**7PAM2000 Applied Data Science 1**

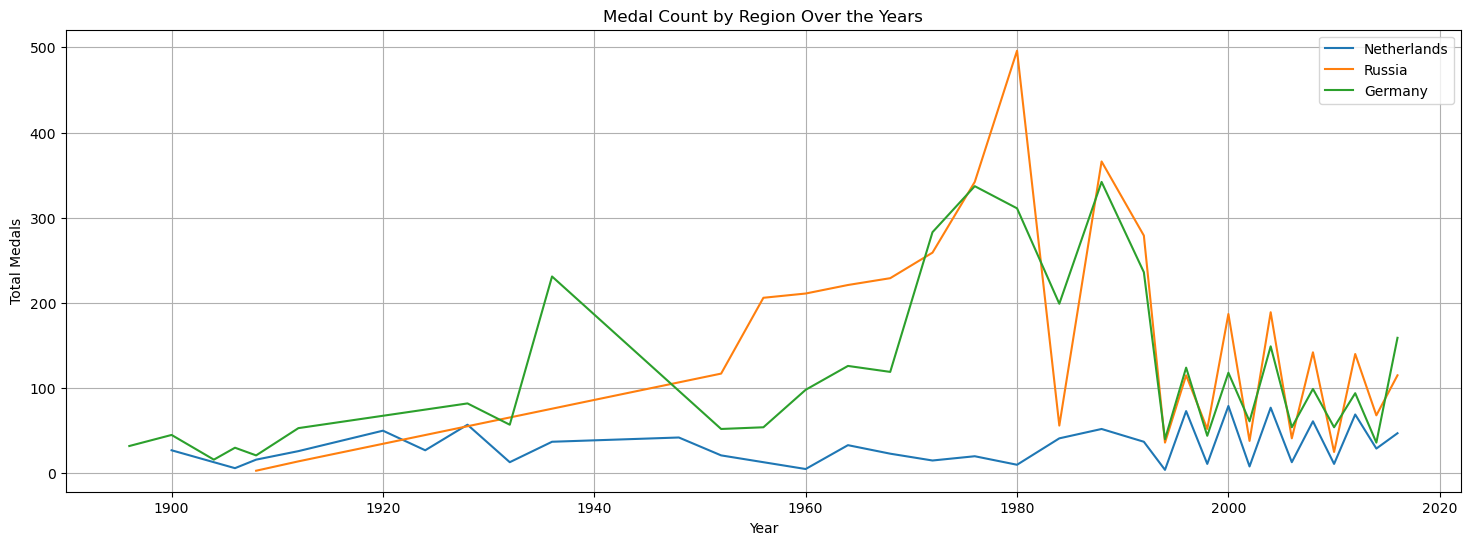
**Assignment 1: Visualisation**

**Name:** JAMAL NASIR

**Github Repository:** <https://github.com/jamalnasir551/jamalnasir551-jamalnasir551-Applied-Data-Science-1>

**Data Source:** <https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results>

**Visualization 1 - Medal Count Over Time by Region**

****

**Line Plot:**

A line plot is a type of graph that displays data points as individual markers and connects these points with lines to show how the data changes over a continuous interval, such as time. In a line plot, the x-axis typically represents the continuous variable (e.g., time), and the y-axis represents the value of interest.

Line plots are commonly used to visualize and analyze trends, changes, or patterns over time.

**Function's Purpose and Line Plot**:

The function plot\_medal\_count\_by\_region\_over\_years uses a line plot to display the distribution of Olympic medals (Gold, Silver, Bronze) by region over multiple years.The primary reason for using a line plot in this case is to visualize and compare the temporal trends in medal counts for various regions. It helps in understanding how their performance has changed over time.

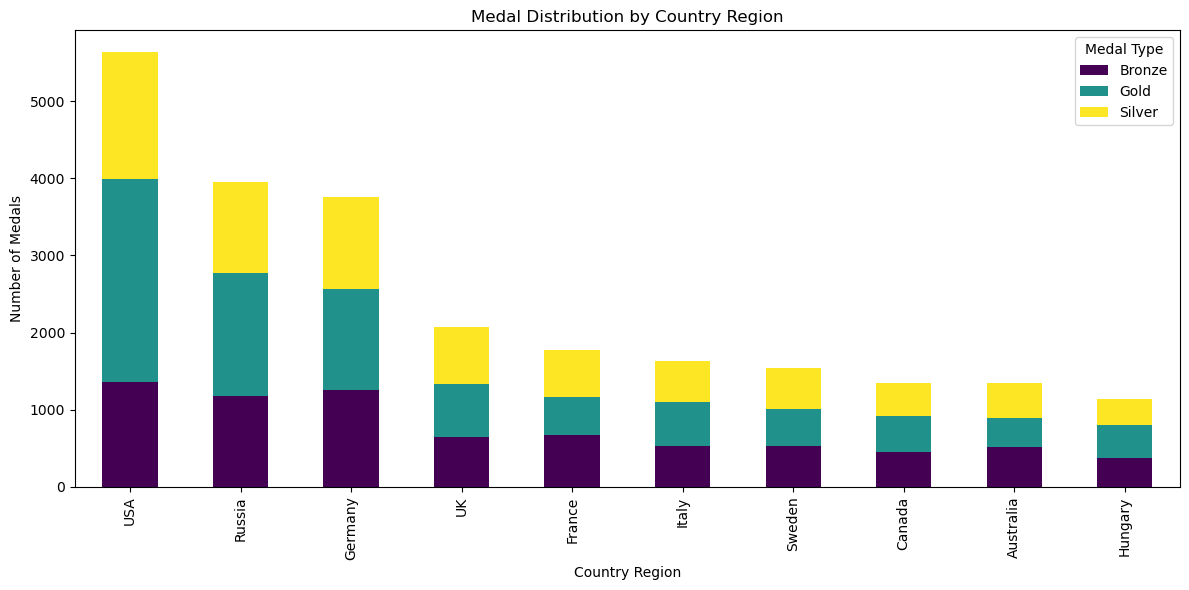
By using a line plot, the function enables a clear and effective representation of how these three regions have fared in the Olympics over the years.

**Insights and Sample Run:**

A sample run using the regions Netherlands, Russia, and Germany reveals that Russia had an unusually large number of medals in the year 1980. This insight might not have been immediately apparent without the line plot.

The line plot allows you to see a spike in Russia's performance in 1980, indicating a remarkable year for the country in terms of Olympic success.

**Visualization 2 - Age Distribution of Olympic Athletes**



**Stacked Bar Plot:**

A stacked bar plot is a type of chart that represents data using rectangular bars. Each bar is divided into segments, with each segment representing a different category or sub-component of the data. The height of each segment corresponds to the quantity or proportion of that category within the whole.

Stacked bar plots are commonly used to display the composition of a whole in terms of its components.

**Function's Purpose and Stacked Bar Plot:**

The function plot\_medal\_distribution\_by\_country\_region employs a stacked bar plot to visualize the distribution of gold, silver, and bronze medals by country region.

It focuses on the top N country regions (e.g., top 10) with the most total medals to provide a clear overview of how medals are distributed among these regions.

The primary reason for using a stacked bar plot in this case is to illustrate the relative contributions of each type of medal (Gold, Silver, Bronze) within each top country region. It helps in understanding the medal distribution within the selected regions.

By using a stacked bar plot, the function effectively conveys the composition of medals by type for the chosen country regions.

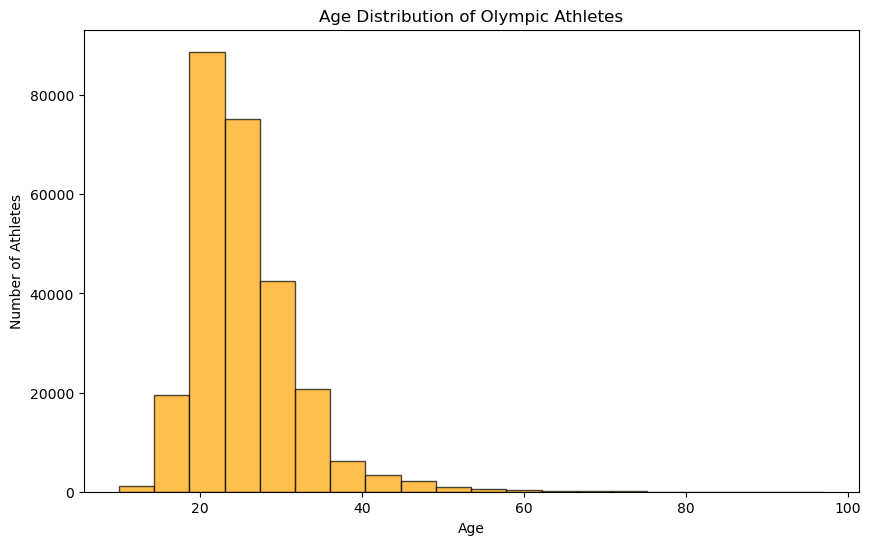
**Insights and Sample Run:**

A sample run using the top 10 countries reveals that the USA has the most medal wins in terms of Gold, Silver, and Bronze combined. This insight is quickly discerned through the stacked bar plot.

The stacked bar plot clearly shows the proportional distribution of medal types for each of the top regions, making it easy to identify the region with the highest overall medal count.

In this case, the visualization highlights that the USA has been highly successful across all types of Olympic medals.

**Visualization 3 - Medal Distribution by Country Region**



**Histogram Plot:**

A histogram is a graphical representation of the distribution of a dataset. It divides the data into "bins" or intervals, counts the number of data points that fall into each bin, and then represents this information as vertical bars. The height of each bar corresponds to the frequency or count of data points in that bin.

Histograms are used to understand the underlying data distribution and to visualize the frequency or probability distribution of continuous or discrete data.

**Function's Purpose and Histogram Plot:**

The function plot\_age\_distribution utilizes a histogram plot to display the age distribution of Olympic athletes.It's employed to show the frequency or count of athletes within different age groups. In this case, the data is grouped into "bins" representing specific age ranges (e.g., 20-24, 25-29, 30-34, etc.).

The histogram is ideal for revealing patterns in the age distribution, such as identifying peaks, modes, or variations in the dataset.

By using a histogram, the function effectively presents the distribution of athletes' ages in a clear and interpretable manner.

**Insights and Sample Run:**

In a sample run using your dataset, it was found that the majority of athletes were in the age group of 20, which indicates a strong representation of young athletes. Additionally, there were substantial numbers of athletes under the age of 18.

The visualization also showed that there were relatively fewer athletes in the age group of 50 and above, suggesting a lower presence of older athletes in the dataset.

Overall, the histogram provides insights into the age demographics of Olympic athletes, including the prevalence of younger participants and the distribution of athletes across different age ranges.