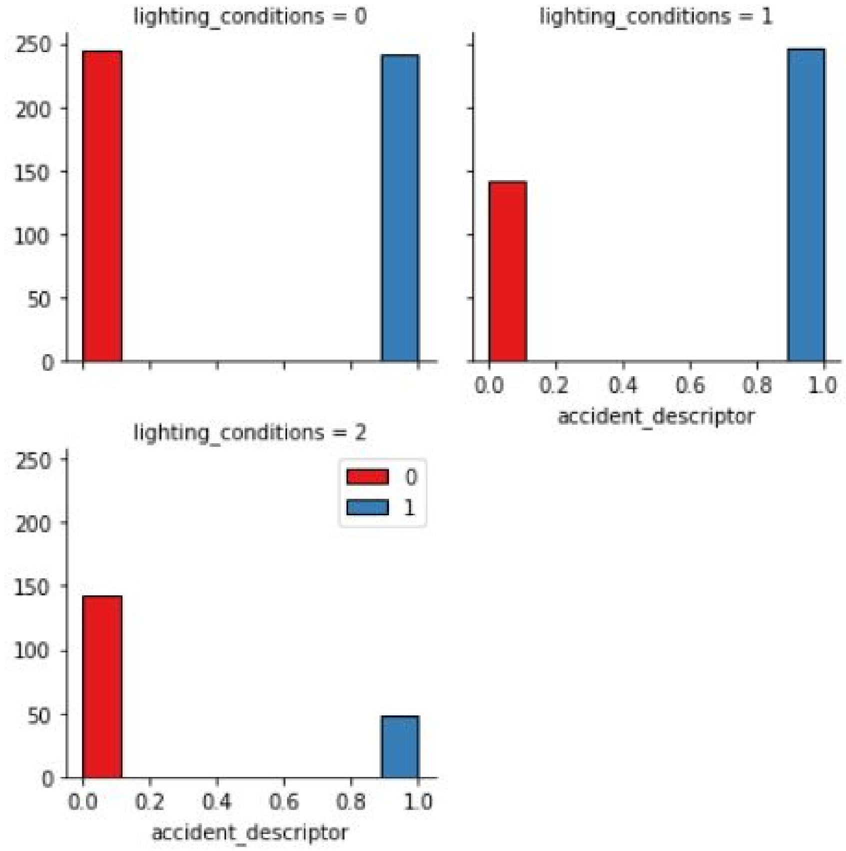
Coursera Capstone Project
Car Accident Severity

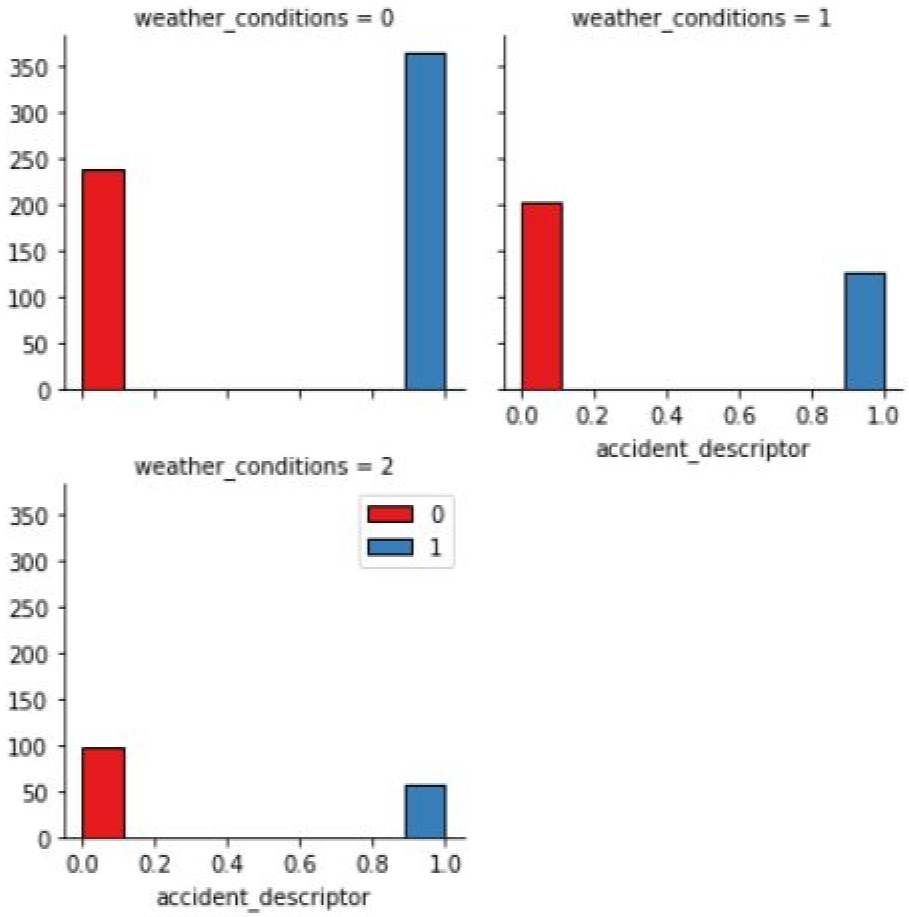
Car accidents caused over 30,000 deaths per year in the U.S. alone. This is in addition to the enormous cost of injuries and property damage. Many of the victims of these accidents receive some form of compensation form insurance companies. How do these insurance companies know which accidents really happened, and which are fraudulent? By using data science we can build a model which takes into account the circumstances of the accident and reports the likely severity of the crash. This can then find outliers which can be investigated by insurance companies. This could save insurance companies money and time that would otherwise be wasted on investigating more reasonable accidents.

To accomplish the task of building a data science model to predict how reasonable the severity of a car crash is we will use the Traffic Crashes dataset from the City of Las Vegas. This dataset includes a large number of instances, and has information on the severity of the crash as well as the circumstances, such as lighting, crash type, injury type, and road condition.

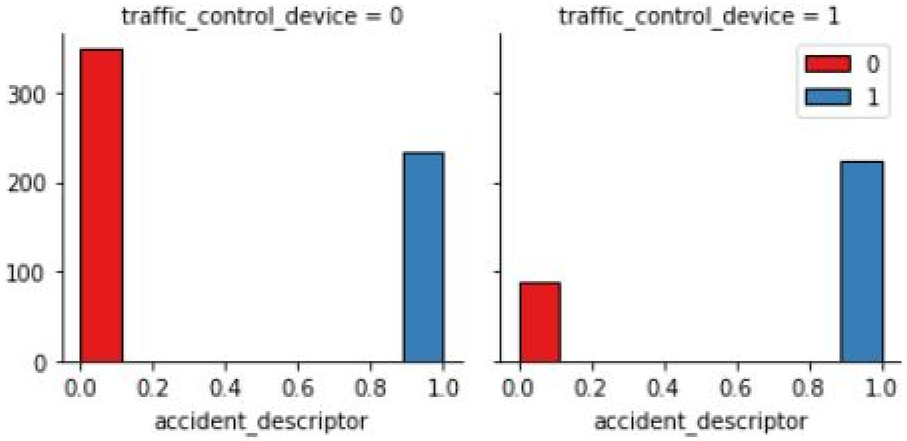
The dataset chosen included many columns which were not relevant to the analysis, such as county, date, police report, and others. These columns were removed, leaving accident description as the dependant variable, and lighting conditions, road conditions, weather conditions, road surface conditions, traffic control device, and crash type as the independent variables. Many of these variables are categorical and were changed to numeric in order to run machine learning algorithms on them.



In this graph you can see that lighting condition 1, a dark lighted road, results in more injuries than lighting option 0, daylight.



In the above graph you can see how weather condition affects accident severity, with snowy conditions resulting in more injuries and fatalities than dry or wet roads.



This graph shows that the presence of a stop sign makes the crash much more likely to involve an injury.

Four machine learning algorithms were used to model the data on car crashed; KNN, SVM, Decision Tree, and Logistic Regression. All of these models had accuracy scores of between 55% and 65%, with SVM being the most accurate.

	F1	Jaccard	Log Loss
KNN	0.6293	0.6301	NA
SVM	0.6300	0.6301	NA
Decision Tree	0.6063	0.6070	NA
Logistic Regression	0.5873	0.5894	0.6812

In conclusion, machine learning can be used to assess how likely a car crash is given relevant circumstances, such as road condition, weather, lighting, and the presence of signs or lights. However the accuracy of our models was lower than expected, at 65%. Further work could refine the data used and add additional variables to increase accuracy.