ANA 515 Assignment 2: Loading/Saving/Describing Driving Data

Kevin M Rabidou

2023-06-15

# Driving Data Description

The dataset contains information on all 50 states as well as the District of Colombia for driving data. The intent of setting up this data was to look at what states have the best and worst drivers. The focus of the data is broken down into two sections; accident reports and insurance. Accidents are broken down by their potential causes and insurance looks at claims and payouts. Through this assignment, the focus is going to be on the accidents and looking at the differences of probabilities of a certain event whether alcohol, speeding, etc is more common in fatal accidents. The bad drivers data file was sourced as a csv file therefore it has a comma delimiter. As the file was found on GitHub, it can easily be read into RStudio through the url link with taking notice of the delimiter properly.

# Uploading and Cleaning Data

# Section 2  
# Name the link then create a new set after uploading the file  
url <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/bad-drivers/bad-drivers.csv"  
DrivingData <- read\_csv(url)  
# As the dataset was set as a csv file from GitHub, I found it best to use read.csv to read in the data. This reading function requires the "readr" package. This function also takes care of noticing the headers (row/column names) as well as the delimiter which in this case was a comma. Using read\_csv() also allows for the column names to not be changed if there are spaces.

# Section 3  
# I am first going to rename some of the columns and use short hand as they have the majority of the same beginning phrase just a different category of driving.   
# Short hand examples:   
# Number of drivers involved in fatal collisions per billion miles = NoDiFC\_perBillionMiles  
# Percentage Of Drivers Involved In Fatal Collisions Who Were Speeding = PoDiFC\_Speeding ...(PoDIFC will be the same for the columns with the cuase cahnging after underscore)  
  
renamedDrivingData <- DrivingData %>% rename("NoDiFC\_perBillionMiles" = "Number of drivers involved in fatal collisions per billion miles", "PoDiFC\_Speeding" = "Percentage Of Drivers Involved In Fatal Collisions Who Were Speeding", "PoDiFC\_AlcoholImpaired" = "Percentage Of Drivers Involved In Fatal Collisions Who Were Alcohol-Impaired" , "PoDiFC\_NotDistracted" = "Percentage Of Drivers Involved In Fatal Collisions Who Were Not Distracted" , "PoDiFC\_NoPreviousAccidents" = "Percentage Of Drivers Involved In Fatal Collisions Who Had Not Been Involved In Any Previous Accidents")  
  
# To simplify the data and look at only the percentages I am going to remove the insurance premiums and payouts (subsetting the renamed data file). Just looking to focus on the amount of people involved in accidents and there causes.   
  
subsettedDrivingData <- renamedDrivingData %>% select(-`Car Insurance Premiums ($)` , -`Losses incurred by insurance companies for collisions per insured driver ($)`)

# Data Characteristics

This dataframe has 51 rows and 6 columns. The names of the columns and a brief description of each are in the table below.

#Section 4  
library(knitr)  
columns\_summary <- data.frame(Columns = c(colnames(subsettedDrivingData)),Description = c("States/Districts in the United States", "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 Were Not Distracted", "Percentage Of Drivers Involved In Fatal Collisions Who Had Not Been Involved In Any Previous Accidents" ))  
  
kable(columns\_summary, caption = "Measures of Driving In the United States")

Measures of Driving In the United States

| Columns | Description |
| --- | --- |
| State | States/Districts in the United States |
| NoDiFC\_perBillionMiles | Number of drivers involved in fatal collisions per billion miles |
| PoDiFC\_Speeding | Percentage Of Drivers Involved In Fatal Collisions Who Were Speeding |
| PoDiFC\_AlcoholImpaired | Percentage Of Drivers Involved In Fatal Collisions Who Were Alcohol-Impaired |
| PoDiFC\_NotDistracted | Percentage Of Drivers Involved In Fatal Collisions Who Were Not Distracted |
| PoDiFC\_NoPreviousAccidents | Percentage Of Drivers Involved In Fatal Collisions Who Had Not Been Involved In Any Previous Accidents |

# Summary of Subsetted Data

# Section 5  
sectionFiveSubset <- select(subsettedDrivingData, PoDiFC\_Speeding, PoDiFC\_AlcoholImpaired, PoDiFC\_NoPreviousAccidents)  
  
speedingSubset <- select(subsettedDrivingData, PoDiFC\_Speeding)  
alcoholSubset <- select(subsettedDrivingData, PoDiFC\_AlcoholImpaired)  
accidentSubset <- select(subsettedDrivingData, PoDiFC\_NoPreviousAccidents)

summarySection5 <- summary(sectionFiveSubset)  
  
kable(summarySection5, caption = "Summary of Accident Report Variables")

Summary of Accident Report Variables

|  | PoDiFC\_Speeding | PoDiFC\_AlcoholImpaired | PoDiFC\_NoPreviousAccidents |
| --- | --- | --- | --- |
|  | Min. :13.00 | Min. :16.00 | Min. : 76.00 |
|  | 1st Qu.:23.00 | 1st Qu.:28.00 | 1st Qu.: 83.50 |
|  | Median :34.00 | Median :30.00 | Median : 88.00 |
|  | Mean :31.73 | Mean :30.69 | Mean : 88.73 |
|  | 3rd Qu.:38.00 | 3rd Qu.:33.00 | 3rd Qu.: 95.00 |
|  | Max. :54.00 | Max. :44.00 | Max. :100.00 |

The probability of a fatal accident from speeding has 0 missing values. The probability of a fatal accident from being Alcohol Impaired has 0 missing values. The probability of a fatal accident from having no previous accidents has 0 missing values.