607 Project 2 Bus Routes

Lu Beyer

Overview

This is an extremely large dataset with over 650,000 entries logging bus breakdowns and delays for New York City schools. I want to organize the data in ways that can help me determine the route difficulty of these NYC school buses. Upon first looking at this data, I am interested in what years have the most breakdowns and delays, which routes have the most breakdowns and delays, and the biggest cause of breakdown or delays on these routes. I'm considering exploring which boros are most impacted as well, and what the length of delays tend to be, but I will have a clearer idea of how I want to analyze the data once I begin to clean the dataset.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
data <- read.csv("Bus_Breakdown_and_Delays_20240225.csv")
```

```
a <- data %>%
  select(School_Year, Route_Number, Busbreakdown_ID, Reason, Boro, Bus_Company_Name, How_Long_Delayed, I
```

```
b <- a %>%
  filter(Route_Number != "") %>%
  # filter(Reason == "Heavy Traffic") %>%
  filter(School_Year == "2017-2018") %>%
  group_by(Reason) %>%
  summarise(count = n())
```

```
c <- a %>%
  select(School_Year, Reason)
```

```
table(c)
```

```
##            Reason
## School_Year       Accident Delayed by School Flat Tire Heavy Traffic
##   2015-2016     0      626               617      1709         35343
##   2016-2017     0      725               824      2817         51255
##   2017-2018     2      782               591      2770         55898
##   2018-2019     0      916               496      2592         72288
##   2019-2020     0      643               335      1534         43491
##   2020-2021     0      289               105       538         12744
##   2021-2022     0      674               704      1041         39531
##   2022-2023     0      599               483      1317         61712
##   2023-2024     0      349               187       806         31435
##            Reason
## School_Year Late return from Field Trip Mechanical Problem Other Problem Run
##   2015-2016                        1742                5516 11127        1030
##   2016-2017                        1712                8231 10190        1154
##   2017-2018                        1707               10193 11429         708
##   2018-2019                        1739               10523 13585        2128
##   2019-2020                         806                6467  8649        1604
##   2020-2021                           9                2493  3179          47
##   2021-2022                         278                5570 14792        5391
##   2022-2023                         981                6909 19107        1354
##   2023-2024                         419                4493  8992        6315
##            Reason
## School_Year Weather Conditions Won‘t Start
##   2015-2016               2529        2945
##   2016-2017               1995        4238
##   2017-2018               1612        3743
##   2018-2019               1366        3426
##   2019-2020                764        1705
##   2020-2021                421         613
##   2021-2022               1197        1001
##   2022-2023               1140        1049
##   2023-2024               1927         874
```

Because I am working with such a large dataset, I wanted to focus on a single year and look at that data more closely. I decided to organize my data so I would be able to determine which year had the most delays and breakdowns. Grouping by year and getting the count of entries per year allowed me to see which year had the most number of break downs and delays documented.

```
summary <- a %>%
  group_by(School_Year) %>%
  summarise(count = n())
```

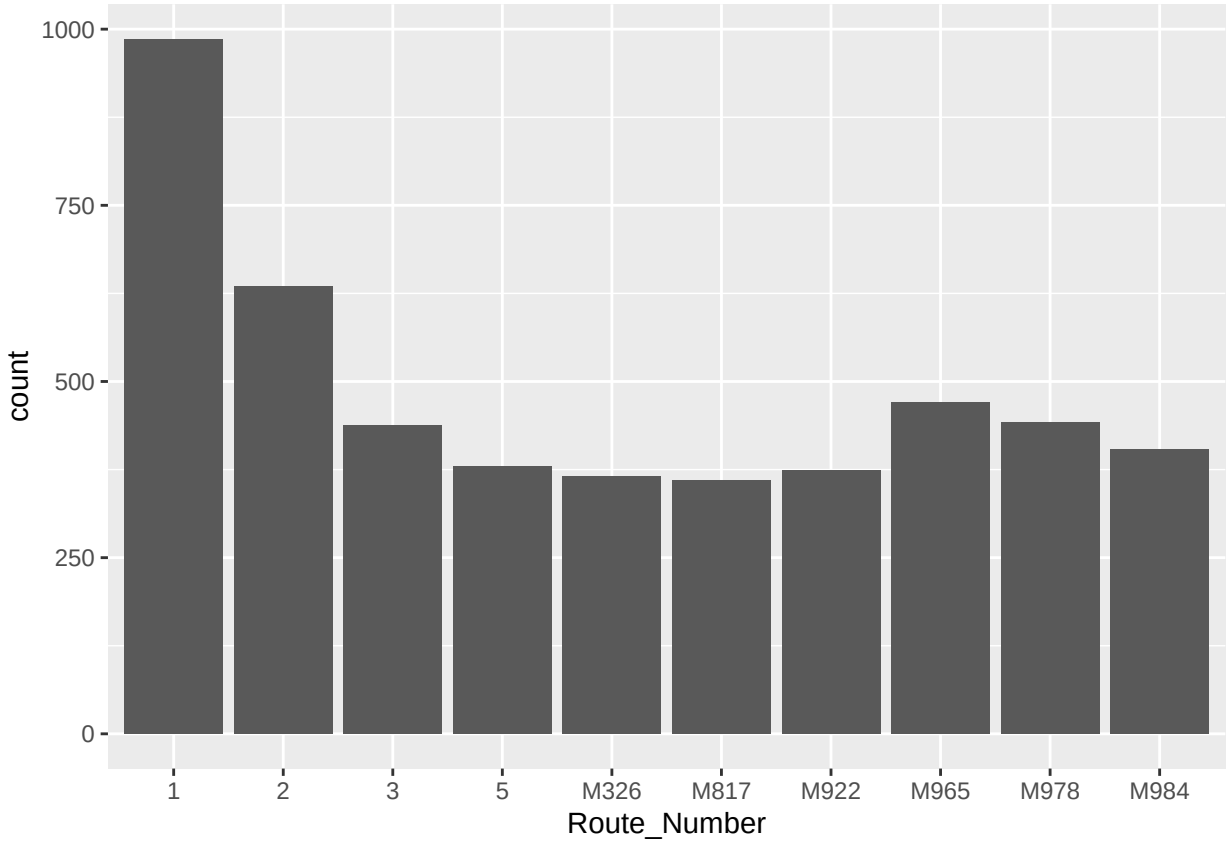
Once I determined 2018 had the most delays and breakdowns in my dataset, I wanted to see which route number had the most entries. Filtering by school year and grouping by route number, I was able to get the counts for each route. Because there were over 6000 bus routes, I wanted to focus on the 10 routes with the highest number of delays and breakdowns so I would be able to create clearer visualizations for that data.

```
Year_2018_2019 <- a %>%
  filter(School_Year == "2018-2019") %>%
  group_by(Route_Number) %>%
  summarise(count = n()) %>%
  arrange(.,count) %>%
  tail(10)
```

```
d <- a %>%
  filter(Route_Number != "") %>%
  group_by(Reason, School_Year) %>%
  summarise(count = n())
```

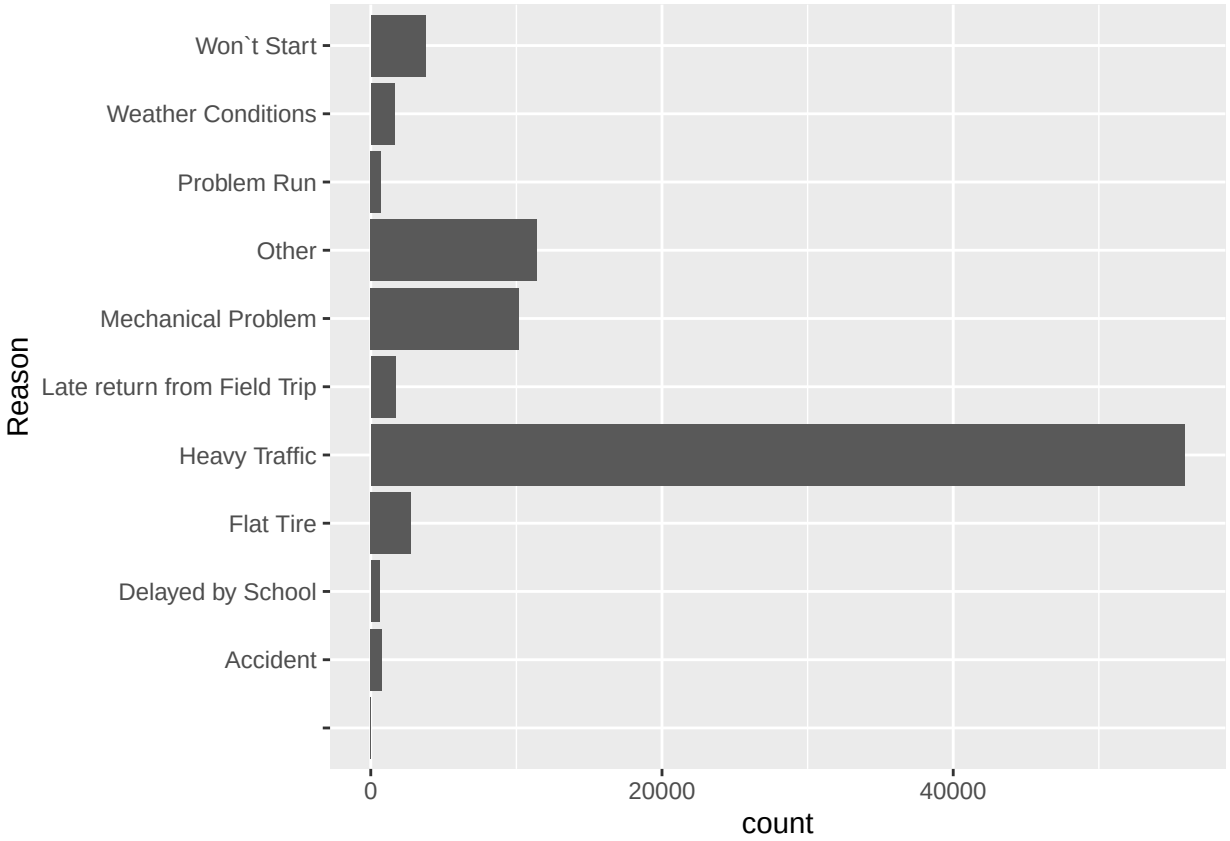
```
## `summarise()` has grouped output by 'Reason'. You can override using the
## `.groups` argument.
```

```
ggplot(data = Year_2018_2019, aes(x = Route_Number, y = count)) +
  geom_bar(position = "dodge", stat = "identity")
```

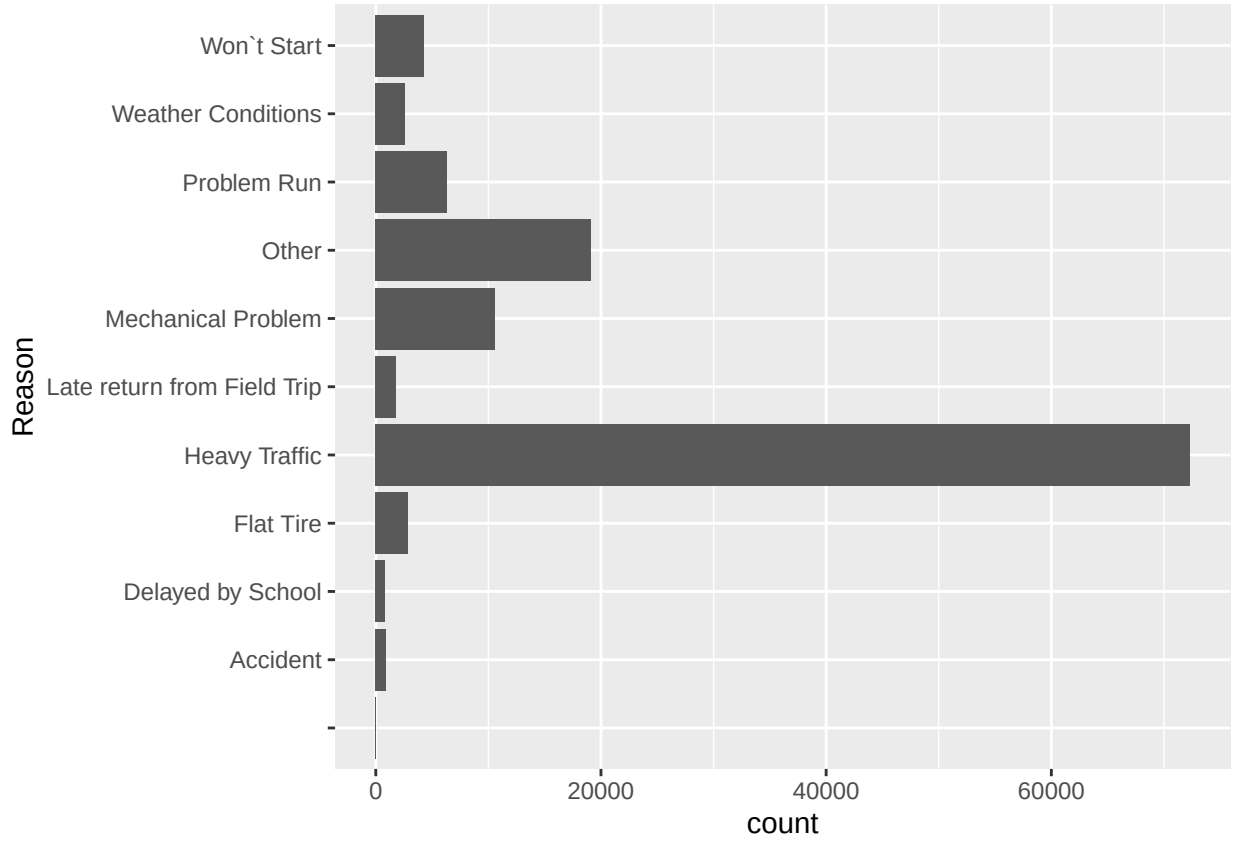


I wanted to look at what the biggest cause for delays in 2018 and compare that to the cause of delays over all of the years documented in this dataset.

```
ggplot(data = b, aes(x = count, y = Reason)) +
  geom_bar(position = "dodge", stat = "identity")
```

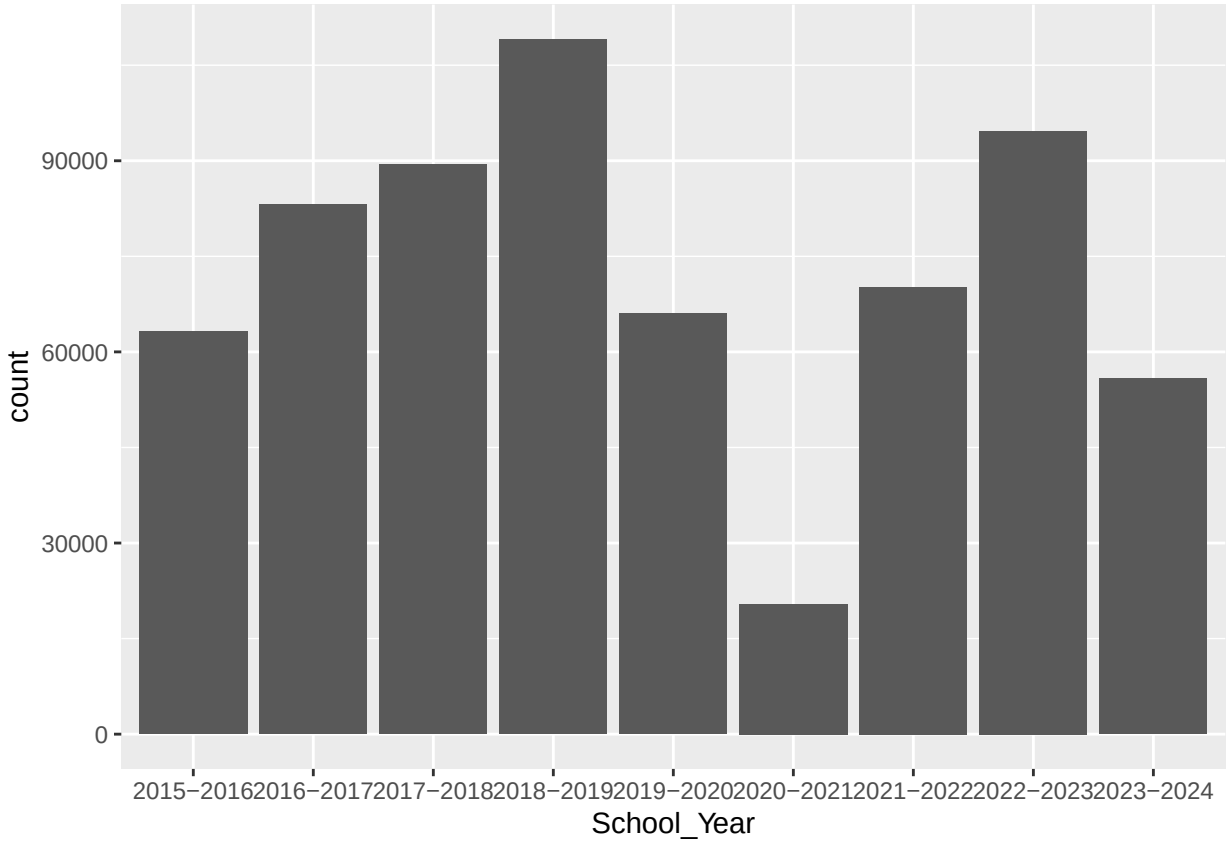



```
ggplot(data = d, aes(x = count, y = Reason)) +
  geom_bar(position = "dodge", stat = "identity")
```



Finally, I decided to look at which years had the most documented incidents.

```
ggplot(data = summary, aes(x = School_Year, y = count)) +
  geom_bar(position = "dodge", stat = "identity")
```



Conclusion

There is more that can be done with this data set, but with the variables I chose to analyze, I was able to determine which year had the highest number of incidents (2018), which 10 bus routes within that year were most impacted (1, 2, 3, 5, M326, M817, M922, M965, M978, and M984), and what was the highest cause for delay reported (heavy traffic). If I wanted to delve deeper into the data I chose to analyze, I could look at the specific route each route number takes to see what parts of the city they run through. If heavy traffic is the main cause of bus delays as this data shows, it's likely these routes run through areas of the city that are more densely populated or have higher rates of commuters in general. It could be interesting to see this data to be able confirm this, or see what other causes impacts these traffic on these specific routes. Additionally, looking at the reported incidents by year, I see that 2020-2021 had by far the lowest reported incidents. Knowing that this was the year of the start of the pandemic and many schools shifted to remote learning at that time, this dip in incidents makes a lot of sense, and as the 2024 school year has not yet ended, the lower count of reported incidents will likely change.

I would be interested to look at population data throughout 2015 - 2019 to see if there was an increase in NYC population to cause the steady increase of incidents over that time (given that heavy traffic is the main cause of incidents). It could also be helpful to look at commuter data to see if there is a trend of more drivers vs public transition commuters. The dip in incidents between 2019 and 2022 makes sense to me because of the pandemic, but I would be curious to know the reasoning for the trends in the data otherwise.