Code ▾

Data606 Lab2:Intro to Data

Sean Connin

2021-03-08

Hide

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

Exercise 1

Insert any text here.

Hide

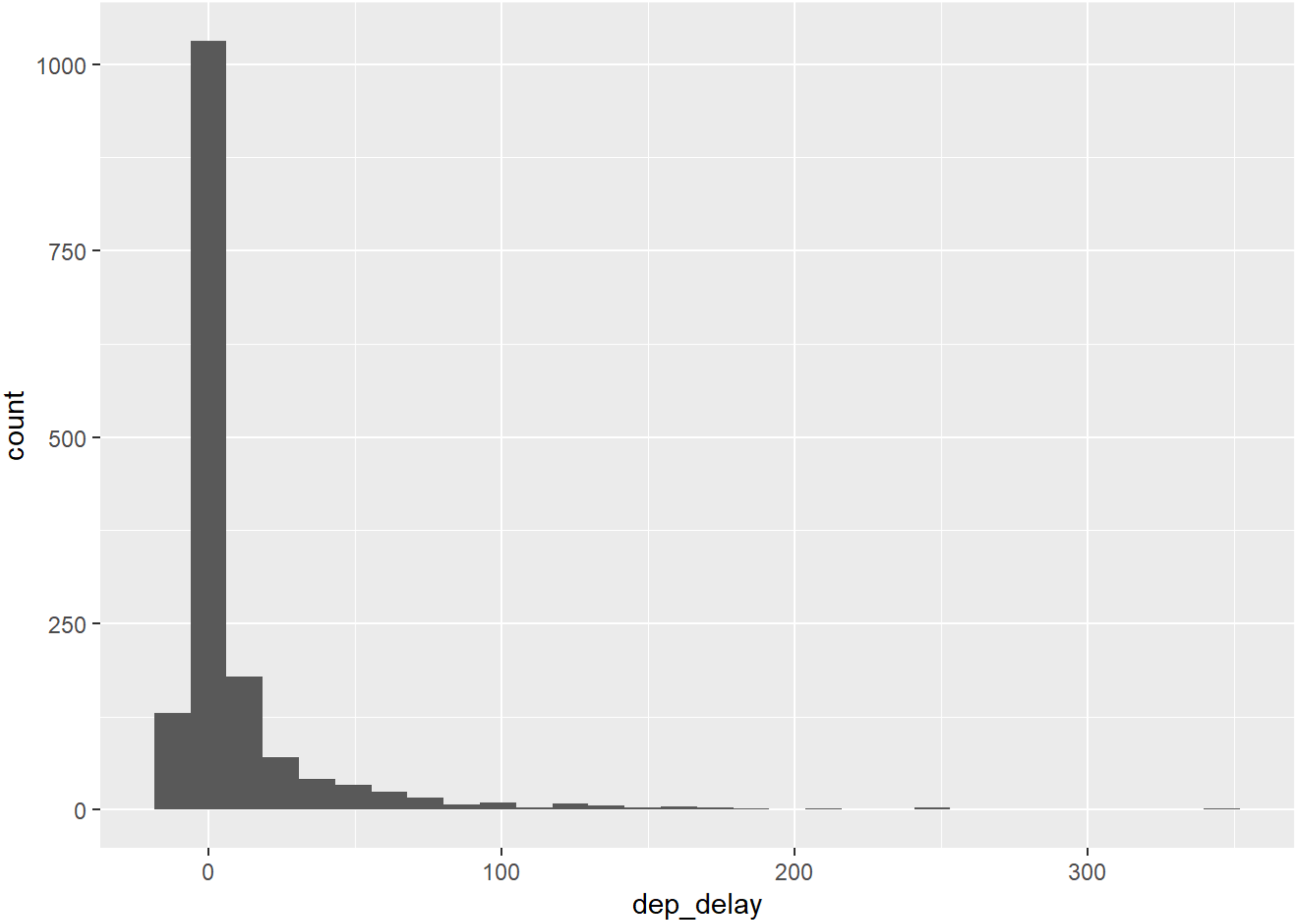
```
names(nycflights)
```

```
##  [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)

# 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)
```

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)

#count rows

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

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

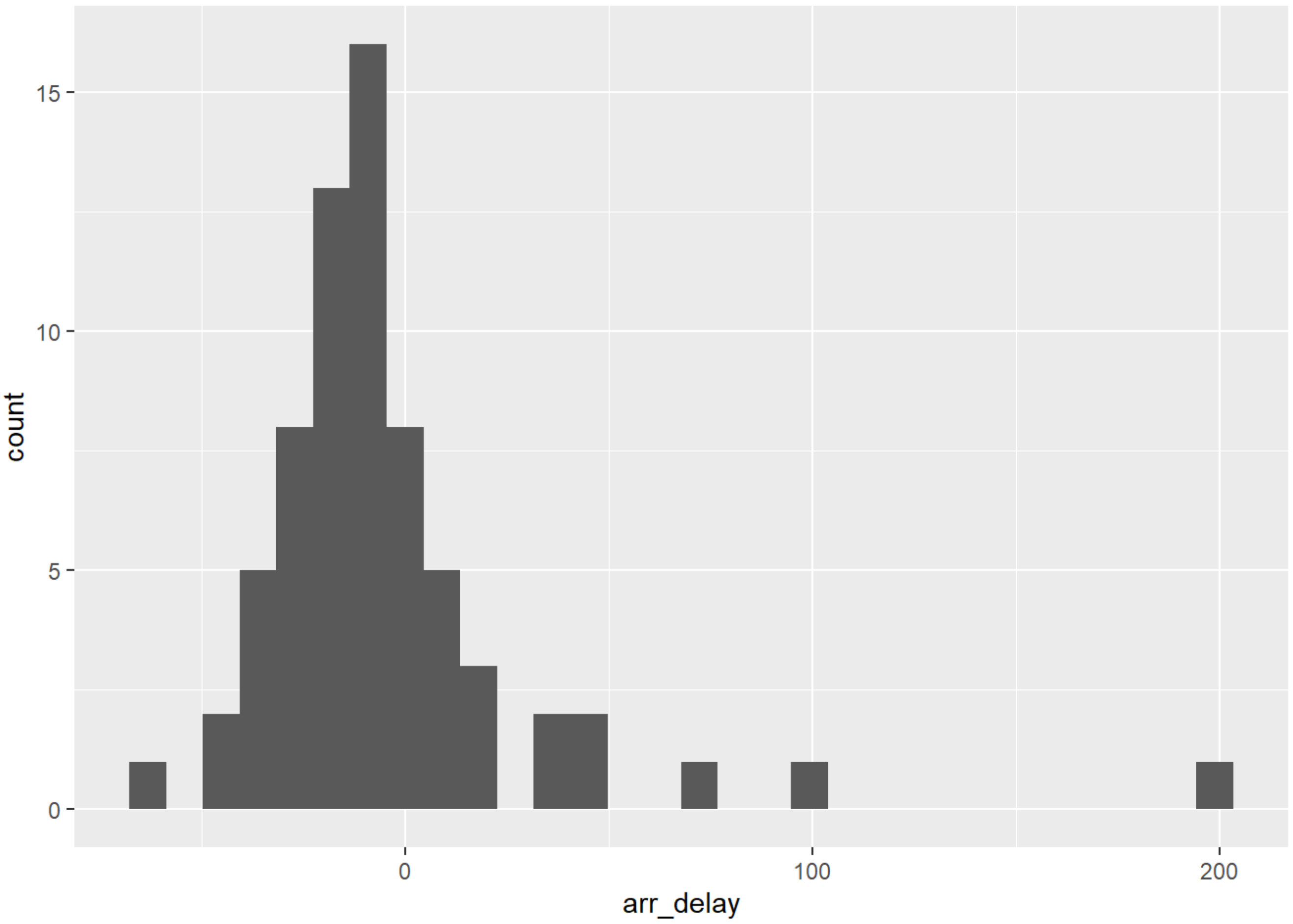
Exercise 3

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
```

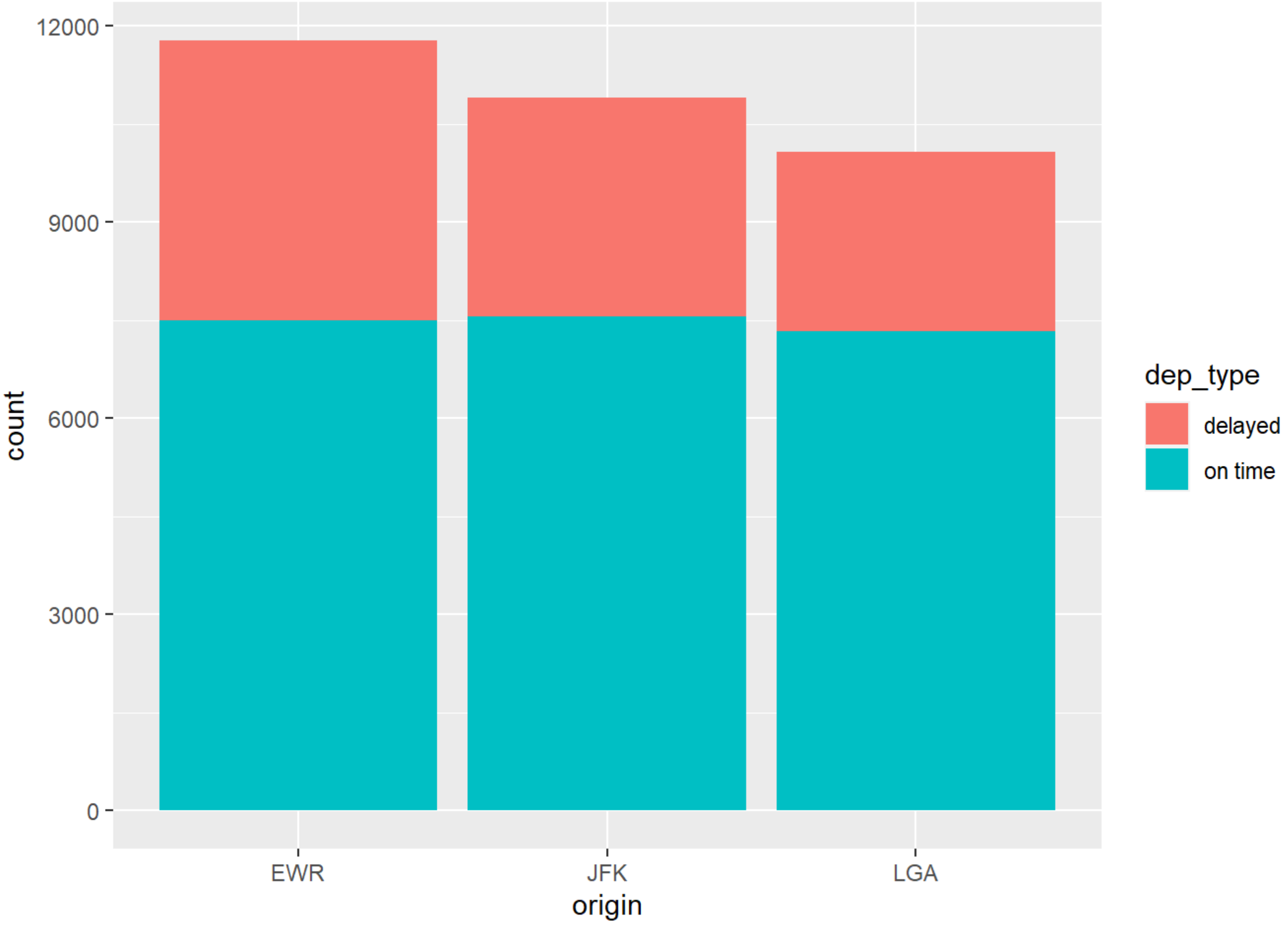
Exercise 6

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))
```

###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.

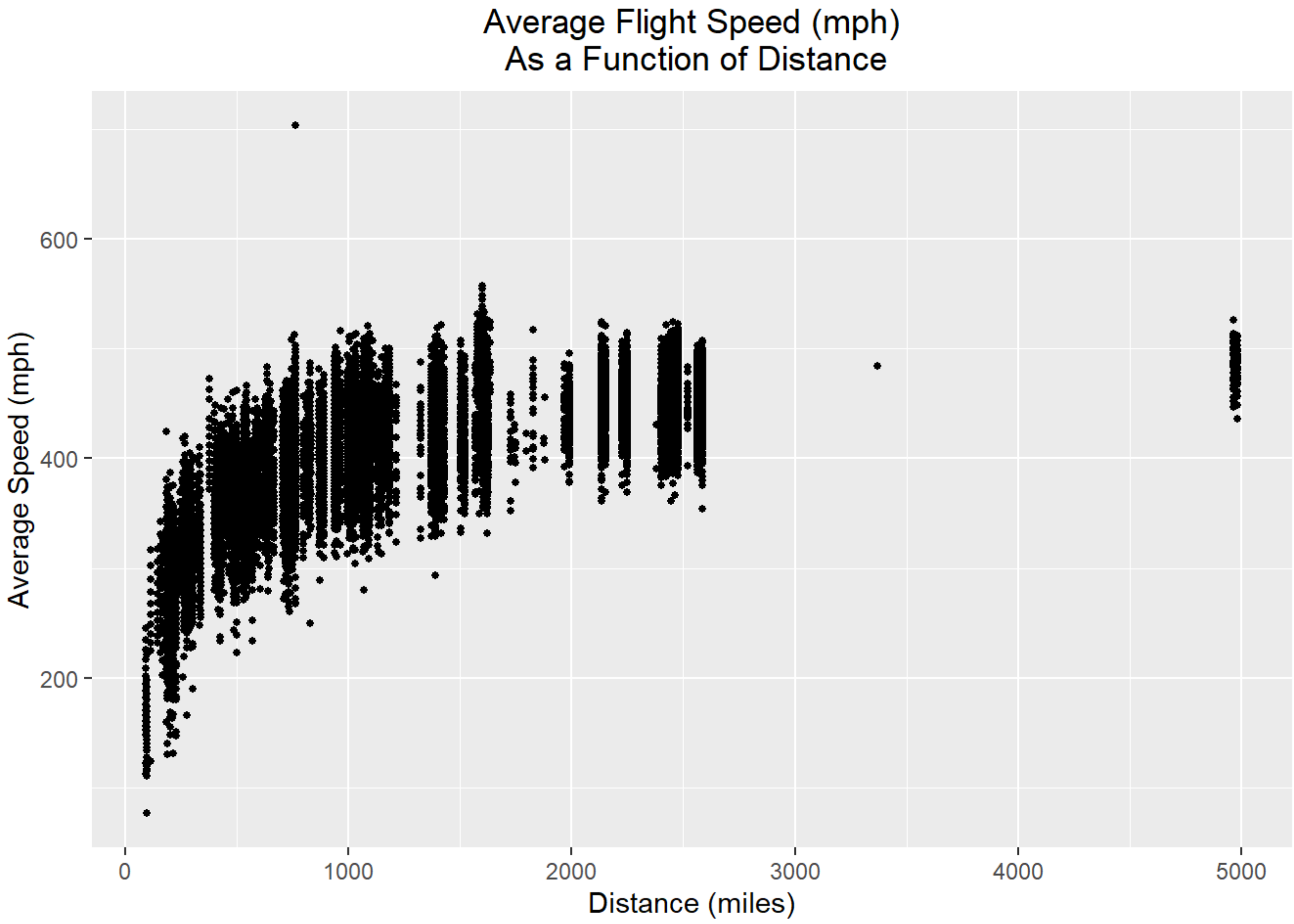
Hide

```
# 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 Speed
          (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.
```

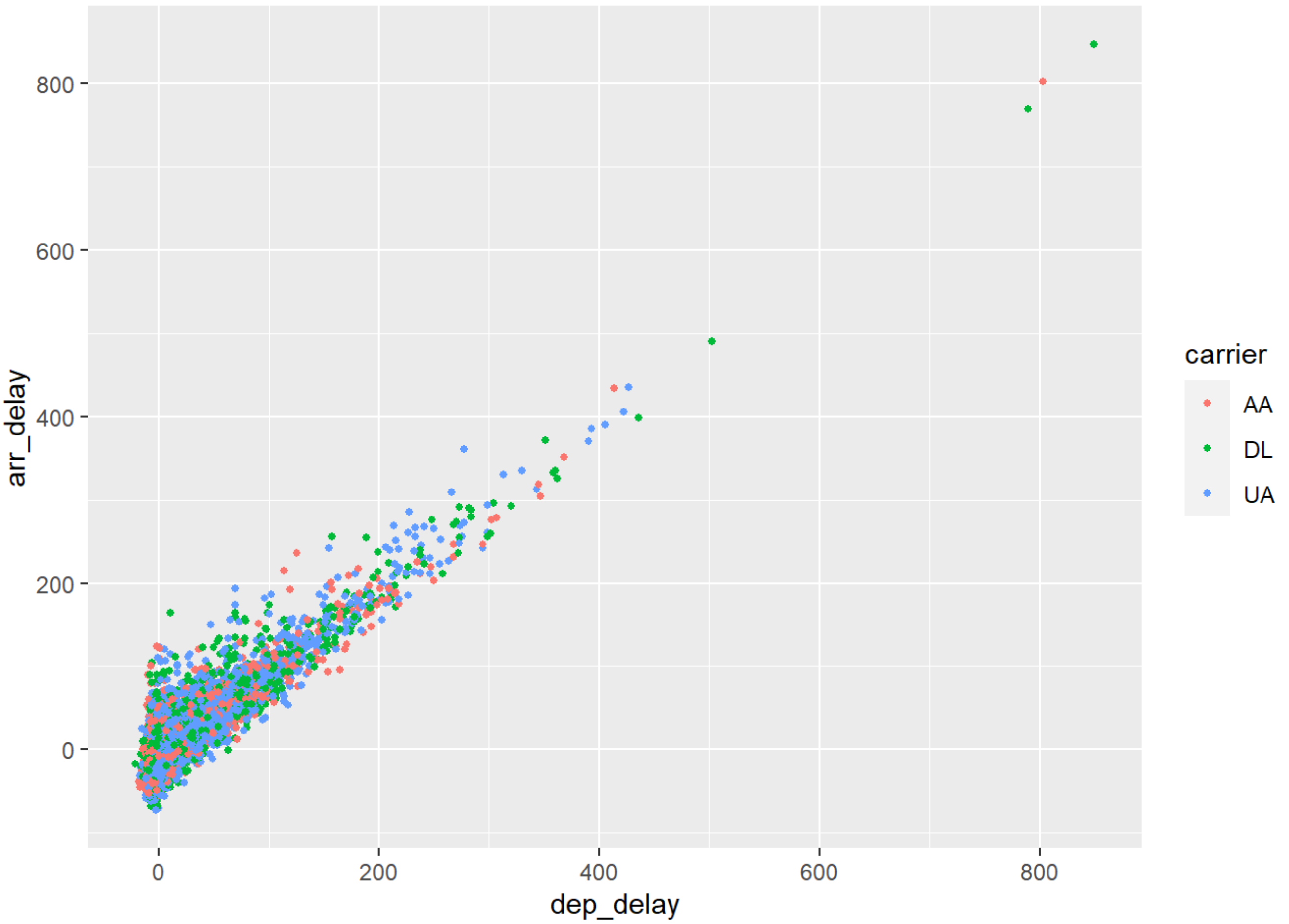
Exercise 9

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
```