INFX 573: Problem Set 1 - Exploring Data

Charudatta Deshpande

Due: Thursday, October 12, 2017

Collaborators: N/A

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: Installed Rcpp (0.12.10) different from Rcpp used to build dplyr (0.12.11).
## Please reinstall dplyr to avoid random crashes or undefined behavior.
```

```
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 -

```
# Load Data
data(flights)
```

Source - This data was collected from Bureau of transportation statistics, http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236. This is On-time data for all flights that departed NYC (i.e. JFK, LGA or EWR) in 2013.

Following fields can be found in the data -

year,month,day - Each field is part of Date of departure.

dep_time,arr_time - Actual departure and arrival times in local timezone.

sched_dep_time,sched_arr_time - Scheduled departure and arrival times in local timezone.

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 name (airline name) abbreviation.

tailnum - Plane tail number.

flight - Assigned Flight number.

origin,dest - Origin and destination airport codes.

air_time - Amount of time spent in the air, in minutes.

distance - Distance between airports, in miles.

time_hour - Combination of Scheduled date and hour of the flight as a single field.

```
# Following commands will convert the data into data.table format, and display first
#and last few lines of the dataset.
library("data.table")
data(flights)
as.data.table(flights)
```

```
##         year month day dep_time sched_dep_time dep_delay arr_time
##      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
##     ---
## 336772: 2013     9  30       NA           1455        NA       NA
## 336773: 2013     9  30       NA           2200        NA       NA
## 336774: 2013     9  30       NA           1210        NA       NA
## 336775: 2013     9  30       NA           1159        NA       NA
## 336776: 2013     9  30       NA            840        NA       NA
##         sched_arr_time arr_delay carrier flight tailnum origin dest
##      1:            819        11      UA   1545  N14228    EWR  IAH
##      2:            830        20      UA   1714  N24211    LGA  IAH
##      3:            850        33      AA   1141  N619AA    JFK  MIA
##      4:           1022       -18      B6    725  N804JB    JFK  BQN
##      5:            837       -25      DL    461  N668DN    LGA  ATL
##     ---
## 336772:           1634        NA      9E   3393      NA    JFK  DCA
## 336773:           2312        NA      9E   3525      NA    LGA  SYR
```

```
## 336774:           1330        NA      MQ   3461  N535MQ    LGA  BNA
## 336775:           1344        NA      MQ   3572  N511MQ    LGA  CLE
## 336776:           1020        NA      MQ   3531  N839MQ    LGA  RDU
##         air_time distance hour minute           time_hour
##      1:      227     1400    5     15 2013-01-01 05:00:00
##      2:      227     1416    5     29 2013-01-01 05:00:00
##      3:      160     1089    5     40 2013-01-01 05:00:00
##      4:      183     1576    5     45 2013-01-01 05:00:00
##      5:      116      762    6      0 2013-01-01 06:00:00
##     ---
## 336772:       NA      213   14     55 2013-09-30 14:00:00
## 336773:       NA      198   22      0 2013-09-30 22:00:00
## 336774:       NA      764   12     10 2013-09-30 12:00:00
## 336775:       NA      419   11     59 2013-09-30 11:00:00
## 336776:       NA      431    8     40 2013-09-30 08:00:00
```

```
head(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>
```

```
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 <dttm>
```

Observations -

1. The data appears to be sorted by year, month and day. The year is constant at value 2013.
2. Departure delay ranges from 1301 minutes to -43 minutes (early flight).
3. Total number of records is 336776. This indicates a little less than 1000 flights depart from these three airports everyday.

(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.

Answer -
I find these three questions interesting for reasons specified below.

1. What is the relationship between 'carrier' (airline) and arrival delay? I find this interesting because I would like to find out if there any airlines that frequently cause delays, or if there are any airlines that are always on time.
2. Are there specific months where arrival delay is more, or less? I am interested in this because I would like to find out if delays are more during winter weather, holiday season, summer etc. And if one can choose a particular month to travel to minimize the delay.
3. Is there a relationship between time of the day the flight is scheduled to leave, and arrival delay? This is interesting since if a relationship is proven, one can take flights during a specific time of the day and experience minimum amount of delay.

(c) Exploring Data:

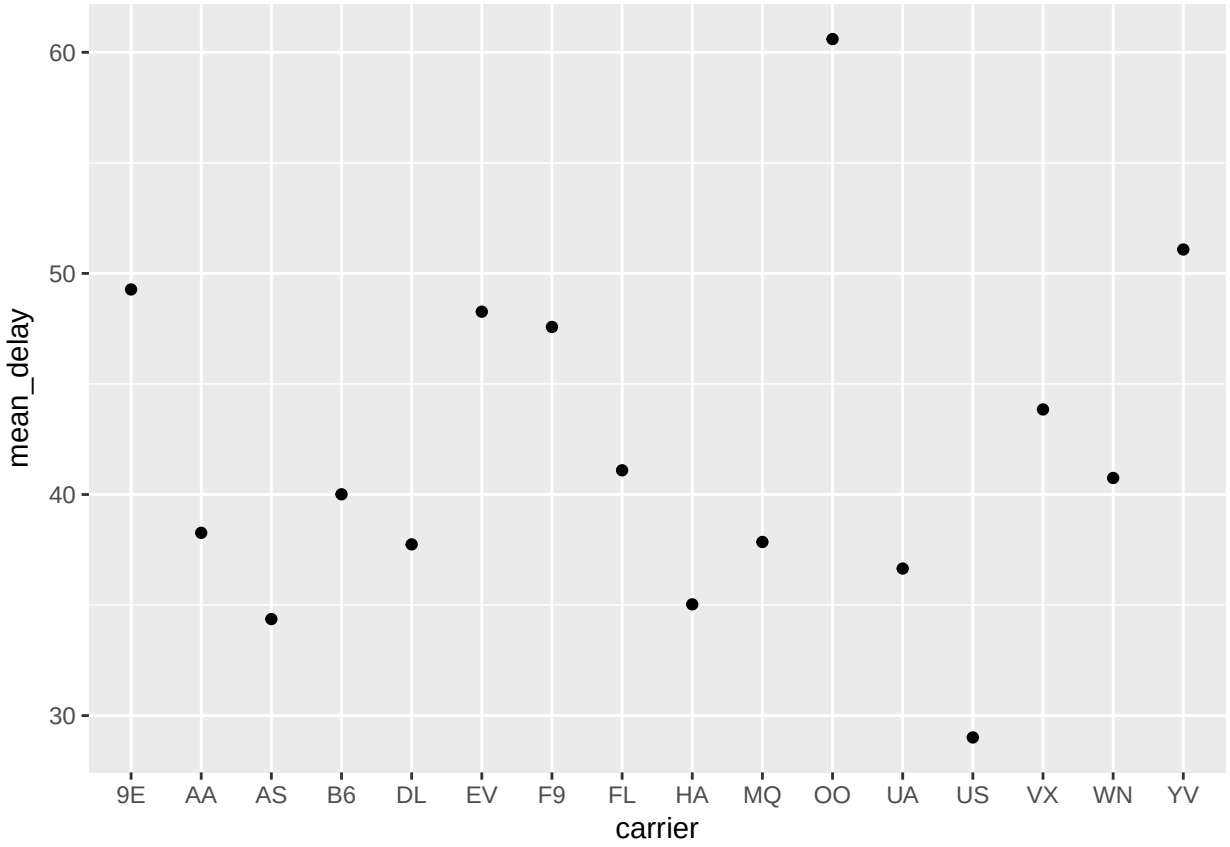
For each of the questions you proposed in Problem 1b, perform an exploratory data analysis designed to address the question. At a minimum, you should produce two visualizations related to each question. Be sure to describe what the visuals show and how they speak to your question of interest.

Answer -
1. What is the relationship between 'carrier' (airline) and arrival delay? I find this interesting because I would like to find out if there any airlines that frequently cause delays, or if there are any airlines that are always on time.

```
# Following code creates a new dataset delay which stores the carrier name and average arrival
#delay for each carrier. Then a scatter plot is created to visualize the results.
delay <- flights %>%
         filter(arr_delay > 0) %>%
         group_by(carrier) %>%
         summarize(mean_delay=mean(arr_delay)) %>%
         arrange(desc(mean_delay))
print(delay)
```

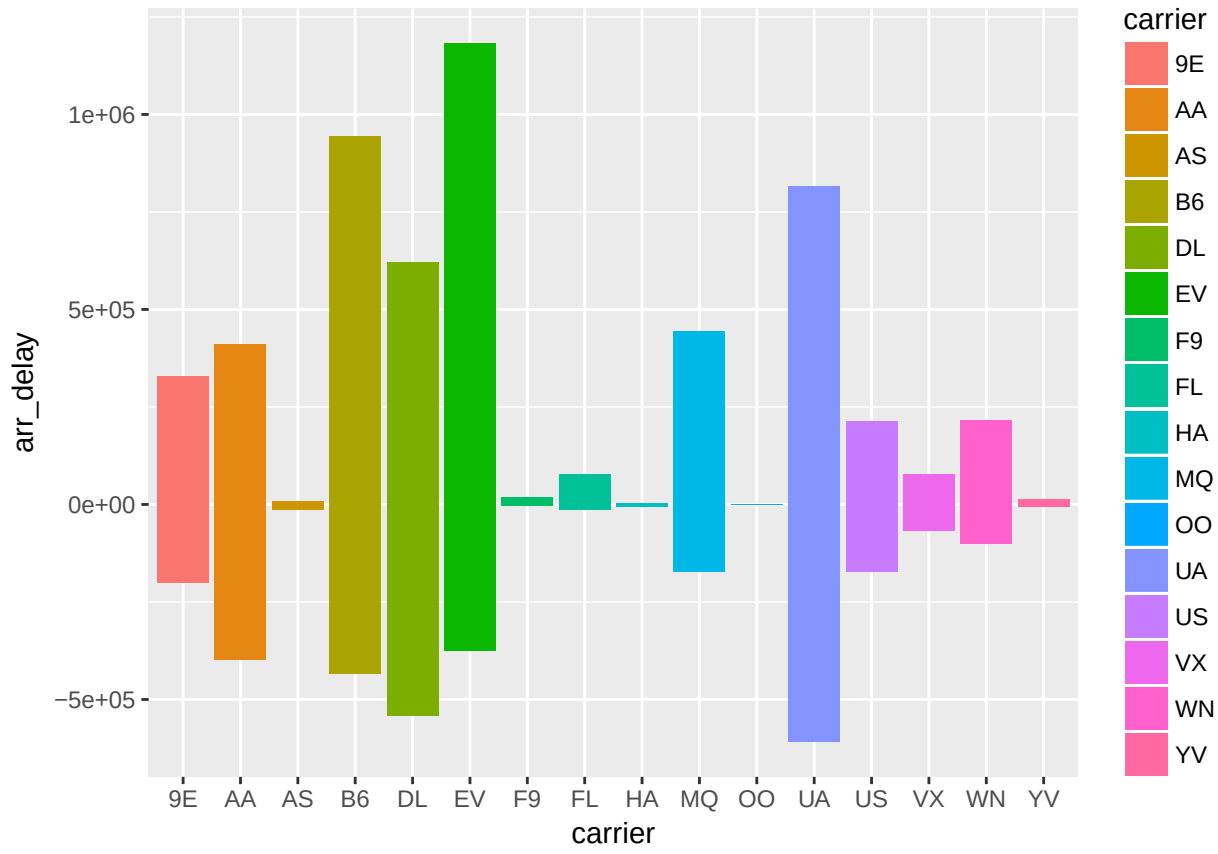
```
## # A tibble: 16 x 2
##    carrier mean_delay
##      <chr>      <dbl>
##  1      OO   60.60000
##  2      YV   51.08140
##  3      9E   49.27271
##  4      EV   48.26858
##  5      F9   47.57908
##  6      VX   43.84708
##  7      FL   41.09446
##  8      WN   40.74755
##  9      B6   40.00906
## 10      AA   38.26555
## 11      MQ   37.85205
## 12      DL   37.74356
## 13      UA   36.65098
## 14      HA   35.03093
## 15      AS   34.36508
## 16      US   29.01157
```

```
ggplot(delay, aes(carrier, mean_delay)) + geom_point()
```



Comments -
Above plot indicates that 'OO' (SkyWest Airlines Inc.) has the most average arrival delay at 60.6 minutes, while 'US' (US Airways Inc.) has the least amount of delay at 29 minutes. Most of the well known airlines seem to have less mean delay. This could indicate that bigger airlines are given priority helping them minimize delays.

```
# Following code creates a bar chart of carrier and arrival delay.
ggplot(flights, aes(carrier, arr_delay, fill=carrier)) + geom_bar(stat="identity")
```



Comments -
Above plot indicates that 'OO' (SkyWest Airlines Inc.) though has the maximum arrival delay, the volume of its flights is very low compared to some others. 'UA' (United Air Lines Inc.) on the other hand seems to have a large number of flights, both late and early, thereby resulting in reduced mean delay. Same can be observed for DL, AA, US, VX etc. For 'EV'(ExpressJet Airlines Inc.), more number of flights seem delayed, thereby increasing the delay.

Conclusion -
Based on above two plots, it can be concluded that though mean delay varies from airline to airline, there isn't a pattern that can be clearly established. It is therefore concluded that choosing a specific airline does not indicate reduced possibility of delay.

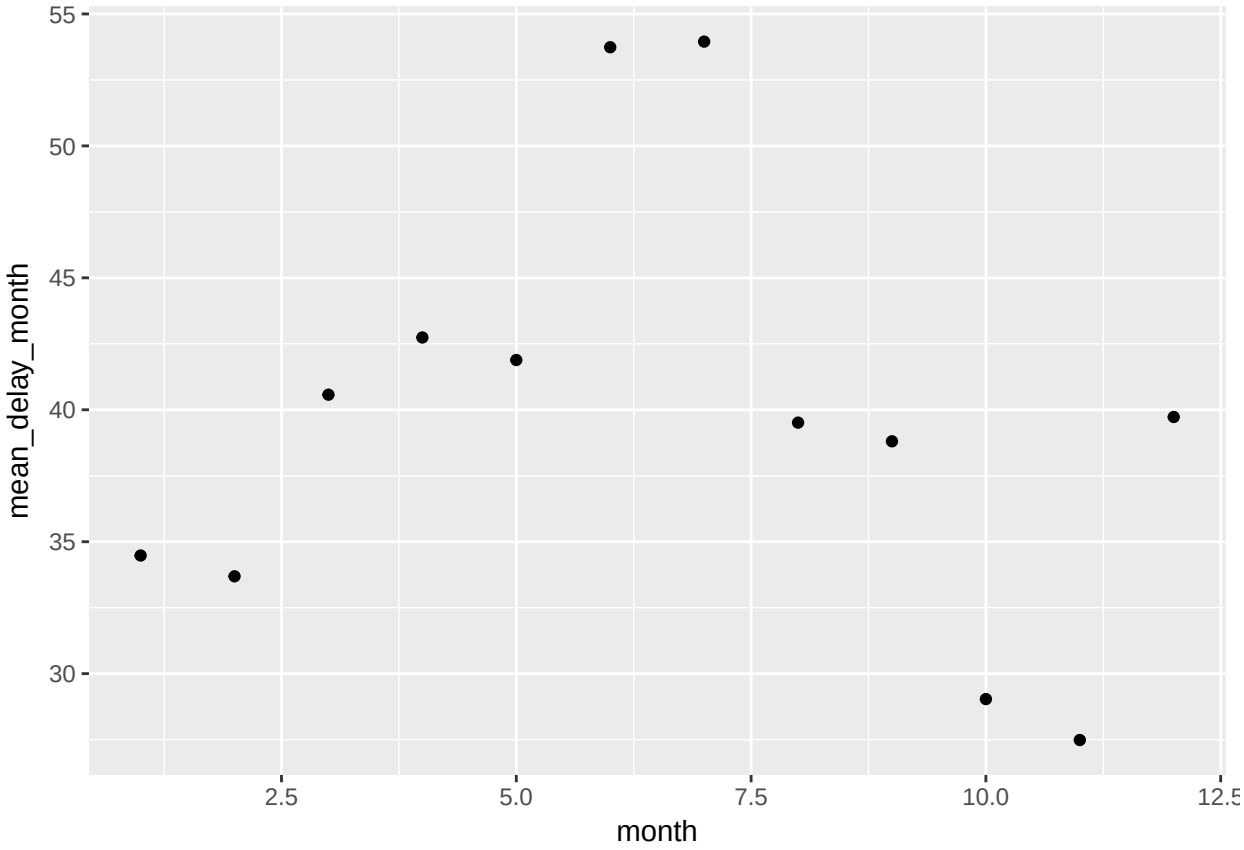
2. Are there specific months where arrival delay is more, or less? I am interested in this because I would like to find out if delays are more during winter weather, holiday season, summer etc. And if one can choose a particular month to travel to minimize the delay.

```
# Following code creates a new dataset delay which stores the month name and average arrival
#delay for each month. Then a scatter plot is created to visualize the results.
delay_month <- flights %>%
         filter(arr_delay > 0) %>%
         group_by(month) %>%
         summarize(mean_delay_month=mean(arr_delay)) %>%
         arrange(desc(mean_delay_month))
print(delay_month)
```

```
## # A tibble: 12 x 2
##    month mean_delay_month
##    <int>            <dbl>
```

```
##  1     7         53.95152
##  2     6         53.73827
##  3     4         42.73958
##  4     5         41.88586
##  5     3         40.57166
##  6    12         39.72725
##  7     8         39.51294
##  8     9         38.80555
##  9     1         34.47749
## 10     2         33.68921
## 11    10         29.03665
## 12    11         27.48459
```

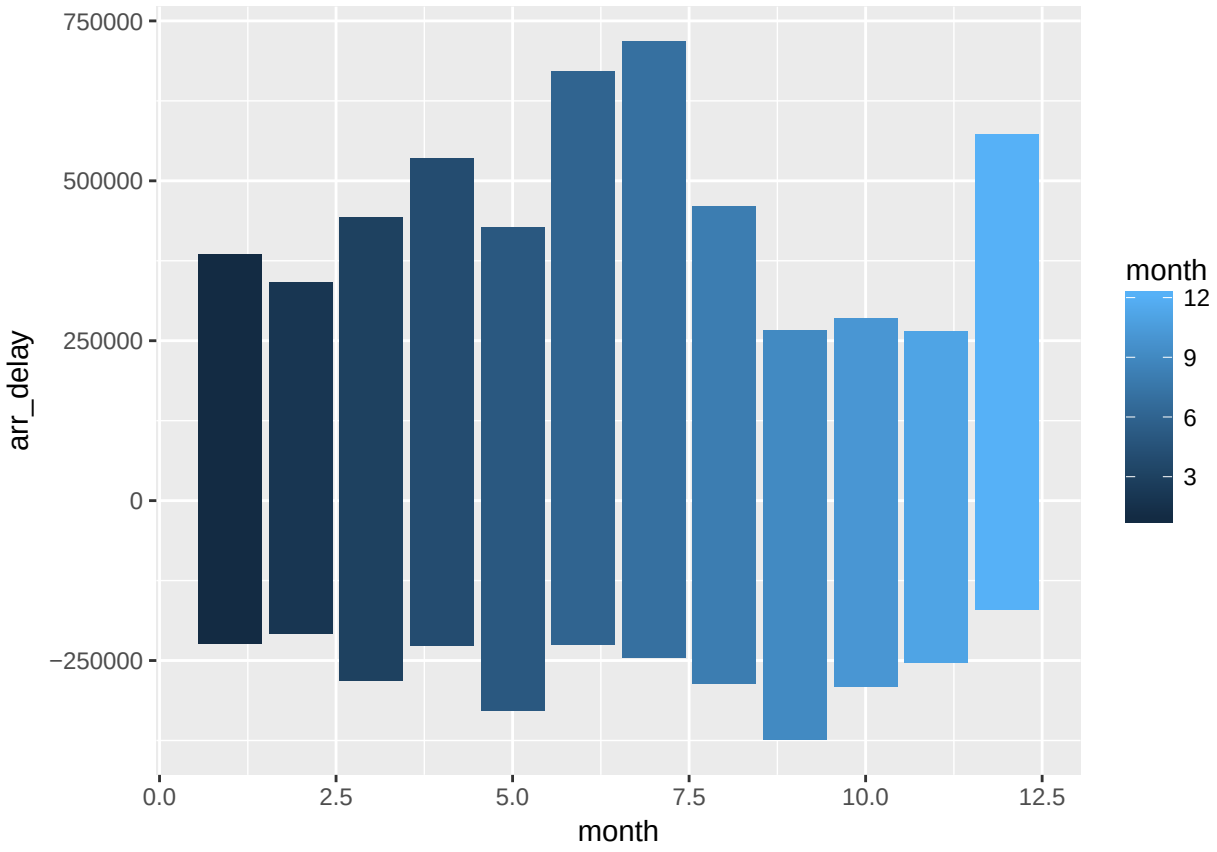
```
ggplot(delay_month, aes(month, mean_delay_month)) + geom_point()
```



Comments -

Above plot indicates that June and July are the months with highest delay, both at about 54 minutes. This would correspond to peak summer. Next in sequence are spring months, which would indicate travel congestions during spring break. The delays are the lowest in Winter months, indicating that Winter weather does not cause increase of delay time.

```
# Following code creates a bar chart of month and arrival delay.
ggplot(flights, aes(month, arr_delay, fill=month)) + geom_bar(stat="identity")
```

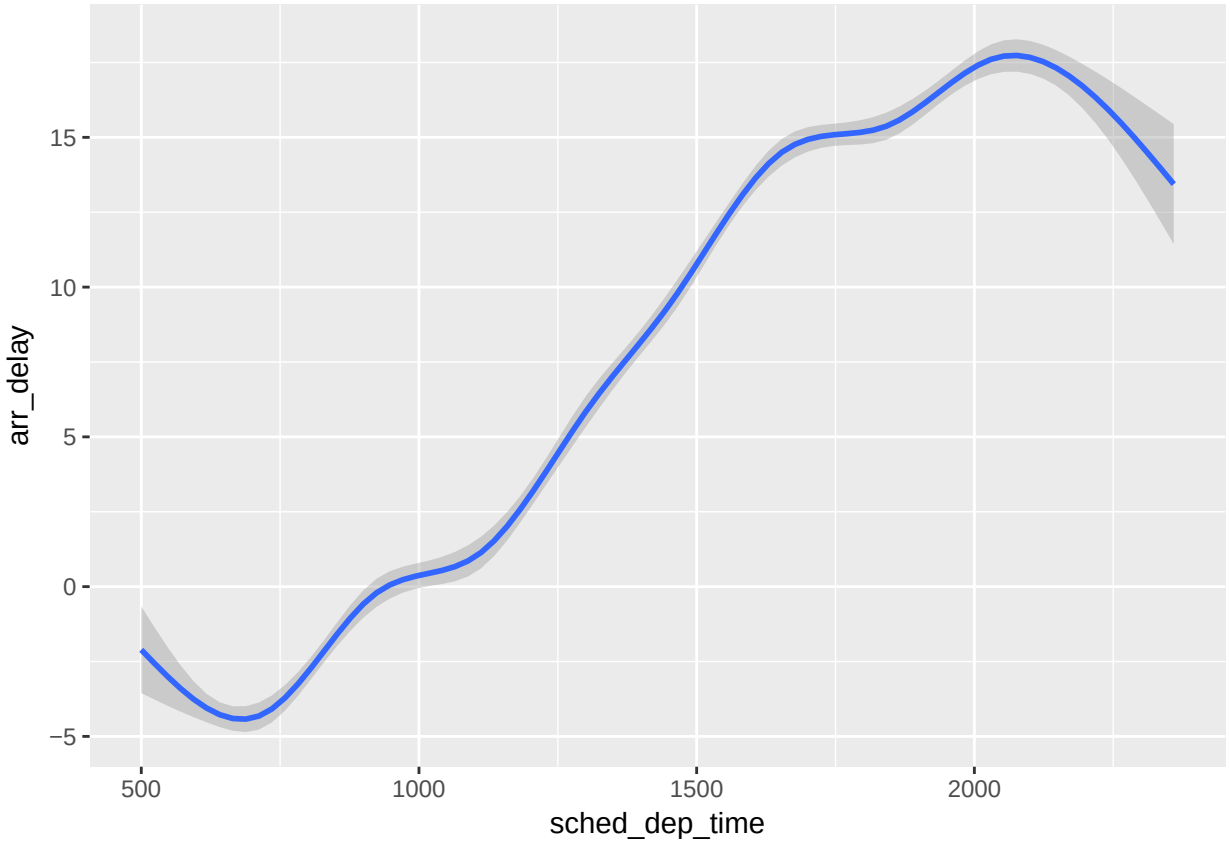


Comments -
Above plot indicates that all months have positive and negative delays. However, in June and July, the number of flights are more than other months, and causes a higher net delay. The least number of flights occur in November and February, and the net delay is lower. December is the most popular winter month for travel, and has the highest delay among winter months.

Conclusion -
Based on above two plots, it can be concluded that a relationship exists between month of travel and arrival delay. It is possible to minimize arrival delays by choosing a specific month.

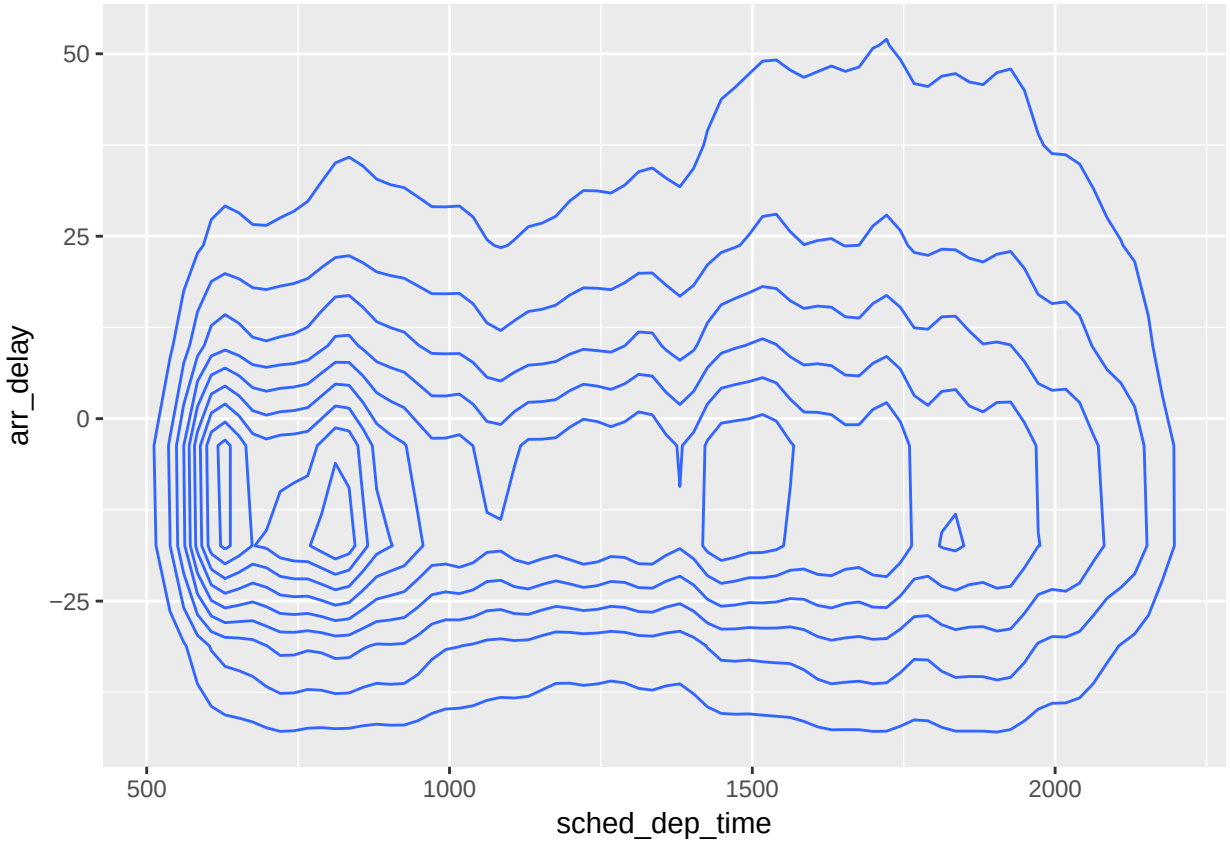
3. Is there a relationship between time of the day the flight is scheduled to leave, and arrival delay? This is interesting since if a relationship is proven, one can take flights during a specific time of the day and experience minimum amount of delay.

```
# Following code creates a plot of Scheduled Departure Time and arrival delay.
ggplot(flights, aes(sched_dep_time, arr_delay)) + geom_smooth(method="auto")
```



Comments -
Above plot indicates that the delays are the least if the flight is scheduled to leave between 12 am to 6 am. After 6 am, the delay gradually increases, with the peak at about 9 PM.

```
# Following code creates a plot of Scheduled Departure Time and arrival delay.
ggplot(flights, aes(sched_dep_time, arr_delay)) + geom_density2d()
```



Comments -
Above plot indicates almost the same fact observed by earlier visualization. If the flight is scheduled to leave between 12 am to 6 am, the delays are the least.

Conclusion -
Based on above two plots, it can be concluded that there is a relationship between scheduled departure time and arrival delay. It is possible to leave at specific times of the day and minimize delays.

(d) Challenge Your Results:

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

Answer -
The answers to Question # 2 and #3 are as expected. I would have expected the relationship to exist between 'month and delay' and 'time of the day and delay'. And the relationships are as I expected.

Answer to #1, relationship between airline and delay was unknown to me. I wasn't sure if it existed, and if delay is something that airlines can control. Some airlines do advertise as always being on time, but the analysis showed no relationship between an airline and arrival delay. There were variations, but there wasn't a clear pattern that could be established.