# A MINI PROJECT REPORT
# On
# Multiple Regression and Model Building

**Submitted by**


**Name: Anujay Jain**
**Roll No: 161500106**

**Name: Nihit Jain**
**Roll No: 161500350**

**Name: Utkarsh Rai**
**Roll No: 161500599**


**To**
**Mr. Rahul Pradhan**


Department of Computer Engineering & Applications

**Institute of Engineering & Technology**

**GLA University**
**Mathura- 281406, INDIA**


**December, 2018**

**Department of Computer Engineering and Applications**
**GLA University, Mathura**
**17 km. Stone NH#2, Mathura-Delhi Road, P.O. – Chaumuha,**
**Mathura – 281406**

# ***Declaration***

*We hereby declare that the work which is being presented in the Mini Project "Multiple Regression and Model Building**",** in partial fulfillment of the requirements for Mini-Project LAB, is an authentic record of our own work carried under the supervision of **Mr. Rahul Pradhan, Assistant Professor, GLA University, Mathura**.*

**Anujay Jain**

**Sign:_________**

**Nihit Jain**

**Sign:_________**

**Utkarsh Rai**

**Sign:_________**

Department of Computer Engineering and Applications
GLA University, Mathura
17 km. Stone NH#2, Mathura-Delhi Road, P.O. – Chaumuha,
Mathura – 281406

# CERTIFICATE

*This is to certify that the project entitled **"Multiple Regression and Model Building"** carried out in Mini Project – I Lab is a bonafide work done by **Anujay Jain (161500106), Nihit Jain (161500350) and Utkarsh Rai (161500599)**and is submitted in partial fulfillment of the requirements for the award of the degree Bachelor of Technology (Computer Science & Engineering).*

**Signature of Supervisor:**

**Name of Supervisor:**

**Date:**

# ACKNOWLEDGEMENT

*It gives us a great sense of pleasure to present the report of the B. Tech Mini Project undertaken during B. Tech. Third Year. This project in itself is an acknowledgement to the inspiration, drive and technical assistance contributed to it by many individuals. This project would never have seen the light of the day without the help and guidance that we have received.*

*Our heartiest thanks to* ***Dr. (Prof). Anand Singh Jalal,*** *Head of Dept., Department of CEA for providing us with an encouraging platform to develop this project, which thus helped us in shaping our abilities towards a constructive goal.*

*We owe special debt of gratitude to* ***Mr. Rahul Pradhan,*** *Assistant Professor Department of CEA, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. He has showered us with all his extensively experienced ideas and insightful comments at virtually all stages of the project & has also taught us about the latest industry-oriented technologies.*

*We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind guidance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.*

Anujay Jain
Nihit Jain
Utkarsh Rai

# Abstract

In this project, we wanted to analyse on Car Crash dataset and predict Head Injury using Multiple Regression. In our project, we first build a regression model considering all other variables as predictors. Thus, it results in errors due to multicollinearity, between the variables. Thus, in our next model, we eradicate multicollinearity using Backward Elimination and Stepwise Elimination. We calculate and find that our model is better from the previous one and the related values say it all. We, find from the dataset, the indicator variables and also the variables that are superfluous. The project is quite accurate in its predictions and results, and can be used for future references. In a world, where AI is in trends, it will rule the world in future, cars will be automated and driverless. In such a scenario, these results will be influential and beneficial for making such cars.

# Table of Contents

# CHAPTER 1

## 1. Business Understanding

A crash test is a form of destructive testing usually performed in order to ensure safe design standards in crashworthiness and crash compatibility for various modes of transportation (automobiles) or related systems and components.

## 1.1 Motivation

The main motivation for us to go for this project was that a lot of accidents happen due to various reasons like not following the rules, increase in traffic, failure of a certain model of a car etc., it is not possible to completely stop these accidents but if we can find out what causes the most dangerous type of injury i.e. head injury then we can minimize the overall injury caused in an accident and that is our goal.

## 1.2 Scope

The scope of our analysis is that it can be used to reduce the severity of the injury caused in a car accident. This is so, as based on our analysis we can find out those attributes which are most significant for the injury thus helping us to focus on only those attributes rather than wasting our effort and time on other unrelated areas. Thus it will help as follows:

1. Redefine the criteria for passing of the car before production
2. Indirectly reduce medical expenditure of the country.
3. Reduce fatality rate

## 1.3 Drawbacks in existing system

- These days more importance is given to shape, color and speed of the car than their safety features.

- The test a car need to pass are now outdated and need to be revised based on current advancement in designs of cars.

# CHAPTER 2

## 2. Description

- Data collected from online collaborative repository of Car Assessment Program
- A sample data is recorded in a file - Crash.dat

## 2.2 Project Plan

### 2.2.1 Objective

- Build the best multiple regression model that can predict head injury severity, using all the other variables as the predictors.
- Determine which variables must be made into indicator variables.
- Determine which variables might be superfluous.
- Build two parallel models, one where you account for multicollinearity, and another where you don't consider multicollinearity. For which purpose may each of these models be used?
- Continuing with the Crash data set, combine the four injury measurement variables into a single variable, defending your choice of combination function.
    - Build the best multiple regression model for the purpose of predicting injury severity, using all the variables as the predictors.
    - Build two parallel models, one where you account for multicollinearity, and another where you don't consider multicollinearity. For which purpose may each of these models be used?

## 2.2.2 Goals

- To help the lecturers, improve and organize the process of track and manage student attendance.
- Provides a valuable attendance service for both teachers and students.
- Reduce manual process errors by provide automated and a reliable attendance system.
- Increase privacy and security which student cannot present him or his friend while they are not.
- Produce monthly reports for lecturers.
- Flexibility, Lectures capability of editing attendance records.

# CHAPTER 3

## 3. Project Implementation

It includes the steps taken to implement the project.

## 3.1 Understanding the Dataset

| MAKE | MODEL | CARID | CARID_YR | HEAD_INJ | CHEST_IN | LLEG_INJ | RLEG_INJ | DRIV_PAS | PROTECT | DOORS | YEAR | WEIGHT | SIZE | SIZE2 | PROTECT2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Acura | Integra | Acura Integra | Acura Integra 87 | 599 | 35 | 791 | 262 | Driver | manual belts | 2 | 87 | 2350 | lt | 2 | 1 |
| Acura | Integra RS | Acura Integra RS | Acura Integra RS 90 | 585 | | 1545 | 1301 | Driver | Motorized belts | 4 | 90 | 2490 | lt | 2 | 2 |
| Acura | Legend LS | Acura Legend LS | Acura Legend LS 88 | 435 | 50 | 926 | 708 | Driver | d airbag | 4 | 88 | 3280 | med | 3 | 4 |
| Audi | 80 | Audi 80 | Audi 80 89 | 600 | 49 | 168 | 1871 | Driver | manual belts | 4 | 89 | 2790 | comp | 1 | 1 |
| Audi | 100 | Audi 100 | Audi 100 89 | 185 | 35 | 998 | 894 | Driver | d airbag | 4 | 89 | 3100 | med | 3 | 4 |
| BMW | 325i | BMW 325i | BMW 325i 90 | 1036 | 56 | 865 | | Driver | d airbag | 2 | 90 | 2862 | comp | 1 | 4 |
| Buick | Century | Buick Century | Buick Century 91 | 815 | 47 | 1340 | 315 | Driver | passive belts | 4 | 91 | 2992 | comp | 1 | 3 |
| Buick | Elect. Park Ave | Buick Elect. Park Ave | Buick Elect. Park Ave 88 | 1467 | 54 | 712 | 1366 | Driver | manual belts | 4 | 88 | 3360 | med | 3 | 1 |
| Buick | Le Sabre | Buick Le Sabre | Buick Le Sabre 90 | | 35 | 1049 | 908 | Driver | passive belts | 2 | 90 | 3240 | med | 3 | 3 |
| Buick | Regal | Buick Regal | Buick Regal 88 | 880 | 50 | 996 | 642 | Driver | passive belts | 2 | 88 | 3210 | med | 3 | 3 |
| Cadillac | De Ville | Cadillac De Ville | Cadillac De Ville 90 | 423 | 39 | 541 | 1629 | Driver | d airbag | 4 | 90 | 3500 | hev | 4 | 4 |
| Chevrolet | Astro | Chevrolet Astro | Chevrolet Astro 88 | 1603 | 72 | 1572 | 700 | Driver | manual belts | - | 88 | 3787 | van | 6 | 1 |
| Chevrolet | Astro | Chevrolet Astro | Chevrolet Astro 89 | 1849 | 64 | 2737 | 1043 | Driver | manual belts | - | 89 | 4002 | van | 6 | 1 |
| Chevrolet | Beretta | Chevrolet Beretta | Chevrolet Beretta 91 | 343 | 37 | 659 | 523 | Driver | d airbag | 2 | 91 | 2671 | comp | 1 | 4 |
| Chevrolet | Beretta GT | Chevrolet Beretta GT | Chevrolet Beretta GT 88 | 864 | 50 | 1692 | 1052 | Driver | passive belts | 2 | 88 | 2890 | comp | 1 | 3 |
| Chevrolet | Camaro | Chevrolet Camaro | Chevrolet Camaro 87 | 733 | 39 | 736 | 353 | Driver | manual belts | 2 | 87 | 3070 | med | 3 | 1 |
| Chevrolet | Camaro | Chevrolet Camaro | Chevrolet Camaro 91 | 585 | 39 | 717 | 150 | Driver | d airbag | 2 | 91 | 3191 | med | 3 | 4 |
| Chevrolet | Caprice | Chevrolet Caprice | Chevrolet Caprice 89 | 1328 | 64 | 406 | 493 | Driver | manual belts | 4 | 89 | 3693 | hev | 4 | 1 |
| Chevrolet | Caprice | Chevrolet Caprice | Chevrolet Caprice 91 | 533 | 54 | 1529 | 613 | Driver | d airbag | 4 | 91 | 3990 | hev | 4 | 4 |
| Chevrolet | Cavalier | Chevrolet Cavalier | Chevrolet Cavalier 90 | 770 | 49 | 775 | 531 | Driver | passive belts | 4 | 90 | 2540 | comp | 1 | 3 |

The above image is a snippet of the dataset on which this project is done. It has 16 columns and 352 rows, i.e. there are 352 records each having 16 variables.

## 3.2 Describe data

The description of the column is as follows:

- MAKE - It tells the company of the car.
- MODEL - It tells the model of the car.
- CARID - It is the unique name by which we can know the company and the model of the company. It uniquely identifies a certain type of car from another.

- CARID_YR - It also contains the CARID, but along with it also tells the year in which it was released.
- HEAD_INJ - Number of head injuries in an accident in a certain type of car.
- CHEST_IN - Number of chest injuries in an accident in a certain type of car.
- LLEG_INJ - Number of injuries in left leg in an accident in a certain type of car.
- RLEG_INJ - Number of injuries in right leg in an accident in a certain type of car.
- DRIV_PAS - Whether the passenger got hurt or the driver.
- PROTECT - What type of protection was available in the vehicle.
- DOORS - Number of doors in the vehicle.
- YEAR - In what year the car was launched.
- WEIGHT - Tells the weight of the car.
- SIZE - Categorize the cars by their size.
- SIZE2 - Size in numeric value for the category of size to which the car belongs.
- PROTECT2 - Numeric value for the protection in the car.

## 3.3 Data cleaning

The data which is obtained may need to be processed before it can be actually used, like there may be some values missing which need to be filled otherwise they will cause problem when doing the analyses on the data.

| HEAD_INJ | CHEST_IN | LLEG_INJ | RLEG_INJ | DRIV_PAS | PROTECT | DOORS | YEAR | WEIGHT | SIZE | SIZE2 | PROTE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 599 | 35 | 791 | 262 | Driver | manual be | 2 | 87 | 2350 | lt | 2 | |
| 585 | | 1545 | 1301 | Driver | Motorizec | 4 | 90 | 2490 | lt | 2 | |
| 435 | 50 | 926 | 708 | Driver | d airbag | 4 | 88 | 3280 | med | 3 | |
| 600 | 49 | 168 | 1871 | Driver | manual be | 4 | 89 | 2790 | comp | 1 | |
| 185 | 35 | 998 | 894 | Driver | d airbag | 4 | 89 | 3100 | med | 3 | |
| 1036 | 56 | 865 | | Driver | d airbag | 2 | 90 | 2862 | comp | 1 | |
| 815 | 47 | 1340 | 315 | Driver | passive be | 4 | 91 | 2992 | comp | 1 | |
| 1467 | 54 | 712 | 1366 | Driver | manual be | 4 | 88 | 3360 | med | 3 | |
| | 35 | 1049 | 908 | Driver | passive be | 2 | 90 | 3240 | med | 3 | |
| 880 | 50 | 996 | 642 | Driver | passive be | 2 | 88 | 3210 | med | 3 | |
| 423 | 39 | 541 | 1629 | Driver | d airbag | 4 | 90 | 3500 | hev | 4 | |
| 1603 | 72 | 1572 | 700 | Driver | manual be - | | 88 | 3787 | van | 6 | |
| 1849 | 64 | 2737 | 1043 | Driver | manual be - | | 89 | 4002 | van | 6 | |

In the above image we can see that there are some missing values so we need to fill them before we do anything else.

## 3.4 Data Insights

Here are certain insights of the data which may help in analyses:

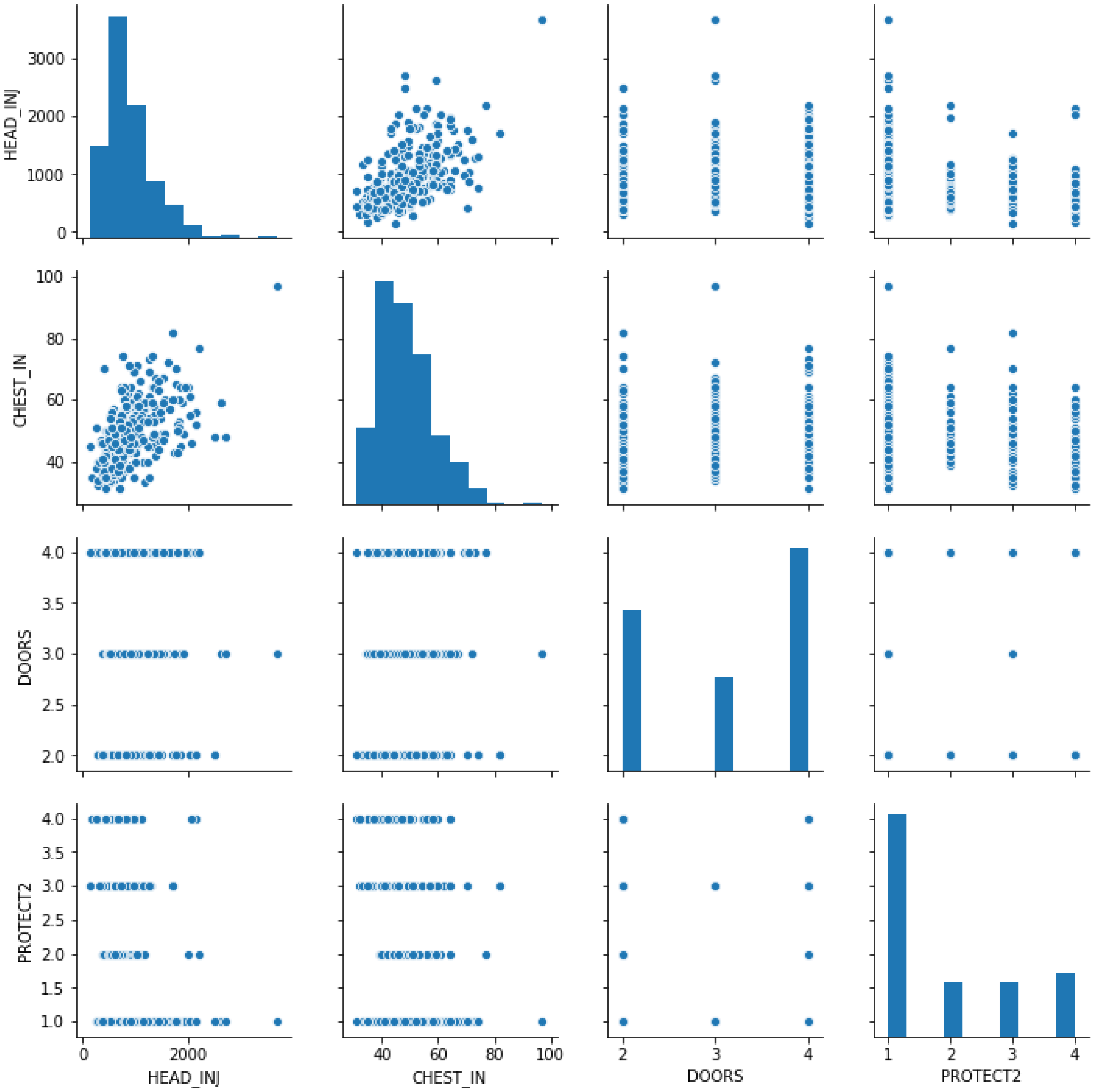
1. First five rows of the dataset

| | MAKE | MODEL | CARID | CARID_YR | HEAD_INJ | CHEST_IN | LLEG_INJ | RLEG_INJ | DRIV_PAS | PROTECT | DOORS | YEAR | WEIGHT | SIZE | SIZE2 | PR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Acura | Integra | Acura Integra | Acura Integra 87 | 599 | 35 | 791 | 262 | Driver | manual belts | 2 | 87 | 2350 | lt | 2 | 1 |
| 1 | Acura | Integra RS | Acura Integra RS | Acura Integra RS 90 | 585 | | 1545 | 1301 | Driver | Motorized belts | 4 | 90 | 2490 | lt | 2 | 2 |
| 2 | Acura | Legend LS | Acura Legend LS | Acura Legend LS 88 | 435 | 50 | 926 | 708 | Driver | d airbag | 4 | 88 | 3280 | med | 3 | 4 |
| 3 | Audi | 80 | Audi 80 | Audi 80 89 | 600 | 49 | 168 | 1871 | Driver | manual belts | 4 | 89 | 2790 | comp | 1 | 1 |
| 4 | Audi | 100 | Audi 100 | Audi 100 89 | 185 | 35 | 998 | 894 | Driver | d airbag | 4 | 89 | 3100 | med | 3 | 4 |

2. Summary of the dataset after preprocessing

| | HEAD_INJ | CHEST_IN | LLEG_INJ | RLEG_INJ | DOORS | YEAR | WEIGHT | SIZE2 | PROTECT2 |
|---|---|---|---|---|---|---|---|---|---|
| count | 338.000000 | 338.000000 | 338.000000 | 338.000000 | 338.000000 | 338.000000 | 338.000000 | 338.000000 | 338.000000 |
| mean | 900.568047 | 48.523669 | 1058.073964 | 740.180473 | 3.130178 | 88.881657 | 2902.917160 | 3.553254 | 1.896450 |
| std | 465.049823 | 9.556689 | 542.827741 | 424.225893 | 0.888890 | 1.398671 | 592.878968 | 2.416279 | 1.159883 |
| min | 157.000000 | 31.000000 | 120.000000 | 89.000000 | 2.000000 | 87.000000 | 1590.000000 | 1.000000 | 1.000000 |
| 25% | 583.000000 | 42.000000 | 691.750000 | 450.000000 | 2.000000 | 88.000000 | 2465.000000 | 2.000000 | 1.000000 |
| 50% | 790.500000 | 47.000000 | 1012.500000 | 656.500000 | 3.000000 | 89.000000 | 2845.000000 | 3.000000 | 1.000000 |
| 75% | 1069.500000 | 54.000000 | 1365.500000 | 943.500000 | 4.000000 | 90.000000 | 3284.000000 | 6.000000 | 3.000000 |
| max | 3665.000000 | 97.000000 | 3347.000000 | 2856.000000 | 4.000000 | 91.000000 | 5103.000000 | 8.000000 | 4.000000 |

3. Pairplot of significant variables



4. OLS Regression Results after considering all variables as predictors

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | HEAD_INJ | R-squared: | 0.390 |
| Model: | OLS | Adj. R-squared: | 0.375 |
| Method: | Least Squares | F-statistic: | 26.28 |
| Date: | Fri, 14 Dec 2018 | Prob (F-statistic): | 2.35e-31 |
| Time: | 02:06:40 | Log-Likelihood: | -2471.6 |
| No. Observations: | 338 | AIC: | 4961. |
| Df Residuals: | 329 | BIC: | 4996. |
| Df Model: | 8 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| intercept | 1889.2188 | 1372.098 | 1.377 | 0.169 | -809.973 | 4588.411 |
| CHEST_IN | 25.2415 | 2.304 | 10.955 | 0.000 | 20.709 | 29.774 |
| LLEG_INJ | -0.0203 | 0.039 | -0.518 | 0.605 | -0.097 | 0.057 |
| RLEG_INJ | -0.0194 | 0.050 | -0.390 | 0.697 | -0.117 | 0.078 |
| DOORS | -16.5654 | 24.189 | -0.685 | 0.494 | -64.149 | 31.019 |
| YEAR | -24.9043 | 15.856 | -1.571 | 0.117 | -56.095 | 6.287 |
| WEIGHT | 0.0682 | 0.042 | 1.630 | 0.104 | -0.014 | 0.150 |
| SIZE2 | 8.0180 | 11.164 | 0.718 | 0.473 | -13.943 | 29.979 |
| PROTECT2 | -73.1042 | 21.115 | -3.462 | 0.001 | -114.641 | -31.567 |

| | | | |
|---|---|---|---|
| Omnibus: | 88.562 | Durbin-Watson: | 1.770 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 229.007 |
| Skew: | 1.240 | Prob(JB): | 1.87e-50 |
| Kurtosis: | 6.180 | Cond. No. | 2.22e+05 |

5. Which company make had highest head injury count

| | MAKE | HEAD_INJ | CHEST_IN | LLEG_INJ | RLEG_INJ | DOORS | YEAR | WEIGHT | SIZE2 | PROTECT2 | intercept |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Chevrolet | 41903 | 1980 | 51539 | 23507 | 116 | 3386 | 129905 | 171 | 72 | 38 |

6. Which company make had lowest head injury count

| | MAKE | HEAD_INJ | CHEST_IN | LLEG_INJ | RLEG_INJ | DOORS | YEAR | WEIGHT | SIZE2 | PROTECT2 | intercept |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | Cadillac | 423 | 39 | 541 | 1629 | 4 | 90 | 3500 | 4 | 4 | 1 |

7. Which year car make had the highest head injury

| | YEAR | HEAD_INJ | CHEST_IN | LLEG_INJ | RLEG_INJ | DOORS | WEIGHT | SIZE2 | PROTECT2 | intercept |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 87 | 70044 | 3479 | 65088 | 46651 | 212 | 193924 | 245 | 115 | 73 |

8. Which year car make had the lowest head injury

| | YEAR | HEAD_INJ | CHEST_IN | LLEG_INJ | RLEG_INJ | DOORS | WEIGHT | SIZE2 | PROTECT2 | intercept |
|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 91 | 46416 | 2907 | 66819 | 46660 | 196 | 171189 | 181 | 155 | 59 |

9. Which types of protection have highest influence on head injury

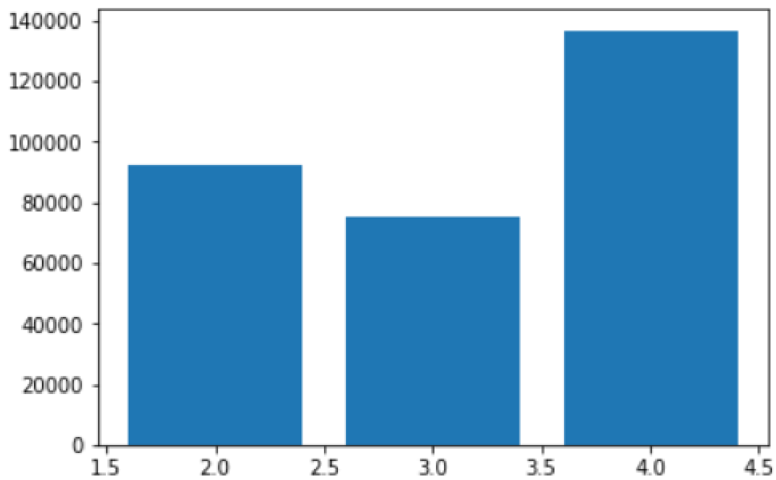
| | PROTECT2 | HEAD_INJ | CHEST_IN | LLEG_INJ | RLEG_INJ | DOORS | YEAR | WEIGHT | SIZE2 | intercept |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 197512 | 9495 | 181948 | 120332 | 582 | 16884 | 559452 | 885 | 191 |

10. Which types of protection have least influence on head injury

| | PROTECT2 | HEAD_INJ | CHEST_IN | LLEG_INJ | RLEG_INJ | DOORS | YEAR | WEIGHT | SIZE2 | intercept |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 34358 | 2191 | 59704 | 49825 | 156 | 4141 | 128480 | 89 | 46 |

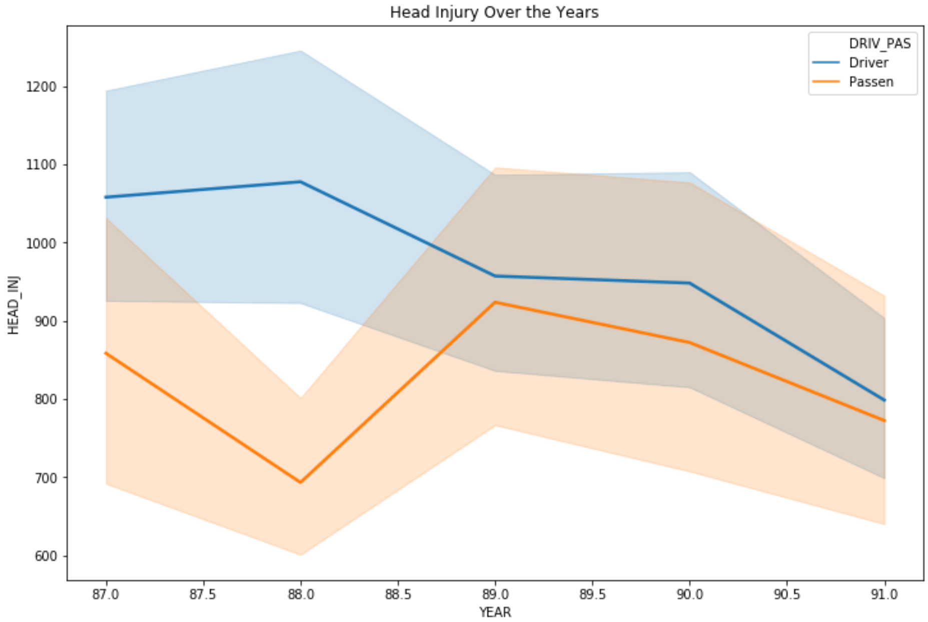
11. Significance of doors on head injury.

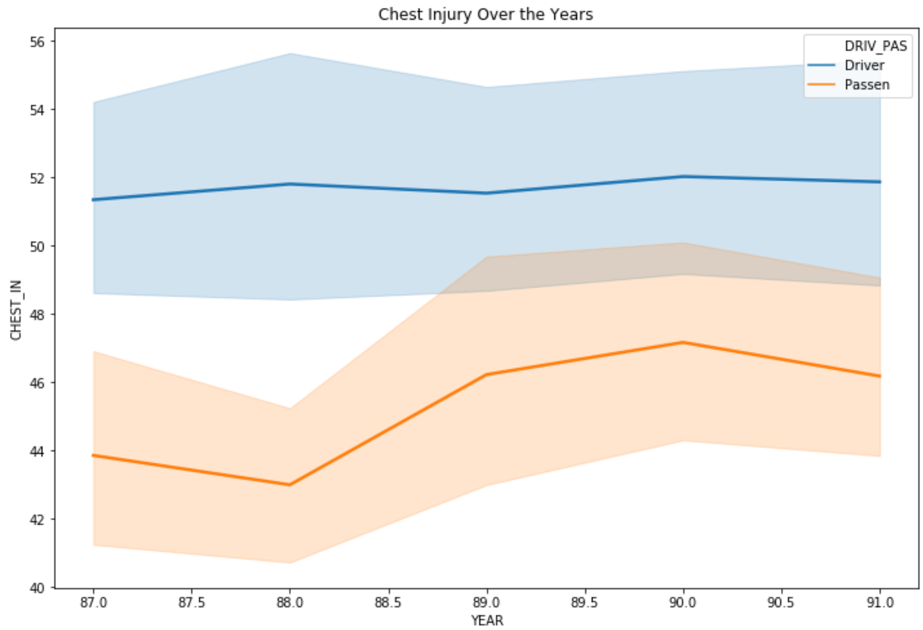
| | DOORS | HEAD_INJ | CHEST_IN | LLEG_INJ | RLEG_INJ | YEAR | WEIGHT | SIZE2 | PROTECT2 | intercept |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 4 | 136752 | 7759 | 169763 | 123737 | 14087 | 462894 | 426 | 341 | 158 |

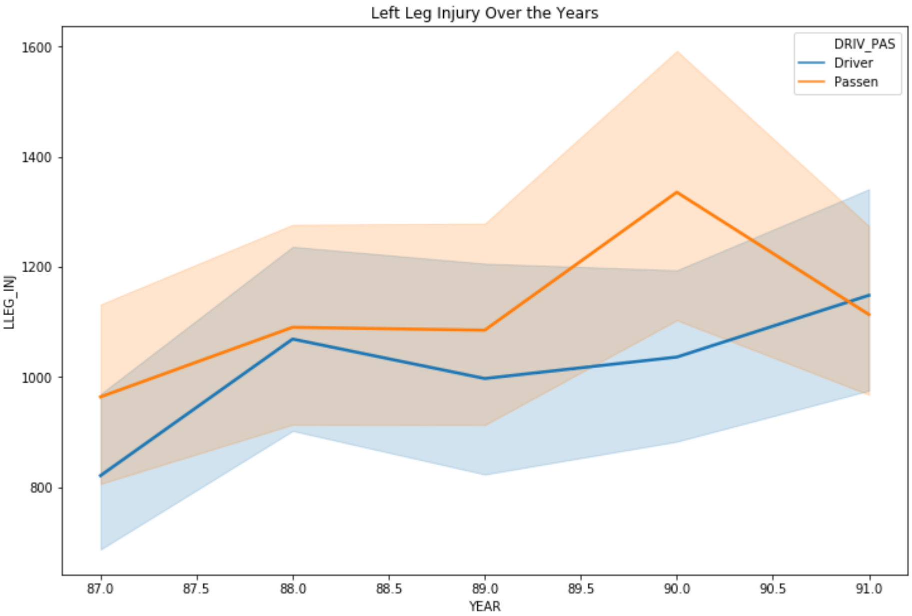


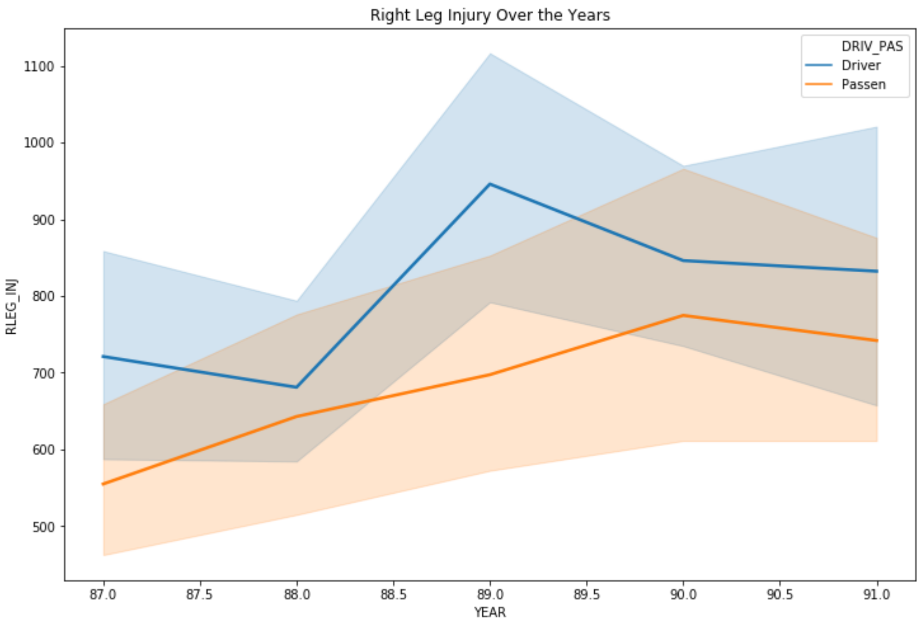
12. Driver had more injury or passenger (line graph plot of driver and passenger based on various factors)

| | DRIV_PAS | HEAD_INJ | CHEST_IN | LLEG_INJ | RLEG_INJ | DOORS | YEAR | WEIGHT | SIZE2 | PROTECT2 | intercept |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Driver | 973.298851 | 51.712644 | 1010.316092 | 801.867816 | 3.143678 | 88.919540 | 2913.109195 | 3.551724 | 1.919540 | 1.0 |
| 1 | Passen | 823.402439 | 45.140244 | 1108.743902 | 674.731707 | 3.115854 | 88.841463 | 2892.103659 | 3.554878 | 1.871951 | 1.0 |




Left Leg Injury Over the Years
DRIV_PAS
Driver
Passen
1600
1400
1200
1000
800
LLEG_INJ
87.0
87.5
88.0
88.5
89.0
89.5
90.0
90.5
91.0
YEAR


Right Leg Injury Over the Years
DRIV_PAS
Driver
Passen
1100
1000
900
800
700
600
500
RLEG_INJ
87.0
87.5
88.0
88.5
89.0
89.5
90.0
90.5
91.0
YEAR

# CHAPTER 4

## 4. Conclusion

The analysis help to predict these things:

I. Head injury based on all other variables

II. Head injury based on all other variables when coliniarity is taken in to consideration.

III. Total injury based on all other variables

IV. Total injury based on all other variables when coliniarity is taken in to consideration.

V. The indicator variables are MAKE and DRIV_PAS

VI. The superfluous variable is CHEST_INJ

# CHAPTER 5

## 5. Appendix

**#Read the dataset**

```
ds=read.csv("F:/crash-dat-analysis-master/crash-dat-analysis-master/Crash.csv",sep =
'\t')
View(ds)
```

**#Data cleaning**

```
ds$CHEST_IN[is.na(ds$CHEST_IN)]<-mean(ds$CHEST_IN,na.rm=TRUE)
ds$HEAD_INJ[is.na(ds$HEAD_IN)]<-mean(ds$HEAD_IN,na.rm=TRUE)
ds$LLEG_INJ[is.na(ds$LLEG_IN)]<-mean(ds$LLEG_IN,na.rm=TRUE)
ds$RLEG_INJ[is.na(ds$RLEG_IN)]<-mean(ds$RLEG_IN,na.rm=TRUE)

ds$DOORS=as.numeric(ds$DOORS)

ds$DOORS[is.na(ds$DOORS)]<-mean(ds$DOORS,na.rm=TRUE)
ds$WEIGHT[is.na(ds$WEIGHT)]<-mean(ds$WEIGHT,na.rm=TRUE)
ds$SIZE2[is.na(ds$SIZE2)]<-mean(ds$SIZE2,na.rm=TRUE)
ds$PROTECT2[is.na(ds$PROTECT2)]<-mean(ds$PROTECT2,na.rm=TRUE)

#ds$SIZE2=as.numeric(ds$SIZE2)
```

**#Indicator variables**

```
lv = levels(ds$MAKE)
lb = length(levels(ds$MAKE))

ds$MAKE=as.numeric(factor(ds$MAKE,
          levels=lv,
          labels = c(1:lb)) )
```

```
ds$DRIV_PAS=factor(ds$DRIV_PAS,
                   levels=c('Driver','Passen'),
                    labels=c(1,2))
```

**#Data spliting in training set and validation set**

```
ind=sample(2,nrow(ds),replace=TRUE,prob = c(0.8,0.2))
```

**#traing set made of 80% of the data**

```
tdata=ds[ind==1,]
```

**#validation set made of 20% of the data**

```
vdata=ds[ind==2,]
```

**#Multiple regression**

**#regression model using all the variables**

```
result1=lm(HEAD_INJ~CHEST_IN+MAKE+LLEG_INJ+RLEG_INJ+DRIV_PAS+
DOORS+YEAR+WEIGHT+SIZE2+PROTECT2,tdata)
```

**#for all variables**

```
y_pred=predict(result1,newdata = vdata)
```

**#Using backward elimination to get the best regression model**

```
step(result1, direction = "backward")
```

**#based on backward elimination best model is:**

```
result1=lm(formula = HEAD_INJ ~ CHEST_IN + DOORS + WEIGHT +
PROTECT2,
   data = tdata)
y_pred=predict(result1,newdata = vdata)
View(y_pred)
```

**#creating a new column which is based on all the four type of injury**

```
ds$TOTAL_INJ=ds$HEAD_INJ+ds$CHEST_IN+ds$LLEG_INJ+ds$RLEG_INJ
```

**#Data spliting in training set and validation set**

```
ind=sample(2,nrow(ds),replace=TRUE,prob = c(0.8,0.2))
```

**#traing set made of 80% of the data**

```
tdata=ds[ind==1,]
```

**#validation set made of 20% of the data**

```
vdata=ds[ind==2,]
```

**#Multiple regression**

**#regression model using all the variables**

```
result1=lm(TOTAL_INJ~MAKE+DRIV_PAS+DOORS+YEAR+WEIGHT+SIZE2+
PROTECT2,tdata)
```

**#for all variables**

```
y_pred=predict(result1,newdata = vdata)
```

**#Using backward elimination to get the best regression model**

```
step(result1, direction = "backward")
```

**#based on backward elimination best model is:**

```
result1=lm(formula = TOTAL_INJ ~ MAKE + DRIV_PAS + WEIGHT, data = tdata)
y_pred=predict(result1,newdata = vdata)
View(y_pred)
```