N.Y.C. Traffic Accidents: Location, Frequency, and Factors

A Crash Course in Urban Accident Analysis

Prepared by: Alex Shannon, Anupama Santosh, Lan Mi, Yubo Li

Abstract

Our team set out to investigate the city's robust NYPD Motor Vehicle Collisions dataset, combining it with other datasets to investigate what factors contribute to a severity of accidents. Our research is driven by two primary questions: 1. *What are the external factors that affect the severity of accidents in New York City?* and 2. *Can the effects be quantified into a street safety score?* Question 1 takes up the majority of our exploratory analysis and modeling, however we believe that it is essential to keep Question 2 in mind while conducting this research, as it provides a cohesive framework for further decision-making by the city.

From our analysis, in addition to the obvious human factors, we were able to identify other non-subjective features which have a significant effect on the accident severity using Chi-square test for association and Cramer's V statistic. These results motivated us to build a composite score for each street based on site characteristics and traffic /crash volumes using the widely accepted technique of Factor analysis and Principal Component analysis. It can be used for comparison across streets or settings, or over time, provided the separate indices are calculated with the same variables.

Introduction

A typical year in New York City witnesses roughly 4,000 severe traffic accidents and 250 traffic related deaths. In line with Mayor de Blasio's *Vision Zero* initiative to eliminate all NYC traffic deaths by 2024. Upon analysis of leading contributing factors, it is evident that human error plays a vital role in most, but not all, incidents (Figure 1). We believe the most promising way to reduce human error is to further automate vehicles, thus eliminating possible situations (e.g. distraction or fatigue) in which human error is likely to occur. This process which is rapidly occurring in the auto-industry at present, and is not so much tied to city policies as it is

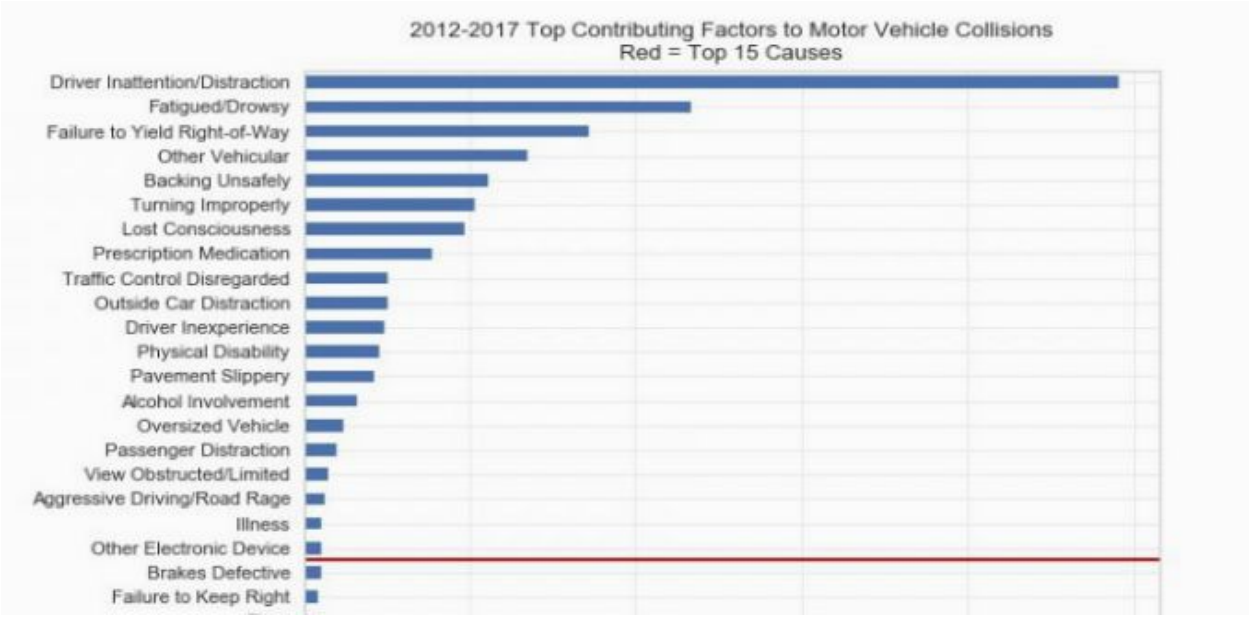

Figure 1. Top contributing factors to motor vehicle collisions

to the advancement of tech and the adoption of these vehicles on a larger scale. For our analysis, we find the exploration and analysis of non-human factors contributing to accidents to be a much more fruitful pursuit in-line with Vision Zero objectives, as these are factors that can potentially be influenced by the city through means such as road repair, traffic patterns, new

laws and other means of influence, thus leading to tangible reductions in traffic deaths in the coming years.

Literature Review

The *Vision Zero* initiative has prompted many investigations into NYC collision data. Kaggle released a cleaned-up version of the dataset, covering the range of 2015 to early 2017. This lead to a large amount of publicly available analysis on this dataset, investigating lead causes of accidents, types of injuries that occurred, and general exploration of the correlations between the variables included in the dataset.

In addition, many New York area data bloggers have attempted investigations into the dataset. We reviewed reports such as time-series analyses of accident causes, attempts to view the data spatially, and insight from New York Car accident reports. While the dataset itself has been largely exhausted in terms of analytics, what we found lacking in the domain is a combination of the dataset with outside factors, such as weather conditions, traffic density, and other factors that may (or may not be) associated with accident occurrence, and particularly looking at these data from a spatial perspective, attempting to identify key locations that are prone to accidents. This is the area we found lacking, and thus decided to pursue further analysis to see what the inclusion of additional datasets might reveal.

Datasets & Preparation

NYPD Motor Vehicle Collisions Data: The central dataset for our analysis is, without question, the NYPD Motor Vehicle Collision dataset. It can be accessed through NYC Open Data project and can also be found through Google's BigQuery Public Datasets This dataset contains over 1.17 million rows, consisting of wide-ranging features such as date-time information, accident

location (at borough, zip, street name, and GPS detail), injury and death counts broken down into categories such as motorists, cyclists, and pedestrians; more granular metrics are also included, such as 'bridge direction' and 'vehicle make.' We cleaned this data to ensure that all numbers were represented in the correct format, and accounted for null and missing values.

Weather Data: We acquired historical weather data from the National Centers for Environmental Information; we needed to request this dataset, and it was received via email. These data required a good deal more cleaning for our purposes than did the collision data; for example, did the *amount* of precipitation matter? Or just the fact that it was raining? To accommodate these discrepancies, we created both continuous variables (snow depth, rainfall, etc.) and binary variables to demarcate whether there was snow, fog, rain, etc. on a particular date. Because weather does not vary significantly between boroughs (let alone zip codes and streets!), we chose to use a single dataset for the whole city, using weather data collected in Central Park. We then merged this data with collision data by date.

Traffic Volume Data: Traffic volume counts was one feature that would be integral to our analysis, but strenuous to quantify. We looked for various proxies for traffic counts like NYC taxi data, pedestrian count data etc. Either the data was sparse or it was difficult to get it at the street level. At this juncture, although crude, we decided to use the New York State DoT's Average Annual Daily Count as the proxy, the shapefile for which was obtained from Traffic Viewer.

NYC Street base map: Our objective necessitated us to map all collisions onto the street and analyze it at a street level. The Department of Planning manages the LION dataset, a single line street base map representing the city's streets and other linear geographic features, along with feature names and address ranges for each addressable street segment. It also makes

available information on Street Usage Class, Direction of Traffic etc. This file had to be cleaned of all pedestrian paths, non-vehicular segments and non-street factors. All other datasets were merged with this to conduct street level analysis.

Parking Regulation Signs: The DOT manages over one million traffic signs in New York City. DOT distributes the data that underlies the search tool in shapefile format. The file includes the location and a description of parking signs throughout the city, and is updated monthly.

Street Pavement Rating: The New York City Department of Transportation is responsible for keeping the City's streets in good repair. The Agency performs ongoing assessment of New York City streets. Ratings are based on a scale from 1 to 10, and results are grouped in the following categories: Good (%) - ratings of 8 to 10, Fair (%) - ratings of 4 to 7, and Poor (%) - ratings of 1 to 3.

Data on Speed Limits: Department of Transportation maintains the speed limit data of all segments in NYC. The data can be downloaded as a shapefile. This is part of the Vision Zero data feeds the city has made public

Data on Street Humps: Department of Transportation maintains the data on humps across all segments in NYC. The data can be downloaded as a shapefile. This is part of the Vision Zero data feeds the city has made public

Most of the datasets had to be spatially joined and for the collision data geometry had to explicitly specified. Same epsg units were maintained and GeoPandas package was used to manipulate and join datasets.

Exploratory Analysis

We performed initial descriptive and exploratory analyses of the motor vehicle collision data in order to attend the obvious trends/patterns and biases. We examined a relatively consistent trend in total accidents per year from 2013-2017 (notably, 2013 was lower than later years, but this is due to the transition to the practice of recording the data by NYPD officers and not an anomaly of a low-accident year). More revealingly, we plotted collisions by hour (Figure 2), revealing that significantly more accidents take place in evening hours than in morning hours, particularly during the prime commuting hours of 14:00-18:00. One question that begs further investigation is *why do these hours, which experience roughly the same traffic as morning hours result in more collisions than morning hours*?

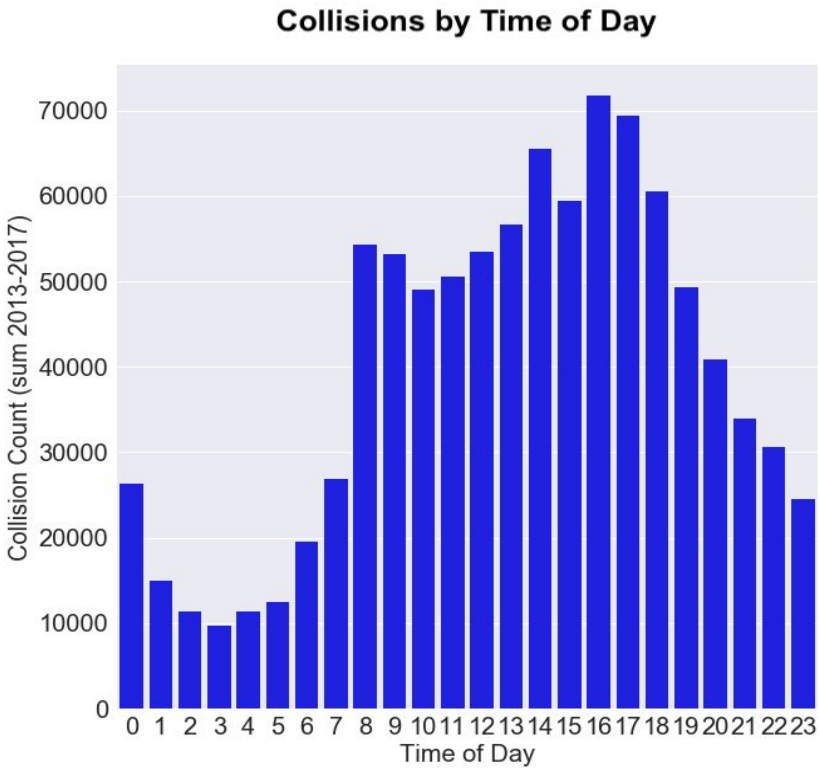

Figure 2. Collisions by Hour of Day: we can see a clear peak in collisions during evening hours.

In addition to temporal plots, we investigated our data spatially, creating a coarse plot of all recorded incidents (Figure 3). The points roughly correspond to the street map of New York City. You see very few collisions near bridges, as opposed to street crossings and curves, where there is a noticeably higher density of collisions. This also shows that the data is not biased by geography, and all areas are equally represented.

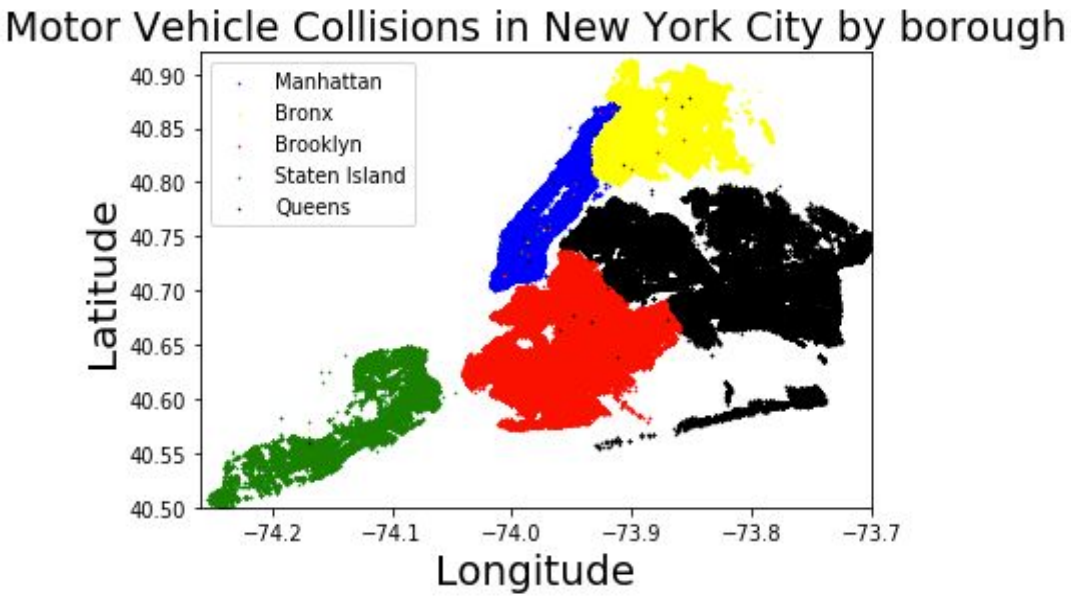

Figure 3. Coarse Plot of All Collision Points (2013-2017)

To visualize the severity of motor vehicle collision in New York City, we divided the dataset into three main collision types (fatal, personal injury, car-body damage) grouping the points by their most severe category (e.g. a fatality that also involved significant damage to the car-body would simply be classified as a fatality). The resulting image provides a good proxy for locations in which fatal incidents are more likely to occur than simple damage or injuries, thus providing an intuitive lens through which to view our later analysis.

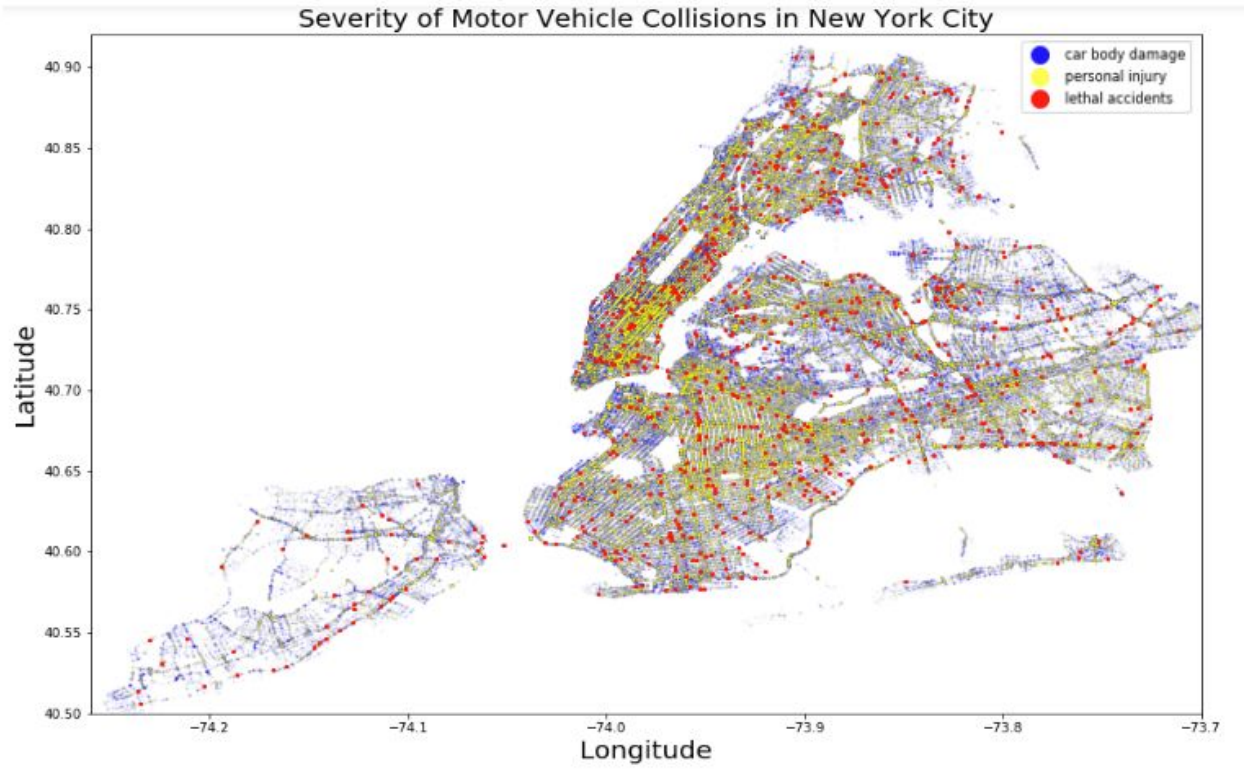

Figure 4: Plot of Accidents by Severity

Methodology

Variable definition and selection: The purpose of this initiative is to identify the factors that influence the severity of a collision. To conduct this analysis, we make use of the 'Collision Severity Metric' that was defined before based on 3 categorical levels:

- Fatal: At least one person involved with the collision passed away.
- Severe: Accidents above the 50th percentile based on number of injuries and damage
- Slight: Accidents below the 50th percentile based on number of injuries and damage

The possible factors under consideration to test the influence on a collision can be broadly defined as:

- Human factors
- Non-human factors

The human factors such as driver's attentiveness, reckless driving, etc. are extremely difficult to adequately quantify and represent and are virtually non existent and are ignored for the purpose of this analysis. The entire focus is on the non-human factors, which are as follows:
Quarter of the day, Weekday, Time of accident, Type of Vehicle, Speed Limit, Traffic density, Street Usage Class, Presence of Humps, and Presence of Traffic Regulatory Signs. We observe that a large number of variables are in fact, categorical. For uniformity of analysis, we convert the other continuous variables into categorical as well, through bucketing.

Chi-squared test: Now, we move forward to conduct an association analysis by making use of the widely used chi-squared test for association. Each variable is tested against the collision severity metric under three different weather conditions:

- Clear
- Rain
- Snow

Briefly, the chi-squared test of association works through the following steps:

- Tabulate the data of the 2 categorical variables in such a way that the levels of one variable occupy the rows and the levels of the other occupy the columns with a corresponding total row and total column to stand for the sum of each row and column respectively.
- For every level of each variable, the proportion of that level in the total is calculated by dividing the corresponding row/column sum with the overall total number of observations in the table.
- The expected values of each level of a variable with respect to a level of the other variable is calculated by multiplying the corresponding proportion by the row/column sum of the level of interest.

- The contribution of this to the test-statistic is:

$$\frac{(observed - expected)^2}{expected}$$

The test statistic is the sum of this metric across all cells in the table.

Cramer's V: In order to quantify the results of the chi-squared test into a usable association value, we used the Cramer's V measure. The Cramer's V association between 2 variables is defined as follows:

$$V = \sqrt{\frac{\frac{\chi^2}{n}}{min(k-1, r-1)}}$$

Here,

- χ^2 is the value of the test statistic that we just deduced
- *n* is the sample size of this analysis(number of collisions in our data for the particular weather type)
- *k* is the number of columns in the chi-squared table
- *r* is the number of rows in the chi-squared table

The denominator is essentially an indicator towards the *degrees of freedom* of a chi-squared test. We found out that there were many major associations with the collision severity metric, which were mainly covered by the properties of the street in which the collision occurred, such as:

- Speed limit
- Number of humps
- Traffic regulation signs
- Traffic direction

This indicated that a single variable which could incorporate the effects of all such street parameters would possibly be valuable.

Bartlett's test of sphericity: The Bartlett's test of sphericity is used to test whether the correlation matrix is significantly deviant from the corresponding identity matrix (which represents the null hypothesis). Further dimensionality-level analysis can be performed only if the null hypothesis is rejected.

The test is conducted in the following way:

- Calculate the correlation matrix of the entire data (we choose the original forms of the bucketed variables that were used for the chi-squared test, thus they are continuous)
- We compute the determinant of the correlation matrix. Under the null hypothesis, the identity matrix has a determinant of 1. If the variables are highly correlated, the determinant is close to 0.
- We define the Bartlett's statistic as follows:

$$\chi^2 = -\left(n - 1 - \frac{2p + 5}{6}\right) ln|R|$$

Here,

- n is the number of samples being used to prepare the correlation matrix
- p is the total number of variables under consideration
- $|R|$ is the determinant of the correlation matrix

This statistic indicates the amount of deviance from the null hypothesis. Once it is confirmed that the correlations are good enough to proceed, we implement factor analysis as a way to build the common street parameter variable that we talked about previously.

Factor Analysis: Factor Analysis is primarily used as a method of dimensionality reduction, but here, we additionally use it for weighting our variables. The combination of the indicator variables with their corresponding weights can be used as a 'Composite Safety Index'. The stepwise procedure is outlined as follows:

- Make sure that the Bartlett's test of sphericity rejects the null hypothesis (i.e., proves significant correlation between the variables).
- Convert each variable into a linear combination of a set of factors where the weights of each are the factor loadings and act as a measure of the correlation between the variable and the particular factor. There are many ways to do this. We choose the method of Principal Component Analysis where the factors are called Principal Components.
- Choose a specific number of the first principal components that can explain most of the variance in the data.
- Rotate the factors (principal components) chosen(alter their values directionally) to enhance their interpretability, i.e. to maximise the loadings of individual variables on individual factors for a good approximate one-to-one correspondence(as many variables as there are factors).
- The square of such factor loadings represent the proportion of the total unit variance of the variable explained by the factor. We make intermediate composite factors by scaling each of these to unity sum.
- These composites are aggregated by assigning a weight to each, which is equal to the proportion of the total variance explained by each of them in the data.

Results

The results of the chi square test for association are tabulated in Table 1. It can be seen that various factors have significant associations with collision severity. Also, the dependence is influenced by weather characteristics as well.

Table 1: Measure of association for various features with collision severity

	Clear		**Rain**		**Snow**	
	χ2(p-value)	**Cramer's V**	**χ2(p-value)**	**Cramer's V**	**χ2(p-value)**	**Cramer's V**
Quarter of the year	0.034	0.043	0.012	0.029	0.024	0.014
Weekday	0.042	0.039	0.038	0.026	0.048	0.033
Time of Accident	0.014	0.129	0.009	0.103	0.016	0.190
Type of Vehicle	0.037	0.098	0.043	0.086	0.056	0.075
Speed Limit	0.006	0.109	0.012	0.113	0.007	0.092
Traffic Density	0.015	0.070	0.032	0.089	0.041	0.067
Street Usage Class	0.023	0.064	0.031	0.075	0.043	0.066
Presence of Humps	0.017	0.053	0.022	0.042	0.018	0.031
Traffic Regulatory Signs	0.012	0.057	0.023	0.041	0.012	0.038

Significant associations: Time of the day has a major influence on severity of collision irrespective of the weather characteristics. In addition, on a clear day, type of vehicle, speed limit and traffic density are influential. When it is raining, speed limit and traffic density tend to be major factors. Accidents happening on snowy days can be characterised by Speed limit, Type of vehicle, Speed limit and Street Usage class. Time of accident on a snowy day has the highest Cramer's V coefficient of all associations.

These significant associations prompted us to further explore the possibility of creating a safety index for a street dependent of site characteristics and traffic/collision patterns. As mentioned in the methodology, we used the widely accepted technique of Factor Analysis coupled with Principal Component Analysis to develop a weighted index.

It is to be noted that only continuous variables can be used for Factor Analysis. Considering the results from the Chi-square analysis, we selected the following features collated at street level:

- A measure of collision density: Total number of collisions in a street were normalized by the average annual traffic density along each street. This would mask the bias associated with traffic volume
- Humps, Traffic Regulation Signs: Earlier these were used as binary variables indicating the presence of absence of such features. For factor analysis, the total number of these features were considered
- Street Ratings, Speed Limit, No:of lanes,Width of segment.

The methodology described earlier was followed taking into account the numerous considerations and general practises, and the factor loadings after varimax rotation is presented below.

Table 2: Factor loadings after Varimax rotation

Features	Factor Loadings	
	Factor 1	Factor 2
Measure of Collision Density	-0.62	0.14
Speed Limit	0.10	-0.06
Traffic Regulation Signs	0.03	0.19
Humps	0.01	0.36
Street Ratings	0.01	0.51

No:of lanes	0.07	0.04
Width of segment	0.23	-0.04
Percent of variance explained	0.64	0.2

The last step deals with the construction of the weights from the matrix of factor loadings after rotation, given that the square of factor loadings represents the proportion of the total unit variance of the indicator which is explained by the factor. The approach used by Nicoletti et al., (2000) is that of grouping the individual indicators with the highest factors loadings into intermediate composite indicators (Table 3). In our case, there were two intermediate composites. The first includes Measure of Collision Density (with a weight of 0.86), and Width of segment (weight 0.11). Likewise the second intermediate is formed by street ratings, humps and traffic regulation signs (worth 0.58, 0.29 and 0.08 respectively). The four intermediate composites are aggregated by assigning a weight to each one of them equal to the proportion of the explained variance in the data set: 0.63 for the first, 0.21 for the second. Note that different methods for the extraction of principal components imply different weights, hence different scores for the composite (and possibly different street rankings).

Table 3: Squared factor loadings scaled to unity

Features	Squared Factor Loadings (Scaled to unity)	
	Factor 1	Factor 2
Measure of Collision Density	**0.86**	0.04
Speed Limit	0.02	0.00
Traffic Regulation Signs	0.00	0.08
Humps	0.00	**0.29**
Street Ratings	0.00	**0.58**

No:of lanes	0.00	0.00
Width of segment	**0.11**	0.00
Percent of variance explained	0.63	0.21

The weights/scoring coefficients of each feature extracted using PCA is listed in Table 4.

Table 4: Weights/Scoring coefficients from PCA

Features	Scoring Coefficients(PCA)
Measure of Collision Density	0.4
Traffic Regulation Signs	0.05
Humps	0.17
Street Ratings	0.31
Width of segment	0.07

Composite Score: Using the scores from the principal components as weights, a dependent variable was then constructed for each street. Each of the features were standardized and then multiplied with the weights and the cumulative sum was then scaled to the range 0-10 for easy interpretability. This dependent variable can be regarded as the street safety score, and the higher the household socio-economic score, the higher the implied safety of street.

Conclusion

In addition to the obvious human factors, we were able to identify other non-subjective features which have a significant effect on the accident severity. These results motivated us to build a composite score for each street based on site characteristics and traffic /crash volumes. It can be used for comparison across streets or settings, or over time, provided the separate indices are calculated with the same variables.

Furthermore, features like time of the day, weather data etc. could not be incorporated as that would necessitate the modelling of a dynamic index which was computationally beyond the scope of our work. As next step, this kind of a real time index can be envisioned taking into account more varied factors.