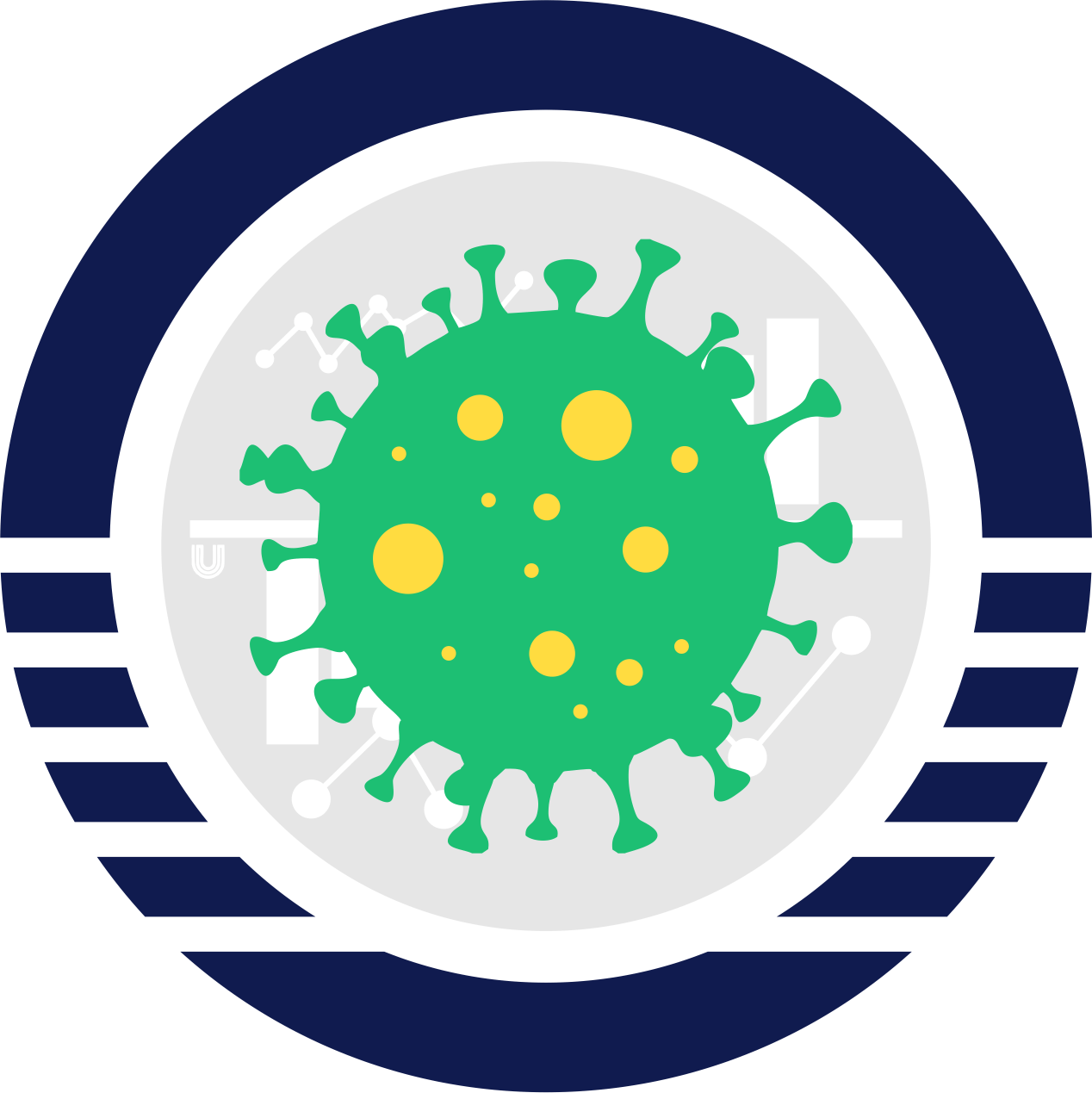
**HEART DISEASE PREDICTION ANALYSIS AND MODELLING**

****

**By**

**ADEPOJU Yusuf Abiodun**

**USTACKY DATA SCIENCE CAPSTONE PROJECT**

**Jan 2023**

INTRODUCTION

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

**Data Source**

This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

* Cleveland: 303 observations
* Hungarian: 294 observations
* Switzerland: 123 observations
* Long Beach VA: 200 observations
* Stalog (Heart) Data Set: 270 observations

At the end of this report, hidden facts about Heart disease will be revealed and a model for prediction for future prediction

**DATA OVERVIEW**

Attribute Information and Data

1.  Age: age of the patient [years]

2.  Sex: sex of the patient [M: Male, F: Female]

3.  ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]

4.  RestingBP: resting blood pressure [mm Hg]

5.  Cholesterol: serum cholesterol [mm/dl]

6.  FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]

7.  RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]

8.  MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]

9.  ExerciseAngina: exercise-induced angina [Y: Yes, N: No]

10. Oldpeak: oldpeak = ST [Numeric value measured in depression]

11. ST\_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]

12. HeartDisease: output class [1: heart disease, 0: Normal]

**METHODOLOGY**

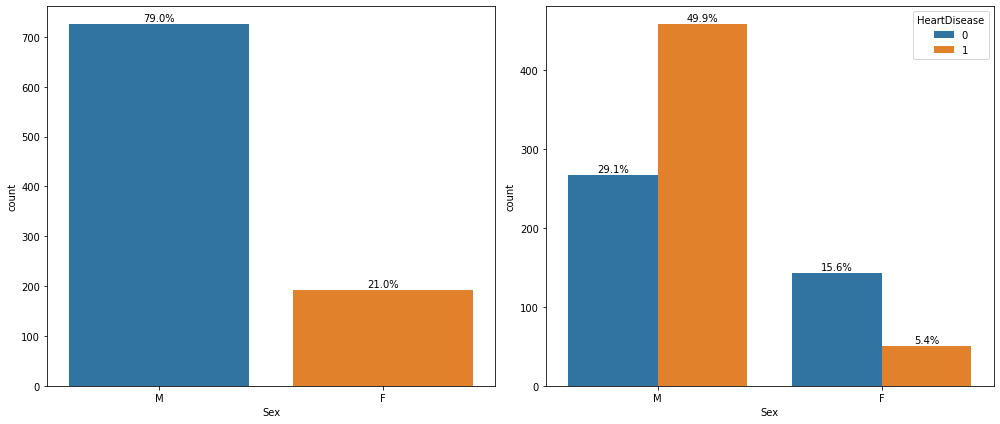
The analysis was carried out with the following method

* Data Cleaning and wrangling: Data cleaning is the process of removing incorrect or otherwise erroneous data from the dataset. These errors include incorrectly formatted data, redundant entries, mislabeled data, and null values. Data cleaning improves the quality of our data as well as any insights or decisions that you draw based on the data.
* Exploratory Data Analysis: Exploratory data analysis (EDA) is used to analyze and investigate data sets and summarize their main characteristics, also used in data visualization methods. It helps determine how best to manipulate data to get the answers to business problems, making it easier to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

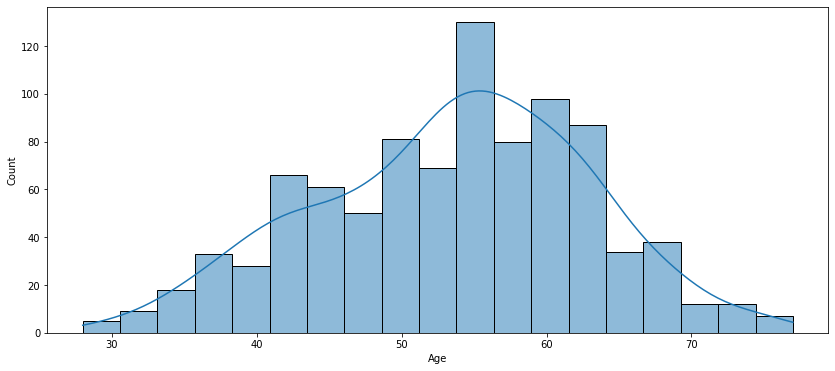
The analysis include:

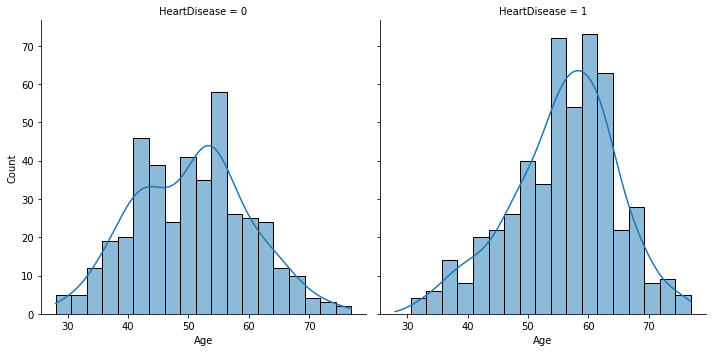
Distribution of Heart Disease by Sex

Approach: The whole patients and non-patients of heart disease are grouped based on sex

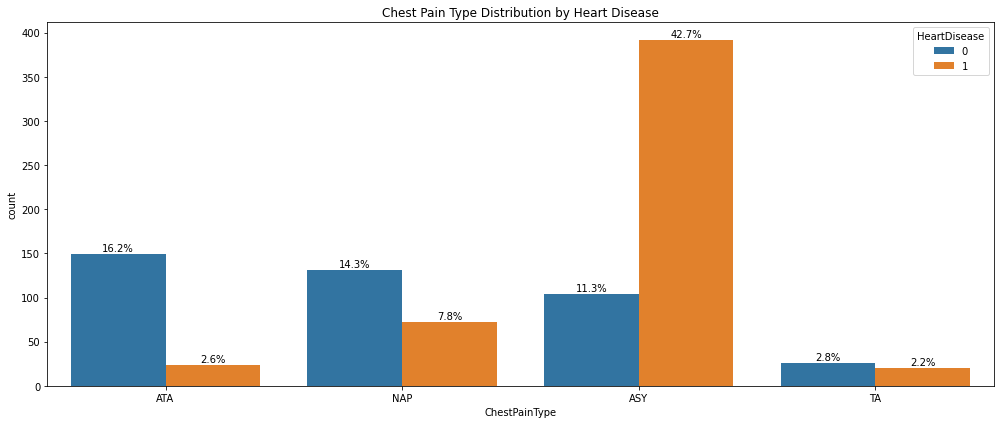


Age Distribution Analysis

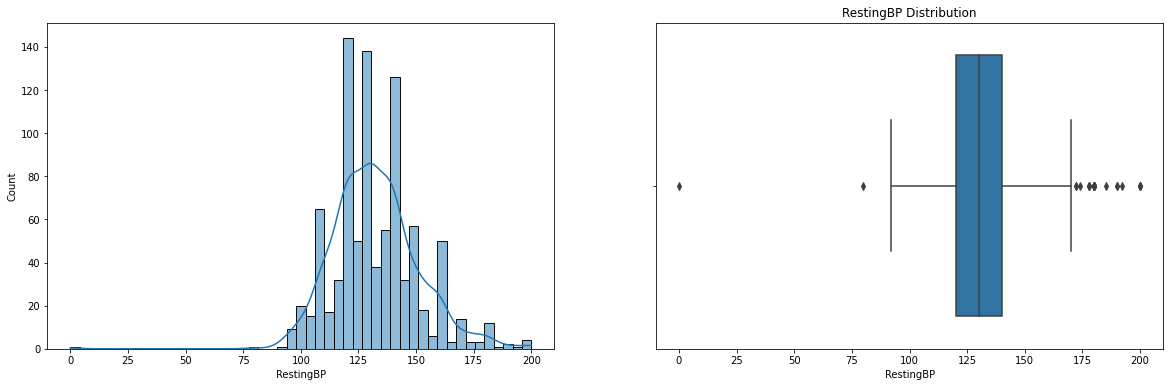




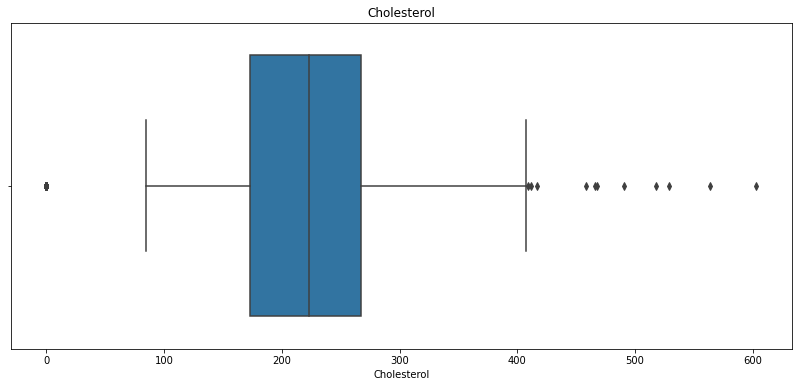
Chest Pain Type Distribution by Disease



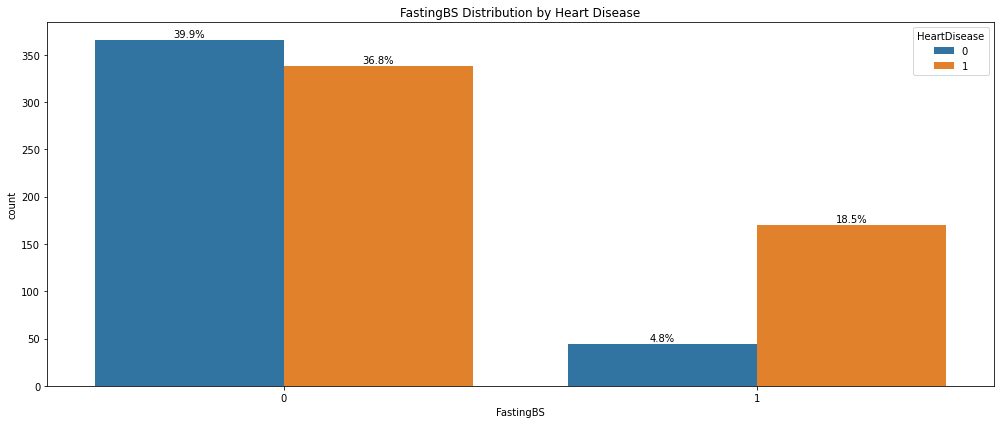
RestingBP Distribution



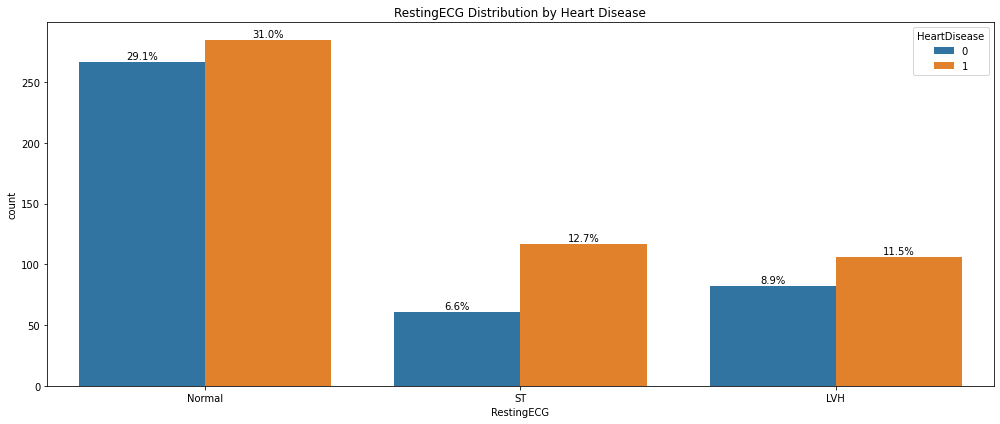
Cholesterol Distribution



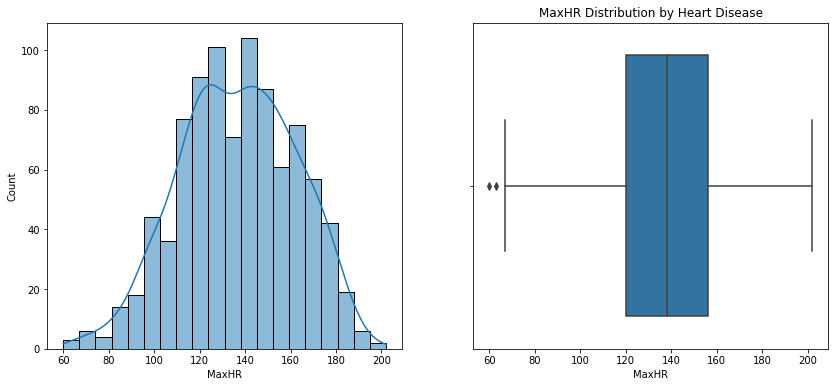
FastingBS Distribution by Heart Disease



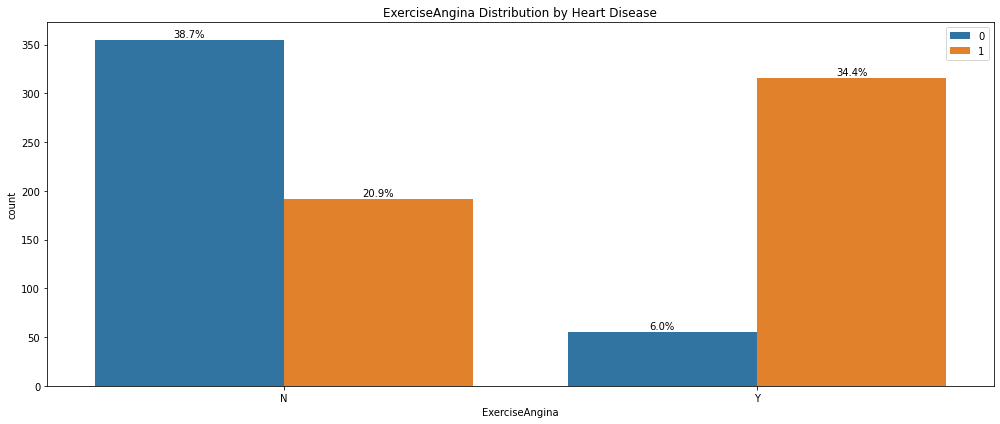
Resting Electrocardiogram

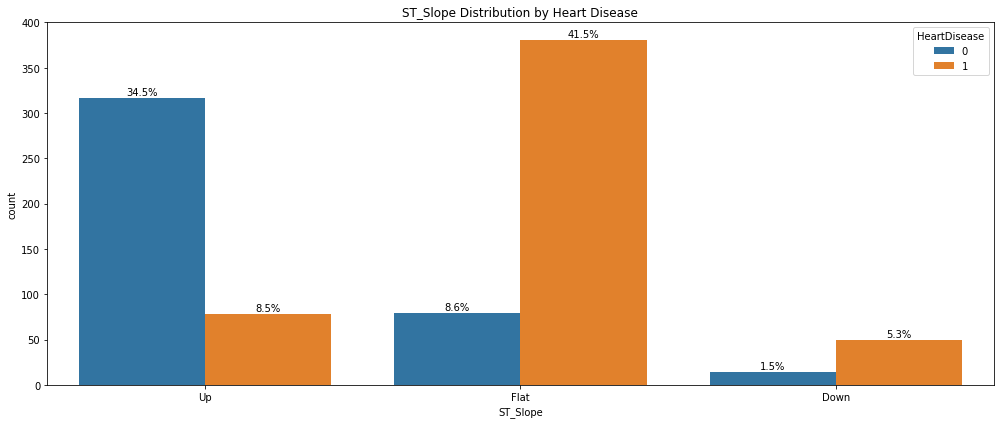


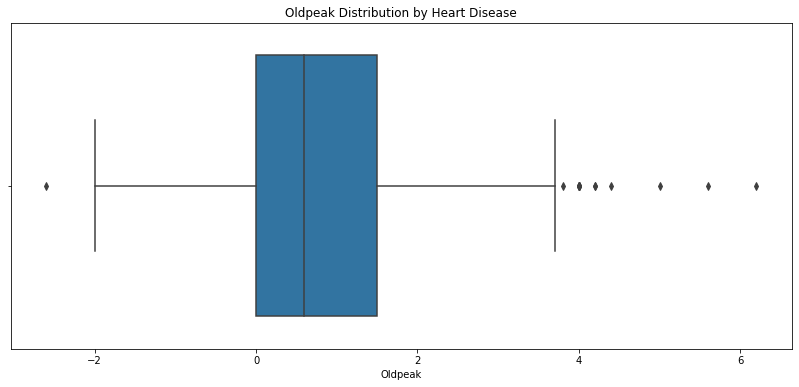
MaxHR Distribution by Heart Disease



Exercise Angina Distribution by Heart Disease







* Lagos state has the highest number of covid-19 cases, leading the second highest by FCT by 71% while Ogun and Kano are the 9th and 10th respectively.
* Lagos state has the highest number of discharged cases, leading the second highest FCT by 71% while Kano and Delta are the 9th and 10th state with discharged cases respectively.
* Lagos state has the highest number of death cases of covid-19 patients, leading the second highest Edo state by 58% while Osun and Kaduna state are the 9th and 10th state with death cases of covid-19 patients.

- On 2022-11-10 highest number of active confirmed cases in Nigeria was recorded with 6158 numbers of active patients, on 2022-11-10, 93 death was also recorded as the highest.

**CONCLUSIONS**

The analysis is carried out to investigate the cases of covid-19 in the country, emphasis are laid on top 10 states are and the following conclusions were drawn.

This analysis consists of explanatory and exploratory data analysis, the basic information was provided and overview of covid-19 was explored to identify missing gaps and areas where further analysis might be needed especially in this era of subsequent waves of covid-19 to make provision to ratify its effect.

Lagos state has the highest number of covid-19 cases, leading the second highest by FCT by 71% while Ogun and Kano are the 9th and 10th respectively.

Lagos state has the highest number of discharged cases, leading the second highest FCT by 71% while Kano and Delta are the 9th and 10th state with discharged cases respectively.

Lagos state has the highest number of death cases of covid-19 patients, leading the second highest Edo state by 58% while Osun and Kaduna state are the 9th and 10th state with death cases of covid-19 patients.

On 2022-11-10 highest number of active confirmed cases in Nigeria was recorded with 6158 numbers of active patients, 2022-11-10 number of death also recorded.

On 2022-11-10 highest number of active confirmed cases in Nigeria was recorded with 6158 numbers of active patients, on 2022-11-10, 93 death was also recorded as the highest.

Gombe and Yobe has the overall highest CCVI index of 1, followed by Sokoto, Jigawa, Zamfara, Borno with CCVI of 0.9, Akwa Ibom, Cross River, Lagos, Rivers are the states with highest difference in revised and initial budgets.

There is a fluctuation in GDP at first quarters over the years, rise across the second, third and fourth quarters with a sharp fall in 2020.