INFX 573: Problem Set 1 - Exploring Data

TAPASVI BANSAL

Due: Thursday, October 12, 2017

Collaborators: Ishaan Srivastava

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 the PDF file on Canvas.

stress more visualization, dplyr, less questions/ethics, etc

Setup:

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

```
# Load standard libraries
library("tidyverse")
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
library("nycflights13")
```

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.

```
library(nycflights13)
library (magrittr)
```

```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library (dplyr)
library(tidyverse)
?nycflights13::flights
?nycflights13::airports
destsea2 <- structure(list())
head(nycflights13::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     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 <dttm>
```

```
View(nycflights13::planes)
table(nycflights13::planes$speed)
```

```
##
##  90  95 105 107 108 112 126 127 162 167 202 232 432
##   2   1   2   1   1   1   1   1   2   1   1   1   8
```

```
View(nycflights13::weather)
table(nycflights13::weather$origin)
```

```
##
##  EWR  JFK  LGA
## 8708 8711 8711
```

```
nycflights13::flights %>% filter(dest == "SEA" & month == 12)
```

```
## # A tibble: 293 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013    12     1      700            705        -5     1005
##  2  2013    12     1      740            745        -5     1037
##  3  2013    12     1      848            845         3     1202
##  4  2013    12     1     1459           1500        -1     1813
##  5  2013    12     1     1832           1825         7     2200
##  6  2013    12     1     1849           1849         0     2208
##  7  2013    12     1     1854           1759        55     2230
```

```
##  8  2013    12     1     1922           1820        62     2222
##  9  2013    12     1     1933           1909        24     2255
## 10  2013    12     1     2011           1850        81     2331
## # ... with 283 more rows, and 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 <dttm>
```

The package data describes the on-time flights data that departed out of NYC airports. The data has 5 individual tabular data with individual detailed focus. The airlines table has carrier code and assoicated airlines name. The planes table talks about the details about planes like manufacturer, type, engine, etc. With majority of planes belonging to Airbus (336), Airbus Industries (400),Bombardier (368), Embraer(299) and majority of "speed" data is missing. The airports table details on location, FAA airport code, etc. The weather data talks about the weather in relation to the three origin airports (EWR, JFK, LGA). The flights data provides in depth information about the various flights that took off at the three airports.

(b) Formulating Questions:

Consider the NYC flights data. Formulate three motivating questions you want to explore using this data and explain why they are of interest.

Question 1: a)Numbers are always interesting, the most elementary information that motivates me for this partiuclar data is what is the frequency of flights to Seattle, flying out of NYC in different months? b)Another close question that fairly affects a travelers decision is the flight frequency by various carriers i.e what is the frequency of flights to Seattle offered by various existing air carriers?

Question 2: When taking flights, it could be fruitful to know the limitations and general numbers on arrival delays exhibited by flights belonging to various air carriers i.e What is the pattern/distribution of arrival delay in Seattle exhibited by various airline carriers?

Question 3: In general we look into the information that appears to directly affect us, the airlines organizations serve merely as a channel between passesngers and airplanes. It would be interesting to know about What are the current and past manufacturers of airplanes, the nature of planes they produced and the years they have been operating in?

(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.

Question 1:

```
head(nycflights13::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     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 <dttm>
```

```
nycflights13::flights %>% filter(dest == "SEA" & month == 8)
```

```
## # A tibble: 428 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     8     1      715            720        -5     1010
##  2  2013     8     1      728            730        -2     1034
##  3  2013     8     1      835            842        -7     1206
##  4  2013     8     1      859            901        -2     1215
##  5  2013     8     1      930            930         0     1227
##  6  2013     8     1     1459           1459         0     1804
##  7  2013     8     1     1647           1529        78     1927
##  8  2013     8     1     1729           1735        -6     2045
##  9  2013     8     1     1828           1729        59     2139
## 10  2013     8     1     1832           1835        -3     2147
## # ... with 418 more rows, and 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 <dttm>
```

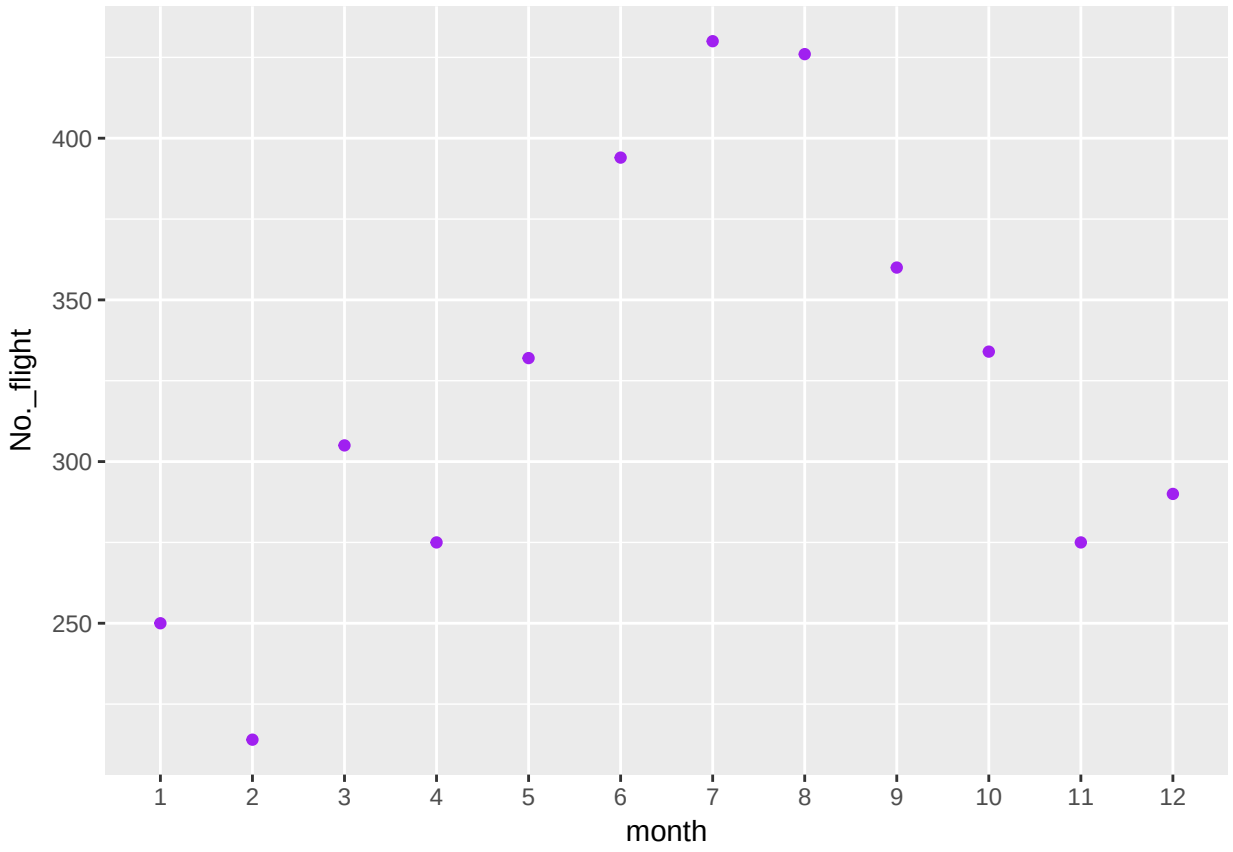
```
nycflights13::flights %>% filter(dest == "SEA" & carrier== "AA")
```

```
## # A tibble: 365 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     1824           1830        -6     2203
##  2  2013     1     2     1835           1830         5     2145
##  3  2013     1     3     1832           1830         2     2152
##  4  2013     1     4     1830           1830         0     2148
##  5  2013     1     5     1827           1830        -3     2128
##  6  2013     1     6     1827           1830        -3     2201
##  7  2013     1     7     1846           1830        16     2216
##  8  2013     1     8     1829           1830        -1     2201
##  9  2013     1     9     1829           1830        -1     2230
## 10  2013     1    10     2006           1830        96     2300
## # ... with 355 more rows, and 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 <dttm>
```

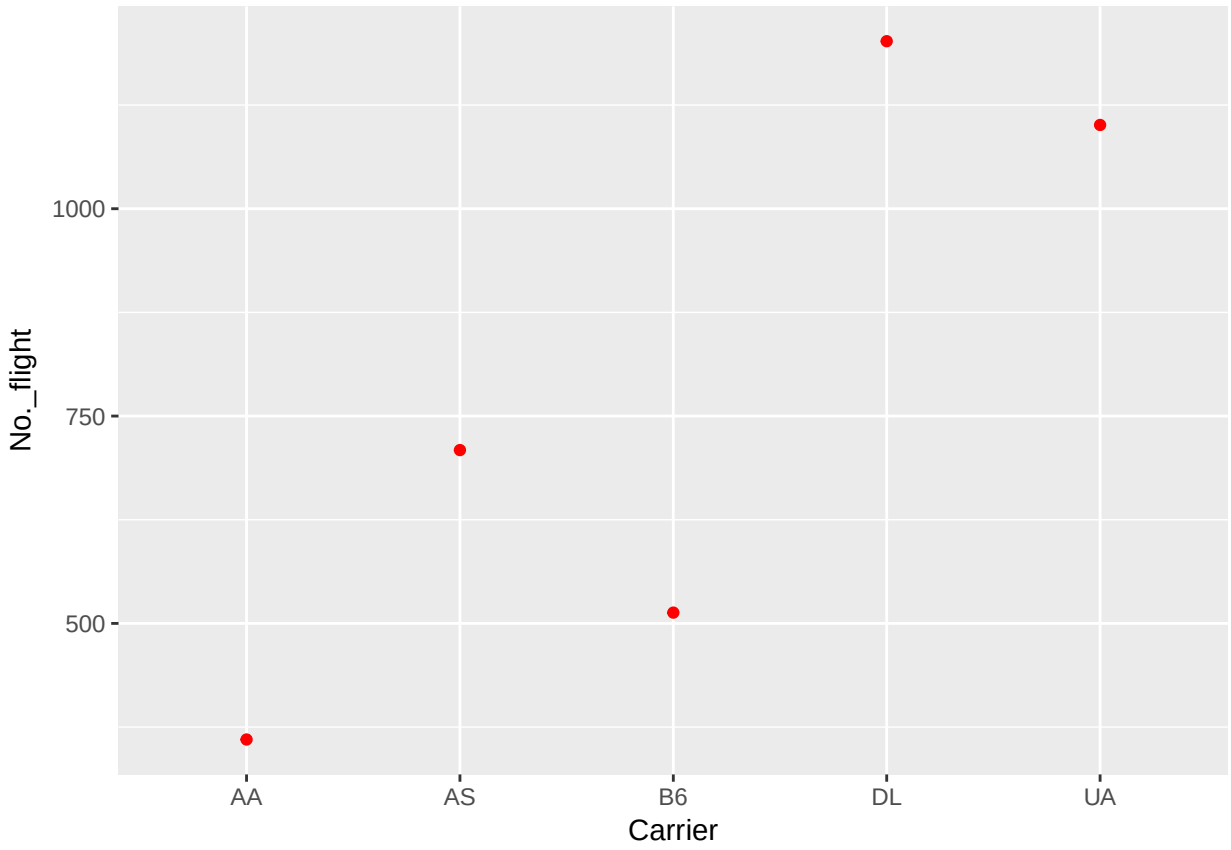
```
destsea <- na.omit (nycflights13::flights %>% filter(dest == "SEA"))

x <- data.frame(table(destsea$month))
y <- data.frame(table(destsea$carrier))

ggplot(data= x %>% mutate (month = Var1, No._flight=Freq)) + geom_point(aes(x= month, y= No._flight), c
```



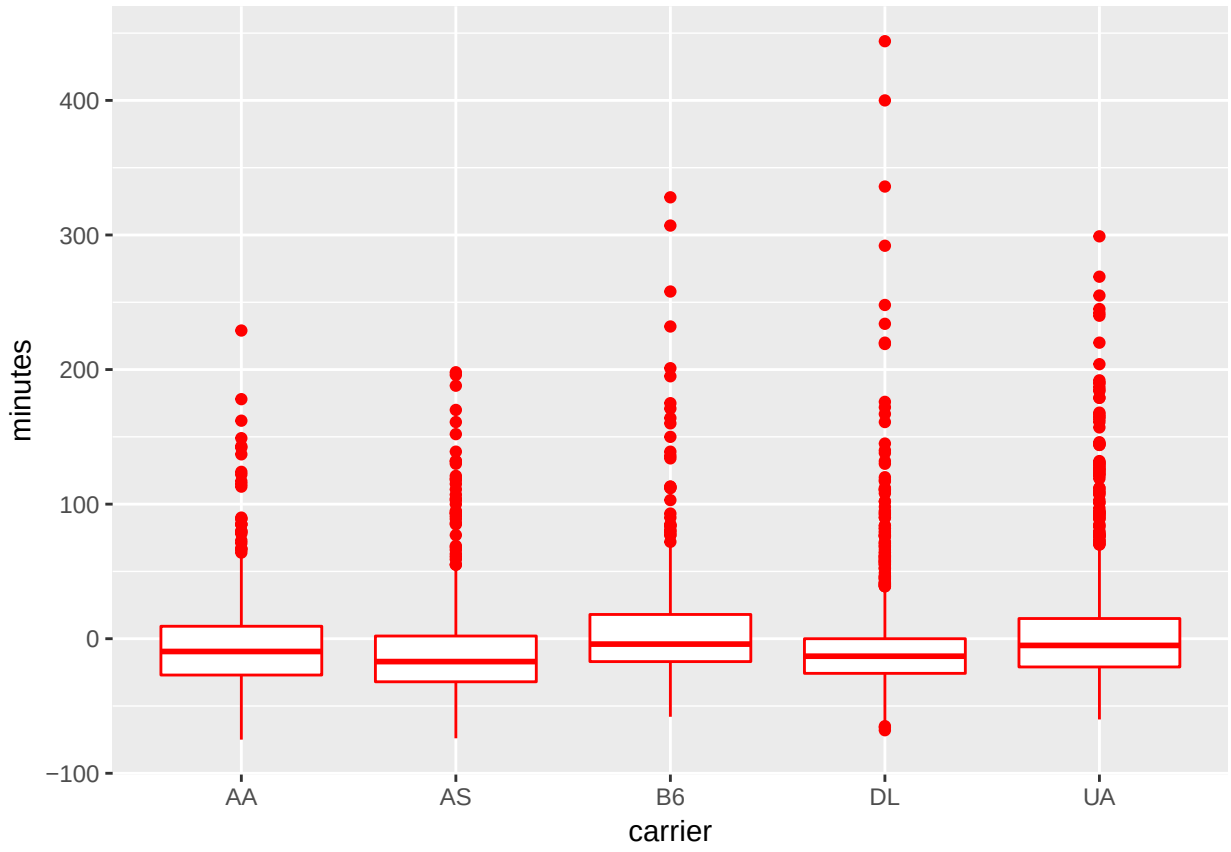

```
ggplot(data= y %>% mutate (Carrier = Var1, No._flight=Freq)) + geom_point(aes(x= Carrier, y= No._flight)
```



a)Through the analysis of the plot above that is based on the data provided, we witness that the frequency of flights to Seattle, flying out of NYC is maximum in the month of July (431) and it is lowest in the second month of the year i.e February (224). b)Out of the 5 major carriers, we observe that Delta Air Lines Inc. offers the highest frequency of flights to Seattle from NYC (1213) whereas American Airlines Inc. offers the least number of flights to Seattle from NYC (365).

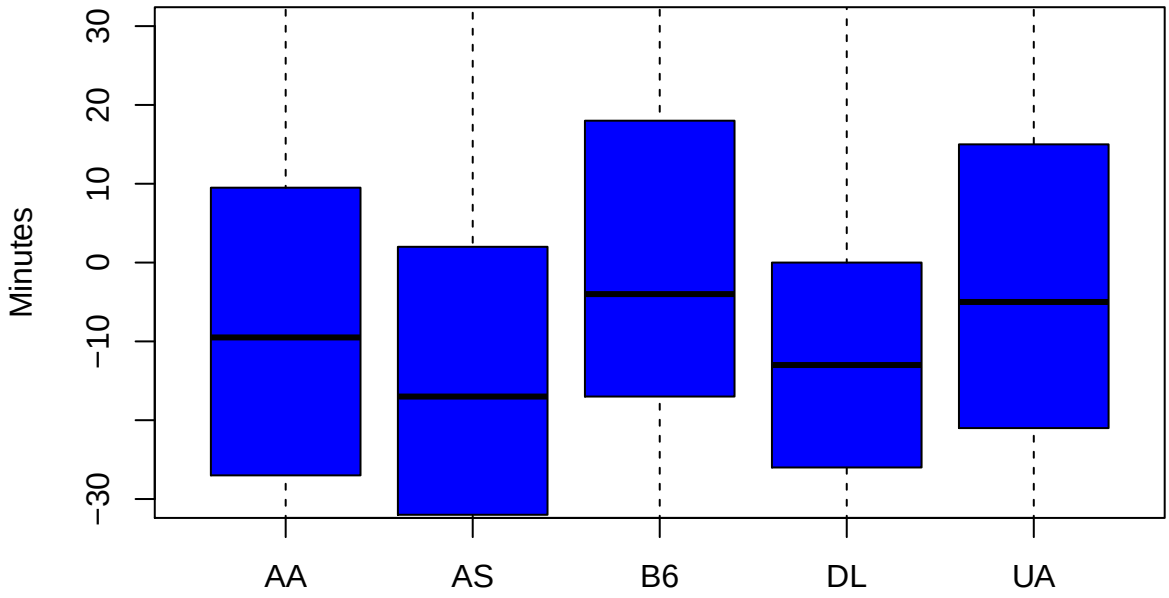
Question 2:

```
ggplot(data= destsea %>% mutate (minutes=arr_delay)) + geom_boxplot(aes(x= carrier, y= minutes), col= "
```




```
# Plotting graph between the arrival delays and various carriers for the flights from NYC to Seattle.

boxplot(destsea$arr_delay~destsea$carrier, ylim = c(-30,30), ylab = "Minutes", col="blue")
```




```
# Using boxplot to Plot graph between the arrival delays and various carriers for the flights from NYC
```

In order to observe the pattern/distribution of arrival delay in Seattle exhibited by various airline carriers, I used box-whisker plot to observe the variability outside the upper and lower quartiles in addition to observe the central tendency of the data. JetBlue Airways (B6) appears to display the highest of all delays. Also, the

relatively higher median projects that Jetblue is more consistent at arrival delays from NYC to Seattle. With the upper quartile for arrival delay data falling on 0 and the Maximum of box-whisker plot falling relatively below of all the other airlines involved, Delta Air Lines Inc. seems to be more consistent at arriving on time in Seattle from NYC. Delta also seems to have more spreaded out outliers above the maximum and reflects outliers below the minimum of box-whisker plot.

Question 3:

```
planesdata <- nycflights13::planes
ggplot(data= planes ) + geom_point(aes(x= year, y= manufacturer, col= engine))
```

```
## Warning: Removed 70 rows containing missing values (geom_point).
```

It seems interesting to know about manufacturers in this data, Embraer, Bombardier Inc, Boeing, Airbus appears to be the major and continuous manufacturers of Turbo-fan type engine planes recently. Like Boeing that has been continuously operating over two decades and has been majorly producing Turbo-fan and Turbo-jet type of engine based airplanes. The manufacturer Beech seems to be the only Turbo-Prop type engine planes manufactuer whose production existed mostly between late 1960s to early 1970s.

(d) Challenge Your Results:

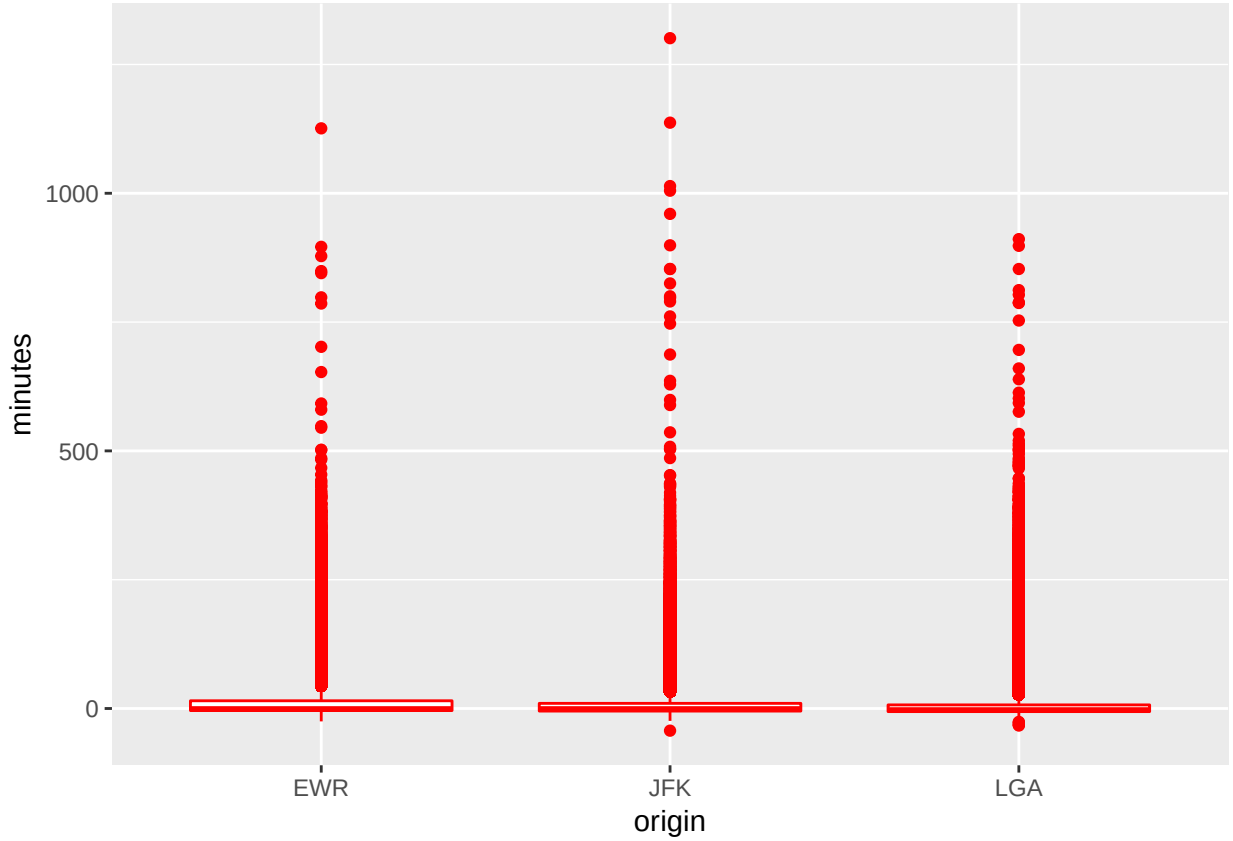
After completing the exploratory analysis from Problem 1c, do you have any concerns about your findings?

Due to limitations of time, I missed exploring about the various factors that are related to each other like the weather patterns and/or time of day associated with the delays in departure and arrival of flights. Due to limitation of current skillset, I missed connecting dots between the location of airports and/or weather to arrival or departure delays.

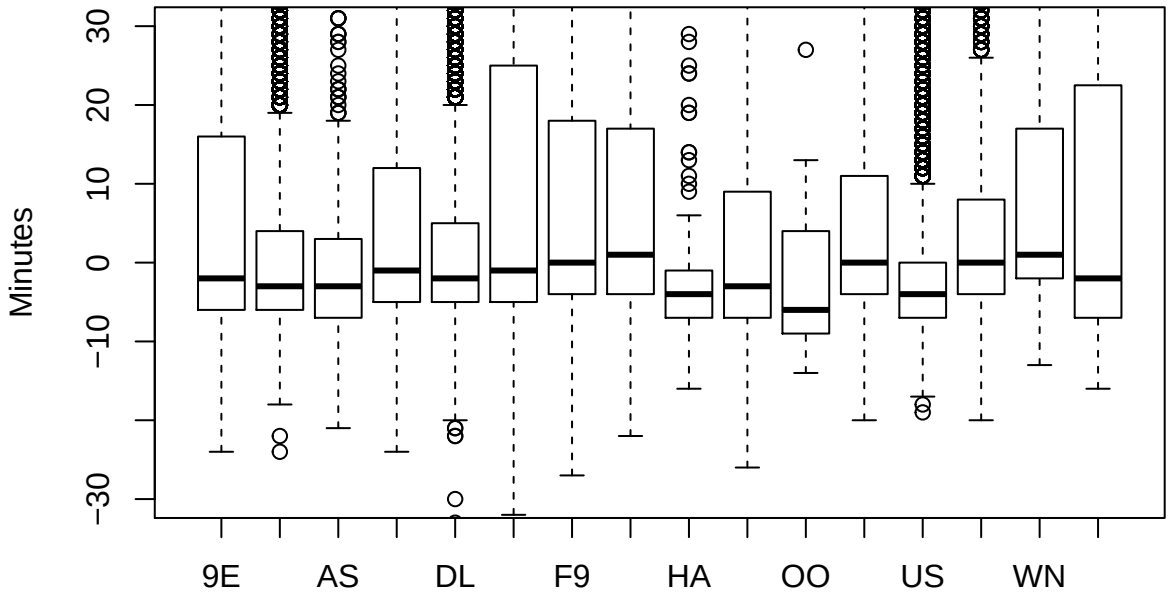
Some of my feeble attempts of making a sense out of departure delays and various carriers or departure delays and flight origin.

```
nycflights<-na.omit(nycflights13::flights)

ggplot(data= nycflights %>% mutate (minutes=dep_delay)) + geom_boxplot(aes(x= origin, y= minutes), col=
```




```
boxplot(nycflights$dep_delay~nycflights$carrier, ylim = c(-30,30) ,ylab = "Minutes")
```



```
boxplot(nycflights$dep_delay~nycflights$origin, ylim = c(min(nycflights$dep_delay),max (nycflights$dep_
```