Fatal accidents from the NHTSA

Carlos Arauz

2021-05-02

1 Introduction

The National Highway Traffic Safety Administration (NHTSA) is an agency of the U.S. federal government, part of the Department of Transportation. It describes its mission as saving lives, preventing injuries, and reducing economic costs due to road traffic crashes through education, research, safety standards, and enforcement activity. The NHTSA statement regarding a child's death in a car reads, "Nearly 900 children died of heatstroke since 1998 because they were left or became trapped in a hot car. Everyone needs to understand that children are more vulnerable to heatstroke and that all hot car deaths are preventable. We — as parents, caregivers, and bystanders — play a role in helping to make sure another death doesn't happen".

1.1 Dataset

Data for the analysis is obtained from the FARS website. FARS is a nationwide census providing NHTSA, Congress, and the American public yearly data regarding fatal injuries suffered in motor vehicle traffic crashes.

It provides information detailed below:

1. Fatalities and Fatality Rates in the United States between the year 1994 – 2018
2. Fatal Crashes in the United States between the year 1994 – 2018.
3. Vehicles Involved in Fatal Crashes.
4. Persons Killed by Person Type and Vehicle Type.
5. Persons Killed, by Highest Driver Blood Alcohol Concentration (BAC) in the Crash.
6. 2018 Traffic Fatalities by State and Percent Change from 2017.
7. Person killed by sex (2018 and 2017).
8. Fatal Crashes and Percent Alcohol-Impaired Driving, by the time of day and crash type.

2 Initial Hypothesis

The following are suggested hypotheses

1. Higher Alcohol consumption by drivers will result in fatal crashes. According to findings by (Ramstedt, 2008), the development or increase in a male fatal accident can be partly explained by per capita alcohol consumption. Also," Alcohol is associated with nearly half of all traffic accident deaths in the city of Sao Paulo, especially for days and times associated with parties and bars (weekends between 12 am and 6 am)". (de Carvalho Ponce et al.,2011).

2. Also, traveling in rainy weather conditions and dark light condition increases the risks of traffic accidents. # Exploratory Data Analysis Loading the pertinent libraries

```
library(tidyverse);library(readxl);library(reshape2);library(plotly);library(modelr);
library(mapdata)
```

Loading the dataset

```
fatal.crashes <- read_excel('Dataset.xlsx',sheet='Fatal crashes')
fatalities <- read_excel('Dataset.xlsx',sheet='Fatalities and Fatality Rates') fatalities.state <-
read_excel('Dataset.xlsx',sheet='Traffic Fatalities by STATE') Vehicles.fatalities <-
read_excel('Dataset.xlsx',sheet='Vehicles Invd. in Fatal C.') fatalities.alcohol <- read_excel('Dataset.xlsx',sheet='BAC')
fatalities.gender <- read_excel('Dataset.xlsx',sheet='person killed by sex')
crashes.alcohol<-read_excel('Dataset.xlsx',sheet='Fatal crashes- alcohol impaired',skip=1)
```

Viewing the Dataset

```
head(fatal.crashes,2)
```

```
## # A tibble: 2 x 2##
      Year Total
##    <dbl> <dbl>
## 1   1994 36254
## 2   1995 37241
```

```
head(fatalities,2)
```

```
## # A tibble: 2 x 10
##    Year Fatalities `Resident Populatio~ `Fatality Rate per 1~ `LicensedDrivers~ ##    <dbl>
    <dbl>      <dbl>                <dbl>                 <dbl>
## 1   1994      40716               260327                 15.6            175403
## 2   1995      41817               262803                 15.9            176628
## # ... with 5 more variables: Fatality Rate per 100,000 Licensed Drivers<dbl>, ## #   Registered Motor
Vehicles (Thousands) <dbl>,
## #   Fatality Rate per 100,000 Registered Vehicles <dbl>, ## #
    Vehicle Miles Traveled (Billions) <dbl>,
## #   Fatality Rate per 100 Million VMT <dbl>
```

```
head(fatalities.state,2)
```

```
## # A tibble: 2 x 6
##   State  `state code`  lat  long `2018` `2017`   ##  <chr>  <chr>
<dbl> <dbl> <dbl> <dbl>  ## 1 Alabama AL       28.1 -118.   953
948
## 2 Alaska   AK              32.7  -86.8    80     79
```

```
head(Vehicles.fatalities,2)
```

```
## # A tibble: 2 x 13
## Year `Number of Passe~ `Involvement Rate~ `Involvement Rate~ `Number of Ligh~ ##   <dbl>     <dbl>
                        <dbl>              <dbl>              <dbl> ## 1  1994      30273
                                           2.07               24.8            16353
## 2  1995             30940              2.09               25.1            17587
## # ... with 8 more variables: Involvement Rate per 100 Million VMT...6<dbl>, ## #   Involvement
Rate per 100,000 Registered Vehicles...7 <dbl>,
## #   Number of Large Trucks <dbl>,
## #   Involvement Rate per 100 Million VMT...9 <dbl>,
## #   Involvement Rate per 100,000 Registered Vehicles...10 <dbl>,
```

```
## #       Number of Motorcycles <dbl>,
## #       Involvement Rate per 100 Million VMT...12 <dbl>,
## #       Involvement Rate per 100,000 Registered Vehicles...13 <dbl>
```

```
head(fatalities.alcohol,2)
```

```
##  #  A tibble: 2 x 9
##      Year `BAC = .00`      `Percent   (BAC  =   0.00~  `BAC  =    .01-.07`  `Percent(BAC   =   .01-.0~
##     <dbl>          <dbl>                        <dbl>             <dbl>                      <dbl>
##  1   1994          24948                           61              2236                          5
##  2   1995          25768                           62              2416                          6
##  #       ... with 4  more variables: BAC = .08+ <dbl>,            Percent(BAC    = .08+) <dbl>,
##  #      BAC=.01+  <dbl>, Percent(BAC = .01+) <dbl>
```

```
head(fatalities.gender,2)
```

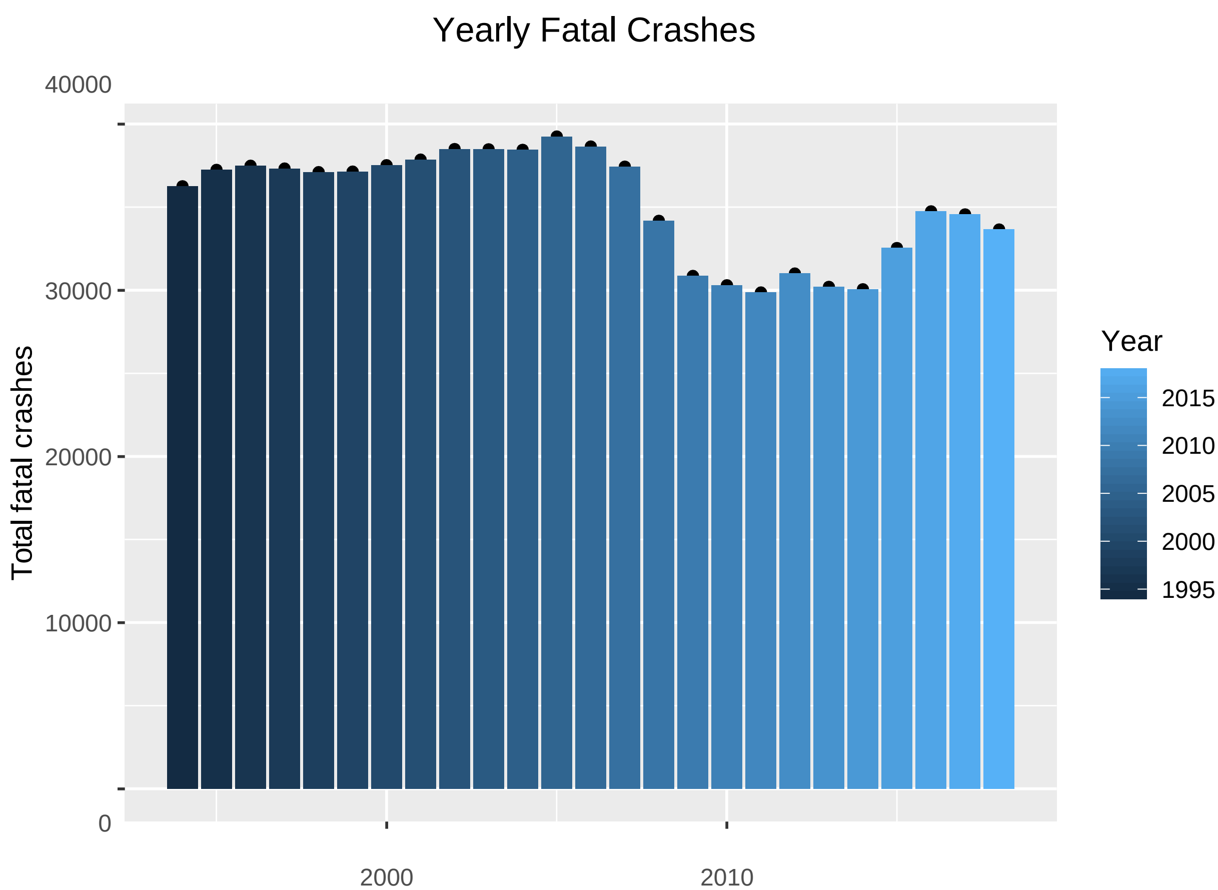
```
## # A tibble: 2 x 3
##   Gender   Total    Year  ##
<chr>  <dbl> <dbl> ## 1 Male
25841   2018
## 2 Female 10676   2018
```

```
head(crashes.alcohol,2)
```

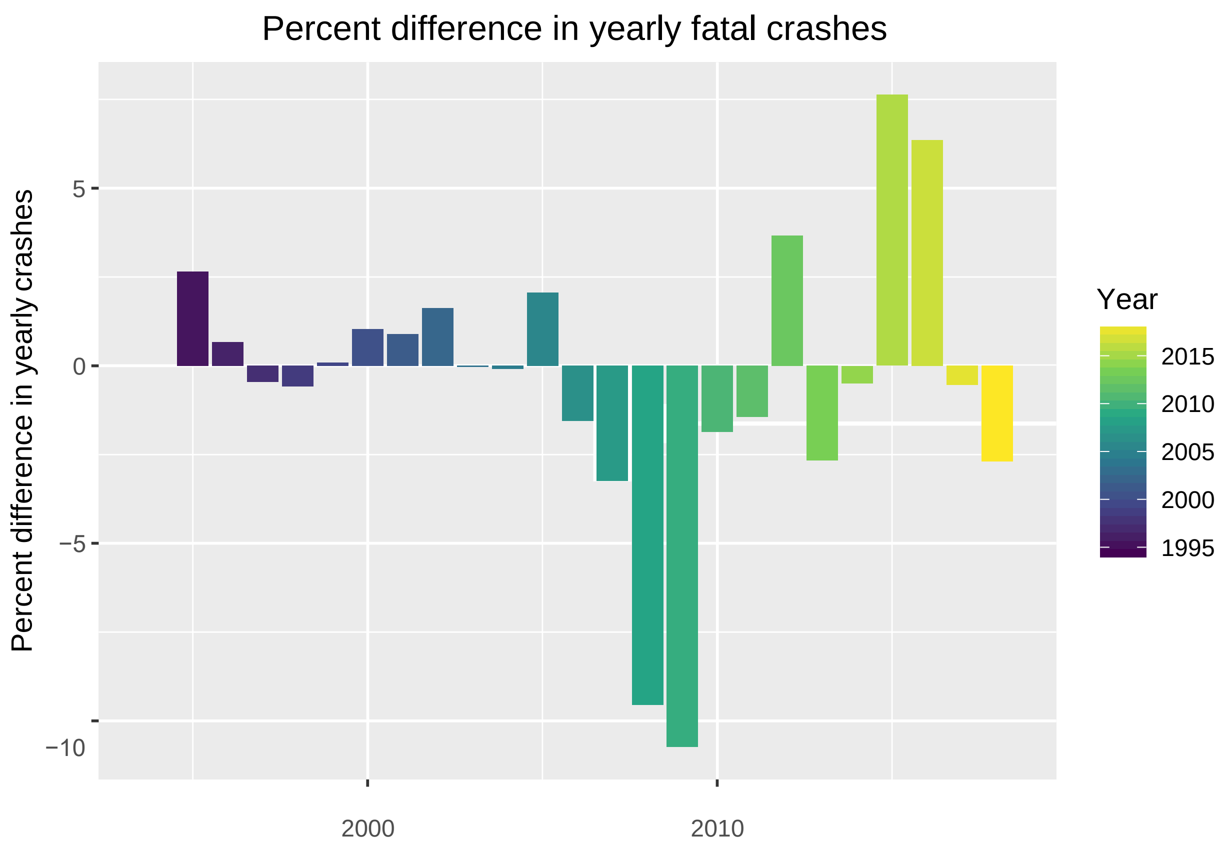
```
##  #  A tibble: 2 x 10
##     Period          `Number SV` `Alcohol-Impaired ~ `Percent Alcohol-Impa~            `Number MV`
##      <chr>                 <dbl>                    <dbl>                    <dbl>         <dbl>
##  1  Midnight t~            2696                     1482                       55          1010
##  2  3 a.m. to ~            1852                      727                       39           952
##  #  ... with 5 more variables: Alcohol-Impaired Driving MV <dbl>,
##  #     Percent Alcohol-Impaired Driving MV <dbl>, Number <dbl>,
##  #                          Alcohol-Impaired Driving <dbl>, Percent Alcohol-Impaired Driving  <dbl>
```

There was an increase in the total number of fatal crashes between the year 1994 to 2005. There was a decline in fatal crashes between 2006 and 2014. Fatalcrashes remain reasonably constant between the year 2009 and 2014. The total number of fatal crashes was seen to increase between the year 2014 and 2017. The drop in fatal crashes between 2017 and 2018 is 2.69%. There was a constant drop in fatal crashes between the year 2005 and 2009.

```
ggplot(data = fatal.crashes, aes(x=Year,y= Total, fill = Year))+geom_point()+geom_bar( stat='identity')+labs(y="Total
    fatal crashes",title = "Yearly Fatal Crashes",x="")+
    theme(plot.title = element_text (hjust=0.5))
```



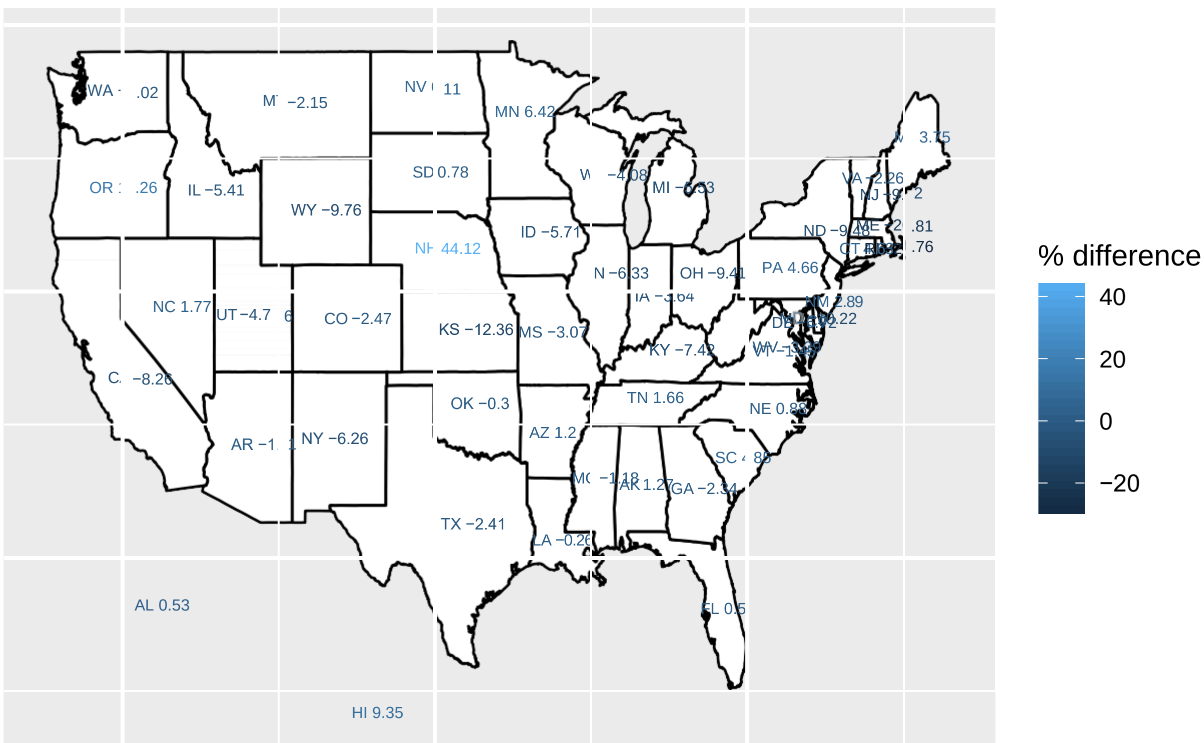

```
fatal.crashes <- fatal.crashes                    %>%
    mutate(percent_difference = round(c(0,diff(Total)/Total[-1]*100),2) ) ggplot(data = fatal.crashes,
aes(x=Year, y= percent_difference,fill = Year))+
    geom_bar(stat='identity')+labs(y = 'Percent difference in yearly crashes',x="")+ scale_fill_viridis_c()+
    ggtitle("Percent            difference in yearly fatal crashes")+
    theme(plot.title = element_text (hjust=0.5))
```




```r
states_map <- map_data("state")
fatalities.state$State <-tolower(fatalities.state$State) fatalities_by_state
= fatalities.state %>%
  transmute(State,`state code`,lat,long,`% difference`=round(((`2018`-`2017`)/`2017`)*100
,2))
ggplot(data=fatalities_by_state,aes(color=`% difference`)) +
  geom_polygon( data=states_map, aes(x=long, y=lat,group=group),colour="black", fill="white"
)+geom_text(data=fatalities_by_state,aes(x=long,y=lat,
  label=`% difference`),size=2,hjust=-0.08)+geom_text(data=fatalities_by_state, aes(x=long,
  y=lat,label=`state code`), size=2, hjust=1)+
  ggtitle("Percent difference in fatalities",subtitle="(2018 - 2017)")+
  theme(axis.title.x=element_blank(),axis.ticks.x=element_blank(), axis.ticks.y=element_blank(),axis.text.x
  =element_blank(),axis.text.y = element_blank(),axis.title.y=element_blank(),plot.title = element_text (hjust=0.5),
  plot.subtitle = element_text (hjust=0.5))
```

Percent difference in fatalities

(2018 − 2017)



New Hampshire has the highest percentage increase in fatalities between 2017 and 2018, while Rhode Island has the lowest percentage drop in total fatalities between 2018 and 2017

```
#sorting in descending order by % difference
head(arrange(fatalities_by_state, desc(`% difference`)),5)
```

```
## # A tibble: 5 x 5
##   State          `state code`   lat   long `% difference`
##   <chr>          <chr>        <dbl>  <dbl>          <dbl>
## 1 new hampshire  NH            41.5  -99.8          44.1
## 2 oregon         OR            43.9 -121.           15.3
## 3 hawaii         HI            24.2 -104.            9.35
## 4 minnesota      MN            46.6  -94.6           6.42
## 5 nevada         NV            47.5 -100.            6.11
```

```
#sorting in descending order by % difference
head(arrange(fatalities_by_state, `% difference`),5)
```

```
## # A tibble: 5 x 5
##   State        `state code`   lat   long `% difference`
##   <chr>        <chr>        <dbl>  <dbl>          <dbl>
## 1 rhode island RI            41.6  -71.5         -29.8
## 2 maine        ME            42.4  -71.5         -20.8
## 3 kansas       KS            38.5  -98.4         -12.4
## 4 maryland     MD            39.0  -76.3         -10.2
## 5 wyoming      WY            43.0 -108.           -9.76
```

About 18 States recorded an increase In fatalities in 2018 when compared to 2017. New Hampshire (41.5%) and Oregon (43.9%) are the two states with the most significant percentage increase in the number of fatalities between 2017 and 2018. States with the most percentage decrease in the number of fatalities between 2017 and 2018 are: Rhode Island (-29.76%) and Maine (-20.81%)

```
fatalities_by_state %>% filter(`% difference`>0) %>%
  arrange(desc(`% difference`)) %>%
  head(3)
```

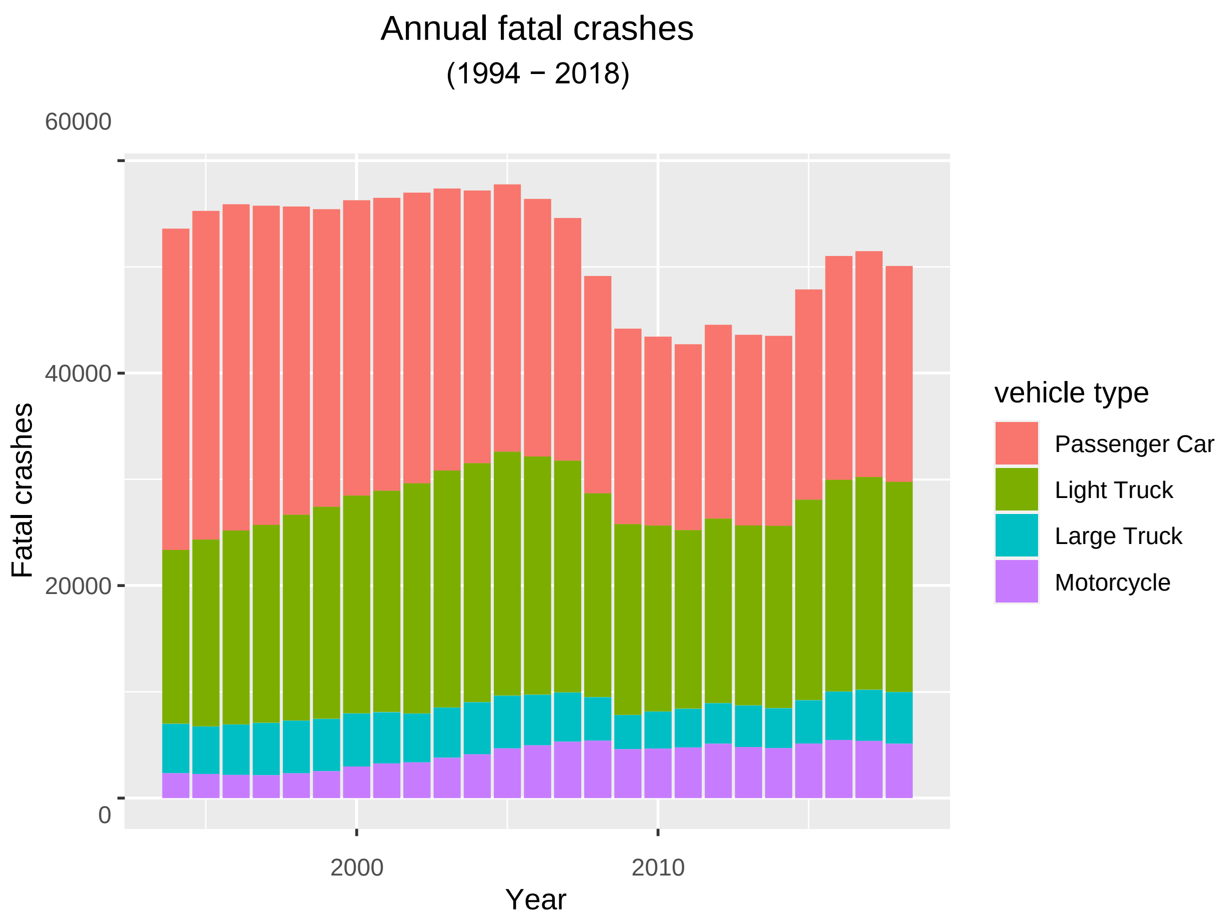
```
## # A tibble: 3 x 5
##   State         `state code`   lat   long `% difference`
##   <chr>         <chr>        <dbl>  <dbl>          <dbl>
## 1 new hampshire NH            41.5  -99.8          44.1
## 2 oregon        OR            43.9 -121.           15.3
## 3 hawaii        HI            24.2 -104.            9.35
```

```
fatalities_by_state %>% filter(`% difference`<0) %>%
  arrange(`% difference`) %>%
  head(3)
```

```
## # A tibble: 3 x 5
##   State        `state code`   lat  long `% difference`
##   <chr>        <chr>        <dbl> <dbl>          <dbl>
## 1 rhode island RI            41.6 -71.5          -29.8
## 2 maine        ME            42.4 -71.5          -20.8
## 3 kansas       KS            38.5 -98.4          -12.4
```

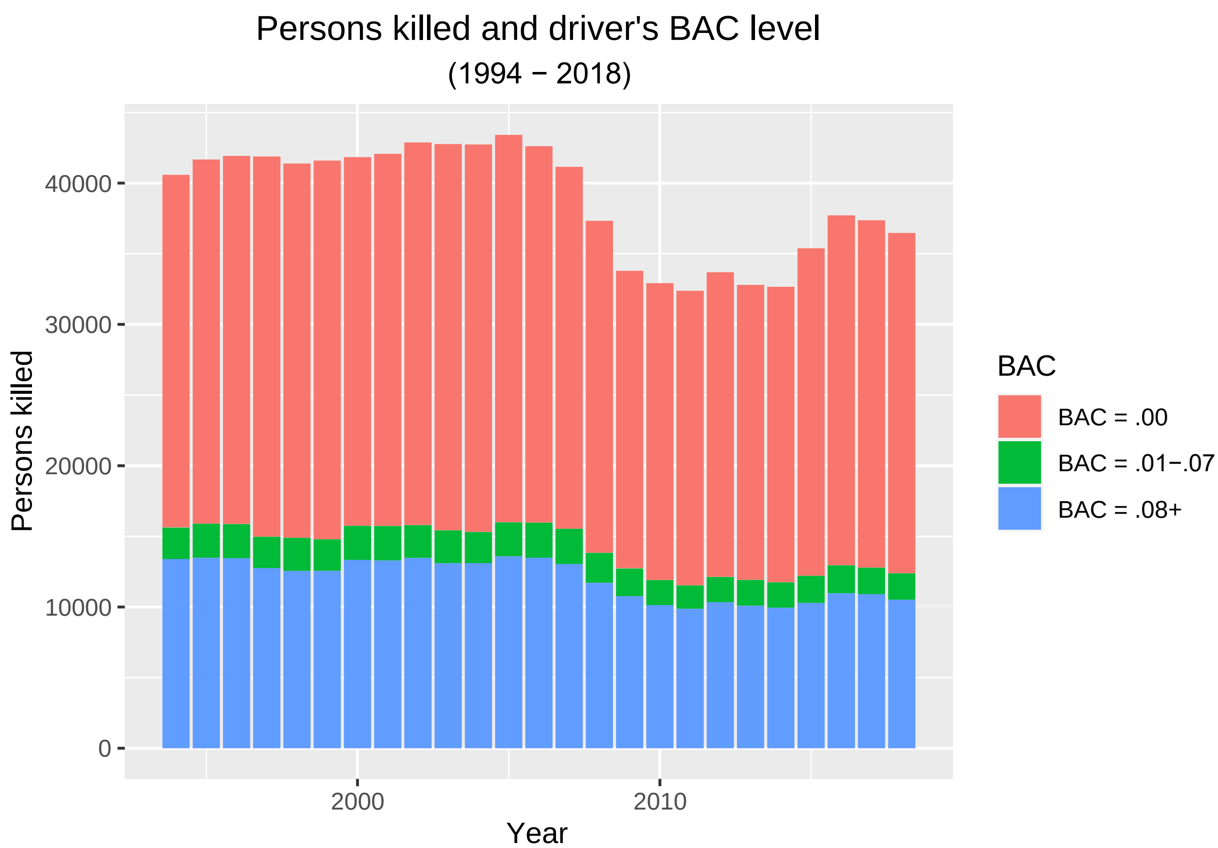
Most crashes are caused by Passenger cars and Light trucks since 1994 – 2018.

```
vehicle_type.fatalities <- Vehicles.fatalities %>% select(Year,`Number of Passenger
  Cars`,`Number of LightTrucks`,
        `Number of Large Trucks`,`Number of Motorcycles`) %>%
  rename(`Passenger Car`=`Number of Passenger Cars`,`Light Truck`=
          `Number of Light Trucks`,`Large Truck`=`Number of Large Trucks`,
        `Motorcycle`=`Number of Motorcycles`)
Pivot_vehicle_type.fatalities <- vehicle_type.fatalities %>%
  melt(id.vars=c("Year"),value.name = "Fatal crashes") %>% rename(`vehicle
  type` = variable)
ggplot(data = Pivot_vehicle_type.fatalities, aes(fill=`vehicle type`))+ geom_bar(mapping = aes(x =
  Year, y = `Fatal crashes`), stat = "identity")+ ggtitle("Annual fatal crashes", subtitle = "(1994 -
  2018)")+
  theme(plot.title = element_text (hjust=0.5),plot.subtitle = element_text(hjust=0.5))
```



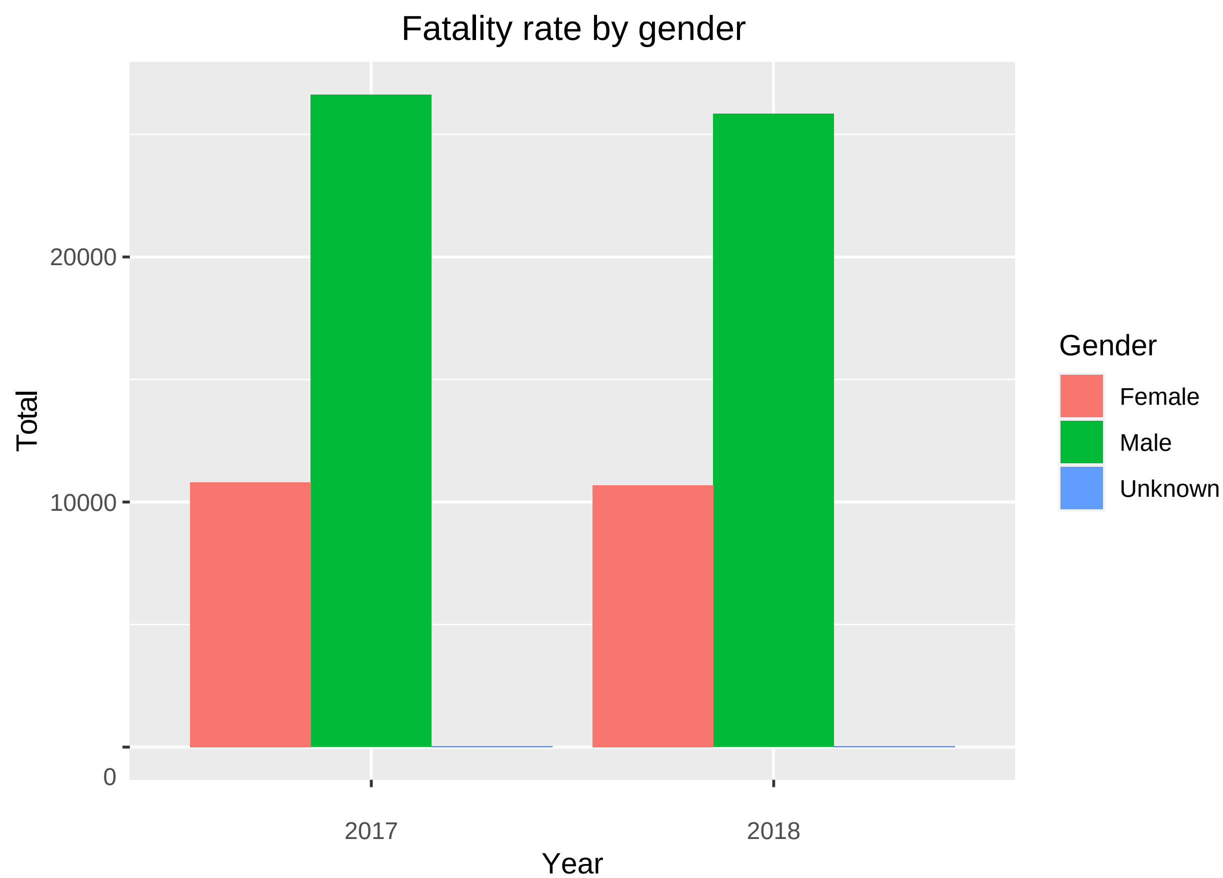
For most of the persons killed between 1994 and 2018, the driver's BAC level was 0. However, more fatalities resulted from drivers with BAC level of 0.08+ than drivers with BAC level of 0.01-0.07.

```
Persons_Killed.highestBAC <- fatalities.alcohol                    %>%
  select(Year,`BAC = .00`,`BAC = .01-.07`,`BAC = .08+`)        >     % %
  melt(id.vars=c("Year"),value.name = "Persons killed")              % %
  rename(BAC = variable)
ggplot(data = Persons_Killed.highestBAC, aes(fill=BAC))+
  geom_bar( mapping = aes(x = Year, y = `Persons killed`), stat = "identity")+ ggtitle("Persons killed and driver's
  BAC level",subtitle = "(1994 - 2018)")+ theme(plot.title = element_text (hjust=0.5),plot.subtitle =
  element_text(hjust=0.5))
```

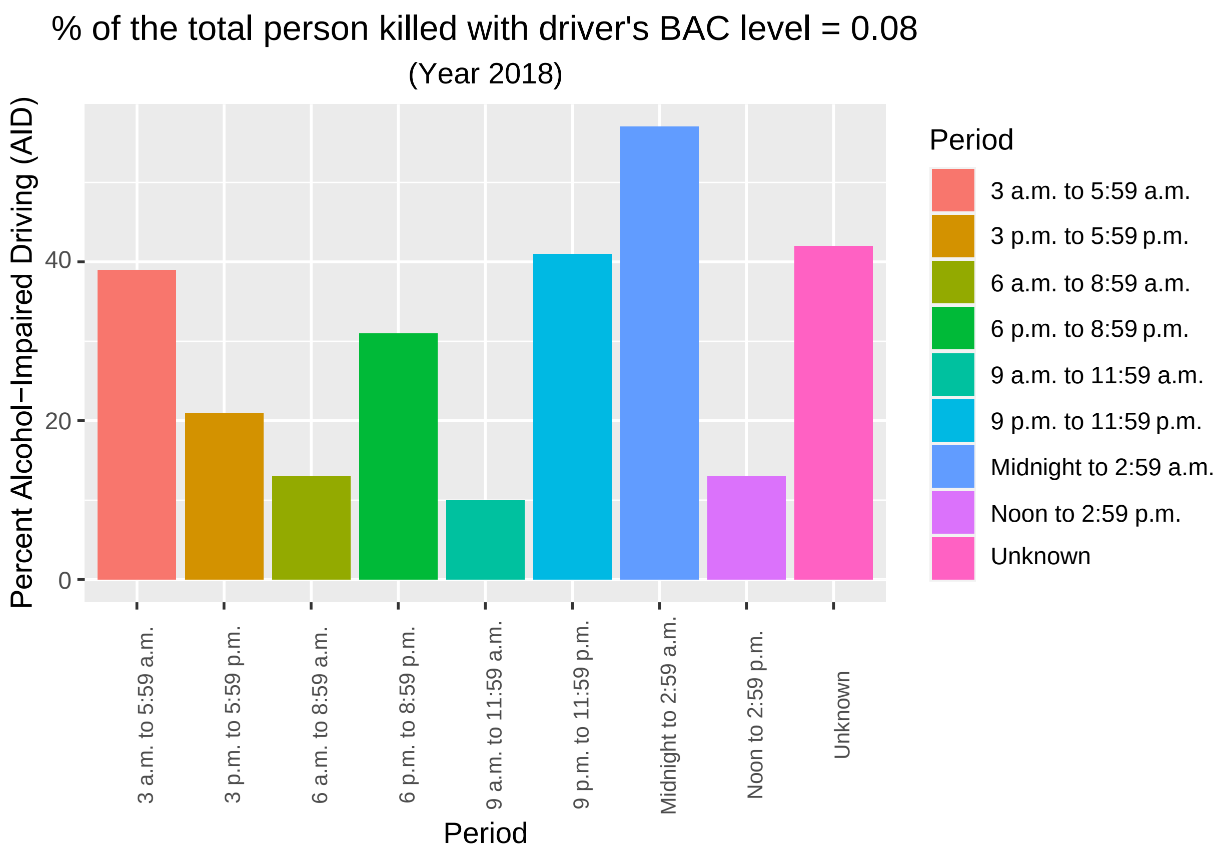


fatality rate was lower for females than for males

```
fatalities.gender$Year <- as.factor(fatalities.gender$Year) ggplot(data=fatalities.gender,aes(fill=Gender))+
  geom_bar(position = "dodge",mapping = aes(x = Year, y = `Total`), stat = "identity")+
  ggtitle("Fatality rate by gender")+ theme(plot.title = element_text (hjust = 0.5))
```




```
ggplot(data=crashes.alcohol,aes(fill=Period))+
  geom_bar( mapping = aes(x =Period, y = `Percent Alcohol-Impaired Driving`),
       stat = "identity")+labs(x='Period',y='Percent Alcohol-Impaired Driving (AID)', title="% of the total person
  killed with driver's BAC level = 0.08",
  subtitle = "(Year 2018)")+ theme(axis.text.x=element_text(angle=90, size=8),
  plot.title = element_text (hjust = 0.5),plot.subtitle = element_text (hjust = 0.5))
```




```
tot=sum(crashes.alcohol$`Alcohol-Impaired Driving`)/sum(crashes.alcohol$Number)*100
cat("percentage of fatal crashes in 2018 involving AID ",paste(round(tot,2)," ",sep=''))                    %
```

```
## percentage of fatal crashes in 2018 involving AID                    28.4 %
```

Twenty-eight percent of all fatal crashes in 2018 involved alcohol-impaired driving. The highest blood alcohol concentration (BAC) among drivers involved in the crash was .08 grams per deciliter (g/dL) or higher. For fatal crashes occurring from midnight to 3 am, Fifty-seven percent involved alcohol-impaired driving. Similarly, Fourty-one percent involved alcohol-impaired driving for fatal crashes between 9 pm to 11:59 pm.

```
res.pop = cor(fatalities$Fatalities,fatalities$`Resident Population (Thousands)`)
lic.driver=cor(fatalities$Fatalities,fatalities$`Licensed Drivers (Thousands)`) vmt =
cor(fatalities$Fatalities,fatalities$`Vehicle Miles Traveled (Billions)`)
cat("Correlation between fatalities and Resident Population (Thousands)=",res.pop,"\n")
```

```
## Correlation between fatalities and Resident Population (Thousands)= -0.7082962
```

```r
cat("Correlation between fatalities and Licensed Drivers(Thousands)=",lic.driver,"\n")
```

```
## Correlation between fatalities and Licensed Drivers (Thousands)= -0.6779238
```

```r
cat("Correlation between fatalities and Vehicle Miles Traveled (Billions)=",vmt)
```

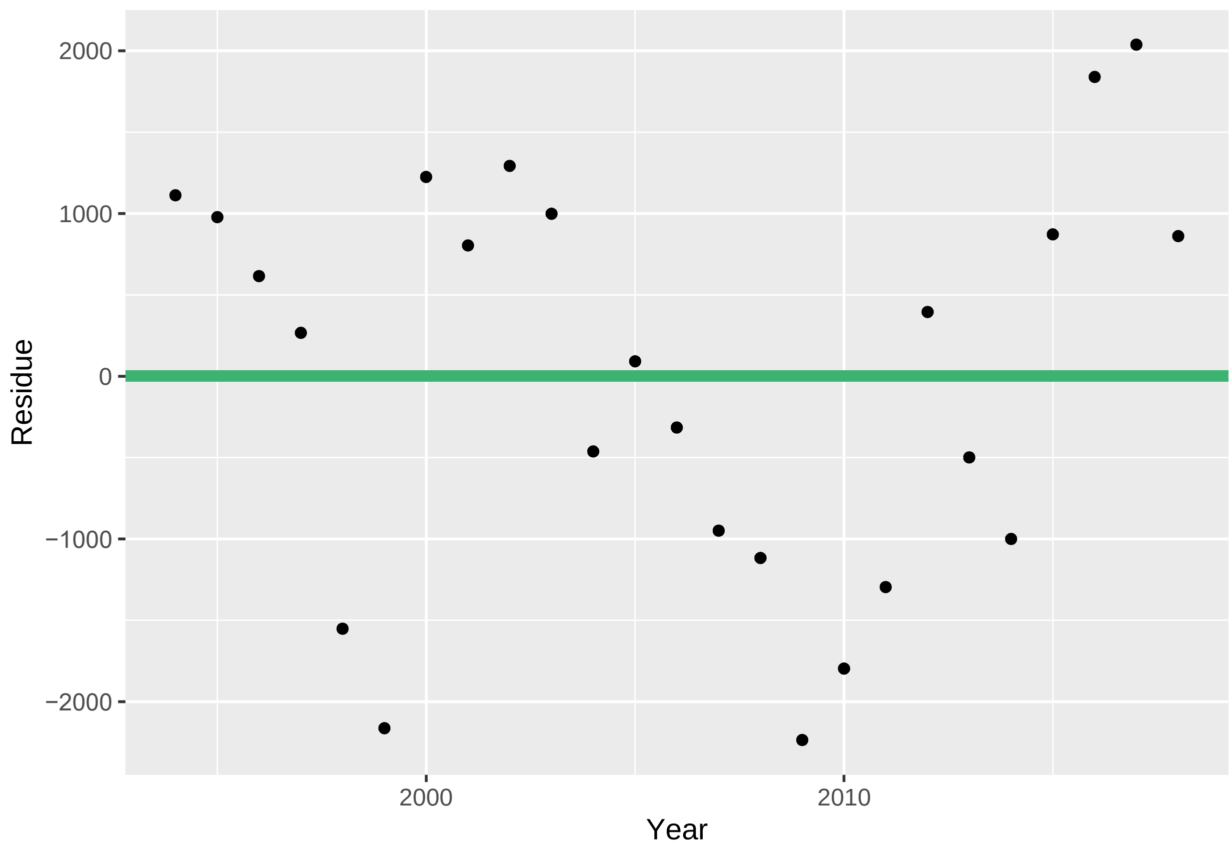
```
## Correlation between fatalities and Vehicle Miles Traveled (Billions)= -0.4665822
```

There is a strong negative correlation between fatalities and Resident Population (Thousands). Correlation also also exists between fatalities and Licensed Drivers (Thousands) in the opposite direction and not too strong. Similarly, there exists a weak correlation between fatalities and Vehicle Miles Traveled (Billions) in the opposite direction. However, all these correlations do not imply causation.

We can build a multiple linear regression model to estimate the possible total number of fatalities using

1. Resident Population (Thousands)
2. Licensed Drivers (Thousands)
3. Vehicle Miles Traveled (Billions)

```r
model = lm(Fatalities~`Resident Population (Thousands)`+`Licensed Drivers (Thousands)`+
  `Registered Motor Vehicles (Thousands)`+`Vehicle Miles Traveled (Billions)`, data=fatalities)
# Residual standard error: 1367 on 20 degrees of freedom
# Multiple R-squared:       0.9016,     Adjusted R-squared:      0.882
# F-statistic: 45.84 on 4 and 20 DF,      p-value: 8.485e-10 #Next we
compute and visualize the residuals:
lm_model1 <- fatalities          %>%
  add_residuals(model)
lm_model1    %>%
  ggplot(aes(Year, resid)) +geom_ref_line(h = 0,colour = 'mediumseagreen') + geom_point()+ylab("Residue")
```

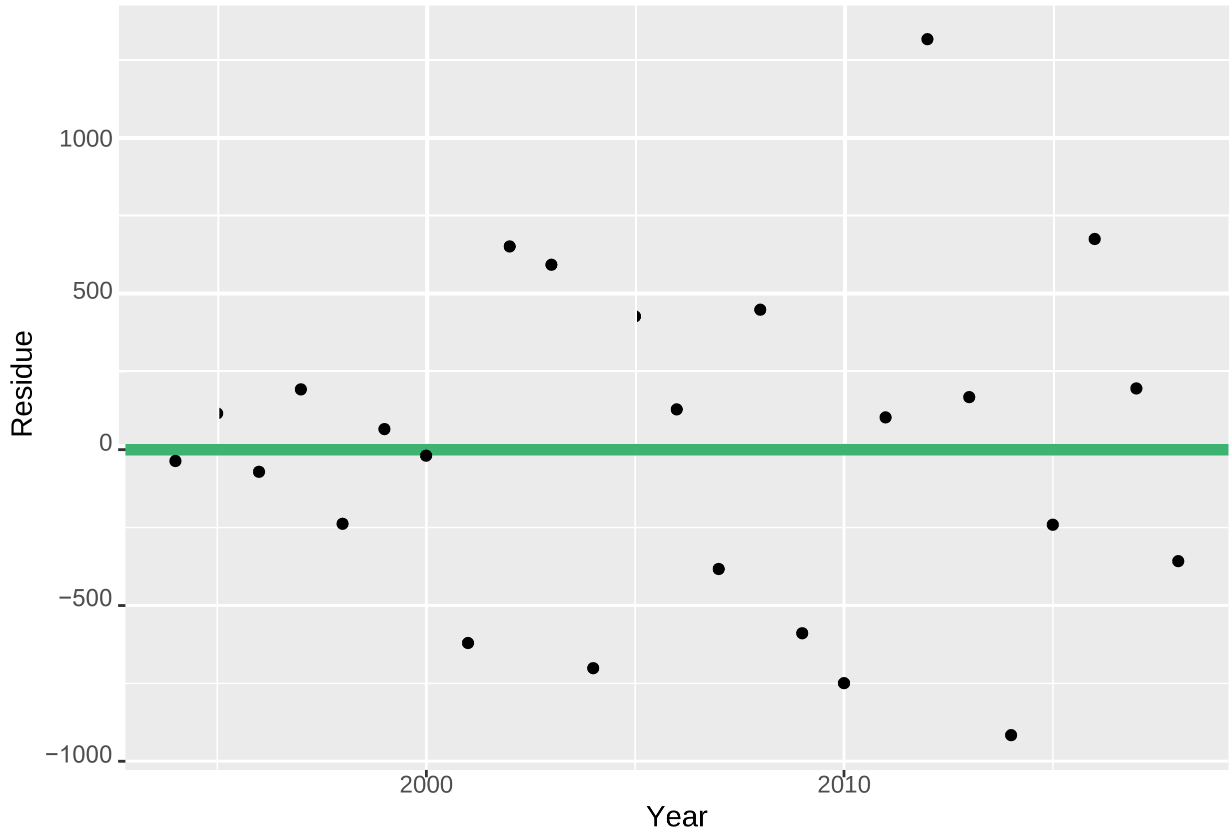


From the residue plot above, the residues vary between +2000 fatalities and -2000 fatalities.

Incorporating interaction between the variables into the model

```
model2=lm(Fatalities~`Resident Population (Thousands)`*`Licensed Drivers (Thousands)`*
                    `Registered Motor Vehicles (Thousands)`*
                    `Vehicle Miles Traveled (Billions)`, data = fatalities)
# Residual standard error: 846.7 on 9 degrees offreedom
# Multiple R-squared:        0.983,      Adjusted R-squared:       0.9547
# F-statistic: 34.72 on 15 and 9 DF,      p-value: 4.041e-06
```

```
#Next we compute and visualize the residuals:
lm_model2 <- fatalities          %>%
  add_residuals(model2)
lm_model2  %>%
  ggplot(aes(Year, resid)) +geom_ref_line(h = 0,colour = 'mediumseagreen') + geom_point()+ylab("Residue")
```



Incorporating the interaction factor gave a better model as this was obvious in the R^2 and Adjusted R^2. The R^2 value is 0.983 compared to 0.9016 without considering the interaction effect. Also, from the residue plot, The residue plot when the interaction effected was considered clustered around the 0 reference line more than the residue obtained without considering the interaction effect. The residues vary between approximately -1000 fatalities and +1000 fatalities.

3 Discussion

1. Blood alcohol concentration (BAC) among drivers increases fatal crash
2. Fatality rate is lower for females across all ages than for males.
3. There is a decrease in fatal crashes between year 2017 and 2018 (-2.69%).

References

1. de Carvalho Ponce, J., Muñoz, D. R., Andreuccetti, G., de Carvalho, D. G., & Leyton, V. (2011).Alcohol-related traffic accidents with fatal outcomes in the city of Sao Paulo. *Accident Analysis & Prevention*, 43(3),

782–787.https://doi.org/10.1016/j.aap.2010.10.025

2. *FARS Encyclopedia (dot.gov)*, https://www-fars.nhtsa.dot.gov/Main/index.aspx,
3. Ramstedt, M. (2008). Alcohol and fatal accidents in the United States—A time series analysis for 1950–2002. *Accident Analysis & Prevention*, 40(4), 1273–1281. ‘ https://doi.org/10.1016/j.aap.2008.01.008