**A PROJECT WORK PRESENTED IN FULFILLMENT OF STATISTICAL ANALYSIS OF COVID-19 DATA**

**SAIL INNOVATION LAB**

**BY: GROUP 9**

**CHISOM CHIBUIKE**

**MAYOWA ORESAJO**

**BLESSING TASIBO**

**OMOLARA ADEBAMBO**

**IDOWU ADESANYA**

**OPEYEMI AMOBI**

**DATE: 6TH OF JULY, 2024**

**INTRODUCTION**

Covid-19 is an infectious disease caused by the SARS-CoV-2 virus which is very deadly and can easily spread. This analysis focuses mainly on the Regions(states) in Nigeria showing the death rate, total confirmed cases, new confirmed cases, total recoveries, new deaths and also recovered cases.

**AIM AND OBJECTIVES**

In this work, we aim to carry out hypothesis testing to compare the mean of various regions in order to determine if there is a significant difference between the means.

Our objectives are to;

1) Summarize the covid data by regions

2) Test each region against other to test if their average death number from COVID-19 is same

**THE DATASET**

The dataset, covid-19 dataset was downloaded from [ekoanalytics](https://ekoanalytics.net/data-catalogue.html) as an excel file and comprises 4100 rows, and 11 columns. The key variables include death rate, total confirmed cases, new confirmed cases, total recoveries, new deaths and also recovered cases. The data spans of this report covers the year 2020. Our dataset also has no null value and no duplicate values.

Our dataset comprises of the following columns in Table 1.1 below:

|  |  |
| --- | --- |
| Columns | Meaning |
| States | List of states or regions |
| total\_confirmed | List of total confirmed cases for each state |
| new\_confirmed | List of new confirmed cases for each state |
| total\_recoveries | List of total recoveries for each state |
| new\_recoveries | List of new recoveries for each state |
| total\_deaths | List of total deaths for each state |
| new\_deaths | List of new deaths for each state |
| week | List of weeks (assuming week number) |
| day | List of days |
| month | List of months |

**Table 1.1**

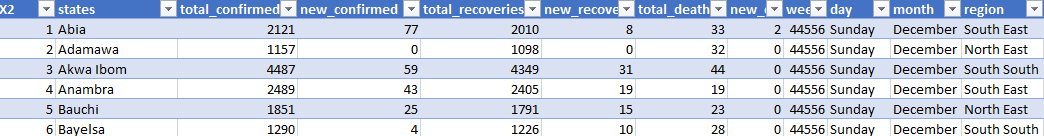
Lastly, our dataset contains 7 integer columns and 4 character columns.

**METHODOLOGY**

In this work, we start by taking a look at our data structure, checking for duplicates, number of columns, null values, and data types. By doing this, we are able to deal with inconsistencies in our dataset (if any), handle null values and duplicates.

Furthermore, for the purpose of our aim and objectives, we examine the distribution of our total\_death column using histogram.

We summarized our data by region as seen below. This was done in order to achieve one of the objective of this work



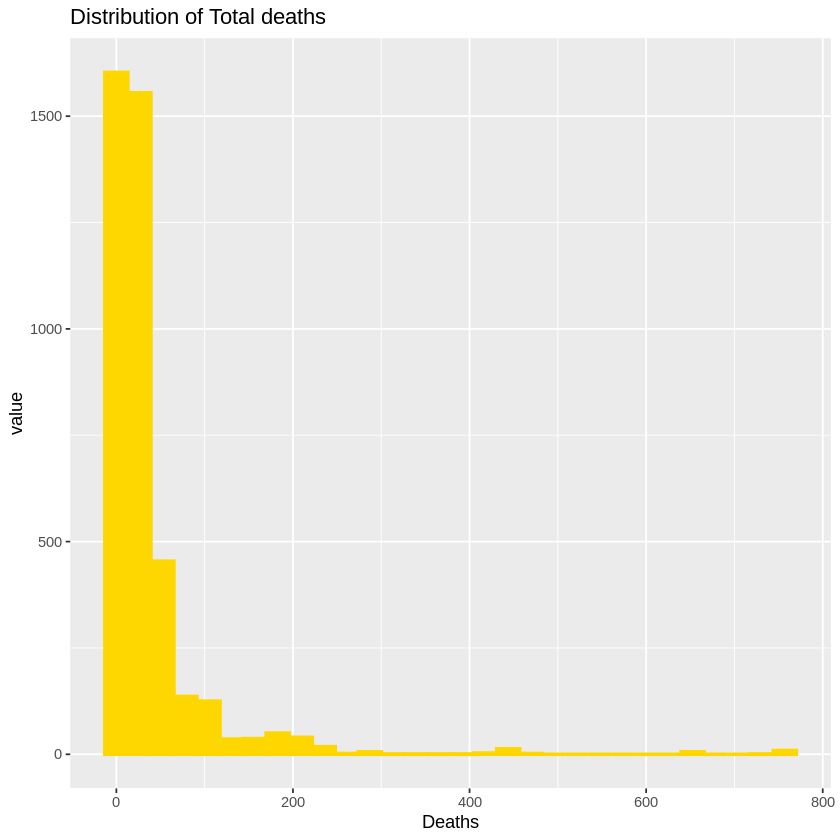
**Image 1.1**

We carried out bivariate analysis to examine the distribution of the total\_deaths in each region and then finally performed hypothesis testing to see if there is a significant difference between the mean of each region.

We further employed pos hoc tests using turkey HSD to understand specific differences between the mean of each region.

**RESULTS**

* From the result of the Distribution of our total deaths below:

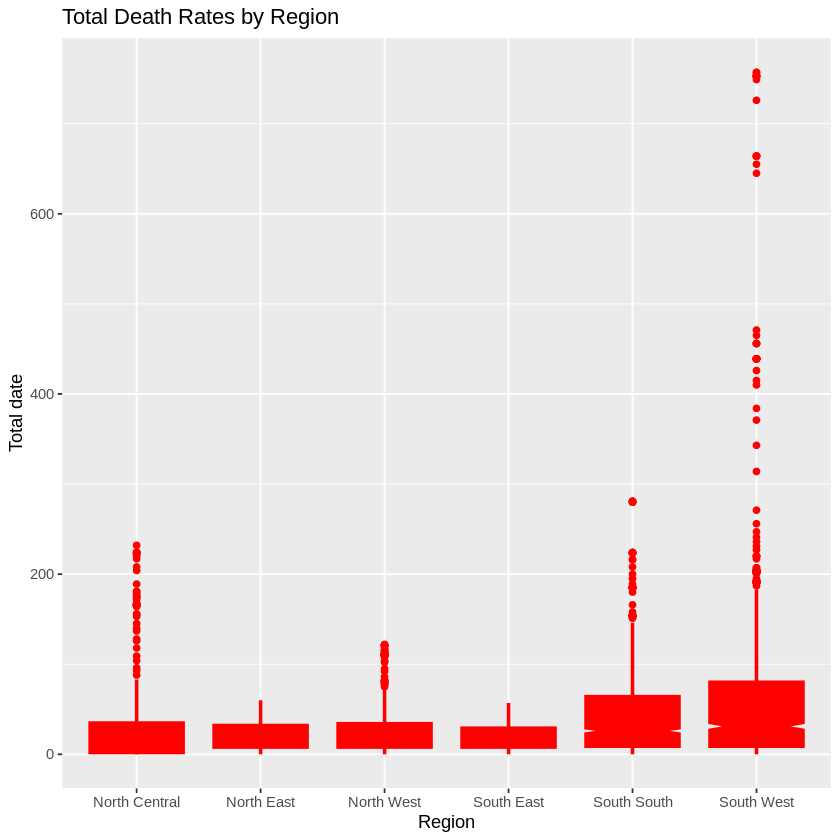


**Image 1.2**

We; see that the distribution of the total deaths is Right skewed and does not follow a normal distribution, meaning that there are outliers in our total death for each day. With the minimum death being **0** and the maximum death being **757.**

We further conclude that there were days where we recorded no death amd on some other days, more deaths than usual.

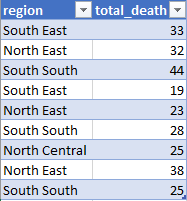
* In addition, by conducting bivariate analysis as seen below:



**Image 1.3**

We see that in regions like South west, North central, South south and North west the death rates had outliers indicating that some days experienced death rates ranges which were above average.

* To carry out our hypothesis, we subset the column of interests which are region and total deaths as seen below.



Using anova, we test if there is a significant difference between the average death of the different regions, i.e

* H0 (null hypothesis) : There is no difference between the average deaths in each region.
* H1 (alternate hypothesis): There is a difference between the average deaths in each region.

We obtain the following results:

* Degree of freedom of 5 (i.e K-1 degree of freedom), indicating that there are 6 groups in our dataset.
* Sum of squares of 1786769, indicating that there is a variability of 1786769 between groups.
* F-statistic of 91.22, indicating that there is a greater degree of variation between the group means relative to the variation within the groups.
* A p-value of 2e-16 (using a confidence interval of 99%). This means that at 99% confidence interval, we reject the null hypothesis (H0) that there no difference between the average deaths in each region and conclude that there is a difference in the average deaths in each region. Our hypothesis is also valid at 95% and 90% respectively. This indicates that the difference in average deaths in these regions are actually significantly different.

To further understand the result of our hypothesis result using anova, we conducted a pos hoc test using using turkey HSD to understand specific differences between the mean of each regions.

Here we obtain the following results:

Region diff lwr CI upr CI p adj

North East-North Central -10.115722 -19.574273 -0.657172 0.0279505

North West-North Central -3.483692 -12.543109 5.575725 0.8830174

South East-North Central -11.592658 -21.534032 -1.651284 0.0115155

South South-North Central 19.139010 9.664873 28.613147 0.0000002

South West-North Central 48.654239 39.282231 58.026248 0.0000000

North West-North East 6.632030 -2.815332 16.079392 0.3415416

South East-North East -1.476936 -11.773077 8.819205 0.9985382

South South-North East 29.254732 19.408978 39.100487 0.0000000

South West-North East 58.769962 49.022441 68.517482 0.0000000

South East-North West -8.108966 -18.039695 1.821764 0.1829583

South South-North West 22.622703 13.159736 32.085669 0.0000000

South West-North West 52.137932 42.777215 61.498649 0.0000000

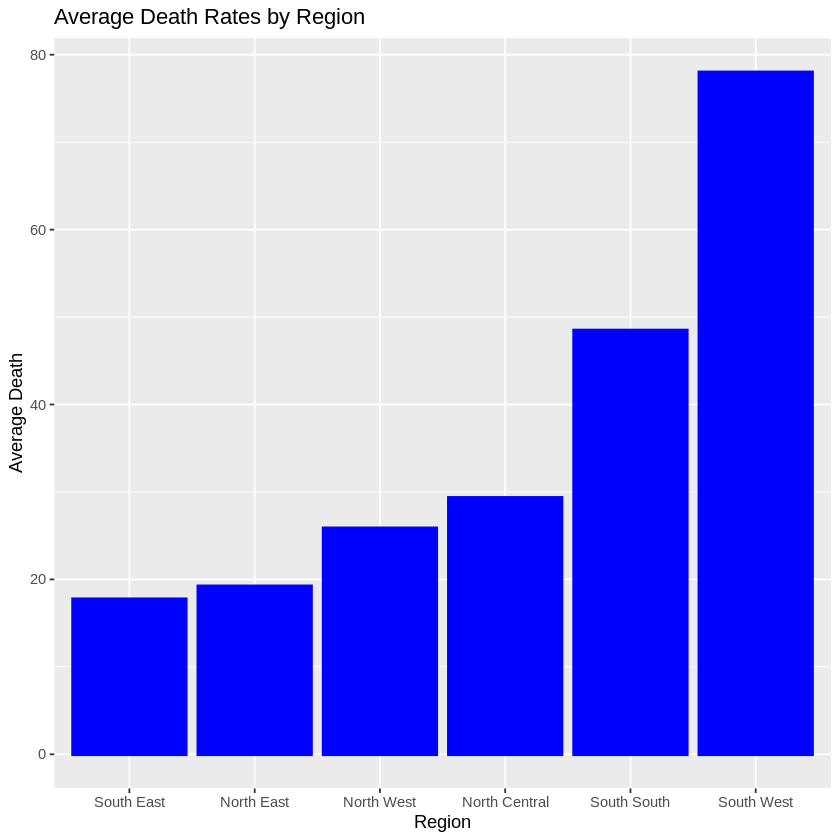
South South-South East 30.731668 20.421207 41.042130 0.0000000

South West-South East 60.246897 50.030200 70.463595 0.0000000

South West-South South 29.515229 19.752583 39.277875 0.0000000

This means that at 99% confidence interval, there is a significant difference (comparing between regions) in the mean deaths of the regions whose p adj values are less than 0.01 (P<0.01). I.e at P (0.050 less than 0.01 we reject our null hypothesis.

To see the mean distribution, we plot the graph below.



**CONCLUSION**

In this work we conclude that there is a significant difference between the mean deaths in certain region as seen above.