Using Clustering to Locate Accident Hotspots

Introduction

The aim of this project is to use density based clustering (DBSCAN) to identify accident hotspots in central London. In this way, we can identify the most dangerous zones in the city for traffic and make sure the municipality can take action according to our analyzed data. This solution will provide valuable insight to municipality teams to take the necessary action. In conclusion, there will be less property damage and more safety in traffic.

Data

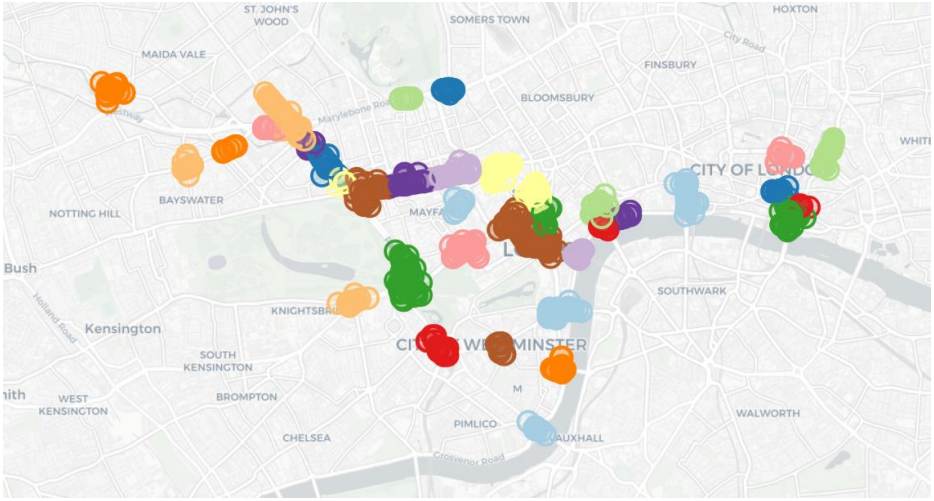
DBSCAN groups points that are closely packed together and marks points outside of these groups as noise. Therefore, using this algorithm, locations in which a high density of accidents take place will be highlighted as clusters. We can then plot the location of these clusters using folium.

Methodology

DBSCAN requires a metric to use when calculating the distance between points. To do this we shall create a function that takes the latitude and longitude of two points and calculates the distance between them in meters. The great_circle function from geopy is used to ensure that the curvature of the Earth is taken into account when calculating these distances.

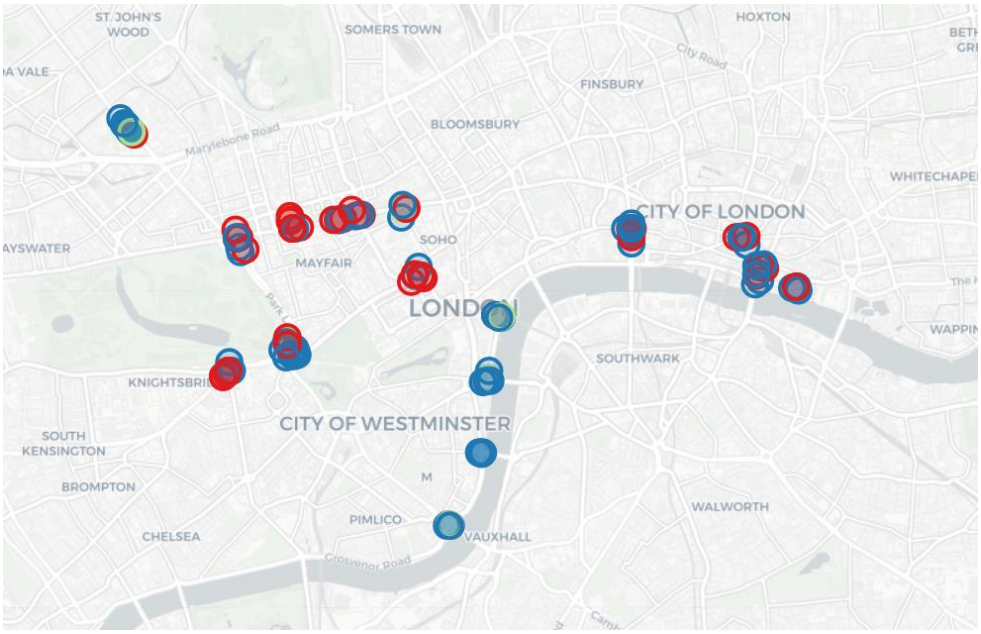
The eps parameter determines the maximum distance between two samples for one to be considered as in the neighborhood of the other. This and the min_samples parameter can be adjusted to change the size and number of points contained in each cluster. Points not located within clusters are given the label -1.

Now that we have found the clusters we can plot them on a map using folium. We can assign colours to each cluster so that they can be more easily identified.



We have found the spacial clusters, but what if we want to find the regions and times that accidents are most likely to occur? To achieve this we need to perform spaciotemporal clustering. This can be done by creating a new distance metric that takes both space and time into account. Used this metric to determine the clusters.

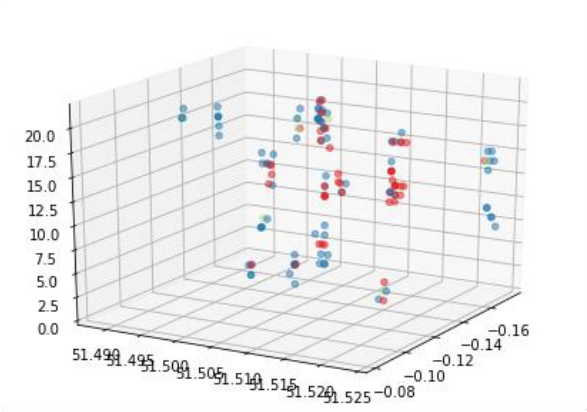
Once again we plot these clusters onto a map using folium. In this case I thought it would be interesting to colour the clusters according to the number of vehicles involved in the accident.



Result

This shows that accidents involving two or more vehicles are more likely to take place at busy junctions such as those near bridges. Conversly, accidents involving one vehicle are generally found in areas containing large numbers of pedestrians such as Oxford Street.

We can try and visualise the hours in which these accidents take place using a space-time cube:



Conclusion

Unfortunately this representation is fairly confusing but it does provide a general idea of the cluster locations in spacetime.

These are just some basic ideas for analysing accident hotspot location, but I hope it provides some inspiration for a more indepth analysis. For example, the DBSCAN parameters in each instance can be adjusted to find clusters of different sizes, and the clusters can be coloured according to different accident conditions.