IMT 573: Problem Set 2

Exploring Data

Jenny Skytta

Due: April 10, 2022

Collaborators: ***Independent Work***

Instructions: Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Cloud.

1. Download the `02_ps_exploringdata.Rmd` file from Canvas or save a copy to your local directory on RStudio Cloud. Supply your solutions to the assignment by editing `02_ps_exploringdata.Rmd`.
2. Replace the "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. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and we encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors you can do so with the `eval=FALSE` option as follows:

```
a + b # these object don't exist
# if you run this on its own it with give an error
```

7. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`, download and rename the knitted PDF file to `ps2_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

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

```
# Load standard libraries
library(tidyverse) # This library gives us access to all the functions we will use
library(nycflights13) # This library provides the data we will use
```

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.

```
# Load the nycflights13 library which includes data on all
# flights departing NYC
data(flights)
# Note the data itself is called flights, we will make it into a local df
# for readability
flights <- tbl_df(flights) #creating a tibble and titling it "flights"
```

```
## Warning: `tbl_df()` was deprecated in dplyr 1.0.0.
## Please use `tibble::as_tibble()` instead.
```

```
# Look at the help file for information about the data
?flights #querying flights data
flights
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      533            529         4      850            830
##  3  2013     1     1      542            540         2      923            850
##  4  2013     1     1      544            545        -1     1004           1022
##  5  2013     1     1      554            600        -6      812            837
##  6  2013     1     1      554            558        -4      740            728
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## # ... with 336,766 more rows, and 11 more variables: 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>
```

```
summary(flights) #running a summary of flights to view the basic summary stats
```

```
##       year          month             day           dep_time    sched_dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 106
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.: 906
##  Median :2013   Median : 7.000   Median :16.00   Median :1401   Median :1359
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349   Mean   :1344
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744   3rd Qu.:1729
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400   Max.   :2359
##                                                  NA's   :8255
##    dep_delay          arr_time    sched_arr_time   arr_delay
##  Min.   : -43.00   Min.   :   1   Min.   :   1   Min.   : -86.000
##  1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124   1st Qu.: -17.000
##  Median :  -2.00   Median :1535   Median :1556   Median :  -5.000
##  Mean   :  12.64   Mean   :1502   Mean   :1536   Mean   :   6.895
##  3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945   3rd Qu.:  14.000
##  Max.   :1301.00   Max.   :2400   Max.   :2359   Max.   :1272.000
##  NA's   :8255      NA's   :8713                  NA's   :9430
##    carrier              flight       tailnum             origin
##  Length:336776      Min.   :   1   Length:336776      Length:336776
```

```
##  Class :character   1st Qu.: 553   Class :character   Class :character
##  Mode  :character   Median :1496   Mode  :character   Mode  :character
##                     Mean   :1972
##                     3rd Qu.:3465
##                     Max.   :8500
##
##      dest              air_time        distance         hour
##  Length:336776      Min.   : 20.0   Min.   :  17   Min.   : 1.00
##  Class :character   1st Qu.: 82.0   1st Qu.: 502   1st Qu.: 9.00
##  Mode  :character   Median :129.0   Median : 872   Median :13.00
##                     Mean   :150.7   Mean   :1040   Mean   :13.18
##                     3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00
##                     Max.   :695.0   Max.   :4983   Max.   :23.00
##                     NA's   :9430
##      minute        time_hour
##  Min.   : 0.00   Min.   :2013-01-01 05:00:00
##  1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
##  Median :29.00   Median :2013-07-03 10:00:00
##  Mean   :26.23   Mean   :2013-07-03 05:22:54
##  3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00
##  Max.   :59.00   Max.   :2013-12-31 23:00:00
##
```

```
# of the tibble
#View(flights) Viewing the dataframe as a table
```

(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.

The *Flights* dataset is a collection of on-time data for all flights that departed NYC in 2013. The dataframe has 19 variables with 33,6776 observations or rows of data. Within the variables, the data collected captures flight scheduled departures, actual departures times, delay time, flight distance, carrier, date of travel, airport flight origin location and destination, and tail number of plane. The source for the data is RITA, Bureau of transportation statistics. In reviewing the FAQ's page from the Bureau of transportation statistics, it states that the 17 carriers listed are the mainline carriers within the US. Many of these mainline carriers operate as code-shares wherein they fly under a mainline carrier's name. To some degree, this muddies the data as we cannot know if a mainline carrier or a code-share carrier is responsible for a given data point. Additionally, its difficult to generalize this data to US domestic carriers overall, as there are about 59 domestic carriers in the US of which this only represents a subset by holdings.

(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.

When viewing the variables, I would like to explore the following questions:

- *Does one carrier have more total delays?-*
- *Is there a time of year when there are more frequent delays?*

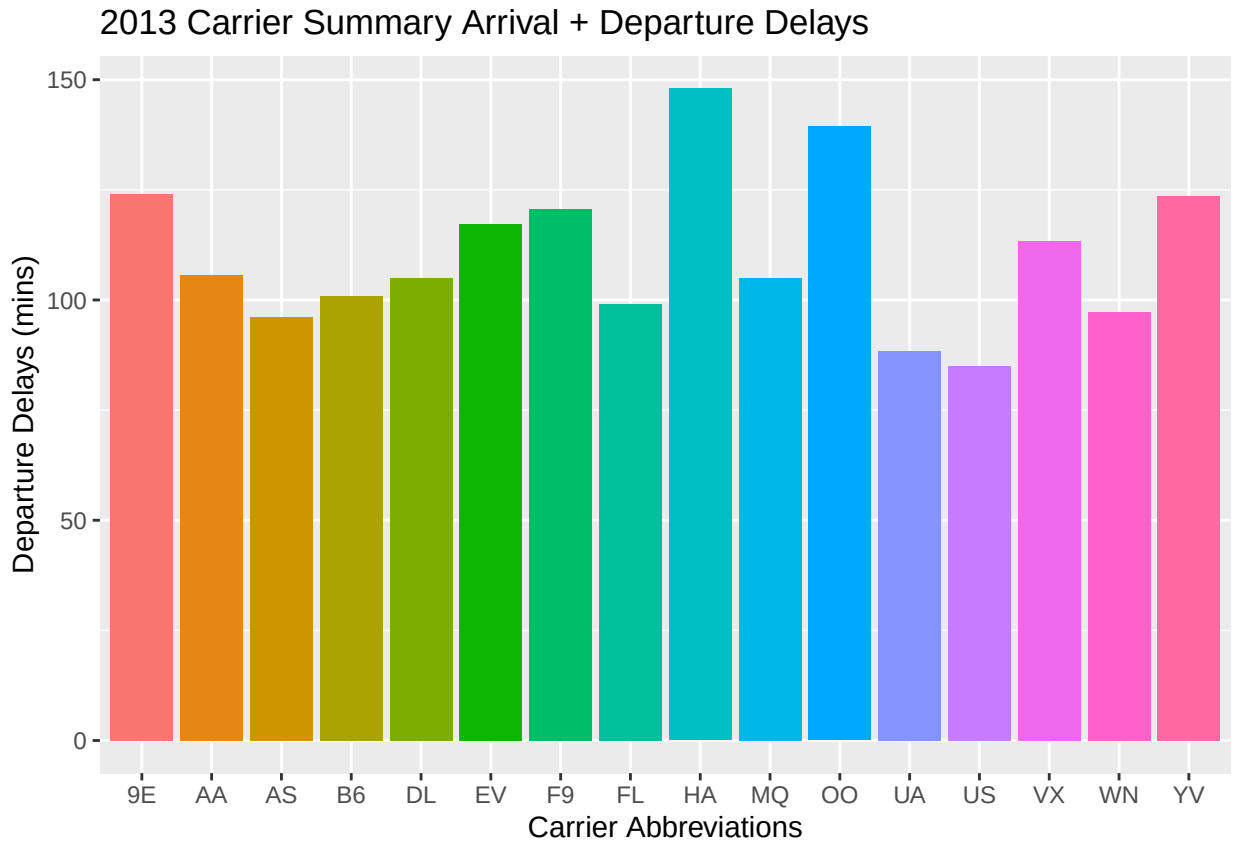
(c) Exploring Data For each of the questions you proposed in Problem 1b, perform an exploratory data analysis designed to address the question. At a minimum, you should produce two visualizations related to each question. Be sure to describe what the visuals show and how they speak to your question of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

Does one carrier have more total delays?

```
delays_by_carrier <- flights %>%
  filter(dep_delay >= 1) %>%
  filter(arr_delay >= 1) %>%
  group_by(carrier) %>%
  summarise("total_delay" = mean(arr_delay + dep_delay))

#View(delays_by_carrier)

ggplot(delays_by_carrier) +
  geom_col(mapping = aes(x = carrier, y = total_delay, fill = carrier), position = "dodge", show.legend
  scale_y_continuous(labels = scales::comma) +
    labs(
    title = "2013 Carrier Summary Arrival + Departure Delays",
    x = "Carrier Abbreviations",
    y = "Departure Delays (mins)", # y-axis label
    fill = "Months"
  )
```



Is there a time of year when there are more frequent delays?

```
freq_delays_bymonth <- flights %>%
  filter(dep_delay >= 1) %>%
  filter(arr_delay >= 1) %>%
```

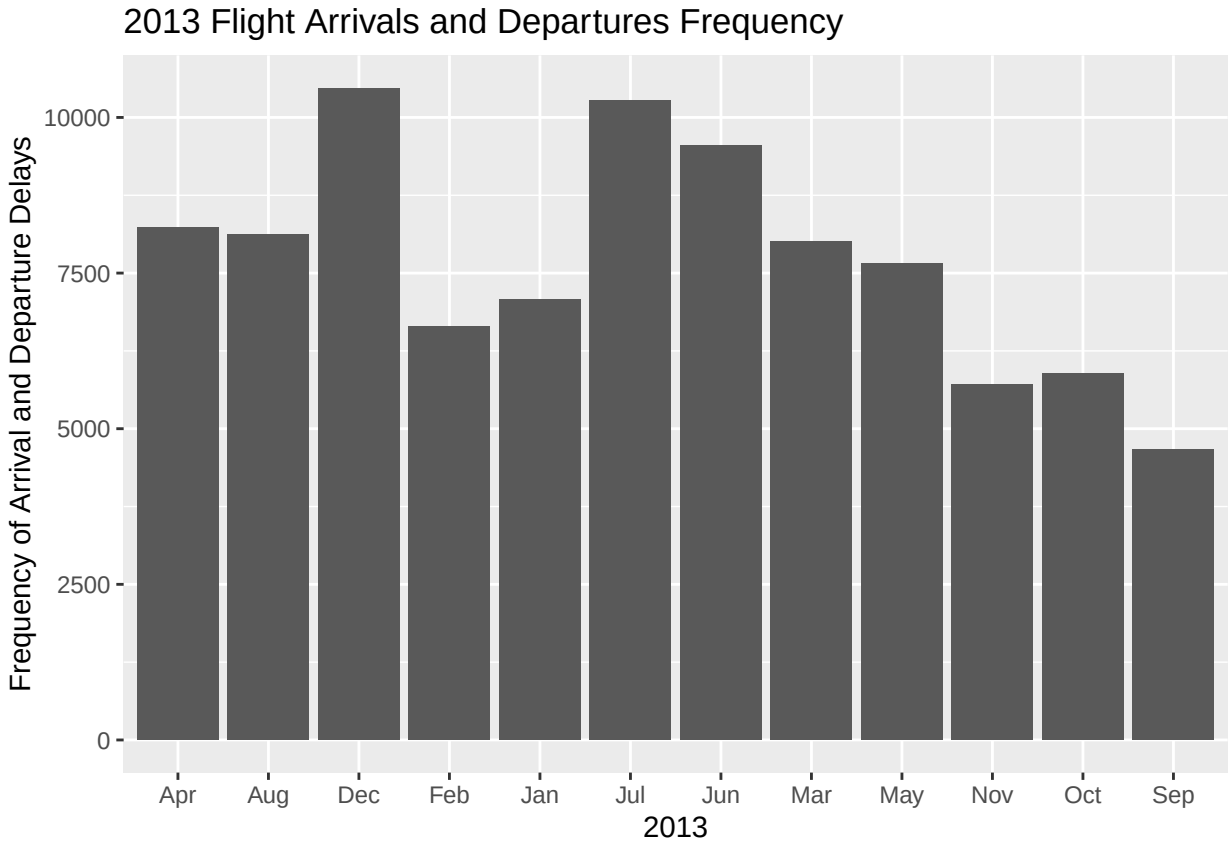
```
  group_by(month) %>%
  select(dep_delay, arr_delay, month) %>%
  summarise(total = n())

#View(freq_delays_bymonth)

freq_delays_bymonth$month <- as.character(freq_delays_bymonth$month)
#changing months to charcter type

freq_delays_bymonth$month[freq_delays_bymonth$month == 1] <- "Jan"  #changing months to charcter names
freq_delays_bymonth$month[freq_delays_bymonth$month == 2] <- "Feb"
freq_delays_bymonth$month[freq_delays_bymonth$month == 3] <- "Mar"
freq_delays_bymonth$month[freq_delays_bymonth$month == 4] <- "Apr"
freq_delays_bymonth$month[freq_delays_bymonth$month == 5] <- "May"
freq_delays_bymonth$month[freq_delays_bymonth$month == 6] <- "Jun"
freq_delays_bymonth$month[freq_delays_bymonth$month == 7] <- "Jul"
freq_delays_bymonth$month[freq_delays_bymonth$month == 8] <- "Aug"
freq_delays_bymonth$month[freq_delays_bymonth$month == 9] <- "Sep"
freq_delays_bymonth$month[freq_delays_bymonth$month == 10] <- "Oct"
freq_delays_bymonth$month[freq_delays_bymonth$month == 11] <- "Nov"
freq_delays_bymonth$month[freq_delays_bymonth$month == 12] <- "Dec"

ggplot(data = freq_delays_bymonth, mapping = aes(x = month, y = total)) +
  geom_col() +
        labs(
    title = "2013 Flight Arrivals and Departures Frequency",
    x = "2013",
    y = "Frequency of Arrival and Departure Delays") # y-axis label
```



(d) Challenge Your Results After completing the exploratory analyses from Problem 1c, do you have any concerns about your findings? How well defined was your original question? Do you still believe this question can be answered using this dataset? Comment on any ethical and/or privacy concerns you have with your analysis.

During the wrangling of my data, I realized that the questions I posed were not well defined. For my first question regarding which carrier has more total delays, I needed an operational definition of "more total delays". For the purposes of the visualization, I defined total delays as a combination of arrival and departure delays combined. This then was a sum of total minutes which makes little sense as one carrier could have one enormous delay while another carrier has more frequent delays throughout the year and the carrier with just one large delay could appear to have "more total delays" by virtue of summing the total. I opted to change to an average of the delays which feels like it captures the essense of the question much better. This way we are looking at the total number of delays (arrival and departure) and the average of those delays to arrive at a number representing the data.For my other question regarding whether there was a time of year wherein there were more frequent delays, I went back and forth about how to visualize and assess this question. The data set again combines arrival and departure delays to capture the total frequency of delays by month. These data appear to have the ring of truth in that one would expect December to have the most frequent delays due to weather and holiday travel. Regarding ethical and privacy concerns, the main concern I have with these data are that they aren't generalizeable in a very meaningful way since they're one year from one city in the US. While there may be privacy concerns regarding flight data in general given its been a target for terrorists, I being that we should have this information for public discourse and review. Most recently, a twitter user was posting flight data for Elon Musk and was asked to remove the information as a security risk. In most realms, one must do a comparative cost (risk)/ benefit assessment. Does sharing these data pose a greater threat than the benefit? I would argue that the transparency of these data benefit society more than they pose any risk. With these data, we could ostensibly calculate validity of carrier cost increases.

Citations

RITA, Bureau of transportation statistics, https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236

Code written above is from the previous course IMT 511 which used the below text to support class scripts. https://www.google.com/books/edition/Programming_Skills_for_Data_Science/BnB6DwAAQBAJ?hl=en&gbpv=1&printsec=frontcover