Predicting Car Accident severity

Applied Data Science Capstone

**Andres Michel**

github.com/andresmichel92

**Introduction.**

**Background**

During 2019, almost 40 thousand people lost their lives to car crashes solely in the U.S. and even though in recent years the mortality rate has been declining, it is still a leading cause of death. The odds of dying in a car crash in the US are of one in 103.

Being able to identify and predict when car crashes will occur, allows people to avoid certain roads, take extra precautions when driving under certain weather conditions and even alert authorities and manufacturers of trends and risk factors that could translate into policies and design improvements to help save lives.

**Problem**

Motor vehicle road accidents is a leading cause of death, and having enough data accounting for location, vehicle characteristics, road and weather conditions as well as the outcome (severity of car crash) can be used to build a model that predicts the chances of a high severity car crash to occur under certain conditions.

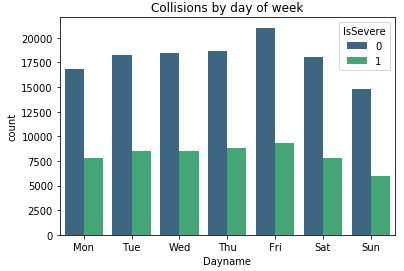
**Data**

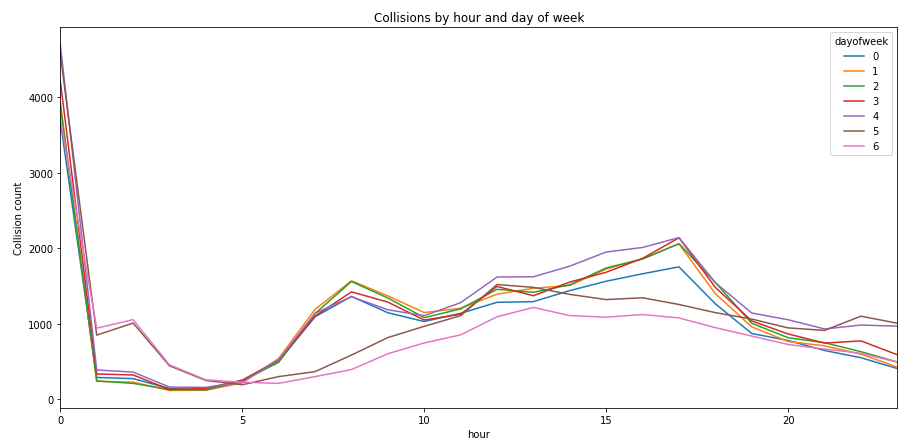
We’ll be using the Seattle GIS collision data, provided by the City of Seattle all data was obtained by SPD and recorded by Traffic Records. This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present.

The data consists of 37 attributes and 194,673 rows (or accidents), with information ranging from Incident ID to Weather and Light conditions. Thus, in order to work with the data we’ve proceeded to clean it up, removing rows with missing data and removing redundant columns. After cleansing the data, we ended up with 182,660 rows.

**Data EDA**

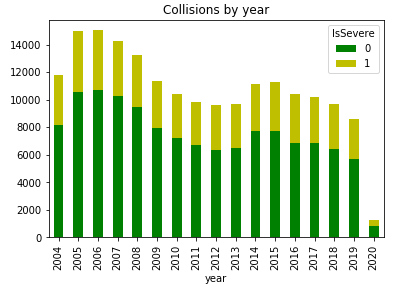
We proceeded to make some sense of the data by trying to identify the seasonality of the severe collisions:



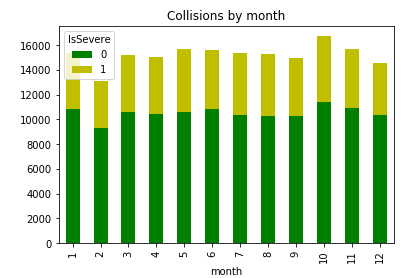


Given these two graphs we can identify a common trend, as more collisions are registered during Friday, and less on Sunday. Same can be said from Hour of day, where most accidents occur during midnight as well as in the afternoon.

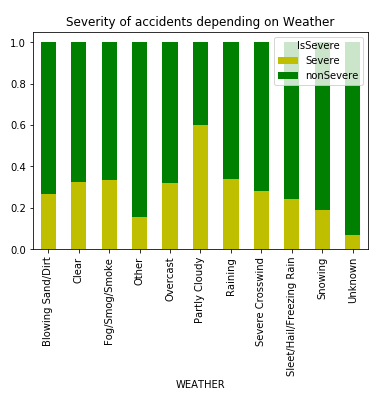
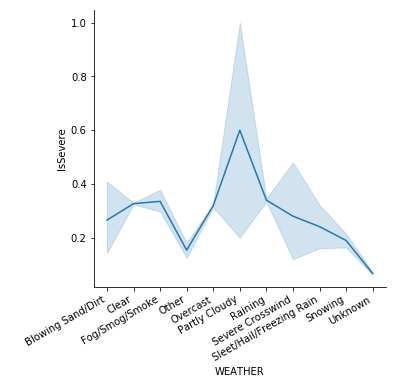
Analyzing the information by Year, we notice that the total number of car accidents has been in decline over the last five years.



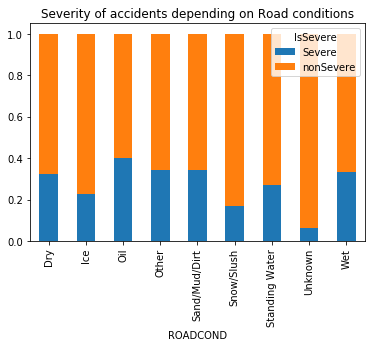
As per seasonality, we’ve identified that October is the month with more collisions registered.



In relationship with the weather conditions, we can note that the chances of a Collision increases with a cloudy sky:



A similar assessment can be made for the condition of the road:



Finally, we’ve plotted the correlation matrix in a heatmap for some of the attributes selected as feature variables.

