Exercise 3: Assignment

Philipp Kling

2020-06-15

Contents


```
library(knitr)

### Global options
options(max.print="75")
opts_chunk$set(echo=FALSE,
                 cache=FALSE,
               prompt=FALSE,
               tidy=TRUE,
               comment=NA,
               message=FALSE,
               warning=FALSE)
opts_knit$set(width=75)
rm(list = ls())
```

Preparation

Install the 'nycflights13' package and load the data into R.

```
library(nycflights13)
```

Overview

You can get a basic overview of the dataset with these functions

```
# How many rows and columns?
dim(flights)  #  or: nrow(flights)    ncol(flights)
```

```
[1] 336776     19
```

```
# What are the names of the variables/columns?
colnames(flights)
```

```
 [1] "year"           "month"          "day"            "dep_time"
 [5] "sched_dep_time" "dep_delay"      "arr_time"       "sched_arr_time"
```

```
 [9] "arr_delay"      "carrier"        "flight"         "tailnum"
[13] "origin"         "dest"           "air_time"       "distance"
[17] "hour"           "minute"         "time_hour"
```

```r
# Summary statistics
summary(flights)
```

```
      year          month             day           dep_time
 Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1
 1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907
 Median :2013   Median : 7.000   Median :16.00   Median :1401
 sched_dep_time   dep_delay          arr_time    sched_arr_time
 Min.   : 106   Min.   : -43.00   Min.   :   1   Min.   :   1
 1st Qu.: 906   1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124
 Median :1359   Median :  -2.00   Median :1535   Median :1556
   arr_delay          carrier              flight       tailnum
 Min.   : -86.000   Length:336776      Min.   :   1   Length:336776
 1st Qu.: -17.000   Class :character   1st Qu.: 553   Class :character
 Median :  -5.000   Mode  :character   Median :1496   Mode  :character
    origin              dest              air_time        distance
 Length:336776      Length:336776      Min.   : 20.0   Min.   :  17
 Class :character   Class :character   1st Qu.: 82.0   1st Qu.: 502
 Mode  :character   Mode  :character   Median :129.0   Median : 872
      hour           minute        time_hour
 Min.   : 1.00   Min.   : 0.00   Min.   :2013-01-01 05:00:00
 1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
 Median :13.00   Median :29.00   Median :2013-07-03 10:00:00
 [ reached getOption("max.print") -- omitted 4 rows ]
```

Assignment 1: Subsetting and alterations with dplyr

(a) Create a new variable

Use dplyr to create a variable 'caught_up' that only consists of values that are TRUE or FALSE and which indicates whether a flight *caught up* with a departure delay. I.e. it should be TRUE if the delay at arrival was less than the delay of the departure and FALSE otherwise.

```r
solution <- ""
```

(b) Extraction of observations

Use dplyr to filter the dataset to include only flights that had a delayed departure. Report which percentage of all the flights had a delayed departure. How many of those delayed flights also had a delayed arrival?

```r
library(dplyr)
solution <- ""
```

Assignment 2: Summary statistics

(a) Summary statistics 1

Do flights from JFK have a greater departure delay than flights from EWR on average? Use dplyr to find out.

```
library(dplyr)
solution <- ""
```

(b) Summary statistics 2

Which airport is the most common to get to Chicago O'Hare International Airport (ORD)? Use dplyr to find out.

```
library(dplyr)
solution <- ""
```

Assignment 3: Rewriting

Piping

Rewrite the following statement with a pipe operator (%>%).

```
library(dplyr)
sum(select(sample_n(filter(flights, origin == "JFK", dest == "PHX"), 200), air_time),
    na.rm = T)
```

```
[1] 57904
```

```
solution <- ""
```

dplyr and data.table

Rewrite the following statement with dplyr and in data.table format.

- "Average departure delay for every flight to Phoenix (PHX) differentiated by carrier and airport of origin."

```
library(dplyr)
library(data.table)

solution_dplyr <- ""

solution_dtable <- ""
```