DATA VISUALIZATION

Submitted to: OISIN CREANER

National College of Ireland

Submitted by: Divyang Jain
MSc. Data Analytics
Student ID: 16110323

Rio de Janeiro Olympic Games Analysis

Abstract: In today's century sports has become the necessity for a human being, many talented players are participating in distinct activities and Olympic game is the place which serves them as a platform. They are the foremost global sporting event, which fascinates huge worldwide interest from people of various social positions, races and ages throughout the world and participate in different kinds of sports like football, rugby, archery, handball etc. The major purpose of the Olympics is to unite the different countries and playing on the same arena. It was started in 1896 and took place in Greece, from that era till now it takes place every 4 years and in the latest period it happened in 2016, Rio de Janeiro (Brazil). In this project, a pristine effort has been put in to analyse the Rio Olympics data on athletes, their home Countries, medals, Facebook likes and many attributes which will convert it into information by visualizing it. This information can be useful for betting to predict the next Olympics Country winner and many other aspects.

Introduction:

Data visualization is very significant in the development of modern data analytics. "Visualization as the communication of information using graphical representations" (Ward et.al 2010, p.1). It explores and discovers the relationship between the hidden attributes because a single picture can be processed quickly then large a number of words and numeric values. It is not just about the graphics it is about statistics, data mining technique, perception and psychology. Nowadays, it is used by every big organization in decision making by observing the trends. In other words, the major role of visualization is to discover the knowledge by spotting the patterns, highlighting the outliers and uses human perception like Galaxyzoo (Creaner, 2017). Data is increasing generated by people through institution, social sites and many organization data which cannot be depicted from words. Hence, data visualization takes places to represent it into graphical formats. Many processes of visualization have been done like-data selection, pre-processing, mapping and image rendering.

This project gives the analysis on Rio Olympics happened in 2016 throughout the world with the help of data visualization techniques. As importance of sports is increasing day by day throughout the world which is helping and making the world united by the event of Olympics so it is very important part of the study to know from which country more athletes are coming, which country is dominates the game then other countries can work on the performance of their athletes, many such information can be grabbed from visualizing the Olympics data. According to Pop, 2013 many research has been done on Olympics from 1896-2010 data to enhance the existing sports database. The Olympics are the biggest sport events around the world but their value is still less in many countries as they are not participating and giving coaching appropriately. By studying and visualizing country wise and sports wise data, everyone can get to know the importance of sports popularity throughout the world from which many more countries can do hard work to get the medal and also to predict the winning from the patterns which will help in betting field.

The aim of this study is to analyse the Olympic games for a national committee in making correct decisions on the patterns discovered. There are many queries which are built to analyse it like Which Country dominates the Olympic game in 2016 happened in Rio de Janeiro. Also, to analyse which sport is mostly played and popular in Olympics. There are GDP and population given in the dataset which can help us to analyse the countries with the maximum population and GDP with the relation of winning gold medals. Description of players with their average height, weight, date of birth participated in different sports. Nationality of the winners and discipline of sports are also being analysed to give the proper information of Rio Olympics.

Dataset Background:

Rio Olympics 2016 data for this project is provided by Olympic Data feed -International Olympics committee and it is available on Kaggle also which is published under CC0: Public Domain License. According to my research the visualization I have represented is not done previously so I am taking the opportunity to represent this visualization. The data is for Olympics games happened in Rio for the year 2016 only and joined by tableau. Except this dataset, I have fetched the **Facebook likes** of each sport from **R studio** and merged it with Rio 2016 data with the help of inner join which is provided by **tableau**. Facebook likes is fetched by searching the different sport pages and getting the likes hit by public.

Data visualization Process

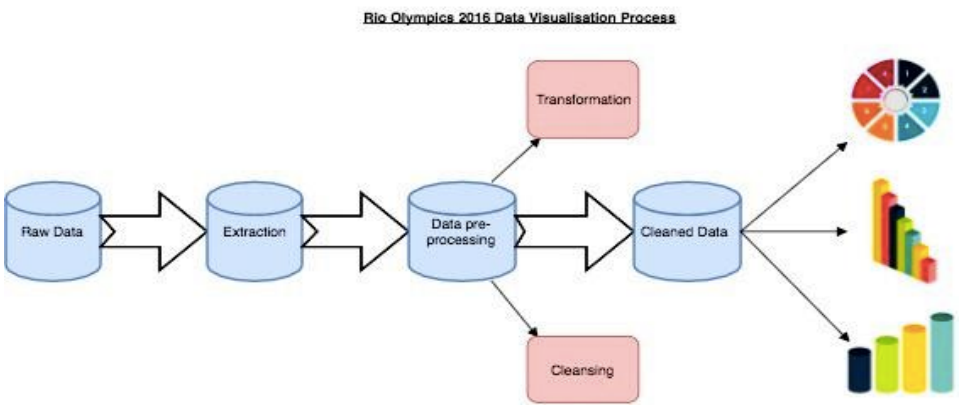


Fig 1. Data Visualization Process (made from draw.io)

Original Raw Data Set

Data consists of the official statistic on the 306 events and 11,538 athletes in Rio de Janeiro. These files are updated by the international Olympic committee every 4 years to get the latest records. All relevant information is provided related to athletes, countries and GDP popularity which took place in Rio 2016 Olympics.

There are 3 CSV files which is fetched from international Olympic committee and one CSV for

Facebook likes:

1.Athletes, 2. Countries ,3.Event, 4.Facebook Likes

FEATURES IN THE DATASET

1. Athlete ID- Unique Identification of Athlete
2. Athlete Name- Player Name
3. Athlete Nationality- To which nation athlete belongs
4. Athlete Sex- Athlete gender –male or female
5. Athlete D. O. B- Age of Athlete
6. Athlete Height- Height of athlete
7. Athlete Weight- Weight of athlete
8. Sport- Olympics sport like football, weightlifting
9. Gold Medal- Quantity of gold medals
10. Silver Medal- Quantity of silver medals
11. Bronze Medal- Quantity of Bronze medals
12. Country- Countries participated in Olympics
13. Code- Short Code of Countries
14. Population- Population in a particular Country
15. GDP- Gross Domestic Product per capita which can be useful to detect the amount its validating
16. Event ID- Event Identification Number
17. Sport- Olympics sport like football, weightlifting relating to event
18. Discipline- Discipline of sport, rules or diversions like diving, freestyle, butterfly
19. Event Name- Name of the event like Men's Team Pursuit, Women's Kayak Single 500m
20. Sex- Gender of the player according to the event game.
21. Venue- Location of the Olympics stadiums
22. Sport- Olympics sport like football, weightlifting related to Facebook likes.
23. **Facebook likes**- Public choice for a particular sport

Data Extraction with the help of R Studio

There are 3 CSV files which are gathered from International Olympics committee but to know the popularity of a particular sport, I have run the code in R studio to fetch the likes on Facebook by each sport and write that file into CSV and later on, it was linked with the files in tableau after cleaning. Here is the code- https://gist.github.com/divyang7666/e029997c0a16e8cb4f37b8a038a14f5b.

Data Pre-processing

1. Cleaning

2. Transformation

In most circumstances, the original raw data in the real world is really dirty, it is preferable to clean the data. As per the data scientist research more than half of the time is required to collect the data and pre-process it. Data scientist also need to filter or smoothen the data then see the missing values, outliers which may result in error in computation or input and also necessary to remove the duplicate values (Keller, Keller, 2013). Many steps can be performed to match the attributes to be included in tableau or any visualization tool sometimes normalization, segmentation, dimension reduction is

also required before starting the process (Keim, 2002). Hence, it is essential to prepare quality data to get accurate results to visualize. For this project, extracted data holds missing values, duplicate values and unknown values in GDP and some other attributes which were removed with the help of **Rapid Miner** software. Below is the diagram for the same:

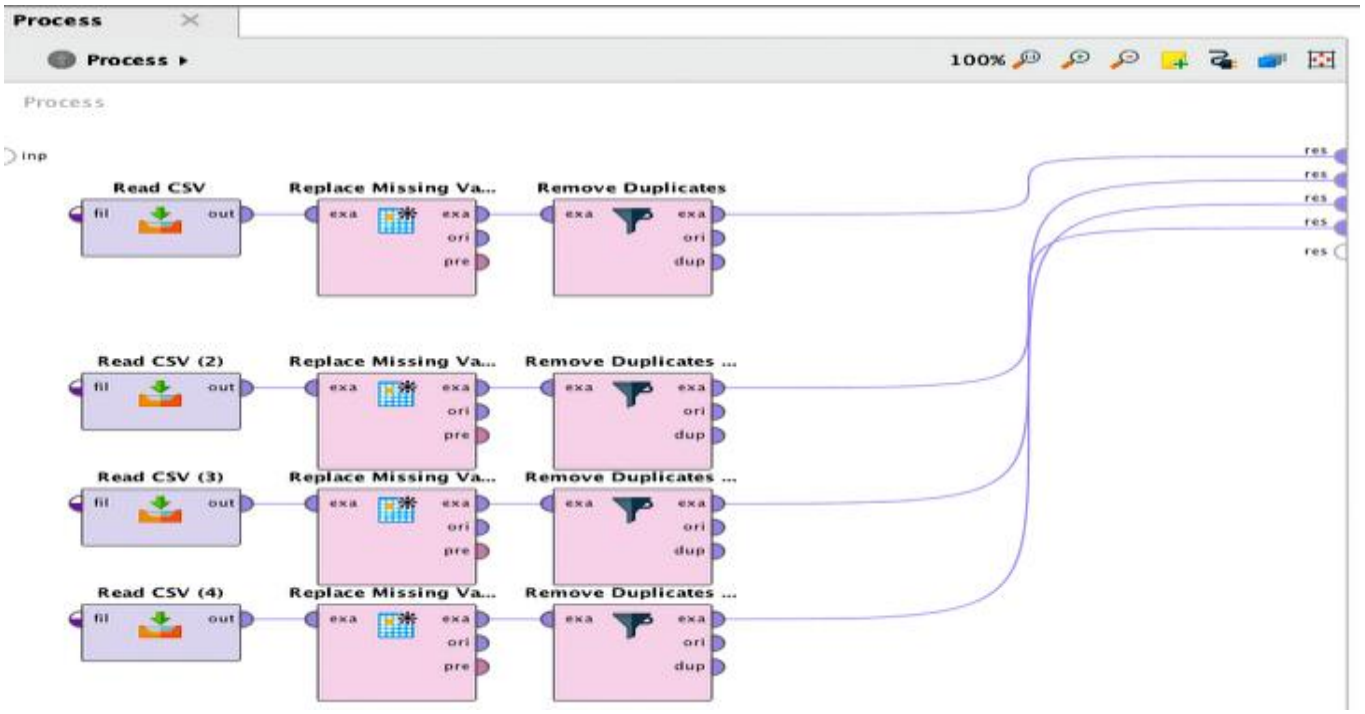
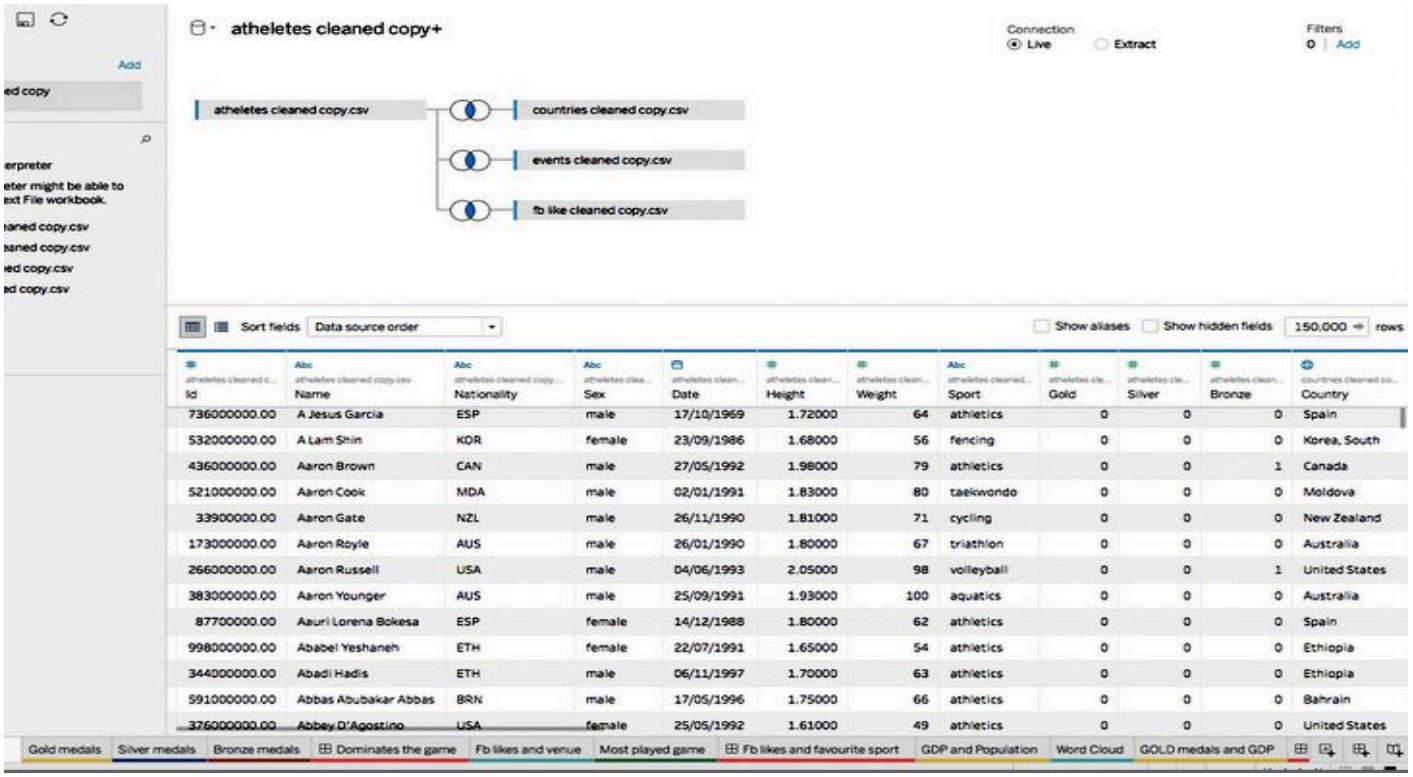


Fig 4. Data Pre- Processing flow

Cleaned Data

Cleaning is done with the help of replacing the missing values and removing the duplicates values with the help of Rapid Miner tool which is made in JAVA and result is the cleaned output of four CSV files. Now files are ready to put into **tableau** which is a data visualization tool. Tableau is an amazing tool for understanding, seeing and making important decisions on your cleaned dataset. It allows us to accomplish data connection, data exploration, data visualization, data analysis and data storytelling with many features of making dashboard and filtering with animation (Milligan ,2016). Import files into tableau and link them with the help of inbuilt **SQL** joins as it has **VQL** which makes the files join.



Id	Name	Nationality	Sex	Date	Height	Weight	Sport	Gold	Silver	Bronze	Country
736000000.00	A Jesus Garcia	ESP	male	17/10/1969	1.72000	64	athletics	0	0	0	Spain
532000000.00	A Lam Shin	KOR	female	23/09/1986	1.68000	56	fencing	0	0	0	Korea, South
436000000.00	Aaron Brown	CAN	male	27/05/1992	1.98000	79	athletics	0	0	1	Canada
521000000.00	Aaron Cook	MDA	male	02/01/1991	1.83000	80	taekwondo	0	0	0	Moldova
33900000.00	Aaron Gate	NZL	male	26/11/1990	1.81000	71	cycling	0	0	0	New Zealand
173000000.00	Aaron Royle	AUS	male	26/01/1990	1.80000	67	triathlon	0	0	0	Australia
266000000.00	Aaron Russell	USA	male	04/06/1993	2.05000	98	volleyball	0	0	1	United States
383000000.00	Aaron Younger	AUS	male	25/09/1991	1.93000	100	aquatics	0	0	0	Australia
87700000.00	Aauri Lorena Bokesa	ESP	female	14/12/1988	1.80000	62	athletics	0	0	0	Spain
998000000.00	Ababel Yeshaneh	ETH	female	22/07/1991	1.65000	54	athletics	0	0	0	Ethiopia
344000000.00	Abadi Hadis	ETH	male	06/11/1997	1.70000	63	athletics	0	0	0	Ethiopia
591000000.00	Abbas Abubakar Abbas	BRN	male	17/05/1996	1.75000	66	athletics	0	0	0	Bahrain
376000000.00	Abbey D'Agostino	USA	female	25/05/1992	1.61000	49	athletics	0	0	0	United States

Business Queries/Case study

After the files are imported and linked, it time to make visualizations.

Below are some case studies on which analyses has been done

1. Which country dominates the Olympic game in Rio de Janeiro 2016?

1.1 Countries won the maximum gold medals.

1.2 Countries won the maximum silver medals

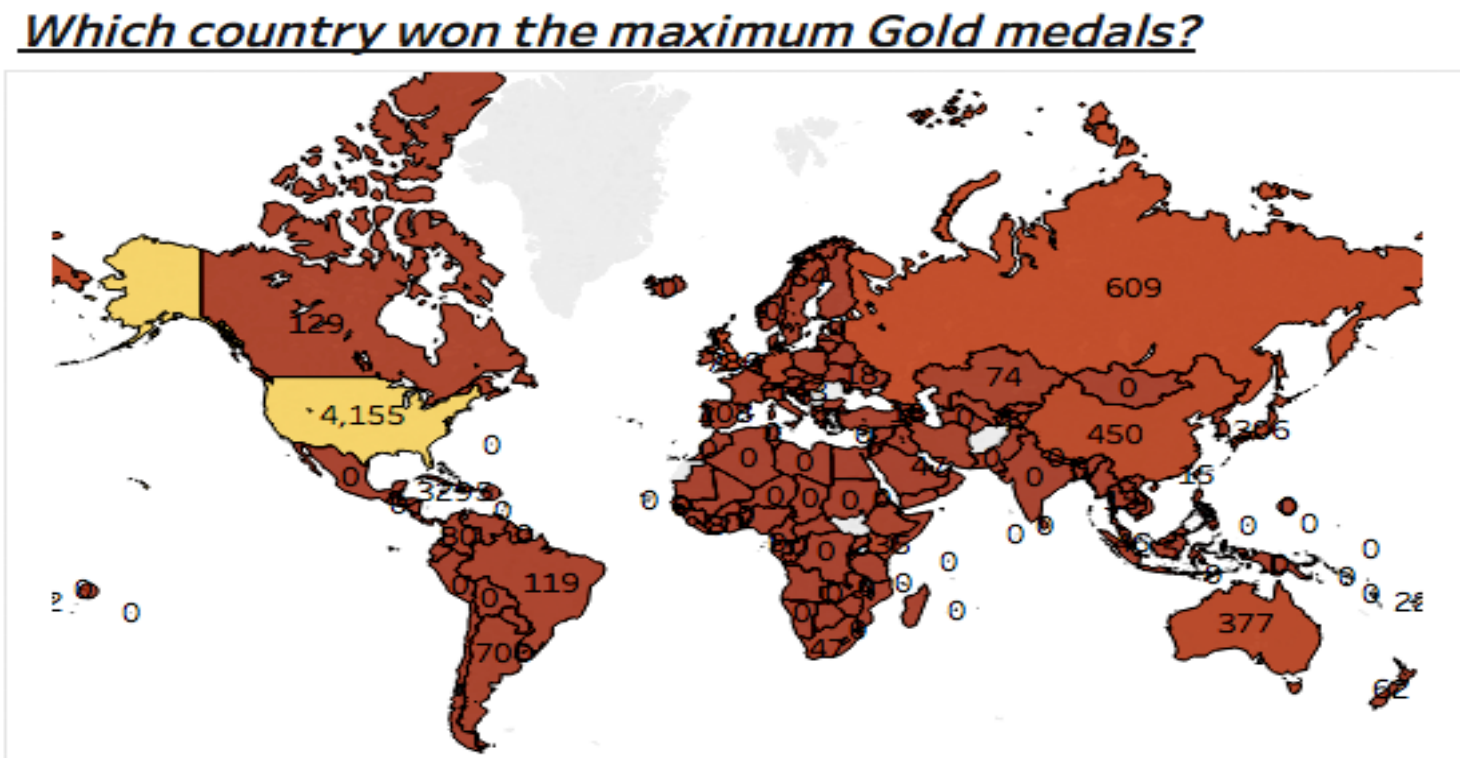
1.3 Countries won the maximum bronze medals

Objective of the analyses is to check which country has the maximum number of medals in total with the help of **filled map and symbol map.**

Preferred Visualization-Filled map is used to get the information about the countries which have won the medals in one view with the proper location in the map (geographically).

Steps to design the graph:

1.Drag the l**ongitude** measure drop into the columns pane
2.Drag the **latitude** measure and drop into the rows pane.
3. Drag the **country** dimension and drop into the detail in the marks pane
4. Drag the **gold measure** and drop into the color in the marks pane and set the color and opacity according to the gold medal colour to represent in a more significant manner.

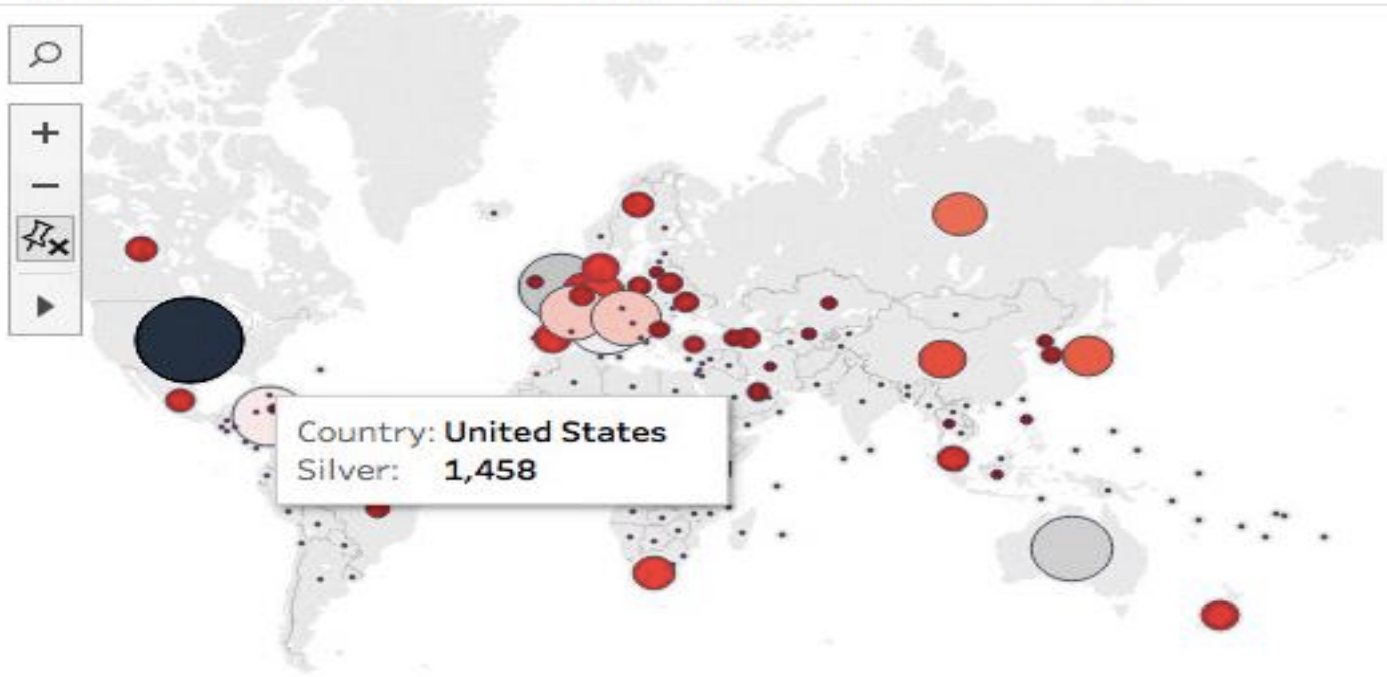


Analyses: We can see that many countries are shaded with dark colour showing 0 medals like India, Nigeria etc. and some countries are shaded little bit lighter showing they own some medals like China, Russia (450,609). Finally, gold colour represents the highest number of medals won and it is own by U.S.A- 4155.

Symbol maps-It is used to represent the size of the silver medals won by the different countries like a filled map.

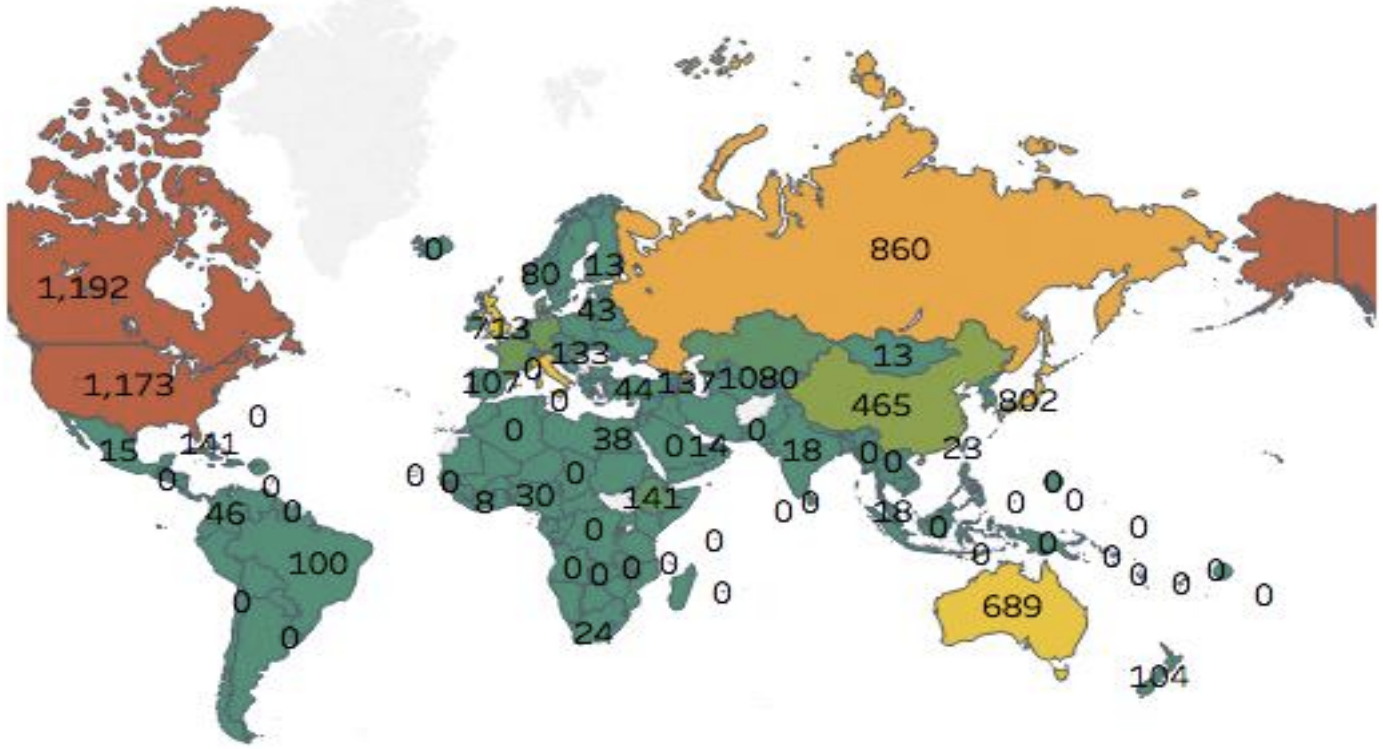
Same steps are followed except dragging the sum of the silver measure into the size in marks pane.

Which country won the maximum silver medals?



Analyses- In this symbol map, the countries are sized according to the silver medals and here also U.S.A takes the crown for the silver medals by having the biggest circle and colored black. However, other countries have won less medals then United States and are smaller in size with red- color.

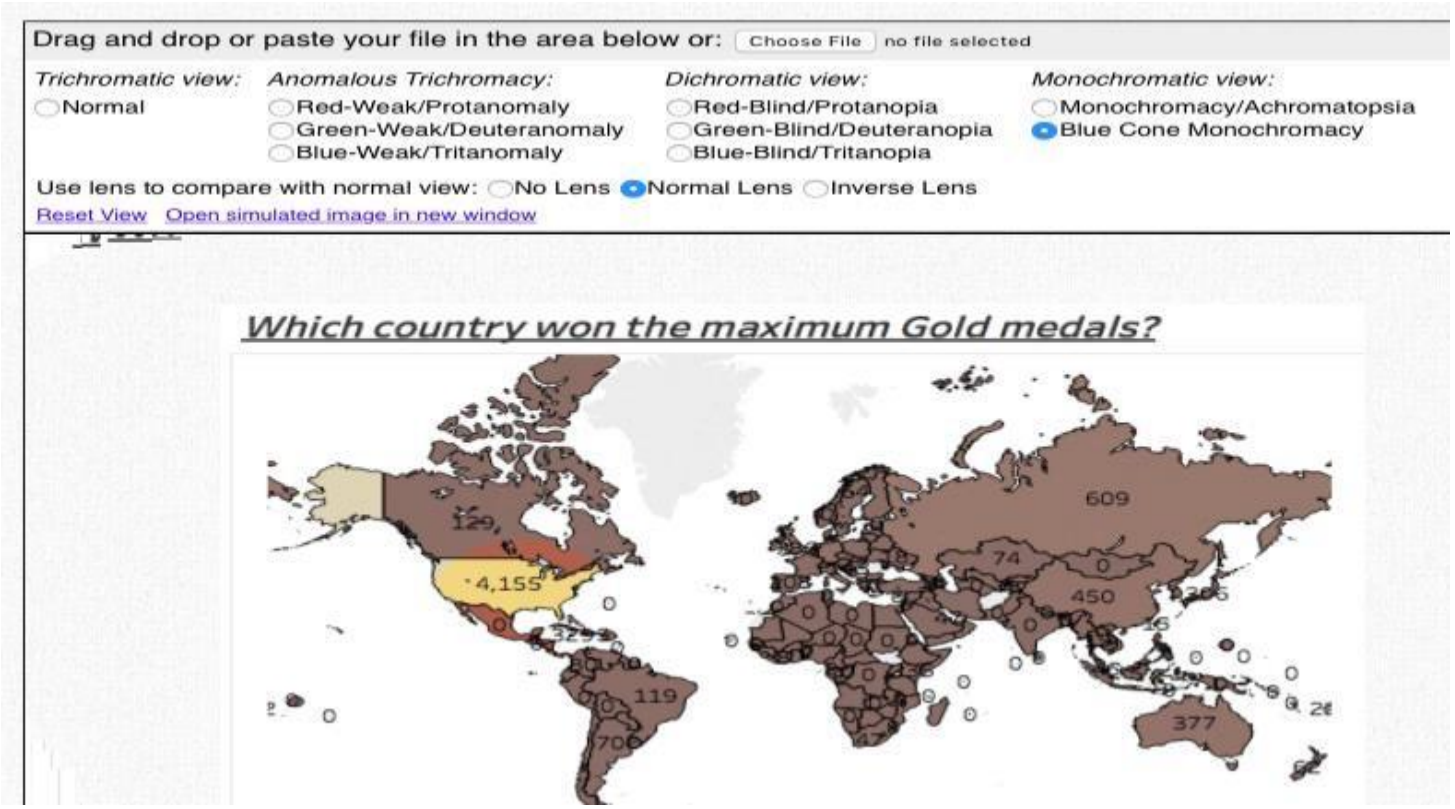
Which country won the maximum bronze medals?



Analyses-As we can see the filter goes from zero to 1,192 with green the minimum and dark bronze (brown colour) for maximum. Here, Canada takes the crown with having slightly more number of medals then U.S.A and many other countries have won as shown in light green like India having 18 medals, brazil 100.

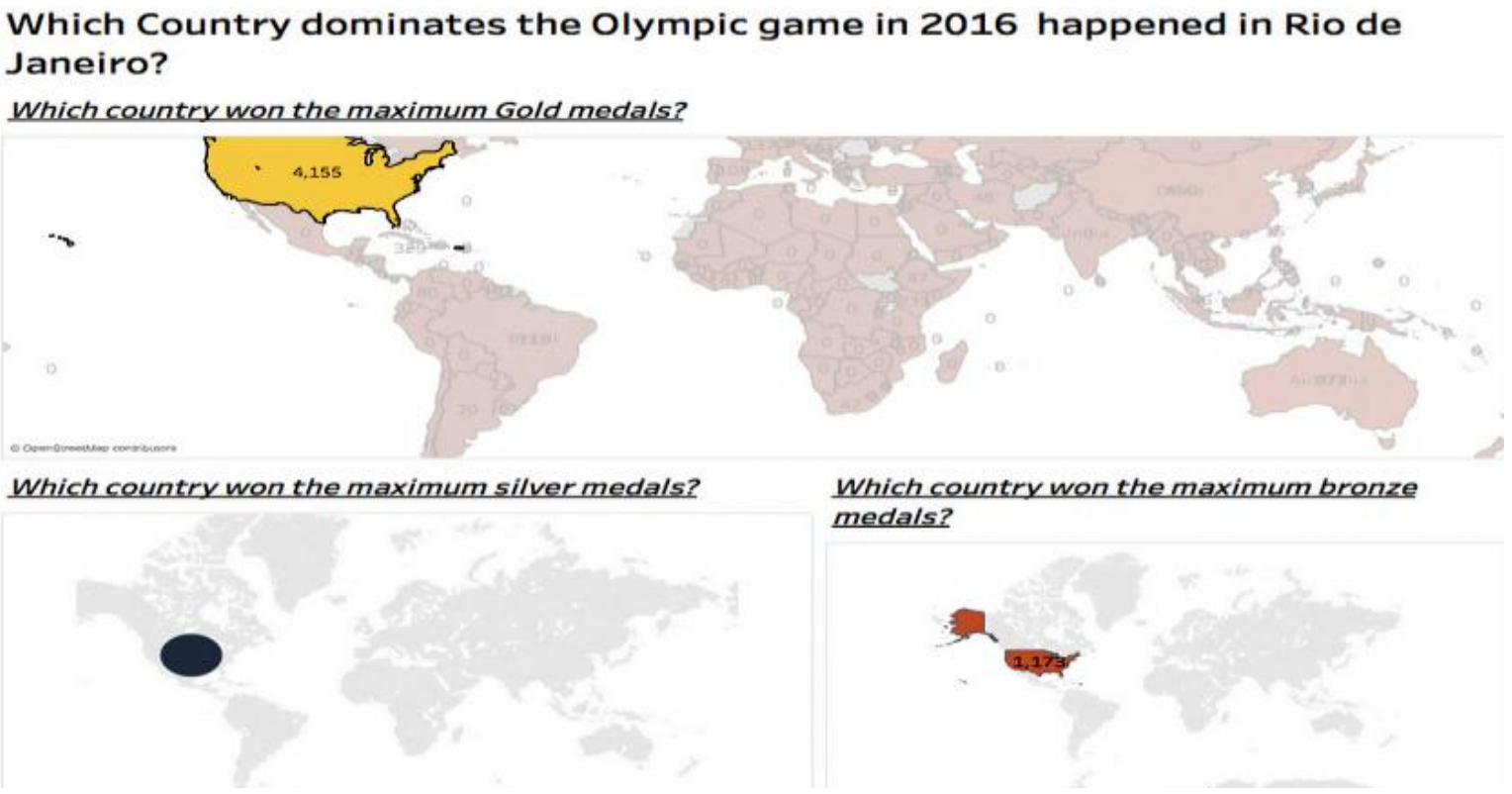
Coblis — Color Blindness Simulator

I have put the file into the simulator to check the colour for blind people into other nature of colour as it should be universally excepted.



http://www.color-blindness.com/coblis-color-blindness-simulator

Dashboard 1



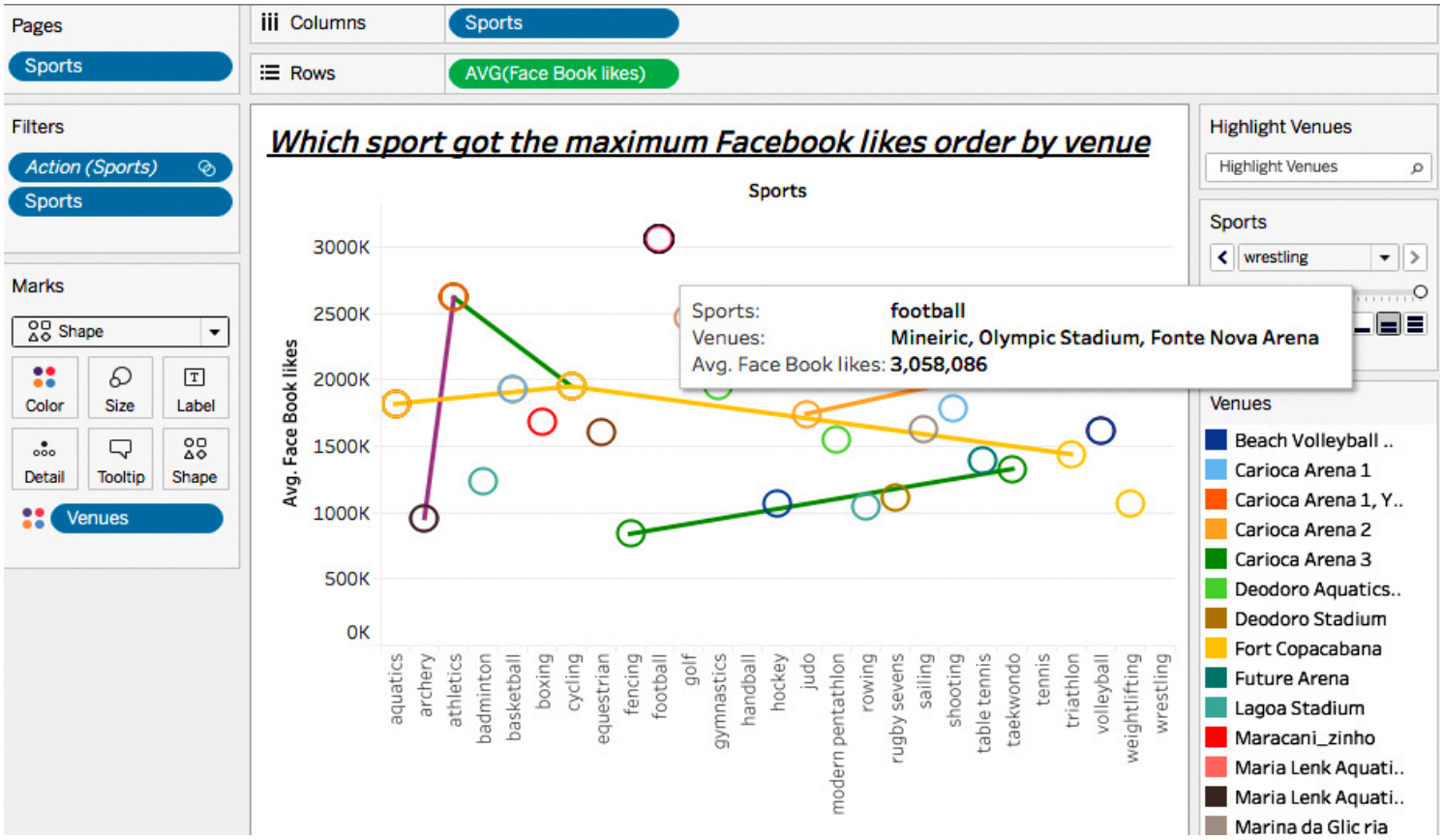
Analyses: As by applying the filters in the dashboard we can see U.S.A dominates the Olympics game in 2016 as being topped in gold medals and silver medals. In bronze medals, Canada takes the crown but the difference was very less. Countries which have got less medals should work upon their skills to get some medals in the next Olympics which will be taking place in 2020.
Result: Hence **U.S.A has the most perfect athletes in the world.**

2. Can we judge which sport is mostly played and popular in Olympics?

2.1 Which sport got the maximum Facebook likes order by venue?

2.2 Which sport is mostly played in Rio Olympics 2016?

2.1 Objective is to find which sport is mostly played and popular in the public happened in Olympics 2016.
Preferred Visualization-Bar chart is used with the shapes chosen from the marks pane.
Animation-Animation by sports linking with the Facebook likes and different venues.
Steps to add a bar chart with shapes:
1. Drag the sports dimension and drop it to columns pane.
2. Drag the Facebook likes measure and drop into the rows pane and take the average of it.
3. Drag the venues from the dimension and add that to color in marks pane.
4. Filter the sports and add that into pages to start the animation
5. On the right side start the animation



Animation completed with different likes by sports order by venue

Analyse- The graph shows the different types of sport liked by public through Facebook. Football got the maximum like which is 3,058,086 which took place in Mineiric, Olympic stadium and least likes are hit on fencing. Hence, fencing should be enhancing in the world. The lines connection shows the same venues where all of the linked sports took place so according to that, Fort Copacabana is the busiest stadium.

2.2 Objective is to find the mostly played game in Olympics 2016.

Preferred Visualization- Word cloud is built to show the most common game as it is the best way to show the popularity by itself i.e. sport**.** Colour palette is color blind to read easily by blind people
Steps to make word cloud:
1. Drag the sports dimension into the color of the marks pane.
2. Drag other sports dimension, change it to the measure count and add that to the size pane.
3. Drag sports dimension and add to the label
4. Change the drop down from automatic to text

Analyze-As per the word cloud, athletics sport has taken place mostly in Rio Olympics 2016 with the count of 93,060 as many players are for athletics only and very less for golf. Players of the golf should be trained more to get participated in the Olympics which will increase their importance of the game too.

Dashboard 2

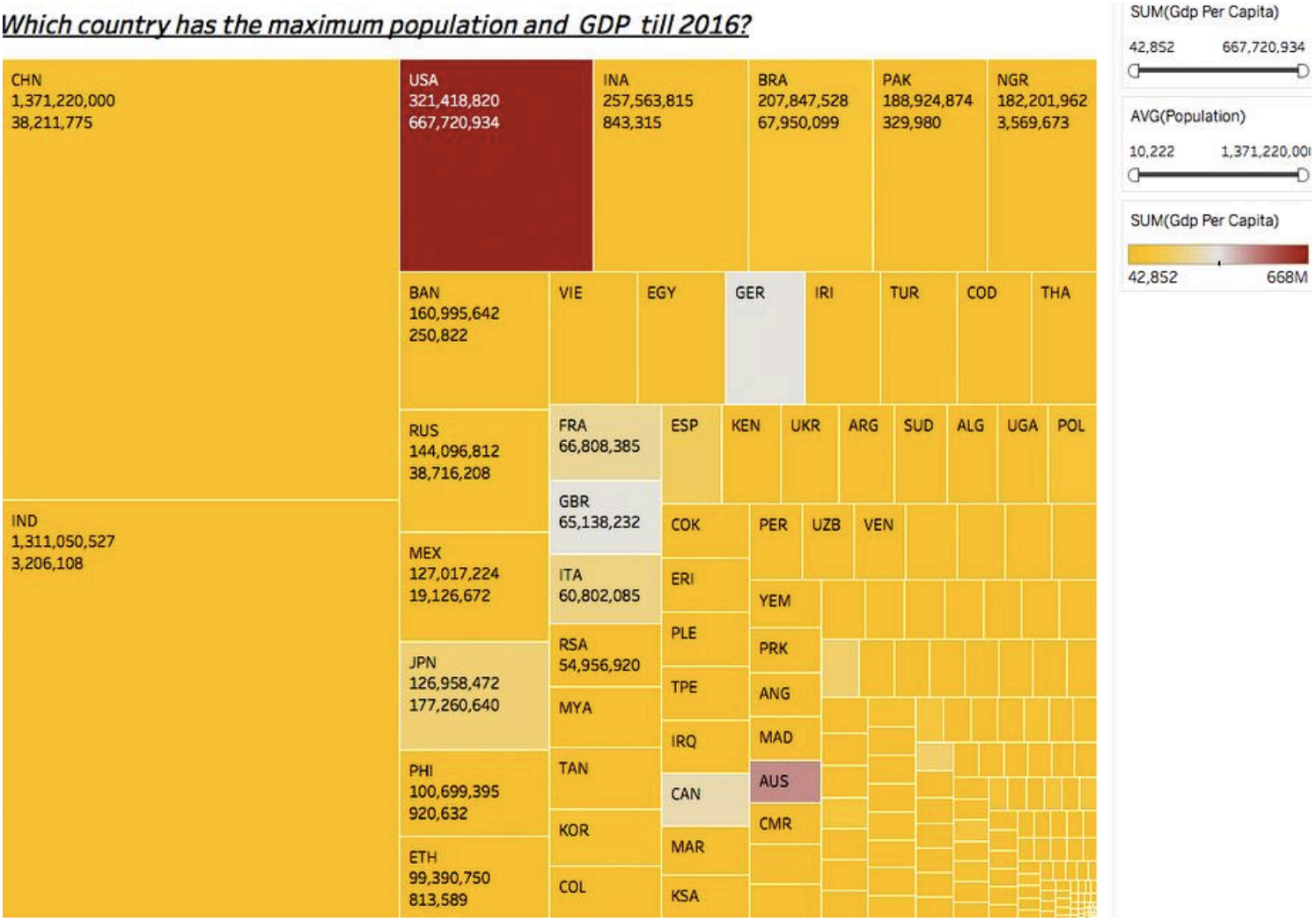
Filtering has been done to select the most common game.

Analyze- Athletics is the most played game in Rio Olympics 2016 with 2,612,342 Facebook likes though football has more number of likes around 3,12,246 but if its compared with the most played game then its counting is around 1,023. Aquatic has also 56,324 counts but it doesn't have the Facebook likes hence athletics is the most prevalent game.

3. Which country has the maximum population and GDP till 2016?

Objective is to find the maximum number of population and GDP per capita in different countries.

Preferred Visualization-Tree maps has been used to visualize different nationalities with their size of the population and coloured by GDP.
Steps to obtain tree map:
1.Drag the GDP from the measure and drop it to the color of the mask pane.
2.Drag the population and drop into the size with changing it to the average
3.Drag the nationality/code from the dimension pane to label of the mask pane.
4.Press control and click AVG population and drop it into the label to get other duplicate of population in the label form.
5.Do same for the GDP by picking from the mask

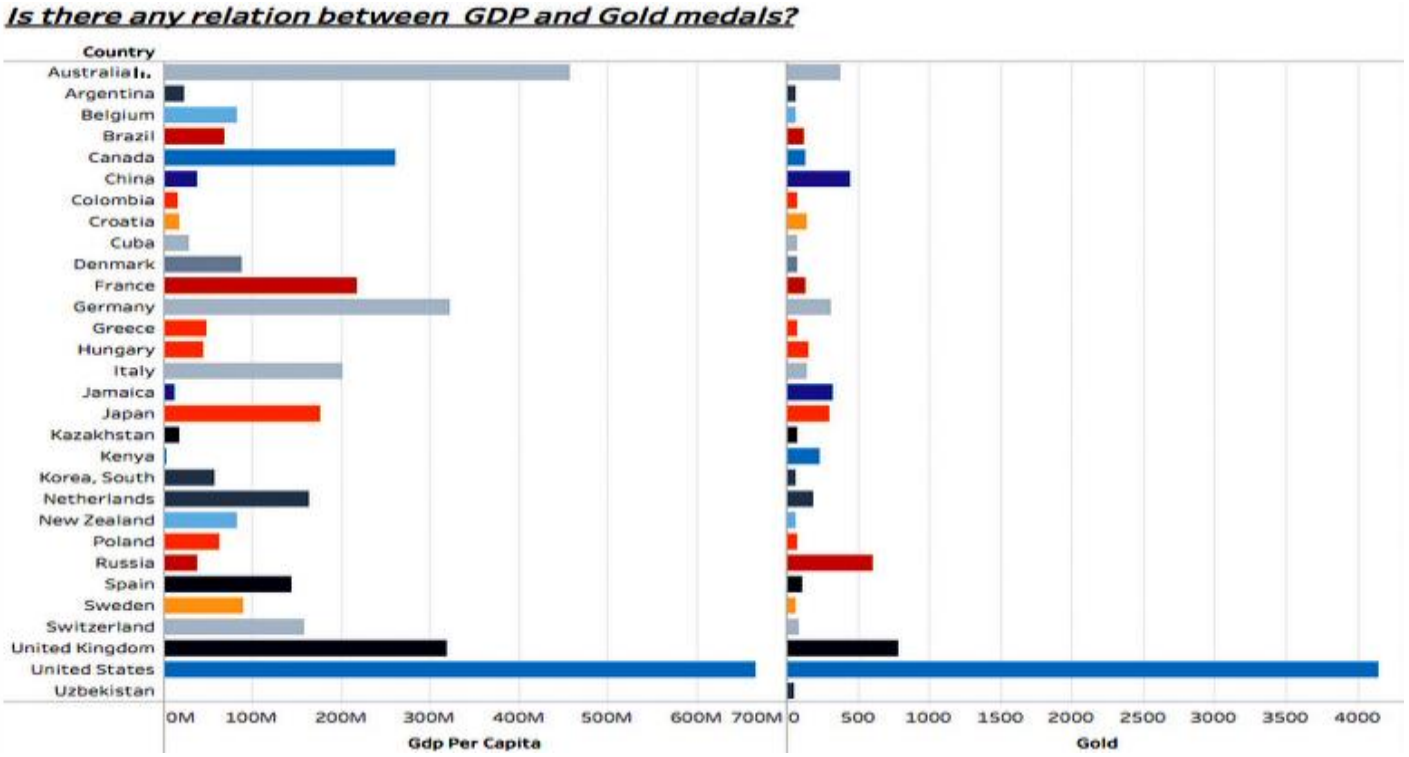


Analyse- Tree map shows China as the most populated Country and second is India followed by U.S.A till the less populated country. In case of GDP, the red colour filter shows that United States has the maximum GDP followed by little bit light purple of Australia but its population is less, Germany takes the third place in case of GDP. Many filters can be applied which is shown on the right tab.

4.Is there any relation between GDP and Gold medals?

Objective is to find if there is any relation with having high GDP and gold medals as government can invest more on trainers and coaches to send their players to represent their country.

Preferred Visualization is **horizontal bar** to view the relation of gold medals and GDP.
Steps involved in making the horizontal bar of Olympics data:
1. Drag the GDP measure and drop it into the columns pane.
2. Drag the gold measure and drop it into the columns pane.
3.Drag country dimension and drop it into the rows pane.
5. Make duplicate of country to add into the color from the mask pane.

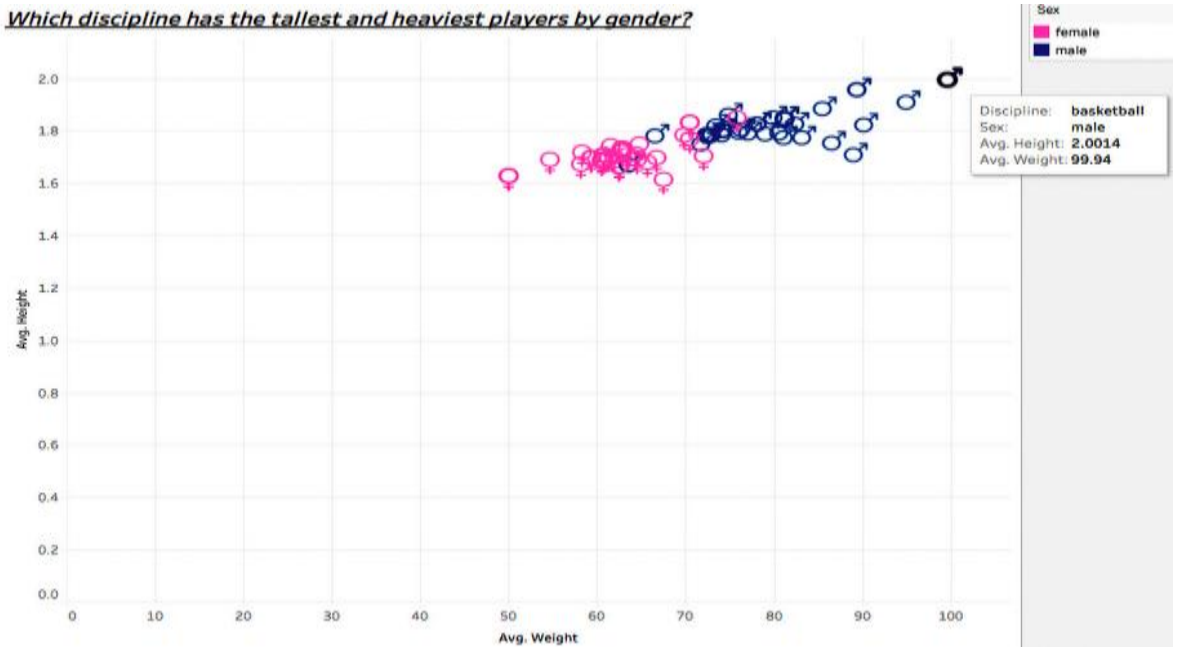


Note-Filtered data to get top 30 countries record

Analyze: Yes, if we study the graph properly we can see there is much significant relation between these GDP and winning gold medals except for Australia and Germany which has high GDP but poor performance in Olympics to get the gold medal. U.S.A, U.K, Canada and many other countries shows that there is much relation between GDP and winning gold medals.

5.1. Which discipline has the most tall and heavy players?

Objective is to find in which discipline there are tallest and heaviest person
Preferred Visualization is **scatter plot** which can be helpful in distributing the discipline by taking the height and weight-different gender.
Steps to make Scatter Plot:
1. Drag the weight from measure and drop into the columns pane and convert it into average.
2. Drag the height from measure and drop into the rows pane taking its average.
3. Drag sex from the dimension and drop into the color in marks pane.
4. Drag discipline from the dimension pane and drop into the detail of marks pane with colors blue and pink to represent the sex.
5. Make duplicate of sex by pressing control and dropping into the shape of the marks pane and convert the shape according to “MALE symbol” and “Female symbol”.



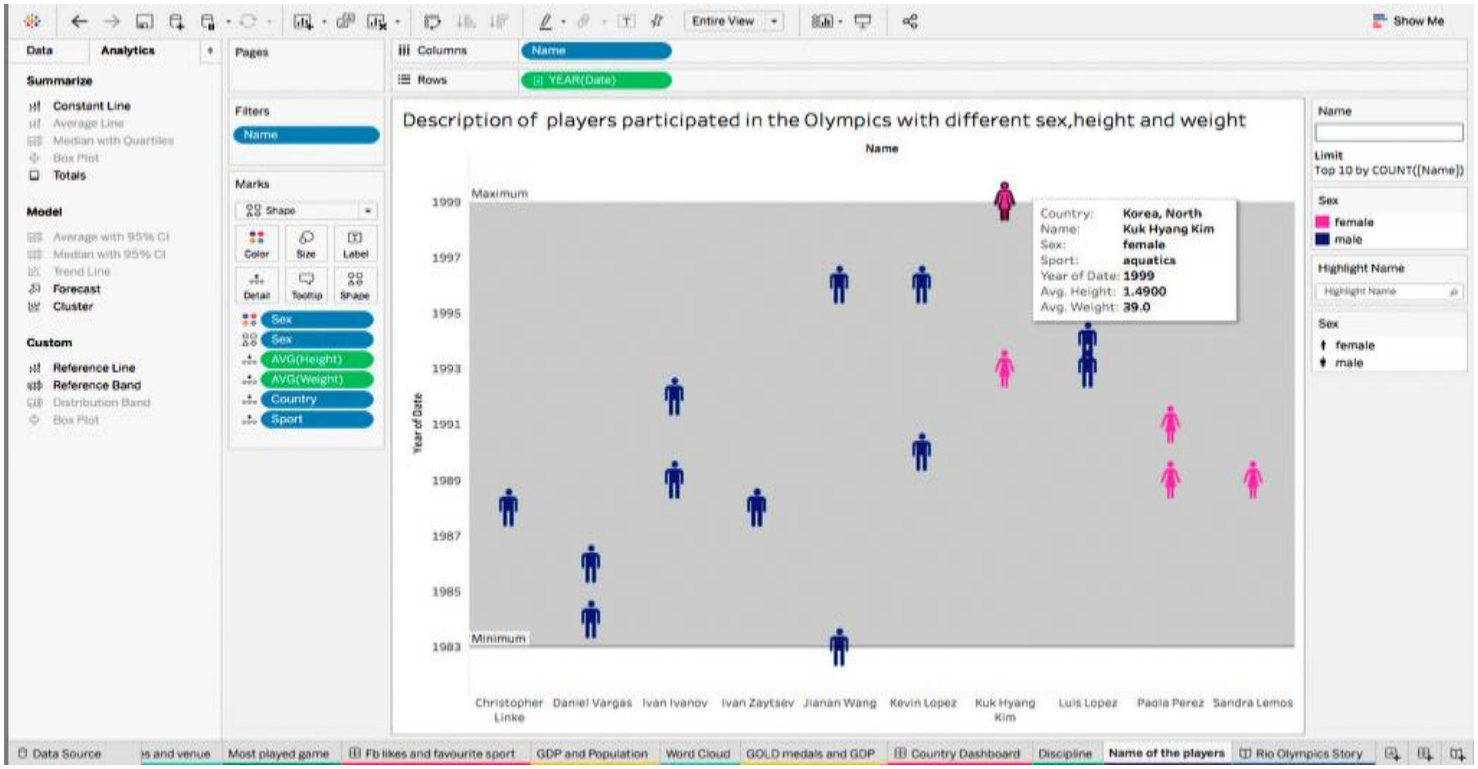
Scatter plot is distributed on the basis of discipline by average heights and weight of different gender.

Analyze- From the scatter plot we can see basketball has the tallest and heaviest male players having 99.94 Kgs weight and 2.01 -meter height. In case of female players, the lowest average height and weight is around 50 Kgs weight and 1.6-meter height.

5.2 Which athlete is the youngest in Olympics with the description of height and weight and country.

Objective is to plot the different number of athletes in plot described by the date of birth (age), weight and height and gender.
Preferred Visualization is scatter plot to see the proper view of the description for the top 10 youngest athletes.
Steps to analyse:
1.Drag sex dimension to the color pane with duplicating it with shape.
2.Drag height and weight to marks pane taking its average.
3. Drag Country from dimension pane to marks pane.
4. Drag name of the athlete's dimension to the columns pane.
5.Drag and drop the sport dimension to marks pane (detail)
6.Drag the year dimension to the rows and making it a measure

Analytics has been done by taking the reference line and reference band for the maximum and minimum to show the athletes age between that era.



(Filtered to top 10 youngest players)

Analyse -Analytics has been done to show the frame. As we can see from the scatter plot youngest player the Olympics 2016 is Kuk Hyang Kim (female as its pink and the shape is for female symbol) from north Korea with the date of birth :1999. She is for aquatics sport.
Maximum and minimum line is taken from analytics pane through distribution band showing the different athletes from 1983-1999 and Jianan Wang from Congo took birth in 1983(eldest), rest of the other athletes took birth between 1983-1999.

Conclusion

In a nutshell, our analysis infers that the evaluations can be helpful for the National committee pertaining to individual countries to grow the importance of Olympics games and prepare their individual sport to have more chances to win medals in 2020 Olympics.

References:

Creaner, O.,2017. *Data Visualization,* Moodle, National College of Ireland

Healey, C.G., 1996, October. Choosing effective colours for data visualization. In *Visualization'96. Proceedings.* (pp. 263-270). IEEE.

Keim, D.A., 2002. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, *8*(1), pp.1-8.

Keller, P.R. and Keller, M.M., 1993. *Visual cues: practical data visualization* (Vol. 2). Los Alamitos, CA: IEEE Computer Society Press.

Milligan, J.N., 2016. *Learning Tableau 10*. Packt Publishing Ltd.

Pop, C., 2013. The Modern Olympic Games–A Globalised Cultural and Sporting Event. *Procedia-Social and Behavioural Sciences*, *92*, pp.728-734.

Ward, M.O., Grinstein, G. and Keim, D., 2010. *Interactive data visualization: foundations, techniques, and applications*. CRC Press.