EXPLORATORY DATA ANALYSIS ON MOTOR VEHICLE COLLISIONS IN NEW YORK

Submitted by

Nelson Joseph (1002050500)

Jeeva Manavalan (1002067528)

in partial fulfilment for the award of the degree of

M.S. IN DATA SCIENCE

THE UNIVERSITY OF TEXAS AT ARLINGTON

DASC 5300/CSE 5300 FOUNDATIONS OF COMPUTING

INSTRUCTOR: Prof. Sharma Chakravarthy

22 SEPTEMBER 2022

Honor Code:

Nelson Joseph and Jeeva Manavalan did not give or receive any assistance on this project, and the report submitted is wholly on our own.

Table of Contents

Introduction

The main objective of the project was to analyse the Motor Vehicle Collisions in New York City. The data contains a total of around 3.7 million with 26 column labels and we analysed every piece of data by doing proper pre- processing for all the columns that comes into our analysis. We cleaned the null values as well as corrected the values that got entered incorrectly while pre-processing. We obtained our analysis results in different visualization plots that includes: -

A. Analysis 1 of VEHICLE MAKE vs YEARS for TOYT, CADI, GMC, SUBA using bar Graph.
B. Analysis 2 of VEHICLE MAKE vs MONTHS for TOYT, CADI, GMC, SUBA using Line Graph.
C. Analysis 3 of VEHICLE TYPE vs Frequency of Accidents using Pie Chart.
D. For our team 30, we were asked to analyse the data between the dates 01-oct-2018 to 31-aug-2020.

File Description

File Type	File Name
.docx	**REPORT.docx** -Final analysis report
.xlsx	**Work_flow.xlsx** – The excel file that shows the workflow of project.
CSV file	**Motor_Vehicle_Collisions_-_Vehicles, Cleaned_data.csv, MVC (1).csv** Initial data, Cleaned data and Vehicle specific data respectively.
Jupyter Notebook	**DASC5300_Proj1_Fall22_team__30_ (2). Ipynb –** Analysis Notebook
.ppt and .pdf files	**Presentation-project1-DASC5300_v7.pptx**, **state_prefixes.pdf** - where we checked for all the state registration validity. **DASC5300_Proj1_Fall22_team__30_ - Colaboratory.pdf** – PDF showing the visualization.

Division of Labour

- Initial pre-processing together
- Pre-processing of columns -Nelson
- Analysis 1 - Jeeva
- Analysis 2 and 3 - Nelson
- Report - Together

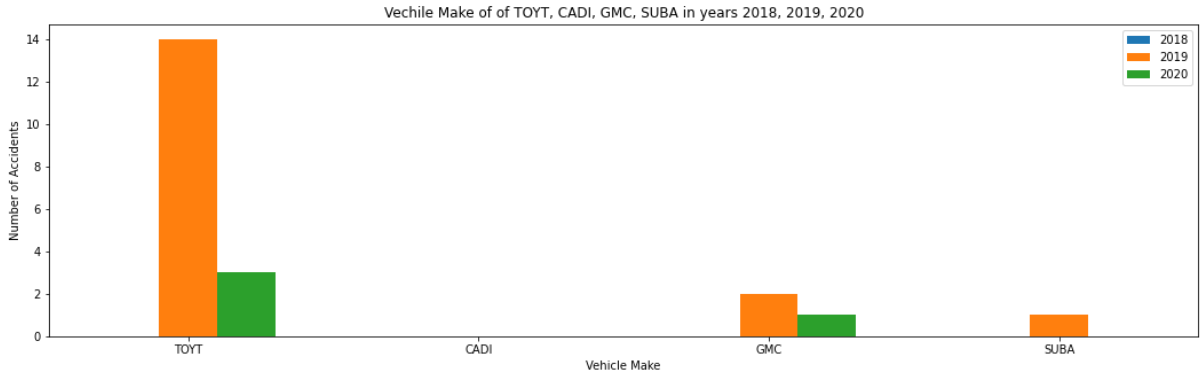
Pre-processing

- After reading the CSV file, we took the data given to our team which is from 01-09-2018 to 31-08-2020 and stored the crash data in a CSV file.
- Crash data between 01-09-2018 to 31-08-2020 consist of 741086 accidents. that is 20 % of the accidents that happened during this period
- There can be chances of empty values that are not **np. NaN," ??"," –"** in the dataset so we have to fill those with **np. NaN** and finally remove it for proper analysis
- We are using VECHILE_MAKE for our analysis of the missing values in the column is useless to us. We cannot replace the data with any random values as it affects the analysis directly. So, we removed the 85791 rows of data in which the **VECHILE_MAKE** column is a null value.
- Pre-processing for **PUBLIC_PROPERTY_DAMAGE** Column
- Pre-processing for **DRIVER_SEX** Column
- Pre-processing for **POINT_OF_IMPACT** Column
- Pre-processing for **STATE_REGISTRATION** Column
  - There is no need of removing the state registration or replacing it since the NYC accident data include the vehicle from the USA and from outside the country. Several vehicles from Canada and nearby countries come into the USA. For example, we found that vehicles with state registration ON (Ontario, Canada) have 261 accidents that occurred in with in NYC.
  - We found that majority of accidents happened by the vehicle within 'NY'- New York and on the second place is 'NJ' - New Jersey' which is a nearby state.
- Pre-processing for **DRIVER_LICENSE_JURISDICTION** Column
- Pre-processing for **DRIVER_LICENSE_STATUS** Column
  - It is identified that most of the vehicle accidents are by people with a driving license which is around 80% of the accidents.
- Pre-processing for VEHICLE_MAKE Column
  - Vehicle make was represented by a different name for the same maker. In order to properly analyse the data, we converted all the similar values to their group.
  - We replaced the errors of the vehicle make of TOYT, CADI, SUBA, and GMC to their original forms for analysis using replace method in pandas.
  - We also be corrected the mistakes of HONDA, NISSAN, FORD and CHEVROLET too as majority of the accidents are happened by these vehicles after TOYOTA.
- Pre-processing for VEHICLE_TYPE Column
  - Several cleaning approaches were done in order to identify the different ways in which a VEHICLE TYPE is represented in the data.
  - All kinds of utility vehicle and Sports utility vehicle were grouped together to SPORT UTILITY VEHICLE
  - There were several representations of UNKNOWN values too. So, we grouped them first and cleaned it.
- Data is cleaned now, and we used the cleaned data for each of the following Analysis.

Analysis 1

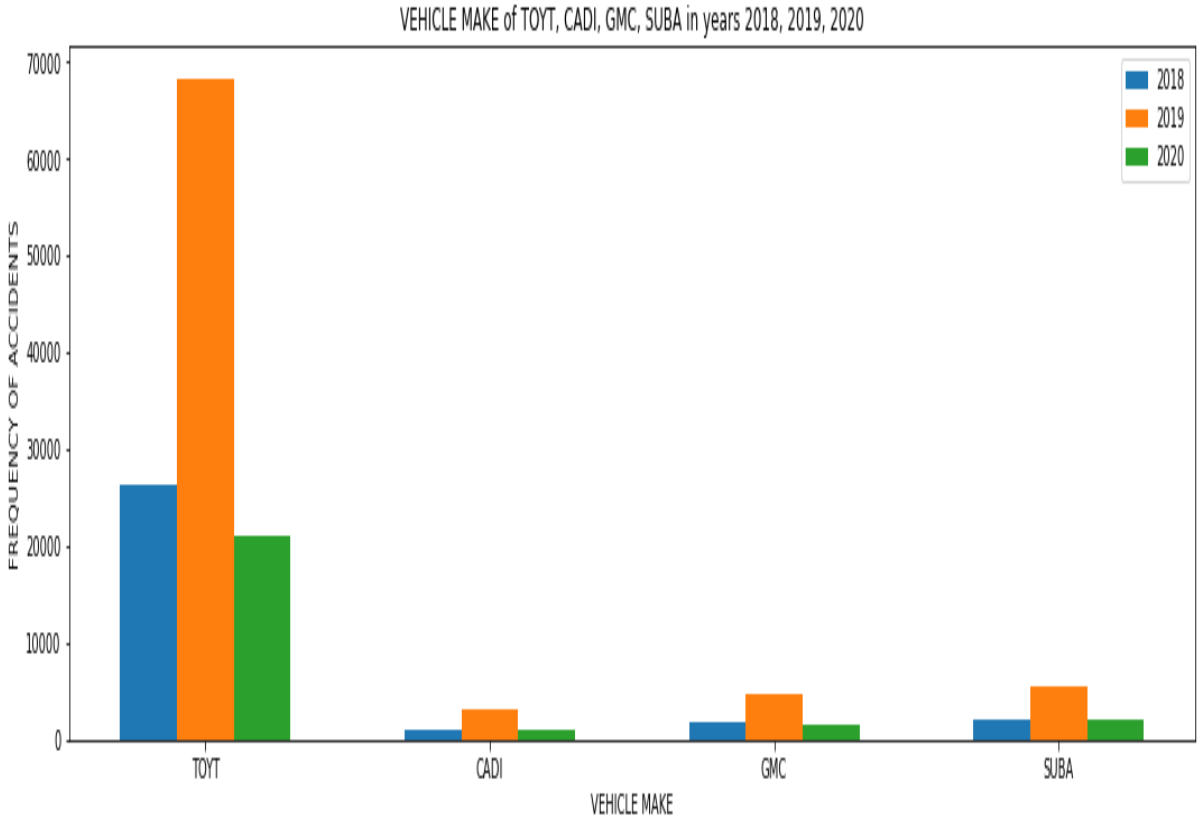
Analysis 1 for Sample Data

We analysed the data first by taking the sample using the .sample method of pandas using one of Data of Birth (07/30/1998) which generated random 100 rows of data.



We verified the obtained results using value_counts () function from the sample data. The results shows that TOYOTA is having the maximum number of accidents for the sample data.

Analysis 1 for Entire Data

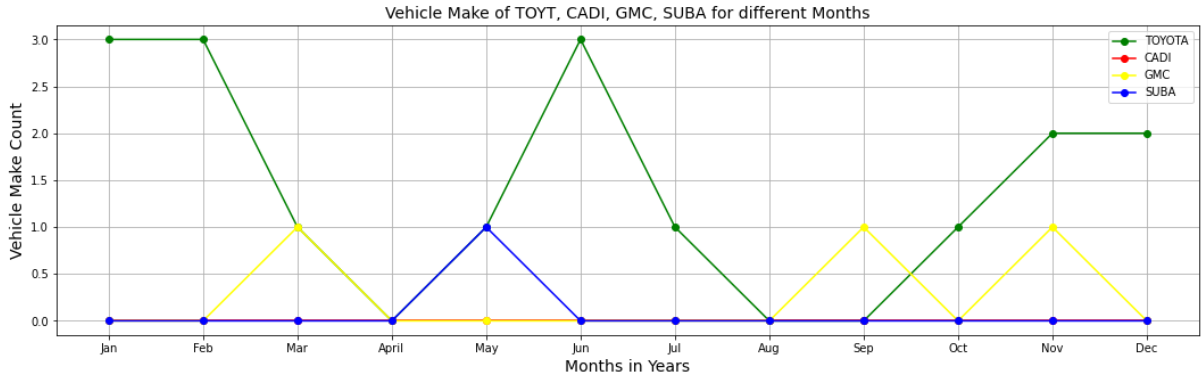


The accident for TOYOTA is maximum in 2019 but we can't conclude that since we only took 4 months of data in 2018 and 8 months for 2020. The conclusion that we can get is that TOYOTA is the vehicle that got into accidents mostly when compared with CADI, GMC, and SUBA with a high margin. The lockdown was executed in 2020 so driving vehicles are mostly used by 1 layer of professionals like doctors, police, and health-related person.

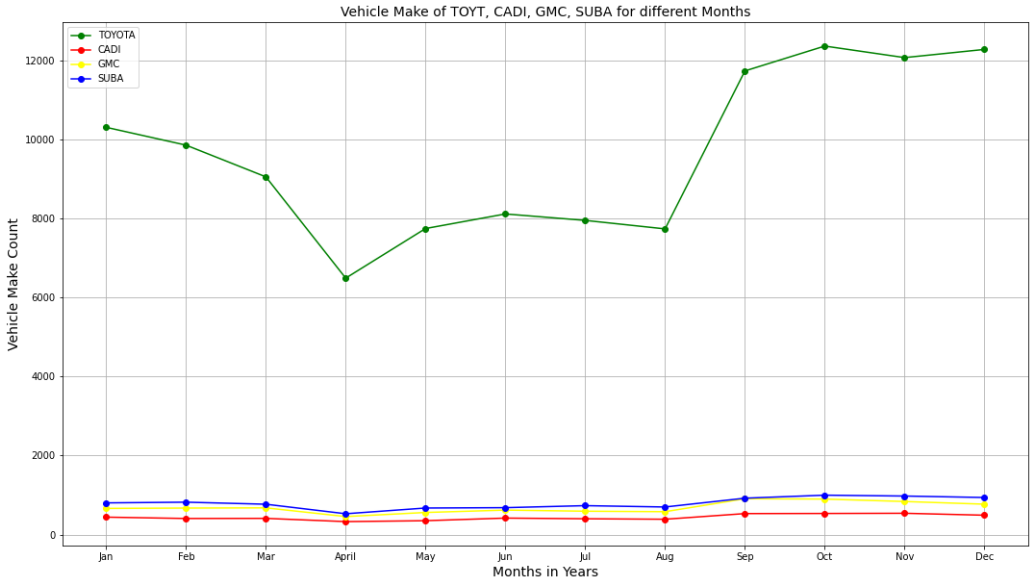
Analysis 2

Analysis 2 for Sample Data

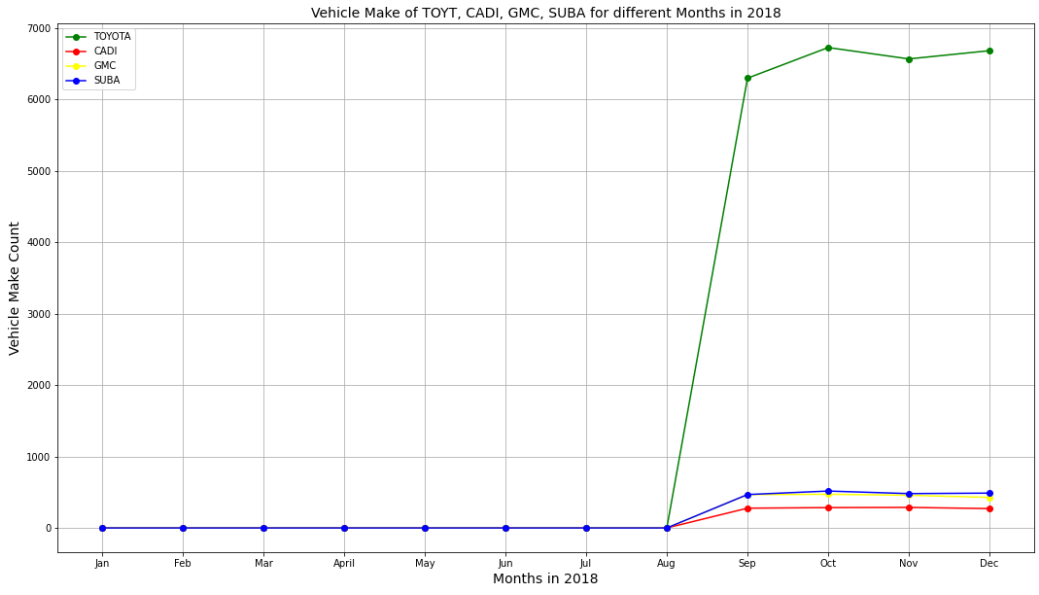
For Analysis 2, We splitted the data into 4 different data frames using the VEHICLE MAKE conditions for TOYT, CADI, GMC, and SUBA and analysed the values using value_counts () function. The input to the line graph for each VEHICLE MAKE was taken by a function that we made that returns frequency of accidents in each month and zeros for months were no accidents occurred. We verified the validity of the graph by checking accidents of CADI which was not present in sample data.



Analysis 2 for Entire Data

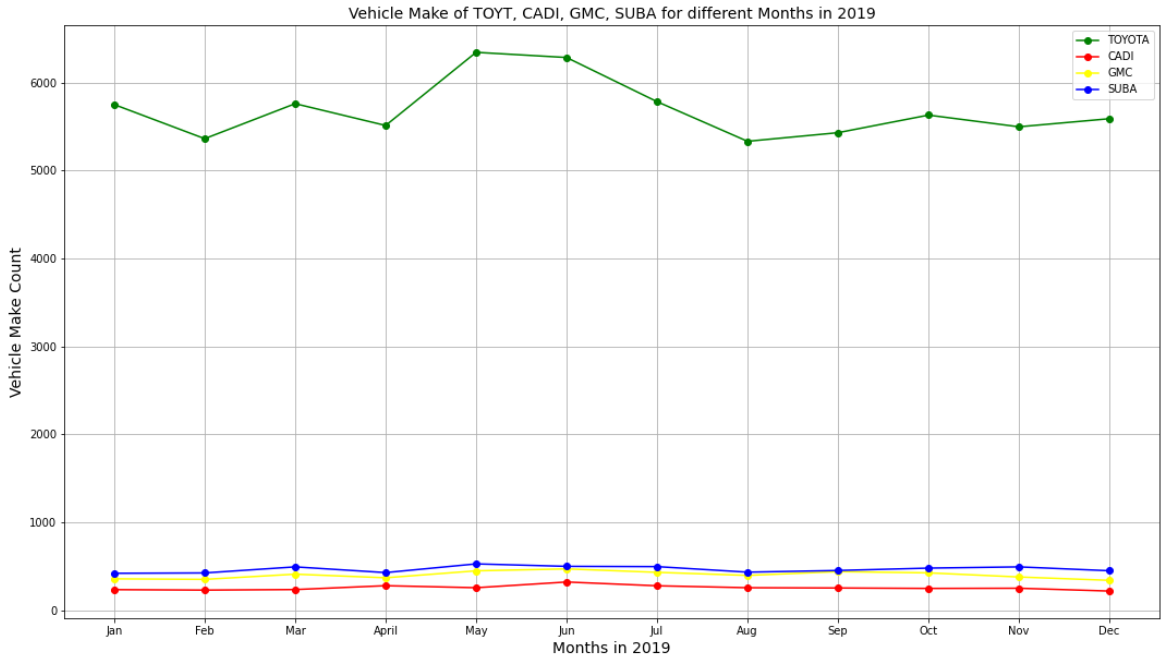


We identified from the line graph that VEHICLE MAKE TOYATA has the maximum no of accidents in this period and other VEHICLE MAKE are almost similar. VEHICLE MAKE CADI is the one which has least accidents. We are not able to properly tell which month has more accidents since the data interval we took contains 4 months from 2018, All the months of 2019, and 8 months of 2020. So, we need to plot year wise monthly analysis



- Toyota is the VEHICLE MAKE that got into accidents mostly in the months of 2018.
- There is increase in accidents in the month of September, October, November, and December since these are the months in which there will harsh climate. The road will be slippery since New York is a place where snowfall usually happens.

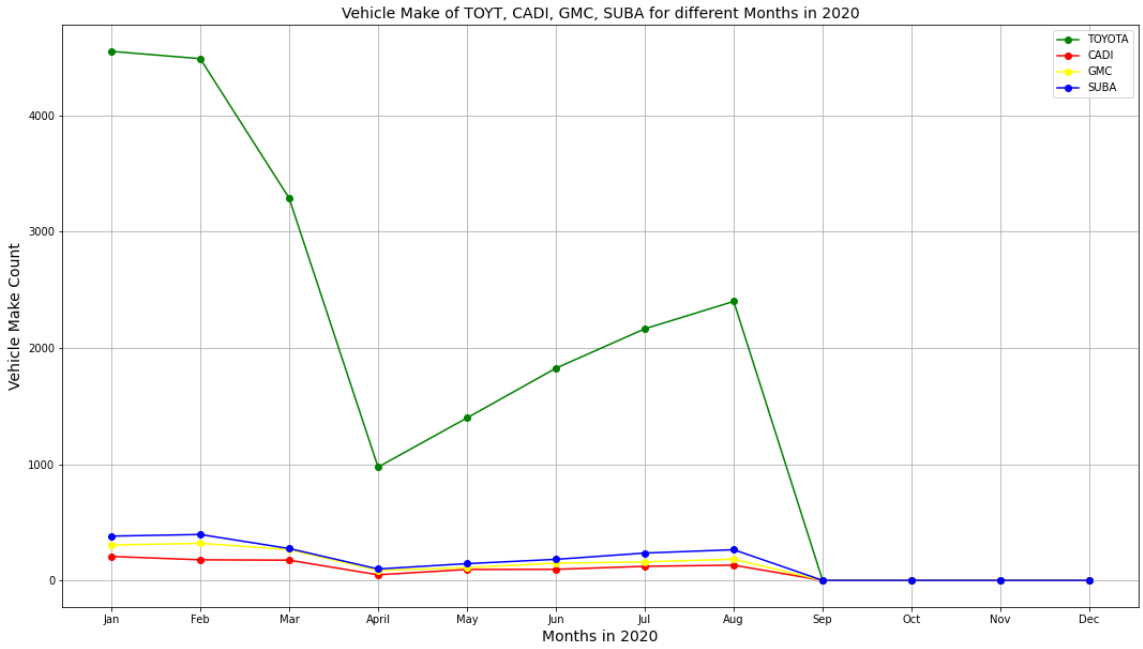
https://weatherspark.com/h/s/23912/2019/2/Historical-Weather-Fall-2019-in-New-York-City-New-York-United-States#Figures-ObservedWeather



- Maximum accidents occurred in the month of May in 2019 for TOYOTA, but for all the months the accident of TOYOTA is very high compared to all other VEHICLE MAKE.
- May is the month with the highest number of reported accidents (20,551), followed closely by June (20,479) and October (20,470).
  https://mirmanlawyers.com/new-york-car-accident-lawyer/statistics/
  The Information is verified from internet so our analysis corrects.

- Toyota is very popular vehicle in USA as it builds solid, efficient, and reliable vehicles as per consumer reports. This can be the main reason for increased no of accidents as the VEHICLE MAKE 'TOYATA' is used by a major population. Thus, our analysis of the data is valid.

  https://www.driversautomart.com/why-is-the-toyota-brand-so-popular-among-consumers/



We identified that in initial months of 2020, we can see that all the VEHICLE MAKE accidents got declined rapidly. This decline in accidents is due to impact of COVID-19 pandemic. We can see from the above graph that, the accidents started declining from January and reached a bottom threshold in April.

https://en.wikipedia.org/wiki/COVID-19_pandemic_in_New_York_City

According to the data from internet we can see that the coronavirus has been spreading in New York city from January.

1. **By March 29, over 30,000 cases were confirmed**
2. **Starting March 16, New York City schools were closed.**
3. **On March 20, the New York State governor's office issued an executive order closing "non-essential" business.**

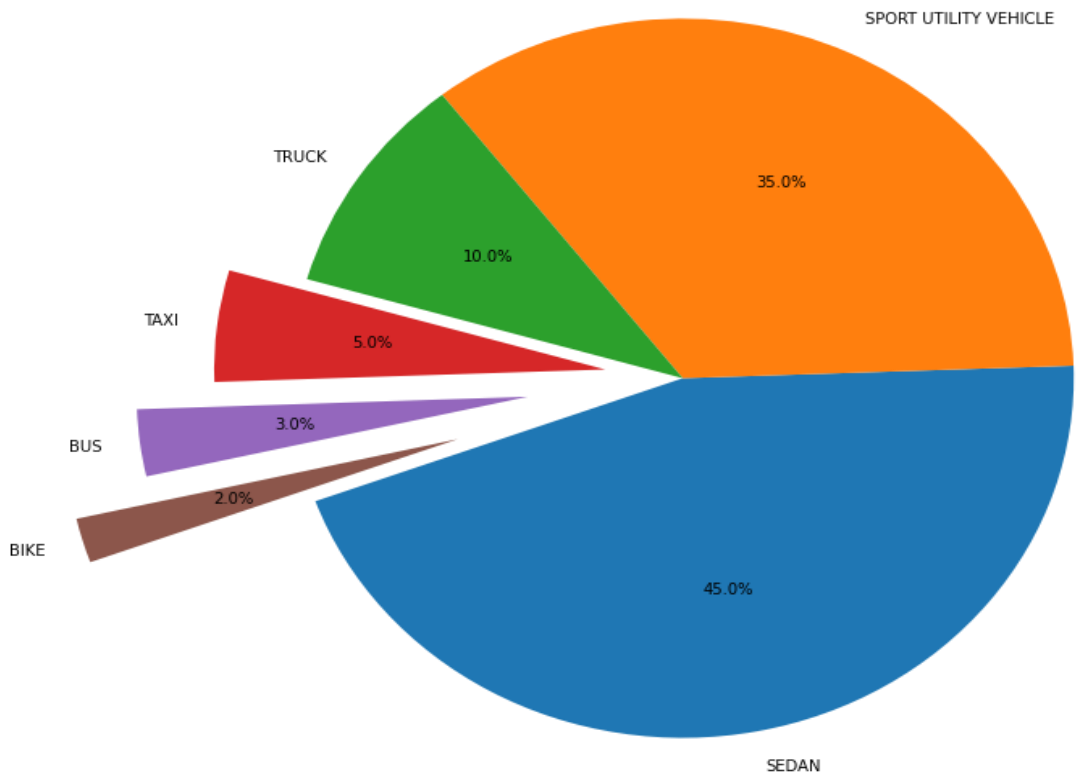
These were the reasons for maximum rate of decline in accidents in the month of March and April, people started to go out only for survival needs.

The vehicles were mostly used by emergency and health workers during those time periods.

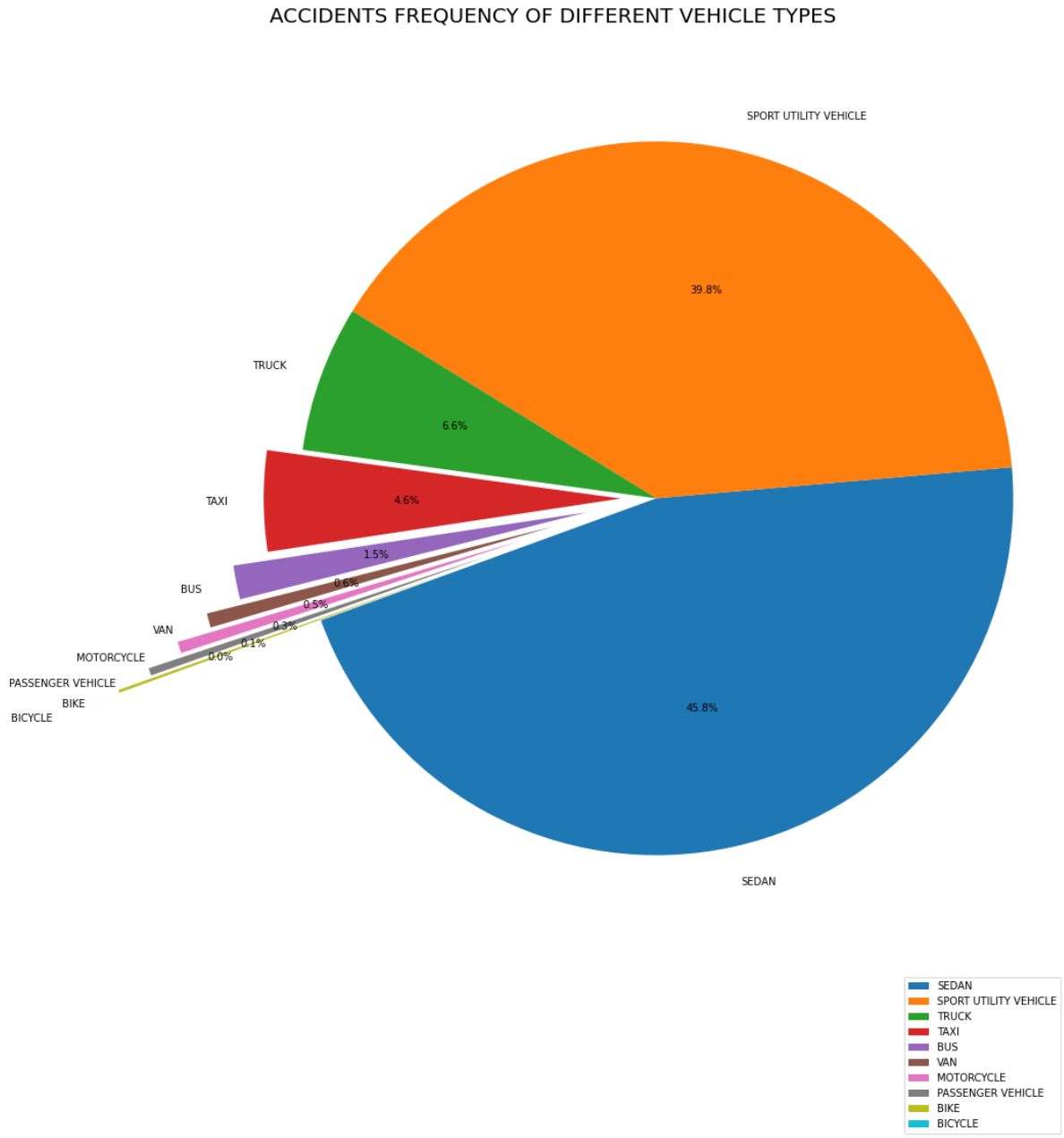
Analysis 3

Analysis 3 for Sample Data

For Analysis 3, we plotted a pie chart with the frequency of accidents occurred for the sample data for different VEHICLE TYPE. It is observed that the SEDAN is one which having the greatest number of accidents for the sample data.



Analysis 3 for Entire Data



- From the pie chart we can see that the maximum accidents were occurred by SEDAN and the least by BICYCLE.
- From the data from internet, we can see most of the accident-prone vehicles in USA are SEDAN.

  https://www.carlsonattorneys.com/news-and-update/10-dangerous-vehicles

- Thus, our analysis is correct for VEHICLE TYPE vs Frequency of Accidents by them.

Problems Encountered

- ✓ We were able to identify the state prefix that is not from the USA using the state prefix PDF of states since there was an additional state code in state registration that was not a part of us, but rather was on a vehicle from one of our neighbours' countries. The United States has 50 states and 5 union territories. However, there were 66 entries in our data. Further investigation revealed that these are the vehicles that were involved in accidents in NYC, albeit the cars may have come from anywhere. Ontario, Canada's state prefix, "ON," includes 261 accidents that occurred in the USA.

- ✓ There were numerous months without any accidents during the extraction of accidents that occurred in certain months. Thus, we had to deliver a list that included all the frequency of accidents for a given month with zeros for those months when no accidents were present. With keys ranging from 1 to 12, we created a dictionary and gave zeros priority and created another dictionary from the value counts of the data frame column ['MONTH'] for a particular vehicle and then changed the values comparing the two dictionaries.

- ✓ The preprocessing of the VEHICLE MAKE, and VEHICLE TYPE columns presented the biggest challenge. There were multiple entries with various counts for different vehicles for a same vehicle make and vehicle type. Some merely received the model's name. To group them, we had to conduct an internet search. Because various people will use different short forms, there may be differences in the data that are 12abelled differently as a result of entry errors. To find comparable patterns and replace them with the original label, we experimented with various algorithms.

Conclusion

According to our comprehensive investigation, the accident was primarily caused by the car that most individuals were traveling in. According to our research, the TOYOTA vehicle brand had the highest accident frequency, followed by HONDA and NISSAN. Most of the accidents were caused by authorized workers. We discovered that May is the busiest month for accidents in New York City. Out of the four vehicle manufacturers we had to analyse, TOYOTA was responsible for most accidents, while CADI had the fewest. Due to the COVID-19 pandemic's effects, there was a decrease in accidents in the month of April in 2020. We discovered that the decline began in the first few months of 2020 and peaked in April. The number of collisions then began to rise until it returned to normal.

The years mentioned were both pre- and post-pandemic, which had an impact on the accident by reducing people's freedom to move. Due to online learning, which further reduced the need for automobile transportation, academic activities were suspended at this time. It also had an impact on the production of autos because fewer people would purchase vehicles if they couldn't drive. On the other hand, weather changed during the winter because of snowfall, fog, and the possibility of slick roads, which had an impact on accident data. As a result, collisions are affected now. Thus, we draw the conclusion that a number of interrelated, related factors combine to cause automobile collisions.