Declaration on Plagiarism

|  |  |
| --- | --- |
| **Name:** | Tejal Vijay Nijai |
| **Student Number:** | 19210412 |
| **Programme:** | MSc in Computing(Data Analytics) |
| **Module Code:** | CA682 |
| **Assignment Title:** | Data Visualisation |
| **Submission Date:** | 16 Dec 2019 |
| **Module Coordinator:** | Dr Suzanne Little |

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found at <http://www.dcu.ie/info/regulations/plagiarism.shtml>, <https://www4.dcu.ie/students/az/plagiarism> and/or recommended in the assignment guidelines

Name: Tejal Vijay Nijai Date: 16/12/2019

**120 year of Olympic Games Data Visualization**

The database contains information about 120 years of summer and winter games Olympic data with various events and descriptions of male and female athletes. I'm curious to see how male and female athlete participation has evolved over the years. I'm trying to find out if they have a similar turnout trend in contribution. Furthermore, what are the popular events or sports? I am also trying to find the inclination of males and females in various events held in popular athletics sport over the years.

Through my graphs, there is some trend followed in the initial years of Olympic(1890-1980) and after 1980 same kind of pattern is followed for both genders. Athletics is the summer sport in case of sports popularity and cross-country skiing is famous winter sport. The trend of contribution of male and female in athletics sport’s various events such as Athletics Men's Shot Put, Athletics Women's Shot Put, Athletics Men's High Jump held over the years.

**1.Dataset Description:**

The size of the dataset is 40MB. It has 271116 rows and 15 columns and each record corresponding to athlete detail competing in the Olympic event. This is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. The columns with the data types are as follows:

1. ID - Unique number for each athlete (Integer)
2. Name - Athlete's name (String)
3. Sex - M or F (Character)
4. Age – Integer (Integer)
5. Height - In centimetres (Integer)
6. Weight - In kilograms (Integer)
7. Team - Team name (String)
8. NOC - National Olympic Committee 3-letter code (String)
9. Games - Year and season (String)
10. Year – Integer
11. Season - Summer or Winter (String)
12. City - Host city (String)
13. Sport – Sport (String)
14. Event – Event (String)
15. Medal - Gold, Silver, Bronze, or NA (String)

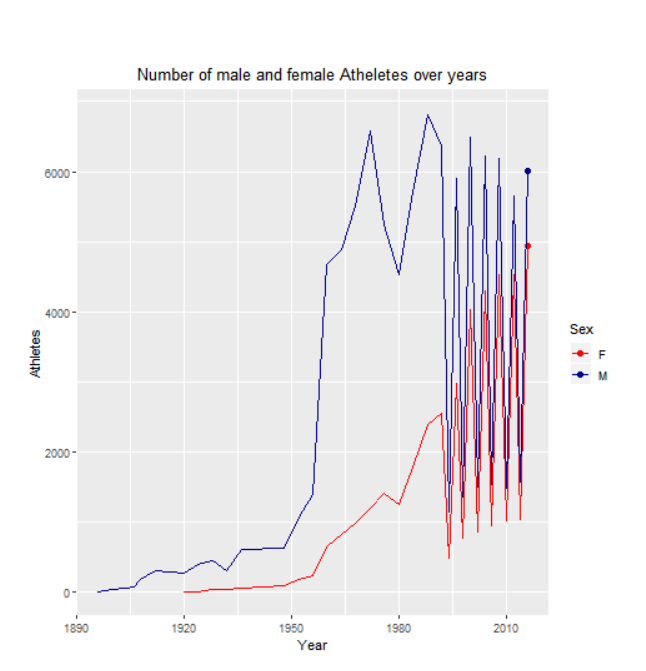
As it has 2lakhs+ rows with 15 different descriptive attributes, my dataset comes under the volume dimension.

**2. Data Exploration, Processing, Cleaning and/or Integration:**

Dataset had incomplete information about the athletes' age, height, and weight. I cleaned the dataset by using dropna function in python. I wanted to make a story around gender performance or participation in different sports or events in different regions if possible. Despite that, columns like age, height, weight, and medal have been omitted. Because the dataset is about 120 years of players, which made me think of showing a pattern in participation and also provides gender-specific data that we can compare over the years. Additionally, there was data about sports or games in which we could find out the favourite sport.

**3. Visualisations**

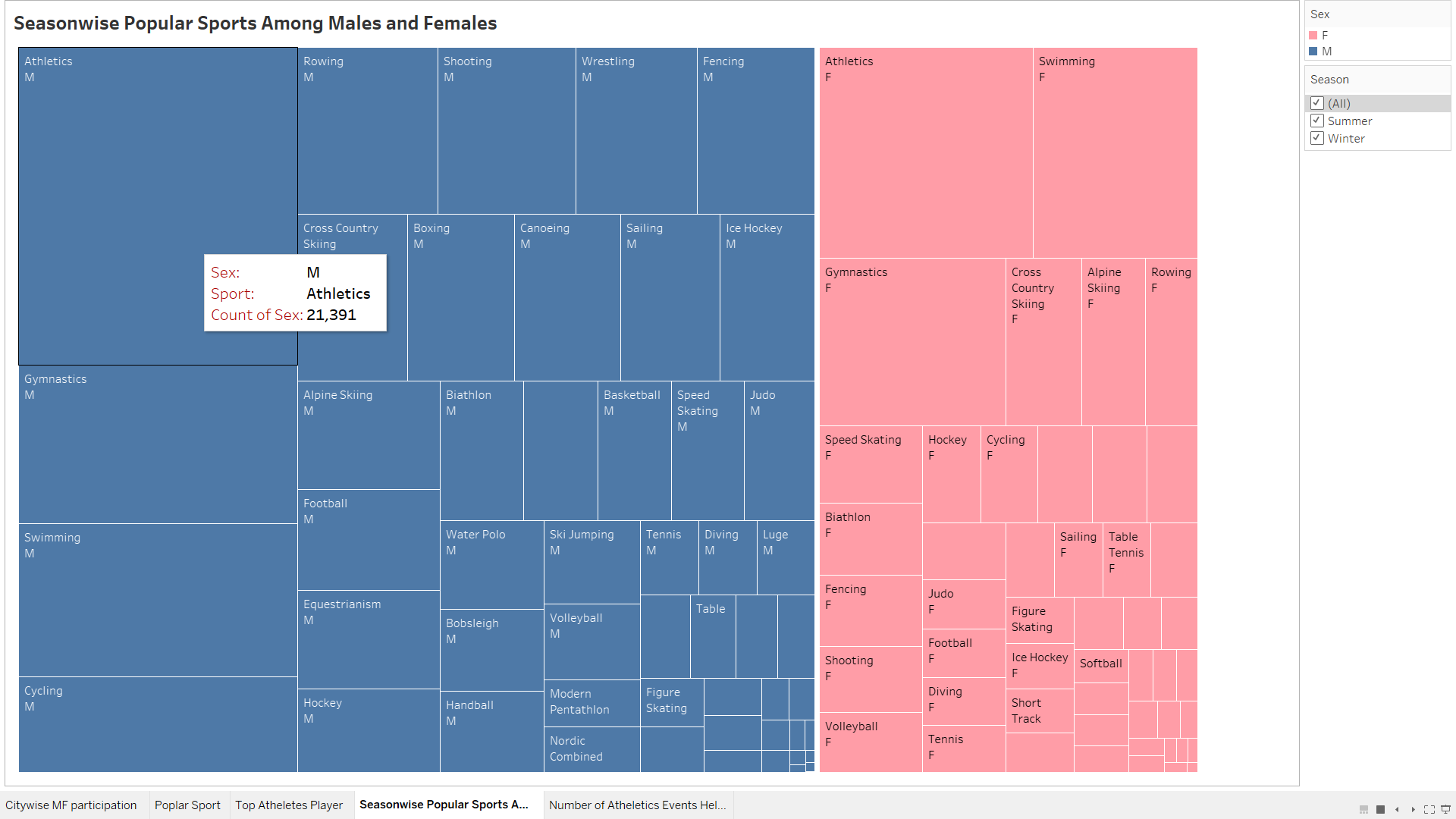
Graph 1:



The graph above shows the pattern of male and female participation in the Olympic Games for 120 years, i.e. 1896 to 2016. From this graph, we discovered that women started to contribute from 1920 Olympic Games. Male participation has grown rapidly from 1950-1980, although female participation is slowly gaining momentum. In 1980, there was a sharp decrease in both genders' participation. In 1980, the United States led a boycott of the Summer Olympic Games in Moscow to protest the late 1979 Soviet invasion of Afghanistan. In total, 65 nations refused to participate in the games, whereas 80 countries sent athletes to compete[2].Since 1980, competing in both sexes ' Olympic games is quite similar.

Line graphs are used for monitoring changes over short and long periods of time. These are also used to compare improvements over the same time period for more than one category. I constructed this chart in R and showed it in animation. I used red and blue colours respectively to represent the female and male line. We might work out some points in my diagrams quickly by animating the line graph. Once animation begins, we can discern immediately that for female graph started at 1920. Going further there is a sudden drop in 1980 and after 1980 the pattern is similar for both sexes. This helped examine and compare the pattern of the phenomenon.

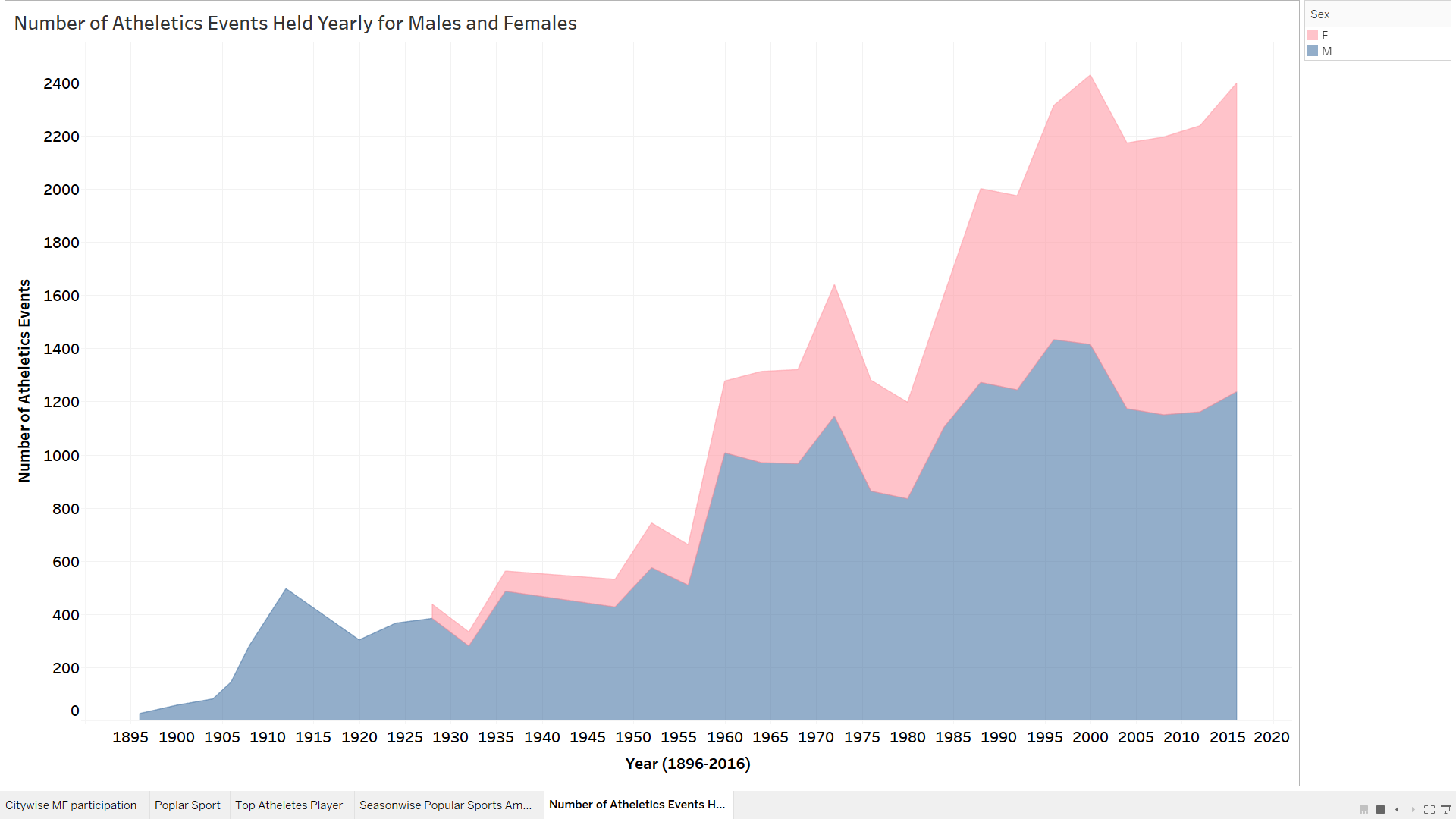
Graph 2:



The chart above shows common games / sports played by men and women of the season based on the number of people who played the game. Using this graph, we discovered that athletics is the summer season's popular game among males and females. Cross-country skiing is the most popular sport among males and females for the winter season. We get to know that, irrespective of season same sports is adored by both genders.

Treemap graphs are used to view hierarchical data and relationships in part. It is automatically ordered in descending order by the width of the rectangle which helps to classify the values from the highest to the lowest.

I created this graph and used common colours to represent blue and pink for male and female to realize that data visualization is about the distribution of sex. I have used tableau interactive treemap feature to filter data season wise.

Graph 3:By referring second graph, we got to know that athletics sport is famous among men and women. To drill down further how the athletics sport with its various events has men and women inclination. This graph not only tells us count of athletics events held for male and female individually but also the total count over gender. Graph has been created in tableau and generic colours for male and female have been used. Area chart is used to show trend of population over time.

We got to know that athletics is popular among men and women by referring to the second graph. To further drill down how men and women are drawn to the athletics sport with its various events, I created area chart. The chart shows us not only the number of individually held athletic events for male and female, but also the overall count by sex. Graph was developed in the table and male and female standard colors were used. Area map is used to display

population pattern over time.

**4. Conclusion:**

I tried to plot animation of line graph in python, but it was taking longer time. Whereas in R , I could plot the graph with a smaller number of lines of code with the help of R’s gganimate libraries. The data from treemap graph could have been shown in bubble chart also. The con of bubble chart is it shows data in different size of circles indicating the highest and lowest of data. To distinguish among these circles to get the largest one is difficult or time consuming when count of parameter is nearby as the size would seem same. In contradiction to this , data is displayed in descending order automatically.

**References:**

1. **Dataset retrieved from following Link:**

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

2. <https://2001-2009.state.gov/r/pa/ho/time/qfp/104481.htm>

3. <https://rpubs.com/elifdemir/olympics_analysis_report> (Referred this for the design of my Graph 1)