**Why are Traffic Fatalities Rising?**

**Milestone Report**



**INTRODUCTION**

Following a 50 year trend of decline, 2015 saw a 7% increase in traffic accident related deaths over 2014. In August 2016, the National Highway Traffic Safety Administration along with the Whitehouse put out a call to action to investigate this alarming increase. Further research is need to understand what might have contributed to this significant increase. Moreover, recent advances in big data analytics could allow for new types of analysis that were not available in the past which could help uncover differences in specific causes of traffic fatalities year over year.

In this project I explore what factors may have contributed to increasing the likelihood of a less common types of fatal traffic accidents (e.g. distracted drivers, drowsy drivers) as well as how these factors intersect. (e.g. Is drowsiness as a factor contributing to fatal crashes worse among truck drivers/bus drivers? Are traffic fatalities associated with distracted driving higher among young drivers?)

**PROJECT OBJECTIVES**

More specifically, I am interested in predicting (1) factors associated with less common types of fatal accidents, such as those involving distracted drivers and drowsy drivers, for example, as well as (2) whether I can predict the actual incidence of these less common types of accidents with much accuracy.

I have chosen to focus on distracted drivers because the smartphone market has exploded over the past 5 years with the number of smartphone owners in the US rising from 62 million in 2010 to more than 207 million in 2016. While growth is slowing, the number of smartphone users in America is still trending upward.

(<https://www.statista.com/statistics/201182/forecast-of-smartphone-users-in-the-us/>)

This is a potential area of concern for increased traffic fatalities and. Potential stakeholders include insurance companies, paramedics, hospitals, drivers, and government road safety policy makers. Findings could result in policy changes, including further restrictions on cell phone use while driving or more specific public awareness campaigns among subgroups of the country, such as region.

**DATASET**

The primary dataset used for this project is the 2015 US Traffic Fatalities datset

<https://www.kaggle.com/anokas/2015-us-traffic-fatalities>

Each year has three datasets that comprise the Traffic Fatalities data released by the National Highway Traffic Safety Administration in conjunction with the U.S. Department of Transportation:

1)      Dataset of accidents (38 variables,  >32,000 rows of data)

2)     Dataset of persons (22 variables, >80,000 rows of data)

3)     Dataset of vehicles (16 variables,>40,000 rows of data )

Limitations of the datasets:

Following the initial exploration of the each dataset, I discovered that, unfortunately, there are multiple rows per case in both the *Persons* and *Vehicles* datasets that make it somewhat difficult to link them to the Accidents dataset cleanly. Unfortunately this means that I cannot link demographic information (apart from age) or vehicle information to the distracted driver. Therefore, since the *Accidents* data is the primary dataset I want to use to investigate distracted drivers, I have decided to focus on this dataset alone for this project.

The accidents dataset has 38 columns and 32,167 rows each representing a traffic accident that involved at least one fatality. The dataset includes a case number for each accident as well as information about the accident. While I will explore all the variables in the dataset, I am particularly interested in how factors like, age, region, urban rural setting, interstate, intersection, manor of collision, day of the week, time of day, and alcohol, are related to traffic fatalities involving a distracted driver.

**INITIAL DATA CLEANING AND WRANGLING**

Since the data were released by the National Highway Traffic Safety Administration in conjunction with the U.S. Department of Transportation, they were very tidy and clean from the outset. The data were already arranged in columns, with cases as rows.  All of the data in the *Accidents* dataset were discrete, apart from the number of fatalities involved in the accident.

Nearly all of the variables in the *Accidents* dataset were discrete integers.  I converted them to factor variables for the initial data exploration.

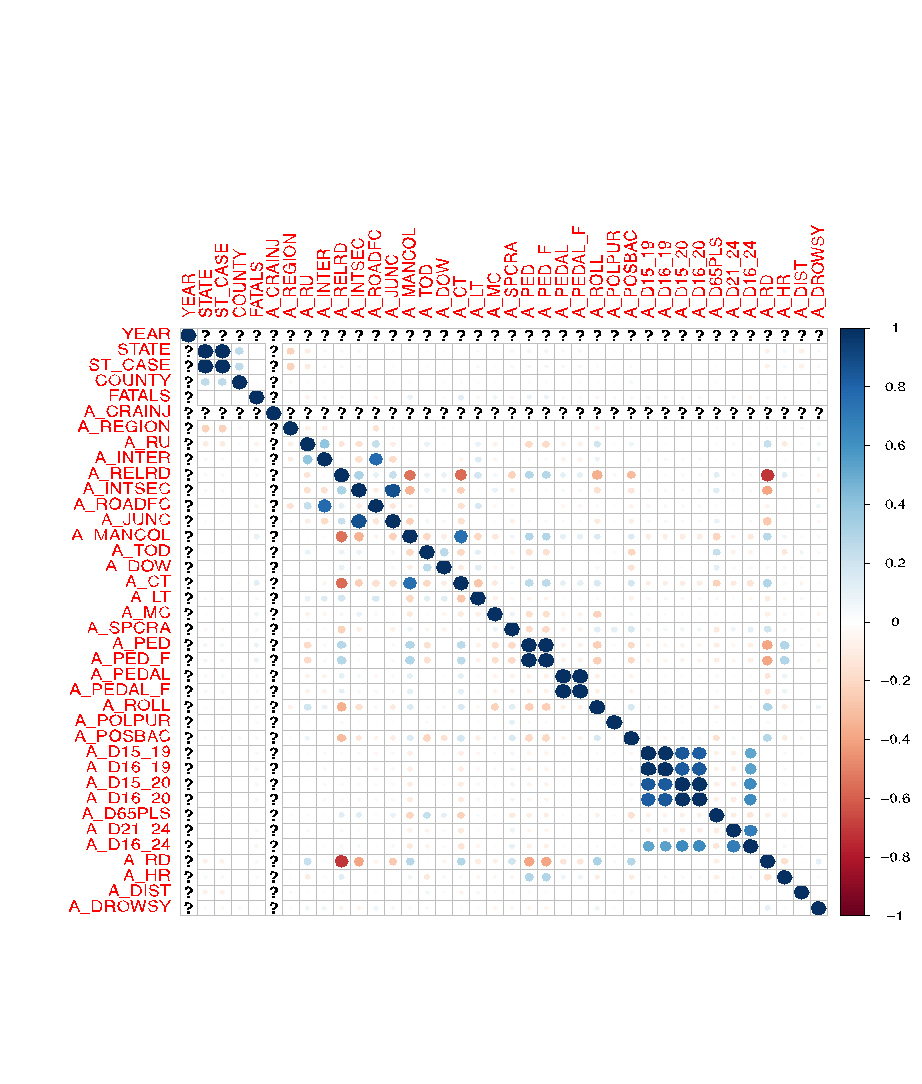
Even though I didn’t have to perform rigorous data cleaning on this particular dataset, I would like to note a few techniques I could use in the future for data cleaning/tidying needs. For example, for I could replace missing data/null values in a continuous variable with the mean or median of that variable. I could also group categories without many cases from a discrete variable into fewer categories. If the data did not come in a column format, I could reshape the data, to o make new variables by uniting variables, as well as many other things with the dplyr and tidyr packages in R.

DATA EXPLORATION

Correlations

I performed correlations on all the variables in each dataset of the three datasets using the corrplot package in R.  Since I was interested in exploring specific uncommon outcomes like distracted driving and drowsy driving, for example, the correlation plots provided me with initial ideas about which variables I might want to include as part of other initial data exploration, such as bar plots, and eventually modelling.

Correlations

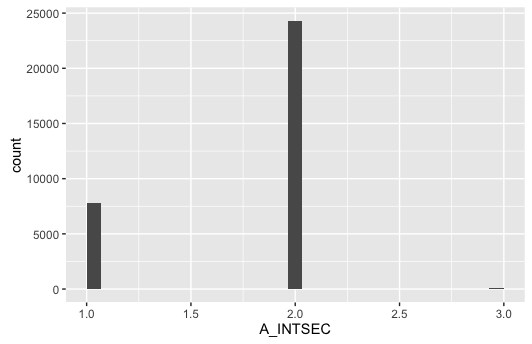
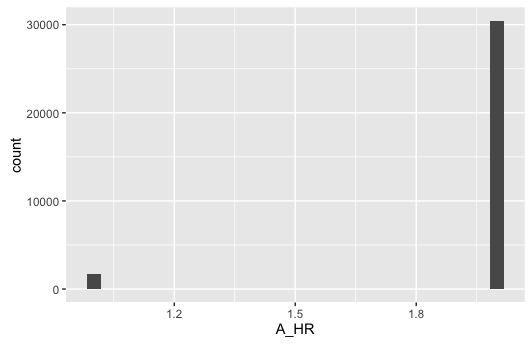
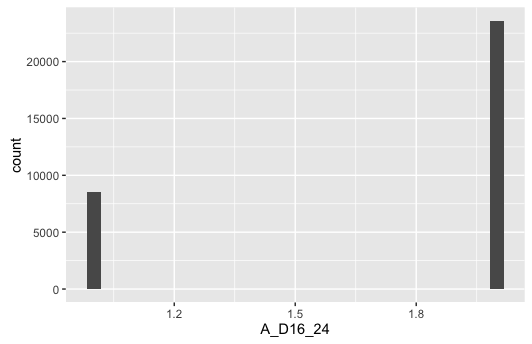


From this correlation matrix initially I see that not many variables are strongly correlated with distracted driving. However, I can still see that state is positively correlated with distracted driving fatalities and time of day is negatively correlated with distracted driving fatalities. This shows me that there is at least some variation in distracted driving fatalities that I want to explore further.

Histograms

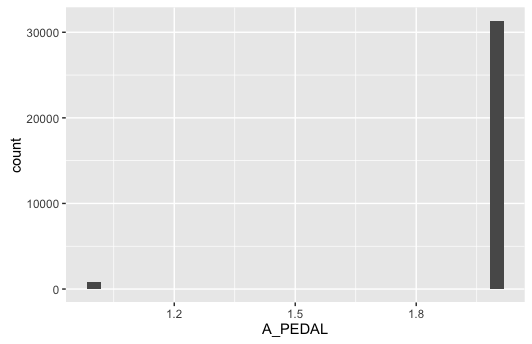
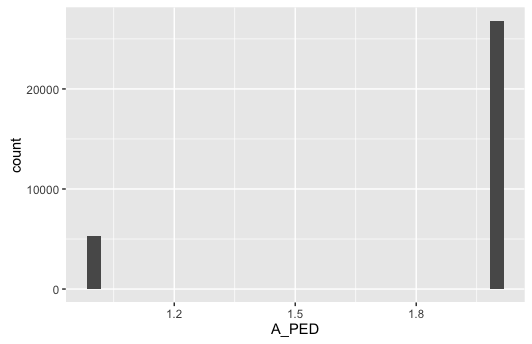
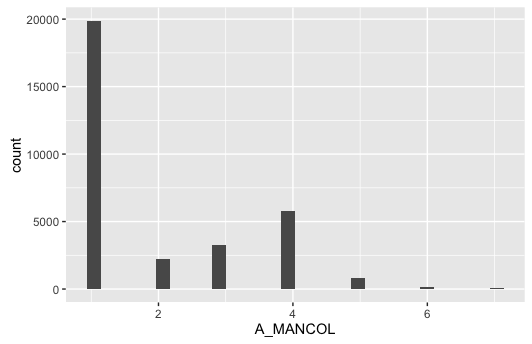
I ran standard histograms on all variables to see the distribution of each variable. Here I present a selected group of variables.

Age under 25 Hit and Run Intersection



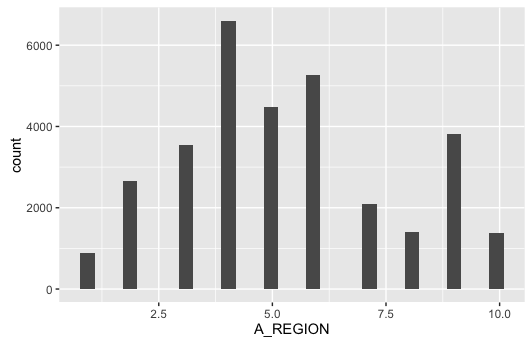
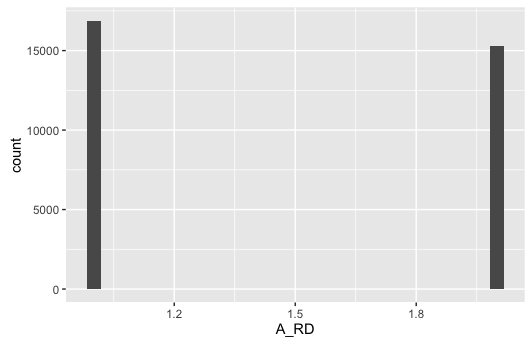
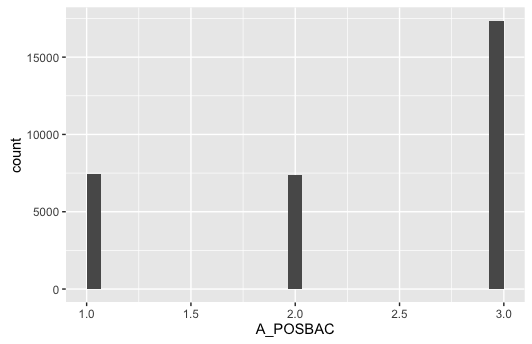
From the three histograms shown above, it clear that a relatively large proportion of fatal accidents in 2015 involve drivers under 25 years of age, as well as intersections. From this I want to investigate whether distracted driving fatalities were more common in younger drivers. I would also like to see whether distracted driving fatalities occur at intersections.

Manor of Collison Involving a Pedestrian Involving a Cyclist



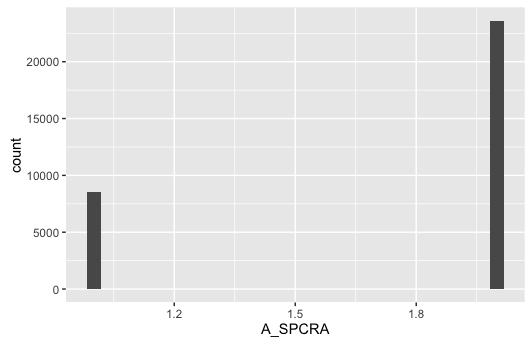
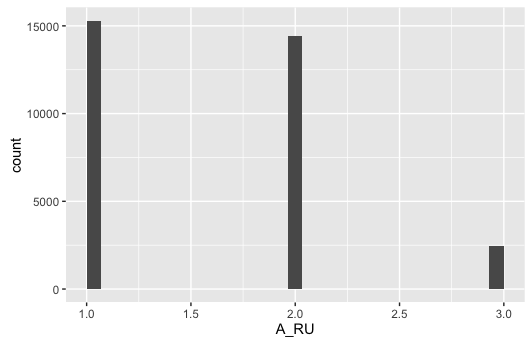
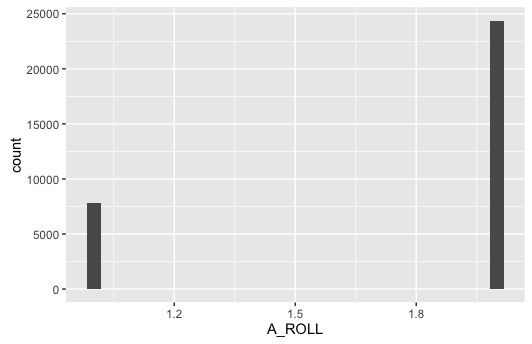
Manor of collision is a very interesting variable. I was somewhat surprised to note that most fatal accidents do not involve another vehicle, that is, they are single vehicle accidents. Moreover, rear-end, head-on, angle, and sideswipe collisions accounted for thousands of accidents. I suspect distracted driving collisions are included here. I am also interested to explore whether distracted drivers are more likely to hit and kill pedestrians and cyclists.

Involved alcohol Roadway departure     Region



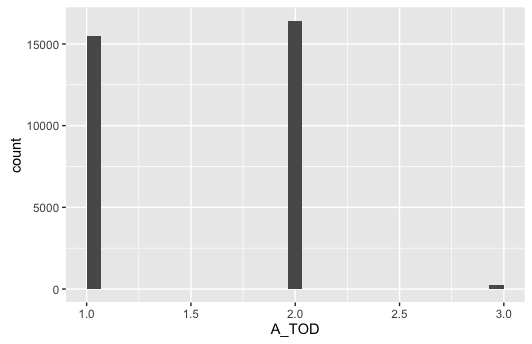
I am surprised by the number of accidents that involve alcohol. I am also surprised at the number of accidents where the influence of alcohol is unknown. Roadway departure is also very common in fatal accidents, while fatal accidents appear to vary across region, with the most fatal accidents occurring in the southern states of Alabama, Florida, Georgia, South Carolina, and Tennessee.

Rollover Rural or Urban Involving speeding



Vehicle Rollover and speeding also appear to be involved in a significant number of fatal accidents in 2015. It would appear that fatal traffic accidents happened in urban and rural areas at approximately the same rate. The same is true for daytime and nighttime fatal accidents (see below)

Time of Day

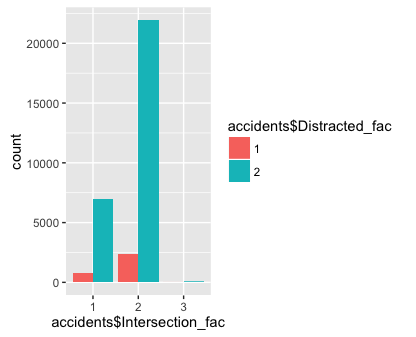
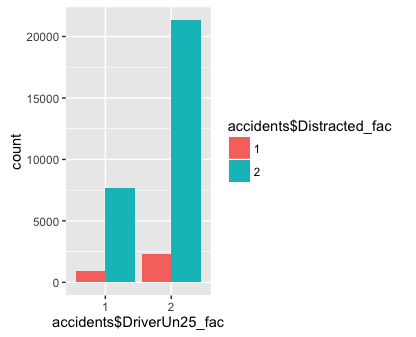
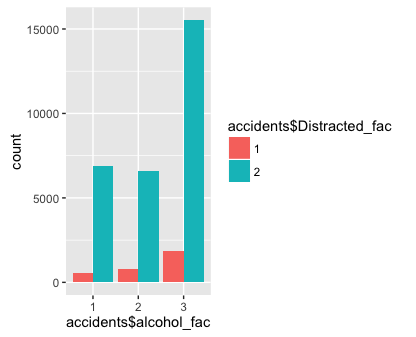


Bar plots with 2 variables

I also produced some bar plots to examine the relationship between key variables and potential outcome variables. Here I’ve included examples of some bar plots with the distracted driver variable.

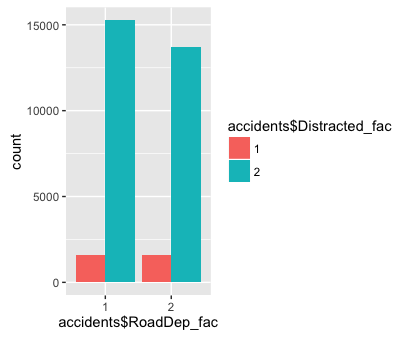
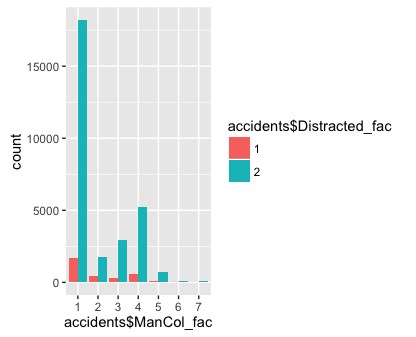
Bar plots by Distracted Driver

Alcohol Age under 25 Intersection



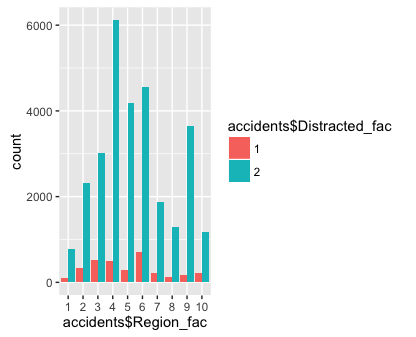
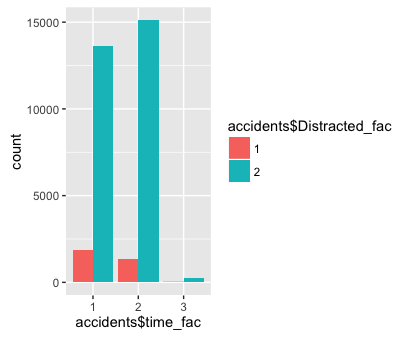
From these bar plots we begin to see some interesting relationships. Based solely on raw counts, distracted driving related traffic fatalities were less likely to involve alcohol or an intersection, and more likely to involve drivers age 25 and older. This part of about age is counter-intuitive to my initial thought that perhaps younger people were texting and driving. Perhaps it’s actually older drivers!

Manor or collision Roadway Departure



Distracted driver fatalities happened across all manors of collision and equally involved a roadway departure vs not.

Time of Day Region



Finally, it is interesting to note that more distracted driver fatalities occurred during the daytime as opposed to nighttime. Distracted driving fatalities varied widely across region, with the most taking place in region category 6, which includes Louisiana, Mississippi, New Mexico, Oklahoma, and Texas.

**APPROACH BASED ON EXPLORATORY ANALYSIS**

Initially I had planned to examine many reasons for the increase in traffic fatalities in 2015. I will build a few different logistic regression models using different outcomes to explore the data. However, at this point, I have narrowed my project to exploring traffic fatalities involving distracted drivers.

I will focus on building a logistic regression model that predicts the likelihood of a traffic fatality involving a distracted driver.

I will also build a random forest model that predicts the occurrence of a traffic fatality involving a distracted driver so that I can compare the performance of both types of models.

I will compare the performance as well as the accuracy of both models.