Code ▾

Lab 2

Kevin White

1/26/2023

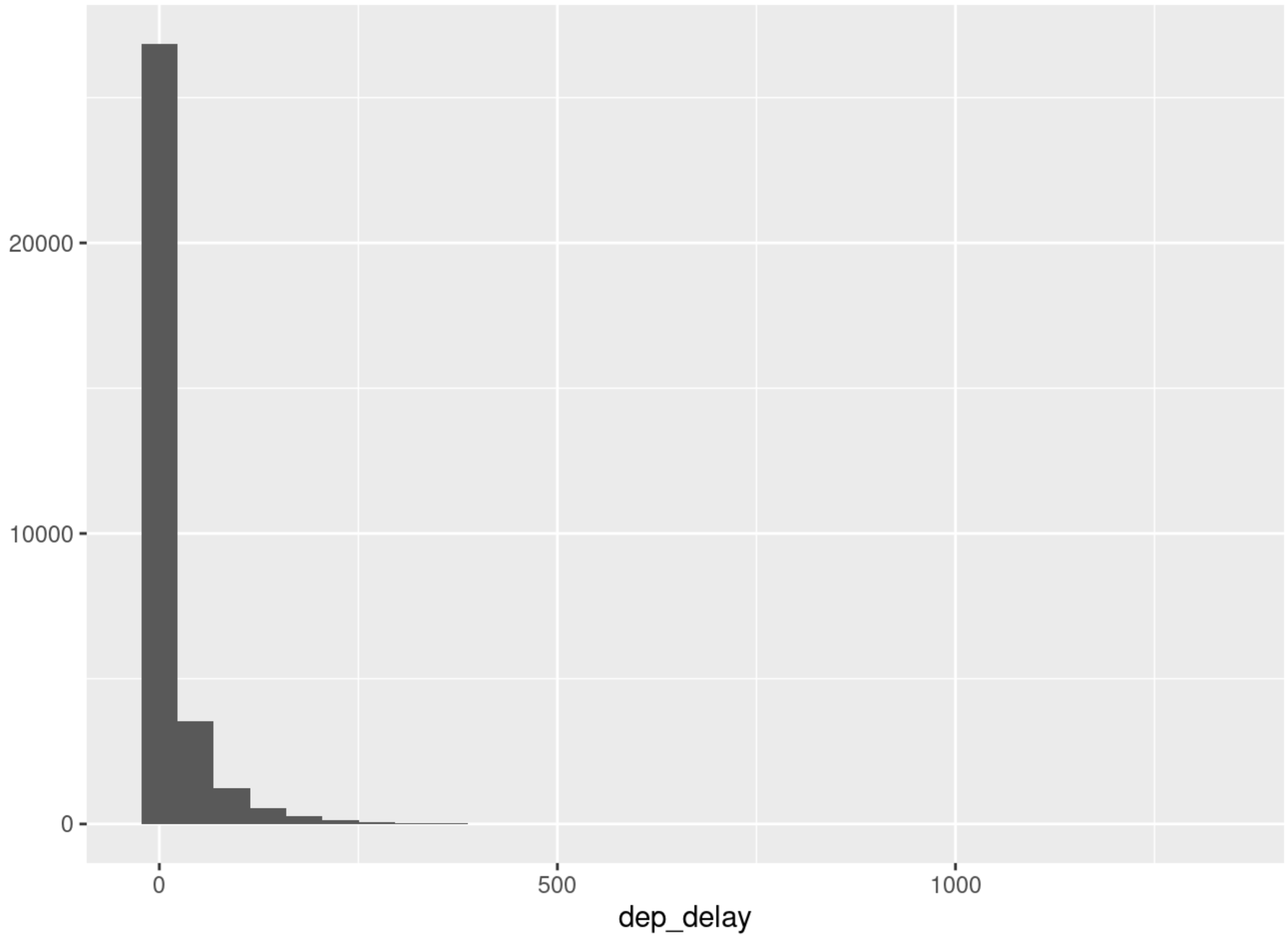
Exercise 1:

Hide

```
qplot(x = dep_delay, data = nycflights, geom = "histogram")
```

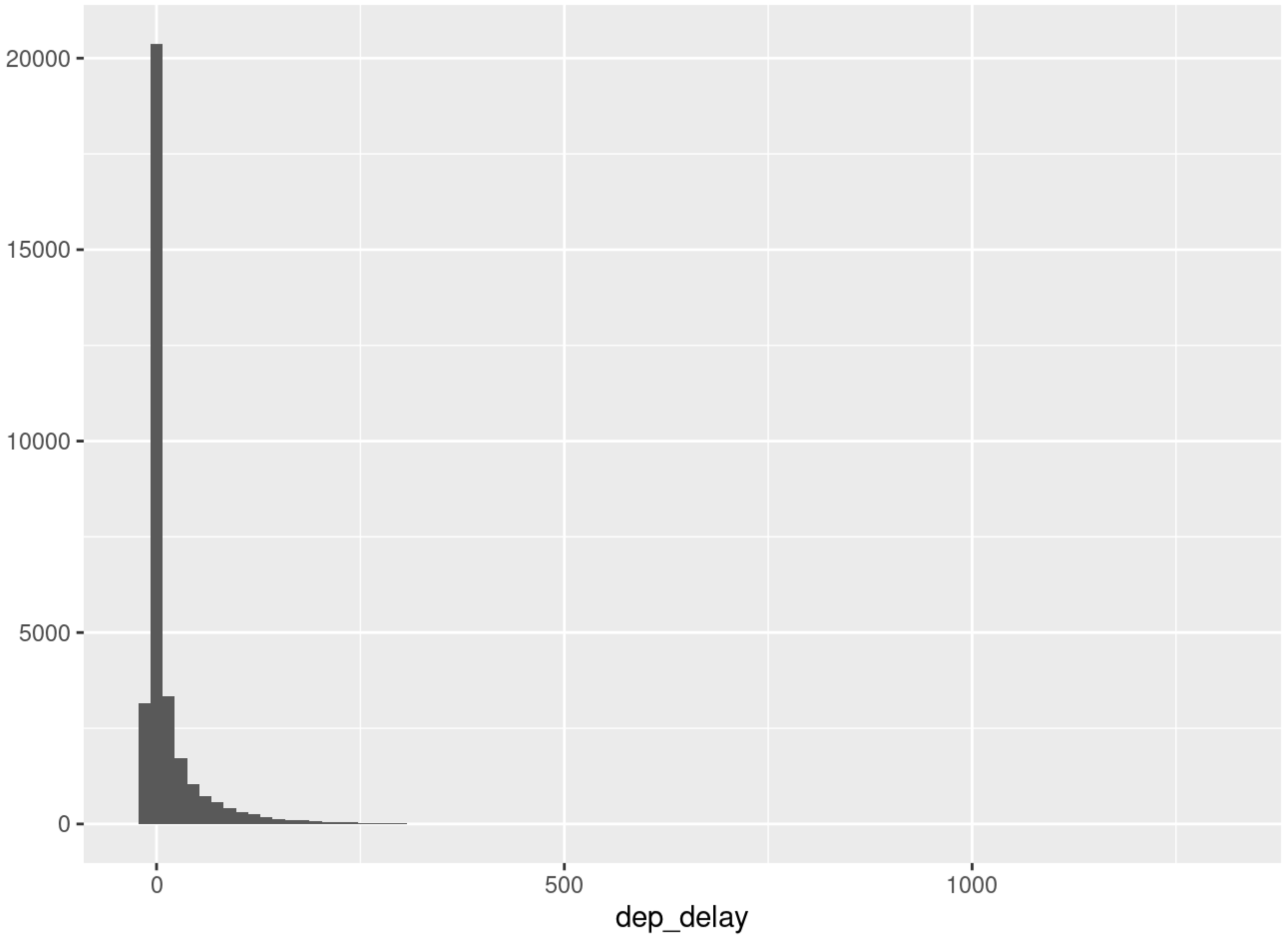
```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
```

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



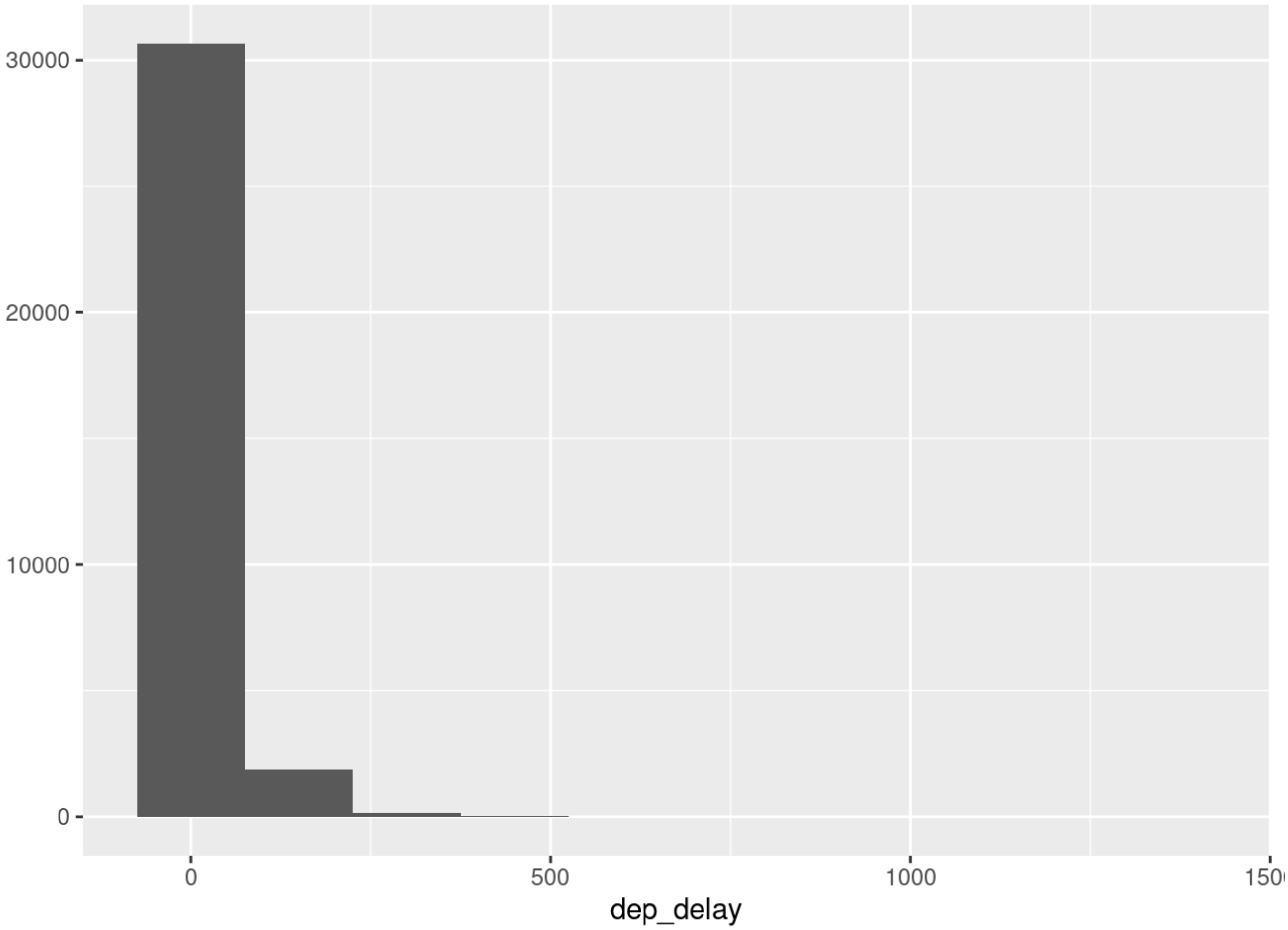
Hide

```
qplot(x = dep_delay, data = nycflights, geom = "histogram", binwidth = 15)
```



Hide

```
qplot(x = dep_delay, data = nycflights, geom = "histogram", binwidth = 150)
```



The first and 3rd histograms do not show a wide verity of data while the second hisogram with a block size of 15 we can see more information about the data.

Exercise 2:

Hide

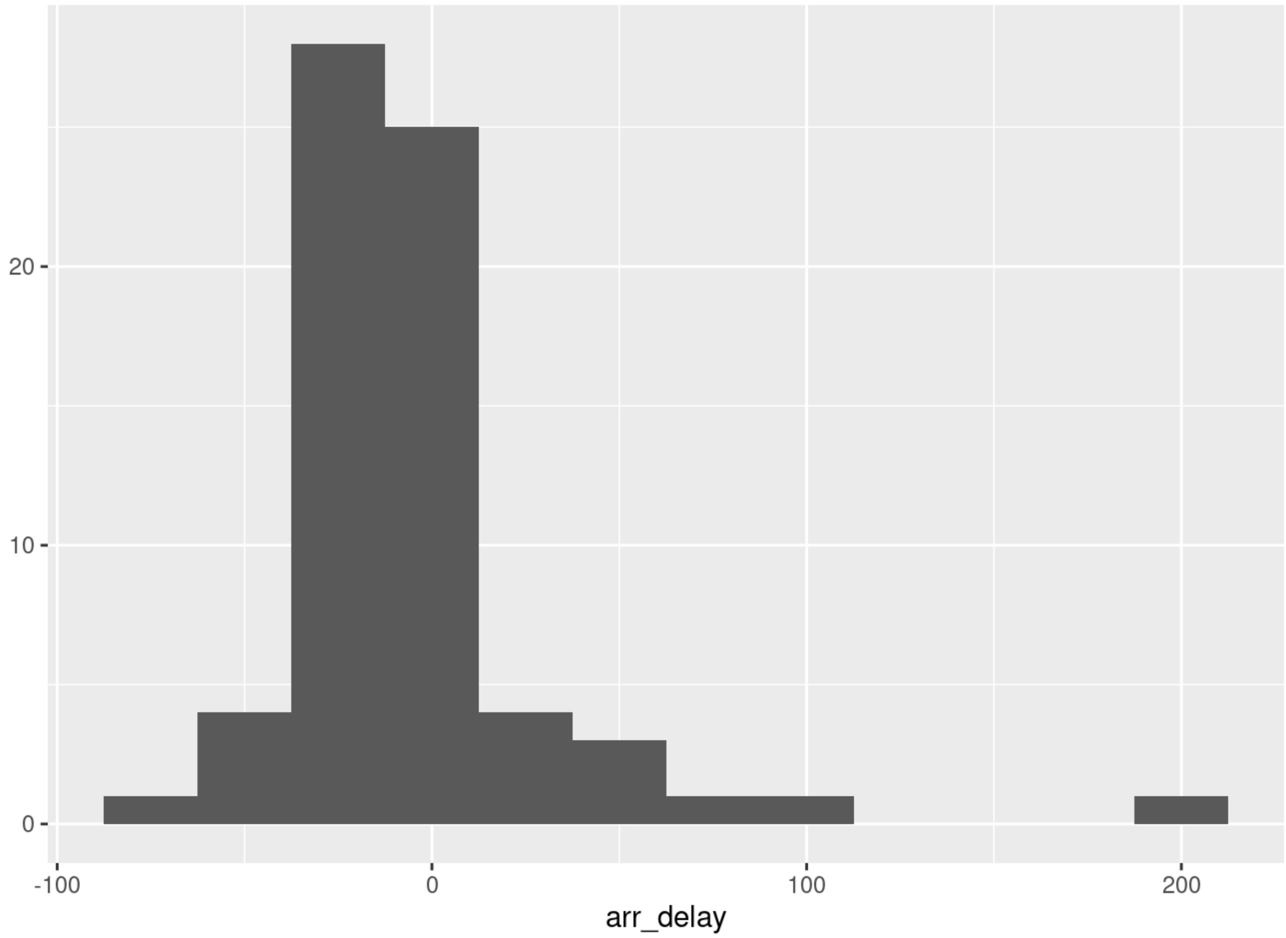
```
  sfo_feb_flights <- nycflights %>%
  filter(dest == "SFO", month == 2)
```

68 Flights fit this criteria

Exercise 3:

Hide

```
qplot(x = arr_delay, data = sfo_feb_flights, geom = "histogram", binwidth = 25)
```



Many of the Flights arive on time or early, with only a few landing late, and a rare condition of a few landing up to 200 min late

Exercise 4:

Hide

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

```
## # A tibble: 5 × 4
##   carrier median_apd iqr_apd 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
```

The Carrier DL and AU have the most variable arrival delays at 22.

Exercise 5:

Hide

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

```
## # A tibble: 12 × 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
```

Hide

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

```
## # A tibble: 12 × 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
```

Mean Pro: Because is is the average of ALL the data, all data is represented in this number Con: Because all data is represented is is more easily skewed by outliers than median

Median: Pro: It is less likey to be skewed by outliers because it takes the middle of the data set Con: it does not give as well representation of the data as a whole.

Exercise 6:

Hide

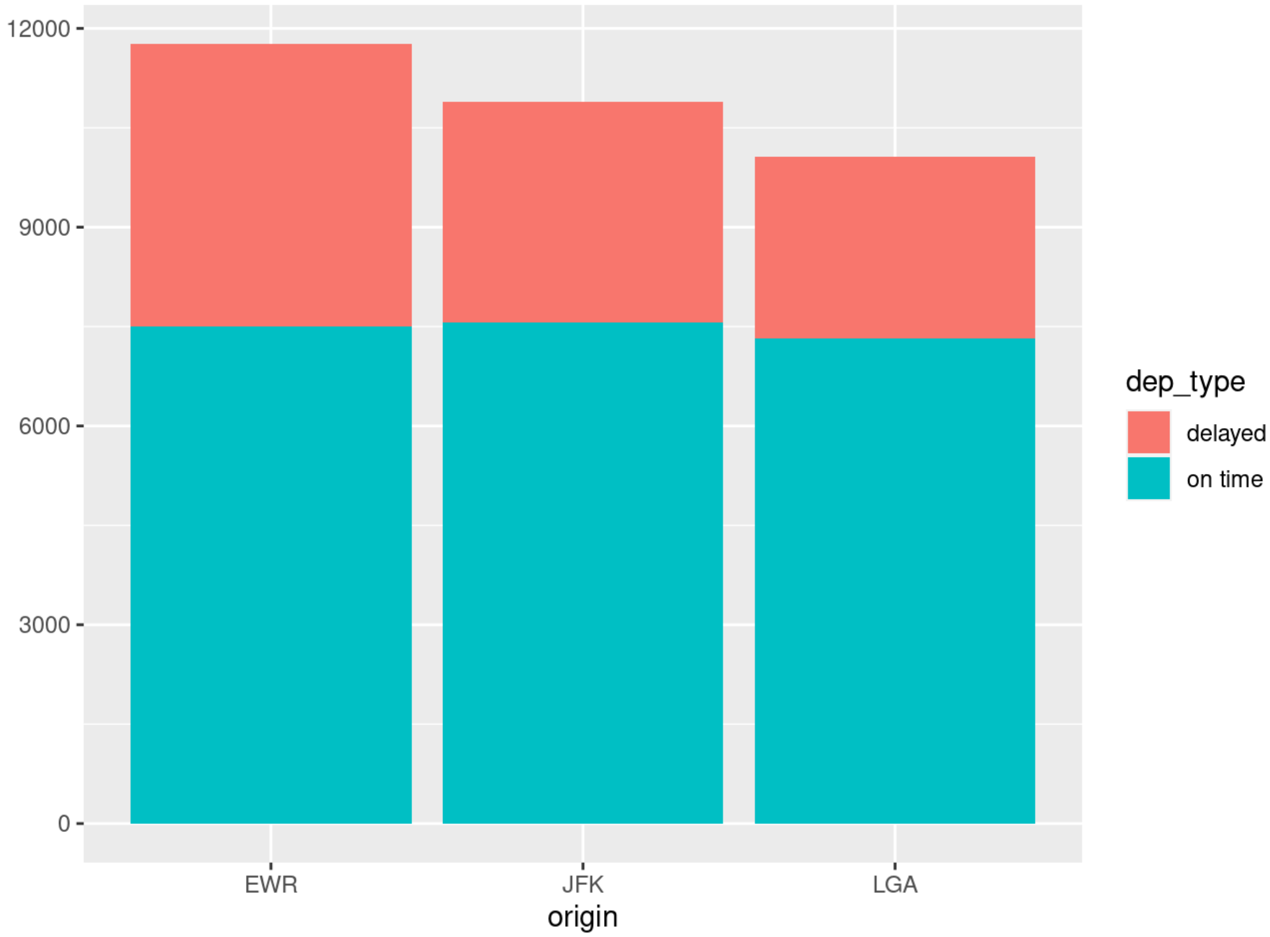
```
nycflights <- nycflights %>%
  mutate(dep_type = ifelse(dep_delay < 5, "on time", "delayed"))

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

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

Hide

```
qplot(x = origin, fill = dep_type, data = nycflights, geom = "bar")
```



It would be best to fly out of LGA airport

Exercise 7:

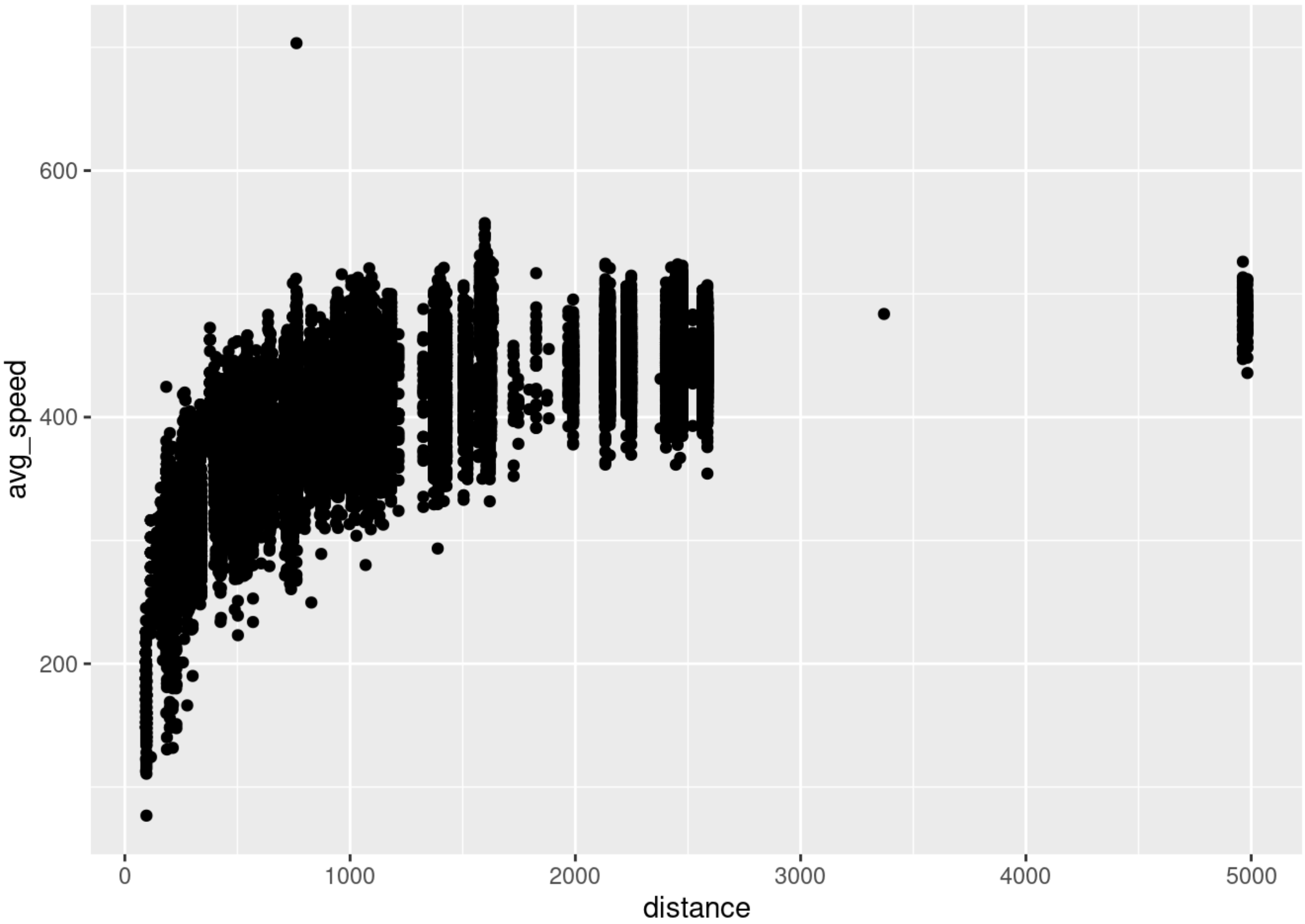
Hide

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

Exercise 8:

Hide

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

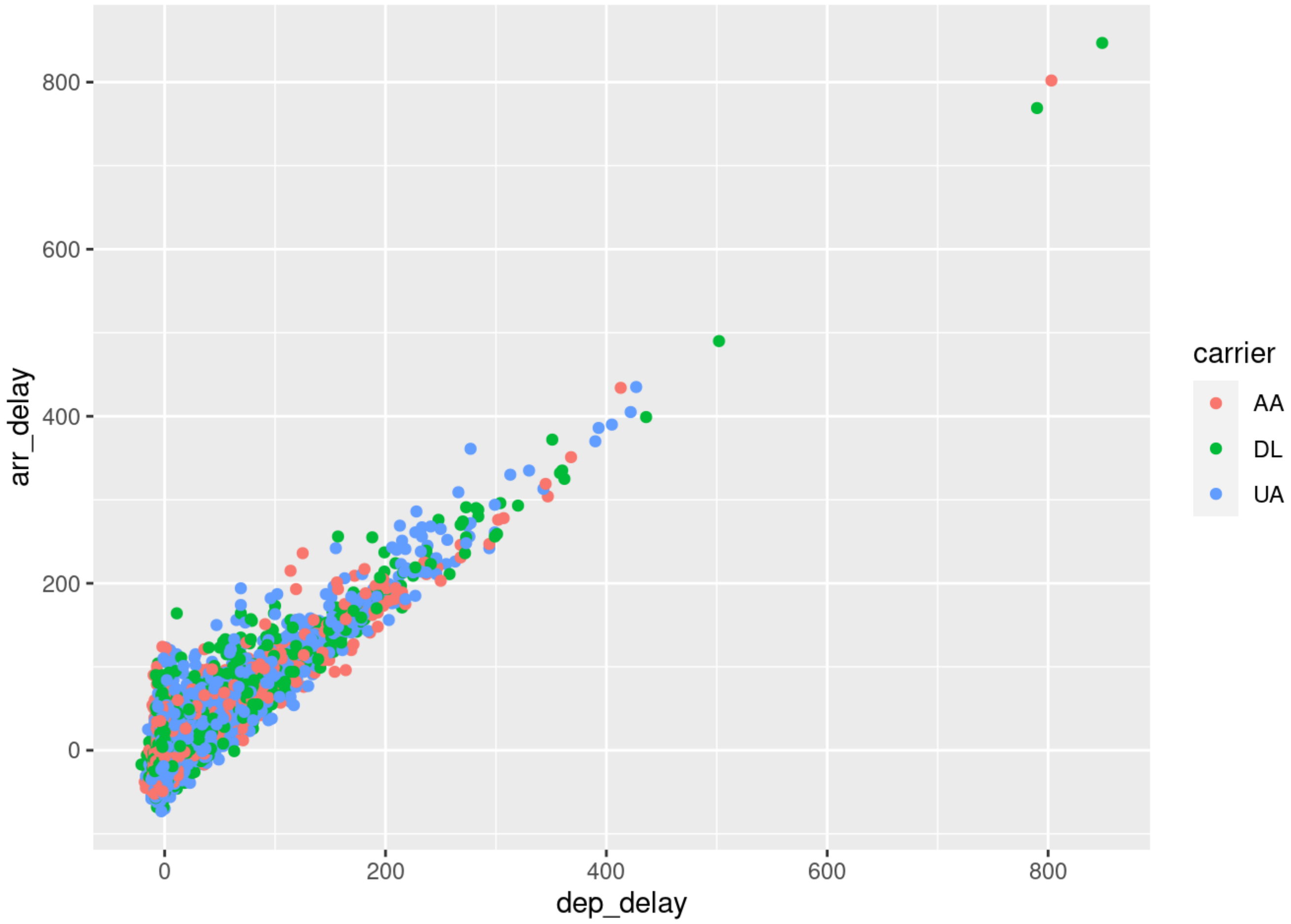


As Distence increses so does averages speed, however the change in average speed gets exponentially slower as distance incresses.

Exercise 8:

Hide

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



Exercise 8 (Part 2):

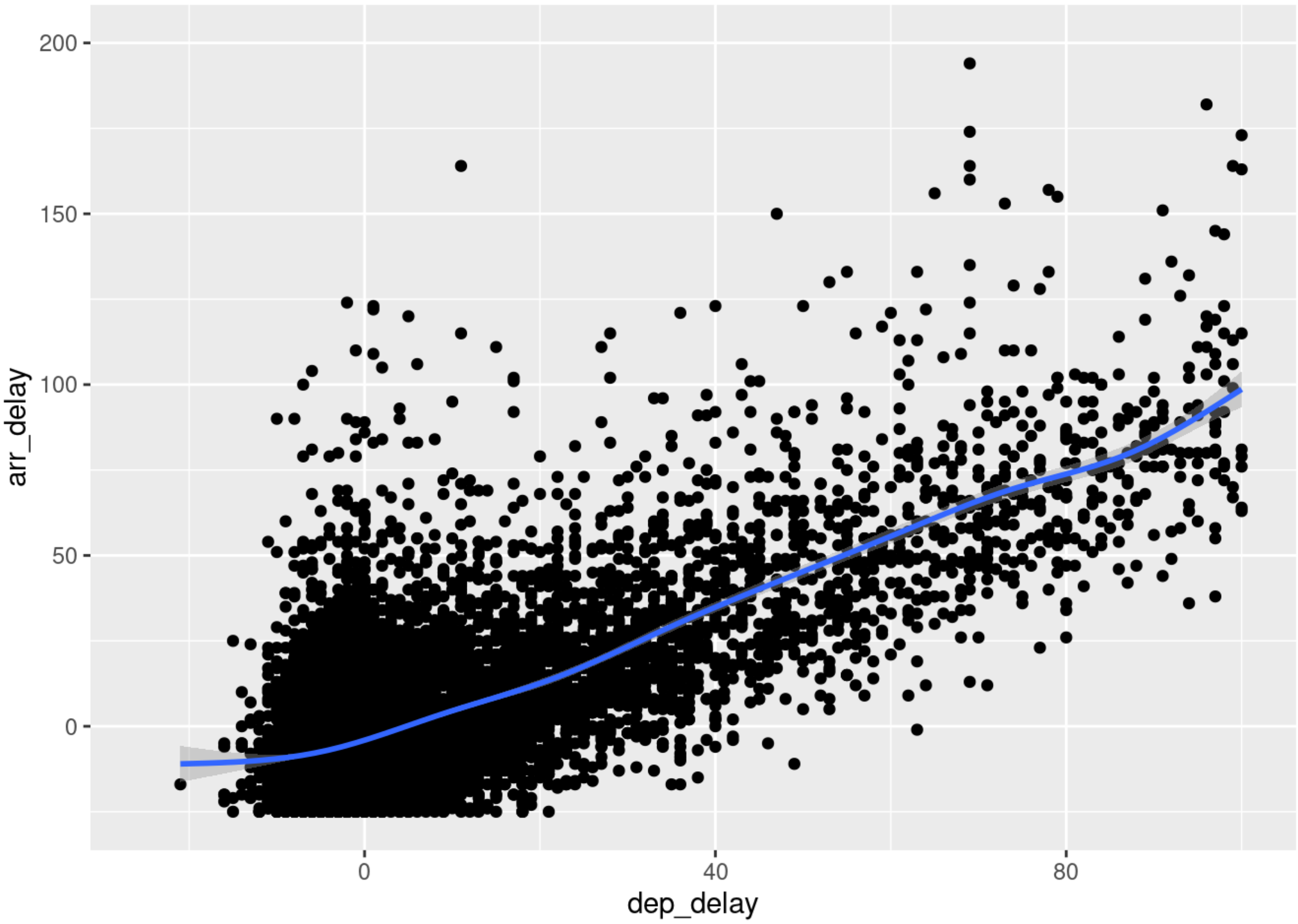
Hide

```
Replicate <- nycflights %>%
  filter(carrier == "AA" | carrier == "DL" | carrier == "UA")
Replicate %>%
ggplot(aes(x = dep_delay, y = arr_delay)) +
  xlim(-25,100) +
  ylim(-25, 200) +
geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 2490 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 2490 rows containing missing values (`geom_point()`).
```



If there is a departure delay delay you can expect to get to get to your destination on time if the delay is no longer than 20-30 min