IMT 573: Problem Set 1 - Exploring Data

Alexander Davis

Due: Tuesday, October 13, 2020 by 9am PT

Collaborators:

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. Collaboration shouldn't be confused with group project work (where each person does a part of the project). Working on problem sets should be your individual contribution. More on that in point 8.
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. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment. In particular, note that Stack Overflow is licenses as Creative Commons (CC-BY-SA). This means you have to attribute any code you refer from SO.
5. Partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. But please **DO NOT** submit pages and pages of hard-to-read code and attempts that is impossible to grade. That is, avoid redundancy. Remember that one of the key goals of a data scientist is to produce coherent reports that others can easily follow. 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 dont' 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 PDF`, rename the knitted PDF file to `ps1_ourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.
8. Collaboration is often fun and useful, 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.

Problem 1: Basic R Programming Write a function, `calculate_bmi` to calculate a person's body mass index, when given two input parameters, 1). weight in pounds and 2) height in inches.

NOTE: You would have to go to external sources to find the formula of bmi. In your response, before presenting your code for the function, tell us your official reference for the BMI formulae.

Insert Response first https://www.bmi-calculator.net/bmi-formula.php

```
calculate_bmi <- function(weight, height) {
  return((weight / (height * height)) * 703)
}
```

Insert code. Your code should appear within R Code Chunks.

Problem 2: 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.

Setup: Problem 2 You will need, at minimum, the following R packages. The data itself resides in package *nycflights13*. You may need to install both.

```
# Load standard libraries
library(dplyr)
library(tidyverse)
library('nycflights13')
```

```
# 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()` is deprecated as of dplyr 1.0.0.
## Please use `tibble::as_tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_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>
```

```
# summary(flights)
```

(a) Importing Data Load the data and describe in a short paragraph how the data was collected and what each variable represents.

This is the data collected over the period of a year from flights going into and out of NYC. The variables include; YEAR, MONTH, DAY as well as departure time (dep_time) which would be in a time format of 5:17 for example, a scheduled departure time (shed_dep_time) in similar format, departure delay (dep_delay) in minutes either positive or negative values depending if the flight departed early or late, an arrival time (arr_time) in a similar time format, a scheduled arrival time (shed_arr_time), and an arrival delay (arr_delay) which is also in a minutes format. There is then the carrier which is just 2 letters, the flight number (numbers), a tail number which is a combination of letters and numbers referencing the specific plane, the origin airport and the destination airport. Total Air_time which is in minutes, distance in miles, and then hour and minute column followed by time_hour.

(b) Inspecting Data Perform a basic inspection of the data and discuss what you find. Inspections may involve asking the following questions (the list is not inclusive, you may well ask other questions):

- How many distinct flights do we have in the dataset?
- How many missing values are there in each variable?
- Do you see any unreasonable values? *Hint: Check out min, max and range functions.*

```
unique(flights$carrier)
```

```
##  [1] "UA" "AA" "B6" "DL" "EV" "MQ" "US" "WN" "VX" "FL" "AS" "9E" "F9" "HA" "YV"
## [16] "OO"
```

```
## Not sure what carrier operates as "OO"... this could be an error.

max(flights$distance)
```

```
## [1] 4983
```

```
min(flights$distance) ## some possible issues with a 17 mile trip??
```

```
## [1] 17
```

```r
flight_time_wo <- flights %>%
  filter(!is.na(air_time), !is.na(air_time))
## Removing NA values...
## these could potential be canceled flights.

min(flight_time_wo$air_time)
```

```
## [1] 20
```

```r
## Seeing a min of 20 min which could be abnormal...
##  I would want to dig further into this specific flight.
```

(c) 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.

Example questions:

- Which airport, JFK or LGA, experience more delays?
- What was the worst day to fly out?
- Are there seasonal patterns

Do origin airports experience seasonal delays more than others? This would be interesting as one could make an informed decision to choose to fly from one airport over another to avoid delays. This kind of data would be useful also for busiensses and/or goverment bodies to better organize how flights are scheduled to try and optimize airport usage or to find which airport might need some remodeling/optimization to support increased traffic

Which airline experiences more delays and by how much?

(d) Exploring Data For each of the questions you proposed in Problem 1c, perform an exploratory data analysis designed to address the question. Produce visualizations (graphics or tables) to answer your question. * You need to explore the data from the point of view of the questions * Depending on the question, you would need to provide precise definition. For example, what does "more delays" mean. * At a minimum, you should produce two visualizations (graphics or tables) related to each question. Be sure to describe what the visuals show and how they speak to your question of interest.

```r
## Do origin airports experience seasonal delays more than others?
airport_data <- flights %>%
  group_by(origin, month) %>%
  summarize(mean = mean(dep_delay, na.rm = TRUE))
```

```
## `summarise()` regrouping output by 'origin' (override with `.groups` argument)
```

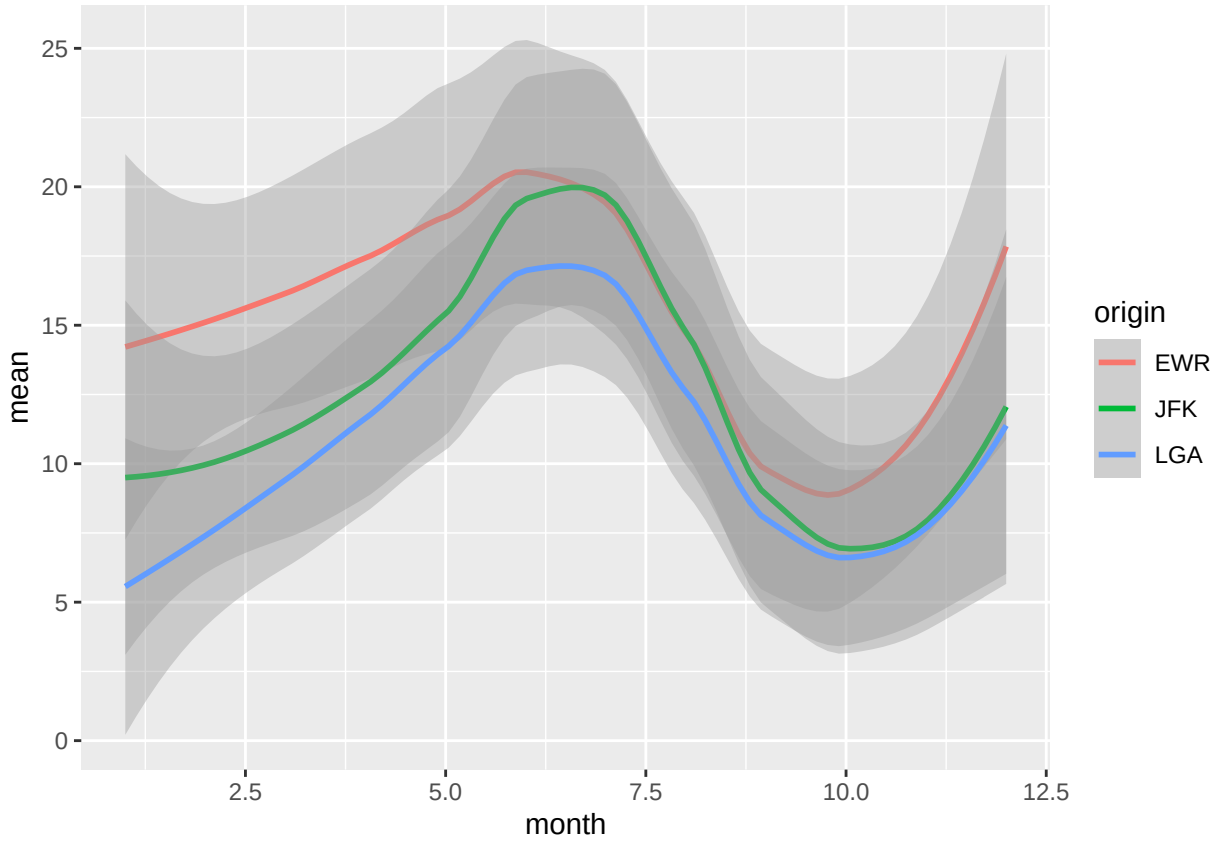
```r
airport_data
```

```
## # A tibble: 36 x 3
## # Groups:   origin [3]
##    origin month  mean
##    <chr>  <int> <dbl>
```

```
##  1 EWR        1 14.9
##  2 EWR        2 13.1
##  3 EWR        3 18.1
##  4 EWR        4 17.4
##  5 EWR        5 15.4
##  6 EWR        6 22.5
##  7 EWR        7 22.0
##  8 EWR        8 13.5
##  9 EWR        9  7.29
## 10 EWR       10  8.64
## # ... with 26 more rows
```

```
ggplot(data = airport_data) +
  geom_smooth(
    mapping = aes(x = month, y = mean, color = origin)
  )
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
## The graph shows a clear difference between EWR and the other two airports.
## It would appear that there is a seasonality for all airports. This we would
## expect as the summer months see more travelers and probably more families
## traveling which could account for the increase in delays. Furthermore, we
## see delays during typically winter months which could be attributed to snowy
## conditions causing delays. One could use this data to to pick an optimal time
```

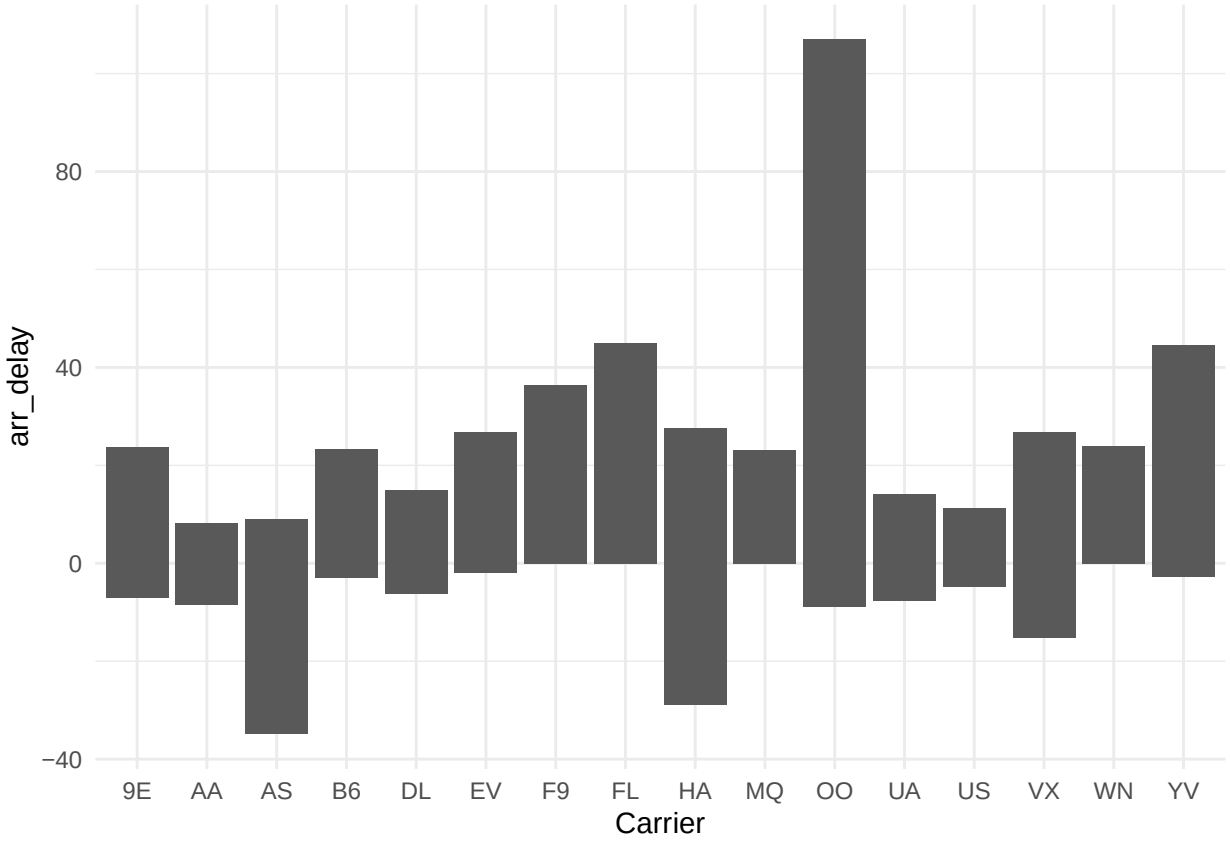
```r
## to fly and airlines could use this data to scrutinize how flights are
## scheduled and account for expected delays during specific months.
```

```r
# Which airline experiences more delays and by how much?

## Selecting the necessary columns and then removing the NA rows
airline_data <- flights %>%
  select(carrier, month, arr_delay) %>%
  na.omit()

## Creating a new data frame wich finds the average for each carrier each month.
new_df <- aggregate(airline_data[,3], list(airline_data$carrier,
                                           airline_data$month), mean) %>%
  rename(
    Carrier = Group.1,
    Month = Group.2
  )
head(new_df)
```

```
##   Carrier Month  arr_delay
## 1      9E     1 10.2074324
## 2      AA     1  0.9823789
## 3      AS     1  8.9677419
## 4      B6     1  4.7171992
## 5      DL     1 -4.4046512
## 6      EV     1 25.1601917
```

```r
## I tried graphing this to show how each carrier performs against eachother
ggplot(new_df, aes(x = Carrier, y = arr_delay)) +
 geom_col(stat="identity", position=position_dodge()) +
  theme_minimal()
```

```
## Warning: Ignoring unknown parameters: stat
```




```
## The graph is pretty simple and does not do a great job of providing clearer
## data with regards to seasonality. We can see how certain airlines perform
## better than others, but this is only for the year 2013 and thus is a small
## snapshot of the industry.
```

(e) Challenge Your Results After completing the exploratory analyses from Problem 1d, do you have any concerns about your findings? How well defined was your original question? Do you have concerns regarding your answer? Is additional analysis/different data needed? Comment on any ethical and/or privacy concerns you have with your analysis.

I don't believe there is sufficient privacy data of individuals for there to be any ethical concerns. If there were ticket information attached to each flight I could see potential privacy issues. I think the major problem would be assigning more context to some of the assertions made about the data being used to answer the questions. There could be many factors that can cause delays of a flight and to rely on just a few attributes to show which airline or airport does better might not provide the whole picture. I think the questions I provided are good starting points but could be refined further or expanded on to allow for more attributes or variables to provide more insight into what causes delays.