IMT 573: Problem Set 2

Exploring Data

Ali Qazi

Due: July 4, 2021

Collaborators: Akeel Qazi

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
# 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.
```

```
# 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>
```

```
# summary(flights)
```

```
vi
```

(a) Importing and Inspecting Data

```
## function (name = NULL, file = "")
## edit.default(name, file, editor = "vi")
## <bytecode: 0x563413b1d6b0>
## <environment: namespace:utils>
```

```
glimpse(flights)
```

```
## Rows: 336,776
## Columns: 19
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2~
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

```
## $ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558, ~
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 600, ~
## $ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2, -1~
## $ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 849,~
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745, 851,~
## $ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7, -1~
## $ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6", "~
## $ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301, 4~
## $ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N394~
## $ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LGA",~
## $ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IAD",~
## $ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149, 1~
## $ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 733, ~
## $ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6, 6~
## $ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 0, 59, 0~
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 0~
```

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.

Looking at the data, there is information on all 336,776 flights that is found in the nycflights13 package. There are 19 variables for each flight from the year 2013. Using the glimpse() function we can explore a data frame with the first few entries. Each column includes a data type: int, dbl, chr, and dttm. Quantitative variables within the data frame are dbl and int. Dttm is a time_hour variable that gives the time of the day and the date. ##### (b) Formulating Questions

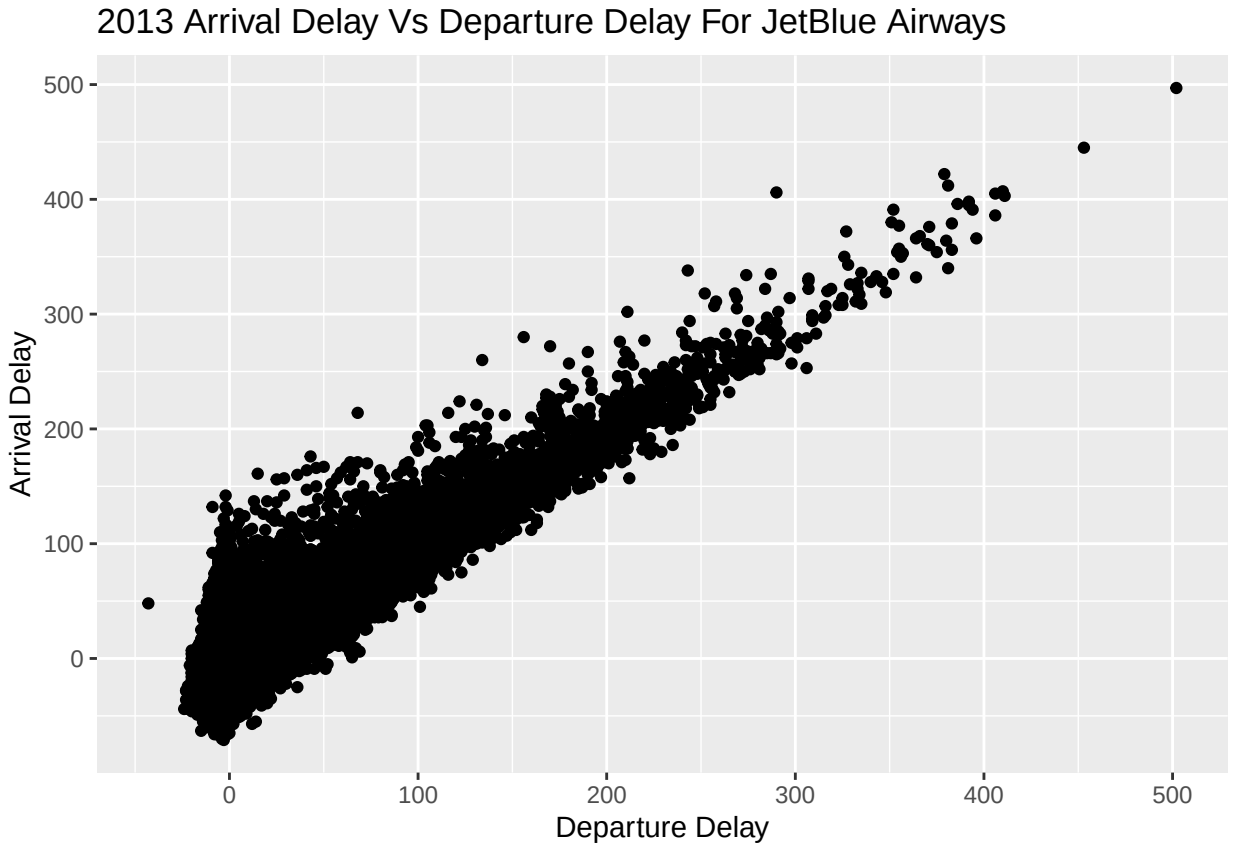
Consider the NYC flights data. Formulate two motivating questions you want to explore using this data. Describe why these questions are interesting and how you might go about answering them.

1. How does the arrival delays relate to the departure delays for Jet Blue Airways?
2. How often was the arrival delays in 2013 for all NYC flights?

(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.

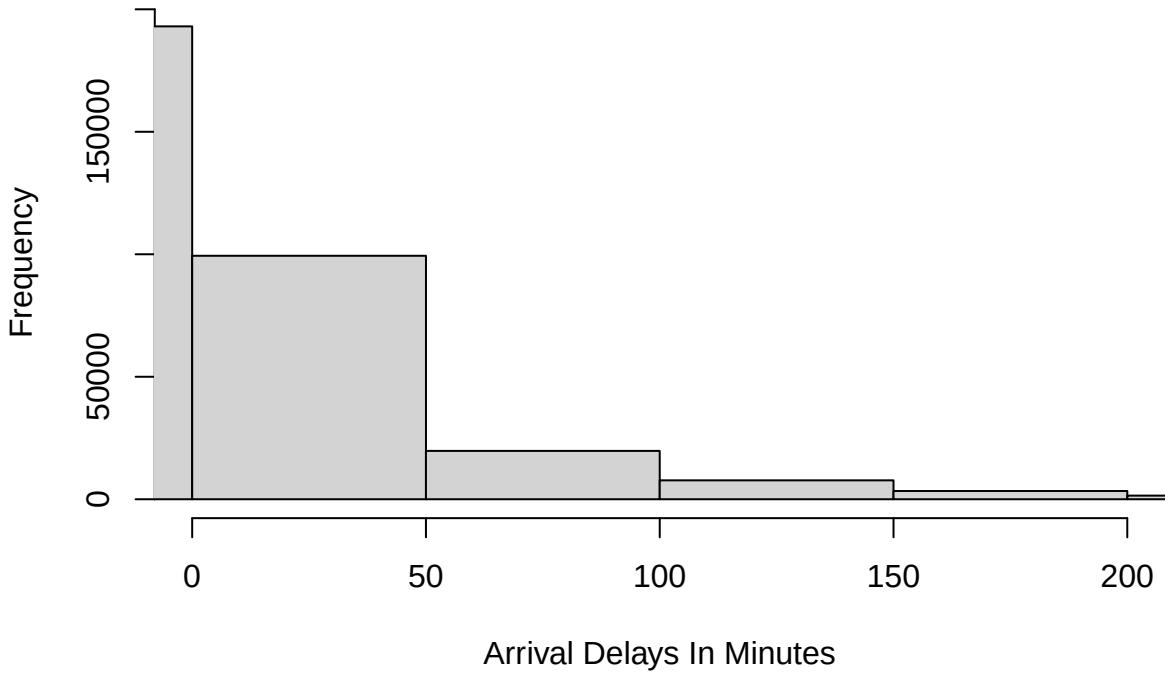
```
JetBlue_flights <- flights %>%
filter(carrier == "B6")
ggplot(data = JetBlue_flights,
mapping = aes(x = dep_delay, y = arr_delay)) +
ggtitle("2013 Arrival Delay Vs Departure Delay For JetBlue Airways") +
xlab("Departure Delay") +
ylab("Arrival Delay") +
geom_point()
```

```
## Warning: Removed 586 rows containing missing values (geom_point).
```



The visual above shares the relationship between arrival and departure delays for Jet Blue Airways and New York flights in the year 2013. With the scatter plot we can conclude that there is a positive relationship between the x and y axis that represents the arrival delays and departure delays. To answer the first question above, as the arrival delays increase, the departure delays increase as well.

```
hist(flights$arr_delay, main = "Amount of Arrival Delays",
xlab = "Arrival Delays In Minutes",
xlim = c(0,200))
```



This histogram answers the question of how often NYC flights had arrival delays in the year of 2013. Looking at the x axis, we can see that most of the arrival delays were in the negative. This means that the plane arrived before the scheduled arrival time!

(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.

Some concerns I had with my findings was the 586 rows that were removed due to missing values for the first visualization (geom_point). The data set has 300,000 plus observations meaning that there is a high chance for missing values. My original questions were specific to certain variables such as arrival time and a type of airline carrier. I had to first look up what carrier B6 as I did not know it was Jet Blue Airways. If the data set did not have missing values, then I believe my questions can be answered correctly and the data could be more accurate. The dataset did not include any passenger or crew member information. Including passenger and crew member information with each flight would be redundant and a privacy issue.