**Objective -**

* Build the best multiple regression model that you can for the purpose of predicting 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 we account for multicollinearity, and another where we don't consider multicollinearity.
* 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 you can 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?

**Success Criteria 🡪** predicting the severity of injury

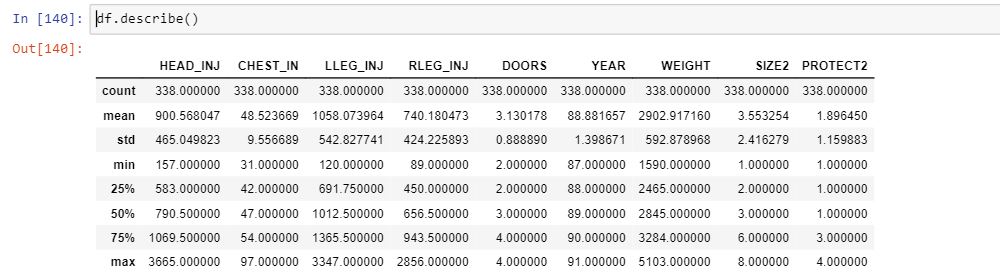
**Resources 🡪** Crash.dat

**Technology** 🡪 R studio

**Data mining success criteria**🡪Multiple regression

**Data Understanding**

Data summary🡪



**Data preparation**

Fields to be included 🡪

MAKE object

MODEL object

CARID object

CARID\_YR object

HEAD\_INJ object

CHEST\_IN object

LLEG\_INJ object

RLEG\_INJ object

DRIV\_PAS object

PROTECT object

DOORS object

YEAR int64

WEIGHT int64

SIZE object

SIZE2 int64

PROTECT2 object

Clean data🡪

Number of null values

HEAD\_INJ 🡪 14

CHEST\_IN 🡪 11

LLEG\_INJ 🡪 9

RLEG\_INJ 🡪 11

DOORS 🡪 66

PROTECT2 🡪 4

To integrate the data we formatted the data types as per required, which are as follows:-

HEAD\_INJ object🡪int64

CHEST\_IN object🡪int64

LLEG\_INJ object🡪int64

RLEG\_INJ object🡪int64

DOORS object🡪int64

PROTECT2 object🡪int64

**Modeling**

model technique used🡪 multiple regression

**Evaluation**

Evaluate results🡪 asses of data mining result w.r.t business criteria

Approve Model🡪

Review process🡪

Determine next steps🡪list of possible action

**Deployment**

Plan deployment🡪

Maintenance🡪

Produce final report🡪

Review project🡪

np.corrcoef(data[‘TV’],data[Sales])

mapplot🡪axes3d

rss

gradient (multiple linear)and leastsquare(linear)

rmse(root mean square error)

sklearn.svm(package)🡪

web scrapping

BeautiFulSoup(package)

SoupStrainer(package)

**Itroduction**

The formula for Simple Linear Regression:

Y=B0+B1x

But, for Multiple Linear Regression, it is

Y=B0+B1x1+B2x2+B3x3+……...+Bnxn

Due to the nature of the regression equation, your **x** variables have to be continuous as well.

Thus, you’ll need to look into changing your categorical variables into continuous ones.

Continuous variables are simply put, running numbers. Categorical variables are categories.

Categorical Variables are also referred to as discrete

or qualitative variables. There are 3 types:

Nominal : more than 2 types. e.g color

Dichotomous: 2 types e.g yes or no

Ordinal: more than 2 types, but have a rank/order e.g Below average, Average, Above Average

### **Feature Selection**

Having too many variables could potentially cause your model to become less accurate, especially if certain variables have no effect on the outcome or have a significant effect on other variables.

*Basic step of Feature Selection: Use your Common and/or Business Sense(s)*

*One other way to select features is to*

*use the p-values. As we*[*last discussed*](https://medium.com/swlh/super-simple-machine-learning-simple-linear-regression-part-3-validation-65b8c11fa36b)*,*

*p-values tell you how statistically*

*significant the variable is.*

*Removing variables with high pvalues*

*can cause your accuracy/R squared to*

*increase, and even the p-values of the*

*other variables to increase as well — and*

*that’s a good sign.*

This action of omitting variables is part of [stepwise regression](https://en.wikipedia.org/wiki/Stepwise_regression). There are 3 ways to do this:

* Forward Selection: Start with 0. Run the model over and over, trying each variable. Find the variable that gives the best metric (E.g p-values, Adjusted R-squared, SSE, % accuracy) and stick with it. Run model again with the chosen variable, trying one of the remaining variables at each time and sticking with the best one. Repeat process until the adding does not improve the model any more.

Simple linear regression, is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

* One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.
* The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

Checking for collinearity helps you get rid of variables that are skewing your data by having a significant relationship with another variable.

Multicollinearity exists when two or more of the

predictors (x variables) in a regression model are

moderately or highly correlated (different column).

When one of our predictors is able to strongly

predict another predictor or have weird

relationships with each other (maybe x2 = x3

or x2 = 2(x3) + x4), then your regression equation

is going to be a mess.

The simplest method to detect collinearity would be to plot

it out in graphs or to view a correlation matrix to check out

pairwise correlation (correlation between 2 variables).