Solution 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)
```

```
## # A tibble: 6 x 21
##   School_Year Busbreakdown_ID Run_Type          Bus_No Route_Number Reason
##   <chr>                 <dbl> <chr>             <chr>  <chr>        <chr>
## 1 2015-2016           1224901 Pre-K/EI          811    1            Other
## 2 2015-2016           1225098 Pre-K/EI          9302   1            Heavy Traff~
## 3 2015-2016           1215800 Pre-K/EI          358    2            Heavy Traff~
## 4 2015-2016           1215511 Pre-K/EI          331    2            Other
## 5 2015-2016           1215828 Pre-K/EI          332    2            Other
## 6 2015-2016           1225671 Special Ed AM Run 12568  P640         Heavy Traff~
## # i 15 more variables: Schools_Serviced <chr>, Occurred_On <chr>,
## #   Created_On <chr>, Boro <chr>, Bus_Company_Name <chr>,
## #   How_Long_Delayed <chr>, Number_Of_Students_On_The_Bus <dbl>,
## #   Has_Contractor_Notified_Schools <chr>,
## #   Has_Contractor_Notified_Parents <chr>, Have_You_Alerted_OPT <chr>,
## #   Informed_On <chr>, Incident_Number <chr>, Last_Updated_On <chr>,
## #   Breakdown_or_Running_Late <chr>, School_Age_or_PreK <chr>
```

The following code creates new features

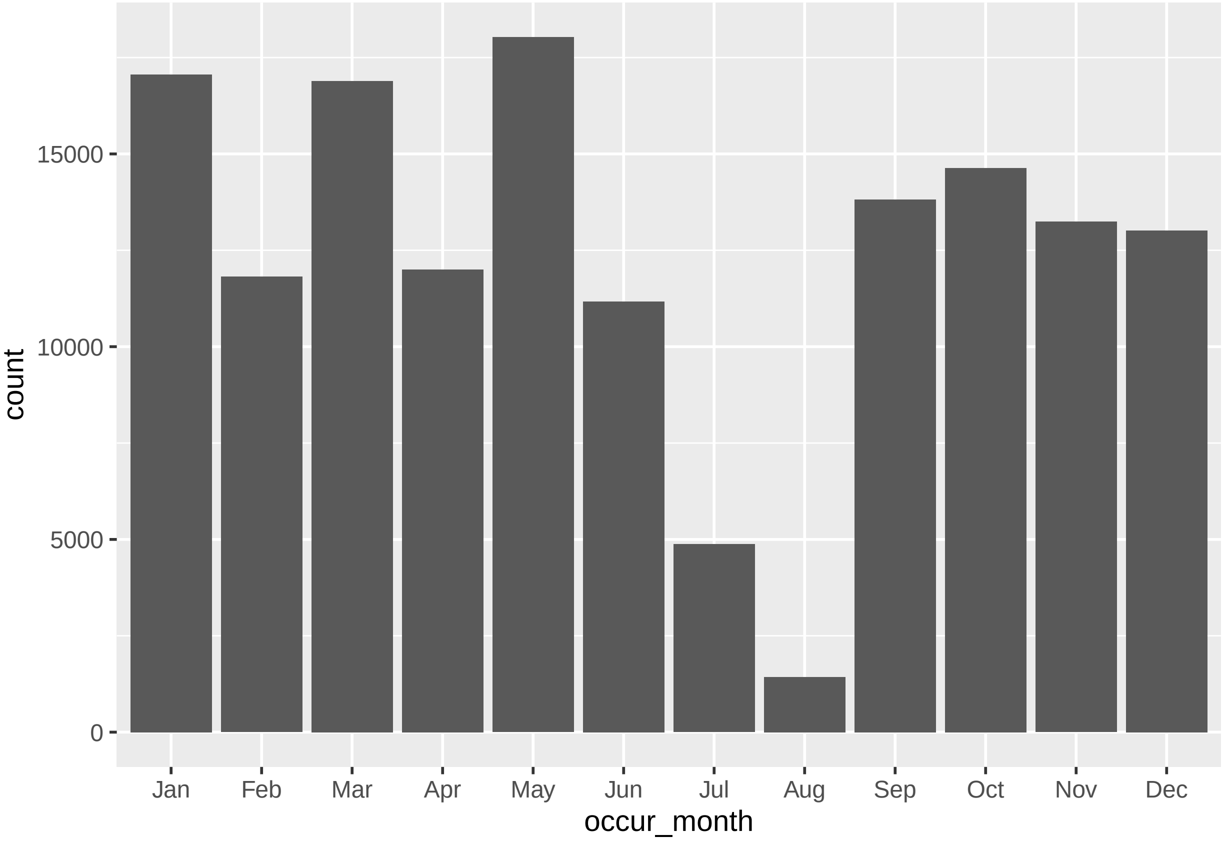
```
library(lubridate)

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

```r
# 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.

```r
#library(ggplot2)
# tidyverse
bus.delay %>%
  group_by(occur_month) %>%
  summarise(count = n()) %>%
  ggplot(aes(x = occur_month, y = count)) + #axis
  geom_bar(stat = 'identity') # stacked
```

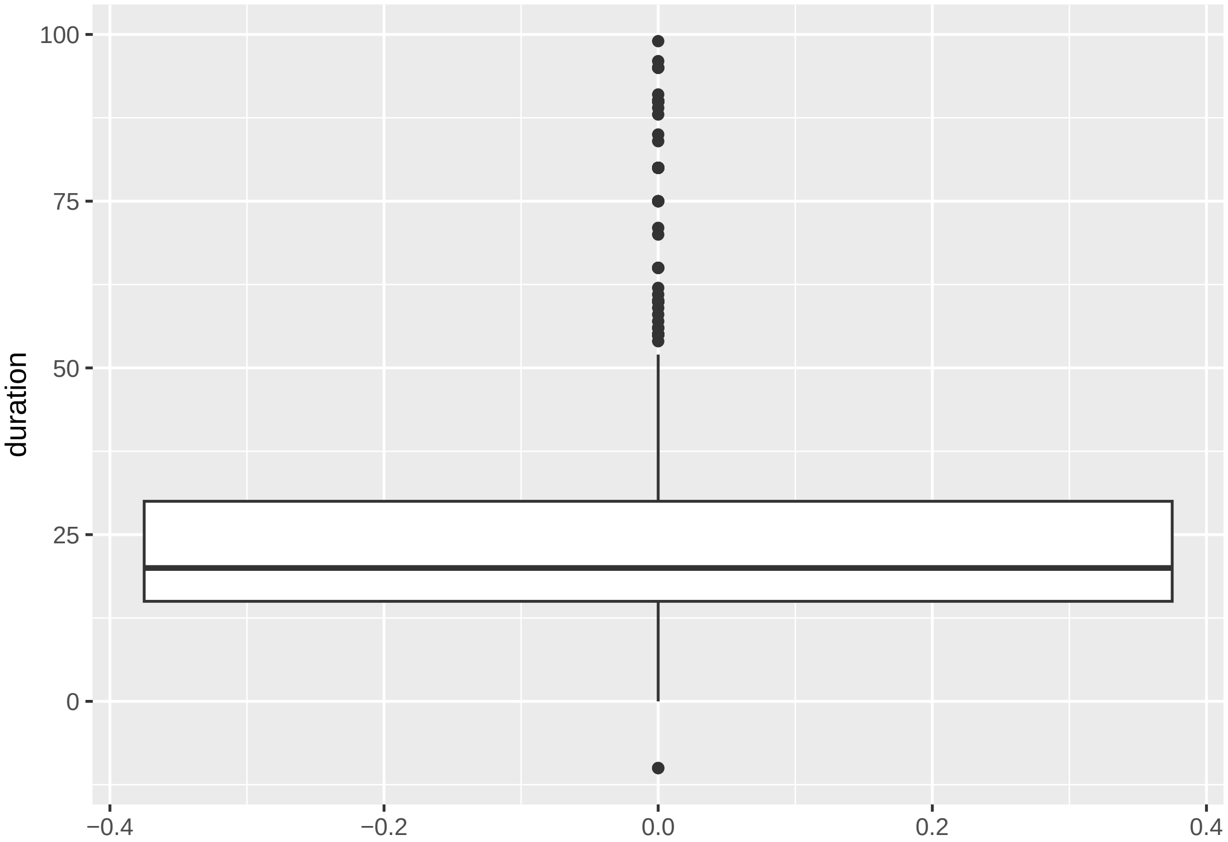


The following code generates a new variable called "duration" in mins.

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

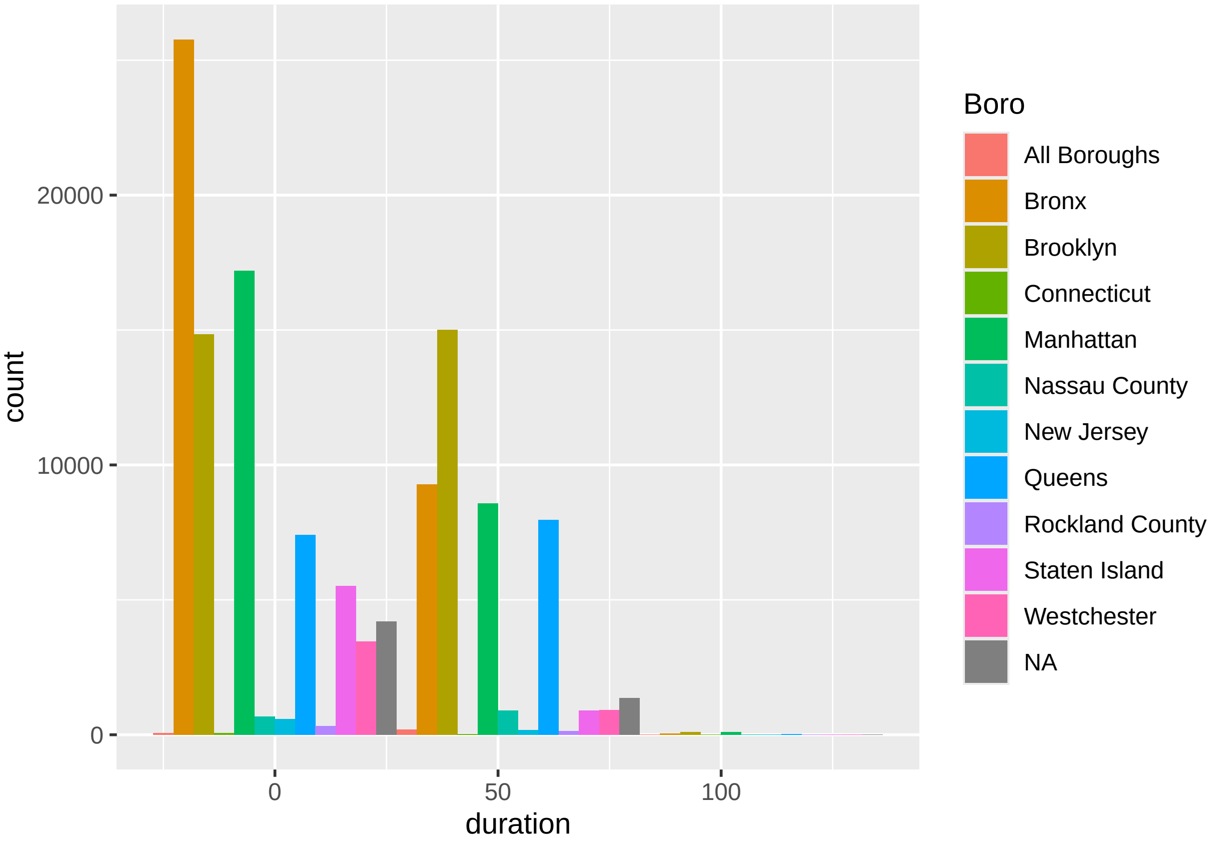
2. Plot the boxplot of duration.

```r
bus.delay %>%
  ggplot(aes(y = duration)) +
  geom_boxplot()
```



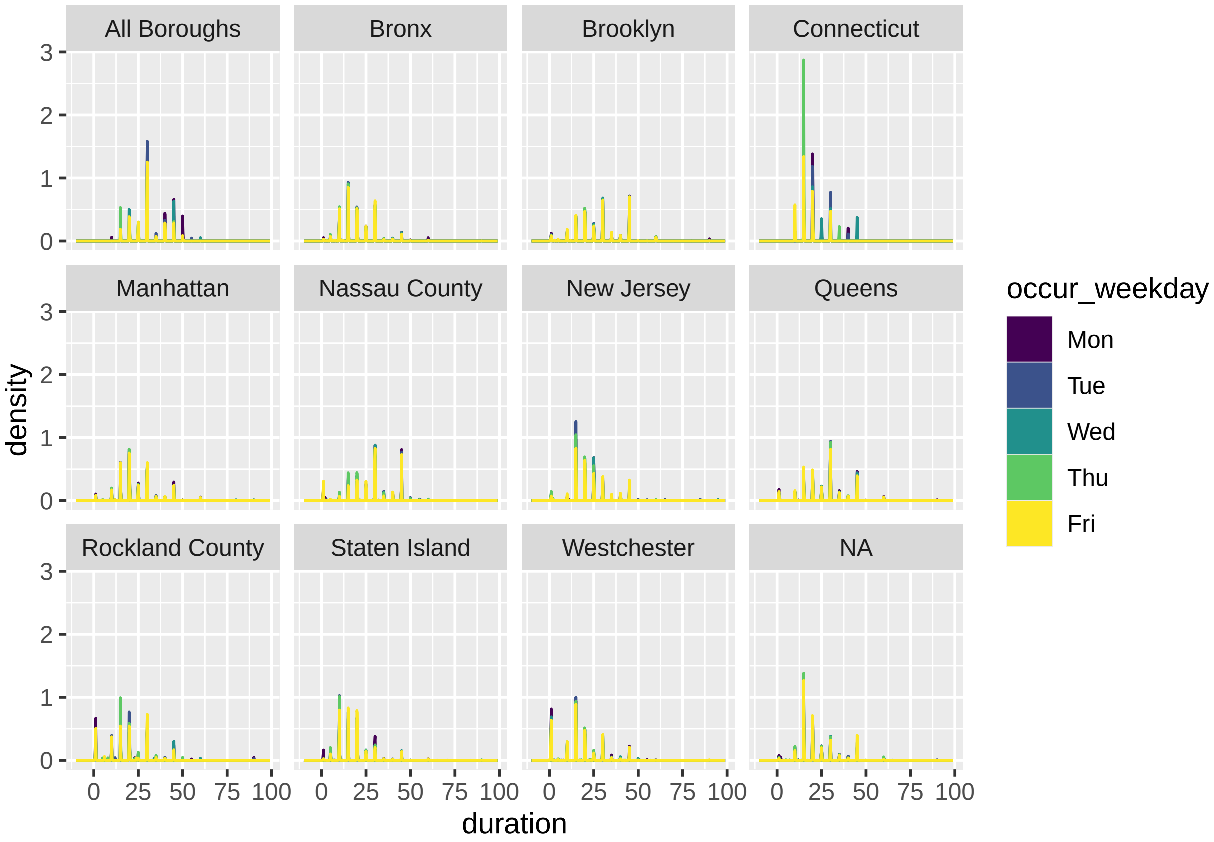
3. Plot a histogram of durations and group by “Boro”.

```r
bus.delay %>%
  ggplot() +
  geom_histogram(aes(x = duration,
                     fill = Boro),
                 position = 'dodge',
                 bins = 3)
```



4. Plot the density of duration, group by “occur_weekday” and use facet to stratify on” “Boro”.

```
bus.delay %>%
  ggplot() +
  geom_density(aes(x = duration,
                   color = occur_weekday,
                   fill = occur_weekday),
               bw = 0.08) +
  facet_wrap(~Boro, ncol = 4)
```



5. Print top five Boro by mean delay.

```
bus.delay %>%
  group_by(Boro) %>%
  summarise(mean_delay = mean(duration,
                              na.rm = TRUE)) %>%
  arrange(desc(mean_delay)) %>%
  head(5)
```

```
## # A tibble: 5 x 2
##   Boro           mean_delay
##   <chr>               <dbl>
## 1 All Boroughs         31.7
## 2 Brooklyn             27.8
## 3 Nassau County        27.8
## 4 Queens               26.5
## 5 Manhattan            24.1
```