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 influenced 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.
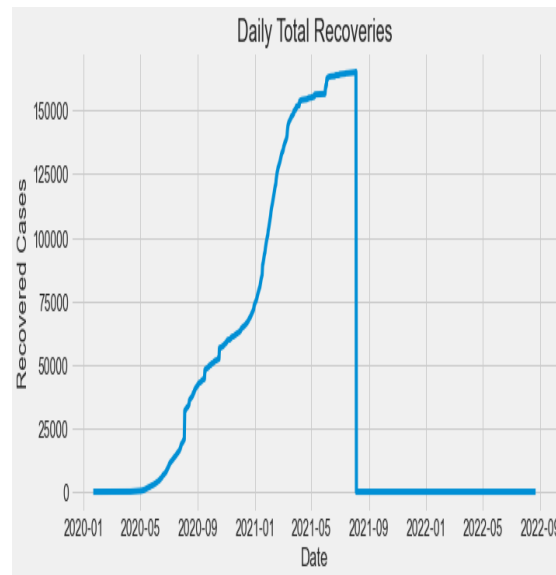
Cumulative Daily COVID Related Cases
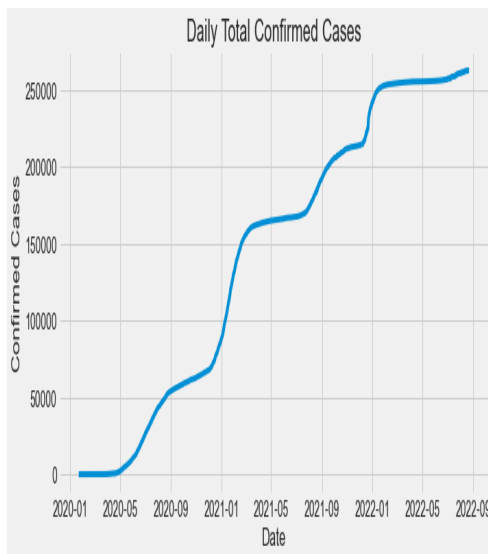
Fig a.
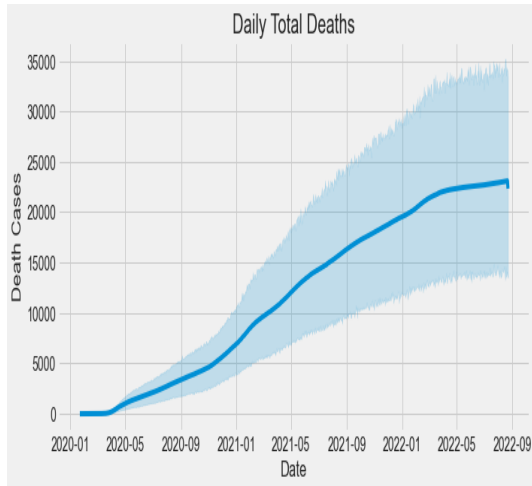

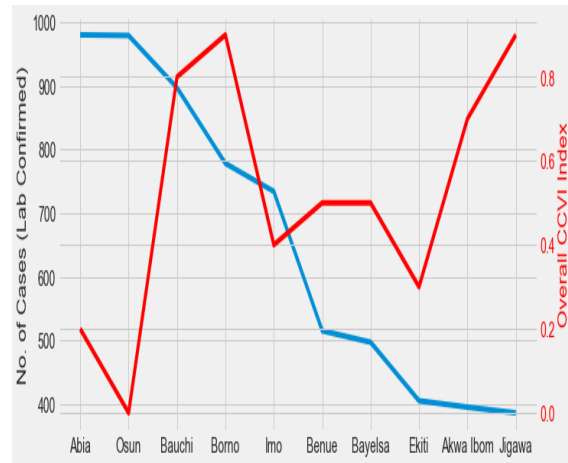**Fig b.**

Fig. c.   **Fig d.**

- Record for daily total deaths were cummulatively increasing as at the time of data retrieval in August 2022.

- Increased daily recovery cases were being recorded. These records again were cumulative values. No record of recovery in Nigeria was recorded from 2021-09 because only death cases were recorded as cumulatively confirmed cases increased as time of data retrieval in August 2022.

- Record for daily total deaths were cumulatively increasing as at the time of data retrieval in August 2022.

- Benue, Borno, Bayelsa, Ekiti, Akwa-Ibom and Jigawa all indicated inverse relationship between the No. of Cases (Lab Confirmed) and Overall CCVI Index i.e as Number of Cases (Lab Confirmed) decreased, Overall CCVI Index tended to increase across these states with Jigawa having the highest overall ccvi index but with lowest lab confirmed cases. For Abia, Osun, Imo, overall ccvi index were lower at higher confirmed lab cases. Bauchi state showed that the high number of confirmed cases was at par with its high overall ccvi index i.e the impact of covid crisis can be directly linked to the number of confirmed cases.

- Inverse relationships for other states further suggests that impact of covid crisis is unconnected to number confirmed lab cases alone.

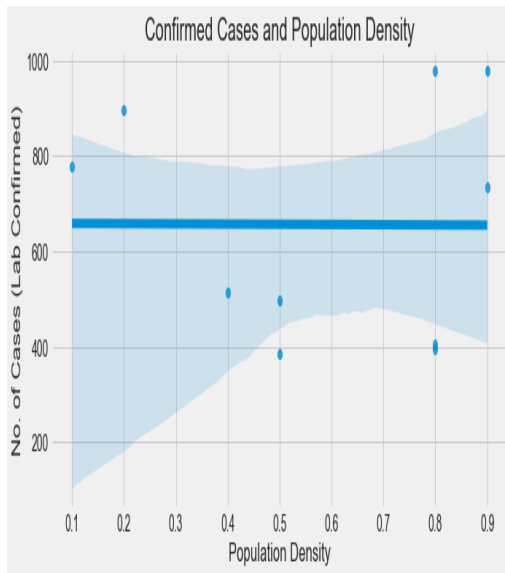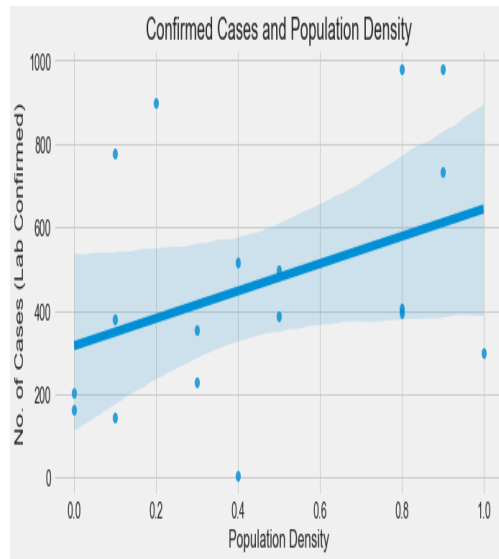# Relationship between NCDC Data and Vulnerability Index Sub-Theme
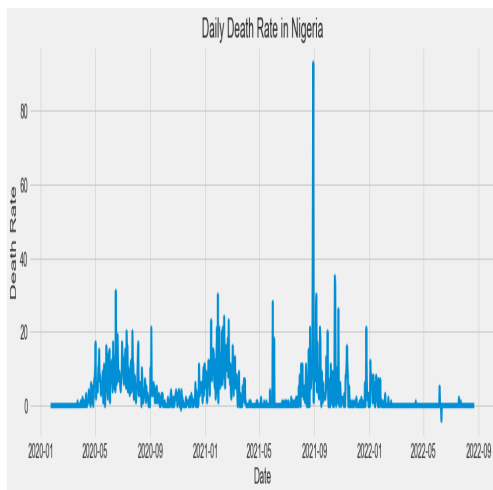

Fig e.
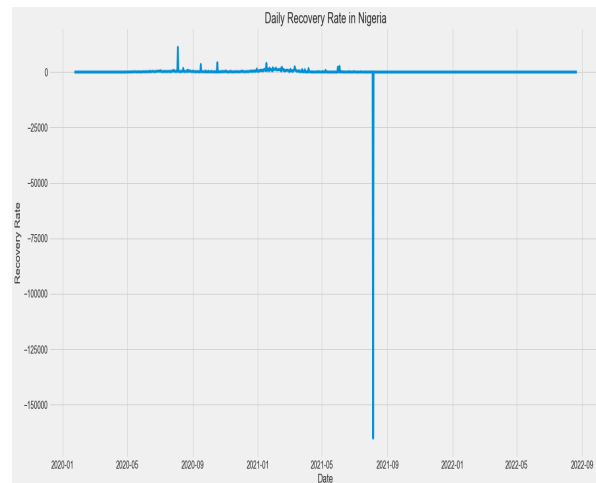

Fig f.

Fig g.

Fig.h

- For the top ten (10) highest confirmed cases, it is generally observed that there is higher uniformity in distribution of confirmed cases across most population densities in states in Nigeria. However, a few outliers (exception to this case) exists: There are five 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.

- The general trend indicates that higher record of Lab confirmed cases is associated with higher population densities across all states. There are also many outlier cases that do not align with this trend i.e does not show that higher or lower confirmed cases connected with higher or lower population densities.

- 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.

- 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.

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.
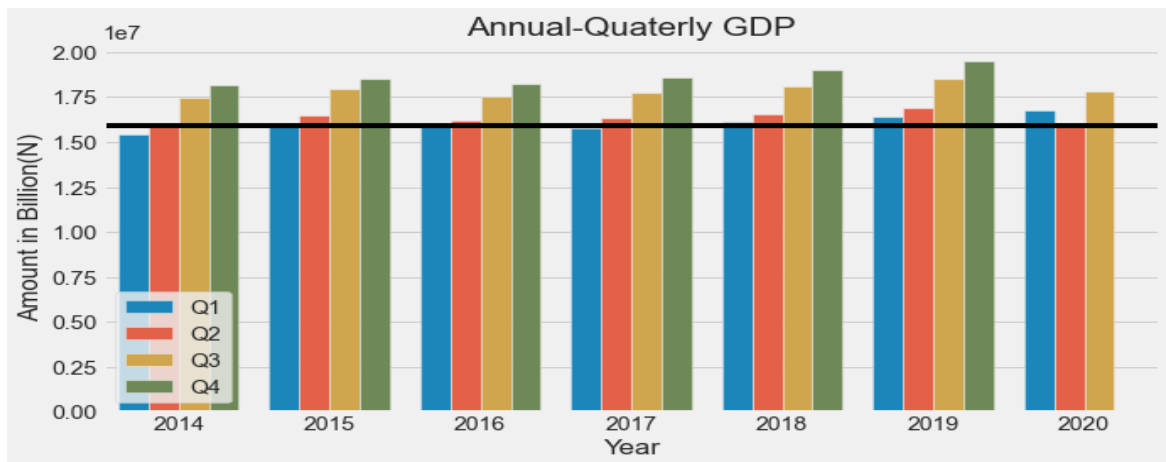

Fig. i

- 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.

- There is no record for the fourth quarter's GDP in 2020

- 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.

- 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.

- 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.

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.

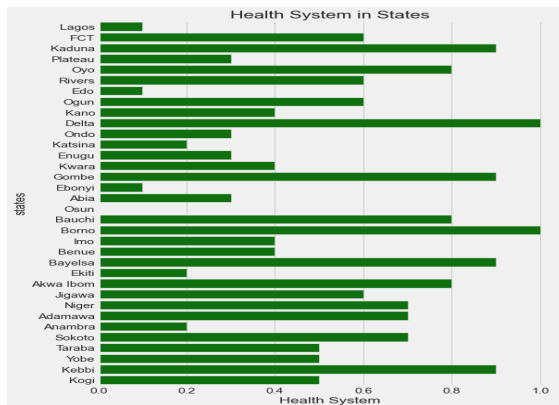**Comparing Vulnerabilities: Relationship between some vulnerability indices**
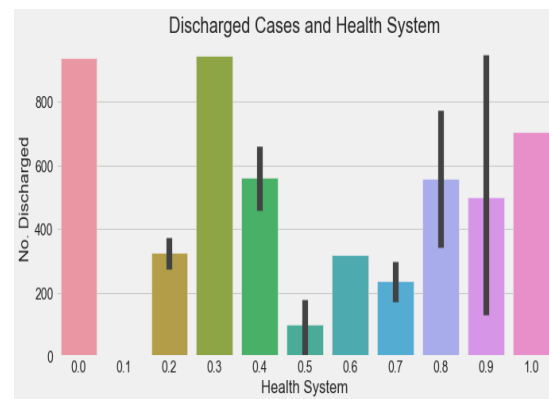

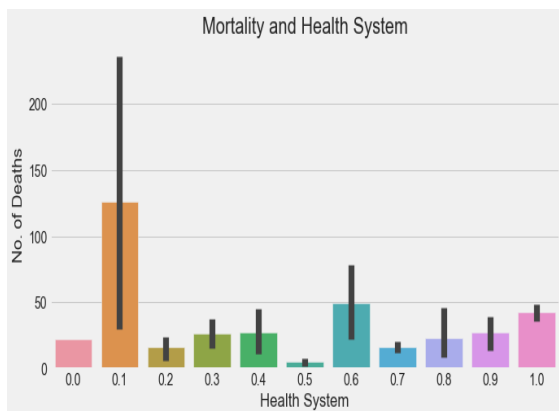
Fig. j

Fig. k

Fig. l

- Health systems of a country may influence the overall health impact of the coroner virus. Higher system indices show that health crisis management is more effective and therefore there are chances that there will be a higher number of discharged cases. 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.

- From the bar plots above, it is not clear whether health systems in states in Nigeria had any significant influence on the number of deaths recorded during the covid crisis.
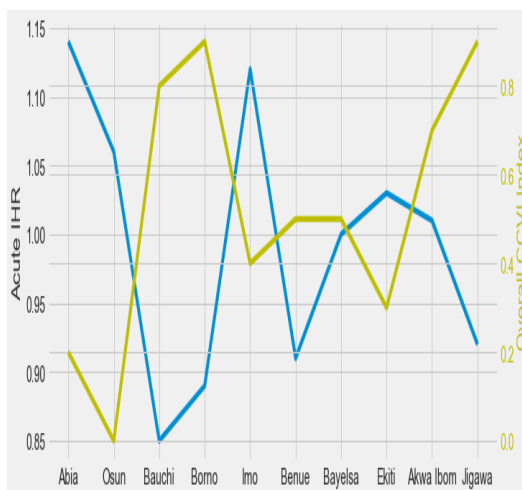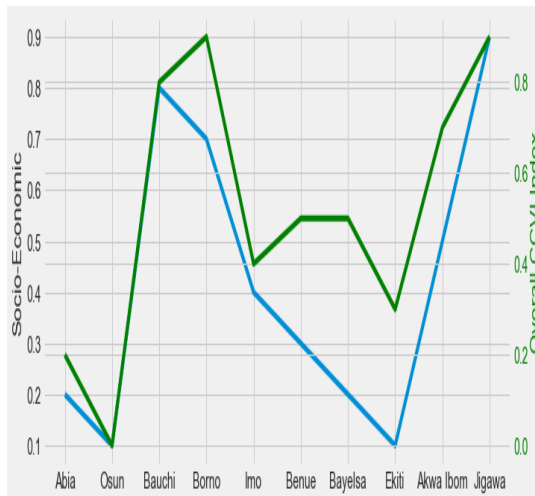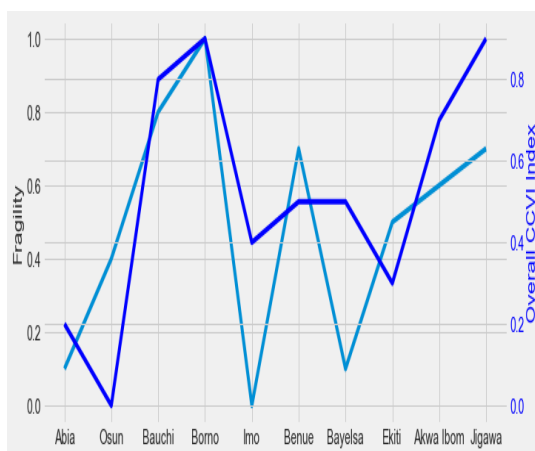
Fig. m

Fig. n

Fig. o

- For the top 10 states, here we see some direct correlation between Socio-Economic indices and overall CCVI index in most states. Generally Overall CCVI indexes are higher than Socio-Economic indices such as Osun and Bauchi and Jigawa. Osun state had the lowest and equal Socio-Economic indices and Overall CCV index. Jigawa had the highest and equal Socio-Economic index and Overall CCVI index. It can be inferred that the impact of the covid crisis is generally higher and can be traced to Socio-Economic indices such as poverty, unemployment, e.t.c

- Inverse correlation between Overall CCVI index and Acute IHR were observed in all top 10 states. Abia, Osun, Imo and Ekiti states had lower Overall CCVI index with relatively higher Acute IHR. Abia had the highest Acute IHR index and Jigawa the lowest. Borno and Jigawa had the highest Overall CCVI index. Bauchi, Borno, Benue and Jigawa States have the higher Overall CCVI index with correspondingly lower Acute IHR index. Borno and Jigawa had the highest Overall CCVI index. This observation suggests that the impact of covid crisis cannot be traced to Acute IHR

alone.

- Inverse correlation is observed between Fragility index and Overall CCVI index in almost all the top 10 states being considered here. While Osun, Benue and Ekiti all had lower Overall CCVI index with a correspondingly higher Fragility index, Abia, Bauchi, Imo, Bayelsa, Akwa Ibom and Jigawa states all had higher Overall CCVI index with correspondingly lower Fragility index. Only Borno showed equal Fragility and Overall CCVI indices. It is therefore inferable that Fragility alone can be responsible for the recorded impact of the covid crisis in Borno. This is however not applicable to other members of the top 10 states where Fragility alone may not be responsible for the recorded impact of covid crisis.
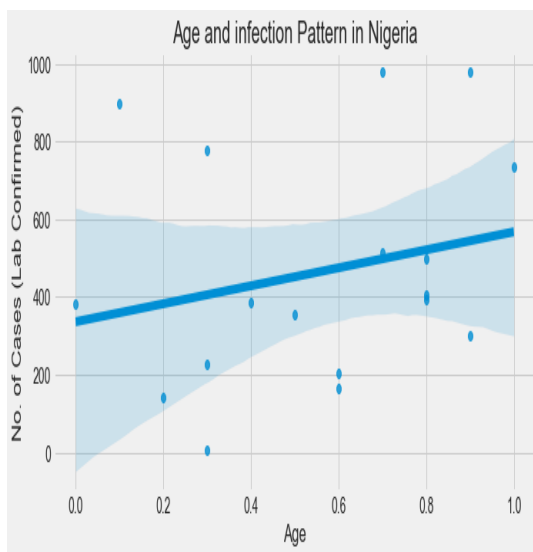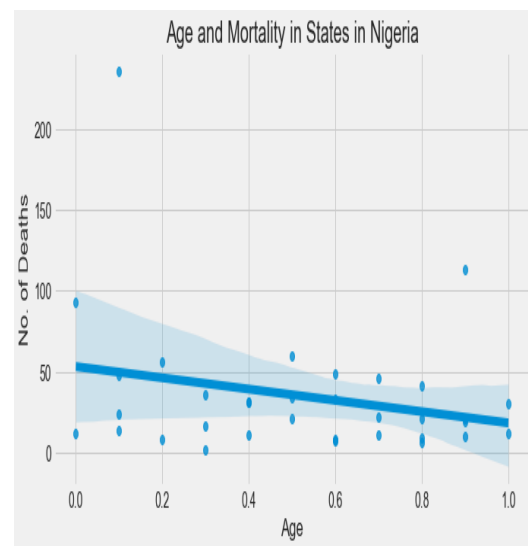

Fig. p

Fig. q

- The above plot displays the association between number of confirmed cases and age of populations. Though it can be inferred from the regression trend that susceptibility to the covid infections are generally associated with elderly groups in populations, many outliers indicate that many other age categories also were captured in high and low numbers of recorded lab confirmed cases.

- Apparently, the number of deaths tends to be more in the younger population because they are more in the overall population compared to the elderly group. The trendline of the regression does however indicate that the number of deaths recorded among the elderly group is still significant because they are less in population compared to the younger populace.

Conclusion

The truth whether the impact of the arrival of the coronavirus affected 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:

- Jigawa and Borno Had the highest overall Covid-19 Community vulnerability index (OCCVI) score.
- Northern States of Nigeria tend to have higher vulnerability indices.

- It can be inferred that the impact of the covid crisis is generally higher and can be traced to Socio-Economic indices such as poverty, unemployment, e.t.c

- It is conclusive to say that the impact of covid crisis for most states is unconnected to clinical indicators such as number of confirmed lab cases, health systems, death cases, population densities, alone. Overall vulnerability is determined by a combination of multiple indices. Except for Bauchi state which showed that the high number of confirmed cases was at par with its high overall ccvi index i.e 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, many outliers indicate that many other age categories also were captured in high and low numbers of recorded lab confirmed cases.

- The number of deaths tends to be more in the younger population because they are more in the overall population compared to the elderly group. The trendline of the regression does however indicate that the number of deaths recorded among the elderly group is still significant because they are less in population compared to the younger population.

- The general trend indicates that a higher record of Lab confirmed cases is not associated with higher population densities across all states. There are many outlier cases that do not align with this trend i.e does not show that higher or lower confirmed cases connected with higher or lower population densities.

- 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).

- 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.

Recommendations:

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

1.	Since socio-economic 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 of this index such as poverty, unemployment, illiteracy and information dissemination gaps. This 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 vulnerability data to explain past events. 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 arrival of such a crisis. Also information about possible areas of impact such as job availability, duration of future economic impacts can be obtained

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

2.   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-9-and-oil-prices