Olympic Data Analysis Project

Tanya Riemann / Alyson Nicol

December 2, 2020

Purpose/Hypothesis:

The purpose of our analysis was to answer some, if not all, of the following questions:

1) Is there a correlation between being a host country and increasing a country's medal count?
2) How many athletes participate per country and what is the correlation to increased medal count.
3) Is the medal count per country closely correlated to GDP and/or Population?
4) How many athletes participate in both summer and winter games and are there any trends around the sports that they participate in? Are there any stand-outs?

We hypothesized the following:

1) We expect that host countries generally have more participants, therefore increasing the likelihood of making it to the podium.
2) We expect that a larger number of participants, on average, does lead to more medals.
3) We expect that countries with a higher GDP are more likely to get to the podium due to more funding available for training.
4) We expect that there are a number of athletes that participate in both games and that there is a strong correlation between cycling and speed skating for sure (e.g. Clara Hughes, Canada).

Based on the theme of the above questions and hypotheses, the overriding main question that our analysis is looking to answer is:

What are some of the major factors contributing to a country's success at the Olympics?

Datasets:

There were a number of datasets needed in order to pull together all of the information required to answer the questions above.

1) Olympic athlete dataset contained athlete and event-level data for every Olympic event. It included the following:
   (Source: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results)

   a) Athlete ID
   b) Athlete Name
   c) NOC Region Code (National Olympic Committee)
   d) Team name
   e) Athlete height, weight
   f) Athlete gender and age
   g) The Olympic year
   h) The season
   i) Host City
   j) Sport
   k) Event
   l) Medal

2) NOC Regions dataset provided the actual region name associated with the NOC Region code which was included on the athlete dataset
   (Source: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results)

3) GDP data by country dataset showing Income Category, by year starting from 1960
   (Source: http://api.worldbank.org/v2/en/indicator/NY.GDP.MKTP.CD?downloadformat=excel)

4) World Cities dataset – provided the corresponding country to major world cities to allow us to pull in the country for the Olympic host cities
   (Source: https://simplemaps.com/static/data/world-cities/basic/simplemaps_worldcities_basicv1.71.zip

5) Country codes dataset - this provided the ISO code by country, which was used to cross-reference to the NOC regions. The data was copied directly from the webpage and into a blank Excel file.
   (Source: https://www.iban.com/country-codes)

Process – Initial Clean Up and Formatting of Data:

Our team split the analysis using different approaches. Tanya was focusing on getting the data into MySQL Workbench, while Alyson did most of her analysis in Excel. This was helpful in cross-checking the results and ensuring that the resulting analysis was accurate.

NOTE: Tanya experienced some challenges with the last column in some of the data imported into MySQL Workbench through the Load Data Infile process would add extra characters to the text string. Tanya worked with Andrew to find a solution however in the end the solution only worked some of the time. Tanya chose to add a dummy column to the end of each data set which was causing issues in order to be able to continue forward with the final project analysis. This will be the reason there is an ABCD column in the datasets at times. This extra column did not negatively affect the analysis but it was an effective workaround due to the tight timelines.

1) **Olympic Athlete Dataset**

    a) In order to bring down the size of the dataset and to match it to the other datasets we were working with (e.g. years), the first step was to remove unneeded data.

    b) Since the GDP data started at 1960, we decided to remove all Olympic data prior to that year. 56 years of Olympic data was deemed to be enough to answer the questions we were looking at.

    c) We also knew that we weren't answering any questions related to the athlete's size, so we deleted the columns related to that. We probably could have also removed gender and age, but left those in, just in case our analysis lead us down a different path.

    d) For SQL purposes, some "NA" cells were changed to "NULL"

    e) Data columns were reviewed by filters to ensure that the information was appropriate for that column. In the event data did not seem correct, it was reviewed and updated as required.

2) **GDP Dataset:**

    a) The GDP dataset had a few tabs in it.

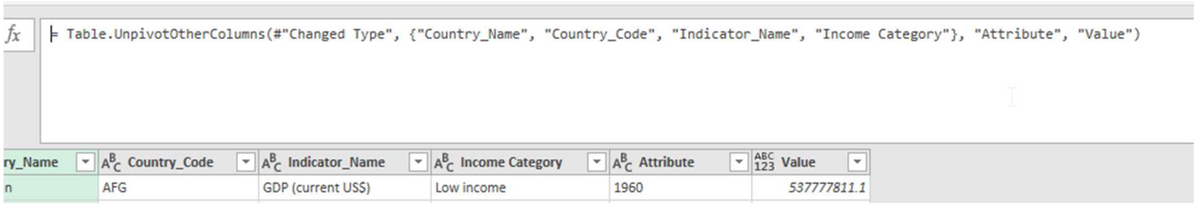
        i) The main tab had the country name and country code as rows and then the year was across the top, with the relevant GDP figure by country under each year's column.

        ii) A different tab had the county name, country code and Income Category

        iii) We made the decision that we wanted to analyze by Income Category and not straight GDP, so we had to pull in the Income Category into the first sheet, using a vlookup

        iv) Following that, in order to make the data more useful, the table years needed to be unpivoted in order to have them flow as rows instead of columns

v) Using Excel power query, we used the Unpivot option (shown below) and then saved the new file as an Excel workbook **GDP Data.xlsx**.



```
= Table.UnpivotOtherColumns(#"Changed Type", {"Country_Name", "Country_Code", "Indicator_Name", "Income Category"}, "Attribute", "Value")
```

ry_Name	Country_Code	Indicator_Name	Income Category	Attribute	Value
n	AFG	GDP (current US$)	Low income	1960	537777811.1

vi) A bit later in the analysis it was determined that for certain countries, the income category was not populated for all years. There was some manual manipulation required to copy the income category for those countries to fill the missing rows (e.g. Germany, Russia).

3) NOC Regions Dataset:

a) Because it was found that the NOC Region codes did not match up to the ISO country codes, we had to analyze the differences.

i) First we pulled in the country codes and country names for those that did match using a vlookup in Excel

ii) Based on this, we were able to identify those that did not match based on those that did not have a corresponding country code pulled in.

iii) Through manual review and update, a Country Code column was added to the NOC Regions dataset by comparing the country name to the NOC Region name.

(1) For example Russia had three different NOC Region codes (RUS, EUN, URS) in the NOC Region file. We made the decision to map these all to RUS, so that when cross-referencing with the other datasets, the proper data would be pulled in (e.g. the GDP data file had RUS)

4) World Cities Dataset:

a) This data was used to pull in the relevant country for the host cities for each Olympics

b) The athlete dataset didn't have anything other than city name for host city, so we couldn't cross-reference to the country code listing

c) An initial attempt was made to cross-reference using a vlookup from the athlete dataset to the world cities dataset based on city name to pull in the country name.

d) The issues found were:

i) Host city names were not in the world cities dataset

ii) Host city names didn't match the name in the world cities dataset

iii) Country name from the world cities dataset didn't match the country name from the Country Codes dataset

e) As a result of the issues, there was some manual manipulation that needed to occur in order to ensure that all host cities had a country properly available to pull in to the analysis

i) Those that did not have a match were identified and analyzed to determine whether it was just a spelling (e.g. Athina vs Athens) or whether it just wasn't on the world cities dataset (likely because this dataset only included major cities)

(1) Once the above was determined, the world-cities dataset was updated to add the missing cities / countries or to amend the city_ascii column to the name that matched the Olympic dataset

(2) Further on in the analysis, country names were also modified to ensure the country name matched the country name per the country code dataset

5) Country Code Dataset:

a) There was no manual manipulation required on this dataset.

b) It was used to pull in for comparison to the NOC Regions and the world cities

Process – Subsequent Analysis of Data:

Now that we had all of the data in the format required and most data issues identified and corrected, we were able to start actually analyzing the data in order to answer our questions.

This analysis required merging and summarizing the data in a number of different ways, which may be best outlined by the questions that we were trying to obtain answers for or confirmation of our hypothesis on. As noted, we used different methods for analysis and then compared our results for accuracy.

For Questions 1 and 3, Alyson used Excel and Power Query within Excel to summarize and manipulate the data before pulling it into Tableau.

For Questions 2 and 4, Tanya used MySQL to manipulate and summarize the data before pulling it into Tableau.

Each of us also did some analysis on each other's questions using our respective tools and we compared results to make sure we were on the right path.

We will outline the analysis that led to the ultimate results for each of the questions below.

Question 1) Impact of Being the Host Country on Olympic Success

We hypothesized that being a host of the Olympics would tend to lead to greater participation by the host country and also improve their chances of making the podium.

In order to answer this question, we needed to determine the following:

- All host countries for the summer or winter Olympics for the period being analyzed, so that we could isolate these countries from all other countries participating
- the countries that were the host for each specific Olympic event
- the number of medals each country won for each specific Olympic games
- the number of entrants each country had in each specific Olympic games

The datasets required for this question included:

- Athlete dataset
- NOC Regions file
- Table mapping host city to country (World Cities)

<u>Steps required to obtain the answer:</u>

Note that this answer and analysis was done in Excel and Tableau. In Excel, instead of using standard vlookup and other functions, Power Query was used.

There are a number of layers to this analysis, resulting in multiple queries and tabs which contribute to the ultimate set of data used to answer the question. It took a bit of analysis to formulate a clear path, so in hindsight, some of the analysis could have been done more efficiently.

The relevant tabs/queries in the **Olympic Data Analysis.xlsx** file are listed below and the detailed steps taken in each query are included in the Appendices noted. Below I will give a brief summary of the analysis that was completed. Note that we have also provided the above Excel file with the queries removed, as they will likely cause issues when you open the file.

- the **Olympic Mega Table** query (Appendix 1)
- the **Athlete Total Counts** query (Appendix 2)
- the **Medal Count by Country** query (Appendix 3)
- the **Summary Host Countries** query (Appendix 4)
- the **Host Country Data** query (Appendix 5)

1. The Olympic Mega Table was used to format the original athlete dataset and then to pull in accurate country data for the NOC regions and the host countries. The details of the query used to complete this are included in Appendix 1.

2. The Athlete Total Counts table was created to summarize the data by the count of entrants by country for each Olympic games. The Olympic mega table was duplicated and then the rows were grouped by country by games to count the total athlete ID's (basically counting all rows in the grouping). Refer to Appendix 2 for the detailed steps.

3. The Medal Count by Country table was created to summarize the medal count by country for each Olympic games. This table started by referencing the Olympic mega table and then it goes through a series of steps (outlined in Appendix 3) to summarize the data in order to get the medal counts, merge in the GDP (and population) data and also the total entrant counts from the Athlete Total Counts table.

   This table is also used for the GDP question in Question 3 below, but for this particular question, this table is used as the initial source for creating the Summary Host Countries Table.

4. The Summary Host Table was created to provide a lookup of all countries that ever hosted an Olympic games for each season. It basically takes the Medal Count by Country table and filters on host countries and then groups the data to give one distinct entry for each country and season, indicating that the country was a host at some point. For example, there are two line items showing the USA – one for the summer and one for the winter – as they have hosted both.

5. The Host Country Data query was created to summarize the medal counts and entrant data for only host countries, so that we could perform analysis specific to host countries. It was created as a duplicate of the Medal Count by Country table, but then additional steps were added to isolate only data for host countries. The details of these steps are outlined in Appendix 4).

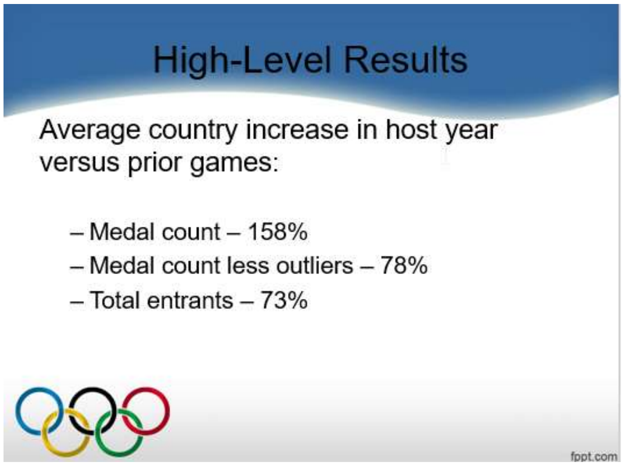
   a. The Host Country Data table was saved to a separate Excel workbook, called Host Country Analysis.xlsx to do some additional review of the data to help support our hypothesis. Appendix 6 provides the two queries included in this analysis and the detailed logic to manipulate the data.

b. In this file, an analysis was created to calculate the trend in medals won and total entrants for each host country based on the year they hosted and the games immediately prior to that

c. The goal was to calculate the impact of hosting in the number of medals won and the total number of entrants

d. The assumption was that a significant increase in one or both of these indicators would support our hypothesis that hosting the Olympics does boost your success at the Olympics

e. The result was that while not all hosts experienced improved results in their host year, on average, the numbers clearly indicated that hosting improved a country's overall success.

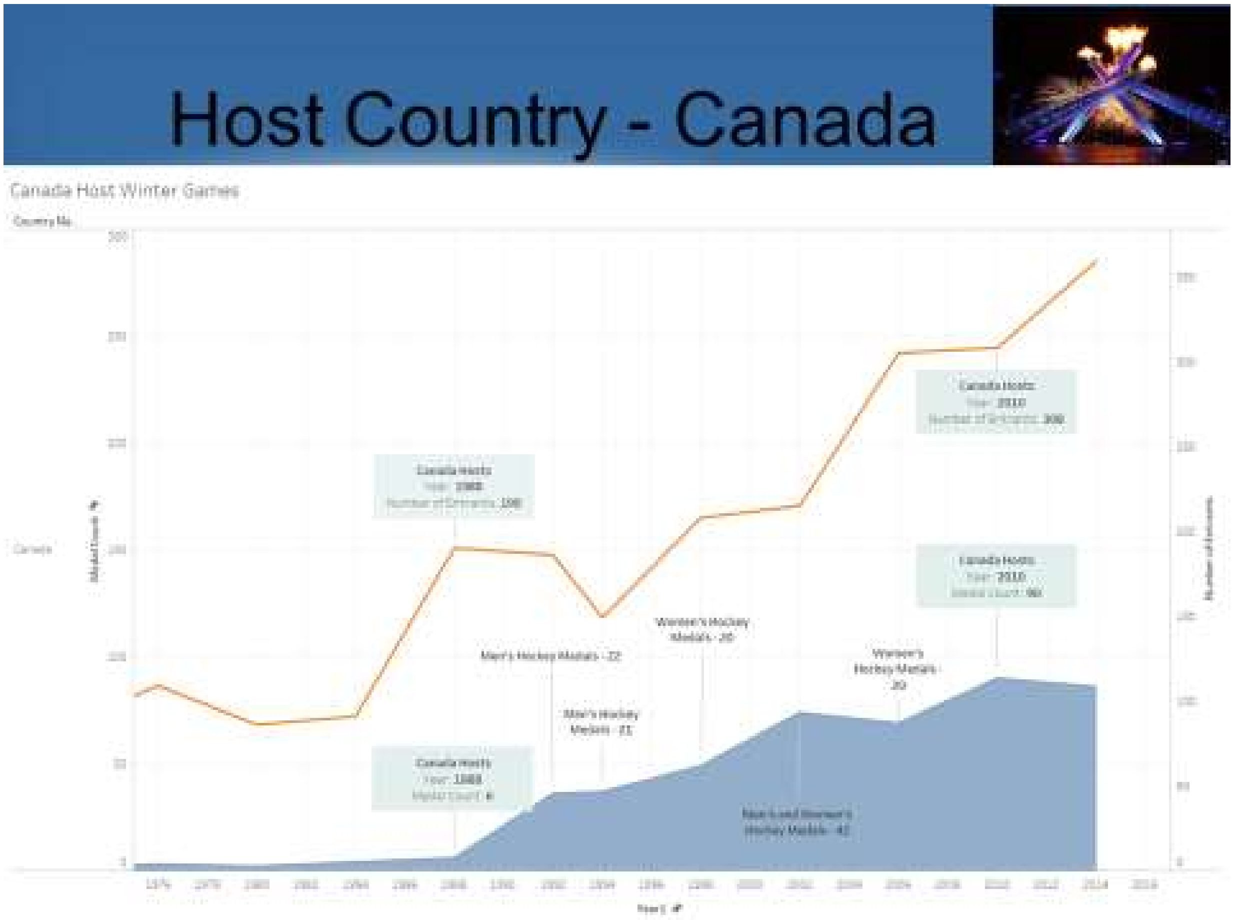


High-Level Results

Average country increase in host year versus prior games:

- – Medal count – 158%
- – Medal count less outliers – 78%
- – Total entrants – 73%

6. The Host Country Data table was also imported into Tableau. In Tableau, we focused on Canada's results for the winter games and the US's results for the summer games to create a visual to illustrate our conclusion. We chose to include the Canadian visual in our presentation.



Question 2) Correlation of number of athletes per country and their medal counts

We hypothesized that countries with larger Olympic teams would increase their chances of earning medals.

Early on in our analysis we determined that rather than count the number of athletes who participated per country, we needed to look at athletes as event entrants. The reason for this is we needed to count the times in which an athlete could medal. For instance, one athlete could compete in a variety of events in speed skating as such it was critical to count event entrants to ensure our numbers were reflective of the situation.

As well for this question, the answers were based on NOC teams, rather than country. In the timeframe of 1960 to 2016, there were a few countries that were under different governmental rule with more intense athlete training programs and it was important to review the effectiveness of the NOC teams on their individual merit. Examples of this include the Soviet Union and Eastern Germany.

In order to answer this question, we needed to determine the following:

- the total number of entrants per country and NOC inclusive from 1960 to 2016
- the total number of medals won per country and NOC inclusive from 1960 to 2016

The datasets required for this question included:

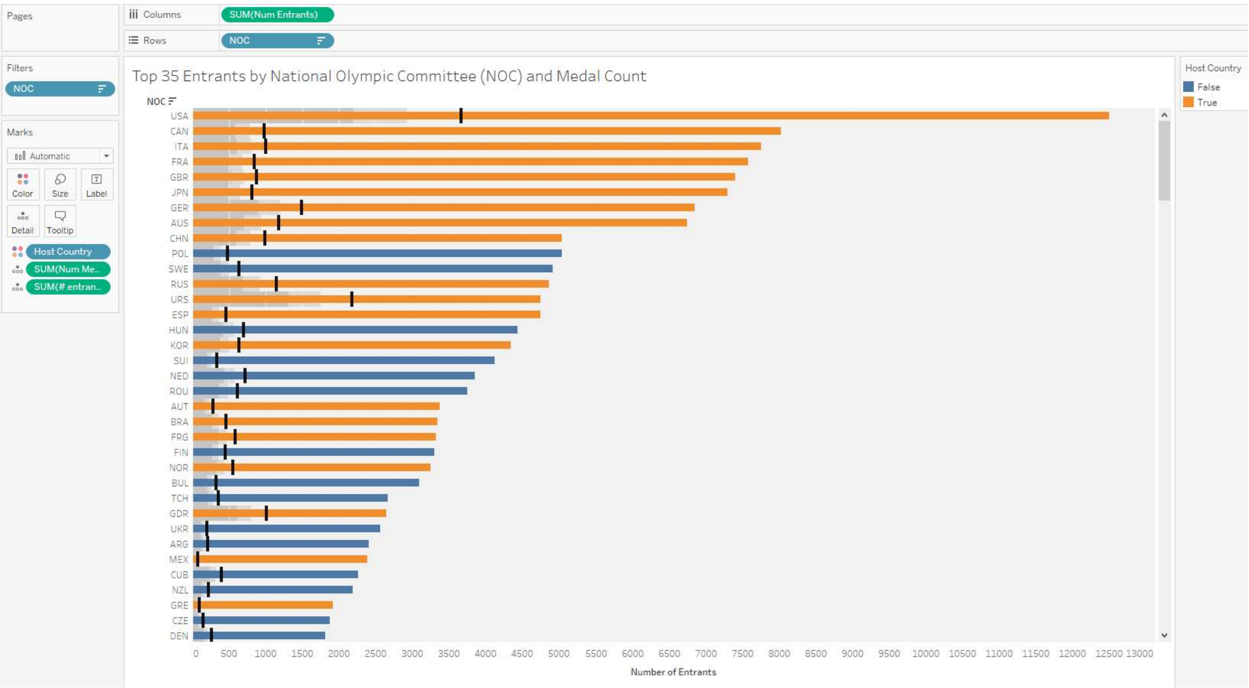
- olympic_athlete_events_table_setup.csv
- olympic_athlete_events_1960_to_present.txt
- olympic_athlete_events_ALL.txt
- noc_table_setup.csv
- noc_code_region_countrycode_country.txt
- HostCities_Countries_Years.csv
- Q1_Olympics-year-country-season-total-medals.sql
- Q2_Olympics-year-country-season-total-athletes.sql
- Q2_Entrants_and_medals.xlsx
- Q2_Entrants_and_medals.twb

Steps required to obtain the answer:

This answer and analysis were done using MySQL Workbench, Excel and Tableau.

1. In MySQL Workbench an Olympics schema was created and the Table Data Import Wizard used the .csv file to setup the table and columns. When those were in place the data in the table was deleted.
2. The Load Data Infile process was used to import the olympic_athlete_events_1960_to_present.txt file data into MySQL Workbench was process would take seconds to import, versus 10+ days if the Table Data Import Wizard feature was used. The same process was used to get the NOC regions into the Olympics database as a table.
3. A couple of queries were written in SQL to extract the necessary data. SQL files contain commented out codes and tests which were used to verify the data results. (Appendix 7)
4. The query result data was then copied from MySQL and pasted into Excel
5. In Excel a variety of techniques were used in order to produce the final Excel sheet named Q2_Entrants_and_medals.xlsx which would be used in Tableau for further analysis.
    a. Data from noc_code_region_countrycode_country.txt file and HostCities_Countries_Years.csv file were added into separate sheets in the Excel file
    b. The SQL data located in “total athletes entrants medals” sheet was enhanced in the following ways:
        i. Vlookup to bring in country codes which correlated to the NOC code
        ii. Countifs to determine if the country (based on country code) was a host country or not
        iii. Vlookup to bring in the medal count based on the correlating NOC

6. In Tableau a couple of bar charts were created in order to conduct the analysis.



Columns: SUM(Num Entrants)

Rows: NOC

Filters: NOC

Marks: Automatic

Host Country
SUM(Num Me..
SUM(# entran..

Top 35 Entrants by National Olympic Committee (NOC) and Medal Count

The first bar chart (Appendix 8) depicts

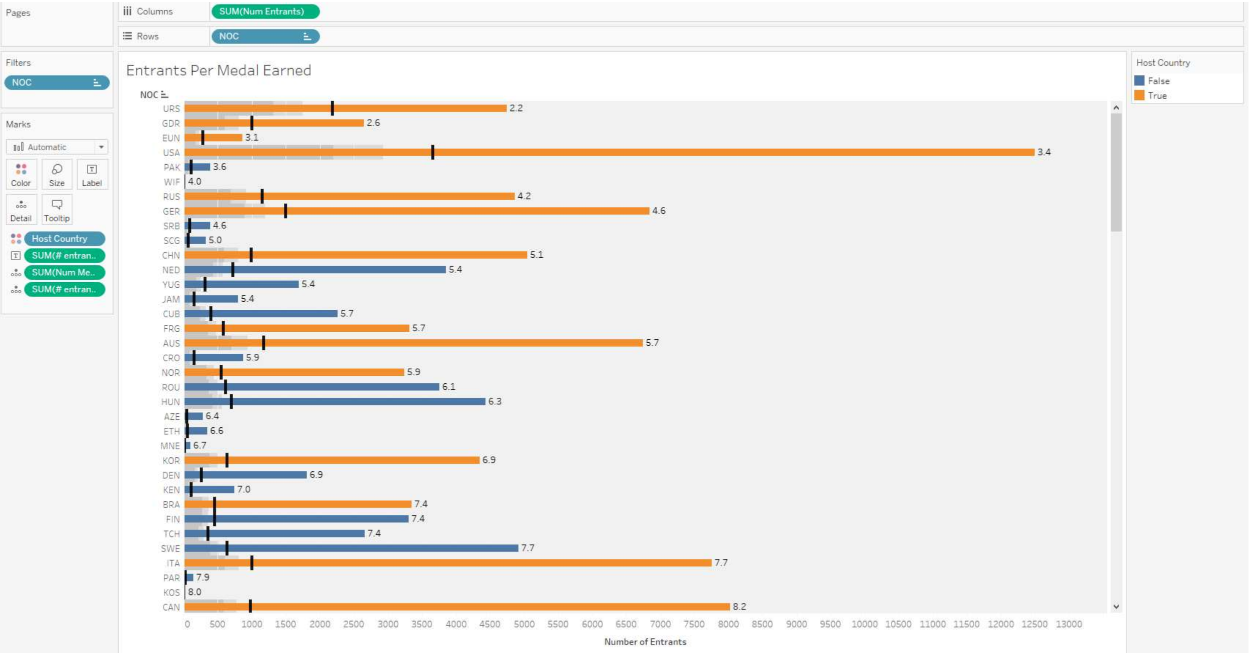
a. In descending order, the NOC teams and the total number of entrants each had for the Olympic games, summer and winter, from 1960 to 2016

b. Number of medals each NOC team won (the black tick on the bars)

c. If at one point, they were a host country (orange) or not (blue)

The USA, Canada and Italy had the most entrants in Olympic sporting events. Respectively their ratio of entrants to one medal is, USA 3.4:1, Canada 8.2:1 and Italy 7.7:1.

Interestingly nearly all NOC teams whose country hosted one or more Olympic games appear in the top 35 NOC teams for number of entrants. It should be noted that one NOC team was not in the top 35. The EUN, known as the Unified Team, was representing the former Soviet Union in the summer and winter Olympics in 1992.



Columns SUM(Num Entrants)

Rows NOC

Filters NOC

Host Country
SUM(# entran..
SUM(Num Me..
SUM(# entran..

The second bar chart (Appendix 8) depicts

a. In descending order, the NOC team and ratio of entrants to one medal (the number at the end of the bar)

b. The total number of entrants each had for the Olympic games, summer and winter, from 1960 to 2016

c. Number of medals each NOC team won (the black tick on the bars)

d. If at one point, they were a host country (orange) or not (blue)

With the help of a calculated field in Tableau, we were able to determine ratio of entrants to medals and sort the data by this ratio.

The top three NOC teams were very successful in gaining the highest number of medals with less entrants in Olympic sporting events. These three NOC teams no longer exist. These teams were:

a. URS – Soviet Union until 1991

b. EUN – Unified team competing as former Soviet Union in 1992

c. GDR – Eastern Germany under Soviet rule until October 1990

The sports regime under Soviet rule was known to be intensive, deliberate, and clearly effective. Soviet rule over Russian and East Germany ceased over 24 to 26 years ago and yet they still hold the highest medal success statistics today as seen by the bar chart. URS saw 2.2 entrants per medal, GDR 2.6 entrants, and the EUN 3.1.

Question 3) Correlation between medal count per country and GDP

For our analysis on GDP, we figured that there would be a strong correlation between a country's GDP and their success at the Olympics. In order to analyze this, we had to pull in the GDP data into the Olympic data table and then summarize the results by country. In order to complete the analysis, we had to determine the following:

- The income category for each country
- The total cumulative medals won by each country
- The total cumulative entrants by country

The datasets required for this question included:

- Athlete dataset
- NOC Regions file
- Table mapping host city to country (World Cities)
- GDP Data file

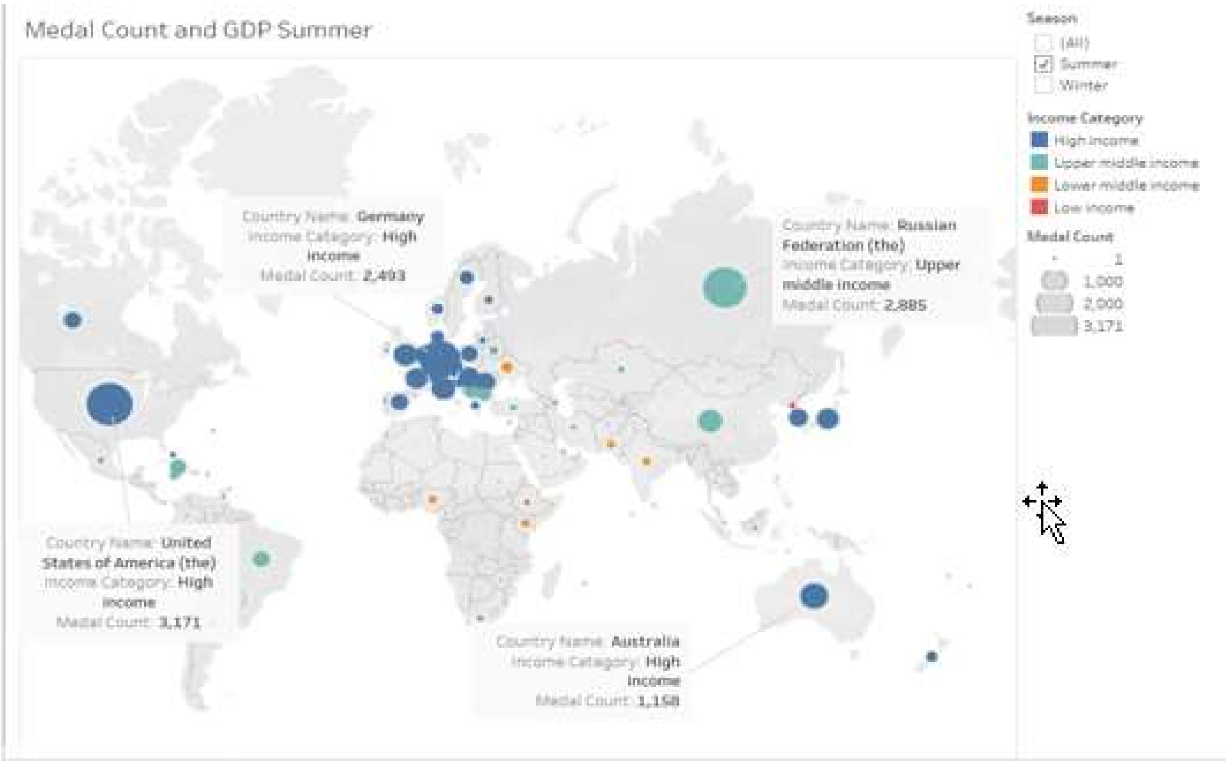
The queries used in the Olympic Analysis.xlsx file are:

- the **Olympic Mega Table** query (Appendix 1)
- the **Athlete Total Counts** query (Appendix 2)
- the **Medal Count by Country** query (Appendix 3)
- the **GDP by Country by Year** query

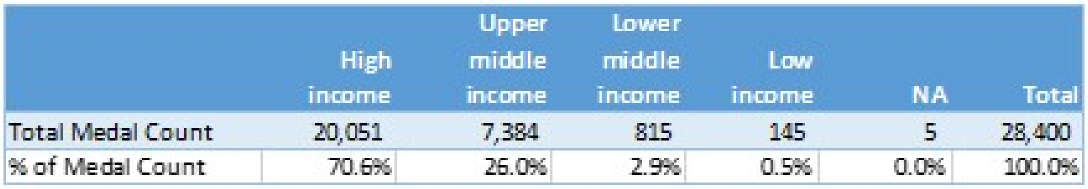
The first three queries were described in question 1 above.

The GDP by Country by Year query basically just pulls the GDP Data.xls file into the overall Olympic Analysis.xlsx file. From there, the GDP data was merged into the Medal Count by Country query to pull in the GDP and Income Category by country. These steps are outlined in Appendix 3 for Medal Count by Country.

The resulting table of data was saved in a separate workbook called Medal County by Country.xlsx and this was used to load into Tableau to create a visual showing the correlation between total medal counts and GDP income category. The resulting map supports the strong correlation between the two.



In addition to the visual, the below chart summarizes the overall distribution of medal count by income category (using a Pivot Table), again supporting the theory that medal success at the Olympics is highly correlated to higher income countries.



	High income	Upper middle income	Lower middle income	Low income	NA	Total
Total Medal Count	20,051	7,384	815	145	5	28,400
% of Medal Count	70.6%	26.0%	2.9%	0.5%	0.0%	100.0%

Question 4) Athletes who participated in summer and winter Olympic sports

Dual sport athletes were also a topic which interested us. We were interested in finding out how many athletes completed in both summer and winter Olympic games throughout their career. As well we wanted to determine what combination of summer and winter sports were most popular and why.

In order to answer this question, we needed to determine the following:

- the total number of athletes who participate in both summer and winter Olympics
- the summer and winter sport each dual sport athlete competed in
- if dual sport athletes win medals in both sports

The datasets required for this question included:

- olympic_athlete_events_table_setup.csv

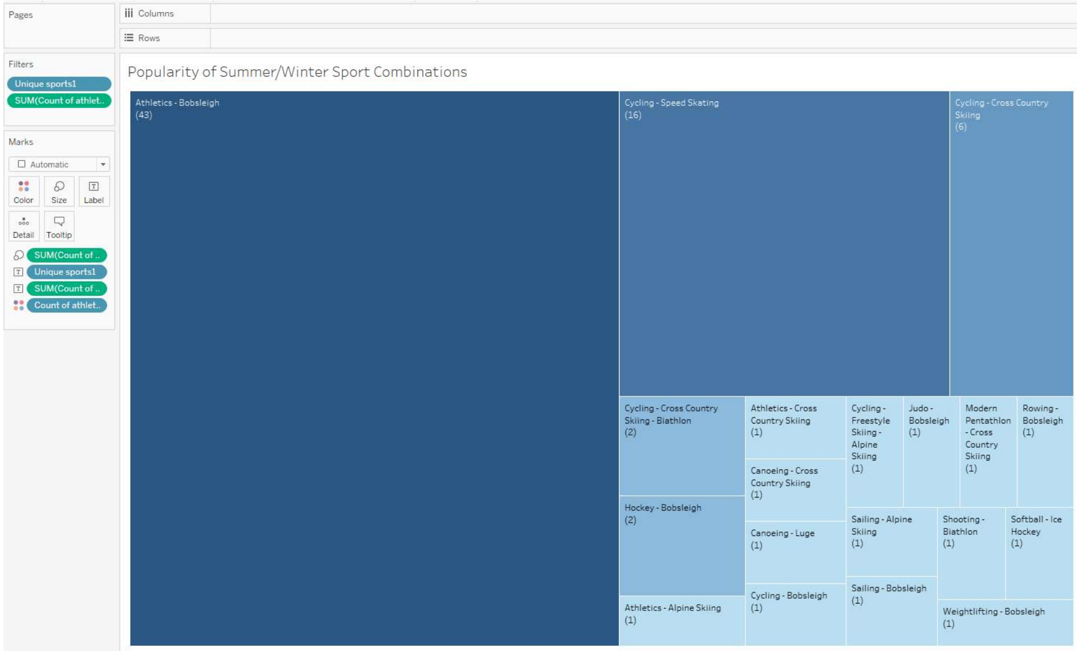
- olympic_athlete_events_1960_to_present.txt
- Q4_Dual_Sport_Athletes.sql
- Q4_Dual_Sport_Athletes.xlsx
- Q4_Dual_Sport_Athletes.twb

Steps required to obtain the answer:

This answer and analysis were done using MySQL Workbench, Excel and Tableau.

1. As per question 2, the data was cleaned and imported/loaded into the Olympics database in MySQL Workbench.
2. A more complex SQL query was written using a WITH clause to extract the necessary data. (Appendix 9)
3. The query result data was then copied from MySQL and pasted into Excel
4. In Excel a variety of techniques were used in order to produce the final Excel sheet named Q4_Dual_Sport_Athletes.xlsx which would be used in Tableau for further analysis.
    a. The SQL data located in “Dual Sport Athletes” sheet was enhanced in the following ways:
        i. Filtering sport column to review results and clean up any duplicates. Some sports were listed summer then winter, and others were winter then summer.
        ii. UNIQUE function to extract a list the unique sport combinations
        iii. Countif to determine the number of times the sports combination appeared in the data array
        iv. Text to columns to take original concatenated data of “sport,sport” and split into two columns and bring back together as “sport – sport” which worked better when used in Tableau

7. In Tableau a couple of bar charts were created in order to conduct the analysis.


Pages
Columns
Rows
Filters
Unique sports1
SUM(Count of athlet..
Marks
Automatic
Color
Size
Label
Detail
Tooltip
SUM(Count of ..
Unique sports1
SUM(Count of ..
Count of athlet..
Popularity of Summer/Winter Sport Combinations
Athletics - Bobsleigh (43)
Cycling - Speed Skating (16)
Cycling - Cross Country Skiing (6)
Cycling - Cross Country Skiing - Biathlon (2)
Hockey - Bobsleigh (2)
Athletics - Alpine Skiing (1)
Athletics - Cross Country Skiing (1)
Canoeing - Cross Country Skiing (1)
Canoeing - Luge (1)
Cycling - Bobsleigh (1)
Cycling - Freestyle Skiing - Alpine Skiing (1)
Judo - Bobsleigh (1)
Modern Pentathlon - Cross Country Skiing (1)
Rowing - Bobsleigh (1)
Sailing - Alpine Skiing (1)
Sailing - Bobsleigh (1)
Shooting - Biathlon (1)
Softball - Ice Hockey (1)
Weightlifting - Bobsleigh (1)

The treemap chart (Appendix 10) depicts

a. The popularity of the various summer and winter sport combinations

Of the 507 athletes who competed in the Olympics since 1960 to 2016, there were 83 athletes who competed both in the summer and winter games.

The top three sport combinations include:

a. Athletics / Bobsleigh (43 athletes)

b. Cycling / Speed skating (16 athletes)

c. Cycling / Cross country skiing (6 athletes)

The athletes that competed under the Athletics sport category were typically runners.

With such popularity for athletes who compete in Athletics / Bobsleigh and Cycling / Speed skating the question was why?

Athletics / Bobsleigh

Nearing the late 1950s, as the sport was developing into what it is today it looks like it was not uncommon to recruit athletes from athletics (track and field) to bobsleigh. Athletes who compete in Athletics, particularly runners, seem more likely to convert over to bobsleigh. This duality is attractive to athletes as it elongates their athletic career.

(source: https://news.medill.northwestern.edu/chicago/changing-lanes-track-athletes-switch-to-bobsled-to-prolong-athletic-careers/ and https://www.ibsf.org/en/our-sports/bobsleigh-history)

Cycling / Speed skating

The commonalities between cycling and speed skating are well documented. Among the reasons for athletes to compete in both sports includes:

a. Physiology
The body position of a cyclist and speed skater are very similar in that both are hunched over for increased aerodynamics. Their legs are their source of power however the rest of their body needs to be conditioned to sustain the hunched position for short to long periods of time depending on the sporting event being competed in.

b. Drafting techniques
The way athletes can benefit from the athletes competing against them is similar in that they can position themselves strategically which uses less energy in the hopes that the energy can be used closer to the finish line.

c. Suffering
In both sports it is the athlete that can endure the most in the fastest time who wins. In

addition to the physical boundaries that are pushed these athletes need to be mentally tough.

(source: http://www.bikeroar.com/articles/speed-skating-and-cycling-the-similarities-are-uncanny)

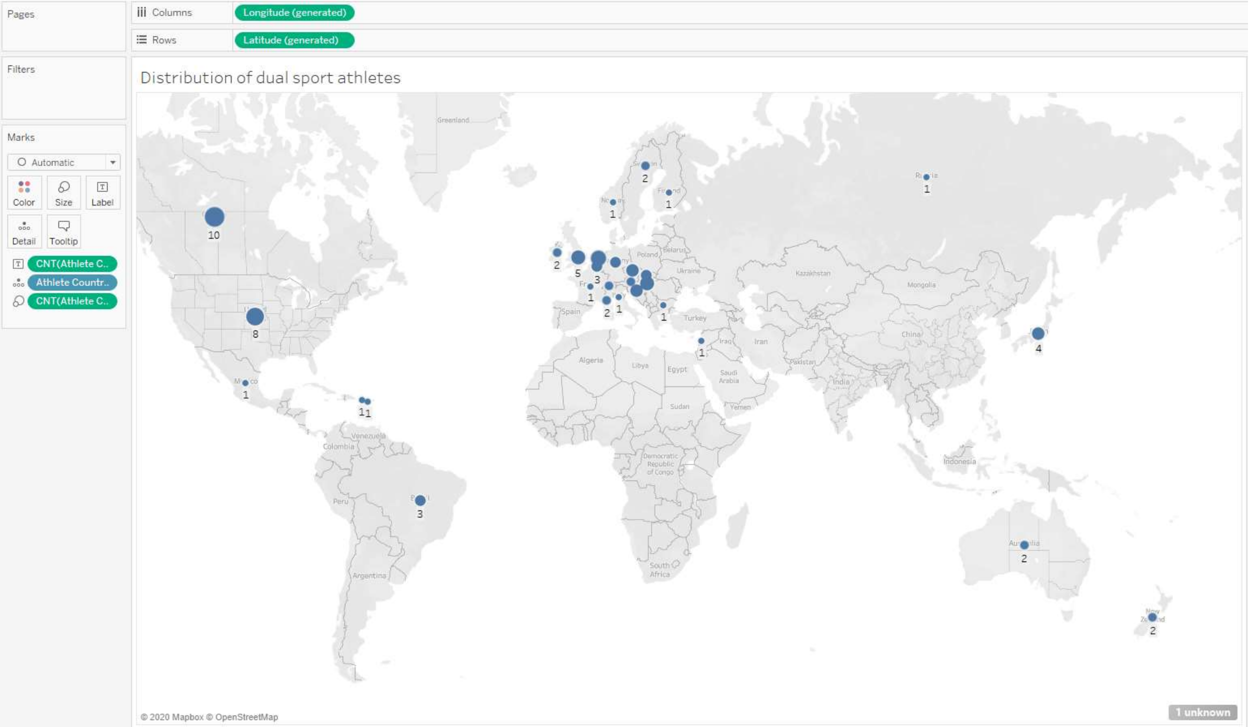
Interestingly, very few athletes medal in both seasons. If they do medal, they will be stronger in one sport than the other. This was analyzed in Alyson's **Olympic Analysis_Queries Removed.xlsx** file, under the Dual Athlete Medals tab.

Athletes who have medaled in both a summer and winter Olympics include:

a. Clara Hughes, Canada, Cycling / Speed Skating

b. Christa Rothenburg, Germany, Cycling / Speed Skating

c. Lauryn Chenet Williams – USA, Athletics (100m) / Bobsleigh



Columns: Longitude (generated)

Rows: Latitude (generated)

Distribution of dual sport athletes

Marks: Automatic

CNT(Athlete C..

Athlete Countr..

CNT(Athlete C..

The map (Appendix 10) depicts where in the globe dual sport athletes are located. From 1960 to 2016, Canada had the most athletes who completed in both the summer and winter games. USA had 8, and the Netherlands with 6.

In the map there is 1 unknown. This is not a mistake. This is a record of Aleksandar Milenkovi who completed in Cycling, Cross Country Skiing and Biathlon. He competed as an IOA (Individual Olympic Athlete) as Yugoslavia dissolved and Serbia came to be. He competed as YUG, IOA and SRB in his years in the Olympics.

IOAs are athletes who compete as Independent Olympians for various reasons, including political transition, international sanctions, suspensions of National Olympic Committees, and compassion. (source: https://en.wikipedia.org/wiki/Independent_Olympians_at_the_Olympic_Games)

Conclusion

Through the data analysis we see that, on average, when a country is awarded the right to host the Olympic games, their medal counts in the year they host increase from the previous games. As being awarded the Olympic games to host happens 7 years in advance of the actual event, it gives the host country time to focus on and provide extra supports for their athletes to perform at their best when the time comes.

While it doesn't always hold true, our analysis did support our hypothesis that a country's success at the Olympics is highly correlated to the number of entrants that they bring to the Olympics. The USA in particular has the largest cumulative number of entrants over time and by far the largest number of medals won. However the countries under old soviet rule somewhat go against this theory in that they have the highest number of medals won to entrants (e.g. the Soviet Union required only 2.2 entrants for every medal won). So while there are definitely exceptions, the correlation of medals won to entrants is quite high (0.84 on a per country/per games analysis).

There is a strong correlation between per country medal count and GDP. Of the countries that participate and earn medals at the Olympics, for the vast majority (97%), the GDP is listed as high income and upper middle income.

Lastly, the Olympics have had a number of athletes who participate in both the summer and winter Olympics. The most common sport combination are athletes who participate in Athletics (particularly running) in the summer and changed over to Bobsleigh for the winter. The next popular combination is cycling and speed skating. Dual sport athletes are not highly likely to medal in both sports and there seems to always be one sport which is their strength. Dual sport athletes therefore do not have a significant impact on a country's improved success in either Olympic games.

Appendix 1 – Olympic Mega Table Query Steps

```
let
   Source = Excel.CurrentWorkbook(){[Name="Table1"]}[Content],
   #"Changed Type" = Table.TransformColumnTypes(Source,{{"ID", Int64.Type},
{"athelete_name", type text}, {"Sex", type text}, {"Age", type any}, {"Team", type text}, {"NOC",
type text}, {"Year", Int64.Type}, {"Season", type text}, {"City", type text}, {"Sport", type text},
{"Event", type text}, {"Medal", type text}}),
   #"Replaced Value" = Table.ReplaceValue(#"Changed
Type","NULL",0,Replacer.ReplaceValue,{"Age"}),
   #"Changed Type1" = Table.TransformColumnTypes(#"Replaced Value",{{"Age", Int64.Type}}),
   #"Merged Queries" = Table.NestedJoin(#"Changed
Type1",{"NOC"},noc_regions,{"NOC"},"noc_regions",JoinKind.LeftOuter),
   #"Expanded noc_regions" = Table.ExpandTableColumn(#"Merged Queries", "noc_regions",
{"region", "Country Code", "Country Name"}, {"region", "Country Code", "Country Name"}),
   #"Reordered Columns" = Table.ReorderColumns(#"Expanded noc_regions",{"ID",
"athelete_name", "Sex", "Age", "Team", "NOC", "Year", "Season", "City", "region", "Sport",
"Event", "Medal"}),
   #"Renamed Columns" = Table.RenameColumns(#"Reordered Columns",{{"region",
"NOC_Region"}, {"City", "Host_City"}}),
   #"Reordered Columns1" = Table.ReorderColumns(#"Renamed Columns",{"ID",
"athelete_name", "Sex", "Age", "Team", "NOC", "NOC_Region", "Year", "Season", "Host_City",
"Sport", "Event", "Medal"}),
   #"Merged Queries1" = Table.NestedJoin(#"Reordered Columns1",{"Host_City"},#"World
Cities",{"city_ascii"},"World Cities",JoinKind.LeftOuter),
   #"Expanded World Cities" = Table.ExpandTableColumn(#"Merged Queries1", "World Cities",
{"country"}, {"country"}),
   #"Renamed Columns1" = Table.RenameColumns(#"Expanded World Cities",{{"country",
"Host_Country"}}),
   #"Reordered Columns2" = Table.ReorderColumns(#"Renamed Columns1",{"ID",
"athelete_name", "Sex", "Age", "Team", "NOC", "NOC_Region", "Country Code", "Country
Name", "Year", "Season", "Host_City", "Host_Country", "Sport", "Event", "Medal"}),
   #"Replaced Value1" = Table.ReplaceValue(#"Reordered Columns2","Bosnia And
Herzegovina","Serbia",Replacer.ReplaceText,{"Host_Country"})
in

   #"Replaced Value1"
```

Appendix 2 – Athlete Total Counts

```
let
    Source = #"Olympic Mega Table",
    #"Grouped Rows" = Table.Group(Source, {"Country Code", "Country Name", "Year", "Season", "Host_Country"}, {{"Number of Athletes", each Table.RowCount(_), type number}}),
    #"Sorted Rows" = Table.Sort(#"Grouped Rows",{{"Country Code", Order.Ascending}})
in
    #"Sorted Rows"
```

Appendix 3 – Medal County by Country Power Query Steps

```
let
   Source = #"Olympic Mega Table",
   #"Replaced Value" = Table.ReplaceValue(Source,"NA","0",Replacer.ReplaceText,{"Medal"}),
   #"Replaced Value1" = Table.ReplaceValue(#"Replaced
Value","Bronze","1",Replacer.ReplaceText,{"Medal"}),
   #"Replaced Value2" = Table.ReplaceValue(#"Replaced
Value1","Silver","1",Replacer.ReplaceText,{"Medal"}),
   #"Replaced Value3" = Table.ReplaceValue(#"Replaced
Value2","Gold","1",Replacer.ReplaceText,{"Medal"}),
   #"Changed Type2" = Table.TransformColumnTypes(#"Replaced Value3",{{"Medal",
Int64.Type}}),
   #"Grouped Rows" = Table.Group(#"Changed Type2", {"Season", "Country Code", "Country
Name", "Year", "Host_City", "Host_Country"}, {{"Medal Count", each List.Sum([Medal]),
Int64.Type}}),
   #"Added Custom" = Table.AddColumn(#"Grouped Rows", "Flag Host Country", each if
[Host_Country] = [Country Name] then "Yes" else "No"),
   #"Changed Type1" = Table.TransformColumnTypes(#"Added Custom",{{"Flag Host Country",
type text}}),
   #"Changed Type" = Table.TransformColumnTypes(#"Changed Type1",{{"Flag Host Country",
type text}}),
   #"Merged Queries" = Table.NestedJoin(#"Changed Type",{"Country Code", "Year"},#"GDP by
Country_Year",{"Country_Code", "Year"},"GDP by Country_Year",JoinKind.LeftOuter),
   #"Expanded GDP by Country_Year" = Table.ExpandTableColumn(#"Merged Queries", "GDP by
Country_Year", {"Income Category", "GDP"}, {"Income Category", "GDP"}),
   #"Merged Queries1" = Table.NestedJoin(#"Expanded GDP by Country_Year",{"Country Code",
"Year"},#"Population by Country_Year",{"Country_Code", "Attribute"},"Population by
Country_Year",JoinKind.LeftOuter),
   #"Expanded Population by Country_Year" = Table.ExpandTableColumn(#"Merged Queries1",
"Population by Country_Year", {"Population"}, {"Population"}),
   #"Merged Queries2" = Table.NestedJoin(#"Expanded Population by Country_Year",{"Country
Code", "Year", "Season"},#"Athlete Total Counts",{"Country Code", "Year", "Season"},"Athlete
Total Counts",JoinKind.LeftOuter),
   #"Expanded Athlete Total Counts" = Table.ExpandTableColumn(#"Merged Queries2", "Athlete
Total Counts", {"Number of Athletes"}, {"Number of Athletes"}),
   #"Renamed Columns1" = Table.RenameColumns(#"Expanded Athlete Total
Counts",{{"Number of Athletes", "Number of Entrants"}})
in
```

```
#"Renamed Columns1"
```

Appendix 4 – Summary Host Countries

```
let
    Source = #"Medal Count by Country",
    #"Filtered Rows" = Table.SelectRows(Source, each ([Flag Host Country] = "Yes")),
    #"Grouped Rows" = Table.Group(#"Filtered Rows", {"Flag Host Country", "Country Code", "Country Name", "Season", "Host_Country"}, {{"Count", each Table.RowCount(_), type number}})
in
    #"Grouped Rows"
```

Appendix 5 - Host Country Data Power Query Steps

```
let

    Source = #"Olympic Mega Table",

    #"Replaced Value" = Table.ReplaceValue(Source,"NA","0",Replacer.ReplaceText,{"Medal"}),

    #"Replaced Value1" = Table.ReplaceValue(#"Replaced
Value","Bronze","1",Replacer.ReplaceText,{"Medal"}),

    #"Replaced Value2" = Table.ReplaceValue(#"Replaced
Value1","Silver","1",Replacer.ReplaceText,{"Medal"}),

    #"Replaced Value3" = Table.ReplaceValue(#"Replaced
Value2","Gold","1",Replacer.ReplaceText,{"Medal"}),

    #"Changed Type2" = Table.TransformColumnTypes(#"Replaced Value3",{{"Medal", Int64.Type}}),

    #"Grouped Rows" = Table.Group(#"Changed Type2", {"Season", "Country Code", "Country Name",
"Year", "Host_City", "Host_Country"}, {{"Medal Count", each List.Sum([Medal]), Int64.Type}}),

    #"Added Custom" = Table.AddColumn(#"Grouped Rows", "Flag Host Country", each if [Host_Country]
= [Country Name] then "Yes" else "No"),

    #"Changed Type1" = Table.TransformColumnTypes(#"Added Custom",{{"Flag Host Country", type
text}}),

    #"Changed Type" = Table.TransformColumnTypes(#"Changed Type1",{{"Flag Host Country", type
text}}),

    #"Merged Queries" = Table.NestedJoin(#"Changed Type",{"Country Code", "Year"},#"GDP by
Country_Year",{"Country_Code", "Year"},"GDP by Country_Year",JoinKind.LeftOuter),

    #"Expanded GDP by Country_Year" = Table.ExpandTableColumn(#"Merged Queries", "GDP by
Country_Year", {"Income Category", "GDP"}, {"Income Category", "GDP"}),

    #"Merged Queries1" = Table.NestedJoin(#"Expanded GDP by Country_Year",{"Country Code",
"Year"},#"Population by Country_Year",{"Country_Code", "Attribute"},"Population by
Country_Year",JoinKind.LeftOuter),

    #"Expanded Population by Country_Year" = Table.ExpandTableColumn(#"Merged Queries1",
"Population by Country_Year", {"Population"}, {"Population"}),

    #"Merged Queries2" = Table.NestedJoin(#"Expanded Population by Country_Year",{"Country Code",
"Year", "Season"},#"Athlete Total Counts",{"Country Code", "Year", "Season"},"Athlete Total
Counts",JoinKind.LeftOuter),

    #"Expanded Athlete Total Counts" = Table.ExpandTableColumn(#"Merged Queries2", "Athlete Total
Counts", {"Number of Athletes"}, {"Number of Athletes"}),

    #"Renamed Columns1" = Table.RenameColumns(#"Expanded Athlete Total Counts",{{"Number of
Athletes", "Number of Entrants"}}),
    #"Merged Queries3" = Table.NestedJoin(#"Renamed Columns1",{"Country Code",
"Season"},#"Summary Host Countries",{"Country Code", "Season"},"Summary Host
Countries",JoinKind.LeftOuter),

    #"Expanded Summary Host Countries" = Table.ExpandTableColumn(#"Merged Queries3", "Summary
Host Countries", {"Flag Host Country"}, {"Flag Host Country.1"}),

    #"Renamed Columns" = Table.RenameColumns(#"Expanded Summary Host Countries",{{"Flag Host
Country.1", "Ever Host Country"}}),

    #"Replaced Value4" = Table.ReplaceValue(#"Renamed
Columns",null,"No",Replacer.ReplaceValue,{"Ever Host Country"}),

    #"Renamed Columns2" = Table.RenameColumns(#"Replaced Value4",{{"Flag Host Country", "Event
Host Country"}}),

    #"Filtered Rows" = Table.SelectRows(#"Renamed Columns2", each ([Ever Host Country] = "Yes")),

    #"Reordered Columns" = Table.ReorderColumns(#"Filtered Rows",{"Country Code", "Country Name",
"Year", "Season", "Host_City", "Host_Country", "Medal Count", "Event Host Country", "Ever Host
Country", "Income Category", "GDP", "Population", "Number of Entrants"})

in

    #"Reordered Columns"
```

Appendix 6 - Host Country Analysis

Host Country Stats Query:

```
let
    Source = Excel.CurrentWorkbook(){[Name="Table2"]}[Content],
    #"Changed Type" = Table.TransformColumnTypes(Source,{{"Country Code", type text}, {"Country Name", type text}, {"Year", Int64.Type}, {"Season", type text}, {"Host_City", type text}, {"Host_Country", type text}, {"Medal Count", Int64.Type}, {"Event Host Country", type text}, {"Ever Host Country", type text}, {"Income Category", type text}, {"GDP", Int64.Type}, {"Population", Int64.Type}, {"Number of Entrants", Int64.Type}, {"Event Prior to Host Year", type any}}),
    #"Reordered Columns" = Table.ReorderColumns(#"Changed Type",{"Country Code", "Country Name", "Year", "Event Prior to Host Year", "Season", "Host_City", "Host_Country", "Medal Count", "Event Host Country", "Ever Host Country", "Income Category", "GDP", "Population", "Number of Entrants"}),
    #"Replaced Value" = Table.ReplaceValue(#"Reordered Columns","NA","",Replacer.ReplaceValue,{"Event Prior to Host Year"}),
    #"Changed Type1" = Table.TransformColumnTypes(#"Replaced Value",{{"Event Prior to Host Year", Int64.Type}})
in
    #"Changed Type1"
```

Host Country Analysis Query:

```
let
    Source = Excel.CurrentWorkbook(){[Name="Table2"]}[Content],
    #"Changed Type" = Table.TransformColumnTypes(Source,{{"Country Code", type text}, {"Country Name", type text}, {"Year", Int64.Type}, {"Season", type text}, {"Host_City", type text}, {"Host_Country", type text}, {"Medal Count", Int64.Type}, {"Event Host Country", type text}, {"Ever Host Country", type text}, {"Income Category", type text}, {"GDP", Int64.Type}, {"Population", Int64.Type}, {"Number of Entrants", Int64.Type}, {"Event Prior to Host Year", type any}}),
    #"Reordered Columns" = Table.ReorderColumns(#"Changed Type",{"Country Code", "Country Name", "Year", "Event Prior to Host Year", "Season", "Host_City", "Host_Country", "Medal Count", "Event Host Country", "Ever Host Country", "Income Category", "GDP", "Population", "Number of Entrants"}),
    #"Replaced Value" = Table.ReplaceValue(#"Reordered Columns","NA","",Replacer.ReplaceValue,{"Event Prior to Host Year"}),
    #"Changed Type1" = Table.TransformColumnTypes(#"Replaced Value",{{"Event Prior to Host Year", Int64.Type}}),
    #"Filtered Rows" = Table.SelectRows(#"Changed Type1", each ([Event Host Country] = "Yes")),

    #"Merged Queries" = Table.NestedJoin(#"Filtered Rows",{"Country Code", "Country Name", "Event Prior to Host Year", "Season"},#"Host Country Stats",{"Country Code", "Country Name", "Year", "Season"},"Host Country Stats",JoinKind.LeftOuter),

    #"Expanded Host Country Stats" = Table.ExpandTableColumn(#"Merged Queries", "Host Country Stats", {"Medal Count", "Number of Entrants"}, {"Medal Count.1", "Number of Entrants.1"}),

    #"Renamed Columns" = Table.RenameColumns(#"Expanded Host Country Stats",{{"Medal Count.1", "Prior Medal Count"}, {"Number of Entrants.1", "Prior Number of Entrants"}}),

    #"Removed Columns" = Table.RemoveColumns(#"Renamed Columns",{"Ever Host Country"}),

    #"Reordered Columns1" = Table.ReorderColumns(#"Removed Columns",{"Country Code", "Country Name", "Year", "Event Prior to Host Year", "Season", "Host_City", "Host_Country", "Medal Count", "Prior Medal Count", "Number of Entrants", "Prior Number of Entrants", "Event Host Country", "Income Category", "GDP", "Population"}),

    #"Added Custom" = Table.AddColumn(#"Reordered Columns1", "Change in Medal Count", each ([Medal Count]-[Prior Medal Count])/[Prior Medal Count]),

    #"Changed Type2" = Table.TransformColumnTypes(#"Added Custom",{{"Change in Medal Count", Percentage.Type}}),

    #"Reordered Columns2" = Table.ReorderColumns(#"Changed Type2",{"Country Code", "Country Name", "Year", "Event Prior to Host Year", "Season", "Host_City", "Host_Country", "Medal Count", "Prior Medal Count", "Change in Medal Count", "Number of Entrants", "Prior Number of Entrants", "Event Host Country", "Income Category", "GDP", "Population"}),

    #"Added Custom1" = Table.AddColumn(#"Reordered Columns2", "Change in Entrants", each ([Number of Entrants]-[Prior Number of Entrants])/[Prior Number of Entrants]),

    #"Changed Type3" = Table.TransformColumnTypes(#"Added Custom1",{{"Change in Entrants", Percentage.Type}}),

    #"Reordered Columns3" = Table.ReorderColumns(#"Changed Type3",{"Country Code", "Country Name", "Year", "Event Prior to Host Year", "Season", "Host_City", "Host_Country", "Medal Count", "Prior Medal Count", "Change in Medal Count", "Number of Entrants", "Prior Number of Entrants", "Change in Entrants", "Event Host Country", "Income Category", "GDP", "Population"}),

    #"Removed Columns1" = Table.RemoveColumns(#"Reordered Columns3",{"GDP", "Population"})

in

    #"Removed Columns1"
```

Appendix 7 – Entrant to Medals Analysis

File: Q1_Olympics-year-country-season-total-medals.sql

```
-- Shows sum medals per season for each country per year

SELECT a.oly_year, a.season, a.noc, COUNT(a.medals) AS MedalCount

FROM athletes a

WHERE medals <> 'NULL'

GROUP BY oly_year, NOC, season

ORDER BY oly_year, NOC, season;
```

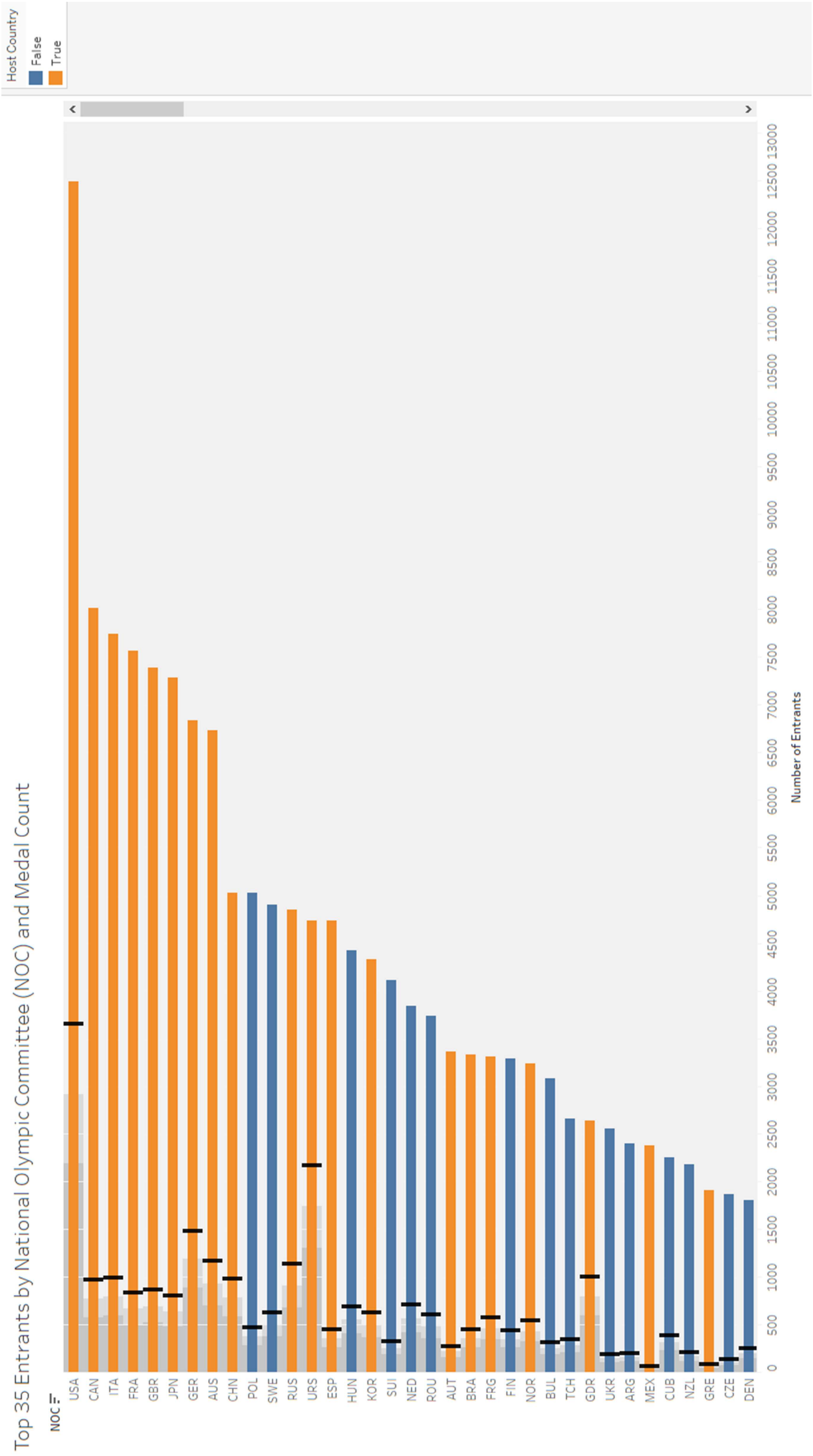
File: Q2_Olympics-year-country-season-total-athletes.sql

```
-- Shows number of athletes per noc TOTAL from 1960-2016.
-- Cross checked and validated data with example CAN in 1960 has 129 total WHERE oly_year = '1960'. Received correct numbers.

SELECT noc,COUNT(DISTINCT athlete_id) AS athletes

FROM athletes

GROUP BY noc;

-- Same as above but without DISTINCT on athlete-id as we needed entrants not athletes.

SELECT noc,COUNT(athlete_id) AS athletes

FROM athletes

GROUP BY noc

ORDER BY noc;


-- Medal count per noc TOTAL from 1960-2016
-- Cross checked and validated data with example WHERE medals <> 'NULL' AND noc = 'CAN' AND oly_year = '1960'  = 30.  Received correct numbers.

SELECT a.noc, COUNT(a.medals) AS MedalCount

FROM athletes a

WHERE medals <> 'NULL'

GROUP BY noc

ORDER BY noc;
```

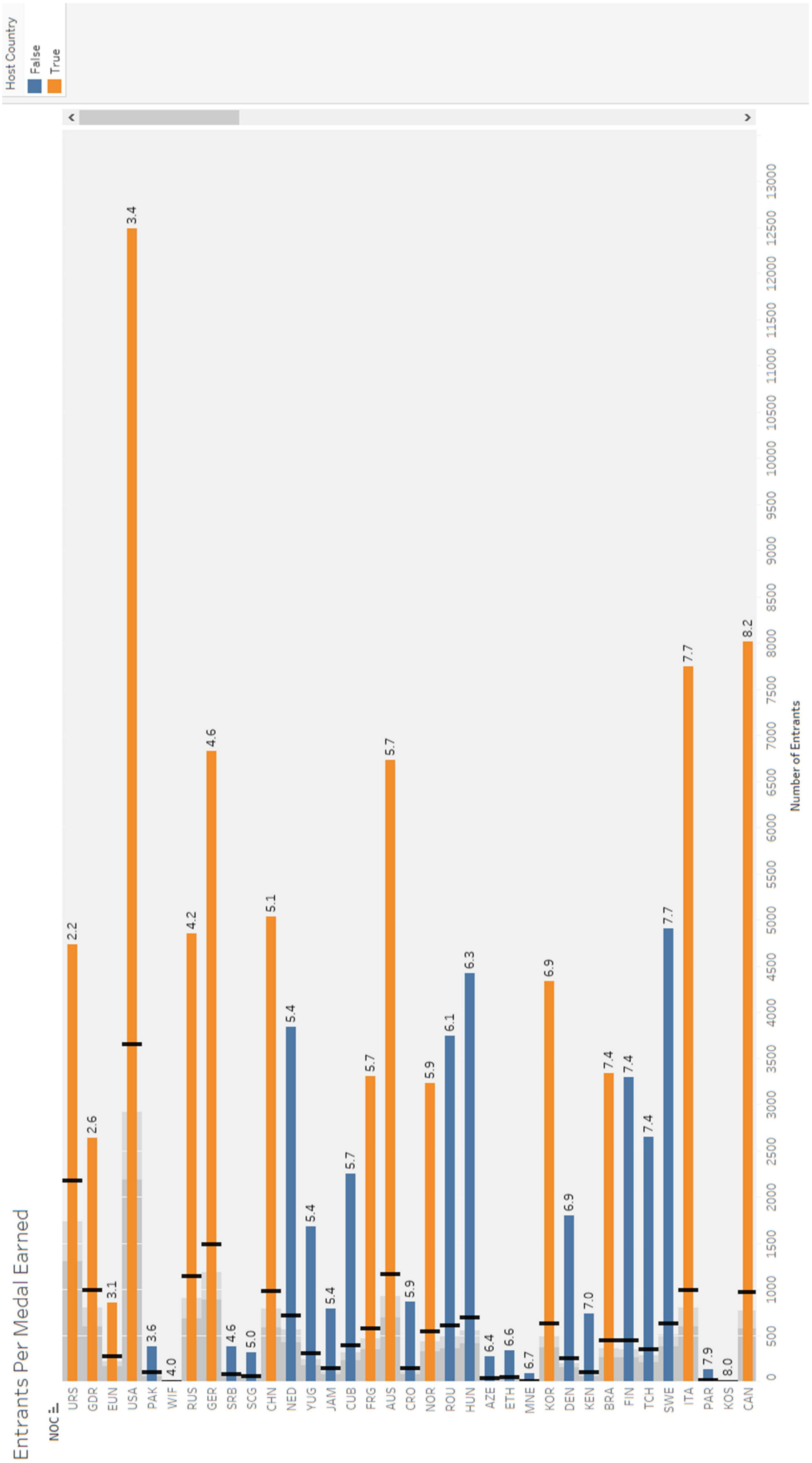
```
-- Testing what CAN had for number of athletes for 1960 summer
SELECT DISTINCT athlete_id, season, oly_year, noc
FROM athletes
WHERE oly_year = '1960' AND noc = 'CAN' AND season = 'Summer';

-- Testing number of athletes total for CAN
SELECT noc, COUNT(DISTINCT athlete_id)
FROM athletes
WHERE noc = 'CAN';
```

Appendix 8 – Entrant to Medals Visuals

First Bar Chart



Second Bar Chart



Appendix 9 – Dual Sport Athletes Analysis

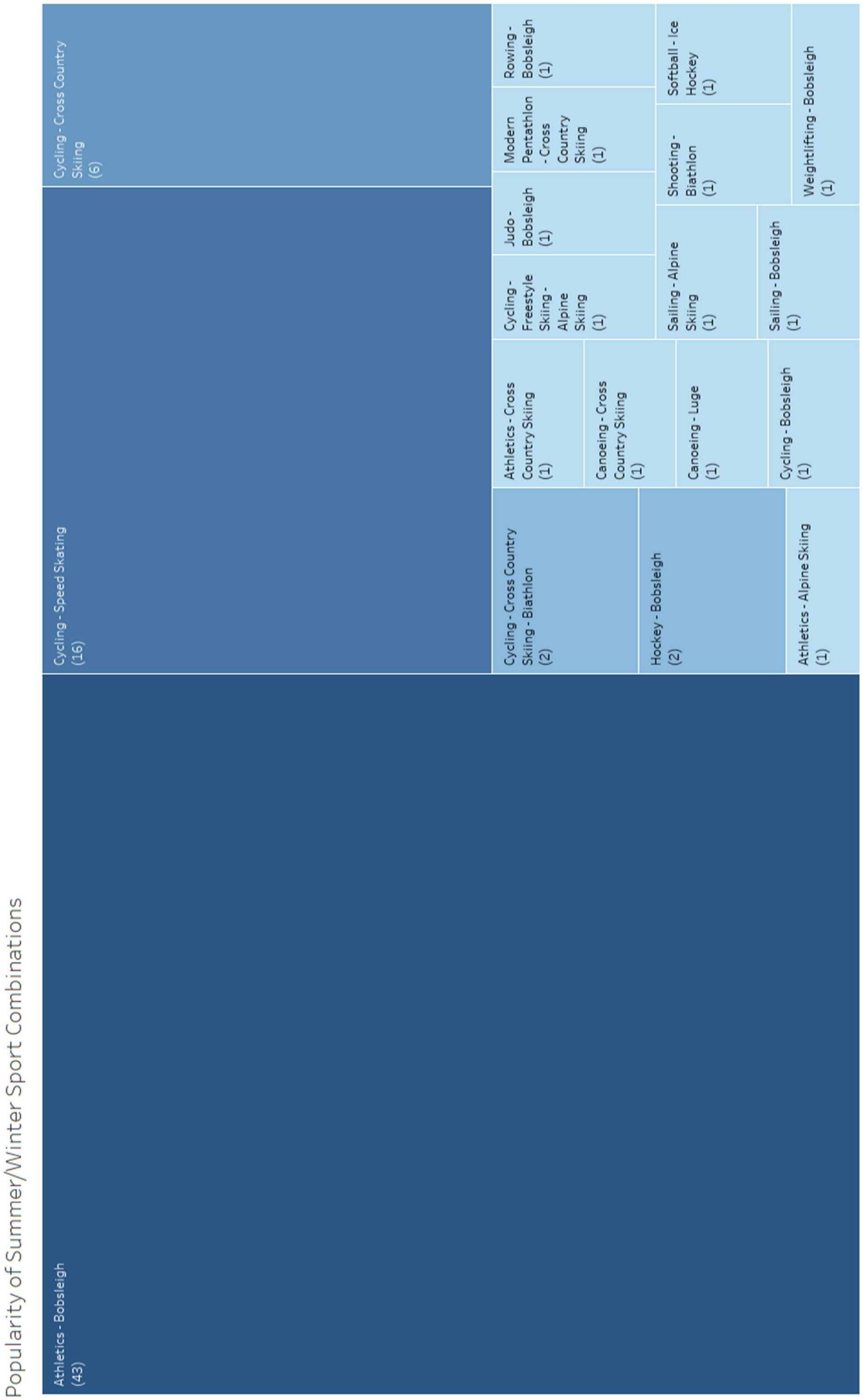
File: Q4_Dual_Sport_Athletes.sql

```
-- To get the count of athletes who were Summer AND Winter

SELECT DISTINCT a.athlete_id, a.athlete_name

FROM athletes a

WHERE athlete_id IN

        (SELECT athlete_id FROM athletes WHERE season = 'Summer')

                AND athlete_id IN

                        (SELECT athlete_id FROM athletes WHERE season = 'Winter')

ORDER BY a.athlete_id, a.athlete_name ASC;


--To get instances of when athlete was SUMMER AND WINTER

SELECT athlete_id, athlete_name, sport, oly_events

FROM athletes a

WHERE athlete_id IN

        (SELECT athlete_id FROM athletes WHERE season = 'Summer')

                AND athlete_id IN

                        (SELECT athlete_id FROM athletes WHERE season = 'Winter')

ORDER BY athlete_id;


--  SELECT per athlete what sport did they do in summer and winter (mostly pairs)

SELECT DISTINCT sport, athlete_id, athlete_name

FROM athletes a

WHERE athlete_id IN

        (SELECT athlete_id FROM athletes WHERE season = 'Summer')

                AND athlete_id IN

                        (SELECT athlete_id FROM athletes WHERE season = 'Winter')

ORDER BY athlete_id;


-- Table with id, name and sports concatted for xls analysis
```

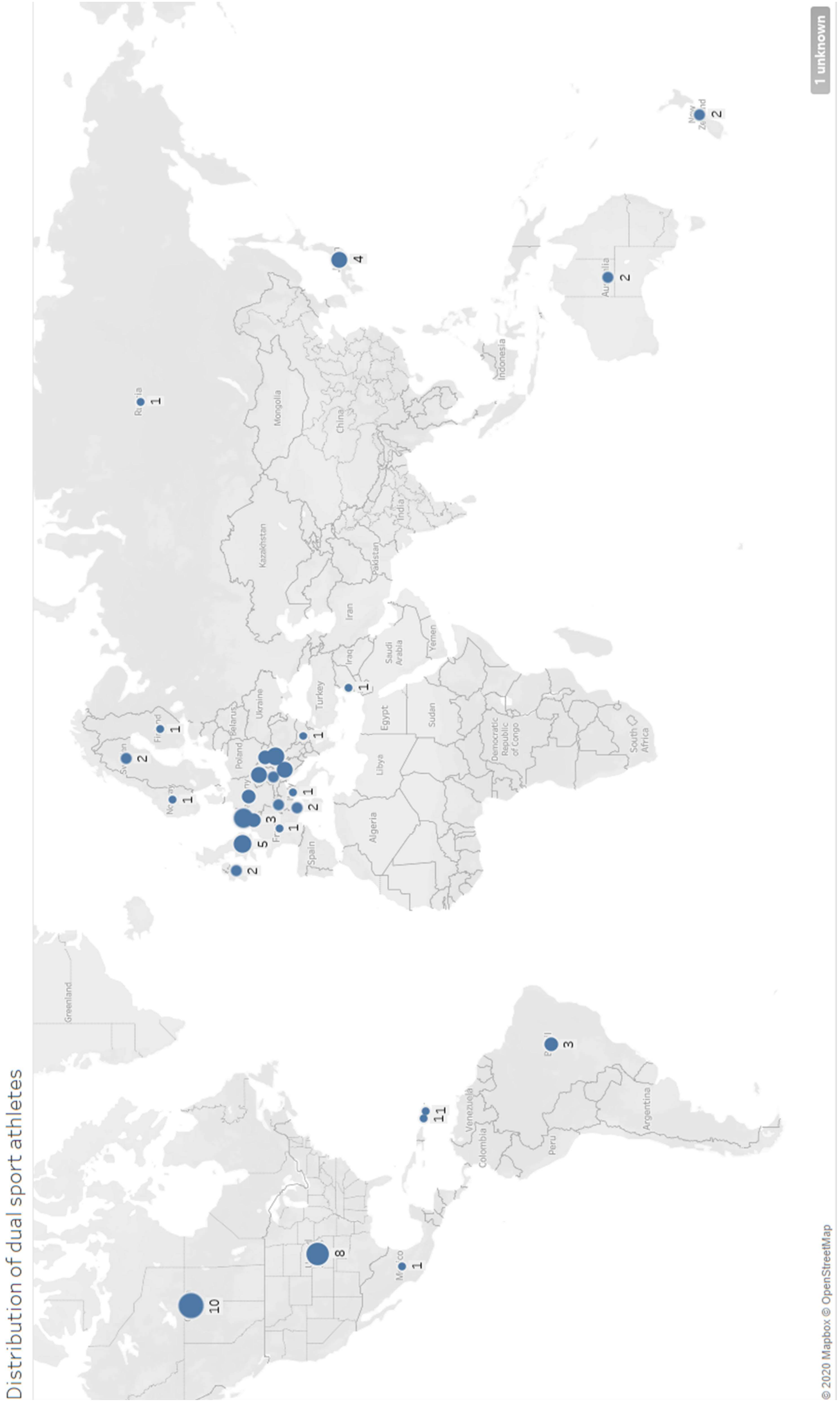
```
WITH discsports AS (
SELECT DISTINCT sport, athlete_id, athlete_name, athlete_country_code
FROM athletes a
WHERE athlete_id IN
        (SELECT athlete_id FROM athletes WHERE season = 'Summer')
                AND athlete_id IN
                        (SELECT athlete_id FROM athletes WHERE season = 'Winter')
ORDER BY athlete_id)
SELECT
        athlete_id,
        athlete_name,
        athlete_country_code,
       GROUP_CONCAT(sport) AS 'all_sports'
FROM discsports
GROUP BY athlete_id;
```

Appendix 10 – Dual Sport Athletes Visuals

Treemap Chart



Geographic Chart



Distribution of dual sport athletes

© 2020 Mapbox © OpenStreetMap