Lab 2: Introduction to Data

Sam Reeves

2021-02-10

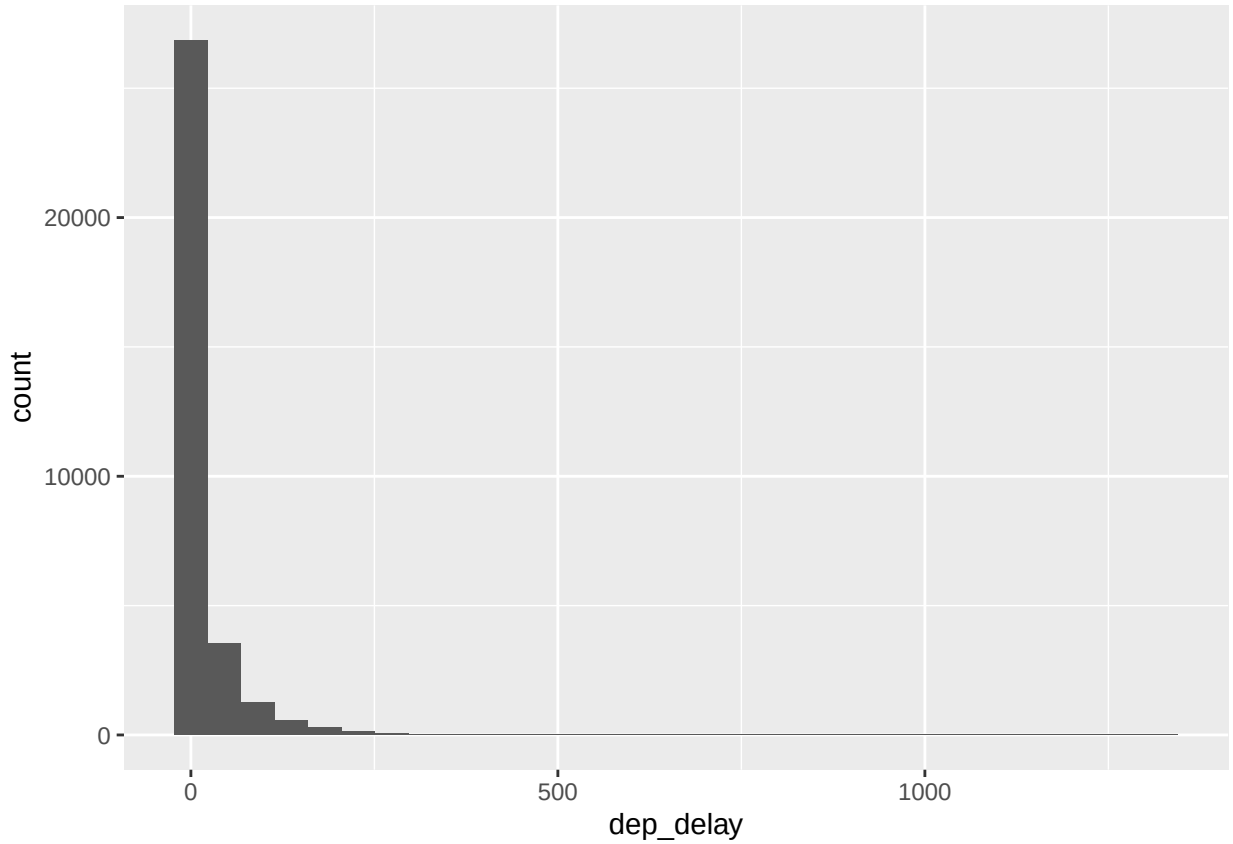
```
library(tidyverse)
library(openintro)
data(nycflights)
```

Exercise 1

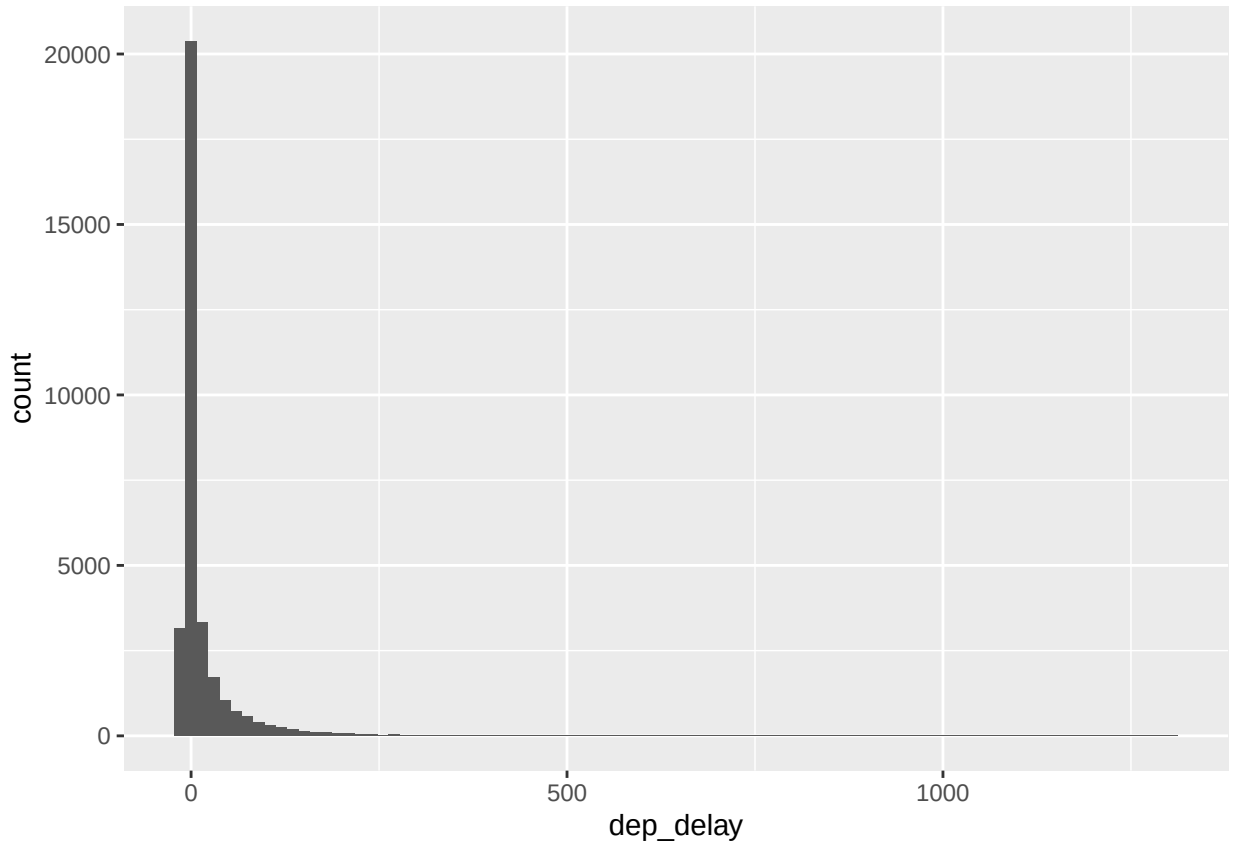
Look carefully at these three histograms. How do they compare? Are features revealed in one that are obscured in another?

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

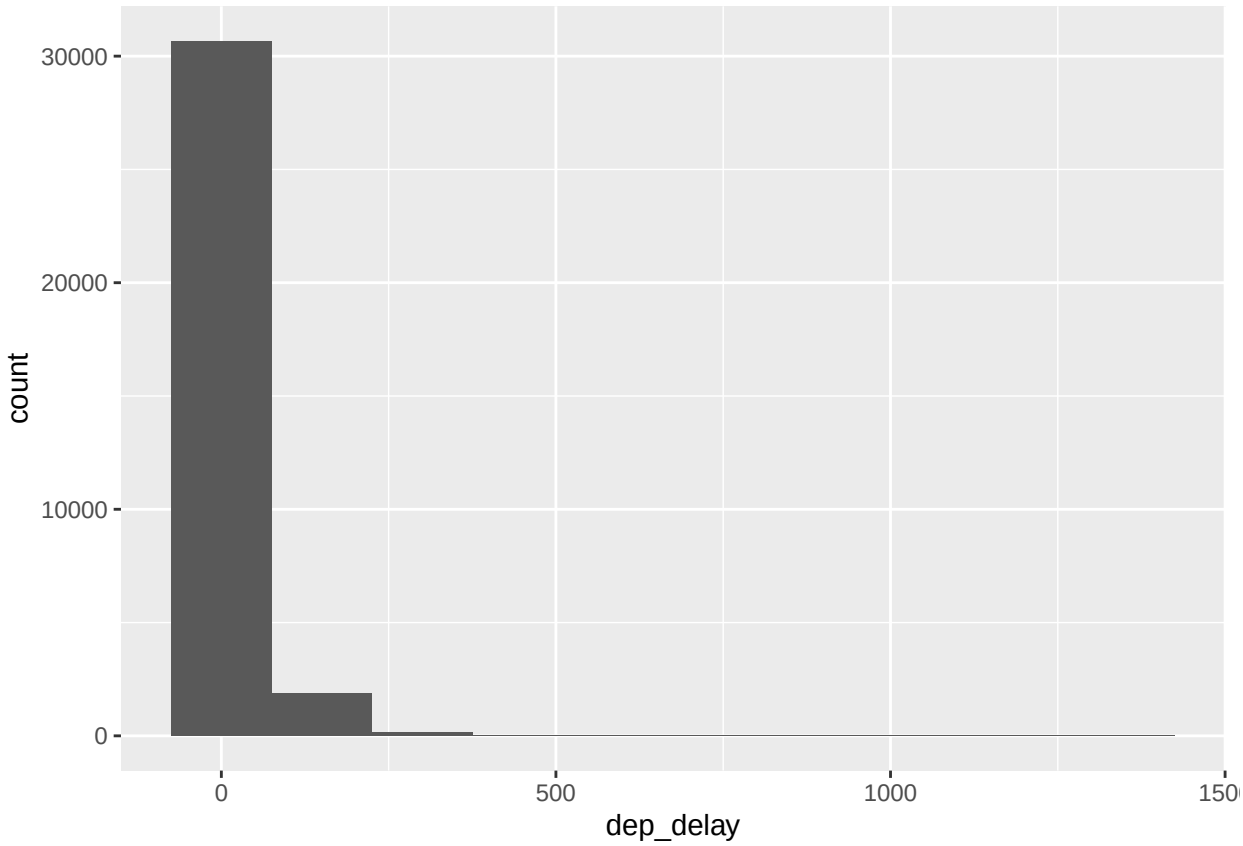
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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




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



These plots cut the same data into bins of different width. The larger bins are perhaps easier to read, but the smaller bins reveal with greater specificity that most flights left with a delay of 15 minutes or less. In the large binwidth graph, it's unclear that many flights leave a bit early.

Exercise 2

Create a new data frame that includes flights headed to SFO in February, and save this data frame as sfo_feb_flights. How many flights meet these criteria?

```
sfo_feb_flights <- nycflights %>%
  filter(month == 2,
         dest == 'SFO')
```

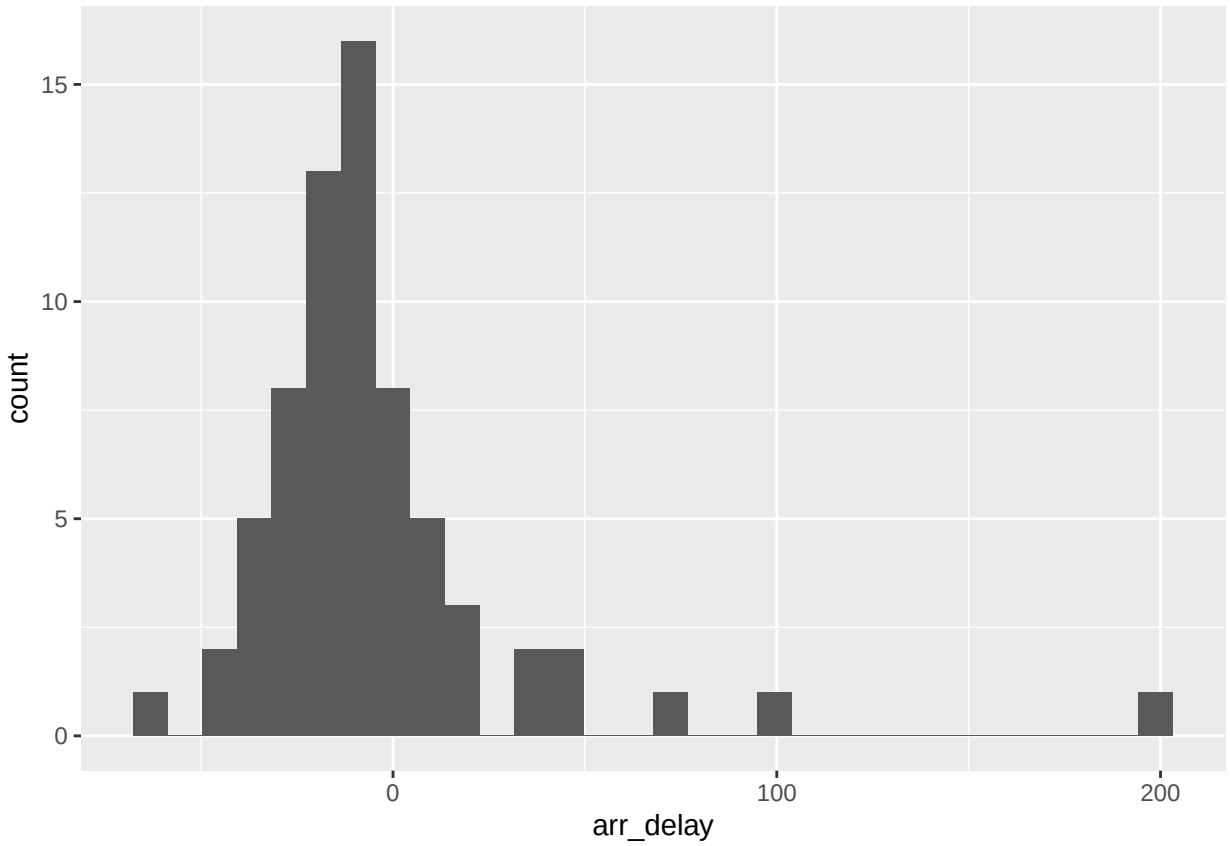
68 Flights meet the criteria.

Exercise 3

Describe the distribution of the arrival delays of these flights using a histogram and appropriate summary statistics. Hint: The summary statistics you use should depend on the shape of the distribution.

```
ggplot(sfo_feb_flights, aes(x = arr_delay)) + geom_histogram()
```

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




```
sfo_stats <- sfo_feb_flights %>%
  summarise(mean_arr_del = mean(arr_delay),
            med_arr_del = median(arr_delay),
            sd_arr_del = sd(arr_delay),
            var_arr_del = var(arr_delay),
            iqr_arr_del = IQR(arr_delay))
```

This group is distributed monomodally and skewed right. The majority of observations arrived early.

Exercise 4

Calculate the median and interquartile range for arr_delays of flights in in the sfo_feb_flights data frame, grouped by carrier. Which carrier has the most variable arrival delays?

```
sfo_feb_flights %>%
  group_by(carrier) %>%
  summarize(var_arr_delay = mean(var(arr_delay))) %>%
  arrange(desc(var_arr_delay))
```

```
## # A tibble: 5 x 2
##   carrier var_arr_delay
##   <chr>           <dbl>
## 1 UA              2335.
## 2 VX              1669.
## 3 AA               868.
## 4 DL               485.
## 5 B6               121.
```

United Airlines has the highest variance for its arrival times.

Exercise 5

Suppose you really dislike departure delays and you want to schedule your travel in a month that minimizes your potential departure delay leaving NYC. One option is to choose the month with the lowest mean departure delay. Another option is to choose the month with the lowest median departure delay. What are the pros and cons of these two choices?

The mean will give you an average amount of waiting time for the whole set, but the median can tell you more about how likely it is for a flight to be delayed for a given amount of time.

Exercies 6

If you were selecting an airport simply based on on time departure percentage, which NYC airport would you choose to fly out of?

```
nycflights %>%
  group_by(origin) %>%
  summarize(avg_departure = mean(dep_delay)) %>%
  arrange(avg_departure)
```

```
## # A tibble: 3 x 2
##   origin avg_departure
##   <chr>          <dbl>
## 1 LGA             10.1
## 2 JFK             12.3
## 3 EWR             15.3
```

I would select La Guardia!

Exercise 7

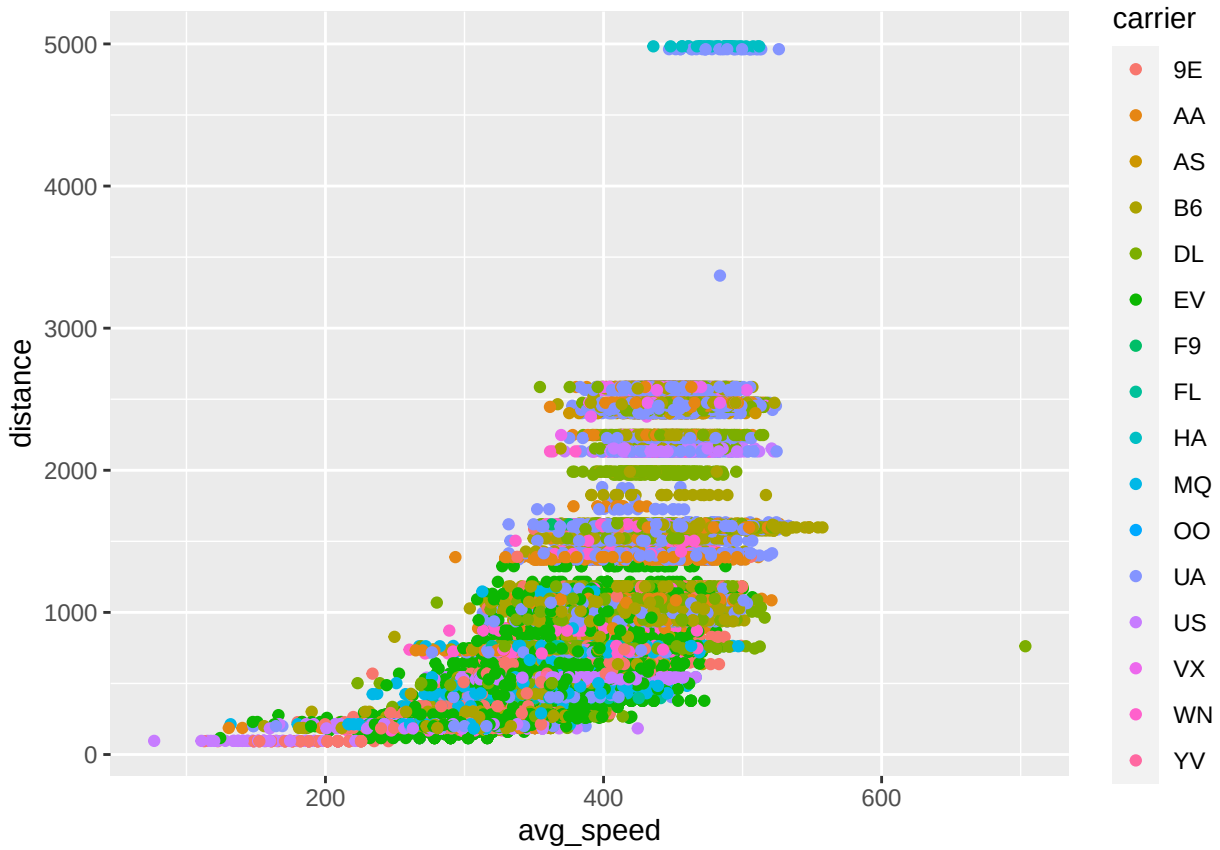
Mutate the data frame so that it includes a new variable that contains the average speed, avg_speed traveled by the plane for each flight (in mph). Hint: Average speed can be calculated as distance divided by number of hours of travel, and note that air_time is given in minutes.

```
nycflights <- nycflights %>%
  mutate(avg_speed = distance / air_time * 60)
```

Exercise 8

Make a scatterplot of avg_speed vs. distance. Describe the relationship between average speed and distance. Hint: Use geom_point().

```
nycflights %>% ggplot() +
  geom_point(aes(x = avg_speed, y = distance, color = carrier))
```



The relationship is non-linear and skewed left.

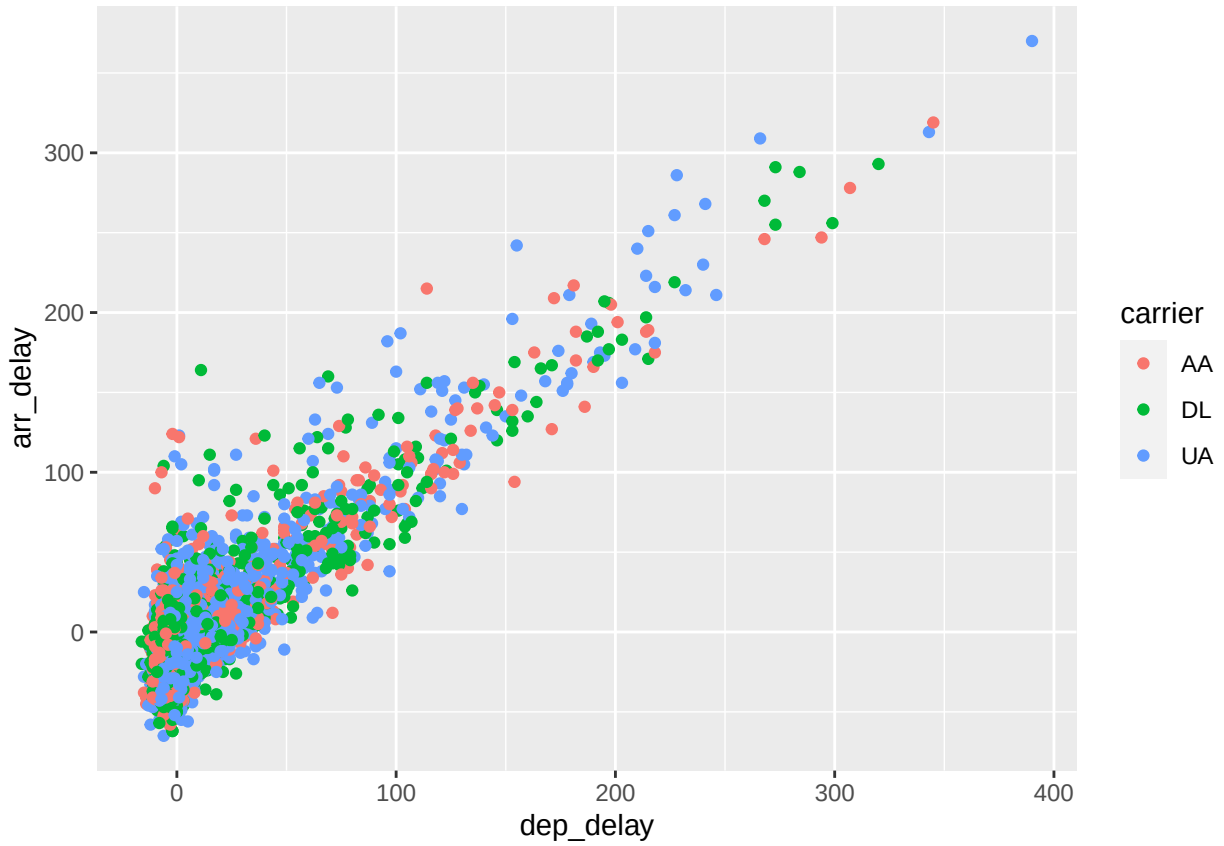
Exercise 9

Replicate the following plot. Hint: The data frame plotted only contains flights from American Airlines, Delta Airlines, and United Airlines, and the points are colored by carrier. Once you replicate the plot, determine (roughly) what the cutoff point is for departure delays where you can still expect to get to your destination on time.

```
special_group <- nycflights %>%
  filter(carrier == c('AA', 'DL', 'UA'))
```

```
## Warning in carrier == c("AA", "DL", "UA"): longer object length is not a
## multiple of shorter object length
```

```
special_group %>%
  ggplot() +
  geom_point(aes(x = dep_delay, y = arr_delay, color = carrier))
```




```
special_group %>%
  filter(arr_delay <= 0) %>%
  group_by(carrier) %>%
  summarize(cutoff = mean(dep_delay)) %>%
  arrange(desc(cutoff))
```

```
## # A tibble: 3 x 2
##   carrier cutoff
##   <chr>    <dbl>
## 1 UA      -0.274
## 2 DL      -2.17
## 3 AA      -2.88
```

It appears that these three companies only tend to arrive on time if they leave early.