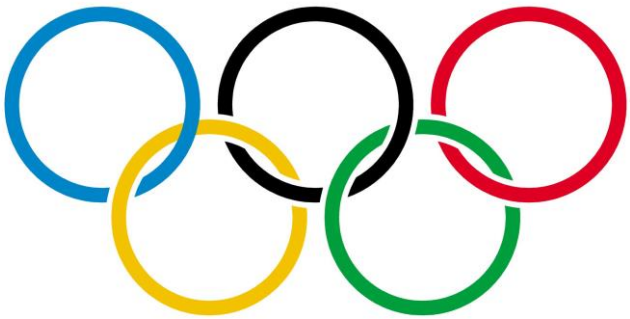
STUDY OF OLYMPIC HISTORY OF ATHENS 1896 TO RIO 2016



NAME: SNEHA RAVIKUMAR

Table of Contents

I. INTRODUCTION

One of the defining features of human beings is Sport. Olympics Games are the most acknowledged international sporting event of the world. Olympics was initially started in ancient Greece which occur every four years (Georgetown Law Library, 2019). The Summer and Winter Olympic Games occurs every four years once alternatively, but both occur two years apart. The first modern Olympics was held in 1896 and Athens was the first host city of Olympics.

Women have come a long way in terms of gender equality at the Olympics games. This is still an on-going trend. Only in 1900, women participated in the Olympic Games. This was four years after the modern Olympics was held in Athens. Since then, female participation is steadily increasing. International Olympics Committee (IOC) and National Olympics Committee (NOC) are committed towards gender equality in sports. Hence, IOC has expanded the women's events in Olympic programme (International Olympic Committe, 2016).

1. Message to be conveyed by the narrative visualisation

The best and biggest achievements of women and men athletes in Olympics who come from various parts of the globe was in Rio 2016. For exposing the professionalism of women to the world, Olympics has been a good platform. Over the Olympic history of past 120 years, there was a multitude of firsts for women. Only by taking a closer look at the trend, we can notice that female participants are trying hard to play a long game to catch up with the male participants in Olympics (Hills, 2019).

Visualising the number are always for the better. Data visualisations changes the organisations and affects the decision-making process. Through this narrative visualisation, we can create a new discussion and spread awareness about the women's role in Olympics. To find out more about women in Olympics, let's do some analysis and visualize certain aspects related to participation and performance of women.

2. Intended Audience

Since, women's participation and performance are not discussed that often, the main intension is to create more awareness about the same among the public and peer classmates as well. Also, the visualisations will be helpful for the NOC to get to know how women's participation and performance in Olympics has changed over the years and take data-driven decisions regarding the same.

The focus of this report will be answering the following questions for all the intended audience as mentioned earlier:

- How women's participation has changed from Athens 1896 to Rio 2016 Olympic games?
- How women's performance has changed from Athens 1896 to Rio 2016 Olympic games?
- How the number of athletes participating from each county has changed from Athens 1896 to Rio 2016 Olympic games?

II. DESIGN

1. Data Sources:

The main dataset of Olympics games is obtained from Kaggle website. This has two datasets, namely, athlete_events and noc_regions. The athlete events dataset has 271,116 rows and 15 columns. These columns are namely, ID, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event, Medal. The noc regions dataset has 3 columns and 230 rows. These columns are namely, regions, NOC and notes. These datasets were merged and wrangled for the analysis and building visualisation purposes. The merged dataset contains 271116 rows and 19 columns.

Also, for building a specific visualisation, the country's latitude and longitude co-ordinates are also required. Hence, co-ordinates dataset is obtained from Developers google website. This dataset consists of 245 rows and 4 columns. The columns are namely, country code, latitude, longitude and name of the country. This dataset is also merged and wrangled with the original dataset.

The final wrangled dataset consists of 267,538 rows and 23 columns. Since the dataset is quite large, the application is built only on selective columns, namely, ID, Sex, Age, Height, Weight, Team, NOC, Games, Year, City, Sport, Event, Medal, regions, latitude and longitude. The data is aggregated at each step to build the appropriate application. The links for the data sources is given below:

- https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results#noc_regions.csv
- https://developers.google.com/public-data/docs/canonical/countries_csv

2. Five Design Sheet Methodology:

A useful way to plan and consider alternative solutions is by sketching designs. Few low-fidelity prototyping methods, mainly paper-based sketching saves a lot of money and time and provides recommended solutions rapidly. Five Design Sheet methodology helps to manage the ideas and thoughts, initially it allows to think broadly and then converge to an appropriate solution (Roberts, Headleand and Ritsos, 2016).

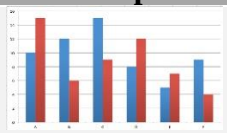
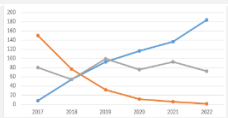
The Five Design Sheet consists of five pages. The sheet-1 allows the user to plan and sketch their broad ideas, helps the user to think about the appropriate solutions to the assigned task. The sheets- 2,3 and 4 allows the user to develop three principle designs. The sheet-5 allows the user to consider the various options mentioned in sheets-2,3 and 4 and converge to the most appropriate final design which can be implemented (Roberts, Headleand and Ritsos, 2016).

i. Sheet-1:

The sheet-1 allows the user to generate ideas, filter them, categorize them, combine and refine them. Finally, allows the user to question if the design provides a solution to the intended task. In this sheet, various ideas are sketched to address the main goal of this report, i.e., gain more insights about women in Olympics.

The columns that are of focus from the final dataset are listed down to generate various ideas to represent them. Various ideas that were generated were filtered out to remove duplicate or impossible ideas. Initially, an idea was popped up to merge the GDP and population data along with the final dataset obtained, to find out more insights about how the GDP and population is affecting the participation and performance. Since, the final dataset is already quite huge, and many attributes are considered for achieving the goal of this report, this idea was filtered out as it is impossible to represent everything at once. The various ideas generated was categorized, combined and refined as shown in the figure- 8.

The various ideas for representation that were considered are mentioned below:

Type of Chart	Example
Bar Graph	
Line Charts	

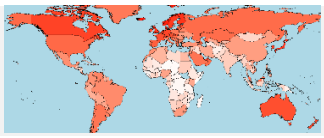
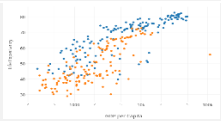
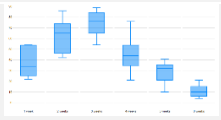
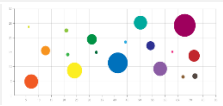
Choropleth Map	
Scatter Plot	
Box Plot	
Bubble Chart	
Population Pyramid	

Table 1: Table representing different ideas

ii. Sheet-2:

The sheet-2 contains the layout of a design proposed to meet the requirements of the user. The layout consists of 3 menu items to choose form, namely, Participation and performance, Overall Trend and Overall medal count. Upon choosing each menu by the user, the respective graphs show up as shown in the figure- 9. A line graph with a filter option to choose the sex appear when the user clicks 'Participation and Performance' menu. The line graph shows the trend of men and women participation in Olympics across all the countries from 1896 to 2016. A bar graph with a filter to choose country appears when the user clicks 'Overall Trend' menu. This graph shows the overall trend of athletes count across all the years for each country. A stacked bar chart appears when the user clicks 'Overall Medal Count' menu option. This graph shows the medal count across all years for each NOC. This also shows the count of gold, silver and bronze medal as shown in the figure-9. When user hovers over the line chart and the bar chart, respective information will appear in a tooltip.

The drawbacks of this design are that it is a simple solution which doesn't represent any depth information as per the user requirements. All the requirements of the user are not covered as well.

iii. Sheer-3:

The sheet-3 contains the layout of one of the design solutions to meet the requirements of the user. This layout contains all the visualisations on a single page as shown in the figure-10. The line graph shows the male and female trend of participation over the years in all countries on the same graph. There is an additional information provided to the user about the correlation between the height and weight of the athletes if various sports. The population pyramid graph changes according to the user's selection from the country drop down provided. This graph is used for comparison of age group of the athletes participating. This comparison is done for both male and female athletes. The bar graph represents the trend of medal count for each NOC. This graph shows the trend for both male and female participants. When user hovers over the scatter plot, respective information will appear in a tooltip.

This design provides a good overview of a solution to the user's requirement. Bar charts require additional explanation and it can be manipulated easily. Also, the comparison of height and weight for all the athletes doesn't provide a good visualisation to understand the relation. It can be filtered out according to the sex. This solution is less interactive.

iv. Sheet-4:

The sheet-4 contains the layout of one of the design solutions to meet the requirements of the user. This layout also contains all the visualisations on a single page as shown in the figure-11. The line graph shows the participants trend over the years for all countries. This solution provides a filter for the line graph, from which the user can choose the sex of the athletes and according the graph automatically changes. It also provides the user with a 'Year' slider for the line graph, with minimum value of 1896 and maximum value of 2016. The user can slide up to the year of interest and check the trend of the athlete participation (especially women). The box plot shows the variation of height and weight of the athletes across all the countries for each year. This graph contains a filter, from which user can select to view the height variation trend or the weight variation trend over the years. The choropleth map shows how many athletes are coming from which region of the world. This map contains a 'year' filter, from which the user can select the required year and observe how many athletes are coming from each region. When hovered over each region, the user is provided with a tooltip displaying the required information. The trend of the medal counts of participants in each year is represented with a Graphics

Interface Format (GIF). This graph automatically changes according to each year from 1896 to 2016 displaying the medal count from each region.

The main focus of this design would be the slider input provided for the line chart and the map with bubble chart in Graphics Interface format. This design is highly interactive and GIF provides the users with a good animation. This design is very difficult to implement as the dataset is huge and this design requires lots of aggregation methods. Displaying all the graphs on one page would make it look uncoordinated.

v. Sheet-5:

The sheet-5 shows the final realisation after comparing all the design ideas presented in sheets- 2,3 and 4. This sheet consist of a layout of the final design which includes three menu items (idea taken from sheet-1). This is represented in the figure-12. The menu items are 'Participation and Performance', 'Overall Participation Trend', 'Overall Performance Trend' and 'Medal Count Trend'.

When the user selects the 'Participation and Performance' menu option, three graphs appears along with few sidebar layouts for the filters. This menu option includes a common filter for all the three visualisations, i.e., a 'sex' filter which allows the user to choose between male or female options. The first graph shows the participation trend of the athletes over the year across all countries with the help of a line graph. This line graph also has an 'year' input slider, by changing which the line graph changes accordingly. The next graph is the scatter plot that shows the change in height v/s change in weight over the time for the athlete's sports. Since the dataset is large, this graph is built by eliminating few events, namely, weightlifting, boxing and wrestling. Also, a separate linear regression line is fit for weight and year, height and year. The colours in this graph represent different sports. The bubble chart with the map represents the athlete's medal counts trend across Countries from Athens 1896 to Rio 2016 Olympic Games. This graph has 2 more filters apart from the main 'sex' filter, namely, 'year' filter and 'medal type' filter. Also, due to the size of the dataset, this visualisation is created for every 40 years, starting from 1896. This graph changes according to the options chosen from the filter by the user. The colours in this graph represent the countries and the sizes of the bubbles represent the medal count.

When the user choses the 'Overall Trend' menu option, two graphs appear along with a side bar for the filter. This is a 'year' filter for both the graphs. The first graph is a choropleth map that represents the number of athletes count coming from various

countries. As mentioned earlier, since the dataset is huge, this visualisation is built for every 40 years of Olympics data starting from 1896 to 2016. When the user selects a particular year, the graph changes accordingly. The darker colour represents more athletes count. The second graph is the population pyramid graph that represents the comparison of the athlete's age group, for both male and female athletes.

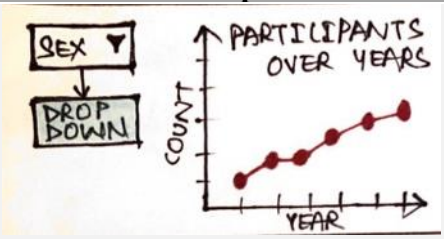
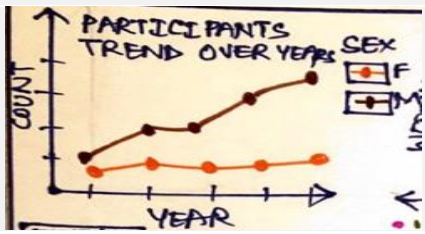
When the user choses the 'Medal Count Trend' menu option, a GIF appears, that represents the trend of medal count in each region for all the year from 1896 to 2019. This graphs automatically runs as it is a GIF. The colours represent different countries and the sizes represent the medal count of that region.

When the user hovers over all the graphs, the respective information will be displayed. The focus of this design will be all the drop-down options as discussed above, input slider, map GIF and interactive choropleth maps. The time estimated to build this application is 2 weeks. This application is built in R-studio. A shiny application will be developed. The software requirements are a browser that supports R.

All the designs discussed above provide the user with an option for exploring the data more and helps them to compare the results to take data-driven decisions.

3. Alternatives Considered:

i. For representing the women's participation, the following alternatives were considered:

Alternative charts	Example
Alternative-1: **Line Graph with a drop down for selecting between male/female (from sheet-2).**	
Alternative-2: **Line Graph representing for both male and female participation trend in the same graph (from sheet-3).**	

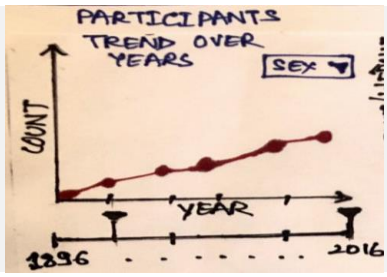
Alternative-3: **Line Graph with a drop down for selecting between male/female and also a slider for the Years. The user can choose from which year to which year they want to see the trend of participant count (from sheet-4).**	

Table 2: Table representing alternative visualization options for women's participation

ii. For representing the women's performance, the following alternatives were considered:

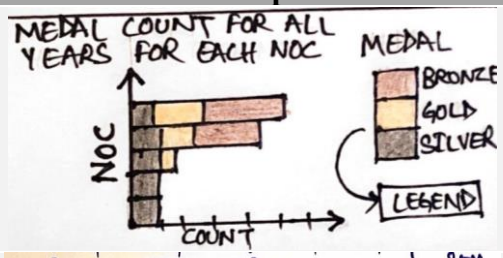
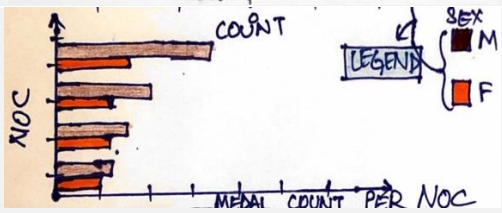
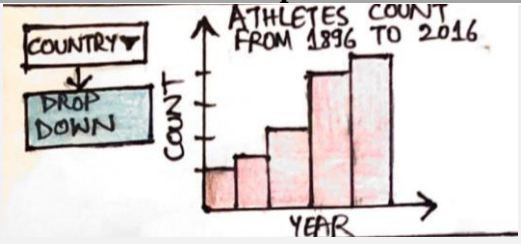
Alternative charts	Example
Alternative-1: **Stacked bar graph representing the medal count for all years in each NOC (from sheet-2).**	
Alternative-2: **Bar chart representing medal count for all years in each NOC for both male and female participants (from sheet-3).**	
Alternative-3: **Bubble chart with world map in a GIF format representing the medal counts variation in each region for all years from 1896 to 2016 (from sheet-4).**	

Table 3: Table representing alternative visualization options for women's performace

iii. For representing the variation in athlete count coming from different regions, the following alternatives were considered:

Alternative charts	Example
Alternative-1: **Bar graph with a drop down for selecting different country. (from sheet-2).**	

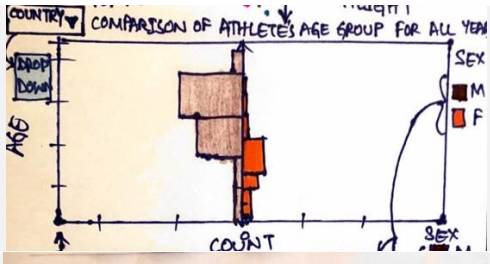
Alternative-2: **A population pyramid graph with a drop down for choosing different country. This graph shows the age of the athletes, for both male and female participants (from sheet-3).**	
Alternative-3: **A choropleth map with a drop down for selecting a particular year of Olympics (from sheet-4).**	

Table 4: Table representing alternative visualization options for athletes count from different region

4. Justification for final design:

i. Layout:

The final design has the menu options similar to the Design sheet-1. When compared to the other design sheets (without different menu options), the sheet-1 provides a better overview and clear differentiation between the topics. It is more readable for the user when the visualisations are presented in different menu options or tabs. Also, the application creates an interaction with the user.

ii. Grouping of charts according to menu options:

The 1st menu option 'Participation and Performance' mainly focuses on women's participation and performance. Hence, all the graphs related to this topic are put together. It is easier for the user to compare the women's participation trend and performance trend. The user is also provided with a filter option (drop-down) to compare men's participation and performance as well.

The 2nd menu option 'Overall Trend' mainly focuses on all the athlete's count from various regions. Hence, all the graphs related to this topic are put together. The user is also provided with additional visualisation, to compare the athletes age group as well. It is easier for the user to compare them and analyse.

The 3rd menu option 'Overall Medal Count' mainly focuses on the medal count trend for all the athletes irrespective to sex. Hence, it is better to show it in a separate menu option. The way, it reduces confusion for the user and it is easier for the users to interpret.

iii. Final visualisation for women's participation:

The alternative-3 as mentioned in table-2 is chosen for the final design, as the is more interactive and provides the user with multiple options to analyse deeper. This is done by providing the input 'sex' filter as well as input 'year' slider. Both these options are present only in alternative-3 from table-2.

Along with this, the user is provided with an additional plot that compared the change in height v/s change in weight of the athletes in various sports. This provides the user with additional information to get to know for which sport, change in height and change in weight is more.

iv. Final visualisation for women's performance:
The idea from alternative-3 as mentioned in table-3 is chosen in the final design. Instead of GIF, the graph is made interactive by providing 'year' and 'medal type' drop down options. This graph also provides a tooltip with useful information for the user. Also, this graph has an additional drop down to choose between male or female. Since, the GIF in alternative-3 from table-3 only represents for all the athletes irrespective of sex, the GIF is placed in a different menu option to represent the overall medal count trend.

v. Final visualisation for variation in athlete count coming from different regions:
The alternatives-2 and 3 from table-4 are chosen in the final design as this provides the user with additional information than the intended ones. The map provides the user with the global trend from various geographical regions.

The final design provides the user with useful content. The fonts, colours and backgrounds are appropriate in a way that users can easily understand as the design is very clear and precise. The additional visualisations provided makes the application even more interactive and attractive as well.

III. IMPLEMENTATION

The implementation of the final design was done using R as mentioned earlier. The data is taken from the respective sources as mentioned earlier. After merging and wrangling of the datasets, they were aggregated according to each visualisation, followed by which the visualisations were built. Initially, all the visualisations on the final design sheet was built separately in a R script file and then implemented in a shiny application, which is a web-based application.

1. Reasons for Implementation design:

R is a software environment and programming language used for graphics and statistical computing, graphical representations and reporting. RStudio is an IDE (Integrated Development Environment) used to develop R scripts (Mehta, 2019).

- RStudio and R: The main reason for choosing R is that the author is familiar with R programming when compared to D3. The author could provide good visualisation with the help of RStudio. R connects with different languages as well (Vries and Meys, 2019).
- Shiny Dashboard: The final design is implemented in the form of a dashboard. This is done with the help of shiny dashboard. It is one of the packages that R provides, which helps in building interactive web-based applications from R. Dashboards are widely accepted as they help businesses to derive insights from the data that is existing (AMR, 2019).
- User Interactions: The application contains the necessary information that allows the user to drive more insights from the visualisations and gain more awareness about women in Olympics. All the graphs have been made interactive, which helps the user not to get confused. Additional information has been provided as well through the visualisations.

2. Libraries Used:

R packages is a group of R data, complied code in a good format and functions. These packages are stored in a directory called library. R provides the user with a standard package set. Other packages can be used by downloading and installing them. Once, the packages are installed, the user needs to load the library in the session that is being used (Kabacoff, 2019).

The libraries used for building this application are mentioned below:

Package Name	Description	Usage
shinydashboard	▪ Helps in building a dashboard in R. ▪ Contains 3 parts, namely, header, sidebar and body.	▪ Used to create filters. ▪ Used to display all the graphs.
shiny	▪ Helps in building interactive web-based applications from R. ▪ Easy to write. No skills like web development required	▪ Used to build the application for producing the intended visualisations

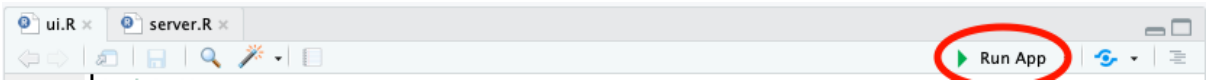
ggplot2	▪ Provides graphics language to create complex and elegant graphs. ▪ Grouping can be done as well. ▪ Most popular R packages.	▪ Line graphs, scatter plot, bubble chart, choropleth map, population pyramid, map GIF
plotly	▪ Used to produce an interactive web-based visualisation. ▪ The interactions are tooltips, zooming.	▪ Used to create interactive charts. ▪ All the graphs are made interactive in this application.
maps	▪ Used to display maps ▪ Requires co-ordinates of the regions for plotting.	▪ Used to plot map GIF and the bubble chart with map.
gganimate	▪ Helps to declare animations with the use of an API known to ggplot2 users. ▪ Key features: Views, transitions and shadows.	▪ Used to ceate the bubble chart with map GIF.

Table 5: Table representing the various libraries used to build the application

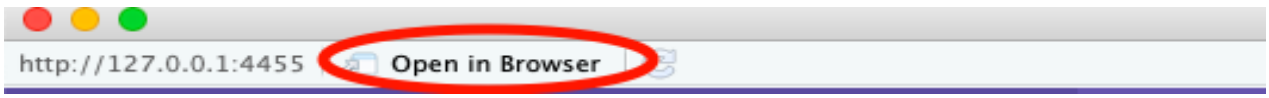
IV. USER GUIDE

This section provides an overview of how to run the application that was built.

- After downloading the zip file provided, open the RStudio software. Place the unzipped files in the working directory of RStudio.
- Open the ui.R and script.R files found in the folder provided.
- To run the application, click the 'Run App' option that appears once you open the ui.R and script.R files as shown below.



- After running the application, click on 'Open in Browser' option available in the pop-up as shown below. Once you click this option, it takes a few seconds to load as the dataset is huge and it needs to be aggregated.



- Once the application is open in the browser, a dashboard appears on the screen as shown below. This is the shiny dashboard that is built to address all the main goals of this report.

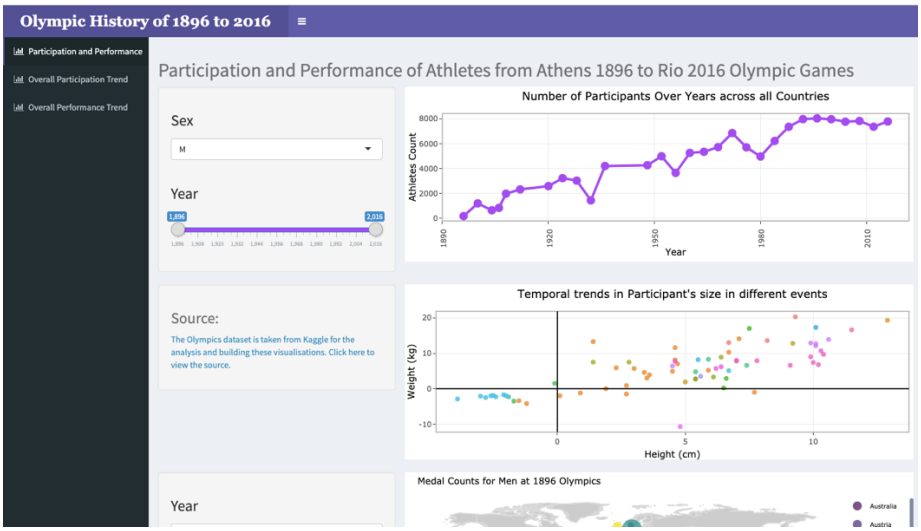


Figure 1: Figure showing the final application that was built

This dashboard provides us with 3 menu options/tabs. If the user wishes to see only the graphs without displaying the menu options, then click the below mentioned icon, which hides the menu items and displays only the graphs with the filters.

The dashboard without the menu options is as shown below.

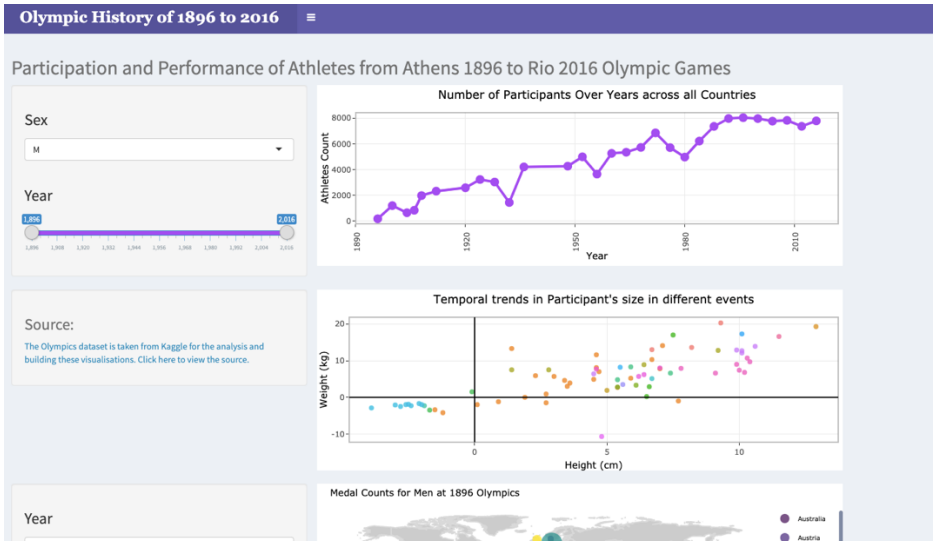
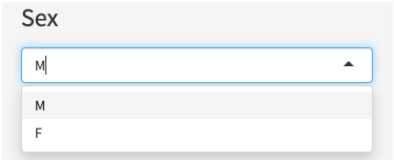


Figure 2: Figure showing the final dashboard without the menu options

Scroll up and down to see all the graphs displayed on the first menu option. Three graphs appear in the first menu option.

- As mentioned earlier in the report, the 'Sex' filter is for all the 3 graphs.



The graphs change according to the sex input chosen. The 'Year' slider input is for the first line graph. Slide through the years and the graph changes respectively as shown below.

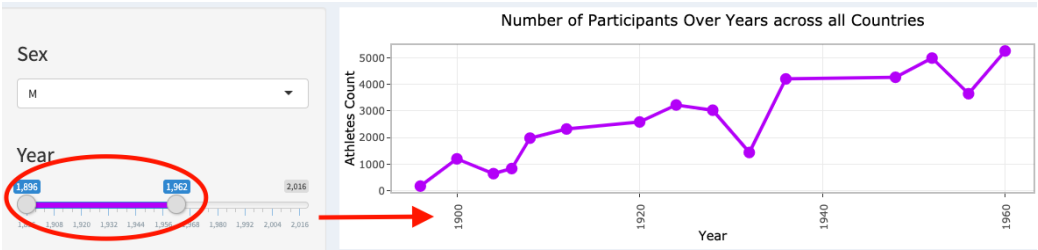


Figure 3: Figure representing the change in the graph according to the change in the slider input

The source option is displayed in all the menu options. Clear instructions are specified as well on the dashboard. Click the source text to see the data source page as shown below.

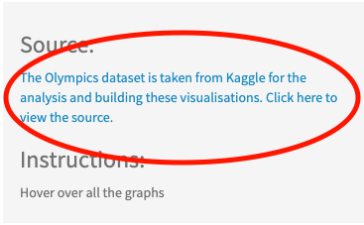
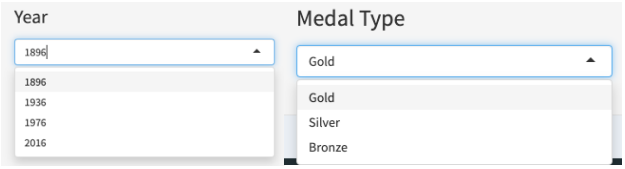


Figure 4: Figure showing the source option displayed on the dashboard

The 'Year' and 'Medal Type' filter shown below are for the bubble chart with map. This map changes according to the inputs chosen from these filters.



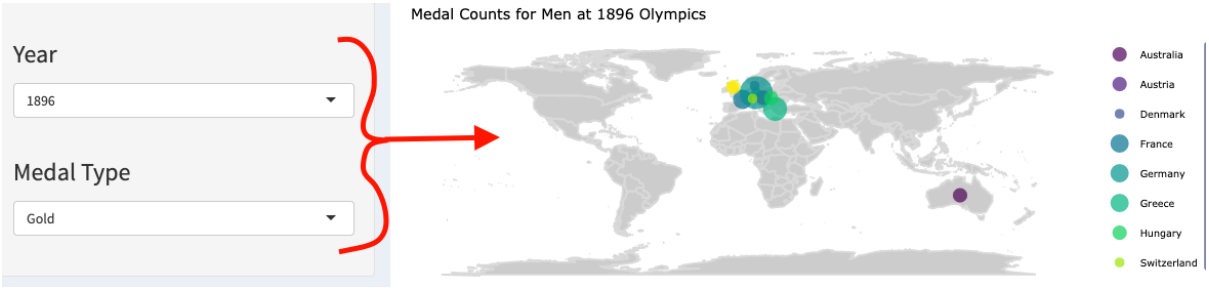


Figure 5: Figure representing the respective filters for the graph on dashboard

- Click the second menu option 'Overall Participation Trend' for selecting the next menu. Two graphs appear as shown below along with a 'Year' filter.

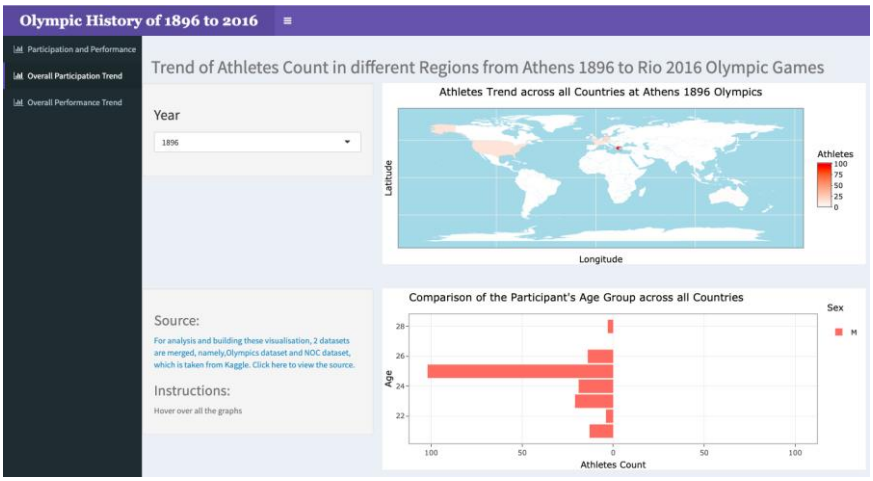
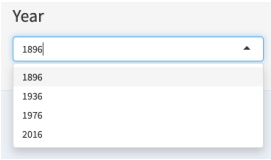


Figure 6: Figure showing the second menu option of the dashboard

The 'Year' drop down is shown below.



Clear instructions are provided on the dashboard for the user to navigate.

- Click the second menu option 'Overall Performance Trend' for selecting the next menu. This takes about a minute to load as this requires analysis of the whole dataset. Also, the dataset is huge, hence, it takes some time to run. Clear instructions are mentioned for the user to interpret the graphs.

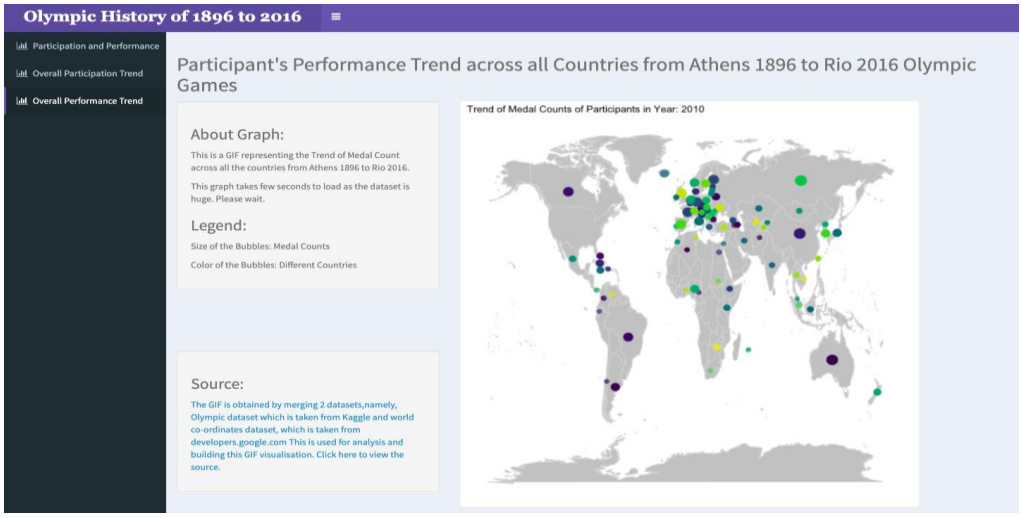


Figure 7: Figure showing the third menu option of the dashboard

V. CONCLUSION

1. Achievements:

- The findings of this report have been communicated with the help of shiny dashboard application in RStudio.
- The displayed visualisations are suitable for the intended audience as mentioned earlier.
- Three datasets have been merged to obtain the final dataset.
- The final datasets consist of 267,538 rows and 23 columns, which is really huge. Analysis have been done on this dataset.
- The GIF created was a really difficult task as it involves many aggregation steps.

2. Reflection:

One of the major information that can be drawn is that there were no women participants in 1896 Olympics. From then, the women's participation has been increasing dramatically. We also could draw the information that many regions of the world still haven't taken part in Olympics from the geographical representation. The countries include, Africa, Asia and some of Middle East and South America. The age group of the greatest number of participants can also be found.

3. What in hindsight that can be done differently to improve the result:

The implementation can be done using D3 instead of using shiny dashboard, which provides a platform for learning different tools and software. Also, the interactions could be produced in a different manner. The graphs in the second menu option could be more interactive, specifically mentioning, when the user clicks on a region on the map, a corresponding population pyramid graph should appear. Also, the GIF produced could be more interactive with few filter options.

VI. REFERENCES

AMR (2019). *Build your first web app dashboard using Shiny and R*. [online] freeCodeCamp.org News. Available at: https://www.freecodecamp.org/news/build-your-first-web-app-dashboard-using-shiny-and-r-ec433c9f3f6c/ [Accessed 11 Jun. 2019].

Georgetown Law Library (2019). *Guides: Olympics and International Sports Law Research Guide: Introduction*. [online] Guides.ll.georgetown.edu. Available at: https://guides.ll.georgetown.edu/lawolympics [Accessed 11 Jun. 2019].

Hills, L. (2019). *Women athletes are still put in second place at the Olympics – it's time to sprint towards equality*. [online] The Conversation. Available at: http://theconversation.com/women-athletes-are-still-put-in-second-place-at-the-olympics-its-time-to-sprint-towards-equality-63296 [Accessed 11 Jun. 2019].

International Olympic Committe (2016). *FACTSHEET WOMEN IN THE OLYMPIC MOVEMENT*. Switzerland: International Olympic Committe, p.5.

Kabacoff, R. (2019). *Quick-R: R Packages*. [online] Statmethods.net. Available at: https://www.statmethods.net/interface/packages.html [Accessed 11 Jun. 2019].

Mehta, R. (2019). *What is the difference between R and RStudio?*. [online] Quora. Available at: https://www.quora.com/What-is-the-difference-between-R-and-RStudio [Accessed 11 Jun. 2019].

Roberts, J., Headleand, C. and Ritsos, P. (2016). Sketching Designs Using the Five Design-Sheet Methodology. *IEEE Transactions on Visualization and Computer Graphics*, 22(1), pp.419-428.

Vries, A. and Meys, J. (2019). *The Benefits of Using R - dummies*. [online] dummies. Available at: https://www.dummies.com/programming/r/the-benefits-of-using-r/ [Accessed 11 Jun. 2019].

Appendix A

Five Design Sheet

1. Sheet-1:

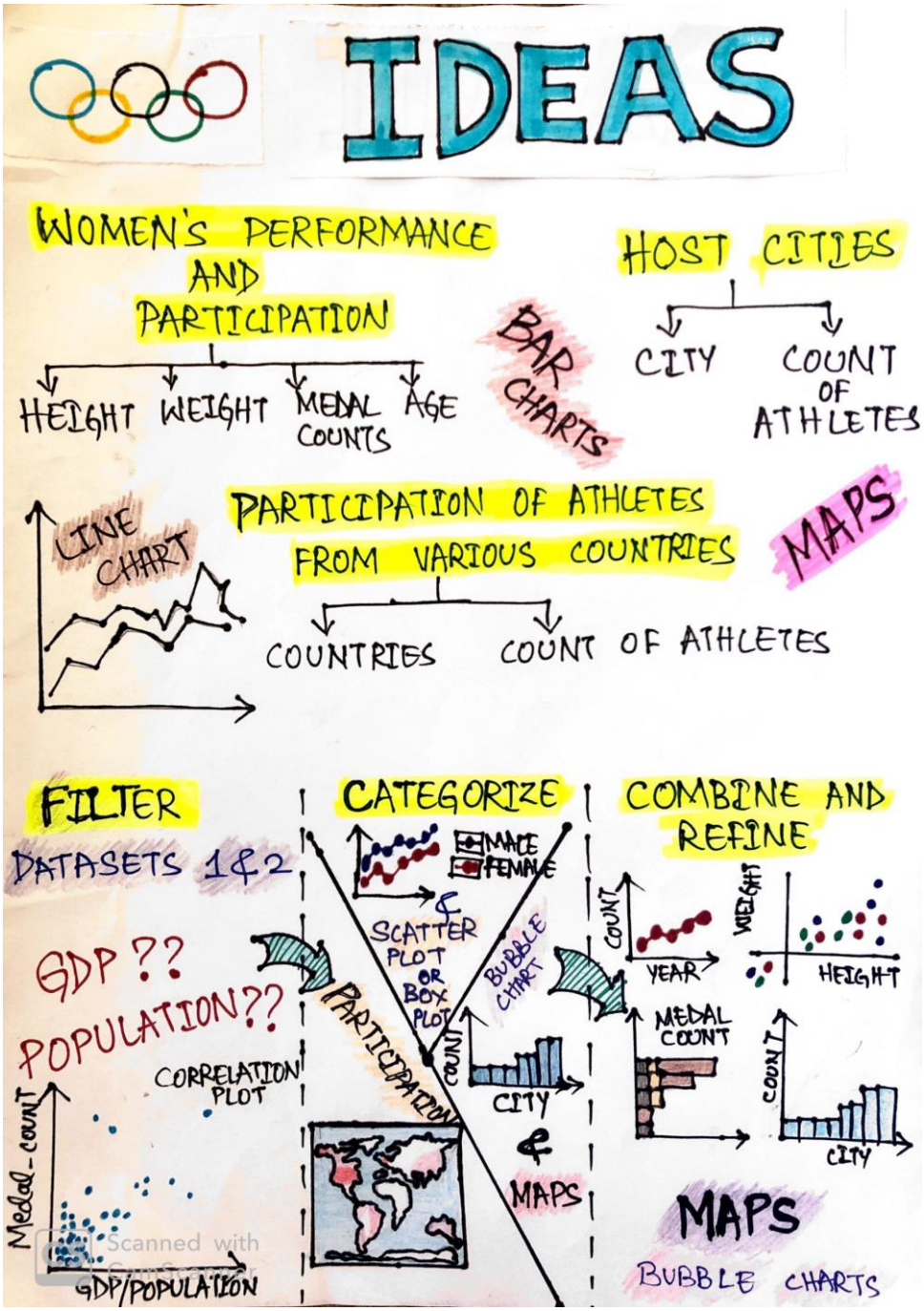


Figure 8: Figure showing the sheet-1 of Five Design Sheet

2. Sheet-2:

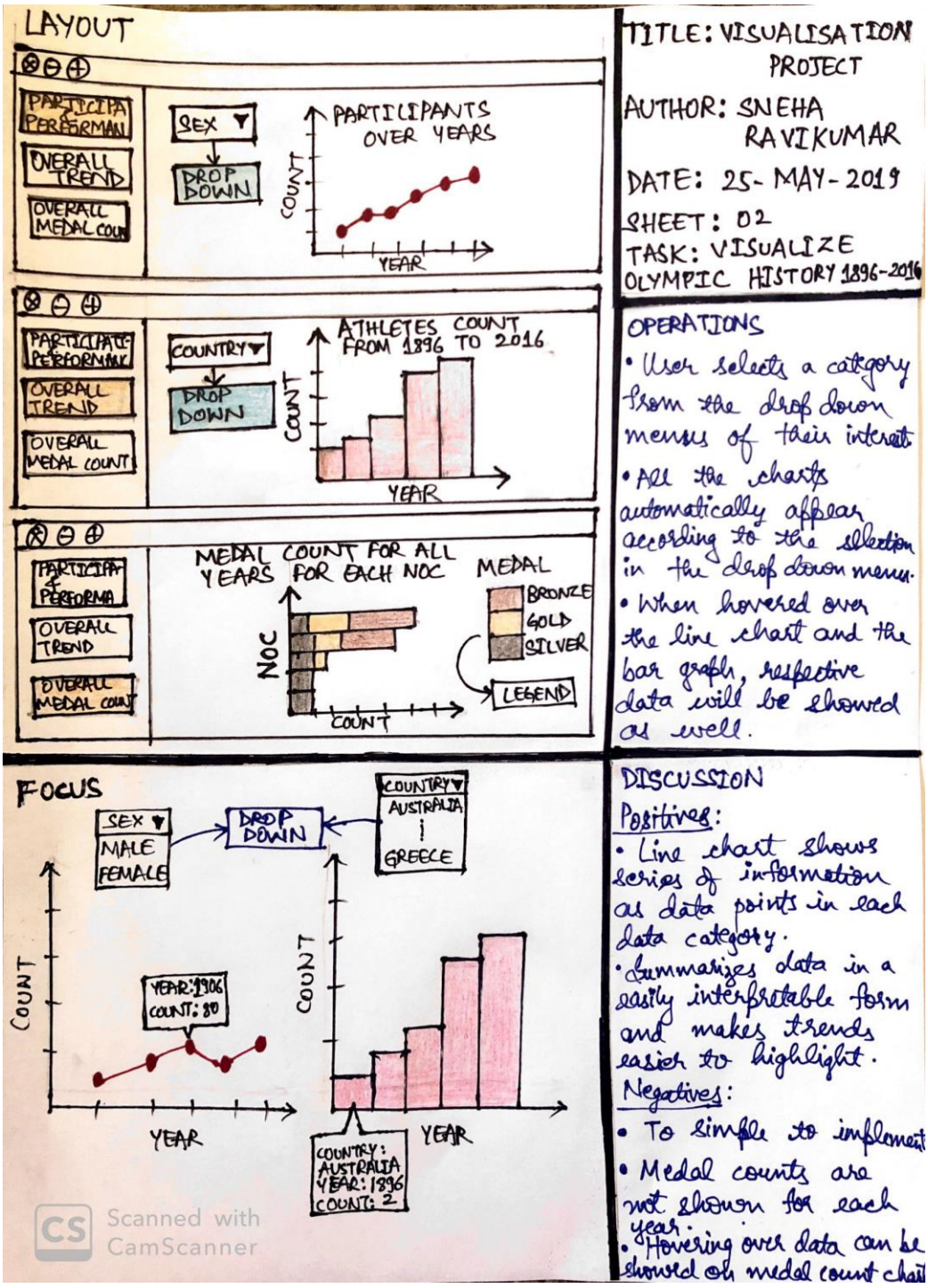


Figure 9: Figure showing the sheet-2 of Five Design Sheet

3. Sheet-3:

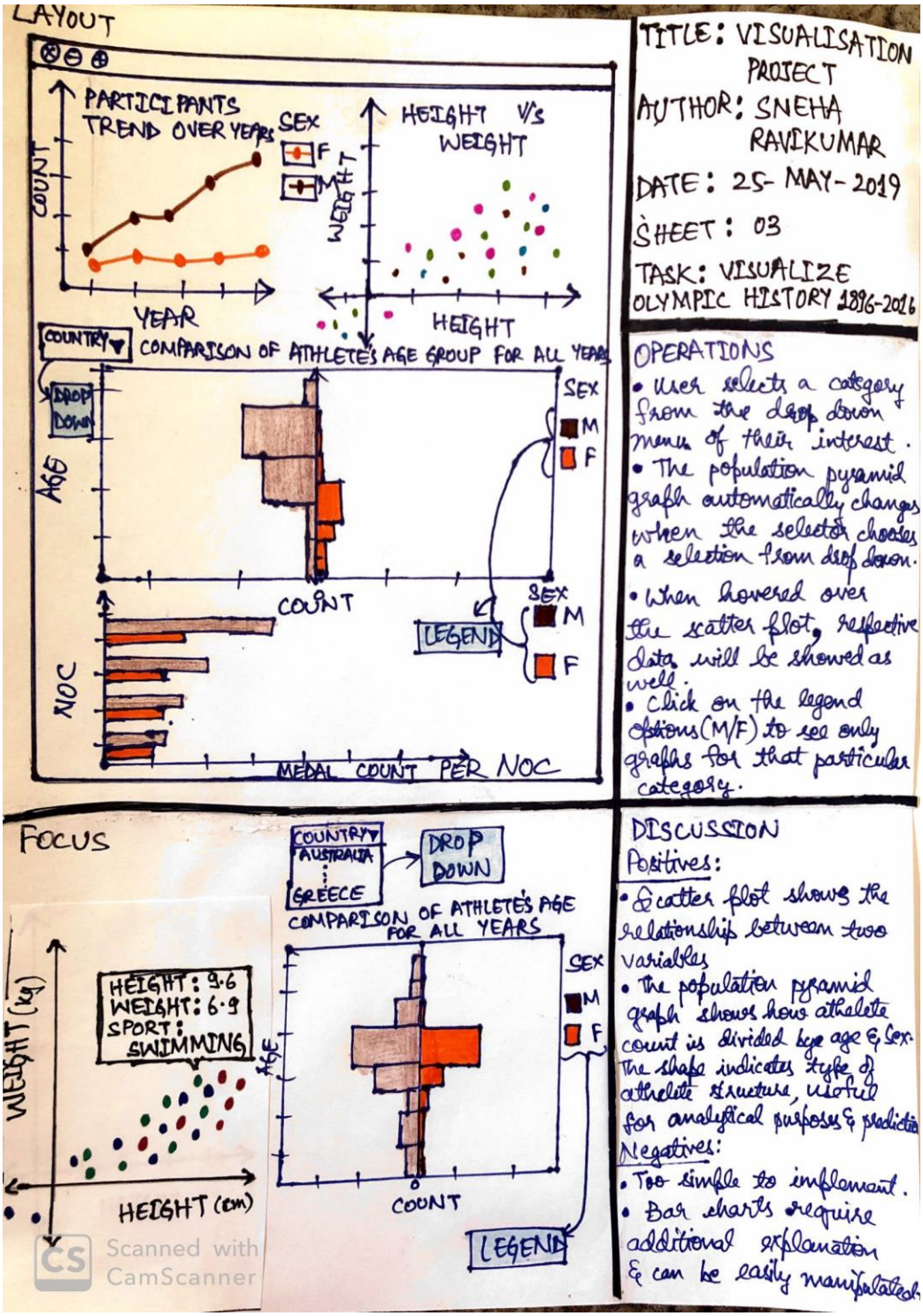


Figure 10: Figure showing the sheet-3 of Five Design Sheet

4. Sheet-4:

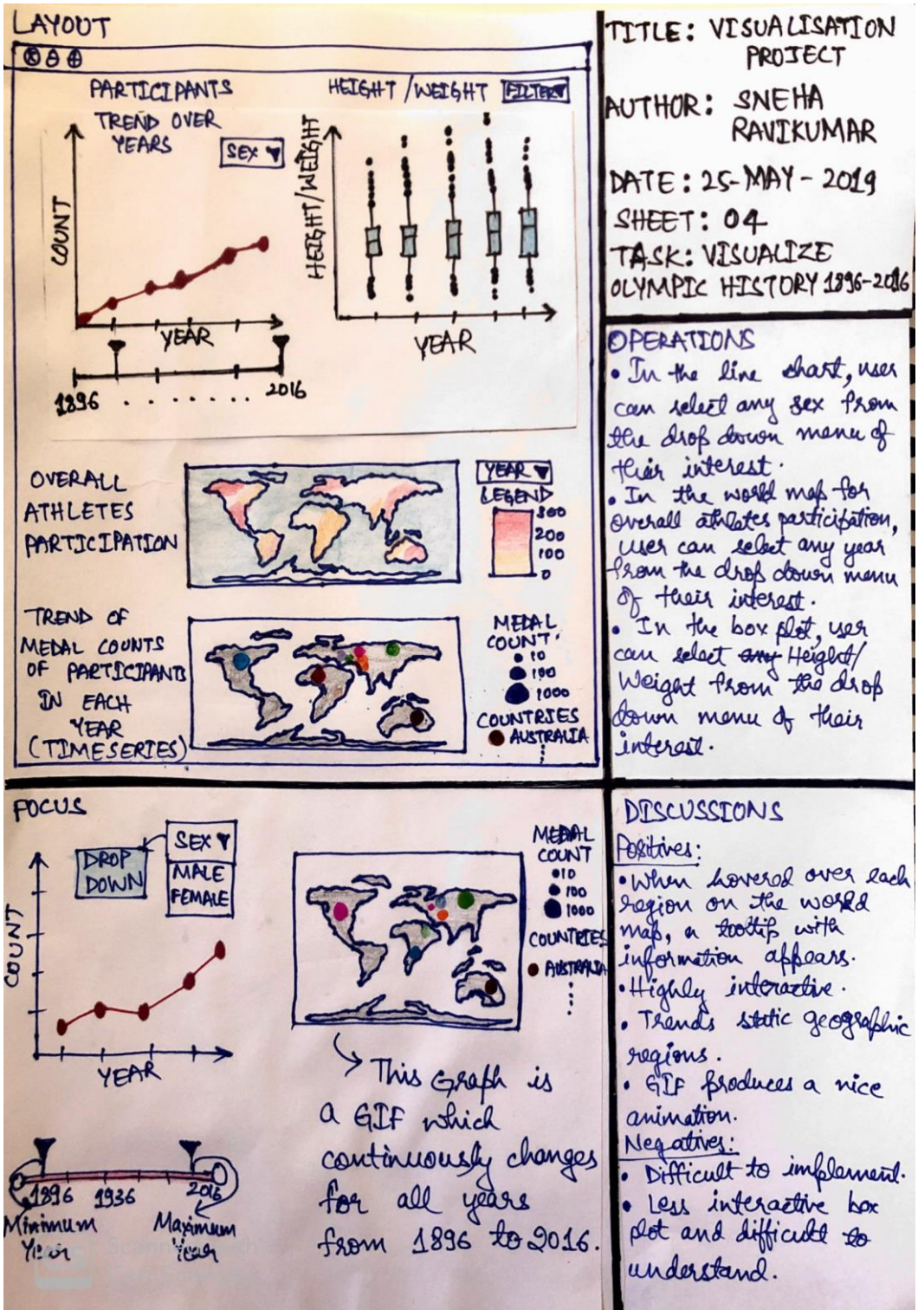


Figure 11: Figure showing the sheet-4 of Five Design Sheet

5. Sheet-5:

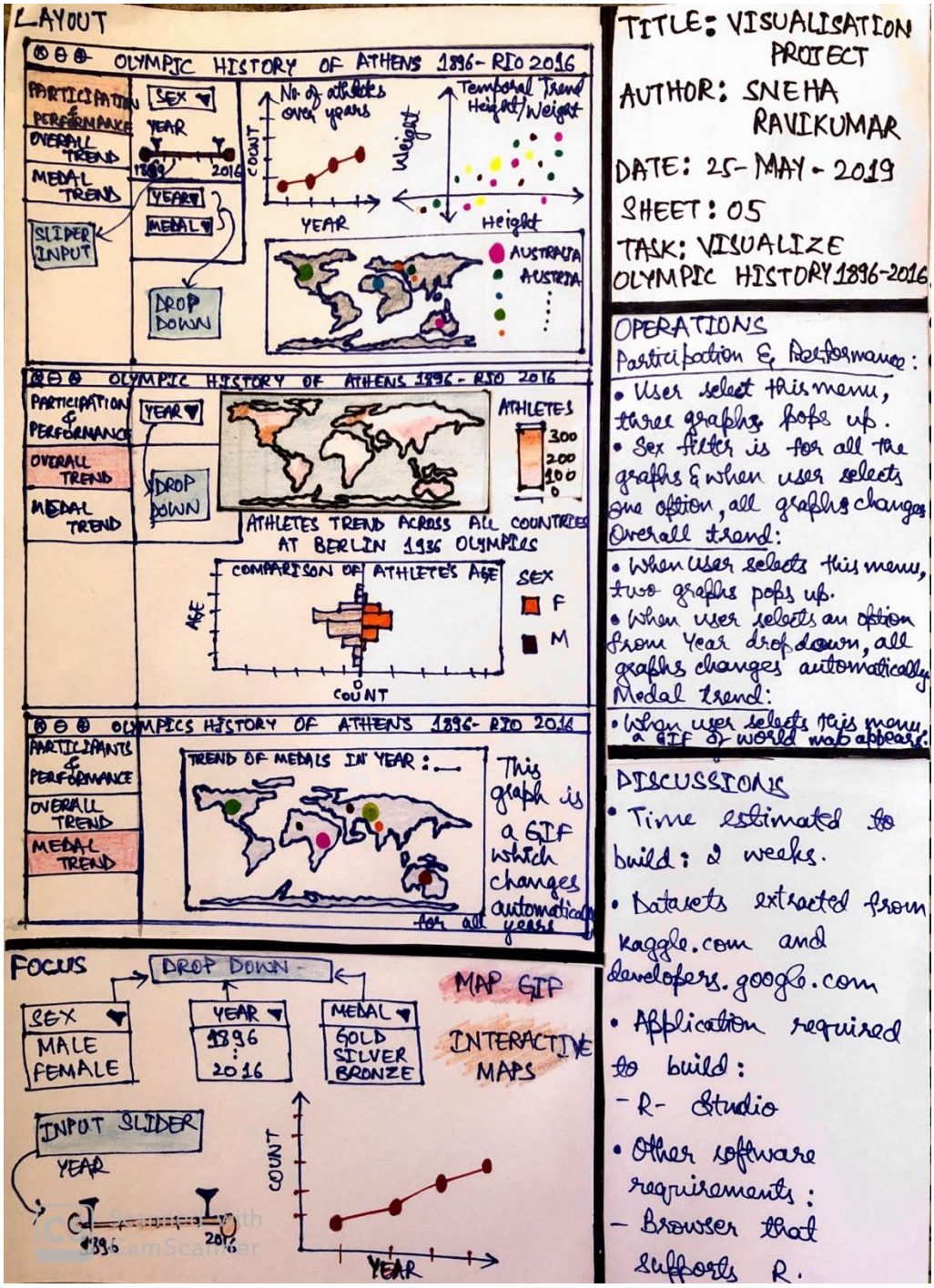


Figure 12: Figure showing the sheet-5 of Five Design Sheet