Lab 1 - Math 58 / 58b: Introduction to Data

Jo Hardin

due Jan 28, 2020

The Bureau of Transportation Statistics (BTS) is a statistical agency that is a part of the Research and Innovative Technology Administration (RITA). As its name implies, BTS collects and makes available transportation data, such as the flights data we will be working with in this lab.

```
data(flights)
```

To Turn In

5. Mutate the data frame so that it includes a new variable that contains the average speed, `avg_speed` traveled by the plane for each flight (in mph). **Hint:** Average speed can be calculated as distance divided by number of hours of travel, and note that `air_time` is given in minutes.

Solution

```
flights %>%
  names()
```

```
##  [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"
```

6. Another useful `dplyr` filtering helper function is `between`. What does it do? Use it to find flights that arrived between 0 and 60 minutes late. How many such flights are there?

Solution

Some words here describing what I see below.

```
flights_ORD <- flights %>%
  dplyr::filter(dest == "ORD") %>%
  select(dep_time, dep_delay, arr_time, arr_delay)

summary(flights_ORD, na.rm=TRUE)
```

```
##     dep_time      dep_delay          arr_time      arr_delay
##  Min.   :   1   Min.   : -20.00   Min.   :   1   Min.   : -62.000
##  1st Qu.: 853   1st Qu.:  -5.00   1st Qu.:1015   1st Qu.: -20.000
##  Median :1329   Median :  -2.00   Median :1448   Median :  -8.000
##  Mean   :1310   Mean   :  13.57   Mean   :1449   Mean   :   5.877
##  3rd Qu.:1721   3rd Qu.:  11.00   3rd Qu.:1900   3rd Qu.:  13.000
##  Max.   :2400   Max.   :1126.00   Max.   :2359   Max.   :1109.000
##  NA's   :641    NA's   :641       NA's   :676    NA's   :717
```

7. Suppose you really dislike departure delays, and you want to schedule your travel in a month that minimizes your potential departure delay leaving NYC. One option is to choose the month with the lowest mean departure delay. Another option is to choose the month with the lowest median departure delay. What are the pros and cons of these two choices? Which month do you choose?

Solution

Some words here describing what I see below.

```
flights %>%
  group_by(carrier) %>%
  summarize(min_flight = min(air_time, na.rm = TRUE),
            mean_flight = mean(air_time, na.rm = TRUE),
            med_flight = median(air_time, na.rm = TRUE),
            max_flight = max(air_time, na.rm = TRUE))
```

```
## # A tibble: 16 x 5
##    carrier min_flight mean_flight med_flight max_flight
##    <chr>        <dbl>       <dbl>      <dbl>      <dbl>
##  1 9E              21        86.8       83          272
##  2 AA              29       189.       169          426
##  3 AS             277       326.       324          392
##  4 B6              29       151.       142          413
##  5 DL              26       174.       145          490
##  6 EV              20        90.1       87          286
##  7 F9             195       230.       229          278
##  8 FL              53       101.       109          161
##  9 HA             580       623.       622.         691
## 10 MQ              33        91.2       83          236
## 11 OO              50        83.5       68          177
## 12 UA              23       212.       197          695
## 13 US              21        88.6       76          359
## 14 VX             264       337.       337          406
## 15 WN              31       148.       122          362
## 16 YV              32        65.7       56.5        122
```

8. Which month has the highest average arrival delay from an NYC airport? What about the highest median arrival delay? Which of these measures is more reliable for deciding which month(s) to avoid flying if you really dislike delayed flights.

Solution

Some words here describing what I see below.

```
flights %>%
  group_by(carrier, origin) %>%
  summarize(n())
```

```
## # A tibble: 35 x 3
## # Groups:   carrier [16]
##    carrier origin `n()`
##    <chr>   <chr>  <int>
##  1 9E      EWR     1268
##  2 9E      JFK    14651
##  3 9E      LGA     2541
##  4 AA      EWR     3487
##  5 AA      JFK    13783
##  6 AA      LGA    15459
##  7 AS      EWR      714
##  8 B6      EWR     6557
##  9 B6      JFK    42076
## 10 B6      LGA     6002
## # ... with 25 more rows
```