|  |  |
| --- | --- |
| Name: | Nihar Sudhanshu Limaye |
| Student Number: | 18210876 |
| Email Address: | nihar.limaye3@mail.dcu.ie |
| Module Code: | CA682 |
| Date of Submission: | 16/12/2018 |
| Assignment Title | Visualization of Olympic data |
|  |  |
| Module coordinator | Dr. Suzanne Little |

A report submitted to Dublin City University, School of Computing for module CA682 Data management and visualization, 2018/2019.

I understand that the University regards breaches of academic integrity and plagiarism as grave and serious. I have read and understood the DCU Academic Integrity and Plagiarism Policy.

I accept the penalties that may be imposed should I engage in practice or practices that breach this policy.

I have identified and included the source of all facts, ideas, opinions, viewpoints of others in the assignment references. Direct quotations, paraphrasing, discussion of ideas from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references.

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.

By signing this form or by submitting this material online I confirm that 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.

By signing this form or by submitting material for assessment online I confirm that I have read and understood DCU Academic Integrity and Plagiarism Policy (available at <http://www.dcu.ie/registry/examinations/index.shtml>

Name(s): Nihar Sudhanshu Limaye

Date: 16/12/2018

**Introduction**

The report contains visualization of Olympic data. The purpose of this visualization is to show the Olympic data from year 1896-2016. The visualization includes tools such as spreadsheet, open refine and tableau. The all-time medal table for all [Olympic Games](https://en.wikipedia.org/wiki/Olympic_Games) from 1896 to 2016, including [Summer Olympic Games](https://en.wikipedia.org/wiki/Summer_Olympic_Games), [Winter Olympic Games](https://en.wikipedia.org/wiki/Winter_Olympic_Games), and a combined total of both, is shown below [1].

**Dataset**

This is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. I got the dataset from Kaggle. It is good website containing multiple datasets from different purpose.

The hyperlink for the dataset is given below:

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

The data contains two files with name as athelete.csv and noc\_regions.csv. Both the files are .csv so easy to handle and understand. The data which I got from Kaggle contains some duplicate values so I have to clean the dataset.



Both the files contains the column with name NOC. The second file contains team name with NOC column.

Content

The file athlete\_events.csv contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event (athlete-events) [2] . The columns are:

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

**Process**

To do the visualization I have taken Openrefine,Spreadsheet and Tableau as my toolkit.

1. Getting the dataset:

I got the dataset from Kaggle. But it contains some dirty data with name inconsistency and duplication so that to clean the dataset Openrefine is good.

1. Cleaning the data

The data contains some duplicates and inconsistencies in the name of the athlete. Using open refine I did some of cleansing of the data.

For eg.

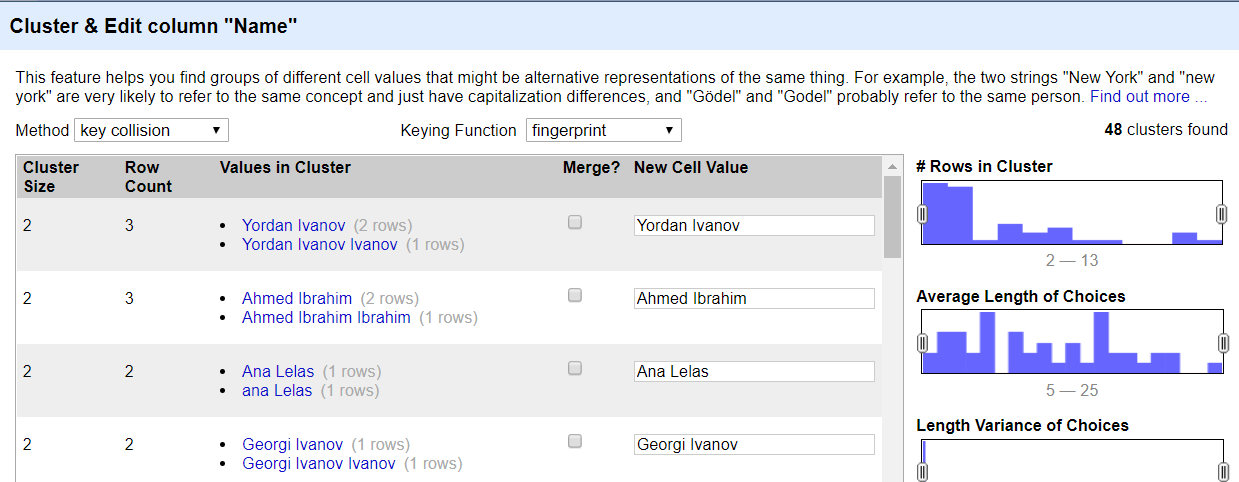


Fig 1. Data Cleaning using Openrefine

The Ana Lelas & ana Lelas both are same but written differently. So Openrefine analyses it and merge it with the name Ana Lelas. We have to this cleaning for all the columns of both files. To remove duplicate data I used spreadsheet.

1. Connection with tableau

To connect with the tableau both the files are imported in it and then both the files are connected with each other with primary key as NOC. I have taken the left outer join because our first file athlete\_events.csv contains major data.

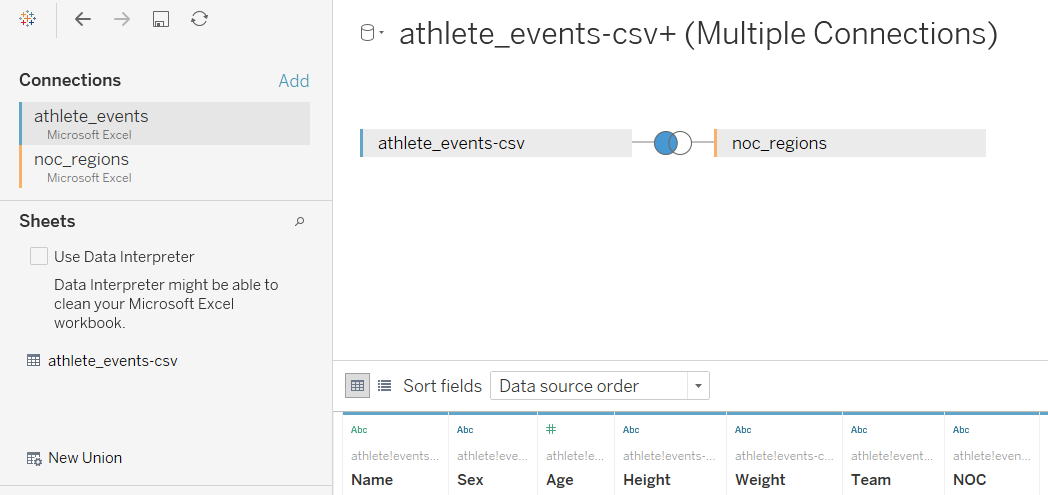


Fig 2. Tableau data source

1. Designing the dashboard.

The dashboard shows the Olympic history with slides such as hosted cities, event count vs years, Medal table and Top 10 sports medal count.

The map shows hosted cities and nations. The color chosen in medal table represents the colors of medals of Olympic i.e Bronze, Gold and Silver.

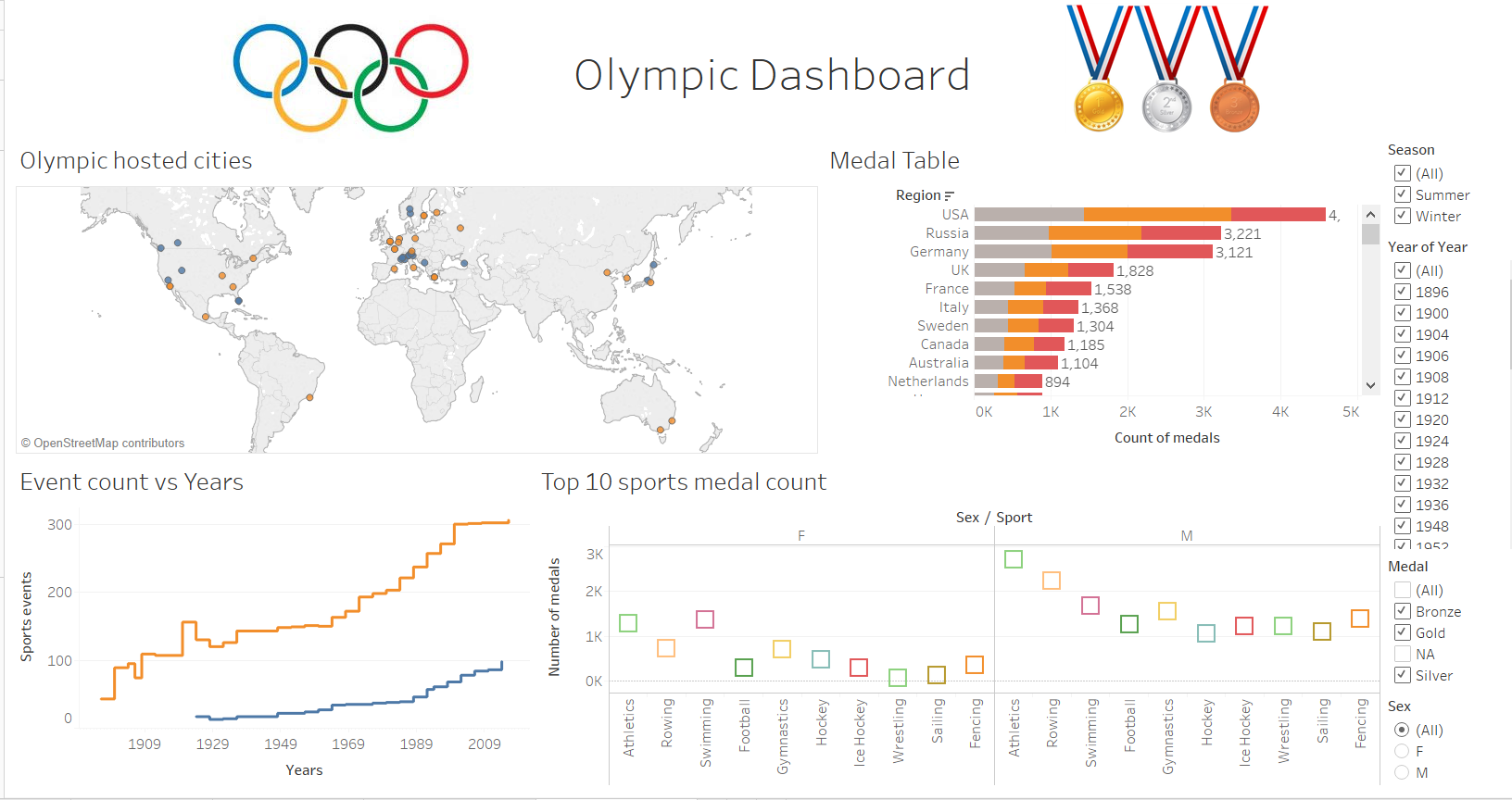


Fig 3. Dashboard of Tableau

The line graph in event count vs years shows sports played in the Olympics respective of their years. The color orange and blue denotes the summer and winter Olympics. And the last quadrant shows the circular graph with popular sports with medal count.

1. Interactive dashboard

The tableau gives us interactive dashboards so with the help of it I have designed some interactive graphs. It tells the story of the Olympic 2016 which was hosted in Rio de Janeiro

The USA was in the top with most of medals 239. The second was Germany with 154 medals. The top 5 countries with most medals were the USA,Germany,UK,Russia and China.

But we can see most females won medals for USA rather than male athletes. The females of USA won 135 medals compare to male which was only 104. The events were 306 in that Olympic.

I have taken top 10 sports of the Rio Olympic and came to know that female athletes won 97 medals in swimming which was highest for a sport and males won 99 in athletics sport.

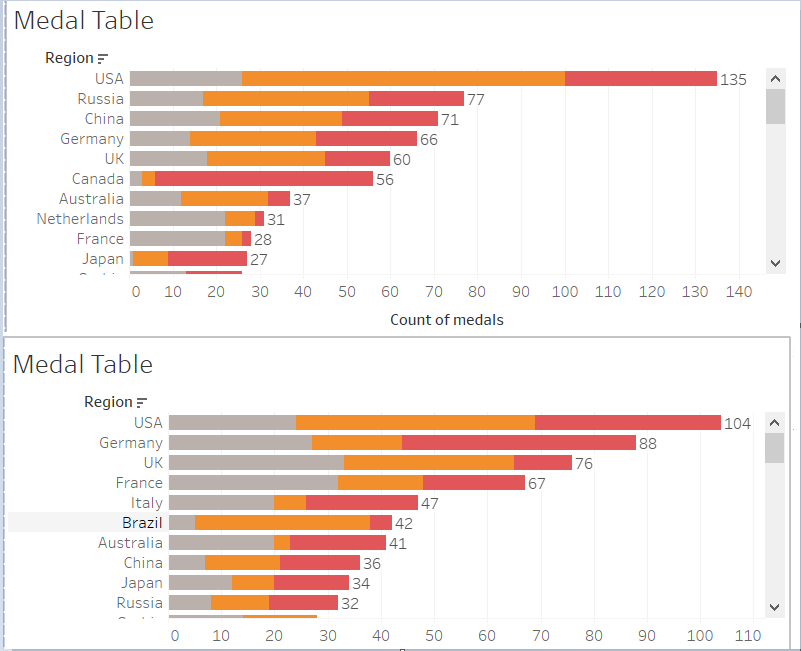


Fig 4. Rio 2016 Olympic data

When it compare to gold medal count comparison for summer and winter Olympics. The USA is far behind Russia in overall medal count for most medals secured. This says Russia do well in winter Olympics and USA do in summer Olympics.

Various analysis we can do with the dashboard some of them are discussed above.

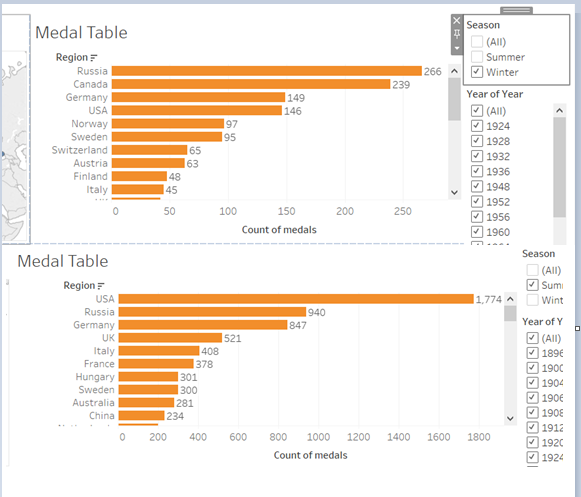


Fig 5. Gold medal of Olympics

**Conclusion**:

It can be seen that Tableau is very good visualization tool once it gets the cleaned data. For data cleaning we can use Openrefine tool. Tableau dashboards are very interactive and easy to understand.

From above visualization I have considered following data visualization principles.

1. Chose the proper graphs for the selected category.
2. The color selection was proper with easy to understand and self-explanatory.
3. Aligned the data and displayed it right.
4. Direct Attention to the Important Details
5. Used clear axes and numbers

# **References**

[1] (n.d.). Retrieved from https://en.wikipedia.org/wiki/Olympic\_Games

[2] Griffin, R. H. (2018, 6 17). *120 years of Olympic history: athletes and results*. Retrieved

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