# DM2 Pratical Assigment 1

*Pedro Belem, Rui Fonseca*

*18 March 2017*

## Data pre processing

After reading the data from the csv file, we decided that it would be more readable if we converted all the values in the data frame to their corresponding label in the xls file, so we decided to make it dynamically. We loaded the xls file and made a function that receives the column name and the correspondent excel section name and it returns the nominal values instead of the numeric codes.

```
# Reads information from excel guide, used to label the data
read_excel_allsheets <- function(filename) {
  sheets <- readxl::excel_sheets(filename)
  sheets <- sheets[3:length(sheets)]
  x <- lapply(sheets, function(X) readxl::read_excel(filename, sheet = X))
  names(x) <- sheets
  x
}


# Converts code values of col col_name to labels in excel_section_name
convert_codes <- function(df, col_name, excel_section_name) {
  df[[col_name]] <- sapply(df[[col_name]], function(i) if(!is.na(i)) subset(mysheets[[excel_section_nam
  df
}

mysheets <- read_excel_allsheets("Road-Accident-Safety-Data-Guide.xls")
```

After analising the data, we found some attributes that will be irrelevant to our task. Those are:

- Accident Index: Each number is unique to each accident so it doesn't contribute to any meaningful rules.
- Location Easting OSRG, Location Northing OSGR, Longitude, Latitude: all these attributes represent information about the local of the accident. We chose LSOA (Lower Layer Super Output Area) as the attribute representative of space. This will also help to group accidents in the same area together.
- Local Authorities !?!?
- 1st and 2nd roads !?!?
- Carriageway Hazards, Pedestrian Crossing (Human Control), Special Conditions at Site: We notice all these attributes had one class with >97% of frequency, so we decided to remove them from the data because there isn't frequent itemsets with classes different from the most frequent, if we consider a support bigger than 0.03% which we think is low.
- Did Police Officer Attend Scene of Accident: We think is irrelevant to the problem.
- Juncion Control: 40% of the data is missing

We decided to convert the date attribute into two separate columns (Day and Month), and to remove the minutes from the time, because we think it's not important to rule generation

```
# Converts date and time content into separated columns
convert_date_and_time <- function(df) {
  df$Date <- as.Date(df$Date, "%d/%m/%Y")
  df$Day <- day(df$Date)
  df$Month <- month(df$Date)
```

```
  df$Date <- NULL

  df$Time <- hm(df$Time)
  df$Hour <- hour(df$Time)
  df$Time <- NULL
  df
}
data.discretized <- convert_date_and_time(data.discretized)
```

In the excel it's explained that the missing values in the data is represented with -1 so we converted them to NA

```
data.discretized[data.discretized==-1] <- NA
```

## Data discretization

To effectively generate rules the data needs to be discretized

- Number of Vehicles: 1, 2, 3, 4, 5+
- Number of Casualties: 1, 2, 3, 4, 5+
- Day: split in 4 intervals, representing the weeks of the month.
- Hour: Hours of the day will be split in intervals of 3 hours each, starting at midnight.

```
# Discriteze all data



# Longitude
data.discretized$Longitude<- discretize(data.discretized$Longitude, method = "interval", categories = 1
# Latitude
data.discretized$Latitude<- discretize(data.discretized$Latitude, method = "interval", categories = 10)
# Police Force
data.discretized <- convert_codes(data.discretized, "Police_Force", "Police Force")
data.discretized$Police_Force <- factor(data.discretized$Police_Force)
# Accident Severity
data.discretized <- convert_codes(data.discretized, "Accident_Severity", "Accident Severity")
data.discretized$Accident_Severity <- factor(data.discretized$Accident_Severity)
# Number of Vehicles
n_vehicles.range <- c(1,2,3,4,5,Inf)
data.discretized$Number_of_Vehicles <- discretize(data.discretized$Number_of_Vehicles, method = "fixed"
# Number of Casualties
n_casualties.range <- c(1,2,3,4,5,Inf)
data.discretized$Number_of_Casualties <- discretize(data.discretized$Number_of_Casualties, method = "fi
# Day of Week
data.discretized <- convert_codes(data.discretized, "Day_of_Week", "Day of Week")
data.discretized$Day_of_Week <- factor(data.discretized$Day_of_Week)
# Local Authority District
data.discretized <- convert_codes(data.discretized, "Local_Authority_.District.", "Local Authority (Dis
data.discretized$Local_Authority_.District. <- factor(data.discretized$Local_Authority_.District.)
# Local Authority Highway
data.discretized <- convert_codes(data.discretized, "Local_Authority_.Highway.", "Local Authority (Highw
data.discretized$Local_Authority_.Highway. <- factor(data.discretized$Local_Authority_.Highway.)
# 1st Road Class
data.discretized <- convert_codes(data.discretized, "X1st_Road_Class", "1st Road Class")
data.discretized$X1st_Road_Class <- factor(data.discretized$X1st_Road_Class)
```

```r
# 1st Road Number
data.discretized$X1st_Road_Number <- factor(data.discretized$X1st_Road_Number)
# Road Type
data.discretized <- convert_codes(data.discretized, "Road_Type", "Road Type")
data.discretized$Road_Type <- factor(data.discretized$Road_Type)
# Speed Limit
data.discretized$Speed_limit <- factor(data.discretized$Speed_limit)
# Junction Detail
data.discretized <- convert_codes(data.discretized, "Junction_Detail", "Junction Detail")
data.discretized$Junction_Detail <- factor(data.discretized$Junction_Detail, exclude = NULL)
# Junction Control
data.discretized <- convert_codes(data.discretized, "Junction_Control", "Junction Control")
data.discretized$Junction_Control <- factor(data.discretized$Junction_Control, exclude = NULL)
# 2nd Road Class
data.discretized <- convert_codes(data.discretized, "X2nd_Road_Class", "2nd Road Class")
data.discretized$X2nd_Road_Class <- factor(data.discretized$X2nd_Road_Class, exclude = NULL)
# 2nd Road Number
data.discretized$X2nd_Road_Number <- factor(data.discretized$X2nd_Road_Number)
# Ped Cross - Human
data.discretized <- convert_codes(data.discretized, "Pedestrian_Crossing.Human_Control", "Ped Cross - Hu
data.discretized$Pedestrian_Crossing.Human_Control <- factor(data.discretized$Pedestrian_Crossing.Human_
# Ped Cross - Physical
data.discretized <- convert_codes(data.discretized, "Pedestrian_Crossing.Physical_Facilities", "Ped Cros
data.discretized$Pedestrian_Crossing.Physical_Facilities <- factor(data.discretized$Pedestrian_Crossing
# Light Conditions
data.discretized <- convert_codes(data.discretized, "Light_Conditions", "Light Conditions")
data.discretized$Light_Conditions <- factor(data.discretized$Light_Conditions)
# Weather Conditions
data.discretized <- convert_codes(data.discretized, "Weather_Conditions", "Weather")
data.discretized$Weather_Conditions <- factor(data.discretized$Weather_Conditions)
# Road Surface
data.discretized <- convert_codes(data.discretized, "Road_Surface_Conditions", "Road Surface")
data.discretized$Road_Surface_Conditions <- factor(data.discretized$Road_Surface_Conditions)
# Special Conditions
data.discretized <- convert_codes(data.discretized, "Special_Conditions_at_Site", "Special Conditions a
data.discretized$Special_Conditions_at_Site <- factor(data.discretized$Special_Conditions_at_Site)
# Carriageway Hazards
data.discretized <- convert_codes(data.discretized, "Carriageway_Hazards", "Carriageway Hazards")
data.discretized$Carriageway_Hazards <- factor(data.discretized$Carriageway_Hazards)
# Urban Rural
data.discretized <- convert_codes(data.discretized, "Urban_or_Rural_Area", "Urban Rural")
data.discretized$Urban_or_Rural_Area <- factor(data.discretized$Urban_or_Rural_Area)
# Police Officer Attend
data.discretized <- convert_codes(data.discretized, "Did_Police_Officer_Attend_Scene_of_Accident", "Poli
data.discretized$Did_Police_Officer_Attend_Scene_of_Accident <- factor(data.discretized$Did_Police_Offi
# Day
data.discretized$Day <- discretize(data.discretized$Day, method="interval", categories=4)
# Month
data.discretized$Month <- factor(data.discretized$Month)
# Hour
hour.range <- c(3,6,9,12,15,18,21,Inf)
data.discretized$Hour <- discretize(data.discretized$Hour, method="fixed", categories=hour.range)
```

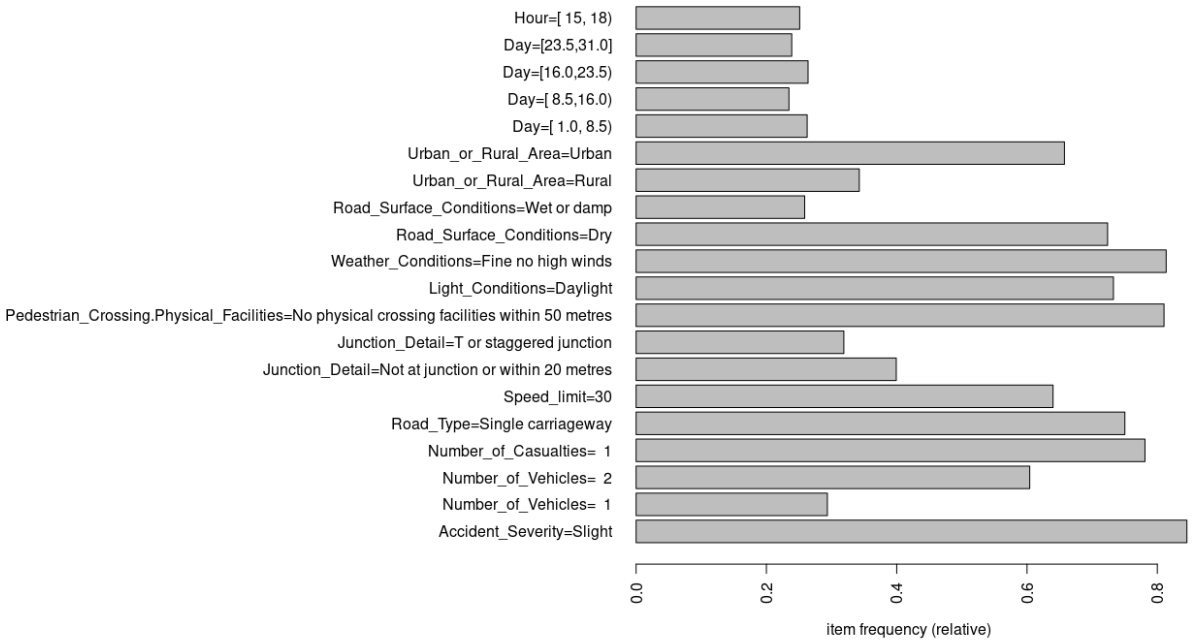


Figure 1: Frequent Items considering support of 20%

## Data Analysis

```
itemFrequencyPlot(as(data.discretized, "transactions"), support = 0.2, horiz = TRUE)
```

We can see that most accidents happen between 15h and 18h, and that the number of accidents is well distributed in each quarter of the month. There's more accidents in a urban area than in a rural, but they are both frequent. It's also frequent to have accidents in a dry or a wet condition but there's more in a dry condition.