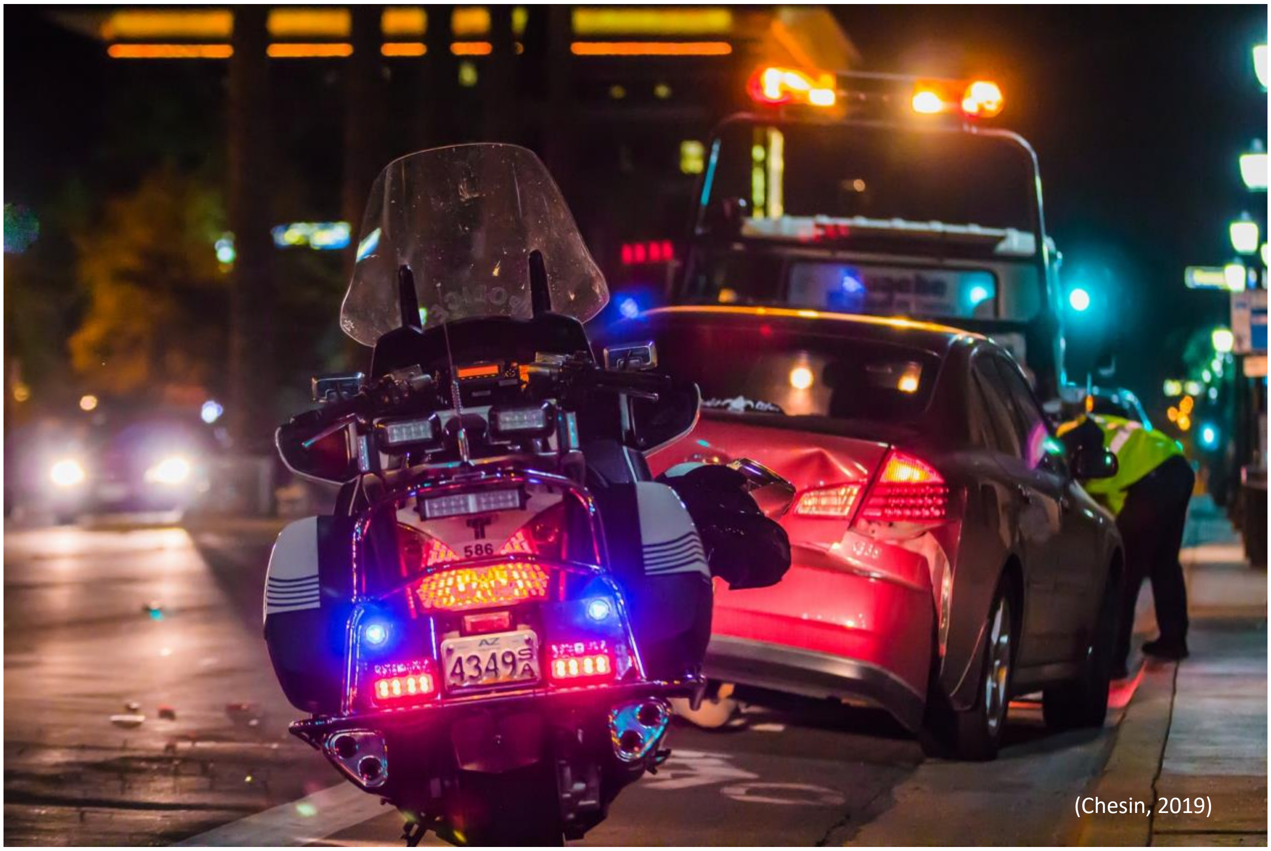
Descriptive Analysis for Car Accidents in Washington D.C


(Chesin, 2019)

ZAID ALTUKHI, TEAM# 9
DR. LIAO, PH.D.

George Mason University
AIT 614: Big Data Essentials

January 2, 2022

Descriptive Analysis for Car Accidents in Washington D.C.

Abstract

The descriptive analysis could be used to find the trends and patterns in historical data. In this project, we will use the descriptive analysis to describe the car accidents in Washington, D.C, between 2009 to 2020. The dataset was downloaded from the District Department of Transportation (DDOT), the department responsible for car accidents in Washington D.C. We have applied multiple analytics and statistical models to find the relationships between different variables and the patterns and trends among the data, such as correlation analysis, confidence interval, One-Way-ANOVA, decision tree, and visualizations. The project aims to find the common reasons for the accident and help DDOT find ways to reduce and eliminate the accidents in the area. The statistical and analytical tools examine multiple features to find the patterns and trends among the datasets. Four main findings were found after analyzing the data. First, the main reason for most crashes is drunk people, either drivers or pedestrians. The second finding is that the top reason which causes deadly accidents is speed. Also, we have found that most of the accidents are not dangerous. In addition, we found the top ten streets that contain the highest accident number, and we found that they are located on the north side of the town.

Introduction

What are the reasons that car accidents are one of the most dangerous events on roads? World Health Organization says that around 1.3 million persons die because of car accidents and cost around 3% of most nations' gross domestic (2021). According to the National Highway Traffic Safety Administration of the United States Department of Transportation, this is the highest six-month rise ever recorded in the Fatality Analysis Reporting System's history. In the first half of 2021, a projected 20,160 individuals died in car accidents, rising 18.4 percent over the same period in 2020. Since 2006, this has been the highest number of expected fatalities in that period (National Highway Traffic Safety Administration, 2021). Many two kinds of reasons that lead to crashes on roads. First, external effects are the effects that the driver cannot avoid or is hard to avoid, such as the weather conditions and road situations. Second, internal effects are related to the car's driver, such as the driver's health condition, distractions, or car's tier issues. According to Templeton Smithee Hayes Heinrich & Russell, LLP (2020), Driver distraction is the leading cause of automobile accidents. Intoxicated drivers, speeding, hostile conduct, rain, failure to obey traffic signs, night driving, vehicle troubles, tailgating, and improper turns and driving lean are all factors to consider.

According to data acquired from the District Department of Transportation, about 258,000 accidents in Washington, D.C., between 2009 and 2020. (District Department of Transportation, 2021). Because it is the capital of the United States, Washington D.C. is one of the most important locations in the country. It includes all government offices, tourist attractions, and educational institutions. Furthermore, according to demographic data, the population in 2020 will be 689,545 people living in 68.34 square miles (Data Commons, n.d.).

This project explores and performs a descriptive-analytical analysis of the car accidents done in Washington, D.C, between 2009 and 2020. In order to find insights and patterns in those accidents and understand the reasons and the relationships between different variables that lead to those crashes.

Objective

This project will analyze the car accidents in the Washington, D.C area from 2009 to 2020. The researcher will examine the common factors between accidents, the locations of the accidents, the car crashes factors that may cause deaths or injuries compared to the number of accidents. Also, find the factors that significantly correlate with the number of injuries and accident elements. The accident elements are any vehicle involved in an accident. In addition, to rank the accidents into groups.

Finding patterns in many incidents gives a clear view of the likely causes that lead to an accident. It will also reveal whether any issues need to be addressed to limit the number of incidents. Because automobile accidents result in numerous injuries, the causes of such injuries will be investigated. On the other hand, many accidents result in merely automobile damage and no human injuries. In addition, assessing the automobile collision location will offer helpful information about areas where authorities should focus their efforts. It can also determine which areas have a high number of injuries or accidents that result in fatalities.

Dataset

Original Dataset

District Department of Transportation (DDOT) provides a high-quality dataset containing the accidents in Washington D.C. The data were collected by DDOT and Metropolitan Police Department (MPD). The data were downloaded contains 258,122 records and 60 features. However, 19 features will be dropped because it has no relation to the kind of analysis performed on the data. Also, all car accidents that happened before 2009 were dropped. After dropping the unrelated column and data before 2009, those accidents have no vital data. There are seven columns added to the datasets. Because the dataset is very detailed, those columns were added to aggregate some columns to calculate the number of injuries per accident.

Preprocessing And Explore the Dataset

The shape of the final dataset that will be analyzed contains 237,193 and 55 columns. The following table shows the features with a brief description for each column (District Department of Transportation, 2021).

Attributes Description

Column Name	Data Type	Description
X	Float	X point is a spatial point for the accidents
Y	Float	Y point is a spatial point for the accidents
OBJECTID	Object	Unique ID generated by a system.
CRIMEID	Integer	Unique ID for MPD crash report.
CCN	Object	Criminal complaint number.
REPORTDATE	Date- time	Date the crash was reported.
ROUTEID	Object	Unique ID that defines each street.
MEASURE	Float	Distance (in meters) from the start of the street/route along the DDOT centerline
OFFSET	Float	The distance (in meters) between the original MPD latitude/longitude location and the point along the DDOT centerline.
FROMDATE	Date-time	Date the crash was happened
MARID	Object	Master Address Repository Web services ID.
ADDRESS	Object	crash location
LATITUDE	Float	latitude of the crash location
LONGITUDE	Float	longitude of the crash location
XCOORD	Float	longitude of the address location, which is supplied by the MAR.

Column Name	Data Type	Description
YCOORD	Float	latitude of the address location, which is supplied by the MAR.
EVENTID	Object	Internally ID for the crash location
MAJORINJURIES_BICYCLIST	Integer	Total number of bicyclists with major injuries
MINORINJURIES_BICYCLIST	Integer	Total number of bicyclists with minor injuries
UNKNOWNINJURIES_BICYCLIST	Integer	Total number of bicyclists with unknown injuries
FATAL_BICYCLIST	Integer	Total number of bicycle fatalities
MAJORINJURIES_DRIVER	Integer	Total number of car occupants with major injuries
MINORINJURIES_DRIVER	integer	Total number of car occupants with minor injuries
UNKNOWNINJURIES_DRIVER	Integer	Total number of car occupants with unknown or undefined injuries
FATAL_DRIVER	integer	Total number of car occupants with fatalities
MAJORINJURIES_PEDESTRIAN	Integer	Total number of pedestrians with major injuries
MINORINJURIES_PEDESTRIAN	Integer	Total number of pedestrians with unknown or undefined injuries
UNKNOWNINJURIES_PEDESTRIAN	Integer	Total number of pedestrians with unknown or undefined injuries
FATAL_PEDESTRIAN	Integer	Total number of pedestrian fatalities
TOTAL_VEHICLES	Integer	Total number of any cars type involved in the crash

Column Name	Data Type	Description
TOTAL_BICYCLES	Integer	Total number of bicyclists involved in the crash
TOTAL_PEDESTRIANS	Integer	Total number of pedestrians involved in the crash
PEDESTRIANSIMPAIRED	Integer	Total number of pedestrians who were impaired
BICYCLISTSIMPAIRED	Integer	Total number of bicyclists who were impaired
DRIVERSIMPAIRED	Integer	Total number of motorists who were impaired
TOTAL_TAXIS	Integer	Total number of taxis involved in the crash.
TOTAL_GOVERNMENT	Integer	Total number of DC, federal or other governmental vehicles involved in the crash
SPEEDING_INVOLVED	Integer	if the reporting officer believed speeding was a factor in the crash.
NEARESTINTROUTEID	Object	The unique centerline id of the nearest intersecting roadway.
NEARESTINTSTREETNAME	Object	The name or the nearest intersecting roadway
INTAPPROACHDIRECTION	Object	The cardinal direction to the nearest intersecting roadway
FATALPASSENGER	Integer	Total number of fatalists involved in the crash.
MAJORINJURIESPASSENGER	Integer	Total number of major injuries involved in the crash.
MINORINJURIESPASSENGER	Integer	Total number of minor injuries involved in the crash.
UNKNOWNINJURIESPASSENGER	Integer	Total number of unknown injuries involved in the crash.

Column Name	Data Type	Description
TOTAL_FATAL	Integer	Total number of fatalists involved in the crash.
TOTAL_MAJORINJURIES	Integer	Total number of major injuries involved in the crash.
TOTAL_MINORINJURIES	Integer	Total number of minor injuries involved in the crash.
TOTAL_UNKNOWNINJURIES	Integer	Total number of unknown injuries involved in the crash.
TOTAL_ACCEIDENTELEMENTS	Integer	Total number of cars, taxis, governmental cars involved in the crash.
TOTAL_INJURIES	Integer	Total number of injuries involved in the crash (minor, major, unknown).
TOTAL_INJURIES_ELEMENTS	Integer	Total number of injuries and vehicles involved in the crash (minor, major, unknown).
RATE	Category	Ranking the accident based on specific algorithm between lowest, low, medium, high, very high, extreme based on the number of injuries and vehicles in the accident.
FATAL	Boolean	Indicates if the accident has any fatal case or not

TABLE 1: ATTRIBUTES DESCRIPTION

After exploring the dataset and determining the needs, many processes were perfume to clean and prepare the data. We need to use some offered software to perform the data analysis, statistical models, and visualizations.

First, import the packages used to clean and prepare the data. There are eight libraries that were added as follow:

- Anaconda (Anaconda Software Distribution ,2020)
- Pandas (The pandas development team, 2020)
- Numpy (Harris et al., 2020)
- Sklearn (Varoquaux et al., 2015)

- SciPy (Virtanen et al., 2020)
- Altair (VanderPlas et al., 2018)
- Matplotlib (Hunter, 2007)
- Seaborn (Waskom et al., 2017)

Feature Engineering

After importing the original dataset, there are 18 columns dropped because either they have too many null values or have not relevant to our analysis. Then there are nine columns added. They all aggregate multiple numeric columns except the rate column, which classifies the crashes into categories.

Column name	Description
TOTAL_FATAL	Sum of all attributes that contain fatalities data, which are: driver, bicyclist, pedestrian, and passengers.
TOTAL_MAJORINJURIES	Sum of all attributes that contain major injuries data, which are: driver, bicyclist, pedestrian, and passengers.
TOTAL_MINORINJURIES	Sum of all attributes that contain minor injuries data, which are: driver, bicyclist, pedestrian, and passengers.
TOTAL_UNKNOWNINJURIES	Sum of all attributes that contain unknown injuries data, which are: driver, bicyclist, pedestrian, and passengers.
TOTAL_ACCEIDENTELEMENTS	Sum of all attributes that contain any type of objects involved in the accident, which are: vehicles, bicycles, pedestrians, taxis, and government cars.
TOTAL_INJURIES	Sum the new columns were add that related to injuries which are, TOTAL_MAJORINJURIES, TOTAL_MINORINJURIES, and TOTAL_UNKNOWNINJURIES
TOTAL_INJURIES_ELEMENTS	Sum the column TOTAL_ACCEIDENTELEMENTS and TOTAL_INJURIES
RATE	Divided the accidents to six levels based on the total injuries and accident elements.
FATAL	Indicates if the accident has any fatal case or not

TABLE 2: NEW COLUMNS

Then, there are some columns need to edit their data types from numerical to other type of data type:

Column name	Old data type	New data type
REPORTDATE	String	Date/time
FROMDATE	String	Data/time

Column name	Old data type	New data type
OBJECTID	Integer	String
CRIMEID	Integer	String
ROUTEID	Integer	String
MARID	Integer	String

TABLE 3: EDITED DATA TYPES

After converting the data type for the columns that need to be edited, we found that the data contains old accident information, but it is few, and they contain many null values, so they were removed. The data will have remained the accidents that happened between 2009 to 2020. The original dataset contains accidents information up to Aug 2021. However, it decided to remove the accidents that happened in 2021 because it may affect the analysis results. Since they just for eight months.

Null Values

After performing feature engineering, we found five columns containing missing values: FROMDATE, ADDRESS, LATITUDE, LONGITUDE, and EVENTID. All these values were removed Because it is hard to fill them.

However, some accidents have zero accident elements and zero injuries. Those accidents were removed.

Classify Accidents

To analyze the accidents, it must be categorized. In this project, the author decided to divide the accidents into five categories lowest, low, medium, high, extreme. The categories are done by using the cut point function provided in Panda's library. The cut points range is shown in the table below. However, these numbers were picked based on the observations and data distribution. The result of this rank is as follow:

Rank	Range	Number of accidents
lowest	0-2	115,755
low	3-10	121,254
medium	11-13	129
high	13-30	49
extreme	>30	6

TABLE 4: ACCIDENTS RATES WITH TOTAL NUMBER OF ACCIDENTS IN EACH CATEGORY

Explore the Data

To understand the data and the distribution, we need to present statistical information and illustrate visualizations.

First, using describe function on the data frame that contains the data provide vital information about the data. The table below shows the count, mean, standard deviation, min, max, 25%, 50%, and 75% for the added columns.

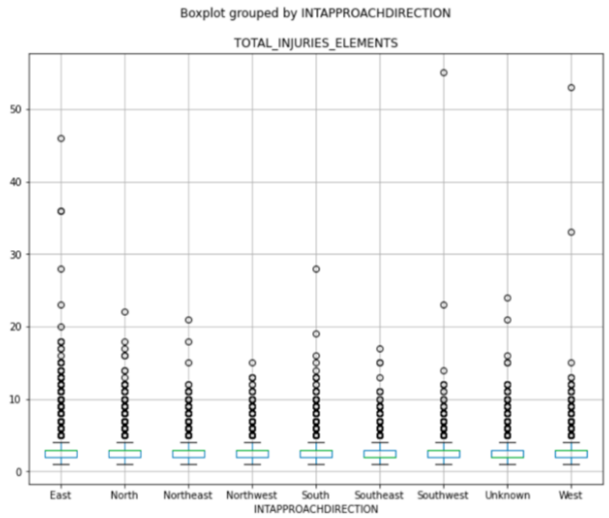


FIGURE 1: TOTAL ACCIDENTS PER IN APPROACH DIRECTION

	TOTAL_FATAL	TOTAL_MAJORINJURIES	TOTAL_UNKNOWNINJURIES	TOTAL_ACCEIDENTELEMENTS	TOTAL_INJURIES	TOTAL_INJURIES_ELEMENTS
count	237193	237193	237193	237193	237193	237193
Sum	476	25,980	18,161	532,660	114,316	646,976
mean	0.002007	0.109531	0.076566	2.245682	0.481954	2.727635
std	0.045961	0.459874	0.311264	0.66367	0.838257	1.096224
min	0	0	0	0	0	1
25%	0	0	0	2	0	2
50%	0	0	0	2	0	3
75%	0	0	0	3	1	3
max	2	51	16	17	51	55

TABLE 5: NEW COLUMNS DESCRIPTION

The previous table shows the data distribution. By reading the number, we can find much helpful information that makes understanding the data more straightforward. For example, it can be noticed that the maximum number of fatal is two, which is a good indicator that although the number of accidents is high, the fatal cases are too few.

Also, visualization is considered one of the best ways to explore and understand the data. There are some charts generated to understand the data range and distribution.

This chart in **Error! Reference source not found.** shows the total number of elements and injuries based on the sector. It can be noticed that there are some outliers and the car crashes distribution based on these regions.

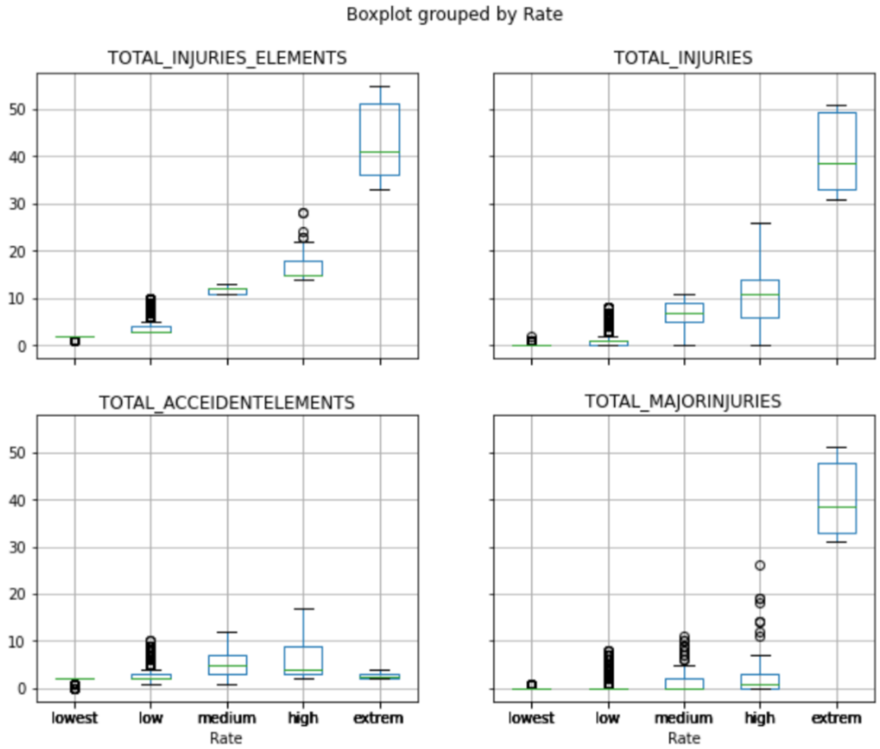


FIGURE 2: DISTRIBUTION OF TOTAL NUMBER OF ACCIDENTS AND INJURIES BASED ON ACCIDENT RATE

These box plots show the relationships and distribution between the total number of injuries and elements, total injuries, total accident elements, and total major injuries based on the accident category rank.

The System

This project examined many statistical and analytical models to find the patterns and trends in data. The first model that will be used is the correlations between features. This model can provide a good idea about how columns interact with each other. If there are high correlations this indicated, we can analyze these columns and find if there is an actual relationship or not. As known, correlation does not mean causation, but this phrase has been wrong in some cases. Also, a decision tree analysis is performed to find the factors that lead to fatal accidents. To understand in which cases the accidents could have any fatal people. In addition, visualizations are a great way to explore the data and find the outliers and data distribution. Understanding the data make the analyzing process done in the right way. If data were understood well and in context, this would help to understand the result better. In addition, some statistical models could be performed to test if there is a significant relation between features, such as Spearman correlation and ANOVA analysis.

The statistical and analytical tests that will be performed can answer many questions. For example, is there a relationship between the location and the number of accidents or injuries? Is there a relationship between the weak days or weekend days and the number of accidents or injuries? What are the common factors between accidents cause deaths, major or minor cases? These are some questions this search will answer using data analytics tools. In addition, answering those questions contributes positively to identifying the car accident problem and finding solutions to reduce the number of accidents.

On the other hand, using visualizations will make understanding the data much more straightforward by visualizing the data. It makes the data complexity present in a way many people can understand. Also, charts and maps can visually answer several questions. Because we have geospatial data, we can present the accident on maps, quickly understanding the locations that hold a high or a low number of accidents. Also, we can find the locations of the significant injuries or where precisely the taxis had accidents. Besides, a timeline chart, which is an excellent way to find the number of incidents within a time range, can provide many answers to understand the issue.

System architecture

In the analyzing phase, many steps are performed to get the analyzed results and find the patterns and the relationships between different variables.

First, to prepare the data for some analysis, we need to add a column to indicate if the accident is fatal. This will help to perform the decision tree analysis.

Second, convert string and categorical data into numbers which makes applying statistical model applicable. All statistical models cannot work with non-numerical data. Five columns were converted from string to number: ADDRESS, NEARESTINTSTREETNAME, NEARESTINTROUTEID, INTAPPROACHDIRECTION, and Rate.

Finally, we need to group the data to apply the statistical models. In this project, the data were grouped by accident rates.

Software and Hardware Development Platforms

We need to use some offered software to perform the data analysis, statistical models, and visualizations. This project mainly uses the Python programming language. To use Python, the researcher will work on Microsoft Visual Studio, and MS VS is software that can run many programming languages.

The hardware used to clean, prepare, and process the data has 16 Gb RAM and a 2.3 GHz Quad-Core Intel Core i5 process.

Data Analytics Algorithms

The data visualizations were done in this project show the relationships and trends among the datasets. Most of the charts were generated by Tableau software, and the charts were implemented after the data got cleaned and prepared for analysis.

Also, some data analytics algorithms are used to prepare the dataset for the statistical and analytical models. The first algorithm was used is the cut function for binning (Jain, 2020). This function allows us to classify the accidents using the total accident elements and injuries, making the analytics operations more resealable. In our case, we apply the statistical models to five groups.

The second algorithm was used to label the categorical variables. Because the statistical models cannot understand the string data types, we need a way to convert the string values into numerical values.

Data Analytics and Statistical Models

Seven statistical and analytical methods were applied to the data to understand the relationships between the different variables and answer the questions we asked in the introduction part. Some of these tools are descriptive, inferential, and advanced tools.

Descriptive Models:

The first model was used is confidence intervals. Confidence intervals give the estimated value in a variable to have happened with 90% and 95% probability. This model was examined multiple variables to understand that the most number might appear in most cases. For example, what are the total injuries and accident elements that could happen in 90% of the accidents accrued in Washington D.C, and what are the total injuries in 95% of the car crashes?

Then, we performed the correlation analysis to find the relationship between the different features. There are 15 features were used in this analysis: total vehicles, total pedestrians, pedestrians impaired, drivers impaired, total taxis, total government, speeding involved, fatal passenger, total fatal, total major injuries, total minor injuries, total unknown injuries, total accident elements, total injuries, and total injuries and elements.

Inferential Models

The third statistical model was used is ANOVA to examine an independent variable with two dependent variables.

The fourth, MANOVA which allows us to examine an independent variable with more than two dependent variables.

Advanced Models

The fifth one is the decision tree. We used decision tree analysis to find the reasons that lead to fatal accidents.

Visualizations

Finally, to see the model results, it needs to visualize them into charts, making it easy to understand the patterns and identify any relationships. Data visualizations could present patterns and trends in the dataset that are hard to find by looking at the values shown in the data, especially if the dataset was relatively large. Many types of visualizations could be used to illustrate the data. For example, the bar chart shows the frequency of the categorical variables. The map shows the geospatial points on a map, which helps find helpful information that could be used to find the car crashes patterns.

Experimental results and analysis

Confidence Interval

First, we perform the confidence intervals on multiple features: total injuries and elements, total injuries, total accident elements, total fatal, total major injuries, and total minor injuries. This can give us the chance of total injuries and vehicles in 90%-95% in the accidents. The results as follow:

Confidence Interval based on rates

Rate	count	TOTAL_INJURIES_ELEMENTS		TOTAL_INJURIES		TOTAL_ACCEIDENTELEMENTS		TOTAL_FATAL	
		ci95_hi	ci95_lo	ci95_hi	ci95_lo	ci95_hi	ci95_lo	ci95_hi	ci95_lo
extreme	6	50.76	35.56	47.96	33.03	3.32	2.01	0	0
high	49	17.94	15.97	12.46	9.20	7.34	4.89	0.14	-0.02
medium	129	11.84	11.58	7.08	6.15	5.55	4.63	0.07	0.005
low	121254	3.48	3.47	0.90	0.89	2.58	2.58	0.001	0.001
lowest	115755	1.91	1.91	0.03	0.03	1.88	1.88	0.002	0.001

Table 6: Confidence Interval based on accident rates

Confidence Interval based on variables

Variable	90%		95%	
	Low	High	Low	High
Total injuries and elements	2.723932	2.731337	2.723223	2.732046
Total injuries	0.479122	0.484784	0.478580	0.485326

Variable	90%		95%	
	Low	High	Low	High
Total accident elements	2.243440	2.247923	2.243010	2.248352
Total fatal	0.001851	0.002162	0.001821	0.002191
Total major injuries	0.107977	0.111084	0.107680	0.111381
Total minor injuries	0.293638	0.298074	0.293213	0.298498

TABLE 7: CONFIDENCE INTERVAL BASED ON VARIABLES

Correlation Analysis

The second model that was performed is the correlation analysis. The correlation analysis helps find the relationships between different variables to understand how the data related to each other and what factors have come together. There are two kinds of correlations: the positive correlation between 0.5 to 1.0 and the negative correlation between -0.5 to -1.0. The more significant number indicates to high correlation while the smaller number indicates weak correlations.

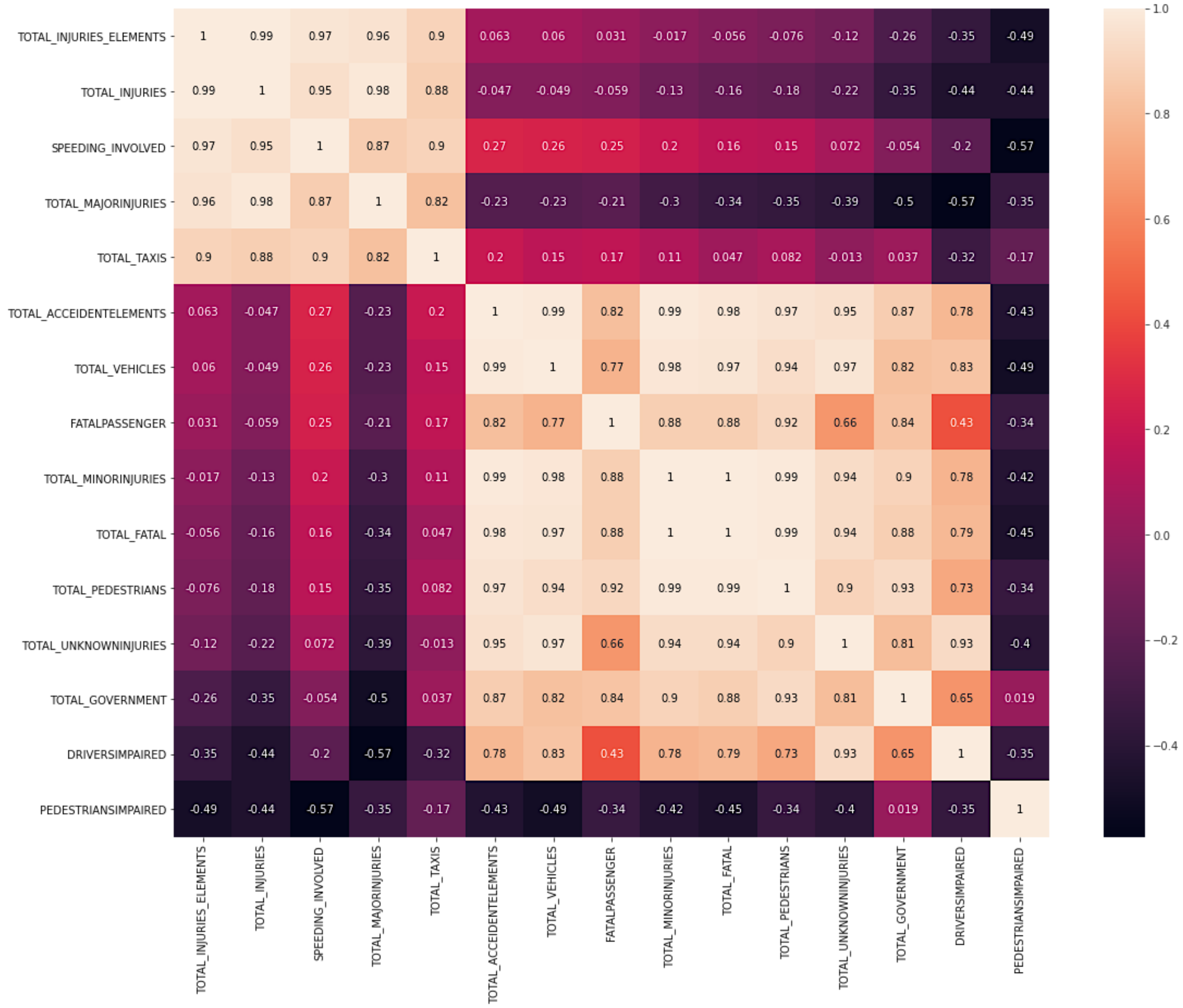


Figure 3: Correlation analysis

This visualization shows the correlation between the different variables. The correlation between 0.5 to 1.0 indicates a high positive correlation. On the contrary, the correlation between -0.5 to -1.0 indicates to high negative correlation. The chart gives vital information about how the feature relates to each other. However, if we ignore the similar features such as the total vehicles and total taxis because all of them are cars, it is normal to see a high correlation between these columns. The exciting result could be seen from the columns that have no relationships. For example, the total injuries and total taxis have 0.88, which indicates a high positive correlation. We can say that as the number of taxis accident accrue affects the number of injuries positively. In other words, taxis accidents cause more injuries than other vehicles involved in an accident. Nonetheless, we can conclude from this chart that there are high and low significant relationships between the different variables as follow:

There are many significant relations between different variables, here are the most notable once:

- the total number of injuries and accident elements with taxis.
- the total number of injuries and speed and taxis.
- the speed and
  - total accident and injuries
  - total injuries
  - total major injuries
  - total taxies
- major injuries with taxies
- Total number of accident elements with
  - fatal passenger
  - minor injuries
  - total fata
  - pedestrians
  - government
  - driver impaired
- Total number of vehicles involved in the accidents with:
  - Fatal passenger
  - minor injuries
  - total fatal

One-Way-ANOVA

The third statistical model that was applied is the One-Way-ANOVA algorithm tests the differences between one independent variable and two dependent variables. This can help find if the data are random or there is a significant relationship between the independent and dependent variables. Here are the results that we found:

Independent variable	Dependent variable 1	Dependent variable 2	p-value
TOTAL_INJURIES_ELEMENTS	DRIVERSIMPAIRED	SPEEDING_INVOLVED	**0.03978**
TOTAL_INJURIES_ELEMENTS	TOTAL_FATAL	SPEEDING_INVOLVED	**0.03974**
TOTAL_ACCEIDENTELEMENTS	DRIVERSIMPAIRED	PEDESTRIANSIMPAIRED	**0.000154**
TOTAL_FATAL	TOTAL_TAXIS	DRIVERSIMPAIRED	**0.01528**
TOTAL_FATAL	TOTAL_GOVERNMENT	DRIVERSIMPAIRED	0.05858
TOTAL_FATAL	TOTAL_TAXIS	SPEEDING_INVOLVED	**0.04698**
TOTAL_FATAL	TOTAL_GOVERNMENT	SPEEDING_INVOLVED	0.11406
FATALPASSENGER	TOTAL_TAXIS	SPEEDING_INVOLVED	**0.03778**
TOTAL_MAJORINJURIES	PEDESTRIANSIMPAIRED	SPEEDING_INVOLVED	0.28874

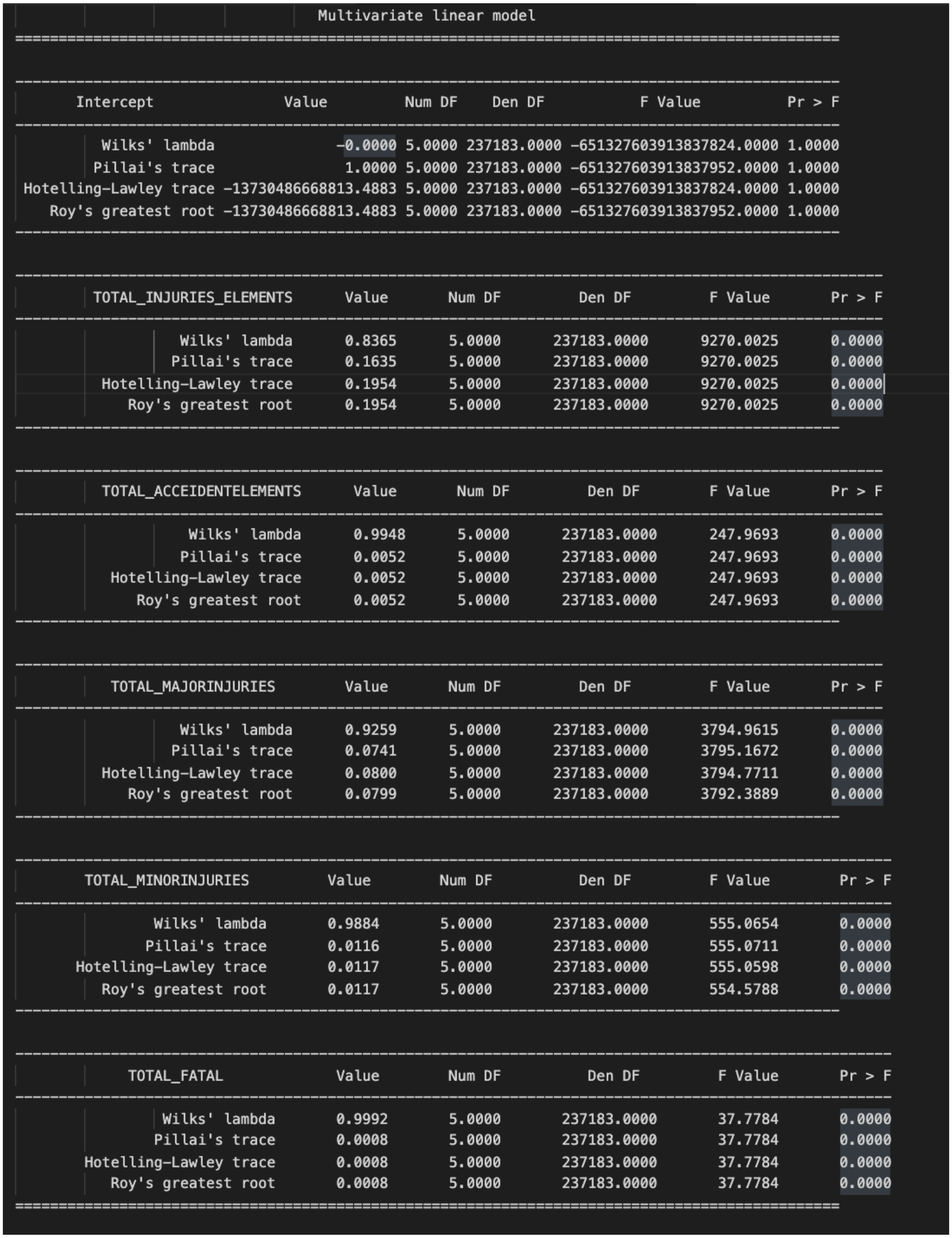
Table 8: One-Way-ANOVA results

The table above shows significant relationships between the independent and dependent variables where the p-value is less than 0.05. It can be seen that the total injures and elements have a significant relationship between impaired drivers and speed. Also, the total injuries have a significant relationship with the total fatal and speed. However, the total accident elements variable has a significant relationship with drivers impaired and pedestrians impaired. In addition, it can be noticed that the total fatal has significant relationships between multiple dependent variables: total taxis & drivers impaired, total taxis & speed. Nevertheless, there is no significant relationship between total government & drivers impaired and total government and speed.

MANOVA

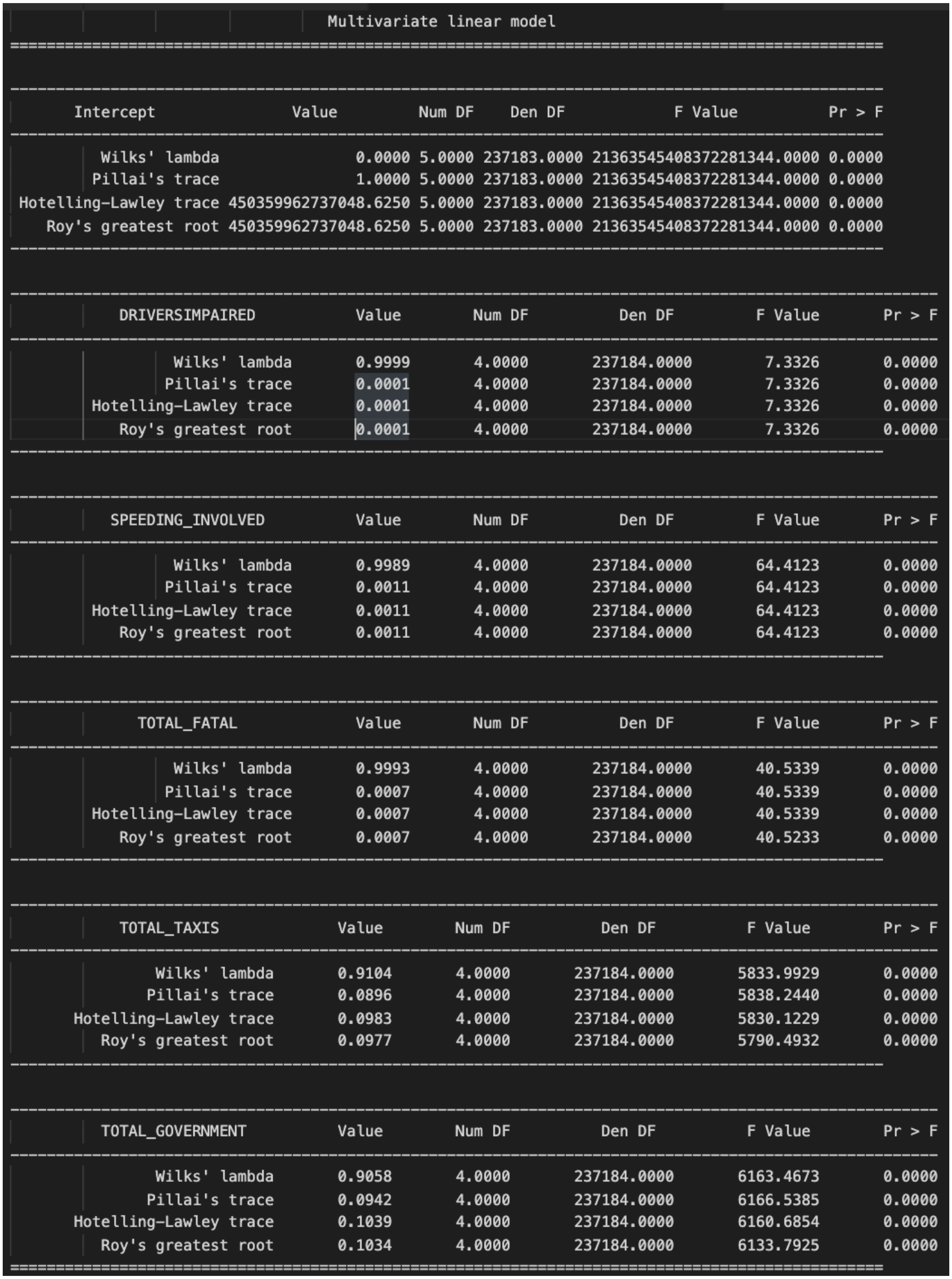
Fifth, MANOVA, this model is like the ANOVA. Nevertheless, the difference is examining more than one dependent variable to fit MANOVA, which is in our project is RATE column because it contains five groups. The following images show the results for the following variables:

- The first one shows Rate and:
    - TOTAL_INJURIES_ELEMENTS
    - TOTAL_ACCEIDENTELEMENTS
    - TOTAL_MAJORINJURIES
    - TOTAL_MINORINJURIES
    - TOTAL_FATAL
- The second image shows the Rate and:
    - DRIVERSIMPAIRED
    - SPEEDING_INVOLVED
    - TOTAL_FATAL
    - TOTAL_TAXIS
    - TOTAL_GOVERNMENT



```
                      Multivariate linear model
===============================================================================

-------------------------------------------------------------------------------
     Intercept             Value         Num DF    Den DF          F Value         Pr > F
-------------------------------------------------------------------------------
        Wilks' lambda              -0.0000 5.0000 237183.0000 -651327603913837824.0000 1.0000
       Pillai's trace               1.0000 5.0000 237183.0000 -651327603913837952.0000 1.0000
Hotelling-Lawley trace -13730486668813.4883 5.0000 237183.0000 -651327603913837824.0000 1.0000
   Roy's greatest root -13730486668813.4883 5.0000 237183.0000 -651327603913837952.0000 1.0000
-------------------------------------------------------------------------------

-----------------------------------------------------------------------------------
      TOTAL_INJURIES_ELEMENTS     Value      Num DF        Den DF        F Value      Pr > F
-----------------------------------------------------------------------------------
                 Wilks' lambda    0.8365     5.0000     237183.0000    9270.0025      0.0000
                Pillai's trace    0.1635     5.0000     237183.0000    9270.0025      0.0000
        Hotelling-Lawley trace    0.1954     5.0000     237183.0000    9270.0025      0.0000
           Roy's greatest root    0.1954     5.0000     237183.0000    9270.0025      0.0000
-------------------------------------------------------------------------------

-----------------------------------------------------------------------------------
       TOTAL_ACCEIDENTELEMENTS     Value      Num DF        Den DF        F Value      Pr > F
-----------------------------------------------------------------------------------
                  Wilks' lambda    0.9948     5.0000     237183.0000    247.9693       0.0000
                 Pillai's trace    0.0052     5.0000     237183.0000    247.9693       0.0000
         Hotelling-Lawley trace    0.0052     5.0000     237183.0000    247.9693       0.0000
            Roy's greatest root    0.0052     5.0000     237183.0000    247.9693       0.0000
-------------------------------------------------------------------------------

-----------------------------------------------------------------------------------
        TOTAL_MAJORINJURIES       Value      Num DF        Den DF        F Value      Pr > F
-----------------------------------------------------------------------------------
                 Wilks' lambda    0.9259     5.0000     237183.0000    3794.9615      0.0000
                Pillai's trace    0.0741     5.0000     237183.0000    3795.1672      0.0000
        Hotelling-Lawley trace    0.0800     5.0000     237183.0000    3794.7711      0.0000
           Roy's greatest root    0.0799     5.0000     237183.0000    3792.3889      0.0000
-------------------------------------------------------------------------------

------------------------------------------------------------------------------------
     TOTAL_MINORINJURIES        Value       Num DF         Den DF          F Value       Pr > F
------------------------------------------------------------------------------------
              Wilks' lambda     0.9884      5.0000      237183.0000      555.0654       0.0000
             Pillai's trace     0.0116      5.0000      237183.0000      555.0711       0.0000
     Hotelling-Lawley trace     0.0117      5.0000      237183.0000      555.0598       0.0000
        Roy's greatest root     0.0117      5.0000      237183.0000      554.5788       0.0000
-------------------------------------------------------------------------------

------------------------------------------------------------------------------------
         TOTAL_FATAL            Value       Num DF         Den DF          F Value       Pr > F
------------------------------------------------------------------------------------
               Wilks' lambda    0.9992      5.0000      237183.0000      37.7784        0.0000
              Pillai's trace    0.0008      5.0000      237183.0000      37.7784        0.0000
      Hotelling-Lawley trace    0.0008      5.0000      237183.0000      37.7784        0.0000
         Roy's greatest root    0.0008      5.0000      237183.0000      37.7784        0.0000
===============================================================================
```

FIGURE 4: FIRST MANOVA RESULTS



```
                                   Multivariate linear model
================================================================================================

------------------------------------------------------------------------------------------------
       Intercept               Value         Num DF    Den DF           F Value          Pr > F
------------------------------------------------------------------------------------------------
          Wilks' lambda              0.0000 5.0000 237183.0000 21363545408372281344.0000 0.0000
         Pillai's trace              1.0000 5.0000 237183.0000 21363545408372281344.0000 0.0000
 Hotelling-Lawley trace 450359962737048.6250 5.0000 237183.0000 21363545408372281344.0000 0.0000
    Roy's greatest root 450359962737048.6250 5.0000 237183.0000 21363545408372281344.0000 0.0000
------------------------------------------------------------------------------------------------

------------------------------------------------------------------------------------------------------
            DRIVERSIMPAIRED           Value        Num DF          Den DF          F Value       Pr > F
------------------------------------------------------------------------------------------------------
                  Wilks' lambda       0.9999       4.0000       237184.0000        7.3326       0.0000
                 Pillai's trace       0.0001       4.0000       237184.0000        7.3326       0.0000
         Hotelling-Lawley trace       0.0001       4.0000       237184.0000        7.3326       0.0000
            Roy's greatest root       0.0001       4.0000       237184.0000        7.3326       0.0000
------------------------------------------------------------------------------------------------

------------------------------------------------------------------------------------------------------
           SPEEDING_INVOLVED          Value        Num DF          Den DF          F Value       Pr > F
------------------------------------------------------------------------------------------------------
                  Wilks' lambda       0.9989       4.0000       237184.0000       64.4123       0.0000
                 Pillai's trace       0.0011       4.0000       237184.0000       64.4123       0.0000
         Hotelling-Lawley trace       0.0011       4.0000       237184.0000       64.4123       0.0000
            Roy's greatest root       0.0011       4.0000       237184.0000       64.4123       0.0000
------------------------------------------------------------------------------------------------

------------------------------------------------------------------------------------------------------
               TOTAL_FATAL            Value        Num DF          Den DF          F Value       Pr > F
------------------------------------------------------------------------------------------------------
                  Wilks' lambda       0.9993       4.0000       237184.0000       40.5339       0.0000
                 Pillai's trace       0.0007       4.0000       237184.0000       40.5339       0.0000
         Hotelling-Lawley trace       0.0007       4.0000       237184.0000       40.5339       0.0000
            Roy's greatest root       0.0007       4.0000       237184.0000       40.5233       0.0000
------------------------------------------------------------------------------------------------

------------------------------------------------------------------------------------------------------
            TOTAL_TAXIS             Value        Num DF          Den DF          F Value         Pr > F
------------------------------------------------------------------------------------------------------
                Wilks' lambda       0.9104       4.0000       237184.0000       5833.9929       0.0000
               Pillai's trace       0.0896       4.0000       237184.0000       5838.2440       0.0000
       Hotelling-Lawley trace       0.0983       4.0000       237184.0000       5830.1229       0.0000
          Roy's greatest root       0.0977       4.0000       237184.0000       5790.4932       0.0000
------------------------------------------------------------------------------------------------

------------------------------------------------------------------------------------------------------
          TOTAL_GOVERNMENT          Value        Num DF          Den DF          F Value         Pr > F
------------------------------------------------------------------------------------------------------
                Wilks' lambda       0.9058       4.0000       237184.0000       6163.4673       0.0000
               Pillai's trace       0.0942       4.0000       237184.0000       6166.5385       0.0000
       Hotelling-Lawley trace       0.1039       4.0000       237184.0000       6160.6854       0.0000
          Roy's greatest root       0.1034       4.0000       237184.0000       6133.7925       0.0000
================================================================================================
```

Figure 5: Second MANOVA results

Decision Tree

The sixth model used is a kind of machine learning model, a decision tree algorithm. The decision tree can help find the factors that lead to a specific event. In this project, we use a decision tree to find the factors that lead to fatal accidents, which could help understand the causes that might lead to deadly accidents.

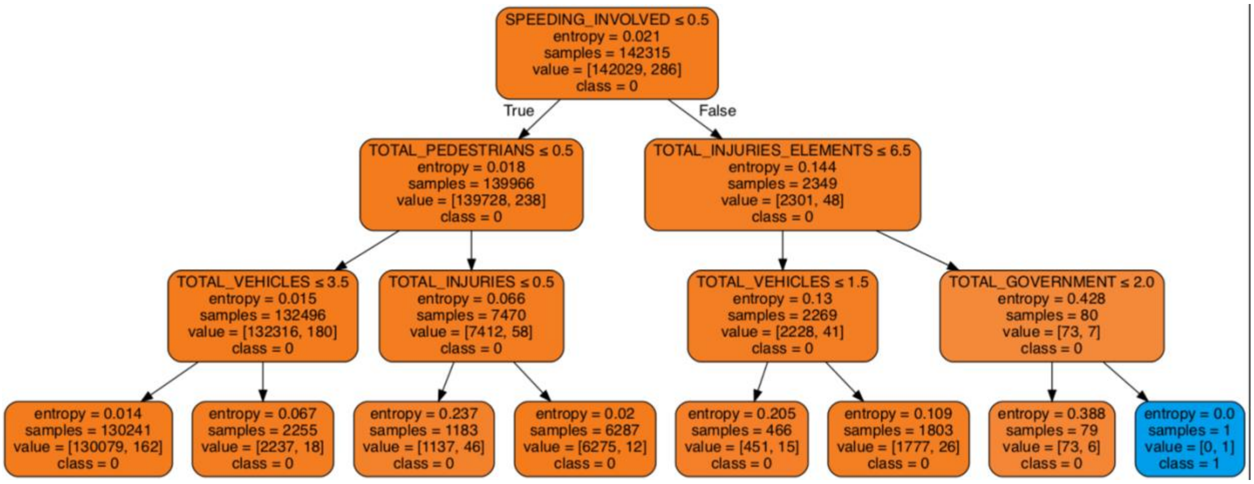


FIGURE 6: DECISION TREE

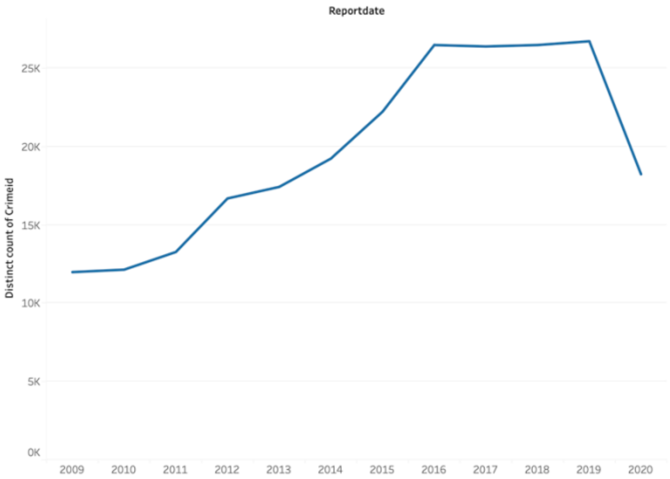
This decision tree shows that speed is a significant cause of deadly accidents. This chart shows four-level depth, which gives a scenario if the accident has the speeding case and the total injuries and accident elements less than 6.5 and there are government cars involved. The chance the accident has a fatal case is high. The model accuracy is 99.8% which is high accuracy.

If we want to see the entire scenario, the perfect depth level is five. This chart can be found in the Appendix C: Decision Tree 5 levels.

Data Visualization

As known, data visualizations are a valuable way to understand the data. All charts in this section were generated by Tableau software (Chabot et al., 2021), which makes creating the chart straightforward and provides a feature that allows the developer to create attractive visualizations. This part contains four charts describing the data distribution to find the trends and valuable information. The first chart shows the car accidents over time, and the second shows the crashes over time but not the lowest and low ones. The third visualization is bar chart demonstrations to top 20 streets hold the most accidents more than others. The final one shows the map of the District of Columbia with the top 10 streets and all the accidents show in points.

Total number of accidents over time



This chart shows the accident number from 2009 to 2020. It can be seen that the number of accidents rose from 2009 to 2016 from 11,982 to 26,470 respectively. After, the numbers changed slightly from 2016 until 2019. However, the accident number hit the peak in 2019 with 26,711 accidents. In contrast, car crashes dropped significantly in 2020, with 18,230 accidents. The reason for that drop because of the Covid-19 lockdown.

FIGURE 7: TOTAL NUMBER OF ACCIDENTS OVER TIME

Total number of extreme, high, and medium accidents

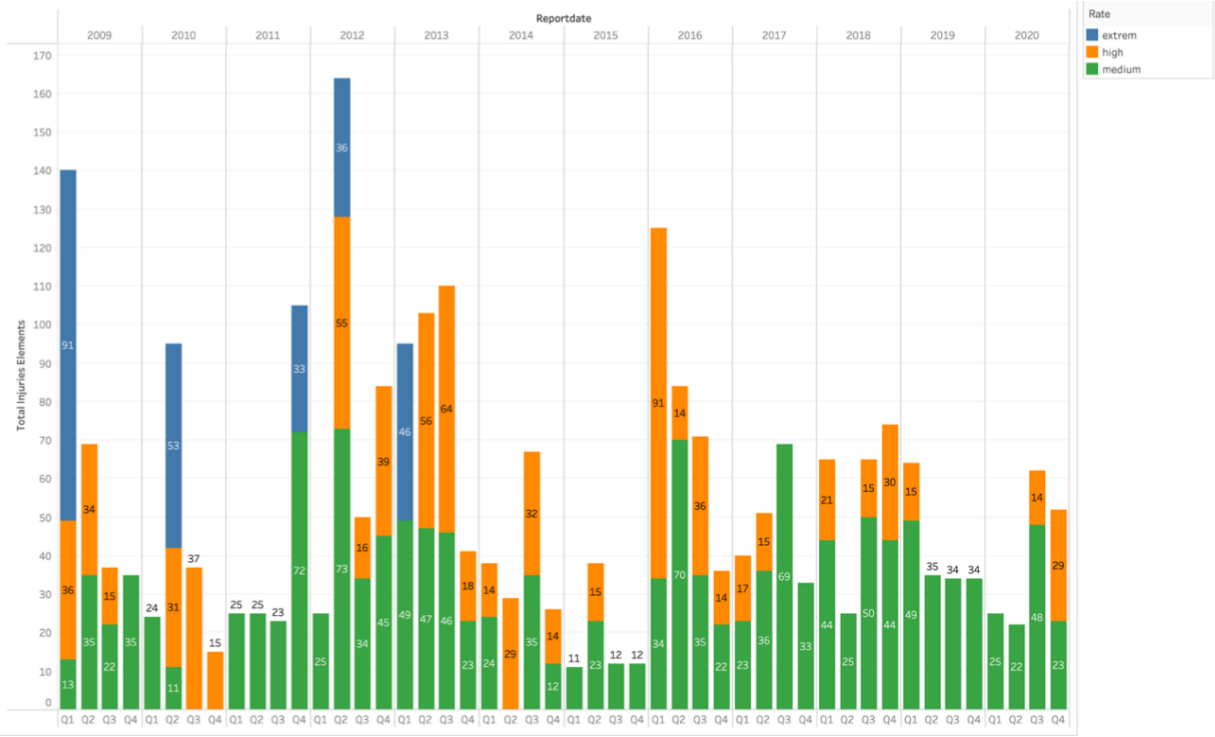


FIGURE 8: NUMBER OF EXTREME, HIGH, AND MEDIUM ACCIDENTS

This chart shows the number of extreme, high, and medium accidents from 2009 to 2020 divided by quarters of each year. The number shown in the bars represents the total number of injuries and the number of vehicles involved in those accidents. It can be seen that the extreme accidents stopped in 2013. Also, the number of these accidents hit the peak in 2012 with around 163 accidents with 144 in the total injuries and vehicles in the second quarter. Also, the highest number of extreme accidents happened in 2009 in the first quarter, with 90 accidents. This chart helps to find accidents distribution through the time series and understand the relationship between the kind of accident and the year those accidents happened.

Top 20 streets with the number of accidents

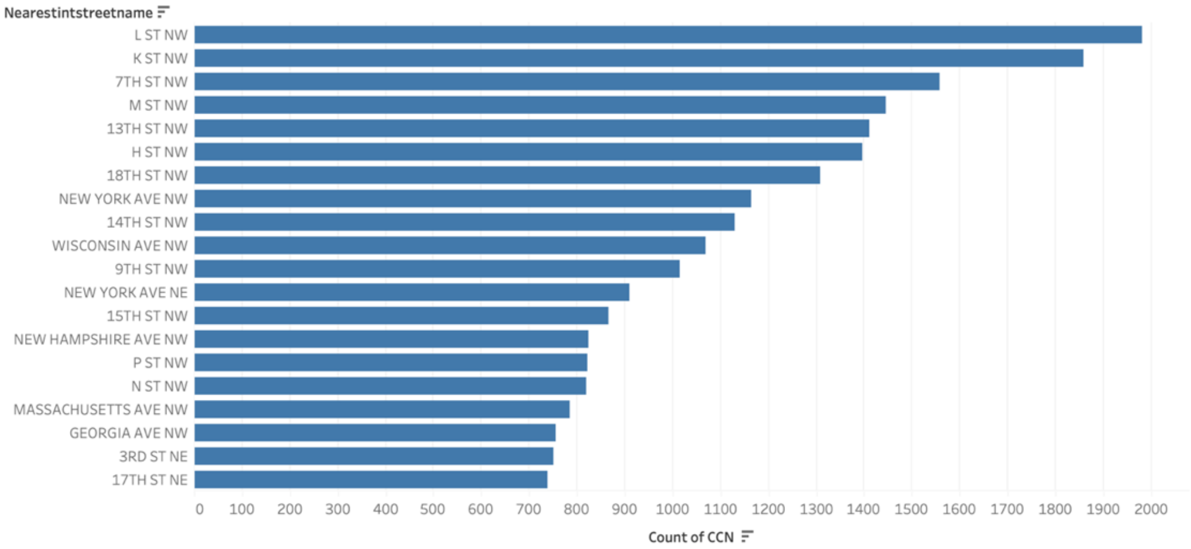


FIGURE 9: TOP 20 STREETS WITH THE NUMBER OF ACCIDENTS

This bar graph represents the top 20 streets with the number of accidents in those streets sorted ascending. It can be noticed that most accidents happen in the northwest regions. From this chart, we can conclude that the streets located in the northwest have the highest chance of having accidents more than the other regions in Washington D.C.

Top 10 streets that contained accidents

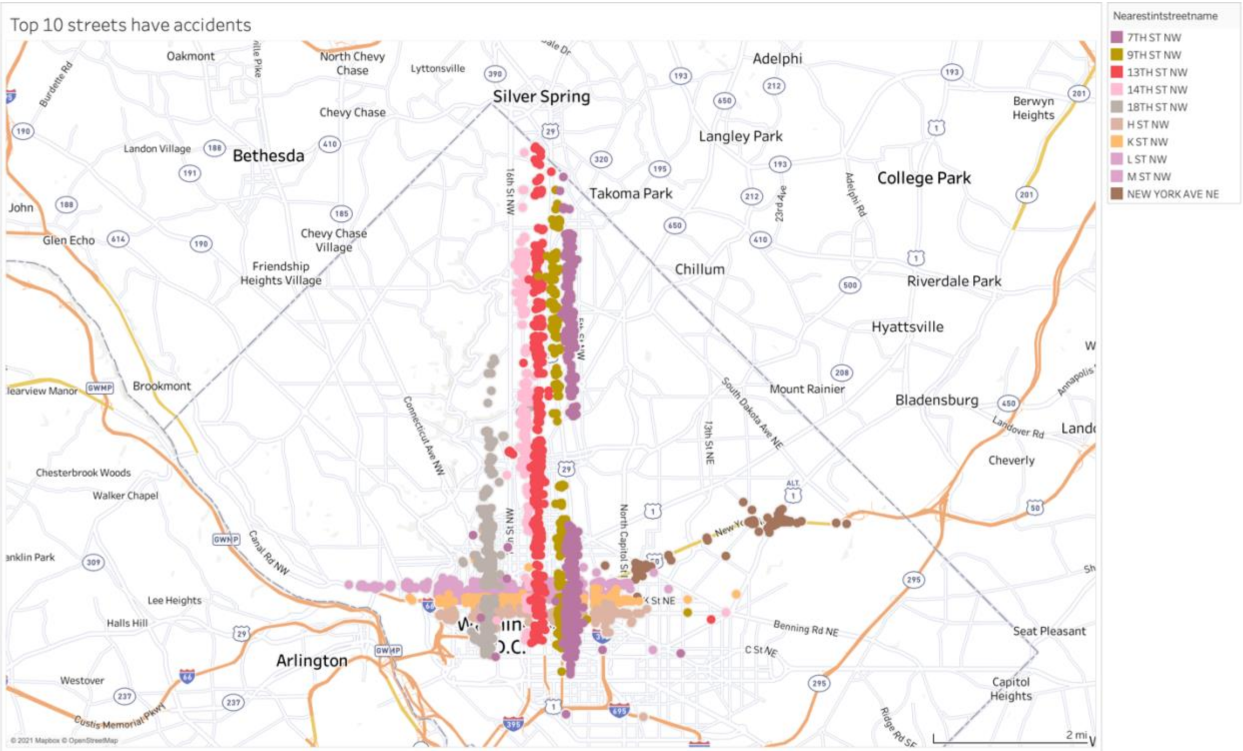


FIGURE 10: MAP OF TOP 10 STREETS THAT CONTAINED ACCIDENTS

The map shows the top 10 streets that contained accidents between 2009 and 2020. It can be seen from this map that the accidents focused on the city center and in the whole streets that lead to out of town from the northeast side. However, there are some accidents on both sides (northeast and southwest), but they are much fewer than those on the northwest side. This diagram helps find the most streets with the highest number of accidents to study the reasons and find ways to solve them.

Conclusions

In conclusion, this project studies the dataset related to the car accidents in Washington, D.C, between 2009 to 2020. The data has multiple processes to be ready for analytical models. First, we cleaned the data by dropping the unrelated columns and null values that the known ways could not fill. Then the data were explored to understand the data distribution and find the columns that may benefit this analysis. After, the data was prepared to be ready for the analytical and statistical models. The preparation processes include classifying the accidents into five groups (lowest, low, medium, high, extreme) and creating a separate dataset that contains the categorical data mapped the string values into numbers. Then we have applied correlation analysis, ANOVA, confidence interval, decision tree model, and visualizations. These models were used to find the trend and patterns among the data and find any significant relationships between different variables. This project aims to help the authorities to understand the car accidents within the area so they can find the proper solutions to reduce the number of accidents and fix any issues that can be fixed. We have found that most crashes in the area are caused by impaired persons, while the deadly accidents happen if one of the accident cars was driving fast or a government car is involved in the accidents. Also, we have found a significant relationship between the number of fatal and the number of taxis involved in the accidents. In addition, most roads that hold the accidents are located in the north area of the district.

On the other hand, we believe that this project could be enhanced by analyzing the crash address and zip codes. There are some issues in the address feature that need to be handled. For instance, some addresses are not complete, and others contain missing numbers or street names to name a few. Also, we have found that there is no information about the impaired people before 2015. This issue needs to find the proper way to solve. If these problems are fixed, we think the results could be better.

Appendix A: Data sample

CCN	REPORTDATE	FROMDATE	LATITUDE	LONGITUDE	UNKNOWNINJ	TOTAL_VEHIC	TOTAL_BICYCL	TOTAL_PEDES	PEDESTRIANSI	BICYCLISTSIM	DRIVERSIMPA	INTAPPROACHDIRECTION	TOTAL_FATAL	TOTAL_MAJO	TOTAL_MINO	TOTAL_UNKN	TOTAL_ACCEI	TOTAL_INJURI
21093410	2021-07-08 17:24:01+00:00	2021-07-08 04:00:00+00:00	38.865871	-76.980153	1	2	0	0	0	0	0	West	0	0	0	1	2	1
21093404	2021-07-08 17:29:31+00:00	2021-07-08 04:00:00+00:00	38.905186	-77.061788	0	1	1	0	0	0	0	East	0	0	1	0	2	1
21093438	2021-07-08 17:41:24+00:00	2021-07-08 04:00:00+00:00	38.916443	-77.02231	0	2	0	0	0	0	0	Northwest	0	0	0	0	2	0
21093392	2021-07-08 17:43:49+00:00	2021-07-08 04:00:00+00:00	38.961973	-77.02795	0	2	0	0	0	0	0	East	0	0	0	0	2	0
21093440	2021-07-08 17:45:38+00:00	2021-07-08 04:00:00+00:00	38.933908	-77.022805	2	3	0	0	0	0	0	East	0	0	0	2	3	2
21093458	2021-07-08 18:22:06+00:00	2021-07-07 04:00:00+00:00	38.92388	-77.03557	0	2	0	0	0	0	0	North	0	0	0	0	2	0
21090593	2021-07-09 04:10:25+00:00	2021-07-03 04:00:00+00:00	38.873629	-76.977491	0	3	0	0	0	0	0	East	0	0	0	0	3	0
21093448	2021-07-09 18:18:35+00:00	2021-07-08 04:00:00+00:00	38.915381	-77.020455	0	2	0	0	0	0	0	Northwest	0	0	0	0	2	0
21095310	2021-07-12 04:03:36+00:00	2021-07-11 04:00:00+00:00	38.920091	-77.027036	0	1	0	0	0	0	1	East	0	0	1	0	1	1
21095285	2021-07-12 03:32:46+00:00	2021-07-11 04:00:00+00:00	38.823287	-76.999585	1	3	0	0	0	0	1	South	0	0	0	1	3	1
21095336	2021-07-12 01:44:44+00:00	2021-07-10 04:00:00+00:00	38.958513	-77.036771	0	2	0	0	0	0	0	North	0	0	0	0	2	0
21095334	2021-07-12 03:40:18+00:00	2021-07-11 04:00:00+00:00	38.905352	-76.996288	0	1	1	0	0	0	0	Southeast	0	0	1	0	2	1
21095313	2021-07-12 02:08:29+00:00	2021-07-11 04:00:00+00:00	38.871022	-76.9704	0	2	0	0	0	0	0	East	0	0	0	0	2	0
21095283	2021-07-12 02:22:23+00:00	2021-07-11 04:00:00+00:00	38.896874	-77.006443	0	2	0	0	0	0	0	East	0	0	0	0	4	0

CCN	REPORTDATE	FROMDATE	LATITUDE	LONGITUDE	UNKNOWNINJ	TOTAL_VEHIC	TOTAL_BICYCL	TOTAL_PEDES	PEDESTRIANSI	BICYCLISTSIM	DRIVERSIMPA	INTAPPROACHDIRECTION	TOTAL_FATAL	TOTAL_MAJO	TOTAL_MINO	TOTAL_UNKN	TOTAL_ACCEI	TOTAL_INJURI
21095344	2021-07-12 02:30:13+00:00	2021-07-11 04:00:00+00:00	38.852412	-76.968901	0	1	0	0	0	0	0	Northwest	0	0	0	0	2	0
21095360	2021-07-12 03:31:30+00:00	2021-07-11 04:00:00+00:00	38.906117	-77.014415	1	2	0	0	0	0	0	North	0	0	0	1	2	1
21095227	2021-07-12 03:09:52+00:00	2021-07-11 04:00:00+00:00	38.936056	-77.037149	0	2	0	0	0	0	0	South	0	0	1	0	2	1
21095356	2021-07-12 03:33:43+00:00	2021-07-11 04:00:00+00:00	38.871965	-76.989817	0	5	0	0	0	0	0	East	0	0	0	0	5	0
21054232	2021-04-28 23:29:10+00:00	2021-04-28 04:00:00+00:00	38.839678	-76.98942	0	2	0	0	0	0	0	West	0	0	0	0	2	0
21054244	2021-04-29 00:00:21+00:00	2021-04-28 04:00:00+00:00	38.839773	-77.009088	0	2	0	0	0	0	0	East	0	0	0	0	2	0
21054258	2021-04-29 00:06:52+00:00	2021-04-28 04:00:00+00:00	38.93669	-77.028014	0	2	0	0	0	0	0	East	0	0	0	0	2	0
21054162	2021-04-29 00:12:27+00:00	2021-04-28 04:00:00+00:00	38.909012	-77.002573	0	2	0	0	0	0	0	Northwest	0	0	0	0	2	0
21054272	2021-04-29 00:51:40+00:00	2021-04-28 04:00:00+00:00	38.87812	-76.930619	0	2	0	0	0	0	0	South	0	0	0	0	2	0
21054285	2021-04-29 00:51:24+00:00	2021-04-28 04:00:00+00:00	38.888415	-76.937449	0	2	0	0	0	0	0	North	0	0	0	0	2	0
21054268	2021-04-29 01:05:00+00:00	2021-04-28 04:00:00+00:00	38.869029	-77.006108	0	1	0	0	0	0	0	East	0	0	0	0	2	0
21054298	2021-04-29 03:29:20+00:00	2021-04-28 04:00:00+00:00	38.895312	-76.94847	0	2	0	0	0	0	1	Northwest	0	0	0	0	2	0

Table 9: Data sample

Appendix B: Table of Figures and tables

Table of figures

Table of Tables

Appendix C: Decision Tree 5 levels

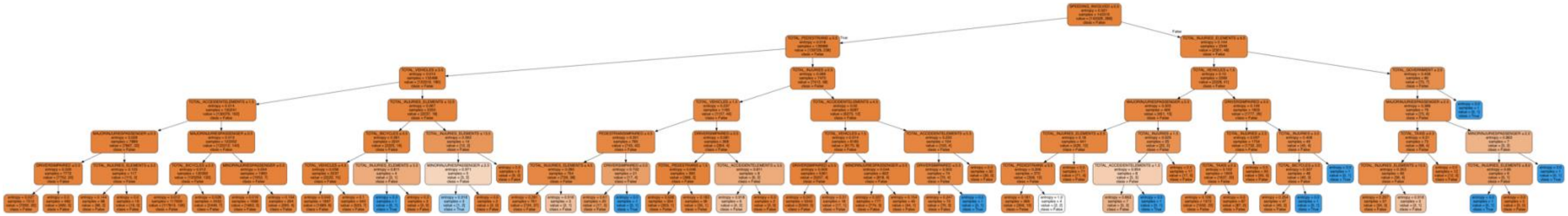


FIGURE 11: DECISION TREE 5 LEVELS

References:

Anaconda Software Distribution. (2020). Anaconda Documentation. Anaconda Inc. Retrieved from https://docs.anaconda.com/

Chabot, C., Beers, A., & Hanrahan, P. (2021). *Tableau* (2021.3.3) [Computer software]. Tableau. https://www.tableau.com

Chesin, M. (2019, December 10). *Motor Vehicle Accident* [Photograph]. Unsplash. https://unsplash.com/photos/ZI-vWZBbwj8

Data Commons. (n.d.). *Washington, D.C. Demographics - Place Explorer - Data Commons*. Retrieved November 18, 2021, from https://datacommons.org/place/geoId/11001?topic=Demographics

District Department of Transportation. (2021). *Crashes in DC* [These data represent the crash locations associated along the DDOT centerline network within the District of Columbia. In addition to locations, a related table consisting of crash details is available for each crash]. District Department of Transportation. https://opendata.dc.gov/datasets/DCGIS::crashes-in-dc/about

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science &amp; Engineering, 9(3), 90–95

Jain, A. (2020, June 26). *Pandas In Python | Data Manipulation With Pandas*. Analytics Vidhya. Retrieved November 30, 2021, from https://www.analyticsvidhya.com/blog/2016/01/12-pandas-techniques-python-data-manipulation/

National Highway Traffic Safety Administration. (2021, October 28). *USDOT Releases New Data Showing That Road Fatalities Spiked in First Half of 2021*. NHTSA. Retrieved November

18, 2021, from https://www.nhtsa.gov/press-releases/usdot-releases-new-data-showing-road-fatalities-spiked-first-half-2021

Templeton Smithee Hayes Heinrich & Russell, LLP. (2020). *Amarillo Car Accident Lawyers | 200+ Years of Combined Experience*. Retrieved November 18, 2021, from https://www.templetonsmithee.com/personal-injury/car-accidents/

The pandas development team. (2020). *pandas-dev/pandas: Pandas* (3.8.8) [Python library]. Zenodo. https://doi.org/10.5281/zenodo.3509134

VanderPlas, J., Granger, B., Heer, J., Moritz, D., Wongsuphasawat, K., Satyanarayan, A., Lees, E., Timofeev, I., Welsh, B., & Sievert, S. (2018). Altair: Interactive Statistical Visualizations for Python. *Journal of Open Source Software*, *3*(32), 1057. https://doi.org/10.21105/joss.01057

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn. *GetMobile: Mobile Computing and Communications*, *19*(1), 29–33. https://doi.org/10.1145/2786984.2786995

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . van Mulbregt, P. (2020). Author Correction: SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 352. https://doi.org/10.1038/s41592-020-0772-5

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., . . . Qalieh, A. (2017, September 3). mwaskom/seaborn: v0.8.1 (September 2017). Zenodo. Retrieved December 2, 2021, from https://zenodo.org/record/883859

Visual Studio Code (1.61.0). (2019). [Software]. Microsoft. https://code.visualstudio.com

Wilson, L. T., & Wilson, L. T. (n.d.). *Statistical Correlation*. Explorable. Retrieved October 15, 2021, from https://explorable.com/statistical-correlation

World Health Organization. (2021, June 21). *Road traffic injuries*. Retrieved November 17, 2021, from https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries