Assignment 3: Solution

Karsten Donnay

16.6.2021

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. These data cover the airline on-time data for all flights departing NYC in 2013. It also includes useful 'metadata' on airlines, airports, weather, and planes.

```
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
```

```r
# 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    sched_dep_time
 Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 106
 1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.: 906
 Median :2013   Median : 7.000   Median :16.00   Median :1401   Median :1359
   dep_delay          arr_time    sched_arr_time   arr_delay
 Min.   : -43.00   Min.   :   1   Min.   :   1   Min.   : -86.000
 1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124   1st Qu.: -17.000
 Median :  -2.00   Median :1535   Median :1556   Median :  -5.000
   carrier              flight       tailnum             origin
 Length:336776      Min.   :   1   Length:336776      Length:336776
 Class :character   1st Qu.: 553   Class :character   Class :character
 Mode  :character   Median :1496   Mode  :character   Mode  :character
     dest              air_time        distance         hour
 Length:336776      Min.   : 20.0   Min.   :  17   Min.   : 1.00
 Class :character   1st Qu.: 82.0   1st Qu.: 502   1st Qu.: 9.00
 Mode  :character   Median :129.0   Median : 872   Median :13.00
     minute        time_hour
 Min.   : 0.00   Min.   :2013-01-01 05:00:00
 1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
 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 at departure and FALSE otherwise.

```r
flights <- dplyr::mutate(flights, caught_up = arr_delay < dep_delay)
```

(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
del_dep <- dplyr::filter(flights, dep_delay > 0)
print(paste("Overall, ", round(nrow(del_dep) * 100/nrow(flights), 1), "% of the flights had a delayed d
```

```
[1] "Overall,  38.1 % of the flights had a delayed departure."
```

```r
del_dep_del_arr <- dplyr::filter(del_dep, arr_delay > 0)
print(paste("Overall, ", round(nrow(del_dep_del_arr) * 100/nrow(del_dep), 1), "% of the flights with del
```

```
[1] "Overall,  71.9 % of the flights with delayed departure also had a delayed arrival."
```

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)
flights %>%
    dplyr::group_by(origin) %>%
    dplyr::summarise(avg_dep_delay = mean(dep_delay, na.rm = T))
```

```
# A tibble: 3 x 2
  origin avg_dep_delay
  <chr>          <dbl>
1 EWR             15.1
2 JFK             12.1
3 LGA             10.3
```

```
print("EWR has a higher departure delay (15.1) than JFK (12.1)")
```

```
[1] "EWR has a higher departure delay (15.1) than JFK (12.1)"
```

(b) Summary statistics 2

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

```
library(dplyr)
flights %>%
    dplyr::filter(dest == "ORD") %>%
    dplyr::group_by(origin) %>%
    dplyr::summarise(freq = n())
```

```
# A tibble: 3 x 2
  origin  freq
  <chr>  <int>
1 EWR     6100
2 JFK     2326
3 LGA     8857
```

```
print("LGA is the airport where the most flights to ORD depart from.")
```

```
[1] "LGA is the airport where the most flights to ORD depart from."
```

Assignment 3: Rewriting

Piping

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

```
library(dplyr)
set.seed(12345)
sum(select(sample_n(filter(flights, origin == "JFK", dest == "PHX"), 200), air_time),
    na.rm = T)
```

```
[1] 58202
```

```r
set.seed(12345)
flights %>%
    filter(origin == "JFK", dest == "PHX") %>%
    sample_n(200) %>%
    select(air_time) %>%
    sum(na.rm = T)
```

```
[1] 58202
```

dplyr and data.table

Write 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."

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

solution_dplyr <- flights %>%
    filter(dest == "PHX") %>%
    group_by(carrier, origin) %>%
    summarise(avg_dep_delay = mean(dep_delay, na.rm = T)) %>%
    select(carrier, origin, avg_dep_delay)

solution_dtable <- as.data.table(flights)[dest == "PHX", .(avg_dep_delay = mean(dep_delay,
    na.rm = TRUE)), by = list(carrier, origin)]
```