Exercise 3: Graphing

Jianhui Gao

July 10, 2025

```
library(tidyverse)
library(skimr)
library(visdat)
library(janitor)
```

NYC bus delays

The data is from the kaggle dataset "Bus Breakdown and Delays NYC - When and why a bus was delayed? Bus delays 2015 to 2017". (https://www.kaggle.com/anthobau/busbreakdownanddelays).

The Bus Breakdown and Delay system collects information from school bus vendors operating out in the field in real time. Bus staff that encounter delays during the route are instructed to radio the dispatcher at the bus vendor's central office. The bus vendor staff are then instructed to log into the Bus Breakdown and Delay system to record the event and notify OPT. OPT customer service agents use this system to inform parents who call with questions regarding bus service. The Bus Breakdown and Delay system is publicly accessible and contains real time updates. All information in the system is entered by school bus vendor staff.

You can find data for years 2015 to 2017.

```
bus.delay <- read_csv("Bus_Breakdown_and_Delays.csv")
head(bus.delay)
```

The following code creates new features

```
library(lubridate)

bus.delay$occur_date <- as.POSIXct(bus.delay$Occurred_On, format= "%m/%d/%Y %H:%M:%S %p")

# Day of the week
bus.delay$occur_weekday <- wday(bus.delay$occur_date, label = T)
bus.delay$occur_month <- month(bus.delay$occur_date, label = T)
bus.delay$occur_year <- year(bus.delay$occur_date)
#head(bus.delay)
```

1. Group by `occur_month`, and generate a bar plot that showing the number of delays by month.

The following code generates a new variable called "duration" in mins.

```
bus.delay$duration <- as.numeric(gsub("([0-9]{1,2}).*$", "\\1", bus.delay$How_Long_Delayed))
```

2. Plot the boxplot of duration.
3. Plot a histogram of durations and group by "Boro".
4. Plot the density of duration, group by "occur_weekday" and use facet to stratify on "Boro".
5. Print top five Boro by mean delay.