Casey Carr
Math 7608
Assignment 3
03/12/2018

1. How many airports in NYC? What are they? Which one has the most flights? Use a bar graph to show the number of flights in each airport.

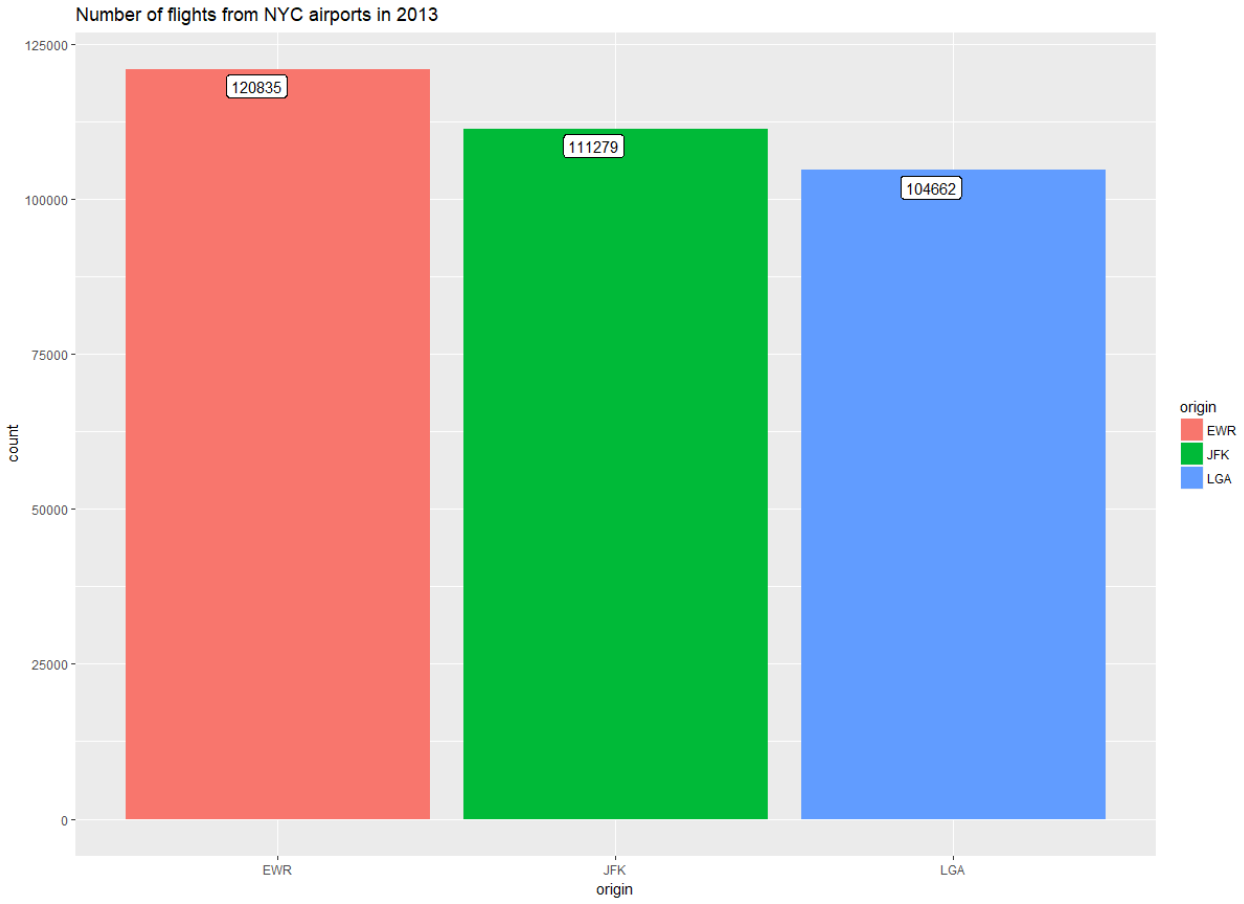
(a) There are 3 airports in NYC. Namely, JFK, LGA, and EWR. EWR has the most flights (surprisingly!)

```
most_flights <- flights %>% group_by(origin) %>% summarise(count = n())
# summarizes the total number of flights from NYC per airport
```

To return **most flights** enter: most_flights[1,2] ➔ return is EWR: 120835 flights

(b)

```
ggplot(data = flights) +
  geom_bar(mapping = aes(x = origin, fill = origin))+
    labs(title = paste("Number of flights from NYC airports in 2013")) +
      ggrepel::geom_label_repel(aes(x = origin, y = count, label = count), data = most_flights)
```



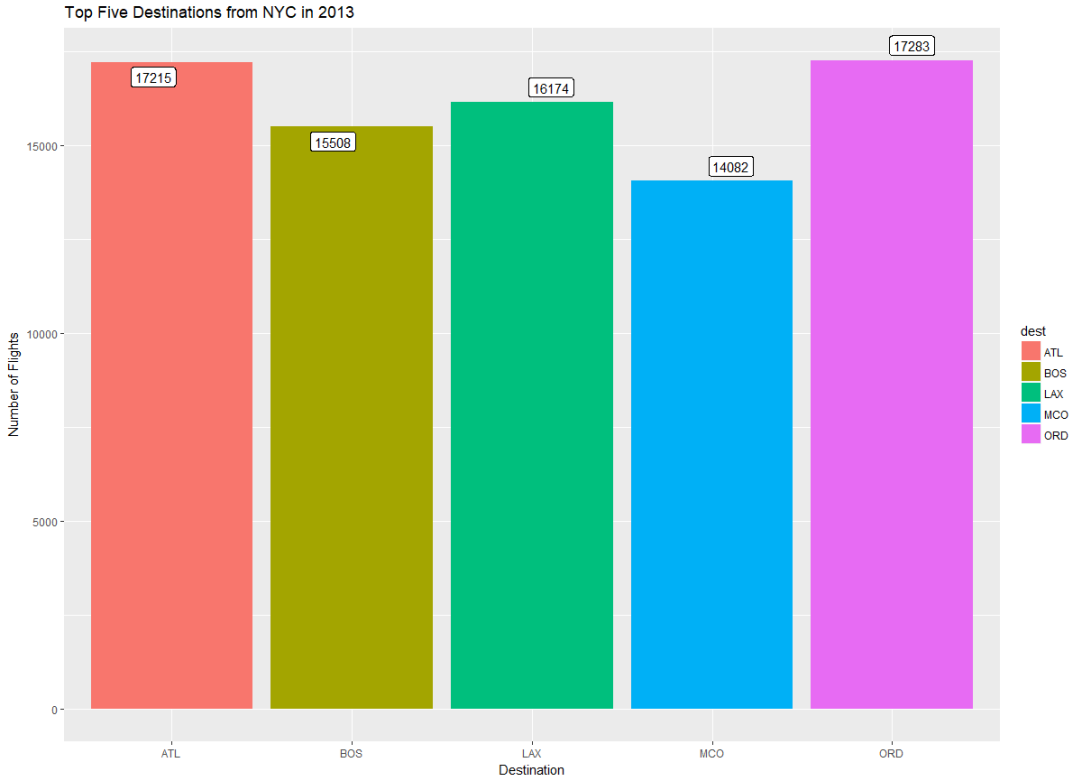
2. Find top five destinations in terms of the number of flights. Use a bar graph to show the number of flights from various NYC airports (origin).

```
top_five <- flights %>% group_by(dest) %>% summarise(count = n()) %>% arrange(desc(count))
top_five <- top_five[1:5,]
```

Extracts the top five destinations by count of number of flights to said destination:

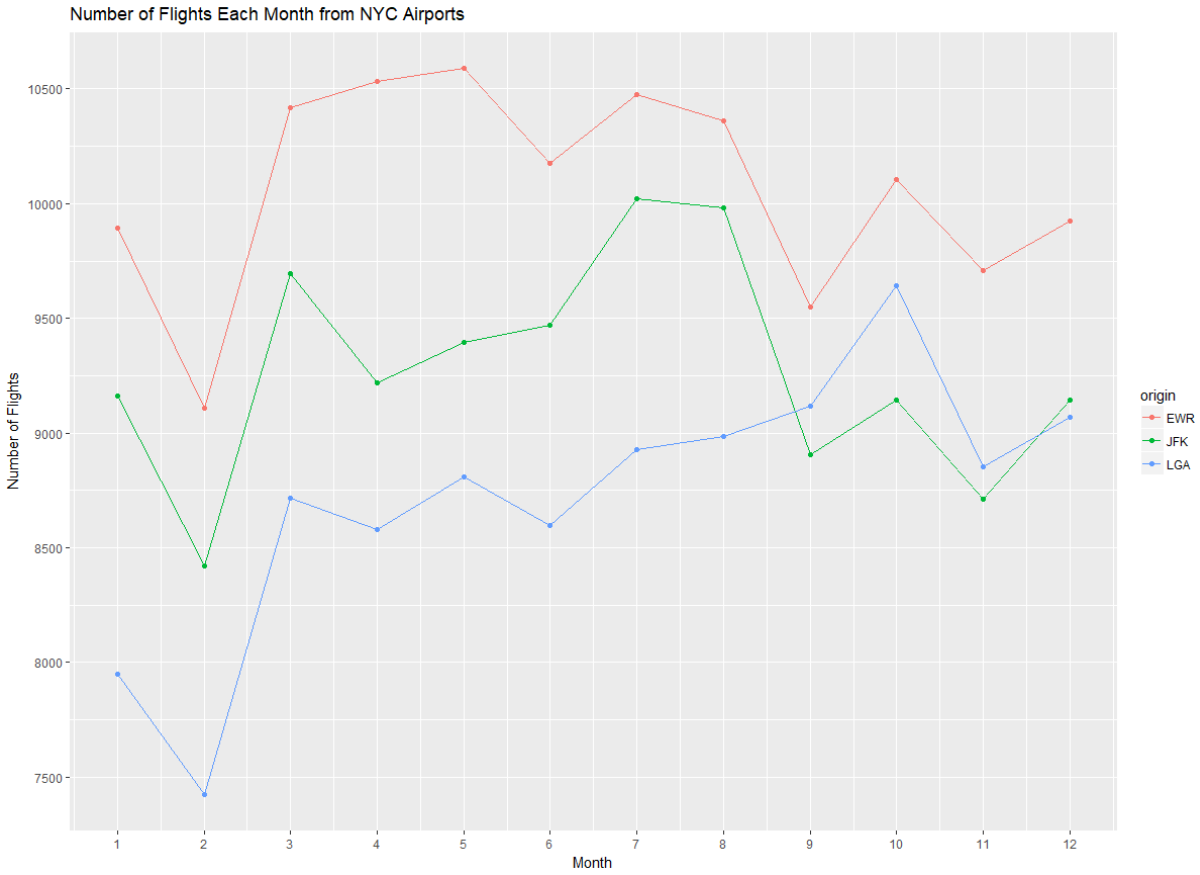
1	ORD	17283
2	ATL	17215
3	LAX	16174
4	BOS	15508
5	MCO	14082

```
top_dest_df <- filter(flights, dest == "ATL" | dest == "BOS" | dest == "LAX" | dest == "MCO" | dest == "ORD")
 ggplot(data = top_dest_df) +
   geom_bar(mapping = aes(x = top_dest_df$dest,  fill = dest)) +
     labs(title = paste("Top Five Destinations from NYC in 2013")) +
       ggrepel::geom_label_repel(aes(x = top_five$dest, y = count, label = count), data = top_five)
```



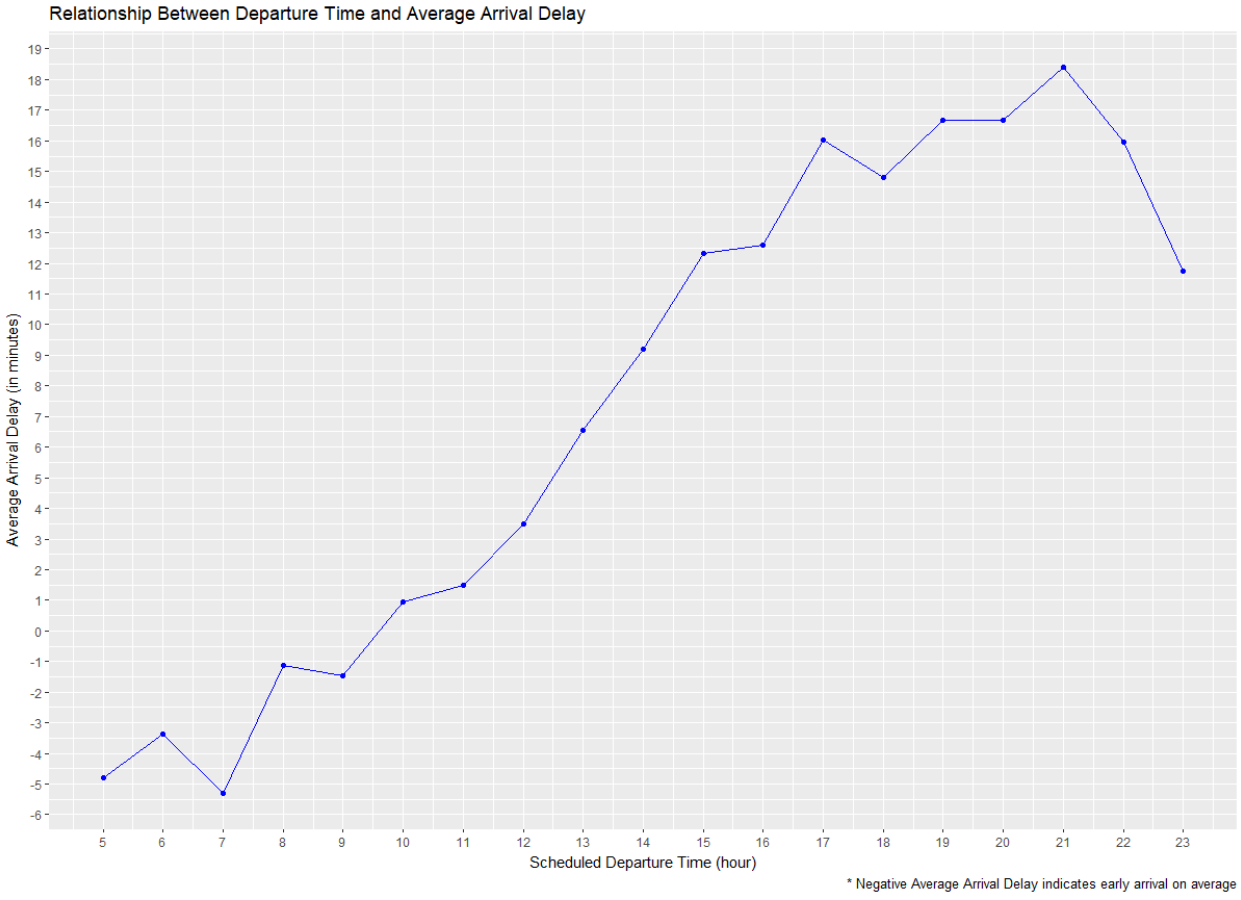
3. Find the number of flights each month. Use a graph to show the number of flights each month from various NYC airports(origin).

```
flights_by_month_df <- flights %>% group_by(month)
 flights_by_month_origin <- flights %>% group_by(month, origin) %>% summarise(count = n())

 ggplot(data = flights_by_month_origin, mapping = aes(x = month, y = count), group = 1 ) +
        # setting group = 1 in aes() ensures that all values are treated as one group
  geom_point(mapping = aes(color = origin)) +
   geom_line(aes(color = origin)) +
    ylab("Number of Flights") + xlab("Month") +
    labs(title = paste("Number of Flights Each Month from NYC Airports")) +
     scale_x_continuous(breaks = seq(1,12, by = 1)) +
      scale_y_continuous(breaks = seq(7000, 12000, by = 500))
```



4. What time of day should you fly if you want to avoid delays as much as possible? Use a graph to show the relationship between its departure time (in hour) and average arrival delays.

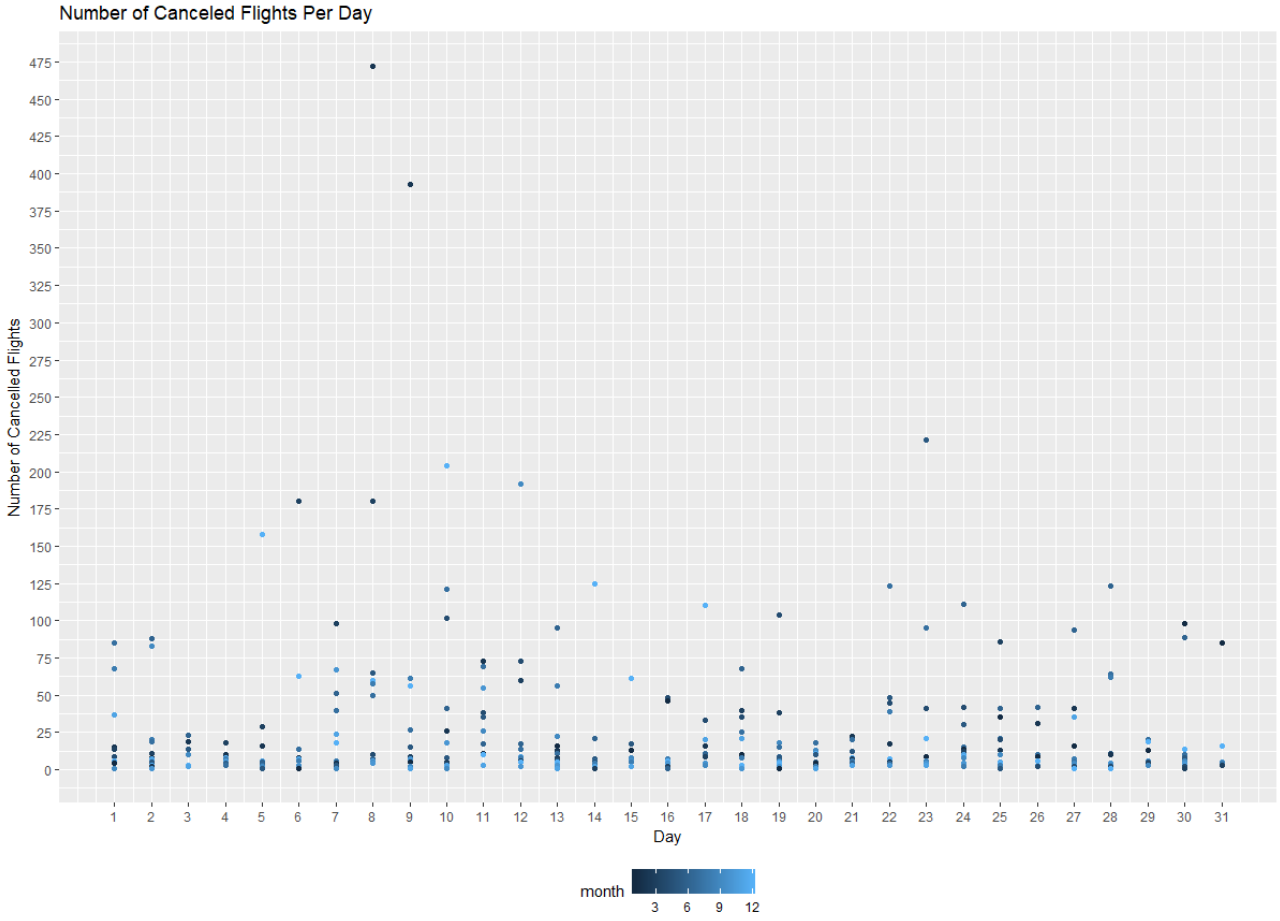
```
schedDepTime_ArrDelay <- tibble(schedDepTime = flights$sched_dep_time %/% 100, arrDelay = flights$arr_delay) %>% arrange(schedDepTime_ArrDelay$schedDepTime)

schedDepTime_avgArrDelay <- schedDepTime_ArrDelay %>% group_by(schedDepTime) %>% summarise(avgArrDelay = round(mean(arrDelay, na.rm = TRUE), 2)) %>% na.omit()
# removes NA i.e. NaN values from output

ggplot(data = schedDepTime_avgArrDelay) +
geom_point(mapping = aes(x = schedDepTime_avgArrDelay$schedDepTime,
        y = schedDepTime_avgArrDelay$avgArrDelay), color = "blue") +
geom_line(mapping = aes(x = schedDepTime_avgArrDelay$schedDepTime, y = schedDepTime_avgArrDelay$avgArrDelay), color = "blue") +
    xlab("Scheduled Departure Time (hour)") + ylab("Average Arrival Delay (in minutes)") +
    labs( title = paste("Relationship Between Departure Time and Average Arrival Delay"),
    caption = "* Negative Average Arrival Delay indicates early arrival on average") +
    scale_x_continuous(breaks = seq(5,23, by = 1)) + scale_y_continuous(breaks = seq(-10, 20, by = 1))
```



Thus, the best time of day to fly to avoid delays is before 10 AM. After 10 AM, delays increase exponentially.

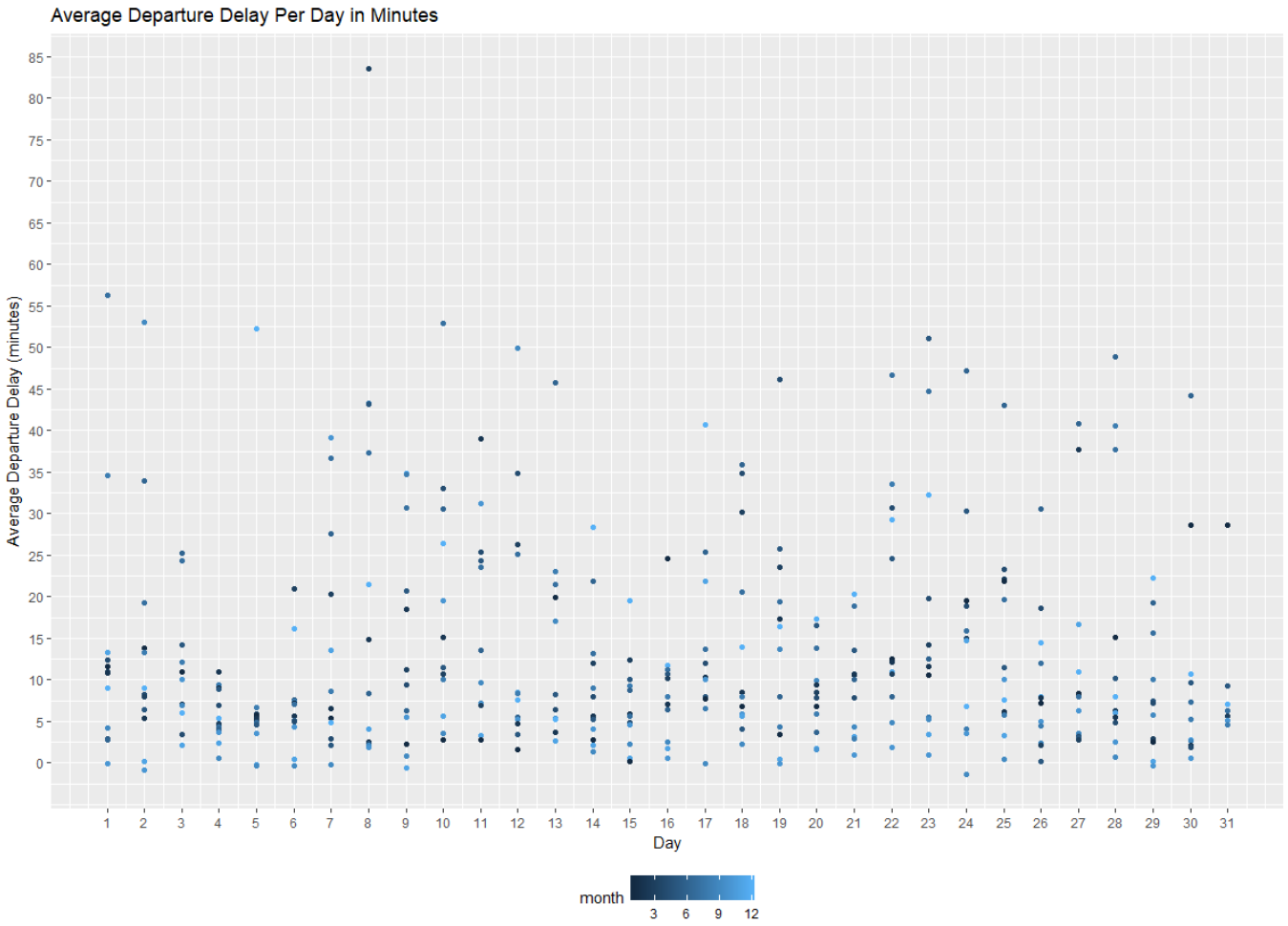
5. Look at the number of cancelled flights per day. Is there a pattern? Is the proportion of cancelled flights related to the average delay? Use a graph to show the relationship.

```
cancelledFlights <- flights %>% select(month, day, dep_delay) %>% filter(is.na(dep_delay)) %>%
group_by(month, day) %>% summarise(numCancelledFlightsPerDay = n()) %>%
arrange(desc(numCancelledFlightsPerDay))

ggplot(data = cancelledFlights) +
  geom_point(mapping = aes(x = cancelledFlights$day,
                           y =  cancelledFlights$numCanceledFlightsPerDay, color = month)) +
  scale_x_continuous(breaks = seq(1,31, by = 1)) + scale_y_continuous(breaks = seq(0, 600, by = 25)) +
  theme(legend.position = "bottom") +
  xlab("Day") + ylab("Number of Cancelled Flights") +
  labs(title = paste("Number of Canceled Flights Per Day"))
```



There seem to be significantly more canceled flights around holidays i.e. Christmas, New Years Eve and there must have been a winter weather storm around February 8 and 9 given the approximate 400 canceled flights on each day.

```
departureDelays <- flights %>% select(month, day, dep_delay) %>% filter(!is.na(dep_delay)) %>%
group_by(month, day) %>% summarise(avgDepartureDelay = round(mean(dep_delay),2)) %>%
arrange(desc(avgDepartureDelay))

 ggplot(data = departureDelays) +
  geom_point(mapping = aes(x = departureDelays$day, y = departureDelays$avgDepartureDelay,
          color = month)) +
  scale_x_continuous(breaks = seq(1,31, by = 1)) + scale_y_continuous(breaks = seq(0, 100, by = 5)) +
  theme(legend.position = "bottom") + guides(color = guide_colorbar(order = 1)) +
  xlab("Day") + ylab("Average Departure Delay (minutes)") +
  labs(title = paste("Average Departure Delay Per Day in Minutes"))
```



It makes since that days on which there are the highest number of flight cancellations, there are also the highest average departure delays. For example, February 8 has the highest average departure delay (83 minutes) and the highest number of cancellations (472).

6. Which carrier has the worst delays? Challenge: can you disentangle the effects of bad airports vs. bad carriers? Why/why not?

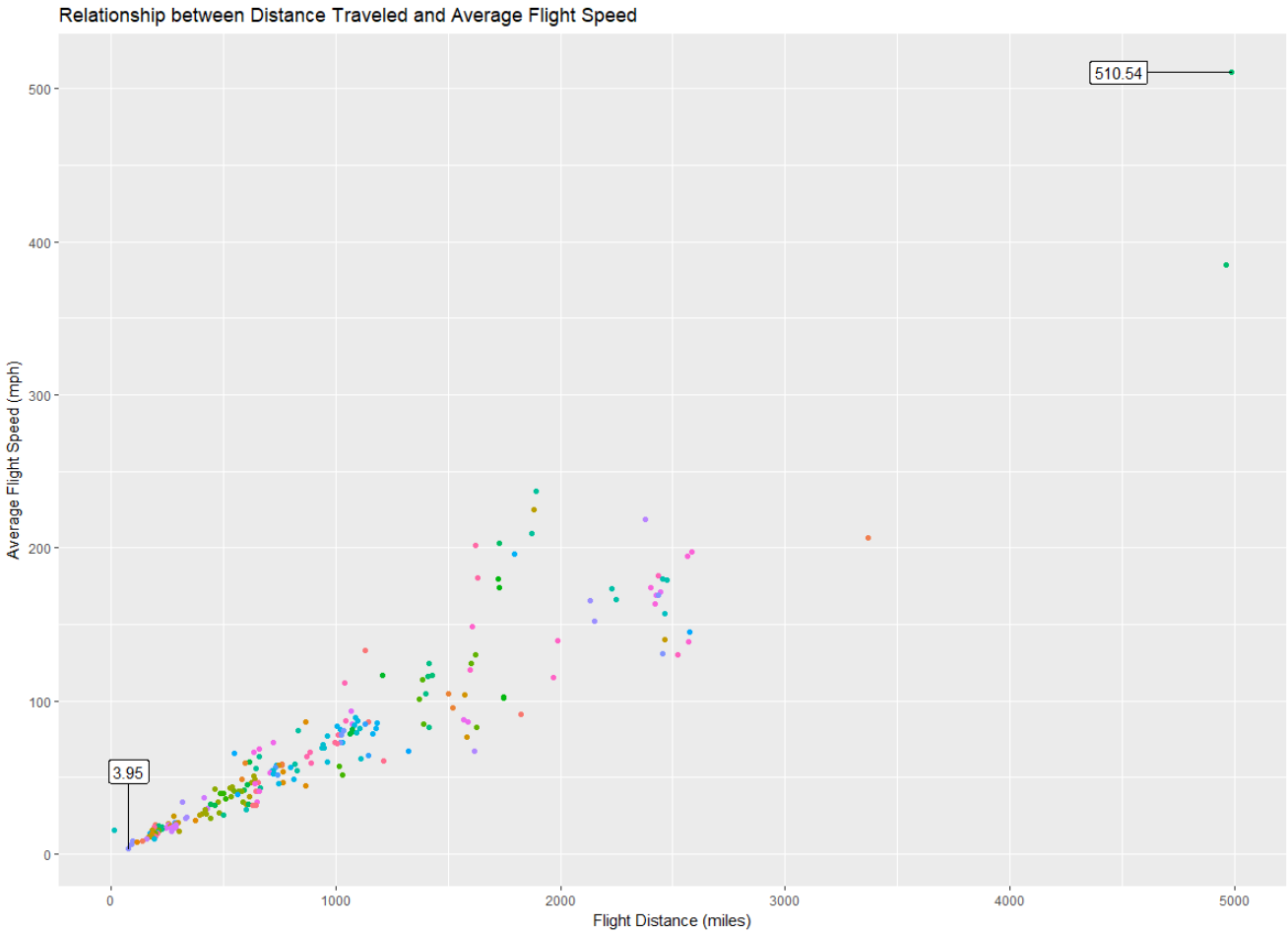
It is unclear how we can manage to disentangle the effects of bad airports vs. bad carriers as there is no direct correlation between Average Departure Delay and carriers, airports. There must be other factors involved such as weather, time of departure, and distance traveled.
For example, the carrier DL has an average departure delay of 168.77 minutes when traveling to ATL, whereas, when traveling to PHL, DL has an average departure delay of only 5.00 minutes.

```
delay <- flights %>% group_by(carrier, dest, dep_delay) %>% summarise(n()) %>% filter(dep_delay > 0)

delayAvg <- delay %>% group_by(carrier, dest) %>% summarise(AvgDepDelay =
round(mean(dep_delay),2)) %>% arrange(desc(AvgDepDelay))
avgCarrierDelays <- delay %>% group_by(carrier) %>% summarise(AvgDepDelay =
round(mean(dep_delay),2)) %>% arrange((desc(AvgDepDelay)))
avgDestinationDelays <- delay %>% group_by(dest) %>% summarise(AvgDepDelay =
round(mean(dep_delay),2)) %>% arrange((desc(AvgDepDelay)))
```

7. Which flights traveled the longest (in terms of distance)? Which traveled the shortest? Identify the destinations. Is there a relationship between the travel distance and its speed (mph, miles per hour)? Use a graph to show the relationship.

longestFlights <- flights %>% group_by(flight,dest) %>% summarise(AvgDistance = round(mean(distance),2)) %>% arrange(desc(AvgDistance))

flights <- mutate(flights, flightTime = round(hour + (minute / 60),2))
flightsTime <- flights %>% group_by(dest, distance) %>% summarise(AvgFlightTime = mean(flightTime)) %>% arrange(desc(AvgFlightTime))
flightsMPH <- flightsTime %>% group_by(dest, distance) %>% summarise(AvgFlightSpeed = round(distance / AvgFlightTime,2)) %>% arrange(desc(AvgFlightSpeed))
maxAvgFlightSpeed <- flightsMPH[1,]
minAvgFlightSpeed <- flightsMPH[length(flightsMPH$AvgFlightSpeed),]



Flights 15 and 51 traveled the longest distances of 4963 and 4983 miles, respectively. Flight 1632 to LGA traveled the shortest distance of 17 miles. There is a clear relationship between travel distance and average flight speed. Namely, the shorter the flight distance the lower the average speed, and the longer the flight distance the higher the average speed.