Data Analytics Bootcamp
Final Project
Gary Schulke, James Bruhns

Project Report
Automotive Fatalities and Analysis

March 2, 2020

Introduction

This report examines the fatal automobile accidents and the factors that may contribute to them. Contributing factors include location, distractions, alcohol, drugs, and age. The initial belief was that there would be one or two factors acting as major contributors and all the rest would be minor contributors. With the majority of media attention focused on drunk and distracted driving it would be logical to conclude these are major contributors, but this is shown to be false.

Installation

There are no specialized installation procedures. The folder structure create by a Git clone supports running of Collaboratory and Juniper Notebooks and Tableau Workbook.
Clone the repository: Git Repository - https://github.com/Gary-Schulke/Project3Startup.git

Open the Tableau Workbook "Project3Startup/Final Workbook.twbx".
If prompted to connect to the Postgress database, login as user: "root" and pwd: "finalproject"
If prompted to connect to a csv file, use the Tableau prompts to connect to: acc_dis_joined_cleaned.csv
Path "Project3Startup\TrafficData\ExtractionZone\Cleaned\acc_dis_joined_cleaned.csv"

The RDS Postgres database is located at "myprojectdb.cqbbih9hnlbh.us-east-1.rds.amazonaws.com"

Jupyter Notebooks can be run from the "Project3Startup" folders. These files are documented in the Submitted Files section below.

The Presentation

- Decision and Random Forest Classification was used to determine which factors are most closely related to fatalities. Time and location data are shown to be the biggest contributors. The data was analyzed a second time with time and location data removed to better show the human contribution to fatal accidents.

- Lighting conditions are charted and shows that there isn't any single or pair of lighting conditions that contribute heavily. Dark Unknown Lighting is an equal contributor for many states while other states don't use this designation at all. How lighting gets classified is highly subjective dependent upon the person recording the data.

- Drug and Alcohol fatalities are compared. Trend line show a decrease in both alcohol and drug related fatalities. The results are distorted by exceptionally high number for alcohol in year 2013 in Ohio. With or without Ohio 2017 and 2018 show improvement for both drugs and alcohol. The raw data has many sub-categories for drugs, and they were grouped together in this chart.

- Fatalities by Month show a slight rise during the warmer months but nothing dramatic. Drunk driving by month and year are consistent for the 6 year period examined. Many states changed their drunk driving laws in the late 90s and early 2000s so their affect cannot be seen in these charts.

- The top 20 city fatalities with alcohol involved are shown. The magnitude of fatalities is heavily dependent upon how urban areas have been grouped together. For example, Menomonee Falls, Wisconsin is a suburb of Milwaukee and represents more people than its name recognition would indicate.

- Alcohol related incidents in the range of lighting conditions is surprising. It was assumed that the darker conditions would be higher than in daylight. The opposite is true. Alcohol involved incidents during daylight house is significantly higher than any of the others. Day drinking is fun but dangerous.

- Cell phone related distracted driving has gotten a lot of attention in recent years and the results show why. Fatalities due to cell phones is more than twice the next closest which is Other People and more than all the other combined. In the raw data, the fatalities that are cell phone related are subdivided into categories such as taking on phone, texting, and fumbling for phone.

Conclusion

While alcohol, drugs, and distracted driving all contribute to driving fatalities so do passengers in the car and operating the car's heater and radio controls. It is more accurate to say that alcohol, drugs, and distracted driving are 100 percent preventable. A better analysis could be performed if non-fatal accidents and their contributing causes were included in the data. The last chart showing distractions clearly shows that cell phones contribute significantly to fatalities. Enforcement of distraction laws, particularly cell phone laws is important in the effort to bring down preventable automotive fatalities.

Lessons Learned

- Even modest amounts of "Big Data" can be difficult to munge into a useable form and can take enormous time and computing resources to analyze.
- Putting a lot of thought into including and excluding data as it can affect what the results are. Basically, think big, work your way smaller.
- If grouping data as was done with distraction data, putting the data in meaningful groups is important and possibly the easiest part. You also need to consider how the groups will be analyzed and visualized in a meaningful way.

Software Languages and Packages

Python Pandas
Python Scikit-Learn
Postgres DB and SQL
AWS S3 and RDS
Tableau
Collaboratory

About the Data

The data was acquired and organized by the National Highway Traffic Safety Administration and focuses exclusively on traffic accidents with fatalities and the related circumstances that my have

contributed to the accident. The data is available for download from a FTP site and is in csv format. Date for years 2012 to 2018 was downloaded (223Mb). The data was well formatted and documented. Differences in data format versions were resolved and all column heading named consistently , 35 files (81.5Mb) was loaded into the AWS hosted Postgres Database.

Data Sources

National Highway Traffic Safety Administration
https://www.nhtsa.gov/ (Home)
ftp://nhtsa.gov/FARS/ (Data Repository)

U.S. General Services Administration
https://www.gsa.gov/reference/geographic-locator-codes/glcs-for-the-us-and-us-territories

References

National Conference of State Legislatures
https://www.ncsl.org/research/transportation/cellular-phone-use-and-texting-while-driving-laws.aspx

National Center for Biotechnology Information
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4001667/

Submitted Files

Final Workbook.twbx Tableau Workbook, charts and stories used to develop the project and show results.

HighwaySafetyMunging.ipynb – The development version of the Jupyter Notebook for munging the NHTSA FARS csv data.

HighwaySafetyMungingSingleCell.ipynb – The final implementation for data munging. Same as the above and has final changes and merged to a single cell for better automation.

hwy_safety_helpers.py – Supplies helper methods and dictionaries for categorizing distract and impairment numbers.
city_lookup.py – City, state lookup helper dictionary.
SQL_queries.ipynb – Implements the project models.
FARSmunge2018.ipynb – Combines csv files and inserts them into the Postgres database using *findspark*.
Project3Startup\TrafficData\ExtractionZone\Cleaned* – The output files from data munging.

Project3Startup\TrafficData\ExtractionZone* – The working area for zipped and extracted folders with the downloaded files.
acc_dis_to_cvs_query.sql – Script to merge accident and distract data.

SQL Table Creation
– create_accident.sql
– create_distract.sql
– create_drimpair.sql
– create_person.sql
– create_violation.sql