Olympics Dataset (1976-2008) Analysis

AN EXPLORATORY DATA ANALYSIS USING EXCEL, R AND POWER BI

Overview

The Olympics dataset is a list of all medal winners from the year 1976 Montreal to 2008 Beijing. In this project, our main focus is to prepare the data, work on the EDA , as well as to provide summary, insights and recommendation related to various sports, countries, event, medal over the given time.

The tools implemented in this analysis are Microsoft Excel, R and Power BI.

Objectives

To find out which countries and athletes won the most medals in the Olympics from 1976 to 2008.

To See how medal wins changed over time that is, which years were more successful and which sports became more popular.

To compare male and female participation.

To interpret and visualize the outcome of analysis.

Steps Involved

Data collection

Data cleaning and preparation

Exploratory data analysis

Visualization

Insights

Summary

Data Collection

The data has been collected in the form CSV file named "Olympics dataset.csv"

The CSV file contains data of 15316 participants from 1976 to 2008. It also contain relevant columns like Year, Event, Sport, Gender, Medal, City, Discipline, Athlete, Country, Country_code, Event_gender in their respective order in the data important for the analysis.

Data Cleaning and Preparation

The data was cleaned and prepared mostly using Microsoft excel. Firstly, the data was imported into excel where I checked for missing values using the **countblank()** function, which enabled us to know that there are no missing values in the data.

Secondly, I imported the csv file into R using the **read.csv** function were I analyzed the structure of data to understand the fact that almost all columns in the data were of character type which was converted to factor in order to make the analysis easy, that is to treat the data as categories and not just text.

Exploratory Data Analysis

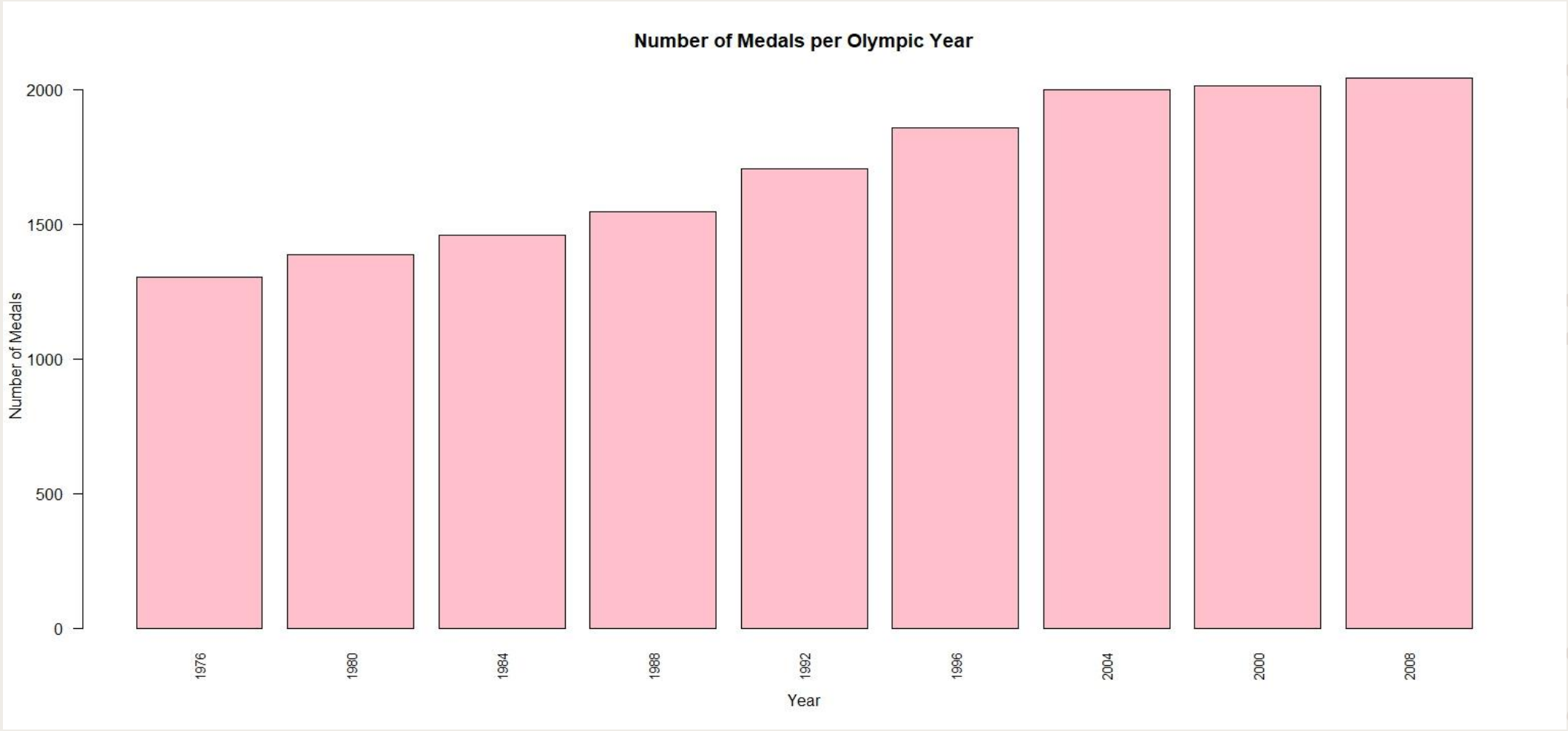
In this part, I analyzed and visualized the data using R, where the data was manipulated and transformed using the 'dplyr' package and visualized using the 'ggplot' package.

The main questions analyzed are:

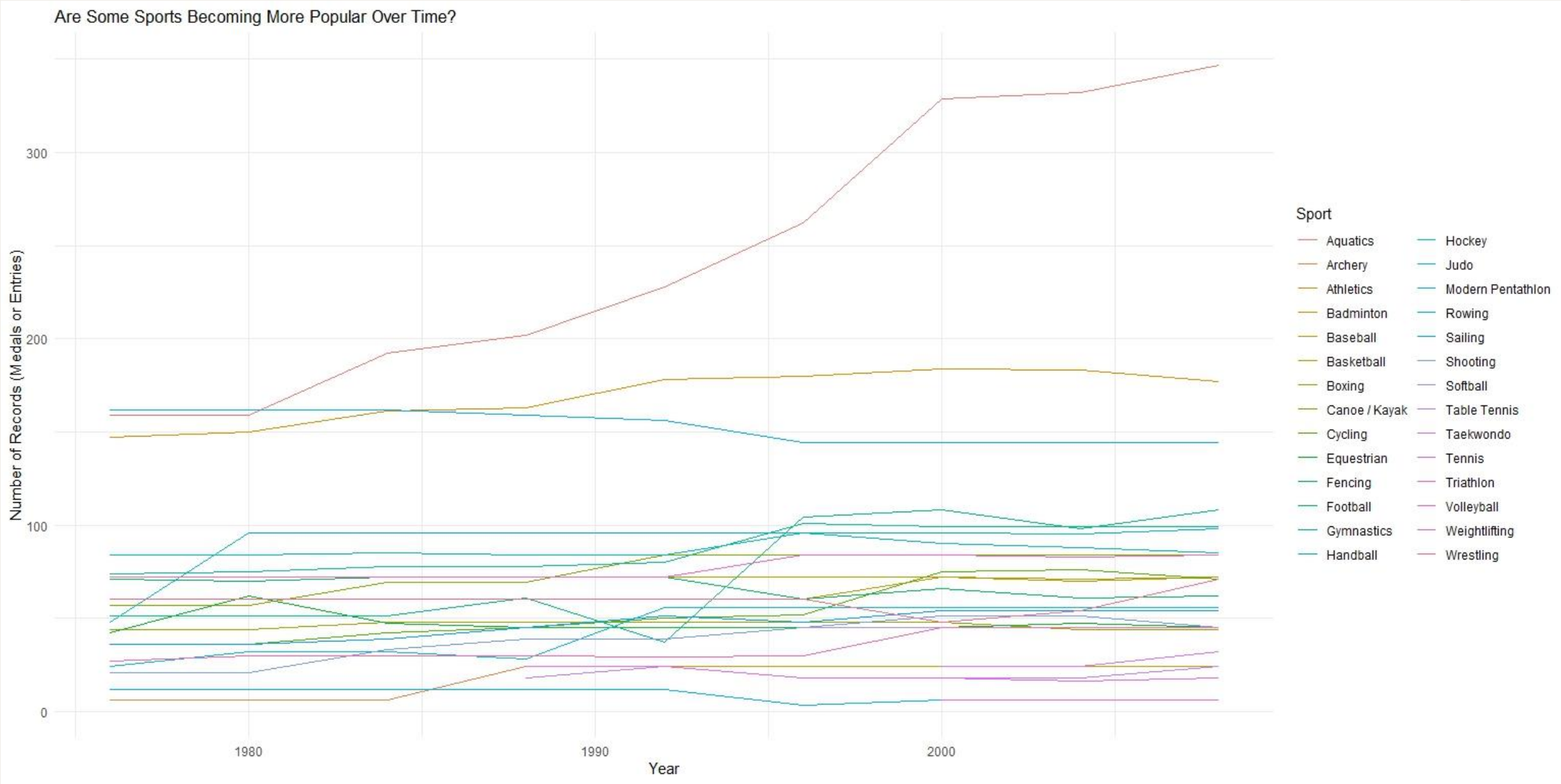
- **The number of medals won in each year**
- **Popularity of sports over time**
- **Countries that won the most medal**
- **Most successful athletes**
- **Top sports for each country**
- **The number of male vs female athletes who won the medal**
- **Sports and events that are more balanced in terms of gender**

1) Number of medals won in each year

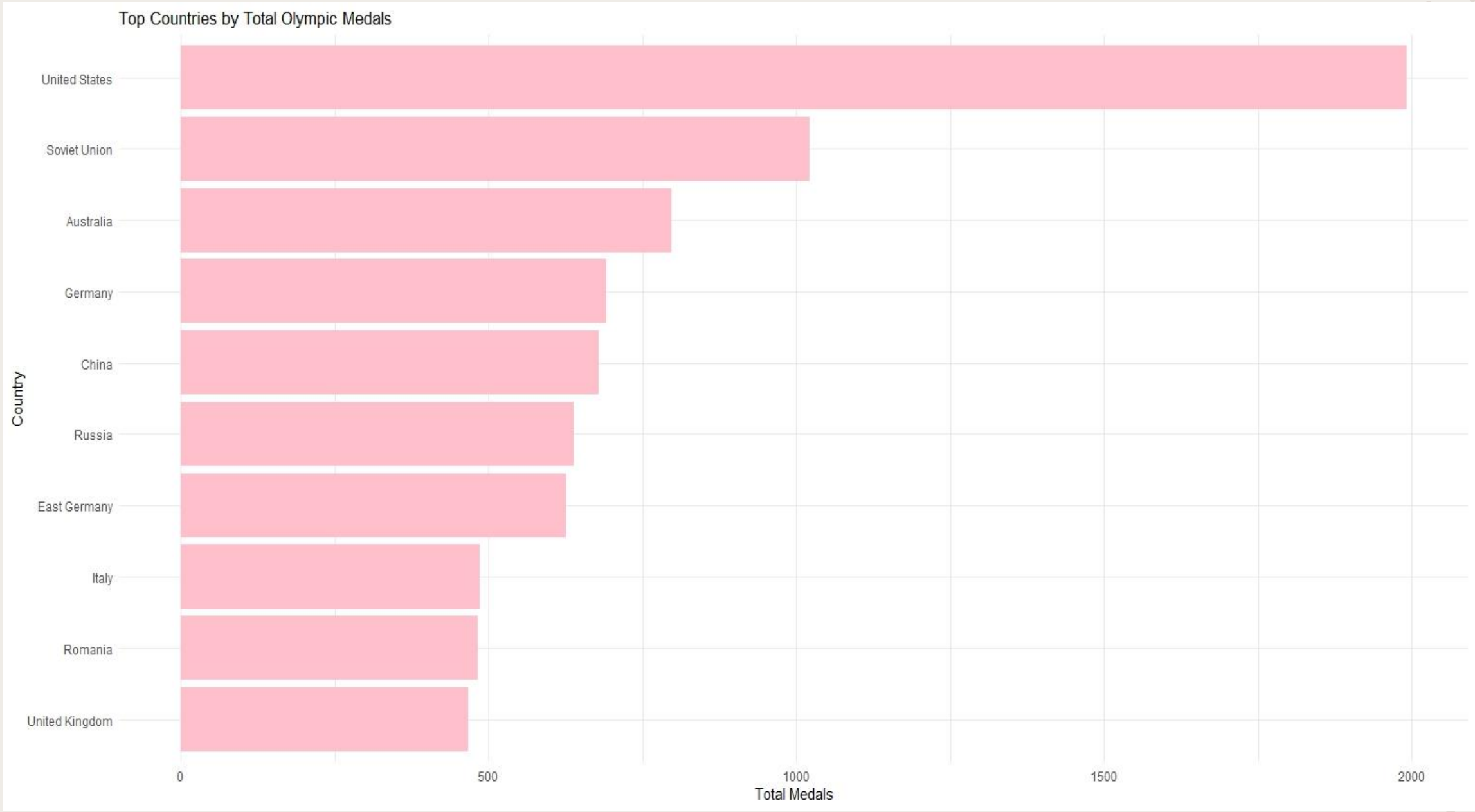
1976	1980	1984	1988	1992	1996	2000	2004	2008
1305	1387	1459	1546	1705	1859	2015	1998	2042



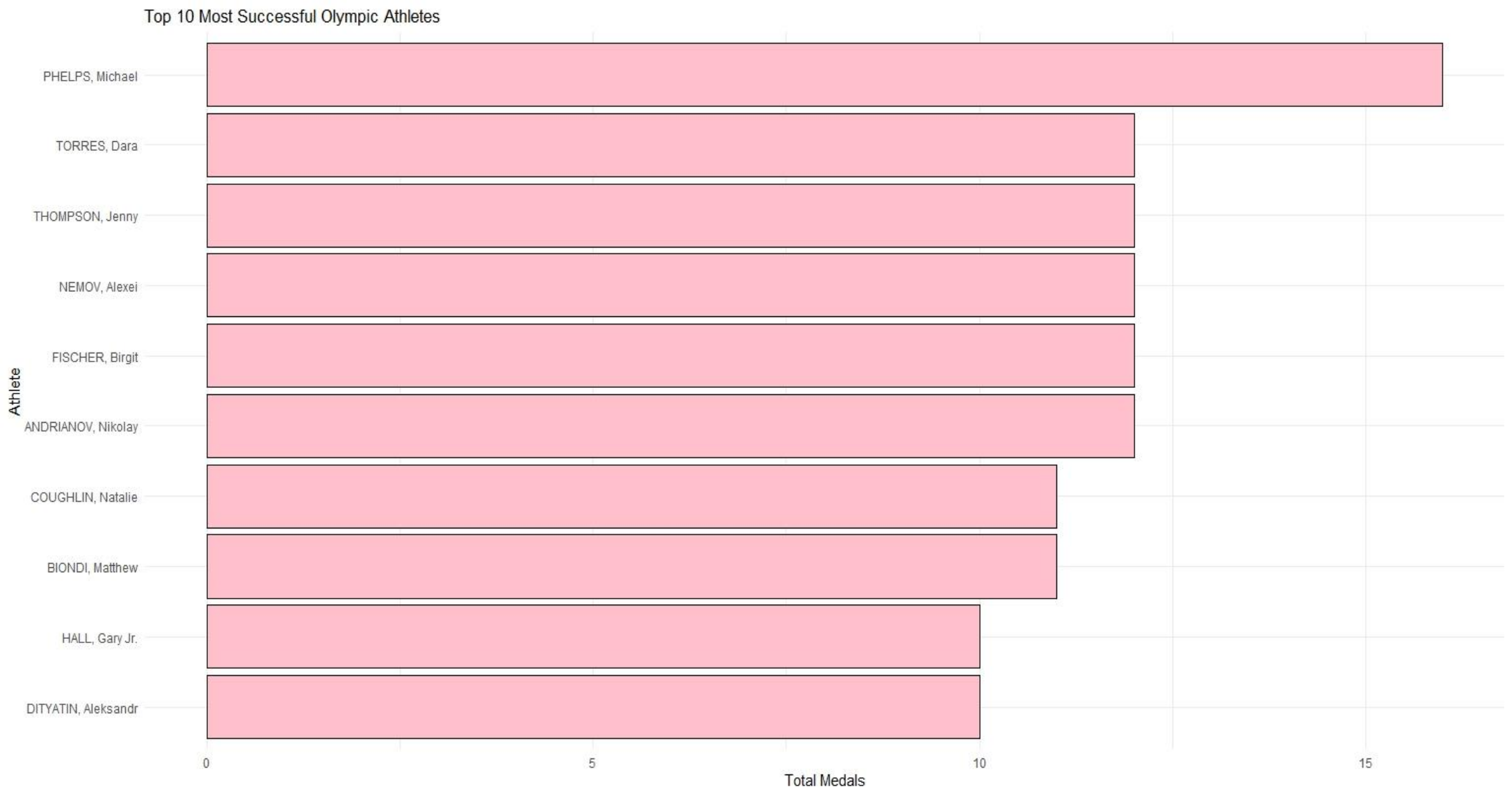
2)Popularity of sports over time



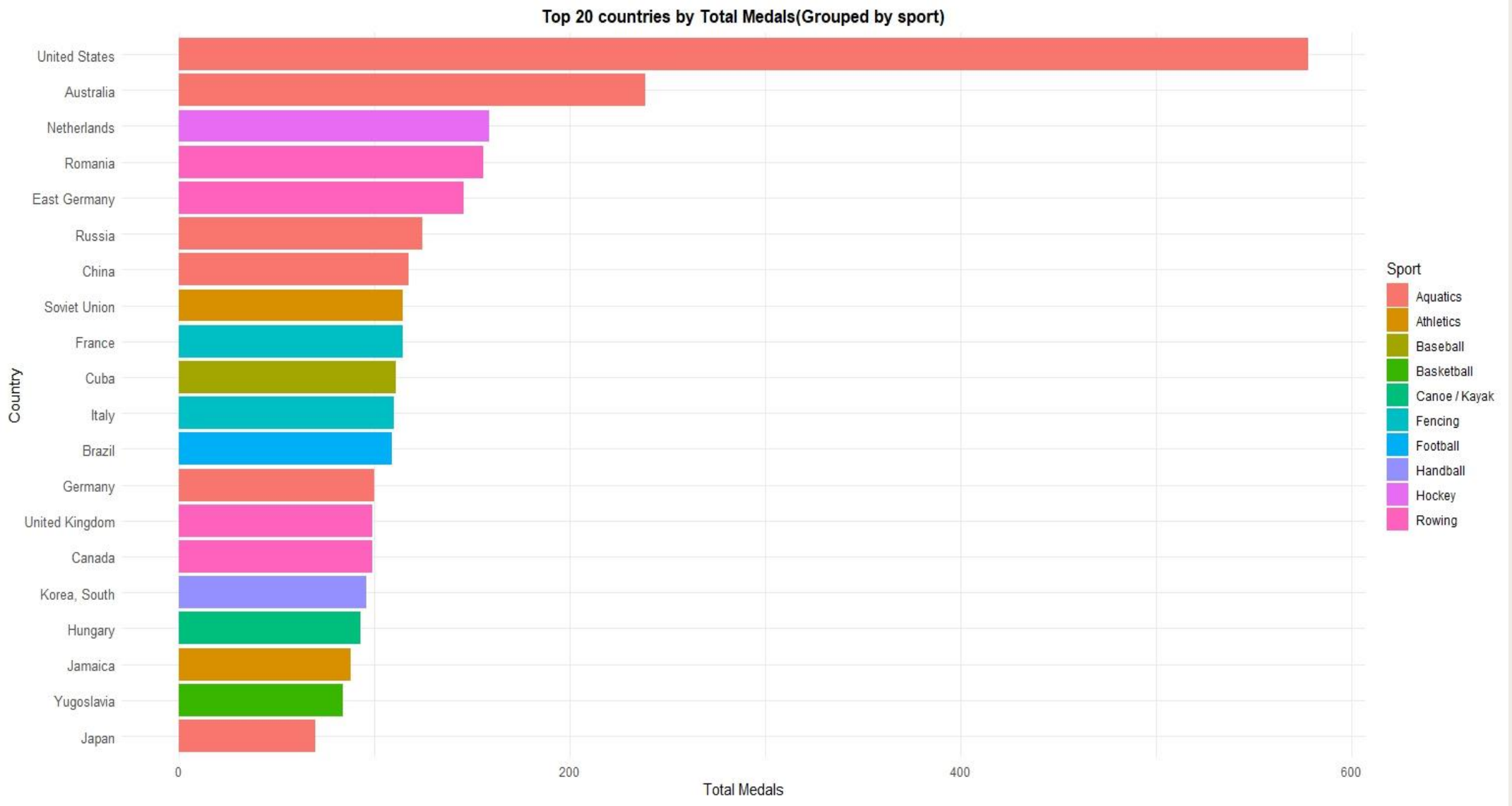
3)Countries that won the most medal



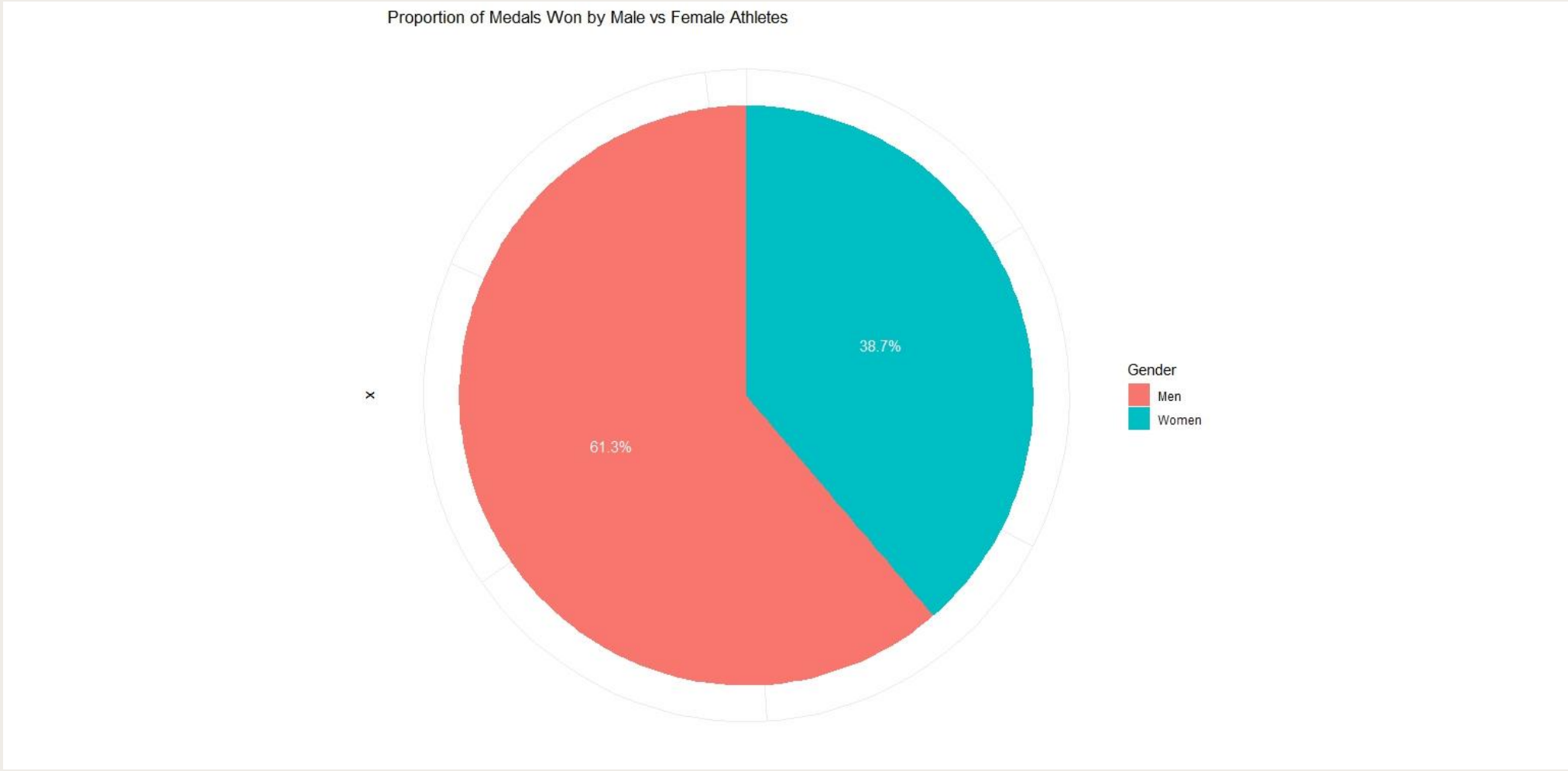
4)Most successful athletes



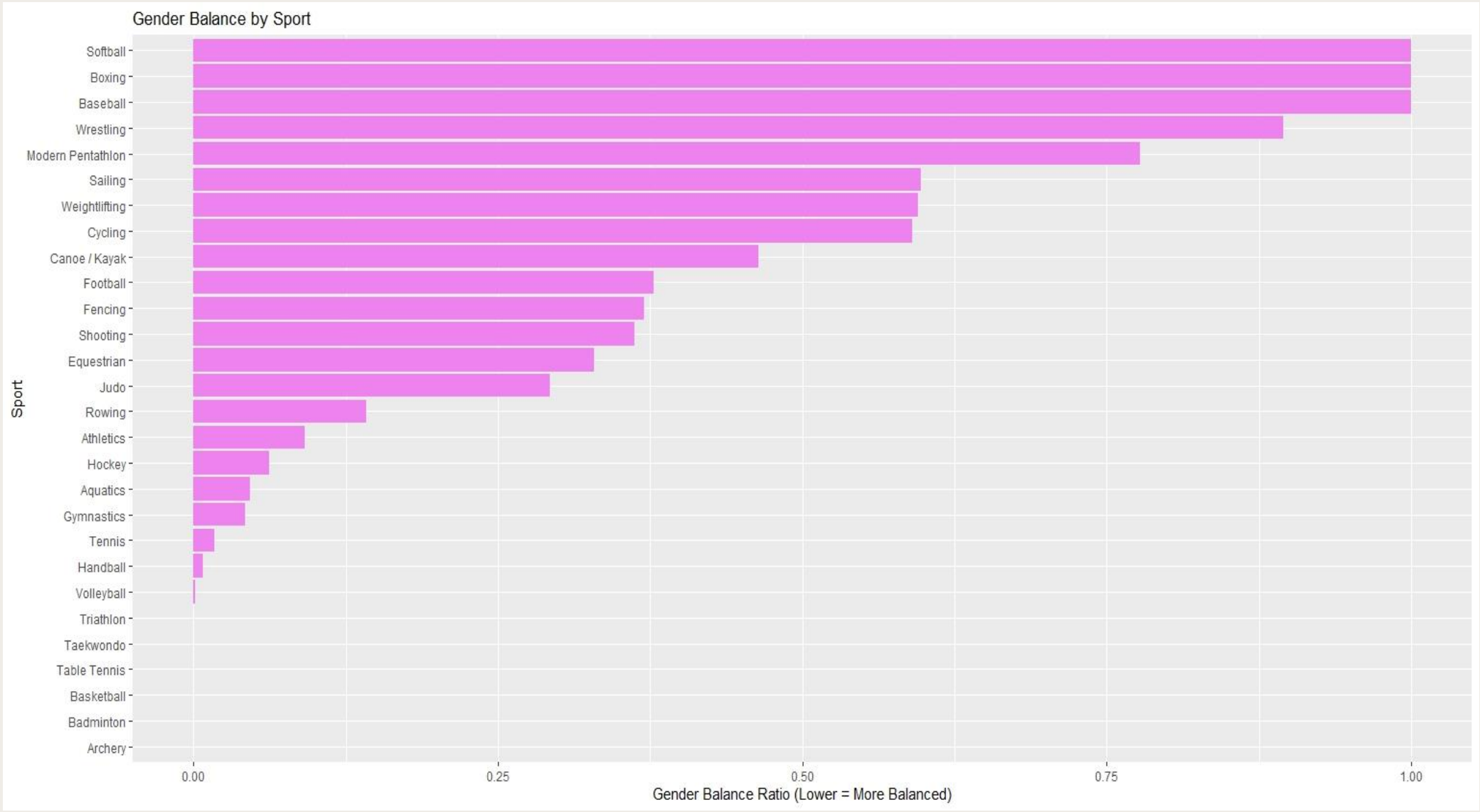
5)Top sports for each country



6)The number of male vs female athletes that won the model



7)Sports/Events that are more balanced in terms of gender



Insights

- The total number of medals received are **15316,** the **total number of countries participated** is **127** and the **total number of events** held are **289**.
- From the analysis, it is clear that the year **2008** recorded the most number of medals received , that is **2042** medals whereas the least number of medals received was **1305** during the year **1976.**
- **Aquatics** gained the most popularity over the years. It shows a **sharp upward trend** with a net growth of additional 188+ entries from 1976-2008. Similarly, **Football, Judo, Hockey** gained popularity as well. Sports such as **Rowing, Fencing, Tennis** showed a **decline** while **Badminton, Basketball, Boxing** showed **stagnant growth.**

- **United States** won the most number of medals compared to other countries, with a **total of 1992 medals** including **928 gold, 583 silver, 481 bronze** medals respectively and the **Soviet Union** being the **second** with a total of **1021 medals.**
- Among the most successful players, **Michael Phelps** secured a total of **16 medals** stating him as the top performer in the Summer Olympics.
- For the **US and Australia**, the top sport is **Aquatics**, securing a total of 578 and 239 medals respectively for the same meanwhile **Football** is the most top sport for **Brazil**. **Rowing** is the top sport for **Canada, East Germany, Romania and UK.**
- Out of the total number of medals, a proportion of **61.3% male** won the medal compared to **female** who only won **38.7%** representing that men dominated the medal winning.
- In terms of gender, the sport which showed **balanced gender distributions** are **Archery, Basketball, Table Tennis** and the sport which showed least balanced distributions are **Baseball, Boxing, Softball**. This can be interpreted from the graph that if the gender ratio (= number of male/number of female) of a particular sport is 0 it means it has the most balanced male-female participant distribution that is the **lower the gender ratio, the more balanced participation.**

Summary

- This project helped us to understand the various trends and highlights involved in the Summer Olympics Data, where I analyzed the number of medals won in each year, the popularity of sport over time, gender distribution among sport and the top performing athletes.
- Aquatics was the most dominated sport over the year.
- USA, Russia and China consistently reported to be the top performing countries.
- The number of medals received among male was significantly high compared to female.
- A few number of sport (for e.g., Rowing, Fencing) declined in popularity while sport like Football, judo saw significant increase in their participation.
- As mentioned earlier, Excel was used to clean and prepare the data, R was used majorly for analyzing the data and Power BI was used to implement the interactive dashboard of the overview of the data which will be attached along with PPT report for having a particular knowledge of the situation.

Thankyou