Analysis of “UK Road Safety: Traffic Accidents and Vehicles” dataset

Submitted By:
Andrey Feygelman

Supervised By:
Muhammad Fahim

Summer Internship Report 2019

Table of Contents

Abstract

This paper made for learning purposes. This is my first dataset exploration and work with big data in general. It contains an analysis of dataset of road accidents and involved vehicles in the UK (2005-2017). In this work, were used association (Cramér's V and Theil's U), clustering (k-modes) algorithms and was made simple analysis.

Introduction

Recently I have been interested in data analysis and applied for the internship "Data analysis using machine learning algorithms". For my first dataset examination, I decided to go with "UK Road Safety: Traffic Accidents and Vehicles" dataset from the kaggle to gain some experience.

The UK government collects and publishes (usually on an annual basis) detailed information about traffic accidents across the country. This information includes, but is not limited to, geographical locations, weather conditions, type of vehicles, number of casualties and vehicle maneuvers.

This dataset consists of two csv files (dataframes): "accident Information" and "vehicle information" with date ranges 2005-2017 and 2004-2016 respectively.

Related Work

Since this dataset was published on kaggle a year ago before this research, there were some works already done on it. They include different data visualization, preprocessing and predicting number of casualties.

They turn out to be useful for my own work. For example, one kernel changes datatype of "Date" column from a string to "Date" type and introduce "Hour" and "Daytime" from old "Time" column, which is repeated in this research.

Methodology

0) Preprocessing

First, changed datatype from a string to categorical for every column that has not so many unique values (<600) in both dataframes

Second, made changes in two columns in "accident Information" dataframe:

- Change 0.0 to NaN in "2nd_Road_Number" and
- Change "1","2","3" codes into real meaning in "Did_Police_Officer_Attend_Scene_of_Accident"

Third, introduced new column "Hour of accident" in "accident Information" dataframe based on "Time" column

1) Simple analysis

This part contains countplots (bar chart) and histograms for some interesting columns.

1.5) Preprocessing

Both dataframes are merged into single one via “Accident Index” column. Observations with accident Index appearing only in one dataframe are dropped (dataframes have different date range). Each row in merged dataframe represents a vehicle in accident, but not an accident itself. If there is a several driver involved in accident, then general information about this accident is repeated for each of them.

For next parts used data only from 2016 year

2) Association matrices

Association is simply speaking “correlation in the world of categorical data”

In this part used two algorithms: Cramér’s V and Theil’s U. Both of them output number in the range of [0,1], where 0 means no association and 1 is full association. To run them, I transform every column to type categorical that possible and drop the others.

Cramér’s V is symmetrical (V(a,b)=V(b,a)) and based on a nominal variation of Pearson’s Chi-Square Test

	col1	col2
0	A	c
1	A	d
2	A	c
3	B	f
4	B	g
5	B	f
6	C	h

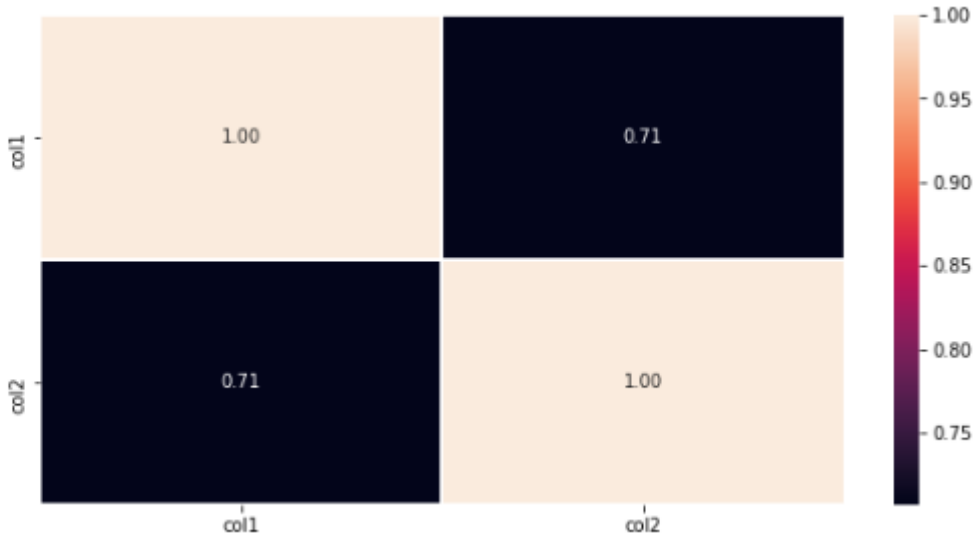


Fig 1. 7x2 toy dataset and its Cramér’s V association matrix. Values on main diagonal are always 1.00.

Theil’s U is asymmetrical (U(a,b) ≠ U(b,a)) and based on conditional entropy between x and y. Putting it simply: “given the value of x, how many possible states does y have, and how often do they occur”.

	col1	col2
0	A	c
1	A	d
2	A	c
3	B	f
4	B	g
5	B	f
6	C	h

Fig 2. same 7x2 toy dataset and its Theil's U association matrix. For each value of "col2" there is only one value in "col1 ", but not vice versa. In other words, knowing "col2", you can predict "col1"with 100% accuracy, not vice versa. Values on main diagonal are always 1.00.

3) Clustering

This part uses k-modes algorithm, which is similar to k-means one but for categorical data.

Algorithms work like this: Assume we want to separate data with N numerical columns into k clusters using k-means. Then K-means will represent each observation as point in N-dimensional space. Then it will try to find k points called "centroids", such that total distance from each observation to the closest centroid was minimal. Only difference between k-modes and k-means is that k-modes instead of Euclidean distances uses dissimilarities (quantification of the total mismatches between two objects: the smaller this number, the closer these two objects)

Similarly, to association matrices I transform every column to type categorical that possible and drop the others. Then I run k-modes with 4 clusters to identify most important half of columns via Cramér's V and drop others. After that, it is a good idea to find suitable number of clusters, so I calculated and plot cost of K-modes over number of clusters.

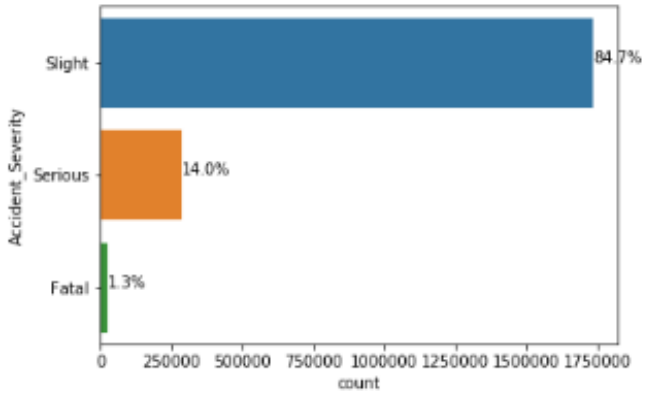
Analysis and discussion

1) Simple analysis

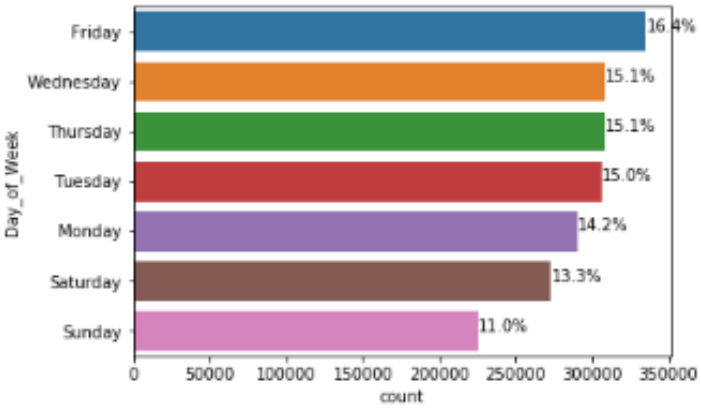
Even simplest analysis can give some interesting result:

- In 85% of accidents people had only slight injures
- The least amount of accidents happened on Sunday: 11%, most on Friday: 16% (1/7=14%)
- 80% of accidents happened in fine weather without high winds
- 46% of the drivers get in accident when tried to go ahead of other vehicle
- 49% of hits was on front part of a vehicle (column "X1st_Point_of_Impact", Fig 9)

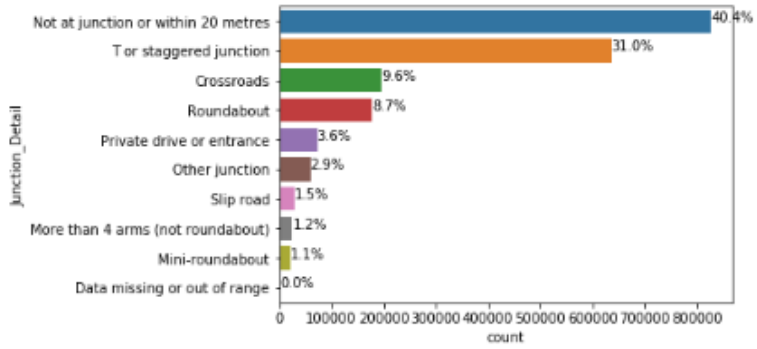
Accident_Severity
number of nulls is 0.0%



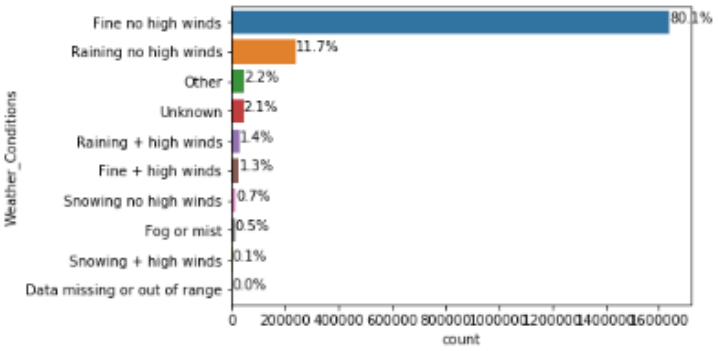
Day_of_Week
number of nulls is 0.0%



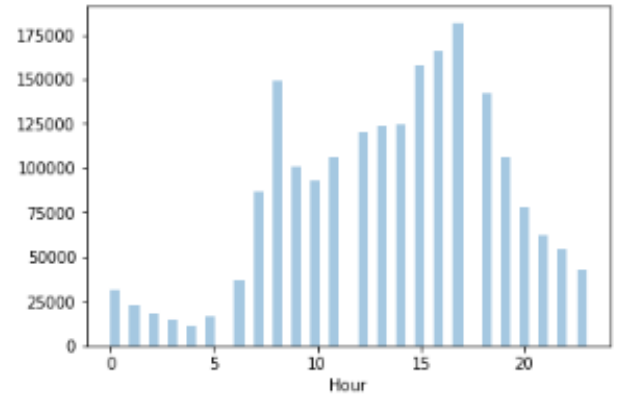
Junction_Detail
number of nulls is 0.0%



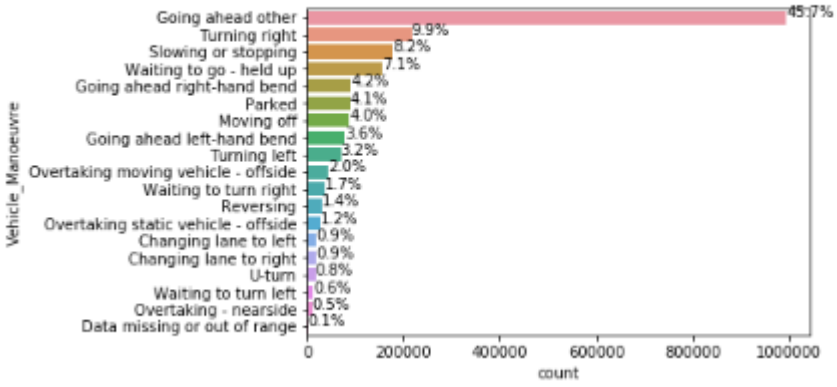
Weather_Conditions
number of nulls is 0.0%



Hour
number of nulls is 0.0%



Vehicle_Manoeuvre
number of nulls is 0.0%



X1st_Point_of_Impact
number of nulls is 0.0%

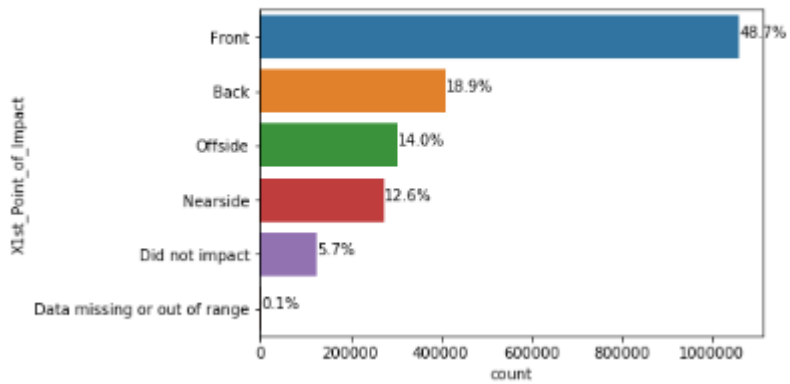


Fig 2-9. Countplots on different columns

2) Association matrices

Comparing two matrices (Fig. 10& 11), we can see some differences. For example, Theil's U's matrix is less noisy and have many zeroes, when in Cramer's V's one does not have zeroes at all. It makes Theil's U's matrix somewhat easier to analyze. Despite this, they have some similarities, high/low values in one matrix usually correspond to high/low values in another one. Most of the high values in both matrices have more or less obvious cause; however, lack of association can give information too. For example, separately none of parameters presented here heavily associated with number of casualties and vehicles per single accident.

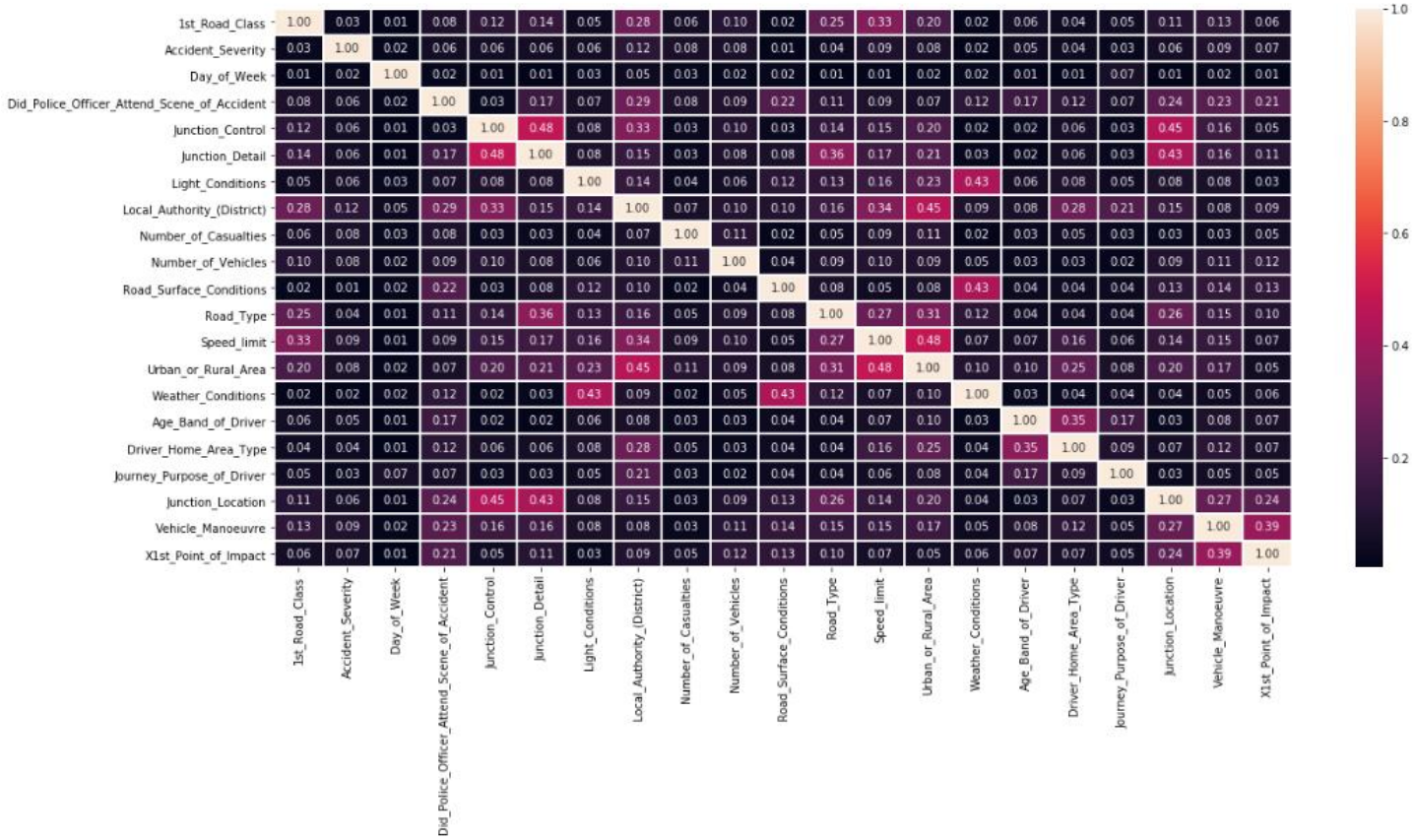


Fig 10. Matrix from Cramer's V algorithm

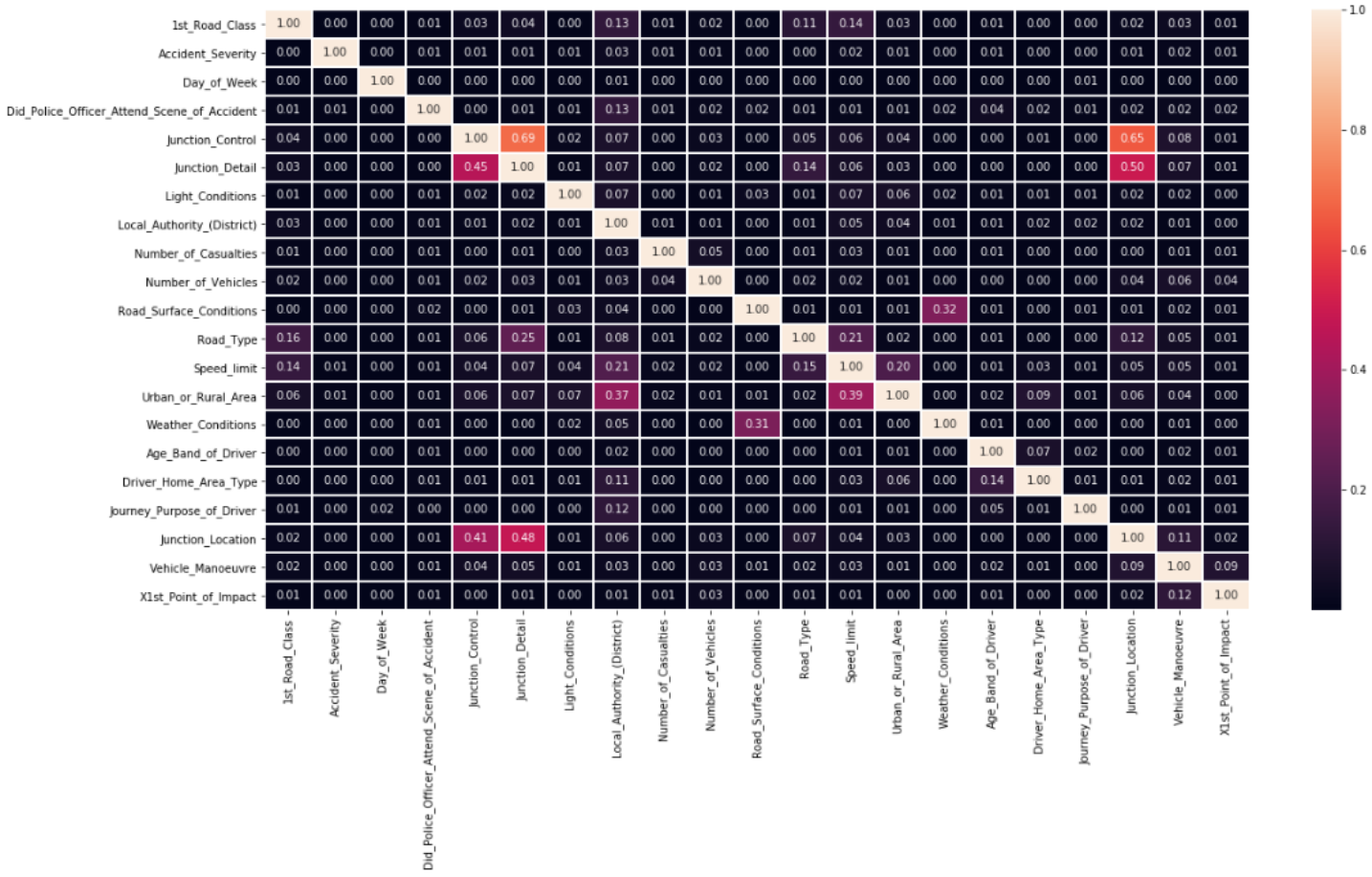


Fig 11. Matrix from Theil's U algorithm

3) Clustering

After running K-modes with 4 clusters, importance of columns for clustering looks like this:

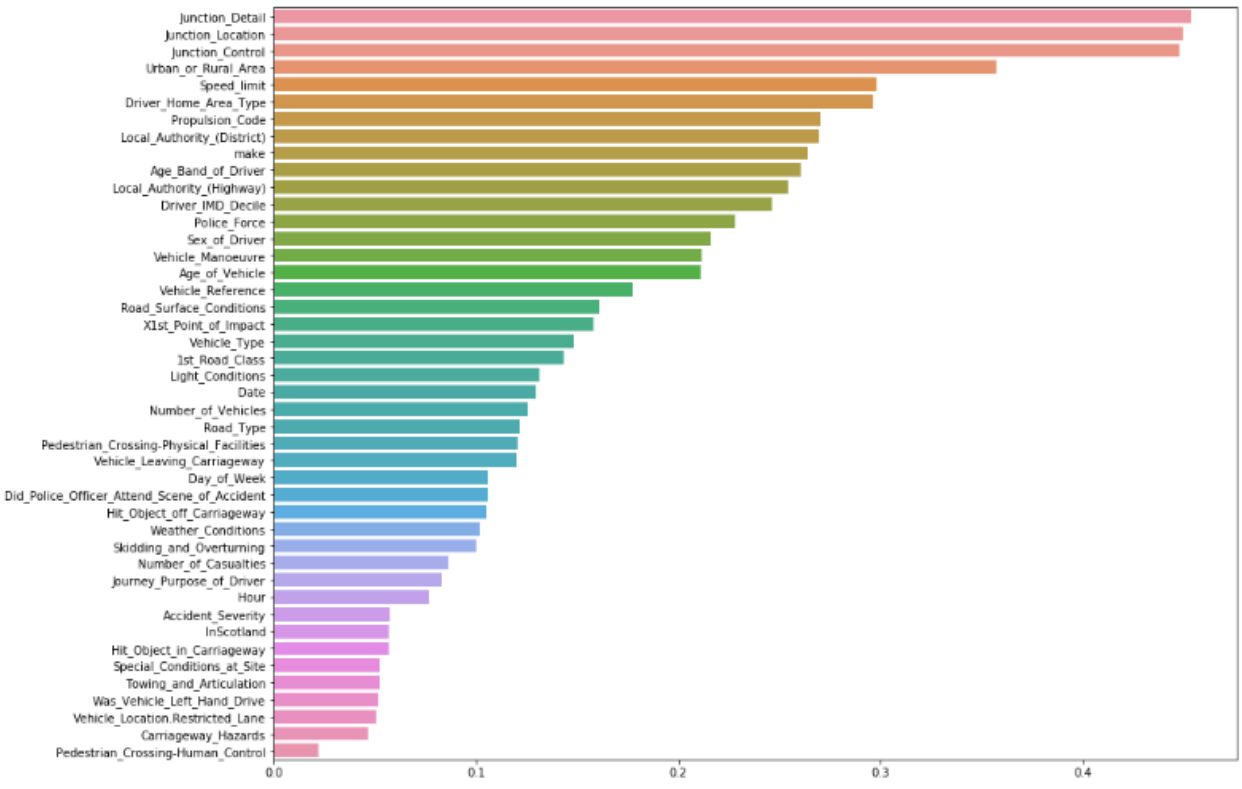


Fig 12. Importance of columns before dropping half of columns obtained by Cramer's V algorithm

After dropping half of columns starting from 'Vehicle Reference' and running K-modes with 4 clusters again it changes to this:

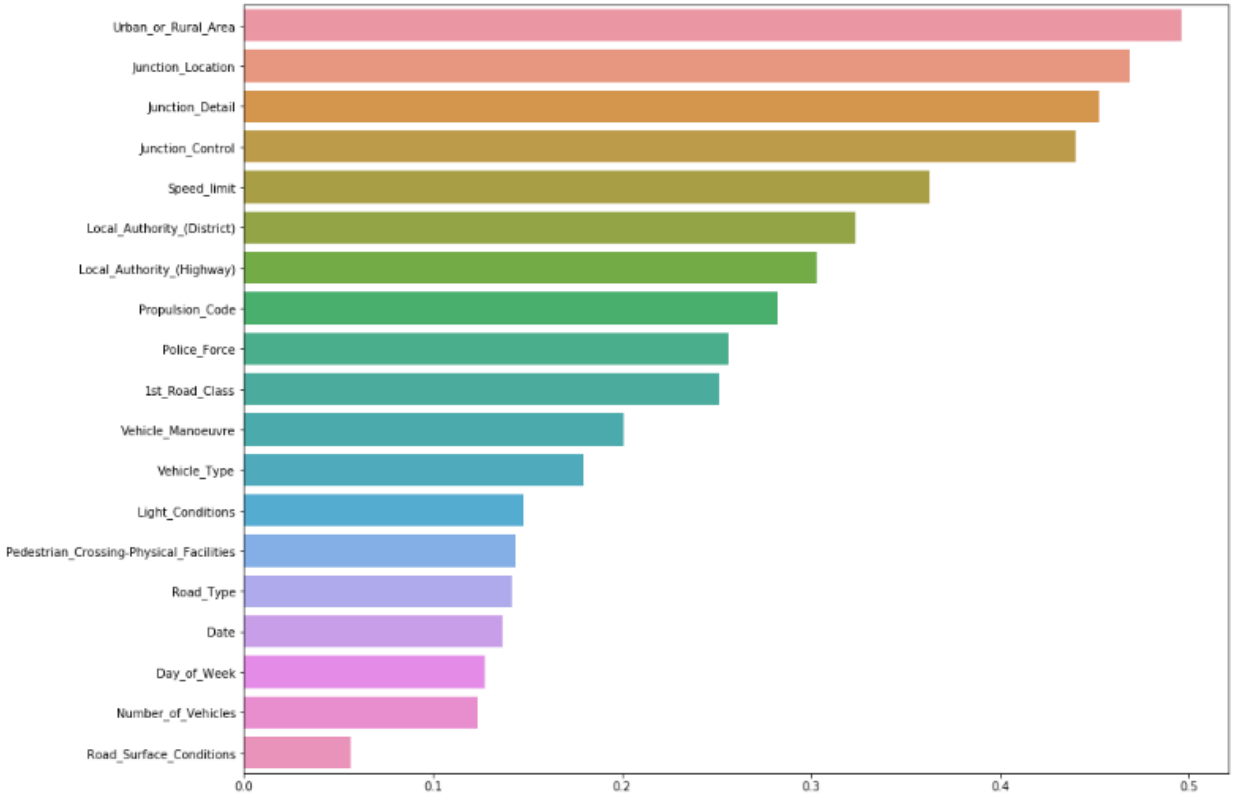


Fig 13. Importance of columns after dropping half of columns obtained by Cramer's V algorithm

Next plot represents K-modes cost (that total distance from each observation to the closest centroid) over the number of clusters. This is useful to find appropriate number of clusters. Here we see interesting pattern: we have local minimums at 5,9,13. It is 4+1, 8+1, 12+1.

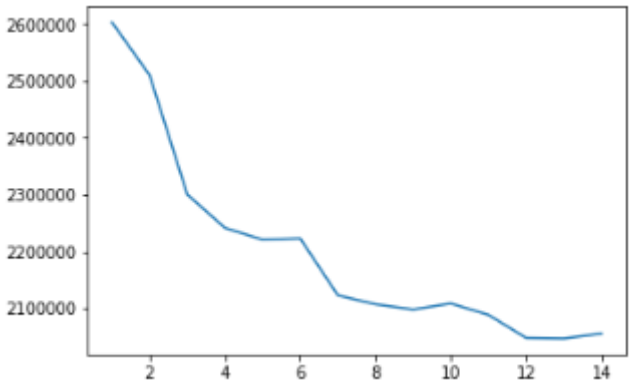


Fig 14. K-modes cost over the number of clusters

Run K-modes with 5 clusters to try to understand what they represent.

For this, we can check centroids

Reference: 30 mph =48 km/h; 70 mph =113 km/h

	1	2	3	4	5
Urban_or_Rural_Area	Urban	Urban	Urban	Urban	Rural
Junction_Location	Approaching junction or waiting/parked at junc...	Not at or within 20 metres of junction	Not at or within 20 metres of junction	Mid Junction - on roundabout or on main road	Not at or within 20 metres of junction
Junction_Detail	T or staggered junction	Not at junction or within 20 metres	Not at junction or within 20 metres	Roundabout	Not at junction or within 20 metres
Junction_Control	Give way or uncontrolled	Data missing or out of range	Data missing or out of range	Give way or uncontrolled	Data missing or out of range
Speed_limit	30.0	30.0	30.0	30.0	70.0
Local_Authority_(District)	Birmingham	Westminster	Birmingham	Leeds	Stafford
Local_Authority_(Highway)	Birmingham	Surrey	Kent	Leeds	Staffordshire
Propulsion_Code	Heavy oil	Petrol	Petrol	Petrol	Heavy oil
Police_Force	Metropolitan Police	Metropolitan Police	Metropolitan Police	Metropolitan Police	Staffordshire
1st_Road_Class	Unclassified	A	A	A	Motorway
Vehicle_Manoeuvre	Going ahead other	Going ahead other	Going ahead other	Going ahead other	Going ahead other
Vehicle_Type	Car	Car	Car	Car	Car
Light_Conditions	Daylight	Daylight	Daylight	Daylight	Daylight
Pedestrian_Crossing-Physical_Facilities	0.0	0.0	0.0	0.0	0.0
Road_Type	Single carriageway	Single carriageway	Single carriageway	Roundabout	Dual carriageway
Date	2016-11-30	2016-11-29	2016-10-15	2016-07-25	2016-08-19
Day_of_Week	Wednesday	Tuesday	Saturday	Monday	Friday
Number_of_Vehicles	2	2	1	2	2
Road_Surface_Conditions	Dry	Dry	Dry	Dry	Dry

Fig 15. Centroids

Check how many elements of each cluster there are

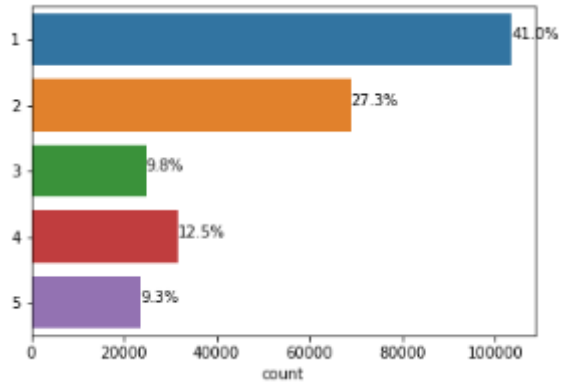
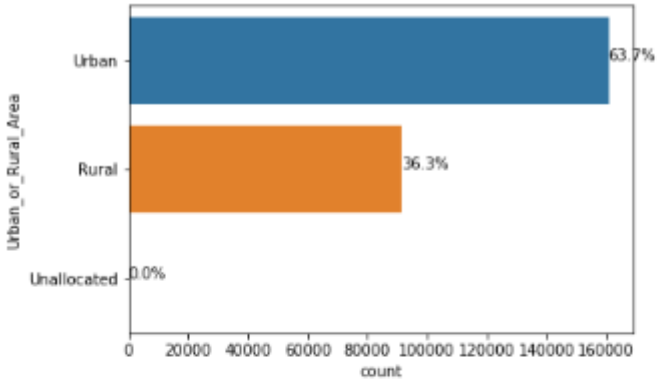
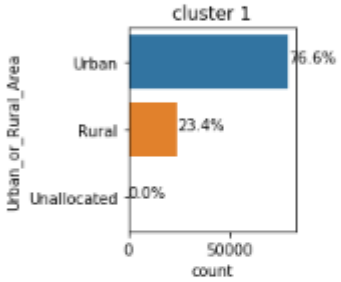


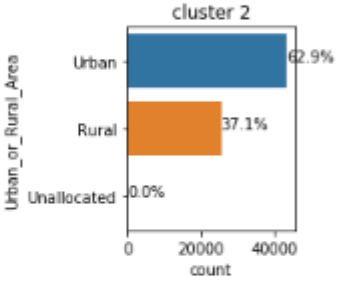
Fig 16. Quantity of elements of each cluster

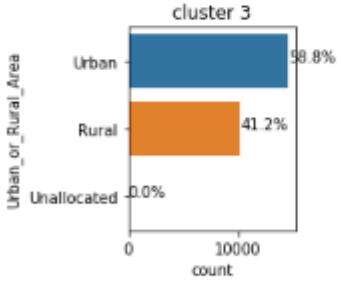
Check the distribution of observation over the cluster in some columns

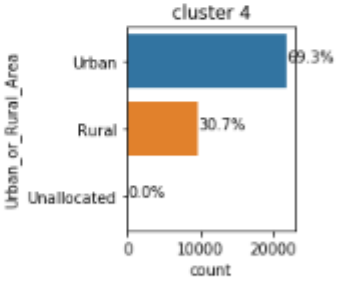
Urban_or_Rural_Area
overall

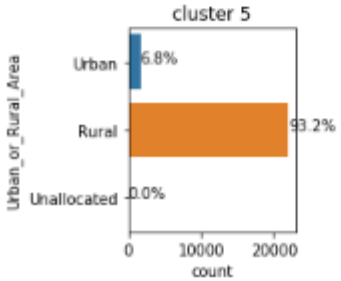


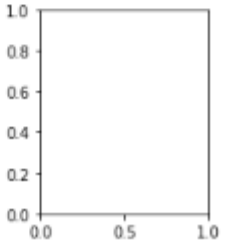




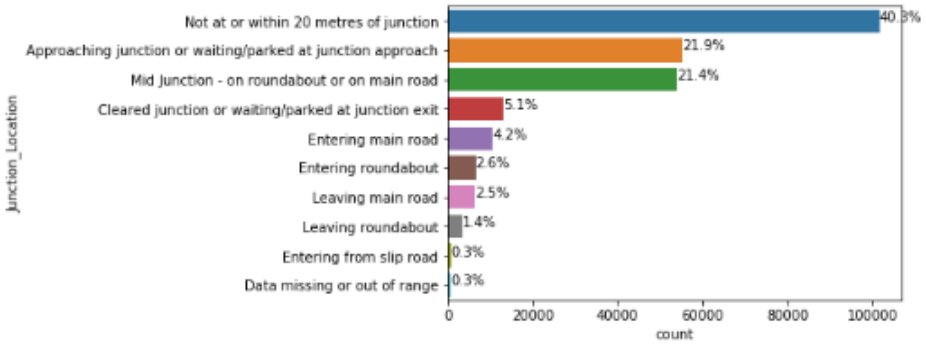


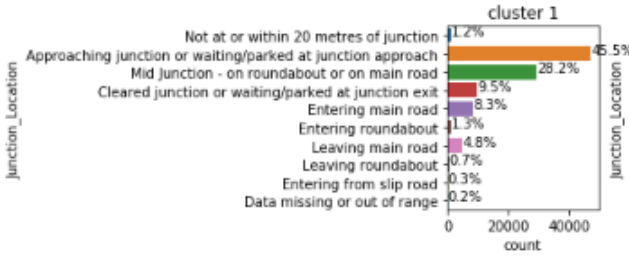


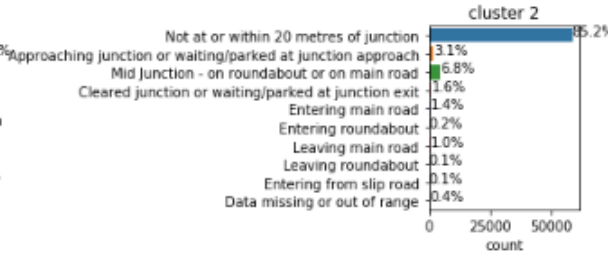


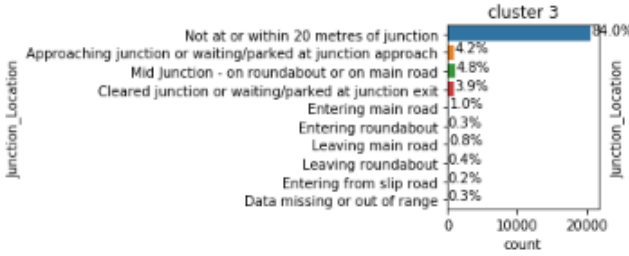


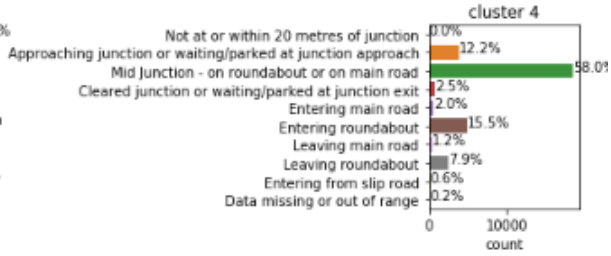
Junction_Location
overall

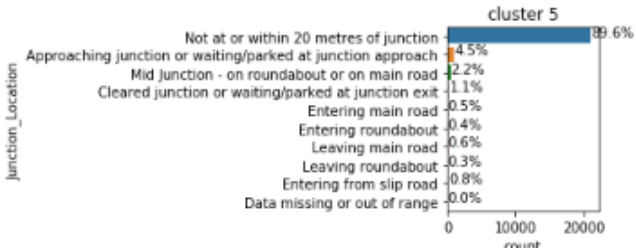


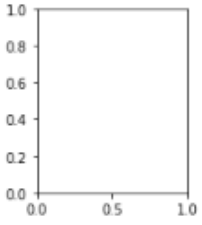












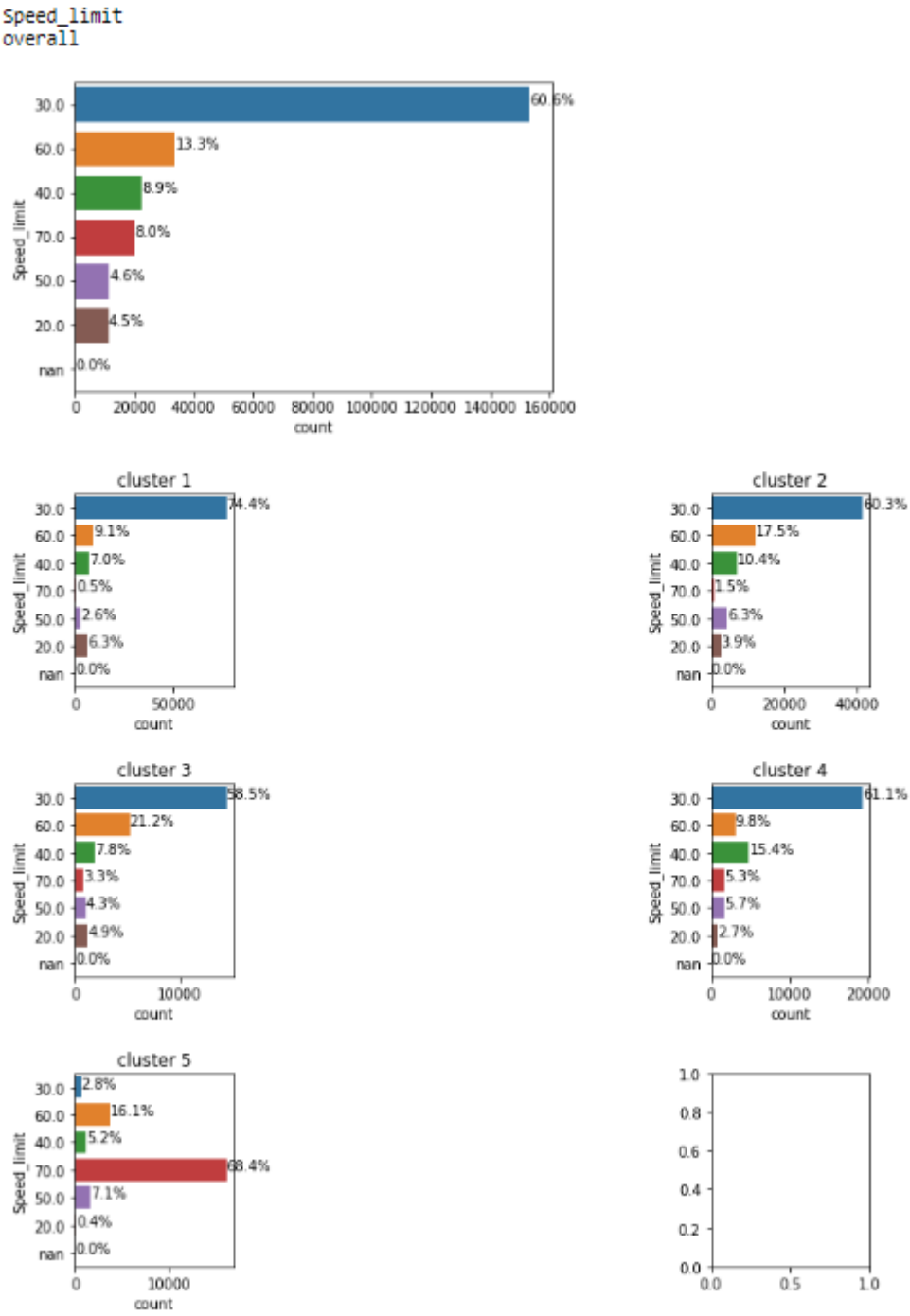


Fig 17-19. Distribution bar-plots

We can see some interesting patters.

- 1st cluster (41%) represent accidents mostly on low speed give way/uncontrolled T/ staggered junction. About 77% of this accident was in urban area
- 2nd and 3rd clusters (33.5% total) are very similar. They represent accident that happened far from junction and didn't fit in other clusters
- 4th cluster (12.5%) stands for accident Mid or right before roundabout or junction
- 5th cluster (9%) represent accidents mostly on rural areas, high speed road (Motorway or A class road), far from junction. 60% of vehicles had Heavy oil engine

From the fact that there is two very similar clusters, we can conclude that, four would be more suitable number of clusters than five. This is supported by pattern we found earlier, when tried to find appropriate number of clusters (Fig 14). The reason may be that dataset consist of four types of drivers gotten into accident or the fact that we dropped the least important half of columns for clustering with k=4

Conclusion and future direction

In this dataset investigation, I conducted an analysis of road vehicle collisions in the UK using open data provided by kaggle. My work consists of three parts. Simple analysis gave us some general patterns in dataset. Association matrices gave us understanding of dependencies between some columns. Clustering showed from what part consist this dataset. Finally, this investigation gave me substantial experience in big data analysis for my future projects.