UNIVERSITÉ CÔTE D'AZUR

Processing Large Datasets with R

Exam presentation (exam 1)

Joris LIMONIER

December 7, 2021

Table of Contents

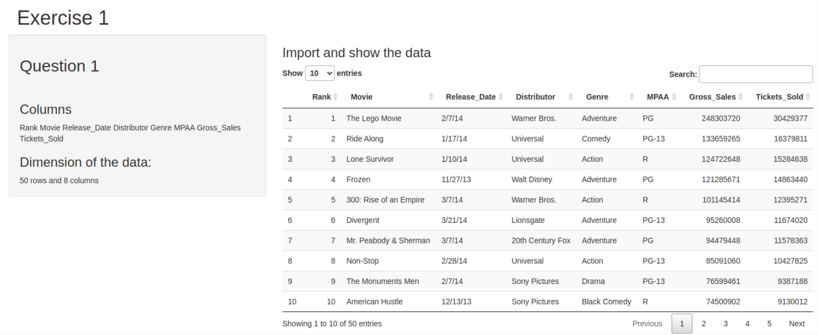
1. Exercise 1 - Shiny
(Movies dataset)

2. Exercise 2 - RMarkdown
(Winter dataset)

3. Exercise 3 - Data Analysis
(Summer-Winter dataset)

1. Exercise 1 - Shiny (Movies dataset)

Import and overview the data



Exercise 1

Question 1

Columns

Rank Movie Release_Date Distributor Genre MPAA Gross_Sales Tickets_Sold

Dimension of the data:

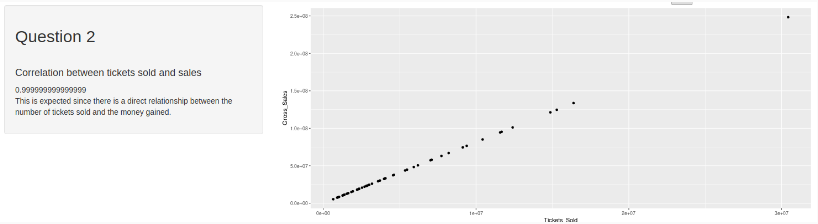
50 rows and 8 columns

Import and show the data

	Rank	Movie	Release_Date	Distributor	Genre	MPAA	Gross_Sales	Tickets_Sold
1	1	The Lego Movie	2/7/14	Warner Bros.	Adventure	PG	248303720	30429377
2	2	Ride Along	1/17/14	Universal	Comedy	PG-13	133659265	16379811
3	3	Lone Survivor	1/10/14	Universal	Action	R	124722648	15284638
4	4	Frozen	11/27/13	Walt Disney	Adventure	PG	121285671	14863440
5	5	300: Rise of an Empire	3/7/14	Warner Bros.	Action	R	101145414	12395271
6	6	Divergent	3/21/14	Lionsgate	Adventure	PG-13	95260008	11674020
7	7	Mr. Peabody & Sherman	3/7/14	20th Century Fox	Adventure	PG	94479448	11578363
8	8	Non-Stop	2/28/14	Universal	Action	PG-13	85091060	10427825
9	9	The Monuments Men	2/7/14	Sony Pictures	Drama	PG-13	76599461	9387188
10	10	American Hustle	12/13/13	Sony Pictures	Black Comedy	R	74500902	9130012

Showing 1 to 10 of 50 entries

Plot ticket sales vs gross sales



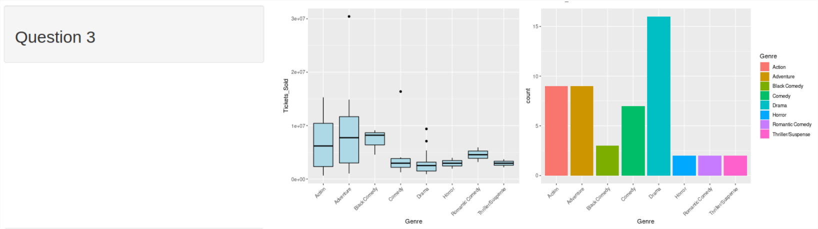
Question 2

Correlation between tickets sold and sales

0.999999999999999

This is expected since there is a direct relationship between the number of tickets sold and the money gained.

Box plot and count type of film



Question 3

Tickets sales histogram - play with number of bins

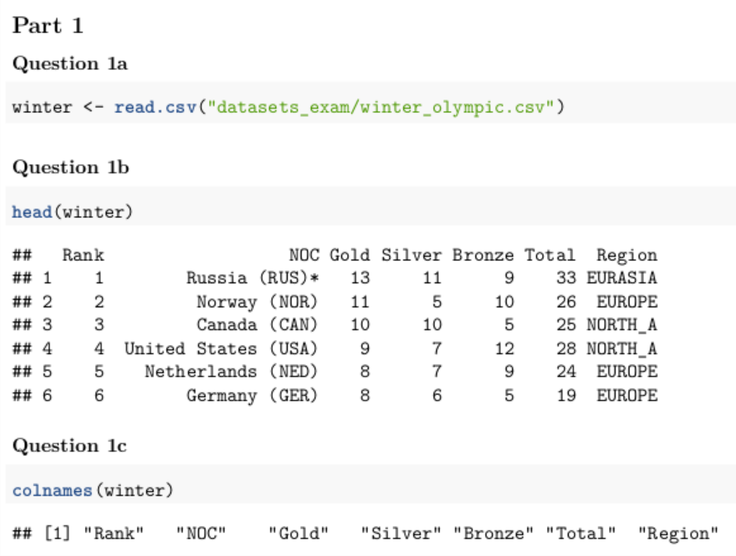
Watch video
Backup link: https://youtu.be/NTgGG7UvRRU

Tickets and gross sales by genre and distributor

Watch video
Backup link: https://youtu.be/w_QQVsRoOpA

2. Exercise 2 - RMarkdown (Winter dataset)

Import and overview of the data



Part 1

Question 1a

```
winter <- read.csv("datasets_exam/winter_olympic.csv")
```

Question 1b

```
head(winter)
```

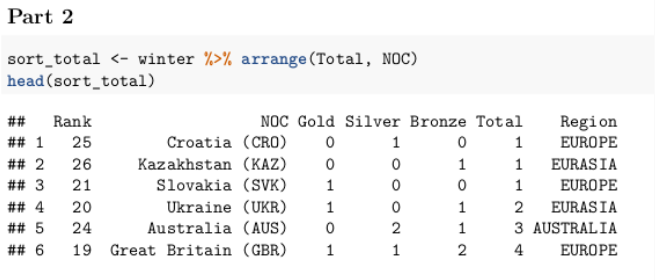
```
##   Rank                     NOC Gold Silver Bronze Total  Region
## 1    1          Russia (RUS)*   13     11      9    33 EURASIA
## 2    2           Norway (NOR)   11      5     10    26  EUROPE
## 3    3           Canada (CAN)   10     10      5    25 NORTH_A
## 4    4  United States (USA)     9      7     12    28 NORTH_A
## 5    5    Netherlands (NED)     8      7      9    24  EUROPE
## 6    6          Germany (GER)    8      6      5    19  EUROPE
```

Question 1c

```
colnames(winter)
```

```
## [1] "Rank"   "NOC"    "Gold"   "Silver" "Bronze" "Total"  "Region"
```

Sort by total medals

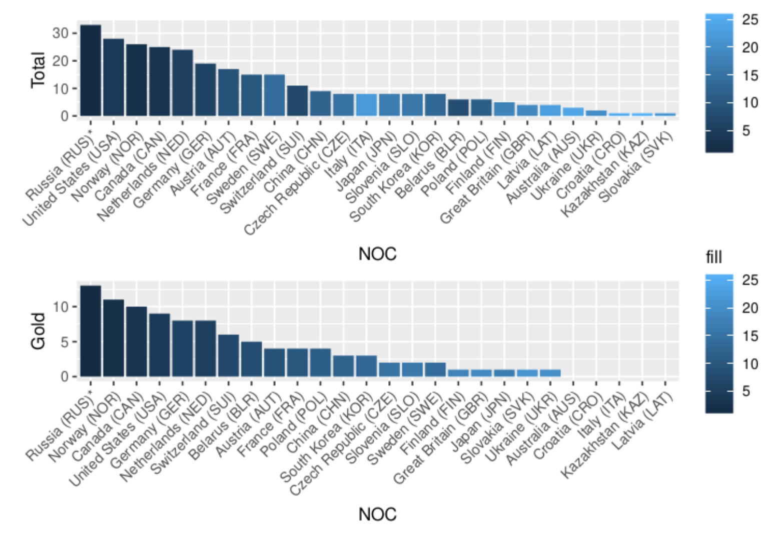


Part 2

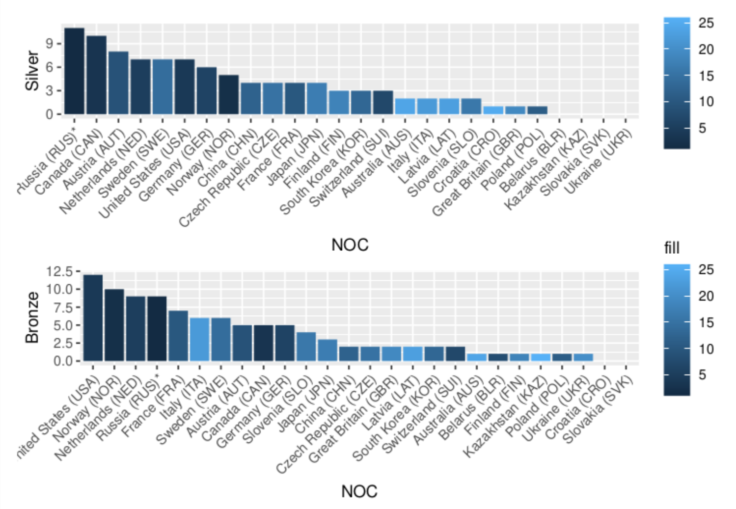
```
sort_total <- winter %>% arrange(Total, NOC)
head(sort_total)
```

```
##   Rank                 NOC Gold Silver Bronze Total    Region
## 1   25       Croatia (CRO)    0      1      0     1    EUROPE
## 2   26    Kazakhstan (KAZ)    0      0      1     1   EURASIA
## 3   21      Slovakia (SVK)    1      0      0     1    EUROPE
## 4   20       Ukraine (UKR)    1      0      1     2   EURASIA
## 5   24     Australia (AUS)    0      2      1     3 AUSTRALIA
## 6   19 Great Britain (GBR)    1      1      2     4    EUROPE
```

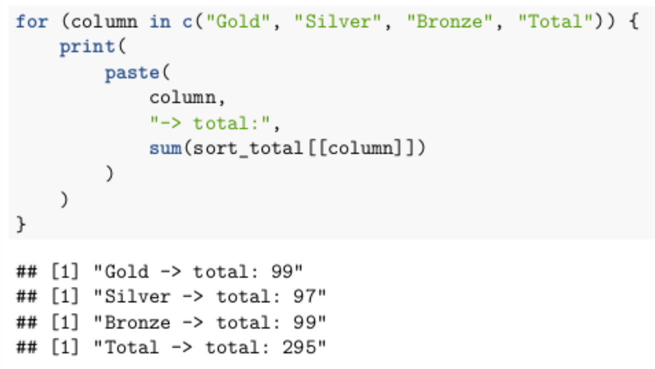
Total and Gold bar plots



Silver and Bronze bar plots



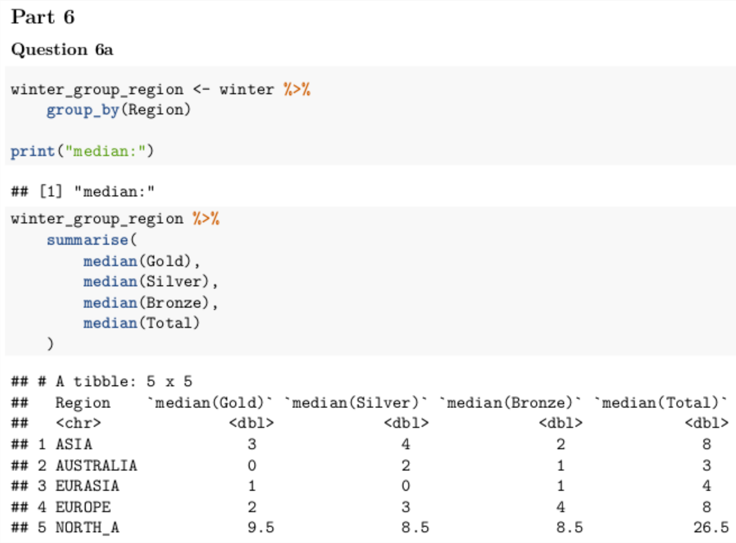
Total of medals



```
for (column in c("Gold", "Silver", "Bronze", "Total")) {
    print(
        paste(
            column,
            "-> total:",
            sum(sort_total[[column]])
        )
    )
}
```

```
## [1] "Gold -> total: 99"
## [1] "Silver -> total: 97"
## [1] "Bronze -> total: 99"
## [1] "Total -> total: 295"
```

Medians of medals per region



Part 6

Question 6a

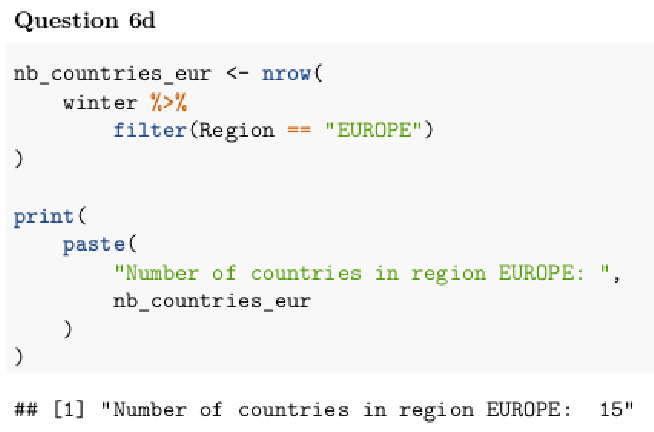
```
winter_group_region <- winter %>%
    group_by(Region)

print("median:")
```

```
## [1] "median:"
```

```
winter_group_region %>%
    summarise(
        median(Gold),
        median(Silver),
        median(Bronze),
        median(Total)
    )
```

```
## # A tibble: 5 x 5
##   Region    `median(Gold)` `median(Silver)` `median(Bronze)` `median(Total)`
##   <chr>              <dbl>            <dbl>            <dbl>           <dbl>
## 1 ASIA                 3                4                2               8
## 2 AUSTRALIA            0                2                1               3
## 3 EURASIA              1                0                1               4
## 4 EUROPE               2                3                4               8
## 5 NORTH_A              9.5              8.5              8.5            26.5
```

Number of European countries in the dataset

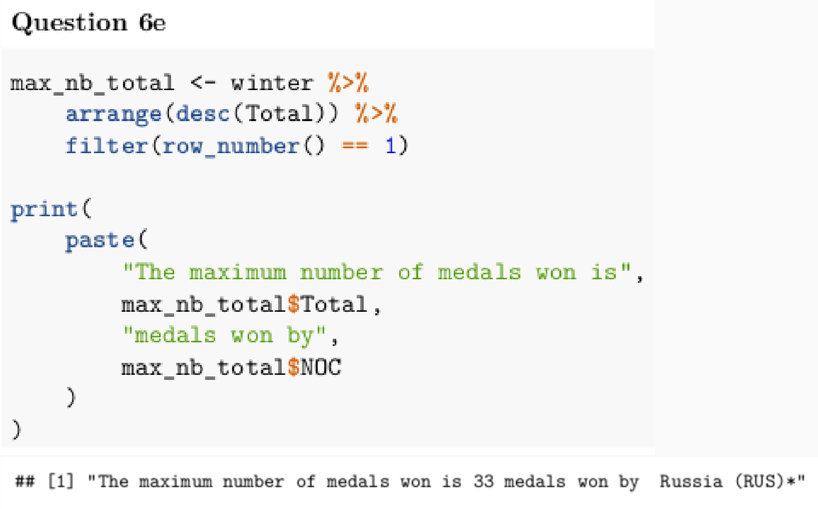


Question 6d

```r
nb_countries_eur <- nrow(
    winter %>%
        filter(Region == "EUROPE")
)

print(
    paste(
        "Number of countries in region EUROPE: ",
        nb_countries_eur
    )
)
```

```
## [1] "Number of countries in region EUROPE:  15"
```

Country with most medals



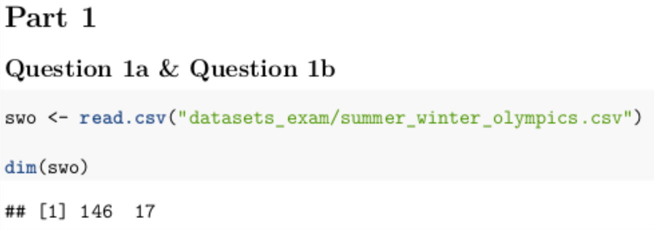
Question 6e

```r
max_nb_total <- winter %>%
    arrange(desc(Total)) %>%
    filter(row_number() == 1)

print(
    paste(
        "The maximum number of medals won is",
        max_nb_total$Total,
        "medals won by",
        max_nb_total$NOC
    )
)
```

```
## [1] "The maximum number of medals won is 33 medals won by  Russia (RUS)*"
```

3. Exercise 3 - Data Analysis (Summer-Winter dataset)

Import dataset



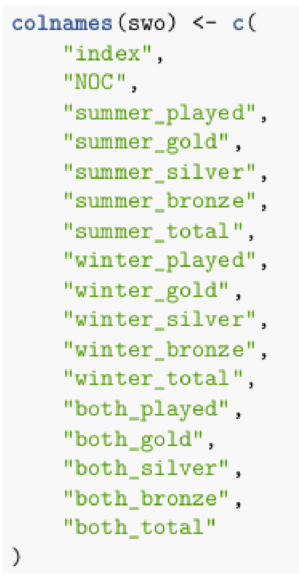
Part 1

Question 1a & Question 1b

```
swo <- read.csv("datasets_exam/summer_winter_olympics.csv")

dim(swo)
```

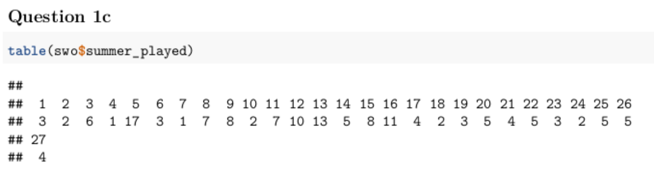
```
## [1] 146  17
```

Rename columns



```
colnames(swo) <- c(
    "index",
    "NOC",
    "summer_played",
    "summer_gold",
    "summer_silver",
    "summer_bronze",
    "summer_total",
    "winter_played",
    "winter_gold",
    "winter_silver",
    "winter_bronze",
    "winter_total",
    "both_played",
    "both_gold",
    "both_silver",
    "both_bronze",
    "both_total"
)
```

Frequency counts

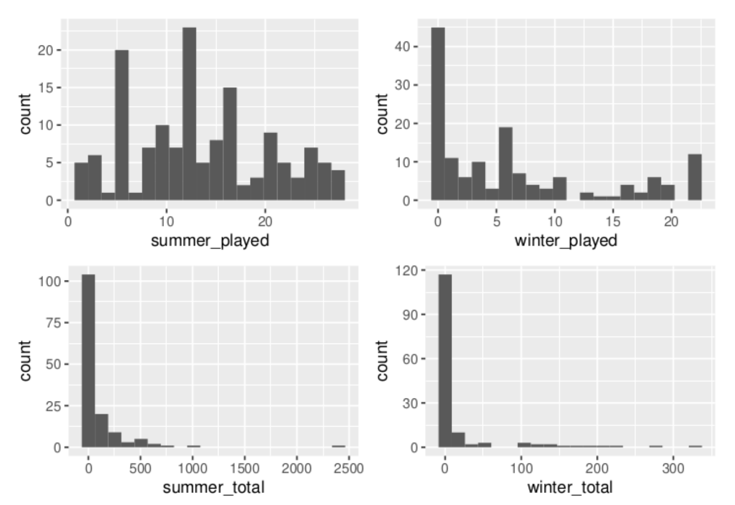


Question 1c

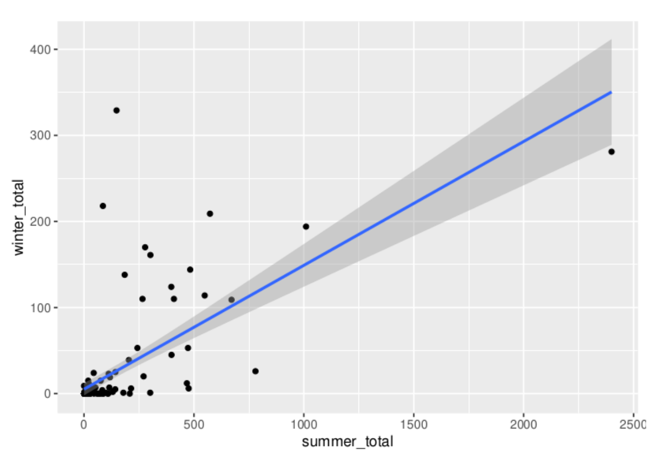
```
table(swo$summer_played)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
##  3  2  6  1 17  3  1  7  8  2  7 10 13  5  8 11  4  2  3  5  4  5  3  2  5  5
## 27
##  4
```

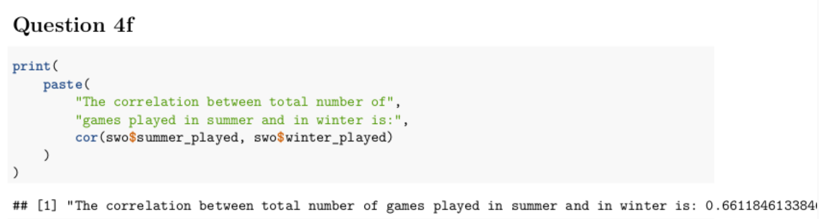
Compare summer vs winter & played vs total



Plot winter vs summer total



Correlation between winter vs summer total



Question 4f

```
print(
    paste(
        "The correlation between total number of",
        "games played in summer and in winter is:",
        cor(swo$summer_played, swo$winter_played)
    )
)
```

```
## [1] "The correlation between total number of games played in summer and in winter is: 0.661184613384
```

Thank you

Questions?