Analysing factors influencing motor vehicle accidents and localising their hotspots in Glasgow (2005-2014).

CS989: Big Data Fundamentals

Coursework - Processing and Analysing Data

WORDCOUNT: 3260

Contents

List of figures:

List of tables:

Introduction:

Last year, 2020, saw a marked decrease in the number of motor vehicle accidents (MVA) due to the unforeseeable circumstances that were multiple covid-19 related lockdowns. Unsurprisingly, this led to a 68% decrease in MVA for the month of April and made national news at the time which inspired much debate about road safety (HRM Department for Transport, 2020).

Here, we will examine almost a decade's worth of MVAs from 2005-2014 for the local authority area of Glasgow - S12000043 (today: S12000049) during that period (seen in figure 1.).

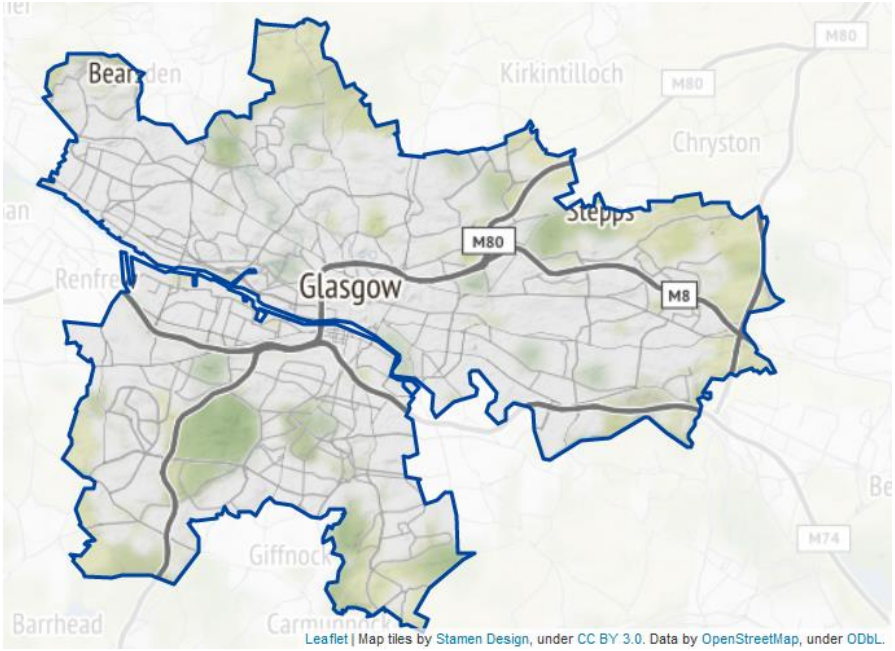


Figure 1. Modern day S12000049 (Local Authority Glasgow).

Image sourced from FindthatPostcode and used under creative commons license CC BY-NC-ND.

Over 13,000 MVA were localised in S12000043 between 2005 and 2014. This averages out to just over 1,450 MVAs per year and stands in contrast to more recent data that suggests just under 1,000 MVAs per year (Transport Scotland, 2019).

This indicates that there is still much that can be ascertained about the frequency of MVAs for specific locations and at specific times. Using this data on road MVAs and examining the greater Glasgow area we can perhaps pinpoint accident hotspots and areas of greater concern. This will allow for improved future planning of road construction and regulations, accident management and the implementation of road and vehicular safety measures and precautions to protect both drivers and passengers.

Introduction to the dataset:

This dataset is created from data supplied by the Government of the United Kingdom through the Department of Transport (Department of Transport, 2021) and made accessible on Kaggle (Fisher-Hickey, 2017). It encompasses a wide variety of traffic data for the years of 2005-2014 (but missing data from 2008). It contains public sector information that is licensed under the Open Government Licence 3.0 (The National Archives, 2021).

Originally this dataset contained over 1.6 million MVA's with 33 unique variables. Much of this data was not relevant to the target area (S12000043) and was therefore excluded. This left the filtered dataset with 13 variables (20 variables were dropped) and 13,355 overall unique MVAs for the period between 2005 to 2014 in the Glasgow area. Some variables were easily justified in their removal, such as Lower Super Output Area only affected England and Wales, therefore the removal would have no impact on our results. Similarly, local motorway authority was only used to filter the dataset, it had no further analytic value.

Equally, the accident index was only valuable for removing duplications after initial indexing. There was missing data in various parts of this dataset, and this further justified dropping more variables; for instance, the "junction detail" variable was missing data for Glasgow in its entirety and the junction control variable was missing almost 50% of the data. Therefore, the choice was made to exclude these from the scope of this report's analysis. This process is further illustrated in Appendix 1.

Ignoring the accident index and the local authority variables as well as grouping the latitude and longitude variables (since these together make up the geodetic location in question) this leaves the dataset with 10 variables. Eight were numeric (including a datetime format), and two others were character based (weather and road conditions) but later converted to numeric. Appendix 2 also provides context for the discrete variables with a table of explanations and their respective total counts within the dataset.

This filtered dataset was used for all further analysis to provide valuable insights into the objectives of this report as outlined previously.

Results & Discussion:

Descriptive statistics:

To gather and provide basic insights into the data and lay a framework for potential in-depth analysis, the value of the different variables can be quantified in a correlation map (figure 2.) that pivots the variables against each other.

This indicates that there are very many variables that are seemingly not effective predictors of others; it would be presumable that accident severity is impacted by road surface conditions (and weather conditions), but this is apparently false. Some are unsurprisingly highly correlated, such as the weather conditions and road conditions. None of the variables showed a significant negative correlation; in fact, most variables trended towards neutrality.

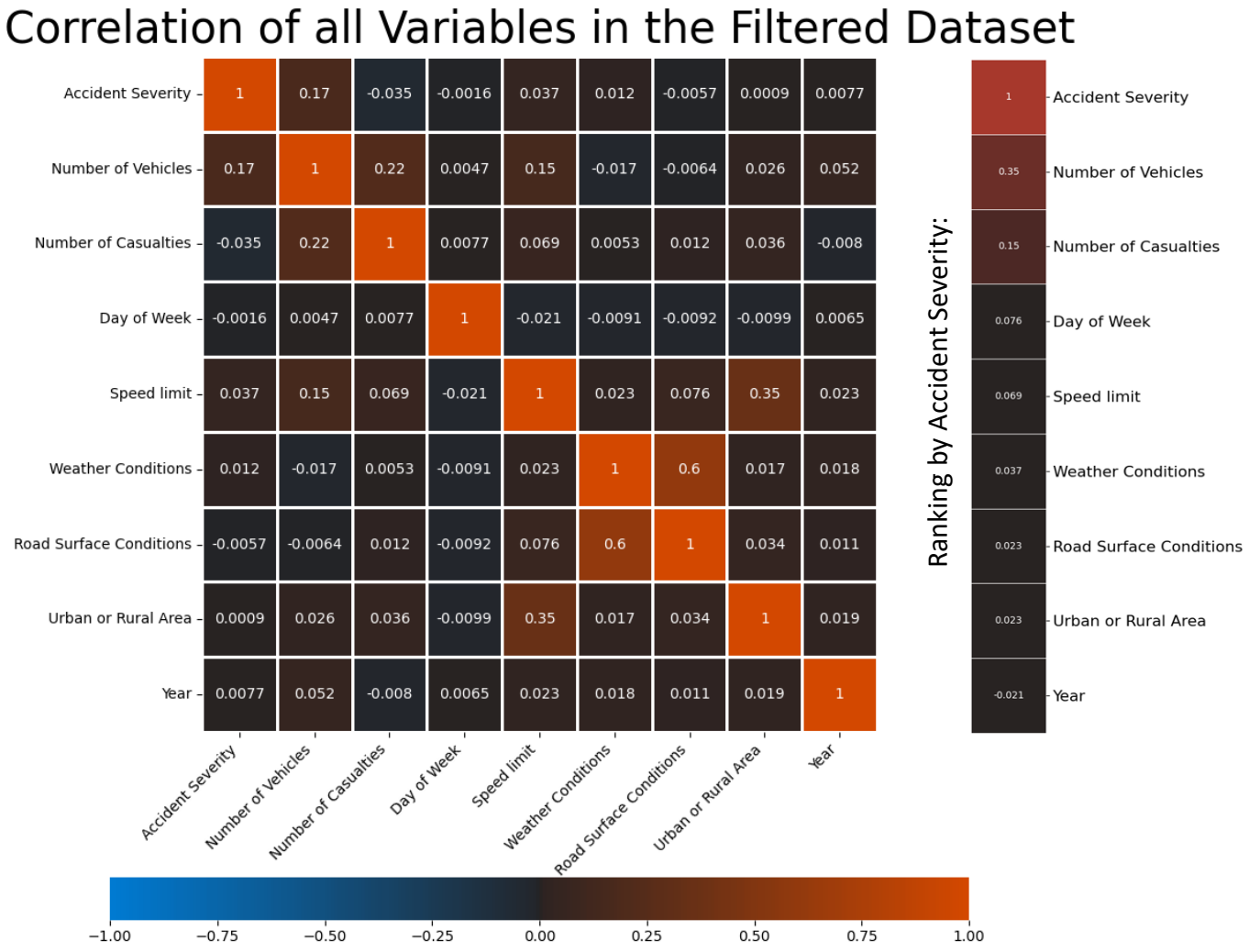


Figure 2. Correlation heatmaps.

The separate ranked correlation map with Accident Severity as the dependent variable showed that the independent variable with the most significant impact was the number of vehicles involved in the accident which ultimately, when visualised using Kernel Density Estimation (KDE), informed us about the number of casualties as well.

KDE visualises the spread of our data amongst the specific variables as they are pitted against each other. This aided in the understanding of how different speed limits were related to each of the variables and followed realistic expectations: vehicles moving at higher speeds (on roads with higher speed limits) and on worse road conditions correlated with more vehicles being involved in the accident and a higher casualty count typically but it also shows us that these are not mutually exclusive, accidents with high casualty count and numerous vehicles can still happen at lower speed limits. It also displayed that wet road surface conditions (that stand in linear relation to rainy weather conditions) correlate with more high-speed accidents when compared to dry road surface conditions (pairplot visualisation added as Appendix 3.).

It is important to bear in mind that throughout this dataset the predominant Accident Severity was "slight" (>85% of the data, 11422 cases); the rest were either serious (1809 cases) or fatal (124 cases) and therefore we can establish that there will be an imbalance of other independent variables upon the 3 possibilities of Accident Severity, largely skewing the data if not accounted for. Intriguingly, within these 124 fatal cases, the vast majority occurred within the 30MPH speed limit albeit the vehicle might have been travelling in excess of that speed (table detailing fatalities per speed limit added as Appendix 4.).

Equally, the weather and road conditions were predominantly "*good*" (fine without high winds & dry road conditions) even though one would expect Glasgow, in extension of Scotland, to have more numerous and severe accidents during the "*bad*" weather conditions. Interestingly, this might point to the adaptability of Scottish drivers as they might be accustomed to worse conditions and therefore are not impacted by it significantly or alternatively, less people drive during bad weather. Thus, the natural conclusion from these variable relationships is that they precipitate a "*worse"* accident.

To analyse if these factors might have a lessened impact over time, for instance due to changes in road policies or improvements to motor vehicle technology, total MVA's between 2005 and 2014 can be studied (seen in figure 3.) and a general downward trend will be seen.

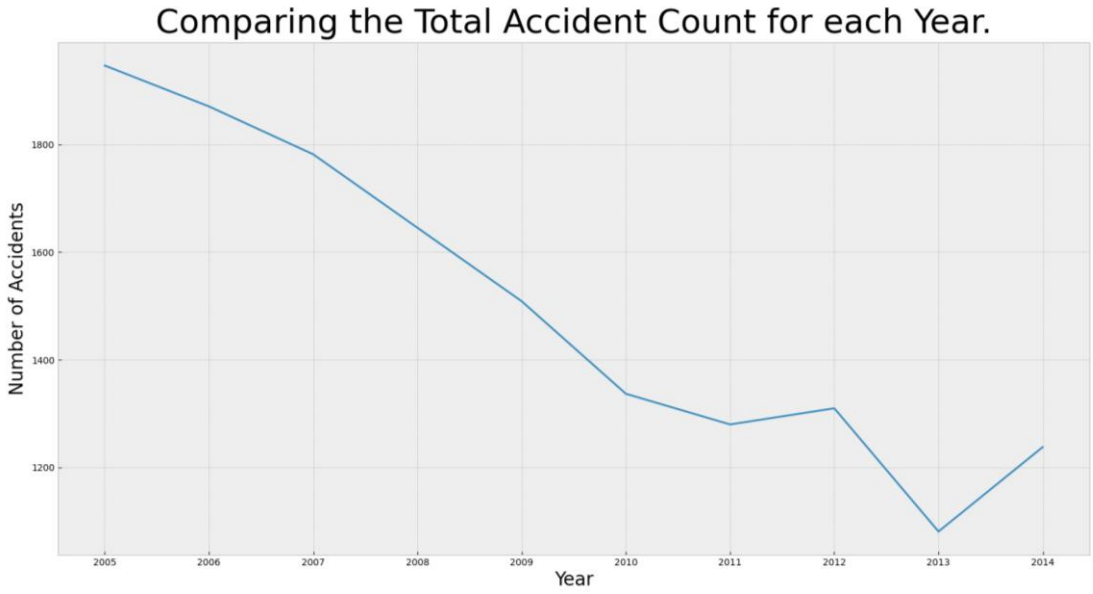


Figure 3. MVA's per year.

To contrast only the two extremes of this period, figure 4. aims to compare accident frequency on a 24-hour timescale for both years using their hourly average.

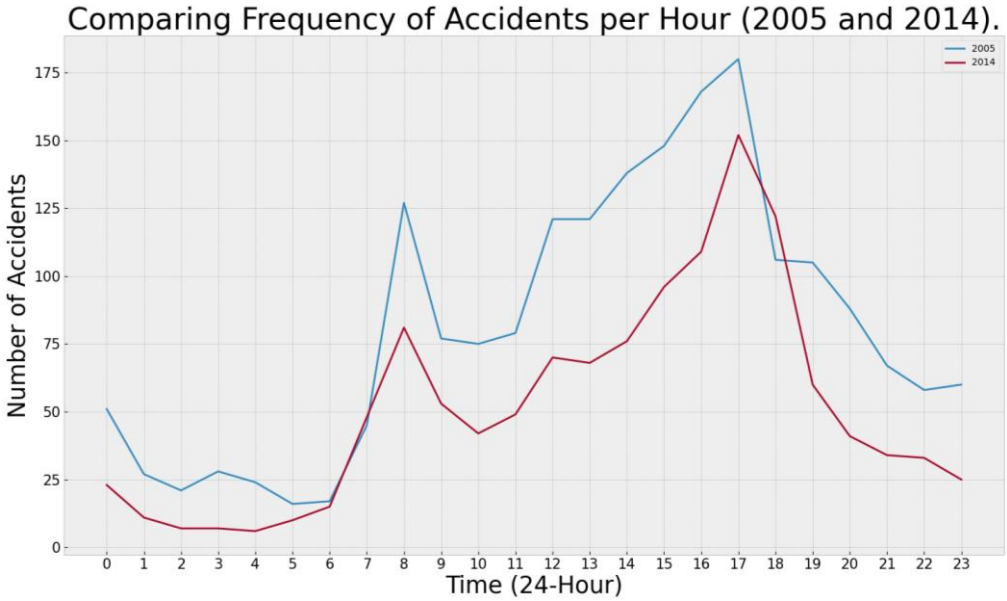


Figure 4. Annual 24-hr MVA frequency.

Visibly, fewer accidents occurred at almost all time points for 2014 when compared to the data from 2005 with the only exception being the total accidents for that year at 6PM (perhaps due to a rise in vehicle popularity or affordability).

Further exploration of the data analysed which hour had the highest amount of MVA's on average and how many vehicles were typically involved, this can be seen in figure 5. (the time of the accident was rounded down to the nearest hour).

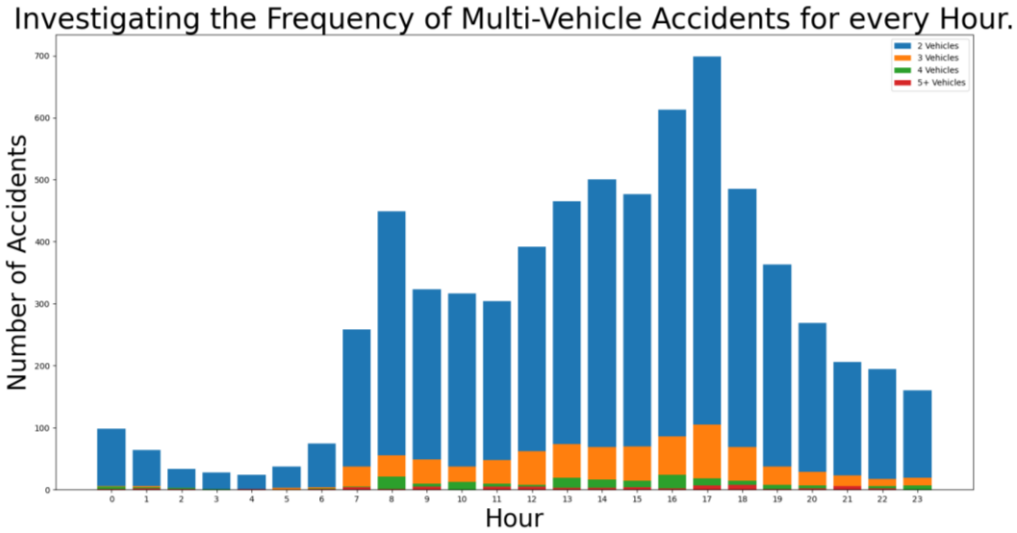


Figure 5. MVA's per hour (2005-2014).

The entire 24-hour period shows large numbers of multi-vehicle accidents whilst the daytime hours, particularly the *"rush-hours"* (8-9AM and 4-6PM), show a spike from the preceding hour and an overall large increase in multi-vehicle accidents that slowly declines again. This is in-line with expectations; more accidents occur as people are travelling to work and returning home after work.

In comparison to data on single-vehicle accidents (figure 6.), it can be said that the trend is mostly similar. However, there are more single-vehicle accidents during the early morning hours (0AM-5AM) whilst it remains dark. Single-vehicle accidents are more likely than multi-vehicle accidents during the early hours of morning presumably because of the bad visibility in which humans are harder to spot than vehicles. Alternatively, the bad visibility also causes off-road accidents in which no other vehicle is involved. Equally, driving while tired or under the influence of alcohol might also play a role.

However, outside of rush-hour times, the total multi-vehicle accident counts largely supersede single-vehicle ones.

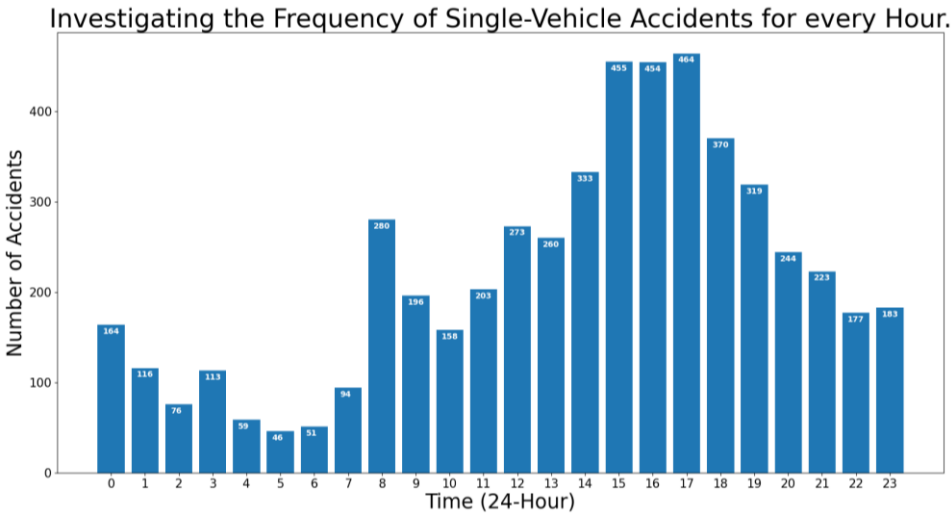


Figure 6. Single-vehicle accidents per hour (2005-2014).

Unsupervised Learning:

K-Means

Traditionally, K-means relies on user input to determine the numbers of clusters. To circumvent random guessing of an adequate number of clusters, the elbow method was used (see figure 7. for result) and this estimated that 4 clusters would be sufficient, as the line flattens dramatically beyond 4.

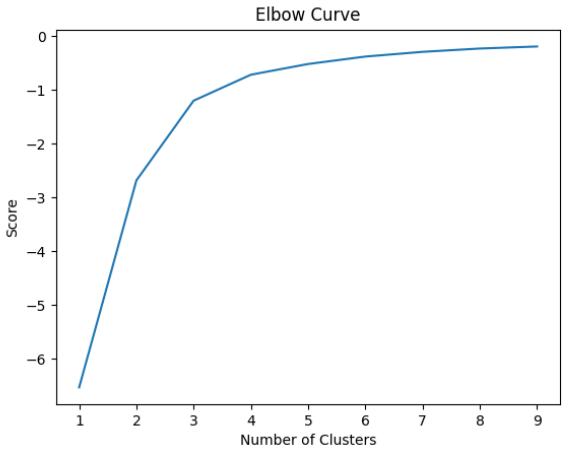


Figure 7. Elbow Curve for Kmeans algorithm.

As Glasgow had over 13,000 accidents and there are hundreds of junctions, from a rational standpoint, 4 clusters would not be enough, and this is evidenced further by the result of K-Means clustering in figure 8. Here we see all the accidents for Glasgow clustered into 4 unique clusters albeit there is no discernible correlation between them, they are arbitrarily split into roughly equal quarters with centroids given in black. Even with increased cluster count inputs, the grouping would be "arbitrarily" decided. This is due to the K-Means algorithm aiming to reduce variance but as the analysis requires the use of geodetic distances rather than variables with some form of linear relationship to one another, it fails to be of any analytical benefit. Therefore, K-Means is simply not suitable to handle this data analysis.

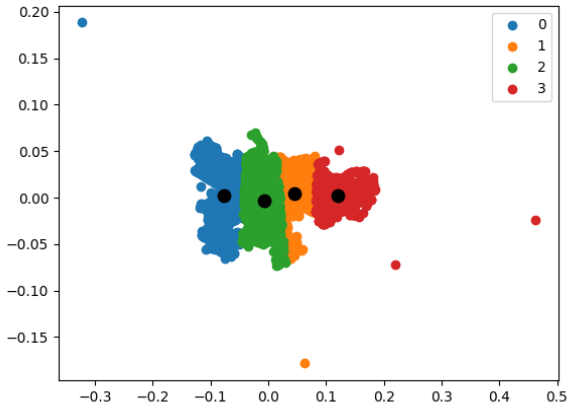


Figure 8. K-Means clustering of all accidents.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Pairing latitude and longitude of the individual accidents to one another and relating these to other accident coordinates, thus establishing "*hotspots*" where accidents were most likely to happen, a function from Pythons GeoPy package was used to accurately compute geodetic distances (further overcoming the K-Means limitations). One major benefit of this distance function over others of a similar nature is that it accounts for the curvature of the earth when calculating distances from Longitude and Latitude.

The main feature that compels the use of DBSCAN over other forms of hierarchical single linkage clustering algorithms and applications was that it does not rely on user input to form the numbers of

clusters, instead it computes these itself. The algorithm is also more finely tuned for the clustering of data such as in this report, particularly when compared to the previous K-Means clustering method. This was highly important as roads, particularly in the city centre, would be highly interconnected and the accidents would not be spaced far apart by nature of the city's design.

With these constraints in mind, initial analysis (as seen in figure 9.) concentrated on clustering all data in the Glasgow region and then focusing on the densest clusters throughout the 10-year period. First DBSCAN analytics used a *minimum-point* (MP) parameter of 10 (this is used to form "*dense regions*" by DBSCAN) and a *distance between* [accidents] *points* parameter of 30 and 50 meters (henceforth DBP, in the documentation: *eps*).

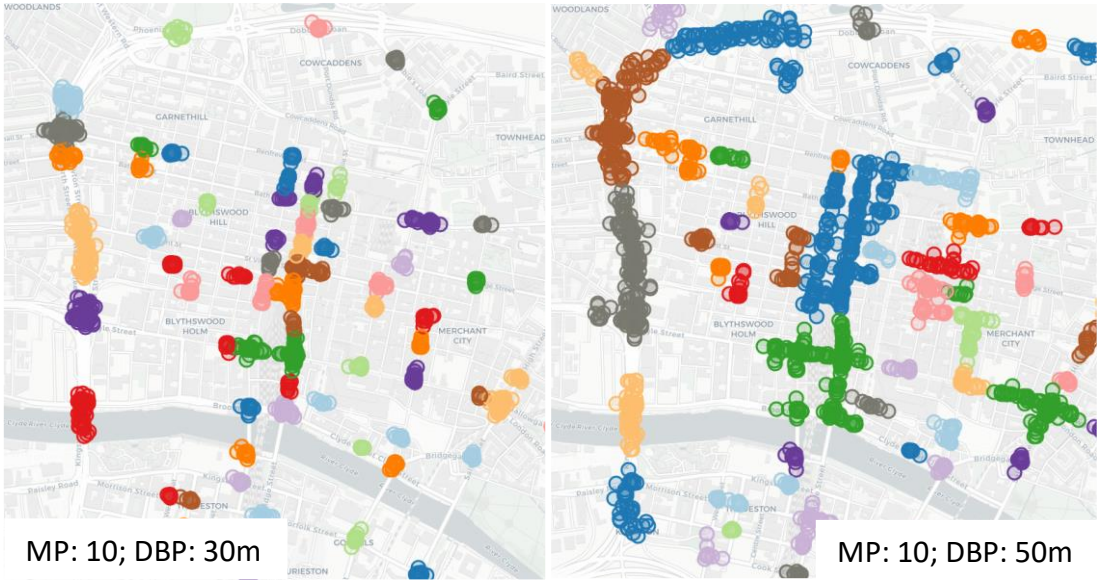


Figure 9. Comparing DBSCAN outcomes for all years combined.

It becomes apparent that the DBP is too large of a leeway (even at 30m) for the algorithm at this clustering level. This justifies further parameter adjustment, especially considering the city's layout. Reducing the DBP to 10m yielded clearer results, as seen in figure 10., and helped to ascertain that there were indeed pockets of clusters spread throughout Glasgow but that it would be difficult to pinpoint the most significant ones, particularly with such a small DBP specified.

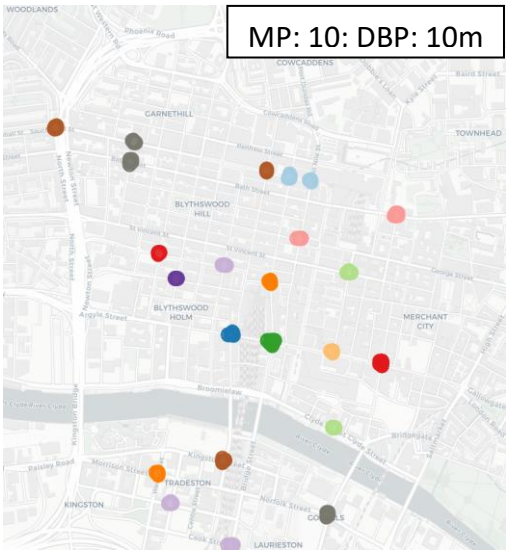


Figure 10. Further DBP lowering.

Nonetheless, a trend is seen with junctions at busy roads such as those in the city centre (Sauchiehall St. or Union St. for instance) and areas around Glasgow's central railway station (GCRS).

In figure 11. the MP and DBP parameters were both adjusted to 30 and returned only two clusters (wherein 30 accidents within 30 meters were clustered together)- This can be interpreted as two hotspots with some of the most accidents occurring within a moderately concentrated area.

Figure 11. Increasing MP and increasing DBP.

The light-blue cluster is known locally as "*the four corners*" and is infamous for seeing a lot of traffic during the day and being a remarkably busy junction during the night due to its vicinity to restaurants (particularly fast-food restaurants) and nightlife venues such as bars and clubs. This might also precipitate many single-vehicle accidents at this junction as there is a very high footfall in this area.

Figure 12. is the culmination of both parameters being gradually fine tuned to tailor the algorithm to our objective This provides a very good overview MVAs in the Glasgow city centre region when DBP and MP are tuned to avoid over- and underfitting.

Obviously, most clusters are seen at the intersections that are known to be busy with pedestrians or see high amounts of vehicular traffic, such as the Trongate junction.

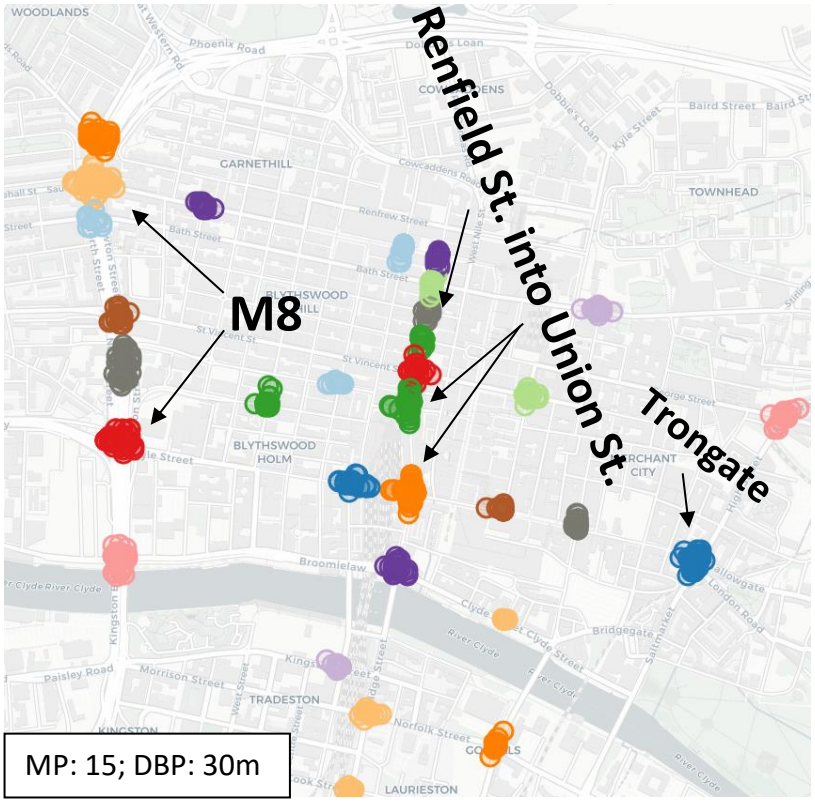


Figure 12. Accident hotspots in Glasgow city centre.

Equally, the M8 motorway and its intersections have very clearly defined groups of hotspots populating it. Some of Glasgow's busiest roads are also home to the most accidents (Renfield St. and Union St.). As there are undoubtedly very many accident hotspots in Glasgow, it necessitates asking how many of these are single-vehicle accidents and where are they predominantly located? Unsurprisingly, the most single-vehicle accidents occur near GCRS which also includes the aforementioned *four corners* (visualised in figure 13.).

Two other clusters can be seen near the pedestrian crossings at Bath St. and Sauchiehall St. with another one at a busy pedestrian crossing near George Square and finally, one more cluster at the end of the pedestrian-only zone towards the east end of Argyle St.

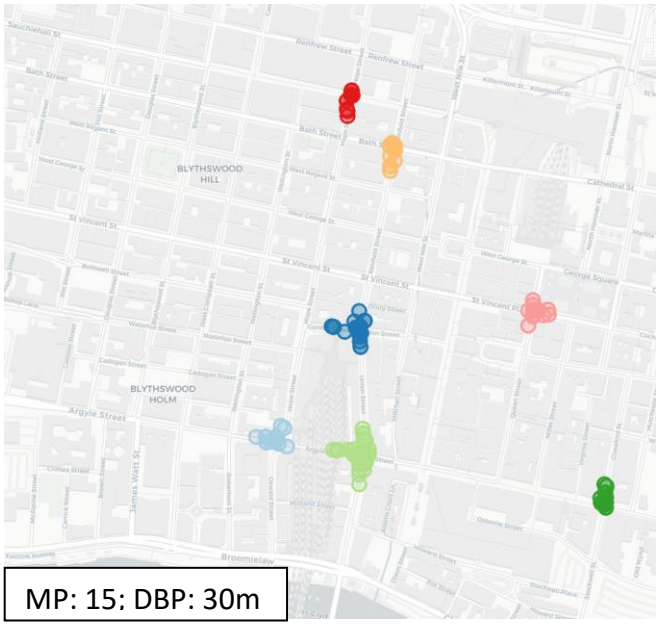


Figure 13. Clustering of single-vehicle accidents.

DBSCAN enabled an excellent visual representation of the analysis objective and furthermore, using the parameters *MP:15* and *DBP:30,* it also allowed for the computation of clustering-specific statistical metrics as seen in table 1.

Metric:	Outcome:
Estimated number of clusters	57
Estimated number of noise points	12088
Homogeneity	0.073
Completeness	1
V-measure	0.135
Adjusted Rand Index	0
Adjusted Mutual Information	0
Silhouette Coefficient	-0.219

Table 1. Metrics for MP: 15 & DBP: 30m parameters in DBSCAN.

A homogeneous clustering is achieved when all the data points within a cluster belong to a unique class (Shannon's Entropy). This is demonstrably low, however, the completeness score is a perfect 1. Completeness is a measure that considers how many data points that belong to the same class are clustered together. V-measure (also known as normalised mutual info score) was low, as it is a function of both homogeneity and completeness.

The adjusted rand index is 0, implying there is no overlap in the samples between clusters (as these are unique events, this was presumable but also displays that there are no duplications in our data). Similarly, adjusted mutual information was also 0. This accounts for chance and is an alternative to the adjusted rand index. The silhouette score being -0.219 implies that there is little significance in the distance between the clusters which confirms that accidents in one particular hotspot might not necessarily precipitate an accident at an alternative hotspot.

Supervised Learning:

Logistic Regression

Applying a logistic regression machine learning algorithm to our data provided an accuracy of over 85% across 20 runs. The Broyden–Fletcher–Goldfarb–Shanno algorithm was used as a solver, to account for the nonlinear issues that are faced in this dataset. Traditionally, logistic regression is not the most suitable to method to analyse nonlinear outcomes nor to analyse outcomes that are non-binary. Therefore, alternatives needed to be sought out.

AdaBoost: Ensemble-based Sequential Learning.

The supervised learning method used was Adaptive Boosting (AdaBoost). This is a statistical classification meta-algorithm that uses “weak-learners” in an adaptive manner that aims to reclassify results that were incorrectly classified by the previous learner (as it runs iteratively). The benefit of AdaBoost is that its weak-learners would systematically improve as the dataset was being analysed and this ultimately produces what is in effect a strong-learner.

AdaBoost allows parameters for its application to be defined. Here, these included the total number of estimators (default 50) at which boosting was terminated and the learning rate (0-1). A high estimator number was used (1000) and a very low learning rate (0.1) to encourage optimal machine learning.

When the dataset is split into two parts, one being a test set (30% of the total data) and the other being a training set (70%) of the total data, AdaBoost returned an accuracy result of approximately 0.86 (86%) across 100 unique runs for which it was aiming to determine accident severity based on the number of vehicles, casualties, speed limit, urban or rural area, and year. Since the initial premise is that weak-learners should be at least somewhat better than a guess (suggesting a 50/50 chance of being correct), this showcases AdaBoost’s immense predictive power as accuracy is based on correct predictions divided by the total number of predictions made by the algorithm.

XGBoost: Extreme Gradient Boosting.

XGBoost (XGB) is a regularising gradient booster that is very popular for being highly efficient (and thus very fast). When compared with AdaBoost it ran roughly 50% faster (AdaBoost ran at approximately 12 seconds per run in comparison to XGB's 8 seconds). It was necessary to use the multiclass classification methodology of XGB as our outcome variable, Accident Severity, had 3 levels. Table 2. showcases the results of 10 XGB runs.

Metric:	Outcome:
Precision	0.29
Recall	0.33
Accuracy	0.86

Table 2. Average results of 10 XGBoost runs.

Precision is a measure of true positives in the selection made by the algorithm and Recall is a measure of true positives selected by the algorithm out of the entire available true positives in the sample dataset. Both are considerably low when averaged out (Precision 29% & Recall 33%). The accuracy of the XGBoost algorithm was however still at 86%, so much like AdaBoost, its predictive power was significant in determining the dependent variable from the independent variables. A measure of classifier precision and robustness is the F1-score, and this was calculated to be approximately 0.3.

Supervised learning might be a reliable method of ascertaining which hotspots have a higher correlation to some of this dataset's variables. However, this is a multifactorial problem that requires more overall considerations than just what is outlined in this report or analysed in this section.

Reflection

Retrospectively, I would organize my data in a more meaningful manner early on and instantly drop variables that I don't need as well as convert those that I want to use later. I also lacked a good plan for how to approach coding and handling the data, as such I had to "learn it on the go" for much of the project.

The dataset was incredibly large and encompassed many interesting variables but ultimately some were simply too vague and difficult to illustrate well, such as the spike between slight and serious/fatal injury – almost 85% of the dataset is populated with slight injury and therefore accident severity as a variable becomes difficult to use in a meaningful manner.

Overall, my methodology was well suited to the problem and in particular unsupervised learning analysis via DBSCAN proved to be the correct choice. It was a critical decision that allowed me to establish a great understanding of what this data represented when visualised (with ample credit to the Folium package for making the visualisation easily accessible).

Appendix.

Appendix 1.

Variable	Brief Description:
Accident_Index	Unique accident identifier
Location_Easting_OSGR	Grid reference system
Location_Northing_OSGR	Grid reference system
Longitude	Geographic coordinate
Latitude	Geographic coordinate
Police_Force	Relevant authority in the area
Accident_Severity	Accident severity
Number_of_Vehicles	Number of vehicles
Number_of_Casualties	Number of casualties
Date	Date
Day_of_Week	Day of the week
Time	Time
Local_Authority_(District)	Local district authority
Local_Authority_(Highway)	Local motorway authority
1st_Road_Class	Road classification
1st_Road_Number	Unique road identifier
Road_Type	Road type
Speed_limit	Road speed limit
Junction_Detail	Type of junction (or vicinity to one)
Junction_Control	Traffic control at junction
2nd_Road_Class	Road classification
2nd_Road_Number	Unique road identifier
Pedestrian_Crossing-Human_Control	Pedestrian crossing with human controller
Pedestrian_Crossing-Physical_Facilities	Pedestrian crossing with automated systems or no human controller
Light_Conditions	Light conditions at time of accident
Weather_Conditions	Weather conditions at time of accident
Road_Surface_Conditions	Road surface conditions at time of accident
Special_Conditions_at_Site	Special conditions at site at time of accident
Carriageway_Hazards	Carriageway hazards at time of accident
Urban_or_Rural_Area	Type of area
Did_Police_Officer_Attend_Scene_of_Accident	Did police officer attend scene of accident
LSOA_of_Accident_Location	Lower Super Output Area (England & Wales only)
Year	Year

Brief Description:
Unique accident identifier
Geographic coordinate
Geographic coordinate
Accident severity
Number of vehicles
Number of casualties
Time
Local motorway authority
Road speed limit
Weather conditions at time of accident
Road surface conditions at time of accident
Type of area
Year

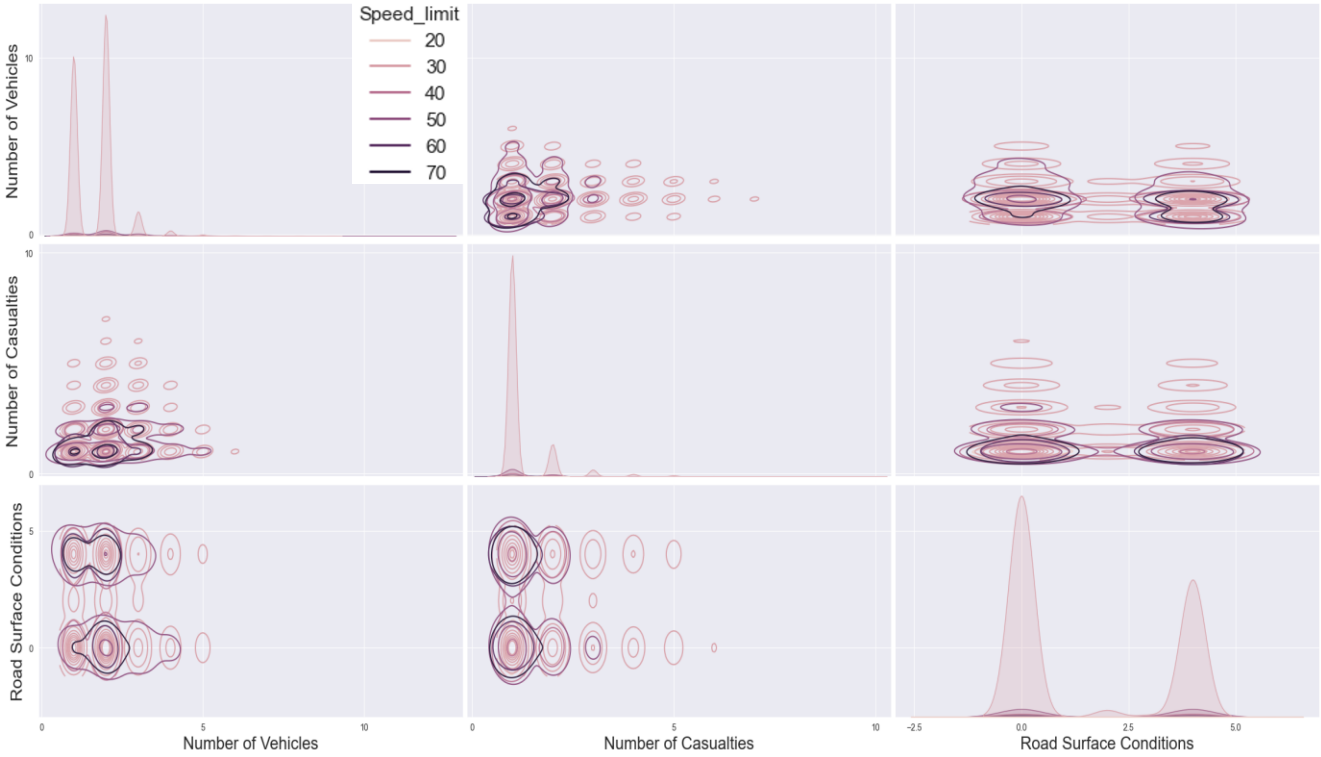
These two tables showcase the variables in the original dataset (left) and the final subset of variables (right) that were used in the analysis throughout this report. Green fields indicating their primary use was as an index.

Appendix 2.

Variable:	Numeric value:	Description:	Occurences:
Weather conditions:	0	Fine with high winds	142
	1	Fine without high winds	9771
	2	Fog or Mist	26
	3	Other	241
	4	Raining with high winds	463
	5	Raining without high winds	2458
	6	Snowing with high winds	13
	7	Snowing without high winds	90
	8	Unknown	151
Road Surface Conditions:	0	Dry	7802
	1	Flood (Over 3cm of water)	14
	2	Frost/Ice	274
	3	Snow	74
	4	Wet/Damp	5191
Accident Severity:	1	Fatal	124
	2	Serious	1809
	3	Slight	11422

This table provides context for the discrete variables in the dataset wherein it was not self-explanatory. It is also shown how often they occurred in the dataset

Appendix 3.



This KDE pairplot showcases the relationship between some of the most important variables in the dataset and distinguishes varying speed limits according to hue.

These notable variables relate to accident severity either directly or indirectly.

Accident severity being (presumably) determined by road surface conditions which in turn (again, presumably) determines the number of casualties and vehicles involved in the accident.

Appendix 4.

Speed Limit	Total Accidents
30	103
40	8
50	4
60	3
70	6

A detailed overview of the total number of fatal accidents per each speed limit is provided in this table.

Environment and Bibliography.

Environment & Packages:

PyCharm 2021.2.3 (Community Edition)

Build #PC-212.5457.59, built on October 19, 2021

Runtime version: 11.0.12+7-b1504.40 amd64

VM: OpenJDK 64-Bit Server VM by JetBrains s.r.o.

Python Version: 3.10.0

Additional packages:

- Folium 0.12.1
- GeoPy 2.2.0
- Matplotlib 3.4.3
- NumPy 1.21.2
- Pandas 1.3.3
- SciKitLearn 1.0
  - DBSCAN
  - AdaBoost
- Seaborn 0.11.2
- XGBoost 1.5.0

References:

Fisher-Hickey, D. (2017) *1.6 million UK traffic accidents | Kaggle*. Available at: https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales (Accessed: November 5, 2021).

HRM Department for Transport (2020) *Reported road casualties Great Britain, annual report: 2020*. Available at: https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2020 (Accessed: October 24, 2021).

Department of Transport (2021) *Road traffic statistics - Download data*. Available at: https://roadtraffic.dft.gov.uk/downloads (Accessed: November 5, 2021).

The National Archives (2021) *Open Government Licence*. Available at: https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/ (Accessed: October 24, 2021).

Transport Scotland (2019) “Reported Road Casualities Scotland.” Available at: http://www.transportscotland.gov.uk/analysis/statistics/publications/reported-road-casualties-scotland- (Accessed: October 24, 2021).