Edoardo Pastorino: 5169595, Kozy-Korpesh Tolep: 5302354

Project of Data Warehousing - Part 2

Task 1:

Hypotheses (Business Question):

1) Do countries with higher GDP / Capita perform better at the Olympics?
2) Is the number of women winners low in the Iraq, Iran, Afghanistan Nations?
3) How is the trend about the winning of medals in the United States over the years?
4) Is it true that the countries with higher populations perform better at the Olympics?
5) Do states that win a lot of medals in the summer olympics also win a lot of medals in the winter olympics?
6) What is the sport in which the Nordic European people (female and male) win more?
7) What are the ten best nations in terms of medal's number?
8) Is it true that the rich olympic nations have a higher number of athletes?
9) What is the sport with the max number of events in 2016 and in 1896?

Explored Datasets:

Dataset1: This is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016.

https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results/version/2?select=athlete_events.csv
The file athlete_events.csv contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event.
https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results/version/2?select=noc_regions.csv
This table of this dataset is composed by 3 columns and 230 rows:

1. **NOC**
2. **regions**
3. **notes**

Dataset2: This dataset includes a row for every Olympic athlete that has won a medal since the first games of 2020.

https://www.kaggle.com/amirba/olympic-sports-and-medals-18962021?select=summer.csv
This table is redundant because all the information is already in the events.csv, if you consider the data until 2016, by the way it is composed of 9 columns and 31200 rows.
https://www.kaggle.com/amirba/olympic-sports-and-medals-18962021?select=dictionary.csv
This table of this dataset is composed of 4 columns and 201 rows.

1. **Country**
2. **Code**
3. **Population**
4. **GDP Per Capita**

https://www.kaggle.com/amirba/olympic-sports-and-medals-18962021?select=winter.csv
This table is redundant because all the information is in the previous dataset, by the way it is composed of 3274 rows, and 9 columns.

Edoardo Pastorino: 5169595, Kozy-Korpesh Tolep: 5302354

Dataset3: This dataset includes a row for every Olympic athlete that has won a medal from the first games to 2012.

https://www.kaggle.com/the-guardian/olympic-games?select=summer.csv
This table is redundant because all the information is already in the events.csv, by the way it is composed of 9 columns and 31200 rows.
https://www.kaggle.com/the-guardian/olympic-games?select=dictionary.csv
This table of this dataset is composed of 4 columns and 201 rows.

1. **Country**
2. **Code**
3. **Population**
4. **GDP Per Capita**

https://www.kaggle.com/the-guardian/olympic-games?select=winter.csv
This table is redundant because all the information is already in the events.csv, by the way it is composed of 3274 rows, and 9 columns, and it is equal to the previous winter.csv except for the fact that this one stops at 2012.

Dataset4: The file olympic_hosts.csv contains all Olympic Game host city. Each row corresponds to an individual Olympic games and city.

https://www.kaggle.com/piterfm/olympic-games-hosts

Dataset5: https://data.world/johayes13/summer-winter-olympic-games

Selected Datasets:

Dataset1: We have chosen this one because it contains the most accurate information for the athletes and for the events, and it is used for answering almost all the business questions.

https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results/version/2

Dataset2: We have chosen this dataset because it contains the information for the GDP per capita and the population of the countries related to the nationality of the athletes, from 1896 to 2020, and for this reason is useful for the business question Nbr. 1, 3, 4 and 8.

https://www.kaggle.com/amirba/olympic-sports-and-medals-18962021

Dataset4: We have also selected this dataset because it contains all the info about the host of the olympic games and it is useful for responding to the business question number 8.

https://www.kaggle.com/piterfm/olympic-games-hosts

Dataset5: We have chosen this dataset because it stores a lot of additional information for each country that participated at the olympic games, and for example can be used for answering at the question number 5 and 6.

https://data.world/johayes13/summer-winter-olympic-games

Edoardo Pastorino: 5169595, Kozy-Korpesh Tolep: 5302354

Integration:

For integrating the different data sources we decided to use only the *nation* word for all the dataset and remove the use of *country,* the same for the usage of *code* (for example ITA for Italy) of a nation instead *NOC*. We decided to remove redundant and unuseful attributes (columns) for the analysis, and sometimes an entire table, in the various datasets, and also we have renamed attributes for readability's reasons. We have decided to use the syntax *Nbr.* for the attributes that represent a number of stuff. We have chosen to split one table, the one relative to the olympic events, in another two tables, for separating all the information that we need about the athletes and the events.
For doing the integration phase we have used the Tableau Prep application and we have obtained in output the CSV files relative to the integrated and transformed tables. A big problem was in the different number of nations that we have in the different tables, this is why we consider different temporal spans in tables. In the table relative to the host city the temporal span goes to 2028 and so we have eliminated manually (in the CSV file) the rows that refer to the years after 2016 to 2028. Another two problems were the different names of the cities, sometimes they were in English, sometimes in the language of the nation of belonging, and so we have unified the city's names, and also we have fixed some syntax problems in the names of the athletes.

Logical Schema

Olympic_events (Id_olympic_event, Games, Year, City, Medal**,** id_host^{Olympic_hosts}, athlete_name^{Olympic_athletes}**,** Id_athlete^{olympic_athlete}, Id_event^{event}, id_host_city^{olympic_host}**)**

Olympic_athletes(Id_athletes, Name, Sex, Team**,** Code^{Nation}, Id_nation^{Nation}**)**

Nations(Id_nations, Nation, Code, Population, GDP_capita)

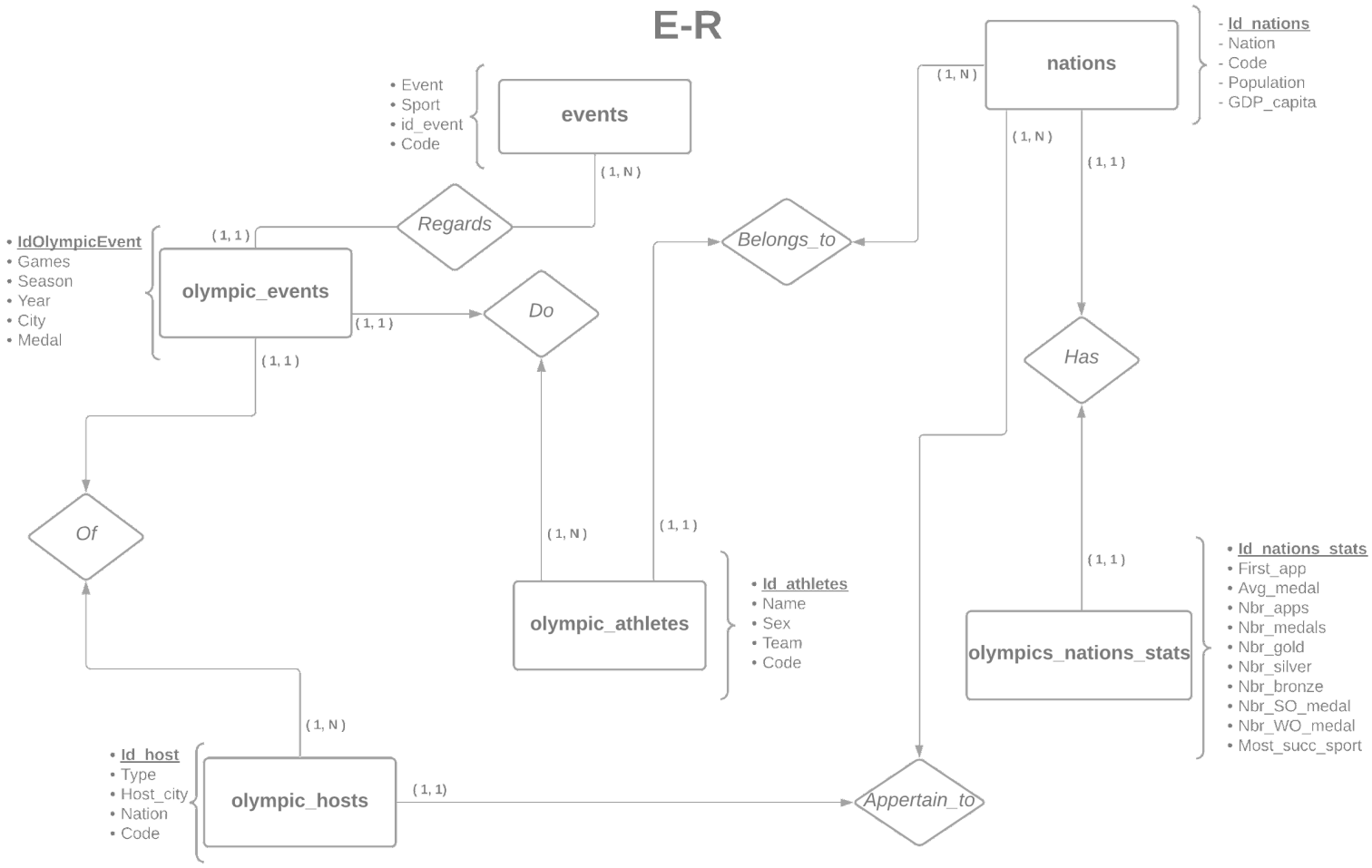
Olympic_nations_stats(Id_nations_stats, Name^{nations}, Code^{nations}, First_app, Avg_medal, Nbr_apps, Nbr_medals, Nbr_gold, Nbr_silver, Nbr_bronze, Most_succ_sport **)**

Olympic_hosts(Id_host, Type, Games_URL, Nbr_disciplines, Name^{Nations}, Start_date, End_date, Nbr_athletes, Nbr_countries, Nbr_events, Host_city, Year **)**

Event(Id_event**,** Sport, Event)

Edoardo Pastorino: 5169595, Kozy-Korpesh Tolep: 5302354

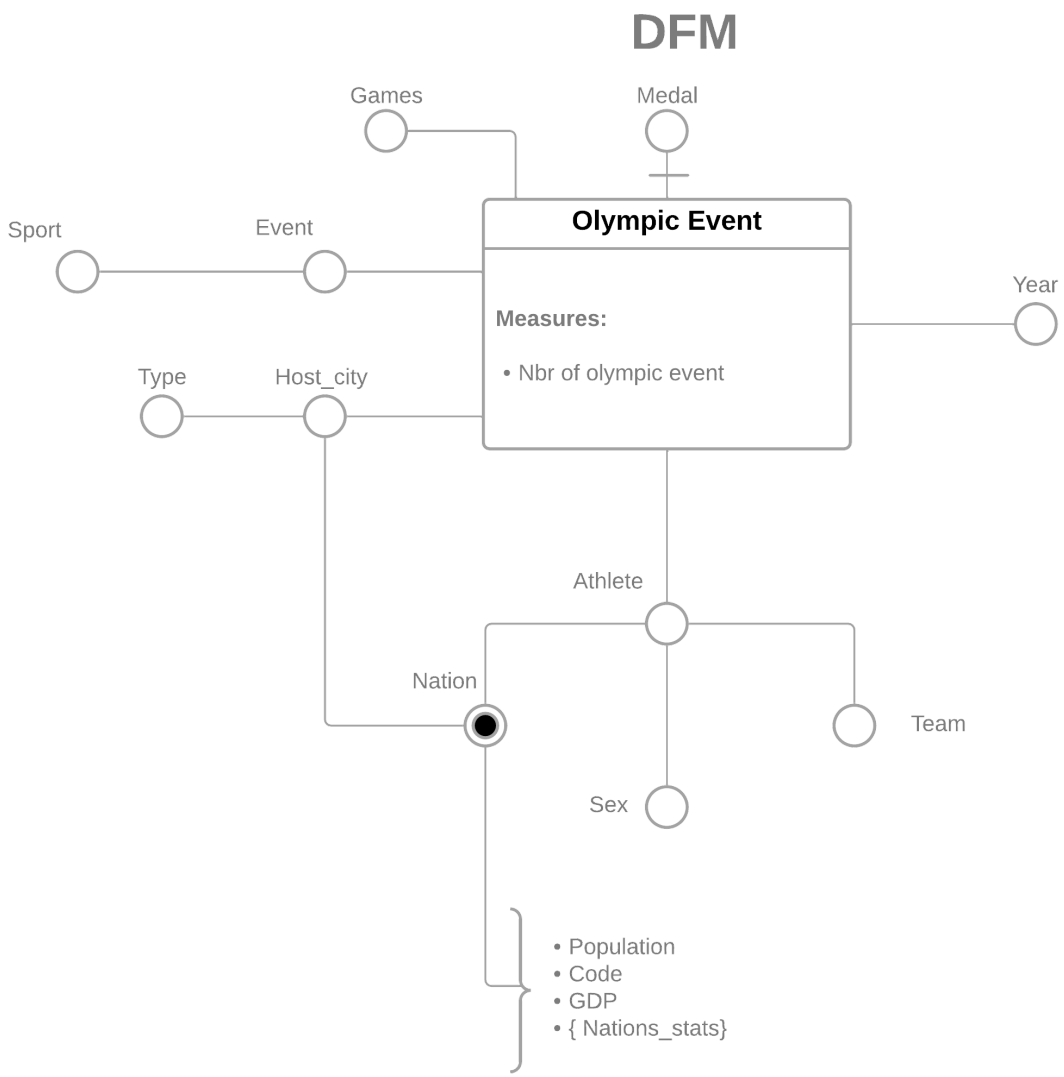
E-R



Task 2:

Edoardo Pastorino: 5169595, Kozy-Korpesh Tolep: 5302354

DFM - Conceptual Schema:



According to the DFM conceptual model we have identified the Olympic Events as the fact of interest in the analysis.
We have inserted only one measure, the classical count of instance of the fact, called the number of olympic events. This measure is additive and so aggregable over the sum function. We didn't find any other measures for the analysis because with the data in the chosen datasets there wasn't any information, any columns that seemed like measure.
We have introduced 6 dimensions: the event dimension (with these functional dependencies: event → sport). This dimension answers the question: what?; the host city dimension (with a functional dependence to the attribute type and another dependence to the attribute nation) that answers the question where?; the year dimension. This dimension responds to the when? question; the athlete dimension (linked to nation, team and sex with three dependencies) answering the who? question;The last dimension are the medal dimension and the games dimension, these are degenerate dimensions.

As regards the dynamicity in the dimensions we can say that there are few coarser level dimension attributes that can change, but for instance the sport relative to a certain event may change in the future because this event instead of belonging to "athletics" belongs to "gymnastic", or to a new sport. Another example is the changing of an athlete's nationality over time. We can interpret this dynamicity

in different ways, for our case we have decided to use the “yesterday or today” configuration and so analyze the event of the fact according to the configuration the hierarchies had at the time when the event occurred.

Workload:

Each component of this workload refers to the business questions, and it is written following the syntax related to the aggregation patterns used for the relative query.

1) *Olympic_participation(Nation.Nation, Nation.GDP, Nations.Nbr_medals)*
   For each nation shows the GDP and number of medals, ordered by GDP

2) *Olympic_participation(Athlete.name, Athlete.Nation, Event.Medal; Athlete.Sex = “ F ” & Athlete.Nation = “Iraq” & Athlete.Nation = “Iran” & Athlete.Nation = “Afghanistan”)*
   Count the number of medals of the female athletes that belong to Iraq, Iran and Afghanistan.

3) *Olympic_participation(Nation.Nbr_medal, Year.Year; Nation.Code = “USA”)*
   Select for each year the number of medals of the USA

4) *Olympic_participation(Nation.Nation, Nation.Population, Nations.Nbr_medals; Nations.Populations >= 80.000.000)*
   For the nations that have a population greater than 80 millions show the number of medals

5) *Olympic_participation(Nation.Nation,Nation.GDP,Nations.Nbr_SO_medal,Nations.Nbr_WO_medal; Nations.Nbr_SO_medal >= 100)*
   For each nations that has won more than 50 medals in the summer olympics show the medals relative to the winter olympics

6) *Olympic_participation(Nation.most_succ_sport; Athlete.code=”NOR” & Athlete.code=”SWE” & Athlete.code=”FIN” & Athlete.code=”DEN” & Athlete.code=”ISL” & Athlete.code=”IRL”)*
   Show the sport for which the athletes of the nordic states have won more medals

7) *Olympic_participation(Nation.Nations,Nation.nbr_medal)*
   Order by number of medals and select the ten best Nations and the relative number

8) *Olympic_participation(Nation.nation, Athlete.name; Nation.nation.GDP >= 20.000)*
   Select the nation that has a GDP greater than 20.000 and show the relative number of athletes.

9) *Olympic_participation(Event.Sport, Day.Year; Day.Year= “2016” & Day.Year=”1896”)*
   Count the number of each sport in the year 2016 and 1896 pick, for each of the two years, the sports with the largest number of events.

Data Volume:

- 271000 olympic events (number of events of the fact)
- 198 nations
- 50 Host city and 2 types
- 135000 athletes of 2 sex and 1184 teams
- 765 event, 66 sports
- 4 kind of medals

- 51 games
- 40 years, approximately 20 days of duration of each olympics

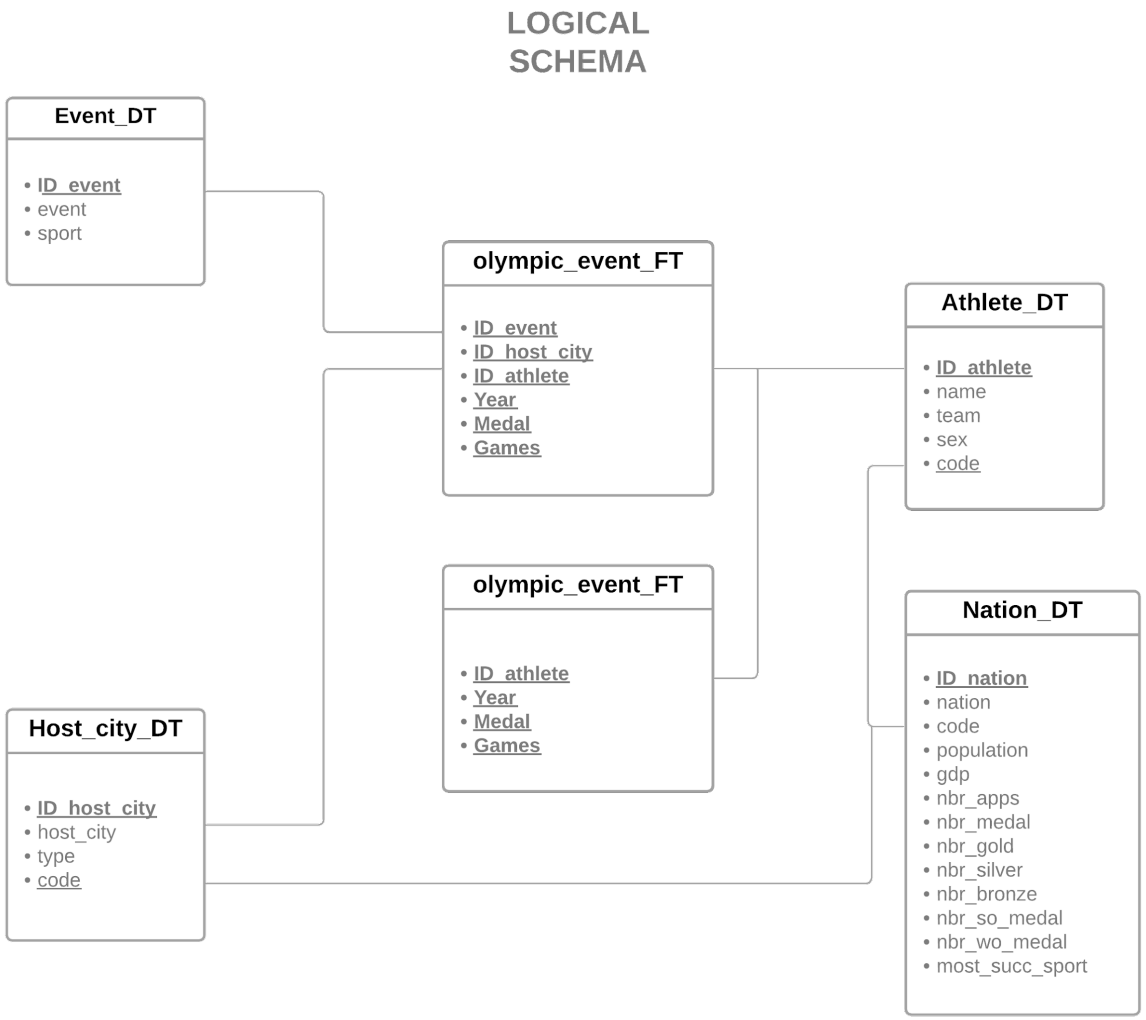
0) P0.(event, day, athlete, host_city) = 217 * 10^3
1) P1.(nation) = 198
2) P2.(athlete, nation, medal) = 135000 * 198* 3 = 801,90 * 10^5
3) P3.(nation, year) = (20*40)* 198=158 * 10^3
4) P4.(nation) = 198
5) P5.(nation) = 198
6) P6.(athlete, nation) = 198 * 135000 = 297,30 * 10^5
7) P7.(nation) = 198
8) P8.(athlete, nation)= 135000* 198 = 297,30 * 10^5
9) P9.(year, sport) = (20*40) * 66 = 52,8 * 10^3

Task 3:

Logical Schema

For the logical schema at first we have imported the CSV files in the tables in postgresql regarding the E-R Schema. For this part of the project we have created a CSV file with all the cities in the world with the relative nation and code of nation and thank to this we were able to insert the field code of the nation in the other tables that don't have it yet (by the use of the V-Lookup function of Excel) , this because the code of nation is the common column between all the tables and always with the V-Lookup function we have imported the id of the athlete table, event table and the host city table in the olympic event table for creating in Postgresql the foreign key constraints and especially for the further creation of the fact table.
After this stage we have filled the dimension tables based on the DFM Schema, created like a materialized view in Postgresql, starting by the tables relative to the E-R Schema. After we have also filled a primary fact table and a secondary fact table (a view) with this aggregation pattern: *medal, year, games, id_athlete*. We have decided to create this secondary fact table because the dimensional attributes athlete and nation are the most used. The degenerate dimensions are represented directly in the fact table and so we have chosen to not use a junk dimension table for storing them, for simplicity reasons and because the year and medal attribute are not too long. The logical schema in question is a snowflake schema, in particular we have chosen to snowflake in the nation attributes and so create a nation's secondary dimension table for managing the shared hierarchy nation between athlete and host_city, but also because the pattern with nation and so also athlete is used very often in the queries for answering the hypothesis. We want to advise that we have used the code (of nation) attribute like foreign key for linked the host city dimension table and the athlete dimension table instead of the primary key id_nation only for simplicity, because on the other and we have to put the id of nation inside two CSV files by the use of V-lookup function and this process may create problems.

Edoardo Pastorino: 5169595, Kozy-Korpesh Tolep: 5302354



Task 4:

Olap Queries

we have structured the olap queries in two directories, one divided in two subdirectories, one with the sql queries for the creation of the tables regarding the E-R schema and one with the queries for the creation of the fact and dimension tables, regarding the DFM schema. The other directory is also divided in two other subdirectories one with the queries about the workload and one with the queries about the group by extension and partitioning, ordering and framing.

Ex. of query for creating E-R's table

```
CREATE TABLE olympic_athletes
(
   ID_athlete serial PRIMARY KEY,
   Name varchar(256),
   Sex varchar(10),
```

```
    Team varchar(256),
    Code varchar(10)
);
```

Ex. of query for creating dimension table

```
CREATE MATERIALIZED VIEW Athlete_DT
(
    ID_athlete,
    Name,
    Team,
    Sex,
    Code

)

AS

SELECT
    "LogicalSchema_Project_DW".olympic_athletes.Id_athlete,
    "LogicalSchema_Project_DW".olympic_athletes.Name,
    "LogicalSchema_Project_DW".olympic_athletes.Team,
    "LogicalSchema_Project_DW".olympic_athletes.Sex,
    "LogicalSchema_Project_DW".olympic_athletes.Code
FROM
    "LogicalSchema_Project_DW".olympic_athletes
```

Ex. of workload query

```
SELECT nation, gdp, nbr_medal FROM  "LogicalSchema_Project_DW".nation_dt  where gdp is not null order by Gdp desc
```

Ex. of OLAP extension query

```
Select team, year, count(medal) as nbr_medal, RANK() over (PARTITION BY team ORDER BY count(medal) desc ) as ranking
FROM
    "LogicalSchema_Project_DW".olympic_events_ft2
    join "LogicalSchema_Project_DW".athlete_dt on
    "LogicalSchema_Project_DW".olympic_events_ft2.id_athlete = "LogicalSchema_Project_DW".athlete_dt.id_athlete
    where (medal != 'NA' and Team= 'Italy') or (medal != 'NA' and Team= 'Germany') or (medal != 'NA' and Team= 'France')
Group by (team,year)
```

Edoardo Pastorino: 5169595, Kozy-Korpesh Tolep: 5302354

Task 5:

Spark SQL

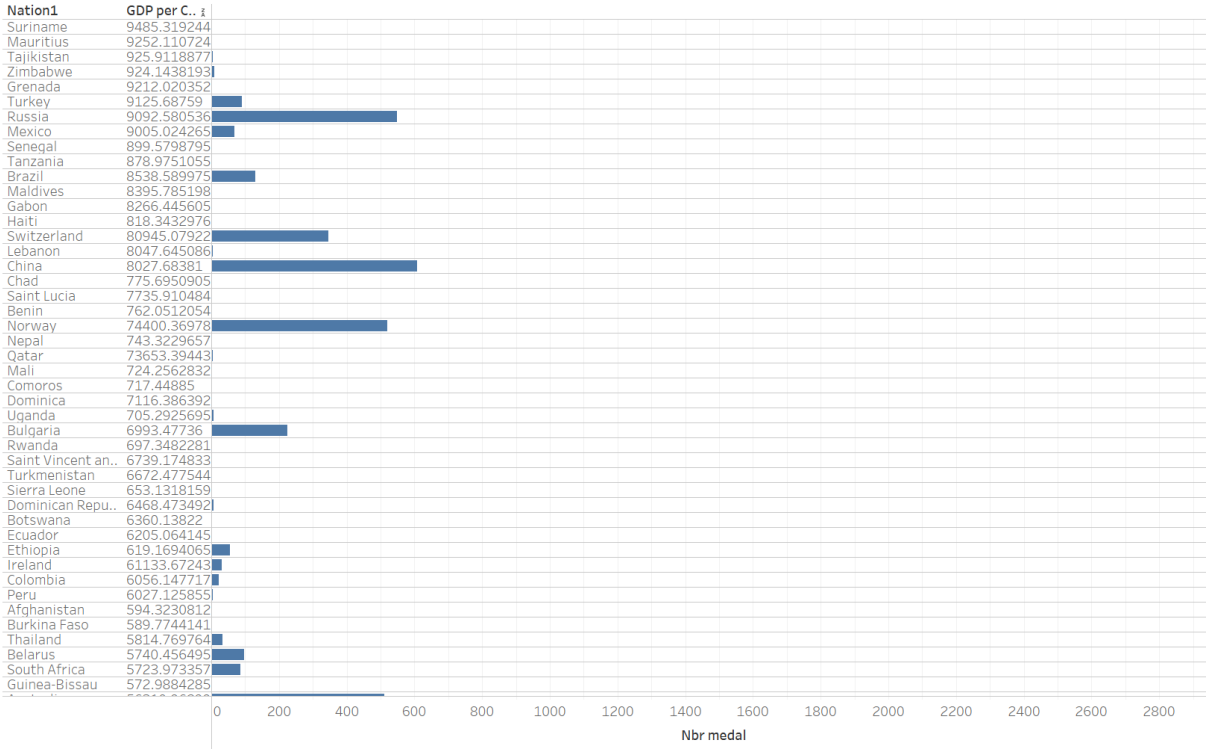
We specified olap queries as operations on our datasets. Firstly, we have converted them to dataframes and used the Spark Structured API to convert queries to methods that do the same operations with data. Below is the example of the 5th Workload question (*For each nations that has won more than 50 medals in the summer olympics show the medals relative to the winter olympics*). Spark queries that were used were written in the word file in the directory WorkloadPySpark_queries. If you want the results you can find them in Colab Notebook in the same folder.

```
nation_datasetDF.select('Nation', 'Nbr_SO_medal',
    'Nbr_WO_medal').filter(col('nbr_SO_medal')>='50').show()
```

So, in this example, 3 methods from the Spark Structured API: select(), filter() and show(). Data was already converted to dataframes. We have used select() method to leave out columns that we need, filter() method for filtering the data according to condition and show() method to illustrate the result.

Task 6:

Tableu Report



Report 1:

Do countries with higher GDP / Capita perform better at the Olympics?

Edoardo Pastorino: 5169595, Kozy-Korpesh Tolep: 5302354

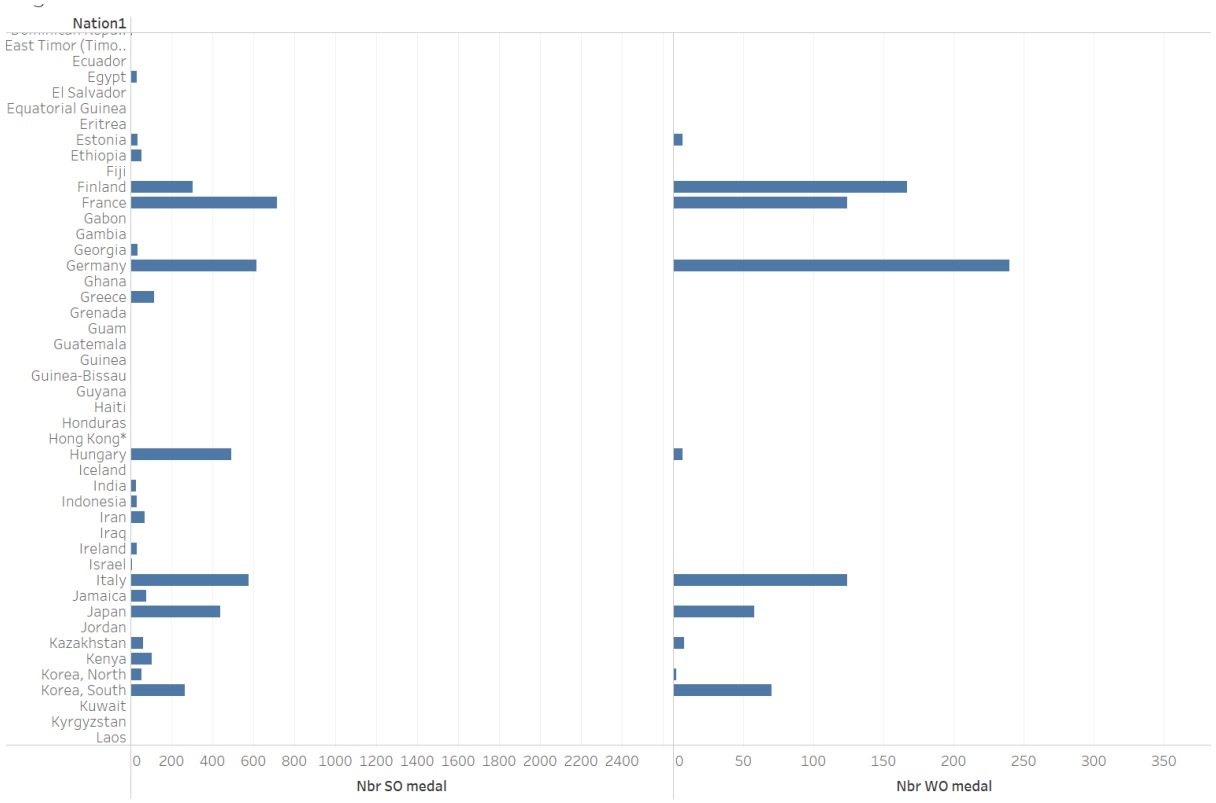
In this first Tableau report we try to show if there is a correlation between the GDP per capita of the nation and the number of medals won by each country from 1986 to 2016 and so show in some way if the richest countries are also the most winning countries or not. As the image (we have to precise that this image, and also the next ones similar to this one, are a partial representation of the entire graph, in which we shows the information for all the nations) shows there isn't a real correlation between these two parameters because sometimes it is true that a rich country is also a winning country, but sometimes this it isn't true and so we can't give a correct answer to the business question as well as saying that it depends from nation to nation.

Report 2:

How is the trend about the winning of medals in the United States over the years?

Like the graph of this second report shows us, we can say that the trend of winning medals in the United States, both for summer olympics and winter olympics is pretty good. In the years corresponding the olympic summer games in fact we almost always had a great number of medals, except for the two first olympic games, as expected for all the nations because the number of athletes and of sports were very low, and also for the year 1980 in which there was the boycott of all the olympic games of Mosca performed by the united states team. For what concerns the winter olympics games we can see that we have also had a very well performance in terms of medals and the trend during the years has an improvement.

Edoardo Pastorino: 5169595, Kozy-Korpesh Tolep: 5302354

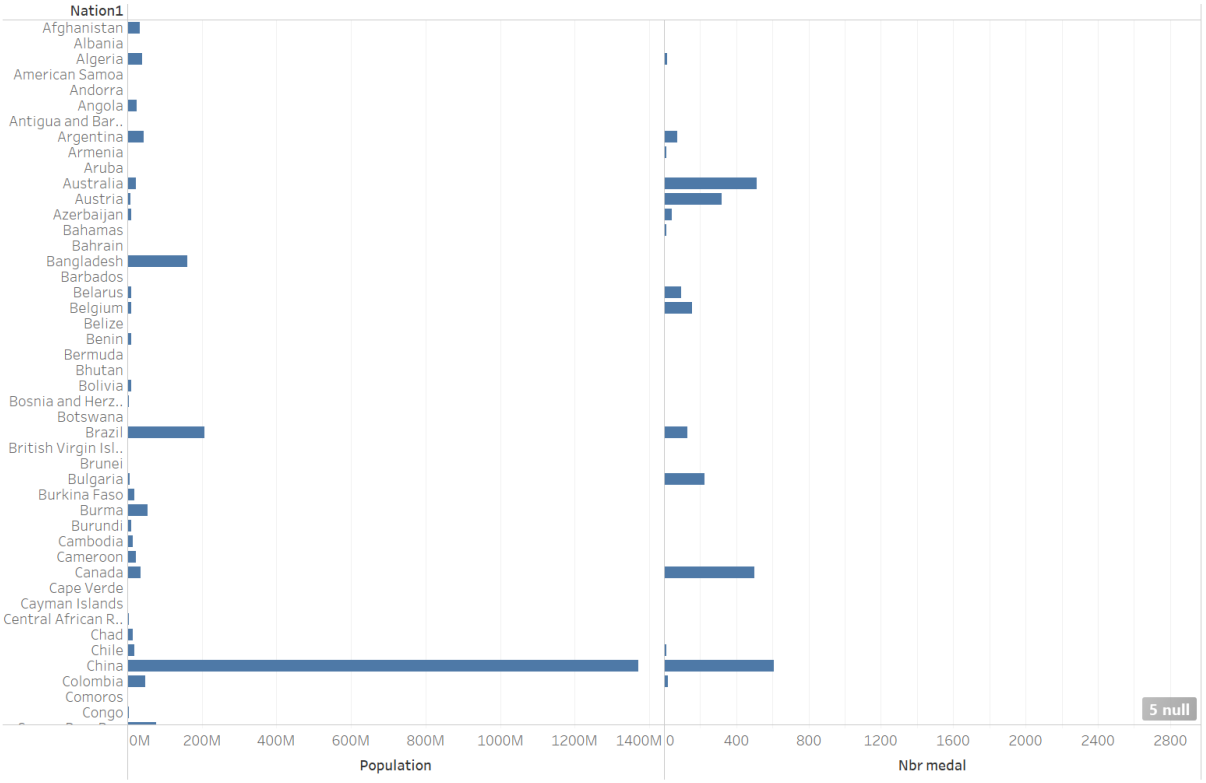


Report 3:

Do states that win a lot of medals in the summer olympics also win a lot of medals in the winter olympics?

The answer to this business question in this case is yes, in general the nations that are winning during the summer olympics games are also winning during the winter olympics games. However there are few cases in which some nations, especially in the north of Europe, have a larger number of medals regarding the winter olympics than the summer olympics, for instance Finland, Norway, Austria and Germany.

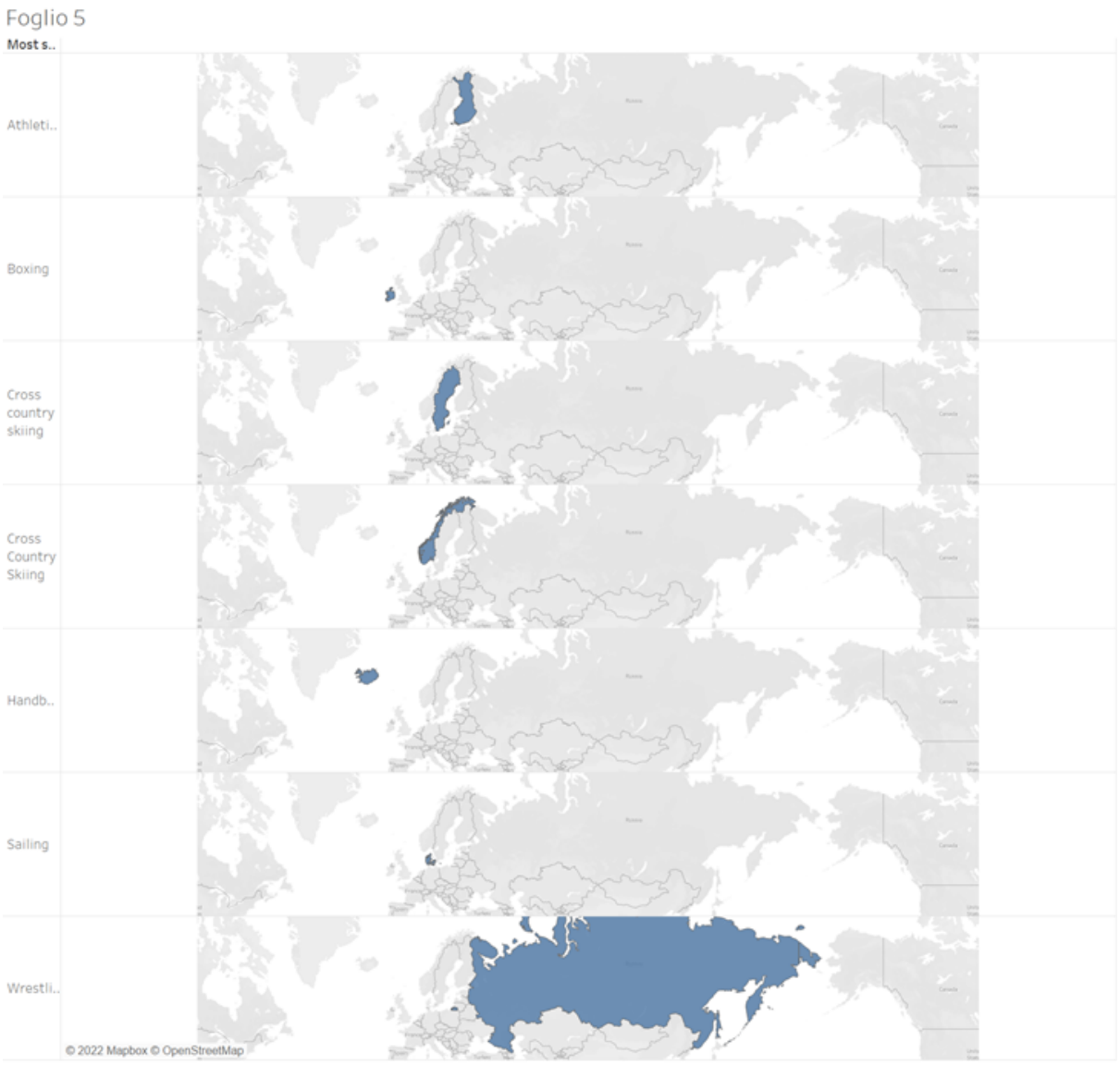
Edoardo Pastorino: 5169595, Kozy-Korpesh Tolep: 5302354



Report 4:

Is it true that the countries with higher populations perform better at the Olympics?

This fourth Tableau report allows us to see if there is another correlation between the number of medals and, this time, the population of the countries. Analyzing this graph we are able to see that often the nations, for example the european countries, that doesn't has a great number of population are winning in the olympics, on the other hand there are countries, like India and other Asian nations, that has a very large population but a very few number of medals.



Report 5:

What is the sport in which the Nordic European people (female and male) win more?
In this last report we want to simply show what are the most successful sports for the “Nordic European Countries”, that are for our opinion Sweden, Norway, Finland, Denmark, Ireland and Russia and Iceland. The outcomes of this report is that, contrary to what we expected, only two of these countries, Norway and Sweden, are better at winter sports.

Final Considerations:

Critical Analysis

About the software used for the integration phase, so the Tableau Prep application, we have noted that it isn’t intuitive the way for making the correspondent v-lookup function on the CSV, in the table of Tableau Prep and also we didn’t find a method for insert a serial number as the primary key of a table.

Edoardo Pastorino: 5169595, Kozy-Korpesh Tolep: 5302354

With Lucid Chart, the online tools for designing the schemas, PostgresSQL, the software for creating tables according to a DB and a DW, and Tableau Desktop, we had no particular difficulties.

Hours Effort

The effort for finishing the entire project in terms of hours is approximately 52 hours.
The first task, so the part relative to the choice of the business hypothesis, the exploration and selection of the datasets, the integration phase and the creation of the E-R schema and the relative logical schema was about 14 hours of work.
The second task, the one regarding the creation of the DFM, and the production of the workload and data volume was also about 12 hours of work.
The third task, the one about the design of the logical schema required 8 hours of work, instead the task number four in which we had to write OLAP SQL queries required approximately 6 hours of work.
The last two parts, task number five and six required respectively about 2 hours and 10 hours.