Using Descriptive Analysis to Find Patterns and Trends: A Case of Car Accidents in Washington D.C.

Zaid M. Altukhi, Nasser F. Aljohani
Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia

***Abstract*—The descriptive analysis could be used to find the trends and patterns in historical data. In this article, descriptive analysis has been used to describe the car accidents in Washington, D.C., between 2009 and 2020. The dataset was downloaded from the District Department of Transportation (DDOT), the department responsible for car accidents in Washington, D.C. Multiple analytics and statistical models have been applied 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 article aims to find the common reasons for accidents and help DDOT find ways to reduce and eliminate 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 drunken 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.**

Keywords*—*Descriptive analysis; trends; patterns; analytics; statistics; car accidents

I. Introduction

What are the reasons that car accidents are one of the most dangerous events on roads? The World Health Organization says that around 1.3 million persons die because of car accidents which cost around 3% of most nations' gross domestic [1]. 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 [2]. In the first half of 2021, a projected 20,160 individuals died in car accidents, rising 18.4 per cent over the same period in 2020, and since 2006, this has been the highest number of expected fatalities in that period [2]. There are two main reasons that can cause crashes on roads. First, external effects are the effects that the driver cannot avoid or are hard to avoid, such as 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 tier issues. According to [3], 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, [4] improper turns and driving lean are all factors to consider.

According to data acquired from the District Department of Transportation, about 258,000 accidents occurred in Washington, D.C., between 2009 and 2020 [4]. 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 [5].

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

II. Objective

This article has analyzed the car accidents in the Washington, D.C. area from 2009 to 2020. The researchers have examined the common factors between accidents, the locations of the accidents, and the car crashes factors that may cause deaths or injuries compared to the number of accidents; also, the factors that significantly correlate with the number of injuries and accident elements. The accident elements are vehicles involved in an accident. In addition, we 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.

III. Dataset

A. 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) [4]. The data that was downloaded contains 258,122 records and 60 features. However, 19 features have been dropped because they have 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.

B. Preprocessing and Exploring the Dataset

The shape of the final dataset that will be analyzed contains 237,193 and 55 columns. After exploring the dataset and determining the needs, many processes were performed to clean and prepare the data. The researchers have used 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 follows:

- Anaconda [6].
- Pandas [7].
- Numpy [8].
- Sklearn [9].
- SciPy [10].
- Altair [11].
- Matplotlib [12].
- Seaborn [13].

C. Feature Engineering

Once the original dataset had been imported, there 18 columns were dropped because either they had too many null values or have not relevant to our analysis. Then there, nine columns were added, as shown in Table I below. They all aggregate multiple numeric columns except the rate column, which classifies the crashes into categories.

Then, there were some columns need to edit their data types from numerical to other type of data as shown in Table II.

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 remain on the accidents that happened between 2009 and 2020. The original dataset contains accident information up to August 2021. However, it decided to remove the accidents that happened in 2021 because it may affect the analysis results since they were just for eight months.

D. 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 had zero accident elements and zero injuries. Those accidents were also removed.

E. Classify Accidents

To analyze the accidents, it must be categorized. In this article, the authors have decided to divide the accidents into five categories lowest, low, medium, high, and extreme. The categories are done by using the cut-point function provided in Panda's library [7]. The cut points range is shown in the Table III. However, these numbers were picked based on the observations and data distribution. The result of this rank is as follows:

F. Explore the Data

To properly understand the data and distribution, it is necessary to present statistical information and visualize the data [14]. First, using the describe function on the data frame that contains the data we provide vital information about the data. The Table IV shows the count, mean, standard deviation, min, max, 25%, 50%, and 75% for the added columns.

The data distribution is displayed in the Table IV. A richness of details assists in deciphering the data that can be acquired from simply reading the number. 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 data. To better comprehend the data range and distribution, various graphics have been generated.

TABLE I. NEW COLUMNS

Column Name	Description
TOTAL FATAL	Sum of all attributes that contain fatalities data, which are: driver, bicyclist, pedestrian, and passengers.
TOTAL MAJOR INJURIES	Sum of all attributes that contain major injuries data, which are: driver, bicyclist, pedestrian, and passengers.
TOTAL MINOR INJURIES	Sum of all attributes that contain minor injuries data, which are: driver, bicyclist, pedestrian, and passengers.
TOTAL UNKNOWN INJURIES	The sum of all attributes that contain unknown injury data, which are: driver, bicyclist, pedestrian, and passengers.
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 II. EDITED DATA TYPES

Column name	Old data type	New data type
REPORTDATE	String	Date/time
FROMDATE	String	Data/time
OBJECTID	Integer	String
CRIMEID	Integer	String
ROUTEID	Integer	String
MARID	Integer	String

TABLE III. ACCIDENT RATES WITH THE TOTAL NUMBER OF ACCIDENTS IN EACH CATEGORY

Rank	Range	Number of accidents
Lowest	0-2	115,755
Low	3-10	121,254
Medium	11-13	129
High	14-30	49
Extreme	>30	6

TABLE IV. NEW COLUMNS DESCRIPTION

	TOTAL FATAL	TOTAL MAJOR INJURIES	TOTAL UNKNOWN INJURIES	TOTAL ACCEIDENT ELEMENTS	TOTALINJURIES
count	237193	237193	237193	237193	237193
Sum	476	25,980	18,161	532,660	114,316
mean	0.002007	0.109531	0.076566	2.245682	0.48195
std	0.045961	0.459874	0.311264	0.66367	0.83825
min	0	0	0	0	0
25%	0	0	0	2	0
50%	0	0	0	2	0
75%	0	0	0	3	1
max	2	51	16	17	51

The chart in Fig. 1 shows the total number of elements and injuries based on the sector. It can be noticed that there are some outliers and the car crash distribution based on these regions.

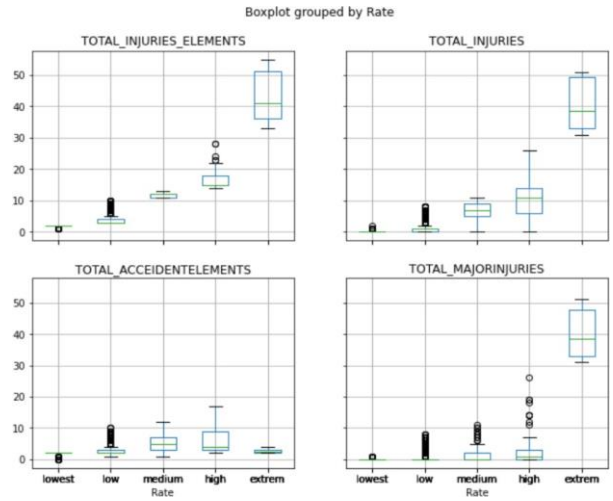

Fig. 1. Distribution of the total number of accidents and injuries based on the accident rate.

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.

IV. THE SYSTEM

This article has examined many statistical and analytical models to find the patterns and trends in data. The first model that has been used is the correlations between features. This model can provide a good idea of 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 accidents could be fatal to people. In addition, visualizations are a great way to explore the data and find the outliers and data distribution. Understanding the data makes 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 week days or weekend days and the number of accidents or injuries? What are the common factors between accidents that 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 such a way that 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.

A. 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, a new column has been added to show whether the accident resulted in death. This can 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.

B. 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 [15]. 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.

C. Data Analytics Algorithms

The data visualizations done in this project show the relationships and trends among the datasets. Most charts were generated by Tableau software [16], and the charts were implemented after the data was 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 used is the cut function for binning [17]. 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.

D. Data Analytics and Statistical Models

Several 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 [18] [19] [20].

1) Descriptive models: The first model used is confidence intervals. Confidence intervals give the estimated value of a variable to have happened with 90% and 95% probability [18]. This model 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, a correlation analysis was conducted to determine the relationship between the various features. There are 15 features 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.

2) Inferential models: The third statistical model used is ANOVA to examine an independent variable with two dependent variables. The fourth, MANOVA, allows us to examine an independent variable with more than two dependent variables [19].

3) Advanced models: The fifth one is the decision tree. We used decision tree analysis to find the reasons that lead to fatal accidents.

4) 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 is 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 crash patterns.

V. Experimental Results and Analysis

A. 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% of accidents. The results are presented in Table V and Table VI.

TABLE V. Confidence Interval based on Accident Rates

Rate	count	TOTAL INJURIES		TOTAL ACCEIDENT ELEMENTS		TOTAL_FATAL	
		ci95 High	ci95 Low	ci95 High	ci95 Low	ci95 High	ci95 Low
extreme	6	47.96	33.03	3.32	2.01	0	0
high	49	12.46	9.20	7.34	4.89	0.14	-0.02
medium	129	7.08	6.15	5.55	4.63	0.07	0.005
low	121254	0.90	0.89	2.58	2.58	0.001	0.001
lowest	115755	0.03	0.03	1.88	1.88	0.002	0.001

TABLE VI. 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
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

B. Correlation Analysis

The second model that was performed was correlation analysis. The correlation analysis helps find the relationships between different variables to understand how the data relate to each other and what factors have come together. There are two kinds of correlations: the positive correlation between 0.5 and 1.0 and the negative correlation between -0.5 and -1.0. The more significant number indicates high correlations, while the smaller number indicates weak correlations [19].

The 20 highest correlations between the variables are displayed in Table VII. The correlation between 0.5 and 1.0 indicates a high positive correlation. On the contrary, the correlation between -0.5 to -1.0 indicates a high negative correlation. The table gives vital information about how the features relate to each other. However, if we ignore 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 taxis have 0.88, indicating a highly positive correlation. We can say that the number of taxis accident accrue affects the number of injuries positively. In other words, taxi accidents cause more injuries than other vehicles involved in an accident. Nonetheless, we can conclude from this table that there are highly significant relationships between the different variables as follow.

TABLE VII. THE TOP 20 HIGHEST CORRELATIONS

Variable 1	Variable 2	Correlation
Total Fatal	Total Minor Injuries	0.996284
Total Injuries Elements	Total Injuries	0.993960
Total Accident Elements	Total Vehicles	0.992893
Total Pedestrians	Total Minor Injuries	0.991245
Total Minor Injuries	Total Accident Elements	0.990081
Total Pedestrians	Total Fatal	0.988103
Total Injuries	Total Major Injuries	0.983547
Total Accident Elements	Total Fatal	0.978477
Total Vehicles	Total Minor Injuries	0.977746
Speeding Involved	Total Injuries Elements	0.974481
Total Vehicles	Total Unknown Injuries	0.974045
Total Fatal	Total Vehicles	0.971432
Total Pedestrians	Total Accident Elements	0.966891
Total Major Injuries	Total Injuries Elements	0.957828
Total Unknown Injuries	Total Accident Elements	0.954655
Speeding Involved	Total Injuries	0.945711
Total Pedestrians	Total Vehicles	0.941978
Total Unknown Injuries	Total Minor Injuries	0.938114
Total Fatal	Total Unknown Injuries	0.936562
Total Government	Total Pedestrians	0.931648

There are many significant relations between different variables; here are the most notable ones:

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

C. 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 if there is a significant relationship between the independent and dependent variables. Here are the results that we found:

TABLE VIII. ONE-WAY-ANOVA RESULTS

Independent Variable	Dependent variable 1	Dependent variable 2	*p-value*
Total Injuries Elements	Drivers impaired	Speeding Involved	**0.03978**
Total Injuries Elements	Total Fatal	Speeding Involved	**0.03974**
Total Accident elements	Drivers Impaired	Pedestrians Impaired	**0.000154**
Total Fatal	Total Taxis	Drivers Impaired	**0.01528**
Total Fatal	Total Government	Drivers Impaired	0.05858
Total Fatal	Total Taxis	Speeding Involved	**0.04698**
Total Fatal	Total Government	Speeding Involved	0.11406
Fatal passenger	Total Taxis	Speeding Involved	**0.03778**
Total Major injuries	Pedestrians Impaired	Speeding Involved	0.28874

The Table VIII shows significant relationships between the independent and dependent variables where the p-value is less than 0.05 [21]. It can be seen that the total injuries and elements have a significant relationship between impaired drivers and speed. Also, total injuries have a significant relationship with total fatalities 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 and drivers impaired and total government and speed.

D. MANOVA

The MANOVA model is similar to the ANOVA. Nevertheless, the difference is examining more than one dependent variable to fit MANOVA, which is in our project's RATE column because it contains five groups [19]. We examine four statistical tests that use in MANOVA, which are Wilk's lambda, Pillai's trace, Hotelling-Lawley trace, and Roy's greatest root. The number of degrees of freedom (DF) is five, and the denominator degrees of freedom (Den DF) is 237183. Where the probability of obtaining an F-ratio is zero in all tests. The Table IX shows the results for the different variables examined.

TABLE IX. FIRST MANOVA ANALYSIS RESULTS FOR WILK'S LAMBDA, PILLAI'S TRACE

Variable	Wilks' lambda		Pillai's trace	
	Value	F Value	Value	F Value
Total Injuries Elements	0.8365	9270.002	0.1635	9270.002
Total Accident Elements	0.9948	247.9693	0.0052	247.9693
Total Injuries Elements	0.8365	9270.002	0.1635	9270.002
Total Minor Injuries	0.9884	554.5788	0.0116	554.5788
Total Fatal	0.9992	37.7784	0.0008	37.7784

TABLE X. FIRST MANOVA ANALYSIS RESULTS FOR HOTELLING-LAWLEY TRACE, AND ROY'S GREATEST ROOT

Variable	Hotelling-Lawley trace		Roy's greatest root	
	Value	F Value	Value	F Value
Total Injuries Elements	0.1954	9270.002	0.1954	9270.002
Total Accident Elements	0.0052	247.9693	0.0052	247.9693
Total Injuries Elements	0.1954	9270.002	0.1954	9270.002
Total Minor Injuries	0.0117	554.5788	0.0117	554.5788
Total Fatal	0.0008	37.7784	0.0008	37.7784

- The above Table X shows Rate column with:
  - TOTAL_INJURIES_ELEMENTS
  - TOTAL_ACCIDENTELEMENTS
  - TOTAL_MAJORINJURIES
  - TOTAL_MINORINJURIES
  - TOTAL_FATAL

TABLE XI. SECOND MANOVA ANALYSIS RESULTS FOR WILK'S LAMBDA, PILLAI'S TRACE

Variable	Wilks' lambda		Pillai's trace	
	Value	F Value	Value	F Value
Drivers impaired	0.9999	7.3326	0.0001	7.3326
Speeding Involved	0.9989	64.4123	0.0011	64.4123
Total Fatal	0.9993	40.5339	0.0007	40.5339
Total Taxis	0.9993	5833.992	0.0896	5838.244
Total Government	0.9058	6163.467	0.0942	6166.538

TABLE XII. SECOND MANOVA ANALYSIS RESULTS FOR HOTELLING-LAWLEY TRACE, AND ROY'S GREATEST ROOT

Variable	Hotelling-Lawley trace		Roy's greatest root	
	Value	F Value	Value	F Value
Drivers impaired	0.0001	7.3326	0.0001	7.3326
Speeding Involved	0.0011	64.4123	0.0011	64.4123
Total Fatal	0.0007	40.5339	0.0007	40.5233
Total Taxis	0.0983	5830.1229	0.0977	5790.4932
Total Government	0.1039	6160.6854	0.1034	6133.7925

- The second Table XI and Table XII show the results of MANOVA of the Rate column with:
  - DRIVERS IMPAIRED
  - SPEEDING INVOLVED
  - TOTAL FATAL
  - TOTAL TAXIS
  - TOTAL GOVERNMENT

E. 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 [14]. In this project, we use a decision tree to find the factors that lead to fatal accidents, which could help us understand the causes that might lead to deadly accidents.

The decision tree in Fig. 2 shows that speed is a significant cause of deadly accidents. This chart shows four-level depth, which gives a scenario if the accident has a speeding case, the total injuries and accident elements are less than 6.5, and there are government cars involved. The chance of an accident having a fatal case is high. The model accuracy is 99.8% which is high accuracy.

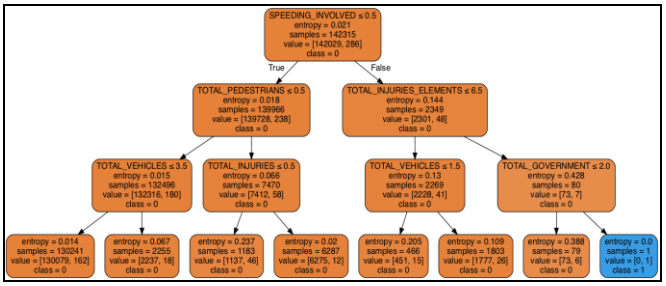

Fig. 2. Decision tree.

F. Data Visualization

Data visualization is a critical tool for evaluating and learning from enormous datasets. They are especially helpful in spotting patterns, trends, and linkages that may not be visible from raw data. In the context of data analysis, visualizations can provide a more natural and accessible approach for stakeholders to communicate complicated information.

The visualizations in this scenario were created with Tableau software [16]. This section's four charts give a detailed overview of the data distribution, allowing the viewer to discover trends and other significant insights immediately.

Fig. 3 depicts the pattern of automobile accidents over time, giving a clear picture of how the frequency of accidents has evolved over time. Fig. 4 shows crashes across time but removes the lowest and lowest, allowing the viewer to focus on the most relevant trends. Fig. 5 is a bar chart displaying the top 20 streets with the most accidents, offering a more thorough look at the data and aiding in the identification of places that may require more attention. Finally, Fig. 6 depicts a map of the District of Columbia with the top ten streets and all accidents depicted as dots, allowing for a visual depiction of the geographical distribution of accidents.

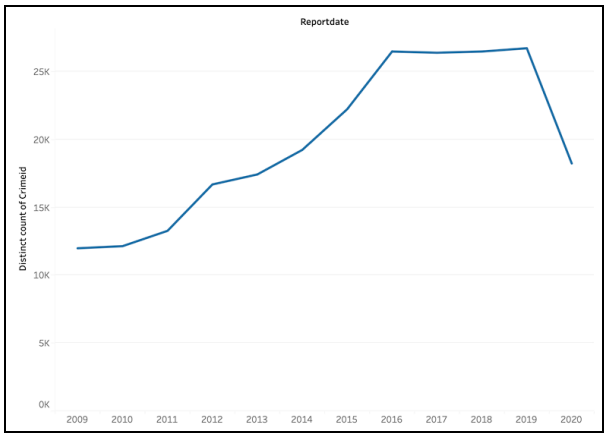

Fig. 3. Total number of accidents over time (2009-2020).

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 that, the numbers changed slightly from 2016 until 2019. However, the accident number hit its peak in 2019 with 26,711 accidents. In contrast, car crashes dropped significantly in 2020, with 18,230 accidents. The reason for that drop is the Covid-19 lockdown.

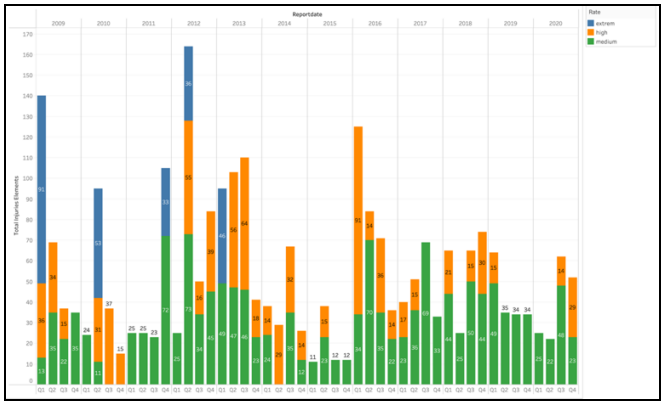

Fig. 4. Number of extreme, high, and medium accidents.

The above 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. The number of accidents reached its highest point in 2012, with approximately 163 accidents resulting in 144 injuries and damage to vehicles in the second quarter. Furthermore, the first quarter of 2009 saw the highest number of severe accidents, with a total of 90. This chart provides insight into the distribution of accidents over time and helps to understand the correlation between the type of accident and the year it occurred.

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.

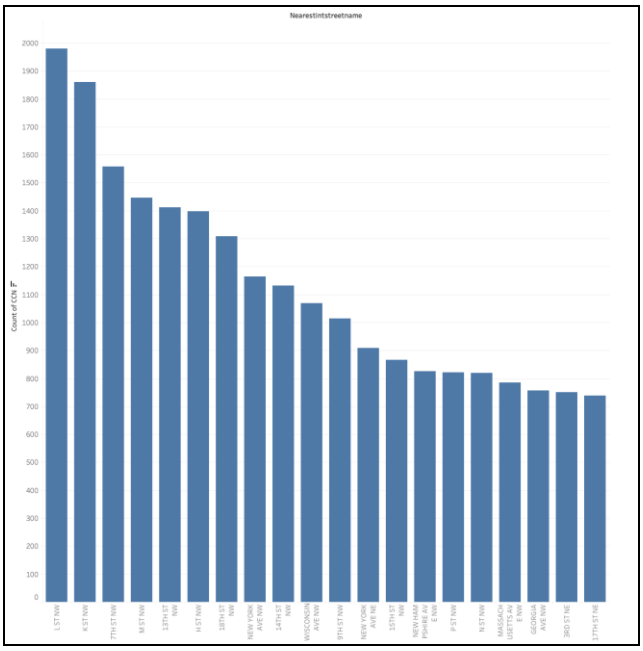

Fig. 5. Top 20 streets with the number of accidents.

The map (Fig. 5) belowdepicts the top ten streets in the District of Columbia where accidents happened between 2009 and 2020. By mapping the geographic distribution of incidents, this visualization can assist in identifying patterns and trends that may not be immediately obvious from raw data alone. Closer examination reveals that most incidents happened in the city center and on routes going out of town from the northeast. This shows that certain issues, such as heavy traffic, bad road conditions, or insufficient signs, may contribute to the high incidence of accidents in these regions.

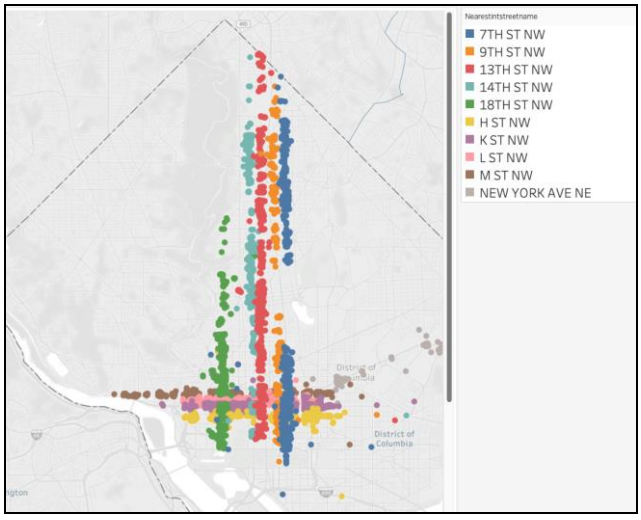

Fig. 6. Map of the top 10 streets that contained accidents.

It is also worth mentioning that while there were some incidents on the city's northeast and southwest sides, the number of accidents on the northwest side was substantially lower. This might imply that characteristics peculiar to the northwest side contribute to the high incidence of accidents in that area. Overall, this visualization might be useful for

identifying the streets with the most accidents and researching the causes to design remedies. Policymakers and city planners may strive to improve road conditions, change traffic patterns, and take other steps to minimize the frequency of accidents and enhance overall safety for cars, cyclists, and pedestrians alike by identifying high-accident zones.

VI. CONCLUSION

In conclusion, this paper studies the dataset related to car accidents in Washington, D.C., between 2009 and 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 that, 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, and extreme) and creating a separate dataset that contains the categorical data mapped the string values into numbers. Then we 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 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 deadly accidents happen if one of the accident cars is 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 accidents are located in the northern 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 found that there was no information about impaired people before 2015. This issue needs to be found the proper way to solve. If these problems are fixed, we think the results could be better.

REFERENCES

[1] W. H. Organization, "Road traffic injuries," World Health Organization, 21 June 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. [Accessed March 2022].

[2] N. H. T. S. A. o. t. U. S. D. o. Transportation, "NHTSA," October 2021. [Online]. Available: https://www.nhtsa.gov/press-releases/usdot-releases-new-data-showing-road-fatalities-spiked-first-half-2021. [Accessed March 2022].

[3] T. S. H. Heinrich and Russell, LLP, "Amarillo Car Accident Lawyers | 200+ Years of Combined Experience," 9 December 2022. [Online]. Available: https://www.templetonsmithee.com/personal-injury/car-accidents/. [Accessed January 2022].

[4] D. D. o. Transportation, "Crashes in DC [These data represent the crash locations associated along the DDOT centerline network within the District of Columbia," 2021. [Online]. Available: https://opendata.dc.gov/datasets/DCGIS::crashes-in-dc/about. [Accessed January 2022].

[5] D. Commons, "Washington, D.C. Demographics - Place Explorer - Data Commons," [Online]. Available: https://datacommons.org/place/geoId/11001/?category=Demographics. [Accessed 18 November 2022].

[6] A. S. Distribution, "Anaconda Documentation," Anaconda Inc, 2022. [Online]. Available: https://docs.anaconda.com/. [Accessed January 2022].

[7] T. p. d. team, "pandas-dev/pandas," Pandas (3.8.8) [Python library]. Zenodo, 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3509134. [Accessed January 2022].

[8] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357-362, 2022.

[9] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa and A. Mueller, "Scikit- learn," *GetMobile: Mobile Computing and Communications*, vol. 19, no. 1, pp. 29-33, 2015.

[10] P. Virtanen, R. Gommers, T. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern and Lars, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261-272, 2020.

[11] J. VanderPlas, B. Granger, J. Heer, D. Moritz, K. Wongsuphasawat, A. Satyanarayan, E. Lees, I. Timofeev, B. Welsh and S. Sievert, "Altair: Interactive Statistical Visualizations for Python," *Journal of Open Source Software*, vol. 3, no. 32, p. 1057, 2018.

[12] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science Engineering*, vol. 9, no. 3, pp. 90-95, 2007.

[13] M. Waskom, O. Botvinnik, D. O'Kane, P. Hobson, S. Lukauskas, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, H. Warmenhoven, J. Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant and Martin, "mwaskom/seaborn: v0.8.1," 3 September 2017. [Online]. Available: https://zenodo.org/record/883859#.ZGLQtC8RqDU. [Accessed March 2022].

[14] Y. Wang, Z. Sun, H. Zhang, W. Cui, K. Xu, X. Ma and D. Zhang, "Datashot: Automatic generation of fact sheets from tabular data," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 895-905, 2019.

[15] "Visual Studio Code (1.61.0)," Microsoft, 2019. [Online]. Available: https://code.visualstudio.com. [Accessed November 2021].

[16] C. Chabot, A. Beers and P. Hanrahan, "Tableau (2021.3.3) [Computer software]," Tableau, 2021. [Online]. Available: https://www.tableau.com. [Accessed March 2022].

[17] A. Jain, "Pandas In Python | Data Manipulation With Pandas," Analytics Vidhya, 26 June 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/01/12-pandas-techniques-python-data- manipulation. [Accessed 30 November 2021].

[18] M. Wood, "Bootstrapped confidence intervals as an approach to statistical inference," *Organizational Research Methods*, vol. 8, no. 4, pp. 454-470, 2005.

[19] Y. Xia, J. Sun and D. G. Chen, "Statistical analysis of microbiome data with R," *Singapore: Springer*, vol. 847, 2018.

[20] Q. Zhang, H. Abel, A. Wells, P. Lenzini, F. Gomez, M. A. Province and I. B. Borecki, "Selection of models for the analysis of risk-factor trees: leveraging biological knowledge to mine large sets of risk factors with application to microbiome data," *Bioinformatics*, vol. 31, no. 10, pp. 1607-1613, 2018.

[21] S. Mcleod, "P-Value And Statistical Significance: What It Is & Why It Matters," SimplyPsychology, 15 May 2023. [Online]. Available: https://www.simplypsychology.org/p-value.html#:~:text=A%20p%20%2Dvalue%20less%20than,and%20accept%20the%20alternative%20hypothesis..