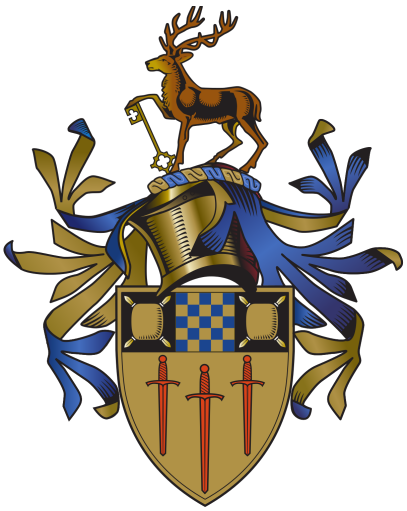
COMM053 Coursework:

Team Dataholics

UK ROAD ACCIDENTS

2005-2015



University of Surrey

Word Count: 5330

Contents

1 Introduction

1.1 Aims

The aim of this report is to analyse car accidents that have occurred in the UK in the period 2005-2015 to give insight into the question: "How can the UK government use its collected data to improve road safety?"

The UK government provides detailed road safety data with respect to personal injury, road accidents, the types of vehicles involved and casualties (if any). The collected data only includes incidences that are reported to the police and recorded using the STATS19 accident reporting form. The data used in this analysis was obtained from the UK government website.[10]

This report considers 'Cross-Industry Standard Process for Data Mining' (CRISP-DM), widely used for data mining and business analytics projects. It is worth noting, before a more detailed description of the data, the obvious problem presented is the total number of accidents happening in the UK every day. More specifically, the number of high severity accidents with casualties. It is known than accidents do not happen randomly without a cause. Human errors are the most common reasons as to why accidents occur. As a result, this report's hypotheses are based on the assumptions that a combination of a variety of internal and externals factors presented in the data can cause a car journey to result in an accident:

- Environmental factors including the location and quality of the road and weather information directly influence the likelihood of an accident to happen.
  - This can help the UK government to better allocate resources for road improvement in order to decrease the number of accidents.
- Seasonality of the number of accidents.
  - This can help the government to identify parts of the year when additional support and control is needed on UK streets to prevent accidents.
- Age of the driver and accident characteristic
  - This can help identify what accident features are common amongst different age groups and thus help ways to improve the education of drivers and thus prevent accidents.

1.2 Assessing Usability of Data

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.In the project plan, the initial proposal was to incorporate all of the features from each dataset however, once data preparation began it was soon realised that there would be a very high amount of dimensions to the data and thus it was decided to not use the Casualties dataset and to only incorporate selected fields from Vehicles dataset (such as Age_of_driver and Age_of_Vehicle). Therefore, the core of the analysis is based on the Accidents dataset. Mappings for each of the column values can be found in the supplied excel spreadsheet Road-Accident-Safety-Data-Guide.

Dataset	Unique Identifier	# of Attributes	# of Rows
Accidents	Accident Index	32	1,780,653
Casualties	Casualty Reference	15	2,589,098
Vehicles	Vehicle Reference	22	3,520,115

Table 1: Unprocessed raw data provided.

The accidents dataset provides information on each accident. Some example data include the coordinates of the location, time, severity, number of vehicles involved, road type etc. Most of these attributes are categorical and any non-integer values such as strings were categorised and given an integer value.

1.2.1 Exploratory Data analysis

Before processing the data for modelling, exploratory analysis was performed to identify initial trends and data characteristics. Figure 1 shows the development of the number of accidents per year from 2005-2015. There is an overall decrease in the number of accidents year-on-year (except in 2014 where there was a slight increase from 2013). Further analysis will be needed to identify why this is the case, but historical improvements in the overall reduction in the number of accidents can be attributed to advancements in road safety which include: updating the theory and practical tests for new drivers [7], multi-million pound investments in improving dangerous roads around the country [9] and stricter laws on punishments for the use of mobile phones [8] (post 2003).

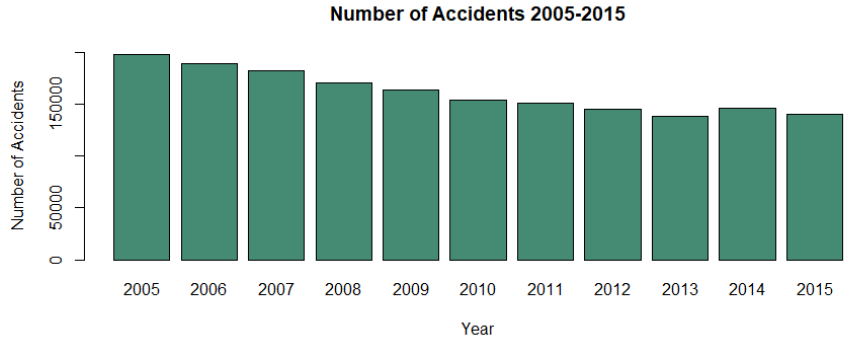


Figure 1: Number of Accidents 2005-2015.

The Accidents dataset has a column that defines the severity of the accident: fatal, serious or slight. Figure 2 shows the year-by-year breakdown of the number of accidents where the accident severity is either serious or fatal. Although the number of serious and fatal accidents has decreased since 2005, the overall reduction has been 23% (21,237 serious or fatal accidents in 2015, compared to 27,439 in 2005).

Figure 2: Number of Accidents 2005-2015 where Accident Severity Serious or Fatal.

1.2.2 Hypothesis 1: Environmental Factors and Accident Severity

For the first hypothesis, the aim was to predict the severity of an accident using the environmental factors of the accidents dataset. Some of these factors include weather conditions, road type and light conditions and their frequencies of accidents are shown in figures 3, 4 and 5 below.

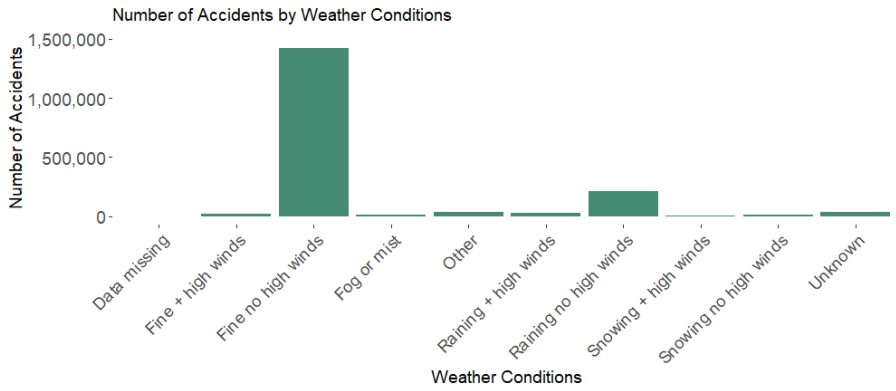


Figure 3: Number of Accidents by Weather Conditions.

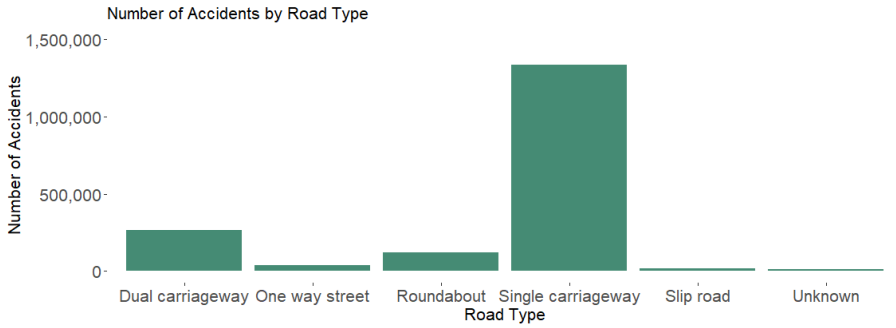


Figure 4: Number of Accidents by Road Type.

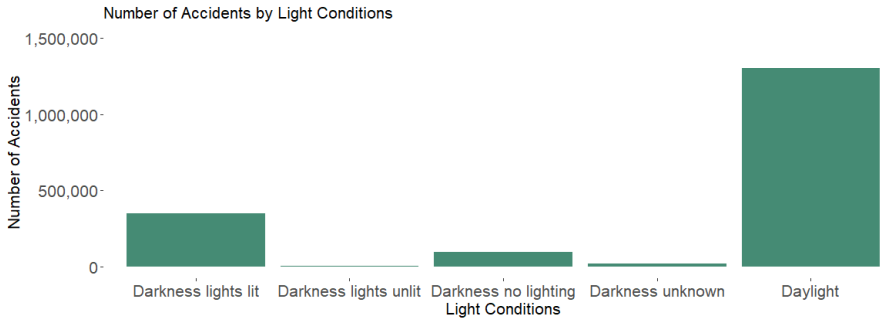


Figure 5: Number of Accidents by Light Conditions.

1.2.3 Hypothesis 2: Number of Accidents Time Series Analysis

An initial look at the number of accidents by year and month is shown in figure 6. As established earlier, the number of accidents has reduced considerably year-on-year however, figure 6 illustrates that accident numbers are higher during the last few months of the year compared to earlier months. Further analysis will be required to find out whether this is because of factors such as weather and light conditions.

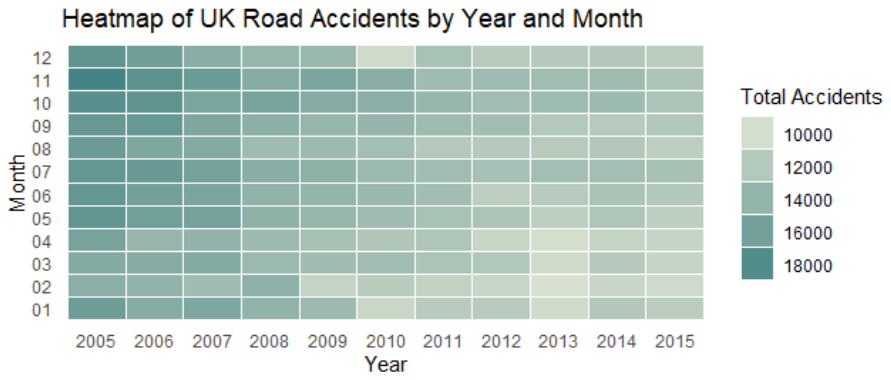


Figure 6: Heatmap of the Number of Accidents by Year and Month. [2]

1.2.4 Hypothesis 3: Age of the Driver and Accident Characteristics

Figure 7 shows a histogram of age against the frequency of accidents. The expected result was to see a higher number of accidents for ages between 18-25 and as the age further increases, a lower frequency of accidents. This is seen in the figure, however it is clearly not a linear relationship.

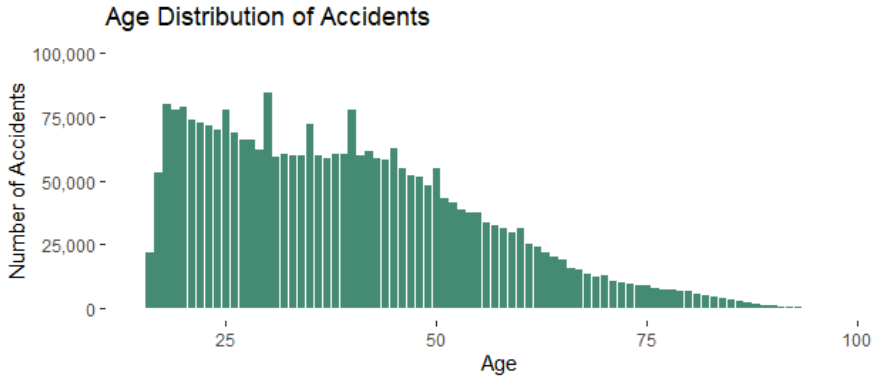


Figure 7: Histogram of Age Against Frequency of Accidents.

2 Data Preparation

2.1 Hypothesis 1: Environmental Factors and Accident Severity

Excluding the location from the accident dataset, this hypothesis aims to gain insight on which environmental parameters play a significant role in the Severity of an accident. More specifically, looking at the Weather, Light Conditions, Road Surface Conditions, Carriageway Hazards, Conditions on Site, Road type, Junction Control and if the accident took place in an Urban or Rural area.

Even though we are looking at all those factors, we are still making a lot of assumptions, for example humidity affects the road surface friction. That in combination with weather conditions and special conditions at cite have all been utilised quite broadly. In other words, since the dataset was given with categorical index values for all the columns, i.e. Weather Conditions: 1 = fine, 2 = Raining with no wind. No specific values for friction, humidity and other parameters are given, and are assumed negligible as they are unknown, for the entirety of this report.

2.1.1 Cleaning & Aggregating

Cleaning:

Once the columns that were inconsequential to our analysis were dropped, further processing needed to be done, in order to derive more meaningful information from the dataset. Initially certain values that were listed as "Other" or "Unknown" were dropped in the columns: {Road_Type, Speed_Limit, Junction_Detail, Light_Conditions, Weather_Conditions, Urban_or_Rural_Area, Junction_Detail and Junction_Control, Pedestrian_Crossing.Physical_Facility, Pedestrian_Crossing.Human_Control}. After the latter processing there was only an $\sim$ 8 % observed data loss. In fact, looking at the proportions of the data redacted see Table 2, we can see that there is an insignificant loss.

Accident_Severity	Accident_Index	Accident_index_Clean
1	0.01291549	0.01319934
2	0.13595013	0.13793461
3	0.85113439	0.84886605

Table 2: Proportions after Cleaning data. Initial Accident_index and after cleaning.

Aggregating:

Once preprocessing took place, we decided to re-index a few columns in an integer ascending scale $0 \rightarrow n$ rather than the arbitrary scale they were provided in. This was done to improve readability and binning. Most columns did not need re-indexing or combining as they were already indexed in integer ascending order. The ones that needed combining and re-indexing are listed below:

- Road_Type: 6→4, 7→5, 12→6
- Pedestrian_Crossing.Physical_Facility: 4→2, 5→3, 7→4, 8→5
- Speed_Limit: 10→1, 15→2, 20→3, 30→4, 40→5, 50→6, 60→7, 70→8
- Light_Conditions: 4→2, 5 & 6 →3, 7→4
- Junction_Detail: 5→4, 6→5, 7→6
- Special_Conditions_at_Site: 1 & 2→1, 3→2, 4→3, 5→4, 6 & 7→5
- Carriageway_'s: 1 & 2 → 1, 3→2, 4 & 5 & 7 →3, 6→4

After both preprocessing and the aggregating process, we are left with a total of $\sim$ 92.6% of the initial data, with the biggest amount of data lost per column amounting to $\sim$ 4% in the Weather_Conditions column. This can also be seen by the proportions in Table 3.

Accident_Severity	Accident_Index	Accident_index_Agg
1	0.01291549	0.01319615
2	0.13595013	0.13794053
3	0.85113439	0.84886333

Table 3: Proportions after aggregating data. Initial Accident_index and after aggregation.

2.1.2 Prediction Parameter

In the original dataset Number_of_Vehicles and Numer_of_Casualties included a few outlier values. Instead of dropping them and replacing them with a mean filling method, we chose to bin them. This made it easier to distinguish between them and run the desired modelling techniques. The new dictionaries are as follows:

	Accident_Severity	Number_of_Casualties	Number_of_Vehicles
Value	**Meaning**	**Meaning**	**Meaning**
1	Single (1)	Single (1)	Single Vehicle Accident (1)
2	Low Count (2, 3, 4)	Low Count (2, 3, 4)	Two Vehicle Accident (2)
3	Medium Count (5 - 10)	Medium Count (5 - 10)	Medium Accident (3, 4)
4	—	High Count (11 - 15)	Big Accident (5 - 15)
5	—	Extreme Count (>15)	Extreme Accident (>15)

Table 4: New Indices after merging.

The data set gives us an overall categorisation of accidents, called Accident_Severity. Using that as a starting point, we had a look at all the unique combinations between Accident_Severity, Number_of_Vehicles and Number_of_Casualties. The combination of Accident_Severity, Number_of_Casualties and Number_of_Fatal_Accidents (serious/slight), was redundant and was not checked for fit. Using a number of clusters between 2 and 15 for each combination, we ran the k-means method. The best clustering, with the least variance and widest spread of new parameters, came from the combination of Accident_Severity, Number_of_Vehicles which can be seen in figure 8.

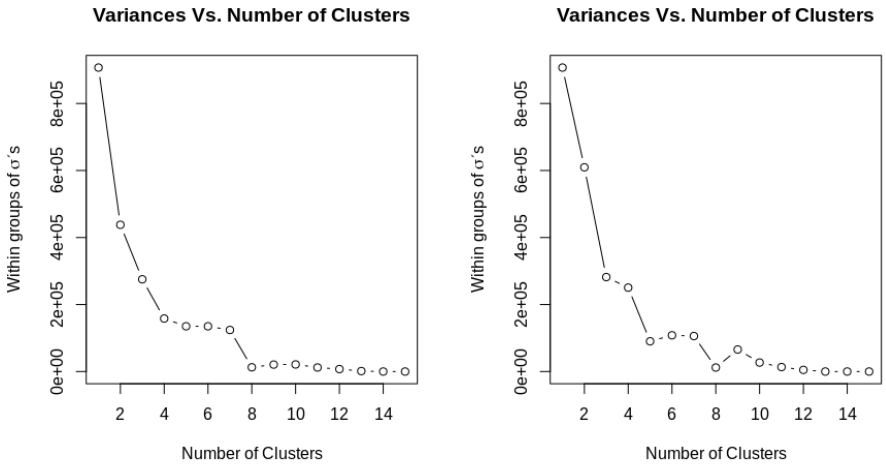


Figure 8: Variance σ against number of Clusters, using k-means method on the number of clusters converging to a nearest mean value of 8.

The above figure, Figure 8 shows the two solutions, both converging, with a nearest mean of 8. We then executed the k-means cluster with 8 clusters and 2500 starting points. The model was saved such that the integrity was maintained for the sake of reproducibility and later analysis. Then re-ordering of the data-frame with a random seed was done, before inputting the required data-frame columns for clustering and modelling.

2.2 Hypothesis 2: Number of Accidents Time Analysis

2.2.1 Selecting & Cleaning Data

After reviewing the available attributes in the dataset Accidents, the columns Date and the count of the accident indexes were selected for pre-processing before modelling as the goal was to predict the number of accidents occurring each day. There were no missing or null date values within the datasets. The attribute Date was provided as a factor in the format "YYYY/MM/DD," which was then converted to the type date using the standard R library function. As there can be multiple accidents on a given date, the number of accidents was aggregated by each date providing the final data frame containing the columns Date and Number of Accidents.

The identified modelling techniques, see section 3.2.1, are Deep Learning and Recurrent Neural Networks which require transforming the data to an appropriate form and then input into the respective model. However, before carrying this out, the time series was analysed for trend and seasonality.

Initially, the decomposition plot of the time series was carried out to identify trend and seasonality and to then remove them from the dataset. This is shown in figure 9, there is a trend of a decrease in the number of accidents over the period 2005-2015. There is a seasonal component, with sharp peaks and troughs towards the end and start of a year respectively. The bottom panel shows the time series without the trend and seasonality. An additive model was produced as the amplitude of the seasonal effect is the same each year.

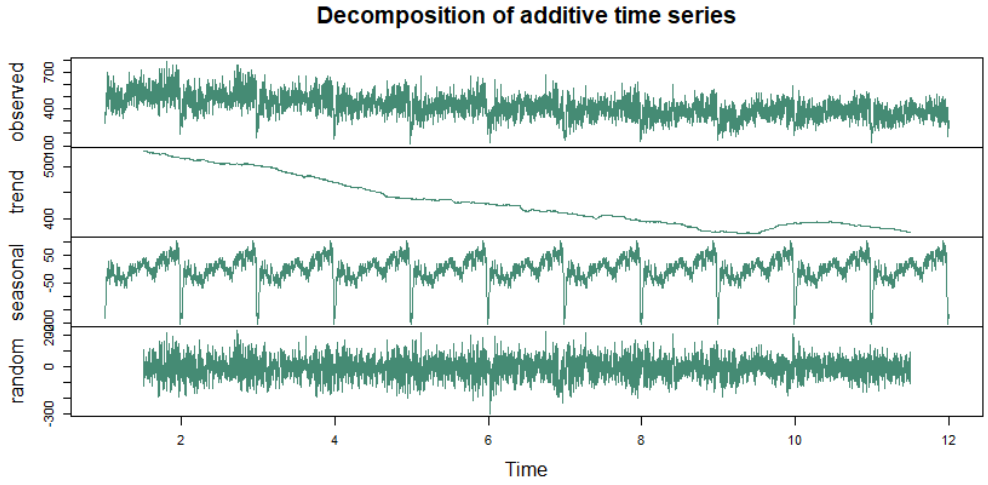


Figure 9: Decomposition of the Time Series.

The next step was to carry out de-trending and scaling. This works by defining a time slot of value n and then de-trending each time slot row in the times series by dividing each of the values for the number of accidents by the mean of all the values in the time slot, and then scaling to have a range [0,1]. Using the sliding window method [12], a window of specified length n (time slot) moves over the data, sample by sample, and the scaled value is computed using the data in the window. The output for each input sample is the scaled value divided by the window of the current sample and n - 1 previous samples. After producing the sliding window, the data was split into train and test. This is further discussed in the modelling section as the test and train data for the deep learning model and the RNN was different.

2.3 Hypothesis 3: Age of the driver and Accident Characteristics

2.3.1 Selecting & Cleaning Data

In the interest of how age plays a role in road accidents, this hypothesis required extracting the age of the driver and vehicle from the Vehicles dataset and then merged with the accidents dataset. The columns of this merged dataset were checked to identify whether they would be suitable for k-means clustering. This was carried out by plotting histograms of each column and where most of the data was unknown for a column it was not selected for processing. The main reason for this was to reduce the dimensionality of the data. However, the drawback of the method used was the loss of information. For example, the column Junction_Control was removed as over half the data was missing. The selected columns for processing for the modelling were: {Police_Force, Accident_Severity, Number_of_Vehicles, Road_Type, Speed_limit, Light_Conditions, Weather_Conditions, Road_Surface_Conditions, Age_of_Driver, Age_of_Vehicle, Urban_or_Rural_Area}.

The dataset consisted of 2.1 million rows however, because of the large amounts of data, the k-means clustering algorithm did not work for larger values of k. Therefore, the dataset was filtered using the column Police_Force to only select road accidents that occurred in the City of London or Greater London (which will now be established as London). This yielded a dataset of 265,585 rows. Another dataset was produced to only include fatal or serious accidents in London. This data set contained 28,249 rows. A summary of the two datasets is shown in table 5:

Dataset	Number of Rows
London	265,585
London with Accident Severity	28,249

Table 5: Datasets for the k-means Clustering Analysis

Both datasets required cleaning for k-means clustering. The first step was to separate the ordinal fields from the categorical ones. The ordinal columns in the dataset were Age_of_Driver and Age_of_Vehicle. To clean these columns, they were initially scaled to obtain the z-scores and then normalised to have a range [0,1]. For the categorical columns one-hot-encoding was used and finally both ordinal and categorical columns were combined to produce the input for the model. To further reduce dimensionality, correlation between the fields were assessed and where correlation was greater than 0.9, those columns were removed.

3 Modelling

3.1 Hypothesis 1: Environmental Factors and Accident Severity

3.1.1 Models used

The aim of this hypothesis is to find out if there are certain factors in the environment which have an influence on the severity of the accident. The field of interest for this hypothesis was Accident Severity, for more information, see section 2.1. Based on the low number of certain categories and the extremely uneven distribution of those values, we created a new measure for the severity of an accident called Custom Severity. Several models were tested with both severities in order to identify parameters with a high influence on the outcome of an accident, including: multivariate linear regression, multivariate adaptive regression splines (mars) and linear classifier model. [3]

3.1.2 Assessing the Models

The linear regression model is implemented with a simple multivariate linear function. It is supposed to give a general insight about linear relationships in the data and information about variable importance in those linear relationships. The multivariate adaptive regression splines (mars) method is an industry known, non - parametric regression method. It considers non-linearity and interactivity between variables and can therefore be seen as an extension to linear regression. The earth package [3] in R delivers a function for a mars regression also called earth().

3.2 Hypothesis 2: Number of Accidents Time Analysis

3.2.1 Models used

For the modelling stage of hypothesis 2 several modelling techniques were identified as part of the formation of the hypothesis. These included: Deep Learning and a Vanilla Recurrent Neural Network . RNNs are useful with time series data because each neuron can use its internal memory to retain information about the previous input. A deep learning model was also used to see whether indeed an RNN is better or not. Deep learning for time series has the advantage that it makes no assumption about the underlying pattern in the data, it is more robust to noise [11] and can also learn longer patterns than the RNN. It is worth noting that as we are interested in machine learning applications, a classical time series model was not considered.

Before running each model, train and test data were formed from the cleaned data. For deep learning, it was decided to split the data as 70:30 for train and test respectively and then rounded to the nearest year. Therefore, the training dataset spanned from 2005-2012 and the testing dataset spanned from 2013-2015. The data was now ready for modelling. To carry out the modelling the h2o deep learning function was used from the h2o library. [4]

The initial parameters used for deep learning were chosen arbitrarily and are as follows:

```
BASICNN_EPOCHS <- 150          #Number of Epochs
DEEP_HIDDEN <- c(30,2)        #Number of neurons in each layer
DEEP_STOPPING <- 30            #Number of times no improvement before stop
DEEP_TOLERANCE <- 1e-4         #Error threshold
DEEP_ACTIVATION <- "Tanh"      #Non-linear activation function
DEEP_REPRODUCABLE <- FALSE     #Set to FALSE to test training for each run
TIME_SLOTS <- 20               #Number of days for Moving Average
```

For the second model, Vanilla Recurrent Neural Network (RNN), a different set of train and test datasets were used. The initial train and test period selected was 365 days. The rnn library was selected for modelling using the function trainr.

The initial parameters for the RNN were as follows:

```
RNN_SLOTS <- 365          #Number of days to use to train the RNN
RNN_NEURONS <- 25         #Number of hidden neurons
RNN_EPOCHS <- 5000        #Number of training cycles
NETWORK_TYPE <- "rnn"     #Recurrent Neural Network
SIGMOID <- "logistic"     #Activation Function
LEARNINGRATE <- 0.05      #Gradient Descent Step Size
```

3.2.2 Assessing the Models

The deep learning model was run multiple times with a combination of different parameters. This is summarised in table 6 where the top 7 runs are shown. The first four model runs were tested using a time slot of 10 days and perhaps this could have been a limiting factor in the neural network. Therefore, an increased time slot of 20 days was selected and three extra model runs were carried out to try and improve the model. The combination of the lowest metrics was in model run 7 which had an RMSE of 40.55,

MAE 31.34 and R^2 66.85%. Unfortunately, given time constraints further model runs or model improvements could not be carried out to try and identify how to improve the performance of the model.

Deep Learning Model	Epochs	Hidden	Time Slot	RMSE	MAE	$\mathbf{R^2}$ (%)
1	100	30, 10	10	41.10	32.38	66.54
2	50	30, 10	10	41.77	33.01	65.54
3	150	30, 10	10	41.11	32.38	66.56
4	150	30, 2	10	41.12	32.37	66.49
5	100	30, 10	20	40.60	31.38	66.86
6	150	30, 10	20	40.55	31.34	66.95
7	150	30, 2	20	40.59	31.40	66.93

Table 6: Deep Learning Model Runs

The recurrent neural network was also run multiple times with a combination of different parameters. This is summarised in table 7 where the top 3 runs are shown. The first point to make is that the RNN performs worse than the deep learning – illustrating immediately the weakness of RNN models (the lack of ability to learn long term patterns). For all model runs in table 4, a time slot of 365 days was used to train the RNN. Changing the training time of the RNN decreased performance and thus those runs have not been included.

RNN Model	Hidden Dim	# of Epochs	RMSE	MAE	$\mathbf{R^2}$ (%)
1	25	3500	41.92	32.38	59.77
2	25	5000	40.73	31.06	62.99
3	50	3500	41.58	31.91	60.99

Table 7: RNN Model Runs

3.3 Hypothesis 3: Age of the driver and Accident Characteristics

3.3.1 Models Used

For hypothesis 3 the aim was to group the features of the accidents dataset and identify what combination of features cause accidents amongst different ages. For this analysis k-means clustering was chosen. The advantages of using k-means clustering is its simplicity and high speed performance. However, the drawbacks include the manual

selection of k and the randomness of clustering. When run many times, different clusters are produced and thus affect what trends can be identified from the data.

To implement the k-means clustering the k-means function from the R stats library was used. Unfortunately, due to the vast size of each dataset and lack of computing power it was not possible to use the Elbow method nor the Silhouette method to identify the optimal number of clusters to use. Therefore, as a work around each dataset was run separately using different values of k and then using the plot of each cluster to select the appropriate number for k. The parameter used for the k-means clustering were as follows:

- centers $= i$ where $0 < i < 10$
- nstart $= 25$

3.3.2 Assessing the Models

For the first dataset (London), the optimal value of k was found to be 6. Figure 10 visualises the k-means clusters where cluster 6 shows a fairly clear distinction from the rest of the clusters. Plots with values of k= 4, 5 and 8 can be found in appendix 6.1.

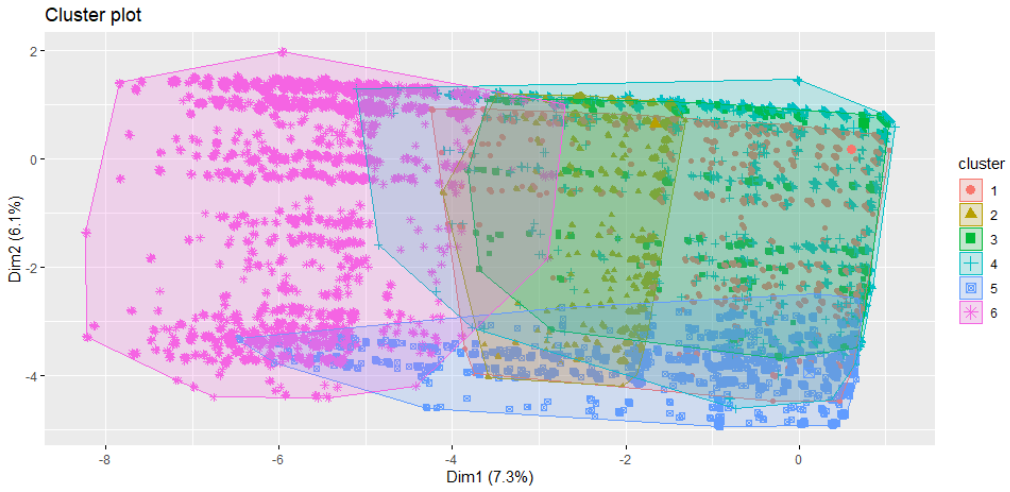


Figure 10: k-Means Clustering with $k = 6$.

For the second dataset (London with accident severity fatal or serious), the optimal value of k was found to be 3. Figure 11 visualises the k-means clusters where cluster 3 shows a fairly clear distinction from the rest of the clusters. Plots with values of k= 4, 5 and 8 can be found in appendix 6.2.

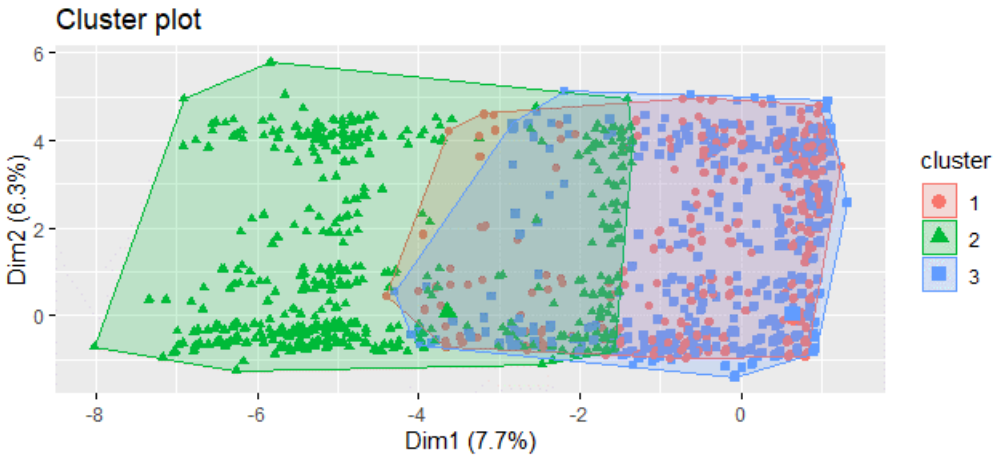


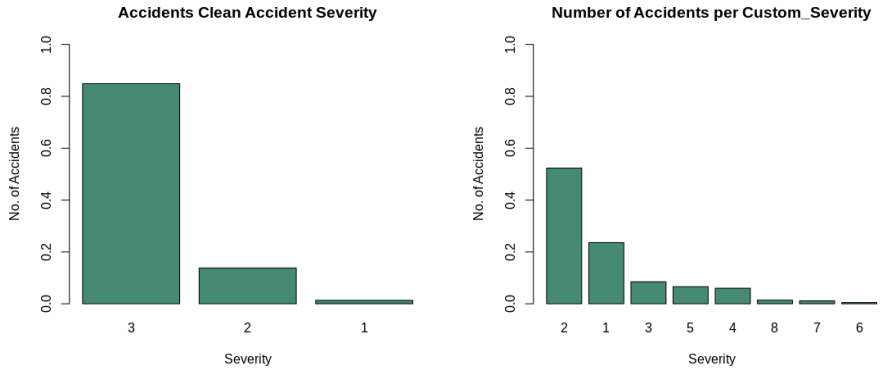
Figure 11: k-Means Clustering with $k = 3$.

4 Results and Evaluation

4.1 Hypothesis 1: Environmental Factors and Accident Severity

The linear Regression is mainly used to show the general importance of the given variables. It is also of interest to see if there were major changes in importance by creating a new severity parameter. This was tested on Accident_Severity and implemented with a multivariate linear function. The metrics of the model implied a pretty good prediction (MAE $= 0.27$, RMSE $= 0.4$, $R^2 = 0.02$). The same model, with the same functions, when tried on the Custom_Severity returned worse results. (MAE $= 0.95$, RMSE $= 1.37$, $R^2 = 0.2$).

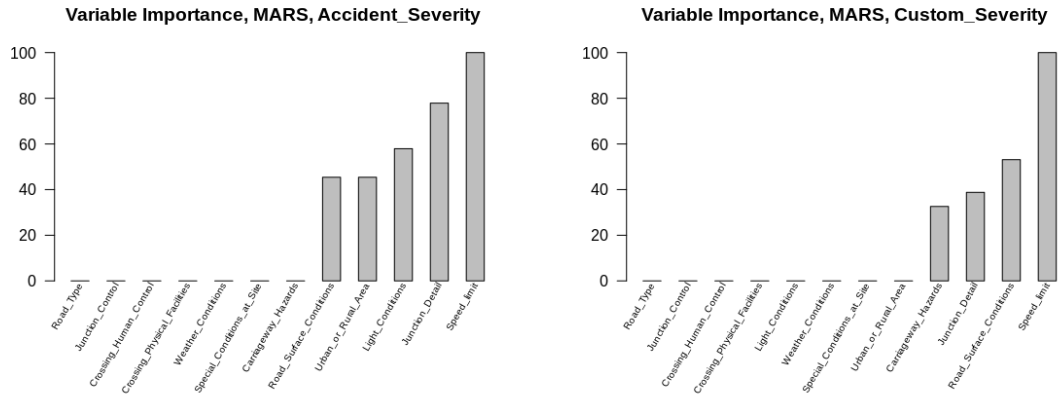
By taking a look at the actual data, it is clear that those metrics are misleading. Since there are only three categories in Accident_Severity, and around 85% of the data is in one category. The linear regression model always predicts results in category 3 and is close to the truth in most cases. The introduction of a wider spread severity, Custom Severity, changed that. The major part of the data was still grouped into one category, therefore the model only predicts that category. The RMSE and MAE metrics got worse because our prediction parameter was wider spread, and therefore the prediction error was bigger than before.



(a) Distribution of Accident_Severity. (b) Distribution of Custom_Severity.

The chosen degree of interaction is 6, this is to make sure the algorithm detects interaction between parameters. Another way of running a mars regression is with the train() function of the crane package. This function gives the option to include further analysing practices. For validation purposes another mars regression, this time with cross validation, was performed on the data. The mars regressions on Accident_Severity and Custom_Severity return the same results as the linear regressions. It is safe to say that regression is not the right method for this dataset. This is mainly due to the low number of accident severities and the fact that most of the data is grouped into one category. Although the introduction of a new severity parameter changed that slightly, it is still too small of a number of possible outcomes and an uneven distribution of accidents over the possible severity values.

The results regarding the variable importance produced interesting results. Using the varImp function [5] of the caret package [1], all of the previous regression models were used to give the parameters of the model an importance value.The parameter importance for each regression per severity are very similar. There are even big similarities between the importance of the two severities.



(a) Distribution of Accident_Severity. (b) Distribution of Custom_Severity.

Figure 13: Variable importance, MARS.

The Speed Limit, in both cases, has the biggest influence on the severity. That implies that there must be some kind of relationship mainly between slight accidents and the speed limit on the scene of the accident. Similar results can be seen for Junction Detail and Road Surface Conditions. Although some of the parameters are getting less important, the introduction of a new parameter did not change those relationships entirely.

In summary, it is quite clear that the linear regression and how the dataset was utilised, was not the most optimal way. It was only in the interest of time that we did not run any further models or other techniques that came to mind as we were processing the dataset. For a more fleshed out discussion on what could have been done, see section 5.

4.2 Hypothesis 2: Number of Accidents Time Analysis

As part of the model assessment for the deep learning model, table 3 showed all the model runs. The best model out of those runs was the sixth run, with an RMSE of 40.55, MAE of 31.34 and R^2 of 66.95%. Figure 14 shows the RMSE plotted against the number of epochs of the deep learning model for the training and validation stage. RMSE decreases as the number of epochs increases for both validation and test and thus the optimal number of epochs was selected as 150.

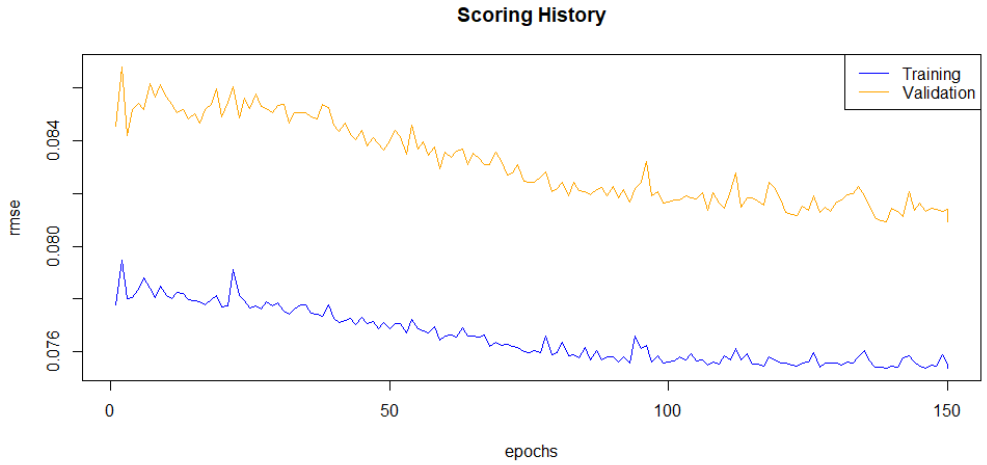


Figure 14: Number of Epochs vs RMSE for Training Validation.

Figure 15 illustrates the time series plot of the expected and predicted number of accidents for model run 6. The red lines indicate the predicted number of accidents from January 2013 to December 2015. The lack of accuracy of the predictions is apparent when comparing the peaks and troughs of the expected number of accidents. The model has failed to learn this pattern.

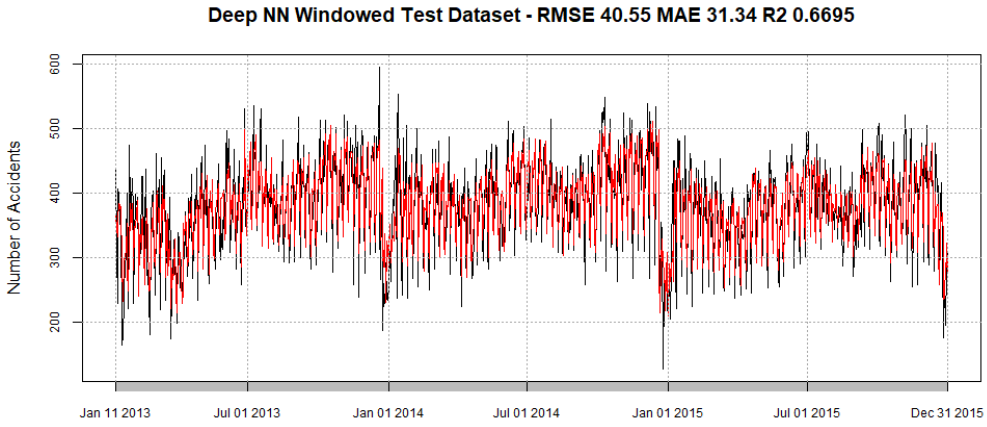


Figure 15: Expected and Predicted Values for Test Dataset (2013-2015).

For the RNN, figure 16 depicts the comparison between the expected and predicted (red line) number of accidents. The model performs poorly in comparison to the deep learning model – the RMSE, MAE of the deep learning model and RNN are very similar however, the R^2 for the RNN is 63% compared to 67% for the deep learning neural network.

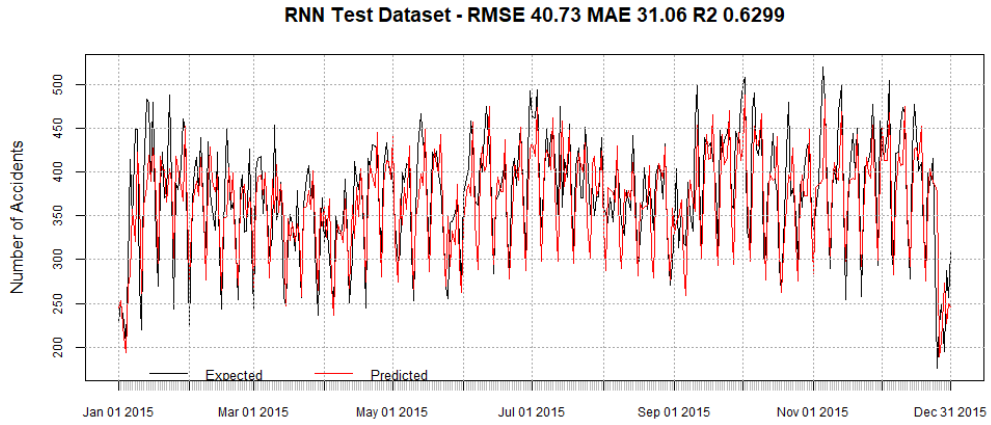


Figure 16: Expected and Predicted Values for Test Dataset (2015).

In summary, two models were created to predict the number of accidents in a day – deep learning and a RNN. Although RNNs have the advantage of retaining information about the previous input, for this dataset, the deep learning model performed better with the value of R^2 being much larger. Reasons for this include: the lack of ability for the RNN to learn long-term patterns, a smaller number of days trained on the RNN and no evaluation of the performance of the training data was taken leading to either under-fitting or over-fitting of the data.

4.3 Hypothesis 3: Age of the driver and Accident Characteristics

As part of the model assessment for dataset 1 (London only) the optimal number of k for k-means clustering was identified as 6 (figure 10). Further analysis of this cluster plot is illustrated in table 8.

Cluster	# of Records
1	42,167
2	16,681
3	99,830
4	55,193
5	30,755
6	20,959

Table 8: Number of Records by Cluster for $k = 6$

Cluster 3 had the largest number of records, whereas cluster 6 had the lowest number

of records. An example of the cluster analysis was identifying what accident features were common for drivers aged 25 and under. Figure 17 illustrates cluster 1's age distribution as it had the highest proportion of the number of accidents for drivers aged 25 and under.

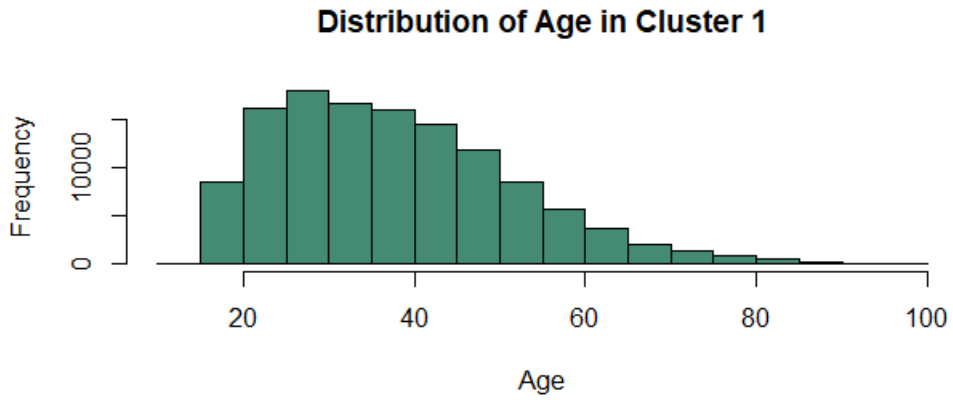


Figure 17: Distribution of Age in Cluster 1.

The next step was to find accident characteristics for drivers aged 25 and under to identify what types of accidents occur the most within this group. Examples include: the road type, weather conditions and light conditions for drivers under the age of 25. Figure 18 uses the example of light conditions to assess which categories of light resulted in a higher number of accidents. The histogram tells us that most accidents occur in categories 1 and 4, which translate to daylight and darkness-lit respectively. Therefore, given the high proportion of accidents occurring in darkness, it is possible to infer that younger drivers are prone to accidents occurring at night in London. However, further analysis would still be required to confirm this, which is outside the scope of this report.

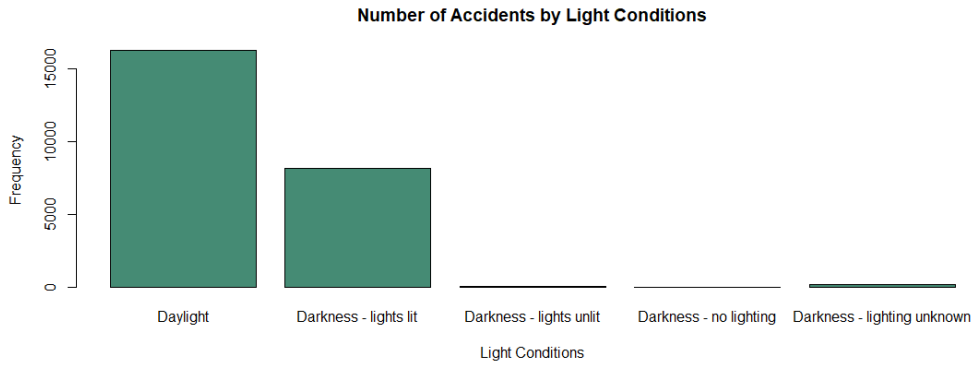


Figure 18: Number of Accidents by Light Conditions for Drivers Aged 25 and Under within Cluster 1.

As part of the model assessment for dataset 2 (London with accident severity fatal or serious) the optimal number of k for k-means clustering was identified as 3 (figure 11). Further analysis of this cluster plot is illustrated in table 9.

Cluster	# of Records
1	13,613
2	4,247
3	10,389

Table 9: Number of Records by Cluster for $k = 3$

Cluster 1 had the largest number of records, whereas cluster 3 had the lowest number of records. Figure 19 illustrates cluster 1's age distribution as it had the highest number of accidents for drivers aged 25 and under.

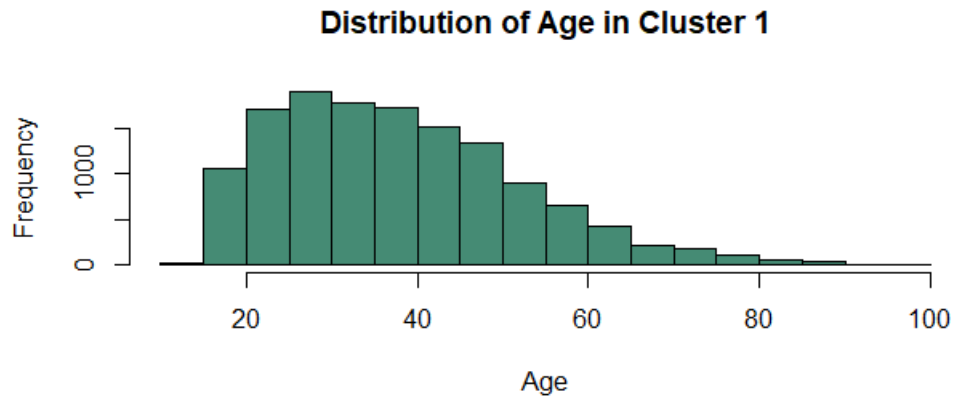


Figure 19: Distribution of Age in Cluster 1.

The next step was to find accident characteristics for drivers aged 25 and under to identify what types of accidents occur the most within this group. Continuing with the example used for dataset 1, the distribution of accidents by light conditions is shown in figure 20. Although the histogram tells us that most accidents occurred in category 1 (daylight), there is again a repeated pattern that a significant amount of accidents occurred in category 4 (darkness lit). Therefore, it is again possible to assume that younger drivers may be more prone to fatal or serious accidents occurring at night in London. Again, further analysis would still be required to confirm this which is outside the scope of this report.

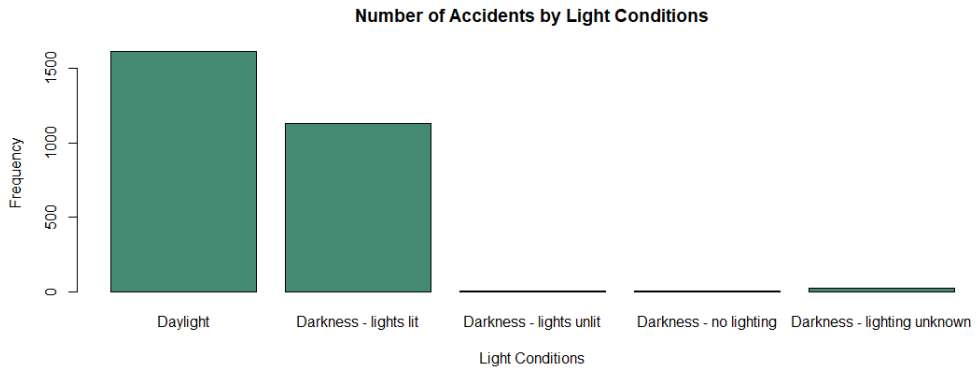


Figure 20: Number of Accidents by Light Conditions for Drivers Aged 25 and Under within Cluster 1.

In summary, the k-mean clustering algorithm was used to find a pattern in the features of a road accident in London. The example used was for drivers aged 25 or younger and

the accident feature used was Light Conditions. Furthermore, it could be possible to find ways to increase road safety by aiming to reduce accidents occurring in darkness. Perhaps more road lighting or additional training for driving at night for younger drivers can be example proposals to the relevant authorities.

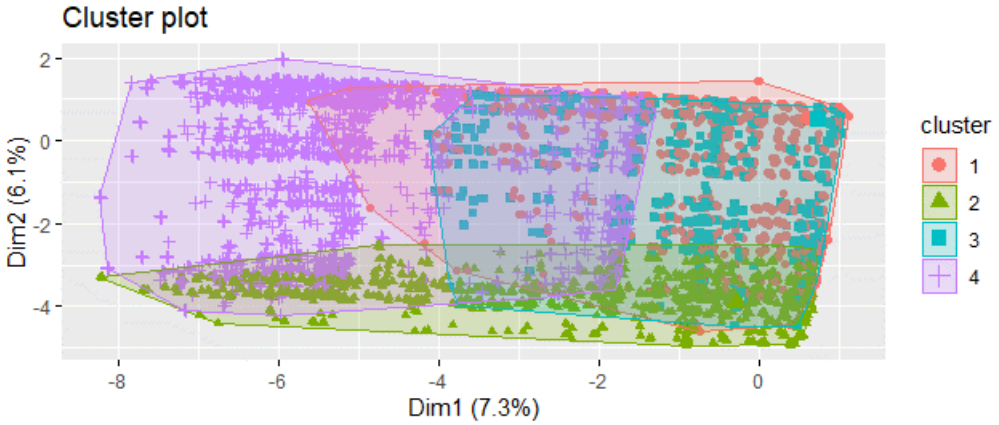
5 Conclusion and Further Work

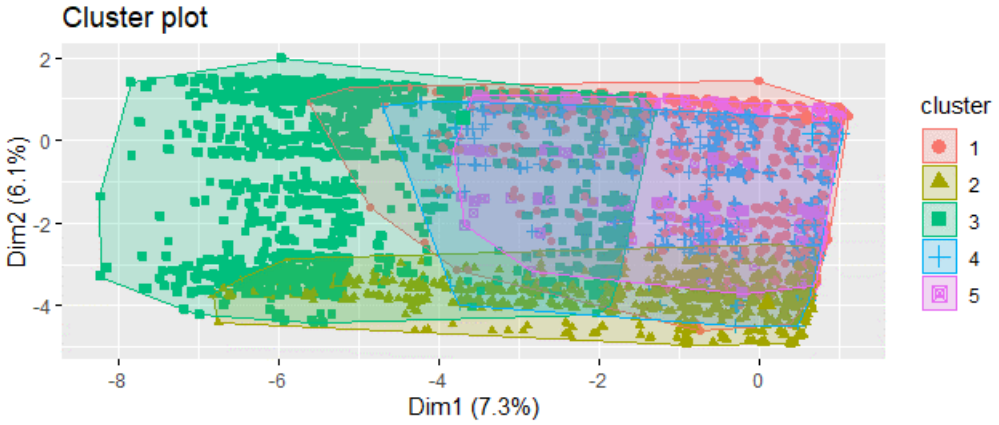
In conclusion, this analysis provided us with useful insights on different ways showing how the UK government can use its collective data to improve road safety. The first hypothesis, even though unsuccessful in its function to predict the severity of any given accident, it provided us with crucial information on the most important factors of an accident - the number of vehicles through the use of k-means clustering and speed limit via the use of variable importance method. With respect to hypothesis 2, a better functioning model could be deployed to communicate the number of anticipated accidents to local authorities. With this information, planning and budgeting can be adequately resourced and can result in improved road safety across the UK and more importantly reducing the number of road accident victims. For hypothesis 3, clustering of road accident features amongst different age groups could be communicated to the Metropolitan Police and the City of London.

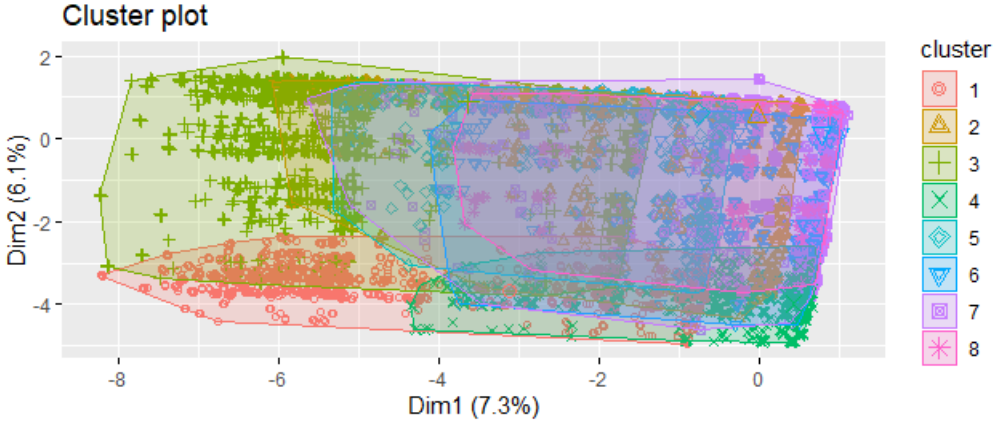
Further work that would be done on all three hypotheses involves ways on how to improve model performance. More specifically with hypothesis 1, a Self-Organizing maps technique could have been employed to better visualise the relationships of the combined subset of factors in interest. A DatabionicSwarm method [6] could have been used to better utilise any interactions between different factors and create a better prediction for Accident Severity per accident. For hypothesis 2 both models failed to learn the pattern of the data. One approach to improve this would be the re-training of the data every year by calculating the average between the expected and predicted values. In this way, the number of predicted accidents would converge towards the expected number of accidents. With hypothesis 3, only drivers aged 25 and under were analysed to find ways to improve road safety. Also, the method of k-means clustering is very random and thus other clustering algorithms should be utilised and compared. Lastly, the models used in this report were generalised for the UK - the next step would be to have a variety of models that are location specific.

6 Appendix

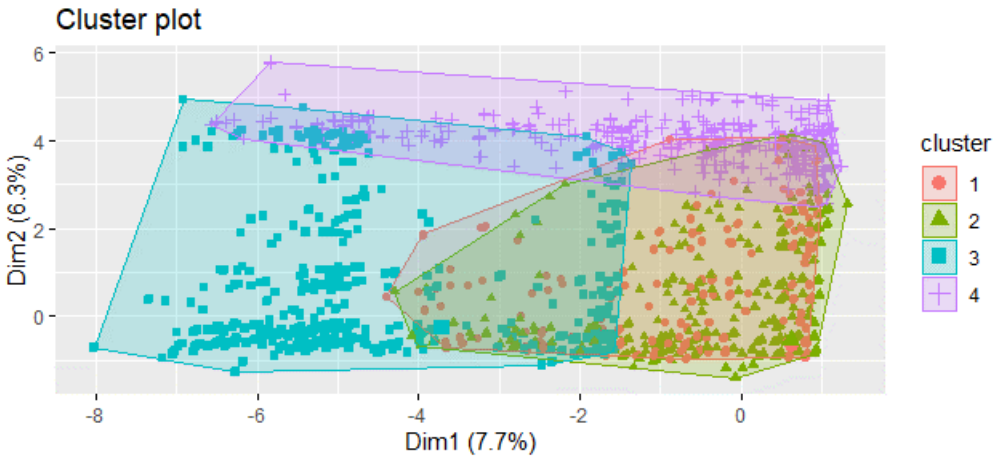
6.1 Appendix 1: Clusters for different Values of k (Dataset 1)

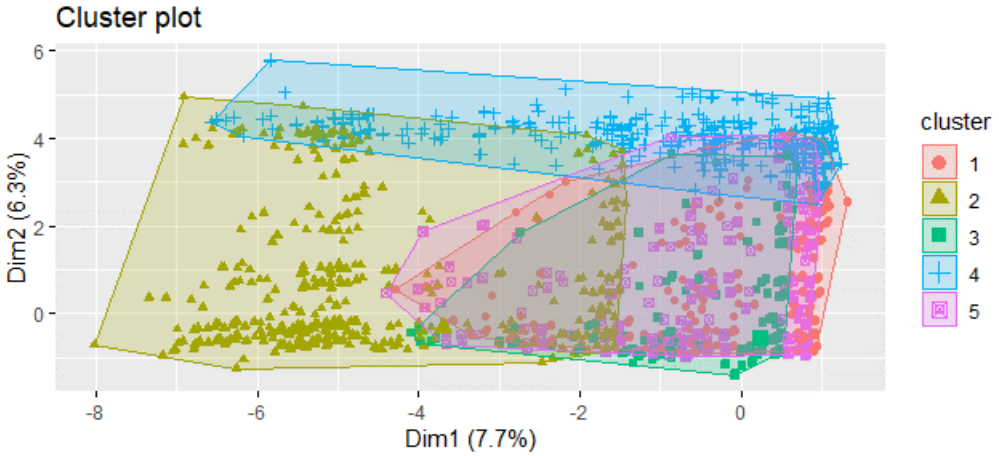


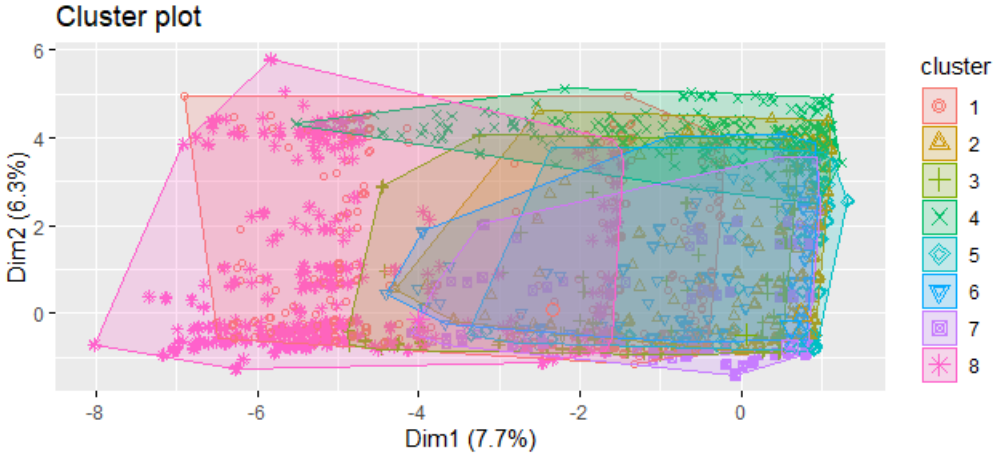




6.2 Appendix 2: Clusters for different Values of k (Dataset 2)







References

[1] The caret package. "http://topepo.github.io/caret/index.html".

[2] Create elegant data visualisations using the grammar of graphics. "https://cran.r-project.org/web/packages/ggplot2/index.html".

[3] Package 'earth'. "https://cran.r-project.org/web/packages/earth/earth.pdf".

[4] R interface for 'h2o'. "https://cran.r-project.org/web/packages/h2o/index.html".

[5] Rf variable importance for arbitrary measures. "https://cran.r-project.org/web/packages/varImp/index.html".

[6] Swarm intelligence for self-organized clustering. "https://cran.r-project.org/web/packages/DatabionicSwarm/DatabionicSwarm.pdf".

[7] Driver Vehicle Standards Agency. History of road safety, the highway code and the driving test. "https://www.gov.uk/government/publications/history-of-road-safety-and-the-driving-test/history-of-road-safety-the-highway-code-and-the-driving-testa".

[8] Driver Vehicle Standards Agency. History of road safety, the highway code and the driving test. "https://www.gov.uk/government/publications/history-of-road-safety-and-the-driving-test/history-of-road-safety-the-highway-code-and-the-driving-test", 2019.

[9] Driver Vehicle Standards Agency. The road safety statement 2019 a lifetime of road safety. "https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/817695/road-safety-statement-2019.pdf", 2019.

[10] Department for Transport. Road safety data. "data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data", 2016.

[11] Prakhar Ganesh. Time series analysis with deep learning : Simplified. "https://towardsdatascience.com/time-series-analysis-with-deep-learning-simplified-5c444315d773", 2019.

[12] Nick Ryman-Tubb. Laboratory notes. "https://surreylearn.surrey.ac.uk/d2l/le/content/188750/Home", 2019.