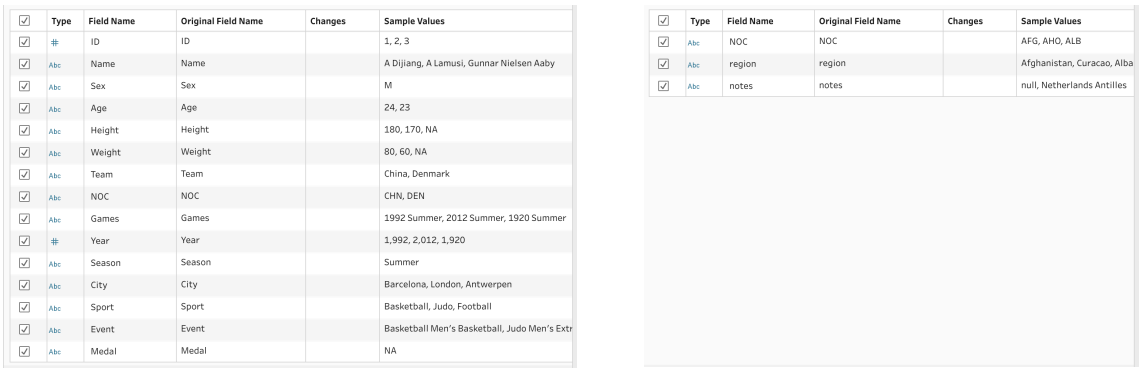
I. INTRODUCTION

The first Olympics was held in Athens in 1896 after that 28 summer Olympics and 23 winter Olympics held. In this report, I will explain an exploratory 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.

For the analysis, the tools like Tableau Desktop and Tableau Prep is used. The final presentation is shared in the Tableau public website which is accessible in any JavaScript enabled web browser.

II. DATASET

The data for the analysis is obtained from the Kaggle website [1]. 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.



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 Extr
Abc	Medal	Medal		NA

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

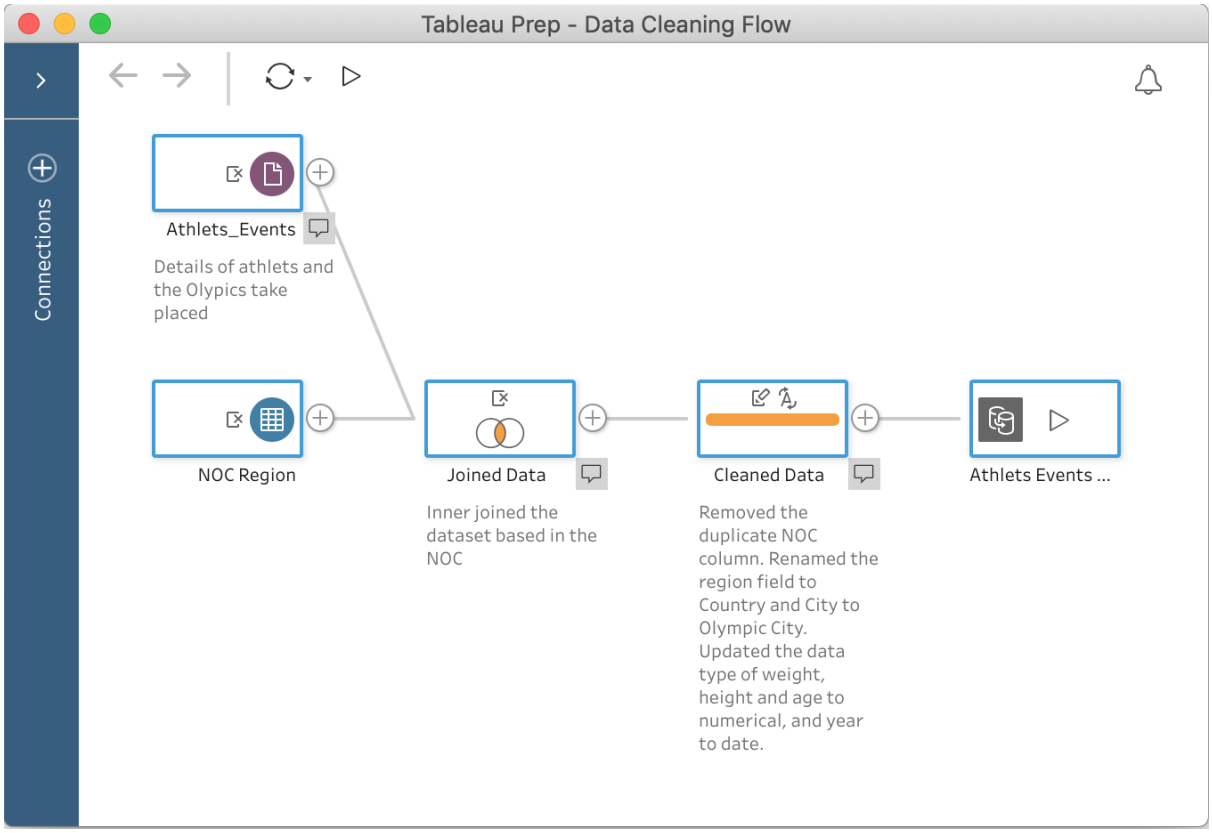
athlete_events.csv and noc_regions.csv dataset structure in Tableau Prep

III. PROCESS

The exploratory analysis of the data is carried out using the Tableau Prep and Tableau Desktop software and for the presentation, the Tableau public website is used. The trial version of both software is available to download in the Tableau website [2]. They also provide a full version of both software to students free of cost providing all credential.

For the data cleaning and pre-processing both the dataset are 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

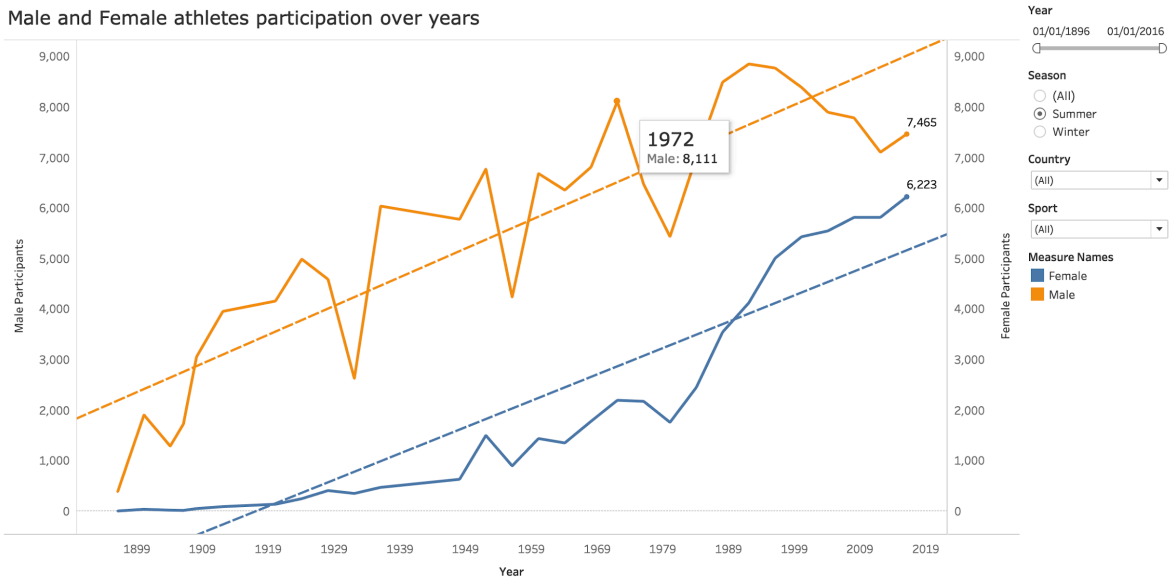
and Olympic city to geographical type. In the end, the final dataset is exported as a hyper file, which will make the data extraction faster in the Tableau Desktop.



Data pre-processing in Tableau Prep

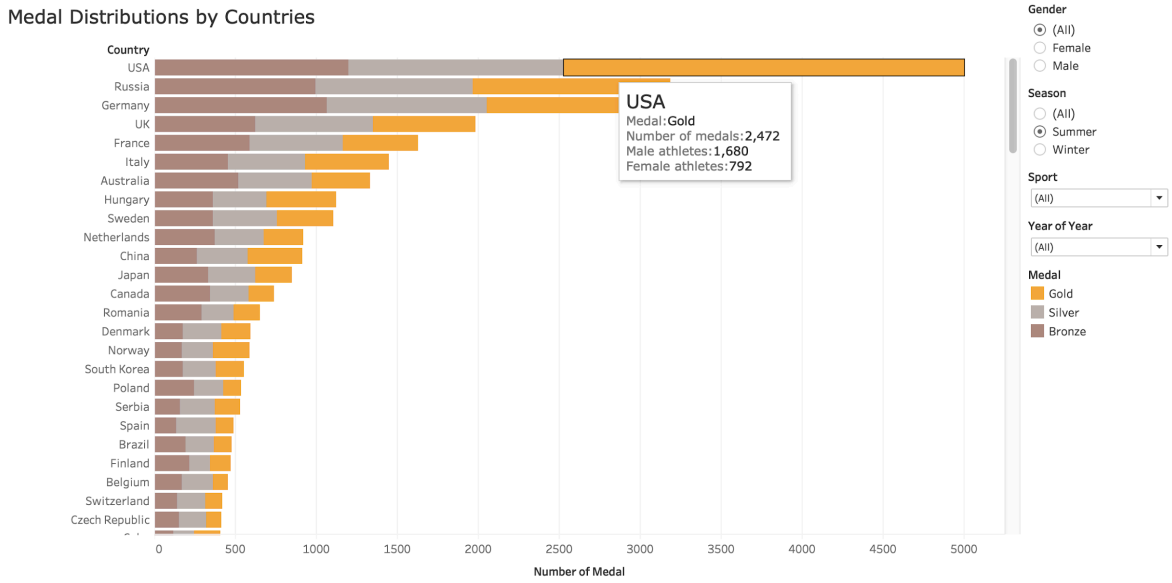
The hyper file is then imported in the Tableau Desktop for the visualisation. The final visualisation of the data is implemented in a story format in Tableau Desktop. For the analysis, I have created multiple visualisations which include graphs like the line graph, bar graph, scatter plot, map but for the final visualisation contains only 3 graphs rest of the graphs will be mentioned in the video report.

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 orange colour and female in blue 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.



Male and Female athletes participation over years in Tableau Desktop

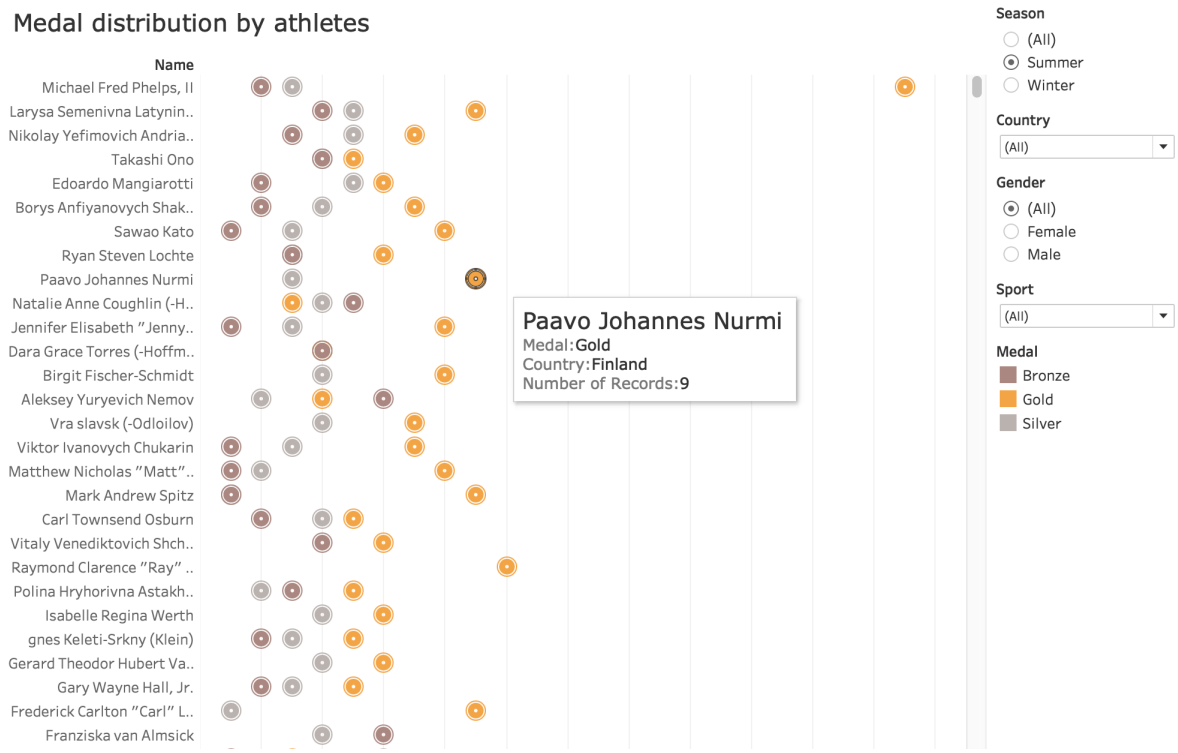
The medal distribution by countries is represented in the second graph, which is plotted in a horizontally stacked bar graph because to compare the quantity among different group bar graph is best. The Y-axis contains the countries and the X-axis contains the number of medals. The graph is sorted in 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.



Medal Distributions by Countries in Tableau Desktop

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.



Medal distribution by athletes in Tableau Desktop

In the video report the medal distribution across different age groups and their average weight, athlete weight and age distribution, body mass index by country as a Key Performance Index(KPI), medal distribution across different sport are also shown.

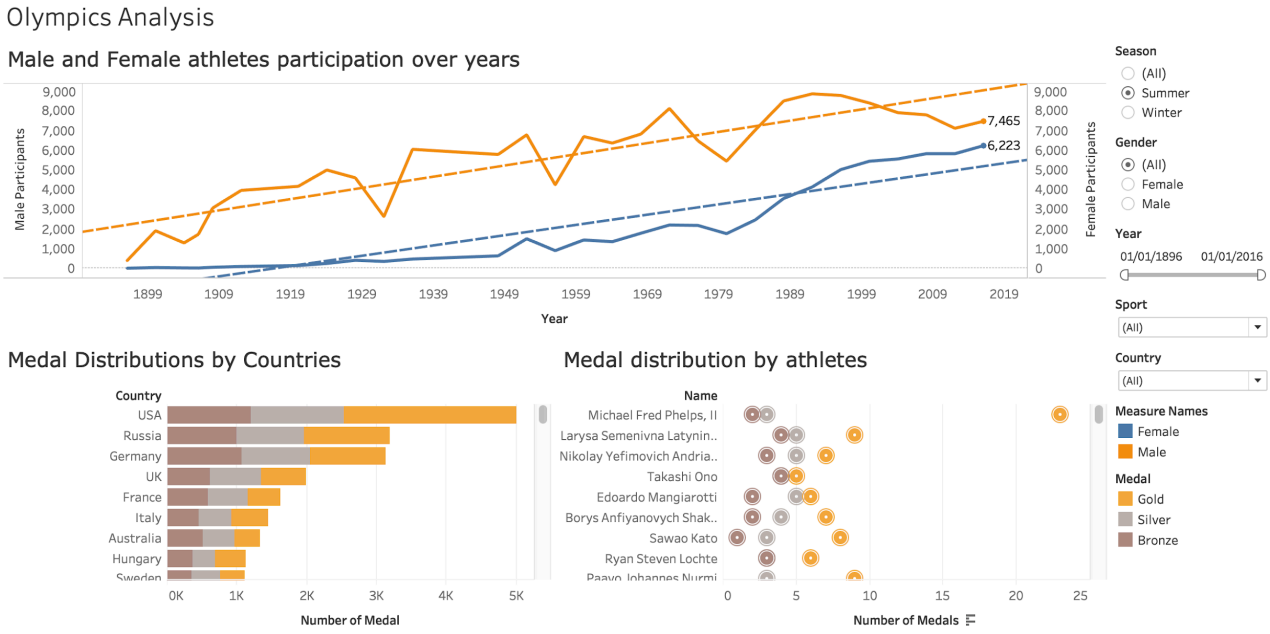
IV. RESULT

The visualisation has used preattentive 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 three graphs were connected together. In the first graph, the continuity principle was used to show the trend in the participation of the athletes and the colour attribute was used to distinguish male and female athletes. In the second graph, the proximity principle was used to align the medal together and the colour attribute to distinguish medals. In the last graph, the similarity principle was used to show the similarity between medals and the colour attribute to distinguish medals.

The complete story of the analysis is available on the Tableau Public website [3]. Some of the information noticed from the analysis are

- There was no Summer Olympics in 1916, 1940 & 1944 and Winter Olympics in 1940 & 1944, later on, the research found that it is because of World War.
- Only the first edition of the Olympics didn't 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.
- The USA has the most medal in Summer Olympics, but Russia holds that position in Winter Olympics.

- For the age group, 20-25 have the most number of medals.
- American Samoa has the highest average body mass index of 28.18 and Ethiopia have the least of 19.59.
- Athletics events have the most 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.



Interaction between multiple graphs in Tableau Desktop

The analysis could come up with much more meaningful information if more data about the countries were available, like population, GDP, per capita income for the time period of 1896 to 2016. If I was analysing for the last 50 years then it could happen but my aim was to analysis the 120 years of history of Olympics.

V. REFERENCES

[1] Tableau. [Online]. Available: https://www.tableau.com/. [Accessed 16 December 2018].

[2] J. M. Kalarickal, "Tableau Public," 16 December 2018. [Online]. Available: https://public.tableau.com/profile/jagajith.monappan.kalarickal#!/vizhome/Olympics_210/Story. [Accessed 16 December 2018].

[3] R. H. Griffin, "Kaggle," 15 June 2018. [Online]. Available: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results. [Accessed 16 December 2018].