Data science with R: tidyverse

VII Data Wrangle: dplyr for relational data

Assignment

Create *R* script called *assignment_7.R*. In given exercises we will use **nycflights13** database.

Exercise 1

In the first exercise we would like to inspect, if carriers use different type of planes:

- we would like to see the distribution of planes per **manufacturer** for each **carrier**
- first, add **plane info** and **carrier name** to **flights** table
- second, count number of different planes, break down by **manufacturer** for each **carrier** (you can also add number of flights)
- lastly, create a bar plot (**manufacturer** on x-axis, **number of planes** on y-axis, bar fill color is defined by **carrier**)

Exercise 2

In the second exercise we would like to use data and try to answer the following question: "Do weather conditions have affect on flight arrival delay time?"

Here you have more freedom, choose the data you want and use the plot you think is the most adequate, for answering the question.

Exercise 3

The last exercise will be completed in two parts. First part is covered in this exercise, where we will create two different tables:

- **table 1** - "**distance_per_date**":
  - first create **date** column in **flights** table
  - then calculate total distance flown for each **carrier** per **date**
- **table 2** - "**dates_span**":
  - generate table with single column **date**
  - in given column dates are a sequence from minimum date found in flights table
  - and up to maximum date found in the flights table

Exercise 4

This is the second part of the last exercise, here we will:

- join tables from previous exercise
- we would like to get a single table
- each **date** must be shown per each **carrier**
- if 365 different dates, then 365 rows per carrier
- probably some carriers did not fly on some dates
- fill these blanks with 0 (0 miles flown)
- then calculate **cumulative sum** of miles flown per each carrier
- rows must be sorted by **date** per each **carrier**
- then draw a line chart:
  - line per **carrier**
  - each line represents **cumulative distance flown**
  - x-axis is represented with **date**
  - y-axis is represented with **cumulative distance flown**
  - color of line is represented by **carrier**