Motor vehicle colisions in NYC

Group 1 (605 - S1)

September 26, 2025

Contents

1 Introduction

We first load the model and create new columns:

```
setwd("C:/Users/louis/Documents/Academics/Fall 2025/CMOR 605/STAT 605 Project") #SET TO YOUR DIRECTORY
nyc <- read.csv(file = "data/Motor_Vehicle_Collisions_-_Crashes.csv")
nyc$CRASH.DATE <- as.Date(nyc$CRASH.DATE, format = "%m/%d/%Y")
nyc$BOROUGH <- factor(nyc$BOROUGH)
nyc$YEAR_MONTH <- format(nyc$CRASH.DATE, "%Y-%m")
nyc$YEAR <- format(nyc$CRASH.DATE, "%Y")
```

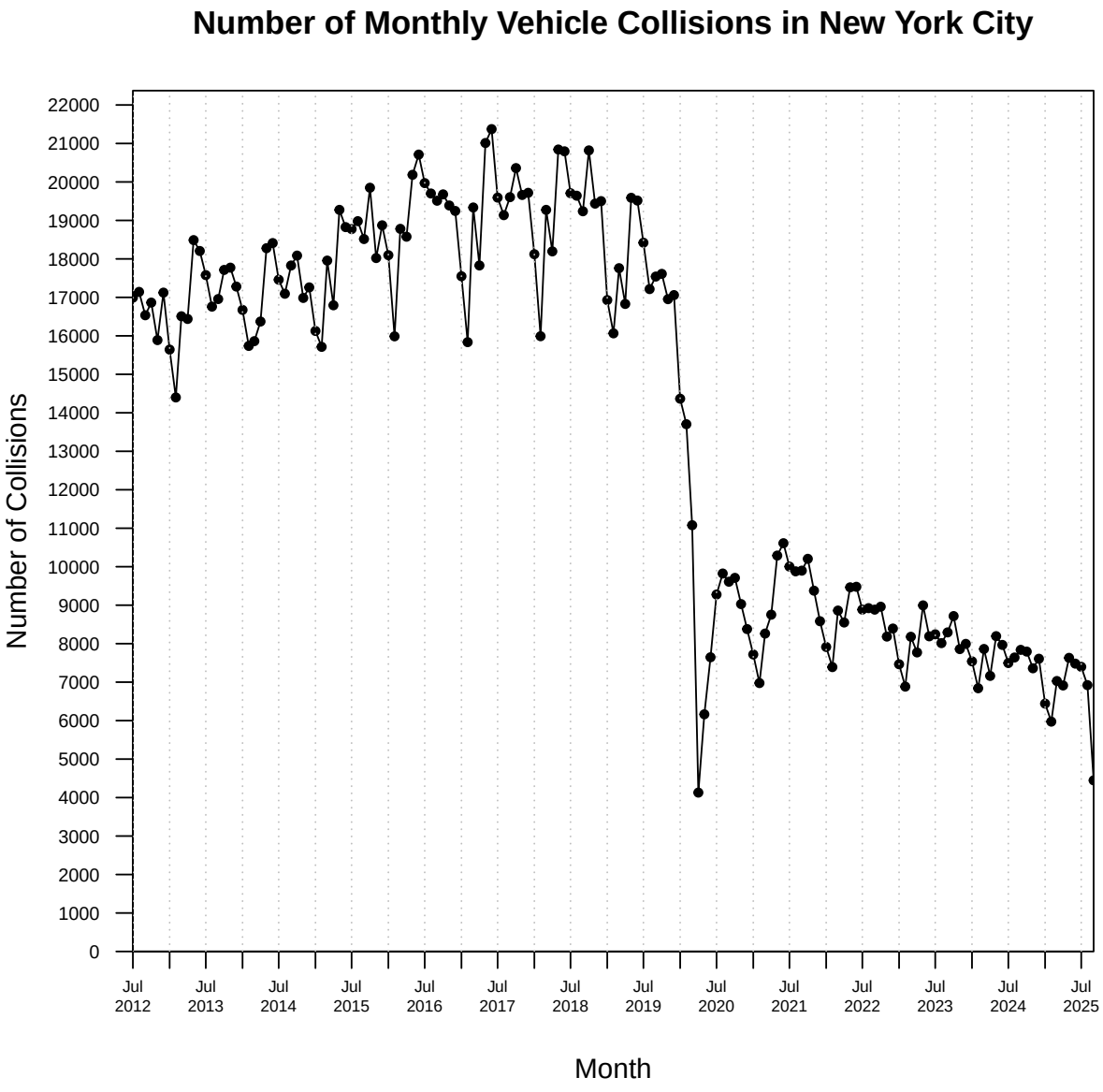
2 Visualizations

2.1 Modelling Vehicle Collisions over Time

By grouping the collisions into each month in which they occured, we can plot a line graph showing the number of collisions over the past 13 years.

```
monthly_counts <- table(nyc$YEAR_MONTH)
months <- names(monthly_counts)
colspermonth <- as.numeric(monthly_counts)
plot_months <- as.Date(paste0(months, "-01"))
plot(plot_months, colspermonth,
     main="Number of Monthly Vehicle Collisions in New York City",
     type="o",
     pch=20,
     xlab="Month",
     ylab="Number of Collisions",
     xaxt="n",
     yaxt="n",
     xaxs="i",
     yaxs="i",
     ylim=c(0,max(colspermonth)+1000))
```

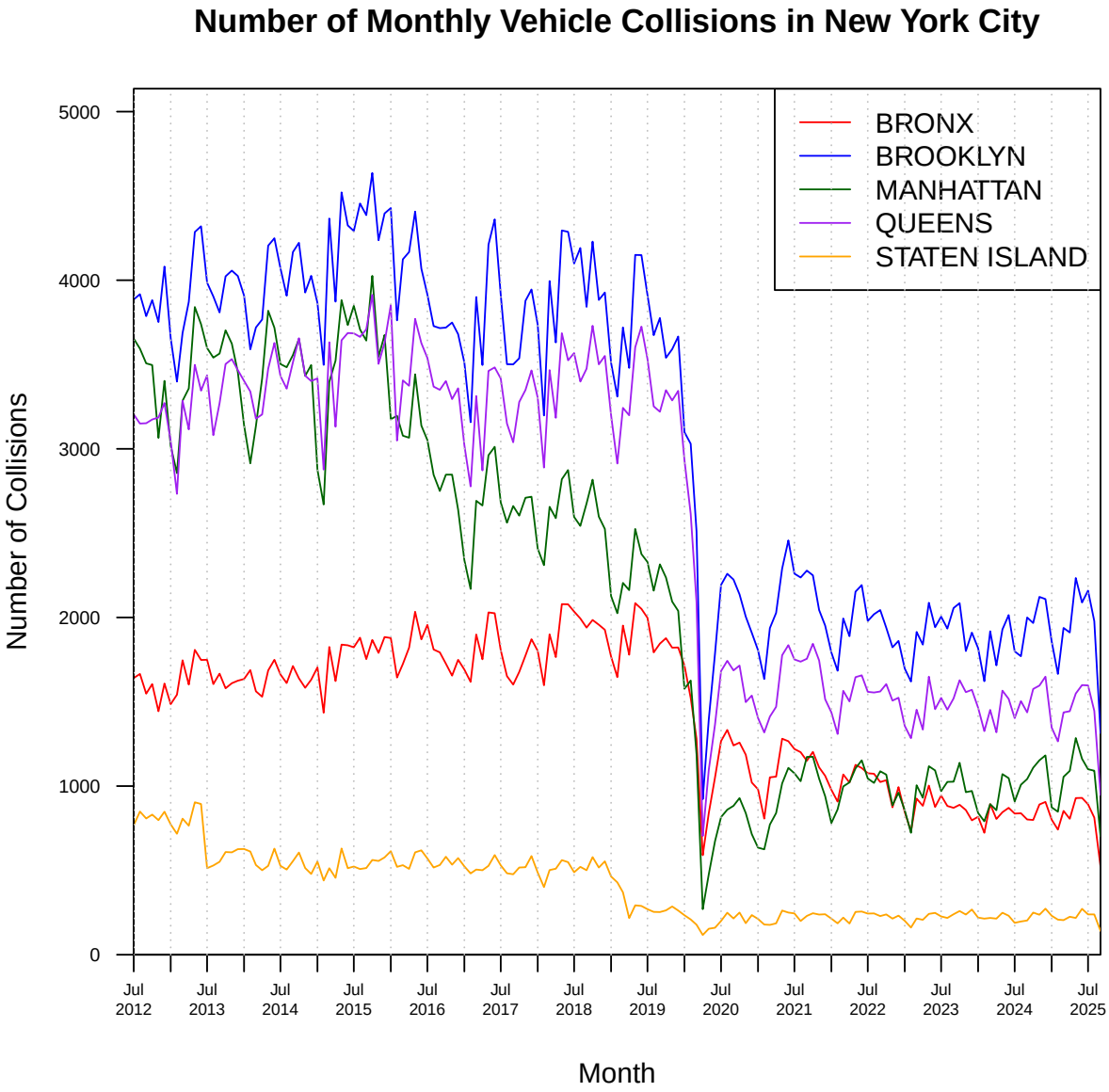
```
xticks <- seq(min(plot_months), max(plot_months)+365, by = "6 months")
axis.Date(1, at = xticks,
          format = "%b\n%Y", cex.axis = 0.6)
axis(2, at = seq(0, max(colspermonth)+1000, by = 1000), cex.axis = 0.7, las=1)
abline(v = xticks, col = "gray", lty = "dotted")
```



A notable trend is the sudden drop in collisions at the beginning of 2020. This is clearly due to the Covid-19 pandemic, which limited the time people spent outside their homes. What is more notable is how the number of collisions seems to even out to a much lower level than pre-pandemic. More research is needed into why we see this pattern.

Next, we explore the number of collisions over time within each borough. Note that a large quantity of entries from the dataset do not have a location attached, so those results are not considered for this analysis.

```r
borough <- levels(nyc$BOROUGH)
borough <- borough[borough != ""]
colors = c("red", "blue", "darkgreen", "purple", "orange")
plot_months <- list()
colspermonth <- list()
for (b in borough){
  monthly_counts <- table(nyc$YEAR_MONTH[nyc$BOROUGH==b])
  months <- names(monthly_counts)
  colspermonth[[b]] <- as.numeric(monthly_counts)
  plot_months[[b]] <- as.Date(paste0(months, "-01"))
}

max_col <- max(sapply(colspermonth, max))
min_dat <- as.Date(min(sapply(plot_months,min)))
max_dat <- as.Date(max(sapply(plot_months,max)))

plot(plot_months[[1]], colspermonth[[1]],
     main="Number of Monthly Vehicle Collisions in New York City",
     type="l",
     pch=20,
     col= colors[1],
     xlab="Month",
     ylab="Number of Collisions",
     xaxt="n",
     yaxt="n",
     xaxs="i",
     yaxs="i",
     ylim=c(0,max_col+500))

for (i in 2:(length(borough))){
  lines(plot_months[[i]], colspermonth[[i]], type="l", pch=20, col=colors[i])
}

legend("topright",
       legend = borough,
       col = colors,
       lty = 1)


xticks <- seq(min_dat, max_dat+365, by = "6 months")
axis.Date(1, at = xticks,
          format = "%b\n%Y", cex.axis = 0.6)
axis(2, at = seq(0, max_col+1000, by = 1000), cex.axis = 0.7, las=1)
abline(v = xticks, col = "gray", lty = "dotted")
```



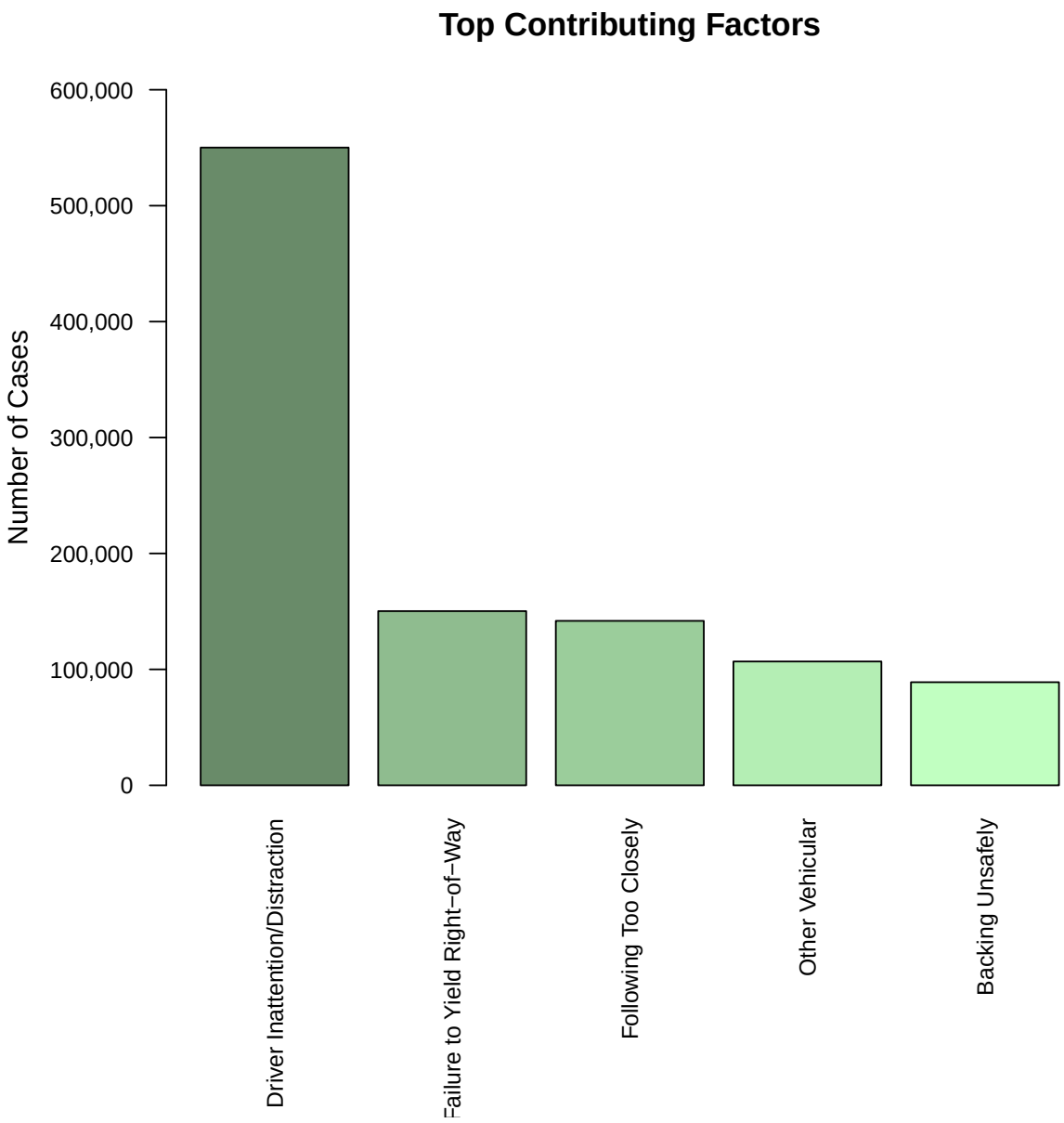
Each borough follows very similar shapes in their respective line. Further analysis to normalize by some factor, e.g. by population or road network density may provide more insight.

2.2 Causes of Collisions

In the dataset every entry has up to 5 'contributing factors'. Below, We see these results aggregated and the top 5 most prevalent causes displayed in a bar chart.

```
cfs <- unique(nyc$CONTRIBUTING.FACTOR.VEHICLE.1)
cont_fact <- c()
for (f in cfs){
  s <- 0
  for (i in 1:5){
    colmn <- paste0("CONTRIBUTING.FACTOR.VEHICLE.",i)
    s <- s + sum(nyc[[colmn]] == f)
```

```
  }
  cont_fact <- append(cont_fact,s)
}
names(cont_fact) <- cfs
fivemost <- sort(cont_fact, decreasing=TRUE)[3:7] #1st and 2nd are unspecified or null

par(mar = c(10, 5, 4, 2))
barplot(fivemost,
        main = "Top Contributing Factors",
        col = c("darkseagreen4", "darkseagreen", "darkseagreen3", "darkseagreen2", "darkseagreen1"),
        las = 2,
        cex.names = 0.8,
        yaxt = "n",
        ylim = c(0,ceiling(max(fivemost)/100000)*100000)
        )

title(ylab = "Number of Cases", line=4)

axis(2, at = seq(0, ceiling(max(fivemost)/100000)*100000, by = 100000),
     labels = format(seq(0, ceiling(max(fivemost)/100000)*100000, by = 100000),
                     big.mark = ",",
                     scientific = FALSE),
     cex.axis = 0.8,
     las = 1,
     )
```



An overwhelming cause of accidents are due to driver distraction, more then 3 times the next most common.