Introduction to data: **LAB2**

Gabriel Campos

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

Basic R Markdown with an OpenIntro Lab

Lab report

To record your analysis in a reproducible format, you can adapt the general Lab Report template from the **openintro** package. Watch the video above to learn how.

```
names(nycflights)
```

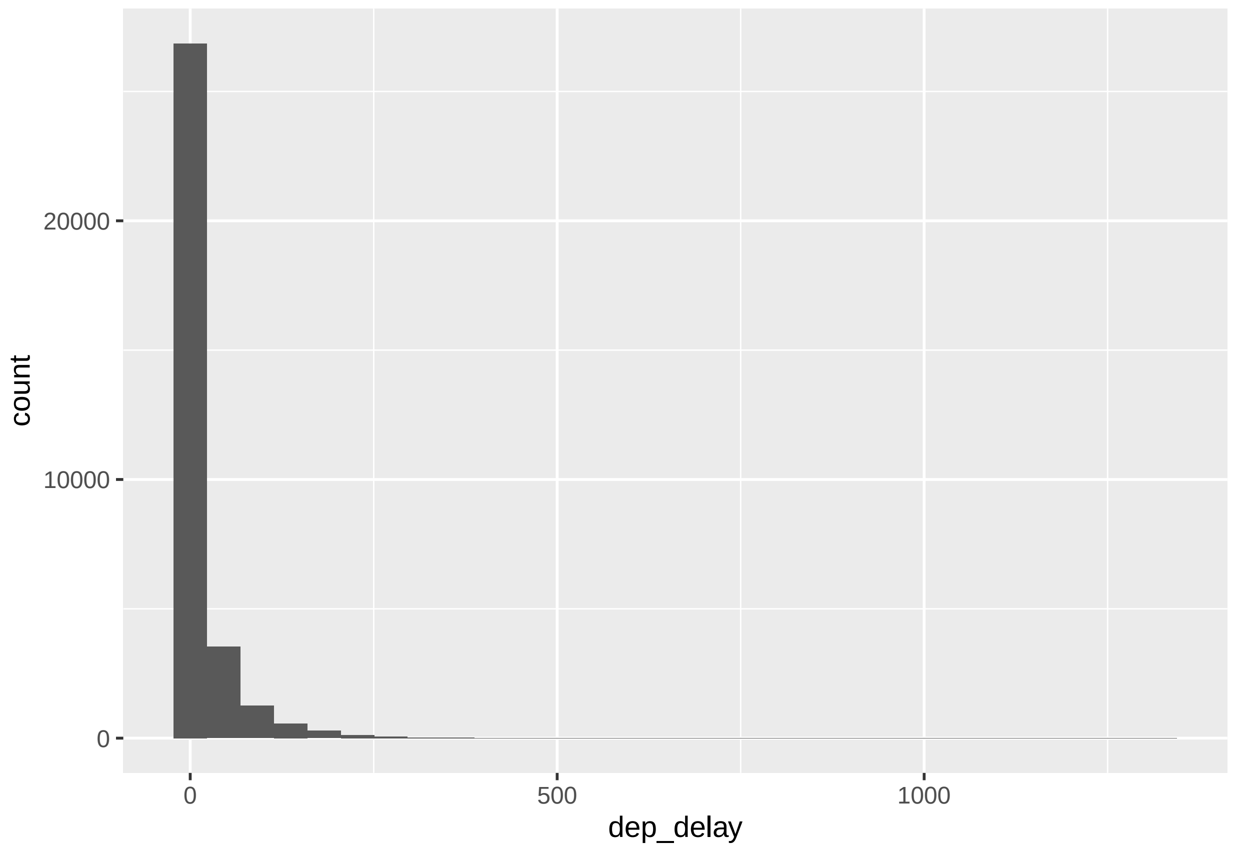
```
##  [1] "year"      "month"     "day"       "dep_time"  "dep_delay" "arr_time" 
##  [7] "arr_delay" "carrier"   "tailnum"   "flight"    "origin"    "dest"     
## [13] "air_time"  "distance"  "hour"      "minute"
```

Exercise 1

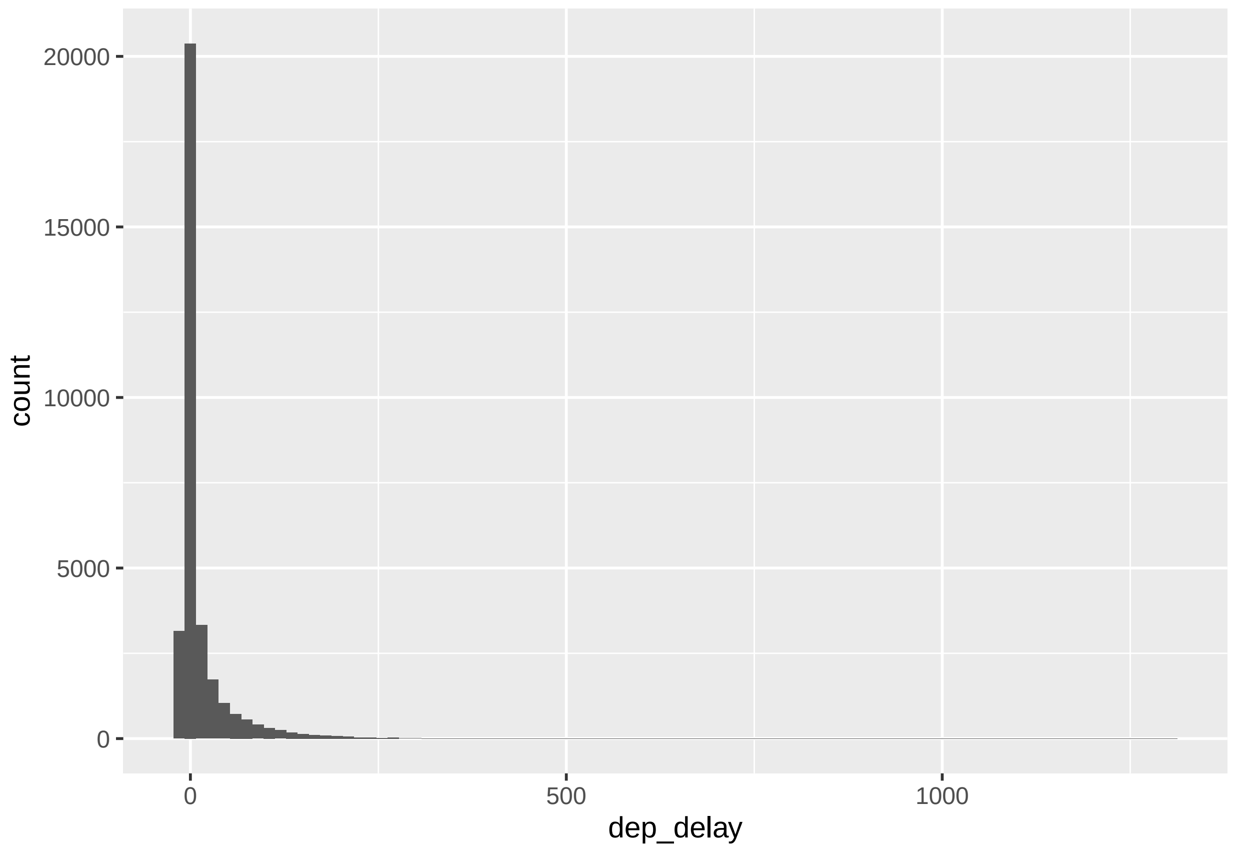
1. Look carefully at these three histograms. How do they compare? Are features revealed in one that are obscured in another?

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

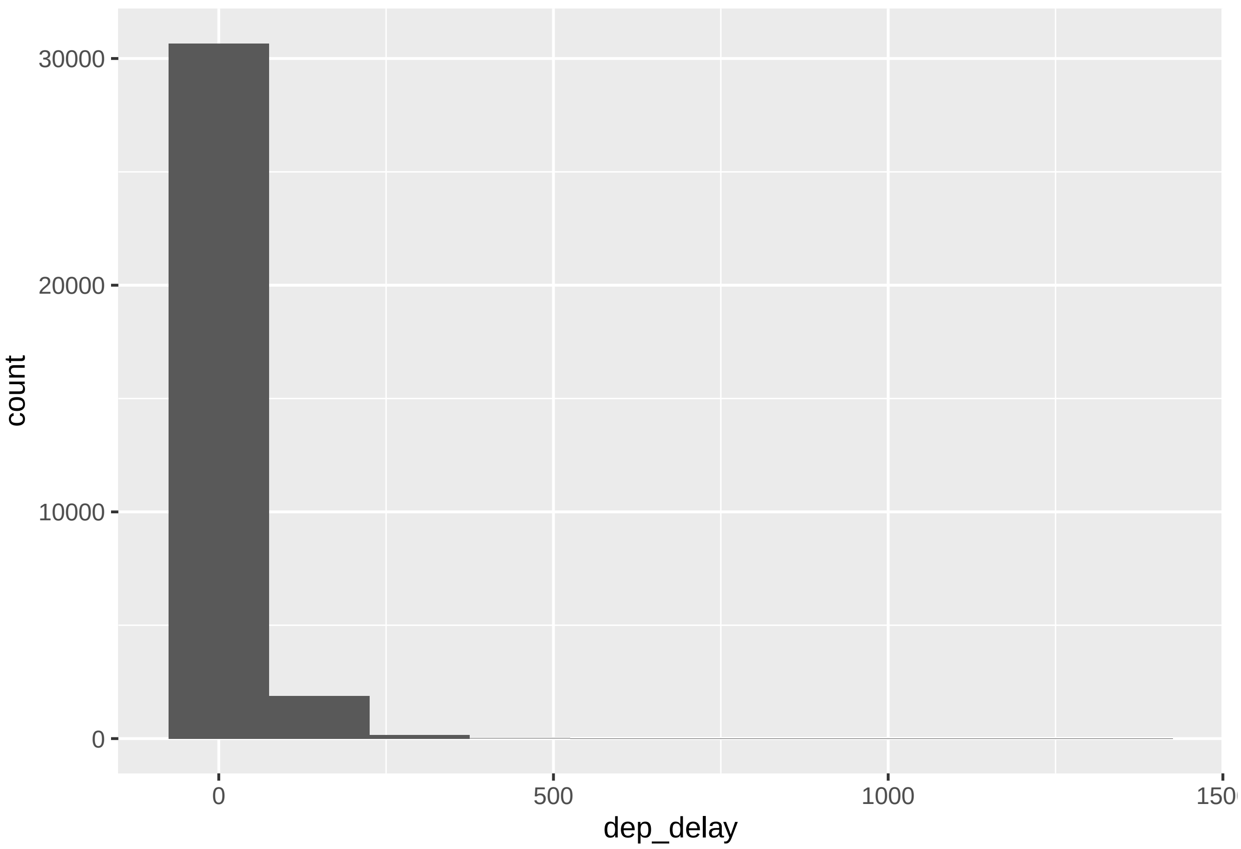
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



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



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



Exercise 1 Answer:

i) binwidth = 15 shows departure delays before the peak amount and is symmetrical

ii) binwidth = 150 shows a larger count for number of delay flights. Approximately 100 more and is left skewed.

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, 201...
## $ month     <int> 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, ...
## $ dep_time  <int> 1527, 613, 955, 1928, 1340, 1415, 1032, 1805, 1056, 656, ...
## $ dep_delay <dbl> 57, 14, -5, 15, 2, -10, 1, 20, -4, -4, 40, -2, -1, -6, -7...
## $ arr_time  <int> 1903, 1008, 1313, 2239, 1644, 1737, 1352, 2122, 1412, 103...
## $ arr_delay <dbl> 48, 38, -28, -6, -21, -13, -10, 2, -13, -6, 2, -5, -30, -...
## $ 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, 27...
## $ origin    <chr> "JFK", "JFK", "JFK", "EWR", "EWR", "JFK", "JFK", "JFK", "...
## $ dest      <chr> "SFO", "SFO", "SFO", "SFO", "SFO", "SFO", "SFO", "SFO", "...
## $ air_time  <dbl> 358, 367, 338, 353, 341, 355, 359, 338, 347, 361, 332, 35...
## $ distance  <dbl> 2586, 2586, 2586, 2565, 2565, 2586, 2586, 2586, 2586, 258...
## $ hour      <dbl> 15, 6, 9, 19, 13, 14, 10, 18, 10, 6, 19, 8, 10, 18, 7, 17...
## $ minute    <dbl> 27, 13, 55, 28, 40, 15, 32, 5, 56, 56, 10, 33, 48, 49, 23...
```

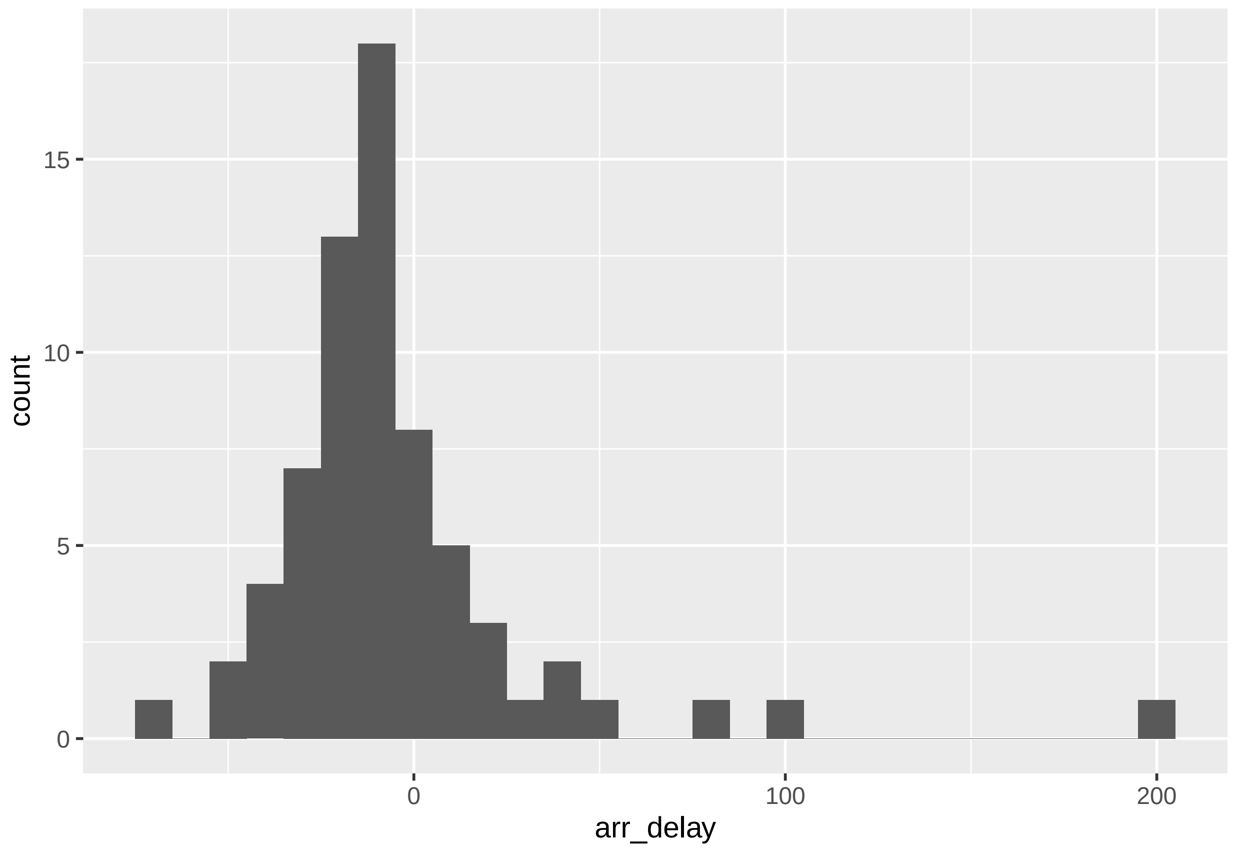
Exercise 2 Answer:

68 flights meet these criteria.

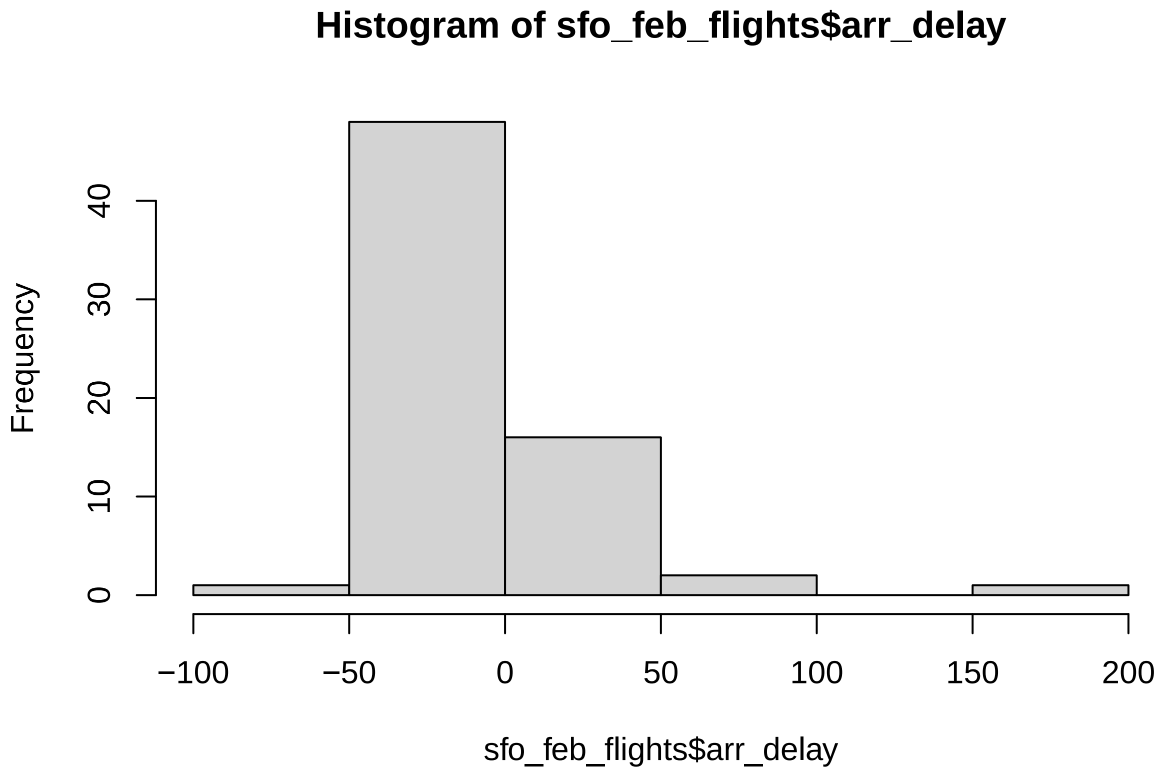
Exercise 3

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.

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



```
### Alternative method based off of https://www.statmethods.net/graphs/density.html
hist(sfo_feb_flights$arr_delay)
```



Note: Histogram is symmetric and unimodal with one potential outlier.

Exercise 3 Answer:

For the bell curve created by the histogram, the appropriate summary statistics are the numerical stats mean, media, interquartile range, standard deviation and number of values "n".

```
sfo_feb_flights %>%
  summarise(mean_ad   = mean(arr_delay),
            median_ad = median(arr_delay),
            iqr_ad = IQR(arr_delay),
            sd_ad = sd(arr_delay),
            n_ad = n())
```

```
## # A tibble: 1 x 5
##   mean_ad median_ad iqr_ad sd_ad  n_ad
##     <dbl>     <dbl>  <dbl> <dbl> <int>
## 1    -4.5       -11   23.2  36.3    68
```

Exercise 4

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?

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

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

```
## # A tibble: 5 x 3
##   carrier median_ad iqr_ad
##   <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
```

Exercise 4 Answer:

United Airlines Inc. (UA) and Delta Airlines Inc. (DL) have the highest interquartile range. IQR is a measure of variability. Hence the answer is UA and DL

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), median(dep_delay))
```

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

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

Exercise 5 Answer:

Mean is the optimal choice. An average gives a calculated value of a the potential delay for the month. The median is uncalculated, robust and less sensitive to outliers in the data set for the month (e.g. 1/2013 saw at least one triple digit delay) hence is not the best choice.

On time departure rate for NYC airports

Suppose you will be flying out of NYC and want to know which of the three major NYC airports has the best on time departure rate of departing flights. Also supposed that for you, a flight that is delayed for less than 5 minutes is basically "on time."" You consider any flight delayed for 5 minutes of more to be "delayed". In order to determine which airport has the best on time departure rate, you can

- first classify each flight as "on time" or "delayed",
- then group flights by origin airport,
- then calculate on time departure rates for each origin airport,
- and finally arrange the airports in descending order for on time departure percentage.

Let's start with classifying each flight as "on time" or "delayed" by creating a new variable with the `mutate` function.

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

The first argument in the `mutate` function is the name of the new variable we want to create, in this case `dep_type`. Then if `dep_delay < 5`, we classify the flight as `"on time"` and `"delayed"` if not, i.e. if the flight is delayed for 5 or more minutes.

Note that we are also overwriting the `nycflights` data frame with the new version of this data frame that includes the new `dep_type` variable.

We can handle all of the remaining steps in one code chunk:

```r
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

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

Exercise 6 Answer:

I would chose LGA or LaGuardia which has a 72.3% on time arrival percentage

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.

Exercise 7 Answer:

```
# This chunk was in order to view code before storing on table
#nycflights%>%
#   mutate(avg_speed = (distance/(air_time/60)))
```

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

```
# to verify storage
glimpse(nycflights)
```

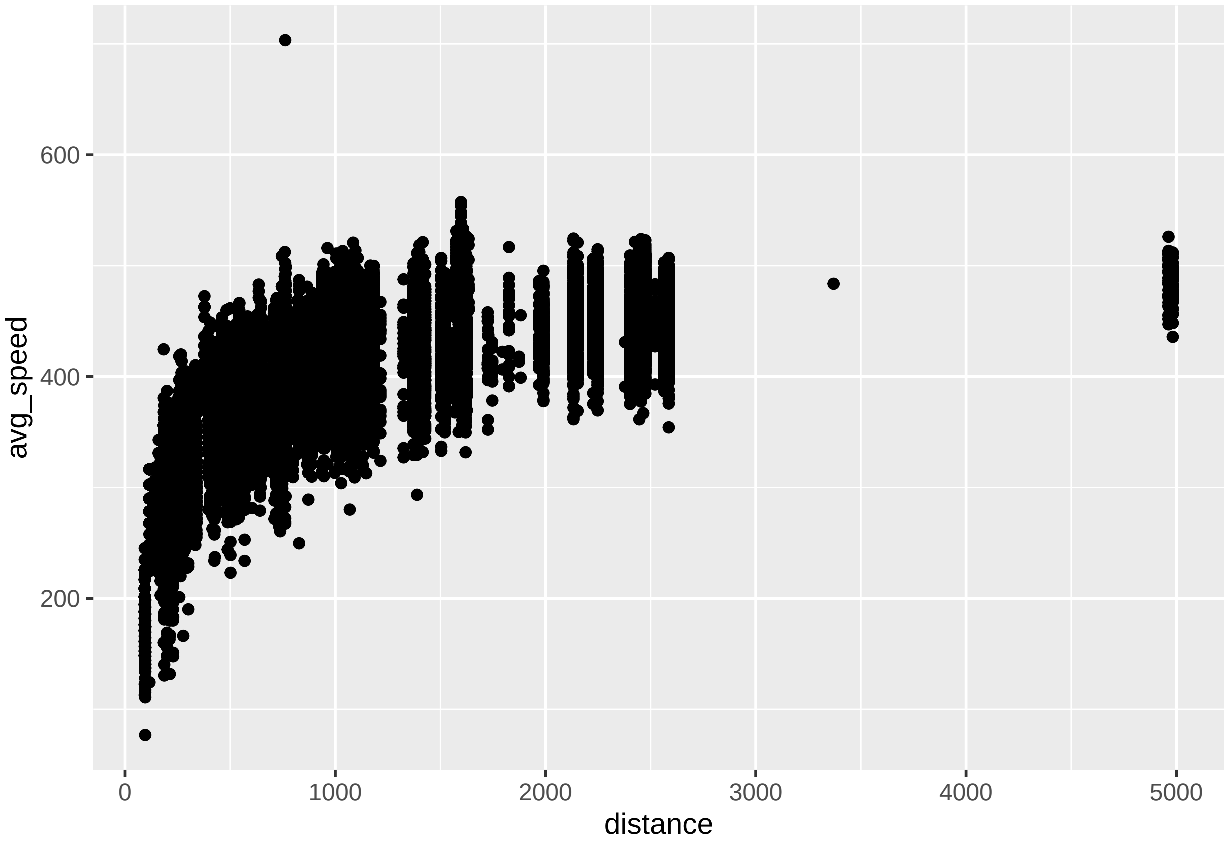
```
## Rows: 32,735
## Columns: 18
## $ year      <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 201...
## $ month     <int> 6, 5, 12, 5, 7, 1, 12, 8, 9, 4, 6, 11, 4, 3, 10, 1, 2, 8,...
## $ day       <int> 30, 7, 8, 14, 21, 1, 9, 13, 26, 30, 17, 22, 26, 25, 21, 2...
## $ dep_time  <int> 940, 1657, 859, 1841, 1102, 1817, 1259, 1920, 725, 1323, ...
## $ dep_delay <dbl> 15, -3, -1, -4, -3, -3, 14, 85, -10, 62, 5, 5, -2, 115, -...
## $ arr_time  <int> 1216, 2104, 1238, 2122, 1230, 2008, 1617, 2032, 1027, 154...
## $ arr_delay <dbl> -4, 10, 11, -34, -8, 3, 22, 71, -8, 60, -4, -2, 22, 91, -...
## $ 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, 2...
## $ origin    <chr> "JFK", "JFK", "JFK", "JFK", "LGA", "LGA", "EWR", "JFK", "...
## $ dest      <chr> "LAX", "SJU", "LAX", "TPA", "ORF", "ORD", "HOU", "IAD", "...
## $ air_time  <dbl> 313, 216, 376, 135, 50, 138, 240, 48, 148, 110, 50, 161, ...
## $ distance  <dbl> 2475, 1598, 2475, 1005, 296, 733, 1411, 228, 1096, 820, 2...
## $ hour      <dbl> 9, 16, 8, 18, 11, 18, 12, 19, 7, 13, 9, 13, 8, 20, 12, 20...
## $ minute    <dbl> 40, 57, 59, 41, 2, 17, 59, 20, 25, 23, 40, 20, 9, 54, 17,...
## $ dep_type  <chr> "delayed", "on time", "on time", "on time", "on time", "o...
## $ avg_speed <dbl> 474.4409, 443.8889, 394.9468, 446.6667, 355.2000, 318.695...
```

Note the last variable is avg_speed

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, mapping = aes(x = distance,y = avg_speed)) + geom_point()
```



Exercise 8 Answers

There is a positive association with avg_speed and distance

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.

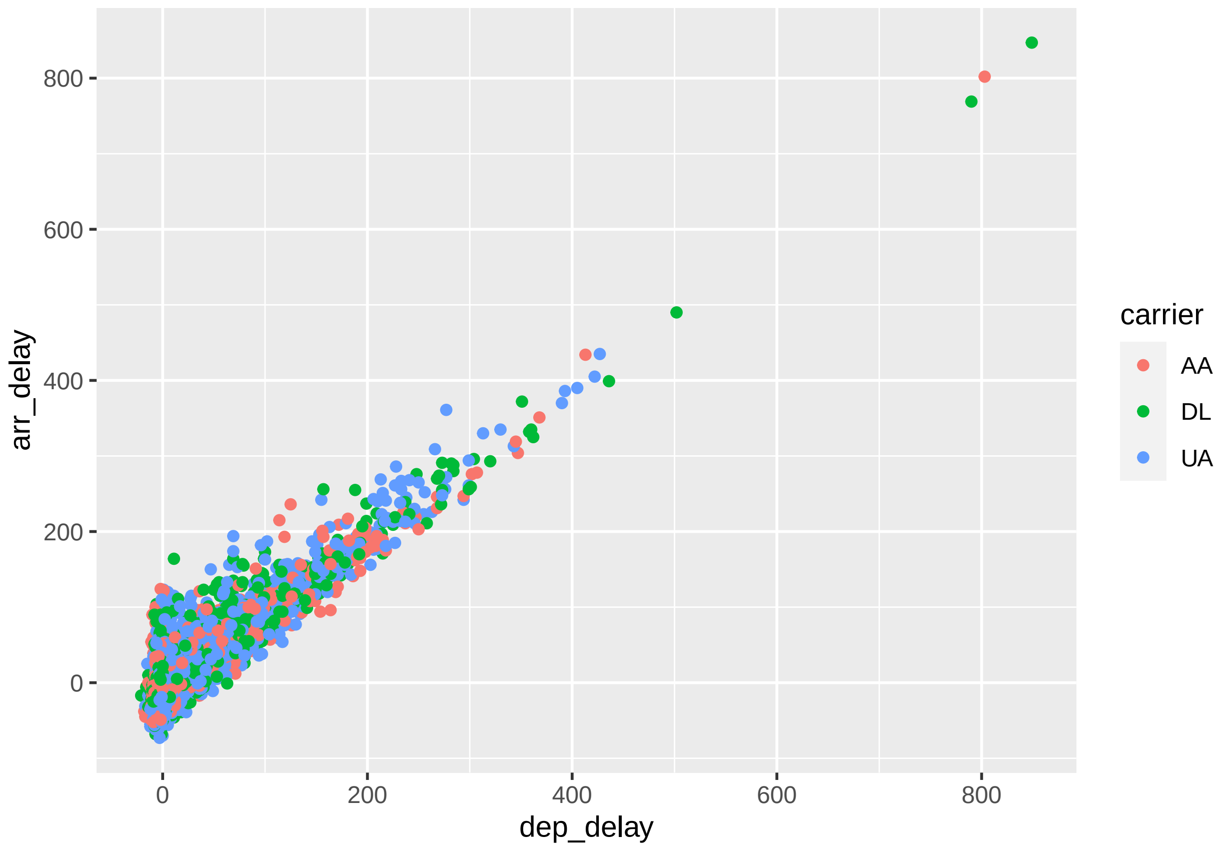
Exercise 9 Answer

First create the data frame with the appropriate data

```
ex9_data<-nycflights%>%
    filter(carrier=="AA"|carrier=="DL"|carrier=="UA")
```

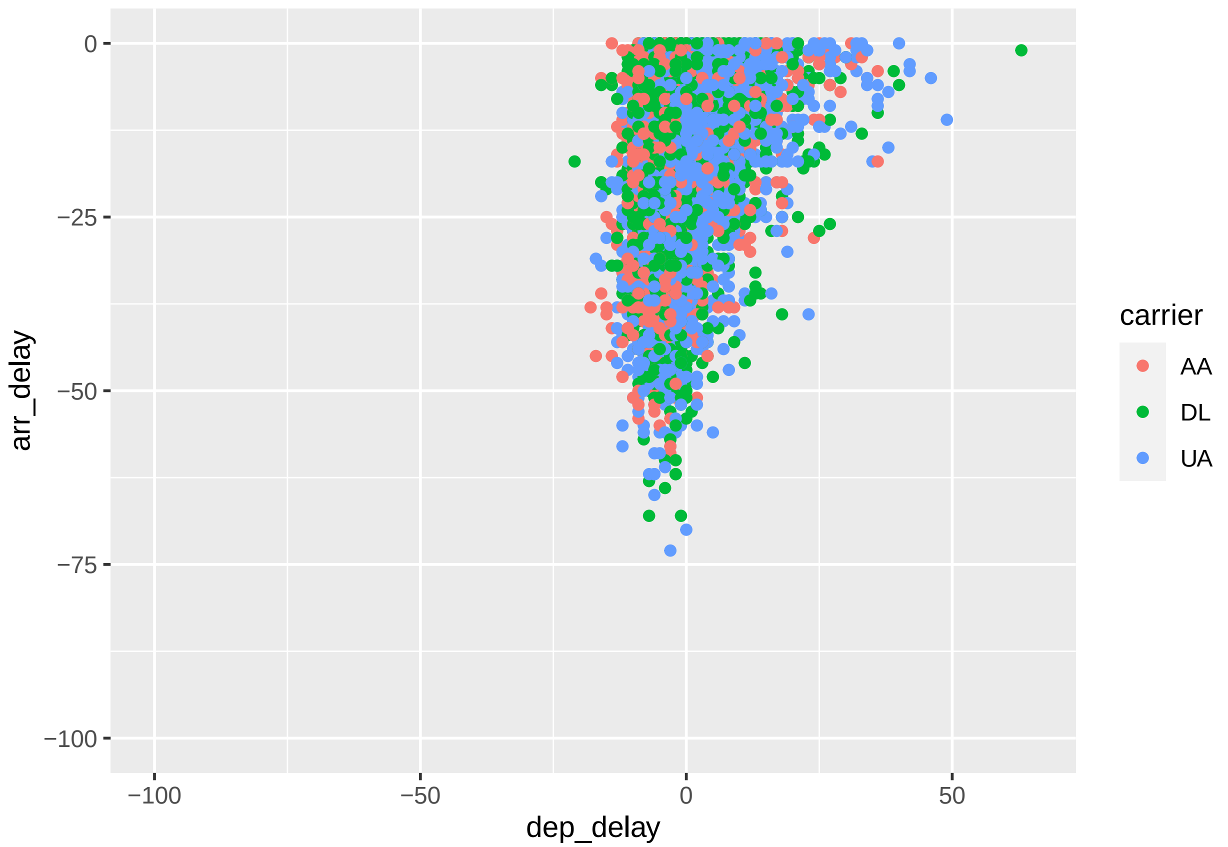
then create the scatter plot with color configuration using dep_delay and arr_delay based on graph provided.

```
ggplot(data  = ex9_data, mapping = aes(x = dep_delay, y = arr_delay, color = carrier))+ geom_point()
```



```
ggplot(data = ex9_data, aes(group = 1, x = dep_delay, y = arr_delay, color = carrier))+ geom_point() +
```

```
## Warning: Removed 5001 rows containing missing values (geom_point).
```



```
ex9_data%>%
    summarise(max_dep = max(dep_delay[arr_delay==0]))
```

```
## # A tibble: 1 x 1
##   max_dep
##     <dbl>
## 1      40
```

Exercise 9 Answers Part II

Assessing the latest a person can depart and still arrive on time can be viewed as largest departure delay value where the arrival delay value equals 0. Therefore I set the app_delay limit value between -100 to 0. From there I decreased the x-limit value down to 65 (as decreasing any more would make my largest plot point disappear).
To check my result I also searched the largest departure delay where arrival delay equals 0 using the max function.
I found that (roughly) the cutoff point for departure is between **40-65** minutes.

. . .