REPORT OF ANALYSIS OF 2019 UK ROAD ACCIDENTS

BY

John Abutu Ujah

CONTENTS

1. Introduction
2. Analysis
3. Predictions
4. Report
5. commendations

**INTRODUCTION**

The aim of this Project is to analyze car accidents that have occurred in the UK in 2019 to give insight into the questions posed in this report, ultimately leading to the recommendations to government agencies to improve road safety.

The overall data is divided into three datasets: accidents, vehicles and casualties. A summary of each of these datasets is presented in Table 1. The ‘Accident Index’ is provided in each dataset to identify an accident. Mappings for each of the column

values can be found in the supplied excel spreadsheet ‘Variable Lookup’.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Unique Identifier** | **# of Attributes** | **# of Rows** |
| Accidents | Accident Index | 32 | 117536 |
| Casualties | Accident Index | 16 | 153158 |
| Vehicles | Vehicle Reference | 23 | 216381 |

Table 1: Unprocessed raw data provided.

# ANALYSIS

**DATA CLEANING/ EXPLORATORY DATA ANALYSIS**

# Data Cleaning

The median of the distribution of each column was used to fill the missing values in the column when the values in the column are clustered (i.e. a small Interquartile range and Standard Deviation). The approach was applied to the columns, Age\_of\_Vehicle, Engine\_Capacity\_(CC), Propulsion\_Code,

I filled the missing values with the mode of the distribution of values in a column when there are a few unique values in the column and the mode represents a high percentage of the distribution. By so doing, the distribution would not skew after filling NaN. The approach was applied to the columns, Driver\_Home\_Area\_Type, Casualty\_Home\_Area\_Type, 2nd\_Road\_Class and Junction\_Control.

Forward fill method was adopted on columns where the distribution of the values in the column seem even. This also addresses alteration to the distribution statistics after filling missing values. This method was used to fill NaN in the columns, Vehicle\_IMD\_Decile, Driver\_IMD\_Decile, Casualty\_IMD\_Decile and LSOA\_of\_Accident\_Location.

Interpolation function was used to fill missing values in columns where they are in the range of 5,000 and 8,000 since they represent only about 4% of the data in their respective columns. Those columns are kidding\_and\_Overturning', ‘Hit\_Object\_in\_Carriageway', 'Vehicle\_Manoeuvre', 'Vehicle\_Leaving\_Carriageway',

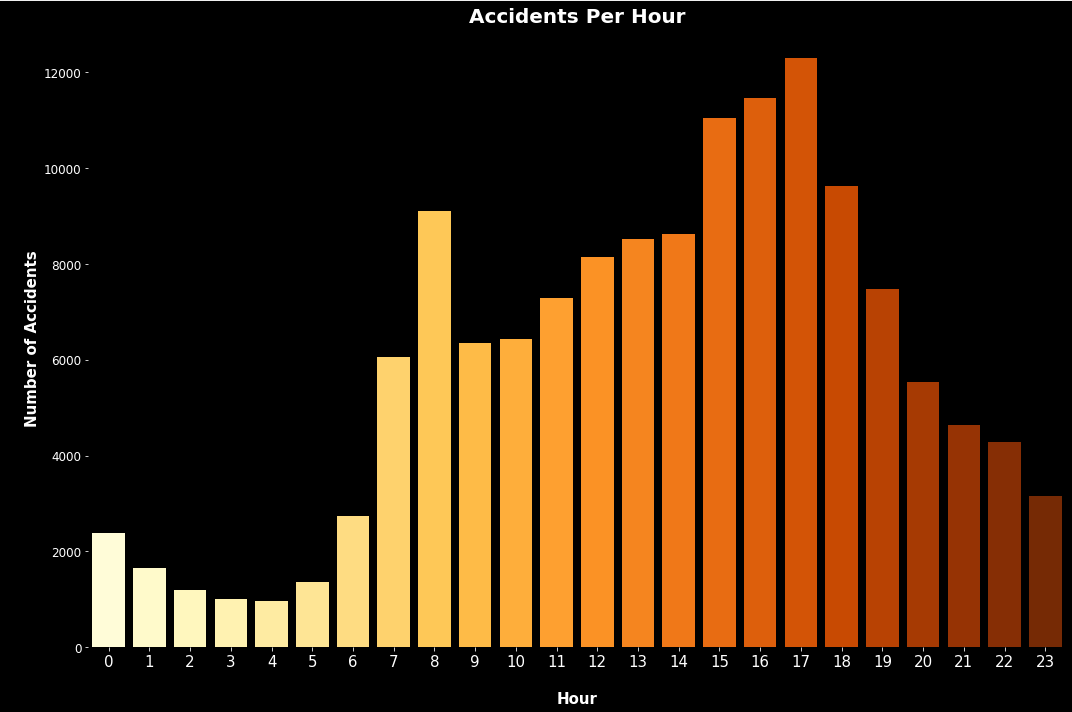
'Vehicle\_Location\_Restricted\_Lane', 'Hit\_Object\_off\_Carriageway', 'Age\_of\_Driver','Age\_Band\_of\_Driver', 'Junction\_Location', 'Was\_Vehicle\_Left\_Hand\_Driv?'.

I dropped Location\_Easting\_OSGR and Location\_Northing\_OSGR because they are almost perfectly correlated with Longitude and Latitude respectively. I also dropped the remaining rows with NaN values since they constitute only about 5% of the entire rows in the dataset.

# This analysis aims to find out the hours of the day and days of the week which have the greatest number of accidents.

**- Hour of The Day Which Have the Greatest Number of Accidents:**

By grouping the data by the hour column I created, I obtained a count of accidents grouped by the hour column and I did a count plot visualization of hour against count. This is shown in fig4 below.



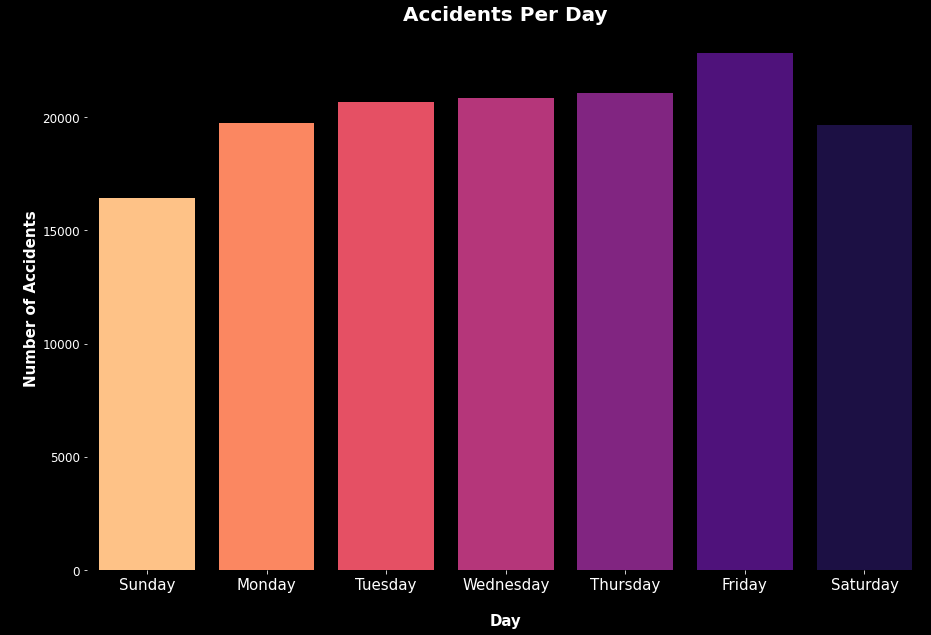
**Fig4:** Accidents per Hour

It can be observed from the distribution that there is a progressive increase in cases of recorded accidents from the early hours of the morning till 08:00am where it peaks for the morning records, and declines by about a third of the morning peak through 08:00am to 10:00am. Afterwards, records of cases begin to rise again to a record daily peak at about 17:00pm following which recorded cases begin to an almost daily low at the end of the day.

# Day of The Week Which Have the Greatest Number of Accidents:

By grouping the data by the day of week column, I obtained a count of accidents grouped by the day of week column and I did a count plot visualization of day of week against count.

This is shown in fig5 below.



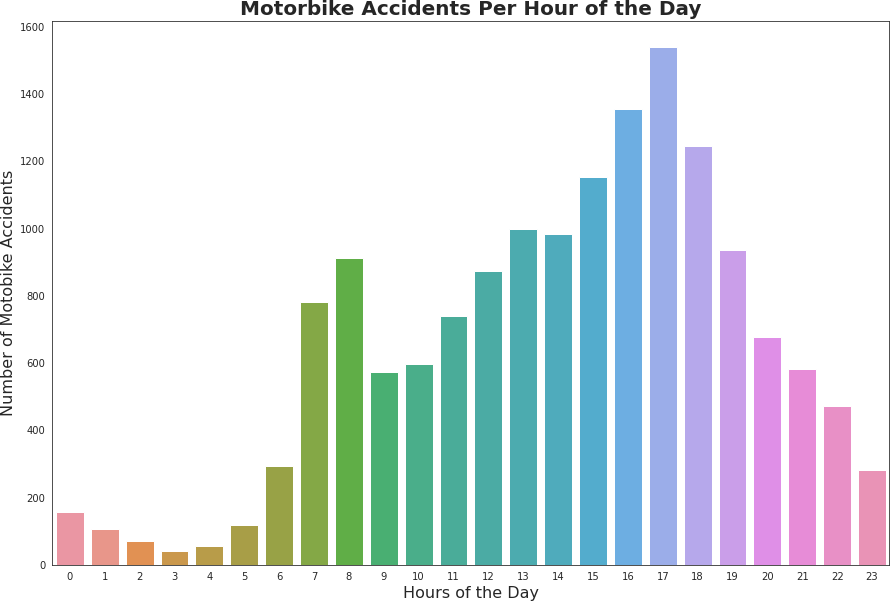
**Fig5:** Accidents per Day

From Fig5, it can be observed that number of accidents remained similar during the week days except Friday where it the total accidents on the day for the year, increased significantly over the other days. Sundays recorded the least accident occurrences for the year.

# This analysis aimed at discovering whether there are significant hours of the day, and days of the week, on which accidents occur for motorbikes.

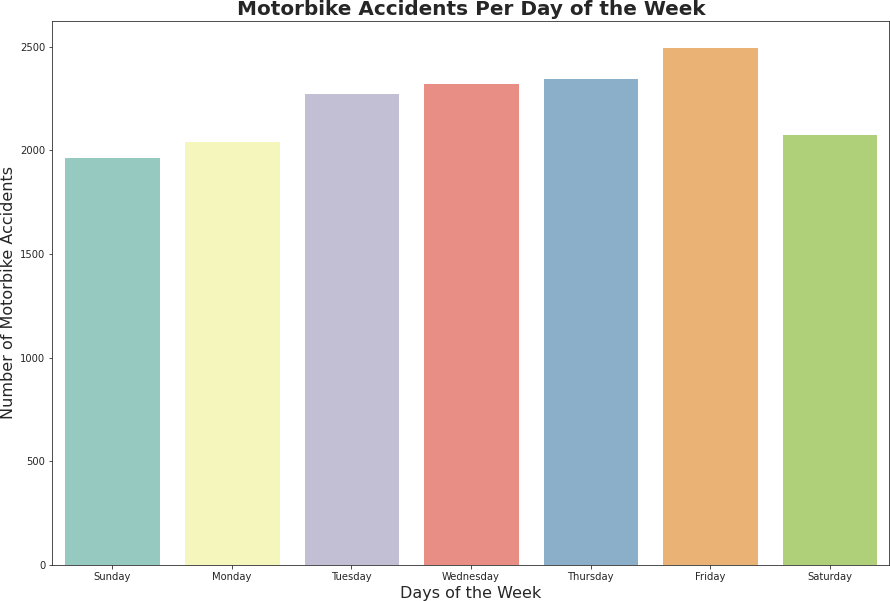
* **Hour of The Day Which Have the Greatest Number of Motorbike Accidents:**

As can be observed from fig6, a plot of motorbike accidents per hour of the day, the total accidents in each hour of the year clearly stands highest at 5pm and generally higher at the period of 3pm to 6pm. Accidents happened least at the early hours of the morning and increased with the day, peaking at 5pm. The number began to decrease into the night in similar fashion it rose.



**Fig6:** Motorbike Accidents per Hour of the Day

# Day of The Week Which Have the Greatest Number of Motorbike Accidents:

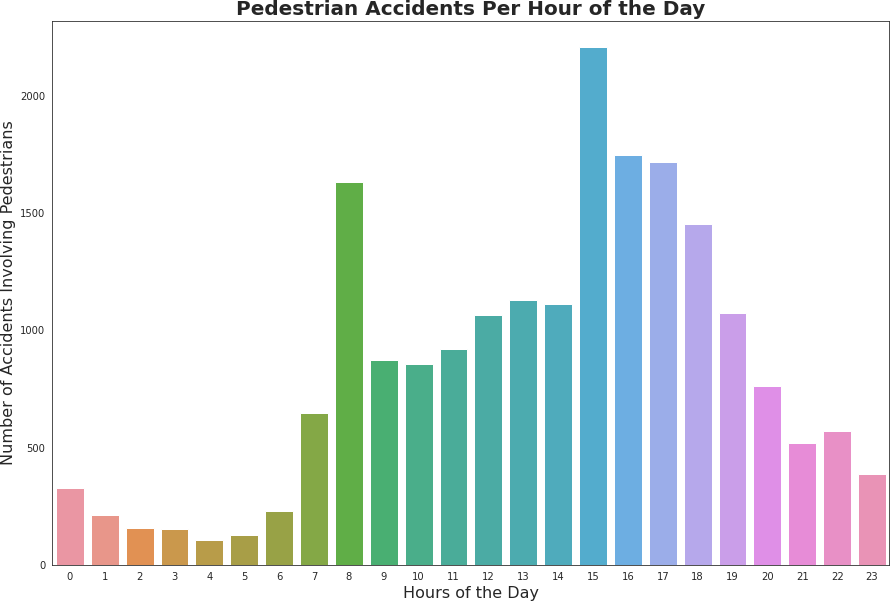


**Fig7:** Motorbike Accidents per Day

From Fig7, while there is a slight increase in motorbike accidents across the week, it can still be observed that they occurred most on Fridays and least on Sunday.

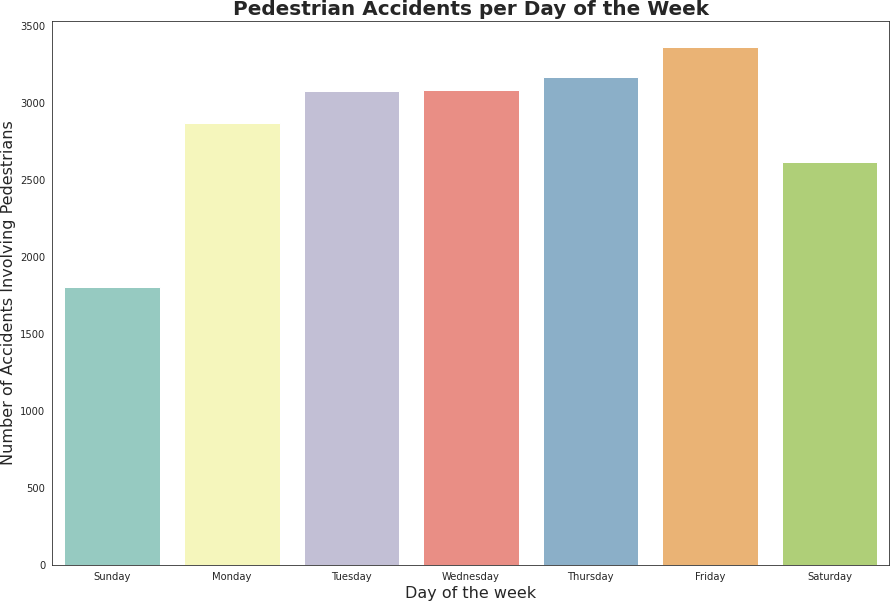
1. **An analysis of hours of the day, and days of the week, on which pedestrians are more likely to be involved accidents.**

* **Hour of The Day Which pedestrians are more likely to be involved accidents:**

It can be observed from fig8 that the hours of 8, 15, 16 and 17 are times when pedestrians are more involved in accidents but worthy of note is that the numbers are significantly higher at 17hour while they are least at 2 to 5 in the morning.

**Fig8:** Pedestrian Accidents per Hour of the Day

* **Day of The Week Which pedestrians are more likely to be involved accidents:**

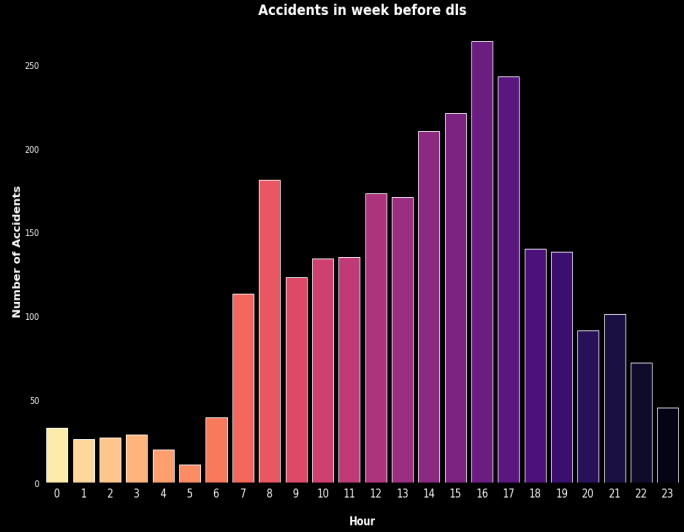
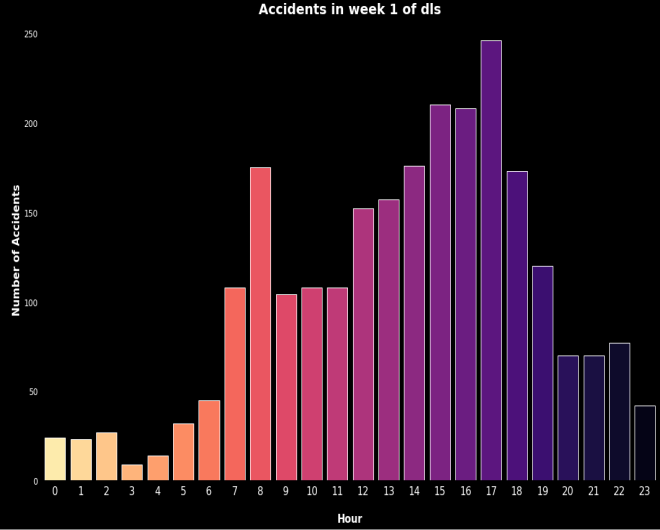
It stands out that pedestrians are more involved in accidents on Fridays but the numbers are considerably high across the other days except Sundays.

**Fig9:** Pedestrian accidents per Day of Week

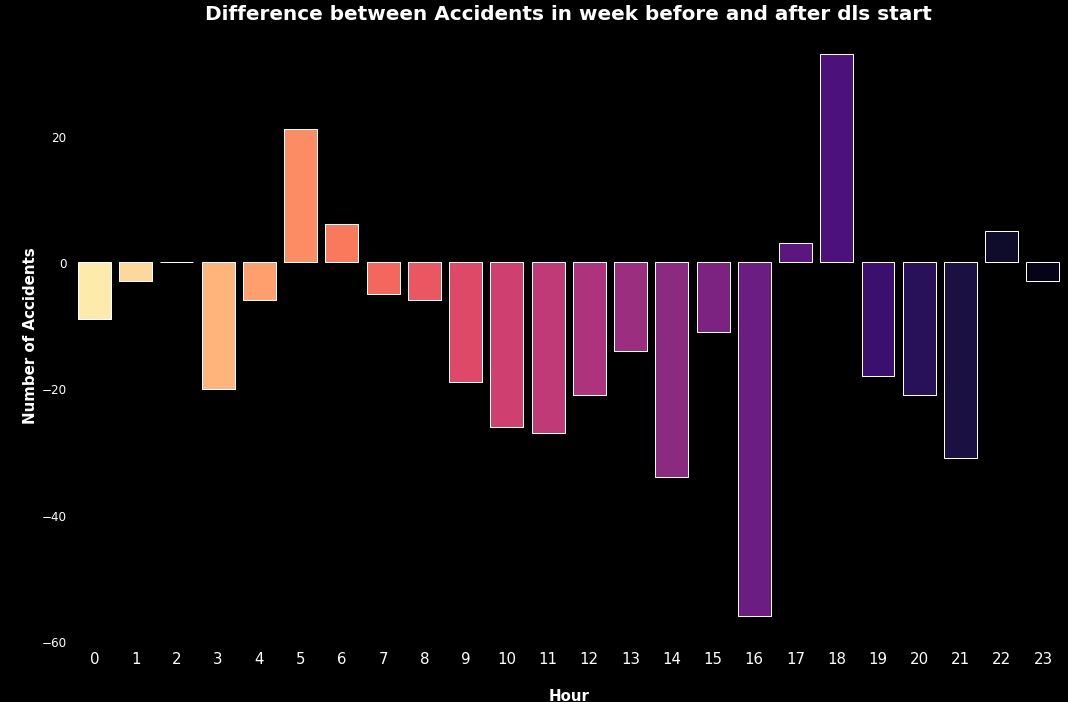
# Impact of Daylight savings on road traffic accidents in the week after it starts and stops.

* **Impact of Daylight savings on road traffic accidents in the week after it starts:**

The start and end date of daylight savings in 2019 were 31st March and 30th October respectively (Barr, 2019) and every analysis was made with reference to these dates. Figures 10 and 11 shows a visualization of the accidents in the week1 and week before the start of day light savings respectively.



**Fig10** Accident in week1 of Day Light Savings**Fig**11 Accident in the week before Day Light Savings



**Fig12:** Difference between accidents in the week before and after day light savings start

The distribution of occurrences in the week before day light savings started was subtracted from those of after it started and from fig12, it can be observed that accidents reduced in the week after the start of day light savings. This claim is however supported by some statistics, **Pooled Variance Method of hypothesis testing.**

# Null Hypothesis(H0):

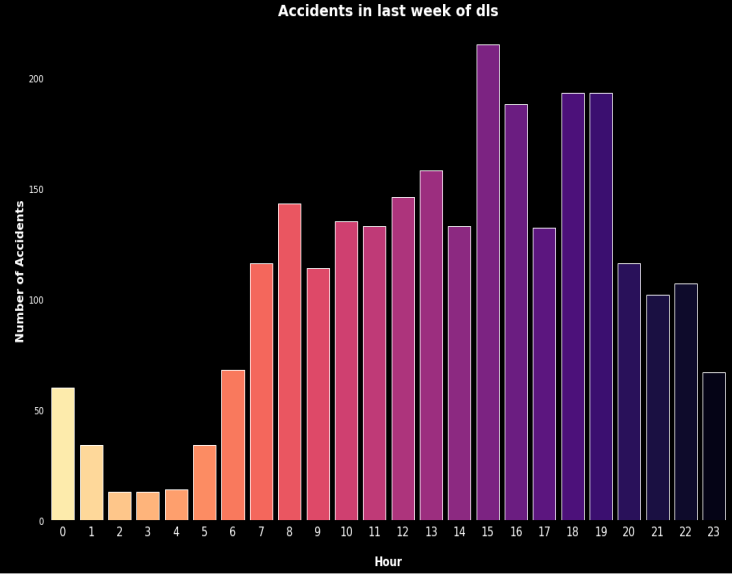
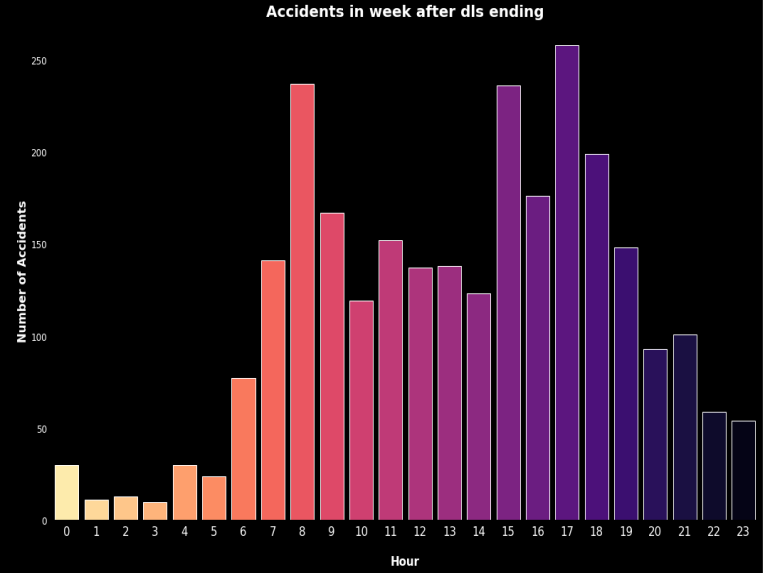
Daylight savings had no impact on road traffic accidents in the week after it starts

# Alternative Hypothesis(H1):

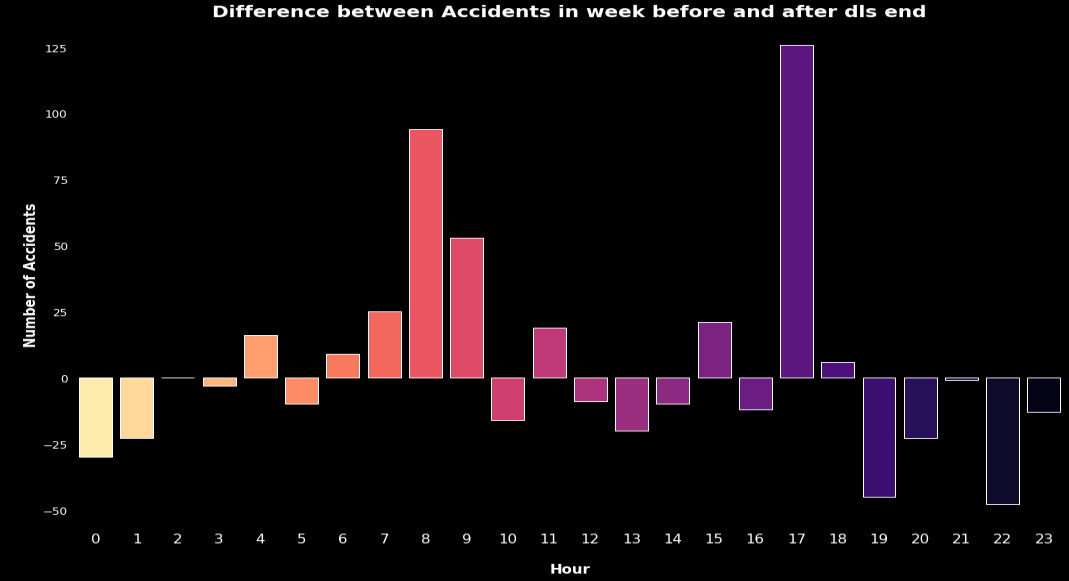
Daylight savings had an impact on road traffic accidents in the week after it starts

The Null hypothesis was rejected since at both significance level of 0.01 and 0.05, the difference in the means of the two distributions was less that twice and thrice the pooled variance of the two distributions respectively.

* **Impact of Daylight savings on road traffic accidents in the week after it stops:** Figures 13 and 14 shows a visualization of the accidents in the week after and week before the end of day light savings respectively.



**Fig13** Accident in week after Day Light Savings Ending**Fig14** Accident in last week of Day Light



**Fig15:** Difference between accidents in the week before and after day light savings end

Fig15 is a plot of the difference in the distribution of week after day light savings ended and the last week of day light savings. It is observed that accidents increased after the end of day light savings. I supported this claim by the hypothesis test: Pooled Variance Method.

# Null Hypothesis(H0):

Daylight savings had no impact on road traffic accidents in the week after it ended.

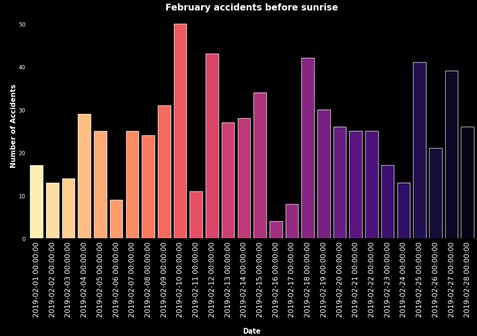
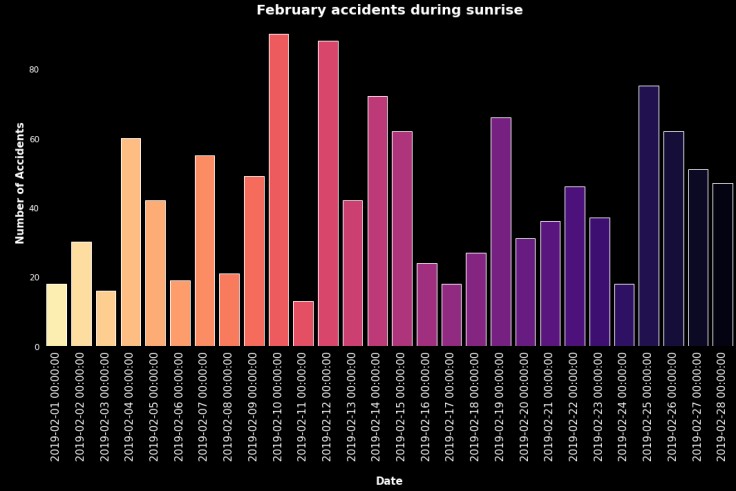
# Alternative Hypothesis(H1):

Daylight savings had an impact on road traffic accidents in the week after it ended.

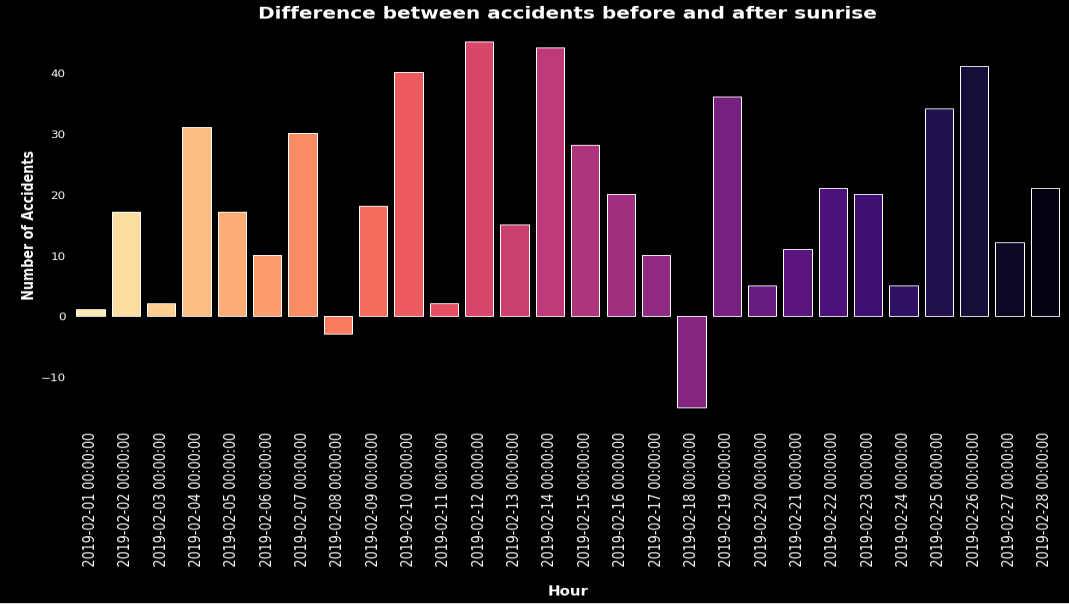
The Null hypothesis was rejected since at both significance level of 0.01 and 0.05, the difference in the means of the two distributions was less that twice and thrice the pooled variance of the two distributions respectively.

# Analyzing the impact, if any, of sunrise and sunset times on road traffic accidents

**- Impact of sunrise time on road traffic accidents:**



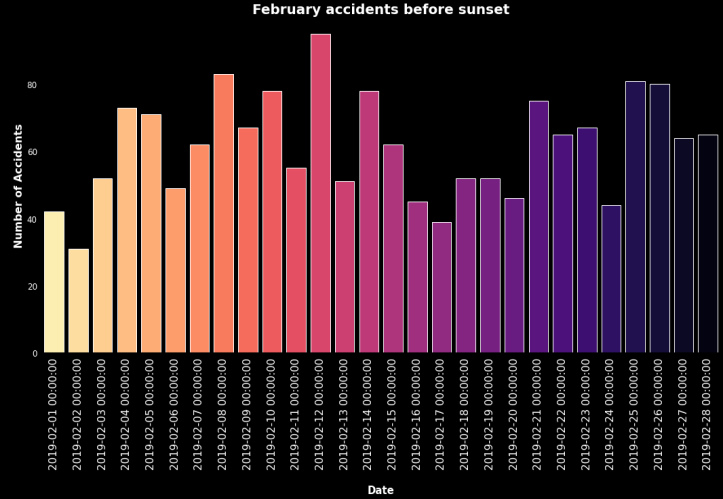
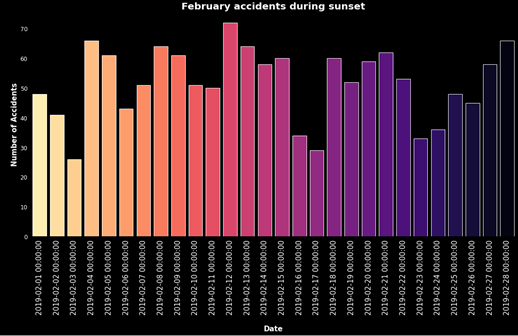
**Fig17** Accident in February during sunrise**Fig18** Accident in February before sunrise



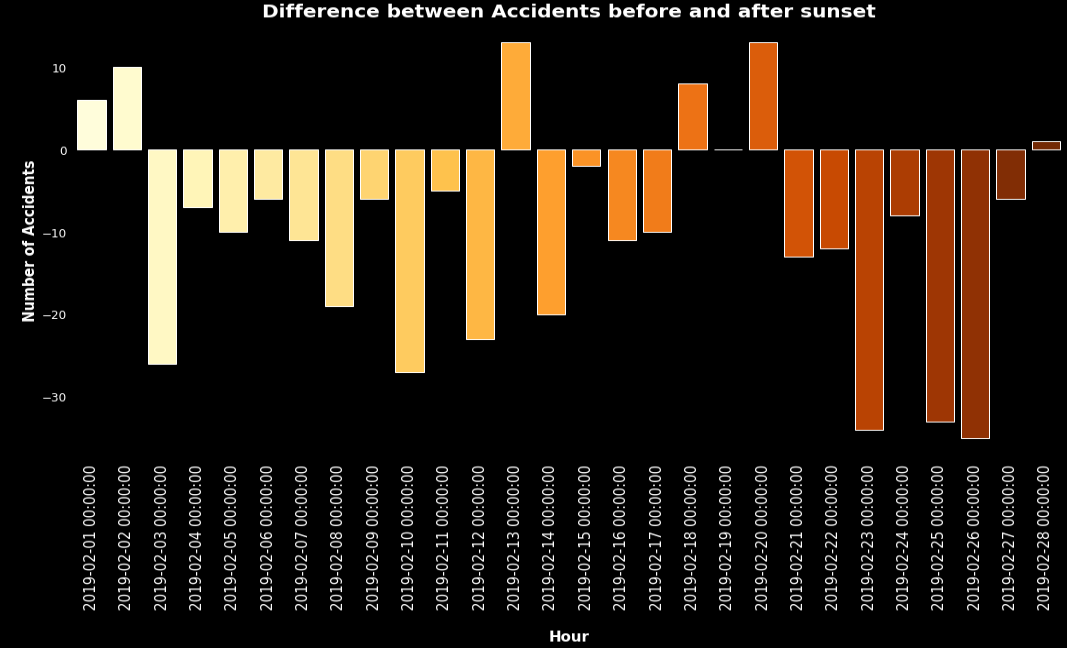
**Fig19:** Difference between accidents in February before and during sunrise

The fig19 shows the difference between accidents in February before and during sunrise and from this, accidents during sunrise was higher than those before, hence suggesting that sunrise time had impact on accidents.

# - Impact of sunset time on road traffic accidents:



**Fig20** Accident in February during sunset**Fig21** Accident in February before sunset



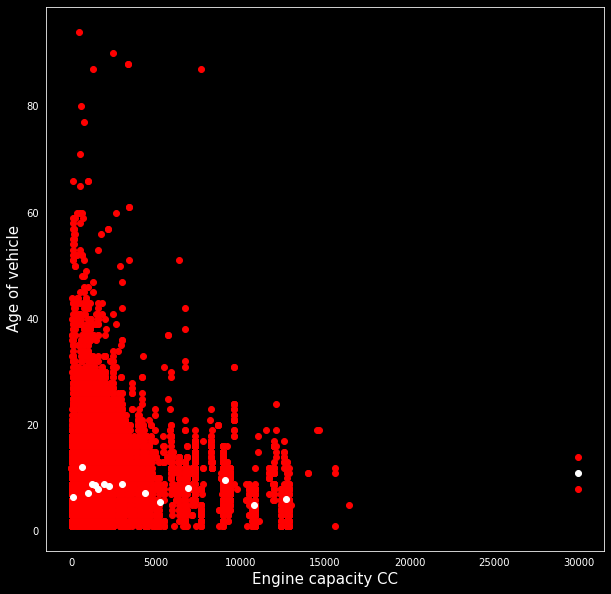
**Fig22:** Difference between accidents in February before and during sunset

The fig22 shows the difference between accidents in February before and during sunset and from this, accidents before sunset was higher than those during sunset, hence suggesting that sunset time had impact on accidents.

# Analyzing particular types of vehicles (engine capacity, age of vehicle, etc.) that are more frequently involved in road traffic accidents.

**i, K-Means clustering of Age of Vehicle and Engine Capacity**

From fig23, recently manufactured vehicles with lower engine capacities are more frequently involved in accidents.

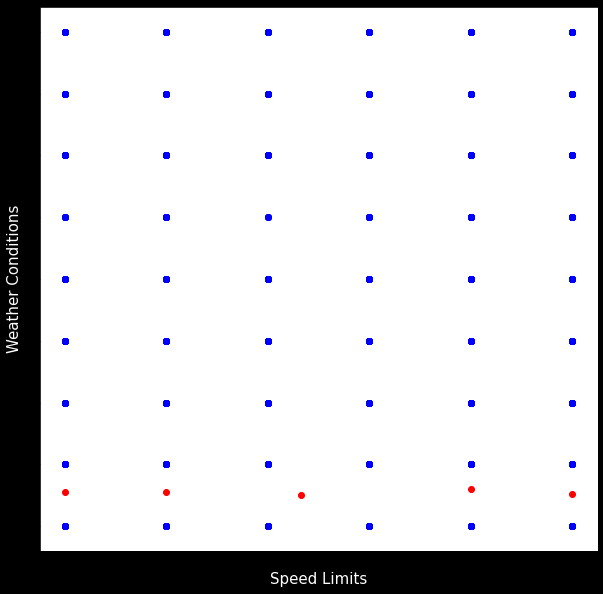


**Fig23:** K-Means clustering of Age of Vehicle and Engine Capacity

# Analyzing conditions (weather, geographic location, situations) that generate more road traffic accidents

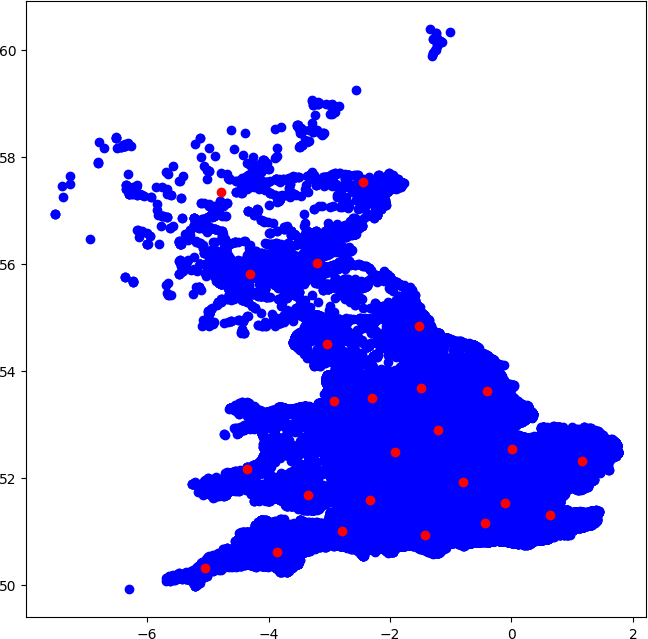
**i, K-Means clustering of Age of Weather Condition and Speed Limit**

From the plot of clustering analysis shown in fig24, conditions of lower weather generate more accidents. From the variable lookup, these weathers are raining and snowing with no high winds.



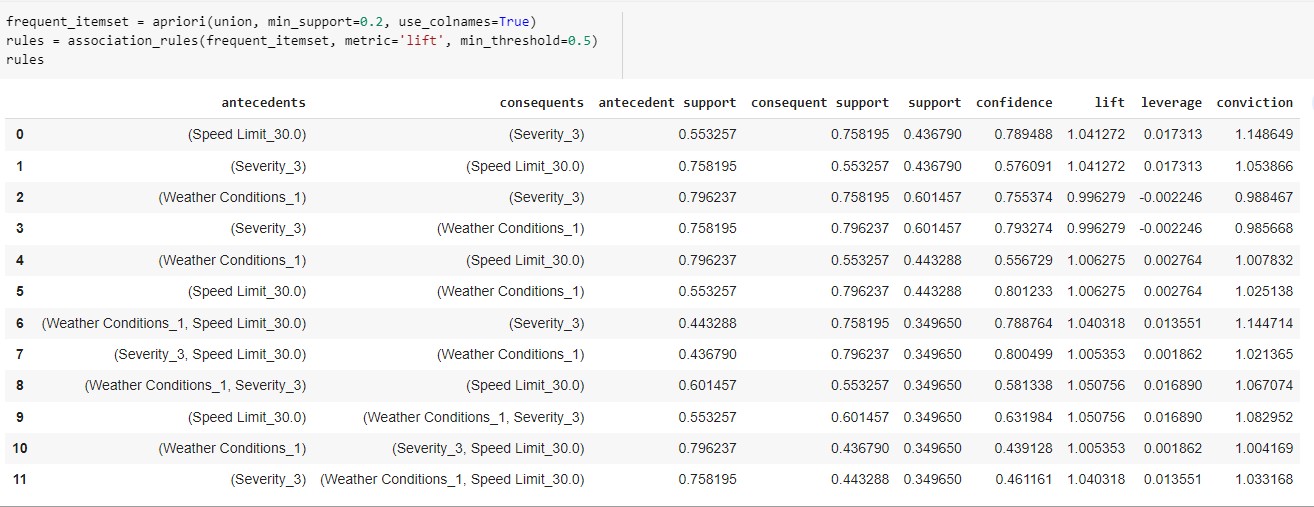
**Fig24:** K-Means clustering of Age of Weather Condition and Speed Limit

# ii, K-Means clustering of Longitude and Latitude

Accidents are clustered more in the south of UK, regions with lower latitude but evenly across longitude as shown in fig25.

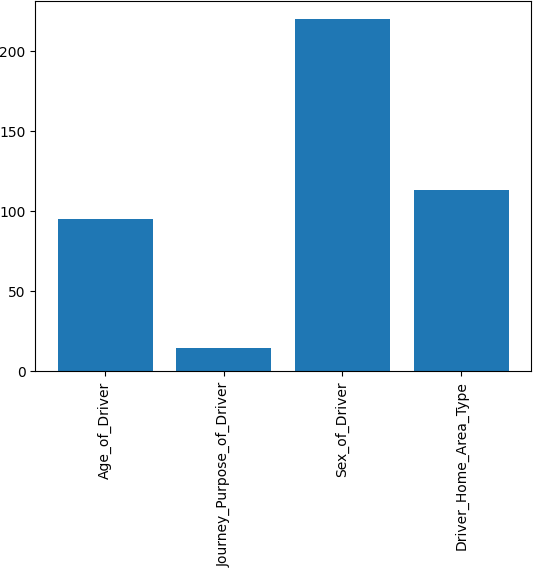
**Fig25:** K-Means clustering of Longitude and Latitude

# iii, Relationship between and severity accidents using apriori.

From fig26 below, slight accidents are related to lower speed limits as they are to fine weather

**Fig26:** Relationship between and severity accidents using apriori

**G. Analyzing how driver related variables affect the outcome (e.g., age of the driver, and the purpose of the journey)**



**Fig27:** Ranking the impact of driver related variable on accident severity

Fig27 is a plot of KBest analysis of the ranking the impact of driver related variables to accident severity and here, sex of driver had the highest rank. Driver’s age an home area type rank next while journey purpose ranked least.

# I. Addressing predictions about when and where accidents will occur, and the severity of the injuries sustained from the data supplied to improve road safety? How well do our models compare to government models?

**-Predicting when and where accidents will occur.**

Using the Kmeans clustering algorithm and the plot of its outcome in the fig28, it can be observed that when the weather conditions are fine(1), accidents occur across many speed limits and a cluster notable is when weather is fogy or misty(7) and speed limit is high.

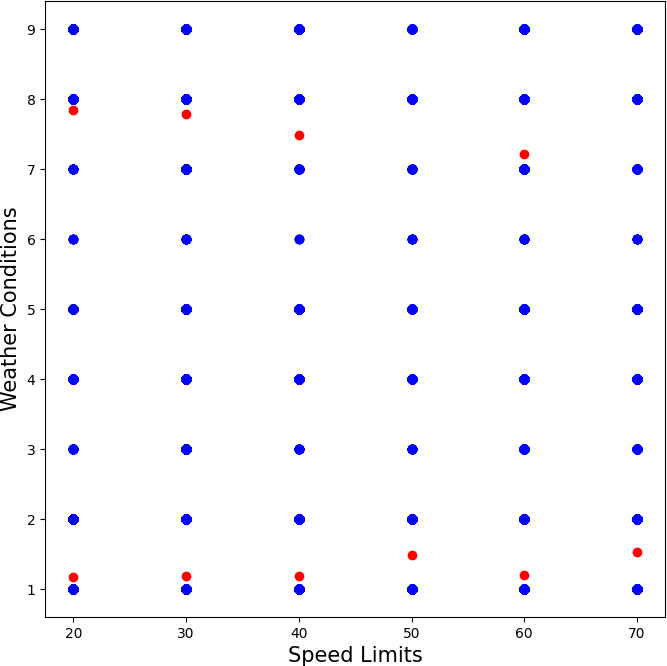


Fig28

# Predicting the severity of the injuries sustained from the data supplied to improve road safety

Below is the plot of the training accuracies of my models and I chose the best model for my validation- Decision tree(93%). The validation accuracy score of my model was 95%. Model was tested against the Government model and it scored 88%

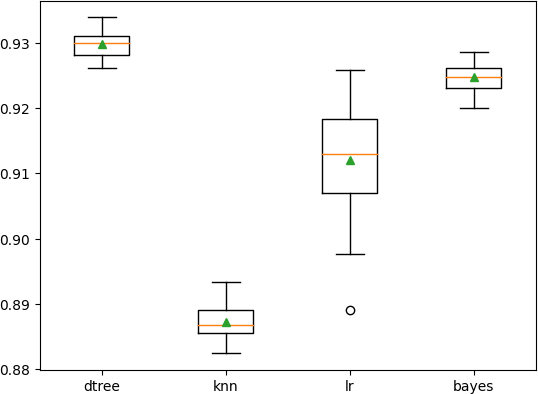


Fig29

RECOMMENDATION

Accidents occurrence is mostly under normal weather and lighting conditions which suggests that most road accidents are caused by human factors. Government should monitor road usage with a view to reduce human factors contributing to accident. One such way is to up police road surveillance and close circuit cameras.

* Another way is to make public transport efficient, provide good infrastructure and encourage their usage since they are less involved in accidents.
* Also, more should be done by the government to constantly provide awareness to road user on the danger of disobeying traffic laws.
* Special attention, in the form of increase surveillance and provision of standby emergency aid should be provided on days and times identified to have more accidents.