Data Analysis Project on

Traffic Violations Recorded in Montgomery County

About Dataset:

The dataset that I have selected is the traffic violations that are recorded in the county of Montgomery county of Maryland. The dataset contains all the violations that are recorded from approximately past 25years along with the cause and details of the violation. The size of the file is approximately 500MB with total of 1411992 rows and 35 columns.

1) **Who (company, organization, agency) collected the data?**
   a) **Who they are, what do they do?**
   "The data.gov is the government website which aims to improve public access to high value machine readable datasets which are developed by few officials of federal government. It contains datasets related to various domains at state, county, and tribal level of the country U.S." [1]

   b) **What is their role/purpose?**
   The role of this data.gov is that it provide various real time datasets belonging to various domains which can be used for research and analysis of particular problem or situation. Also, "it aims at providing data analysts with various insights about the big-data problems that may arise due to various reasons and provide solutions for these problems. It also helps in analyzing a particular situation in state or county of the country U.S."[3]

2) **Need**
   a) **Why did they collect this data?**
   The main purpose of collecting this data is to uniquely identify the person who have frequently involved in traffic violations and find out the cases filed against him/her and take necessary actions. Also, it is collected to analyze the various outcomes of the violations like accidents, loss of life's, etc.

3) **Questions that can be answered by studying this data?**

   ➔ How many traffic violations are recorded in respective years?
   ➔ How many times traffic violations led to accidents?
   ➔ How many times violations led to property damage or personal injury or fatal conditions?
   ➔ How many males or females are involved in violations?
   ➔ Which state has got more number of traffic violations?

Requirements & Resources needed:

Software Requirements:

➔ Installing SQL and setting up the environment for performing SQL operations on the dataset.

- ➔ R- Studio for analysis and visualizations of the results.
- ➔ PyCharm for running the python script.

Hardware Requirements:

- ➔ Laptop or PC with 8GB RAM.
- ➔ Intel I-5 processor
- ➔ L2 Cache- 256 MB
- ➔ Processor speed 1.6GHz

a) Issues with data:

-- Privacy:

The only privacy issues of the data are that when fall into false hands there is a chance of overriding the records in the dataset and the dataset is public.

--Quality:

The quality issues with this data is that it contains raw data and extra columns which are not required for the analysis, so we need to filter the dataset. Also the dataset contains many blank fields (i.e NA's) which are to be removed from the dataset for preparing the dataset suitable for extracting the results. The dataset can uniquely identify the vehicle. [1]

b) Metadata Definitions: [1]

The dataset describes the number of traffic violations recorded in the county of Montgomery, Maryland in United States. It contains records for past 10 years which are recorded by traffic department of the county.

Dataset Description:

1) **Date of Stop:** The date on which the traffic violation is recorded.**[Datatype: Ordinal]**
2) **Time of Stop:** The time at which the violation is recorded. **[Datatype: Ratio]**
3) **Agency:** The agency where the violation is recorded. **[Datatype: Nominal]**
4) **Sub Agency:** The district where the violation is recorded. **[Datatype: Nominal]**
5) **Description:** The description of rule that he/she has violated. **[Datatype: Nominal]**
6) **Location:** The address of street where the violation is recorded. **[Datatype: Nominal]**
7) **Latitude:** Latitude of location where violation has occurred. **[Datatype: Ordinal]**
8) **Longitude:** Longitude of location where the violation has occurred. **[Datatype: Ordinal]**
9) **Accident:** Whether the violation led to accident or not. **[Datatype: Nominal]**
10) **Belts:** Whether the violation is of not putting the seat belt. **[Datatype: Nominal]**
11) **Personal Injury:** Whether the violation led to personal injury or not. **[Datatype: Nominal]**
12) **Property Damage:** Whether the violation led to any property damage or not.**[Datatype: Nominal]**
13) **Fatal:** Whether the violation led to accident where the condition is fatal or not. **[Datatype: Nominal]**
14) **Commercial License:** Whether the violation is caused by person who has commercial license or not. **[Datatype: Nominal]**
15) **HazMat:** Whether there is any hazardous material in the vehicle or not. **[Datatype: Nominal]**

16) **Commercial Vehicle:** Whether the vehicle is commercial or not. **[Datatype: Nominal]**
17) **Alcohol:** Whether the violation is caused due to consumption of alcohol or not. **[Datatype: Nominal]**
18) **Work zone:** Whether the violation has occurred at location where the construction is going on or not. **[Datatype: Nominal]**
19) **State:** State in which the violation has recorded. **[Datatype: Nominal]**
20) **Vehicle type:** Whether the rules are violated by 2 wheeler or4 wheeler or other type of vehicles. **[Datatype: Nominal]**
21) **Year:** Year in which the vehicle involved in violation is purchased. **[Datatype: Interval]**
22) **Make:** Name of the manufacturer/company of the vehicle. **[Datatype: Nominal]**
23) **Model:** Type/Model of the vehicle involved in violation. **[Datatype: Nominal]**
24) **Color:** Color of the vehicle involved in violation. **[Datatype: Nominal]**
25) **Violation:** Where the punishment for violation is referred to. **[Datatype: Nominal]**
26) **Charge:** What is the charge sheet filed against person involved in violation. **[Datatype: Nominal]**
27) **Article:** Article type from where the record of violation is collected. **[Datatype: Nominal]**
28) **Contributed to accident:** Whether the violation led to accident of other vehicles or not. **[Datatype: Nominal]**
29) **Race:** Race of the person involved in violation. **[Datatype: Nominal]**
30) **Gender:** Whether the rules are violated by male or female. **[Datatype: Nominal]**
31) **Driver city:** Name of driver's city. **[Datatype: Nominal]**
32) **Driver state:** Name of driver's state.**[Datatype: Nominal]**
33) **DL state:** State which issued driving license to driver. **[Datatype: Nominal]**
34) **Arrest type:** Type of arrest warrant issued. **[Datatype: Nominal]**
35) **Geolocation:** Co-ordinates of the location where the accident is recorded. **[Datatype: Nominal]**

SQL schema and SQL based operations:

SQL SCHEMA:

```
create database traffic;

use traffic;

#creating the table in database(schema)

create table violations(dateofstop varchar(100), timeofstop varchar(100), agency varchar(100), subagency varchar(100), description varchar(100), location varchar(100), lat float, longi float, accident varchar(100), belts varchar(100), personalinjury varchar(100), propertydamage varchar(100), fatal varchar(100), commlic varchar(100), hazmat varchar(100), commvechile varchar(100), alcohol varchar(100), workzone varchar(100), state varchar(100), vechiletype varchar(100), year int, make varchar(100), model varchar(100), color varchar(100), violationtype varchar(100), charge varchar(100), article varchar(100), contributetoacc varchar(100), race varchar(100), gender varchar(100), drivercity varchar(100), driverstate varchar(100), dltype varchar(100), arresttype varchar(100), geoloc varchar(100));

#loading csv data into database
```

load data infile '/media/sf_VM-files/Traffic_Violations.csv' into table violations fields terminated by ',' lines terminated by '\n'
(dateofstop,timeofstop,agency,subagency,description,location,lat,longi,accident,belts,personalinjury,propertydamage,fatal,commlic,hazmat,commvechile,year,make,model,color,violationtype,charge,article,contributetoacc,race,gender,drivercity,driverstate,dltype,arresttype,geoloc);

#SQL operations selecting only first row

select * from violations LIMIT 1;

```
mysql> select * from violations LIMIT 2,1;
+-----------+-----------+--------+-------------+------------+----------------------------------------------------------------------------------------
+------+---------+-----------------+-------+-------------+--------------+-------+---------+--------+-----------+---------+---------+
------+------------+------+------+-------+-------+---------------+--------+---------+---------------+-------+---------+-----------+---
-------------------+--------+-----------+--------+
| dateofstop | timeofstop | agency | subagency     | description | location
| lat  | longi   | accident         | belts | personalinjury | propertydamage | fatal | commlic | hazmat | commvechile | alcohol | workzone |
 state | vechiletype | year | make | model | color | violationtype   | charge | article | contributetoacc | race  | gender   | drivercity | dr
iverstate           | dltype | arresttype | geoloc |
+-----------+-----------+--------+-------------+------------+----------------------------------------------------------------------------------------
+------+---------+-----------------+-------+-------------+--------------+-------+---------+--------+-----------+---------+---------+
------+------------+------+------+-------+-------+---------------+--------+---------+---------------+-------+---------+-----------+---
-------------------+--------+-----------+--------+
| 08/29/2017 | 10:19:00   | MCP    | "2nd district |  Bethesda"  | DRIVER FAILURE TO OBEY PROPERLY PLACED TRAFFIC CONTROL DEVICE INSTRUCTIONS
|    0 | 38.9817 | -77.0927566666667 | No    | No            | No            | No    | No     | No     | No          | NULL    | NULL     |
 NULL  | NULL       |    0 | No   | No    | VA    | 02 - Automobile | 2001   | TOYOTA  | COROLLA        | GREEN | Citation | 21-201(a1) | Tr
ansportation Article | No     | WHITE     | F     |
+-----------+-----------+--------+-------------+------------+----------------------------------------------------------------------------------------
+------+---------+-----------------+-------+-------------+--------------+-------+---------+--------+-----------+---------+---------+
------+------------+------+------+-------+-------+---------------+--------+---------+---------------+-------+---------+-----------+---
-------------------+--------+-----------+--------+
1 row in set (0.00 sec)
```

Reading the data for analysis:
The libraries and data present in csv file is read using the following R-script snippet.

```
library(tidyverse)
library(ggplot2)
library(plyr)
data <- read.csv('F:/study/masters/1sem/ait-580/final project/Tra
```

The above data read has many unused columns and blank columns which are to be prepared for data analysis.

Data preprocessing:
The blank data is to be removed and only the required columns are to be selected, also we need to handle the categorical data by converting categorical data to numerical data i.e(1 and 0's).This is done using the following code snippet.

```
#remove empty rows and rows with NA's
data <- na.omit(data)

#select required columns
data <- subset(data,select=c(7:19,21,28:30,32))
data <- data.frame(data)

#Handling categorical variables i.e converting Yes ->1, No->0
data$Accident <- ifelse(data$Accident == "Yes",1,0)
data$Belts <- ifelse(data$Belts == "Yes",1,0)
data$Personal.Injury <- ifelse(data$Personal.Injury == "Yes",1,0)
data$Property.Damage <- ifelse(data$Property.Damage == "Yes",1,0)
data$Fatal <- ifelse(data$Fatal == "Yes",1,0)
data$Commercial.License <- ifelse(data$Commercial.License == "Yes",1,0)
data$HAZMAT <- ifelse(data$HAZMAT == "Yes",1,0)
data$Commercial.Vehicle <- ifelse(data$Commercial.Vehicle == "Yes",1,0)
data$Alcohol <- ifelse(data$Alcohol == "Yes",1,0)
data$Work.Zone <- ifelse(data$Work.Zone == "Yes",1,0)
```

Results and Findings:

Descriptive statistics and visualizations for selected data items:

Since I have converted categorical variables into numerical the summary statistics values ranges between 0-1.

Statistics of column violations because of belts:

Min value	1st Quartile	Median	Mean	3rd Quartile	Max
0.00000	0.00000	0.00000	0.03335	0.00000	1.00000

The mean of the column is 0.03335 which tells us that average of 33.35% cases are registered due to not putting their seat belts.

Statistics of column violations led to personal injury:

Min value	1st Quartile	Median	Mean	3rd Quartile	Max
0.00000	0.00000	0.00000	0.01119	0.00000	1.00000

The mean of the column is 0.01119 which tells us that average of 11.19% cases violations have led to personal injury.

Statistics of column violations led to property damage:

Min value	1st Quartile	Median	Mean	3rd Quartile	Max
0.00000	0.00000	0.00000	0.01784	0.00000	1.00000

The mean of the column is 0.01784 which tells us that average of 17.84% cases violations have led to property damage.

Statistics of column violations led to fatal injury:

Min value	1st Quartile	Median	Mean	3rd Quartile	Max
0.00000	0.00000	0.00000	0.0001654	0.00000	1.00000

The mean of the column is 0.0001654 which tells us that average of 0.0164% cases violations have led to fatal injuries i.e very minute cases.

Statistics of column violations is of carrying hazardous material:

Min value	1st Quartile	Median	Mean	3rd Quartile	Max
0.00e+00	0.00e+00	0.00e+00	8.62e-05	0.00e+00	1.00e+00

The mean of the column is 8.62e-05 which tells us that average of 0.00862% cases violations are of carrying hazardous material.

Statistics of column violations by commercial vehicle:

Min value	1st Quartile	Median	Mean	3rd Quartile	Max
0.000000	0.000000	0.000000	0.004648	0.000000	1.000000

The mean of the column is 0.004648 which tells us that average of 0.4648% cases violations are caused by commercial vehicle.

Statistics of column violations are of drinking alcohol and driving:

Min value	1st Quartile	Median	Mean	3rd Quartile	Max
0.000000	0.000000	0.000000	0.001461	0.000000	1.000000

The mean of the column is 0.001461 which tells us that average of 0.1461% cases violations are of drunk and driving.

Statistics of column violations are at work zone:

Min value	1st Quartile	Median	Mean	3rd Quartile	Max

0.000000	0.000000	0.000000	0.0001954	0.000000	1.000000

The mean of the column is 0.0001954 which tells us that average of 0.01954% cases violations have occurred at work zone.

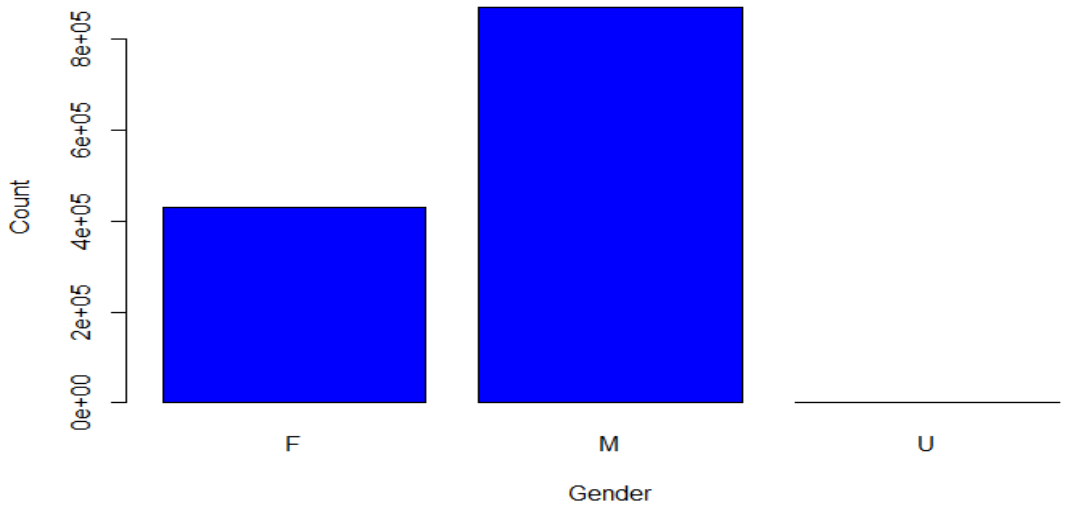
Statistics of column violations that have contributed in accidents:

Min value	1st Quartile	Median	Mean	3rd Quartile	Max
0.000000	0.000000	0.000000	0.02276	0.000000	1.000000

The mean of the column is 0.02276 which tells us that average of 2.276% cases violations have contributed to accidents.

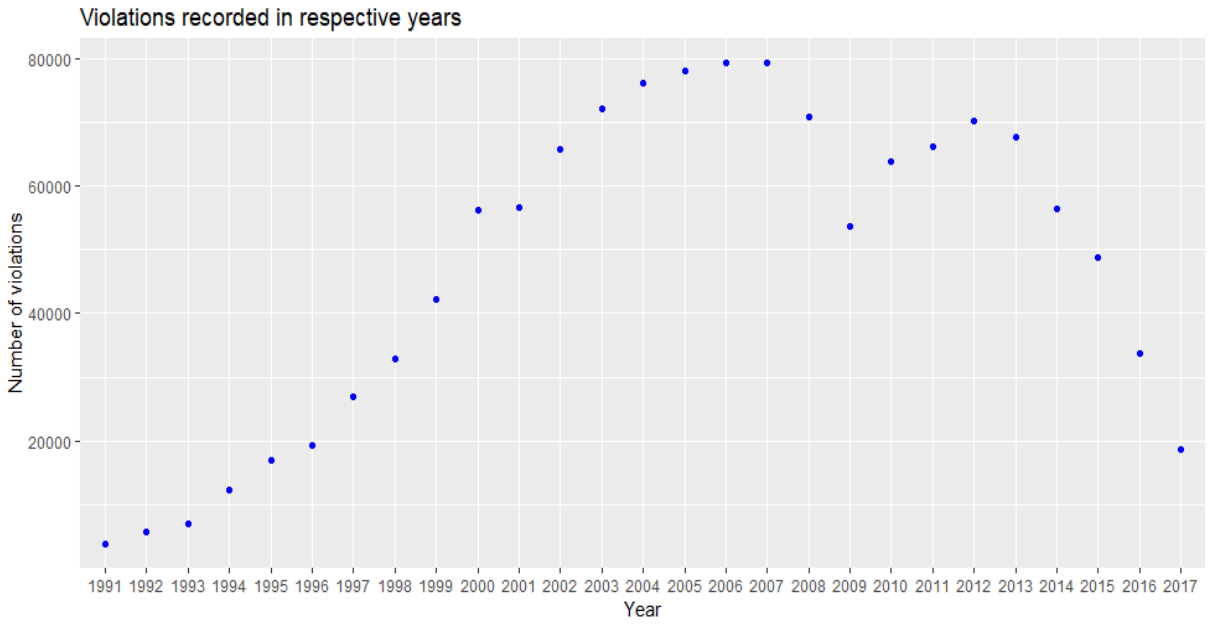
Interpretation of results from visualizations:

Representing the count of number of males, females and trans-genders who have violated traffic rules.



The above graph shows that more than 400000 females have violated the rules whereas more than 800000 males have violated the rules and less than 100000 trans-genders have violated the traffic rules.

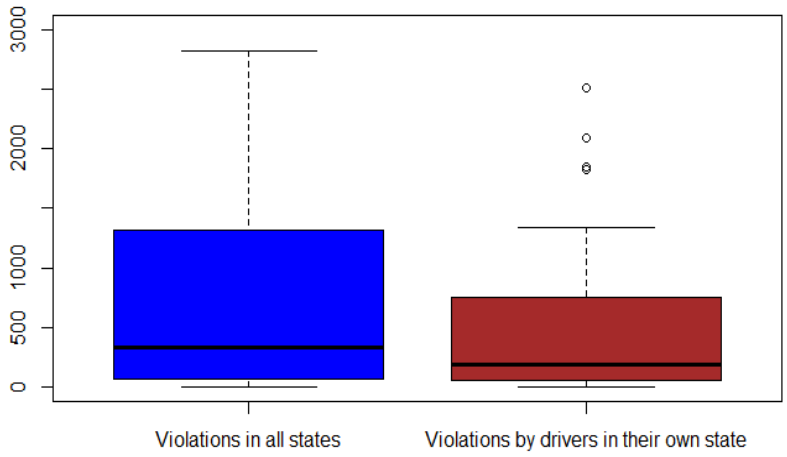
Representing number of violations recorded in respective years:



From the above graph you can observe an increasing trend from year 1991 to 2007 after which there is slight decrease in violations for 2 years and again increased for 3 years after which the number of violations have gradually decreased. The highest number of violations are recorded in the year 2007 whereas the least is recorded in the year 1991.

Representing violations in all the states and comparing them to violations by drivers in their own state:

Since the dataset don't contain any numerical values we need to perform some data processing and attain frequencies of violations occurred in each state and drivers state and merge the data frames based on the states and now make a box plot on the two columns, violations in all states (total violations by driver) and violations by drivers in their own state. The following boxplot is the output.

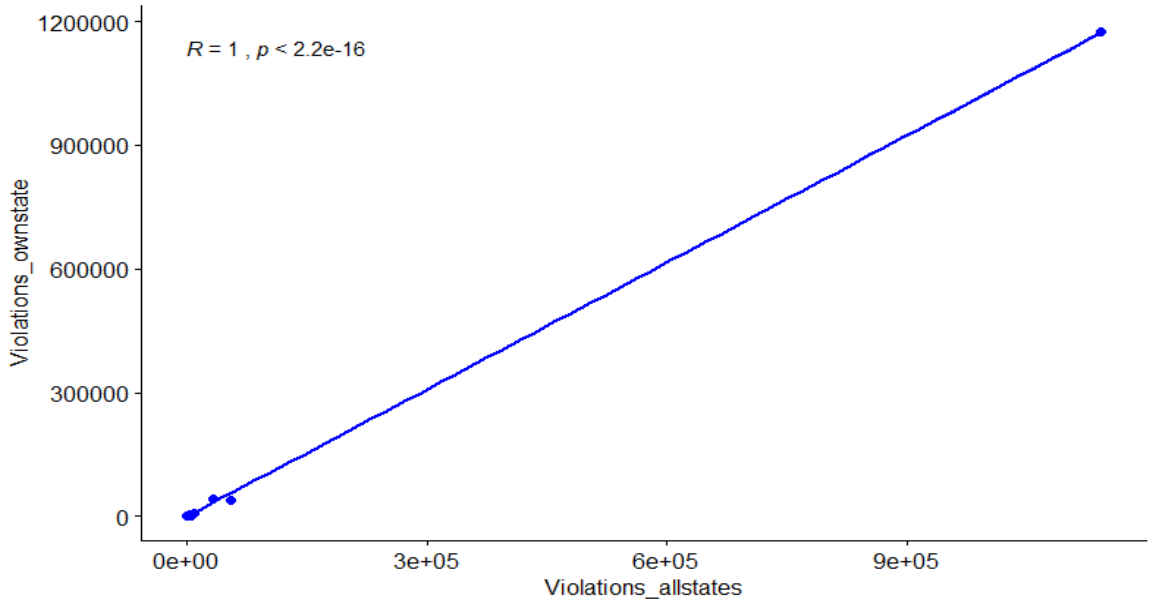


The above box plot tells us that the on an average most of the violations occurred in respective states are violated by the drivers who are out of the state.

Correlation analysis between above variables and hypothesis test:

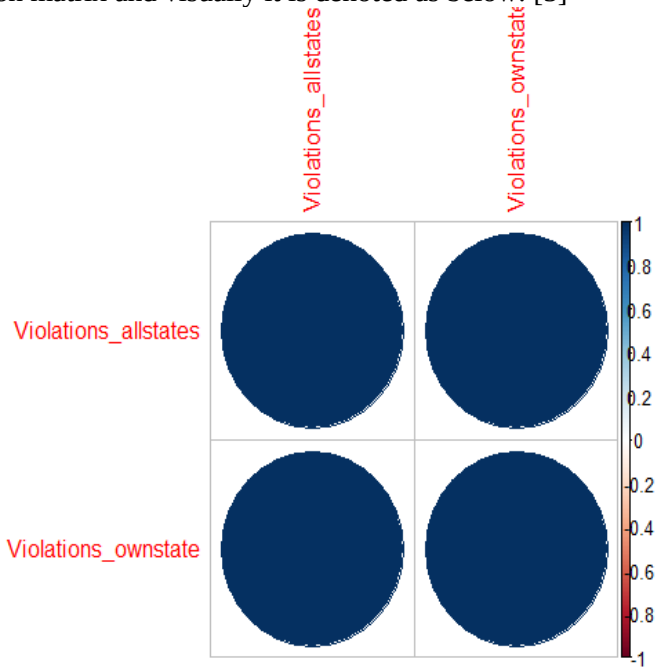
```
> cor.test(df$`Violations in all states`,df$`Violations by drivers in their
own state`)

        Pearson's product-moment correlation

data:  df$`Violations in all states` and df$`Violations by drivers in their
own state`
t = 496.94, df = 65, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9997852 0.9999194
sample estimates:
      cor
0.9998684

>
```

When we do a correlation test in R the above output is obtained. It says that the variables has 95 confidence interval with p-value < 2.26e-16 which is very minute, so we can reject the null hypothesis with the obtained confidence interval. The correlation coefficient is 0.9998684 which says that if one value increase by one unit the other increase by 0.9998684 times.
The null hypothesis here is the violations in own state and other states are independent. This can rejected as the p-value obtained is very less. So, we can tell that violations in own state and in different states are dependent.



Visualization of correlation analysis

Also the correlation between all the variables in data-frame df which is considered for correlation analysis is represented as correlation matrix and visually it is denoted as below. [5]



From this we can say that the two variables are equally correlated (i.e) Violations in all states and violations in their own state are equally correlated.

Logistic Regression model and hypothesis test: [2]

The regression model is generated between columns Accidents and Alcohol which gives a relation between violations of drunk and driving that has led to accidents.

```
> summary(glm(Accident~Alcohol,family = binomial(link=logit)))

Call:
glm(formula = Accident ~ Alcohol, family = binomial(link = logit))

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.657e+01  3.126e+02  -0.085    0.932
Alcohol     -9.222e-13  8.178e+03   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 1299924  degrees of freedom
Residual deviance: 7.5416e-06  on 1299923  degrees of freedom
```

The above model gives us the intercept (-2.657e+01) and coefficient of remaining variables (-9.222e-13) along with the Z-value and probability.
It also gives us the summary statistics of the model i.e mean, median, min, max, 1st quartile, 3rd quartile.
If you observe the p-value here it is very small, so we can say that it is unlikely that two variables Accident and Alcohol on which we have done regression model are independent (i.e) if a person is drunk it is more likely that he is involved in accident than not.

Conclusion:

The above analysis on traffic violations tells us that, most of the violations are by male with nearly more than 800000+ cases recorded in 10 years. Also we have visualized the trend of violations between 1991 and 2018 where we observed increasing trend till 2017 and then decreased gradually. The boxplot generated tells us that most of the violations recorded in respective states are due to drivers of other states, and also correlation tells us that these two are dependent. The logistic regression model generated tells us that the Accident occurred due to violation and violation of drunk and drive are dependent based on p-value obtained.

Technical terms:

Hypothesis test: "Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true. "[6]

Correlation coefficient: "The correlation coefficient is a statistical measure that calculates the strength of the relationship between the relative movements of the two variables. The range of values for the correlation coefficient bounded by 1.0 on an absolute value basis or between -1.0 to 1.0. "[7]

Appendix:

```
library(tidyverse)
library(ggplot2)
library(plyr)
library(corrplot)
data <- read.csv('F:/study/masters/1sem/ait-580/final project/Traffic_Violations.csv')

#remove empty rows and rows with NA's
data <- na.omit(data)

#select required columns
data <- subset(data,select=c(7:19,21,28:30,32))
data <- data.frame(data)

#Handling categorical variables i.e converting Yes ->1, No->0
data$Accident <- ifelse(data$Accident == "Yes",1,0)
data$Belts <- ifelse(data$Belts == "Yes",1,0)
data$Personal.Injury <- ifelse(data$Personal.Injury == "Yes",1,0)
data$Property.Damage <- ifelse(data$Property.Damage == "Yes",1,0)
data$Fatal <- ifelse(data$Fatal == "Yes",1,0)
data$Commercial.License <- ifelse(data$Commercial.License == "Yes",1,0)
data$HAZMAT <- ifelse(data$HAZMAT == "Yes",1,0)
data$Commercial.Vehicle <- ifelse(data$Commercial.Vehicle == "Yes",1,0)
data$Alcohol <- ifelse(data$Alcohol == "Yes",1,0)
data$Work.Zone <- ifelse(data$Work.Zone == "Yes",1,0)
data$Contributed.To.Accident <- ifelse(data$Contributed.To.Accident == "Yes",1,0)
attach(data)
summary(Accident)
summary(Belts)
summary(Personal.Injury)
summary(Property.Damage)
summary(Fatal)
summary(Commercial.License)
summary(HAZMAT)
summary(Commercial.Vehicle)
summary(Alcohol)
summary(Work.Zone)
summary(Contributed.To.Accident)

#count number of male,female violate dthe traffic rules
gen <- count(Gender)
#Generating visualizations
#ggplot(gen,aes(x=x,y=freq))+geom_point()
plot(Gender,col="Blue")+title(xlab='Gender',ylab = 'Count')

#How many traffic violations are recorded in respective years
```

```
yrs <- subset(data,Year>1990 & Year<2018)
filter_yrs <- yrs$Year
plt <- as.data.frame(table(filter_yrs))
ggplot(plt,aes(x=filter_yrs,y=Freq))+geom_point(col="blue")+labs(title ="Violations recorded in
respective years",x="Year",y="Number of violations")

#Box plot representing count of violations in all states compared to drivers violated in their own state
st <- as.data.frame(table(State))
drst <- as.data.frame(table(Driver.State))
st <- st[sample(1:nrow(st),69,replace= FALSE),]
names(drst)[1] <- paste("State")
df <- merge(st,drst,by="State")
df <- na.omit(df)
df <- df[,2:3]
names(df)[1] <- paste("Violations_allstates")
names(df)[2] <- paste("Violations_ownstate")
boxplot(df,ylim=c(0,3000),col=c("blue","brown"))

#correlation test & hypothesis test
cor.test(df$`Violations_allstates`,df$`Violations_ownstate`)
library(ggpubr)
ggscatter(df,x='Violations_allstates',y='Violations_ownstate',add = "reg.line",conf.int = TRUE,cor.coef =
TRUE,cor.method = "pearson",color = 'blue')

co_matrix <- cor(df)
corrplot(co_matrix)

#Generating logistic reression model & hypothesis test
summary(glm(Accident~Alcohol,family = binomial(link=logit)))
```

References

[1] Data.gov, "Traffic Violations," [Online]. Available: https://catalog.data.gov/dataset/traffic-violations-56dda.

[2] A. Gupta, "Logistic Regression output interpretation in R," [Online]. Available: https://analyticsdataexploration.com/logistic-regression-output-interpretation-in-r/.

[3] Data.gov, "Aims of Data.gov," [Online]. Available: https://www.google.com/search?q=What+does+data.ggov+do&oq=What+does+data.ggov+do&aqs=chrome..69i57.6926j1j7&sourceid=chrome&ie=UTF-8.

[4] Data.gov, "It's all about the data," [Online]. Available: https://www.data.gov/developers/blog/its-all-about-data.

[5] STHDA, "Visualize correlation matrix using correlogram," [Online]. Available: http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram.

[6] Wolfarm, "Hypothesis Testing," [Online]. Available: http://mathworld.wolfram.com/HypothesisTesting.html.

[7] INVESTOPEDIA, "Correlation coefficient," [Online]. Available: https://www.investopedia.com/terms/c/correlationcoefficient.asp.