INFX 573: Problem Set 1 - Exploring Data

Pamela Chakrabarty

Due: Monday, October 11, 2016

Collaborators: None

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)
```

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.

Answer:The data was collected by installing the nycflights13 R package.Next I used the library(nycflights13) function and then view function to see all the datasets in tabular forms.

```r
# Load standard libraries
library(tidyverse)
library(nycflights13)
flights # shows the data loaded
```

```
## # A tibble: 336,776 × 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
## 7   2013     1     1      555            600        -5      913
## 8   2013     1     1      557            600        -3      709
## 9   2013     1     1      557            600        -3      838
## 10  2013     1     1      558            600        -2      753
## # ... with 336,766 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>
```

```r
# the view command shows each of the following datasets in tabular format
View(airlines)
View(airports)
View(planes)
View(weather)
View(flights)
summary(flights) # shows a quick summary of the variables present in the flights dataset
```

```
##       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
##
```

Understandong of the different datasets under the package:

The data was collected using the nycflights13 R package. This package contains data about all flights departing from NYC (e.g. EWR, JFK and LGA) in 2013. There were total 336,776 flights observations(rows). 1. The "flights" dataset has a total of 19 variables here are as follows: year,month,day: date of departure dep_time,arr_time: the actual departure and arrival times, local tz. sched_dep_time,sched_arr_time: Scheduled departure and arrival times, local tz. dep_delay,arr_delay: Departure and arrival delays, in minutes(Negative times represent early departures/arrivals). hour,minute: Time of scheduled departure broken into hour and minutes. carrier: Two letter carrier abbreviation (We can refer to the "airlines" dataset to locate the carrier names) tailnum: Plane tail number flight: Flight number origin,dest Origin and destination. air_time: Amount of time spent in the air distance : Distance flown time_hour: Scheduled date and hour of the flight as a POSIXct date. Along with origin, can be used to join flights data to weather data.

The package contains other useful datasets as follows:

2."airlines" dataset contains the 2 columns(variables) namely Carrier (stands for carrier codes) and names (names of airlines)

2. "Airports" dataset comprising of 6 columns (variables) namely faa : the FAA airport code Name :name of airport lat: latitide of airport lon: longitide of airport) alt: altitude in feet) tz: timezone based off GMT dst: Daylight savings from timezone: A= Standard US DST starting on second Sunday of March and ending on first Sunday of November, U= unknown and N= no dist

3."Planes" dataset having construction data all planes in the FAA aircraft registry. It contains 9 variables tailnum : Tail number year : Year manufactured type : Type of plane model: model name manufacturer: Manufacturer name engines: Number of engines seats : Number of seats speed : Average cruising speed in mph engine : Type of engine

4."Weather" dataset conatining hourly meterological data for each airport.It has 15 variables which are as follows: origin: denotes Weather station and is named origin to help merging this with flights data year,month,day,hour: Time of recording
temp: Temperature in F dewp: Dewpoint in F humid: Relative humidity wind_dir: Wind direction (in degrees), wind_speed,wind_gust: speed and gust speed (in mph) precip Preciptation, in inches pressure : Sea level pressure in millibars visib : Visibility in miles time_hour : Date and hour of the recording as a

POSIXct date
where POSIXct class stores date/time values as a list of componenst(hour, min, sec, mon etc)

(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.

Answer: Two questions that I would like to explore using this data would be

1.What is the trend of arrival delays versus departure delays for airline carrier US (US airways Inc.) flights departing from NYC in 2013? This question interests me as I will be able to see the trends in delays for the US Airways Inc which is one of the most popular carriers in the list.

Approach : For the first question, I will start by joining the data sets "flights" and "airlines" using the commin variable "carrier". Next, I will filter a new dataset named USairways_filter_new using the function filter() with a condition carrier=="US". Next I will use the function ggplot() on this new dataset with the departure delays on the x- axis and the arrival delays on the y-axis.

2. What is the trend in the number of flights departing from NYC that show arrival delays and departure delays? What is the trend for the same against each month? This question interests me as I want to understand the number of flights that have been arriving opr departing late from NYC and I am also curious to know what this observation trends over the 12 months.

Approach: For the second question, I will load the "flights" dataset and then use then use the function ggplot() on the dataset with the origin as the x axis and the delayed arrivals as the y axis. I will use another ggplot() on the same dataset with the origin as the x axis and the delayed arrivals as the y axis.

(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.

Answer: I have mentioned in the last question how I plan to perform EDA on the dataset

Vizualization for Question 1

```
# Joining flights dataset with airlines dataset using common variable "carrier"
joined_flights <- inner_join(x = flights, y = airlines, by = "carrier")
# viewing the newly joined dataset
View(joined_flights)
# Filtering the new dataset by applying a condition of carrier as US and then assigning the output to a
USairways_flights_new <- filter(joined_flights, carrier == "US")
#Viewing the new dataset USairways_flights_new
View(USairways_flights_new)
```

This below graph (capturing arrival and departure delays for only US carrier airline flights) shows me that there seems to be a positive relationship between arr_delay and dep_delay as departure delays increase,

arrival delays tend to increase as well. We also note that the majority of points fall near the point (0, 0) here. There is a large mass of points clustered there and I have tried to thin out this overploting by using the argument alpha to the geom_point function.The graph also shows that a lot of these delays are showing strongly for the LGA airport particularly, followed by the JFK.

```
ggplot(data= USairways_flights_new, aes(x = dep_delay, y = arr_delay, color = origin)) + geom_point( si
```

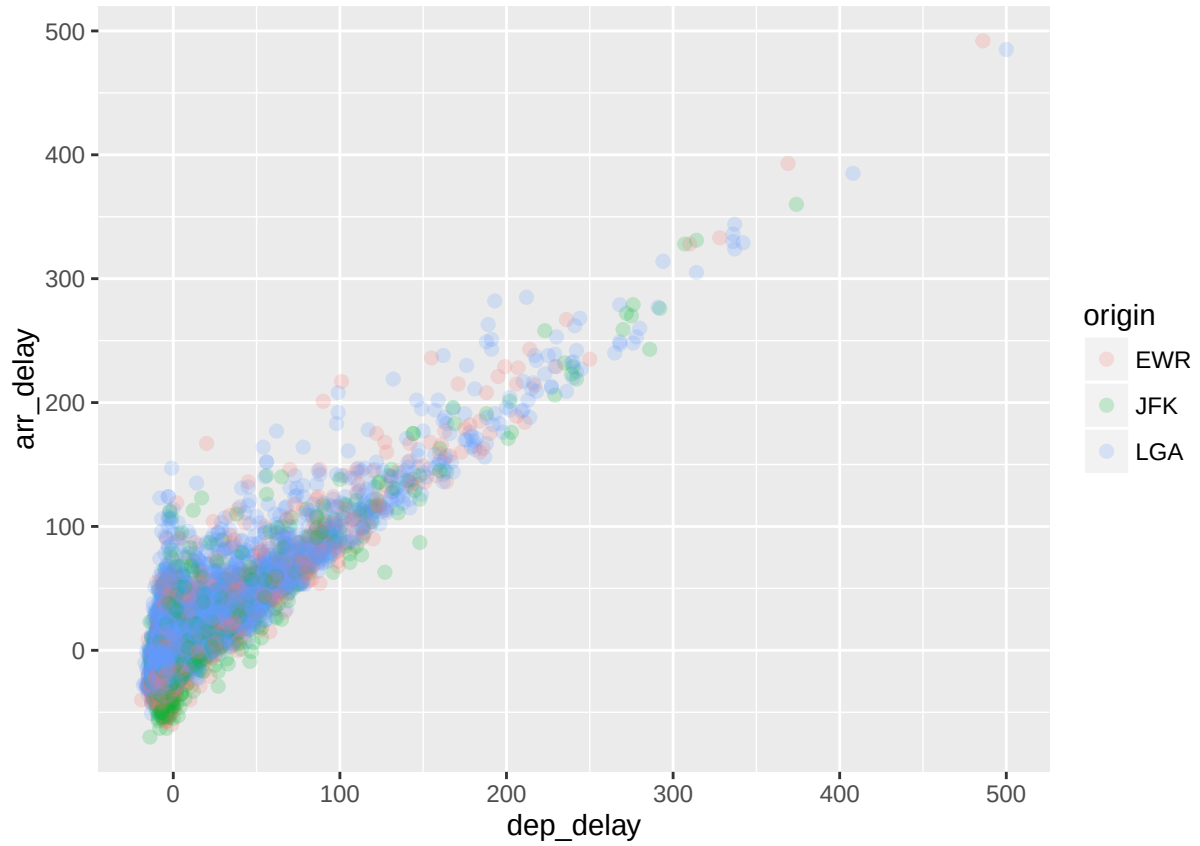

Figure 1: Arrival delays versus departure delays for all US Airways Inc flights departing from nyc in 2013

This below graph shows the arrival delays for the total number of flights of all carrier airlines. The findings indicate that while the time period of arrival delays are moderately around 200 or less, there are a good number of outliers that record above 200 as well. The graph shows that the LGA airport has the highest number of flights with reported delayed arrivals starting at about 900 to about 2500. This is followed by the JFK airport.

```
ggplot(data= USairways_flights_new, aes(x = flight, y = arr_delay, color = origin)) + geom_point(size =
```

This below graph shows the departure delays for the total number of flights of all carrier airlines. The findings indicate that while the time period of arrival delays are moderately around 150 or less, there are a good number of outliers that record above 150 as well. As compared to the above graph (recording arrival delays), this graph shows that the LGA airport again has the highest number of flights with reported delayed departures starting at about 950 to about 2500. However, unlike the above grapg this one shows a close match between number of delayed departures from the EWR and those from JFK. Overall, this graph looks less dense than the above graph which tells me that the number of arrival delays are more than departure delays for all flights departing from NYC.

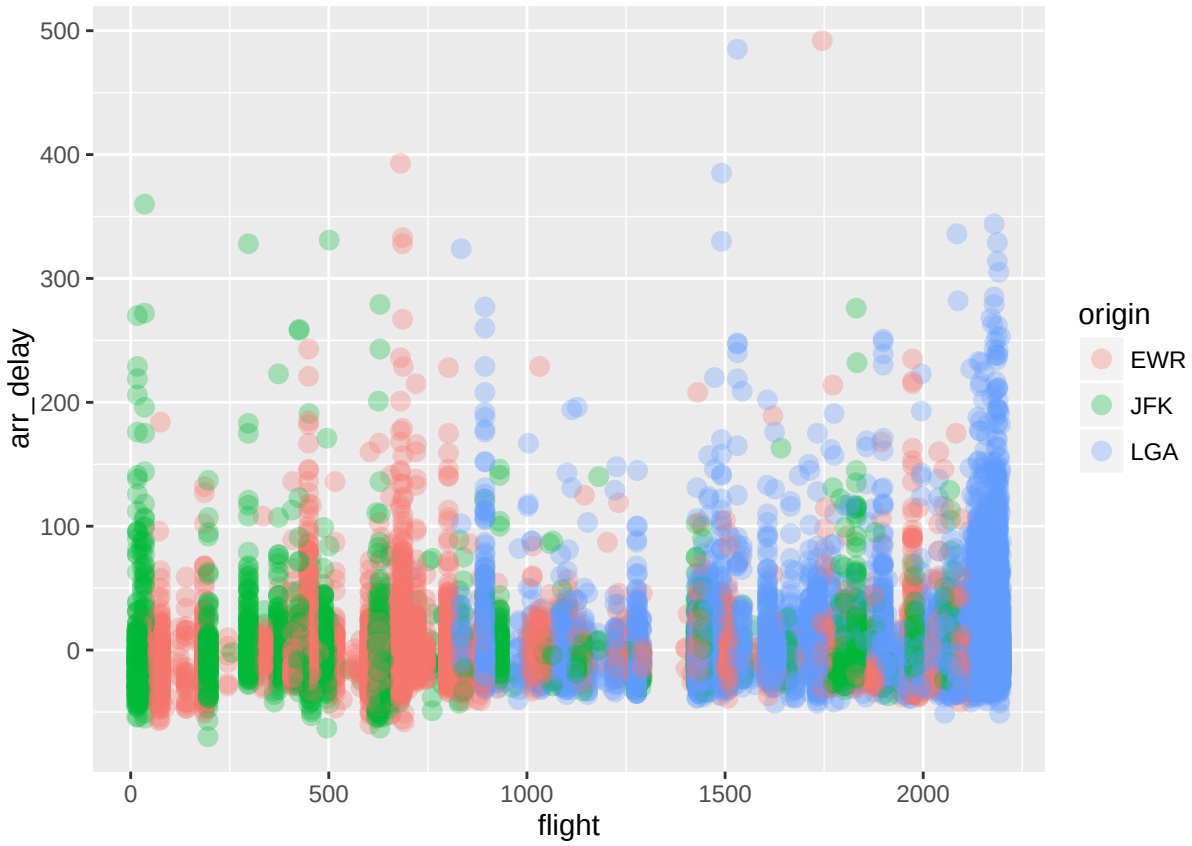

Figure 2: Trend in Arrival delays for total number of US Airways Inc flights departing from NYC

```
ggplot(data= USairways_flights_new, aes(x = flight, y = dep_delay, color = origin)) + geom_point(size =
```

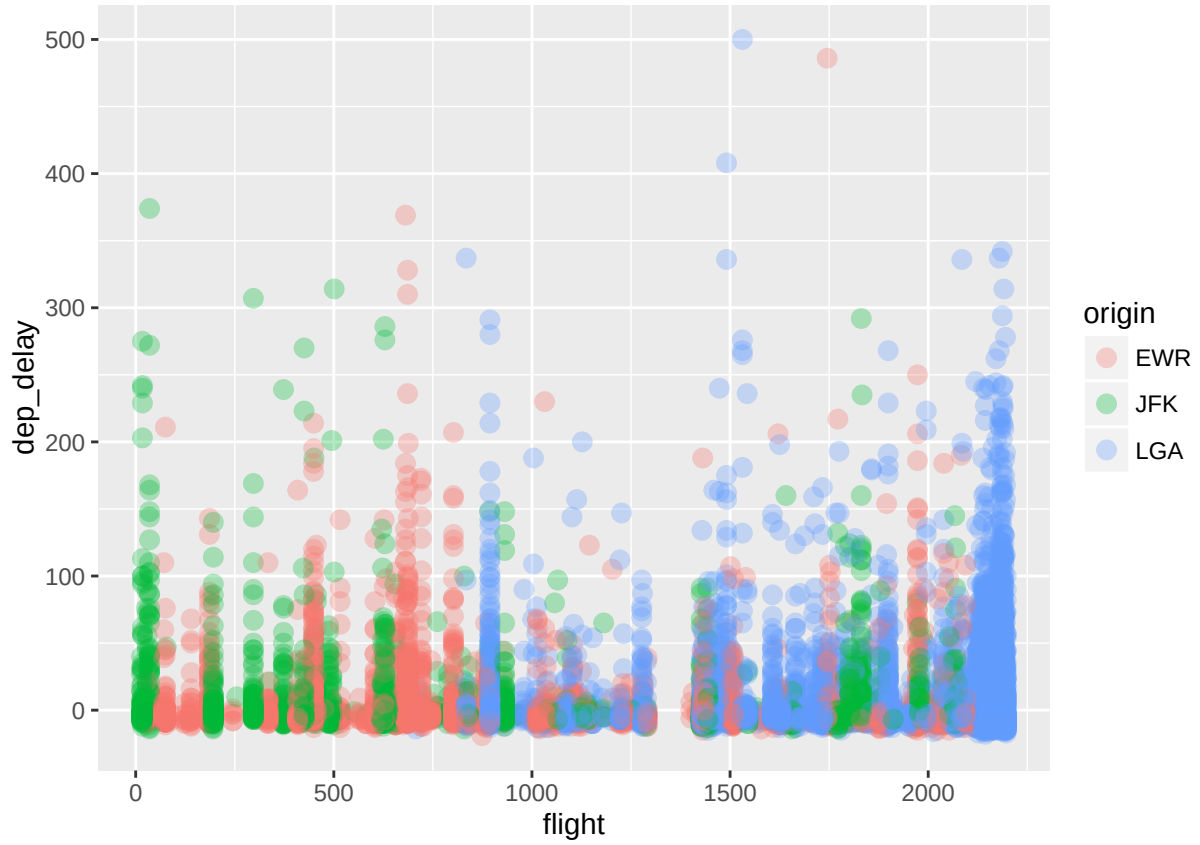

Figure 3: Trend in Departure delays for total number of US Airways Inc flights departing from NYC

Vizualization for Question 2

The below graph shows that Out of the three NYC airports, the flights arrived at the JFK airport are recording higher delayed arrival times for all flights, followed by EWR and then LGA. This finding tells me that one reason could be that JFK is the most popular airport in NY and is expected to be very busy with a huge number of connecting and direct flights (my assumption). I would need to do more analysis on this finding to see if that is indeed the case.

```
ggplot(data= flights, aes(x = origin, y = arr_delay, color= origin)) + geom_line(stat="identity")
```

The below graph shows that Out of the three NYC airports, the flights deaprting from the JFK airport are recording higher delayed arrival times for all flights, followed by EWR and then LGA. This finding is very close to the above finding for arrival delays and hence this tells me that that JFK must be operating more intensely with a huge number of connecting and direct flights (again my assumption). I would need to do more analysis on this finding to see if that is indeed the case.

```
ggplot(data= flights, aes(x = origin, y = dep_delay, color= origin)) + geom_line(stat="identity")
```

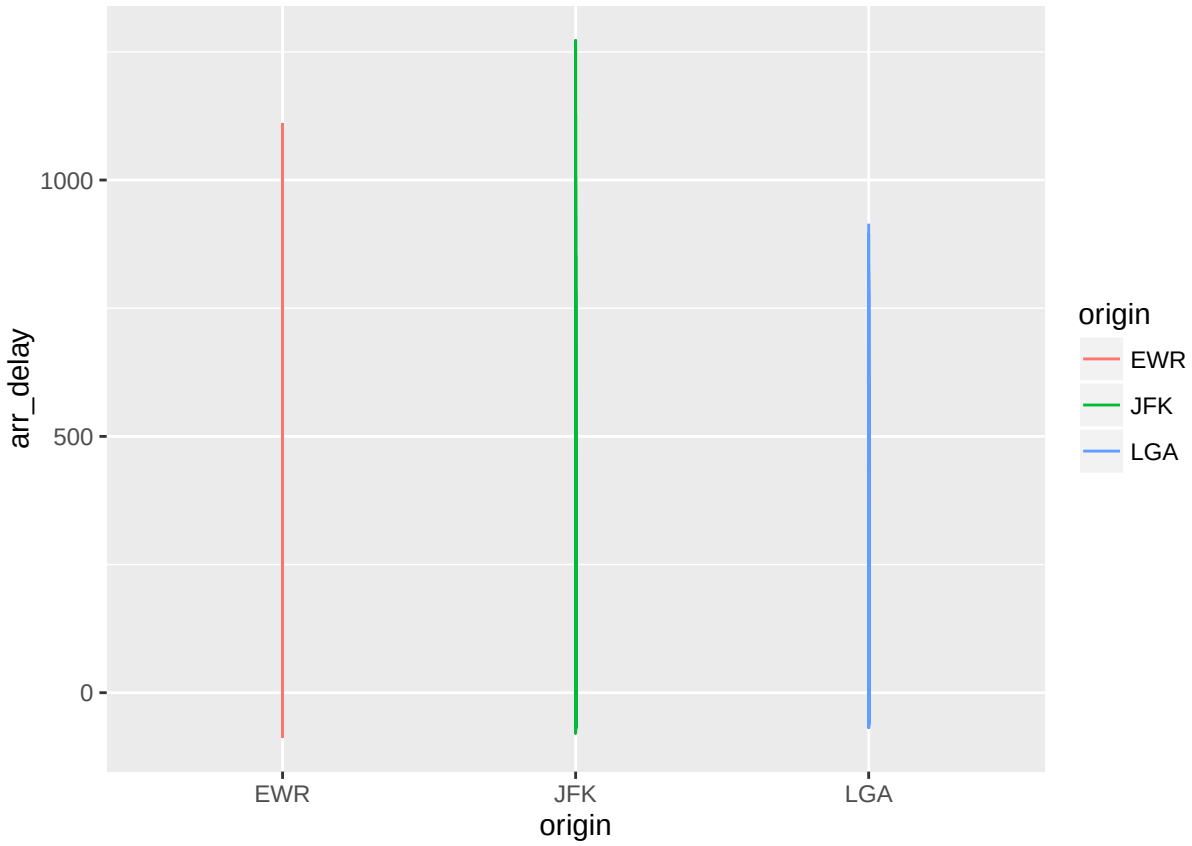

Figure 4: Number of all-carrier flights from NYC with delayed arrivals

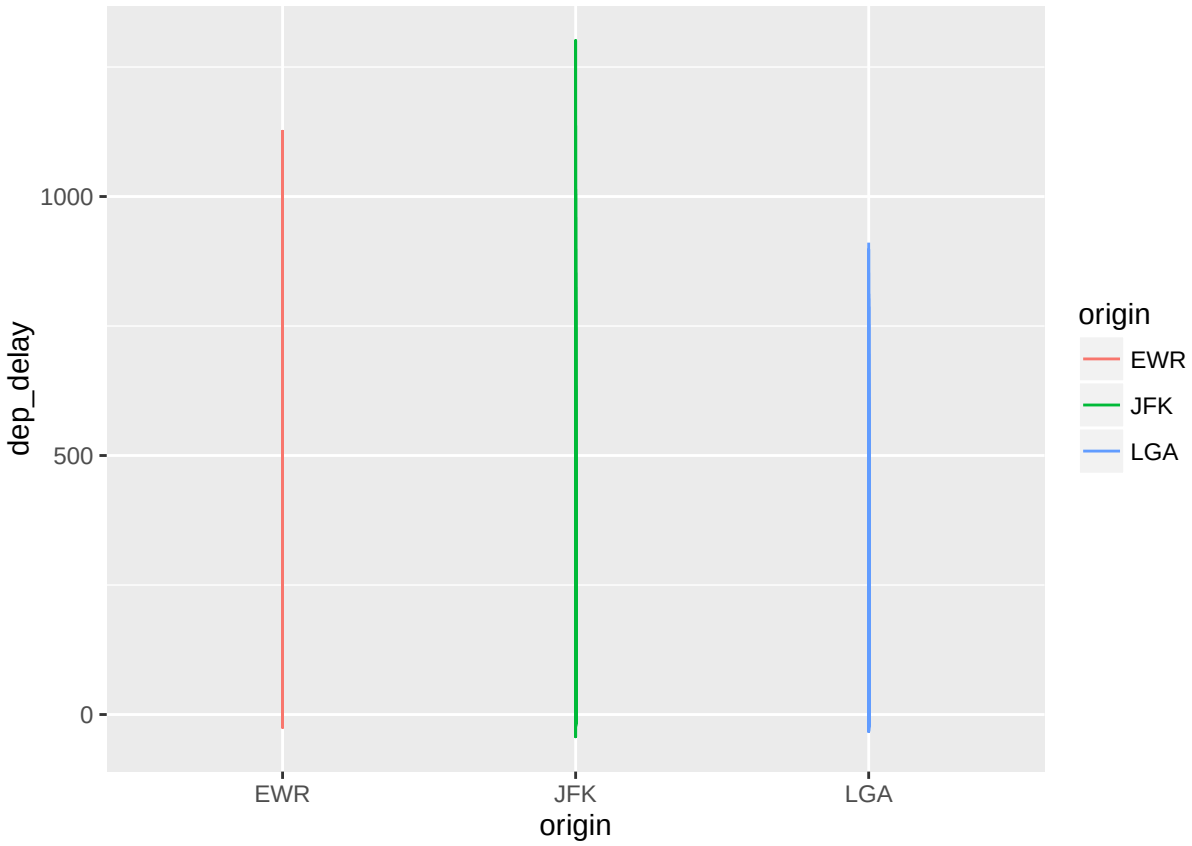

Figure 5: Number of all-carrier flights from NYC with delayed departures

The below graph shows that most of the flights have arrival delays of less than 500 minutes and that there is a relative increase in the arrival delays in the months of June, July and December, with a few outliers in each month. This finding tells me that summer might be a time when flights are arriving late. I am curious to exploremore around the dataset to find out what might be causing this trend.

```
flight_arr_times <- select(flights, month, day, arr_delay)

ggplot(data= flight_arr_times, aes(x = arr_delay, y = month)) + geom_point(alpha=0.2) + geom_smooth(se=
```

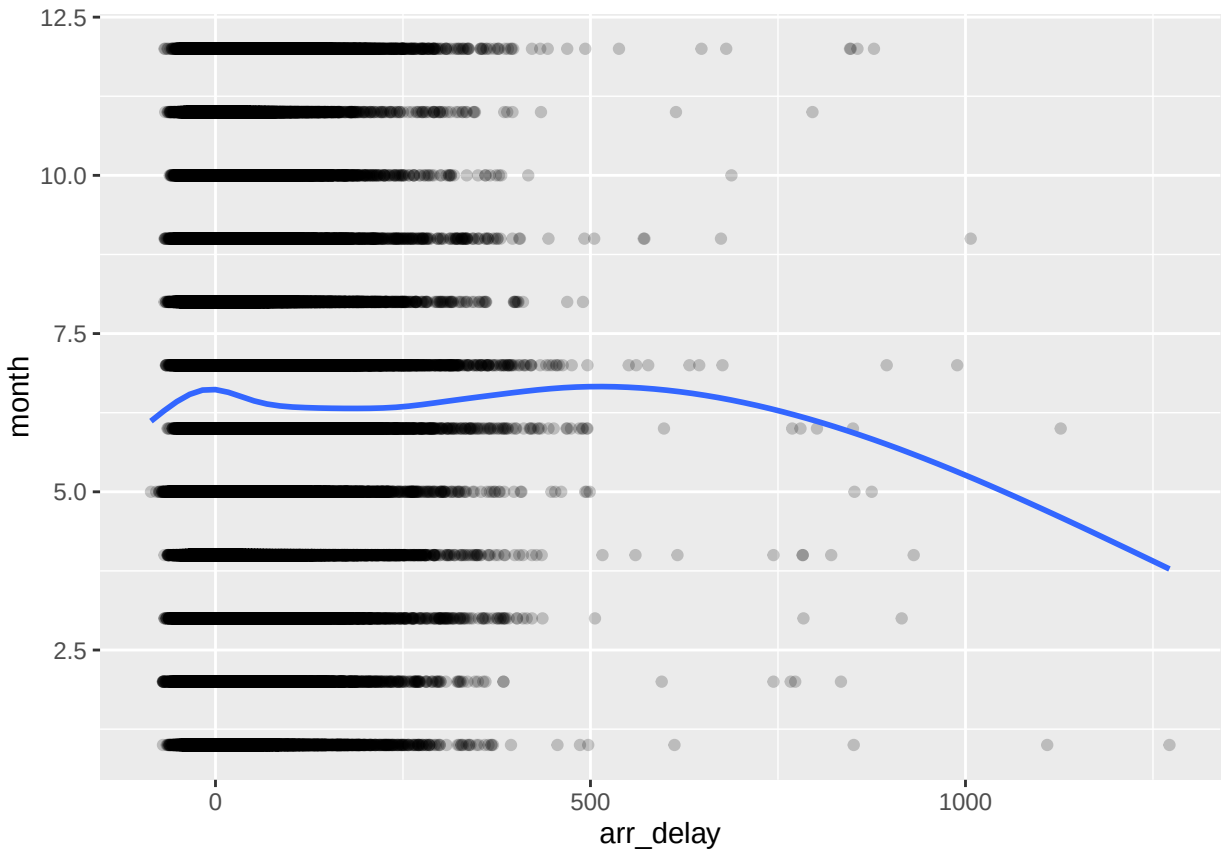

Figure 6: Number of delayed arrivals for all flights against each month

The below graph shows that most of the flights have departure delays of less than 500 minutes and that there is a relative increase in the arrival delays in the months of June, July and December, with a few outliers in each month. January, February and October shows the least departure delay time. This finding tells me that winter might be the time when flights are departing relatively faster as compared to the other months.

```
flight_dep_times <- select(flights, month, day, dep_delay)

ggplot(data= flight_dep_times, aes(x = dep_delay, y = month)) + geom_point(alpha=0.2) + geom_smooth(se=
```

(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.

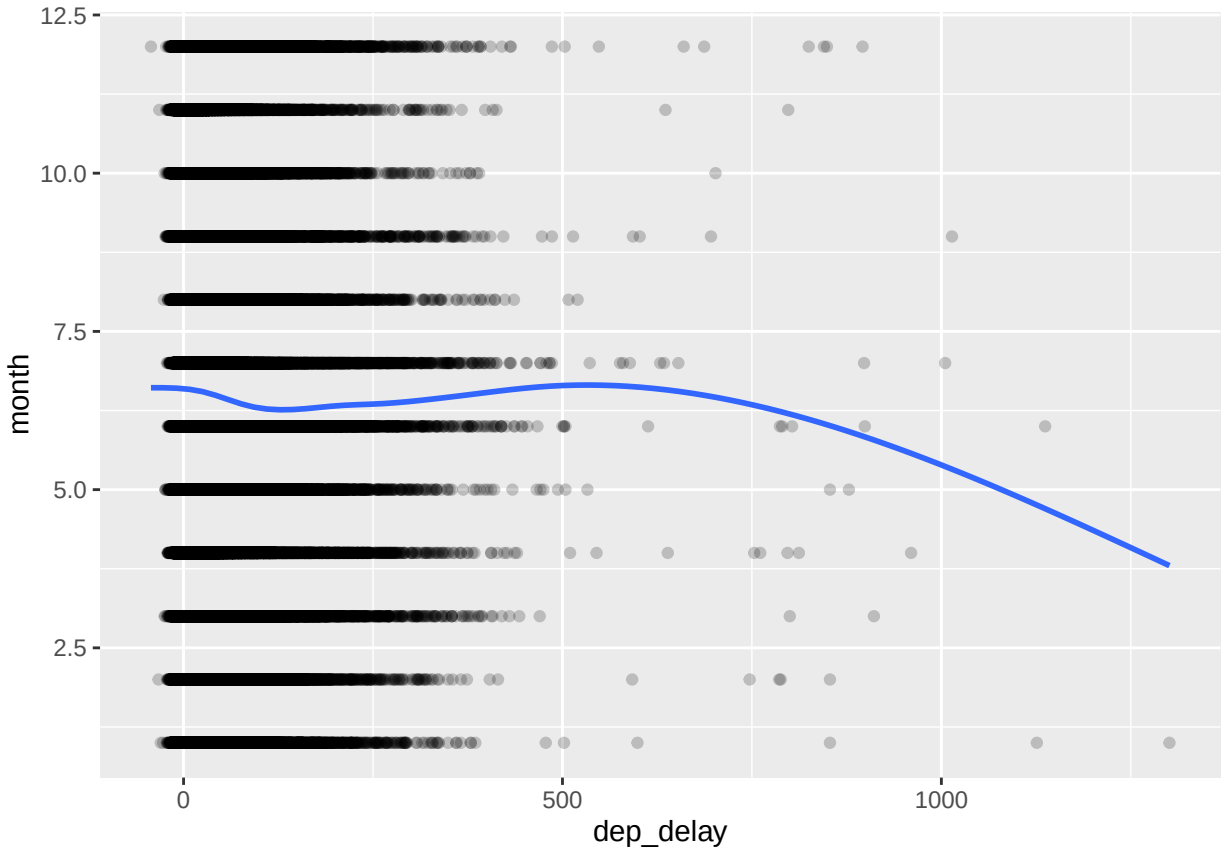

Figure 7: Number of delayed departures for all flights against each month

Answer: Yes, some of my concerns around my inferences are as follows:

1. Based on my findings, I said that one of the graph shows that for most of the flights, there is a relative increase in the arrival delays in the months of June, July and December, with a few outliers in each month. Based on June and July, I inferred that summer might be a time when flights are arriving late. But then December was way far from summer and summer doesnt include December. So, again there is a scope of forming assumptions here which I am sure could be very faulty while trying to make accurate inferences from the data.

2. Whie answering the second question, I inferred based on my plots that the flights arriving at the JFK airport are recording higher delayed arrival times for all flights, followed by EWR and then LGA. This finding tells me that one reason could be that JFK is the most popular airport in NY and is expected to be very busy with a huge number of connecting and direct flights (my assumption). This might be a bias from my end as in my knowledge, JFK tends to be busier than the other two. If this is not true, then my findings might not be that accurate. That is once case that I am concerned about with respect to my findings.

3. While answering the second question, I inferred based on my plots that the number of arrival delays are more than departure delays for all flights departing from NYC. However, there is no record of any flights arrival or departure delays recorded on my graph plots. This might indicate that some of the observations(rows) might have been habing missing values. So, here my concern is that an outsider observation might think that there had been no flight arival or departure delays when the number of flights was > 1250 and < 1450. But the case might be that some values were missing. How to resolve this conflict.