1. INTRODUCTION / BUSINESS PROBLEM

1.1 Background

The outbreak of the nouvelle coronavirus (COVID-19) has been having devastating effects on the health status of humans across the globe. The virus has been identified as a very contagious disease which has more health impact on the aged and those with cardiovascular infections – making them more vulnerable to being severally affected by the virus.

Though, the virus is reported to have less than 1% infection rate, the rate of spread of the infection across borders and cities all over the nations has called for a state of emergency so as to combat this virus. The virus has been more prevalent in the non-African countries in past few months, but more recently, the virus has attacked Africa nations and there is fear for the upending of these nations due to their poor healthcare facilities and social infrastructures.

1.2 Problem

Nigeria, the most populous country with the best economy in Africa, is one of the affected nations with over 500 cases of the Nouvelle Coronavirus. The country has been given a lot of donations and raised a huge amount of money to combat the virus. But searching through the portals of National Bureau of Statistics, Nigeria, there is no adequate information about the trends and the future predictions of this corona virus as well reports on its relationship based on the demographic distribution. These pieces of information could inform the government of the best approaches that could be harnessed to prevent the spread and curtail the death toll across the states in the nation.

1.3 Interest

This project aims at the identifying the relationships between the different factors that could affect the spread of the virus as well as the death toll in case of a widespread outbreak such as average age, population density, population, literacy rate, GDP (Gross Domestic Product – indicator of economic prowess) and the current COVID-19 statistics. The analysis would be based carried out per state and the number of hospitals present in the central city of these states would be identified. This would aid in recommending to the government, the states that would be in more dire need of their support.

2. DATA ACQUISITION AND CLEANING

2.1 DATA COLLECTION

The datasets to be used for the analysis are collected across different platforms which is depended on the data requirement. The data on the status of the nouvelle coronavirus cases in Nigeria are collected from the National Centre for Disease Control through [this link](https://covid19.ncdc.gov.ng/) which highlights the confirmed cases, active cases and death toll across the nation. The population forecast of Nigeria for 2019 is obtained from the National Bureau of Statistics through [this link](https://nigeria.opendataforafrica.org/ifpbxbd/state-population-2006). Other datasets such as the Gross Domestic Products (GDP), the literacy level and the geographical locations are obtained from [here](https://en.wikipedia.org/wiki/List_of_Nigerian_states_by_GDP), [here](https://en.wikipedia.org/wiki/List_of_Nigerian_states_by_literacy_rate) and [here](https://www.researchgate.net/figure/The-36-states-of-Nigeria-and-their-coordinates_tbl1_273009753) respectively.

These data are not collected for exactly the same period of time but they are all obtained within the last five years, thence, it could be assumed that there have not been much changes in the demographics of the country.

The data on the location of hospitals at the central cities of each state would be collected using the Foursquare Application Programming Interface (API). The collection would be based on a radius of 5000m from the central location and a limit of 500 responses would be imposed.

## 2.2 DATA CLEANING

The datasets were obtained from different sources and hence, were merged to a single table on a column which is common to all the table, i.e. name of states. During the merging process, a number of problems were encountered and are dealt with accordingly.

The states that don’t have any confirmed case of coronavirus and other COVID-19 statistics, returned *NaN* (Not a number). These states are filled with ‘0’ using the *pandas.DataFrame.fillna()* function. Nasarawa state also returned a missing value for *Literacy Rate* and Nasarawa state is recognized as one of the poorest states in terms of education and commerce. Hence, the missing value is replaced with the value of the 25% quartile of the *Literacy Rate*. The Federal Capital Territory, Abuja, did not report any *GDP* value and this region is one of the second most hit in the nation, hence, cannot be dropped. The *GDP* of Abuja was estimated as the 75% quartile of the *GDP* of all states due to its high commerce.

The *.info()* method was used to check the nature of the datasets and it was observed that the number of confirmed cases were returned as o*bjects* rather than *np.float64*. To correct this, pd.to\_numeric() function was used to convert all the values to *np.float64*. And the .describe() method was used to check the summary of the descriptive statistics of the dataset.

Boxplot was used to check for outliers in the datasets and the outliers were left so as to emphasized on the extreme cases of coronavirus infection recorded in the nation.

## 2.3 FEATURE SELECTION

None of the features were dropped as they are all important for the data analysis.

# 3. METHODOLOGY

## 3.1 EXPLORATORY DATA ANALYSIS

The correlation between the identified variables was uncovered using the *.corr()* method and thereafter visualized using a heatmap. The heatmap is as shown in Figure 1 below.

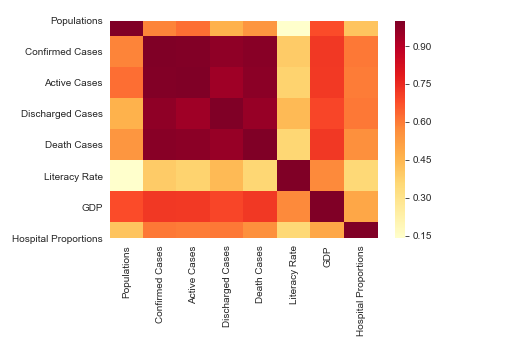


Figure 1: Heatmap showing relationship between the identified variables.

The above heat map indicates a strong relation amongst the features relating to COVID-19. It also inferred that GDP has a strong relation while hospital proportion has a moderate relationship with the COVID-19 variables. The Literacy rate has a poor relationship with the COVID-19 variables.

Then, the relationship between the death cases against population, GDP and literacy rate need further check to affirm the nature of their relationship.

### 3.1.1 RELATIONSHIP BETWEEN DEATH CASES AND OTHER FEATURES SUCH AS POPULATION, GDP AND LITERACY RATE

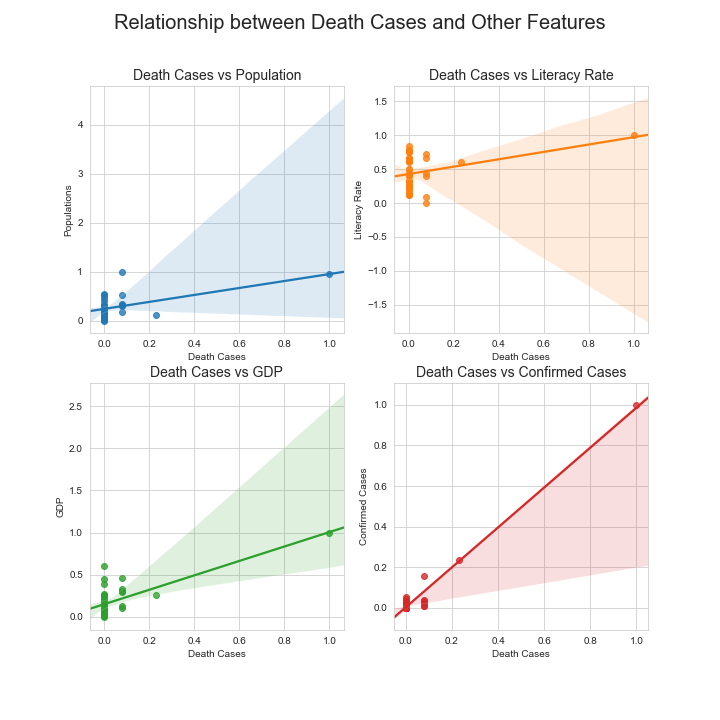


Figure 2: Relationship between Death Cases and four other variables (Population, Literacy Rate, GDP and Confirmed Cases).

The above plot signifies a positive relationship between the death cases and all other features. But an outlier is visible from the plots. As in all the cases, there is an extreme plot.

Also, from the plots, it is observed that majority of the plots are skewed. Hence, grouping the observation in clusters would signify the different clusters based on the cases.

Since, all of the cases are correlated, the death cases could be predicted using Regression Model. But before the models are deployed, the model could be checked for the most appropriate model (either Linear, Polynomial or other non-linear regression models) to be deployed using Residual plots as shown in Figure 3.

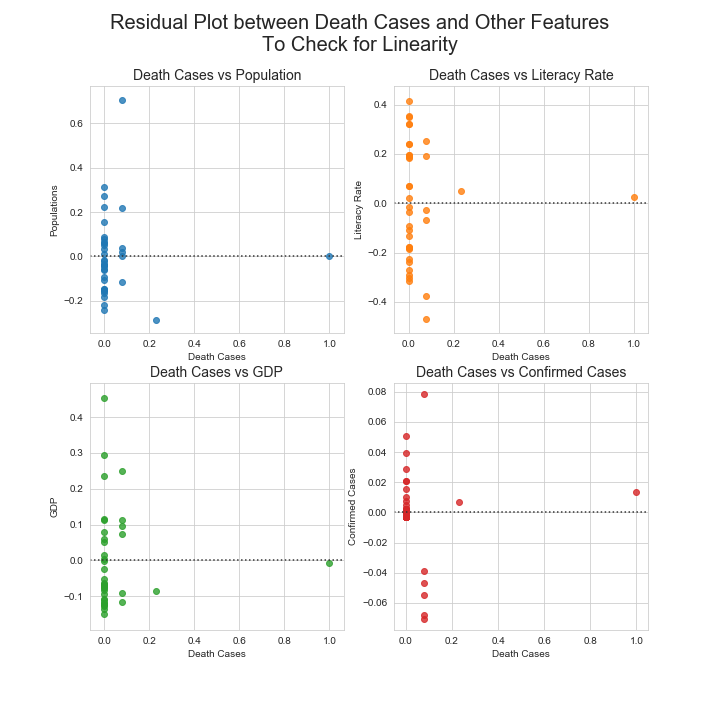


Figure 3: Residual Plot of the variables.

Hence, from the above descriptive statistical plots generated, it is the observed that the residual plots of the regression lines are not all randomly distributed (e.g. the plot against GDP and the plot against Confirmed Cases). Hence, a polynomial regression would best fit the relation between the features and the confirmed cases.

## 3.2 MODEL DEVELOPMENT

### 3.2.1 REGRESSION MODEL (FOR DEATH CASE PREDICTION)

#### 3.2.1.1 LINEAR REGRESSION MODEL

The Linear Regression Model was first developed to see the relative significance of each variable by comparing their corresponding coefficients. This model couldn’t be deployed for the actual prediction of the death cases because the residual plots suggested (most likely) a non-linear model.

The Linear Regression Model was developed, trained and tested and the output is given in Table 1.

Table 1: The results obtained from the Linear Regression Model

|  |  |  |
| --- | --- | --- |
| S/N | Result Description | Value |
| 1 | Intercept | 0.0199 |
| 2 | Coefficient for Population | -0.1141 |
| 3 | Coefficient for Literacy Rate | -0.0405 |
| 4 | Coefficient for GDP | 0.1331 |
| 5 | Coefficient for Confirmed Cases | 0.9878 |

Hence, from Table 1, the Linear Regression Model gives the model equation, Equation 1.

y = 0.01989 + 0.98777a + 0.13314b - 0.11411c - 0.04046d

where y = Death Cases, a = Confirmed Cases, b = GDP, c = Population, d = Literacy Rate.

#### 3.2.1.2 POLYNOMIAL REGRESSION

The Polynomial Regression Model was deployed on the datasets to obtain a closer prediction of the amount of death cases recorded per state from the COVID-19 pandemic. The Polynomial Regression Model is tested with different degrees of polynomial and the degree with the minimal error and maximal score was selected for the model. The result of the regression is depicted in Figure 4 and Figure 5.

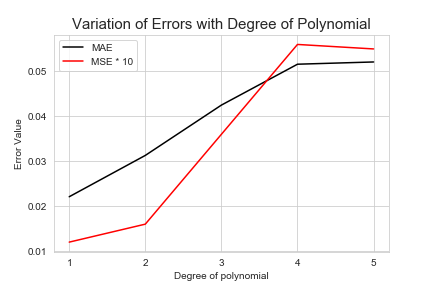


Figure 4: Variation of Error with Degree of Polynomial

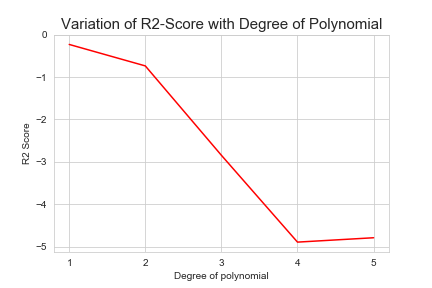


Figure 5: Variation of the R2 Score with Degree of Polynomial

From the above visualizations, the best degree of polynomial to be deployed is n = 1. Hence, we define our testing model as such. Based on the Polynomial Regression Model with degree, n = 1, the COVID-19 death cases for Kwara State and Lagos State are predicted as shown in Figure 6 and Figure 7 respectively. Similar predictions could be made for other states.

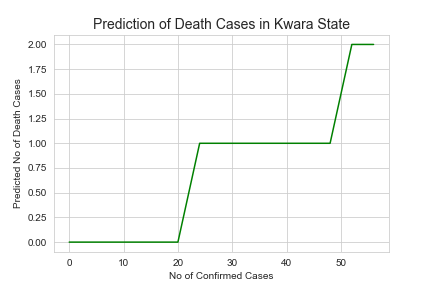


Figure 6: Prediction of COVID-19 Death Cases for Kwara State.

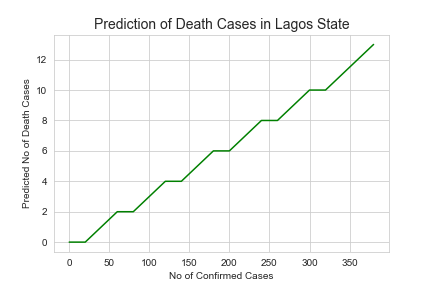


Figure 7: Prediction of COVID-19 Death Cases for Lagos State.

### 3.2.2 CLUSTERING MODEL

The clustering model deployed for the datasets is the k-Means Clustering Model which is used to group the datasets into k distinct clusters based on their features. The model was deployed with k = 4 so as to divide the states into groups with minimal infections, developing infections, serious infections and extreme infections. The output of the k-Means Model was generated using Folium Map as shown in Figure 8. Colour code for the cluster is thus: yellow = “Minimal Cases”, blue = “Developing Cases”, red = “Serious Cases” and black = “Extreme Cases”.

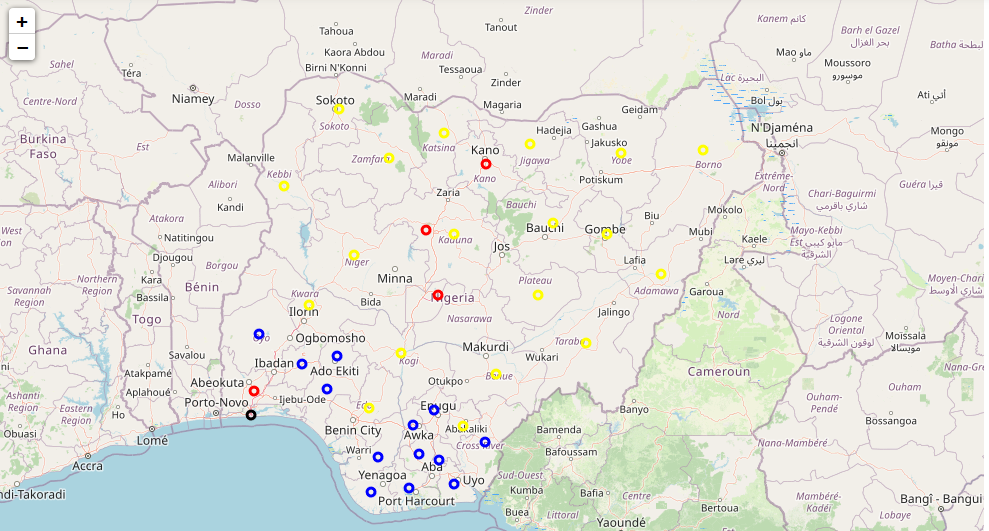


Figure 8: Map of Nigeria with Markers Indicating the Different Clusters based on COVID-19 statistics.

# 4. RESULTS AND DISCUSSIONS

## 4.1 Analysing the Linear Regression

### 4.1.1 Relative Strength of Each Variable on the Amount of Death Cases Recorded

The Linear Regression gave the following value as the co-efficient of the variable in the regression. The co-efficient give the relative strenth of each variable with respect to the target variable 'Death Cases'.

The coefficients values are as follows:

* Population = -0.1141
* Literacy Rate = -0.0404
* GDP = 0.1331
* Confirmed Cases = 0.9878

Hence, from the above, the Death Case is strongly correlated with the number of COVID-19 confirmed cases. The other features have very low relationship with the death cases recorded. Though, the state commercial status (GDP) also shows a very weak positive relation to the number of cases recorded.

### 4.1.2 Model Development to Predict Number of Death Cases

The Linear Regression has a score, R2 Score = 0.2271.

This implies that the relationship between Death Cases and other features outline namely Confirmed Cases, State Income Level (GDP), Literacy Level and Population has low linear correlation. Also, some of the residual plots obtained from the variables against Death Cases gave an ordered scatter-plot as an output which signifies that Linear Regression Model might not be deployable. Hence, polynomial regression need be used to obtain the model desired.

The Polynomial Regression, however, returns polynomial of degree, n = 1, as the best model to predict the death cases since it has the smallest error and the highest accuracy score. This negated the initial assumption because Confirmed cases has a very high relationship with Death Cases while other have close to zero relationship.

## 4.2 Discussion on k-Means Clustering

The goal of the analysis is to group the data based on the features selected into four clusters. And we make prediction based on the results obtained as follows:

* The first cluster are those that have zero or very low cases of coronavirus recorded. These are states that are sparsely or sparingly infected. It's also observed that majority of them have very low hospital proportions and relatively low value of GDP.
* The second cluster consists of state with low cases of coronavirus which range between 1 to 4 cases with exception of Osun, Oyo and Akwa-Ibom which could be termed as mis-classified states. The are states that are mildly or mininally infected and the spread of the virus is just developing. The states in this group are not correlated on GDP (a number of them has high values and a number also has small values) but majority of them have low hospital proportions.
* The third cluster (with exception of Nasarawa state which is mis-classified) consists of states with high amount of coronavirus cases as compared to the earlier discussed clusters. They also have relatively high proportion of hospitals. It also observed that these states have a high rate of commerce (GDP) compared to states in the previously discussed clusters. Hence, the states in this cluster are seriously infected and the virus is spreading.
* The fourth cluster is the case of an outlier. It is a point which has a very high amount of cases confirmed. It is observed that this state has a very high hospital proportions and also the best state for commerce (having the highest GDP). Hence, this state could be said to be extremely affected and the virus could be spreading exponentially.

### 4.2.1 Cluster Map Interpretation

The following could be observed from the map:

* From the Folium Map generated, it is observed that majority of the states in the Northern region are having minimal cases (yellow markers) while the states in the Southern region are having developing cases (blue markers). Hence, spread could be due to inter-state movements.
* States with serious cases (red marker) are not distributed based on location on the map. But these are regions with high commerce and hence, spread could be due to the that.
* The last case has been adequately categorized as an outlier with severe condition.

# 5. CONCLUSIONS AND RECOMMENDATIONS

The conclusions and consequent recommendation could be from the results and analysis as indicated in Table 2.

Table 2: Conclusions and Recommendations.

|  |  |  |
| --- | --- | --- |
| S/N | CONCLUSION | RECOMMENDATION |
| 1 | Northern States are sparingly affected. They have poor infractures to combat the virus if infected. | Measures should be taken against travelling to these states. |
| 2 | Southern States are increasingly infected. | Restrictions on movement should be made to curtail the spread. |
| 3 | Commercial States are seriously infected. | Commercial States are seriously infected. | Governmental intervention needed to provide facilities to treat infected persons. And restrictions on movement and commerce need be made to minimize spread. |
| 4 | Lagos State is relatively extremely infected. It is the home to the best commerce and also has good hospital proportion. | State of emergency need be declared and there should be total movement lockdown. The health facilities need be improved and more medical equipments and materials need be deployed. |

Hence, Lagos State and other states in the third cluster are in dire need of governmental support.