US Vehicular Accidents

Kirsten Miller

5/4/2020

Table of Contents

[Introduction 2](#_Toc39596377)

[Methods 3](#_Toc39596378)

[Results 3](#_Toc39596379)

[Number of People Involved In Vehicular Accidents 3](#_Toc39596380)

[Severity of Injury Sustained 14](#_Toc39596381)

[Age of People Involved In Vehicular Accidents 19](#_Toc39596382)

[Age By Group: Alcohol Involvement & Injury Severity 25](#_Toc39596383)

[Discussion 29](#_Toc39596384)

[References 30](#_Toc39596385)

# Load packages  
library(dplyr)  
library(tidyr)  
library(lubridate)  
library(ggplot2)  
library(forcats)  
library(lattice)

# Load data  
dat <- readRDS("../Data/farsp.RDS")  
  
# Load files containing column descriptions  
states\_df <- read.csv("../Data/state\_code.csv") %>%   
 rename(state = state\_code)  
alcohol\_df <- read.csv("../Data/alcohol.csv") %>%   
 rename(drinking = alcohol,  
 drinking\_desc = description)  
body\_type\_df <- read.csv("../Data/body\_typ.csv") %>%   
 rename(body\_type\_desc = description)  
inj\_sev\_df <- read.csv("../Data/inj\_sev.csv") %>%   
 rename(inj\_sev\_desc = description)  
man\_coll\_df <- read.csv("../Data/man\_coll.csv") %>%   
 rename(man\_coll\_desc = description)  
sex\_df <- read.csv("../Data/sex.csv") %>%   
 rename(sex\_desc = description)  
  
# create date object, month column, day of week column  
dat2 <- dat %>%   
 mutate(Date = ymd\_hm(paste(year, month, day, hour, minute)),  
 Mon = month.abb[month],  
 Day = wday(Date, label = TRUE, abbr = TRUE))  
  
# qdd column descriptions to data  
# state -join tables  
dat3 <- left\_join(dat2, states\_df, by= "state")  
rm(dat2)  
# drinking -join tables  
dat4 <- left\_join(dat3, alcohol\_df, by= "drinking")  
rm(dat3)  
# body type -join tables  
dat5 <- left\_join(dat4, body\_type\_df, by= "body\_typ") # remove NAs?  
rm(dat4)  
# injury severity join tables  
dat6 <- left\_join(dat5, inj\_sev\_df, by= "inj\_sev")  
rm(dat5)  
# manner of collision  
dat7 <- left\_join(dat6, man\_coll\_df, by= "man\_coll") # remove NAs?  
rm(dat6)  
# sex  
dat8 <- left\_join(dat7, sex\_df, by= "sex")   
rm(dat7)  
  
# remove old dataframes  
rm(states\_df, alcohol\_df, body\_type\_df, inj\_sev\_df, man\_coll\_df, sex\_df)

## Introduction

I investigated vehicular accidents in the United States over a 21-year period (1996 to 2016). There are many potential ways to explore this dataset, so I tried to understand it broadly while also focusing on some more specific aspects and potential relationships. I first explored the data broadly by visualizing the number of people involved in vehicular accidents over varying timescales. I then looked at the breakdown of the number of people involved in accidents by many of the variables included to get a sense of the dataset. After this more general visualization, I attempted to uncover some of the relationships between specific variables and the number of people involved in accidents. In particular, I focused on injury severity, age, and alcohol involvement. I also focused on local states in portions of my analysis (Maine, New Hampshire, Vermont, and Massachusetts).

Some questions that guided this data exploration include:

\*How has the number of vehicular accidents changed over different timescales (years, months, hours)?

\*What is the breakdown of types of injury severity and has this changed over time?

\*Are there significant differences between the ages of people involved in certain types of accidents?

While answering these questions, I also considered:

\*How do vehicular accidents vary among regions of the country and local states and how does this compare to the United States as a whole?

\*Could involvement of alcohol be a predictor for certain types of accidents?

## Methods

This data was collected by the NHTSA for the years 1996 through 2016. The data include the following information for each person involved in a vehicular accident: state, county, month, day, year, hour, minute, manner of collision, number of vehicles involved, type of vehicle involved, number of people involved, age of driver, sex of driver, involvement of alcohol, and severity of injury.

An important note is that the data are broken down by individuals involved in vehicular accidents, not the unique accidents themselves. Therefore, my analysis is based on the number of inviduals involved in vehicular accidents, not the number of vehicular accidents.

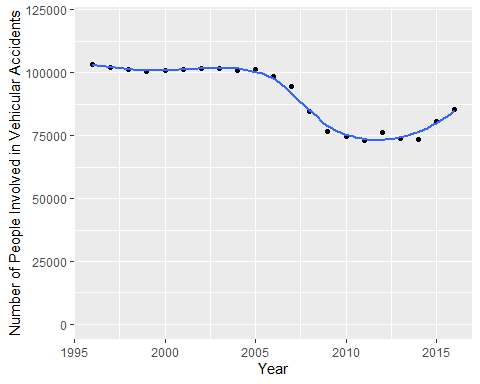
## Results

### Number of People Involved In Vehicular Accidents

First, I examined the number of people involved in vehicular accidents in the US over the entire 20 year time period for the data (1996 - 2016). I used a loess fit and found that a second-degree loess fit was best.

#### Number of people involved in vehicular accidents in the US over a 21-year period

# group data by year  
dat\_us <- dat8 %>%   
 group\_by(year) %>%   
 count()  
# plot number over time with loess fit  
ggplot(dat\_us, aes(year, n)) + geom\_point() + ylim(0, 120000) +   
 stat\_smooth(method = "loess", se = FALSE, span = 0.5, method.args = list(degree = 2)) +  
 ylab("Number of People Involved in Vehicular Accidents") + xlab("Year")

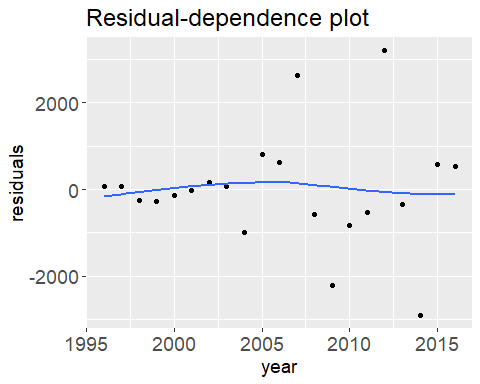


The number of people involved in vehicular accidents was fairly constant (around 100,000 accidents) over the first 10 years of the dataset from 1995 to 2005. This was followed by a decrease in the number of accidents to a low number of 73073 in 2011, followed by a recent increase to 85496 accidents in 2016. The data is fitted with a loess degree 2 fit.

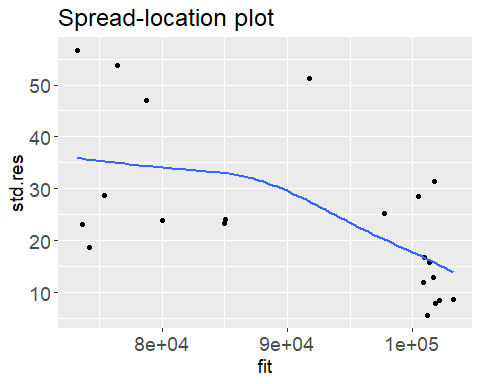
#### Analytical plots for residuals

The loess fit does approximate a horizontal line; however, there appears to be a potential fanning pattern of the residuals (the points become more scattered when moving from left to right) seen in the residual-dependence plot. The spread-location plot shows that there is somewhat of a systematic descreasing trend in the residuals. I then checked the residuals for normality by comparing to the normal distribution. The residuals align somewhat well with a theoretic distribution, although there may be some level of skew to the left.

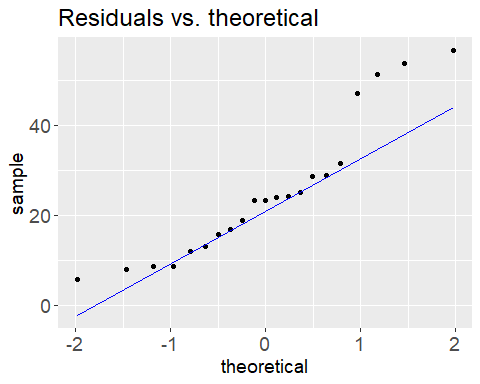
##### Residual-dependence plot  
# calculate residuals for the loess fit  
lo <- loess(n ~ year, dat\_us, span = 0.5, method.args = list(degree = 2))  
dat\_us$residuals <- residuals(lo)  
# plot the residuals for the loess fit  
ggplot(dat\_us, aes(x = year, y = residuals)) + geom\_point() +  
 stat\_smooth(method = "loess", se = FALSE, span = 1,   
 method.args = list(degree = 2) ) + ggtitle("Residual-dependence plot") + theme(axis.title=element\_text(size=14), plot.title = element\_text(size = 18), axis.text = element\_text(size =14))



# Spread-location plot  
# create dataframe for spread-location plot  
sl2 <- data.frame( std.res = sqrt(abs(residuals(lo))),   
 fit = predict(lo))  
# create spread-location plot  
ggplot(sl2, aes(x = fit, y =std.res)) + geom\_point() +  
 stat\_smooth(method = "loess", se = FALSE, span = 1,   
 method.args = list(degree = 1) ) + ggtitle("Spread-location plot") + theme(axis.title=element\_text(size=14), plot.title = element\_text(size = 18), axis.text = element\_text(size =14))

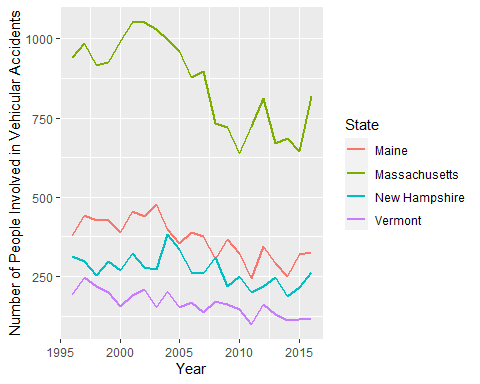


# Checking residuals for normality  
# check residuals for normality by comparing to normal distribution  
ggplot(sl2, aes(sample = std.res)) + geom\_qq(distribution = qnorm) +  
 geom\_qq\_line(distribution = qnorm, col = "blue") + ggtitle("Residuals vs. theoretical") + theme(axis.title=element\_text(size=14), plot.title = element\_text(size = 18), axis.text = element\_text(size =14))



#### Number of people involved in accidents over time in local states

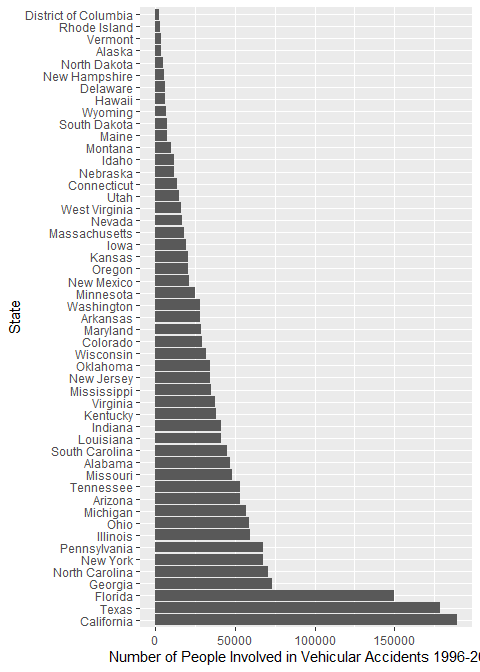
# create dataframe for local states  
dat\_local\_states <- dat8 %>%   
 filter(state\_name == "Maine" | state\_name == "New Hampshire" |   
 state\_name == "Vermont" | state\_name == "Massachusetts") %>%   
 group\_by(state\_name, year) %>%   
 count()  
  
# plot number over time for local states  
ggplot(dat\_local\_states, aes(x=year, y=n, color=state\_name)) + geom\_line(size = 1) +  
 ylab("Number of People Involved in Vehicular Accidents") + xlab("Year") + labs(color = "State")



The number of people involved in accidents over time in local states appears to follow a similar pattern to the overall trend for the United States. However, with smaller numbers of people involved in accidents in these states, the year-to-year variability is more apparent. This also indicates the varying number of people involved in accidents by state, which is examined further below:

#### Cumulative number of people involved in accidents in each state from 1996 - 2016

# bar chart for number of people involved in accidents in each state  
ggplot(dat8, aes(fct\_infreq(state\_name, ordered = TRUE))) + geom\_bar() +  
 xlab("State") + ylab("Number of People Involved in Vehicular Accidents 1996-2016") + coord\_flip()

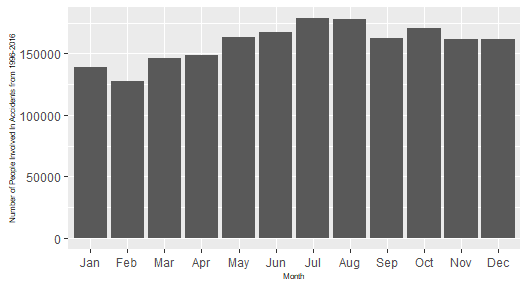


The number of accidents in each state varies widely. The top 5 states with the greatest cumulative number of accidents over the 21-year period were California, Texas, Florida, Georgia, and North Carolina, while the 5 states with the least number of accidents were the District of Columbia, Rhode Island, Vermont, Alaska and North Dakota.

To further examine how the number of accidents change over different timescales, I looked at the distribution of the total number of people involved in accidents over each month of the year.

#### Number of people involved in accidents per month of the year

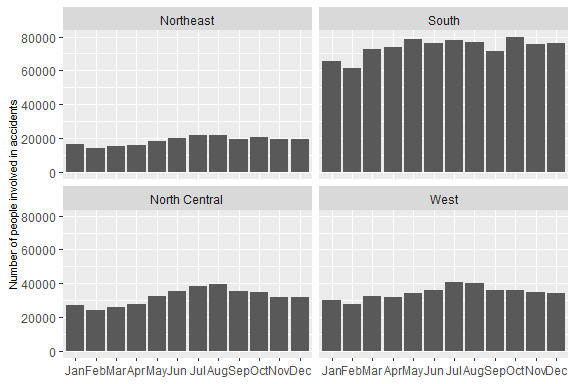
# set levels for months  
dat8$Mon <- factor(dat8$Mon, levels=c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"))  
# bar chart for n per month  
ggplot(dat8, aes(Mon)) +geom\_bar() +   
 ylab("Number of People Involved In Accidents from 1996-2016") + xlab("Month") + theme(axis.title=element\_text(size=6))



I was surprised to see that the cumulative number of people involved in accidents was higher during the summer months (with the highest number occuring in July and the lowest number occuring in February). I would have predicted that there would be a greater number of accidents during the winter because of more weather events and difficult driving conditions. To examine this further, I looked at the number people involved in accidents each month by region (assuming that northern regions would experience more weather that could impact driving in the winter).

#### Number of people involved in accidents per month by region

# add region to   
st.reg <- data.frame(state\_name = state.name, Region = state.region)  
st.reg <- rbind(st.reg, data.frame(state\_name="District of Columbia", Region="South") )  
# add region to datframe  
dat\_region <- left\_join(dat8, st.reg, by= "state\_name")  
# bar plot for n per month faceted by region  
ggplot(dat\_region, aes(Mon)) +geom\_bar() + facet\_wrap(~Region) +   
 ylab("Number of people involved in accidents") + xlab(NULL) + theme(axis.title=element\_text(size=8))

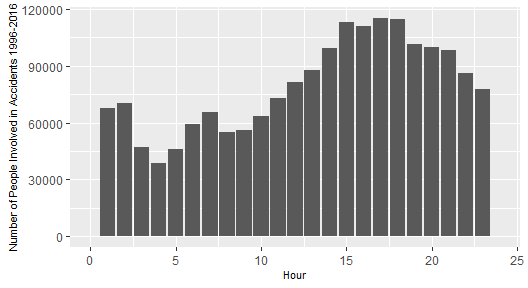


Here, I was again surprised to see that the number of people involved in accidents was still higher in the summer in both the Northeast and North Central regions. However, the greater number people involved in accidents during the summer could be due to increased travel during these months. It was interesting to see that many more people were involved in accidents overall in the South Region (althought this is likely because this region has a higher population).

Next, I looked at the distribution of the total number of people involved in accidents per hour of the day:

#### Number of people involved in accidents per hour of the day

# bar chart for n per hour of the day  
ggplot(dat8, aes(hour)) +geom\_bar() + xlim(0,24) +   
 ylab("Number of People Involved in Accidents 1996-2016") + xlab("Hour") + theme(axis.title=element\_text(size=8))

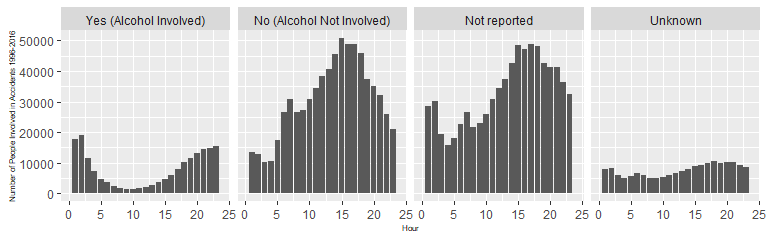


This shows the most people were involved in accidents that occurred around 17:00, which is intuitive, as this is around the time when many people return home from work, as well as when it may be getting dark. It is also interesting to note that there is a small peak in the observed distribution at 6:00-7:00, which might be explained by the morning commute.

I was interested to see how alcohol involvment might be related to this distribution:

#### Number of people involved in accidents during each hour of the day, by alcohol involvement

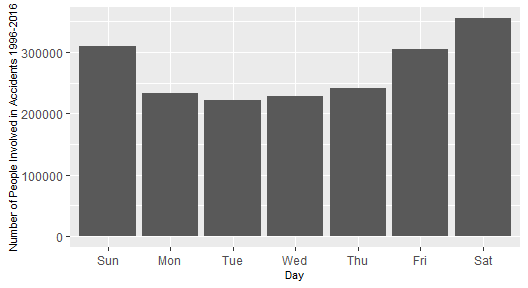
# set levels for alcohol involvment  
dat8$drinking\_desc <- factor(dat8$drinking\_desc, levels=c("Yes (Alcohol Involved)" ,"No (Alcohol Not Involved)", "Not reported", "Unknown"))  
# bar chart for n per hour of the day faceted by alcohol involvment  
ggplot(dat8, aes(hour)) +geom\_bar() + xlim(0,24) + facet\_wrap(~drinking\_desc, nrow = 1) +   
 ylab("Number of People Involved in Accidents 1996-2016") + xlab("Hour") +  
 theme(axis.title=element\_text(size=6))



When alcohol was not involved, not reported, or unknown, the distribution of the number of people involved in accidents per hour is similar to the overall distribution. However, when alcohol is involved, the distribution is near opposite, with the peak of accidents occuring during the nightime hours.

#### Number of people involved in vehicular accidents by day of the week

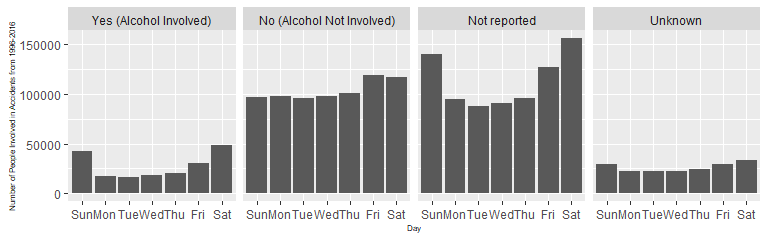
# set scipen so figures display not in scientific notation  
options(scipen=1000000)  
# remove NA values from day of the week column for plot  
dat\_day <- dat8 %>%   
 filter(Day != "N/A")  
# bar plot for n per day of the week  
ggplot(dat\_day, aes(Day, na.rm =FALSE)) +geom\_bar() + ylab("Number of People Involved in Accidents 1996-2016") + theme(axis.title=element\_text(size=8))



Finally, I examined the number of people involved in vehicular accidents by day of the week. The most people were involved in accidents that occured on Friday and the weekend, while less people were involved in accidents that occured during the week. The greatest number of people were involved in accidents on Saturdays (356068) whereas the least number of people were involved in accidents that occured on Tuesdays (221969).

#### Number of people involved in vehicular accidents by day of the week, by alcohol involvement:

# bar plot for n per day of the week, faceted by alcohol involvment  
ggplot(dat\_day, aes(Day)) +geom\_bar() + facet\_wrap(~drinking\_desc, nrow =1) + ylab("Number of People Involved in Accidents from 1996-2016") + theme(axis.title=element\_text(size=6))



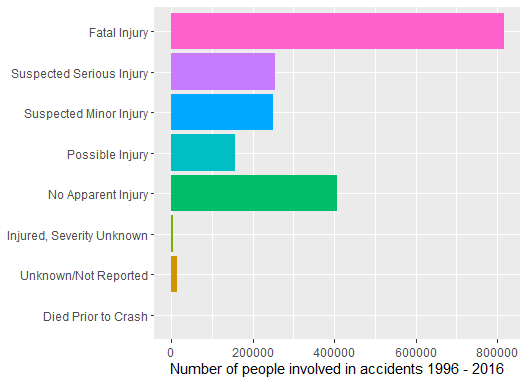
Again, I also looked at whether this breakdown by day of the week changed at all with alcohol involvement. When alcohol was involved (“Yes”), the distribution appears to be similar to the overall distribution, although the number of people involved in accidents on the weekend appears to be proportionally higher. When alcohol was not involved (“No”), the number of people involved in accidents is fairly consistent from Sunday through Thursday, and slightly larger on Friday and Saturday.

### Severity of Injury Sustained

After an initial investigation of some more of the dataset variables, (including vehicle type, manner of collision, and number of people involved), I decided that I was most interested in looking at injury severity. Below is the overall breakdown of injury severity for the cumulative number of people involved in accidents:

#### Injury Severity

# breakdown of injury types  
dat\_injury <- dat8 %>%   
 filter(inj\_sev\_desc != "N/A")  
# set levels for injury severity  
dat\_injury$inj\_sev\_desc <- factor(dat\_injury$inj\_sev\_desc, levels=c("Died Prior to Crash", "Unknown/Not Reported","Injured, Severity Unknown", "No Apparent Injury", "Possible Injury",   
"Suspected Minor Injury", "Suspected Serious Injury", "Fatal Injury", "N/A" ))  
# bar plot for injury severity  
ggplot(dat\_injury, aes(x= inj\_sev\_desc, fill = inj\_sev\_desc)) + geom\_bar() + coord\_flip() + xlab(NULL) + ylab("Number of people involved in accidents 1996 - 2016") + theme(legend.position = "none")

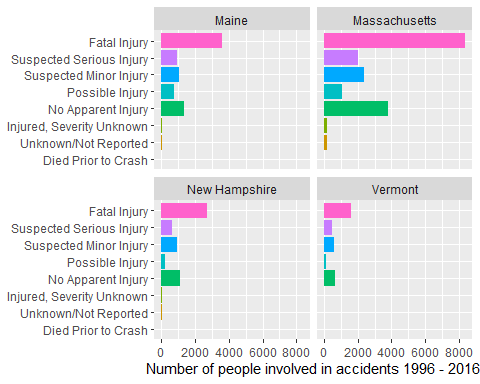


Out of all the the people involved in accidents, fatal injuries makes up the largest category of injury level, with 815926 total fatalities recorded in this dataset. This is over twice as many as those who were reported to have no apparent injury. Those with possible, minor, or serious injuries also made up significant portions of the total number of people involved in accidents.

I also was curious to see if there were any evident differences in injury severity in local states, shown below:

#### Severity of injury by state

# create dataframe with local states only (from injury dataframe)  
dat\_injury\_local <- dat\_injury %>%   
 filter(state\_name == "Maine" | state\_name == "New Hampshire" |   
 state\_name == "Vermont" | state\_name == "Massachusetts")   
# bar plots for injury severity faceted by state  
ggplot(dat\_injury\_local, aes(x= inj\_sev\_desc, fill = inj\_sev\_desc)) + facet\_wrap(~state\_name) +  
 geom\_bar() + coord\_flip() + xlab(NULL) + ylab("Number of people involved in accidents 1996 - 2016") + theme(legend.position = "none")

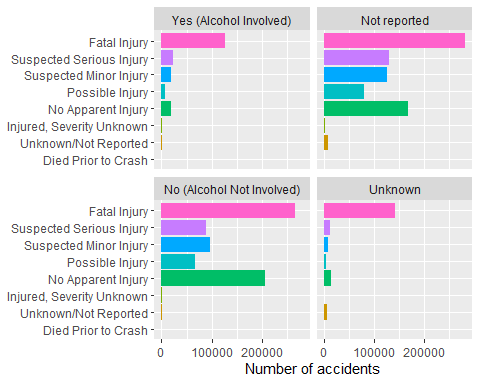


The breakdown of the categories of injury severity appears to be fairly consistent across these states, and fairly consistent with the overall breakdown for the United States above.

Next, I looked at injury severity by alcohol involvement:

#### Severity of injury by alcohol involvment

# set levels for alcohol involvement  
dat\_injury$drinking\_desc <- factor(dat\_injury$drinking\_desc, levels=c("Yes (Alcohol Involved)" , "Not reported","No (Alcohol Not Involved)", "Unknown"))  
# bar plots for injury severity faceted by alcohol involvement  
ggplot(dat\_injury, aes(x= inj\_sev\_desc, fill = inj\_sev\_desc)) +   
 facet\_wrap(~drinking\_desc) + geom\_bar() + coord\_flip() + xlab(NULL) + ylab("Number of accidents") +  
 theme(legend.position = "none")

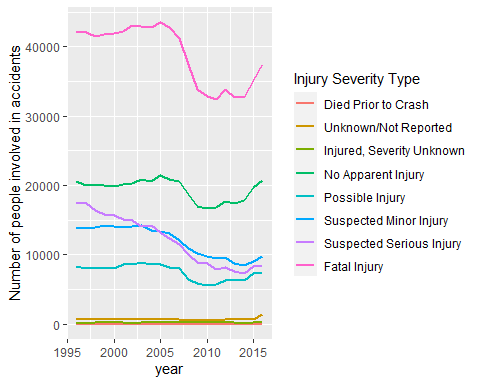


When comparing the accidents in which it was known whether alcohol was involved (“Yes”) or not involved (“No”), it appears that the percentage of fatal injuries was much higher when alcohol was involved. This is especially evident when comparing the magnitude of the “Fatal Injury” category to the “No Apparent Injury” category.

I also looked at how the different types of injury had changed over time:

#### Severity of injury over time

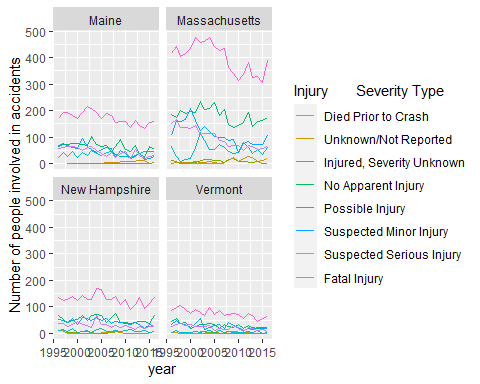
# dataframe grouping injury levels by year  
dat\_injury2 <- dat\_injury %>%   
 group\_by(inj\_sev\_desc, year) %>%   
 count()  
# plot of injury severity levels over time  
ggplot(dat\_injury2, aes(x=year, y=n, color=inj\_sev\_desc)) + geom\_line(size =1) +  
 ylab("Number of people involved in accidents") + labs(color = "Injury Severity Type")



All types of injuries appear to have decreased over time, although after more substaintial decreases between 2005 and 2010, injuries (as well as the no injury category) have increased again in the last 5 years of the dataset. Recalling the trend for overall number of people involved in accidents (the first figure in the Results), this decrease followed by a recent increase is very similar.

#### Severity of injury over time in local states

# selecting local states only  
dat\_injury\_local2 <- dat\_injury\_local %>%   
 group\_by(state\_name, inj\_sev\_desc, year) %>%   
 count()  
# plot of injury severity levels over time faceted by state  
ggplot(dat\_injury\_local2, aes(x=year, y=n, color=inj\_sev\_desc)) + geom\_line() +  
 ylab("Number of people involved in accidents") + facet\_wrap(~state\_name) + labs(color = "Injury Severity Type")



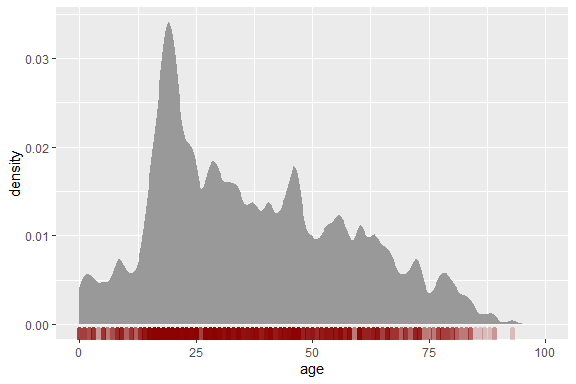
Again, the patterns over time in local states were relatively consistent and similar to the overall trend.

### Age of People Involved In Vehicular Accidents

I was also interested in investigating the distribution of the age of people involved in vehicular accidents. I first looked at the distribution of the ages of people involved in vehicular accidents, and whether their average age has changed over time.

#### Density plot for age of people involved in vehicular accidents

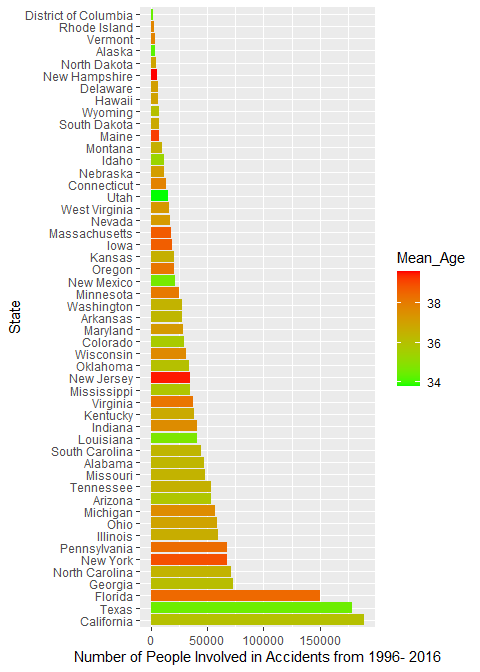
# filter NAs from age  
dat\_age <- dat8 %>%   
 filter(age < 200) %>%   
 filter((age != 99) & sex !=9) # removed additional NA values for age  
# density plot of age  
ggplot(sample\_n(dat\_age, 1000), aes(x = age)) + geom\_density(fill = "grey60", col = NA, bw = 1) +  
 geom\_rug(col = "darkred", size = 2, alpha = 0.2) + xlim(0,100)



This density plot shows the distribution of the age of people involved in vehicular accidents. The median age of people involved in vehicular accidents over this time period was 33. However, looking at the density plot for age, the peak density occurs in the low 20’s. This is followed by a somewhat consistent density level from age 25 to age 50, followed by a decrease in density as age continues to increase.

#### Number of people involved in accidents for each state and mean age

# calculate median age for each state  
dat\_med\_age <- dat\_age %>% group\_by(state\_name) %>%   
 summarize(Mean\_Age = mean(age))  
dat\_med\_age2 <- left\_join(dat8, dat\_med\_age, by = "state\_name")  
  
# add median age to plot of n for each state  
ggplot(dat\_med\_age2, aes(x= fct\_infreq(state\_name, ordered = TRUE), fill= Mean\_Age)) + geom\_bar() +  
 scale\_fill\_gradient(low = "green", high = "red") +  
 xlab("State") + ylab("Number of People Involved in Accidents from 1996- 2016") + coord\_flip()

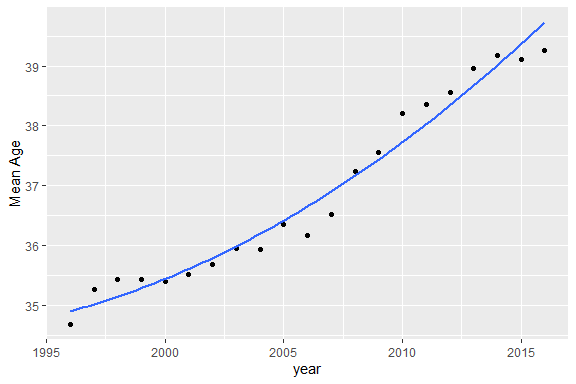
 The plot above displays the number of people involved in accidents in each state, as well as the mean age in each state. There does not appear to be a relationship between the mean age and the number of people involved in accidents in each state. The mean age of people involved in accidents does vary noticably among the states, which suggests that it might be related to the mean age of the overall population.

Next, I examined whether there has been a change in the age of people involved in vehicular accidents over time by plotting the mean age by year:

#### Mean Age of People Involved In Vehicular Accidents Over Time

# calculate mean age by year  
dat\_age\_mean<- dat\_age %>%   
 group\_by(year) %>%   
 summarise(Mean = mean(age))

# create polynomial fit model  
M2 <- lm(Mean ~ year + I(year^2) , dat = dat\_age\_mean)  
  
# plot polynomial fit residuals  
ggplot(dat\_age\_mean, aes(x = year, y = Mean)) + geom\_point() +   
 stat\_smooth(method = "lm", se = FALSE, formula = y ~ x + I(x^2) ) + ylab("Mean Age")

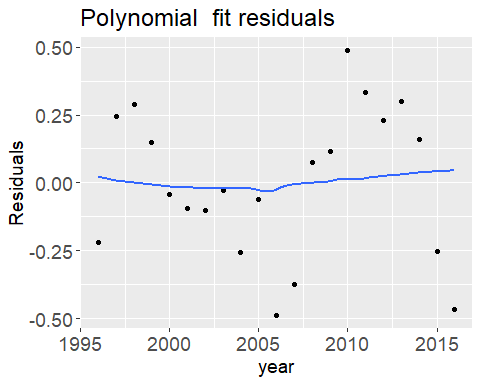


The mean age of people involved in vehicular accidents increased over time from 34.6877506 in 1996 to 39.2625147 in 2016. Although there was some inconsistency in the residuals, the data was best approximated by second degree polynomial fit, defining the relationship as follows:

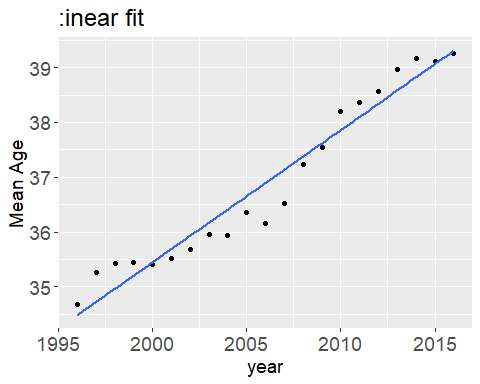
It could be also be approximated with a linear fit as follows (although the residual values were less consistent):

#### Analytical plots for fits and residuals

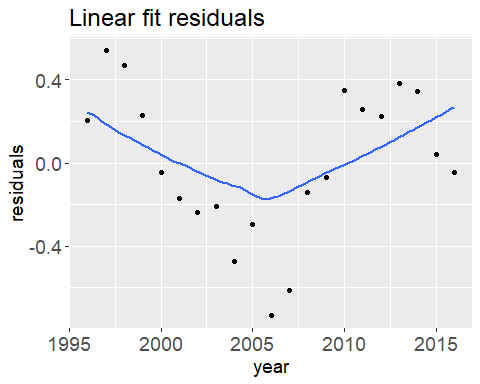
# residuals for polynomial fit  
dat\_age\_mean$residuals2 <- residuals(M2)  
# plot residuals for polynomial fit  
ggplot(dat\_age\_mean, aes(x = year, y = residuals2)) + geom\_point() +  
 stat\_smooth(method = "loess", se = FALSE, span = 1,   
 method.args = list(degree = 1) ) + ylab("Residuals") +   
 ggtitle("Polynomial fit residuals") + theme(axis.title=element\_text(size=14), plot.title = element\_text(size = 18), axis.text = element\_text(size =14))



# plot mean age by year with linear fit  
ggplot(dat\_age\_mean, aes(x = year, y = Mean)) + geom\_point() +   
 stat\_smooth(method ="lm", se = FALSE) + ylab("Mean Age") + ggtitle(":inear fit") + theme(axis.title=element\_text(size=14), plot.title = element\_text(size = 18), axis.text = element\_text(size =14))



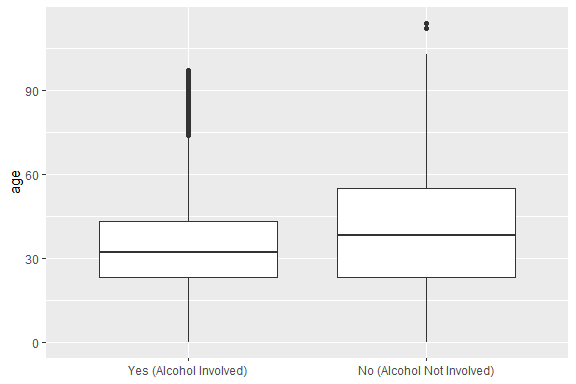
# create linear fit model  
M <- lm(Mean ~ year, dat = dat\_age\_mean)  
  
# plot linear fit residuals  
dat\_age\_mean$residuals <- residuals(M)  
ggplot(dat\_age\_mean, aes(x = year, y = residuals)) + geom\_point() +  
 stat\_smooth(method = "loess", se = FALSE, span = 1,   
 method.args = list(degree = 1) ) + ggtitle("Linear fit residuals") + theme(axis.title=element\_text(size=14), plot.title = element\_text(size = 18), axis.text = element\_text(size =14))



### Age By Group: Alcohol Involvement & Injury Severity

#### Age of people involved in accidents, by alcohol involvement

# set levels for alcohol involvment in new dataframe for age  
dat\_age\_alc <- dat\_age %>%   
 filter(drinking\_desc == "Yes (Alcohol Involved)" | drinking\_desc == "No (Alcohol Not Involved)") %>%  
 group\_by(drinking\_desc)  
  
# boxplot for alcohol involvment and age  
ggplot(dat\_age\_alc, aes(y=age, x=drinking\_desc)) +   
 geom\_boxplot() + xlab(NULL)



# summary table for boxplot  
age\_alc\_table <- dat\_age\_alc %>% summarise(median = median(age), count = n() )

##### T-Test for difference in age between alcohol involved or alcohol not involved groups

# create values for alcohol involved/not involved groups  
alc\_involved <- filter(dat\_age\_alc, drinking\_desc == "Yes (Alcohol Involved)") %>% pull(age)  
alc\_notinvolved <- filter(dat\_age\_alc, drinking\_desc == "No (Alcohol Not Involved)") %>% pull(age)  
# t-test for age of alcohol involved vs. age of alcohol not involved  
t.test(alc\_involved, alc\_notinvolved, alt = "two.sided")

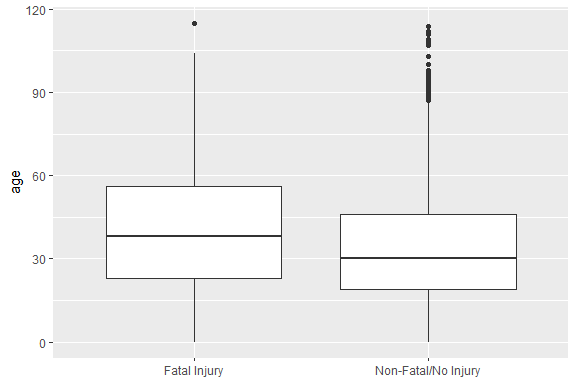
##   
## Welch Two Sample t-test  
##   
## data: alc\_involved and alc\_notinvolved  
## t = -154.67, df = 465241, p-value < 0.00000000000000022  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -6.069848 -5.917936  
## sample estimates:  
## mean of x mean of y   
## 34.53314 40.52703

The median age of those involved in vehicle accidents when alcohol was involved was 32, and when alcohol was not involved the median age was 38. The t-test above compared the two groups and found that there was a significant difference between the age of those involved in vehicle accidents when alcohol was involved and when it was not involved (p < 0.001). This suggests that the age of people involved in vehicle accidents with alochol is significantly lower than the age of people involved in accidents without alcohol.

Next, I investigated whether there was a significant difference between the age of people involved in accidents who experienced fatal or nonfatal injuries.

#### Age of people who suffered fatal injuries versus non-fatal/no injuries

# create new dataframe and set new levels of fatal/non-fatal inuries  
dat\_injury3 <- dat\_injury%>%   
 filter(inj\_sev\_desc != "Died Prior to Crash") %>%   
 mutate(Inj\_fatality = case\_when(inj\_sev\_desc == "Fatal Injury" ~ "Fatal Injury",  
 inj\_sev\_desc == "Unknown/Not Reported" ~ "Unknown/Not Reported",  
 TRUE ~ "Non-Fatal/No Injury")) %>%   
 filter(age < 200) %>%   
 filter((age != 99) & sex !=9) %>% # removed additional NA values for age  
 select(age, Inj\_fatality, state\_name) %>%   
 filter(Inj\_fatality == "Fatal Injury" | Inj\_fatality == "Non-Fatal/No Injury")  
   
  
# create boxplot of age by fatal or non-fatal injury  
ggplot(dat\_injury3, aes(y=age, x=Inj\_fatality)) +   
 geom\_boxplot() + xlab(NULL)

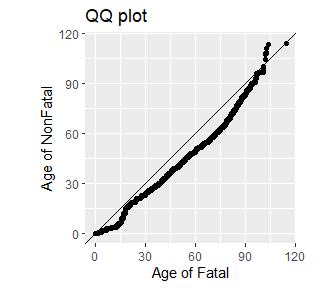


# summary table for fatal vs. non fatal injury  
inj\_table <- dat\_injury3 %>% group\_by(Inj\_fatality) %>% summarise(median = median(age), count = n() )

##### Analytical plot

The qq plot below shows that there is a systematic difference between the age of people who experienced fatal injuries and those who did not, with the age of fatal injuries being higher.

# create values for fatal and non-fatal groups  
Fatal <- filter(dat\_injury3, Inj\_fatality == "Fatal Injury") %>% pull(age)  
NonFatal <- filter(dat\_injury3, Inj\_fatality == "Non-Fatal/No Injury") %>% pull(age)  
# create qq data table  
qq.out <- as.data.frame(qqplot(x=Fatal, y=NonFatal, plot.it=FALSE))  
# set x and y limits  
xylim <- range( c(qq.out$x, qq.out$y) )  
# create qq plot  
ggplot(qq.out, aes( x= x, y = y)) + geom\_point() +   
 geom\_abline( intercept=0, slope=1) +  
 coord\_fixed(ratio = 1, xlim=xylim, ylim = xylim) +  
 xlab("Age of Fatal") + ylab("Age of NonFatal") + ggtitle("QQ plot")



##### T-Test for difference in age between fatal and non-fatal injury severity groups

t.test(Fatal, NonFatal, alt="two.sided")

##   
## Welch Two Sample t-test  
##   
## data: Fatal and NonFatal  
## t = 257.47, df = 1645318, p-value < 0.00000000000000022  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 7.582410 7.698737  
## sample estimates:  
## mean of x mean of y   
## 41.02294 33.38236

The median age of those who suffered fatal injuries was 38, compared to 30 for those who did not experience a fatal injury. The t-test above found that there was a significant difference in the age of the two groups (p < 0.001).

## Discussion

This exploratory analysis investigated various aspects of vehicular accidents in the United States over a 21-year period. The main areas of focus were change over varying timescales, injury severity, age of people involved, and alcohol involvement.

By analyzing this data, I found that the number of people involved in vehicle accidents, as well as the number of fatalities (in the US as a whole and in local states), has decreased over time. However, both measures have increased in the last few years of the data examined. The number of accidents also varies by month of the year, hour of the day, and day of the week; accidents are higher during the summer months, evening hours, and weekend days.

Fatalities make up a significant component of the people who experienced vehicular accidents recorded in this dataset, and proportions of fatal injury as well as change over time of fatal injury are relatively consistent in local states.

The mean age of people involved in vehicular accidents has increased over time. The age of people who experience fatal injuries through vehicular accidents is higher than the age of those who do not.

Accidents involving alcohol are especially prevalent during nighttime hours and on weekend days. Fatal injuries are higher for people in accidents for whom alcohol was involved. The data also suggests that alcohol involvement is more common in younger people involved in accidents.

These results identify certain time periods during which more people are usually involved in accidents, which is helpful for understanding risk, as well as in consideration of how vehicle accidents can be managed and prevented. These findings also suggest that both young people and older people may be at risk for vehicle accidents due to differing reasons: young people may have more accidents related to alcohol use, while older people may be more at risk for fatal injuries if they are to be involved in a vehicular accident.

There are many ways that this data could be explored more to highlight these potential relationships, as well as to uncover further trends and patterns. Perhaps most pertinently, it will be important to understand why the number of people involved in accidents has again increased over the last 5 years of this dataset in order to reduce this number in the future. To further this analysis, I would suggest examining the trends over time in comparison to overall population trends to see if some of the findings could be explained (such as increasing age). I would also suggest that more studies consider alcohol involvement in vehicle accidents, since this analysis found that it may be connected to fatalities as well as accidents involving young people.

## References

NHTSA Data: <https://www.nhtsa.gov>

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.