Predicting Motor Vehicle Collisions in New York City

by Alex Fung, Viswesh Krishnamurthy, Tony Lee, Patrick Osborne

```
#code below IS shown in final document
```

Abstract

Technological progress in the world has unarguably improved the quality of life for the average person in many ways. The age of the automobile has shaped the way in which work, play and live our lives. Roadways, buildings, cities and entire countries have been designed to accommodate motor vehicles. As automobile technology has advanced making cars faster and capable of more advanced maneuvers, so has our concern with the safety of these vehicles. Entire disciplines such as traffic management are devoted to optimizing numerous factors to ensure the safe and efficient movement of people and goods. As we move into the age of data, all stakeholders in the automobile industry must effectively collect and utilize the wealth of information available to better meet their goals if progress is to continue. In this project, we take the position of a law enforcement agency, the New York City Police Department, as they seek to best utilize their resources in the context of responding to traffic collisions in the city.

Background

At the end of 2017 in New York City, there were 1,923,041 cars registered to residents of the city. (https://nyc.streetsblog.org/2018/10/03/car-ownership-continues-to-rise-under-mayor-de-blasio/) This already-significant number does not include the heavy flow of vehicles of those who visit the city or are simply passing through. By contrast, the New York City Police Department (NYPD) budgets for a headcount of 35,822 uniformed officers (http://council.nyc.gov/budget/wp-content/uploads/sites/54/2017/03/056-NYPD.pdf - page 4), distributed across 77 police precincts (geographic divisions of the city). On-duty officers/traffic enforcement agents are allocated to each precinct to enforce traffic laws and handle emergency and administrative response to traffic incidents (such as collisions). NYC has been collecting traffic data, including specific data on vehicle collisions since 2014 to support "Vision Zero" , a traffic safety initiative which has the goal of eliminating traffic fatalities. (https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/data)

Objective

The objective of our analysis is to develop a supervised, prediction model using Machine Learning techniques and the CRISP-DM framework (cite textbook) on the available collision data to predict whether there will be a collision in a specified police precinct at a specified time. The intent of predicting this data is to inform the NYPD's optimal assignment of limited officers and resources across the 77 police precincts.

Data Analysis

The data set that supports this analysis is sourced from the NYC Open Data project. The title of the data set is "Motor Vehicle Collisions – Crashes". It contains entries for every collision recorded within New York City limits by NYPD agents beginning July 1st, 2012 up to the present day. There are approximately 1.65 million entries in the data set.

Data Dictionary

[NOTE ON EXCLUDED VARIABLES TO BE INCLUDED]

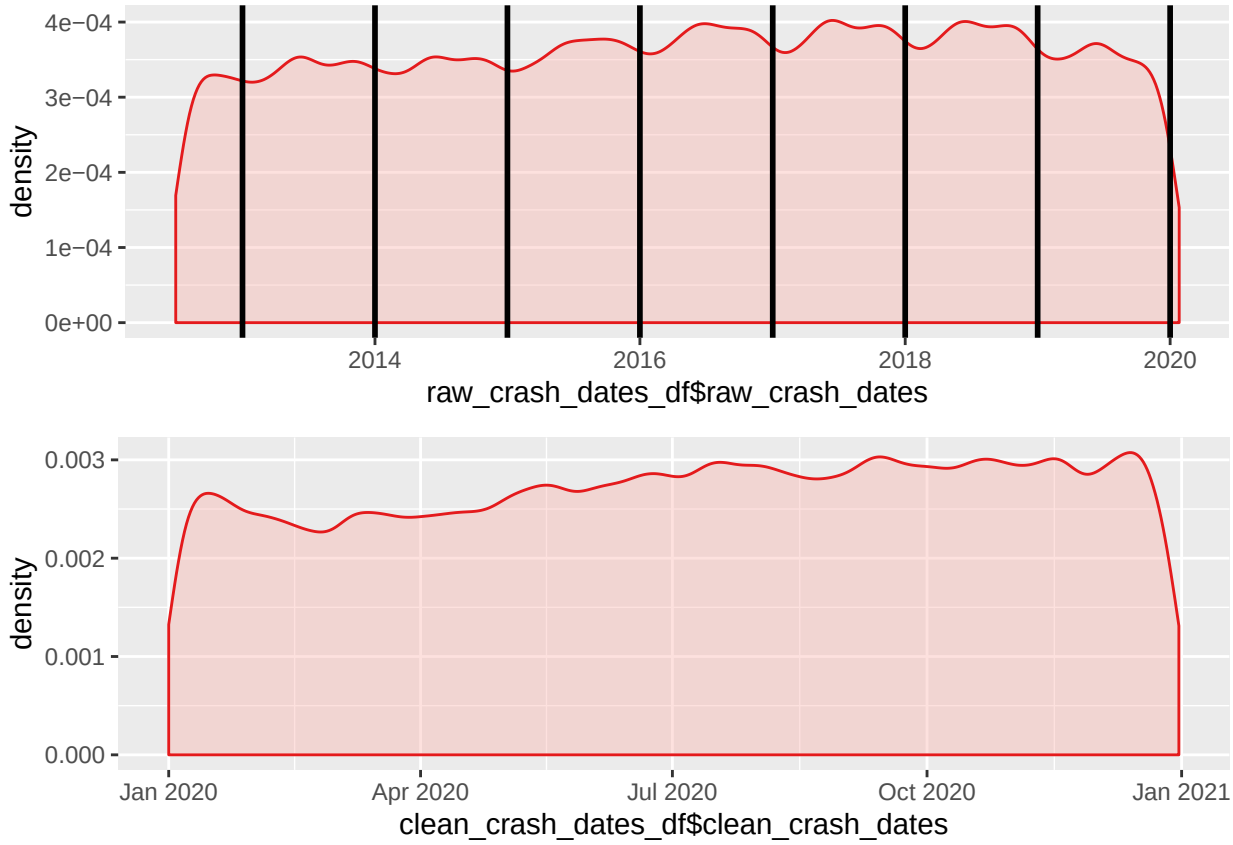
Data dictionary sourced from https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/data - "MVCollisionsDataDictionary_20190813_ERD.xlsx".

Initial Data Exploration and Cleaning

Based on what we know about the data set from the specifications at NYC Open Data and the data dictionary, we have decided to perform some initial cleaning steps.

Our analytics problem is to predict whether there will be a collision at a specific time (time including a time of the day, day of the year and calendar year). In this context, we will first look at the "CRASH.DATE" graph. Thinking about the scheduling of police resource, we assume that this happens in advance, on a hour-by-hour and day-by-day basis. We assume that resources are not scheduled on a year-by-year basis due to uncertainty in staffing, budget, etc. We therefore examine the data to see whether we should include the year at all. Including the year would treat the data set as a time-series, years ranging from 2012-2020. Alternatively, we could drop the year and group all occurrences on the same day in the same bin, possibly enhancing our prediction.

To decide, we plot the dates and look for trends. If trends repeat annually, we will drop the year as this trend will be preserved when we combine. If the trend does not repeat annually (extends over the whole range of dates) then we will not combine year as we will lose this information when dropping year.



Cleaned Data Exploration

[TEXT ABOUT EXPLORING THE DATA]

Table 1: Data Dictionary - Motor Vehicle Collisions – Crashes

Feature	Feature.Description
COLLISION_ID	Unique record code generated by system
ACCIDENT_DATE	Occurrence date of collision
ACCIDENT_TIME	Occurrence time of collision
BOROUGH	Borough where collision occurred
ZIP CODE	Postal code of incident occurrence
LATITUDE	Latitude coordinate for Global Coordinate System, WGS 1984
LONGITUDE	Longitude coordinate for Global Coordinate System, WGS 1984
LOCATION	Latitude , Longitude pair
ON STREET NAME	Street on which the collision occurred
CROSS STREET NAME	Nearest cross street to the collision
OFF STREET NAME	Street address if known
NUMBER OF PERSONS INJURED	Number of persons injured
NUMBER OF PERSONS KILLED	Number of persons killed
NUMBER OF PEDESTRIANS INJURED	Number of pedestrians injured
NUMBER OF PEDESTRIANS KILLED	Number of pedestrians killed
NUMBER OF CYCLIST INJURED	Number of cyclists injured
NUMBER OF CYCLIST KILLED	Number of cyclists killed
NUMBER OF MOTORIST INJURED	Number of vehicle occupants injured
NUMBER OF MOTORIST KILLED	Number of vehicle occupants killed
CONTRIBUTING FACTOR VEHICLE 1	Factors contributing to the collision for designated vehicle
CONTRIBUTING FACTOR VEHICLE 2	Factors contributing to the collision for designated vehicle
CONTRIBUTING FACTOR VEHICLE 3	Factors contributing to the collision for designated vehicle
CONTRIBUTING FACTOR VEHICLE 4	Factors contributing to the collision for designated vehicle
CONTRIBUTING FACTOR VEHICLE 5	Factors contributing to the collision for designated vehicle
VEHICLE TYPE CODE 1	Type of vehicle based on the selected vehicle category
VEHICLE TYPE CODE 2	Type of vehicle based on the selected vehicle category
VEHICLE TYPE CODE 3	Type of vehicle based on the selected vehicle category
VEHICLE TYPE CODE 4	Type of vehicle based on the selected vehicle category
VEHICLE TYPE CODE 5	Type of vehicle based on the selected vehicle category

```
#crash_dates <- as.Date(raw_crashes_data$CRASH.DATE, "%m/%d/%Y")
#ggplot(raw_crashes_data, aes(x=CRASH.DATE))
```

DEFAULT R MARKDOWN CODE BELOW

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

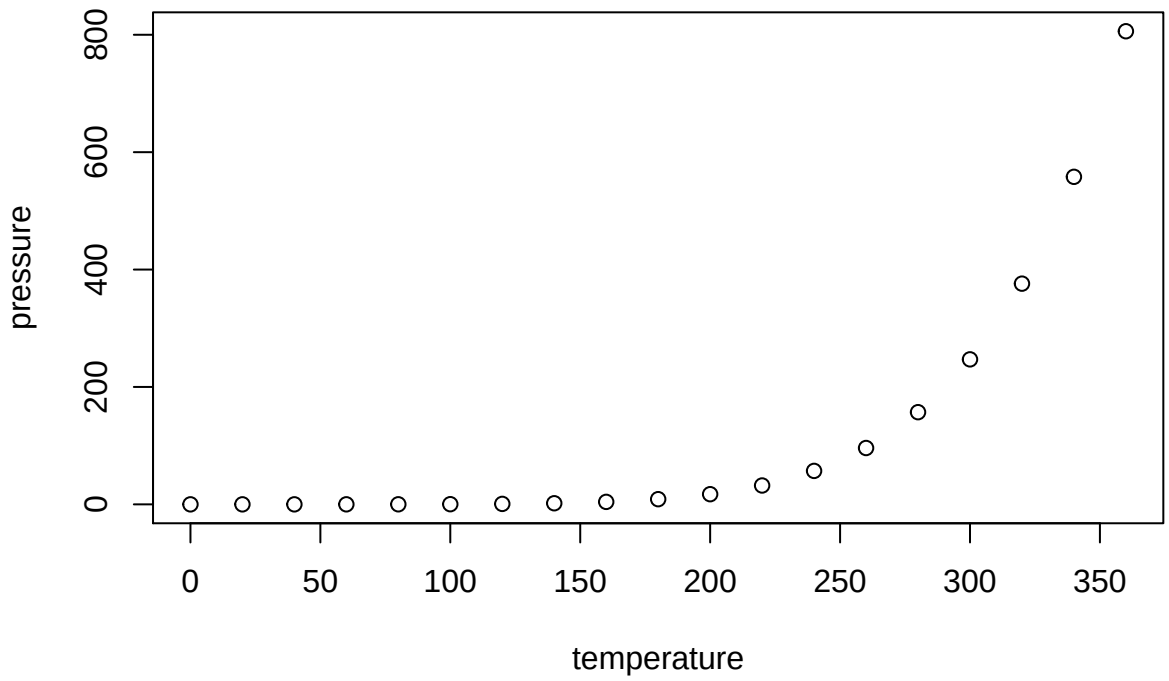
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
#>      speed           dist
#>  Min.   : 4.0   Min.   :  2.00
#>  1st Qu.:12.0   1st Qu.: 26.00
#>  Median :15.0   Median : 36.00
#>  Mean   :15.4   Mean   : 42.98
#>  3rd Qu.:19.0   3rd Qu.: 56.00
#>  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Document Style Attribution

This document was generated using a modified version of the "RJournal.sty" file provided by the The R Foundation at https://journal.r-project.org/submissions.html.

The document can be regenerated in RStudio by Knitting the provided "R Markdown-Group 10-Assignment 1.Rmd" file with the provided "RJournal.sty" file in the same directory.

Alex Fung

Viswesh Krishnamurthy

Tony Lee

Patrick Osborne