Crash Data from Town of Cary, North Carolina

Uyen Huynh - G01287682 - AIT 580

Abstract

Traffic accidents are causing a substantial amount of unnecessary deaths and injuries every year in the United States. Having a better understanding about the leading causes that contribute to vehicle crashes on road is a very important step towards reducing the occurrence of these unfortunate events. In this paper, we investigate the most common reasons that lead to traffic accidents in the Town of Cary (North Carolina) in particular, and then generalize our findings to the entire US country in general. Through our analysis, we effectively show the following interesting observations: (1) COVID-19 pandemic substantially reduced traffic congestion as well as vehicle crashes on road during 2020 compared to prior years; (2) the most common reasons that have led to crashes is due to drivers' carelessness and ignorance of their own vehicle conditions; (3) most traffic accidents occur in the afternoon during rush hour, and very few accidents occur at night but most of them are fatal; (4) Taxicab has the highest crash percentage during bad weather condition, which is possibly due to tax drivers are often in a rush; (5) vehicle crashes often happen in dark lighting condition on two-way roads without medians separating the two traveling sides.

1. Introduction

Motorized vehicles have become the main transportation means of people in the United States. As more and more people are using their own cars to commute or truck to transport goods everyday, the number of traffic accidents is also increasing considerably. In order to mitigate this problem, it is absolutely necessary to investigate the main contributing factors that lead to vehicle crashes and formulate appropriate research questions to these factors. One way to perform this investigation is to look into available data sources on the internet that are pertinent to vehicle crashes on road. With insights drawn from these data, it is possible to come up with answers and solutions that can help avoid these unfortunate traffic deaths and injuries. The dataset that we select for this study is the **Crash Data from Town of Cary, North Carolina** (Town of Cary Open Data Portal), which provides statistical information about vehicle crashes in the Town of Cary, North Carolina in the years from 2016 to 2021.

There are multiple underlying factors to traffic accidents. In this paper, we will specifically target the following five research questions. First, we will investigate the most common factors that led to traffic accidents in the Town of Cary of North Carolina, and generalize that finding to the US as a whole. Second, we will examine the time that these accidents occurred in order to

find out whether there was a particular time during a day that the majority of traffic accidents happened. Third, we will inspect the location information of these crashes to determine whether there exists some particular locations that had a high frequency of accidents. Fourth, we look into what kind of vehicles that were often involved in vehicle crashes. Finally, we analyze the correlation between environmental factors such as light condition and road configuration. Answering these research questions will allow us to better understand vehicle crashes in terms of the time, location, and environmental factors that are often overlooked by most people. Readers will benefit from this research by becoming more aware of these critical factors and turning into better drivers on roads.

2. Literature Review

Because there have not been prior works that work on the same dataset **Crash Data from Town of Cary, North Carolina**, we explore three other reports that also investigate vehicle crashes but on different datasets. This gives us an overview about how existing relevant works are investigating this problem.

First, we look at the article *Association between Crash Attributes and Drivers' Crash Involvement: A Study Based on Police-Reported Crash Data* (Li, Lai, & Qu, 2020). In this article, the authors present multiple factors that could have an influence on drivers' actions, including environmental elements and drivers' behaviors and characteristics. Even though there are multiple factors, the focus of the article is placed upon the driver's age which is found to be very important towards the actions of drivers in vehicle crashes. For example, the authors of the article show that young and middle-aged people have a lower rate of crashing in rainy weather than in sunny weather, which is not the case for old drivers. This shows that younger drivers tend to have better driving actions in bad weather scenarios, possibly due to their better vision and reflex capability. Based on the results of this work, we will also focus on the age attribute of drivers when conducting our analysis on the **Town of Cary** crash dataset.

Second, we explore the work *Inattention and Distraction in Fatal Road Crashes - Results from In-depth Crash Investigations in* Norway (Sundfør, Sagberg, & Høye, 2019). The focus of this article is to analyze different types of inattention which can lead to traffic crash deaths and identify the ones that cause the most crashes. The results of the article show that the three types of distraction contributing to the majority of fatal crashes are: (1) drivers unable to detect pedestrians in time, (2) drivers fail to check their blind spots, (3) drivers use mobile phones while driving. Based on the results, the authors provide various solutions to prevent fatal crashes, for example, drivers should enable "driving mode" on their mobile phone while driving, or the government should put warning signs at dangerous spots. Similarly, in our main work on the **Town of Cary** crash dataset, we will also explore how crucial is inattention in vehicle crashes.

Third, we analyze the work *Injuries to 15-19-year olds in Road Traffic Crashes: a Cross Sectional Analysis of Police* Crash Data (Thomas & Jones, 2014). The crash dataset used in this work contains information about accident records in England, Scotland and Wales from 2008 to 2010. The work focuses on cases where individuals aged from 15 to 19 were involved in traffic accidents, since this age group particularly has a higher mortality rate than the others. Some interesting findings show that young drivers tend to be involved in traffic crashes when they carry a passenger who is about the same age as them, or young and new drivers are at a higher risk if they drive from 9pm to 5.59am. The authors also propose a possible solution to reduce the rate of crashing for young people, by introducing a special step-by-step learning-to-drive curriculum so that young people can gradually gain experience before having full driving privileges. This article successfully shows that time is indeed one of the main factors that leads to vehicle crashes, which is aligned with one of our three research questions. Different from this work, we will not limit ourselves to any particular age group but will also explore the others.

3. Methodology

- **Dataset description**

The dataset consists of 23892 vehicle crashes with detailed information including the time, location, types of vehicles, weather condition, light condition, injuries/fatalities, etc. of each crash. The dataset is represented as a table comprising 23892 rows and 47 columns where each row denotes one instance of a car crash incident and each column depicts one particular information about the crash.

- **Dataset preprocessing**

Before conducting our investigation into the dataset, we first explore and preprocess the dataset. We first found that there exists some columns with duplicate information. For example, the columns *Vehicle1, Vehicle2, Vehicle3, Vehicle4, Vehicle5* all provide the type of vehicle that was involved in a traffic collision, while column *Vehicle Type* also provides that same information by concatenating the content in the columns *Vehicle1-5*. More specifically, if there are only two vehicles in the crash, the first two columns *Vehicle1* and *Vehicle2* are used to describe them while the remaining columns *Vehicle3-5* are filled with **NA** values. Because *Vehicle Type* contains the same information as in *Vehicle1-5*, it is redundant so we remove them.

Second, one problem in the dataset is that there are some categorical attributes with a large number of possible values which makes it difficult to visualize along other variables. For instance, the column *Vehicle1* has up to 20 distinct values while the column *Weather* consists of 8 distinct values. This means that if we want to group *Vehicle1* and *Weather* together to display the correlation between them, the number of possible groups becomes extremely large from which it is difficult to extract meaningful insights. Therefore, we will transform and regroup similar values of these two columns *Vehicle* and *Weather* into the same category. Figure 1 shows

20 unique values of column *Vehicle1* in the original dataset before and after being transformed, and similarly figure 2 describes a similar transformation process for column *Weather*.

```
PASSENGER CAR
SPORT UTILITY
PICKUP
VAN
UNKNOWN
LIGHT TRUCK (MINI-VAN- PANEL)
SINGLE UNIT TRUCK (2-AXLE- 6-TIRE)
POLICE
FIRETRUCK
TRUCK/TRAILER
MOPED
OTHER *
EMS VEHICLE- AMBULANCE- RESCUE SQUAD
MOTOR HOME/RECREATIONAL VEHICLE
SINGLE UNIT TRUCK (3 OR MORE AXLES)
UNKNOWN HEAVY TRUCK
SCHOOL BUS
ACTIVITY BUS
TAXICAB
TRACTOR/SEMI-TRAILER
```

Figure 1: Unique values of column 'Vehicle1' in the original dataset

```
PASSENGER CAR
SPORT UTILITY
PICKUP
VAN
TRUCK
POLICE
MOPED
EMS VEHICLE- AMBULANCE- RESCUE SQUAD
MOTOR HOME/RECREATIONAL VEHICLE
BUS
TAXICAB
TRACTOR/SEMI-TRAILER
FARM EQUIPMENT
ALL TERRAIN VEHICLE (ATV)
AUTOCYCLE
```

Figure 2: Unique values of column 'Vehicle1' after transformation

In addition to the above preprocessing steps, whenever we visualize a particular column, we also remove the **NA** values from that column. Finally, after this dataset preprocessing step, the dataset contains 18 attributes including all four data types Nominal, Ordinal, Interval and Ratio. The number of observations in the dataset still remains at 23892 rows with missing values converted to the string "None" for easy management.

- **Overview of used methods & tools**

The process of data cleaning and data wrangling is mainly executed by the programming language Python. In the next steps, I use Python, R and SQL to perform exploratory data analysis. After that, R is the main language used for data visualization which gives us a clear understanding about important trends, patterns and relationships between variables. Furthermore, I also apply Natural Language Processing (NLP) methods on the text describing the vehicle crash occurrence locations to extract information about potentially *dangerous* places.

4. Results

4.1. Summary Statistics Information

To provide an overview about vehicle accidents in the Town of Cary, North Carolina in recent years from 2016 to 2021, we present the yearly number of crashes in Figure 3 As we can see, during the time between 2016 and 2019, the number of motor vehicle collisions is very similar to each other, ranging from 4731 to 5034 cases per year. However, the 2020 data shows a significant decline in the number of crashes. The number drops nearly 36% from 5034 reported cases in 2019 to 3196 cases in 2020. This is largely due to the impacts of COVID-19 lockdown in 2020 which makes people mainly work from home and restricted from commuting or traveling (MORRIS, 2021).Since then, the traffic congestion in North Carolina in particular and in the US in general is also reduced considerably. From this observation, we can deduce that there exists a positive correlation between traffic volumes and traffic accidents. This makes sense since the larger the number of vehicles commuting on road, the more number of traffic accidents that can possibly happen. For the year 2021, the data has been collected for only the first four months, so the number is far less than the previous years.

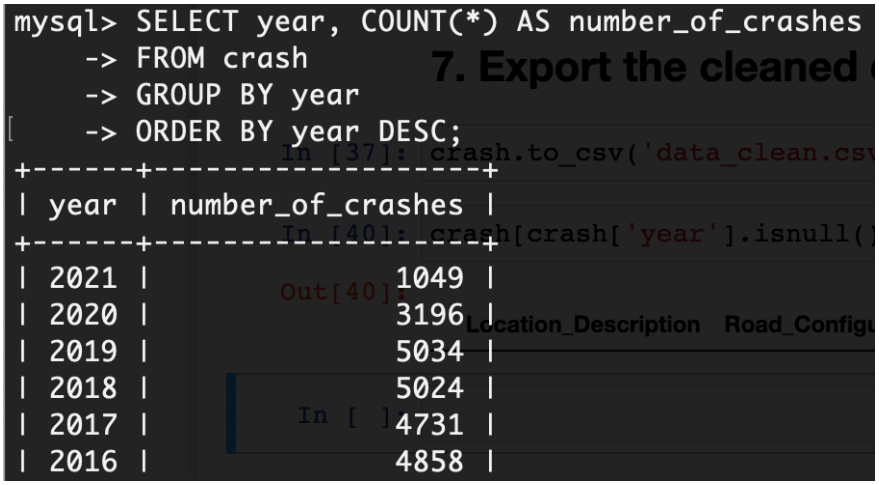

Figure 3: Yearly number of crashed during the years 2016-2020

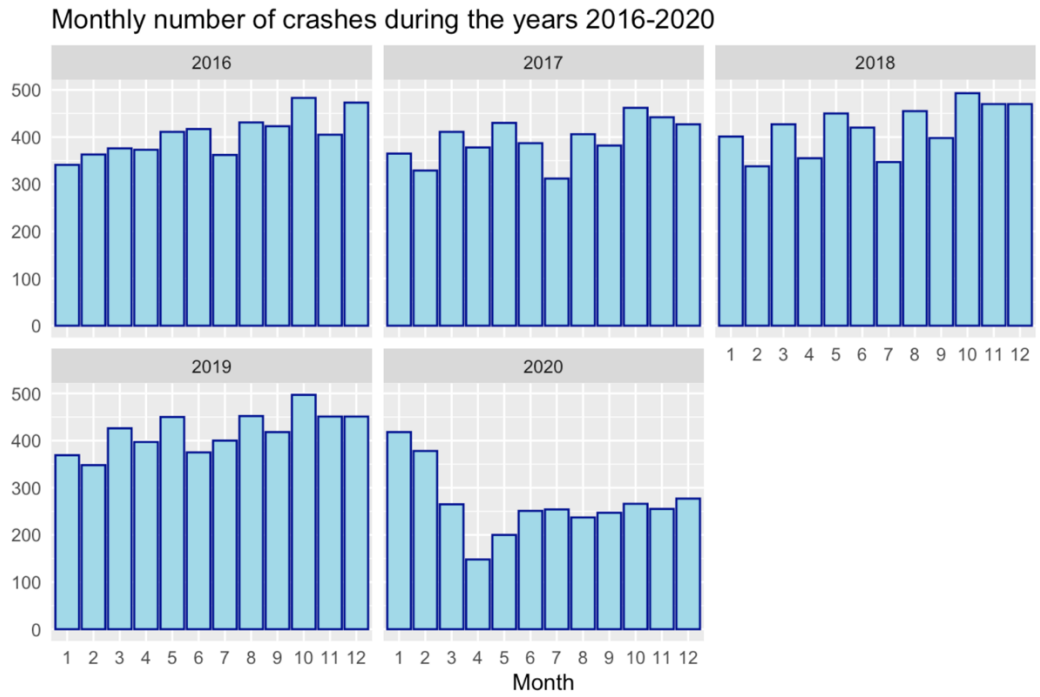

Figure 4: Monthly number of crashes during the years 2016-2020

Next, we will take a deeper look at the monthly number of crashes during the same timespan which is displayed in Figure 4. As we can see, from 2016 to 2019, the number of accidents is quite steady during the year with only the last three months (October to December) being slightly higher than the previous months. In 2020, it can be seen that April was the month having the least number of traffic accidents.

4.2. The most common contributing factors to traffic accidents in Town of Cary, North Carolina

To find the factors responsible for most vehicle crashes that happened in Town of Cary of North Carolina, we used SQL to extract the most common values from the column *Contributing Factor* and present the results in Figure 5. The results show that the top-2 most frequent reasons are *coming from behind parked vehicles* and *darting*. For the first reason, it is mostly due to the carelessness of drivers that fail to check their blind spots, the road condition and give warning signals to other traveling vehicles. Carelessness mostly comes from each person's intrinsic personality and it is very difficult to change one person's intrinsic behavior. One solution to improve drivers' awareness is to put warning banners at some locations on every street to encourage everyone to be more careful while driving. For the second reason, it is mostly due to car deficiency. One way to mitigate this is also to encourage/require drivers to check their car or bring their car for maintenance more often.

```
mysql> SELECT Contributing_Factor, COUNT(*) AS number_of_crashes
    -> FROM crash
    -> WHERE Contributing_Factor != ''
    -> GROUP BY Contributing_Factor
    -> ORDER BY number_of_crashes DESC
    -> LIMIT 6;
+-----------------------------------+--------------------+
| Contributing_Factor               | number_of_crashes  |
+-----------------------------------+--------------------+
| COMING FROM BEHIND PARKED VEHICLE |                210 |
| DARTING                           |                116 |
| FAILURE TO YIELD RIGHT OF WAY     |                 88 |
| LYING AND/OR ILLEGALLY IN ROADWAY |                 18 |
| NOT VISIBLE (DARK CLOTHING,ETC.)  |                 15 |
| INATTENTIVE (TALKING,EATING,ETC.) |                 12 |
+-----------------------------------+--------------------+
```

Figure 5: The most common contributing factors to traffic accidents in Town of Cary, North Carolina

4.3. Is there any specific time in a day when the majority of traffic accidents happen?

To investigate whether there is any specific time in a day when the majority of traffic crashes happen, in Figure 6, we plot the chart showing the number of accidents in each hour of a day during the year from 2016 to 2019 (the bar at hour 15 provides the total number of crashes from 15:00 to 15:59). It can be observed that most accidents tend to occur in the afternoon. In particular, the time from 15:00 to 17:59 has the highest number of vehicle crashes in a day. It makes sense because this is the rush hour traffic when there is a very high number of cars on the roads. In addition, the earlier AMs from 00:00 to 06:00 have the least number of accidents during a day since the traffic volume during this time is very low.

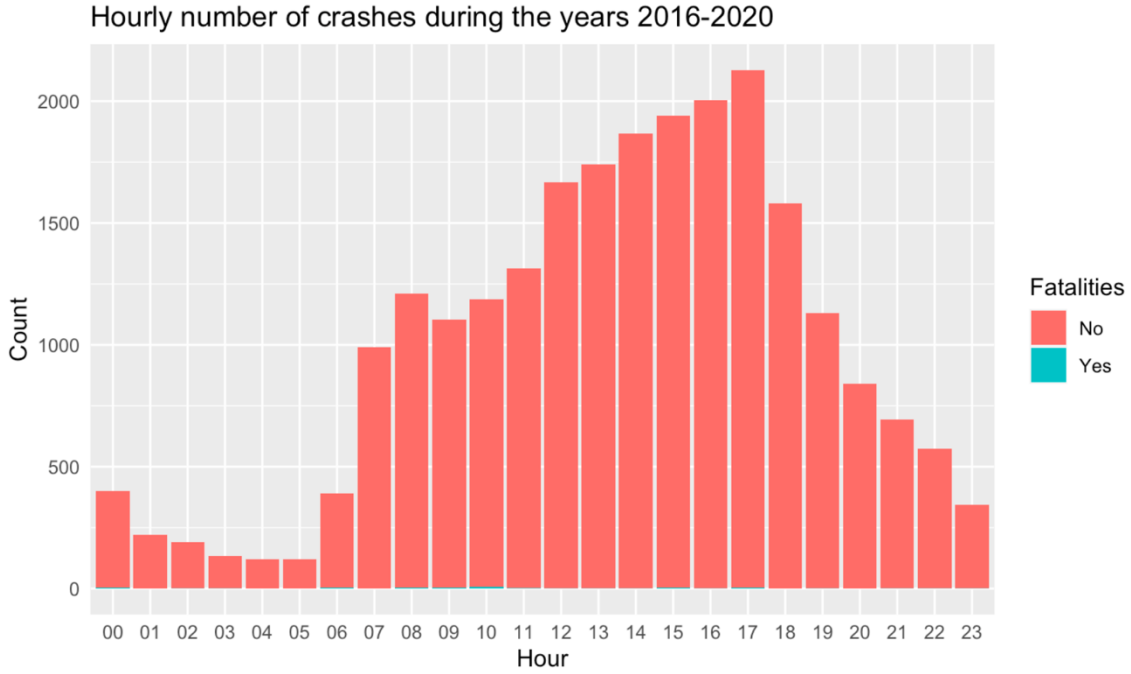

Figure 6: Hourly number of crashes during the years 2016-2020

As mentioned previously, according to the yearly number of crashes, it shows that the correlation between traffic volumes and crash volumes is positive. The same thing can also be applied to the number of crashes counted for each hour. It makes sense because the number of cars running in the early morning is far fewer than that in the working hour or in the early evening. However, the correlation between traffic volumes and crash severity might not be as large as the one between traffic volumes and crash volumes. Intuitively, the more vehicles commuting on road, the more crowded the road is and the less severe traffic accidents are since vehicles are not traveling at high speed. To answer this question, we calculated the fatality rate per hour to explore the time when most accidents are most severe and plot it in Figure 7. The rate is defined by the number of fatal crashes divided by the total number of crashes during the hour. Looking at the chart, we can see that the time from 00:00 to 06:00 which has the least number of accidents is the most dangerous time period with the highest fatality crash rate. Moreover, the highest injuries rate, which is calculated by the number of injury crashes divided by the total number of crashes, is in the time between 05:00 and 05:59 (as Figure 8). These might be due to the fact that when there is no traffic congestion, people tend to drive at higher speed. Therefore, the chance of severe accidents happening during nighttime is higher.

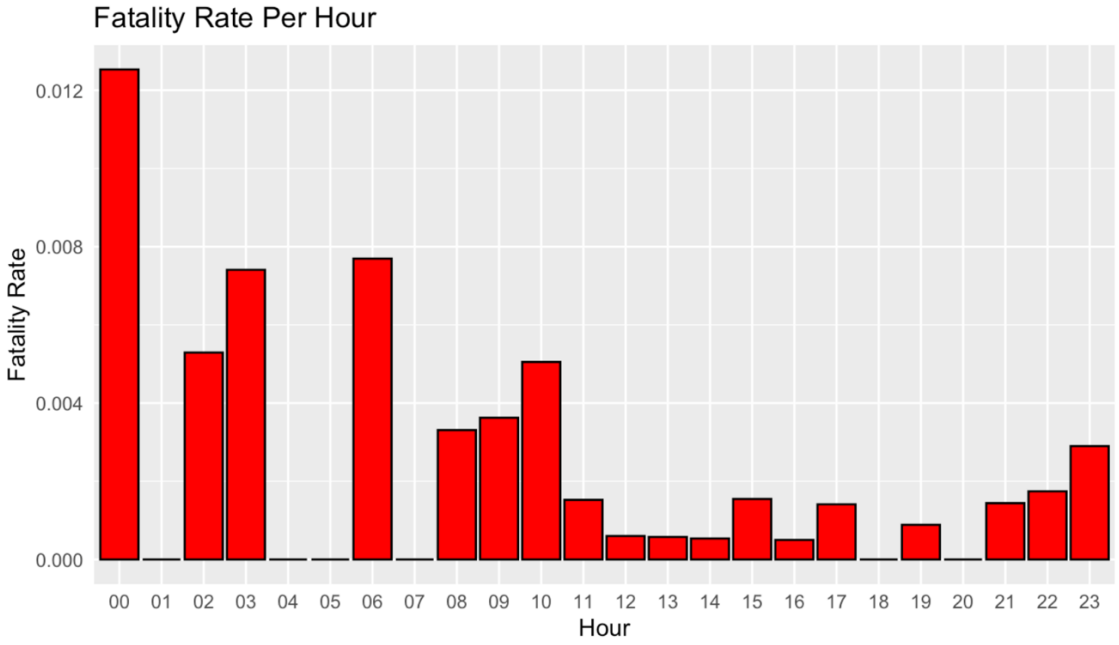

Figure 7: Fatality rate per hour

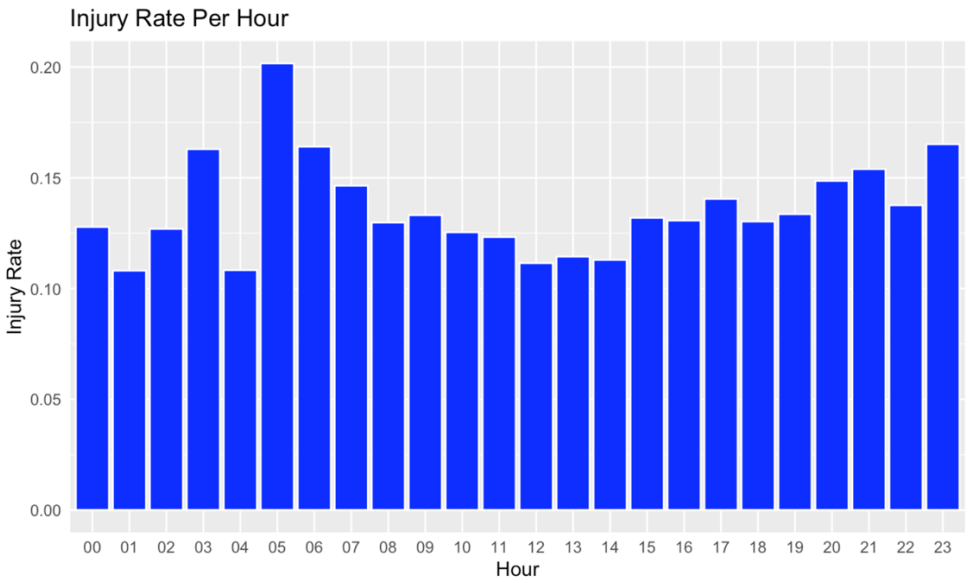

Figure 8: Injury rate per hour

4.4. Are there any particular locations that had a high frequency of accidents?

The dataset has an attribute that provides detailed information about the location where a crash happened. To leverage this information, I used NLP methods to remove stopwords and unnecessary words in the text (street and road abbreviations such as 'rd', 'ave', 'pkwy'... or measurements like 'miles', 'feets'). We then display the result showing the roads with the most vehicle accidents occured. Figure 9 and Figure 10 show that most accidents happened on these following roads and streets: Cary Parkway, Walnut St, Kildaire Farm Rd, SW Maynard Rd, Harrison Av, Tyron Rd,...

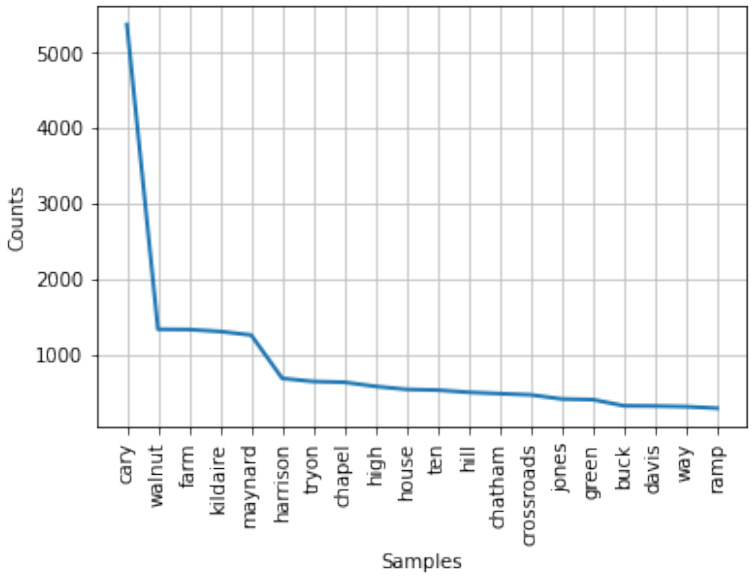

Figure 9: Locations that had the highest frequencies of accidents

```
.1 MILES FROM CARY PARKWAY
.1 MILES FROM P.V.A. (1830 WALNUT ST)
60 FEET FROM SR 1300 (KILDAIRE FARM RD)
70 FEET FROM PVA 760 SW MAYNARD RD
200 FEET FROM SR 1652 (N HARRISON AV)
00 FEET FROM SR 1009 (TYRON RD)
75 FEET FROM SR3081 (CHAPEL HILL RD)
360 FEET FROM SR 1615 (HIGH HOUSE RD)
30 FEET FROM SR 1011 (W CHATHAM ST)
```

Figure 10: Examples about the locations where the most accidents happened

4.5. What kind of vehicles that were often involved in vehicle crashes?

Figure 11 represents the number of crashes for each type of vehicle during the same year period. It appears that passenger cars and sport utility vehicles are the two types that are frequently involved in crashes. To figure out whether there is any pattern between the types of vehicles and the impacts of environments like weather, we plot a chart visualizing the number of accidents of every vehicle type in every weather condition. As we can observe from Figure 12, most accidents happened when the weather was clear. This may be explained by the fact that when the weather is not good, people tend to avoid commuting.

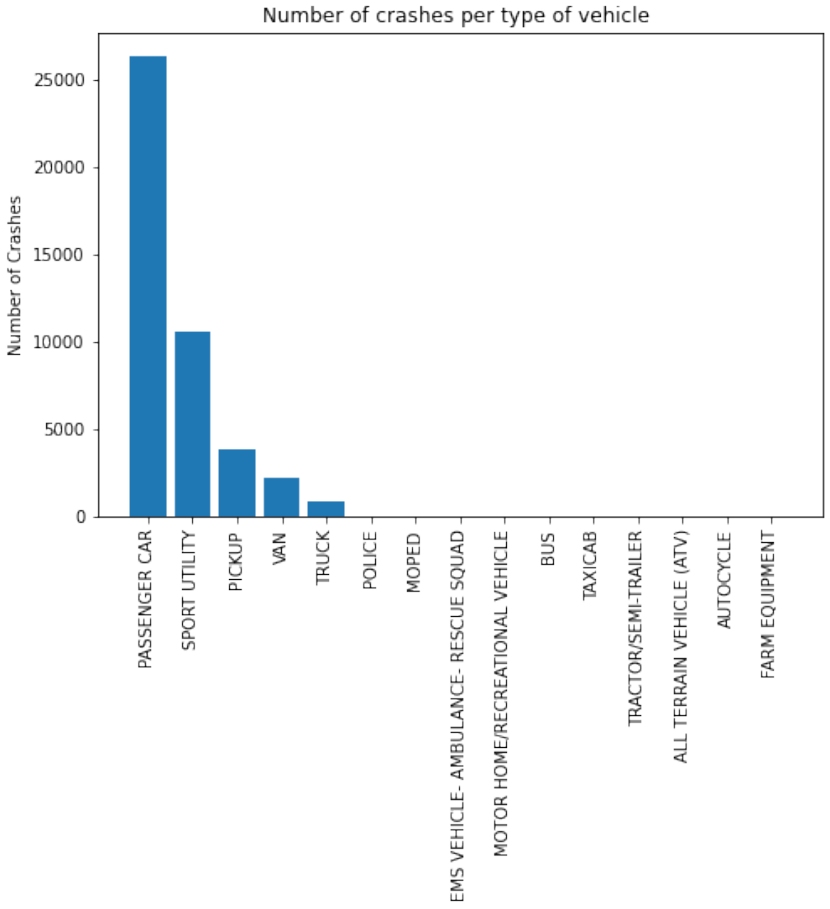

Figure 11: Number of crashes per type of vehicle

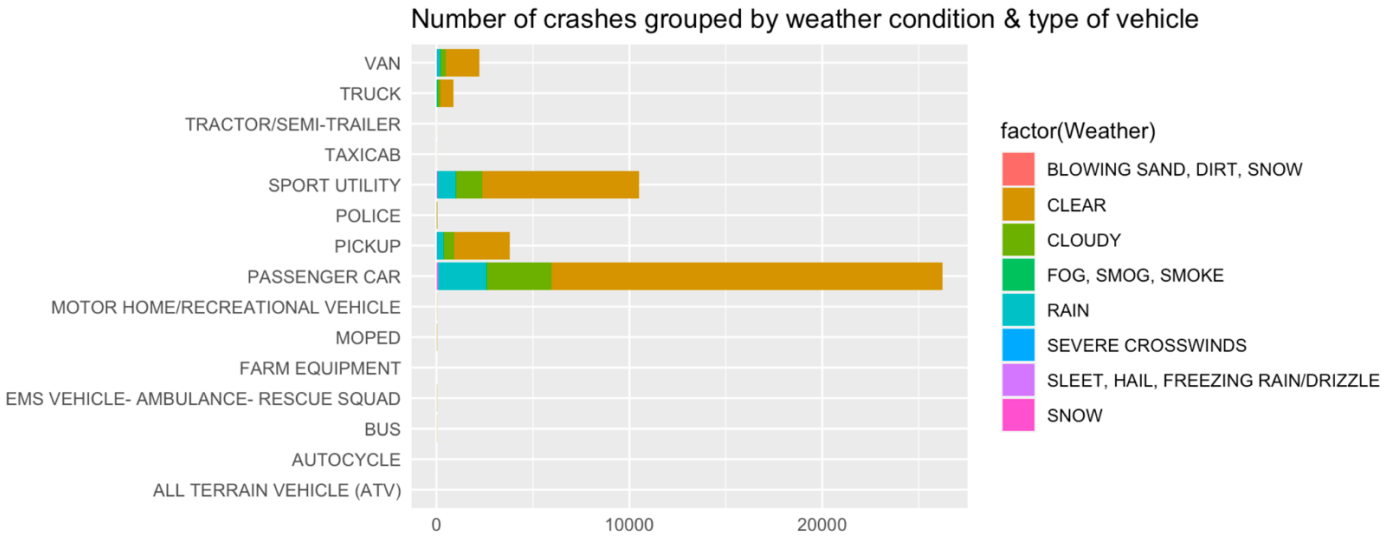

Figure 12: Number of crashes grouped by weather condition & type of vehicle

Next, in order to better understand the relationship between vehicle types and weather conditions, we will apply a simple transformation to convert all weather values into only two categories: 'Good' and 'Bad' condition. The number of possible values for weather conditions in the dataset is too many, which hinders us from extracting meaningful insights. Specifically, we transformed the following weather conditions 'Clear', 'Cloudy', 'Fog, Smog, Smoke', 'Rain', 'Severe Crosswinds', 'Sleet, Hail, Freezing Rain/Drizzle' and 'Snow' into category 'Bad', and rename the 'Clear' condition with a new name 'Good'. After this transformation, the *weather* attribute has only two possible values: 'Bad' and 'Good'. Then, we continued to calculate the crash percentage in 'Good' and 'Bad' weather for each vehicle type. The crash percentage of a vehicle type for the 'Good'/'Bad' weather condition is calculated by the number of crashes happening in the 'Good'/'Bad' weather divided by the total number of crashes of that vehicle type. Looking at Figure 13**,** it can be seen that the taxicab had the highest crash percentage in bad weather compared to the remaining types.

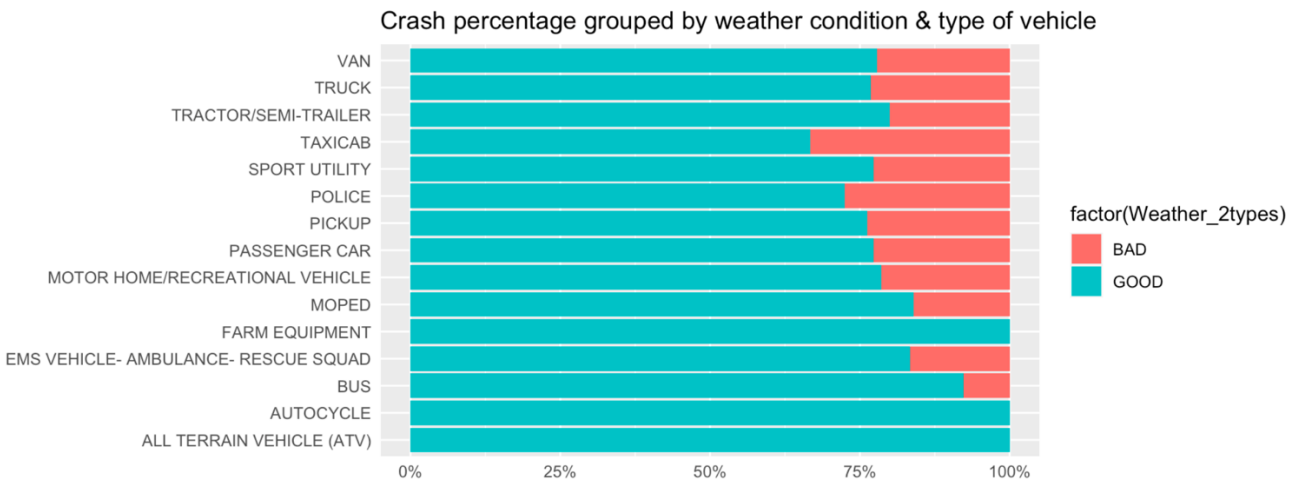

Figure 13: Crash percentage grouped by weather condition & type of vehicle

4.6. Correlation between environment factors (light condition and road confirguation)

Besides weather conditions, other environmental factors such as the configuration of roads and the light condition are also very important. Figure 14 shows the crash percentage for each combination of light condition and road configuration. To give an example, for each type of light condition, we divide the crashes into 4 groups (representing 4 types of road configuration) and calculate the crash percentage for each group. From the result in Figure 14, it can be observed that when the light condition was dark, there were more crashes happening on the 'two-way, not divided' roadways. The 'two-way, not divided' roadway means that there is no median that physically separates the two opposite directions (Massachussetts Law Enforcement Crash Report E-Manual, n.d.). To explore the correlation between the two variables 'Light_Condition' and 'Road_Configuration', we conduct the Chi-squared independence test. The result in Figure 15 shows that the p-value is 2.2e-16, which is significant enough to reject the null hypothesis that light condition is independent of road configuration.

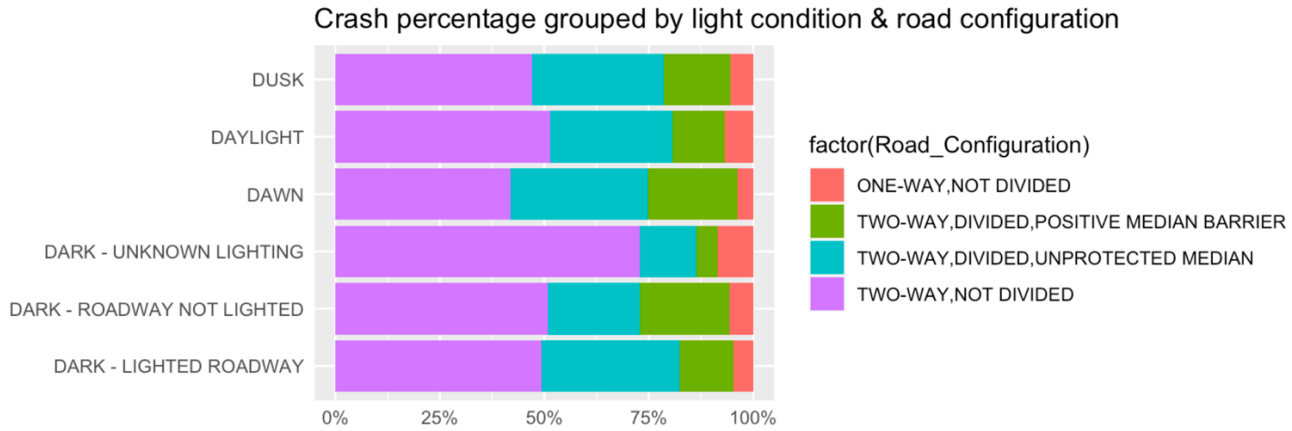

Figure 14: Crash percentage grouped by light condition & road configuration

```
        Pearson's Chi-squared test

data:  light_configuration_chi
X-squared = 123.77, df = 15, p-value < 2.2e-16
```

Figure 15: Chi-squared test result

5. **Limitations and needed future research**

- **Limitation:**

There are two limitations to the dataset that hinder us from exploring further into this problem. The first one is that the dataset does not provide any demographic information about the drivers that are involved in the vehicle crashes. For example, we are unable to conduct investigations about whether age is an important factor because it can determine a driver's behavior. Having said that, the dataset authors have to be careful if they want to provide demographic information due to the recent attention of people into dataset bias. For example, a machine learning prediction model trained on a dataset with racial or other demographic information may be biased against the minorities and lead to very bad consequences. Secondly, the dataset does not provide information about physical damage of these vehicle crashes, which hinders us from inspecting how severe a crash was at the time.

- **Needed future research:**

There is one additional data analysis that we could have done was to visualize the locations of these vehicle crashes onto a map. This could potentially help us to determine if some specific locations are prone to having traffic accidents.

6. **Conclusion**

In this study, we have performed data analysis into the dataset of vehicle crashes in Town of Cary of North Carolina to better understand what factors contribute to traffic accidents in North Carolina in particular and the US as a whole. Our findings show some interesting observations, such as carelessness and ignorance to their own vehicles' condition of drivers are the major reasons that led to most crashes.

References

Li, G., Lai, W., & Qu, X. (2020). Association between Crash Attributes and Drivers' Crash Involvement: A Study Based on Police-Reported Crash Data . *International Journal of Envinronmental Research and Public Health* .

Massachussetts Law Enforcement Crash Report E-Manual. (n.d.). *Trafficway Description*. Retrieved from Massachussetts Law Enforcement Crash Report E-Manual: https://masscrashreportmanual.com/crash/trafficway-description/

MORRIS, D. Z. (2021). *These U.S. cities led the world in reduced car traffic last year*. Retrieved from Fortune: https://fortune.com/2021/01/12/us-car-traffic-2020-covid-19-tomtom/

Sundfør, H. B., Sagberg, F., & Høye, A. (2019). Inattention and distraction in fatal road crashes – Results from in-depth crash investigations in Norway . *Accident Analysis & Prevention* .

Thomas, J., & Jones, S. (2014). Injuries to 15–19-year olds in road traffic crashes: a cross sectional analysis of police crash data . *Journal of Public Health*.

Town of Cary Open Data Portal. (n.d.). *Crash Data.* Retrieved from Town of Cary Open Data Portal: https://data.townofcary.org/explore/dataset/cpd-crash-incidents/export/?disjunctive.rdfeature&disjunctive.rdcharacter&disjunctive.rdclass&disjunctive.rdconfigur&disjunctive.rdsurface&disjunctive.rdcondition&disjunctive.lightcond&disjunctive.weather&disjun