Session 3 Markdown

Robert Atkinson

04/11/2020

Introduction

In this exercise flight data for planes departing from New York City airport will be used to calculate which destinations have on average the longest delay time.

Load R Libraries

```
library(tidyverse)
```

```
## -- Attaching packages ----------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts -------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
```

Add Flight Data

- Load flight data for planes arrving and departing New York City airport.
- Select just the flights from this data

Pre-requisites: ensure nycflights13 package has been installed, if not run: **install.packages("nycflights13")**

```
library(nycflights13)


flights_from_nyc <- nycflights13::flights
```

Analysis

To show the results of the analysis a table will be created,

illustrating the average departure delay for the 5 most delayed destinations.

1. Start from the flights_from_nyc data.

2. Select origin, destination, departure delay, year, month, and day.
3. Filter only rows referring to flights in November.
4. Filter only rows where departure delay is not NA.
5. Group by destination.
6. Calculated the average delay per destination.
7. Add a column with the delay calculated in hours (minutes over 60).
8. Sort the table by descending delay (note that - is used before the column name).
9. Only show the first 5 rows.
10. Create a well-formatted table.

```
flights_from_nyc %>%
  # Select origin, destination, departure delay, year, month, and day.
  dplyr::select(origin, dest, year, month, day, dep_delay) %>%

  # Filter only rows referring to flights in November.
  dplyr::filter(month == 11) %>%

  # Filter only rows where departure delay is not NA.
  dplyr::filter(!is.na(dep_delay)) %>%

  # Group by destination.
  dplyr::group_by(dest) %>%

  # Calculate the average delay per destination.
  dplyr::summarize(
    avg_dep_delay = mean(dep_delay)
  ) %>%

  # Add a column with the delay calculated in hours (minutes over 60).
  dplyr::mutate(
    avg_dep_delay_hours = avg_dep_delay / 60
  ) %>%

  # Sort the table by descending delay
  dplyr::arrange(-avg_dep_delay_hours) %>%

  # Only show the first 5 rows.
  dplyr::slice_head(n = 5) %>%

  # Create a well-formatted table.
  knitr::kable()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

dest	avg_dep_delay	avg_dep_delay_hours
SBN	67.50000	1.1250000
BDL	26.66667	0.4444444
CAK	19.70909	0.3284848
BHM	19.61905	0.3269841
DSM	16.14815	0.2691358

Conclusion

To conclude, the 5 most delayed flights from New York City airport are to the destinations: South Bend Internation (SBN), Bradley International (BDL), Akron-Canton (CAK), Birmingham (BHM), and Des Moines International (DSM). The average of 4 of these is less than 30 minutes. However, South Bend International (SBN) is, on avergae, delayed by more than twice any other destination. With an avergage departure delay of 1.125, equating to 67.5 minutes.