Data Mining - Report

James Donohue (16024143)

Abstract

This is the abstract.
Exercise choices: classification and clustering

Contents

1 Data Choice

1.1 Data Set

The open data set used in this report is the Department for Transport's 2016 release of Road Safety Data (Department for Transport, 2017), which gives "the circumstances of personal injury road accidents … the types (including Make and Model) of vehicles involved and the consequential casualties". The data are published annually on the data.gov.uk website under a licence that permits non-commercial exploitation with attribution (The National Archives, 2017).

The source of the data is the 'STATS19' accident reporting form used by police and so the data is also referred to by this name. The STATS19 data has been recorded (with some changes in attribute semantics) since 1979 (Administrative Data Liaison Service, 2018). The Department for Transport also produces a

'STATS20' document which provides guidance on completing the form defines some of its terminology (Department for Transport, 2004).

The 2016 data set is distributed as three separate ZIP-compressed comma-separated variable (CSV) files, listed in Table 1 along with the number of data rows in each (each file also includes a single 'header' row providing metadata (Han, Kamber and Pei, 2012, p. 92) in the form of field names).

Table 1: Overview of 2016 data files

Link	Decompressed filename	Data rows
Road Safety Data - Accidents 2016	`dftRoadSafety_Accidents_2016.csv`	136621
Road Safety Data - Vehicles 2016	`Veh.csv`	252500
Road Safety Data - Casualties 2016	`Cas.csv`	181384

1.2 Attribute types

Every data row in each data file represents an object with a number of attributes that describe its features or characteristics (Han, Kamber and Pei, 2012, p. 40). The types of each attribute in the data set are summarised in Appendix A.

1.3 Coding of nominal data

Many attributes in the data set are of nominal data type (that is, representing names of things) but contain integer values. For example, an object in the *Accidents* file may have a value of *20* for the attribute `Police_Force`, but this number is not intended to be used quantitively, instead it represents a nominal value (Han, Kamber and Pei, 2012, p. 41). In this case, *20* represents the West Midlands police force.

An accompanying Excel spreadsheet `Road-Accident-Safety-Data-Guide.xls` (linked under *Additional resources*) provides lookup tables of all possible values for such 'coded' nominal types. This means such attributes have an *enumerated value domain* (OECD, 2006). The spreadsheet also explains that the special value `-1` represents "NULL or out of range values".

1.4 Data relationships

The Excel spreadsheet that accompanies the data set also explains the relations between the individual files:

> The ACC_Index field give a unique index for each accident and links to Vehicle and Casualty data. Casualties are linked to vehicles by "VEHREF".

Each of the three files therefore constitutes a relation with its own primary key (Codd, 1970), with `ACC_Index` (given as `Accident_Index` in the CSV header rows) as a unique accident identifier and `Casualty_Reference` and `Vehicle_Reference` identifying each casualty and the vehicle with which they are associated, respectively. The primary keys for each relation are shown in the simplified Entity Relationship (ER) diagram in Figure 1.

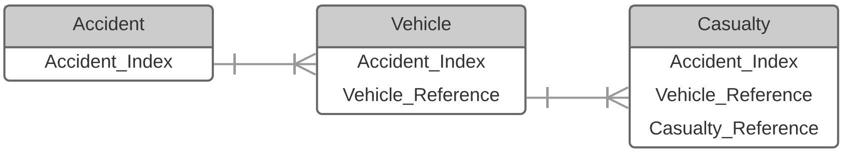

Figure 1: Simplified Entity Relationship diagram for data set

The `Number_of_Vehicles` and `Number_of_Casualties` fields in the *Accidents* file indicate the number of related rows in the other two files. For example, if `Number_of_Vehicles` is 2, there will be two rows in the *Vehicles* file with the same `Accident_Reference`, having a `Vehicle_Reference` of 1 and 2 respectively. The same applies to casualties.

1.5 Applicability of data mining techniques

1.5.1 Literature review

Hill (2005) evaluated the usefulness of several cluster analysis methods in identifying relatively homogenous groups of accidents from the STATS19 data for further investigation. Two-step, CHAID and cross-tabulation were considered, with the last two yielding the most promising results. The author found two-step cluster analysis as requiring too much subjective intervention by the analyst and producing less 'transparent' results than the other two methods.

2 Data Analysis

As the data set contains over 70 attributes this section will focus only on selected attributes as illustrations of the process of data analysis.

2.1 `Number_of_Vehicles`

The attribute `Number_of_Vehicles` (in the *Accidents* file) represents how many vehicles were involved in each accident. It is a numeric ratio-scaled attribute (that is, has an inherent zero-point) however as a typical 'count' attribute its values are only integers. In the 2016 data set it has no missing values.

Using the *Statistics* node in KNIME we can calculate some basic statistical descriptions for this attribute. The minimum and maximum values are 1 and 16 respectively. The **mean** number of vehicles in an accident is 1.8482, however as the data are positively skewed for this attribute (skew = 1.5756) a better measure of the centrally tendency is the **median** (Han, Kamber and Pei, 2012, p. 46). Here we can say that the median accident involved 2 vehicles. The **mode** is also 2. The histogram in Figure 2 shows the distribution of values for this attribute. Values of 6 and above are grouped together as they account for only a very small proportion of the data.

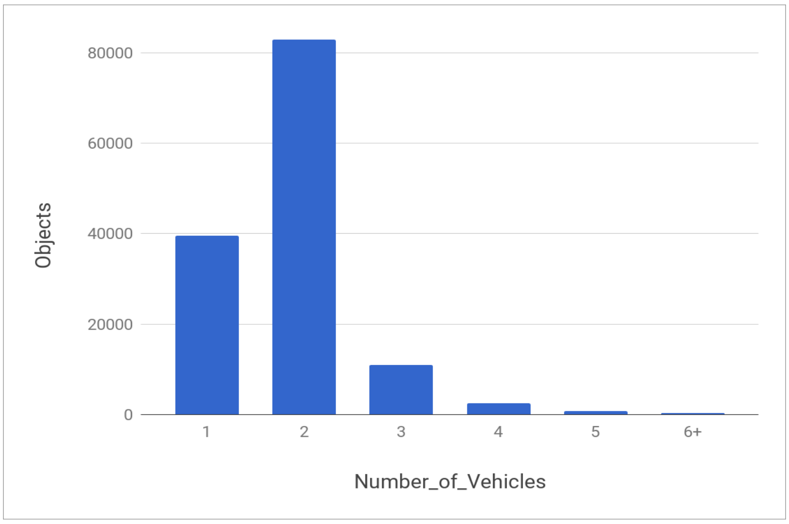

Figure 2: Histogram for `Number_of_Vehicles`

2.2 `Vehicle_Type`

The `Vehicle_Type` attribute in the *Vehicles* file indicates the type of each vehicle involved in an acident. It is a nominal (categorical) attribute. Although represented as an integer, it does not make sense to perform mathematical operations on it so we cannot calculate the mean vehicle type (Han, Kamber and Pei, 2012, p. 78), however the mode (most commonly occuring) type is 9 (Car). The integer values correspond to values in the *Vehicle Type* sheet of Excel spreadsheet that accompanies the data.

The KNIME workflow *Vehicle type correlation matrix* produces a correlation matrix for vehicle types in accidents involving two vehicles. In the resulting table, both the rows and columns represent a given vehicle type and the intersecting cell gives the number of accidents involving that combination of vehicles. This can be further visualised as a 'heat map' such as Figure 4 in order to reveal possible patterns. Cells are highlighted in different shades of red depending on their value; the highlighting midpoint was set to the 85th percentile after some experimentation to produce the clearest differentiation between cells.

2.3 `Date`

The `Date` attribute in the *Accidents* file is a string representation (with the format `MM/dd/YYYY`) of the calendar date between 1 January and 31 December 2016 that each accident occurred. It considered interval-scaled because it is measured on a scale of equal-size units (here days) but no true zero-point (Han, Kamber and Pei, 2012, p. 80).

Every date occcurs at least once in the data set. Using the *Statistics* node in KNIME we can determine that the date that occurs the most (i.e. with the most accidents) was 25 November, with 566 accidents.

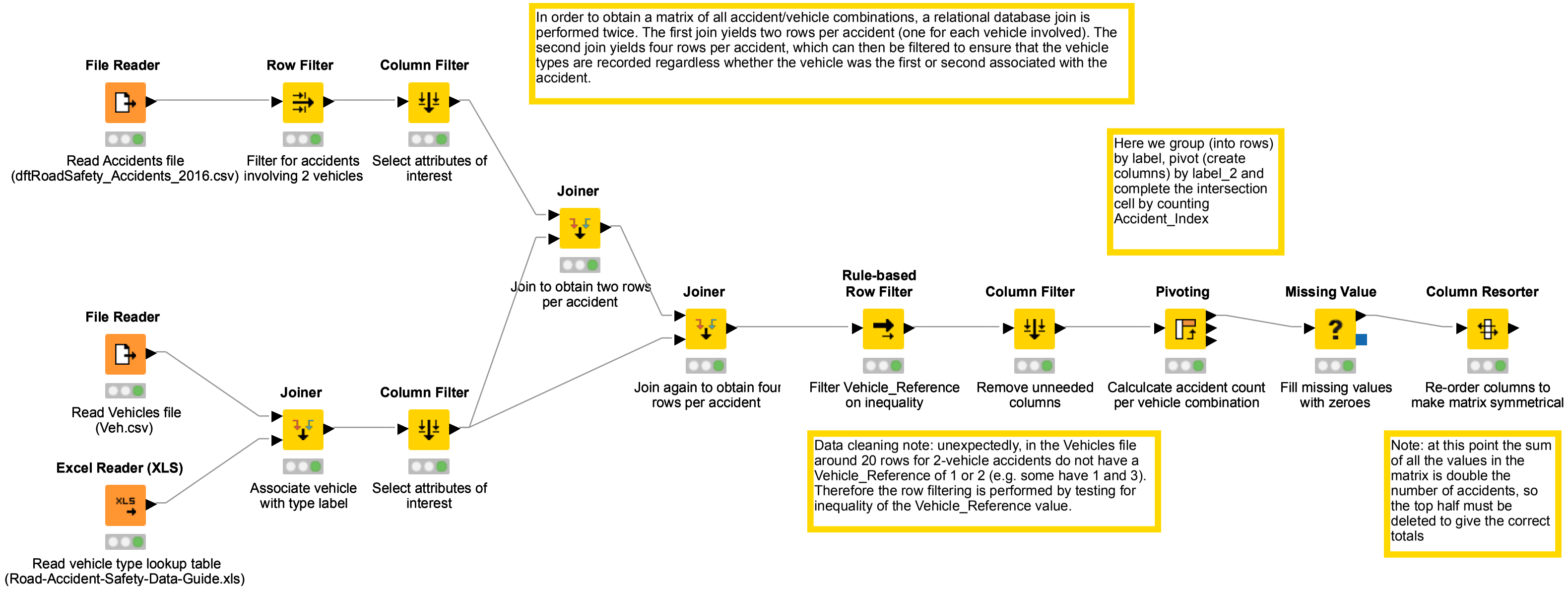

Figure 3: Workflow for vehicle type correlation matrix

Vehicle type	Agricultural vehicle	Bus or coach (17 or more pass seats)	Car	Data missing or out of range	Electric motorcycle	Goods 7.5 tonnes mgw and over	Goods over 3.5t. and under 7.5t	Goods vehicle - unknown weight	Minibus (8 - 16 passenger seats)	Mobility scooter	Motorcycle - unknown cc	Motorcycle 125cc and under	Motorcycle 50cc and under	Motorcycle over 125cc and up to 500cc	Motorcycle over 500cc	Other vehicle	Pedal cycle	Ridden horse	Taxi/Private hire car	Tram	Van / Goods 3.5 tonnes mgw or under	
Agricultural vehicle	0																					
Bus or coach (17 or more pass seats)	4	54																				
Car	262	1303	37684																			
Data missing or out of range	0	11	113	1																		
Electric motorcycle	0	0	38	1	1																	
Goods 7.5 tonnes mgw and over	5	32	1975	0	1	83																
Goods over 3.5t. and under 7.5t	4	14	641	0	0	32	19															
Goods vehicle - unknown weight	0	12	498	1	2	18	5	8														
Minibus (8 - 16 passenger seats)	0	4	167	0	0	5	4	0	0													
Mobility scooter	0	4	149	2	0	7	4	2	0	0												
Motorcycle - unknown cc	1	0	170	0	0	2	2	5	1	0	3											
Motorcycle 125cc and under	10	57	5450	11	0	62	34	25	12	1	4	74										
Motorcycle 50cc and under	1	15	1090	0	0	14	7	6	4	0	2	7	13									
Motorcycle over 125cc and up to 500cc	4	11	1297	0	0	20	10	2	3	0	0	8	4	16								
Motorcycle over 500cc	30	35	3598	7	0	61	35	30	10	0	2	29	8	13	78							
Other vehicle	1	24	582	0	0	14	9	4	6	0	1	38	8	7	29	22						
Pedal cycle	36	275	13968	50	3	141	130	95	37	2	23	146	54	30	88	86	109					
Ridden horse	1	3	62	2	0	3	1	1	0	0	1	0	1	1	0	1	2	2				
Taxi/Private hire car	0	51	1628	2	0	31	15	2	8	2	1	307	41	56	106	18	663	0	107			
Tram	0	0	9	0	0	0	0	0	0	0	0	1	0	0	0	1	3	0	1	0		
Van / Goods 3.5 tonnes mgw or under	26	156	4897	5	0	187	64	30	9	12	9	503	103	128	351	48	1204	8	131	0	279	
Total number of accidents	**385**	**2061**	**74016**	**82**	**7**	**680**	**339**	**210**	**90**	**17**	**46**	**1113**	**232**	**251**	**652**	**176**	**1981**	**10**	**239**	**0**	**279**	**82866**

Figure 4: Heat map for `Vehicle_Type` in two-vehicle accidents

This is therefore the mode date. However least-occurring date (i.e. with the least accidents) was 25 December, with only 138, which is likely due to fewer people travelling on Christmas Day. For comparison, since 2016 was a leap year with 366 days and the *Accidents* file contains data about 136621 accidents, we can say that the mean number of accidents per day was 373.28.

Although analysis of accident numbers over the year might reveal seasonal trends, to be sound any conclusions would need to be supported by comparison with data from other years. Since reviewing multiple years' data is out of scope for this report, an alternate approach is to look for trends within the year under study. For example, we may wish to know if significantly different numbers of accidents occur at weekends. Figure 5 uses a box plot to visualise the distribution of of accidents per day of the week. Each bar represents a different day. Data points that are more than 1.5 times the inter-quartile range (IQR) are plotted separately, all of which are classified by KNIME as 'mild' outliers.

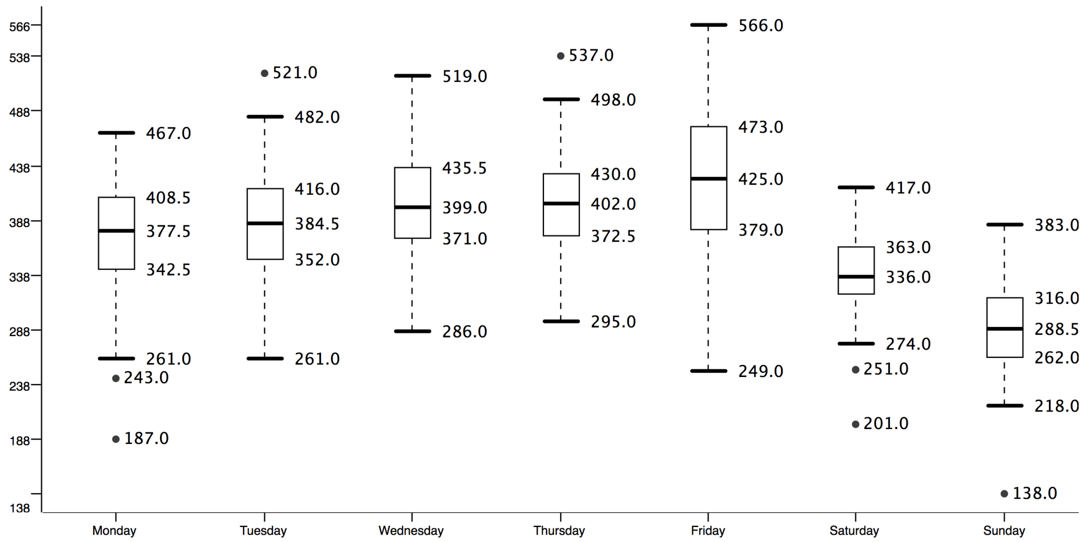

Figure 5: Box plot of accidents by day of week

From this we can see that the number of accidents increases slightly as the week progresses, and that weekends have significantly fewer accidents. It is also notable that the 'whiskers' (representing the maximum and minimum values) for Friday are much longer for other days. Table 2 confirms that Friday has a much larger standard deviation than other days of the week, meaning that it has the greatest dispersion of values.

Table 2: Distribution of accidents by day of week

Day of week	Min	Max	Mean	StdDev
Monday	187	467	370.69	57.34
Tuesday	261	521	386.62	46.42
Wednesday	286	519	401.37	47.73
Thursday	295	537	402.87	47.43
Friday	249	566	426.02	63.82
Saturday	201	417	336.19	42.73
Sunday	138	383	288.92	42.34

Note that this analysis does not allow us to conclude that weekends are somehow inherently safer - there may simply be less traffic. In order to know for sure we would need to cross-reference data about how many vehicles are on the road on each day, which is beyond the scope of this study.

2.4 Time

The Time attribute in the *Accidents* file represents the time of day that an accident occurred. Like Date, it is an interval-scaled attribute with values ranging from 00:00 (midnight) to 23:59. By extracting the hour and minute components from each time we can produce histograms that reveal patterns in the data.

Figure 6 shows the number of accidents that occur during each hour of the day, from 0 to 23, across the whole data set. This shows that the fewest accidents occur between 04:00-04:59, while there are noticeable peaks between 08:00-08:59 and 17:00-17:59. These peaks could be predicted as they correspond to the usual 'rush hours' at the start and end of the working day. Nearly 25% of accidents occurred between 8am and 9am or between 4pm and 6pm.

Similarly, Figure 7 shows the number of accidents that occur during each minute of the day, from 0 to 59, across the whole data set. It is apparent that the most frequent times are on the hour and half-past the hour, with smaller peaks every five minutes. Although this could mean that acccidents happens more frequently at these points, a more likely explanation is that the police offers reporting the accident often round their approximation of the accident time to the nearest 5- or 30-minute boundary. This 'snapping' effect implies that attempting to analyse the Date and Time attributes with a resolution of less than 5 minutes is unlikely to be successful. A comparable level of accuracy in police reporting road accident times was inferred for road accident data collected about Helsinki's Ring Road (Innamaa et al., 2014, p. 18).

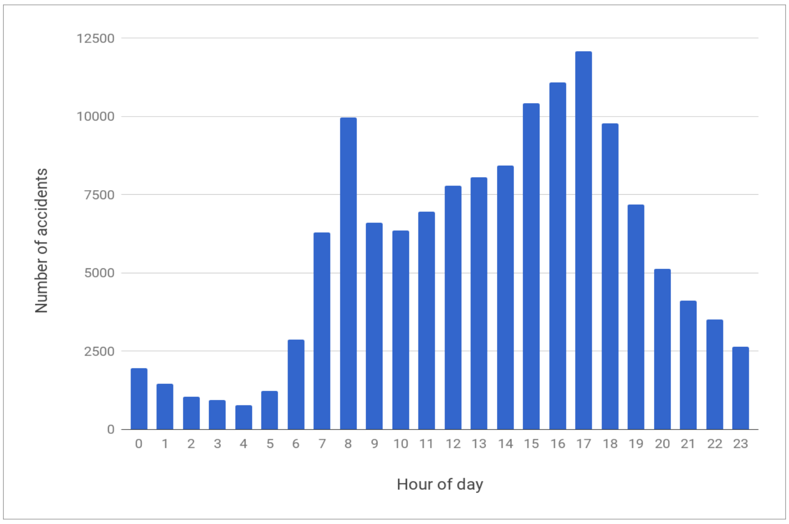

Figure 6: Histogram of number of accidents occuring during each hour of day

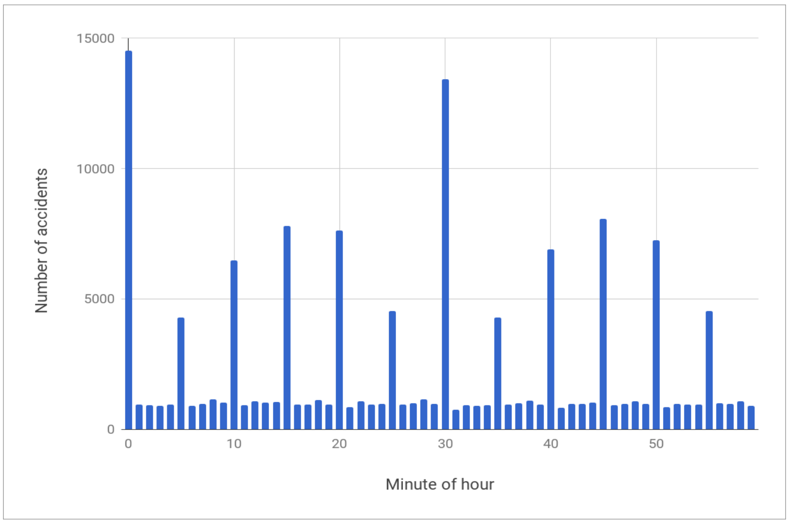

Figure 7: Histogram of number of accidents occuring during each minute of hour

2.5 Missing values

2.6 Data Preprocessing

We can use the knowledge that the majority (around 60%) of accidents involve two vehicles to focus this study and eliminate some of the complexity that arises from one-to-many data relationships. By considering only accidents involving two vehicles, we can 'denormalise' the *Accident* and *Vehicle* files to allow for further processing.

3 Appendix: Summary of software and KNIME workflows

The software used in producing the figures and tables in this report is:

- KNIME 3.5.1
- Google Sheets (histograms, heat maps)

Table 3 summarises the KNIME workflows included with this report.

Table 3: Summary of KNIME workflows

Workflow	Purpose
Accident date distribution	Creates a box plot for accidents by day of the week and statistics table

Workflow	Purpose
Vehicle Type correlation matrix (Figure 3)	Generates a correlation matrix of vehicle types in two-accident collisions suitable for rendering as a heat map
Accident date distribution	Generates a box plot showing distribution of accident dates and statistical summary

References

Administrative Data Liaison Service (2018) *STATS19 Road Accident dataset details*. http://www.adls.ac.uk/department-for-transport/stats19-road-accident-dataset/?detail Accessed 18 May 2018.

Codd, E. (1970) A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, 13(6), pp. 377–387.

Department for Transport (2004) *Instructions for the Completion of Road Accident Reports*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/230597/stats20-2005.pdf Accessed 18 May 2018.

Department for Transport (2017) *Road Safety Data - 2016*. https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data Accessed 29 April 2018.

Han, J., Kamber, M. and Pei, J. (2012) *Data Mining: Concepts and Techniques*. 3rd edition, Morgan Kaufmann.

Hill, J. P. (2005) *The innovatory analysis of road traffic accident data*. https://trl.co.uk/reports/PPR056 Accessed 18 May 2018.

Innamaa, S., Norros, I., Kuusela, P., Rajamäki, R. and Pilli-Sihvola, E. (2014) *Road traffic incident risk assessment*. https://www.vtt.fi/inf/pdf/technology/2014/T172.pdf Accessed 18 May 2018.

OECD (2006) *Glossary of Statistical Terms: Value Domain*. https://stats.oecd.org/glossary/detail.asp?ID=2849 Accessed 30 April 2018.

The National Archives (2017) *Open Government Licence v3.0*. http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/ Accessed 29 April 2018.