**PYTHON PROGRAMMING AND DATA ANALYSIS**

of

**NEW YORK VEHICLE MOTOR COLLISION**

By

Bharath Raj Varma

Student ID: 1001848518

&

Akshata Nitin Agine

Student ID: 1002065578

**OVERALL STATUS:**

1. We mounted the google drive to the colab and read the CSV file from the problem statement. After reading the CSV file, we extracted the data according to the given years.
2. For the preprocessing part, we extracted relevant portions of data like “Vehicle Make” and “Vehicle Type” and removed NaN values.
3. By grouping the “Vehicle Make” and “Year” column values, we took the count of accidents that happened in a year (2016, 2017, 2018) for a particular Vehicle Make (HOND, ISU, FORD, CADI).
4. For analysis, we plotted Bar Graph for 3 years for a particular vehicle make and also plotted a Line Graph for every month (October 2016 - September 2018).
5. Furthermore, we plotted a Pie Chart for only 6 Vehicle Types.

**CODE DEVELOPED:**

The code developed for Preprocessing and for Analysis is mentioned in a [.ipynb file](https://colab.research.google.com/drive/1yLjqG7kOmLr5pIGiL3uXcULhgu0unJmg?authuser=1#scrollTo=5KQKvZwC-i81).

**PREPROCESSING:**

* IMPORTED LIBRARIES FOR PREPROCESSING

1. Pandas
2. Datetime
3. Re

* READ/WRITE CSV FILE

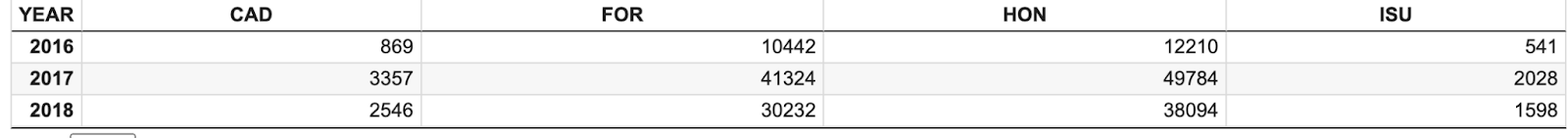
1. The function used to read the data given in the problem statement from the CSV

* MISSING VALUES
  1. Used the **dropna()** function to remove all the NaN values from “Vehicle Make” and “Vehicle Type”.
* CATEGORIZING THE DATA
  1. By using the **to\_datetime()** and **query()** functions we have extracted data of the years pertaining to our dataset.
  2. To separate month and year from the dates provided we used the **datetimeIndex()**  function.
* DATA CLEANING
  1. By using the re library, we have extracted only the first three string values from the “Vehicle Make” column.

Ex: For Instance, Initially, it was, “FORD - CAR/SUV”. After using the function mentioned above and only considering the first three string values, we got “FOR”.

* 1. Additionally, as per the data given for our project group, we need to extract data only pertaining to four Vehicle Makes i.e., HOND, FORD, ISU, and CADI.
* COUNT OF THE VEHICLE MAKE

1. By using **groupby()** and **count()** we get a count of accidents that happened by a “Vehicle Make” for that particular year as mentioned below.



**ANALYSIS:**

* LIBRARIES USED
  + 1. Numpy
    2. Matplotlib
* BAR GRAPH

**Chart, bar chart

Description automatically generated**

* The above bar graph is plotted using the ‘count of the vehicle make’ table above.
* Comparatively, Honda Accord and Ford F-150 series vehicles are the most accident-prone vehicles in New York. But whereas ISU vehicles are less accident-prone vehicles.
* LINE GRAPH

A picture containing diagram

Description automatically generated

* The above Line graph is depicted using the table below which we got from counting vehicles made per month.

**Table

Description automatically generated**

* By contemplating the graph, we can say that Honda and Ford’s vehicles are more accident-prone than other vehicles. If we compare accidents in summer vs winter, the accidents count is more in the summer months than in winter. Even though summer has numerous accidents count, winter months accidents count is still high comparatively
* PIE CHART

Chart, pie chart

Description automatically generated

* The above-depicted pie chart shows the percentage of accidents that happened for different kinds of vehicle types.
* From observing the pie chart, we can say that sedans are more most accident-prone vehicles in New York City. And next on the list are sport utility vehicles.

**DIVISION OF LABOUR:**

* Programming parts were equally shared, and the report was majorly done by Akshata Nitin Agine.
* The programming was done in one week.
* The project report was completed 3 days before the submission date.

**PROBLEMS ENCOUNTERED:**

* We were asked to extract Vehicle Make values up to 4 characters as in “FORD - SUV/CAR” to “FORD” using the **re** library
* However, we are getting some redundant values apart from FORD, HOND, ISU, and CADI. Therefore, we had to extract only the first three string values as ISU has three letters in it.
* For separating month and year from ‘CRASH\_DATE’, we tried functions that are mentioned in the TA slides. But we have encountered an error. So, we have used **DatetimeIndex** function to separate month and year.
* In analysis, we are asked to plot the pie chart for 10 different Vehicle Types, but we are only able to plot 6 of them.

**REFERENCES:**

* Some code snippets are taken from the problem statement.

[**https://itlab.uta.edu/courses/dasc-cse-5300/current-offering/project-1-nyc- accident-analysis/Fall2022\_project1\_v4.pdf**](https://itlab.uta.edu/courses/dasc-cse-5300/current-offering/project-1-nyc-%20%20%20%20%20%20%20accident-analysis/Fall2022_project1_v4.pdf)

* DatetimeIndex functions and datetime libraries references from:

**https://pandas.pydata.org/docs/reference/api/pandas.DatetimeIndex.html**