IMT 573: Problem Set 1 - Exploring Data

Stephen V. Tucker

October 14, 2022 PST

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 withouth errors you can do so with the `eval=FALSE` option as follows:

```
a + b # these objects 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 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.

BMI Calculation

formula BMI : 703 * weight(lbs) / [height(in)]^2

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

bmi_function(209,69)
```

```
## [1] 30.86053
```

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

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(tidyverse)
library('nycflights13')
```

```
# 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 <- tibble(flights)
# Look at the help file for information about the data
# ?flights
?flights
# summary(flights)
```

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

The data was

(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?

```
distinct_flights <- distinct(flights)
 # checking work
```

```
distinct_flights <- sum(unique(flights$flight)) # checking work
```

How many missing values are there in each variable?

The total number of missing values is 46595. To locate all the missing values in each variable I used the summary(fun) where it said all the missing values for quantatative data dep_time: 8255 NA's, dep_delay: 8255, arr_time: 8713, arr_delay:9430, air_time: 9430

```
missing_values <- apply(flights, 1, function(x) sum(is.na(x)))
sum(missing_values) # 46595
```

```
## [1] 46595
```

```
missing_values <- apply(flights, 1, function(x) (is.na(x)))
sum(missing_values)
```

```
## [1] 46595
```

```
summary(flights) # min max checks  & range
```

```
##       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
##
```

- Do you see any unreasonable values? *Hint: Check out min, max and range functions.*

(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.

This analysis is to explore trends analyzing two seperatedependent with the same independent variables

What is the top destinations?

Question 1. whats the most traveled to and origin within the dataset

Question 2. what are the delay times? In comparison

Within the data set the worst day to travel is on the 8th with 1,500 minute of total delay time The top destination is XNA, which is a airport in northwest Arkansas its total dep delay time was 5754, total arrival delay time was 7406, and its total delay time was 13160. This information becomes powerful for many reasons. one it answers

This question is important because it allows us to understand how a specific NYC airport handles there most traveled destination, along with aggregate data we can use to also analyze and use as a comparison tool.

Destination

The destinatio traveled the most.

```
top_destination <- flights %>%
  drop_na() %>%
  filter(dest == max(dest)) %>%
  select(year,day,month,arr_delay,dep_delay,dest)
```

```
top_destination_delay_time <- top_destination %>%
  group_by(total_dep_delay = sum(dep_delay)) %>%
  group_by(total_arr_delay = sum(arr_delay)) %>%
  group_by(total = total_dep_delay + total_arr_delay)


average_delay <- mean(top_destination_delay_time$arr_delay)
average_dep_delay <- mean(top_destination_delay_time$dep_delay)
```

origin

The origin departed from the most

```
top_origin <- flights %>%
  drop_na() %>%
```

```
  filter(origin == max(origin)) %>%
  select(year,day,month,arr_delay,dep_delay,origin)
```

Origin Delay Time

I wanted to get a break doen of delay time to see how they compared along with getting averages.

```
top_origin_delay_time <- top_origin %>%
  group_by(total_dep_delay = sum(dep_delay)) %>%
  group_by(total_arr_delay = sum(arr_delay)) %>%
  group_by(total = total_dep_delay + total_arr_delay)

# XNA
# total departure delay time 5754
#  total arrival delay time 7406
#  delay time in total 13160

average_arr_delay <- mean(top_destination_delay_time$arr_delay)
average_dep_delay <- mean(top_destination_delay_time$dep_delay)



## table 1
destination <- c("XNA")
dep_delay <- c(5754)
arr_delay <- c(7406)
total_delay<- c(13160)

xna_df <- data.frame(destination, dep_delay,arr_delay,total_delay, average_dep_delay, average_arr_delay

knitr::kable(xna_df, "pipe")
```

destination	dep_delay	arr_delay	total_delay	average_dep_delay	average_arr_delay
XNA	5754	7406	13160	5.800403	7.465726

table 2

```
## LGA
## total departure delay time 1040385
## total arrival delay time  584942
## total delay time 1625327

origin <- c("LGA")
dep_delay <- c(1040385)
arr_delay <- c(584942)
total_delay<- c(1625327)
origin_average_arr_delay <- mean(top_origin$arr_delay)
origin_average_dep_delay <- mean(top_origin$dep_delay)

lga_df <- data.frame(origin, dep_delay,arr_delay,total_delay, origin_average_dep_delay,
                     origin_average_arr_delay )
```

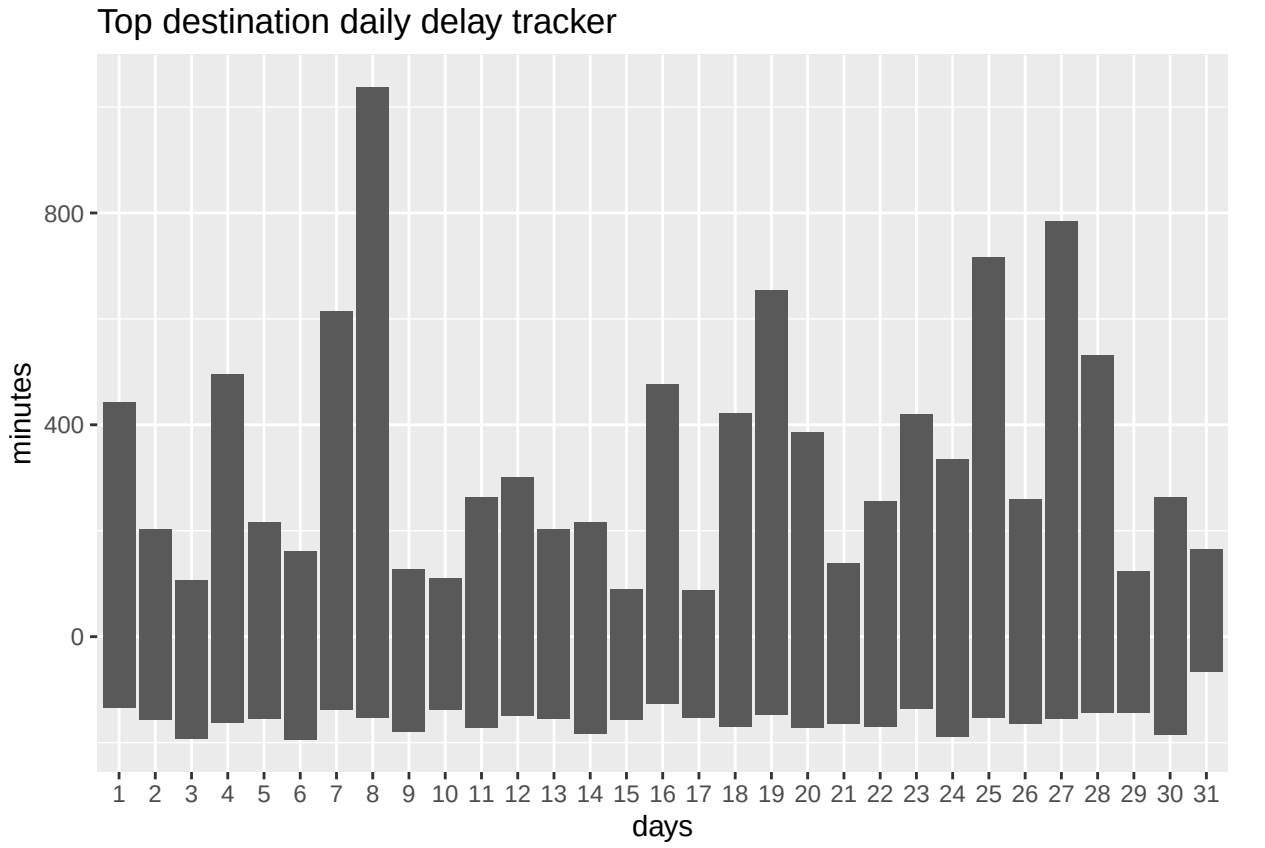
```
knitr::kable(lga_df, "pipe")
```

origin	dep_delay	arr_delay	total_delay	origin_average_dep_delay	origin_average_arr_delay
LGA	1040385	584942	1625327	10.28658	5.783488

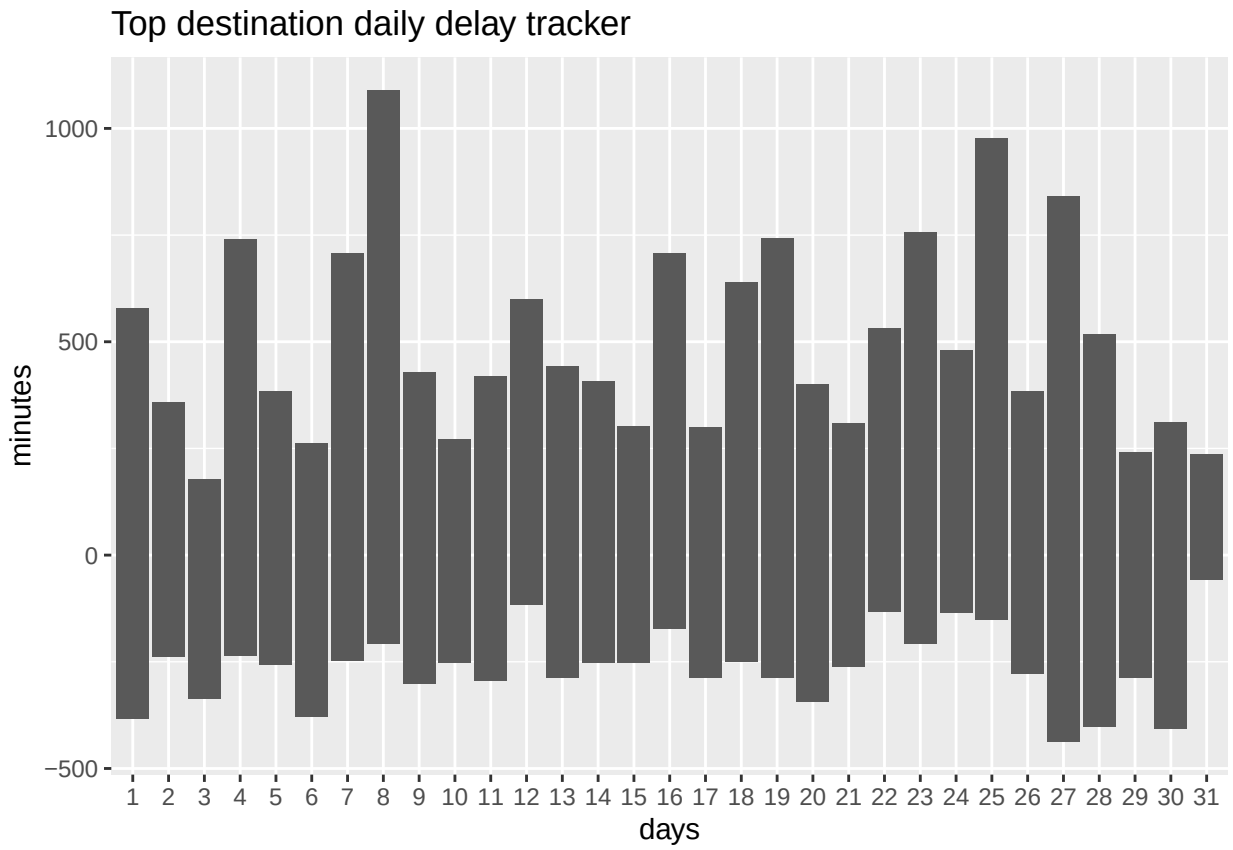
Destination Daily Delay Tracker

I wanted to explore at a granular level, doing so allows one to see if there were any similarities when compared. I utilized the data and used different dependent variables top destination and the top origin. The reason I gravitated towards isolated places of opposite locations is to see if i can identify any trends. Also choosing something straight forward that's easy to understand and to be analyzed further if wanted. I think its interesting that

```
ggplot(top_destination,
       aes(x = factor(day), y = dep_delay)) +
  geom_col() +
  labs(title = "Top destination daily delay tracker", x = "days", y = "minutes")
```



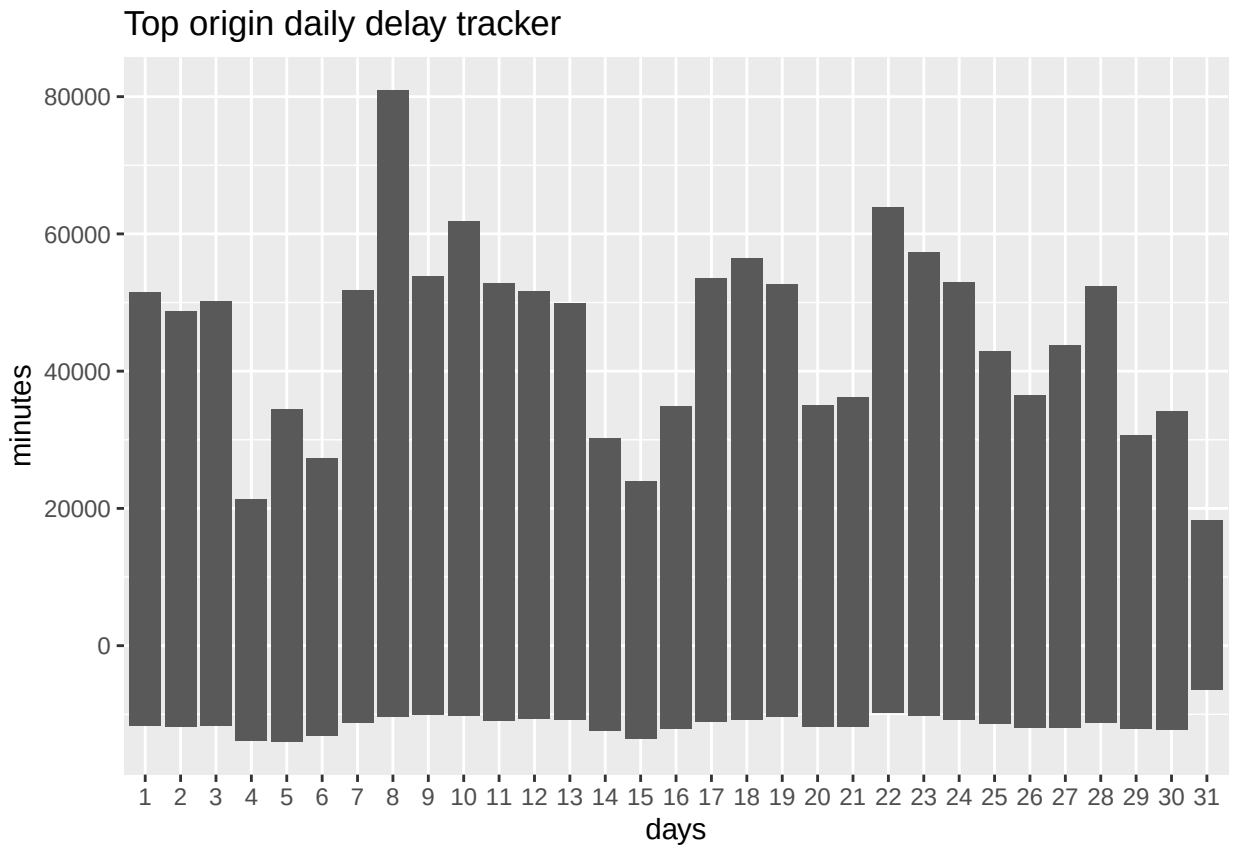

```
ggplot(top_destination,
       aes(x = factor(day), y = arr_delay)) +
  geom_col() +
  labs(title = "Top destination daily delay tracker", x = "days", y = "minutes")
```



Origin Daily Delay Tracker

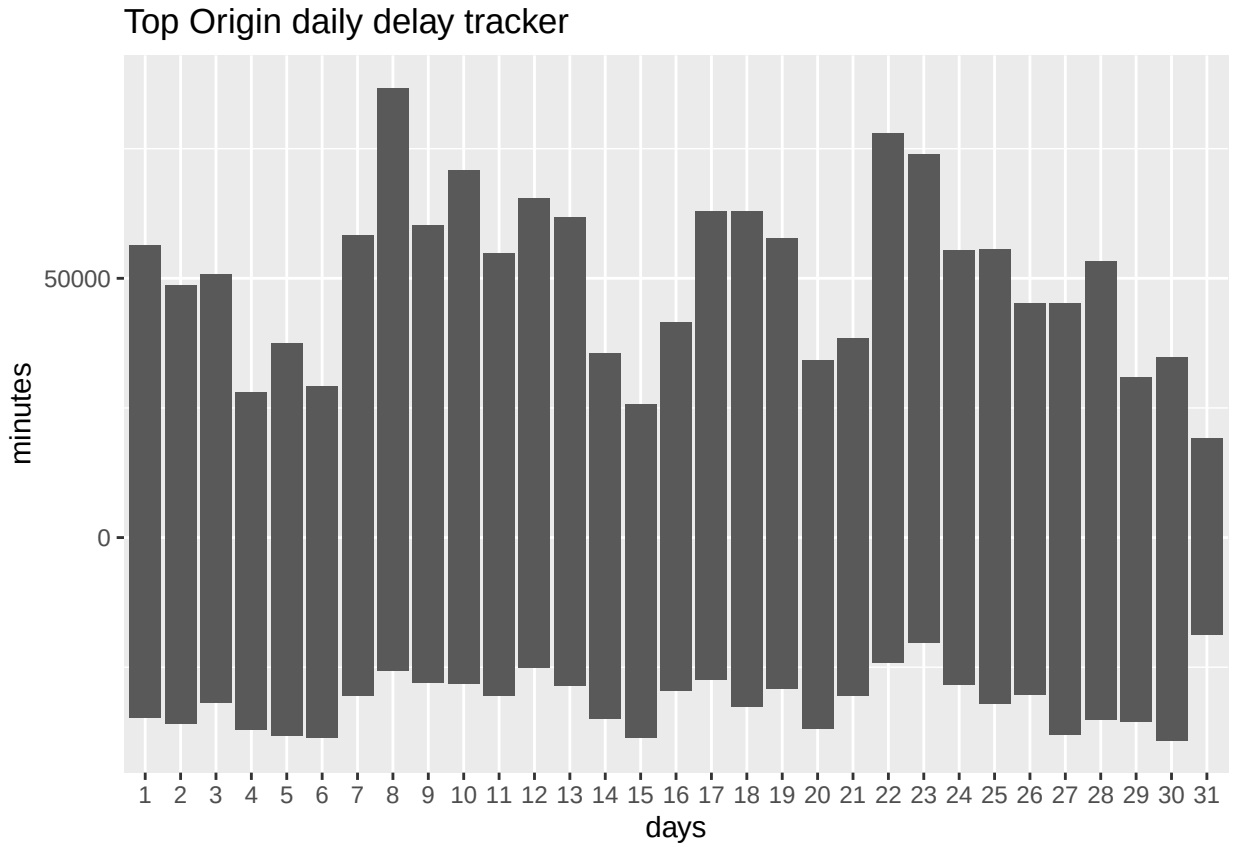
This is a bar plot showing the origin who has the most flights showing there departure delays.

```
ggplot(top_origin,
       aes(x = factor(day), y = dep_delay)) +
  geom_col() +
  labs(title = "Top origin daily delay tracker ", x = "days",
       y = "minutes")
```



Arrival

```
ggplot(top_origin,
       aes(x = factor(day), y = arr_delay)) +
  geom_col() +
  labs(title = "Top Origin daily delay tracker ", x = "days",
       y = "minutes")
```



(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 will need to provide a more 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.

(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.

Missing Data

Yes me only concern is how reliable is the data because i cannot visully see moths 10-12

When conducting range i notticed it had calculated for all 12 months but when i look at the data frame I can only see data for 9 months this makes me suspicious on whether I have reliable data or not. i was able to notice this when I hit used the tail FUN.

My concerns are only how do they compare with other variables also whether its consistent across years to think if we possibly leverage it at an operational level. In regards of questions and concerns with scope I would want to compare my answers to other parts of the data to see if this is a sign of a larger trend. In regards to privacy concerns i dont have any because its public data.