Math 300 Lesson 6 Notes

group_by, mutate, and arrange

YOUR NAME HERE

June, 2022

Contents

Objectives 1
Reading 1
Lesson 1
Documenting software 4

Objectives

1. Use the `group_by()` function to create aggregated data frames to use with other functions, in particular `summarize()`, to explore, explain, and visualize.
2. Use the `mutate()` function to create new variables in a data frame in order to explore, explain, and visualize.
3. Use the `arrange()` function to sort data frames to explore, explain, and visualize.

Reading

Chapter 3.4 - 3.5

Lesson

Work through the learning checks LC3.5 - LC3.12. Complete code as necessary.

- It is important to note that the `group_by()` function doesn't change data frames by itself. Rather it changes the meta-data, or data about the data, specifically the grouping structure. It is only after we apply the `summarize()` function that the data frame changes. The book does a good job explaining meta-data.
- The `group_by()` can be used on more than two variables but they must be in the same call to `group_by()`.
- Using `arrange()` is straightforward except for the use of `desc()` within the `arrange()` call to sort in decreasing order.
- As a rough rule of thumb, as long as you are not losing original information that you might need later, it's acceptable practice to overwrite existing data frames with updated ones.

Setup

```
library(nycflights13)
library(ggplot2)
library(dplyr)
```

LC 3.5 (Objective 1)

(LC 3.5) Recall from Chapter 2 when we looked at plots of temperatures by months in NYC. What does the standard deviation column in the `summary_monthly_temp` data frame, which we need to create from the code at the section 3.4, tell us about temperatures in New York City throughout the year?

Solution:

```
# Complete the code and then remove comment symbols
# summary_temp_by_month <- _________ %>%
#   group_by(______) %>%
#   summarize(
#     mean = mean(_______, na.rm = TRUE),
#     std_dev = sd(_______, na.rm = TRUE)
#   )
```

LC 3.6 (Objective 1)

(LC 3.6) What code would be required to get the mean and standard deviation temperature for each day in 2013 for NYC?

Solution:

```
#Complete the code and then remove comment symbols
# summary_temp_by_day <- weather %>%
#   group_by(________, ________) %>%
#   summarize(
#     mean = mean(temp, na.rm = TRUE),
#     std_dev = sd(temp, na.rm = TRUE)
#   )
```

```
#head(summary_temp_by_day)
```

LC 3.7 (Objective 1)

(LC 3.7) Recreate `by_monthly_origin`, but instead of grouping via `group_by(origin, month)`, group variables in a different order `group_by(month, origin)`. What differs in the resulting dataset?

Solution:

```
# Complete the code and then remove comment symbols
# by_origin_monthly <- flights %>%
#   group_by(________, __________) %>%
#   summarize(count = ______)
```

```
# Complete the code and then remove comment symbols
# by_monthly_origin <- flights %>%
#   group_by(________, __________) %>%
#   summarize(count = ______)
```

LC 3.8 (Objective 1)

(LC 3.8) How could we identify how many flights left each of the three airports for each `carrier`?

Solution:

```
# Complete the code and then remove comment symbols
# count_flights_by_airport <- flights %>%
#   group_by(_________, _________) %>%
#   summarize(count = _______)
```

```
# Complete the code and then remove comment symbols
#head(count_flights_by_airport,n=______)
```

LC 3.9 (Objective 1)

(LC 3.9) How does the `filter` operation differ from a `group_by` followed by a `summarize`?

Solution:

LC 3.10 (Objective 2)

(LC 3.10) What do positive values of the `gain` variable in `flights` correspond to? What about negative values? And what about a zero value?

Solution:

LC 3.11 (Objective 2)

(LC 3.11) Could we create the `dep_delay` and `arr_delay` columns by simply subtracting `dep_time` from `sched_dep_time` and similarly for arrivals? Try the code out and explain any differences between the result and what actually appears in `flights`.

Solution:

```
# Complete the code and then remove the comment symbols
# LC3.11<- flights %>%
#   mutate(time_gain=dep_time-arr_time, gain = _____________) %>%
#   select(air_time,dep_time,arr_time,time_gain,dep_delay,arr_delay, gain)
```

LC 3.12 (Objective 2)

(LC 3.12) What can we say about the distribution of `gain`? Describe it in a few sentences using a boxplot and the `gain_summary` data frame values.

Solution: We must create the data frame from the notes. We have to copy and combine two chunks of code.

```
# Complete the code and then remove comment symbols
# gain_summary <- ______ %>%
#   mutate(gain = _______ - __________) %>%
#   summarize(
#     min = min(_________, na.rm = TRUE),
#     q1 = quantile(gain, 0.25, na.rm = TRUE),
#     median = quantile(gain, 0.5, na.rm = TRUE),
#     q3 = quantile(gain, 0.75, na.rm = TRUE),
#     max = max(gain, na.rm = TRUE),
#     mean = _________(gain, na.rm = TRUE),
#     sd = _________(gain, na.rm = TRUE),
#     missing = sum(is.na(__________))
#   )
```

```
# Complete the code and then remove the comment symbols
# flights %>%
#   mutate(gain = _____________ - ____________) %>%
#   ggplot(aes(x=1,y=gain)) +
#   __________() +
#  theme_classic()
```

Documenting software

- File creation date: 2022-06-16
- R version 4.1.3 (2022-03-10)
- `ggplot2` package version: 3.3.6
- `dplyr` package version: 1.0.9
- `nycflights13` package version: 1.0.2