Project of Datawarehousing

Master degree in computer science - UNIGE - 2021/2022

Pastorino Edoardo
Kozy-Korpesh Tolep

Task 1: Inspection and Profiling

Hypothesis

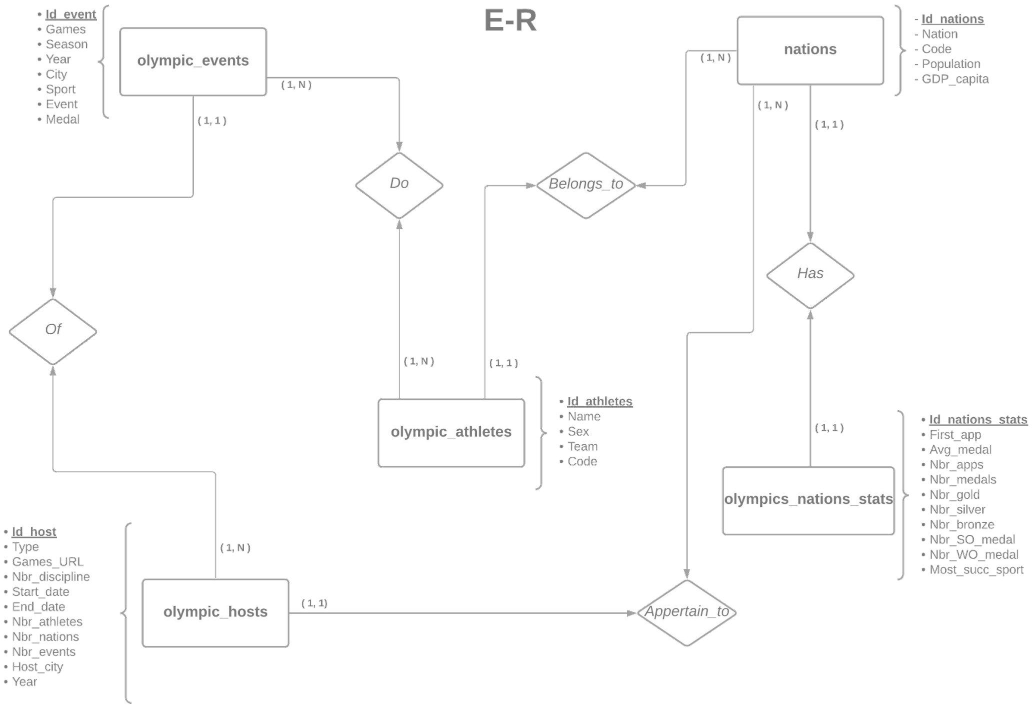
1) Do countries with higher GDP / Capita perform better at the Olympics?
2) Is the number of women winners low in the Iraq, Iran, Afghanistan Nations?
3) How is the trend about the winning of medals in the United States over the years?
4) Is it true that the countries with higher populations perform 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 average?
8) Is it true that the rich olympic host has a higher number of disciplines and athletes?
9) What is the sport with the max number of events in 2016 and in 1896?

Datasets

1) Dataset1: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results/version/2
2) Dataset 2: https://www.kaggle.com/amirba/olympic-sports-and-medals-18962021
3) Dataset3: https://www.kaggle.com/piterfm/olympic-games-hosts
4) Dataset4: https://data.world/johayes13/summer-winter-olympic-games

Integration and E-R Schema

For integrating the different data sources we decided to remove redundant and unuseful attributes (columns) for the analysis, and sometimes remove an entire table. We have also changed the name of the columns for avoiding mismatch.



E-R

olympic_events

- Id_event
- Games
- Season
- Year
- City
- Sport
- Event
- Medal

nations

- Id_nations
- Nation
- Code
- Population
- GDP_capita

olympic_athletes

- Id_athletes
- Name
- Sex
- Team
- Code

olympics_nations_stats

- Id_nations_stats
- First_app
- Avg_medal
- Nbr_apps
- Nbr_medals
- Nbr_gold
- Nbr_silver
- Nbr_bronze
- Nbr_SO_medal
- Nbr_WO_medal
- Most_succ_sport

olympic_hosts

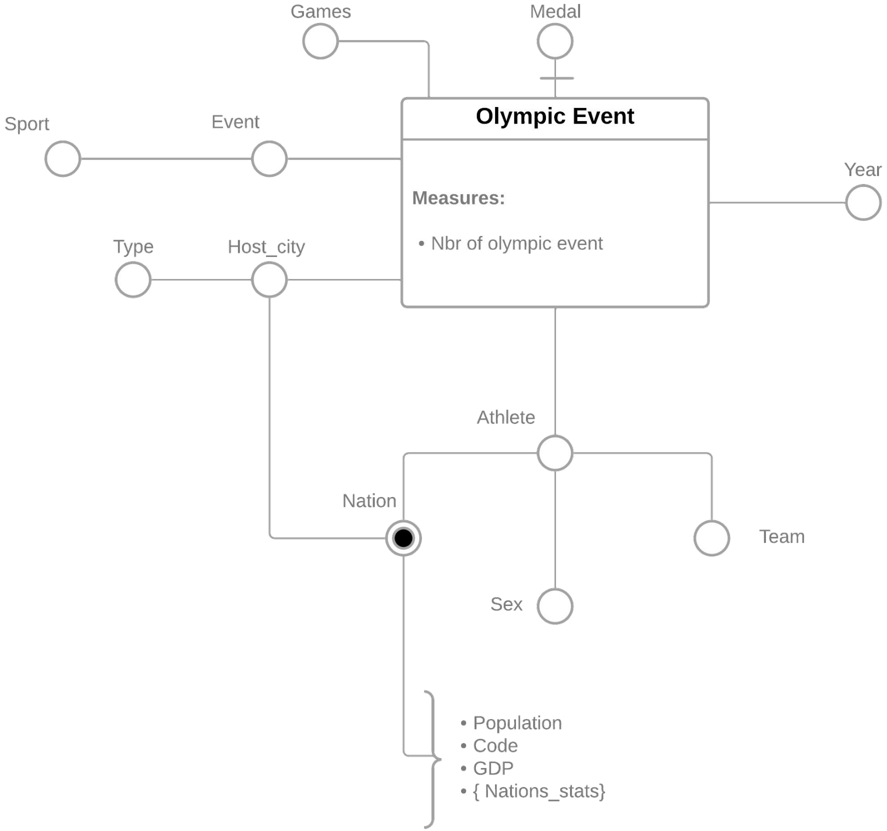
- Id_host
- Type
- Games_URL
- Nbr_discipline
- Start_date
- End_date
- Nbr_athletes
- Nbr_nations
- Nbr_events
- Host_city
- Year

Task 2: Conceptual Design

Dimensional Fact Model

According to the DFM conceptual model we have identified the Olympic Events as the fact of interest in the analysis.
We have inserted one measure, the classical count of instance of the fact.
We have identified 6 dimensions: Event, Athlete, Medal, Year, Host City and games.

DFM



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

. . .

Data Volume

271000 olympic events

201 nations

50 Host city and 2 types

135000 athletes of 2 sex and 1184 teams

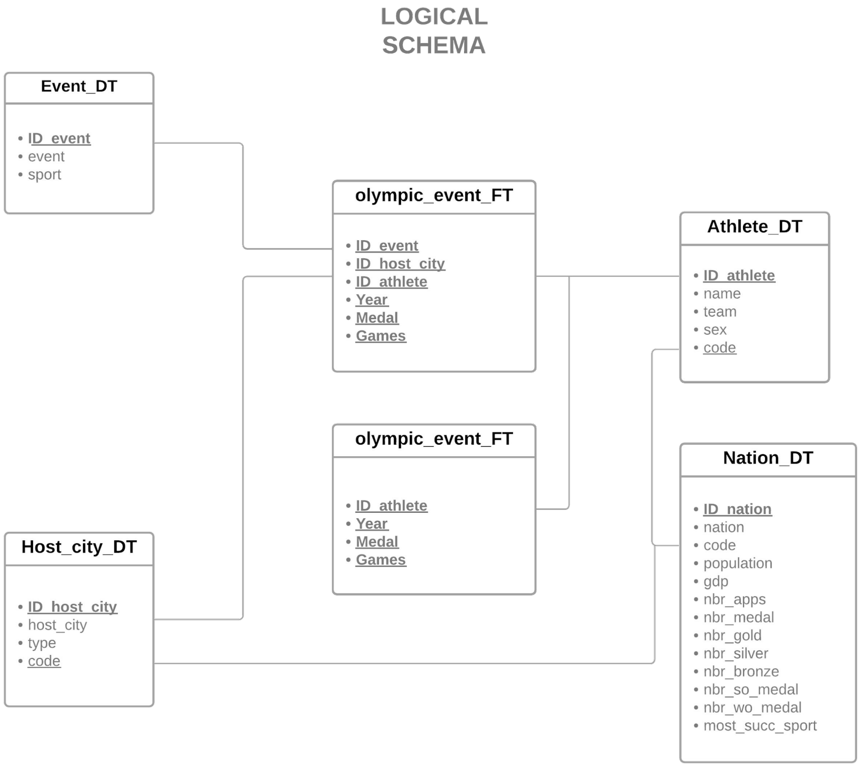
765 event, 3 kind of medals, 66 sports and 51 games

40 years, approximately 20 days of duration of each olympics

Task 3: Logical Design

Logical Schema

For the logical schema at first we have imported the csv files in the tables in postgresql regarding the E-R Schema.
After we have filled the dimension tables and primary fact table and a secondary fact table.
 The degenerate dimensions are represented directly in the fact table.
The logical schema in question is a snowflake schema



LOGICAL SCHEMA

Event_DT
- ID_event
- event
- sport

olympic_event_FT
- ID_event
- ID_host_city
- ID_athlete
- Year
- Medal
- Games

Athlete_DT
- ID_athlete
- name
- team
- sex
- code

olympic_event_FT
- ID_athlete
- Year
- Medal
- Games

Nation_DT
- ID_nation
- nation
- code
- population
- gdp
- nbr_apps
- nbr_medal
- nbr_gold
- nbr_silver
- nbr_bronze
- nbr_so_medal
- nbr_wo_medal
- most_succ_sport

Host_city_DT
- ID_host_city
- host_city
- type
- code

Task 4: Olap Queries

Queries for creating tables

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
```

Queries for the Workload and OLAP Extension

Example of workload query

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

Example of OLAP extension query

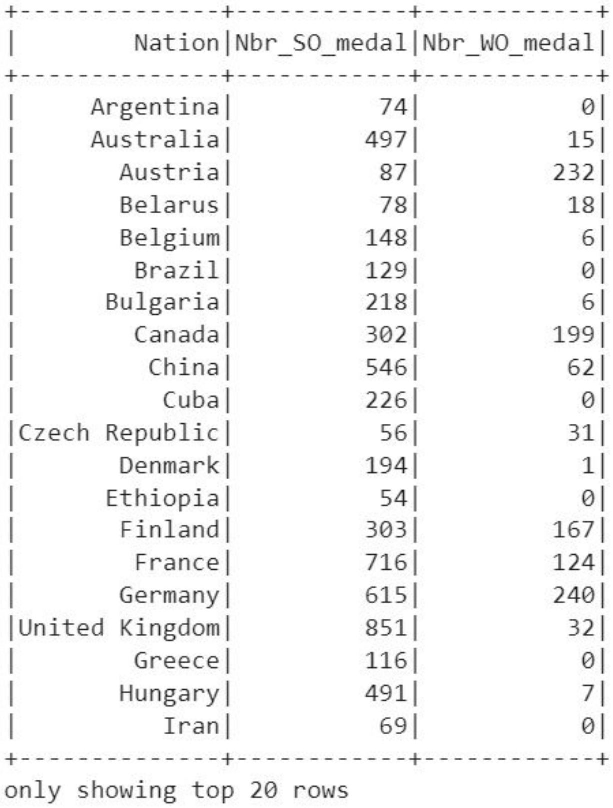
```
Select team, year, count(medal) as nbr_medal, RANK() over
(PARTITION BY team ORDER BY count(medal) desc ) as
ranking
FROM
   "LogicalSchema_Project_DW".olympic_events_ft2
   join "LogicalSchema_Project_DW".athlete_dt on

"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)
```

Task 5:Hive/SparkSQL

Example of workload query

```
data = datasetDF.select('Nation', 'GDP per Capita', 'Nbr_medal ').filter(col('GDP per Capita').isNotNull()).sort(desc('GDP per Capita')).show()
```

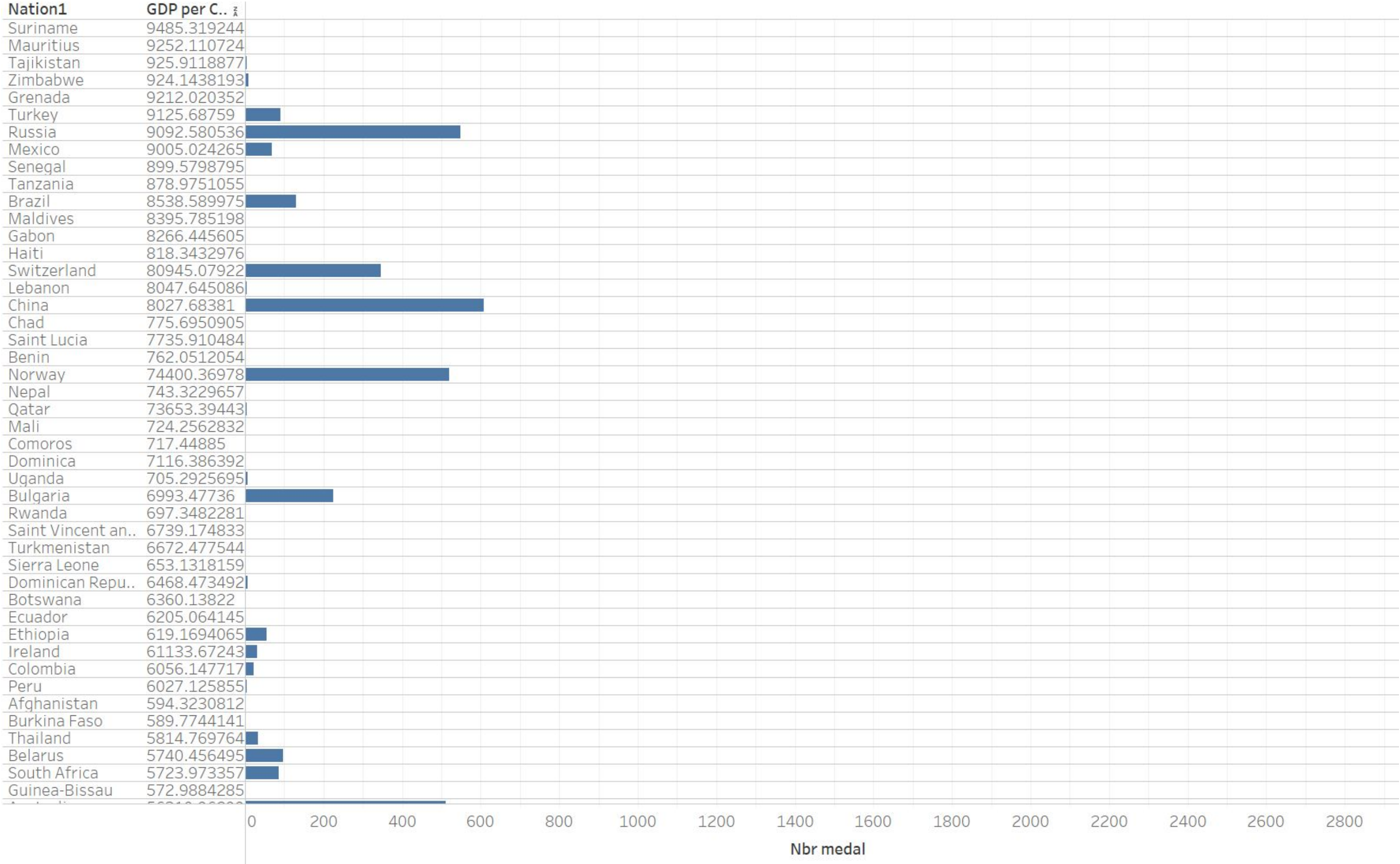


```
+--------------+------------+------------+
|        Nation|Nbr_SO_medal|Nbr_WO_medal|
+--------------+------------+------------+
|     Argentina|          74|           0|
|     Australia|         497|          15|
|       Austria|          87|         232|
|       Belarus|          78|          18|
|       Belgium|         148|           6|
|        Brazil|         129|           0|
|      Bulgaria|         218|           6|
|        Canada|         302|         199|
|         China|         546|          62|
|          Cuba|         226|           0|
|Czech Republic|          56|          31|
|       Denmark|         194|           1|
|      Ethiopia|          54|           0|
|       Finland|         303|         167|
|        France|         716|         124|
|       Germany|         615|         240|
|United Kingdom|         851|          32|
|        Greece|         116|           0|
|       Hungary|         491|           7|
|          Iran|          69|           0|
+--------------+------------+------------+
only showing top 20 rows
```

Result of the query

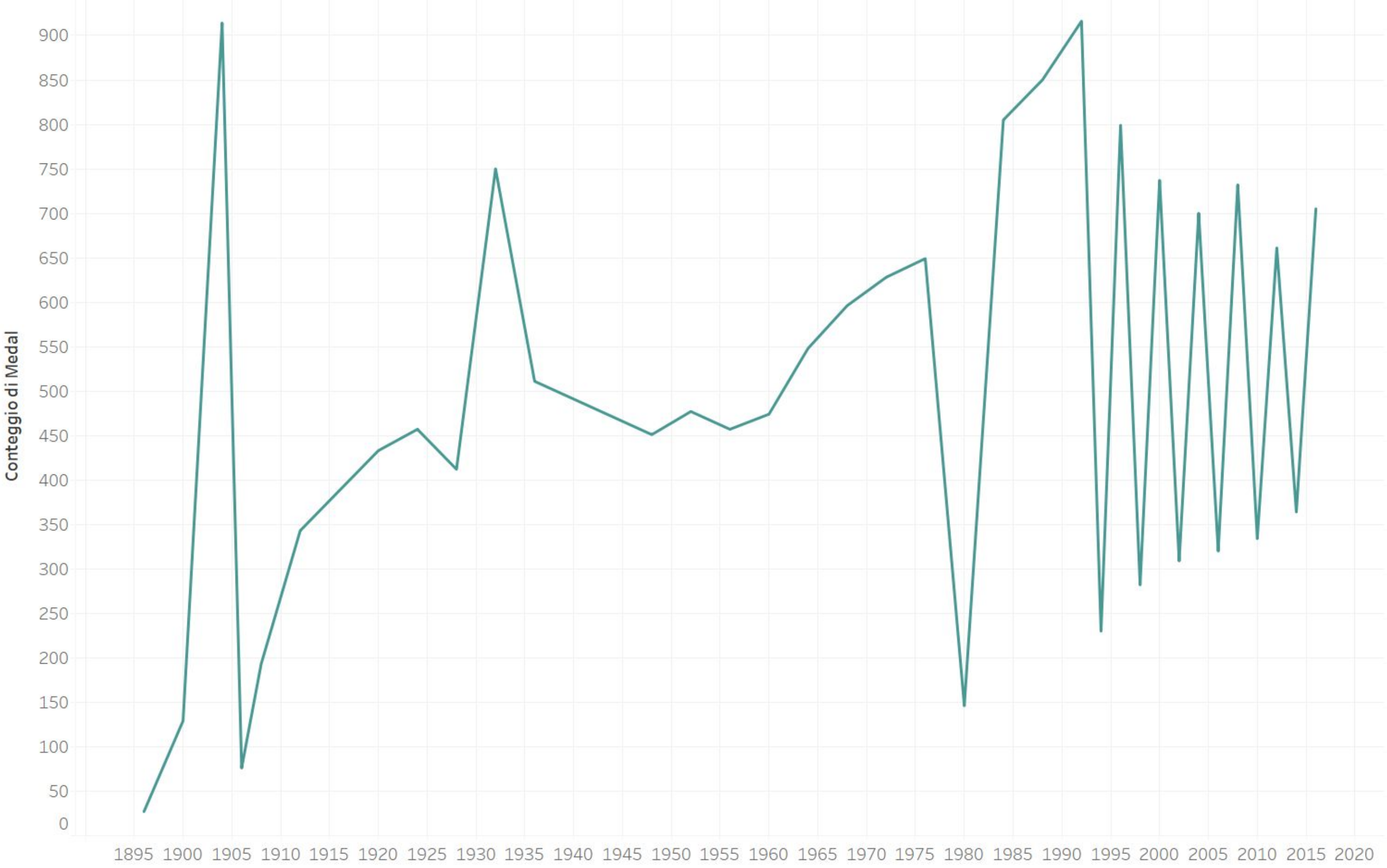
Task 6: Tableau Report

Report 1



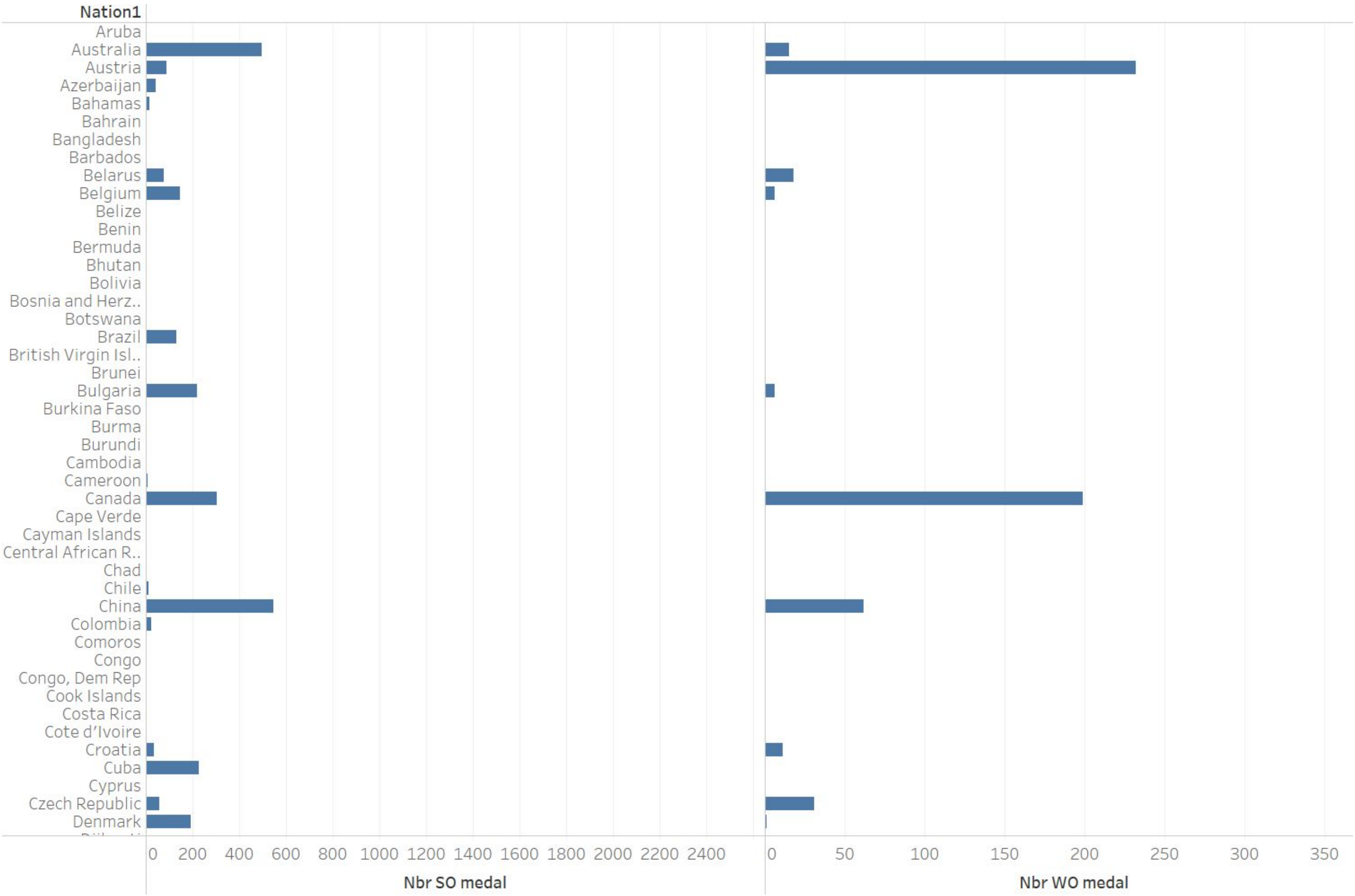
Nation1	GDP per C..
Suriname	9485.319244
Mauritius	9252.110724
Tajikistan	925.9118877
Zimbabwe	924.1438193
Grenada	9212.020352
Turkey	9125.68759
Russia	9092.580536
Mexico	9005.024265
Senegal	899.5798795
Tanzania	878.9751055
Brazil	8538.589975
Maldives	8395.785198
Gabon	8266.445605
Haiti	818.3432976
Switzerland	80945.07922
Lebanon	8047.645086
China	8027.68381
Chad	775.6950905
Saint Lucia	7735.910484
Benin	762.0512054
Norway	74400.36978
Nepal	743.3229657
Qatar	73653.39443
Mali	724.2562832
Comoros	717.44885
Dominica	7116.386392
Uganda	705.2925695
Bulgaria	6993.47736
Rwanda	697.3482281
Saint Vincent an..	6739.174833
Turkmenistan	6672.477544
Sierra Leone	653.1318159
Dominican Repu..	6468.473492
Botswana	6360.13822
Ecuador	6205.064145
Ethiopia	619.1694065
Ireland	61133.67243
Colombia	6056.147717
Peru	6027.125855
Afghanistan	594.3230812
Burkina Faso	589.7744141
Thailand	5814.769764
Belarus	5740.456495
South Africa	5723.973357
Guinea-Bissau	572.9884285

Report 2

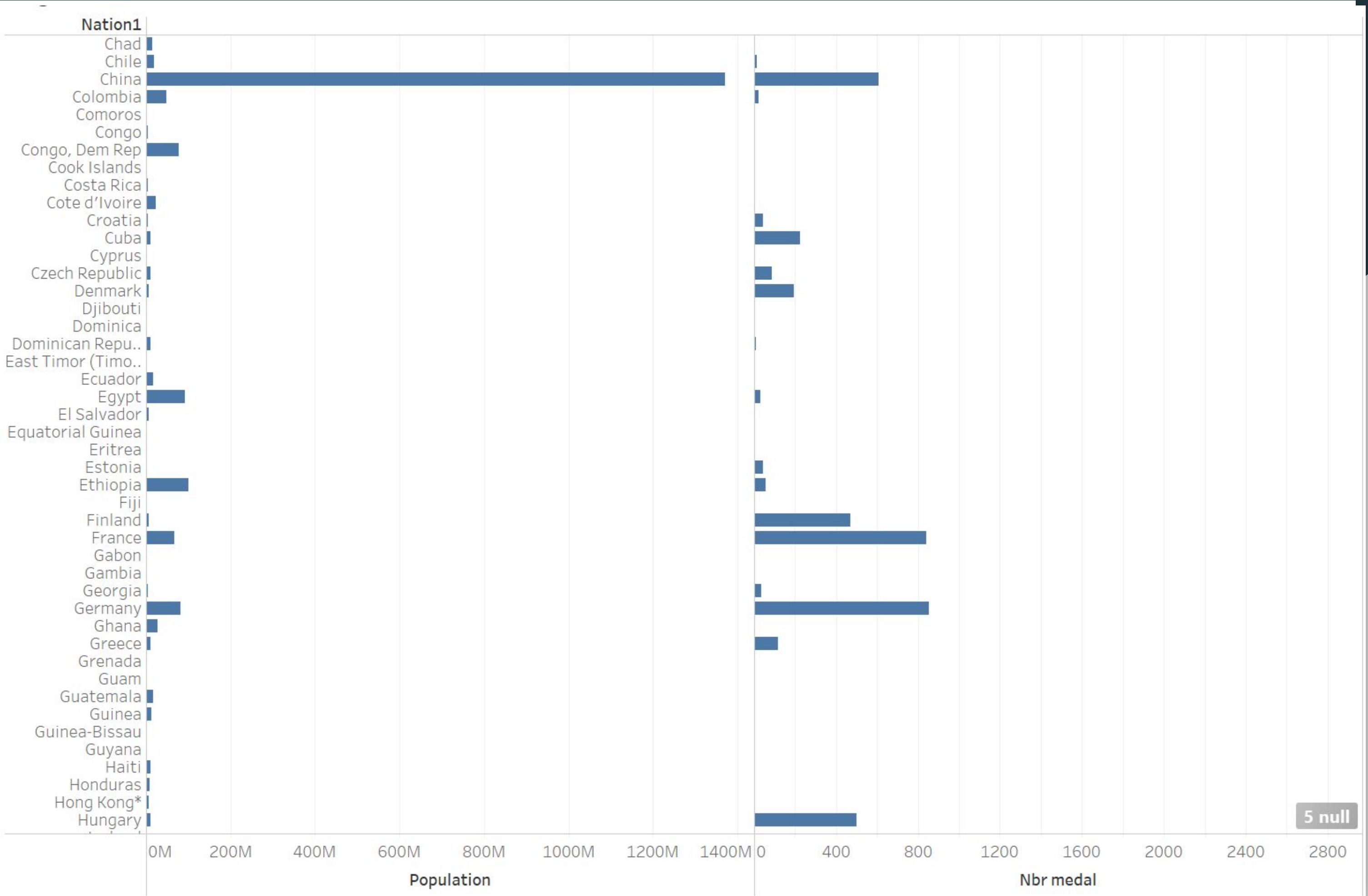


Conteggio di Medal
900
850
800
750
700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
1895
1900
1905
1910
1915
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020

Report 3


Nation1
Aruba
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Bangladesh
Barbados
Belarus
Belgium
Belize
Benin
Bermuda
Bhutan
Bolivia
Bosnia and Herz..
Botswana
Brazil
British Virgin Isl..
Brunei
Bulgaria
Burkina Faso
Burma
Burundi
Cambodia
Cameroon
Canada
Cape Verde
Cayman Islands
Central African R..
Chad
Chile
China
Colombia
Comoros
Congo
Congo, Dem Rep
Cook Islands
Costa Rica
Cote d'Ivoire
Croatia
Cuba
Cyprus
Czech Republic
Denmark
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
Nbr SO medal
0
50
100
150
200
250
300
350
Nbr WO medal

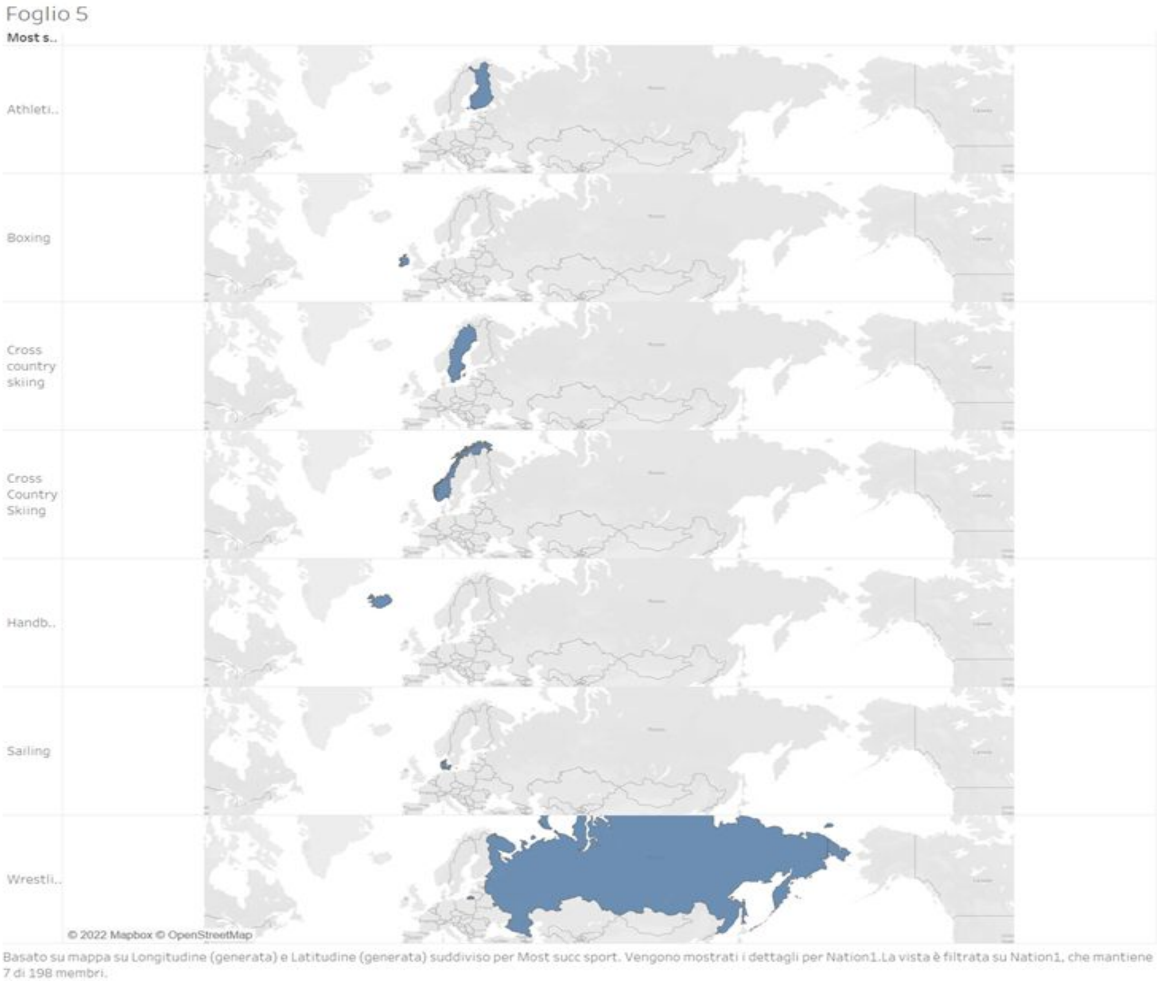
Report 4



Nation1
Chad
Chile
China
Colombia
Comoros
Congo
Congo, Dem Rep
Cook Islands
Costa Rica
Cote d'Ivoire
Croatia
Cuba
Cyprus
Czech Republic
Denmark
Djibouti
Dominica
Dominican Repu..
East Timor (Timo..
Ecuador
Egypt
El Salvador
Equatorial Guinea
Eritrea
Estonia
Ethiopia
Fiji
Finland
France
Gabon
Gambia
Georgia
Germany
Ghana
Greece
Grenada
Guam
Guatemala
Guinea
Guinea-Bissau
Guyana
Haiti
Honduras
Hong Kong*
Hungary
0M
200M
400M
600M
800M
1000M
1200M
1400M
Population
0
400
800
1200
1600
2000
2400
2800
Nbr medal
5 null

Report 5

Foglio 5



Basato su mappa su Longitudine (generata) e Latitudine (generata) suddiviso per Most succ sport. Vengono mostrati i dettagli per Nation1.La vista è filtrata su Nation1, che mantiene 7 di 198 membri.

Final Considerations

Critical Analysis

About 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.

Hours Effort

The effort for finishing the entire project in terms of hours is approximately 50 hours.

The first task was about 14 hours of work.

The second task was also about 12 hours of work.

The third task required 8 hours, instead the task number four required approximately 6 hours of work.

The last part, the task number six required about 10 hours.

Organization of the work

We have tried to split the amount of work based on our engagements (other exam and personal appointment). We often spoke in chat and occasionally also via meetings. We have work on shared files and directories in Google Drive.

The first task, relative to the inspection and profiling of the data, it has been done by Edoardo Pastorino and checked and correct by Kozy-Korpesh Tolep, the same for the task number two, relative to conceptual design and task number four, the one relative to the OLAP Sql queries.

Instead the 5th task has been done by Kozy-Korpesh Tolep alone and only after checked by Edoardo Pastorino.

The task number three and six, the final task, they have been designed and done together.

THE END

THANK YOU FOR YOUR ATTENTION