Milestone 3 report

SQL for Data Science Capstone Project

Justas Mundeikis
mundeikis@gmx.de

2023-02-18

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 testing

I have raised following hypotheses:

3.1 Hypotheses

- 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

3.2 Testing

Exploratory data analysis (see milestone 2 in Github Repo) suggested, that there might a a statically significant increase in Age. Height and Weight appeared not to have changed over time.

I tests 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

3.3 Conclusions

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