Olympic Games Over The Years

Deployed App URL:
https://share.streamlit.io/idsf21/assignment-2-urvashipriyam/main/main.py

Goals of the Project

This project enables a user to visualise the performance of countries at the Olympic Games over a span of 130 years.

It also helps a user visualize any trends between the countries' economic development (per-capita GDP) or population and the number of the medals it wins in the Olympics.

The application also enables a user to easily visualise the number of medals filtered on particular sports. The user can select the list of sports that they would like to see the visualization for, and the application would filter the results and show the number of medals won by each country in only the selected sports. The user can select the year of the games using a slider.

The application consists of 2 pages:

i) The 1st page shows the aggregate data of all the sports held in a particular Olympic Games. It consists of 2 parts, the first part is a heatmap visualization of the total number of medals won by each participating country (scaled by the type of medal) in that Olympic game on an interactive map of the world. In the second part, we colour code each country into its region and then use a bubble graph to plot the total number of medals won by each country, with the countries' per-capita GDP on the X-axis and the Population on the Y-axis. The size of the bubbles denotes the total number of medals won by each country (hence, scaling is needed to denote which country wins more Golds). This allows an easy way to visualise lots of information:

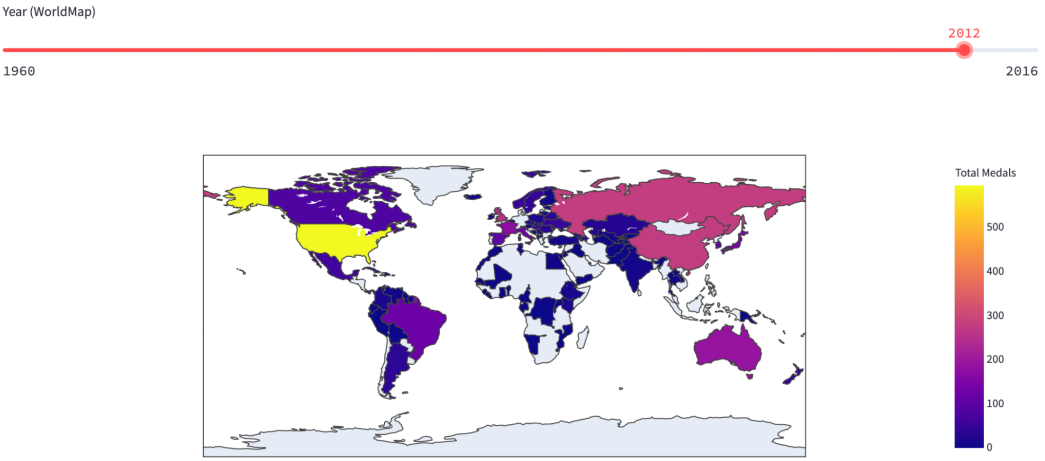
- Which regions of the world are the most and least developed?
- Which regions have the most and least population?
- Do developed countries usually win more Olympic medals every year?
- Do more populous countries usually win more medals each year?

Both the parts have a slider which can be used to change the year of the Olympic game.

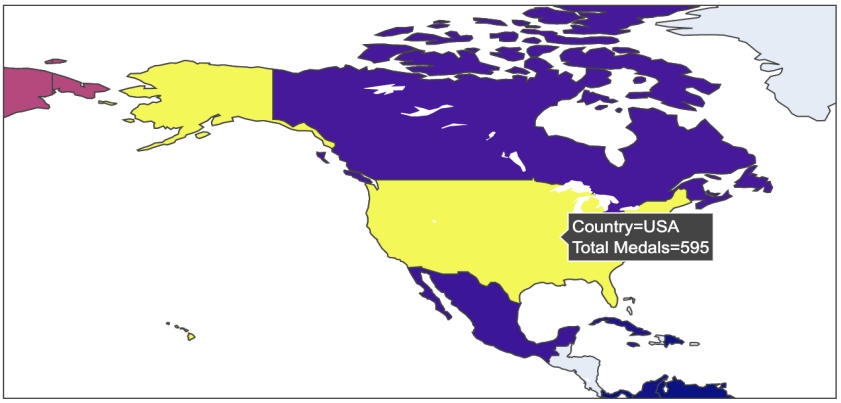
ii) The 2nd page allows the user to filter the data by particular sports. Multiple sports can be selected by the user and the application returns a visualization of the countries that won medals in those selected sports. Hovering over a country returns the exact number of Gold, Silver or Bronze medals won by that country.

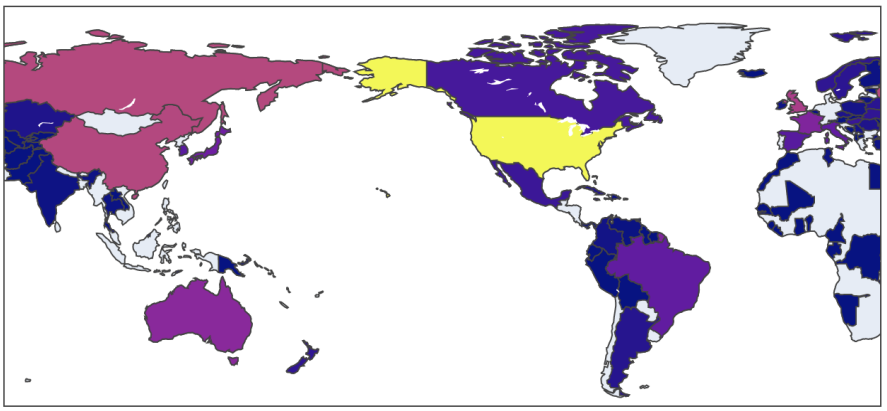
In greater detail, the application has the following visualisations and interactions:

1. An interactive Global Heat Map which shows the total scaled number of medals won by each country in that year of the Olympic Games.

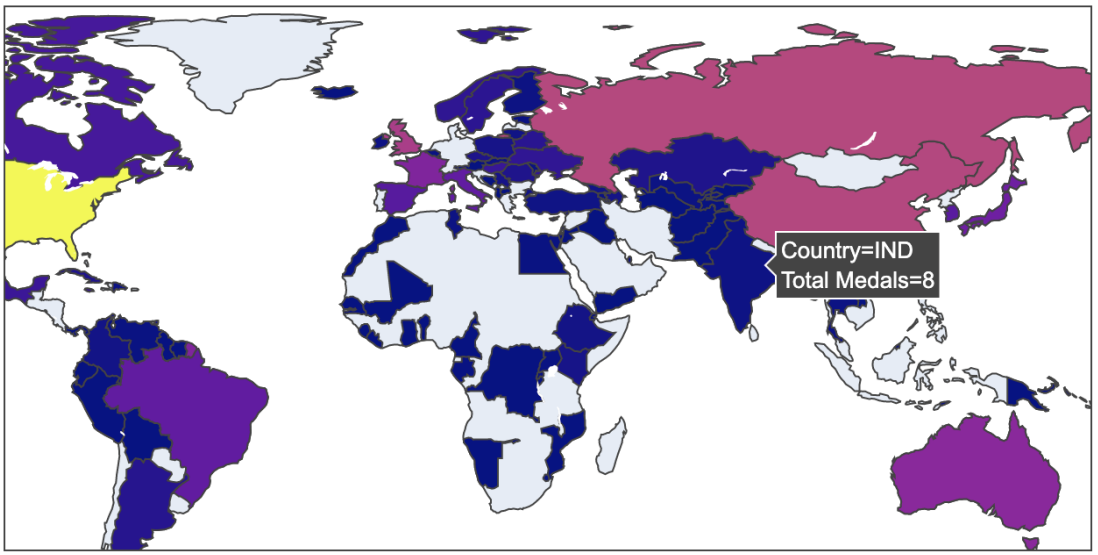


2. The map is interactive, which allows the user to zoom-in, zoom-out, pan or move over the map to show different parts of the map in greater detail or to focus on any particular region.



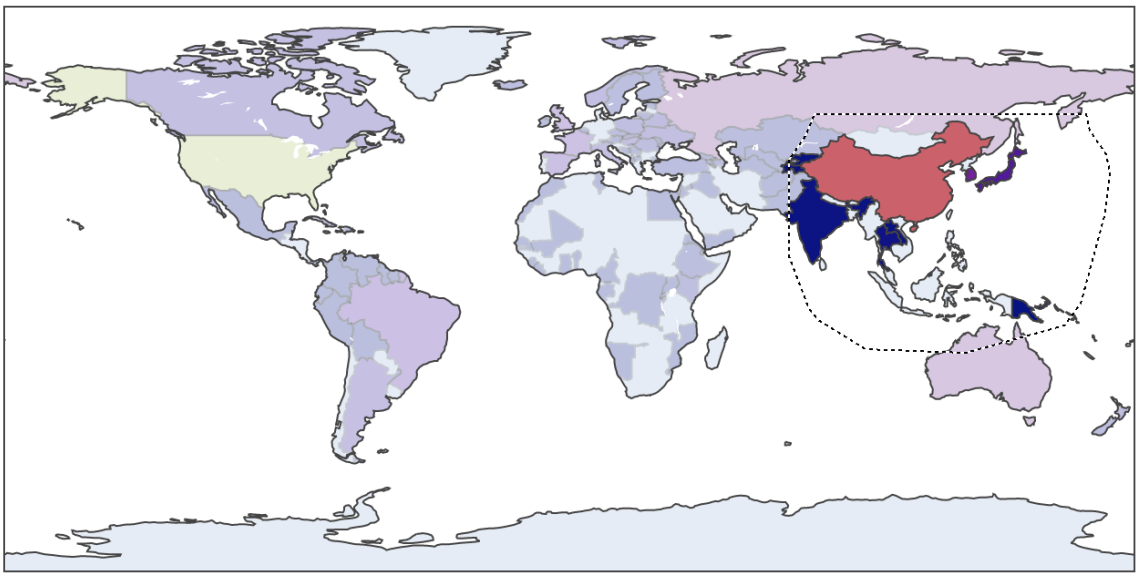


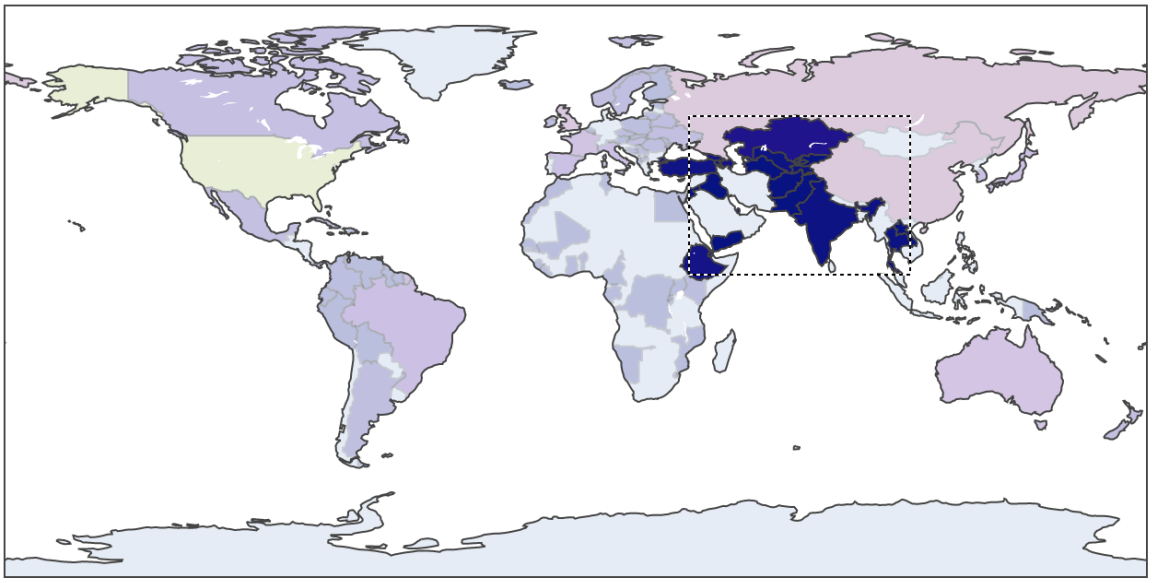
3. Hovering over a particular country in the map shows the total number of medals that the country won that year.



4. The following options are available:
   i) Download plot as a PNG file
   ii) Pan
   ii) Box Select an area of the map
   iv) Lasso Select an area of the map
   v) Zoom in
   vi) Zoom out
   vii) Reset
   viii) Toggle showing data on hover

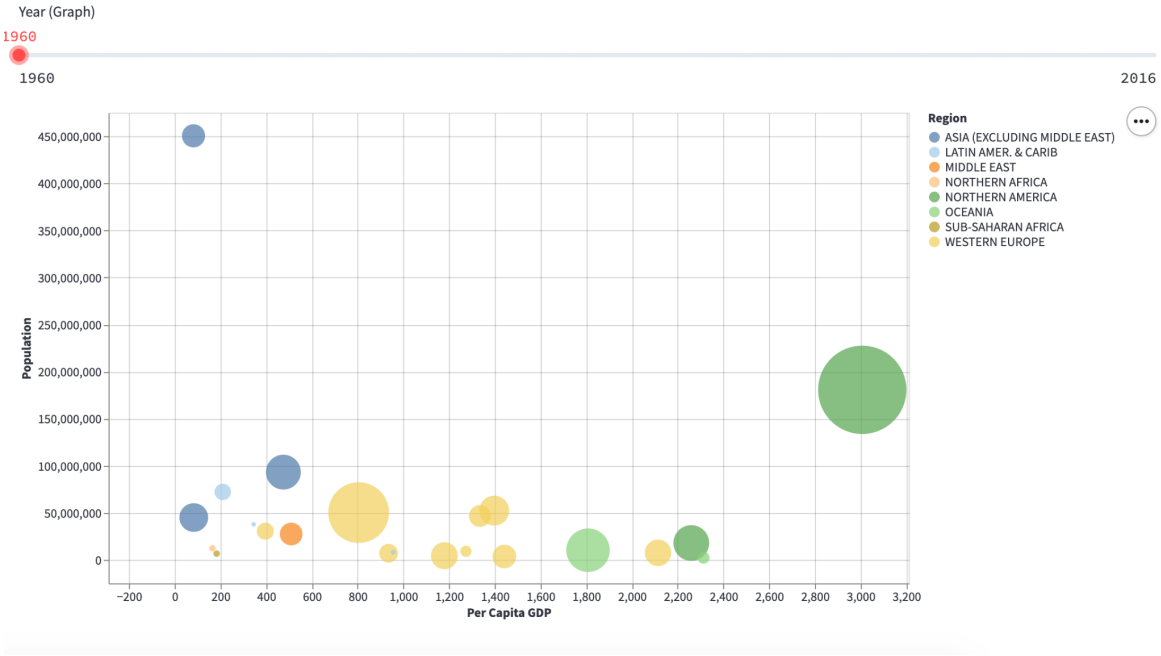
   Some screenshots of Lasso Select and Box Select are shown below:



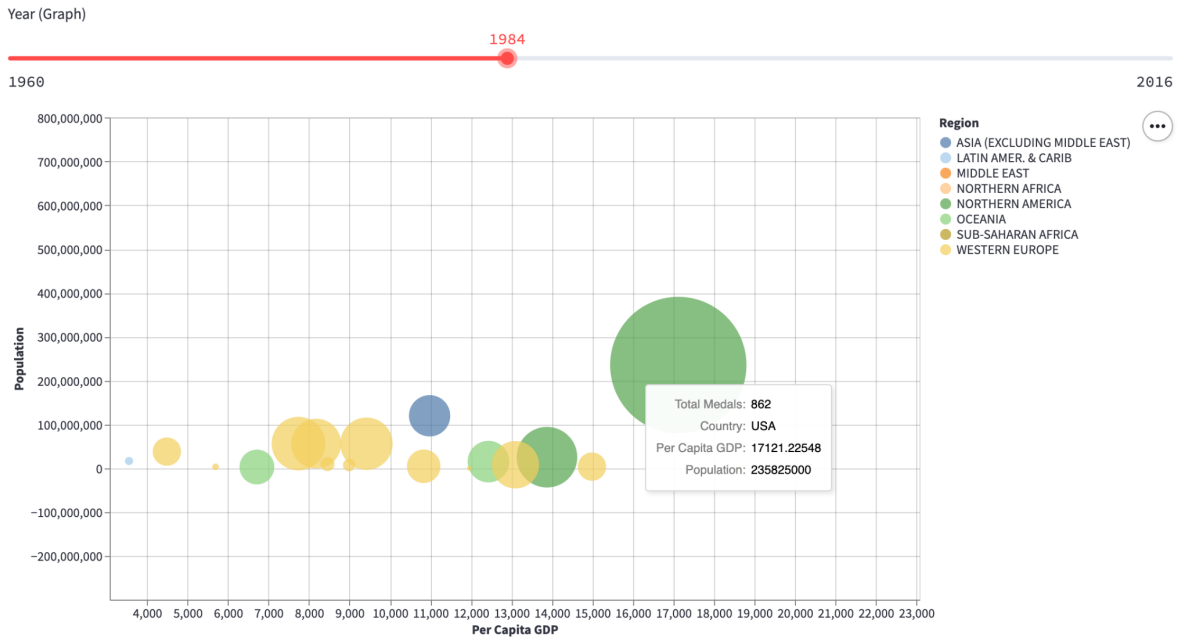


5. A slider to change the year of the Olympic Games

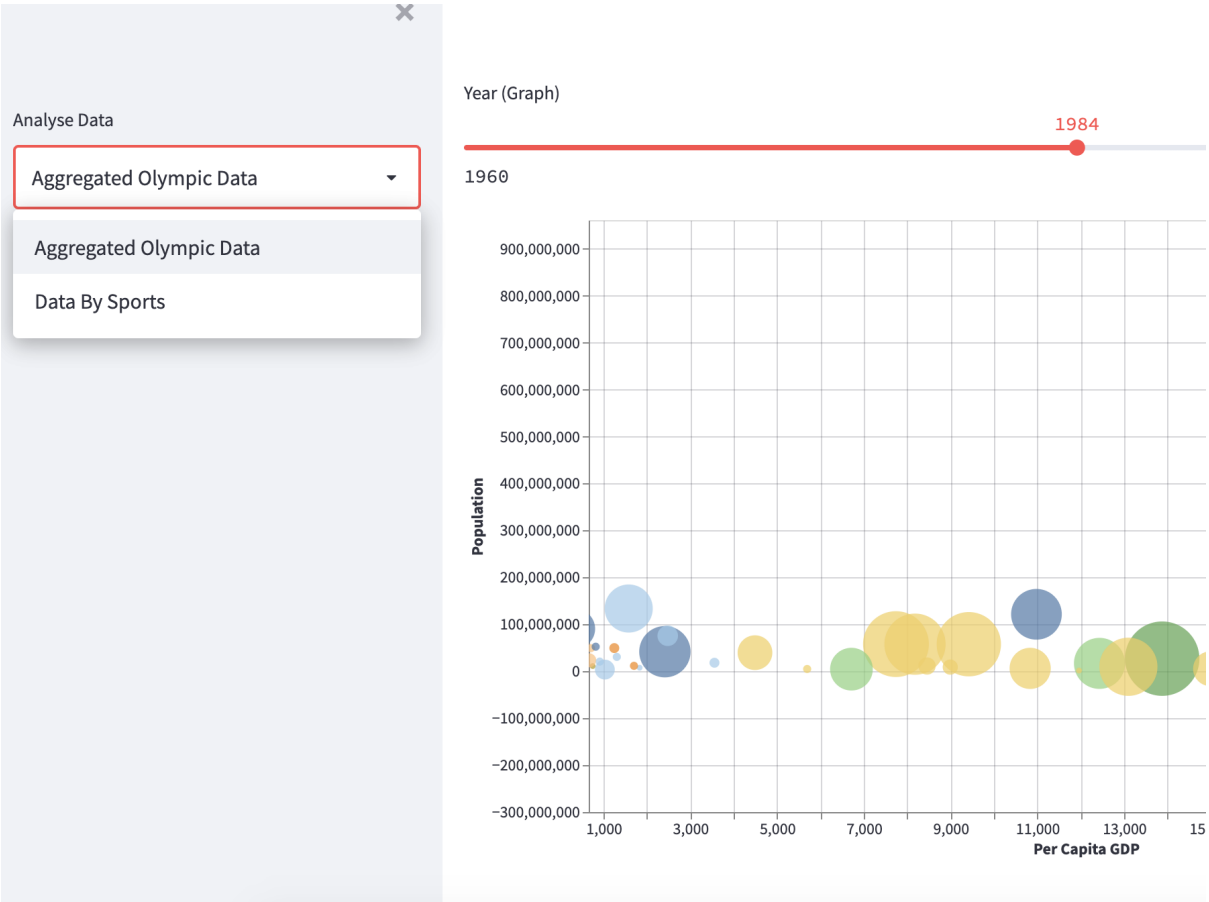
6. A plot showing the total number of medals won by each country along with the Per Capita GDP vs Population.



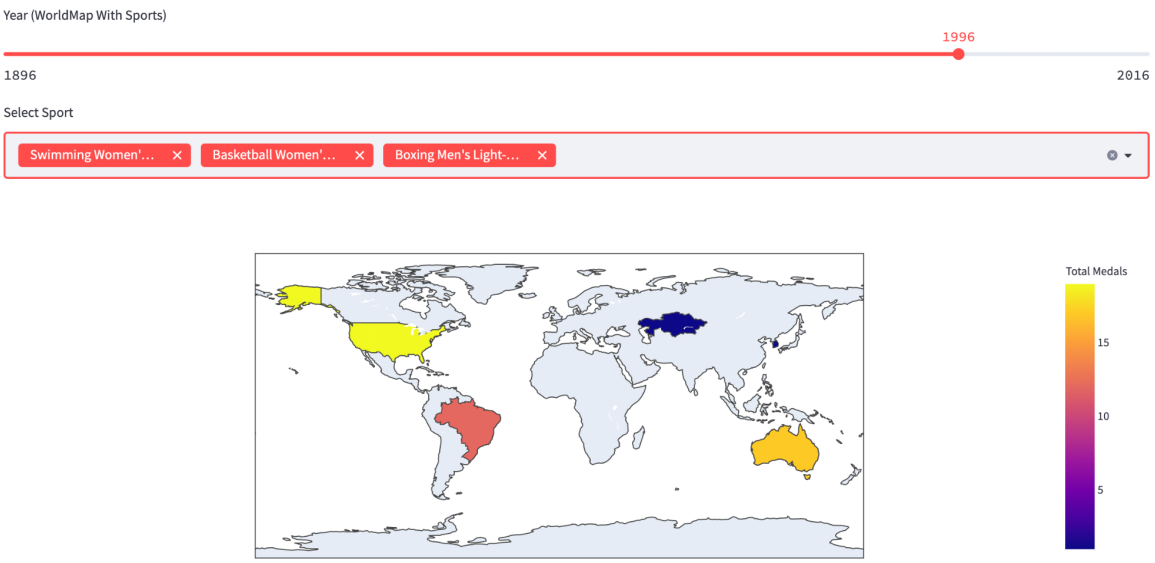
7. The countries are colour coded by region.
8. The graph is interactive, allowing the user to move around, zoom in or zoom-out.
9. Hovering over a bubble shows more details like the country name, the total number of medals won that year, the per capita GDP of the country and its population in that year.



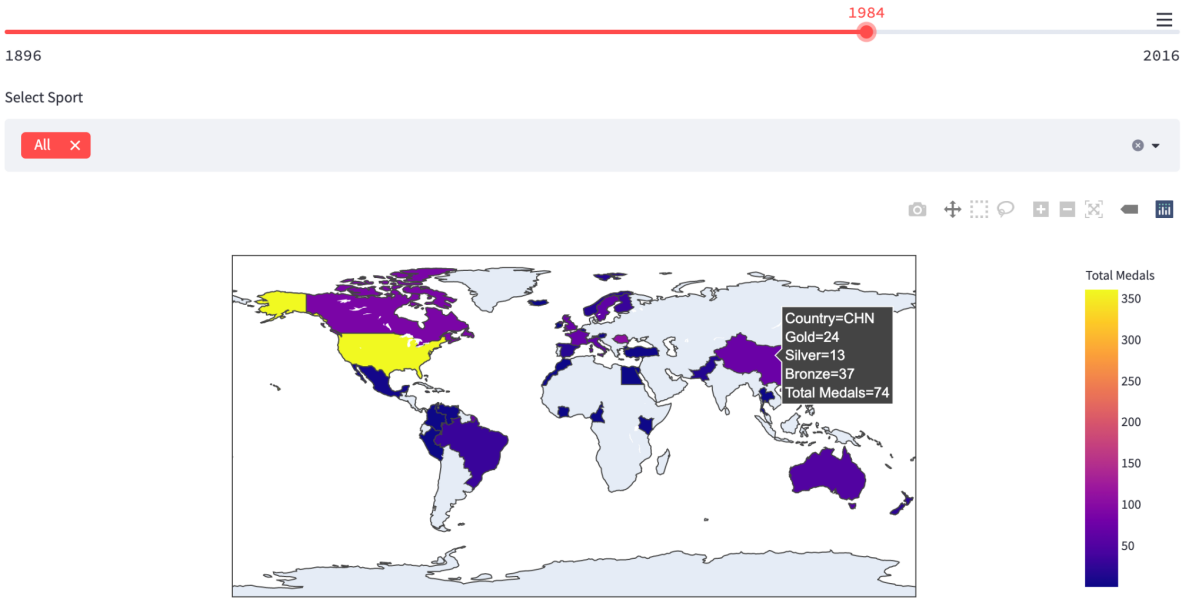
10. The option to save the graph as an image.
11. A selection pane to allow the user to switch between the 2 pages.



12. The second page allows a user to filter the data by sports. The users can select one, multiple or all sports.



Hovering over any country also shows details about the number of medals won for the sports selected.



A user should be able to answer the following questions from the application:

1. Which country won the most medals in a particular year at the Olympic Games?
2. Which country won how many medals in one or multiple sports at a particular Olympic Games?
3. Do specific geographical regions usually win more at the Olympics?
4. Do developed countries usually win more medals at the Olympics?
5. Is the population of a country related to how many medals a country wins at a particular sport?
6. Is a country usually better at particular sports at the Olympics?
7. What was the per capita GDP of a participant country in a particular Olympic year?
8. What was the population of a participant country in a particular Olympic year?

Design Decisions

I chose a global heatmap (heatmap over a world map) to show the Olympic Medal Distribution. This seemed the most effective and visually appealing way to showcase the large amount of information that I wanted to convey. Users can easily see where a country is located and how the global distribution of medals occurred for that year. Some alternatives that I considered were a bar-graph or a pie-chart to display this information but none of them were able to visually show the geographical location of the countries. Therefore, I ended up on this design.

For choosing the Olympic Game Year, I used a slider because it was sleek and easy-to-use as compared to a dropdown.

For the 2nd part, I used a bubble graph to be able to display 4 kinds of information in 1 plot - GDP, Population, Number of Medals as well the region of the country. I made the graph interactive so that users could focus on a particular area if they so desired or if they wanted to download the graph for a particular set of points as an image. I also color-coded the country bubbles by region.

I chose to add a second page for the visualization which allows filtering of data and the exact medal distribution so that users can distinguish between the purposes of the two pages. To allow the users to select the sports, I added a dropdown from which the users can select the sports. I also added an 'All' option here so that the users can see the data for all the sports here without having to manually select each sport.

Hovering over any part of the graph in all these plots shows more detailed information to the user. This allows a clean interface and makes it easy-to-use.

Overview of the Development Process

This was a solo project. It took me about 25 hours to complete this project. This involved finding the right datasets, analyzing them, writing preprocessing scripts, joining datasets, data cleaning and then working on the visualization aspect of the data.

Here is the breakdown of specific tasks in my project:

1. Sourcing the dataset
    a. This involved sourcing different kinds of datasets from different sources such as Kaggle, WorldBank.org and Geolocation data (to map every country to its corresponding boundaries on the world map).
        i. Olympic dataset was sourced from Kaggle (link)
        ii. Historical GDP data was sourced from WorldBank.org (link)
        iii. Historical population data was sourced from WorldBank.org (link)
        iv. Central latitudes/longitudes for each country were derived from real time APIs provided through the geopy python package.
        v. The country to polygon shape mapping data was sourced from https://geojson-maps.ash.ms
2. Data cleaning and aggregation
   This is where I cleaned and joined different datasets.
    a. Notebook (*gdp_vs_medals_data_preprocessing.ipynb*) takes the following steps to combine data:
        i. Convert 'Gold', 'Bronze', 'Silver' to ordinal values of 3, 2, 1 and aggregate medal count by country and year.
        ii. This was then joined with the GDP and population data and the missing data was dropped.

        iii. I used geopy to convert country codes to central latitude, longitude which was later on mapped to a world map through country based polygons.
    b. Notebook (*medals_per_sports.ipynb*) aggregates the medals by sports, country and year.
    c. Final data is stored in csv dumps which is then loaded during visualization.
3. Visualization
   This is where I used streamlit to load and visualize the processed data. We have majorly divided the visualization over two sections:
    a. Section1 (Aggregated Olympic Data) has two plots, one which shows a heatmap distribution of medals won by countries over different years and the second one shows a bubble graph of medals won by country plotted with respect to their per capita GDP and population.
    b. Section 2 (Data By Sports) could be used to see a heatmap distribution of the exact number of medals won by countries over different years, grouped over different sports.

The visualization part took me the most amount of time, especially trying to plot the countries on the worldmap as it is not straightforward and requires augmentation of data. It took me around 14 hours to complete the visualization aspect of this project.