**DATA VISUALIZATION**

**(CMP020X302A) - (UG)**

**Autumn Term 2023-2024**

**Coursework 1**

**A diagram of a graph

Description automatically generated with medium confidence**

**Student Name: Mehraj Harish Gaud**

**Student ID: GAU20508418**

|  |  |  |
| --- | --- | --- |
| Sr No. | Topic | Page |
| 1. | Introduction | 3 |
| 2. | Coursework Research | 3 |
| 3. | Data Pre-processing | 4 |
| 4. | Task 1:  Python Visualization | 8 |
| 5. | Task 2: Tableau | 20 |
| 6. | Conclusion | 27 |

**INTRODUCTION:**

Data visualization is a powerful tool in today’s digital world where huge amounts of data can be presented in a concise form to get meaningful information from it. Today where there is the limitation of time everywhere we need to consider the quickest, easiest, and most convenient way of presenting or delivering data so that we can get the most out of it, data visualization techniques can be a game changer. Moreover, looking over big data one by one manually can be stressful and there is a likelihood of missing out on some key points from the dataset, whereas using data visualization techniques helps to visualize those key points, trends, important references, etc. without fail while focusing on minute details.

**COURSEWORK RESEARCH:**

To build this visualization coursework, a lot of research has been done, including research from academic resources like lecture slides and lab sessions, and external sources like the video tutorial links given in the description of the Data Visualization module page of Moodle from LinkedIn and YouTube. This research work helped in enhancing the knowledge about the use of Python for pre-processing the data and visualization of it. Similarly, the video tutorials helped to understand Tableau from basic to advanced due to which these top-notch visualizations are possible and being delivered along with an interactive dashboard for the overall analysis of Tableau visualizations.

**DATA PREPROCESSING**

Step 1:

A close up of a sign

Description automatically generated

Changed the WHO region column name from WHO Region to WHO\_REGION in the COVID-19 dataset. The reason to change the column name is that we have a similar column in the vaccination dataset, so whenever we need to merge those two datasets in any situation while plotting a graph, we can join based on this column.

Step 2:

A screenshot of a computer

Description automatically generated

The values in the column WHO\_REGION for the COVID-19 dataset changed to its short for example:

* AMRO: American Region
* SEARO: South-East Asia Region
* WPRO: Western Pacific Region
* EURO: European Region
* EMRO: Eastern Mediterranean Region
* AFRO: African Region

This pre-processing step will help to match the column values in vaccination and COVID datasets while merging the data using this column.

Step 3:

Some missing values in the vaccination dataset made data inconsistent and vague, so that column needed to be filled to make the plot more efficient and look better for visualization. The columns in the dataset that were incomplete:

* PERSON\_VACCCINATED\_1PLUS\_DOSE
* PRESONS\_VACCINATED\_1PLUS\_DOSE\_PER100
* PERSONS\_FULLY\_VACCINATED\_PER100
* FIRST\_VACCINE\_DATE
* NUMBER\_VACCINES\_TYPES\_USED
* PERSONS\_BOOSTER\_ADD\_DOSE
* PERSONS\_BOOSTER\_ADD\_DOSE\_PER100

After pre-processing those columns, the data in that column made more sense was easy to use for analysis and provided meaningful visualizations.

Step 4:

I added a few columns to the COVID-19 dataset that could provide a good and deeper understanding of the dataset and would make visualization more interesting:

* Number of cases recovered.
* Date updated.
* Number of female cases.
* Number of male cases.
* Total number of adult cases.
* Total number of child cases.
* Number of female deaths.
* Number of male deaths.
* Total deaths of children.
* Total deaths of adults.
* ISO3.
* Recovery rate.

Also, I added a few columns in the vaccination dataset to enhance the understanding of the vaccination dataset:

* Number of females fully vaccinated.
* Number of males fully vaccinated.
* Fully vaccinated adults.
* Fully vaccinated children.
* Booster dose in children.
* Booster dose in adults.

I added all the above columns using Python and its operations.

Step 5:

A screenshot of a calendar

Description automatically generated A screenshot of a calendar

Description automatically generated A screenshot of a medical record

Description automatically generated

Vaccination dataset COVID dataset Vaccination dataset

I manipulated the dates in the DATE\_UPDATED columns of vaccination and the COVID-19 dataset, also the first vaccination date in the vaccination dataset using Excel formulas.

Step 6:

I have used four different files for the coursework that are:

* 2 CSV files: covid 19 and vaccination files (Python visualization).
* 2 excel files: COVID-19 and vaccination files (Tableau).

**TASK 1: PYHTON VISUALIZATION**

**IMPORTANT LIBRARIES FOR VISUALIZATION:**

**A screenshot of a computer

Description automatically generated**

**VISUALIZATION 1:**

**Question 1:**

Compare and contrast the number of populations with one plus dose per 100 with a fully vaccinated rate per 100 to understand the trend of vaccination for the top 10 countries.

**CODE SOLUTION:**

**A screenshot of a computer code

Description automatically generated**

**GRAPH SOLUTION:A graph of vaccination

Description automatically generated**

**EXPLANATION:**

This is a line graph between person one plus dose per 100 and person fully vaccinated per 100 which gives us a clear comparison between those two. The data for this comparison is arranged according to the number of people fully vaccinated in descending order and only the top 10 are used for the comparison, we can see that a huge number of the people got vaccinated with the one-plus but after that, they didn’t get fully vaccinated with the further dose. This graph gives us a distinct picture of all the top countries with no fully vaccinated population per 100 and gives the least importance to being vaccinated with further doses.

From this visualization, we can observe that only China and Japan have a close rate of one plus dose and fully vaccinated per 100, unlike other countries like the USA, Pakistan, and Indonesia. From this visualization, we can also say that those countries need more awareness about being fully vaccinated and its benefits towards the protection against the virus.

**VISUALIZATION 2:**

**Question 2:**

How does the cumulative total of COVID-19 cases per 100,000 population correlate with the cumulative total of deaths per 100,00 population for countries in different WHO regions?

**CODE SOLUTION:**

A screenshot of a computer code

Description automatically generated

**GRAPH SOLUTION:**

**A graph with numbers and a red box

Description automatically generated**

**EXPLANATION:**

The scatter plot comparing COVID-19 cases per 100,000 population with deaths per 100,000 for different WHO regions offers a brief yet powerful visual summary of the pandemic’s impact. The color-coding data points based on WHO regions give a quick and intuitive comparison of how different regions are affected by the pandemic in terms of both cases and deaths. The circles with blue colour represent the EMRO region, red coloured circles represent EURO, green coloured circles represent AFRO, purple-coloured circles represent WPRO, orange-coloured circles represent AMRO, and the sky-blue coloured circles represent SEARO. If you hover over the circles, you will get more information about the number of cases and deaths in that region. The outliers in the plot may represent a country in that region that is doing well or facing significant challenges, these outliers can help for further investigation that the area requires additional support.

It offers a comprehensive snapshot of regional dynamics, highlighting not only the sensitivity of the pandemic in different areas but also the potential and opportunities for collaborative responses to global health challenges. It also gives an instant comparison between different WHO regions and what regions need to be given priority.

**VISUALIZATION 3:**

**Question 3:**

Measuring the fluctuations and stability for the recovery rate and persons fully vaccinated per 100 in the top 15 countries with maximum death rate.

**CODE SOLUTION:**

A screenshot of a computer code

Description automatically generated

**GRAPH SOLUTION:**A graph of different colored bars

Description automatically generated with medium confidence

**EXPLANATION:**

This is a visualization of the recovery rate and persons fully vaccinated per 100 for the top 15 countries sorted in descending order by death rate. This visualization is a quick summary for the decision makers to go through and conclude which country with a high death rate has a good recovery and fully vaccinated rate per 100. It is a useful visualization to get insight into a country and compare the recovery rate with the fully vaccinated rate per 100. The green bars in the visualization are for the recovery rate and the maroon bars are for the fully vaccinated persons per 100.

We can clearly see that the recovery rate for most of the countries (Yemen, Syrian Arab Republic, Somalia, Egypt, etc.) is independent of the person's full vaccination rate. This can also give us an assumption that immunity among the people in those countries is doing well against the virus. This can be due to various factors such as their diet, their daily routine, or anything they include to protect themselves from the virus. Still, for a few of them (Sudan, Mexico, etc.), the person getting fully vaccinated is not enough and still has the worst recovery rate. This can influence people to not get vaccinated as it does not help to boost their immunity against the virus.

**VISUALIZATION 4:**

**Question 4:**

Evaluate the trends for different age groups and genders for their respective cumulative death totals in the top 15 countries where the overall death cumulative total is sky-high.

**CODE SOLUTION:**

**A screenshot of a computer code

Description automatically generated**

**GRAPH SOLUTION:**

**The death rate for different genders:**

A graph with red and blue lines

Description automatically generated

**The death rate for different age groups:**

A graph with numbers and a green rectangle

Description automatically generated

**EXPLANATION:**

The main purpose of this visualization is to provide an overview of the number of deaths across different age groups and genders in the top 15 countries with a high number of deaths in different countries. Using the above visualization, we can say which gender and age group needs to follow high-level safety measures and should be alerted about it. Also, vaccination should be mandatory for that age group and gender.

The above visualization is divided into two sections the death rate of different age groups and the death rate of different genders. The top visualization is for the death rate of genders where the red line is for the number of female deaths and the blue line is for the number of male deaths. As we can observe from the visualization the number of female deaths is higher than the number of male deaths in most countries like India, Peru, United Kingdom, etc. In the bottom visualization, the green line represents the number of children’s deaths, and the purple line represents the number of adult deaths. From this visualization, we can say that the number of children’s deaths is more than the number of adult deaths in most countries like the United States, India, United Kingdom, etc. So, females and children should be given priority to be vaccinated in those countries and various methods or remedies should be given to boost their immunity. This also gives us an overview of the immunity of the females and children being very poor against the attack.

**VISUALIZATION 5:**

**Question 5:**

Visualizing the rate of total vaccinations administrated to people per 100 along with the total number of booster doses per 100 and the number of types of vaccines used in different countries?

**CODE SOLUTION:**

A screenshot of a computer code

Description automatically generated

**GRAPH SOLUTION:**

A screen shot of a computer

Description automatically generated

**EXPLANATION:**

This is a geographical visualization of the total vaccination per 100, the total number of booster doses per 100, and the number of vaccine types administrated in different countries you can animate it according to different WHO regions using the slider at the bottom of the graph solution provided below. The colour scale in this visualization is according to the total number of vaccinations per 100 and the colour intensity changes according to the total number of vaccinations in different WHO regions. As we hover over the different countries on the map, we can see the information it holds.

This can be a key visualization to infer the total vaccinations per 100 and booster dose per 100 according to different countries, as we can see in the above visualization the total number of vaccinations per 100 is way more than the number of booster doses per 100 and might vary depending on the country and region. And using the number of types of vaccination used we can analyze what vaccination is used the most worldwide, and how many types of vaccinations are used on an average across the globe. This can also help to track the most successful vaccine that helped in reducing the number of cases and how many vaccinations should at least be given to achieve good recovery and reduced number of deaths.

**TASK 2: TABLEAU VISUALIZATION**

**A screenshot of a computer

Description automatically generated**

**INTERACTIVE DASHBOARD:**

A screenshot of a computer

Description automatically generated

This is an interactive dashboard for global COVID and vaccination analytics, which is a combination of five different visualizations that are a pack bubble, a two-bar plot, a map, and a line graph detailed description of each plot is given below. The colour scale and parameters for them are in the middle of the dashboard. The colour scale is a useful indicator of the information that a visualization holds. Similarly, the parameters for those visualizations can be used to extract necessary information from the visualization and focus on a specific detail of that visualization. The reason to adjust the color scale and parameters in the middle of the dashboard is that the colour scale for different visualizations can be close to their respective visualizations and the parameters for easy access and maximum usage. There are two actions assigned to the dashboard: filter action and highlight action. Filter action is used to filter the data according to the selection made on the dashboard and will only keep the information for the selection. Likewise, there is a highlight action used to highlight the data from the visualization while keeping the rest of the information as it is in the visualization in the dashboard. To implement filter action, you must select the region on the bar or bubble or country on the map, and for highlight action, you must hover over a bubble or a country on the map or bar.

**SHEET 1:**

A screenshot of a computer screen

Description automatically generated

The above visualization is a pack bubble that shows the total number of COVID-19 cases, deaths, and total number of vaccinations supplied to different countries. The circles are coloured according to the countries and the size of the circles is based on the number of cases in each country smaller the size of the bubble smaller the number of cases, and the larger the size of the bubble higher the number of cases. There is a filter used which is set according to the WHO regions to find out the totals according to countries belonging to that region. The other information for the cumulative totals like death cumulative, total vaccinations, and persons fully vaccinated are given as details. This is a colourful and interactive way to depict the individual total of all parameters in the dashboard while hovering over bubbles and countries with a high number of covid cases can be easily identified.

**SHEET 2:**

A map of the world

Description automatically generated

This is a geographical plot that depicts the total number of active cases and the total number of cases recovered for all the countries the color intensity of the plot is according to the number of cases recovered where the dark green color represents the countries with a higher number of recovered cases, gradually when it gets light yellow the number of recovered cases reduces and when it gets dark orange the number of recovered cases is extremely less. The plot can be filtered according to different WHO regions to get deeper insights into a specific country or group of countries. This can be a useful summary to visualize the trend of recovered cases in different countries or across different WHO regions. As you can see from the example on the map India, the USA, and France are coloured in green as they have a high number of recovered cases and are getting light yellow and turn orange like in the African region.

**SHEET 3:**

A graph of a number of cases

Description automatically generated with medium confidence

This visualization compares the total number of deaths and the total number of cases for adults in the top 20 countries and is currently sorted according to the total adult cases. The colour of the bars is set according to their respective WHO region, and the number of countries can be changed using the top ‘n’ parameter on the right-hand side of the visualization. This parameter allows you to enter any number of countries and will present the data in descending order according to the total number of adult cases for your selection. This is an interactive way through which user can interpret the data according to their need and requirements. The takeaway from this plot can be that the countries with record-level numbers of cases always do not have high death rates for adults whereas, countries with comparatively a smaller number of cases may have exceptional death rates. So, we can say that the countries with high cases are doing well in getting their people recovered or people there are doing well against the virus when attacked.

**SHEET 4:**

**Adult booster doses are given across WHO regions:**

A graph with different colored bars

Description automatically generated

**Children's vaccination doses given across WHO regions:**

A screen shot of a graph

Description automatically generated

This is a bar plot to visualize the number of cases, vaccinations, and booster doses for various WHO regions and across different age groups and genders. To make this possible in a single bar chart the parameter and a calculation field have been used, which is on the right-hand side of the visualization. It is a drop-down list where you can select from various options as mentioned in the list. The colour scale for the bars changes according to the parameter selection, as you change the parameter the colour scale will adjust itself according to the parameter. Also, the x-axis label will change according to the parameter selection. The dark red colour in the visualization represents the maximum number of counts for the selection, and the blue colour represents the minimum number of counts for the selection. It can be a useful, easiest, and quickest source of information for the audience who are interested in various stats about cases, vaccination, or booster doses for different age groups and genders.

**SHEET 5:**

**Daily Deaths reported:**

A graph on a screen

Description automatically generated

**Daily Cases reported:**

A graph on a white background

Description automatically generated

Finally, we have this line chart to represent the daily cases and deaths reported in the last 24 hours monthly for different countries in the years 2021 and 2022. There is a drop-down menu in the top right corner of the visualization to assemble the line graph according to the total number of daily cases and deaths reported. The colour scale for this graph is next to the drop-down menu, the color scale will adjust itself as per the selected parameter from the list. For example, if you select cases reported then the range for the colour scale change as you can see in the bottom image for this section, and similarly, if you select deaths reported the range for the colour scale change as you can see in the top image for this section. As visible in the colour scale, the higher the peak darker the red colour indicating a maximum number of counts, and the darker the green colour smaller the number of counts. In October 2021 the death rate was higher, and it reached its peak in September 2022, in October 2021 the cases reported were slightly higher than before and reached their peak in September 2022.

**CONCLUSION:**

To conclude the report, the above visualizations give a vivid picture of the pandemic around the globe covering COVID-19 cases, deaths, types of vaccinations used, number of cases recovered, and booster doses, for distinct genders and age groups. These visualizations are a blend of excellent colour schemes and meaningful visualizations for appropriate questions. The dashboard is the overall and concise view of the COVID and vaccination data with an absolute focus on low-level details, along with dynamic aspects of the dashboard like parameters and actions for great user interactions allowing them control over the visualizations and dataset for best possible outcomes.

**REFERENCES:**

* LinkedIn Learning
* YouTube
* Lecture Slides