Final Report: Predicting and preventing US traffic fatalities

Background and Problem:

Fatalities on US highways in 2015 increased by 7.2% from 2014 to 35092 individuals. The goal of this project is to analyze traffic fatality and supporting data to identify spatial patterns and factors that are related to fatal traffic incidents. Additionally, the 2014 and 2015 data will be compared in detail to identify specific segments of the population that saw increased fatalities.

Client:

This project is interesting to state and local planners, law enforcement, and driving instructors. This results will identify communities that are at higher risk of fatal crashes and also identify contributing risk factors likely to be involved, such as speeding, intoxicated driving, and adverse weather. These insights could inform targeted campaigns to reduce the risk factors or help police be better prepared.

Dataset(s):

Main Dataset:

The main dataset used in this project was the Fatality Analysis Reporting System (FARS) database, collected by the National Highway Traffic Safety Administration (NHTSA). The FARS database spans the years of 1979-2016 and is a complied list of features describing every traffic accident with at least one fatality occurring on US public roads. The database is available as a collection of DBF files, one for each year, and can be downloaded from the NHTSA FTP site (ftp://ftp.nhtsa.dot.gov/fars). Relevant features include the conditions of the accident, such as weather, location, date/time, speeding, and alcohol involvement; information about the driver(s) and any passengers, such as age and sex; and information about the involved vehicle(s), including make, model, and year.

Supplementary datasets:

Data from the US Census was used to standardize the statewide statistics by state population, giving per capita metrics.

Data relating to state traffic laws was obtained from the Insurance Institute for Highway Safety (IIHS). Statewide laws regulating alcohol-related offenses in particular were compiled, specifically whether a drivers license is suspended, the length of time before driving privileges are restored, and whether an interlock device is required.

Statewide alcohol tax data were obtained from the Alcohol Policy Information System. Taxes rates are reported by year and divided into taxes for beer, wine, and spirits.

Alcohol consumption data were obtained from the National Institute on Alcohol Abuse and Alcoholism and describe the consumption rate of various alcohol types by state and year.

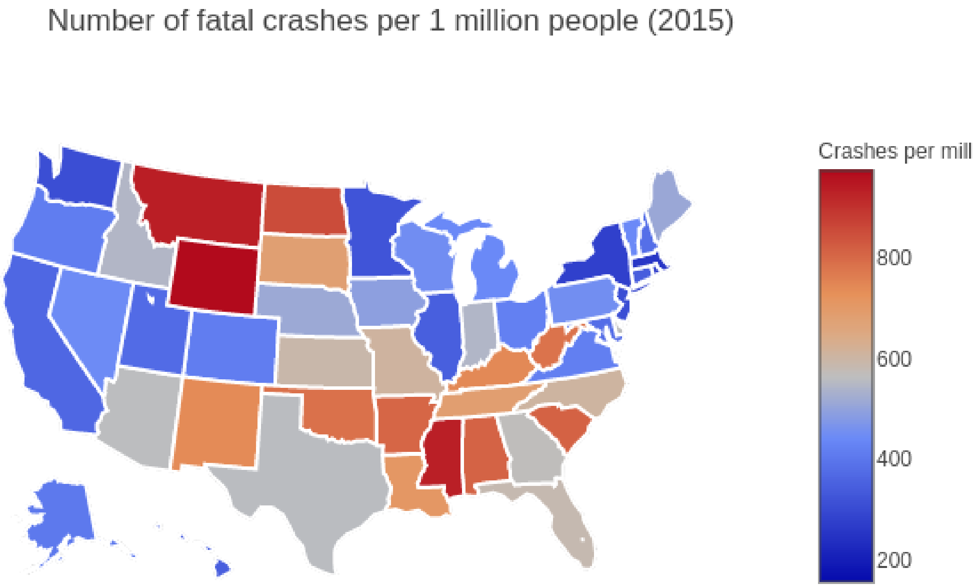
Data Wrangling:

The majority of the datasets are available as DBF or CSV files and easily readable into Python. The exception are the IIHS traffic law data, which were scraped with BeautifulSoup. The datasets were then combined by state and year. Any parameters that were not already expressed as per capita values were converted using the census-derived population estimates.

Exploratory analysis:

Initial data exploration focused on describing the spatial and temporal patterns in fatal traffic accidents in the US. The predominate causes of fatal accidents were also examined.

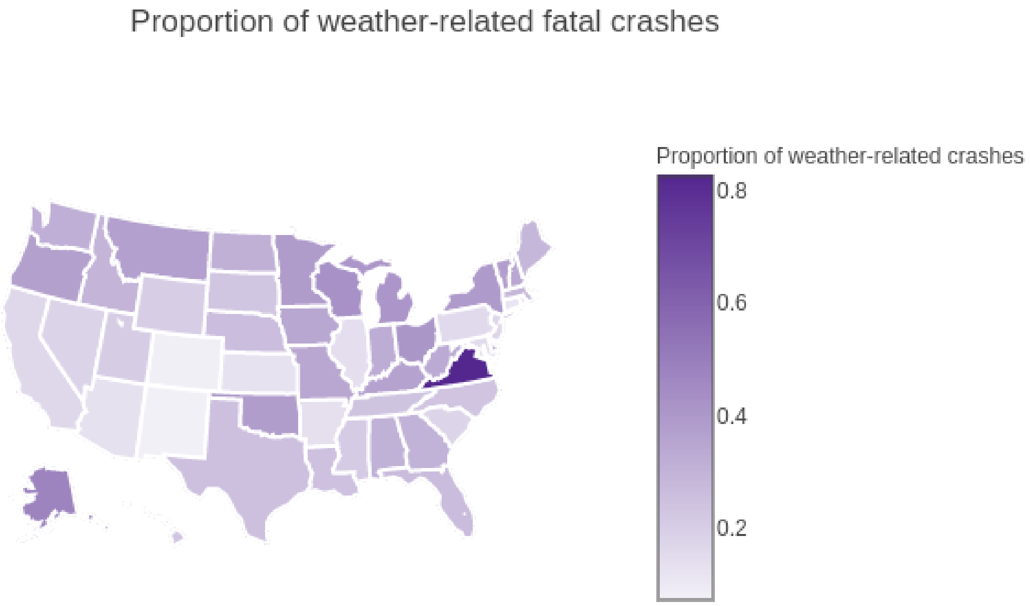
Where do fatal accidents occur?



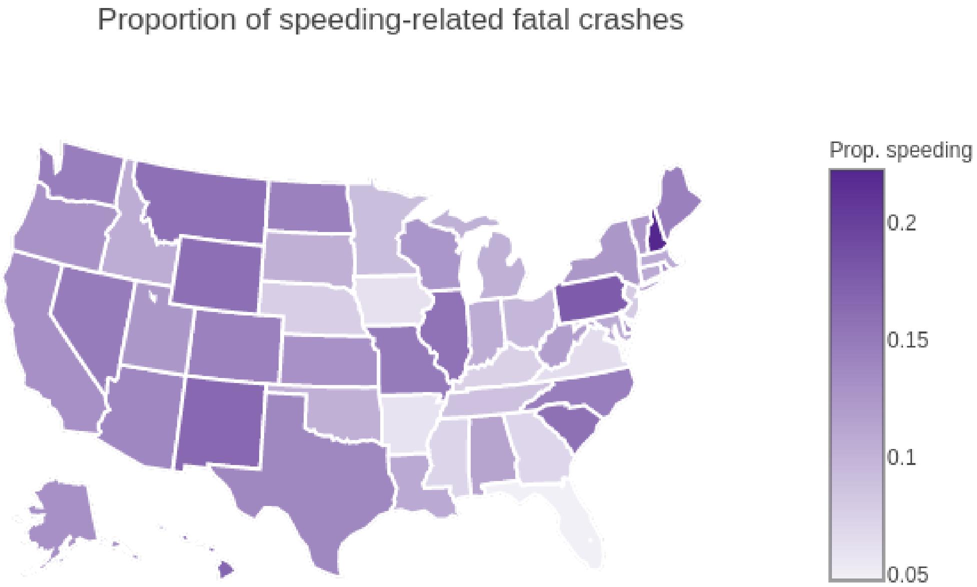
When adjusted by statewide population, the most fatal accidents occur in southern sates and the northern midwest. This figure contains data from 2015, and the pattern is representative of all years examined (2010-2015).

What are the predominate causes of fatal accidents?

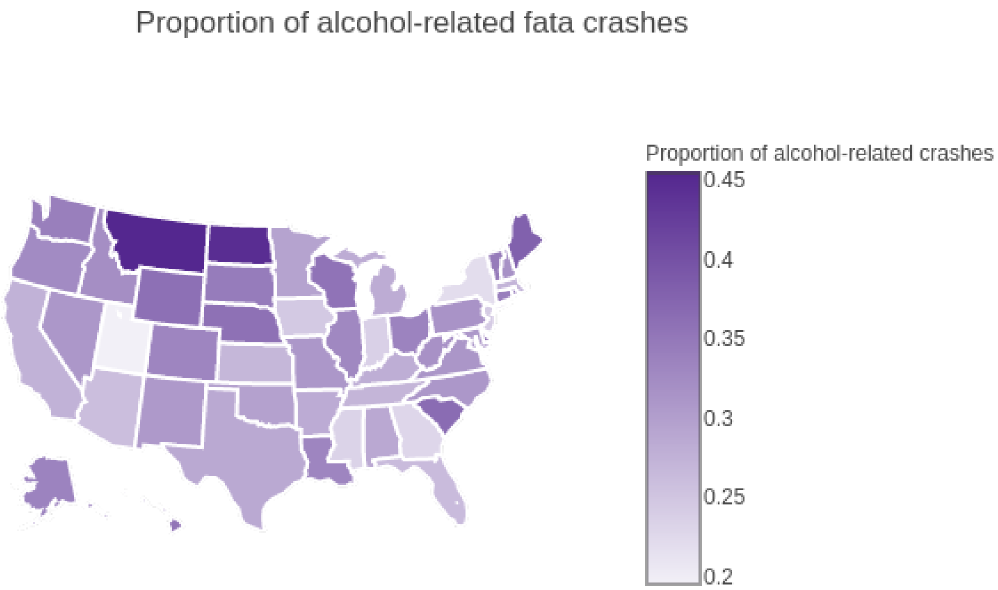
Three conditions were investigated as predominate causes of fatal traffic accidents: weather, speeding, and alcohol involvement. These causes are not mutually exclusive.


Proportion of weather-related fatal crashes
Proportion of weather-related crashes
0.8
0.6
0.4
0.2

Weather was typically involved in less than 40% of the fatal crashes within a state, with the exception of Virginia, which reported a percentage of 76% or fatal crashes involving weather.



Speeding was involved in a relatively low percentage of fatal accidents, ranging from 6% to 22%.

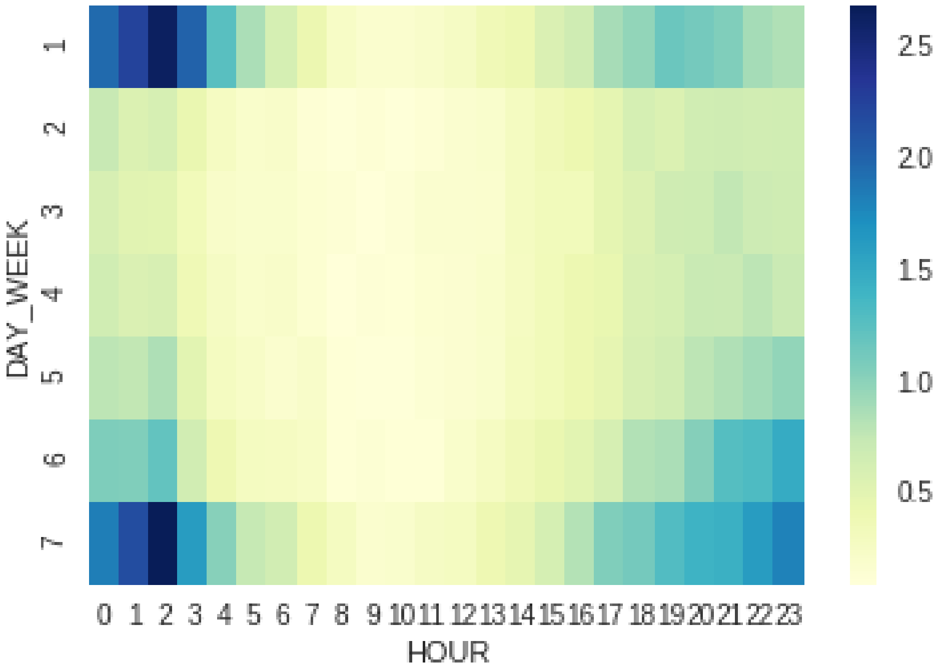

Proportion of alcohol-related fata crashes
Proportion of alcohol-related crashes
0.45
0.4
0.35
0.3
0.25
0.2

Alcohol was involved in 22-44% of fatal accidents by state. Because of the high values in Montana and North Dakota, two of the highest total fatality states, and the ability to potentially regulate alcohol usage and driving, the subset of crashes involving alcohol is the focus of the rest of the analysis.

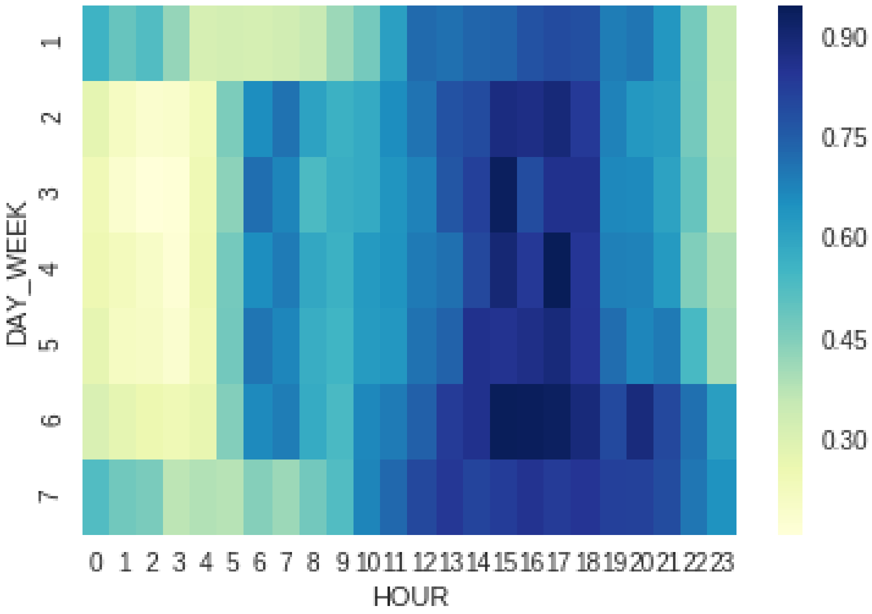
When do fatal accidents occur?

To determine when law enforcement officers should target there efforts against intoxicated driving, the timing of fatal accidents involving alcohol and not involving alcohol was compared.

Alcohol-involved fatal accidents were heavily concentrated during the early morning hours of Saturday and Sunday. This intuitively makes sense because these are times when people would be traveling home after spending the night out on the weekend.



In contrast, fatal crashes not involving alcohol are concentrated in the afternoon and early evening, particularly of weekdays, when people are likely traveling home from work.



Predictive Modeling:

Three modeling approaches were used to explore the relationship between statewide predictor variables and the prevalence of alcohol-involved fatal crashes. Initially a simple multiple regression model was used. This approach was used as a baseline to compare the performance of decision tree, which has the added advantage of being able to capture nonlinear relationships between the outcome and predictors, and a random forest approach, which is an ensemble of decision trees.

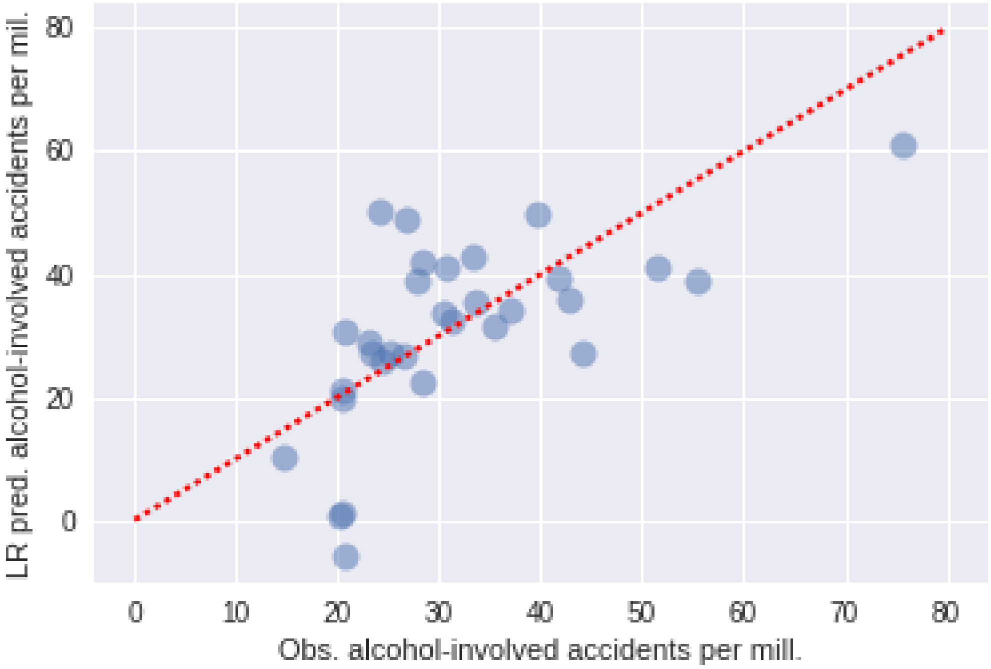
For each model, model predictions were generated for the training dataset, a randomly selected 4/5 of the dataset used to develop the model, and the training dataset, the remaining 1/5 of the dataset. Cross-validation with five folds was used to prevent over fitting and representative plots for five folds are shown below.

Model evaluation:

The modeling approaches were compared and evaluated by plotting the relationship between the observed number of alcohol-involved fatal crashes (standardized by population size) and the number predicted by the various modeling methods.   A strong 1:1 relationship between the observed and model-predicted data indicates a good model fit, which is indicated by the red dashed line. Root mean square error (RMSE) was also used to evaluate model performance. This metric is the square root of the variance of the residuals and describes how close the modeled values are to the actual observed values. Lower RMSE indicates a better model fit.

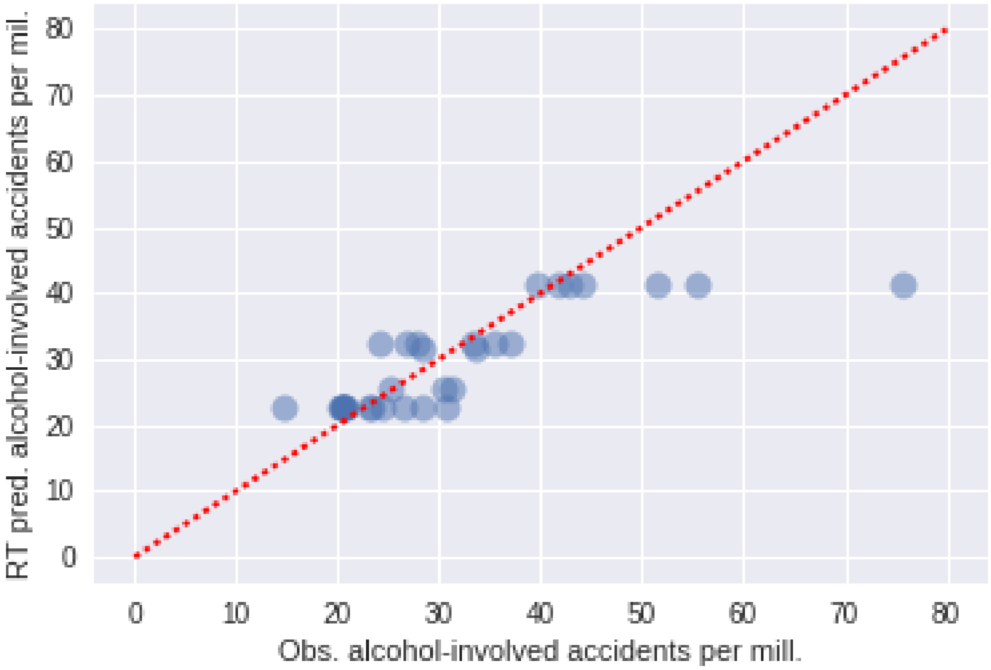
Model performance:

Multiple regression:



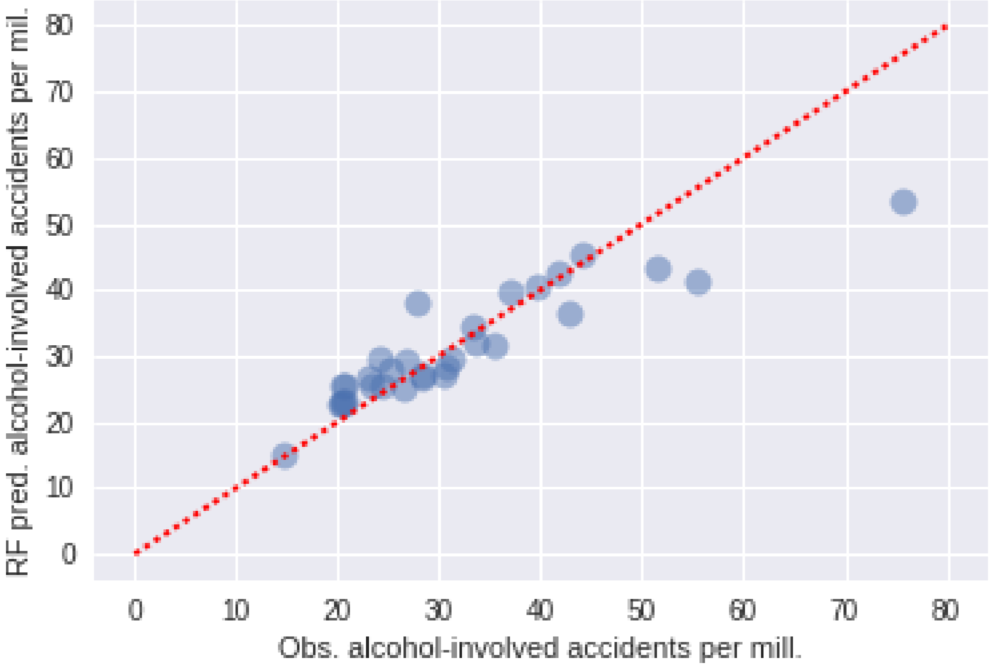
Predicted vs. actual alcohol-involved crashes based on linear regression model for test data.

Decision tree:



Predicted vs. actual alcohol-involved crashes based on decision tree model for test data.

Random forest:



Predicted vs. actual alcohol-involved crashes based on random forest model for test data.

Modeling Approach	**RMSE**
Multiple linear regression	11.948
Decision tree (regression)	7.909
Random forest	5.920

Based on the RMSE values for the three modeling approaches, the random forest model produces the best fit for the test dataset. Therefore, the random forest model will be used to identify important predictor variables and develop recommendations.

The best-fit RF model explained 73% of the variation in the test dataset. The top three predictor variables contributing to the model, identified by Gini importance score, are alcohol consumption rate, beer tax rate, and whether an interlock is required following an alcohol-related offense. Variables with high importance scores contribute more to the model and should therefore be targeted to reduce alcohol-involved traffic fatalities.

Recommendations:

1) To reduce alcohol-involved fatal crashes, law enforcement efforts should be concentrated between midnight and 3am on Saturday and Sunday mornings.

2) Beer tax rate is a key predictor of alcohol-involved crash rate. Increasing the beer tax rate may reduce this group of fatal crashes, likely by reducing consumption.

3) Interlock requirement laws are key predictors of alcohol-involved crash rate. Requiring interlocks after alcohol-related offenses may reduce this group of fatal crashes, likely by reducing the prevalence of reoffenders.