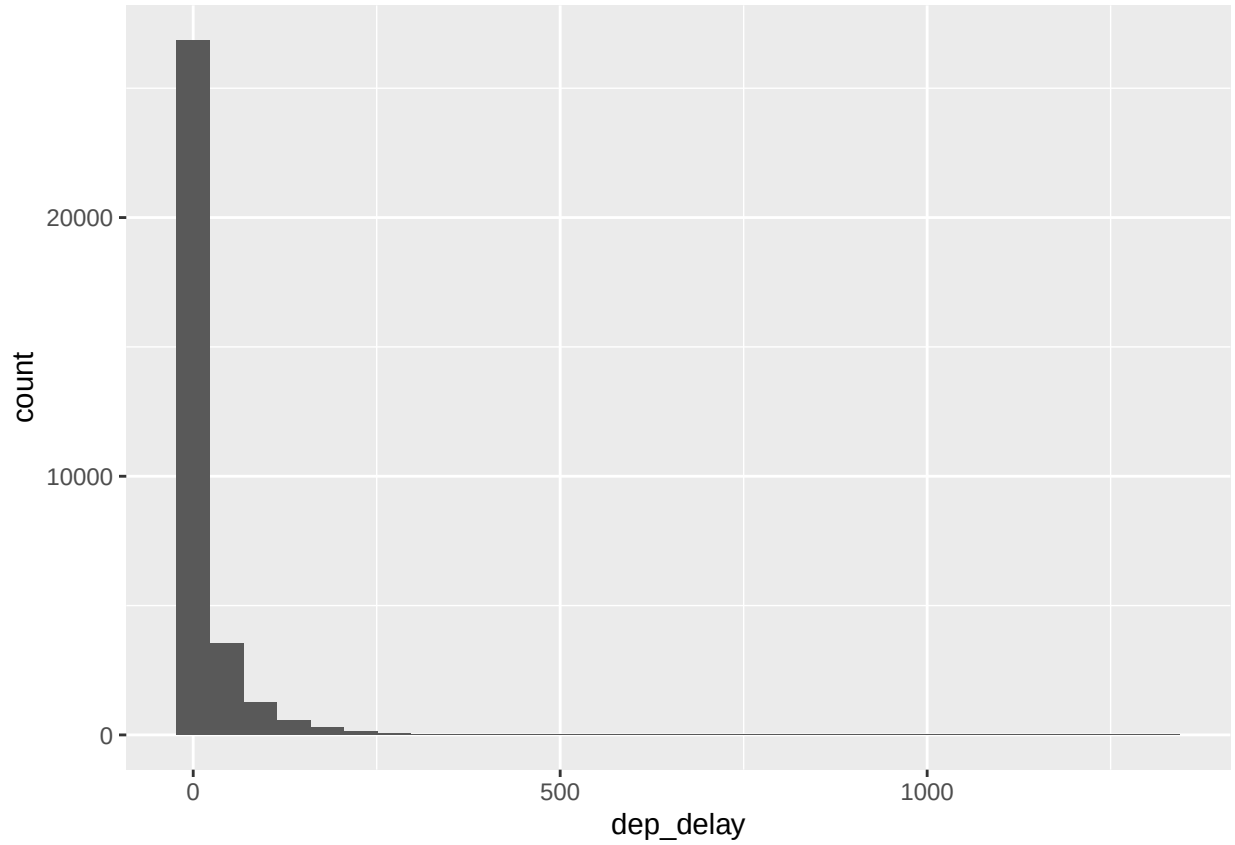
Lab 2

Alice Ding

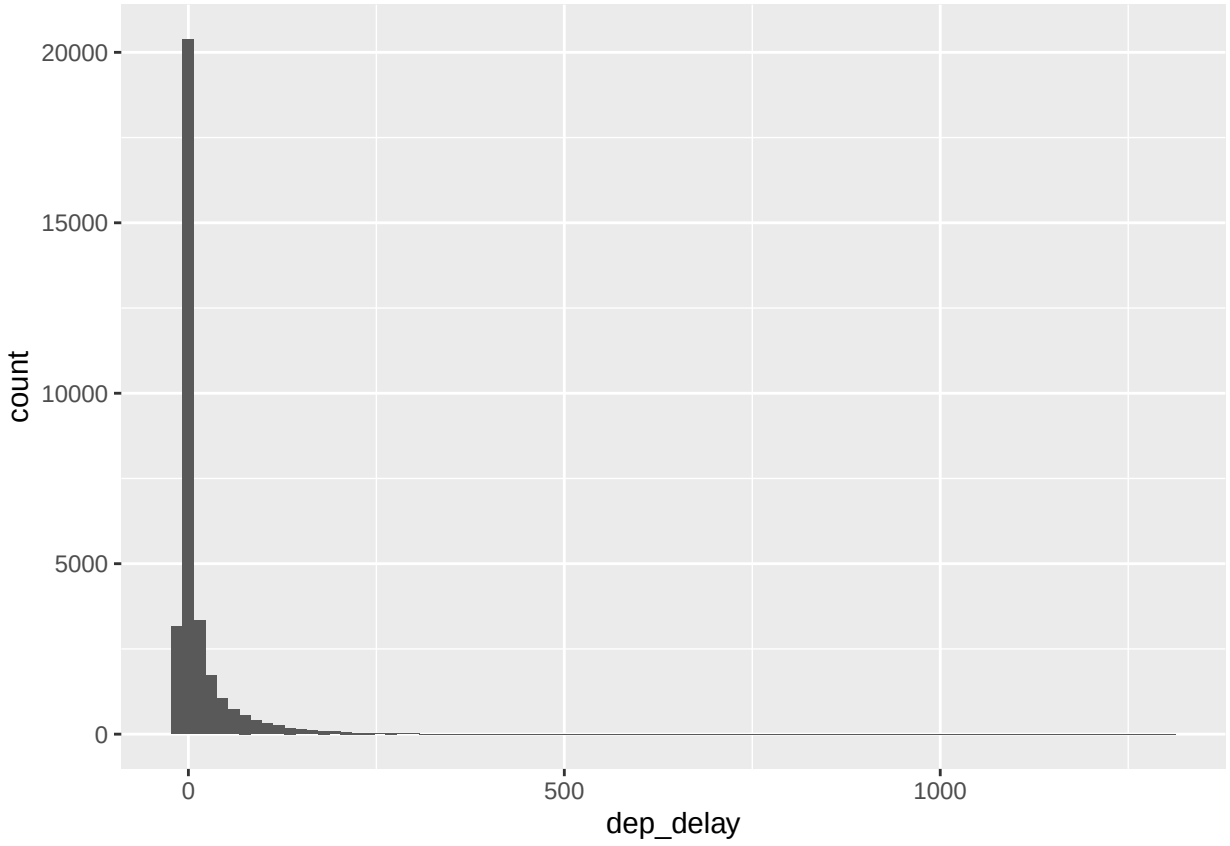
Exercise 1

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

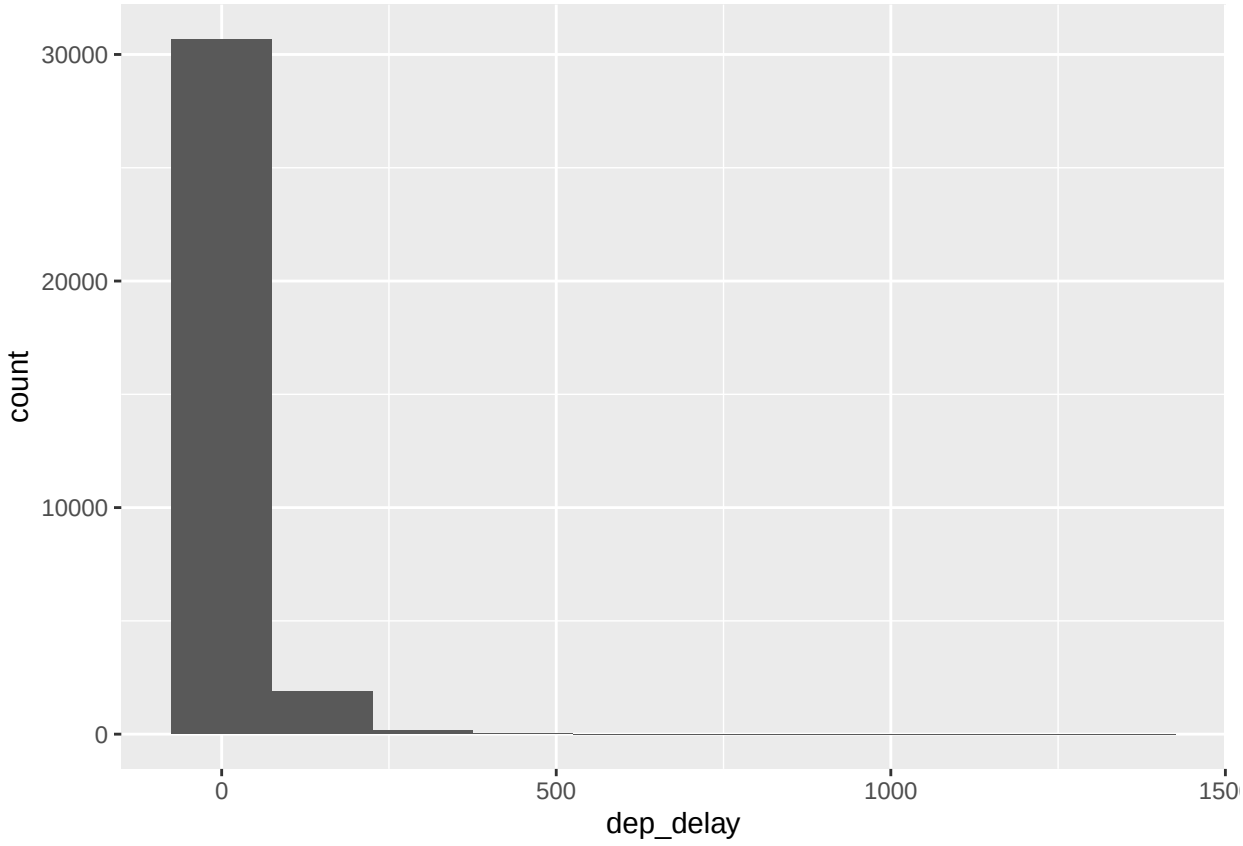
```
ggplot(data = nycflights, aes(x = dep_delay)) +
  geom_histogram()
```




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




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



The first histogram shows the fast majority of flights being in the first bin (almost 30k flights) with the second bin having less than 5k bins and then trailing downwards. The max minutes late flights seem to be based on this chart can be almost 400 minutes. The first bin looks like it contains all flights that land early as well as ones that are a little delayed.

The second histogram is more granular and highlights where a majority of flights fall which is actually at 0, meaning flights are usually on time and this holds a little over 20k flights. The second largest bin is around 3k and holds flights that are a little late while the third largest bin that is just slightly less than the second actually has flights that are less than 0 minutes late, meaning they're early. The trend as seen in the first histogram is then mirrored, but at a more granular scale as it tapers off a at around 300 minutes late.

The third histogram with a bin size of 150 has even less granularity and only shows 4 bins, the largest at above 30k flights that encompasses early to late (probably an hour or more), then the rest of the bins taper off in size as we get later and later. The latest bin goes past 500 minutes.

When comparing the three of them, the second histogram reveals a trend that is not visible in the first or third as it shows a vast majority of flights are on time. The third histogram also seems to imply that the max minutes late could be past 500, but in the other two histograms, they seem to stop at around 300 in the second histogram and 400 for the first.

Exercise 2

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(dest == "SFO", month == 2)
glimpse(sfo_feb_flights)
```

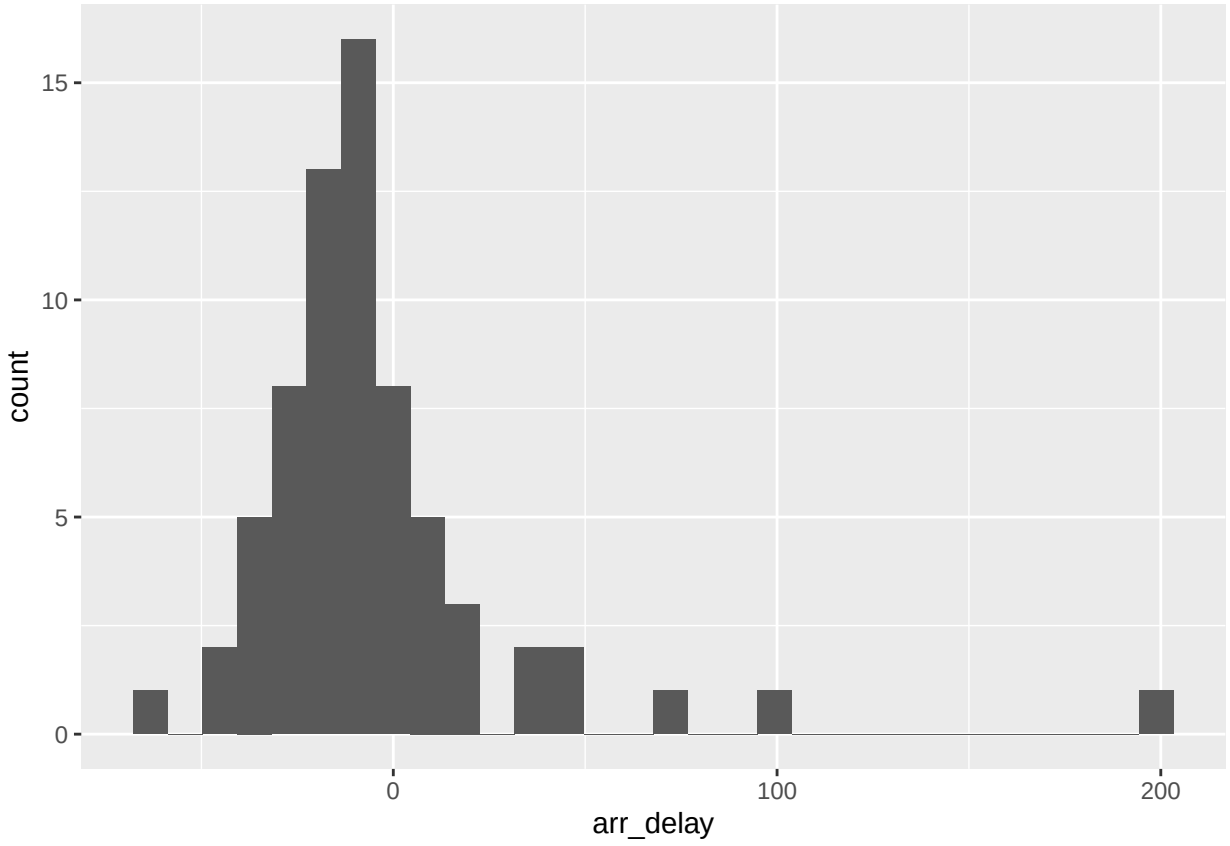
```
## Rows: 68
## Columns: 16
## $ year      <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, ~
## $ month     <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ day       <int> 18, 3, 15, 18, 24, 25, 7, 15, 13, 8, 11, 13, 25, 20, 12, 27,~
## $ dep_time  <int> 1527, 613, 955, 1928, 1340, 1415, 1032, 1805, 1056, 656, 191~
## $ dep_delay <dbl> 57, 14, -5, 15, 2, -10, 1, 20, -4, -4, 40, -2, -1, -6, -7, 2~
## $ arr_time  <int> 1903, 1008, 1313, 2239, 1644, 1737, 1352, 2122, 1412, 1039, ~
## $ arr_delay <dbl> 48, 38, -28, -6, -21, -13, -10, 2, -13, -6, 2, -5, -30, -22,~
## $ carrier   <chr> "DL", "UA", "DL", "UA", "UA", "UA", "B6", "AA", "UA", "DL", ~
## $ tailnum   <chr> "N711ZX", "N502UA", "N717TW", "N24212", "N76269", "N532UA", ~
## $ flight    <int> 1322, 691, 1765, 1214, 1111, 394, 641, 177, 642, 1865, 272, ~
## $ origin    <chr> "JFK", "JFK", "JFK", "EWR", "EWR", "JFK", "JFK", "JFK", "JFK~
## $ dest      <chr> "SFO", "SFO", "SFO", "SFO", "SFO", "SFO", "SFO", "SFO", "SFO~
## $ air_time  <dbl> 358, 367, 338, 353, 341, 355, 359, 338, 347, 361, 332, 351, ~
## $ distance  <dbl> 2586, 2586, 2586, 2565, 2565, 2586, 2586, 2586, 2586, 2586, ~
## $ hour      <dbl> 15, 6, 9, 19, 13, 14, 10, 18, 10, 6, 19, 8, 10, 18, 7, 17, 1~
## $ minute    <dbl> 27, 13, 55, 28, 40, 15, 32, 5, 56, 56, 10, 33, 48, 49, 23, 2~
```

Based on the glimpse(), it looks like there are 68 flights in this subset of data.

Exercise 3

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



There is a bell curve for this chart that tops at 0 which makes sense as most flights seem to be on time in terms of leaving, however this is an outlier at 200 minutes which skews the data. This makes data more prone to outliers less valuable, such as standard deviation or mean. It also looks like more flights seem to leave a little earlier as the second largest bin is to the left of 0.

```
sfo_feb_flights %>%
  summarise(mean_an   = mean(arr_delay),
            median_ad = median(arr_delay),
            iqr_ad    = IQR(arr_delay),
            min_ad    = min(arr_delay),
            max_ad    = max(arr_delay),
            n         = n())
```

```
## # A tibble: 1 x 6
##   mean_an median_ad iqr_ad min_ad max_ad     n
##     <dbl>     <dbl>  <dbl>  <dbl>  <dbl> <int>
## 1    -4.5       -11   23.2    -66    196    68
```

The histogram implies that more flights seem to leave early and this is validated by the median being at -11. We see that the mean was more impacted here by that outlier at ~200 (revealed to be 196) as it's -4.5 vs. -11.

Exercise 4

1. 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) %>%
  summarise(median_dd = median(arr_delay), iqr_dd = IQR(arr_delay), n_flights = n())
```

```
## # A tibble: 5 x 4
##   carrier median_dd iqr_dd 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
```

DL and UA both have an IQR of 22, meaning their spread sees the widest variability. DL has less flights which would imply that out of the two, it has the most variable arrival delays.

Exercise 5

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?

```
nycflights %>%
  group_by(month) %>%
  summarise(mean_dd = mean(dep_delay)
            , med_dd = median(dep_delay)
            , iqr_dd = IQR(dep_delay)
            , sd_dd = sd(dep_delay)
            , max_dd = max(dep_delay)
            , min_dd = min(dep_delay)) %>%
  arrange(desc(mean_dd))
```

```
## # A tibble: 12 x 7
##    month mean_dd med_dd iqr_dd sd_dd max_dd min_dd
##    <int>   <dbl>  <dbl>  <dbl> <dbl>  <dbl>  <dbl>
##  1     7   20.8       0     26  47.8    392    -17
##  2     6   20.4       0     25  53.5    803    -18
##  3    12   17.4       1     25  43.0    849    -18
##  4     4   14.6      -2     16  43.4    427    -18
##  5     3   13.5      -1     17  40.3    393    -20
##  6     5   13.3      -1     19  38.3    351    -19
##  7     8   12.6      -1     15  39.2    436    -21
##  8     2   10.7      -2     15  33.1    319    -20
##  9     1   10.2      -2     12  42.4   1301    -17
## 10     9    6.87     -3      8  35.3    473    -21
## 11    11    6.10     -2     10  27.6    413    -21
## 12    10    5.88     -3      9  29.4    272    -18
```

Mean is more influenced by outliers thus it would probably be less reliable if the data was super noisy. A pro though is that it would be the average you'd have to wait given a day for that particular month. For the median, it is not as influenced by outliers and thus a little more reliable. It however is not a good representative for how the data is distributed.

For this dataset, I'd probably go with the median out of the two if I had to choose one. The median though only ranges from -3 to 0 so it's not super insightful - to utilize more data, bringing in IQR and standard deviation could help as they show the spread of the data. Based on these two numbers, I would say October is the best month as it has the lowest median (tied with September) and its IQR is second lowest while its standard deviation is also second lowest.

Exercise 6

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

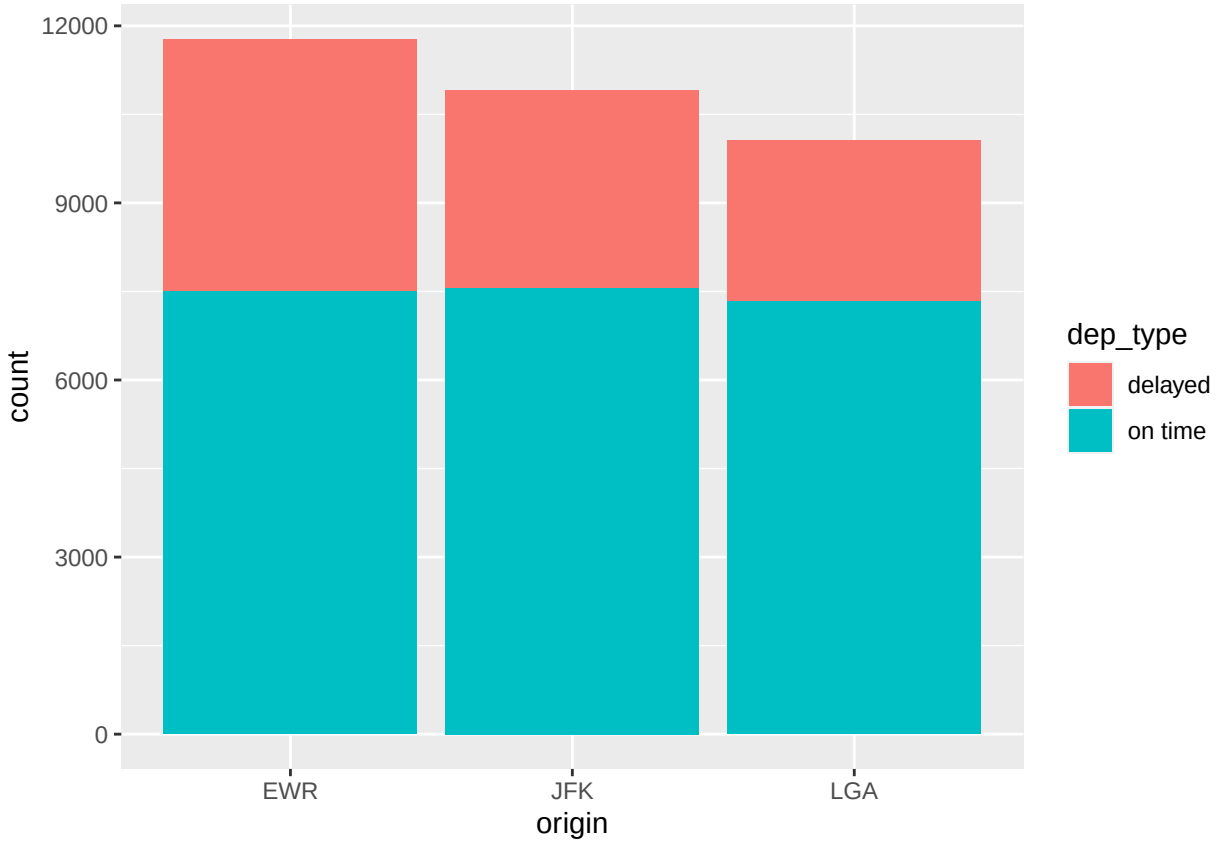
```
nycflights <- nycflights %>%
  mutate(dep_type = ifelse(dep_delay < 5, "on time", "delayed"))

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

```
## # A tibble: 3 x 2
##   origin ot_dep_rate
##   <chr>        <dbl>
## 1 LGA          0.728
## 2 JFK          0.694
## 3 EWR          0.637
```

You can also visualize the distribution of on on time departure rate across the three airports using a segmented bar plot.

```
ggplot(data = nycflights, aes(x = origin, fill = dep_type)) +
  geom_bar()
```



LGA has the highest on time departure rate at ~73%.

Exercise 7

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)

select(nycflights, distance, air_time, avg_speed) %>% head(10)
```

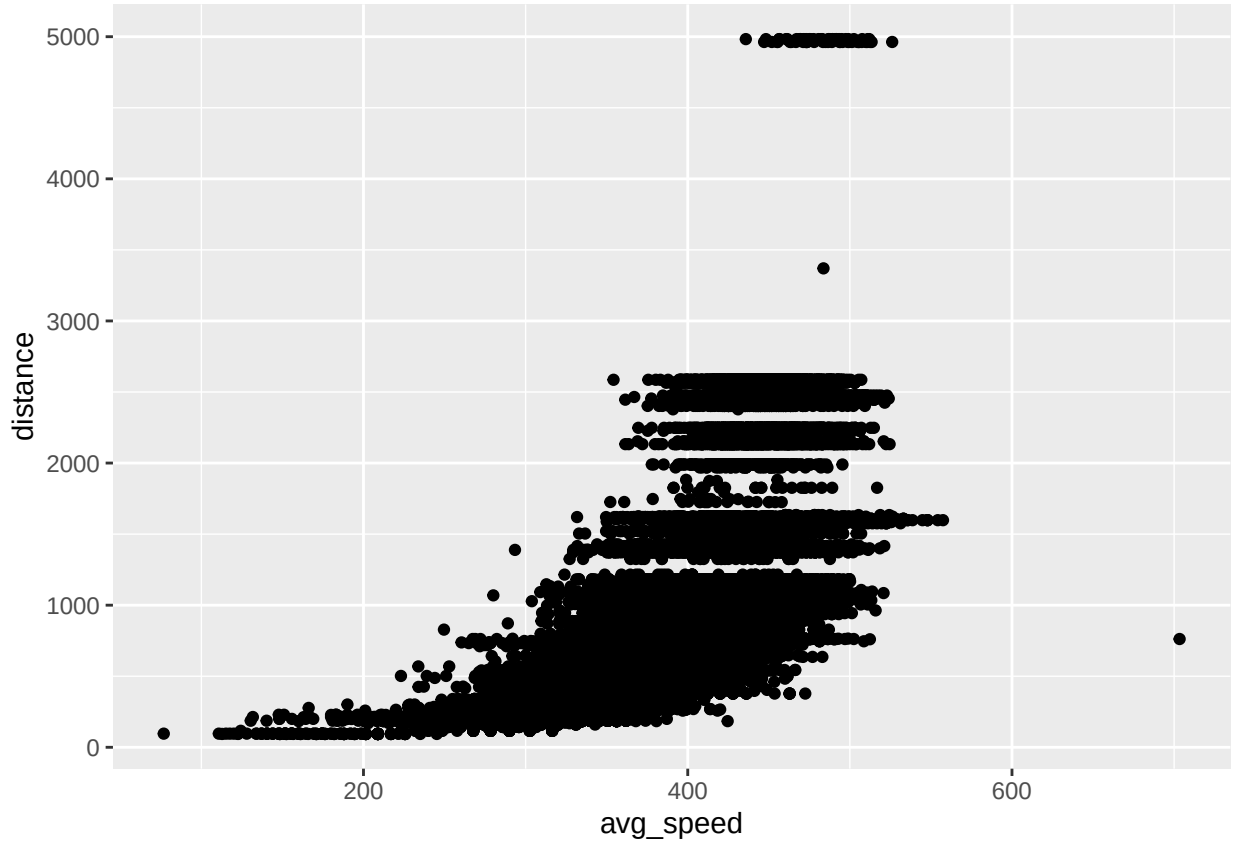
```
## # A tibble: 10 x 3
##    distance air_time avg_speed
##       <dbl>    <dbl>     <dbl>
##  1     2475      313      474.
##  2     1598      216      444.
##  3     2475      376      395.
##  4     1005      135      447.
##  5      296       50      355.
##  6      733      138      319.
##  7     1411      240      353.
##  8      228       48      285
```

```
##  9     1096      148      444.
## 10      820      110      447.
```

Exercise 8

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

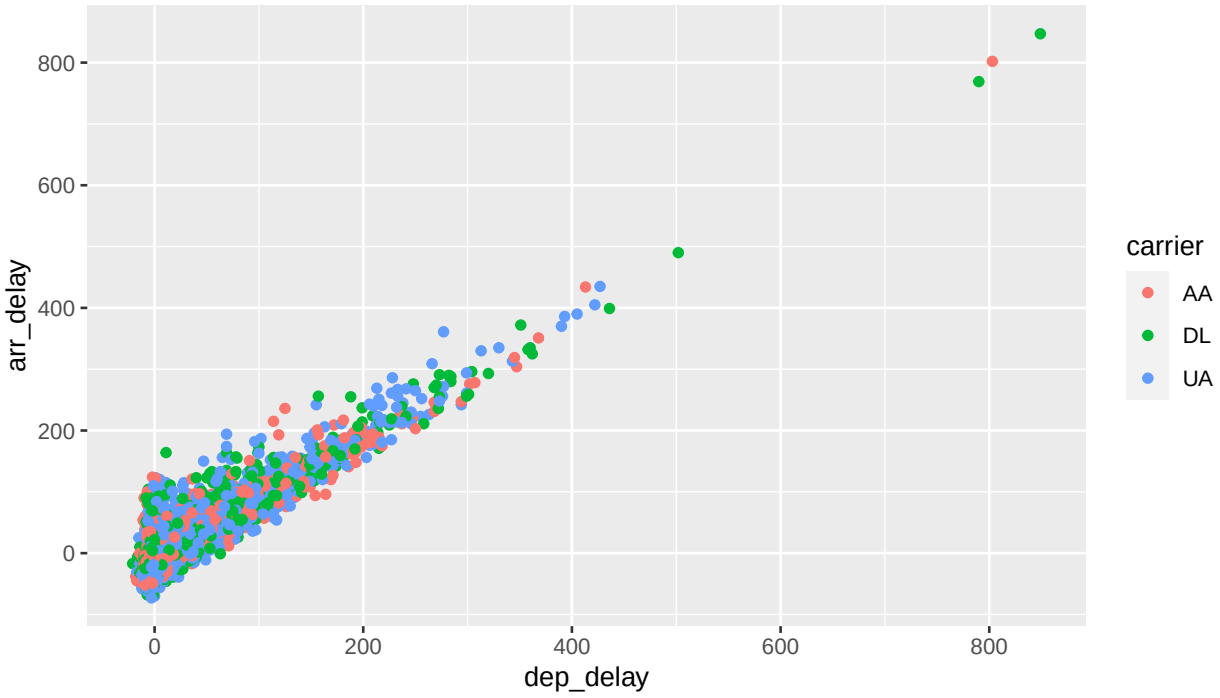
```
ggplot(data = nycflights, aes(x = avg_speed, y = distance)) +
  geom_point()
```



There does seem to be a positive relationship between the two (the larger the distance, the faster the speed), however it seemes to plateau after a certain distance which seems to suggest planes can only go so fast. This also would mean that in general, shorter distance flights will probably be a little slower - this makes sense as shorter flights usually have smaller planes, thus less power. There also seems to be a bit of variability between same distances and average speed which makes sense as well due to weather conditions and other variables that affect travel.

Exercise 9

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.



It looks like that the latest you can depart would be at maybe 60 minutes - it looks like that 2/3 to 100, you can still have an `arr_delay` of 0.