IMT 573: Problem Set 2 - Working with Data

Vighnesh Misal

Due: Tuesday, October 15, 2019

Collaborators: Ashish Anand

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 Server.

1. Download the `problemset2.Rmd` file from Canvas or save a copy to your local directory on RStudio Server. Open `problemset2.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset2.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. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do not 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 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 withouth errors you can do so with the `eval=FALSE` option.

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 PDF`, 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)

## Warning: package 'tidyverse' was built under R version 3.6.1

## Warning: package 'dplyr' was built under R version 3.6.1

## Warning: package 'stringr' was built under R version 3.6.1

library(nycflights13)

## Warning: package 'nycflights13' was built under R version 3.6.1

library(dplyr)

flightdata <- nycflights13::flights
flight_df1 <- nycflights13::flights
flight_df2 <- tbl_df(flightdata)
```

Problem 1: Describing the NYC Flights Data

In this problem set we will continue to use the data on all flights that departed NYC (i.e. JFK, LGA or EWR) in 2013. Recall, you can find this data in the `nycflights13` R package. Load the data in R and ensure you know the variables in the data. Keep the documentation of the dataset (e.g. the help file) nearby.

In Problem Set 1 you started to explore this data. Now we will perform a more thorough description and summarization of the data, making use of our new data manipulation skills to answer a specific set of questions. When answering these questions be sure to include the code you used in computing empirical responses, this code should include code comments. Your response should also be accompanied by a written explanation, code alone is not a sufficient response.

(a) Describe and Summarize

Answer the following questions in order to describe and summarize the `flights` data.

\begin{enumerate}

How many flights out of NYC are there in the data?

```
nrow(flight_df1)

## [1] 336776
```

How many NYC airports are included in this data? Which airports are these?

```
res_unique_origin <- unique(flight_df1$origin)
res_unique_origin
```

```
## [1] "EWR" "LGA" "JFK"
```

```
length(res_unique_origin)
```

```
## [1] 3
```

Into how many airports did the airlines fly from NYC in 2013?

```
res_unique_dest <- unique(flight_df1$dest)
res_unique_dest
```

```
##   [1] "IAH" "MIA" "BQN" "ATL" "ORD" "FLL" "IAD" "MCO" "PBI" "TPA" "LAX"
##  [12] "SFO" "DFW" "BOS" "LAS" "MSP" "DTW" "RSW" "SJU" "PHX" "BWI" "CLT"
##  [23] "BUF" "DEN" "SNA" "MSY" "SLC" "XNA" "MKE" "SEA" "ROC" "SYR" "SRQ"
##  [34] "RDU" "CMH" "JAX" "CHS" "MEM" "PIT" "SAN" "DCA" "CLE" "STL" "MYR"
##  [45] "JAC" "MDW" "HNL" "BNA" "AUS" "BTV" "PHL" "STT" "EGE" "AVL" "PWM"
##  [56] "IND" "SAV" "CAK" "HOU" "LGB" "DAY" "ALB" "BDL" "MHT" "MSN" "GSO"
##  [67] "CVG" "BUR" "RIC" "GSP" "GRR" "MCI" "ORF" "SAT" "SDF" "PDX" "SJC"
##  [78] "OMA" "CRW" "OAK" "SMF" "TUL" "TYS" "OKC" "PVD" "DSM" "PSE" "BHM"
##  [89] "CAE" "HDN" "BZN" "MTJ" "EYW" "PSP" "ACK" "BGR" "ABQ" "ILM" "MVY"
## [100] "SBN" "LEX" "CHO" "TVC" "ANC" "LGA"
```

```
length(res_unique_dest)
```

```
## [1] 105
```

How many flights were there from NYC to Seattle (airport code)?

```
flight_df1 %>% filter(dest == "SEA") %>% NROW()
```

```
## [1] 3923
```

```
flight_df1 %>% filter(dest == "SEA") %>% summarise(count = n())
```

```
## # A tibble: 1 x 1
##   count
##   <int>
## 1  3923
```

Were there any flights from NYC to Spokane ?

```
flight_df1 %>% filter(dest == "GEG") %>% summarise(count = n())
```

```
## # A tibble: 1 x 1
##   count
##   <int>
## 1     0
```

Are there missing destination codes? (i.e. are there any destinations that do not look like valid airport codes (i.e. three-letter-all-upper case)?)

```
res_incorrect_length <- nchar(flight_df1$dest) == 3
length(res_incorrect_length)
```

```
## [1] 336776
```

```
res_incorrect_codes <- str_detect(flight_df1$dest, "[[:lower:]]")
length(res_incorrect_codes)
```

```
## [1] 336776
```

```
res_final <- nchar(flight_df1$dest) == 3 && !str_detect(flight_df1$dest,
"[[:lower:]]")
length(res_final)
```

```
## [1] 1
```

```
unique(nchar(flights$dest))
```

```
## [1] 3
```

All codes are correct. \end{enumerate}

(b) Reflect and Question

What are your thoughts on the questions (and answers) so far? Were you able to answer all of these questions? Are all questions well defined? Is the data suitable for answering all these?

ANS: Yes, all the questions are well defined and I was able to answer all of them. The data is sufficient for answering the questions.

Problem 2: NYC Flight Delays

Flights are often delayed. Let's look at closer at this topic using the NYC Flight dataset. Answer the following questions about flight delays using the `dplyr` data manipulation verbs we talked about in class.

(a) Typical Delays

What is the typical delay for a flight in this data?

```
flight_df2 <- flight_df2 %>% mutate(total_delay = dep_delay + arr_delay)
mean1 <- mean(flight_df2$total_delay, na.rm = TRUE)
mean1
```

```
## [1] 19.45053
```

(b) Defining Flight Delays

What definition of flight delay did you use to answer part (a)? Did you do any specific exploration and description of this variable prior to using it? If no, please do so now. Is there any missing data? Are there any implausible or invalid entries?

The original flight data had two types of delays, arrival delay and departure delay. I did a sum of the delays and created a new column called total delay.

The data in the delay columns has "NA" and "negative" values. These either signify a cancelled flight or an early flight. Although the NA values were handled by the na.rm argumnent. The negative values can impact the observed mean as we need to find average delay for flights that were delayed and not the ones that came in before time.

Also, I feel that departure delay is insignificant in the analysis as total delay would always be how much time the flight was late with respect to the actual landing time. Time lost in departure delay could be covered in transit.

(c) Delays by Destination

Now, compute flight delay by destination. Which are the worst three destinations from NYC if you don't like flight delays? Be sure to justify your delay variable choice.

```
res <- flight_df2 %>% group_by(dest) %>% summarise(Avg_delay =
median(total_delay, na.rm = TRUE)) %>% arrange(desc(Avg_delay))
res
```

```
## # A tibble: 105 x 2
##    dest  Avg_delay
##    <chr>     <dbl>
##  1 CAE        42
##  2 JAC        33
##  3 OKC        23
##  4 TUL        20.5
##  5 SBN        12
##  6 EYW        10
##  7 ANC         7.5
##  8 SMF         6
##  9 HDN         5.5
## 10 CAK         3
## # ... with 95 more rows
```

(d) Seasonal Delays

Flight delays may be partly related to weather, as you may have experienced yourself. We do not have weather information here but let's analyze how delays are related to season.

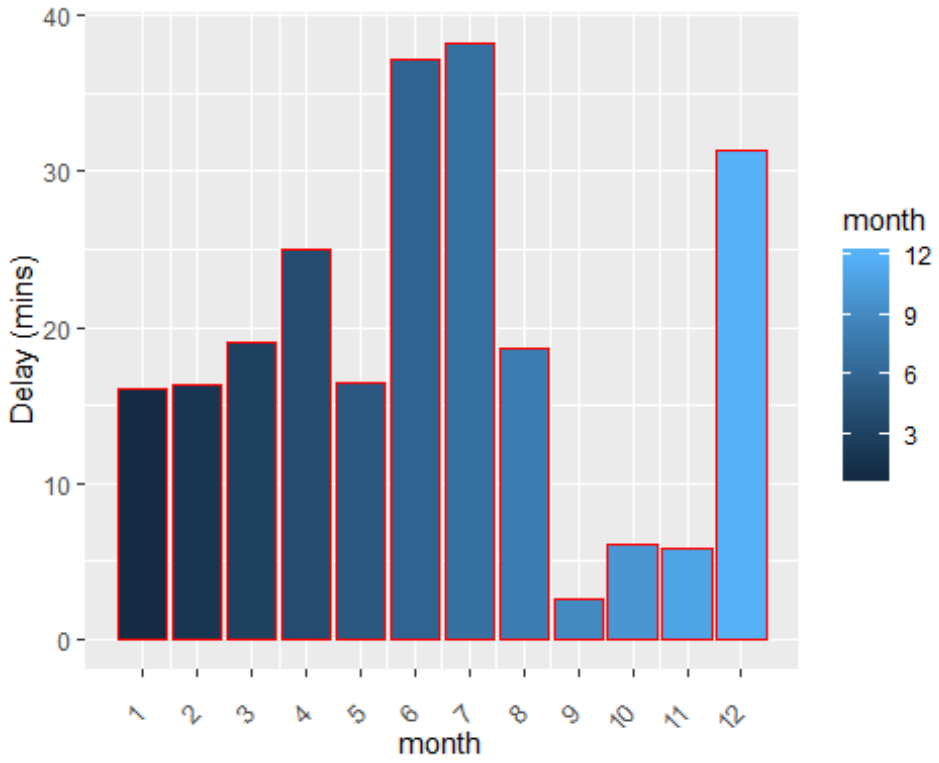
Which seasons have the worst flights delays? Why might this be the case? In your communication of your analysis, use one graphical visualization and one tabular respresentation of your findings.

```
res2 <- flight_df2 %>% group_by(month) %>% summarise(Avg_delay2 =
mean(total_delay, na.rm = TRUE)) %>% arrange(desc(Avg_delay2))
res2 <- tbl_df(res2)
res2
```

```
## # A tibble: 12 x 2
##    month Avg_delay2
##    <int>      <dbl>
##  1     7      38.2
##  2     6      37.2
##  3    12      31.4
##  4     4      25.0
##  5     3      19.0
##  6     8      18.6
##  7     5      16.4
##  8     2      16.4
##  9     1      16.1
## 10    10       6.07
## 11    11       5.88
## 12     9       2.61
```

```
plot1 <- ggplot(res2, mapping = aes(month, Avg_delay2, fill = month)) +
geom_bar(stat = "identity", color ="red") + theme(axis.text.x =
element_text(angle = 45,hjust=1,vjust=0.3)) + xlab("Month") +  ylab("Delay
(mins)")


plot1 + scale_x_continuous("month", labels = as.character(res2$month), breaks
= res2$month)
```



The flights were delayed the most in Jun/JUly (summer) and december(winter)

(e) Challenge Your Results

After completing the exploratory analyses from Problem 2, 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 am concerned about how data in the delay column can cause confusion about the net delay that should be calculated as well as using rudimentary statistics to reach conclusions about this data is a big mistake.

Problem 3: Let's Fly Across the Country!

(a) Describe and Summarize

Answer the following qeustions to describe and summarize the `flights` data, focusing on flights from New York to Portland, OR (airport code `PDX`).

\begin{enumerate}

How many flights were there from NYC airports to Portland in 2013?

```
res_NYC_Portland_flight <- flight_df1 %>% filter(dest == "PDX", year ==
"2013") %>% summarise(count = n())

res_NYC_Portland_flight
```

```
## # A tibble: 1 x 1
##   count
##   <int>
## 1  1354
```

How many airlines fly from NYC to Portland?

```
res_NYC_Portland_airline <- flight_df1 %>% filter(dest=="PDX")
unique(res_NYC_Portland_airline$carrier)
```

```
## [1] "DL" "UA" "B6"
```

```
length(unique(res_NYC_Portland_airline$carrier))
```

```
## [1] 3
```

Which are these airlines (find the 2-letter abbreviations)? How many times did each of these go to Portland?

```
res_NYC_Portland_airline <- flight_df1 %>% filter(dest=="PDX")
unique(res_NYC_Portland_airline$carrier)
```

```
## [1] "DL" "UA" "B6"
```

How many unique airplanes fly from NYC to PDX? \

```
res_NYC_Portland_airplane <- flight_df1 %>% filter(dest == "PDX")
unique(res_NYC_Portland_airplane$tailnum) %>% length()
```

```
## [1] 492
```

How many different airplanes flew from each of the three NYC airports to Portland?

```
res_NYC_Portland_airplane <- flight_df1 %>% group_by(origin) %>% filter(dest
== "PDX") %>%summarise(count = n())
res_NYC_Portland_airplane
```

```
## # A tibble: 2 x 2
##   origin count
##   <chr>  <int>
## 1 EWR      571
## 2 JFK      783
```

What percentage of flights to Portland were delayed at departure by more than 15 minutes?

```
res_NYC_Portland_depart <- flight_df1 %>% filter(dest == "PDX", dep_delay > 15) %>% summarise(count = n())
res_NYC_Portland_depart_total <- flight_df1 %>% filter(dest == "PDX") %>% summarise(count = n())

(res_NYC_Portland_depart/res_NYC_Portland_depart_total)*100
```

```
##      count
## 1 26.66174
```

Is one of the New York airports noticeably worse in terms of departure delays for flights to Portland than others?

```
res_NYC_Portland_depart1 <- flight_df1 %>% group_by(origin) %>% filter(dest == "PDX", dep_delay > 15) %>% summarise(count = n())

res_NYC_Portland_depart_total1 <- flight_df1 %>% group_by(origin) %>% filter(dest == "PDX") %>% summarise(count = n())

res_NYC_Portland_depart1
```

```
## # A tibble: 2 x 2
##   origin count
##   <chr>  <int>
## 1 EWR      168
## 2 JFK      193
```

```
res_NYC_Portland_depart_total1
```

```
## # A tibble: 2 x 2
##   origin count
##   <chr>  <int>
## 1 EWR      571
## 2 JFK      783
```

```
final_res <-
full_join(res_NYC_Portland_depart1,res_NYC_Portland_depart_total1,by = "origin" )
final_res <- final_res %>% mutate(significant_delays = (final_res$count.x/final_res$count.y)*100)
final_res
```

```
## # A tibble: 2 x 4
##   origin count.x count.y significant_delays
##   <chr>    <int>   <int>              <dbl>
## 1 EWR        168     571               29.4
## 2 JFK        193     783               24.6
```

We can observe from the above table that 24.6% of the flights that flew to Portland from JFK had been delayed by more than 15 mins as opposed to 29.42% from EWR. So JFK is slightly better than EWR.

\end{enumerate}

(b) Reflect and Question

What are your thoughts on the questions (and answers) examining flights to Portland? Were you able to answer all of these questions? Are all questions well defined? Is the data suitable for answering all these?

I was able to answer most of them in terms of writing code. I am not sure if the measures I used are apt representations and if there are additonal measures that would give a more accurate description of what's happening under the hood.