Name: Akhil Koppera
akhilkop@buffalo.edu

Project 1: Data Analysis

For this project Annual average daily Traffic counts data since 1977 of New York State from New York State database has been considered. Dataset has many missing values due to poor data collection in the early years. So, only data of year 2019 has been considered for this project as the original dataset excess data (2593803 entries) and missing values.

Part 1:

1. Dataset has 60321 entries with 17 variables

2. Type of data included:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2593803 entries, 0 to 2593802
Data columns (total 17 columns):
Year                      int64
Station ID                int64
County                    object
Signing                   object
State Route               object
County Road               object
Road Name                 object
Beginning Description     object
Ending Description        object
Municipality              object
Length                    int64
Functional Class          int64
Ramp                      object
Bridge                    object
Railroad Crossing         object
One Way                   object
Count                     float64
dtypes: float64(1), int64(4), object(12)
memory usage: 336.4+ MB
```

The description of each column can be found at https://data.ny.gov/Government-Finance/Annual-Population-Estimates-for-New-York-State-and/krt9-ym2k

3. We do have missing data:

```
data_2019.isnull().sum(axis=0)
```

```
Year                        0
Station ID                  0
County                      1
Signing                 52055
State Route             50980
County Road             48700
Road Name                5862
Beginning Description     199
Ending Description        318
Municipality              711
Length                      0
Functional Class            0
Ramp                    55271
Bridge                  45999
Railroad Crossing       57698
One Way                 51968
Count                       0
dtype: int64
```

Here columns like Ramp, Bridge, Railroad Crossing, One way has a lot of missing values but from the metadata/column description we can know that these columns have possibly two values, one of them is 'Y' and the other was just left as null (which got reflected as a missing value). It is similar to the case of column Signing. So Missing values of above five columns are not considered.

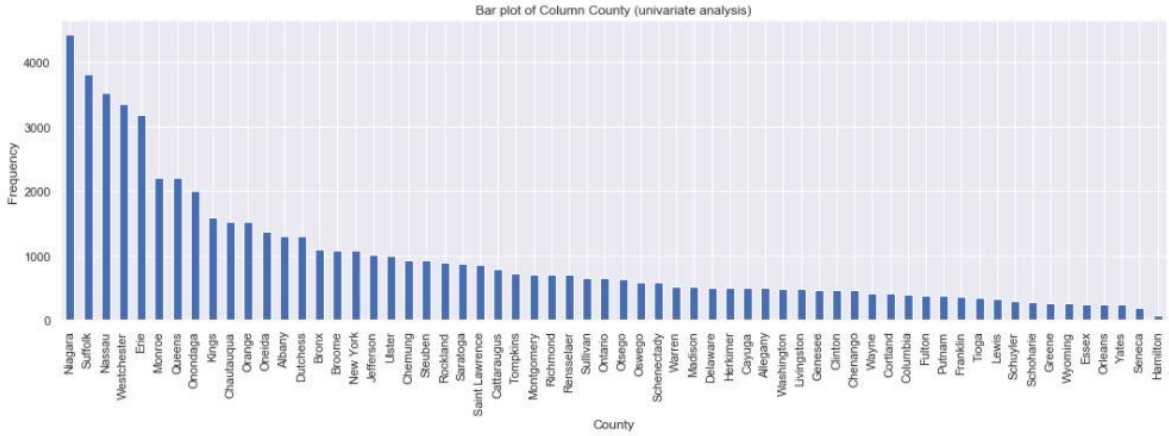
Ramp		Signifies by either a "Y" for yes or blank for no if a ramp is included in the roadway segment.	Plain Text	T
Data Type Text	API Field Name ramp			
Bridge		Signifies by either a "Y" for yes or blank for no if a bridge is included in the roadway segment.	Plain Text	T
Data Type Text	API Field Name bridge			
Railroad Crossing		Signifies by either a "Y" for yes or blank for no if a railroad crossing is included in the roadway segment.	Plain Text	T
Data Type Text	API Field Name rr_xing			
One Way		Signifies by either a "Y" for yes or blank for no if the roadway segment is a one-way direction.	Plain Text	T
Data Type Text	API Field Name oneway			
Signing		Designates what kind of a route or roadway the segment of roadway is. • Interstate: Interstate designation commonly known as the Interstate Highway System, a network of controlled-access highways that forms a part of the National Highway System. • US: US designation of a set of roads typically called U.S. Routes or U.S. Highways which forms an integrated network of roads and highways numbered within a nationwide grid in the United States. • NY: New York State Highway designation assigned at the State level. A blank field designates a 'Local' town, municipality or city roadway.	Plain Text	T

4. Main statistics about the entries of the dataset:

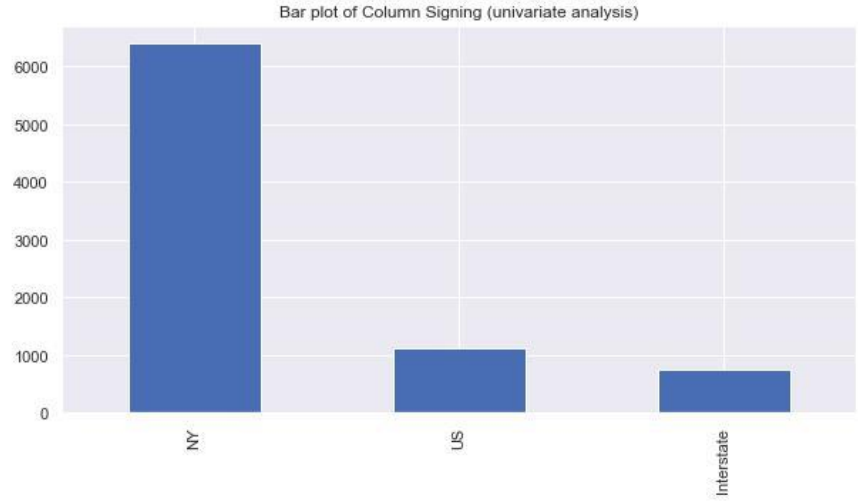
```
data_2019.describe()
```

	Year	Station ID	Length	Functional Class	Count
count	60321.0	60321.000000	60321.000000	60321.000000	60321.000000
mean	2019.0	438404.470218	86.561761	14.100612	5432.958157
std	0.0	308530.825040	124.785483	4.592381	13241.867994
min	2019.0	10001.000000	1.000000	1.000000	0.000000
25%	2019.0	115248.000000	18.000000	9.000000	290.000000
50%	2019.0	440047.000000	42.000000	16.000000	1468.000000
75%	2019.0	726064.000000	101.000000	19.000000	5520.000000
max	2019.0	978608.000000	1856.000000	19.000000	283686.000000

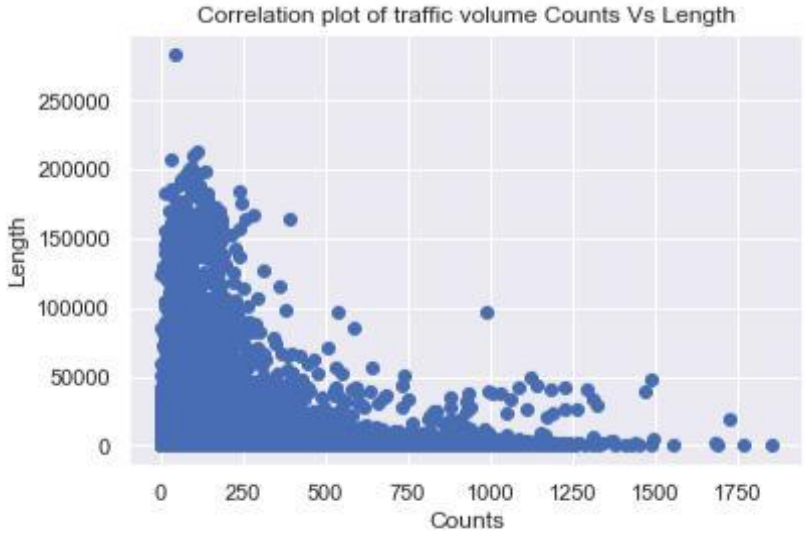
5. Visualizations of the data:



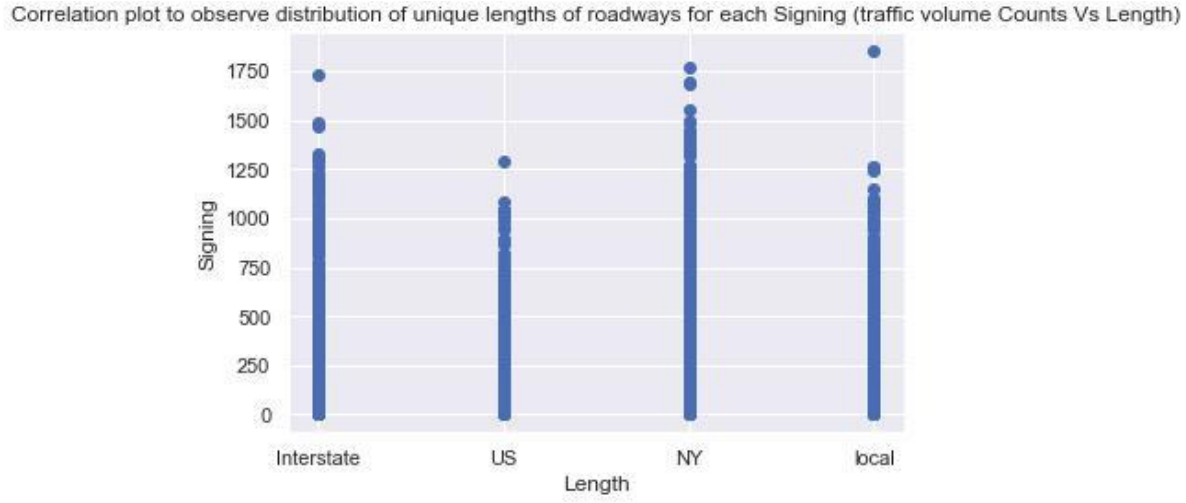
We can see that Niagara county has comparatively more traffic than the rest.



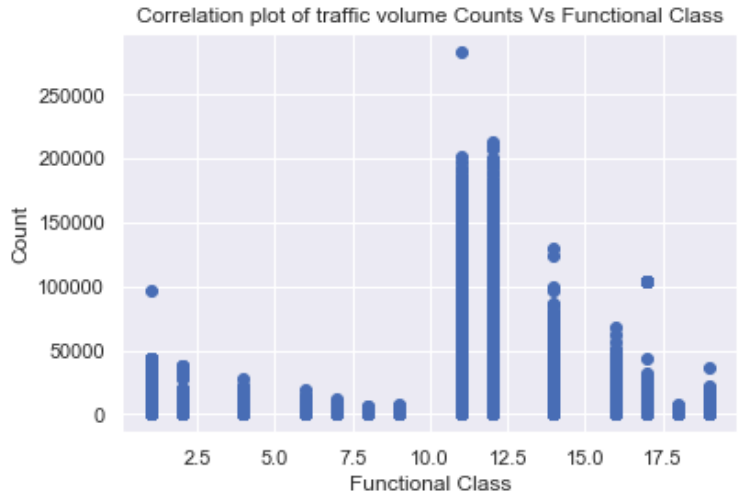
We can observe that New York State Highway designation assigned at the State level is comparatively very large than US and interstate designated highway system. But according to the metadata null's (which are 52055) correspond to local town, municipality or city roadway. So local roadway is much larger than even NY highway.

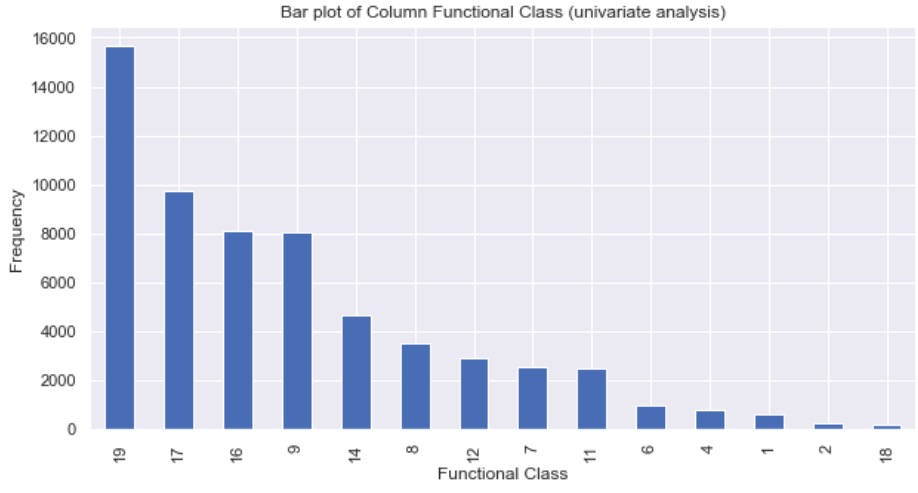


This implies that roads with shorter lengths are more likely to have heavy volumes of traffic.



We can observe that length of road for Interstate and NY highways are a little higher than that of US highways and local roadways





FUNCTIONAL CLASSIFICATION CODES	NYS Codes Urban	NYS Codes Rural	FHWA Codes
Principal Arterial - Interstate	11	01	1
Principal Arterial - Other Freeway/Expressway	12	02	2
Principal Arterial - Other	14	04	3
Minor Arterial	16	06	4
Major Collector	17	07	5
Minor Collector	18	08	6
Local	19	09	7

From the above univariate analysis of column Functional Class and metadata about the column, we can observe that Local NYS Codes Urban has comparatively high volumes of traffic. PFB the link for further details. https://www.dot.ny.gov/gisapps/functional-class-maps

Part 2:

1. Choose the features and targets in the dataset :

From the observations made out of data analysis and visualization and also taking the metadata into consideration, Below are the features and targets in the dataset:

Features:

1. County
2. Station ID
3. Signing
4. Municipality
5. Length
6. Functional Class
7. Ramp
8. Bridge
9. Railroad Crossing
10. One Way

Target/Response variable:

Count (Annual Average Daily Traffic volume value)

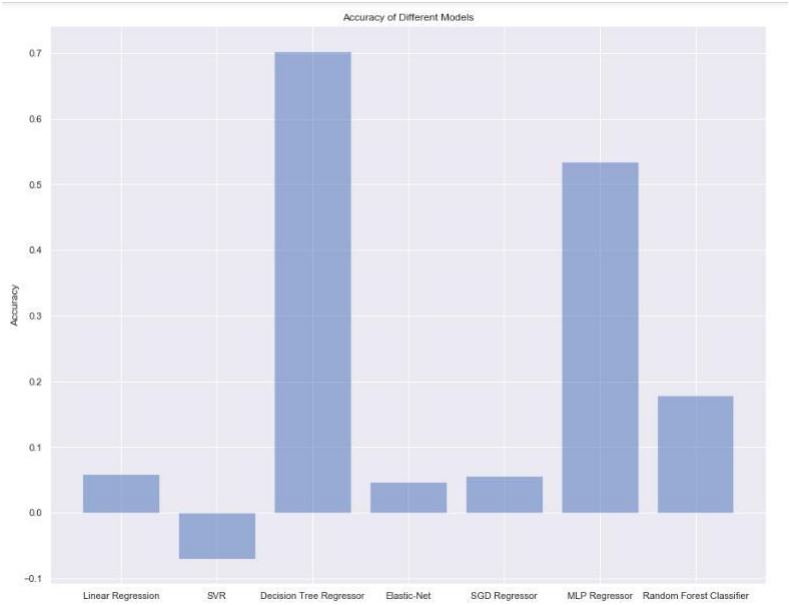
2. Data has been cleaned and missing values were imputed.

```
Null values in the data :
 Station ID           0
County                0
Signing               0
Municipality          0
Length                0
Functional Class      0
Ramp                  0
Bridge                0
Railroad Crossing     0
One Way               0
Count                 0
dtype: int64
```

This Pre-processed data is now split into train and test data.
But We can see that there are 651 unique values in 'Municipality' which have less than 20 individual entries. Splitting this will result in test data which is unreachable to the predictions of the trained model as the model which was trained will have no information about many of the labels in 'Municipality' column. So, to handle this case the entries with lesser unique values must be provided to training data rather than splitting.

3. Machine Learning Algorithms like Linear Regression', 'SVR', 'Decision Tree Regressor', 'Elastic-Net', 'SGD Regressor', 'MLP Regressor', 'Random Forest Classifier were used to model the target variable.

4. Comparison of results from each of the above ML model:



We can observe that Decision Tree Regressor works very efficiently in predicting the traffic counts for the test data compared to remaining models. And even MLP Regression does a decent job in predicting counts of test data with increase in number of iterations.

The type of data we trained is a hybrid data with both continuous and categorical variables. But now it is apparent that categorical variables like Ramp, Bridge, Railroad Crossing, One way and Signing in the data plays a crucial role than continuous variables like length.

Part 3:

1. Choosing and combining a related dataset to the existing traffic dataset.

Population dataset of New York State from United Census Bureau has been considered as related dataset. https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html

But this dataset has population of every year since 1970 till 2019 for every county in NY state but we do not need this data. We only need data which belong to year 2019 as our existing traffic dataset has the data for year 2019.

```
population_data_2019
```

	FIPS Code	Geography	Year	Program Type	Population
0	36000	New York State	2019	Postcensal Population Estimate	19453561
1	36001	Albany County	2019	Postcensal Population Estimate	305506
2	36003	Allegany County	2019	Postcensal Population Estimate	46091
3	36005	Bronx County	2019	Postcensal Population Estimate	1418207
4	36007	Broome County	2019	Postcensal Population Estimate	190488
...	...	...	...	...	...
58	36115	Washington County	2019	Postcensal Population Estimate	61204
59	36117	Wayne County	2019	Postcensal Population Estimate	89918
60	36119	Westchester County	2019	Postcensal Population Estimate	967506
61	36121	Wyoming County	2019	Postcensal Population Estimate	39859
62	36123	Yates County	2019	Postcensal Population Estimate	24913

63 rows × 5 columns

County wise population for year 2019.

New column 'County' has been added to our existing traffic dataset in order to make it compatible to the population dataset.
And had dropped the entry with 'Geography' = New York State from population dataset as we do not need it at all. Changing column name from Geography to County in order to make it compatible.
Before combining them, we need to make them compatible as each county in traffic dataset is sub divided into many small traffic stations based on many factors. So we need to group the dataset by County and combine the traffic counts and any other meaningful variables which makes sense when compressed.

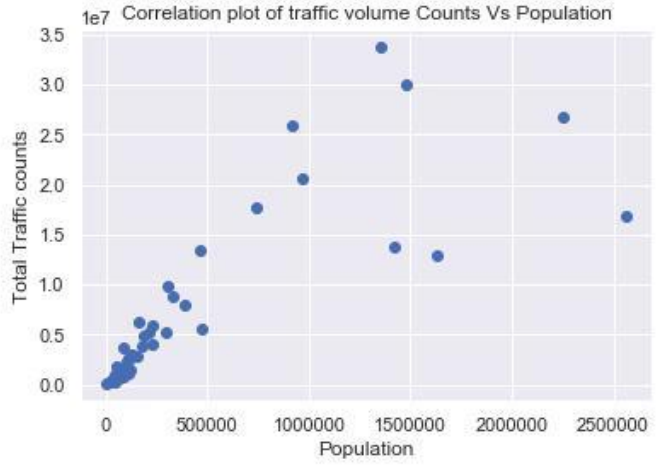
A discrepancy has been noticed from both traffic and population datasets as the combined dataset is giving 61 rows instead of 62 rows(expected): i.e., St. Lawrence County in population_data_2019 and Saint Lawrence County from traffic dataset. So, we need to rename one of these.

Combined dataset has 62 entries each representing each County with its respective traffic volume, Length of roads and Population.

2. Choosing correlated variables from both datasets:

Variables like 'Count', 'Length' from traffic dataset may be correlated to 'population' variable from population dataset.

3. Performing Statistical analysis on finding the correlation between selected features from both datasets:



Total traffic counts are seen increased with an increase in population in the county.

4. Analysis from the results and few interesting patterns:

1. We see that Total traffic counts increasing with an increase in population in the county

2. From the above County wise bar plot of population and Traffic counts, we can see that Except in Kings County Traffic volumes are very much correlated to Population of the county. More the population, more is the Traffic volume.

3. Even Pie chart supports the same above point with respect to the correlation between population and traffic counts. We see Kings county not complying with the above correlation, having less traffic counts in spite of high population.

References:

https://data.ny.gov/Transportation/Annual-Average-Daily-Traffic-AADT-Beginning-1977/6amx-2pbv

https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html

https://data.ny.gov/Government-Finance/Annual-Population-Estimates-for-New-York-State-and/krt9-ym2k/data