Car Crash EDA

Matt Oehler

March 12, 2018

1 Description

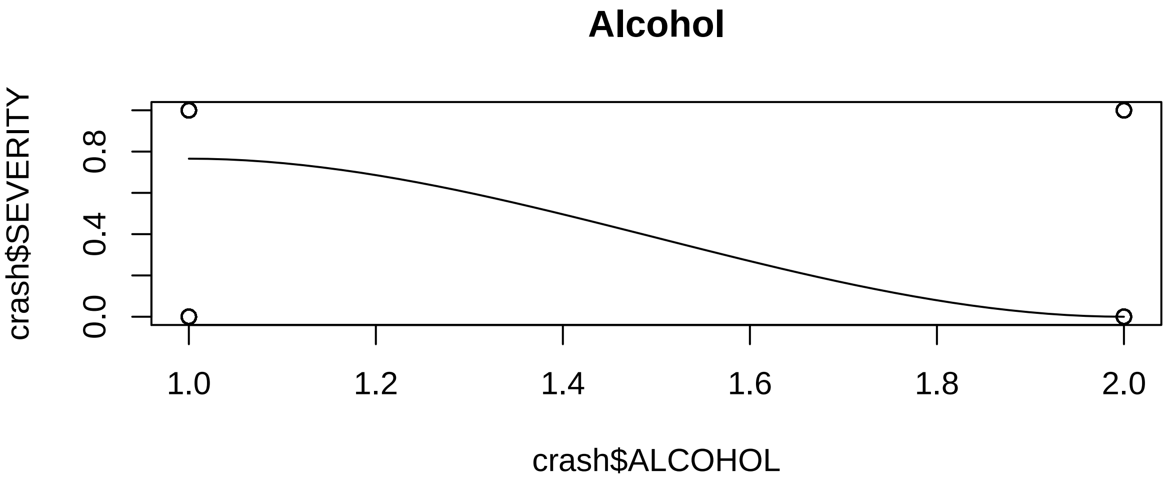
Traffic accidents are the cause of several thousand deaths each year. The FHWA is responsible for increasing the safety of highways and road ways. The GES collects data about car crashes. Using the data collected by the GES we hope to be able to see which independent features relate with car crashes. This will hopefully lead to better policies that can help to improve road conditions and save lives.

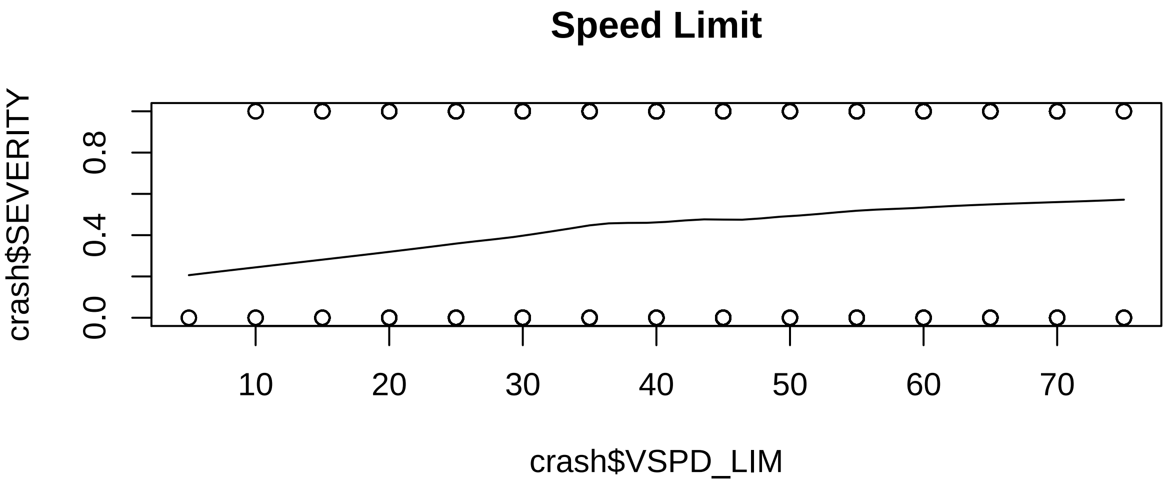
2 Data

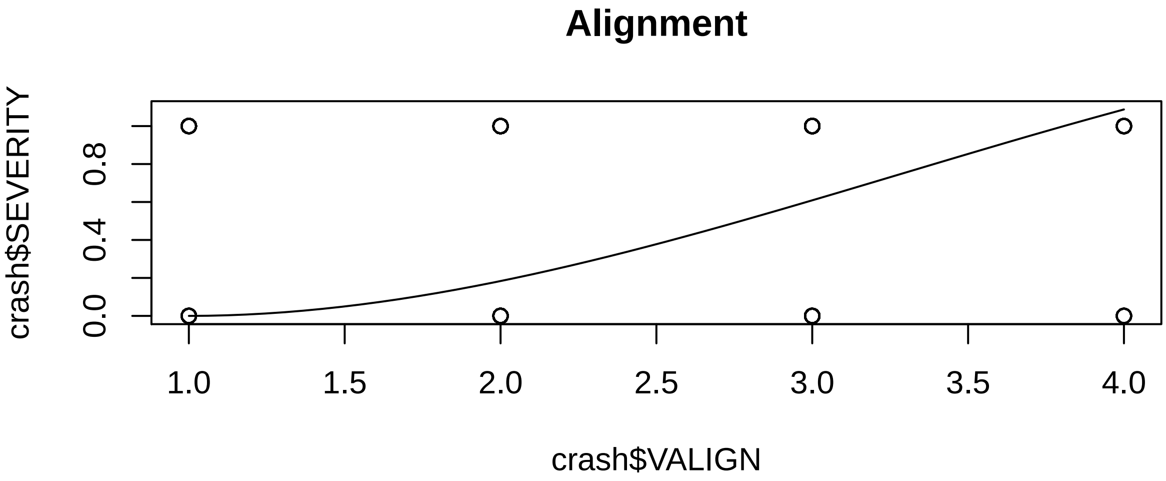
The data consists of over 8000 observed accidents, and has characteristics about each accident such as speed limit, type of intersection, involvement of alcohol, and etc. The severity of the accident is recorded as a binary variable, and the table below shows that the data set is very balanced between severe and non severe accidents.

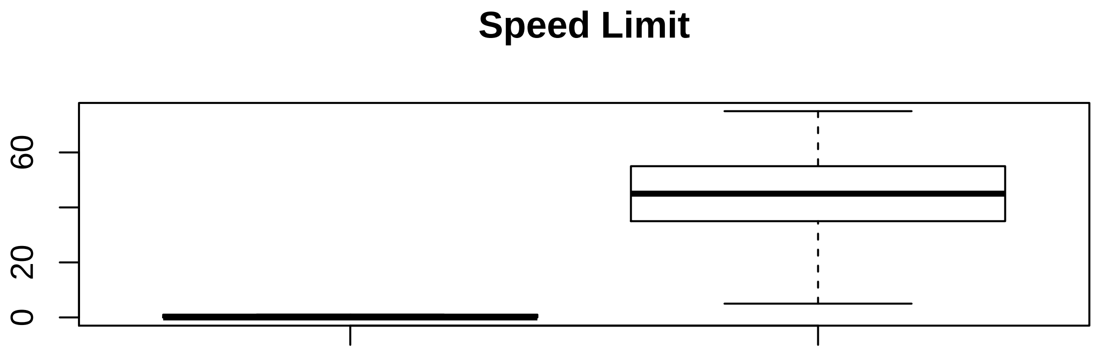
	Severity
0	4549
1	4054

Since the response variable is categorical we can't use standard scatter plots to view the data. But by using scatterplots with a smooth local regression line, we can see over all trend of certain features in how they relate to the response. Box plots can also be useful when assessing how quantitative variables relate to categorical variables. The plots below show how alcohol, speed limit, and road alignment relate to crash severity.




Speed Limit
crash$SEVERITY
0.0
0.4
0.8
10
20
30
40
50
60
70
crash$VSPD_LIM


Alignment
crash$SEVERITY
0.0
0.4
0.8
1.0
1.5
2.0
2.5
3.0
3.5
4.0
crash$VALIGN



3 Method

Logistic regression would be an appropriate method for this analysis because it can be used to classify data. Using this method we can probabilities that a certian accident is severe or not severe.

4 Something I don't know

I suggested logistic regression as the appropriate methodology, but that is only because I am familiar with it. I have never used other methods. I am curious to know what the strengths and weaknesses are of each method. Then I would be able to pick a method that I feel best applies to this problem.