INFX 573: Problem Set 1 - Exploring Data

Shuyang Wu

Due: Monday, October 11, 2016

Collaborators:

Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset1.Rmd` file from Canvas. Open `problemset1.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset1.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text.
4. Collaboration on problem sets is acceptable, and even encouraged, but each student must turn in an individual write-up in his or her own words and his or her own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the R Markdown file to `YourLastName_YourFirstName_ps1.Rmd`, knit a PDF and submit both the PDF file on Canvas.

Setup:

In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(nycflights13)
library(ggplot2)
```

Problem 1: Exploring the NYC Flights Data

In this problem set we will use the data on all flights that departed NYC (i.e. JFK, LGA or EWR) in 2013. You can find this data in the `nycflights13` R package.

(a) Importing and Inspecting Data:

Load the data and describe in a short paragraph how the data was collected and what each variable represents. Perform a basic inspection of the data and discuss what you find.

```
# Inspecting data
head(flights) # did the same for weather, airlines, and airports
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1  2013     1     1      517            515         2      830
## 2  2013     1     1      533            529         4      850
## 3  2013     1     1      542            540         2      923
## 4  2013     1     1      544            545        -1     1004
## 5  2013     1     1      554            600        -6      812
## 6  2013     1     1      554            558        -4      740
## # ... with 12 more variables: sched_arr_time <int>, arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,
## #   time_hour <time>
```

```
tail(flights)
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1  2013     9    30       NA           1842        NA       NA
## 2  2013     9    30       NA           1455        NA       NA
## 3  2013     9    30       NA           2200        NA       NA
## 4  2013     9    30       NA           1210        NA       NA
## 5  2013     9    30       NA           1159        NA       NA
## 6  2013     9    30       NA            840        NA       NA
## # ... with 12 more variables: sched_arr_time <int>, arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,
## #   time_hour <time>
```

```
str(flights)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    336776 obs. of  19 variables:
##  $ year          : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
##  $ month         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ day           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ dep_time      : int  517 533 542 544 554 554 555 557 557 558 ...
##  $ sched_dep_time: int  515 529 540 545 600 558 600 600 600 600 ...
##  $ dep_delay     : num  2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
##  $ arr_time      : int  830 850 923 1004 812 740 913 709 838 753 ...
##  $ sched_arr_time: int  819 830 850 1022 837 728 854 723 846 745 ...
##  $ arr_delay     : num  11 20 33 -18 -25 12 19 -14 -8 8 ...
##  $ carrier       : chr  "UA" "UA" "AA" "B6" ...
##  $ flight        : int  1545 1714 1141 725 461 1696 507 5708 79 301 ...
##  $ tailnum       : chr  "N14228" "N24211" "N619AA" "N804JB" ...
##  $ origin        : chr  "EWR" "LGA" "JFK" "JFK" ...
##  $ dest          : chr  "IAH" "IAH" "MIA" "BQN" ...
##  $ air_time      : num  227 227 160 183 116 150 158 53 140 138 ...
##  $ distance      : num  1400 1416 1089 1576 762 ...
##  $ hour          : num  5 5 5 5 6 5 6 6 6 6 ...
##  $ minute        : num  15 29 40 45 0 58 0 0 0 0 ...
##  $ time_hour     : POSIXct, format: "2013-01-01 05:00:00" "2013-01-01 05:00:00" ...
```

```r
summary(flights) #look at min, mean, max of all quantitative columns, did the same for weather
```

```
##       year          month             day           dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907
##  Median :2013   Median : 7.000   Median :16.00   Median :1401
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400
##                                                  NA's   :8255
##  sched_dep_time   dep_delay          arr_time    sched_arr_time
##  Min.   : 106   Min.   : -43.00   Min.   :   1   Min.   :   1
##  1st Qu.: 906   1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124
##  Median :1359   Median :  -2.00   Median :1535   Median :1556
##  Mean   :1344   Mean   :  12.64   Mean   :1502   Mean   :1536
##  3rd Qu.:1729   3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945
##  Max.   :2359   Max.   :1301.00   Max.   :2400   Max.   :2359
##                 NA's   :8255      NA's   :8713
##    arr_delay          carrier              flight       tailnum
##  Min.   : -86.000   Length:336776      Min.   :   1   Length:336776
##  1st Qu.: -17.000   Class :character   1st Qu.: 553   Class :character
##  Median :  -5.000   Mode  :character   Median :1496   Mode  :character
##  Mean   :   6.895                      Mean   :1972
##  3rd Qu.:  14.000                      3rd Qu.:3465
##  Max.   :1272.000                      Max.   :8500
##  NA's   :9430
##     origin              dest              air_time        distance
##  Length:336776      Length:336776      Min.   : 20.0   Min.   :  17
##  Class :character   Class :character   1st Qu.: 82.0   1st Qu.: 502
##  Mode  :character   Mode  :character   Median :129.0   Median : 872
##                                        Mean   :150.7   Mean   :1040
##                                        3rd Qu.:192.0   3rd Qu.:1389
##                                        Max.   :695.0   Max.   :4983
##                                        NA's   :9430
##       hour            minute        time_hour
##  Min.   : 1.00   Min.   : 0.00   Min.   :2013-01-01 05:00:00
##  1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
##  Median :13.00   Median :29.00   Median :2013-07-03 10:00:00
##  Mean   :13.18   Mean   :26.23   Mean   :2013-07-03 05:02:36
##  3rd Qu.:17.00   3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00
##  Max.   :23.00   Max.   :59.00   Max.   :2013-12-31 23:00:00
##
```

```r
summary(weather)
```

```
##     origin               year          month             day
##  Length:26130       Min.   :2013   Min.   : 1.000   Min.   : 1.00
##  Class :character   1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00
##  Mode  :character   Median :2013   Median : 7.000   Median :16.00
##                     Mean   :2013   Mean   : 6.506   Mean   :15.68
##                     3rd Qu.:2013   3rd Qu.: 9.000   3rd Qu.:23.00
##                     Max.   :2013   Max.   :12.000   Max.   :31.00
```

```
##
##       hour             temp             dewp           humid
##  Min.   : 0.00   Min.   : 10.94   Min.   :-9.94   Min.   : 12.74
##  1st Qu.: 6.00   1st Qu.: 39.92   1st Qu.:26.06   1st Qu.: 46.99
##  Median :12.00   Median : 55.04   Median :42.08   Median : 61.66
##  Mean   :11.52   Mean   : 55.20   Mean   :41.39   Mean   : 62.35
##  3rd Qu.:18.00   3rd Qu.: 69.98   3rd Qu.:57.92   3rd Qu.: 78.62
##  Max.   :23.00   Max.   :100.04   Max.   :78.08   Max.   :100.00
##                  NA's   :1        NA's   :1       NA's   :1
##     wind_dir       wind_speed         wind_gust            precip
##  Min.   :  0.0   Min.   :   0.000   Min.   :   0.000   Min.   :0.000000
##  1st Qu.:120.0   1st Qu.:   6.905   1st Qu.:   7.946   1st Qu.:0.000000
##  Median :220.0   Median :   9.206   Median :  10.594   Median :0.000000
##  Mean   :198.1   Mean   :  10.396   Mean   :  11.963   Mean   :0.002726
##  3rd Qu.:290.0   3rd Qu.:  13.809   3rd Qu.:  15.892   3rd Qu.:0.000000
##  Max.   :360.0   Max.   :1048.361   Max.   :1206.432   Max.   :1.180000
##  NA's   :418     NA's   :3          NA's   :3
##     pressure          visib          time_hour
##  Min.   : 983.8   Min.   : 0.000   Min.   :2012-12-31 16:00:00
##  1st Qu.:1012.9   1st Qu.:10.000   1st Qu.:2013-04-01 14:00:00
##  Median :1017.6   Median :10.000   Median :2013-07-01 07:30:00
##  Mean   :1017.9   Mean   : 9.205   Mean   :2013-07-01 12:07:20
##  3rd Qu.:1023.0   3rd Qu.:10.000   3rd Qu.:2013-09-30 07:45:00
##  Max.   :1042.1   Max.   :10.000   Max.   :2013-12-30 15:00:00
##  NA's   :2730
```

There are four datasets in the nycflight13 library. Flights contains information on the date, time, time delayed and location of both departure and arrival, also flight number, air time and distance. The weather dataset has information on date, time, temperature, humidity, wind pressure and visibility. Airports and airlines datasets serves more as an description or explaination of the abbreviations in the flights and weather datasets. The flight dataset seems more interesting because delayed time is highly skewed. But there is quite a few missing data under the departure time, departure delay, arrival time, arrival delay and airtime columns, represented by NAs.

(b) Formulating Questions:

Consider the NYC flights data. Formulate two motivating questions you want to explore using this data. Describe why these questions are interesting and how you might go about answering them.

Question 1: Is there any pattern with the number of flights leaving NYC during the year? I'm interested in finding out if there is a particular busy time, perhaps during holidays. I plan on group the flights dataset by days and count the flights leaving NYC each day, possibly also each month and then visualize it.

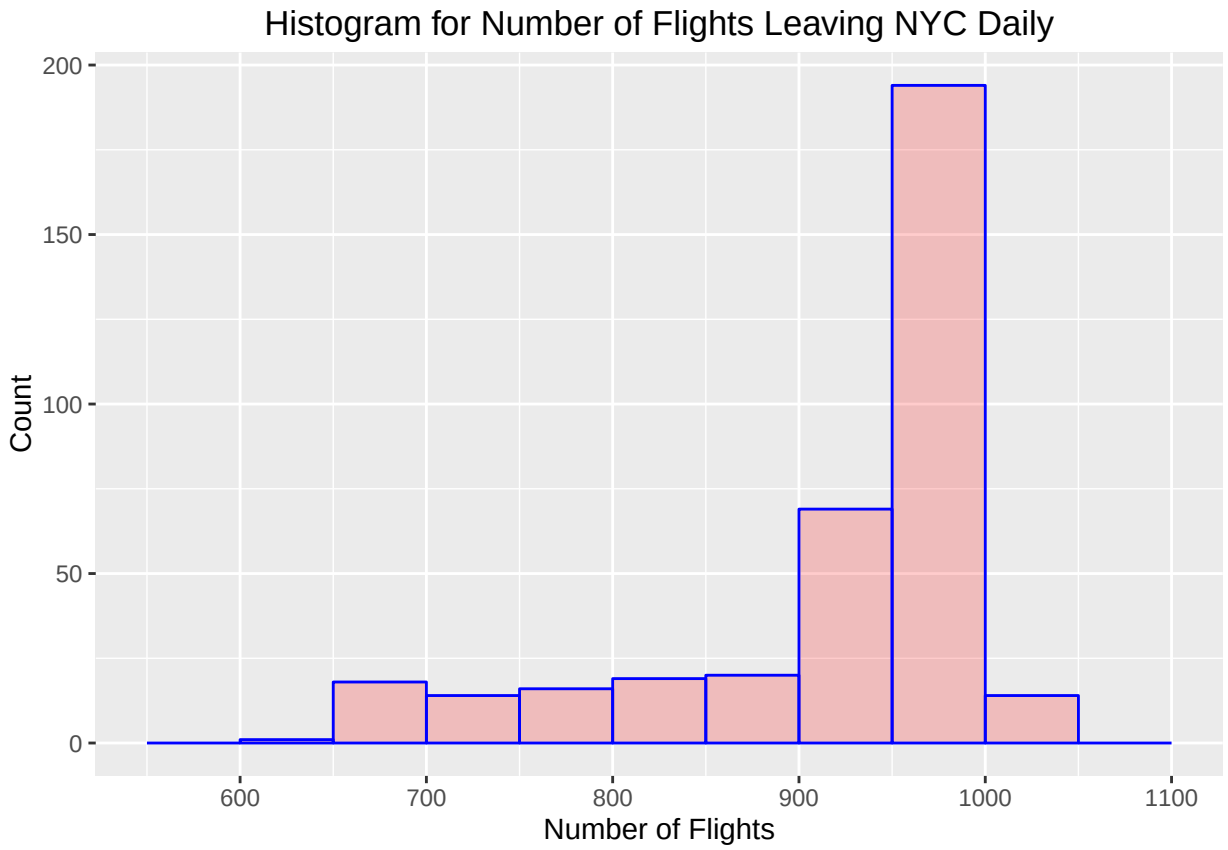
Question 2: Is wind speed and/or visibility correlated with more delayed departure or arrival? I'm interested in predicting flight delay time given weather information, but I want to try the easy solution first and see if wind speed and visibility alone has any relationship with delay times. I plan on subsetting the delayed (>30 min) flights and then adding the wind speed and visibility at corresponding times and locations.

(c) Exploring Data:

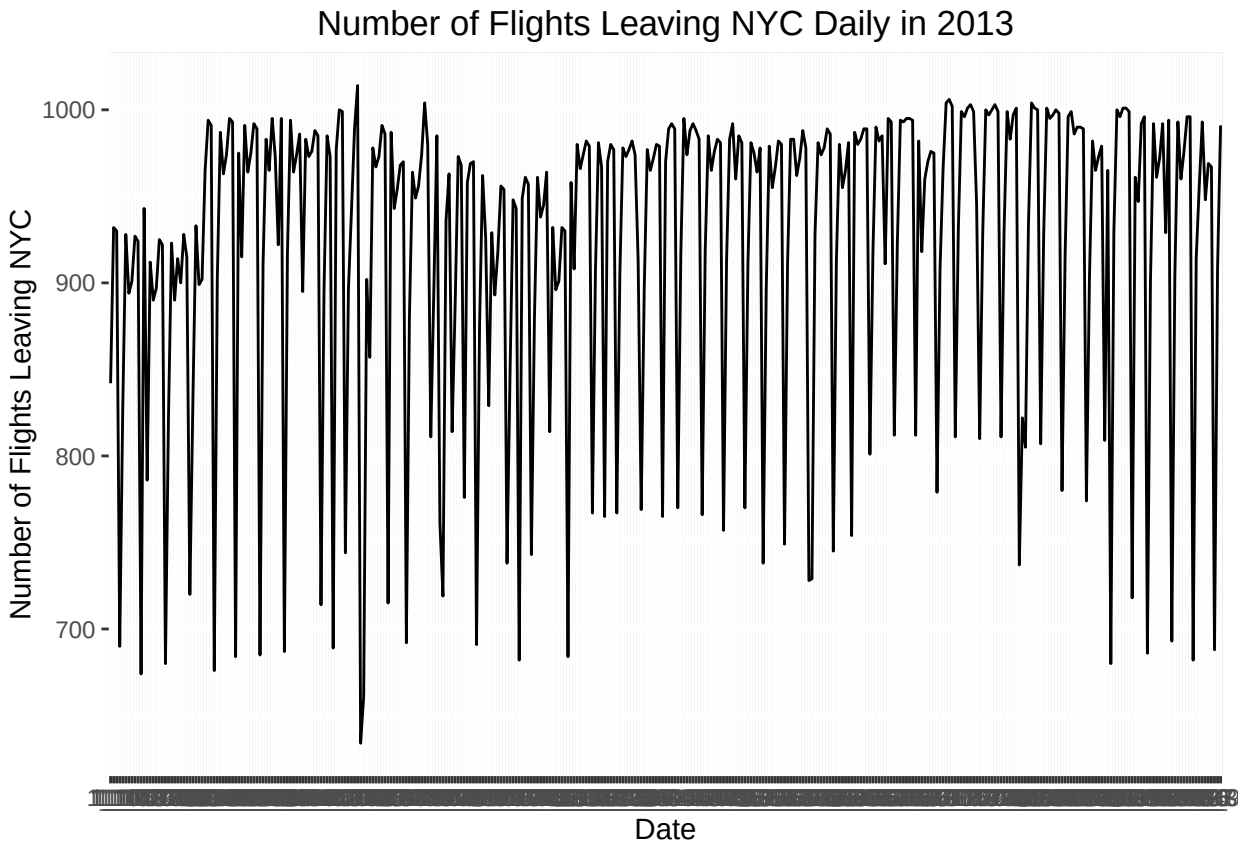
For each of the questions you proposed in Problem 1b, perform an exploratory data analysis designed to address the question. At a minimum, you should produce two visualizations related to each question. Be sure to describe what the visuals show and how they speak to your question of interest.

```
### Exploratory data analysis for Q1
daily <- group_by(flights, year, month, day) #group flights by day
flightperday   <- summarise(daily, flights = n()) #number of flights per day
flightperday$date = paste(flightperday$month, flightperday$day, sep="_") #dates in 2013

#Histogram of number of flights:
ggplot(data=flightperday, aes(flightperday$flights)) +
  geom_histogram(breaks=seq(550, 1100, by =50),
                 col="blue",
                 fill="red",
                 alpha = .2) +
  labs(title="Histogram for Number of Flights Leaving NYC Daily") +
  labs(x="Number of Flights", y="Count")
```

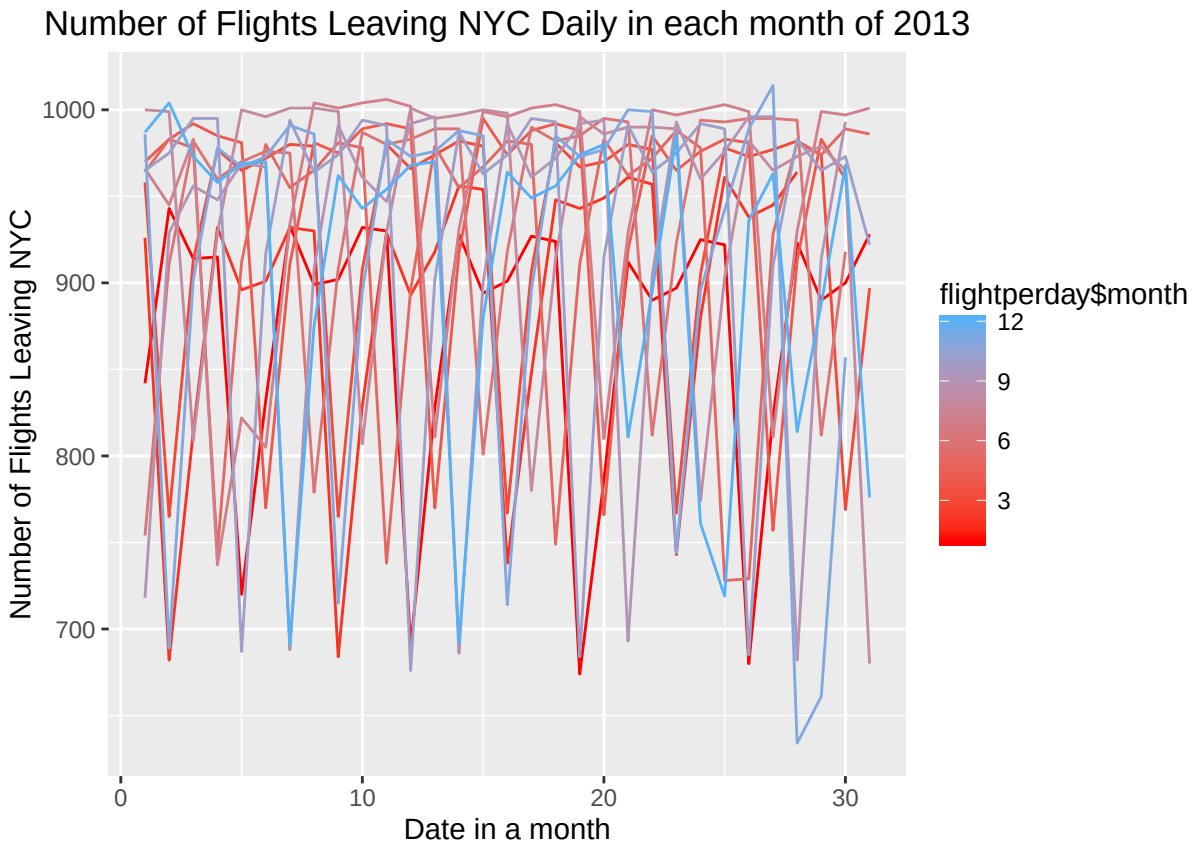



```
#Plot of number of flights daily in 2013
ggplot(data=flightperday, aes(x = flightperday$date, y = flightperday$flights, group = 1)) +
        geom_line() +
        labs(x = "Date", y = "Number of Flights Leaving NYC",
             title = "Number of Flights Leaving NYC Daily in 2013")
```




```
#Plot of number of flights monthly in 2013
flightpermonth <- summarise(flightperday, flights = sum(flights)) #number of flights per month
ggplot(data=flightpermonth, aes(x = flightpermonth$month, y = flightpermonth$flights, group = 1)) +
        geom_line() +
        labs(x = "Month", y = "Number of Flights Leaving NYC",
             title = "Number of Flights Leaving NYC Monthly in 2013")
```

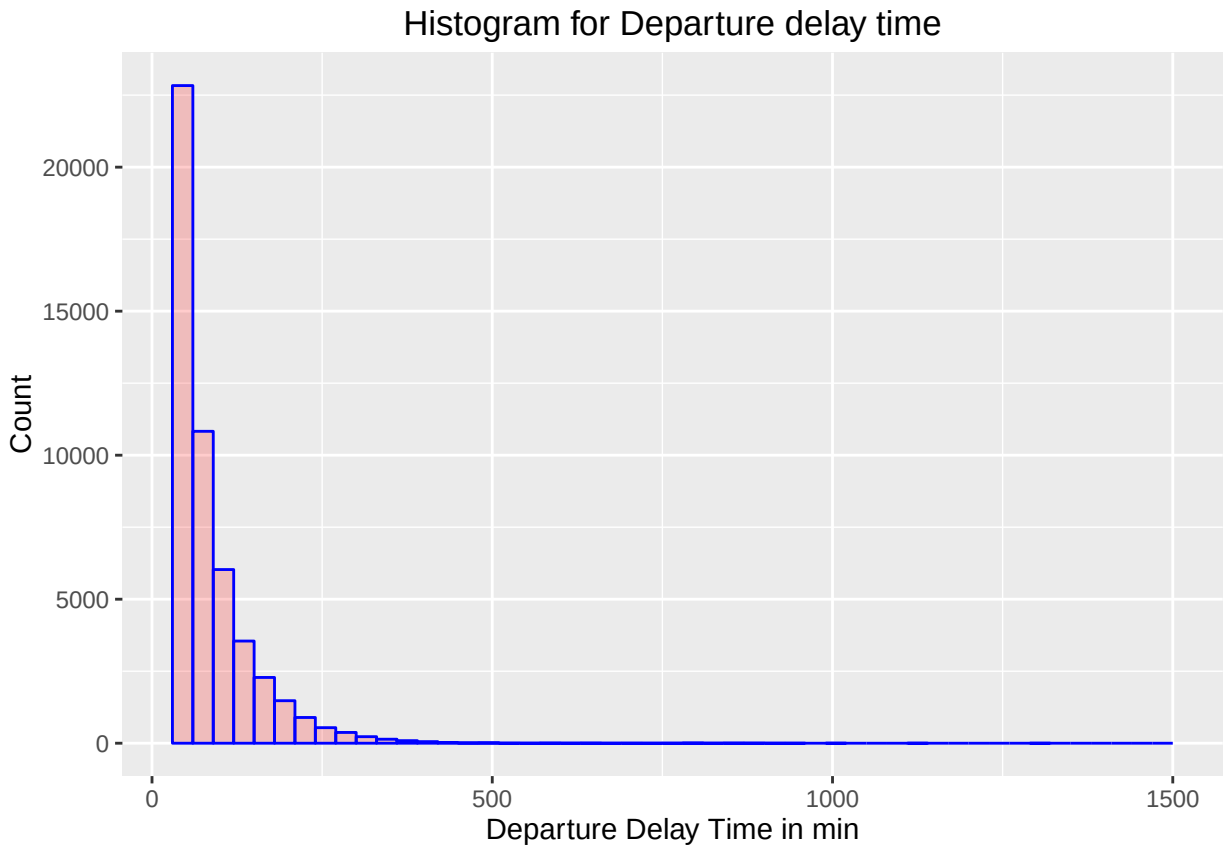

```
#Plot of number of flights daily in each month (overlapping to see if there is any pattern)
ggplot(data=flightperday, aes(x = flightperday$day, y = flightperday$flights, group = flightperday$month
  geom_line(aes(colour = flightperday$month)) + scale_colour_gradient(low="red") +
  labs(x = "Date in a month", y = "Number of Flights Leaving NYC",
       title = "Number of Flights Leaving NYC Daily in each month of 2013")
```



From the histogram for number of flights, we can see that the values mostly fall into 950~1000 range, but there are also a wide distribution of values from 600 to 1050. From the plot of number of flights daily, the numbers of flights drop by two to three hundred in every couple of days and then rise back up. Such pattern continues throughout the year. The first two months of 2013 have fewer numbers of flights on average and the following three to four months have more variation (drop in number of flights) than in the summer months. The plot of number of flights monthly also shows that the number of flights leaving NYC is the lowest during January and February. Plot of number of flights daily in each month overlapped shows that the variation pattern in each month is very similar although not completely the same.

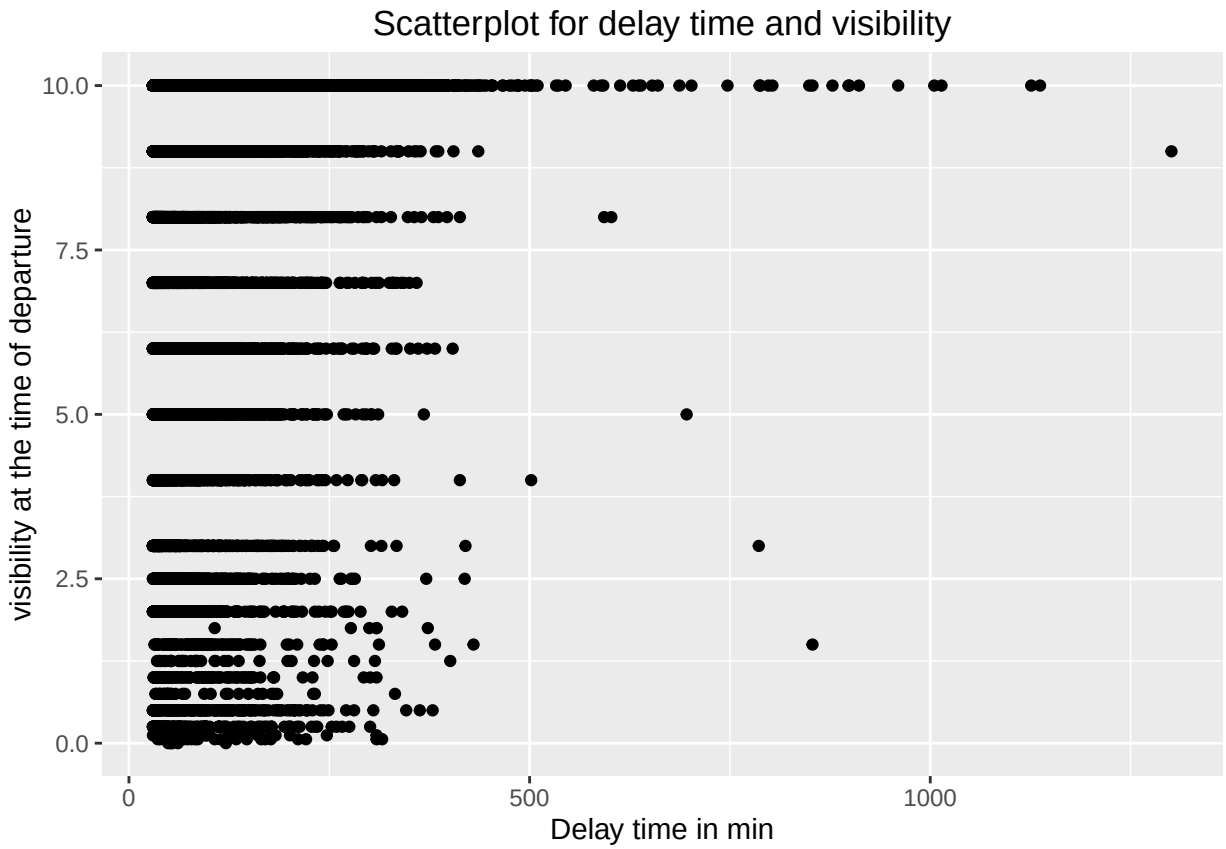
```
### Exploratory data analysis for Q2
#subset delayed flights to be departure_delay >= 30 min
delay<- subset(flights, flights$dep_delay >= 30, select=year:time_hour)

#histogram of delayed departure time
ggplot(data=delay, aes(delay$dep_delay)) +
  geom_histogram(breaks=seq(30, 1500, by =30),
                 col="blue",
                 fill="red",
                 alpha = .2) +
  labs(title="Histogram for Departure delay time") +
  labs(x="Departure Delay Time in min", y="Count")
```

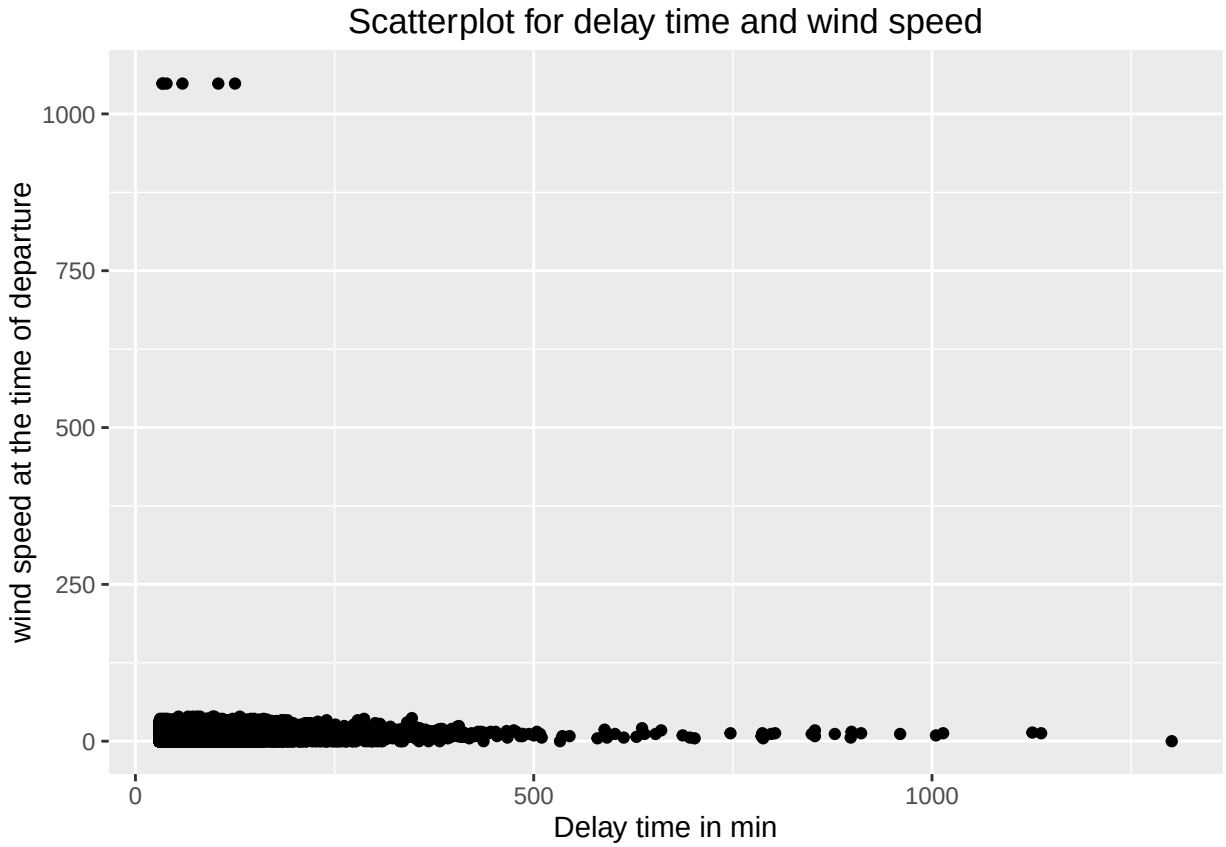



```
#merge flights with weather based on the same time_hour and origin
merged <- merge(delay, weather, by = c("time_hour", "origin"))
merged <- na.omit(merged) # remove rows containing NAs

#relationship between visibility and delay departure time
ggplot(data=merged, aes(x = merged$dep_delay, y = merged$visib)) +
        geom_point() +
        labs(x = "Delay time in min", y = "visibility at the time of departure",
             title = "Scatterplot for delay time and visibility")
```




```
#relationship between wind speed and delay departure time
ggplot(data=merged, aes(x = merged$dep_delay, y = merged$wind_speed)) +
        geom_point() +
        labs(x = "Delay time in min", y = "wind speed at the time of departure",
             title = "Scatterplot for delay time and wind speed")
```



From the histogram I see that the delay time is highly skewed with the majority of delay times being in 30 mins to an hour. From the plot of visibiliy and delay time, data does not show enough evidence that visibility and delay time is negatively correlated as I had hypothesized. In fact based on the plot, longer delay time is associated with higher visibility. From the plot of wind speed and delay time, firstly there are a few outliers for the very high wind speed but it's not really correlated with longer delay time. Besides the outliners, the rest of the wind speeds are mostly below 50 and account for most of the delay cases. Data does not suggest any strong correlation between wind speed and delay time.

(d) Challenge Your Results:

After completing the exploratory analysis from Problem 1c, do you have any concerns about your findings? Comment on any ethical and/or privacy concerns you have with your analysis.

For my first question and analysis, I think privacy of people's traveling patterns could be at risk. In an unfortunate case, leak of the flights dataset and information on patterns of number of flights leaving NYC to terrorists could result in them planning attacks at the most crowded season/times at the airports. As for my second question and analysis, because I did not find any significant correlation between my perceived predictor and delay time, I wonder what could be the reason for some of the relatively long delays. If there is any pattern other than the weather being discovered from the datasets, for example, particular airline or maintenance team, there might be conflict of interests or privacy violations if accusations were to be made to those companies.