

# Solution 3: Graphing

Siyue Yang

05/10/2022

```
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("../Data/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~
## # ... with 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")
```

```

# 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)

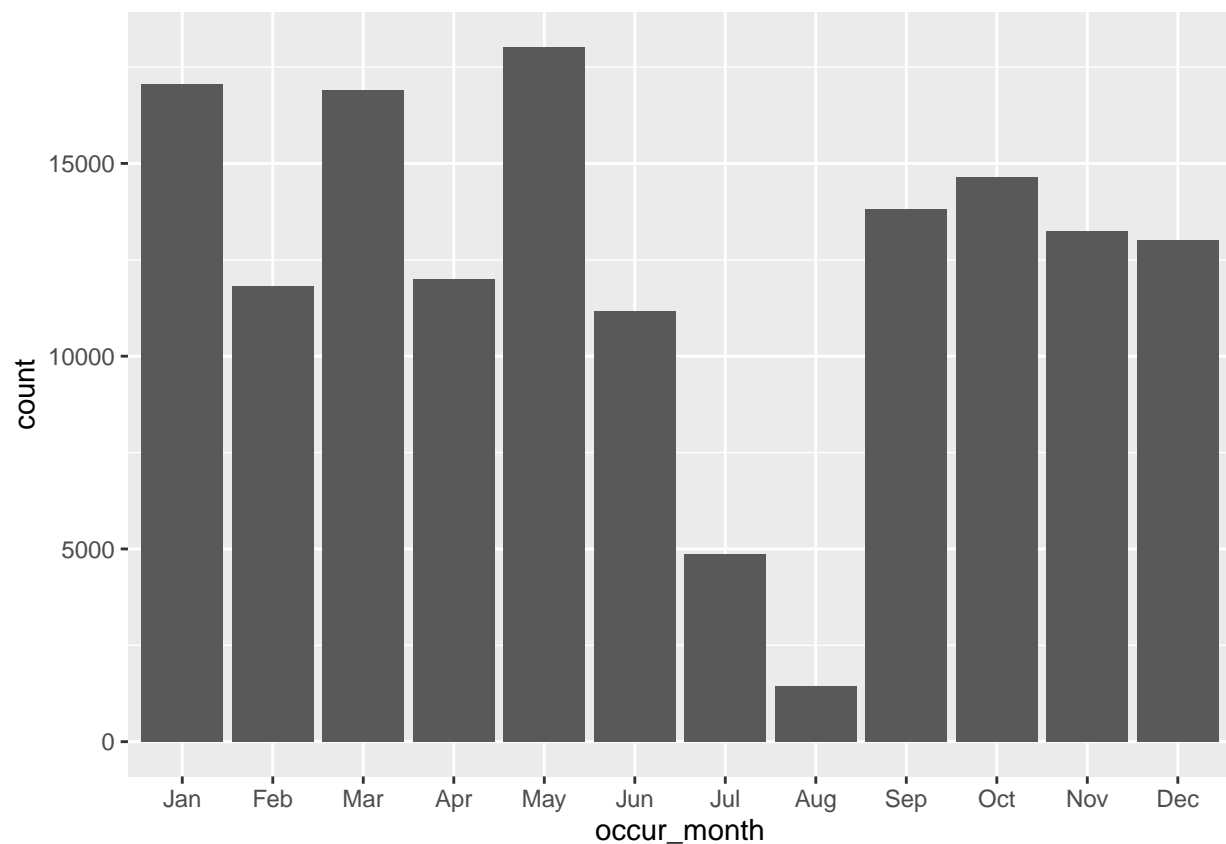
```

1. Group by occur\_month, and generate a bar plot that showing the number of delays by month.

```

#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.

```

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

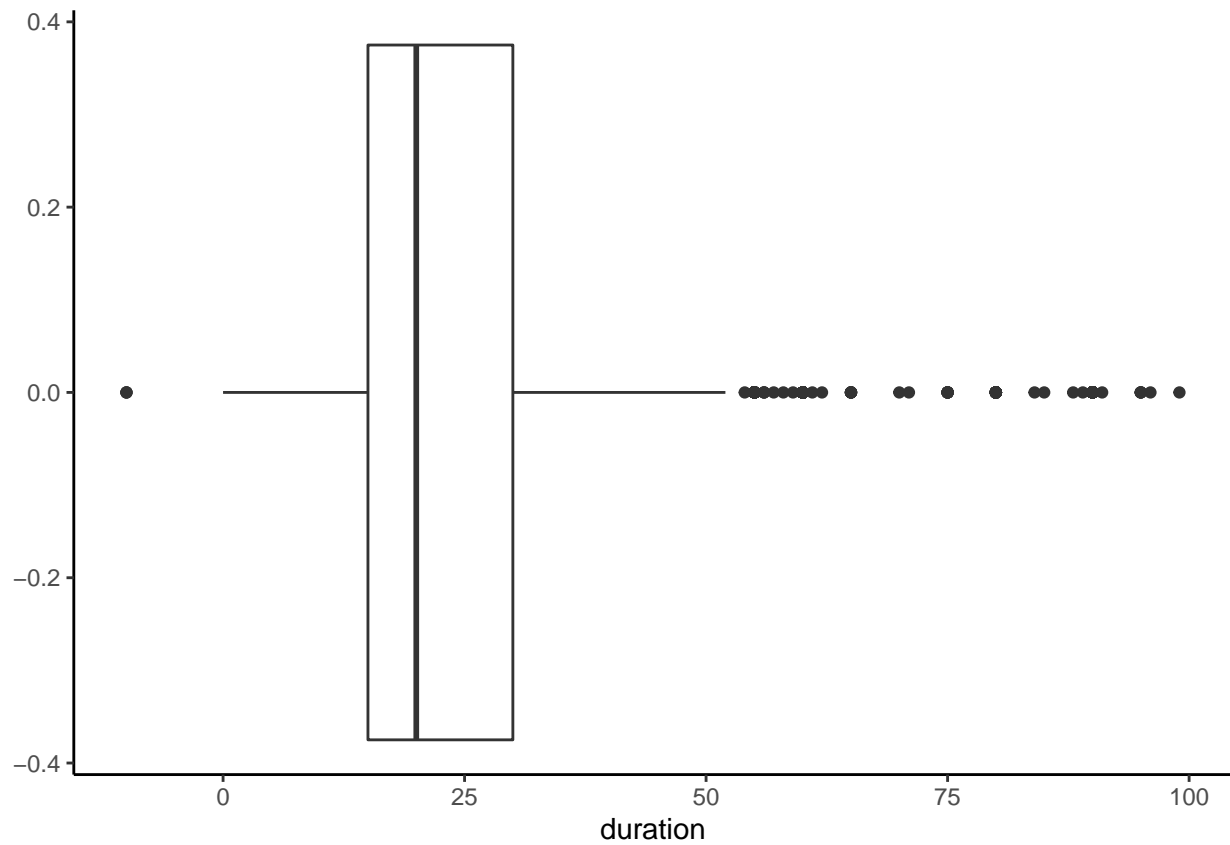
```

2. Plot the boxplot of duration.

```

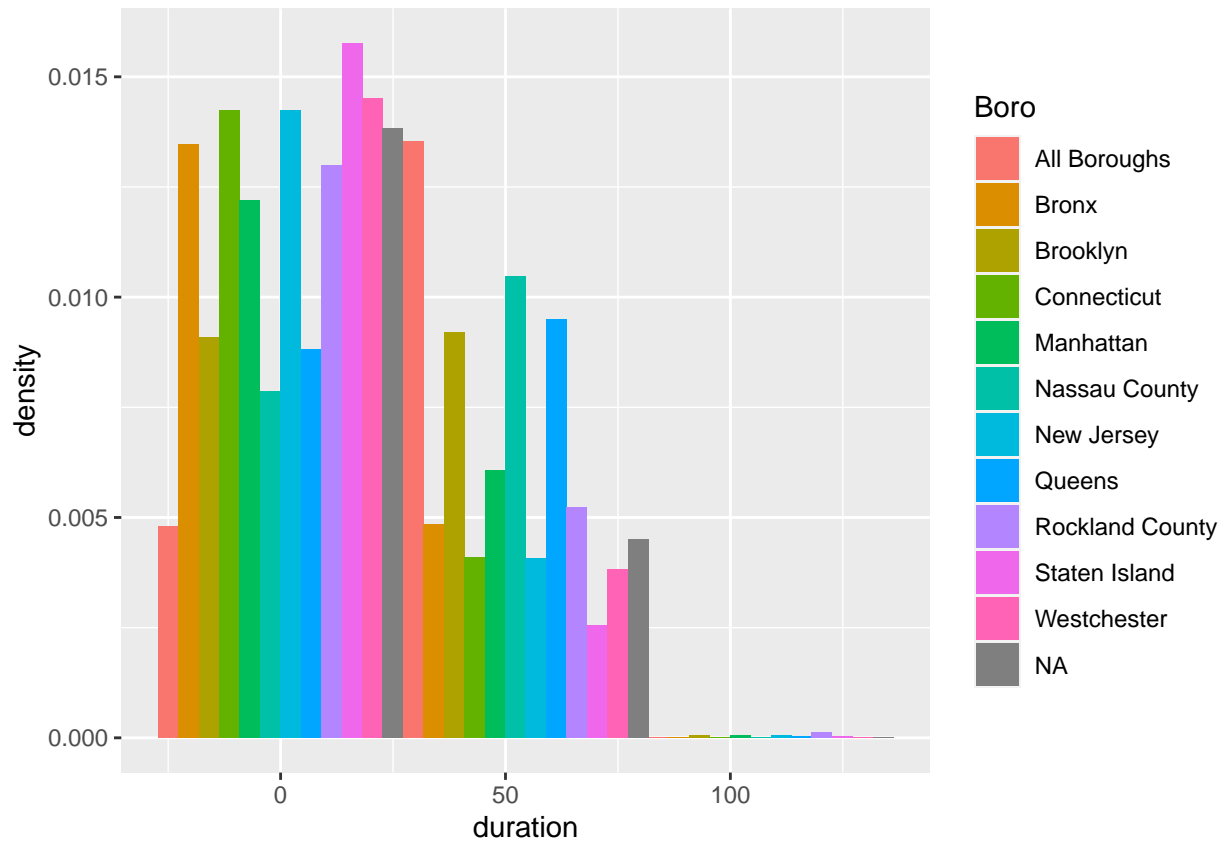
ggplot(data = bus.delay) +
  geom_boxplot(aes(x = duration)) + theme_classic()

```



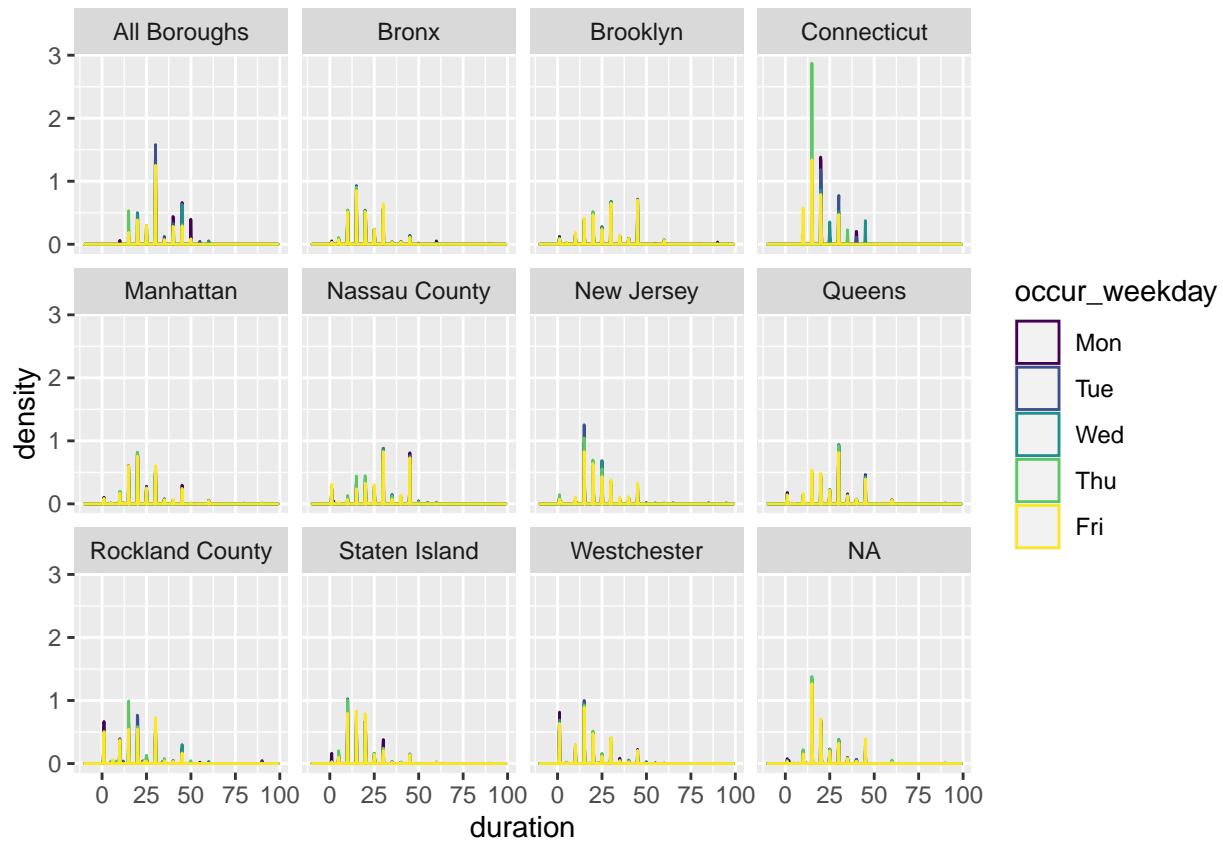
3. Plot a histogram of durations and group by “Boro”.

```
ggplot(data = bus.delay) +  
  geom_histogram(aes(x = duration, y = ..density.., fill = Boro),  
                 position = 'dodge',  
                 bins = 3)
```



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

```
ggplot(data = bus.delay) +
  geom_density(aes(x = duration, color = occur_weekday), bw = .08) +
  facet_wrap(~Boro, ncol = 4)
```



5. Print top five Boro by mean delay.

```
bus.delay %>%
  group_by(Boro) %>%
  summarise(mean_delay = mean(duration), n_obs = n())
```

```
## # A tibble: 12 x 3
##   Boro          mean_delay n_obs
##   <chr>          <dbl> <int>
## 1 All Boroughs      NA    275
## 2 Bronx             NA  40995
## 3 Brooklyn          NA  35632
## 4 Connecticut      NA    102
## 5 Manhattan         NA  28675
## 6 Nassau County     NA   1970
## 7 New Jersey        NA    887
## 8 Queens            NA  21169
## 9 Rockland County   NA    491
## 10 Staten Island    NA   6923
## 11 Westchester      NA   4535
## 12 <NA>             NA   6318
```

6. Look by week to see if there's any seasonality. i.e. create a plot that group by "Boro" and use facet to stratify by "Boro".

```
bus.delay %>%
  filter(duration > 0) %>%
  mutate(week = week(occur_date)) %>%
  group_by(week, Boro) %>%
```

```

summarise(mean_delay = mean(duration)) %>%
  ggplot(aes(week, mean_delay, color = Boro)) +
  geom_point() +
  geom_smooth() +
  facet_wrap(~Boro, ncol = 4)

```

