Check-in 2

Jonathan Manzano

2024-02-14

A. Flights Data

1. For this lab, we'll be using the `flights` data. You can load this data using the following code:

```
library(tidyverse)
library(nycflights13)
data(flights)
```

2. Use the `mutate()` function to turn `origin` into a factor.

```
flights <- flights |>
  mutate(origin = factor(origin))

str(flights$origin)
```

```
 Factor w/ 3 levels "EWR","JFK","LGA": 1 3 2 2 3 1 1 3 2 3 ...
```

3. Compute the mean `arr_delay` for each origin airport. Which airport has the longest delays on average? You can drop missing rows.

```
origin_arr_delay_means <- flights |>
  filter(!is.na(arr_delay)) |>
  group_by(origin) |>
  summarize(mean_arr_delay = mean(arr_delay, na.rm = TRUE))

print(origin_arr_delay_means)
```

```
# A tibble: 3 x 2
  origin mean_arr_delay
  <fct>           <dbl>
1 EWR              9.11
2 JFK              5.55
3 LGA              5.78
```

```
longest_avg_delay <- origin_arr_delay_means |>
  arrange(desc(mean_arr_delay)) |>
  top_n(1, mean_arr_delay)

print(longest_avg_delay)
```

```
# A tibble: 1 x 2
  origin mean_arr_delay
  <fct>           <dbl>
1 EWR              9.11
```

The airport that has the longest delays on average is EWR.

4. Use the `mutate()` function to turn `month` into a factor.

```
flights <- flights |>
  mutate(month = factor(month))

str(flights$month)
```

```
 Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
```

5. Compute the mean `arr_delay` by month. What do you notice?

```
month_arr_delay_means <- flights |>
  filter(!is.na(arr_delay)) |>
  group_by(month) |>
  summarize(mean_arr_delay_month = mean(arr_delay, na.rm = TRUE))

print(month_arr_delay_means)
```

```
# A tibble: 12 x 2
   month mean_arr_delay_month
   <fct>                <dbl>
 1 1                     6.13
 2 2                     5.61
 3 3                     5.81
 4 4                    11.2
 5 5                     3.52
 6 6                    16.5
 7 7                    16.7
 8 8                     6.04
 9 9                    -4.02
10 10                   -0.167
11 11                    0.461
12 12                   14.9
```

```
longest_avg_delay_month <- month_arr_delay_means |>
  arrange(desc(mean_arr_delay_month)) |>
  top_n(1, mean_arr_delay_month)

print(longest_avg_delay_month)
```

```
# A tibble: 1 x 2
  month mean_arr_delay_month
  <fct>                <dbl>
1 7                     16.7
```

The longest `arr_delay` is July. The difference is that since we are computing the mean based on different groupings, the means will differ. In Q4 it was based on the location of distinct airports origin regardless of the month. For this question its is based specific months but from all airport origins.

6. Compute the mean `arr_delay` by month AND `origin`. What do you notice?

```
month_origin_means <- flights |>
  filter(!is.na(arr_delay)) |>
  group_by(month, origin) |>
  summarize(mean_arr_delay_month_origin = mean(arr_delay, na.rm = TRUE), .groups = 'drop')

print(month_origin_means)
```

```
# A tibble: 36 x 3
   month origin mean_arr_delay_month_origin
   <fct> <fct>                         <dbl>
 1 1     EWR                           12.8
 2 1     JFK                            1.37
 3 1     LGA                            3.38
 4 2     EWR                            8.78
 5 2     JFK                            4.39
 6 2     LGA                            3.15
 7 3     EWR                           10.6
 8 3     JFK                            2.58
 9 3     LGA                            3.74
10 4     EWR                           14.1
# i 26 more rows
```

```
highest_avg_temp_location <- month_origin_means |>
  arrange(desc(mean_arr_delay_month_origin)) |>
  slice(1)

print(highest_avg_temp_location)
```

```
# A tibble: 1 x 3
  month origin mean_arr_delay_month_origin
  <fct> <fct>                         <dbl>
1 7     JFK                            20.2
```

The JFK airport in the month of July has the longest `arr_delay`. I notice not that that are it has outputted a table with each month and for each origin, it has computed for that month and origin's mean `arr_delay`.