Georgia Tech

MGT6203 – Data Analytics in Business
Course Project
Summer 2023

Final Report

Traffic Collision Analysis

Marco LoConte, Mohammed Al-Desouky, Majed AlOtaibi, Kangwa Ng'omalala, Sarah Hasanen

Georgia Institute of Technology
OMSA – MGT6203 Course Project

ABSTRACT

Road traffic crashes are a leading cause of death in many countries, especially those with people living in high-population density environments. In this project, due to data availability issues, we will be focusing on the city of New York City in the United States.

This project is aimed at testing for the existence of a relationship between several predictors and the occurrence of death/injury. Such findings can help our client, MGTTA, to adjust insurance premiums accordingly.

The team has been successful in proving the existence of a relationship between:

1. Driver attributes (sex, age group, being licensed)
2. Vehicle attributes (instate/out-of-state license, safety category, weight, type)
3. Weather (temperature, precipitation, wind speed)
4. Time components (time of day, season)
5. Population density of crash location

and the probability of injury/death. To successfully achieve this result, extensive data manipulation had to take place.

DATASETS

To get a complete picture of the circumstances surrounding a car crash on one hand, and not to overwhelm our model with large number of predictors, the following data sets were used:

1. Crashes
2. Vehicles
3. Persons
4. Zip Codes
5. Weather
6. Solar

Crashes [3], Vehicles [4], and *Persons [5]* datasets contain attributes of crashes and those involved. *Weather* dataset [9] contains weather conditions obtained at the coordinate level for every collision in *Crashes* dataset. *Zip Codes* [6] dataset is a spatial (geolocation) dataset. It was used to obtain correct values of zip codes or replace missing ones. *Solar* [8] dataset was used to find sunrise/sunset times. Datasets cover the period from 2016-2021.

DATA WRANGLING AND PREPARATION

Crashes, *vehicles*, and *persons* datasets required extensive pre-processing. They have millions of records that are a direct raw input by police officers at the various crash scenes. The following steps were applied to clean the data up:

1. Removed all crashes without vehicles or persons or coordinates (latitude, longitude).
2. Kept only crashes that occurred between two vehicles.
3. Kept only records for which there is a single driver.
4. For categorical variables with many categories, only the top n categories were kept. Everything else was put under "other" category.
5. Remove predictors with missing values.
6. `DRIVER_AGE` predictors were limited to between 10 and 110 years.
7. Drivers with unidentified `SEX` were removed.
8. `VEHICLE_TYPE_CODE` was manipulated to correct spelling mistakes and reduce the number of types without altering the data.
9. `ZIPCODE` was obtained by intersecting *Zip Codes* geolocational data set with crash coordinates.

Some variables were modified to make analysis more computationally efficient:

1. `VEHICLE_MAKE` was reduced to only the top 20 makes. Every other make was grouped into one category called "other".

2. VEHICLE_CATEGORY was reduced to the top 14 categories. Others are grouped under "other" category.
3. CRASH_DATETIME was broken into YEAR, MONTH, DAY, HOUR variables. Some of these were used categorically.
4. TIME_OF_DAY was introduced by breaking daytime into four periods:
    - DAWN: To indicate the period of sunrise time ±0.5 hours.
    - DUSK: To indicate the period of sunset time ±0.5 hours.
    - DAY: The period between sunrise and sunset.
    - NIGHT: The period after sunset until sunrise.
5. CRASH_SEASON to tell which season the crash took place in (SUMMER, WINTER, FALL, SPRING).

A few calculated fields were created to support the analysis:

1. VEHICLE_WEIGHT was created based on VEHICLE_CATEGORY to calculate weight differential (VEHICLE_WEIGHT_DIFFERENTIAL).
2. DRIVER_AGE_GROUP was created to allocate driver ages into bins of 10 years each.
3. DENSITY is the population density in zip code area. Created by dividing POPULATION by AREA.

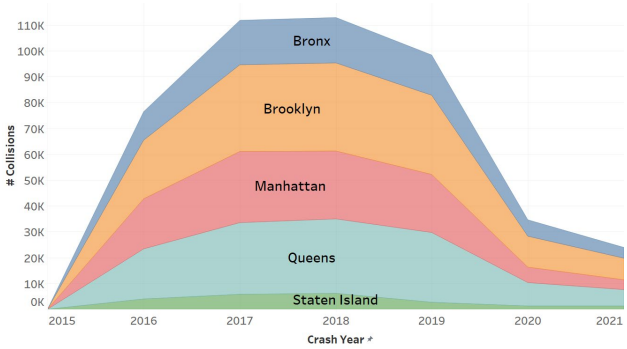

Figure 1: Yearly Collision Distribution per Borough

Wrangling took place using both Microsoft SQL Server and Tableau Data Prep. Some preparations were embedded in the code. Nevertheless, most of the work occurred in SQL Server.

HYPOTHESIS

We anticipated the existence of a strong correlation between the occurrence of injury/death and weather conditions (temperature, visibility, wind, precipitation). It was also anticipated that vehicle type is correlated with the response (injury/death occurrence). Human-related factors were also expected to have influence (age, sex).

Another predictor we expected to have a great effect on the response is weight differential. We expected that when bigger cars collide with small ones, injuries should increase.

This project brought interesting findings that we really did not expect, but data does not lie!

MODELING

We tried to make our models simple. Simple models allow for good and understandable interpretations, and they present a good starting point for going forward with more detailed in-depth modeling. Logistic regression was the natural choice.

It was decided to combine the occurrence of death and injury in one single variable IOD, which stands for Injury or Death. IOD will equal 1 if injury and/or death occurred, otherwise 0.

Car safety is a definite factor. After extensive thought, it was decided to use car model as a proxy for car safety. To clearly identify which car makes are safer than others, a logistic regression was used to identify which makes are correlated with IOD. Car makes that have high p-values are not correlated with IOD and are considered safe.

It was natural to think of automated variable selection techniques to utilize the best features we have. However, we needed to understand what the data is telling us. Therefore, we decided to follow a different route. We decided to follow our intuition and inspect features in the following feature subsets:

Table 1: Feature Subsets used in Modeling

Feature Set	# Terms	R Model Name
Sex of Drivers*	3	model_sex
Age of Drivers*, **	5	model_age
License Status*	3	model_lic
License Location*	3	model_instate
Car Safety*	3	model_car_safe
Density	1	model_density
Weather	3	model_weath
Time of Day	1	model_TOD
Vehicle Weights, Weight Diff.**	3	model_weight
Crash Season	1	model_season
The Hero Model*, **	23	model_fin

* Contains interaction term(s)
** Contains transformations (log and/or polynomial)

After completing analysis using the subsets above, we constructed what we thought was the best model. R was the ideal choice for modeling and analysis.

Key Variables

The choice of key variables came after modeling with subset features as indicated earlier. The following table contains the final list of features:

Table 2: Key Variables

Predictor	Type*	# Categories
DRIVER_SEX_1	C	2
DRIVER_SEX_2	C	2
DRIVER_AGE_1	N	-
DRIVER_AGE_2	N	-
DRIVER_LICENSE_STATUS_1**	C	3
DRIVER_LICENSE_STATUS_2**	C	3
DRIVER_LICENSE_JURISDICTION_1**	C	65
DRIVER_LICENSE_JURISDICTION_2**	C	65

Predictor	Type*	# Categories
VEHICLE_MAKE_1**	C	21
VEHICLE_MAKE_2**	C	21
POPULATION	N	-
AREA	N	-
FEELSLIKE	N	-
PRECIPROB	N	-
WINDSPEED	N	-
CRASH_TIME_OF_DAY	C	4
VEHICLE_WEIGHT_1	N	-
VEHICLE_WEIGHT_2	N	-
CRASH_SEASON	C	4

* N = Numerical, C = Categorical
** These variables were re-categorized before they were included in the model

Modeling Challenges

Size! Most of the datasets were huge and required extensive computing power. Tableau Data Prep proved to be lightning fast in dealing with large datasets. Combined with Microsoft SQL Server, we were able to counter the size challenge.

Wrangling was time-consuming and exhausting. Most of data sets had multiple errors in inputs due to human factor. There was no way around this. We had to invest all the effort possible to ensure having clean and modeling-ready dataset.

Getting weather data was not as easy as anticipated. National Oceanic and Atmospheric Administration (NOAA) provides amazing datasets [8], but they were unsuited for analysis. New York city is huge, and location-specific weather data was needed, which was unavailable in NOAA's dataset archive. The only way forward was to use a paid weather service [9] to conserve time.

The number of categories within categorical variables was huge and caused R to crash. The best solution was to group all categories with a small number of records into one category, usually called "other".

Wrangling and Modeling were iterative. Without these iterations, a solid conclusion was not to be expected.

Model Performance

We decided to add model performance, yet as an appendix so it does not clutter the report. The following table contains all models created and their performance metrics.

Table 3: Model Performance

Model	# Predictors Outside 95% CI	AUC
model_sex	0	0.521
model_age	0	0.548
model_lic	0	0.512
model_instate	0	0.523
model_car_safe	1	0.516
model_density	0	0.546
model_weath	0	0.524
model_TOD	0	0.532
model_weight	1	0.556
model_season	0	0.522
model_fin	**0**	**0.608**

All models performed nicely. The final model, model_fin, performed even more nicely.

RESULTS

This section will discuss results obtained either by looking at the data itself or by modeling. It is divided into subsections by feature group.

Vehicle Safety and Weight

Despite dramatic improvements in safety equipment in vehicles in the past 30 years (anti-locking brakes, air bags, and auto-braking systems), we could not identify a relationship between the age of a car and IOD. One possible explanation is that confounding variables make the analysis difficult; for example, older vehicles, which may be less safe, tend to be driven by older, more experienced drivers, who tend to be safer. Another possible explanation is that drivers adjust their behavior to the level of protection afforded by the vehicle. For example, a driver with auto-braking technology may give other drivers less space, thus mitigating the value of the protection.

One of the most important risk factors is travelling in an unprotected vehicle such as a motorcycle or a bicycle. In the data we reviewed, bicyclists and motorcyclists had a risk of injury in a collision three to five times greater when compared to a car or SUV, and a risk of death about 50 times greater per collision.

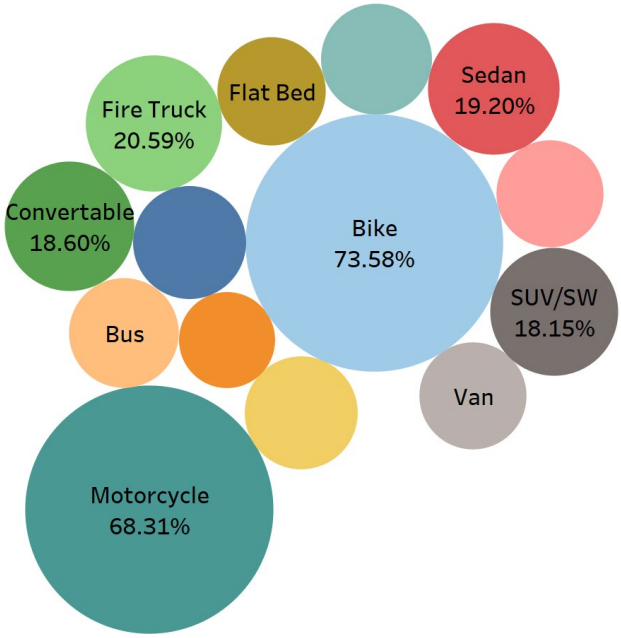

Figure 2: Average Injuries per Vehicle Category

We expected weight differential to be a significant predictor of injury. It seems reasonable that an incident where a large

vehicle collides with a small vehicle would have a greater risk of injury than a collision where two large vehicles or two small vehicles collide. This assumption, however, was not borne out by the data. This is especially interesting because the relationship between vehicle weight and injury risk is very clear.

A lot of thought (and advertising dollars) have gone into positioning auto brands as safer. While our study does indicate that certain brands are meaningfully associated with reduced accident severity, it is hard to determine if the effect is due to the brand itself. Are "safe" brands safer, or do safe drivers tend to choose "safe" brands for themselves? One potential way to answer this question would be to look at the accident severity of taxicabs, since taxicab drivers typically lease their vehicle on a day-to-day basis and do not get to choose the brand of vehicle that they drive.

Age Group

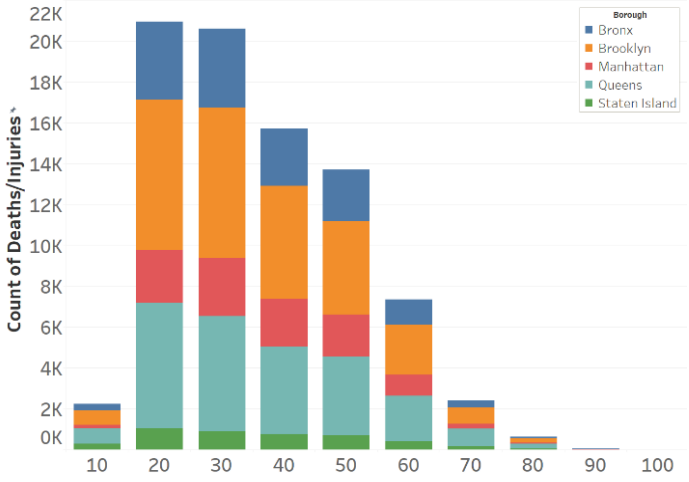

Figure 3: IOD per Age Group per Borough

Age groups 20 to 50 years of age have the highest count of `IOD`. Brooklyn is the leading borough across the board.

Weather

The weather effects were less powerful than expected. For example, there was no relationship between reported visibility or snowfall and accident severity. There are some explanations, one is that drivers tend to adjust their behavior in response to dangerous conditions. Another is that rain reduces traffic because people generally tend to avoid driving in bad weather. *Precipitation* did influence accident severity, but the sign of the coefficient was the opposite of what was expected: rainy conditions tend to be associated with fewer severe collisions.

While it should be considered that responsible drivers may adjust their behavior according to dangerous conditions (as suggested above), one should also consider the more cynical explanation that inclement weather simply keeps high-risk vehicles (i.e. motorcycles and bicycles) in the garage.

Warmer weather tends to be associated with more severe collisions, which can be seen in both the predictive relationship with season and with temperature. While part of this relationship is due to the motorcycle/bicycle effect mentioned above, the relationship still exists and is significant after adding vehicle weight to the model. While it would be unreasonable to expect a driver to stay home during the summer, the responsible driver should be aware of the increased risk during these times and adjust his/her behavior accordingly. High windspeed is significantly associated with worse collisions.

Time of Day

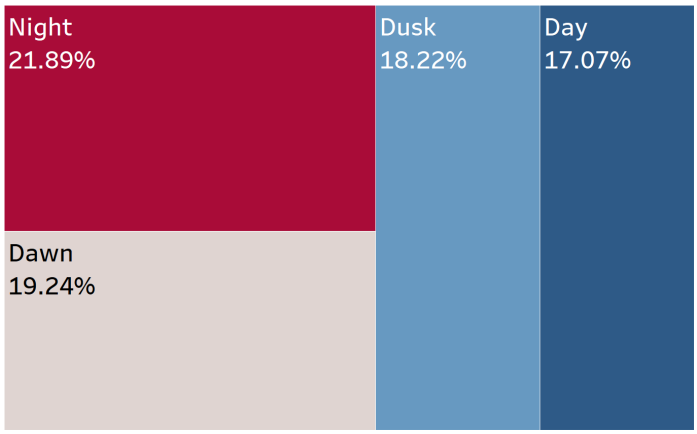

Figure 4: IOD per Time of Day

Driving at night carries significantly greater risk than driving during the day: In the case of this model, driving at night (as opposed to during the day) increases the log odds of injury/death by 0.31, which is equivalent to increasing the odds of a severe collision by 36%. A possible explanation is that reduced visibility at night reduces the amount of time available to react to a potential collision. It is also worth considering the possibility that people are more likely to be impaired at night, either due to insobriety, drowsiness, or some other explanation.

Miscellaneous

Driver choice has limited impact on accident severity. Two variables that a driver may have in his or her control are when to drive and what to drive. In addition, there are several variables that are significantly associated with accident severity but are outside of the driver's control. These include age, sex, license status, and the population density of driving environment.

Ultimately, a large portion of traffic risk comes down to individual driver behavior and cannot be modelled using the data we have available, though we have used our data to draw valuable insights.

RECOMMENDATIONS

Regarding auto liability premiums, rates are already tied to the type of vehicle being driven and the age and sex of the driver. These findings, however, could be quite valuable when associated with an *auto telematics program*, which can inform the insurance company where, when, and how much the policyholder is driving.

For example, a telematics transceiver could charge a rate per hour or per mile driven that was different in summer versus

winter, or different during the day versus at night. There could be a surcharge on the rate if the driver chooses to go out during a windstorm or on a very hot day. Similarly, the rate could adjust if the driver chooses to travel in a less densely populated area that increases the risk of an injurious collision. These premium adjustments would allow the insurance company to offer a better price to less risky drivers and would also allow riskier drivers to lower their premiums by changing their behavior. The potential of a lower premium for drivers would give the insurance company an advantage in acquiring new policyholders and keeping preferred existing policyholders.

REFERENCES

[1] Stewart, T. (2023, April). *Overview of Motor Vehicle Traffic Crashes in 2021 (Report No. DOT HS 813 435)*. National Highway Traffic Safety Administration. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813435

[2] Yang Mao, et. al. (1997, March). *Factors Affecting The Severity Of Motor Vehicle Traffic Crashes Involving Young Drivers In Ontario.* Injury Prevention Journal. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1067816/pdf/injprev00003-0031.pdf

[3] NYC OpenData (2023, June). *Motor Vehicle Collisions - Crashes*. City of New York. https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes

[4] NYC OpenData (2023, June). *Motor Vehicle Collisions - Vehicles*. City of New York. https://catalog.data.gov/dataset/motor-vehicle-collisions-vehicles

[5] NYC OpenData (2023, June). *Motor Vehicle Collisions - Persons*. City of New York. https://catalog.data.gov/dataset/motor-vehicle-collisions-person

[6] NYC OpenData (2023, June). *Zip Code Boundaries*. City of New York. https://catalog.data.gov/dataset/zip-code-boundaries

[7] NCEI DSI 3505 (2023, June). *Global Surface Summary of the Day*. National Centers for Environnemental Information (NCEI/NOAA). https://www.ncei.noaa.gov/data/global-summary-of-the-day/archive/

[8] National Oceanic and Atmospheric Administration, *NOAA Solar Calculator*. https://gml.noaa.gov/grad/solcalc/

[9] *Visual Crossing Weather API Documentation*. https://www.visualcrossing.com/resources/documentation/weather-api/timeline-weather-api/