Contents

1 Business understanding 3

1.1 Problem definition 3

1.2 Project stakeholders 3

1.3 Determine data mining goals 3

1.4 Refining the problem statement 4

1.5 Project success criteria 4

2 Data understanding 5

2.1 Kaggle data 5

2.2 Additional data collection 5

2.3 Data description 6

2.4 Data quality report 8

2.5 Data exploration 18

3 Data preparation 27

3.1 Selecting data 27

3.2 Data cleaning 27

3.3 Data construction 33

3.4 Data integration 33

3.5 Data formatting 33

4 Model Building 34

4.1 Select modelling technique 34

4.2 Generate test design 36

4.3 Build model 36

4.4 Assess model 37

4.5 Time series analysis and forecasting 39

5 Deployment 47

6 Conclusions & Recommendations 48

6.1 Conclusions 48

6.2 Recommendations for future work 49

Appendix A: Project proposal 50

Appendix B: Scatter plot with high covariance 51

Appendix C: Scatter plots with high correlation (r >= 0.7) 53

Appendix D: R code for correlation analysis and modelling 55

Appendix E: R Code for time series analysis and forecasting 59

1 Business understanding

1.1 Problem definition

A competition was recently launched on Kaggle[1] with data on accidents and traffic count data across the UK. The data sets consisted of 1.6 million accidents, from 2005 to 2014, and 16 years of traffic flow (2000 – 2016). The authors of the competition are interested in understanding what factors affect the number of traffic-related accidents.

1.2 Project stakeholders

Stakeholders are those with any interest in your project's outcome. Thus, the project stakeholders will be any individual or organisation that has an interest in and/or makes use of road networks in the UK. This includes:

- Local authorities
- Department for Transport (UK)
- Policing Authority
- Road users (drivers, passengers, pedestrians and cyclists)
- Insurance companies
- Car manufacturers[2]

The outputs of this study will inform stakeholders of the features which have the biggest influence on the number of accidents that occur, with a focus on the accident severity. This knowledge base can be used to establish policies and actions plans that will improve road safety.

1.3 Determine data mining goals

Examples of factors which could affect the number and severity of accidents include:

- Road characteristics (speed limit, road type, surface conditions, lighting)
- Weather conditions
- Traffic volume
- Location and time of day
- Driver and vehicle characteristics
- Police presence

1 https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales

[2] While car manufacturers will have an interest in any insights that can improve road and vehicle safety, the scope of this assignment will not focus on aspects of vehicle safety.

1.4 Refining the problem statement

For the given problem, we would like to answer the following questions:

Descriptive

1. How observed changes in road traffic volumes compare with changes in road traffic accidents?
2. Common features of accidents?

Predictive

3. Can we predict accident severity using other attributes (classification)?
4. Can we predict accident rates over time (time series)?

The project proposal (**Appendix A**) referenced a number of other areas for analysis. These are discussed in more detail in Section 6.

1.5 Project success criteria

A successful outcome of this project would be to identify key factors that influence the severity of road traffic accidents in the UK. Furthermore, the development of an accurate prediction model would assist road safety groups (and policy makers) implement various measures at both local and national level to improve road safety e.g. motorway upgrades, improved junction types, reduced-speed zones, targeted road gritting strategy, additional lighting.

Reducing the number of accidents, as well as the severity, would have additional economic benefits:

- Less road traffic delays, which means less time lost for individuals and businesses
- Police resources can be re-deployed in other community areas
- Hospitals will have greater capacity due to reduced A&E demand
- Safer roads will encourage more people to cycle (which is more environmentally friendly) , which in turn will increase road capacity and reduce the requirement to provide additional parking facilities

2 Data understanding

2.1 Kaggle data

The files made available on Kaggle were grouped into two areas (see Table 1).

Table 1: Kaggle files

Area	File Name	Description
Traffic	ukTrafficAADF	Average annual daily flow values from traffic counters across the UK (from 2000 to 2016).
Accident	accidents_2005_to_2007 accidents_2009_to_2011 accidents_2012_to_2014	Records of all reported traffic accidents in the UK from 2005 to 2014 (2008 is missing).

All files provided were in comma separated value (CSV) format.

2.2 Additional data collection

On acquiring the data available in Kaggle, it became evident that additional information would be required for two reasons:

Missing data: there was no information available on vehicles or drivers involved in accidents, and

Missing names: in some instances categorical attributes were provided in numeric format, with no explanation on the numbering system employed. Without a corresponding name to match the number, the attribute was not useful for descriptive analysis.

In addition to the Kaggle files (Section 2.1), the following data was obtained:

- For each accident, corresponding vehicle and driver details were acquired from the UK government's open-source database[3] for the period 2005 to 2014.
- Data was acquired from *OpenDataSoft*[4] to match local authority ID codes with local authority names, as well as their corresponding region and country.
- STATS19 police accident report form, which contained corresponding name terms to match numerical values of categorical attributes within the Kaggle files.

All files acquired were in CSV format, apart from the STATS19 form, which was in PDF format. The STATS19 form was manually transferred into a CSV file format to allow replacement of numerical values with name terms.

[3] https://data.gov.uk/

[4] https://data.opendatasoft.com/pages/home/

2.3 Data description

A description of the information contained in each of the main files acquired is provided in Table 2.

Table 2: Data description

Area	File Name	Events (No. of rows)	Attributes (No. of fields)	Attribute Names
Traffic	ukTrafficAADF	~ 275,000	29	**AADFYear** CP Estimation_method Estimation_method_detailed **Region** **LocalAuthority** **Road** **RoadCategory** Easting Northing **StartJunction** **EndJunction** **LinkLength_km** **LinkLength_miles** **PedalCycles** **Motorcycles** **CarsTaxis** **BusesCoaches** **LightGoodsVehicles** V2AxleRigidHGV V3AxleRigidHGV V4or5AxleRigidHGV V3or4AxleArticHG V5AxleArticHGV V6orMoreAxleArticHGV **AllHGVs** **AllMotorVehicles** **Lat** **Lon**
Accident	merged_accidents	~ 1.6 million	33	**Accident_Index** Location_Easting_OSGR Location_Northing_OSGR **Longitude** **Latitude** **Police_Force** **Accident_Severity** **Number_of_Vehicles** **Number_of_Casualties** **Date** **Day_of_Week** **Time** **Local_Authority_(District)** **Local_Authority_(Highway)** 1st_Road_Class 1st_Road_Number

Area	File Name	Events (No. of rows)	Attributes (No. of fields)	Attribute Names
				Road_Type **Speed_limit** **Junction_Detail** Junction_Control 2nd_Road_Class 2nd_Road_Number **Pedestrian_Crossing-Human_Control** **Pedestrian_Crossing-Physical_Facilities** **Light_Conditions** **Weather_Conditions** **Road_Surface_Conditions** **Special_Conditions_at_Site** **Carriageway_Hazards** **Urban_or_Rural_Area** **Did_Police_Officer_Attend_Scene_of_Accident** LSOA_of_Accident_Location **Year**
Vehicles	vehicles0514	~ 3 million	22	**Accident_Index** Vehicle_Reference **Vehicle_Type** Towing_and_Articulation **Vehicle_Manoeuvre** Vehicle_Location-Restricted_Lane **Junction_Location** Skidding_and_Overturning Hit_Object_in_Carriageway Vehicle_Leaving_Carriageway Hit_Object_off_Carriageway 1st_Point_of_Impact Was_Vehicle_Left_Hand_Drive? **Journey_Purpose_of_Driver** **Sex_of_Driver** **Age_of_Driver** **Age_Band_of_Driver** **Engine_Capacity_(CC)** Propulsion_Code **Age_of_Vehicle** Driver_IMD_Decile Driver_Home_Area_Type

Based on the problem statements outlined in Section 1.4, attributes which were of interest to addressing the problems were highlighted (in **bold**) in the table above for further analysis.

File formats are detailed in Section 2.1 and 2.2.

2.4 Data quality report

The following philosophy was adopted in determining how features should be handled (shown in Table 3).

Table 3: Data handling philosophy

Data Quality Issue	Handling strategy
Missing values	If the number of values missing for a feature is greater than 10-15% of the data provided, then the feature would be removed. In cases where less than 10-15% of values are missing, an impute function would be applied to replace missing values with the average. There were some exceptions to this e.g. dates, geographic coordinates. In these particular instances, missing values would be retained as blanks.
Irrelevant features	Where the team consider that a feature will not assist with addressing the problem statements (see Section 1.4), the feature would be removed.
Duplicate features	There are two sets of coordinates in each data set: • Latitude/Longitude • Easting/Northing For this analysis, a single set of coordinates will suffice; Easting/Northing values would be removed. Similarly for the 'traffic' file, road lengths were provided in miles and kilometres. The miles attribute would be removed.
Negative values	Where negative values occur for categorical features, these would be replaced with average values.
Features outside of range	In some instances, numeric values for categorical features would fall outside of the allowable reporting range e.g. the STATS19 report form might specify six vehicle types, but the value entered is greater than six. In these instances, the outlier value would be replaced with the average.

An analysis was undertaken on the continuous and categorical features in accident, traffic and vehicles files, which is detailed in Table 4 to Table 9.

Table 4: Continuous features in 'Traffic' file

Feature	Distribution	Min Max	Average (Stan.Dev.)	Data quality issue	Handling Strategy
AADFYear	Uniform	2000 2016	2007.968 (4.863)	None – time period is correct	Retain
CP	Uniform	60 99967	47277.773 (27004.784)	Feature not relevant to analysis	Remove
Easting	Normal – skewed left	69987 655040	425674.683 (98687.704)	None – data plotted and appears to geographically correct. However, no need to have TM and NG coordinate data	Remove
Northing	Normal – skewed right	76250 1205400	356182.956 (186992.933)	None – data plotted and appears to geographically correct. However, no need to have TM and NG coordinate data	Remove
LinkLength_km	Exponential	0.07 55.5	2.757	None – outer values appear to be realistic	Retain
LinkLength_miles	Exponential	0.04 34.49	1.713	None – outer values appear to be realistic. However, no need to have both km and miles	Remove
PedalCycles	Exponential	0 18,629	123.315	Zero value is unlikely – possibly an error with the counter.	Retain
Motorcycles	Exponential	0 9,815	222	Zero value is unlikely – possibly an error with the counter.	Retain

Feature	Distribution	Min Max	Average (Stan.Dev.)	Data quality issue	Handling Strategy
CarsTaxis	Exponential	0 207,133	16,813	Zero value is unlikely – possibly an error with the counter.	Retain
BusesCoaches	Exponential	0 11,359	249	Zero value is unlikely – possibly an error with the counter.	Retain
LightGoodsVehicles	Exponential	0 38,449	2,613	Zero value is unlikely – possibly an error with the counter.	Retain
V2AxleRigidHGV	Exponential	0 10,942	489	Zero value is unlikely – possibly an error with the counter.	Retain
V3AxleRigidHGV	Exponential	0 5,968	84	Zero value is unlikely – possibly an error with the counter.	Retain
V4or5AxleRigidHGV	Exponential	0 3,684	89	Zero value is unlikely – possibly an error with the counter.	Retain
V3or4AxleArticHGV	Exponential	0 3,949	74	Zero value is unlikely – possibly an error with the counter.	Retain
V5AxleArticHGV	Exponential	0 11,034	248	Zero value is unlikely – possibly an error with the counter.	Retain
V6orMoreAxleArticHGV	Exponential	0 13,758	272	Zero value is unlikely – possibly an error with the counter.	Retain

Feature	Distribution	Min Max	Average (Stan.Dev.)	Data quality issue	Handling Strategy
AllHGVs	Exponential	0 27,095	1,256	Zero value is unlikely – possibly an error with the counter.	Retain
AllMotorVehicles	Exponential	0 262,842	21,153	Zero value is unlikely – possibly an error with the counter.	Retain
Lat	Normal – skewed right	5,584 60.727	53.093		Retain
Lon	Normal – skewed left	-7.443 1.755	-1.655		Retain

Table 5: Categorical features in 'Traffic' file

Feature	1st mode	1st mode (%)	2nd mode	2nd mode (%)	Data Quality Issue	Handling Strategy
Estimation_method	Estimated	73.8	Counted	26.2	Missing values (15%)	Remove
Estimation_method_detailed	Estimated using previous year's AADF on this link	67.1	Manual count	23.3	Missing values (15%)	Remove
Region	Scotland	13.9	South East	13.6		'Merseyside' changed to 'North West'
LocalAuthority	Kent	2.7	Lancashire	2.1		Retain
Road	A1	1.3	A6	1.2		Retain
RoadCategory	PU	48.9	PR	31.8	Cardinality issues – some labels appear to be identical but have case issues	Replace lower case values with upper case

Feature	1st mode	1st mode (%)	2nd mode	2nd mode (%)	Data Quality Issue	Handling Strategy
StartJunction	LA boundary	8.6	A6	0.4		Retain
EndJunction	LA boundary	9.4	A6	0.5		Retain

Table 6: Continuous features in 'Accident' file

Feature	Distribution	Min Max	Average(Stan.Dev).	Data Quality Issue	Handling Strategy
Location_Easting_OSGR	Normal – skewed left	64950 655370	439621.405	Missing values (0.01%)	Remove – no need to have TM and NG coordinate data
Location_Northing_OSGR	Normal – skewed right	10290 1208800	300158.404	Missing values (0.01%)	Remove – no need to have TM and NG coordinate data
Longitude	Normal – skewed left	-7.516 1.759	-1.437	Missing values (0.01%)	Retain as is, inappropriate to assign avg. values to coordinate
Latitude	Normal – skewed right	49.913 60.758	52.589	Missing values (0.01%)	Retain as is, inappropriate to assign avg. values to coordinate
Police_Force	Uniform	1 98	30.205		Retain
Accident_Severity	Exponential - reverse	1 3	2.838		Retain
Number_of_Vehicles	Exponential	1 67	1.832		Retain
Number_of_Casualties	Exponential	1 93	1.351		Retain

Feature	Distribution	Min Max	Average(Stan.Dev).	Data Quality Issue	Handling Strategy
Day_of_Week	Uniform	1 7	4.119		Retain
Local_Authority_(District)	Uniform	1 941	347.615		Remove feature, not required for analysis
1st_Road_Class	Uniform	1 6	4.088		Retain
1st_Road_Number	Exponential	-1 9999	1009.919		Retain
Speed_limit	Normal – skewed right	10 70	39.005		Retain
2nd_Road_Class	Uniform	-1 6	2.675		Retain
2nd_Road_Number	Exponential	-1 9999	381.568		Retain
Urban_or_Rural_Area	Exponential	1 3	1.354	'3' should not be in dataset	Replace 3 with 2 (closest value)
Year	Uniform	2005 2014	2009.37 (3.013)		Retain
Junction_Detail	-	-	-	Missing values (100%)	Remove

Table 7: Categorical features in 'Accident' file

Feature	1st mode	1st mode (%)	2nd mode	2nd mode (%)	Data Quality Issue	Handling Strategy
Accident_Index	2.01E+12	25.4	2.00913E+12	0.04	None – primary key	Retain

Feature	1st mode	1st mode (%)	2nd mode	2nd mode (%)	Data Quality Issue	Handling Strategy
Date	21/10/05	0.01	18/11/05	0.01	None – dates within defined period	Retain
Time	17:00	1.0	17:30	0.9	Missing values (0.01%)	Replace with mean values
Local_Authority_(Highway)	E10000016	2.7	E10000030	2.5		Replace ID values with names (from external spreadsheet)
Road_Type	Single carriageway	74.9	Dual carriageway	14.7		Retain
Junction_Control	Giveway or uncontrolled	81.4	Automatic traffic signal	17.3	Missing values (40%)	Remove
Pedestrian_Crossing-Human_Control	None within 50 metres	99.4	Control by other authorised person	0.03	Missing values (<0.01%)	Replace with mean values
Pedestrian_Crossing-Physical_Facilities	No physical crossing within 50 meters	83.3	Pedestrian phase at traffic signal junction	0.=6.7	Missing values (<0.01%)	Replace with mean values
Light_Conditions	Daylight: Street light present	73.3	Darkness: Street lights present and lit	19.7		Retain
Weather_Conditions	Fine without high winds	80.0	Raining without high winds	11.8	Missing values (0.01%)	Replace with mean values
Road_Surface_Conditions	Dry	68.9	Wet/Damp	28.2	Missing values (0.13%)	Replace with mean values
Special_Conditions_at_Site	None	97.6	Roadworks	1.1		Retain
Carriageway_Hazards	None	98.2	Other object in carriageway	0.08		Retain

Feature	1st mode	1st mode (%)	2nd mode	2nd mode (%)	Data Quality Issue	Handling Strategy
Did_Police_Officer_Attend_Scene_of_Accident	Yes	81.2	No	18.8	Missing values (0.19%)	Replace with mean values
LSOA_of_Accident_Location	E01000004	0.01	E01011365	0.01	Missing values (7.2%)	Remove feature, not required for analysis

Table 8: Continuous features in 'Vehicles' file

Feature	Distribution	Min Max	Average(Stan.Dev).	Data Quality Issue	Handling Strategy
Vehicle_Reference	Exponential	-1 91	1.555	Negative values	Replace negative values with average
Vehicle_Type	Exponential	-1 98	9.651	Negative values	Replace negative values with average
Towing_and_Articulation	Exponential	-1 5	0.031	Negative values	Remove – not required for analysis
Vehicle_Manoeuvre	Normal – skewed left	-1 18	12.687	Negative values	Replace negative values with average
Vehicle_Location-Restricted_Lane	Exponential	-1 9	0.133	Negative values	Remove – not required for analysis
Junction_Location	Exponential	-1 8	2.533	Negative values	Replace negative values with average
Skidding_and_Overturning	Exponential	-1 5	0.214	Negative values	Remove – not required for analysis
Hit_Object_in_Carriageway	Exponential	-1 12	0.305	Negative values	Remove – not required for analysis

Feature	Distribution	Min Max	Average(Stan.Dev).	Data Quality Issue	Handling Strategy
Vehicle_Leaving_Carriageway	Exponential	-1 8	0.374	Negative values	Remove – not required for analysis
Hit_Object_off_Carriageway	Exponential	-1 11	0.572	Negative values	Remove – not required for analysis
1st_Point_of_Impact	Normal – skewed right	-1 4	1.765	Negative values	Remove – not required for analysis
Was_Vehicle_Left_Hand_Drive?	Exponential	-1 2	0.989	Negative values	Remove – not required for analysis
Journey_Purpose_of_Driver	Exponential - reverse	-1 15	8.714	Negative values 75% of values fall outside of the report conditions	Remove
Sex_of_Driver	Exponential	-1 3	1.399	Negative values	Replace negative values with average
Age_of_Driver	Normal – skewed right	-1 100	34.328	Negative values	Replace negative values with average
Age_Band_of_Driver	Normal – skewed left	-1 11	5.85	Negative values	Remove – not required for analysis
Engine_Capacity_(CC)	Exponential	-1 99999	1411.705	Negative values	Replace negative values with average
Propulsion_Code	Normal – skewed right	-1 12	0.752	Negative values	Remove – not required for analysis
Age_of_Vehicle	Exponential	-1 111	4.841	Negative values	Replace negative values with average
Driver_IMD_Decile	Uniform	-1 10	3.559	Negative values	Remove – not required for analysis

Feature	Distribution	Min Max	Average(Stan.Dev).	Data Quality Issue	Handling Strategy
Driver_Home_Area_Type	Exponential	-1 3	0.882	Negative values	Remove – not required for analysis

Table 9: Categorical features in 'Vehicles' file

Feature	1st mode	1st mode (%)	2nd mode	2nd mode (%)	Data Quality Issue	Handling Strategy
Accident_Index	2013460234852	<0.01	2011520104001	<0.01	None – primary key	Retain

2.5 Data exploration

2.5.1 Data insights - Descriptive

Based on the cleaned data, approximately 1.47 million accidents were reported between 2005 and 2014. Almost 90% of accidents occurred in England, with the remainder divided between Scotland (6.7%) and Wales (4.3%). The South-East and London accounted for 30% of accidents, with the fewest occurring the North East (3.8%).

Accidents occurred most frequently in the local authority areas of Kent, Surrey, Lancashire, Essex and Hampshire; each accounting for more than 2% of accidents. The fewest number of accidents, between 2005 and 2014, occurred in the Isles of Scilly (19 accidents). Police attended 81% of reported accidents.

One of our objectives (Objective 1 in Section 1.4) was to compare traffic volume and accidents trend for the analysis period. A time series of total accidents, by accident severity, and traffic volumes is shown in Figure 1.

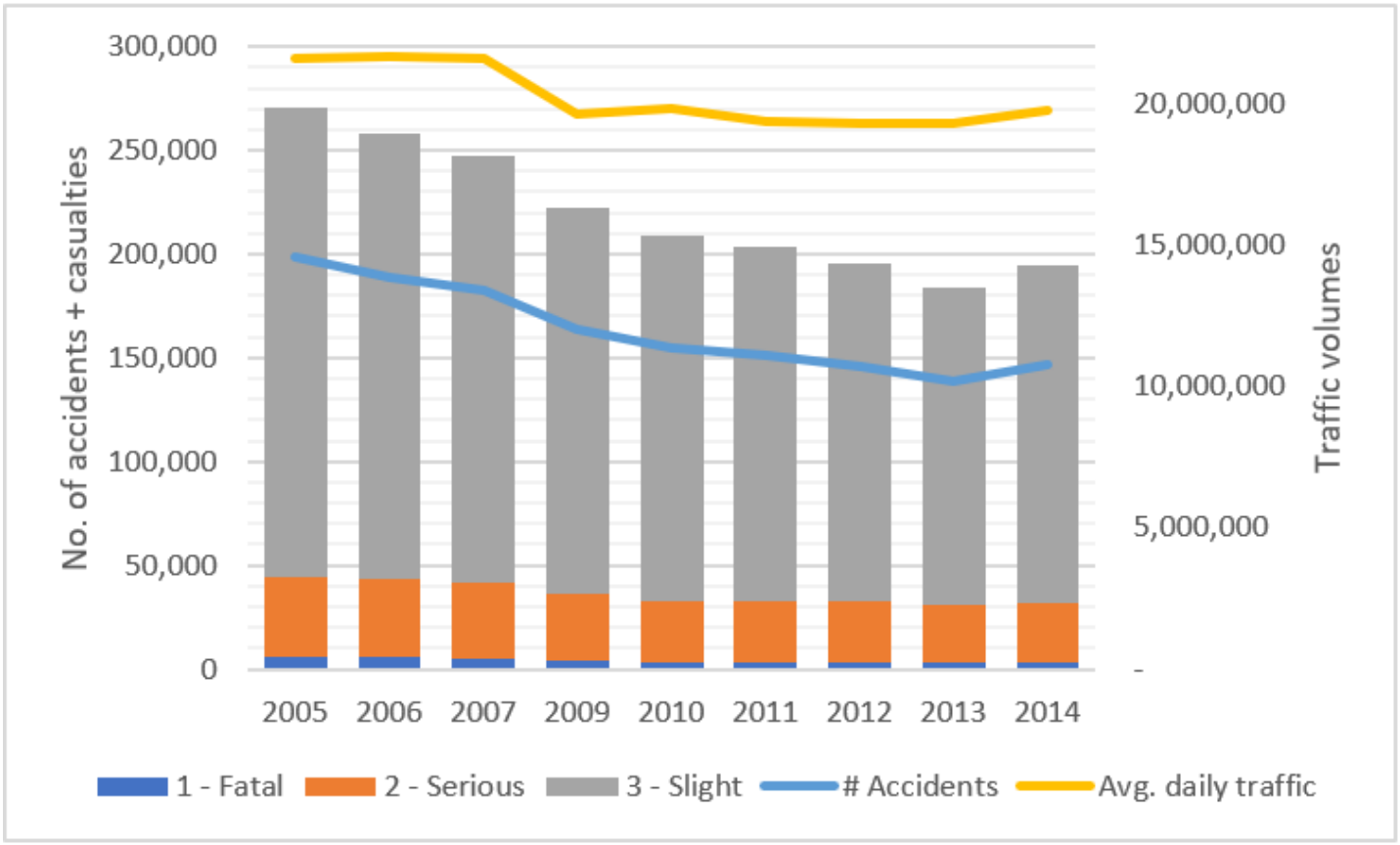


Figure 1: Trend of total accident casualties (by accident severity) and traffic volumes

Traffic volumes fell 8% between 2005 and 2014. The most significant reduction was in 2009 (-9%), without further detail this is presumably due to the economic downturn. During the same period (2005 to 2014), the number of accidents fell 26% and the number of casualties fell 28%.

The most significant reduction, in percentage terms, in accident casualties is observed in accident severity category 1 (fatal). There was a 50% reduction between 2005 and 2014. In total, there were 2,898 fatal casualties (1.5% of all casualties in 2014.

However, there has been an upward trend in traffic volumes since 2013. This trend has continued into 2015 and 2016 (not shown in Figure 1). Both accident severity categories 2 (serious) and 3 (slight) have experienced increases in

casualty numbers in 2014 (up 5% and 6%, respectively). It seems likely there will be a higher number of casualties in 2015 (this is investigated in Section 4).

The accident data was reviewed to understand the most likely conditions under which an accident would occur (as per objective 2 in Section 1.4).

The accident and vehicles data files were analysed to better understand common features associated with accidents in the UK. The following information was discerned from the data.

General

- Approximately 1.8 vehicles were involved in an accident, resulting in 1.3 casualties (on average). This suggests that a significant number of single vehicle, single occupant accidents occurred.
- Saturday was the most common day for accidents.
- There is an urban to rural accident ratio of 1.72.
- 17:00 was the most common time for an accident (1%), followed by 17:30 (0.9%).

Site Conditions

- 30 MPH was the most common speed at which accidents occurred.
- 75% occurred on single carriageways, with 15% on dual carriageways.
- 73% occurred during daylight hours.
- 80% occurred in normal weather conditions.
- 69% occurred when the road surface was dry.
- 98% occurred where there were no special road conditions e.g. roadworks, oil or mud on road.
- 98% occurred where there were no carriageway hazards.
- 99% had not pedestrian crossing within 50 m of the accident.

Vehicles[5]

- 76% of vehicles were cars, with 7.4% motorcycles and 6.2% bicycles.
- The typical engine size was 1,800 CC.
- The average vehicle age was 7.1 years
- 47% of vehicles were moving straight ahead when an accident occurred; 10% were turning right.

Drivers

[5] Of the 1.47 million accidents that occurred between 2005 and 2014, almost 3 million vehicles were involved.

- 72% of drivers were male.
- The average driver age was 38.7.

2.5.2 Relationship analysis

2.5.2.1 Covariance

The covariance pairwise tables are shown in Table 10 (positive) and Table 11 (negative).

Table 10: Ten highest 'positive' covariances

Attribute 1	Attribute 2	Covariance
1st_Road_Number	2nd_Road_Number	479450.59535350744
Age_of_Driver	Engine_Capacity_(CC)	2529.4670235468075
Engine_Capacity_(CC)	Speed_limit	1199.3232464071666
Police_Force	2nd_Road_Number	891.5280844618181
Police_Force	1st_Road_Number	766.9069019140735
2nd_Road_Class	2nd_Road_Number	582.5637139245904
1st_Road_Number	2nd_Road_Class	298.5555253097407
Latitude	1st_Road_Number	102.84973590061605
Age_of_Vehicle	1st_Road_Number	92.49243632113172
Police_Force	Speed_limit	92.1921664159574

Some pairs have a strong positive covariance (1st road number/2nd road number, driver age/engine capacity and engine capacity/speed limit), from which we can surmise that:

- road numbers are area specific i.e. high 1st road numbers in an area correspond with high 2nd road numbers at junctions,
- younger drivers tend to driver less powerful (smaller engine size) vehicles, and
- larger vehicles (e.g. HGVs) tend to drive on roads with higher speed limits (i.e. motorways).

Table 11: Ten highest 'negative' covariances

Attribute 1	Attribute 2	Covariance
Engine_Capacity_(CC)	1st_Road_Number	-42452.10820484739
Engine_Capacity_(CC)	2nd_Road_Number	-11536.748838371226
1st_Road_Number	Speed_limit	-1931.807487251036
Speed_limit	2nd_Road_Number	-979.3784116832256
Engine_Capacity_(CC)	Age_of_Vehicle	-351.38178877336003
1st_Road_Class	1st_Road_Number	-245.87059264747984
Engine_Capacity_(CC)	2nd_Road_Class	-233.08115254169218
Longitude	1st_Road_Number	-220.43430789777102
Age_of_Driver	1st_Road_Number	-161.9601044763916
Engine_Capacity_(CC)	1st_Road_Class	-150.15385751775847

A number of pairs have strong negative covariance (such as engine capacity/1st road number, engine capacity/2nd road number, 1st road number/speed limit and speed limit/2nd road number). The large negative covariance tells us that there is an inverse relationship between these attributes and suggests that as engine capacity and speed limits increase, the road number decreases.

However, the issue with covariance is that it can take any numerical value; because of this it is very sensitive to scale. If two attributes have different ranges (such as engine capacity and first road number), then the value is very difficult to interpret.

Charts showing examples of pairs of attributes with strong covariance (both positive and negative) are shown in **Appendix B**. Due to the number of attributes and volume of data in the dataset, we did not have sufficient memory to generate scatter matrix plots for all covariance pairs.

2.5.2.2 Correlation

Correlation is more useful at determining how strong the relationship is between pairs of attributes due to its small numerical range (-1 to 1). Correlations were generated in RapidMiner, as shown in Figure 2. The correlation pairwise tables are shown in Table 12 (positive) and Table 13 (negative). If a strong relationship does exist between 'predictor' pairs, we can remove these as part of the feature selection process. These attributes need to be removed because:

- they do not add additional information to the model, and
- can result in unstable or counter-intuitive estimates of model coefficients (multi-collinearity).

However, attributes that are strongly correlated with dependent (label) attributes will be retained.

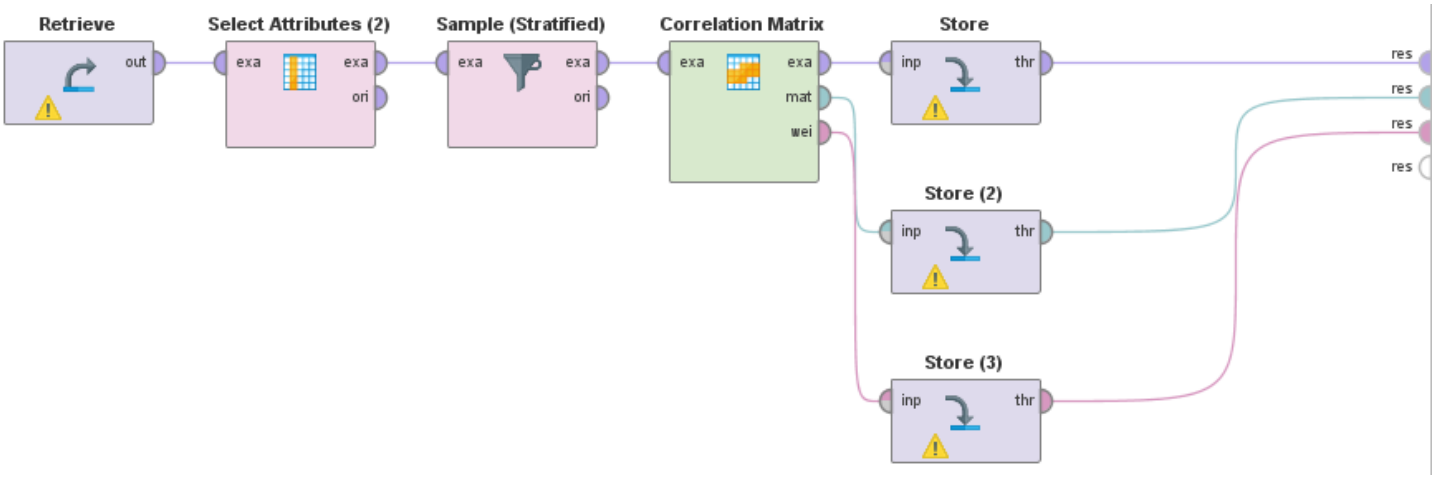


Figure 2: RapidMiner Correlation Process

Table 12: Ten highest 'positive' correlations

Attribute 1	Attribute 2	Correlation
Accident_Index	Year	0.99075
Year	Date	0.99026
Accident_Index	Date	0.98577
Local_Authority	Region	0.96211
Region	Police_Force	0.95517
Local_Authority	Police_Force	0.91115
Junction_Location	2nd_Road_Class	0.72561
Country	Police_Force	0.69213
Urban_or_Rural_Area	Speed_limit	0.68017
Vehicle_Reference	Number_of_Vehicles	0.60060

Several pairs in Table 12 have strong positive relationships within each other. As the majority of these are predictor attributes, the following (where r >= ± 0.7) can be removed from the modelling process:

- Date
- Local Authority
- Police Force
- 2nd Road Class
- Country

While it makes sense to remove either Year or Accident Index (as they are highly correlated), Year will be retained for modelling purposes and Accident Index will be retained as the ID for each row in the dataset (but will not be used in the predictive models).

Table 13: Ten highest 'negative' correlations

Attribute 1	Attribute 2	Correlation
Country	Longitude	-0.54163
Local_Authority	Longitude	-0.44798
Longitude	Latitude	-0.44367
Longitude	Police_Force	-0.44277
Region	Longitude	-0.41468
1st_Road_Class	Speed_limit	-0.39362
Speed_limit	2nd_Road_Class	-0.35088
Urban_or_Rural_Area	2nd_Road_Class	-0.29239
Junction_Location	Speed_limit	-0.24492
Urban_or_Rural_Area	1st_Road_Class	-0.22010

Table 13 shows correlated negative attribute pairs in the dataset. As $r < -0.7$ in all cases, none of the pairs can be classified as "highly" correlated; none can be removed for modelling purposes.

The distribution of pairwise correlations for each is shown in Figure 3.

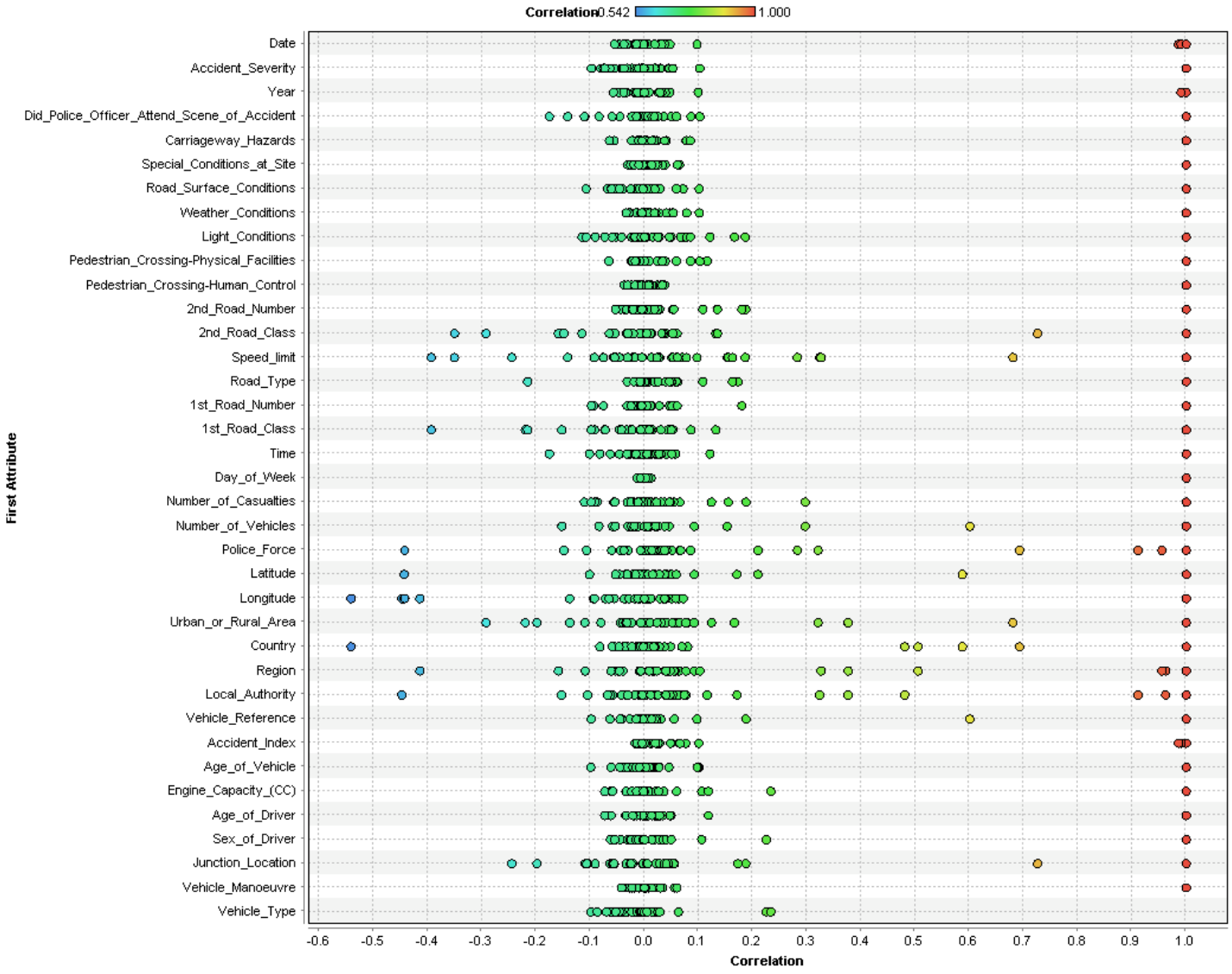


Figure 3: Distribution of pairwise correlations for each attribute

Charts showing examples of pairs of attributes with strong correlation (both positive and negative) are shown in **Appendix C.** Due to the number of attributes and volume of data in the dataset, we did not have sufficient memory to generate scatter matrix plots for all correlation pairs.

We also used R to generate correlation matrix plots for the dataset. A number of additional variables were removed prior to undertaking correlation analysis in R (the reasons for removing are discussed in Section 4.1.)

The correlation plot for the remaining variables is shown in Figure 4.

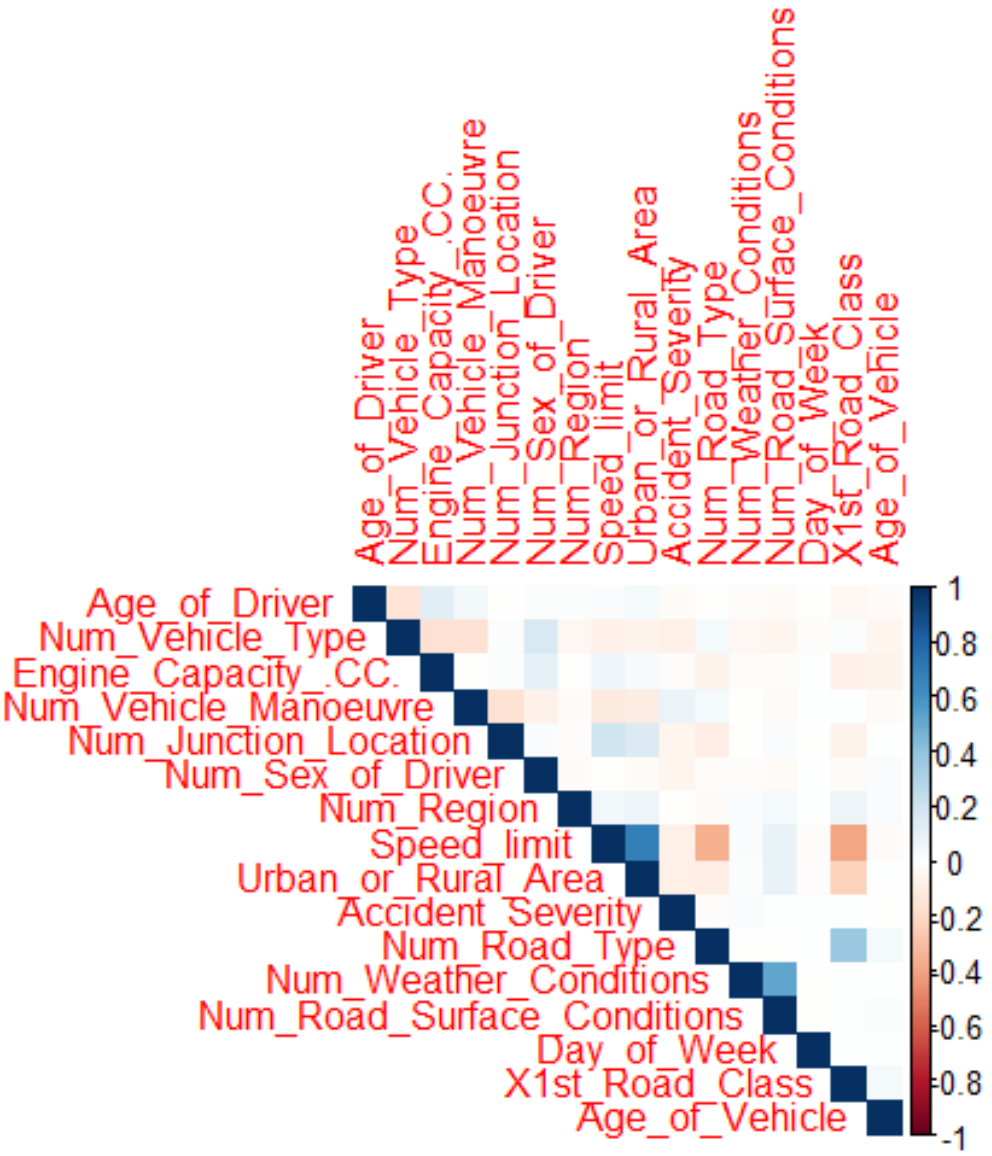


Figure 4: Correlation plot in R

The correlation pairs match the results shown in the RapidMiner process, with speed limit and rural (r = 0.68) returning the highest correlation for the remaining attributes. R allows you to combine correlation plots with significant tests, which removes matching pairs that are not highly correlated (as shown in Figure 5). This allows the user to easily identify and analyse highly correlated variables.

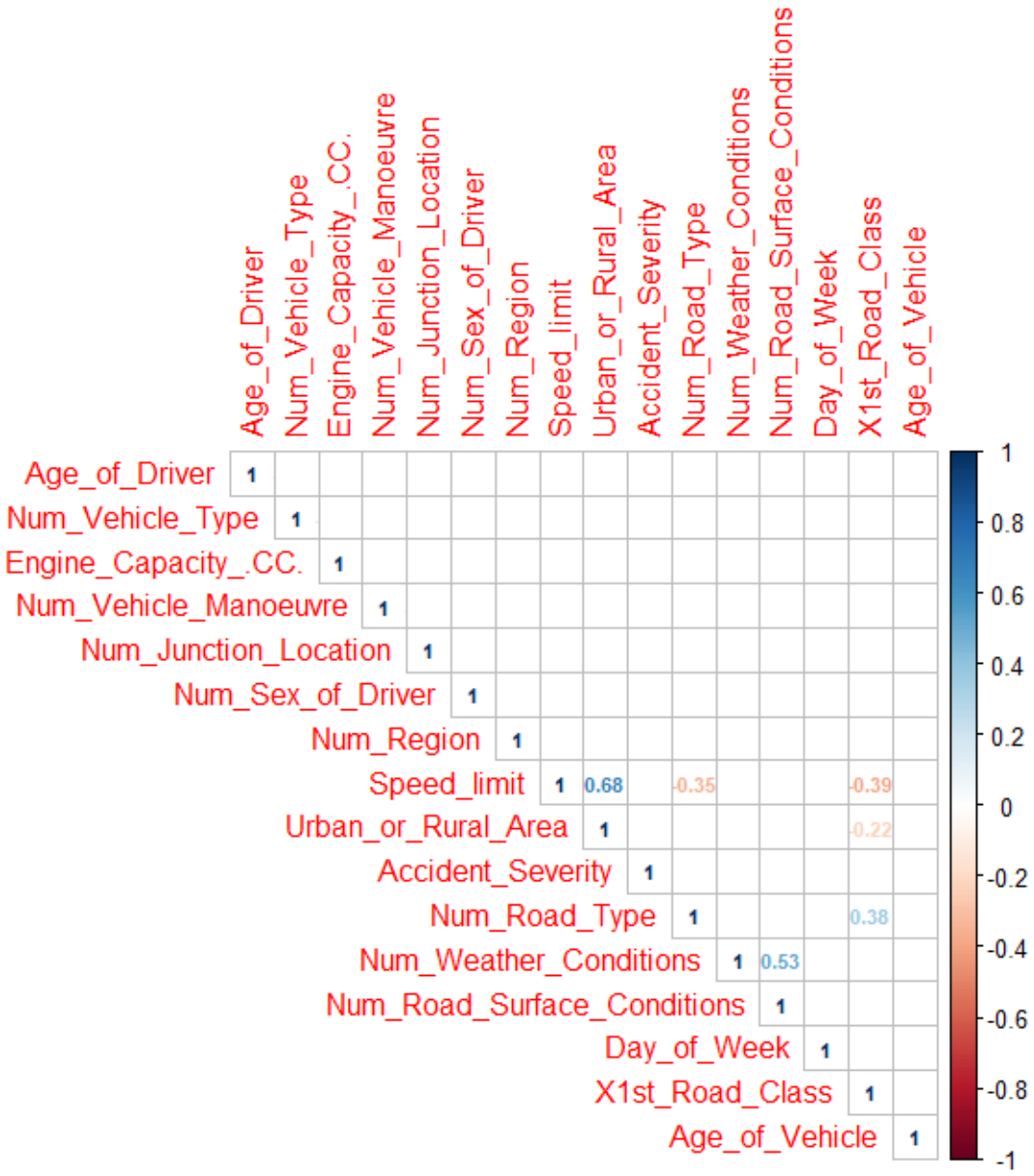


Figure 5: Correlation plot in R following test for significance

3 Data preparation

3.1 Selecting data

The files selected for the assignment are detailed in Section 2.1 and 2.2. In summary, there are five files (3x accident files, 1x traffic file and 1x vehicle file). A separate cleaning process has been prepared for each, as shown in Figure 6.

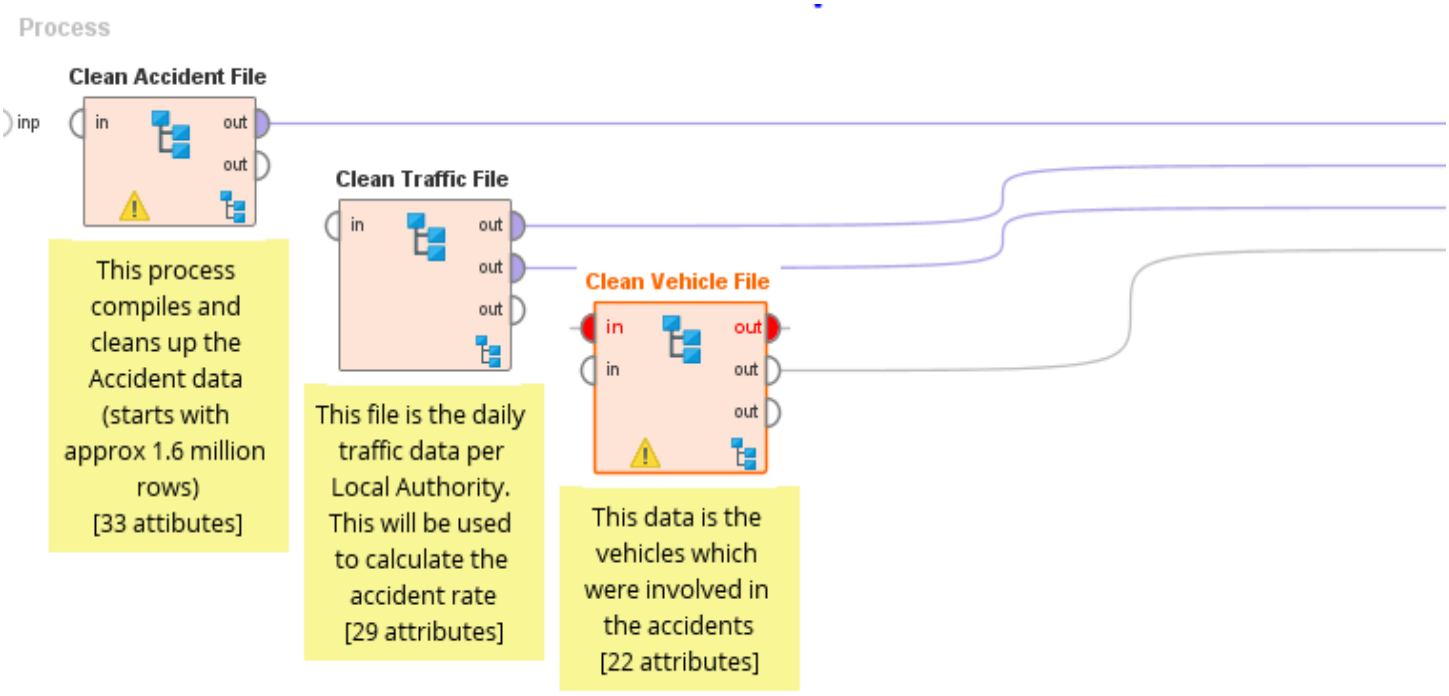


Figure 6: Separate cleaning process for each file

RapidMiner was chosen as tool of choice for the cleaning process due to its ability to easily read in CSV data and its vast library of operators available for cleaning data. Each process involved three main steps:

- Reading data
- Cleaning data
- Preparing data

3.2 Data cleaning

3.2.1 Accident files

Data preparation is a very important process for a data scientist. Very often data needs to be located in one place, and can require complex cleaning and transformation steps to achieve this. This section sets out the cleaning process we went through for the assignment.

As mentioned above, there are three accident files (split into three separate time periods). The first step in the process was to merge these files.

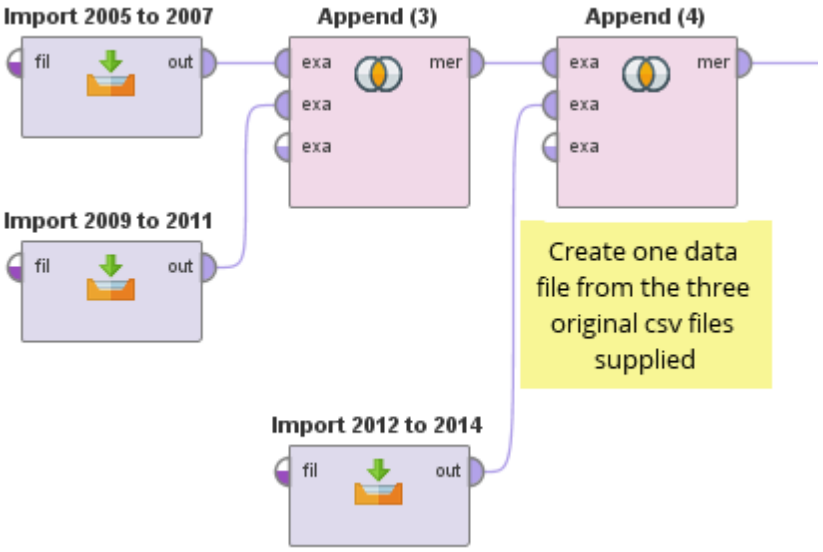


Figure 7: Merging accident files

Once merged, attributes which were highlighted for removal in the data quality report (Section 0) were omitted, namely Junction Control, Junction Detail, LSOA of Accident Location, Local Authority (District), Local Easting and Local Northing.

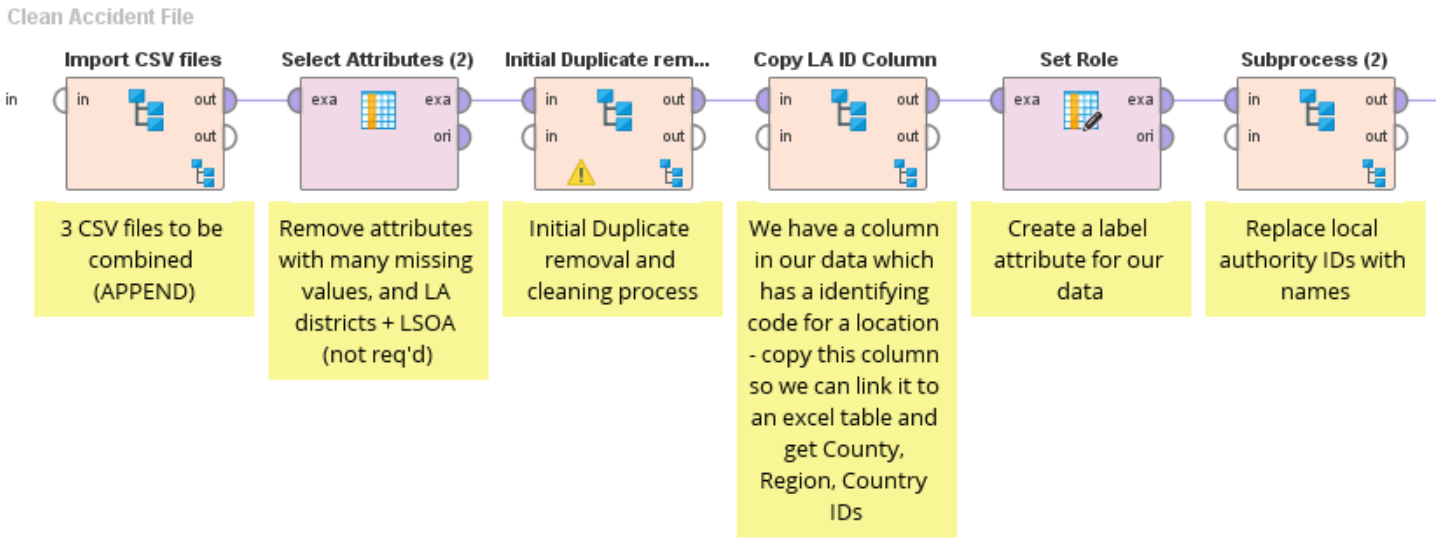


Figure 8: Accident cleaning process (part 1)

Duplicate values were also removed from the data, and missing values were imputed with average values for each attribute selected, as shown in Figure 9. We have also replaced the '3' values with a representative value in Urban/Rural field; it is binomial attribute and cannot have more than 2 values.

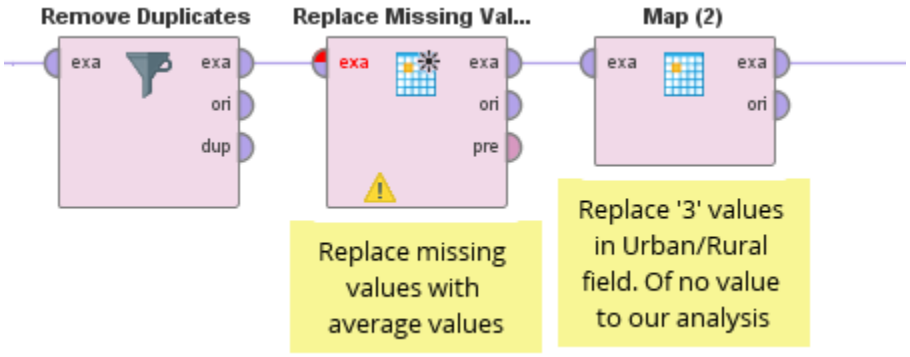


Figure 9: Duplicates and missing values

The file contains a column, Local Authority (Highway), which holds a numerical identifier for the local authority in which each accident occurred. We sourced additional data online which gave the matching local authority name, region, and country for each identifier. To include region and country details in the dataset, new attributes were generated (see Figure 10).

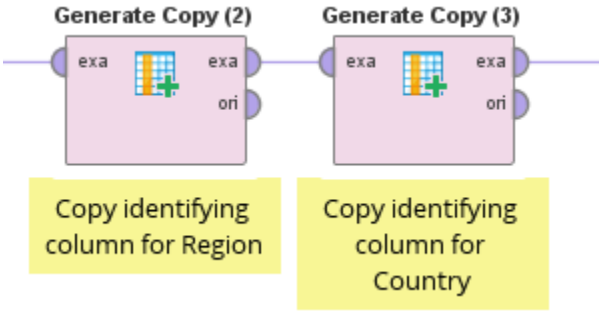


Figure 10: Generating 'Region' and 'County' attributes

The Set Role operator was used to create a label attribute for our data; Accident Severity was chosen . This operator will overwrite any existing attribute with the same role.

Sub-process 1, 2 and 3 are used to replace the local authority identifier key in the Local Authority, Region and Country columns. An example of the process adopted in shown in Figure 11. Two replace operators are used in each sub-process because some local authorities are district councils and other city councils.

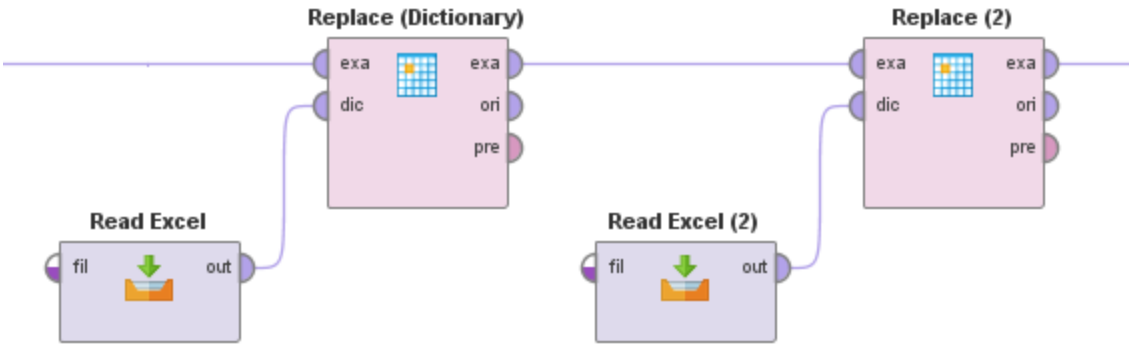


Figure 11: Example of sub-process used to replace local authority identity keys

The second part of the cleaning process for the accident file is shown in Figure 12.

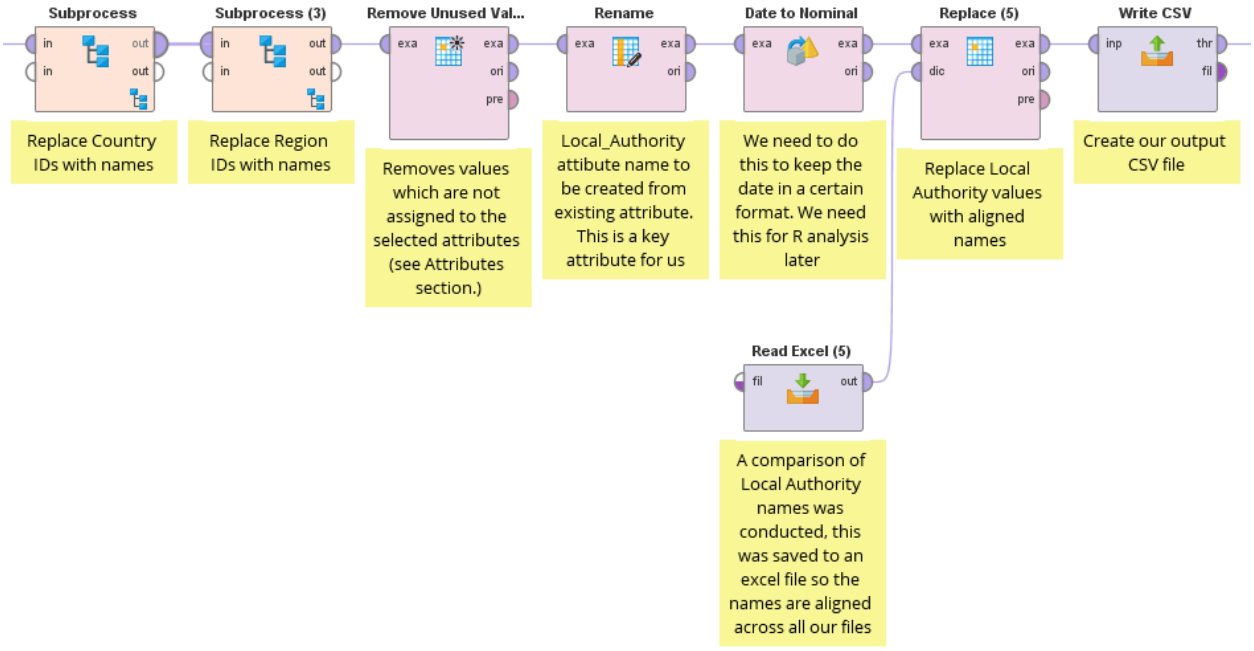


Figure 12: Accident cleaning process (part 2)

Unused values which were assigned in the previous step (sub-process 1, 2 and 3) are removed. The Local Authority attribute is renamed for consistency with other files, and a number of local authority names are replaced (again for consistency with other files).

The Date to Nominal process was used convert all dates to a specific format. This was required so that the data can then be analysed in R at a later stage.

Once all cleaning operations are completed, a single, cleaned accident file is saved as a new CSV file (merged_accidents_clean.csv).

3.2.2 Traffic file

The traffic file follows a similar cleaning process to the accident file. An overview of the cleaning operations undertaken is shown in Figure 13 and Figure 14.

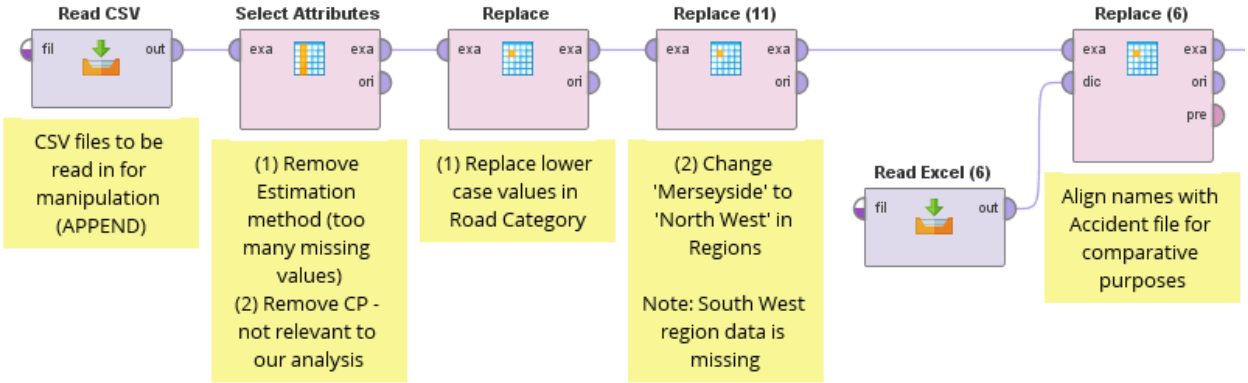


Figure 13: Traffic cleaning process (Part 1)

Attributes which were highlighted for removal in the data quality report (Section 0) were not selected. There were some spelling errors in the Road Category attribute, which resulted in more categories than was intended. This was rectified using the Replace operator.

For the Region attribute, 'Merseyside' was replaced with 'North West' to ensure consistency with the accident file regions. (Merseyside was previously classified as a region in England, but was later changed to the North West region.) We also noted that data for the South West region was missing.

A number of local authority names are replaced (for consistency with other files) using the Replace (6) operator.

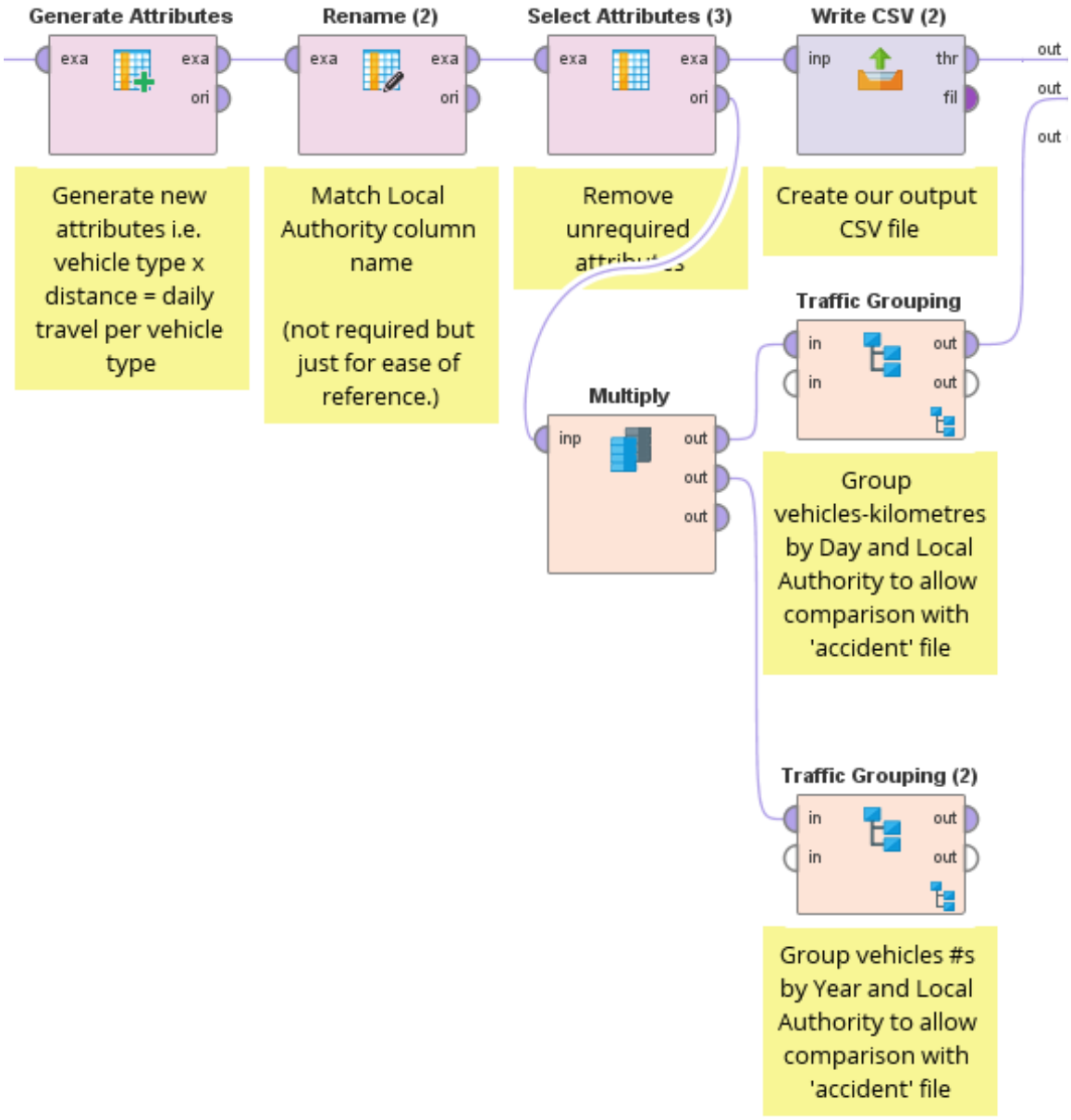


Figure 14: Traffic data cleaning process (Part 2)

A number of new fields were created using the Generate Attributes operator to show the total distance travelled each day by different vehicle types. A second select attribute operator was included at a later stage to remove the sub-categories of heavy goods vehicles in the dataset.

Finally, the cleaned data was saved as a new CSV file (ukTrafficAADF_clean.csv).

An additional step was added at the end of the process to aggregate vehicle numbers and distance travelled by vehicle type on a yearly basis. This information will be used to analyse traffic trends relatively to the number of accidents that occur annually.

3.2.3 Vehicle file

An overview of the vehicle cleaning process is shown in Figure 15. The process follows the sample philosophy as the previous cleaning processes (accidents and traffic).

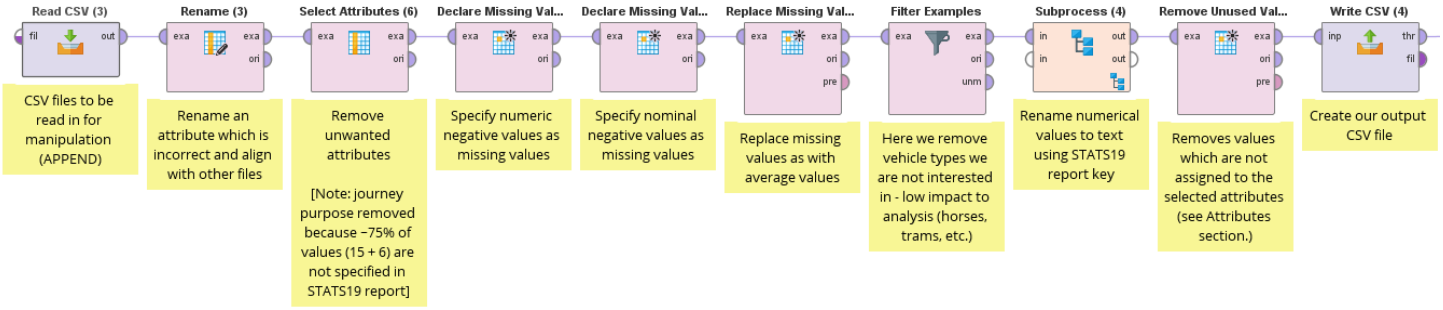


Figure 15: Vehicle cleaning process

The Rename operator is used to correct a misspelling in the Accident Index attribute. Attributes which were highlighted for removal in the data quality report (Section 0) were not selected.

A lot of the attributes had negative numeric values (possibly a default entry in the police system where a value was not recorded). To overcome this issue, we used the Declare Missing operator to instruct RapidMiner that negative values were in fact missing values (for both numeric and categorical data). The missing values were then imputed with average values (as previously described in the accident cleaning process).

A significant number of attributes held numeric inputs (e.g. vehicle type used a reference system from 1 to 6). While this is good practice in database storage systems, it does not inform us which vehicles were most commonly involved in accidents. Thus, we acquired the STATS19 police report form[6] and replaced numeric values with text values (names).

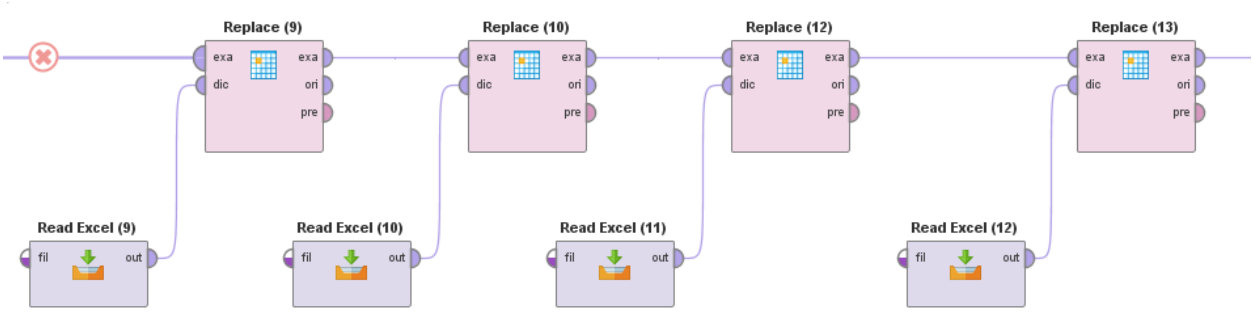


Figure 16: Replacing numeric values with text from STATS19 police report form

[6] http://docs.adrn.ac.uk/888043/mrdoc/pdf/888043_stats19-road-accident-injury-statistics-report-form.pdf

Four attributes were replaced in the above operation: Junction Location, Vehicle Manoeuvre, Vehicle Type and Driver Gender. Unused value were removed (as described in the accident file).

The cleaned data was saved as a new CSV file (vehicles_clean.csv).

3.3 Data construction

Several constructions were undertaken as part of the data preparation process:

- Two new fields were generated in the accident file: Region and Country (see Section 3.2.1);
- Distance travelled by vehicle type fields were calculated and created in the traffic file (see Section 3.2.2);
- Missing values were imputed with average values in all files (unless an attribute had too many missing values; in this case it was removed).

3.4 Data integration

As discussed in Section 3.2.1, the accident files were originally three separate files. These were merged as part of the cleaning process. Several other external sources were integrated into the different datasets during the cleaning process:

- local authority ID codes were replaced with local authority, region and country names in the accidents file, and
- numeric values in the vehicles file were replaced with names (text) using the STATS19 report form.

The accidents file was joined to the vehicles file prior to undertaking the exploratory analysis (Section 2.5) and modelling (Section 4).

3.5 Data formatting

Other than replacing numeric values with text values (as described in this section of the report) no other formatting was undertaken.

4 Model Building

4.1 Select modelling technique

The basis for accurate modelling is to ensure that only relevant attributes remain in the dataset. Based on the reviews conducted in Section 2.3, Table 2 and Section 2.5.2.2 (Correlation) any attributes remaining in the dataset which had been deemed to be irrelevant to the analysis were removed prior to beginning the modelling process.

The remaining attributes were reviewed to assess their value to the process; it was found that a number of them had an overwhelming amount of a single value i.e. 85% ≥ of the attribute values were of a single denomination. We determined that these attributes should be removed from the dataset. As the majority of attributes remaining were categorical, the focus would be on classification models in order to build a predictive model (to meet objective 3 of Section 1.4).

The following assumptions were made about our data prior to modelling (see also Table 3):

1. There are no missing values → missing values were handled in the data preparation stage
2. All attributes are categorical → the remaining numerical attributes were transformed to polynomials e.g. Speed Limits, Age and Engine Capacity. Each were converted to polynomial ranges for analysis purposes.
3. All values are reasonable → all data ranges were reviewed to ensure that the values within (e.g. Engine Capacity, values greater than 15,000cc were filtered as they were determined to be outliers.)
4. Irrelevant and duplicate features (attributes) have been removed.

From the correlation and data review process outlined in previous section, a number of variables remaining in the dataset were highlighted to be removed (see Figure 17). The attributes to be removed were selected, following which the 'invert selection' option was ticked to ensure only those attributes which were considered relevant to our model remained.

The attributes selected for removal were:

- 1st Road Number – this is a locator ID, not relevant to our analysis
- 2nd Road Class – highly correlated to Junction Location which remains in the dataset
- 2nd Road Number – this is a locator ID, not relevant to our analysis
- Accident Index – individual accident ID, not relevant to our analysis
- Carriageway Hazards – 98% single value
- Country – highly correlated with Police Force and Region
- Did Police Officer Attend Scene of Accident – Police attending after an accident has occurred is not relevant to an accident occurring
- Latitude – this is a locator ID, not relevant to our analysis
- Local Authority – highly correlated to Region which remains in the dataset
- Longitude – this is a locator ID, not relevant to our analysis

- Number of Casualties – not considered to be a key attribute to cause an accident, suggestion that this be included in future analysis to determine the effect of its inclusion
- Number of Vehicles – not considered to be a key attribute to cause an accident, suggestion that this be included in future analysis to determine the effect of its inclusion
- Pedestrian Crossing-Human Control – 99% single value
- Pedestrian Crossing-Physical Facilities – 82% single value
- Police Force – highly correlated to the Region which remains in the dataset
- Special Conditions at Site 97% single value
- Vehicle Reference – individual car ID, not relevant to our analysis
- Year – highly correlated to Date which remains in the dataset

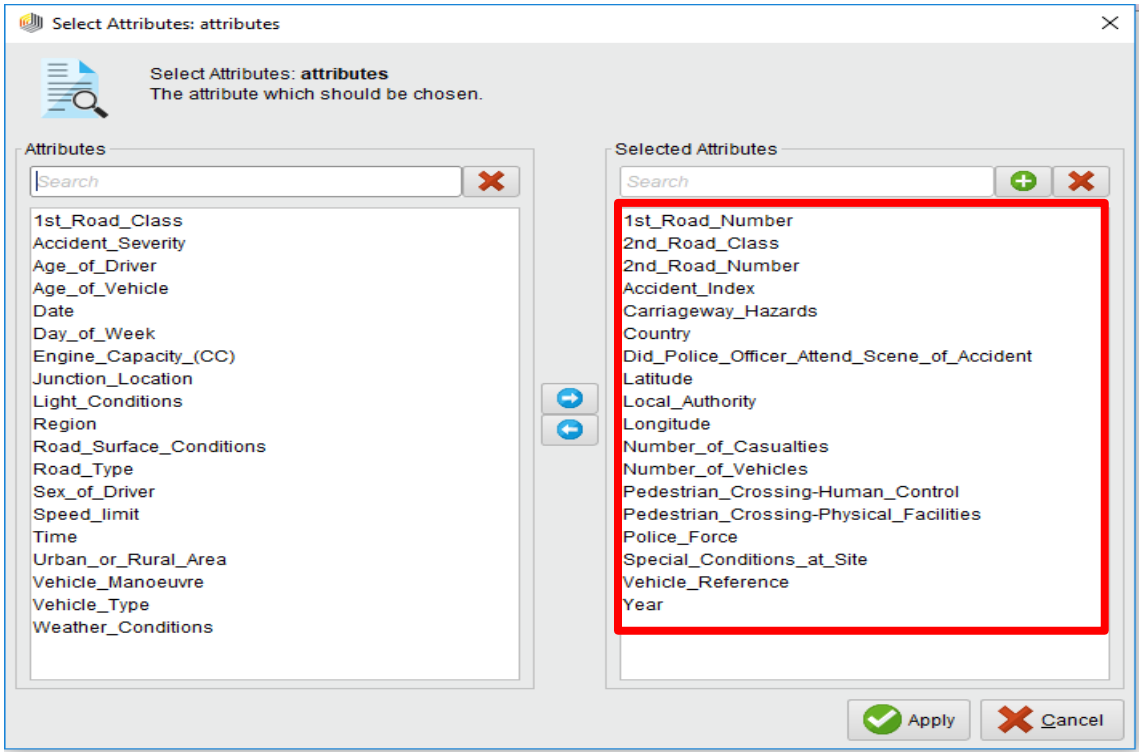


Figure 17: Attributes in dataset selected for removal before modelling

R was considered to create the model for the dataset; R code was written (refer to **Appendix D**). However, we found that RStudio allocates memory statically, and on running the R code in RStudio we were presented with an insufficient memory error ("*Error: cannot allocate vector of size 3141.2 Gb"*) owing to the size of our dataset. Thus, RapidMiner was selected as it allocates memory for the tasks it completes dynamically and was capable of completing the analysis we required.

4.2 Generate test design

Owing to system resource restrictions, a single Split Validation model was utilised for the analysis of the dataset. (While the added benefit of a Cross Validation sub-process would have been the preferred route, the time taken to complete even a single validation on the team's personal equipment was prohibitive.)

A 70% training and 30% test dataset split was determined to be optimal for the Split Validation.

A Performance module was added inside the Validation Process to evaluate models under the following criteria:

- Accuracy
- Error Factor(s)[7]
- Correlation

4.3 Build model

The process built in RapidMiner for modelling our dataset can be seen in Figure 18. The following settings were used:

- Discretize operator
    - Attribute filter type = subset [Age of Driver, Age of Vehicle, Engine Capacity, Time]
    - Number of Bins = 10
    - Range Name Type = Interval
- Numerical to Polynomial operator
    - Attribute filter type = subset [1st_Road_Class, Accident Severity, Date, Day of Week, Speed limit, Urban or Rural Area]
- *Set Role* operator
    - Attribute Name = Accident Severity
    - Target role = label
- Validation operator
    - Split = relative
    - Split Ratio = 0.7
    - Sampling Type = Stratified Sampling

[7] Error factors were determined as appropriate for the modelling task at hand.

- Local Random Seed = 1992

- Model Learning Process
  - Default settings as defined by RapidMiner were used for the below Prediction (classification) Modelling operators:
    - Decision Tree
    - Deep Learning
    - Generalized Linear Model
    - Naïve Bayes
    - Gradient Boosted Trees

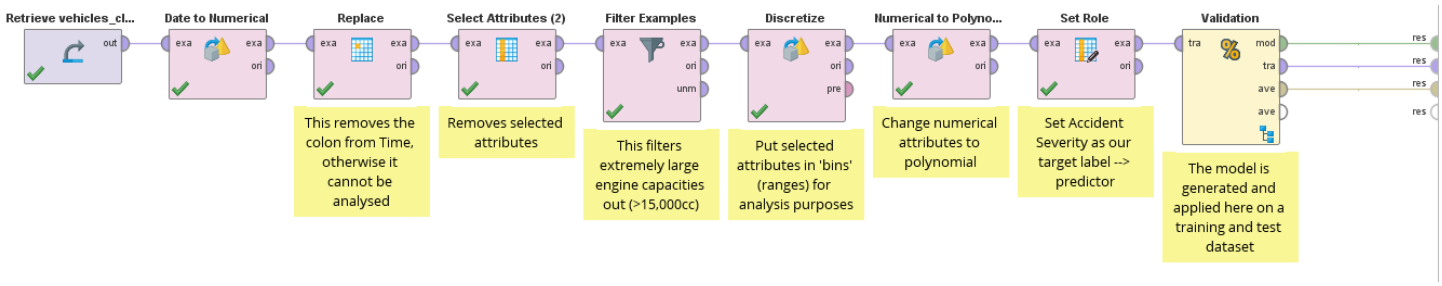


Figure 18: RapidMiner Data Modelling Process

Accident Severity was selected as our label as this is the output we wish to predict in our Test dataset. As outlined previously, Accident Severity is a polynomial attribute and consists of three categories (with the following proportions):

- Category 3 (Slight) = 87%
- Category 2 (Serious) = 12%
- Category 1 (Fatal) = 1%

Given the quantity of Category 3 Accident Severity values, it is likely that this could have a negative impact on our model. Modelling will proceed with the full dataset to assess if there is sufficient data to generate a good model with all data present.

Several ensemble models were also constructed; upon review it was noted that no measurable improvement in model performance was noted and as such they were not included in this report.

4.4 Assess model

The classification models outlined in Section 4.3 were executed and are compared in Table 14 below against the criteria listed in Section 4.2:

Table 14: Model Performance Comparison Table (Entire Dataset)

Label: Accident Severity	Decision Tree	Deep Learning	Generalized Linear Model	Naïve Bayes	Gradient Boosted Trees
Accuracy	86.86%	86.88%	86.87%	85.12%	86.87%
Normalized absolute error	2.122	2.016	2.027	2.158	2.402
Root mean squared error	0.345 +/- 0.000	0.338 +/- 0.000	0.339 +/- 0.000	0.361 +/- 0.000	0.353 +/- 0.000
Squared error	0.119 +/- 0.262	0.114 +/- 0.250	0.115 +/- 0.249	0.130 +/- 0.255	0.125 +/- 0.242
Correlation	0	0.034	0.023	0.022	0.026

The **Deep Learning model** was selected as the best performing model against our dataset based on all compared values.

accuracy: 86.87%

	true 2	true 3	true 1	class precision
pred. 2	147	107	30	51.76%
pred. 3	60277	436734	5637	86.89%
pred. 1	0	0	0	0.00%
class recall	0.24%	99.98%	0.00%	

Figure 19: Deep Learning Confusion Matrix

As discussed in Section 4.3, there is a preponderance of one category of Accident Severity (Category 3) in the dataset, the proportion of which closely corresponded with the accuracy readings achieved and seen in Table 14, and also in the confusion matrix in Figure 19. To overcome this inequality in representation it was deemed prudent to add a sampling operator into the process before the validation step. This Sampling operator was added to ensure that all levels were compared on an equal footing; as a result Category 2 and 3 values were undersampled, and Category 1 was oversampled, relative to the entire dataset. See Figure 20 for the updated process flow.

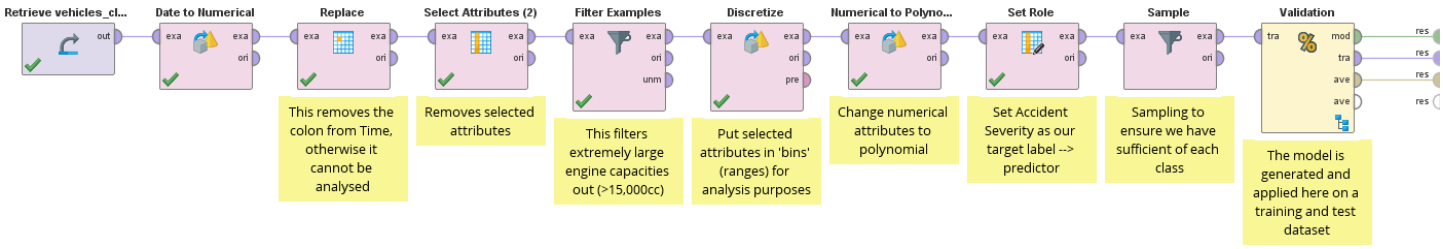


Figure 20: RapidMiner Data Modelling Process with Sampling

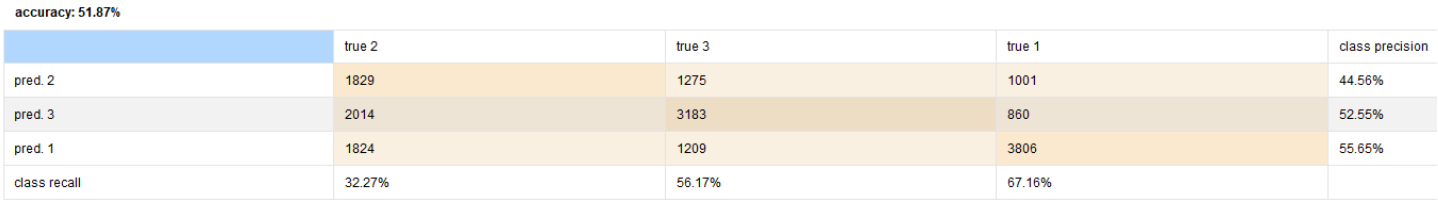
The classification models outlined in Section 4.3 were re-executed and compared in Table 15 against the criteria listed in Section 4.2:

Table 15: Model Performance Comparison Table (Oversampled Dataset)

Label: Accident Severity	Decision Tree	Deep Learning	Generalized Linear Model	Naïve Bayes	Gradient Boosted Trees
Accuracy	33.33%	50.57%	51.60%	50.16%	51.87%
Absolute error	0.667 +/- 0.000	0.575 +/- 0.194	0.583 +/- 0.170	0.542 +/- 0.284	0.602 +/- 0.130
Correlation	0.000	0.236	0.240	0.224	0.258

The Gradient Boosted Trees model was selected as the best performing model against our oversampled dataset based on Accuracy and Correlation values. The confusion matrix in Figure 21 also indicates that the model is now taking all three categories into consideration.

However, based on the absolute error (average absolute deviation of the prediction from the actual value) it could be considered that the Generalized Linear Model is also worthy of consideration as the best model for the dataset.



accuracy: 51.87%

	true 2	true 3	true 1	class precision
pred. 2	1829	1275	1001	44.56%
pred. 3	2014	3183	860	52.55%
pred. 1	1824	1209	3806	55.65%
class recall	32.27%	56.17%	67.16%	

Figure 21: Gradient Boosted Trees Confusion Matrix

This addresses objective 3 of Section 1.4: we can predict accident severity using other attributes with an accuracy of 51.87%.

4.5 Time series analysis and forecasting

4.5.1 Overview

We set out to determine if the number of accidents in 2016 could be predicted based on the time series data available for the years 2005 – 2014 (objective 4 in Section 1.4). Using both RapidMiner and R, we were able to predict the number of traffic accidents in the 2015 and 2016 periods. We acquired the 2015 and 2016 data (in csv format) to compare recorded data with forecasting results from the UK government site referenced in Kaggle.

4.5.2 Time series forecasting in RapidMiner

Time series forecasting can be undertaken in RapidMiner using a two phase data transformation approach: windowing and prediction. A forecasting process was built in RapidMiner and is shown in Figure 22.

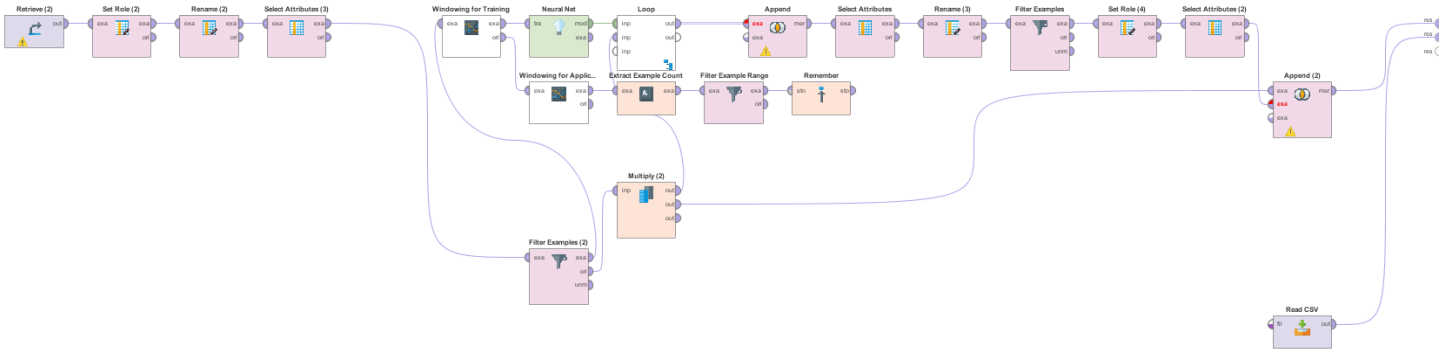


Figure 22: RapidMiner Time Series Forecasting Process

The output in RapidMiner is shown in Figure 23. The values for 2005 – 2014 are actual values of the number of accidents officially recorded for each year; the years 2015 and 2016 are the predicted figures.

ExampleSet (11 examples, 1 special attribute, 1 regular attribute)

Row No.	Year	AccCount
1	2005	198732
2	2006	189161
3	2007	182115
4	2009	163553
5	2010	154414
6	2011	151470
7	2012	145584
8	2013	138659
9	2014	146320
10	2015	147285.525
11	2016	138993.119

Figure 23: Forecast vales table for time series data

Table 16 shows the Predicted vs. Actual number of accidents for the two years we have predicted. The margin of error when comparing predicted and actual values for 2016 was 1.7%. (Forecasting prediction results in R are shown in Table 18.

Table 16: Predicted vs. Actual Accident Values

Year	Actual	Predicted	Margin of Error
2015	140,057	147,286	5.2%
2016	136,621	138,993	1.7%

Figure 24 shows the time series graph generated in RapidMiner. The red dotted line indicates where the existing dataset ends, and the extended line shows where the forecast has predicted the next two years will be. The red circle highlights the year 2016 i.e. our target year.

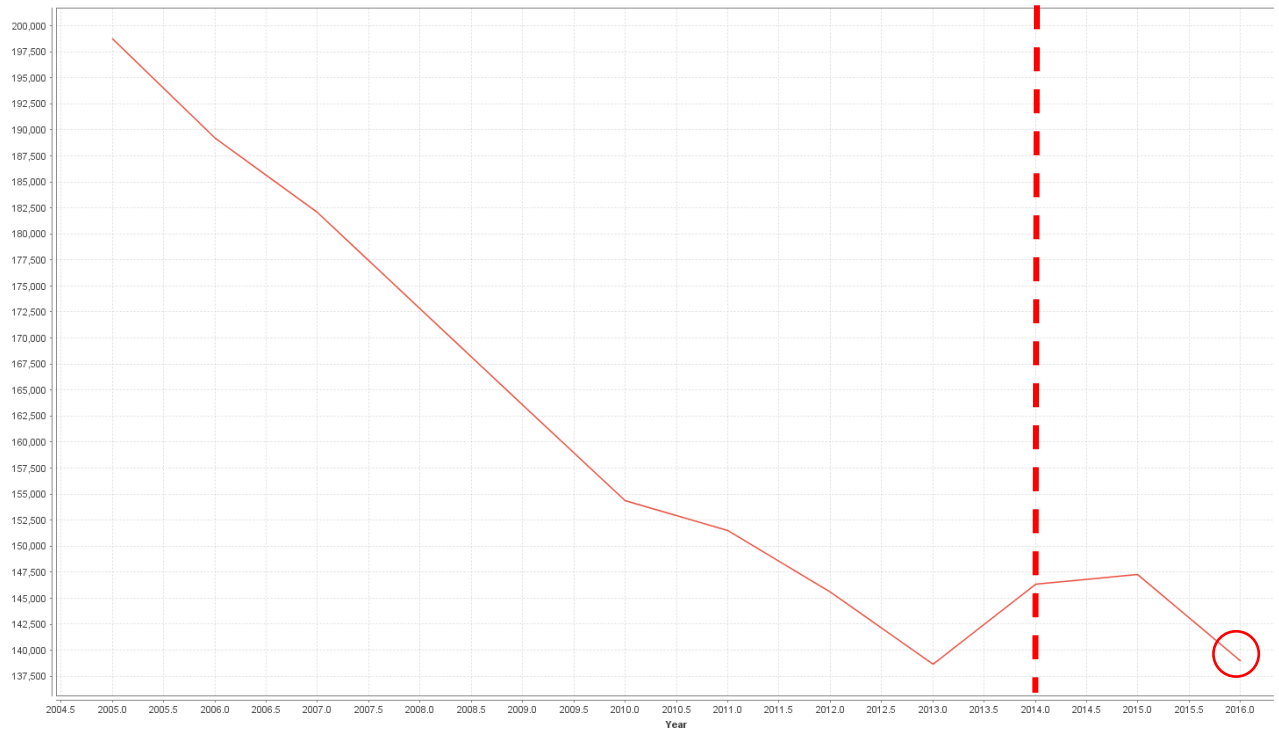


Figure 24: Time series forecast in RapidMiner

4.5.3 Time series forecasting in R

A time series for the number of accidents is shown in Figure 25.

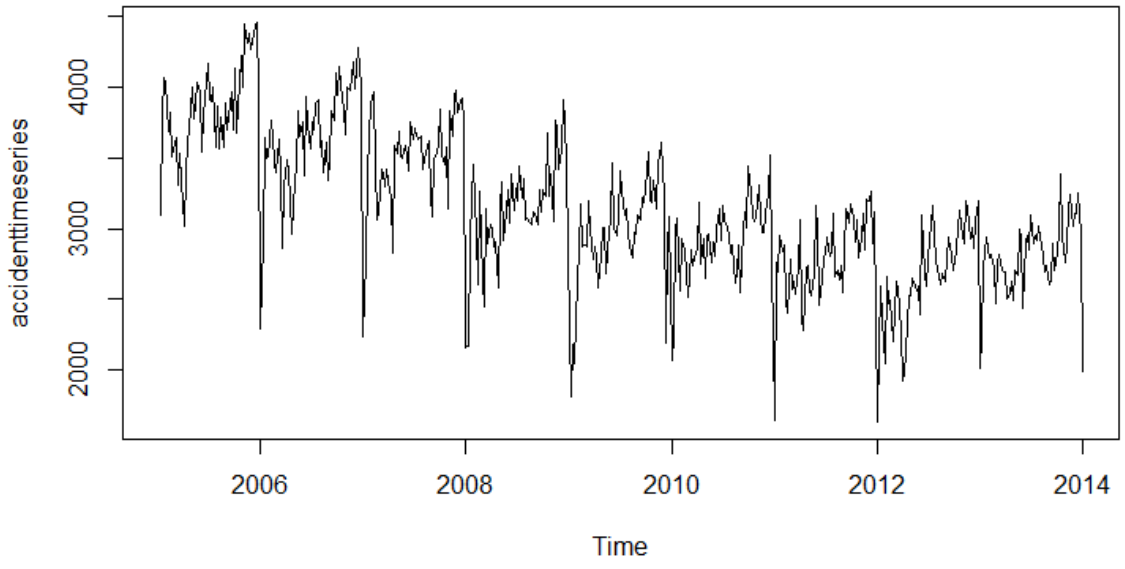


Figure 25: Time series of weekly accidents in UK

It appears from the chart that the time series is an additive model i.e. it has a trend, seasonal and random (noise) component. If you have a seasonal time series that can be described using an additive model, you can seasonally adjust the time series by estimating the seasonal component, and subtracting the estimated seasonal component from the original time series. R allows you to also decompose a time series into its constituent parts (as shown in Figure 26).

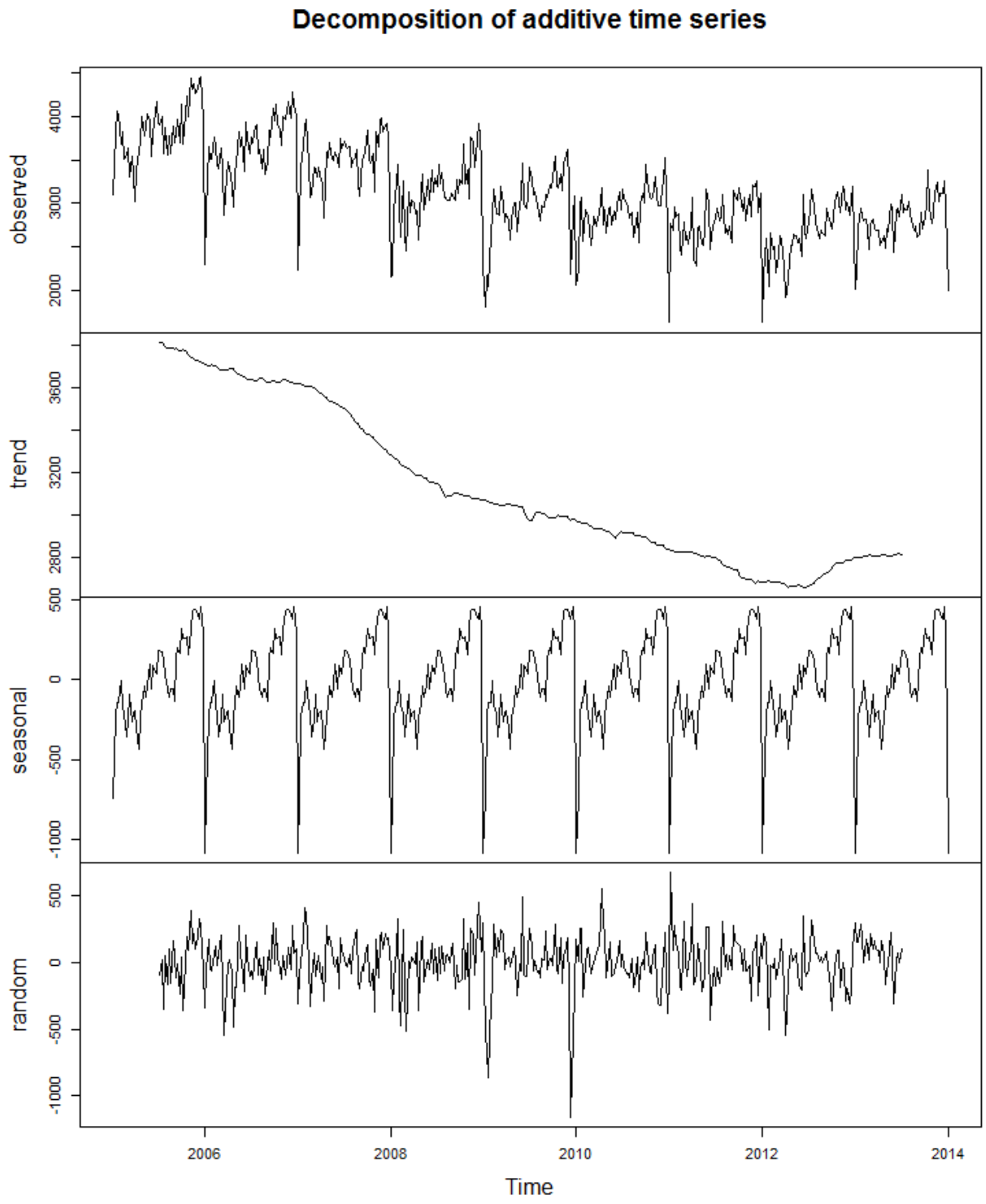


Figure 26: Decomposition of accidents time series

Several observations can be made about the time series:

- The number of accidents have been trending downwards from 2005 to 2013, but are showing indications of a rising trend in 2014.
- It appears that accidents in the UK are highly seasonal: there is a gradual increase at the start of every year, with two relatively minor drops early and midway through the year, followed by a sharp drop at the end of each year.
- There are a number of unexplained variations in the noise component (in 2009 and 2010).

R can also be used to make short-term forecasts for time series data. For comparative purposes with the results in Section 4.5.2, we will forecast the time series for 104 weeks (2 years – 2015 and 2016). While this would not be considered a "short-term" forecast, it will allow us to compare the two tools (RapidMiner and R).

The three forecasting techniques we will use are:

1. **Simple exponential smoothing**: common forecasting method which uses previous forecast value for given period to predict value for next period, but not really suited to long horizon forecasts as it does not consider trend and seasonality;
2. **Holt's two-parameter exponential smoothing:** extension of previous forecasting technique, it considers slope as well as average values;
3. **Holt**-Winters 3 parameter exponential smoothing: useful where a time series contains seasonality in addition to trend.

The forecasts for the three modelling techniques are shown in Figure 27 to Figure 29.

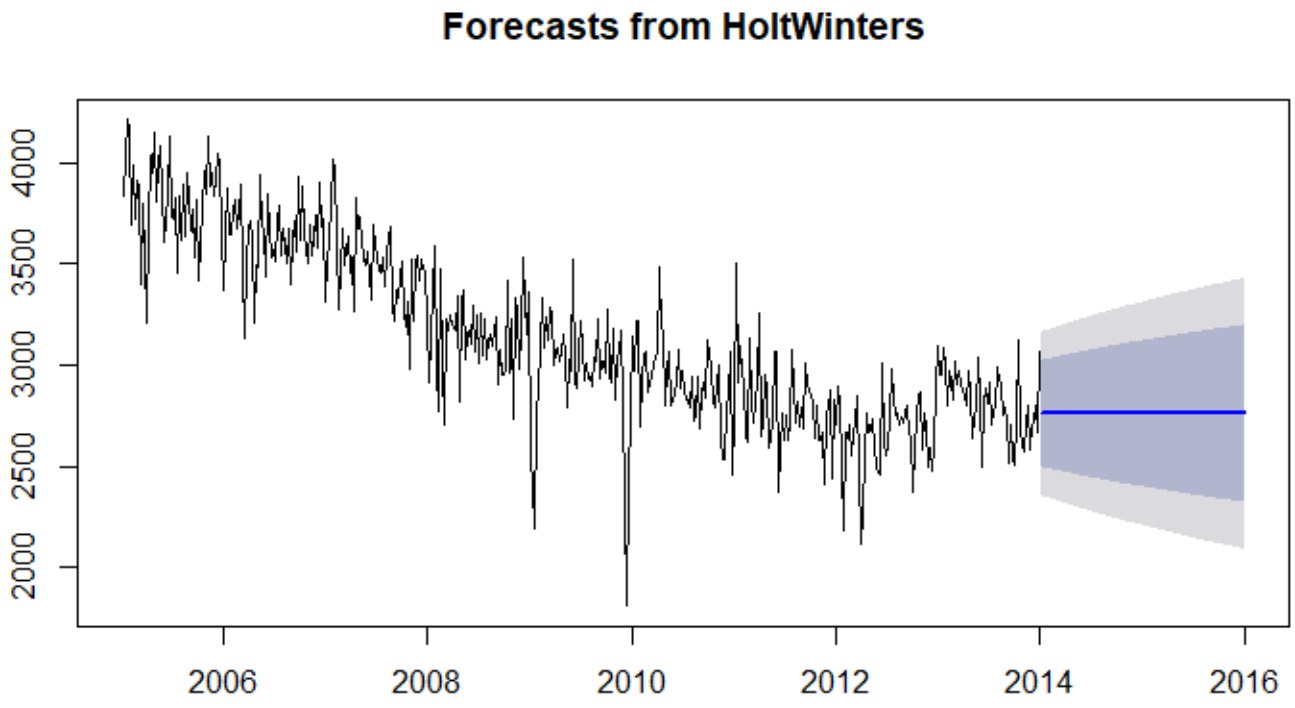


Figure 27: Simple exponential smoothing forecasting model

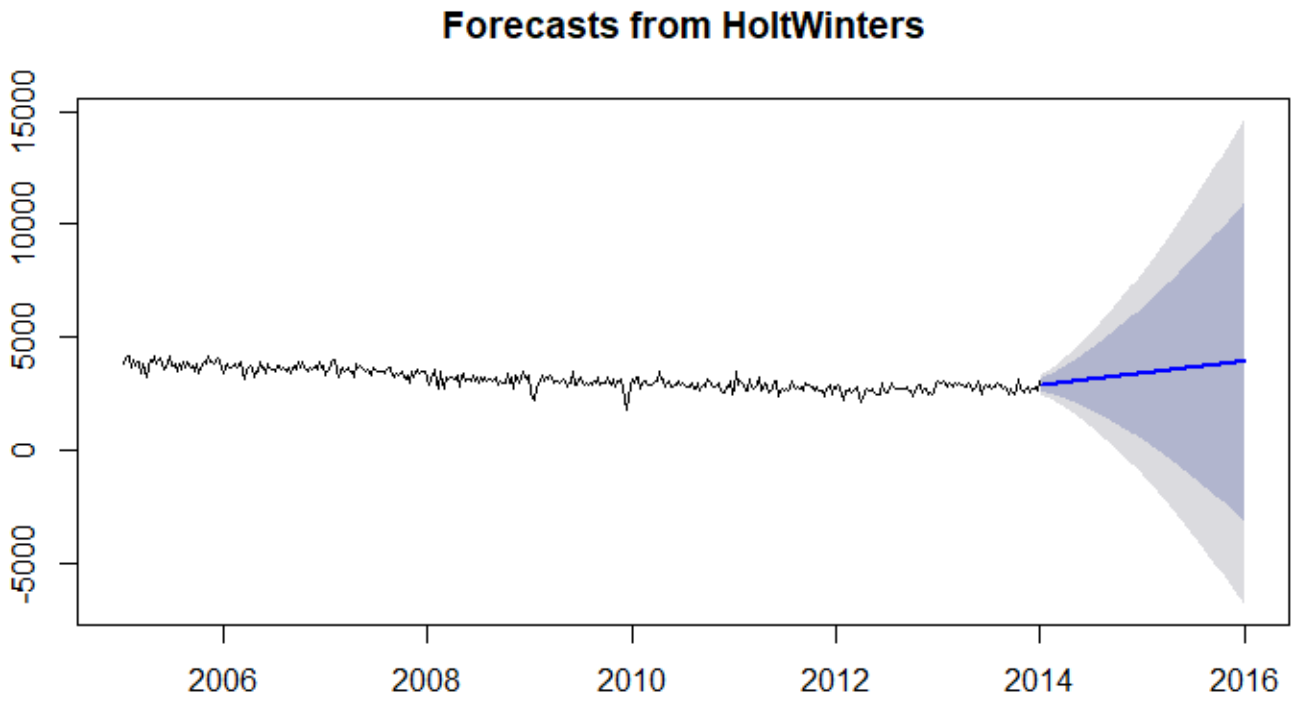


Figure 28: Holt' exponential smoothing forecasting model

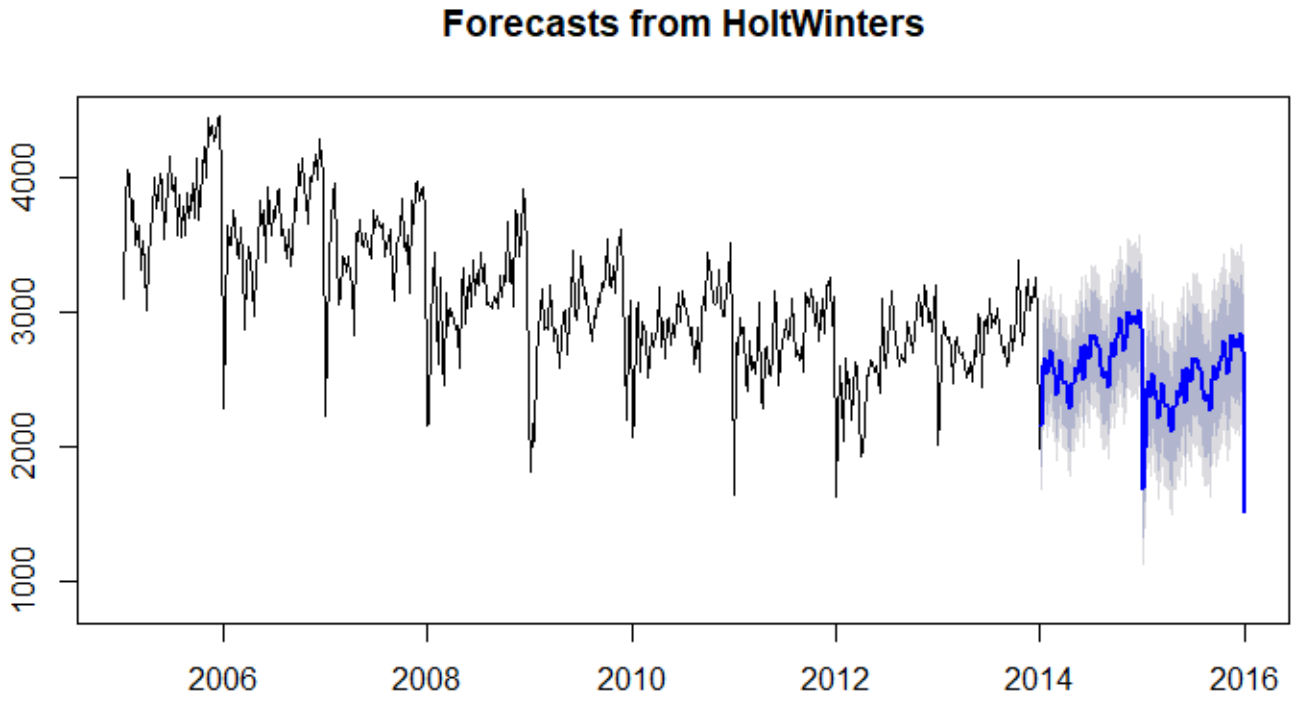


Figure 29: Holt-Winters exponential smoothing forecasting model

The models are evaluated using the sum of squared errors (SSE) and the Box-Ljung test statistic (as shown in Table 17).

Table 17: Evaluation of forecasting models in R

		Simple exponential smoothing	Holt's exponential smoothing	Holt-Winters exponential smoothing
Sum of squared errors (SSE)		19,916,304	24,066,092	25,074,592
Box-Ljung test (df = 20)	X-Squared	44.811	26.421	38.672
	p-value	0.00117	0.1524	0.007321

If the predictive model cannot be improved upon, there should be no correlations between forecast errors for successive predictions. In other words, if there are correlations between forecast errors for successive predictions, it is likely that the model could be improved upon by another forecasting technique. With the Box-Ljung test, a high p-value indicates that a model has low-auto-correlation (and vice versa).

Based on the results for Table 18 above, the results suggest the **Holt's exponential smoothing** model is the best fit, although it has a slightly higher SSE value than the simple exponential smoothing model. As we've already discussed, this technique does not consider trend or seasonality, so it is unlikely to be a reliable model.

A correlogram (Figure 30) and histogram (Figure 31) of forecast errors has also been provided.

- The autocorrelation at lag 0.6 in the correlogram is very close to the significance bounds. Coupled with the high p-value from the Box-Ljung test, there is little evidence of non-zero autocorrelations in the sample forecast errors.
- The histogram shows that the distribution of forecast errors is roughly centred on zero, with little or no skew; thus the errors are normally distributed.

The outputs (predictions) from the three forecasting techniques, including model coefficients, are shown in Table 18. 80% and 95% confidence intervals were also calculated in R, but these have not been shown in the report.

Table 18: Forecasting predictions

Year	Actual	Simple exponential smoothing	Holt's exponential smoothing	Holt-Winters exponential smoothing
Coefficients	-	a = 2761.0	a = 2872.4 b = 10.2	a = 2714.5 b = 3.3 (also values for s1 to s52)
2015	140,057	143,572	163,269	137,547
2016	136,621	143,572	190,753	128,559

The chosen model, **Holt's exponential smoothing**, produced very poor predictions for both years, with the two alternative modelling techniques returning relatively good predictions. As we discussed earlier, data driven models are generally very poor at longer horizon predictions. Further analysis using ARIMA techniques and other model-driven approaches may produce more accurate results.

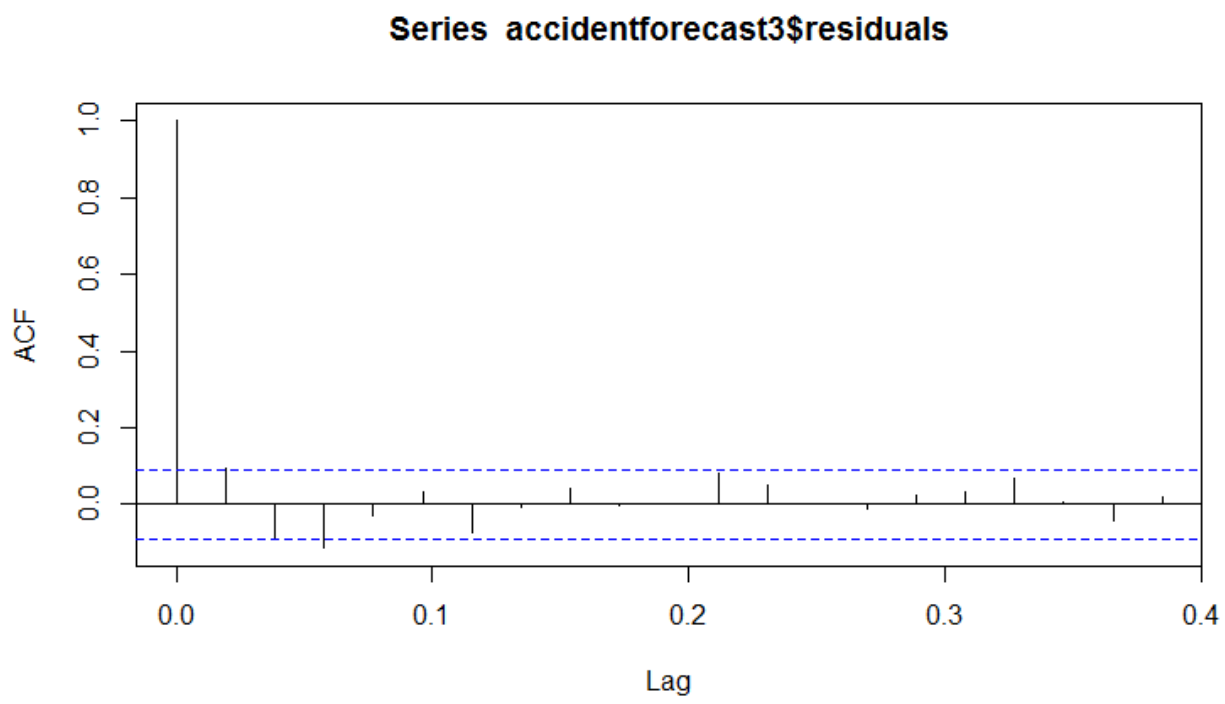


Figure 30: Correlogram of forecast errors

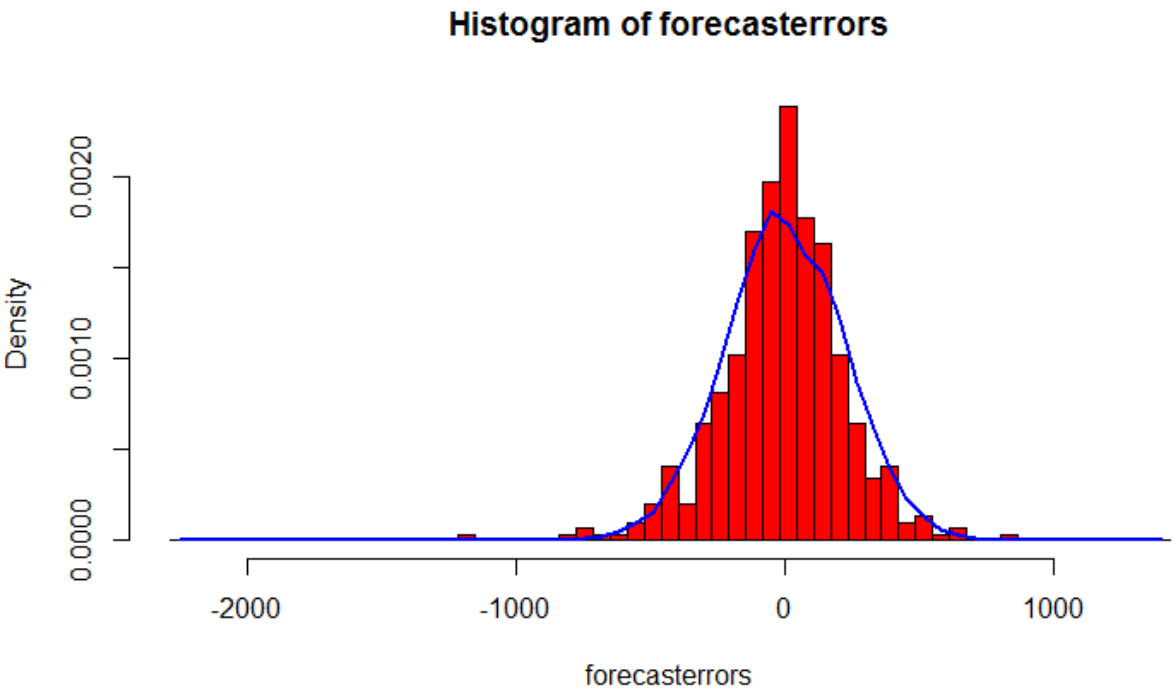


Figure 31: Histogram of forecast errors (with overlaid normal distribution curve)

It is important to note that univariate time series forecasting was undertaken; that is, we did not consider any other attributes which could influence the predictor attribute.

5 Deployment

For the models developed to be useful, they must be outputted in a form that the predictive analysis can be utilised. The currently agreed standard in industry is PMML (Predictive Model Mark-up Language). An extension for RapidMiner exists which allows the user to write a given model to an XML file in PMML 3.2 or 4.0 format. This file can be understood by many databases and this allows a data mining model to be applied directly on a dataset. Thus, PMML allows us to use our model with no dependence on RapidMiner following the development stage. At the time of writing this extension will only support the following model types:

- Decision Tree Models
- Rule Models
- Naive Bayes models for nominal attributes
- Linear Regression Models
- Logistic Regression Models
- Centroid based Cluster models like models of k-means and k-medoids

The model developed on our dataset, Gradient Boosted Trees, is not currently supported by this extension. However, there is an additional service offered by RapidMiner which deploys the process on a server and allows the data mining model to be operationalised. RapidMiner Server does not merely deploy the model, but allows the site administrator to monitor model performance, detect model degradation and automatically react to changes in model performance.

Most data science platforms now offer options to deploy models, usually through an API. Web APIs have made it easy for cross-language applications to work well and integrate models into databases. Since most models are built in R and Python initially, it is time and cost effective to take the working model and supply the necessary details to it rather than rewriting the code in JavaScript, etc. (which may not have the required libraries to complete the task successfully.) Flask is the most widely used web framework in Python, however there is tools such as Django, Falcon and Hug. For R, there is a package called 'plumber' which can be used to deploy a model. This ensures modelling can be an effective tool in data mining, as it is not restricted by the model development package.

The Data Mining Group, the standards group which helped develop PMML, have drafted a successor standard called the Portable Format for Analytics (PFA), which addresses several of PMML's shortcomings. Rather than using XML, PFA uses JSON and YAML as the basis. As a result, where PMML had gaps where models could not be supported as the modelling approach was not canonised in the PMML standard (as can be seen with our model,) PFA addresses these issues and allows all current model types to be deployed.

Once we have deployed our model the performance must be monitored. As it is such a complex model it will require review at regular intervals to ensure the model basis is still accurate, and that the model accuracy has not degraded. Since a high number of the factors which were used to construct the model change on a low frequency (i.e.: Road Type, Speed Limits, Urban or Rural Area, etc.), this interval could be once a year or every second year.

6 Conclusions & Recommendations

6.1 Conclusions

By cleaning and transforming data obtained from Kaggle and other open source repositories, such as the UK government, we were able to address the four objectives we set ourselves (outlined in Section 1.4 of this report). The objectives and the outcomes are summarised below.

1. **How observed changes in road traffic volumes compare with changes in road traffic accidents?**

   Based on our analysis of traffic volumes and accidents for the period 2005 to 2014, we observed that both features reduced from 2005 levels (8% and 28%, respectively). The most significant year-on-year reductions were seen during the 2009 period. However, upward trends were observed from 2014 onwards. Furthermore, the predictive elements of the assignment show this upward trend continuing in 2015 and 2016. The reasons for this are likely increased economic activity, with usually coincides with an increase in traffic volumes.

   On a positive note, accident casualties in severity category 1 (fatal) reduced by 50% over the studied period. There were less than 3,000 victims in 2014. This will be welcome news to stakeholders, such as safety campaigners and local authorities, in the transport industry.

2. **Common features of accidents?**

   Exploring the data allowed us to find a significant number of interesting (and surprising) insights, such as:

   - Saturday was the most common day for accidents and most accidents occur between 4.30PM and 5.30PM (5PM is the most common time),
   - most accidents occurred in 30 MPH zones,
   - nearly 80% of accidents occur in normal weather conditions, and
   - the average age of a driver, who were predominantly (72%) male, involved in an accident was 38.7 years old.

3. **Can we predict accident severity using other attributes (classification)?**

   It was determined that it was possible to predict accident severity, using an oversampled dataset, to assess the attributes which can influence the three categories of accident severity. The highest model accuracy achieved was 51.87%. This model could be utilised by Local Authorities to determine which factors could be varied in order to reduce accident severity and improve road safety.

   As there are a finite (and recorded) number of accidents each year, there was no option to increase our sample size to improve model accuracy. Two possible options include:

   - increasing the scope to additional countries (assuming data has been gathered in a similar manner there), or
   - reducing the number of attributes (to find the critical factors that affect accident severity).

   For example, we know most accidents occur at low speed in urban settings, thus attributes such as engine size and light conditions (most urban areas are lit by street lights) may not be relevant. A local authority

deploying the model may determine that certain factors are not applicable in their area of influence and so reduce the attribute set further.

4. **Can we predict accident rates over time (time series)?**

   Using time series forecasting techniques, it was shown that the number of accidents for a given year could be predicted with a low margin of error. This time series model does not take into account other features i.e. univariate analysis. A nationwide investment by the national government into road infrastructure or urban planning could reduce the accident rate significantly, which simple univariate forecasting techniques would not capture. Suggested next steps would be to build on our predictive model, which takes these factors into account, to create a forecasting method which could be take any changes into account and maintain an accurate forecast.

As mentioned previously it was attempted to use RStudio to carry out a modelling exercise, however we discovered that as memory is allocated statically by RStudio that we did not have the resources to execute this task. Likewise a drawback of RapidMiner is that it is memory intensive, especially for big data. For the *UK Traffic and Accidents* project we found that it was very challenging to use RapidMiner on 8GB and 16GB RAM machines due to memory over-loading. This might not be an issue for large organisations which are resourced for intensive modelling operations, however it definitely looks like the program was not designed for entry level laptops. We did not explore parallel processing options using Spark or Hadoop to execute our calculations, however this could be an option for future analysis.

6.2 Recommendations for future work

Given the timeframe for our project, we did not have time to address some of the questions posed in the Kaggle project. As a continuation of the work completed as part of this report, recommendations for building upon the final dataset created and the model developed are listed below:

- Are accident rates a function of factors such road/vehicle/driver characteristics and weather conditions?
- Is there a significant difference in accident rates between different countries and regions?
- Do accident rates differ significantly between rural and urban areas?
- Do certain areas remain static over time i.e. highly correlated factors appear to have little influence over there accident rates?
- Does a greater police presence result in fewer accidents?
- Given the existing data, with no changes to infrastructure, can future accident figures be accurately predicted?
- Does altering a variable(s) results in a change to accident rates?
- Predict no. of accidents, casualties or vehicles based on other attributes (using regression analysis).

Furthermore, it would be useful to produce interactive maps of changing trends, such as accident hot spots, traffic volumes and other features that would be of interest to stakeholders.

Appendix A: Project proposal

DBS Dublin Business School
excellence through learning

Module Code: B8IT108
Module Title: B8IT108 Data and Web Mining
Lecturer: Terri Hoare

Assessment Title: Application of Data Mining Tools & Techniques

Team Members:

- Colum Kenny
- Maitiú Baxter
- Tatiana Borta

Assessment Task:

- Select an open data source(s)
- Determine what questions are to be answered
- Apply the CRISP-DM methodology to analyse the data

Dataset selected:

- https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales

Aims of Project

Dataset consists of 1.6 million accidents and 16 years of traffic flow in the UK.

- How has changing traffic flow impacted accidents?
- Differences between England, Scotland, and Wales

Challenges presented by the project

- Can we predict accident rates over time?
- What might improve accident rates?

Learning Outcomes

- Plot interactive maps of changing trends
- Identify infrastructure needs, failings and successes
- How have Rural and Urban areas differed (See RoadCategory column)

Project Timetable

- End Week 2 (8/10 23:55) Project Proposal
- End Week 4 (16/10 23:55) CRISP-DM Business and Data Understanding
- End Week 10 (10/12 23:55) CRISP-DM Data Preparation, Modelling, Evaluation
- Week 11 (20/12) A Presentation to be prepared for the final lecture week
- End Week 12 (22/12 23:55) CRISP-DM Deployment, Full Report All Stages

Appendix B: Scatter plot with high covariance

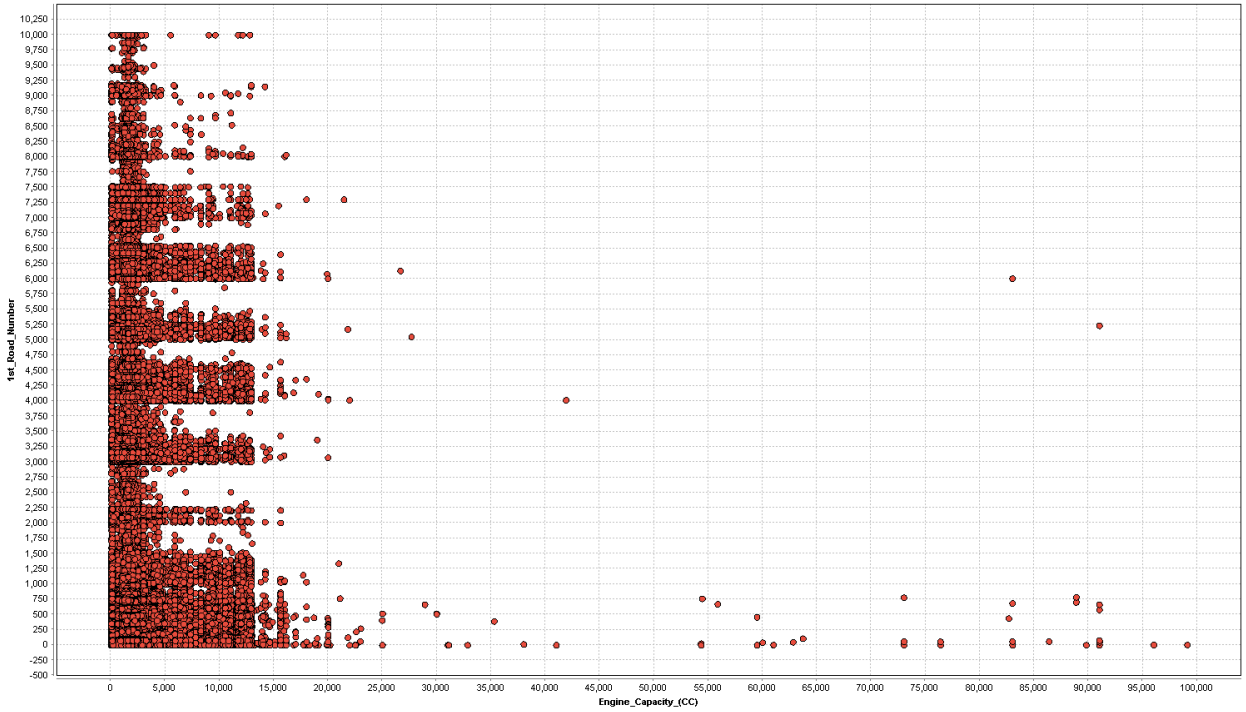


Figure 32: Scatter plot of engine capacity vs 1st road number

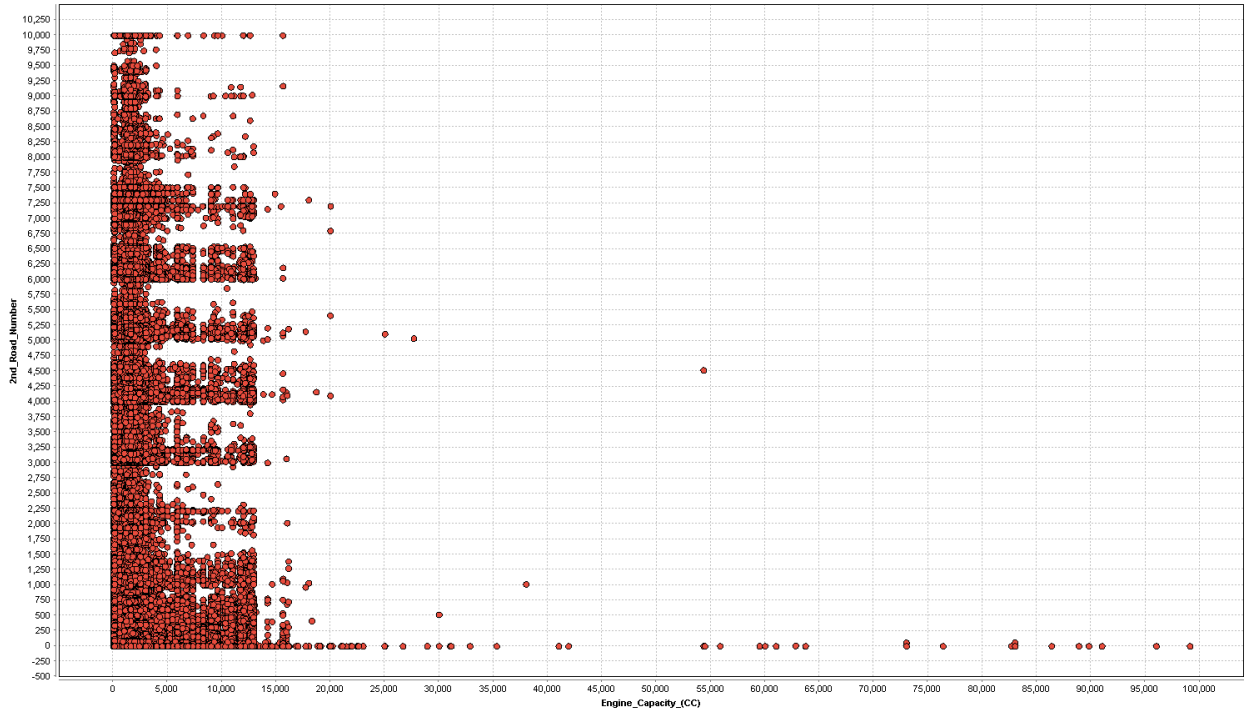


Figure 33: Scatter plot of engine capacity vs 2nd road number

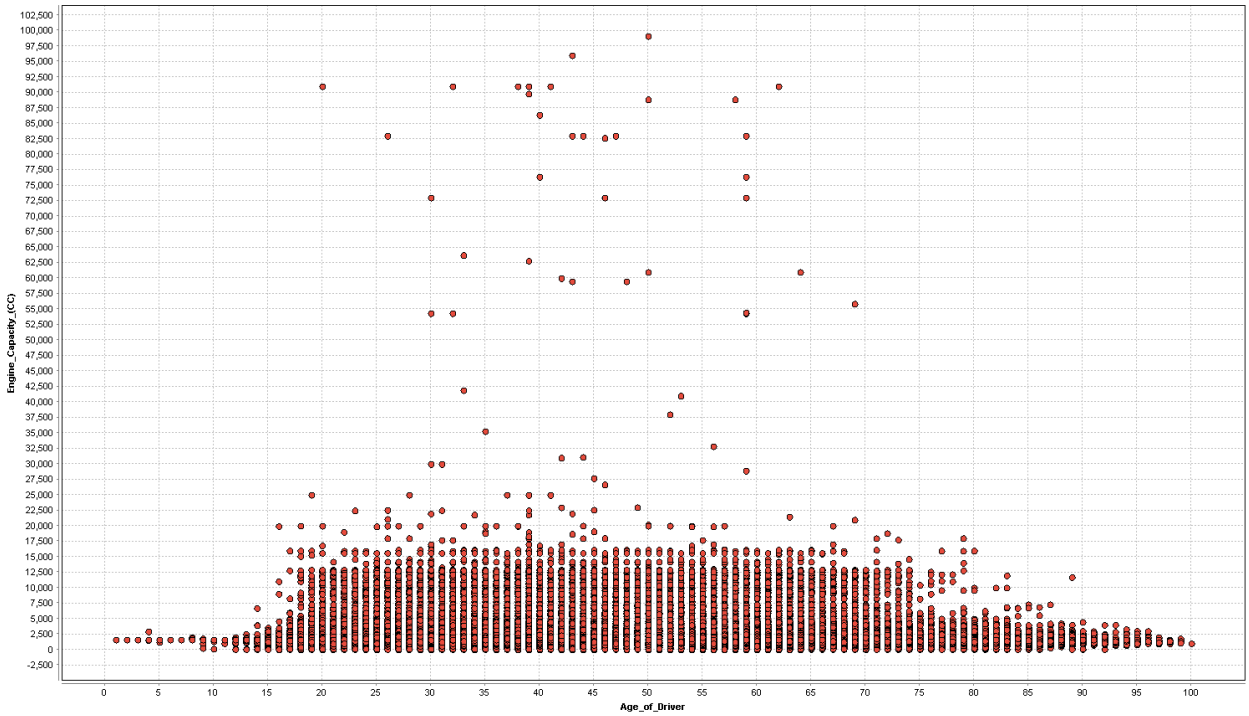


Figure 34: Scatter plot of engine capacity vs age of driver

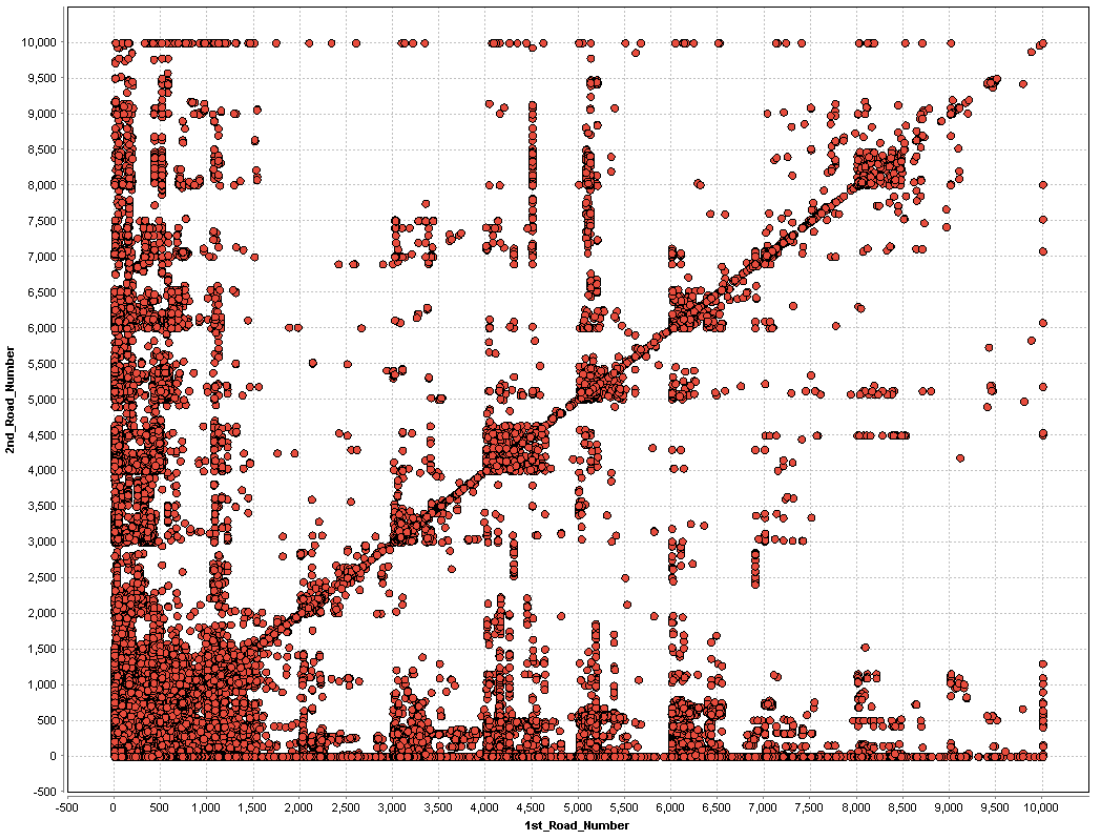


Figure 35: Scatter plot of 1st road number vs 2nd road number

Appendix C: Scatter plots with high correlation (r >= 0.7)

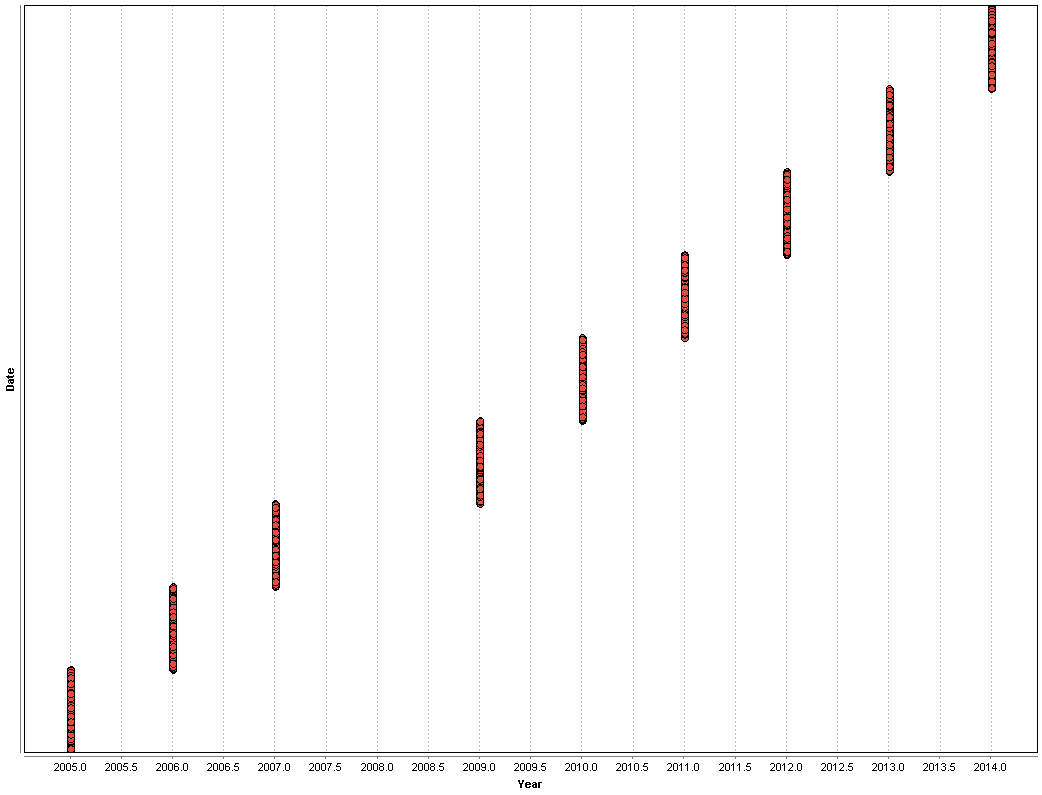


Figure 36: Scatter plot of date vs year

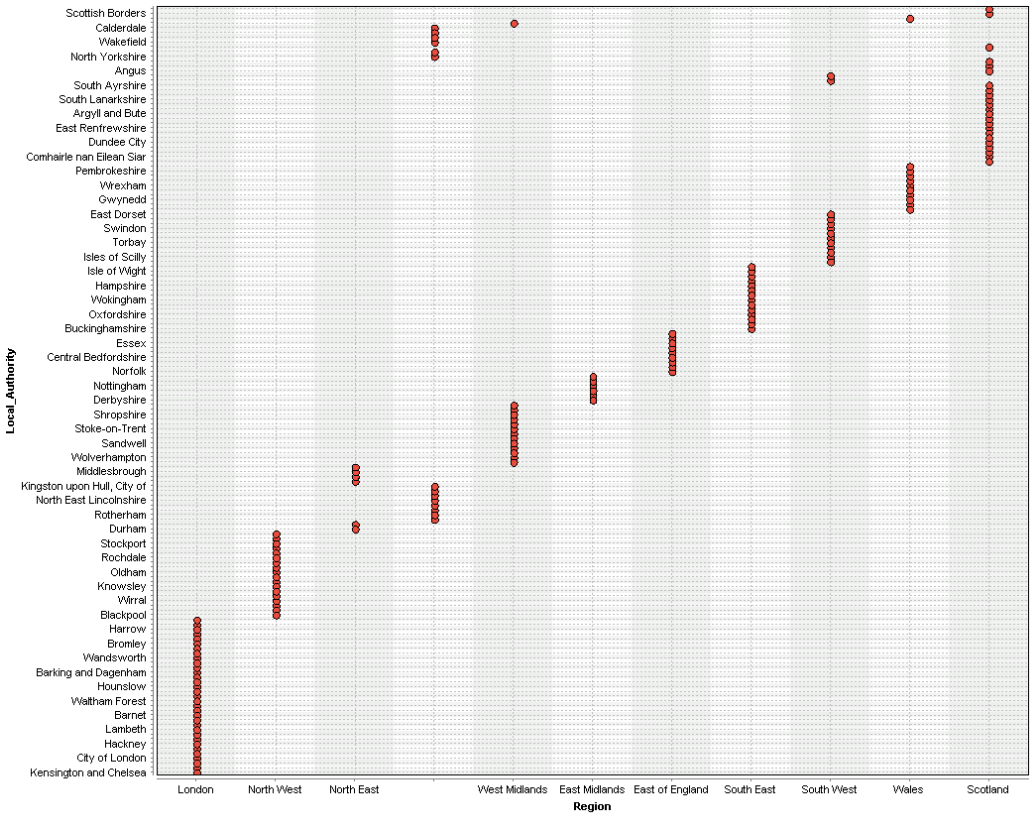


Figure 37: Scatter plot of region vs local authority

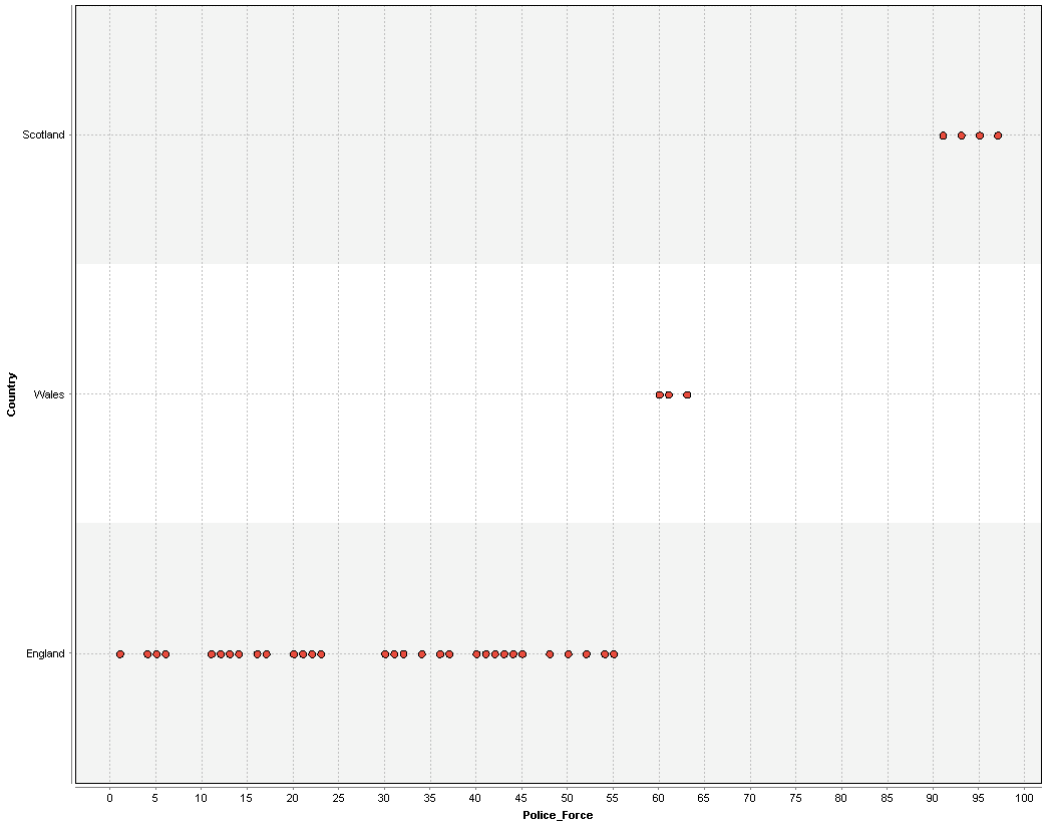


Figure 38: Scatter plot of country vs police force

Appendix D: R code for correlation analysis and modelling

```
# ipak function: install and load multiple R packages.

# check to see if packages are installed. Install them if they are not, then load them into the R session.


ipak <- function(pkg){

 new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]

 if (length(new.pkg))

  install.packages(new.pkg, dependencies = TRUE)

 sapply(pkg, require, character.only = TRUE)

}

# usage

packages <- c("ggplot2", "dplyr", "reshape2", "RColorBrewer", "scales", "grid","rgdal","plotly","corrplot","DEoptimR","caret")

ipak(packages)


setwd("XXXXX")


#### input csv files ####

AccidentData <- read.csv(file="merged_accidents_clean.csv", header=TRUE, sep=",")

TrafficData <- read.csv(file="ukTrafficAADF_clean.csv", header=TRUE, sep=",")

TrafficDataGroupedDailyKM <- read.csv(file="ukTrafficAADF_grouped_daily_km_local_authority.csv", header=TRUE, sep=",")

TrafficDataGroupedYear <- read.csv(file="ukTrafficAADF_grouped_year_local_authority.csv", header=TRUE, sep=",")

VehicleData <- read.csv(file="vehicles_clean.csv", header=TRUE, sep=",")

#### prepare the data for analysis ####

# create a data table for grouping

CData <- tbl_df(AccidentData)

country_group <- group_by(CData, Country,Year,Accident_Severity)

country_summary <- summarise(count =n(),country_group)


LA_group <- group_by(CData, Local_Authority,Year)
```

```
LA_summary <- summarise(count =n(),LA_group)

vehicle_group <- group_by(VehicleData, Vehicle_Type)

vehicle_summary <- summarise(count=n(),vehicle_group)


# tie each car with the accident it was involved in

VehicleDataWithLA <- VehicleData %>% left_join(AccidentData, c("Accident_Index" = "Accident_Index"))


summary(VehicleDataWithLA)

summarise(VehicleDataWithLA)

colnames(VehicleDataWithLA)

colnames(AccidentData)

head(VehicleDataWithLA)


#### Data exploration ####

#### Correlation Graphs ####

study <- tbl_df(VehicleDataWithLA)

study$Num_Vehicle_Type = as.integer(as.factor(study$Vehicle_Type))

study$Num_Vehicle_Manoeuvre = as.integer(as.factor(study$Vehicle_Manoeuvre))

study$Num_Junction_Location = as.integer(as.factor(study$Junction_Location))

study$Num_Sex_of_Driver = as.integer(as.factor(study$Sex_of_Driver))

study$Num_Local_Authority = as.integer(as.factor(study$Local_Authority ))

study$Num_Region = as.integer(as.factor(study$Region))

study$Num_Country = as.integer(as.factor(study$Country))

study$Num_Road_Type = as.integer(as.factor(study$Road_Type))

study$Num_PC.Human_Control = as.integer(as.factor(study$Pedestrian_Crossing.Human_Control))

study$Num_PC.Physical_Facilities = as.integer(as.factor(study$Pedestrian_Crossing.Physical_Facilities))

study$Num_Light_Conditions = as.integer(as.factor(study$Light_Conditions))

study$Num_Weather_Conditions = as.integer(as.factor(study$Weather_Conditions))

study$Num_Road_Surface_Conditions = as.integer(as.factor(study$Road_Surface_Conditions))
```

```
study$Num_Special_Conditions = as.integer(as.factor(study$Special_Conditions_at_Site))

study$Num_Carriageway_Hazards = as.integer(as.factor(study$Carriageway_Hazards))

study$Num_Police_Attendance = as.integer(as.factor(study$Did_Police_Officer_Attend_Scene_of_Accident))



# calculate the correlation matrix

cm <- subset(study,select = c(Age_of_Driver,Num_Vehicle_Type, Engine_Capacity_.CC.,Num_Vehicle_Manoeuvre,
Num_Junction_Location,Number_of_Vehicles,Num_Sex_of_Driver,Num_Country,Num_Region,Num_Local_Authority,Number_of_C
asualties,Speed_limit,Urban_or_Rural_Area,Accident_Severity,Police_Force,Num_Road_Type,Num_PC.Human_Control,

Num_PC.Physical_Facilities,Num_Weather_Conditions,Num_Road_Surface_Conditions,Num_Special_Conditions,Num_Carriageway_
Hazards,Year,Day_of_Week,Num_Police_Attendance))

cm_matrix <- cor(cm,method = "pearson", use = "complete.obs")



#correlation plot

corrplot(cm_matrix,method = "color") # nice, better chart follows

corrplot(cm_matrix,method = "number",type="upper", col=brewer.pal(n=8, name="RdYlBu"), number.cex = 0.7) # nice chart for
display

corrplot.mixed(cm_matrix, lower.col = "black", number.cex = .7) # not useful, remove?



#Combining correlogram with the significance test

res1 <- cor.mtest(cm_matrix, conf.level = .95)

corrplot(cm_matrix, p.mat = res1$p, sig.level = .05)

corrplot(cm_matrix, method="number",number.cex = .7, p.mat = res1$p, insig = "blank")



# Reducing variables to plot correlation again

cm_short <- subset(study,select =
c(Age_of_Driver,Num_Vehicle_Type,Num_Sex_of_Driver,Number_of_Casualties,Speed_limit,Urban_or_Rural_Area,Accident_Severi
ty,Num_Road_Type,Num_Weather_Conditions,Num_Road_Surface_Conditions))

cm_matrix_short <- cor(cm_short,method = "pearson", use = "complete.obs")

res1_short <- cor.mtest(cm_matrix_short, conf.level = .95)

corrplot(cm_matrix_short, p.mat = res1_short$p, sig.level = .05)

corrplot(cm_matrix_short, method="number",number.cex = .7, p.mat = res1_short$p, insig = "blank")
```

```
corrplot(cm_matrix_short,method = "number",type="upper", col=brewer.pal(n=8, name="RdYlBu"), number.cex = 0.7, insig =
"blank",sig.level = .05)


#### Model Building ####

inTrain <- createDataPartition(AccidentData$Local_Authority, p = .7, list = FALSE)

str(inTrain)

training <- AccidentData[ inTrain,]

testing <- AccidentData[-inTrain,]


set.seed(123)

ctrl <- trainControl(method = "repeatedcv",repeats = 3,classProbs = TRUE,summaryFunction = twoClassSummary)

plsFit <- train(Accident_Severity ~ .,data = training,method = "pls",metric = "ROC",preProc = c("center", "scale"))#trControl = ctrl,

plsFit <- train(Accident_Severity ~ .,data = training,method = "glm", preProc = c("center", "scale"))


plsFit

gc()


model<-glm(Accident_Severity~. , data=training)

prediction<- predict(model, test)
```

Appendix E: R Code for time series analysis and forecasting

```
library(dplyr)

library(TTR)

library(forecast)


file <- read.csv("...../merged_accidents1.csv") #insert correct file location

file <- as.data.frame(file)

summary(file)


##TIME SERIES ANALYSIS


##1. SIMPLE EXPONENTIAL SMOOTHING

#time series of all attributes

filetimeseries <- ts(file,frequency = 52,start = c(2005,2))

plot.ts(filetimeseries)


#create dataframe with 'accidents' attribute

accidents <- file[,1]

head(accidents)


#row 1 + another row deleted (due to v. low values) - start at week 2

#freq=52 (weekly)

accidenttimeseries <- ts(accidents,frequency = 52,start = c(2005,2))

plot.ts(accidenttimeseries)


#using SMA function to smooth ts data by specifying order (span) of the simple moving average

accidenttimeseriesSMA3 <- SMA(accidenttimeseries,n=3)

plot.ts(accidenttimeseriesSMA3)
```

```
#change order of moving average to smooth again

accidenttimeseriesSMA5 <- SMA(accidenttimeseries,n=5)

plot.ts(accidenttimeseriesSMA5)


#estimate trend, seasonal and irregular components of seasonal ts data

accidenttimeseriescomponents <- decompose(accidenttimeseries)

accidenttimeseriescomponents$trend

plot(accidenttimeseriescomponents)


#seasonal adjusting - seasonally adjust the time series by estimating the seasonal component, and

#subtracting the estimated seasonal component from the original time series

accidenttimeseriesseasonallyadjusted <- accidenttimeseries - accidenttimeseriescomponents$seasonal

plot(accidenttimeseriesseasonallyadjusted)


##TIME SERIES FORECASTING

##1. SIMPLE EXPONENTIAL SMOOTHING

#additive model with constant level and no seasonality - don't have constant level so unlikely to be useful


#for HoltWinters simple exponential smoothing, we need to set the parameters beta=FALSE and gamma=FALSE

accidentforecast <- HoltWinters(accidenttimeseriesseasonallyadjusted, beta = FALSE, gamma = FALSE)

accidentforecast

plot(accidentforecast)


#check accuracy of forecast

#A value closer to 0 indicates that the model has a smaller random error component, and that the fit will be more useful for
prediction.

accidentforecast$SSE


#implement forecast function with specifying h value (i.e. no. of points = weeks)
```

```
accidentforecast1 <- forecast(accidentforecast,h=104)

plot(accidentforecast1)

accidentforecast1


#dealing with missing values in the residuals

is.na(accidentforecast1$residuals)

accidentforecast1$residuals <- na.omit(accidentforecast1$residuals)


#calculate a correlogram (autocorrelation plot) of the in-sample forecast errors

acf(accidentforecast1$residuals)

acf(accidentforecast1$residuals, lag.max = 20)


Box.test(accidentforecast1$residuals, lag = 20, type = "Ljung-Box")


plot.ts(accidentforecast1$residuals)


#plot forecase errors

plotForecastErrors <- function(forecasterrors)

{

 # make a histogram of the forecast errors:

 mybinsize <- IQR(forecasterrors)/4

 mysd   <- sd(forecasterrors)

 mymin  <- min(forecasterrors) - mysd*5

 mymax  <- max(forecasterrors) + mysd*3

 # generate normally distributed data with mean 0 and standard deviation mysd

 mynorm <- rnorm(10000, mean=0, sd=mysd)

 mymin2 <- min(mynorm)

 mymax2 <- max(mynorm)

 if (mymin2 < mymin) { mymin <- mymin2 }
```

```
 if (mymax2 > mymax) { mymax <- mymax2 }
 # make a red histogram of the forecast errors, with the normally distributed data overlaid:
 mybins <- seq(mymin, mymax, mybinsize)
 hist(forecasterrors, col="red", freq=FALSE, breaks=mybins)
 # freq=FALSE ensures the area under the histogram = 1
 # generate normally distributed data with mean 0 and standard deviation mysd
 myhist <- hist(mynorm, plot=FALSE, breaks=mybins)
 # plot the normal curve as a blue line on top of the histogram of forecast errors:
 points(myhist$mids, myhist$density, type="l", col="blue", lwd=2)
}

plotForecastErrors(accidentforecast1$residuals)
## roughly centred on zero and normally distributed with slight left skew

##2. HOLT'S EXPONENTIAL SMOOTHING
# additive model with increasing or decreasing trend and no seasonality

accidentforecast2 <- HoltWinters(accidenttimeseriesseasonallyadjusted, gamma = FALSE)
accidentforecast2
plot(accidentforecast2)

accidentforecast2 <- HoltWinters(accidenttimeseriesseasonallyadjusted, gamma = FALSE, l.start = 3838.213)
accidentforecast2
plot(accidentforecast2)
accidentforecast2$SSE

accidentforecast3 <- forecast(accidentforecast2, h=104)
plot(accidentforecast3)
accidentforecast3
```

```
accidentforecast3$residuals <- na.omit(accidentforecast3$residuals)

acf(accidentforecast3$residuals, lag.max = 20)

Box.test(accidentforecast3$residuals, lag = 20, type = "Ljung-Box")

plot.ts(accidentforecast3$residuals)

plotForecastErrors(accidentforecast3$residuals)


##3. HOLT-WINTERS EXPONENTIAL SMOOTHING

#additive model with increasing or decreasing trend and seasonality


accidentforecast4 <- HoltWinters(accidenttimeseries)

accidentforecast4

accidentforecast4$SSE

plot(accidentforecast4)


accidentforecast5 <- forecast(accidentforecast4,h=104)

plot(accidentforecast5)

accidentforecast5


accidentforecast5$residuals <- na.omit(accidentforecast5$residuals)

acf(accidentforecast5$residuals)

Box.test(accidentforecast5$residuals, lag = 20, type = "Ljung-Box")

plot.ts(accidentforecast5$residuals)

plotForecastErrors(accidentforecast5$residuals)


##4. ARIMA Models

# include an explicit statistical model for the irregular component of a time series, that allows for non-zero autocorrelations in the irregular component
```

```
accidenttimeseriesdiff1 <- diff(accidenttimeseries,differences = 1)

plot.ts(accidenttimeseriesdiff1)


accidenttimeseriesdiff2 <- diff(accidenttimeseries,differences = 2)

plot.ts(accidenttimeseriesdiff2)


acf(accidenttimeseriesdiff1, lag.max = 20)

acf(accidenttimeseriesdiff1, lag.max = 20, plot = FALSE) #get auto-correlation values

##correlogram is zero after lag 2


pacf(accidenttimeseriesdiff1, lag.max = 20)

pacf(accidenttimeseriesdiff1, lag.max = 20, plot = FALSE) #get auto-correlation values

##partial correlogram tails off to zero after lag 7 (really 16?)


#find appropriate ARIMA model

accidentarima <- auto.arima(accidenttimeseriesdiff1)

##ARIMA(2,0,4)(1,0,0)[52] with zero mean

#accidentarima <- arima(accidenttimeseries,order = c(2,0,4))

accidentarima


accidentforecast6 <- forecast(accidentarima,h=52)

plot(accidentforecast6)


acf(accidentforecast6$residuals, lag.max = 20)

Box.test(accidentforecast6$residuals, lag = 20, type = "Ljung-Box")


plot.ts(accidentforecast6$residuals)

plotForecastErrors(accidentforecast6$residuals)

accidentforecast6
```