

Final Project Report Draft 1

Motor vehicle collisions in NYC

Group 1 (STAT 605 - S1)

Authors:

Astha Singh - as677@rice.edu

Franco Bosetti - fb57@rice.edu

Louis Clarke - lc160@rice.edu

Lindsey Russ - ltr1@rice.edu

Contents

1	Introduction	2
2	Dataset Description	2
3	Visualizations	3
3.1	Vehicle Collisions over Time	3
3.2	Causes of Collisions	6
3.3	Crashes by Time of Day	8
3.4	Injuries by Person Type	10
3.5	Borough VS Injury Severity	11
3.5.1	Contingency Table	11
3.5.2	Barplot of Row Proportions (by Borough)	12
3.5.3	Cramer's V	13
3.6	Total Persons Injured vs Killed	14
4	Preliminary Statistical Modeling & Interpretation	15
4.1	Code and Results	16
4.2	Interpretation	20

1 Introduction

Motor vehicle collisions are a major concern in major populated urban areas like New York City (NYC), where millions of people use diverse transportation systems every day. Understanding these incidents through data analysis can help us identify patterns and make safety improvements. On the other hand, social media platforms such as Twitter have become spaces where people share their real-time information about traffic incidents and congestion. This project will analyze motor vehicle collision data in NYC and explore the relationship between actual incidents and social media traffic discussions.

2 Dataset Description

Primary Dataset: NYC Motor Vehicle Collisions

(https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data)

Our primary dataset contains police-reported motor vehicle collision records from New York City, maintained by the NYPD through their TrafficStat initiative. The dataset includes 2.21 million collision records with 29 variables.

Source: NYC OpenData (2025)

Key variables include:

- Location & Time: Borough, coordinates (latitude & longitude), street names, zip codes, crash date/time
- Casualties: Injuries and fatalities by category (pedestrians, cyclists, motorists)
- Contributing Factors: Up to five causal factors per collision
- Vehicle Types: Vehicle type classifications (ATV, bicycle, car/SUV, e-bike, e-scooter, truck/bus, motorcycle, other) for each vehicle involved

Preprocessing: We prepared the dataset for analysis by converting crash dates to a standard date format, organizing borough names as categorical variables, and creating additional time-based variables to explore patterns across years and months.

Secondary Dataset: Twitter Traffic Classifications (<https://data.mendeley.com/datasets/c3xvj5snvv/1>)

Our secondary dataset contains classified tweets collected from Twitter’s API, with each tweet categorized by its traffic-related content.

Source: Research dataset by Sina Dabiri (2018)

Variables include:

- Tweet ID
- Classification label:
- Non-Traffic (Class 0): General tweets unrelated to traffic
- Traffic Incident (Class 1): Reports of crashes, breakdowns, road closures
- Traffic Conditions (Class 2): Congestion reports, traffic advisories, flow conditions
- Tweet text content

Preprocessing: We haven’t processed the Twitter data yet, but plan to rename variables and clean the text data to prepare it for our analysis and comparison with the collision records.

3 Visualizations

Prepare the data for processing:

```
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")
```

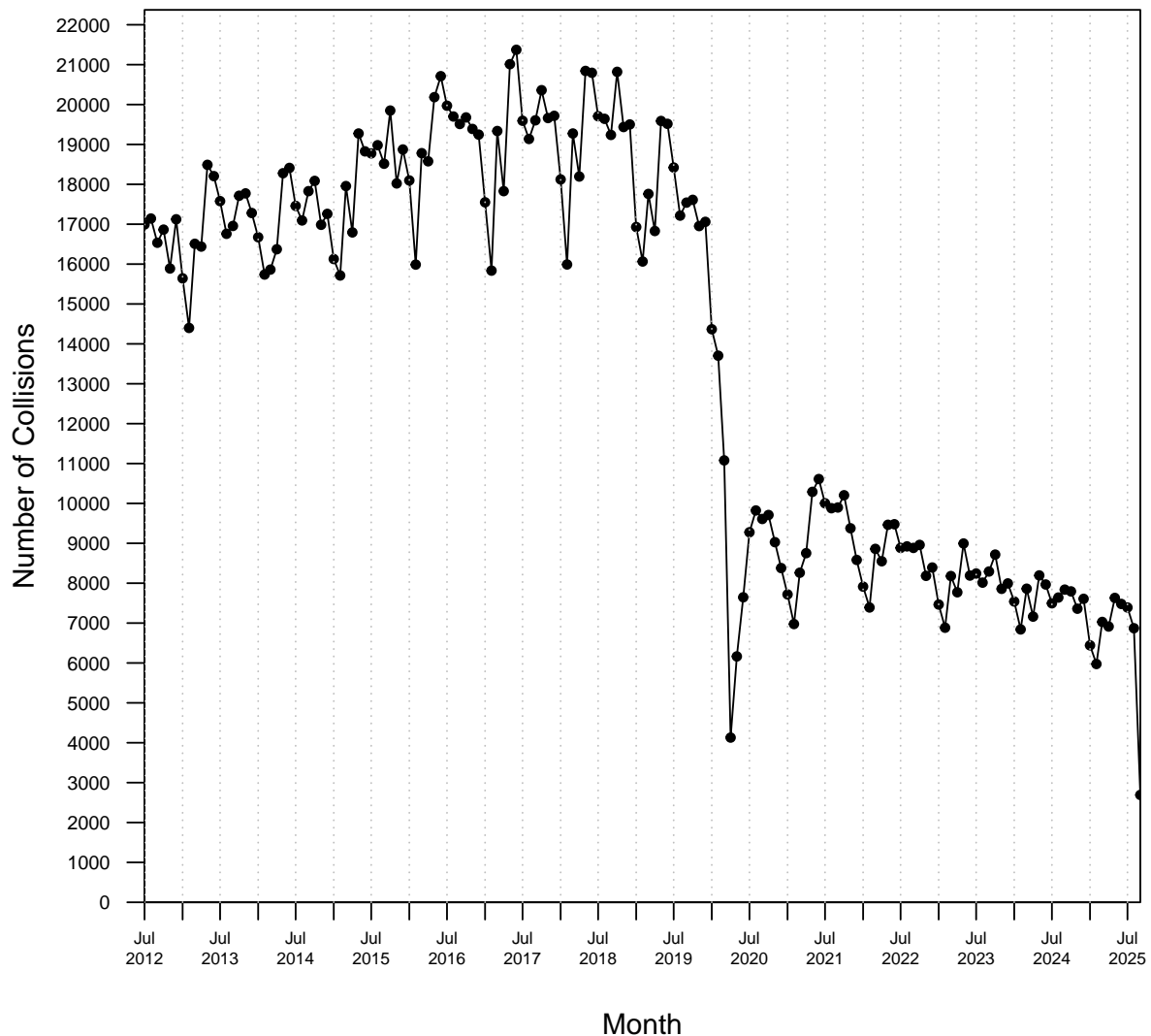
3.1 Vehicle Collisions over Time

By grouping the collisions into each month in which they occurred, 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")
```

Number of Monthly Vehicle Collisions in New York City



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.

```
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))

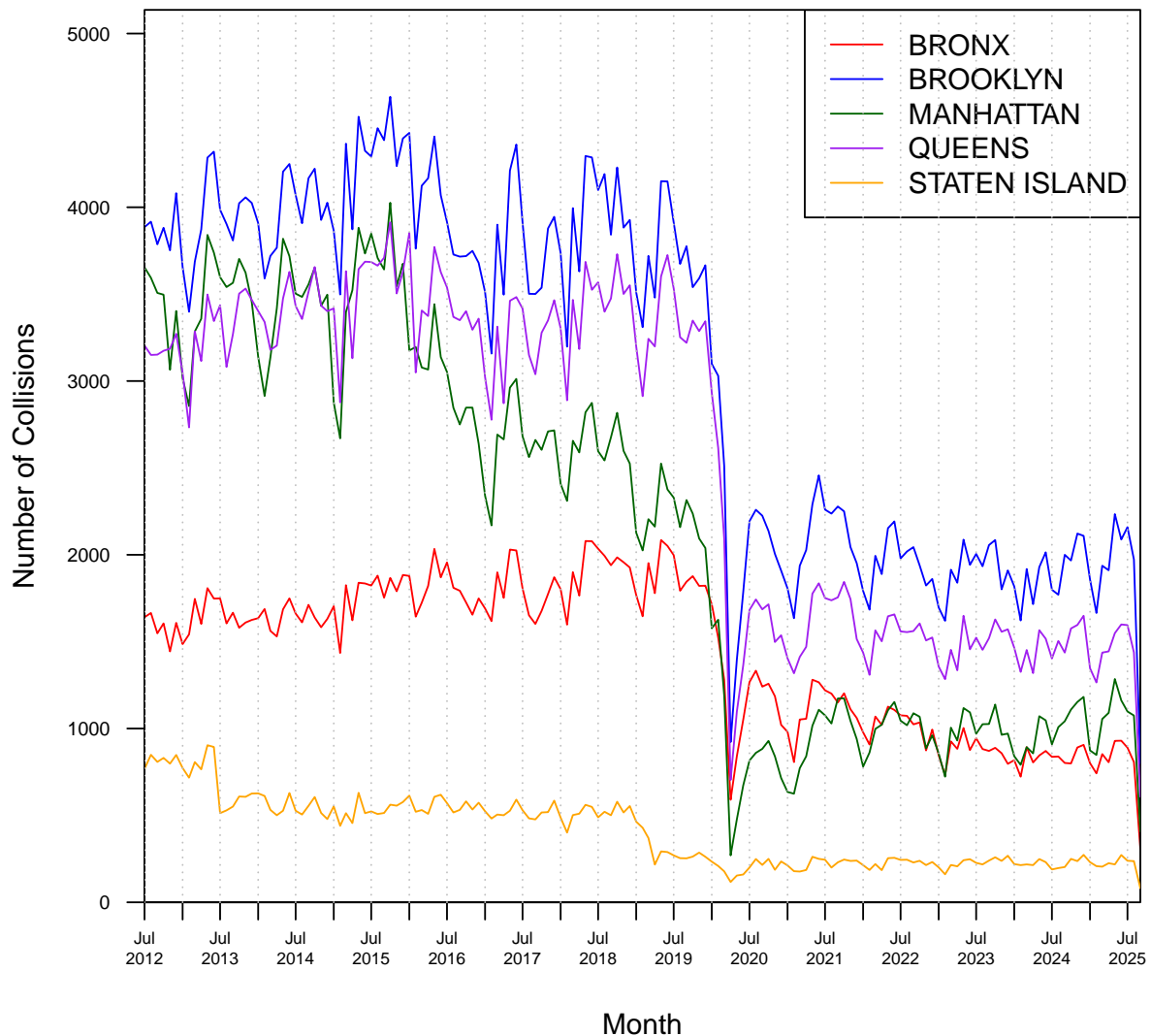
for (i in 2:(length(borough))){
  lines(plot_months[[i]], colspermonth[[i]], type="l", pch=20, col=colors[i])
}

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")

```

Number of Monthly Vehicle Collisions in New York City



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.

3.2 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){
    column <- paste0("CONTRIBUTING.FACTOR.VEHICLE.",i)
    s <- s + sum(nyc[[column]] == f)
  }
}
```

```

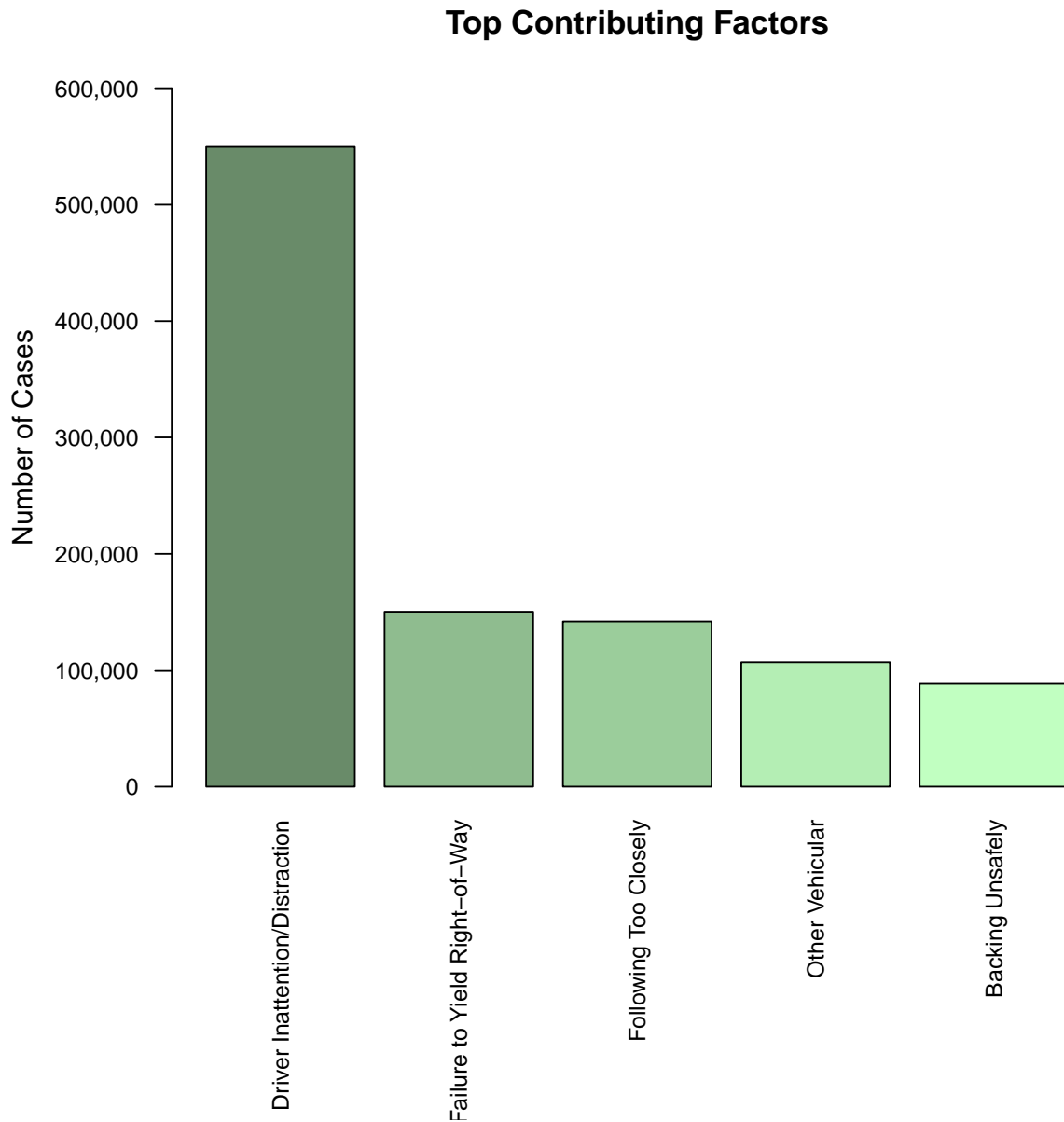
}
cont_fact <- append(cont_fact,s)
}
names(cont_fact) <- cfs
fivemost <- sort(cont_fact, decreasing=TRUE)[3:7]
#1st and 2nd are unspecified or null

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 than 3 times the next most common.

3.3 Crashes by Time of Day

The histogram below shows the total number of accidents in the dataset for each time of day.

```
layout(mat = 1)
times <- nyc$CRASH.TIME
hours <- as.numeric(sub("\\:.*", "", times))
par(tck = -0.015, mgp = c(1.5, 0.4, 0), mar = c(3, 3, 3, 1.5))
hist(hours, xlab = "Time of day", ylab = "Number of accidents", axes = FALSE,
     main="Total Number of Accidents per Time of Day",
     breaks=seq(-0.5, 23.5, 1))

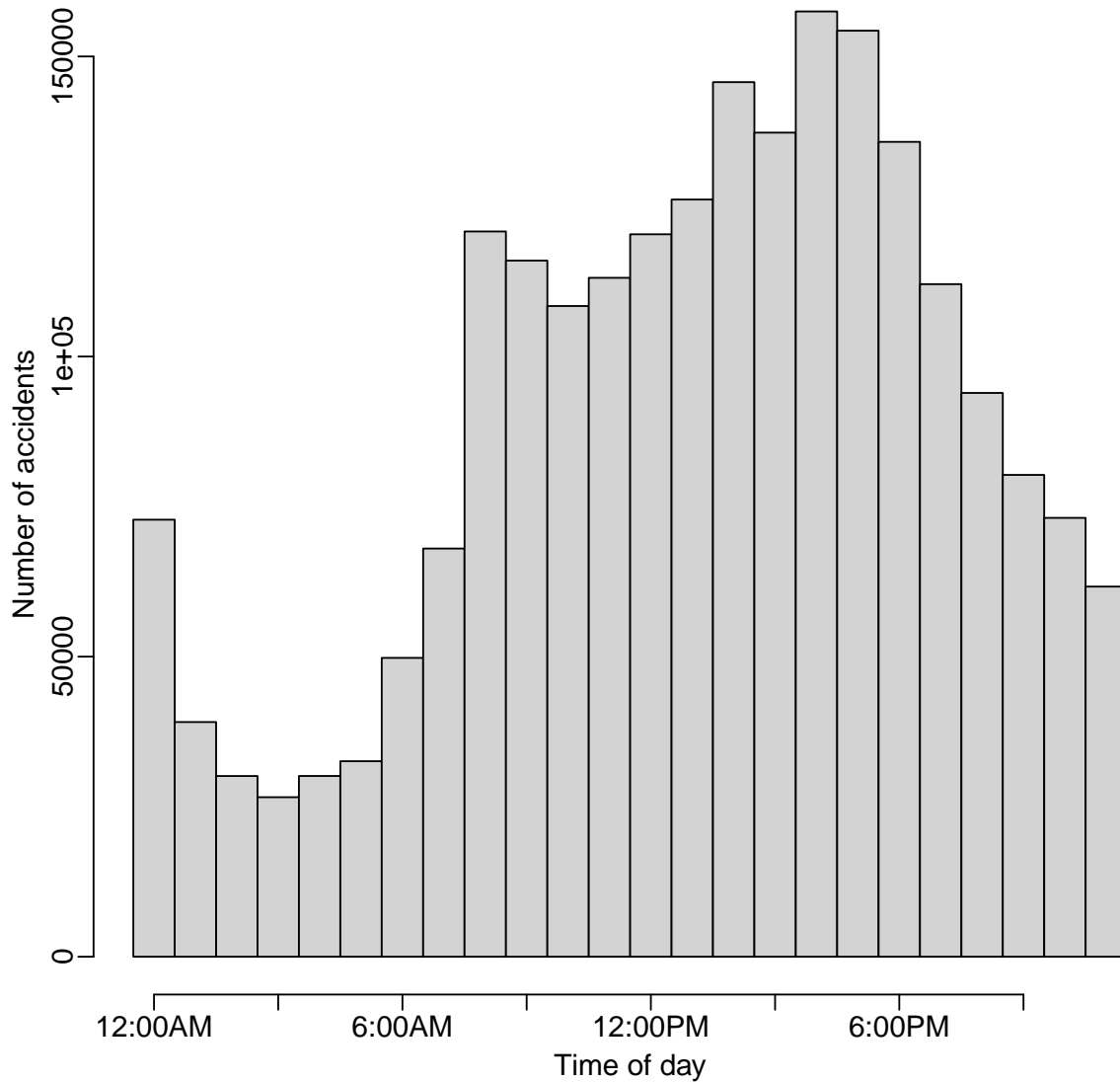
labels <- c("12:00AM", "3:00AM", "6:00AM", "9:00AM",
```

```

      "12:00PM", "3:00PM", "6:00PM", "9:00PM")
axp <- seq(0, 23, by=3)
axis(1, at=axp, labels = labels)
axp2 <- seq(0, 150000, by=50000)
axis(2, at = axp2, labels = axp2)

```

Total Number of Accidents per Time of Day



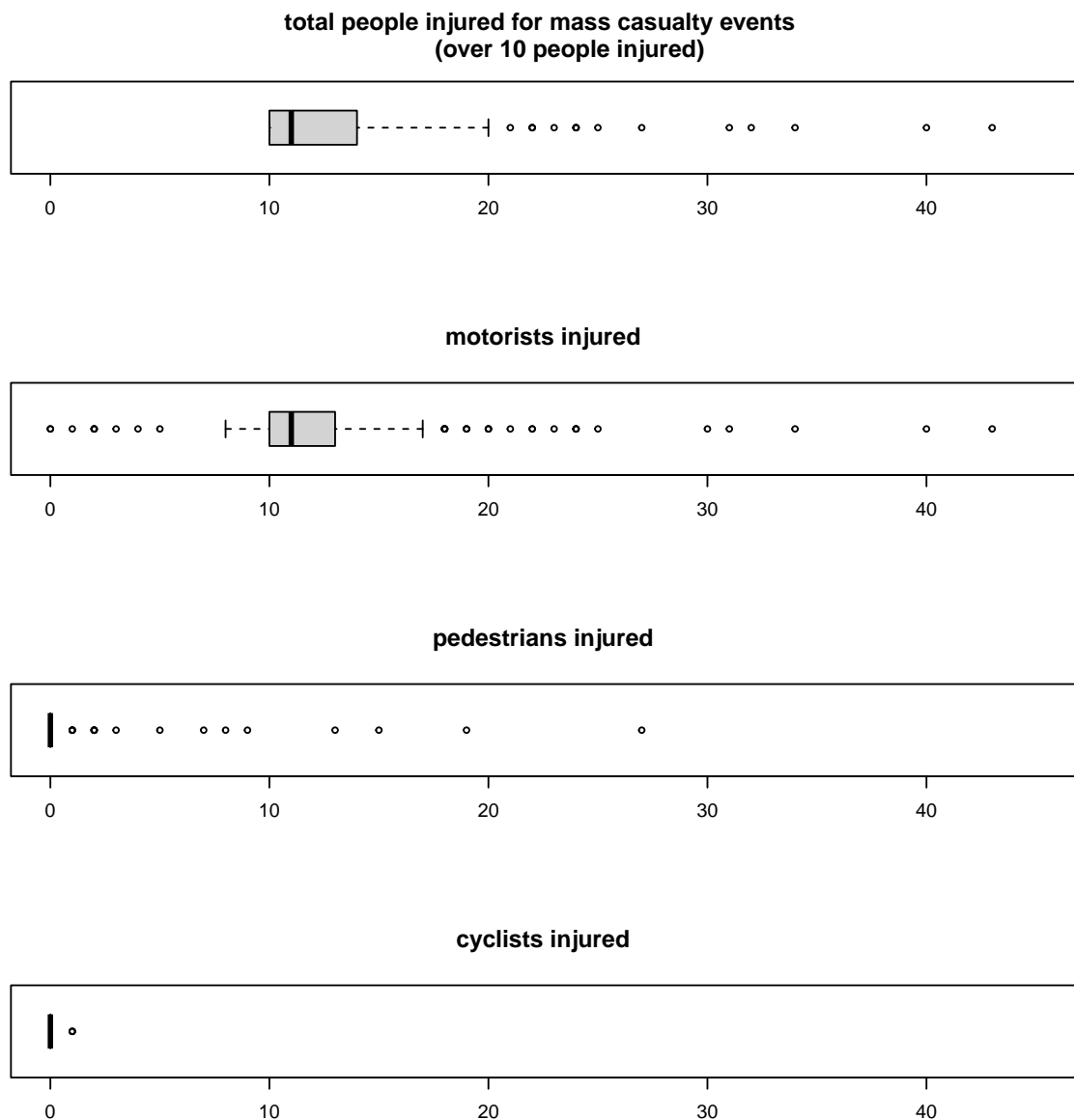
We can see that most of the accidents occur during the daytime, with a semi-continuous increasing and decreasing structure. The distribution reaches its lowest value at 3:00AM, and peaks around 4:00PM. There are slight spikes around 8-9AM, when people would be driving to work, and also around 4-5:00PM, when people would be leaving work. There is also a small spike around midnight.

3.4 Injuries by Person Type

This series of box plots focuses on collisions that resulted in at least 10 total injuries. We compare the total number of people injured with the amounts of people injured of different types: pedestrians, motorists, and cyclists.

```
people_injured <- nyc$NUMBER.OF.PERSONS.INJURED
pedestrians_injured <- nyc$NUMBER.OF.PEDESTRIANS.INJURED
motorists_injured <- nyc$NUMBER.OF.MOTORIST.INJURED
cyclists_injured <- nyc$NUMBER.OF.CYCLIST.INJURED

mass_factor <- people_injured >= 10
people_injured_mass <- people_injured[mass_factor]
motorists_injured_mass <- motorists_injured[mass_factor]
pedestrians_injured_mass <- pedestrians_injured[mass_factor]
cyclists_injured_mass <- cyclists_injured[mass_factor]
layout(mat = matrix(c(1, 2, 3, 4), 4, byrow = TRUE))
boxplot(people_injured_mass,
        main="total people injured for mass casualty events
              (over 10 people injured)",
        horizontal=TRUE, ylim = c(0, 45))
boxplot(motorists_injured_mass, main="motorists injured",
        horizontal=TRUE, ylim = c(0, 45))
boxplot(pedestrians_injured_mass, main="pedestrians injured",
        horizontal=TRUE, ylim = c(0, 45))
boxplot(cyclists_injured_mass, main = "cyclists injured",
        horizontal=TRUE, ylim = c(0, 45))
```



The vast majority of these collisions resulted in motorists being the main affected group. This is unsurprising, as every collision must involve motorists, but the other possible parties may not even be present. Pedestrians were mostly uninjured by these collisions, with some exceptions that result in the amount of pedestrian injuries to go into the 20s. Cyclists were the most unaffected by these mass injury collisions, with a couple of collisions injuring one cyclist, but otherwise 0 cyclists were injured in any of the other collisions.

3.5 Borough VS Injury Severity

3.5.1 Contingency Table

This contingency table examines whether crash severity (fatal, injury, or no injury) varies across NYC boroughs.

```

nyc$injury_severity <- ifelse(nyc$NUMBER.OF.CYCLIST.KILLED > 0,
                             "FATAL/DIED",
                             ifelse(nyc$NUMBER.OF.PERSONS.INJURED > 0,
                                     "INJURY", "NO INJURY"))

# removing NA's
borough_injury_data <- nyc[!is.na(nyc$BOROUGH) & !is.na(nyc$injury_severity), ]

# Contingency Table
(borough_injury_table <- table(borough_injury_data$BOROUGH,
                              borough_injury_data$injury_severity))

##
##           FATAL/DIED INJURY NO INJURY
##           75 170997    506210
## BRONX          30  57186    169143
## BROOKLYN       75 127827    361518
## MANHATTAN      48  64000    274797
## QUEENS         40  97339    312283
## STATEN ISLAND   4  14005     50023

# Chi-square test
(chi_test1 <- chisq.test(borough_injury_table))

##
## Pearson's Chi-squared test
##
## data:  borough_injury_table
## X-squared = 6988.6, df = 10, p-value < 2.2e-16

```

Since $p\text{-value} \leq \alpha = 0.05$, we reject the null hypothesis and conclude that we have sufficient evidence to prove the true population of boroughs are not evenly distributed or in other words, that the distribution of fatal, injury, and property-damage-only crashes varies significantly across NYC boroughs (Chi-square, $df=8$, $p\text{-value} \leq 0.0001$). Furthermore, there is a massive test statistic of $X^2 = 6377.3$ which shows the huge differences between what we'd expect if boroughs were all the same vs. what we actually observe.

3.5.2 Barplot of Row Proportions (by Borough)

This barplot shows crash severity proportions within each borough.

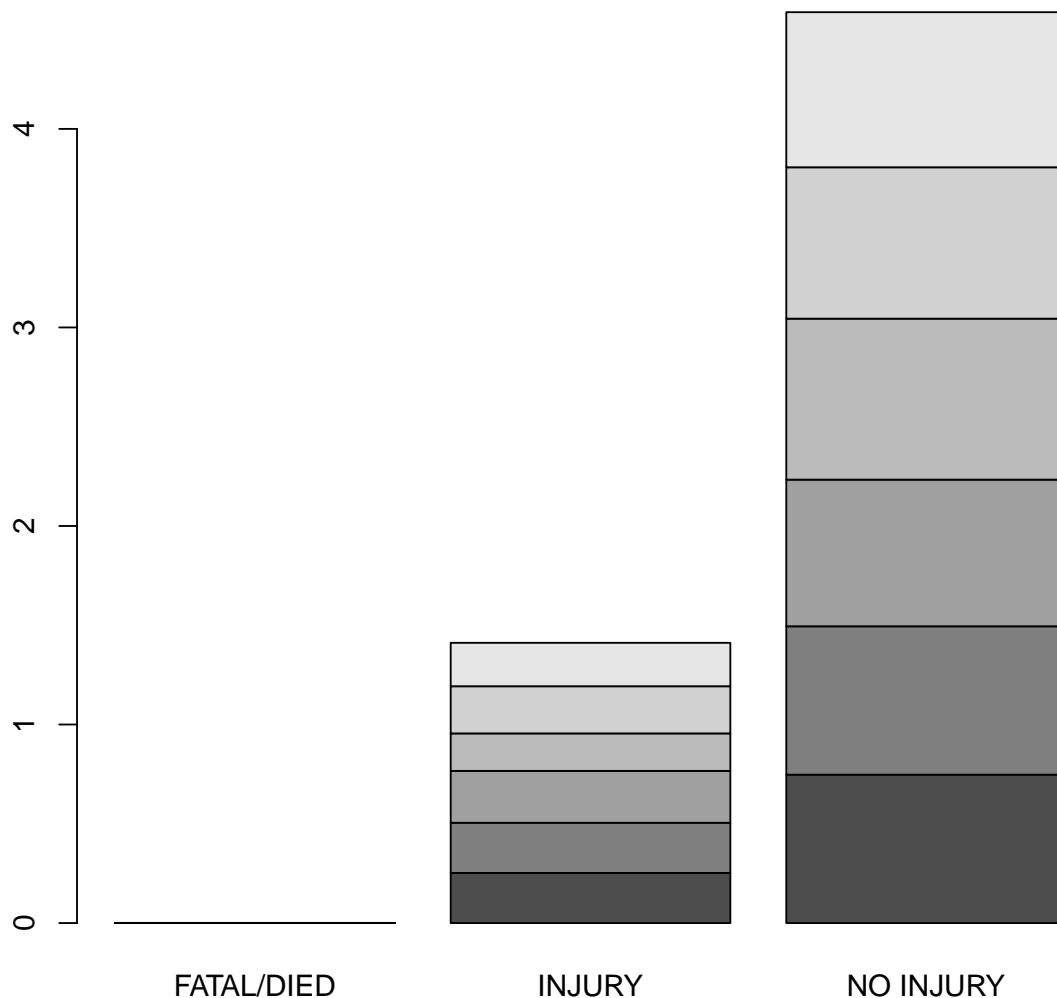
```

(prop_table1 <- prop.table(borough_injury_table, margin = 1))

##
##           FATAL/DIED      INJURY    NO INJURY
## 1.107367e-04 2.524753e-01 7.474139e-01
## BRONX        1.325328e-04 2.526341e-01 7.472334e-01
## BROOKLYN     1.532426e-04 2.611806e-01 7.386662e-01
## MANHATTAN    1.416577e-04 1.888769e-01 8.109814e-01
## QUEENS       9.764147e-05 2.376081e-01 7.622943e-01
## STATEN ISLAND 6.246877e-05 2.187188e-01 7.812188e-01

barplot(prop_table1)

```



While all boroughs have low fatal crash rates (around 0.1-0.16%), Staten Island has the highest fatality rate at 0.159%, followed by Brooklyn at 0.142%. Manhattan appears to be the safest borough with the lowest rates of both fatal (0.107%) and injury crashes (18.9%). Brooklyn and the Bronx have the highest injury rates at 26.1% and 25.3% respectively, while Manhattan has significantly more no injury crashes (80.1%) compared to other boroughs.

3.5.3 Cramer's V

```
cramerV(borough_injury_table)
```

```
## Cramer V
## 0.0398
```

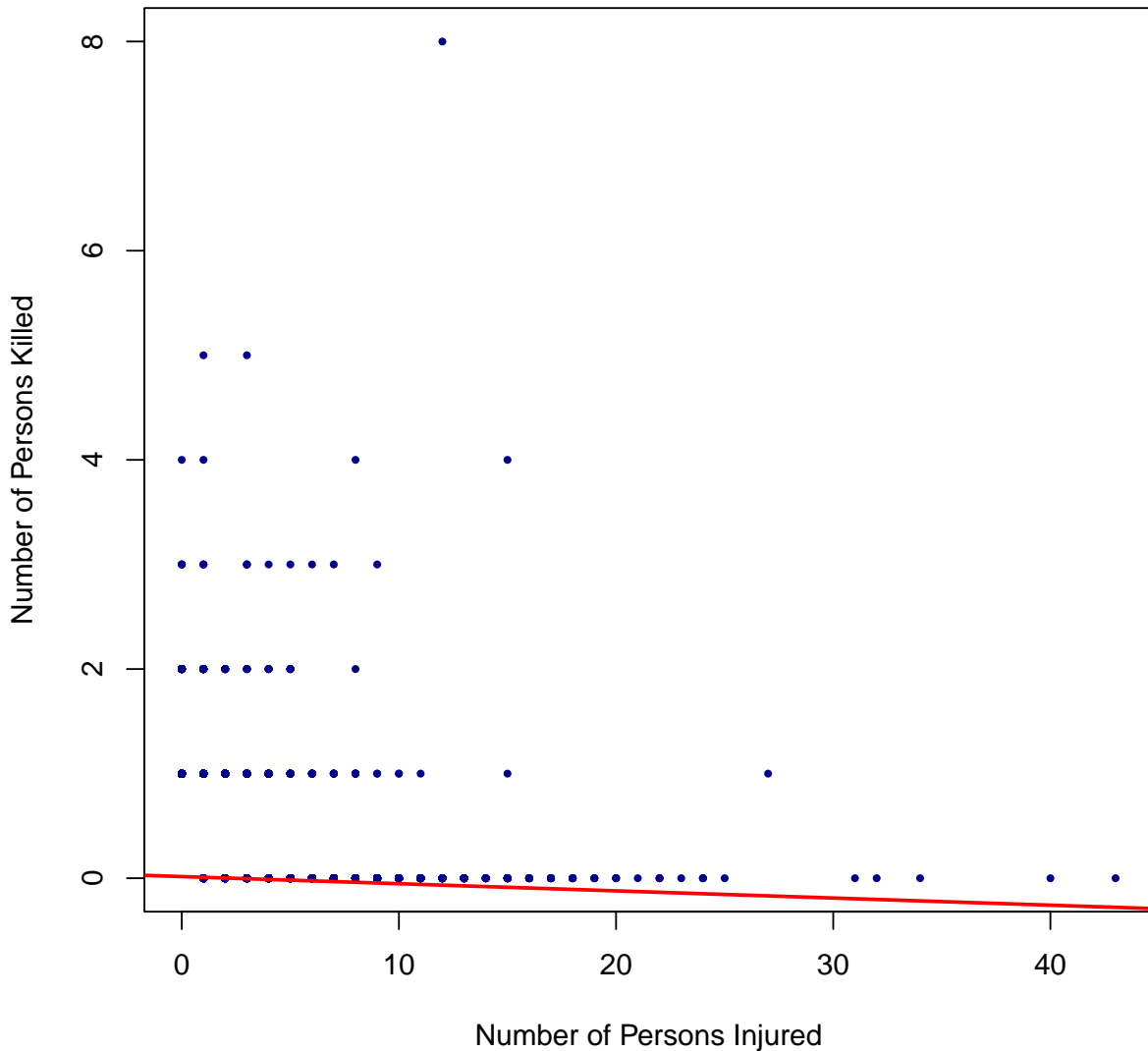
The chi-square test concluded that there is a statistically significant association between borough and crash severity. However, the effect size was very small (Cramer's $V = 0.046$), indicating that while borough differences are statistically detectable, they explain less than 1% of the variation in crash severity. This suggests that factors other than borough location are much more important in determining crash outcomes.

3.6 Total Persons Injured vs Killed

This scatterplot examines the relationship between injuries and fatalities in NYC vehicle crashes.

```
injury_data <- nyc[nyc$NUMBER.OF.PERSONS.INJURED > 0 |  
                  nyc$NUMBER.OF.PERSONS.KILLED > 0, ]  
  
plot(injury_data$NUMBER.OF.PERSONS.INJURED,  
     injury_data$NUMBER.OF.PERSONS.KILLED,  
     main = "Relationship Between Injuries and Fatalities in NYC Crashes",  
     xlab = "Number of Persons Injured",  
     ylab = "Number of Persons Killed",  
     pch = 16,  
     col = "darkblue",  
     cex = 0.6)  
  
abline(lm(injury_data$NUMBER.OF.PERSONS.KILLED ~  
          injury_data$NUMBER.OF.PERSONS.INJURED),  
       col = "red", lwd = 2)
```

Relationship Between Injuries and Fatalities in NYC Crashes



The plot shows that most crashes result in injuries but no deaths which can be seen by the concentration of points along the bottom. Fatal crashes are rare and scattered across injury levels, with no clear correlation. The flat regression line shows that there is a minimal relationship between injuries and deaths, suggesting fatalities depend more on crash circumstances than the number of people involved.

4 Preliminary Statistical Modeling & Interpretation

To better understand the factors that influence the severity of motor vehicle collisions in New York City, we developed a preliminary statistical model using logistic regression. The outcome of interest is whether a crash was fatal (at least one person killed) or non-fatal. Logistic regression is appropriate in this context because the outcome is binary (fatal vs. non-fatal), and it allows us to estimate how different factors change the odds of a fatal crash while holding other conditions constant.

The predictors included in the model were the borough of occurrence, time of day (day vs. night), vehicle

type, and the primary contributing factor reported by police. By examining the estimated odds ratios and their statistical significance, we can identify which circumstances or behaviors are most strongly associated with fatal outcomes.

4.1 Code and Results

Data cleansing and transformations:

```
nyc_small <- nyc %>%
  filter(BOROUGH != "" &
         CONTRIBUTING.FACTOR.VEHICLE.1 != "" &
         CONTRIBUTING.FACTOR.VEHICLE.1 != "1" &
         CONTRIBUTING.FACTOR.VEHICLE.1 != "80") %>%
  transmute(
    is_fatal = ifelse(NUMBER.OF.PERSONS.KILLED > 0, 1, 0), # outcome
    borough = BOROUGH,
    crash_hour = hour(hm(CRASH.TIME)),
    contributing_factor = CONTRIBUTING.FACTOR.VEHICLE.1,
    vehicle_cat = factor(case_when(
      grepl("MOTORCYCLE|MOTORBIKE", VEHICLE.TYPE.CODE.1, ignore.case = TRUE) ~
        "motorcycle",
      grepl("TRUCK|BUS", VEHICLE.TYPE.CODE.1, ignore.case = TRUE) ~
        "truck_bus",
      grepl("BICYCLE|BIKE|SCOOTER", VEHICLE.TYPE.CODE.1, ignore.case = TRUE) ~
        "bike_scooter",
      TRUE ~ "car_suv"
    ))
  )

nyc_small <- nyc_small %>%
  mutate(hour = as.numeric(substr(crash_hour, 1, 2)),
         tod = ifelse(hour >= 6 & hour < 18, "day", "night"))

nyc_small$BOROUGH <- droplevels(as.factor(nyc_small$borough))

# set baselines for model
nyc_small$borough <- relevel(as.factor(nyc_small$BOROUGH),
                             ref = "STATEN ISLAND")
nyc_small$tod <- relevel(as.factor(nyc_small$tod),
                         ref = "day")
nyc_small$vehicle_cat <- relevel(as.factor(nyc_small$vehicle_cat),
                                ref = "car_suv")
nyc_small$contributing_factor <- relevel(
  as.factor(nyc_small$contributing_factor), ref = "Unspecified")
```

Creating the logistic regression model:

```
model <- glm(is_fatal ~ borough + tod + vehicle_cat + contributing_factor,
             data = nyc_small,
             family = binomial)

summ <- summary(model)
```

```

coefs <- summ$coefficients
est <- coefs[, "Estimate"]           # log-odds
se <- coefs[, "Std. Error"]          # standard errors

# compute odds ratios + 95% confidence intervals
ci_low <- exp(est - 1.96 * se)
ci_high <- exp(est + 1.96 * se)

results_table <- data.frame(
  term      = rownames(coefs),
  odds_ratio = exp(est),
  ci_low     = ci_low,
  ci_high    = ci_high,
  significant = ifelse(coefs[, "Pr(>|z|)"] < 0.05, "yes", "no"),
  row.names = NULL
)

# filter out the intercept/baseline to keep table cleaner
results_table <- results_table[results_table$term != "(Intercept)", ]

results_table$term <- gsub("vehicle_cat", "Vehicle: ",
                          results_table$term)
results_table$term <- gsub("borough", "Borough: ",
                          results_table$term)
results_table$term <- gsub("tod", "Time of Day: ",
                          results_table$term)
results_table$term <- gsub("contributing_factor", "Contributing Factor: ",
                          results_table$term)

results_table
##                                term
## 2                      Borough: BRONX
## 3                      Borough: BROOKLYN
## 4                      Borough: MANHATTAN
## 5                      Borough: QUEENS
## 6                      Time of Day: night
## 7                      Vehicle: bike_scooter
## 8                      Vehicle: motorcycle
## 9                      Vehicle: truck_bus
## 10                     Contributing Factor: Accelerator Defective
## 11                     Contributing Factor: Aggressive Driving/Road Rage
## 12                     Contributing Factor: Alcohol Involvement
## 13                     Contributing Factor: Animals Action
## 14                     Contributing Factor: Backing Unsafely
## 15                     Contributing Factor: Brakes Defective
## 16                     Contributing Factor: Cell Phone (hand-held)
## 17                     Contributing Factor: Cell Phone (hand-Held)
## 18                     Contributing Factor: Cell Phone (hands-free)
## 19                     Contributing Factor: Driver Inattention/Distractio
## 20                     Contributing Factor: Driver Inexperience
## 21                     Contributing Factor: Driverless/Runaway Vehicle
## 22                     Contributing Factor: Drugs (illegal)
## 23                     Contributing Factor: Drugs (Illegal)

```

```

## 24          Contributing Factor: Eating or Drinking
## 25          Contributing Factor: Failure to Keep Right
## 26          Contributing Factor: Failure to Yield Right-of-Way
## 27          Contributing Factor: Fatigued/Drowsy
## 28          Contributing Factor: Fell Asleep
## 29          Contributing Factor: Following Too Closely
## 30          Contributing Factor: Glare
## 31          Contributing Factor: Headlights Defective
## 32          Contributing Factor: Illnes
## 33          Contributing Factor: Illness
## 34          Contributing Factor: Lane Marking Improper/Inadequate
## 35          Contributing Factor: Listening/Using Headphones
## 36          Contributing Factor: Lost Consciousness
## 37          Contributing Factor: Obstruction/Debris
## 38          Contributing Factor: Other Electronic Device
## 39          Contributing Factor: Other Lighting Defects
## 40          Contributing Factor: Other Vehicular
## 41          Contributing Factor: Outside Car Distraction
## 42          Contributing Factor: Oversized Vehicle
## 43          Contributing Factor: Passenger Distraction
## 44          Contributing Factor: Passing or Lane Usage Improper
## 45          Contributing Factor: Passing Too Closely
## 46          Contributing Factor: Pavement Defective
## 47          Contributing Factor: Pavement Slippery
## 48 Contributing Factor: Pedestrian/Bicyclist/Other Pedestrian Error/Confusion
## 49          Contributing Factor: Physical Disability
## 50          Contributing Factor: Prescription Medication
## 51          Contributing Factor: Reaction to Other Uninvolved Vehicle
## 52          Contributing Factor: Reaction to Uninvolved Vehicle
## 53          Contributing Factor: Shoulders Defective/Improper
## 54          Contributing Factor: Steering Failure
## 55          Contributing Factor: Texting
## 56          Contributing Factor: Tinted Windows
## 57          Contributing Factor: Tire Failure/Inadequate
## 58          Contributing Factor: Tow Hitch Defective
## 59          Contributing Factor: Traffic Control Device Improper/Non-Working
## 60          Contributing Factor: Traffic Control Disregarded
## 61          Contributing Factor: Turning Improperly
## 62          Contributing Factor: Unsafe Lane Changing
## 63          Contributing Factor: Unsafe Speed
## 64          Contributing Factor: Using On Board Navigation Device
## 65          Contributing Factor: Vehicle Vandalism
## 66          Contributing Factor: View Obstructed/Limited
## 67          Contributing Factor: Windshield Inadequate
##      odds_ratio      ci_low      ci_high significant
## 2  7.800684e-01  6.224261e-01  9.776368e-01      yes
## 3  8.206692e-01  6.657245e-01  1.011677e+00      no
## 4  6.641701e-01  5.319389e-01  8.292720e-01      yes
## 5  7.858335e-01  6.353490e-01  9.719606e-01      yes
## 6  1.817549e+00  1.661621e+00  1.988110e+00      yes
## 7  2.772972e+00  2.206805e+00  3.484394e+00      yes
## 8  1.405029e+01  1.191286e+01  1.657121e+01      yes
## 9  3.284954e+00  2.856664e+00  3.777456e+00      yes

```

## 10	8.865442e-01	1.243119e-01	6.322489e+00	no
## 11	1.190088e+00	6.860713e-01	2.064378e+00	no
## 12	1.892357e+00	1.420115e+00	2.521638e+00	yes
## 13	6.312020e-01	8.850369e-02	4.501688e+00	no
## 14	4.015649e-01	2.816217e-01	5.725920e-01	yes
## 15	2.616205e-06	1.905631e-136	3.591739e+124	no
## 16	2.626197e-06	0.000000e+00	Inf	no
## 17	2.543068e-06	0.000000e+00	Inf	no
## 18	2.749615e-06	0.000000e+00	Inf	no
## 19	6.469233e-01	5.588991e-01	7.488109e-01	yes
## 20	1.161200e+00	8.433343e-01	1.598874e+00	no
## 21	2.148334e+00	6.890774e-01	6.697853e+00	no
## 22	9.603969e+00	4.742405e+00	1.944925e+01	yes
## 23	7.909210e+00	2.525750e+00	2.476714e+01	yes
## 24	2.560611e-06	0.000000e+00	Inf	no
## 25	6.905380e-01	1.721712e-01	2.769584e+00	no
## 26	2.159502e+00	1.872314e+00	2.490741e+00	yes
## 27	3.462195e-02	4.869166e-03	2.461776e-01	yes
## 28	1.533115e+00	7.625250e-01	3.082444e+00	no
## 29	2.334913e-01	1.399643e-01	3.895149e-01	yes
## 30	1.263585e+00	4.720836e-01	3.382127e+00	no
## 31	1.901810e-06	0.000000e+00	Inf	no
## 32	1.298353e+01	8.180360e+00	2.060692e+01	yes
## 33	2.819541e-06	1.596854e-214	4.978422e+202	no
## 34	2.507140e-06	0.000000e+00	Inf	no
## 35	1.400264e-06	0.000000e+00	Inf	no
## 36	1.389860e+00	9.383349e-01	2.058659e+00	no
## 37	2.667570e-01	3.747518e-02	1.898838e+00	no
## 38	9.700006e-01	2.417603e-01	3.891876e+00	no
## 39	5.432201e+00	7.560456e-01	3.903047e+01	no
## 40	2.350121e-01	1.383635e-01	3.991710e-01	yes
## 41	2.838662e-01	9.126375e-02	8.829356e-01	yes
## 42	1.336765e-01	3.330782e-02	5.364926e-01	yes
## 43	4.403438e+00	3.155602e+00	6.144714e+00	yes
## 44	4.848878e-01	3.275762e-01	7.177451e-01	yes
## 45	1.873880e-02	2.635163e-03	1.332527e-01	yes
## 46	2.926542e-01	4.103354e-02	2.087230e+00	no
## 47	1.300188e-01	3.244025e-02	5.211082e-01	yes
## 48	4.425379e+00	3.342686e+00	5.858754e+00	yes
## 49	2.721378e+00	1.823227e+00	4.061975e+00	yes
## 50	3.930060e-01	1.758458e-01	8.783475e-01	yes
## 51	2.870500e-06	2.105411e-282	3.913617e+270	no
## 52	4.253698e-01	1.902874e-01	9.508750e-01	yes
## 53	2.567836e-06	0.000000e+00	Inf	no
## 54	2.354771e-06	2.074365e-203	2.673082e+191	no
## 55	2.294992e-06	0.000000e+00	Inf	no
## 56	1.005801e+01	2.451410e+00	4.126751e+01	yes
## 57	2.454227e-06	9.819917e-260	6.133687e+247	no
## 58	3.491929e+00	4.865653e-01	2.506050e+01	no
## 59	2.618604e-06	0.000000e+00	Inf	no
## 60	5.208127e+00	4.443561e+00	6.104245e+00	yes
## 61	2.923187e-01	1.721300e-01	4.964284e-01	yes
## 62	1.249143e-01	4.017034e-02	3.884357e-01	yes

##	63	5.107109e+00	4.259591e+00	6.123254e+00	yes
##	64	2.437281e-06	0.000000e+00	Inf	no
##	65	2.612111e-06	0.000000e+00	Inf	no
##	66	1.239972e+00	7.757244e-01	1.982058e+00	no
##	67	2.630546e-06	0.000000e+00	Inf	no

4.2 Interpretation

The revised model confirms several strong predictors of fatal crashes. Compared to Staten Island, crashes in the Bronx, Queens, and Manhattan show significantly lower odds of being fatal, while Brooklyn is not statistically different. Nighttime crashes remain much riskier, with nearly 1.8 times higher odds of fatality compared to daytime. Vehicle type shows especially pronounced differences: motorcycle crashes are about 14 times more likely to be fatal than car/SUV crashes, while trucks and buses carry over 3 times the risk and bicycles/scooters nearly 3 times the risk.

Reported contributing factors highlight behaviors that dramatically elevate risk. Crashes linked to alcohol are almost 2 times more likely to result in death, while those involving drug use show odds 8–10 times higher than the baseline. Dangerous driving behaviors such as speeding, traffic signal violations, unsafe lane changes, and pedestrian/bicyclist error all significantly increase fatal crash likelihood, often by a factor of 4–5 or more. By contrast, mechanical defect categories (e.g., “brakes defective”) produce unstable estimates due to rare occurrences and should be interpreted cautiously.

Overall, these results suggest that time of day, vehicle type, and risky human behaviors are the dominant predictors of fatal outcomes in NYC crashes, while borough differences are smaller in magnitude.