Data Visualisation Coursework

This project explores the global impact of COVID-19 on public health and economy using data analysis and visualization. By combining datasets on cases, deaths, vaccinations, and economic indicators like GDP growth, the aim is to uncover key trends and relationships. The analysis uses Python and Power BI to deliver clear insights for both technical and non-technical audiences, while also evaluating existing public dashboards.

Data Collection

There are many sources of data we can use in our data analysis.

For our COVID-19 health data, Johns Hopkins University (JHU) collected data on the pandemic from 22/1/2020 to 10/3/2023. They tracked confirmed cases, deaths, recoveries and later, vaccination data.

Economic data was collected from the World Bank Group. They offer thousands of datasets and have an interactive dashboard for public use to download specific series from specified datasets. This project will use GDP growth (annual %), Inflation (Consumer Prices, annual %) and Unemployment figures (total % of labour force).

Due to formatting differences between the datasets, additional work was required to standardize elements such as column names, to ensure effective merging.

Data Cleaning

A screenshot of a computer program

AI-generated content may be incorrect.After collecting the datasets, extensive cleaning was performed to ensure they were able to be merged and suitable for analysis.

Within the project folder, there are folders for the code used in the project, folders for housing the data and a folder for the generated visualizations. In total, six datasets required cleaning and reformatting before being merged into a single dataset.

We will use the Pandas library to handle our datasets. The first function in the *clean.py* will load our .csv files into pandas dataframes:

A screen shot of a computer program

AI-generated content may be incorrect.Next, we must deal with empty values within the datasets. For the first three datasets, empty values will be filled with ‘0’. Initially, I planned to fill all missing values with ‘0’, but this was later avoided to prevent data distortion.

A screenshot of a computer program

AI-generated content may be incorrect.The next function dropped columns that were not required for our analysis. This will streamline our datasets and make it easier to merge them together.

The next couple of functions dealt with formatting the ‘Country/Region’ column. Three of the datasets used alternative names for that column, which needed changed to allow us to merge our datasets without duplicating the data. It also tweaked the year columns in the economic dataset.

A screenshot of a computer program

AI-generated content may be incorrect.

Additionally, the country names need to be examined to ensure there are no duplicate countries, for example ‘US’ and ‘United States’. There is a function that finds similar names and renames them for consistency. With the consistent country names, we will be able to merge our datasets using the Country names as indicators.

A screenshot of a computer

AI-generated content may be incorrect.The structure of the datasets also needs to be reworked for consistency. For us to merge effectively on the Country and the Date columns, we must ‘melt’ our tables to convert it from wide to long format.

The above function uses the ‘Country/Region’ column as constant while the other columns that are dates will be reshaped into a new column called ‘Date’. The value\_name creates a new column that will hold the value being measured. Below is the before and after of the structure of the .csv files:

Wide Format:

|  |  |  |  |
| --- | --- | --- | --- |
| Country/Region | 1/22/20 | 1/23/20 | 1/24/20 |
| Afghanistan | 0 | 0 | 0 |
| Albania | 0 | 0 | 0 |

Long Format:

|  |  |  |
| --- | --- | --- |
| Country/Region | Date | Confirmed |
| Afghanistan | 1/22/20 | 0 |
| Afghanistan | 1/23/20 | 0 |
| Afghanistan | 1/24/20 | 0 |

This restructuring simplifies the merging process and enhances compatibility for visualization.

Finally, with the datasets looking more consistent, the final function converts the date to a ‘datetime’ object. This will enable accurate sorting, grouping, analysis and makes merging the datasets easier.

A screen shot of a computer program

AI-generated content may be incorrect.

In the datetime conversion, it is specified what format the date will be displayed in (mm/dd/yy) for consistency.

After these functions have run, the dataframes are saved as new .csv files in the ‘data/cleaned’ folder, ready for merging.

Data Merging

The ‘*merge.py’* script takes the datasets we cleaned in the previous step and combines them into a single dataset for analysis after additional formatting.

Again, the cleaned .csv datasets are loaded into dataframes using pandas. The datetime conversion is also ran again, to ensure that the format is correct.

The economic data is then reshaped using the pandas function, pivot\_table(). This is necessary as we require the series of population, GDP, inflation and unemployment to be their own columns.



The next code snippet begins to merge our datasets together, starting with confirmed and deaths. At this point, it was decided not to continue with the recovered dataset as the vaccine rate would be used instead in the analysis.

A screen shot of a computer code

AI-generated content may be incorrect.

The above code merges the datasets together on the ‘Country/Region’ and ‘Date’ columns with outer joins to preserve all available data. Currently, these datasets have been merged together to form the ‘df\_merged’ dataframe.

A screen shot of a computer code

AI-generated content may be incorrect.

The above code creates a ‘Year’ column, extracting the year from the datetime object in the ‘Date’ column. This is essential as we will use it to merge the economic dataset which only has years as a measure of time. The economic dataset is then merged into larger dataset.

Now that the data is merged, there was some additional cleaning that was required after inspection of the dataset. The .*info*() function in pandas gave us details of the format of the columns. When importing into Power Bi, our percentage columns were not being interpreted correctly, the below code fixed that by changing the decimal place of the data.

There are also some ‘Countries/Regions’ that have irrelevant data, for example, the Summer and Winter Olympics are present in the dataset, so those and several other entries were dropped from the dataset.

Now that the datasets are merged into one file, we can break down and analyse the data.

Data Analysis and Visualisations

The ‘*analysis.py’* adds feature engineering, clustering analysis, regression modelling and generates a range of static and interactive visualizations.

This python script uses pandas and numpy, while also adding matplotlib and seaborn for creating for the visualisations and Sklearn for machine learning clustering and regression.

A screen shot of a computer code

AI-generated content may be incorrect.Again, the dataset is loaded into a dataframe using pandas. New columns are created that normalize the data so countries with different population sizes can be compared fairly.

The dataframe is then saved to a final dataset to be used for the rest of our analysis and visualisations.

A screen shot of a computer program

AI-generated content may be incorrect.K-Means Clustering

The first analysis is KMeans clustering. To explore the patterns among countries based on their public health and economic recovery, K-means clustering was applied using two key indicators, Vaccination Rate and GDP growth (annual %) the most recent data was extracted from the date column with the corresponding data entries per country for vaccination rate and GDP growth.

The KMeans model groups countries into three clusters based on their average vaccination rate and GDP growth. The clustering analysis produced the following results in the console:

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | Average Vaccination Rate % | Average GDP Growth (Annual %) | Countries in Cluster |
| 0 | 34.27 | 0.03 | 71 |
| 1 | 78.52 | -0.01 | 52 |
| 2 | 73.85 | 0.05 | 63 |

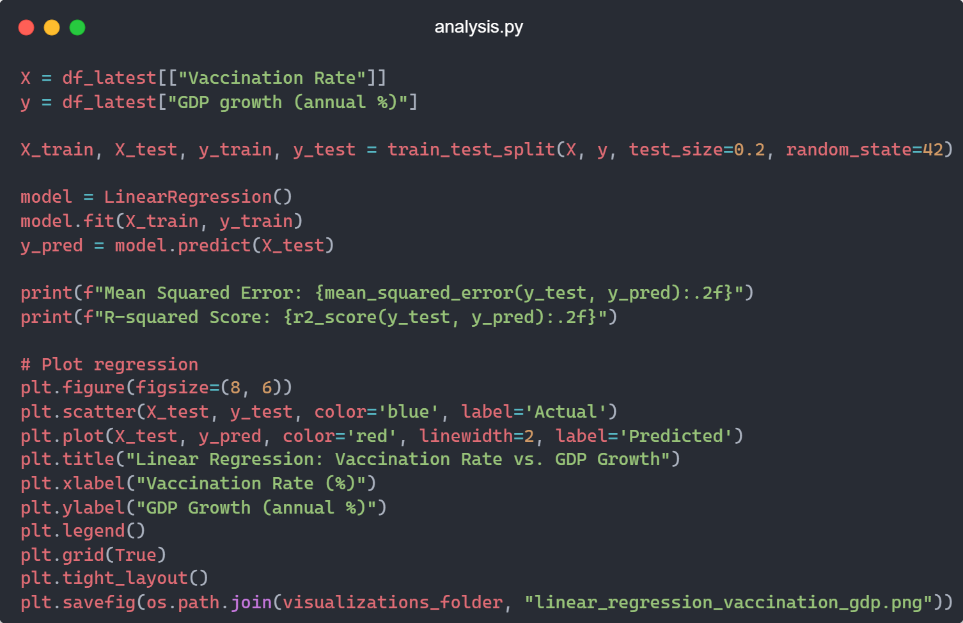
A graph with colored dots

AI-generated content may be incorrect.These results give us insights into the three groups. Cluster 0 suggests that lower-income countries or those with limited health care resources were both struggling with slow vaccine rollout and economic stagnation. Cluster 1 indicates that while higher income countries may have successfully implemented vaccination rollout, they still faced substantial economic challenges, possibly due to high debt, supply chain disruptions or reduced trade. Cluster 2 represents emerging economies or countries with more balanced responses, where moderate vaccination rates have been accompanied by strong economic recovery, possibly due to factors like resilient industries, strong domestic markets and effect government support. A .csv file is produced giving a country-by-country breakdown along with the cluster it fell into. It also produces a visualisation as seen below.

To evaluate the clustering, I added measurements for the silhouette and inertia scores. With a score of 0.35, the silhouette score tells us that the clustering has some structure with a bit of overlap in the data. The inertia score is the sum of squared distances between datapoints, measuring compactness. With a score of 188.3, the clusters have a moderate level of compactness.

Overall, the K-Means clustering analysis highlights that while vaccination rates are an important factor in economic recovery, they do not fully explain the variations in economic performance.

Linear Regression

To further examine the potential relationship between vaccination rates and economic recovery, we can use linear regression to evaluate whether a country’s COVID-19 vaccination rate could predict it’s GDP growth.

We used vaccination rate as the independent variable (x) and GDP growth as the dependent variable (y). The dataset was split into training (80%) and testing (20%) subsets using *train\_test\_split.* After training the model, the machine learning generated predictions for the test set and calculated two performance metrics that were printed in the console:

Mean Squared Error: 0.00 - suggests that the predicated GDP values are very close to actual values in terms of magnitude, likely due to the 4 years of GDP we have to work with in the data set.

R-squared Score: -0.02 – The negative squared score means the model fails to capture any meaningful trend between vaccination rates and GDP growth.

A graph with blue dots and a red line

AI-generated content may be incorrect.This indicates that the vaccination rate alone does not explain the variation in GDP growth, showing that economic recovery is driven by multiple factors beyond vaccination rate alone. A scatter plot is produced to demonstrate this, notice the red line, suggesting there’s no strong linear pattern, supporting the conclusion that vaccination rates alone are not a reliable predictor of GDP growth.

Correlation Matrix

To expand beyond the relationship between vaccination rates and GDP growth, a correlation matrix was generated and visualized using a heatmap. This matrix provides a picture of how strongly pairs of variables a linearly related, helping to identify patterns that might not be immediately obvious from raw data.

The above code takes the columns in out dataset and compares numerical values to find relationships. The visual below is produced:

A close-up of a computer screen

AI-generated content may be incorrect.The correlation matrix gives us some insights into our data and shows it in the heatmap. There are strong positive correlations in confirmed case and deaths (0.88), People with at least one dose and doses administered (0.89) and People with at least one dose and Population (0.85). This indicates that countries with high numbers of COVID- 19 cases also experienced high deaths. It also shows that vaccination proportionally scaled with country population size.

There were also moderate correlations, between Vaccination Rate and Cases per 100k (0.50) indicating that countries with higher case rates tended to have higher vaccination rates, likely to combat these high case numbers. There is also a logical link between cases per 100k and Deaths per 100k (0.48).

There is also some weak correlations as supported by our previous clustering and linear regression. GDP and Vaccination rate (-0.16) shows a weak negative correlation, suggesting the higher vaccine rates did not link to economic recovery. GDP Growth and Deaths (-0.01) or Confirmed Cases (-0.05) also suggests that the number of deaths and confirmed cases did not correlate with a country’s economic performance.

The matrix highlights that health outcomes and economic indicators are not linearly dependent. Countries that had higher vaccination numbers did not all have positively correlated GDP performance. This reinforces that economic recovery of countries during the pandemic did not hinge entirely on vaccination rollout, but by a wider array of factors, such as trade policy, supply chains and other geopolitical aspects.

These three analysis visuals will be insightful for data scientists looking at our dataset.

Critique of Existing Visualisations

World Health Organization COVID-19 Dashboard- <https://data.who.int/dashboards/covid19/cases>

A screenshot of a computer screen

AI-generated content may be incorrect.The first visualization we will examine is the number of COVID-19 reported to WHO:

The visual is comprised of a map with bubbles scaled to the number of cases reported. To the right is a KPI card showing a total case for the world with an annotation showing increase on previous 7 days. Below that is a country by country break down.

While the visual has clear labels, colour coding by region (Africa, Americas, Eastern Mediterranean, Europe, SE Asia and Western Pacific) and a list of countries and their corresponding case numbers, there are improvements that could be made.

While there are time measurements for the last 7 and 28 days, you cannot look beyond that without looking at the total cumulative. The list of countries is also not interactive, it would be ideal for the list to show additional details or perhaps even an expanded time scale to look at. However, when you change the time measurement to something other than total cumulative, the map does change, showing some interactivity.

While the colour coded bubbles are helpful, data can be distorted due to scaling. This scaling can also cause issues with overlap if other countries.

There is also no way to look at the rate per 100k, even though there is a button present there.

A screenshot of a computer screen

AI-generated content may be incorrect.JHU- <https://gisanddata.maps.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6>

This is the Johns Hopkins University dashboard that uses the dataset we have used in this project for our health metrics. While the WHO dashboard is several separate visuals on a webpage, the JHU dashboard is non scrollable page with several visualisations that are connected to the country slicer on the left-hand side. By clicking on a country, the visuals change to country specific data.

A screenshot of a computer screen

AI-generated content may be incorrect.

The visualization is very comprehensive, showing multiple metrics with clear numerical values. There is use of colour coding, red for cases, white for deaths and green for vaccine doses administered. Comparing it to the world heat map on the WHO visualization discussed previously, the scale of the bubbles is smaller to prevent bubbles overlapping into other areas. For some countries, there is an additional breakdown by region. For example, Australia in the above image has several bubbles to represent different areas within the country. This is expanded upon when looking at ‘US Vaccinations’ which shows state by state metrics for the whole country.

A problem that the JHU visual suffers from however is information overload. A user could find themselves to be overwhelmed with the amount of information being presented to them by default. There are many additional toggles and buttons, nested within other views. This could lead to a use clicking about before finding the data they wish to see.

A screenshot of a graph

AI-generated content may be incorrect.Visual Capitalist’s Post-Pandemic GDP recovery. <https://www.visualcapitalist.com/visualized-post-pandemic-gdp-growth-recovery-by-region/>

This visual shows the predicted GDP and actual GDP of countries before, during and after the pandemic. The visual is comprised of four-line graphs showing the GDP of the United States, China, Euro Area and EMDEs (excluding China and Brazil). While the visual shows line graphs for two individual countries, it does not state what countries make up the additional Euro Area and EMDE groups. Stacking two y axis on top of each other while keeping a consistent x-axis is an effective way to show two separate line charts at the same time without data distortion or cluttered. While the charts show the data value for the 2020 gap and 2025 gap, they could have shown the percentage points for the other years. There are no interactive features of this visual as it is a static image. Being able to choose between countries for each of the four charts would have allowed the user to customize which countries they wish to compare.

Creation of Additional Visualisations

In order to show more varied visualisations, Power BI was incorporated into this project. Power BI allows for interactive dashboards to be created from our data, offering a low-code/no-code approach to data visualization and analysis. Ideal for the public consumption.

A screenshot of a computer screen

AI-generated content may be incorrect.

This is the main dashboard, which users can interact with using the ‘Country/Region’ and ‘Date’ slicers to hone in on specific data. For example, clicking on the ‘United Kingdom’ brings up specific data on the UK:

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.On this main screen is a heatmap visual showing a comparative look at the average cases per 100k by Country/Region while also showing a line/bar chart showing GDP against Vaccination Rate over time. This visual helps the user surmise that while GDP did go up in 2021, the Vaccination rate did not have a strong correlation as the GDP then proceeded to fall afterwards.

The second page shows additional economic comparisons in a matrix. Users can scroll through to find a country to see the average GDP growth, Average inflation and Unemployment. Because the economic data is annual, there is a year slicer to see specific years. Below are individual line graphs for vaccination rate and GDP. Holding ‘Control’ on the keyboard allows for multiple countries to be selected for comparison.

A screenshot of a graph

AI-generated content may be incorrect. For example, I have selected the top 5 countries with the highest average inflation over the 4 years in our data, giving comparative graphs for Vaccination Rate and GDP Growth.

While we have provided audience specific visualisations for data scientists and for public consumption, we currently have little for that would appeal to policymakers. For this, we will create an additional visual to focus on actionable insights, such as showing that countries with early vaccine rollouts saw lower death rates.

For this, I opted to use pandas for more control over the visual and dataset manipulation.

The above code takes the date from the dataset, converts from a string in the .csv file to a datetime object, enabling us to perform sorting and date-based operations. We then filter the dataset to rows where vaccination has started. Then for each country, we take the earliest date that a dose was recorded, saving it in a new column. The dataset is then sorted by date, keeping only the last row per country which gives the most recent data point available, extracting the deaths per 100k at the most recent date. The two datasets are then merged on Country/Region to produce a summary table of the date of the first vaccine dose and the final recorded death rate. This is then passed into matplotlib to create a scatter chart with a correlation.

A graph showing a vaccine

AI-generated content may be incorrect.

The correlation of -0.29 shows a negative relationship, suggesting that countries that began vaccinating their populations earlier tended to report lower cumulative COVID-19 death rates. While the correlation is not strong, it does suggest that as the number of days before first dose increases, the deaths per 100k also tended to increase. The above chart shows two anomalies to far right of the chart, when inspected using code, the two Countries/Regions were Micronesia and the Marshall Islands, both recording their first vaccination on 22/02/2023 in our dataset.

We have now given three different audience specific visualizations.

Conclusion

This project focused on data cleaning, analysis, and visualization. The quality of our insights heavily depended on how well the data was cleaned and merged. Due to inconsistencies across countries, many values were left empty to avoid excessive data distortion. In particular, I preserved empty vaccine data to avoid falsely implying early vaccine availability.

Our analysis found that while higher vaccination rates correlated with lower infection and death rates, they did not strongly align with GDP growth, showing that economic recovery involved more complex, multifactorial influences.

Visualizations were tailored for different audiences:

1. **Policymakers**: Scatter plots linking vaccination timing to death rates.
2. **Data scientists**: Correlation matrices and clustering visuals.
3. **The Public**: Interactive Power BI dashboards for intuitive exploration.

Compared to WHO, JHU and Visual Capitalist visualizations, our work integrated health and economic data while avoiding information overload.

In summary, our analysis shows that public health efforts such as lockdowns were key to reducing COVID-19’s health impact, but economic recovery required broader, more integrated analysis across multiple sectors.