THE CLUSTERING ALGORITHMS TO CLUSTER THE FACTORS OF GREAT BRITAIN ROAD TRAFFIC ACCIDENTS

TAN SHU ZHEN

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

STATUS CONFIRMATION FOR UNDERGRADUATE PROJECT REPORT

THE CLUSTERING ALGORITHMS TO CLUSTER THE FACTORS OF GREAT BRITAIN ROAD TRAFFIC ACCIDENTS

ACADEMIC SESSION 2016/2017

I, **TAN SHU ZHEN** agree to allow this Undergraduate Project Report to be kept at the Library under the following terms:

1. This Undergraduate Project Report is the property of the Universiti Tun Hussein Onn Malaysia.
2. The library has the right to make copies for educational purposes only.
3. The library is allowed to make copies of this report for educational exchange between higher educational institutions.
4. ** Please Mark (√)

□ CONFIDENTIAL (Contains information of high security or of great importance to Malaysia as STIPULATED under the OFFICIAL SECRET ACT 1972)

□ RESTRICTED (Contains restricted information as determined by the Organization/institution where research was conducted)

□ FREE ACCESS

Approved by,

(TAN SHU ZHEN)

Permanent Address:
303B, JALAN QUEK KAI KEE,
85000 SEGAMAT,
JOHOR.

(DR SABARIAH BINTI SAHARAN)

FACULTY OF SCIENCE,
TECHNOLOGY AND HUMAN
DEVELOPMENT, UNIVERSITI
TUN HUSSIEN ONN MALAYSIA

Date : ________________________

Date: ________________________

NOTE:** If this Undergraduate Project Report is classified as CONFIDENTIAL or RESTRICTED, please attach the letter from the relevant authority/organization stating reasons and duration for such classifications.

This Undergraduate Project Report has been examined on date 5th December 2016 and is sufficient in fulfilling the scope and quality for the purpose of awarding the degree of Bachelor of Science (Industrial Statistics) with Honour.

Examiners:

DR. ROHAYU BINTI MOHD SALLEH
Faculty of Science, Technology and Human Development
Universiti Tun Hussein Onn Malaysia

PN. NORHAIDAH BINTI MOHD ASRAH
Faculty of Science, Technology and Human Development
Universiti Tun Hussein Onn Malaysia

THE CLUSTERING ALGORITHMS TO CLUSTER THE FACTORS OF GREAT BRITAIN ROAD TRAFFIC ACCIDENTS

TAN SHU ZHEN

A final year project submitted in
fulfillment of the requirement for the award of the Degree of
Bachelor (Science) Industrial Statistics with Honour

Faculty of Science, Technology and Human Development
Universiti Tun Hussein Onn Malaysia

JANUARY 2017

I hereby declare that the work in this project report is my own except for quotations and summaries which have been duly acknowledged.

Student : ……………………………………………….

TAN SHU ZHEN

Date : ……………………………………………….

Supervisor : ………………………………………………

DR. SABARIAH BINTI SAHARAN

For my beloved parents and my supervisor, Dr Sabariah Saharan

ACKNOWLEDGEMENT

Firstly, I would like to give my deep gratitude to my beloved lecturer, Dr Sabariah Saharan, who gave me advices and helps in this project, especially when I could not make a progress. Secondly, my parents also deserve my gratitude because they always give me mental and economic supports to pursue my goals. Last but not least, I should thank my coursemates who were willing to give me a favor when I was in difficulty.

ABSTRACT

There was an unusual peak in November and December 2014 in terms of fatalities resulting from road traffic accident because normally most traffic accidents happened in summer. Since the dataset of road traffic accident of Great Britain was large, *K*-means and *K*-modes clustering methods were applied due to their efficiency of dealing with complex and huge dataset. Hence, in this study, both methods were adopted to partition the variables for June (summer) and December (winter) and investigate the combination of factors that led to road accident. The results showed that there were 8 fixed elements that were always assigned into same cluster, regardless of season as well as *K*-means and *K*-modes models.

ABSTRAK

Kematian kemalangan jalan raya pada bulan November and Disember 2014 (musim sejuk) menunjukkan puncak yang luar biasa kerana biasanya kebanyakan kemalangan jalan raya selalu berlaku pada musim panas. Disebabkan saiz data kemalangan jalan raya di Great Britain yang amat besar, kaedah *K*-min dan *K*-mod telah digunakan berdasarkan keberkesanan kedua-dua kaedah ini semasa mengendalikan data yang kompleks dan besar. Oleh yang demikian, dalam kajian ini, kedua-dua kaedah ini telah digunakan bagi mengelompokkan data pada bulan Jun (musim panas) dan Disember (musim sejuk) untuk mengenalpasti kombinasi faktor yang menyebabkan kemalangan jalan raya. Keputusan menunjukkan bahawa terdapat 8 elemen tetap yang sentiasa diberikan ke dalam kelompok yang sama, tanpa mengira musim, *K*-min dan *K*-mod model.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

In this chapter, background of the study and problem statement are introduced and discussed. These sections are followed by objectives and scope of study. The main objective of this chapter is to give a better understanding on the road traffic accident in Great Britain and realize the importance of finding the major factors of the problem using data mining algorithm.

1.1 Background of Study

A road traffic accident is defined as any vehicles accident occurring on a public highway. It includes collisions between vehicles and animals, vehicles and pedestrians, or vehicles and fixed obstacles. Single vehicle accidents, which involve a single vehicles that means without other road user, are also included (Ogwueleka *et al.*, 2014)

World Health Organization (WHO) (2013) stated that about 1.25 million people die in the road traffic crashes each year and people aged 15-29 years make up the 48% of global road traffic accidents. Furthermore, although there are 90% of deaths resulting from the accidents happen in low- and middle- income countries, people from lower socio-economic backgrounds in high-income countries are also more likely to be involved in the accidents. WHO (2013) also stated that if there is no any action taken, road traffic accidents are projected to increase to become 7th leading cause of death by 2030. Besides, road accidents give influential impact on the economy and well-being of the society in the countries as WHO estimated that road incidents cost the countries approximately 3% of their gross national product.

Likewise, Great Britain government has given a big concern on the road traffic accidents. As an evidence, STATS19 road accident data set (Administrative Data Liaison Service, 2014) was built to record the road accidents on the public highway in Great Britain. Police fill the STATS19 report form at either the scene of accident or when the accident is reported. This dataset provides the detailed information of the road traffic incidents, including accident circumstances, vehicles involved and resulting casualties and can be used to conduct research and formulate sustainable policies. Besides, Reported Road Casualties Great Britain (RRCGB), which is the official statistical publication of the Department of Transport, is based on the STATS19 data. Nevertheless, the contributory factors are largely subjective, reflecting the opinion of the reporting police officer. Evidence may be unavailable after the event. Therefore, some factors are impossible to be recorded on STATS19.

Furthermore, a 5-year road safety strategy for 2011-2015 was launched by Great Britain government on 11 May 2011 and the overall number of fatalities in 2012 decreased about 38%, compared with the 2005-2009 average. Hence, Great Britain has been one of the five European countries including Denmark, Norway, Sweden, and Iceland that managed to reduce their annual road fatalities (International Traffic Safety Data and Analysis Group, 2014).

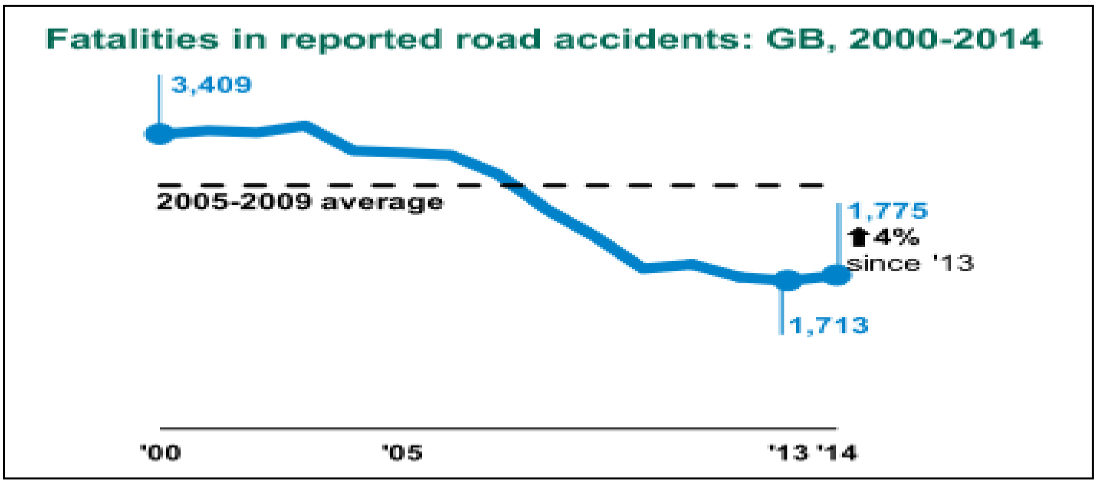


Figure 1.1: Fatalities in Reported Road Accidents, Great Britain from 2000 to 2014. (Department for Transport, 2015)

In spite of this, Figure 1.1 from Department for Transport(2015) demonstrates that there were 1,775 reported death rate in 2014, which increased 4% compared with 2013. Although it was the third lowest value of death rate after 2012 and 2013, the number of people killed and injured was 194,477 and it was the first increase in road traffic accident toll since 1997. In fact, pedestrians were the most vulnerable road users, which accounted for 3 quarters of the rise in fatalities between 2013 and 2014. Likewise, International Traffic Safety Data and Analysis Group (IRTAD) (2014) reported that pedestrians are also the largest group of vulnerable road users in most countries and constitute 19% of all fatalities in IRTAD countries, including Great Britain.

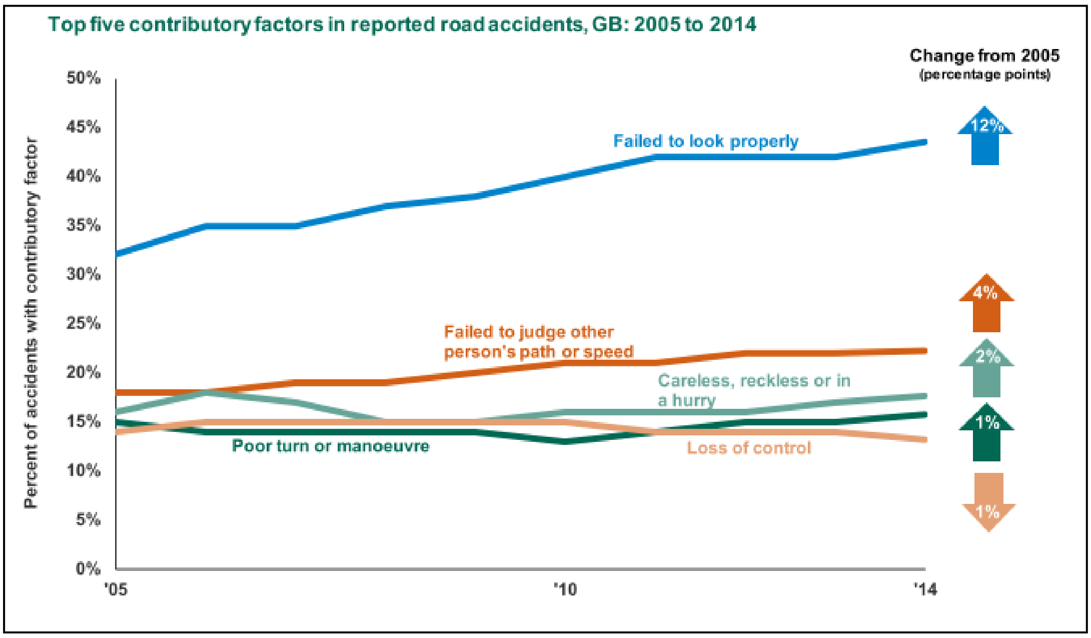


Figure 1.2: Top 5 Contributory Factors in Reported Road Accidents, Great Britain from 2005 to 2014.(Department of Transport, 2015)

In Figure 1.2, Department of Transport(2015)shows that the top five contributory factors to the road traffic accident, from 2005 to 2014 were “failed to look properly”, “failed to judge other person’s path or speed”, “careless, reckless or in a hurry”, “loss of control”, “poor turn or manoeuvre”. “failed to look properly” occupied 44% of the contributory factors to road traffic accidents in 2014 and has remained as the most frequently reported contributory factors since 2005.

Table 1.1: Reported Casualties by Month and Severity, Great Britain, 2014. (Department for Transport, 2015)

	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
Killed	128	117	131	140	128	160	153	146	158	145	170	199
Killed or Seriously Injured	1901	1673	1970	1938	2033	2262	2272	2101	2114	2212	2111	1990
All Severities	15826	14267	15800	14844	16464	16715	17412	16610	15442	17712	17408	15972

In addition, RRCGB Annual Report 2014 (Department for Transport, 2015)reported that the end of the year, which are November and December 2014 were the worst months of the year in terms of fatalities. It was an unusual pattern because the peak time for fatalities is normally over the summer months, rather than the end of the year as shown in Table 1.1. Furthermore, as compared with 2013, it was the largest increase (27.56%) for the fatalities over December, among the rest of the months. However, the reports revealed that there is no clear reason for this pattern. It guessed that it might be due to the warmer temperature in December than 30-year long term average, covering 1981 to 2010 because road users tend to be careful during bad weather and it results in a reduction of road traffic accidents, while good weather encourages people to travel and always has an effect of increasing casualties.

However, when working on huge data, simple statistical analysis has no significant evidence to show all valuable information of a given data (Tesema *et al.*, 2005). Hence, data mining algorithms have been applied to traffic accident data analysis. For example, data segmentation, such as *K*-modes clustering has been applied widely to handle the heterogeneity of the accident data (Kumar and Toshniwal, 2015). Moreover, datasets of traffic accidents in Great Britain are able to be accessed openly, with the Open Government License. Therefore, since an entire population of data is available, the notions of sampling and hypothesis testing lose their point (Hand, 1999). Furthermore, the basic assumption of this study is that the road traffic accidents are not distributed randomly (Beshah and Hill, 2010). Hence, data mining techniques are applied on the datasets instead, with all those reasons.

Besides, with the aim of dealing with categorical dataset effectively, Huang (1997) proposed *K*-modes clustering algorithm after *K*-prototypes which is adopted to handle the datasets with mixed numeric and categorical values. *K*-modes clustering technique is a modified-version of *K*-means due to the limitation of *K*-means on dealing with categorical datasets. Hence, its feature is to cluster the datasets on categorical scale by using its new dissimilarity measures as well as to give characteristic descriptions of clusters when interpreting the clusters. Besides, it inherits the characteristics of *K*-means clustering, including its advantages and disadvantages.

In this study, the *K*-modes and *K*-means clustering algorithms were applied to cluster the factors of Great Britain road traffic accidents, in terms of similarity. Instead of the most important single factor, the combinations of factors that lead to road traffic accident were investigated.

1.2 Problem Statement

Department for Transport (2015) has reported that there was a 4% increase in number of reported road deaths in 2014, compared with 2013. Besides, there were unusual peaks on the end of the year, particularly in November and December compared to 2013 because the worst season for road traffic accidents normally happened in summer. Moreover, there was a 27.56% increase in terms of fatalities over December 2014, as compared with 2013. It was the largest change among the rest of the months in 2014. However, Department of Transport has not discovered the reasons behind these unusual patterns. Thus, the intention of this study is to discover the causes of road fatalities in summer and winter.

On the other hand, large road traffic accident dataset that consist of high heterogeneity have made the analysis hard to discover the hidden message and important relationships among variables (Kumar and Toshniwal, 2015). Besides, Department of Transport (2015) neglected the fact of the limitation of multiple-way contingency tables applied to the high-dimensional dataset to identify the interaction between the factors. Hence, data mining algorithms are one of the methods to deal with large and complex datasets due to its scalability.

1.3 Objectives

Thus, in this study, the objectives are shown below:-

(i) To determine the combinations of the causes that led to road traffic accident during summer and winter in 2014, by using *K*-means algorithm.

(ii) To identify the combinations of the causes that led to road traffic accident during summer and winter in 2014, by using *K*-modes algorithm.

(iii) To study the differences of results generated by *K*-means and *K*-modes.

1.4 Scope of Study

There are three types of datasets, namely accident circumstances, vehicle and casualties in the official website of United Kingdom government (Data.gov.uk, 2015). However, only dataset of accident circumstances at Great Britain in June and December 2014 were selected as representation of summer and winter, respectively in the study, since the whole-year dataset cannot be analyzed due to its large total number of data objects, more time required to run the iteration by using standard computer processor.

June and December 2014 were selected based on the purpose of studying the causes of deaths resulting from road traffic accidents. The reasons are that June 2014 was the month that had the largest number of fatalities as compared with the months during summer, July and August and the only month that experienced the increase in term of fatalities, while the rest decreased. On the other hand, December 2014 was chosen due to its unusual peak. Besides, it was the month that had the largest increase in term of fatalities among the months in 2014 (Department for Transport, 2015). Thus, the results for both seasons can be compared whether there are any grouping differences of the road traffic accident factors. Hence, there are 12532 and 12036 reported road traffic accidents for June and December 2014, respectively, in Great Britain.

Moreover, the variables in the dataset were selected in line with the objectives of the study. Variable that has too many missing data or observation with unknown information on the variables were excluded due to the disadvantages of the methods that cannot deal with the missing values. Thus, this disadvantage can be furthered up for the future research. Lastly, there were 14 categorical variables to be included in the study.

1.5 Conclusion

Although Great Britain have reduced successfully the occurrence of road traffic accident, Reported Road Casualties Great Britain (RRCGB) showed that there were unusual peaks on November and December in 2014 compared to 2013. The reasons have not been revealed yet. In this study, *K*-modes and *K*-means clustering algorithm are introduced to discover the leading factors of the traffic road accidents with the categorical dataset.

CHAPTER 2

LITERATURE REVIEW

Various modifications have been done on the *K*-means clustering algorithm so that the modified algorithms have the ability to scale well with the large data sets in categorical scale and the power to recover the underlying structure of the data. In this chapter, extension to *K*-means clustering algorithm, namely *K*-modes and various kind of clustering methods that are able to handle the datasets in nominal scale are explained. Besides, various studies on the road traffic accident data analysis using data mining techniques are also introduced in this chapter.

2.1 Cluster Analysis on Categorical Data

Data mining is used to uncover the unsuspected and hidden information from a large datasets. From the view of statistics, it can be treated as computer automated exploratory data analysis of complex large dataset. Besides, it intersects with fields, such as Data Base Management, Artificial Intelligence, Machine Learning, Pattern Recognition as well as Data Visualization (Friedman, 1997).

The distinct characteristic of the data mining is the ability to deal with large datasets. Hence, the scalability of the algorithm is always the concern of data mining researches. However, most algorithms do not scale well with the large datasets because they were initially developed for other fields in which small datasets are involved (Huang, 1997). This case also occurs in statistical methods. In particular, traditional statistical algorithms are too slow for data mining when dealing with large datasets which are huge by statistical standards (Hand, 1999).

Clustering analysis is one of the statistical analysis methods offered by data mining packages, apart from decision tree induction, rule induction, nearest neighbors, association rules, feature extraction and visualization (Friedman, 1997). Clustering is widely used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics (Madhulatha, 2012). It deals with unsupervised learning problems when the classes in the data are not already known and the training data is not available (Gupta, 2011). It groups the data into cluster in which the object are similar between each other and dissimilar to the object belonging to other clusters according to the distance measure (Madhulatha, 2012).

K-means clustering is a partitional clustering that is suitable for data mining because of its efficiency in clustering large datasets(Huang, 1997). Nevertheless, this algorithm is limited to the numeric data because the categorical data also account for part of real-world data. This drawback has drawn the attentions among the scholars. Clustering categorical data with *K*-means clustering was developed by Ralambondrainy (1995). This algorithm consists of converting categorical data with multiple attribute values into binary attributes and treating them as numeric data in the process. However, it was criticized by Huang (1997) who stated that it was not practical when handling datasets of categorical data with hundreds or thousands of categories. Besides, the cluster means which are 0 or 1, could not represent the characteristics of the clusters (Huang, 1997).

Thus, Huang (1997) has proposed the modified-version of *K*-means clustering, named *K*-modes clustering. Literally, this method replaces means of the clusters with modes. The other modifications are introduction of new dissimilarity measure and frequency-based method to update modes in the process. The new dissimilarity measure is the simplification of the dissimilarity measure from *K*-prototype by only considering the categorical attributes. The major advantage of *K*-modes clustering is that it scales well with the large datasets due to its parentage of *K*-means clustering. Besides, Huang (1997) stated that it is faster than *K*-means because less iteration to converge is needed. In contrast with means, the modes generate characteristic descriptions of cluster when working on categorical data.

Besides, latent class models are applied for cluster analysis of polytomous variables with the assumption that the observed variables are mutually independent given the class variable(Zhang, 2004). Furthermore, Chaturvedi *et al.*(2001) have conducted Monte Carlo simulation to compare the performance of latent class clustering with that of *K*-modes clustering. The results revealed that both of them have the equal ability of recovering a known underlying cluster structure. However, *K*-modes is faster in speed and suffers less to local optima than latent class clustering does. Besides, latent class clustering becomes computationally slow and infeasible when dealing with the datasets consisting of numerous categorical variables. Finally, Chaturvedi*et al*. (2001) suggested that these two algorithms to be complementary in performing cluster analysis.

Despite of this, Khan and Ahmad (2012) stated that *K*-modes clustering inherits the same drawbacks from its parent, *K*-means clustering. Similar to *K*-means, *K*-modes clustering assumes that the number of cluster, *k* is known in advance. Furthermore, *K*-modes clustering also cannot handle non-globular data of different sizes and densities. Another drawback is that it does not guarantee unique clustering owing to random initialization of cluster centers because it may generate different groupings for each run. Last but not least, these two algorithms are sensitive dependent on the choice of initial centers.

Hence, Khan and Ahmad (2012) proposed multiple attribute clustering to initialize cluster center based on the rationales that some of the objects do not change cluster membership, disregarding the choice of initial cluster centers and individual categorical variables with few categories may provide information about the cluster structures. Hence, prominent attributes are introduced. The features of prominent attributes are that number of attributes are less than or equal to *k* and these attributes shall have higher differentiating power. Therefore, prominent attributes are believed to be able to initialize the modes. The datasets are then divided iteratively based on the attribute values of prominent attributes. The results of the proposed algorithm revealed that repeatable and better cluster structures can be obtained and fixed cluster centers can be generated.

Likewise, He (2006) also related the sensitivity to the initial centers to the reliability and accuracy of clustering results. Hence, he proposed farthest-point heuristics based initialization methods for *K*-modes clustering algorithm. Besides, He (2006) also stated that non-randomized initialization algorithm leading to non-repeatable clustering results is much desirable for clustering. However, random initialization is widely used for *K*-modes clustering and re-runs may be needed to generate a meaningful interpretation and conclusion. On the other hand, considering the problem of unknown true value of *k*, Chaturvedi *et al.* (2001) suggested the idea of using latent class procedure with Akaike Information Criterion (AIC) to determine the true value of *k*.

2.2 Previous Studies on Road Traffic Accidents

Various researches on road traffic accidents have been conducted using different methods of data analysis in order to build a model that predict accurately the major factors of road traffic accidents. Furthermore, in order to obtain the accurate and reliable answers from the analysis, some researchers have attempted to modify the previous researches using different state-of-art data mining algorithms. In this section, the previous studies stated are the experiments whose datasets are in categorical scale only so that comparison between methods can be easily seen and are listed in Table 2.1.

In the study conducted by Shanthi and Ramani (2012), various types of classification algorithms were applied on the in predicting the factors which lead to road traffic accidents specific to injury severity. There were 33 attributes involved with variable "Injury Severity' as the class attribute. Different feature selection algorithms were attempted to remain the important variables and reduce the error rate, prior to classification and the results were compared with the experiments with only classification algorithms conducted. The study illustrated that Random Tree based on features selected by Feature Ranking algorithm and Arc-X4 Meta classifier outperformed the other individual approaches. It also concluded that manner of collision, seating position, first harmful event occurred during accidents, type of protection system used, age range of the person involved and police reported drug involvement were the causes driver to road traffic accidents, based on the road traffic accident datasets from the Fatality Analysis Reporting System (FARS).

Chong *et al.* (2004) evaluated the performance of artificial neural network and decision trees in modeling the severity of injury in head-on front impact point collisions. Only categorical variables were involved in the study. Contrast to the previous research focusing on binary classification, they extended the research to multiple category classification, including no injury, possible injury, non-incapacitating injury, incapacitating injury, and fatal injury. The results concluded that the decision tree outperforms the Artificial Neural Network in terms of accuracy.

Besides, in order to build a model that predicts more accurately the injury severity, support vector machines and hybrid decision tree-artificial neural network were added in the previous comparison including artificial neural network and decision trees in their another paper proposed in 2005 (Chong *et al.*, 2005). The results showed that hybrid decision tree-neural network outperformed the other methods.

With the similar objective with other studies, Kashani*et al.* (2010)used Classification and Regression Tree and Variable Importance Measure to identify the main factors affecting injury severity of drivers involved in traffic crashes on the main two-lane and two-way roads in Iran. Unlike the generalized linear regression models whose assumptions, such as functional relationships between the injury severity and the crash-related factors are linear, CART was used due to its non-parametric feature in modeling. Furthermore, the aforementioned measure was adopted in order to select the variables having the major contribution in predicting the target variable. The result showed that although the overall accuracy of the model decreased compared with previous studies, its prediction accuracy for fatality class increased.

On the other hand, clustering algorithms have also been widely adopted as preprocessor or data miner in recent years. For instance, Kumar and Toshniwal (2015) proposed a data mining framing in which *K*-modes clustering, association rule mining and trend analysis were included, to analyze road accident data. *K*-modes algorithm was used to reduce the heterogeneous among the categorical data objects in the study. Then, association rule mining was applied to both entire data set and the clusters resulting from *K*-modes clustering. The results showed that the combination of *K*-modes clustering and association rule performed well in terms of uncovering hidden patterns.

Besides, one of the clustering variant, Latent Class Clustering was used as a preliminary tool for segmentation of the categorical dataset in the study conducted by López *et al.*(2013). The performance of Bayesian Networks on the clusters resulting from Latent Class Clustering was then compared, in terms of identifying the major factors associated with accident severity, with the performance of Bayesian Network on the entire database. The results revealed that the combination of Latent Class Clustering and Bayesian Network outperformed the latter one when handling data with heterogeneous characteristics. Furthermore, in order to find the number of clusters, Bayesian Information Criterion, Akaike Information Criterion and

Consistent Akaike Information Criterion were adopted so that the model resulted was able to explain the data well.

Table 2.1: List of Previous Studies on Road Traffic Accidents.
(Department of Transport, 2016)

Title	Author(s), Year	Methods
Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques	(Shanthi and Ramani, 2012)	Classification algorithms (including C4.5, CR-T, ID3, CS-CRT, CS-MC4, Naïve Bayes and Random Tree) with feature selection algorithms (including CFS, FCBF, MIFS, MODTree and Feature Ranking algorithms)
Traffic Accident Analysis using Decision Trees and Neural Networks	(Chong *et al.*, 2004)	Decision Trees and Neural Networks
Traffic Accident Data Mining using Machine Learning Paradigms	(Chong *et al.*, 2005)	Decision Trees, Neural Networks, Support Vector Machines and Hybrid Decision Tree-Artificial Neural Network
A Data Mining Approach To Identify Key Factors of Traffic Injury Severity	(Kashani *et al.*, 2010)	Classification and Regression Tree and Variable Importance Measure
A Data Mining Framework to Analyze Road Accident Data	(Kumar and Toshniwal, 2015)	*K*-modes Clustering, ,Association rule and trend analysis
Analysis of Traffic Accidents on Rural Highways Using Latent Class Clustering and Bayesian Networks	(López *et al.*, 2013)	Latent Class Clustering and Bayesian Networks

2.3 Conclusion

In this study, various types of clustering algorithms, such as *K*-means, *K*-modes, hierarchical and latent class clustering methods are introduced on how to deal with large datasets on qualitative scale in terms of their advantages and disadvantages. Besides, apart from clustering techniques, the previous studies stated in this paper also provide the examples of the ability of other data mining algorithms when handling with road traffic accident datasets on categorical scale, such as decision trees, neural network, association algorithms and so on.

CHAPTER 3

METHODOLOGY

This chapter explains the description of the variables and the clustering techniques that were applied in this study, namely *K*-means and *K*-modes clustering algorithms. Besides, Silhouette index is also mentioned in order to select a suitable model. A flowchart is also illustrated in order to have a better understanding on the framework of this study.

3.1 Data Selection and Data Cleaning

In this study, secondary data analysis was conducted due to the open access to the road traffic accident dataset in Great Britain (Data.gov.uk, 2015). There are 3 datasets about Great Britain road traffic accidents, namely, accident circumstances, vehicle and casualty. The accident circumstances dataset was used and only data of June and December 2014 were selected to conduct the experiment. There are 12532 and 12036 accident events for June and December, respectively.

Originally, there are 32 variables in the original dataset but 9 variables that describe mainly the location (e.g. longitude, latitude), number of vehicles and people involved, absence or presence of police in the incident, were excluded since the objective of the study is to determine the combination of contributory factors leading to road traffic accidents. Instead of variable about the location, the variables of road type, road class, junction detail gives more specific information about the accident circumstance.

Besides, Junction Control was removed as it contains too many missing values. Furthermore, list-wise deletion was also applied by removing those accident cases with one or more missing values. Hence, there were 12283 for June and 11492 for December with 14attributes being included in the study. Description of the selected variables is shown in Table 3.1

Table 3.1: Description of Selected Variables.
(Department for Transport, 2015)

Attribute	**Attribute Values**
a) Day of Week	1- Weekday 2- Weekend
b) Time	1- 00:00 – 11:59 2- 12:00 – 23:59
c) Road Class	1- Motorway 2- A(M) 3- A 4- B 5- C 6- Unclassified
d) Road Type	1- Roundabout 2- One-way Street 3- Dual carriageway 4- Single carriageway 5- Slip Road
e) Speed Limit	1- Less than and equal to 40 mph 2- More than 40 mph
f) Junction Detail	1- Not at or within 20 metres of junction 2- Roundabout 3- Mini-roundabout 4- T or Staggered junction 5- Slip road 6- Crossroads 7- Junction – more than 4 arms (not a roundabout) 8- Using private drive of entrance 9- Other junction
g) Pedestrian Crossing (Human Control)	1- None within 50 metres 2- Control by school crossing patrol 3- Control by other authorized person
h) Pedestrian Crossing (Physical Facilities)	1 No physical crossing facilities within 50 metres 2 Zebra crossing 3 Pelican, puffin, toucan or similar non-junction pedestrian light crossing 4 Pedestrian phase at traffic signal junction 5 Footbridge or subway 6 Central refuge

Table 3.1: (continued)

Attribute	**Attribute Values**
i) Light Conditions	1- Daylight 2- Darkness – lights lit 3- Darkness – lights unlit 4- Darkness – no lighting
j) Weather	1- Fine without high winds 2- Raining without high winds 3- Snowing without high winds 4- Fine with high winds 5- Raining with high winds 6- Snowing with high winds 7- Fog or mist 8- Other
k) Road Surface Condition	1- Dry 2- Wet or damp 3- Snow 4- Frost or ice Flood (surface water over 3cm deep
l) Special Conditions at Site	1- None 2- Automatic traffic signal out 3- Automatic traffic signal partially defective 4- Permanent road signing or marking defective or obscured 5- Roadworks 6- Road surface defective 7- Oil or diesel 8- Mud
m) Carriageway Hazards	1- None 2- Dislodged vehicle load in carriageway 3- Other object in carriageway 4- Involvement with previous accident 5- Pedestrian in carriageway- not injured 6- Any animal in carriageway (except ridden horse)
n) Urban or Rural	1- Urban 2- Rural

3.2 Data Transformation

After data cleaning, 14 multi-categorical variables were transformed into binary variables. In other words, number of binary variable for each categorical variable was as many as number of classes it had.

Then, data was transposed because the objective was to cluster variables to study the similarity between variables by using *K*-means and *K*-modes clustering algorithms. Hence, value of ‘1’ for binary variable denoted its presence when accident occurred while ‘0’ denoted absence.

Proximity measures between the variables that were applied in this study, were Hamming distance and simple dissimilarity measure for *K*-means and *K*-modes clustering algorithms, respectively. Hence, both Hamming distance and simple dissimilarity measure indicated how many times both variables were not present together when accident happened.

3.3 Data Mining

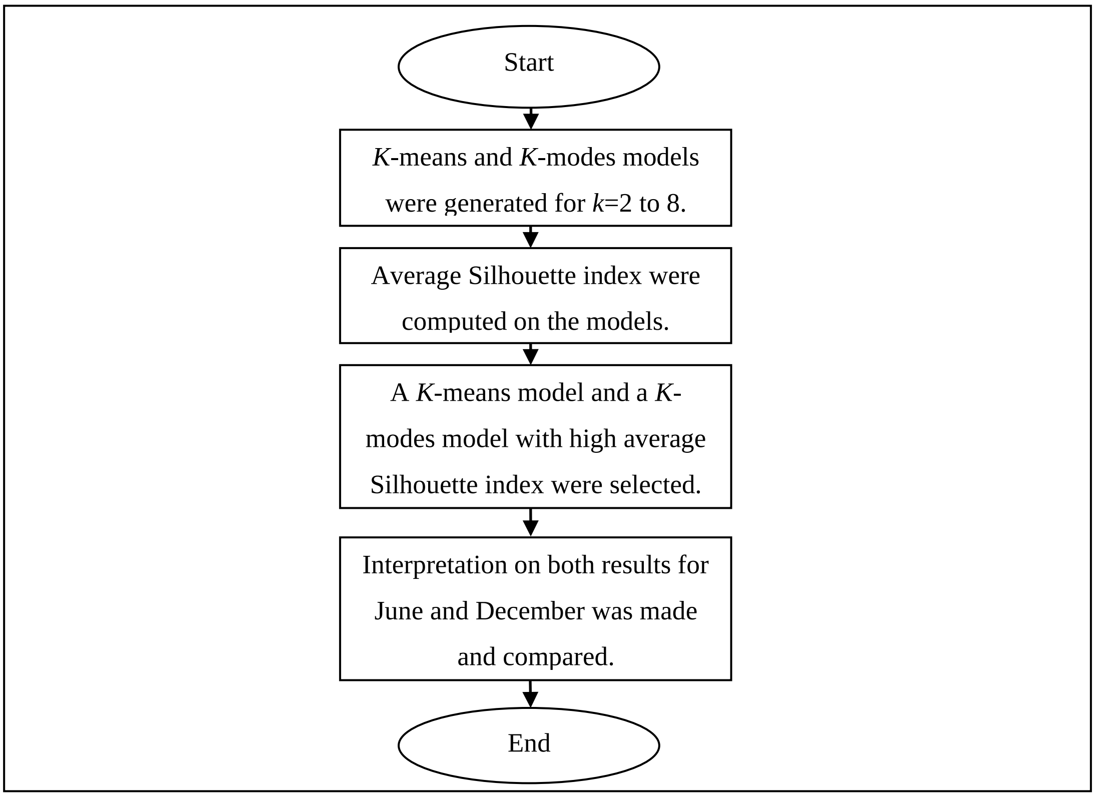

Start
K-means and *K*-modes models were generated for *k*=2 to 8.
Average Silhouette index were computed on the models.
A *K*-means model and a *K*-modes model with high average Silhouette index were selected.
Interpretation on both results for June and December was made and compared.
End

Figure 3.1: Flow Chart of Proposed Framework for Analysis.

Figure 3.1 gives a brief introduction of the framework that was conducted on 2 datasets (June and December) in this study and each step is explained in detail one-by-one in the following sub-topics. The selected data mining software were R 3.3.1 and Matlab R2012b.

3.3.1 *K*-means Clustering

When dealing with categorical variables with more than 2 values, *K*-means clustering can be adopted by mapping them into large number of new binary variables and treating them as numeric data (Ralambondrainy, 1995). Therefore, there were 68 binary variables when employing this algorithm. In this study, the algorithm was run by Matlab R2012b.

The procedure of *K*-means clustering is stated below (Gupta, 2011).

(i) Select k initial means randomly for k clusters.

(ii) Calculate the dissimilarity between an object with all cluster means.

(iii) Allocate an object to the cluster whose mean is nearest to the object.

(iv) Re-calculate the cluster mean after the assignation of the object.

(v) Repeat Step (i) to (iv) until the cluster membership is unchanged.

The procedure above shows the dissimilarity measure needed. In this paper, since the variables were defined in binary form, Hamming distance is one of the suitable proximity measures specifically for binary variables. It deals with pairs of binary strings of equal length by comparing their corresponding bits as stated below.

$$d(X,Y) = \sum_{m=1}^{M} \delta(X_m, Y_m) \tag{3.1}$$

where

$$\delta(X_m, Y_m) = \begin{cases} 0 & (X_m = Y_m) \\ 1 & (X_m \neq Y_m) \end{cases}$$

While X and Y represents M-dimensional binary variables (Huang, 1997).In this study, X_m and Y_m represented presence or absence of the variable on the m-th accident spot since the focus was on the dissimilarity of the elements of the accident circumstances.

3.3.2 *K*-modes Clustering

K-modes clustering algorithm was performed, according to the simple matching dissimilarity measure. The formula of the dissimilarity measure is

$$d(X,Y) = \sum_{m=1}^{M} \delta(X_m, Y_m) \tag{3.2}$$

where

$$\delta(X, Y_m) = \begin{cases} 0 & (X_m = Y_m) \\ 1 & (X_m \neq Y_m) \end{cases}$$

while *X* and *Y* are two *M*-dimensional categorical objects.

Thus, *K*-modes clustering algorithm is shown below (Huang, 1997):-

i. Assign data objects to the cluster whose initial cluster mode is nearest to it according the simple matching dissimilarity measure shown above.

ii. Update the cluster modes after the allocation of the objects and calculate the dissimilarity between each object and the updated modes.

iii. Repeat Step 1 and 2 with the replacement of initial cluster modes with updated modes, until no object has moved to another cluster.

3.3.3 Selection of *K*-means and *K*-modes Models

Silhouette coefficient plays the role of providing a graphical display to the result of the clustering algorithms, which is similar with the function of dendrograms of hierarchical method, and evaluating the clustering validity. It is based on the concept of tightness and separation， which indicates the within-variability between objects of a cluster and the distance between the clusters, respectively. Thus, it is suitable to find the appropriate number of cluster, *k*. The formula of Silhouette coefficient, $s(x_i)$ of each clustered object $x_i$is

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} \qquad (3.3)$$

where

$a(x_i)$ is the average dissimilarity between x_i and other objects of the same cluster;

$d(x_i, C)$ denotes the average dissimilarity of x_i to all objects of any other cluster, C.and

$\mathrm{b}(x_i) = \min\ d(i, C)$, the minimum average dissimilarity of x_i with the other clusters apart from the cluster $x_i$it belongs to(Rousseeuw, 1987).

The range of $s(x_i)$ is between -1 and 1. When $s(x_i)$ is positive and close to 1, x_i is well-clustered. However, when it is negative, it means that $x_i$is placed in the wrong cluster and $s(x_i)$ is zero when x_i is between the clusters.

In this study, average Silhouette value were computed on both *K*-means and *K*-modes models for*k*=2 to8. The *K*-means and *K*-modes models with interpretable and highest value of the coefficient, which is close to 1, were selected. Furthermore, instead of using the dissimilarity matrix on a ratio scale, such as Euclidean distance, simple dissimilarity matrix and Hamming distance for *K*-modes and *K*-means clustering, respectively were used when calculating Silhouette index. Explanation of both distance measures are explained in subtopics 3.3.1 and 3.3.2, respectively.

To sum-up, both selected *K*-means and *K*-modes models, with high average Silhouette index were chosen and both clustering results were compared between each other.

3.4 Conclusion

K-means and *K*-modes clustering were adopted to deal with categorical accident factor dataset. Both models were selected based on average Silhouette index. Lastly, interpretation was conducted on the clustering result to identify the similarities within the clusters and dissimilarities between the clusters.

CHAPTER 4

RESULTS AND DISCUSSION

Great Britain’s accident circumstances dataset in June and December 2014 were clustered separately by both *K*-modes and *K*-means clustering methods, with the objective of studying the combinations of factors that lead to road traffic accident in Great Britain. Then, the interpretation on the results was made and the results from both methods were also compared between each other.

4.1 Explanation of Data

As stated in Chapter 3, data preprocessing was conducted before clustering data. Irrelevant variables, such as location, number of vehicles and people involved, absence or presence of police on the spot, were excluded. Besides, Junction Control was removed due to too many unknown data. Then, list-wise deletion was made on the dataset because *K*-means and *K*-modes clustering method cannot analyze data that contains missing data values

It resulted that there were 12283 for June and 11492 for December with 14 attributes being included in the study. Refer to Table 3.1and Appendix A. Then, these 14 attributes were transformed into binary form. Hence, there were 68 binary variables. Refer to Chapter 3.1 and 3.2 that give clear explanation on this part, and Appendix B that states the codes for data-processing. There are 2 tables (Table 4.1 and 4.2) below that show number of accident cases in June and December in terms of accident severity.

Table 4.1: Number of Accident Cases in June in terms of Accident Severity.

Accident Severity	**Number of Cases**	**Percentage (%)**
Fatal	143	1.1642
Serious	1874	15.2569
Minor	10266	83.579
Total	**12283**	

Table 4.2: Number of Accident Cases in December in terms of Accident Severity.

Accident Severity	**Number of Cases**	**Percentage (%)**
Fatal	174	1.5141
Serious	1541	13.4093
Minor	9777	85.0766
Total	**11492**	

4.2 Research Findings

In this part, results for *K*-means and *K*-modes clustering on June and December dataset were analyzed and interpreted. Only 3 Silhouette plots are displayed when discussing suitable *k* for each method and dataset, while the rest are shown in Appendix C. Then, the difference of results from both *K*-means and *K*-modes models were compared in terms of how many clusters each model generated and the elements of accident circumstances each cluster contained.

4.2.1 *K*-means Clustering Method on June Dataset

The average Silhouette index table for *K*-means clustering on accident circumstances datasets in June for *k*=2 until 8 is illustrated below.

Table 4.3: Average Silhouette Index of *K*-means Clustering Method on June Dataset.

Number of Clusters, *k*	Average Silhouette Index
2	0.7906
3	0.7218
4	0.7013
5	0.7051
6	0.7072
7	0.7343
8	0.5204

Table 4.3 shows that average Silhouette value for *k*=2 is the highest values, followed by *k*=7 and 3. Silhouette plots for *k*=2, 3 and 7 are shown in Figure 4.1, 4.2 and 4.3, respectively.

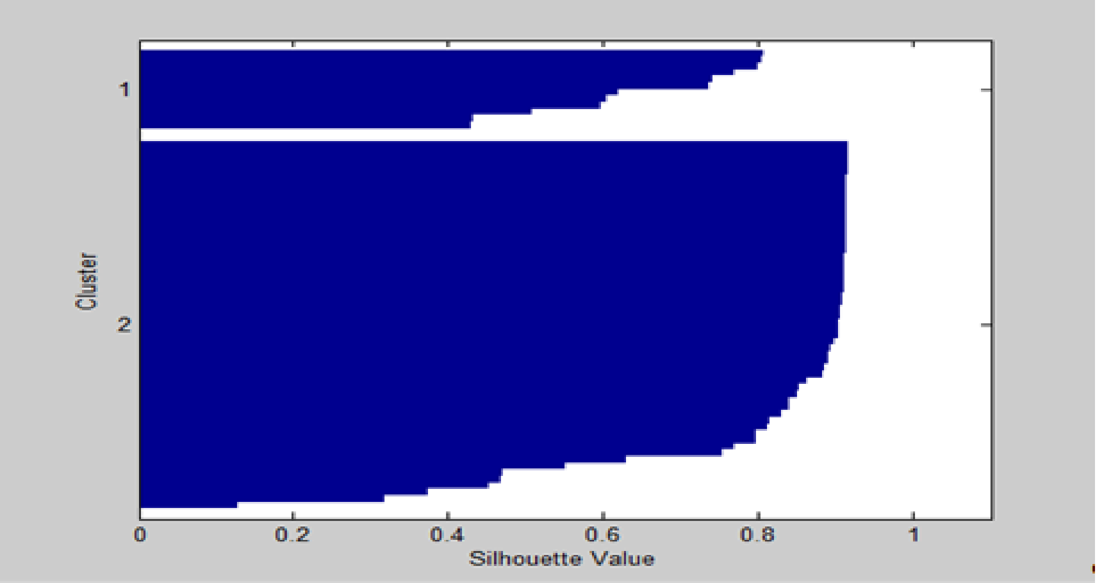


Figure 4.1: Silhouette Plot of June Dataset for *k*=2 after *K*-means Clustering.

Figure 4.1 shows that most items of accident circumstances dataset were well-clustered because their Silhouette indexes are high. Therefore, this model had the highest average Silhouette index among the models.

First large cluster in Figure 4.1 cannot be considered as combination of factors because originally, all points were the attribute value of multi-categorical variable, before they were transformed into binary form. Some points were originally from same multi-categorical variable but they were clustered into the same cluster. Therefore, there was only one interpretable cluster, which is Cluster 1.

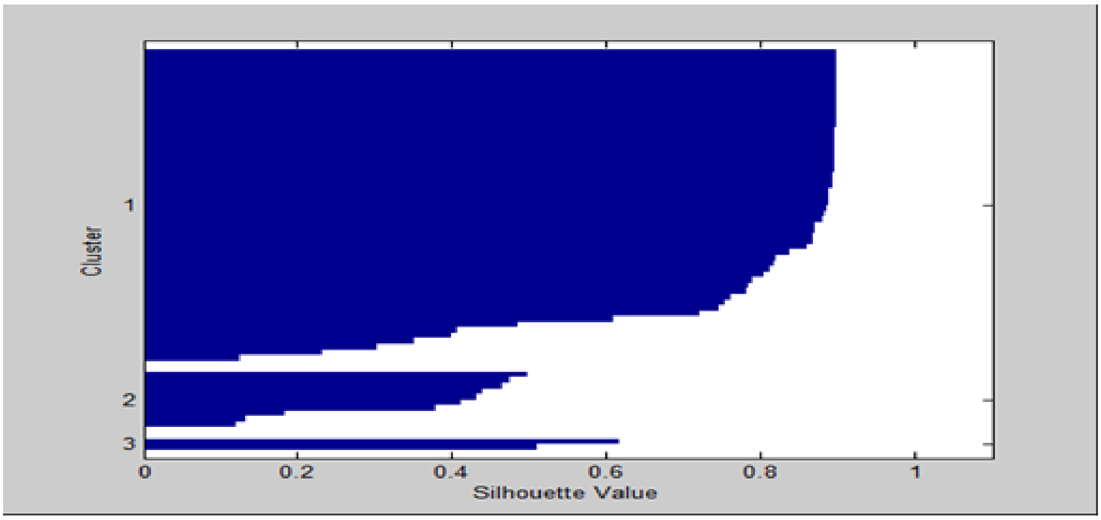


Figure 4.2: Silhouette Plot of June Dataset for *k*=3 after *K*-means Clustering.

Figure 4.2 illustrates that although most Silhouette indexes of the points in second and third clusters were moderately positive, as compared to the first cluster, there was no misclassification in this model.

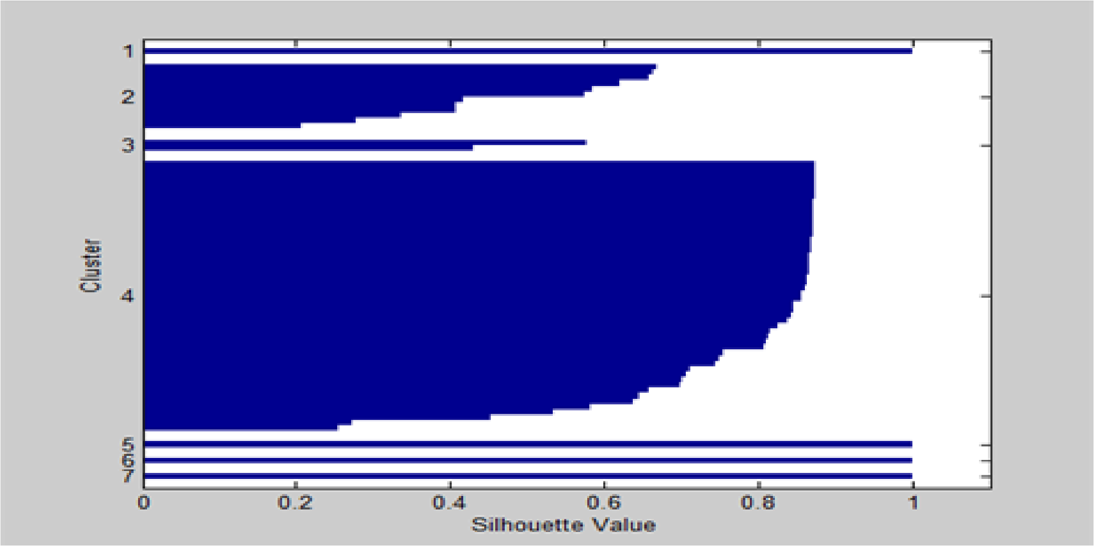


Figure 4.3: Silhouette Plot of June Dataset for *k*=7 after *K*-means Clustering.

When clustering the 68 elements into 7 groups, there were 4 groups that contained an element only. It meant that this element was independent of others and could be present or absent, regardless of the presence of others. Hence, there were 4 Silhouette value of ‘1’. This resulted that the average Silhouette index was high, compared to others.

As a result, model for *k*=3 was selected as the final *k*-means model on June dataset because there was no misclassification and no single-element cluster that might affect the average Silhouette value. Compared to model with *k*=2, it had one more group of factors, while to model with *k*=7, it had no cluster that contained independent element. Table 4.4displays the cluster label for each accident circumstances' element.

Table 4.4: *K*-means Cluster Label on June Dataset.

Cluster Label	Elements of Accident Circumstances
1	1 Weekday 2 12:00-23:59 (Time) 3 Single Carriageway 4 No Human Control for Pedestrian Crossing within 50m 5 No Physical Crossing Facilities within 50m 6 Daylight 7 Fine without Wind 8 Dry Road Surface Condition 9 No Special Condition at Site 10 No Carriageway Hazard
2	1 Speed Limit of Less than 40mph 2 Urban
3	Rest of the elements

Table 4.5: Number of Cases in June when All Elements in Each *K*–means Cluster were Present and Absent.

		Number of Cases
Elements in Cluster 1	Present	3089
	Absent	0
Elements in Cluster 2	Present ('1')	7744
	Absent ('0')	2706

The factors were clustered together by calculating total number of mismatches between binary values of factors in every case (one is present '1' while the other is absent, '0'). Hence, the similarity, not only included the number of time when both variables were present, '1', and also when both were absent, '0'. Thus, the question of whether Cluster 1 and Cluster 2 were sources that led to road traffic accident in June 2014 could not be answered. Refer to Table 4.5.

As displayed in Table 4.4 and 4.5, during the time period of between 12:00 and 23:59 of weekday, there were 3089 accidents happening on the dry single carriageway without any human control and physical facilities for pedestrian crossing, special condition and any hazard around. These aforementioned cases happened in the daylight when the weather was fine without wind. For the second cluster, 7744 road traffic accidents happened on the road where speed limit is less than 40mph' in Urban area.

4.2.2 *K*-means Clustering Methods on December Dataset

The mean Silhouette index table for *K*-means clustering on accident circumstances datasets in December for *k*=2 until 8 is illustrated below.

Table 4.6: Average Silhouette Index of *K*-means Clustering Method on December Dataset.

Number of Clusters, *k*	Average Silhouette Index
2	0.7013
3	0.6234
4	0.5812
5	0.6008
6	0.5899
7	0.6159
8	0.6479

Average Silhouette values in Table 4.6 can be concluded that when 68 binary variables were clustered into 2, 3 or 8 clusters, their average values, which were 0.7013, 0.6234 and 0.6479, respectively, are higher, than others. Figure 4.4, 4.5 and 4.6 display Silhouette plots for *k*=2, 3 and 8, respectively, after *K*-means clustering on December dataset.

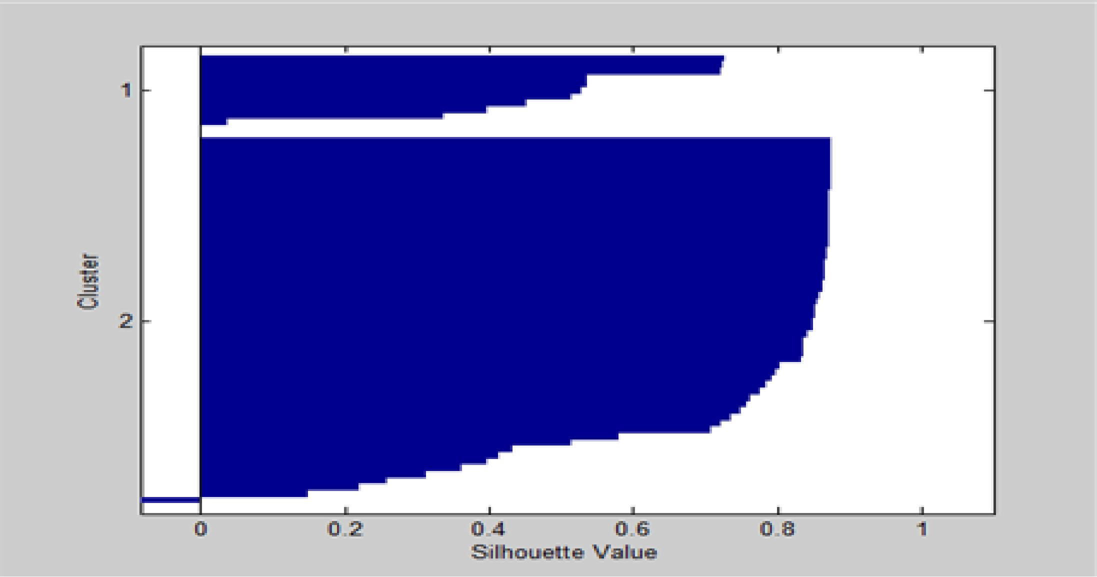


Figure 4.4: Silhouette Plot of December Dataset for *k*=2 after *K*-means Clustering.

From Figure 4.4, it can be concluded that most points were well-clustered, except that there was a point in second cluster that was assigned into wrong cluster.

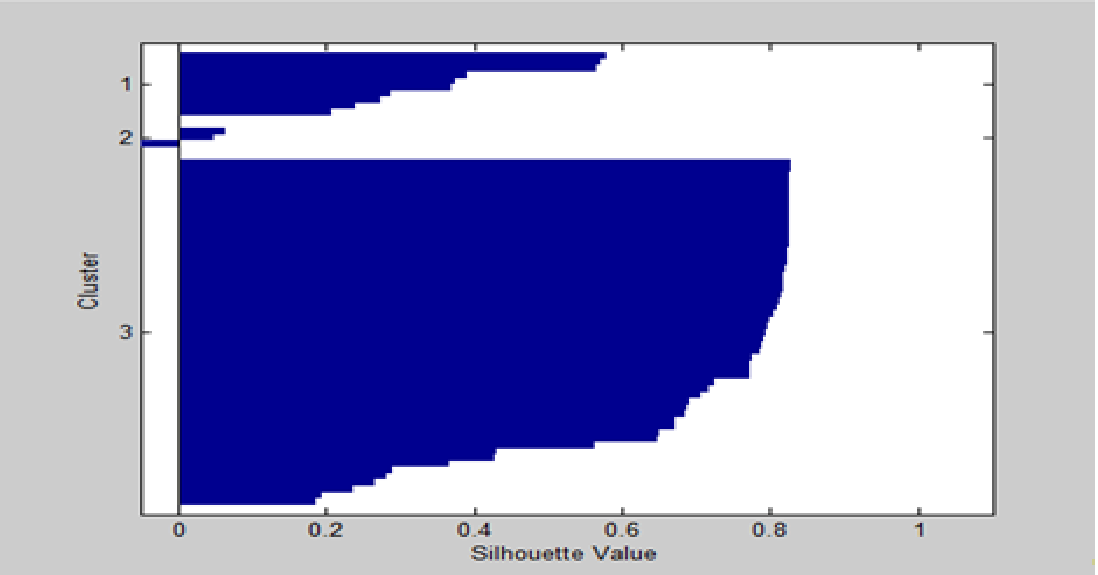


Figure 4.5: Silhouette Plot of December Dataset for *k*=3 after *K*-means Clustering.

Figure 4.5 shows that there were a misclassified point and points with very low Silhouette indexes in second cluster, while for first and third clusters, most points had positive high or moderate values.

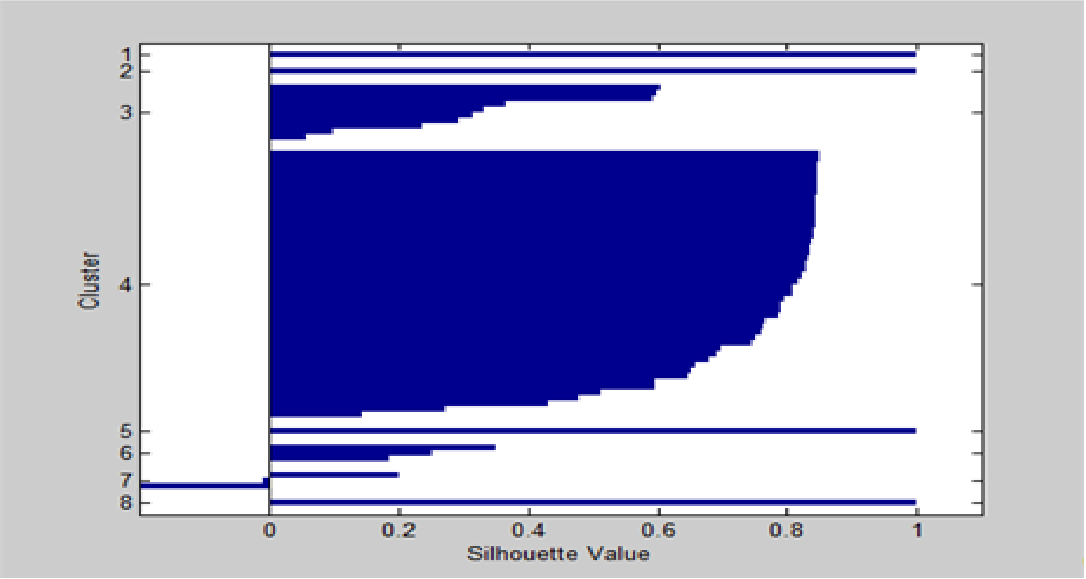


Figure 4.6: Silhouette Plot of December Dataset for *k*=8 after *K*-means Clustering.

Figure 4.6 presents that 7th cluster had 2 misclassified points due to their negative Silhouette values. Besides, there were 4 independent points and it might be a reason of why the average Silhouette value could be high.

So, suitable number of cluster for *K*-means model on December dataset was decided to be 2. Although there was a misclassified point, it was located in Cluster 2 which could not be interpreted logically. Furthermore, most points had high or moderate Silhouette indexes. Undeniably, this model partitioned the elements well, as compared to the other 2 models. Cluster label for element is illustrated below.

Table 4.7:*K*-means Cluster Label on December Dataset.

Cluster Label	Elements of Accident Circumstances
1	1 Weekday 2 1200-2359 3 Single Carriageway 4 Speed Limit of Less than 40mph 5 No Human Control for Pedestrian Crossing within 50m 6 No Physical Crossing Facilities within 50m 7 Daylight 8 Fine without Wind 9 No Physical Crossing Facilities within 50m 10 Daylight 11 Fine without Wind 12 No Special Condition at Site

Table 4.7: (continued).

Cluster Label	Elements of Accident Circumstances
	13 No Carriageway Hazard 14 Urban
2	Rest of the elements

Table 4.8**:** Number of Cases in December when All Elements in Each *K*-means Cluster were Present and Absent.

		Number of Cases
All Elements in Cluster 1	Present	730
	Absent	0

As referred to Table 4.4 (June) and 4.7 (December), Cluster 1 of both clustering models contained 9 similar factors, namely, ‘Weekday’, ‘12:00-23:59’, ‘Single Carriageway’, No Human Control for Pedestrian Crossing within 50m’, ‘No Physical Crossing Facilities’, ‘Daylight’, ‘Fine without Wind’, ‘No Special Condition at Site’ and ‘No Carriageway Hazard’. The non-overlapped elements in Cluster 1 for December were ‘Speed Limit of Less than 40mph’ and ‘Urban’. To conclude, in December, there were 730 road traffic accidents in Great Britain due to the combination of factors, as listed in Cluster.1of Table 4.8.

4.2.3 *K*-modes Clustering Method on June Dataset

Table 4.9: Average Silhouette Index of *K*-modes Clustering Method on June Dataset.

Number of Clusters, *k*	Average Silhouette Index
2	0.7906
3	0.7224
4	0.6917
5	0.6430
6	0.5537
7	0.5797
8	0.5650

First 3 *K*-modes models on June dataset in Table 4.9 have the highest average Silhouette values. See Figure 4.7, 4.8 and 4.9.

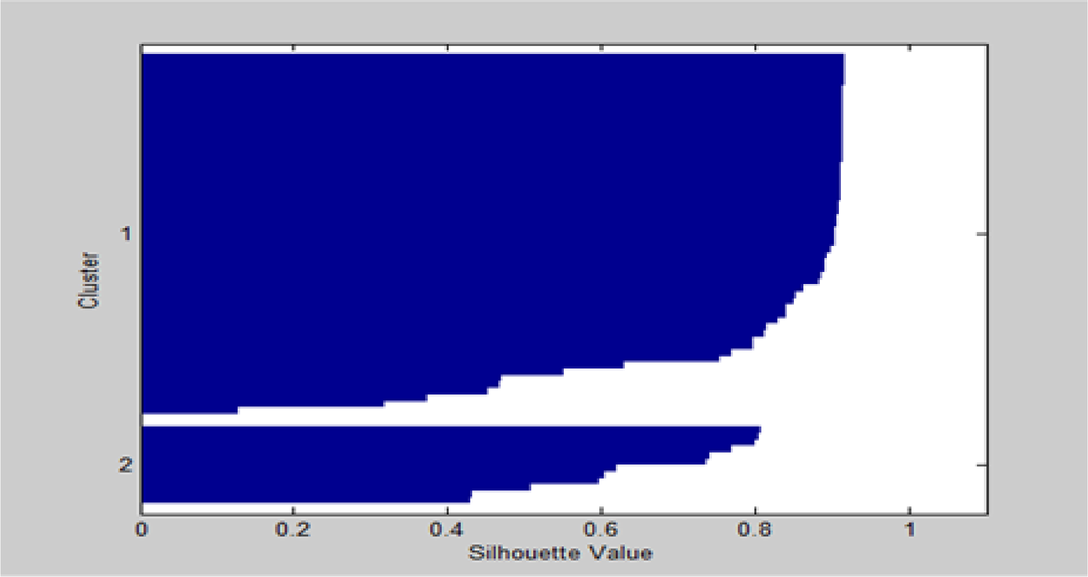


Figure 4.7: Silhouette Plot of June Dataset for *k*=2 after *K*-modes Clustering.

Most points were assigned well when clustering element into 2 groups because they had high Silhouette values. Refer to Figure 4.7.

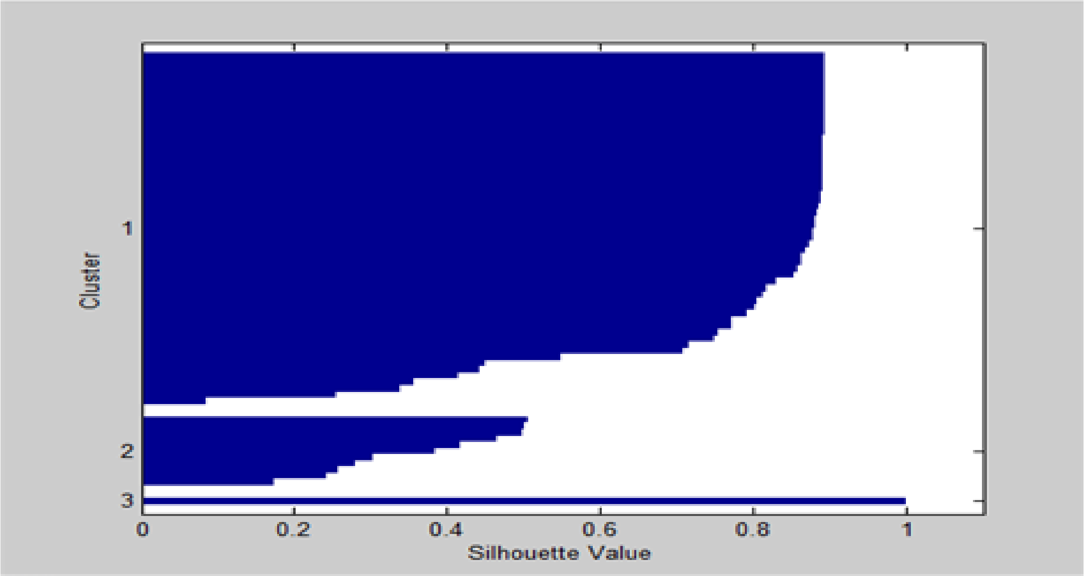


Figure 4.8: Silhouette Plot of June Dataset for *k*=3 after *K*-modes Clustering.

Figure 4.8 illustrates that, Cluster 2 had points with moderate or low Silhouette values, while Cluster 3 was an independent element because its Silhouette value was 1.

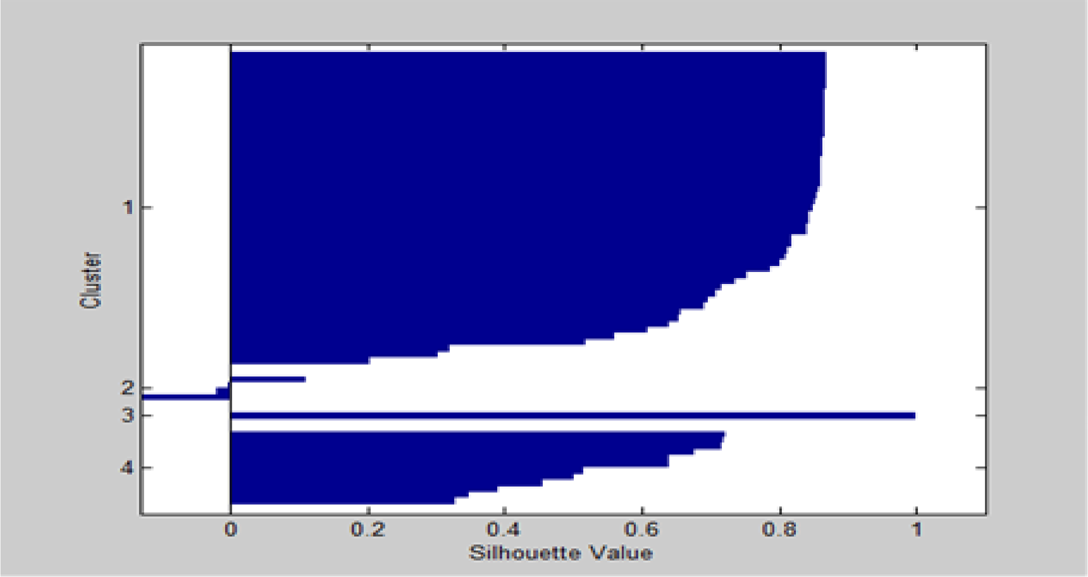


Figure 4.9: Silhouette Plot of June Dataset for *k*=4 after *K*-modes Clustering.

Second cluster in Figure 4.7 contained 3 misclassified points, out of 4 points. A single-element cluster in this model offset the negative Silhouette values, so average value of this model could be high.

Thus, when clustering June data by *K*-modes clustering methods, *k* was decided to be 2 because most points were well-assigned. Besides, there was no single-element cluster and misclassified point. Table 4.10 exhibits the cluster label for each element of accident circumstances.

Table 4.10: *K*-modes Cluster Label on June Dataset.

Cluster Label	Elements of Accident Circumstances
1	1 Weekday 2 1200-2359 3 Single Carriageway 4 Speed Limit of Less than 40mph 5 No Human Control for Pedestrian Crossing within 50m 6 No Physical Crossing Facilities within 50m 7 Daylight 8 Fine without Wind 9 Dry Road Surface Condition 10 No Special Condition at Site 11 No Carriageway Hazard 12 Urban
2	Rest of the elements

Table 4.11: Number of Cases in June when All Elements in Each *K*–modes Cluster were Present and Absent.

		Number of Cases
Elements in Cluster 1	Present	1980
	Absent	0

From this model, the characteristics of the accident site of 1980 accidents was a dry urban single carriageway on which there was no human control and physical facilities for human crossing, special condition as well as hazard. Its speed limit was less than 40 mph. These cases occurred when the weather was fine without wind, during the daylight in the period of ‘1200-2359’ of weekday.

Furthermore, although there were only 2 clusters in *k*-modes model, its elements in its first cluster included all elements of first and second clusters of *k*-means model on June dataset. Refer to Table 4.4.

4.2.4 *K*-modes Clustering Method on December Dataset

Table 4.12: Average Silhouette Index of *K*-modes Clustering Method on December Dataset.

Number of Clusters, *k*	Average Silhouette Index
2	0.7055
3	0.5991
4	0.5737
5	0.5499
6	0.4261
7	0.5607
8	0.4876

For December dataset, *K*-modes models with *k*=2, 3 and 4 have the highest values among the models. Refer to Table 4.12.

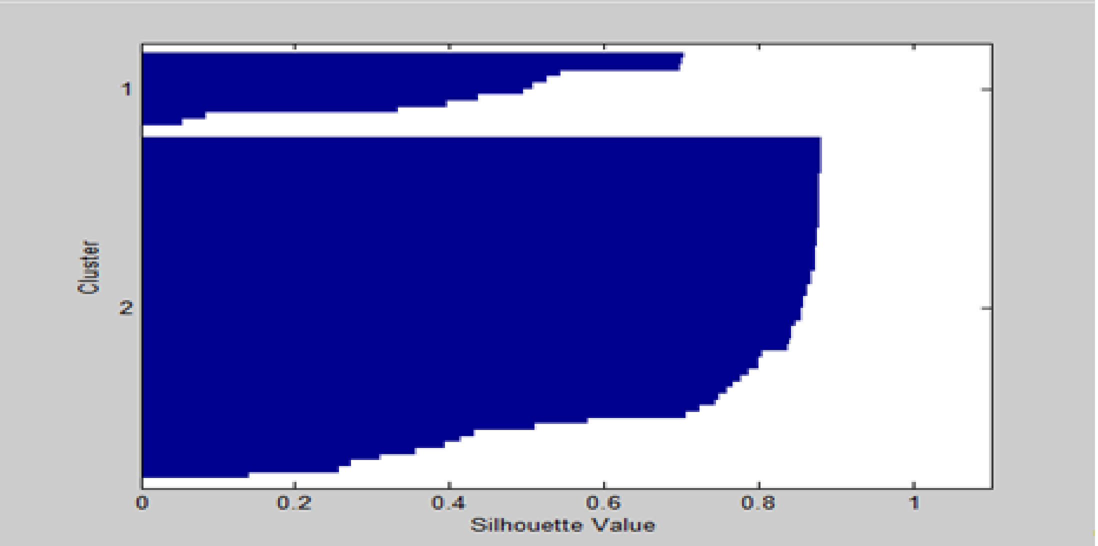


Figure 4.10: Silhouette Plot of December Dataset for k=2 after K-modes Clustering.

Figure 4.10 exhibits that most points were well-assigned into their respective clusters.

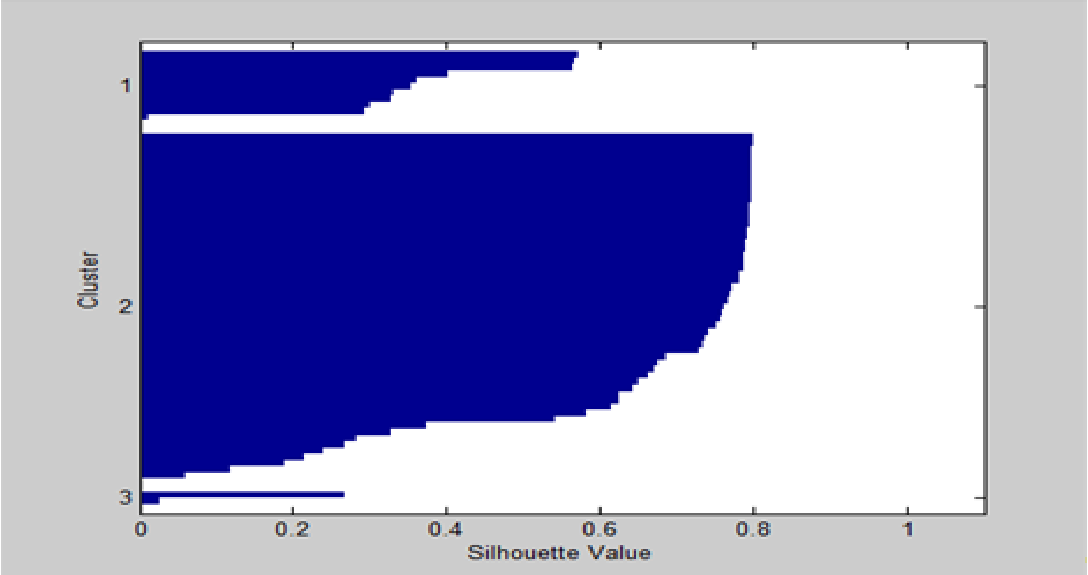


Figure 4.11: Silhouette Plot of December Dataset for k=3 after K-modes Clustering.

In Figure 4.11, apart from most well-assigned points in first 2 clusters, points from 3rd cluster had relatively low Silhouette values, as compared to Silhouette values of other points. However, there was no misclassification in this model.

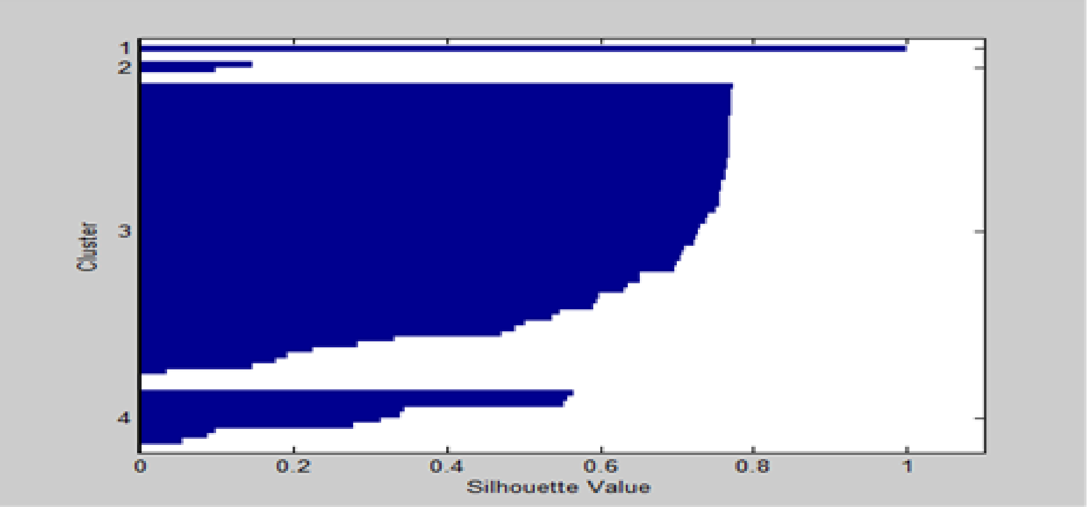


Figure 4.12: Silhouette Plot of December Dataset for *k*=4 after *K*-modes Clustering.

Figure 4.12 illustrates that elements in 2nd cluster had low Silhouette index and Silhouette value of the only element in first cluster was value of 1 because this cluster was made up of an element only. Most elements in 3rd and 4 clusters were grouped well due to their respective high or moderate Silhouette value.

To sum-up, *k* for *K*-modes clustering on December dataset was decided to be 3 because the objective of investigating the combinations of factors that led to road traffic accidents had to be considered. Hence, as compared to *k*=2, it had one more group of factors. Unlike *k*=4, the model did not have single-element cluster that might raise the value of average Silhouette value. Table 4.13 shows cluster label for element of accident circumstances.

Table 4.13: *K*-modes Cluster Label on December Dataset.

Cluster Label	Elements of Accident Circumstances
1	1 Weekday 2 12:00-23:59 3 Single Carriageway 4 Speed Limit of Less than 40mph 5 No Human Control for Pedestrian Crossing within 50m 6 No Physical Crossing Facilities 7 Fine without Wind 8 Dry Road Surface Condition 9 No Special Condition at Site 10 No Carriageway Hazard 11 Urban
2	1. 00:00-11:59 2. Daylight
3	Rest of the elements

Table 4.14: Number of Cases in December when All Elements in Each *K*–modes Cluster were Present and Absent.

		Number of Cases
Elements in Cluster 1	Present	1089
	Absent	0
Elements in Cluster 2	Present	2829
	Absent	4529

As illustrated in Table 4.13 and 4.14, it can be concluded that in December 2014, 2829 accidents happened during ‘Daylight’ and in the period of ‘00:00-11:59’. Besides, 1089 cases happened on single carriageway in the urban area, where speed limit was less than 40 mph and road condition was dry, during ‘12:00-23:59’ of Weekday. These cases occurred on the fine weather without wind and there was no human control and physical crossing facilities for pedestrian crossing within 50m, special condition and hazard around the accident spot.

4.3 Conclusion

In overall, each model were chosen, not just based on their average Silhouette value and also how many combinations of elements of accident circumstance each model had, regardless of the method. It was because that this study focused on studying the combinations of factors that led to accidents.

Table 4.15: Differences between Results of *K*-means and *K*-modes Models on June and December Datasets.

***K*-means Clustering Method**	***K*-modes Clustering Method**
June Cluster 1 1 Weekday 2 1200-2359 3 Single Carriageway 4 No Human Control for Pedestrian Crossing within 50m 5 No Physical Crossing Facilities 6 Daylight 7 Fine without Wind 8 Dry Road Surface Condition 9 No Special Condition at Site	**June** Cluster 1 1 Weekday 2 1200-2359 3 Single Carriageway 4 Speed Limit of Less than 40mph 5 No Human Control for Pedestrian Crossing within 50m 6 No Physical Crossing Facilities 7 Daylight 8 Fine without Wind 9 Dry Road Surface Condition

Table 4.15: (continued).

***K*-means Clustering Method**	***K*-modes Clustering Method**
10 No Carriageway Hazard <u>Cluster 2</u> 1 Speed Limit of Less than 40mph 2 Urban	10 No Special Condition at Site 11 No Carriageway Hazard 12 Urban
December <u>Cluster 1</u> 1 Weekday 2 1200-2359 3 Single Carriageway 4 Speed Limit of Less than 40mph 5 No Human Control for Pedestrian Crossing within 50m 6 No Physical Crossing Facilities 7 Daylight 8 Fine without Wind 9 No Special Condition at Site 10 No Carriageway Hazard 11 Urban	**December** <u>Cluster 1</u> 1 Weekday 2 12:00-23:59 3 Single Carriageway 4 Speed Limit of Less than 40mph 5 No Human Control for Pedestrian Crossing within 50m 6 No Physical Crossing Facilities 7 Fine without Wind 8 Dry Road Surface Condition 9 No Special Condition at Site 10 No Carriageway Hazard 11 Urban <u>Cluster 2</u> 1. 00:00-11:59 2. Daylight

Tables 4.15 exhibits that for June dataset, *K*-modes clustering algorithms included all elements of cluster 1 and cluster 2 of *K*-means model, while for December dataset, cluster 1 of *K*-modes model contained all the elements almost as same as the cluster 1 from *K*-means model. The exceptions were the ‘Daylight’ that was assigned into cluster 2 of *K*-modes model as well as ‘Dry Road Surface Condition’ that was included in cluster 1 of *K*-modes model but not in cluster 1 of *K*-means model.

Furthermore, regardless of the season (winter or summer) and methods, there were some of the elements of accident circumstance that were always assigned into the same group. There were ‘Weekday’, ‘12:00-23:59’, ‘Single Carriageway’, ‘No Human Control for Pedestrian Crossing within 50m’, ‘No Physical Crossing Facilities within 50m’, ‘Fine without Wind’, ‘No Special Condition at Site’ as well as ‘No Carriageway Hazard’. It also brought the meaning that these factors had high degree of homogeneity between each other because these 8 elements existed together, or otherwise (all of them were not on the spot), in most accident cases in June and December in Great Britain.

CHAPTER 5

CONCLUSIONAND FUTURE RESEARCH

5.1 Conclusion

Department for Transport of Great Britain (2015) stated that there was a 4% increase in number of road death in 2014. Furthermore, the peak normally occurred in summer. However, unusual peaks in number of being killed in road accident happened on the end of this year, particularly in the last 2 months, November and December. This issue rose the interest to conduct this study.

Besides, since Great Britain's road traffic accident dataset is large, statistical clustering methods, such as *K*-means and *K*-modes clustering techniques are very suitable to work on this data. These methods have already offered by data mining package due to their efficiency to handle with large dataset, Thus, the study was conducted in order to investigate the combinations of elements of road circumstances (factors) that led to road traffic accidents in Great Britain in summer (June) and winter (December), by *K*-means clustering and *K*-modes clustering methods.

The feature of the study is to cluster the factors of the road accident, instead of accident cases, based on dissimilarity metrics – Hamming distance and simple dissimilarity measure. The word 'dissimilarity' in this study means that how usual 2 elements were not present together at the accident spot. Thus, all the factors were transformed into binary form.

As a result, among these 4 models, there was no any large difference between each others. Apart from the slight alteration on one or two elements, there were 8 fixed elements of accident circumstance, that were always grouped together, namely

‘Weekday’, ‘1200-2359’, ‘Single Carriageway’, ‘No Human Control for Pedestrian Crossing within 50m’, ‘No Physical Crossing Facilities’, ‘Fine without Wind’, ‘No Special Condition at Site’ and ‘No Carriageway Hazard’ that were always grouped together.

5.2 Recommendations

There are some drawbacks in this study that can be furthered up deeply in future. The first one is the random initialization of centers (means or modes) in *k*-means and k-modes clustering methods. This led to many times of iterations to get a good average Silhouette value. As a result, it consumed a lot of time in this study, especially, during running the code in software R.

Another one that can be researched deeply in the future is that clustering accident cases, instead of accident circumstance’s variables. After that, association rules mining can be applied on each cluster, so that list of frequent combinations of factors and their supports as well as confidence can be known and calculated. Then, the result is compared with that of this study. The reason was that, in this study, the similarity measures included the frequency of both occurrence together and non-occurrence together of 2 variables. Hence, the result tells us the similarity between 2 variables, but cannot exactly tell us that whether the combination of factors generated by the models led to a road traffic accident in Great Britain.

REFERENCES

Administrative Data Liaison Service. (2014). Retrieved on March 31, 2016 from http://www.adls.ac.uk/deparment-for-transport/stats19-road-accident-dataset/?detail.

Beshah, T., & Hill, S. (2010). Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in Ethiopia. *AAAI Spring Symposium - Technical Report, 10* (*1*) 14–19.

Chaturvedi, A., Green, P. E.,& Carroll, D. (2001). K-modes clustering. *Journal of Classification, 18* (*1*), 35-55.

Chong, M., Abraham, A., & Paprzycki, M. (2004). Traffic accident analysis using Decision Trees and Neural Networks. *IADIS International Conference on Applied Computing, 18,* 39–42.

Chong, M., Abraham, A., & Paprzycki, M. (2005). Traffic accident analysis using machine learning paradigms. *Informatica, 29* (*1*), 89-98.

Data.gov.uk. (2015). Retrieved on March 31, 2016 from https://data.gov.uk/dataset/ road-accidents-safety-data.

Department for Transport.(2015). Contributory factors to reported road accidents 2014. *Reported Road Casualties Great Britain: 2014 Annual Report, 1,* 1–13.

Friedman, J. H. (1997). Data mining and statistics: What's the connection? In D. Scott (Eds.), *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics,* 1-7.

Gupta, G. K. (2011). *Introduction to data mining with case studies* (2^nd^ed.). New Delhi: PHI Learning Private Limited.

Hand, D. (1999). Statistics and data mining: Intersecting disciplines. *ACM SIGKDD Explorations Newsletter, 1*(*1*), 16–19.

He, Z. (2006). Farthest-point heuristic based initialization methods for K-Modes Clustering. *Computing Research Repository, 6,* 1-7.

Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical datasets in data mining. *Research Issues on Data Mining and Knowledge Discovery*, *13*, 1–8.

International Traffic Safety Data and Analysis Group. (2014). *Road safety annual report 2014.* Paris: OECD.

Kashani, A. T., Shariat-Mohaymany, A., & Ranjbari, A. (2010). Data mining approach to identify key factors of traffic injury severity. *Promet-Traffic & Transportation*, *23*, 11–17.

Khan, S. S., & Ahmad, A. (2012). Cluster center initialization for categorical data using multiple attribute clustering. In E. Mülle, T. Seidl, S. Venkatasubramanian, & A. Zimek (Eds.), *Workshop proceedings of the 3rd multiclust workshop: discovering, summarizing and using multiple clusterings*, 3–10.

Kumar, S., & Toshniwal, D. (2015).A data mining framework to analyze road accident data. *Journal of Big Data*, *2*(*1*), 26.

López, G., Mujalli, R., Calvo, F. J., & De, O. J. (2013). Accident analysis & prevention - Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks.*ScienceDirect.com*, *51*, 1–3.

Madhulatha, T. S. (2012). An overview on clustering methods. *IOSR Journal of Engineering*, *2*(*4*), 719–725.

Ogwueleka, F. N., Misra, S., & Ogwueleka, T. C. (2014). An artificial neural network model for road accident prediction: A case study of a developing country. *Acta Polytechnica Hungarica*, *11* (*5*), 177-197.

Ralambondrainy, H. (1995). A conceptual version of the K-means algorithm. *Pattern Recognition Letters*, *16*(*11*), 1147–1157.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65.

Shanthi, S., & Ramani, G. (2012). Feature relevance analysis and classification of road traffic accident data through data mining techniques. *Proceedings of the World Congress on Engineering and Computer Science (WCECS 2012)*, *1*, 122-127.

Tesema, T. B., Abraham, A., & Grosan, C. (2005). Rule mining and classification of road traffic accidents using adaptive regression trees. *International Journal of Simulation Systems*, *6*(*10*), 80–94.

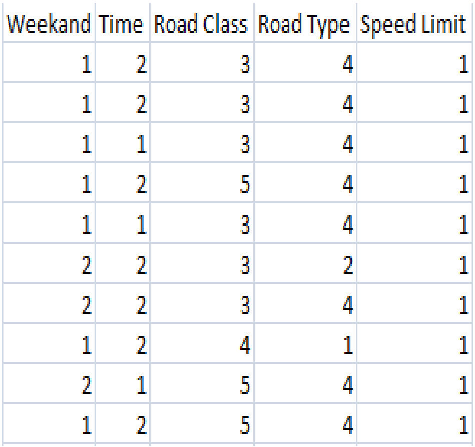
WHO.(2013). WHO | Road traffic injuries. Retrieved on June 13, 2016 from http://www.who.int/mediacentre/factsheets/fs358/en/.

Zhang, N. L. (2004). Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, *5*, 697–723.

APPENDIX A

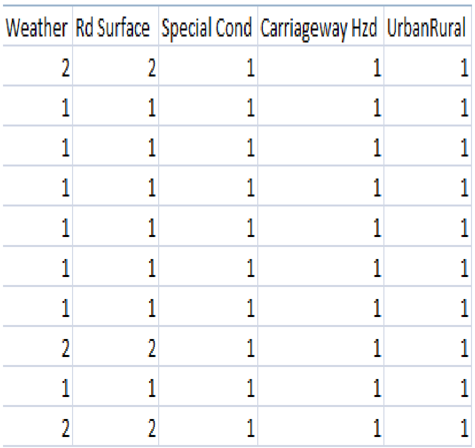
Great Britain’s Accident Circumstances Dataset in 2014

(i) Original form (first 10 rows)



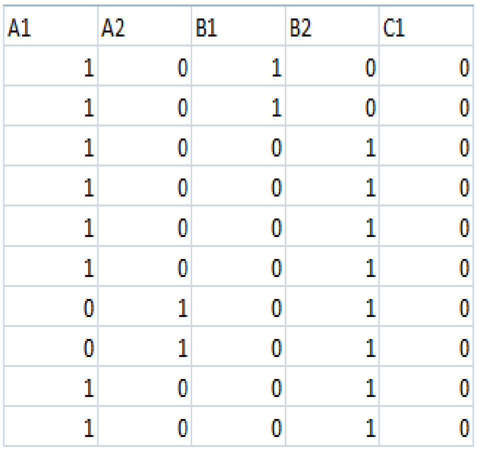
Weekand	Time	Road Class	Road Type	Speed Limit
1	2	3	4	1
1	2	3	4	1
1	1	3	4	1
1	2	5	4	1
1	1	3	4	1
2	2	3	2	1
2	2	3	4	1
1	2	4	1	1
2	1	5	4	1
1	2	5	4	1

…



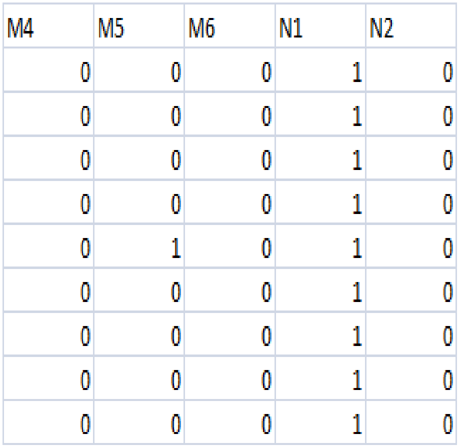
Weather	Rd Surface	Special Cond	Carriageway Hzd	UrbanRural
2	2	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
2	2	1	1	1
1	1	1	1	1
2	2	1	1	1

(ii) Binary form (first 10 rows)



A1	A2	B1	B2	C1
1	0	1	0	0
1	0	1	0	0
1	0	0	1	0
1	0	0	1	0
1	0	0	1	0
1	0	0	1	0
0	1	0	1	0
0	1	0	1	0
1	0	0	1	0
1	0	0	1	0

…



M4	M5	M6	N1	N2
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	1	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0

APPENDIX B

(i) Codes for Data Preprocessing

Matlab Coding:

```
%Import data
%Listwise deletion
NoMiss=(Data.RdTy~=6) & (Data.SpecCon~=9) & (Data.LightCon~=5)
& (Data.WeaCon~=9) & (Data.Surf~=6) & (Data.CrgHzd~=7);
data=data(NoMiss,:);
Data=Data(NoMiss,:);

%Transformation to Binary
fori=1:size(Data,1)
A(i,Data.Wkd(i))=1;
B(i,Data.Time(i))=1;
C(i,Data.RdCls(i))=1;
D(i,Data.RdTy(i))=1;
E(i,Data.SpdLim(i))=1;
F(i,Data.JunDet(i))=1;
G(i,Data.HumCon(i))=1;
H(i,Data.PhyFaci(i))=1;
I(i,Data.LightCon(i))=1;
J(i,Data.WeaCon(i))=1;
K(i,Data.Surf(i))=1;
L(i,Data.SpecCon(i))=1;
M(i,Data.CrgHzd(i))=1;
N(i,Data.UrbRrl(i))=1;
end;
Bin=[A,B,C,D,E,F,G,H,I,J,K,L,M,N];
```

```
%Take Jun data and transpose them
June=Data.Mon==6;
JBin=Bin(June,:);
JBin=JBin';
dlmwrite('JBin.txt',JBin)

%Take December data and transpose them
Dec=Data.Mon==12;
DBin=Bin(Dec,:);
DBin=DBin';
dlmwrite('DBin.txt',DBin)
```

(ii) Codes for *K*-means clustering algorithm

Matlab Coding:

```
%K-means clustering for k=2 until 8
[idx2, cm2]=kmeans(JBin , 2 , 'dist','Hamming');

%{
Plot Silhouette index and calculate average Silhouette index
for both June and December data.
Example of clustering June Data into 2 clusters is shown
below.
%}
[silh2,h2]=silhouette(JBin,idx2,'Hamming')
mean(silh2)
```

(iii) Codes for *K*-modes Clustering Algorithm

(a) R coding:

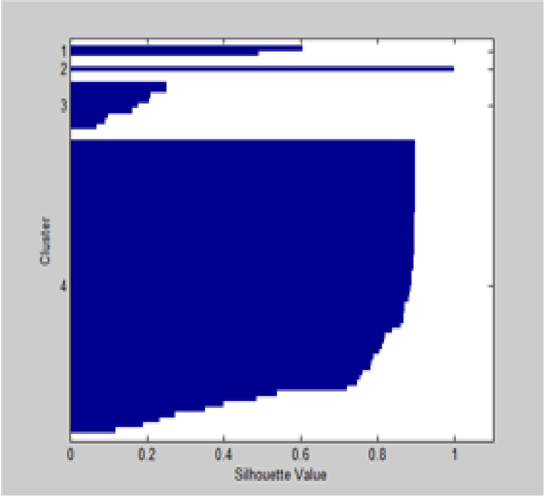
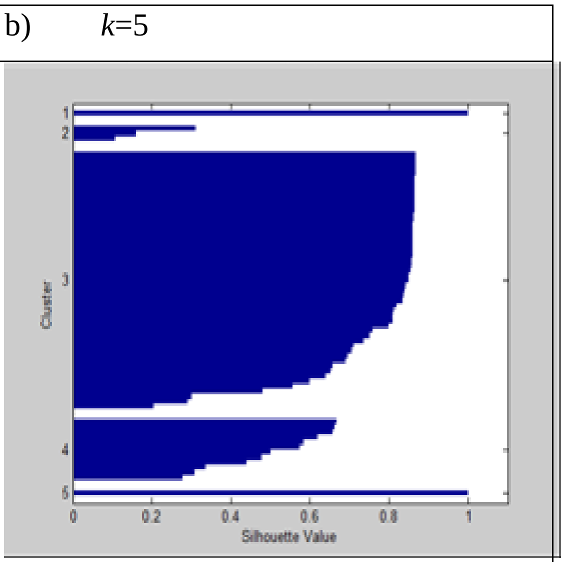
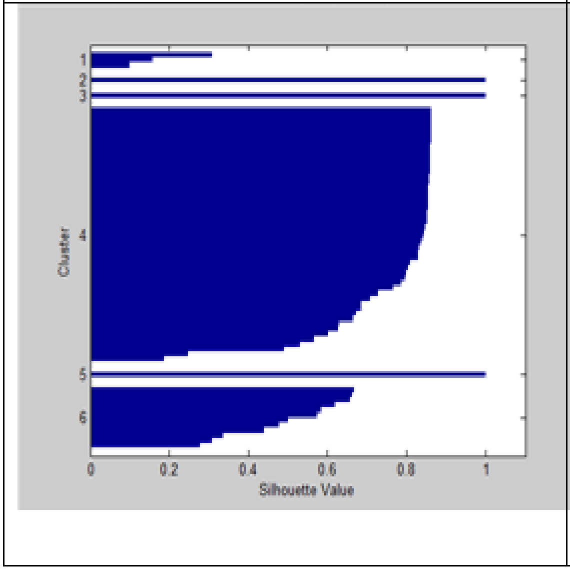
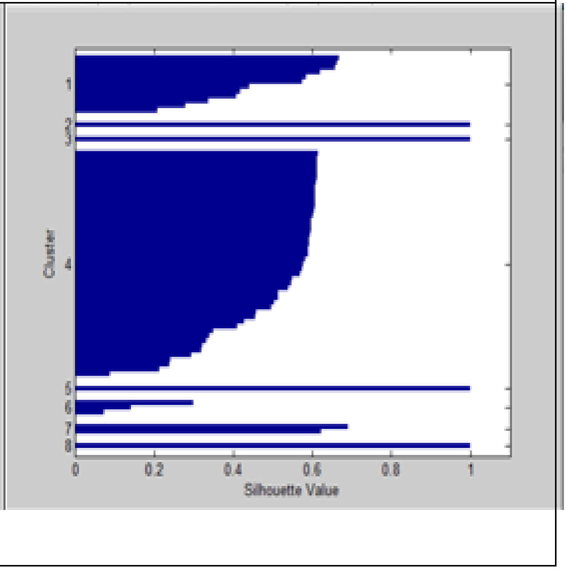
```
# Change directory to where the data locates
# Import them into R
JBin=read.csv("JBin.txt",header=FALSE)
DBin=read.csv("JBin.txt",header=FALSE)

#Install 'klaR' package and open this library
library(klaR)

#K-modes clustering for k=2 until 8 for both June and December data
km2=kmodes(JBin,2)
clus2=km2$cluster

#Save the cluster labels in a file
write.csv(clus2,"clus2.csv")
```

(b) Matlab coding:

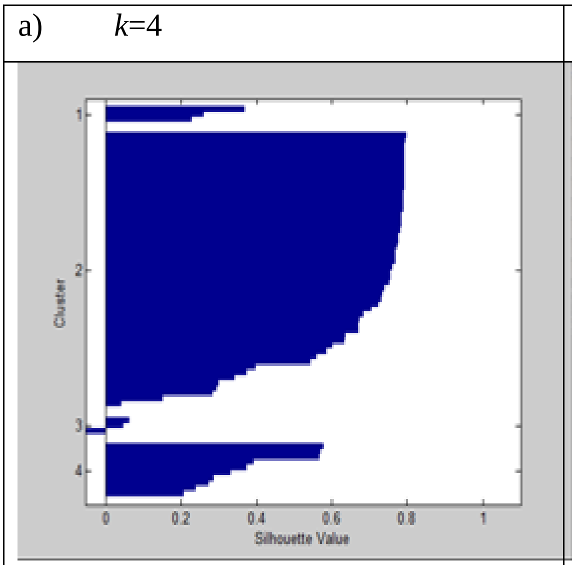
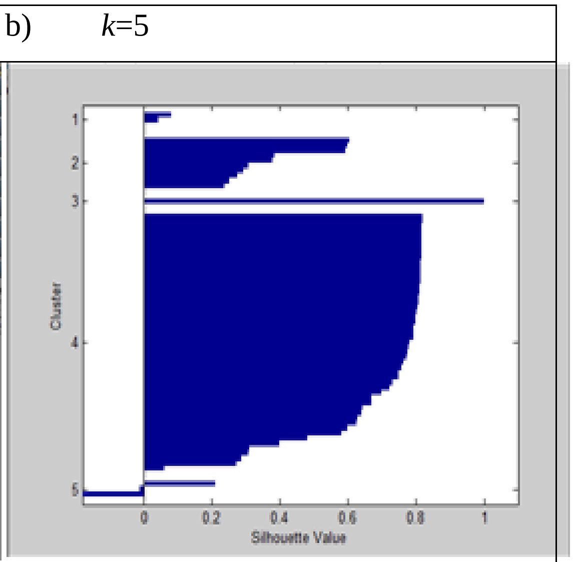
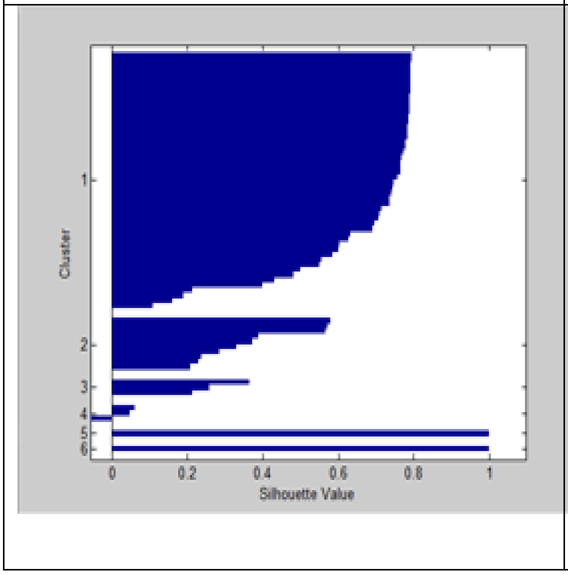
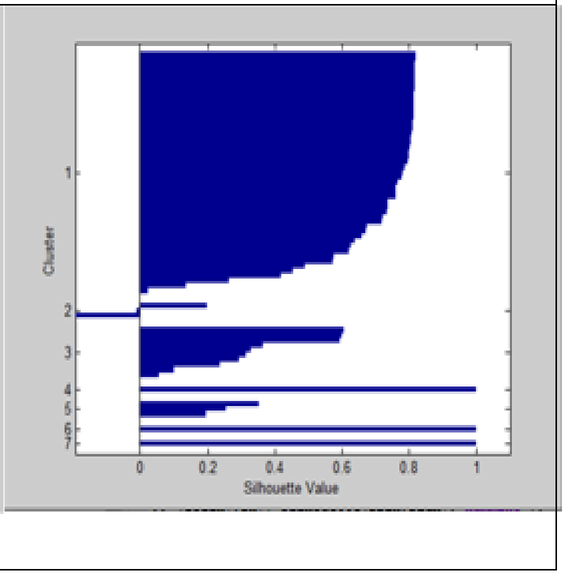
```
%{Import cluster label vector because Matlab runs faster than
R.
Plot Silhouette index and calculate average Silhouette index
both June and December data.
Example of k-modes clustering June Data into 2 clusters is
shown below.
%}
[silh2,h2]=silhouette(JBin,clus2,'Hamming')
mean(silh2)
```

APPENDIX C

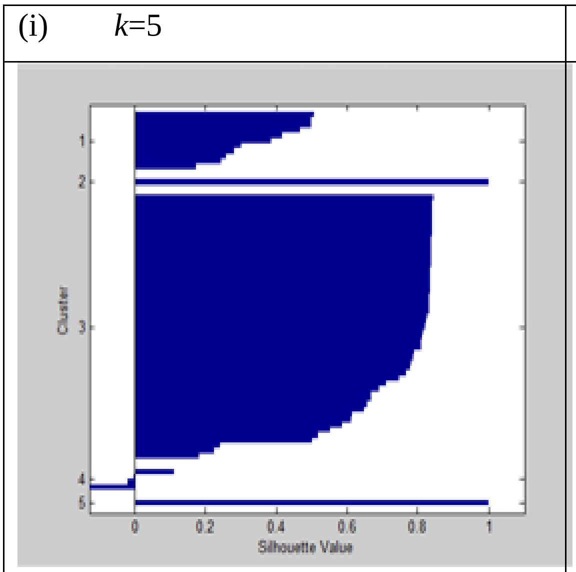
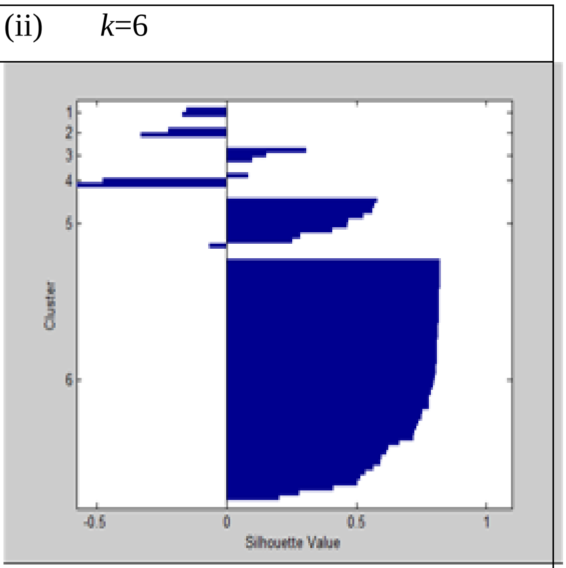
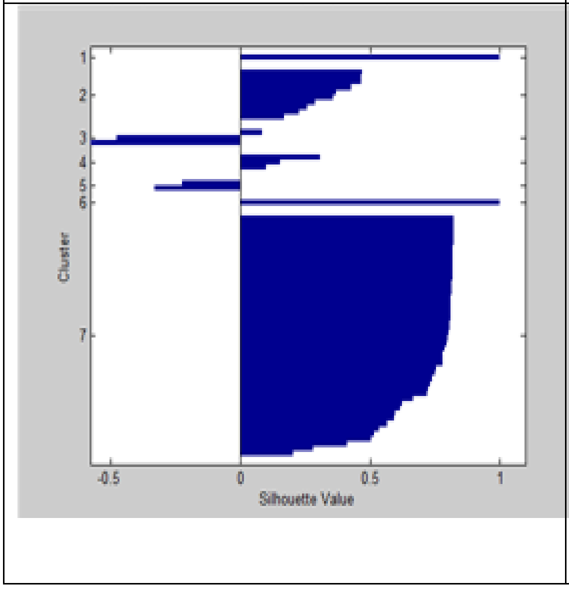
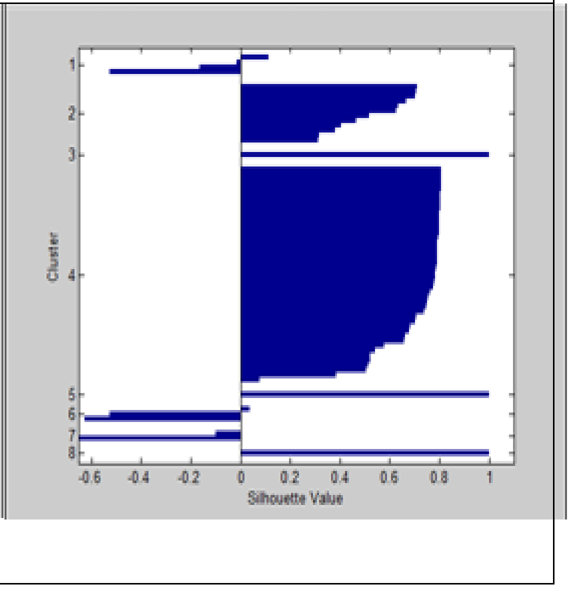
(i) Silhouette plots of *K*-means model on June dataset:

a) *k*=4	b) *k*=5
	
c) *k*=6	d) *k*=8
	

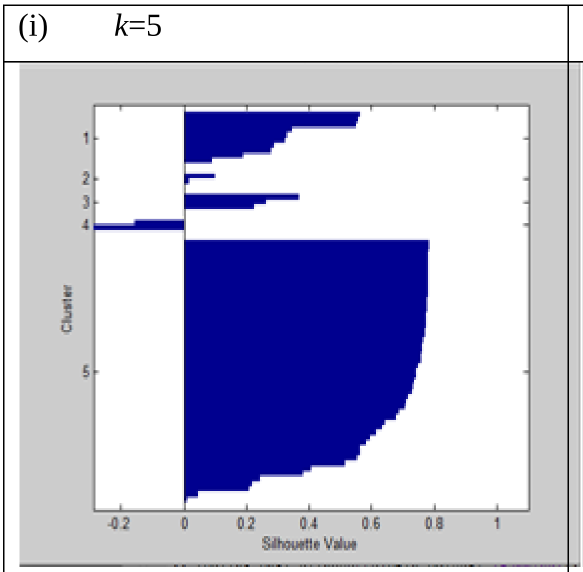
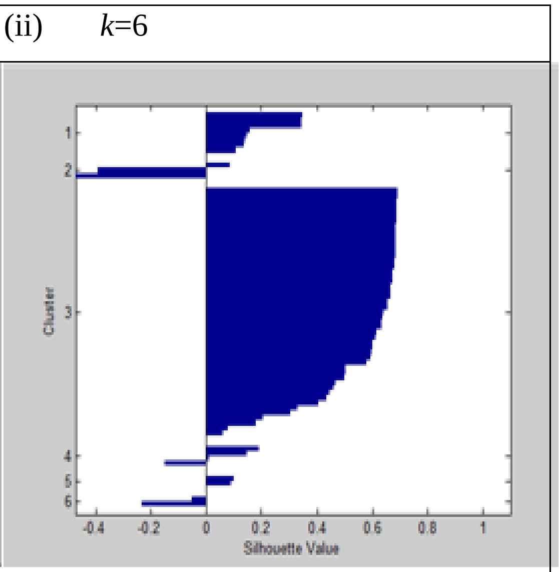
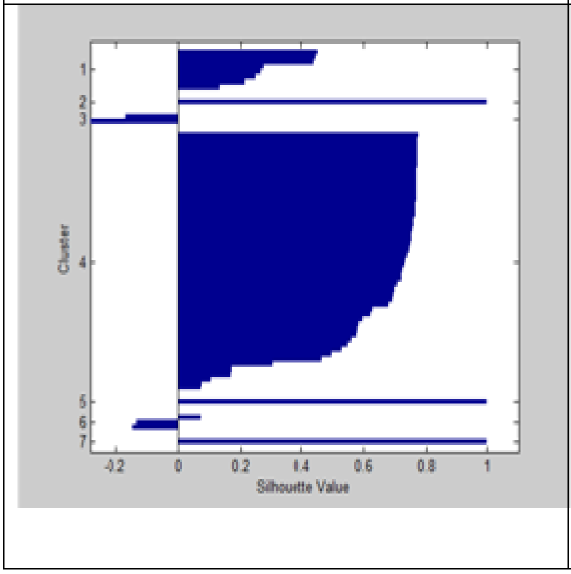
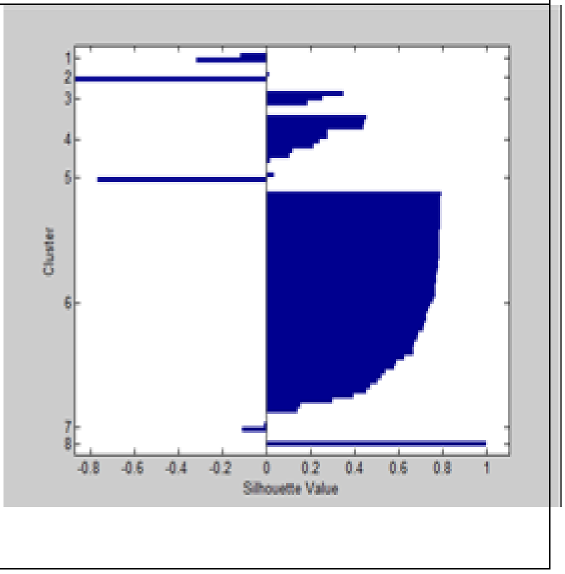
(ii) Silhouette plots of *K*-means model on December dataset:

a) *k*=4	b) *k*=5
	
c) *k*=6	d) *k*=7
	

(iii) Silhouette plots of *K*-modes model on June dataset:

(i) *k*=5	(ii) *k*=6
 	
(iii) *k*=7	(iv) *k*=8
 	

(iv) Silhouette plots of *K*-modes model on December dataset:

(i) *k*=5	(ii) *k*=6
	
(iii) *k*=7	(iv) *k*=8
 Cluster 1 2 3 4 5 6 7 -0.2 0 0.2 0.4 0.6 0.8 1 Silhouette Value	 Cluster 1 2 3 4 5 6 7 8 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6 0.8 1 Silhouette Value