Data606 Lab:Intro to Data

Code

Sean Connin

2021-02-10

Hide

```
knitr::opts_chunk$set(echo=TRUE)
library(tidyverse)
library(ggplot2)
library(openintro)
data('nycflights')
```

Exercise 1

Insert any text here.

Hide

```
(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"
```

Hide

```
(glimpse(nycflights))
```

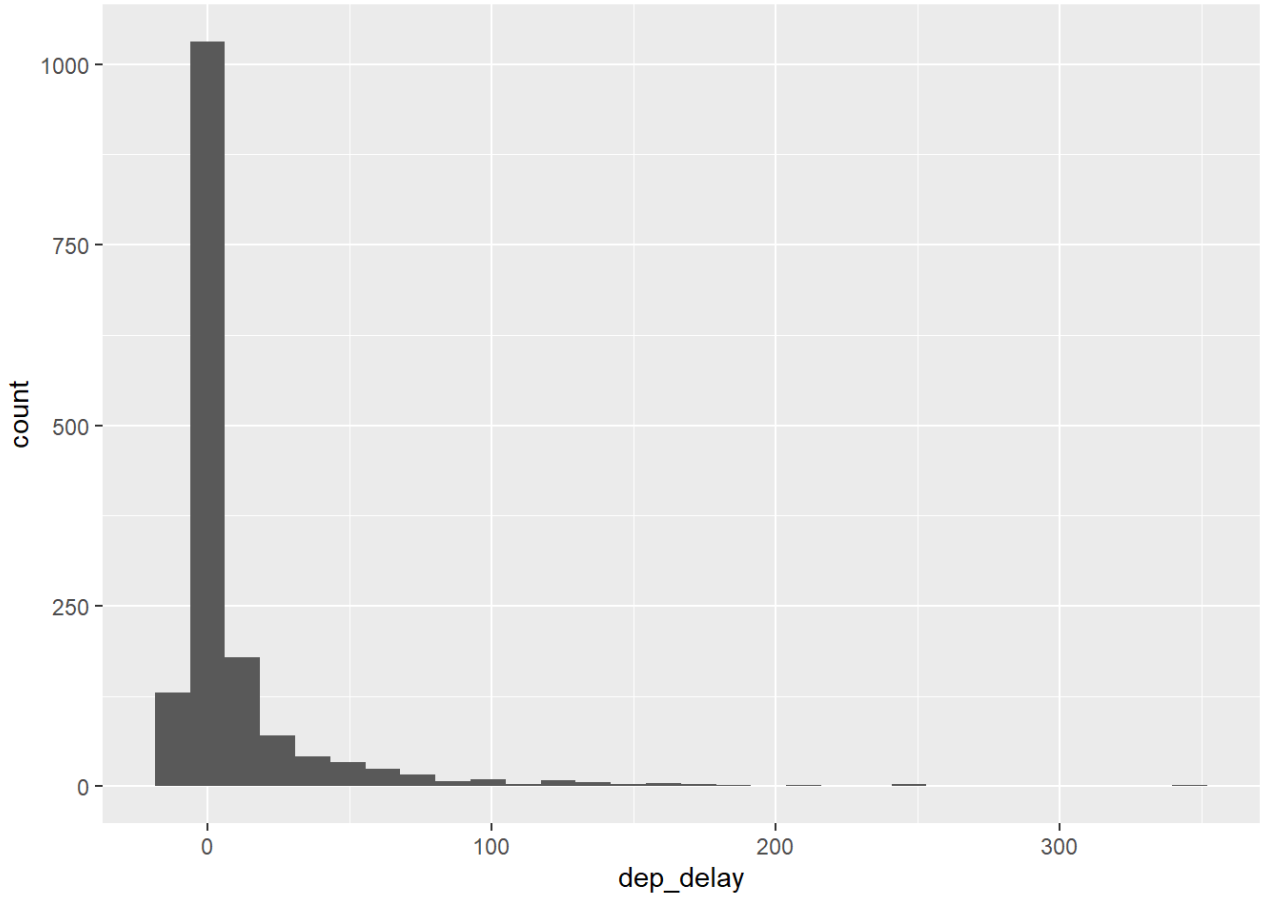
```
## Rows: 32,735
## Columns: 16
## $ year      <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 201...
## $ month     <int> 6, 5, 12, 5, 7, 1, 12, 8, 9, 4, 6, 11, 4, 3, 10, 1, 2, 8,...
## $ day       <int> 30, 7, 8, 14, 21, 1, 9, 13, 26, 30, 17, 22, 26, 25, 21, 2...
## $ dep_time  <int> 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,...
```

```
## # A tibble: 32,735 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     6    30      940        15     1216        -4 VX      N626VA
##  2  2013     5     7     1657        -3     2104        10 DL      N3760C
##  3  2013    12     8      859        -1     1238        11 DL      N712TW
##  4  2013     5    14     1841        -4     2122       -34 DL      N914DL
##  5  2013     7    21     1102        -3     1230        -8 9E      N823AY
##  6  2013     1     1     1817        -3     2008         3 AA      N3AXAA
##  7  2013    12     9     1259        14     1617        22 WN      N218WN
##  8  2013     8    13     1920        85     2032        71 B6      N284JB
##  9  2013     9    26      725       -10     1027        -8 AA      N3FSAA
## 10  2013     4    30     1323        62     1549        60 EV      N12163
## # ... with 32,725 more rows, and 7 more variables: flight <int>, origin <chr>,
## #   dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>
```

Hide

```
# delay of flights to LAX

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`.
```



Hide

```
#summary statistics include(mean, median,sd, var, IQR, min, max)

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
```

Hide

```
#flights to SFO in February

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

```
## # 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>
```

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?

Ans: 68 flights …

Hide

```
#create dataframe

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

```
##    year month day dep_time dep_delay arr_time arr_delay carrier tailnum flight
## 1  2013     2  18     1527        57     1903        48      DL  N711ZX   1322
## 2  2013     2   3      613        14     1008        38      UA  N502UA    691
## 3  2013     2  15      955        -5     1313       -28      DL  N717TW   1765
## 4  2013     2  18     1928        15     2239        -6      UA  N24212   1214
## 5  2013     2  24     1340         2     1644       -21      UA  N76269   1111
## 6  2013     2  25     1415       -10     1737       -13      UA  N532UA    394
## 7  2013     2   7     1032         1     1352       -10      B6  N627JB    641
## 8  2013     2  15     1805        20     2122         2      AA  N335AA    177
## 9  2013     2  13     1056        -4     1412       -13      UA  N532UA    642
## 10 2013     2   8      656        -4     1039        -6      DL  N710TW   1865
## 11 2013     2  11     1910        40     2204         2      UA  N532UA    272
## 12 2013     2  13      833        -2     1210        -5      UA  N73259   1739
## 13 2013     2  25     1048        -1     1401       -30      UA  N37293   1436
## 14 2013     2  20     1849        -6     2218       -22      VX  N641VA     29
## 15 2013     2  12      723        -7     1035       -40      VX  N839VA     11
## 16 2013     2  27     1721        21     2048        -1      DL  N718TW     31
## 17 2013     2   1     1436         6     1758       -17      DL  N705TW   2126
## 18 2013     2  20     1629        -1     1951       -24      VX  N845VA     27
## 19 2013     2  25     1508        38     1836        21      DL  N718TW   1322
## 20 2013     2  12     1901         2     2216       -13      UA  N27205   1139
## 21 2013     2   7      952        -8     1336        -5      DL  N710TW   1765
## 22 2013     2   1      754        -6     1119        -6      UA  N525UA    397
## 23 2013     2  21     1925        47     2244        34      UA  N512UA    389
## 24 2013     2  26     1857        -3     2216       -45      DL  N718TW   1967
## 25 2013     2  24      733        -3     1057       -18      B6  N821JB    643
## 26 2013     2  28     1938        43     2226       -14      VX  N635VA     29
## 27 2013     2  11      658        -2     1034       -11      DL  N706TW   1865
## 28 2013     2   4     1107        37     1440        45      AA  N343AA    179
## 29 2013     2  21     1852        -8     2213       -48      DL  N723TW   1967
## 30 2013     2  27     1830        45     2128         8      AA  N329AA    177
## 31 2013     2  22     1830        -8     2207        -3      UA  N512UA    389
## 32 2013     2   7     1741        -4     2117        -3      AA  N335AA    177
## 33 2013     2  20      726        -4     1052       -23      VX  N840VA     11
## 34 2013     2  27     1056        -3     1406       -35      UA  N87512   1120
## 35 2013     2  19     1210       100     1554        99      VX  N842VA     23
## 36 2013     2  20     1327        -2     1646       -18      UA  N35260   1641
## 37 2013     2   4      654        -6     1028       -17      DL  N721TW   1865
## 38 2013     2  24     1547        17     1928        18      AA  N381AA     85
## 39 2013     2   3     1428        -2     1810        -5      DL  N722TW   2126
## 40 2013     2  15     1858        -2     2241       -20      DL  N706TW   1967
## 41 2013     2   3     1026        -5     1414        11      B6  N583JB    641
## 42 2013     2   1      756         9     1129         3      B6  N789JB    643
## 43 2013     2   3      724        -6     1113        -2      VX  N844VA     11
## 44 2013     2   5      744        -1     1133         8      AA  N383AA     59
## 45 2013     2  17      655        -5     1018       -27      DL  N710TW   1865
## 46 2013     2  17     1027        -3     1345       -30      VX  N839VA     23
## 47 2013     2  14     1434         4     1745       -30      DL  N709TW   2126
## 48 2013     2  27     1951        56     2256        16      VX  N642VA     29
## 49 2013     2  28     1624        -6     1909       -66      VX  N842VA     27
## 50 2013     2  12     2159       209      118       196      UA  N508UA    272
## 51 2013     2  25      916        91     1241        76      AA  N335AA     59
## 52 2013     2  11      753        -7     1115       -10      UA  N510UA    397
```

```
## 53 2013     2   5     1030        -1     1351       -11      B6  N821JB    641
## 54 2013     2  25     1030         0     1356         1      AA  N367AA    179
## 55 2013     2  19      652        -8     1038        -7      DL  N706TW   1865
## 56 2013     2  20     1032        -3     1351       -13      B6  N658JB    641
## 57 2013     2  11     1539         9     1844       -26      AA  N352AA     85
## 58 2013     2  10      955        -5     1332        -9      DL  N722TW   1765
## 59 2013     2   4      657        -3     1034         9      UA  N510UA    799
## 60 2013     2   4     1719         9     2043         7      UA  N29124   1178
## 61 2013     2  19     1857        -3     2246       -15      DL  N723TW   1967
## 62 2013     2  14     1725         0     2015       -35      UA  N554UA    512
## 63 2013     2  21     1107        -3     1420       -20      UA  N508UA    642
## 64 2013     2  21     1745         0     2106       -14      AA  N329AA    177
## 65 2013     2   3     1055        -5     1405       -20      UA  N510UA    642
## 66 2013     2  25     1855         0     2220       -20      VX  N624VA     29
## 67 2013     2   6     1654        -6     2015       -34      DL  N624AG     31
## 68 2013     2  25     1023        -7     1336       -39      VX  N845VA     23
##    origin dest air_time distance hour minute
## 1     JFK  SFO      358     2586   15     27
## 2     JFK  SFO      367     2586    6     13
## 3     JFK  SFO      338     2586    9     55
## 4     EWR  SFO      353     2565   19     28
## 5     EWR  SFO      341     2565   13     40
## 6     JFK  SFO      355     2586   14     15
## 7     JFK  SFO      359     2586   10     32
## 8     JFK  SFO      338     2586   18      5
## 9     JFK  SFO      347     2586   10     56
## 10    JFK  SFO      361     2586    6     56
## 11    JFK  SFO      332     2586   19     10
## 12    EWR  SFO      351     2565    8     33
## 13    EWR  SFO      355     2565   10     48
## 14    JFK  SFO      362     2586   18     49
## 15    JFK  SFO      349     2586    7     23
## 16    JFK  SFO      327     2586   17     21
## 17    JFK  SFO      357     2586   14     36
## 18    JFK  SFO      350     2586   16     29
## 19    JFK  SFO      352     2586   15      8
## 20    EWR  SFO      342     2565   19      1
## 21    JFK  SFO      376     2586    9     52
## 22    JFK  SFO      349     2586    7     54
## 23    JFK  SFO      339     2586   19     25
## 24    JFK  SFO      318     2586   18     57
## 25    JFK  SFO      345     2586    7     33
## 26    JFK  SFO      330     2586   19     38
## 27    JFK  SFO      354     2586    6     58
## 28    JFK  SFO      360     2586   11      7
## 29    JFK  SFO      351     2586   18     52
## 30    JFK  SFO      329     2586   18     30
## 31    JFK  SFO      358     2586   18     30
## 32    JFK  SFO      367     2586   17     41
## 33    JFK  SFO      359     2586    7     26
## 34    EWR  SFO      317     2565   10     56
## 35    JFK  SFO      353     2586   12     10
## 36    EWR  SFO      356     2565   13     27
## 37    JFK  SFO      352     2586    6     54
```

```
## 38    JFK  SFO      354     2586   15     47
## 39    JFK  SFO      360     2586   14     28
## 40    JFK  SFO      348     2586   18     58
## 41    JFK  SFO      370     2586   10     26
## 42    JFK  SFO      369     2586    7     56
## 43    JFK  SFO      373     2586    7     24
## 44    JFK  SFO      366     2586    7     44
## 45    JFK  SFO      373     2586    6     55
## 46    JFK  SFO      362     2586   10     27
## 47    JFK  SFO      345     2586   14     34
## 48    JFK  SFO      335     2586   19     51
## 49    JFK  SFO      328     2586   16     24
## 50    JFK  SFO      344     2586   21     59
## 51    JFK  SFO      356     2586    9     16
## 52    JFK  SFO      340     2586    7     53
## 53    JFK  SFO      361     2586   10     30
## 54    JFK  SFO      360     2586   10     30
## 55    JFK  SFO      369     2586    6     52
## 56    JFK  SFO      359     2586   10     32
## 57    JFK  SFO      347     2586   15     39
## 58    JFK  SFO      366     2586    9     55
## 59    JFK  SFO      362     2586    6     57
## 60    EWR  SFO      355     2565   17     19
## 61    JFK  SFO      348     2586   18     57
## 62    JFK  SFO      335     2586   17     25
## 63    JFK  SFO      356     2586   11      7
## 64    JFK  SFO      348     2586   17     45
## 65    JFK  SFO      351     2586   10     55
## 66    JFK  SFO      367     2586   18     55
## 67    JFK  SFO      355     2586   16     54
## 68    JFK  SFO      359     2586   10     23
```

Hide

```
#count rows

sfo_feb_flights%>%nrow()  #--> 68 rows
```

```
## [1] 68
```

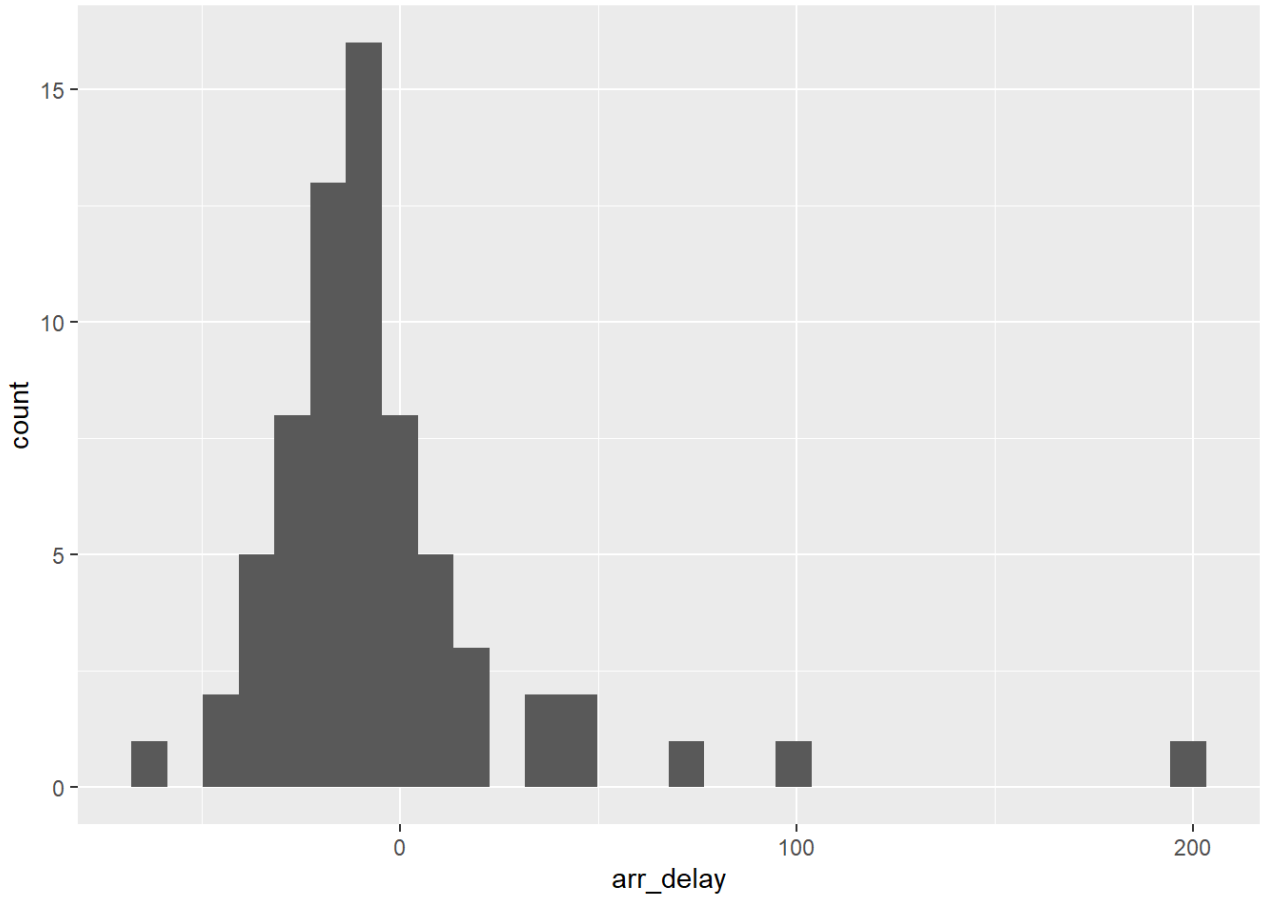
Exercise 3

Describe the distribution of the arrival delays of these flights using a histogram and appropriate summary statistics.

Hide

```
# histogram of sfo_feb_flights delays

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

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



Hide

```
# sfo_feb_flight summary stats

#summary statistics include(mean, median,sd, var, IQR, min, max)

summary<- sfo_feb_flights %>%
    summarise(median = median(arr_delay),
    interquartile_range = IQR(arr_delay),
    minimum = min(arr_delay),
    maximum = max(arr_delay),
    n=n())
```

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?

Ans: United Airlines has the most arrival delays

Hide

```
(sfo_feb_flights%>%
    group_by(carrier)%>%
    summarise(median = median(arr_delay),
    interquartile_range = IQR(arr_delay),
    maximum = max(arr_delay))%>%
    arrange(desc(maximum), .by_group =TRUE))
```

```
## # A tibble: 5 x 4
##   carrier median interquartile_range maximum
##   <chr>    <dbl>               <dbl>   <dbl>
## 1 UA       -10                  22       196
## 2 VX       -22.5                21.2      99
## 3 AA         5                  17.5      76
## 4 DL       -15                  22        48
## 5 B6       -10.5                12.2      11
```

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?

Hide

```
#classify lights by 'on time' or 'delayed'

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

#group flights and on time departure rates

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
```

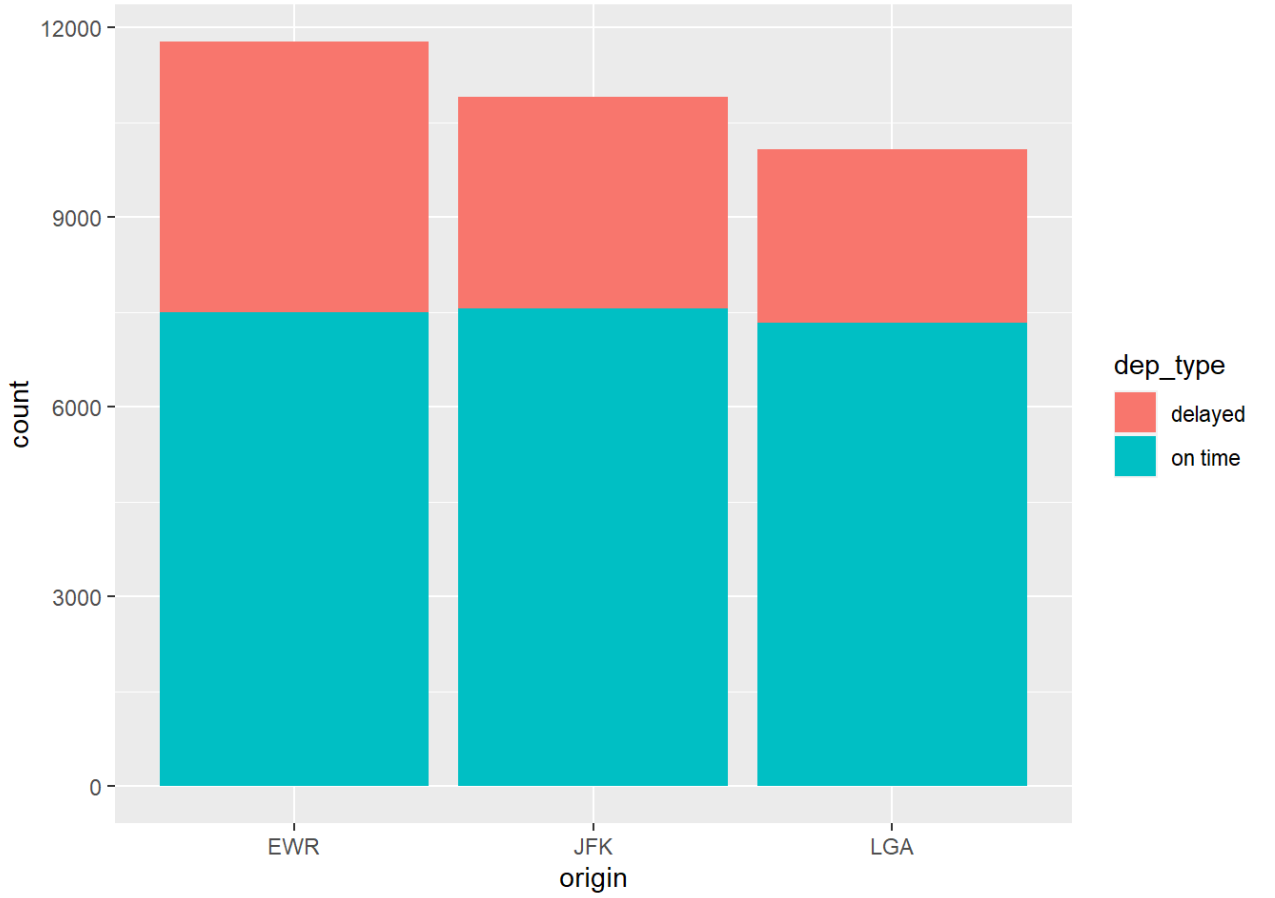
Exercise 6

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

I would pick LGA –> assuming that I could get to the airport on time.

Hide

```
#bar plot of departure rates by carrier

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



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.

Hide

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

```
## # A tibble: 32,735 x 18
##     year month   day dep_time dep_delay arr_time arr_delay carrier tailnum
##    <int> <int> <int>    <int>     <dbl>    <int>     <dbl> <chr>   <chr>
##  1  2013     6    30      940        15     1216        -4 VX      N626VA
##  2  2013     5     7     1657        -3     2104        10 DL      N3760C
##  3  2013    12     8      859        -1     1238        11 DL      N712TW
##  4  2013     5    14     1841        -4     2122       -34 DL      N914DL
##  5  2013     7    21     1102        -3     1230        -8 9E      N823AY
##  6  2013     1     1     1817        -3     2008         3 AA      N3AXAA
##  7  2013    12     9     1259        14     1617        22 WN      N218WN
##  8  2013     8    13     1920        85     2032        71 B6      N284JB
##  9  2013     9    26      725       -10     1027        -8 AA      N3FSAA
## 10  2013     4    30     1323        62     1549        60 EV      N12163
## # ... with 32,725 more rows, and 9 more variables: flight <int>, origin <chr>,
## #   dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,
## #   dep_type <chr>, avg_speed <dbl>
```

###Exercise 8

Make a scatterplot of avg_speed vs. distance. Describe the relationship between average speed and distance. Hint: Use geom_point().

Average flight speeds are relatively constant (avg. 317 mph)for travel distances < 500 miles and then increase rapidly between a travel distance of 500 and 1000+ miles - leveling off at an average speed of 511 mph. The data indicate at least one flight reaching 703 mph. This may be an outlier or a unique instance owing to that particular airplane.

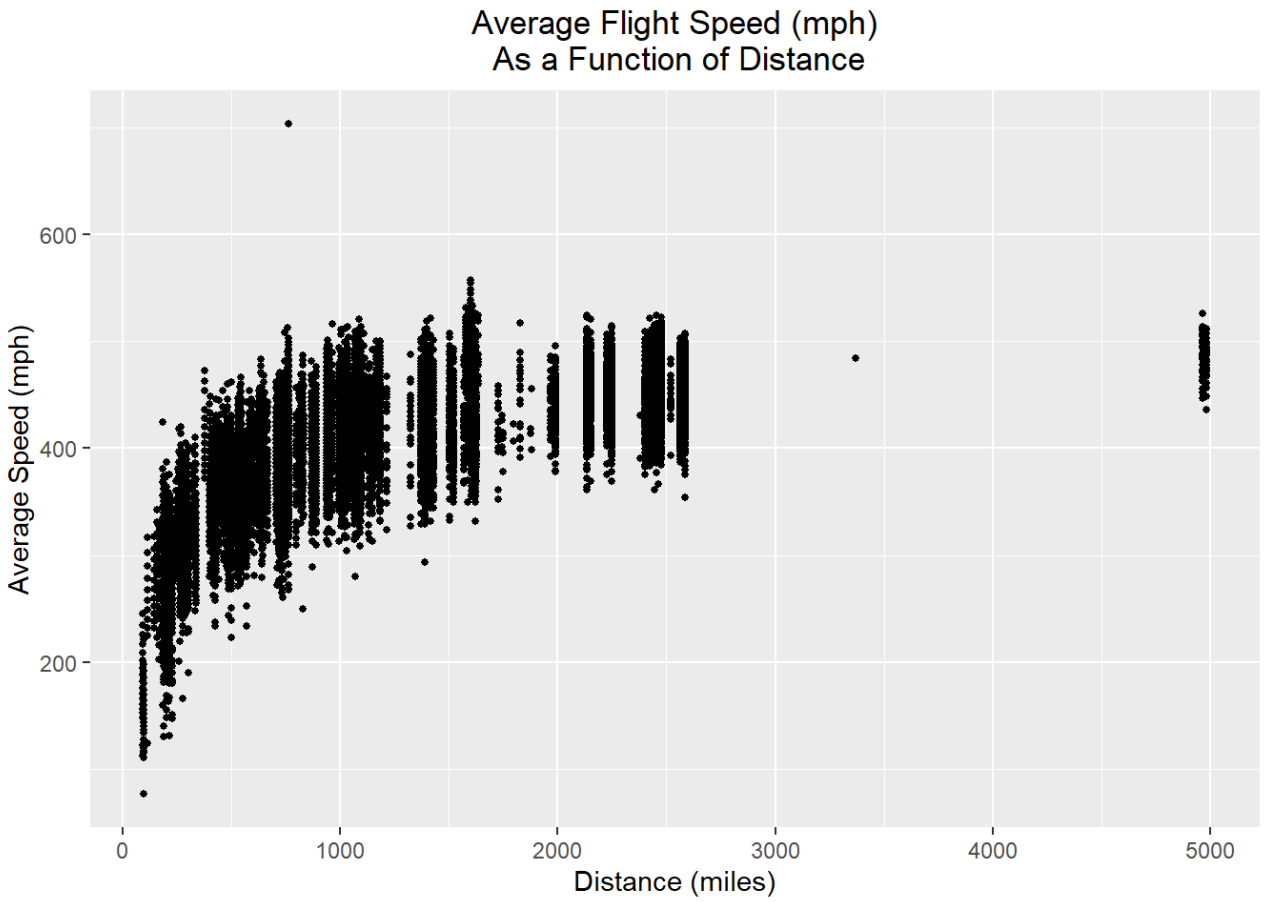
Hide

```
# compute summary stats for avg_speed

avg_df%>%summarize(mean=mean(avg_speed),
                   max_speed=max(avg_speed),
                   min_speed=min(avg_speed))
```

```
## # A tibble: 1 x 3
##    mean max_speed min_speed
##   <dbl>     <dbl>     <dbl>
## 1  394.      703.      76.8
```

Hide

```
#plot average_speeds vs. distance

avg_df%>%ggplot(aes(x=distance,y=avg_speed))+
    geom_point(size=1)+
    labs(title="Average Flight Speed (mph)\n As a Function of Distance", y="Average Sp
         eed (mph)", x = "Distance (miles)")+ theme(plot.title=element_text(hjust=0.5
         ))
```



Hide

```
#find average speed at distances < 500 miles

(avg_df%>%filter(distance<500)%>%summarize(mean=mean(avg_speed)))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  317.
```

Hide

```
#find average speed at between 500 - 600 miles

(avg_df%>%
        filter(avg_speed >500 & avg_speed <600)%>%
        summarize(mean=mean(avg_speed)))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  511.
```

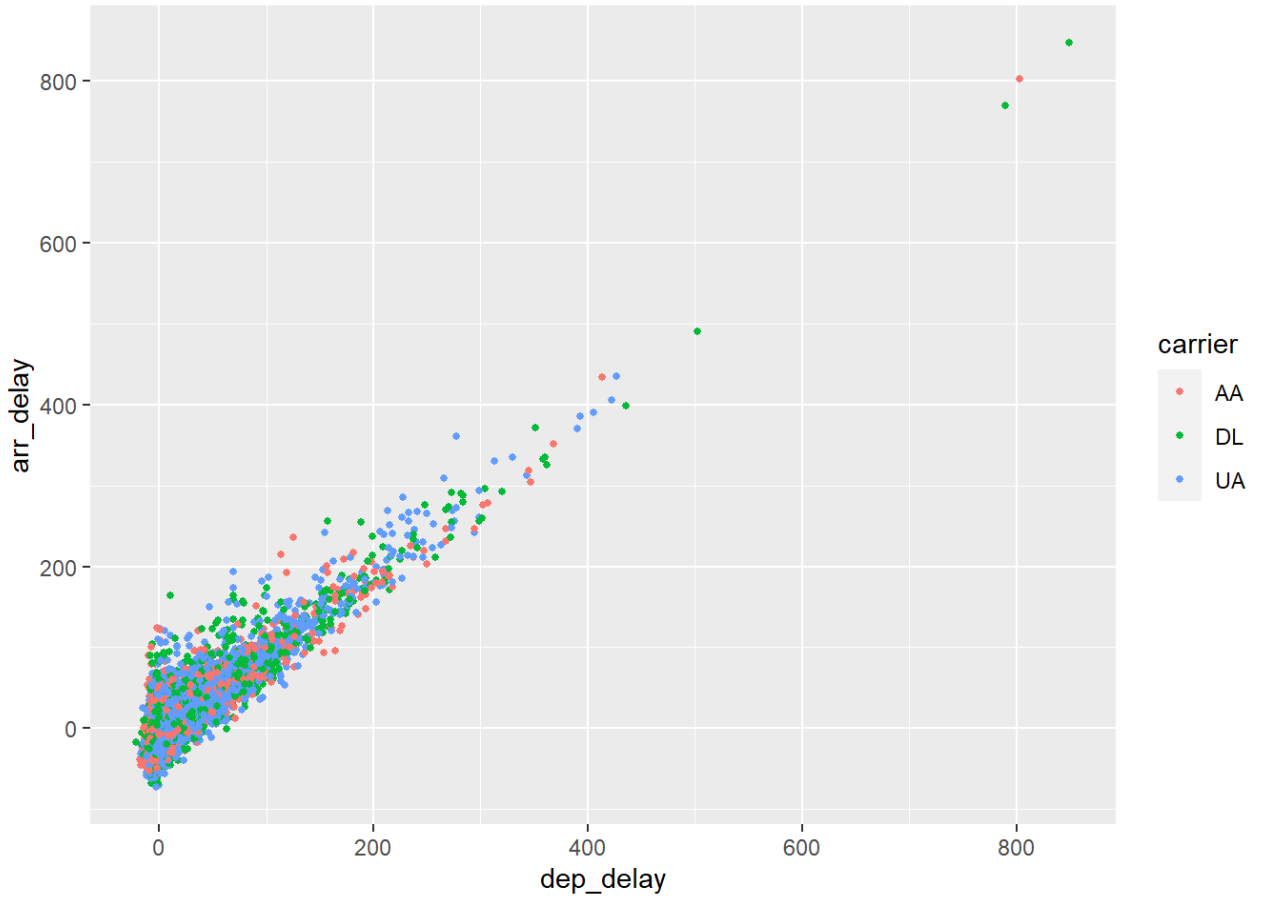
Exercise 9

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.

My maximum departure delay cutoff would be approximately 1 hr.

Hide

```
(nycflights%>%data.frame()%>%
    filter(carrier == 'AA' | carrier=='DL'|
    carrier== 'UA')%>%
    ggplot(aes(x=dep_delay,y=arr_delay,
               color=carrier))+geom_point(size=1))
```



Hide

```
# calculate delay cut-off point for on-time arrivals

nycflights%>%filter(arr_delay <=0)%>%summarise(max=max(dep_delay))
```

```
## # A tibble: 1 x 1
##     max
##   <dbl>
## 1    63
```