Math 300 Lesson 3 Notes

Histograms and Facets

YOUR NAME HERE

June, 2022

Contents

Objectives

1. Understand and wrangle data in preparation for plotting.
2. Explain when it is appropriate to use a linegraphs, histogram, and facets.
3. Create and interpret a linegraphs using the `ggplot()` function.
4. Create and interpret a histogram using the `ggplot()` function.
5. Use facets to improve the visual presentation of data and then interpret plots that use facets.

Reading

Chapter 2.4 - 2.6

Lesson

Work through the learning checks LC2.9 - LC2.21. There appear to be a large number of learning checks but many of them are linked together. Complete the code when necessary.

- We are looking a different plots but the need for a particular plot changes based on the nature of our data. For the linegraph, the variable we choose to put on the x-axis has a sequential nature, usually time. For the histogram we want to understand the distribution of a single quantitative variable. Facets allow us to bring in another variable, usually categorical.
- We will be using `filter()` again to create a subset of data but we are putting in an **and** condition. This is required for some of the learning checks. Complete the setup code.

Setup

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

We need to create the `early_january_weather` data object.

Use ?weather to understand the variables.

```
glimpse(weather)
```

```
## Rows: 26,115
## Columns: 15
## $ origin     <chr> "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EW~
## $ year       <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,~
## $ month      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ day        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ hour       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, ~
## $ temp       <dbl> 39.02, 39.02, 39.02, 39.92, 39.02, 37.94, 39.02, 39.92, 39.~
## $ dewp       <dbl> 26.06, 26.96, 28.04, 28.04, 28.04, 28.04, 28.04, 28.04, 28.~
## $ humid      <dbl> 59.37, 61.63, 64.43, 62.21, 64.43, 67.21, 64.43, 62.21, 62.~
## $ wind_dir   <dbl> 270, 250, 240, 250, 260, 240, 240, 250, 260, 260, 260, 330,~
## $ wind_speed <dbl> 10.35702, 8.05546, 11.50780, 12.65858, 12.65858, 11.50780, ~
## $ wind_gust  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 20.~
## $ precip     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ pressure   <dbl> 1012.0, 1012.3, 1012.5, 1012.2, 1011.9, 1012.4, 1012.2, 101~
## $ visib      <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,~
## $ time_hour  <dttm> 2013-01-01 01:00:00, 2013-01-01 02:00:00, 2013-01-01 03:00~
```

```
# This code finds the flights from EWR between January 1 up
# to and including January 15

#early_january_weather <- weather %>%
#  filter(origin == "______" & month == ______ & day <= 15)
```

LC 2.9 (Objective 1)

(LC 2.9) Take a look at both the `weather` and `early_january_weather` data frames by running `View(weather)` and `View(early_january_weather)` in the console. In what respect do these data frames differ?

Solution:

LC 2.10 (Objective 1)

(LC 2.10) `View()` the `flights` data frame again. Why does the `time_hour` variable uniquely identify the hour of the measurement whereas the `hour` variable does not?

Solution:

LC 2.11 (Objective 2)

(LC 2.11) Why should linegraphs be avoided when there is not a clear ordering of the horizontal axis?

Solution:

LC 2.12 (Objective 2)

(LC 2.12) Why are linegraphs frequently used when time is the explanatory variable?

Solution:

LC 2.13 (Objective 3)

(LC 2.13) Plot a time series of a variable other than `temp` for Newark Airport in the first 15 days of January 2013.

First let's create the plot in the book.

```
# Complete code and remove comment symbol
#ggplot(data = early_january_weather,
#       mapping = aes(x = __________, y = ________)) +
#  geom_line()
```

Solution:

LC 2.14 (Objective 4)

- A list of the available colors

```
head(colors())
```

```
## [1] "white"         "aliceblue"     "antiquewhite"  "antiquewhite1"
## [5] "antiquewhite2" "antiquewhite3"
```

(LC 2.14) What does changing the number of bins from 30 to 40 tell us about the distribution of temperatures?

Solution:

```
# Complete code and remove comment symbol
# ggplot(data = weather, mapping = aes(x = temp)) +
#   geom_histogram(bins=____,color = "white", fill = "steelblue")
```

LC 2.15 (Objective 4)

(LC 2.15) Would you classify the distribution of temperatures as symmetric or skewed?

Solution:

LC 2.16 (Objective 4)

(LC 2.16) What would you guess is the "center" value in this distribution? Why did you make that choice?

Solution:

LC 2.17 (Objective 4)

(LC 2.17) Is this data spread out greatly from the center or is it close? Why?

This plot may help you answer the question.

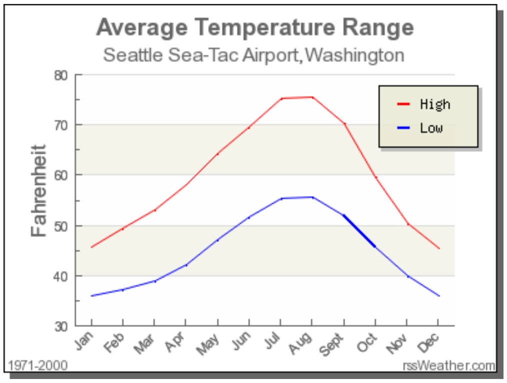


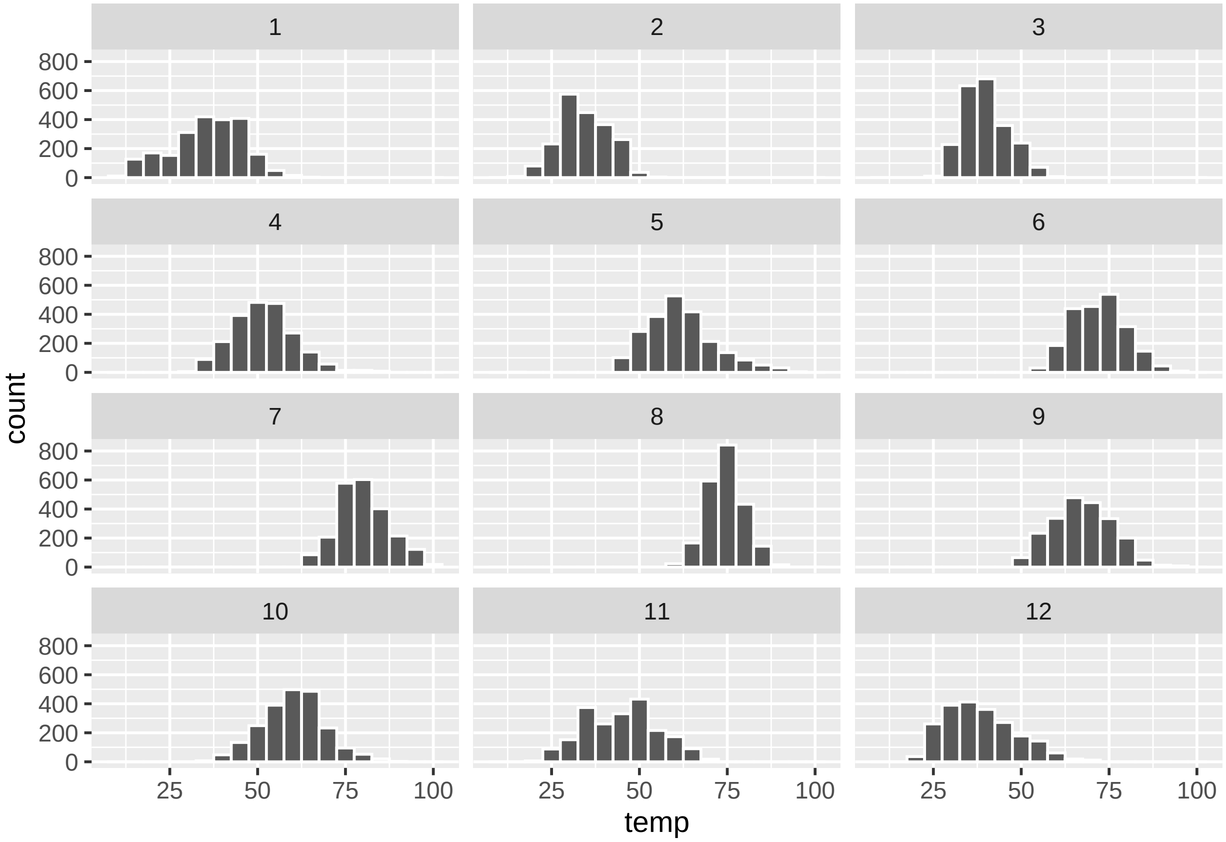
Figure 1: Annual temperatures at SEATAC Airport.

Solution:

LC 2.18 (Objective 5)

```r
ggplot(data = weather, mapping = aes(x = temp)) +
  geom_histogram(binwidth = 5, color = "white") +
  facet_wrap(~ month, nrow = 4)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



(LC 2.18) What other things do you notice about the faceted plot above? How does a faceted plot help us see relationships between two variables?

Solution:

LC 2.19 (Objective 5)

(LC 2.19) What do the numbers 1-12 correspond to in the plot above? What about 25, 50, 75, 100?

Solution:

LC 2.20 (Objective 2, 5)

(LC 2.20) For which types of datasets would these types of faceted plots not work well in comparing relationships between variables? Give an example describing the nature of these variables and other important characteristics.

Solution:

LC 2.21 (Objective 5)

(LC 2.21) Does the `temp` variable in the `weather` dataset have a lot of variability? Why do you say that?

Solution:

Documenting software

- File creation date: 2022-06-04
- 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