DATA 606 Lab 2

Joao De Oliveira

2025-09-15

Load packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(openintro)
```

```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

The Data

```
data(nycflights)
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"
```

```
?nycflights
glimpse(nycflights)
```

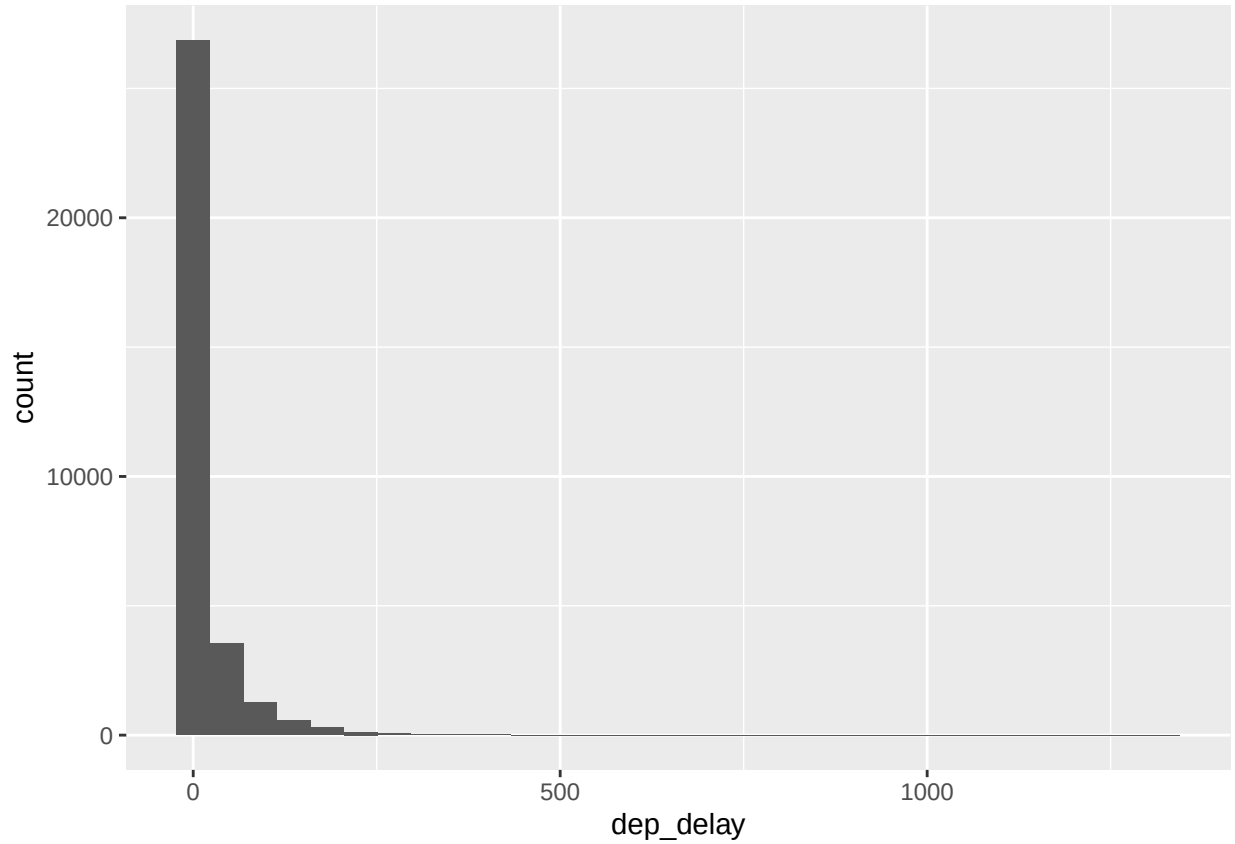
```
## Rows: 32,735
## Columns: 16
## $ year      <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, ~
```

```
## $ month     <int> 6, 5, 12, 5, 7, 1, 12, 8, 9, 4, 6, 11, 4, 3, 10, 1, 2, 8, 10~
## $ day       <int> 30, 7, 8, 14, 21, 1, 9, 13, 26, 30, 17, 22, 26, 25, 21, 23, ~
## $ dep_time  <int> 940, 1657, 859, 1841, 1102, 1817, 1259, 1920, 725, 1323, 940~
## $ dep_delay <dbl> 15, -3, -1, -4, -3, -3, 14, 85, -10, 62, 5, 5, -2, 115, -4, ~
## $ arr_time  <int> 1216, 2104, 1238, 2122, 1230, 2008, 1617, 2032, 1027, 1549, ~
## $ arr_delay <dbl> -4, 10, 11, -34, -8, 3, 22, 71, -8, 60, -4, -2, 22, 91, -6, ~
## $ carrier   <chr> "VX", "DL", "DL", "DL", "9E", "AA", "WN", "B6", "AA", "EV", ~
## $ tailnum   <chr> "N626VA", "N3760C", "N712TW", "N914DL", "N823AY", "N3AXAA", ~
## $ flight    <int> 407, 329, 422, 2391, 3652, 353, 1428, 1407, 2279, 4162, 20, ~
## $ origin    <chr> "JFK", "JFK", "JFK", "JFK", "LGA", "LGA", "EWR", "JFK", "LGA~
## $ dest      <chr> "LAX", "SJU", "LAX", "TPA", "ORF", "ORD", "HOU", "IAD", "MIA~
## $ air_time  <dbl> 313, 216, 376, 135, 50, 138, 240, 48, 148, 110, 50, 161, 87,~
## $ distance  <dbl> 2475, 1598, 2475, 1005, 296, 733, 1411, 228, 1096, 820, 264,~
## $ hour      <dbl> 9, 16, 8, 18, 11, 18, 12, 19, 7, 13, 9, 13, 8, 20, 12, 20, 6~
## $ minute    <dbl> 40, 57, 59, 41, 2, 17, 59, 20, 25, 23, 40, 20, 9, 54, 17, 24~
```

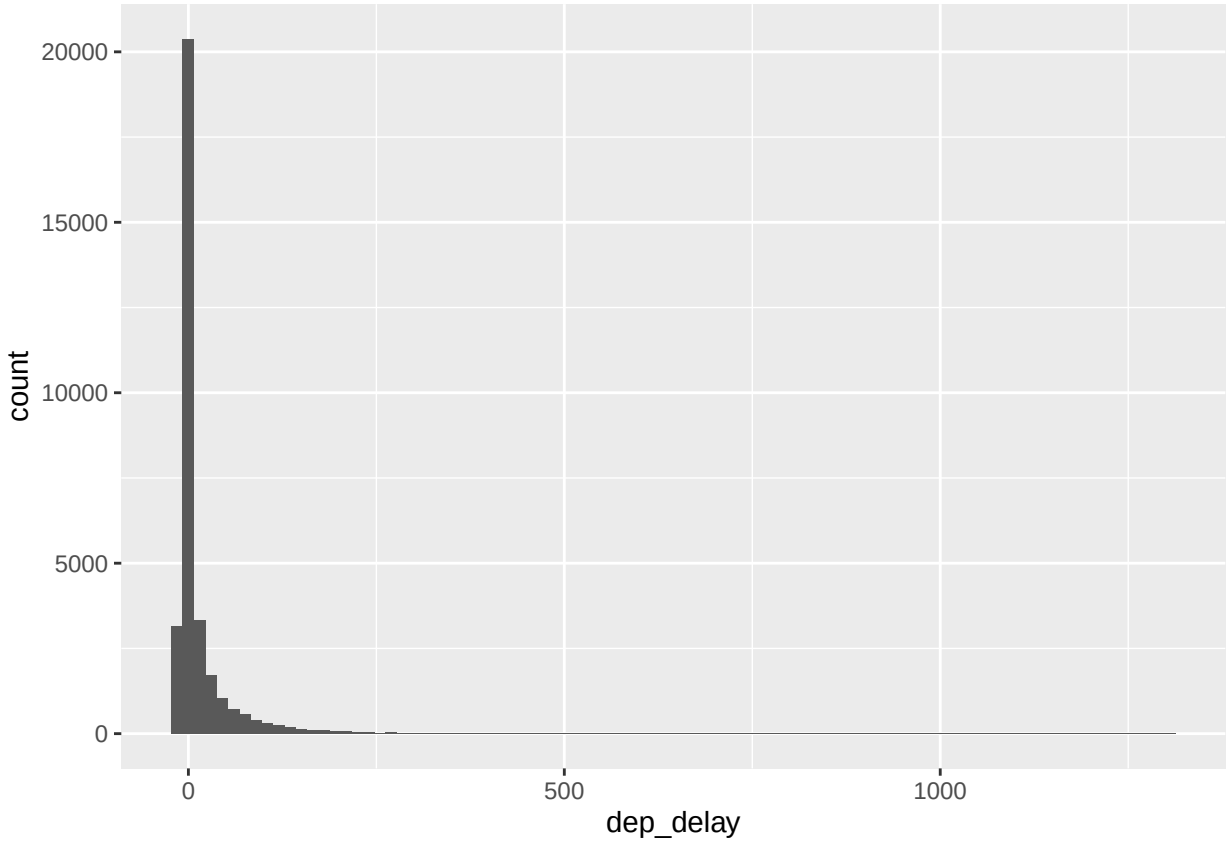
Plotting

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

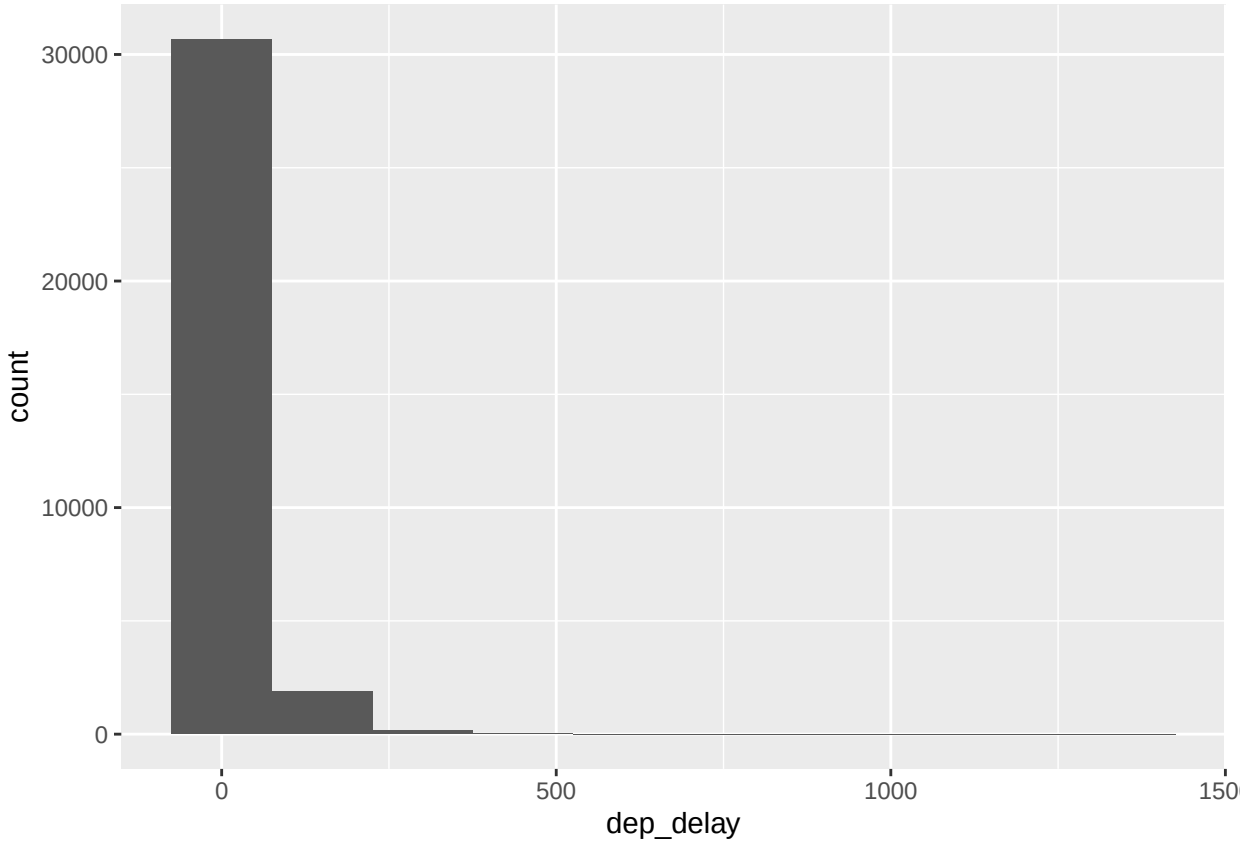
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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




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



Exercise 1

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

Answer: They all seem to show a similar structure, meaning that most flights depart on time (or close). However, the second histogram allows us to identify early flights as well, which histograms 1 and 3 can't really provide because of the interval of minutes (early, on time, and slightly late flights are all included in the same interval/bar of the histogram). All 3 histograms show that there are barely any flights with over 4 hours late departure (240min).

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(dest == "SFO", month == 2)

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

```
## # A tibble: 1 x 1
##       n
```

```
##   <int>
## 1    68
```

```
sfo_feb_flights
```

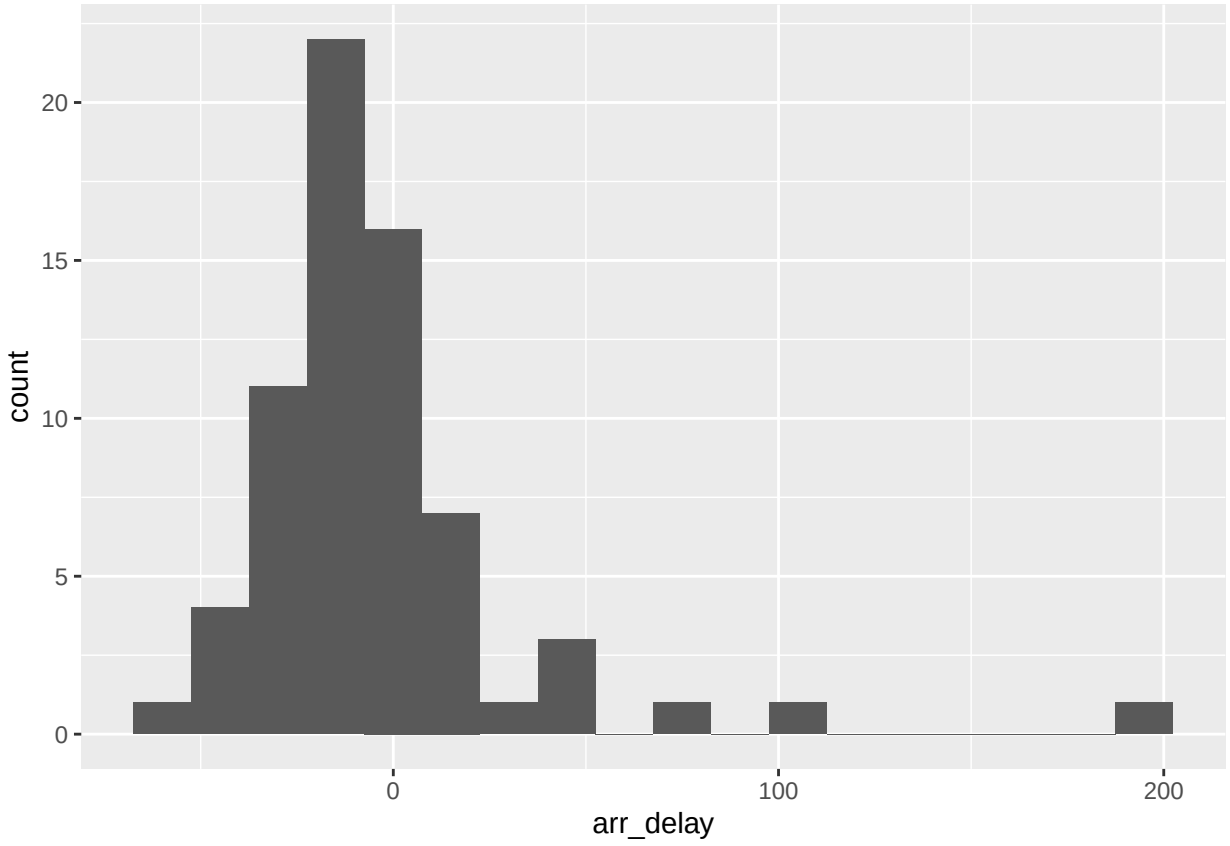
```
## # A tibble: 68 x 16
##     year month   day dep_time dep_delay arr_time arr_delay carrier tailnum
##    <int> <int> <int>    <int>     <dbl>    <int>     <dbl> <chr>   <chr>
##  1  2013     2    18     1527        57     1903        48 DL      N711ZX
##  2  2013     2     3      613        14     1008        38 UA      N502UA
##  3  2013     2    15      955        -5     1313       -28 DL      N717TW
##  4  2013     2    18     1928        15     2239        -6 UA      N24212
##  5  2013     2    24     1340         2     1644       -21 UA      N76269
##  6  2013     2    25     1415       -10     1737       -13 UA      N532UA
##  7  2013     2     7     1032         1     1352       -10 B6      N627JB
##  8  2013     2    15     1805        20     2122         2 AA      N335AA
##  9  2013     2    13     1056        -4     1412       -13 UA      N532UA
## 10  2013     2     8      656        -4     1039        -6 DL      N710TW
## # i 58 more rows
## # i 7 more variables: flight <int>, origin <chr>, dest <chr>, air_time <dbl>,
## #   distance <dbl>, hour <dbl>, minute <dbl>
```

Answer: As we can see using n(), there 68 flights that meet these criteria.

Exercise 3

Question: 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.

```
sfo_feb_flights %>%
  filter(!is.na(arr_delay)) %>%
  ggplot(aes(x = arr_delay)) +
  geom_histogram(binwidth = 15)
```




```
# summary statistics
sfo_feb_flights %>%
  summarise(
    mean_sfo_delay = mean(arr_delay, na.rm = TRUE),
    median_sfo_delay = median(arr_delay, na.rm = TRUE),
    sd_sfo_delay = sd(arr_delay, na.rm = TRUE),
    min_sfo_delay = min(arr_delay, na.rm = TRUE),
    max_sfo_delay = max(arr_delay, na.rm = TRUE),
    var_sfo_delay = var(arr_delay, na.rm = TRUE),
    iqr_sfo_delay = IQR(arr_delay, na.rm = TRUE)
  )
```

```
## # A tibble: 1 x 7
##   mean_sfo_delay median_sfo_delay sd_sfo_delay min_sfo_delay max_sfo_delay
##            <dbl>            <dbl>        <dbl>         <dbl>         <dbl>
## 1           -4.5              -11         36.3           -66           196
## # i 2 more variables: var_sfo_delay <dbl>, iqr_sfo_delay <dbl>
```

Answer (question 3): Looking at the graph, it seems like most flights arrived on time or early. It is skewed to the right, since the tail on the right side is longer than the left one.

Exercise 4

Question: 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?

```r
sfo_feb_flights %>%
  group_by(carrier) %>%
  summarise(
    median_delay = median(arr_delay, na.rm = TRUE),
    iqr_delay = IQR(arr_delay, na.rm = TRUE),
    .groups = "drop"
  )
```

```
## # A tibble: 5 x 3
##   carrier median_delay iqr_delay
##   <chr>          <dbl>     <dbl>
## 1 AA               5        17.5
## 2 B6             -10.5      12.2
## 3 DL             -15        22
## 4 UA             -10        22
## 5 VX             -22.5      21.2
```

Answer: As we can see using the code above, American Airlanes has the most variavle arrival delay.

Exercise 5

Question: 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?

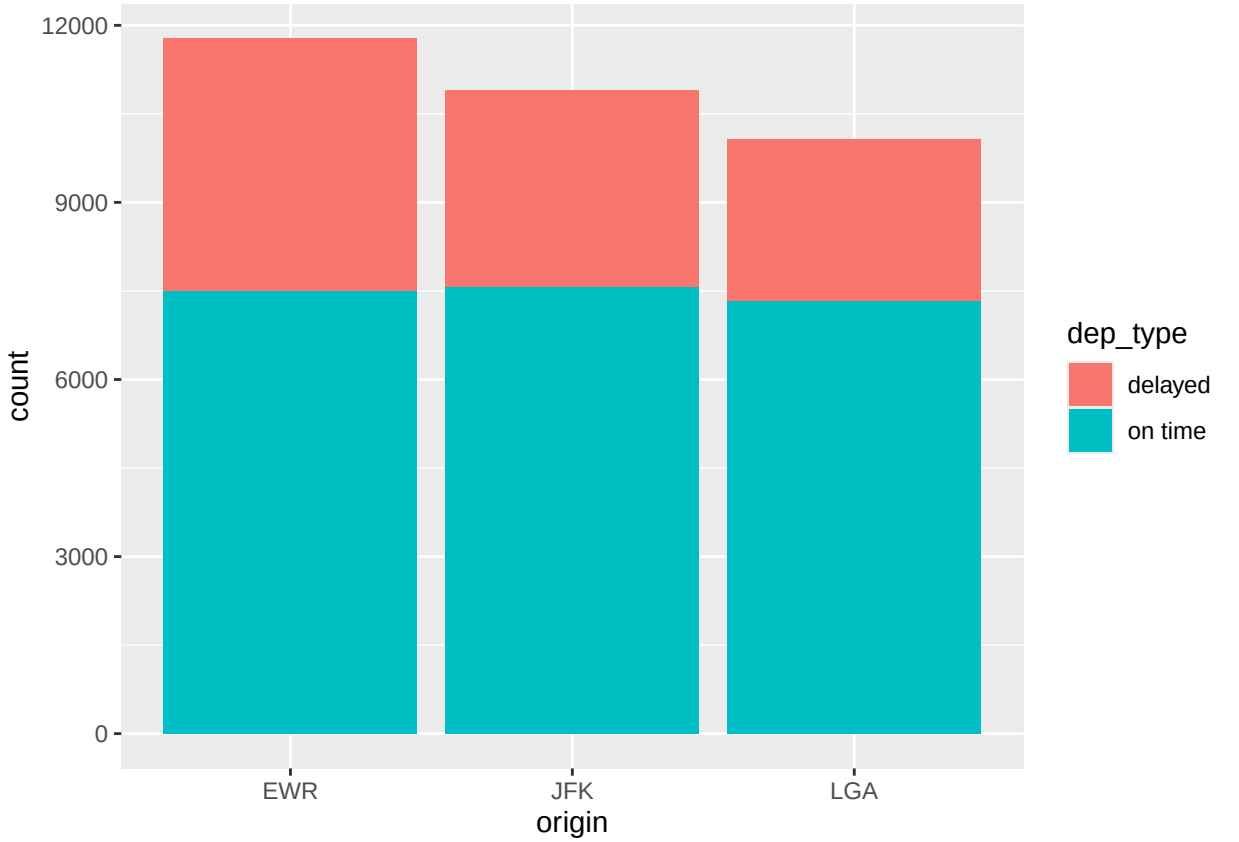
Answer: Since the graph of departure delays is skewed due to the right it has some flights with very long delays, which are clear outliers, which indicates that the median is the best choice here. The median is not distorted by outliers. The median defines the point where half of the flights departed earlier and half later than that. The median has the downside of dismissing (or hiding) the extreme outliers (late flights). The mean, however, can be extremely impacted by extreme outliers, which sometimes can be a poor representation of reality. The mean has the advantage of taking into consideration all the data, which might be relevant for someone who cares a lot about departure delays.

On time departure rate for NYC airports

```r
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
```

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



Exercise 6

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

Answer: Looking at the resulting table and plot from the code above, we can see that LaGuardia airport is the NYC airport with the highest rate of departures on time, so that would be the choice.

Exercise 7

Question: 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.

Answer:

```r
nycflights <- nycflights %>%
  mutate(air_time_h = air_time / 60) %>%
  mutate(
    average_speed = if_else(
      !is.na(air_time_h) & air_time_h > 0,
```

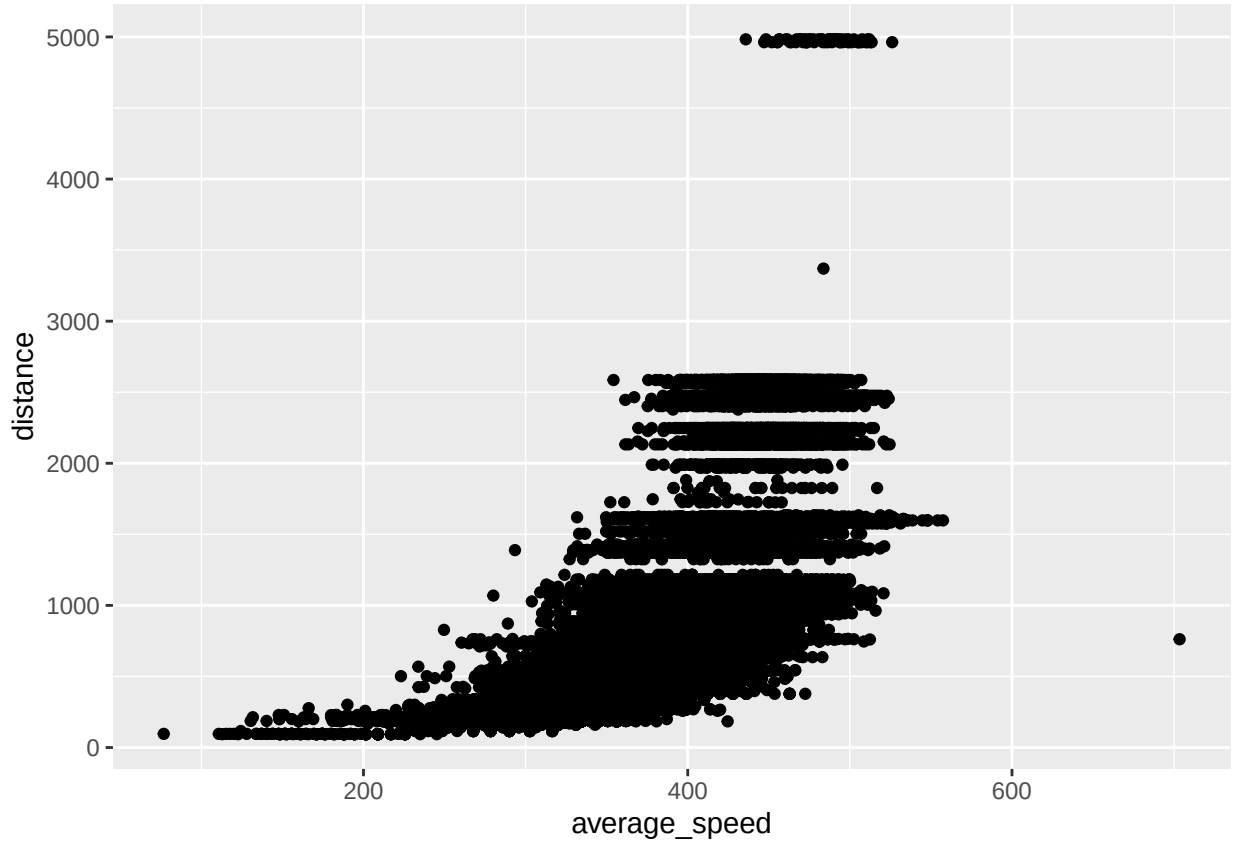
```
      distance / air_time_h,
      NA
    )
  )

nycflights %>%
  select(origin, dest, distance, air_time, air_time_h, average_speed) %>%
  slice_head(n = 10)
```

```
## # A tibble: 10 x 6
##    origin dest  distance air_time air_time_h average_speed
##    <chr>  <chr>    <dbl>    <dbl>      <dbl>         <dbl>
##  1 JFK    LAX       2475      313      5.22           474.
##  2 JFK    SJU       1598      216      3.6            444.
##  3 JFK    LAX       2475      376      6.27           395.
##  4 JFK    TPA       1005      135      2.25           447.
##  5 LGA    ORF        296       50      0.833          355.
##  6 LGA    ORD        733      138      2.3            319.
##  7 EWR    HOU       1411      240      4              353.
##  8 JFK    IAD        228       48      0.8            285
##  9 LGA    MIA       1096      148      2.47           444.
## 10 EWR    JAX        820      110      1.83           447.
```

Exercise 8

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

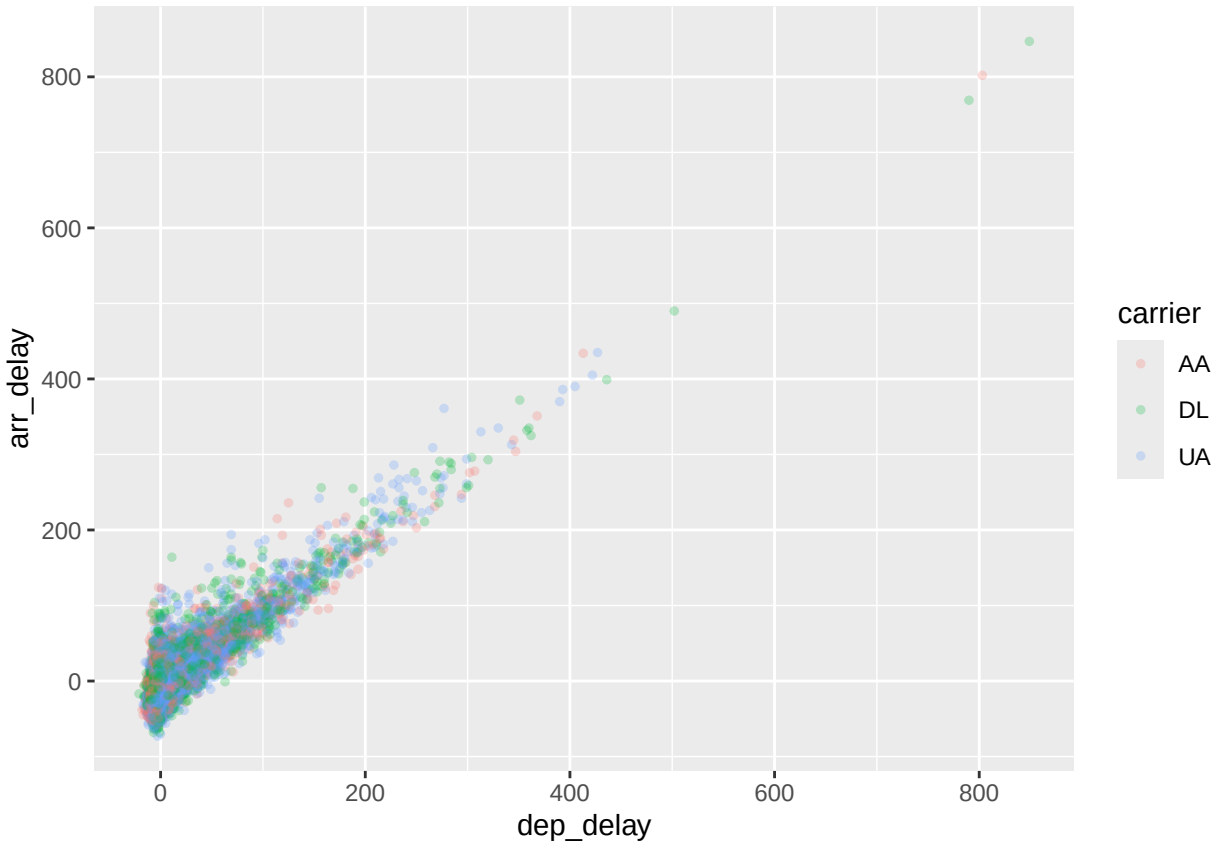
```
ggplot(data = nycflights, aes(x = average_speed, y = distance)) +
  geom_point()
```



Exercie 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.

```
nyc_top3 <- nycflights %>%
  filter(carrier %in% c("AA", "DL", "UA")) %>%
  filter(!is.na(dep_delay), !is.na(arr_delay))

#scatter plot
ggplot(nyc_top3, aes(x = dep_delay, y = arr_delay, color = carrier)) +
  geom_point(alpha = 0.25, size = 1) +
  labs(x = "dep_delay", y = "arr_delay", color = "carrier")
```




```
# data-driven cutoff
cutoff_t <- nyc_top3 %>%
  mutate(dep_bin = floor(dep_delay / 5) * 5) %>%
  group_by(dep_bin) %>%
  summarise(ontime_rate = mean(arr_delay <= 0, n = n(), .groups = "drop")) %>%
  arrange(dep_bin)

cutoff_dep <- cutoff_t %>%
  filter(ontime_rate >= 0.5) %>%
  summarise(cutoff = max(dep_bin, na.rm = TRUE)) %>%
  pull(cutoff)

cutoff_dep
```

```
## [1] 5
```

Answer: We can see a positive relationship between departure and arrival times. This means that beyond the cutoff point, which is 5 minutes, the probability of arriving late increases.