Dalhousie University

Project Report

CSCI 6612 - Visual Analytics

Visualization and Analysis of Traffic Accidents

Author(s)
Meeta Chanchalani
Mehul Kulshreshtha
Mohamed Muzamil H
Rahul Shewani

Supervisor(s)
Dr. Fernando Paulovich

May 27, 2021

https://git.cs.dal.ca/mmh/va-final-project

DALHOUSIE UNIVERSITY
Faculty of Computer Science

Abstract

The road infrastructure which has oiled the economic engine of the world is also responsible for causing many deaths. Analysing the factors that lead to accidents is a major concern for the authorities looking to minimize the collateral damage. Generally, in any country the Traffic Department, the Health Department and the Police Force work together to achieve this and any visual insight on the contributing factors makes their task easier. In this report, we document the characteristics of graphic tool made over accident data, which is handy for doing an exploratory and explanatory analysis using a variety of chart types and report its findings.

Keywords— Visualisation, Time Series, Heatmap, DBSCAN

1 Introduction

According to the World Health Organization, approximately 1.25 million people suffer from accidents and die each year. This accounts for an average of 3,287 deaths per day. In UK alone, over a period of 2009 and 2010, a staggering total of about 300000 accidents have occurred. In an 8 year period between 2009 and 2017, the number shoots up to 1.3 million.[2]

This project aims to use visualization techniques and machine learning to gain an understanding on the problem at hand, and develop insights and prevention mechanisms for Traffic Accidents and Road Safety. If traffic hazards are pin-pointed accurately, then, an effective mechanism can be devised to prevent their occurrence and minimize damage. Hence, this analysis is of immense utility to various Government Departments/Authorities like Police, Healthcare and Transport for policy formulation.

2 Problem Formulation

The problem under consideration is analysis of factors leading to road accidents to minimize their frequency and severity using UK Road Safety dataset.

2.1 The UK Traffic Dataset

Here, we use U.K Road Safety Data collected over a period of 2005 to 2017. The dataset is published by Department of Transports, under Open Government Licence. The datasets house detailed information related to vehicles, average annual daily flow of traffic and a variety of other factors related to road accidents in UK. All accident data comes from police reports, so this data does not include unreported incidents.

2.2 Analysis Questions

Some of the questions that can be answered using the dataset are as follows:

1. What is the trend of accidents that have occurred between 2005 and 2017?
2. What is the severity of accidents between 2005 and 2017?
3. When do accidents usually happen?
4. Under which circumstances do accidents happen? Is there any correlation between these features?
5. What are the main factors causing an accidents, and can we predict the severity based on these factors?
6. What are the accident-prone areas (hotspots) across UK in this period? [1]

3 Exploratory Data Analysis

3.1 Tabular Analysis

Year	No. of Road Accidents (in thousand)	No. of Road Accident Deaths (in thousand)
2005	198.485	269.388
2009	163.418	220.986
2011	151.362	202.783
2013	138.568	182.742
2015	139.932	185.281
2017	129.866	170.227

Table 1: Comparison of number of accidents and casualties

3.2 Line Plots

3.2.1 Accident Pattern over the Years

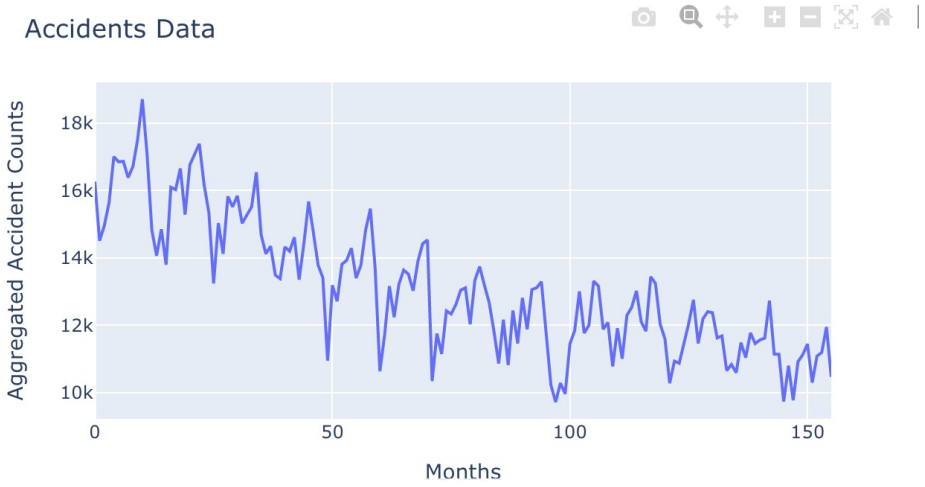

Figure 1: Accident Pattern over the Years

The chart above shows for each month over a period of 2005 to 2017. The time-series plot shows a downtrend with accident count decreasing from about 16000 in the first month of 2005 to about 11000 in the last month of 2017. Another noticeable aspect is the seasonal ups and downs every year.

3.2.2 Casualty Pattern over Months

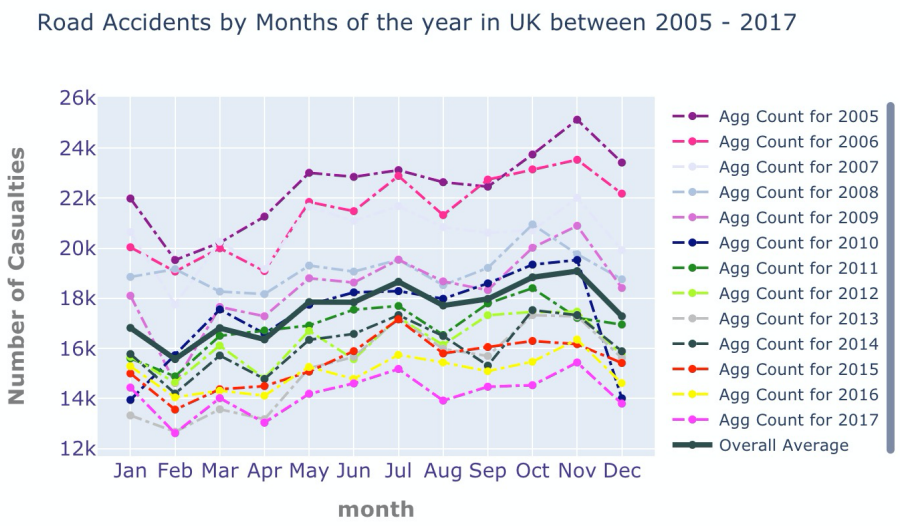

Figure 2: Casualty Pattern over Months

The chart above shows casualty count grouped on monthly basis over a period of 2005 to 2017. The number of deaths in February is minimum. In contrast, November accounts for the maximum fatalities. This is however, followed by a seasonal decrease going into the holidays and New Years. The period of summers reports average number of deaths with slight ups and downs.

3.2.3 Casualty Pattern over Days of the Week

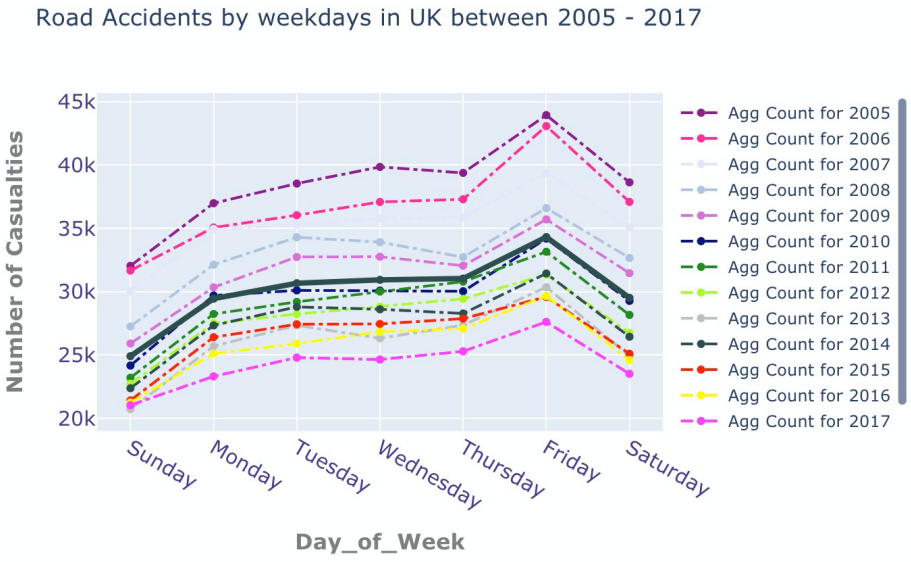

Figure 3: Casualty Pattern over Days of the Week

This graph above shows accidents count grouped by days of the week. We can see a spike in fatalities on Friday. On the other hand, Sunday is least lethal, while other days reported average accidents.

3.2.4 Casualty Pattern over Hour of the Day

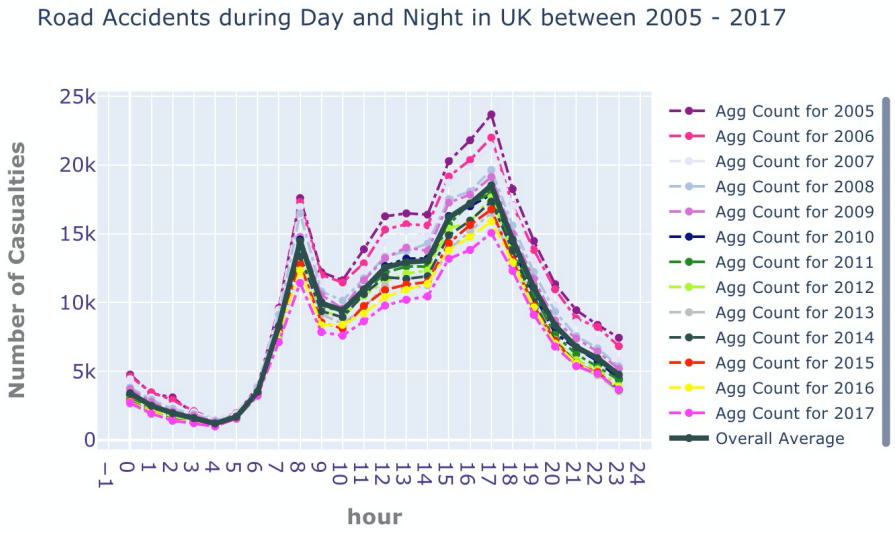

Figure 4: Casualty Pattern over Hour of the Day

The graph above shows fatality count plotted against hour of the day. Its evident that casualties peak in office hours between morning 9 am to 10 am and evening 6 pm to 7pm. In contrast, tragedies are minimum during the night hours indicating a correlation of number of deaths to traffic congestion.

3.2.5 Number of Casualties on different Road Types

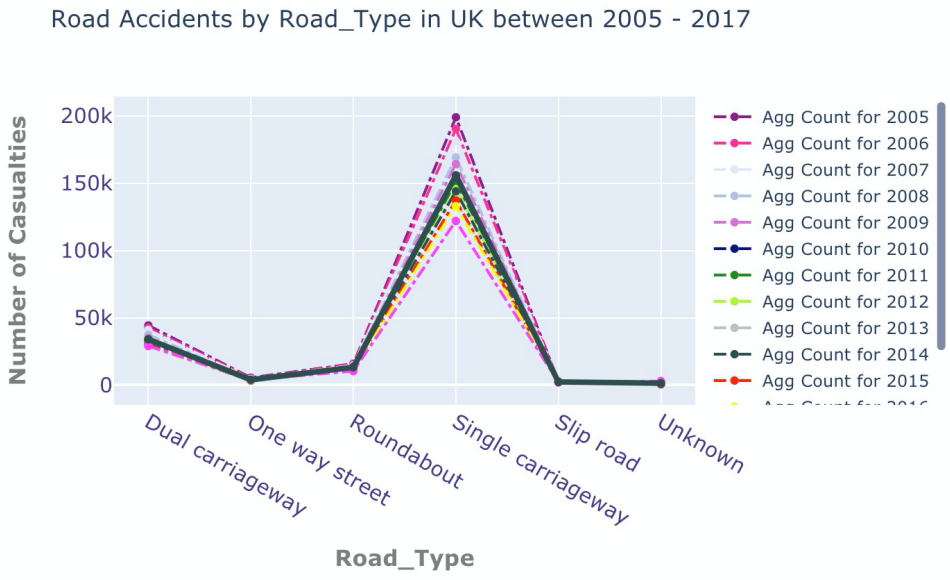

Figure 5: Number of Casualties on different Road Types

The above chart maps deaths to road type. Although its clear that single carriageway roads unfortunately account for most fatalities, followed by roundabouts, as the total number of different road types are

unknown, it does not give us a causal relationship. On the other hand, one way streets and dual carriageway roads seem to be a safer alternative, however, it can not be concluded with certainty.

3.2.6 Number of Casualties by Speed Limit in UK

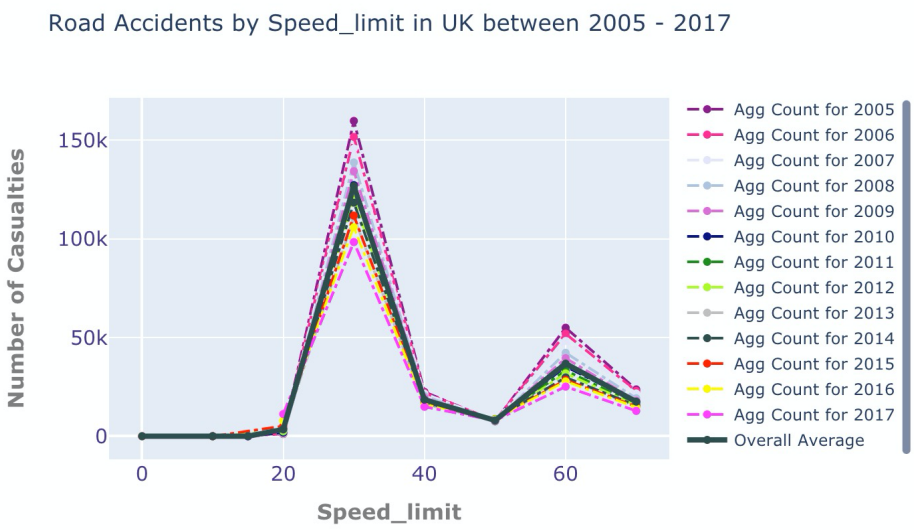

Figure 6: Number of Casualties by Speed Limit in UK

The above chart relates fatalities to roads of a certain speed limit. In a trend which has persisted over the years, deaths peak on roads with speed limit of 30 miles/hour, which are largely inside cities and usually have street lighting, in contrast to expressways with higher speed limits where the reported casualties are about one third of highest.

3.2.7 Number of Casualties by Light Conditions in UK

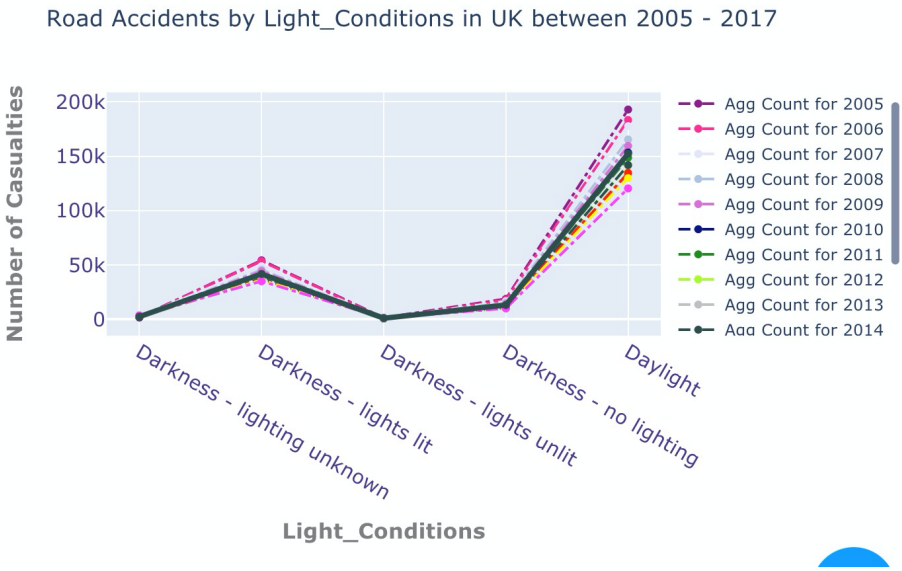

Figure 7: Number of Casualties by Light Conditions in UK

Given any type of light condition, the number of casualties have reduced from 2005 to 2017. Most of the fatal accidents have occurred in broad daylight, while the least in dark conditions. Intuitively, as the number of vehicles in broad daylight are likely to be more, it seems to be the reason of this trend.

3.2.8 Number of Casualties by Weather Conditions

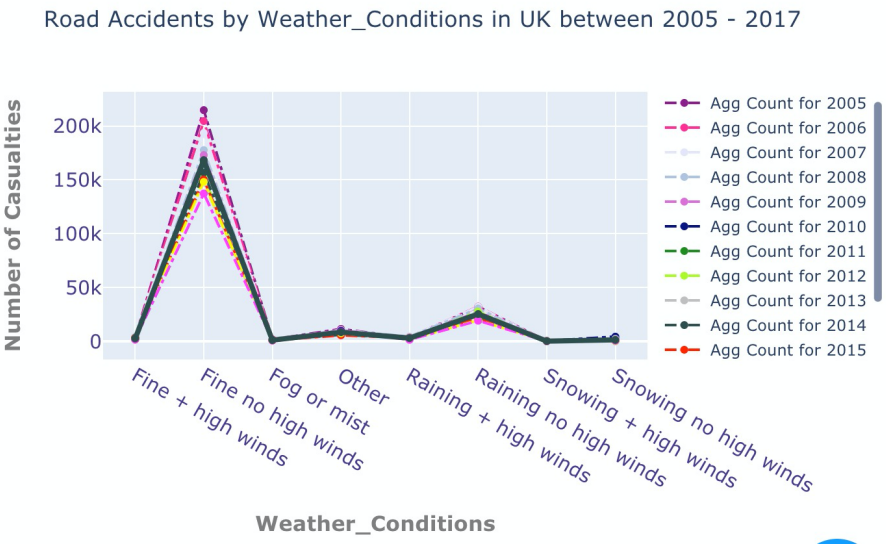

Figure 8: Number of Casualties by Weather Conditions

Given any type of weather condition, the number of fatalities have reduced from 2005 to 2017. Most of the casualties have occurred in fine weather with no winds, while the least in fine weather with high winds. Interestingly, the number of deaths is higher when it was raining and snowing without winds, as opposed to when it was raining and snowing and windy.

3.2.9 Number of Casualties by Road Surface Conditions

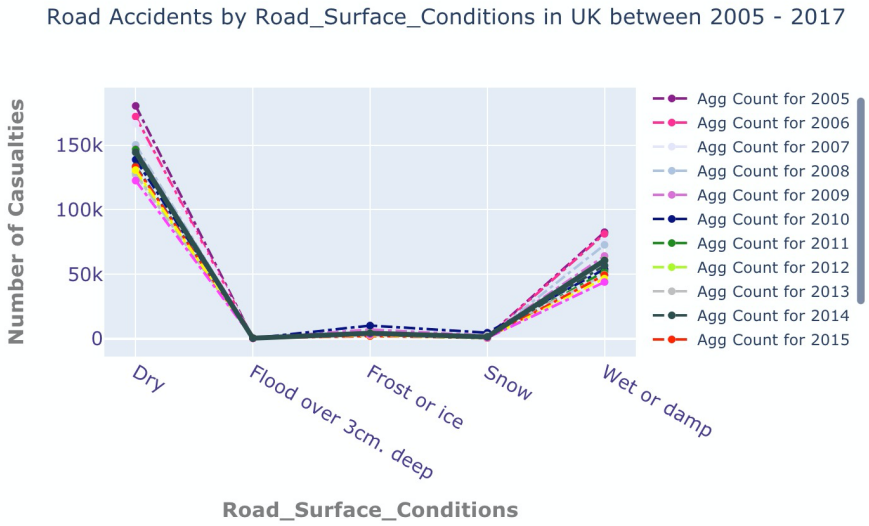

Figure 9: Number of Casualties by Road Surface Conditions

The chart portrays that most of the deaths have occurred on dry road surface. Also, wet or damp roads report significantly more number of deaths as compared to when the roads had snow, frost or ice. When the roads were flooded, the number of casualties due to road accidents in police records were minimum which is in line with intuition.

3.2.10 Number of Casualties by Urban or Rural Areas

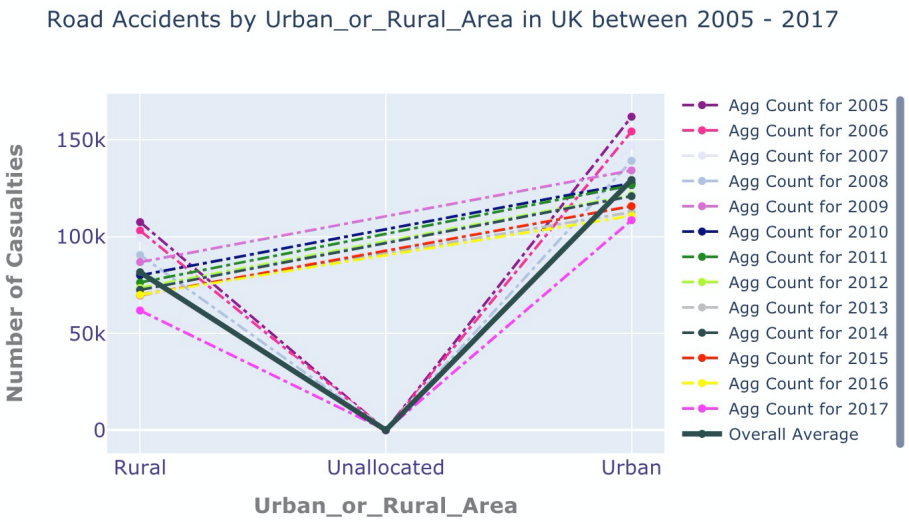

Figure 10: Number of Casualties by Urban or Rural Areas

The above chart relates fatalities to area type. As expected the number of deaths in urban areas due to road accidents are higher as compared to rural areas regardless of the year. Although the number of fatalities reduced as the years progressed, interestingly, the difference in deaths between urban and rural areas remains unchanged. This might hint that whatever has worked to reduce accidents in urban areas has worked equally well for rural areas as well. Although this is not conclusive evidence.

3.3 Tree Map

A treemap is a chart type that displays hierarchical or part-to-whole relationships via rectangles. In case of hierarchical (tree-structured) data these rectangles are nested. The space in the view is divided into rectangles that are sized and ordered by a measure. Nested rectangles mean that hierarchy levels in the data are expressed by larger rectangles (above in the hierarchy) containing smaller ones (below in the hierarchy). [3]

The rectangles in the treemap range in size from the top left corner of the chart to the bottom right corner, with the largest rectangle positioned in the top left corner and the smallest rectangle in the bottom right corner. In case of hierarchical data – when the rectangles are nested -, the same ordering of the lower level rectangles is repeated within each higher level rectangle in the treemap. So the size, and thus the position of a rectangle that contains other rectangles is determined by the sum of the areas of the contained rectangles. [3]

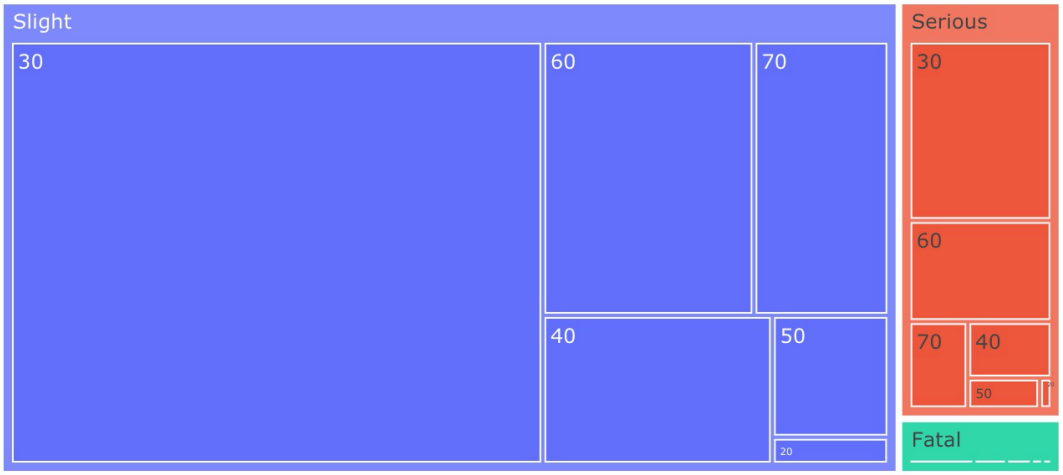

Figure 11: Accident Severity By Speed Limit

The above chart depicts the fraction of roads of varying speed limits for each accident severity type. It can be seen that the roads with a speed limit of 30 miles per hour account for maximum accidents regardless of their severity.

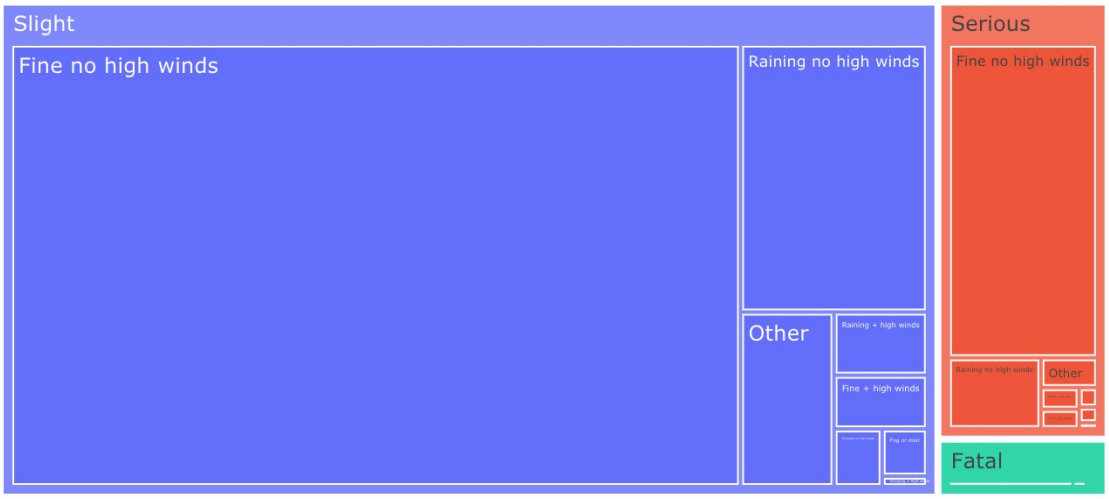

Figure 12: Accident Severity By Weather Conditions

It is also interesting to chart the distribution of weather conditions for each accident severity type. It can be seen that when the weather conditions were fine and it was not windy, the number of accidents peaked regardless of accident severity level.

3.4 Correlation between Columns

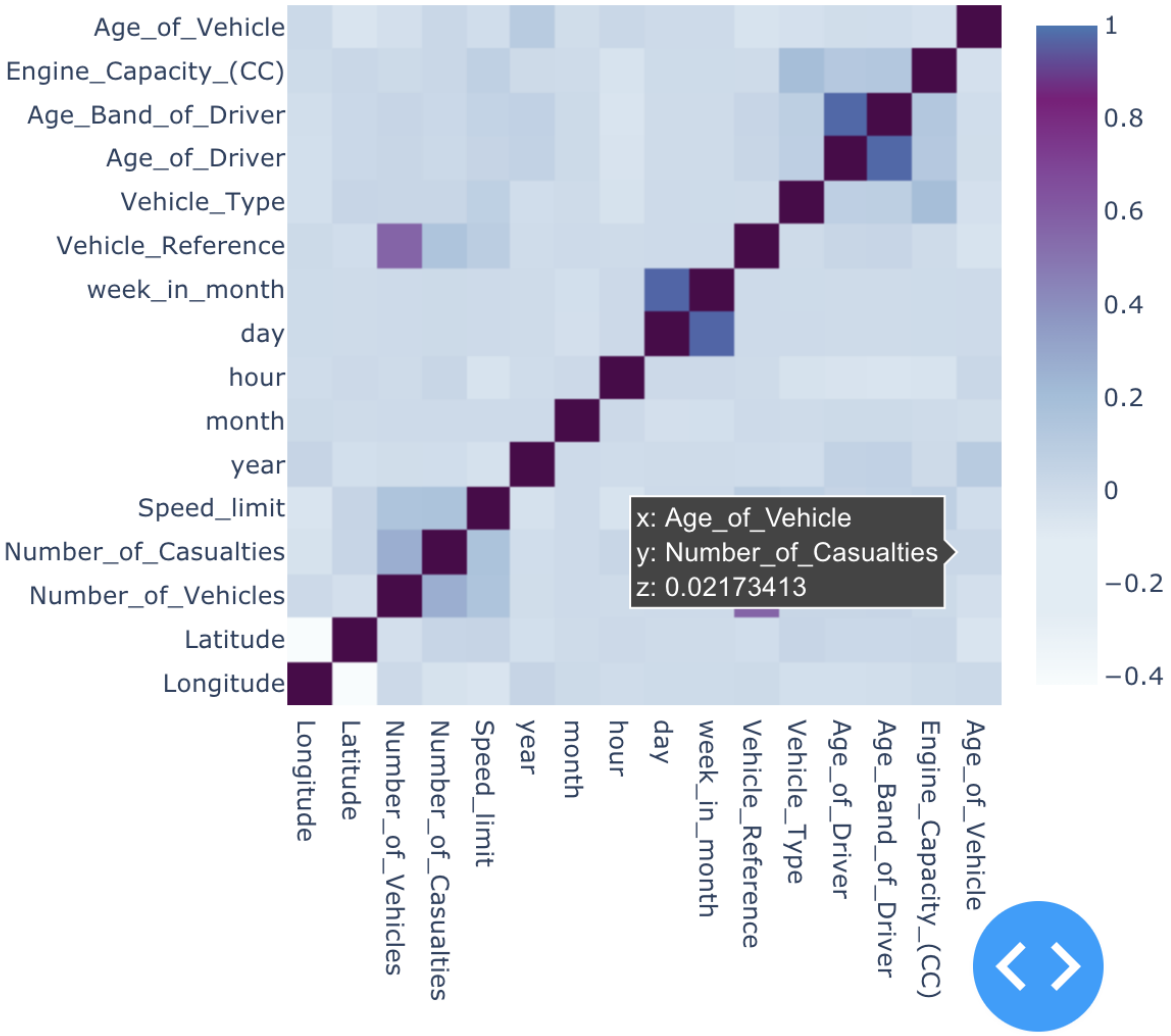

Figure 13: Correlation between Columns

The above figure shows the correlation of various features of the dataset with each other. The darker the spot, more is the correlation. One such interesting correlation is between Age of the Vehicle and Number of Casualties. This indicates that as the vehicle gets older it was prone to more casualties. Another meaningful dark spot is between Speed Limit and Number of Casualties.

4 Time Series Prediction

4.1 Time Series

Time Series is a set of observations on the values that a variable takes at different times. For example Stock Market Price, Weather Forecast etc. In Time Series analysis, this data is used to predict the future values based on the observations obtained previously. [2]

4.2 Components of a Time Series

1. **Trend:** Trend may show the growth or decline in a time series over a long period. This is the type of tendency which continues to persist for a very long period. Prices and export and import data, for example, reflect obviously increasing tendencies over time.

2. **Seasonality:** These are short term movements occurring in data due to seasonal factors. The short term is generally considered as a period in which changes occur in a time series with variations in weather or festivities.

3. **Irregularity:** These are sudden changes occurring in a time series which are unlikely to be repeated. They are components of a time series that cannot be explained by trends, seasonal or cyclic movements. These variations are sometimes called residual or random components.

4. **Cyclic:** These are long term oscillations occurring in a time series. These oscillations are mostly observed in economic data and the periods of such oscillations are generally extended from five to twelve years or more. These oscillations are associated with well-known business cycles. [2]

4.3 ARIMA Model

ARIMA stands for Auto Regressive Integrated Moving Average. There are seasonal and Non-seasonal ARIMA models that can be used for forecasting. An ARIMA model is characterized by 3 terms: p, d, q where p is the order of the AR term, q is the order of the MA term and d is the number of differences required to make the time series stationary. If a time series, has seasonal patterns, then you need to add seasonal terms and it becomes SARIMA, short for 'Seasonal ARIMA'. [2]

4.4 Evaluation of Forecast

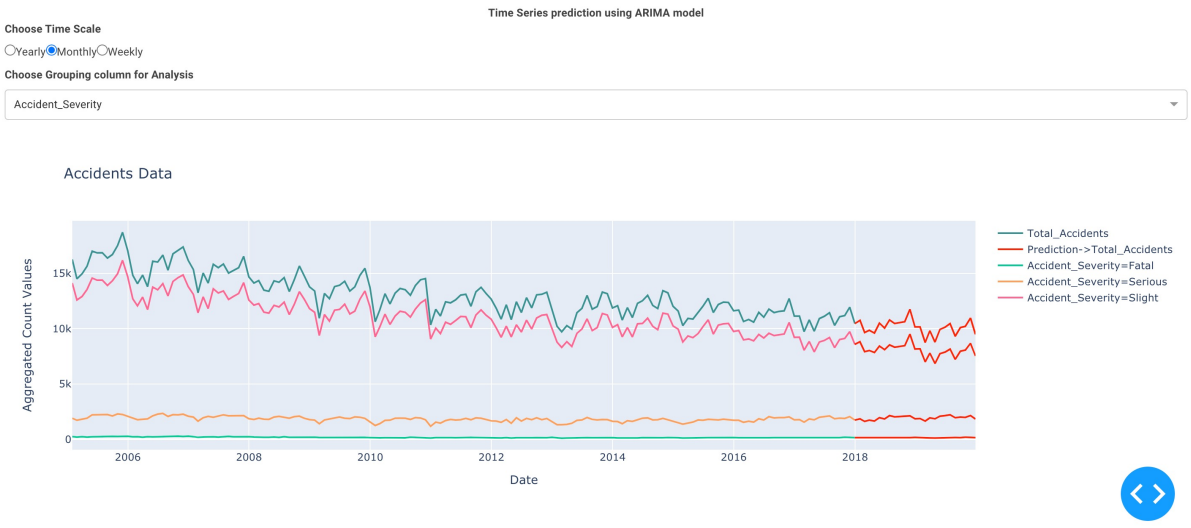

Figure 14: Forecast

5 Clustering

5.1 DBSCAN

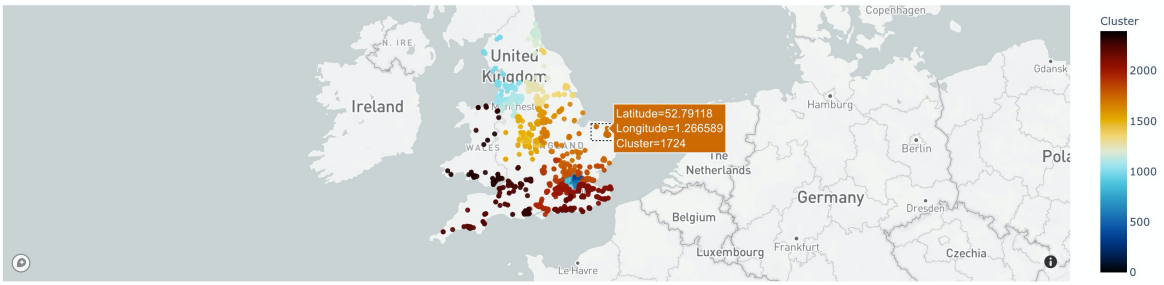

Figure 15: Accident Hotspots

We use density based clustering (DBSCAN) to identify accident hotspots in UK. 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 Plotly plot_map.

DBSCAN requires a metric to use when calculating the distance between points. To do this we shall use ball_tree algorithm which takes the latitude and longitude of two points in radians and calculates the distance between them in meters.

Next, we find the clusters. 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 Plotly plot_map. We can assign colours to each cluster so that they can be more easily identified. [4]

One of the important feature of this visualization is that the user has the option to select the distance between the points and the number of points to be considered while forming clusters. By default the map shows all the accidents without clustering.

5.2 Selection of a cluster and its features

We can box-select some points in the DBScan clustering result graph shown in Figure 15 and there are three sub graphs generated as a result of this selection.

The scatter plot in the below figure is a visualization graph which shows how the Age of the Vehicle(in years), Age of the Driver and the Sex of the Driver are correlated for the points selected in the cluster shown in Figure 15.

The histogram graph on the other hand shows the number of accidents versus the day of the week on which it occurred.

The sunburst graph depicts the visualization of the points selected in the DBScan clustering graph. It depicts attributes such as Sex of the Driver, Speed Limit, Severity of the Accidents, and the day of the week when the accident had occurred.

Figure 16: Selection of a cluster and its features

6 Heatmap

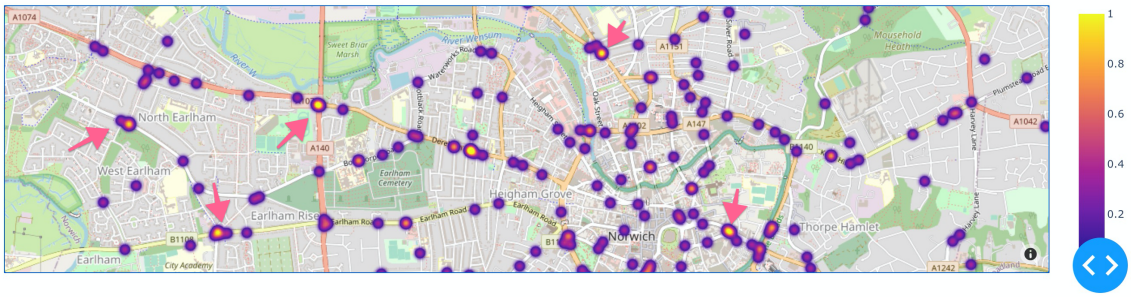

Figure 17: Heatmap

The heatmap in the above diagram is used to show the intensity of accidents in UK. It majorly helps in identifying the traffic hotspots in a particular area. The frequency of accidents in a particular area is directly correlated to the intensity/darkness of the colour shown in the map. The darker the colour depicted in the heatmap the less is the frequency of accidents in that area. In this graph we used the latitude and longitude given in the dataset to plot them on the map. From this map we can observe that majority of the accidents in UK occur at junctions and roundabouts.

7 Suggested Improvements - Implemented

As suggested by professor during the presentation we have made the following changes in the project.

Bar/Line chart

We have updated the chart to show year wise stacked data for a clearer view as shown in figure 18. We have also updated the x axis ticks to reflect Month Names rather than indexes. This has provided the user with a clearer view of the data being represented by the charts.

Figure 18: Bar/Line Chart

Time series

We have updated the time series graph to accommodate the prediction and training data. Now 10 percent of the Actual Data is also being predicted to highlight the accuracy of the prediction model. As seen in the figure 19, the Red line depicts the prediction and we can see a uniform flow in the graph. Also, the predicted and the actual values overlap which shows the accuracy of the model is very high. We have also added Root mean Square error value of the ARIMA Model's Training.

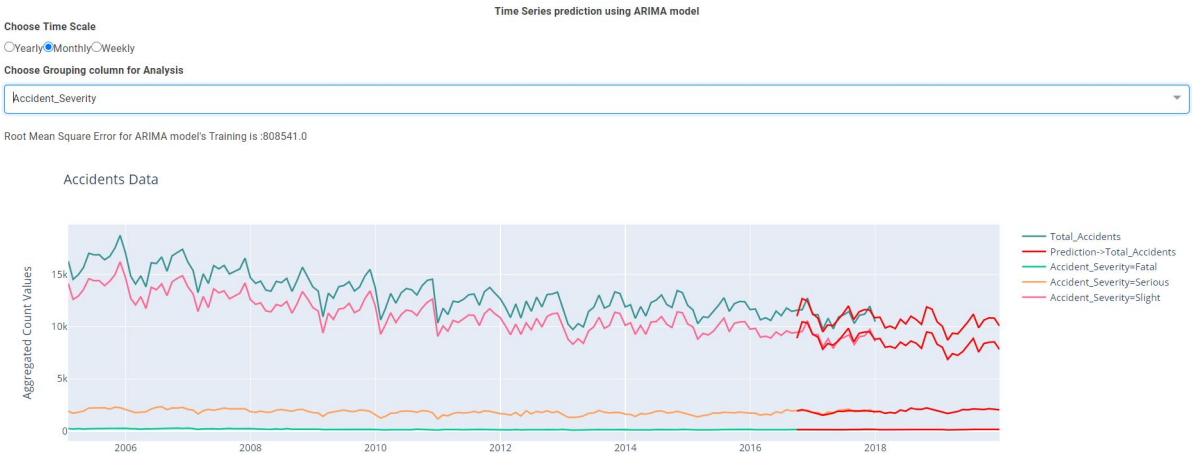

Figure 19: Time Series Prediction

Accident Cluster Map

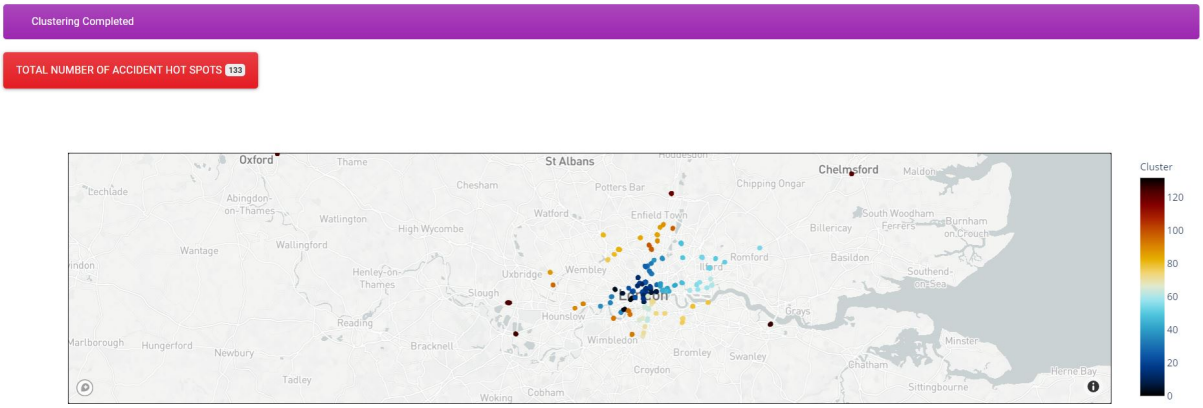

Figure 20: Clustered Map

We have updated the clustered Accident Map to a better color scale. Now the individual Clusters can be identified based on their colors. This has helped users get a clarity on the various clusters present in the given section of the map. The updated graph is shown in the figure 20.

Accident Heat Map

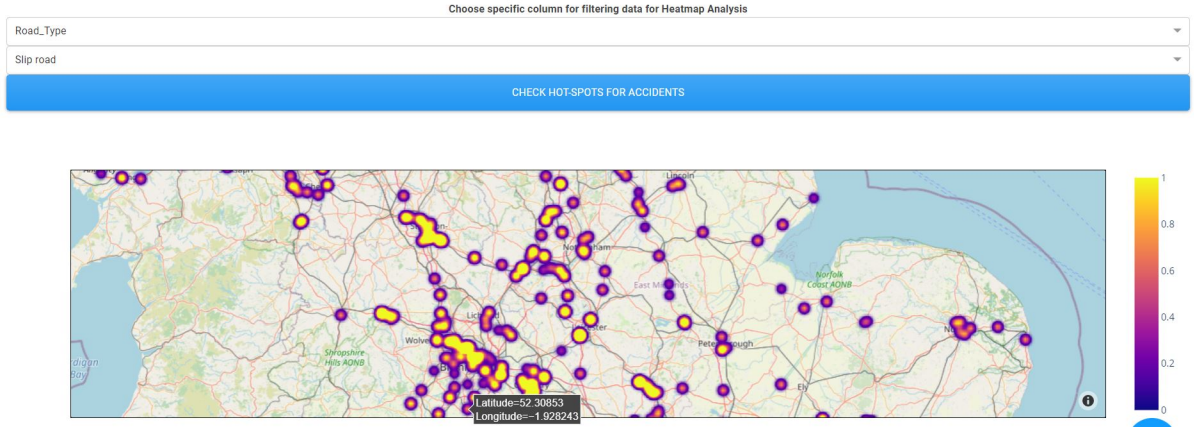

Figure 21: Heat Map

We have updated the Accident Heat Map to include more interactivity. Now the Heat Map has various drop down options to customise the Map as per the user's needs. The user can select various drop downs to investigate a particular area or a specific symptom in the data set. As seen in the updated graph is shown in the figure 21, the user can investigate the accidents occurring due to a specific road type.

8 Final Thoughts

In this journey, we understood the problem through many visualizations, yet leaving so much to uncover. We can summarize our findings as follows:

1. From the line plots, it is clear that most of the accidents occur in broad daylight during office hours in fine weather conditions on highways of 30 mph speed limit which have street light. This indicates that any preventive measure that works well in times of most traffic congestion will have maximum impact on decreasing accidents.
2. As many of the accident hot-spots seem to lie on junctions and roundabouts, the traffic department in coordination with Police can work on measures for enhancing safety for these locations.
3. From the correlation graph we could identify the relation of various features with each other and how factors like Age of Vehicle, Age of Driver, Light Conditions, Weather Conditions and many others play a significant role in the occurrence of accidents.
4. The visualizations helped us to understand the factors affecting severity of the accidents and the number of casualties.

Finally, we really appreciate the UK government efforts to provide this open data in very organized and well-documented format. We would like to explore similar dataset for other countries too!

References

[1] Rawan Almohimeed. *U.K. Traffic Accidents — Data Analysis (10+years)*. URL: https://medium.com/@rawanme/u-k-traffic-accidents-data-analysis-10-years-c81293180ee5. (accessed: 12.06.2020).

[2] Sadegh Jalalian. *Time Series Forecasting For Road Accidents in UK*. URL: https://towardsdatascience.com/time-series-forecasting-for-road-accidents-in-uk-f940e5970988. (accessed: 12.06.2020).

[3] Unknown. *Treemap vs Bar chart – The end of Treemap*. URL: https://www.theinformationlab.co.uk/2014/12/16/treemap-vs-bar-chart-end-treemap/. (accessed: 12.06.2020).

[4] Unknown. *Using clustering to locate accident hotspots*. URL: https://www.kaggle.com/hjnotter/using-clustering-to-locate-accident-hotspots. (accessed: 12.06.2020).