Lab - 2

Irene Jacob

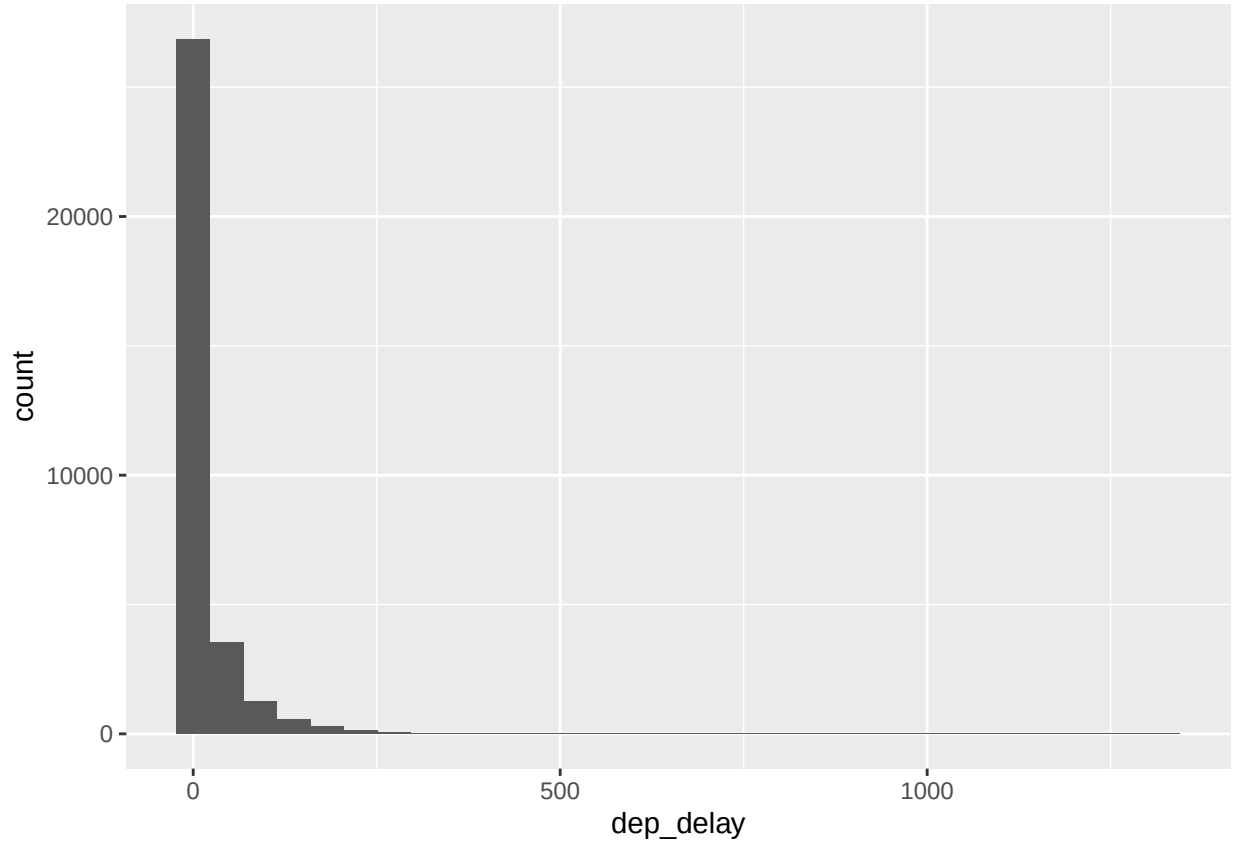
2020-09-06

Exercise 1

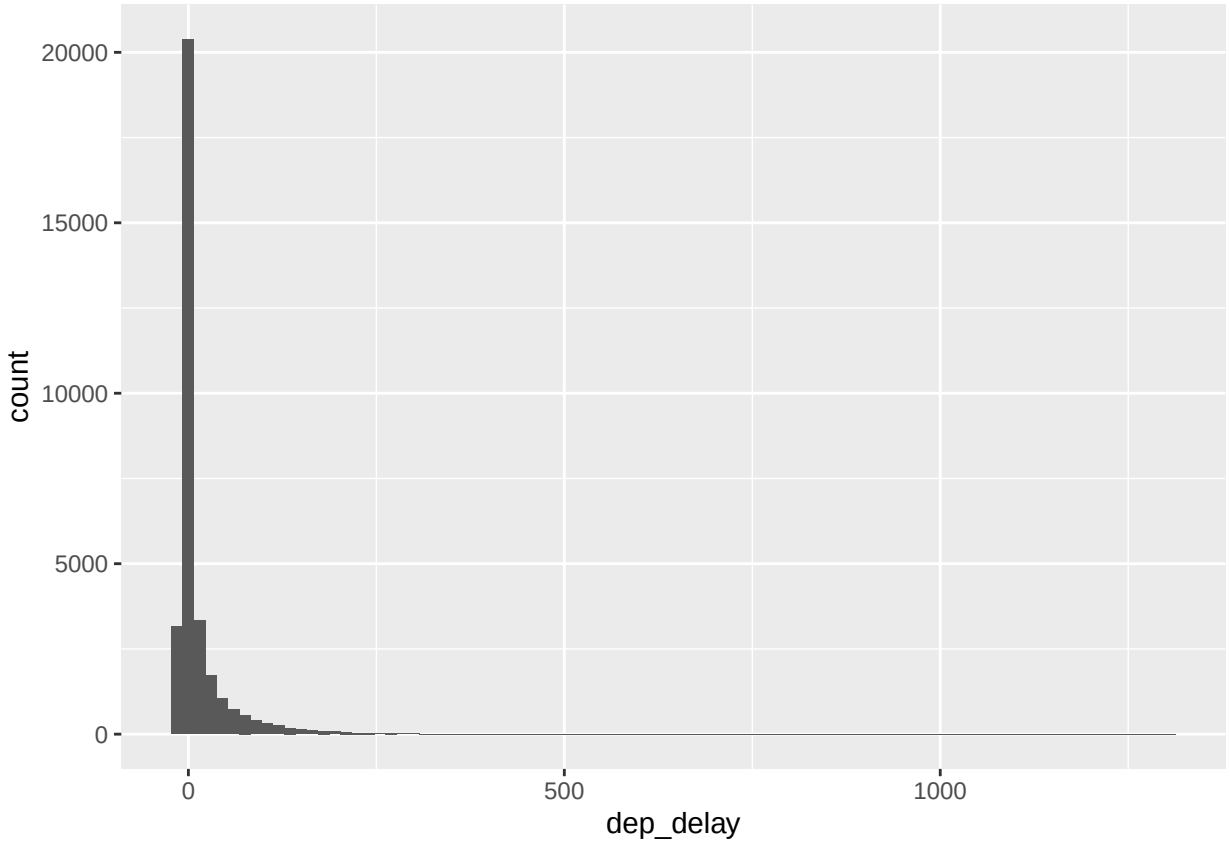
What can be noticed from the 3 histograms given below is that as the bin width increases the accuracy of the data represented decreases. When the bin width is 15 the accuracy level is maximum.

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

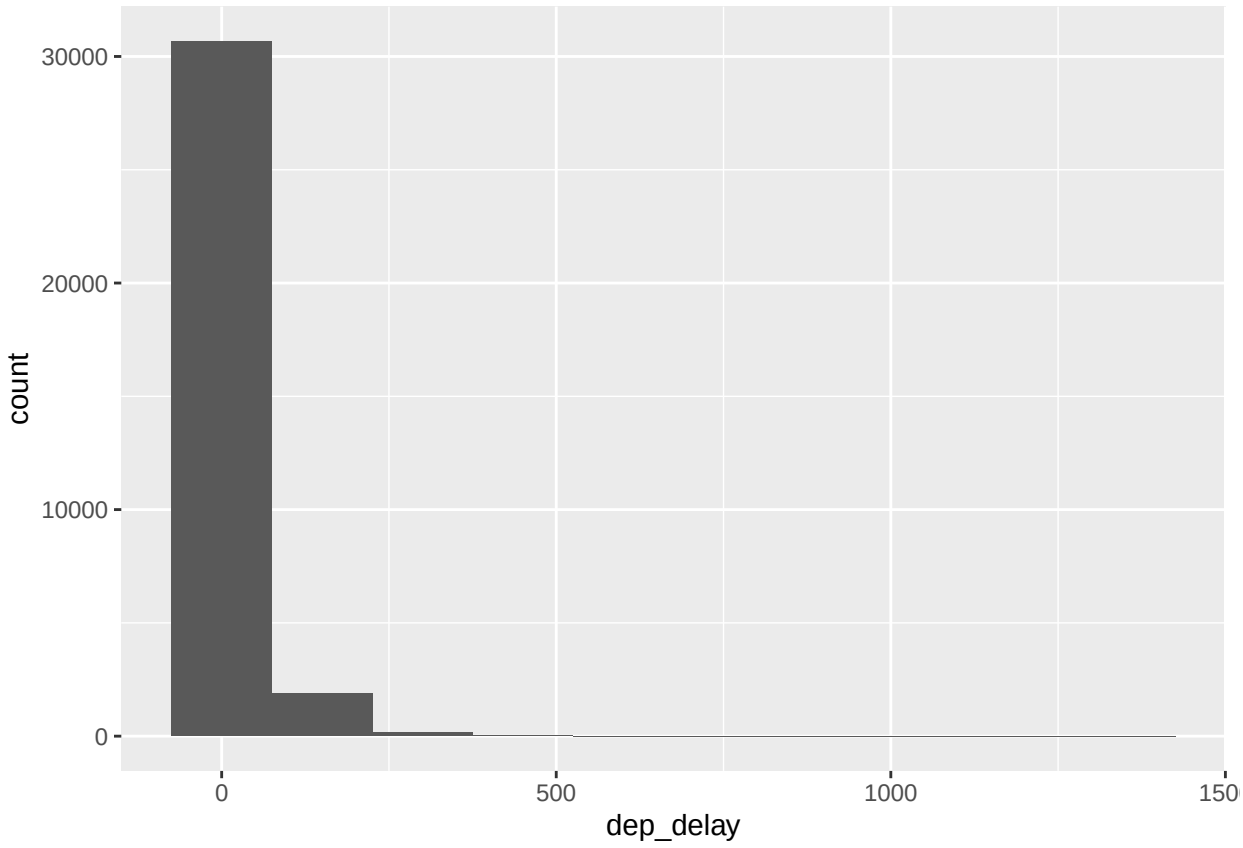
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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




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


30000
20000
10000
0
count
0
500
1000
150
dep_delay

Exercise 2

All the flights that were headed to SFO in February are as follows:

```
sfo_feb_flights <- nycflights %>%
  filter(dest == "SFO", month == 2)
print(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
## # ... with 58 more rows, and 7 more variables: flight <int>, origin <chr>,
## #   dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>
```

The total number of flights meeting the given criteria are:

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

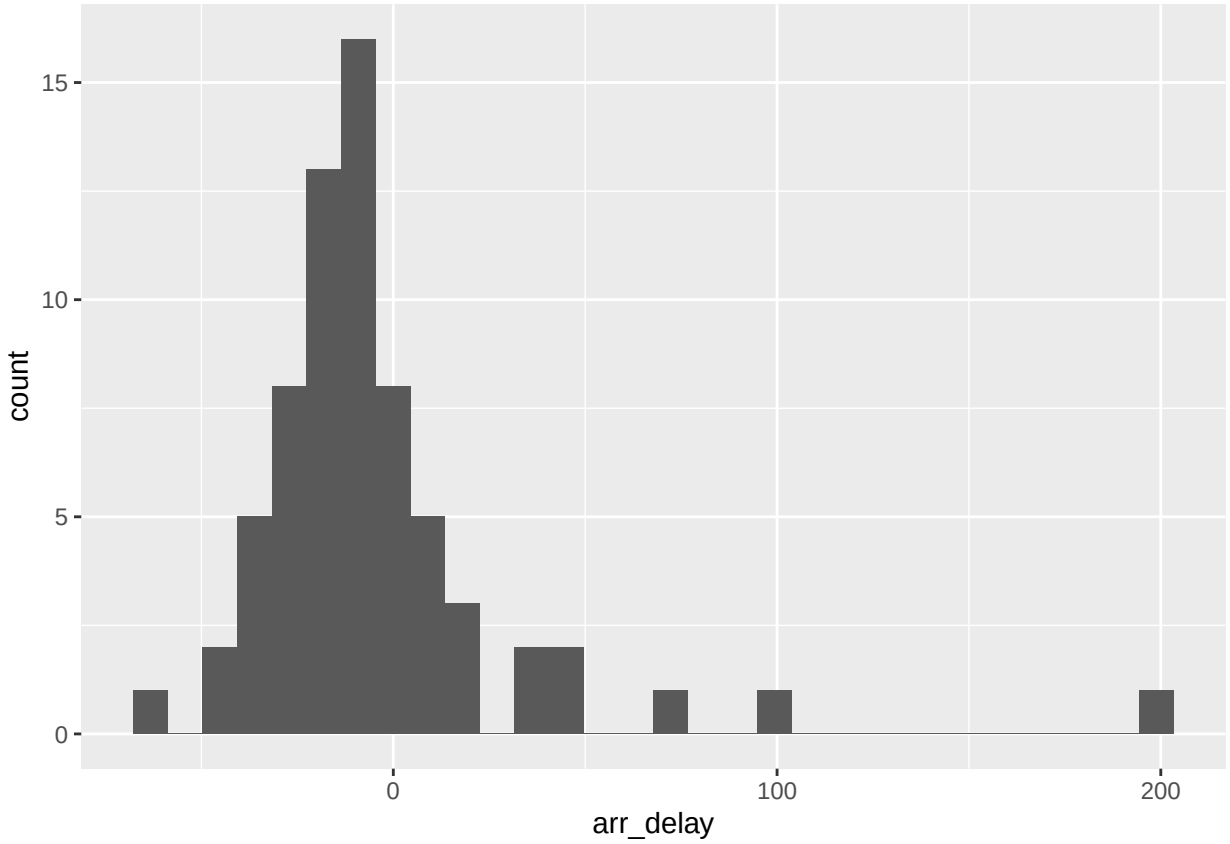
```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    68
```

Exercise 3

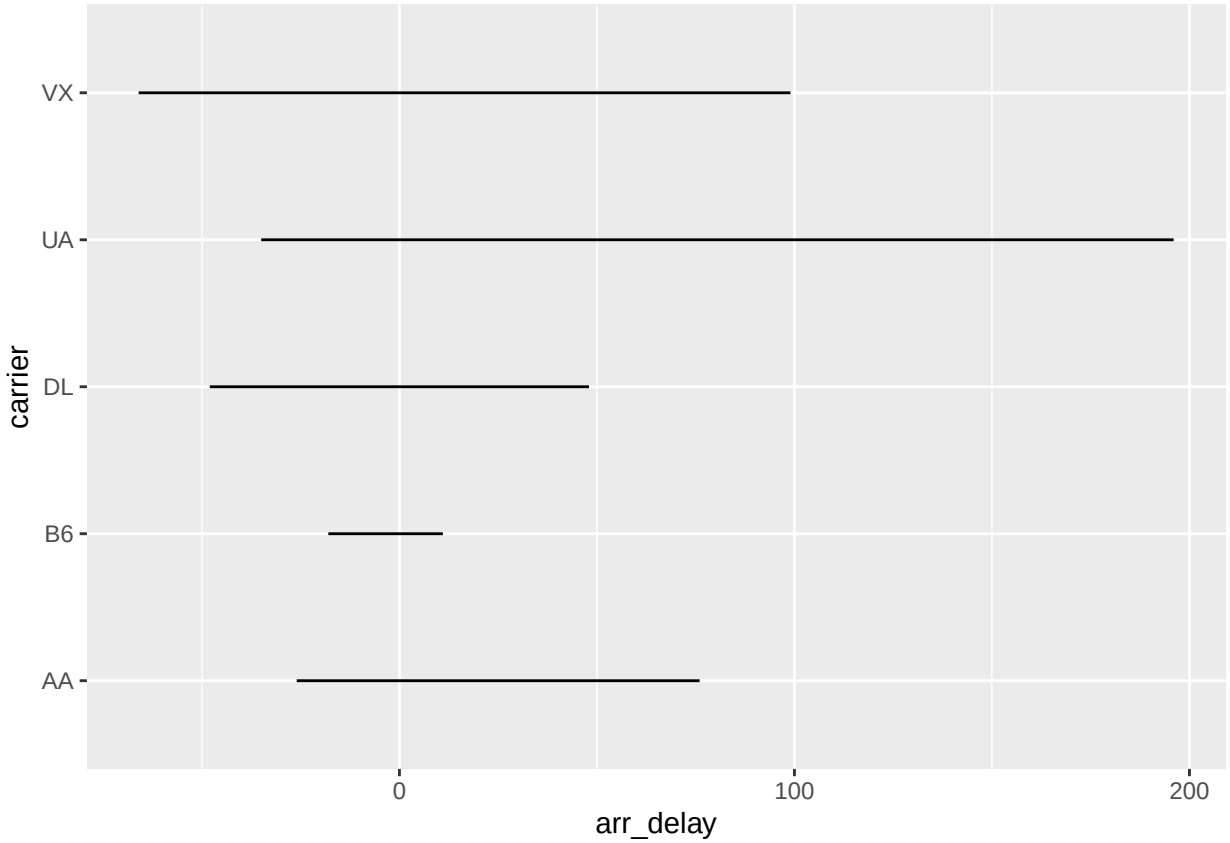
From the below statistics it is clear that "United Air Lines Inc." has the highest delays and "Virgin America" has a reputation of arriving early.

```
ggplot(data = sfo_feb_flights, mapping = aes(arr_delay)) +
  geom_histogram()
```

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




```
ggplot(data = sfo_feb_flights, mapping = aes(arr_delay, carrier)) +
  geom_line()
```


VX
UA
DL
B6
AA
carrier
0
100
200
arr_delay

Exercise 4

From the below result it is clear that Delta Airlines and United Airlines has the most variable arrival delays as they have the highest IQR values.

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

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

```
## # 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
```

Exercise 5

Median is a strong method when compared to mean. That being said the outliers seen in the histograms would make it not that powerful in this case. So this suggests that mean is a better option here.

Exercise 6

From the below result I would choose LGA to fly out of.

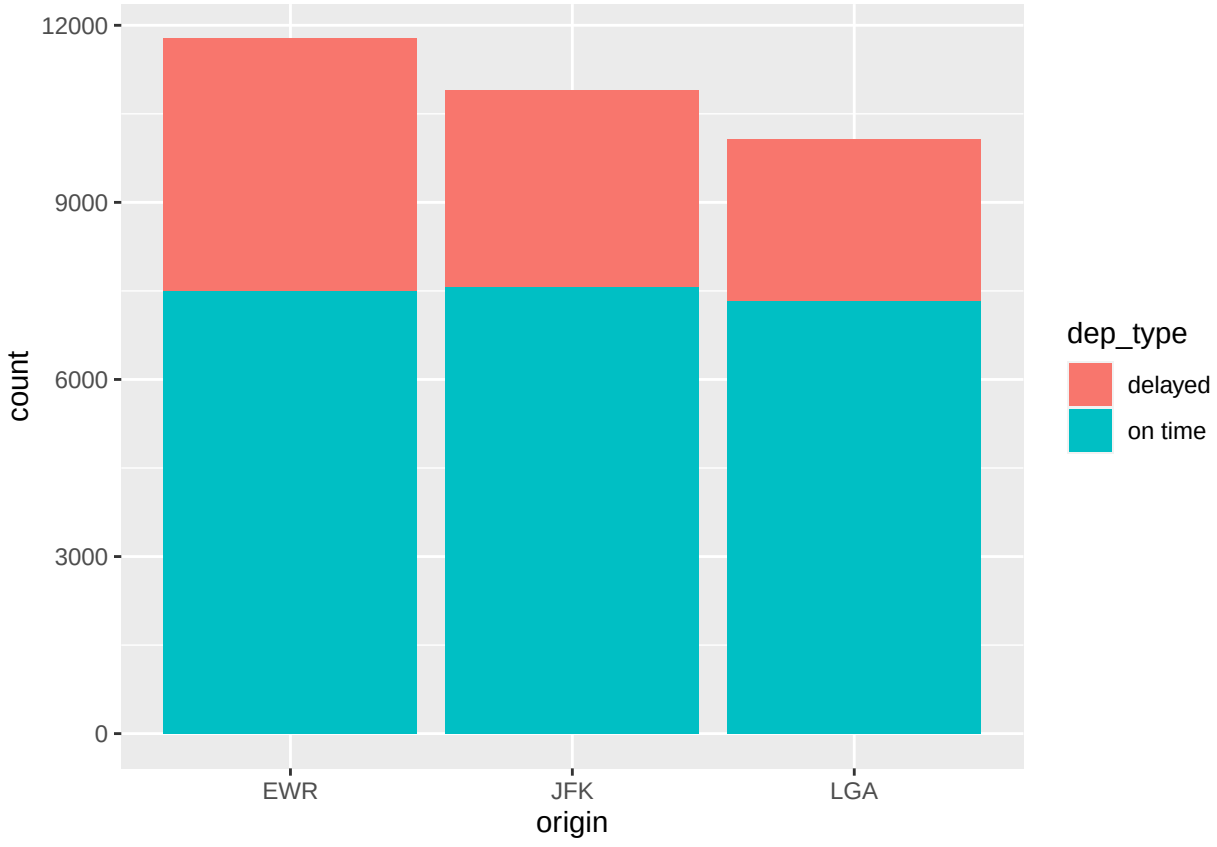
```
nycflights <- nycflights %>%
  mutate(dep_type = ifelse(dep_delay < 5, "on time", "delayed"))

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
```

The visuals of the ontime flights from the 3 airports are as follows:

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

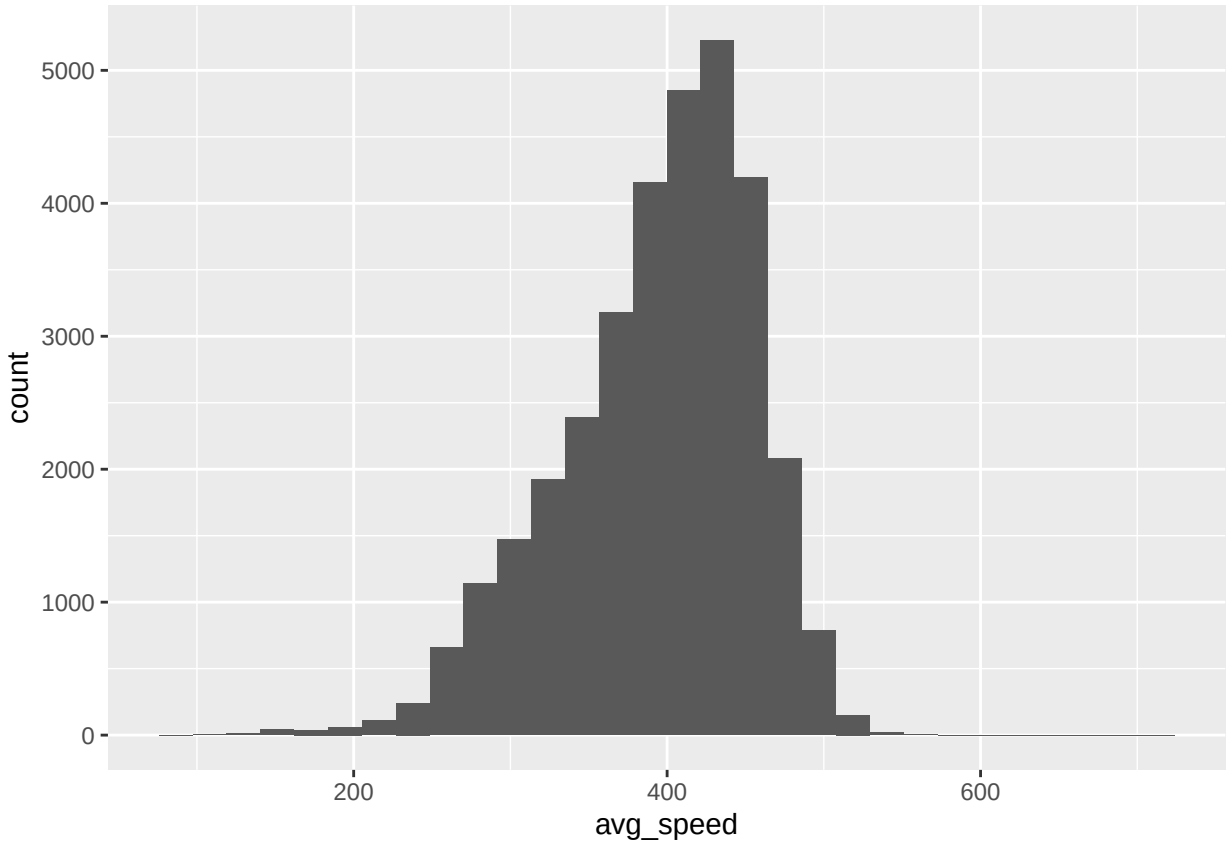


Exercise 7

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

Histogram of avg_speed is as follows:

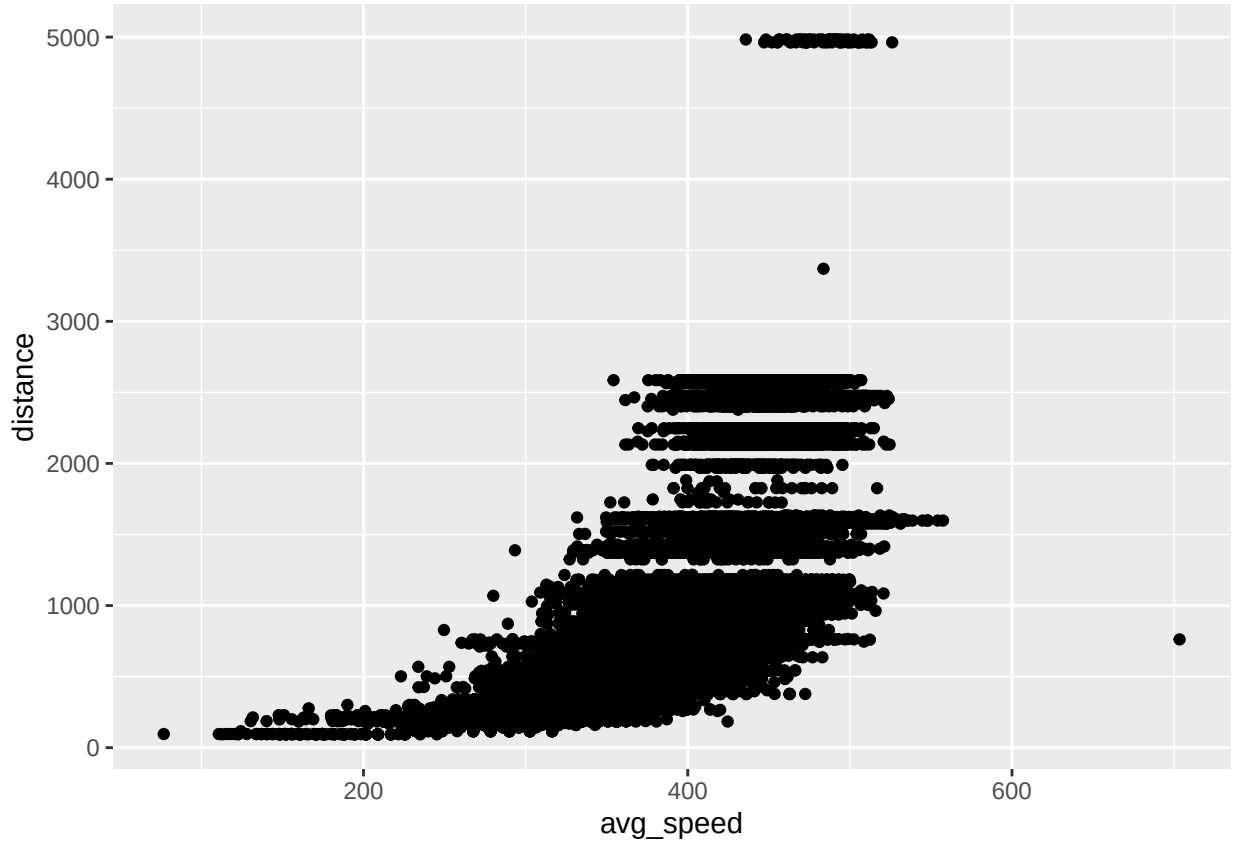
```
ggplot(data = nycflights, aes(avg_speed)) +
  geom_histogram()
```



Exercise 8

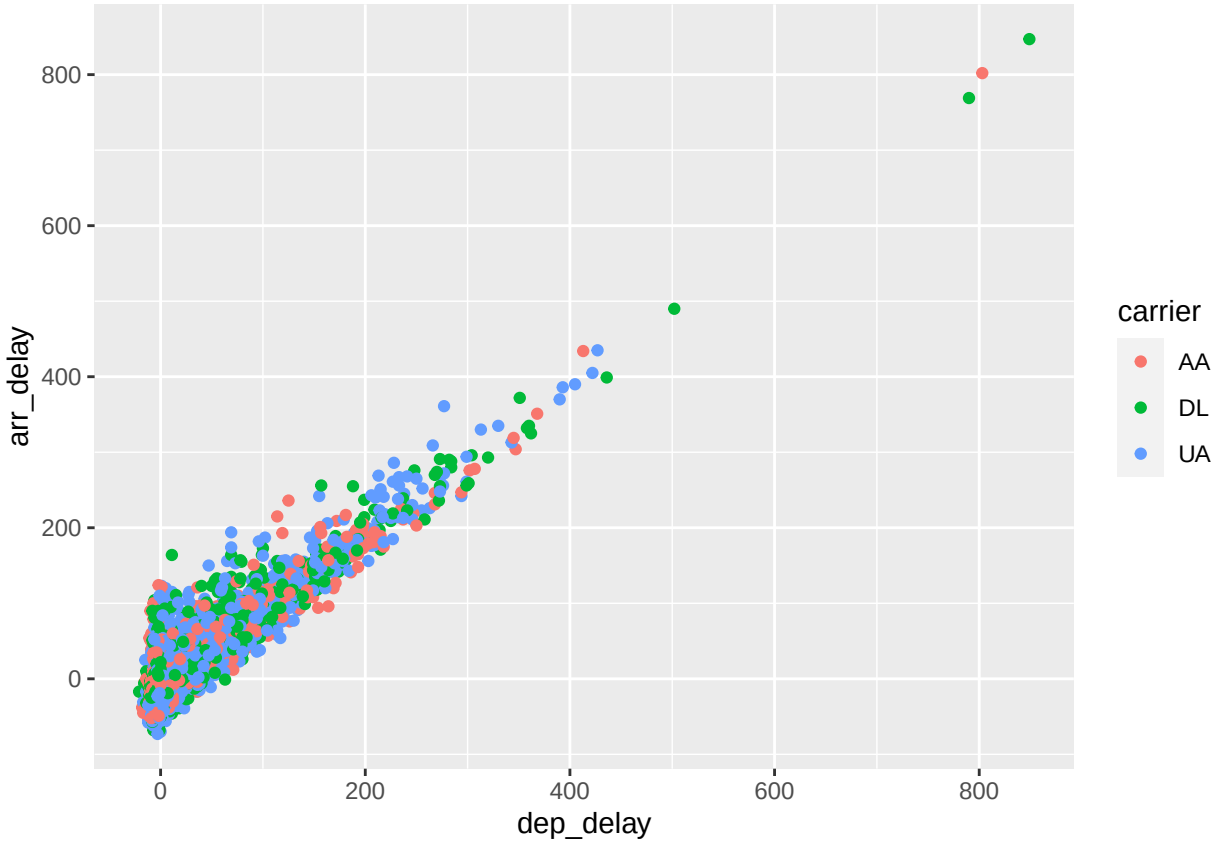
From the below scatterplot it can be seen that the average speed changes a little with the increase in distance.

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



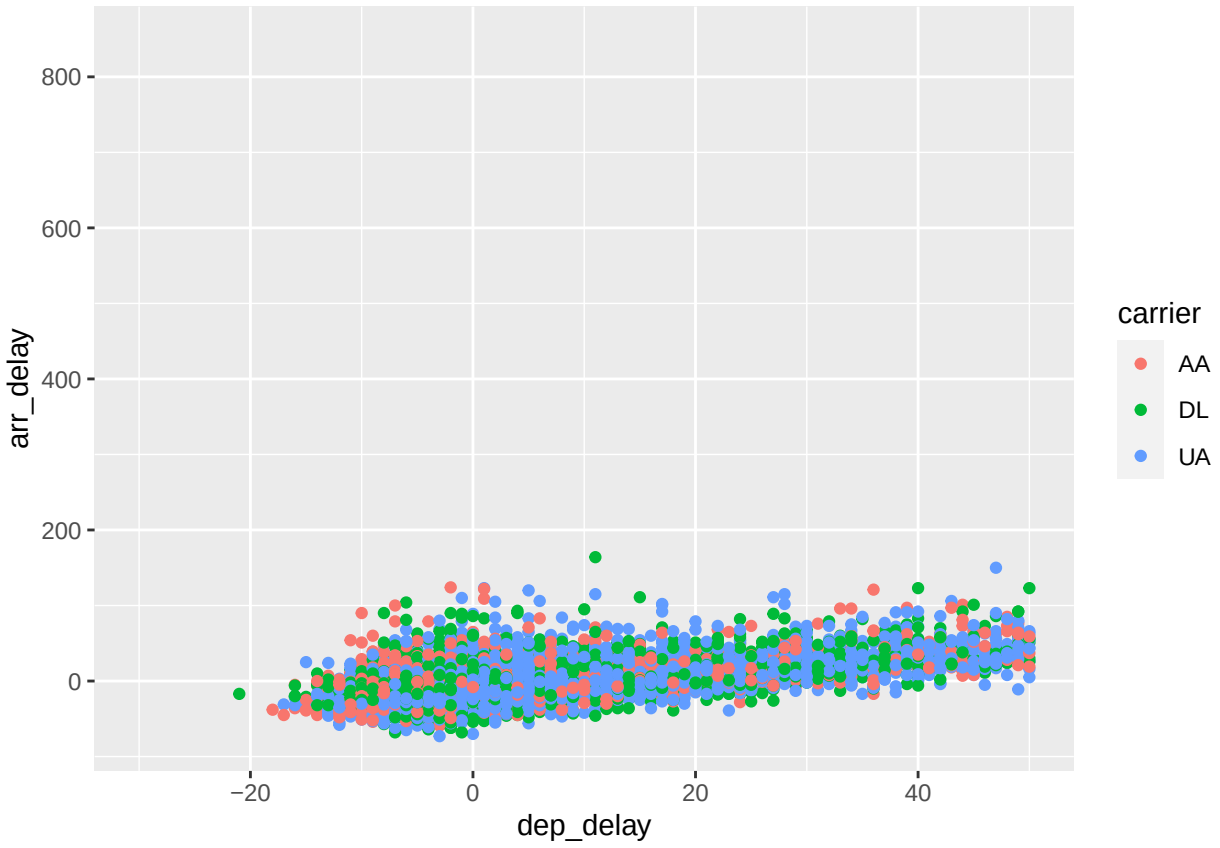
Exercise 9

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




```
ggplot(nycflights, aes(dep_delay, arr_delay, color = carrier )) +
  xlim(-30, 50) +
  geom_point()
```

```
## Warning: Removed 1033 rows containing missing values (geom_point).
```



The cutoff point is for departure delays where you can still expect to get to your destination on time is around 20 to 25 minutes.This can be understood better from the above plot.