Intelligent Systems



by Wladyslaw Eysymontt, Mateusz Klimas,

Juan Luis, Artjom Poljakow and Jorge Martín Lasaosa

*Intelligent Systems*

*Dr. X*

*Universidad Politécnica de Madrid*

*December 10, 2019*

# Part 1

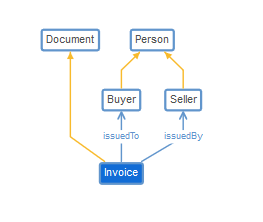
The following situations must be modeled using n-ary relation pattern:

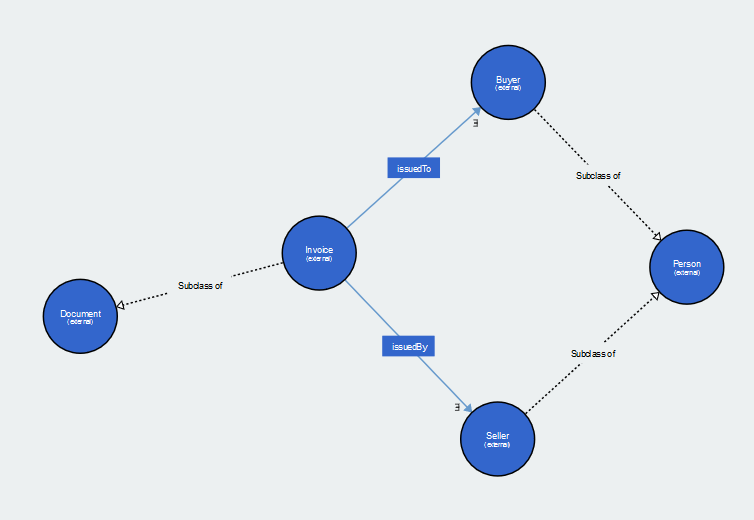
## S1: An invoice is a commercial document issued by a seller to a buyer

### Functional representation

Invoice(Buyer,Seller)

### Graphical representation

****

****

### Owl code

The different links to access the code are:

* [WebProtegé S1](https://github.com/jomartla/IntelligentSystemsAssignment/blob/master/Part%201/Turtle/WebProtege_S1.ttl)
* [WebVOWL S1](https://github.com/jomartla/IntelligentSystemsAssignment/blob/master/Part%201/Turtle/WebVOWL_S1.ttl)

# Data analysis plan

The used dataset is flagged as cleaned by the offering source, nevertheless adjustments needed to be done considering the task.

In order to answer the questions of the next step of the assignment, we are going to use the features bla, bla, blas which are blab la bla.

To work with the severity and see how it is influenced by the age of the car and car type the following measures must be taken. To perform numerical operations on categorical variables these were transformed appropriately. The most important was the severity which is given as string values and needs to be transformed to 0 and 1 for statistical operations. The given vehicle categories are categorical as well. These need to change to 1 for cars and taxis and 2 for motorcycles. Both categories contain over 90 % of the given samples. As a conclusion the minorities of vehicle types were omitted as a cleaning measurement.

To inspect the influence of the seasons the category of them and years needs to be combined and elaborated. Based on that the amount of accidents depending on the season may help to see if there is a significant context between them.

# Analysis for each research question

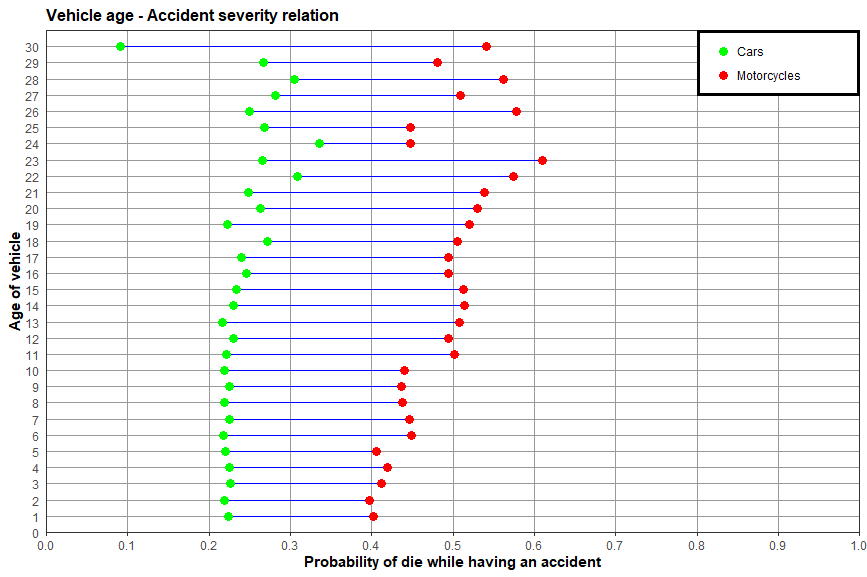
## What useful knowledge can we extract analysing this dataset?

This question will be answered with the rest of the research questions. The goal is to explore the dataset, understand better the variables we have and try to recognize the patterns. This way we will be able to make a data analysis plan based not only on our presumptions, but also on extracted knowledge. Two different approaches will be used to analyse the data:

* Death rate, meaning the probability for driver to die while having an accident.
* The distribution of the number of accidents through the time.

## How is the severity of an accident affected by the age of the vehicle? Or by the type of the vehicle?

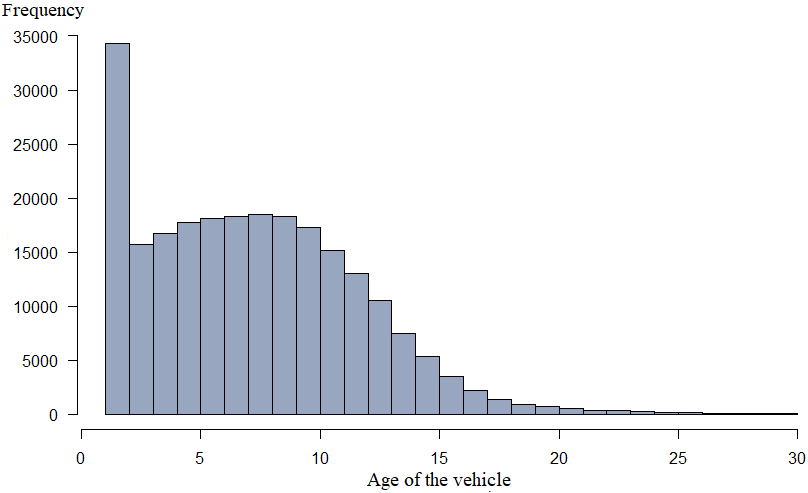
For analyse these relations we are going to use a lollipop type and a bar type charts, one for the age of the vehicle – death rate and one for the speed limit – death rate relations, where the death rate is calculated as a sum of the lethal accidents in a sub dataset divided by a total number of the accidents of this same sub dataset. In both graphics we will plot the results for cars and motorcycles.



Graph 1: Accident severity – vehicle age relation

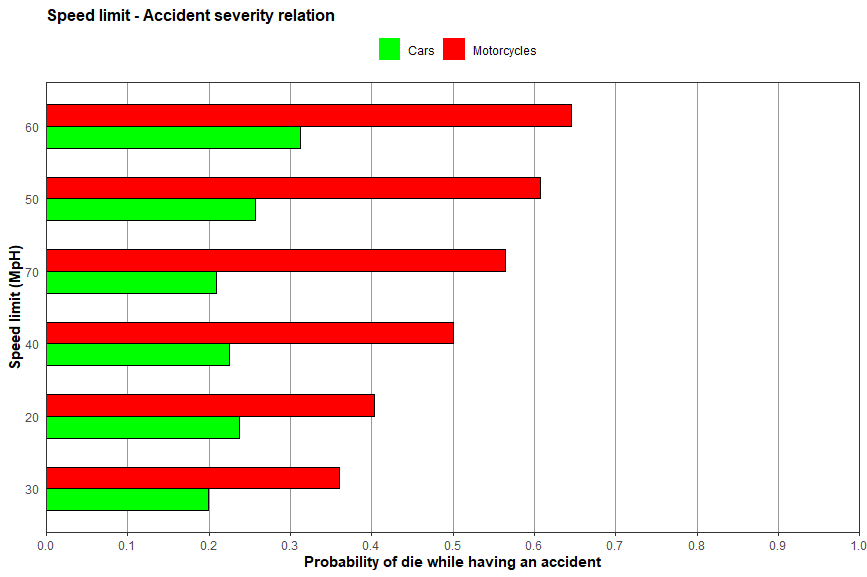
In the first chart we can observe how the probably of lethal result of an accident is distributed across the age of a vehicle for both, motorcycles and cars. We had been expecting more pronounced positive correlation between age of the cars and the death rate because of the technological advances in security systems, but, as we can see, there are no clear differences comparing the cars no older than 17 year. On the other hand, the graphic shows us that the older the motorcycle is, the more dangerous it becomes, although we could have been thinking they should be more or less the same. It is shown by an increase of the probability to die from around 40% up to 60%. An interesting spike is shown by 30 year old cars. The probability drops from around 25% to less than 10%. An assumption can be make that this cars are considered historical and therefore taken care of by their drivers. At the same time the majority of this drivers are at least 45 years old and have more practice. This assumptions need to be verified, with a bigger dataset which was not given here.

From 17 years old vehicles on we have much less data, so the points could be less representative, the clear example of this is the 30 years old cars point. We can see it easily in a simple data distribution chart for the age of the vehicle variable:



Graph 2: Vehicle age distribution

## How does the speed limit affect the death rate depending on the vehicle?

****

Graph 3: Accident severity - speed limit relation

Here we are using speed limit variable in miles per hour instead of age of a vehicle and we also tried to group the results by the mean death rate for motorcycle so we could see intervals of the roads with what speed limits are more dangerous for users of this type of vehicles. It is visible that the probability is around to times higher if you use a motorcycle. As expected, the three most dangerous ones are the highest ones: 50, 60 and 70 miles per hour.

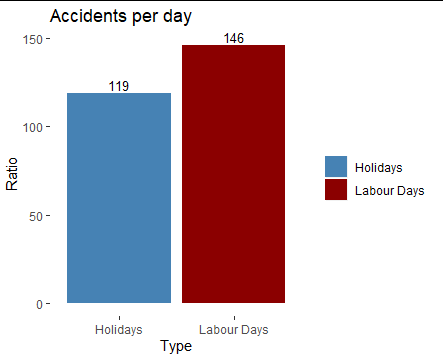
As it was in the “Vehicle age - Accident severity” chart, the correlation for the cars is not so clear as expected.

## Do more accidents occur during labour days?

In order to answer the question, we must define the difference between working days and holidays. To do so, we use a REST API service from [www.clarendarific.com](http://www.clarendarific.com) through which we send a GET request (HTTP protocol) with the country and the year as parameters. As a response we obtain the holidays in a data frame with much more information such as which regions are affected, the religious origin if it exists, etc.

Once we have obtained the answer from the service, we can use the information to divide the original accidents data in two groups. The first one has the accidents occurred on holidays, considering that weekends are holidays too. The second group contains the labour days accidents.

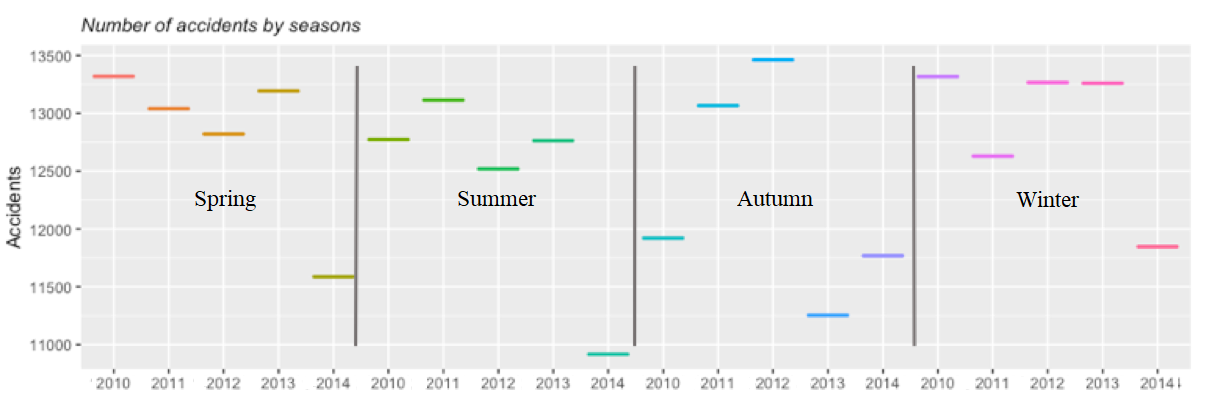
Finally, we can use the results to obtain the accidents per day ratio in both working days and holidays. The results are represented in the following graph:



Graph 4: Accidents per day in holidays and labour days

As we can see, there is a 22% increase in the number of accidents produced on labour days. So, we can conclude that more accidents occur during these days.

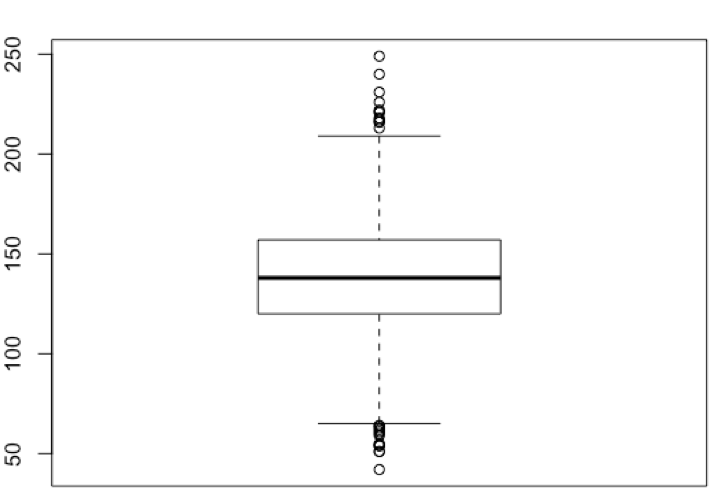
## How is the number of accidents distributed through time?



Graph 5: Number of accidents by seasons

The graph at the top (Graph 5) is a plot which represents the number of accidents that have occurred in each of the seasons of the year, being 1: spring, 2: summer, 3: autumn and 4: winter. The boxplot boxes are stacked as only one variable is used in the analysis.

The performance of spring, summer and autumn was similar, except that in spring the number of accidents was lower than in the other seasons and in winter it was the highest. And all of the years follow the same pattern in all the seasons except the year 2014 that have always the minimum number of accidents except in Autumn.

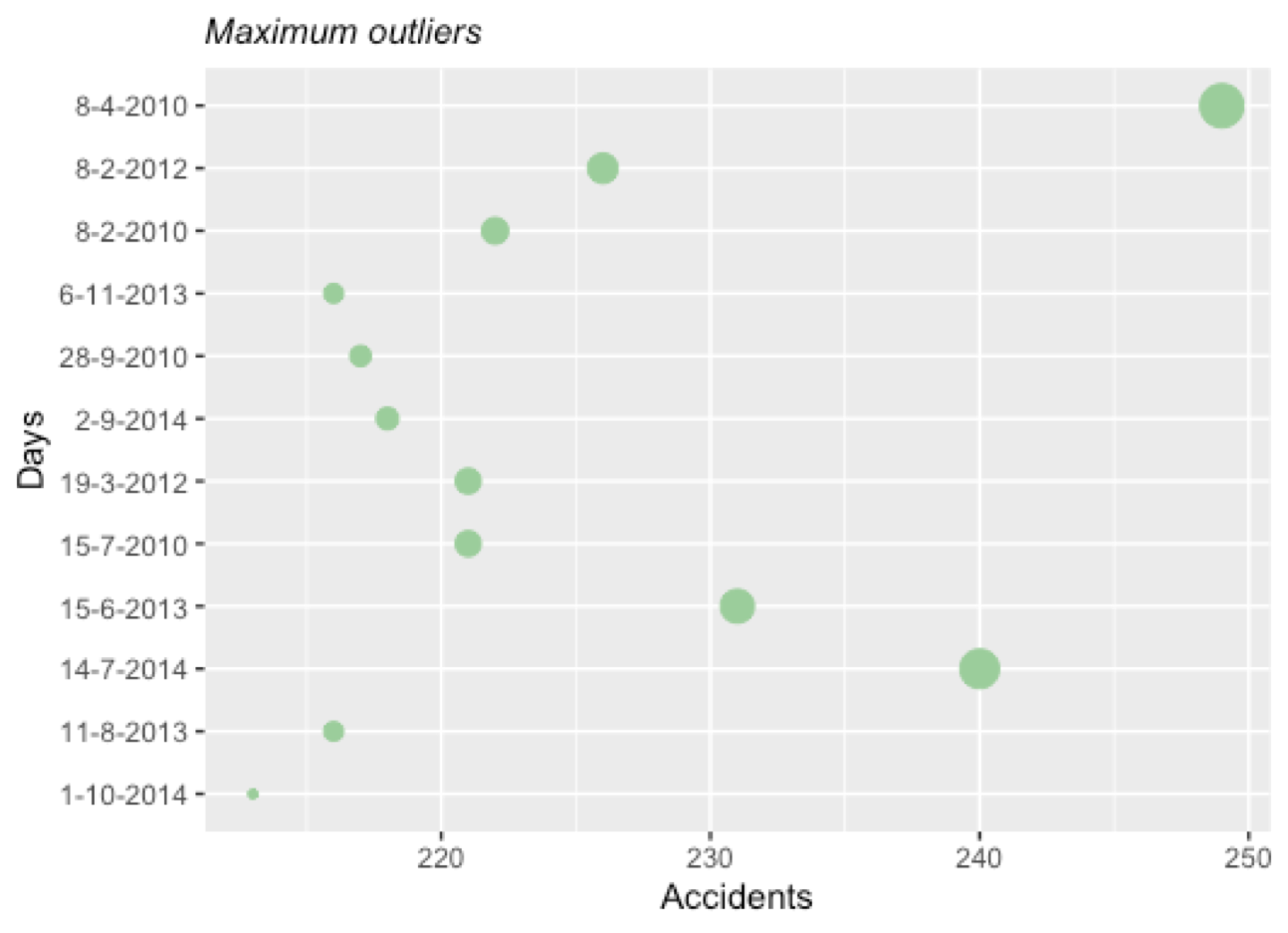
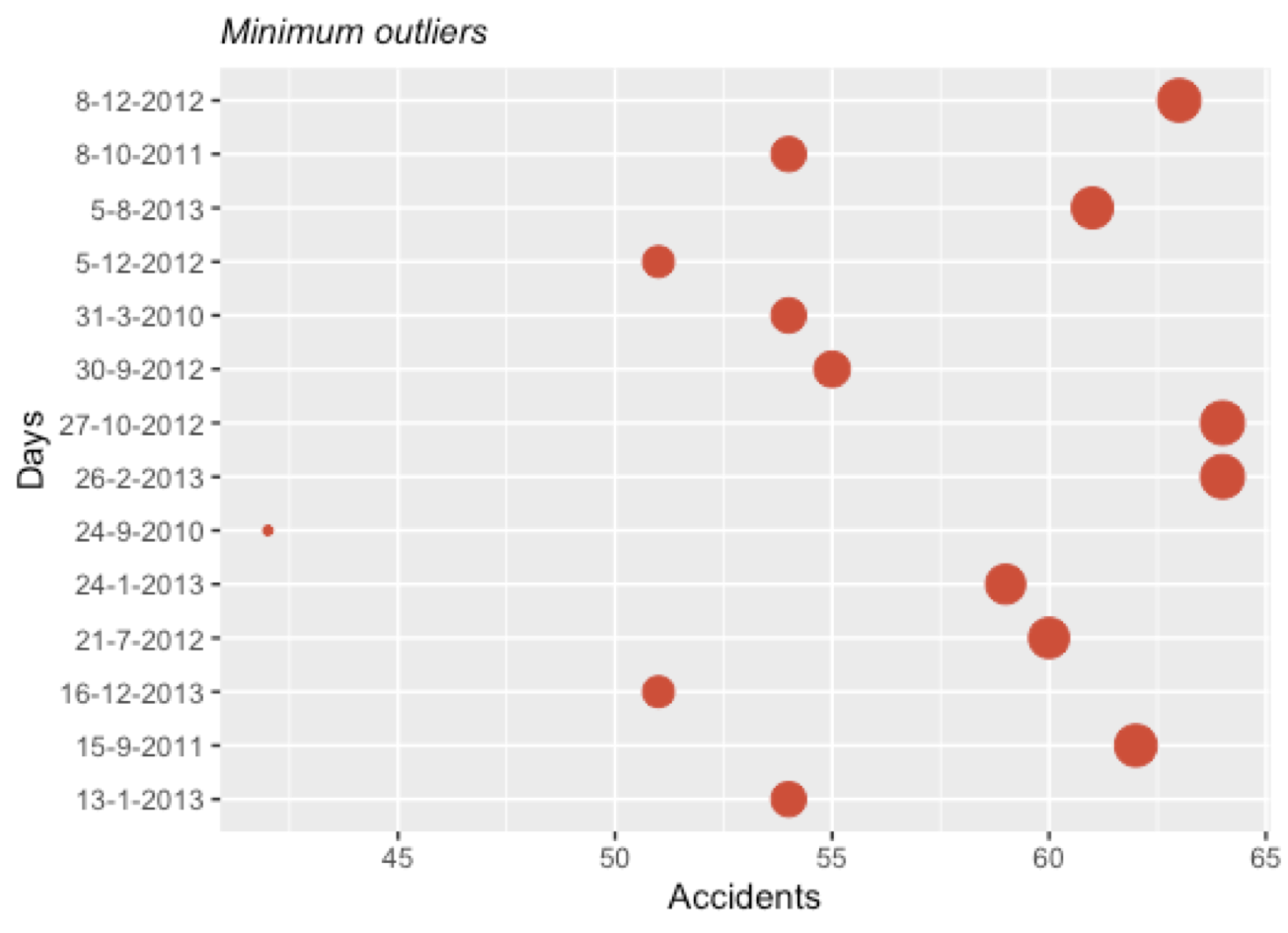


Graph 6: Boxplot of the number of accidents per day in United Kingdom

In graph 6, it can be seen how the interval of number of accidents in the UK is around 120-157 accidents, this represents the 50% of the days. Reaching an average number of accidents per day of 138, a maximum of 249 accidents and a minimum of 42 accidents per day.

To analyse the outliers in our database, we use the following formulas:

* Maximum Value in the Data: Q3 + 1.5\*IQR
* Minimum Value in the Data: Q1 + 1.5\*IQR



Graph 7: Minimum outliers Graph 8: Maximum outliers

Graph 8 shows the maximum outliers, ergo, all those days that have had more than 212 accidents per day. The day with the highest number of accidents is 2010-04-08. Furthermore, most of these outliers are between 212 and 225 accidents. And finally, it should be noted that not every month during the years 2010-2014 there have been atypical values, the months of January, May and December do not present any outlier.

Graph 7 shows the minimum outliers, which have reached a value of less than 64 accidents per day. The day on which the lowest value of accidents was reached was

2010-09-24. Most of these outliers are in the range of 55-65 accidents per day, with 2012 being the year in which the fewest accidents occurred. Another noteworthy fact is that not all the months of the year have outlier’s data, in the months of April, May, June and November there were no accidents below normal.

In 2014, there was no below average accidents either.

# References

The dataset is available under the following link:

<https://www.kaggle.com/stefanoleone992/adm-project-road-accidents-in-uk>

The base code for plots were taken from:

<https://www.r-graph-gallery.com/index.html>

<http://shinyapps.stat.ubc.ca/r-graph-catalog/>

How to create a data analysis plan:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4552232/>