Exercise 2: Reporting, Data Wrangling and Graphing

Jianhui Gao

07/08/2025

- Quick R
- Rstudio cheatsheet
- Rstudio for beginners

Part 1: Analyze NYC flight delays.

Install the "nycflits13" package. The data comes from the US Bureau of Transportation Statistics. Using the data, complete the following tasks:

1. Find all flights that had an arrival delay of >4 hours, return the first 5 row. (Note: `arr_delay` is in mins)
2. Find all flight names that flew from JFK to IAH, i.e. return only unique values of "flight" variable after filtering. Hint: `unique()` would help.
3. Find how many flights were operated by UA.
4. Find how many unique flights were operated by UA.
5. Sort flights that have the most delayed flights. Show the first 5 row.
6. Generate a scatter plot with x-axis `dist` and y-axis `delay`, where each dot is a unique flights and destination, `dist` is the average distance of each destination `dest`, and `delay` is the average delay time `arr_delay`, with the size of `dot` equals to the count of delay records.

```
library(nycflights13)
head(flights)
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1     1      517            515         2      830            819
## 2  2013     1     1      533            529         4      850            830
## 3  2013     1     1      542            540         2      923            850
## 4  2013     1     1      544            545        -1     1004           1022
## 5  2013     1     1      554            600        -6      812            837
## 6  2013     1     1      554            558        -4      740            728
## # i 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```