Group 3: Crash Analysis NYC

Kevin Lei, Micah Kepe, Zachary Kepe, Giulia Costantini

2024-02-15

Table of contents


```
data <- read.csv("../data/Motor_Vehicle_Collisions_-_Crashes_20240213.csv")
data <- as.data.frame(data)
```

Introduction

The NYC OpenData dataset contains details of motor vehicle collision occurrences reported by the NYPD from 2012 to 2024. Collisions are only reported if a person was injured or killed, or if there was more than $1000 in damage. Our project seeks to determine the most important factors that contribute to potentially fatal accidents. By analyzing the data, we hope to identify to provide recommendations for reducing the number of accidents.

Data Description

The dataset contains over 2 million records, each detailing a motor vehicle collision. The data includes the date, time, location (latitude, longitude, borough), and the number of individuals affected (injured or killed) among drivers, pedestrians, and cyclists. The dataset also includes information about the vehicles involved, such as the type of vehicle and the contributing factors to the crash. Our analysis will focus on understanding patterns and trends within these incidents, and identifying the most important factors that contribute to potentially fatal accidents.

```
head(data)
```

```
  CRASH.DATE CRASH.TIME  BOROUGH ZIP.CODE LATITUDE LONGITUDE
1 09/11/2021       2:39                NA       NA        NA
2 03/26/2022      11:45                NA       NA        NA
3 06/29/2022       6:55                NA       NA        NA
4 09/11/2021       9:35 BROOKLYN    11208  40.6672 -73.86650
5 12/14/2021       8:13 BROOKLYN    11233  40.6833 -73.91727
6 04/14/2021      12:47                NA       NA        NA
                LOCATION              ON.STREET.NAME CROSS.STREET.NAME
1                               WHITESTONE EXPRESSWAY         20 AVENUE
2                             QUEENSBORO BRIDGE UPPER
3                                 THROGS NECK BRIDGE
4   (40.667202, -73.8665)
5 (40.683304, -73.917274)              SARATOGA AVENUE    DECATUR STREET
6                         MAJOR DEEGAN EXPRESSWAY RAMP
         OFF.STREET.NAME NUMBER.OF.PERSONS.INJURED NUMBER.OF.PERSONS.KILLED
1                                                2                        0
2                                                1                        0
3                                                0                        0
4 1211      LORING AVENUE                        0                        0
5                                                0                        0
6                                                0                        0
  NUMBER.OF.PEDESTRIANS.INJURED NUMBER.OF.PEDESTRIANS.KILLED
1                             0                            0
2                             0                            0
3                             0                            0
4                             0                            0
5                             0                            0
6                             0                            0
  NUMBER.OF.CYCLIST.INJURED NUMBER.OF.CYCLIST.KILLED NUMBER.OF.MOTORIST.INJURED
1                         0                       0                          2
2                         0                       0                          1
3                         0                       0                          0
4                         0                       0                          0
5                         0                       0                          0
6                         0                       0                          0
  NUMBER.OF.MOTORIST.KILLED CONTRIBUTING.FACTOR.VEHICLE.1
1                         0  Aggressive Driving/Road Rage
2                         0              Pavement Slippery
3                         0         Following Too Closely
4                         0                   Unspecified
5                         0
6                         0                   Unspecified
  CONTRIBUTING.FACTOR.VEHICLE.2 CONTRIBUTING.FACTOR.VEHICLE.3
1                   Unspecified
2
3                   Unspecified
4
5
6                   Unspecified
  CONTRIBUTING.FACTOR.VEHICLE.4 CONTRIBUTING.FACTOR.VEHICLE.5 COLLISION_ID
1                                                                 4455765
2                                                                 4513547
3                                                                 4541903
4                                                                 4456314
```

```
5                                                                    4486609
6                                                                    4407458
  VEHICLE.TYPE.CODE.1 VEHICLE.TYPE.CODE.2 VEHICLE.TYPE.CODE.3
1               Sedan               Sedan
2               Sedan
3               Sedan       Pick-up Truck
4               Sedan
5
6                Dump               Sedan
  VEHICLE.TYPE.CODE.4 VEHICLE.TYPE.CODE.5
1
2
3
4
5
6
```

Data Cleaning

```
# Load required libraries
# install.packages("chron")
library(chron)

# Remove rows with missing values in critical columns
cleaned_data <- data[complete.cases(data[c("BOROUGH", "LATITUDE", "LONGITUDE", "CRASH.DATE", "CRASH.TIM

# Convert CRASH.DATE to Date format
cleaned_data$CRASH.DATE <- as.Date(cleaned_data$CRASH.DATE, format="%m/%d/%Y")

# Correct outlier LATITUDE and LONGITUDE values (assuming NYC coordinates)
cleaned_data <- cleaned_data[cleaned_data$LATITUDE > 40 & cleaned_data$LATITUDE < 41 & cleaned_data$LON

# convert CRASH.TIME to time format
cleaned_data$CRASH.TIME <- times(format(strptime(cleaned_data$CRASH.TIME,
                                                 format="%H:%M"), "%H:%M:%S"))

# Change empty string values in BOROUGH to 'UNKNOWN'
cleaned_data$BOROUGH[cleaned_data$BOROUGH == ""] <- "UNKNOWN"

# Change all empty string values in street names to 'UNKNOWN'
cleaned_data$ON.STREET.NAME[cleaned_data$ON.STREET.NAME == ""] <- "UNKNOWN"
cleaned_data$CROSS.STREET.NAME[cleaned_data$CROSS.STREET.NAME == ""] <- "UNKNOWN"
cleaned_data$OFF.STREET.NAME[cleaned_data$OFF.STREET.NAME == ""] <- "UNKNOWN"

# Display the cleaned data
head(cleaned_data)
```

```
   CRASH.DATE CRASH.TIME   BOROUGH ZIP.CODE LATITUDE LONGITUDE
4  2021-09-11   09:35:00  BROOKLYN    11208 40.66720 -73.86650
5  2021-12-14   08:13:00  BROOKLYN    11233 40.68330 -73.91727
7  2021-12-14   17:05:00   UNKNOWN       NA 40.70918 -73.95682
8  2021-12-14   08:17:00     BRONX    10475 40.86816 -73.83148
```

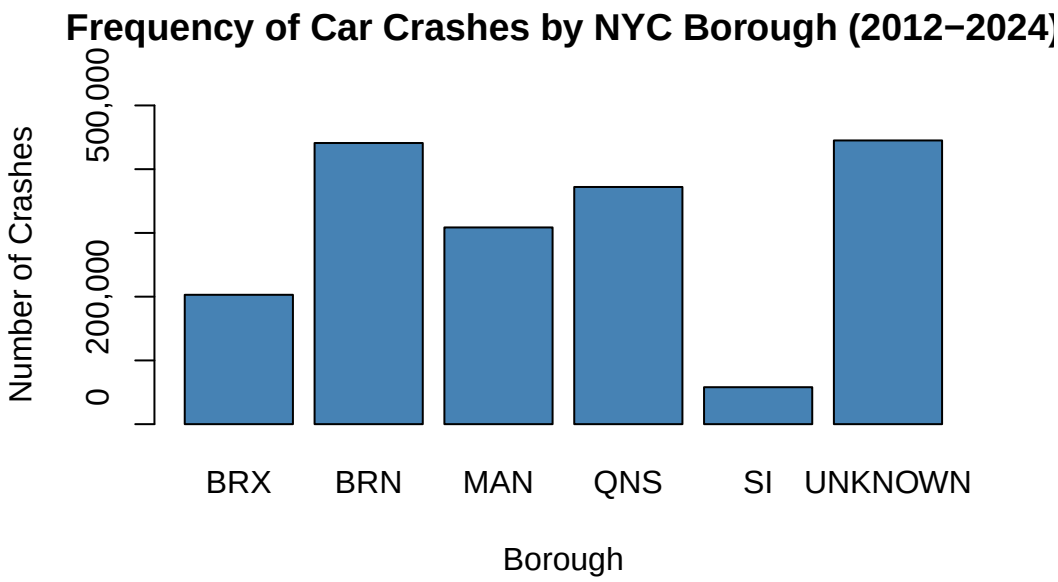
```
9  2021-12-14   21:10:00  BROOKLYN    11207 40.67172 -73.89710
10 2021-12-14   14:58:00 MANHATTAN    10017 40.75144 -73.97397
                  LOCATION             ON.STREET.NAME CROSS.STREET.NAME
4    (40.667202, -73.8665)                    UNKNOWN           UNKNOWN
5  (40.683304, -73.917274)             SARATOGA AVENUE    DECATUR STREET
7  (40.709183, -73.956825) BROOKLYN QUEENS EXPRESSWAY           UNKNOWN
8    (40.86816, -73.83148)                    UNKNOWN           UNKNOWN
9     (40.67172, -73.8971)                    UNKNOWN           UNKNOWN
10   (40.75144, -73.97397)                  3 AVENUE    EAST 43 STREET
              OFF.STREET.NAME NUMBER.OF.PERSONS.INJURED
4      1211      LORING AVENUE                        0
5                      UNKNOWN                        0
7                      UNKNOWN                        0
8  344       BAYCHESTER AVENUE                        2
9      2047      PITKIN AVENUE                        0
10                     UNKNOWN                        0
   NUMBER.OF.PERSONS.KILLED NUMBER.OF.PEDESTRIANS.INJURED
4                         0                             0
5                         0                             0
7                         0                             0
8                         0                             0
9                         0                             0
10                        0                             0
   NUMBER.OF.PEDESTRIANS.KILLED NUMBER.OF.CYCLIST.INJURED
4                             0                         0
5                             0                         0
7                             0                         0
8                             0                         0
9                             0                         0
10                            0                         0
   NUMBER.OF.CYCLIST.KILLED NUMBER.OF.MOTORIST.INJURED
4                         0                         0
5                         0                         0
7                         0                         0
8                         0                         2
9                         0                         0
10                        0                         0
   NUMBER.OF.MOTORIST.KILLED CONTRIBUTING.FACTOR.VEHICLE.1
4                          0                  Unspecified
5                          0
7                          0           Passing Too Closely
8                          0                  Unspecified
9                          0           Driver Inexperience
10                         0           Passing Too Closely
   CONTRIBUTING.FACTOR.VEHICLE.2 CONTRIBUTING.FACTOR.VEHICLE.3
4
5
7                    Unspecified
8                    Unspecified
9                    Unspecified
10                   Unspecified
   CONTRIBUTING.FACTOR.VEHICLE.4 CONTRIBUTING.FACTOR.VEHICLE.5 COLLISION_ID
4                                                                   4456314
5                                                                   4486609
```

```
7                                                                    4486555
8                                                                    4486660
9                                                                    4487074
10                                                                   4486519
   VEHICLE.TYPE.CODE.1                VEHICLE.TYPE.CODE.2 VEHICLE.TYPE.CODE.3
4                Sedan
5
7                Sedan               Tractor Truck Diesel
8                Sedan                              Sedan
9                Sedan
10               Sedan Station Wagon/Sport Utility Vehicle
   VEHICLE.TYPE.CODE.4 VEHICLE.TYPE.CODE.5
4
5
7
8
9
10
```

Preliminary Plots

Plot 1: Frequency of Car Crashes by NYC Borough (2012-2024)

```r
# Create a table of the frequency of car crashes by borough
borough_cleaned <- table(cleaned_data$BOROUGH)

# Create abbreviated labels for the boroughs
names(borough_cleaned) <- c("BRX", "BRN", "MAN", "QNS", "SI", "UNKNOWN")

# Create the plot
barplot(borough_cleaned,
        main = "Frequency of Car Crashes by NYC Borough (2012-2024)",
        xlab = "Borough",
        ylab = "Number of Crashes",
        yaxt = "n",
        col = "steelblue",
        ylim = c(0, 500000))

# Add the y-axis labels
axis(2, at=seq(0, 500000, by=100000),
     labels=format(seq(0, 500000, by=100000),
     big.mark=",",
     scientific=FALSE))
```



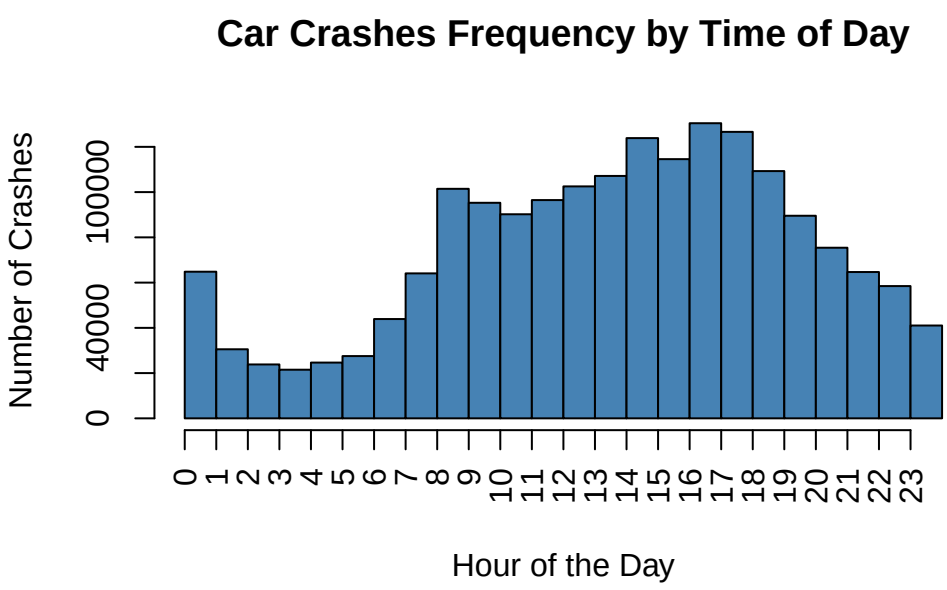
Interpretation

We plotted the frequency of car crashes per borough in NYC. The goal of this graph was to gain a better understanding of which boroughs were more likely to result in car crashes.

The borough with the least car crashes is Staten Island at around 75,000 car crashes, while the borough with the most car crashes is Brooklyn with around 450,000 car crashes. This graph gives us a better understanding of the likelihood that a car crash will occur in a certain borough. However, a consideration that is not specified is the population of each borough and the amount of traffic through them.

Plot 2: Histogram of Car Crashes by Time of Day

```
# Create a histogram of car crashes by time of day
hist((as.numeric(cleaned_data$CRASH.TIME) * 24) %% 24, breaks = 24, col = "steelblue",
     main = "Car Crashes Frequency by Time of Day",
     xlab = "Hour of the Day",
     ylab = "Number of Crashes",
     xaxt = "n")

# Add custom x-axis labels to improve readability
axis(1, at = seq(0, 23, by = 1), labels = seq(0, 23, by = 1), las = 2)
```

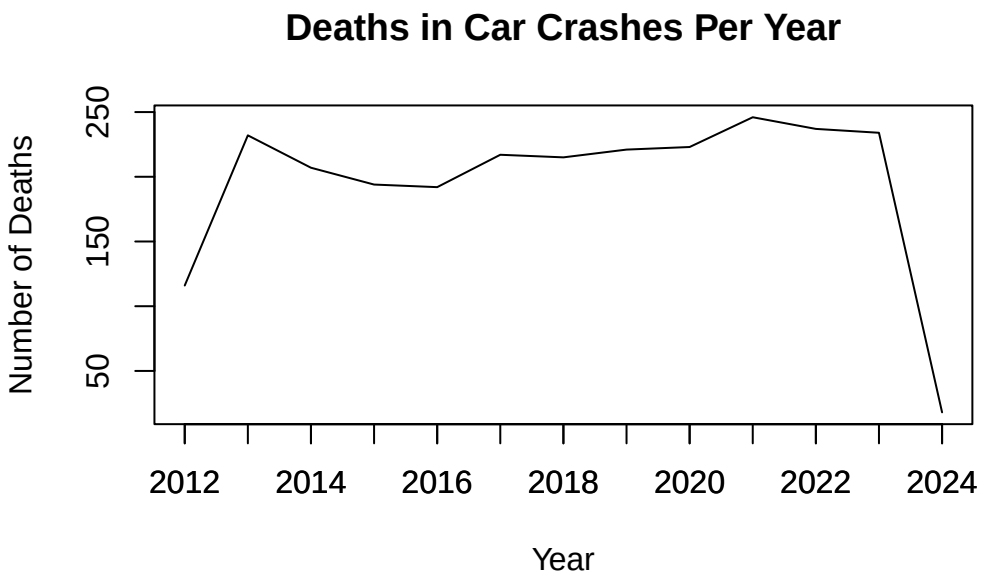


Interpretation

The histogram of car crashes by time of day shows that the most car crashes occur around the hours of 15:00 and 18:00. This is likely due to the fact that these are the hours when people are getting off work and are driving home. The least amount of car crashes occur around the hours of 3:00 and 4:00, which is likely due to the fact that these are the hours when people are sleeping and there is less traffic on the road. This graph gives us a better understanding of when car crashes are most likely to occur. An important consideration that is that traffic patterns may vary seasonally and by day of the week, which could affect the number of car crashes.

Plot 3: Line Graph of Deaths in Car Crashes Per Year

```
# Create a table of the number of deaths per year
deaths_per_year <- aggregate(cleaned_data$NUMBER.OF.PERSONS.KILLED > 0 ~ format(cleaned_data$CRASH.DATE

# Rename columns for clarity
names(deaths_per_year) <- c("Year", "Deaths")

# Convert 'Year' from character to numeric (if necessary)
deaths_per_year$Year <- as.numeric(deaths_per_year$Year)

# Plotting
with(deaths_per_year, {
  plot(Year, Deaths, type = "l",
       xlab = "Year",
       ylab = "Number of Deaths",
       main = "Deaths in Car Crashes Per Year")
  axis(1, at = seq(min(Year), max(Year), by = 1), las = 1)
})
```



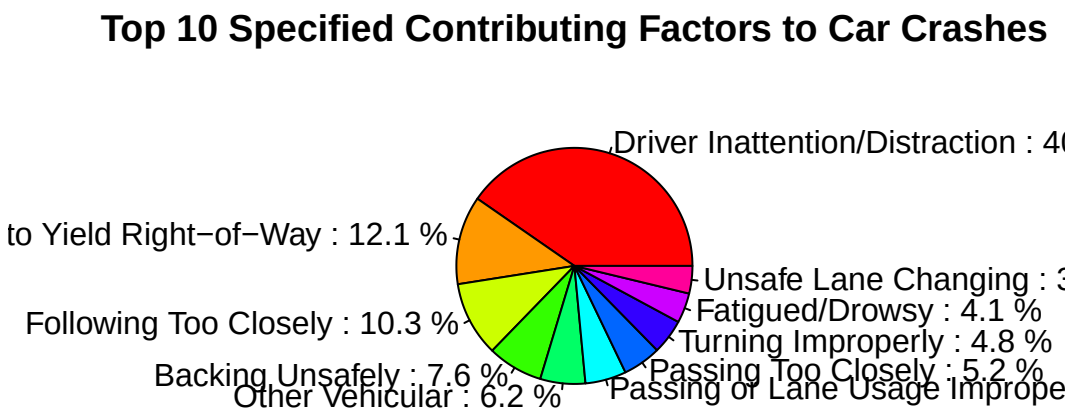
Interpretation

The line graph of deaths in car crashes per year shows that the number of deaths in car crashes seems to stay relatively from 2013 to 2023. The reason for the lower numbers of deaths in 2012 and 2024 are likely due to the fact that the data for each year is incomplete. This graph gives us a better understanding of the number of deaths in car crashes per year. However, a consideration that is not specified is the population of each borough and the amount of traffic through them.

Plot 4: Pie Chart of Contributing Factors to Car Crashes

```
# Create a table of contributing factors excluding 'Unspecified'
factors <- table(cleaned_data$CONTRIBUTING.FACTOR.VEHICLE.1)
factors <- factors[names(factors) != "Unspecified" & names(factors) != ""]

# Sort the table in descending order to get top contributing factors
factors <- sort(factors, decreasing = TRUE)

# Ensure only the top 10 specified factors are considered
top_factors <- factors[1:10]

# Calculate percentages
total_counts = sum(top_factors)
percentages = round((top_factors / total_counts) * 100, 1) # Round to 1 decimal place

# Create labels that include both factor names and percentages
labels_with_percents <- paste(names(top_factors), ":", percentages, "%")

# Create a pie chart for the top 10 specified contributing factors with percentages
pie(top_factors,
    labels = labels_with_percents,
```

```
    main = "Top 10 Specified Contributing Factors to Car Crashes",
    col = rainbow(length(top_factors)))
```



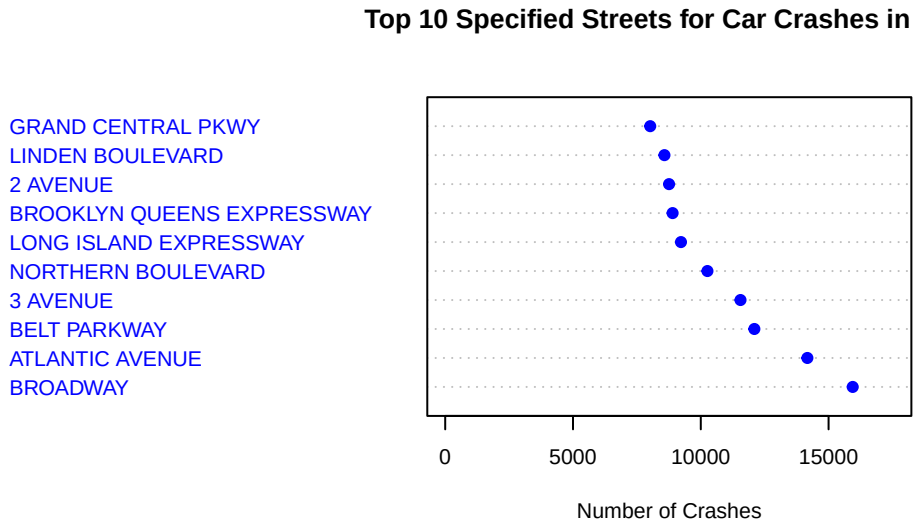
Interpretation

Disregarding the unspecified contributing factors, the top 10 contributing factors to car crashes are: Driver Inattention/Distraction, Failure to Yield Right-of-Way, Following Too Closely, Fatigued/Drowsy, Backing Unsafely, Other Vehicular, Turning Improperly, Passing Too Closely, and Passing or Lane Usage Improper. The pie chart shows that the top contributing factor to car crashes is Driver Inattention/Distraction, which accounts for 40.4% of the observed car crashes.

Plot 5: Top Specified Streets for Car Crashes in NYC

```
# Aggregating crash data by specified street name (excluding 'UNKNOWN')
street_crashes <- table(cleaned_data$ON.STREET.NAME)
street_crashes <- street_crashes[names(street_crashes) != "UNKNOWN"]
street_crashes <- sort(street_crashes, decreasing = TRUE)

# Selecting the top 10 streets with the most crashes
top_streets <- head(street_crashes, 10)

# Convert the table to a numeric vector
top_streets_counts <- as.numeric(top_streets)

# Use the names attribute of the table as labels for the dot plot
top_streets_names <- names(top_streets)

# Creating the dot plot with labels
dotchart(top_streets_counts, labels = top_streets_names, cex = 0.7,
```

```
        main = "Top 10 Specified Streets for Car Crashes in NYC",
        xlab = "Number of Crashes",
        pch = 19,
        col = "blue",
        xlim = c(0, max(top_streets_counts) * 1.1))
```

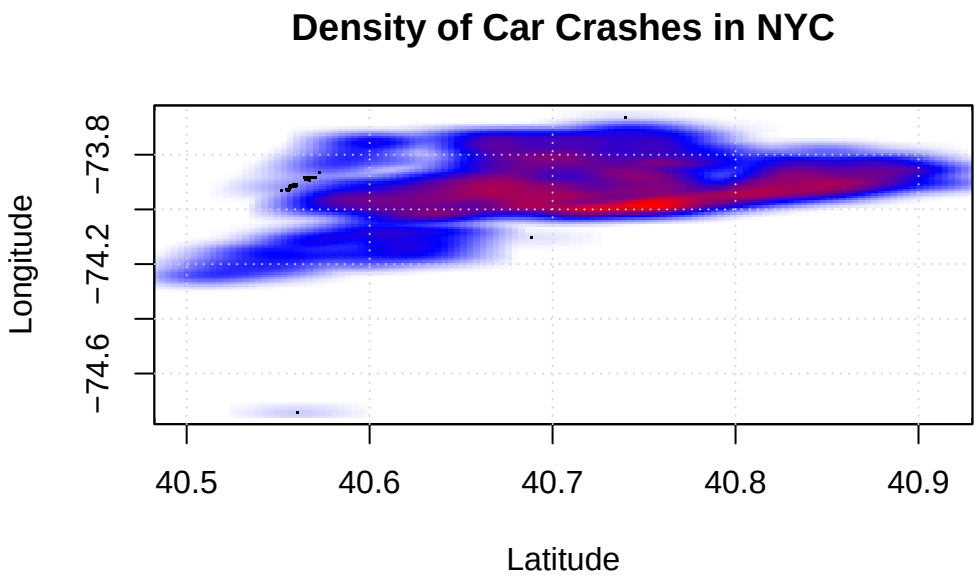


Interpretation

Of the observations where the street name was specified, the top 10 streets with the most car crashes are: BROADWAY, ATLANTIC AVENUE, BELT PARKWAY, 3 AVENUE, NORTHERN BOULEVARD, LONG ISLAND EXPRESSWAY, BROOKLYN QUEENS EXPRESSWAY, 2 AVENUE, LINDEN BOULEVARD, and GRAND CENTRAL PKWY. The dot plot shows that BROADWAY has the most car crashes, with over 15,000 crashes. This graph gives us a better understanding of which streets are more likely to result in car crashes. This information could be used to identify streets that require additional safety measures.

Plot 6: Density of Car Crashes in NYC

```
# Create a heat mapping of car crashes by location
smoothScatter(cleaned_data$LONGITUDE ~ cleaned_data$LATITUDE,
              main = "Density of Car Crashes in NYC",
              xlab = "Latitude",
              ylab = "Longitude",
              colramp=colorRampPalette(c("white", "blue", "red")))

# Add grid lines
grid()
```



Interpretation

The heat map of car crashes in NYC shows that the highest density of car crashes occurs in the center of NYC. This is likely due to the fact that the center of NYC has the most traffic and the most people. The lowest density of car crashes occurs in the outskirts of NYC. This aligns with intuition, as the centers of cities are typically more crowded and have more traffic. This graph gives us a better understanding of where car crashes are most likely to occur in NYC.

Statistical Modeling: Simple Linear Regression

```
# Create a simple linear regression model to predict the number of persons injured
model <- lm(NUMBER.OF.PERSONS.INJURED ~ CRASH.TIME + BOROUGH
            + CONTRIBUTING.FACTOR.VEHICLE.1 + CONTRIBUTING.FACTOR.VEHICLE.2
            + NUMBER.OF.PERSONS.KILLED + NUMBER.OF.PEDESTRIANS.INJURED
            + NUMBER.OF.PEDESTRIANS.KILLED + NUMBER.OF.CYCLIST.INJURED
            + NUMBER.OF.CYCLIST.KILLED + NUMBER.OF.MOTORIST.INJURED
            + NUMBER.OF.MOTORIST.KILLED, data = cleaned_data)

# Display the summary of the model
summary <- summary(model)
summary$r.squared
```

```
[1] 0.9891897
```

Interpretation

The simple linear regression model has an R-squared value of 0.9891897, which indicates that the model explains 98.9% of the variance in the number of persons injured. This suggests that the model is a good fit for the data. However, the model may be overfitting the data, as it includes many predictors. A more

parsimonious model may be more appropriate. Additionally, the model may not be generalizable to other datasets, as it is based on a specific time period and location.

Conclusion

The preliminary analysis of the NYC motor vehicle collision dataset has provided valuable insights into the frequency and patterns of car crashes in NYC. The analysis has revealed that the borough of Brooklyn has the highest frequency of car crashes, and that the most common contributing factor to car crashes is driver inattention/distraction. The analysis has also shown that car crashes are most likely to occur between 15:00 and 18:00, and that the highest density of car crashes occurs in the center of NYC. The simple linear regression model has provided a good fit to the data, explaining 98.9% of the variance in the number of persons injured. However, the model may be overfitting the data and may not be generalizable to other datasets. Future work will involve refining the model and identifying additional factors that contribute to car crashes in NYC.