Assignment 1b

Developing intimacy with data and establishing editorial perspectives

1. Data Examination

In this section, we will examine the meaning and physicality of the Olympics Medalists Data. The data was given as a Microsoft Excel spreadsheet, consisting of 2 data sheets (MedalsData1 and MedalsData2). The first part will describe the meaning and physical properties of MedalsData1, and the second part will describe the meaning and physical properties of MedalsData2. Then, a comparison will be made between the two data sheets to get better ideas for transformation and exploration.

MedalsData1

Representativeness

The Olympics Medalist Data was provided by the lecturer for the assignment. By examining the data, we can infer that MedalsData1 is from the International Olympics Committee or an organization in responsible for recording Olympic Games; it consists of detailed information about winners including their performance (e.g. results in time units). The data would have been collected from the actual events, but however, the integrity and accuracy of the data are not guaranteed; the organization in responsible for the data is not named. It consists of the historical data from the time period of 1896 to 2012 (the whole time period of Olympic Games).

Phenomenon

In a short summary, the data presents actual results of athletes who won Gold, Silver, and Bronze medals in summer Olympic events from 1896 to 2012. The sheet presents columns of years, sports, events, athlete(s), country codes, country names, medal types, results, units for results, and results in seconds. The tone of the spreadsheet is purely exhibitory and reading; it displays the winners of Olympic events, and no emotional engagement is needed to discover insights from the data.

Physicality

MedalsData1 is a data sheet, consisting of 4094 rows and 10 columns. Excluding the first row describing the attributes (e.g. years, sports, events, athlete(s), country codes, country names, medal types, results, units for results, and results in seconds), it presents the data of 4093 items (the medalist/medalists for each event). The columns of Sport, Event, Athletes, CountryCode and CountryName consist of categorical (nominal) variables, and the column of Medal consists of a categorical (ordinal) variable. The column of Game consists of a quantitative (interval) variable, and the columns of Result and ResultsInSeconds consist of quantitative (ratio) variables. Sport, Event, Athletes, CountryCode, CountryName, Medal, and Unit are discrete variables, and Game, Result, and ResultInSeconds are continuous variables. The dataset if typically large as it contains the data from 1896 to 2012 in 4094 rows. The data ranges for qualitative measures of Result and ResultInSeconds are large, and this may influence the design of our visualization.

Developing Intimacy with Data

- The Game column presents years that the Olympic Games were held in the cyclic time period of 4 years. Despite the 4 years cycle, we can see that there are data from 1906, which represents the Intercalated Olympic Games: a series of International Olympic Games half-way between. However, the only incidence of the Intercalated Games was in 1906.
- We can also see that there are no data entry during World War 1 (1914-1918) and World War 2 (1939-1945)
- Events are grouped into Sports categories of Swimming, Athletics, Canoeing, and Swimming. Not all sports categories started from 1896. The category of Canoeing started in 1936.

- Events are divided into gender, and this is coherent to the Olympic Games events. In this form, the events can reflect the gender of corresponding athletes.
- Not all events have data for the whole time period of 1896 to 2012 (e.g. 10000m Men starts in 1912, 10000m Women starts in 1988, and 100m Breaststroke Men starts in 1968). This shows that there was a gender inequality in the implementation of events (e.g. 10000m), and some of the events were recognized as an official Olympic Game later (e.g. 100m Breaststroke).
- Not all events of Olympic Games were represented in the data. It would be interesting to see which games were represented and removed and possible reasons for this.
- The Athlete column contains first name and surname of winner(s). The cell consists of a single winner for individual games, and multiple winners for group events or relays.
- The columns of Country Code and Country Name show multiple encodings. While this may help some audience to understand the data, they are could be merged into a single column of country name of NOC (National Olympic Committee). Also, some countries that constituted as two separate countries (e.g. German Democratic Republic) later reflect as a single country.
- Some events had multiple winners (e.g. 2008 100m Backstroke Men has two winners of Bronze medal)
- The Result and Unit columns are not uniform. Some events recorded the result in M:S:DD, M:SS:DD, H:MM:SS, and #.DD. While this is easier to understand the results, it is harder to make relative comparisons between different results.
- The ResultInSeconds column computes results into seconds and thus deal with the above mentioned issue. However, results presented in this form make it harder for audiences to understand the data when it gets large (e.g. 17946).

MedalsData2

Representativeness

MedalsData2 presents lists of medalists at the Game of the Olympiad (Olympic Games) per edition, sport, discipline, gender, and event. It includes an explicit disclaimer on row 3 that the data is from the International Olympics Committee Research and Reference Service and the service endeavors to provide accurate and up-to-date information. However, it also says that the accuracy or completeness of the information is not guaranteed. By this, we can assume that the integrity and accuracy of the data are guaranteed to a certain extent, but should not be fully trusted. The spreadsheet consists of the historical data from 1920 to 2008.

Phenomenon

In a short summary, the data presents actual results of athletes who won Gold, Silver, and Bronze medals in summer Olympic events from 1920 to 2008 including ice hockey in 1920. The sheet presents columns of city, edition (year), sport, discipline, athlete, National Olympic Committee (country represented), gender, event, event gender, and medal. The tone of the spreadsheet is purely exhibitory and reading; it displays the winners of Olympic events, and no emotional engagement is needed to discover insights from the data.

Physicality

MedalsData2 is a data sheet, consisting of 26395 rows and 10 columns. Excluding the fifth row describing the attributes (e.g. city, edition, sport, discipline, athlete, National Olympic Committee, gender, event, event gender, and medal), it presents the data of 26394 items (the medalist/medalists for each event). The columns of City, Sport, Discipline, Athlete, NOC, Gender, and Event_gender consist of categorical (nominal) variables, and the column of Medal consists of a categorical (ordinal) variable. The column of Edition consists of a quantitative (interval) variable. City, Sport, Discipline, Athlete, NOC, Gender, Event_gender, and medal are discrete variables, and Edition is a continuous variable. The dataset is typically large, containing the data from 1920 to 2008 in 26395 rows. This may influence the design of our visualization.

Developing Intimacy with Data

- The City column presents the city that the Olympic Games were held from 1920 to 2008 in the cyclic time period of 4 years.
- The Edition column presents years that the Olympic Games were held in the cyclic time period of 4 years. However, we can see that there are no data entry during World War 2 (1939-1945)
- The Sport column groups disciplines into Sport categories (e.g. Aquatics... Wrestling). While this column is the highest categorical grouping of events, the range of Sport categories are not small.
- The Discipline column groups events into Discipline categories (e.g. Archery... Wrestling GRE-R). While this column groups events into smaller categories compared to the Sport column, some of the entries are identical (e.g. Taekwondo and Tug of War). A better categorical grouping might be needed.
- Not all disciplines have data for the whole time period of 1920 to 2008 (e.g. Taekwondo starts in 2000, and Archery has no data between 1921 and 1971. This shows that some of the events were recognized as an official Olympic Game later, and the recognition changes over time.
- Not all events of Olympic Games were represented in the data. It would be interesting to see which games were represented and removed and possible reasons for this.
- The Athlete column contains first name and surname of winner(s). The cell consists of a single winner for individual games, and multiple winners for group relays.
- The NOC column consists of values that were only sensible in certain time periods. For an example, EUN is the code for Unified Team, which only exists in 1992.
- The Gender column specifies the gender of the athletes who won the medals. It seems coherent given that there are only two possible values of Men and Women.
- The Event column presents the actual event of Olympic Games. Once again, some of the entries are identical to the Discipline and Sport columns (e.g. Tug of War).
- Some events had multiple winners (e.g. 2008 100m Backstroke Men has two winners of Bronze medal)
- The Event_gender column has 3 possible values of M, W, and X. The gender noted in this column represents gender specific events. Examining X allows us to know that X represents events that are mixed-gender, team, or relay events. These events belong in the sport categories of Badminton, Equestrian, Sailing, Shooting, Skating, and Tennis.

Comparison of MedalsData1 and MedalsData2

MedalsData1 and MedalsData2 both present lists of athletes who won Gold, Silver and Bronze medals in a) *summer Olympic Game events from 1896 to 2012*, and b) *summer Olympic events from 1920 to 2008 including ice hockey in 1920*. Both datasheets have an overlap in the time period (1920 to 2008), and the consistency between them can be proved by looking at a specific overlap of data entry (e.g. both datasheets list Gerald Blitz for 1920 100m Backstroke Men Bronze medalist).

The main differences between the two datasheets are a) time period, and b) MedalsData2 consists of more items (26394 events), while MedalsData1 consists of less items (4093 events). An interesting question on this may be how much of events in MedalsData1 does MedalsData2 cover. If MedalsData2 contains all of the events in MedalsData1, we could add the column attribute not present in MedalData2 from MedalData1 and use MedalsData2 for the main data analysis.

Another observation is that MedalsData1 only consists of events that decide winners based on their time records, which are depicted in the columns of Result, Unit, and ResultInSeconds. In contrast, MedalsData2 also consists of events that decide winners based on non-time based figures such as Boxing, Jumping, and Fencing. This means that MedalsData1 could be effectively used when trying to manipulate the channel of time into our analysis and visualization, while MedalsData2 provides information for more events.

Given that MedalsData2 does not provide information of results (time) for time-based events, MedalsData1 presents more attributes in this regard. This means that we could use MedalsData1 to compare the results (time) of athletes across different times of same events (e.g. medalists of 100m backstroke men in 1988 and 1992). A potential perspective derivable from this is how does the performance of medalists in 4000m Freestyle men change over time? Or how does the performance of a specific athlete change over time (if the athlete participated more than one game).

Given the lists of countries that the medalists represented, we could possibly look at the change of trends in medalists' geographical regions over time. A possible perspective is to compare the trend with the socioeconomic development of each countries, and the amount of support and training for athletes over time. The amount of support and training, and number of medalists can then be utilized as indices for socioeconomic developments of individual countries (e.g. Asian countries). Also, this could be used as an index for the participation and representation of developing countries.

Furthermore, given that both datasets consist of columns representing gender of the athletes. We could examine the participation of female in Olympic Games and relate this perspective to the gender inequality. Yet, the limitation is that we are not sure whether the data we have is an accurate and comprehensive representation of all Olympic events during the time period specified. Also, we events are already gender-specific, thus we cannot really observe the discrimination in participation in this. However, MedalsData2 presents medalists of mixed gender events across time. We can examine the representation and participation of female in mixed-gender or group Olympic Games over time.

2. Data Transformation What are possible ways to transform the data to clean or enhance it? What other data should be sourced to consolidate the data?

MedalsData1

- We should check whether MedalsData1 correctly depicts the information with full resolution. If not, we should think about in which ways the data has been filtered or selected.
- The columns of CountryCode and CountryName show a case of multiple encodings. Based on who the viewer is, this might not be necessary. We can simply use country codes to represent national representations of athletes, and delete the Country Name column.
- The columns of CountryCode and CountryName needs to be cleaned to their current names (e.g. German Democratic Republic should be changed to Germany, and KIT in 1920 should be changed to ITA to represent Italy).
- The CountryCode column consists of a blank cell. The corresponding CountryName column consists of a cell entry called '#N/A'. A closer examination shows that this item entry is for disqualified USA members in 2012 in 4*100m Relay Men. However the corresponding Medal column shows Silver. We have to filter this item and check if this is true. We have to clean the data or delete the item as disqualified members would not have won any medals.
- We can extract gender information from the Event column to represent gender for each medalists. While the gender appended events correctly describe the actual Olympic Games, creating a separate column of gender will strengthen our analysis (e.g. the correlation between performance and gender). Also, by doing this, we can sort the data into gender and this will allow us to compare performance among female athletes and male athletes.
- Deleting the embedded ender from event description can help us to compare performance between female and male athletes.

- Some of the names in the Athlete column consists of symbols (e.g.?). This needs to be cleaned and presented in a consistent format of (Firstname Surname).
- Some of the entries in the Result and ResultInSeconds columns have no result ('No Result). For example, the silver and bronze medalists for 1904 Single Sculls Men have no result recorded. However, seeing that there is a data available for the cold medalist of 1904 Single Sculls Men, the results could be obtained. Obtaining the actual values for those no result entries could significantly improve the accuracy and quality of the dataset.
- The datasheet presents results in seconds. While this is beneficial to compare the results of short events, it is hard to compare the results between long events (e.g. 17946 seconds) and get a real sense of the entry. Thus, adding two more columns to depict the result in minutes and hours could help the readers to understand events with longer timeframes better.
- Adding a list of cities and countries in which the Olympic Games were held in during the time period may help the readers to better picture the data. Also, this will allow us to examine whether athletes tend to perform better in familiar regions or not (e.g. Asian athletes could perform better in games held in Asian countries).
- The column name Game may confuse some of the readers. We could change the name to Game (Year).
- We could obtain data for athletes' age, height, weight at the participation to examine the correlation between physical properties and medal winning (performance).
- We could obtain statistics for the economic performance of countries and amount of support and training for athletes or sports over time. This will allow us to examine the correlation between the performance and economic indices (development). Ideally, the wealthier the country is, the more support the country gives for athletes.
- We could obtain data for the number of medals athletes won and evaluate individual and country performances.

MedalsData2

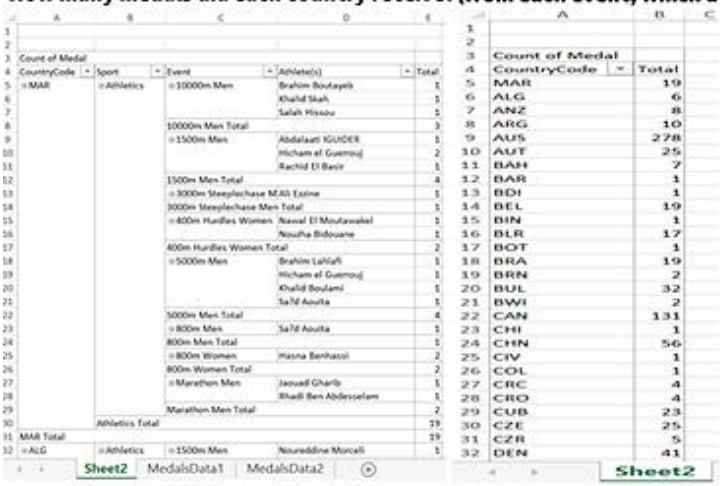
- We should check whether MedalsData2 correctly depicts the information with full resolution. If not, we should think about in which ways the data has been filtered or selected.
- We could add a Country column before the City column and this will give a higher level location information of the Olympic Games. This will allow regional and ethnic comparisons.
- The data is currently updated to 2012 London Olympics. We can search the relevant data from the Official Website of the Olympic Movement and update the data sheet. We could also update data before 1920.
- The National Olympic Committed column containing country codes of athletes should be filtered and cleaned to their latest codes as in MedalsData1. Some codes are no longer recognizable on the official Olympics website e.g. GBR. These should be re-entered with more current codes e.g. GER.
- We could further categorize the Sport and Discipline columns into smaller groups. This will allow us to comprehend the data with a broader perspective in less time.
- The Event_gender column consists of 'X' values, which are very misleading. We could replace this with 'G' with a description that this represents 'Group', or replace with 'Group'.
- Some entries in the Event column are not specific enough (e.g. 100m, 48kg, K-1 10000m). These are hard to comprehend unless you look into discipline and sport. Specifying these entries with more detailed and coherent formats will lessen the time required to comprehend the data.
- The Athlete column present names of medalists in SURNAME, Firstname. There might be cases we make cross-sheet comparisons, so we could present names in the consistent format of Firstname Surname.

- We could add another column with results (e.g. distance for Jumping, time for time-recording events such as swimming). This will allow us to evaluate athletes' performances over time, and compare two or more athletes participated in same events.
- The column name Edition may confuse some of the readers. We could change the name to Year or Game (Year).
- We could obtain data for athletes' age, height, weight at the participation to examine the correlation between physical properties and medal winning (performance).
- We could obtain statistics for the economic performance of countries and amount of support and training for athletes or sports over time. This will allow us to examine the correlation between the performance and economic indices (development). Ideally, the wealthier the country is, the more support the country gives for athletes.
- We could obtain data for the number of medals athletes won and evaluate individual and country performances.
- We could combine MedalsData1 and MedalsData2 to create a more comprehensive datasheet.

3. Data Exploration

In order to explore the data, I have used Microsoft Excel (Pivot Table, Data Analysis, and Charts), and Tableau to better understand the data and search potential ideas for analysis. Both statistical analysis and visualisation idea sketches were conducted to describe MedalsData1 and MedalsData2 and develop editorial perspectives.

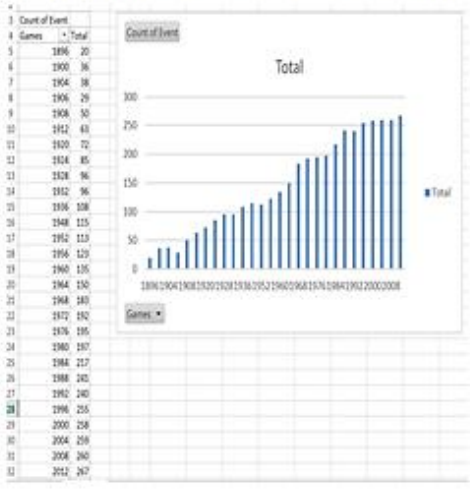
How many medals did each country receive?(from each event, which athlete?)



Who has the world record for time-based events?

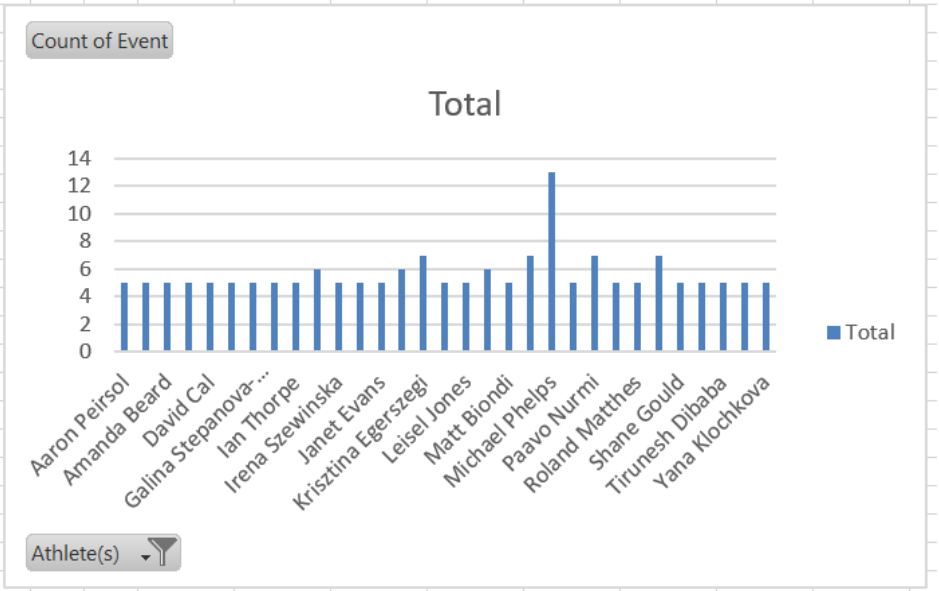
3	Min of ResultInSeconc		
4	Event	Athlete(s)	Total
5	10000m Men	Albin Stenroos	1941.8
6	10000m Men Total		1941.8
7	10000m Women	Lynn Jennings	1879.89
8	10000m Women Total		1879.89
9	100m Backstroke M	Herbert Haresnape	87
10	100m Backstroke Men Total		87
11	100m Backstroke W	Aileen Riggin	88.2
12	100m Backstroke Women Total		88.2
13	100m Breaststroke	Nikolay Pankin	68
14		Vladimir Kosinsky	68
15	100m Breaststroke Men Total		68
16	100m Breaststroke	Sharon Wichman	76.1
17	100m Breaststroke Women Total		76.1
18	100m Butterfly Men	Ross Wales	57.2
19	100m Butterfly Men Total		57.2
20	100m Butterfly Wor	Mary Sears	74.4
21	100m Butterfly Women Total		74.4
22	100m Freestyle Mer	Otto Herschmann	82.8
23	100m Freestyle Men Total		82.8
24	100m Freestyle Wor	Jennie Fletcher	87
25	100m Freestyle Women Total		87
26	100m Hurdles Wom	Kim Turner	13.06
27		Mich?le Chardonnet	13.06
28	100m Hurdles Women Total		13.06
29	100m Men	Alajos Szokolyi	12.6
30		Francis Lane	12.6
31	100m Men Total		12.6
32	100m Women	Ethel Smith	12.3
33		Fanny Rosenfeld	12.3

Did the number of events increase over time?

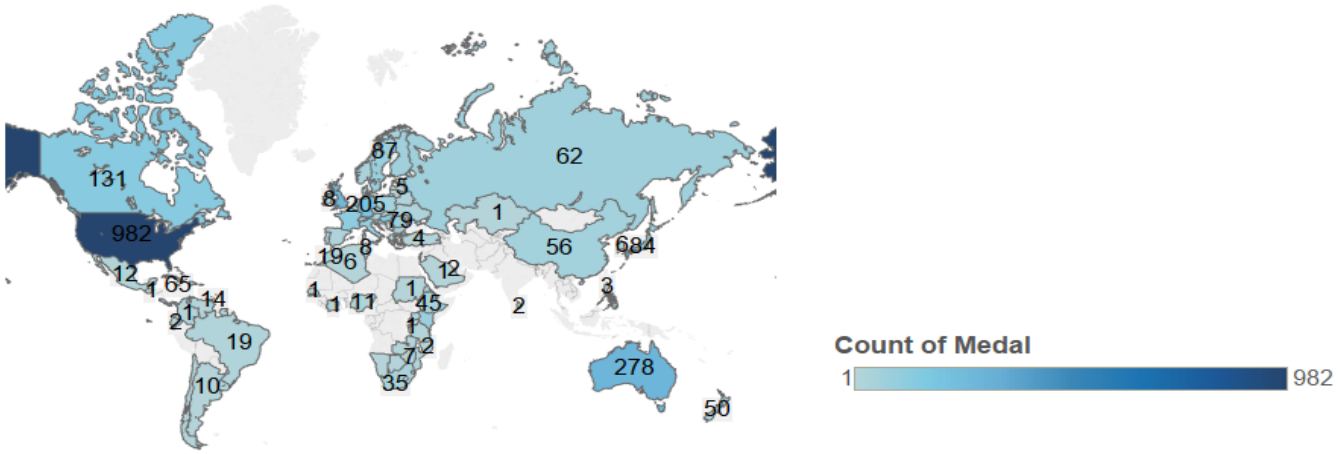


Who are the top medalist since the start of Olympics?

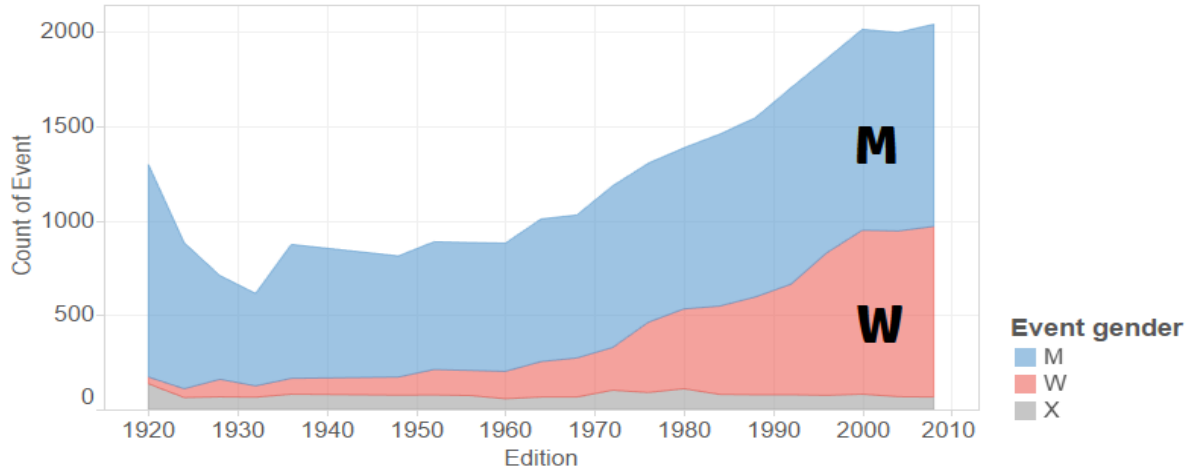
3	Count of Event	
4	Athlete(s)	Total
5	Aaron Peirsol	5
6	Aleksandr Popov	5
7	Amanda Beard	5
8	Dagmar Hase	5
9	David Cal	5
10	Emil Z?topek	5
11	Galina Stepanova-Prozumenshchykc	5
12	Gary Hall Jr.	5
13	Ian Thorpe	5
14	Inge de Bruijn	6
15	Irena Szewinska	5
16	James Lightbody	5
17	Janet Evans	5
18	Knut Holmann	6
19	Krisztina Egerszegi	7
20	Laszlo Cseh	5
21	Leisel Jones	5
22	Mark Spitz	6
23	Matt Biondi	5
24	Merlene Ottey-Page	7
25	Michael Phelps	13



Total Number of Medals NOC Received



Change in Number of Events and Event Gender



- Some of the potential ideas and perspectives are also described in Data Examination and Data Transformation.
- Some examples from Data Exploration is shown above. A lot of exploration can be conducted in this part (20~30 explorations), and these are purposely not included in this report; this would produce a document of a large page number without any substantial meaning. More examples and files for statistical analysis and visualisations in Excel and Tableau can be additionally submitted if necessary.

4. Editorial Perspectives

Some interesting angles of analysis that we can take with the data provided. Some ideas are already mentioned in Data Examination, Transformation, and Exploration.

1) Change in the overall number of Olympic events, disciplines, sports / Change in the number of events for male, female, and group over time

Angle

What is the relationship between two categorical variables of events (or disciplines, or sports) and event gender? How does this change over time (year or edition)?

Framing

Parameters for the time periods of 1920 to 2008 from MedalsData2 or 1896 to 2012 if additional data for gender was consolidated into MedalsData1, and the gender of events (Male, Female, and Group/Mixed).

Focus

Colour Hues can be applied on the gender of events (Male, Female, and Group/Mixed) to draw a particular attention and differentiate event gender, and the visualisation will be depicted in a continuous chart to see the change in the overall trend (e.g. time line chart with segmented areas). This will allow us to examine 1) the change in the total number of events over time, 2) the change in the number of events for female over time, 3) the participation of female athletes in Olympic Games, and 4) the change in the trend and prevalence of gender inequality over time using the Olympic Games data as an index; this will require more data such as socio-political movements against gender inequality, the right of women, or surveys.

2) A comparison of total medals earned by each country

Angle

What is the total number of medals earned by each country over time of all Olympic Games? What does the result indicate?

Framing

Parameters for all countries that have earned at least one Gold, Silver, or Bronze Medal in Olympic Games, and the total number of medals each country earned. If this perspective is further developed to see how the economic wealth of a country affects the performance of Olympic Games athletes, we will also need a comparison of the current or historical GDP among countries. However this will need economic inferences of each country to be consolidated with our Olympic data.

Focus

Colour Saturation can be applied in a tree map or geographic filled map of the world, in which the relative saturation of colours on each country represent the number of medals earned (e.g. Countries with higher numbers of medals such as USA and China will appear with increased saturations). We could include labels for each country's name and current or historical GDP, or economic growths.

3) The Olympic Games records for time-based events

Angle

Who has the fastest record for time-based events (events that decide winner by the shortest time to complete the event)? Are male athletes comparatively advantaged to female athletes?

Framing

The minimum of ResultInSeconds/Minutes/Hours across Game (year) from MedalsData1 will be obtained for all events with corresponding athletes names. We can extract the embedded gender element from events, and create separate lists for Men and Women. This will present the Olympic Games records for time-based events for 1) both men and women, 2) men, and 3) women.

Focus

We should put equal weights on event, athlete and record since the audience must understand, who the athlete is, and what record the athlete has in which event. For this, providing a filtered table will be most efficient. We could selectively filter in the events or sports of our interest.
If we want to compare the result with gender, we should include gender elements in the table (Male and Female) and record two results for same events. For this, providing a side-by-side bar chart will be more efficient to compare male and female to more emphasis on gender.

4) Top 20 medallists for Olympic Games since the first Olympic Game in1896

Angle

Which athletes earned most medals since the first Olympic Game in 1896? Generate a list of top 20 athletes who scored most medals since the first Olympic Game.

Framing

Parameters will involve the top 20 athletes who earned the most medals, listed in the order of most medals, and the number of medals they earned.

Focus

More emphasis should be given to medallists who earned higher numbers of medals. Therefore, the list of athletes will be ordered by the decreasing number of total medals earned by individual athletes, and we will filter out athletes who are not in top 20. It will be helpful to include which events, sports, or disciplines, the athletes participated in. However, the name of the athletes should gain most emphasis in the visualisation.

5) Newly Emerging Winners (Countries)

Angle

Which countries are emerging as new winners in the Olympic Games? Which are the countries that did not earn much medals in the past but not anymore?

Framing

Parameters will involve the list of countries with the total number of medals each country gained in each Olympic Game (year). A trend will be examined.

Focus

More complicated analysis will be required. A regression analysis on the total number of medals gained in each Olympic Game (year) and Game (Year)/edition will be conducted, by each country. This will give how rapidly each country developed its performance (the number of medallists per Olympic Games).
More emphasis will be given to the newly emerging countries in this editorial (e.g. South Korea, China, and Australia). Different colour hues can be used to depict different countries on a geographical map with labels indicating country names and rates of increase in the number of medals in a certain time period (e.g. last 30 years).

6) Participation vs. Medallists

Angle

Is there are correlation between the number of athletes participated in the Olympic Games and the number of athletes who won medals? Are countries who send more athletes to the Olympic Games likely to earn more medals?

Framing

Additional information on the number of athletes participated from each country per year will be added to our current dataset. This is required to examine the participation rate vs. performance. Parameters will be the number of athletes participated from each country and the number of medallists. We do not have to examine the whole time period. We could observe last 5 Olympic Games (20 years).

Focus

Additional Focus on countries that has either 1) high rates of participation (the number of athletes participated) or 2) high rates of medallists, or both. We could use a scatter plot with markers for individual countries, and

markers for countries with high rates of participation or performance (medallists) can gain more magnitude (size). Also, we should use separate colour hues for participation and performance to clearly differentiate them.

7) The development of a particular athlete and his/her performance over time.

Angle

How does the performance of a particular athlete change over time (if the athlete participated in more than one game (e.g. Michael Phelps)?

Framing

Parameters for the years of participation by the particular athlete, and performance review (e.g. results in seconds, which type of medals (Gold, Silver, and Bronze)?).

Focus

Emphasis should be on the change of the athlete's performance over time. Therefore, using a chart with time components (timeline or histogram) is recommended. We could conduct this with a single player or multiple players. With multiple players, we should emphasise the athletes who shows the most rapid development using different colour saturations or different markers.