Big Data Processing- A detailed data analysis on New York Accident Data

University of Derby

100568146@unimail.derby.ac.uk

Abstract.
This paper conducts a detailed data analysis report on New York accident data provided by New York Police Department (NYPD) and maintained by NYC OpenData. The main aim of the project is to model the data from one of the busiest traffic in the world, to ensure traffic safety and to identify the significant features in the data. Different models like K-Mean Clustering, Neural Network, Decision tree are built to predict or classify different factors and a Principal component analysis is used to reduce the number of variables. The results can be used for future references and to expand any new study on the data.

Keywords: Big Data, Linear Regresssion, K-NN, Naïve Bayes, Decision Tree, K-Mean Clustering, Artificial Neural Network, Principal Component Analysis.

Contents

1 Introduction

The United States is one of the busiest countries in terms of traffic with around 280 million vehicles on road. As of 2018, there were 12 million vehicles involved in crashes in the United States alone [Wagner,2020]. In US, the New York city is the fourth largest city with a population of 19 million [World Population Review 2021]. This population on road makes it an extreme traffic challenge [Hopping, 2019].

A detailed data analysis on New York traffic collisions data is built to analysis any trends and patterns for the accidents which could possibly be used for future references and different data analytical models are built to predict any future occurrences in New York city.

Aim and objectives.

Anyone who are currently residing in the city or those who plan to move to the city would like to know more about the traffic situation and the trends of the accidents in the city so that any precautions can be taken. Information like most of the accidents are occurred in time a timeframe between 16:00 and 17:00, and most of the accidents are occurred in Brooklyn can be used for people driving through Brooklyn and especially in those hours to be extra careful. Other information like the most frequent vehicle group in accident, the most frequent contributing factor that results in a collision or that results in filing a police report can be useful.

The main objective of the study is to understand the New York City accident data and extract useful information which can be helpful for the traffic safety and to built models to predict any future occurrences.

2 Dataset

The data set used in the study is Motor Vehicle Collisions- Crashes [NYC OpenData, 2021], which is about the New York city motor vehicle accidents. A police report is necessary for all crashes where someone is injured or killed, or where there is a collateral damage greater than one thousand dollars. The motor Vehicle Collision data contains information from all police reported motor vehicle collisions from the New York Police Department (NYPD). The data is contributed by the Police department and is maintained by NYC OpenData.

The data set was created in April 2014 and is regularly updated, and was last updated on January 3rd, 2022. The dataset has 1.85 million rows and 29 variables including crash date, crash time, type of vehicles in accident, the contributing factors in accidents etc. Figure 1 shows every variable in the dataset and the variable type. Each row in the data set represents an accident.

2.1 Data Variable Description

The dataset contains 29 variables Some of the important variables that are used in most of the study are explained here in short:

- CRASH TIME – Occurrence time of the collision

- BOROUGH - Borough where the collision occurred

- CONTRIBUTING FACTOR 1 - Factors contributing to the collision for designated vehicle

- VEHICLE TYPE CODE 1 - Type of vehicle based on the selected vehicle category

(ATV, bicycle, car/SUV, e-bike, e-scooter, truck/bus, motorcycle, other)
Other variables include THE NUMBER OF PERSONS INJURED, THE NUMBER OF PERSONS KILLED, CRASH DATE, etc.

3 Data Pre-processing

The dataset has a size of 380 MB and has 1.85 million datapoints. Most of the relevant variables like crash time, vehicle types and the contributing factors happens to be a character variable with large number of factors. Furthermore, a large part of the data contains missing values or Na's, altogether making it a challenging problem for data analysis.

3.1 Treating the missing values

The first step in data pre-processing is to treat all the missing values rather than deleting all the rows in the dataset. The missing values in all the character variables are replaced by the most frequent value of the respective column, for example Brooklyn is found to be the borough that most of the accidents occurs and hence 'BROOKLYN' is replaced for all the missing values in the borough column. All the missing values in any numeric variables are replaced by the mean of the respective column. The resulted dataset is used for building models and for data analysis.

It should be noted that the dis-advantages of using this method for missing values are that it can create a bias in the dataset, and it doesn't factor the correlation between the features [Silaparasetty, 2020].

3.2 Removing all the missing values

As replacing all the missing value can possibly bring out a bias inside the dataset, especially for variables with a large number of missing values, a dataset after removing all the missing value is created and few of the tests are simultaneously done on this dataset.

3.3 Sub-setting and dividing into train and test datasets

The large size of the data makes most of the data analysis procedures time consuming and the lack of memory and power for the local pcs make it a necessary for subsetting the data before each algorithm. In addition to avoiding the irrelevant variables like 'collision_id', most of the classification algorithms were tried on dataset which were subsetted for certain values. That is in a study to predict the number of persons injured, the variables like number of persons killed, number of motorists injured etc are avoided since they all are directional propotional.

Diving the dataset into train and test is a necessary before each model building since it allows as to check the model performance. The data after sub setting is divided into train and test in a ratio of 70 :30

4 Data Analysis

4.1 Regression

Regression analysis is one of the popular and oldest method to identify any relation between one or more independent variables and dependent variable. The aim is to find the optimum relationship between the independent variable that could explain most of the information withhold by the dependent variable.

The regression analysis is to estimate the f function in equation 1 that could closely describe the dependent variable [Wikipedia].

$$Y_i = f(X_{i,} \beta) + e_i \tag{1}$$

There are different potential functions that can be used as f in computing regression analysis. The researcher would need to find the potential fitting function f that is best fit for any data.

Linear regression.
One of the most used regression analysis techniques is linear regression or simple (multiple) linear regression. It assumes a linear relationship between the dependent variable and independent variables and propose that it can be explained as some linear combination of the independent variables [Wikipedia].

The model proposes a linear relationship as

$$Y_i = \beta_0 + \beta_{i*} X_i + €, \text{ where } i=1,2,3, \dots, n \text{ (number of data points)} \tag{2}$$

The coefficients are obtained by the popular least square measures that help as to get the optimum equation minimizing the error in the data.

***Linear regression in New York accident data*.**
A linear regression analysis to model the number of persons injured in any accident is build using eight independent variables. Some of the independent variables include Crash time, vehicles in the accident, the borough in which the accident occurred and so on.

The model gave a p-value of 3.062e-06and an adjusted R-square of 0.144, which is not bad given the fact that the model contains more categorical variable than numerical variables and the fact that all the categorical variables had many factors or levels.

Backward and Forward model selection.

The dataset in the study consists mostly of character variable with more than two levels. Regression model was built after factorizing each character variable as R can treat each factor variable as dummy variable by default. However, the large number of variables and its factors make it impossible to identify the most significant variables.

A stepwise algorithm that selects the most significant variables by forward and backward variable selection is used to determine the most significant variable. The borough is identified as the only significant variable in predicting the number of persons injured in an accident. The model built with the one significant factor yielded a p-value of 0.0006616and an adjusted R-square of 0.0073.

4.2 Classification algorithms

Classification is a process of extracting information from a pre-existing training dataset to train the machine to identify different classes or categories for these training datapoints. The basic idea is to train a model so that it can assign any new datapoint into one of the classes [Wolff, 2020] that it have identified.

Classification algorithm is a kind of supervised learning method that is used to identify new observations based on already available data. The already available data is taken as the training dataset and the algorithm is trained using the labels for a particular category. In a classification algorithm the output is a particular category.

K- Nearest Neighbors (K-NN).
K-nearest neighbors is a ‘pattern recognition’ [Wolff, 2020] algorithm that group datapoints to its nearest neighbor. It is a supervised machine learning algorithm that calculates the similarity between the datapoints and group them according to these similarities [Nelson, 2020]. The K-NN algorithm finds the k nearest neighbors of a data point and assigns class to the new data point accordingly.

K-NN algorithm for New York accident data.

A model to classify the time of most accidents occurred is build using K-NN. The data used is a subset which contain the accidents that occurred in 16:00 and 17:00 which is the time of most accidents. The model turns out with an accuracy of 50% in predicting the two categories.

Naïve Bayes.

The naïve Bayes algorithm is a mostly used supervised learning algorithm that classifies the data points on conditional probability values. It is a probabilistic classifier which predicts the probability of an object being in a class. The Bayes' theorem which gives the probability of an event with some prior knowledge is the basis of this mechanism [GeeksforGeeks, 2021].

The Bayes theorem can be stated as :

$$P(A/B) = \frac{\mathrm{P(B/A){*}P(A)}}{P(B)} (3)$$

Equation 3 gives the probability of an event given some information. The naïve Bayes algorithm equivalently uses a certain prior information of a data point to classify its class.

Naïve Bayes algorithm for New York accident data

A model to predict the most frequent type of vehicle in accidents is built using Naïve Bayes algorithm. The first step is to subset the data for two of the most frequent vehicles in accident. The most frequent vehicles are found to be Station Wagon/Sport Utility Vehicle and passenger vehicle. These two are taken as the two classes of the dependent variable for the study. A model is built using Vehicle type 1 as dependent variable and 17 other independent variables. The naïve Bayes was able to predict the test data set with a very low misclassification error 0.004 and 0.009. Naïve Bayes is a good predictor for classifying the contributing factor.

Decision Tree Classification.

Decision tree is an example of supervise learning which uses a set of rules like humans to make decisions [Bento, 2021]. The model learns itself by continuously splitting at each level and classifying at each split [Chakure, 2019].

Decision tree classification for New York accident data.

A model is built to classify the contributing factor for most of the accidents in the city is developed using decision tree. The data used is a subset of data for two of the most contributing factors for accidents. The two of the most frequent contributing factors is Failure to Yield Right-of-Way and Driver Inattention/Distraction.

A random sample of size 300 is used for the study due to large number of factors. A subset containing 18 variables were used for decision tree model. Two models, one using the whole 18 variables and other using a subset of these 18 variables were built and both the model performance were compared. The model with full 18 variables appeared to be a better model. The model had an accuracy of almost 99%.

5 Clustering

Clustering algorithms try to discover natural grouping orders in the given data. These are different from the supervised algorithms where the labels are given. Clusters are often groupings of datapoints with a set of similar features. The boundary of the cluster is usually determined by either measuring the extend of a point from the centre of the clusters or by some mathematical formulations. Clustering algorithms are particularly useful when we are not given priors about the data because it helps to eliminate existing patterns.

K-Means Clustering.

K-Means clustering aims to group a datapoint in one of the either of the k clusters. To do this it measures the distance from the centroid of the cluster. The number of clusters k is defined by the user. The objective of k-means is to group datapoints with similar characteristic together which helps to discover underlying patterns.

K-Means clustering for New York Accident.

The number of clusters is determined using a scree plot using the elbow method. The scree plot suggested a cluster size of 5 in the data and a cluster analysis is done on 5 clusters. A plot (figure 4) for the predicted clusters is plotted and the five clusters are binded back to the dataset

6 Deep Learning

Deep learning is a class of supervised learning techniques which uses artificial neural network to learn from the training data. It is called deep learning because of the multiple hierarchical layers of network.

6.1 Artificial Neural Network (ANN)

Artificial neural networks are algorithms inspired by neurons in animal brain. An artificial neural network consists of many nodes that takes a number of inputs and produces, and output based on a certain threshold value. This threshold value is usually determined by an activation function. An artificial neural network is trained by adjusting the wait associated with each connection. A typical artificial neural network in a dep learning setting consists of many layers of connected nodes which progressively represents data in increasing complexity.

ANN in New York accident Data

An ANN for New York accident data is built to predict the most frequent Boroughs are developed. The initial step was to reduce the number of factors in the dataset.

The data used is a subset the dataset containing just the most frequent observations of crash time, contributing factor 1, vehicle type 1 and borough. The most frequent boroughs are found to be Brooklyn and Queens. These are taken as the two classes in the dependent variable.

The resulted dataset is used for building the model.

The model has a 66% in predicting the borough Brooklyn and a lesser prediction and 52% for predicting the borough Queens. Overall, the neural network classification is good. Figure 4 represents the neural network that was built.

7 Dimensionality Reduction

Dimensionality reduction is an unsupervised learning method. The major goal of dimensionality reduction is to reduce the number of variables representing a data. By reducing the input dimension, we are reducing the degrees of freedom to represent the data. This results in a simpler structure. Dimensionality reduction techniques are specially used to find the variables that are most useful representing the data which in fact help us to prevent the data from overfitting.

7.1 Principal Component Analysis

Principal Component Analysis is a class of dimensionality reducing techniques which reduces the input features by finding the correlation between them and transforming them into a set of unrelated linear features.

Principal Component Analysis for New York Accident Data

The data frame we used for PCA is a subset of the main data that contain 9 variables. A principal component analysis is done to reduce the number of variables and is reduced to 3 dimensions.

The model developed three principal components for the data.

Linear regression for the Principal Components.

A linear regression model is built using the three dimensions from PCA. The resulted model has an R-square of 0.01 and an adjusted r square of 0.005. It was noted that the this model has not improved the value of adjusted R square.

8 Conclusion

.

A detailed data analysis was done on the New York accident dataset and the result were discussed. There is a huge hope for future studies as Traffic statistics need a lot of effort.

The model results from the study can be summarized as follows: -

Linear Regression is finding the best linear fit for the dependent variable using the independent variable. All the regression model showed not so good results.

The three classification algorithms developed were good. Among the three-classification algorithm decision tree to predict the contributing factor had almost 99%accurancy. Naïve Bayes classifier for the vehicle types in an accident also gave a very low mis-classification error. The K-Nearest Neighbour classifier for modelling the most frequent crash times has a 50% accuracy. Overall classification algorithms worked well under the circumstance.

A k- mean cluster algorithm was built. The scree plot suggested a cluster size of five and five clusters were built for the dataset.

A deep learning algorithm for predicting the borough was built fusing a neural network algorithm. The model was able to predict 66% of the time the borough 'Brooklyn' correctly, whereas a 52% in predicting the borough 'Queens'.

And finally, dimensionality reduction using Principal Component Analysis is done and a nine-dimension data is reduced to a three-dimension data. The obtained three components were used to a linear model. However, the model built using these components was statistically insignificant with a p-value>0.05

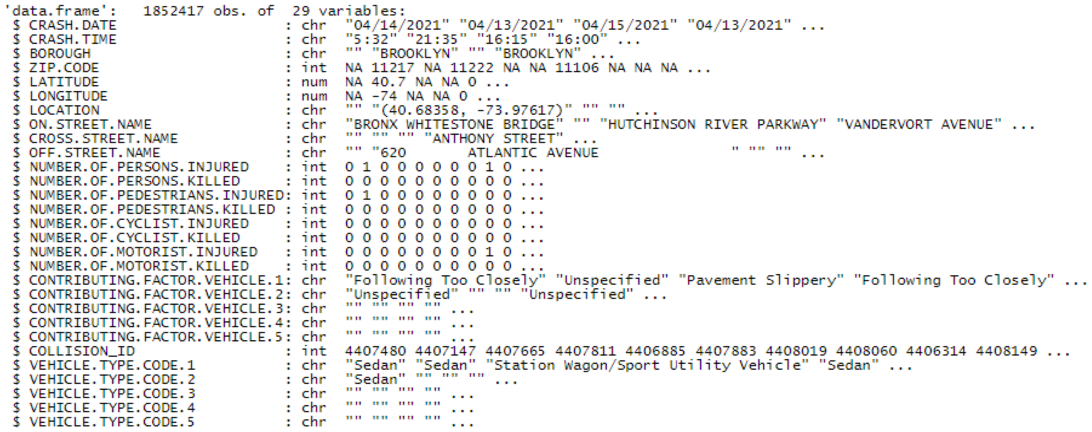
The study revealed several interesting factors. The models do have a hope for more and the study can be continued in future.

9 References

1. Wikipedia contributors, "Regression analysis," Wikipedia, The Free Encyclopedia, Available from : Regression analysis - Wikipedia (accessed January 3, 2022).
2. Wolff. Racheal., (2020),' 5 Types of Classification Algorithms in Machine Learning', Available at: 5 Types of Classification Algorithms in Machine Learning (monkeylearn.com) (Accessed 24 December, 2021).
3. Nelson, Daniel., (2020), 'What is a KNN (K-Nearest Neighbors)? - Unite.AI', Available at What is a KNN (K-Nearest Neighbors)? - Unite.AI, (Accessed: 22 December 2021).
4. Wagner ,I., 'Road accidents in the United States - Statistics & Facts | Statista',Available at: https://www.statista.com/topics/3708/road-accidents-Road accidents in the United States - Statistics & Facts | Statistan-the-us/ (Accessed : 2 January 2022).
5. Hopping, Gene., 'Traffic Accidents in New York City — A Linear Regression Study | by Gene Hopping | Towards Data Science', Available at: Traffic Accidents in New York City — A Linear Regression Study | by Gene Hopping | Towards Data Science (Accessed 5 December 2021).
6. Silaparasetty, Vinita., (2020) 'Guide to Handling Missing Values in Data Science' Available at: Guide to Handling Missing Values in Data Science | by Vinita Silaparasetty | Medium (Accessed: 2 December 2021).
7. GeeksforGeeks, (2021), 'Naive Bayes Classifiers - GeeksforGeeks', Available at: Naive Bayes Classifiers - GeeksforGeeks (Accessed: January 2022)
8. Chakure, Afroz., (2019) 'Decision Tree Classification', Available at: Decision Tree Classification. A Decision Tree is a simple… | by Afroz Chakure | The Startup | Medium (Accessed :10 December 2021).
9. New York OpenData , Motor Vehicle Collisions-Crashes, Available at: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95, Accesed last on :-4-Jan-2022
10. Zack,(2019),'A Complete Guide to Stepwise Regression in R',Available at :https://www.statology.org/stepwise-regression-r/, Accessed on :4-Jan-2022
11. 'Zack,(2019),'A Complete Guide to Stepwise Regression in R',Available at :https://www.statology.org/stepwise-regression-r/, Accessed on :4-Jan-2022

10 Appendix

Fig. 1.



```
'data.frame':   1852417 obs. of  29 variables:
 $ CRASH.DATE                   : chr  "04/14/2021" "04/13/2021" "04/15/2021" "04/13/2021" ...
 $ CRASH.TIME                   : chr  "5:32" "21:35" "16:15" "16:00" ...
 $ BOROUGH                      : chr  "" "BROOKLYN" "" "BROOKLYN" ...
 $ ZIP.CODE                     : int  NA 11217 NA 11222 NA NA 11106 NA NA NA ...
 $ LATITUDE                     : num  NA 40.7 NA NA 0 ...
 $ LONGITUDE                    : num  NA -74 NA NA 0 ...
 $ LOCATION                     : chr  "" "(40.68358, -73.97617)" "" "" ...
 $ ON.STREET.NAME               : chr  "BRONX WHITESTONE BRIDGE" "" "HUTCHINSON RIVER PARKWAY" "VANDERVORT AVENUE" ...
 $ CROSS.STREET.NAME            : chr  "" "" "" "ANTHONY STREET" ...
 $ OFF.STREET.NAME              : chr  "" "620       ATLANTIC AVENUE                " "" "" ...
 $ NUMBER.OF.PERSONS.INJURED    : int  0 1 0 0 0 0 0 0 1 0 ...
 $ NUMBER.OF.PERSONS.KILLED     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NUMBER.OF.PEDESTRIANS.INJURED: int  0 1 0 0 0 0 0 0 0 0 ...
 $ NUMBER.OF.PEDESTRIANS.KILLED : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NUMBER.OF.CYCLIST.INJURED    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NUMBER.OF.CYCLIST.KILLED     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ NUMBER.OF.MOTORIST.INJURED   : int  0 0 0 0 0 0 0 0 1 0 ...
 $ NUMBER.OF.MOTORIST.KILLED    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CONTRIBUTING.FACTOR.VEHICLE.1: chr  "Following Too Closely" "Unspecified" "Pavement Slippery" "Following Too Closely" ...
 $ CONTRIBUTING.FACTOR.VEHICLE.2: chr  "Unspecified" "" "" "Unspecified" ...
 $ CONTRIBUTING.FACTOR.VEHICLE.3: chr  "" "" "" "" ...
 $ CONTRIBUTING.FACTOR.VEHICLE.4: chr  "" "" "" "" ...
 $ CONTRIBUTING.FACTOR.VEHICLE.5: chr  "" "" "" "" ...
 $ COLLISION_ID                 : int  4407480 4407147 4407665 4407811 4406885 4407883 4408019 4408060 4406314 4408149 ...
 $ VEHICLE.TYPE.CODE.1          : chr  "Sedan" "Sedan" "Station Wagon/Sport Utility Vehicle" "Sedan" ...
 $ VEHICLE.TYPE.CODE.2          : chr  "Sedan" "" "" "" ...
 $ VEHICLE.TYPE.CODE.3          : chr  "" "" "" "" ...
 $ VEHICLE.TYPE.CODE.4          : chr  "" "" "" "" ...
 $ VEHICLE.TYPE.CODE.5          : chr  "" "" "" "" ...
```

Fig. 2.

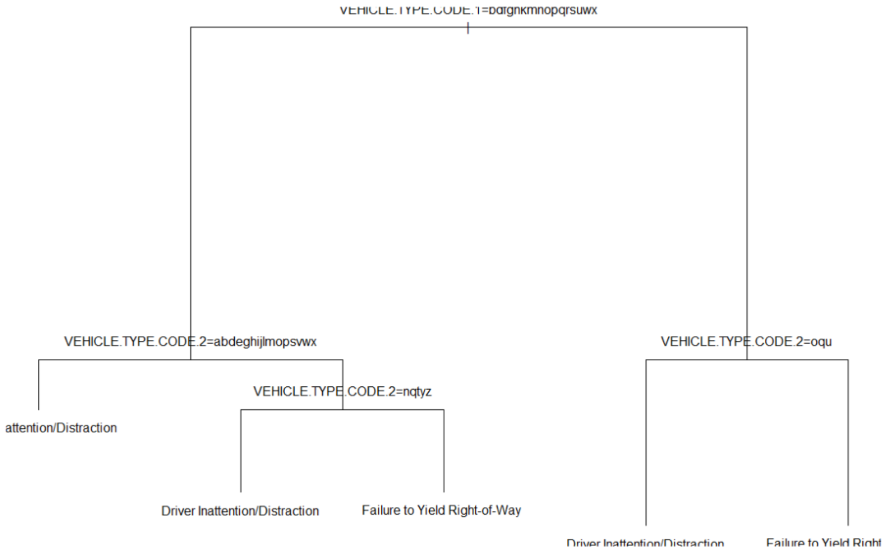

VEHICLE.TYPE.CODE.1=bdfghkmnopqrsuwx
VEHICLE.TYPE.CODE.2=abdeghijlmopsvwx
VEHICLE.TYPE.CODE.2=oqu
VEHICLE.TYPE.CODE.2=nqtyz
attention/Distraction
Driver Inattention/Distraction
Failure to Yield Right-of-Way
Driver Inattention/Distraction
Failure to Yield Right

Fig. 3.

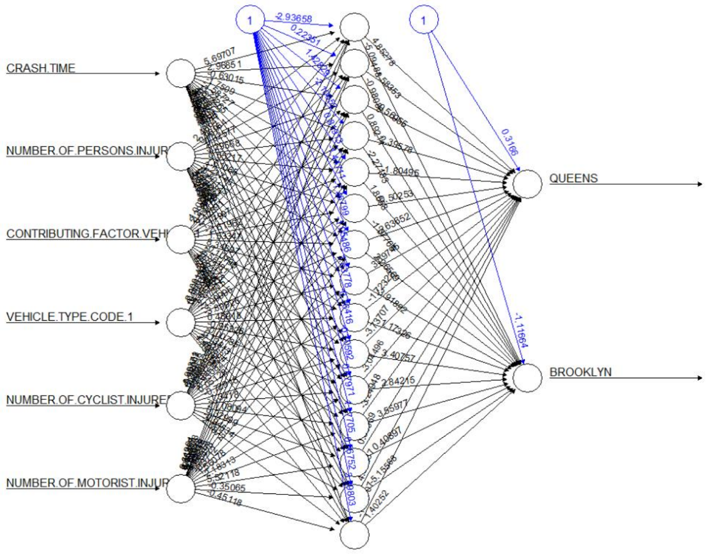
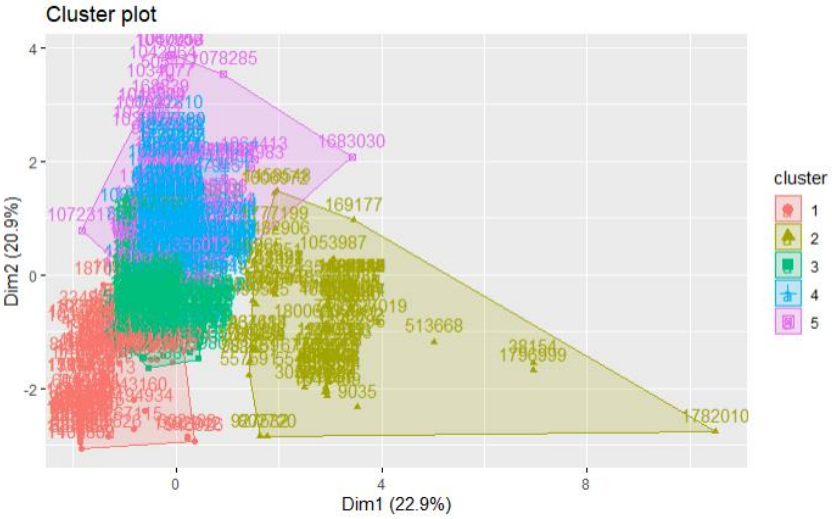

CRASH.TIME
NUMBER.OF.PERSONS.INJUR
CONTRIBUTING.FACTOR.VEH
VEHICLE.TYPE.CODE.1
NUMBER.OF.CYCLIST.INJURE
NUMBER.OF.MOTORIST.INJUR
1
1
-2.93658
0.3166
-1.11664
QUEENS
BROOKLYN

Fig. 4.


Cluster plot
Dim2 (20.9%)
Dim1 (22.9%)
4
2
0
-2
0
4
8
cluster
1
2
3
4
5
1683030
1782010
513668
169177
1072318