Road accidents analytics with data visualization: a case study in Shah Alam Malaysia

N Shaadan¹, M I K Azhar Suhaimi², M I Hazmir³ and E N Hamzah⁴

¹²³⁴Centre for Statistical and Decision Science Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), 40450, Shah Alam, Selangor, Malaysia
¹Business Datalytics Research Group, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Bukit Ilimu, 18500, Machang, Kelantan, Malaysia

Email: shahida@tmsk.uitm.edu.my

Abstract. Road accident has become a serious problem of concern due to the increasing trend of the occurrences in line with the increase in the number of registered vehicles in Malaysia year by year. Causes of road accidents may come from various factors. Information on the causes is important to increase knowledge in assisting various responsible bodies with the theory and framework for establishing appropriate regulation, policy as well as for intervention planning and purposes in controlling and managing road accidents problem. Thus, this study aims to produce a road accidents profile of Shah Alam based on secondary data from MIROS and PDRM using descriptive and inferential data analytics. The application of data visualization methodology and techniques have shown that among all states in Malaysia, Selangor having the highest frequency of road accidents. For the cases of Shah Alam, there exist a temporal time scale pattern of road accidents occurrence. More number of accident happened during the day compared to night. The most critical time the accident took place is during busy hours; between 8.00 to 10.00 am in the morning, during evening hours between 4.00 to 6.00 pm and night hours after 6.00 pm until 8.00 pm. The accidental death cases were found mostly come from motorbike with (90-250cc), followed by car and lorry and were found mostly occurred at straight-ahead road as well as municipal road. The most type of collision is straying, digressing, or skidding followed by back collision and side collision. The severity of accidents impact is found associated with drivers demographic factors in which accidental death cases involving male is higher than female with age group between 21 to 25 years old.

1. Introduction

Road accident is a serious matter as it could result in fatalities, injuries and trauma. Every year, more than one million individuals have died in road crashes while general road users such as pedestrians, motorcyclists and cyclists are particularly susceptible and present in most of fatalities around the world [1]. In Malaysia, road accidents have been identified as the fourth principal causes of death after cerebrovascular, pneumonia, and Ischaemic heart diseases [2]. Year by year the trend on the number of...
registered vehicles and the number of road accidents show an annually increasing pattern in the country. According to Road Transport Department (RTD), Malaysia loses RM9 billion due to road deaths in 2014 alone. In average, the country is burdened with more than RM10 billion ringgits of loss because of road-traffic accidents annually. As Malaysia is a developing country, the transportation system is important to support continuous economic activity development and growth of the nation. Reducing the cases of road accidents and percentage of fatalities has become important concerns in securing the transportation system particularly the safety of the road users. Several efforts to reduce the incidences have been implemented such as road curve improvement, motorcycle lane, overtaking lanes and junction improvement. Various prevention action are also taken by providing pedestrian facilities, push button crossing, paving shoulders, road safety audits, traffic signs as well as line marking. In particular, black spot treatment programme is an example of a successful effort in diminishing the number of injuries up to a certain extent [3].

The scenario of Malaysian road accidents have been investigated and discussed in many aspects and issues. Among the interesting findings include road accident on North-South Expresswayway (NSE) is shown to have a clustering pattern. This means that, the occurrences of the accidents can be classified into spatial locations [4]. Another study has found that there exists a pattern of increasing number of accidents but decreasing number of casualties in most of the states of Peninsular Malaysia in recent years [4]. Several studies also found that the number of accidents is higher during festive seasons [5,6]. Head injuries is the common injury that causes death and almost 50% of fatal road accidents happened during the evening or at night. Most of fatalities on Malaysian roads are linked to motorcyclists [7] in which about 50% occurred within the 16-30 age groups. It is also found that most of fatalities are relatively circles around young adult group. The highest record of motorcycle fatalities occurred within the west coast area of Malaysia, the region with the most number of registered motorcycles and population. Most fatalities happened during daytime and in clear weather conditions [8]. Road accidents may come from various causes. The accidents cases in Malaysia are commonly due to the driver’s factor, which include incompetent drivers, careless, unfocused, reckless, and aggressive behavior [9, 10]. In other word, it is more due to human behaviour factor [15]. It is also point out that the probability of fatalities and severe injuries increases along with a few other factors such as old age, non-utilization of safety belts, vehicle damage, high vehicular speeds and early morning crashes [3]. Other than that, motorcyclist traffic violation also identified as the main risk factor towards road accidents in Malaysia [11].

In relation to the road accidents problem, a few studies were conducted using Malaysia road accidents data [3, 8, 9, 10, 14] including survey and available secondary data. These studies commonly discussed on the qualitative aspect, causes and associated factors, social influence and impact of road accidents and predictive modelling and forecasting. Even though quite a number of statistics on road accidents in Malaysia have been published, but the statistics usually represent the nationwide pattern and behaviour whereby the pattern of road accidents for a specific area or region is less discussed and limited. Road accidents data often contains multiple variables and the data set is complex and big in the amount. In relation to this problem, an effective way to access, understand, and make the data usable is critically important. Furthermore, an efficient analysis technique that able to help users analyse and reason about data and evidence is also required. One of the promising and growing trend in dealing with complex and big data, is the application of data visualization and analytics. Data analytics refers to the science of raw data analysis comprising of qualitative and quantitative techniques and processes used to mine data to obtain information. Visual data analysis makes it easier for individuals to understand big data and gaining meaningful insights [12]. This become the reason why data analytics and visualization play important role in big data and data science area [13]. It is an efficient tool to represent data set structure, distribution, trend, causes, insights and revealing hidden patterns in data.

Thus, generally, this study aims to explore road accidents data in order to reveal potential new pattern to increase understanding and add new knowledge about accidents scenario particularly in an urban city Shah Alam, Malaysia using data analytics and visualization. The following section 2, 3 and 4 discusses on the research methodology, the analysis results and the conclusion respectively.
2. Methodology

2.1 Data and Sources
The data used in this study is obtained from the traffic unit of Royal Malaysia Police (PDRM) and Malaysian Institute of Road Safety Research (MIROS). The study utilizes secondary road accident data that have been recorded in the past 5 years (2013-2017) period within the area of Shah Alam (provided by PDRM) and the nationwide data in general from (2006-2015) provided by MIROS. Shah Alam is located within the district of Petaling and a portion of the district of Klang in the state of Selangor with total area 290.3 sq km. It is also one of the major cities within the Klang Valley region, located at the western part of Peninsular Malaysia.

2.2 Data analytics Framework
General framework of data analytics process for this research is given in figure 1, starting from problem identification, followed by defining research questions, data acquisition, data understanding, data preparation and pre-processing, data analysis, output representation and visualization, synthesizing the results for meaningful insights gaining and finally communicate the findings. A conclusion can be drawn based on the findings to propose pro-active suggestions as input for decision-making purposes.

Data understanding is the most important stage in data analytics process. It is defined as the process for overviewing the quality, the structure and the content of information available in the data set. In this study, during the data pre-processing stage, missing values are removed from the data set and the analysis was conducted based on the remaining complete data set. The variables involved in this study is shown in the following table 1.

![Figure 1. Analytics Process and framework](image-url)

This study employs two types of analytics; descriptive and inferential analytics. The study objectives are achieved by employing data analytics methodology focusing on data visualization tool.
using Tableau software. Meanwhile, inferential analytics focuses more on the process, procedure and methods to finding evidence towards a given hypothesis related to the problem. It involves with statistical hypothesis testing procedures. Statistical Package for Social Science (SPSS) software is used to execute the inferential analysis in this study.

**Table 1. Research Variables**

| Variable name   | Level of measurement          |
|-----------------|-------------------------------|
| Vehicles type   | Nominal (category)            |
| Injuries level  | Nominal (category)            |
| Injuries type   | Nominal (category)            |
| Time occurs     | Ratio (Hour and minutes)      |
| Collision type  | Nominal (category)            |
| Gender          | Nominal (category)            |
| Age             | Nominal (category)            |
| Road type       | Nominal (category)            |

2.2.1 Descriptive Analytics. Descriptive analytics aims to describe the scenario of road accidents in Shah Alam. It consists of procedure and techniques or tools to describe, explore and gaining insight from data. It is the summarization of raw data. The method is capable in showing the results in many kinds of ways including graphical visualization, tables and summary statistics. Graphic visualization involves with the creation of graphical representations of information. This study employs several proposed type of visualization charts with the following corresponding purposes.

(a) **Bubble chart**, which aims to visualize the clustering pattern on the duration hours road accidents occur so that comparison can be made concurrently, between day and night, between morning, afternoon, evening and night and as well as in detailing the time range of the hours it happened.

(b) **Area trend chart with multiple category**, which aims to visualize the trend of accidence cases according to type of injuries of Shah Alam over the past five years from 2013 until 2017. It is also used to see the pattern of accidents cases according to the type of collision by type of injuries.

(c) **Single line chart with indicator mapping**, which aims to visualize the distribution pattern of road accident cases according to type of vehicle’s and also to visualize the distribution of fatalities cases with respect to type of vehicles.

(d) **Category by category line plot chart**, which aims to portray the pattern of road accidents cases according to the type of injury and by road condition.

(e) **Point chart with temporal and category segregation**, which aims to visualize accident cases according to type of road in Shah Alam and according to annual basis.

(f) **Regional map** which aims to visualize the spatial distribution pattern of total number of road accidents over Malaysia using MIROS data from year (2006-2015).

2.2.2 Inferential Analytics. These analytics consist of procedure, techniques, statistical methods and software or tools to conduct inferential analysis to generalize a population based on sample data. Inferential analysis uses statistical tests to investigate the research hypothesis. The inferential analysis method used in this study involved with non-parametric methods. It is a more flexible method that can
be applied for data that have unknown distribution and for ranked data. The application of the methods will be in accordance to the research objectives. Mann Whitney U test is used to test the study hypotheses of whether there is a significant difference in the average number of deaths as well severe injury of road accidents cases between two population groups of gender, male and female. The test statistics is as in equation (1).

\[ T = S - \frac{n_1(n_1+1)}{2} \]  

where:
S is the sum of ranks in the sample
\( n_1 \) is the number of items in the sample

Kruskal Wallis test is used to investigate whether there is a significant difference in the average number of deaths and severe injury of road accidents cases among several groups of age in Shah Alam. Each of these statements represents the alternative hypothesis (\( H_1 \)). The test statistics is given in equation. (2).

\[ W = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N + 1) \]  

where:
\( N \) is the sum of sample sizes for all samples,
\( k \) is the number of samples,
\( R_i \) is the sum of ranks in the \( i^{th} \) sample,
\( n_i \) is the size of the \( i^{th} \) sample.

The tests are conducted using 5% level of significance. The decision to reject or accept the null hypothesis statement is based on p-value. This value is computed based on the sample data. If p-value is less than the significance level, null hypothesis is rejected.

3. Results and Discussion

Based on data from MIROS, within the period between (2006 - 2015), the spatial pattern of total number of recorded road accidents cases in Malaysia is portrayed in figure 2. Indicated by the color dense, the graphs indicates that Selangor has the highest record followed by Johor. These two states in Malaysia have higher population size compared to other states.
Figure 2. Spatial distribution of total road accidents cases

Specific analysis using PDRM within the period between (2013 and 2017) data focusing in Shah Alam city produces the distribution pattern of accidents cases with respect to temporal time range within a day period. Based on the frequency of the coloured bubbles and also the size presented in figure 3, the data shows that more number of accidents occur during the day than the night. Evening hours between (4.00 to 6.00 pm), night hours between (6.00 to 8.00 pm) and morning between (8.00 to 10.00 am) are the most critical time range of the accidents happened and this contribute about 13.65%, 12.98% and 12.76% of the total cases respectively. This scenario is believed to be due to the peak hour where most people go to and get back from work, causing congestion that may lead to accident. Other alarming time the accidents occurred in Shah Alam is during the afternoon hours between 12.00 to 2.00 pm with 11.66% number of cases.
Figure 3. Accident cases at different time range in a day of Shah Alam

Results in figure 4 shows that from year 2013 to 2017, the percentage of total number accidents with death cases having an increasing pattern so as the severe type and damage cases while light injuries have shown a very drastic decrease trend. The percentage of death cases over the total cases is the highest in 2017 and the least in 2015. The ratio between death, severe and light cases is approximately 10: 6: 5 (shows a descending trend) in 2014 while in 2017 the ratio shows an increasing trend approximately 24: 30: 35. Light injuries is the highest in 2013 and being the lowest in 2017 while severe injuries have the opposite records. These results indicate an alarming scenario of road accident cases in Shah Alam.
Figure 4 Pattern of accidence cases according to type of injuries of Shah Alam

Figure 5 (a) indicates that motorcar contributes the biggest numbers, with 72.75% of the total reported accident cases in Shah Alam, followed by (90-250cc) motorcycle, which contribute to 10.37% and the third is MPV with 4.67%. A study of Manan and Varhelyi [13] has stated that in the road accidents nationwide, motorcycle fatalities are 3 times higher than car fatalities, 6 times higher than pedestrian fatalities and nearly 50 times higher than bus passenger fatalities. Figure 5 (b) suggests that accidents motorcycles fatalities with 90-250cc contribute the highest accidents fatalities which is about (50.3%) of the total accidents cases reported in Shah Alam. Then, followed by motorcar with (30.9%) in comparison with other type of vehicles. The detail of ranking of the vehicles type that contribute the most to the total road accidents cases in Shah Alam can be summarized and recorded in table 2.

Table 2. Most dominance vehicles of road accidents in Shah Alam

| Type of vehicle          | Percentage (%) | Ranking |
|-------------------------|----------------|---------|
| Motorcar                | 72.75          | 1       |
| Motorcycle (90-250cc)   | 10.37          | 2       |
| MPV                     | 4.67           | 3       |
| Small lorry             | 3.49           | 4       |
| Four wheel drive        | 3.40           | 5       |
| Private van             | 1.53           | 6       |
| Taxi                    | 0.90           | 7       |
Table 2 shows the ranking distribution of most dominant vehicles of road accidents for the case of Shah Alam, Malaysia, showing that motorcar has recorded the highest cases between year 2013 and 2017. On the other hand, referring to the statistics in 2012, for nationwide data, motorcycles is recorded to be the highest reported cases with 50.5% and 30% are from motorcars [14].
Figure 5. (a) Accidents cases according to vehicles type in Shah Alam
(b) Fatalities distribution according to vehicles type

Figure 6. Accidents cases according to the type of injury and by road condition
Type of injuries with respect to road condition of the accidents cases in Shah Alam can be visualized and assessed using figure 6. It indicates that for all type of injuries, the accidents mostly takes place at the straight-ahead road. With respect to the category of injuries 70.63% of damage injuries, 72.35% death, 54.76% light and 55.22% severe has reported to occur at the straight-ahead road location. Second type of hot location is seen to happen at the four (4) or more cross roads.

Result of figure 7 shows there are three major type of collisions that contribute most to the accidents fatalities in Shah Alam. The first is of type stray or digress or skid, the second comes from back collision and third is side collision. The trend is seen increasing from 2015 until 2017. Majority of the damage cases are due to back collision and the pattern is seen increasing within the duration year 2013 until 2017. Supported by the finding of research by MIROS, the causes of the associated factors mainly rooted from speeding behavior. Speeding is identified as the most risky behavior among drivers in Malaysia. A study by Manan et al. [8] also highlights on the effect of the motorcyclist’s behavior on traffic safety, the study has indicated that the most statistically significant factors affecting traffic safety were traffic volume and speed limits.

In terms of type of road, figure 8 shows that annually between 2013 and 2017, most number of cases occurred at municipal road in the range between 9,000 to 12,000 recorded cases and the second popular road of the occurrence is at the express road within the range of 2,000 and 3,000 recorded cases.
Figure 8. Accident cases according to type of road in Shah Alam

The results of inferential analytics conducted in this study are as follows. There are more number of male than female users involved in road accidents in Shah Alam. This is shown by the rank statistic in table 3. The p-value of Mann-Whitney test which is less than 0.05 (< 5% level of significance) provide the evidence a significant difference in number of severe cases between male and female. The results also similar for death cases. This result is supported by the study of Harun et. al [6]. In their study, based on data from year 2010 until 2015, they also found the same gender distribution pattern for the cases in Klang Valley region. In the samples of 233 victims, about 86% fatalities involving male road users and only 14% are females.
Table 3. Testing gender difference in accident cases

| Hypothesis | Mann-Whitney Test (p-value) | Mean Rank |
|------------|----------------------------|-----------|
| There is gender difference in the average number of severe injuries cases | 0.002 | 9.5 Male, 3.5 Female |
| There is gender difference in the average number of accidents death cases | 0.002 | 9.5 Male, 3.5 Female |

Investigation on whether there is a difference in the number of severe and death cases between several age groups are also conducted. Results in table 4 shows that the p-value of Kruskal-Wallis test is also very small (~0.00) and less than the significance level 5%. Thus, the difference in the number of severe and death cases is significantly different among the tested age groups. It is identified that the age group with the highest rank is of age group 21-25 years then followed group with the age that is less than 15 years old. This result suggest young adults and children are the vulnerable group of road accidents in Shah Alam. According the earlier study conducted for Malaysia nationwide data, those who are male and single and aged between 16-25 years are more prone to risky behavior and they agree more to negative behavior [8].

Table 4. Testing age group difference

| Hypothesis | Kruskal-Wallis Test (p-value) | Group with the 1<sup>st</sup> highest rank | Group with the 2<sup>nd</sup> highest rank |
|------------|-----------------------------|------------------------------------------|------------------------------------------|
| There are age differences in the average number of severe injuries cases | 0.000 | 21-25 | < 15 |
| There are age differences in the average number of accidents death cases | 0.000 | 21-25 | 16-20 |

4. Conclusion

Based on previous studies and the national published reports, road traffic accidents in Malaysia has increased steadily over the past years. In this study, with the application of data analytics, an exploration analysis using data visualization was conducted with the aims to visualize the scenario and identify pattern and behavior of road accidents cases in Shah Alam, Malaysia using secondary data within the duration from year 2013 and 2017. Generally, the results showed an alarming accidents trend and pattern. This study has identified that more number of accident occurs during the day compared to night. The most frequent time of the occurrence can be categorized into three interval groups of temporal time range; with the most frequent occurs between 4.00 to 6.00 pm, followed by between 6.00 to 8.00 pm and during evening hours between 8.00 to 10.00 am. Pattern of vulnerable road users are recognized.
User of motorcycle with 90-250cc and male young adults of age 21-25 years old group are more risky towards fatalities, also children less than 15 years towards severe injuries. The severity of accidents impact and the fatalities are significantly different between genders and between age groups. The most critical type of road of the accidents to occur is at straight-ahead road and at municipal road. While the most common collision types is straying, digressing and skidding. Motorcyclist fatalities account for the largest percentage, even though more motorcar cases were reported. The statistics found in this study represents the information for a particular site or location, which in this case Shah Alam area. It is hoped that this paper helps to provide a better insight on the risk factors that lead to road traffic accidents particularly in Shah Alam. It is also suggested, more research can be conducted by focusing at various specific location in the country, which could include a specific rural, urban, district as well as residential areas. This specific site investigation approach may help to gain more detail statistics and is seen could provide more effective input for the purpose of policy making, prevention strategy and intervention program development.

Acknowledgement
The authors would like to thank the traffic unit of Royal Malaysia Police (PDRM), Bukit Aman, and the Malaysian Institute of Road Safety Research (MIROS) for providing the data on road accidents used in this study.

References
[1] Bhattacharya S, Alberini A and Cropper ML 2007 Journal of Risk and Uncertainty 34(1) pp 21-47
[2] Department of Statistics Malaysia (DOSM), 2020. Statistics on causes of Death, Malaysia 2020.
[3] Darma Y, Karim MR and Abdullah S 2017 Sadhana, 42(9) pp 1605-1615
[4] Shahid S, Minhans A, Puan OC, Hasan SA and Ismail T 2015 Journal Teknologi 76(14)
[5] Harun MH, Hamzah SPA and Hamzah NH 2017 Incidence of Fatal Road Traffic Accidents Involving Motor Vehicles During Festive Seasons in Klang Valley 2010-2015 (Bulletin FSK 1(1) 2017) pp 15-27
[6] Mustafa MN 2005 Overview of current road safety situation in Malaysia (Highway planning Unit, Road Safety Section, Ministry of Works) pp 5-9
[7] Abdelfatah A 2016 Traffic Fatality Causes and Trends in Malaysia, (Malaysia Sustainable Cities Program, Working Paper Series)
[8] Manan MMA, Varhelyi A, Celik AK and Hashim HH 2018 LATSS Research 36(1), pp 30-39
[9] Malaysian Institute of Road Safety Research (MIROS) 2017 MIROS annual report.
[10] Nawawi MN, Nizam A and Samsudin H 2018 Malaysians Journal of Consumer and Family Economics (21) pp 38-49
[11] Harith SH and Mahmud N 2018 Human risks factors and road accident causation among motorcyclist in Malaysia. A Review article. Proc Int. Conf. on Industrial Engineering and Operations Management (Bandung, Indonesia)
[12] Knaflic CN 2015 Storytelling with Data: A Data Visualization Guide for Business Professionals (Wiley-New Jersey)
[13] Prasad YL 2016 Data Analytics Made Easy (Notion Press)
[14] Wai AHC, Seng SY and Fei JLW 2019 Fatality Involving Road Accidents in Malaysia: A comparison between three statistical models. Proc. Int. Conference on Mathematics and Statistics (ICOMS’19) (Prague, Czech Republic)
[15] Sukor NSA, Tarigan AKM and Fujii S 2017 Transportation Research Part F 46 pp 509 – 523