**BUAN 6356 – Business Analytics with R**

**Project Research on Analysis of Car Safety measures vs Fatality**

**Members:** **Group : 12**

* Rajesh Bhattacharjee(rxb210045)
* Maneesh Chandra Mashetty(mxm210088)
* Supriya Padmapriya Suresh (sxp210146)
* Akanksha Rawat (axr210044)
* Gautham Shetty(gxs210012)

**Executive Summary:**

ABC Car Manufacturing company wants to roll out a generic awareness campaign for all drivers to obey speed limits and follow simple safety measures like airbags and seat belts. It is historical fact that in the United States, about 38,000 people die every year in automobile crashes. Many of these accidents are avoidable and are often a result of driver behavior like drunk-driving, speeding, using the cellphone while driving, and driving in precarious weather conditions.

ABC has observed in their recent study of accident records that few drivers believed Seat belts are not effective in saving lives and in few old cars drivers has not purchased cars with air bags or removed them at cases as they felt it is not that of a critical element.

ABC car company has rolled out a project for students of University of Texas at Dallas who will undergo following analysis to find out whether Safety measures like Airbags and Seat belts play any critical role in saving lives in case of any major accident in conjunction of speed limit.

Based on the accurate data points observed during the project, ABC company is planning to drive a nationwide campaign for Safe Driving to protect their brand names. They are observing a negative public sentiment in social media wherever their manufactured cars are involved in recent fatal accidents. Their initial investigation shows it is mostly due to the combined effect of exceeding speed limit and not using safety measures - Air bags and Seat Belts. However, since this is a very sensitive matter, they want to go with a very data based approach except just a gut feeling.

As per the project requirement, University of Texas at Dallas Students will be analyzing the data to find a correlation between Driver conditions after collision and whether Safety measures as well as breached Speed limit. The analysis will be also centered around fatality as that is the most unfortunate turn of event which can happen to human.

The analysis will be taking following Dataset and its’ high level Components are:

* + vehicle impact information- such as whether seat belts and air bags were present and being used at the time of impact. How much was the speed of car at that point of time, which year the car manufactured etc.
  + Driver information- What the condition of driver after impact, gender and whether it is driver or passenger who has been impacted due to the accident.

The project outcome, charts and analysis shall be used to drive the awareness campaign for ABC company.

Additionally, the World Health Organization mentioned that traffic accidents cause economic losses, which result from for example decreased productivity caused by disability or death and the cost of treating injuries. The above has made traffic accidents a complex problem concerning governments, car manufacturers, and people in general. In the pursuit of overcoming this issue, much effort has been invested in making vehicles safer, for instance through the development of seat belts and airbags.

ABC Company wants to make sure above efforts see it’s full potential so they believe the planned advertising campaigns not only enhance their brand image but also they hold this social responsibility towards their customer and citizens.

**PROJECT GOAL:**

The project will be executed in 2 phases.

In Phase I : Project team members will investigate-

* The added effectiveness of combining both safety measures on Fatality rate. Here 2 safety measures are Airbags and Seat Belts.
* In order to arrive the conclusion- team will plot Injury Severity information with Safety measures information and compare bar plots depicting vis-a-vis the impact of collision based on safety measures used.
* Also, it will plot fatality count against safety measures and see the relation between them.
* Part of analysis is to run Logistic Regression Model for Safety Measures as compared to Survival Rate.
* Split the data into Training and Testing set so that we can find accuracy of the model formed.

In Phase II :

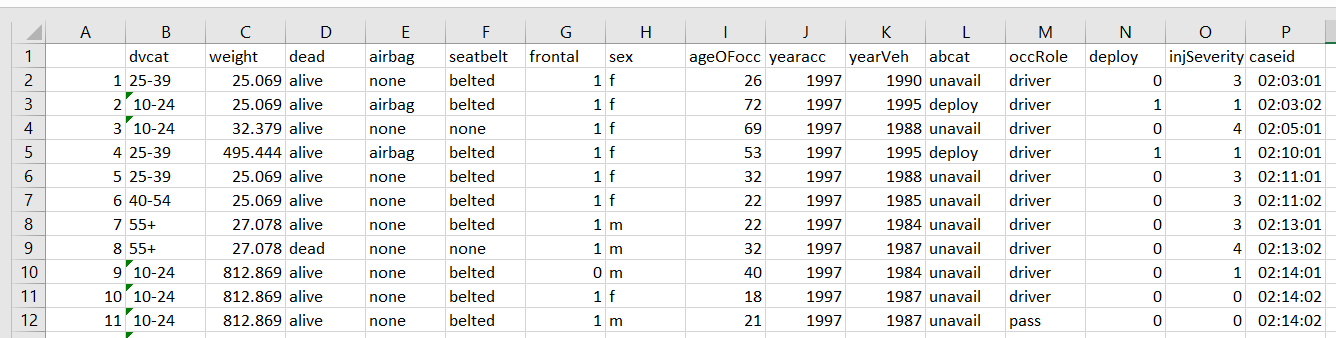
Team further investigate to identify the effect of Vehicle Speed for the scenarios where individuals use Safety measures like Seatbelt and Airbags. In other words, we explore if vehicle speed affects the Fatality rate given the drivers use either or both the Safety Measures (Seatbelt/Airbags). Also predict the threshold speed beyond which accidents tend to often more probably.

**DATASET:**

Dataset used is a second-hand data obtained from online data source Kaggle. The dataset contains 26217 observations with the following 15 variables.

* Dvcat : Estimated impact speeds with categories 1-9km/h, 10-24, 25-39, 40-54, 55+
* Weight: A numerical attribute which represents weights of observations, designed to account for varying sampling probabilities.
* Dead : A categorical attribute to identify if a person is alive or dead
* Airbag: A qualitative attribute to identify airbag is available or not with categories “none/airbag”
* Seatbelt: A qualitative attribute to identity if seatbelt is fastened or not with categories “none/belted”
* Frontal: A binary value where 0 represents non-frontal and 1 represents frontal impact
* Sex: Gender with categories F/M
* AgeOFocc: A numerical attribute representing age of victim
* Yearacc: A numerical vector which represents year of accident
* YearVeh: A numeric vector which represents year of model of vehicle
* Abcat: A qualitative attribute which identifies if one or more (driver or passenger) airbag(s) are deployed with categories “deploy/nodeploy/unavail”
* occRole: A qualitative attribute which represents if victim is driver or passenger with categoriess “driver/pass”
* deploy: A binary attributes where 0 represents an airbag was unavailable or did not deploy; and 1 represents if one or more bags deployed.
* injSeverity: a numeric vector where 0 represents none, 1 represents possible injury, 2 represents no incapacity, 3 represents incapacity, 4 represents killed; 5 represents unknown, 6 represents prior death
* caseid: A string used to uniquely identify the vehicle. It is concatenation of populations sampling unit, the case number, and the vehicle number.

**Sample Screenshot of Data:**



Dataset includes only the people in front seats of the vehicle at the time of accident. The dataset is clean and structured correctly from the onset.

**Data Pre-processing:**

Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format. There are no missing or noisy data in the data set, hence the data is clean. One of the step in data pre-processing is **Attribute Selection**. In order to determine the effect of safety measures (Airbag and Seat belt) on survival rate, an additional attribute “safety\_measure” is added to the data set. The values of “safety\_measure” is determined by two other attributes from the dataset namely, “seatbelt” and “airbag”.

Following are the categories into which “safety\_measure” attribute is divided.

* Accidents with only seat belts protecting the occupants – “SB”
* Accidents with only airbags protecting the occupants – “AB”
* Accidents with both airbags and seat belts protecting the occupants – “ABSB”
* Accidents where neither seatbelts nor airbags were used – “none”

accident\_data$safety\_measure <- 0

accident\_data$safety\_measure[accident\_data$airbag == "none" &

accident\_data$seatbelt == "none"]<-"none"

accident\_data$safety\_measure[accident\_data$airbag == "airbag" &

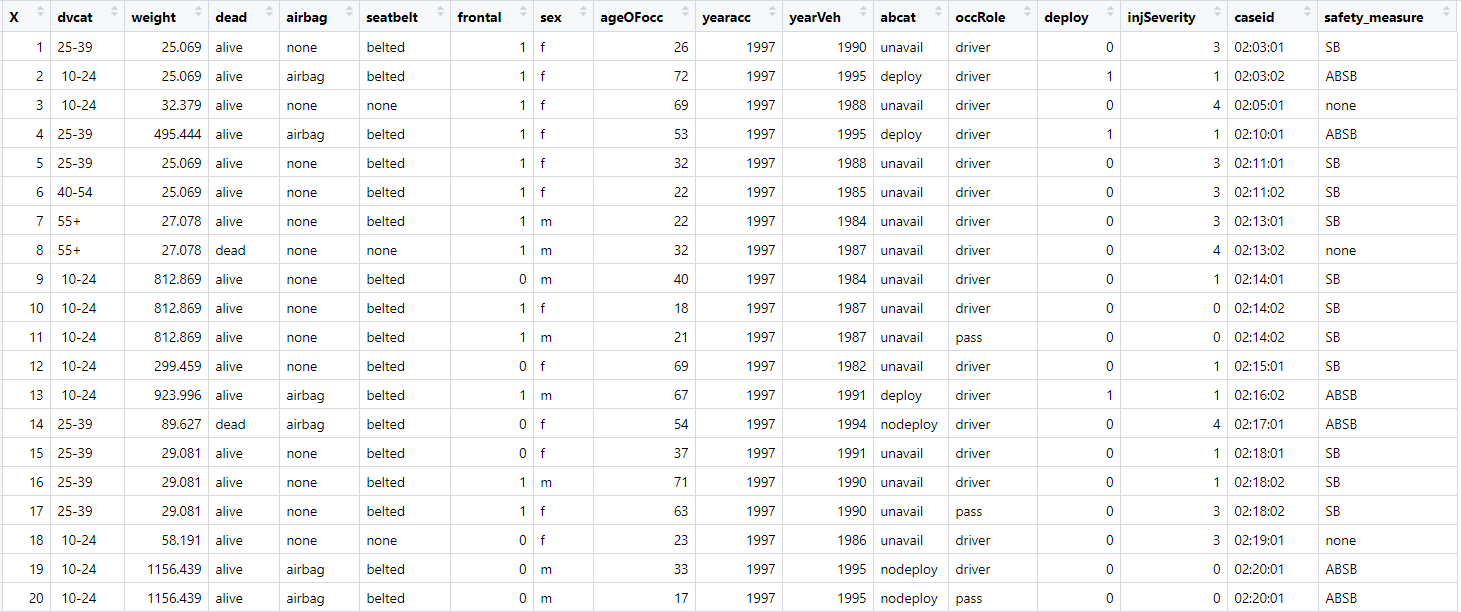
accident\_data$seatbelt == "none"]<-"AB"

accident\_data$safety\_measure[accident\_data$airbag == "none" &

accident\_data$seatbelt == "belted"] <-"SB"

accident\_data$safety\_measure[accident\_data$airbag == "airbag" &

accident\_data$seatbelt == "belted"] <-"ABSB"



For the first objective, the observations are filtered to include only accidents where speed of 55 kmph or more has been used. For the second objective, the data is split in such a way that only observations where both the safety measures “airbag” and “seat belt” were used are taken into consideration.

**Data Mining Technique:**

The data mining technique used on the dataset is “Logistic Regression”. Logistic Regression is a supervised learning algorithm used to predict a dependent categorical target variable.

* **Data Split:** The data set is split into training and testing set in 70:30 ratio. 70% of data is used to train data set.

train\_index = createDataPartition(y = highSpeed\_accidents$dead, p = 0.70, list = FALSE)

train\_data = highSpeed\_accidents[trn\_index, ]

test\_data = highSpeed\_accidents[-trn\_index, ]

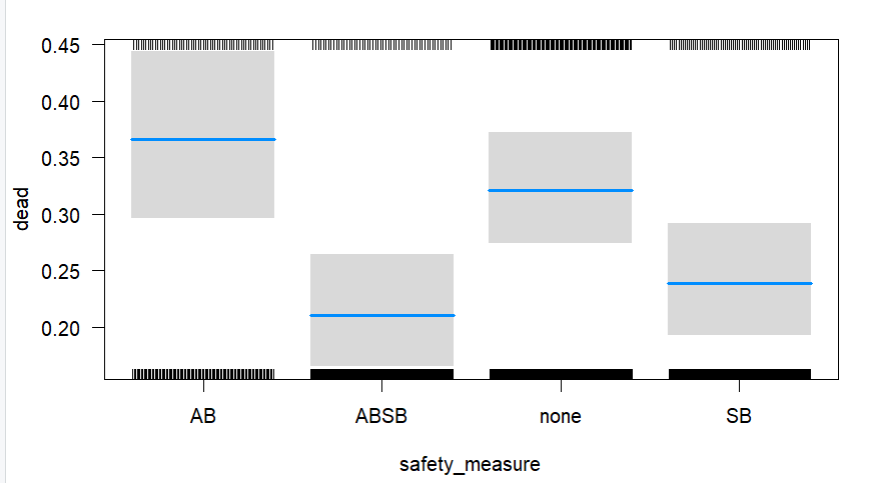
* **Training the model:** There are three main types of logistic regression: binary, multi nominal and ordinal. Since the outcome of the data set “survival\_rate” has two possible values “alive” and “dead”, we use binary logistic regression.

**Objective 1**

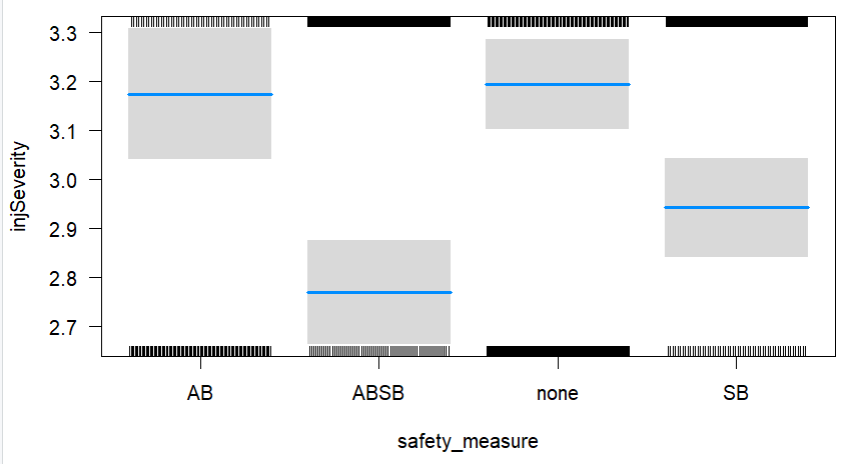
logit\_model\_safety\_measures <- glm(dead ~ safety\_measure, family = binomial, data = train\_data)

**Result:**

When comparing the survival rate accompanying different safety measures, the data shows that survival rate is lowest for “airbag only” group, followed by group with no safety measures, group with only “seat belt” and finally the group with both the safety measures “Airbag” and “Seat Belt”.



To further illustrate the effects of airbags and seat belts, graph below shows the frequency of different levels of injury occurring per group of safety measures.



**Accuracy:**

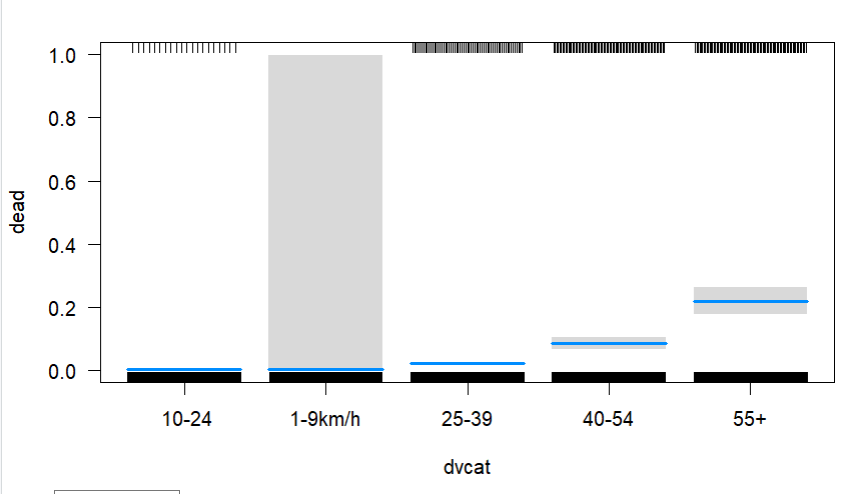
// needs to be done

**Objective 2:**

logit\_model\_speeds<- glm(dead ~ dvcat,family = binomial, data = train\_data)

**Result:**

When comparing the death rate accompanying different speed limit, the data shows that death rate increases with increase in speed.



**Accuracy**

**82%**