



*Predicting Crime in Queens*  
*IST 565: Data Mining*  
*DAT (Data Analytics Team)*

## *Table of Contents*

[Meet DAT:](#)

[Our Objective:](#)

[Our Dataset:](#)

[Problem Faced:](#)

[Data Cleaning and Preprocessing:](#)

[Our Plan:](#)

[Our Results:](#)

[Descriptive Modeling](#)

[Predictive Data Modeling](#)

[Naive Bayes:](#)





[Support Vector Machine:](#)

[Our Conclusion](#)

Please see *DAT\_FinalProjectCode* for all the R code.

If you have any questions about this project, feel free to email William D. Lombardi - [wlombard@syr.edu](mailto:wlombard@syr.edu)

**Meet DAT:**

<i>Photo and Name of Group Member:</i>	<i>Contribution to the project:</i>
 <p><b>William D. Lombardi</b></p>	<ul style="list-style-type: none"> <li>• Found the dataset we used to this final project</li> <li>• Made and organized the presentation</li> <li>• Helped make the whole team work effectively and efficiently</li> </ul>
 <p><b>Adam Miller</b></p>	<ul style="list-style-type: none"> <li>• Dedicated his computer for the R Code</li> <li>• Wrote majority of the R Code for the final project</li> <li>• Decided what 3 Predictive Models we should use for this project</li> </ul>
 <p><b>Max Gertsman</b></p>	<ul style="list-style-type: none"> <li>• Geared us to stay focused on the final project</li> <li>• Created the virtualizations/descriptive statistics for the final project</li> <li>• Reviewed the R Code to make sure everything made sense</li> </ul>
 <p><b>James Lu</b></p>	<ul style="list-style-type: none"> <li>• Compiled the results from the R Code we wrote for each predictive models we made</li> <li>• Analyzed the decision tree and thought it will be smart to use the attributes in the other two predictive models</li> <li>• Looked up ways to make our R code simple and efficient.</li> </ul>

Please see *DAT\_FinalProjectCode* for all the R code.

If you have any questions about this project, feel free to email William D. Lombardi - wlombard@syr.edu

## **Our Objective:**

Since we learned all different predictive modeling techniques although this semester, we thought to use three, decision tree, naive bayes, and support vector machine. Based off three predictive models we decided to make, we agree it is smart to what degree of accuracy can we predict the level of crime that will be committed in the bars and nightclubs of Queens, New York.

## **Our Dataset:**

The dataset our group planned to use and found came from Kaggle. It contains information that describes different crimes that occurred in New York City from the years 2014-2015. There were a total of 24 columns with over 1,000,000 records. These attributes range from nominal to ratio-like in nature. The link for the complete dataset can be found at this link: <https://www.kaggle.com/adamschroeder/crimes-new-york-city/version/1#>

## **Problem Faced:**

The only problem we came across this particular set of data was dimensionality reduction, because we can assume that training a model on over 1,000,000 examples will be prohibitively time-consuming. In order to do this, we first needed to narrow the scope of our data. We plan to do this by focusing our efforts on a subset of the data which represents crimes that occurred in bars/nightclubs in the borough of Queens. Hopefully, with reducing it to a single borough and a specific kind of location, the dataset will be a much more manageable size.

## **Data Cleaning and Preprocessing:**

As a group we decided to clean the data in these following ways:

- We narrowed down and choose to work with the borough, Queens.
- We narrowed down and choose to work with the premise, Bar/Night Club.
- We used the na.omit function to remove any blank or null rows.
- After we did all that, we made a new file called 'Clean\_NYPD\_Data.csv'.
- We chose to work with columns, 3, 4, 9, 12, 13, 16, 17 that includes, date of the crime, time of the crime when it occurred, a description of the crime, the success of the crime, the level of offense, what precinct the crime happened in, and a general description of where the crime took place (inside, outside, in front of, etc.)
- We made the precinct attribute as a factor value since it was a numeric value.
- We mutated the date column into year and month columns, then converted them to factors, so the predictive algorithms could parse them easier
- For the same reason, used the "str\_split" function in R to extract the hour element from the time column, then set that as its own column in factor form.

Please see *DAT\_FinalProjectCode* for all the R code.

If you have any questions about this project, feel free to email William D. Lombardi - wlombard@syr.edu

After processing our dataset based of determining the scope of our problem for the dataset, we cleaned the data, and we got the dataset down to 8 columns with 2,328 records.

## Our Plan:

As a group we decided to do both descriptive modeling and predictive modeling:

For the descriptive modeling we decided to make three different types of models in form of a pie chart and histogram bar chart. We wanted to display the following:

- The 3 different levels of offense in a pie chart
- The level of offense and precinct in a histogram bar chart
- The level of offense and the hour of the day in a histogram bar chart

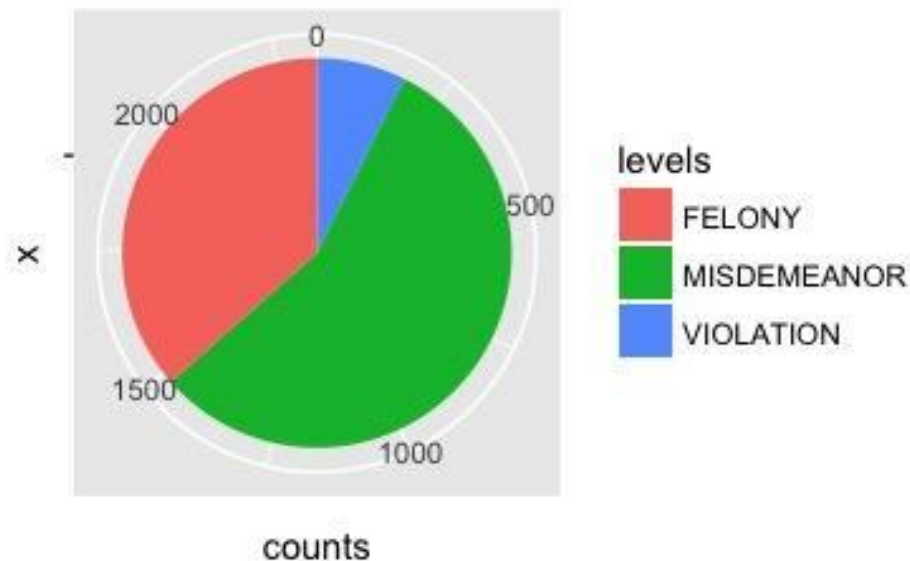
For the predictive modeling we decided to make three different predictive models:

- Decision Tree
  - Based on the attributes that decision tree gives us, we are going to build two alterations (one normal and one with the decision tree predictors) of the other two predictive models we make:
    - Naive Bayes
    - Support Vector Machine

## Our Results:

### Descriptive Modeling

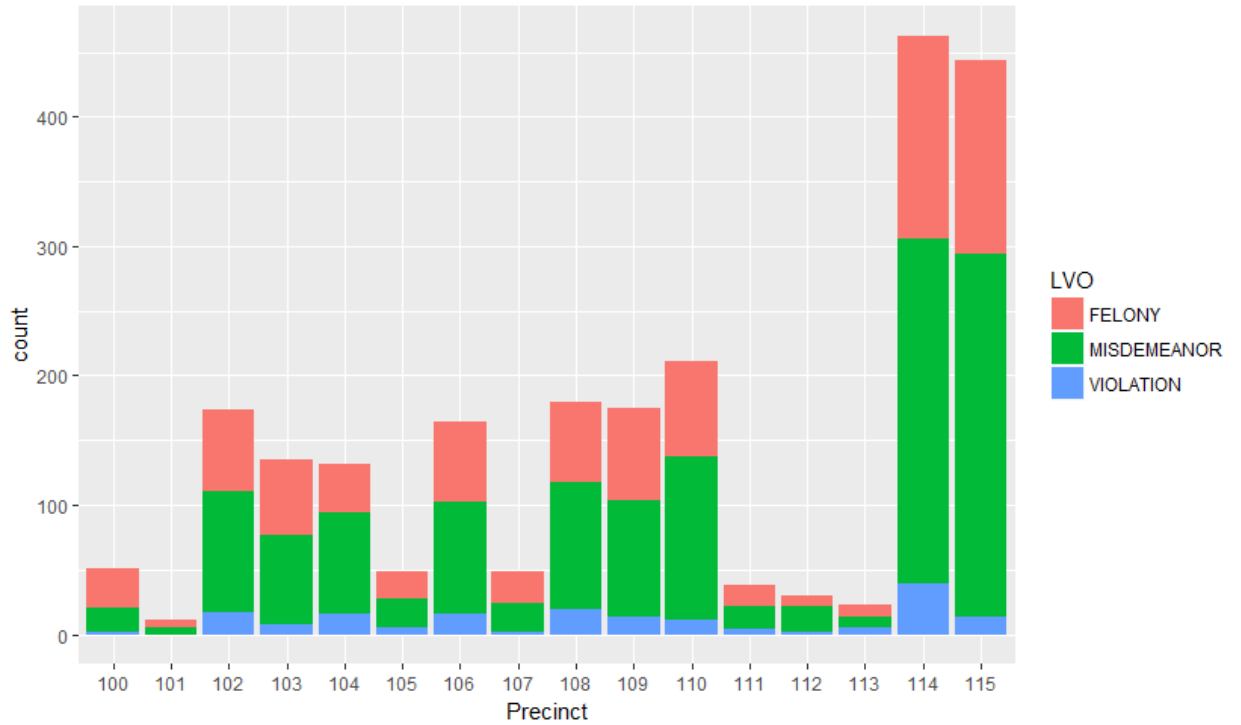
- *The total of the three different levels of offense in a pie chart*



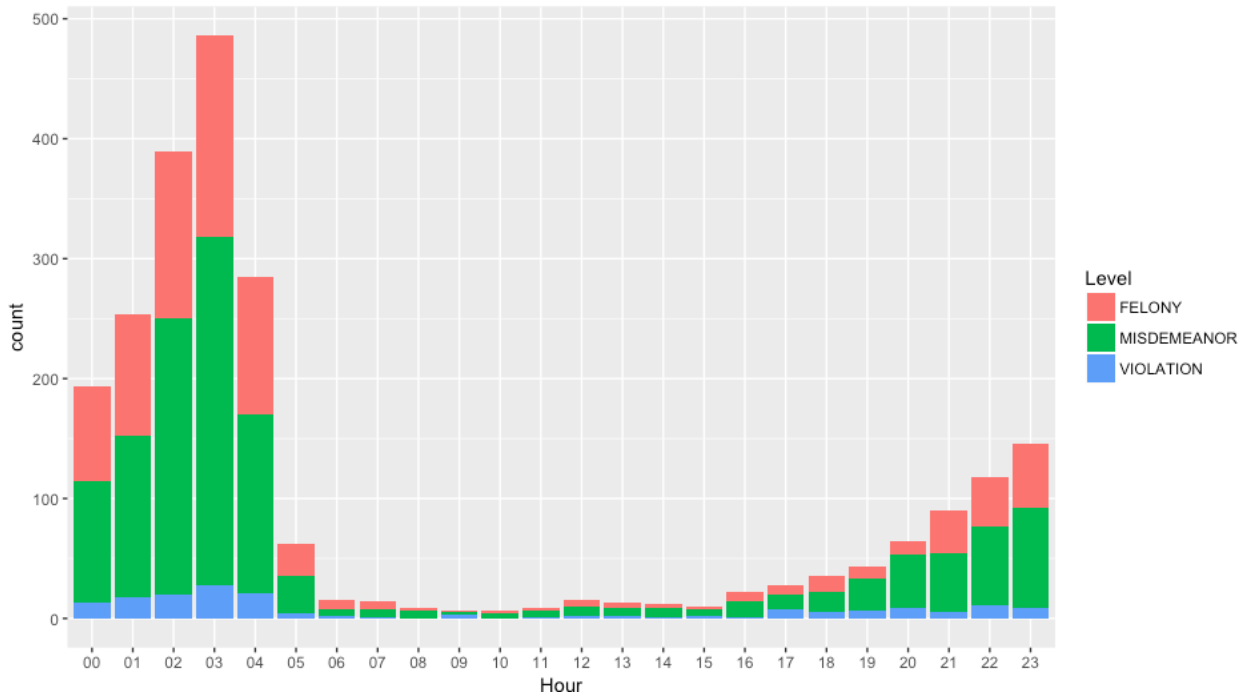
- *The level of offense and precinct in a histogram bar chart*

Please see *DAT\_FinalProjectCode* for all the R code.

If you have any questions about this project, feel free to email William D. Lombardi - wlombard@syr.edu

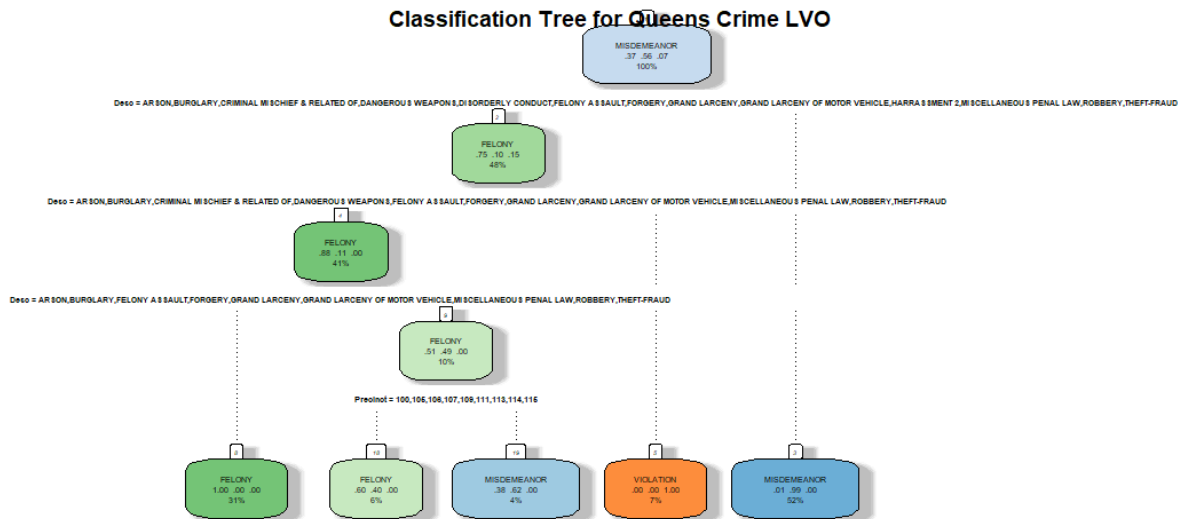


- 
- *The level of offense and the hour of the day in a histogram bar chart*



## Predictive Data Modeling

Because running predictive models with seven explanatory variables can lead to lower accuracy and slower build-time, we decided to pick the most important explanatory variables by building a decision tree using the RPart package in R.



The decision tree model took 78 seconds to build. There were two main attributes that the tree split on: Crime Description and Precinct. Logically, different descriptions of crimes would lead to different offense levels, larceny is more serious than loitering. A more unexpected split was precinct, but the fact that the model splits on precinct suggests that some precincts are more vulnerable to higher level offenses than others.

Out of 466 records in the testing data, the decision tree model correctly classified 442 of them, for a 94.8% accuracy. The model is clearly able to tell which crimes are violations, but has a small amount of confusion when differentiating felonies from misdemeanors.

Prediction	Reference		
	FELONY	MISDEMEANOR	VIOLATION
FELONY	164	12	0
MISDEMEANOR	12	235	0
VIOLATION	0	0	43

Naive Bayes:

- 93.3% without the optimization of the decision tree variables

Actual	Prediction		
	FELONY	MISDEMEANOR	VIOLATION
FELONY	161	18	0
MISDEMEANOR	7	250	0
VIOLATION	0	0	30

- 92% with the optimization of the decision tree variables

Please see *DAT\_FinalProjectCode* for all the R code.

If you have any questions about this project, feel free to email William D. Lombardi - wlombard@syr.edu

Actual	Prediction		
	FELONY	MISDEMEANOR	VIOLATION
FELONY	161	18	0
MISDEMEANOR	8	249	0
VIOLATION	0	0	30

- 
- With Adam tuning it, we learned the decision tree variables didn't really optimize in any shape or form
  - Time to build was really fast

### Support Vector Machine:

- 94.4% without the optimization of the decision tree variables

Prediction	Reference		
	FELONY	MISDEMEANOR	VIOLATION
FELONY	172	21	0
MISDEMEANOR	7	236	0
VIOLATION	0	0	30

- 
- 94.6% with the optimization of the decision tree variables

Prediction	Reference		
	FELONY	MISDEMEANOR	VIOLATION
FELONY	149	0	0
MISDEMEANOR	30	257	0
VIOLATION	0	0	30

- 
- With Adam tuning it, we learned the decision tree variables didn't really optimize in any shape or form
  - Time to build was really fast

## Our Conclusion

After working on this project throughout the semester, we learned that all three data models we made are relatively robust because all of them achieved greater 93%. Also, with exception of decision tree, all the algorithms were built in very little time. When creating the three algorithms, there was either little or no difference between each of the models. We were also able to observe or infer some biases based on both the descriptive as well as the predictive models we built. First, we noticed that crime is much more prevalent in certain precincts over others. Furthermore, we postulated that our data may be skewed due to user bias. In other words, the reported crime description may be subject to the opinion of whoever was the first responder. Finally, from our descriptive modeling we learned that majority of crime in our dataset are committed between the hours of 1-3 AM.

Please see *DAT\_FinalProjectCode* for all the R code.

If you have any questions about this project, feel free to email William D. Lombardi - wlombard@syr.edu