RMarkdown Week 8 & 9 Project

Dipika Sharma

May 16, 2021

Final Project step 1

Introduction

For my final project I want to analyze drivers data. Road accidents have become so common these days that i can surely say I have heard or read about accidents at an average of 1/week. It is sad to know that most of times the reason for accident is very common like driver watching phone, speeding, alcohol consumption and more. I often come across some drivers who do not follow driving rules properly. It is frustrating to see how the mistake of one can harm others on the road. I live in Massachusetts and with this analysis want to observe others states too for bad drivers.

Research questions

I have couple of questions that I want to analysis via this project.

1. Which states have the bad drivers?
2. Which states have more cases of speeding?
3. Which states have more cases of a collision when driver is alcohol impaired?
4. Relationship between the fatal collision and driver with speeding? 5 Relationship between the fatal collision and drive alcohol impaired?

Approach

Data Collection: The best approach to solve any problem is to collect as much as information about the problem we are trying to solve. In our case we want to analyze bad drivers data so I will search for bad drivers data.

Data Cleaning: With the help of different R's function like head, tail, str, summary and others, I will look at the data structure and summary of the collected data set. i will see if any transformation or modification is required before started working on the dataset. It is an important steps as the wrong data can lead our analysis to wrong direction. I will check for blanks or NULL values and will see if all the columns data types are correct.

Planning for solution: Once i have the data to look i will focus on my goals to see what answers I can get using this data. I will look for the problem solution and try to find different ways to answer them. I can look for the ideas that what visualization or the modeling technique would be best to use with bad drivers data set which can give my solution to my problem.

Selection of Solution: After looking at all the solution approaches, I can pen down all the pros and cons of each approach. This will give me better idea which approach to choose for achieving goal.

Execution: as an first time, I will break down the selected solution into multiple steps and will start executing the step one by one. I will make notes of my observation and reading I will get from each steps. This will help me in rectifying errors if i encounter any in the process.

Evaluation: After running all steps of solution, I can go through my notes to write an conclusion. In my project my goal is to look for answer to my research questions. I am expecting to have answer to my question in this stage.

How your approach addresses (fully or partially) the problem.

My first approach is to do one step at a time, I think this is the best approach to solve any problem. This will give me chance of improvement and in case of any error it will make it easier to find it and then solve it. I have tried to include each and every steps in above approaches which will drive me to the solution of my problem. Considering the collection of data to its modification and then to understand the structure of data set will help me to form a strong base to achieve my goal. Once the data is all set I can process the data using different function of R on the data set to determine the results.

Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)

1. I found bad drivers data set in fivethirtyeight.com which I am planning to use in my analysis. Here is the link below:

https://github.com/fivethirtyeight/data/tree/master/bad-drivers

The data is collected by National Highway Traffic Safety Administration in year 2009, 2010, 2011 and 2012. The data is about the motor vehicle crashes in United States during these period. Looking at the data I can see there are 51 cases in this data set and each case represent the data for one state.

I can see two types of variables in this data set, one is dependent variable like the number of fatal collisions which is of numeric data type and other is independent variables like Car Insurance Premiums, states and others. Some of the independent variables are numerical, categorical and respectively.

2. I also find an interesting article by QuoteWizard online which showing the worst and good drivers by states. it would be intersting to compare my analysis with quotewizard study. Here is the link:

https://quotewizard.com/news/posts/the-best-and-worst-drivers-by-state

3. There is one more article by SmartAsset, they spent some time in analyzing the worst drivers by states. Here is the link:

https://smartasset.com/checking-account/states-worst-drivers-2020

4. There is an another data set available from kaggle, below is the link:

https://www.kaggle.com/sobhanmoosavi/us-accidents?select=US_Accidents_Dec20_Updated.csv

This data set have some 47 column which includes both dependent variable and independent variable. I saw we have the accident detail with state or city information so we can also use this data set to find the answer to our question like which state has most accident/worst drivers?

This data set has 49 states accident data for period between February 2016 to Dec 2020.

5. Another data set which caught my attention is data from kaggle about the aviation accident. Here is the link of the data set:

https://www.kaggle.com/khsamaha/aviation-accident-database-synopses

This database contains information from 1962 and later about civil aviation accidents and selected incidents within the United States. This database has some 31 columns which contained the useful information about the avaition accidents.

Add Citations

- R for Everyone (Lander 2014)
- Discovering Statistics Using R (Field, Miles, and Field 2012)

Required Packages

For now, I am planning to us ggplot2, car, lm_beta packages. I might need to add more packages if required to meet the goal.

Plots and Table Needs

Visualization makes easy for us to read the data and to make conclusion. I am thinking of using so many plotslike histogram, scatterplot and may be box plot.

Using histogram I can clearly see which states have more fatal collision during observed period and may be box plot to study the number of fatalities in region which would give us more clear vision of which region has worst drivers. Using the scatterplot for all the region to study the relationship between dependent and independent variables.

I am thinking of using the summary function on data set to have better understanding of the data. I think I will add the summary tables to my final report with my observations.

Questions for future steps

I am not sure how the solution will look like in the end. It would be interesting if I can make some prediction in the starting and to see if the final result matches to my prediction or not. I am eager to see which states has the worst driver in the United States. It would be interested to see the relationship between different dependent variable and independent variable, although i am skeptical if i will be able to find any relation between them using the available data set.

References

Field, A., J. Miles, and Z. Field. 2012. *Discovering Statistics Using r*. SAGE Publications. https://books.google.com/books?id=wd2K2zC3swIC.

Lander, J. P. 2014. *R for Everyone: Advanced Analytics and Graphics*. Addison-Wesley Data and Analytics Series. Addison-Wesley. https://books.google.com/books?id=3eBVAgAAQBAJ.