Math 300 Lesson 2 Notes

Professor Bradley Warner

May, 2022

Contents

Objectives 1
Reading 1
Lesson 1
Documenting software 5

Objectives

1. Create a scatterplot using the `ggplot()` function.
2. Interpret the relationship between variables in a scatterplot.
3. Refine and improve scatterplots to illustrate relevant points by prepossessing the data or using functions such as `alpha()` and `geom_jitter()`.

Reading

Chapter 2 - 2.3

Lesson

Work through the learning checks LC2.1 - LC2.8. Complete the code when necessary.

Setup

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

We need to create the ***alaska_flights*** *data object.* Complete the code and remove the comment symbol #.

```
#alaska_flights <- ____________ %>%
#  filter(carrier == "______")
```

LC 2.1 (Objective 3)

(LC 2.1) Take a look at both the `flights` and `alaska_flights` data frames by running `View(flights)` and `View(alaska_flights)` in the console. In what respect do these data frames differ? For example, think about the number of rows in each dataset.

Solution:

Build the plot for the next set of learning checks. Complete the code and remove the comment symbol #.

```
#ggplot(data = _______________, mapping = aes(x = ________________, y = arr_delay)) +
#  geom_point()
```

LC 2.2 (Objective 2)

(LC 2.2) What are some practical reasons why `dep_delay` and `arr_delay` have a positive relationship?

Solution:

LC 2.3 (Objective 2)

(LC 2.3) What variables in the `weather` data frame would you expect to have a negative correlation (i.e. a negative relationship) with `dep_delay`? Why? Remember that we are focusing on numerical variables here. Hint: Explore the `weather` dataset by using the `View()` function.

Solution:

LC 2.4 (Objective 2)

(LC 2.4) Why do you believe there is a cluster of points near (0, 0)? What does (0, 0) correspond to in terms of the Alaskan flights?

Solution:

LC 2.5 (Objective 2)

(LC 2.5) What are some other features of the plot that stand out to you?

Solution: Different people will answer this one differently. One answer is most flights depart and arrive less than an hour late.

LC 2.6 (Objective 1)

(LC 2.6) Create a new scatterplot using different variables in the `alaska_flights` data frame by modifying the example above.

To insert an `R` code chunk into a markdown, there is the pulldown menu but you can also use Ctrl-Alt-I.

Solution:

```
# Insert plot code here.
```

LC 2.7 (Objective 2)

(LC 2.7) Why is setting the `alpha` argument value useful with scatterplots? What further information does it give you that a regular scatterplot cannot?

Solution:

LC 2.8 (Objective 2, 3)

```
#Plot to use for this problem.
ggplot(data = alaska_flights, mapping = aes(x = dep_delay, y = arr_delay)) +
  geom_point()
```

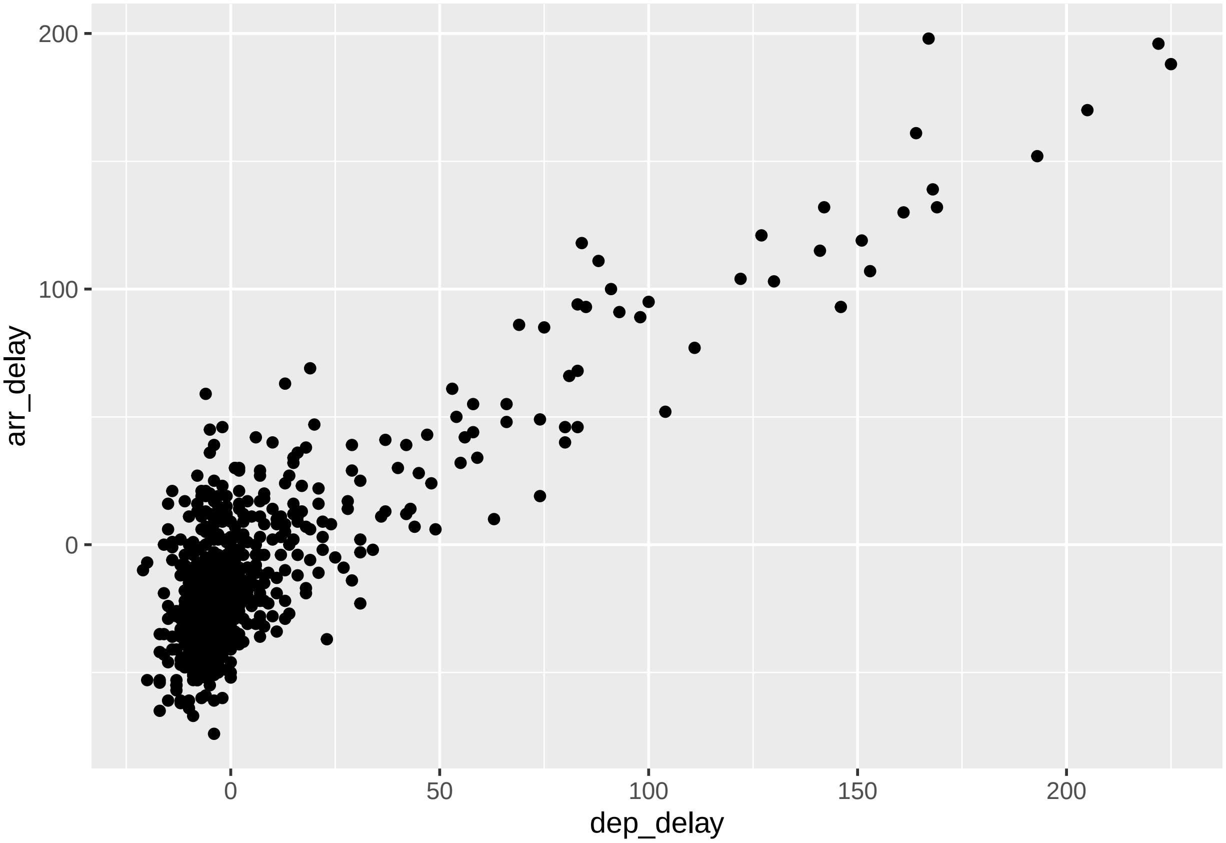

Figure 1: Figure 1: Arrival delays versus departure delays for Alaska Airlines flights from NYC in 2013.

```
#Second Plot to use for this problem.
ggplot(data = alaska_flights, mapping = aes(x = dep_delay, y = arr_delay)) +
  geom_point(alpha = 0.2)
```

(LC 2.8) After viewing the Figure 2 above, give an approximate range of arrival delays and departure delays that occur the most frequently. How has that region changed compared to when you observed the same plot without the `alpha = 0.2` set in Figure 1?

Solution:

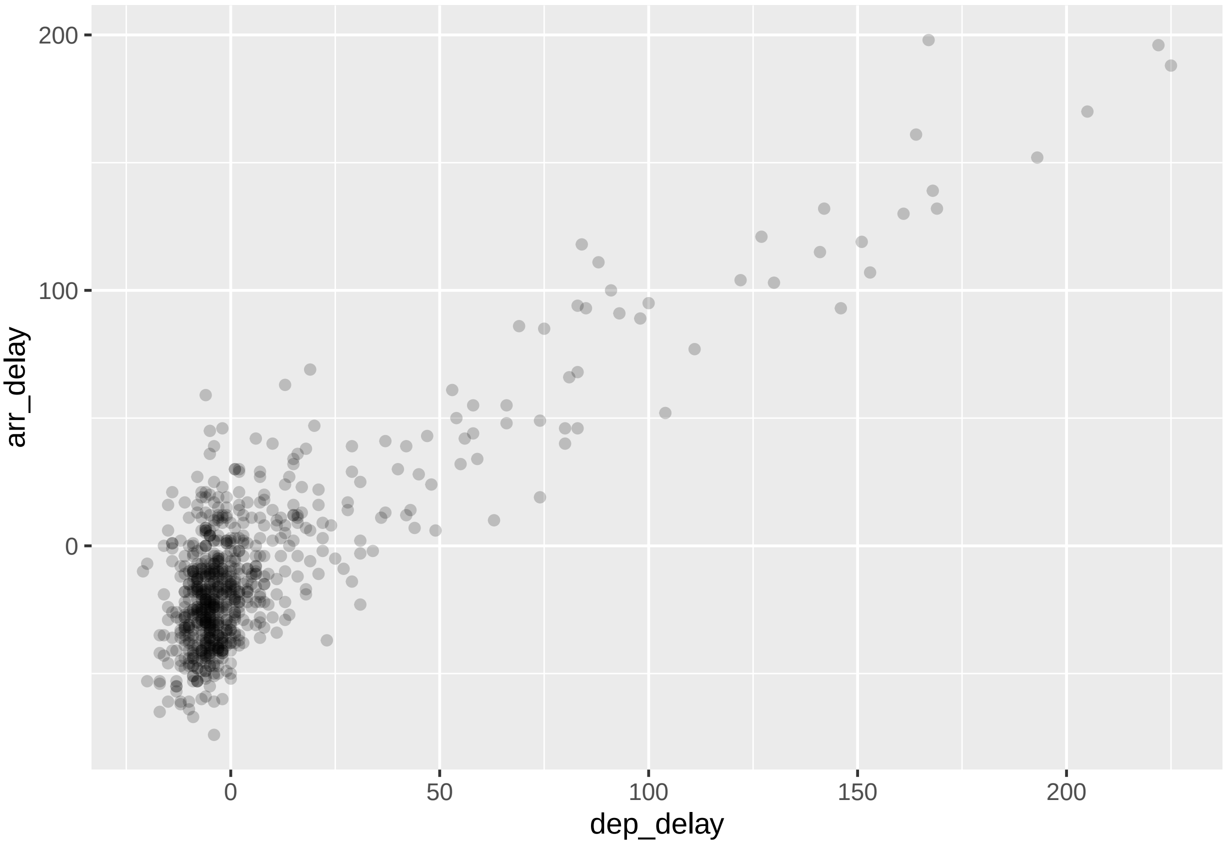

Figure 2: Figure 2: Arrival vs. departure delays scatterplot with alpha = 0.2

Documenting software

- File creation date: 2022-05-20
- 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