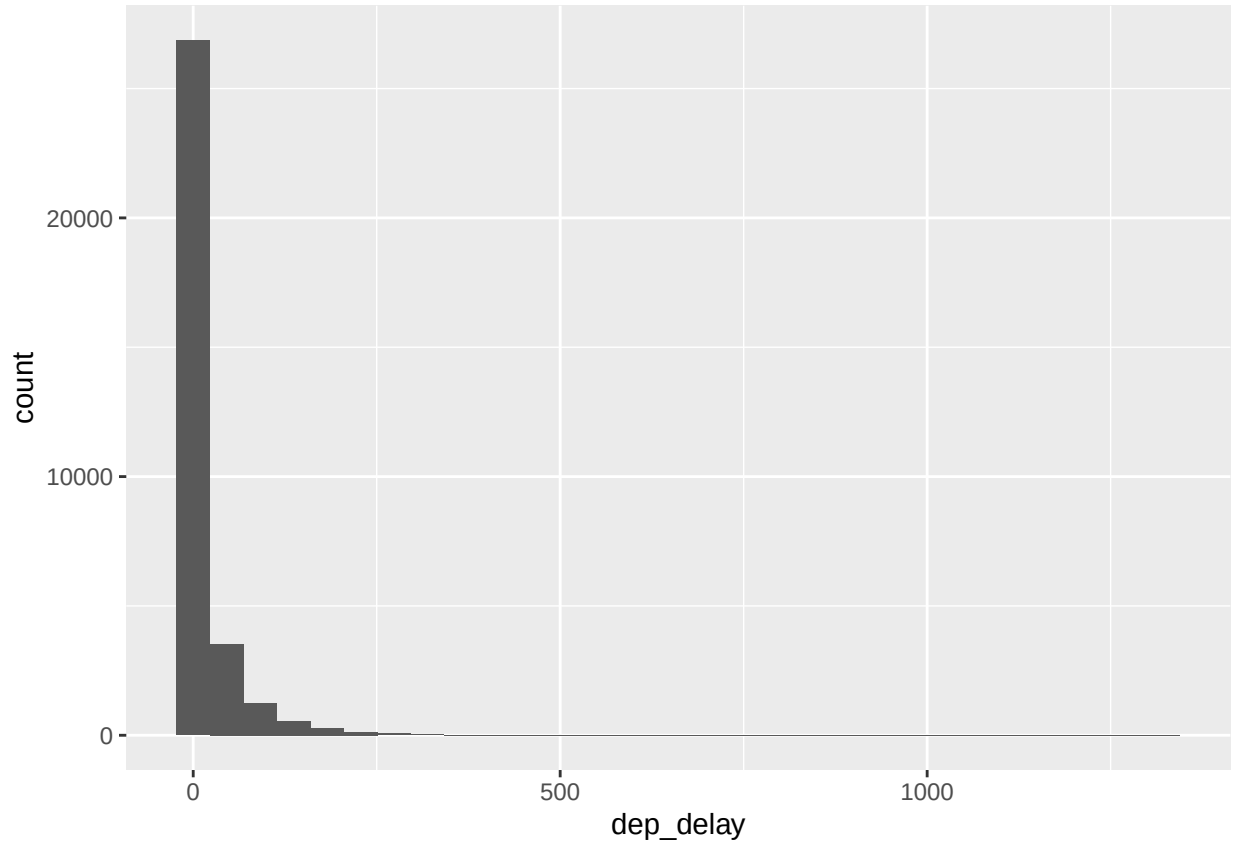
Lab 3 - Exploratory Data Analysis Part II Solution

MATH224 - Intro to Stat

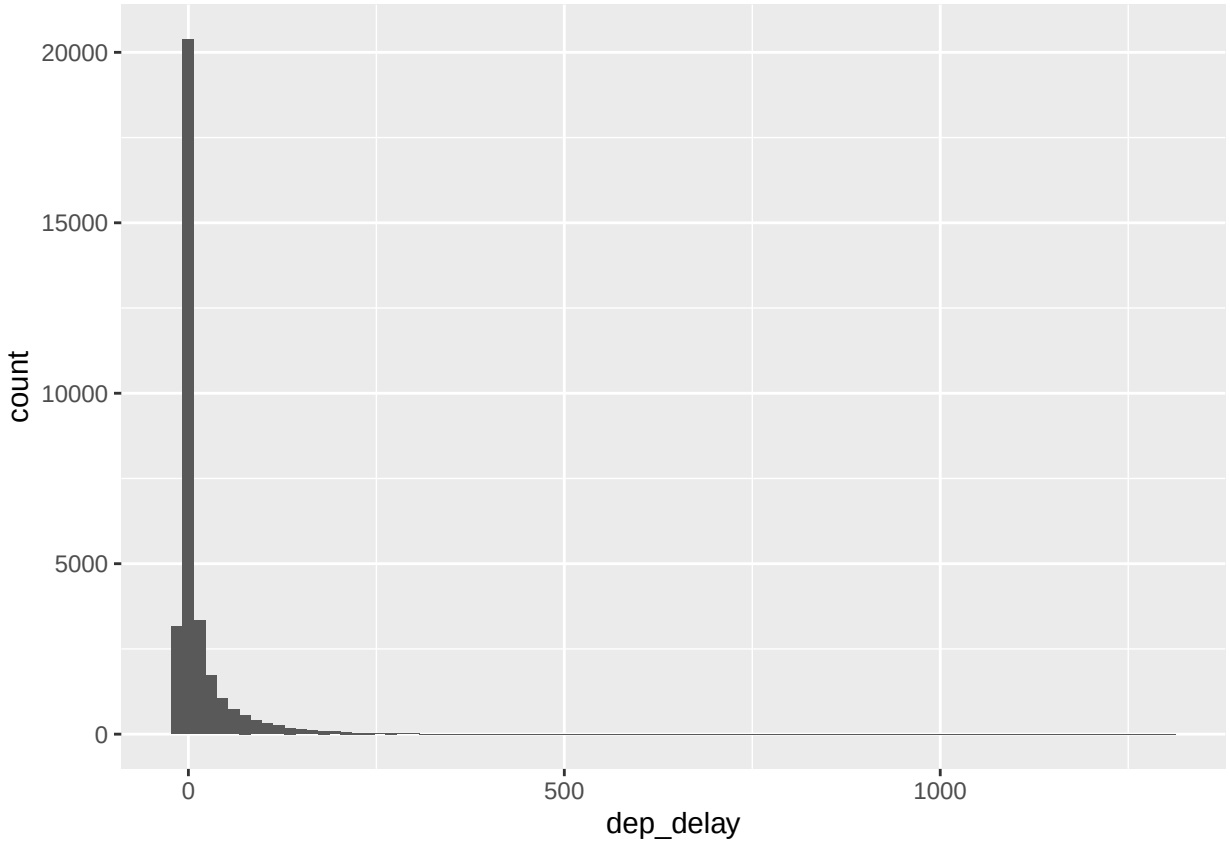
Exercise 1 (3 points)

2 points Explanation: We can see that as we use a smaller binwidth, we see too much of the data whereas a bigger binwidth doesn't provide the vital information about the data.

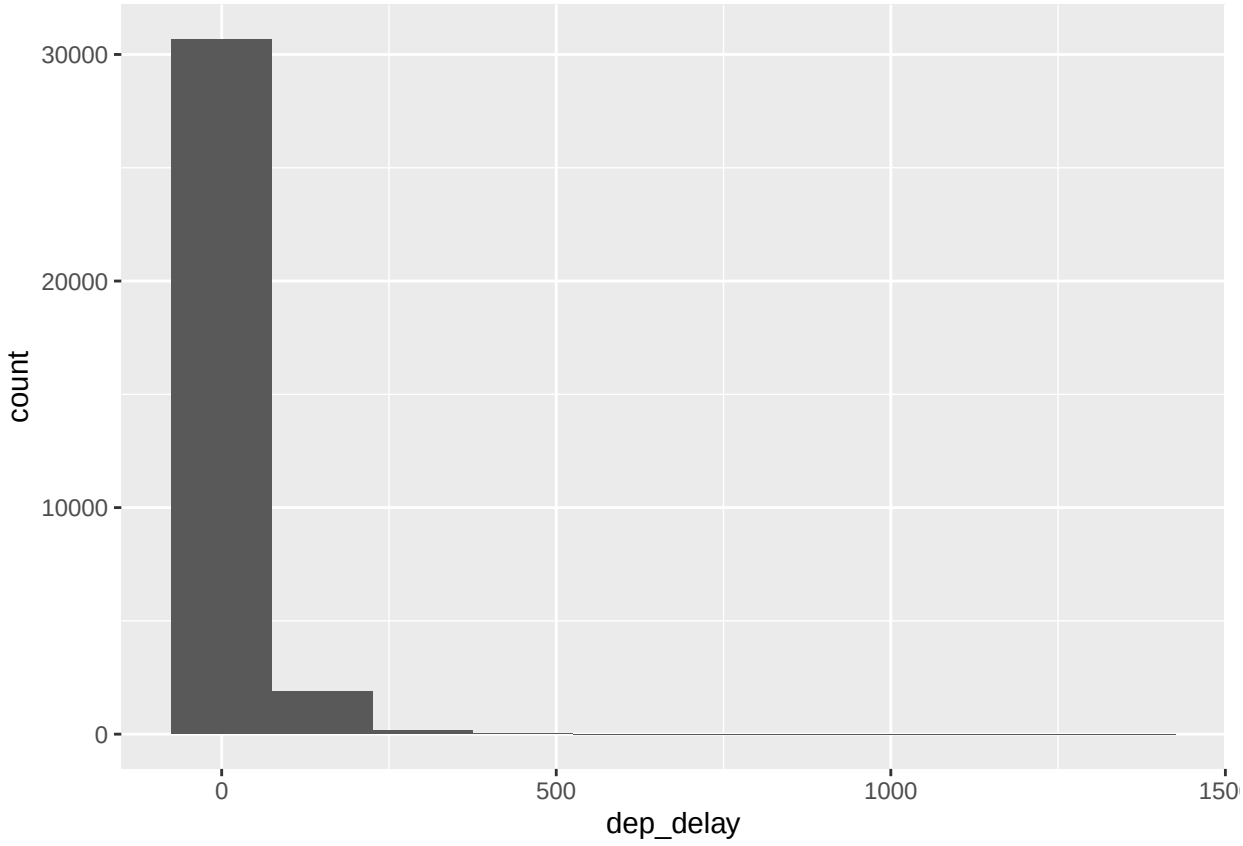
```
#0.34 points
ggplot(data = nycflights, aes(x = dep_delay)) +
  geom_histogram()
```




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




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



Exercise 2 (1 Point)

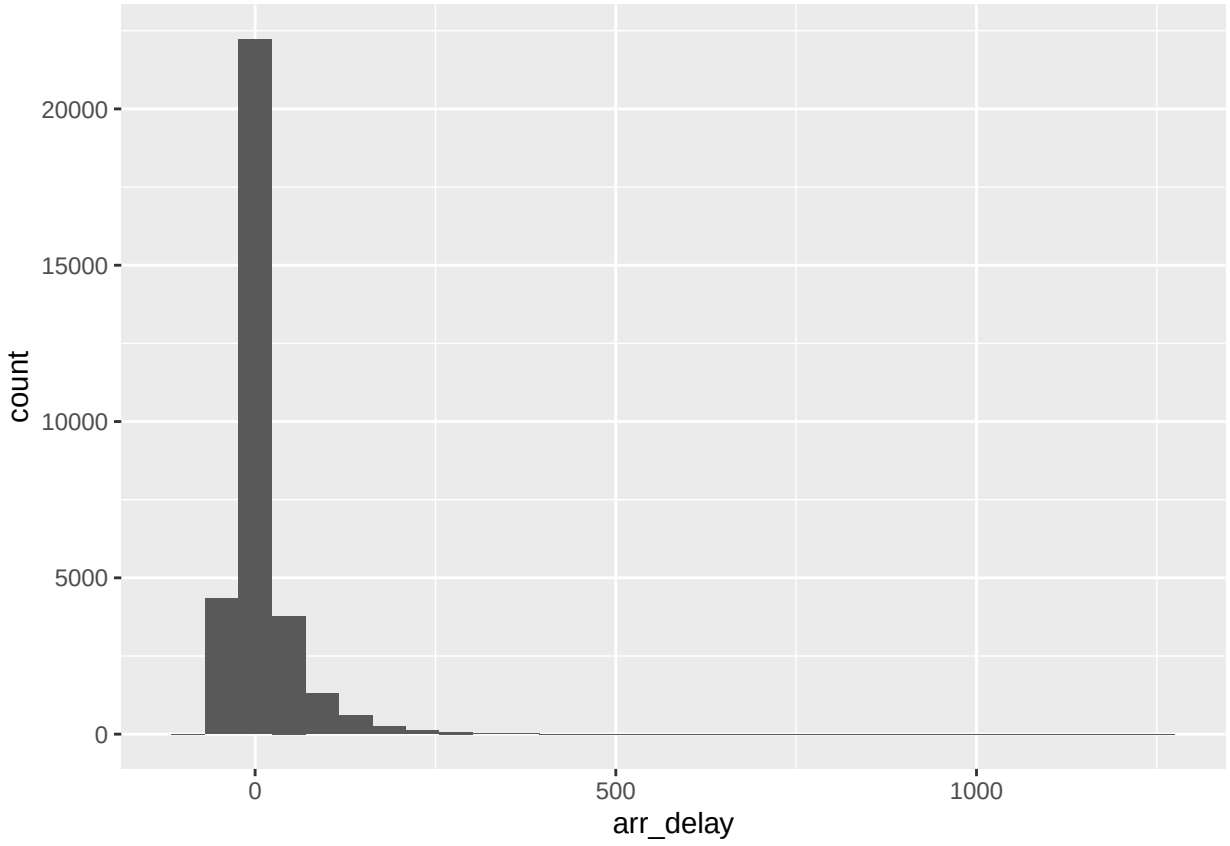
```
sfo_feb_flights <- nycflights %>%
  filter(dest == "SFO", month == 2)
```

Exercise 3 (4 points)

Explanation:

2 Points From the histogram plot, we can see that the arrival delay has a distribution of data that is right skewed. Since the data is right skewed, we will be using median as a summary statistic as it gives us a better understanding of the center of the data because mean is sensitive to outliers.

```
#1 Point
ggplot(data = nycflights, aes(x = arr_delay)) +
  geom_histogram()
```




```
#1 Point
nycflights%>%
  summarise(mean_ad = mean(arr_delay),
            median_ad = median(arr_delay))
```

```
## # A tibble: 1 x 2
##   mean_ad median_ad
##     <dbl>     <dbl>
## 1    7.10        -5
```

Exercise 4 (4 Points)

2 Points Answer: As we know, IQR can be used to understand the variability of data. From the table below, we can see that carriers DL and UA have the highest IQR with an arrival delay of 22 minutes.

```
sfo_feb_flights %>%
  group_by(carrier) %>% #1 Point
  summarise(median_ad = median(arr_delay), #0.5 Point
            iqr_ad = IQR(arr_delay), #0.5 Point
            n_flights = n())
```

```
## # 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
```

Exercise 5 (4 Points)

1 Point We will choose October if we were to consider mean as the summary statistic. We will either choose September or October if we were to use median as they both have a median departure delay of -3 minutes.

2 Points The pros of choosing median is that it gives a better estimate of the center of departure delay as the variable is right skewed. Thus, this makes it a cons for using mean for this particular data.

```
#1 Point (Either finding mean or median or both)
nycflights %>%
  group_by(month) %>%
  summarise(mean_dd = mean(dep_delay)) %>%
  arrange(mean_dd)
```

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

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

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

```
##  9     8        -1
## 10     6         0
## 11     7         0
## 12    12         1
```

Exercise 6 (4 Points)

2 Points We will choose LGA to fly out of as it has a 72.79% of on time departures.

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

#1 Point
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
```

More Practice

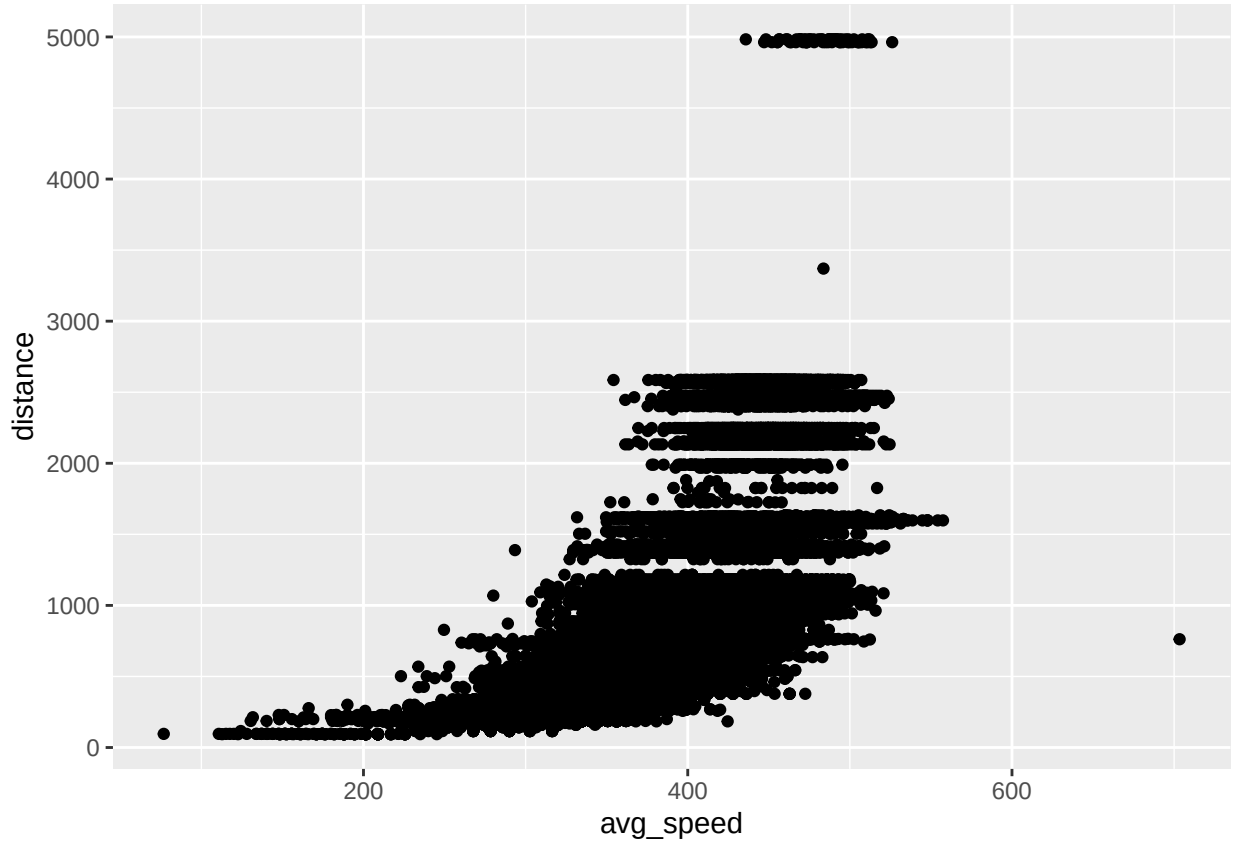
Exercise 7

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

Exercise 8

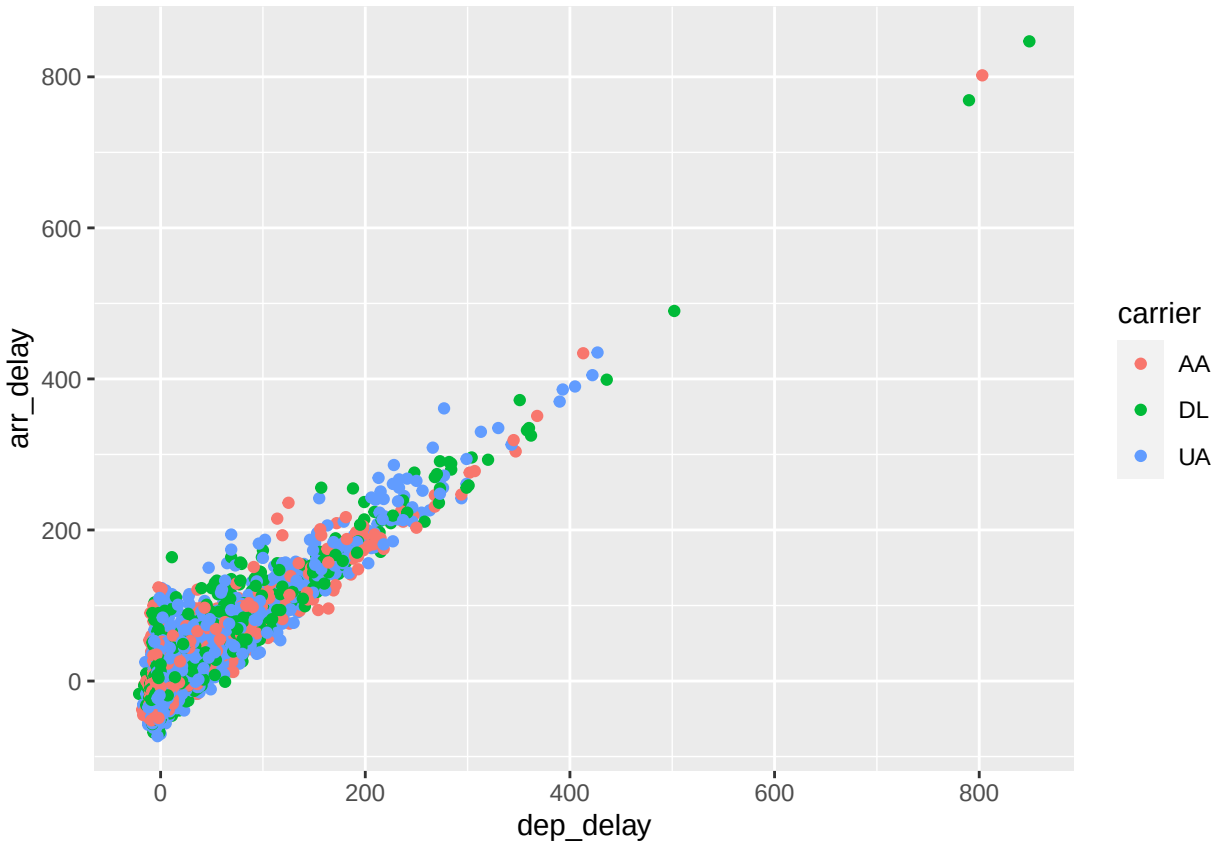
The scatter plot resembles a quadratic relationship between average speed and distance. But average speed doesn't seem to exceed 600 mph except for one outlier.

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



Exercise 9

```
nycflights %>%
  filter(carrier %in% c("AA", "DL", "UA"))%>%
  ggplot(aes(x = dep_delay, y = arr_delay, col = carrier))+
  geom_point()
```


800
600
400
200
0
arr_delay
0
200
400
600
800
dep_delay
carrier
AA
DL
UA