For the capstone project I, I am using the data provided by Montgomery county available in CSV format. This data contains the traffic violation occurred in the Montgomery county. I used pandas *read_csv* method to load the data as DataFrame. For ease of analysis, converted the data to a time series DataFrame.

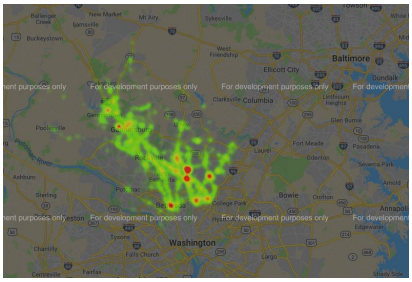
```
violation_df.info()

<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 1384236 entries, 2013-09-24 17:11:00 to 2018-09-27 0
Data columns (total 34 columns):
Date Of Stop_Time Of Stop     1384236 non-null datetime64[ns]
Agency                        1384236 non-null object
SubAgency                     1384226 non-null object
Description                   1384227 non-null object
Location                      1384234 non-null object
Latitude                      1283334 non-null float64
Longitude                     1283334 non-null float64
Accident                      1384236 non-null object
Belts                         1384236 non-null object
Personal Injury               1384236 non-null object
Property Damage               1384236 non-null object
Fatal                         1384236 non-null object
Commercial License            1384236 non-null object
HAZMAT                        1384236 non-null object
Commercial Vehicle            1384236 non-null object
Alcohol                       1384236 non-null object
Work Zone                     1384236 non-null object
State                         1384177 non-null object
VehicleType                   1384236 non-null object
Year                          1375769 non-null float64
Make                          1384179 non-null object
Model                         1384046 non-null object
Color                         1367433 non-null object
Violation Type                1384236 non-null object
Charge                        1384236 non-null object
Article                       1314986 non-null object
Contributed To Accident       1384236 non-null object
Race                          1384236 non-null object
Gender                        1384236 non-null object
Driver City                   1383956 non-null object
Driver State                  1384225 non-null object
DL State                      1383307 non-null object
Arrest Type                   1384236 non-null object
Geolocation                   1283334 non-null object
dtypes: datetime64[ns](1), float64(3), object(30)
memory usage: 369.6+ MB
```

Missing values: The data contains some missing values. However, the fraction of data having missing values is small (~18,000) compared to the large data (~ 1 million). We can drop the missing data without a substantial loss of statistical power.

Outliers: Most of the columns in the data are categorical variables. For each column, I checked the possible list of categorical variables using *DataFrame['column'].unique()* method. There are no outliers in the data.

For the non-categorical columns, we performed an EDA test to check if there are any outliers. An example is the plot below that shows a heat map of the traffic violation in Montgomery county.



Wrangling data: The data contains three types of variables: Categorical, Numeric and text. We should deal with these three separately.

For the categorical variables, we changed the variables into mainly two categories. For example, the *race* column contains five categories *WHITE*, *BLACK, HISPANIC, NATIVE AMERICAN, ASIAN, OTHER,* we changed it into two categories containing *WHITE* vs *non-WHITE.* Then we changed the categorical variables to a binary variable (1 or 0) using **pandas** *get_dummies* method.

We keep the numeric variables in the current form. We will use Normalization option later when we deal with Machine Learning.

For the text data, we converted the data into Vector form using *CountVectorize* tool in python. This method finds the most relevant words in the columns. These relevant words represent the basis of a Vector. Thus each element in the column can be written as a linear combination of basis Vectors.

In the end, we combined all the data into a single DataFrame.