Milestone 3 report

SQL for Data Science Capstone Project

Justas Mundeikis
mundeikis@gmx.de

2023-02-20

Contents

1 Info

This report (milestone 3 documentation) is the third part of the "SQL for Data Science Capstone Project". The porevious report with EDA part can be found in myt Github Repo.

2 Preparation

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

```
## loading libraries requiered
library(tidyverse)
library(gt)
library(RSQLite)
library(DBI)
library(broom)
```

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:

```
## 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 create a temporal local SQLite database.

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

Then I save the dataframes as a SQLite DBs:

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

3 Hypotheses

I have raised following hypotheses:

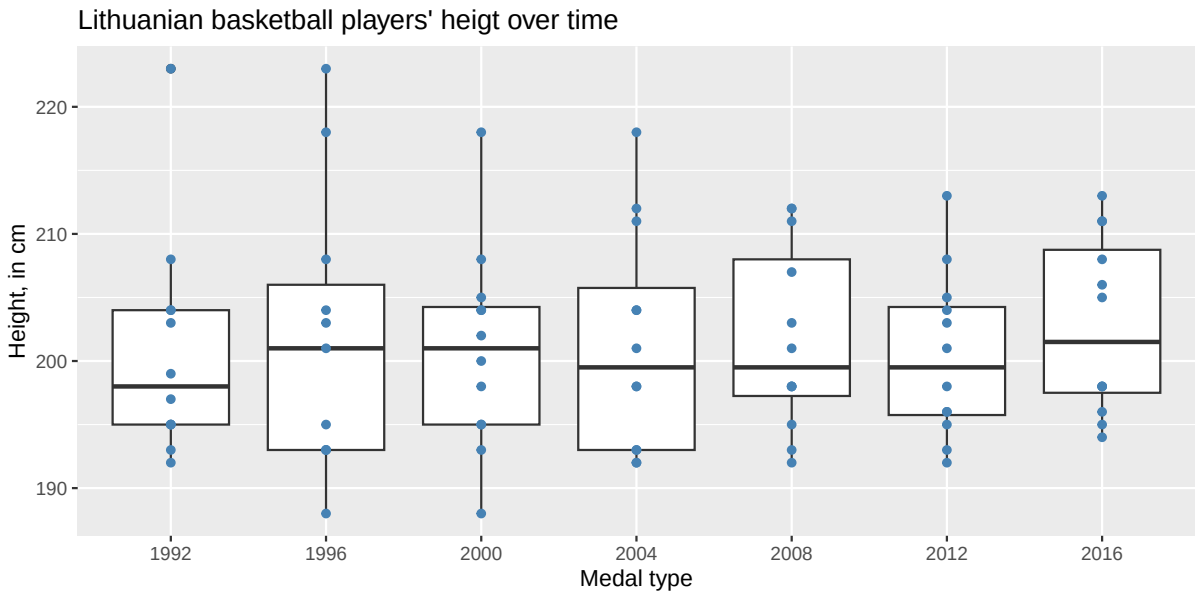
3.1 Hypothesis

- Hypothesis average height **did not change** among Lithuanian players at Olympics in period 1992 vs 2016 due to self selection bias in this sports
- Hypothesis average weight **did not change** among Lithuanian players at Olympics in period 1992 vs 2016 as players have to stay fit
- Hypothesis average age **increased** among Lithuanian players at Olympics in period 1992 vs 2016 as the same players often returned to the Olympic games multiple times

3.1.1 Average height of Lithuanian basketball players has not changed in period 1992 vs 2016

The Boxplot diagram shows that the median has slightly increased over time, while the variation has decreased when comparing to 1996-2004 era.

```
dbGetQuery(con,
"select Year, Height
from athlete_events
where Team='Lithuania'
AND Sport=='Basketball'
")|>
  ggplot(aes(as.factor(Year), Height))+
  geom_boxplot()+
  geom_point(col="steelblue")+
    labs(title="Lithuanian basketball players' heigt over time",
       x="Medal type",
       y="Height, in cm")
```



Summary statistics table:

```
dbGetQuery(con,
"select Year, Height
from athlete_events
where Team='Lithuania'
AND Sport=='Basketball'
")|>
  group_by(Year)|>
  summarise(avg_height=mean(Height),
            sd_height=sd(Height),
            median_height=median(Height))|>
  gt()|>
```

```
  tab_header(
    title = "Summary statistics",
    subtitle = "Lithuanian Basketball players"
  )|>
  fmt_number(columns = ends_with("height"), decimals = 1)
```

Summary statistics

Lithuanian Basketball players

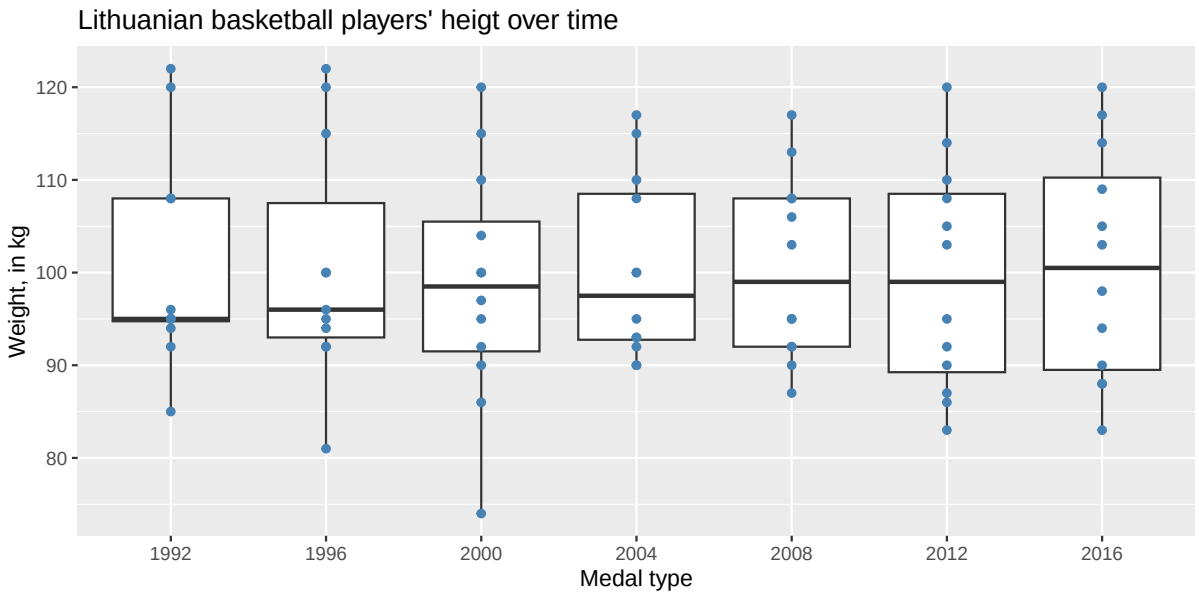
Year	avg_height	sd_height	median_height
1992	200.7	8.7	198.0
1996	201.7	11.1	201.0
2000	200.8	7.9	201.0
2004	201.3	8.7	199.5
2008	201.7	7.3	199.5
2012	200.3	6.4	199.5
2016	202.8	7.0	201.5

This data may suggest, that the hypothesis will be rejected.

3.1.2 Average weight of Lithuanian basketball players has not changed in period 1992 vs 2016

The Boxplot diagram shows that the median weight has slightly increased over time, while the variation has remain approximately the same over time.

```
dbGetQuery(con,
"select Year, Weight
from athlete_events
where Team='Lithuania'
AND Sport=='Basketball'
")|>
  ggplot(aes(as.factor(Year), Weight))+
  geom_boxplot()+
  geom_point(col="steelblue")+
    labs(title="Lithuanian basketball players' heigt over time",
       x="Medal type",
       y="Weight, in kg")
```



Summary statistics table:

```
dbGetQuery(con,
"select Year, Weight
from athlete_events
where Team='Lithuania'
AND Sport=='Basketball'
")|>
  group_by(Year)|>
  summarise(avg_weight=mean(Weight),
            sd_weight=sd(Weight),
            median_weight=median(Weight))|>
  gt()|>
```

```
  tab_header(
    title = "Summary statistics",
    subtitle = "Lithuanian Basketball players"
  )|>
  fmt_number(columns = ends_with("weight"), decimals = 1)
```

Summary statistics

Lithuanian Basketball players

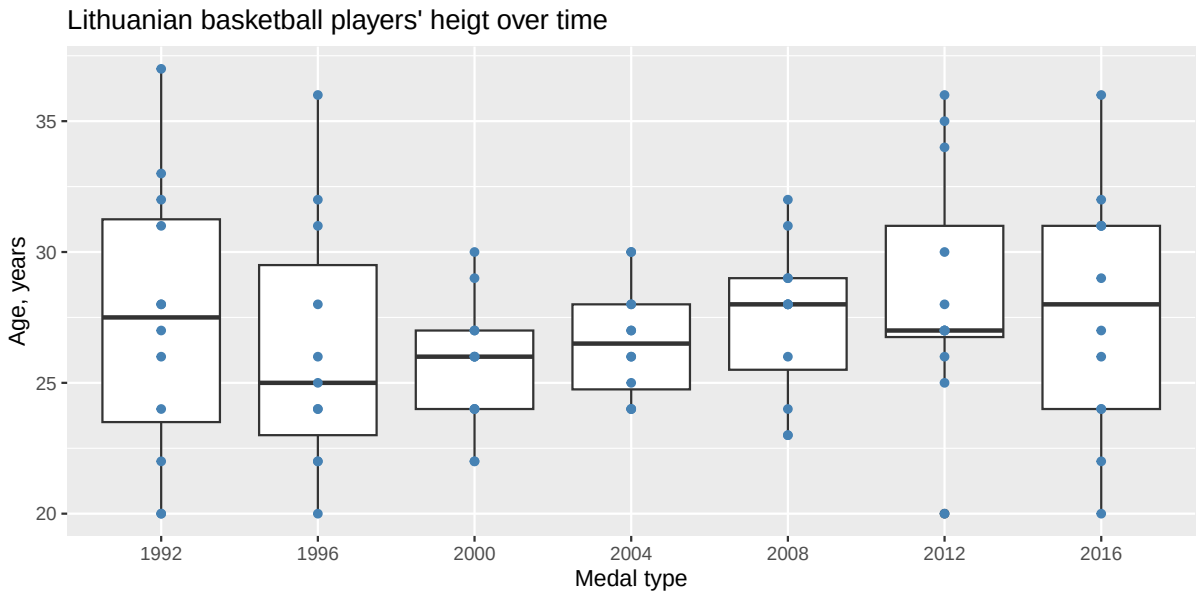
Year	avg_weight	sd_weight	median_weight
1992	100.4	11.5	95.0
1996	100.6	12.9	96.0
2000	98.6	12.7	98.5
2004	100.2	9.8	97.5
2008	100.5	9.9	99.0
2012	99.4	12.2	99.0
2016	100.8	12.5	100.5

This data may suggest, that the hypothesis will be confirmed

3.1.3 Average age of Lithuanian basketball players has not changed in period 1992 vs 2016

The Box plot diagram shows that in 1922 there was the largest variation in data, then in 1996 the median age dropped and every games past that more or less the same players returned to Olympics, the median age increased. In 2016, although completely different players were playing, the median variation appears to be very similar to 1996

```
dbGetQuery(con,
"select Year, Age
from athlete_events
where Team='Lithuania'
AND Sport=='Basketball'
")|>
  ggplot(aes(as.factor(Year), Age))+
  geom_boxplot()+
  geom_point(col="steelblue")+
    labs(title="Lithuanian basketball players' heigt over time",
       x="Medal type",
       y="Age, years")
```



http://127.0.0.1:17887/graphics/6feafc30-bff8-4fab-8f0e-0adc76bb09ed.png Summary statistics table:

```
dbGetQuery(con,
"select Year, Weight
from athlete_events
where Team='Lithuania'
AND Sport=='Basketball'
")|>
  group_by(Year)|>
  summarise(avg_weight=mean(Weight),
            sd_weight=sd(Weight),
```

```
            median_weight=median(Weight))|>
gt()|>
tab_header(
  title = "Summary statistics",
  subtitle = "Lithuanian Basketball players"
)|>
fmt_number(columns = ends_with("weight"), decimals = 1)
```

Summary statistics

Lithuanian Basketball players

Year	avg_weight	sd_weight	median_weight
1992	100.4	11.5	95.0
1996	100.6	12.9	96.0
2000	98.6	12.7	98.5
2004	100.2	9.8	97.5
2008	100.5	9.9	99.0
2012	99.4	12.2	99.0
2016	100.8	12.5	100.5

This data may suggest, that the hypothesis will be rejected, but if the comaprisan was drawn from 1996 and not 1992 as base year, the hypothesis might be confirmed.

3.2 Correlational analysis

Now I test time and the metric variables: height, weight age.

```
dbGetQuery(con,
"select
 Year
,Height
,Weight
,Age
from athlete_events
where Team='Lithuania'
AND Sport=='Basketball'
")|>
  cor()|>
  as.data.frame()|>
  rownames_to_column(var = "Variables")|>
    gt()|>
    tab_header(
        title = "Lithuanian basketball players metrics vs time",
        subtitle = "Correlational analysis"
    ) |>
    fmt_number(col=c(Year,Height,Weight,Age), decimals = 2)
```

Lithuanian basketball players metrics vs time

Correlational analysis

Variables	Year	Height	Weight	Age
Year	1.00	0.04	0.00	0.13
Height	0.04	1.00	0.86	−0.14
Weight	0.00	0.86	1.00	−0.08
Age	0.13	−0.14	−0.08	1.00

Conclusion, year is not correlated with physique of the players.

3.3 Testing

I test the hypothesis for differences in means using student-t statistic.

The data does not allow to reject the H0 hypothesis at $\alpha = 0.05$ level. Although all variables show relative increase, none of the changes in averages is statically significant.

The SQL query results are forwarded into R:

```
dbGetQuery(con,
"select
 Year
,Height
,Weight
,Age
from athlete_events
where Team='Lithuania'
AND Sport=='Basketball'
AND Year in (1992, 2016)
")|>
  gather(variable, values, 2:4)|>
  group_by(variable)|>
      do(tidy(with(data = ., t.test(values[Year == "1992"], values[Year == "2016"]))))|>
    select(variable,  estimate1, estimate2, p.value)%>%
    mutate(p.value=round(p.value, 3))|>
    mutate(rel_change=estimate2/estimate1-1)|>
    mutate(star = case_when(
        p.value <= 0.001 ~ "***",
        p.value <= 0.01 ~ "**",
        p.value <= 0.05 ~ "*",
        p.value <= 0.1 ~ ".",
        TRUE ~ ""
    ))%>%
    mutate(p.value=paste(p.value, star))%>%
    select(-star)|>
    ungroup()|>
    relocate(variable, estimate1, estimate2, rel_change, p.value)|>
    gt()|>
    tab_header(
        title = "Lithuanian basketball players metrics 1992 vs 2016",
        subtitle = "student-t statiscis for height, weight, age"
    ) |>
    fmt_number(c(estimate1, estimate2), decimals = 2)|>
    fmt_percent(rel_change, decimals = 2)|>
    cols_align(
        align = "left",
        columns = p.value
    )
```

Lithuanian basketball players metrics 1992 vs 2016

student-t statiscis for height, weight, age

variable	estimate1	estimate2	rel_change	p.value
Age	27.33	27.75	1.52%	0.841
Height	200.67	202.75	1.04%	0.523
Weight	100.42	100.75	0.33%	0.946

3.4 Conclusions

Despite exploratory data analysis showing some relative increases in basketball players physique - none of it appears to be significant