PROJECT

STAT 575

Namrata Ray

Fatal Road Accidents in the United States: A Multivariate Analysis Exploration

Abstract: Traffic fatalities in the United States have been the highest as compared to most high-income countries. The objective of the current study is to examine the correlates of traffic fatalities in the U.S. and identify groups of states that display higher traffic fatality using statistical clustering techniques. Further, I also try to see the incidence of traffic fatality in the statistical clusters to further validate my results. This project is a multivariate statistical exploration and informs important policy action plan.

Keywords: Principal Components, K-Means Clustering, Traffic Fatalities, United States

Introduction

The recent estimate from the National Safety Council reports an alarming figure of 40,000 traffic fatalities on U.S. roads in 2018, a consistent third year high in a row in which at least that many people died in vehicle crashes (Beene, 2019). Not only within the country, traffic fatality rate in the United States ranks the highest in the developed world. The global report from the World Health Organization—which reviewed laws and crashes in 175 nations—explains that U.S.'s traffic fatality rate is 12.4 deaths per 100,000 — or about 50 percent higher than similar nations in Western Europe, plus Canada, Australia and Japan (StreetsBlog USA, 2018).

Vehicle-related mortality is an increasingly urgent public health concern. There are several underlying factors behind such figures, but weak road-safety laws have been assumed to be the main culprit. Such alarming figures demand a detailed investigation of traffic fatality. The objective of the present study is two-fold- 1) to examine the correlates of traffic fatality rate in the United States and 2) Identify statistical clustering of similar states that display higher incidence of traffic fatality.

Data

The dataset I used in this project was compiled and released as a CSV-file by FiveThirtyEight as an open-source file available on Github. The data was originally collected and compiled by the National Highway Traffic Safety Administration and the National Association of Insurance

Commissioners. The data gives state wise information of all the fatal road accidents that happened during 2014 on U.S. public roadways.

The main study variable is 'Number of drivers involved in fatal collisions per billion miles'. The predictor variables in the study include- 'Percentage Of Drivers Involved In Fatal Collisions Who Were Speeding',' Percentage Of Drivers Involved In Fatal Collisions Who Were Alcohol-Impaired', 'Percentage Of Drivers Involved In Fatal Collisions Who Were Not Distracted', 'Percentage Of Drivers Involved In Fatal Collisions Who Had Not Been Involved In Any Previous Accidents', and 'Car Insurance Premiums (in Dollars)'.

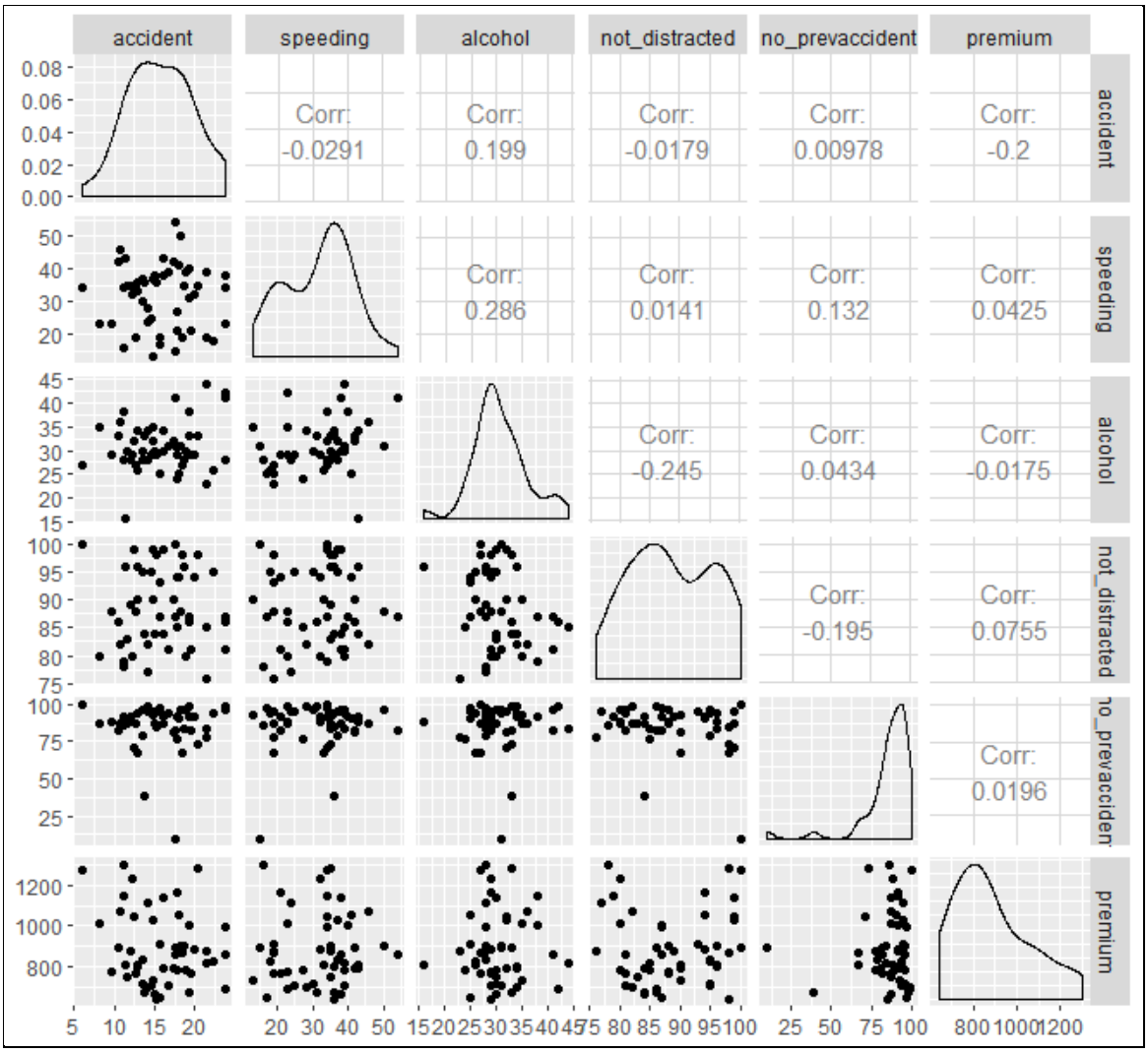
Methods

Principal components analysis (PCA) is a common unsupervised method for reducing data dimensionality, providing information on the overall structure of the analysed dataset. For this dataset, PCA is used to reduce data dimensionality and find features that explain the maximum variance in the dataset. To identify sub-structures in the data and identify groups of states that behave similarly, K-means clustering is performed. To determine the optimal number of clusters, I rely on the 'Elbow Method' and the 'Silhouette Method'. All the analysis has been conducted in R Software Version 3.6.1.

Results

The first figure is a correlation plot to assess the level of correlation among the study variables. The lower half of the diagonal displays the scatter plot while the upper half of the diagonal displays the correlation among the study variables. From the correlation plot, accident is positively correlated with alcohol consumption and negatively correlated with car premium. Further, variables are also correlated among themselves. For example, alcohol consumption is correlated with driving under speeding.

Figure 1: Scatter Plot Matrix and Correlation Matrix of the Study Variables



Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. To determine the correlates of road accidents, I implement a multiple linear regression with road accidents as the outcome variable. The coefficients from the table suggest that driving under alcohol consumption strongly explains road fatalities, after controlling for the effect of other variables. However, a regression often masks the relationship among the independent variables, for example, speeding and alcohol consumption.

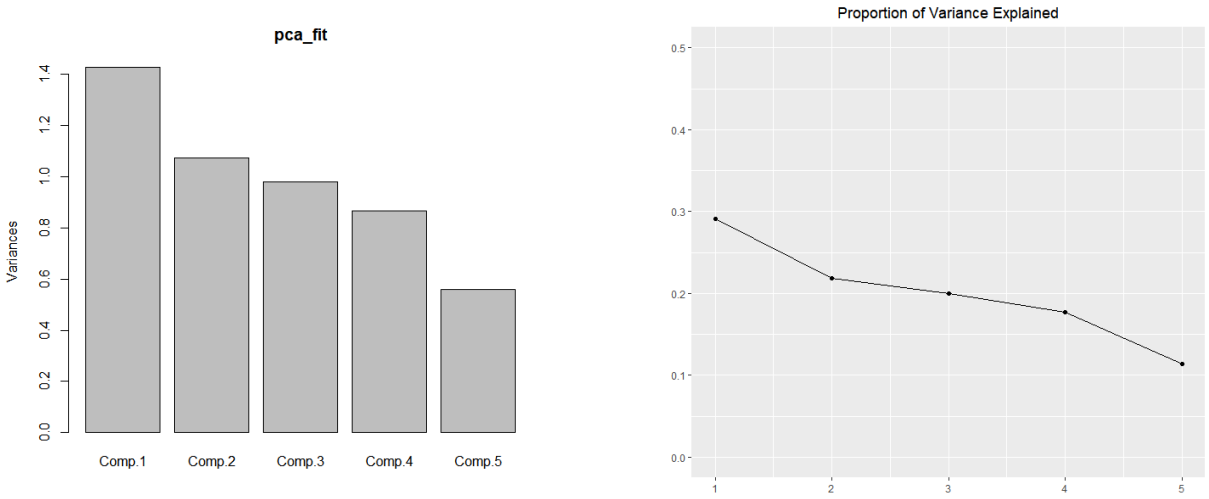
Figure 2: Results from Multiple Linear Regression

Variables	Coefficient
Speeding	-0.03
Alcohol	0.19
Not-distracted	0.03
First time accidents	0.007
Premium	-0.004
Intercept	11.36

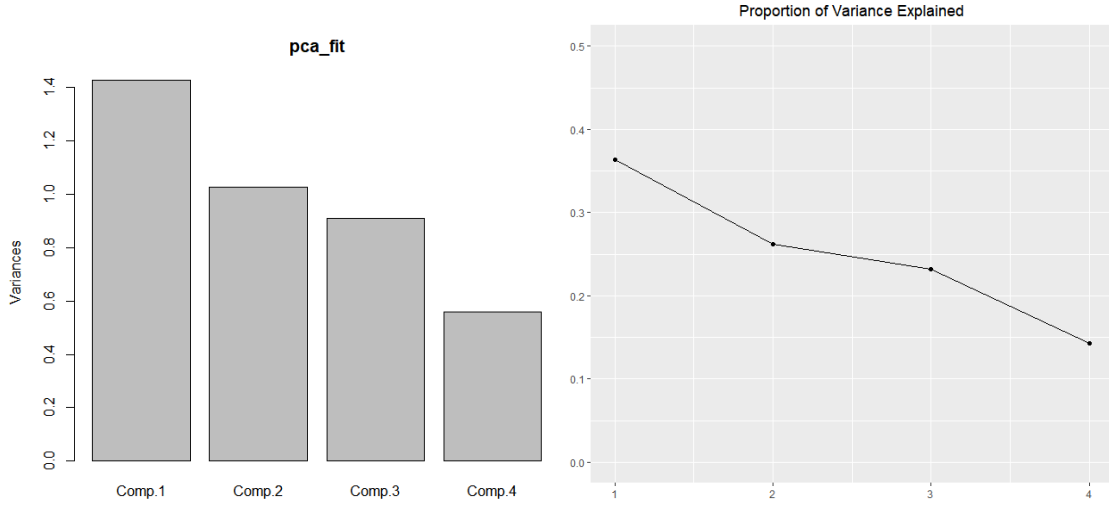
Principal components analysis (PCA) is a common unsupervised method for reducing data dimensionality and reveal strong patterns in the dataset. PCA reduces data by geometrically projecting them onto lower dimensions called principal components (PCs), with the goal of finding the best summary of the data using a limited number of PCs. In this data, I implement PCA to discern those variables that explains the maximum variation in the data. I begin with all the five independent variables and eventually reduce them to see which features explain the maximum variation.

Figure 3: Scree Plot and Proportion of Variance explained in each case

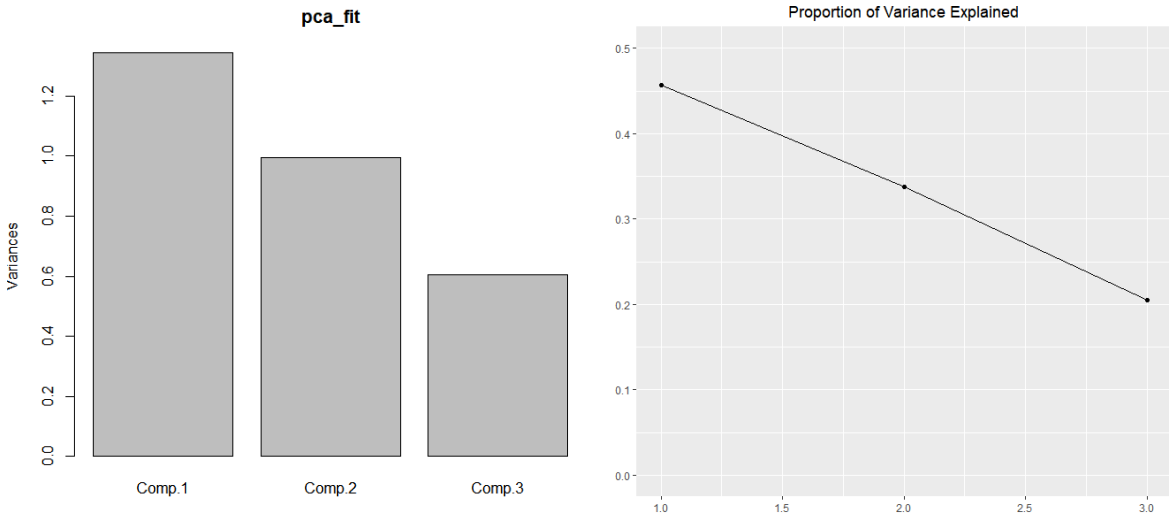
Panel A: Speeding, Alcohol, Not-Distracted, No previous accident, Car Premium



Panel B: Speeding, Alcohol, Not-Distracted, No previous accident



Panel C: Speeding, Alcohol, Not-Distracted

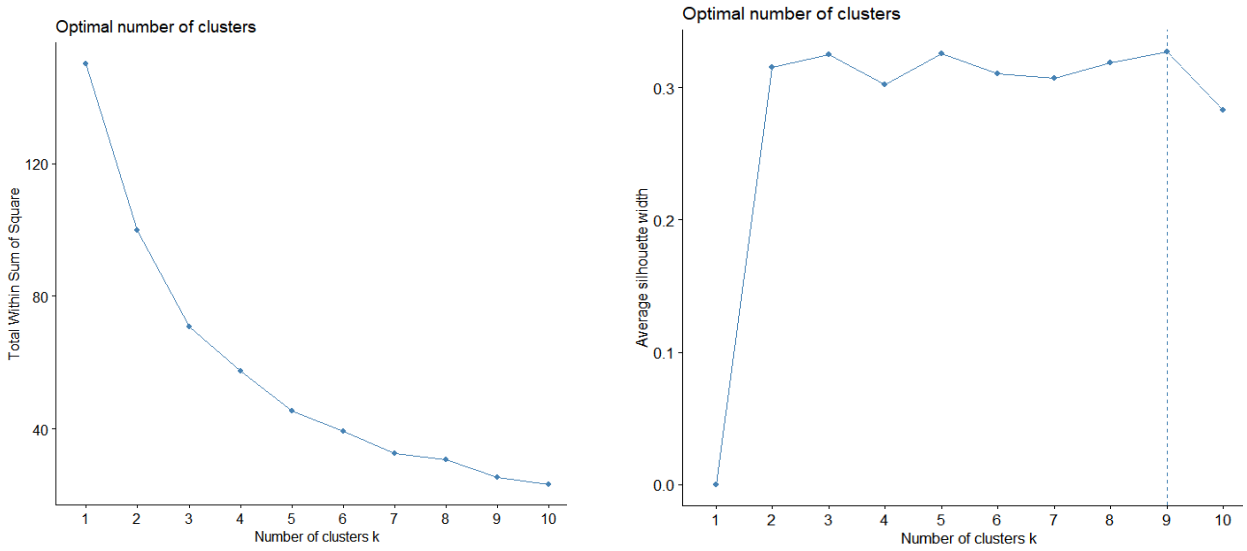


The Cumulative Variance explained by the first two components in Panel A is 51% while Panel B explains 62%. The three features in Panel C explains 79.4% of the variation. The next part employs clustering techniques on the principal components obtained from Panel C. For identifying statistical clusters among the 51 states, I use K-Means Clustering technique. The K-Means algorithm partitions the dataset into 'K' number of distinct clusters where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum.

Before implementing the K-Means method, it is necessary to identify the optimal number of 'k'. I use the 'Elbow Method' to determine the optimal number of clusters, where the elbow is used to determine the number of clusters.

Figure 4: The Elbow Plot and the Silhouette Method

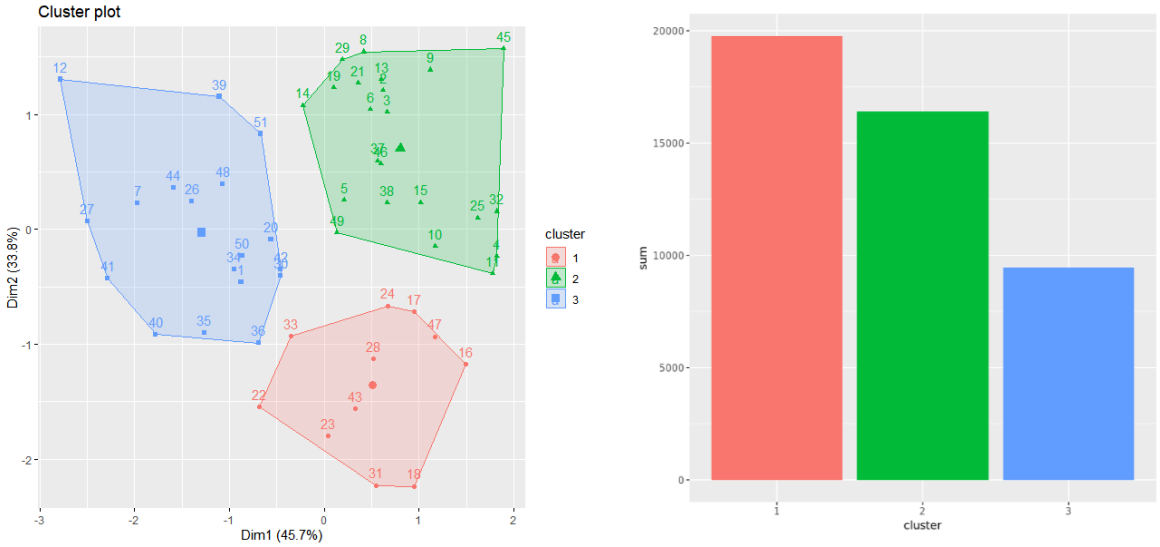
A) Elbow Plot	B) Silhouette Method



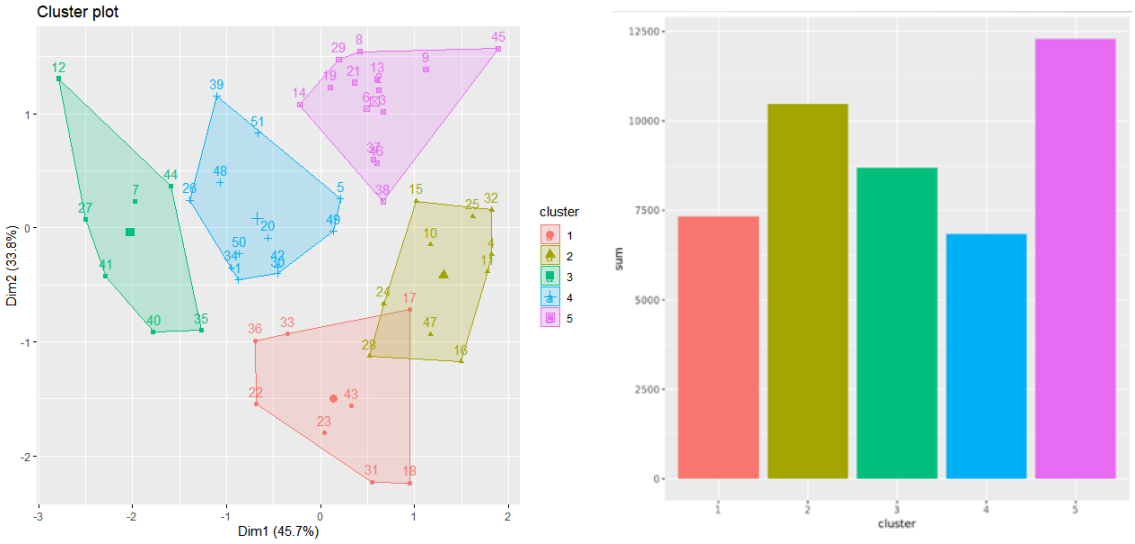
Panel A in the above figure represents the elbow plot. There is no clear elbow in the figure implying that there might be no definitive clustering. Roughly, the slope starts flattening from 3 to 9. While the Elbow method is more of a decision rule, the Silhouette is a metric used for validation while clustering. Thus, it can be used in combination with the Elbow Method. The Silhouette plot reports the global maximum at 9, which makes sense when referring back to the elbow plot. Although, the global maximum is at 9 in the above plot, it can be visually discerned that few local maxima exist around 2 to 3 and 5. I arrive at my clusters using 3 different values of 'k'.

Figure 5: Cluster Plots for different value of K

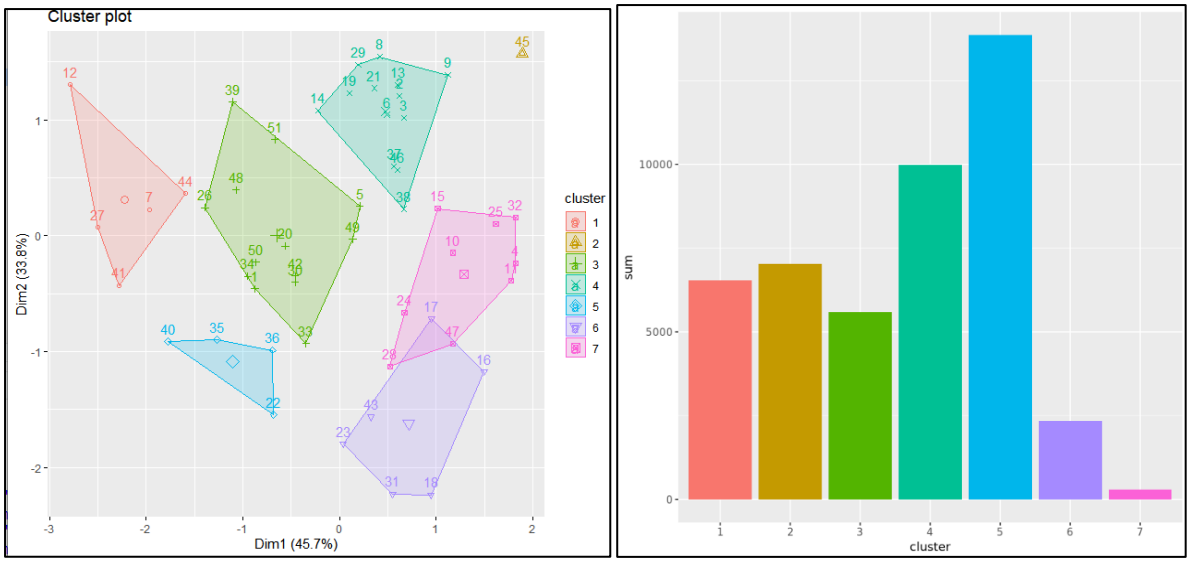
Panel A: K=3



Panel B: K=5



Panel C: K=7



A visual inspection across the panels explains that for three clusters, the number of traffic fatality is highest when K=3. Each cluster captures a different traffic fatality distribution. In reality, the number of clusters selected is a subjective choice. For this study, my goal was to arrive at statistical clusters across states that can partition states based on high traffic fatality. Based on my study results, maximum traffic fatalities are observed in the three-fold cluster but statistically, higher clusters can statistically divide the states based on different similar dimensions which might be more desirable.

Conclusion and Future Directions

My results reveal that the k-means clusters I arrived at captures a higher traffic fatality rate, which was my study objective. The statistical clustering of states has important policy implications for combating high incidence of traffic fatalities. Instead of implementing a costly nationwide plan, implementing policies based on state-wise clusters might be a better way to reduce traffic fatalities.

The current study has limitations. The selection of clustering algorithm and the optimal number of clusters is highly subjective. K-Means clustering tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different as possible based on the distance. This algorithm groups data points to a cluster based on the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster). Hierarchical Clustering is based on another set of algorithms where the

distance metric is calculated using different formulae. Further study can examine clusters using different combinations of distance and linkage metrics.

References

Ryan Beene.2019. “Traffic Fatalities Again Top 40,000; Few Cars on Roads Have Latest Technology”. https://www.insurancejournal.com/news/national/2019/02/13/517563.htm

Angie Schmitt. 2018. https://usa.streetsblog.org/2018/12/13/why-the-u-s-trails-the-developed-world-on-traffic-deaths/

Appendix

The codes used in the analysis are given as follows:

```
library(readr)

library(tidyverse)

library(dplyr)

library(cluster)

library(gridExtra)

#Data manipulation

mydata<-read_csv(url(urlfile))

head(mydata)

car_acc= mydata %>%

select (State, `Number of drivers involved in fatal collisions per billion miles`,`Percentage Of Drivers Involved In Fatal Collisions Who Were Speeding`,`Percentage Of Drivers Involved In Fatal Collisions Who Were Alcohol-Impaired`,`Percentage Of Drivers Involved In Fatal Collisions Who Had Not Been Involved In Any Previous Accidents`,`Percentage Of Drivers Involved In Fatal Collisions Who Were Not Distracted`,`Car Insurance Premiums ($)`)

colnames(car_acc) = c("state","accident","speeding","alcohol","not_distracted","no_prevaccident","premium")

car_acc %>%

  select(-state) %>%

  ggpairs()

corr_col <- car_acc %>%

  select(-state) %>%

  cor()

# Print the correlation coefficient for all column pairs

print(corr_col)

#fitting regression

fit_reg <- lm(accident~speeding+alcohol+not_distracted+no_prevaccident+premium,car_acc)

fit_coef <- coef(fit_reg)

fit_reg

print(fit_coef)

#PCA

car_acc_std <- car_acc %>%

  mutate(speeding=scale(speeding),

         alcohol=scale(alcohol),

         not_distracted=scale(not_distracted),
```

```
        no_prevaccident=scale(no_prevaccident),
        premium=scale(premium))
#pca_fit <- princomp(car_acc_std[,c("speeding", "alcohol", "not_distracted","no_prevaccident","premium")])
#pca_fit <- princomp(car_acc_std[,c("speeding", "alcohol", "not_distracted","no_prevaccident")])
pca_fit <- princomp(car_acc_std[,c("speeding", "alcohol", "not_distracted")])
#pca_fit <- princomp(car_acc_std[,c("speeding", "alcohol", "not_distracted","premium")]) 61%
pca_fit
# Obtain the proportion of variance explained by each principle component
(pr_var <- pca_fit$sdev^2)
(pve <- pr_var / sum(pr_var))
# Plot the proportion of variance explained, draw a point plot connected with lines
data_frame(comp_id=1:length(pve) , pve ) %>%
  ggplot( aes(x=comp_id , y=pve) ) + geom_point() + geom_line() + ggtitle("Proportion of Variance
Explained")+ theme(plot.title = element_text(hjust = 0.5))+
  coord_cartesian(ylim=c(0,0.5)) +
  labs(x="",
       y="")
screeplot(pca_fit)

# cumulative proportion of variance and the variance explained by the first two principal components
cve <- cumsum(pve)
cve_pc2 <- cve[2]
print(cve_pc2)
pcomp1 <- pca_fit$scores[,1]
pcomp2 <- pca_fit$scores[,2]
# Plot of the first 2 principle components in a scatterplot
data_frame(pcomp1,pcomp2) %>%
  ggplot(aes(pcomp1,pcomp2)+geom_point()+ labs(x="Principle Component 1", y= "Principle Component 2")
# k-means
k_vec <- 1:10
inertias <- rep(NA, length(k_vec))
# Initialise empty list to save K-mean fits
mykm <- list()
```

```
set.seed(1)
for (k in k_vec) {
  # for each k, fit a K-mean model with k clusters and save it in the mykm list
  mykm[[k]] <- kmeans(car_acc_std[,c(3,4,5)], centers = k, nstart=50)
  # for each k, get the within-cluster sum-of-squares and save
  inertias[k] <- mykm[[k]]$tot.withinss }
# Plot the within-cluster sum-of-squares against the number of clusters used
data_frame(k_vec,inertias) %>%
  ggplot( aes(k_vec, inertias) ) +
  geom_point() + geom_line() +
  labs(x="Number of clusters", y="Intertias")
cluster_id <- as.factor(mykm[[3]]$cluster)
# Color the points of the principle component plot according to their cluster number
data_frame(pcomp1,pcomp2) %>%
  ggplot(aes(x=pcomp1,y=pcomp2,col=cluster_id)) + geom_point() +
  labs(x="Principle Component 1",
       y="Principle Component 2")
#install.packages("factoextra")
library(factoextra)
head(car_acc_std)
#car_temp =car_acc_std[,c(3,4,5)]
#rownames(car_temp) = car_acc_std$state
fviz_cluster(mykm[[3]], car_acc_std[,c(3,4,5)])
fviz_cluster(mykm[[5]], car_acc_std[,c(3,4,5)])
fviz_cluster(list(mykm[[7]], car_acc_std[,c(3,4,5)]))
fviz_nbclust(car_acc_std[,c(3,4,5)], kmeans, method = "wss")
fviz_nbclust(car_acc_std[,c(3,4,5)], kmeans, method = "silhouette")
```