Project: Capstone Project 1 – (Apply Inferential Statistics section 8.4)

Project name: - Explanatory Analysis of Traffic pullover pattern for Florida v/s Vermont

Student Name: - Jitendra Agarwal

Course: - Springboard cohort Jan2 2018

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

Data provider: - Open policing project by Stanford

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.

Downloaded excel sheet of raw data for VT: - https://github.com/jiagarwa/capstone-project1-Jitendra

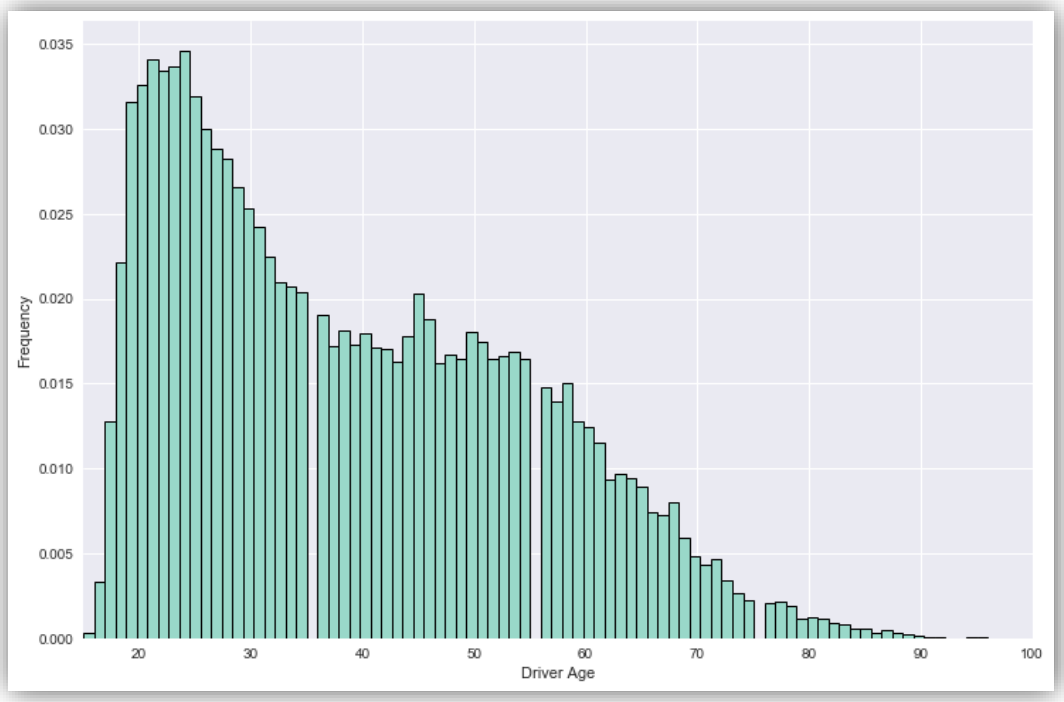
file name: - 'VT-clean.csv.gz'

Some Stats: -

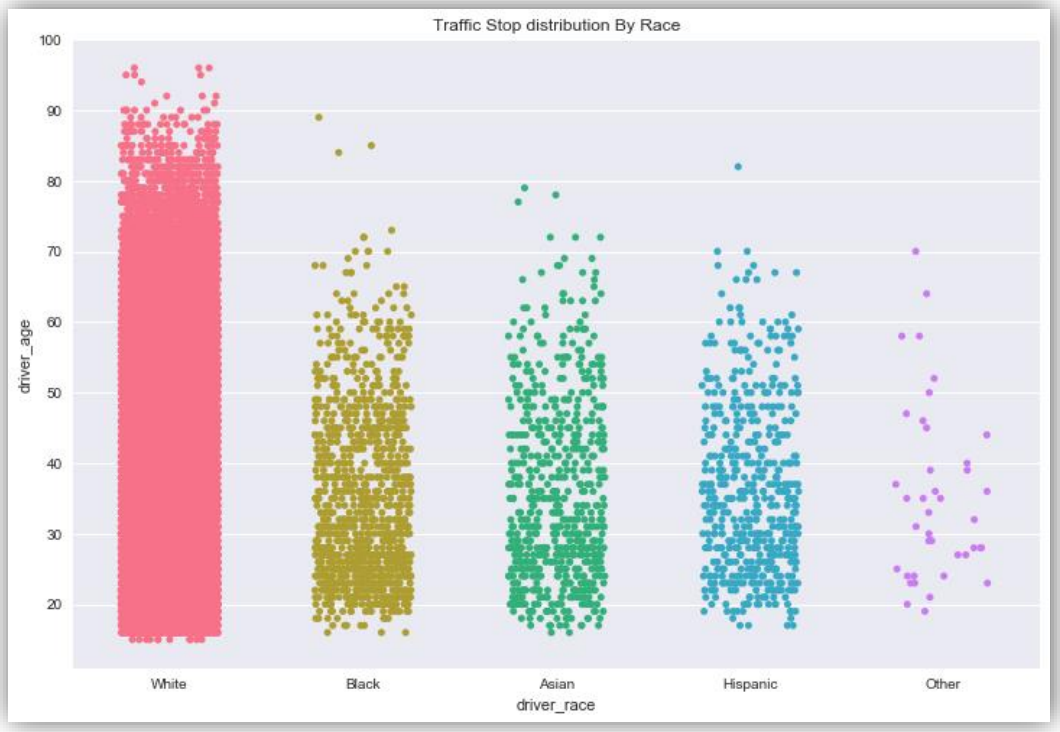
Number of Total Traffic Stops Analyzed: - 45662
Number of Traffic Stops for Minority Race: - 3213
Number of callbacks for Majority Race: - 42449
Number of Registered drivers with Age 16 and older as of end of 2013 in Vermont: - 543057

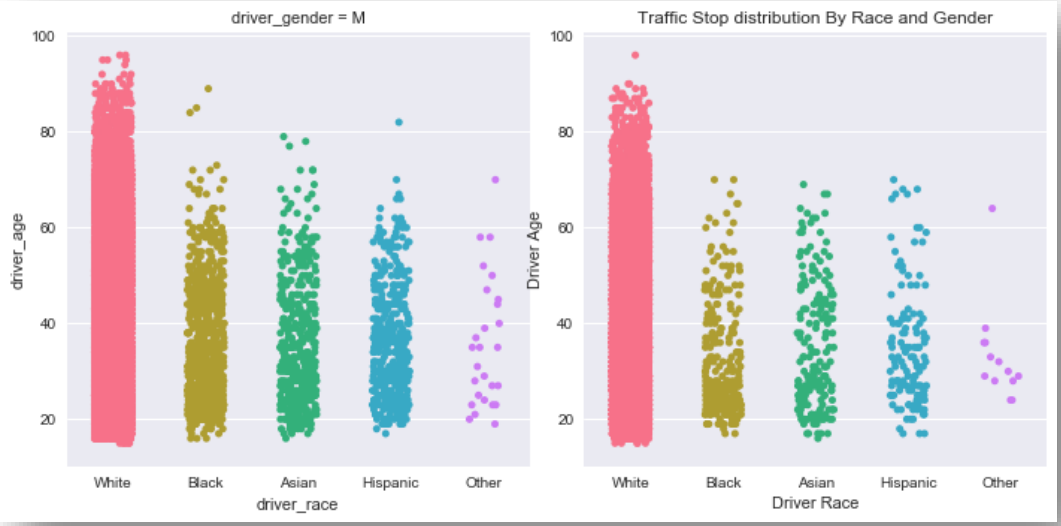
Graphical and Quantitative exploratory data analysis

1. **Histogram to see the frequency of occurrences in continuous data set of Driver age.**

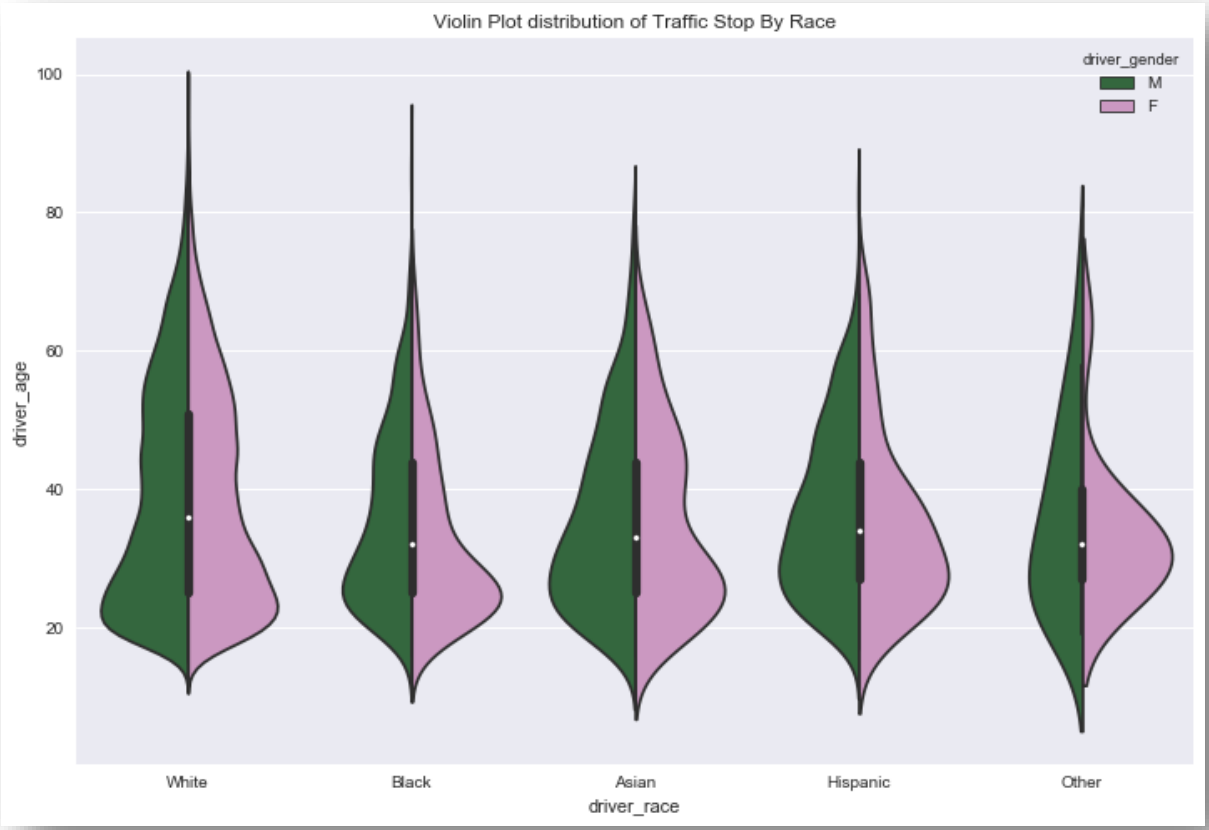


2. **Create Scatterplot and violin plot to see the Distribution of Traffic stop by age across Race, Gender.**





Violin plot of gender V/s the Age split by Race



Hypothesis Test: -

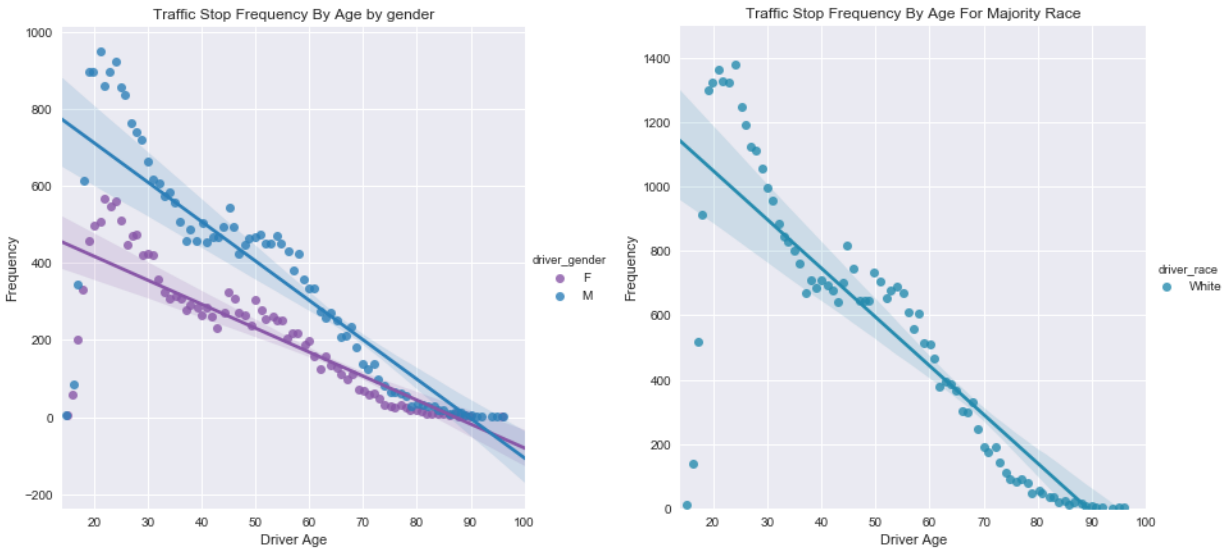
We do not have enough quantitative Parameters to do quantitative test for hypothesis that Gender and Race plays a significant role in traffic pullover.

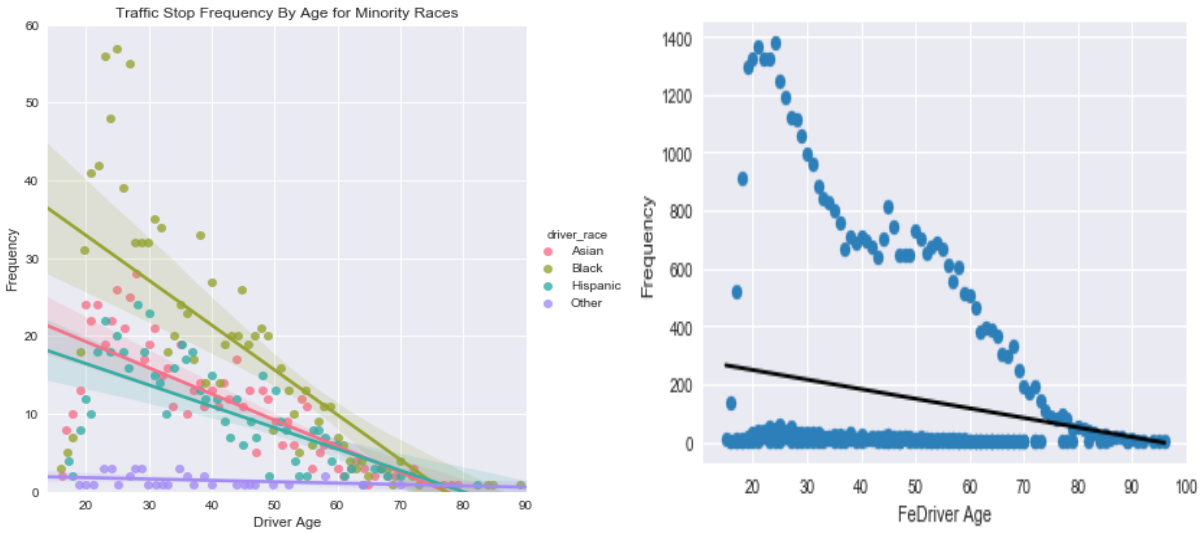
ML Model: - Logistic regression or Liner Regression for Classification

This is a classification problem as we are trying to predict the traffic stops and stop outcome.

We can see that most of the attributes/variable of data set are Categorical however age and time are continuous and Quantitative attribute.

We will first try to do a linear regression and then see if Logistic regression model is more suitable. Logistic regression is most suited for binary classification.





We can see that most Frequent Traffic stop involves the Male Driver of Majority Race and are under 25. We can see this visually who are stopped frequently.

```
Most common Traffic stop were made for User Profile as:

index                  81
driver race         White
driver gender           M
violation       Moving violation
stop outcome     Written Warning
is arrested          False
Frequency            11692


Highest % of Total Traffic Stops for this Top User Profile is at: 25.61
%
```

Mean and Median

we are going to do a comparison between the two proportions of sample by race (White and Non White). we shall also do a comparison between the multiple range of division by age.

According to the description for the collected data, every Traffic stop is different from other, therefore we may consider the samples are independent.

The sample size of each group (divided based on race) is above 30, hence we are safe to consider that CLT holds.

```
- Traffic stop %share for Majority race : 92.96 %
- Traffic stop %share for Minority race : 7.04 %
- Traffic Tickets %for Majority race : 37.75 %
- Traffic Tickets %for Minority race : 43.08 %
- Median age for Traffic stop for Majority race : 36.0
- Median age for Traffic stop for Minority race : 35.0
```

It is interesting to see if there is a 5% more chance to get a ticket after a traffic stop if Driver belong to Min ority race.

Evaluate the Lasso Model: -

Lasso Score is: -1.8953118073e-05

- Lasso score is too low and Negative, so it is possible that there is no significant variable/feature here to choose for correct prediction.
- The Lasso Model plot is entirely Flat. so Lasso model confirms that there is no significantly correlated feature to predict the Traffic stop outcome.

Evaluate the Liner Regression Model: -

The liner regression score is:- 0.0414843836904

- liner regression Score is too low, and as we already saw there is no significant relation with age in the traffic stop or its outcome. so it is possible that there is no significant variable/feature here to choose for correct prediction.

Logistic regression model

We can see that most of the attributes/variable of data set are Categorical. We need to create a Logistic regression model and then try to train and test the model using the subset of data sets.

- For logistic Regression model, we will try to predict Traffic stop outcome as Target variable we need all the feature columns to be in Boolean/binary for this we will use **One Hot Encoding** technique.

-Assumption:- We are only picking records with 2 possible outcome Citation and Written Warning to make our model simpler.

Tuned Logistic Regression Parameters: {'C': 0.43939705607607948}
Best score is 0.6322485207100592

Logistic Model matrix and classification

	precision	recall	f1-score	support
0	0.54	0.29	0.38	5068

1	0.65	0.84	0.74	7970
avg / total	0.61	0.63	0.60	13038

0.654051808825

A popular way of summarizing the discrimination ability of a model is to report the area under the ROC curve. We have seen that a model with discrimination ability has an ROC curve which goes closer to the top left hand corner of the plot, whereas a model with no discrimination ability has an ROC curve close to a 45 degree line. Thus the area under the curve ranges from 1, corresponding to perfect discrimination, to 0.5, corresponding to a model with no discrimination ability. The area under the ROC curve is also sometimes referred to as the c-statistic (c for concordance).

- Our Logistic regression model have the optimistic AUC of ROC as 65.41%