7CCSMSDV Simulation and Data Visualisation

Omar Ahmad
K21052417
Kings College London
Department of Informatics
For Professor Rita Borgo
April 24, 2025

1 Abstract

This data visualisation report explores Olympic performance patterns using three proposed designs. The report analyses national team development trends over time using a multilayered radar chart that encodes medal types and years through spatial positioning and colour. A zoomable circle packing graph reveals performance consistency across Olympic games, using size and colour coding to distinguish between consistent and variable medal retention. And to examine the relationship between economic prosperity and Olympic success, the study will implement an interactive scatter plot matrix that visualises GDP per capita against medal counts across multiple dimensions. Each visualisation design uses intentional design choices in colour, position, size, and interactive features to effectively communicate complex performance patterns in Olympic sports data.

2 Part 1: Analytics

2.1 Exploratory Research Questions

Q1 - Analyse the development of teams' performance over time. Are there detectable trends in the performance of teams over the last decade?

This question was chosen to explore how national teams have evolved in terms of medal achievements and participation in the modern era of the Olympics. This question allows an investigation into the long-term sporting development of these teams. The medals and country performance data is optimal for addressing the question because it provides a continuous historical record of Olympic results across more than a century, This allows for the identification of both long-term and recent trends in team performance with a particular focus on the last decade. The datasets breadth and consistency make it well-suited for time analysis and comparison across nations. The reliability of the Olympic Games dataset is supported by its structural recording of medal counts, athlete participation, and country performance across each Olympics from 1896 to 2024. As an official source, it offers good data integrity, and as it is a comprehensive and consistent dataset, patterns and trends can be detected with minimal data preprocessing [1]. This dataset is highly appropriate for this question however limitations from it include gaps in data due to non participation in certain years and geopolitical changes.

Q2 - Which countries demonstrate the most consistent performance across multiple Olympics versus those showing high variability? By exploring this question, the stability and fluctuations in national athletic performance over time can be studied. Identifying countries with consistent medal outcomes against the countries with high variability within their results can reveal underlying patterns related to long-term investment in sports and changes in national focus, i.e. investing only in specific teams. This question is answered using the same dataset, as it provides the necessary numerical data of medal counts and categorical data of country names and Olympic years needed for the analysis of performance trends across time. This dataset is well suited for this analysis as it provides detailed records of country's achievements at each edition of the Olympics. This historical scope and completeness of the Olympics dataset makes it highly appropriate for identifying long-term trends, however in the same manner for Q1, data sparsity for smaller countries and non participation in some years may affect the results and should be accounted for in the analysis.

Q3 - To what extent does economic prosperity influence a country's Olympic medal success (1984-2020)? This question was chosen to see whether there is a correlation between a country's economic status and its performance in the Olympic Games. This analysis aims to highlight how GDP per capita may influence investment in sports infrastructure, training programs and overall athletic success, internationally. To answer this question, the World GDP per Capita and Annual Growth, and the Summer Olympics datasets are used. The GDP per capita dataset, sourced from the World Bank, includes a range of economic indicators. For the purposes of this analysis, GDP per capita is used as it provides a consistent and quantitive measure of economic conditions across the countries of different population sizes. Both datasets are merged based on country names and overlapping years, resulting in a combined dataset, showing how economic development may relate to Olympic success over time. Both datasets are appropriate for this question as the GDP per capita dataset is sources reliably, covers a sufficient time period to enable long-term trend analysis, and covers a vast range of countries. The Olympic dataset also provides a continuous historical record of athletic performance, which when filtered to match the GDP per capita datasets timespan, supports the analysis.

2.2 Datasets

This project relies on two datasets: The Summer Olympics Medals from 1896 - 2024, and the World GDP per Capita, and Annual Growths from 1960 - 2020. Each provides varied yet complementary data structures and content, making them suitable for answering the answering the proposed exploratory research questions.

2.2.1 Olympics Kaggle Dataset

The Summer Olympics Medals dataset [2] contains structured records of country performance across every Olympic Games since 1896. The dataset includes categorical data such as country names and event types, discrete numerical data in the form of medal counts, and temporal data representing the year of each Olympic Games. This combination of data allows for both cross-sectional and longitudinal analysis. The rich time series allows for finding the patterns in performance over the decades, tracking the rise or decline in medals of specific countries, and comparing the consistency of countries across the years.

In addition to this, the dataset have separate columns for country, year, and medal type, which allows for a targetted study into national sporting trends. This dataset is effective for investigations into teams performance over time, long-term consistency and variability, and gaps in participation due to geopolitical or economic circumstances. The dataset is extensive, and has regular time intervals of 4 years making it reliable for trend identification. When compared with external data, this could also be used to find comparisons across diverse economic contexts too.

2.2.2 World GDP per Capita, and Annual Growths

The GDP per capita dataset [3] provided by Kaggle provides annual economic data for countries globally, from 1960 to 2020. The data type is primarily continuous numerical, representing economic output per person, along with temporal variables. GDP per capita is a widely used measurement for a country's economic status and is particularly useful when comparing countries of differing population sizes.

For this project, the GDP per capita data is cleaned and filtered to cover the years 1984-2020. To clean the data, the steps included linear interpolation, forward and backward filling, and removal of countries with entirely missing data. These preprocessing procedures were chosen to handle missing data by providing reasonable estimates that preserve the overall trend of the various attributes, such as time-series and Olympic performance data.

The cleaned GDP per capita dataset provides a strong foundation for examining how variations in economic capacity can influence a country's investments in sports infrastructure, training programs, athlete development and country-medal performance.

2.3 Data Correlation

To support the analysis required for Q3, which investigates the GDP per capita and Olympic medal performance, two datasets are combined: the Summer Olympics Medals dataset (1896 - 2024) and the World GDP per Capita and Annual Growth dataset (1960-2020). As the datasets span different time periods, to ensure temporal consistency, the overlapping years of 1984-2020 are considered, as this range includes complete data for both GDP per capita and Olympic outcomes following preprocessing.

The combination process involved several key data preparation steps. World GDP per Capita data is cleaned using a script that performs the following transformations [4]:

Linear Interpolation is used for missing middle year values, estimating them based on surrounding data points. Forward Fill is used for missing initial years, by copying the earliest available value backwards. Backward Fill is used for missing end years by copying the latest available value forward. Along with this, the countries with entirely missing GDP per capita data across the relevant years is removed. These steps ensure that the final GDP per capita dataset contains no null values and is ready for integration.

During the pre-processing stage, the datasets are merged using the "Country" and "Year" columns, aligning each country's annual Olympic medal count with its corresponding GDP per capita value, using Pandas [5] Data Frames along with other libraries. This method preserves analytical reliability.

The resulting combined dataset allows for an examination of potential associations between a country's economic status and its success in the Olympic Games.

3 Part 2: Design and Discussion

For each of the three questions, a visualisation is designed which would aid in answering the question. One of these visualisations will be implemented (Q3). All of these visualisations will have important features which aid in understanding the visualisations and adding more dimension to them. These features are size, colour, visualisation type, and can also include scale, other visual properties.

3.1 Question 1

To explore the development of national teams Olympic performance over time, a multilayered radar chart [6] (see figure 1) is created where each axis of the circular chart represents a different country, and each layer corresponds to a specific Olympic year. This design allows for the visualisation of multiple years simultaneously, making it easier to identify trends, such as sustained performance or gradual improvement. The radar chart's circular layout enables clear comparative analysis across countries and across time within a single visual frame.

Medal type is encoded by radial distance from the centre: bronze medals are represented closest to the middle, silver medals in the middle, and the gold medals on the outer ring. This simple spatial arrangement allows the user to immediately understand and see the medal hierarchy. Colour mapping is applied by assigning a distinct colour to each year, helping the viewer identify different years easily. By using translucent layers, the overlapping layers highlight patterns in performance which aids in finding trends in the data over multiple consecutive Olympics.

The performance concentrations shown by the translucent colour overlays allow for a visual identification of which teams are performing best over the years selected by the user. These high contrasting colours allow for an easy depiction of which countries won within specific years. This visualisation technique allows for a powerful representation of multiple data dimensions simultaneously. The radar chart was selected over traditional bar and line charts because of its ability to encode categorical (countries), ordinal (medal type) and temporal (years) data in a compact and insightful layout [7].

The impracticality of 120 years worth of data can instead be fully utilised by managing it through the use of a search bar where the users can define a custom range of years. This gives the user full control over the time period they want to focus on, whist also not overwhelming the visual. By implementing interactive filtering via year selection, and clear visual encodings via translucent colour mapping, the chart is now adaptable to both short term and long term trend analysis, e.g. identifying consistent high-performing teams.

3.2 Question 2

To explore pattern consistency of Olympic performance, specifically medal retention, a zoomable circle packing graph [8] has been produced (see figure 3). This choice allows for layers of data to be shown to the user starting at the country and zooming into individual sports. In this case, to understand the performance over time, the graph has an x axis of time in years and the y axis has number of medals.

Each country will be represented by a circle with size mapping to the consistency of medal wins - larger circles indicate consistent performance across the Olympics, whilst smaller ones show variability in results. This idea was chosen to provide an immediate visual cue of performance stability. As the user zooms in, the circles will dynamically adjust, i.e. unpack, revealing more detailed information about a country's performance in the sports it participated in, all whilst maintaining the same properties of size and consistency. This allows for analysis on larger and smaller scales.

Colour is critical for this visualisation as it indicates the consistency and variability of the teams. Warm tones (see figure 4) like red and yellow represent consistent teams and the warmer the colours (i.e. max. red) show a greater consistency, whilst cool tones such as purple and blue and green indicate variability between each Olympics. The maximum variability here is shown as blue. Using both colour and size encodings to show the consistency enhances clarity for the user. This colour scheme effectively demonstrates good expressiveness for the consistency of medal retention [9].

The performance concentrations shown by the translucent colour overlays allow for a visual identification of which teams are performing best over the years selected by the user. These high contrasting colours allow for an easy understanding of which countries on within specific years. This visualisation technique allows for a powerful representation of multiple data dimensions simultaneously.

An alternative idea was explored (see figure 2) where each circle represented an individual athlete, mapped across the same time and medal axes. However, due to the number of athletes, this approach resulted in lots of overlap in the visualisation, making the zoomable circle packing graph impossible to read properly. This is because many athletes won the same number of medals in the same years, meaning many identical overlapping circles. Due to this the goal of the visualisation could not be reached, and a new idea was explored which would work. This option was eliminated due to poor interpretability.

Overall, the zoomable circle packing graph makes it easy to identify patterns in Olympic medal performance over time. It combines interactivity with a clear visual structure, allowing suers to see both the big picture and specific details. By showing consistency thorough both size and colour, it becomes much easier to compare countries and understand how their performance has changed across different years.

3.3 Question 3

To explore the relationship between economic prosperity and Olympic medal success, a scatter plot matrix is created for the period 1984-2020. This approach allows for a comprehensive examination of the correlation between GDP per capita and Olympic performance. The visualisation is a matrix of scatter plots (see figures 5 and 6), demonstrating high separability by uniquely presenting multiple data dimensions across the multiple connected plots [10]. This reveals nuanced relationships between economic indicators and athletic achievements of a country to the viewer.

It leverages several visual channels such as colour, position and size to encode this multivariate data, enhancing data comprehension, to help viewers quickly identify both micro and macro patterns across different attributes with the help of a zooming feature [11].

Each panel represents a different dimension of the data: average GDP per capita, total medals, and temporal changes by year. The design also allows for brushing and linking, where highlighting a selection in one panel dynamically filters and highlights corresponding data points across all plots [12].

Countries are colour-coded by continent to improve readability, given the number of countries in the dataset. Each mark represents a country, with different colours representing continents and creating distinct data bins [13]. An interactive hover feature enables users to quickly view individual country marks, revealing precise metrics such as specific GDP values and medal counts. Interactive features like mouse-wheel zooming and click/drag selection allow users to explore fine details as well as broader trends, allowing for multiple levels of viewer engagement with the data presented.

By presenting multiple plots at the same time, the visualisation invites users to investigate and draw their own conclusions about the multi-faceted relationship between national economic performance and Olympic medal-winning achievement, telling a visual story of global sporting success.

References

[1] Dorian Pyle. *Data Preprocessing for Machine Learning and Data Science*. Morgan Kaufmann, 1999. 1

[2] Stefany de Oliveira. Summer olympics medals dataset (1896 - 2024). Kaggle Dataset, 2024. 1

[3] ZGRC Meta. World gdp per capita and annual growth dataset (1960 - 2020). Kaggle Dataset, 2024. 2

[4] Zhengping Che, Shuai Yuan, Wei Long, and Fei Liu. Missing data imputation: Methods and applications. *IEEE Transactions on Knowledge and Data Engineering*, 31(6):1011–1027, 2019. Detailed analysis of linear interpolation, forward and backward filling techniques. 2

[5] The Pandas Development Team. Pandas: Powerful python data analysis toolkit, 2024. Version 2.x Documentation. 2

[6] Robert Kosara and Falko Bendix. Multivariate data visualization using multilayered radar charts. *IEEE Transactions on Visualization and Computer Graphics*, 21(7):808–817, 2015. 2

[7] Claus O. Wilke. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O'Reilly Media, 2019. 2

[8] Michael Balzer and Oliver Deussen. Visualization of hierarchical data: Circle packing techniques. In *IEEE Symposium on Information Visualization*, pages 57–64. IEEE Computer Society, 2005. 2

[9] Colin Ware. *Designing with Color: Perception, Interpretation, and Use in Information Visualization*. Morgan Kaufmann, 2021. 2

[10] Scott Murray. *Interactive Data Visualization for the Web*. O'Reilly Media, Inc., Sebastopol, 2017. 3

[11] Tamara Munzner. *Visualization analysis and design*. A.K. Peters visualization series. CRC Press, Taylor & Francis Group, Boca Raton, 2015. 3

[12] Matthew Ward, Georges G. Grinstein, and Daniel Keim. *Interactive data visualization: foundations, techniques, and applications*. CRC Press, Taylor & Francis Group, Boca Raton, second edition edition, 2015. 3

[13] Jacques Bertin and William J. Berg. *Semiology of graphics: diagrams, networks, maps*. ESRI Press, Redlands, Calif, 1st ed. edition, 2010. 3

APPENDIX A

- **Summer Olympics Medals Dataset (1896 - 2024):**
- https://www.kaggle.com/datasets/stefanydeoliveira/summer-olympics-medals-1896-2024
- **World GDP per Capita Dataset (1960 - 2020):**
- https://www.kaggle.com/datasets/zgrcmeta/world-gdpgdp-per-capita-and-annual-growths

.1 Figures of the Prototypes for Part 2: Design and Discussion

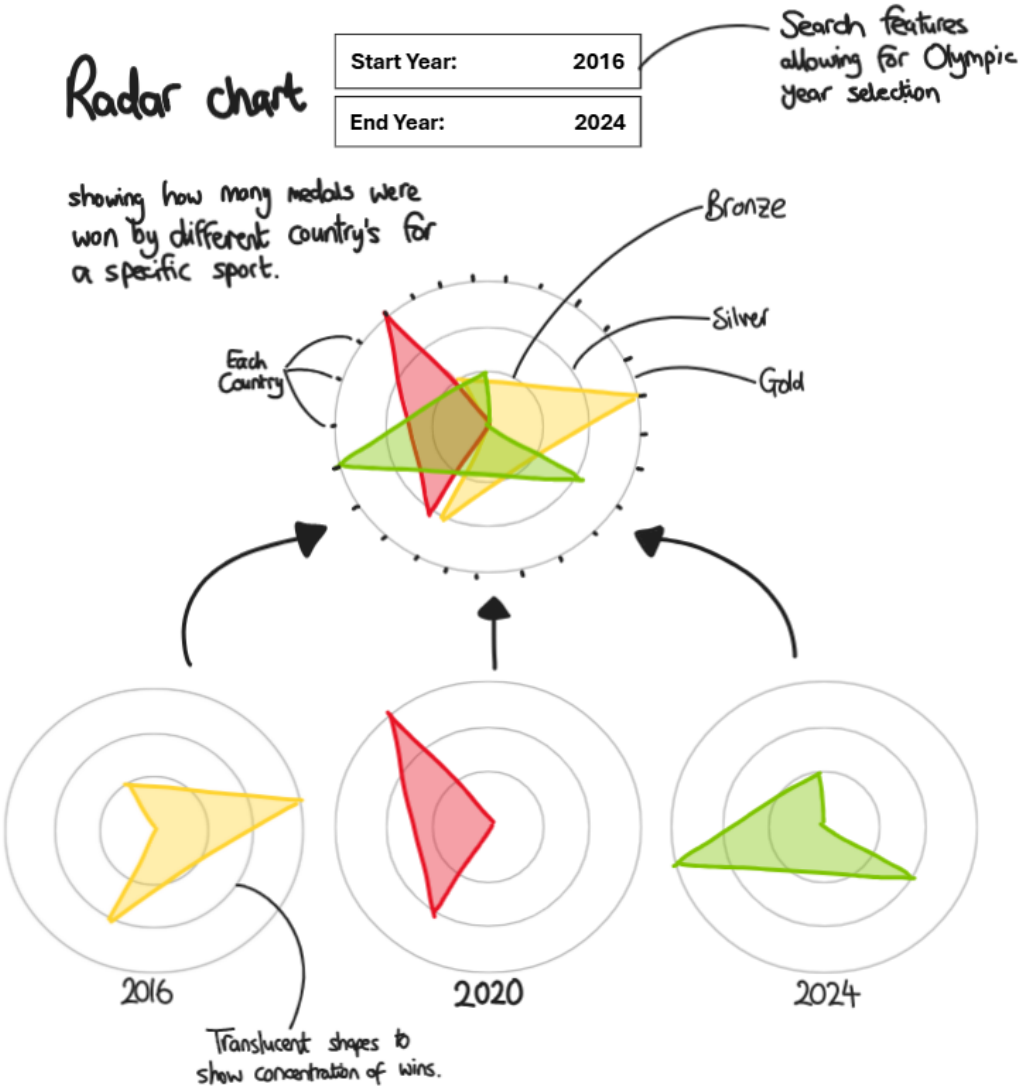

Figure 1: Radar Chart Design

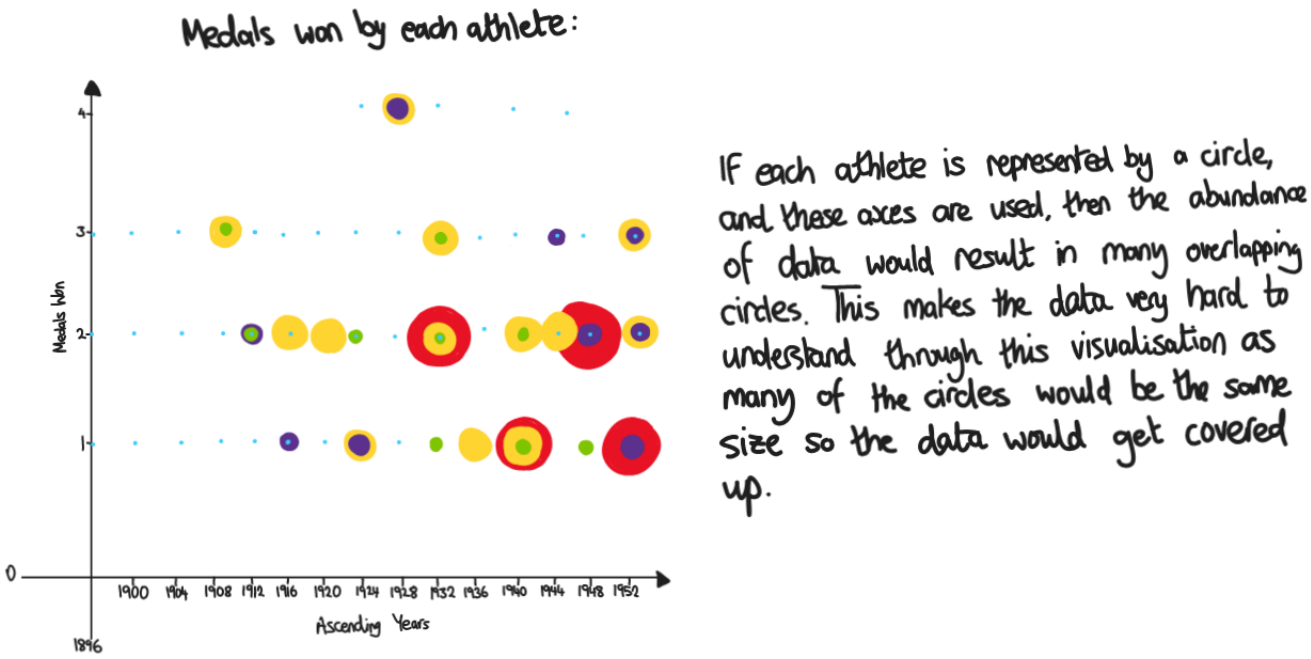

Figure 2: Number of Medals won by Athlete

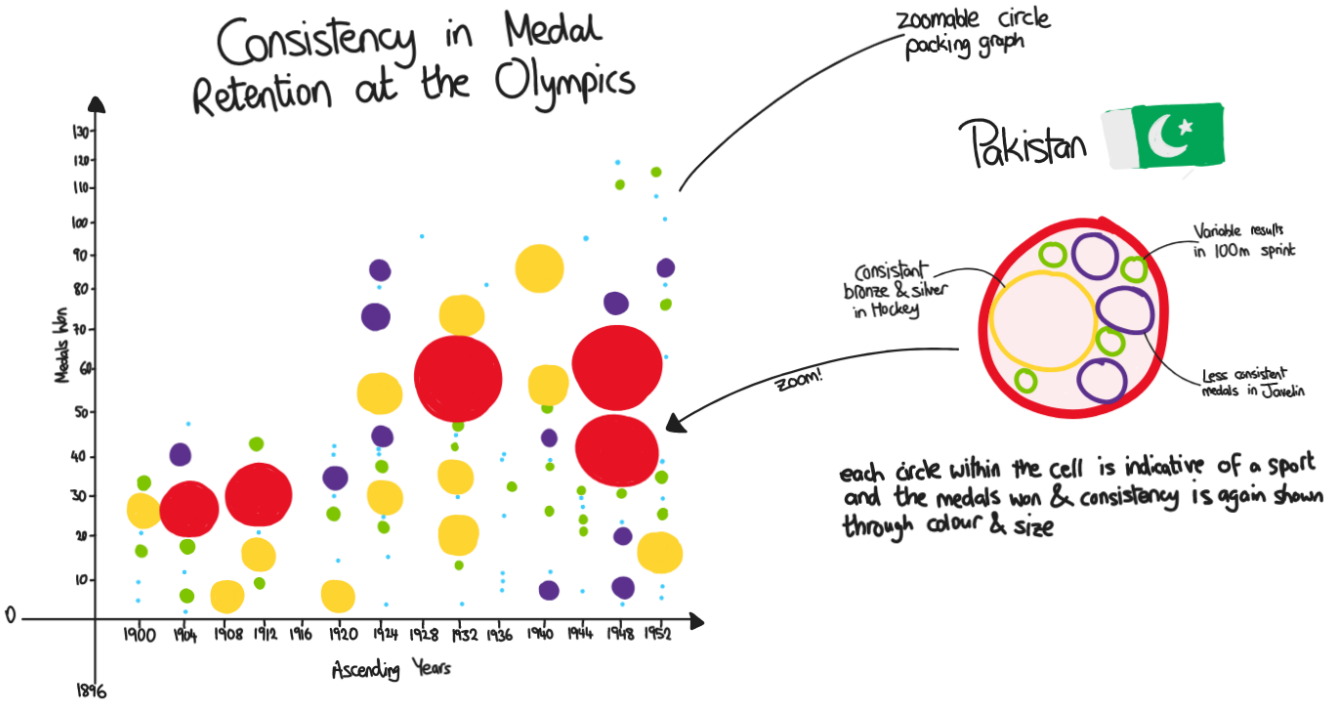

Figure 3: Zoomable circle packing graph design showing consistency in medal retention at the Olympics

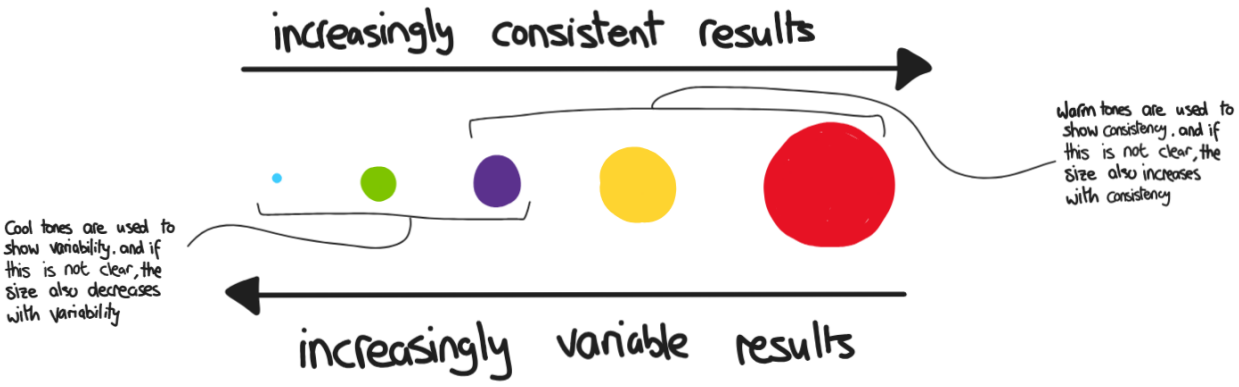

Figure 4: How colour is used to show consistency within the zoomable colour packing graph

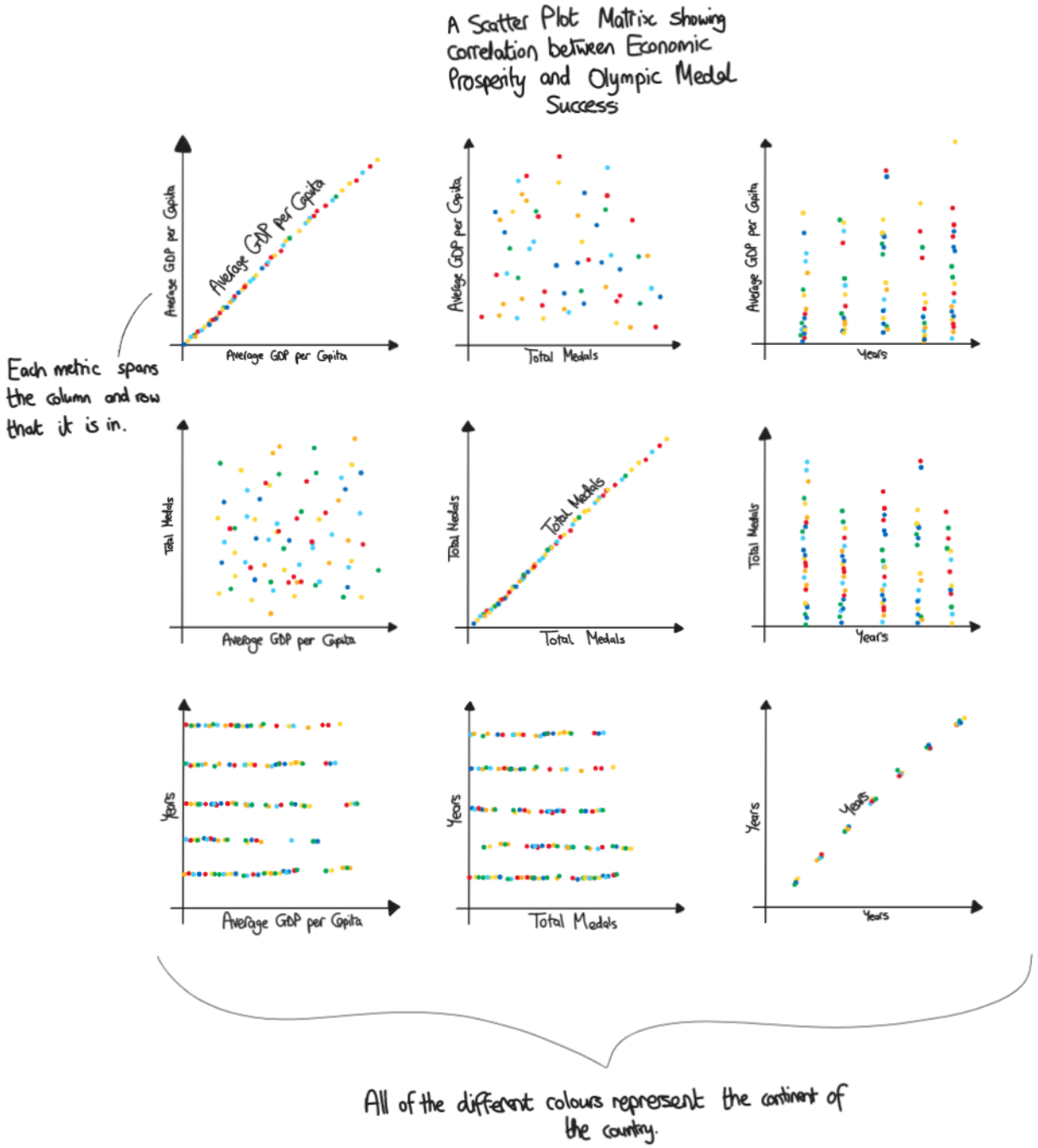

Figure 5: Scatter Plot Matrix Design

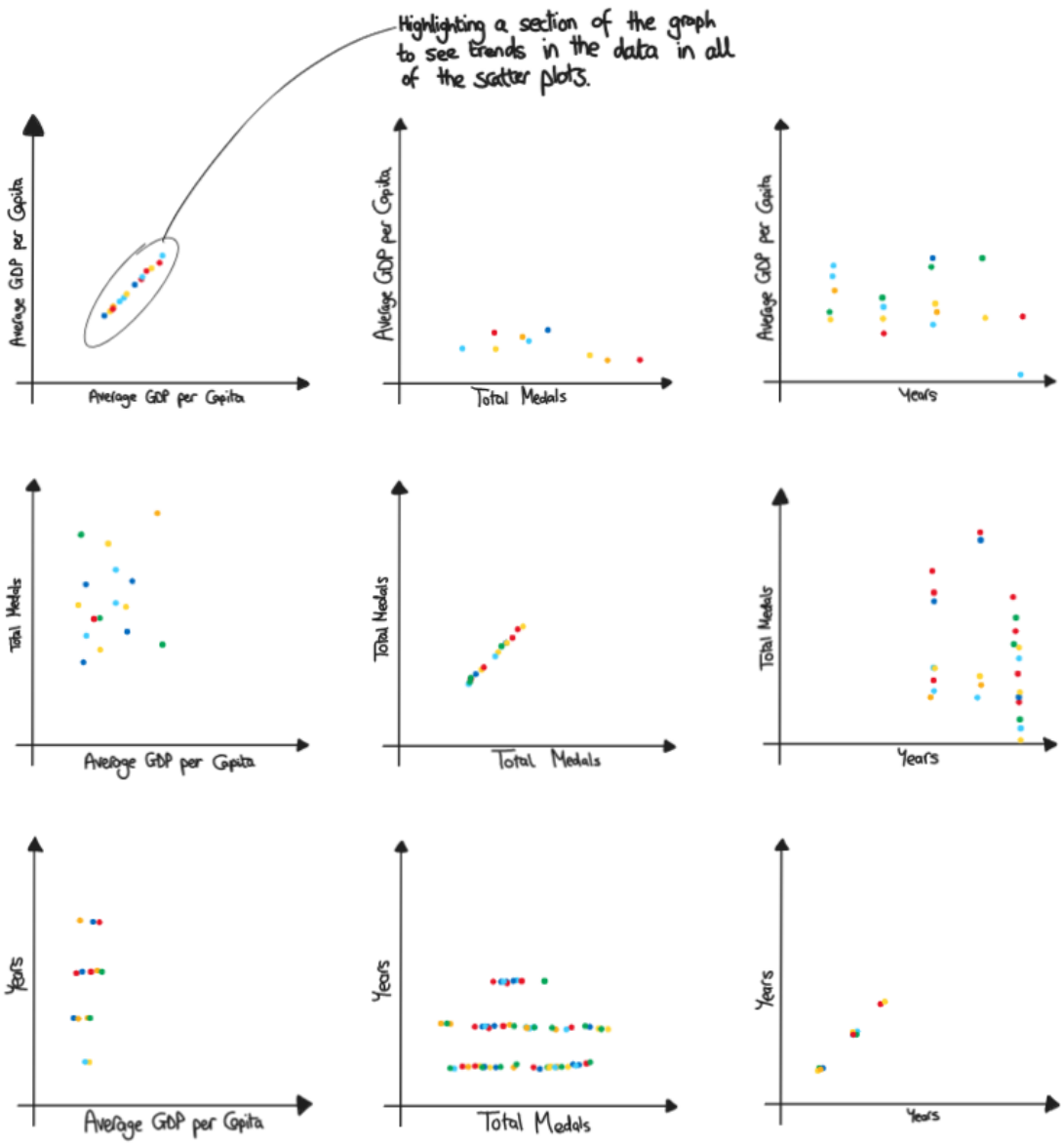

Figure 6: Highlighted area for Scatter Plot Matrix Design