Data Analysis and Visualization in R (IN2339)

Exercise Session 2

Evangelos Theodorakis, Daniela Klaproth-Andrade, Julien Gagneur

Quizzes (from the lecture)

The following quizzes will be solved orally by the students and the professor during the lecture.

1. Let `awesome.dt` be a data.table. Which command produces the following table?

```
##    x     y
## 1: 6  TRUE
## 2: 5  TRUE
## 3: 4  TRUE
## 4: 3 FALSE
## 5: 2 FALSE
## 6: 1 FALSE
```

   a. `awesome.dt <- data.table(x = order(1:6, decreasing = T), y = rep(c(TRUE, FALSE), each = 3))`
   b. `awesome.dt <- data.table(x = order(1:6, decreasing = T), rep(c(TRUE, FALSE), 3))`
   c. `awesome.dt <- data.table(x = order(1:6, decreasing = F), y = rep(c(TRUE, FALSE), each = 3))`
   d. `awesome.dt <- data.table(x = order(1:6, decreasing = F), y = rep(c(TRUE, FALSE), 3))`

2. Find the last 10 flights arriving in LAX on Christmas Eve (24. December). Don't worry about sorting for now, just find the last 10 entries in the table.

   a. `flights[MONTH == 12 & DAY == 24 & ORIGIN_AIRPORT == "LAX"] %>% tail(n=10)`
   b. `flights[MONTH == 12 & DAY_OF_WEEK == 24 & DESTINATION_AIRPORT == "LAX"] %>% head(n=10)`
   c. `flights[MONTH == 12 & DAY == 24 & DESTINATION_AIRPORT == LAX] %>% tail(n=10)`
   d. `flights[MONTH == 12 & DAY == 24 & DESTINATION_AIRPORT == "LAX"] %>% tail(n=10)`

3. Let `iris.dt <- data.table(iris)`. What happens if we run `iris.dt[Species != "setosa" | Sepal.Length <= 5, .N, by = Species]`?

   a. Get the number of rows for each unique value of `Species`.
   b. Get the number of rows for each unique value of `Species` except `setosa`.
   c. Get the number of rows for each unique value of `Species`, for all the rows with `Sepal.Length` less or equal to 5.
   d. Get the number of rows for each unique value of `Species`, for all the rows where `Species` is not `setosa` or with `Sepal.Length` less or equal to 5.

4. Calculate the total number of outbound flights in the summer months (June - August).

a. `flights[MONTH %in% c(6, 7, 8), by = ORIGIN_AIRPORT]`
b. `flights[MONTH %in% 6:8, .N, by = ORIGIN_AIRPORT]`
c. `flights[MONTH == c(6, 7, 8), .N, by = DEPARTURE_AIRPORT]`
d. `flights[MONTH == 6:8, .N, by = ORIGIN_AIRPORT]`

Tutorial

The following exercises will be solved during the tutorial sessions.

Section 00 - Getting ready

1. Make sure you have already installed and loaded the libraries `data.table` and `magrittr` by running the `install.packages("PackageName")` command, where `PackageName` is the package you want to install:

```
library(data.table)
library(magrittr)
```

Section 01 - Reading and cleaning up data

1. Download the datasets needed for the exercise from Moodle, extract them, and put them in a folder called `extdata`. Load the three given datasets as `data.tables` and name them as `users_dt`, `books_dt` and `ratings_dt` accordingly. *Hint:* `fread()` and `file.path()`

2. What are the classes of `users_dt`, `ratings_dt` and `books_dt`. Confirm that these are indeed a `data.table`.

3. What are the column names of the `users_dt data table`? What are the classes of the `users_dt data table`. *Hint:* `str()` or `sapply()`? Then change the type of the `Age` column in `users_dt` to numeric.

4. Produce a summary of the variables in `books_dt`.

5. Return the first 5 and last 5 observations of the table `ratings_dt`.

6. Replace all the `-` in column names by underscores `_` in all three data tables. For example, `Book-Title` should be renamed to `Book_Title`. *Hint:* You can use the function `gsub()` that replaces pattern in a character string by a defined replacement. For example, for replacing `R` by `DataViz` in the following sentence `s` we use:

```
s <- 'R is fun'
gsub('R', 'DataViz', s)
```

```
## [1] "DataViz is fun"
```

7. Delete the columns `Image_URL_S`, `Image_URL_M` and `Image_URL_L` in the table `books_dt`.

8. Create a table `book_dt_2` that contains all the books published between 1900 and 2019 (inclusive) from the table `books_dt`.

Section 02 - Data Exploration

1. How many different authors are included in the table `books_dt`?

2. How many different authors are included for each year of publication between 2000 and 2010 (inclusive) in `books_dt`?

3. In how many observations is the age information missing in the ratings table `users_dt`?

4. What is the maximum rating value in the ratings table?

5. What is the most common rating value larger than 0?

6. Which are the book identifiers (ISBN) with the highest ratings?

7. Reorder the ratings table according to the rating value of each book in descending order. *Hint*: `order()`

Homework

Please solve the exercises below at home. The solutions will be discussed in the central exercise.

Section 03 - Manipulating data tables

1. Add a new column called `High_Rating` to the data table `ratings_dt`. The column has an integer 1 for all observations with a rating value higher than 7. *Hint*: `ifelse()`

2. How many observations are considered to be a high ranking? What is the proportion of high ranked observations among all observations?

3. Which users did not give any rating to any book? Filter these users out from `users_dt`. *Hint*: There's no need to merge `users_dt` with `ratings_dt`, we are simply interested in the users that are not in `ratings_dt`.

4. What is the most common age of users who rated at least one book?

5. On average, how many books did a user rate?

6. What is the title of the first published book with the highest ranking?

7. In which year was a book with the largest number of ratings last published?

8. Add to the table `ratings_dt` the highest ranking that each book received as a new column called `Max_Book_Ranking`.

9. Subset the `ratings_dt` ratings table to contain only books written by the following authors:

```
authors <- c("Agatha Christie", "William Shakespeare", "Stephen King",
             "Ann M. Martin", "Carolyn Keene", "Francine Pascal",
             "Isaac Asimov", "Nora Roberts", "Barbara Cartland", "Charles Dickens")
authors
```

```
##  [1] "Agatha Christie"     "William Shakespeare" "Stephen King"
##  [4] "Ann M. Martin"       "Carolyn Keene"       "Francine Pascal"
##  [7] "Isaac Asimov"        "Nora Roberts"        "Barbara Cartland"
## [10] "Charles Dickens"
```

10. How many ratings has each author from the previous exercise 9? What is their max and average ranking?

Section 04 - Working with Excel formats

1. Using the summer_olympic_medals.xlsx file, which athlete won most bronze medals? *Hint* `read_excel()` from `readxl` package.

2. Are the columns `Gender` and `Event_gender` consistent? Find inconsistent gender entries.

3. Which country won most medals? Which country has the highest ratio of silver medals? Use the data in the country summary sheet starting at row 147 of the summer_olympic_medals.xlsx file.

4. Which countries did participate, but without winning medals? Assume, that all countries listed in the IOC COUNTRY CODES sheet participated. *Hint* you can quick fix the column names with `make.names` and find set differences with `setdiff`.