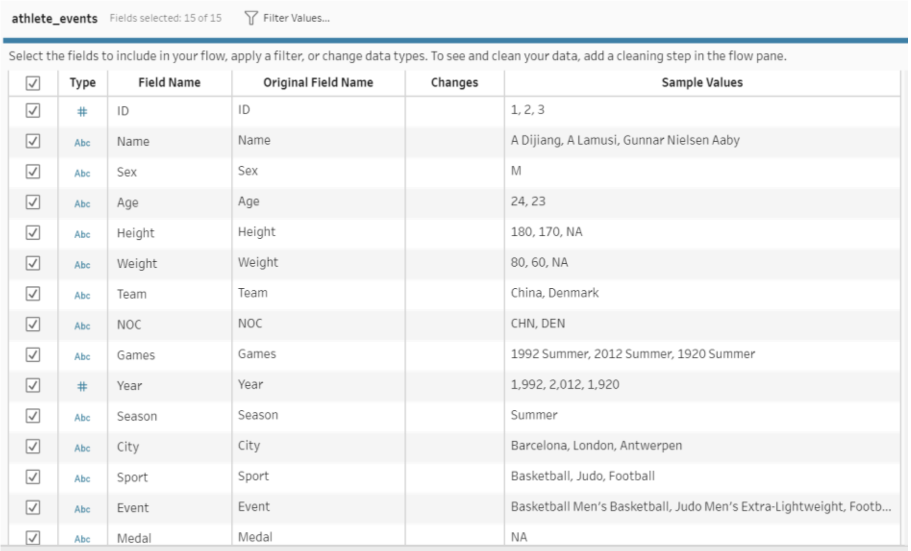
Name: Nivea Nobel Dabre
RUID: 196004350

I. INTRODUCTION

The first Olympics was held in Athens in 1896 after that 28 summer Olympics and 23 winter Olympics held. My topic is exploratory data analysis of 120 years of Olympics data dating from 1896 to 2016 collected from the Kaggle website. The main aim of the analysis is to find the hidden information or pattern in the data. Since the Olympics have evolved over the years, the question will include subjects like the participation and performance of athletes, different nations, and different sports and events. The exploratory analysis of the data is carried out using the Tableau Prep and Tableau Desktop software.

II. DATASET

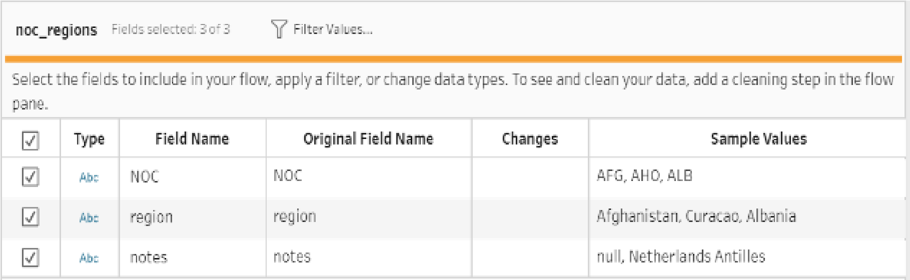
The data for the analysis is obtained from the Kaggle website. The data is spread across two files "athlete_events.csv" and "noc_regions.csv". The first file contains the information about the athletes and the events and the other contains the mapping of the country name to the National Olympic Committees (NOC) code. The athlete and event dataset contain 271116 rows and 15 columns which includes id, name, sex, age, height, weight, team, NOC, games, year, season, city, sport, event and medal. The NOC dataset contains 230 rows and 3 columns which includes NOC code, region and note.



athlete_events Fields selected: 15 of 15 Filter Values...

Select the fields to include in your flow, apply a filter, or change data types. To see and clean your data, add a cleaning step in the flow pane.

☑	Type	Field Name	Original Field Name	Changes	Sample Values
☑	#	ID	ID		1, 2, 3
☑	Abc	Name	Name		A Dijiang, A Lamusi, Gunnar Nielsen Aaby
☑	Abc	Sex	Sex		M
☑	Abc	Age	Age		24, 23
☑	Abc	Height	Height		180, 170, NA
☑	Abc	Weight	Weight		80, 60, NA
☑	Abc	Team	Team		China, Denmark
☑	Abc	NOC	NOC		CHN, DEN
☑	Abc	Games	Games		1992 Summer, 2012 Summer, 1920 Summer
☑	#	Year	Year		1,992, 2,012, 1,920
☑	Abc	Season	Season		Summer
☑	Abc	City	City		Barcelona, London, Antwerpen
☑	Abc	Sport	Sport		Basketball, Judo, Football
☑	Abc	Event	Event		Basketball Men's Basketball, Judo Men's Extra-Lightweight, Footb...
☑	Abc	Medal	Medal		NA



noc_regions Fields selected: 3 of 3 Filter Values...

Select the fields to include in your flow, apply a filter, or change data types. To see and clean your data, add a cleaning step in the flow pane.

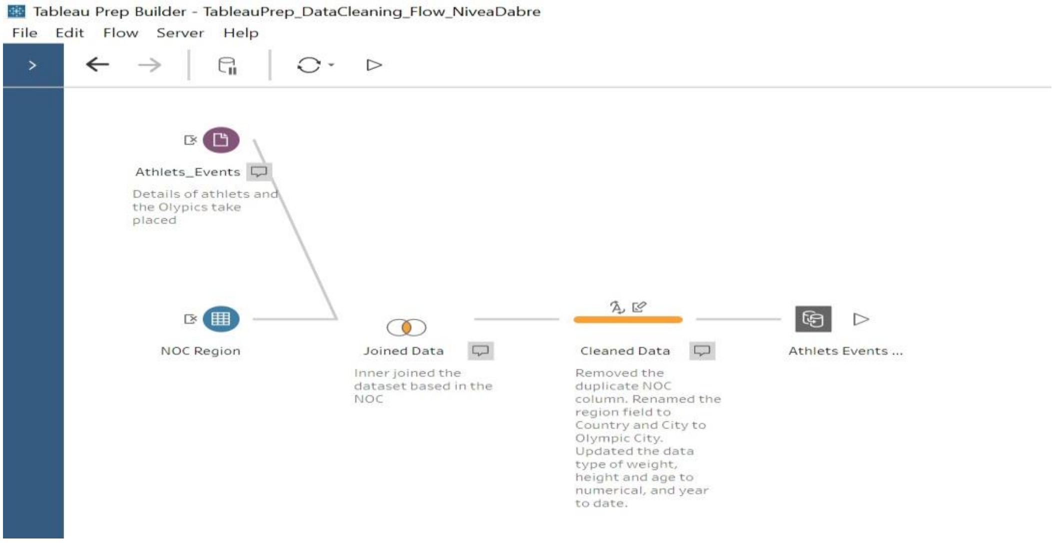
☑	Type	Field Name	Original Field Name	Changes	Sample Values
☑	Abc	NOC	NOC		AFG, AHO, ALB
☑	Abc	region	region		Afghanistan, Curacao, Albania
☑	Abc	notes	notes		null, Netherlands Antilles

athlete_events.csv and noc_regions.csv dataset structure in Tableau Prep

III. PROCESS

For the data cleaning and pre-processing, the dataset is imported in the Tableau Prep, where they are merged and removed the unnecessary columns. Initially, the id field from the athlete and events and note from NOC dataset are removed. Then using NOC from both the dataset are joined and removed the repeating NOC column. After that renamed the columns with meaningful words such as city to Olympic city, region to Country.

Then datatypes of age, height, weight are changed to number type, year to date type and the country and Olympic city to geographical type. In the end, the final dataset is exported as a hyper file, which will make the data extraction faster in the Tableau Desktop.



Details of athlets and the Olypics take placed

Inner joined the dataset based in the NOC

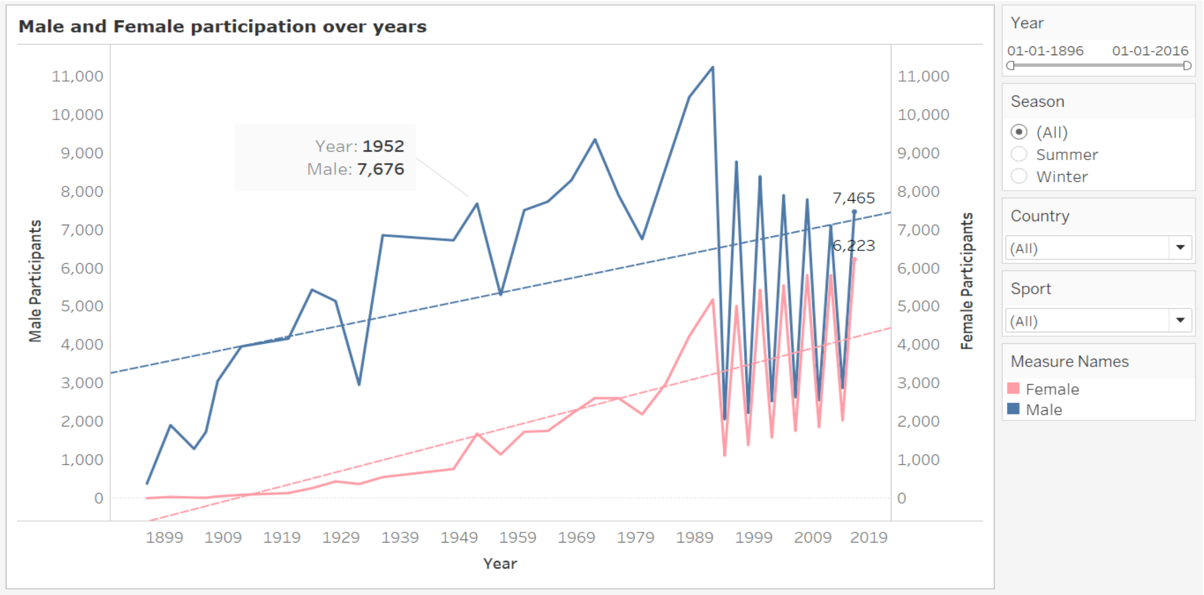
Removed the duplicate NOC column. Renamed the region field to Country and City to Olympic City. Updated the data type of weight, height and age to numerical, and year to date.

Data pre-processing in Tableau Prep

The hyper file is then imported in the Tableau Desktop for the visualisation. For the analysis, I have created multiple visualisations which include graphs like the line graph, bar graph, scatter plot, map.

IV. SOME VISUALIZATION AND ANALYSIS

The first graph compares the participation of the male and female athletes from 1896 to 2016, which is plotted in a line graph. Since it is a time series event the line graph is used. The male and female athlete's count is represented in the left Y-axis and right Y-axis, respectively. The male athletes are denoted in blue colour and female in pink colour as they used as the standard colour for gender distribution. The filters, year, season, country and sport are also attached to the graph. The trend line is added to show the trend in the graph. The tooltip text for the graph includes the year and the number of participants in that year.



Year: 1952
Male: 7,676

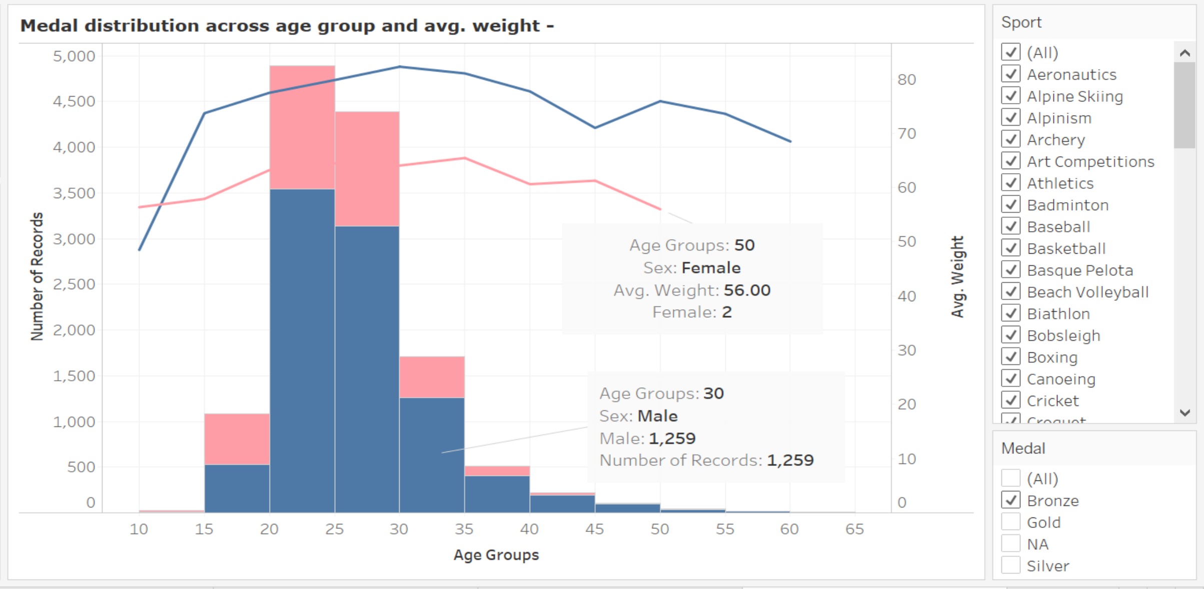
Year
01-01-1896 01-01-2016

Season
● (All)
○ Summer
○ Winter

Country
(All)

Sport
(All)

Male and Female athletes' participation over the years



Age Groups: 50
Sex: Female
Avg. Weight: 56.00
Female: 2

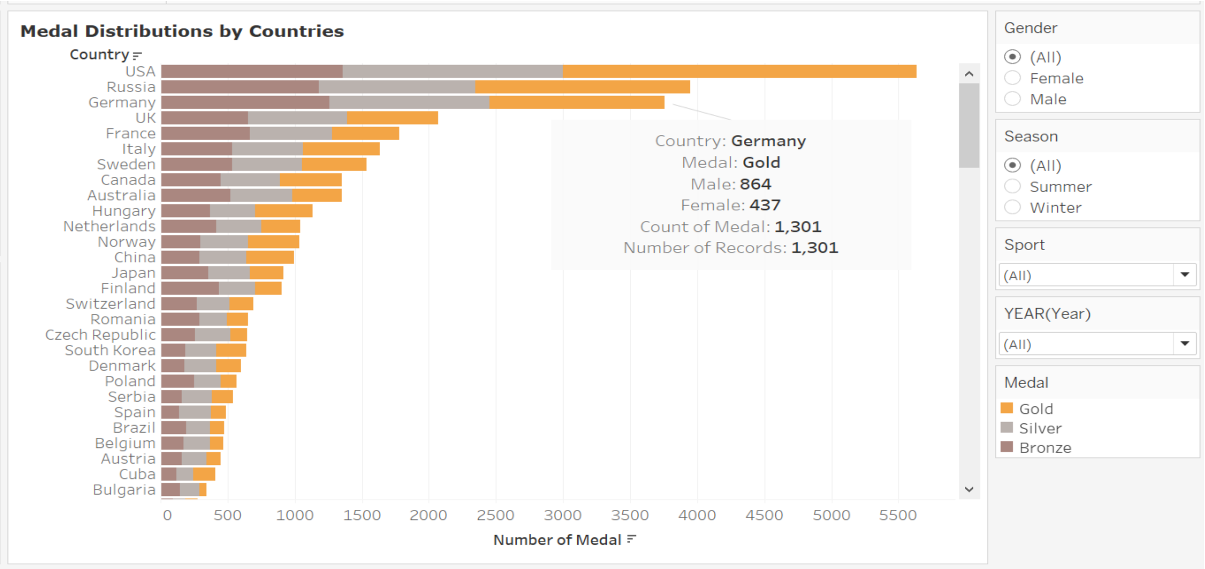
Age Groups: 30
Sex: Male
Male: 1,259
Number of Records: 1,259

Sport
☑ (All)
☑ Aeronautics
☑ Alpine Skiing
☑ Alpinism
☑ Archery
☑ Art Competitions
☑ Athletics
☑ Badminton
☑ Baseball
☑ Basketball
☑ Basque Pelota
☑ Beach Volleyball
☑ Biathlon
☑ Bobsleigh
☑ Boxing
☑ Canoeing
☑ Cricket
☑ Croquet

Medal
☐ (All)
☑ Bronze
☐ Gold
☐ NA
☐ Silver

Medal distribution across different age group and their average weight

The medal distribution by countries is represented in the second graph, which is plotted in a horizontally stacked bar graph because to compare the quantity among different group bar graph is best. The Y-axis contains the countries and the X-axis contains the number of medals. The graph is sorted in descending order of the total number of medals received by the country. The stack contains medal distribution across gold, silver and bronze in their respective colours. The filters, gender, season, sport and year is attached to the graph. The tooltip text contains the name of the country, medal category, the total number of the medal in that category, the number of male medal holders and the number of female medal holders.



Country: Germany
Medal: Gold
Male: 864
Female: 437
Count of Medal: 1,301
Number of Records: 1,301

Gender
● (All)
○ Female
○ Male

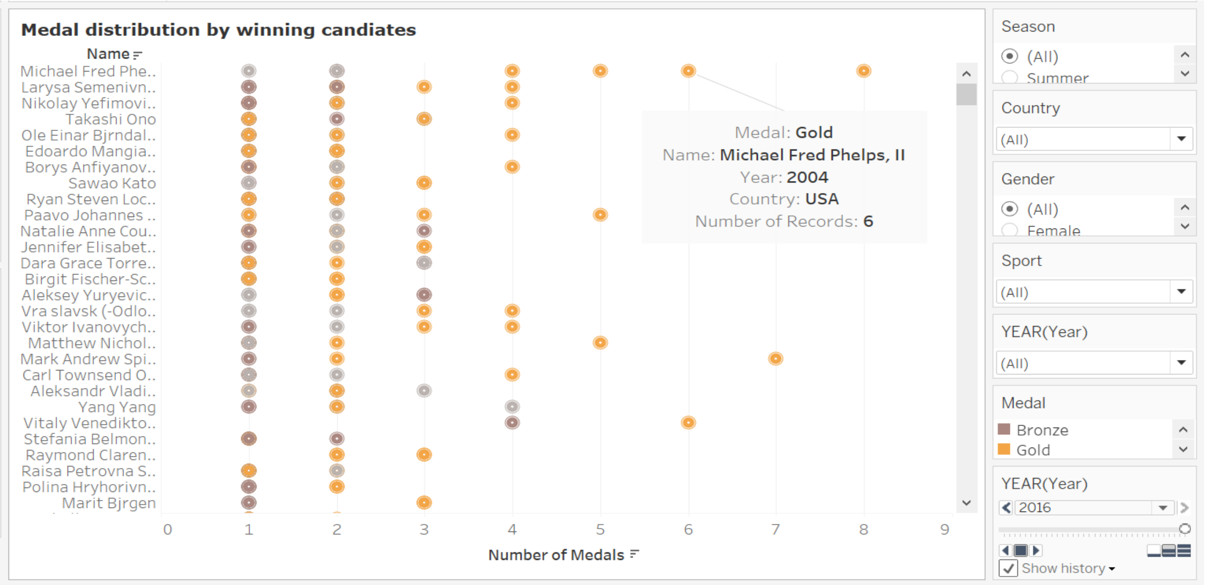
Season
● (All)
○ Summer
○ Winter

Sport
(All)

YEAR(Year)
(All)

Medal Distributions by Countries

The medal distribution by athletes is described in the third graph, which is also plotted in a horizontally stacked graph since it is also a comparison. The shape of each category is in the form of a medal and the colours are applied accordingly. The Y-axis contains the athletes and the X-axis contains the number of medals. The graph is sorted in descending order of the total number of medals received by the athletes. The season, country, gender, sport and year is attached to the graph. The tooltip text contains the name of the athlete, medal category, the country they are representing, the total number of the medal in that category. Paging is used to indicate the category of medals and number of medals achieved by the participants each year.



Medal: Gold
Name: Michael Fred Phelps, II
Year: 2004
Country: USA
Number of Records: 6

Season
● (All)
○ Summer

Country
(All)

Gender
● (All)
○ Female

Sport
(All)

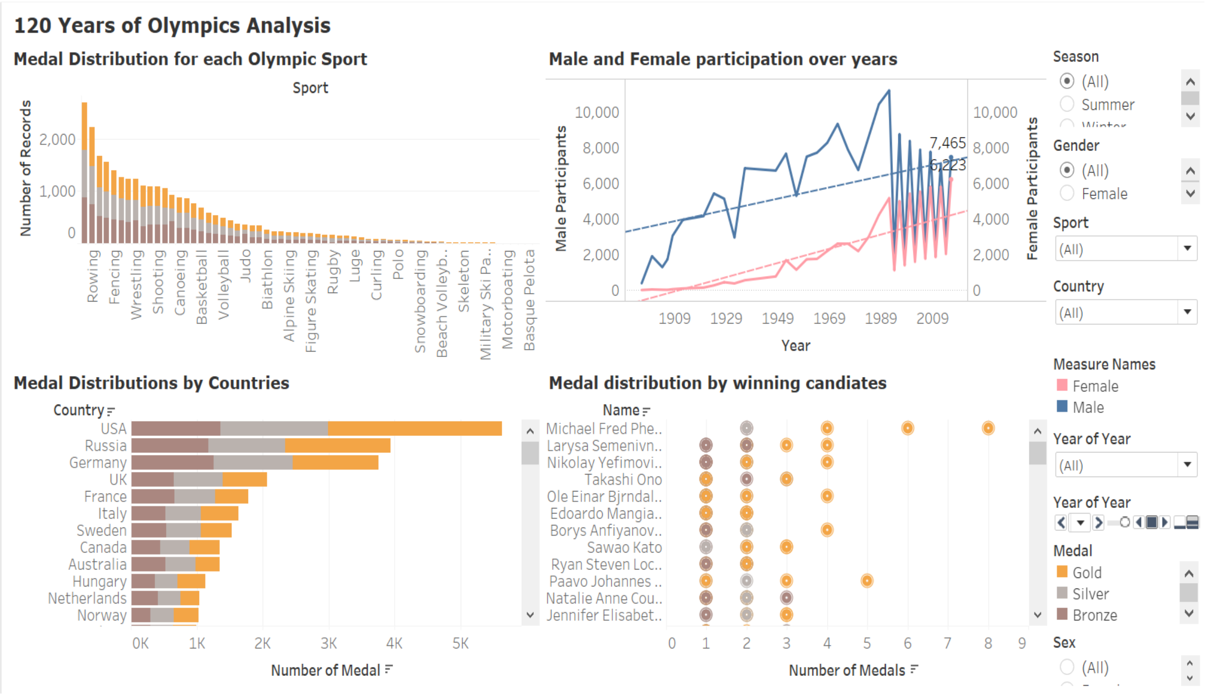
YEAR(Year)
(All)

YEAR(Year)
2016

Medal distribution by winning candidates

V. RESULT

The visualisation has used attributes such as shape, size, orientation, colour, the position, and Gestalt principles such as proximity, similarity, continuity throughout the process. For the final analysis, all the graphs were connected. The continuity principle was used to show the trend in the participation of the candidates and the colour attribute was used to distinguish male and female candidates. The proximity principle was used to align the medal together and the colour attribute to distinguish medals. The similarity principle was used to show the similarity between medals and the colour attribute to distinguish medals.



120 Years of Olympics Analysis

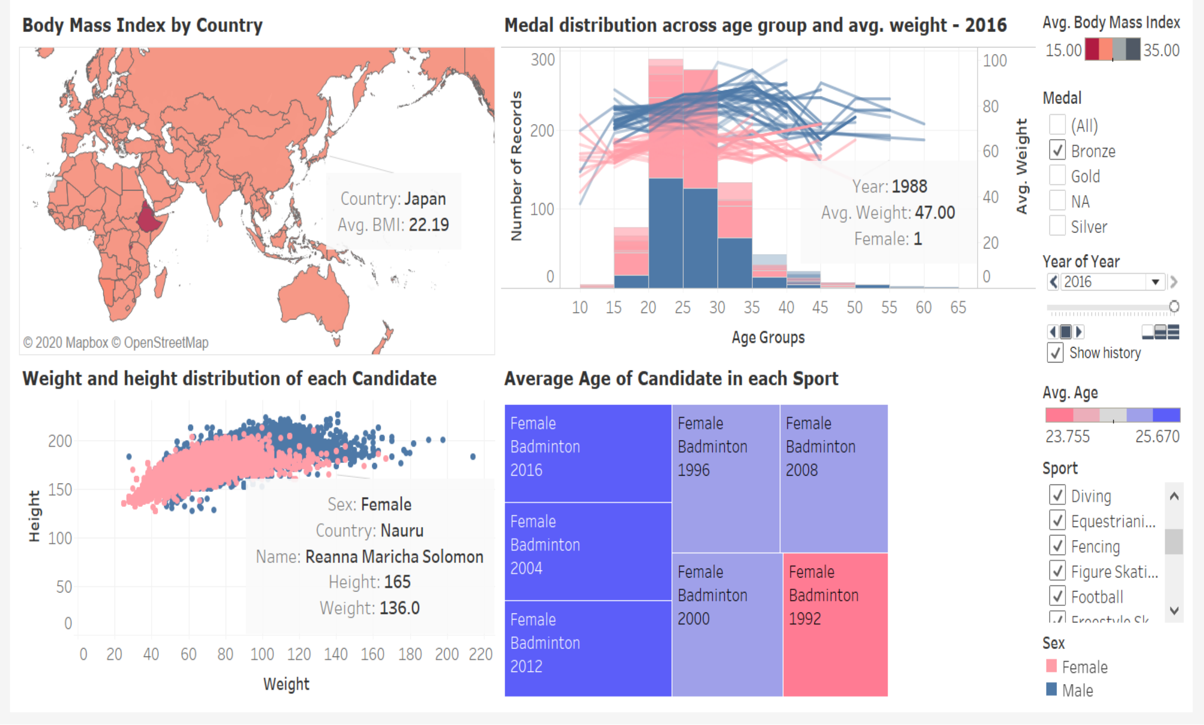
Season: ● (All) ○ Summer ○ Winter
Gender: ● (All) ○ Female
Sport: (All)
Country: (All)
Year of Year: (All)
Sex: ○ (All)

Interaction between multiple graphs in Tableau Desktop

Some of the findings from the visualisation and analysis of the Olympic History is as follows:

- There was no Summer Olympics in 1916, 1940 & 1944 and Winter Olympics in 1940 & 1944, later, the research found that it is because of World War.
- Only the first edition of the Olympics did not have any female athletes. By considering the percentage increase in female participation from the last 30 years of summer Olympics, the female athletes will exceed the male athletes in 2020 Olympics.
- Michael Fred Phelps, II from USA has the greatest number of medals. He holds 23 Gold, 3 Silver and 2 Bronze medals.
- Larisa Semyonovna Latynina from Russia has the greatest number of medals between female athletes. She holds 9 Gold, 5 Silver and 4 Bronze medals.
- The USA has the most medal in Summer Olympics, but Russia holds that position in Winter Olympics.

- Athletics events have the greatest number of medals followed by swimming events considering both male and female athletes together. In the case of male athletes' Athletics events have the most medal followed by Rowing events and in the case of female athletes swimming events have the most medal followed by athletics events.



Country: Japan
Avg. BMI: 22.19

Year: 1988
Avg. Weight: 47.00
Female: 1

Sex: Female
Country: Nauru
Name: Reanna Maricha Solomon
Height: 165
Weight: 136.0

Medal
☐ (All)
☑ Bronze
☐ Gold
☐ NA
☐ Silver

Year of Year
2016

☑ Show history

Sport
☑ Diving
☑ Equestriani...
☑ Fencing
☑ Figure Skati...
☑ Football
☑ Freestyle Sk...

Sex
🟥 Female
🟦 Male

Interaction between multiple graphs in Tableau Desktop

- For the age group, 20-25 have the greatest number of medals.
- American Samoa has the highest average body mass index of 28.18 and Ethiopia have the least of 19.59.
- Average age of each of the Olympic participants fall in between 22.435 to 26.455 years

VI. REFERENCES

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