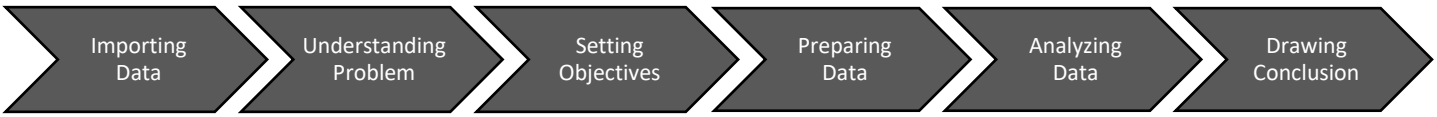
DASC5300- Project 1 (NY Motor Vehicle Collision Data Analysis)

Overall Status

The given project to analyze the NY Motor Collision data is successfully completed under the guidance of Professor Sharma and TAs. In this project, Exploratory Data Analysis was performed on the NYC Motor Vehicle Collision data. The data contains 3.7 million rows and 25 attributes. Each row represented a motor vehicle engaged in crash. The following approach was used to draw out meaningful insights from the data analysis.



Importing the data: The project is completed using Google Colab. The data was uploaded on the google drive. Google Colab allows to mount the drive by running a simple script in colab shell. After mounting the google driver, the data in CSV can be read into Pandas DataFrame using Pandas library in Python.

Understanding the problem: This is the most important step to ensure that the analysis does not go off track. Before starting to work with data, it is crucial to ask right questions and dig deeper to understand the bigger picture. The given task was to extract the two-year data out of the larger data set and use it for the analysis. Using the extracted data set, the following were the three problems that needs to be addressed.

- Number of accidents each Vehicle Make was involved in from June 2018 to May 2019 and from June 2019 to May 2020.

- Count of accidents each Vehicle Make experienced in each month from June 2018 to May 2020
- Crash percentage of each Vehicle Types over the entire span of two years.

Setting the objectives: Defining objectives helped to decide the path that needs to be followed for exploring the data and drawing out trends out of the extracted data. The objective of this exploratory data analysis is to use the visualizations to identify the trends in the data set. To solve the above-mentioned problems, certain visualizations were used to retrieve trends from the dataset.

- Bar plot for number of accidents each Vehicle Make was involved in from June 2018 to May 2019 and from June 2019 to May 2020.
- Line chart depicting the count of accidents each Vehicle Make experienced in each month from June 2018 to May 2020
- Pie chart illustrating the crash percentage of each Vehicle Types over the duration of two years.

At this stage, few parameters were selected out of 25 parameters that are required for analysis. The selected parameters were Crash Date, Vehicle Make, and Vehicle Type. The respective month and year were extracted from Crash Date and were stored into two separate columns. These two columns were highly used in this time series analysis. This step is very important to get rid of unwanted data and focusing on the required attributes.

Preparing the data: Raw data consists of missing values, redundant values, inconsistent values which may compromise the integrity of the analysis. It is very important to clean the data set before using it for the actual analysis. Cleaning data fixes the issues related to raw data for more accurate results. Getting rid of irrelevant data allows to focus more on relevant data points. The missing values can either be dropped off or filled with some value. Dropping values

is never suggested as it may result in loss of significant data. The cleaning of NY Motor Collision data was done in two parts.

- Cleaning Dataset for Analysis 1 and Analysis 2

  Both the analysis depends excessively on Vehicle Make and Crash Date columns. Vehicle Make column consisted of many missing values, null values, and NaN values. Moreover, the Vehicle Make column values had some irrelevant characters and spaces after the specified name of the Vehicle Make. It was necessary to clean the column values to extract the number of collisions for each Vehicle Make.
- Cleaning Dataset for Analysis 3

  This analysis used Vehicle Type and Crash Date columns to calculate the crash percentage for the assigned vehicle types. The Vehicle Type column had numerous irrelevant values. The simple approach was used to handle such values. All the irrelevant column values were replaced with NaN values and then NaN values were dropped off from the Vehicle Make column. After dropping the NaN values, next task was to sort the similar column values into respective categories. Using inbuilt Python functions, the similar data values were merged into separate categories and ambiguous values were ignored

Analyzing the data: After cleaning the data, this step involves slicing the data to extract meaningful results. At this stage, hidden patterns are drawn out. After looking for relationship among data, compelling visualizations are created by using appropriate graphs and charts. In this project Matplotlib library is used for creating data visualizations. Visualizing data is the best way to convey the analysis to the general audience and decision makers.

Drawing out conclusions: This is the last stage and an important one for which the whole analysis was done. Based on the plotted charts and graphs, results are drawn out and conveyed to the stakeholders. In this project, after looking at all the three visualizations, the final results of the NY Motor Collision analysis are interpreted.

Code Developed

The following is the google colab link for the python code for NY Motor Vehicle Collision Data Analysis

https://colab.research.google.com/drive/1i_lo_rIAPzNRZgklagF4ZEckeF7qe4x#scrollTo=wdplfbFnUVPS

The below is the google drive link for the CSV file containing Vehicle Collision data from June 2018 to May 2020

https://drive.google.com/file/d/1KB_CoOxiu-JHemenJo7EloOzQBLDqH34/view?usp=sharing

File Description

DASC5300_Proj1_Fall2022_Team_27.ipynb -- Google Colab Notebook containing the complete code. It contains the complete analysis including cleaning of dataset, analysis of all three problem, and visualization for all the analysis.

Function used to clean data for Analysis 1 and Analysis 2

- Clean_names(vehicle_name): This function searches for hyphen, extract that pattern and return the name excluding the extracted pattern.
- Remove_spaces(vehicle_name): This function searches for space and replace the space with no space.

Extracted_Two_Year_Data.csv – This is CSV file containing the NY Motor Vehicle Collision Data from June 2018 to May 2020**.**

Problems Encountered and Solutions

While doing the project of analyzing NY Motor Collision Dataset, I divided my work into various milestones. As the dataset was quite large, I faced challenges in attaining each milestone. But I tried to overcome these problems by taking help from online resources and python documentation.

Milestone 1: Importing data from CSV and extracting out data for two years in separate CSV

Problem Faced: I successfully imported data from google drive and read the CSV into the Pandas DataFrame. I faced some error while retrieving data for two years i.e. from June 2018 to May 2020.

Solutions: I overcame this problem by converting Crash Date column from string to date.

Milestone 2: Cleaning data for Analysis 1 and Analysis 2

Problem Faced: While I was cleaning data for Analysis 1 and Analysis 2, I faced issue in cleaning the Vehicle Make column. I successfully removed the additional characters after the hyphen. But still the data was not cleaned properly. Still there were some values having extra characters. Moreover, there were some extra spaces in the Vehicle Make column values. Also, some column values were in uppercase and other in uppercase. It was very important to remove these inconsistencies in data for better results.

Solutions: I created a function named Remove_spaces to remove extra spaces. Also, I converted all the lower-case Vehicle Make column values to the uppercase. To remove the extra characters, I replaced the string containing the irrelevant characters with the required string.

Milestone 3: Visualizing data for Analysis 1 and Analysis 2

Problem Faced: I successfully visualized the number of accidents each Vehicle Make was involved in for two years. But I faced some challenges while plotting graph for analysis 2. There was some confusion regarding how to group data based on months and years and then plotting separate line graphs for each Vehicle Make

Solutions: I referred Pandas documentation for grouping the column values in the DataFrame. Along with that I referred some material related to plotting separate line charts for each group. Pandas inbuilt function unstack() helped me to create line charts for each Vehicle Make.

Milestone 4: Cleaning data for Analysis 3

Problem Faced: While cleaning data for Analysis 3, I faced problem in separating the Vehicle Type column values into separate categories. There were so many values related to specific category. It was hard to get all of them initially. Moreover, there were many ambiguous values in Vehicle Type column. I faced some minor issues in handling such values.

Solutions: I overcame this problem by exploring various Pandas functions. I used unique() function to get all the values containing substring of particular category and then replaced all of them with the main category. To handle the ambiguous values, I just considered those column values that are equal to specified categories and then ignored all the ambiguous values.

Milestone 5: Visualizing data for Analysis 3

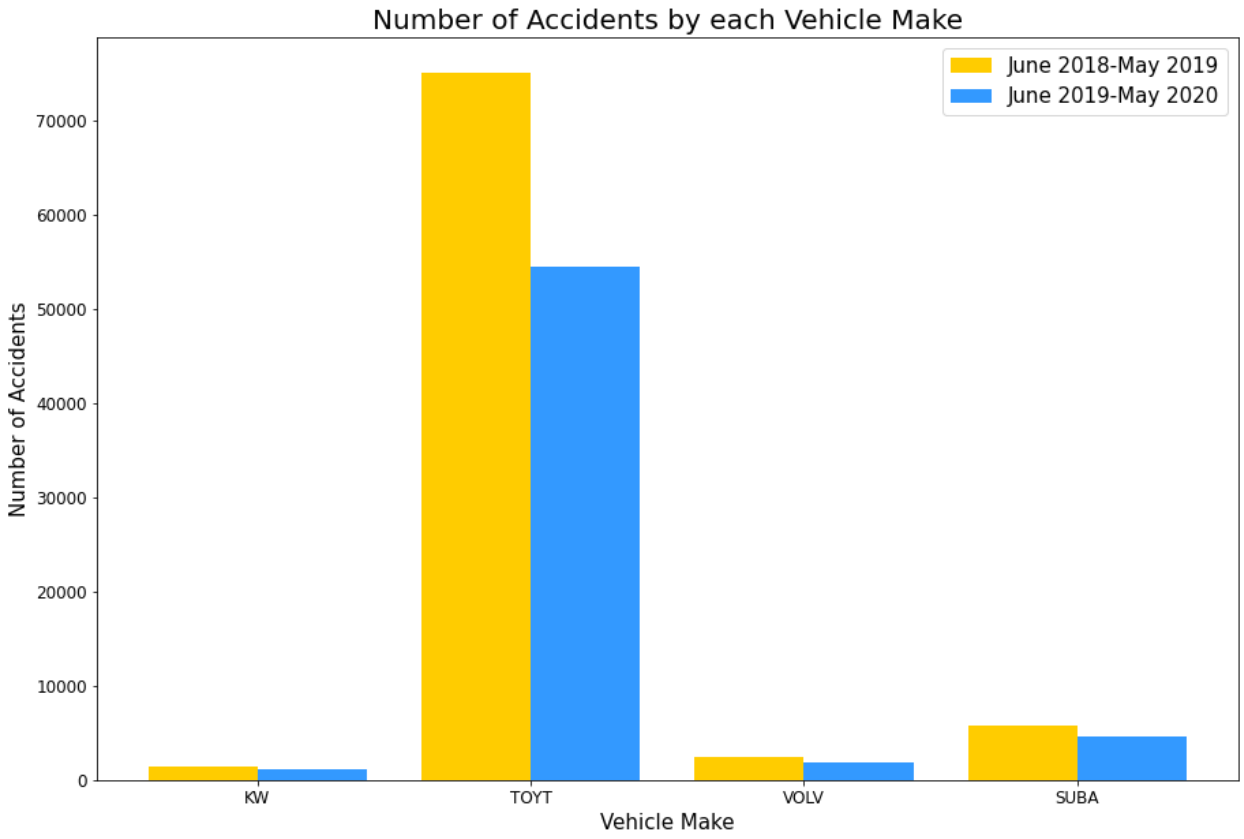
Problem Faced: While plotting pie chart for Analysis 3, I faced challenges with the overlapping label values and percentage values

Solutions: I overcame this problem by taking out the labels and placing them in a separate legend. To avoid overlapping of percentage values, I used explode parameter to make few wedges stand out of the pie chart. This solved the problem of overlapping percentage values.

Analysis and Results

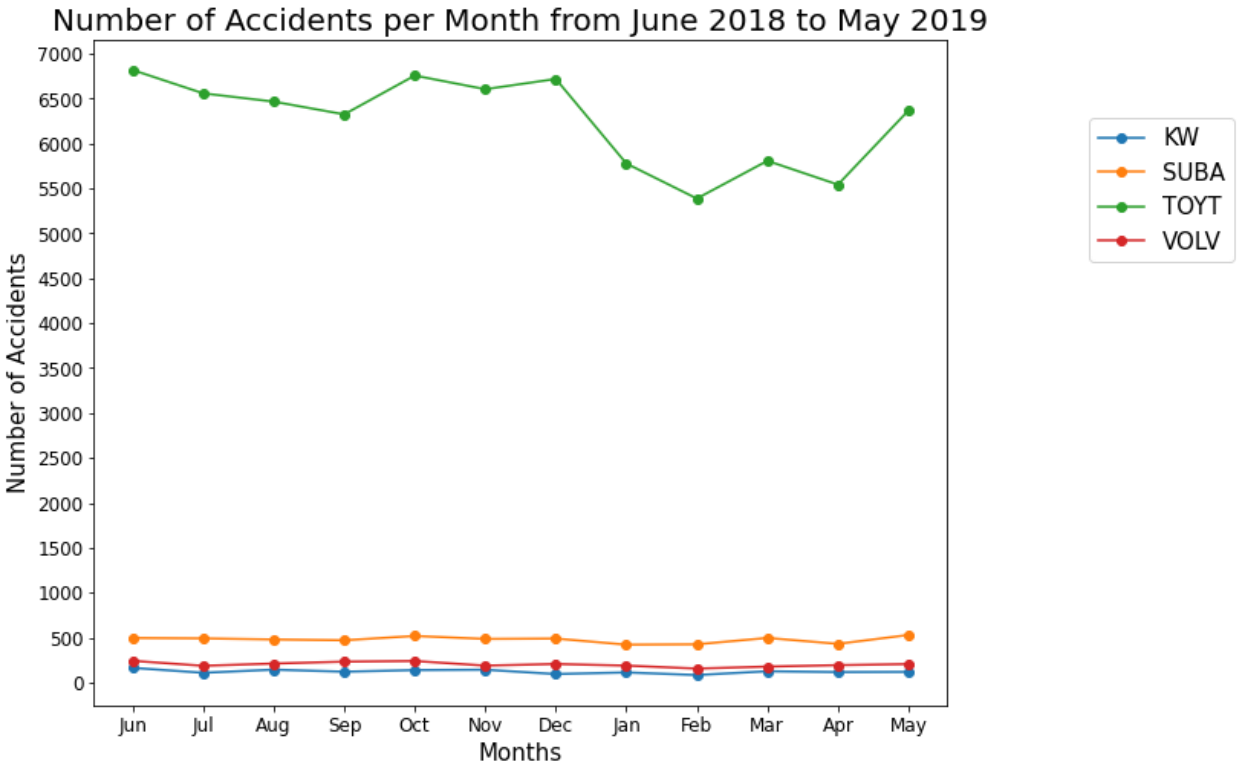
As part of this project, the problem is divided into three sub problems. After analyzing each problem, the results are conveyed through the visualizations. The following is the analysis and results for each problem.

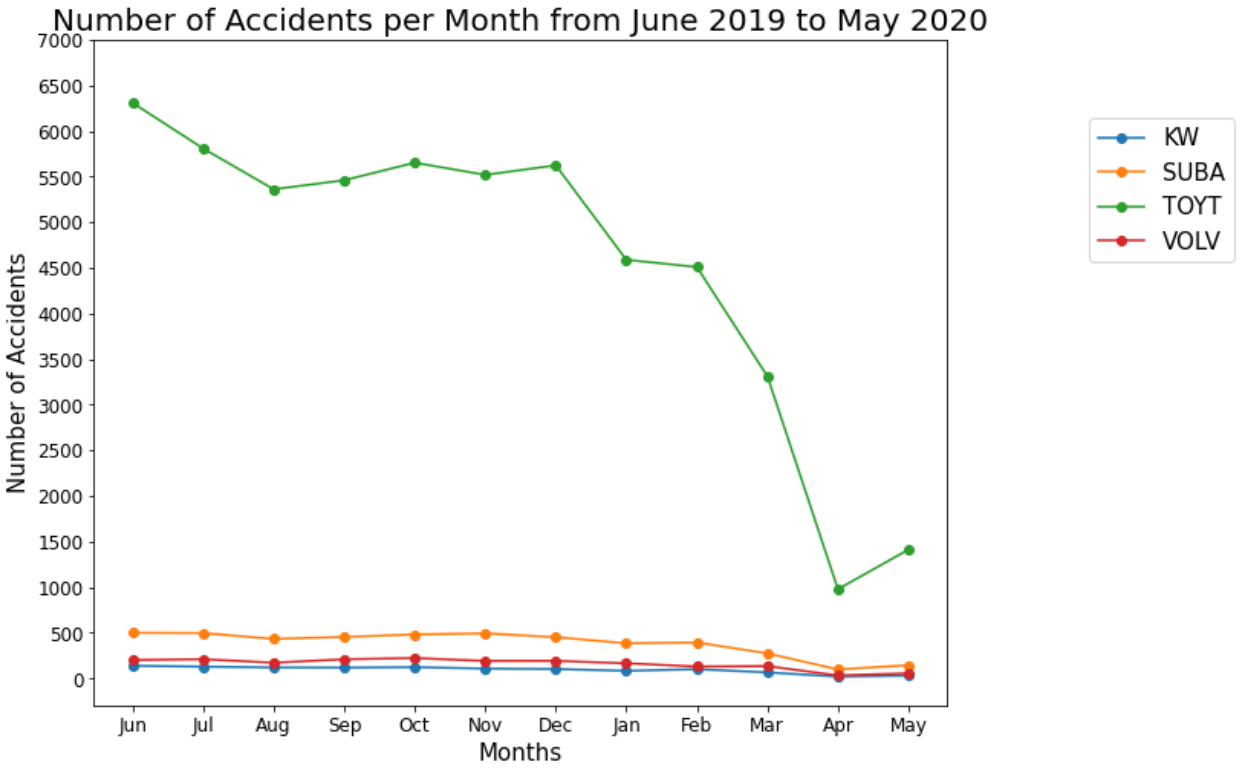
Analysis 1: Number of accidents each Vehicle Make was involved in from June 2018 to May 2019 and from June 2019 to May 2020



Referring the above visualization, it can be conveyed that Toyota was involved in most of the crashes over the span of two years. Also, it was seen that for each vehicle make the number of accidents were more from June 2018 to May 2019. However, it is observed that KW experienced least number of collisions as compared to other vehicle makes.

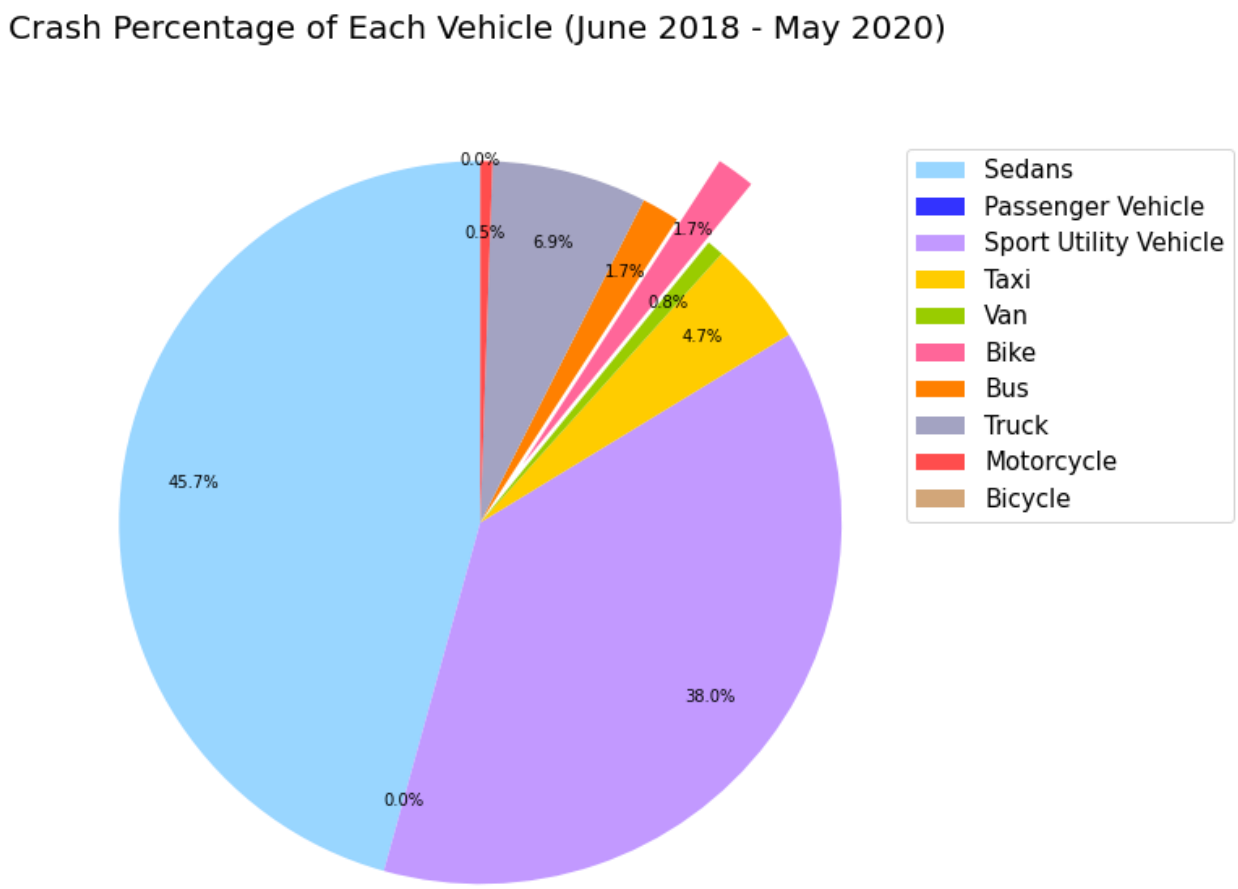
Analysis 2: Count of accidents each Vehicle Make experienced in each month from June 2018 to May 2020





Referring to the above visualizations, it can be concluded that Toyota outstood in number of collisions happening each month from June 2018 to May 2020. Moreover, it was seen that Toyota experienced a drastic fall in number of accidents from February 2020 to April 2020. Apart from this, it was observed that other vehicle makes did not experienced much fluctuation in number of accidents per month for the entire time span of two years.

Analysis 3: Crash percentage of each Vehicle Types from June 2018 to May 2020



The above pie chart illustrates that Sedans were involved in the most number of accidents over the duration of two years. Also, it can be stated that the Sport Utility Vehicles accompanied Sedans in the crash percentage from June 2018 to May 2020. However, the number of accidents experienced by Bicycle and Passenger Vehicle were negligible.

Conclusion

After performing the complete analysis, it can be concluded that Toyota Sedans experienced most number of collisions. Referring to the articles on internet it was observed that Toyota vehicles grabbed most of the positions in the top 20 best-selling cars and trucks. Also, it can be suggested that the vehicles produced by KW are the safest to buy as compared to other manufacturers. Overall, it was seen that people who used bicycles and passenger vehicles were safe from such crashes.

Other References

https://pandas.pydata.org/docs/reference/frame.html

https://www.w3schools.com/python/pandas/default.asp

https://www.w3schools.com/python/matplotlib_bars.asp

https://www.w3schools.com/python/matplotlib_line.asp

https://www.w3schools.com/python/matplotlib_pie_charts.asp

https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.bar.html

https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.plot.html

https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.pie.html

https://www.businessinsider.com/best-selling-cars-and-trucks-in-america-in-2018-2018-8

https://www.caranddriver.com/news/g27041933/best-selling-cars-2019/