# Data science with R: tidyverse

## I Tidyverse essentials (dplyr & tidyr)

### Assignment

Create *R* script called *assignment_1.R*. Try to finish the given exercises by using data transformation techniques!

### Exercise 1

In this assignment you will use **hflights** dataset from the package **hflights**.

Use **dplyr** and **tidyr** and try to answer the following questions:

- How many rows and columns are in table **hflights**?
- How many different carriers are listed in the table (print a table with distinct carrier names)?
- Which and how many airports were involved? Consider both origin and destination airports!
- How many flights were cancelled?

### Exercise 2

First, produce a table where statistics for each carrier is shown:

- number of flights per carrier
- total distance flown in miles per carrier
- total actual elapsed time in hours per carrier
- total air time in hours per carrier
- mean distance per flight for each carrier
- mean actual elapsed time in hours per flight for each carrier
- mean air time in hours per flight for each carrier

Second, calculate the percentage of total distance flown by top 3 performing carriers VS total distance flown by remaining carriers. Execute steps:

- first rank carriers by total distance flown
- top 3 performers are in one group, remaining carriers are in second group
- for each group calculate total distance flown
- for each group calculate %: $\frac{total\ distance\ flown\ per\ group}{total\ distance\ all\ carriers}$

## Exercise 3

Modify your main flights table:

- create **date** column by uniting columns: year, month, day of month
- when uniting columns do not lose source columns (mutate each column - with slightly different name, before unite operation is executed)
- you will need to parse **date** column after unite operation
- also you should add leading zeros to month and day of month column before date is created
- create columns: **quarter**, **week**

**HINT:** you can use **tidyverse** packages **lubridate** (date time related manipulations) and **stringr** (string based manipulations). The usage will be shown in the solution video.

Using your modified table try to answer the given questions:

- Is total number of flights increasing or decreasing quarterly?
- Is total distance increasing or decreasing monthly?
- **HINT: dplyr**'s function **lag** can assist you when calculating the quarterly or monthly differences!
- In the solution video, the visualization of quarterly / monthly differences will be shown using **ggplot2** library.

## Exercise 4

The idea for the last exercise is another data wrangling task, where you will have to use technique called **"pivoting"**. Build a table, that will resemble a heat map by:

- for each carrier and month, calculate total number of flights
- then normalize total number of flights (divide each value with maximum total number of flights, you must get values between 0 and 1!)
- now pivot your table from **long** to **wide** format
- so each row is represented with carrier, and each column is represented with month, normalized total number of flights are values in table cells

You should get a similar output:

| | Month | Month 2 | ... | Month 12 |
|---|---|---|---|---|
| Carrier 1 | $x_{1,1}$ | $x_{1,2}$ | ... | $x_{1,12}$ |
| Carrier 2 | $x_{2,1}$ | $x_{2,2}$ | ... | $x_{2,12}$ |
| ... | | | | |
| Carrier n | $x_{n,1}$ | $x_{n,2}$ | ... | $x_{n,12}$ |

Where $x_{i,j}$ is the normalized value of total flights for carrier $i$ and month $j$. In the solution video, the visualization of heat map will be shown using **ggplot2** library.