ANALYSIS ON TRAFFIC FATALITIES IN THE STATE OF VIRGINIA

JINU KINGCY SEBASTIN

GEORGE MASON UNIVERSITY

STAT 515: APPLIED STATISTICS AND VISUALIZATION FOR ANALYTICS

Abstract

Based on the analysis of data from the United States Department of Transportation's [DOT] Fatality Analysis Reporting System [FARS] the statistics on death rate caused by traffic interprets the fact that fatal crashes are relatively low in Virginia when compared with the 50 states of the U.S. (NHTSA, 2016). Despite the low death rate that the State encounters, Virginia has seen a spike in the total number of fatal crashes over the last few years. The statistics reports show evidence that the Central and Northern region of Virginia has encountered the maximum traffic crashes and fatalities whereas the Southwest and Eastern regions have the least count of traffic death rates. The major components that are analyzed in the causes for traffic related deaths are alcohol consumption, speed crashing and not obeying the state law of buckling up while driving. A regression model is advantageous to predict the variables that are important and are considered to be the leading factors for traffic accidents in the state of Virginia. Conducting awareness programs and educating the public on driving safely plays a vital role in bringing down the number of road accidents and death toll.

ANALYSIS ON TRAFFIC FATALITIES IN THE STATE OF VIRGINIA

Over the past few years, there has been a steady increase in the death rate caused due to traffic accidents in the commonwealth of Virginia. Although there are various elements to be considered for determining the leading cause of traffic related fatalities such as road type, single/multiple collision, driver's age, driver's distraction, injuries on major holidays, etc., the principal factors are Alcohol related crashes, Over speeding and Unrestrained occupants. Virginia has 95 counties and 38 independent cities, composing a total of 133 counties/cities with 8 regions namely, Central, Eastern, Northern, Southside, Southwest, Hampton roads, Valley and West central (see Appendix 2 Figure 13 for Virginia map).

Analyzing the article with misleading information

The article on "Traffic fatalities: How is Virginia doing?" (Peter, 2016) illustrates on the regions of Virginia facing the highest rate of traffic related accidents with Southside ranking first followed by Eastern and the least accidents recorded were in Northern Virginia.

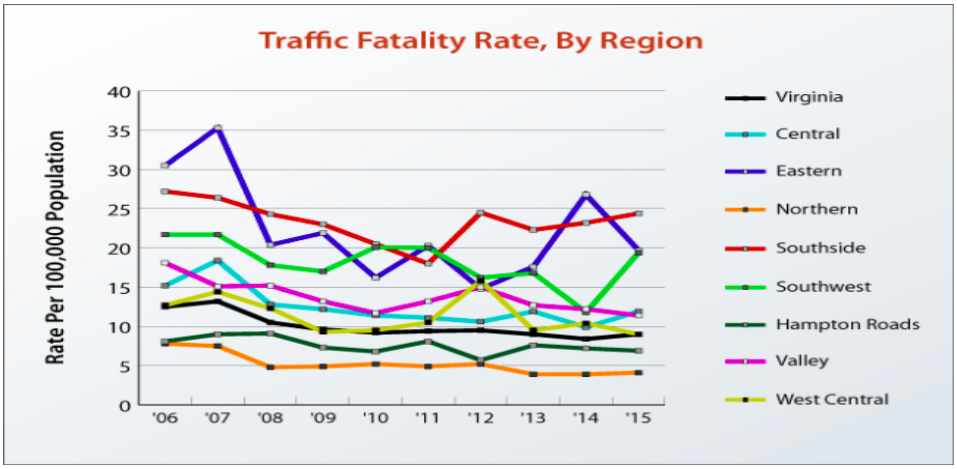

Figure 1. Line chart representation of traffic fatality rate in Virginia by region

On further analysis of the statistics report (VHSO, 2016), it is regarded that the Central and Northern regions of Virginia have suffered the highest fatality rates with the lowest death rates reported in the Eastern region. The data from the article may have missed a detailed study

on the factors causing traffic fatality by regions, thereby conveying a misleading visualization for the readers. Hence, the need for a redesigned graph with current statistics and data analysis is necessary to comprehend the trends of traffic death rates in Virginia by region.

The software used for analysis and Data visualization is RStudio Desktop version 1.1.447 and some of the R packages used are hexbin, ggrepel, reshape2, corrplot, ggplot2, dplyr, lattice and randomForest.

Implementation of redesigned graph

The data on traffic crashes, injuries and fatalities for important components resulting in death such as Alcohol, Speed, Unrestrained and Motorcycle is collected, analyzed and visualized using different packages in R (see Appendix 2 Figure 14 for Dataset column heads).

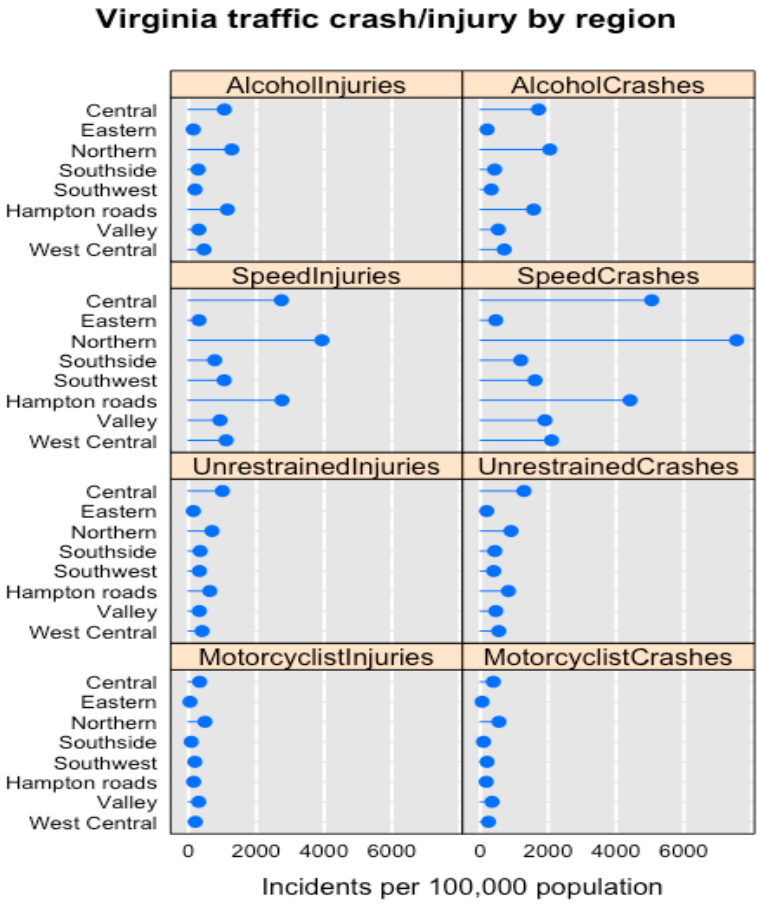

Figure 2. Dot plot representation of traffic crashes and injuries in Virginia by region

The incidents are recorded per 100,000 population in all the 8 regions of Virginia. Using lattice package in RStudio, a dot plot was depicted to study and visualize the rate of crashes and injuries in Virginia by region.

From the graph, we can clearly see that alcohol and speed related injuries/crashes are high in Northern region followed by Central and Hampton roads. While unrestrained injuries and crashes are higher in the Central region and has the least count of similar occurrences in the Eastern part of the state. Northern Virginia has the highest Motorcyclist injuries and crashes in comparison to other regions.

Virginia traffic fatalities based on factors. The traffic fatalities based on the major factors for each region in Virginia is portrayed using a dot plot in RStudio. The graph explains that deaths caused due to Speed, Alcohol and Unrestrained occupants are more in the Central Virginia, while the least of these incidents took place in the Eastern region.

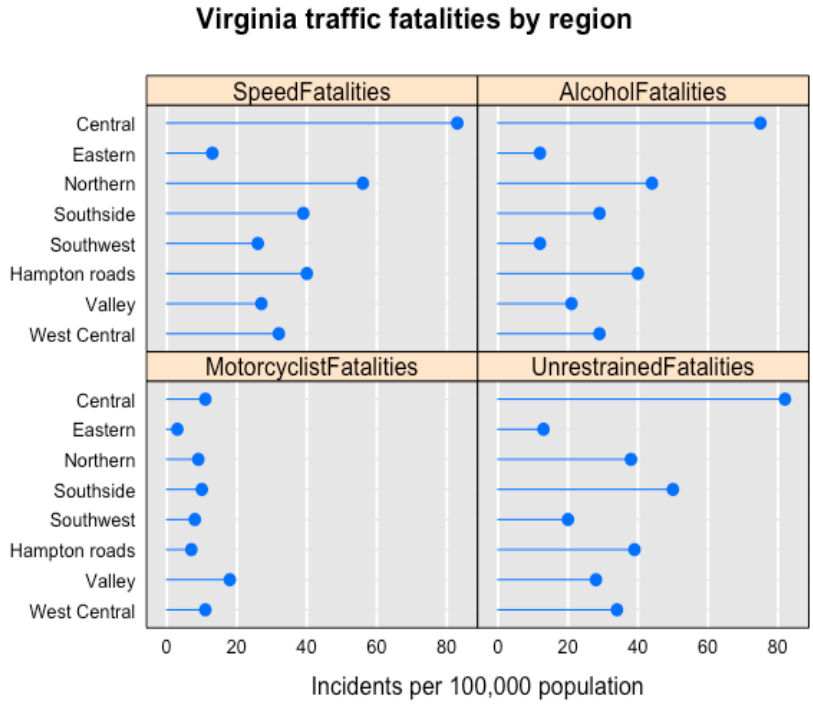

Figure 3. Dot plot representation of traffic fatalities in Virginia by region

Total traffic fatalities. A box plot in R studio with ggplot2 package is used to identify and visualize the total fatality due to traffic in all the regions of Virginia. It is evident that Central region has the highest death rate followed by Northern Virginia and Eastern region had the lowest traffic casualties. A possible reason for high death rate in the Central and Northern Virginia could be because of the Interstate route I-95 that runs along those regions which is highly populated with vehicles since it's the route that connects different states to the north and south bound of the east coast. Eastern region is safer and has record low level traffic fatalities may be due to the fact that it's a shore area and most of the public would prefer ferries and boat rides for commutations.

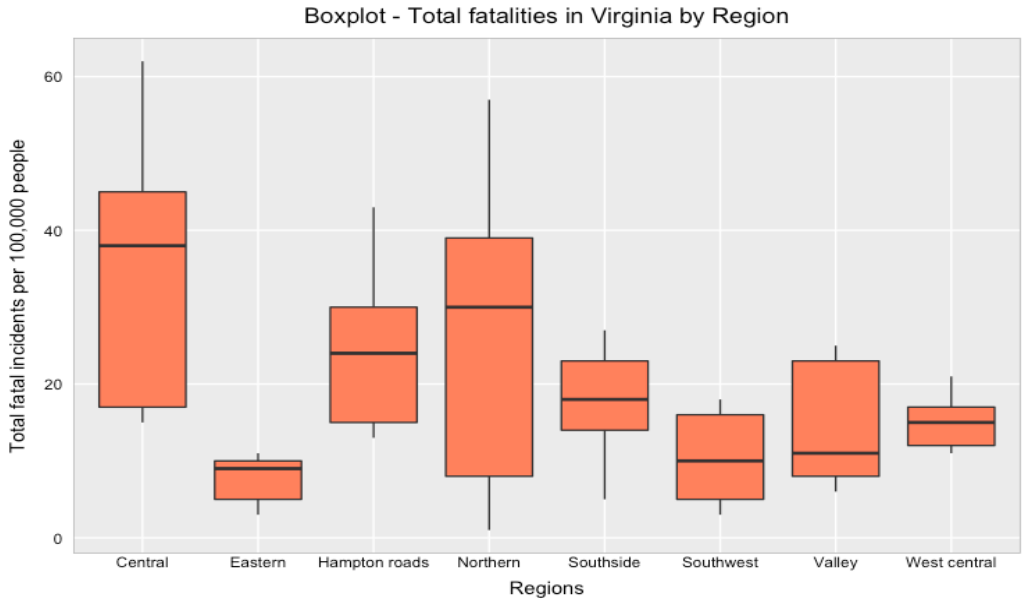

Figure 4. Box plot representation of total fatality rate in Virginia by region

Correlation between various factors leading to death. It is important to determine the mutual relationship between different variables that are leading cause for traffic related death. The correlation matrix plot obtained using the package corrplot in RStudio is used to identify the pair of variables having the highest correlation. From the plot in Figure 5, it is obvious that

unrestrained injuries and unrestrained crashes are highly correlated signifying the fact that passengers that involve in unrestrained crashes are more likely to be injured. Similarly, motorcyclist crashes are highly correlated to motorcyclist injuries representing the fact that motorcyclists involved in crashes are likely to be injured.

The benefits of using a correlation matrix is that it becomes easier to understand the strength of the relationship between any two variables from the data which could not have been inferred directly on the first place.

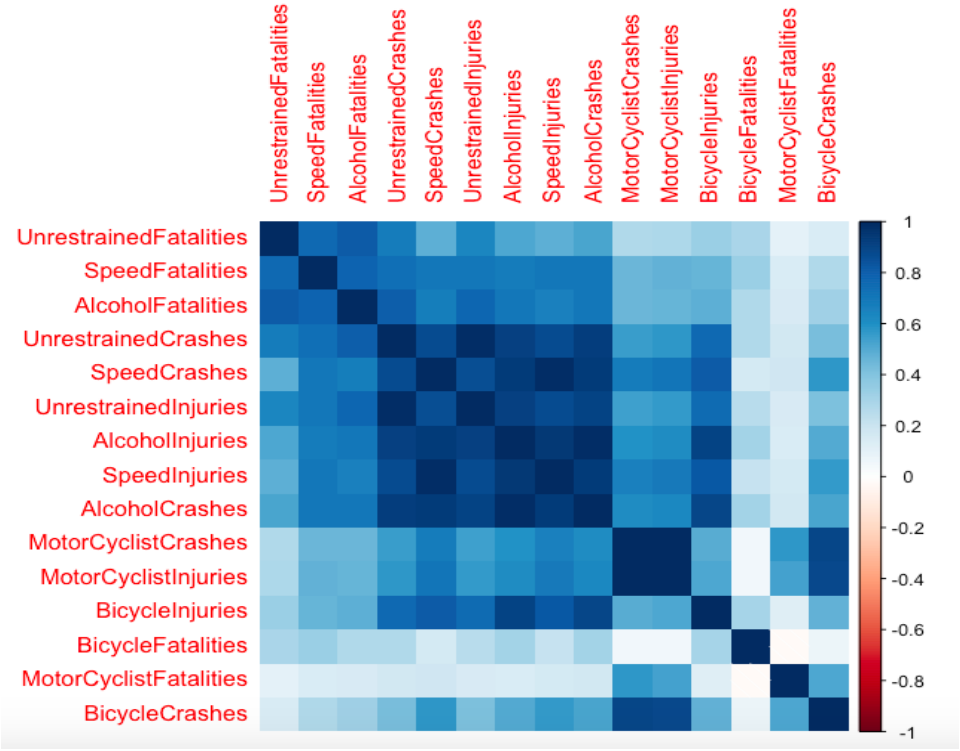

Figure 5. Correlation matrix plot representing the relationship between variables

Labeling counties with the highest and lowest crashes. The highest and lowest ranking counties with total crashes in all the regions can be viewed in scatterplots using the ggrepel package in RStudio. From the plots, we can observe that in the Central region Henrico county has the highest crash rate and Charles city county has the lowest crash rate. Likewise, in

Northern Virginia, Fairfax county has the highest crash rate and Manassas Park city has the lowest. In the Hampton roads, Virginia Beach city ranks first in the number of traffic crashes and Poquoson city stands last.

Similarly, comparisons can be made between the counties/cities from other regions as well. Fredrick county has the highest number of total crashes in the Valley region whereas Buena Vista city has record low crash rate. Also, from the graphs it is vivid that Fairfax county has recorded the highest number of traffic crashes in the state. A general logic may be that Fairfax county is closer in proximity to the D.C area which is overcrowded by government workers, tourists and jammed by traffic. The scatterplot using loess method is essentially used in regression analysis (Stephanie, 2013) creating a smooth line along the scatterplots to understand the trends and view the relationship between the variables.

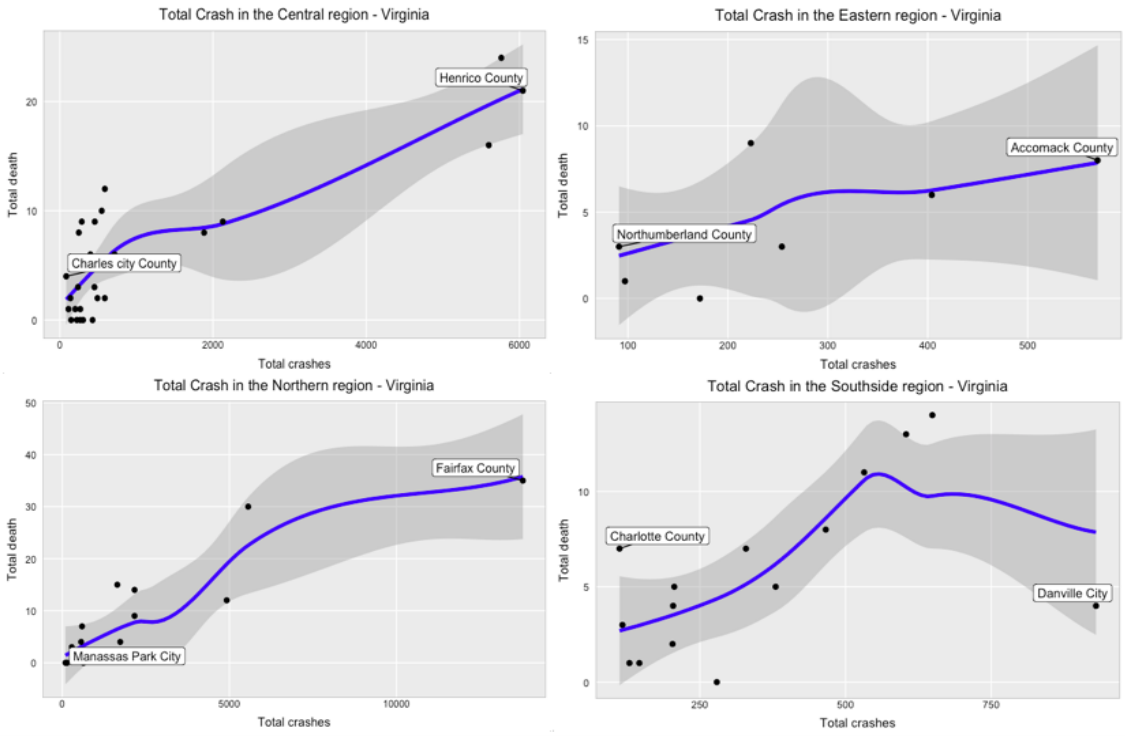

Figure 6. Scatterplot with labels using loess method for Central, Eastern, Northern and Southside regions

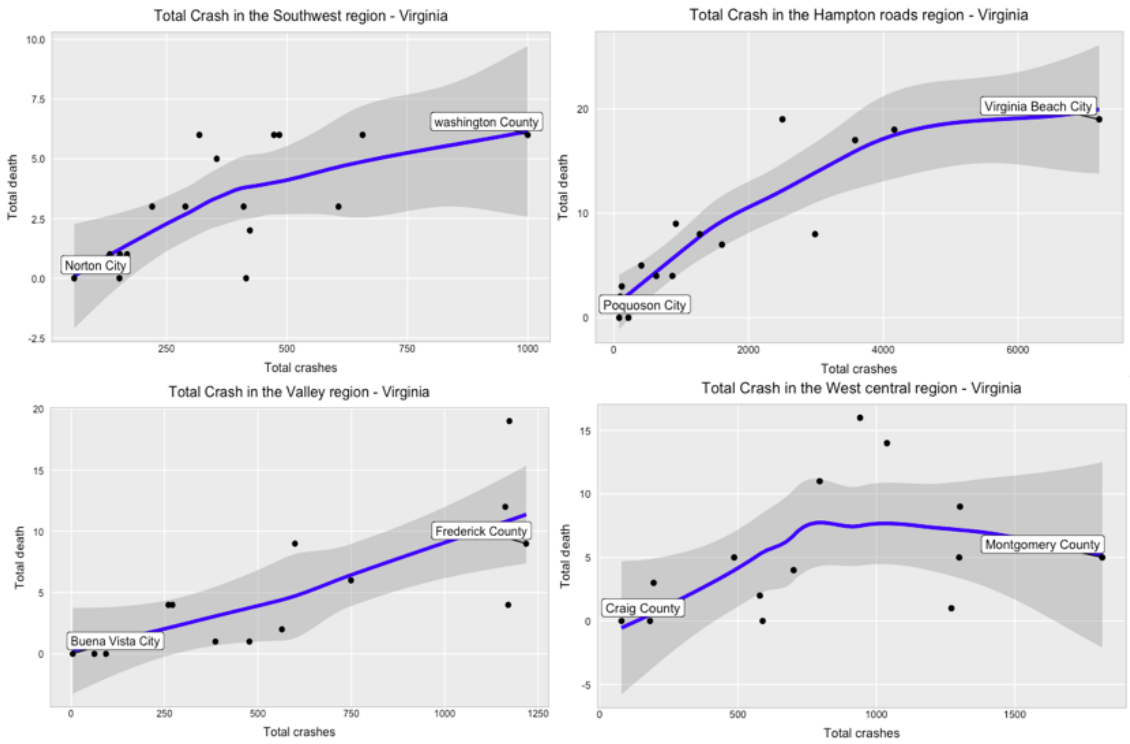

Figure 7. Scatterplot with labels using loess method for Southwest, Hampton roads, Valley and West central regions

Variable Selection method. The purpose of using variable selection technique is to construct a model that predicts well or explains the relationships in the data. Three popular selection methods include backward elimination, forward selection and stepwise selection. For determining the traffic fatalities, we have used backward elimination method for selecting the predictor variables and response variable that best explains the correlations in the data while it drops those variables that has the lowest AIC (Akaike Information Criterion) value one at a time. This process is helpful in selecting the variables that contribute the most to the model.

```
> tra.lm = lm(TotalFatalities~.,data=traffic)
> step(tra.lm,scope=formula(tra.lm),direction = "backward")
```

Figure 8. R script for backward elimination variable selection method

Scatterplot matrix of selected variables. After performing variable selection method, the selected variables are plotted as a scatterplot matrix with hexagon binning and smoothes using the hexbin package in RStudio. This plot helps in identifying the correlation between the independent and the dependent variables such as a fairly strong relationship between total fatalities and unrestrained injuries can be seen. Similarly, a weak relationship between total fatalities and Bicycle crashes can be witnessed.

Using smoothes helps to visualize the trends of each variable in a better manner. A scatterplot matrix using all the variables in the dataset is also plotted (see Appendix 2 Figure 15) to understand the relationship between all the dependent variables. A downside of using all the variables is that the model may not be meaningful as it comprises also of weak correlations which are not required for prediction modeling.

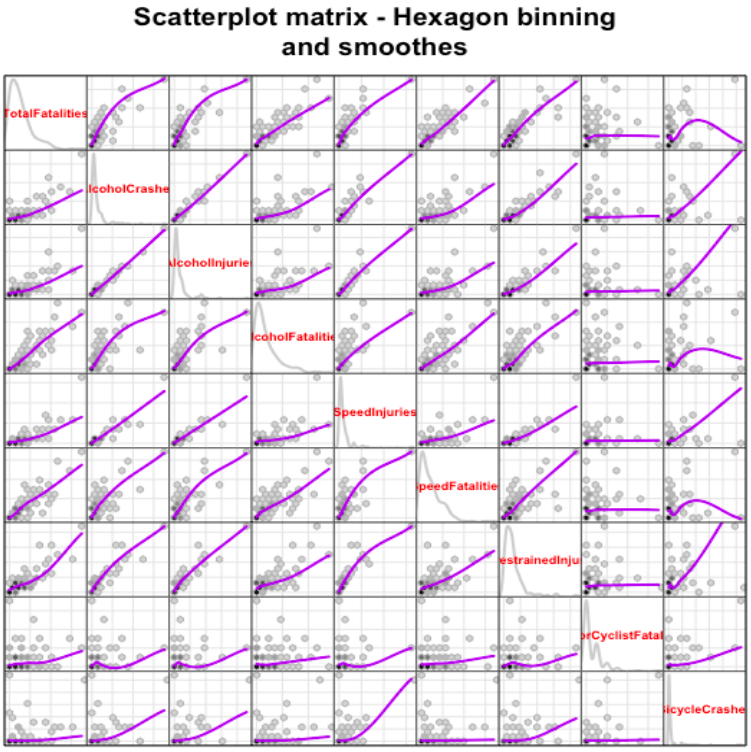

Figure 9. Scatterplot matrix hexagon binning with smoothes of selected variables

For this purpose, variable selection method comes in handy to pick only those predictors and response variable that are a best fit for the model.

Regression using Random Forest algorithm. In order to measure and predict the relative importance of each variable, it is essential to run a regression model which in our scenario we use random forest algorithm. Using the randomForest package in RStudio we can build a predicting model that predicts the factors which contribute most to the total accidents caused in the state of Virginia. The dataset was partitioned in to 60% as training set and 40% as test set. The results obtained are depicted in the figure below.

```
Call:
 randomForest(formula = TotalFatalities ~ ., data = train, mtry = 5,      importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 5

          Mean of squared residuals: 14.6933
                    % Var explained: 63.36
```

Figure 10. MSE and % Variance explained obtained using random Forest

The test set MSE for the data obtained was 8.061. The function importance() in R provides the importance of each variable.

```
> importance(rf.traffic)
                         %IncMSE IncNodePurity
AlcoholCrashes         9.0169069     279.92816
AlcoholInjuries        7.4448300     328.93808
SpeedCrashes          13.8478388     443.85462
SpeedInjuries          8.1879784     283.77986
UnrestrainedCrashes   18.4772723     816.44763
UnrestrainedInjuries  11.4035411     596.74383
MotorCyclistCrashes    2.3127598      61.91826
MotorCyclistInjuries   0.2476721      56.73952
BicycleCrashes         1.9582337      61.12856
BicycleInjuries        0.4513618      64.29227
```

Figure 11. Importance plot with importance of each variable

While the varImpPlot() function is used to plot the importance (James, Witten, Hastie & Tibshirani, 2013,p.330). From our results, we see that unrestrained crashes, unrestrained injuries and speed crashes promote the major cause for total accidents in Virginia.

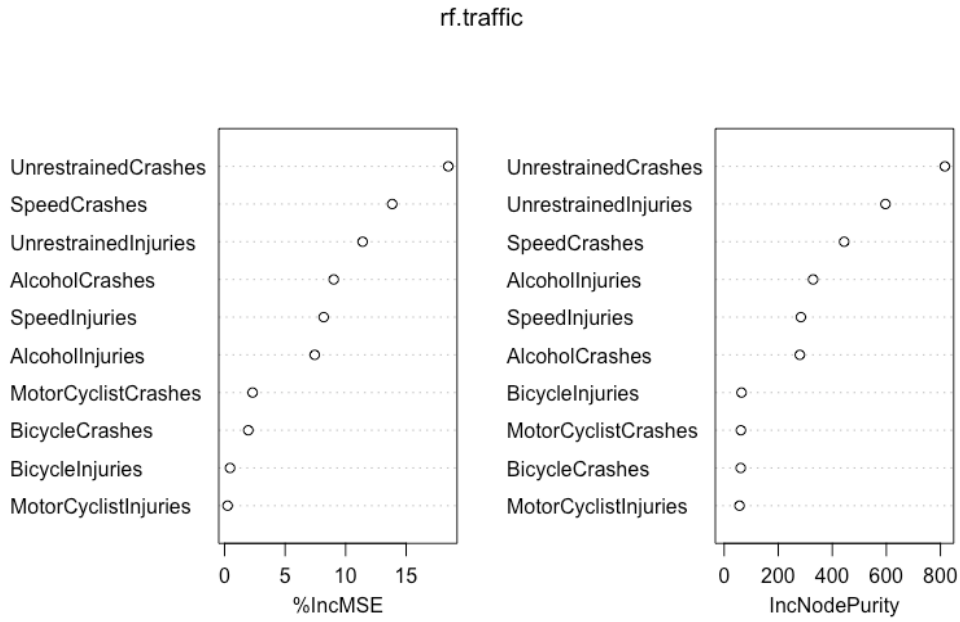

Figure 12. Variable Importance Plot using random forest algorithm

Challenges. The datasets obtained from the Virginia Highway Safety Office and National Highway Traffic Safety Administration contains data on traffic fatalities for the state of Virginia for all the counties/cities under the tab Jurisdiction. In order to understand our data by regions, I had to meticulously breakdown the counties with their data region-wise that required more effort. Including the picture of Virginia map with their regions assisted in understanding the regions where maximum crashes occurred.

Another challenging part was when I tried to build the regression model with few variables, (initially my dataset had 10 attributes) my model was not accurate, and the results were biased (see Appendix 2 Figure 16). I decided to add few more variables that contributed to

the total traffic fatality rate from the Department of Motor Vehicles Crash reports which was originally obtained from FARS and NHTSA to improve the model and make it more meaningful. Running the random forest algorithm for my updated dataset and selected variables produced a good prediction model that was unbiased, purposeful and relevant.

Summary. The roads of Virginia also play a vital role in the traffic death rates rising in the state. Rural areas have roads that are narrow and winding with relatively low traffic creating a hazardous environment for drivers traveling above the posted speed limits. Identifying the reason for traffic casualties in the state rose as a hypothesis. Using a regression model such as random forest, predicted that crashes and injuries caused by failure to restraint or buckle up while driving is the major cause contributing to the death toll in Virginia followed by accidents caused due to uncontrolled speeding.

References

Peter, C. (2016). *Traffic fatalities: How is Virginia doing?* Retrieved from http://vaperforms.virginia.gov/indicators/publicSafety/trafficFatalities.php

Kodali, T. (2016). R-bloggers. *Predicting wine quality using Random Forests*. Retrieved from https://www.r-bloggers.com/predicting-wine-quality-using-random-forests/

Wickham, H., Grolemund, G. (2016). R for Data Science. *Graphics for Communication with ggplot2 (p.446-448)*

James, G., Witten, D., Hastie,T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. *Tree-Based Methods (p.320-330)*

Stephanie (2013). Statistics How To. *Lowess Smoothing in Statistics: What is it?* Retrieved from http://www.statisticshowto.com/lowess-smoothing/

Insurance Institute for Highway Safety (2016). *General Statistics: Fatality Facts*. Retrieved from http://www.iihs.org/iihs/topics/t/general-statistics/fatalityfacts/state-by-state-overview

Department of Motor Vehicles: Virginia Highway Safety Office (2016). *Commonwealth of Virginia Traffic Crash Facts.* Retrieved from https://www.dmv.virginia.gov/safety/crash_data/crash_facts/crash_facts_16.pdf

National Highway Traffic Safety Administration (2016). *Traffic Safety performance measures for Virginia.* Retrieved from https://cdan.nhtsa.gov/SASStoredProcess/guest

Appendix 1

R script

```
library(lattice)
library(ggplot2)
library(hexbin)
library(ggrepel)
library(reshape2)
library(corrplot)
library(randomForest)
traffic <- read.csv(file="TrafficFatality.csv", header=TRUE, row.names=1)
region <- c("Central","Eastern","Northern","Southside","Southwest","Hampton
roads","Valley","West Central")
 reason <-
c("AlcoholCrashes","AlcoholInjuries","SpeedCrashes","SpeedInjuries","UnrestrainedCra
shes","UnrestrainedInjuries","MotorcyclistCrashes","MotorcyclistInjuries")
 matt <-
cbind(AlcoholCrashes,AlcoholInjuries,SpeedCrashes,SpeedInjuries,UnrestrainedCrashes,
UnrestrainedInjuries,MotorcyclistCrashes,MotorcyclistInjuries)
 colnames(matt) <- reason
 rownames(matt) <- region
dotplot(matt,groups=FALSE,
        layout=c(2,4),aspect=.7,
       origin=0,type=c("p","h"),
        main="Virginia traffic fatality",
        xlab="Accidents per 100,000 population",
        scales=list(x=list(tck=0, alternating=FALSE)),
       panel=function(...){
         panel.fill(rgb(.9,.9,.9))
           panel.grid(h=0,v=-1,col="white",lwd=2)
         panel.dotplot(col=rgb(0,.5,1),cex=1.1,...)
```

```
        }
)
dotplot(matt[8:1,8:1],groups=FALSE,
       layout=c(2,4),aspect=.7,
      origin=0,type=c("p","h"),
        main="Virginia traffic crash/injury by region",
       xlab="Incidents per 100,000 population",
        scales=list(x=list(tck=0, alternating=FALSE)),
       panel=function(...){
       panel.fill(rgb(.9,.9,.9))
       panel.grid(h=0,v=-1,col="white",lwd=2)
        panel.dotplot(col=rgb(0,.5,1),cex=1.1,...)+hw
     }
)
reason <-
c("AlcoholFatalities","SpeedFatalities","UnrestrainedFatalities","MotorcyclistFatalities")
matt <-
cbind(AlcoholFatalities,SpeedFatalities,UnrestrainedFatalities,MotorcyclistFatalities)
colnames(matt) <- reason
rownames(matt) <- region
matt
dotplot(matt,groups=FALSE,
       layout=c(2,2),aspect=.7,
       origin=0,type=c("p","h"),
       main="Virginia traffic fatality",
       xlab="Accidents per 100,000 population",
       scales=list(x=list(tck=0, alternating=FALSE)),
       panel=function(...){
        panel.fill(rgb(.9,.9,.9))
        panel.grid(h=0,v=-1,col="white",lwd=2)
        panel.dotplot(col=rgb(0,.5,1),cex=1.1,...)
```

```
        }
)
dotplot(matt[8:1,4:1],groups=FALSE,
        layout=c(2,2),aspect=.7,
        origin=0,type=c("p","h"),
        main="Virginia traffic fatalities by region",
        xlab="Incidents per 100,000 population",
        scales=list(x=list(tck=0, alternating=FALSE)),
        panel=function(...){
          panel.fill(rgb(.9,.9,.9))
          panel.grid(h=0,v=-1,col="white",lwd=2)
          panel.dotplot(col=rgb(0,.5,1),cex=1.1,...)+hw
        }
)
virginia <- read.csv("Book3.csv")
ggplot(data=virginia,aes(x=Regions,y=Total))+ geom_boxplot(fill="salmon1")+
  ylab(label = "Total fatal incidents per 100,000 people")+
  xlab("Regions") + labs(title="Boxplot - Total fatalities in Virginia by Region")+hw
traffic <- read.csv(file="TrafficFatality.csv", header=TRUE, row.names=1)
dim(traffic)
colnames(traffic)
set.seed(37)
trafficRf <- randomForest(x = traffic[ , 3:17], y=traffic[, 20],
                          importance=TRUE,  proximity=FALSE, ntree=500,
                          keepForest=TRUE)
imp <- importance(trafficRf)
n <- 15
ordr1 <- order(imp[, 1],decreasing=TRUE)
name1 <- row.names(imp[ordr1, ])[1:n]
ordr2 <- order(imp[, 2],decreasing=TRUE)
name2 <- row.names(imp[ordr2,])[1:n]
```

```
varyName <- union(name1,name2)
checkCory <- round( cor(traffic[,varyName],
                        method="spearman"),2)
cor(traffic[,varyName])
M <- cor(traffic[,varyName])
corrplot(M, method = "shade")
tra.lm = lm(TotalFatalities~.,data=traffic)
step(tra.lm,scope=formula(tra.lm),direction = "backward")
splom(traffic[,c(20,3:5,7,8,10,14,15)], as.matrix = TRUE,
      xlab = '',main = "Scatterplot matrix - Hexagon binning\n and smoothes ",
      pscale = 0, varname.col = "red",
      varname.cex = 0.56, varname.font = 2,
      axis.text.cex = 0.4, axis.text.col = "red",
      axis.text.font = 2, axis.line.tck = .5,
      panel = function(x,y,...) {
        panel.grid(h = -1,v = -1,...)
        panel.hexbinplot(x,y,xbins = 12,...,
                         border = gray(.7),
                         trans = function(x)x^1)
        panel.loess(x , y, ...,
                    lwd = 2,col = 'purple')
      },
      diag.panel = function(x, ...){
        yrng <- current.panel.limits()$ylim
        d <- density(x, na.rm = TRUE)
        d$y <- with(d, yrng[1] + 0.95 * diff(yrng) * y / max(y) )
        panel.lines(d,col = gray(.8),lwd = 2)
        diag.panel.splom(x, ...)
      }
)
new1 <- read.csv("central.csv")
```

```
plt1 <-  ggplot(new1,
              aes(x = TotalCrash, y = TotalFatality)) +
  geom_smooth(method="loess",span=.90,method.args=list(degree=1),
            size=1.5,color="blue") +
  geom_point(shape=20,size=3,color="black",fill="red") +
  labs(x="Total crashes",
      y="Total death",
      title="Total Crash in the Central region - Virginia") + hw
ptLabs1 <- new1 %>% filter(Jurisdiction %in%
                         c('Henrico County','Charles city County'))
plt1 + ggrepel::geom_label_repel(data = ptLabs1,
                             aes(label = Jurisdiction),
                             nudge_y = .45)
new2 <- read.csv("eastern.csv")
plt2 <-  ggplot(new2,
              aes(x = TotalCrash, y = TotalFatality)) +
  geom_smooth(method="loess",span=.90,method.args=list(degree=1),
            size=1.5,color="blue") +
  geom_point(shape=20,size=3,color="black",fill="red") +
  labs(x="Total crashes",
      y="Total death",
      title="Total Crash in the Eastern region - Virginia") + hw
ptLabs2 <- new2 %>% filter(Jurisdiction %in%
                         c('Accomack County','Northumberland County'))
plt2 + ggrepel::geom_label_repel(data = ptLabs2,
                             aes(label = Jurisdiction),
                             nudge_y = .45)
new3 <- read.csv("northern.csv")
plt3 <-  ggplot(new3,
              aes(x = TotalCrash, y = TotalFatality)) +
  geom_smooth(method="loess",span=.90,method.args=list(degree=1),
```

```
            size=1.5,color="blue") +
  geom_point(shape=20,size=3,color="black",fill="red") +
  labs(x="Total crashes",
       y="Total death",
       title="Total Crash in the Northern region - Virginia") + hw
ptLabs3 <- new3 %>% filter(Jurisdiction %in%
                           c('Fairfax County','Manassas Park City'))
plt3 + ggrepel::geom_label_repel(data = ptLabs3,
                                 aes(label = Jurisdiction),
                                 nudge_y = .45)
new4 <- read.csv("southside.csv")
plt4 <-  ggplot(new4,
                aes(x = TotalCrash, y = TotalFatality)) +
  geom_smooth(method="loess",span=.90,method.args=list(degree=1),
            size=1.5,color="blue") +
  geom_point(shape=20,size=3,color="black",fill="red") +
  labs(x="Total crashes",
       y="Total death",
       title="Total Crash in the Southside region - Virginia") + hw
ptLabs4 <- new4 %>% filter(Jurisdiction %in%
                           c('Danville City','Charlotte County'))
plt4 + ggrepel::geom_label_repel(data = ptLabs4,
                                 aes(label = Jurisdiction),
                                 nudge_y = .45)
new5 <- read.csv("southwest.csv")
plt5 <-  ggplot(new5,
                aes(x = TotalCrash, y = TotalFatality)) +
  geom_smooth(method="loess",span=.90,method.args=list(degree=1),
            size=1.5,color="blue") +
  geom_point(shape=20,size=3,color="black",fill="red") +
  labs(x="Total crashes",
```

```
      y="Total death",
      title="Total Crash in the Southwest region - Virginia") + hw
ptLabs5 <- new5 %>% filter(Jurisdiction %in%
                           c('washington County','Norton City'))
plt5 + ggrepel::geom_label_repel(data = ptLabs5,
                                 aes(label = Jurisdiction),
                                 nudge_y = .45)
new6 <- read.csv("hampton.csv")
plt6 <-  ggplot(new6,
                aes(x = TotalCrash, y = TotalFatality)) +
  geom_smooth(method="loess",span=.90,method.args=list(degree=1),
              size=1.5,color="blue") +
  geom_point(shape=20,size=3,color="black",fill="red") +
  labs(x="Total crashes",
      y="Total death",
      title="Total Crash in the Hampton roads region - Virginia") + hw
ptLabs6 <- new6 %>% filter(Jurisdiction %in%
                           c('Virginia Beach City','Poquoson City'))
plt6 + ggrepel::geom_label_repel(data = ptLabs6,
                                 aes(label = Jurisdiction),
                                 nudge_y = .45)
new7 <- read.csv("valley.csv")
plt7 <-  ggplot(new7,
                aes(x = TotalCrash, y = TotalFatality)) +
  geom_smooth(method="loess",span=.90,method.args=list(degree=1),
              size=1.5,color="blue") +
  geom_point(shape=20,size=3,color="black",fill="red") +
  labs(x="Total crashes",
      y="Total death",
      title="Total Crash in the Valley region - Virginia") + hw
ptLabs7 <- new7 %>% filter(Jurisdiction %in%
```

```
                        c('Frederick County','Buena Vista City'))
plt7 + ggrepel::geom_label_repel(data = ptLabs7,
                            aes(label = Jurisdiction),
                            nudge_y = .45)
new8 <- read.csv("westcentral.csv")
plt8 <-  ggplot(new8,
              aes(x = TotalCrash, y = TotalFatality)) +
  geom_smooth(method="loess",span=.90,method.args=list(degree=1),
            size=1.5,color="blue") +
  geom_point(shape=20,size=3,color="black",fill="red") +
  labs(x="Total crashes",
       y="Total death",
       title="Total Crash in the West central region - Virginia") + hw
ptLabs8 <- new8 %>% filter(Jurisdiction %in%
                        c('Montgomery County','Craig County'))
plt8 + ggrepel::geom_label_repel(data = ptLabs8,
                            aes(label = Jurisdiction),
                            nudge_y = .45)
trafficRan <- read.csv(file="TrafficRFdata.csv", header=TRUE, row.names=1)
dim(trafficRan)
colnames(trafficRan)
set.seed(37)
trafficRf <- randomForest(x = trafficRan[ , 1:10], y=trafficRan[, 11],
                        importance=TRUE,  proximity=FALSE, ntree=500,
                        keepForest=TRUE)
trafficRf
imp <- importance(trafficRf)
varImpPlot(trafficRf,cex=.8)
n <- 10
ordr1 <- order(imp[, 1],decreasing=TRUE)
name1 <- row.names(imp[ordr1, ])[1:n]
```

```
ordr2 <- order(imp[, 2],decreasing=TRUE)
name2 <- row.names(imp[ordr2,])[1:n]
varyName <- union(name1,name2)
checkCory <- round( cor(trafficRan[,varyName],
                    method="spearman"),2)
checkCory
samp <- sample(nrow(trafficRan),0.6*nrow(trafficRan))
train <- trafficRan[samp,]
test <- trafficRan[-samp,]
model <- randomForest(TotalFatalities~.,data=train,mtry=10,importance=TRUE)
model
varImpPlot(model)
pred <- predict(model,newdata = test)
plot(pred,test$TotalFatalities)
abline(0,1)
mean((pred-test$TotalFatalities)^2)
set.seed(37)
rf.traffic <- randomForest(TotalFatalities~.,data=train,mtry=5,importance=TRUE)
rf.traffic
varImpPlot(rf.traffic)
test.rf <- predict(rf.traffic, newdata = test)
plot(test.rf,test$TotalFatalities)
abline(0,1)
mean((test.rf-test$TotalFatalities)^2)
```

Appendix 2

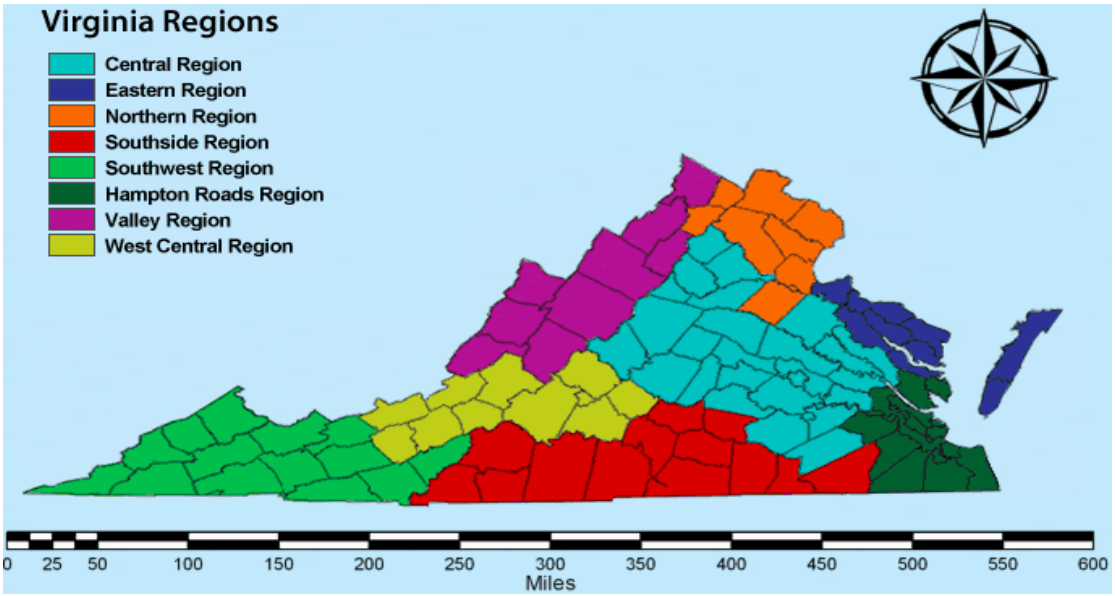

Figure 13. Virginia map by regions. Source: Google images

- Licensed Drivers
- Death rate/1000 drivers
- Alcohol crashes
- Alcohol injuries
- Alcohol fatalities
- Unrestrained crashes
- Unrestrained injuries
- Unrestrained fatalities
- Motorcycle crashes
- Motorcycle injuries
- Motorcycle fatalities
- Speed crashes
- Speed injuries
- Speed fatalities
- Bicycle crashes
- Bicycle injuries
- Bicycle fatalities
- Total crashes
- Total injuries
- Total fatalities

Figure 14. Column heads of the traffic fatality rate dataset

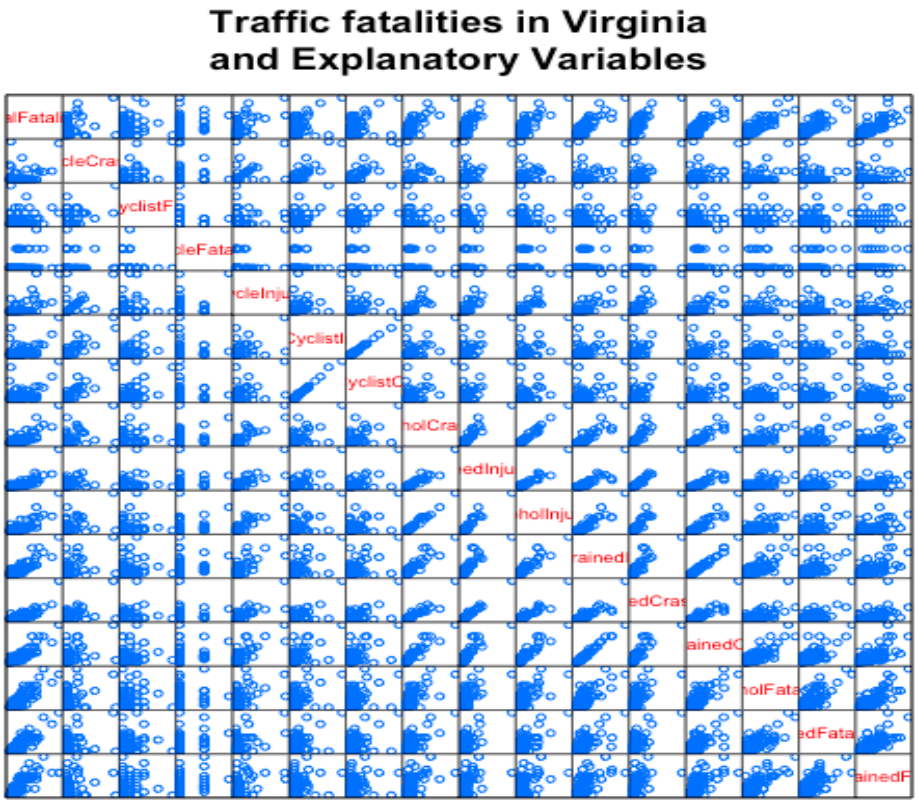

Figure 15. Scatterplot matrix with all the dependent variables

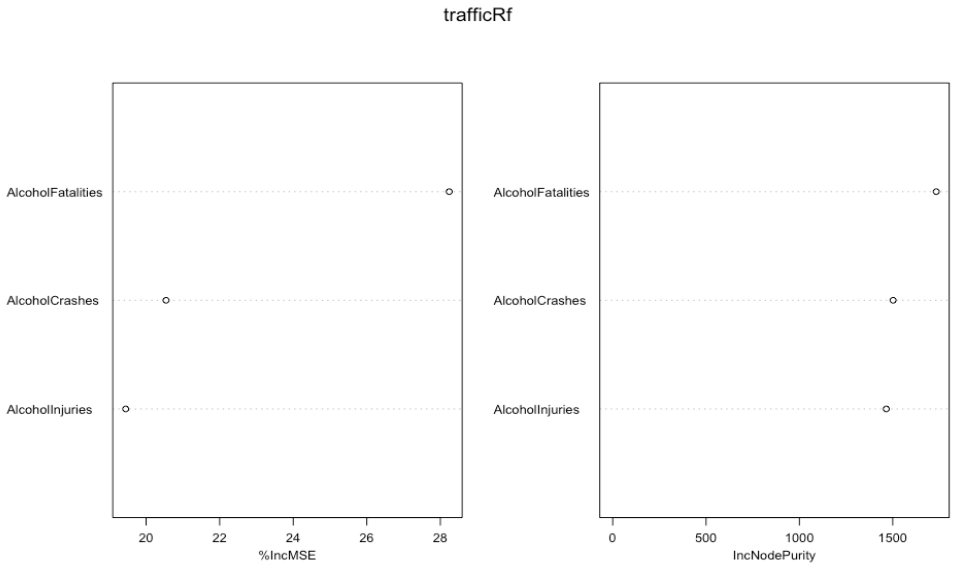

Figure 16. Variable Importance plot using few variables predicting biased results