Introduction to Data

Enid Roman

2022-09-18

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(openintro)
```

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

###**nycflights**

```
data('nycflights', package='openintro')
```

```
nycflights
```

```
## # A tibble: 32,735 x 16
##     year month   day dep_time dep_delay arr_time arr_de~1 carrier tailnum flight
##    <int> <int> <int>    <int>     <dbl>    <int>    <dbl> <chr>   <chr>    <int>
##  1  2013     6    30      940        15     1216       -4 VX      N626VA     407
##  2  2013     5     7     1657        -3     2104       10 DL      N3760C     329
##  3  2013    12     8      859        -1     1238       11 DL      N712TW     422
##  4  2013     5    14     1841        -4     2122      -34 DL      N914DL    2391
##  5  2013     7    21     1102        -3     1230       -8 9E      N823AY    3652
```

```
##  6  2013     1     1     1817        -3     2008        3 AA      N3AXAA     353
##  7  2013    12     9     1259        14     1617       22 WN      N218WN    1428
##  8  2013     8    13     1920        85     2032       71 B6      N284JB    1407
##  9  2013     9    26      725       -10     1027       -8 AA      N3FSAA    2279
## 10  2013     4    30     1323        62     1549       60 EV      N12163    4162
## # ... with 32,725 more rows, 6 more variables: origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, and abbreviated
## #   variable name 1: arr_delay
```

```
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
```

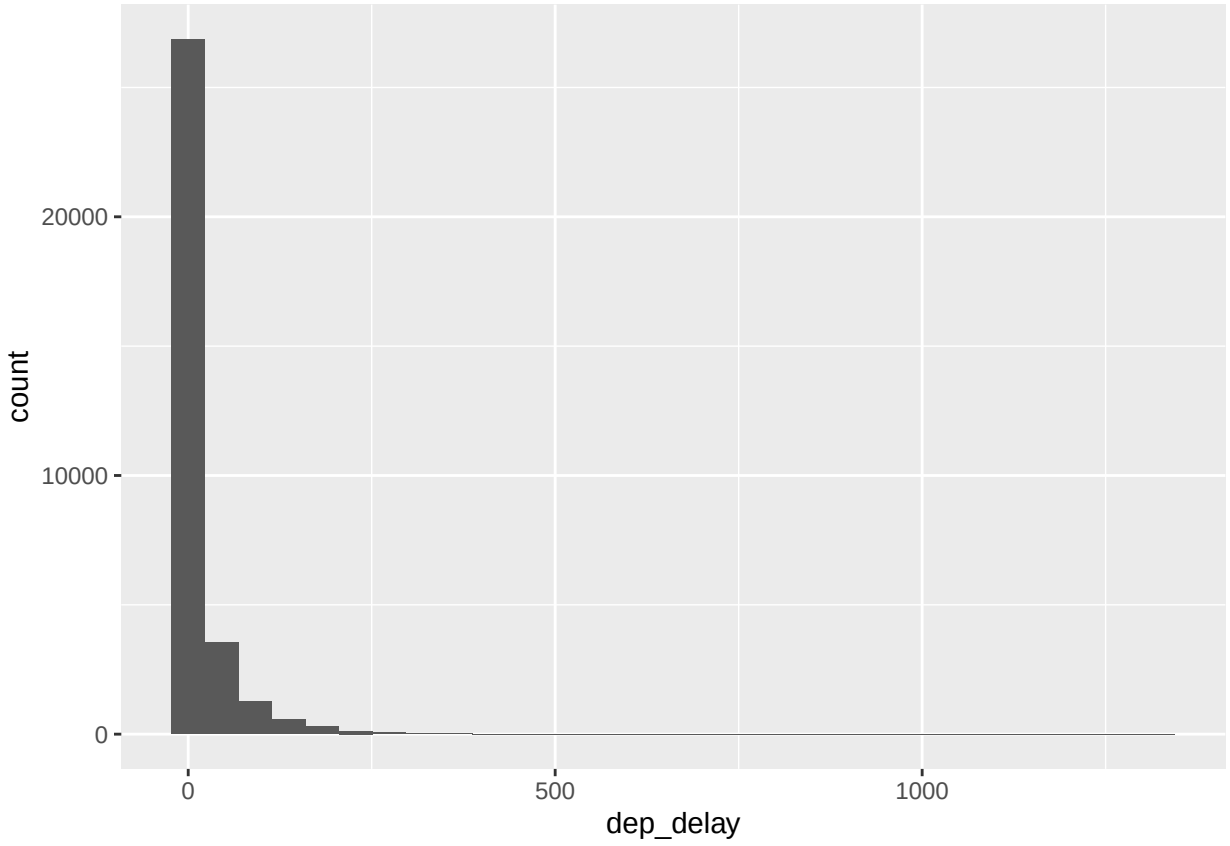
```
## starting httpd help server ... done
```

```
glimpse(nycflights)
```

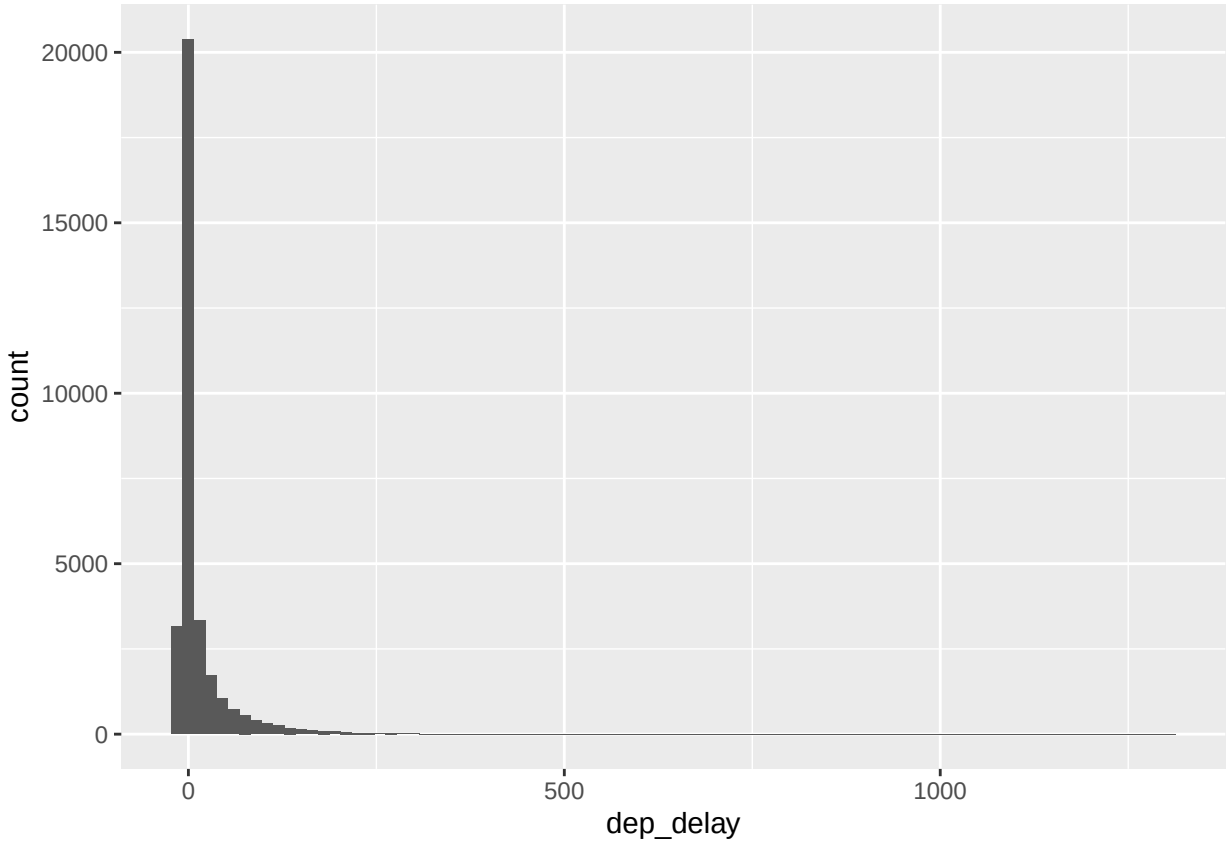
```
## Rows: 32,735
## Columns: 16
## $ year      <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, ~
## $ month     <int> 6, 5, 12, 5, 7, 1, 12, 8, 9, 4, 6, 11, 4, 3, 10, 1, 2, 8, 10~
## $ day       <int> 30, 7, 8, 14, 21, 1, 9, 13, 26, 30, 17, 22, 26, 25, 21, 23, ~
## $ dep_time  <int> 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~
```

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

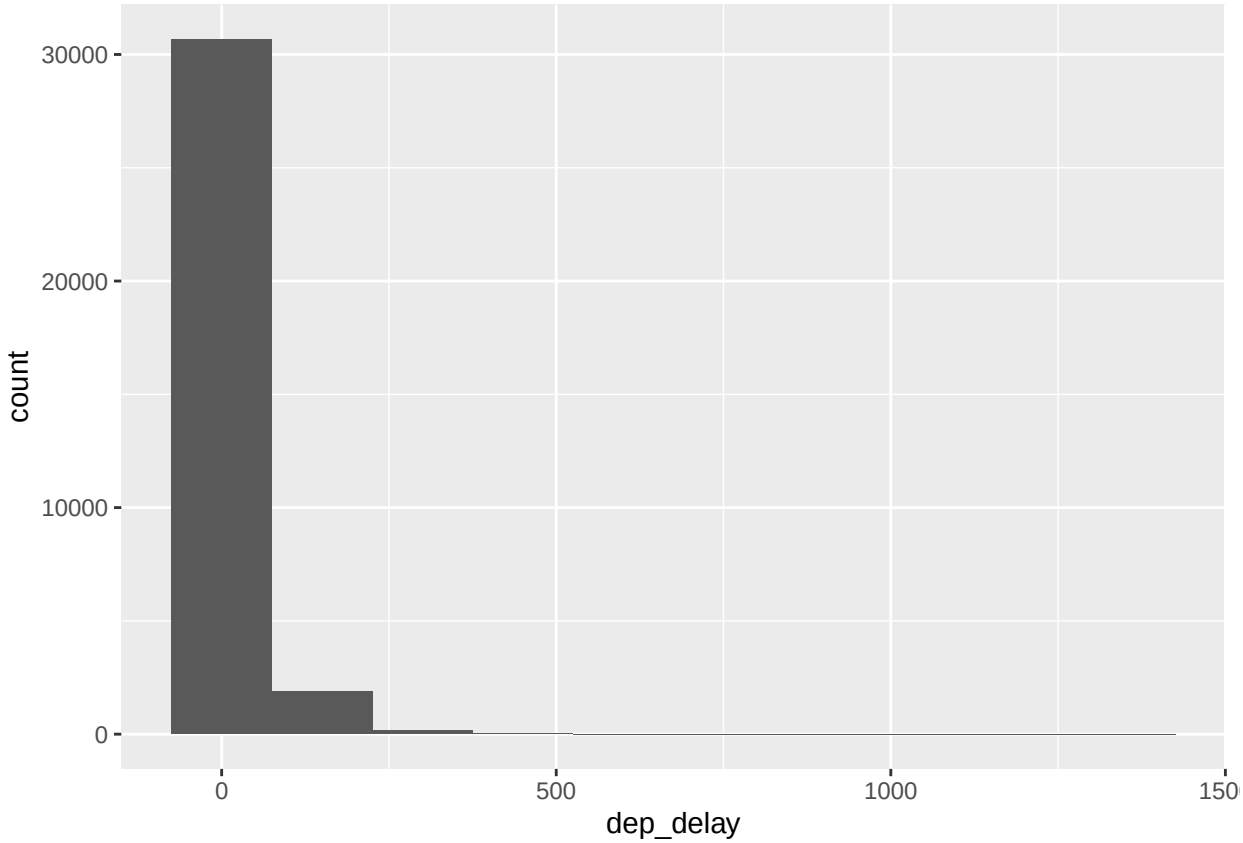
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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




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

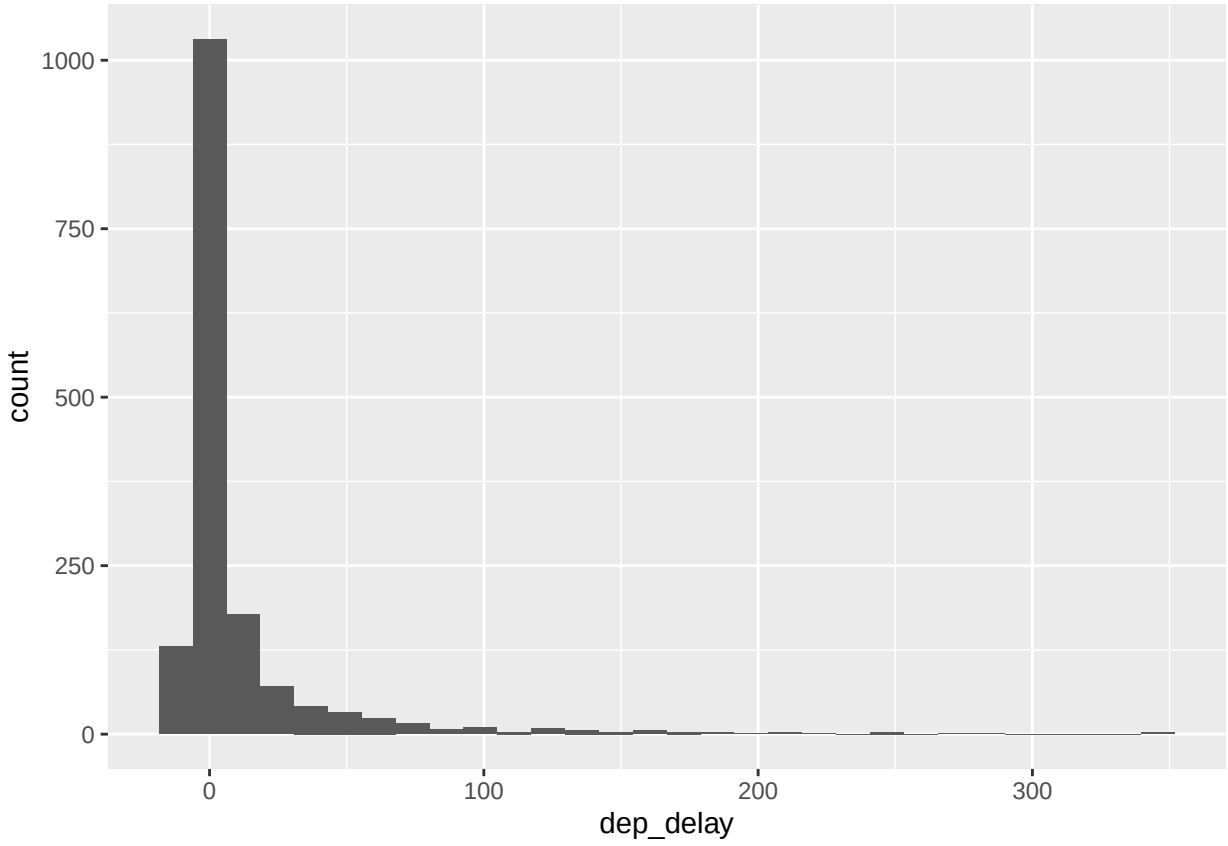


Exercise 1

Yes, more departure delay is revealed in binwidth 15 then the regular histogram and binwidth 150. Binwidth 150 has departure delay that are obsecured.

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

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




```
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
```

Excercise 2

There are 68 flights that meet the below criteria.

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

```
## # A tibble: 68 x 16
##     year month   day dep_time dep_delay arr_time arr_de~1 carrier tailnum flight
##    <int> <int> <int>    <int>     <dbl>    <int>    <dbl> <chr>   <chr>    <int>
##  1  2013     2    18     1527        57     1903       48 DL      N711ZX    1322
##  2  2013     2     3      613        14     1008       38 UA      N502UA     691
##  3  2013     2    15      955        -5     1313      -28 DL      N717TW    1765
```

```
##  4  2013     2    18     1928        15     2239       -6 UA      N24212    1214
##  5  2013     2    24     1340         2     1644      -21 UA      N76269    1111
##  6  2013     2    25     1415       -10     1737      -13 UA      N532UA     394
##  7  2013     2     7     1032         1     1352      -10 B6      N627JB     641
##  8  2013     2    15     1805        20     2122        2 AA      N335AA     177
##  9  2013     2    13     1056        -4     1412      -13 UA      N532UA     642
## 10  2013     2     8      656        -4     1039       -6 DL      N710TW    1865
## # ... with 58 more rows, 6 more variables: origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, and abbreviated
## #   variable name 1: arr_delay
```

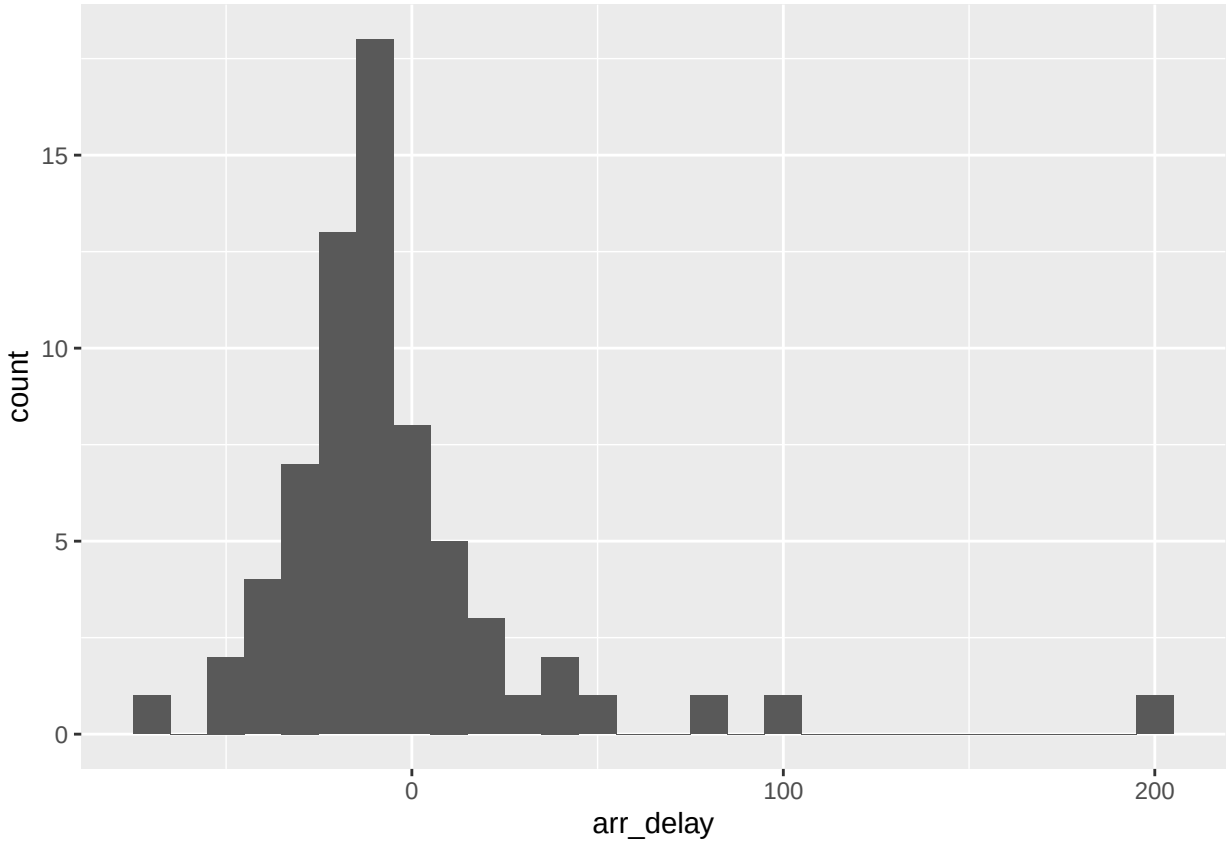
```
sfo_feb_flights %>%
summarise (n = n())
```

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

Excersie 3

The histogram is a right-skewed distribution in which most values are clustered around the left tail of the distribution while the right tail of the distribution is longer.

```
ggplot(data = sfo_feb_flights, aes(x = arr_delay)) +
  geom_histogram(binwidth = 10)
```



Since the distribution is skewed, we would use the median and IQR to describe the distribution

```
sfo_feb_flights %>%
  # summarize(mean(arr_delay), median(arr_delay), max(arr_delay))
  summarise(median_ad = median(arr_delay),
            iqr_ad = IQR(arr_delay),
            n_flights = n())
```

```
## # A tibble: 1 x 3
##   median_ad iqr_ad n_flights
##       <dbl>  <dbl>     <int>
## 1       -11   23.2        68
```

```
sfo_feb_flights %>%
  group_by(origin) %>%
  summarise(median_dd = median(dep_delay), iqr_dd = IQR(dep_delay), n_flights = n())
```

```
## # A tibble: 2 x 4
##   origin median_dd iqr_dd n_flights
##   <chr>      <dbl>  <dbl>     <int>
## 1 EWR          0.5   5.75         8
## 2 JFK         -2.5  15.2         60
```

Exercise 4

Both DL and UA has the most variable arrival delays because their IQR are both the highest which is 22.00.

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

```
## # A tibble: 5 x 4
##   carrier median_arr iqr_arr 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
```

I would expect December to have the highest delays because of the cold and snow. But the data below shows that is actually July.

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

```
## # A tibble: 12 x 2
##    month mean_dd
##    <int>   <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
```

Excerise 5

The pro of using the mean is it actually giving you an average of delays for each month, showing the affect of each delay and showing how the data is distributed. The con of using mean it can be affected by outliers.

The pro of using median is it uses the middle value of the entire data set, so the outliers do not affect the median. The con of using the median is it's not showing the whole data distribution.

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

```
## # A tibble: 12 x 2
```

```
##    month median_dd
##    <int>     <dbl>
##  1    12         1
##  2     6         0
##  3     7         0
##  4     3        -1
##  5     5        -1
##  6     8        -1
##  7     1        -2
##  8     2        -2
##  9     4        -2
## 10    11        -2
## 11     9        -3
## 12    10        -3
```

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

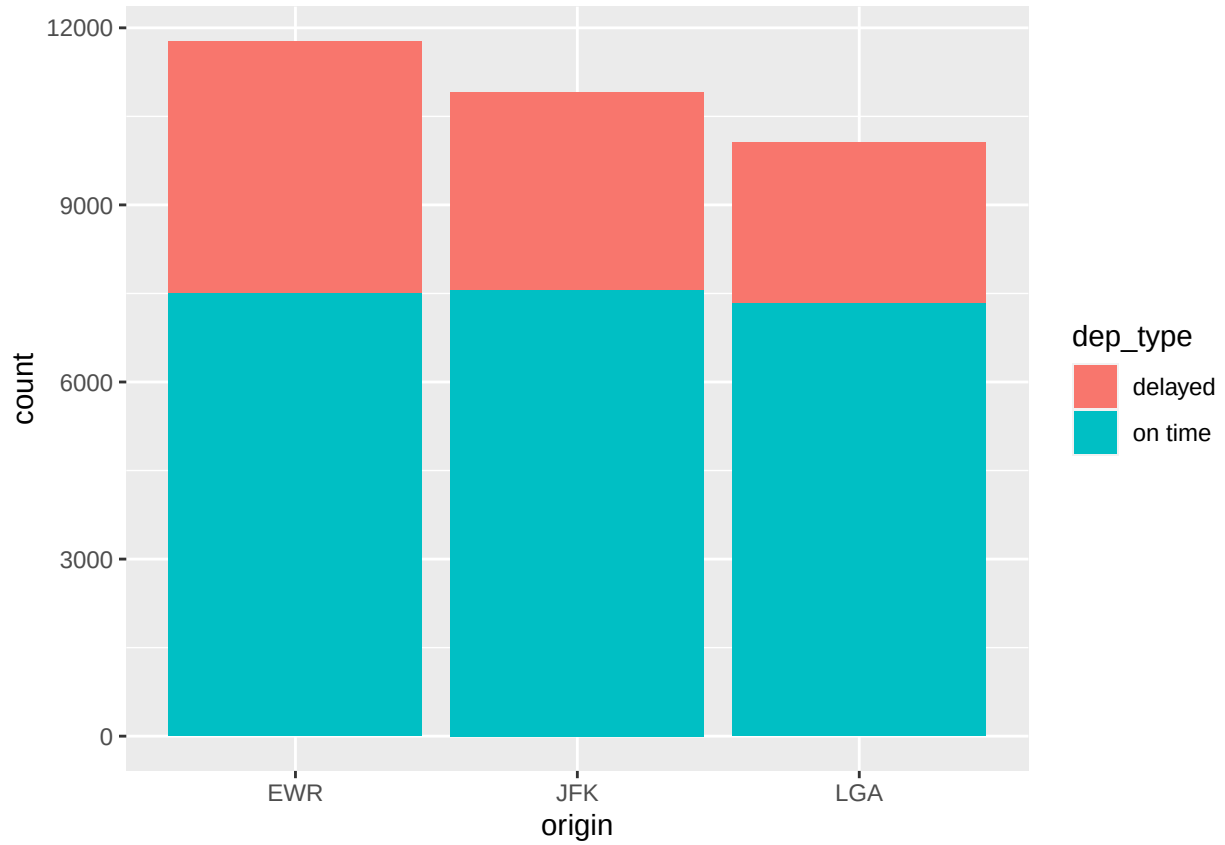
```
nycflights %>%
  group_by(origin) %>%
  summarise(ot_dep_rate = sum(dep_type == "on time") / n()) %>%
  arrange(desc(ot_dep_rate))
```

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

Excerise 6

Based on the above departure rate, LGA at .73 would be the NYC airport I would choose to fly out of. Based on the graph below LGA has the least departure delays then JFK and EWR.

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



Excerise 7

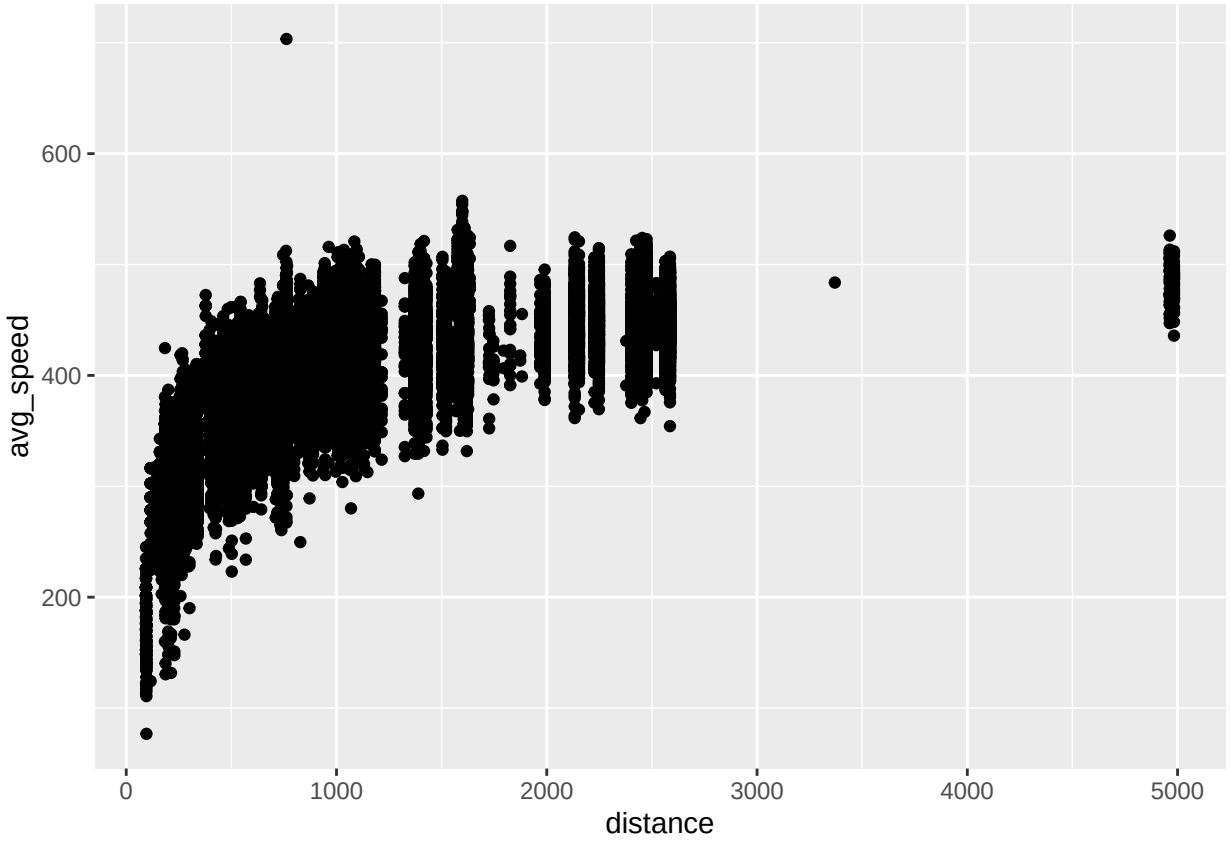
```
nycflights <- nycflights %>%
  mutate(avg_speed = 60*(distance / air_time))
glimpse(nycflights)
```

```
## Rows: 32,735
## Columns: 18
## $ 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~
## $ dep_type  <chr> "delayed", "on time", "on time", "on time", "on time", "on t~
## $ avg_speed <dbl> 474.4409, 443.8889, 394.9468, 446.6667, 355.2000, 318.6957, ~
```

Excerise 8

The relationship between average speed and distance in the scatter plot below is as the distance increases so does the average speed increases. There is a postive association between distance and average speed,

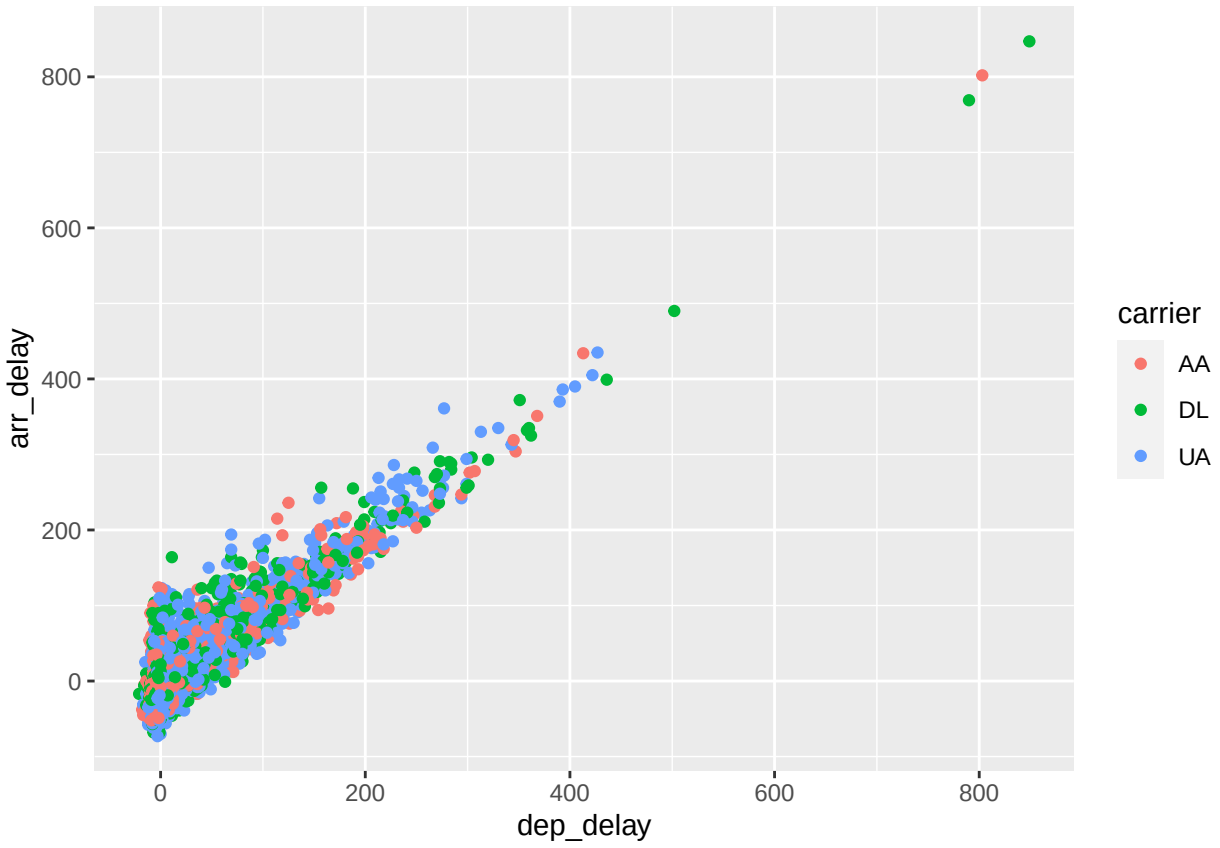
```
ggplot(data = nycflights, aes(distance, avg_speed)) +
  geom_point()
```



Excerise 9

Based on the scatterplot below the cutoff point for departure delays where you can still expect to get to your destination on time is is approximately 6 minutes after departure time, which is very rare.

```
nycflights_carrier <- nycflights %>%
filter(carrier == "AA" | carrier == "DL" | carrier == "UA")
ggplot(data = nycflights_carrier, 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