Project: Capstone Project 1 - Data Wrangling

Data Source: - https://openpolicing.stanford.edu/data/

## ABOUT THE DATA

The Raw data for this project contains the traffic stop data collected for 30+ states for open police project by Stanford research team. Standardized stop data are available to download (by state) from the link above provided by Stanford.

The csv includes a subset of common fields for each state, and indicates whether data are available for at least 70% of records in that state. Some states have more fields.

The original, unprocessed data we collected contain even more information.

The Stanford Open Policing Project data are made available under the [Open Data Commons Attribution License](https://opendatacommons.org/licenses/by/summary/).

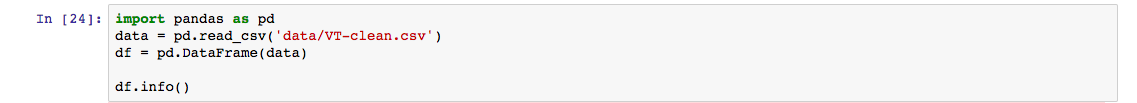
Downloaded excel sheet of raw data for VT: -

<https://github.com/jiagarwa/capstone-project1-Jitendra>

file name: - VT-clean.csv.gz

## Accessing and filtering the Data

1. Read data from csv file VT-clean.csv and Show the columns in data and their type.



Csv has a total of 23 columns

Data columns (total 23 columns):

id 283285 non-null object

state 283285 non-null object

stop\_date 283285 non-null object

stop\_time 283285 non-null object

location\_raw 282591 non-null object

county\_name 282580 non-null object

county\_fips 282580 non-null float64

fine\_grained\_location 282938 non-null object

police\_department 283285 non-null object

driver\_gender 281573 non-null object

driver\_age\_raw 282114 non-null float64

driver\_age 281999 non-null float64

driver\_race\_raw 279301 non-null object

driver\_race 278468 non-null object

violation\_raw 281107 non-null object

violation 281107 non-null object

search\_conducted 283285 non-null bool

search\_type\_raw 281045 non-null object

search\_type 3419 non-null object

contraband\_found 283251 non-null object

stop\_outcome 280960 non-null object

is\_arrested 283285 non-null bool

officer\_id 283273 non-null float64

dtypes: bool(2), float64(4), object(17)

1. Filter data for Year 2015: -

Convert data into a data frame and filter based on year from stop\_date column and save in a separate file Filter data by year 2015. Call it VT\_2015.csv

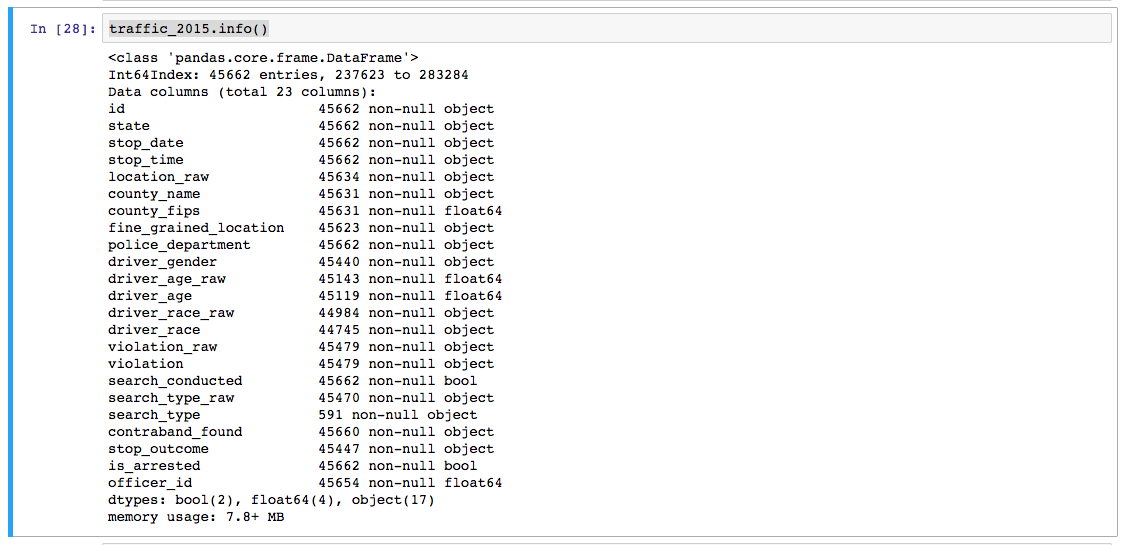


**What are the common problems found in the data?**

* Check Missing Data
* Remove Duplicate rows
* Detect Outliers using Data visualization
* Untidy data
* Drop rows with missing key data

Missing data and checking data types for all data

Using info() method on data frame we checked the summary to see whether any data is missing. As you see below the number of null value for “search\_type” is significantly high and rest all columns are populated more than 95% row.



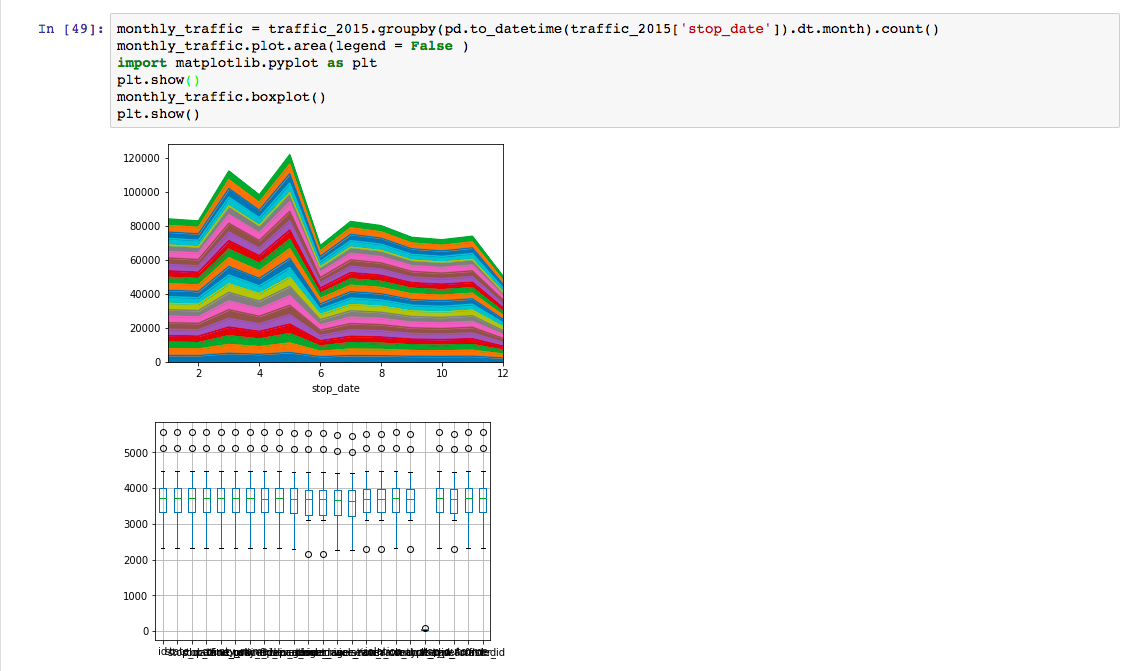
Remove Duplicate rows

As we can see there is a unique column “id” already exists in data and there are no duplicate rows to cleanup further.



Detect Outliers using Data visualization

Using area plot and box plot we checked the data population is consistent across the month (a given window) and the reported incident number of traffic stop in each month does looks fine. There are no outliers and there is no anomalies or no missing data as a big chunk .



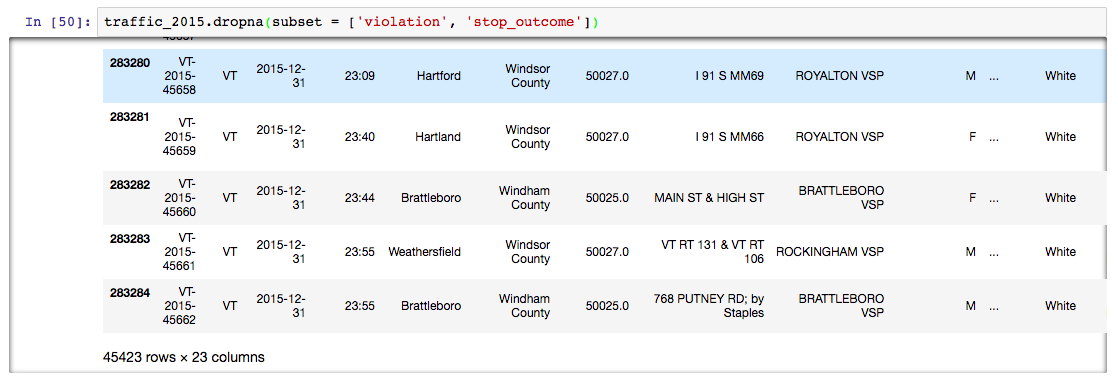
Tidy Data

* Columns represent separate variables
* Rows represent individual observations

Drop rows with missing key data

Drop records where the traffic stop reason or the stop outcome is unknown. These records will not be useful and any meaningful analysis.

As you can see below out of 45662 rows, 45423 rows were remained and 239 records were dropped.



Aggregate data for further analysis

Most of the reports/analysis will be done based on logical grouping of data by time

So aggregating data is useful.

* Aggregate by time of the day (slot of 4 hours)
* Aggregate by day of the week
* Aggregate by month
* Aggregate by age range (slot of 10 years)

We will use pivot table features for aggregating data based on these criteria for analysis.