**Project Report**

**Team 4: Hoan Le, Obalanlege Adewale**

**Analysis of New York Motor Vehicle Collisions**

The following Data Science methodology was adopted in this analysis to answer WHY, WHAT, and HOW in the following parts:

* Business Understanding: WHY Motor Vehicle collision?
* Data understanding: WHAT data?
* HOW to analyze the data: Python data science tools (Pandas, Numpy, Matplotlib, and Seaborn).

**Business Understanding**:

Why is motor vehicle crash the leading cause of injury-related deaths for New York State residents? Our group has been tasked to analyze data related to vehicle accidents in New York City with the goal to uncover potential patterns in the occurrences of accidents. The questions we are looking to investigate are:

1. Which vehicle makes the smallest and the largest accident counts?

2. Are some months more accident-prone than others?

3. Which vehicle type is more likely to be involved in an accident? This analysis has the potential to provide valuable insight for people and organizations who are working to reduce the risk of accidents in New York City.

**Data Understanding**:

To get insights into our questions above, we use the NY Motor Vehicle Collisions dataset file given which contains 3.7M rows and 25 columns. This data was collected between 2012 to today 2022. The Motor Vehicle Collisions table contains details about each vehicle involved in the crash. Each row represents a motor vehicle involved in a crash. We extract and filter the dataset to only contain the attributes relevant to our analysis.

**Overall Status:**

The data was extracted, cleaned, and sorted according to the parameters given. The codes used to process the data and plot it correctly were tested and changed several times to get better virtualization of the data frames and graphs. Overall, the project was completed and met all the requirements though it could be better.

**Data Preprocessing - Cleaning and Manipulation**:

The raw data was unzipped and uploaded to Google drive. A new. ipynb file was created through Google Colab and mounted to Google Drive to read the raw data. The loc function was used to extract data from July 2016 to Jun 2018, and the new data file was saved as a .csv file to Google drive. The purpose of this was to keep the original file in case something went wrong with it while processing data and working on the new data file was also faster because the file was smaller.

**Data Cleaning:**

Two additional columns ‘MONTH’ and ‘YEAR’ were added to the new dataset to generate the monthly and yearly accident occurrences. The dataset was checked if it contained any duplicates, wrong format, or missing values. 17 columns were dropped because they were not relevant to our analysis, and rows that contained nonvalue data were also removed from the dataset. Since only four types of vehicle-makes were needed, rows of ‘VEHICLE\_MAKE’ columns that the first 4 letters of the strings containing HOND, VOLK, FORD, and AUDI were selected for a new data set. Furthermore, we reclassified the vehicle type into 10 different categories by renaming, replacing, and adding ones that have similar categories together. The data now was clean and saved for further analysis. A sample of 100 rows from the clean data was randomly selected and tested for the efficiency of the codes used for counting, grouping, and plotting the charts. After all the codes and charts were satisfied, the same codes were applied to the whole data set.

**Code Developed:**

DASC5300\_Proj1\_Fall22\_team4.ipynb [DASC5300\_Proj1\_Fall2\_team4.ipynb](https://colab.research.google.com/drive/1Pc4IM6ki1wTCjNBNasX1eyxTcQKtIGWg?usp=sharing)

DASC5300\_Proj1\_team4\_Sample.ipynb [DASC5300\_Proj1\_team4\_sample.ipynb](https://colab.research.google.com/drive/1zxfynLZGOe1-_hwIl9bowgwZVEOUX71b?usp=sharing)

**File Descriptions:**

Motor\_Vehicle\_Collisions\_Vehicles.csv: original data set extracted from the zipped folder that contains information about vehicle accidents in New York City for many years.

mvc.csv: new data set extracted from the original data set that only contains the timeframe from 07/01/2016 to 06/30/2018.

clean\_motor\_vehicle\_collision.csv: clean dataset without vehicle type sorted. Data from this file was used to draw tables and charts.

ny\_motor\_vehicle\_collisions.csv: clean dataset after vehicle type has been sorted.

Description\_of\_the\_headers.xlsx: Description of the columns in the datasets

DASC5300\_Proj1\_Fall22\_team4.ipynb: This is a file from Google Colab and contains all the codes used in this project to clean and analyze data.

DASC5300\_Proj1\_team4Sample.ipynb: python Jupyter notebook that contains code developed for the analysis of Motor vehicle collisions.

Project1\_Report.doc: this word document contains tables, charts, and results from analyzing the motor vehicle collision data.

Code\_Develop.doc: a copy of all the codes used in this project.

Figure of plots.

Description of headers/columns.

**Algorithms Used:**

Python Libraries: Pandas, Ploty, Seaborn and Ploty.

Pandas Regular expression: pd.Series.str.contains.

Grouping: pandas goupby.

Python flow control: if, else statements.

**Division of Labor:**

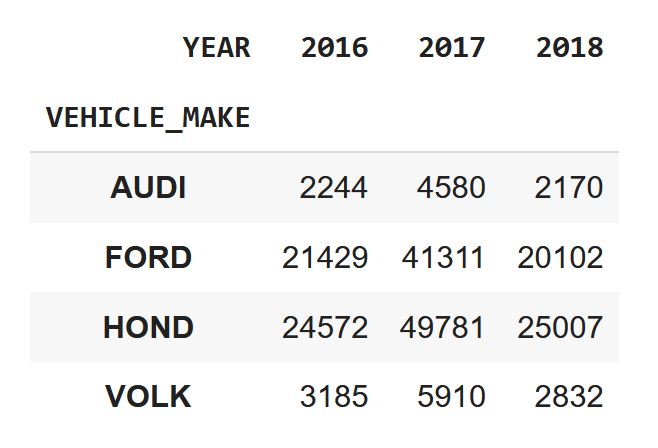
Each team member developed their own codes. After that, the codes were compared and combined to make them better. Each team member also wrote half of the project report.

**Problems Encountered:**

The data cleaning process and sampling were smooth and not too difficult. Plotting the bar chart and the line chart was also straightforward. However, when plotting the pie chart for the whole data set, there were so many vehicle types with different abbreviations. Grouping these vehicles together meant every abbreviation needed to be listed and went over one by one. We were able to overcome this with the help of pandas.unique method. A universal abbreviation for each vehicle type should be used when recording motor vehicle accidents so it helped the analyzing process more accurately. The pie chart also seemed to be crowded using the matplotlib and seaborn because only two vehicle types made up also 90% of the chart. Virtualizing the percentage of the other vehicle types was not very clear because they were too small. Seven or eight vehicle types used for the pie chart might be better. We were able to overcome this challenge with the help of the Plotly visualization library. Another problem we encountered when working with a large data set on Google Colab was reading the whole data instead of just the head or tail, the Ram on Google colab would run out of space and crash, and the whole analyzing process needed to be run again from the start. To avoid this issue, only the head or tail was used to read and check the data set.

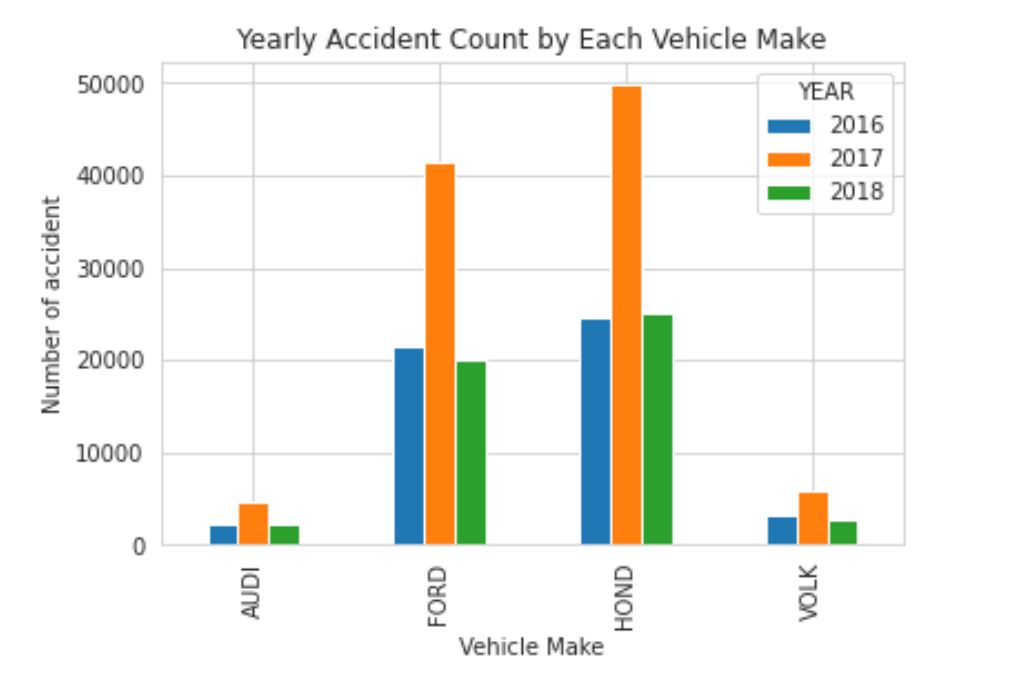
**Results:**

***1. The number of accidents that each one of the vehicle-make was involved in each year:***

******

**Table1:** *Number of accidents each vehicle-make involved from 01 Jul 2016 to 30 Jun 2018*

Table 1 was created after grouping the vehicle-make together and counting the number of accidents per year, then using the unstack function from Python to show the data as a frame.

******

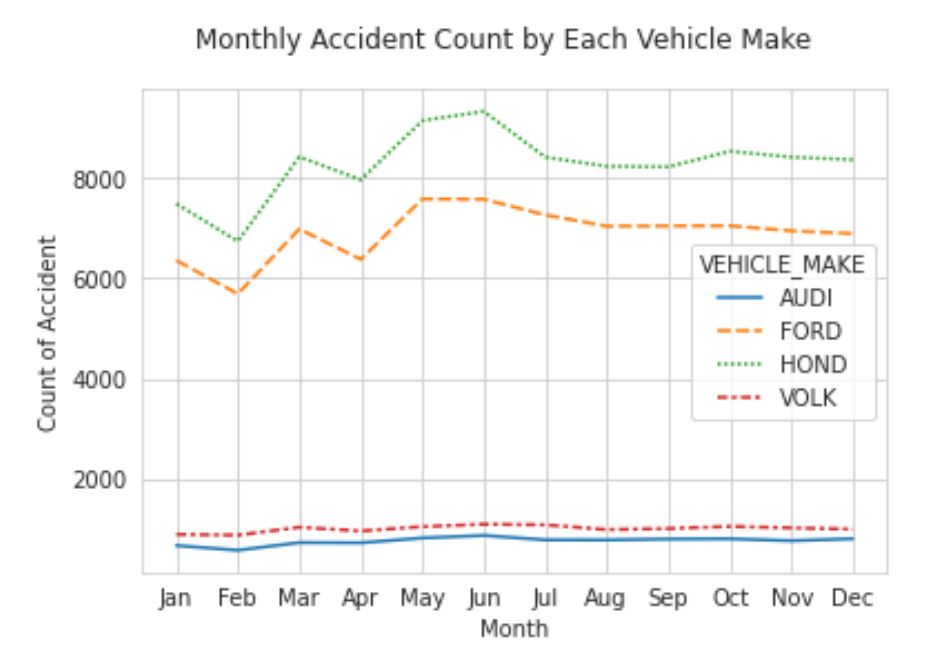
**Figure 1:** *Number of accidents involved with Audi, Ford, Honda, and Volk from Jun 2016 to Jul 2018.*

From July 2016 to June 2018, the number of vehicle accidents involving Ford and Honda was five times more than the number of accidents from Audi and Volkswagen in New York City, and Honda had the most accident counts which were almost 50000 cases in 2017. Ford and Honda vehicles were generally more affordable and popular in the population due to easy maintenance than Audi and Volkswagen, this might explain why there were more accidents involving these vehicles. In 2016 and 2018, the accident counts were only about half the numbers from 2017. This happened because the data collected from 2016 and 2018 was only half year instead of the whole year as in 2017. Therefore, there was no clear conclusion of which year had more vehicle accidents, however, comparing them by vehicle-make gave a better understanding.

***2. Monthly accidents counted from each vehicle-make each year:***

******

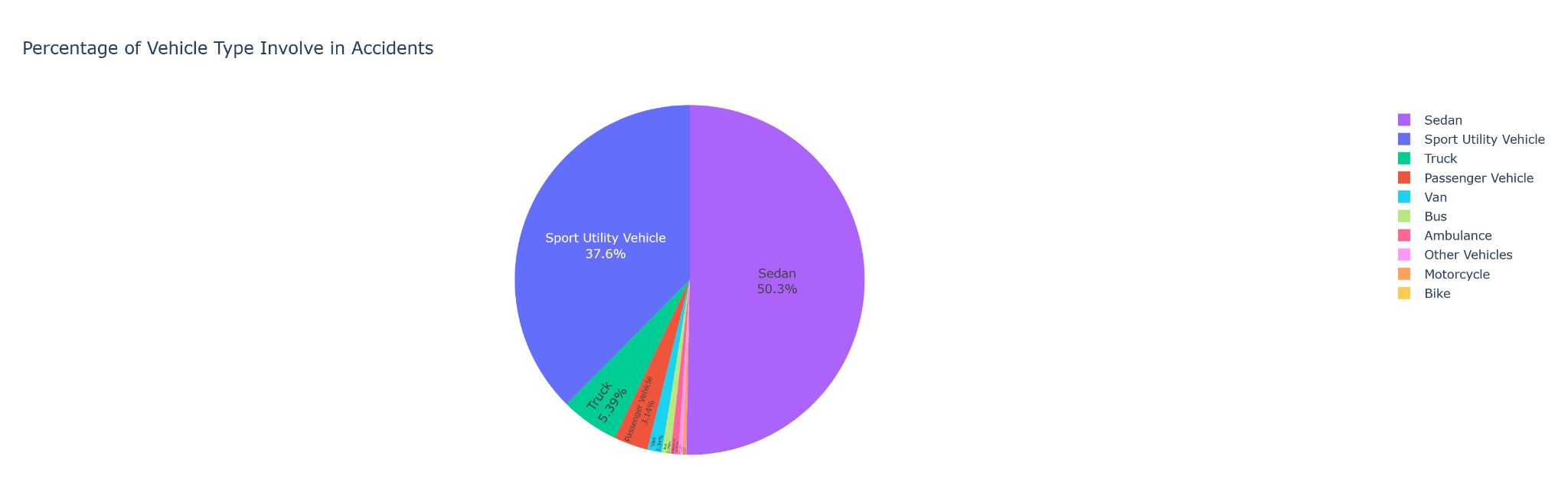
**Table2:** *The number of accidents involving each vehicle-make per month from Jul 2016 to Jun 2018.*

******

**Figure2:** *The number of accidents involving each vehicle-make per month from Jul 2016 to Jun 2018.*

The graph above showed that accident counts were almost the same in all months for Audi and Volkswagen’s vehicles. This was probably because of the small amount of Audi and Volkswagen vehicles that were operated compared to Honda and Ford. The number of accidents from Honda and Ford, however, showed clear trends of lowest counts during the month of Feb, higher counts in March, lower in April, and peak during May and June. This pattern might be explained by more traveling during the summer months of May and June after a long winter. From January to April, there were fewer people traveling because of the cold weather, however, snow often occurred in March so this could be the cause of higher accident counts. An example was the winter storm Stella left many areas of New York City with up to 13 to 18 inches of snow in March 2017 according to the New York Times.

**3. *Percentage of crashes each type of vehicle was involved in the given years*:**

****

**Figure3:** *Percentage of crashes each type of vehicle was involved in from Jul 2016 to Jun 2018.*

According to figure 3, sedan and sport utility vehicles made up almost 90 percent of vehicle accidents. Percent of accidents involved in Sedan vehicles was the biggest, 50 percent, the second one was sport utility vehicles, 38 percent, and the third one was trucks, 6 percent of the total accidents. Accidents involving buses, motorcycles, taxis, vans, and ambulances only accounted for a small percentage. According to several reports from 2017 and 2018 for the best-selling and safest vehicles, sedan and sport utility vehicles were more popular because of their size and safety. This was more likely the reason for the high percentage of accidents involved. Other types of vehicles used mostly for business, they were not popular as sedans and sport utility, and drivers drove more carefully because the vehicles were not for personal use so there would be fewer accidents involving these vehicles.

**Conclusions:**

Vehicle-makes and vehicle-types cannot be used to determine whether they are more likely to cause an accident even though the number of accidents involved in a certain vehicle make or vehicle type is significantly higher than the others. This number depends mostly on the popularity of the vehicles. Motor vehicle accidents are also higher during bad weather conditions and high travel times.

**References:**

[https://europe.autonews.com/automakers/audi-joins-german-rivals-dropping-out-new-york-show accessed 9/17/2022](https://europe.autonews.com/automakers/audi-joins-german-rivals-dropping-out-new-york-show%20accessed%209/17/2022)

<https://www.autoscopecarcare.com/audi-forum/how-much-does-an-audi-cost-to-maintain/#:~:text=Audis%20are%20fantastic%20cars%20to,national%20average%20for%20all%20cars>.

<https://www.nytimes.com/2017/02/09/nyregion/winter-storm-nyc-niko-weather.html>

<https://www.businessinsider.com/best-selling-cars-and-trucks-in-us-2017-2018-1>

<https://www.consumerreports.org/car-safety/iihs-names-safest-cars-2017-top-safety-picks/>