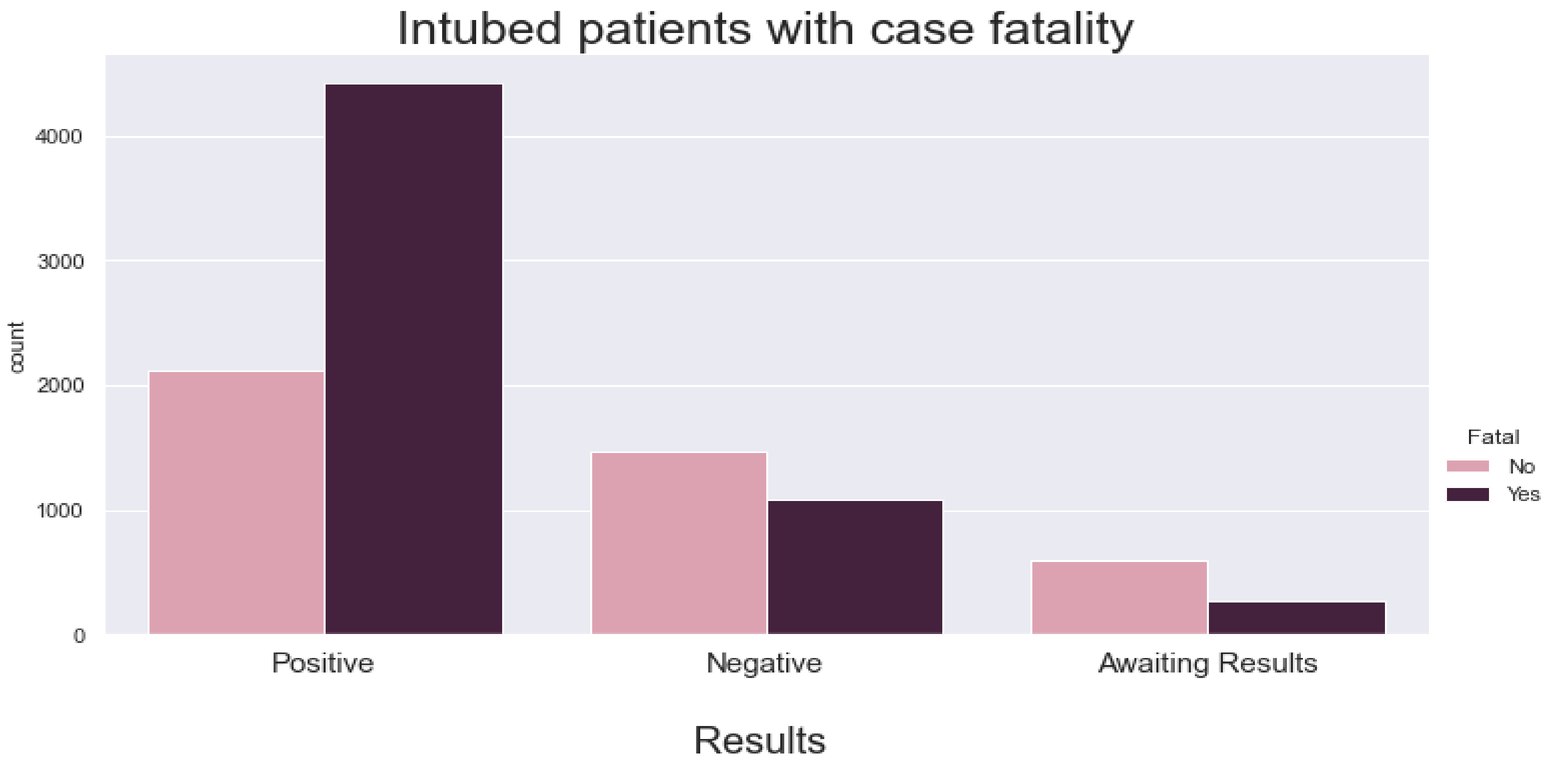
What about intubation ?

Reported intubations



Yes:9965 No:111824

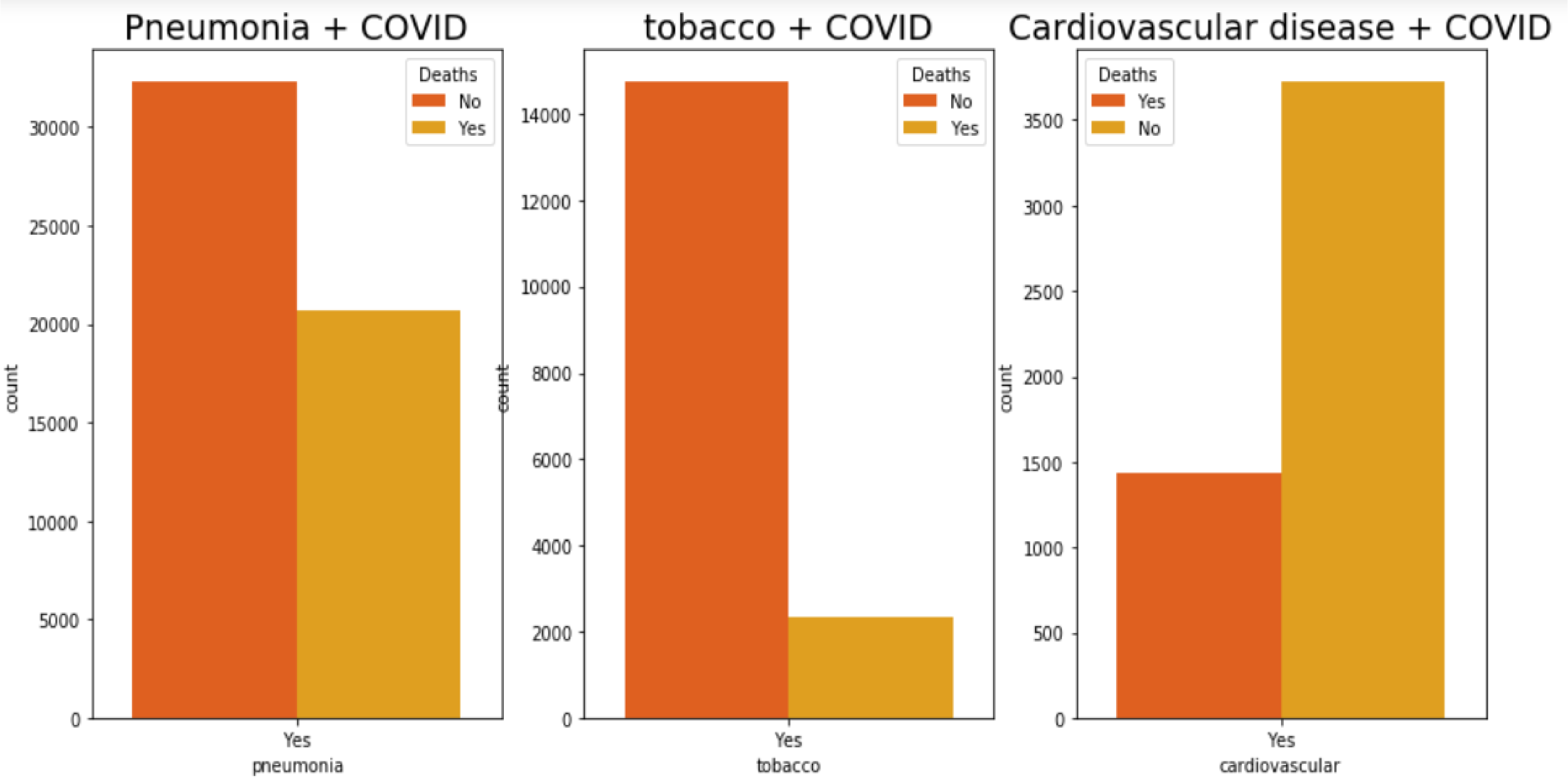




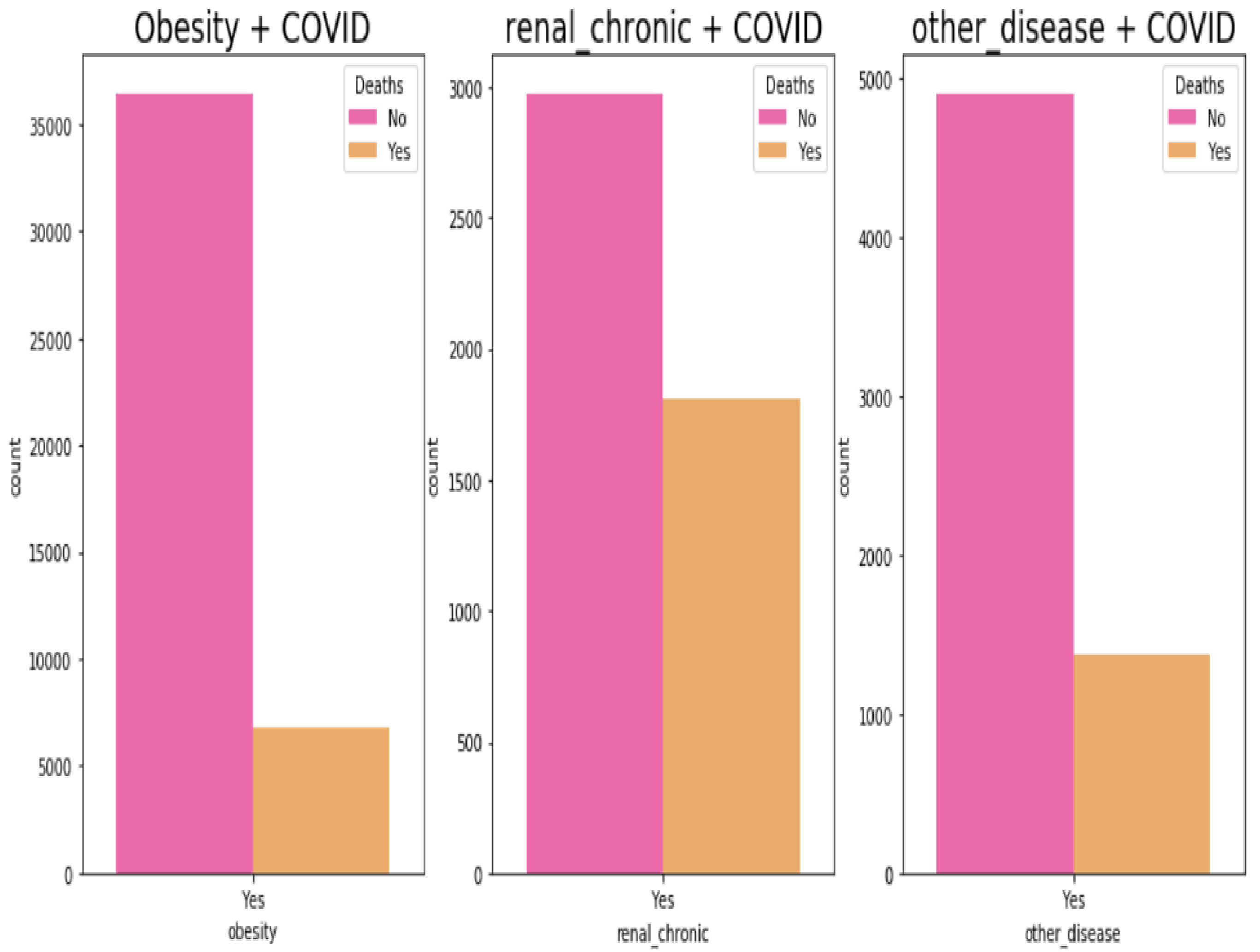
Case Fatality Rate: 58.08 %

COVID +ve Fatality Rate: 67.64 %

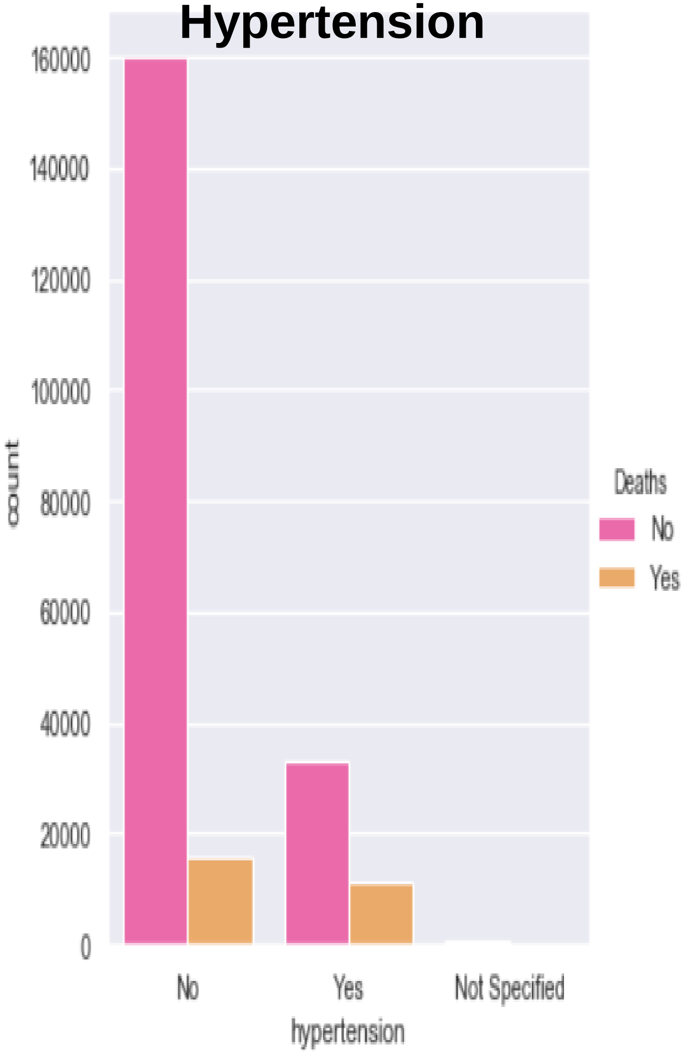
Being Healthy is the 1st Defence wall against Covid.



Conclusion: patient with pneumonia and cardiovascular get big death ratio.

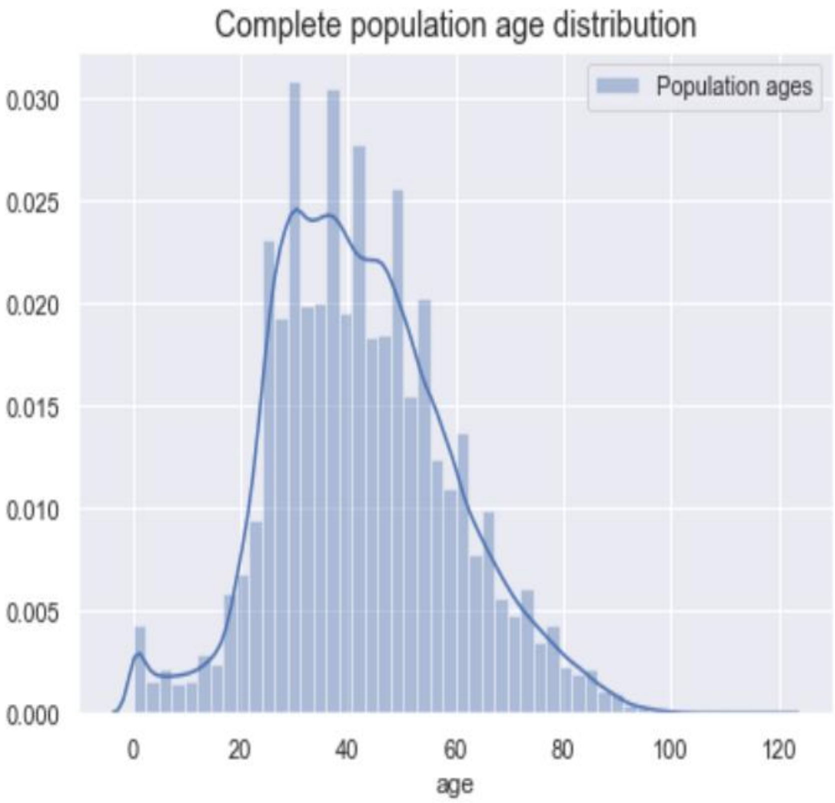


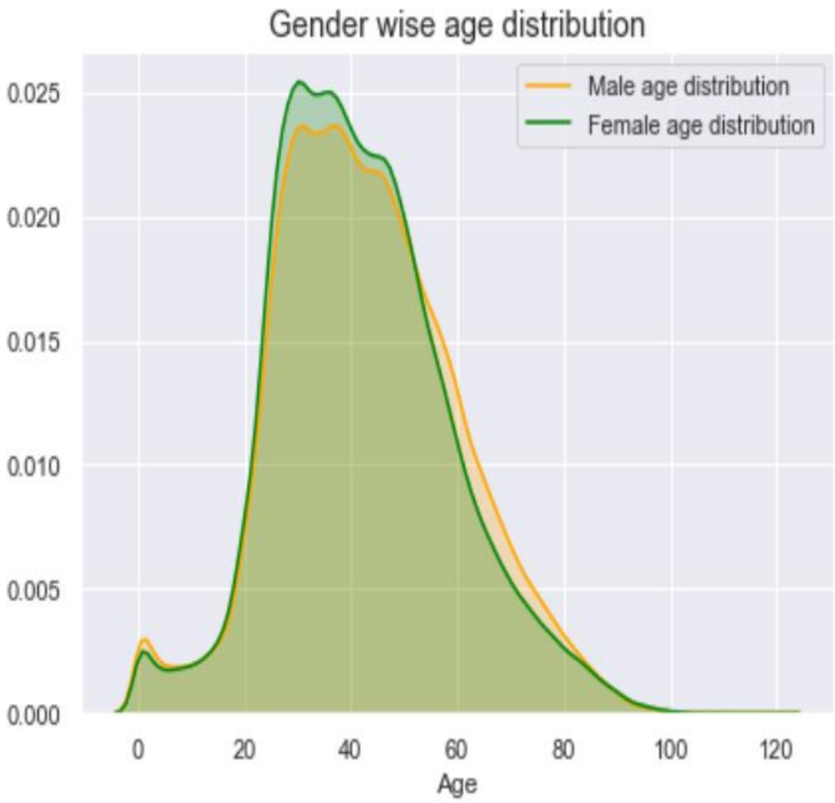
Hypertension



ALL the above diseases are the most common columns which affect our cases

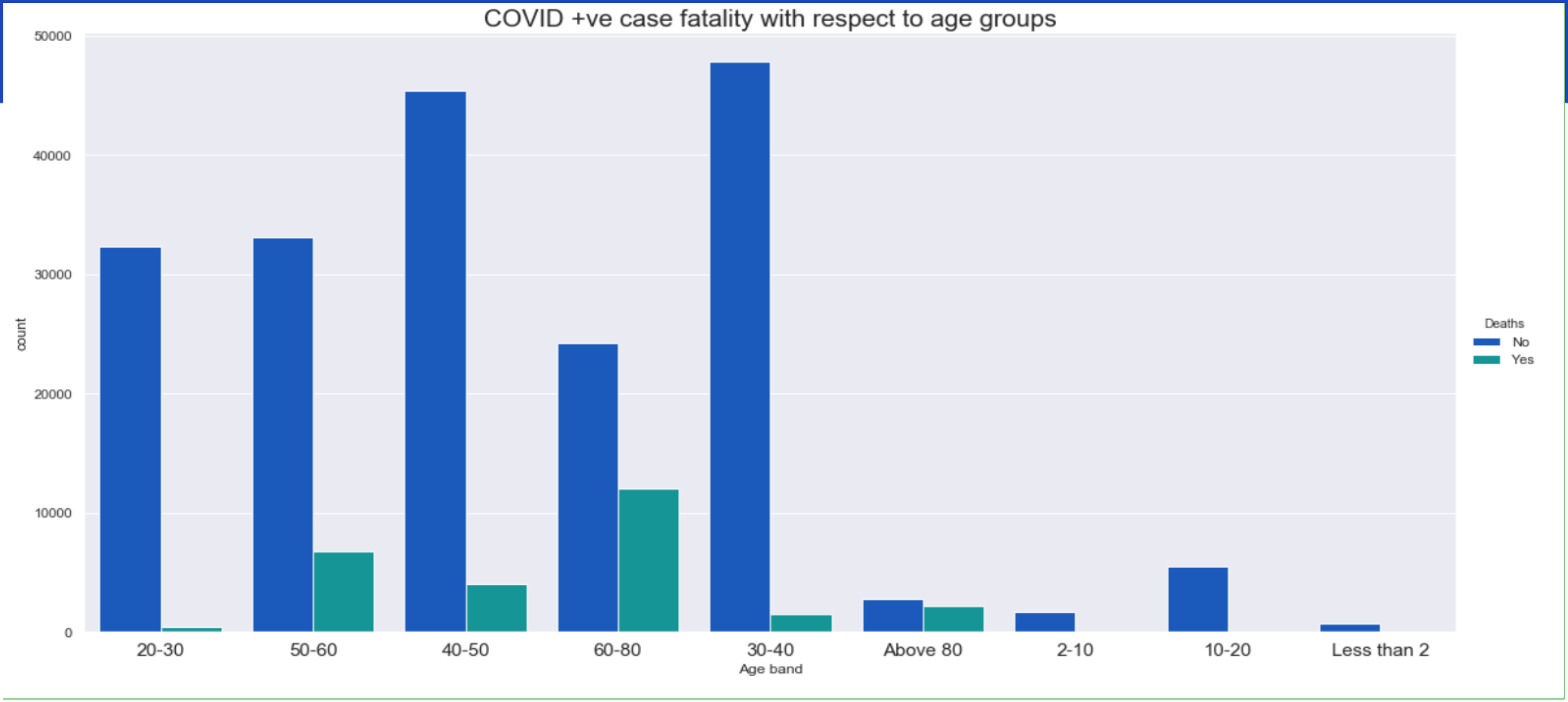
Is really the Age just a number?


Complete population age distribution


Gender wise age distribution

Conclusions: high distribution from 20-60 years, and female age is a little bit greater than male.

Now, we know the distribution. What about the deaths



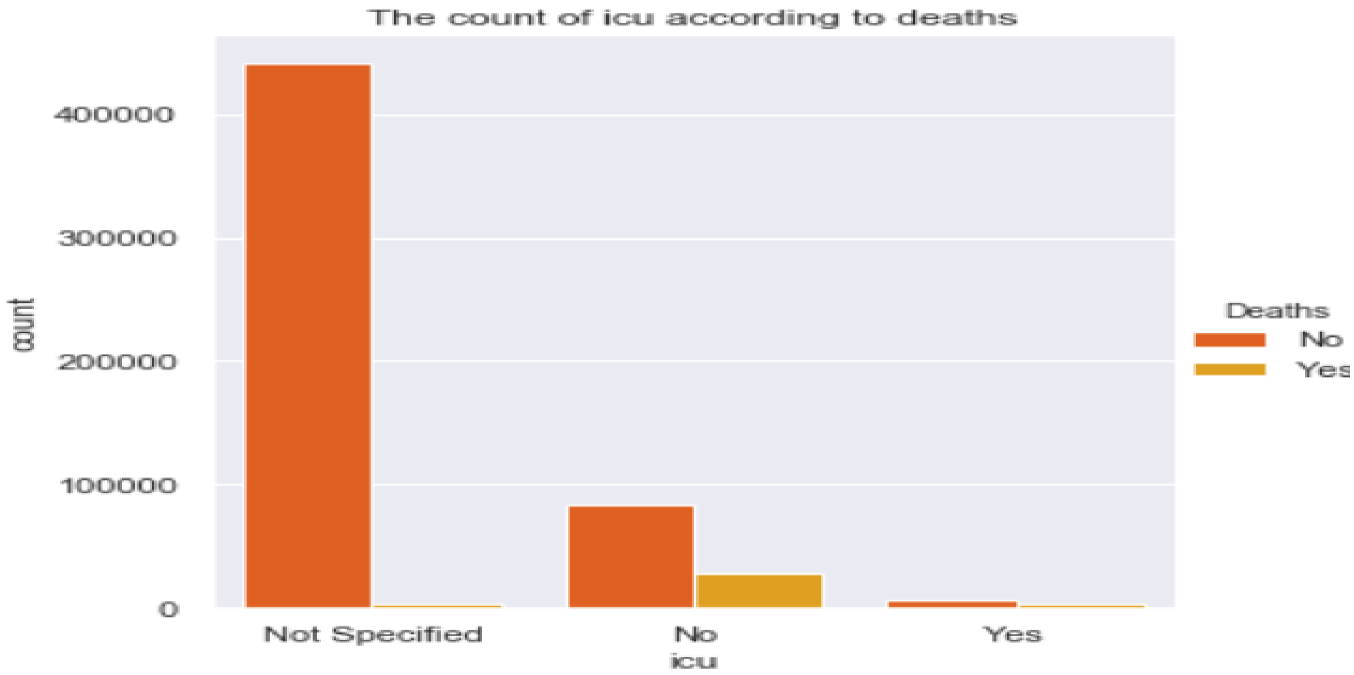
Conclusion: From the above plot, it can be seen that the case fatality is quite high for ages of 60-80 and above 80. This is expected since with body, the immune system becomes weaker and hence, it becomes tough for the body to fight a completely new virus. This is not just true for COVID but for most diseases.

We want to make sure the about the age with +ve and deaths!
# we are 95 confident the mean of the age will be between this intervals: (45.9, 46.54)
# we are 95 confident the mean of the age will be between this intervals: (60.97, 61.5)

ICUs, Clearly Explained

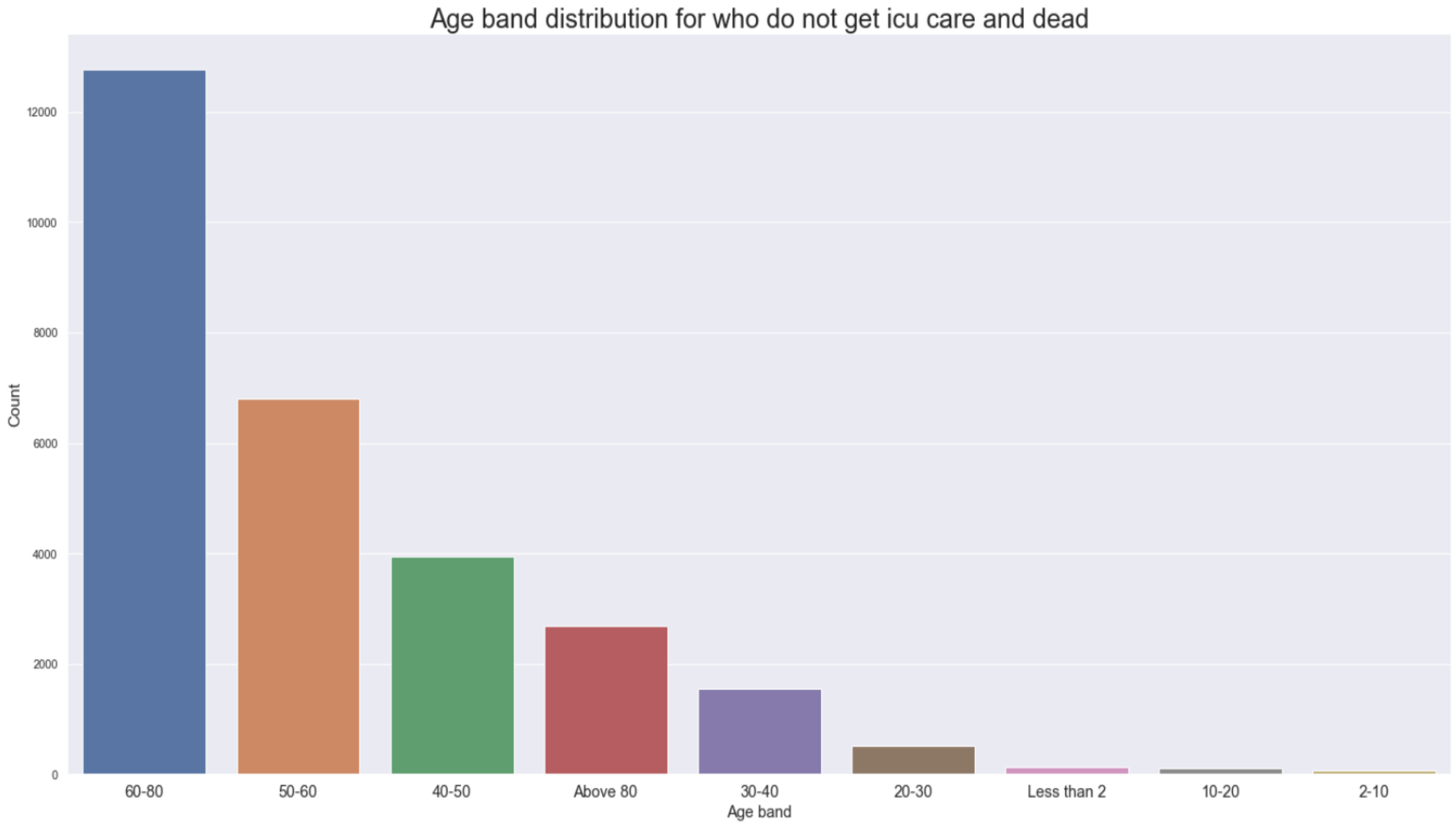
The value counts:

Not Specified:	444814
No	111676
Yes	10112



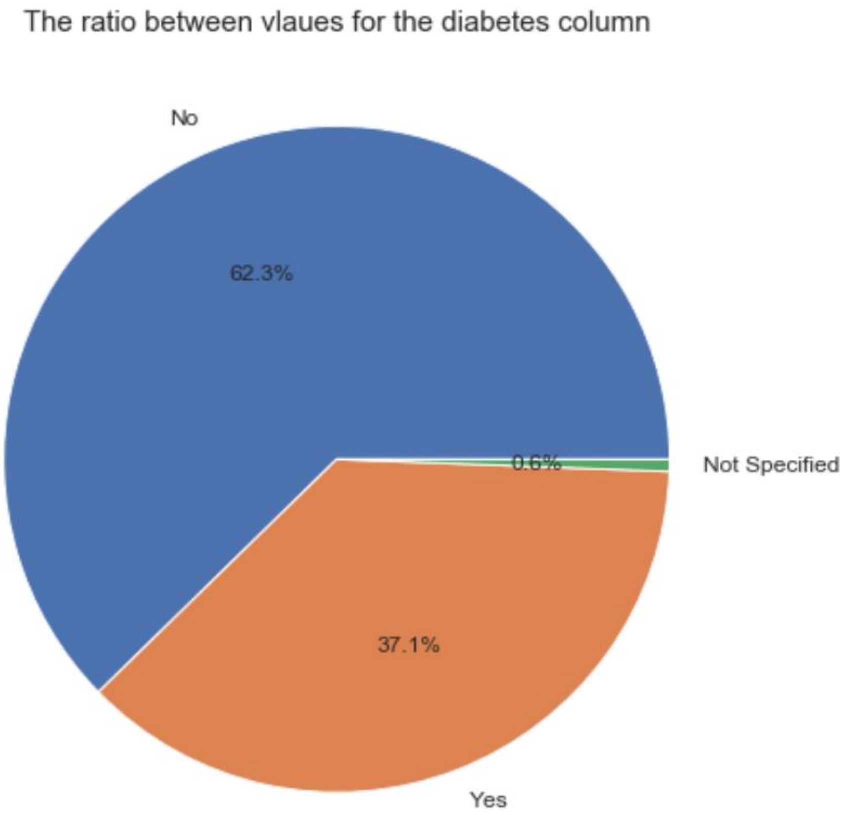
the big ratio of people who didn't put on ICU is died

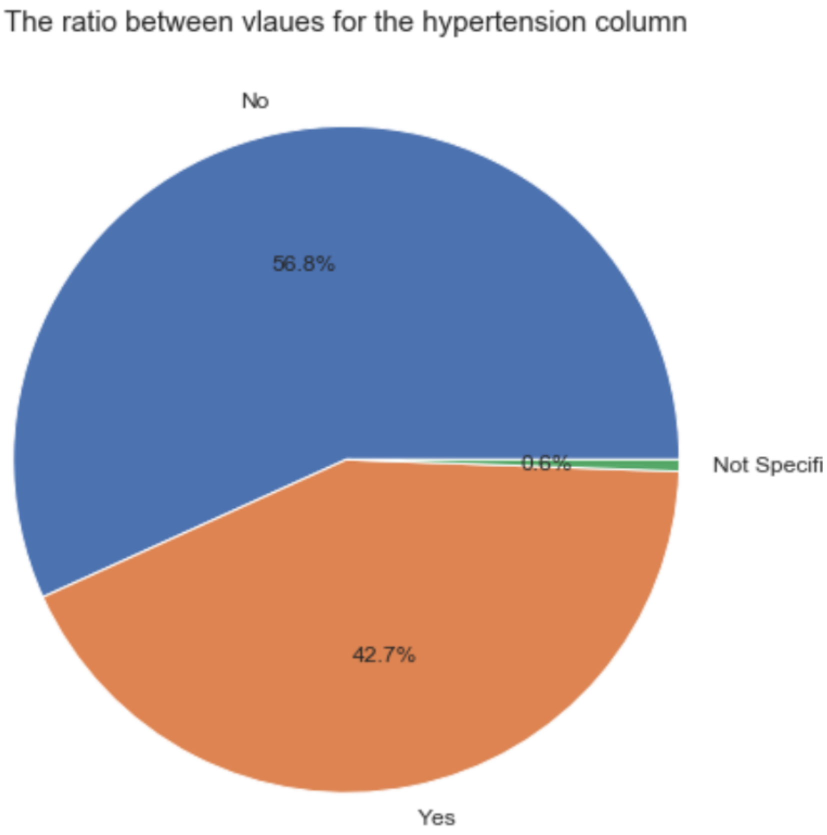
why this happened ?



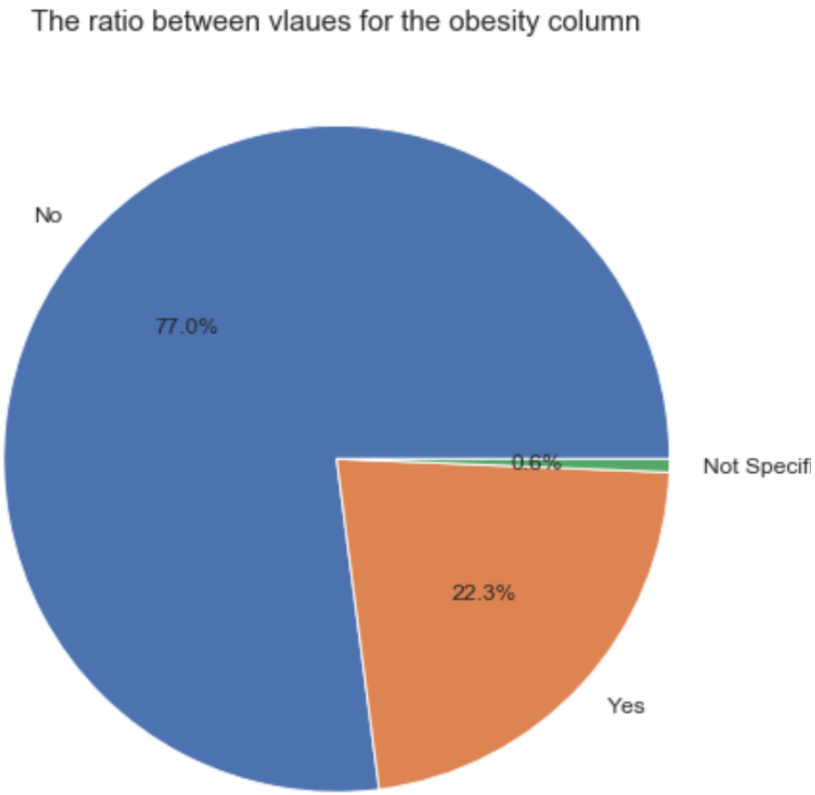
Most of the people who dead are from 60:80, and 50:60

For People who didn’t get ICU care and dead

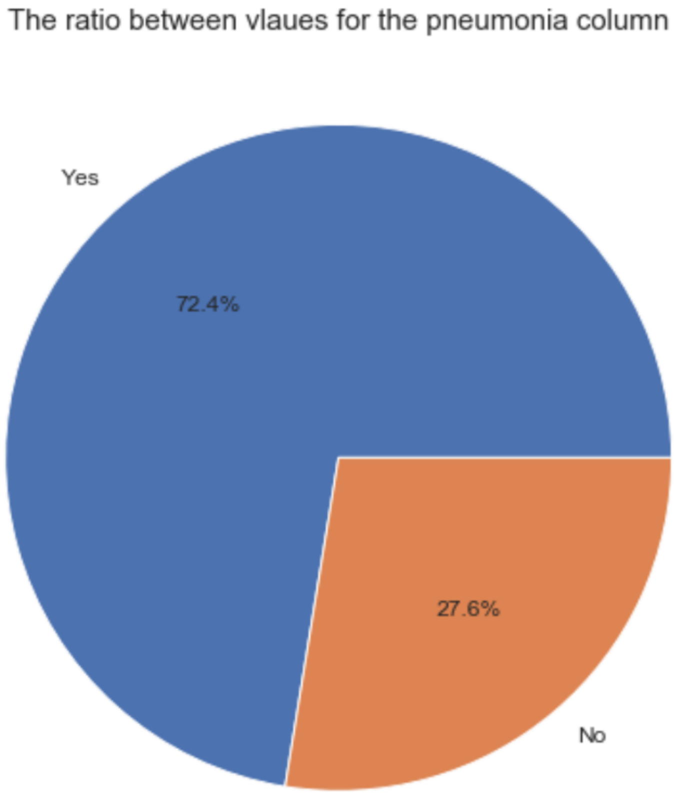




The ratio between vlaues for the obesity column

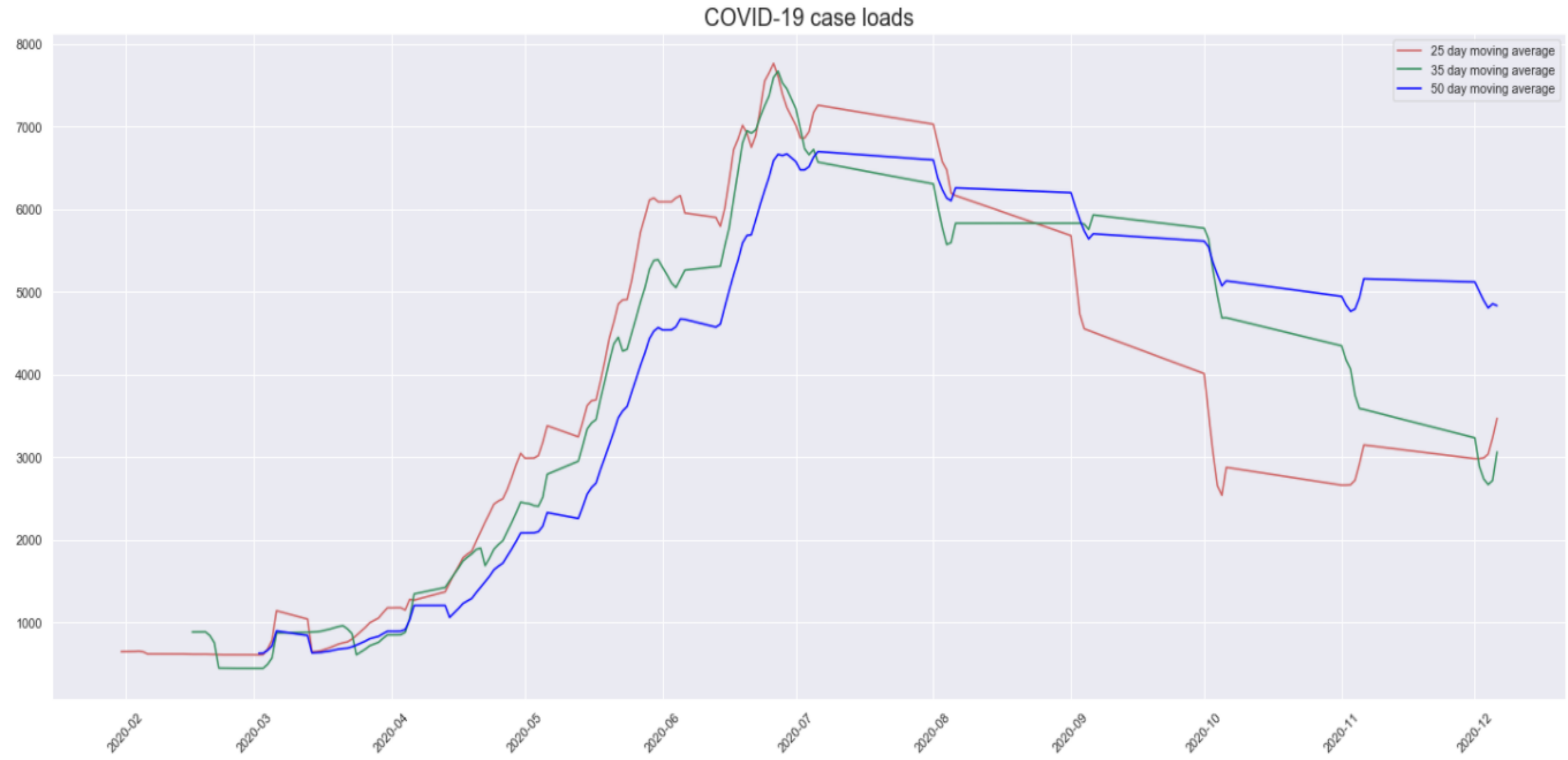


The ratio between vlaues for the pneumonia column

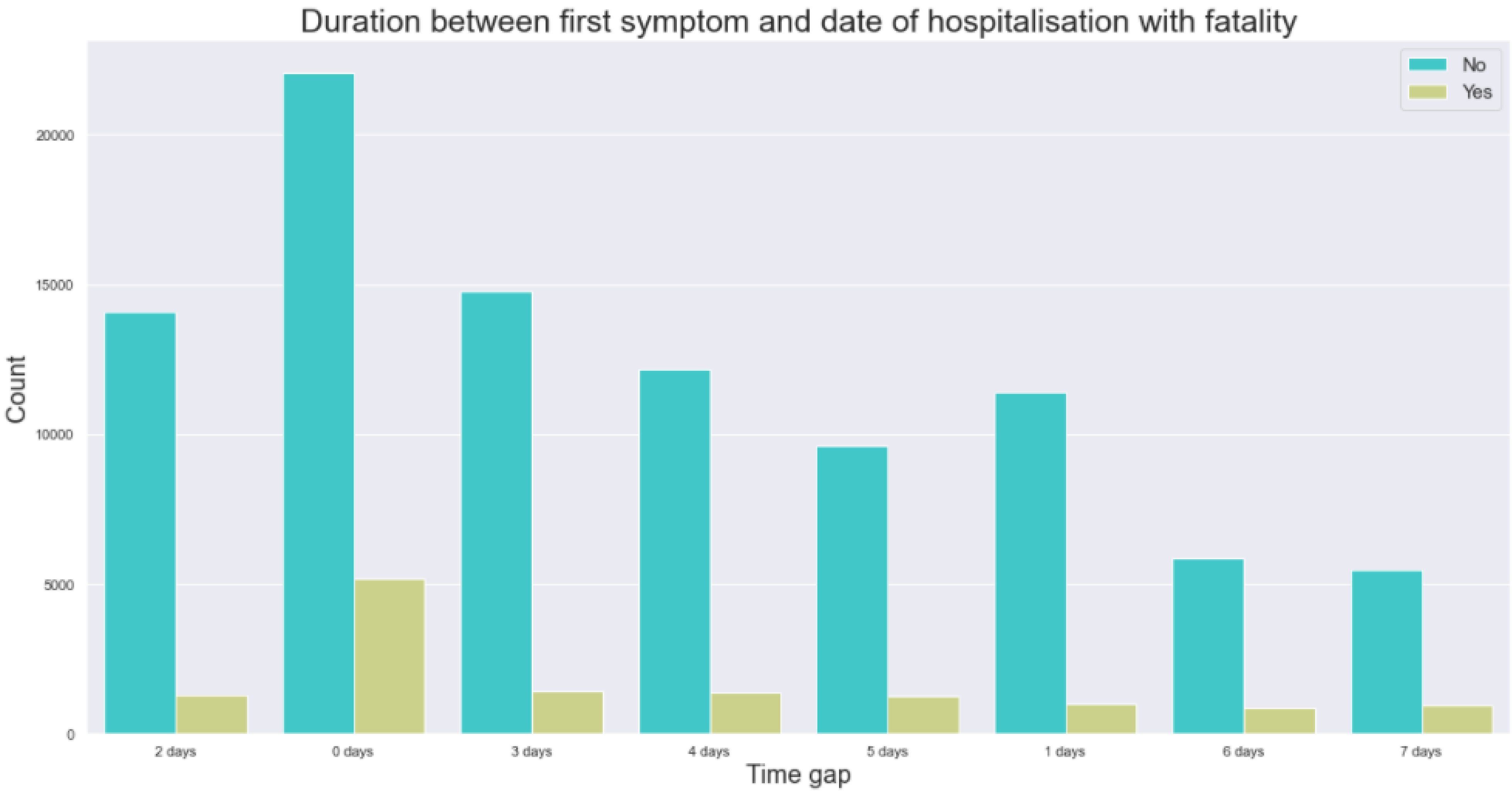


Conclusion: All the last 4 daises is getting the most affects on the patient who didn’t enter the icu and dead.

Let's Crack our time



Conclusion: form the 05_2020: 07_2020 there are a booming in the data


Duration between first symptom and date of hospitalisation with fatality
No
Yes
Count
0
5000
10000
15000
20000
2 days
0 days
3 days
4 days
5 days
1 days
6 days
7 days
Time gap

What about the booming which happened between May and July

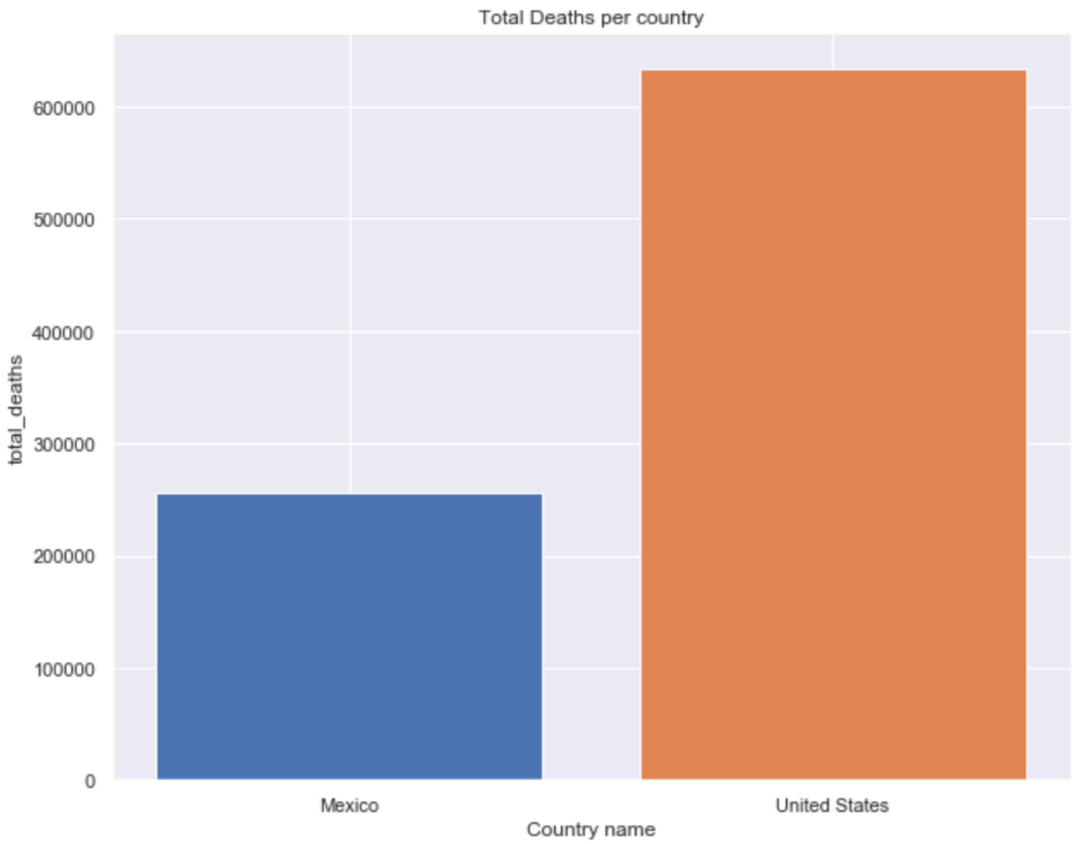
To see the affects of diseases columns on this booming df:

1. the tobacco, pneumonia, obesity, diabetes, and hypertension are the most diseases affects our deaths
2. The most distributed age with +ve case are between 30_40
3. The most distributed age who dead are between 50:60

Let’s evaluate the Mexico medical system by comparing others.

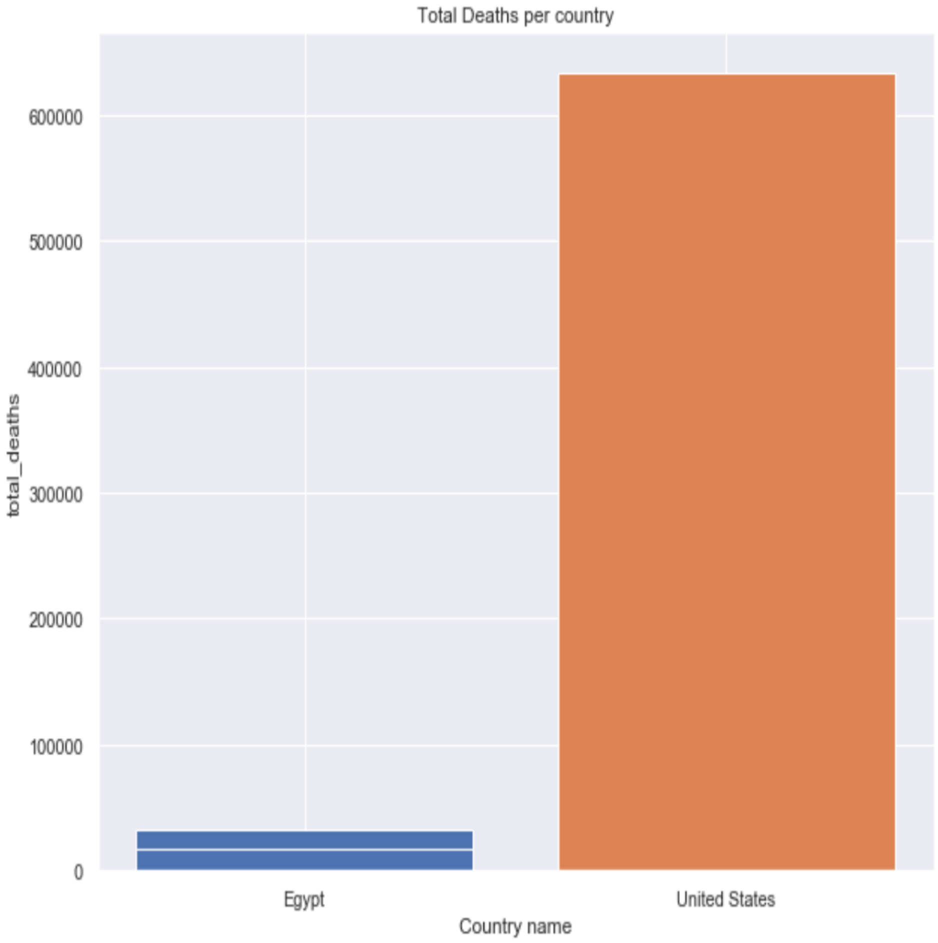
The Global Data Link

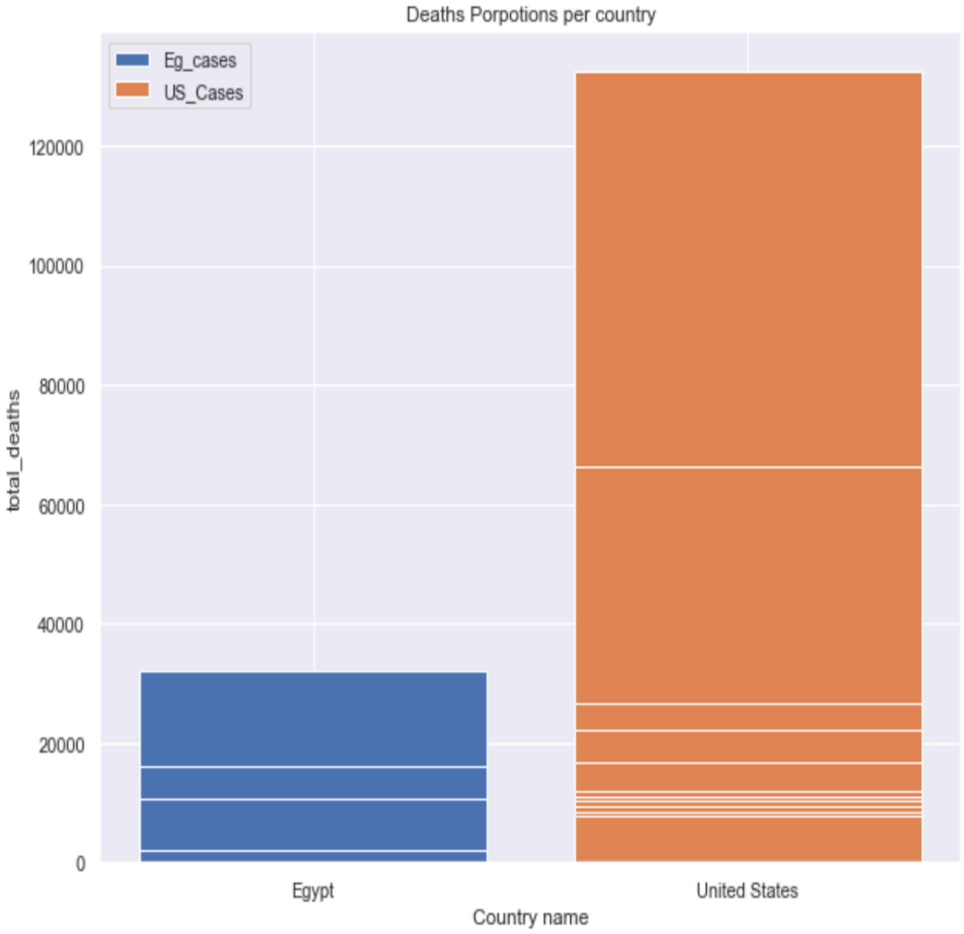
Let’s start by comparing the total deaths



It’s a little bit tricky

Let’s see this example to explain

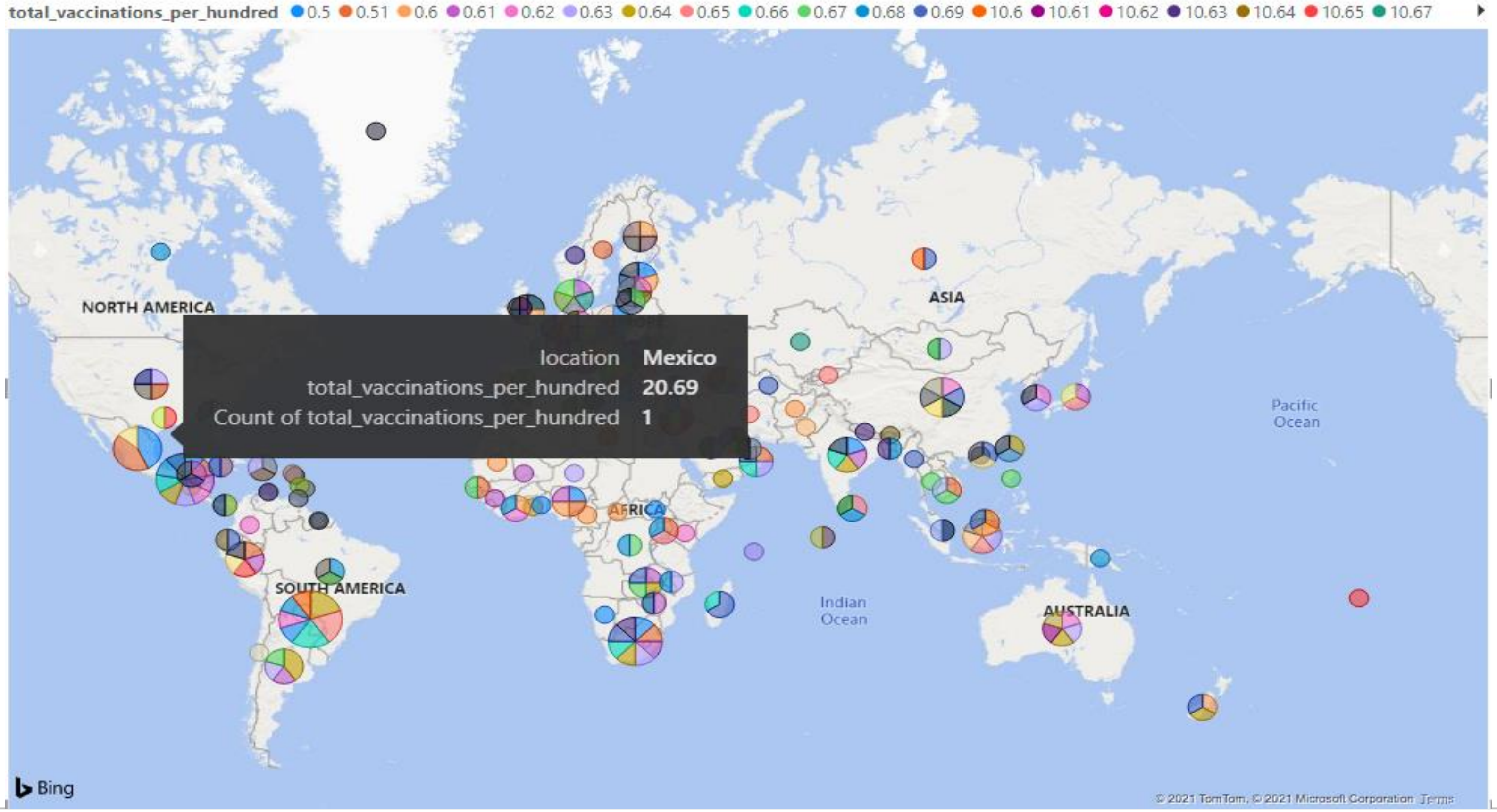

Total Deaths per country


Deaths Porpotions per country

The Death **Proportions** Indicates **the total death / total cases**

Mexico **Death Proportion’s** index: **19th largest index**

About **Vaccinations**


total_vaccinations_per_hundred ● 0.5 ● 0.51 ● 0.6 ● 0.61 ● 0.62 ● 0.63 ● 0.64 ● 0.65 ● 0.66 ● 0.67 ● 0.68 ● 0.69 ● 10.6 ● 10.61 ● 10.62 ● 10.63 ● 10.64 ● 10.65 ● 10.67
NORTH AMERICA
ASIA
AFRICA
SOUTH AMERICA
Pacific Ocean
Indian Ocean
AUSTRALIA
Bing
© 2021 TomTom, © 2021 Microsoft Corporation Terms

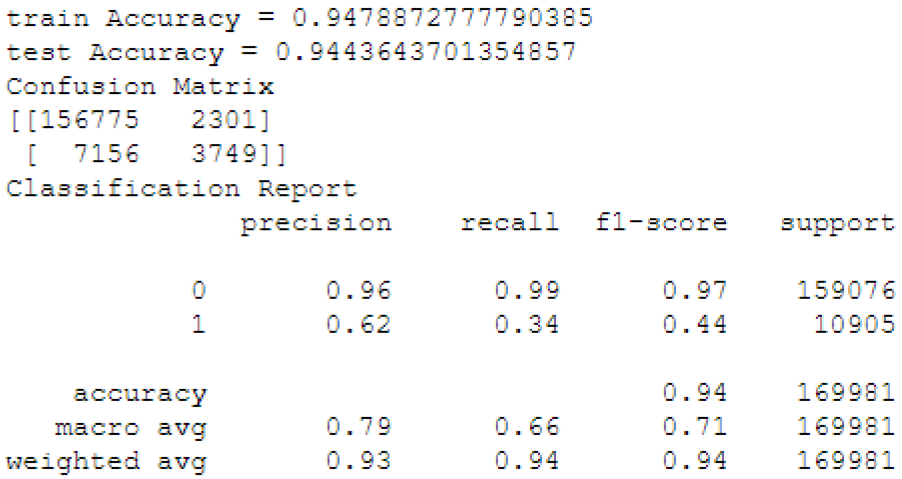
location Mexico
total_vaccinations_per_hundred 20.69
Count of total_vaccinations_per_hundred 1

The classification algorithm we used

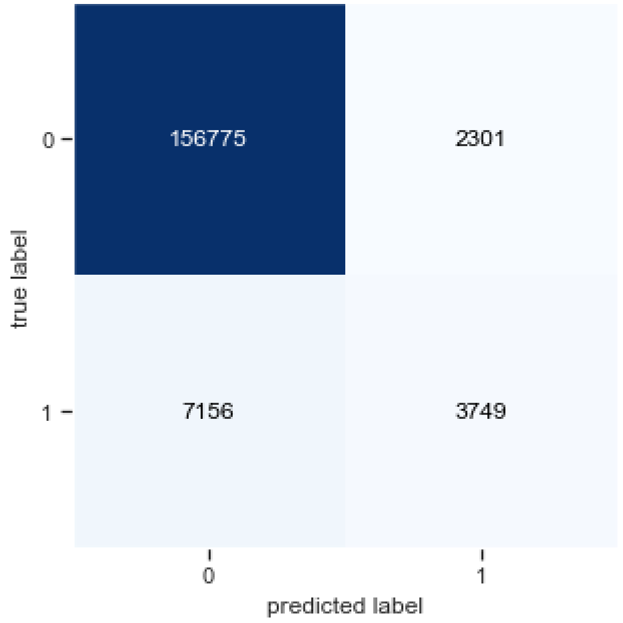
1)Random forest
2) **Naive Bayes Algorithm**
3) **xgboost**
4) **adaboosted**
5) Decision Tree Classifier

which is the highest accuracy ?

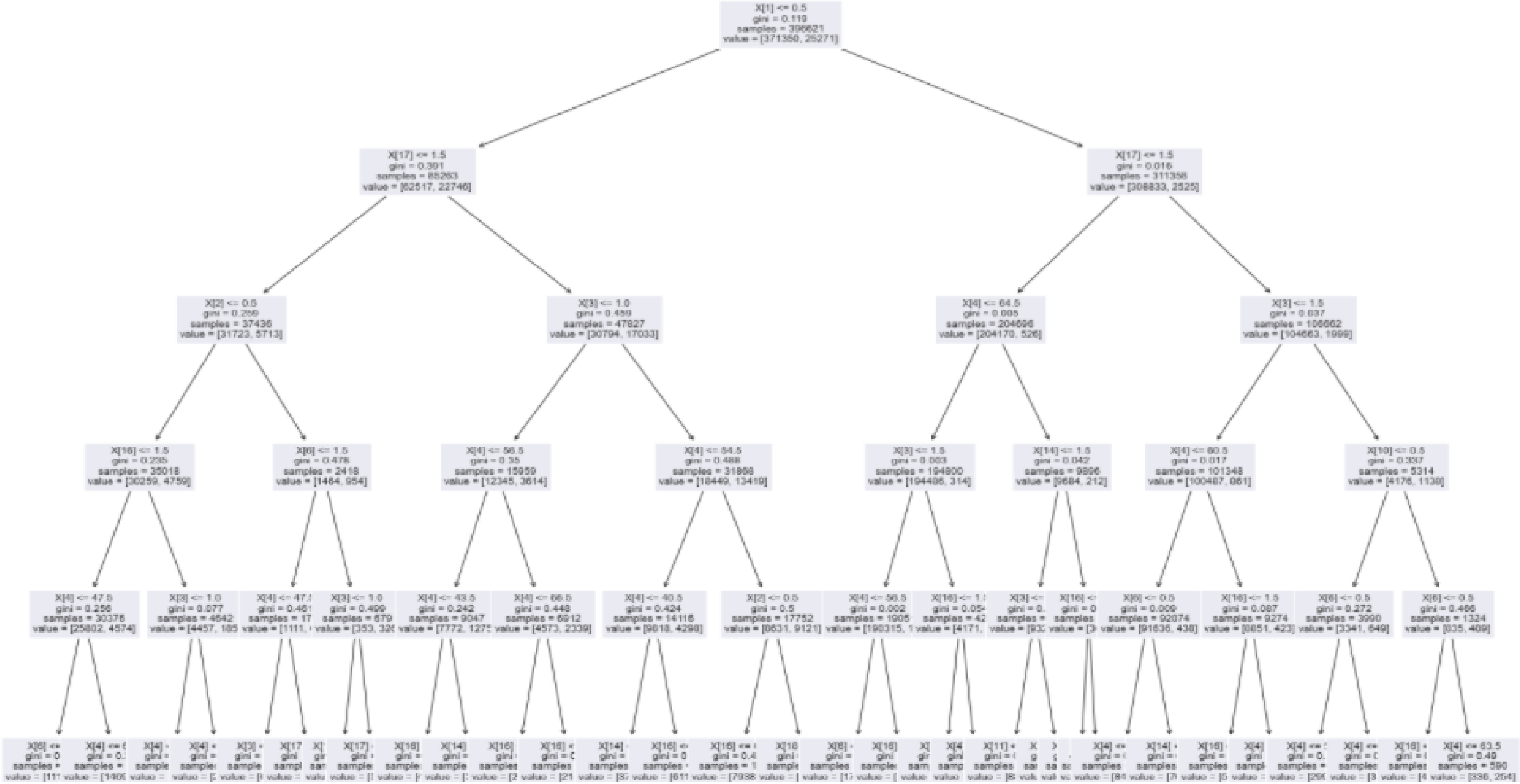
XGBOOST and Decision tree is the highest accuracy
XGBOOST :



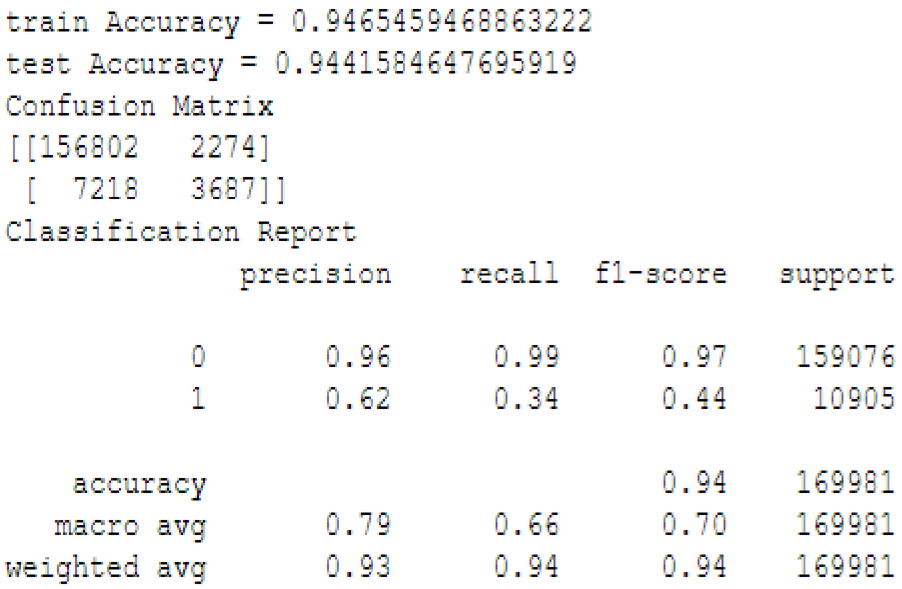
```
train Accuracy = 0.9478872777790385
test Accuracy = 0.9443643701354857
Confusion Matrix
[[156775   2301]
 [  7156   3749]]
Classification Report
              precision    recall  f1-score   support

           0       0.96      0.99      0.97    159076
           1       0.62      0.34      0.44     10905

    accuracy                           0.94    169981
   macro avg       0.79      0.66      0.71    169981
weighted avg       0.93      0.94      0.94    169981
```



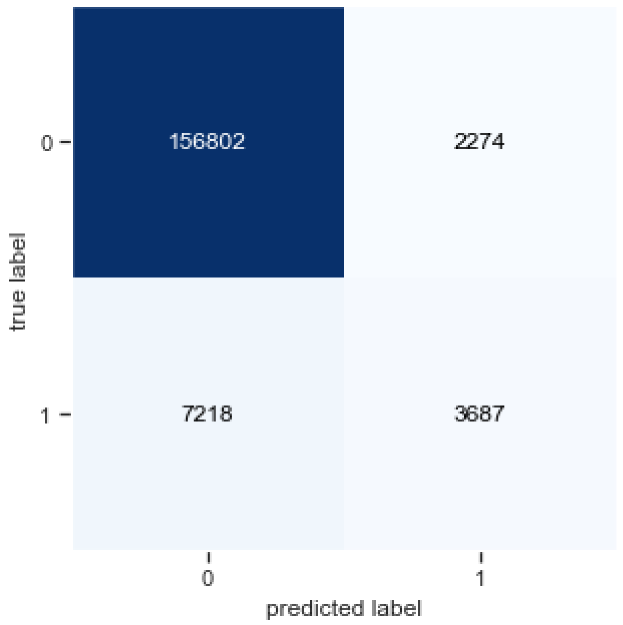
Decision tree



Decision tree



```
train Accuracy = 0.9465459468863222
test Accuracy = 0.9441584647695919
Confusion Matrix
[[156802   2274]
 [  7218   3687]]
Classification Report
              precision    recall  f1-score   support

           0       0.96      0.99      0.97    159076
           1       0.62      0.34      0.44     10905

    accuracy                           0.94    169981
   macro avg       0.79      0.66      0.70    169981
weighted avg       0.93      0.94      0.94    169981
```



SAMSUNG

Together for Tomorrow!

Enabling People

Education for Future Generations