Assignment_2

Victor Torres

2024-02-06

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(openintro)
```

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

```r
data(nycflights)
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"
```

```r
?nycflights
```

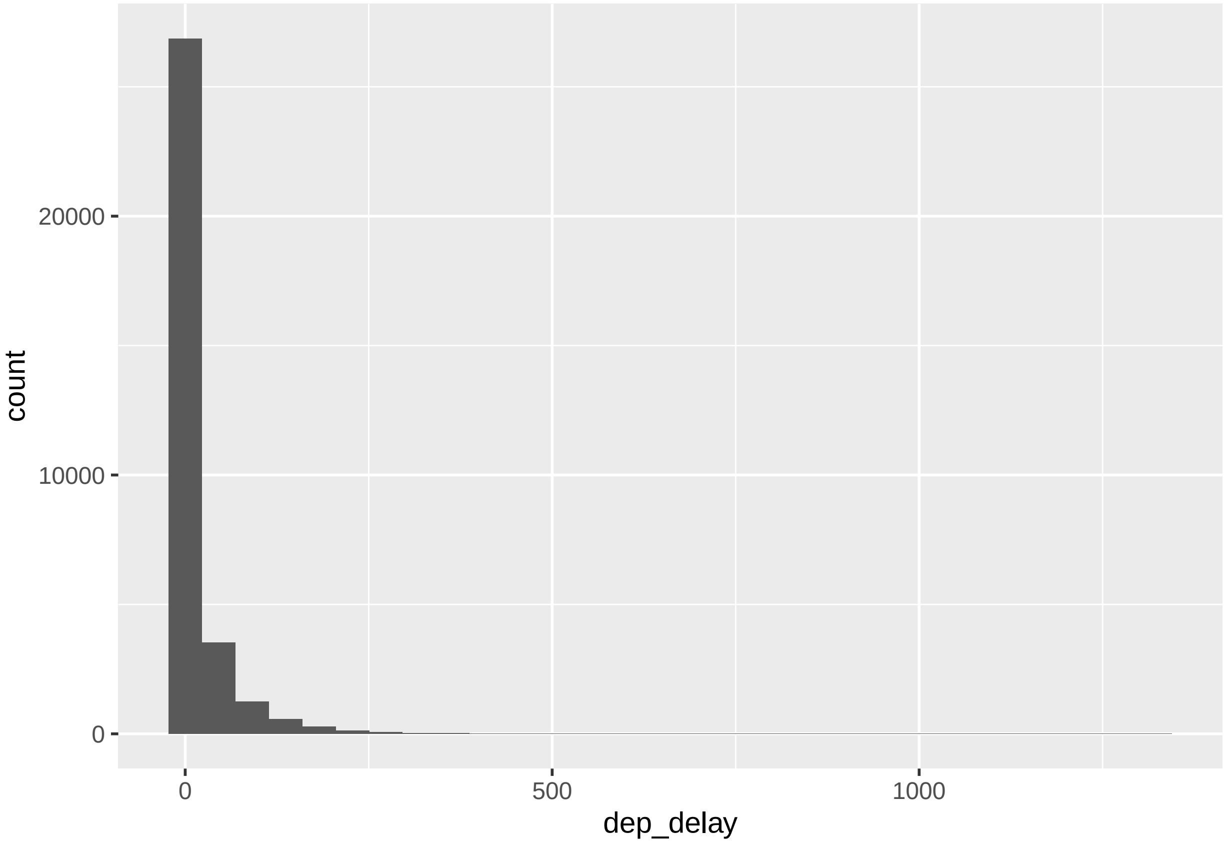
```
## starting httpd help server ... done
```

```r
glimpse(nycflights)
```

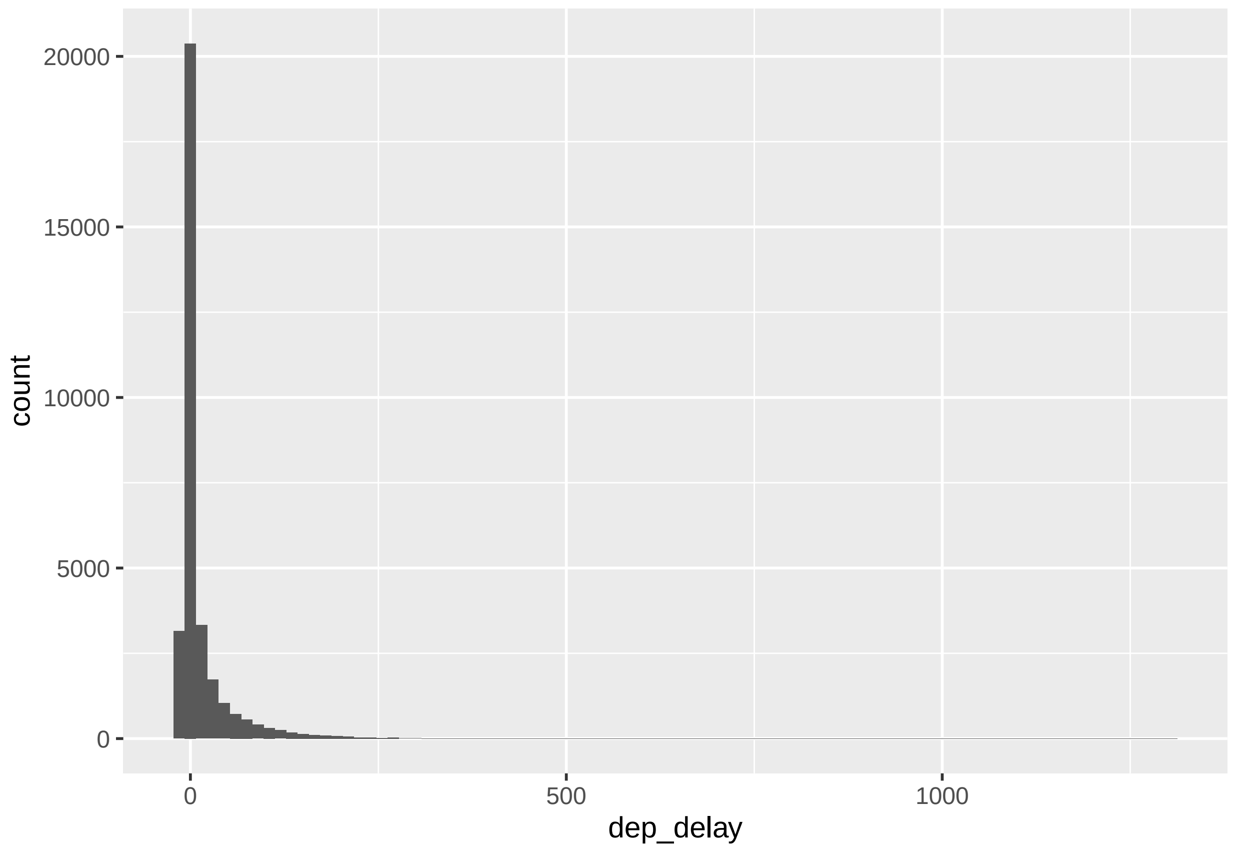
```
## Rows: 32,735
## Columns: 16
## $ year      <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, ~
## $ month     <int> 6, 5, 12, 5, 7, 1, 12, 8, 9, 4, 6, 11, 4, 3, 10, 1, 2, 8, 10~
## $ day       <int> 30, 7, 8, 14, 21, 1, 9, 13, 26, 30, 17, 22, 26, 25, 21, 23, ~
## $ dep_time  <int>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

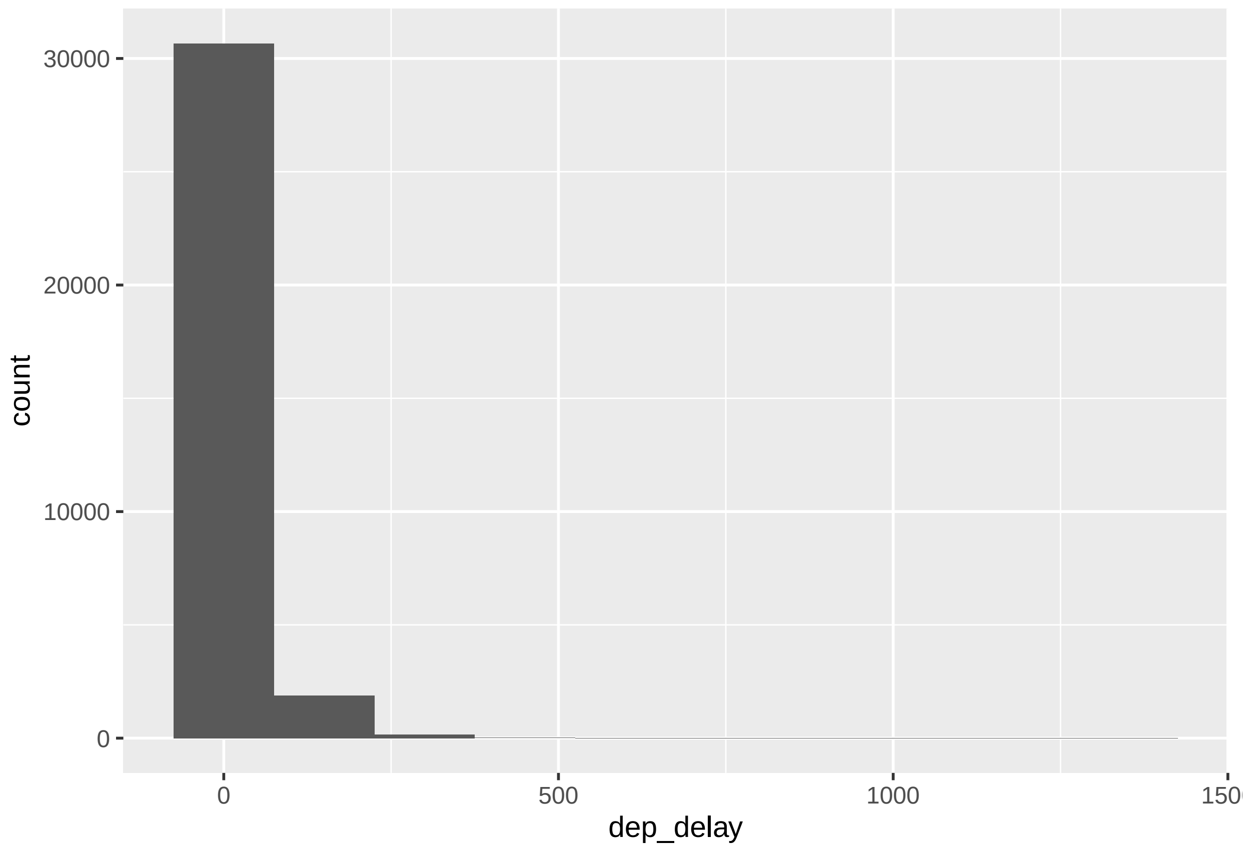
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



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



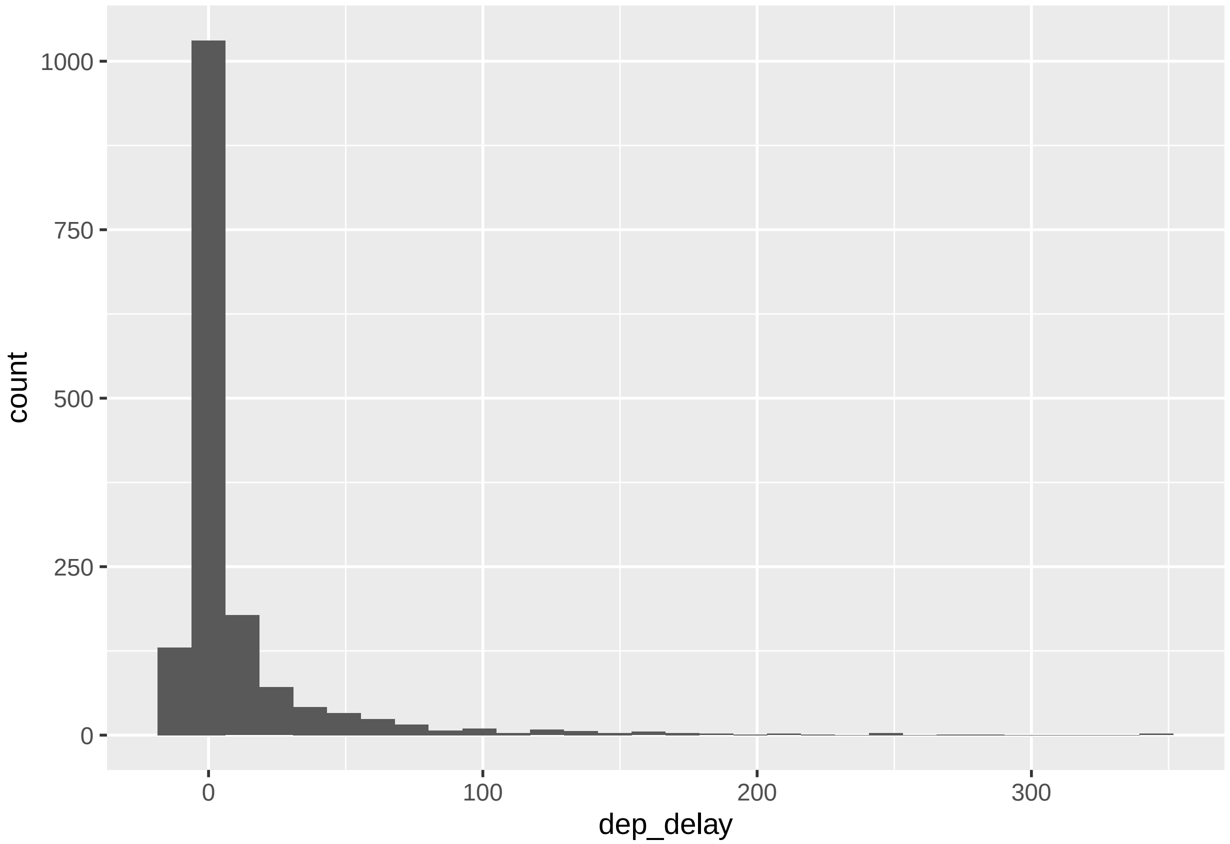
```
ggplot(data = nycflights, aes(x = dep_delay)) +
  geom_histogram(binwidth = 150)
```



#EXERCISE 1 #Look carefully at these three histograms. How do they compare? Are features revealed in one that are obscured in another? #Insert your answer here ##The three histograms contains the same data(nyc flights/dep_delay), the only difference is the visualization of the data. The argument "binwidth" controls the width of the bin representing the X-axis, crearly we can see the difference in the plots above.

```
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
```

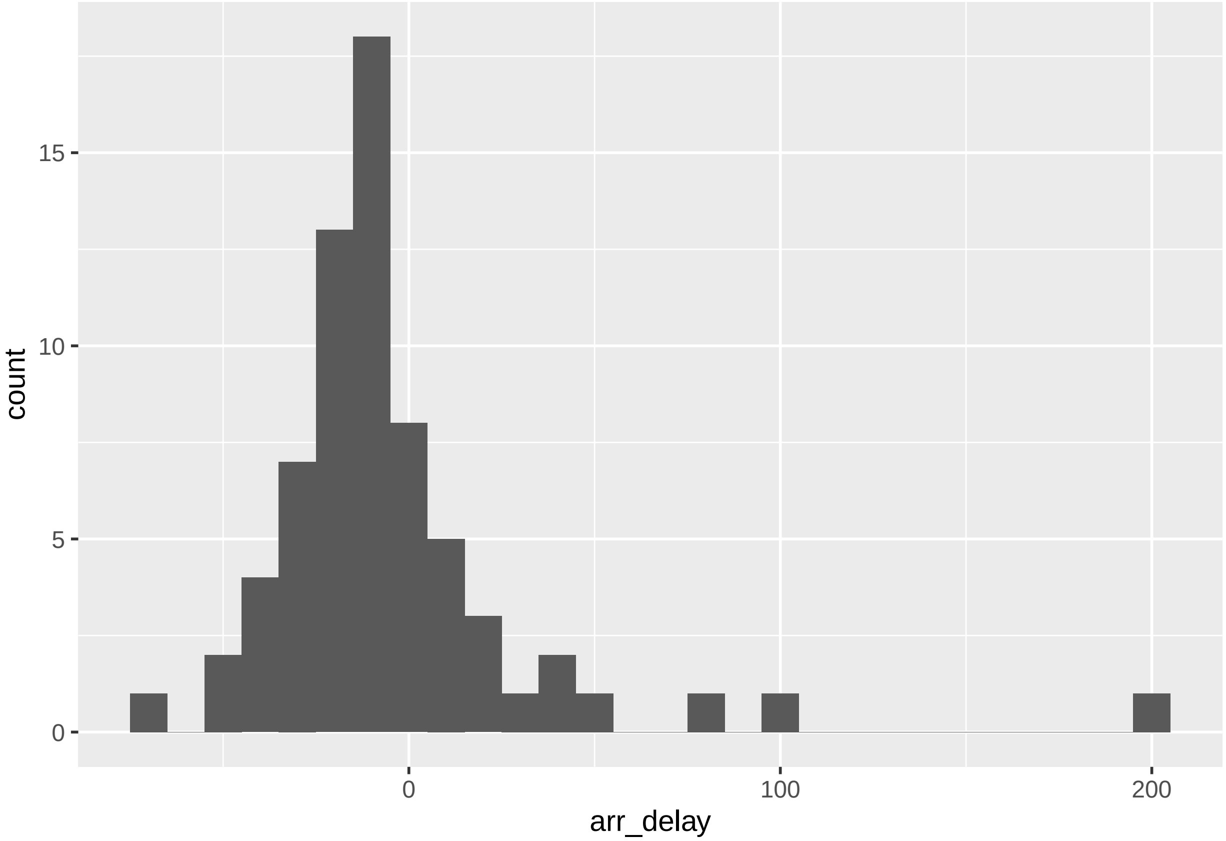
```
sfo_feb_flights <- nycflights %>%
  filter(dest == "SFO", month == 2)
```

#EXERCISE 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? ##Insert your answer here ###We can find 68 flights with this criteria (Flight from NYC to SFO in February)

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

#EXERCISE 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. ##Insert your answer here

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



```r
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 ##Calculate the median and interquartile range for arr_delays of flights in in the sfo_feb_flights data frame, grouped by carrier. Which carrier has the most variable arrival delays? ###The most variable arrival delays bolongs to AA #Insert your answer here

```r
sfo_feb_flights %>%
  group_by(carrier) %>%
  summarise(median_dd = median(arr_delay), iqr_rg = IQR(arr_delay))
```

```
## # A tibble: 5 x 3
##   carrier median_dd iqr_rg
##   <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
```

```
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
```

#EXERCISE 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? ##Insert your answer here ###The prons and cons are several. I did a calculation below with the lowest mean depature delay and lowest median depature delay (months 7 and 6) based on the data it would be de ideal dates to travel, however, the data can vary from other data frames such as arrival time, arrival delays, etc.

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

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

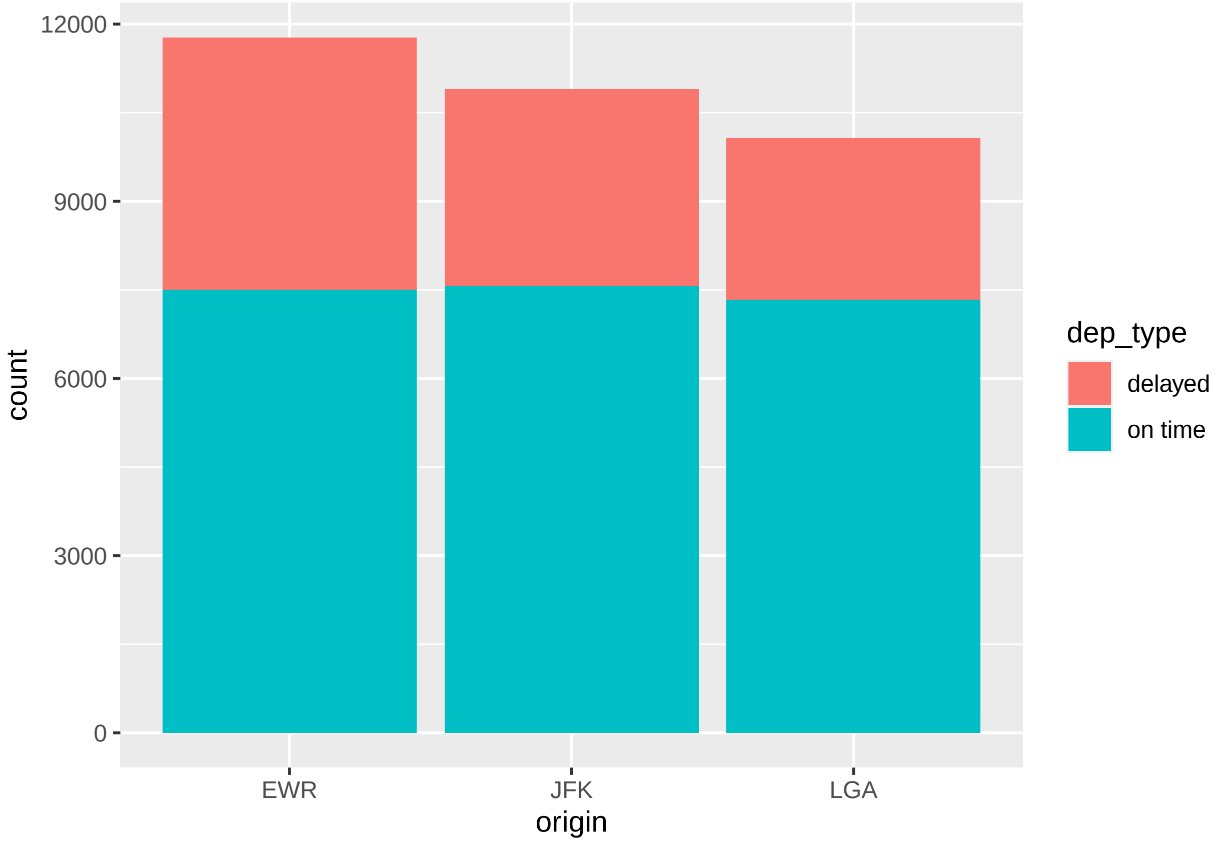
```
nycflights <- nycflights %>%
  mutate(dep_type = ifelse(dep_delay < 5, "on time", "delayed"))
```

```
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
```

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

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



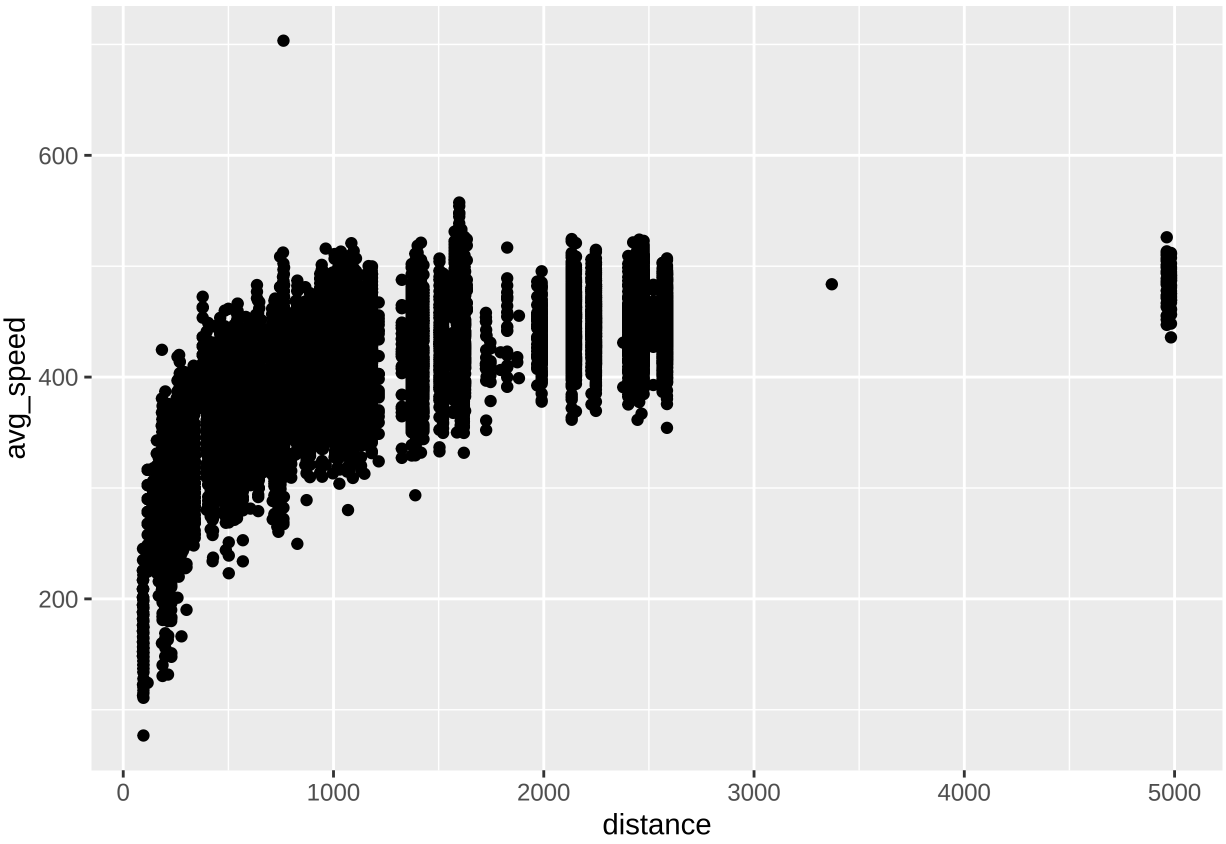
##Insert your answer here ###Based on the data and the chart, the best airport to based on timedepature percentage it will be LGA

#EXERCISE 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. ##insert your answer here

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

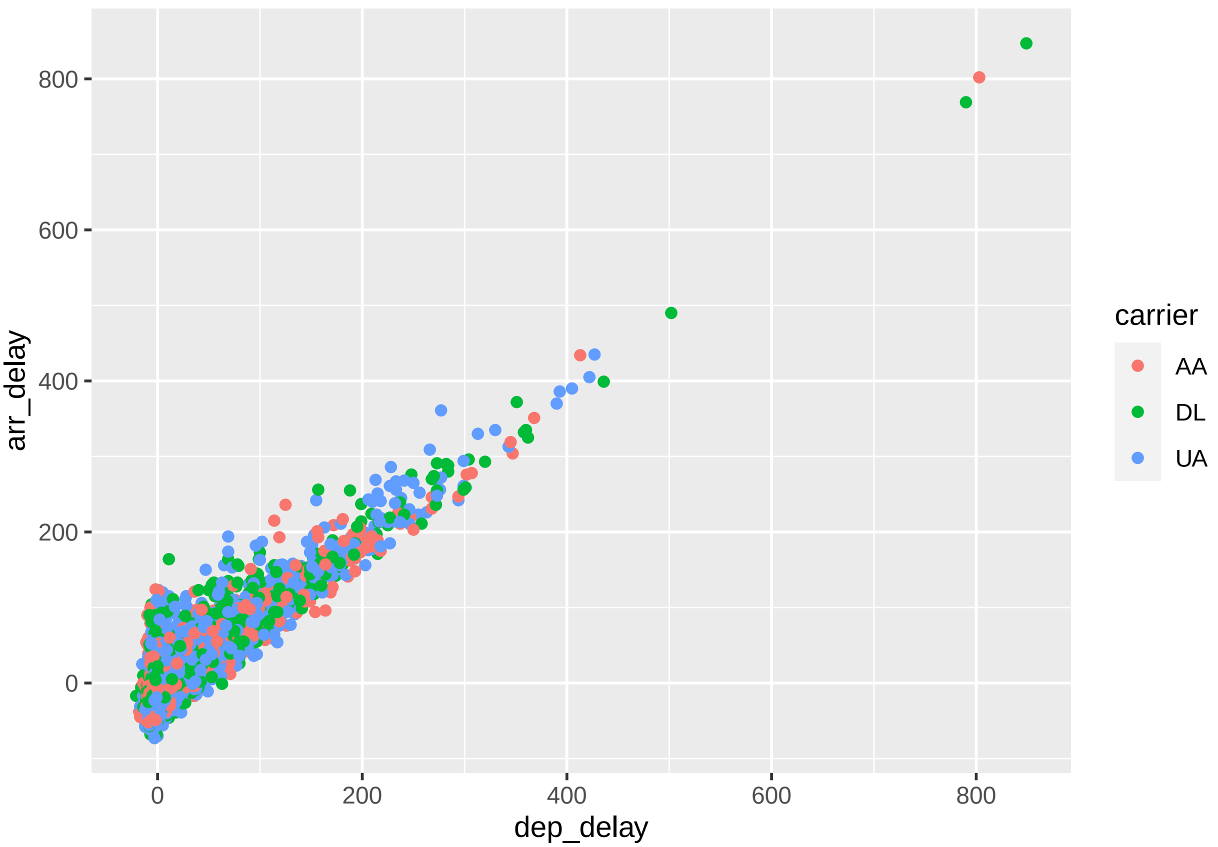
#EXERCISE 8 ##Make a scatterplot of avg_speed vs. distance. Describe the relationship between average speed and distance. Hint: Use geom_point(). ##insert your answer here

```r
ggplot(data = nycflights, mapping = aes(x = distance, y = avg_speed)) +
        geom_point()
```



#EXERCISE 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. ##insert your answer here

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


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