EXAMINATION ASSOCIATION OF ALTERNATIVE CRASH TYPES IN CALIFORNIA

Seyedbamdad Sharifiilierdy, College of Engineering, California State Polytechnic University, Pomona, CA 91768, 949-519-9846, seyedbamdads@cpp.edu
Wen Cheng, College of Engineering, California State Polytechnic University, Pomona, CA 91768 909-869-2957, wcheng@cpp.edu
Yasser Salem, College of Engineering, California State Polytechnic University, Pomona, CA 91768 909-869-4312, ysalem@cpp.edu

ABSTRACT

In the current traffic safety, considerably less studies have been dedicated to evaluating the association of different crash types. This paper aims to assess the fill the gap using seven city-level crash types including alcohol, bicyclists, hit and run, nighttime, pedestrians, speed, and motorcyclists. The goal is achieved through a Pearson's correlation coefficient test along with k-means clustering algorithm. The findings of the paper can assist city planners, policymakers, and pertinent transportation and local agencies in improving safety levels and establishing adequate mitigation measures.

Keywords: Crash Types; Pearson's Correlation; K-means Clustering; Silhouette Method.

INTRODUCTION

For the past several years, road traffic crashes have endured at a high rate taking millions of lives around the world. According to World Health Organization (WHO), approximately 1.3 million people lose their lives, and in addition between 20 to 50 million people experience non-fatal injuries as a result of traffic accidents (Road Traffic Injuries, 2022). In the United State, in addition to the noticeable rate of traffic collisions in metropolitan areas, the statistics suggest a considerable number of crashes nationwide through the first nine months of 2021 (NHTSA, 2022). Many mitigation plans are currently being implemented. The National Roadway Safety Strategy (NRSS) which is supported by President Joe Biden's Infrastructure Bill aims to undertake mitigation efforts to enhance road safety (U.S. Transportation, 2022). Mitigating traffic injuries, Los Angeles County has designed a program, named Vision Zero, to eliminate all fatal accidents by a planned year (Vision Zero, 2019). Concerning the nationwide increase in traffic accidents and related fatalities as well as injuries, supportive research studies are of vital importance to help local transportation agencies improving general safety levels.

Many studies have been conducted to explore the important factors and behaviors causing traffic incidents. Siddiqui et al. (2012) have conducted a macroscopic spatial analysis on a dataset at district level in the state of Florida to inspect pedestrian and bicycle crash components, the result of which indicates the significant difference in predictor variables for pedestrian and bicycle accidents (Siddiqui et al., 2012). Another study was undertaken focusing on the relationship between vehicle types and speed variations using traffic and collision data of urban expressways. The findings of the study reveals that the crash risk of all collision types increases as the average speed of the traffic flow increases (Wang et al., 2022).

Moreover, a combination of convolution neural network (CNN) and gated recurrent units (GRU) using city-level traffic enforcement data was utilized to explore efficient methods for predicting at-fault crash

driver frequency and city-level crash risk. The results of this study indicate that effective methods for predicting, and in addition enhancing safety level methods can be generated (Wu and Hsu, 2021). Another study, using negative binomial regression model, attempted to analyze the relationship between bicycle usage and the number of bike accidents at local level. The findings show that bike collisions increase as cycling usage increases (Yao and Loo, 2016). However, this paper explores the correlation and optimal number of clusters for a city-level dataset by taking advantage of Pearson's correlation test and k-means clustering Silhouette method.

A city-level dataset containing seven crash types which includes accidents related to alcohol, bicyclists, hit and run, nighttime, pedestrians, speed, and motorcyclists, was used for the study. Considering a dataset that is comprised of continuous variables, Pearson's correlation test was chosen to discover the relationship between pairs of crash types. Through Pearson's correlation test, also the positive and negative relationships between variables can be found (Wijaya, 2021). Additionally, k-mean clustering in unsupervised machine learning employing Silhouette method was done to calculate the best number of groups. The silhouette method was preferred to the Elbow method for clustering measures, since the former method considers more factors such as range differences and variances as it assesses the variables (Kumar, 2021). The findings of the paper can support local and transportation related agencies to establish schemes with the aim of improving safety and mitigating crash risks.

DATA DESCRIPTION

The data used for this study was collected from the California Highway Patrol Collision Report (2018). Some redundant variables from the original data, which were not necessary, were omitted from the data, such as average population, city and county ID's, as well as driver vehicle miles traveled (DVMT). The dataset used for the study includes the total number of collisions for the seven crash categories. The crash categories consist of accidents involving alcohol, bicyclists, hit and run, nighttime, pedestrians, speed, and motorcyclists. This set of data is valuable in conducting Pearson's Comprehensive Correlation and K-means clustering in Unsupervised Machine Learning. Previous studies rarely used this kind of dataset including city-level data to perform crash analysis. However, considering the absolute need for safety enhancements at local levels, this paper uses city-level data with seven variables. Some data outliers and missing values were omitted from the dataset in advance of analyzing to insure more precise results and to prevent data alteration. Several data outliers including small and big cities like Dorris, and Los Angeles in California were removed from the dataset using a filtration method which will be discussed further. Table 1 shows a summary of the original dataset before filtering. As shown in Table 1, speed related incidents have the largest maximum count, and incidents involving bikes account for the lowest count. Other measures like mean and median can also be found in Table 1.

Table 1: Summary of Statistics of the Data before Filtration

	Alcohol	Bike	HnR	Night	Ped	Speed	MC
Minimum	0	0	0	0	0	0	0
1st. Quartile	2	1	1	2	2	3	1
Median	12.0	6.0	6.0	11.0	8.0	18.5	4.0
Mean	81.42	35.53	63.10	78.09	52.31	154.91	36.11
3rd Quartile	39	18	22	33	25.25	61.25	15
Maximum	7271	3389	8096	8728	5931	17609	3971

Note: HnR and MC represent hit and run, and motorcycle, respectively. Ped stands for pedestrian.

To obtain a clean dataset and to remove data outliers, a statistical measure of dispersion was used in Microsoft Excel. In addition to quartiles 1 (Q1) and 3 (Q3) for each of the seven categories which was mentioned earlier, the interquartile ranges (IQRs) were calculated. As illustrated in Table 2, motorcyclist and speed had the lowest and highest value of IQR compared to the other categories, respectively.

Table 2: Statistical Measure of Dispersion Results

	Alcohol	Bike	HnR	Night	Ped	Speed	MC
Q1	2	1	1	2	2	3	1
Q3	39	18	22	33	25.25	61.25	15
IQR	37	17	21	31	23.25	58.25	14

Note: Q1 and Q3 represent 1st and 2nd quartiles, respectively. IQR is interquartile range. HnR and MC represent hit and run, and motorcycle, respectively. Ped stands for pedestrian.

Following this measurement, data points that fall outside the defined range were removed from the dataset. The desirable extent of the data points ranges from Q1-1.5*IQR to Q3+1.5*IQR. This filtration was conducted in Microsoft Excel using a pass or fail logical test. All seven collision categories were adapted to meet the criteria for this study. The original dataset included 540 cities which was filtered and lowered into 449 cities meeting the requirements. Table 3 describes a summary of the logical test conducted.

Table 3: Logical Test Results in Microsoft Excel

	Alcohol	Bike	HnR	Night	Ped	Speed	MC
Fail	60	60	63	59	51	57	54
Pass	480	480	477	481	489	483	486
Total Passing	449						

Note: HnR and MC represent hit and run, and motorcycle, respectively. Ped stands for pedestrian.

Lastly, a summary of the final dataset is shown in Table 4. As illustrated, collisions involving speed have the highest maximum value of 138, and all categories have the minimum value of zero. In addition, speed related collisions have the highest value mean and median, while motorcycle collisions account for the lowest value for both median and mean. It can be identified that speed related collisions are more prevalent than the other categories, and motorcycle collisions have the lowest frequency.

Table 4: Table 1: Summary of Statistics of the Final Data

	Alcohol	Bike	HnR	Night	Ped	Speed	MC
Minimum	0	0	0	0	0	0	0
1st. Quartile	2	1	0	1	1	2	0
Median	9	4	3	6	5	13	3
Mean	15.38	7.11	8.12	12.77	9.66	23.24	5.75
3rd Quartile	23	11	12	20	14	34	9
Maximum	91	43	53	75	56	138	36

Note: HnR and MC represent hit and run, and motorcycle, respectively. Ped stands for pedestrian.

METHODOLOGY

Pearson's comprehensive correlation method and K-means in unsupervised machine learning utilizing Silhouette method were conducted using final, filtered data in R programming software.

Pearson's Correlation

Pearson's comprehensive correlation method can be utilized to evaluate the linear relation between the pairs of variables. Using this method reveals that whether two variables are statistically significantly correlated. The correlation coefficient value, r, for this test ranges from -1 to +1. Coefficient value of -1 indicates perfect negative correlation, while +1 reveals perfect positive correlation between variables (SPSS, 2022). The Pearson's correlation coefficient, r, based on two continuous variables can be calculated using equation (1).

$$R = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}} \quad (1)$$

All seven categories were considered as continuous variables. Each possible pair of collision was tested in R resulting in a 7 by 7 matrix. In the mentioned matrix, there exist 21 specific collision combinations among all pairs.

K-means in Unsupervised Machine Learning Silhouette Method

K-means clustering is a popular method to categorize data points into k groups. Considering this, the Silhouette method is a commonly used method to optimize the performance of the test by defining the ideal number of cluster groups. Common purposes of using the Silhouette method include validation and interpretation of the data points in each cluster or category (Kumar, 2021). This method assigns a silhouette coefficient to each data point recognizing the similarity between the data points within each category. The value of silhouette coefficient ranges from -1 to +1. Values near zero indicates overlapping clusters, and coefficient value of 1 shows perfect compatibility with the cluster (Banerji, 2021). Using equation (2), the silhouette coefficient can be calculated.

$$\big(S(i)\big) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}} \quad (2)$$

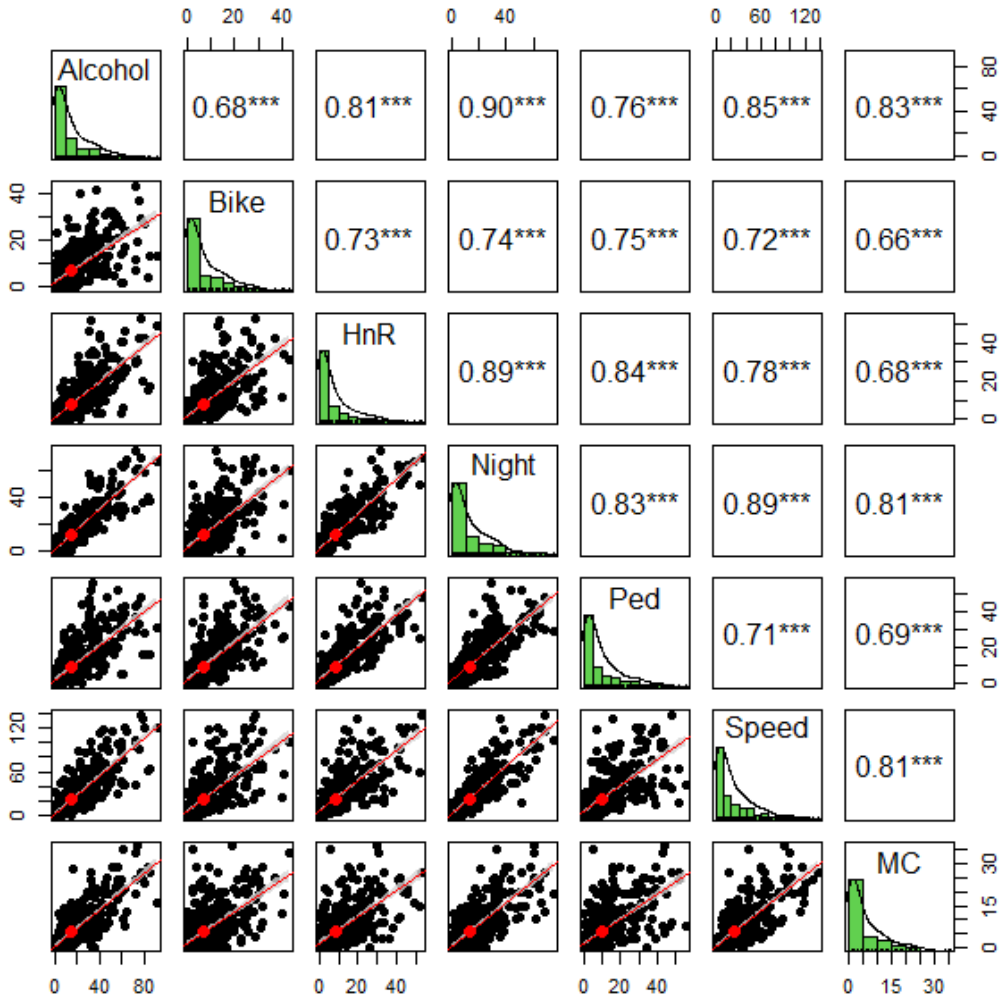
In equation (2), S(i) represents the silhouette coefficient for i^{th} point, b(i) and a(i) denote the average distance of point i with all other points in the same cluster and the other clusters that does not belong to point i, respectively.

RESULTS

The main purpose of this study is to assess collision data at city-level which contains seven crash factors including speed, alcohol, pedestrians, bicyclists, hit and run, nighttime, and motorcyclists. The evaluation was carried out utilizing Pearson's correlation coefficient method and k-means clustering in unsupervised machine learning Silhouette method. The data used in this study was collected from the California Highway Patrol 2018 Collision Report. To achieve more precise results, the raw data was filtered, and some data outliers were removed using a statistical measure of dispersion in Microsoft Excel.

Following the data modification, Pearson's Comprehensive Correlation test was conducted and resulted in 21 graphs and correlation coefficients for each collision pair. As illustrated in Figure 1, a robust, positive correlation between all crash factor pairs is found. The red line (best fit line) describes the relationship between the two variables. The correlation coefficient values for all possible combinations of collision pairs are greater or equal to 0.66 revealing that collisions are correlated with at least two variables among seven crash types. One possible reason for this result is that it may not be doable for drivers to handle more than two collision types. Nighttime and alcohol related accidents account for the highest correction coefficient of 0.90. In addition, it is worth mentioning that collisions happened at nighttime which were associated with speed or hit and run factors demonstrate a high coefficient of 0.89. The lowest coefficient belongs to the bike accidents associated with motorcycles.

Figure 1: Pearson Correlation Test Results in R



Note: HnR and MC represent hit and run, and motorcycle, respectively. Ped stands for pedestrian.

Finally, a k-means clustering using the Silhouette method was performed in R to identify the optimal number of clusters for this study. Through the Silhouette method, average silhouette widths were calculated for different number of clusters. Table 5 and Figure 2 illustrate the average silhouette widths for different values of clusters. As shown in the table and graph, the average silhouette coefficient

decreases as the number of groups increase. The average silhouette coefficient for the highest number of clusters, 10, has a low value of 0.29. However, the cluster value of 2 has the highest silhouette score among all combinations representing the optimal number of clusters for the dataset used in this study.

Figure 2: Number of Clusters (k) vs. Average Silhouette Width scatterplot

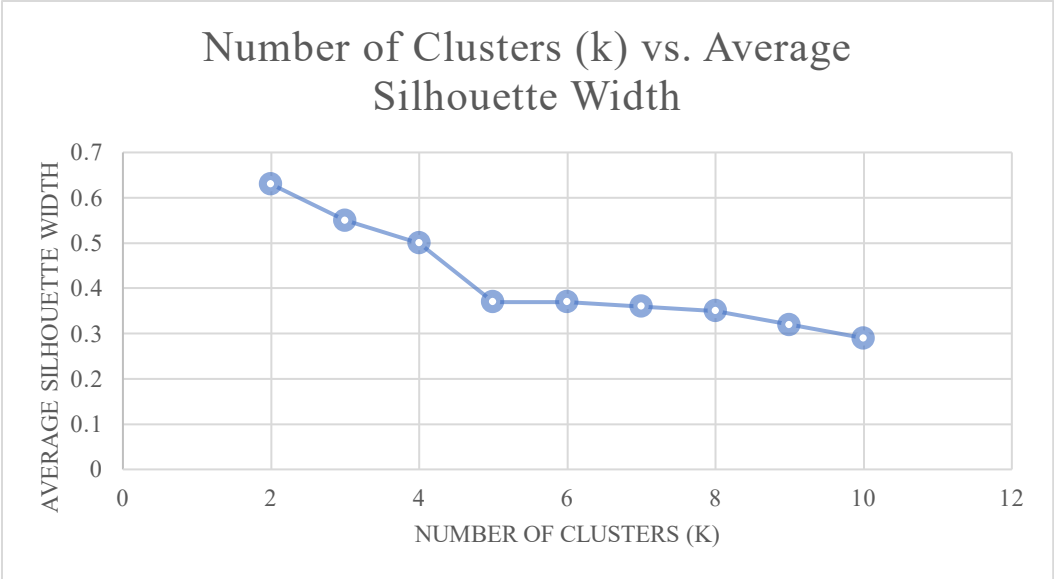


Table 5: the Silhouette Method Results Summary

Number of Clusters (k)	Silhouette Coefficients
2	0.63
3	0.55
4	0.50
5	0.37
6	0.37
7	0.36
8	0.35
9	0.32
10	0.29

CONCLUSION

Considering the constant high rate of traffic accidents across the nation causing fatal or serious injuries each year, the authors conducted research on evaluating seven crash types including speed, alcohol, pedestrians, bicyclists, hit and run, nighttime, and motorcyclists. First, through Pearson's correlation coefficient test, the linear relationship between pairs of variables were calculated. Then, k-mean clustering in unsupervised machine learning using Silhouette method employed to calculate the optimal number of clusters for the dataset. The data used for the study was gathered from the California Highway Patrol Collision Report (2018) followed by a data filtration. The findings show that more than one crash types are involved in the accidents. This can be as a consequence of humans' driving behavior limitation in focusing on more than one driver task or distraction simultaneously. The results can aid city planners and local transportation agencies to improve safety management and mitigation measures to reach the least number of accidents. The following recommendations can be considered in the future: (1)

future manufactured vehicles should have breathalyzer feature to measure the amount of alcohol prior to starting the engine. This test can be obligatory during nighttime, which possibly mitigate the number of nighttime and alcohol related incidents; and (2) dash cameras and sensors must be implemented in the future vehicles. Tracking and finding hit and run vehicles can become easier for highway patrol because of the possibility of tracking license plate numbers. This feature could help reducing hit and run involved accidents.

Although the findings uncover useful information, the study suffers from some limitations. Further studies are recommended to address the following limitations: (1) crash severity types are not involved in the current dataset; and (2) further comparison can be made by employing both supervised and unsupervised clustering.

REFERENCES

Banerji, Ankita. "K-Mean: Getting The Optimal Number Of Clusters." Analyticsvidhya.com, May 18, 2021, Retrieved October 13, 2022, URL: https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/#:~:text=Silhouette%20analysis,for%20each%20value%20of%20K.

California Highway Patrol 2018 Collision Report. State of California Highway Patrol.

Kumar, Satyam. "Elbow Method vs Silhouette Score – Which is Better?" vital flux, 28 Nov. 2021, Retrieved October 15, https://vitalflux.com/elbow-method-silhouette-score-which-better/.

Kumar, Satyam. "Silhouette Method-Better than Elbow Method to Find Optimal Clusters." Medium, Towards Data Science, 21 Sept. 2021, Retrieved October 15, 2022, https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891.

"NHTSA Estimates Traffic Fatalities Continued to Rise at Record Pace in First Nine Months of 2021." U.S. Department of Transportation, February 1, 2022, Retrieved October 14, 2022, https://www.transportation.gov/briefing-room/nhtsa-estimates-traffic-fatalities-continued-rise-record-pace-first-nine-months-2021.

"Road Traffic Injuries." World Health Organization, World Health Organization, June 20, 2022, Retrieved October 14, 2022. URL: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

Siddiqui, Chowdhury, et al. "Macroscopic Spatial Analysis of Pedestrian and Bicycle Crashes." Accident Analysis & Prevention, vol. 45, 2012, pp. 382–391., https://doi.org/10.1016/j.aap.2011.08.003.

"SPSS Tutorials: Pearson Correlation." LibGuides, 2022, Retrieved October 13, https://libguides.library.kent.edu/spss/pearsoncorr.

"U.S. Transportation Secretary Pete Buttigieg Announces Comprehensive National Roadway Safety Strategy." U.S. Department of Transportation, January 27, 2022, Retrieved October 14, 2022,

https://www.transportation.gov/briefing-room/us-transportation-secretary-pete-buttigieg-announces-comprehensive-national-roadway.

"Vision Zero, Los Angeles County A Plan for Safer Roadways 2020 - 2025." Los Angeles County. 2019, Retrieved October 14, 2022. https://pw.lacounty.gov/visionzero/.

Wang, Chen, et al. "Impacts of Real-Time Traffic State on Urban Expressway Crashes by Collision and Vehicle Type." Sustainability, vol. 14, no. 4, 2022, p. 2238., https://doi.org/10.3390/su14042238.

Wijaya, Cornellius Yudha. "What It Takes to Be Correlated." Medium, Towards Data Science, 12 Oct. 2021, Retrieved October 15, 2022, https://towardsdatascience.com/what-it-takes-to-be-correlated-ce41ad0d8d7f.

Wu, Yuan-Wei, and Tien-Pen Hsu. "Mid-Term Prediction of at-Fault Crash Driver Frequency Using Fusion Deep Learning with City-Level Traffic Violation Data." Accident Analysis & Prevention, vol. 150, 2021, p. 105910., https://doi.org/10.1016/j.aap.2020.105910.

Yao, Shenjun, and Becky P Loo. "Safety in Numbers for Cyclists beyond National-Level and City-Level Data: A Study on the Non-Linearity of Risk within the City of Hong Kong." Injury Prevention, vol. 22, no. 6, 2016, pp. 379–385., https://doi.org/10.1136/injuryprev-2016-041964.