IMT 573: Problem Set 2

Exploring Data

Tommy Huynh

Due: July 4, 2021

Collaborators:

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:

```r
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
```

```
## Warning: package ’tidyr’ was built under R version 4.1.1

## Warning: package ’readr’ was built under R version 4.1.1
```

```
library(nycflights13) # This library provides the data we will use
```

```
## Warning: package ’nycflights13’ was built under R version 4.1.1
```

```
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.

```
# Load the nycflights13 library which includes data on all
# lights 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)
```

```
## Warning: ‘tbl_df()‘ was deprecated in dplyr 1.0.0.
## Please use ‘tibble::as_tibble()‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was generated.
```

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

(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 data was collected by the Bureau of Transportation Statistics.The data has all airline on-time data for all flights departing from New York City in 2013. It also has information regarding airlines, airports, weather, and specific planes.

```
summary(flights)
```

```
##       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
##
```

According to the summary statistics, the average departure delay was 12 minutes, and the average arrive delay was close to 7 minutes for the year 2013.

(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.

A good question would be what airlines are operating flights out of New York City. I think this is a good question because it gives us knowledge on what is available to get out of the city.

Another good question I would ask is what airline flew the most flights out of NYC in 2013. I think this is a good question because we can then determine which airline has the largest operating presence in NYC.

(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.

This code removes all duplicates of carriers so that only one will exist in the data frame, giving us a list of all carriers that fly routes in and out of NYC.

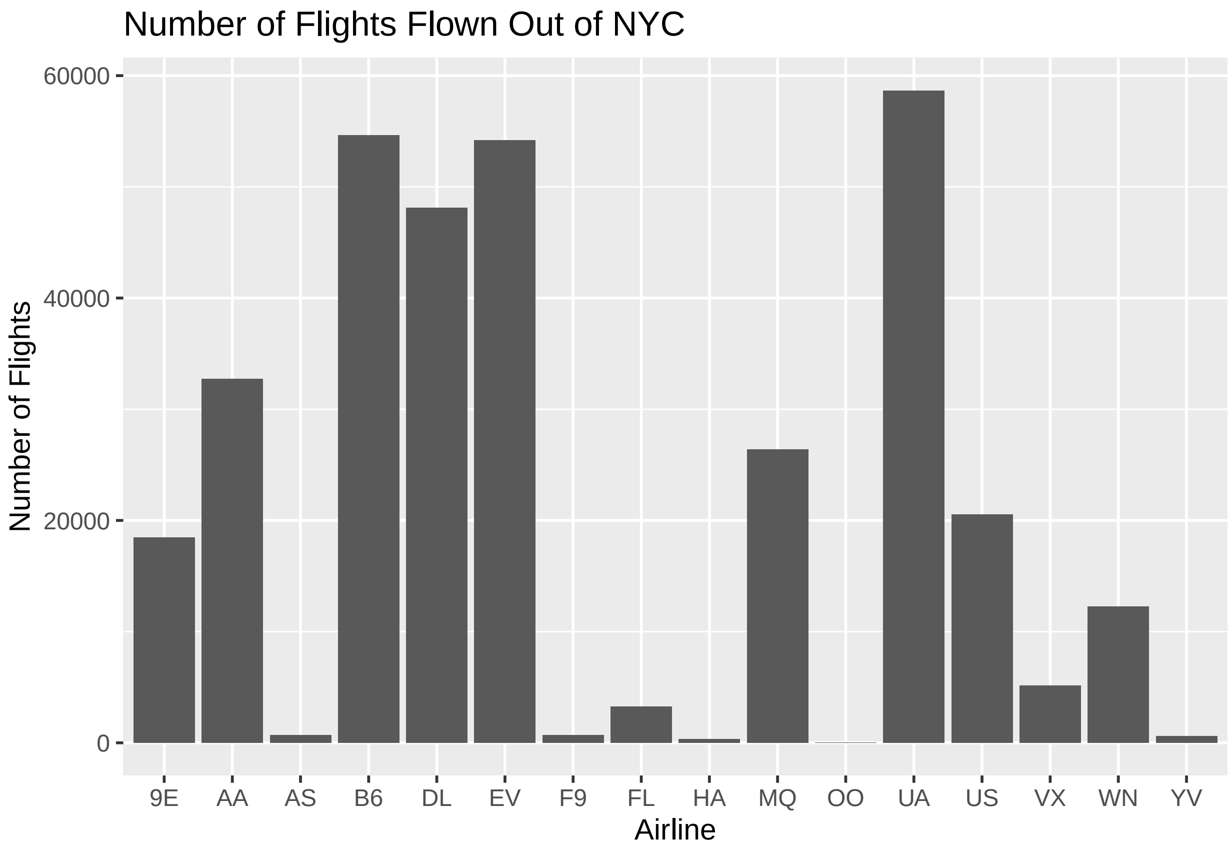
```
airlines <- flights
airlines <- airlines[!duplicated(airlines$carrier),]
airlines <- airlines$carrier
```

This barplot contains the total flight count for the year 2013 for each carrier:

```
flight_count <- count(flights, vars = carrier)
as.data.frame(flight_count)
```

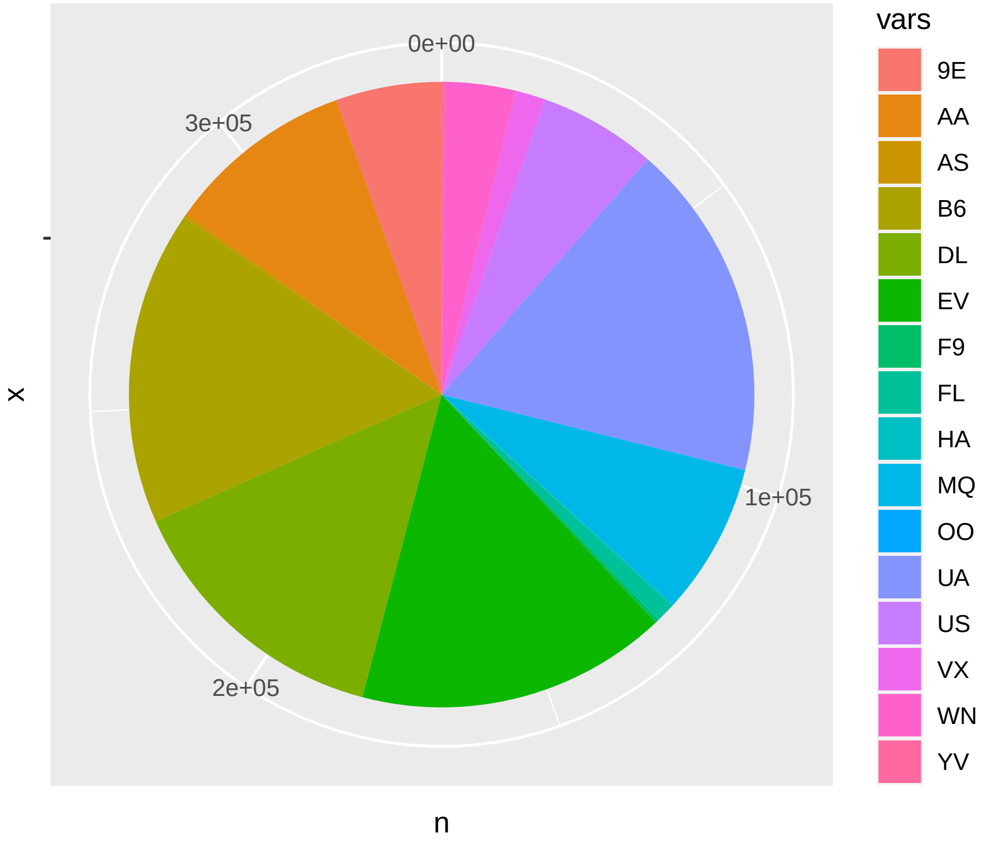
```
##    vars     n
## 1    9E 18460
## 2    AA 32729
## 3    AS   714
## 4    B6 54635
## 5    DL 48110
## 6    EV 54173
## 7    F9   685
## 8    FL  3260
## 9    HA   342
## 10   MQ 26397
## 11   OO    32
## 12   UA 58665
## 13   US 20536
## 14   VX  5162
## 15   WN 12275
## 16   YV   601
```

```
ggplot(flight_count, aes(x=vars, y=n)) +
  geom_bar(stat="identity") +
  labs(title = "Number of Flights Flown Out of NYC", x = "Airline", y = "Number of Flights")
```



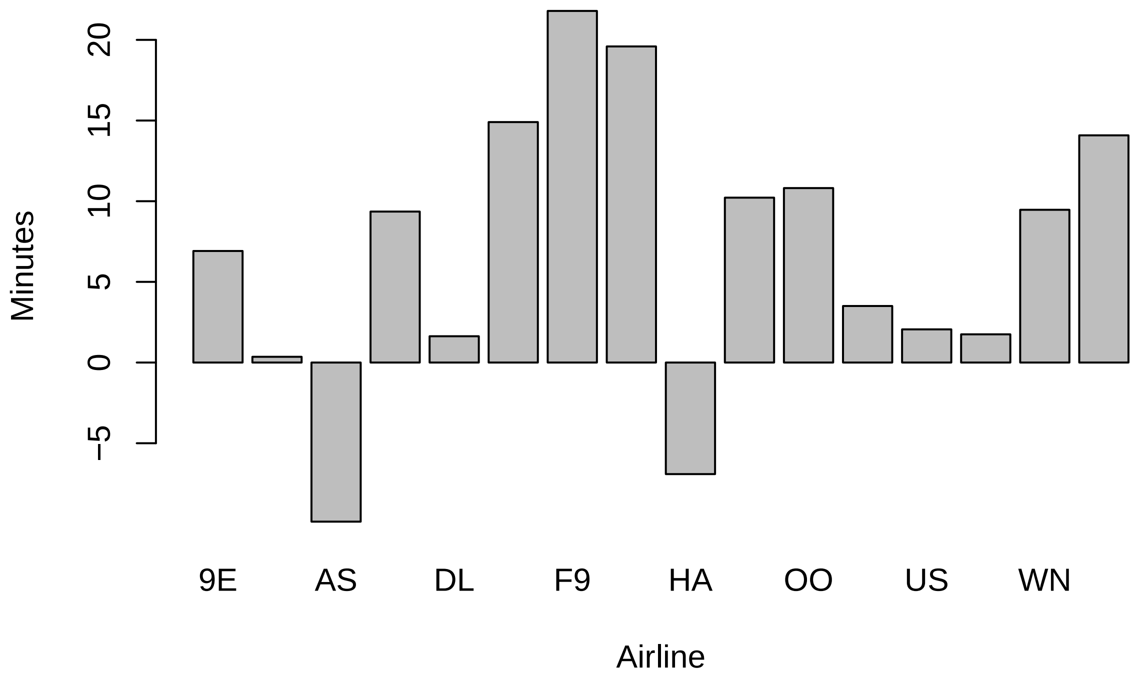
This piechart contains the total flight count for the year 2013 for each carrier:

```
ggplot(flight_count, aes(x="", y=n, fill=vars)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0)
```



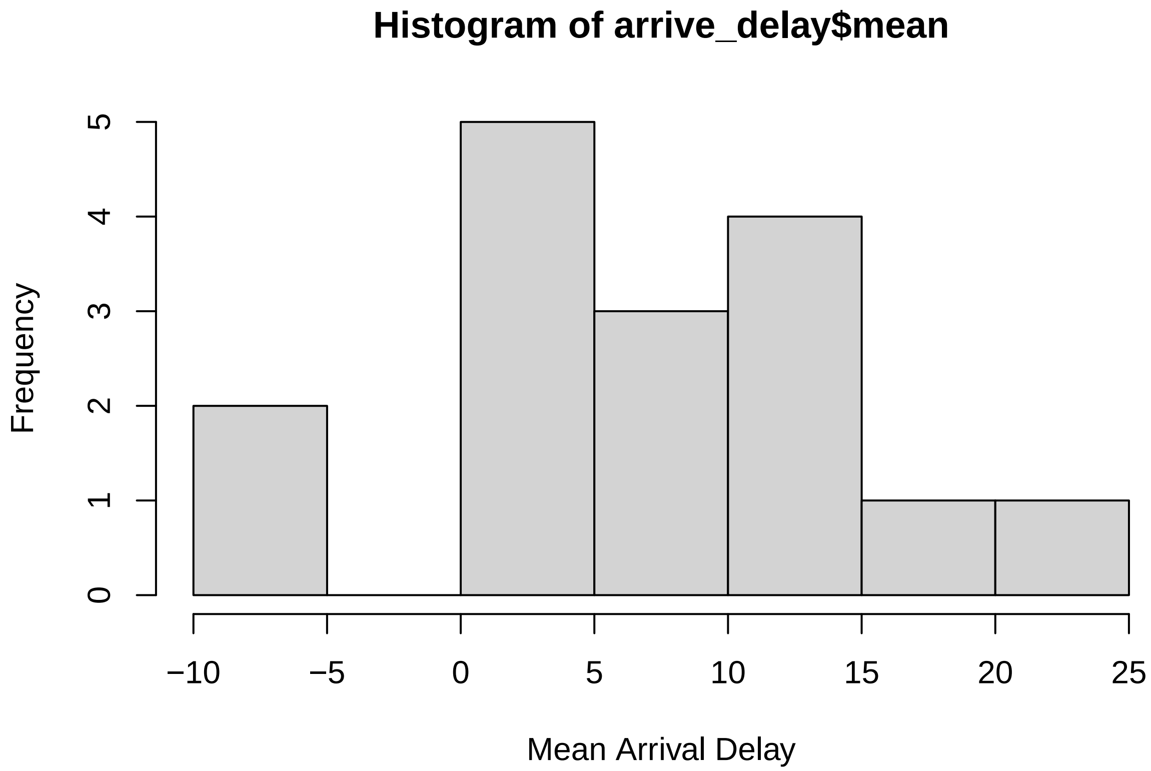
This code aggregates the total minutes of arrival delay for each carrier. It then calculates the total amount of flights for each carrier and calculates the mean arrival delay.

```r
number_flights <- count(flights, carrier)

arrive_delay <- aggregate( cbind( arr_delay ) ~ carrier , data = flights , FUN = sum )

arrive_delay <- merge(arrive_delay, number_flights)
arrive_delay$mean <- arrive_delay$arr_delay/arrive_delay$n


barplot(arrive_delay[ ,4], names.arg = arrive_delay[ ,1], xlab = "Airline", ylab = "Minutes")
```



This is a histogram of the most common mean arrival time relative to the scheduled time of arrival for all airlines.

```
hist(arrive_delay$mean, xlab = "Mean Arrival Delay")
```



##### (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.

I do not have any concerns regarding my findings. I find that the low cost airlines tend to have the longest delays, while the more expensive ones are relatively on time or even early in some cases. I think my original question was well defined. I purposely framed a question that I could answer because I didn't want to pose a question that was out of scope of the data set. I do not have any ethical or privacy concerns with this analysis in specific because it is very public data.