A PROJECT REPORT ON

PYTHON PROGRAMMING AND DATA ANALYSIS: PROJECT 1

BY

ABHISHEK SHRINIVAS JOSHI

ANUSHKA CHANDRAKANT PAWAR

UNDER THE GUIDANCE OF PROF. SHARMA CHAKRAVARTHY

CONTENTS

INTRODUCTION

In this project, we learned and implemented the basics of Data Analysis. We have Motor Vehicle Collisions dataset. The dataset has 3.7M rows and 25 columns for analysis. We have analyzed this motor vehicle collision data for 3 years from May 2016 to April 2018. Python is one of the most suitable languages for data analysis today. The analysis is performed using the programming language Python as it has a variety of libraries available. We used libraries like pandas, matplotlib, and NumPy to analyze data, draw conclusions on the given dataset, and get meaningful results.

DATASET

The Motor vehicle dataset has 3.7M rows and 25 columns. The data set consists of the details of cars or different vehicles involved in crashes in New York City from the last 10 years (2012-2022 daily). The data has a total of 25 attributes for each vehicle crash. Each attribute has various values in each field. There are some 'Nan' values or blank columns in the data as the data is not complete data and needs pre-processing.

WORK DISTRIBUTION

- Data pre-processing is done by Anushka Chandrakant Pawar
- Data Analysis and execution of queries was done by Abhishek Shrinivas Joshi
- Conclusion and Report are Made by both equally.

METHODOLOGY

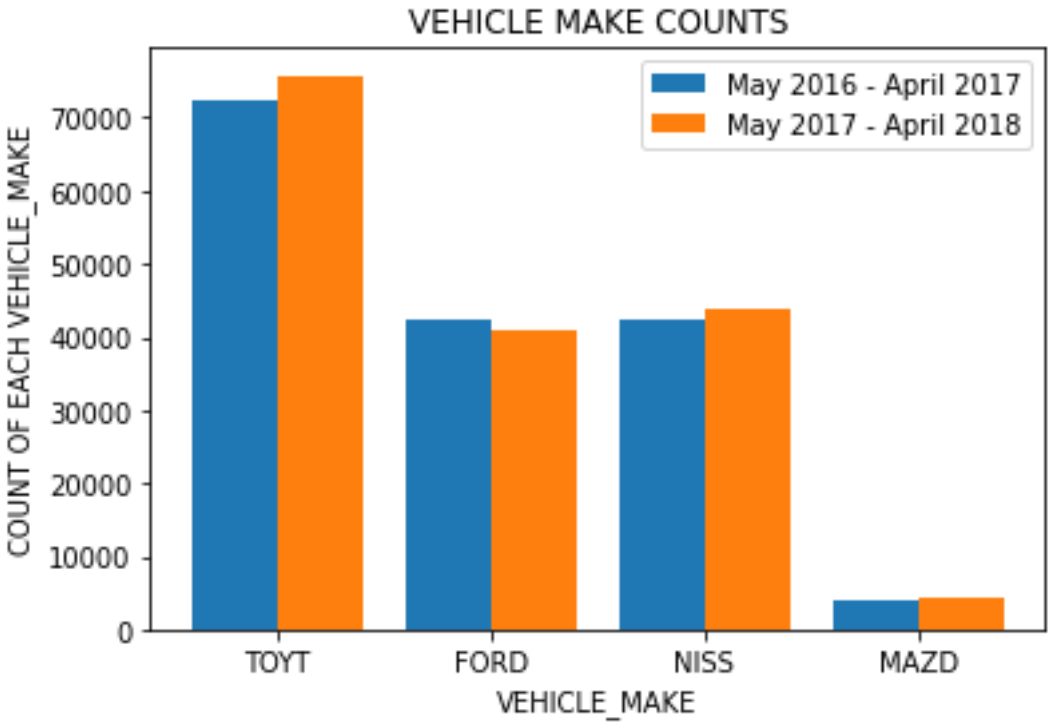
- We used Google Colab to write our code. It allows to write python code and execute it through a browser. Google Colab is a Jupyter notebook service that provides the same application as Jupyter with no setup needed. Google Colab is suitable for Data Analysis and Machine learning.
- Firstly, we imported the libraries required for the project. We imported the pandas library, NumPy library, and regex library.
- Then the Data set is uploaded on Google drive and is fetched from a path given in the program.
- We needed the data between May 2016 to April 2018. Hence, we converted the Crash_Date attribute to data Frames so that we can compare the value of the Crash_Date attribute to perform DateTime operations.
- By giving a start and end date we performed a DateTime operation and extracted data of the years allotted to us.
- The next step is to clean the data that we extracted. We remove Nan values from the dataset by using the .dropna function and obtain a clean dataset to work on.
- As per our given parameters, we extracted four specific vehicle Names from the vehicle_make attribute by using the regex library for operations like find and replace. We got the desired names (ie. TOYT, FORD, NISS, MAZD)
- We took a random sample using seed value as Date of Birth of about 100 tuples and performed an analysis of this smaller set of data.
- As we got meaningful results from the sample, we performed an analysis of the entire data from May 2016 to April 2018.
- For data visualization we plotted three types of charts.
- First is Bar Graph. We plotted 2 separate bar graphs by using the plt.bar function. One is where the graph shows analysis after dividing the data into two parts(years) and the second graph for three years. We plotted the vehical_make attribute on the x-axis and the count of the number of accidents on the y-axis.
- Second is a Line graph. Here, we used the plt.plot function. We plotted months of the accidents on the x-axis and the count of accidents for each vehicle make on the y-axis.
- Third is a Pie chart. We plotted two pie charts as well by using the plot.pie function. First with vehicles given for our group and second where we took other vehicle types to complete 10 vehicle type conditions given in the problem statement.

ANALYSIS AND RESULT

Bar Graph:

First Bar Graph

- According to analysis, TOYT vehicle make is involved in most crashes in the years 2016 to 2018 specifically in 2017-2018 (no. accidents= 75724) when data is divided into two parts form (May 2016 to April 2017) and from (May 2017 to April 2018).



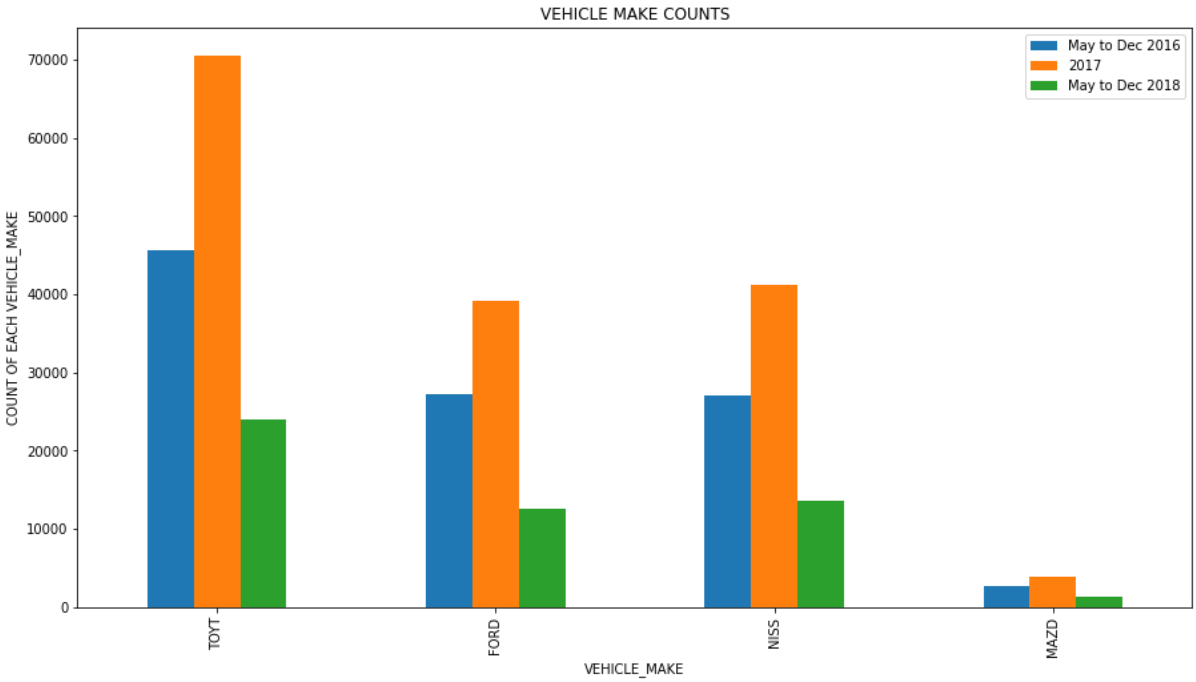
- The vehicles from MAZD and involved in the least accidents.
- Number of accidents in 2017-2018 is more than the number of accidents in 2016-2017.

Reuslt:

	TOYT	FORD	NISS	MAZD
May 2016 - April 2017(No. of Accidents)	72239	42200	42506	4100
May 2017 - April 2018(No. of Accidents)	75724	41027	43914	4253

Second Bar Graph

- Here, we plotted graph for three different years.
- According to this graph, we got the same conclusions as from the previous one.



Result:

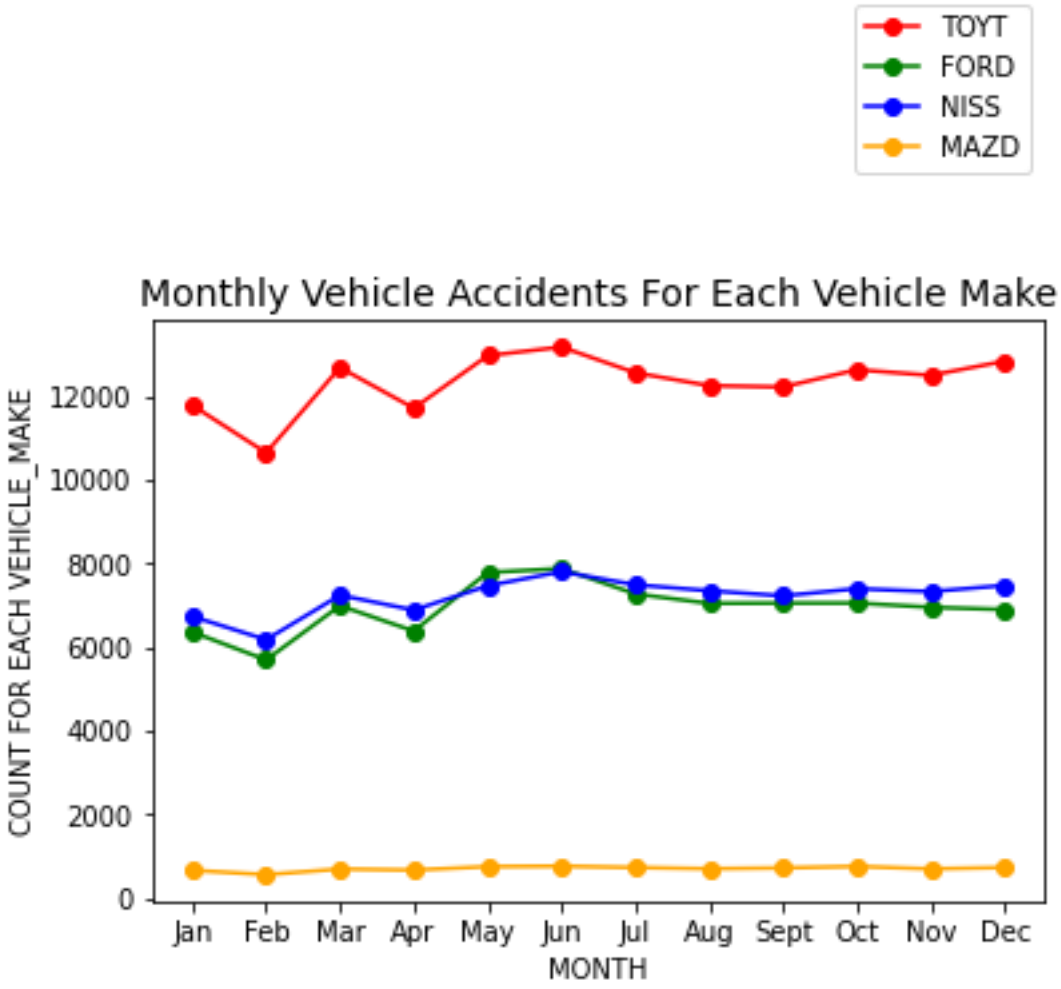
	TOYT	FORD	NISS	MAZD
May to Dec 2016(No. of accidents)	45671	27304	27018	2634
2017 (No. of accidents)	70571	39233	41184	3958
May to Dec 2018(No. of accidents)	23959	12579	13662	1320

Line Graph:

- Number of accidents in the month of June is more than in other months as according to seathecity.com website June is the peak month for tourist visits.
- The vehicle make which is involved in the most number of accidents is TOYT.
- Summer months have most number of accidents than other months whereas winter has the least number of accidents
- Feb has the least number of overall accidents
- There are many festivals like New York Sounds of Summer International Music Festival, Fish Parade, and Summer Festival and so on and hence attracts tourists a lot in the month of June hence it had a large number of accidents
- TOYT vehicle make had a large number of accidents then any other provided makes.

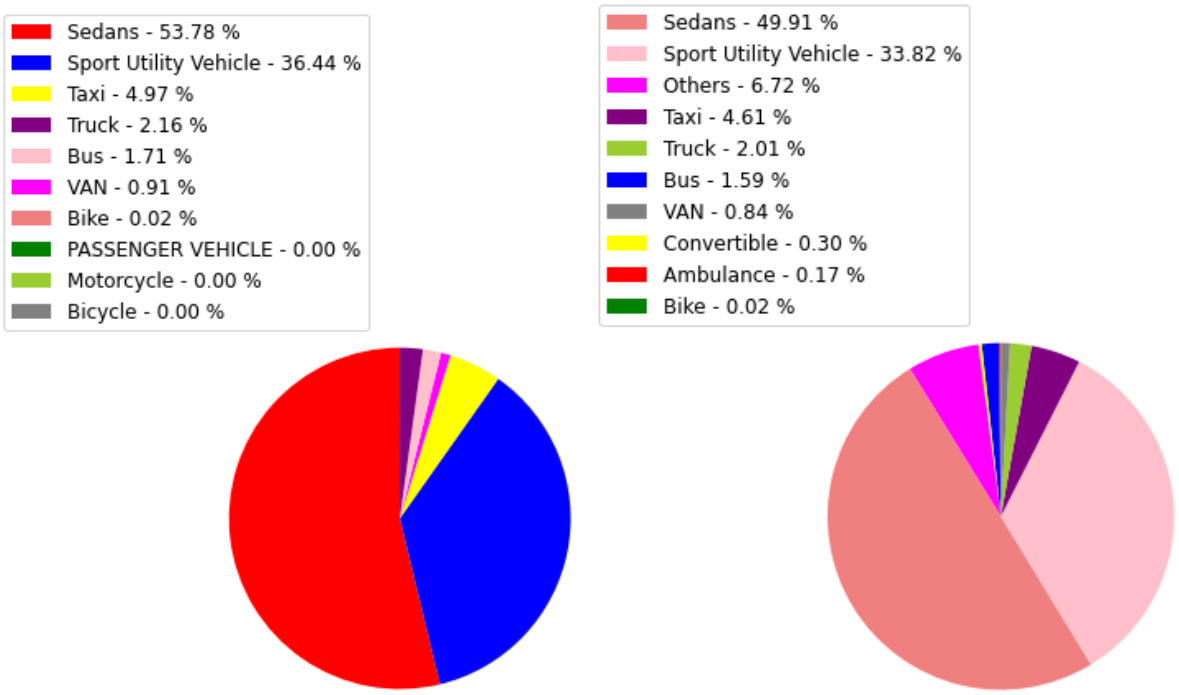
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
TOYT	11782	10645	12685	11712	12969	13176	12560	12252
FORD	6359	5690	6988	6381	7781	7876	7269	7043
NISS	6731	6175	7239	6883	7464	7801	7488	7346
MAZD	657	554	688	664	746	756	725	696

	Sep	Oct	Nov	Dec
TOYT	12219	12632	12498	12835
FORD	7050	7051	6950	6896
NISS	7223	7395	7325	7472
MAZD	715	751	693	725



Pie Chart:

- Sedans is the vehicle type involved in most accidents. And Sports utility vehicles follow it on the second to be involved in crashes.
- Sedans are in 49.91% of crashes in New York City followed by sports utility vehicles are in 36.44% accidents.



Overall Result

From the given data set we have drawn the following conclusion

- June month had the most number of crashes.
- TOYT vehicle make involved in the maximum number of accidents
- Sedan vehicle type had the highest percentage of crashes.
- Crashes that occurred in the years 2017-2018 are more than crashes that occurred in the years 2016- 2017.

DIFFICULTIES FACED IN THE PROJECT:

- We had data from three different years hence we faced difficulties while analyzing the yearly data. Hence we created two bar plots one for three different years and one by dividing extracted data into two equal parts.
- Given parameters were 10 and we didn't have 10 parameters as some parameters had 0 values. So to overcome this, we plotted 2 pie charts. The first included parameters given to us and the other included some random vehicle types from rest of the data.
- We also faced problem for merging the similar categories of vehicle types because some tuple has values different from given parameter files. To overcome this difficulty we created a function named clean_types so that we can merge similar categories of same vehicle make.

REFERENCES:

1. Project Helper Slides
2. Geekforgeeks site
3. Seathcity.com
4. Stack overflow
5. Python Documentation