

Motor Vehicle Collisions in New York City:

Exploratory Data Analysis

Introduction to Data Science (DATS 6101) Sadaf Asrar

Section 10, Group 7 Nelson Foster

6 March 2018 Zhengzheng Yu

**Background**

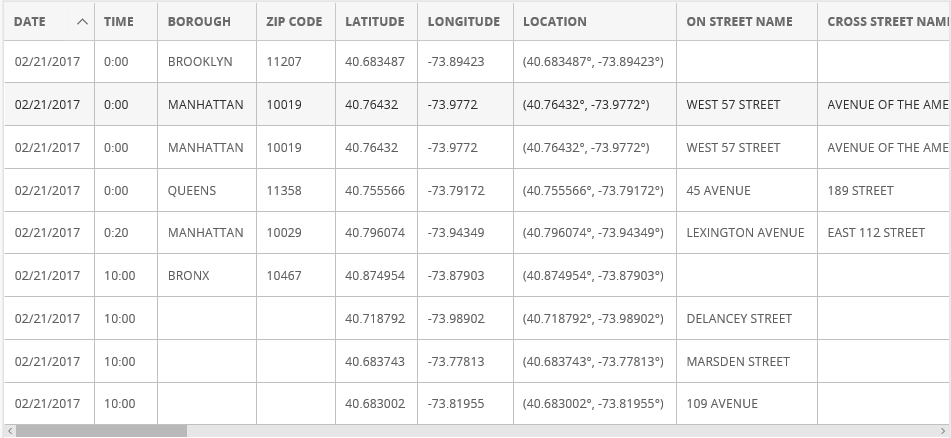
In order to find an interesting, relevant, and up-to-date dataset to conduct exploratory data analysis with, the group turned to [OpenData.gov](https://www.data.gov/open-gov/), a public repository of datasets collected by local, state, and federal government agencies. We found the [New York City Police Department (NYPD) Motor Vehicle Collisions](https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95/data) dataset to be particularly interesting, and decided to research it further.

Data aggregated in NYPD Motor Vehicle Collisions dataset is gathered from the details and metadata collected from NYPD motor vehicle collision reports from July 2012 to present. The data is refreshed weekly, and made available to the public via the [NYC Open Data Project.](https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95)

We began with this initial research-driven question: "What are the most dangerous areas in New York City for traffic collisions?"

**Exploratory Data Analysis and Survey of Existing Research**

The NYPD Motor Vehicle Collisions dataset is in a tabular format, and as of February 19, 2017, it consisted of over 984,000 observations of 29 distinct variables, such as DATE, TIME, ZIP CODE, LOCATION, STREET NAME, NUMBER OF PERSONS INJURED, NUMBER OF PERSONS KILLED, VEHICLE TYPE, etc.:



Despite the fairly substantial amount of observations in the dataset, a few limitations became apparent from the outset. Most notably, as illustrated in the table above, there were several blank fields. This is due to the fact that the quality and completeness of the data is totally reliant on how thorough the responding officer is in their report. Either due to the circumstances of the incident or a lack of diligence in fully completing the paperwork, there is generally a high potential for missing values in datasets sourced from human-logged, “self-report” data. In addition to missing data entries, the group identified several other pieces of information that would have been beneficial to facilitating more in-depth analyses.

While the dataset includes several categorical variables for the number of persons injured and number of persons killed, it does not contain any measure of the severity of the injury. In other words, while we can assess the number of people injured, we cannot identify the extent of the injury. Similarly, we cannot identify if the individuals killed died on the spot or at a later time. A severity index would better enhance our ability to answer the question. That being said, severity of injury is likely recorded in police records and follow-up medical reports. Such data may not be publicly available as it is Protected Health Information (PHI) and as such, it may be difficult to link it at the traffic accident level.

Additionally, variables like proximity to the nearest hospital and traffic conditions could be confounding factors to how severe the injury from the collision was or whether any of the injured party died waiting for medical attention. In other words, the relationship of injury and death from collision versus time could be further explored by controlling for proximity to the nearest hospital and traffic conditions at the time of the collision. Proximity to a hospital might affect the severity of injury or even death, as close proximity could mean that paramedics could not only get at the scene of the collision faster, but also take the injured individuals to the hospital quicker. Similarly, the state of traffic congestion at the time of collision could not only influence likelihood of accidents, but also the response time for paramedics to provide medical support, which in turn can influence the extent of injuries suffered.

While the proximity to the nearest hospital is not a variable included in this dataset, it could be calculated using the addresses of all hospitals in NYC and then added to this dataset, as the dataset include the precise location of the collision. Similarly, the state of traffic conditions at the time of the collision are not included in the dataset. Data on traffic conditions could be researched and included in this dataset. While it is not answerable with the current dataset, we would also like to explore the role if any, proximity to a hospital, and traffic conditions at the time of collision, effect the number of people killed in motor vehicle collisions at different times of the day in New York City

All three factors mentioned in this subsection, namely, severity index, proximity to hospital, and traffic conditions, can help us better answer our research question by using inferential statistics as we could potentially control for these factors in a multiple linear regression model where the dependent variable is number of people injured or killed in traffic collisions and the independent variable of interest is location of collision.

Despite these limitations, a fairly substantive amount of analyses and research has been conducted using the NYPD Motor Vehicle Collisions dataset.

As of February 19, 2017, there were over 50,000 views of the dataset, resulting in over 16,000 downloads. Several research projects and analyses have been completed and published to NYC OpenData page, as well as other sites and blogs. A summary of these projects and analyses is below:

An interactive [visualization](https://data.cityofnewyork.us/NYC-BigApps/NYPD-Motor-Vehicle-Collisions/m666-sf2m) was developed and published on the NYC OpenData page, summarizing several key variables in the dataset. Specifically, the visualization included:

* A run chart of the number collisions by date, July 2012 through present
* A bar chart of the number of collisions by borough
* A map by the “Locations” variable
* A map by Zip Code
* Several searchable analyses where the user can search by
  + Zip code
  + “Off Street”
  + “On Street”
* Several Bar charts summarizing several variables:
  + Number of persons injured
  + Number of persons killed
  + Number of cyclists injured
  + Number of cyclists killed

A [heat map](https://data.cityofnewyork.us/Public-Safety/Heat-map/9dh2-x3up) was developed and published to the NYC OpenData page, created using the details of the collisions dataset, allowing the user to look at the concentration of collisions by location.

An [interactive filter analysis](https://data.cityofnewyork.us/Public-Safety/2015_format/cjfk-j22b) was developed and published to the NYC OpenData page, created using the details of the dataset, allowing the user to filter, pivot, and use conditional formatting to drill down to specific subsets and intersections of the data.

[NYPD Motor Vehicle Collisions Research Part#1](https://rstudio-pubs-static.s3.amazonaws.com/217730_0625ca1f20b34fe983efe07f786a73ee.html)

Asenscio (2008) utilized the Collisions dataset to develop several analyses in a report including

* Collisions by seasons
* Collisions by borough
* Accidents per day of week
* Highest contributing factors
* Contributing factors per borough, by season, by number of persons killed
* Collisions by vehicle type
* “Most Dangerous Boroughs”
* Most dangerous intersections for pedestrians
* Reasons why most collisions occur

[New York City Vehicle Collision Analysis](http://tyokogawa.github.io/blog/NYC_vehicle_collision/)

Tyokogawa (2015) analyzed the collisions dataset by utilizing a clustering algorithm to evaluate the correlations between intersections and weather conditions, identifying the most dangerous locations susceptible to foul weather such as heavy rain or snow.

[NYPD Motor Vehicle Collisions Data](https://cloud.google.com/bigquery/public-data/nypd-mv-collisions)

Google Cloud Platform had a post that analyzed the NYPD Collisions dataset by evaluating the most common factor for motor vehicle collisions, and the most dangerous streets for motor vehicle collisions.

[NYC Fatal Motor Vehicle Collisions hotspots](http://www.residentmar.io/2016/03/19/nyc-motor-vehicle-collisions.html)

Bilogur (2016) utilized the NYPD Collisions dataset to develop a Tableau Visualization highlighting the “inherently dangerous” intersections for motor vehicle collisions within New York City.

[Visualizing Traffic Accidents on Staten Island](http://research.prattsils.org/blog/coursework/information-visualization/visualizing-traffic-accidents-on-staten-island/)

Using the NYPD Collision dataset, Pratt Institute School of Information developed several dashboards evaluating the following factors relating to collisions on Staten Island:

* Top ten causes of accidents
* Accident frequency by time of day
* Accident frequency by time of day per (Staten Island) neighborhood
* Traffic accidents per neighborhood

**Refining the Research Question**

In reviewing several existing analyses, we discovered that our initial research question, "What are the most dangerous areas in New York City for traffic collisions?" had already been approached. Several similar yet more specific questions had already been analyzed, such as what are the most dangerous seasons, boroughs, days of the week, weather conditions, and the leading contributing factors resulting in collisions. However, one variable, time of the collision, only seemed to be analyzed for one borough, Staten Island (Pratt Institute, 2015). The group decided to expand on this by analyzing the relationship of time injuries and death in collisions across New York City, refining our research question to “What is the most dangerous time for motor vehicle collisions in New York City?”

More specifically, we aimed to see if there was a relationship between time of day and injuries and death from traffic collisions, and then identify which time was most “dangerous” for collisions that result in injury or death.

**Setting Expectations**

We expected to have a tabular dataset with several variables and observations relating to motor vehicle collisions in New York City. In analyzing the dataset, we expected there to be some “NA” values due to variances in how diligent Police or dispatch personnel are in being thorough in their reports. We also expected there to be several variables that could be used in exploratory data analysis to compare and draw several correlations from.

**Establishing a S.M.A.R.T. Question**

Is it Specific?

Our research question is specific in that it looks at the time variable to identify the most dangerous times of day for motor vehicle collisions in New York City, in particular those resulting in injury of death for motorists, vehicle occupants, pedestrians, and cyclists.

Is it Measurable?

Due to the high level of granularity in the metadata aggregated by the NYPD Motor Vehicle Collisions dataset, we are afforded 29 unique variables to evaluate nearly a million observations over the course of five years. This allows the time dimension to be applied to several other variables for measurable results.

Is it Achievable?

The dataset is a relatively small tabular formatted dataset, making analyzing the dataset is quite achievable so with conventional resources (RStudio, laptop computer). Utilizing several packages in R to convert various data elements into quantifiable measures makes this achievable.

Is it Relevant?

Our research question expands on existing research by conducting more extensive analysis on the time variable and how it correlates to other variables relating to injury or death due to being involved in a motor vehicle collision.

Is it Time-Oriented?

Our research question is time oriented in that it is an aggregation of motor vehicle collisions from July 2012 to February 2017. Though the dataset is updated weekly on New York City’s OpenData website, we took a snapshot of the data of February 2017 in order to have a finite dataset. Further, our analysis specifically focuses on the time variable, i.e., the time of day a collision occurred.

To sketch an answer to our refined research question of “What is the most dangerous time for motor vehicle collisions in New York City,” we took the steps described below.

**Defining “dangerous”**

As stated earlier, while the dataset includes variables for number of persons injured as well as number of persons killed in the collision, there is no composite variable for all injuries or death. Additionally, the dataset does not include an index for severity of injury, as injury could range from minor cuts and bruises to brain and spinal cord damage. To define “dangerous”, we simply combined the two variables that captured the number of persons injured and number of persons killed in the collision to create a composite variable for number of persons injured and killed in the collision. We attempted to answer our research questions using this variable as a metric for “dangerous” collision.

Exploring the number of people injured and killed for the 24 hours of the day

We generated a table that identified the number of collisions for each of the 24 hours, the number of people injured and killed in each hour, and the mean number of people injured and killed in each hour.

|  |  |  |  |
| --- | --- | --- | --- |
| Hours | Number of collisions in each hour | Number of people injured and killed in each hour | Mean number of people injured and killed in each hour |
| 0:00 | 24,193 | 7,207 | 0.2978961 |
| 1:00 | 15,467 | 4,991 | 0.322687 |
| 2:00 | 11,912 | 4,137 | 0.3472968 |
| 3:00 | 10,087 | 3,568 | 0.3537226 |
| 4:00 | 11,806 | 4,505 | 0.3815856 |
| 5:00 | 12,750 | 4,270 | 0.334902 |
| 6:00 | 19,887 | 5,838 | 0.2935586 |
| 7:00 | 26,632 | 7,735 | 0.2904401 |
| 8:00 | 55,360 | 12,988 | 0.2346098 |
| 9:00 | 55,033 | 11,074 | 0.2012247 |
| 10:00 | 50,599 | 9,958 | 0.1968023 |
| 11:00 | 52,253 | 10,851 | 0.2076627 |
| 12:00 | 55,290 | 11,814 | 0.2136734 |
| 13:00 | 58,351 | 12,866 | 0.2204932 |
| 14:00 | 67,056 | 15,294 | 0.228078 |
| 15:00 | 60,454 | 15,266 | 0.2525226 |
| 16:00 | 73,434 | 17,344 | 0.2361849 |
| 17:00 | 71,026 | 17,743 | 0.2498099 |
| 18:00 | 62,804 | 17,000 | 0.2706834 |
| 19:00 | 51,143 | 14,311 | 0.2798232 |
| 20:00 | 42,624 | 12,965 | 0.3041714 |
| 21:00 | 34,886 | 10,806 | 0.3097518 |
| 22:00 | 31,366 | 10,104 | 0.3221322 |
| 23:00 | 25,452 | 8,660 | 0.3402483 |



**Correlation between number of people injured and killed against time**

The correlation coefficient is 0.0212, which is very weak. The correlation coefficient tells us the relationship between number of people killed and injured versus time in 24 hours, i.e., do the number of injuries and deaths in motor vehicle collisions increase when the time (measured in 24-hour scale) increase. While the correlation found is weak, it actually may not be the best way to test this relationship.

**Defining “Time of Day”**

To test this relationship differently, we decided to split up the day in six different times of day, i.e., late night (0:00-03:59), early morning (04:00-07:59), morning (08:00-11:59), afternoon (12:00-15:59), evening (16:00-19:59), and night (20:00-23:59). For the rest of the analysis we use this time of day variable as a measure of time.

**Are the number of people injured and killed different for different time of day?**

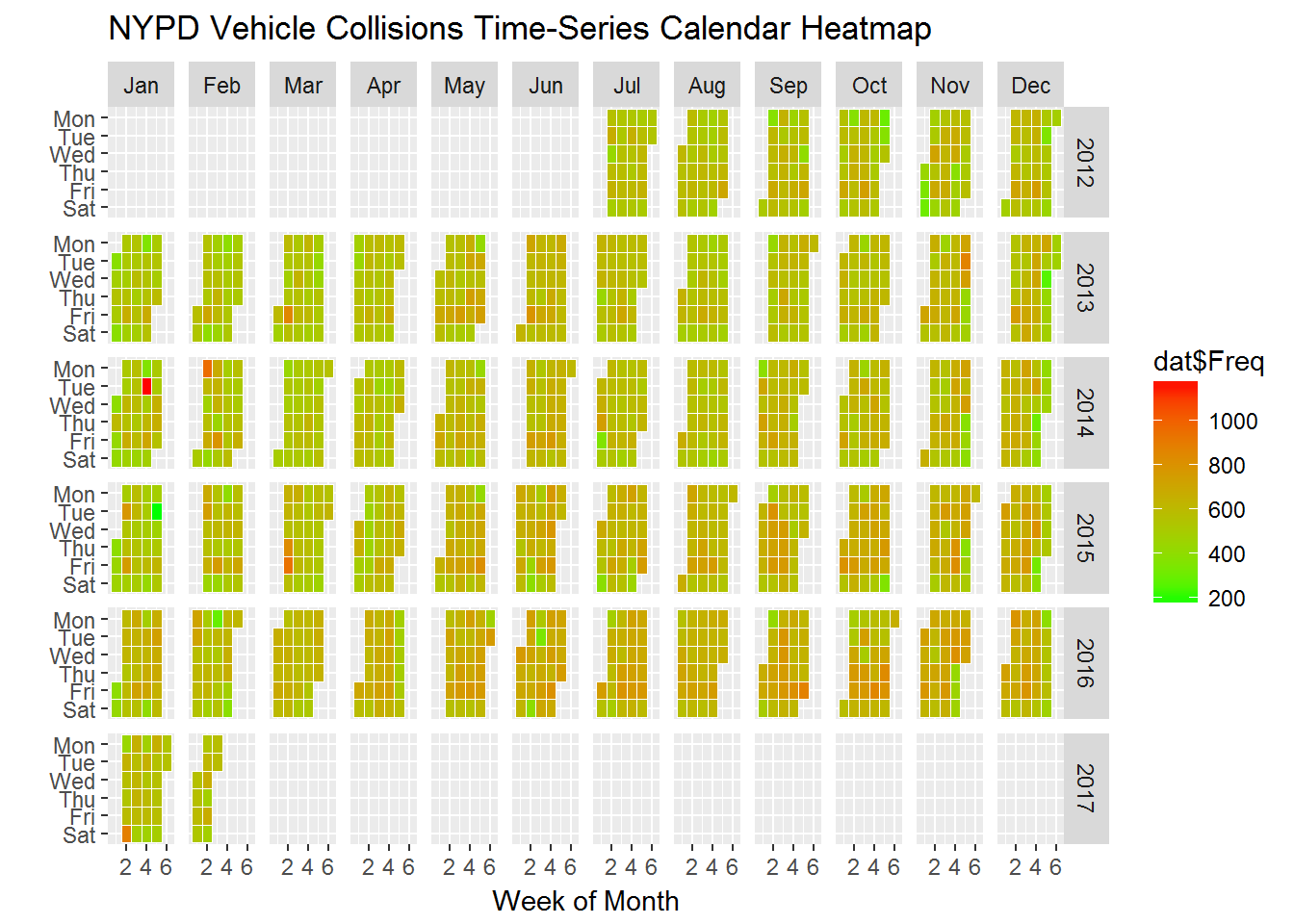
To answer this question, we first conducted a one-way ANOVA, which tests whether the number of people injured and killed are statistically significantly different for the six different times of day.

The one-way ANOVA result produces a p-value of 2.2e-16, which is below the critical value of 0.05, which means, yes, the number of people injured and killed are different for the six different times of day. However, we do not know which time of day has most injuries and death or which time of day has the highest mean injuries or death.

**Which time of day were the largest of people injured and killed?**

|  |  |
| --- | --- |
| Time of day | Number of people injured or killed |
| Afternoon | 55,240 |
| Early morning | 22,348 |
| Evening | 66,398 |
| Late night | 19,903 |
| Morning | 44,871 |
| Night | 42,535 |

In addition, a heat map visualization of collisions per time of day and day of week was created:



According to these analyses, the time of day with the largest number of people injured or killed is the evening, when 66,398 people were injured and killed.

**Which time had largest mean number of people injured and killed?**

|  |  |
| --- | --- |
| Time of day | Mean number of people injured or killed |
| Afternoon | 0.2290681 |
| Early morning | 0.3144284 |
| Evening | 0.2569512 |
| Late night | 0.3227915 |
| Morning | 0.2104199 |
| Night | 0.3166503 |

According to this analysis, the late night time period had the largest mean number of deaths and injuries (0.3227915).

**Are the number of people injured and killed at late night different from all other times of day?**

To answer this question, we conducted t-tests of the number of injuries and death over the binary variable of whether the collision was at late night versus non-late night. The t test of number of people injured and killed during late night vs non-late night is statistically significant at the 5% level, which means that, yes, the number of people injured or killed in traffic collisions at late night are different from collisions at all other times of the day.

**Are the number of people injured and killed at late night different from EACH of the other times of day?**

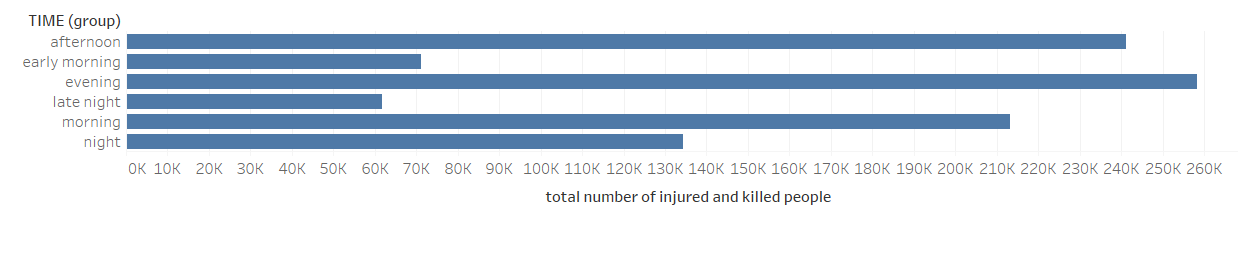
To answer this question, we conducted t tests of the number of people injured or killed in collisions at late night with each of the other 5 times of the day.

The t-test results show that the mean number of people injured and killed at late night is statistically significantly different from each of the other times of day. This means that the mean number of people injured and killed at late night is different from each of the other times of day.

**So, what is the most dangerous time of day?**

Based on the exploratory data analysis we performed, if we are using the raw number or count, it is evening, as evening has the highest number of people who were injured and killed in motor vehicle collisions. However, if we factor in that the number of collisions differs for different times of day, then the most dangerous time for traffic collisions is late night, as it has highest mean number of injuries and death per motor vehicle collision. The number of injuries and death per motor vehicle collision at late night was statistically significantly different for each of the other times of day, as illustrated in the table and visualizations below:

|  |  |  |  |
| --- | --- | --- | --- |
| Time of day | Number of collisions for each time of day | Number of people injured or killed | Mean number of people injured or killed |
| Afternoon | 241,151 | 55,240 | 0.2290681 |
| Early morning | 71,075 | 22,348 | 0.3144284 |
| Evening | 258,407 | **66,398** | 0.2569512 |
| Late night | 61,659 | 19,903 | **0.3227915** |
| Morning | 213,245 | 44,871 | 0.2104199 |
| Night | 134,328 | 42,535 | 0.3166503 |





**Works Cited**

1. Ascencio, Glenda (2016). NYPD Motor Vehicle Collisions Research Part #1. Retrieved from <https://rstudio-pubs-static.s3.amazonaws.com/217730_0625ca1f20b34fe983efe07f786a73ee.html>
2. Bilogur, Aleksey (2016). NYC Fatal Motor Vehicle Collisions Hotspots. Retrieved from <http://www.residentmar.io/2016/03/19/nyc-motor-vehicle-collisions.html>
3. Pratt Institute, (2015). Visualizing Traffic Accidents on Staten Island. Retrieved from: <http://research.prattsils.org/blog/coursework/information-visualization/visualizing-traffic-accidents-on-staten-island/>
4. Google Big Query API, NYPD Motor Vehicle Collisions data. (2016). Retrieved from <https://cloud.google.com/bigquery/public-data/nypd-mv-collisions>
5. New York City Police Department (NYPD), (2008). NYPD Motor Vehicle collisions [Data file]. <https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95/data>
6. Tyokogawa, Tohei(2015). New York City Vehicle Collision Analysis. Retrieved from <http://tyokogawa.github.io/blog/NYC_vehicle_collision/>