Declaration on Plagiarism

This form must be filled in and completed by the student submitting an assignment

Name/s:	Aaron Nolan | Kevin Cogan
Student Number/s:	18423054 | 18421694
Programme:	Masters of Computing: DA (PT)
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	06/12/22
Module Coordinator:	Dr Suzanne Little

I/We declare that this material, which I/we now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I/We have read and understood the Assignment Regulations. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.

I/We have read and understood the referencing guidelines found at http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines

Name: Aaron Nolan Date: 04/12/2022

Name: Kevin Cogan Date: 04/12/2022

Effectiveness of the Covid Vaccine

Abstract

In this assignment, we will use a range of covid related datasets, provided by Google, to explore the following question: '*Is the COVID-19 Vaccine effective?*'. Our investigation utilises 676MB of data to analyse factors such as deaths, infections and vaccine rates as well GDP per capita. We first explore the effectiveness of the vaccine against global death and infection rates. We use exponential modelling to estimate the infection trend rate and then finally explore the effectiveness of the vaccine in different GDP per Capita groups. In our conclusions, we found that there was a trend of death and infection rates lowering as more of the population became vaccinated. We noticed that in comparison to the estimated exponential growth of infections there were fewer infections than expected (showing that the vaccine is effective). Our final conclusion found that death and infection rates decreased in all GDP per Capita groups, however, it was more effective in high GDP per capita groups.

Datasets

The datasets used in this assignment were retrieved from Google Covid-19 Open Data [1]. This dataset is one of the largest aggregations of COVID-19 data with information uploaded daily from hundreds of sources around the world. The Epidemiology, Vaccinations and Demographic datasets were mainly sourced from government official websites with the small exception of some countries using open-source platforms for reporting information. The Economy dataset was sourced from Eurostat, Wikidata, DataCommons and WorldBank. These datasets were accessible in downloadable Common-Separated Values (CSV) formats. We stored them in a directory labelled CSV and imported them via a python script using the Pandas library. The data contained in each of these CSV files are detailed in the table below:

Name	Size	Rows	Cols	Data Types	Main Attributes
Epidemiology	508MB	12525825	10	string, float, datetime	Location, Date, Cumulative Confirmed Cases and Deaths, Daily Confirmed Cases and Deaths
Vaccinations	157MB	2545118	32	string, float, datetime	Location, Date, Cumulative and daily Fully Vaccinated
Economy	10KB	404	11	string, float, double	location, GDP Per Capita
Demographics	1.5MB	1289	4	string, float, double	Population, Population(0-9), Population(10-19)

The datasets outlined above contain the following aspects of big data:

- **Volume**: The total size of all the datasets used is 676MB
- **Variety**: The data comes from various sources and each contains different quantities and data types as discussed above
- **Veracity:** This data was retrieved in real-time for over a year by Google. Based on this we assume the data is coming from a reliable source
- **Velocity:** The data in Epidemiology and Vaccinations, up until September, are based on daily retrieved values

Data Exploration, Processing, Cleaning / Integration

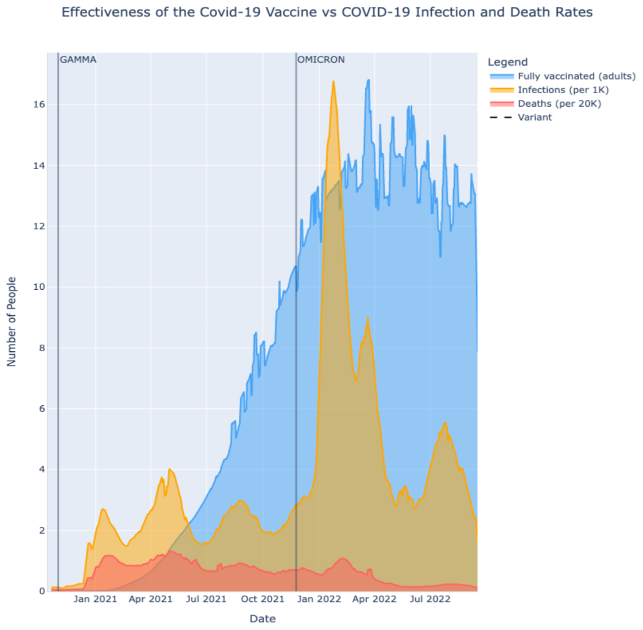
For each of the datasets collected, we had to process and clean the data. We achieved this by following this three-step process:

1. **Drop Data:** This included removing any unnecessary columns that were not being used, or duplicate entries.
2. **Filter Data:** First we removed sub-region rows from the datasets, then if there was a date column set the data type to date time and then only keep dates from 2021-2022.
3. **Sort Data:** Sort the values based on location_key.

Once all the datasets have been cleaned we then explored the data by merging various ones together and visualising them in simple line and bar charts. From here, we chose our graphs and iteratively conducted further data cleaning and transformation for each specific graph. We conducted research on the topic of COVID-19 to find an interesting question to investigate which was the "Investigation of the effectiveness of the COVID-19 vaccine". From the report, we found that the key factor for an effective vaccine is to prevent death and to stop the spread of the virus to the population. This informed us to use mainly the following attributes; population, cumulative_confirmed, new_confirmed, cumulative_deceased and new_deceased. We performed analysis on the distribution of data points to ensure each country has a sufficient number to produce unbiased insight, and on specific countries to see how covid performed in different countries of different GDP ranges.

Visualisation

Shown below are two of the three visualisations created with the COVID-19 datasets.



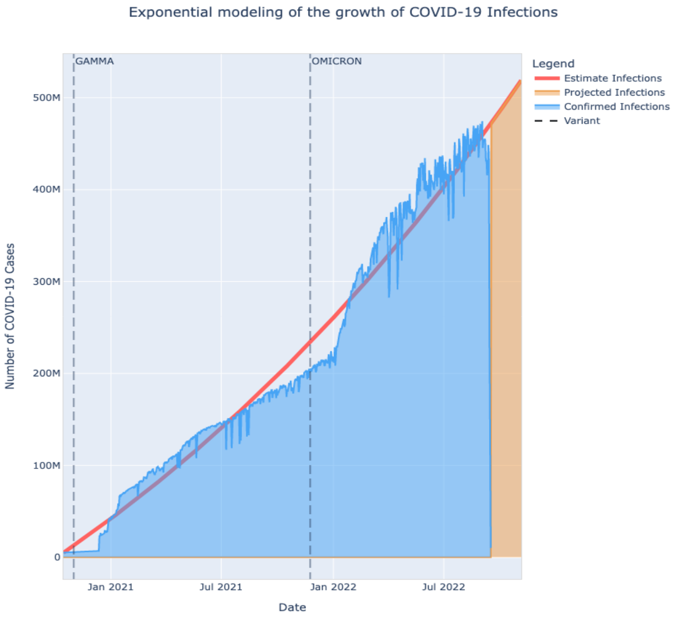


Fig. 1 - Vaccine vs Infection and Death Rates [2][3]

Fig. 2 - Exponential Modelling of Covid-19 Growth [2][4]

For these visualisations we want to compare how effective vaccines are based on deaths and infections over time. We also want to compare the expected infection rate to the actual. To rephrase this, we want to compare ratio values (vaccines, deaths and infections) against interval values (dates). To do this we chose to display the data through area line graphs. This graph type is best suited to compare "*one or more groups' numeric values change over the progression of a second variable, typically that of time*" [5].

To add to these graphs we wanted to consider if the country's GDP per Capita has an impact on the vaccine's effectiveness. This would mean we compare nominal data (GDP per capita groups; low, mid and high) against the ratio values (vaccines, deaths and infections) against interval values (dates), which is why we chose to display this through a vertical bar chart. This chart type is best suited for comparing "*categorical or discrete variables*" and also "*time series data*" [6].

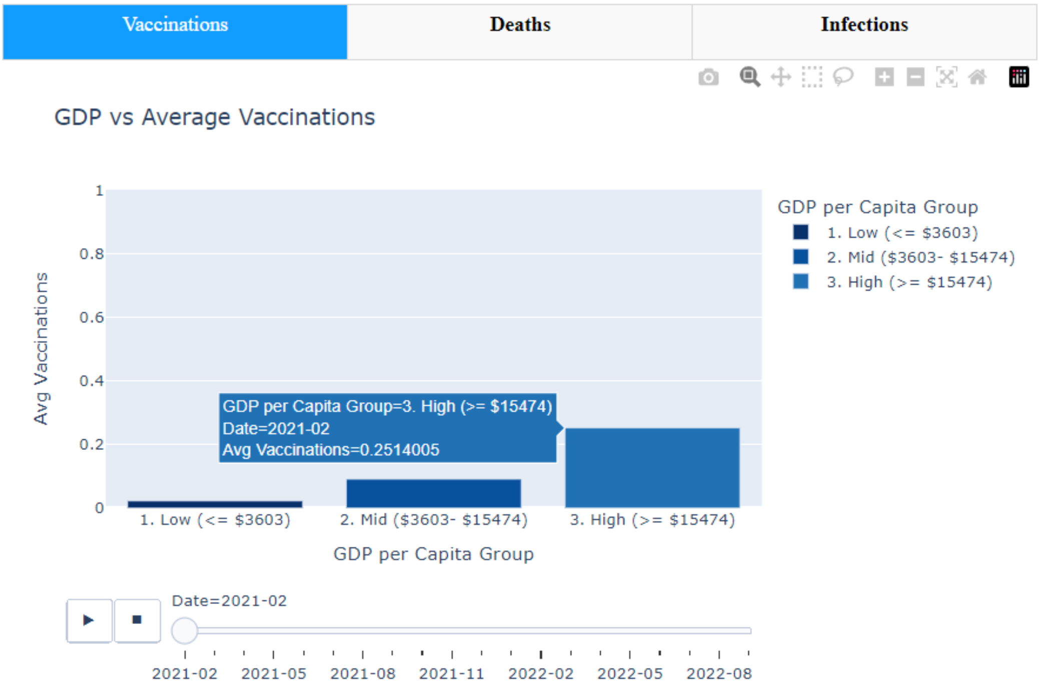


Fig. 3 - Vaccinations, Deaths and Infections based on GDP groups [7]

Design Choices

- For the linear graphs we used a qualitative palette for choosing our colour scheme. We chose this palette because it is used "*when the variable is categorical in nature*" and is "without inherent ordering" [8].

- For the bar chart we used a sequential palette for choosing our colour scheme. We chose this palette because it is used when the variable "*is numeric or has inherently ordered values*" [8].

- In the graphs, to add more context to the events surrounding the development of the COVID-19 variants, we added marks to indicate when the different variants of COVID-19 were first discovered (such as Omicron).

- For all graphs and charts, labelling is an important feature. For our labels, we utilised an Arial font to clearly explain the values of each graph. We added a grid to allow the user easily align the values of the graph to the X and Y axis. This was then aided with a legend that classifies all the information in the graph based on the colour scheme.

- After using the tutorial for building the bar chart [7] and readjusting it based on data, labels, structure and colours. We then brought in Dash to be able to re-render the same bar chart with different y-axis data (Vaccinations, Deaths and Infections). We did this because the difference in values between death and vaccinations meant the fluctuations could not be seen in a grouped bar chart.

Interactivity / Animation Choices

For all of our visualisations we want the user to be able to have as much flexibility over the content so they can make the most out of the visualisations. In regards to the animation, the bar chart allows the user to press play/stop and the chart will adjust its bars (animate them) to reflect each monthly period's data. This allows the user to view the fluctuations of values over a certain period of time. For interactivity, we ensured every graph and chart has a legend that allows the user to select lines and bars they do or do not want in the visualisations. Furthermore, the user can select a region of the graph where they will be provided with specific values. For the bar charts, we also have tab buttons for the user to select which data they would like to see for each of the GDP per capita groups.

Tools and Libraries Used

Although we used downloaded CSV we did not involve the use of Microsoft Excel specifically. We read the datasets using Python Pandas Framework and used this to clean, filter, merge, drop and sort the data needed for each visualisation.For the graphs and bar charts, we used Plotly and Dash to create dynamic animated and interactive visualisations.

Conclusion

Graph 1 (Vaccinations vs Deaths & Cases)
As vaccines are administered to the world population the number of deaths decreases and never exceeds the peak when no vaccines were administered. However, from this graph, it appears that the vaccine does not have an effect on new COVID-19 cases as the number of daily cases continues to fluctuate to new record level highs. To improve our analysis on the chart, we wanted to analyse the number of hospitalisation however all datasets discovered did not contain enough countries and consistent data entries to fairly represent hospitalisation in the world.

Graph 2 (Exponential Modelling For Covid Cases)
We are aware that new COVID-19 cases spread exponentially. When this is modelled on our data it shows that COVID-19 at the end of 2021 should significantly increase over the next 50 days however we see the actual number of new cases decreasing resulting in the vaccine having an effect on reducing the spread of COVID-19. To improve our analysis of the chart, we wanted more data to make a more accurate conclusion about the current trend however the data source has stopped providing daily updated data since September 2022.

Graph 3 (Economic Factors)
Considering countries with more money had faster access to vaccines (shown in the bar chart under Vaccinations) it was clearly seen that slowly as more of the population became vaccinated the infection rate and death declined, however, it was a lot slower in the mid GDP regions. It is also evident that the low GDP per capita regions were the last to receive these vaccines but also had the least amount of deaths and infections in comparison to the other two groups. To improve the chart with more time on the project we would create a more level statistic for measuring vaccination, infections and deaths (such as per 1000 of the population) so as to be able to display the data in a grouped bar chart.

As we worked in a pair, we both first took our own approach to cleaning and exploring the data. Once we had our insights we merged our cleaning process and split the graphs and charts. We also split the report and the screencast up making it an equal workload for each of us. We had meetings every two to three days to update each other as we went through the project.

References

1. COVID-19 Open Data — Google Health. 2022. Google.com. Available from: https://health.google.com/covid-19/open-data/ [Accessed December 3, 2022].

2. Present your data in a scatter chart or a line chart - Microsoft Support. 2019. *Microsoft.com*. Available from: https://support.microsoft.com/en-us/topic/present-your-data-in-a-scatter-chart-or-a-line-chart-4570a80f-599a-4d6b-a155-104a9018b86e#:~:text=Line%20charts%20can%20display%20continuous,evenly%20along%20the%20vertical%20axis. [Accessed December 3, 2022].

3. GitHub - Vaccine Effectiveness. (2022). COVID-19 Open-Data. [online] Available at: https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/examples/uk_vaccination_effectiveness.ipynb.

4. GitHub - Exponential Modeling. (2022). COVID-19 Open-Data. [online] Available at: https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/examples/exponential_modeling.ipynb.

5. Yi, M. (n.d.). A Complete Guide to Area Charts. [online] Chartio. Available at: https://chartio.com/learn/charts/area-chart-complete-guide/.

6. Statistics Canada (2010). Learning resources: Statistics: Power from data! Graph types: Bar graphs. [online] Statcan.gc.ca. Available at: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch9/bargraph-diagrammeabarres/5214818-eng.htm.

7. Plotly.com (n.d.). Intro to Animations. [online] plotly.com. Available at: https://plotly.com/python/animations/ [Accessed 4 Dec. 2022].

8. Yi, M. (n.d.). How to Choose Colors for Data Visualizations. [online] Chartio. Available at: https://chartio.com/learn/charts/how-to-choose-colors-data-visualization/ [Accessed 4 Dec. 2022].