Crash Data Analysis in Practice using Python

Anthony Escalona
Seidenberg School of CSIS
New York, Ny 10038
Email: ae50483p@pace.edu

***Abstract*—Roadway traffic safety is a significant concern for transportation governing agencies as well as ordinary citizens. To provide advice for safe driving, careful analysis of road traffic information is important to identify variables closely related to fatal accidents. In this paper, I apply statistical analysis and data mining algorithms on the NYC Open Data portal dataset as an attempt to address these problems. The relationship between incident rate and other attributes, including collision manner, weather, surface condition, and temperature condition were investigated.**

I. Introduction

It is estimated that around 4,000 New Yorkers are seriously injured in New York and more than 250 people are killed in traffic accidents every year. The automobile is the leading cause of injury-related death for children under 14 years of age and the second leading cause for seniors. On average, every two hours, vehicles severely injure or kill a New Yorker. The cost of these deaths and injuries impacts the city's social and economic growth greatly. New York City should no longer consider traffic crashes as mere "accidents," but as preventable incidents that can be addressed systematically. No degree of fatality is unavoidable or appropriate on the streets of the city. New York City's Vision Zero Action Plan[1] is the foundation to reduce traffic deaths and injuries. City of New York will use every available tool to enhance the safety of our streets. With this action plan, it is making a bold new commitment to improving street safety in every neighborhood and district – with increased enforcement of dangerous moving violations such as speed and failure to yield to pedestrians, new street designs and configurations to improve safety, widespread public access and communications, and a comprehensive legislative agenda to increase penalties.

Data mining is a major step in knowledge discovery. It is the process of extraction of non-trivial, valid and potentially useful information from huge databases. Some of the important data mining techniques are classification, association rule mining, segmentation, and clustering.

Predicting where and when road incidents will occur is complicated. It is possible to analyze traffic injury statistics and identify a correlation between variables based on historical traffic event data. On the other hand, visualization of data from traffic accidents provides detailed insights into how it changes over time. This paper focuses on practical issues related to the project to prevent road accidents. Analysis and visualization of data help observe the occurrence of traffic accidents and take appropriate action to enhance safety. More interestingly, does climate lead to motor vehicle collisions?

The paper is as follows. Section II addresses motivation. Section III addresses related work. Section IV presents data analysis and evaluation lessons learned. Section V concludes and outlines future work.

II. Motivation

NYC's Vision Zero Action Plan was launched in 2014, detailing 63 different programs that are implemented by the Mayor's Office and several City Agencies to minimize mortality and serious injury. The Vision Zero Task Force has since launched 143 new initiatives for a total of 206 initiatives (40 new initiatives were implemented in 2015, 22 were added in 2016, 26 in 2017, 27 in 2018 and 28 in 2019). City Agencies separate these programs and continue to make progress on their following measures [2]:

1) New York City Department of Transportation (DOT):
    a) Launch an integrated speed reducer installation program
    b) Install speed cameras at additional school zone locations

c) Expand and enhance People Priority Streets to improve pedestrian safety and access

2) New York Police Department (NYPD):
   a) Expand outreach and enforcement regarding the safe execution of left and right hand turns by all motorists
   b) Expand NYPD's clear bus routes enforcement action plan
   c) Increase safety within the trade waste and private carting industry through outreach and enforcement
3) New York City Taxi & Limousine Commission (TLC):
   a) Ensure TLC-licensed vehicles with outstanding part recalls are fixed in a timely manner
   b) Reduce use of left turns
   c) Introduce predictive analytics relating to driving behaviors and crashes through CRASHStat and the Fleet Office of Real Time Tracking (FORT)

More vigorous enforcement of dangerous driving behavior by the NYPD and the TLC may help to reduce traffic fatalities and serious injuries. In addition to greater enforcement, efforts are being made to upgrade equipment and technology for speed detection, increase the number of staff on the highway unit, and expand the breadth of information and data captured to preserve crash details better. Finally, the purpose of the data analysis is to derive useful data from the data source and to further use it for visualization by using statistical models.

III. Literature Review

A number of studies have been conducted to determine the factors leading to serious road accidents and to reduce the number and severity of injuries by removing or regulating these factors. As the traffic accident is large and heterogeneous, most scientists adopted data mining methods to carry out their studies.

Chong, Abraham, and Paprzycki [3] have applied artificial neural networks and decision trees to a specific data collection from the National Automotive Sampling Program and General Estimates Systems including traffic incident information from 1995 to 2000. The collection of data was limited only to head-on collisions. The findings revealed that neural networks were outperformed by the decision tree approach. The findings found that seat belt use, highway lighting condition, and driver alcohol use were the most important variables in fatal injuries.

Authors in [4] have applied their work using multivariate logistic regression to determine the independent contribution of driver, crash, and vehicle characteristics to driver' fatality. The result showed that increased use of seatbelts, reduced vehicle speed, and reduced number of and severity of drivers' side-impact might prevent deaths.

The purpose of paper [5] was to employ logistic regression models to develop crash-related injury prediction models. They analyzed traffic crash data in Kentucky during 2001 using logistic regressions. They concluded that the occupant's risk factors for the high level of injury severity were age, gender, and non-use of restraints. In [6], the authors used the same model to quantify the association of driver's age with traffic injury severity. Wisconsin crash data from 2000 to 2004 was used to study 602,694 drivers of a car or truck who were involved in a motor vehicle crash. It was discovered that the oldest drivers, especially those older than 85 drivers 85 years and older, had the highest risks for serious injury or fatality.

IV. Data Analysis and Evaluation

A. Background

The following information should provide enough background for the reader unfamiliar with Linear and Logistic Regressions to understand the points made in this paper. Readers already familiar with them can skip to the next section. Regression analysis is the most dominant force in driving business decisions today. It has many useful characteristics; one is the easy interpretation of results. Regression concepts are widely understood, and the methodology is well developed such that a well-tuned regression model can outperform other complicated algorithms.

Linear regression is a basic form of predictive analysis. Its overall idea is to analyze two things: (1) does a series of predictor variables do a good job of predicting a (dependent) outcome variable? (2) In particular, what variables are important predictors of the outcome variable, and how do they influence the outcome variable, indicated by the magnitude and sign of the beta estimates? These measures of regression are used to explain the relationship between a dependent variable and one

or more independent variables. Linear regression has a form of Y = a + bX equation where X is the explanatory variable and Y is the dependent variable.

Logistic Regression is a statistical method used to evaluate a dataset in which one or more independent variables calculate an outcome. The result is measured by a dichotomous variable in which there are only two possible outcomes. It uses a formula, like linear regression, as the representation. However, the outcome variable is categorical like two-valued outcomes like true/false, pass/fail, or yes/no. To predict value y, input values x are combined linearly using weights or coefficient values as beta. A key difference from linear regression is that the computation output value is a binary number (0 or 1) instead of a numerical value. Below is an example logistic regression equation:

$$y = e\hat{}(b0 + b1*x) / (1 + e\hat{}(b0 + b1*x))$$

Where y is the predicted output, b0 is the bias or intercept term, and b1 is the coefficient for the single input value x. Each column in the input data has an associated b coefficient, a constant real value, that must be learned from data.

B. Methodology

To visualize data, I used the Python programming language. All necessary libraries loaded, and the script required the following libraries: 1) google.cloud to access BigQuery database, 2) pandas for data frame management, 3) Matplot for visuals, and 4) Numpy to perform scientific computing. For making requests, Google authentication is mandatory. The python code must set the google_application_credentials environment variable. Replace [PATH] with the JSON file path that contains the service account key and the filename of [FILE NAME]. The attribute is for the current shell session only.

The collision dataset is made available through the NYC Open Data initiative and can also be searched through the BigQuery Community Datasets offered by Google. For comparison with collision data, historical weather data from NOAA weather data were used. Google BigQuery makes both NOAA weather data and NYPD traffic collisions available publicly.

The first table, is a compilation of daily collision data using years from 2012 to 2019 and limit the return query to the New York City Area. It (Figure 1) contains the number

	reason	year	hour	month	day	dow	incidents
0	Using On Board Navigation Device	2018	17	1	11	5	1
1	Accelerator Defective	2017	16	10	14	7	1
2	Reaction to Other Uninvolved Vehicle	2014	22	8	2	7	1
3	Other Lighting Defects	2018	19	10	10	4	1
4	Steering Failure	2018	4	2	2	6	1
5	Other Electronic Device	2013	11	2	2	7	1
6	Other Electronic Device	2015	19	10	18	1	1
7	Reaction to Other Uninvolved Vehicle	2014	21	1	10	6	1
8	Accelerator Defective	2015	9	9	16	4	1
9	Animals Action	2018	12	1	4	5	1

Figure 1. NYC Motor Vehicle Collisions Crash Table

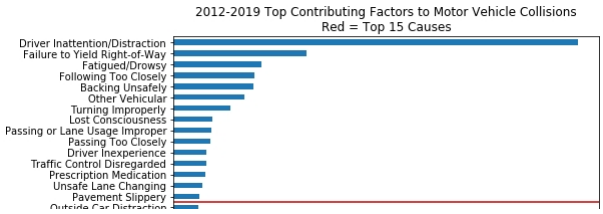

Figure 2. Contributing Factors

of motor vehicle collisions reported per year, month, day, and day of the week (dow). This ranges from 2012 to 2019 and has nearly 30 forms of explanations along with the associated counts of the incidents. As a crash data baseline, figure 2 plots the top contributing factors for collisions grouped by reason and using incident as its sum aggregate. According to the data, the leading cause of crashes is distraction followed by failure to yield right of way.

Next, the second table of data contains weather data reported by element, year, month, day, dow, and location name (Figure 3). The NOAA weather dataset includes over 40 meteorological elements[7], including temperature daily maximum/minimum, the temperature at observation time, precipitation, snowfall, snow depth, evaporation, wind movement, wind maximums, soil temperature, cloudiness, and etc. In this paper, we limit the number of input variables to four elements: 1) TMIN - Minimum temperature, 2) TMAX - Maximum temperature, 4) PRCP – Precipitation and 5) SNWD - Snow depth. The final construction of the weather data required unions of bigquery-public-data.ghcn_d.ghcnd_[year] tables from 2012 to 2019 and finally join the lookup table called, bigquery-public-data.ghcn_d.ghcnd_stations to map the data to New York City.

	id	value	element	year	month	day	dow	date	time	name
0	US1NYQN0026	18.0	PRCP	2018	3	7	4	2018-03-07	None	JACKSON HEIGHTS 0.3 WSW
1	US1NYQN0026	56.0	PRCP	2018	3	13	3	2018-03-13	None	JACKSON HEIGHTS 0.3 WSW
2	US1NYQN0026	254.0	PRCP	2018	9	28	6	2018-09-28	None	JACKSON HEIGHTS 0.3 WSW
3	US1NYQN0026	147.0	PRCP	2018	5	16	4	2018-05-16	None	JACKSON HEIGHTS 0.3 WSW
4	US1NYQN0026	0.0	PRCP	2018	2	10	7	2018-02-10	None	JACKSON HEIGHTS 0.3 WSW
5	US1NYQN0029	3.0	PRCP	2018	12	23	1	2018-12-23	None	QUEENS 2.1 NE
6	US1NYQN0026	15.0	PRCP	2018	10	16	3	2018-10-16	None	JACKSON HEIGHTS 0.3 WSW
7	US1NYQN0026	0.0	PRCP	2018	7	9	2	2018-07-09	None	JACKSON HEIGHTS 0.3 WSW
8	US1NYQN0029	0.0	PRCP	2018	12	26	4	2018-12-26	None	QUEENS 2.1 NE
9	US1NYQN0026	5.0	PRCP	2018	9	14	6	2018-09-14	None	JACKSON HEIGHTS 0.3 WSW

Figure 3. NOAA Weather Table

The third table combines weather and collision data table as one. Using Pandas, groupby operation is performed on the date, day, year, month, and dow to aggregate the weather data. Additionally, new attributes are created from the element column. For the collision data, the top 20 incident reasons are filtered since anything less is insignificant. Pandas pivotable functionality is used to group the data according to reasons performing a sum aggregation on incidents. The data is indexed by day, year, month, and dow. Any non-number inputs are converted to integer value 0. Finally, a merge of weather and collision data is performed thus creating a new data table.

A correlation matrix is used to visually estimate the linear historical relationship between the returns of the weather variables using the merged data. IsWeekend (Sat, Sun) feature was added instead of dow. The built-in .corr() method on the pandas DataFrame calculated the correlation matrix. Correlation ranges from -1 to 1. The matrix shows correlation coefficients between weather, x-axis, and contributing factors. Each cell in the table shows the correlation between the two variables (Figure 4). The represented data in a clustermap form cosine distance based on library from Matplotlib.

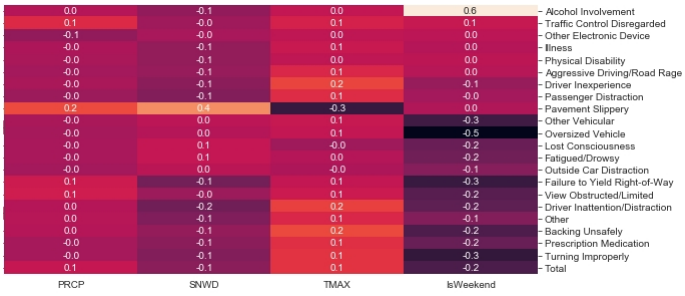

Figure 4. Correlation Matrix

Next we want to measure the effect on the various types of accidents that each parameter has. We iterate through the types of incidents and developed versions of linear regression models with ordinary least squares. For each variable (isWeekend, PRCP, SNWD, TMAX) P-values and coefficient were extracted and plotted as separate graphs. Green bars show a p-value $\geq$ 0.05 (Figure 5).

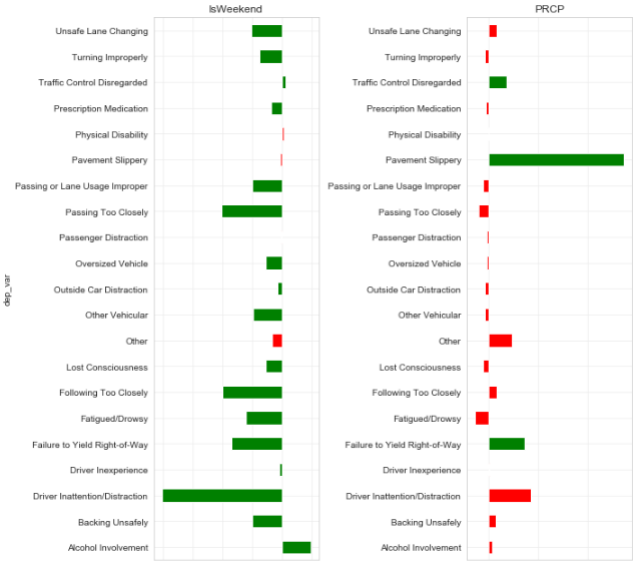

Figure 5. Regression

C. Observations

Based on observations, There are some interesting patterns to see. Alcohol accidents are correlated with the weekend. Illness, physical disability, and road rage seem to occur regardless of weather patterns, maybe slightly influenced by temperature. Driver distraction, passenger distraction, and inexperience are correlated with higher temperatures. The other type of road traffic accidents was drop out of analysis due to the low number of occurrences. Most categories fall off on the weekend, as proven by looking at the data. Traffic & rush hour commuting genuinely indicates lower rates. Precipitation in terms of 1/2 inch of rain adds 6–7 accidents while every 1/2 inch of snow depth adds 1–2 accidents.

V. Conclusion and Future Work

In summary, road traffic accident data analysis and visualization assist investigators in making proper conclusions by observing how weather conditions and would effect the road traffic frequency. The aim was to show how to extract meaningful data from raw database and visualize it. The research may further be expanded by observing how NYC

311 complaints such as ice cream truck noise and location would effect to the road traffic collision.

ACKNOWLEDGMENT

I offer my sincerest gratitude to the Professors of CS802 and CS816 for reviewing the paper.

REFERENCES

[1] "Nyc vision zero," 2019, last access November 2019. [Online]. Available: https://www1.nyc.gov/site/visionzero/index.page

[2] "Vision zero scorecard and initiatives," 2019, last access November 2019. [Online]. Available: https://www1.nyc.gov/site/visionzero/initiatives/initiatives.page

[3] M. M. Chong, A. Abraham, and M. Paprzycki, "Traffic accident analysis using decision trees and neural networks," 2004.

[4] M. Bedard, G. H. Guyatt, M. J. Stones, and J. P. Hirdes, "The independent contribution of driver, crash, and vehicle characteristics to driver fatalities," *Accident Analysis and Prevention*, vol. 34, no. 6, p. 717, 2002.

[5] M. Singleton, H. Qin, and J. Luan, "Factors associated with higher levels of injury severity in occupants of motor vehicles that were severely damaged in traffic crashes in kentucky, 2000-2001," *Traffic Injury Prevention*, vol. 5, no. 2, pp. 144–150, 2004.

[6] R. B. Hanrahan, P. M. Layde, S. Zhu, C. E. Guse, and S. W. Hargarten, "The association of driver age with traffic injury severity in wisconsin." *Traffic Injury Prevention*, vol. 10, no. 4, pp. 361 – 367, 2009.

[7] NOAA, "Readme file for daily global historical climatology network," last access November 2019. [Online]. Available: https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt