**Maryland – Traffic Violations Project Report**

GitHub Url

https://github.com/nisha067/UCDPA\_nishababu

Abstract**:**

The purpose of this project is to understand how much traffic violation occurs in a particular state of a well developed country by concentrating on the top category of vehicle type. With the extracted data I am going to illustrate couple of visualization to understand how the metrics can help public and department.

Introduction:

Traffic violation is a common practice across the Globe and it has different levels of impact for common people, Government properties and Environment. By taking this project I would like to see the top contributors in all angles. The output of the project can be used as a summary view by the government officials to see the details what they collected throughout the year and where they need to focus more to reduce such violations in the future.

Based on the metrics they can potentially introduce some sort of additional speed breakers, Traffic light (or) traffic controller who can potentially reduce the volume of traffic violations.

Also Based on the project analyzes there will be a potential option for the officials to make improve the case registration process and they can understand the data quality.

Dataset**:**

I am opting for the below data source from Kaggle.[**https://www.kaggle.com/rounak041993/traffic-violations-in-maryland-county**](https://www.kaggle.com/rounak041993/traffic-violations-in-maryland-county)

The reason for choosing this data source is because it has all the potential different data formats within the same CSV file. All these data will help to do my code in different angles. Also I am thinking of extending this exercise beyond my project submission to explore more around the latitude and longitude information to see if can connect the information with the real world information’s.

Implementation Process:

As discussed earlier I am going start the exercise with the source as CSV file downloaded from the kaggle data source.

Instead of consuming the entire dataset from CSV file. I will be selecting the below list of fields for my project exercise. The filtering will be done while reading the records from CSV file using panda data frame.

|  |  |
| --- | --- |
| * Date Of Stop * Time Of Stop * Description * Location * Accident * Belts * Personal Injury * Property Damage * Fatal * Commercial License * HAZMAT * Commercial Vehicle | * Alcohol * Work Zone * State * VehicleType * Year * Make * Model * Color * Violation Type * Gender * Driver City * SubAgency |

Once the selected record processed and stored in the variable. I am going to group the full dataset based on “VehicleType” column and identify the top category of vehicle type.

In my use case I am planning to invoke an API call exposed in the internet by a third party with complete list of car details like year, make, model etc returned from the API.

Since the dataset which I downloaded primarily have the Automobile records, the code should work to validate the information with the third party data.

I am handling the API call section with a backup mechanism. If API is up and running we will always pull the records from API and will use it to slice our dataset for valid vehicles. In case if the third party website is down I prepared a JSON file (CarsDetails.JSON) with similar entries as a backup file.

The code will automatically pick the entries from the JSON file locally stored with the project.

Since this is a third party API call, the data returned from the website may or may not be accurate (or) up to date with the models getting released every year. To manage this situation I prepared a sample record “missingCompanyNames” under “application.properties” file.

Any admin who runs this project can configure additional model names which are missing from the API response. So that those vehicles will not be treated as invalid ones part of our data slicing.

Right we will have two dataset (Input Dataset & Dataset extracted from API for lookup), as a next step convert all the dataset to lower case to compare the entries between data frame.

With the help of looping/iteration and if, else condition I will be dividing the data frame into two separate set. One set will have all valid records that have a valid company Name. The other one will have the records where the company Names in the input record not matching with the API response and manually prepared list. This piece of division will be done with the help of a function.

With the valid Records data frame I am planning to show multiple visualization.

The invalid Records set can be used to understand the mistakes made by the officials who created record in the system. This further used to improvise the process of entry.

Results**:**

****

****







Insights**:**

With the help of python code I can able to see the below information’s

* I can able to produce a report based on the top category of vehicles involve in frequent accidents. This insight information will help police to produce a report to top officials for further evaluation.
* Gender based accident ratio extracted in the second image and with this department can have a close look on regular basis to understand if there is a increase in trend or not.
* Generated a visualization which helps department to identify the top locations where more frequent cases getting registered. With this data they can introduce more regulation in those location to avoid more such accidents and can prevent public.
* Alcohol based accident ratio evaluated in the 4th image, this shows the volume of accidents at present is very less. This report can be generated on regular basis by the department to understand if there is a increase in the consumption of alcohol.
* The last plot represents the branch where more number of cases registered under Maryland Police Control. With this information and the 3rd Visualization they can improvise traffic.

References**:**