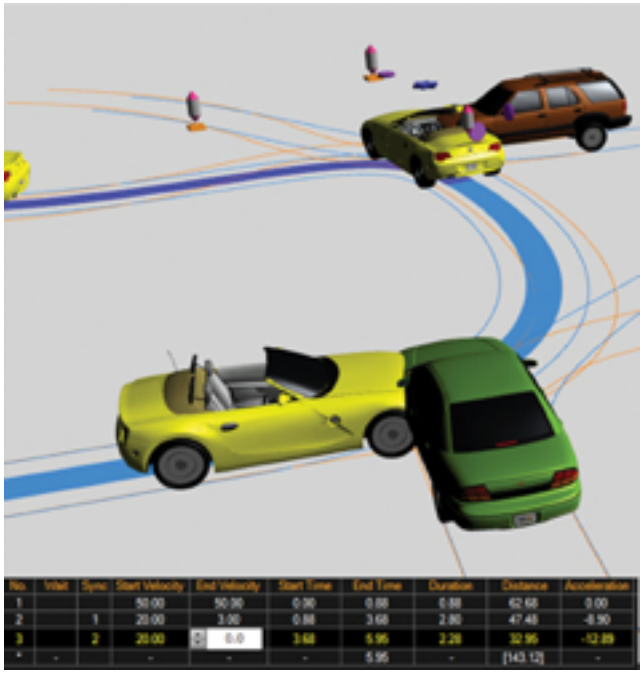
Vehicle Accidents Analysis

Ruchir Vani



Abstract- This article represents the Analysis of Vehicle Accidents in England to find the correlation between various variables associated with accidents. These variables are dependent each-other or some other parameter. I have used the Parallel coordinates with multiple Scatter Plot to find the relations. This article provides output of the chance of the accident with the parameters like date-time, weather conditions and speed limit. This article also represents the detail data of the past accidents.

Index Terms-Accidents Analysis, Weather, Age Band, Date

1 INTRODUCTION

The road is limited but the vehicle population isn't. With the number of vehicles on road increasing exponentially, there is bound to be traffic congestion and even worse- road accidents. With the public transport facilities proving to be inadequate, the increase in number on private vehicles has contributed to the increased traffic. Moreover, with the ever increasing population, traffic situation has spiraled out of control. The following analysis helps us to know more about peak hours for accidents, accident prone areas, what should be the maximum permissible speed, no. of vehicles on road at a time etc. This analysis might be useful for the safety of the personal. Accident is depending on the drivers as well as the physical environment. Accidents are dependent on many parameters like weather conditions, road surface, junction, speed limit of road, gender of the driver, age of the driver, time of the day etc. Currently most system provides the data like number of accidents at this weather conditions, or number of accidents at the time of day etc. I tried to connect this all variables together to find the relation among them. Plotting

These parameters give better view rather than just looking the data.

2 EXPOSITION

I got data from the internet website for the accidents of the England. I have used mysql and java to clean the dataset. I have settled -1 for the NULL values in my database. I have used R and weave for analyze the data. You can have analysis of accidents from plotting the histogram of each variable depends on it. Here you can compare the variables but cannot find the dependency of each-other. By using multiple scatter plot and parallel coordinates you can compare it.

Histogram contains the variable values with different frequency of it. Fig.2. shows the histograms for the various variables. You can have idea of the number of accident related with the particular variables, like histogram for the accidents per hour, accidents per day of week, accident with particular weather conditions. But this information is not enough for analyzing the data. I have plotted scatter plot for the comparing 2-3 variables. For more than 3 variables parallel coordinates (Fig.3.) is good choice, as it is lossless visualization. Though this visualization is not good with large number of tuples I have divided My data into 4 seasons with 3 months each.

Parallel coordinates are good visualization to analyze the data but it is not enough we need to combine this to scatterplot to better understand the dataset. We can also find the correlation among the attributes.

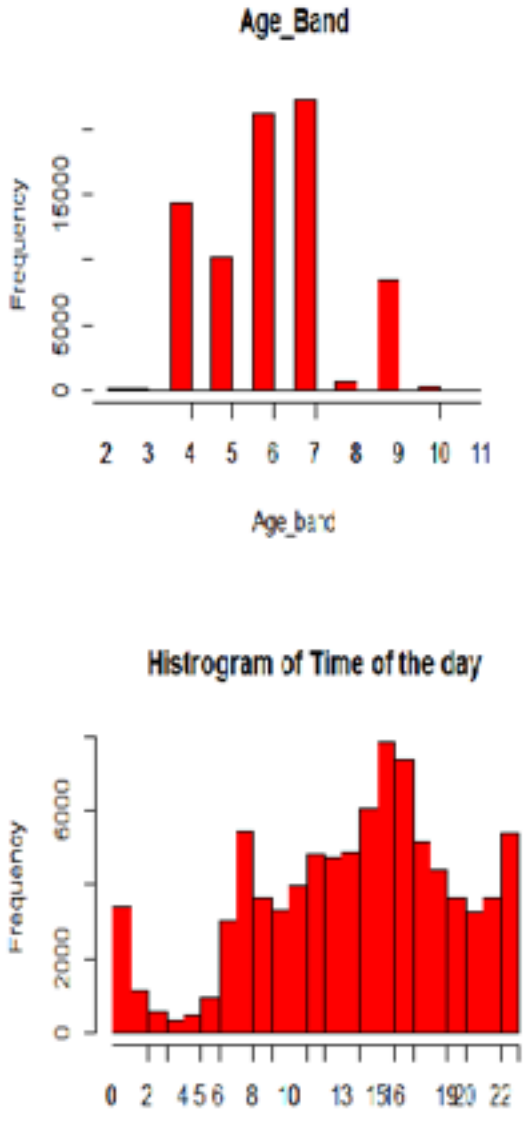


Fig.2. Histogram of Age Band, Speed limit, Time of the day and weather conditions.

Fig.2. give conclusion that age with the age of 36-45 has maximum number of accidents; speed limit 30mph has maximum accidents. Results were as expected, but in case of the time of the day evening has large number of accidents compare to morning. It might be because people feel fresh when they going office and feel tired when coming back to home.

There is lots of parameter associated with accidents. Here in Fig . 3. I have tried to attach weather conditions, type of road, detail of the junction where the accident had occurred and condition of the road surface. Than you can see that at the correlation between variables. Here you can observe that in particular weather conditions probability of accident at some road type and some junctions are high compare to others. If the connecting link is dark than probability is high.

Combination of scatterplots and parallel coordinate with interaction provide you better idea for the outliers. Fig. 4. shows that there is only one accident with 9 vehicles associate with it , you can look at the other parameters associated with it like weather conditions, junction details etc.

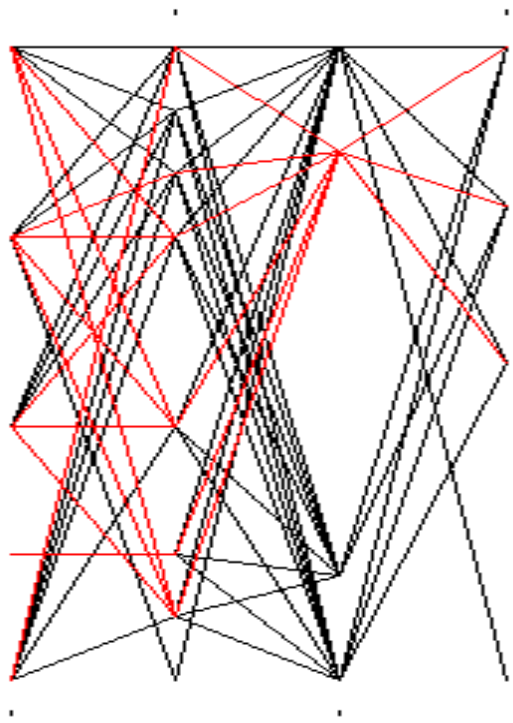


Fig. 3. Parallel coordinates for environment variables.

Scatterplot is generally used for finding the relation between 2 or 3 variables. In scatterplot 2 variables are plotted on axes and other variables can be project using size, color and shape. But this other parameter might not work with wide range of values. I have plotted multiple scatter plot for finding the relations among all other variables. Fig .5. shows some of the scatterplots.

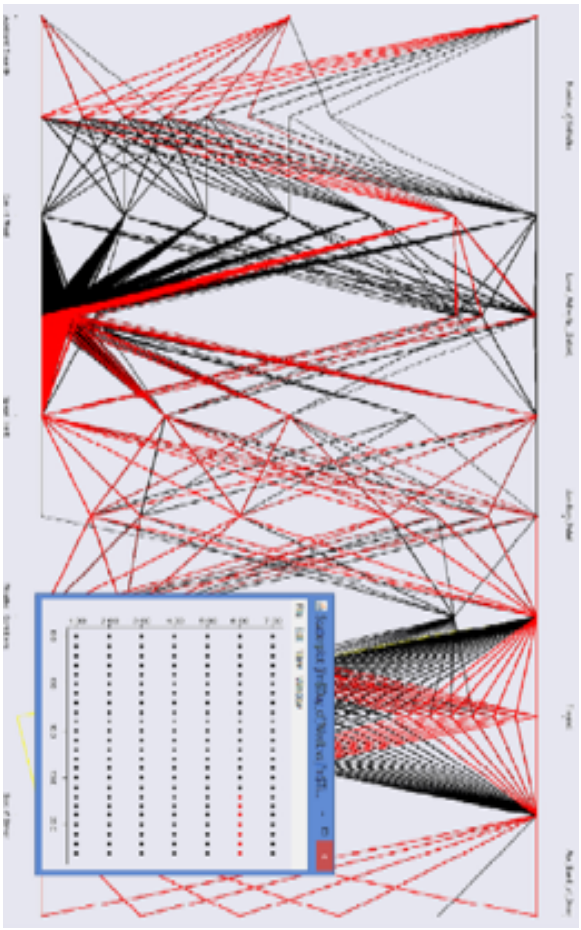


Fig .4. Multiple plots with brushing.

From the various scatterplot I found some of the results which are like this, Accidents in snow consist of less number of vehicles. In fog highway consist of more accidents. When you are looking at the accident with accidents severity , Saturday night has the max number of severe accidents compare to others. Men are having more severe accidents compare to women.

I have try to classify the data using conjunctive rule using Weka. I found this rule for Speed limit with other attributes. Which is as follows: (Junction Detail > 0.5) and (Local_Authority__District_ <= 32.5) and (> 10.5) => Speed_limit = 44.811321. This indicates that if speed limit is around 44.81 than probability will be high in the region of Authority dist 10.5 to 32.5. There is also class for age band of the driver. Which is as follows: (Vehicle_Manoeuvre > 0.5) and (Vehicle Type > 96.5) and (> 27.5) and (Local_Authority__District_ <= 31.5) and (month <= 21.5) and (Junction_Detail <= 8.5) and (Vehicle Type > 106.5) and (Number_of_Vehicles <= 4.5) and (month > 3.5) and (> 123.5) and (Accident Severity > 1.5) => Age_Band_of_Driver = 5.941463. I have also try to find clusters among the database using the clustering algorithms

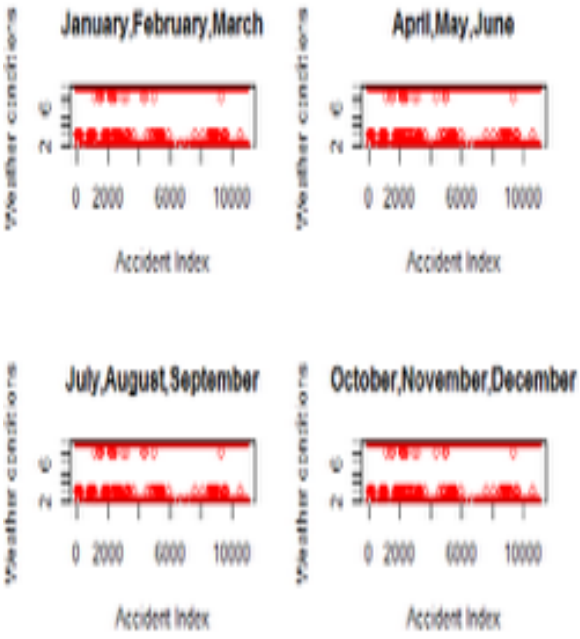


Fig .5. Multiple scatterplots.

Heatmap is also used for the understanding the pattern among the data. It doesn't provide you the exact information but it gives the trend of the data. Fig .6. shows the heatmap for the data with speed-limit, day of week, age band of the driver, sex of the driver and weather conditions. Here you can see that in bad weather conditions male has more accidents compared to women.

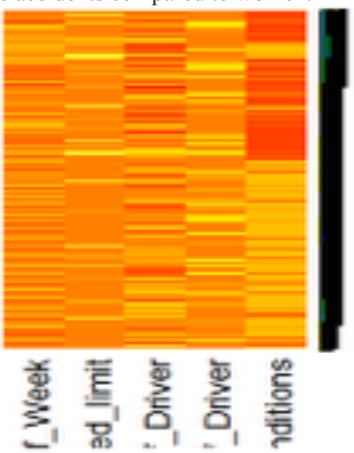


Fig .6. Heat-map for the accident analysis dat

3 CONCLUSION

It can be understood that accident depends on the number of parameters. The weighted of the parameter may vary from conditions. Like in weather conditions like fog the vision is not clear so will have more accidents; Gender of driver is male have more accidents with more severity compare to women. Using interactive visualizations you can analyze every accident for particular period than you can compare it with others.

REFERANCES

[1]Dataset from the following website: http://www.transtats.bts.gov/

[2] Matthew o. ward, Georges Ginstein and Danie Keim. Interactive Data Visualization for concept of parallel coordinates .

[3]Help in the Visualization and analysis tool R: http://www.statmethods.net/