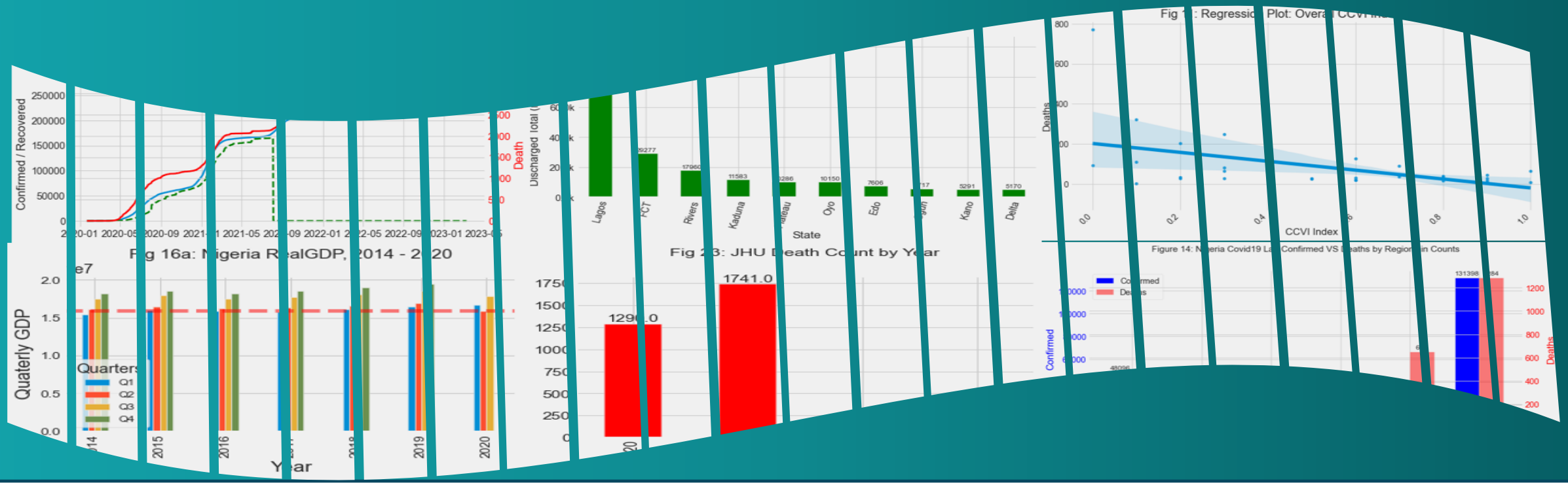
NIGERIA COVID19 DATA ANALYSIS USING PYTHON

BY PAULA UKERUN

FOR USTACKY DATA SCIENCE MICRODEGREE CAPSTONE PROJECT



EXECUTIVE SUMMARY

- This is a summary report for Ustacky Microdegree in Data Science Capstone Project analyzing the effect of COVID-19 on the Nigerian population and economy in terms of infection rate, mortality, economic adjustments, and global to local record comparison. The project included extracting, loading, and transforming the data to prepare it for optimal data quality. Using the ‘TODO’ lists, a series of analyses was carried out on the data, and further analysis to enhance the quality of the results was also carried out. A zip file containing the work process using Python based Jupyter notebook was submitted along with the executive summary report.

DATA EXTRACTION, LOADING AND TRANSFORMATION

- The datasets were read in from links (provided by UStacky for the project) from UStacky and John Hopkins University GitHub repository pages:
  - UStacky Supporting Files
  - JHU COVID files (the global confirmed, death and recovered records sere used)
- Other supporting files used in this project include NCDC data and WHO global data
  - NCDC data path – this webpage was not opening as at the week of September 16, 2023. An earlier downloaded version was uded instead.
  - WHO COVID file

DATA EXTRACTION, LOADING AND TRANSFORMATION

- The suggested import libraries by UStacky guide were read using Jupyter notebook (pandas) and additionally the 'io' library was also read into the environment.
- Using both the direct URL pandas read and request.get() read for the files (all csv files) was effective in reading the data into dataframes;
  - df_url_death = pd.read_csv(url) -------- option 1

    response = requests.get(url) -------- option 2

    file_object = response.content

    df_url_confirmed = pd.read_csv(io.BytesIO(file_object))
- The data read in was viewed using the .head() function, .info() function, .describe() function and .shape

DATA EXTRACTION, LOADING AND TRANSFORMATION

- From the data visualization, it was observed that the JHU data (3) had similar shape for the confirmed and death records (289 rows and 1147 columns), while the recovered file had a different shape (274 rows and 1147 columns).
- From the analyses above, it is obvious that the data is a wide data which is a less convenient form of data for analyses and visualization. Hence the JHU files were transformed to long data using pandas .melt() :
  - df_recovered_long = pd.melt(df_url_recovered, id_vars=['Country/Region', 'Lat', 'Long'], var_name ='Date', value_name='Recovered')
- The Province/Region column had lots of NANs and so was dropped/deleted using .dropna() as it does not have any essential significance in the analysis. Header for the value counts were name Confirmed, Death and Recovered accordingly.
- After the transformation to long and NANs drop, the confirmed and death record files had 328041 rows × 5 columns, while the recovered file had 312039 rows × 5 columns.

DATA EXTRACTION, LOADING AND TRANSFORMATION

- The datasets from JHU was merged to contain records of confirmed, death and recovered in same file as in the WHO and NCDC record.
- The merge was done in two steps, an inner merge of the confirmed and death records and a left merge of the merge1 and recovered record (since the recovered files is least accurate / populated). Pandas merge function was used as shown below:
  - merged_df_confirmed_death = pd.merge(df_confirmed_long, df_death_long, on=['Lat', 'Date'], how='inner')
- The JHU data contained cumulative counts per data. New columns to calculate the difference from each consecutive day difference was used. The name of the daily count columns were; Confirmed_new, Death_new and Recovered_new. The final merge dataframe had a shape of 340614 rows and 10 columns.
- The result was QC'ed and transformed further using head(), sum(), .info(), .describe(), .drop_duplicates(), .rename(), .shape, .astype(), .datetime() etc.

EXTERNAL DATA ANALYSES

- The external data were read using their URLs as already discussed. The file contained records of vulnerability index, population and social infrastructural accessibilities for the 36 States and the FCT of Nigeria.
- The real GDP file consisting of 2014 to 2020 quarterly GDP of Nigeria.
- The 2020 budget file contained initial and revised budgets for all Nigeria States and FCT for the year 2020.
- The COVID data file also contains information on the COVID19 impact in terms of lab confirmed, death and discharged / recovered counts. All the columns except death records were in object type (some having commas) and so the .replace() and .astype('int64') were used to transform the value columns for accuracy.
- The WHO data read in was also cleaned to remove non-essential columns for this analysis, and to rename some columns. The WHO data was already a long data and the file was renamed to match the column names of the JHU data. data structure of the WHO dataframe was 315590 rows and 6 columns.

TO DO LIST PART 1 : EXTRACTING DAILY COUNTS FROM JHU DATA

- The JHU daily record was calculated from differences between consecutive days records.
- The data was first sorted so that the records were ordered by Country to Province to Data. Since the province was deleted due to NANs records, the Latitude and Longitude (Lat and Long columns) were used instead as some countries had repeated day records where they have multiple provinces e.g. China, Russia, USA, Australia etc. Script used to sort:
  - merged_df_CDR.sort_values(by=['Country','Lat', 'Long', 'Date'], inplace=True)
- The difference between records were calculated using example below:

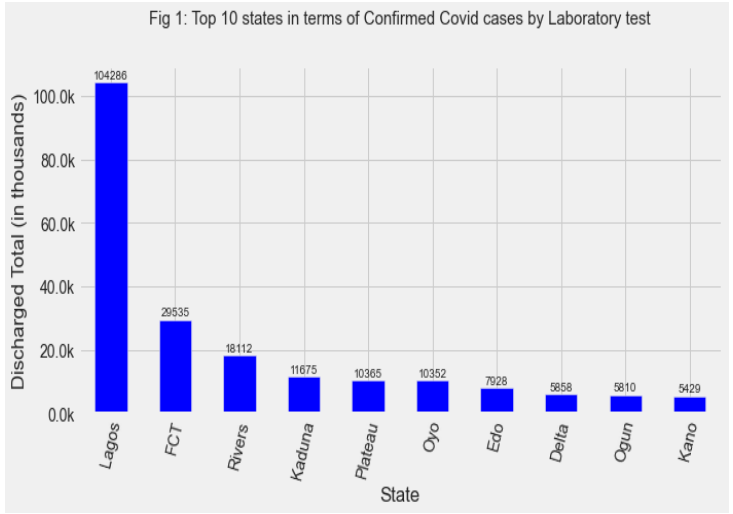
```
df_jhu_sorted['Confirmed_new'] = df_jhu_sorted.groupby(['Country', 'Lat', 'Long'])['Confirmed'].diff()

df_jhu_sorted['Confirmed_new'].fillna(df_jhu_sorted.groupby(['Country', 'Lat', 'Long'])['Confirmed'].shift(1) - df_jhu_sorted['Confirmed'], inplace=True)
```

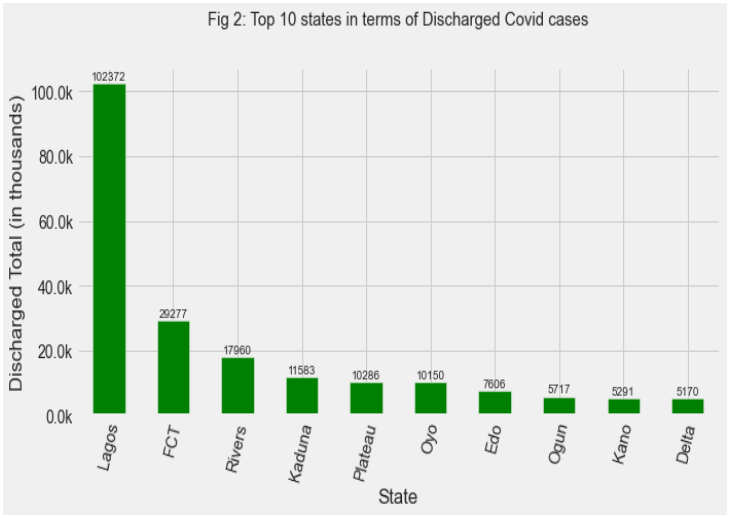
- Values less than zero were set to zero using .clip()
- The Nigeria records were extracted from WHO and JHU global records using the .get_group()

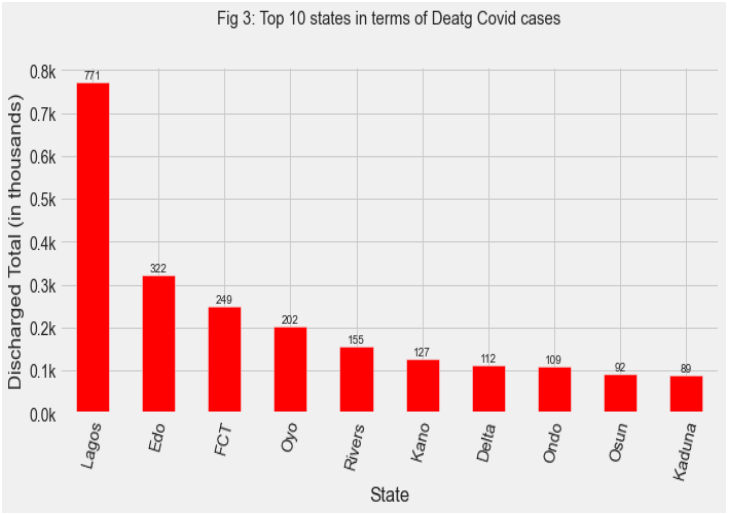
TO DO LIST PART 2: ANALYSIS

Analysis of the Top 10 counts of COVID19 confirmed, discharged and death records for the Nigeria Centre of Disease record, shows that the more industrialized and Urban states had higher counts.

Lagos State has the most counts for laboratory confirmed, discharged / recovered and death records.



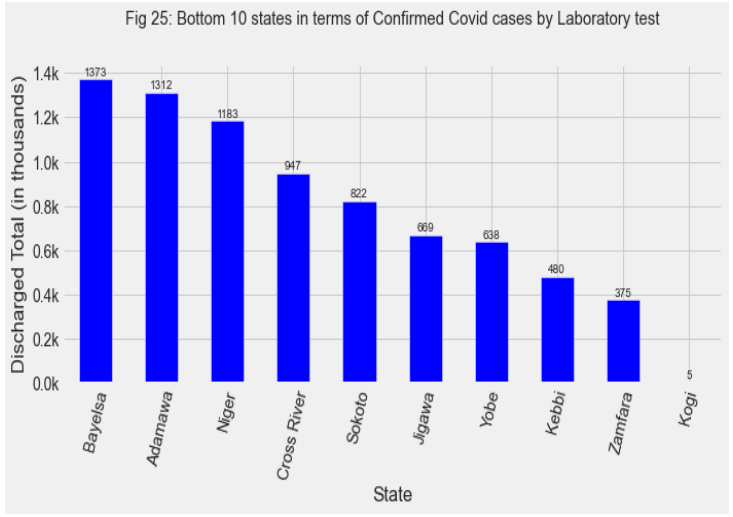


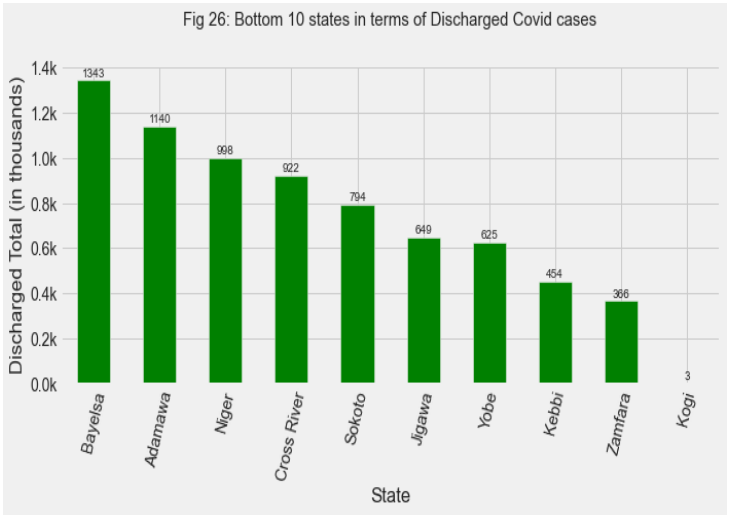


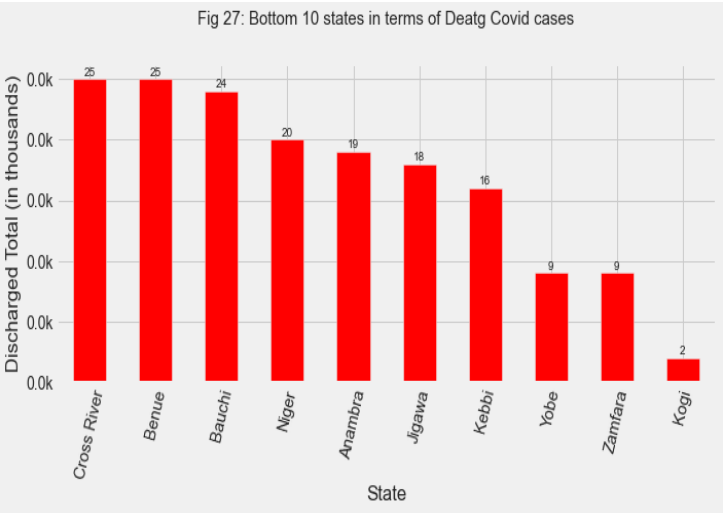
TO DO LIST PART 2: ANALYSIS

Analysis of the Bottom 10 counts of COVID19 confirmed, discharged and death charts (fig 25 – 27) for the Nigeria Centre of Disease record, show that most of the bottom 10 states are less central states.

However, some commercial states like Bayelsa, Cross River and Kogi (a central transit state) are amongst the lowest 10 affected states.





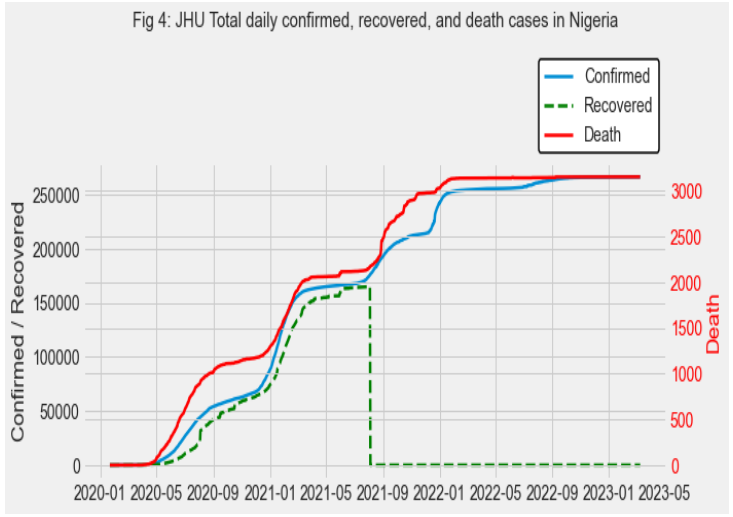


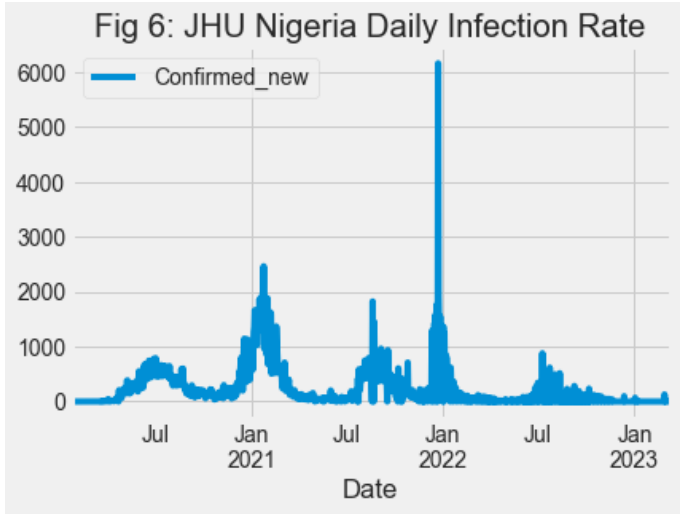
TO DO LIST PART 2: ANALYSIS

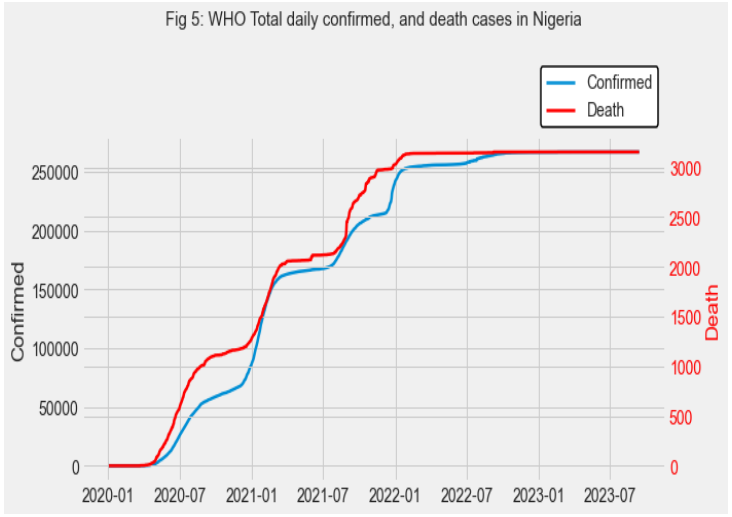
Line plot charts (fig 4 and 5) of Nigeria COVID19 records extracted from the cleaned JHU and WHO files shows correlation in the trend between the confirmed and death cases. The JHU stopped recording data on recover between August and September of 2021.

Date of Highest infection rate in Nigeria based on JHU data is: Dec 22, 2021 with a count of 6158 confirmed cases (figure 6) .

Date of Highest infection rate in Nigeria based on WHO data is: Dec 23, 2021 with a count of 4035 confirmed cases .



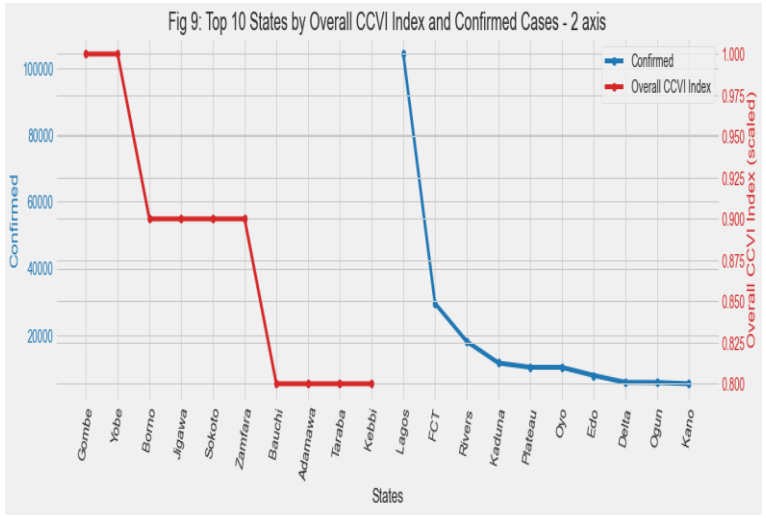


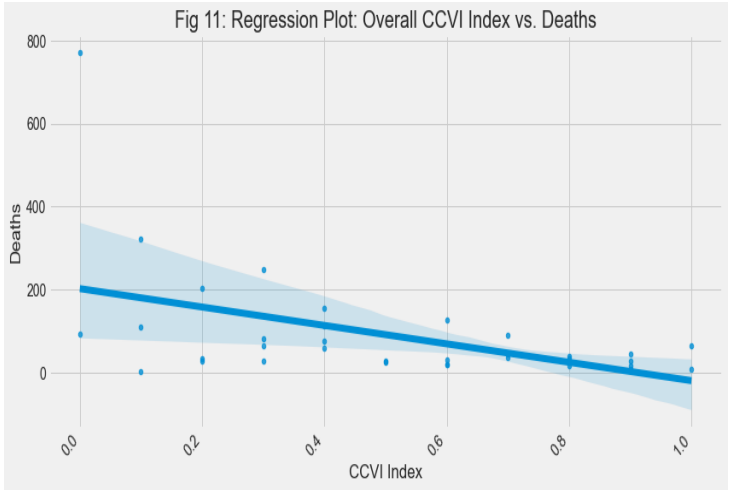


TO DO LIST PART 2: ANALYSIS

Merge of the NCDC data with the eternal data was used to plot line chart of Top 10 Overall CCVI Index vs To 10 Confirmed COVID19 Cases States. The resulting states were different as Top 10 confirmed were more urban States while Top 10 CCVI were more rural States.

The Fig 10 regression curve shows that there is a direct relationship between COVID19 and increasing cases from the increase of confirmed cases with increasing population density.
A regression plot of Fig 11 shows an inverse relationship. Possibly because the rural areas had a better management of the pandemic.

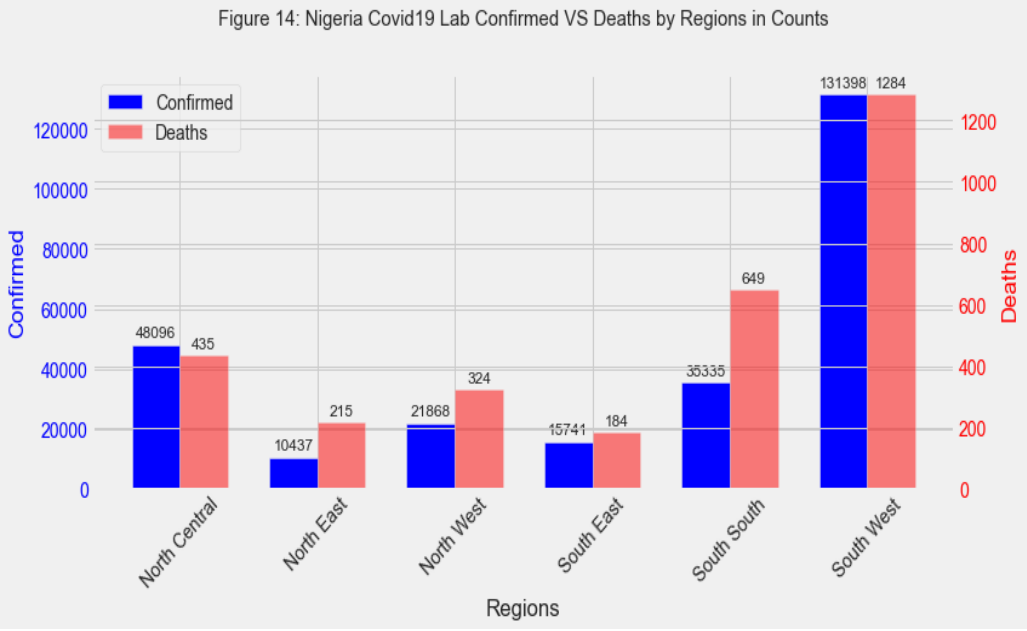


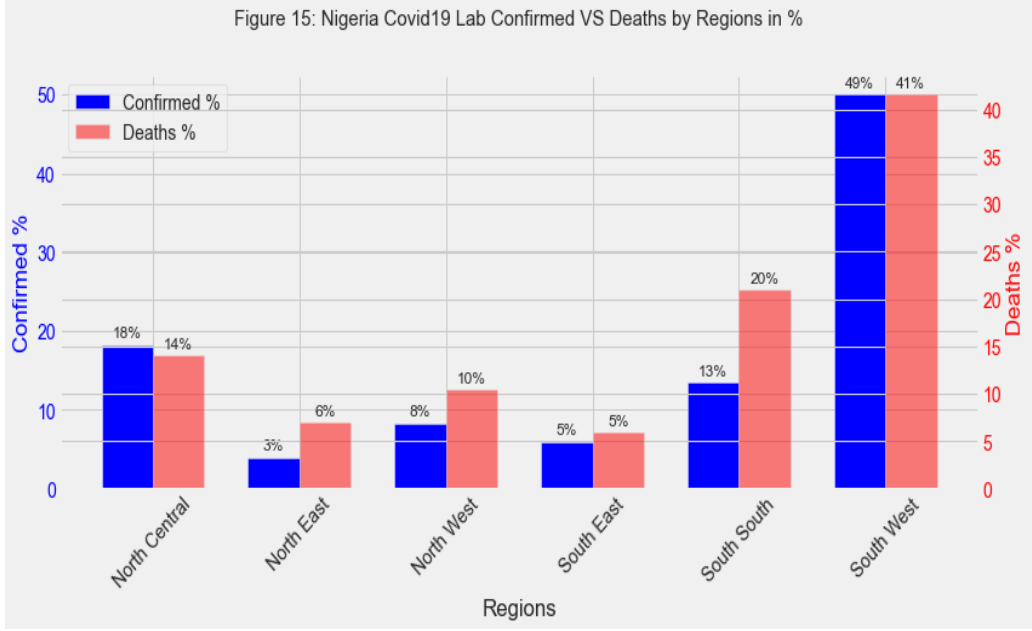


TO DO LIST PART 2: ANALYSIS

Merge of the NCDC data with the eternal data was used to plot line charts of Fig 14, shows that the Highest Confirmed and Death records are from the South West and that the Lowest Death Count is from the South East, while the Lowest Confirmed count is from the North East.

Fig 15 show that the South East has 49% of overall Confirmed cases in Nigeria and 41% of overall Death cases.
South East has the lowest Death percentage of 5% and North East has the lowest Confirmed of 3% of Nigeria total cases.

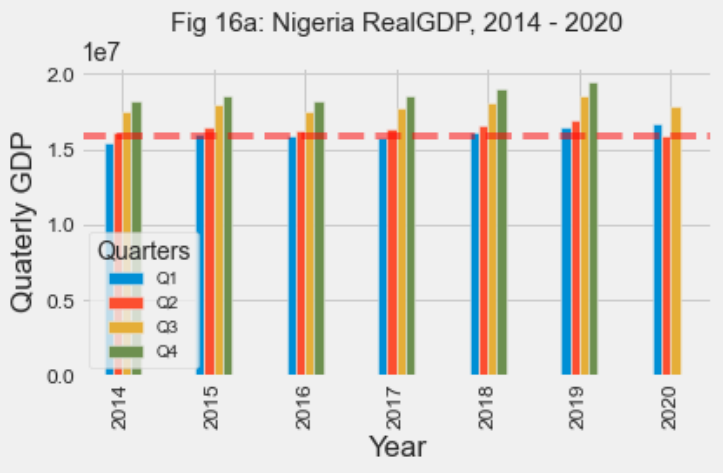


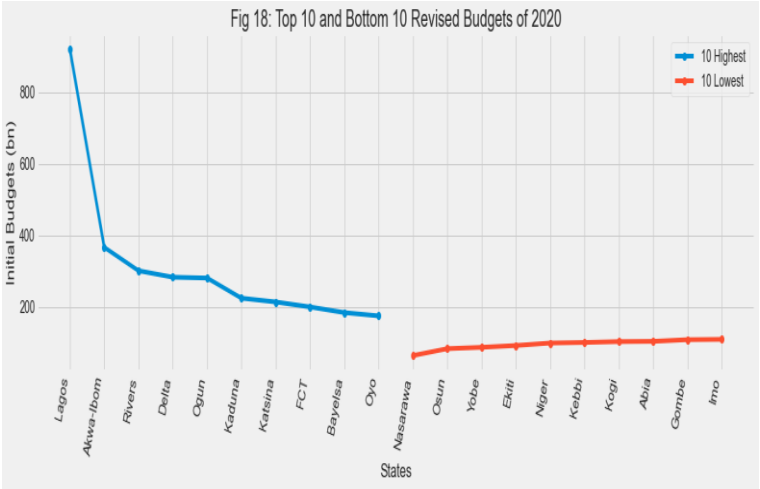


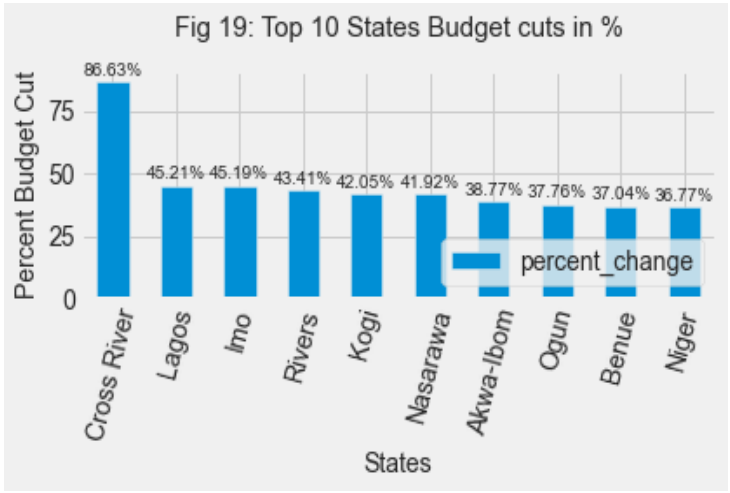
TO DO LIST PART 2: ANALYSIS

Fig 16 shows steady increase of GDP every quarter from 2014 except in the Q2 of 2020 that correspond to time of global quarantine. This appears to have impacted the GDP
The increase of the GDP in the 3rd Quarter that is comparable with past years indicates good adjustment policy or less impact on the Nation as correlates with the percentage death rate in the nation.

Fig 18 and 19 show the budget charts.
Fig 18 shows the top 10 and bottom 10 budgets after cuts.
Fig 19 shows the top 10 percentage cuts from initial budgets. Cross River has the highest cut of 86.63% which removed it from initial top 10 budget perhaps due to it low COVID19 cases shown in fig 25 -27.



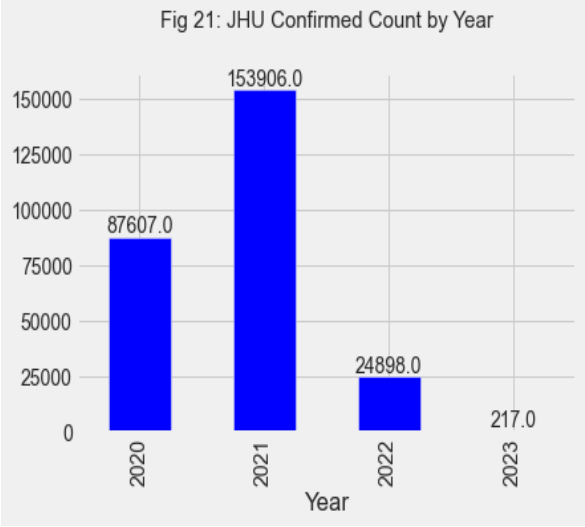


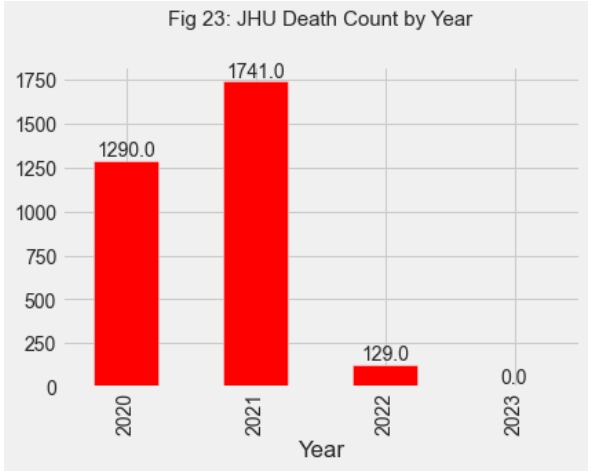


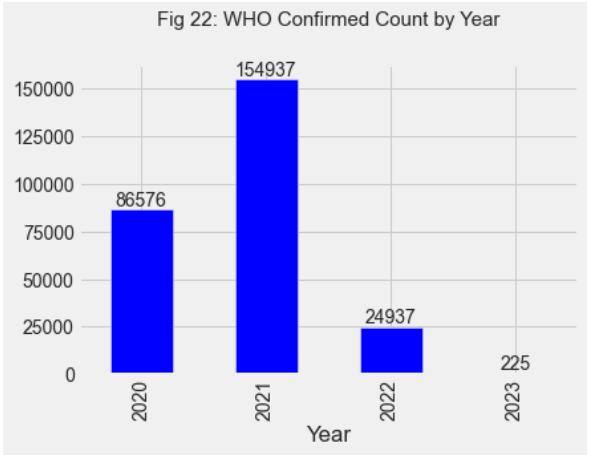
TO DO LIST PART 2: ANALYSIS

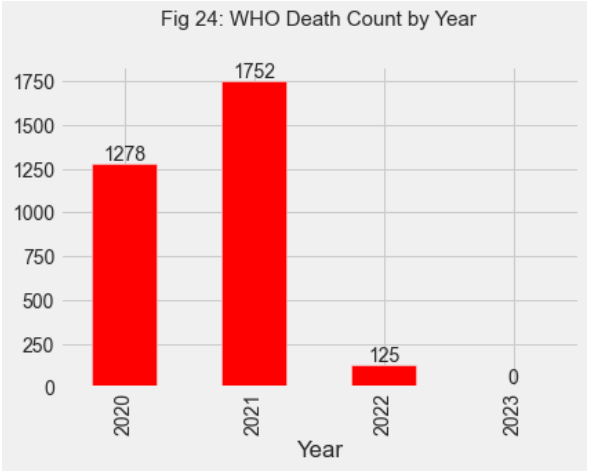
Nigeria records extracted from global records of JHU and WHO were plotted against the extracted year from the Date column.
Fig 21 and Fig 23 show that year 2021 had the most confirmed and death records of COVID19.
JHU had no death record for 2023.

Similarly for the Nigeria Confirmed and Death record from WHO; Fig 22 and 24 show that 2021 had the highest Confirmed and Death counts.
WHO had no death record as well for 2023.









CONCLUSIVE SUMMARY

- Total sum of Confirmed from JHU record is 266628 compared to NCDC record of 262875 (less by 3753). This represents 0.04% of JHU global confirmed count of 671,095,349. Total Nigeria Death sum from JHU is 3160 compared to NCDC record of 3091 (less by 69). This represents 0.046% of global JHU death records of 6,806,448.
- Total sum of Confirmed from WHO record is 266675 compared to NCDC record of 262875 (less by 3800). This represents 0.035% of WHO global confirmed count of 770,563,467. Total Nigeria Death sum from WHO is 3155 compared to NCDC record of 3091 (less by 64). This represents 0.045% of global WHO death records of 6,957,216.
- From NCDC records, Lagos had the highest death count of 771 and Kogi had the lowest death count of 2.
- Regression cut shows inverse relationship between CCVI Index and Death case while showing a direct relationship between population density and confirmed / death cases.
- The Nigeria GDP was observed to suffer a significant fall the 2020 Q2, which correspond to the period of start of global quarantine and border managements.
- 2021 was the most impactful year of the COVID19 infection.