Milestone 1 report

SQL for Data Science Capstone Project

Justas Mundeikis
mundeikis@gmx.de

2023-02-18

Contents

1 Step 1: Preparing for Your Proposal

Which client/dataset did you select and why?

I decided to go with the Olympics Dataset, as I am interested in how did my country men perform in the past compared to other nations representatives. Especially I am interested in analysing the basketball players data on they physique.

Describe the steps you took to import and clean the data.

First I load all required libraries in R (using RStudio)

```r
## loading libraries required
library(tidyverse)
library(RSQLite)
library(DBI)
```

I have downloaded the data as .csv files and put in my working directory, in a sub directory "data". Then I read in the .csv files as dataframes:

```r
## reading in files
athlete_events <- read_csv(file = "data/athlete_events.csv")
noc_regions <- read_csv(file = "data/noc_regions.csv")
```

Then using R I have created a temporal local SQLite database.

```r
## creating a SQLite db in memory
con <- dbConnect(RSQLite::SQLite(),":memory:")
```

Then I have saved the dataframe as a SQLite DB

```r
## deleting if exists
dbRemoveTable(con, "athlete_events")
```

```
Error: no such table: athlete_events
```

```r
dbRemoveTable(con, "noc_regions")
```

```
Error: no such table: noc_regions
```

```r
## saving the dataframe as table
dbWriteTable(con, "athlete_events", athlete_events, overwrite=TRUE)
dbWriteTable(con, "noc_regions", noc_regions, overwrite=TRUE)
```

I can verify that tables exists with:

```r
## saving the dataframe as table
## dbReadTable(con, "athlete_events")
```

I can list columns:

```r
## list name of columns
dbListFields(con, "athlete_events")
```

```
 [1] "ID"      "Name"    "Sex"     "Age"     "Height" "Weight" "Team"    "NOC"
 [9] "Games"   "Year"    "Season"  "City"    "Sport"  "Event"  "Medal"
```

and in coc_regions table:

```r
## list name of columns
dbListFields(con, "noc_regions")
```

```
[1] "NOC"    "region" "notes"
```

Now we can write SQL queries:

```r
## list name of columns
dbGetQuery(con,
"select *
from athlete_events
limit 5")
```

```
  ID                     Name Sex Age Height Weight           Team NOC
1  1                A Dijiang   M  24    180     80          China CHN
2  2                 A Lamusi   M  23    170     60          China CHN
3  3      Gunnar Nielsen Aaby   M  24     NA     NA        Denmark DEN
4  4     Edgar Lindenau Aabye   M  34     NA     NA Denmark/Sweden DEN
5  5 Christine Jacoba Aaftink   F  21    185     82    Netherlands NED
        Games Year Season      City         Sport
1 1992 Summer 1992 Summer Barcelona    Basketball
2 2012 Summer 2012 Summer    London          Judo
3 1920 Summer 1920 Summer Antwerpen      Football
4 1900 Summer 1900 Summer     Paris    Tug-Of-War
5 1988 Winter 1988 Winter   Calgary Speed Skating
                             Event Medal
1      Basketball Men's Basketball  <NA>
2     Judo Men's Extra-Lightweight  <NA>
3          Football Men's Football  <NA>
4      Tug-Of-War Men's Tug-Of-War  Gold
5 Speed Skating Women's 500 metres  <NA>
```

```r
## list name of columns
dbGetQuery(con,
"select *
from noc_regions
limit 5")
```

```
  NOC      region                  notes
1 AFG Afghanistan                   <NA>
2 AHO     Curacao Netherlands Antilles
3 ALB     Albania                   <NA>
4 ALG     Algeria                   <NA>
5 AND     Andorra                   <NA>
```

ERD looks like

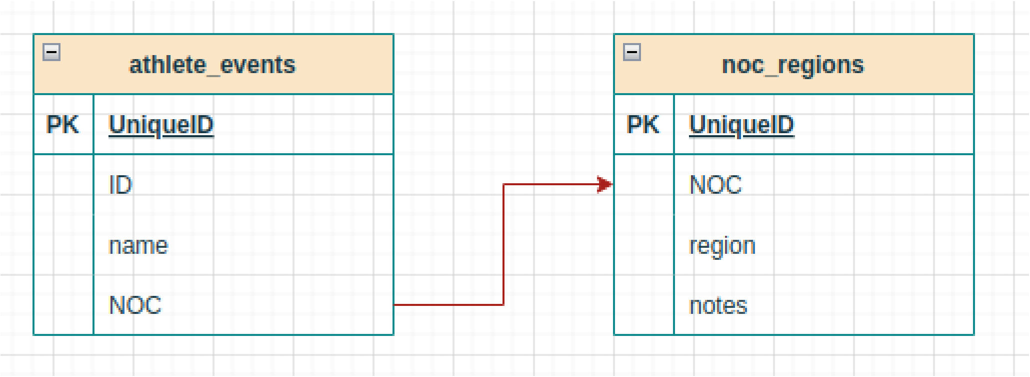
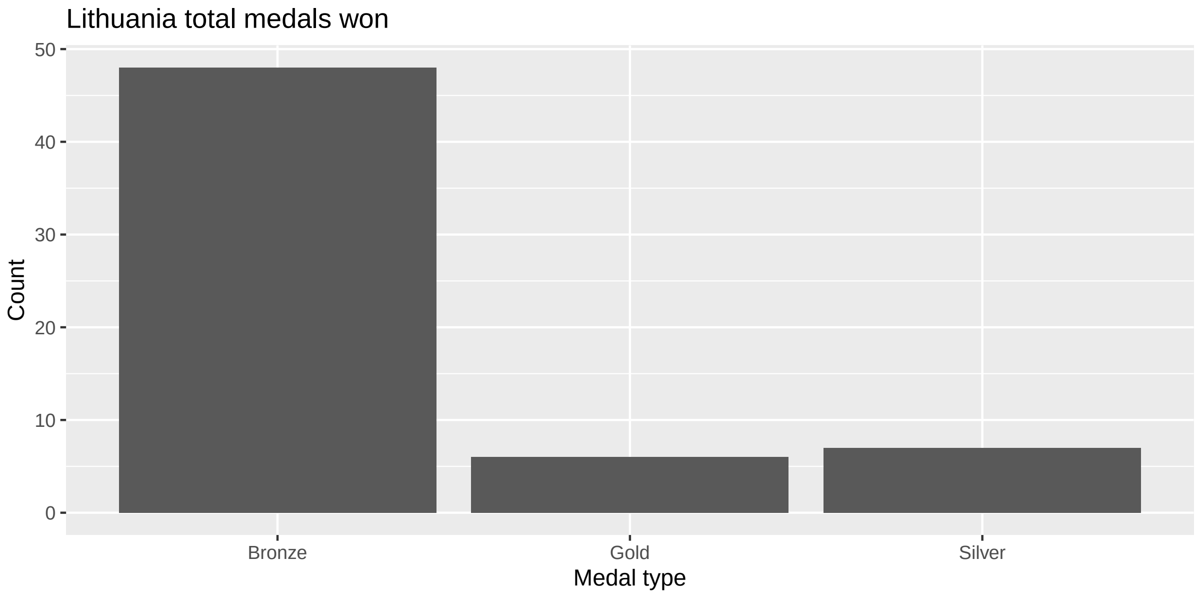


Figure 1: Entity Relationship Diagram (ERD)

2 Exploratory data analysis

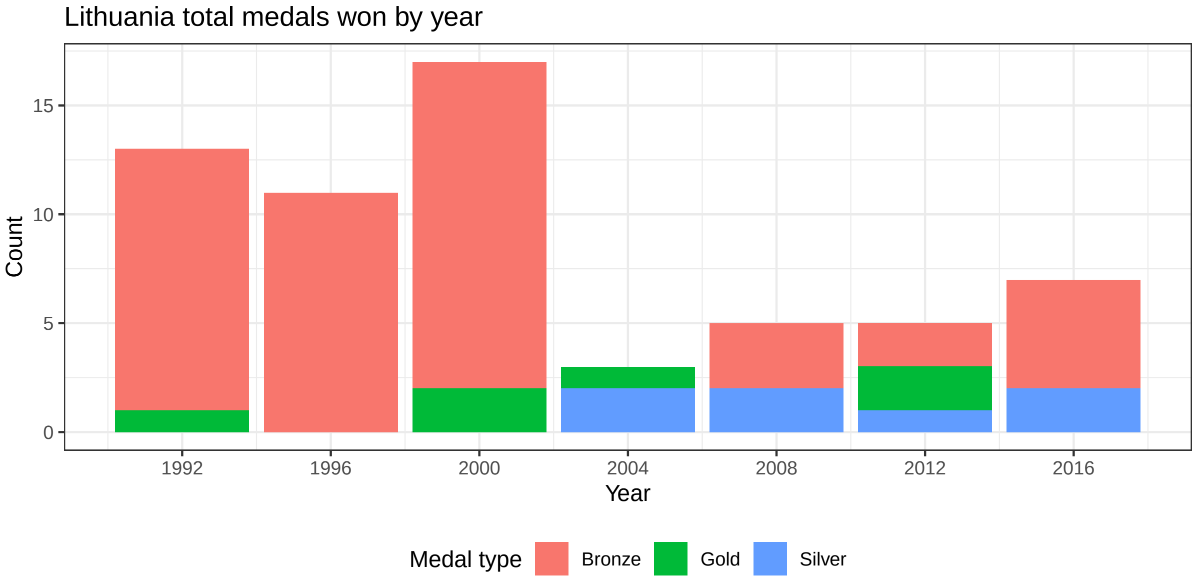
Most medals are bronze

```
dbGetQuery(con,
"select *
from athlete_events")|>
  filter(Team=="Lithuania")|>
  count(Medal)|>
  na.omit()|>
  ggplot(aes(Medal, n))+
  geom_col()+
  labs(title="Lithuania total medals won",
       x="Medal type",
       y="Count")
```



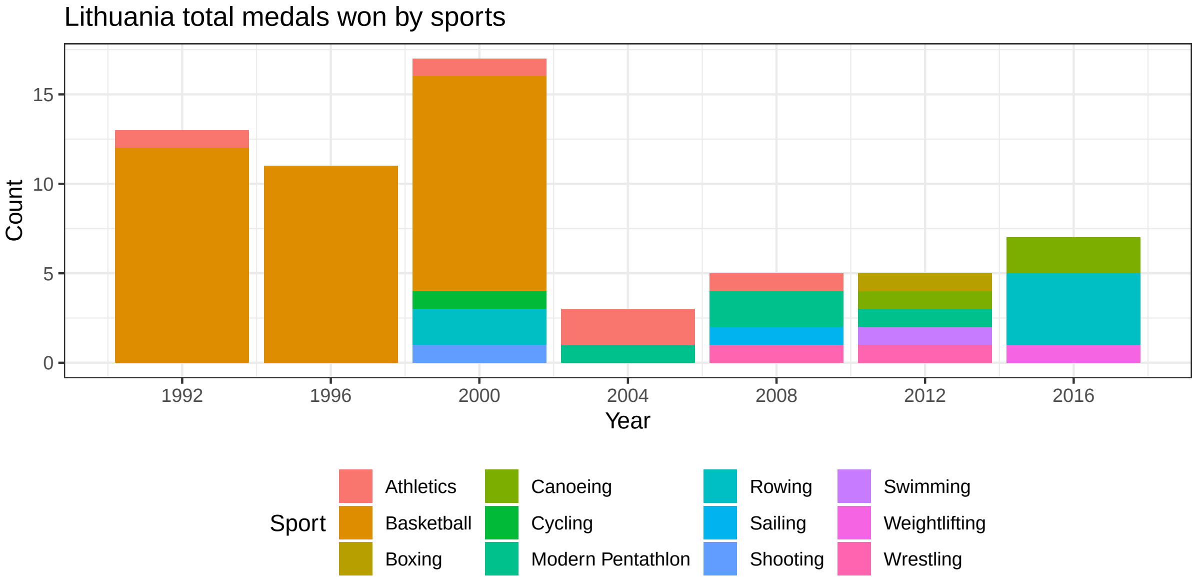
Most medals were won in 1922-2000 period

```
dbGetQuery(con,
"select *
from athlete_events")|>
  filter(Team=="Lithuania")%>%
  select(Year, Medal)%>%
  na.omit() %>%
  count(Year, Medal)%>%
  ggplot(aes(Year, n, fill=Medal))+
  geom_col(position = "stack")+
  scale_x_continuous(breaks = seq(0,3000, by=4))+
  theme_bw()+
  theme(legend.position = "bottom")+
  labs(title="Lithuania total medals won by year",
       x="Year",
       y="Count",
       fill="Medal type")
```



It appears Lithuania has gathered many medals in basketball in 1992,996,2000

```
dbGetQuery(con,
"select *
from athlete_events")|>
  filter(Team=="Lithuania")%>%
  select(Year, Medal, Sport)%>%
  na.omit() %>%
  count(Year, Sport)%>%
  ggplot(aes(Year, n, fill=Sport))+
  geom_col(position = "stack")+
  scale_x_continuous(breaks = seq(0,3000, by=4))+
  theme_bw()+
  theme(legend.position = "bottom")+
  labs(title="Lithuania total medals won by sports",
       x="Year",
       y="Count",
       fill="Sport")
```


Lithuania total medals won by sports
Count
15
10
5
0
1992
1996
2000
2004
2008
2012
2016
Year
Sport
Athletics
Basketball
Boxing
Canoeing
Cycling
Modern Pentathlon
Rowing
Sailing
Shooting
Swimming
Weightlifting
Wrestling

3 Step 2: Develop Project Proposal

3.1 Description

My target audiance is the international sports community. I will report and comapare Lithuanian Olympics perfamnce with other countries. The goal of this analysis is to investigate the basketball players performance with regard to their country of origin and their physique towards winning medals in Olympic games. Further its of interest to my audiance is the change of players physique over time.

3.2 Questions

- Did lithuanian baskettball players experiences a change in physique over time?
- Did they become oler/younger over time? Heavier/ lighter?
- Did more taller players came to pay at olympics?
- How did they compare to other nations: USA?

3.3 Hypothesis

- Hypothesis average heigt **did not change** among lithuanian players at olypmics in period 1992 vs 2016
- Hypothesis average weight **did not change** among lithuanian players at olypmics in period 1992 vs 2016
- Hypothesis average weight **increased** among lithuanian players at olypmics in period 1992 vs 2016
- LTU players compared to USA players during all periods (1992-2016 combined) were significalty lower
- LTU players compared to USA players during all periods (1992-2016 combined) were significalty heavier
- LTU players compared to USA players during all periods (1992-2016 combined) were significalty older

3.4 Approach

- I will use the Name, Sex, Age, Height, Team, Year columns for this analysis.
- I will use stuent-t statiscis to analyse the difference in means, as the variable sof interest are numeric.