# EM-626

# Mid Term Project

AJAY RANA

# Introduction

On Jan 30, 2020, WHO declared a public health Emergency of International Concern, a month after COVID-19 was identified in Wuhan, China. By this point, several mathematical and computational models have already raised the alarm about the potential for severe acute respiratory syndrome.

During the emergency of a novel pandemic, predictive modelling is important in public health planning and response. Relating models to data provides a view into the unseen variable, such as the occurrence of cryptic transmission and the prevalence of infection, and these predictive models allow exploration of counterfactuals and hypothetical interventions.

The main of these models is to support health systems in with COVID-19 strategic decision making, planning, and health policy formulation that help in the fight against COVID-19.

# Data Understanding and Preparation

Dataset is taken from various sources and combined in one file. The dataset includes the number of confirmed cases and number of deaths in all the US counties. The no. of cases in each county is in cumulative form and not of an individual day basis.

The other dataset consists of Annual Estimates of Housing Units for the Counties in the US in 2019. Dataset also consist GDP per each county in the US in 2018 and finally, we have dataset related to the population distribution of each county according to Sex, Age and Race in the US in 2019

A dataset which was used for the predictive modelling was a combination of the all the above dataset and to reduce the complexity and processing of the data many variables have been removed.

The final dataset consists of 22 Columns x 3143 Rows. The variables in the final dataset include total No. of cases in each county which was the value for the most recent date among each county and rest of the date was removed and similarly the total no. of deaths in each county was considered. Total no. of the population in each county, GDP per each county, Housing, Age group were combined to form few age groups and only 4 races were taken with the alone population.

| | State | county | County_state | ... | NA_TOT | TOT_MALE | TOT_FEMALE |
|---|---|---|---|---|---|---|---|
| 0 | AL | Autauga County | Autauga County,AL | ... | 58 | 27092 | 28777 |
| 1 | AL | Baldwin County | Baldwin County,AL | ... | 154 | 108247 | 114987 |
| 2 | AL | Barbour County | Barbour County,AL | ... | 52 | 13064 | 11622 |
| 3 | AL | Bibb County | Bibb County,AL | ... | 26 | 11929 | 10465 |
| 4 | AL | Blount County | Blount County,AL | ... | 67 | 28472 | 29354 |
| 5 | AL | Bullock County | Bullock County,AL | ... | 77 | 5508 | 4593 |
| 6 | AL | Butler County | Butler County,AL | ... | 8 | 9024 | 10424 |
| 7 | AL | Calhoun County | Calhoun County,AL | ... | 127 | 54481 | 59124 |
| 8 | AL | Chambers County | Chambers County,AL | ... | 15 | 15895 | 17359 |
| 9 | AL | Cherokee County | Cherokee County,AL | ... | 8 | 12926 | 13270 |

The Below Output shows that most of the variables are of int and float type and only 3 variables are of the object type which includes State, county and County_state. There are no-Null values as most of them have been replaced with the average age or zero to reduce the errors and for better modelling.

```
RangeIndex: 3142 entries, 0 to 3141
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   State             3142 non-null   object
 1   county            3142 non-null   object
 2   County_state      3142 non-null   object
 3   population        3142 non-null   int64
 4   gdp               3142 non-null   float64
 5   housing           3142 non-null   int64
 6   Cases             3142 non-null   int64
 7   Deaths            3142 non-null   int64
 8   Outcome           3142 non-null   int64
 9   per_age_below_18  3142 non-null   float64
 10  per_age_18_24     3142 non-null   float64
 11  per_age_25_44     3142 non-null   float64
 12  per_age_45_64     3142 non-null   float64
 13  per_age_65_plus   3142 non-null   float64
 14  WA_TOT            3142 non-null   int64
 15  BA_TOT            3142 non-null   int64
 16  IA_TOT            3142 non-null   int64
 17  AA_TOT            3142 non-null   int64
 18  NA_TOT            3142 non-null   int64
 19  TOT_MALE          3142 non-null   int64
 20  TOT_FEMALE        3142 non-null   int64
dtypes: float64(6), int64(12), object(3)
```

Setting the target variable for a predictive model which is the most important part of the modelling. The most suitable variables for the target variable was total no. of the population in each county and no. of cases. As it is more likely that in more populated area COVID-19 is can easily spread if necessary precaution is not taken.

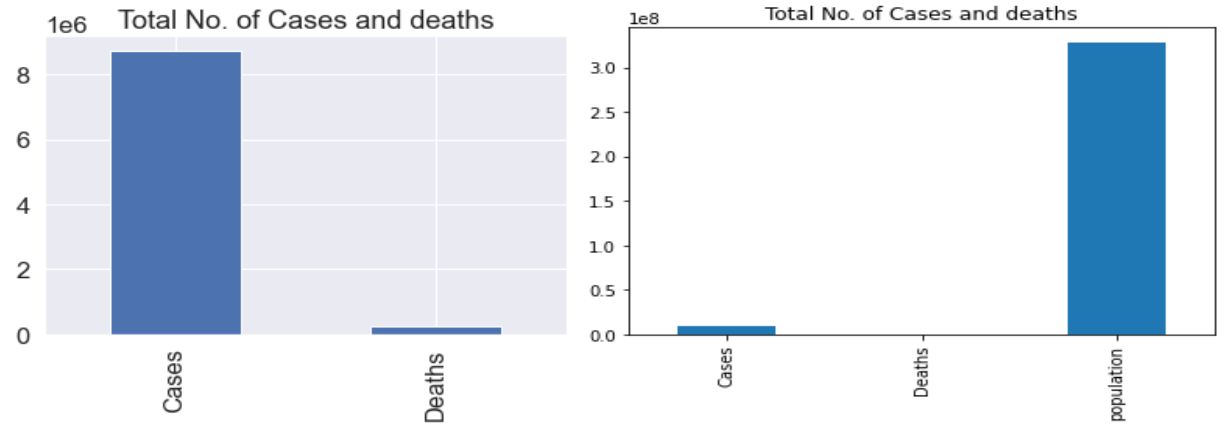
The ratio of cases/population was taken and a threshold of 0.026 (which is the mean of the ratio) was set if the value < 0.26 it is set to 0 else 1. If a person has COVID-19 it was set 1 else 0 and this was set as Outcome variable.

```
count    3142.000000
mean        0.026871
std         0.016986
min         0.000000
25%         0.015003
50%         0.024070
75%         0.035402
max         0.179130
Name: target, dtype: float64
Number of Counties: 3142
```

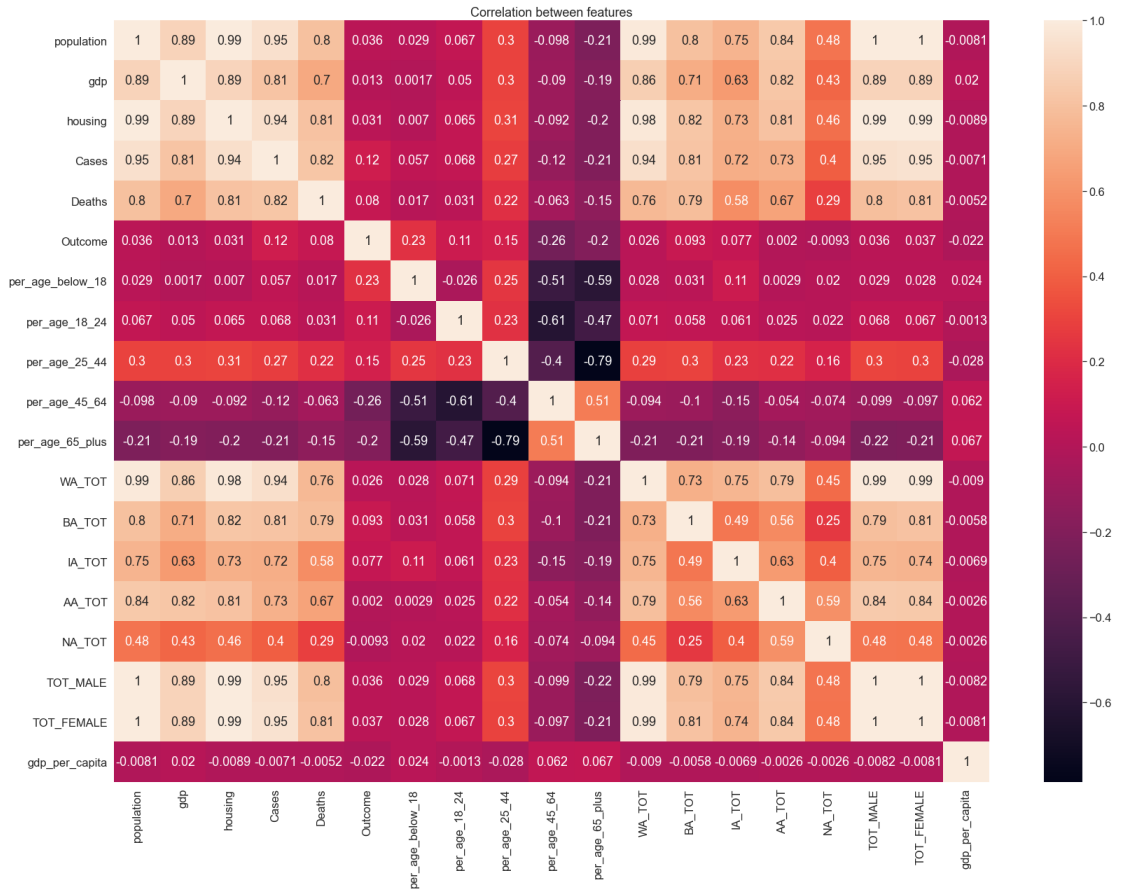
## Data Exploration

The target variable is set, the relations between different variables and correlations between each of the variable is checked.

Below is the bar plot between the total number of cases and deaths of all the counts together. The number of deaths (2,24,807) is high and less compared to the number of confirmed cases (87,33,926).

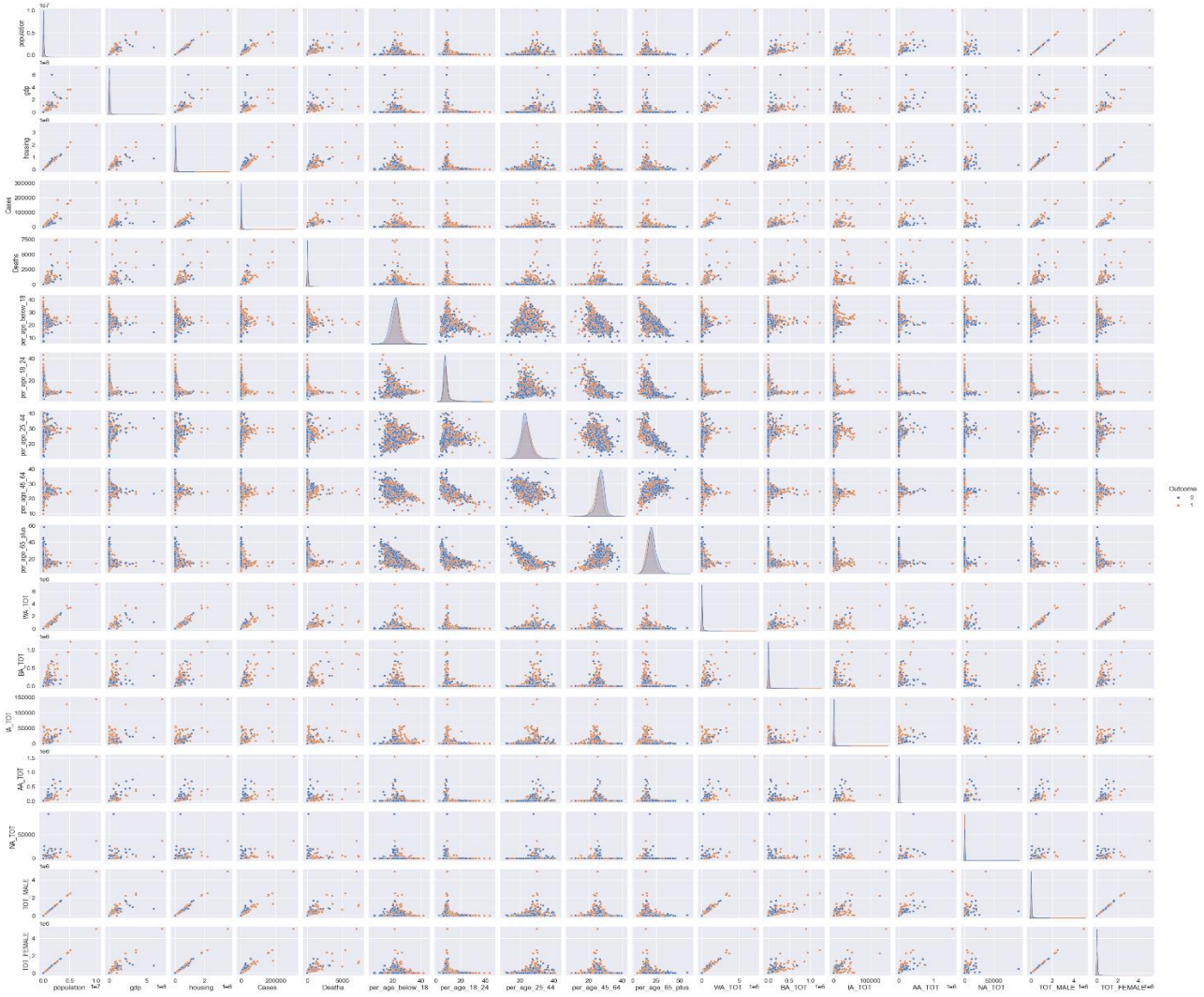



To check on what all variables our target variable is dependent. The correlation matrix gives us more insight into the relationship between each of the variables and how they are affected by each other.

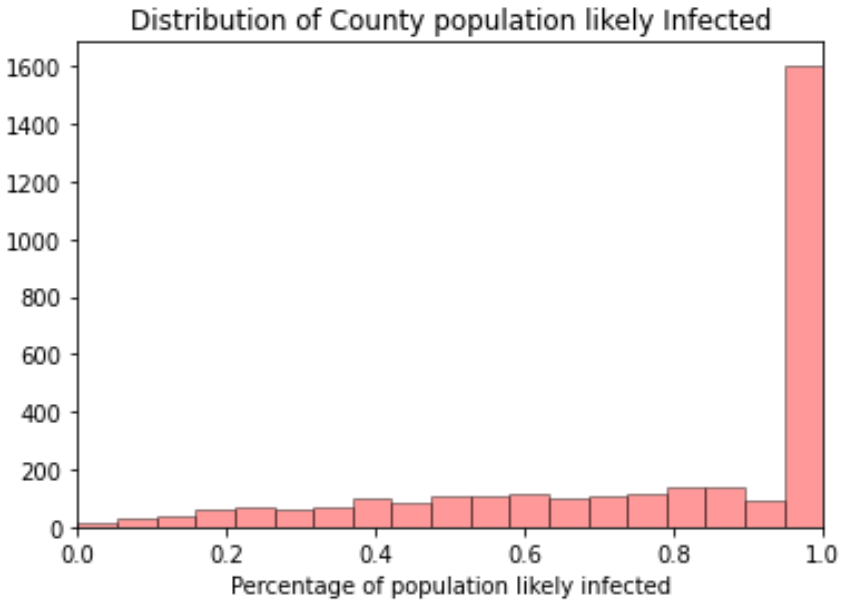


The target variable Outcome is positively correlated to the percentage of the population in each county which are below the age group of 18 and 18-24 and it is negatively correlated to the age group which lies between 45-64 and age group above 65. It is also lightly correlated to the population in each county, housing and among races it highly related to total Black African American(BA_TOT), American Indian and Alaskan Native(IA_TOT), and White American(WA_TOT).

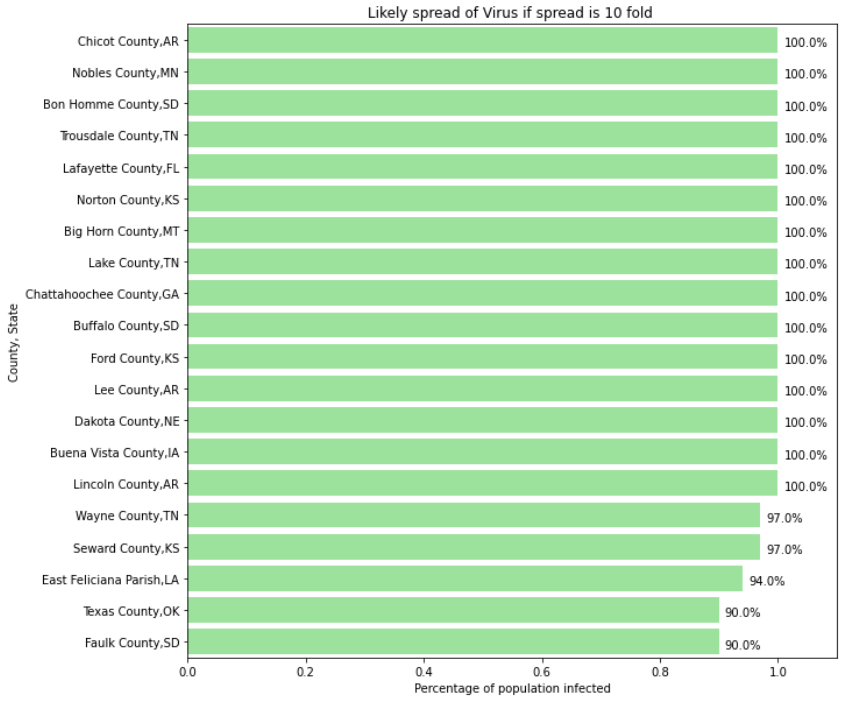
Pair Plot also helps in analyzing the relationship between the two variables like correlation matrix. From the pair plot, it can be seen that most of the variables which were highly correlated in the correlation matrix. They have do not have an overlapping histogram which shows that they are related to the target variable. for example, the percentage of the population in the age group 45-65 do not have an overlapping histogram and this also confirms a high negative value in the correlation matrix.



To check if the number of people likely infected is between 10 to 40 time the number of confirmed cases soon which all counties will be most likely be affected the most. The chances of high likely infected were calculated by finding the ratio of cases to population and multiplied by 40 and low with 10 and then rounded to 0 or 1.



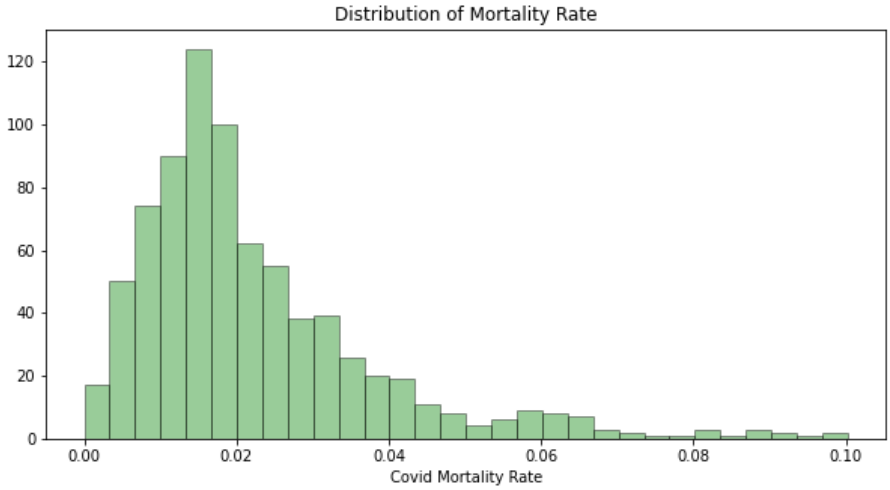

If the spread is 10-fold there are certain counties which will be most likely will be more affected by COVID-19. The counties with the highest percentage of the population likely infected are:



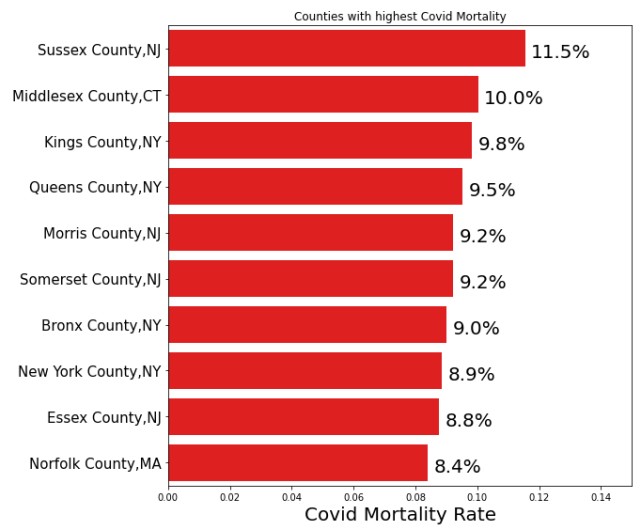
The above results are based on the fact that if the number of cases is high in a county and likely future, most of the population will have antibody generated in them due to the first wave of the COVID-19. So, In 2nd wave, it is most likely the counties which were less affected in the first wave.

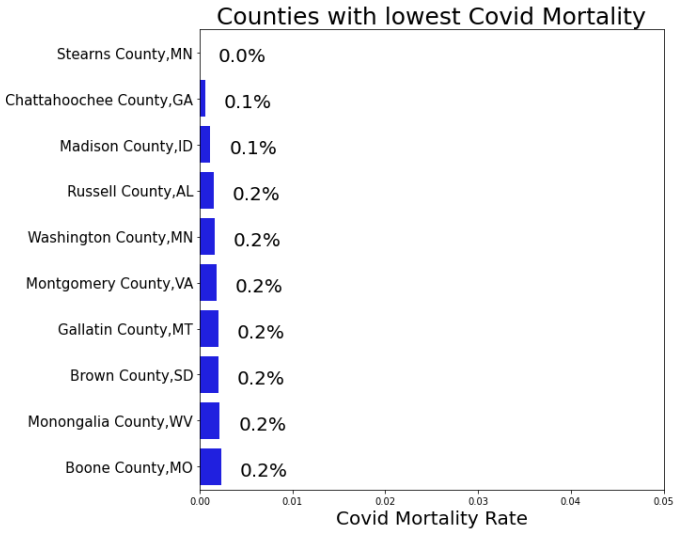
The COVID mortality rate is the ratio of the total number of deaths and the total number of cases in each county. We will explore what does mortality rate depends on and why some counties have a high mortality rate.

Below is the distribution of the mortality rate across counties in which the confirmed cases are greater than 1742 (Chosen as because 75 percentile of confirmed cases lies in this range)




Counties with highest and lowest COVID-19 Mortality rates. I have only taken counties with the confirmed cases > 1742. The high and low Covid mortality could be due to the healthcare facilities, population density, immune system of the county and many other factors.



It could be seen that the population under a certain group are more correlated to the target variable. So, below is the association between the percentage of the population in different age group in the county and COVID mortality. Bar plot shows the age group divided into 4 different quantiles.

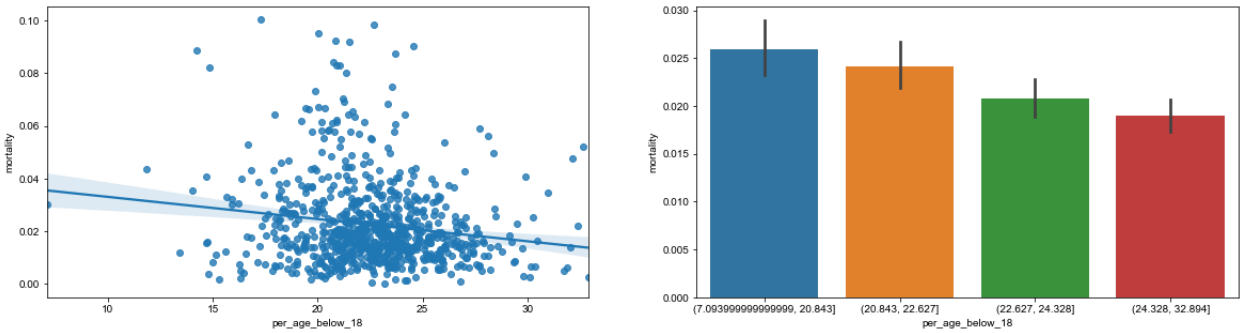



Fig. 1 Population of age below 18 and covid-mortality

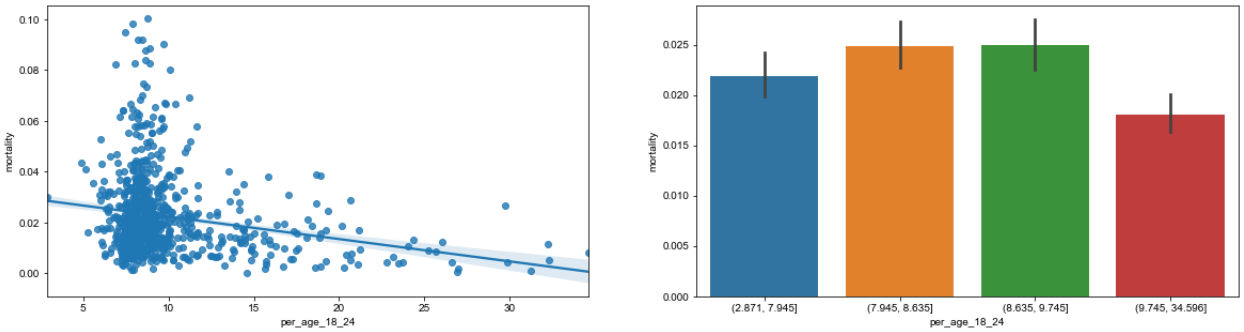



Fig. 2 Population of age between 18-24 and covid-mortality

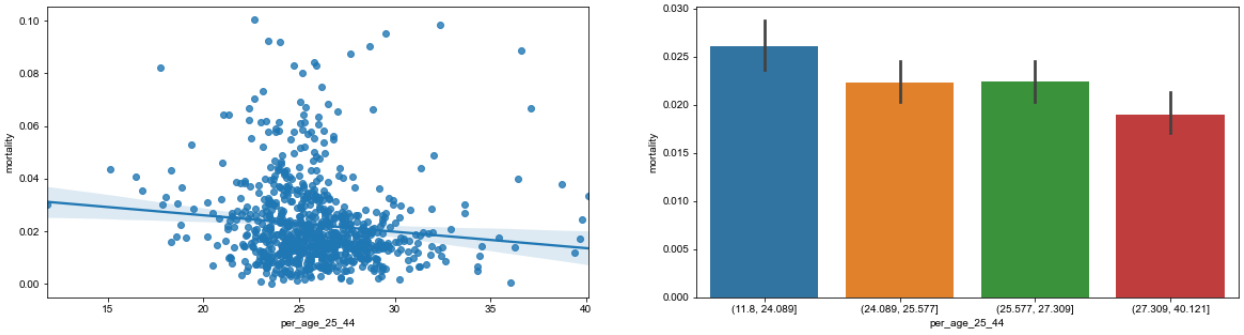



Fig. 3 Population of age between 25-44 and covid-mortality

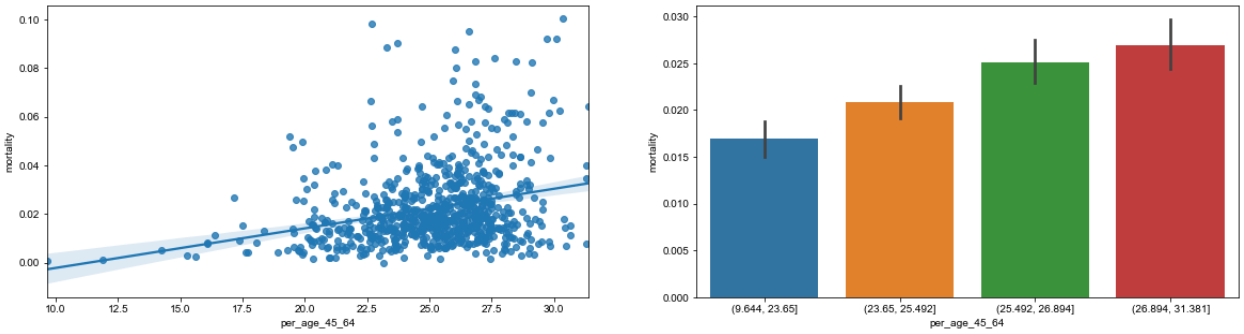



Fig. 4 Population of age between 45-264 and covid-mortality

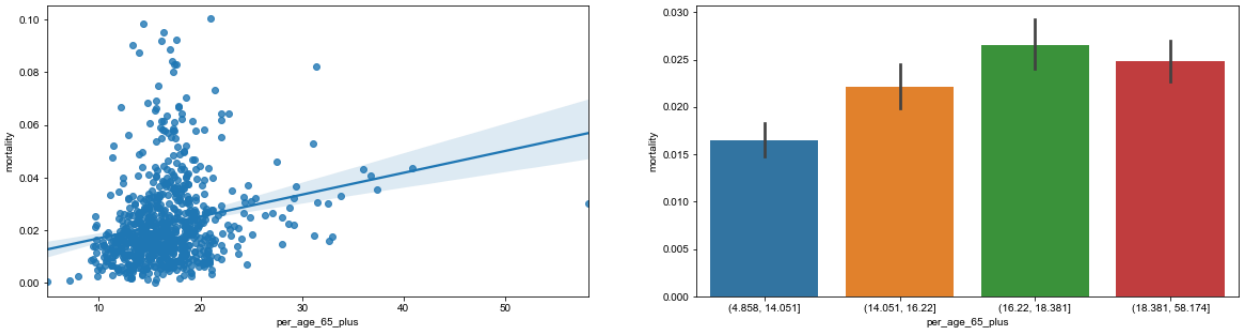



Fig. 5 Population of age above 65 and covid-mortality

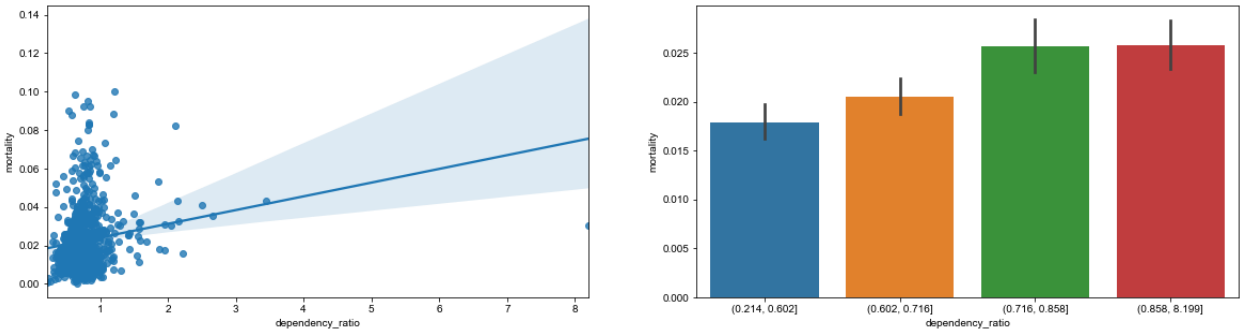



Fig. 6 Population of age (above 65: below 18) and covid-mortality

After the age group, the most important factor was the population of each race in each county. I have calculated the segregation level( It is the ratio of the black population to a white population living in the same neighborhood in a county). Here I have calculated the ratio of the total black population to the white population and Non- white to White American population.

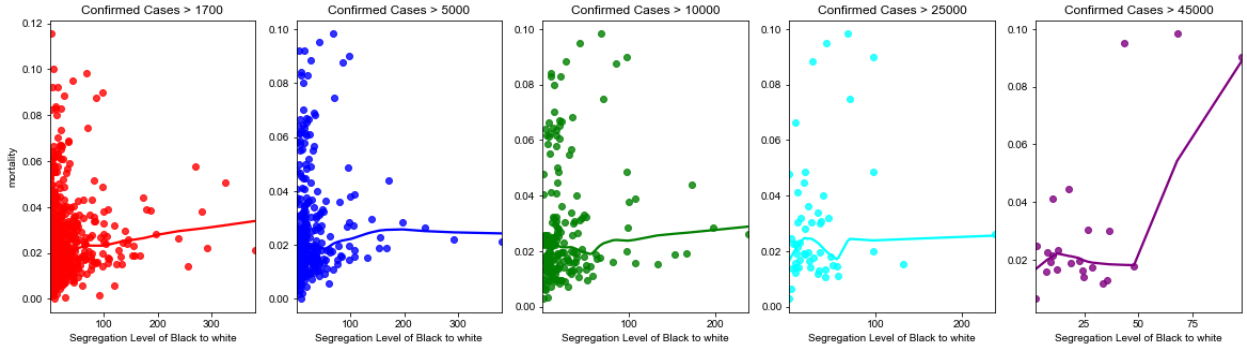



Fig. 1 plot of segregation level of Black to the white population and covid mortality

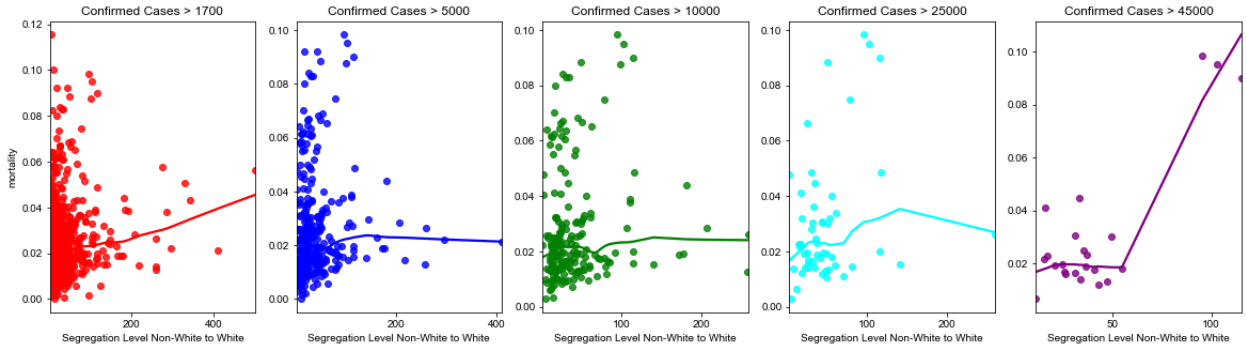


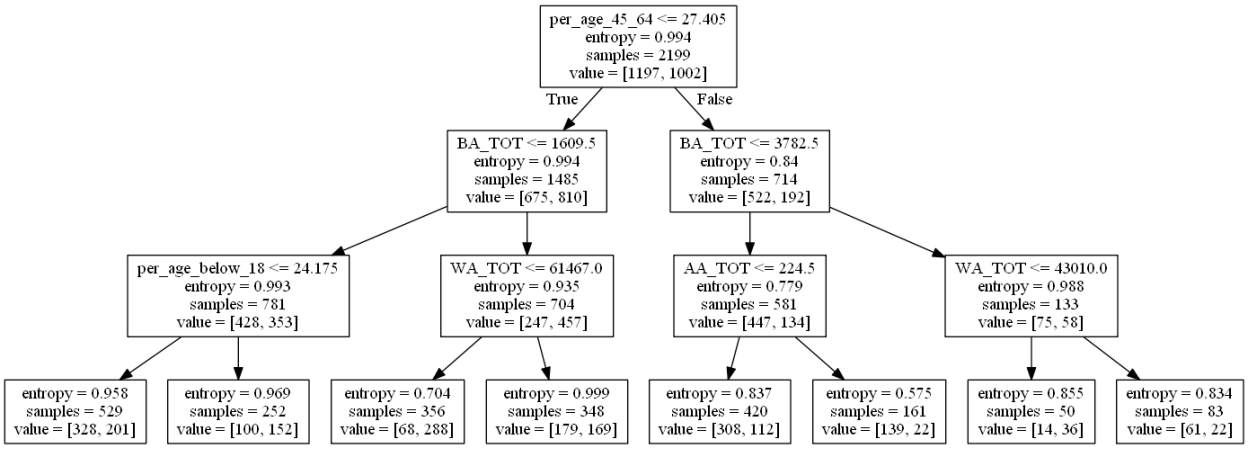

Fig. 2 plot of segregation level of Non-white to the white population and covid mortality

The above plots are various level of confirmed cases in each county, varying from 1700 to 45000 cases. Mostly there is no. much difference as Non-white population also includes the black American population, and the rest of the race population is a small number so we can only see very less variation between the two plot

# Modelling and Evaluation

The modelling of the data was done on the Target variable Outcome and I have compared a few models to check which model gives us the best accuracy.

- **Decision Tree**
  The Decision tree shows us that target variable or Covid-cases are is mostly was related to the percentage of the population between age group 45-64 and below 18, Black American, White and Asian population. These were the variables highly correlated to the target variable in the correlation matrix.

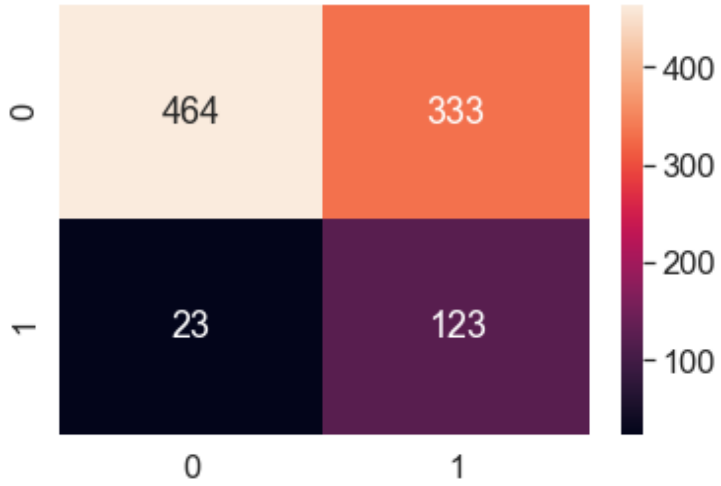



Confusion Matrix:

| n=943 | Predicted: No | Predicted: Yes |
|---|---|---|
| Actual: No | 413 | 253 |
| Actual: Yes | 74 | 203 |
| | Accuracy=(413+203)/943=65.32% | |

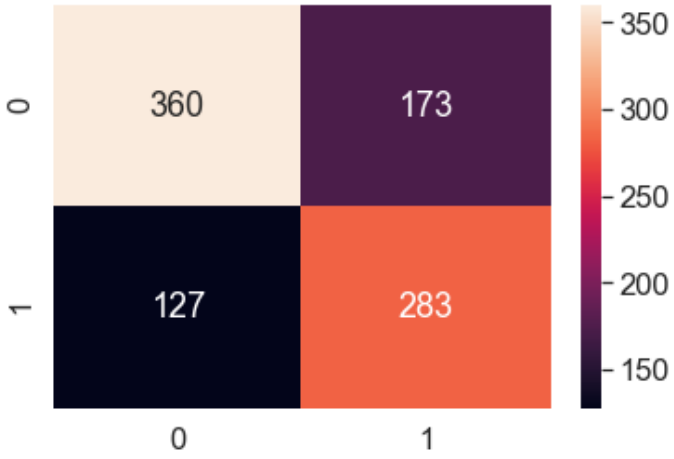
- **Logistic Regression Model**
  Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

  Confusion Matrix:



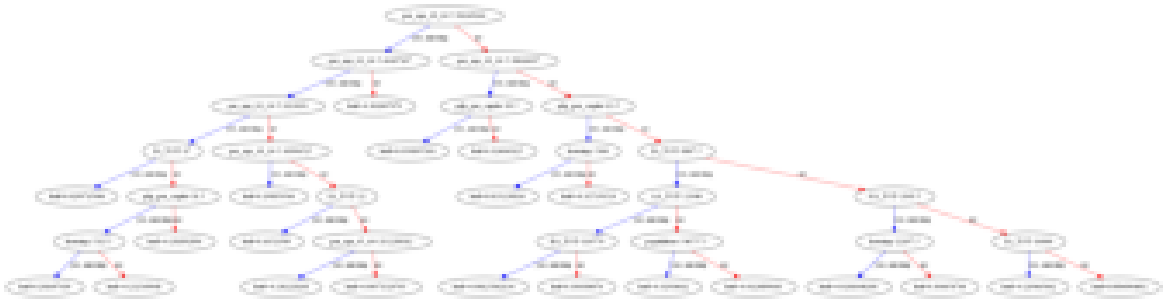

  Accuracy=(464+123)/(464+123+23+333)=62.248%

- **Gradient Boosting Classifier**
  Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.




  Accuracy=(360+283)/943=68.18%

Decision Tree By Gradient Boosting:



## Conclusion

The different predictive models give us different accuracy of the models best among them was the Gradient Boosting Classifier with the accuracy of 68%. All these models use regression and one target binary variable for modelling. In the above covid dataset the accuracy can be seen as little low. For correct accuracy of the model variables such as number hospital beds in the county, No. of ICU, the percentage of people likely earing the mask and population density and various other factors can be used for better accuracy.

Overall our model was able to predict that age group 44-64 are most likely affected by it and it can also be seen that most of the people affected by COIVD-19 were older generation lying in this group range with underlying disease.