**COVID-19 DATA ANALYSIS PROJECT**

**Executive Summary**

**By**

**Ochai Emmanuel Odeh**

**Submitted to**

**Ustacky**

**August, 2022**

**Introduction**

The emergence of a new strain of virus in 2019 led to one of the world's deadliest pandemic in history. The coronavirus infection leads to acute health complications which challenge existing health systems and eventually, normal way of life of people.

The pandemic has had a catastrophic impact on global economies including Nigeria, a populous West-African country, which recorded its first case on 27th February 2020.

With insufficient detailed information on how the COVID crisis had and is influencing socio-economic development within the Nigerian context, this project obtained and analyzed COVID-19 related data to provide relevant authorities and the public domain with some meaningful and useful insights about the relationship between covid crisis indicators and existing socio-economic trends as well as the extent to which a pandemic of this magnitude has disrupted normal life as we know it.

The impact of the pandemic on Nigeria is measured as Overall COVID-19 Community Vulnerability Score (OCCVI) . The term “vulnerability” refers to the impact of the virus on a community after the virus arrives. Vulnerability is expressed at the subnational level e.g states in Nigeria (rather than a single score per country for example in Nigeria). The Africa COVID-19 Community Vulnerability Index (ACCVI) of which Nigeria is one of is modular in that the overall vulnerability score of a region can be broken down into 7 main themes, which themselves are constructed from subthemes. For example: the Socio-economic Status theme consists of the following sub-themes: access to information, education, poverty and unemployment. Each sub-theme is weighted equally when computing the theme score, and each theme is weighted equally for computing the overall covid-19 community vulnerability index score. Subthemes can consist of multiple underlying indicators. The index reflects the risk factors for COVID-19 both in terms of clinical outcomes and socioeconomic impact. The Vulnerability Index is ranked from Very Low(0) to Very High(1+). Example: Borno has a very high fragility score of 0.97 due to civil unrest in the location. It also has a very high health system score of 1.00, due to weak health systems in the location.

The scope of this project that also includes a presentation of a report draws attention to helpful insights obtained from data analyses using python data analytics and visualization tools on the obtained data to verify trends in situational and critical variables during the pandemic. These insights will help determine national economic responses to pandemics such as the one posed by coronavirus.

**Data Overview**

Outsourced data for the analyses were all in csv format structured as series objects from Nigeria Centre for Disease Control (NCDC) official website and John Hopkin’s COVID github data repositories, while external data used was obtained offline and included and was titled: covid\_external, Budget data, RealGDP, all also in csv formats. NCDC data included count cases that were confirmed, on admission and deaths, while John Hopkins COVID data were also counts for daily confirmed, daily recovered and daily deaths cases. Online data obtained contains cumulative records up to August 2022.

**Methods**

Online data extraction from NCDC was by python scraping technique using beautifulsoup library to parse the website’s HTML tree. Steps taken here include locating and inspecting the page of the URL to be scrapped, finding the data to be extracted by parsing, writing the code which will include important lines such as:

html = requests.get(url).content

df\_list = pd.read\_html(html)

nig\_covid\_cases = df\_list[-1]

print(nig\_covid\_cases)

nig\_covid\_cases.to\_csv('ncdcdata.csv')

The code was run and the data extracted and stored in the appropriate format.

Python data importation techniques were applied to John Hopkins repository to retrieve data.

All datasets gathered were in csv format and saved, as same, in respective dataframes. The read function (pd.read\_csv) was used to pass url content into designated dataframes.

Before carrying out analyses, dataframes were viewed to obtain basic information about the data and become acquainted with its contents using the head() and info() method. Dataframes were also cleaned, treated and prepared by:

* Converting column values to appropriate data types.
* Renaming the columns of the scraped data.
* Removing comma(,) in numerical data
* Sorting data for orderliness e.g states in alphabetical order and numerical ranges
* Merging NCDC related dataframes
* Extracting daily data for Nigeria from the Global daily cases data

**Analyses and Results:**

Some analyses were performed on the prepared datasets and results were described to generate corresponding insights. Some of the analyses carried out include:

* Generating and describing line plot outcomes that shows the top 10 states and all states in Nigeria in terms of Confirmed Covid cases by Laboratory, Discharged Covid cases, death cases using sort() and nlargest() method.

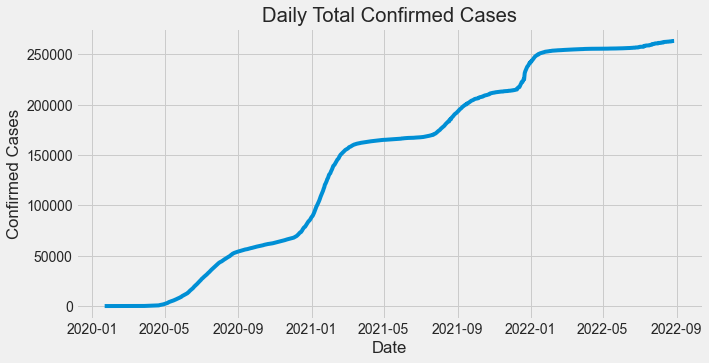
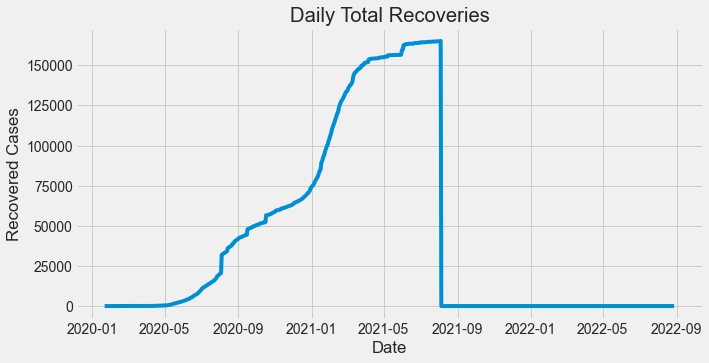
# Determining the relationship between the external dataset and the NCDC COVID-19 dataset where comparative analyses between vulnerability index and other indices was also done using appropriate Seaborns’s visualization methods such as sns.line plot(), sns.barplot(), sns.regplot(), 2-axis plot to illustrate this. Other relationships between other covid impact indicators for states (e.g Mortality and Health Systems), as well as total daily data for confirmed,recovered and death cases in Nigeria were also done respectively.

* Determination of the daily infection, recovery and death rates, using the Pandas *diff* method and to find the derivative of the total cases.
* Determination of highest and lowest values using python’s max() and min() functions for respective variables such as:
* Number for Laboratory confirmed, discharged, recovered and death cases for states and the country (Nigeria).
* Rate of daily infection, recovery and deaths for Nigeria as well corresponding date for each.
* Determination of the effect of the Pandemic on the economy by comparing the Real GDP value Pre-COVID-19 with Real GDP in 2020 (COVID-19 Period, especially Q2 2020). Pandas.melt was used to create DataFrame. This was used to construct a bar plot with subplots.

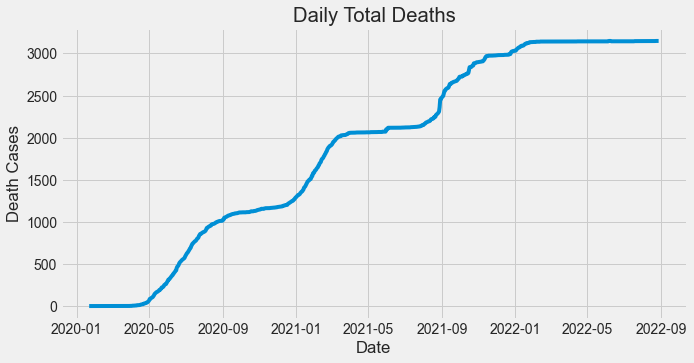
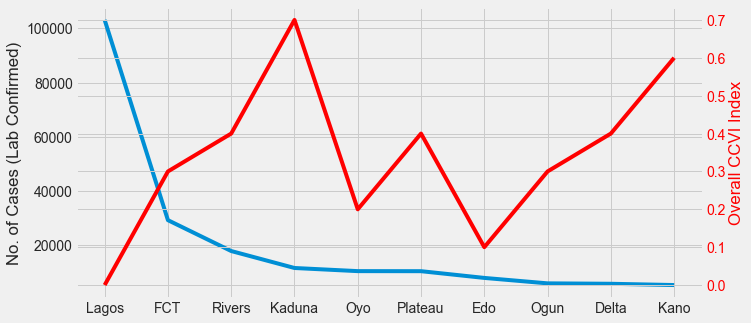
**Results**

Outcome of data analytics shows peculiar trends that provide useful insights when observed and described.

**Cumulative Daily COVID Related Cases**

**** 

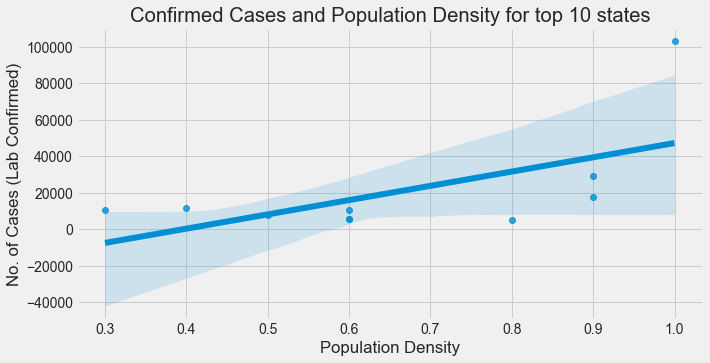
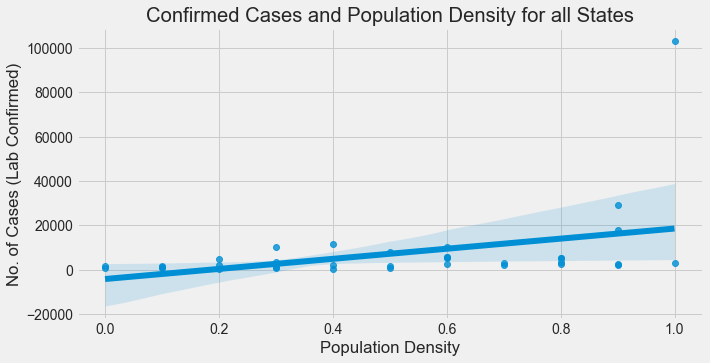
**Fig a. Fig b.**

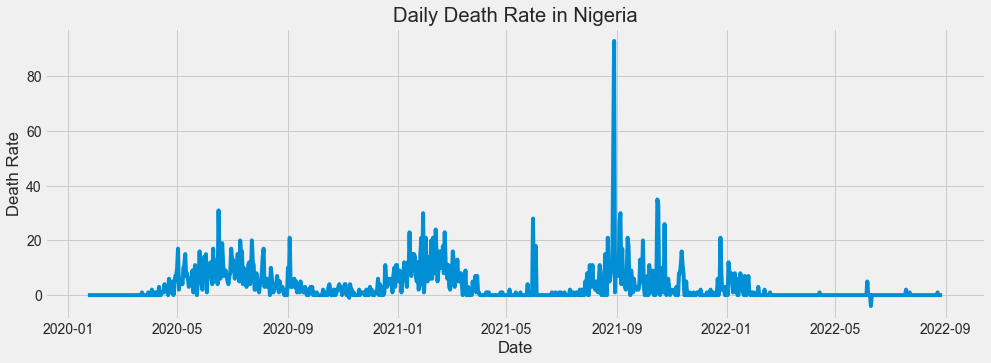
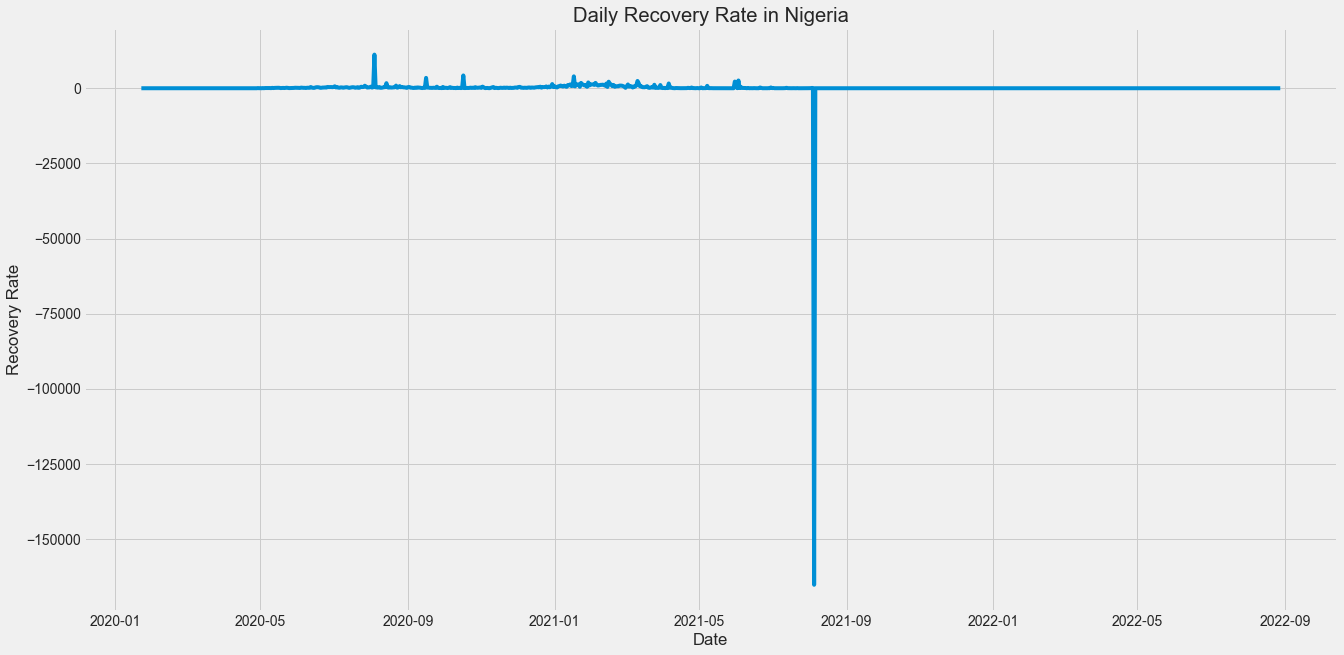
**Fig. c Fig. d**

| * The apparent increase in daily confirmed cases are cumulative values, with brief periods of unchanging records and exponential increments as time of data collection on 29th August, 2022 (fig. a). * No record of recovery in Nigeria was recorded from 2021-09 because only death cases were recorded as cumulative confirmed cases increased as time of data retrieval on 29th August, 2022 (fig. b) * Record for daily total deaths were cumulatively increasing with intervals of unchanging records and exponential increase as at the time of data retrieval on 29th August, 2022. (fig. c) * General progression of relationship between number of confirmed cases and OCCVI for the top 10 states shows an inverse correlation. Southern States like Lagos were not severely impacted by COVID-19 crisis. Northern States showed that they were worst hit by the effect of COVID-19 (fig. d) * This observation might suggest that Southern States like Lagos were not severely impacted by COVID-19 crisis due to some factors like education, better infrastructure and stronger local economy of which most northern states are challenged with (fig. d). |
| --- |

**Relationship between NCDC Data and Vulnerability Index Sub-Themes**

**Fig e. Fig f.**

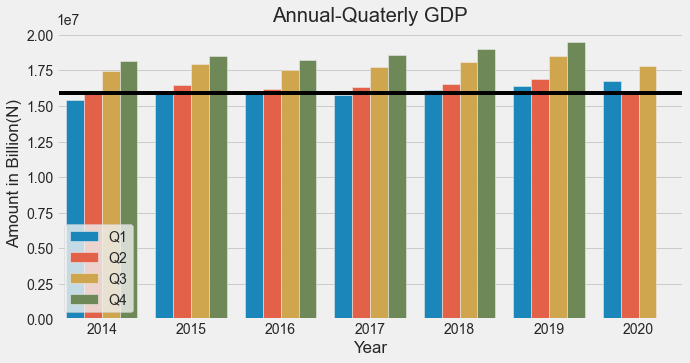
 

**Fig g. Fig.h**

| * There are three occurrences which are above the common distribution value across states and there are five occurrences which are below the common distribution value in states of various population densities. (fig e and f). * This shows that the number of cases confirmed for those three occurrences above the trendline are associated with a higher number of confirmed cases and lower population densities. The five occurrences demonstrate that lower records of confirmed cases are associated with higher population densities.(fig e and f). * The general trend for all states indicates that lower record of Lab confirmed cases is associated with higher population densities across all states. The many outliers below the trend line indicates this. * The highest recorded rate of recovery of 11188.0 was made on 2020-08-04 while the lowest was -165208.0 on 2021-08-05. (fig g). * There was a spike in the death rate in Nigeria on 2021-08-29 with a record high of 93.0. The reason may be because of relaxed preventive measures with apparent ease of local restrictions. The lowest death rate of -4.0 was recorded on 2022-06-10 perhaps due to the success at global control of the coronavirus, especially in the year 2022 (fig. h). |
| --- |

**Annual and Quarterly GDPs during Pre-COVID and Post-COVID(2014-2020)**

Nigeria’s economic contraction in the second quarter of 2020 was no surprise to most analysts as the trends in other climes implied a significant dip would be recorded in the period. The real GDP (rebased in 2014) dipped by 6.1% y/y to N15.9 trillion in Q2 2020 from N16.9 trillion in the corresponding period last year. In nominal terms, the GDP contracted by 3%y/y to N34 trillion in Q2 2020 from N35 trillion in Q2 2019. With this drop, coupled with the average growth rate of 2% per annum (post-2016 recession), the economy may have lost over N1.4 trillion on the back of both demand and supply shocks.



**Fig. i**

| * The second quarter of 2020 (onset of impact of pandemic in Nigeria) was used to apply an axhline threshold and benchmark for comparison of dataset parameters. (fig i) * There is no record for the fourth quarter's GDP in 2020. (fig i). * Economic activities during the fourth quarter were virtually grounded hence there is no record of GDP for the fourth quarter of 2020. This also means that 2020 has the lowest combined GDP in the years under review. (fig i). * The impact of covid 19 was huge on Nigeria's economy around the second quarter, in 2020, because that was when the highest cases of daily infection were recorded. (fig i). * The negative effects of covid on the population and the control measures enforced by the government affected productivity, all compared to other years included in the review. (fig i). |
| --- |

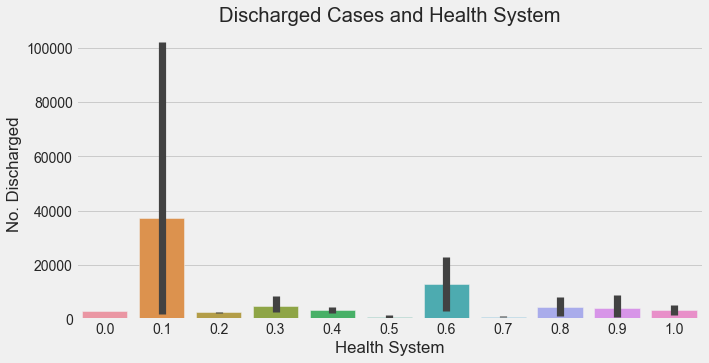
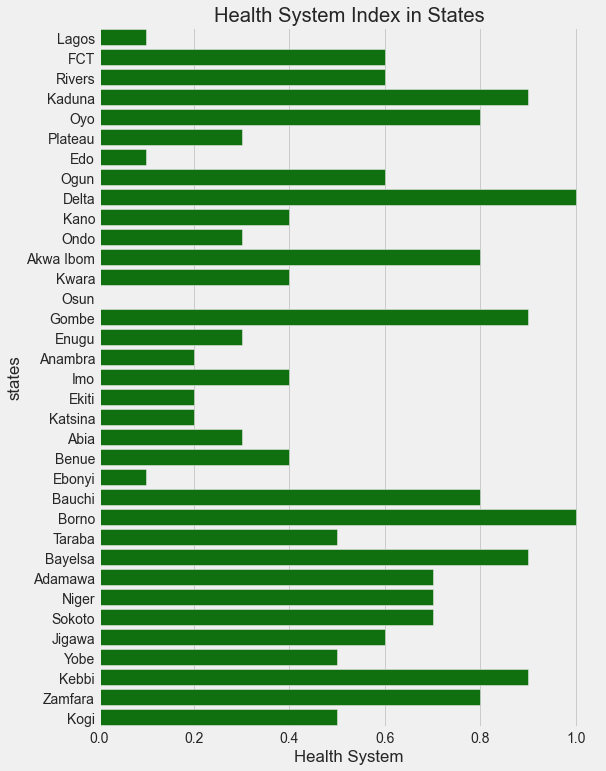
**Further Analyses**

The term “vulnerability” refers to the impact of the virus on a community after the virus arrives. Vulnerability is expressed at the subnational level (rather than a single score per country for example in Nigeria). The index is modular to reflect the reality that vulnerability is a multi-dimensional construct, and two regions could be vulnerable for very different reasons.

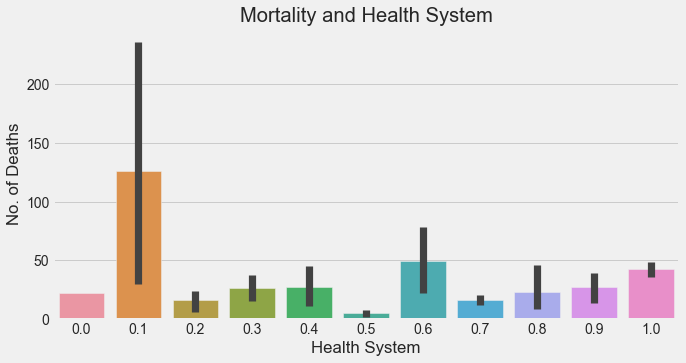
The Vulnerability Index is ranked from Very Low(0) to Very High(1+).

The index reflects the risk factors for COVID-19 both in terms of clinical outcomes and socioeconomic impact.  
Example: Borno has a very high fragility score of 0.97 due to civil unrest in the location. It also has a very high health system score of 1.00, due to weak health systems in the location.

**Comparing Vulnerabilities: Relationship between some vulnerability indices**

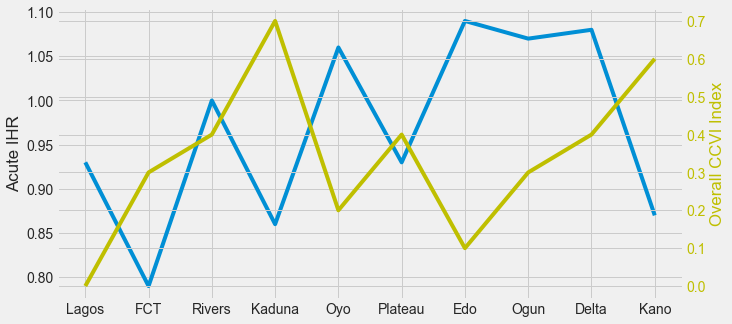
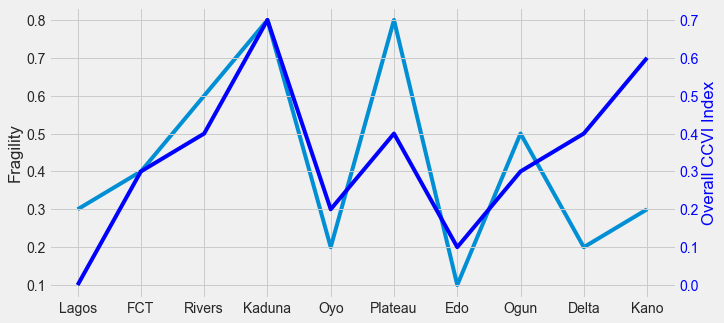


**Fig. j Fig. k**

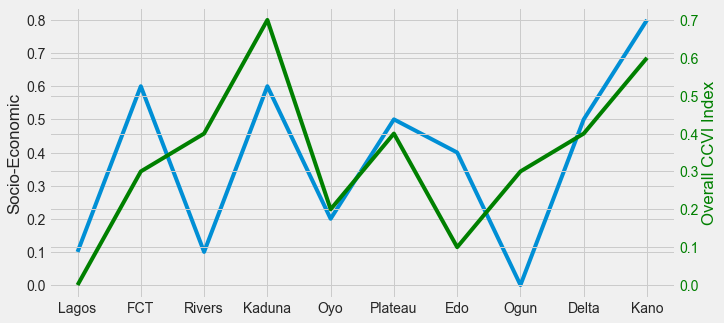


**Fig. l**

| * The barplots above shows that the health System index is diverse for states in Nigeria.Health systems of a country may influence the overall health impact of coronavirus. According to Africa COVID-19 Community guiding principles, higher system indices show that health crisis management is less effective and therefore there are chances that there will be a lower number of discharged cases. Higher health system index scores indicate a weak health system. (fig j and k). * So far from the barplot, it is inconclusive to say that the health system of Nigeria influenced the number of covid discharged cases in Nigeria. This is further highlighted by the large error bars which indicates higher measure of dispersion from any possible trend. (fig. k). * From the bar plots above, lower deaths are associated with various health system index scores. The Health System index of 0.1 (very low) was associated with the highest number of deaths. With this relationship and high standard of deviation, it is therefore not clear whether health systems in states in Nigeria had any significant influence on the number of deaths recorded during the covid crisis. (fig. l). |
| --- |

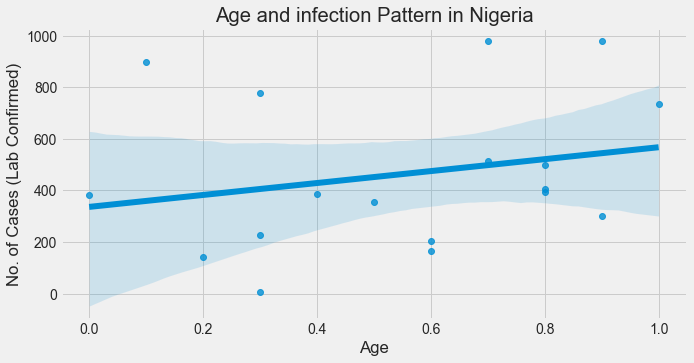
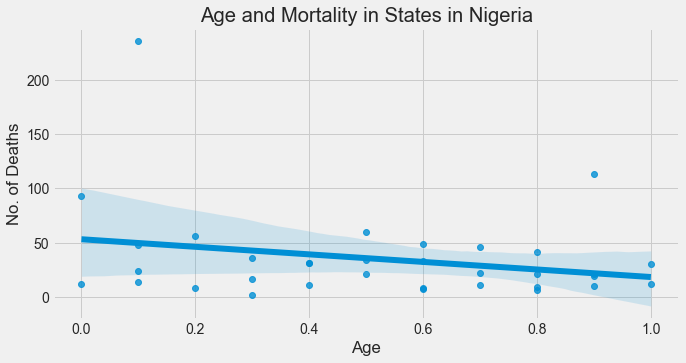
** **

**Fig. m Fig. n**

****

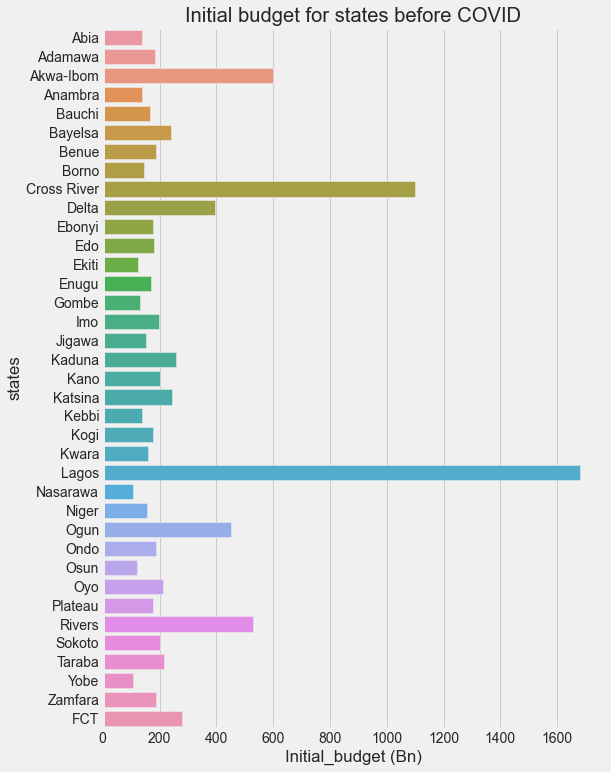
**Fig. o**

| * Inverse correlation between Overall CCVI index and Acute IHR were observed in all top 10 states. * All Southern states have higher Acute IHR and correspondingly low Overall CCVI, while northern states (Kaduna, Plateau and Kano) showed lower Acute IHR and correspondingly higher Overall CCVI index. * Plateau State apparently appears in between the situation with northern and southern states. Lagos has the lowest Overall CCVI with its corresponding high Acute IHR, while Kaduna has the highest Overall CCVI index with its corresponding Acute IHR index. (fig. m). * These observations suggest that the impact of covid crisis cannot be traced to Acute IHR alone (fig. m). * Fragility indices tend to correlate inversely with overall CCVI index. Subfactors considered for fragility index include: Population of concern sites, Food insecurity, Civil Unrest, e.t.c. Northern states (Kaduna, FCT, Plateau) tend to have high fragility scores but only Kaduna and FCT show that their Overall CCVI index is traceable to their corresponding high fragility index. The same cannot be said about Kano and Plateau. (fig. n). * Southern states with comparatively lower fragility index (Edo having the lowest), have slightly higher corresponding Overall CCVI score except for Lagos which shows the overall low fragility score but slightly higher fragility score. (fig. n) * Summarily, the direct correlation in only two northern states (Kaduna and FCT) demonstrates that vulnerability to the covid crisis is indeed tied to fragility scores, while the vulnerability may not be associated to fragility for the others in the top ten states. (fig. n). * Low Socio-economic index reflects the social and economic empowerments such as lower poverty rate, access to information, education and employment, higher index indicates high poverty, illiteracy e.t.c (fig. m). * With the exception of Delta, most southern states have lower Socio-economic index than northern states, with Lagos and Ogun showing the lowest. With Kaduna and Kano again with the highest OCCVI, Kaduna's lower Socio-economic index is the same as Lagos while Kano has the highest Socio-economic index. Lagos shows a low Socio-economic index with also a correspondingly low overall OCCVI. Rivers had high Socio-economic index but high OCCVI index. Both indices are virtually the same for Oyo state, while Edo showed lower OCCVI with high OCCVI. Both indices are roughly the same for Delta and Plateau (fig. o) * Overall, socio-economic indices alone cannot be attributed to the experienced impact of covid crisis. |
| --- |

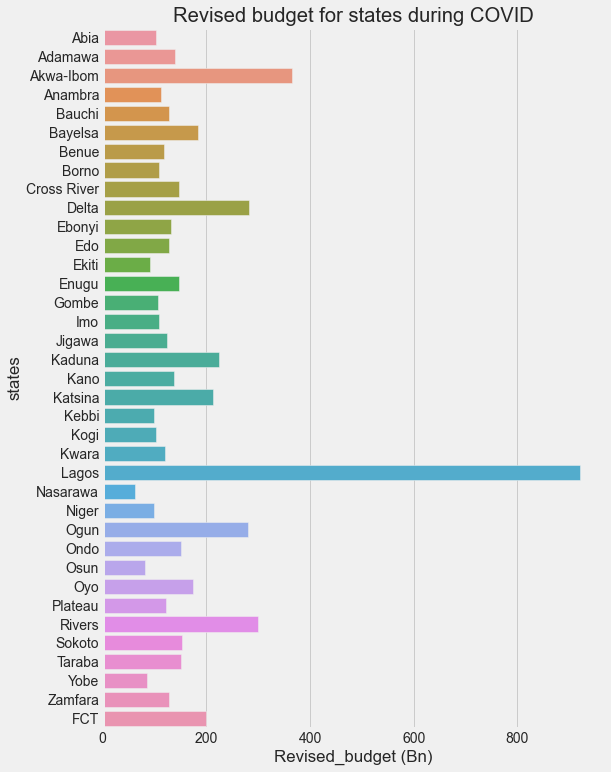
** **

**Fig. p Fig. q**

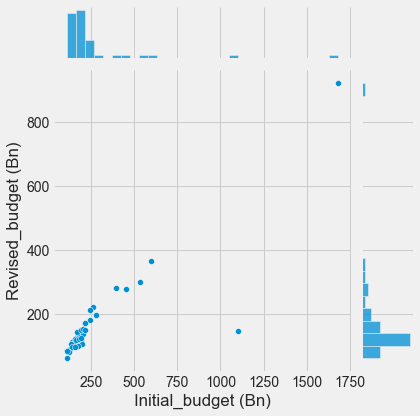
| * The above plot displays the association between number of confirmed cases and age of populations. It can be inferred from the plot that the occurrences along or close to the trendline are higher confirmed cases with higher Age index scores. There are however lesser confirmed cases associated with lower age index. (fig. p). * This shows that aged populace are more susceptible to coronarvirus than the younger population (fig p and q). * Apparently, the number of deaths tends to be more in the elderly population (more occurences tend to cluster close to the trendline at higher indices), even though many age categories are captured in lower covid related deaths. * Indices that can be traced to lower death are however associated with younger population. |
| --- |

****

**Fig. q**

****

**Fig. r**

****

**Fig. s**

| * Lagos had the highest initial and budget. (fig q). * All states undertook a downward review of their budget during COVID.(fig q and r). * Cross-River had the largest downward budget revision of 86.63% while Katsina had the lowest downward budget revision of 12.71% (fig q and r) |
| --- |

**Conclusion**

The truth whether the impact of the arrival of the coronavirus was influenced by existing critical socio-economic and clinical indicators, previously has been unclear. This project connects the dots to help provide the required information that will answer this question.

The key takeaways from these results include:

* Current NCDC cumulative data shows that Lagos still tops the table of laboratory, confirmed discharged and death cases, followed by F.C.T, as at 29th August, 2022.
* Daily cases for confirmed and deaths are still increasing with the exception of recovery where there have been none since September, 2021.
* Northern States of Nigeria tend to have higher vulnerability indices.
* Kaduna had the highest overall Covid-19 Community vulnerability index (OCCVI) score
* It is inconclusive to say that the impact of covid crisis for most states is connected to one index such as clinical indicators (number of confirmed lab cases, health systems, death cases), population densities, Socio-economic status, e.t.c alone. Overall vulnerability is influenced by a combination of multiple indices. Except for states such as Kaduna which showed that the impact of covid crisis can be directly linked to the number of confirmed cases in the state.
* Trend suggests that susceptibility to the covid infections are generally associated with elderly groups in populations, This shows that aged populace are more susceptible to coronarvirus than the younger population.
* Apparently also, the number of deaths tends to be more in the elderly population, while many age categories are captured in lower covid related deaths.
* The general trend for all states indicates that lower record of Lab confirmed cases is associated with higher population densities across all states.
* The impact of covid 19 was huge on Nigeria's economy around the second quarter, in 2020, because that was when the highest cases of daily infection were recorded. This led to economic activities during the fourth quarter to be virtually grounded hence there is no record of GDP for the fourth quarter of 2020. This also means that 2020 has the lowest combined GDP in the years under review (2014-2020).
* All states undertook a downard review of their initial budget in 2020, with Lagos having the highest budgets. While Cross-River had the largest downward review of over 80%, Katsina had the lowest of about 12%
* The negative effects of covid on the population and the control measures enforced by the government affected productivity. This explains the drastic cut in the total GDP in 2020.

**Recommendations:**

With the the insights obtained from this analyses the following is being recommended:

1. Since overall vulnerability to emergencies such as COVID-19 pandemic is influenced by combination several sub-themes and indices, population index during the covid crisis was a key determinant of the overall vulnerability index. Policy makers and relevant stakeholders can direct resources at addressing the sub-themes. They include: inaccessibility to information, unemployment, poverty, illiteracy, and a host of others. Tackling these can mitigate the impact of another wave of similar crisis.

2. Incorporating multiple information sources into a cohesive presentation also helps with pandemic vulnerability analysis, so the accuracy and sufficiency of data is highly recommended to increase the quality and reliability of data analytic outcomes.

**Future Direction of work**

The current approach of this project analysis is descriptive as it analyses past and current vulnerability data to observe real time trends. However a predictive and prescriptive approach to data analytics might help reduce the burden of managing unexpected crises such as the coronavirus pandemic. Preventive and preemptive measures will have been put in place on future arrival of such a crisis. Also data about possible areas of impact such as job availability, duration of future economic impacts can be obtained extrapolated.

This may require utilization of data intelligence methods (AI) to predict trends such as emergence of the new hotspots of infection and its peaks.

**References**

**1.** [**https://covid-static-assets.s3.amazonaws.com/Africa+CCVI+methodology.pdf**](https://covid-static-assets.s3.amazonaws.com/Africa+CCVI+methodology.pdf)

**2.** [**https://www.pwc.com/ng/en/assets/pdf/economic-alert-october-2020.pdf**](https://www.pwc.com/ng/en/assets/pdf/economic-alert-october-2020.pdf)

**3.**[**https://www.aljazeera.com/news/2020/11/21/nigeria-slips-into-recession-blamed-on-covid-19-and-oil-prices**](https://www.aljazeera.com/news/2020/11/21/nigeria-slips-into-recession-blamed-on-covid-9-and-oil-prices)