Application exercises: Flights from New York City

YSC2210 - DAVis with R

Michael T. Gastner

The **nycflights13** package includes a tibble called `flights`, which contains information about the on-time record of all flights that departed from New York City (i.e. JFK, LGA or EWR) in 2013. The columns of `flights` are:

- `year`, `month`, `day`: Date of departure.
- `dep_time`, `arr_time`: Actual departure and arrival times (format HHMM or HMM).
- `sched_dep_time`, `sched_arr_time`: Scheduled departure and arrival times.
- `dep_delay`, `arr_delay`: Departure and arrival delays, in minutes.
- `carrier`: Two-letter carrier abbreviation.
- `flight`: Flight number.
- `tailnum`: Plane tail number.
- `origin`, `dest`: Origin and destination.
- `air_time`: Amount of time spent in the air, in minutes.
- `distance`: Distance between airports, in miles.
- `hour`, `minute`: Time of scheduled departure broken into hour and minutes.
- `time_hour`: Scheduled date and hour of the flight as a POSIXct date. (We will not need this column; thus, please do not worry about its format.)

Tasks

Use functions from the **dplyr** package to perform the following tasks.

(1) Find the number of flights for three different subsets: flights that

  (a) had an arrival delay of two or more hours.
  (b) flew from JFK to Houston (IAH or HOU).
  (c) departed between midnight and 6am (inclusive). Be careful: how is midnight represented in `dep_time`?

(2) (a) How many flights have a missing `dep_time`?
  (b) Do all of these flights also have a missing `arr_time`? Write a pipeline with the `|>`-operator that returns the answer as `TRUE` or `FALSE`.
  (c) What flights might be represented by missing `dep_time`?

(3) Which ten destinations had the highest mean air time (conditional on air time being known)? Make a tibble that shows only two columns: destination and mean air time. Sort the rows in descending order of mean air time.

(4) Which ten flights had the slowest speed? Make a tibble that shows only four columns: air time, distance, speed (in miles per hour) and destination. Sort the rows in ascending order of speed.

(5) How can we use the function `ends_with()` to select the columns for the actual and scheduled departure times?

(6) Is there a similar function that we can use to select actual departure time, scheduled departure time and departure delay?

(7) Compare `dep_time`, `sched_dep_time` and `dep_delay` in the tibble created in (6).

  (a) Append a column `diff_time` with the difference between `dep_time` and `sched_dep_time`.
  (b) How would you expect `diff_time` and `dep_delay` to be related? What do you actually see?
  (c) Fix the problem. Confirm that the relation between `dep_time`, `sched_dep_time` and `dep_delay` is as expected.

(8) Make a scatter plot in which each point represents one day and the coordinates are:

- `x`: the mean departure delay (conditional on the depaprture delay being known).
- `y`: the percentage of cancelled flights. We consider a flight as cancelled if the departure time is `NA`.

Label the axes, give the plot a title and credit the data source. Use code chunk options `fig.width`, `fig.height` and `out.width` to adjust figure dimensions. All parts of the figure should be clearly legible without appearing disproportionately large compared to the font size of the running text in the knitted R Markdown file.

Judging from the plot, is the proportion of cancelled flights related to the mean departure delay?

(9) Make a plot that shows the mean departure delay by hour. What time of day should you fly if you want to avoid departure delays?