MILLIMAN WHITE PAPER

# Predictive analytics using SOLYS software

## Analysis of the U.S. National Highway Traffic Safety Administration's Crash Report Sampling System

Prepared by Eckler:

Dane Grand Maison
Zeyue Niu

Milliman

## 1. Executive Summary

The objective of this white paper is to determine the factors that have the strongest relationship with automobile accidents involving an injury and the number of persons injured in an automobile accident. Using data provided by the Crash Report Sampling System (CRSS) of the U.S. National Highway Traffic Safety Administration (NHTSA) we have created a gradient boosting trees model to determine the most significant factors. Based on our analysis, we discovered that the following five factors have the strongest relationship with injuries:

- Primary sample units (geographical area)
- Most harmful event (i.e., collision with objects, collision with motor vehicles)
- Critical event precrash
- Vehicle body type
- Driver age

Based on the findings from our analysis, here are our main recommendations to help reduce the number of accidents involving injury and the number of persons injured in these accidents:

**Specific geographical areas lead to a higher number of injuries.**

- NHTSA should increase its focus on these areas.

**Accidents with injuries often involve collision with pedestrians.**

- NHTSA should try to find ways to reduce the interactions between vehicles and pedestrians, or reduce speed limits in areas where there is heavy pedestrian presence.

**Heavy vehicles (i.e., trucks) and motorcycles are more prone to be involved in an accident with an injury. However, smaller vehicles tend to be involved in accidents with multiple people injured.**

- NHTSA could improve protection against heavy vehicles while increasing protection of motorcyclists.
- NHTSA could try to find ways to reduce the number of people injured by smaller vehicles by increasing the protection measures of the vehicles.

**Older drivers tend to be more involved in accidents with an injury than younger drivers.**

- NHTSA could provide additional support to older drivers (i.e., additional training and information) in order to decrease their probability of accidents.

**Accidents with injuries often occur early in the morning or late at night.**

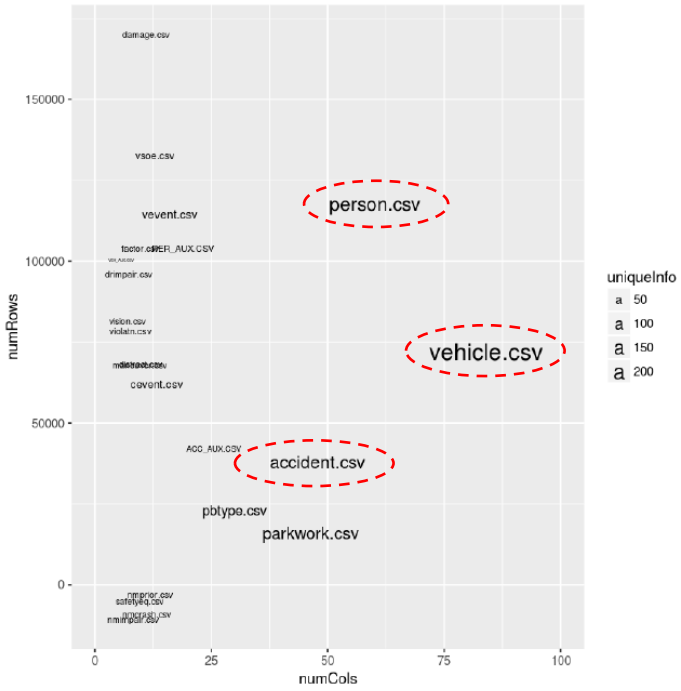
- NHTSA could provide additional guidance or develop awareness programs for the public with respect to the need to be more careful during these periods. NHTSA could also aim to ensure appropriate lighting is available. Additional police officers during these periods could also help reduce the speed of drivers at these times.

## 2. Data and methodology

### 2.1 DATA

Our analysis was done using the data provided by the CRSS, which contains several data tables for accidents at different detail levels (i.e., crash, vehicle, person, event). In order to select tables with the most relevant data, we reviewed the "unique information" defined as the number of unique values for each of the data tables provided. Figure 1 was created using the function ggplot in SOLYS in order to show that three of the tables contain most of the information.

**FIGURE 1: IMPORTANCE OF DATABASES**




Based on this analysis, the data tables vehicle.csv, person.csv, and accident.csv are the most useful files because they contain the largest amount of relevant information. Therefore, we started our analysis using the variables included in these three tables. We also adjusted the data (i.e., eliminated some of the variables) in order to increase the predictability of the model. These adjustments are discussed in section 2.3 below.

### 2.2 METHODOLOGY

We used gradient boosting trees (xgboost function in SOLYS) to fit a predictive model without overfitting. This model was then used to explain and determine the factors that had the strongest relationships with the target variables. Gradient boosting is a machine learning technique that produces a prediction model in the form of chained decision trees.

The boosting trees were used in order to quantify:

1. How the target variable relates to the feature variables that optimize the accuracy of the model.
2. The statistical importance of each variable when calculating predictions.

All the analysis in this white paper was done using the SOLYS Jupyter notebook because it provides the flexibility required for creating the model.

### 2.3 MODELS

Gradient boosting trees aim at minimizing error in every pocket of data. Therefore, caution must be used, and efforts must be made to reduce its tendency to overfit the data, which may lead to inaccurate results.

The model was trained on 60% of the available data and validated on the remaining 40% for each iteration. An early-stopping condition was included to stop the training of the model when accuracy was no longer improving.

For each iteration, only 40% of the training data set was randomly selected and 30% of the feature variables were used to grow the tree.

The maximal depth of trees was selected at five levels.

The objective function being optimized was different from the evaluation metric of the early stopping condition, which monitors when the training should stop.

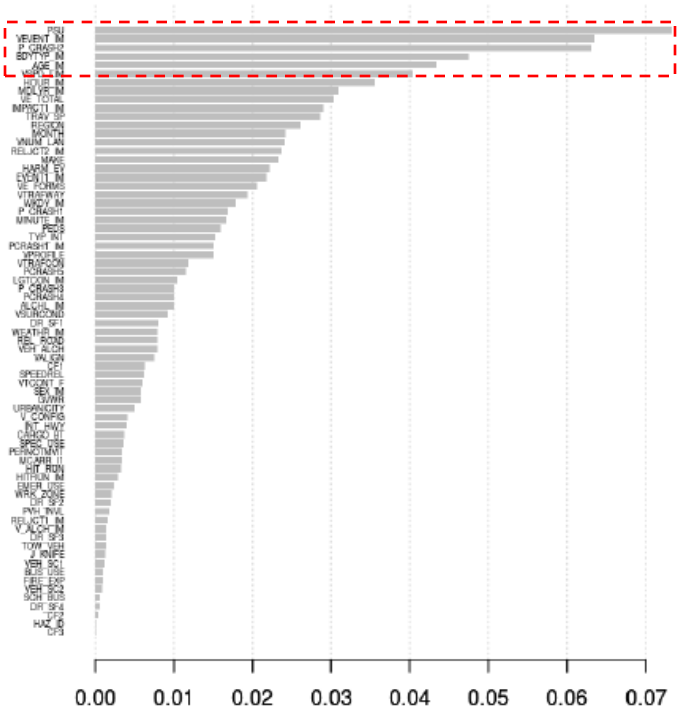
The following types of variables were removed:

- Original variables that had an imputed copy in the raw data.
- Identifications (IDs) and variables that had too many unique values (e.g., CASENUM, VIN, DR_ZIP, etc.) because keeping them may lead to a model's memorization of these specific cases. For instance, if vehicle identification numbers (VINs) were kept in the model, the model may have ended up learning the experience by VIN instead of using other descriptive features.
- Factors that were dependent on or correlated to the target variable (e.g., DEFORMED, TOWED, etc.). These variables tended to be the result of accidents rather than the cause. Their inclusion would have resulted in target information leaking into the model (i.e., concluding that deformed vehicles tend to lead to high degrees of injury, which would not be very insightful).

## Factors predicting accidents that involve an injury (MAXSEV_IM)

The statistical significance of each variable was determined by calculating the average gain increase that combines the univariate effect and interactions with other variables. Using the function xgb.plot.importance in SOLYS, the significance of each variable was calculated and shown in the chart in Figure 2 by order of importance.

**FIGURE 2: SIGNIFICANT VARIABLES**




The five factors with the strongest associations with accidents involving injuries are:

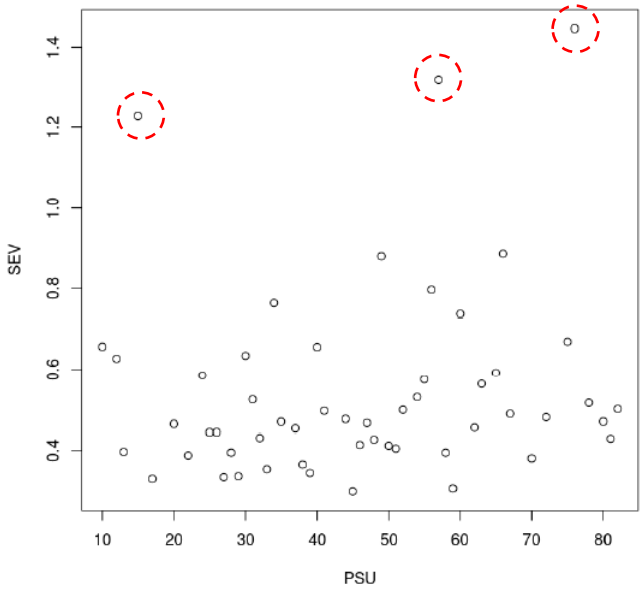
1. Primary sample units
2. Most harmful event
3. Critical event precrash
4. Vehicle body type
5. Driver age

Because there were many variables available, we decided to limit our more detailed analysis to the five most significant variables.

### 3.1 PRIMARY SAMPLE UNITS (PSU)

The general location of the accident was the most significant variable in our model. The chart in Figure 3 shows that, in a few geographical areas, the average severity of accidents is definitely higher than in other areas. It can also be seen that the remaining accidents are distributed randomly across the other areas.

**FIGURE 3: PRIMARY SAMPLE UNITS**

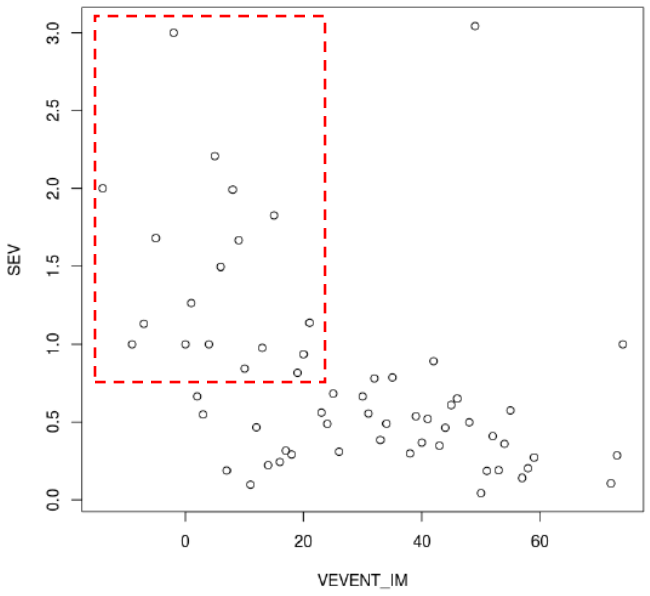



Because only a few PSUs lead to a higher number of injuries, NHTSA should investigate in more detail the reasons why these PSUs have higher severities, in an attempt to make them safer.

### 3.2 MOST HARMFUL EVENT (VEVENT_IM)

The most harmful event helps identify what kind of accident causes injuries (i.e., collision with objects, collision with motor vehicles). The chart in Figure 4 shows that several events lead to a higher severity. The most significant group of events are collision with objects not fixed such as railway (code 10) vehicles, live animals (code 11), or pedestrians (code 8).

**FIGURE 4: MOST HARMFUL EVENT**

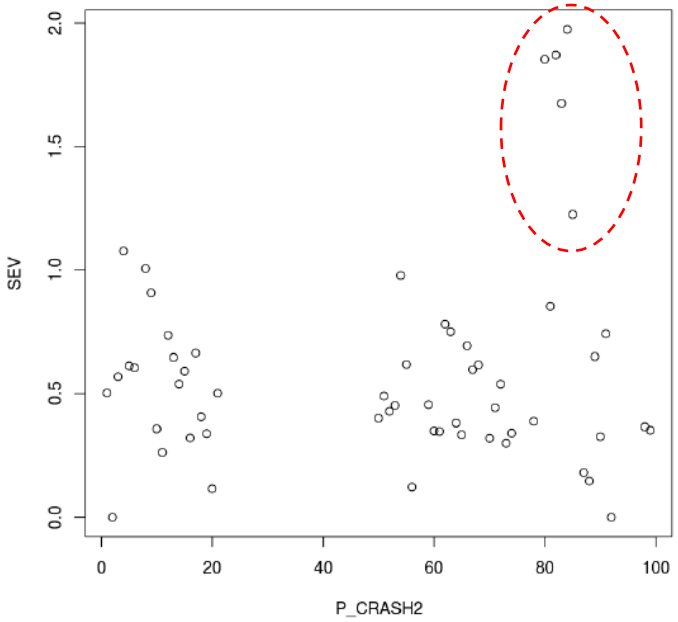


In order to reduce the number of injuries, NHTSA should try to find ways to reduce the interactions between vehicles and pedestrians, or reduce speed limits in areas where there is heavy pedestrian presence.

### 3.3 CRITICAL EVENT PRECRASH (P_CRASH2)

The chart in Figure 5 shows that a collision with a pedestrian tends to lead to a higher chance of injury. This is sensible because a pedestrian would have little or no protection in case of an accident.

**FIGURE 5: CRITICAL EVENT PRECRASH**

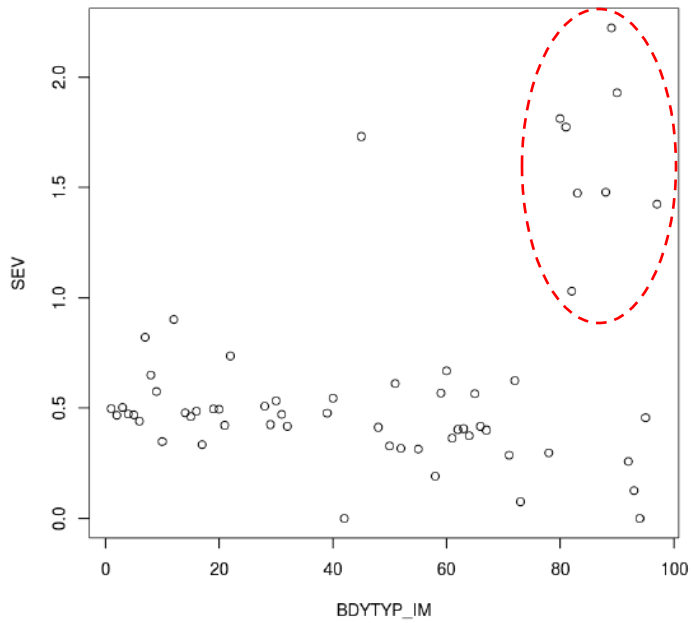



One possible solution would be to improve the design of the roads, intersections, and crosswalks to minimize the chance of vehicles hitting pedestrians. NHTSA could also develop awareness programs for the public to reduce pedestrians' distractions.

### 3.4 VEHICLE BODY TYPE (BDYTYP_IM)

The vehicle body type also has a major impact on accidents involving injuries. The chart in Figure 6 shows that severe injuries occur more frequently when heavy vehicles or motorcycles are involved in an accident. Small vehicles usually lead to lower chances of severe injuries.

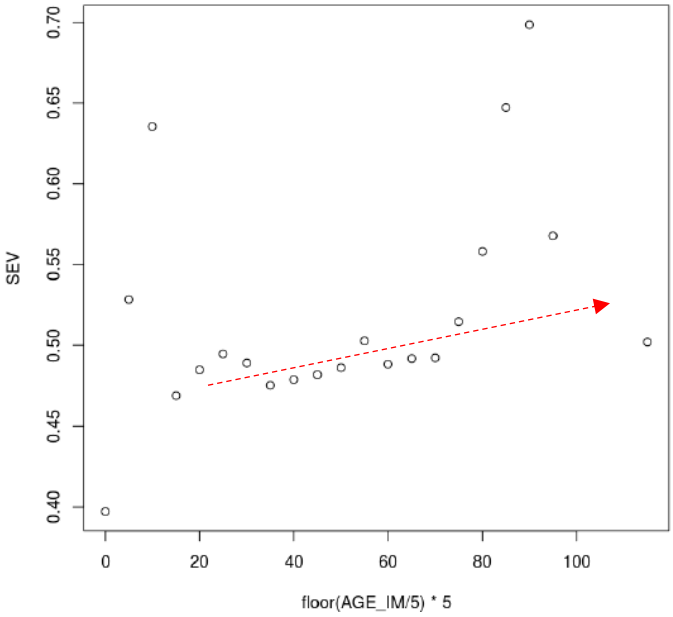
**FIGURE 6: VEHICLE BODY TYPE**




### 3.5 DRIVER AGE (AGE_IM)

The age of the driver is a good predictor of the likelihood of injuries. The chart in Figure 7 shows that the probability of injury increases as the driver gets older, with some nonlinearity around ages 20 to 30. Drivers older than 80 years seem to have a much higher likelihood of being involved in accidents leading to an injury.

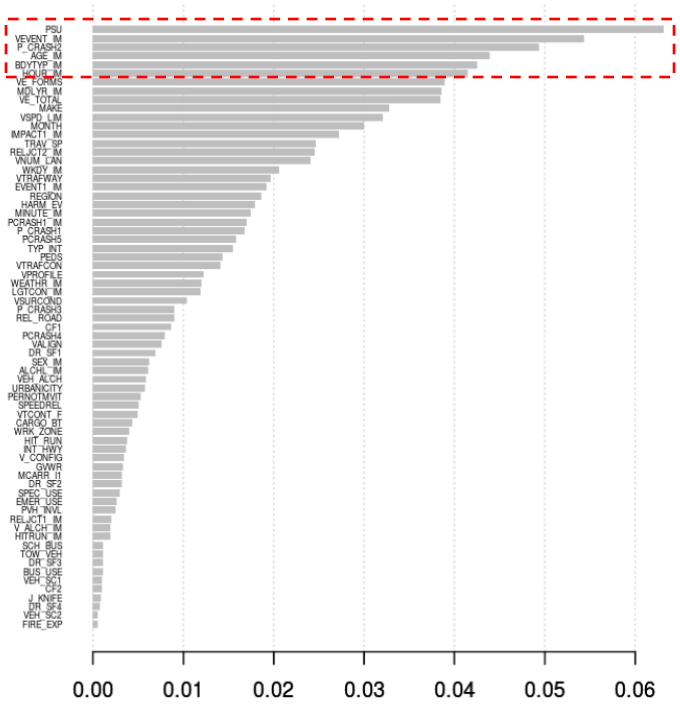
**FIGURE 7: DRIVER AGE**



# Factors predicting the number of persons injured (NO_INJ_IM)

For the number of persons injured, the statistical significance of each variable was determined in the same manner as for the accidents that involve an injury. The five most significant variables are the same as in the previous section but appear in a different order of statistical significance. The chart in Figure 8 confirms that these five variables are important to predict injury and the number of persons injured in these accidents.

**FIGURE 8: SIGNIFICANT VARIABLES**




The five factors that had the strongest associations with the number of persons injured in an accident were:

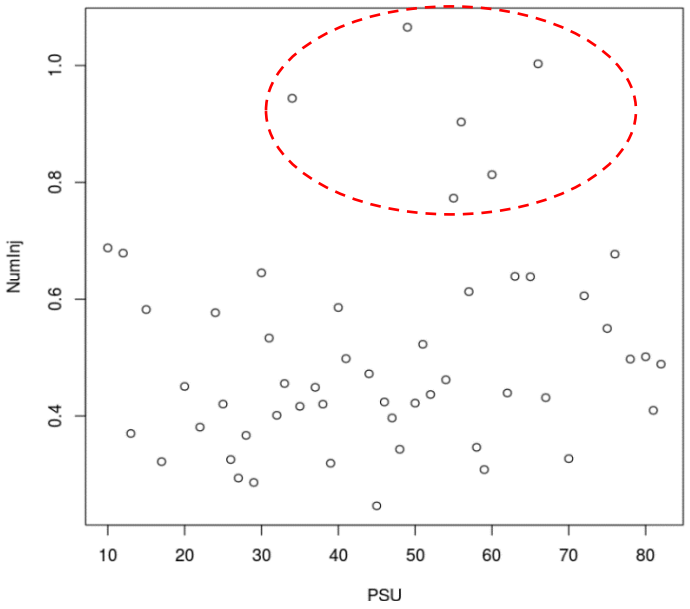
1. Primary sample units
2. Most harmful event
3. Critical event precrash
4. Driver age
5. Vehicle body type

We will discuss these five factors in more detail, with charts using the summarize function in SOLYS.

### 4.1 PRIMARY SAMPLE UNITS (PSU)

The chart in Figure 9 shows that there are some regions (i.e., PSU) with higher likelihoods of having more people injured in an accident. For the remaining territories, the number of injuries in an accident is randomly distributed.

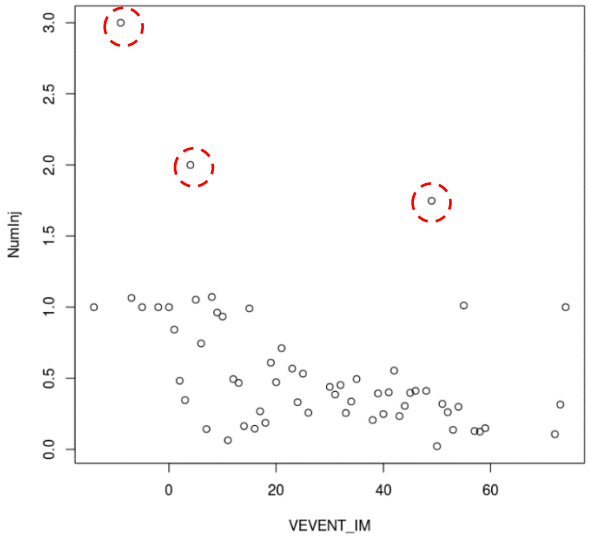
**FIGURE 9: PRIMARY SAMPLE UNITS**



### 4.2 MOST HARMFUL EVENT (VEVENT_IM)

This variable indicates the most harmful event causing an accident. Although it leaks some of the information to the event (e.g., collision with a wall is expected to lead to more injuries than collision with a curb), it also helps identify the type of accident causing each type of injury. There were only a few specific events causing significantly more injuries than other types of events.

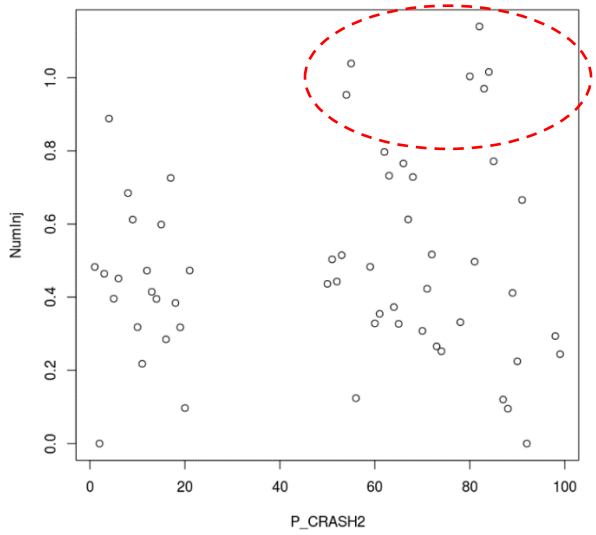
**FIGURE 10: MOST HARMFUL EVENT**




### 4.3 CRITICAL EVENT PRECRASH (P_CRASH2)

The chart in Figure 11 shows that accidents involving a vehicle encroaching into a lane (i.e., code 60 to 80) tended to lead to more people injured. This is due to the fact that accidents between vehicles tend to involve more people than collisions between a vehicle and a pedestrian.

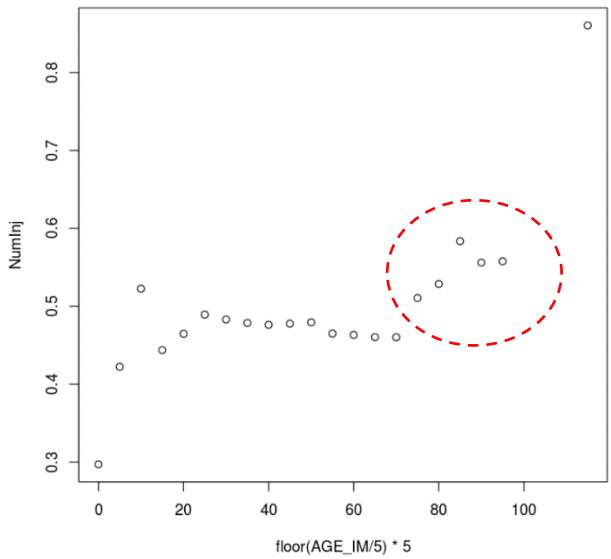
**FIGURE 11: CRITICAL EVENT PRECRASH**




### 4.4 DRIVER AGE (AGE_IM)

The chart in Figure 12 depicts the correlation between drivers older than 70 years old and accidents involving higher numbers of injured persons. Drivers between 30 and 70 seemed to be involved in a similar number of accidents.

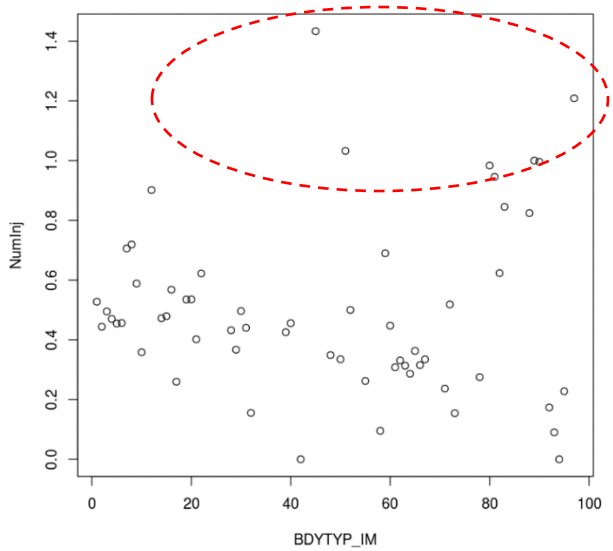
**FIGURE 12: DRIVER AGE**




### 4.5 VEHICLE BODY TYPE (BDYTYP_IM)

The vehicle body type variable shows that smaller vehicles as well as public vehicles tended to have more accidents involving more people injured.

**FIGURE 13: VEHICLE BODY TYPE**



## Conclusion

The U.S. National Highway Traffic Safety Administration's Crash Report Sampling System provides a good source of information in order to predict injuries occurring in an accident. Using the SOLYS software, we analyzed the data and we concluded that:

- Some geographical areas are more prone to accidents with injuries.
- The type of vehicle had an effect on injuries. Heavier vehicles were more often involved in accidents with single severe injuries, whereas smaller vehicles were involved more often in accidents with multiple numbers of injuries.
- Accidents involving more severe injuries often involved pedestrians.
- Older drivers had a higher likelihood of being involved in an accident with more severe injuries.

Accidents involving injuries often occur early in the morning or late at night.