Credit Hours System		Cairo University Faculty of Engineering

Data Mining, Big Data and Data Analytics
Final Project Document
UK Traffic Accidents (2005-2014)

- Submitted by:

  Dina Walid - 1152113
  Nariman Reda - 1152043
  Ahmed Ashraf - 1152032
  Yasmine Hatem - 1152059

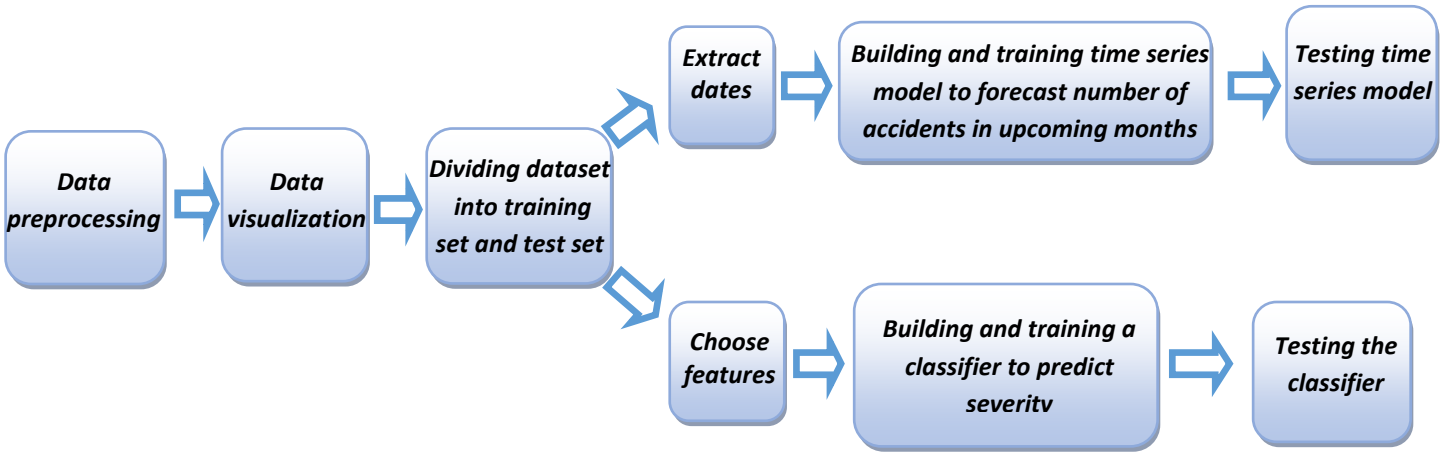
- Date: 01/05/2020

Table of contents:

Problem Description:

Road accidents are a major and growing cause of death, injury and loss of money. So, investigating accidents is a must. This project aims to study accidents in UK over ten years, which would help in numerous applications. One of the applications is to predict the number of accidents that would happen in the next few years, we can use this prediction to compare with the actual number of accidents and check if there is a significant increase which will show us that we have a problem. Another application is to predict severity of the accident based on some attributes, which would help in deciding how many ambulances to send to the accident location for example. We will show some statistics that would help in identifying the problem, like in which districts accidents happen the most. By finding the causes of an accident, taking steps to control or eliminate it can be done to help prevent similar accidents from happening in the future.

Project Pipeline:



Dataset:

The UK government collected traffic data, recording over 1.6 million accidents in the process and making this one of the most comprehensive traffic data sets out there. It is a huge picture of a country undergoing change. Note that all the contained accident data comes from police reports, so this data does not include minor incidents. Accidents data is split across three CSV files. These three files together constitute 1.6 million traffic accidents. The total time period is 2005 through 2014, but 2008 is missing.

Dataset columns:

1	Accident_Index	18	Speed_limit
2	Location_Easting_OSGR	19	Junction_Detail
3	Location_Northing_OSGR	20	Junction_Control
4	Longitude	21	2nd_Road_Class
5	Latitude	22	2nd_Road_Number
6	Police_Force	23	Pedestrian_Crossing-Human_Control
7	Accident_Severity	24	Pedestrian_Crossing-Physical_Facilities
8	Number_of_Vehicles	25	Light_Conditions
9	Number_of_Casualties	26	Weather_Conditions
10	Date	27	Road_Surface_Conditions
11	Day_of_Week	28	Special_Conditions_at_Site
12	Time	29	Carriageway_Hazards
13	Local_Authority_(District)	30	Urban_or_Rural_Area
14	Local_Authority_(Highway)	31	Did_Police_Officer_Attend_Scene_of_Accident
15	1st_Road_Class	32	LSOA_of_Accident_Location
16	1st_Road_Number	33	Year
17	Road_Type		

Analysis and solution of problem:

Data preprocessing:

1- Removing of useless columns that will not affect our insights or models:

a) Accident_index
b) Location_Easting_OSGR
c) Location_Northing_OSGR
We removed (b,c) as we used Latitude and Longitude columns of dataset instead.

2- Removing of columns that we searched a lot and did not understand what they refer to:

a) Police_Force
b) LSOA_of_Accident_Location

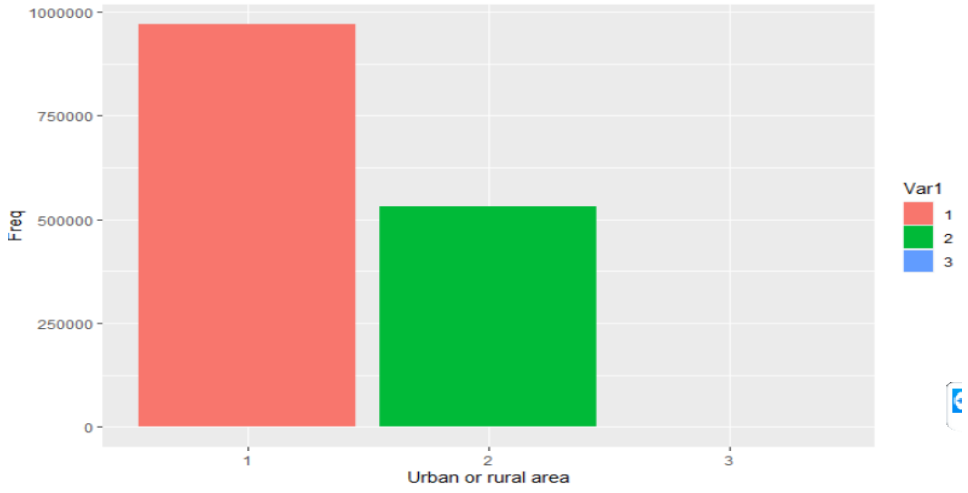
3- Removing of column that had no value for any row:

a) Junction_Detail

4- Removing of rows that had missing values.

5- Removing of outliers:

For example, in Urban_or_Rural column, it had to be 1 for Urban or 2 for Rural but some rows had value of 3, so we removed all rows of value 3 which were only 35/1.6 million records.



6- Categorizing time:

Morning	6:00 -12:00
Afternoon	12:00 – 17:00
Evening	17:00 – 20:00
Night	20:00 - 6:00

7- Categorizing months:

Spring	3 – 5
Summer	6 - 8
Autumn	9 – 11
Winter	12 - 2

Time series analysis model related preprocessing:

8- Creating a new sorted data frame containing only:

Month, Year and Number of accidents.

9- Dividing dataset into training set and test set:

Training set: 2005 – 2011 (1,000,000 records)

Test set: 2012 – 2015 (500,000 records)

Severity prediction model related preprocessing:

10- Dividing dataset into training set and test set:

Training set: 75% of dataset.

Test set: 25% of dataset.

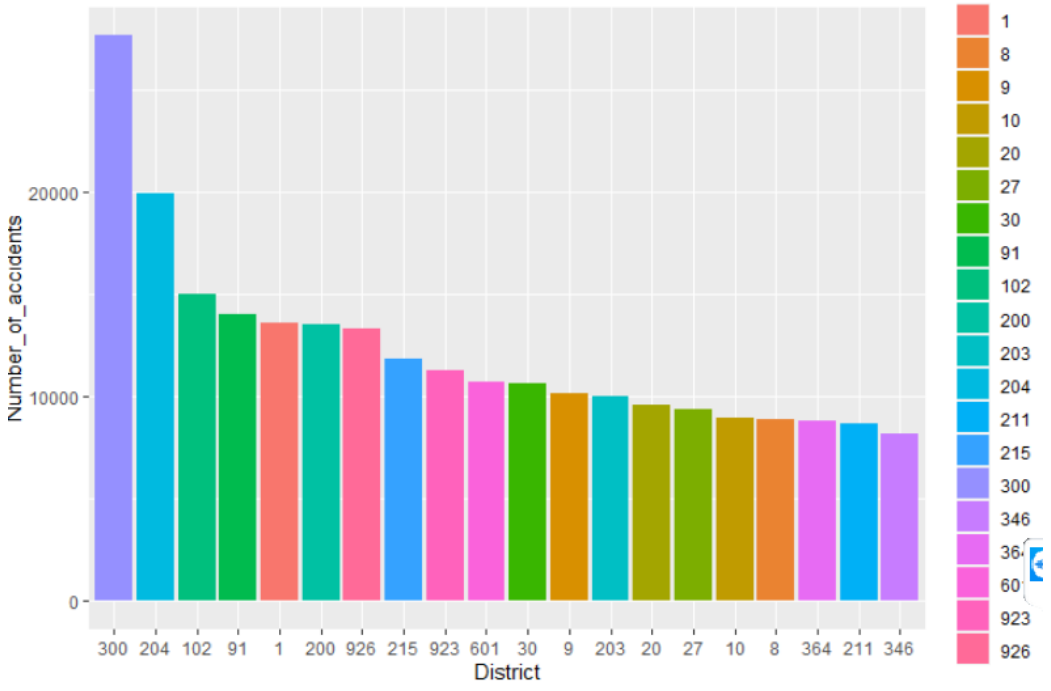
11- Fixing unbalanced training set problem:

The dataset was biased to severity 3, almost 90% of training set. So, this was solved by deleting random rows of severity 3.

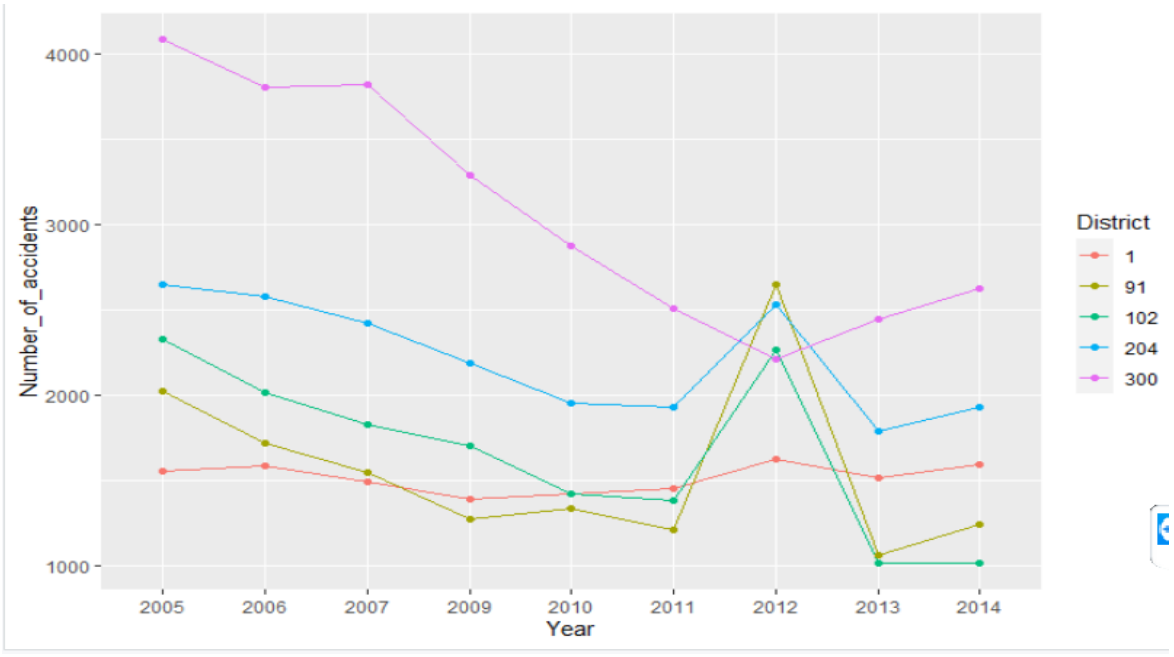
Data visualization:

This part aims to visualize data so as to deeply understand the dataset and recognize the relationships between its columns. This is important because it allows trends and patterns to be more easily seen.

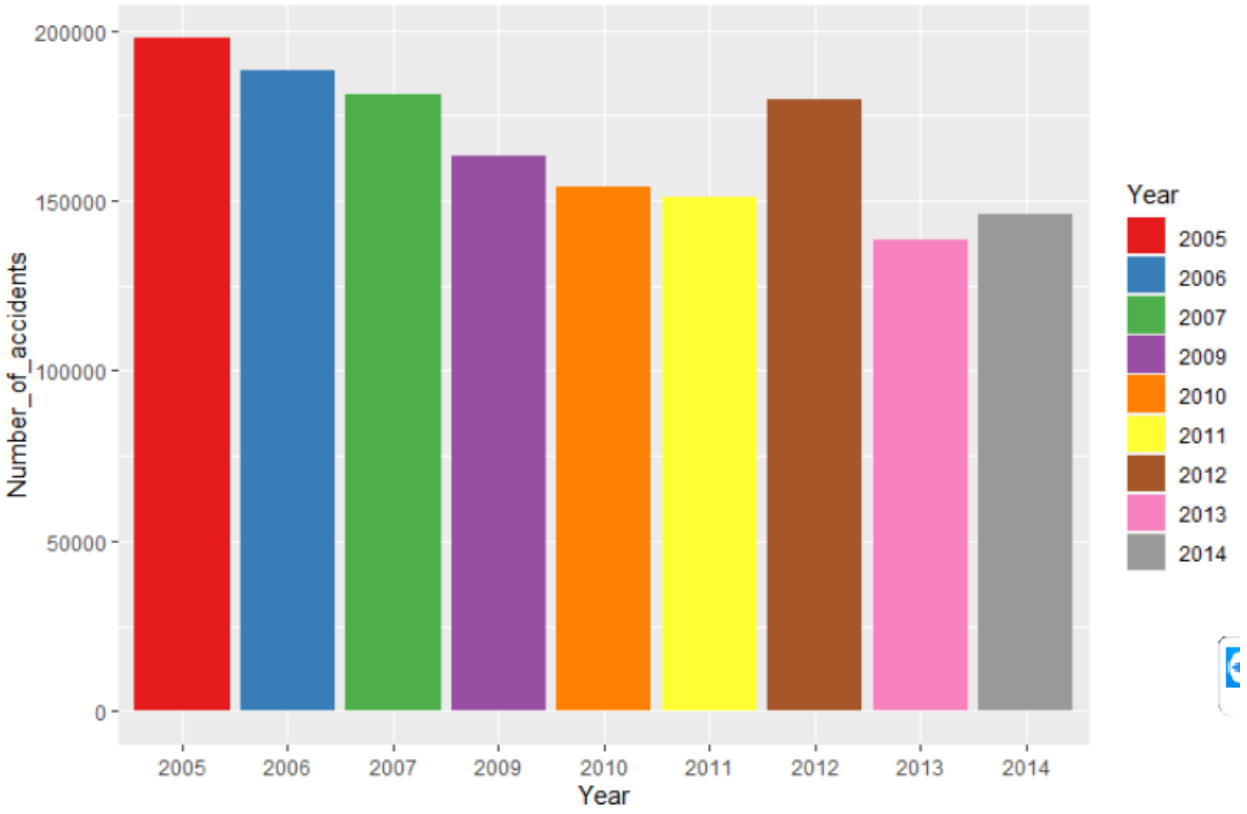
1- Number of accidents in top 20 districts:



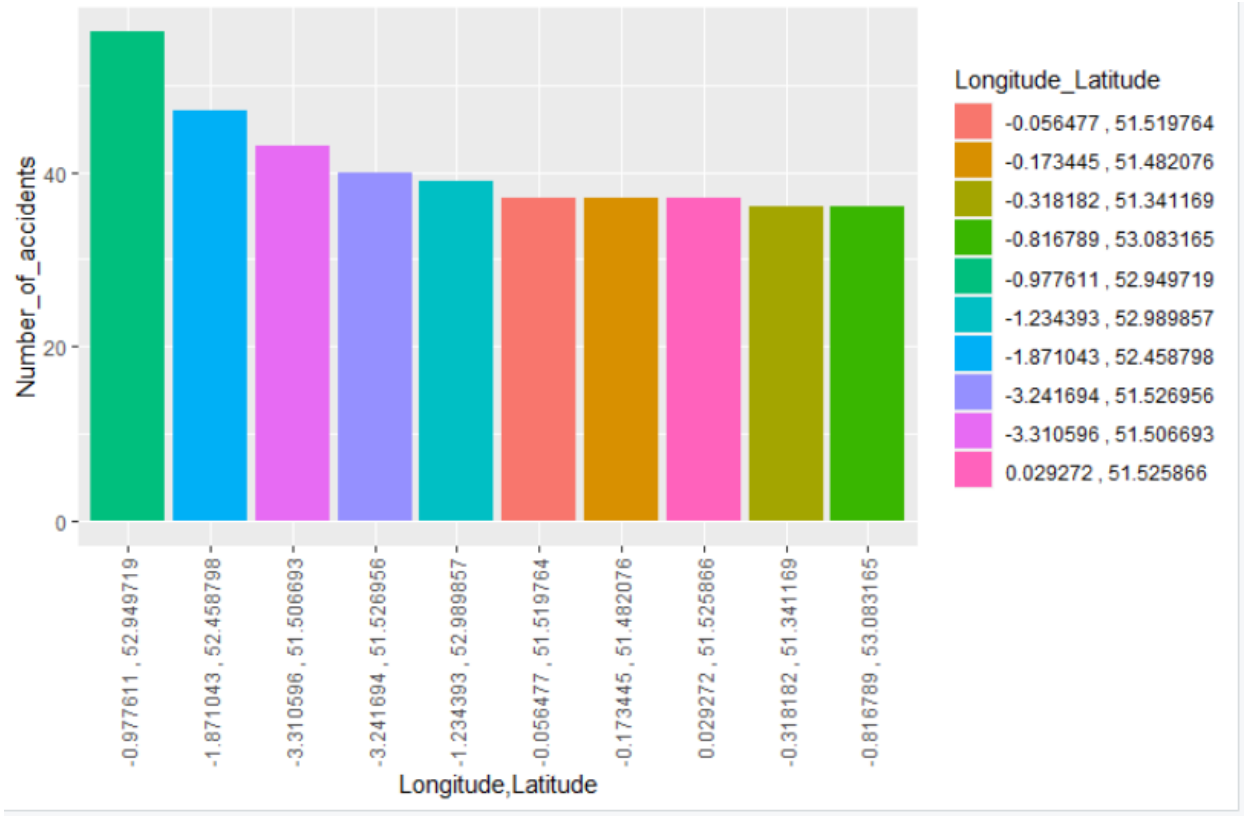
2- Number of accidents in top 5 districts over years:



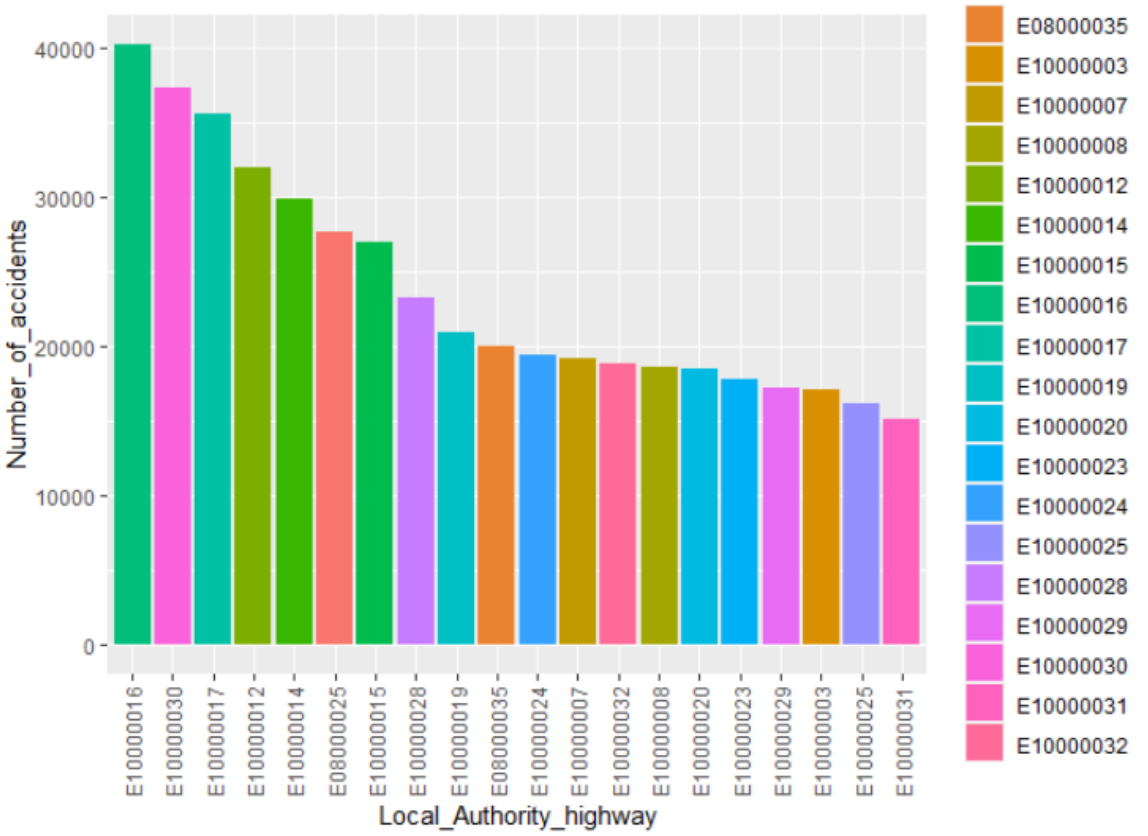
3- Number of accidents over years:



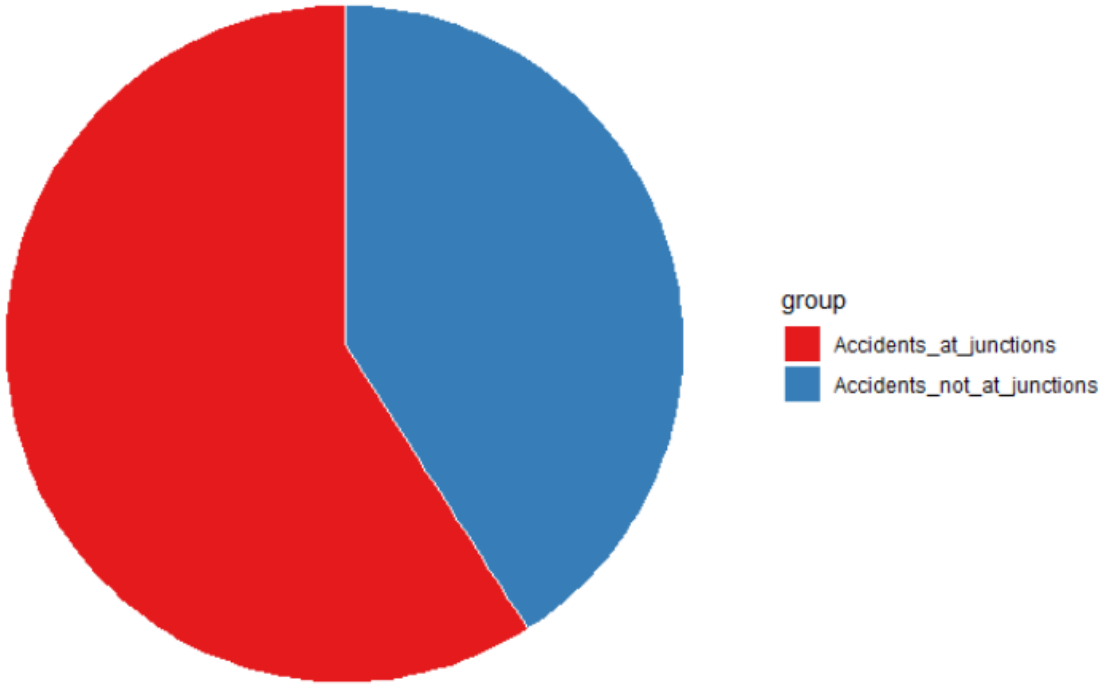
4- Top 10 frequent locations:



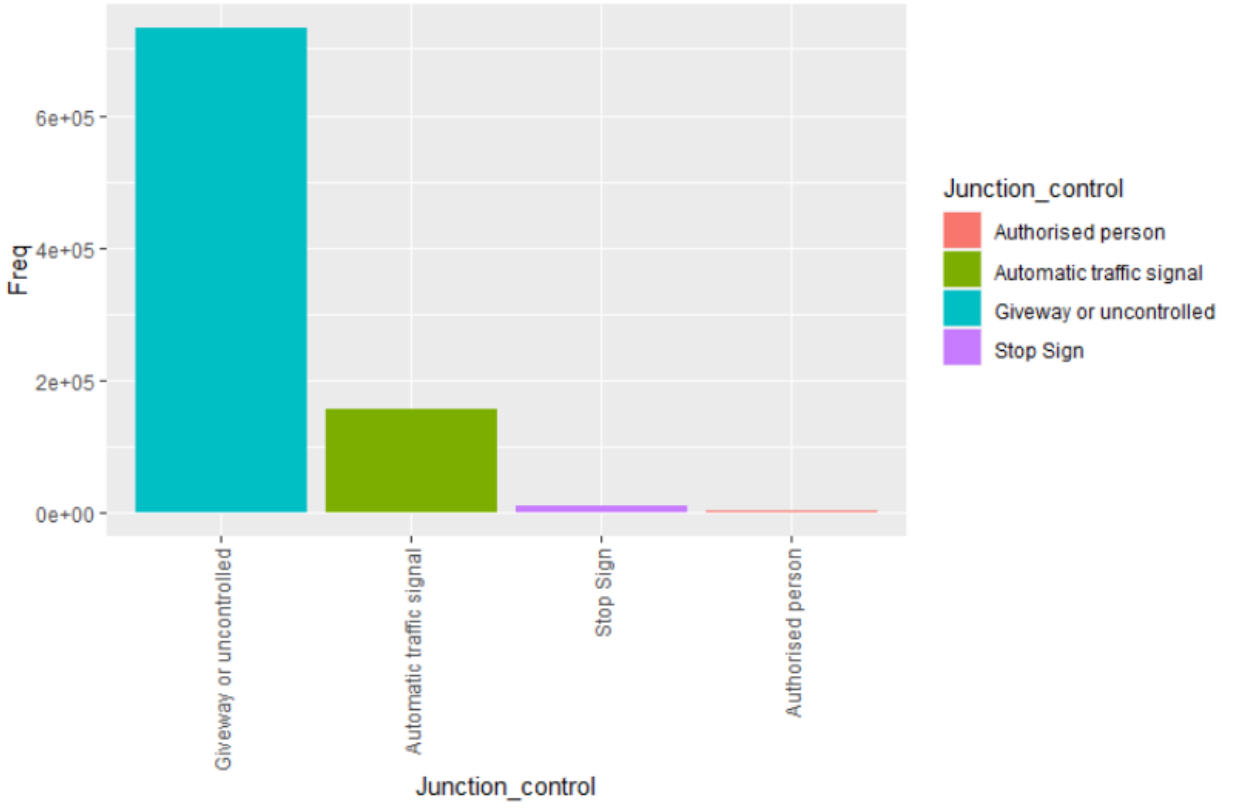
5- Number of accidents of each local authority(Highway) that have the duty of maintaining the roads “Showing only the most frequent 20”:



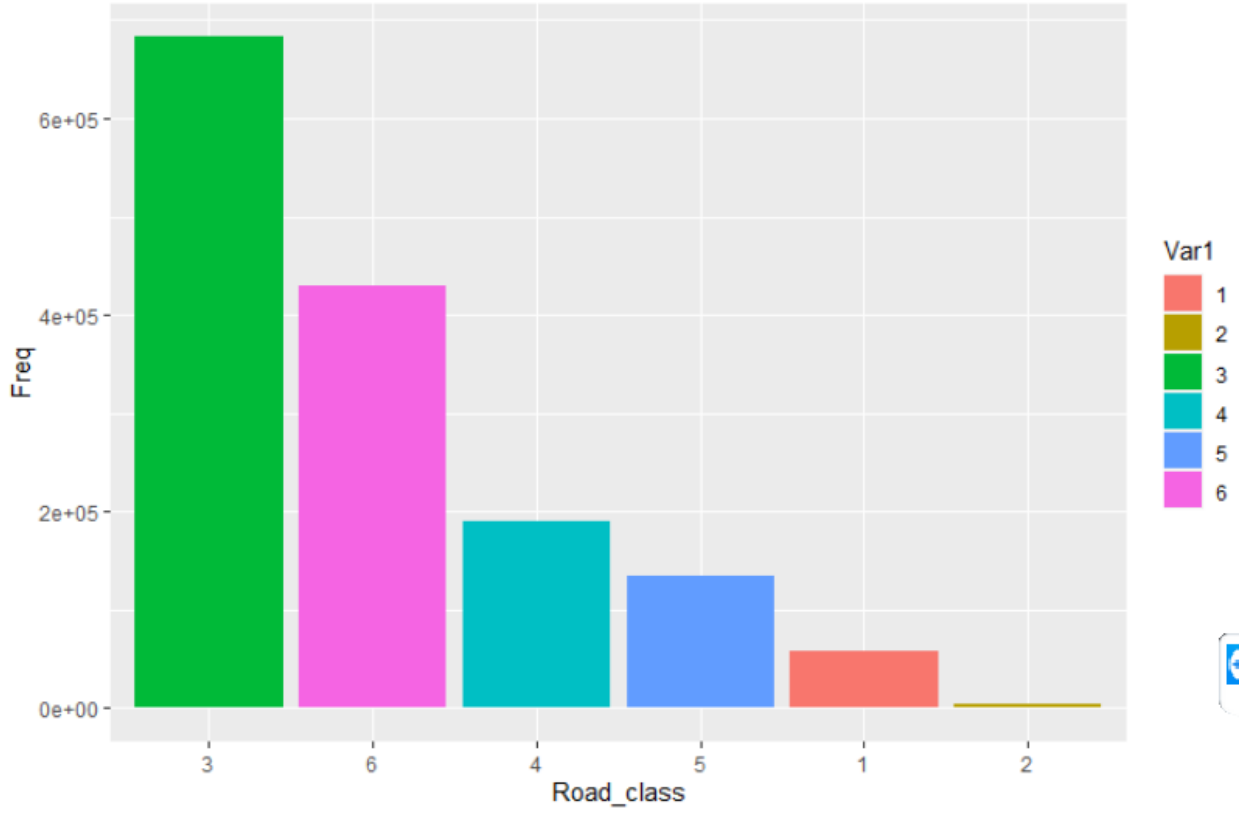
6- Accidents at junctions vs Accidents not at junctions:



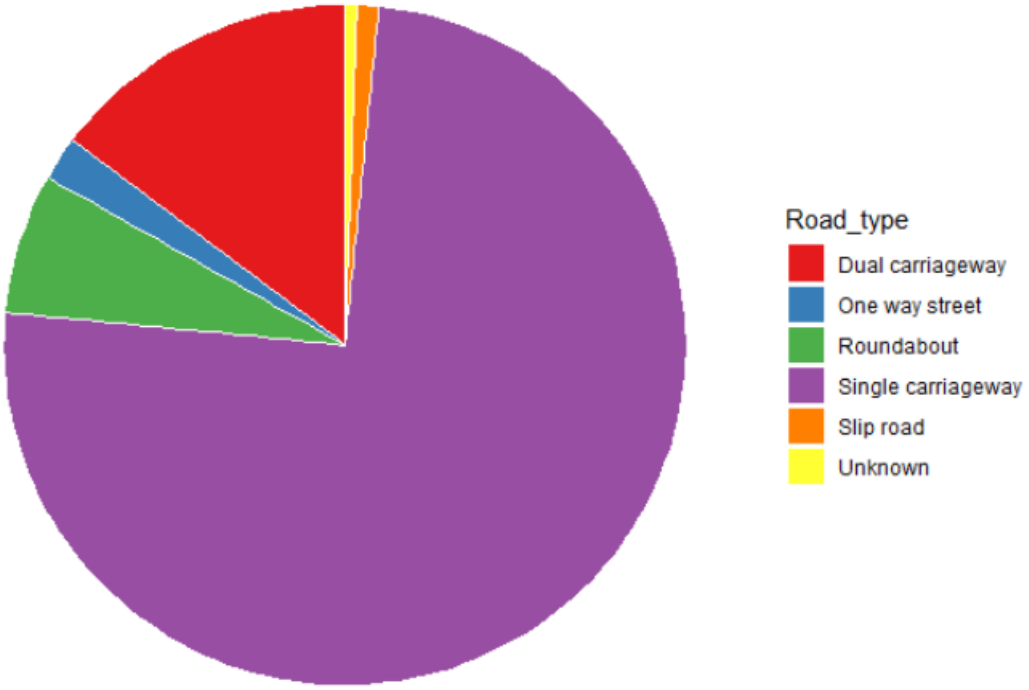
7- Relationship between junction control and number of accidents:



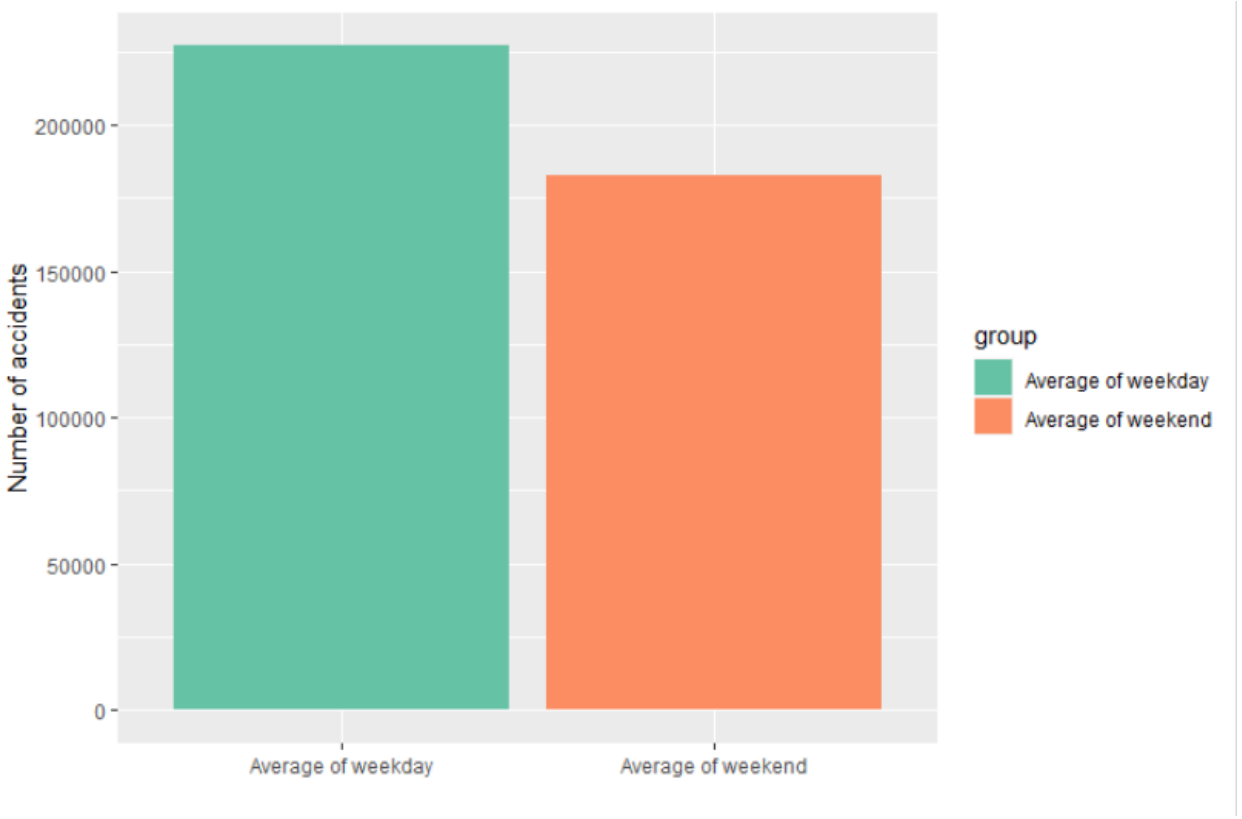
8- Relationship between road class and number of accidents:



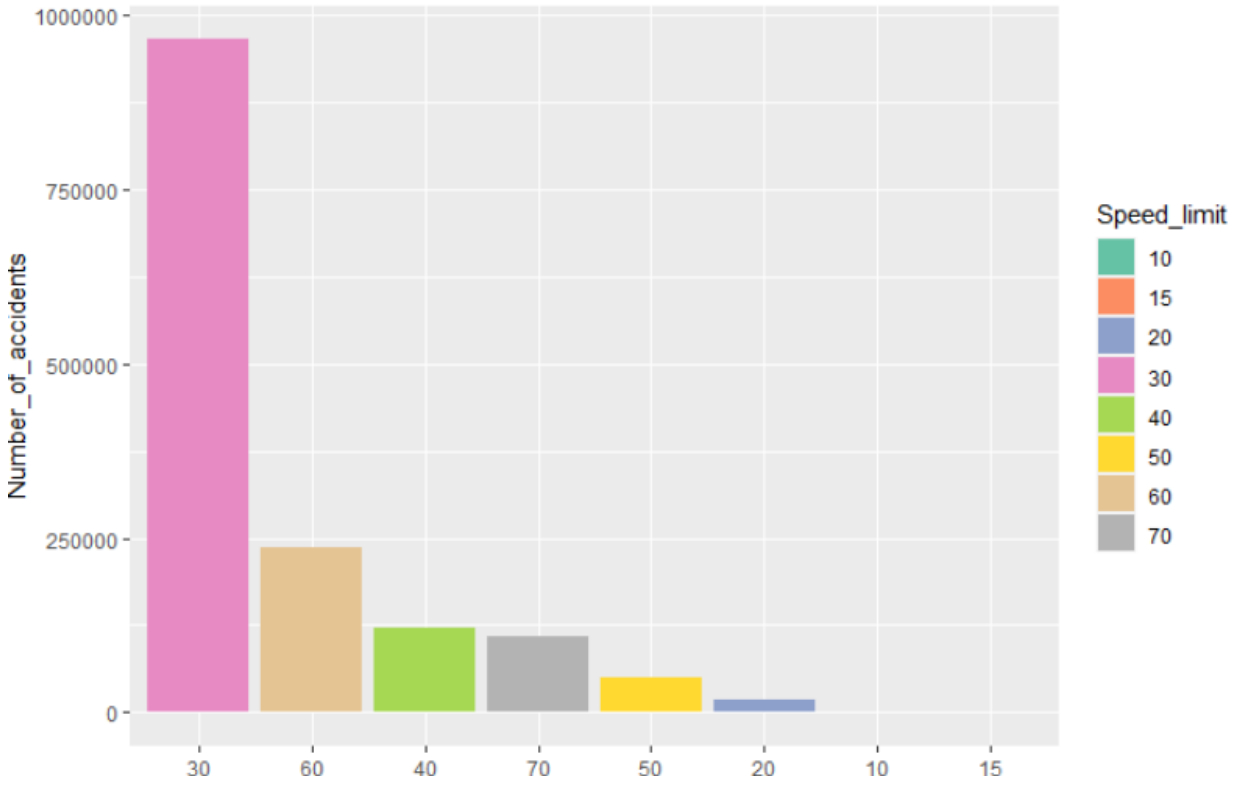
9- Percentage of each road type in accidents:



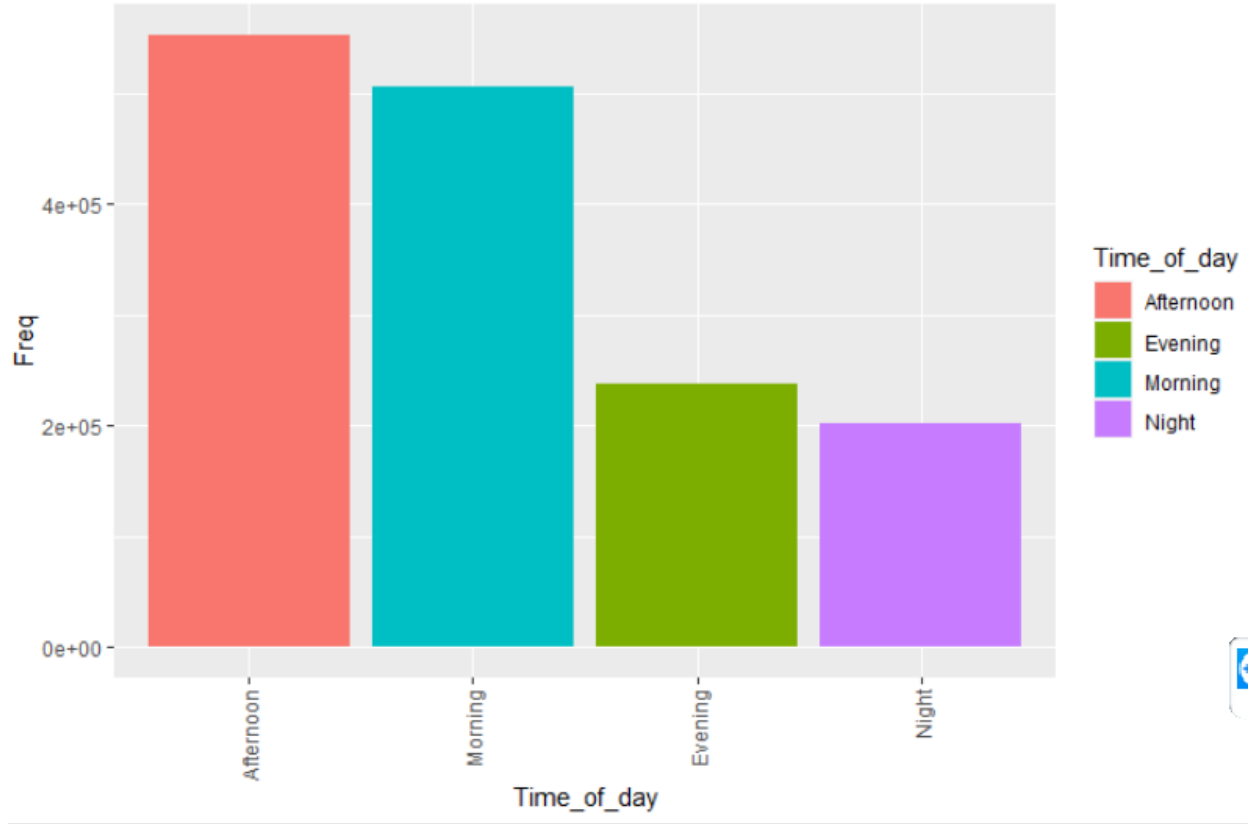
10- Average of weekday vs Average of weekend:



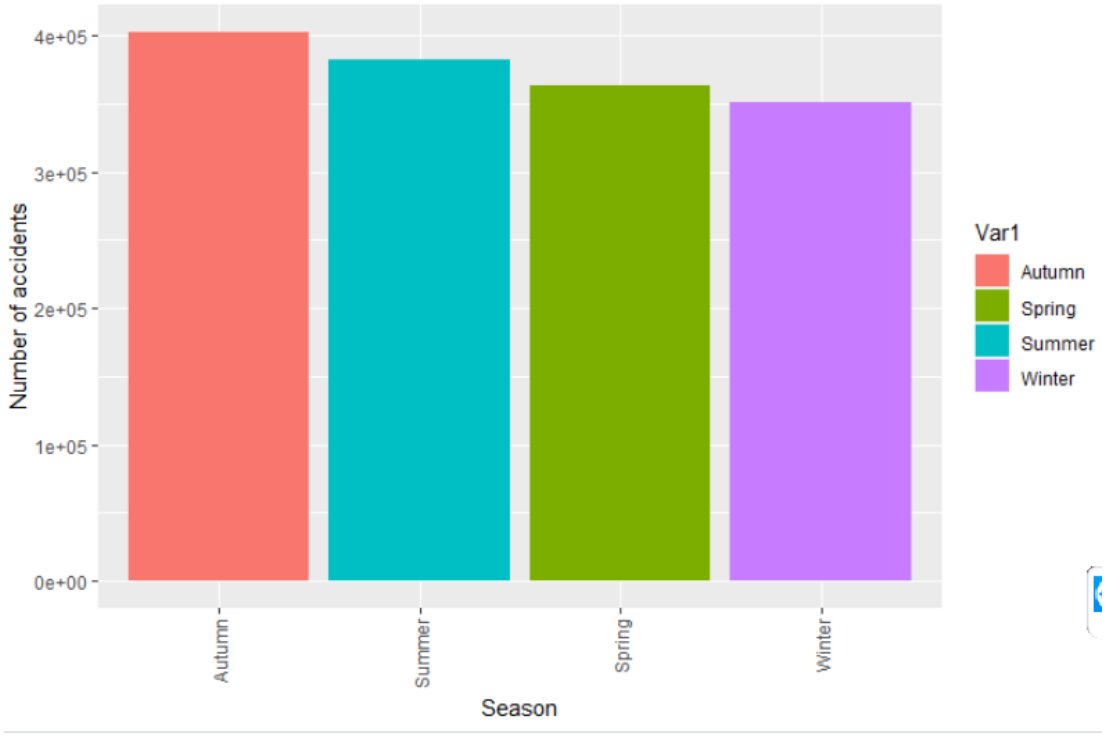
11- Relationship between speed limit and number of accidents:



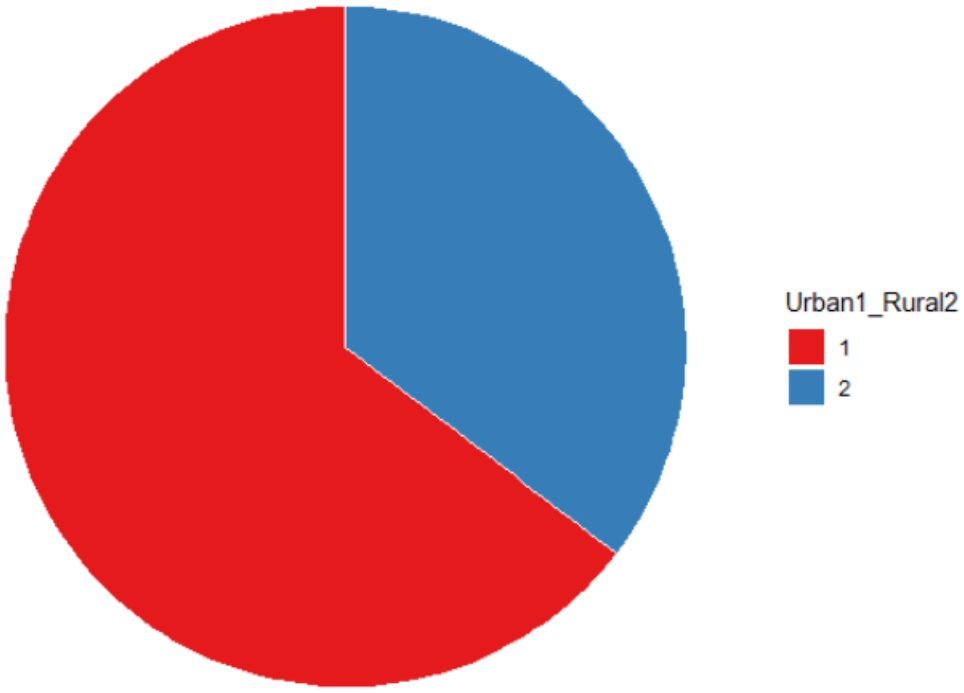
12- Relationship between Number of accidents and time of day:



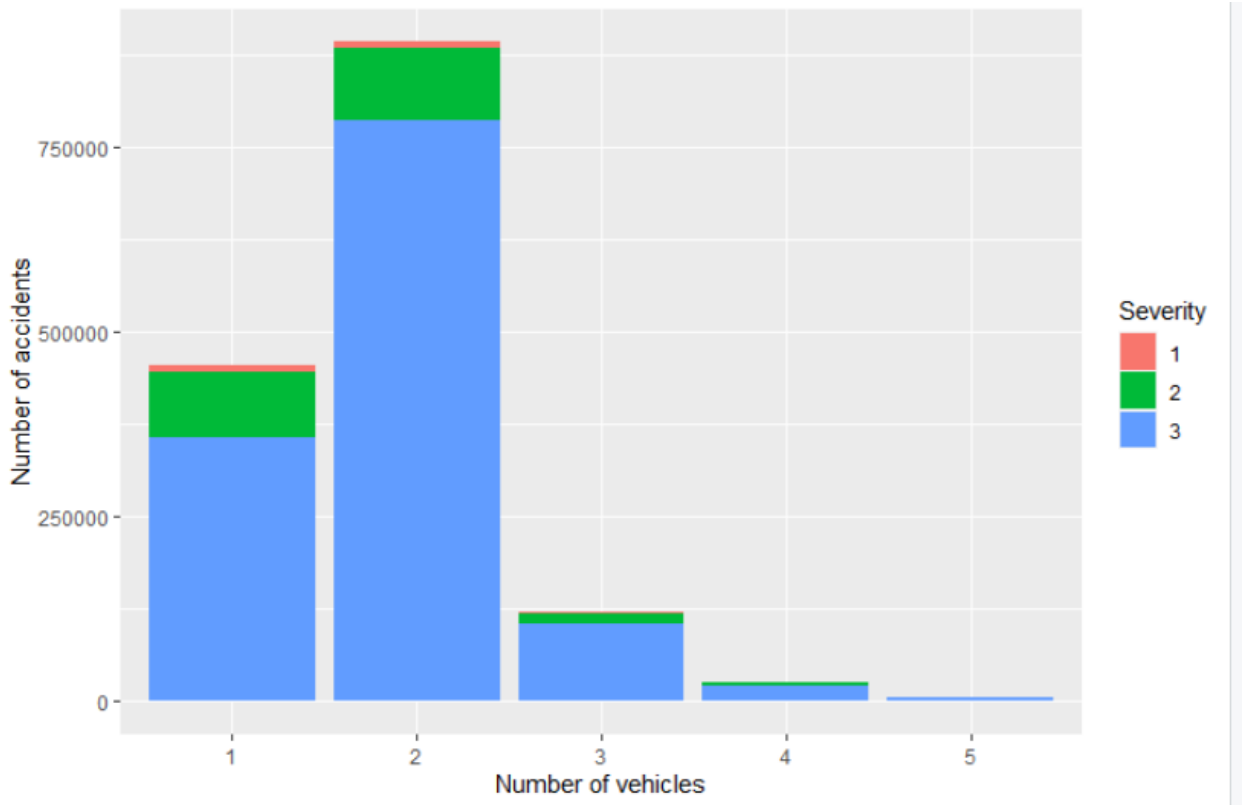
13- Relationship between number of accidents and seasons:



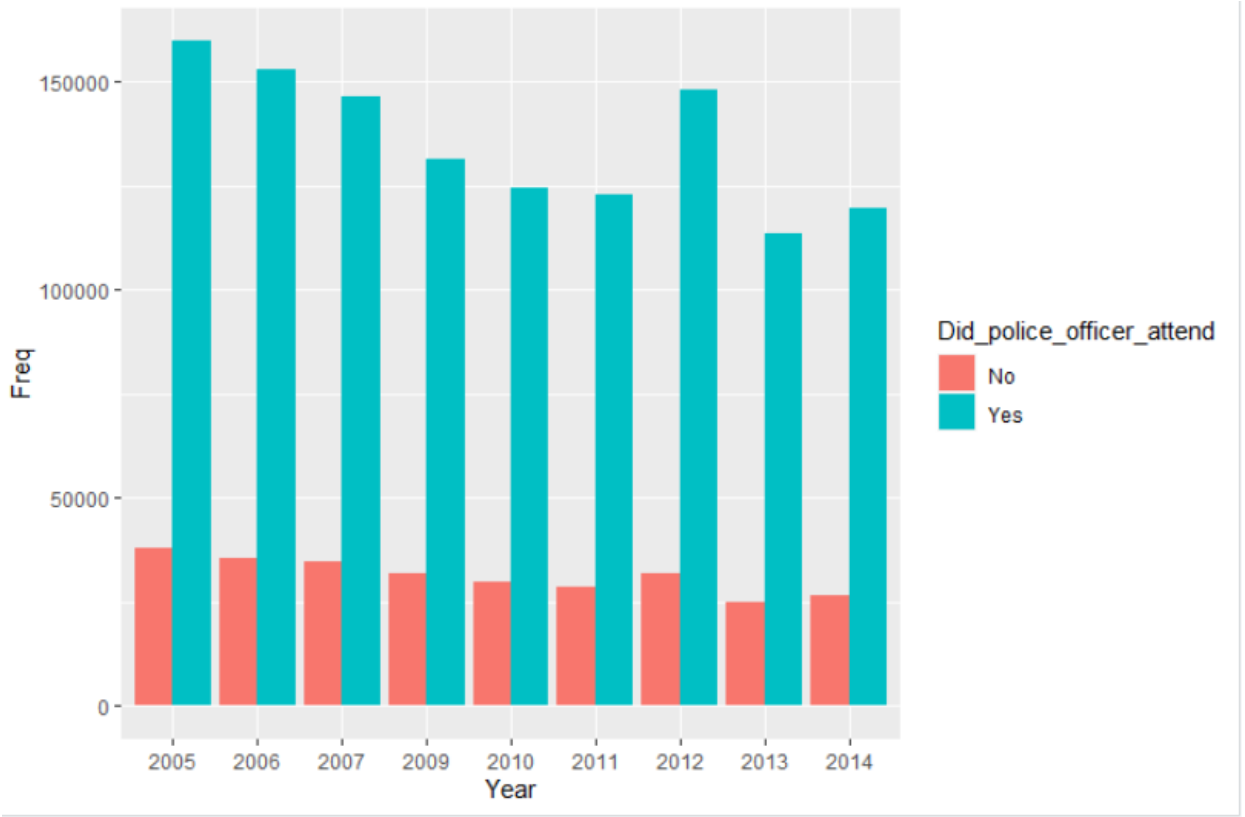
14- Number of accidents in urban areas vs Number of accidents in rural areas:



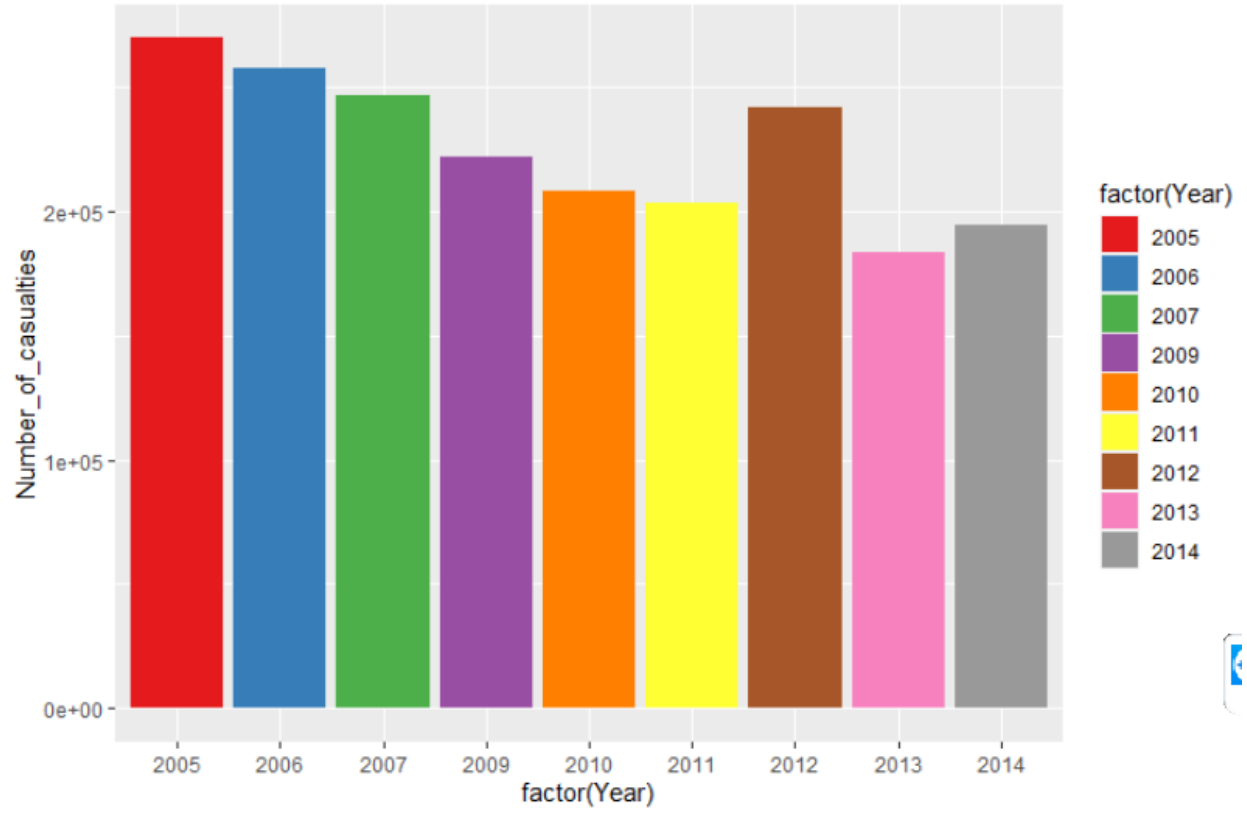
15- Relationship between number of vehicles, severity and number of accidents:



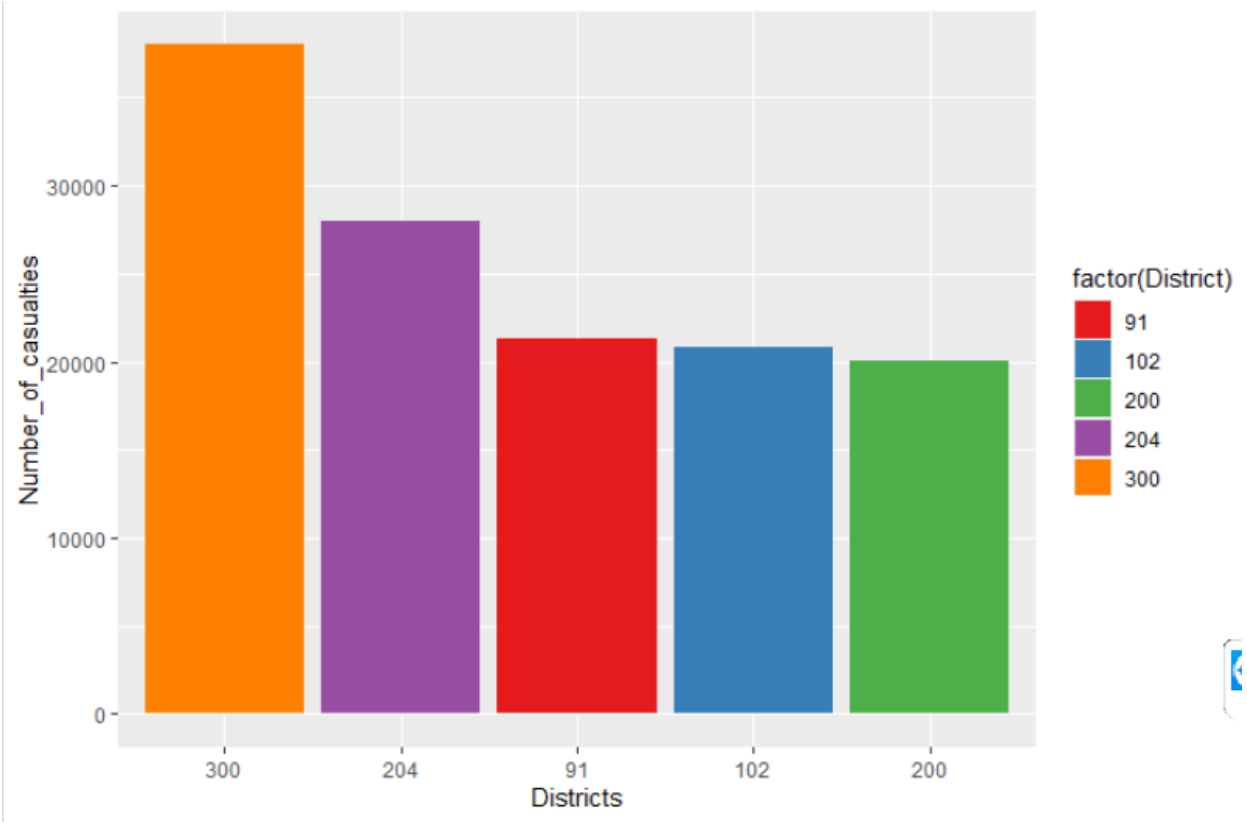
16- “Did police officer attend the scene of the accident” over years:



17- Number of casualties over years:



18- Top 5 districts in number of casualties:



Data insights:

1- District 300 has the highest number of accidents, it is almost two times any of the 20 most frequent districts, but by investigating over years, It was very high till 2007 then it started to decrease since then then became stable, so obviously there was a problem that was solved by 2007.

2- The most frequent 20 districts must be studied deeper to know the causes of their high number of accidents and try to control this problem.

3- By observing the most frequent 5 districts over year, most of the districts had a peak in 2012.

4-Year 2005 had the highest number of accidents then it gradually decreased but had an increase in 2012 then continued decreasing gradually again.

5-Top 10 frequent locations have to be maintained quickly to stop accidents from happening again there, for example the first one had about 70 accidents which is a suspicious number, so it may have infrastructure problems.

6-Local Authority (Highway) is the council responsible of maintaining and protecting public rights of way. The shown most frequent 20 authorities must be asked for maintaining those roads.

7- Accidents happen at junctions more often than accidents not at junctions especially for junctions that are uncontrolled.

8- Junctions have to be controlled to avoid accidents, by observing from dataset the junctions controlled by authorized person have the least number of accidents. There is a significant difference between it and automatic traffic signals.

9- Road class 3 has the highest number of accidents followed by 6, but 1, 2, 4, 5 have significantly less number of accidents than both of them.

10- Road type of single carriageway which is undivided highway with two or more lanes has the highest percentage of accidents (almost 80%) which makes sense as it has to be separated and controlled.

11- Average number of accidents in weekdays is less than in weekends. This is due to the high traffic.

12- Most of the accidents happen in regions of speed limit 30.

13- Most of the accidents happen in Afternoon that includes rush hour, followed by Morning but Evening and Night are significantly less than them.

14- Autumn has the highest number of accidents because of wind and storms, so a solution like Windbreaks must be held.

15- Almost 75% of the accidents happen in Urban areas.

16- Most of the accidents happen between 2 vehicles. These accidents are often of severity 3.

17-Accidents between more than 3 vehicles rarely happen.

18- Severity 3 accidents are the most common.

19- Police attendance to scene of accidents is decreasing over years.

20- Number of casualties is decreasing gradually over years, it had only a slight increase in 2012 then continued decreasing.

21- District 300 has the highest number of casualties.

Time series analysis model:

Model building and training:

Here, the dataset is divided to training set and test set. Training set includes accidents from 2005 to 2011 which are about 1 million rows, and Test set includes accidents from 2012 to 2015 which are about 500,000 rows.

Time series analysis will be done on months. This model will predict the number of accidents in the next few months.

We will construct a model that will be tested with test set first then with a modified test set, then we will compare the results later.

Time series objects:

Training set time series object:

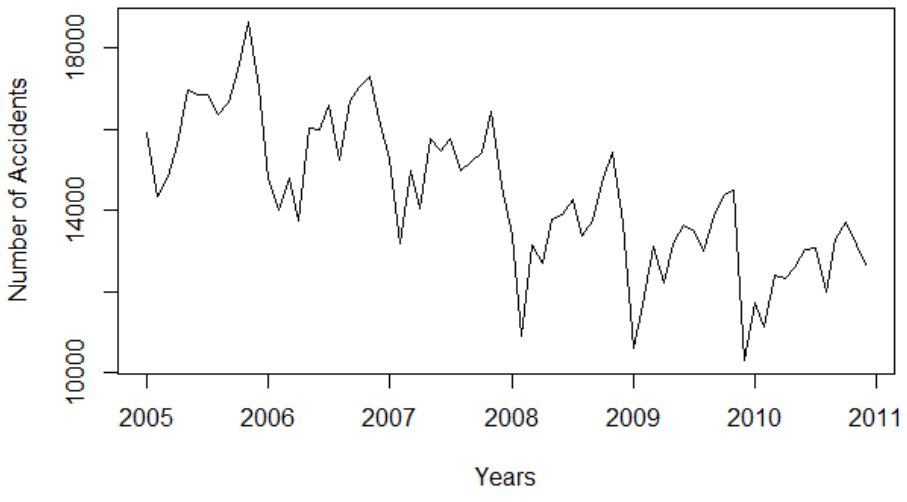
```
      Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
2005 15929 14359 14819 15614 16977 16838 16860 16365 16681 17459 18645 17016
2006 14791 14018 14801 13750 16042 15980 16599 15246 16705 17035 17300 16102
2007 15275 13189 14973 14066 15784 15465 15783 14974 15196 15442 16451 14620
2008 13390 10927 13171 12694 13787 13908 14258 13377 13758 14803 15440 13681
2009 10609 11697 13128 12228 13183 13623 13496 12991 13877 14394 14513 10325
2010 11738 11123 12400 12323 12611 13034 13097 12001 13312 13715 13156 12657
```

Test set time series object:

```
      Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
2012 14969 13726 15365 13607 15246 13994 15623 14399 15915 16010 16189 14392
2013 10196  9692 10257  9947 11425 11828 12996 11748 11985 13300 13143 11875
2014 12071 10766 11906 10992 12281 12509 13012 12089 11822 13421 13209 12021
```

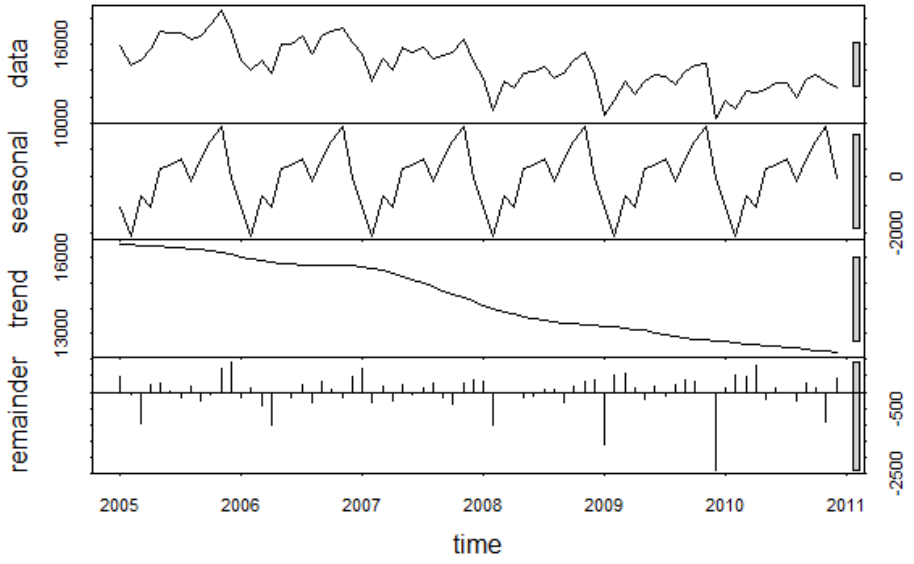
Data visualization:

Training set:



As observed from this plot, both mean and variance are not constant over years.

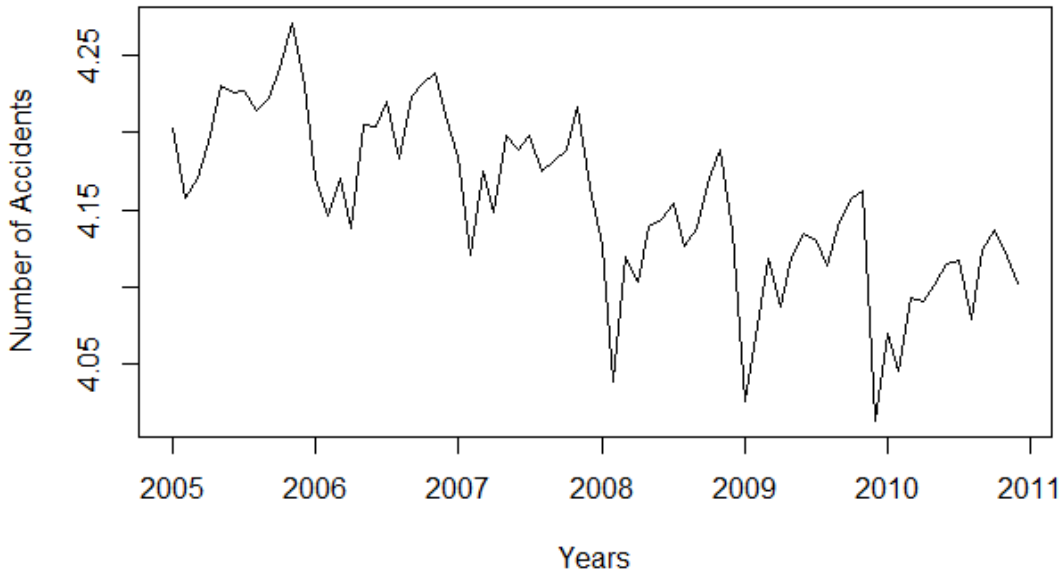
Time series components:



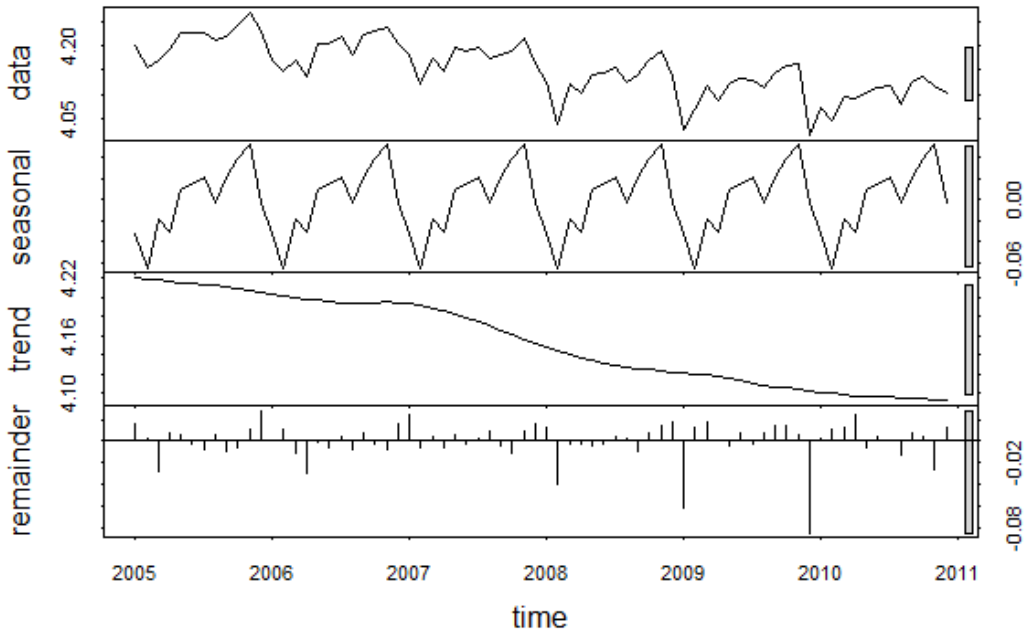
As observed from this plot, there is linear trend as generally it's decreasing linearly and seasonality as the pattern repeats itself each year.

After applying the Logarithmic transformation:

The variance becomes constant over years.



Time series components after applying Logarithmic transformation:



Model Building:

The model used is ARIMA model. We will pass to the model the data after applying logarithmic transformation only, and the model will internally handle the detrending to have constant mean and seasonal adjustment to eliminate seasonality.

```
ARIMA(4,0,0)(2,1,0)[12]                    : -246.5883
ARIMA(4,0,0)(1,1,0)[12]                    : -246.747
ARIMA(4,0,0)(0,1,0)[12]                    : -237.3564
ARIMA(4,0,0)(1,1,1)[12]                    : -246.0171
ARIMA(4,0,0)(0,1,1)[12]                    : -248.2339
ARIMA(4,0,0)(0,1,2)[12]                    : -246.1651
ARIMA(4,0,0)(1,1,2)[12]                    : Inf
ARIMA(3,0,0)(0,1,1)[12]                    : -246.9426
ARIMA(5,0,0)(0,1,1)[12]                    : Inf
ARIMA(4,0,1)(0,1,1)[12]                    : Inf
ARIMA(3,0,1)(0,1,1)[12]                    : Inf
ARIMA(5,0,1)(0,1,1)[12]                    : Inf

Best model: ARIMA(4,0,0)(0,1,1)[12]
```

So, the best candidate is:

(4,0,0): non-seasonal part (p, d, q)

(0,1,1) [12]: seasonal part [P, D, Q] [S]

Model Training:

By evaluating this model, it was found to be good as the measured error is so small.

```
> summary(ARIMAfit)
Series: (log10(data2))
ARIMA(4,0,0)(0,1,1)[12]

Coefficients:
         ar1     ar2     ar3     ar4     sma1
      0.2417  0.1767  0.2822  0.2579  -0.6406
s.e.  0.1519  0.1299  0.1305  0.1311   0.2118

sigma^2 estimated as 0.0007257:  log likelihood=130.91
AIC=-249.82   AICc=-248.23   BIC=-237.25

Training set error measures:
                       ME       RMSE        MAE          MPE      MAPE      MASE
Training set -0.002021495 0.02354526 0.01514859 -0.05030688 0.3684128 0.5082779
```

Model Evaluation:

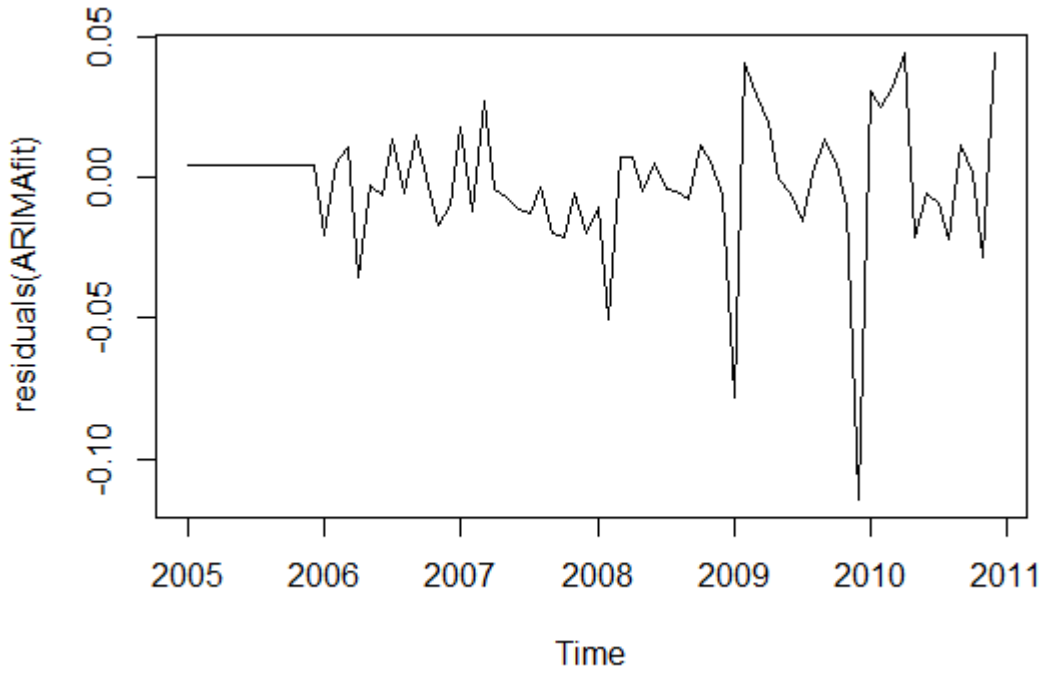
1- On Training set:

Model accuracy on training set:

```
> accuracy(ARIMAfit)
                       ME       RMSE        MAE          MPE      MAPE      MASE         ACF1
Training set -0.002021495 0.02354526 0.01514859 -0.05030688 0.3684128 0.5082779 -0.05869658
>
```

Model residuals:

(This represents the difference between the actual training points and the corresponding fitted values)



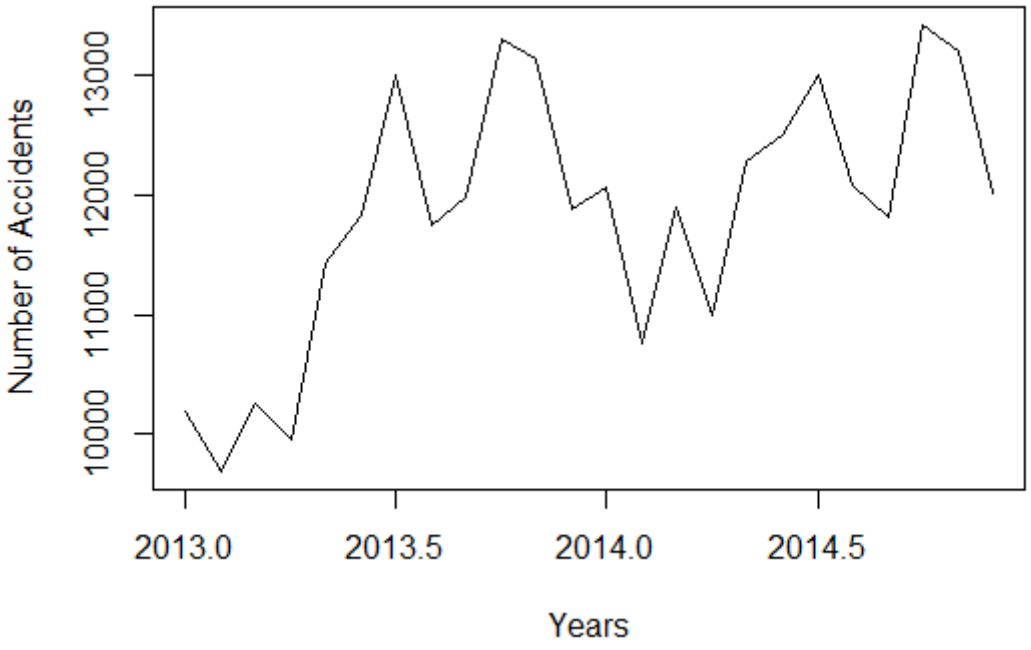
2- On Test set:

Test set:

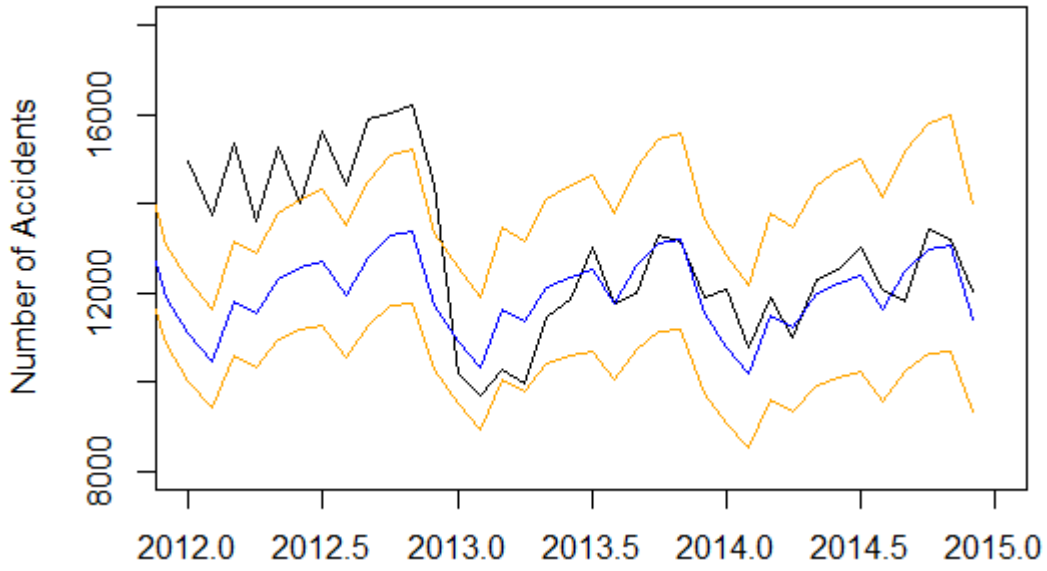
Here, we can clearly notice that 2012 has extremely high number of accidents compared to 2013 and 2014. So, we will have a huge drop between 2012 and 2013, this event will surely affect the accuracy of the model during testing.

So, we will modify the test set by removing year 2012.

Modified Test set:



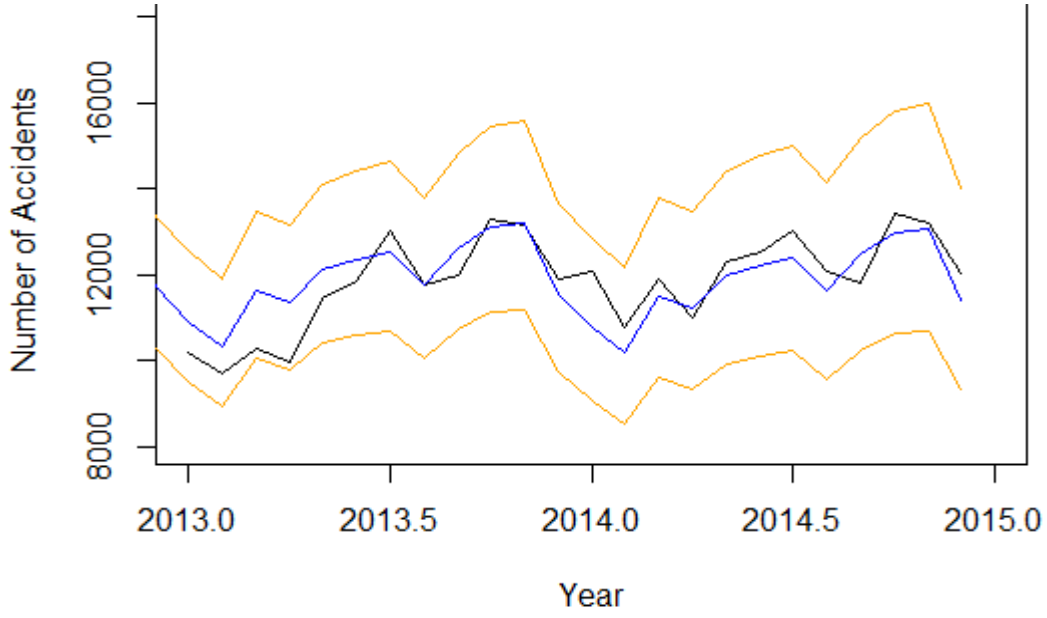
Evaluation on the original test set:



Model accuracy on original Test set:

```
> accuracy(10^(pred$pred),data3)
                ME     RMSE      MAE      MPE     MAPE      ACF1 Theil's U
Test set 920.6641 1752.106 1305.618 5.88102 9.491546 0.7636036  1.198947
```

Evaluation on the modified Test set:



Model accuracy on modified Test set:

```
> accuracy(10^(pred$pred),data3modified)
                ME     RMSE      MAE         MPE     MAPE      ACF1 Theil's U
Test set -30.0976 657.7484 547.333 -0.5960062 4.819783 0.5399315  0.749816
```

- Using both test sets, the accuracy of the model when tested on modified test set is much higher, and the error significantly decreased. So, we can conclude that using accurate test set gives better and more accurate results.

Severity prediction model:

Model building and training:

This model goal is to predict severity of the accident based on some accident information. Severity has 3 values: 1 = Fatal, 2 = Serious and 3 = Slight.

Here, the dataset is divided into training set and test set, 75% for training set and 25% for test set.

But the training set was biased to severity 3, so it was unbalanced which made it predict only severity 3. This problem was fixed by deleting random rows of severity 3 using sample function to have a balanced dataset containing all severity classes.

Three trials were done by three different classifiers:

1- Naïve Bayes Classifier - as the features affecting the severity are categorical not continuous.
2- Neural Networks – using nnet package.

3- Random forest.

We noticed that the accuracy is not high in any model even after fixing the unbalanced dataset issue so this might be due to non-effectiveness of the features in the data itself.
Also , the training set accuracy isn't high , this is an underfitting case due to the high bias and the non-effectiveness of the features in the data itself , which makes the model miss the relevant relations between features and target output .

Model Evaluation:

We will evaluate each model by tabulating the confusion matrix and calculating its accuracy.

1-- Naïve Bayes Classifier:

Confusion matrix:

	Actual 1	Actual 2	Actual 3
Predicated 1	918	1094	2144
Predicated 2	239	342	334
Predicated 2	3611	11149	22522

Training set:

Accuracy:

Accuracy of the training set is equal to 56.53 %.

Test set:

Accuracy:

Accuracy of the testing set is equal to 56.15 %.

Precision:

Precision c1: (918)/(918+1094+2144)=22%
Precision c2: (241)/(239+342+334)=37.4%
Precision c3: (22522)/(11149+22522+3611)=82.5%
Overall precision: (82.5+37.4+22)/3=47.3%

Recall:

Recall c1: (918)/(918+239+3611)=19.3%
Recall c2: (342)/(342+1094+11149)=2.7%
Recall c3: (22522)/(334+2144+22522)=91.3%

Comments:

The accuracy is low, it is biased towards class 3 and this is obvious by the percentage of Recall of c3 which means that the data is classified as c3 91% of the time .
Due to the issues discussed above, we tried to increase accuracy by using Neural Networks.

2—Neural networks:

We monitored the performance and the best one was at decay = 0.2 and iterations = 300 .

Confusion matrix:

	Actual 1	Actual 2	Actual 3
Predicated 1	333	255	282
Predicated 2	482	853	704
Predicated 3	3953	11477	24014

Training set:

Accuracy:

Accuracy of the training set is equal to 56.53 %.

Test set:

Accuracy:

Accuracy of the testing set is equal to 59.5 %.

Precision:

Precision c1: (333)/(333+255+282)=38.2%
Precision c2: (853)/(853+482+704)=41.8%
Precision c3: (24014)/(24014+11477+3953)=40.3%
Overall precision: (38.2+41.8+40.3)/3=40.1%

Recall:

Recall c1: (333)/(333+482+3953)=7%
Recall c2: (853)/(853+255+11477)=7 %
Recall c3 is a very large value .

Comments:

The accuracy is slightly better than that given by the Naïve Bayes Classifier.
We trained the model as following :
modelNN<-*nnet(as.factor(Accident_Severity)~Number_of_Vehicles +Day_of_Week+X1st_Road_Class+Road_Type+Speed_limit+ X2nd_Road_Class+Road_Surface_Conditions,data=training,size =8 ,decay = 0.2,maxit = 300)*
We used the decay as 0.2, we compromised it not be very low so it won't converge slowly and not very high so it won't oscillate around local minima .
For the number of iterations, if we extend it to 700-1000 or decrease it to 100-150 it gives lower accuracy .
The size is the number of the neurons of the hidden layer , the default in the nnet package is 1 hidden layer and the size is left to be decided ;6-8 gives good performance .

3— Random forest:

We monitored the performance and the best one was at decay = 0.2 and iterations = 300 .

Confusion matrix:

	Actual 1	Actual 2	Actual 3
Predicated 1	306	197	255
Predicated 2	435	715	562
Predicated 3	4027	11673	24183

Training set:

Accuracy:

Accuracy of the training set is equal to 60 %.

Test set:

Accuracy:

Accuracy of the testing set is equal to 59.51%.

Precision:

Precision c1: (306)/(306+197+255)=40.36%
Precision c2: (715)(715+435+562)=41.8%
Precision c3: (24183)/(24183+11673+4027)=60.6%
Overall precision: (82.5+37.4+22)/3=47.5

Recall:

Recall c1: (306)/(306+435+4027)=6.4%
Recall c2: (715)/(715+197+11673)=2.7%
Recall c3: (24183)/(24183+562+255)=96.7%

Comments:

The accuracy , Precision and recall are still almost the same as the neural network model .

Conclusion and future work

As is it shown, analyzing the accidents has a great importance. It helped in predicting number of accidents and level of severity. These predictions can be used by the government, police and even hospitals. In the future we can predict the number of casualties according to the predicted severity.

References:

Dataset:

https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales?fbclid=IwAR0dGYRO__sdeIfxN0T-2pLgnHB2wYj1A3IY23ix5siPDjMO4-GvUYDP5QU