NYC Parking Ticket Data Analysis and Prediction

Arpitha Narasinh Alur, Ron Abraham Cherian, Saketh Gourishetty, Shama Zabeen Shaik, Sneha Ramesh Watharkar, Snigdha Bose

Department of Computer Science  
University of North Carolina CharlotteCharlotte, United States

[aalur1@uncc.edu](mailto:aalur1@uncc.edu), [rcherian@uncc.edu](mailto:rcherian@uncc.edu), [sgourise@uncc.edu](mailto:sgourise@uncc.edu), [sshaik3@uncc.edu](mailto:sshaik3@uncc.edu), [swathark@uncc.edu](mailto:swathark@uncc.edu), [sbose1@uncc.edu](mailto:sbose1@uncc.edu).

**Abstract:**

**Big Data projects deal with large volumes of data. It is a booming field where we study the data and the insights drawn help in planning and strategizing business moves.**

**As we know, New York is the city that never sleeps. The city attracts high influx of visitors and has high traffic concentration. This makes the city susceptible to high rates of Parking violations. The city is still making amendments and face-lifting the current rules and policies to better monitor and control the number of violations.**

**In this project, we align Big data analysis and techniques to create a holistic and realistic overview of the New York city parking violation scenarios based from insights derived from the trends demonstrated by the existing data.**

**We have employed an austere approach and recruited effective Big data analysis techniques maintaining a focus on the bottom line problem domain. In this report, we discuss about our approach, techniques employed and results derived closely tied back to our goals. We have further discussed about our findings and mapped them to strategic solutions and recommendations that will help the New York police department conjure effective rules and policies.**

**I. Introduction and Motivation**

New York, being the city that never sleeps and has way too many visitors from all over the world coming over to look at the places in and around it. With tons of people travelling and the sizes of the parking lots being relatively small to accommodate all the visiting and local vehicles, the ticket issuing system has been enforced in and around the busiest areas of New York. Our project deals with one of the datasets from the ticket issuing systems of the New York and we would be aiming at answering few of the common questions people might have with respect to this scenario.

We have devised effective hypotheses that have eventually aided in compelling findings and synthesis of resultant recommendations that could be employed to better monitor the parking violations in and around the city. We have instantiated our analysis on being able to derive insights from the data collected by the NYC Department of Finance over the last few years and to facilitate strategic solutions based upon the nature of violations, the most susceptible time and places of violations.

The insights derived from our research can be effectively employed by the New York police department to monitor and control the number of instances of parking violations as they are derived from real violation records and trend analysis. We have created our hypotheses closely tied back to this goal.

**II. Data Description**

Data is collected by the NYC Department of Finance each year from 07-01 to 06-30. We have taken the dataset from Kaggle website. A total of 42.3 million rows of data is available spread across the time frame of August 2013 to June 2017. We have chosen the 2016 dataset of the complete dataset because of the usage restrictions on data size in XLMiner tool.

**III. Variable Selection**

The complete dataset has table metrics of 9100278 rows and 51 columns. Each column is a feature of our dataset.

There were a few columns which had no value for 99% of the records so those variables(features) have not been considered for this project. The features which have been considered are as follows:

**i. REGISTRATION STATE**

This variable indicates where the vehicle is registered. It tells which state the vehicle belonged.

**ii. PLATE TYPE**

This variable indicates the category of the vehicle. If the vehicle is a commercial vehicle or a passenger vehicle and so on.

**iii. ISSUE DATE**

This variable indicates when the ticket was issued.

**iv. VIOLATION CODE**

This variable tells us about the violation caused by the person to whom the ticket was issued. There are several kinds of violations associated with parking. So, each violation is given a code which acts as its identifier. The details of the violation codes has been derived from the government violation code website. [1]

**v. VEHICLE BODY TYPE**

This variable indicates the type of vehicle. The vehicle shape and size for example. Sedan, hatchback, SUV and so on.

**vi. VEHICLE MAKE**

This variable tells about the manufacturing company of the vehicle.

**vii. VIOLATION LOCATION**

This variable describes about the location in New York where the violation was made.

**viii. VIOLATION TIME**

Time at which the violation has occurred.

**ix. VIOLATION COUNTY**

New York has five counties. New York County (Manhattan), Kings County (Brooklyn), Bronx County (The Bronx), Richmond County (Staten Island), and Queens County (Queens). This variable gives information about the county where the violation is made.

**x. STREET NAME**

Street name where the violation has been made is given by this variable.

**xi. VEHICLE COLOR**

Color of the vehicle is described by this variable.

**xii. VEHICLE YEAR**

This variable tells which year the vehicle was manufactured.

**xiii. VIOLATION IN FRONT OF OR OPPOSITE TO**

This variable tells if the violation is made is front of or opposite to the location that is mentioned.

**IV. Data Preprocessing**

An exploratory data analysis was conducted to generate an overview insight over the data set using Python. This helped explore a summary of insights generated by the same, to lay a foundation for further study.

The data initially had 42.3 million records. We have randomly selected 65,000 records and started with pre-processing. In the data preprocessing stage, making use of python script, we have first dropped the columns that had all missing values, then deleted the rows with missing values and dropped the rows that had unrelated values. This preprocessing step has left us with around 27,000 records, which we fed into the Excel Miner for further stages of analysis. We binned the data as per requirements and partitioned it into training, testing and validation datasets which we used for the further stages of analysis and prediction.

We started with exploratory data analysis (EDA) using python, to find out the possible solutions of the hypotheses questions. The EDA has left us with results as follows:

* Maximum violations on Mondays’ in K county.
* Passenger plate types are causing maximum violations on Mondays’.
* Overall maximum violations were observed in the month of June in county K.
* Violation code 21 was made the most by the passenger plate vehicles.
* Maximum violations were observed in county K during mornings.
* Maximum violations are being caused by passenger plate types during morning.

To understand the insights of the hypotheses better, we

have applied certain machine learning algorithms on the

pre-processed data using XLMiner.

Python was used to process the data and the resultant data in

JSON format was converted to csv format using a delimiter to be able to use the same with XLMiner. We reduced the dataset to number of approximately 35000 rows.

**V. Analysis of modeling techniques**

For our dataset, we have chosen the following data modeling techniques to analyze the data and come up with the one that best suites our dataset and hypothesis.

* Prediction (Classification Trees, Regression)
* Association Rule Mining
* Clustering
* Classification

**Clustering**

Clustering is an unsupervised learning technique that involves grouping of data based on their characteristics and aggregating them based on similarities. It involves segregating data points with similar traits into groups and assigning them to clusters. Clustering technique is best suitable for recommendation systems, Anomaly detection system, Market segmentation and many more. Clustering is mostly suitable for continuous data.

There are different types of clustering methods based on the way the clusters are formed and how the data points are grouped. Two of the well-known clustering methods are K-means clustering and Hierarchical clustering. For our dataset, clustering didn’t yield efficient results because our dataset is mostly categorical and we have discrete data but clustering works well for continuous and numerical data.

**Prediction**

Predictive Analytics is a process of using data analytics to make predictions based on data. Predictive modeling is used to predict future results from the historical data. Regression analysis (Multiple Linear Regression), k-Nearest Neighbors, Regression Tree (Random trees), Neural Networks are some of the well-known predictive analysis methods. We have tried applying Multiple Linear Regression on our dataset.

**Classification**

k-Nearest Neighbors: KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry.

KNN can be used for classification — the output is a class membership (predicts a class — a discrete value). An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. It can also be used for regression — output is the value for the object (predicts continuous values). This value is the average (or median) of the values of its k nearest neighbors.

**VI. Research hypothesis and analysis**

***Hypothesis 1***

*Which week and county are the violations mostly occurring*

An Exploratory Data Analysis was conducted to generate an overview insight over the data set using Python. This helped explore a summary of insights generated by the same, to lay a foundation for further study.

Python was used to process the data and the resultant data in JSON format was converted to csv format using a delimiter to be able to use the same with XLMiner. Additionally, the “Day of Week” variable was generated using Excel formula and the “Date “column as input which was in date/time format. We created Dummies for the “Day of Week”.

In the next stages of the analysis we tried options including:

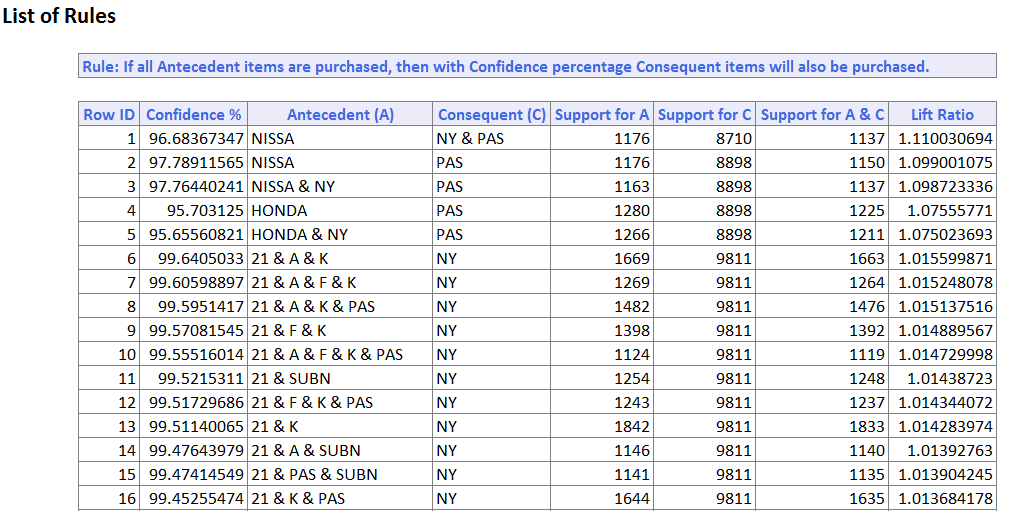
Clustering: Using Clustering to analyze the given data set failed primarily because the clustering results were similar to the frequency analysis. Also, we cannot combine the requisite two different hypothesis to arrive on a result. This was because not all variables were getting clustered effectively.

**Decision Tree**

Using decision tree yielded resultant model with high error rates ranging from 72.7% to 100% with a resultant error rate of approx. 55%. Due to the high error rates the results were not accurate and could not be used to synthesize results.

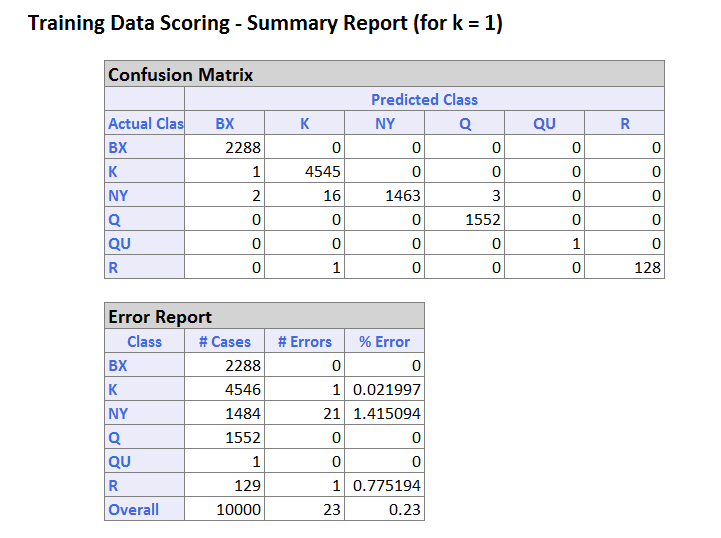
We finally used the Naïve Bayes approach to construct classifiers and to generate a model to assign class labels. This yielded Lower error rates as compared to previous approaches and accurate results.

**Association Rule Mining:**



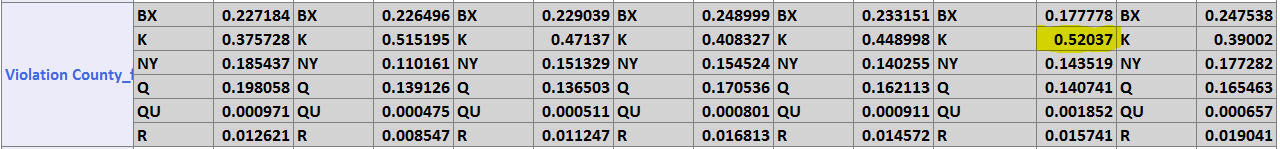
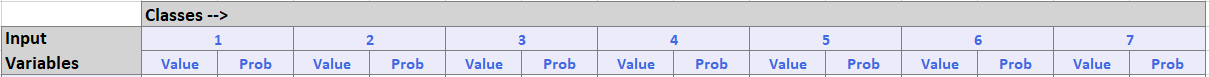
**Figure 1: Association Rule Mining**

**KNN Classification**

****

**Figure 2: KNN Classification Results**

**Naive Bayes**



**Figure 3: Conditional Probabilities (Naive Bayes)**

We can see that the conditional probability output table shows that in day of the week 6 there are more violations observed in county K.

We can see from the above results from the training dataset week of the day 2 has the least error rate, and from the confusion matrix its seen that the misclassified classes are least for this week of the day 2 Monday.

**Hypothesis conclusion drawn**

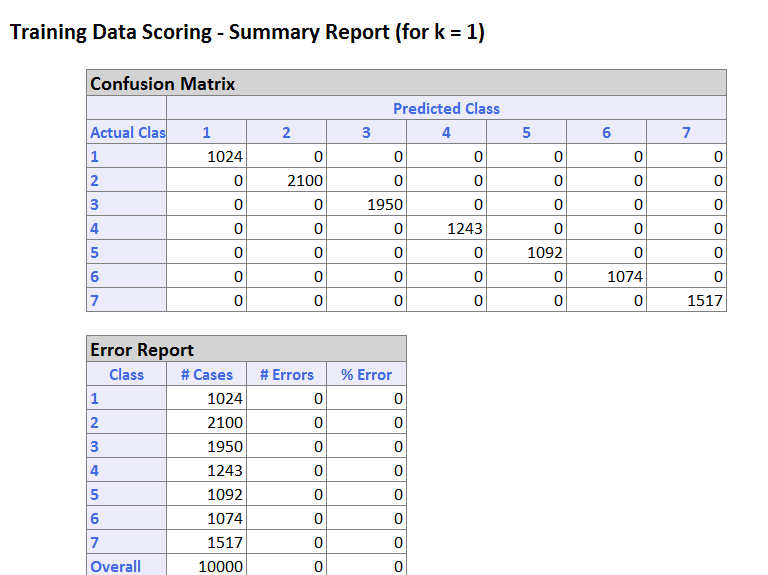
Maximum number of Violations are seen in K- county and on Fridays. These are the screenshots of the results obtained.

***Hypothesis 2***

*Which plate type caused most violations and in which week of the day?*

**KNN Classification**

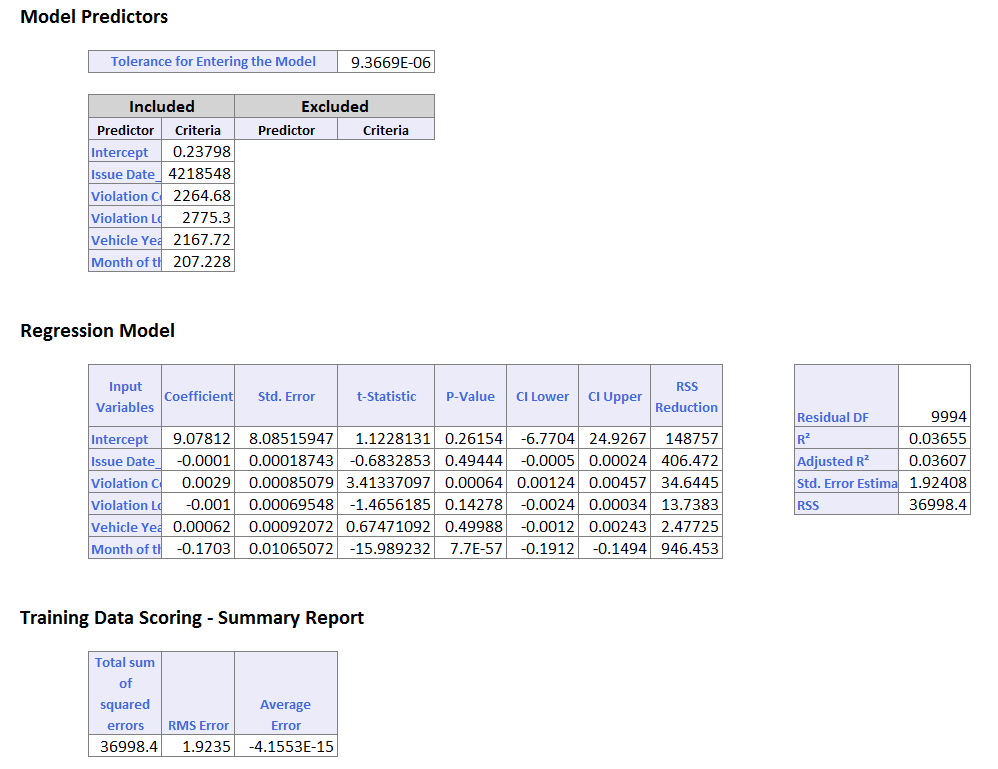
In this hypothesis, we tried using classification trees, that is the Random forests and Regression modeling, but these two approaches did not yield satisfactory results as compared to the Naive Bayes classifier applied which yields better results.

****

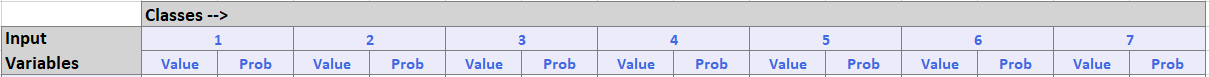
**Figure 4: KNN Report**

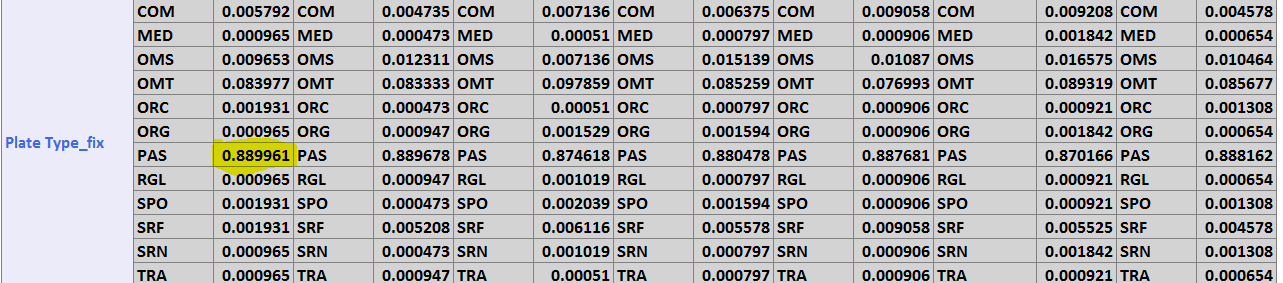
Applying Naive bayes classifier yields following results which answers our hypothesis in which county more violations are observed.

**Multi Linear Regression**

****

**Figure 5: Multi Linear Regression**





**Figure 6: Conditional Probabilities (Naive Bayes)**

Maximum number of Violations are seen in K- county and on sixth month of the year which is June.

***Hypothesis 3***

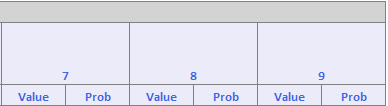
*Which month and which County are most violations observed?*

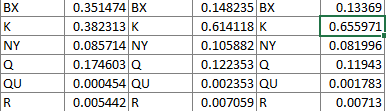
**KNN Classification**

To yield usable results we created Dummies for the “Plate type” variable and further Binned the resultant data.

**Naive Bayes**

In the next stage of the analysis we tried using Classification Tree. This model showed high error rates. This led us to using the Naïve Bayes approach which gave Lower error rates as compared to the previous approach and accurate results.





**Figure 7: Naive Bayes Result**

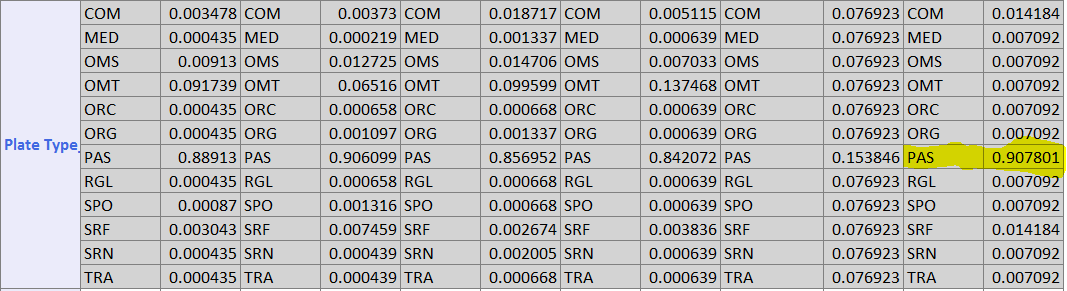
We can see that K county has most violations and in the month of June.

***Hypothesis 4***

*Which plate type is causing most violations in a county?*

**Naive Bayes**

To yield usable results we created Dummies for the “Plate type” variable and further Binned the resultant data.

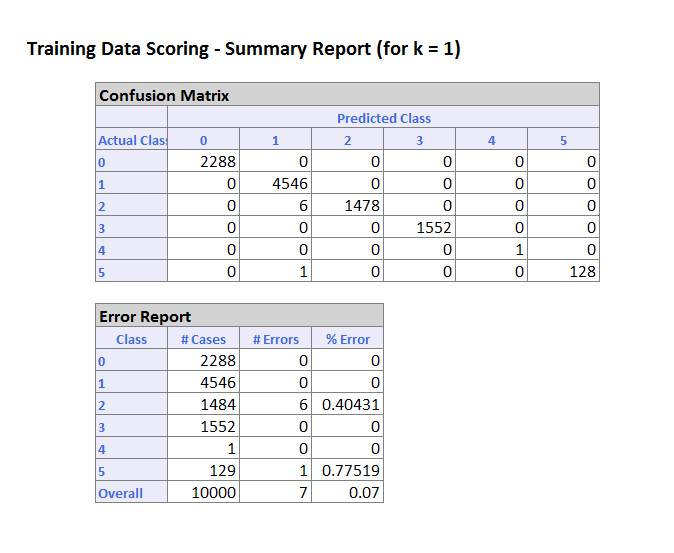
In the next stage of the analysis we tried using Classification Tree. This model showed high error rates. This led us to using the Naïve Bayes approach which gave Lower error rates as compared to the previous approach and accurate result

**Figure 8: Naive Bayes Output**

Passenger Type- Pass were causing most in County-k.

***Hypothesis 5***

Which County Has Maximum occurring violations and during what time of the day are these violations observed. To yield usable results we used the earlier Models developed for Hypothesis 1 through 5.

**KNN Classification**

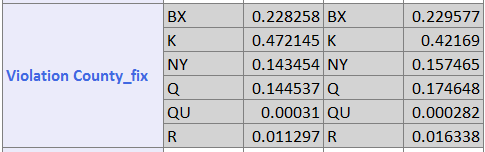
**Figure 9: KNN Result**

**Naive Bayes**

This led us to using the earlier Naïve Bayes model which

gave Lower error rates as compared to the previous approach

and accurate results.



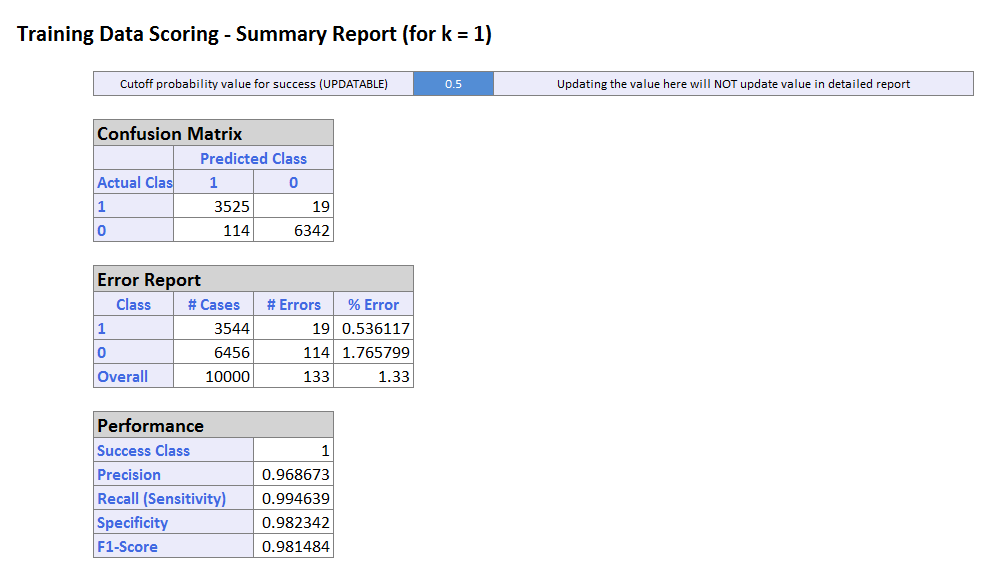
***Figure 10: Naive Bayes Result***

Hypothesis conclusions drawn: Violations are most likely to occur in the Morning in County-K.

***Hypothesis 6***

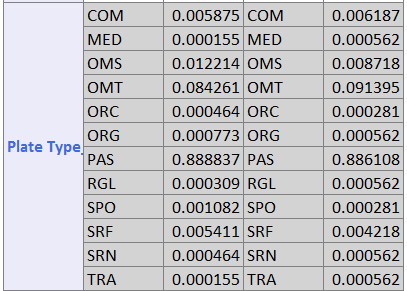
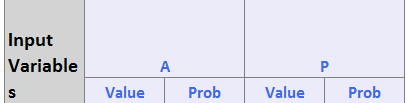
Data analysis to draw insight on When Violations are mostly occurring and are caused by which Plate type.

**KNN Classification**

****

**Figure 11: KNN Classification**

**Naive Bayes**

To yield usable results we used the earlier Models developed for Hypothesis 1 through 5.

**Figure 13:Output of Naive Bayes**

This led us to using the earlier Naïve Bayes model which gave Lower error rates as compared to the previous approach and accurate results.

**Hypothesis conclusions drawn**

Overall, we are able to predict that **Kings County** (**Brooklyn**) is the area most prone to violations especially on Friday, Monday and in the months of **September, August. And it was also observed that Passenger plate types are the ones causing most violations and during the day time(AM).** The NYC police department should do something so that at these times at these places violations can be reduced.

**VII. Modeling**

We started with*Association Rule Mining***,** but the rules generated from this method, were not well aligned with our hypothesis**.** So, we choose the second method which was hypothesis dependent. *KNN, Random tree, Multilinear Regression* were the methods that were generated separately for the respective hypothesis but they too weren’t giving out the results with good precision and accuracy. We then choose **Naive Bayes Classification**, which helped us in deriving the solutions for all the hypotheses with better accuracy and precision.

**Modelling Summary**

We have performed various Machine Learning algorithms on our original dataset in order to identify various relationships among our features and gain insights on the entire dataset and check whether there is considerable difference between our visualizations and our model. The Naive Bayes Classification Algorithm was able to give out a lot of important information to us. It proved that Kings County is most prone to violations, that too on the month of June. Fridays are the days the most

number of violations have occurred. Passenger Plate types have caused the most number of violations on Fridays. Relationship between violation county and plate type. Time of the day - Morning violations are occurring more and win which county Relationship between plate type and time of violations.

**VIII. Recommendations**

We recommend that the policy makers should change the trivial violations and recommend police force to patrol more in places like K county during mornings on Fridays and Mondays to help the passengers. We also recommend installing traffic blackspots. The trailing are the recommendations mapped from our hypotheses conclusions that would facilitate strategic solutions devised from the data and trend analysis:

1. **Help the NYC police department modify the current rules-** The policy makers should consider making amendments to the existing trivial violation rules and control measures. These rules and policies need modification and a complete overhaul based on more realistic trends as shown by our analysis.
2. **Regulate the laws to reduce the number of violations-** As per our findings places like Kings County (Brooklyn) and during the Morning hours, shows the greatest number of violations. Fridays show the worse trends. These findings on the most susceptible places and times, can ease the synthesis of planning and monitoring in accordance to the trends in these areas during these days/times. The department can be more alert and observant during this time at such places, to exercise more control on violations in these areas.
3. **Enforce more patrols based on insights collected-** According to our hypothesis results, the K-County shows the most violations during the Mornings and mostly on the Monday and Friday. This analysis can be mapped to devise better patrolling plans in these areas/times, that show high trends of violations. The monitoring in these areas and times should be more effective.
4. **Create more parking spaces-** Our research has set us to recommend installing more parking spaces and regulatory stations, in areas that show heavy influx of traffic and peaks in the number of violations.
5. **Clear signs while shut down/cleaning-** Clear feedback and signs to re-direct the traffic should facilitate effective control and regulate the number of violations caused due to issues around parking while work is in progress.

**IX. Lessons Learned**

The Naïve Bayes’ algorithm yields best results for categorical data in comparison to the performance of other modeling techniques.

Most of the Output Variables in the given Traffic Violations – Dataset are Categorical. For this reason, The Naïve Bayes algorithm works most effectively for us.

# **X. Results and Discussions**

Based on our Model, for a given County, Vehicle type and Time of a day, we can predict – If a person is causing a violation or not.

Overall, we are able to predict that **Kings County** (**Brooklyn**) is the area most prone to violations especially in the months of June, October and September. Our model also suggests that Friday’s are the worst days where most traffic violations occur. The NYC police department should do something so that at these times at he places violations can be reduced.

# **DISCUSSION**

This dataset since it doesn’t have the decision variable included [2], is not a perfect fit for any of the machine learning modelling. It is a very good dataset for data visualization and data analytics. We have justified our part of trying to apply different modelling and machine learning techniques but were not able to get the results with desired accuracy and precision.

# **Further Work**

We could do data analytics on the other years of data and come up with better analytics [3] which would help solve real time problems faced by the vehicle users and the traffic management of the New York traffic department.

# **Acknowledgement**

Our Thanks to the Professor Dr. Li-Shiang Tsay and both the TA in helping us derive results to meet our analytics objectives. We want to acknowledge the Kaggle - Team for the provision of the Data set that propelled our results and analysis generation.

##### **IV. REFERENCES**

[1] Course Material: ITCS 6100-051Big Data Analytics for Competitive Advantage

[2] G. Shmueli, Peter C. Bruce, and Nitin R. Patel, Data Mining for Business analytics, 3rd Edition, 2016, ISBN: 978-1-118-72927-4.

[3] J. Leskovec, A. Rajaraman, and J. D. Ullman, Mining of Massive Datasets, 2nd, Cambridge University Press, 2014

**Web References:**

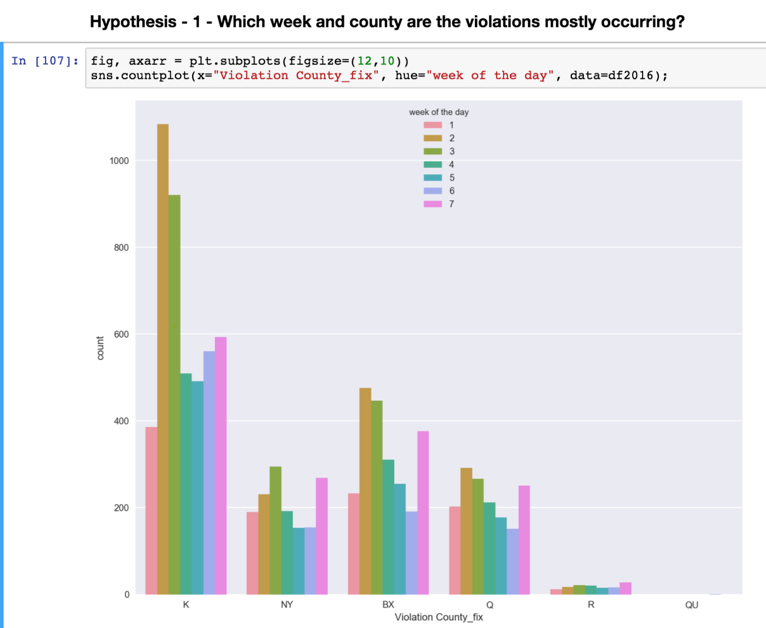
<https://www.solver.com/xlminer/help/classification-using-naive-bayes-example>

[https://seaborn.pydata.org/generated/seaborn.countplot.html](https://www.google.com/url?q=https://seaborn.pydata.org/generated/seaborn.countplot.html&sa=D&source=hangouts&ust=1524105330243000&usg=AFQjCNE9vtTzOnvDi3xmU8uyZe_RvMKKyQ)

**APPENDIX**

1. **Data Visualizations**

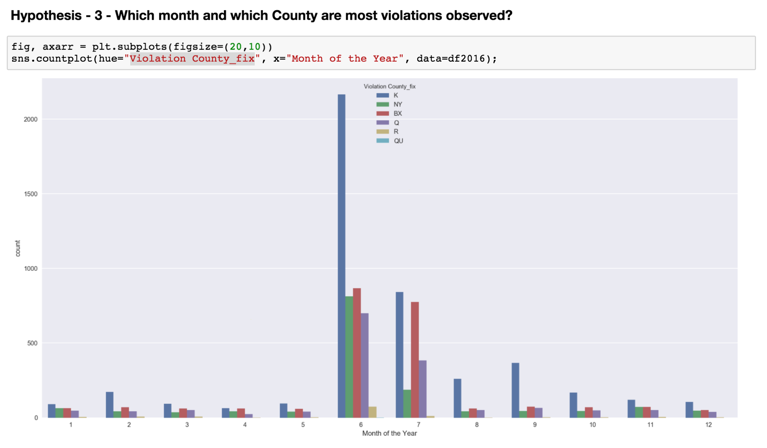
Data visualization is the main theme on the data mining. We have performed various Data Visualizations on our original dataset in order to identify various relationships among our features and gain insights on the entire dataset. The visualizations proved that Kings County is most prone to violations, that too on the month of June. Fridays are the days the most number of violations have occurred. Passenger Plate types have caused the most number of violations on Fridays. Relationship between violation county and plate type. Time of the day - Morning violations are occurring more and win which county Relationship between plate type and time of violations.

****

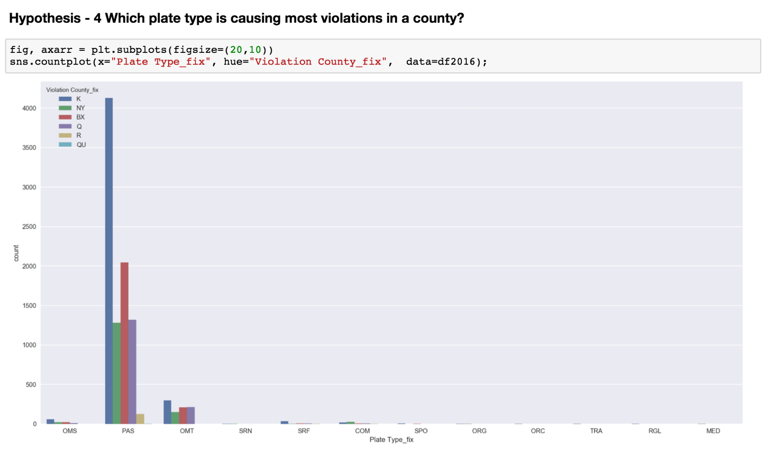
**Figure 1: Hypothesis 1 Visualization**



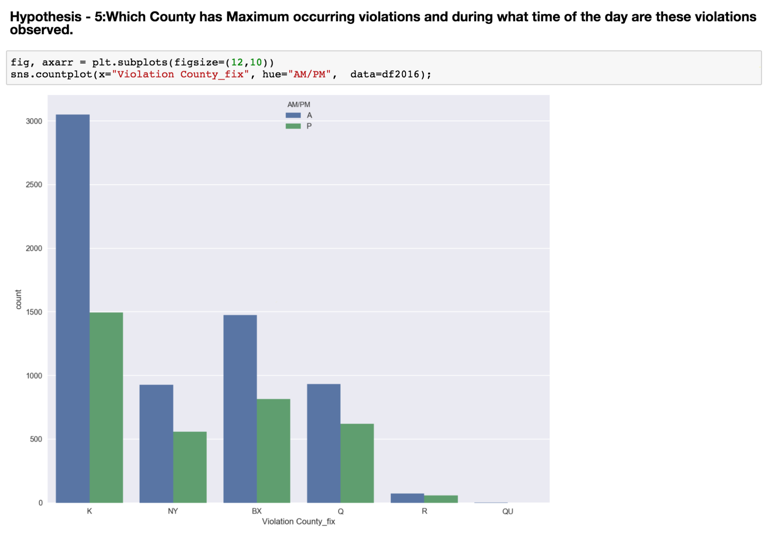
**Figure 2: Hypotheses 2 Visualization**



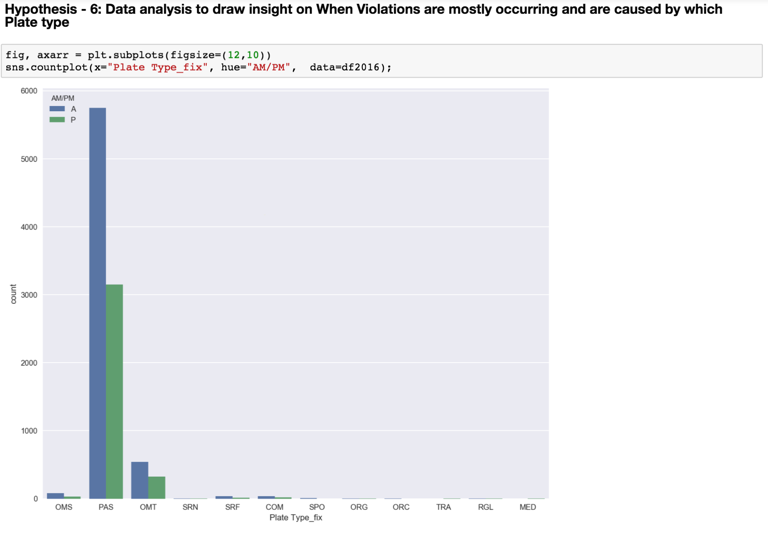
**Figure 3: Hypothesis 3 Visualization**



**Figure 4: Hypothesis 4 Visualization**



**Figure 5: Hypothesis 5 Visualization**



**Figure 6: Hypothesis 6 Visualization**

1. **NEW VARIABLES CREATED**

We have created 3 variables apart from the ones in the dataset.

* Week of the day
* Month
* AM/PM

1. **PSEUDO CODE (PYTHON SCRIPT)**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

sns.set(color\_codes=True)

%matplotlib inline

%config InlineBackend.figure\_format = 'retina'

df2016 = pd.read\_csv("Training dataset.csv")

d = {'Unique Entry': df2016.nunique(axis = 0),

'Nan Entry': df2016.isnull().any()}

pd.DataFrame(data = d, index = df2016.columns.values)

x\_ticks = df2016['Registration State\_fix'].value\_counts().index

heights = df2016['Registration State\_fix'].value\_counts()

y\_pos = np.arange(len(x\_ticks))

fig = plt.figure(figsize=(15,14))

# Create horizontal bars

plt.barh(y\_pos, heights)

pd.DataFrame(df2016['Registration State\_fix'].value\_counts()/len(df2016)).nlargest(10, columns = ['Registration State\_fix'])

# Create names on the y-axis

plt.yticks(y\_pos, x\_ticks)

# Show graphic

plt.show()

x\_ticks = df2016['Month of the Year'].value\_counts().index

heights = df2016['Month of the Year'].value\_counts()

y\_pos = np.arange(len(x\_ticks))

fig = plt.figure(figsize=(15,14))

# Create horizontal bars

plt.barh(y\_pos, heights)

# Create names on the y-axis

plt.yticks(y\_pos, x\_ticks)

# Show graphic

plt.show()

sns.swarmplot(x="Violation County\_fix", hue="Month of the Year", data=df2016);

sns.barplot(x="Violation County\_fix", y="Month of the Year", data=df2016);

fig, axarr = plt.subplots(figsize=(12,10))

sns.countplot(x="Violation County\_fix", hue="week of the day", data=df2016);

df2016.head()

fig, axarr = plt.subplots(figsize=(12,5))

sns.countplot(x="week of the day", hue="Plate Type\_fix", data=df2016);

fig, axarr = plt.subplots(figsize=(20,10))

sns.countplot(hue="Violation County\_fix", x="Month of the Year", data=df2016);

fig, axarr = plt.subplots(figsize=(20,10))

sns.countplot(x="Plate Type\_fix", hue="Violation County\_fix", data=df2016);

fig, axarr = plt.subplots(figsize=(12,10))

sns.countplot(x="Violation County\_fix", hue="AM/PM", data=df2016);

fig, axarr = plt.subplots(figsize=(12,10))

sns.countplot(x="Plate Type\_fix", hue="AM/PM", data=df2016);