A VISUAL EXPLORATION OF UK ROAD TRAFFIC ACCIDENTS IN 2014

Mohd Arifullah

ABSTRACT: This study conducts a detailed analysis of road traffic accident hotspots across the United Kingdom, harnessing multivariate data to uncover the patterns and regularity of such events. Geared towards aiding highway operators and Public Work Department officials, it utilizes data visualization techniques, primarily through Tableau and Python, to transform complex data sets into intelligible visual stories. These visualizations assist in recognizing trends and pinpointing areas of high accident frequency, thereby informing and refining traffic management, and road safety strategies. The insights derived are anticipated to significantly influence the creation of preventative measures, streamline road safety practices, and ultimately enhance the public's welfare and the efficacy of road network management.

PROBLEM STATEMENT

In the realm of road safety and traffic management, the analysis of accident data plays a crucial role in understanding the dynamics and causes of road traffic accidents (RTAs).

In this research we will try to answer the following research questions:

1. How do different road surface conditions, such as dry, wet, or icy conditions, correlate with the frequency and severity of traffic accidents across the United Kingdom?

2. Is there a discernible pattern in the occurrence of road traffic accidents when comparing various light conditions, such as daylight with streetlights present and darkness with no street lighting?

3. Does the type of road (e.g., single carriageway, dual carriageway, roundabout) have a significant impact on the number of casualties and severity of accidents reported?

This study aims to identify accident hotspots, discern temporal trends, and correlate accident severity with various environmental and road factors. Furthermore, authors of [1] suggest that there could be many reasons for traffic congestion such as limited road capacity, time of the day, number of accidents, increase of vehicles at certain time of the day and certain parts of the road, road works that narrow the lanes, and affects traffic flow, as well as bad weather that could result in partial or full traffic congestion.

STATE OF THE ART

The study of road traffic accidents through data analysis has become increasingly sophisticated, as evidenced by several key studies in this field. A notable approach to understanding traffic accidents was demonstrated by Richard and Ray [2], who examined vehicle accidents in Fredericton and Laval, Canada. Their methodology combined historical vehicle accident records with other contextual data such as weather conditions, road features, and geospatial information. Notably, they utilized bar charts and heat maps for temporal and spatial analysis, respectively, and employed the Getis-Ord Gi statistic for identifying significant hotspots.

In my research on UK road traffic accidents, I will leverage a larger dataset and use stacked bar graph, multiline chart, dot distribution map for visualization, which are instrumental in identifying clusters of accidents. I will be using Python (Jupiter) and K-Means Clustering for hotspot identification. Additionally, the research will encompass time analysis to understand the temporal and spatial distribution of accidents.

The predictive aspect of accident analysis was also explored by Richard and Ray [2], who used a random forest model to predict accidents leading to injuries or fatalities. They identified key variables influencing accident severity in each city they studied. In contrast, due to the limitation of features in my dataset, my focus will be on the type of road user affected, light conditions and type of area in the accidents.

Looking at other relevant studies, Gao [3] analyzed traffic accident data from Beijing, using similar visualization techniques (bar charts and heatmaps) to examine the time distribution and density of accidents. They applied Kernel Density Estimation for cluster identification, providing a valuable methodological reference for my work.

Another significant contribution came from Zhou [4], Mao, and Li, who analyzed New York City data to identify the safest streets. They employed a unique approach by matching collision data with the nearest street and analyzing collision frequency by hour. Their use of K-Medoids clustering on collision curves offers an interesting perspective for spatial analysis.

In my analysis of UK road traffic accidents, I aim to synthesize these various methodologies to develop a comprehensive understanding of accident patterns. The use of heat maps, line charts and other different graphs will be instrumental in visualizing spatial clusters, while time analysis will provide insights into the temporal trends of accidents. The research will contribute to a more nuanced understanding of the factors contributing to road accidents in the UK, potentially aiding in the development of targeted safety measures and policy interventions.

This multidimensional approach, combining spatial and temporal analysis with advanced data visualization techniques, represents the state of the art in road traffic accident analysis. It underscores the importance of leveraging diverse datasets and analytical tools to gain a holistic understanding of road safety issues. The insights derived from this study are expected to be valuable for policymakers, urban planners, and safety advocates in designing effective strategies to enhance road safety in the UK.

PROPERTIES OF DATA

The dataset is from Kaggle which consists of around 146,322 records, offers a detailed overview of road traffic incidents for the year 2014 in United Kingdom, with 33 columns, covering a wide geographical scope as indicated by the range in coordinates (Longitude, Latitude) and Ordnance Survey Grid References (Location_Easting_OSGR,Location_Northing_OSGR) . It includes both categorical and numerical data, such as 'Police_Force' and 'Local_Authority_(District)', likely representing various police jurisdictions and districts. Temporally, the data spans the entire year, with the 'Date' and 'Day_of_Week' columns suggesting potential analyses of incident frequencies by time.

Significantly, the dataset provides extensive details about the incidents, roads, and traffic. This includes road classifications ('1st_Road_Class', '2nd_Road_Class'), road identifiers ('1st_Road_Number', '2nd_Road_Number'), and speed limits, which range from 20 to 70 mph. However, the 'Junction_Detail' column is entirely missing, indicating a gap in the data. The number of vehicles involved in incidents (ranging from 1 to 21) and the number of casualties (ranging from 1 to 93) offer insights into the severity of these incidents. The distinction between urban and rural settings is made through the 'Urban_or_Rural_Area' column. However, the datasets contain some nan values and even outliers with no duplicated rows. From the Fig 1 we can see the correlation heatmap of the features.

Statistically, the dataset provides a wealth of information. The mean values, such as an average of 1.84 vehicles involved per incident, offer a general overview. The range of data is captured through the min/max values, and the 25th, 50th (median), and 75th percentiles illustrate the distribution. The standard deviation in columns like 'Number_of_Vehicles' suggests consistency in some aspects of the data.

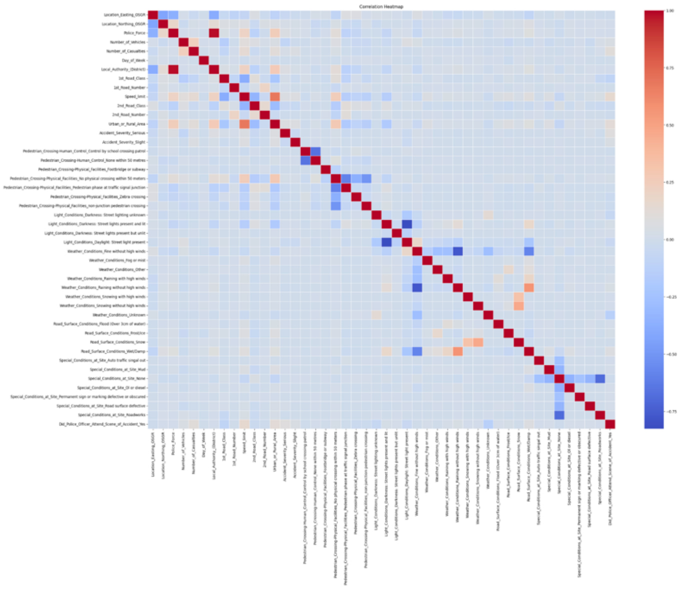

Fig 1 Correlation Heatmap

However, there are indications of data quality issues, such as the complete absence of 'Junction_Detail' which we will drop later and potential outliers or errors in road number columns. Despite these limitations, the dataset is a good enough for analysing traffic safety and patterns, providing a comprehensive view of road incidents over a year with detailed geographical, temporal, and incident-specific information.

ANALYSIS

I. APPROACH

The Fig 2 below shows the steps used in solving the analytical problems of Traffic Accidents by using the different purpose and methods applied in this report. Moving on further we will discuss details of each step.

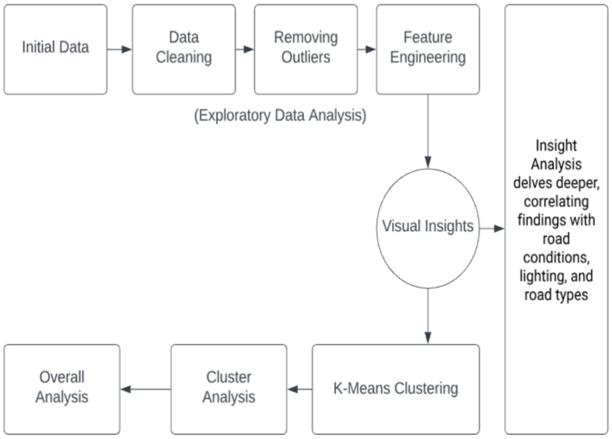
Initial Data
Data Cleaning
Removing Outliers
Feature Engineering
(Exploratory Data Analysis)
Visual Insights
Insight Analysis delves deeper, correlating findings with road conditions, lighting, and road types
Overall Analysis
Cluster Analysis
K-Means Clustering

Fig 2 Flow Chart for the Approach

In studying traffic accident dataset, an exacted method is used starting from a preliminary investigation that reveals the basics about this dataset. This phase entails the use of different types of graphical tools, including histograms for distribution analysis; density map for geographical concentration and scatter plots to view relationships.

Once our dataset is ready, we analyzed data observed some outliers, so to remove them we will using IQR, or Interquartile Range to remove the outliers. For removing outliers, we have the IQR method for removing the outliers. A quartile, in turn, is any of the three values that divides a sorted data set into four parts of equal size. The first quartile Q1 marks the end of the first 25% of data, the third quartile Q3 the beginning of the last 25%. Every value being smaller than Q1–1.5 IQR or larger than Q3+ 1.5 IQR is regarded as outlier [5].

As there were only few nan values and no duplicate values, we filled the categorical data such as Road_Type, 'Urban_or_Rural_Area' with the mode of the column and for the numerical columns we filled it with mean values. The feature engineering phase builds upon the cleaned dataset by removing unwanted columns and introducing calculated columns such as 'Month' extracted from the date entries, and an 'Accident Severity' with Month classification. These newly minted attributes facilitate a more granular examination of the data, particularly within the context of temporal patterns and spatial distributions.

a) Spatial Analysis:

The spatial component examines the physical locations of traffic incidents, revealing clusters and hotspots through enhanced visualizations like dot distribution map and geographically plotted scatter points. This spatial mapping is critical for identifying areas of high incident concentration, informing targeted interventions, and resource allocation.

b) Temporal Analysis

In parallel, the temporal aspect scrutinizes the chronology of events. Time-Day of Week analyses, utilizing line plots and stacked bar charts, unravel patterns and trends over hours, days, and months. This temporal scrutiny exposes fluctuations in accident occurrences, aligning with variables such as commuter cycles and seasonal variations.

To combine these threads, we have further defined accident hotspots using K-Means clustering in python (Jupiter notebook) for distinguishing the subtle groups that could otherwise go undetected. The k-means clustering is one of the simplest clustering methods using only simple geometric distance calculations (e.g. Euclidean distance or others). It is a far more efficient choice when compared to nonlinear methods like spectral clustering which uses a nearest neighbor graph to calculate distance between the clusters [6]. The resulting clusters are interpreted using lifestyle patterns and borough-specific features in addition to being visualized for spatial association.

Conclusively, the spatial-temporal categorization of the analysis provides a comprehensive understanding of the dynamics at play.

II. PROCESS

a) Temporal Analysis

Total Accidents by Severity

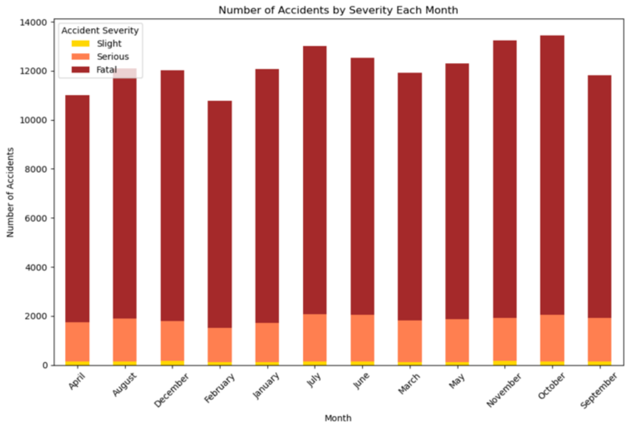

Fig 3 Stack Bar Chart

The above graph (Fig 3) presents a month-by-month breakdown of road traffic accidents, categorized by severity: slight, serious, and fatal. We can conclude that, it is apparent that slight accidents are the most common, maintaining a consistent lead throughout the year, with counts ranging roughly between 10,000 to 12,000 per month. Serious accidents are significantly lower, suggesting a frequency about a quarter to a third of slight accidents. Fatal accidents are the least frequent, depicted by the smallest portion at the top of each bar.

The total number of accidents each month stays relatively stable, indicating no drastic seasonal variation in accident frequency. However, the proportional representation of each severity type remains consistent across months, implying that the underlying risk factors affecting accident severity do not significantly fluctuate throughout the year.

The line graph (Fig 4) displays the number of accidents at different hours throughout the week. There's a noticeable peak during typical commuting hours, particularly on weekdays, indicating a higher risk of accidents during rush hours. The graph depicts the number of accidents plotted against the hours of the day, differentiated by days of the week. It shows two pronounced peaks, typically around 8 AM and between 4 PM to 6 PM, corresponding to morning and evening rush hours when traffic density increases as people commute to and from work.

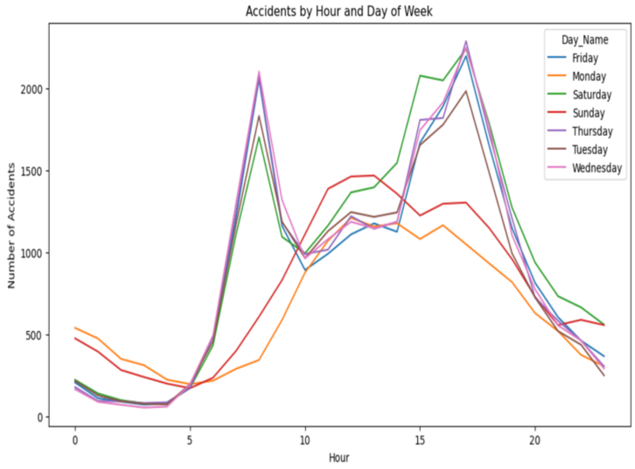

Fig 4 Accidents by hour and days of week

Notably, the pattern is consistent across the weekdays (Monday to Friday), with the number of accidents surging to around or above 1500 during these peak times. The weekend days (Saturday and Sunday) show a different pattern, with a less pronounced morning peak and a gradual increase in accidents throughout the day, peaking in the late afternoon.

The data suggests that the risk of accidents is significantly higher during rush hours on weekdays, which could be attributed to factors such as increased traffic volume, higher travel speeds, or driver fatigue. On weekends, the distribution is more spread out, possibly due to varied travel times and leisure activities.

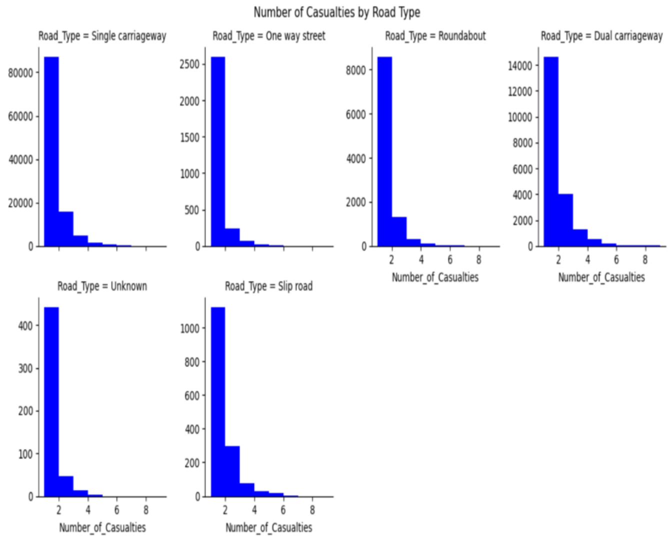

Fig 5 Number of Casualties by Road Type

The above figure (Fig 5) histogram shows that the 'Single carriageway' roads show a significant number of casualties, with the highest bar exceeding 80,000, indicating that most casualties occur on this type of road. We further analyze that the 'Dual carriageways' report fewer casualties, with the highest bar around 14,000. with 'Roundabouts' and 'One way streets' have even fewer casualties, with their tallest bars not surpassing 8,000 and 2,500 respectively. 'Slip roads' and 'Unknown' road types have the fewest, with the number of casualties peaking below 1,000.

From the above graph we can conclude that the single carriageways are the most common site for casualties. This might be due to a greater number of such roads, higher traffic volumes, or potentially more dangerous conditions compared to separated carriageways.

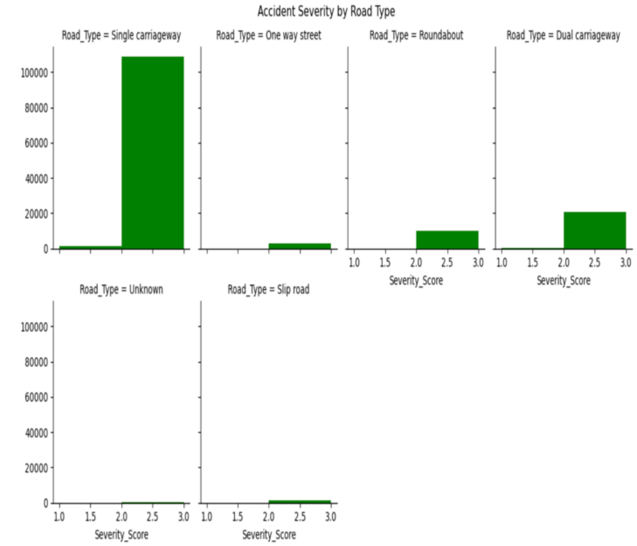

Fig 6 Accident Severity by Road Type

The above bar charts (Fig 6) for 'Single carriageway' and 'Dual carriageway' roads show the severity scores, with a higher concentration of bars at the lower end, indicating a larger number of accidents with 'Slight' severity. The severity distribution is relatively uniform across road types, with most accidents classified as 'Slight'. The 'Dual carriageway' shows a noticeable quantity of accidents with a severity score of 2.0, which could represent 'Serious' accidents.

This data is crucial for traffic safety measures and road design improvements, suggesting a need for enhanced safety protocols on single carriageways and perhaps targeted speed enforcement on dual carriageways.

b) Spatial Analysis

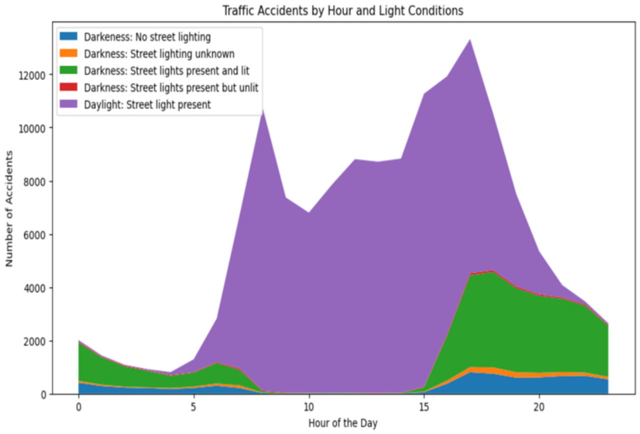

Fig 7 Traffic Accidents by Hour and Lighting Conditions

The stacked area graph (Fig 7) portrays the number of traffic accidents segmented by light conditions throughout different hours of the day. The graph reveals that most accidents occur during daylight with street light present, peaking at two critical points: around 8 AM and between 4 PM to 6 PM, which are typical rush hours. This suggests that the sheer volume of traffic during these times significantly increases the likelihood of accidents, despite the presence of daylight and street lighting.

As daylight fades, represented by the purple area, accidents in darkness with no street lighting increase, peaking after evening rush hour. This indicates that the absence of street lighting may contribute to accidents during the hours of darkness. Interestingly, the number of accidents during dark conditions with streetlights present and lit (green area) remains considerably lower and relatively consistent throughout the hours of darkness, emphasizing the importance of street lighting in preventing accidents.

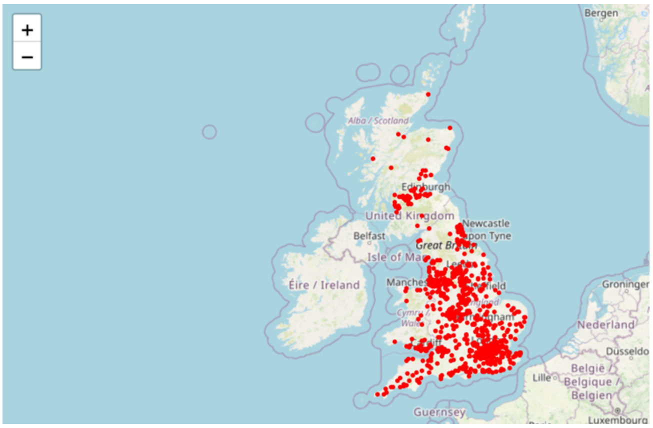

Fig 8 Accidents Hotspots in UK

The provided map (Fig 8) is a dot distribution map representing accident hotspots across the United Kingdom. In the map, each dot typically indicates the location of a traffic accident, and the density of dots suggests areas with higher frequencies of incidents.
From the visualization, we observe a heavy concentration of dots across England, particularly dense in the southeast region, including Greater London, which corresponds to a higher population density and traffic volume. The spread of dots follows the pattern of population centers and major transport routes, with cities like Manchester, Birmingham, and Liverpool also showing significant clustering.
Scotland shows a sparser pattern, with notable concentrations around Edinburgh and Glasgow. The spread in Scotland appears to follow major roadways and population centers as well.
Moving on Further the map indicates several insights such as the Urban areas with higher traffic volumes are likely to report more accidents.

Moreover, this data can inform policy decisions and infrastructure investments to improve road safety.

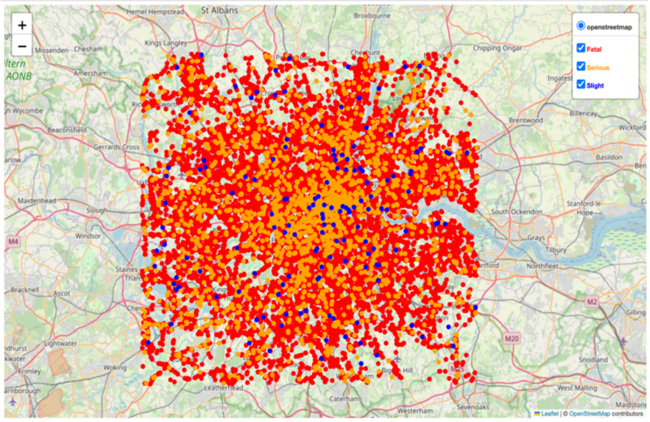

Fig 9 Accidents Hotspots in Central London by Severity

The above map (Fig 9) shows the spatial visualization of traffic accidents specific to the London area, with each dot representing a traffic incident with the accident severity. The colors of the dots likely correspond to the severity of the accidents with red for fatal, blue for slight, and orange for serious accidents. The concentration of dots across the map indicates the distribution and density of traffic accidents. A high density of dots in central London suggests that these areas have a higher frequency of accidents, which could be due to heavy traffic, a complex network of roads, high pedestrian activity, or a combination of these factors. The spread of dots along the main arteries leading out of central London indicates that major roads and intersections are common sites for traffic accidents. The red dots (fatal accidents) are more prevalent on high-speed routes or at complex junctions, this could suggest that these areas are particularly high-risk. This type of visualization is crucial for urban planners, policymakers, and traffic safety analysts as it helps identify areas where traffic accidents are more common.

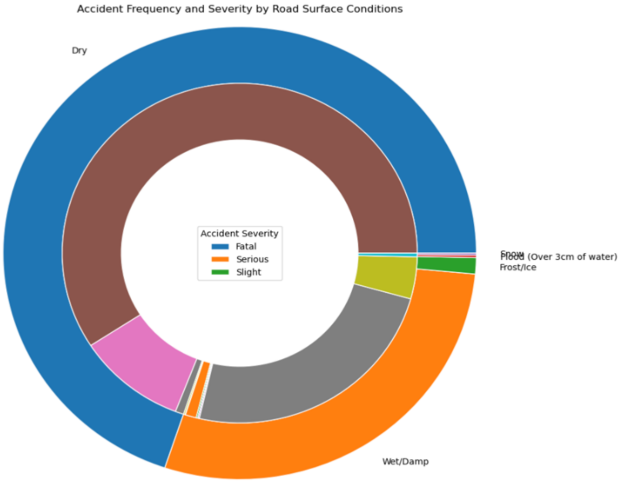

Fig 10 Accident Frequency and Severity by Road Surface

The image is a multi-tiered donut chart representing the frequency and severity of accidents in relation to road surface conditions.

The outer ring likely represents the frequency of accidents under various road surface conditions such as dry, wet/damp, flood (over 3cm of water), and frost/ice. The width of each segment in the outer ring denotes the number of accidents, with wider segments indicating a higher frequency.

The inner ring appears to categorize the severity of accidents within each road surface condition, segmented into fatal, serious, and slight. The color-coding would usually follow a standard pattern: darker colors for more severe, medium intensity colors for serious and lighter colors for slight.

Further we can see that the Dry conditions have the widest segment, suggesting they are the most common scenario for accidents, possibly because they are the most frequent driving conditions. Moreover, Wet, or damp conditions have a significant segment size as well, which could indicate that these conditions contribute substantially to the number of accidents. The segments for flood and frost/ice are noticeably smaller, which may reflect their less frequent occurrence but could still be significant in terms of severity. Also, within each road condition segment, the proportions of accident severity indicate that most accidents are of slight severity, which is typical as minor accidents are more common than fatal ones.

As we can see that a large segment of serious and fatal accidents under frost/ice conditions would underscore the increased danger during such weather. It's a valuable tool for traffic safety analysis, indicating potential areas for intervention, like improving road conditions or increasing driver awareness in adverse weather conditions.

c) Spatial K-Means Clustering

We have used the clusters that are likely formed based on the proximity of accidents to each other. The K-means algorithm minimizes the variance within clusters, meaning it groups accidents that are geographically close together.

We can observe the Geographical Patterns such as the distribution of clusters appears to follow the geographic layout of the UK, with denser clusters in areas that would correspond to higher population densities and urban regions such as London. This suggests a higher frequency of traffic accidents in urban areas, which is a common trend due to higher vehicle concentration.

III. RESULT

Now concluding our results for the research questions which we were discussing, we can see across the United Kingdom, the road surface conditions reveal a distinct correlation with accident frequency and severity, dry conditions prevail in accident occurrences yet exhibit a diverse severity spectrum, whereas icy conditions, though less frequent, often result in graver outcomes.

Now analyzing the clusters, we can see from the (Fig 11), Numerically, Cluster 0 is characterized by a higher mean number of vehicles involved per accident (~2.80) and a concomitant elevated severity score (~2.87), implying a propensity for more severe accidents within this cluster. Conversely, Cluster 4 is distinguished by a notably higher mean speed limit (~62.37), intimating a potential nexus between higher velocity travel zones and the frequency of accidents. Clusters 1 to 3 exhibit lower averages in both vehicular involvement (ranging from ~1.62 to ~1.79).

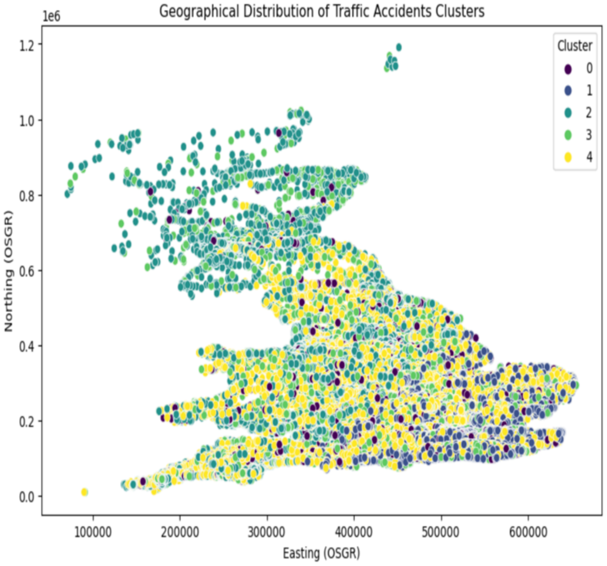

Fig 11 Cluster Results

This clustering pattern furnishes insights into the multifaceted nature of traffic accidents, with indications that road type, traffic density, and speed limits might be influential factors. Such data-driven analysis is imperative for formulating targeted road safety strategies and for the allocation of resources towards areas of higher risk.

CRITICAL REFLECTIONS

One crucial observation is that while dry road conditions are associated with the highest number of accidents, this may be misleading without considering the proportion of dry to adverse weather driving conditions. The severity of accidents in icy or wet conditions, although less frequent, is disproportionately higher, calling for targeted interventions during such weather conditions. Similarly, while daylight with streetlight presence correlates with a higher frequency of accidents, it is the absence of street lighting during dark conditions that significantly increases the risk, emphasizing the need for adequate lighting and possibly reflective road markings to enhance nighttime driving safety.

The type of road also plays a significant role in the frequency and severity of traffic accidents. Single carriageways, being the most common, show a higher number of casualties, which might be due to a combination of higher speeds, the presence of oncoming traffic, and fewer barriers to prevent head-on collisions. In contrast, dual carriageways and roundabouts, possibly due to their design and traffic flow management, report fewer casualties. This insight is particularly critical for urban planning and road design, indicating that road safety measures should not be uniformly applied but rather tailored to specific road types and conditions.

Furthermore, the geographical distribution of accidents, as revealed by the K-means clustering, indicates that accident hotspots are not randomly dispersed but are clustered around urban centers and major traffic routes. This pattern calls for a strategic approach to traffic management and accident prevention efforts, focusing resources and safety measures in these high-risk zones.

In conclusion, the synthesis of insights from the data suggests a need for a multifaceted strategy to improve road safety. Such a strategy should account for the variability in accident causation factors, including environmental conditions, lighting, and road design. It should also consider the temporal patterns of traffic flow and the potential for implementing smart technologies to provide real-time responses to changing road conditions. By doing so, policymakers and traffic safety authorities can better mitigate the risks and reduce the frequency and severity of road traffic accidents.

REFERENCES

[1] S. Haynes, P. C. Estin, S. Lazarevski, M. Soosay and A. -L. Kor, "Data Analytics: Factors of Traffic Accidents in the UK," 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT), Leeds, UK, 2019, pp. 120-126, doi: 10.1109/DESSERT.2019.8770021.

[2] R. Richard and S. Ray. A tale of two cities: Analyzing road accidents with big spatial data, 2017 IEEE International Conference

[3] D. Gao, X. Li, C. Yang and Y. Zhang. Spatial patterns analysis of urban road traffic accidents based on GIS, International Conference on Automatic Control and Artif

[4] E. Zhou, S. Mao and M. Li. Investigating street accident characteristics and optimal safe route recommendation: A case study of New York City, 2017 25th International Conference on Geoinformatics, Buffalo, NY, 2017, pp. 1-7

[5] K. Benkert, E. Gabriel and M. M. Resch, "Outlier detection in performance data of parallel applications," 2008 IEEE International Symposium on Parallel and Distributed Processing, Miami, FL, USA, 2008, pp. 1-8, doi: 10.1109/IPDPS.2008.4536463.

[6] C. Sinclair and S. Das, "Traffic Accidents Analytics in UK Urban Areas using k-means Clustering for Geospatial Mapping," 2021 International Conference on Sustainable Energy and Future Electric Transportation (SEFET), Hyderabad, India, 2021, pp. 1-7, doi: 10.1109/SeFet48154.2021.9375817.

WORD COUNT

Content	**Words**
Problem Statement	**209**
State of Art	**480**
Properties of Data	**330**
Approach	**495**
Process	**1493**
Result	**188**
Critical Reflections	**338**