**Joshua O. Ayodeji**

**Data Scientist Microdegree Capstone Project**

Nigeria COVID-19 Data Analysis Using Python

**Introduction**

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus, and it has affected major parts of the world. Nigeria, a West-African country, has also been affected by the COVID-19 pandemic after recording its first case on 27th February 2020.

Nigeria is a country with 37 states - Federal Capital Territory included- and a fast-growing economic environment with about 200 million citizens. COVID-19 has affected several country activities as the country steadily progressed from its first case to shutting down major airports, state-wide lockdown, curfews, and reviving its economy.

Data science and analysis skills play important roles in providing insights from the Nigerian COVID-19 metrics such as:

* How the states rank in terms of total confirmed cases, total recovered cases, total death cases, the percentage of the confirmed cases that died as a result of the pandemic in the states?
* When were the highest infection rate recorded, maximum daily infection rate (plus the date it was recorded)?
* Is there a linear relationship between number of cases recorded and population density?
* What alarming factors could enhance the impact of the pandemic in each state?
* How can governments, donors, organizations, and companies allocate resources effectively? How more targeted, sector-specific interventions would mitigate the impact of the pandemic in these states?
* Which state is the most vulnerable?
* Did the outbreak have effect on Nigeria’s economy?

**Report Objective**:

This report is meant to communicate project findings, provide a summary of the approaches and explain results of the project.

**Report Outline**

1. Data Overview
2. Methods
3. Analysis
4. Conclusions
5. References
6. Appendices
7. **Data Overview**

The data sources are different, in this project the following data are combined to perform analysis and provide insights.

* The Nigeria Centre for Diseases Control (NCDC) monitors the country’s COVID-19 situation, and releases data on metrics across all the 37 states (included the total number of lab confirmed cases, number of lab confirmed cases that recovered, and number of the lab confirmed cases that died recorded in all 37 states in the country) on their official website. This data is used in this project.
* The Johns Hopkins University Centre for Systems Science and Engineering (JHU CSSE) publishes data on daily confirmed, death, and recovered cases across different countries. Nigerian daily data accessed from their repository was used in this project to derive related insights.
* Nigeria Community Vulnerability Index data (computed by considering several factors such as socio-economic status, population density, housing type, transportation, epidemiological, health system etc.) was used to determine relationships and correlations.
* Gross Domestic Product Data for Nigeria was used to determine the impact of COVID-19 on the economy.
* Nigerian States Budget Data was also used to determine the impact of the pandemic on the economy.

1. **Methods**

* Data Collection

The data on metrics across all the 37 states from NCDCs website was obtained by performing a web scrap of the website with pandas ‘read\_html’ function and saving the data to a dataframe object.

In obtaining data from John Hopkins Repository pandas ‘read\_csv’ function was used to extract data from the csv files on their repository to dataframe objects.

The Nigeria Community Vulnerability Index data, Gross Domestic Product Data, and Nigerian States Budget Data were obtained by using pandas ‘read\_csv’ function to read datasets provided by Ustacky on their ‘Nigeria-COVID-19-Data-Analysis-Using-Python’ repository.

* Approach to data analysis

1. Data Cleaning and Preparation: Involved extracting required data from dataframe objects (using pandas ‘loc’ function to extract Nigeria’s data from John Hopkins data) and creating new dataframe objects off the initial dataframe objects, slicing dataframe objects to filter off unrequired data, setting proper indexes for the dataframe objects, setting proper datatypes for some columns in the dataframes, and renaming column names.
2. Data Analysis and Visualization: Calculations were performed using pandas. Various figures were plot (discussed in detail in next section) on the results.
3. **Analysis**

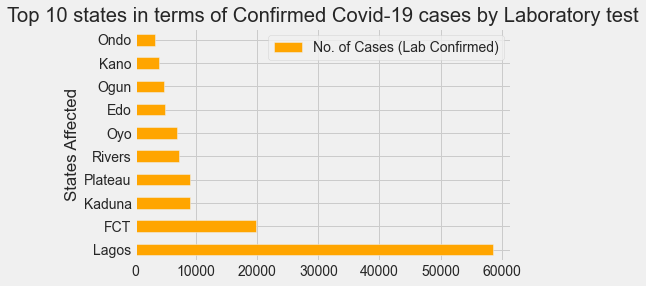


Figure : Horizontal bar plot of the top 10 states measured by the number of lab confirmed COVID-19 cases recorded

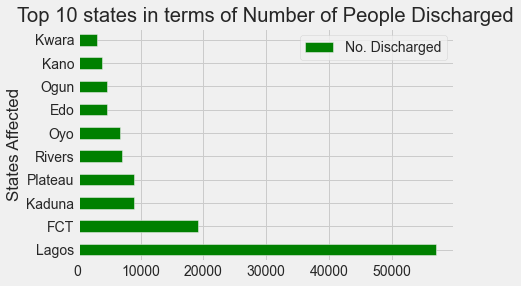


Figure : Horizontal bar plot of the top 10 states measured by the number of COVID-19 patients that were discharged due to recovery

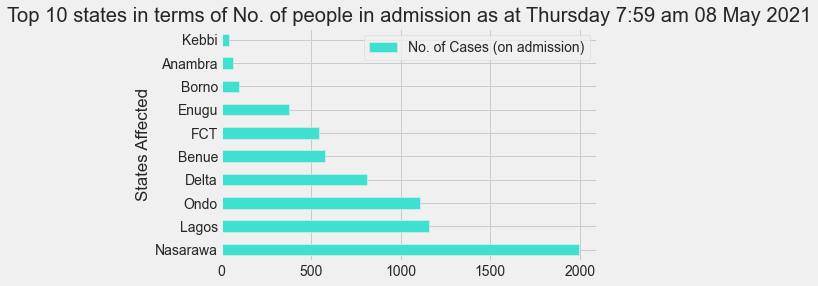


Figure : Horizontal bar plot of the top 10 states measured by the number of COVID-19 patients that are under care.

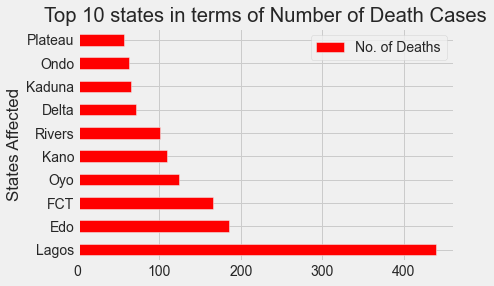


Figure : Horizontal bar plot of the top 10 states measured by the number of COVID-19 caused deaths that were recorded

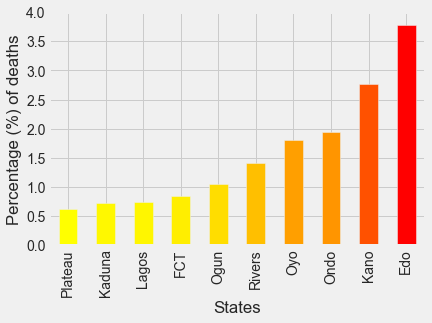


Figure : Bar plot of the percentage of the confirmed COVID-19 cases that resulted in death among the top ten states (measured by cases recorded).

Figure 1 shows that Lagos State recorded the most number of cases, more than two times the number of confirmed cases in FCT (the State with the second highest number of confirmed cases).

Figure 2 shows that nine (the top nine) of the top ten retained the same position on both the recorded cases and the discharged cases.

Figure 3 shows that Ondo State might retake their position as tenth if at least 90% of cases on admission get discharged, it also show that a large percentage of their total recorded cases were recent, since about one-third of the number is under admission.

Figure 4 shows that Edo state moved from number seven to number two, Plateau from fourth to tenth; indicating that Edo state had relatively higher death rate and Plateau state had relatively lower death rate.

Figure 5 shows the relative death rates of the ten states, and confirms the validity of insights from Figure 4.

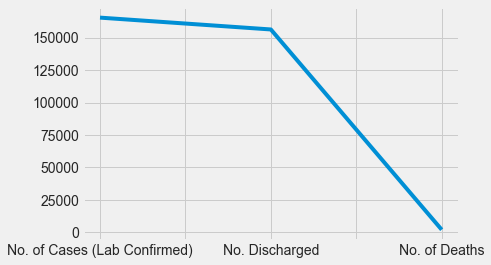


Figure : Line plot of the metrics of COVID-19 cases recorded in the country

From Figure 6, we learnt that a small percentage of the number of cases (lab confirmed) resulted in death, whereas the majority were discharged after treatment.

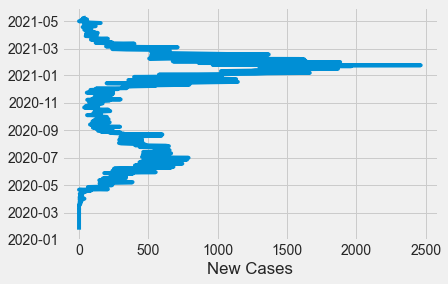


Figure : Line plot of the general overview of COVID-19 infection rate.

From Figure 7, we learn that the all-time maximum infection rates were recorded between December 2020 and January 2021. But the infection rate went on a decline shortly afterwards.

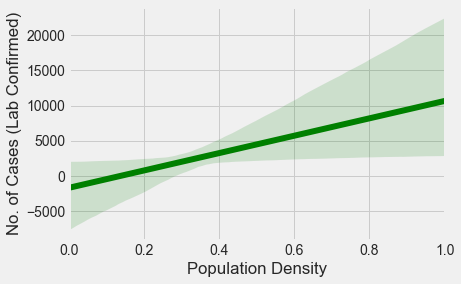


Figure : Regression plot of the linear relationship between the population density and the number of lab confirmed COVID-19 cases.

Figure 8 shows that the Population Density of each State is proportional to the total number of COVID-19 cases (lab confirmed) in each State. Therefore, the higher the Population Density on a scale of 0 to 1 the higher the chances of higher confirmed cases.

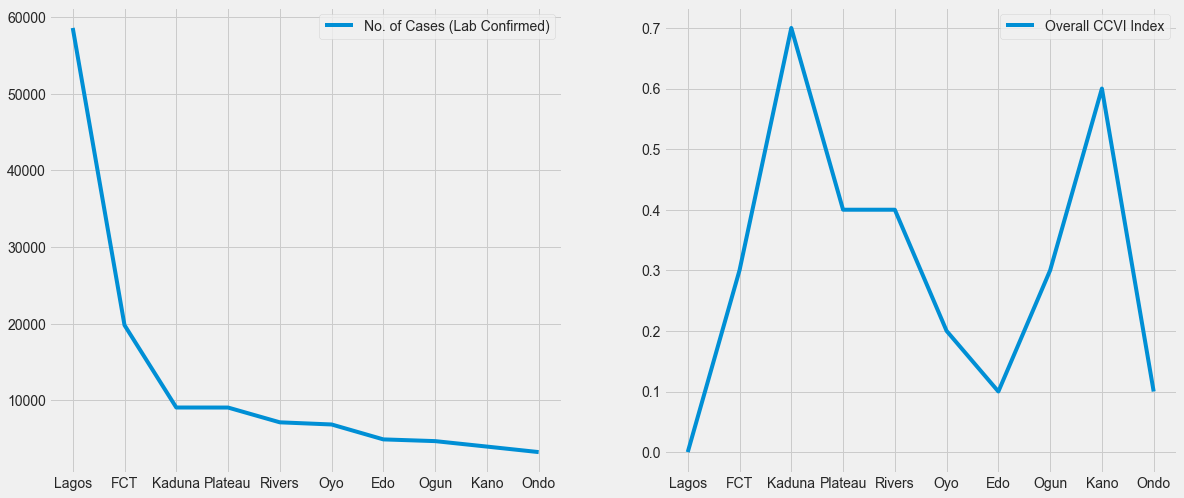


Figure : Line plots of the number of cases (left) against the Overall COVID-19 Community Vulnerability Index Scores (right) of the top ten states with the highest number of confirmed cases.

Figure 9 shows that Kaduna and Kano State, though they haven't recorded relatively high number of cases yet, they have the highest overall CCVI Index score – 0.7 and 0.6 respectively Indicating that the COVID-19 outbreak would have the most impact in the states - if the right measures aren't put in place.

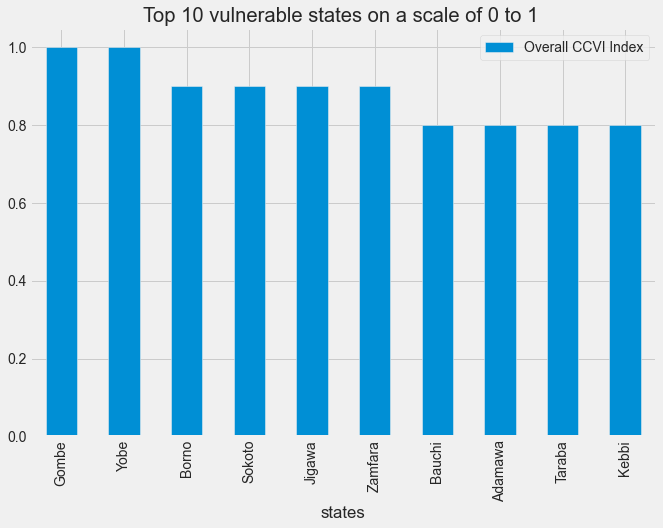


Figure : Bar plot of the top ten states with the highest Overall CCVI Index.

Figure 10 shows that Gombe and Yobe State had the highest Overall CCVI Index score – a measure of the impact of the virus on a community after the virus arrives; and that the ten most vulnerable states are from the Northern region.

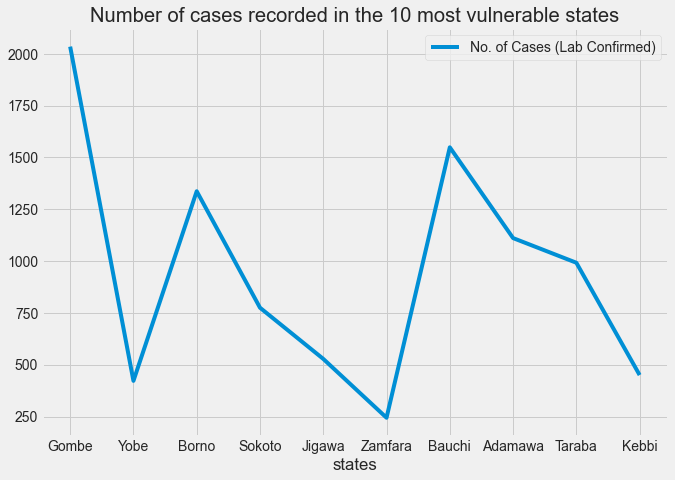


Figure : Line plot of the number of cases recorded in the most vulnerable states.

From Figure 11, the number of cases recorded in these most vulnerable states doesn’t reflect the danger looming – could be because the COVID-19 outbreak is more recent in these states or it could be that proper measures were put in place, it could also mean that the numbers declared were false.

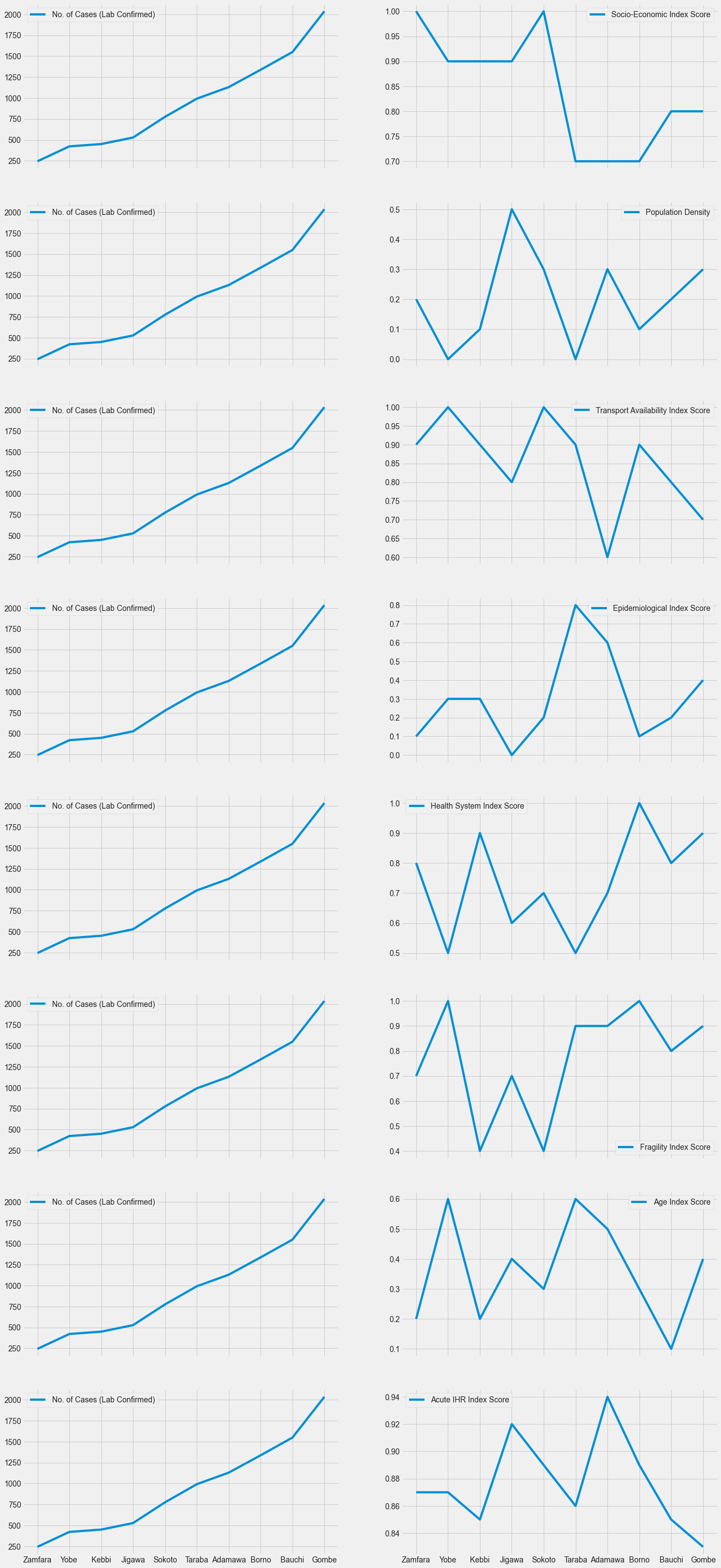


Figure : Line plots of the number of cases recorded in each state (left) against their index scores on the themes (right)

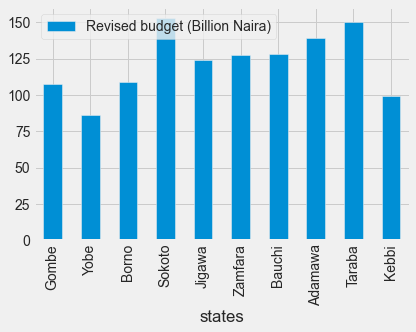
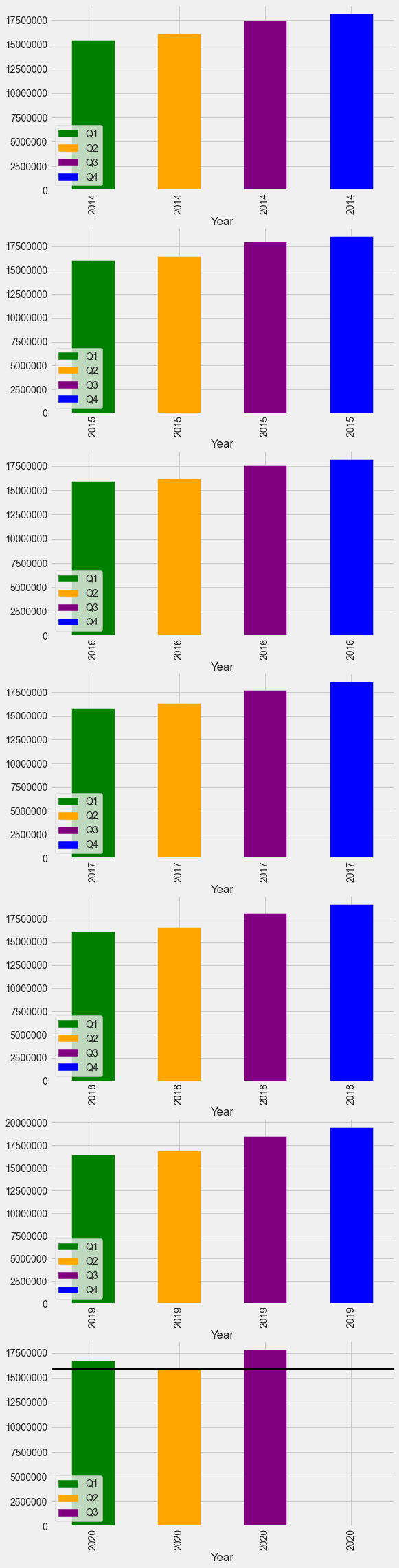


Figure : Bar plot of the revised budget declared in the most vulnerable states.

From figure 12, we see the alarming themes that were responsible for the high Overall index scores. These themes would inform governments, donors, organizations, and companies to allocate resources effectively, to enable more targeted, sector-specific interventions to mitigate the impact of the pandemic in these states.

From Figure 13, it was discovered that Yobe State had the least budget, coupled with the fact that the State had the highest Overall CCVI Index Score indicates that they are the most vulnerable since they might not be able to finance the setting up of proper measures to mitigate the impact of the COVID-19 outbreak.

 Figure : Bar plots of the gross domestic product of Nigeria

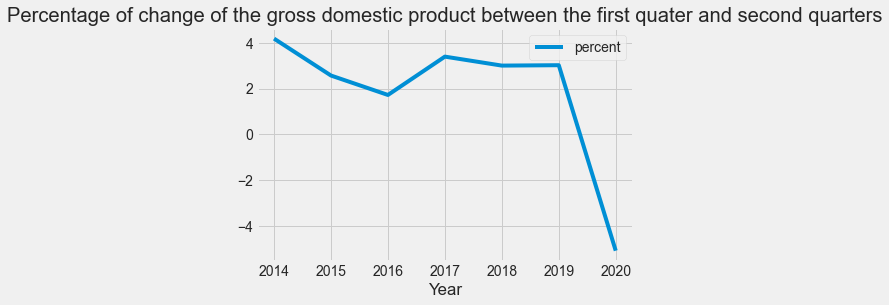


Figure : Line plot of the percentage of change of the gross domestic product.

From Figure 14, it was observed that the Gross Domestic Product of the country contracted in the second quarter of 2020 indicating that amid low oil prices, the COVID-19 outbreak hit the country hard. Moreover, from Figure 15 it was observed that the GDP of Nigeria reduced by around 5% from Q1 to Q2.

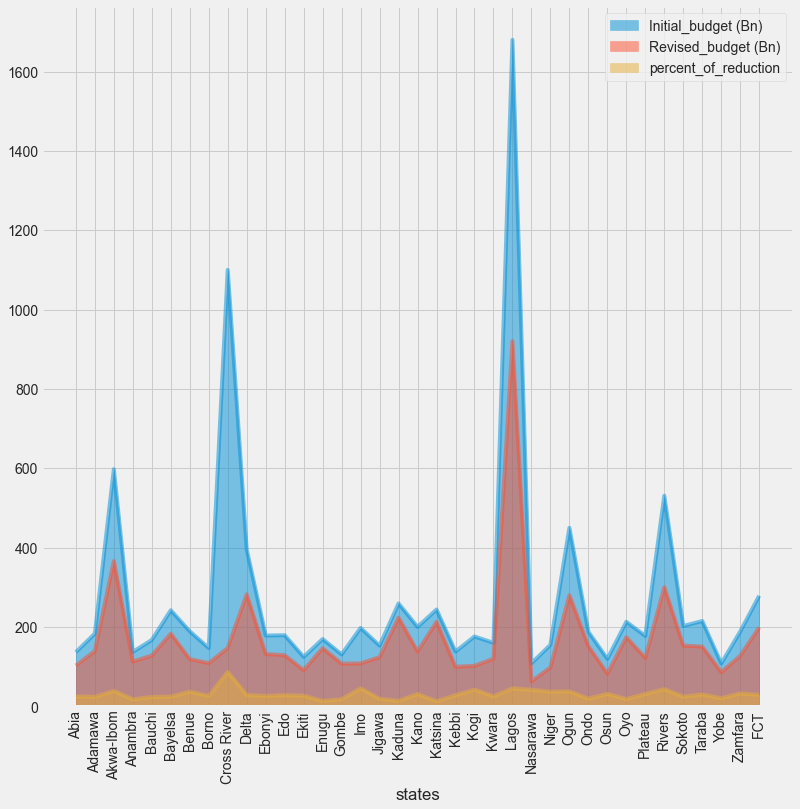


Figure : Area plot of the initial budget, revised budget, and percentage of budget reduction of all 37 states in Nigeria during the COVID-19

Figure 16 shows the budget declared initially by all states in Nigeria and their revised budgets after the outbreak, it was then discovered that their budgets were reduced by an average of about 30% due to the COVID outbreak and low oil prices.

1. **Conclusion**

Before analysis was done on datasets answers to questions like:

* How the states rank in terms of total confirmed cases, total recovered cases, total death cases, the percentage of the confirmed cases that died as a result of the disease in the states?
* When were the highest infection rate recorded, maximum daily infection rate (plus the date it was recorded)?
* Is there a linear relationship between number of cases recorded and population density?
* What alarming factors could enhance the impact of the pandemic in each state?
* How can governments, donors, organizations, and companies allocate resources effectively to enable more targeted, sector-specific interventions to mitigate the impact of the pandemic in these states?
* Which state is the most vulnerable?
* Did the outbreak have effect on Nigeria’s economy?

Weren’t clear but after the analysis, I found that:

* Lagos State recorded the most number of cases (rank 1), more than two times the number of confirmed cases in FCT (the State with the second highest number of confirmed cases); nine (the top nine) of the top ten retained the same position on both the recorded cases and the discharged cases (Ondo State being the one that didn’t make the cut – reason being that their patients were still undergoing treatment); On the discharged cases rank the changes worth mentioning are : Lagos State retained their position as first, Edo State moved from number seven to number two – indicating relatively higher death rate, Plateau State went from fourth to tenth – indicating relatively lower death rate.
* The all-time maximum infection rates were recorded between December 2020 and January 2021, But the infection rate went on a decline shortly afterwards. In addition, the maximum infection rate for a day (2,464) was recorded on the 23rd of January, 2021.
* The higher the Population Density the higher the chances of higher confirmed cases.
* The alarming themes in Zamfara State were: Acute IHR (High), Fragility (High), Health System (High), Socioeconomic Status (Very High), Transport Availability (High).
* The alarming themes in Yobe State were: Acute IHR (High), Fragility (Very High), Socioeconomic Status (High), Transport Availability (Very High).
* The alarming themes in Kebbi State were: Acute IHR (High), Health System (High), Socioeconomic Status (High), Transport Availability (High).
* The alarming themes in Jigawa State were: Acute IHR (High), Fragility (High), Socioeconomic Status (High), Transport Availability (High).
* The alarming themes in Sokoto State were: Health System (High), Socioeconomic Status (Very High), Acute IHR (High), Transport Availability (Very High).
* The alarming themes in Taraba State were: Transort Availability (High), Epidemiological Factors (High), Socioeconomic Status (High), Acute IHR (High), Fragility (High).
* The alarming themes in Adamawa State were: Acute IHR (Very High), Fragility (High), Health System (High), Socioeconomic Status (High).
* The alarming themes in Borno State were: Acute IHR (High), Fragility (Very High), Health System (Very High), Socioeconomic Status (High), Transport Availability (High).
* The alarming themes in Bauchi State were: Socioeconomic Status (High), Acute IHR (High), Fragility (High), Health System (High), Transport Availability (High).
* The alarming themes in Gombe State were: Transport Availability (High), Socioeconomic Status (High), Health System (High), Fragility (High), Acute IHR (High).
* The most vulnerable state in the country was Yobe State.
* COVID-19 had effect on the economy – negative in fact, because the GDP of the country dropped by around 5%, and states reduced their budget by an average of about 30%. Hinting that there was an abnormal reduction in the total monetary or market value of all the finished goods and services produced within the country’s borders in the second quarter of 2020.

**Recommendation**

Governments, donors, organizations, companies, media, and advocacy groups can help mitigate the impact of the pandemic by addressing the subthemes (see Appendix 1) the alarming themes were computed from, for more targeted sector-specific interventions that would help mitigate the impact of the pandemic in these states.

***What Next?***

Data on when the first cases were recorded in the states, travel and tourism data would help improve the current results.

1. **References**
   1. NCDC (2021). Confirmed Cases by State. <https://covid19.ncdc.gov.ng/>
   2. Johns Hopkins University Centre for Systems Science and Engineering (JHU CSSE). <https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv>
   3. Data Dictionary. <https://docs.google.com/document/d/1BqQvTH-dV1xD5QWoMHlcY0PQC39RZ7-0hjqccozaG9k/edit?usp=sharing>
   4. Ustacky Data Science Capstone Project starter code and datasets. <https://github.com/Ustacky-dev/Nigeria-COVID-19-Data-Analysis-Using-Python>
   5. How to use Git and GitHub. <https://www.analyticsvidhya.com/blog/2020/05/git-github-essential-guide-beginners/>
   6. How to write a Professional Summary. <https://docs.google.com/document/d/1QNb0MKhfqcgqjvDSn5wuJ1OyjM7bHtOEyDJcuMbG7vY/edit?usp=sharing>
   7. Africa CCVI methodology. <https://covid-static-assets.s3.amazonaws.com/Africa+CCVI+methodology.pdf>
   8. Benjamin Obi Tayo, Ph.D., DataScienceHub (2021). How to organize your data science project. <https://www.kdnuggets.com/2021/04/how-organize-your-data-science-project-2021.html>
   9. The Kitchen Research Group, Chemical Engineering at Carnegie Mellon University. Plotting two datasets with very different scales (2013). <http://kitchingroup.cheme.cmu.edu/blog/2013/09/13/Plotting-two-datasets-with-very-different-scales/>
   10. Seaborn Official Documentation. Regplot. <https://seaborn.pydata.org/generated/seaborn.regplot.html>
   11. Pandas Official Documentation. <https://pandas.pydata.org/docs/>
   12. Stackoverflow (2019). How to plot pandas bar plot with sequential colormap. <https://stackoverflow.com/questions/53806207/how-to-plot-pandas-bar-plot-with-sequential-colormap>
   13. Matplotlib Official Documentation. <https://matplotlib.org/>
2. **Appendices**

Appendix 1: Themes and subthemes

|  |  |
| --- | --- |
| **Africa CCVI themes** | **Africa CCVI subthemes** |
| Theme 1:  Socioeconomic Status | Access to information |
| Education |
| Poverty |
| Unemployment |
| Theme 2:  Population Density | Population density |
| Theme 3:  Housing type &  Transportation | Access to transportation |
| Connectivity by road |
| Crowding in household |
| Improved housing |
| Sanitation |
| Theme 4:  Epidemiological Factors | HIV |
| Other infectious diseases |
| Obesity |
| Diabetes |
| Hypertension |
| Smoking |
| Theme 5:  Health System Factors | Health Facilities per capita |
| Access to healthcare systems |
| Healthcare system performance |
| Theme 6  Fragility | Civil Unrest |
| Population of concern sites |
| Food insecurity |
| Theme 7:  Old age | Old Age |
| Theme 8: Acute IHR | Percentage of people who got infected that are expected to require  acute care |

Appendix 2: Table of Figures

[Figure 1: Horizontal bar plot of the top 10 states measured by the number of lab confirmed COVID-19 cases recorded 3](#_Toc72089749)

[Figure 2: Horizontal bar plot of the top 10 states measured by the number of COVID-19 patients that were discharged due to recovery 3](#_Toc72089750)

[Figure 3: Horizontal bar plot of the top 10 states measured by the number of COVID-19 patients that are under care. 4](#_Toc72089751)

[Figure 4: Horizontal bar plot of the top 10 states measured by the number of COVID-19 caused deaths that were recorded 4](#_Toc72089752)

[Figure 5: Bar plot of the percentage of the confirmed COVID-19 cases that resulted in death among the top ten states (measured by cases recorded). 4](#_Toc72089753)

[Figure 6: Line plot of the metrics of COVID-19 cases recorded in the country 5](#_Toc72089754)

[Figure 7: Line plot of the general overview of COVID-19 infection rate. 5](#_Toc72089755)

[Figure 8: Regression plot of the linear relationship between the population density and the number of lab confirmed COVID-19 cases. 6](#_Toc72089756)

[Figure 9: Line plots of the number of cases (left) against the Overall COVID-19 Community Vulnerability Index Scores (right) of the top ten states with the highest number of confirmed cases. 6](#_Toc72089757)

[Figure 10: Bar plot of the top ten states with the highest Overall CCVI Index. 7](#_Toc72089758)

[Figure 11: Line plot of the number of cases recorded in the most vulnerable states. 7](#_Toc72089759)

[Figure 12: Line plots of the number of cases recorded in each state (left) against their index scores on the themes (right) 8](#_Toc72089760)

[Figure 13: Bar plot of the revised budget declared in the most vulnerable states. 9](#_Toc72089761)

[Figure 14: Bar plots of the gross domestic product of Nigeria 10](#_Toc72089762)

[Figure 15: Line plot of the percentage of change of the gross domestic product. 11](#_Toc72089763)

[Figure 16: Area plot showing the initial and revised budget of all 37 states in Nigeria during the COVID-19 12](#_Toc72089764)