Lab report

Load data

```
nycflights <- read_csv("https://dyurovsky.github.io/85309/data/lab2/nycflights.csv")
```

```
## Rows: 32735 Columns: 16
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): carrier, tailnum, origin, dest
## dbl (12): year, month, day, dep_time, dep_delay, arr_time, arr_delay, flight...
```

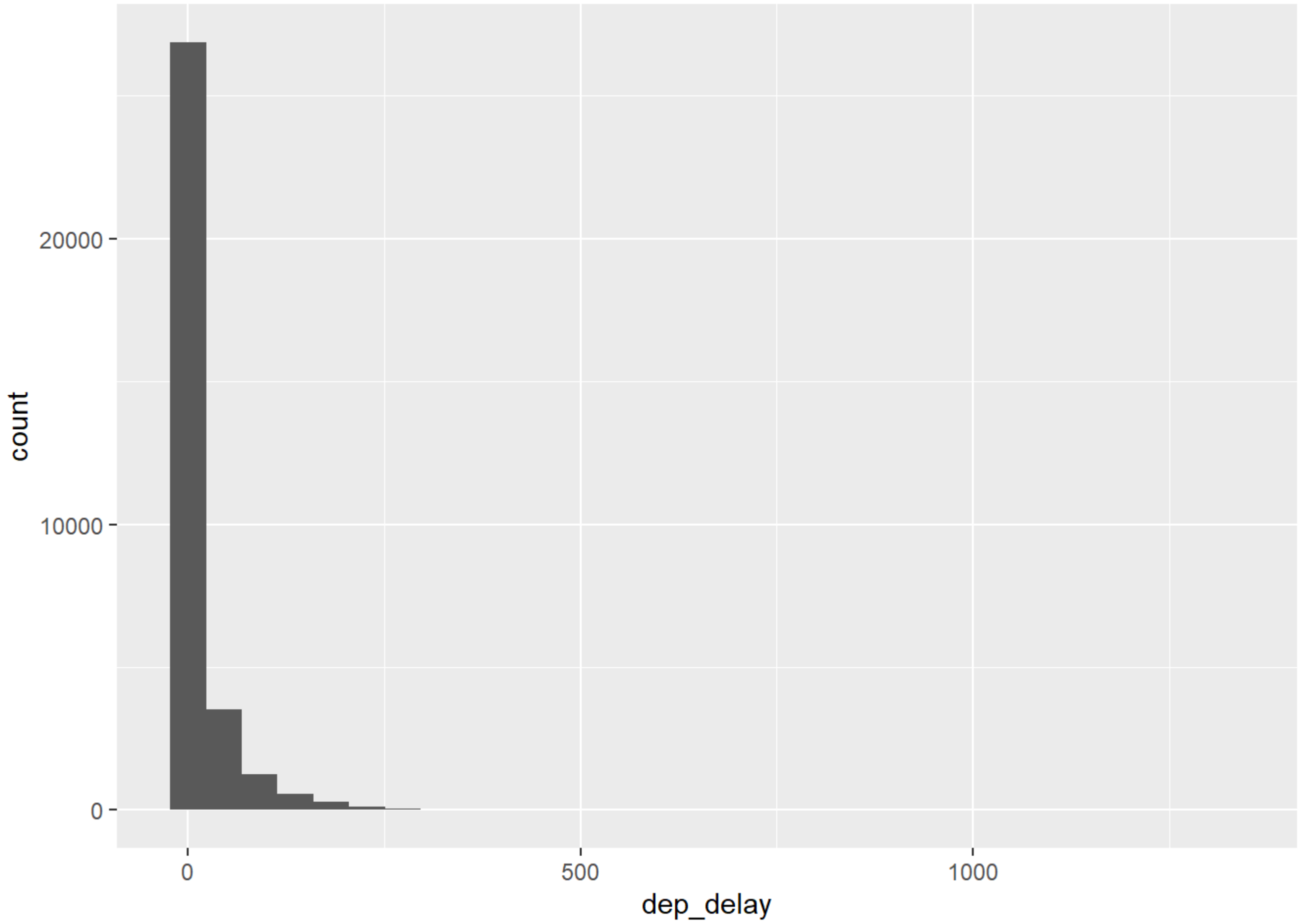
```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Exercise 1:

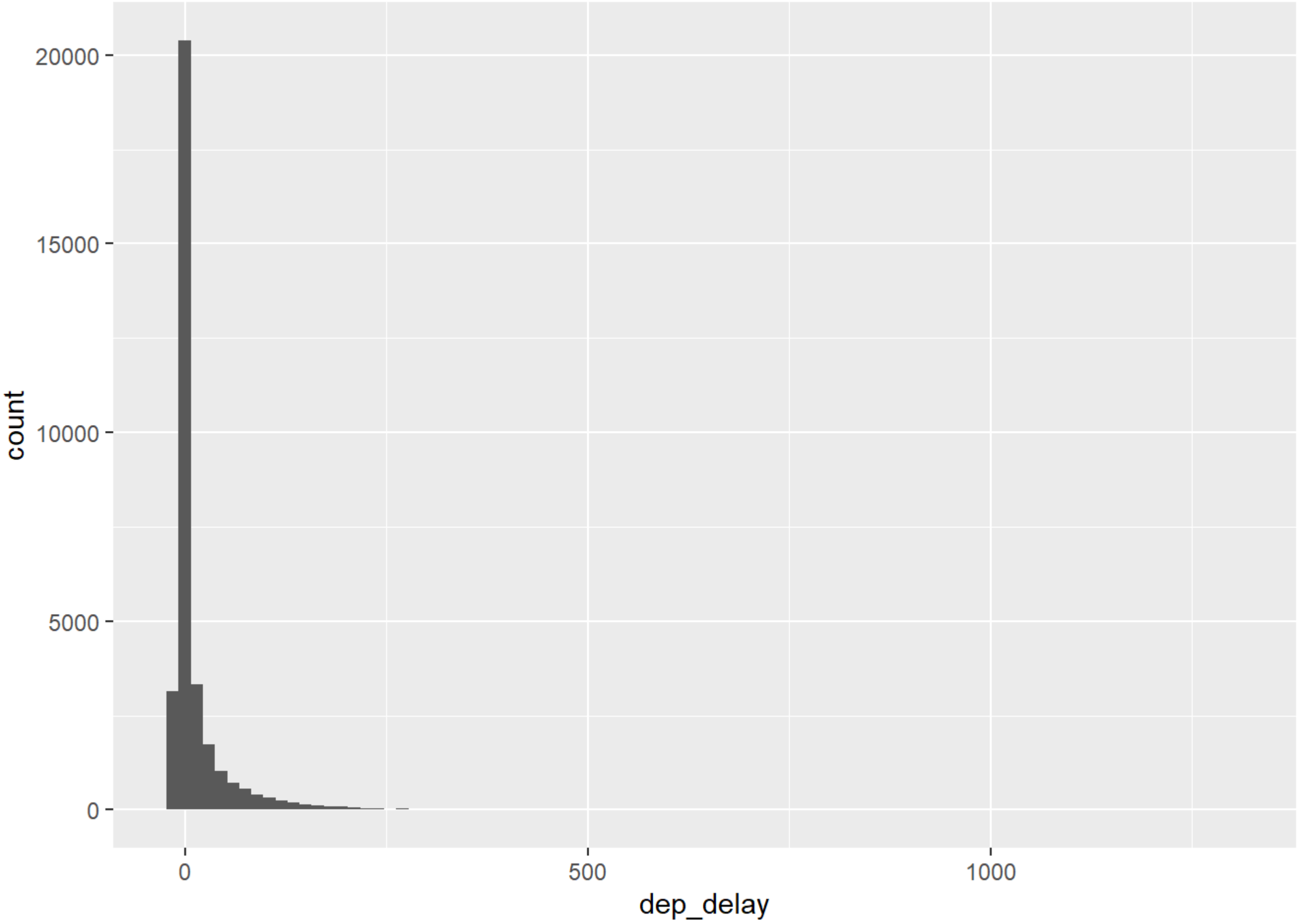
In the binwidth of 15 plots, you can see that some of the flights leave early, and you can see more of the details of the distribution. In the binwidth of 150 plots, almost all of the flights end up in one bin, which maybe doesn’t provide enough detail to get a sense for how they’re distributed. The default plot is in between.

```
# enter your code for Exercise 1 here
ggplot(nycflights, aes(x = dep_delay)) +
  geom_histogram()
```

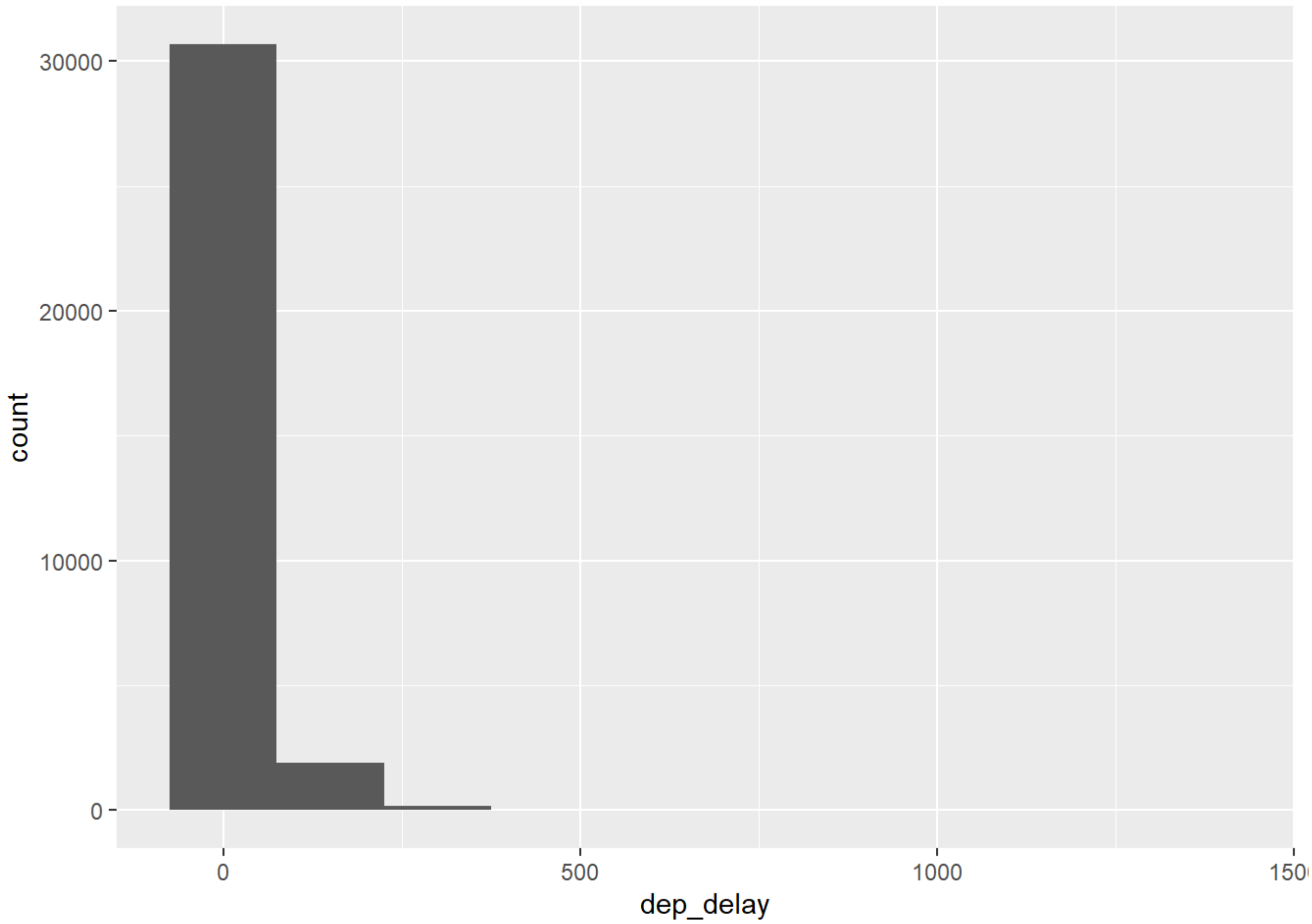
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(nycflights, aes(x = dep_delay)) +
  geom_histogram(binwidth = 15)
```



```
ggplot(nycflights, aes(x = dep_delay)) +
  geom_histogram(binwidth = 150)
```



Exercise 2:

68 flights have headed to SFO in February.

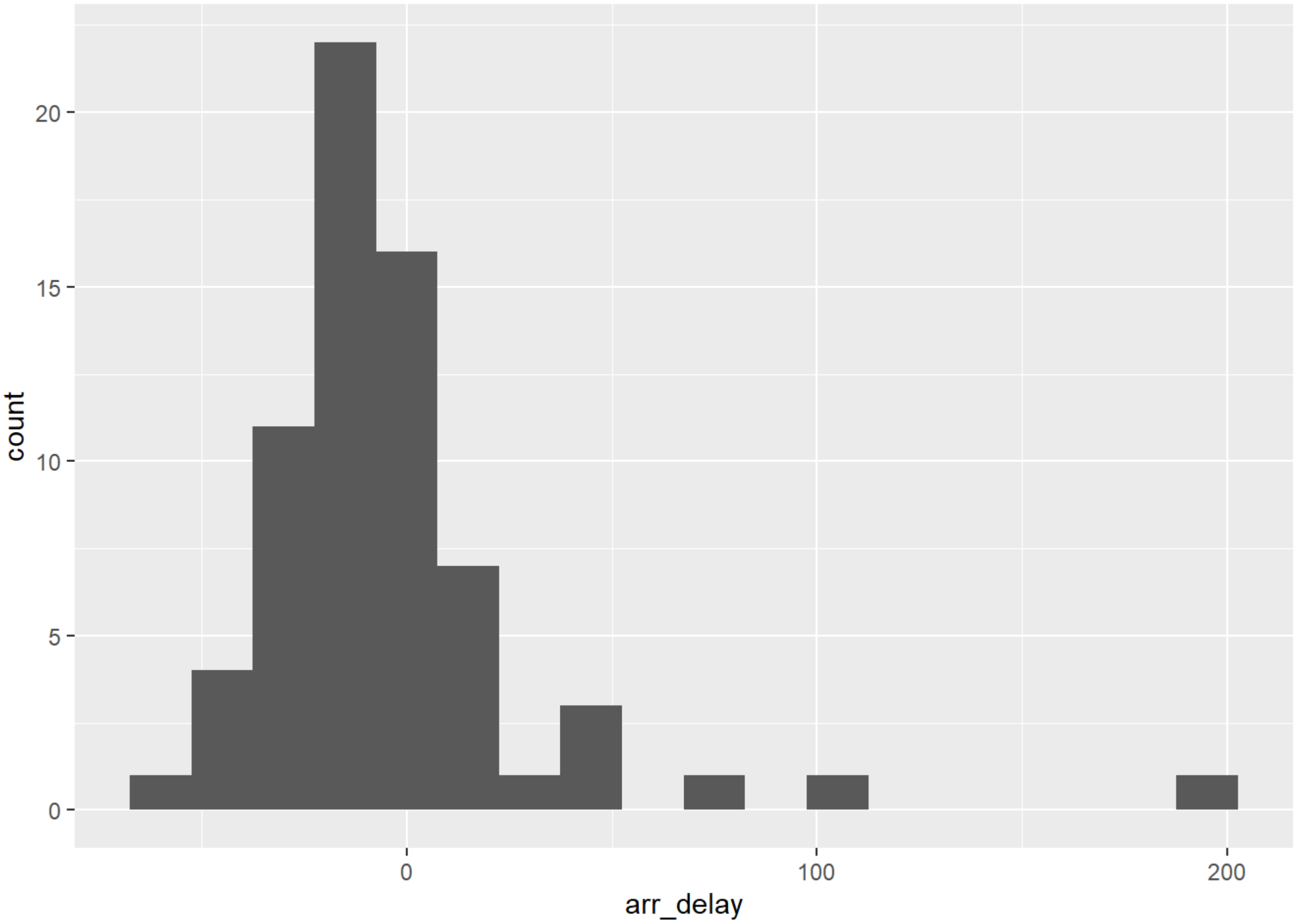
```
# enter your code for Exercise 2 here
sfo_feb_flights <- nycflights %>%
  filter(dest == "SFO", month == 2)
nrow(sfo_feb_flights)
```

```
## [1] 68
```

Exercise 3:

The histogram shows a right-skewed distribution, because of this skewness, the median and IQR would be appropriate to use.

```
# enter your code for Exercise 3 here
ggplot(sfo_feb_flights, aes(x = arr_delay)) +
  geom_histogram(binwidth = 15)
```



```
sfo_feb_flights %>%
  summarise(median = median(arr_delay),
            iqr = IQR(arr_delay))
```

```
## # A tibble: 1 x 2
##   median   iqr
##    <dbl> <dbl>
## 1    -11  23.2
```

Exercise 4:

IQR is a measure of variability. DL and UA have the most variable delays because they have the largest IQRs. we should care IQR not median, because the flight departure has more unpredictability.

```
# enter your code for Exercise 4 here
sfo_feb_flights %>%
  group_by(carrier) %>%
  summarise(median_ad = median(arr_delay),
            iqr_ad = IQR(arr_delay),
            n_flights = n())
```

```
## # A tibble: 5 x 4
##   carrier median_ad iqr_ad n_flights
##   <chr>       <dbl>  <dbl>     <int>
## 1 AA            5     17.5        10
## 2 B6          -10.5   12.2         6
## 3 DL          -15     22          19
## 4 UA          -10     22          21
## 5 VX          -22.5   21.2        12
```

Exercise 5:

The mean of a distribution will be affected more by values in the tails. That means that the mean of arrival day will have information about extreme delay values whereas the median will not. If you want to make sure you don’t have a REALLY long delay, you should use the mean. But if you want to minimize the “typical” delay, you should use the median.so, if you care a lot about delays maybe you need to use mean not median.

```
# enter your code for Exercise 5 here
nycflights %>%
  group_by(month) %>%
  summarise(mean_dd = mean(dep_delay)) %>%
  arrange(desc(mean_dd))
```

```
## # A tibble: 12 x 2
##    month mean_dd
##    <dbl>   <dbl>
##  1     7   20.8
##  2     6   20.4
##  3    12   17.4
##  4     4   14.6
##  5     3   13.5
##  6     5   13.3
##  7     8   12.6
##  8     2   10.7
##  9     1   10.2
## 10     9    6.87
## 11    11    6.10
## 12    10    5.88
```

Exercise 6:

I would select LGA because it has the highest on time departure percentage.

```
# enter your code for Exercise 6 here
nycflights <- nycflights %>%
  mutate(dep_type = if_else(dep_delay < 5, "on time", "delayed"))

nycflights %>%
  group_by(origin) %>%
  summarise(ot_dep_perc = (sum(dep_type == "on time") / n()) * 100) %>%
  arrange(desc(ot_dep_perc))
```

```
## # A tibble: 3 x 2
##   origin ot_dep_perc
##   <chr>        <dbl>
## 1 LGA           72.8
## 2 JFK           69.4
## 3 EWR           63.7
```

```
nycflights %>%
  group_by(origin) %>%
  summarise(dep_num = n()) %>%
  arrange(desc(dep_num))
```

```
## # A tibble: 3 x 2
##   origin dep_num
##   <chr>    <int>
## 1 EWR      11771
## 2 JFK      10897
## 3 LGA      10067
```

More practice:

Exercise 7:

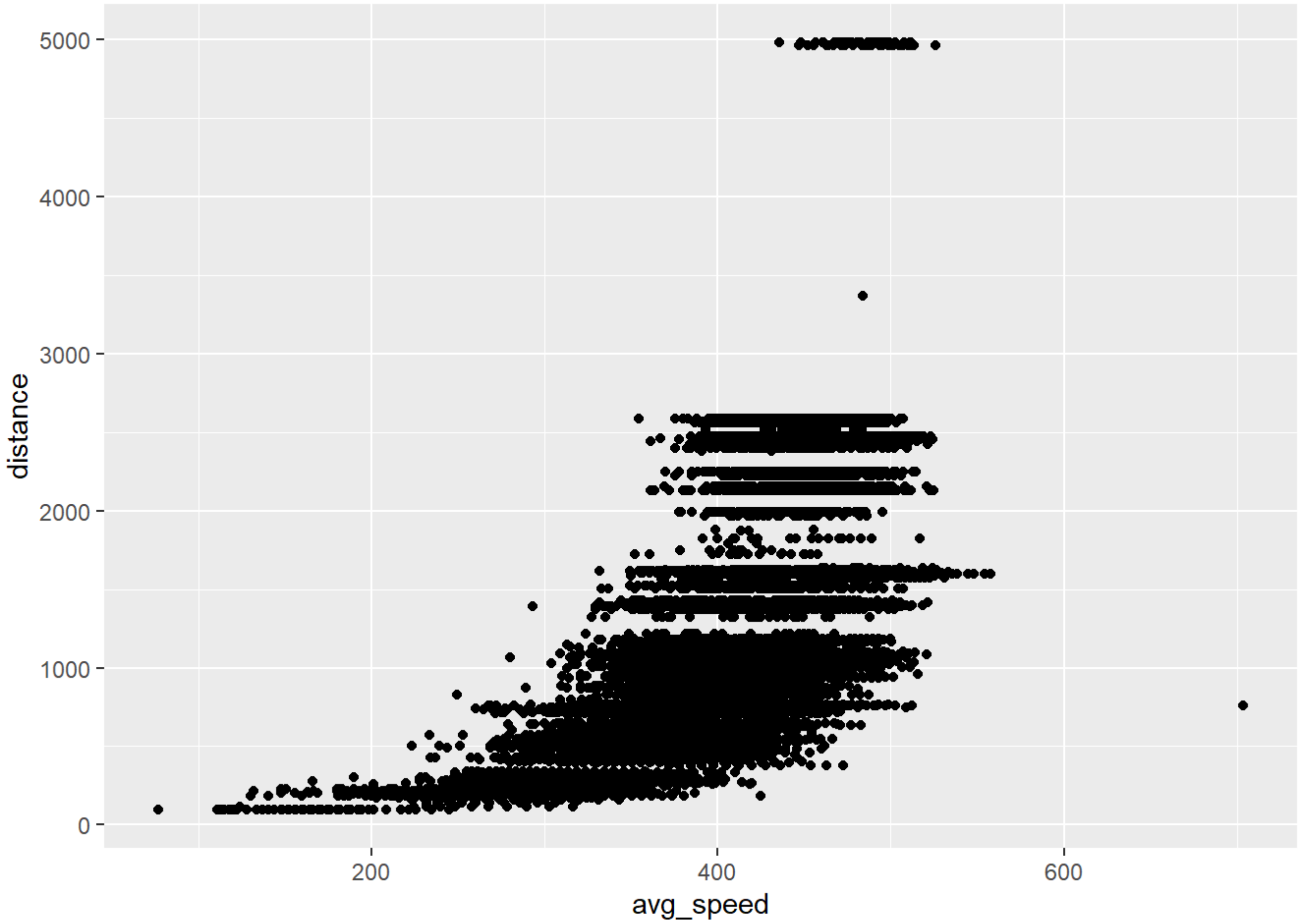
1 mile / hour = 0.0166666667 miles / minute

```
# enter your code for Exercise 7 here
nycflights <- nycflights %>%
  mutate(avg_speed = distance / (air_time/60))
```

Exercise 8:

There is a slightly positive relation between the distance and the average speed. The maximum average speed range is between 300-500, which for almost all of the distances above 1500 has been used. However, the variation of average speed for the distances below 1000 is very high, from 100 to 500.

```
# enter your code for Exercise 8 here
ggplot(nycflights, aes(x = avg_speed, y = distance)) +
  geom_point()
```

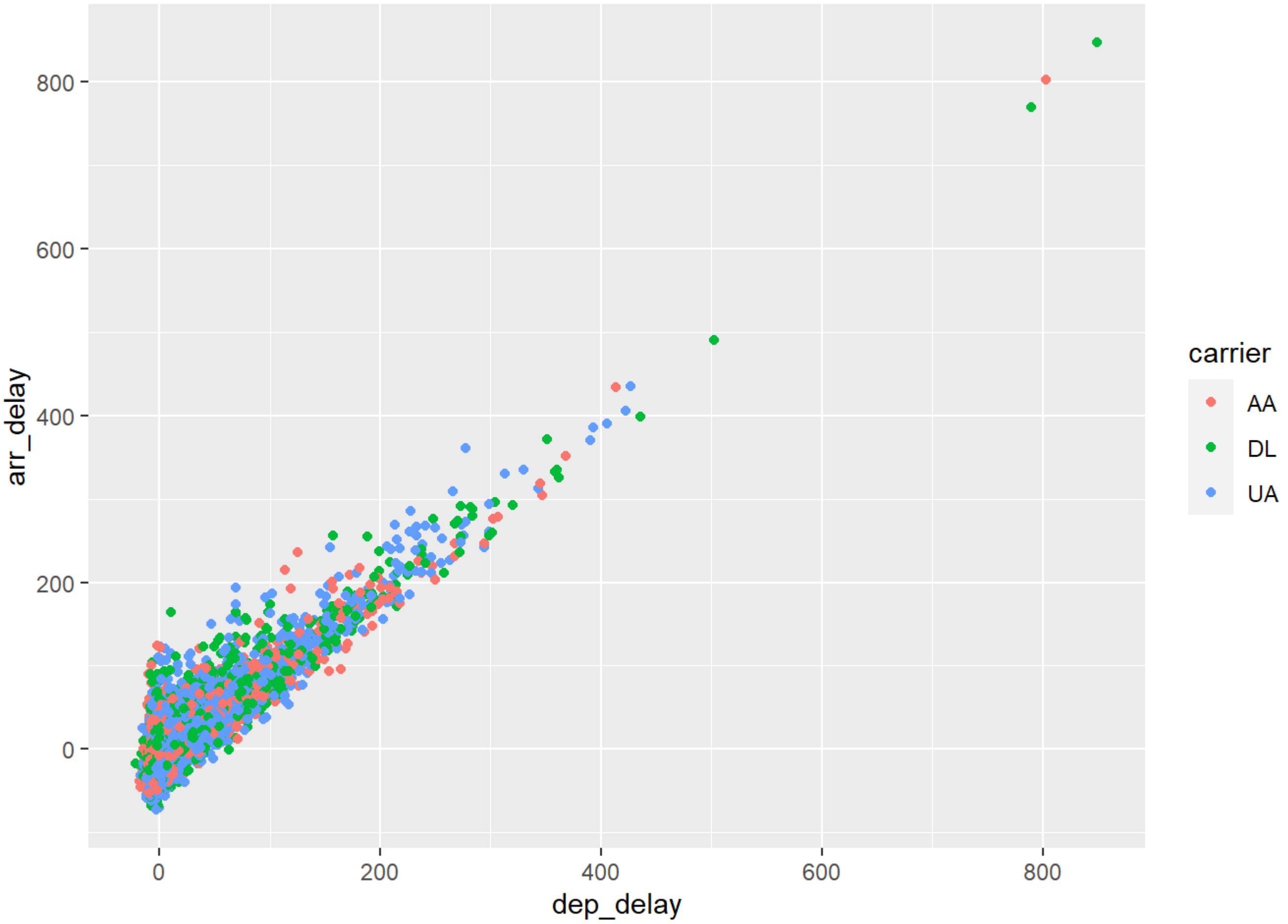


Exercise 9:

So the cutoff can be the time when the departure delay is about 80 (between 50 and 100). The reasoning is the importance of having the arrival delay as the zero. So considering an imaginary horizontal line on when the arr_delay is zero, we can conclude that the maximum departure delay can be about 80.

```
# enter your code for Exercise 9 here
new_flights <- nycflights %>%
  filter(carrier == "AA" | carrier == "DL" | carrier == "UA")

ggplot(new_flights, aes(x = dep_delay, y = arr_delay, color = carrier)) +
  geom_point()
```


800
600
400
200
0
arr_delay
0
200
400
600
800
dep_delay
carrier
AA
DL
UA