Spatial and Temporal Analysis of Road Accidents in Greater London

Leyla Ahmadli

Abstract—We analyse road accident data for Greater London between the years 2018 and 2020 for this study. We applied temporal and spatial analyses to find patterns in the data. Our analyses show that Westminster and Camden boroughs have large density of traffic accident hotspots. We see positive relationship between the speed limit of the road and the costliness of the accidents and report that the slip roads have higher casualties per accidents among the junction types. Temporal analyses of data produce that weekday rush hours have more casualties from vehicle accidents, and we see seasonal maximums of the casualties on Fridays of the summer months.

1 PROBLEM STATEMENT

London is the center of England when it comes to economy, politics, culture and other aspects. There are slightly over 9 million people living in Greater London [1] and many more commute to London. Such a dense population creates congestion, and road accidents are more likely to happen. We are interested in finding patterns about congestion hot spots and traffic accident casualties with data science tools, and to present our findings we employ visualization techniques to facilitate effective consumption by readers. Visualization of data that illustrate areas with high rate of road accidents can help police and London authorities to identify the specific areas that may benefit from better road planning, emergency response access or other measures.

For our purpose, we need data that has times and locations of the road accidents along with information for road type, prevalent road condition and the outcome of the accident, namely, the number of vehicles involved and the number of casualties. Our main goal is to visualize hotspots of accident between 2018 and 2020 and find whether there are any patterns to the cause of these accidents. Which London boroughs have higher number of accidents, and how does it change between 2018 and 2020? Are areas in Central London more likely to have accidents? Furthermore, are there any daily, weekly, or seasonal patterns for accidents occurring? Are the attributes like weather condition, junction type and speed limit helpful in explaining the number of vehicles per accident or number of causalities per accident?

2 STATE OF THE ART

J. Griswold et al have conducted analysis on pedestrian collisions (collisions between a single vehicle and a pedestrian) with the aim of finding time periods where fatal pedestrian collisions are most prevalent. The study examines collisions between 1998 and 2007 from the Fatality Analysis Reporting System, a surveillance system in the United States maintained by the National Highway Traffic Safety Administration. The study reports that, although the evening hours have less pedestrian activity compared to the daytime (presented with a line plot), a large fraction of the single vehicle pedestrian collisions occur at twilight or first hours of darkness. This is presented with a temporal analysis heatmap that clearly shows how the hours corresponding to the highest density of fatal accidents change with seasonally adjusted sunset time. The study also explores whether there is a connection between driver experience (age groups defined as drivers between 16-20 of age vs those who are older) and alcohol consumption. The effect is less pronounced for experienced drivers who are not intoxicated, but it is still visible, especially in the winter months. [2]

In our study, we will also use temporal analysis with different time periods to examine whether there are relationships between these time periods and the number of causalities that happened in the traffic collisions. Finding this can help us understand the times of the day, of the week or of the year when there is a higher risk of casualties due to road accidents. Such knowledge can help the municipalities in their plan for prevention measures.

Michael L. Pack et al have used Incident Cluster Explorer web-based analytics tool to analyze incident dataset. The tool is especially good at clustering geospatial data and is capable of finding relations between different attributes. This application can be used to generate heatmaps on top of a geographical layout as well as histograms, scatterplots and parallel coordinates plots. The study discusses how well the available visualization methods are received by the sample audience and reports that the heatmaps are the desired method to communicate the information to the end users. [3]

We too employ heatmaps in our analysis to create a visual understanding of vehicular collision hotspots. The hotspots in Greater London for our dataset are discovered

through the density-based clustering method. This initial visualization is followed up with a more detailed presentation of the individual junctions and types of junctions where casualties due to road accidents are more likely to occur.

3 PROPERTIES OF THE DATA

The data in this report is taken from the UK (United Kingdom) Government data website data.gov.uk [4] under the subheading Road Safety Data. The data is available for individual annual periods, January through December, we used the most recent 3 datasets corresponding to 2018, 2019 and 2020. For each of the analysis years, we have the same attributes such as latitude and longitude of accident, reference number, accident year and date, day of week, time of accident, number of casualties, number of vehicles, junction type, and weather condition. These datasets represent the entirety of the UK; given we are interested in understanding the road incidents in the Greater London area, we filter the data to include the 32 boroughs of London and the City of London. For the three analysis years, the dataset has 71905 records in total with 36 feature columns. For the purposes of our analysis, the key features of this data are location and time attributes. The data has numeric values such as latitude, longitude, number of vehicles, number of causalities, speed limit, categorical values such as junction type and weather condition, and datetime series which we will use in our analysis.

The datasets for the 3 analysis years between 2018 and 2020 were concatenated to have a single dataset for easiness of analysis process. The rows where latitude or longitude were missing (around 100 rows) were dropped from the analysis in the initial processing because location is a particularly important feature for our analysis process. The categorical values were represented with integers in the raw data set, the excel file available from the same source was used to map the integers to their categories.

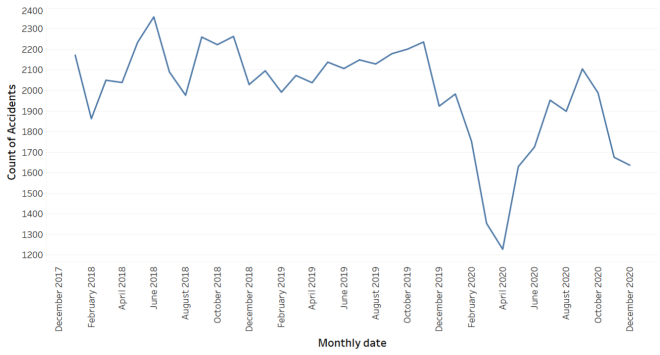

Figure 1. Monthly aggregate number of road accidents in London.

Our data has a lot of records making it useful to perform time and spatial analysis. Figure 1 depicts the monthly aggregate number of accidents for the analysis period. We can see that April of 2020 has the fewest accidents in the analysis period, almost half of the typical monthly number for the years 2018 and 2019. This corresponds to the beginning of the global Covid 19 pandemic, where the government led lockdown and stay at home orders reduced the need for driving. The data shows recovery from May 2020 onwards, but the overall trend for 2020 year is still lower than previous years. Because of this, to look for seasonal, weekly and hourly patterns in our data we compare the analysis years 2018 and 2019 and discard the data corresponding to the analysis year 2020 from this part of the study.

Figure 2 depicts the location of the road accidents present in our dataset, with number of casualties and accident severity overlay. We see that our dataset captures most part of Greater London. Hence, it is beneficial to find dense areas to visualize hotspots in this accident data.

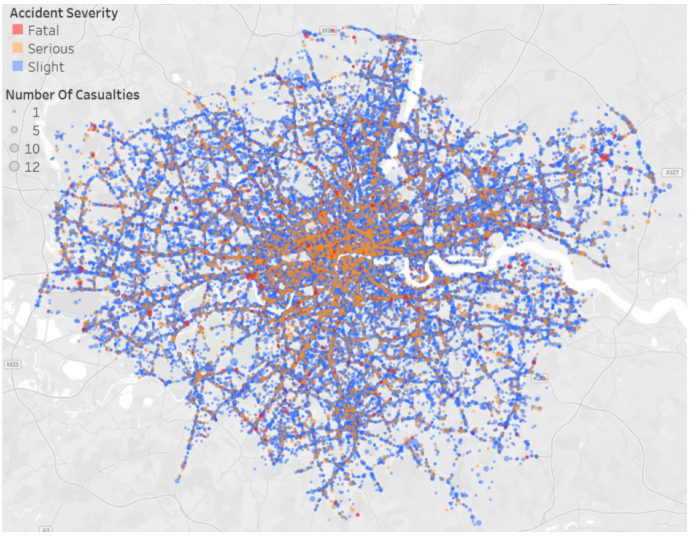

Figure 2. Map of London showing locations of road accidents.

4 ANALYSIS

4.1 Approach

We are interested in identifying patterns that help explain why some road incidents have higher casualties than others. Diagram 1 depicts the flowchart for our analysis process. We start the analysis by examining the locations where accidents are more likely to occur. With the location data, the longitude and latitude information, that we have in our working dataset, clustering methods are particularly useful to discover the problem areas. The method we employ is the density-based clustering method, which separates high

density areas from low density areas and specifies low density areas as noise [5].

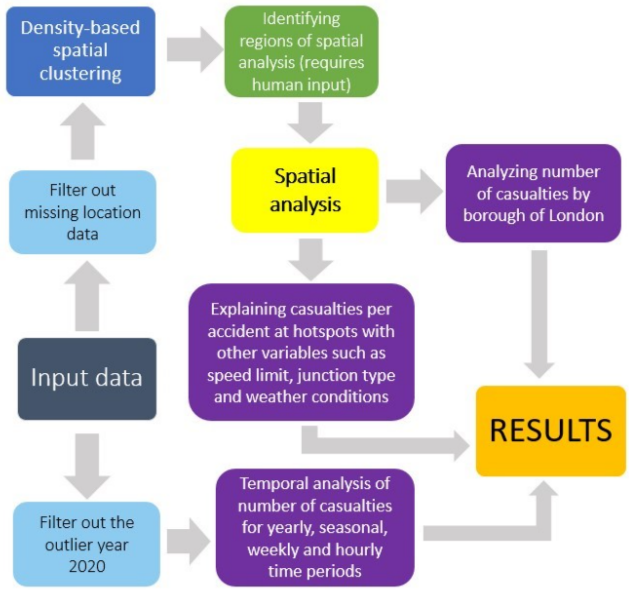

Diagram 1. Flowchart of analysis process.

The visualization of the output of the density-based clustering step helps the user to determine what areas the rest of the analysis should focus on. For the purposes of this paper, we take the entirety of the road accident hotspots in Greater London. Finer area can be selected to tailor the analysis, and consequently the solution to the road incident problem, to the specific portions of the city. With the highest risk areas identified, we look at how factors like weather, speed limit of the road, the junction type affect the outcome of the vehicle crashes. The outcome of vehicle crash metrics we are interested in understanding are designed to illuminate what damage will be given an accident occurs. We define two metrics for this purpose, number of casualties per accident which illustrates the human cost of the accidents, and number of vehicles per accident which illustrates the material cost of the accidents. Right away, one hypothesis that we want to explore is whether the outcome of a vehicle crash is costlier when the speed limit of the road is higher. Linear regression models are developed for this part of the study these relationships and the outputs are presented in regression plots. The regression plots both demonstrate the sensitivity of the outcome of road accidents to the modelled risk factors as well as the estimated uncertainty of the analysis. For the other risk factors, we explore the dependency with combination of line plot and bar chart.

It is also important to study whether there are any temporal patterns in the road accident data. The dataset selected has a wealth of information about the time of accident like the time of the day, day of the week, season and year. We want to see whether there are weekly patterns like commuting to and from work at peak hours, or seasonal patterns that can be attributed to the number of daylights hours. We study the temporal patterns by generating heatmaps for successively more granular time windows (yearly and seasonal, followed by weekly and seasonal, followed by weekly and hourly). A temporal heatmap helps ease the visualization for the end user and makes the comparison of the temporal patterns over multiple analysis years easier. The first temporal heatmap helps guide the analysis by identifying the outlier analysis years that should be removed from the study to ensure the findings are robust.

4.2 Process

We start the analysis process with a temporal analysis of the number of casualties for each season and year. We are looking to see whether there is a significant variation in the number of casualties across the largest time windows of the study. We use the astronomic definitions for the start and end of the seasons (from equinox to solstice or from solstice to equinox) rather than calendar definitions (from 1st of the month to 1st of a month) to divide the year into four quarters. Figure 3 depicts the temporal heatmap of the number of casualties for the four analysis seasons and the three analysis years. Though the spring of 2020 has seen the minimum of the number of casualties in Greater London, the entirety of the analysis year 2020 shows a reduced number of casualties compared to the earlier two analysis years. This is due to the global Covid-19 pandemic, and the lockdowns and stay at home orders created in response to the pandemic which reduced the need for driving. This agrees with the understanding from Figure 1. For this reason, to continue with the rest of the temporal analysis, we limit our data to the analysis years of 2018 and 2019 only.

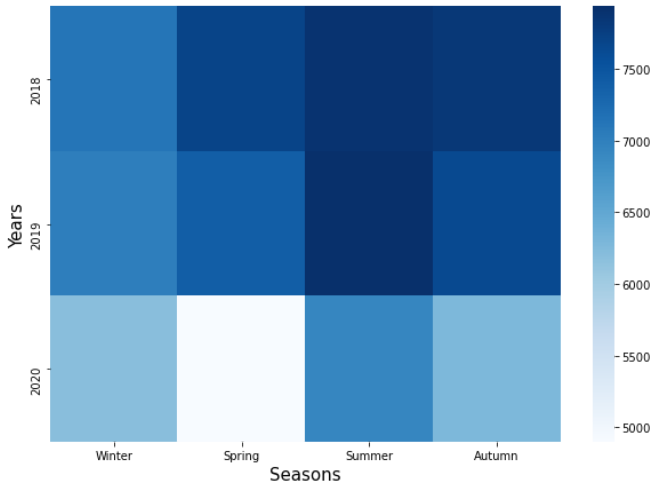

Figure 3. Heatmap for number of casualties vs. years and seasons.

For the selected analysis years, we next explore whether there are weekly patterns in the number of casualties and whether the weekly patterns change with the seasons. Figure 4 depicts the temporal heatmap for 2018 (left) and 2019

(right). The striking patterns emerging from these heatmaps are, (i) Fridays of the Summer months are when the number of casualties due to vehicle accidents are highest and (ii) the Winter season and Sundays of all seasons have lower numbers for casualties than the rest of the year. The former pattern is understood to be due to the higher number of after-work events that happen, especially in the summer months when the days are longer, and the weather is more pleasant to venture outside. The winter months on the other hand see people preferring to spend time inside, which reduces the number of drivers on the public roads and the number of casualties due to road accidents. Sundays are the day before the beginning of the workweek when people are more likely to rest at home and drive less than they do on the other days of the week.

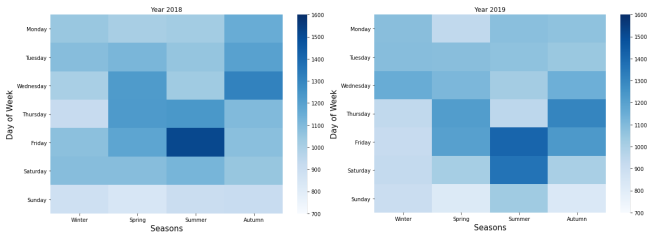

Figure 4. Heatmap number of casualties vs. day of week and seasons (left for year 2018, right for year 2019)

In Figure 5 we compare the hourly and weekly number of casualties aggregated over all four seasons for the analysis years 2018 (on the left) and 2019 (on the right). We see that the weekly and hourly patterns are quite similar between these two years. The daily pattern for the number of casualties shows peaks on weekdays from 7am to 10am and from 3pm to 8pm. These time periods are associated with morning and evening rush hours. This is a different pattern than that observed in [2] where the daily peaks of the casualties occur at twilight and early evening hours. We think the difference of outcome arises since compared to the US (United States) where driving is more prevalent at all times of the day, in the Greater London area driving primarily occurs to commute to work. Besides the clear rush hour patterns, we also see less pronounced increases in the number of casualties on the evenings of Fridays and Saturdays. This is expected to be due to the flexibility to stay outside longer on these days given the days after these are not workdays.

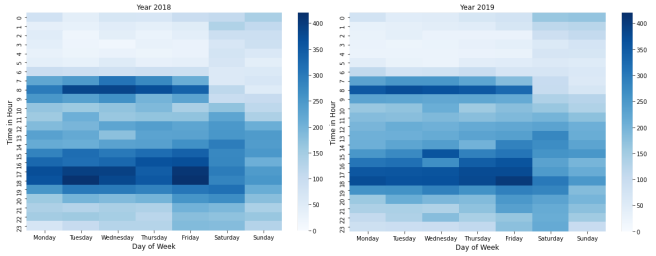

Figure 5. Time in hour vs day of week. Left for year 2018, right for year 2019.

We also compared the number of causalities by boroughs of London. From Figure 6 we see that the analysis years 2018 and 2019 have similar patterns of distribution of the number of casualties across the 32 London boroughs and the City of London. From Figure 6 we can see that for all analysis years, the highest number of casualties from vehicle accidents happens in Westminster while the City of London is the safest with the lowest number of casualties from vehicle accidents. Westminster is a dense area in the center of London, which is also a shopping, tourism and cultural destination. This area saw high vehicle traffic in the analysis years 2018 and 2019 that contributed to the high number of casualties there. Figure 6 shows that in the analysis year 2020 Westminster is no longer the borough with the highest number of casualties, which is understood to be due to the significantly reduced shopping and tourism related trips to this borough since the beginning of the Global Covid-19 pandemic.

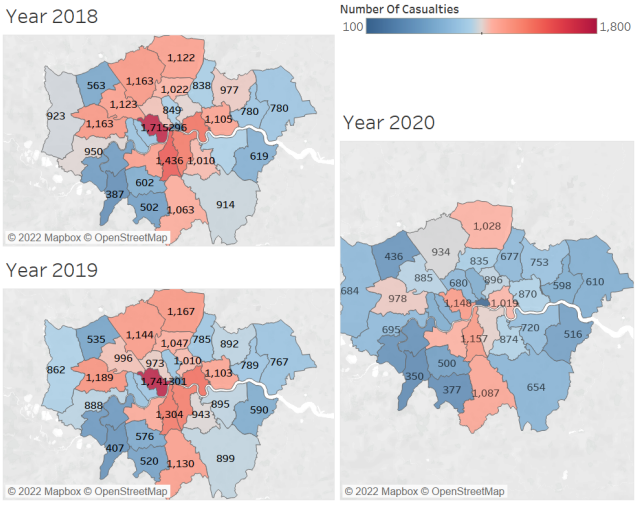

Figure 6. Number of casualties by boroughs.

We will further analyze the road accident data with density-based clustering. The density-based clustering method is good for detecting noise and can cluster the remaining data to produce denser areas while the less dense areas are counted as noise. To find hotspots in London where more accidents happen compared to the other areas, we apply density-based clustering to our location attributes and visualize the result on a map. The clustering method filtered out 32609 of the 71897 record as noise and produced 1173 clusters of various sizes. Since we are interested in identifying the hotspots, we change the labels of clusters with their sizes to find the largest dense area. In Figure 7 we see that the largest dense areas of road accidents are in Westminster and Camden boroughs. This part of the city is central, so it is expected that road accidents will be high because of heavy traffic in this area. We can also see that the motorways connecting the center of London to the north and south have a high density of road accidents.

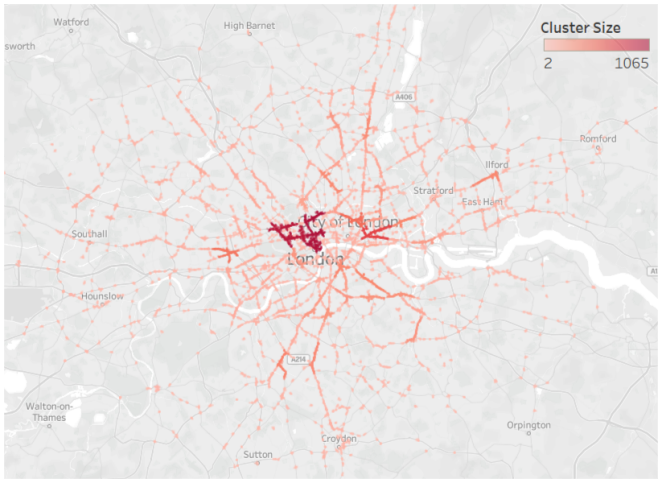

Figure 7. Hotspot of accidents in London.

In Figure 8 we analyze the severity of accidents defined as causalities per accident and vehicles per accident in roads with different speed limits. From the two regression plots, we can see a clear monotonously increasing relationship between the governing speed limit at the severity of the accidents. The shaded regions illustrate the uncertainty in the linear relationship which should be interpreted along with the results of the regression. Although there is sizeable uncertainty in both the regression of the casualties per accident and the vehicles per accident, the slopes in both regression analyses are found to be statistically significant, implying higher potential danger of roads with high speed limits.

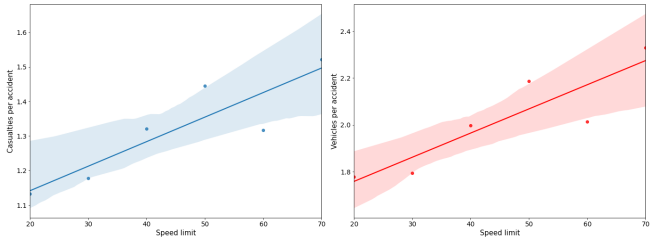

Figure 8. Linear regression plots for speed limit versus casualties per accident on left and speed limit versus vehicles per accident on right

Figure 9 shows how the number of casualties per accident change for different weather conditions (left) and junction types (right). Although it may be expected that hazardous weather conditions would lead to a higher number of casualties, we see that the fine weather has higher casualties per accident than for example snowing weather. This is likely due to the higher alertness that drivers display when driving in worse weather conditions that help make the accidents themselves less dangerous. Analyzing the casualties per accident, we see that the slip roads are the most dangerous type among the junction types. Slip roads are where one car must accelerate from a stopping position to match the speed of the other cars. Miscommunication among the drivers can create conflict points and potential for vehicle collisions with large speed differential resulting in costlier accidents compared to other junctions.

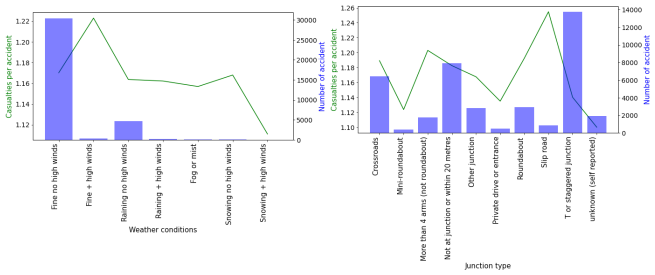

Figure 9. Casualties per accident in different weather conditions (left) and for different junction types (right).The bar charts show number of accidents (right axis) and line plots show casualties per accident (left axis).

4.3 Results

We performed clustering methods to identify the road accident hotspots in Greater London and followed up the analysis to present the results with geospatial and temporal heatmaps, and regression plots. In Figure 10 we are looking at the heatmap of road accident hotspots in Westminster and Camden from the result of our density-based clustering. The junctions with the highest number of accidents can be prioritized for use of the councils' safety funds.

Figure 10. Heatmap of Westminster and Camden boroughs

The temporal heatmaps showed that among the four seasons, summer months have the highest number of casualties due to road accidents and winter months have the lowest. Fridays in summer show an especially high number of casualties compared to the other days of the week. The intraday patterns we discovered show peaks of number of casualties in the morning and evening rush hours. We use

regression plots to highlight how the speed limit affects the number of casualties per accident and report a positive relationship that is statistically significant. The closing findings of this study are that among weather conditions, the fine weather with high winds proved most hazardous when looking at casualties per accident and the slip roads show higher casualties among junction types.

5 CRITICAL REFLECTION

The aim of this study is to understand where, when the road accidents mostly occur, and how the vehicle accidents are different when it comes to their human and material cost. We employed geospatial clustering via density-based clustering [5] method to home in on the problem hotspots and looked at whether the outcome of the accidents be it number of casualties or the number of vehicles damaged from the accident show clear patterns when looking at the weather conditions, the speed limit of the road or the types of junctions. After initial steps of visualization, the approach presented in this study allows human users to tailor the continuation of the analysis based on their needs in terms of selecting which of the vehicle accident hotspot areas are selected for deeper evaluation.

This study can further be enlarged by applying multivariate clustering to better divide road accidents into clusters. In addition, having a new feature in the data that shows whether the accident happened because of the negligence of driver, because of pedestrian or some other reason can help us further investigate the cause of these accidents. The results of geospatial clustering can be used as a guide for boroughs to strategize investments in the safety of the road infrastructure and emergency response in the places where occurrence of accidents is high.

We report that the roads with higher speed limits have costlier accidents compared those with lower speed limits, and the slip roads which create a conflict point between low speed and high speed vehicles can result in higher number of casualties per accident. However, the fine weather with high winds is found to result in the costliest accidents surpassing the foggy, rainy or snowy conditions. This is partially clouded by the fact that people are less likely to drive in hazardous weather conditions and concentrating on the Greater London area only gave us a much smaller sample of accidents in snowy or rainy weather compared to fine weather. The impact of weather on the outcome of accidents can be studied more deeply by looking across the country, especially the northern parts of UK, to get larger sample size for diverse weather conditions.

The temporal heatmaps helped us find that the Fridays in summer months tend to have a higher number of casualties due to vehicle accidents. Refining the visualizations to the hours of different days of the week shows how the hours around morning and evening rush hours are more likely to have vehicle related casualties. Our analysis was done for the most recent three full years of data, but the study can be extended to include the periods before 2018 to see whether the findings of the temporal analysis are stationary or if there are some longer-term trends that a shorter analysis may be missed.

Table of word counts

Problem statement	248/250
State of the art	405/500
Properties of the data	466/500
Analysis: Approach	491/500
Analysis: Process	1177/1500
Analysis: Results	197/200
Critical reflection	467/500

REFERENCES

The list below provides examples of formatting references.

[1] https://data.london.gov.uk/dataset/londons-population

[2] J. Griswold, B. Fishbain, S. Washington and D. R. Ragland, “Visual assessment of pedestrian crashes,” Accident Analysis & Prevention, Volume 43, Issue 1, 2011, Pages 301-306, ISSN 0001-4575, doi: 10.1016/j.aap.2010.08.028.

[3] M. L. Pack, K. Wongsuphasawat, M. VanDaniker and D. Filippova, "ICE--visual analytics for transportation incident datasets," 2009 IEEE International Conference on Information Reuse & Integration, 2009, pp. 200-205, doi: 10.1109/IRI.2009.5211551.

[4] https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

[5] Sander J. (2011) Density-Based Clustering. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_211