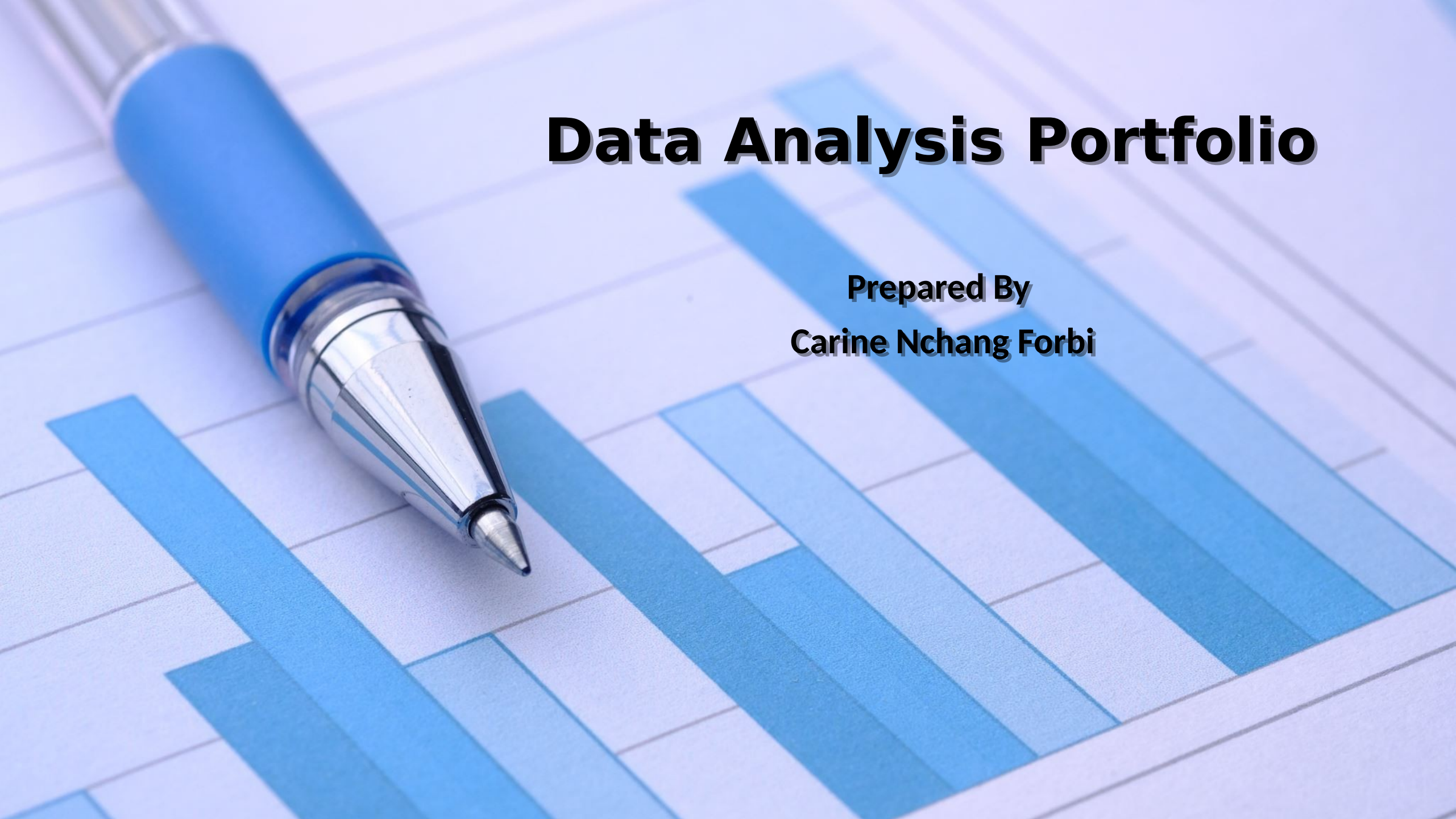


Data Analysis Portfolio

Prepared By

Carine Nchang Forbi

Professional Background

I am a MSc Economics graduate based in Germany. I am passionate about analyzing data and eventually transforming it to a meaningful insight that would impact businesses and result in policy changes.

Throughout my learning path, I have learned and embodied core data analyst skills like asking the right questions, structural thinking, problem solving, data analysis, effective communication of data findings etc.

As a result of that I am currently actively looking for junior analyst roles or analyst intern roles.

Portfolio Outline

Udemy Project Description

- In this project, four csv files were consolidated to form one csv file in excel using excel power query. This consolidated Udemy real world dataset was analysed and likely patterns and trends that may have resulted in the increase or decrease in their course revenue were identified.

- Throughout this analysis, I focused my attention on understanding the reason for price differences as well as subscription differences in courses and to be more specific I was interested in finding out which specific course resulted in a great proportion of Udemy's course revenue.

- Throughout this analysis, I used statistical summaries like mean value and also drew great visuals like pie charts(for parts of a whole relationship) and bar charts(to check patterns) between variables in the dataset. I even went ahead to create a calculated column for the revenue by multiplying the price be the number of subscriptions.

- The result gotten laid emphasis on the need for Udemy in investing in creating more short duration courses as a viable business model to improve on the companies revenue.

The Problem

- To be able to identify the real problem in this task, I used the root cause analysis specifically (the 5 whys). The real business problem here is the difficulty
- I was given an entire duration of 3 weeks to complete my analysis and present my findings.
- I was also not expected to collect anymore data as a quantitative data frame was already provided to me and I was expected to just clean and analyse.
- I findings were to be presented in the form of interactive dashboards as well as writing a final report.
- I am of the opinion that some relevant variables like (Couse Video structure) which might have changed the results of this findings were left out in the data.

Data Design

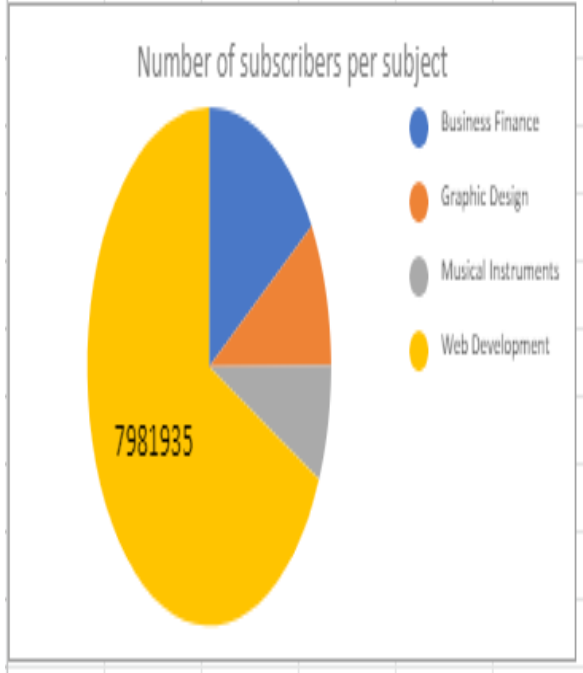
- The dataset came in the form of a data frame. The variables in the dataset were mainly quantitative and a few qualitative data. Also I had the dataset contained a datetime column.

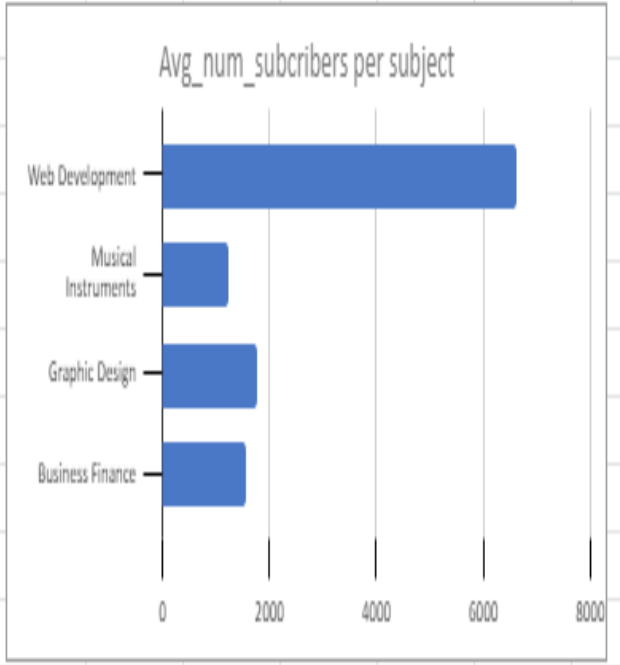
- This dataset contained missing values as well as duplicate values. Also some value and header names were not consistent. Some columns in this dataset were sorted in the descending order. I was able to remove all missing values from the dataset, appropriate header names given to the 3 newly created columns. One of the newly created columns(date) was split from datetime column which contains just the date. Two of the added columns were created with the help of the If function in Excel.

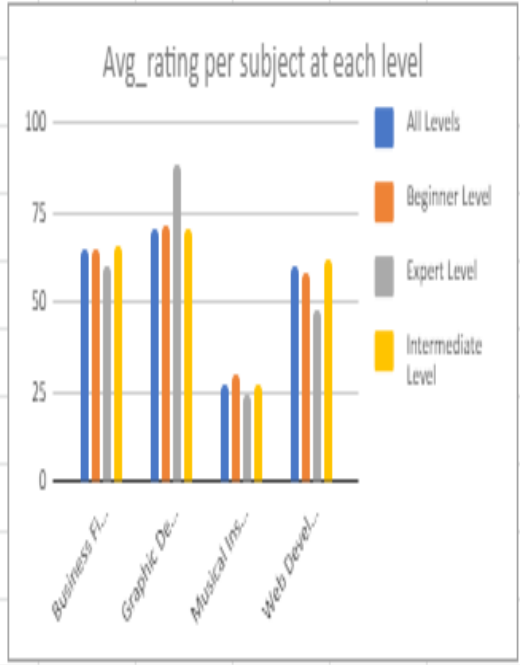
- For the visualisation, I used mainly bar charts and pie charts to check for patterns and parts of a whole relationships.

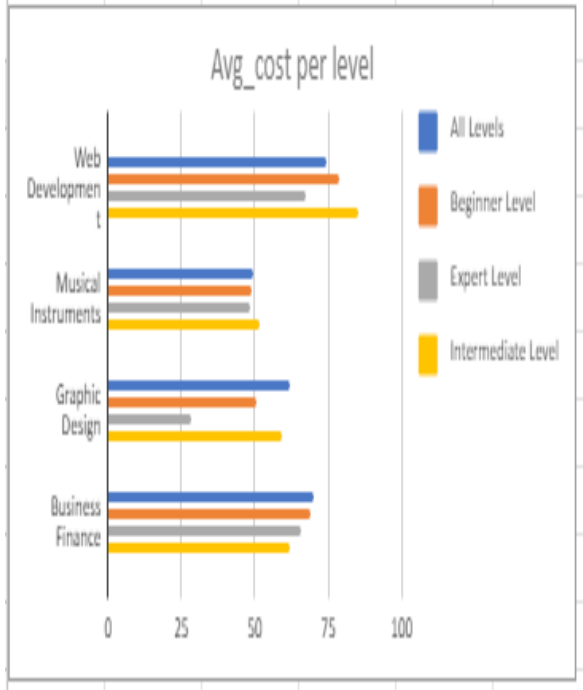
Excel Findings

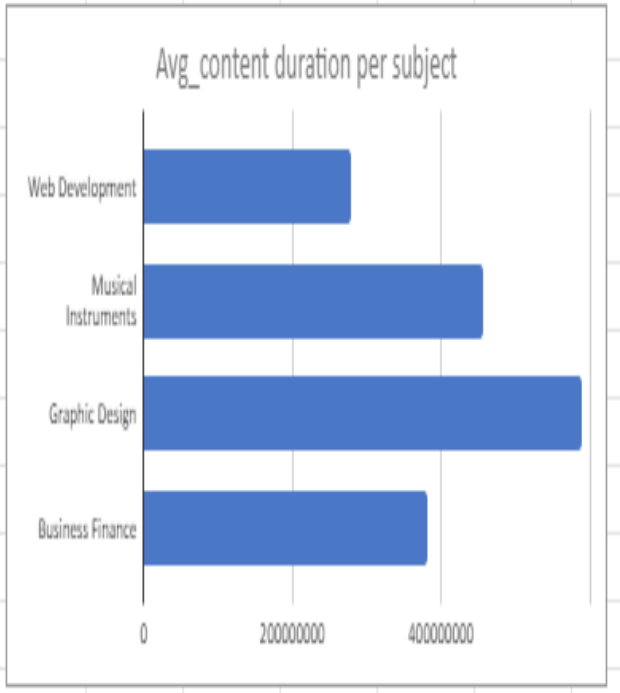
- Web developed had the highest number of subscribers per subject.
- Graphic design had the highest average rating per subject in every level.
- Web development had the highest average cost per level in every level
- Graphic design had the highest average content duration per subject.
- Web development had the highest number of reviews which came mainly from the beginner level and all levels.











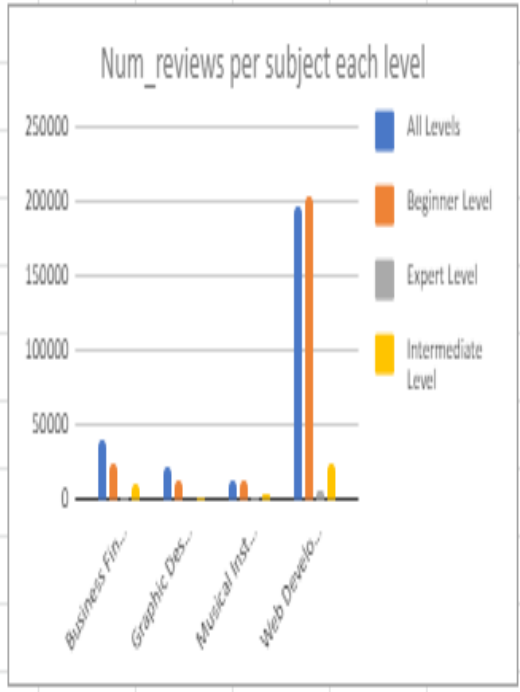
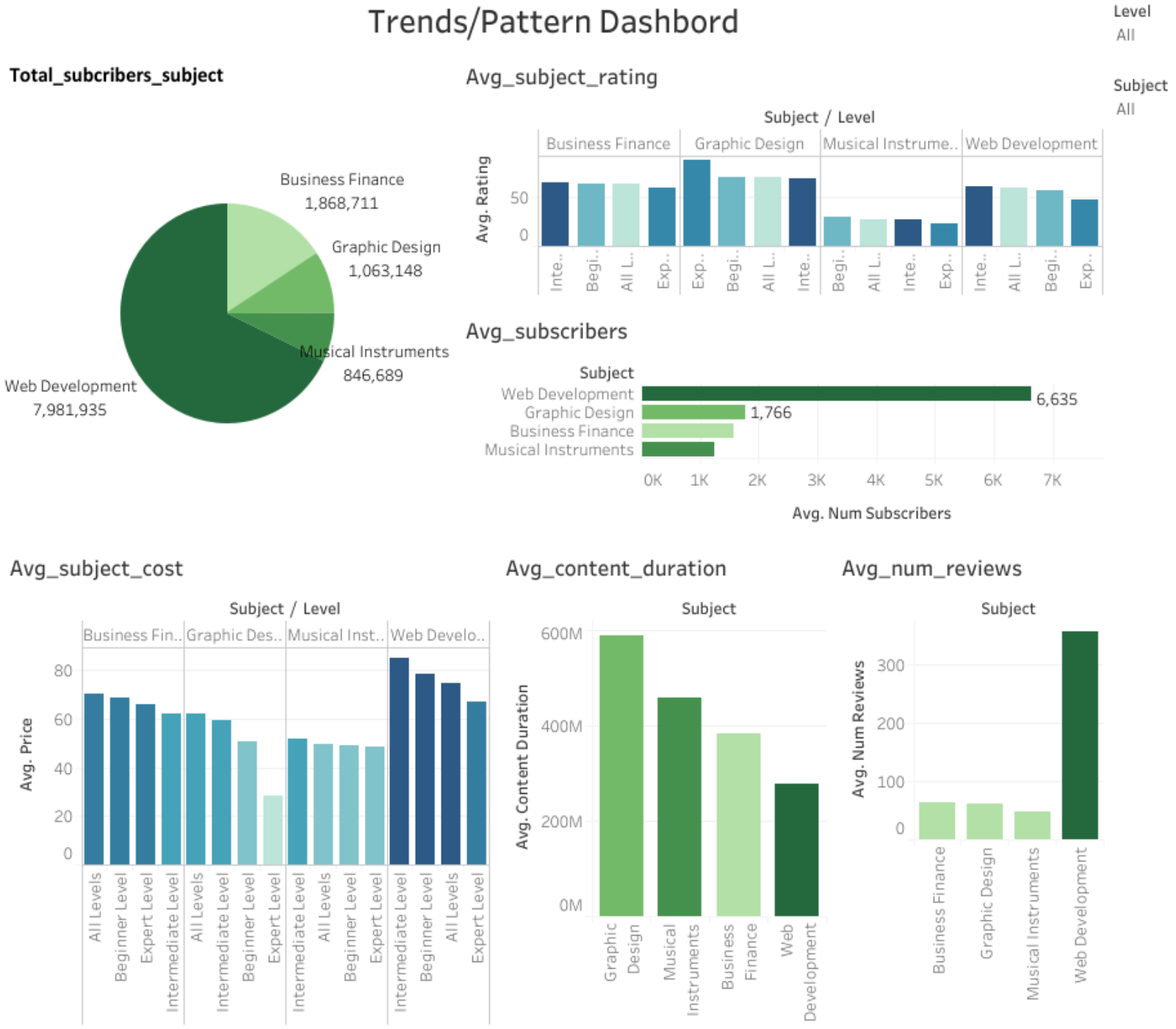


Tableau Findings

- Web developed had the highest total number of subscribers per subject of 7,981,935
- Graphic design had the highest average rating per subject in every level and it was especially high with the expert level.
- Web development had the highest average subscriber number of 6,635
- Graphic design had the highest average content duration per subject.
- Web development had the highest average number of reviews
- I noticed that the expert level had the least cost for all subjects except for business subject whose least costed level was the intermediate level.



Conclusion

- I noticed that people tended to be attracted to short duration courses and do not mind paying any amount for it. Hence Udemy could consider reducing the course duration of its other subjects as a viable business model in increasing its revenue.

- Also further research could be done in this area as there might have been other factors like subject structure which was not present in this dataset but which could in reality affect subject price as well as revenue. Hence this dataset might have omitted an important variable.

COVID AFRICA Project Description

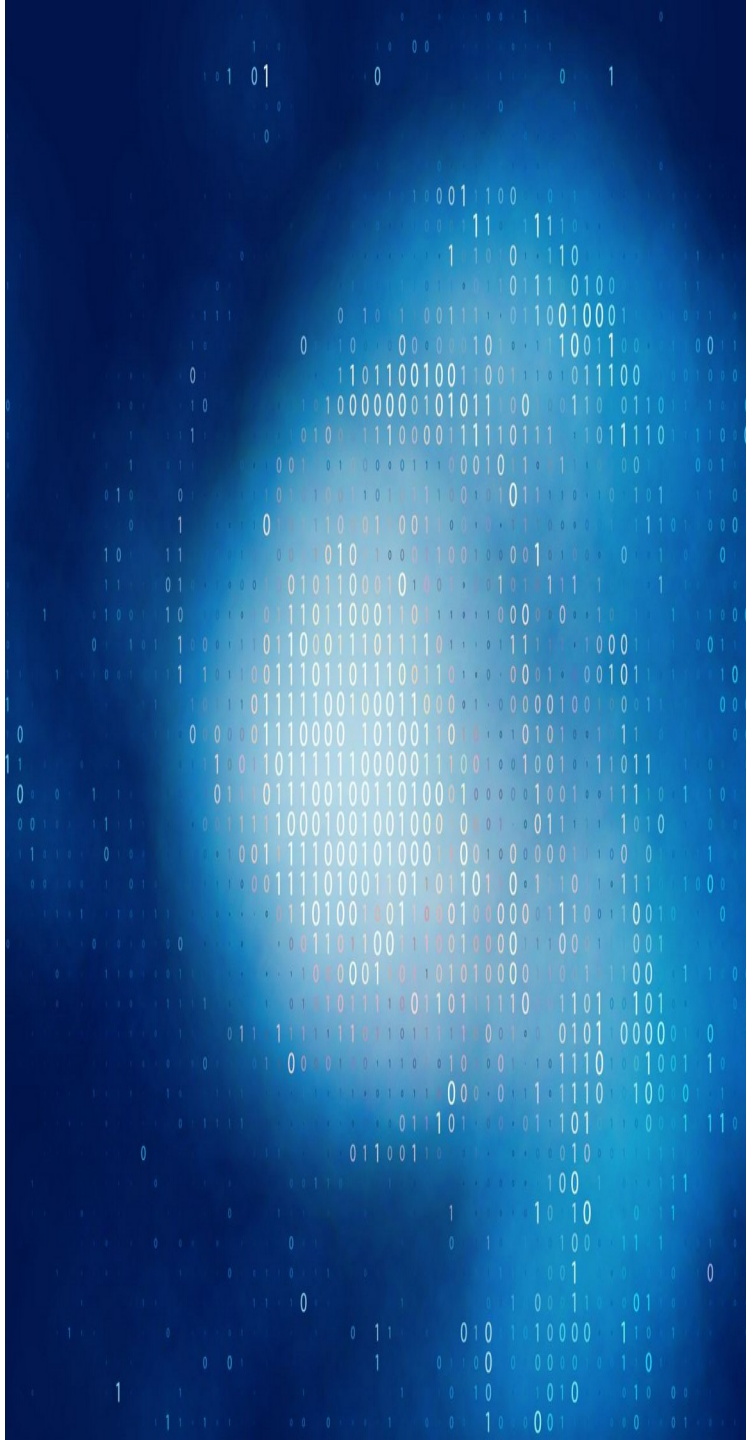
- In this project, I was expected to analyse any dataset of preference and come up with meaningful insights which could be used in making policy changes. The dataset I choose to use is the COVID Africa dataset which I got from Kaggle. This dataset contains information on COVID 19 in Africa for the month May and the year 2022.
- Throughout this analysis, I focused my attention on understanding the reason for high corona active cases and high death rates in some countries.
- Throughout this analysis, I used statistical summaries like mean value and also drew great visuals like line charts(to show trends and relationship) and geographic maps(because of the presence of a geographic column (location)). I even went ahead to create 4 new calculated columns(that is case fatality rate, mortality rate, test per capita, positive case per capita.
- The result gotten laid emphasis on the need for some countries to examine their corona testing policies.

Data Design

- The dataset came in the form of a data frame. The variables in the dataset were mainly quantitative data. Also, I had the dataset contained a datetime column and a geographic location column.

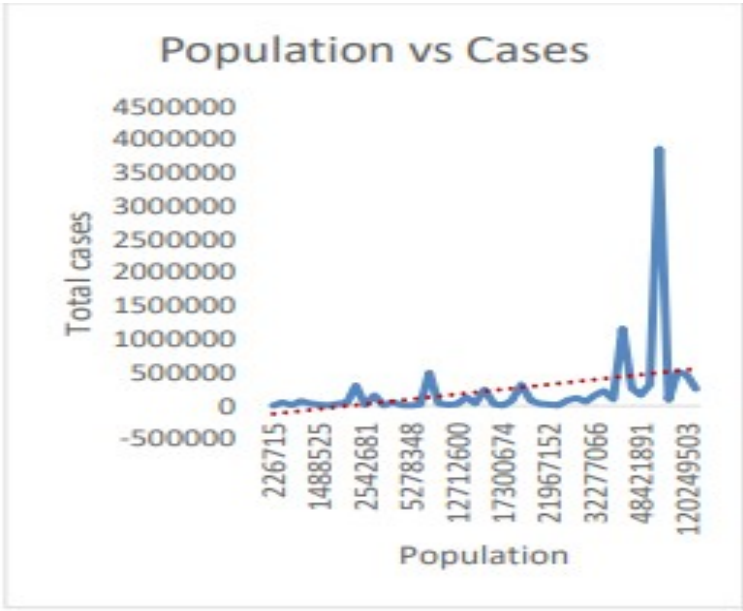
- For the data cleaning process, I began by using the “find & Replace” in Excel to find more than one word header and replace it with an under store to make it stand out more. I checked all columns to confirm the different data types which I realized were mostly string and integer. I also checked the data for duplicates using the “Remove duplicates” on Excel and realized that the data had no duplicate value. I also checked the data for missing values using the filter and realized that the data had missing or null values. I then proceeded to delete 7 rows which contained missing values leaving the data with 48 observations. I then proceeded to add a few calculated columns to the dataset which I felt will be helpful in my analyses. The columns that were added are stated below alongside what would try to explain in the data.

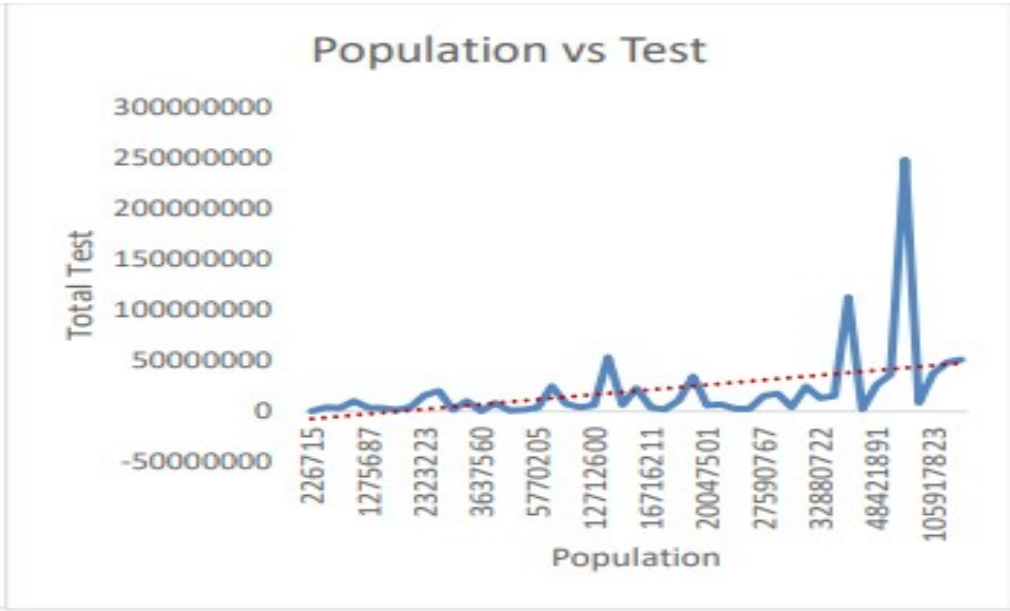
- For the visualization, I used mainly line charts and geographic maps to check for patterns and relationships.

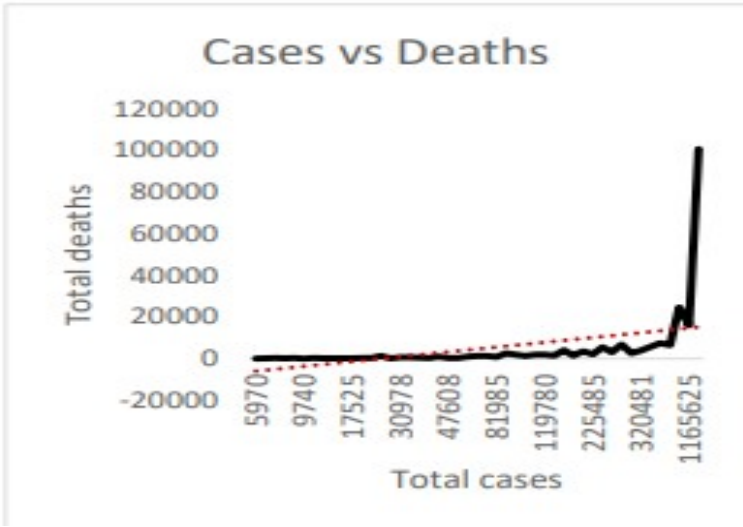


Excel Findings

- The 4 visuals show a positive relationship between the variables in consideration.
- That is there is a positive correlation between total population and total cases meaning that when one increases, the other will increase as well.
- The is also true for population and total test, cases and total deaths, and total test and total recovered.
- These trend lines even follow a somewhat similar pattern.







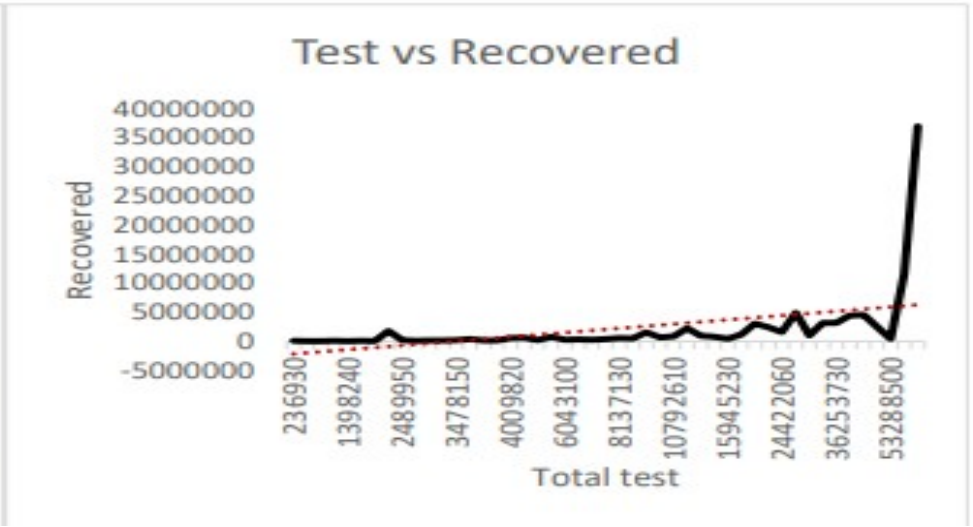


Tableau Findings

All areas in red represent countries at risk and countries of interest for further analysis.

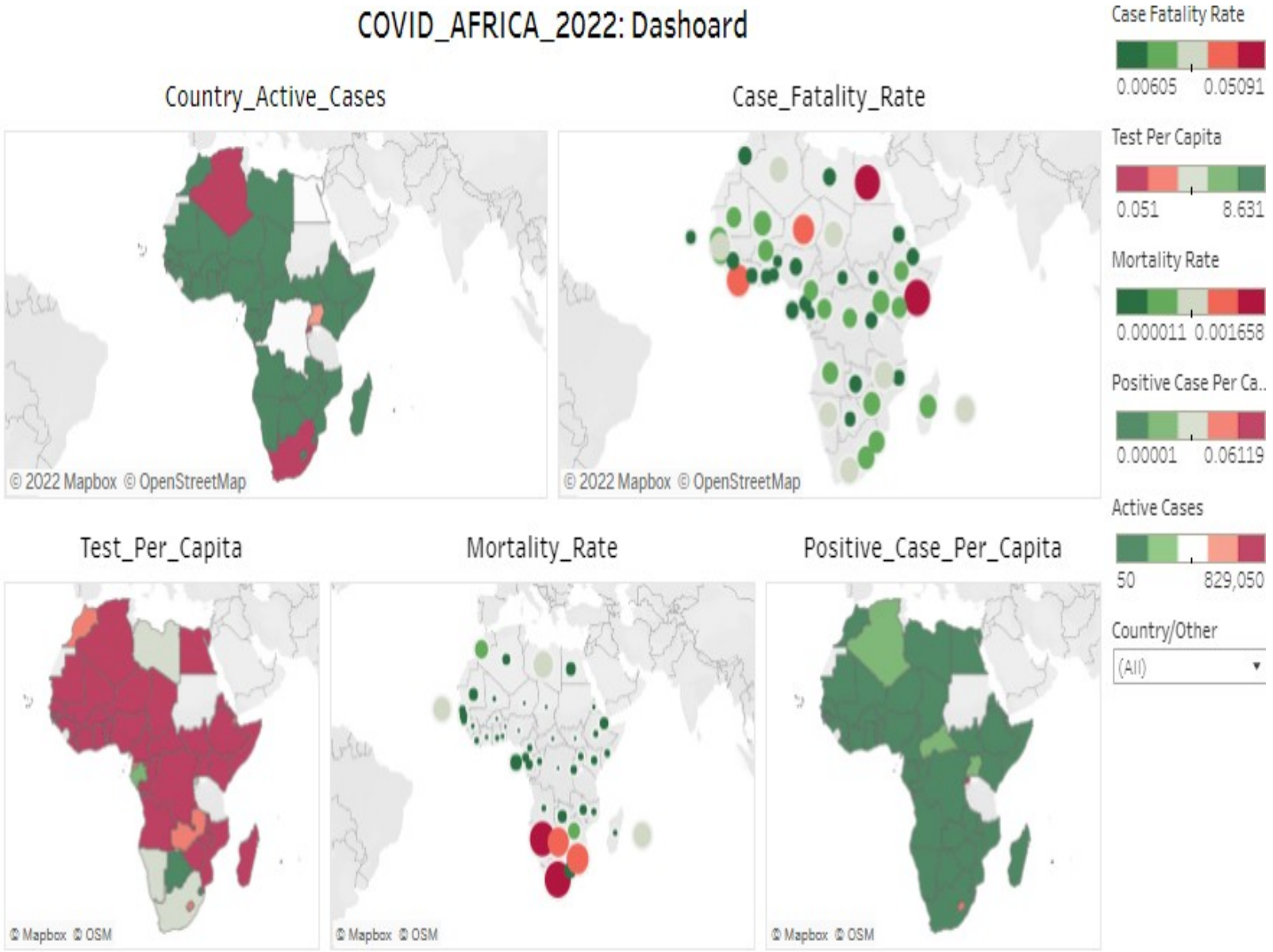
The three countries with the greatest number of active cases are Algeria (805,680) and South Africa (737,240) and Uganda (602,680).

Botswana, Gabon, Morocco, and Zambia really made a lot of efforts in tracking the outbreak by offering free testing to its inhabitants hence resulting to a lot of people testing themselves.

Based on that visual Egypt, Somalia, Niger, and Liberia had on average the highest Fatality rates in Africa ranging between 0.039 to 0.047.

Based on the visual Namibia, South Africa, Botswana and Eswatini had on average the highest mortality rates meaning that they are on high risk.

Rwanda and Lesotho had on average the highest positive case per capita meaning that the virus is spreading rapidly in these countries.



Conclusion

- To conclude, I would recommend that intensive measures in countries like Algeria, South Africa and Uganda should be enforced such that the spread of the virus should be prevented to other neighbouring countries. Other countries should copy the examples of Gabon, Morocco, and Zambia such that people can be tested early on before it gets critical like resulting to breathing issues. One country I'm most interested in further researching on is Botswana because I can see that they kept a good track of the corona outreach but still had a high mortality rate. So, in that case I think there might have been other factors not necessarily mentioned which might have caused these