**BIG DATA AND DATA MINING REPORT**

#### **The Task.**

Imagine that you are a data scientist confronted with this data (this is not far from the truth!). Your task  
is to advise government agencies about how to improve road safety and create a model that would  
predict such accidents and the injuries that they incur.

1. Are there significant hours of the day, and days of the week, on which accidents occur?
2. For motorbikes, are there significant hours of the day, and days of the week, on which accidents occur? We suggest a focus on: Motorcycle 125cc and under, Motorcycle over 125cc and up to 500cc, and Motorcycle over 500cc.
3. For pedestrians involved in accidents, are there significant hours of the day, and days of the week, on which they are more likely to be involved?
4. Using the apriori algorithm, explore the impact of selected variables on accident severity.
5. Identify accidents in our region: Kingston upon Hull, Humberside, and the East Riding of Yorkshire etc. You can do this by filtering on LSOA, or police region or another method if you can find one. Run clustering on this data. What do these clusters reveal about the distribution of accidents in our region?
6. Using outlier detection methods, identify unusual entries in your data set. Should you keep these entries in your data?
7. Can you develop a classification model using the provided data that accurately predicts fatal injuries sustained in road traffic accidents, with the aim of informing and improving road safety measures?

# 1.0 INTRODUCTION

In the UK, approximately 3000 people die in road traffic incidents each year, according to statistics released by (Clarke et al., 2010). There are several fatalities and injuries from traffic accidents. Three different factor types, namely, often cause traffic accidents, human factors, vehicle factors, and external factors (Chen & Jou, 2019; Mohammed et al., 2019). We use analytics and data science to mine road traffic accident data for insights to address this urgent problem. In this research, we examine an extensive dataset that contains 90 columns and 201943 rows and includes facts about vehicles, casualties, accidents, and location-specific information. Utilizing data-driven methods, we hope to find trends, risk factors, and crucial intervals associated with motorbike incidents, pedestrian participation, and accidents in the region of Humberside. Our objective is to offer practical suggestions to government authorities to enhance road safety protocols and reduce fatalities by comprehending the contributing variables and the intensity of accidents.

# ANALYSIS

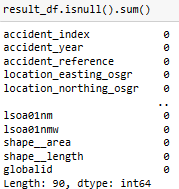
This section will include a synopsis of each analysis as well as potential illustrations and justifications:

# Data Cleaning

To handle missing values and data inconsistencies, I had to ensure data cleaning were conducted, as it is an essential step in data preprocessing.

# Checking for NAN & NULL values

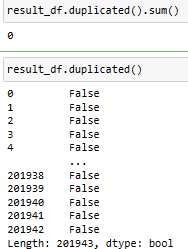
NAN and null values were checked to confirm that there were no missing or inappropriate values, and it was established, as shown in Fig-1, there were none.



**Fig-1.** Dataframe without null values

# Checking for duplicated values

Duplicated values were checked in the data frame to ascertain that we don’t have duplicated values in our dataset, which we checked and reconfirmed that there were none, as shown in Fig 2 below.



**Fig-2.** Dataframe with no duplicated values

# Checking of several columns

As part of my cleaning process, some columns such as Speed Limit, Weather condition, Light conditions, road\_surface\_conditions, age\_of\_driver age\_of\_casualty, junction\_control, sex\_of\_driver which have -1 outliners (not in our stats19 accident statistics documentation checklist) were replaced with the median. I decided to use the median because it is a reliable indicator of central tendency, it will reflect the overall data distribution and is less susceptible to outliers when compared to using mean, mode etc.

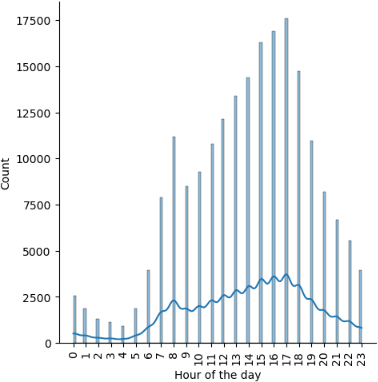
# Checking for data type

It was discovered that the time column must be transformed to date time format(object) to ease date-time interpretation and visualization. The pd.to\_datetime method was used in this conversion.

# Significant Hours of the Day and Days of the Week for Accidents

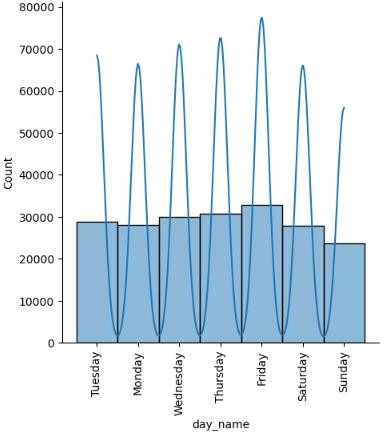
I examined the accident distribution according to the day of the week and the hour of the day after I joined the four tables in the accident database. I went further to make visualizations to help identify the peak hours and days when an accident occurs most likely and understand the trend. The analysis begins with:

* necessary libraries importation
* Data preprocessing ('Time' column is in the datetime format and Extract the 'hour' and 'day\_of\_week' and stored in the result data frame).
* Visualization of accident by the hour of the day and day of the week



**Fig-3.** Accident Distribution by Hour of the Day

The plot shows that accidents peak during specific 17:00 hours of the day, as shown above. This suggests that accidents are more likely to happen during busy commuting hours in the evening, as shown in the plot, likely due to increased traffic associated with people returning from work and school.

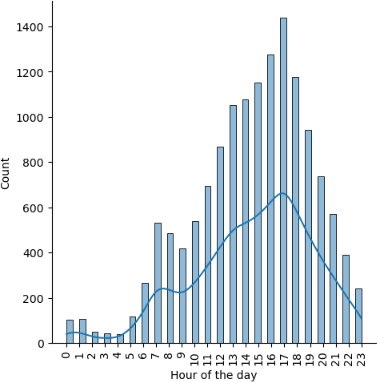


**Fig. 4.** Accident distribution by Day of the Week

Fig-4 Plot demonstrates how the frequency of accidents varies on different days. Friday is the day with the most accidents; we have more weekly accidents (Monday-Friday) than on the weekends (Saturday-Sunday). This suggests there are more weekday accidents, maybe because of increased traffic and the number of individuals travelling to and from work or school.

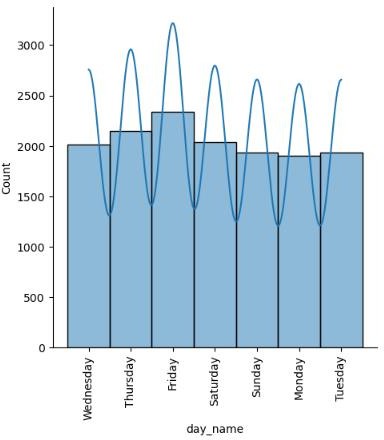
# For motorbikes, are there significant hours of the day, and days of the week, on which accidents occur?

I created a motorcycle dataframe and filtered through the following**-**Motorcycles 125cc and under, Motorcycles over 125cc and up to 500cc, and Motorcycles over 500cc, resulting in 14,311 rows and 83 columns. I went further to make visualizations to identify the peak hours and day when these motorbike types, Motorcycle 125cc and under, Motorcycle over 125cc and up to 500cc, and Motorcycle over 500cc accidents, occurs most likely and understand the trend.



**Fig-5.** Distribution of selected motorbikes Accidents by Hour of the Day

Fig-5 was used to show the distribution of motorbikes accidents by hour of the day. Accidents peak during a particular hour, 17:00. these selected motorbikes accidents are more likely to occur during the evening rush hour, as many working class and students use the road then.



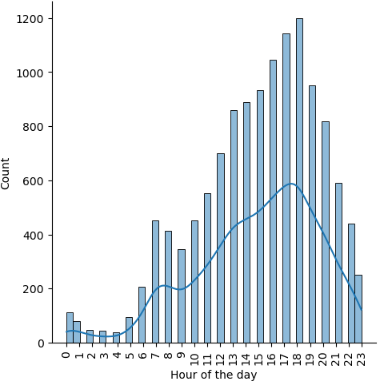
**Fig-6** Distribution of selected motorbikes Accidents by day of the week

Fig-6 illustrates the variation in accident frequency across different days. Friday is the day with the highest number of accidents, as the plot above shows. More motorbikes accidents occur on Friday because many

people use the road that day, ranging from returning from work and school to travelling for the weekend to see our loved ones.

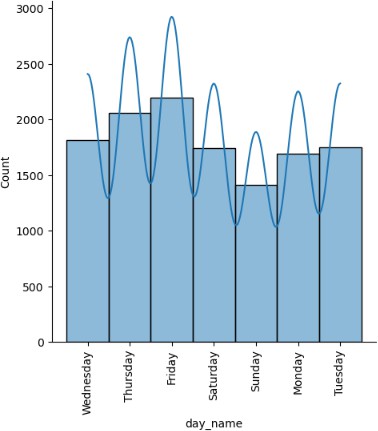
# For pedestrians involved in accidents, are there significant hours of the day, and days of the week, on which they are more likely to be involved?

I filtered through where the casualty type is equal to 3, which had 12,662 rows and 83 columns with variable name called pedestrian\_df. I created visualizations to determine the busiest times of the hour and day of the week when accident associated with pedestrians occurs most likely.



**Fig-7.** Distribution of pedestrian accidents by Hour of the Day

Fig-7 plot was used to show accidents distribution by hour of the day; accidents tend to peak during a particular hour of the day, 18:00. This indicates that pedestrian accidents are more likely to occur during the evening rush hour because of heavier traffic from a lot of working-class and students coming back from work.



**Fig-8.** Distribution of pedestrian accidents by day of the week

Fig-8 shows how events are distributed according to the day of the week. Friday was the most significant number of pedestrians involved in accidents. This suggests that many working-class and-students travel on Friday evenings for the weekend to spend time with their loved ones.

# Using the apriori algorithm, explore the impact of selected variables on accident severity.

To explore the impact of selected variables (["accident\_severity","speed\_limit", "weather\_conditions" "light\_conditions","road\_surface\_conditions”, "urban\_or\_rural\_area", "junction\_detail", "age\_of\_driver", "age\_of\_casualty) on accident severity using the apriori algorithm is to analyze the associations of selected variables and accident-severity level.

Antecedents: These are items that we are observing to find patterns/associations. For instance, the first row "(accident\_severity\_3)" denotes the level 3 accident severity condition.

Consequents: These items represent the outcomes one is interested in based on the presence of specific antecedents. We have a speed limit of 30mph.

Antecedent support: The percentage of occurrences in the dataset that satisfy the antecedent condition(s). For instance, "0.782676" means that around 78.27% of accidents are classified as severity-3

Consequent support: This is the percentage of instances in the dataset that satisfy the corresponding consequent condition(s). For instance, the first-row value "0.543762" means that 54.38% of the time, a speed restriction of 30 mph is observed.

Support: This is the percentage of events that satisfy both the antecedent and the consequent requirements. For instance, "0.441877" implies that a speed restriction of 30 mph and an accident severity level 3 are observed in 44.19% of situations.

Confidence: Given the antecedent condition(s), this is the conditional likelihood of witnessing the consequent condition(s). For instance, "0.564572" implies that there is a 56.46% chance of coming across 30-mph speed limit when accident severity-3 is recorded.

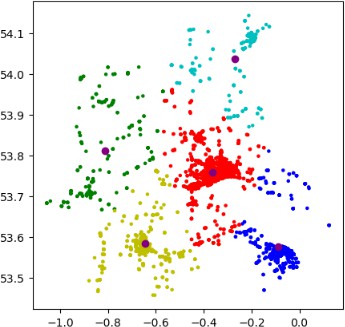
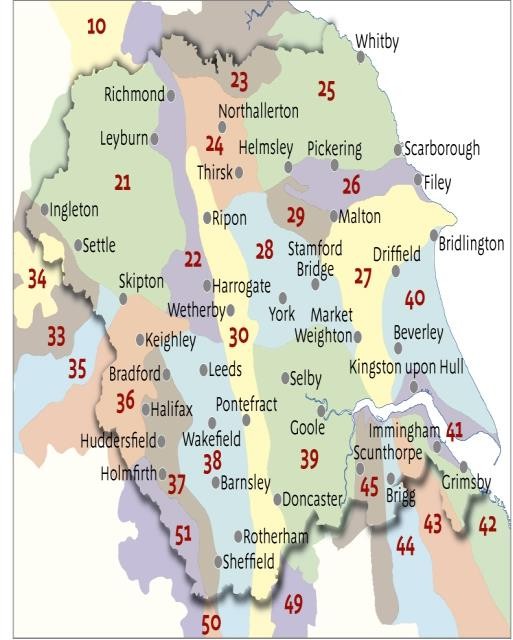
Lift: This a metric that expresses how strongly an antecedent and consequent are associated. Positive associations are indicated by values higher than 1. For instance, "1.038270" indicates a marginally positive association between a 30-mph speed limit and accident severity-3.

Leverage: The difference between the observed frequency of the antecedent and consequent combined and what would be anticipated if they were independent is quantified by leverage. The link between the accident severity-3 and the 30-mph speed restriction is mildly positive, according to the leverage value of 0.016287.

Convictions: Conviction evaluates how dependent the consequent is on the antecedent. A conviction value of 1.047791 indicates a little positive relationship between the occurrence of a 30-mph speed restriction and an accident severity-3 (antecedent).

# 2.6. Identify accidents in our region. Run clustering on this data. What do these clusters reveal about the distribution of accidents in our area?

I selected the police force from the database equal to 16 to filter the Humberside region. I applied KMeans to group accidents based on location (longitude, latitude) because k-means clustering is one of the easiest clustering methods that use simple geometric distance calculations and has good scalability, which has the capacity of clustering large datasets at reduced computational cost (Sinclair & Das, 2021). I went further to Visualize the clusters on a map to understand the spatial distribution of accidents, as shown in Fig-9.

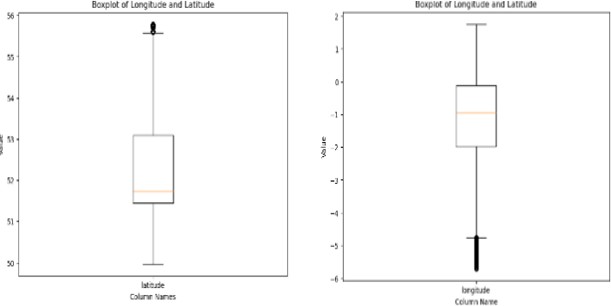
 

**Fig-9.** Accident density map. **Fig-10.** Humberside map.

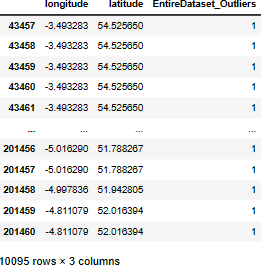
Having compared the plot made with the Humberside map, we can see that the red dense centroids depict Kingston upon hull, the blue clustered centroids show Grimsby, the yellow centroid corresponds to Sculthorpe, the northeast sky blue corresponds to Bridgton cities. The above Fig-9 map shows that most accidents happened in the densely populated urban area in the cities such as Kingston upon Hull, Grimsby, Sculthorpe, and Bridlington.

# 2.7. Identify unusual entries in your data set using outlier detection methods. Should you keep these entries in your data?

I implemented Isolation Forest to identify unusual entries in the dataset because it is a powerful tool for solving the anomaly detection problem (How to Use the Isolation Forest Model for Outlier Detection, 2021). It was observed as shown in fig-12 that there were 10095 outliners, which is approximately 5% of the dataset. It is ideal we keep such outliners since they are recorded accidents with locations.



**Fig-11.** Boxplot of longitude and latitude column showing outliers.



**Fig-12.** Data frame showing number of outliers in the entire dataset.

# PREDICTION

In this section, I developed a classification model to accurately predict fatal injuries sustained in road traffic accidents. The steps I took include:

# Libraries-importation

Needed libraries imported for classification.

# Data Preparation

I declared my target variable(y\_label), the accident severity, when equal to 1. Then said, the independent variable(x), which is df\_acc\_balanced. I further Split the dataset into training and testing sets to train and evaluate the predictive model.80% of the data was used for training, while 20% was used for testing.

# Models Training/Building

Three different machine learning algorithms, namely decision tree classifier, random forest classifier and bagging classifier, were used to train and build the model to predict fatal injuries sustained in road traffic accidents.

# Models Evaluation

The three classification models used performance evaluation metrics like accuracy, precision, recall, and F1-score. Among the three models used in our classification, the random forest classifier performed the best with an accuracy of 89%, as shown in Fig-10. Having completed the best, the random forest classifier model, as shown in Table1, is recommended to accurately predict fatal injuries in road traffic accidents.

**Table1.** Performance metrics of our classification models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Performance metrics | Accuracy | Recall | F1score | precision |
| Random Forest  model | 0.89 | 0.89 | 0.89 | 0.89 |
| Bagging model | 0.85 | 0.85 | 0.85 | 0.85 |
| Decision Tree model | 0.76 | 0.72 | 0.72 | 0.72 |

# 4.0. RECOMMENDATION

Considering the overall analysis and modelling, I recommend the following to government agencies to improve road safety.

* Conduct focused safety programmes to increase awareness and promote safer driving habits throughout the peak accident hours and days in urban areas.
* Enhance pedestrian visibility and crossings to reduce accidents involving pedestrians.
* Put traffic calming measures(signages) in place and speed limits in locations where accidents and motorbike accidents are common, like the urban cities which are most affected.
* Identify and address factors that significantly impact accident severity, such as road, light, and weather conditions, to optimize road safety.
* Strengthen pedestrians, vehicle users and Motorcycle Rider Education and Training to facilitate their awareness of accident-prone hours and safe riding/driving practices.

In conclusion, government agencies can benefit greatly from analyzing this dataset and predictive model to formulate more effective road safety policies to reduce accident fatalities.

**5.0. REFERENCES**

Chen, T. Y., & Jou, R. C. (2019). Using HLM to investigate the relationship between traffic accident risk of private vehicles and public transportation. *Transportation Research Part A: Policy and Practice*, *119*, 148–161. https://doi.org/10.1016/J.TRA.2018.11.005.

Clarke, D. D., Ward, P., Bartle, C., & Truman, W. (2010). Killer crashes: Fatal road traffic accidents in the UK. *Accident Analysis & Prevention*, *42*(2), 764–770. https://doi.org/10.1016/J.AAP.2009.11.008.

*How to use the Isolation Forest model for outlier detection*. (n.d.). Retrieved August 14, 2023, from https://practicaldatascience.co.uk/machine-learning/how-to-use-the-isolation-forest-model-for-

outlier-detection.

Mohammed, A. A., Ambak, K., Mosa, A. M., & Syamsunur, D. (2019). A Review of the Traffic Accidents and Related Practices Worldwide. *The Open Transportation Journal*, *13*(1), 65–83. https://doi.org/10.2174/1874447801913010065.

Sinclair, C., & Das, S. (2021). *ORE Open Research Exeter TITLE Traffic accidents analytics in UK urban areas using k-means clustering for geospatial mapping A NOTE ON VERSIONS Traffic Accidents Analytics in UK Urban Areas using k-means Clustering for Geospatial Mapping*. [http://hdl.handle.net/10871/125145.](http://hdl.handle.net/10871/125145)