Lab 2 - Intro to Data

Peter Gatica

2021-02-17

```
options(tidyverse.quiet = TRUE)
library(tidyverse)
library(openintro)
```

```
names(nycflights)
```

```
##  [1] "year"      "month"     "day"       "dep_time"  "dep_delay" "arr_time"
##  [7] "arr_delay" "carrier"   "tailnum"   "flight"    "origin"    "dest"
## [13] "air_time"  "distance"  "hour"      "minute"
```

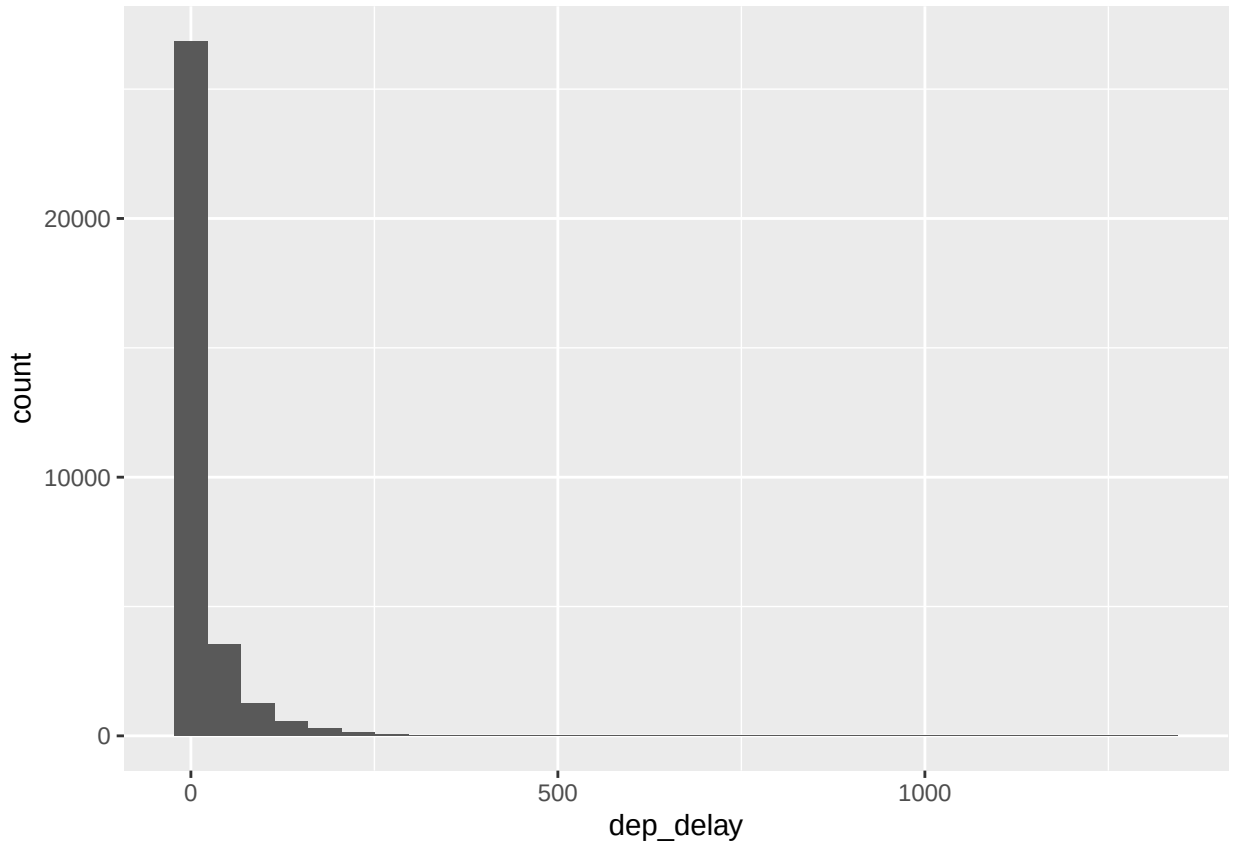
Exercise 1

Import source data nycflights. Examine the distribution of departure delays of all flights with a histogram.

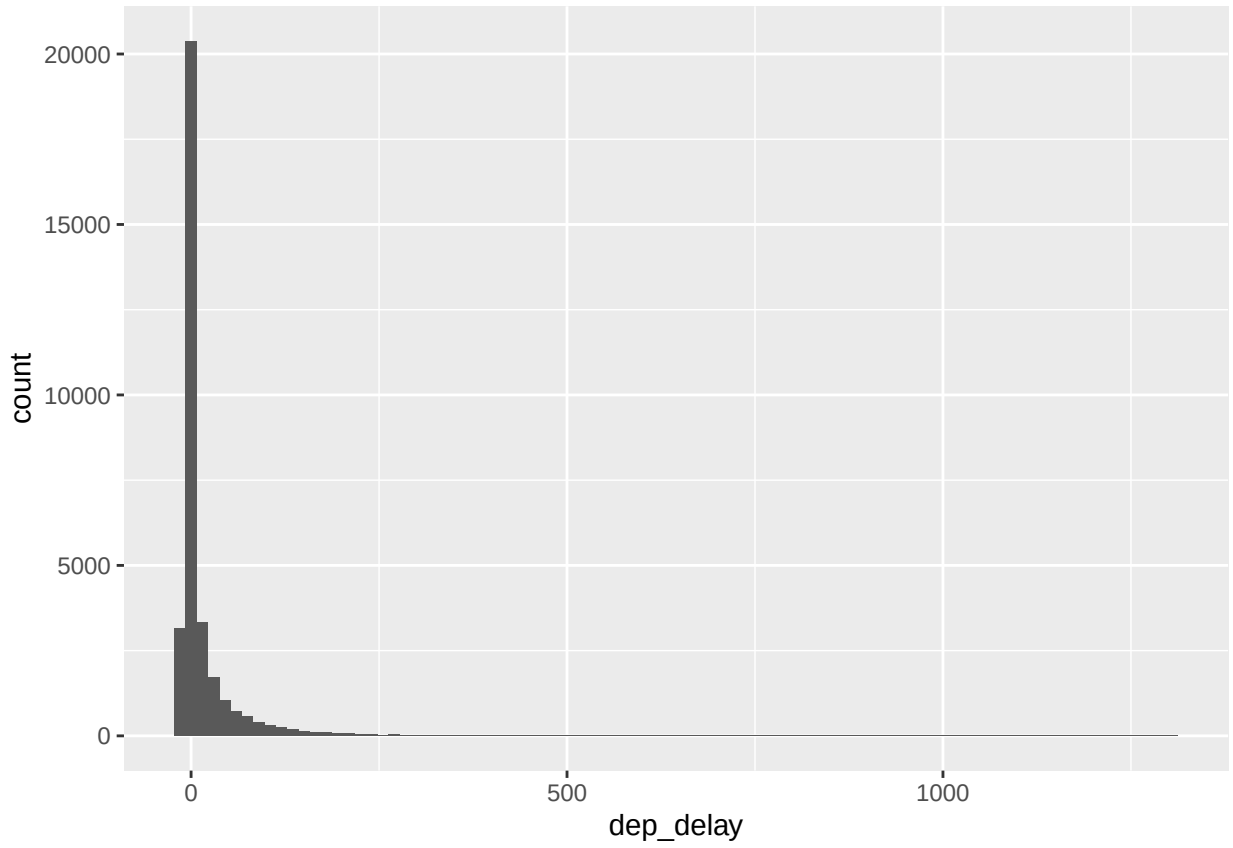
```
data("nycflights")

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

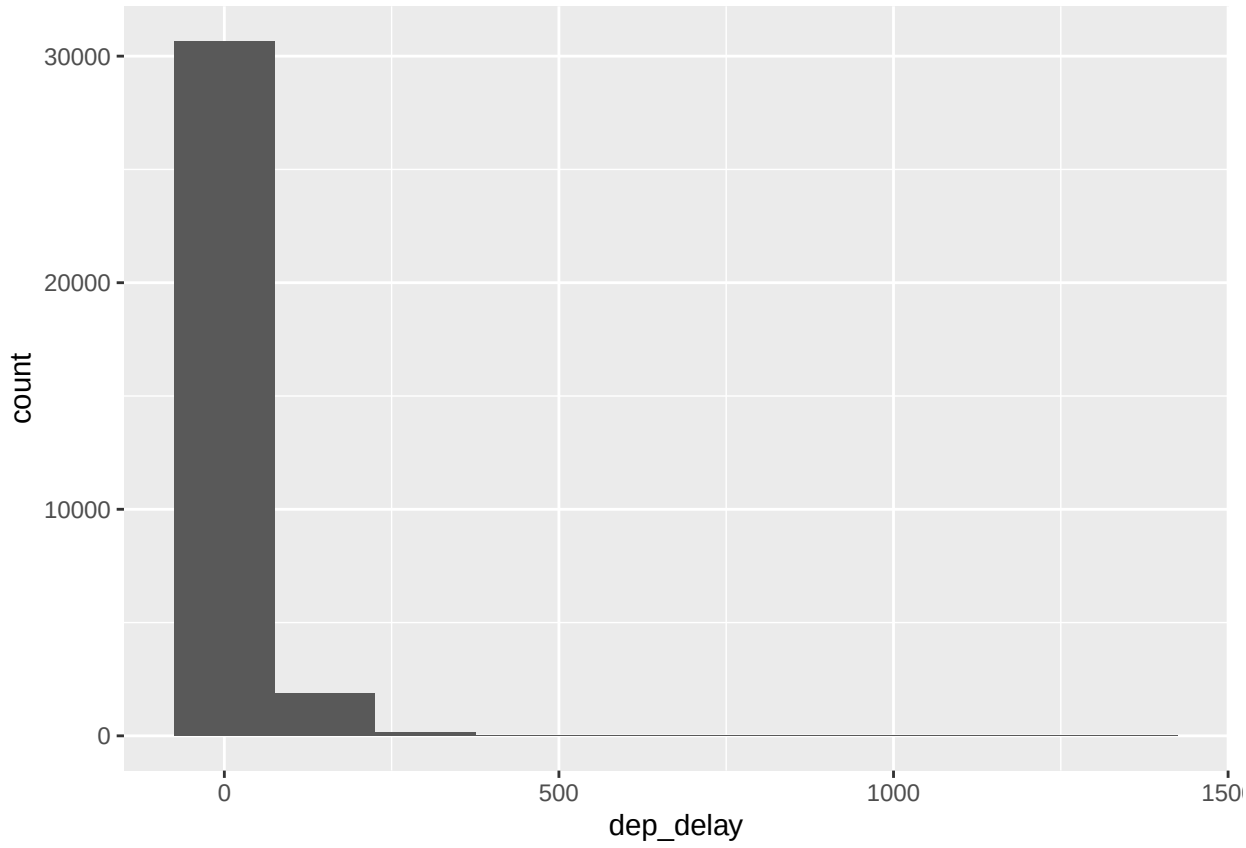
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



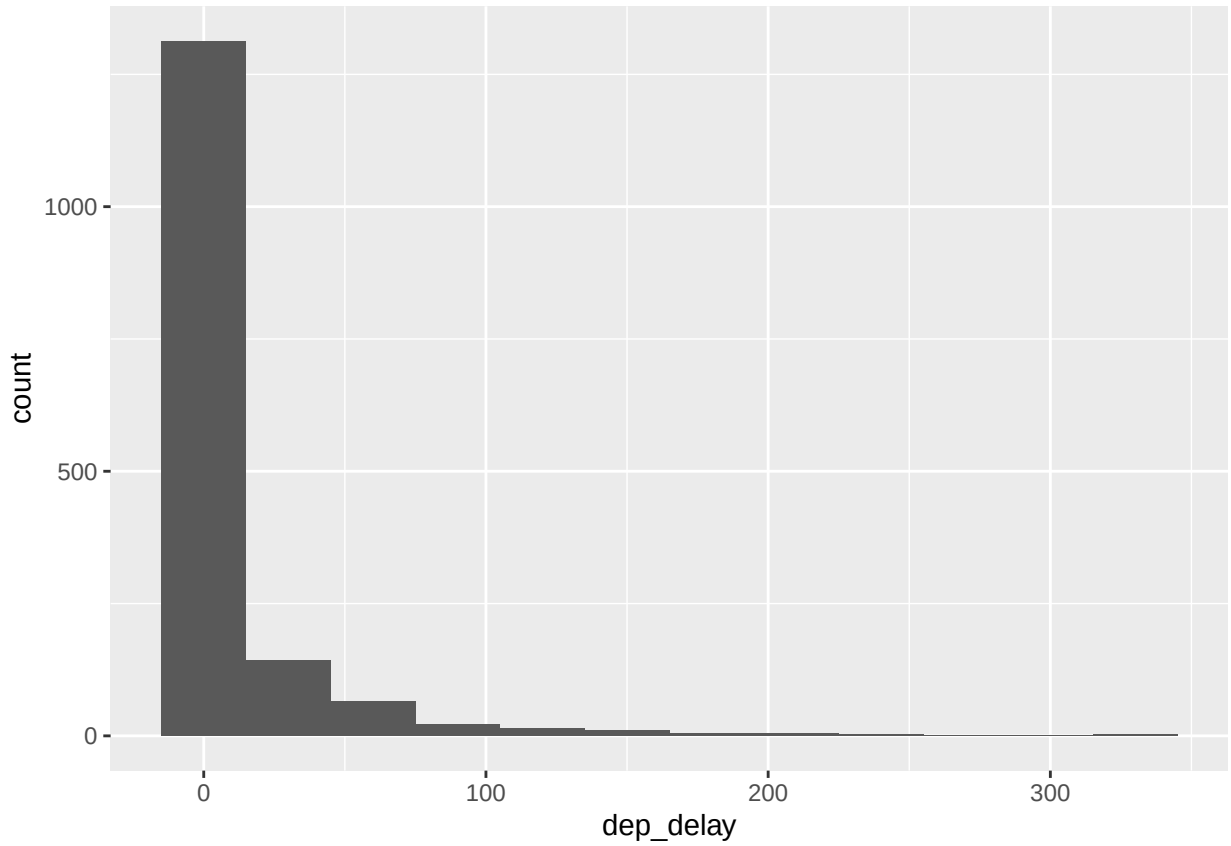

```
ggplot(data = nycflights, aes(x = dep_delay)) +
  geom_histogram(binwidth = 15)
```




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




```
lax_flights <- nycflights %>%
  filter(dest == "LAX")
ggplot(data = lax_flights, aes(x = dep_delay)) +
  geom_histogram(binwidth = 30)
```




```
lax_flights %>%
  summarise(mean_dd   = mean(dep_delay),
            median_dd = median(dep_delay),
            n         = n())
```

```
## # A tibble: 1 x 3
##   mean_dd median_dd     n
##     <dbl>     <dbl> <int>
## 1    9.78        -1  1583
```

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?

Answer: There were 68 flights that arrive into the San Franscisco Airport in February.

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

sfo_feb_flights %>%
  summarise( n = n())
```

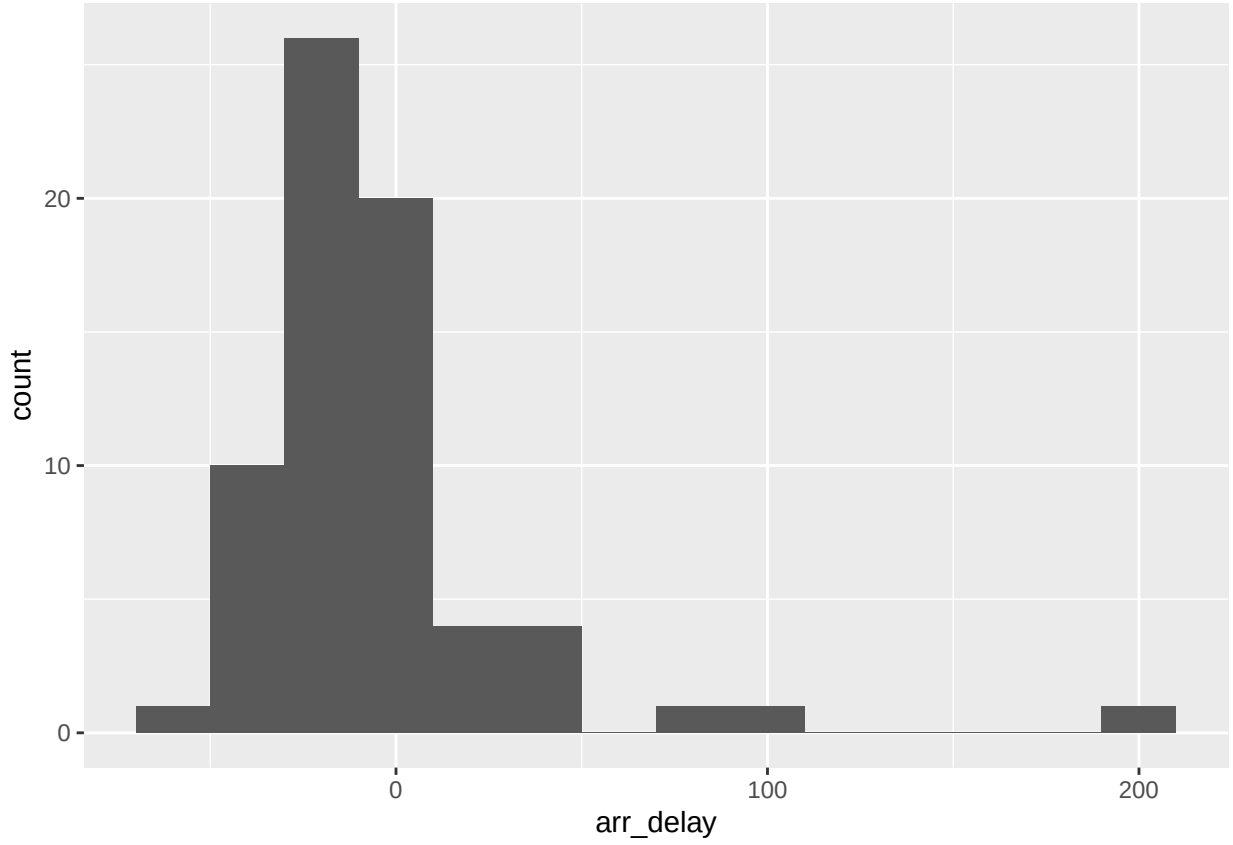
```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    68
```

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.

Answer: The majorty of flights that arrived into the San Francisco Airport in February 2013 arrived early. Approximately 58 of the 68 arrive early or on time. The histogram is right skewed with apparent outliers that were really late in arriving. The distribution also shows that the majority of flights in the dataset were early and on time with a few short delays.

```
ggplot(data = sfo_feb_flights, aes(x = arr_delay)) +
  geom_histogram(binwidth = 20)
```




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

```
## # A tibble: 1 x 5
##   mean_ad median_ad min_ad max_ad     n
##     <dbl>     <dbl>  <dbl>  <dbl> <int>
## 1    -4.5       -11    -66    196    68
```

Exercise 4

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

Answer: While the IQR for United (UA) and Delta (DL) are equal at 22 and almost the same amount of flights, the median arrival delay is lower for Delta. This means that within the IQR of 22 the range of arrive delays is from -21 minutes (early arrivals) to 1 minute delays which tells me that Delta is typically early to SFO while American Airlines (AA) seems to be typically late. United is typically early or on time and occasionally late.

```
sfo_feb_flights %>%
  group_by(carrier) %>%
  summarise(median_ad = median(arr_delay), iqr_ad = IQR(arr_delay), n_flights = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # 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
```

```
nycflights %>%
  group_by(month) %>%
  summarise(mean_dd = mean(dep_delay), median_dd = median(dep_delay)) %>%
  arrange(desc(mean_dd))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 12 x 3
##    month mean_dd median_dd
##    <int>   <dbl>     <dbl>
##  1     7   20.8          0
##  2     6   20.4          0
##  3    12   17.4          1
##  4     4   14.6         -2
##  5     3   13.5         -1
##  6     5   13.3         -1
##  7     8   12.6         -1
##  8     2   10.7         -2
##  9     1   10.2         -2
## 10     9    6.87        -3
## 11    11    6.10        -2
## 12    10    5.88        -3
```

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?

Answer: Looking at the above data, it appears that October would be the best month to take a flight out of NYC if a flyer wants to avoid delays. October's median is the lowest average delay time and the lowest median delay time. A flyer should also look at the median because if it is low and the mean is high then that may indicate that there may be one or more outliers which would then increase the average delay time sometimes significantly. ...

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

head(nycflights, 10)
```

```
## # A tibble: 10 x 17
##     year month   day dep_time dep_delay arr_time arr_delay carrier tailnum
##    <int> <int> <int>    <int>     <dbl>    <int>     <dbl> <chr>   <chr>
##  1  2013     6    30      940        15     1216        -4 VX      N626VA
##  2  2013     5     7     1657        -3     2104        10 DL      N3760C
##  3  2013    12     8      859        -1     1238        11 DL      N712TW
##  4  2013     5    14     1841        -4     2122       -34 DL      N914DL
##  5  2013     7    21     1102        -3     1230        -8 9E      N823AY
##  6  2013     1     1     1817        -3     2008         3 AA      N3AXAA
##  7  2013    12     9     1259        14     1617        22 WN      N218WN
##  8  2013     8    13     1920        85     2032        71 B6      N284JB
##  9  2013     9    26      725       -10     1027        -8 AA      N3FSAA
## 10  2013     4    30     1323        62     1549        60 EV      N12163
## # ... with 8 more variables: flight <int>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, dep_type <chr>
```

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

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

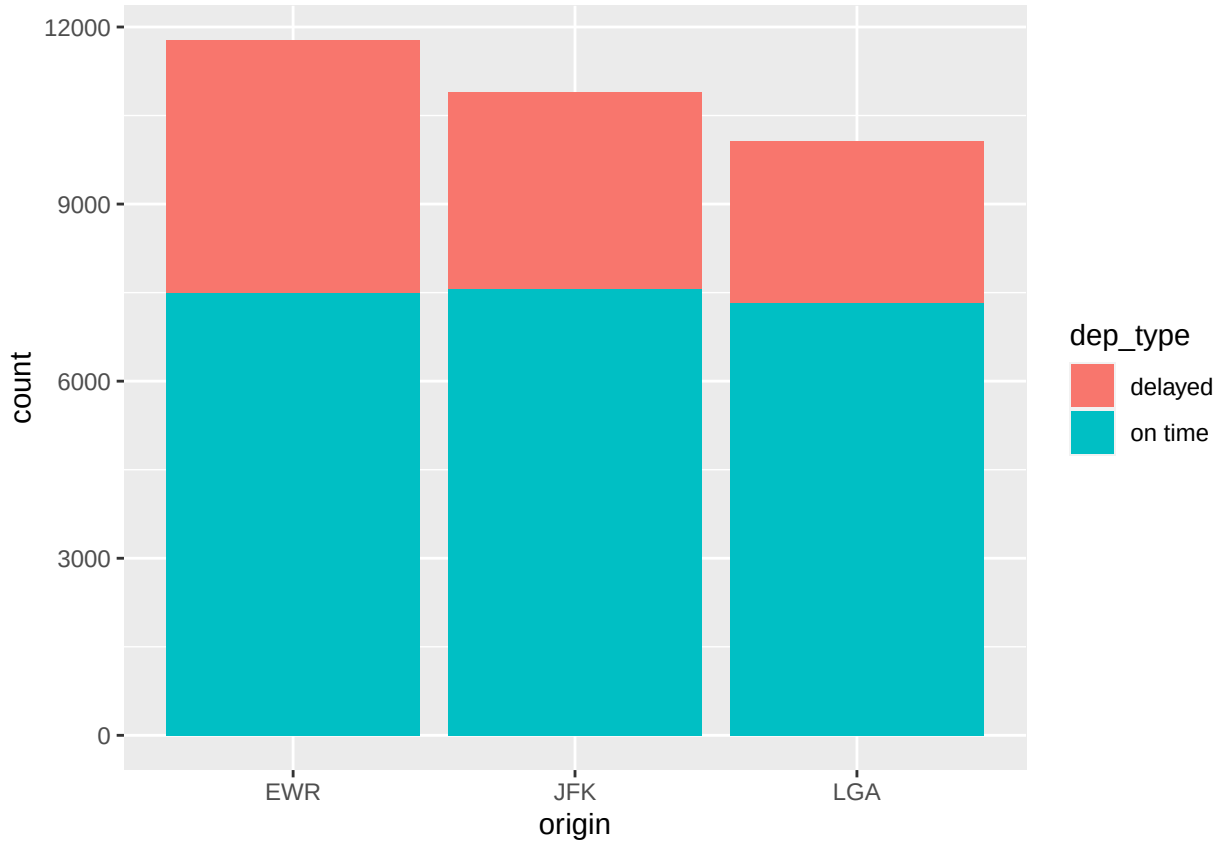
```
## # A tibble: 3 x 2
##   origin ot_dep_rate
##   <chr>        <dbl>
## 1 LGA          0.728
## 2 JFK          0.694
## 3 EWR          0.637
```

Exercise 6

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

Answer: I would choose La Guardia Airport (LGA) to fly out of NYC if I wanted the best chances to depart on time based on percentage.

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



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))
```

```
head(nycflights, 20)
```

```
## # A tibble: 20 x 18
##     year month   day dep_time dep_delay arr_time arr_delay carrier tailnum
##    <int> <int> <int>    <int>     <dbl>    <int>     <dbl> <chr>   <chr>
##  1  2013     6    30      940        15     1216        -4 VX      N626VA
##  2  2013     5     7     1657        -3     2104        10 DL      N3760C
##  3  2013    12     8      859        -1     1238        11 DL      N712TW
##  4  2013     5    14     1841        -4     2122       -34 DL      N914DL
##  5  2013     7    21     1102        -3     1230        -8 9E      N823AY
##  6  2013     1     1     1817        -3     2008         3 AA      N3AXAA
##  7  2013    12     9     1259        14     1617        22 WN      N218WN
```

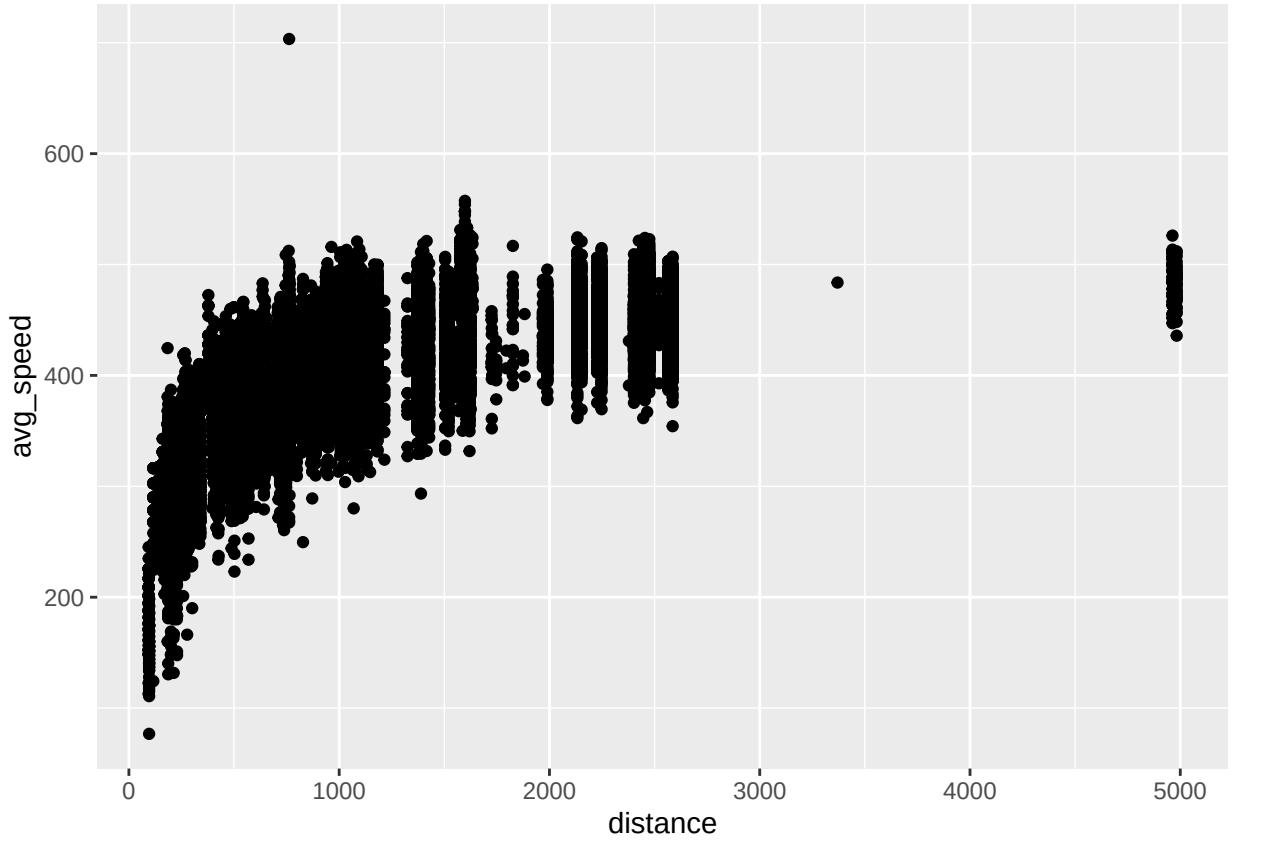
```
##  8  2013     8    13     1920        85     2032        71 B6      N284JB
##  9  2013     9    26      725       -10     1027        -8 AA      N3FSAA
## 10  2013     4    30     1323        62     1549        60 EV      N12163
## 11  2013     6    17      940         5     1050        -4 B6      N351JB
## 12  2013    11    22     1320         5     1628        -2 B6      N526JB
## 13  2013     4    26      809        -2     1030        22 EV      N16559
## 14  2013     3    25     2054       115     2256        91 FL      N919AT
## 15  2013    10    21     1217        -4     1322        -6 B6      N192JB
## 16  2013     1    23     2024        37     2141        29 EV      N17115
## 17  2013     2     8      644        -1      817        20 EV      N14916
## 18  2013     8     5      757        -3     1041       -23 DL      N380DA
## 19  2013    10    21      859        -1     1036        11 UA      N57852
## 20  2013     8    18     1638         8     1942       -17 VX      N849VA
## # ... with 9 more variables: flight <int>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, dep_type <chr>,
## #   avg_speed <dbl>
```

Exercise 8

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

Answer: While the relationship is not linear, it appears that the further distance to fly the plane's average speed increases.

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



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.

Answer: Departure cutoff point to still arrive at your destination is approximately one hour late departure.

```
carrier_flights <- nycflights %>%
  filter(carrier == "UA" | carrier == "AA" | carrier == "DL")
ggplot(data = carrier_flights, aes(x = dep_delay, y = arr_delay, color = carrier)) +
  geom_point()
```