Final Report
Data Mining
April 2020
Omar Qusous, Sertan Akinci, Sam SR

Introduction

The main goal of the project is to clasify crimes based on time, location, victim's sex, and age group in addition to some correlation and patter analysis. We experimented with different kinds of models in order to find the best results, which ended up being a set of decision tree classifiers which predicted the possibility of a given observation of being a felony, misdemeanor, or violation-type crime. Our results were not as accurate as we would like, but through careful feature selection and tweaking available parameters in the models, we were able to get some results.

Dataset

The dataset we're working with comes from the City of New York, and consists of "all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department" from 2006 to the present, last updated in February. [1] We've limited the scope to around 350,000 records for the year 2019, for scaling purposes. Each report contains information such as the demographic information of the victim of the crime, in what borough of the city or where on the property the crime occured, as well as information concerning the suspect.

Cleaning

Each model had the base 2019 dataset cleaned in place for that particular application. For example, for the Decision Tree model, the following cleaning scheme was used:

In order to make the data useful, the features for which we wanted to use in our models had to be categorized. Some of them are numerical and ordinal, such as the various times of day associated with the report, but needed to be made into a python-readable number. Times and dates were transformed quickly in excel into numerical form. Non-numerical but ordinal features such as age group (18-24, 24-45, etc) were mapped to a scale of 1 through 5, representing the order in which they can be grouped. All other categorical variables, in order to be made useful for models like decision tree, were

[1] "NYPD Complaint Data Historic: NYC Open Data." *NYPD Complaint Data Historic | NYC Open Data*, City of New York, 17 Jan. 2020, data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i.

translated into new variables, described as follows: For a variable like “borough” which has a few different possible options, a new feature was added for each option, indicating if the report came from each borough, with a 1 for yes, and 0 for no. This expands the size of the dataset and introduces a lot of sparse-ness, but if they were mapped to a random set of numbers we would be introducing unwanted ordinality.

Null values were handled on a feature-by-feature basis. Features which were already numeric or ordinal were set to the average. Categorical and not ordinal features were filled in based on the given distribution of available values, so if there were 75% male, 25% female, the null or invalid entries were changed to match that distribution. Features with very small numbers of null values had those reports removed, since the scale of our dataset meant that we could accommodate some outright removal.

The final list of features is given below:

Feature in the DataFrame	Description	Type
CMPLNT_FR_DT_CODED (0 = 1/1/1900 for date)	Complaint date	Discrete
COMPLT_FR_TM_CODED (0 = midnight)	Complaint reported time	Discrete
VIC_AGE_GROUP_CODED (<18 = 1, 18-24 =2 etc..)	Victim’s age	Discrete
BORO_NM	Borough’s Name (Brooklyn. Manhattan, etc)	Categorical
VIC_RACE	Victim’s race	Categorical
LOC_OF_OCCUR_DESC	Proximity of crime to location	Categorical
JURIS_DESC	Police department jurisdiction	Categorical
PREM_TYP_DESC	Location of crime (shop, restaurant. Street, etc..)	Categorical

Modeling

Our team split into two, one to create a model to predict the class of the crime committed (Felony, Misdemeanor or Violation) the other two predict the actual type of crime that was committed (robbery, fraud, larceny and approximately 25 other such categories).

The models utilized were kNN, Random Forest, Decision Trees, Gradient Boost (XGBoost library) and SVM. Time permitting, for some models GridSearchCV was utilized to carry out cross validation and hyperparameter tuning.

Our most successful models were decision trees. We have three decision tree models each for predicting if a report is a felony, misdemeanor, and violation. The features we

chose for the model include: time of reported event, date of reported event, victim age, victim sex, victim race, borough in which the event took place, description of the premises, specific location on the premises, and jurisdiction code description Some of these variables were numerical, some were categorical and ordinal, and some were purely categorical. All were encoded into a numerical value of some kind, either a scale where ordinal, or binary 0 and 1 for false and true, in order to accomodate the limitations of the scikit-learn module.

Part of the challenge was to find/reveal a relationship between some of the features and the target because of their high number of unique values of some dependent and independent variables. Looking at the distribution of these variables against the dependent variable suggested a grouping scheme for premise description, month, and the time of the day. These were some of the variables we encoded in a variety of ways to improve the model. For instance, in one logistic model, we encoded time in milliseconds and converted date into a three-digit number to have continuous numerical values as well as categorical (we also standardized them so that the magnitude of one variable wouldn't dominate). In another model, we divided hours of the day into three and months into two categories where the crimes were mostly taking place.

Once the variables were encoded, the modeling phase began with Random Forest (RF). The dependent variable was the crime code and the independent variables were time, date, premise type, premise description, borough, victim's age and sex. After splitting the data into test and train, RF with bootstrap gave disappointing results. Crime description, two different crime codes, and a crime category (felony, misdemeanor, violation) were tested as target variables while the features only varied in the type of encoding, like numerical vs categorical. Trying different combinations of dependent and independent variables, only improved the model slightly but not as expected.

The next method was Multinomial Logistic Regression. A slightly different approach, recursive feature elimination (RFE), was applied to the entire data to see what features it would deem important to use in the model. I tried logistic regression with and without using the RFE and had similar results as RF. While we kept getting undesirable results, we were both improving and converging to a disappointing accuracy with a consistency which could also be an indicator.

The last method was Support Vector Machine. Similar results repeated themselves here too where the most basic model gave the worst results and as we changed the groupings of the features and fine-tuned parameters of the model the results got better and better until converging around 47%.

Results

In order to measure the success of our models, we took the accuracy, F1 score, and confusion matrix of each model, after performing a train-test-split. Where our accuracies were strong, our F1 scores were rather low. This means our dataset is not well balanced and our model was underfitting, we were unable to resolve in the time we had. This is reflected in our confusion matrices.

Target	**Violation**	**Felony**	**Misdemeanor**
Accuracy	84%	70%	58%
Negative Predictive Value	0.13%	6%	61%
F1 Score	F1: 11%	3%	61%
Confusion Matrix	28134 14 5243 7	22903 561 9343 591	8497 6832 7034 11035

Figure 1: Metrics of success, Accuracy, F1 Score, Confusion Matrix for the Decision Tree Models

The average accuracy of our other exploratory models is as follows:

Model	**Target Categories**	**Accuracy**
kNN	Type of crime	30%
Random Forest	Type of crime	40%
Decision	Class of crime	71%
Gradient Boost	Type of crime	40%
SVM	Type of crime	47%
Regression	Type of crime	46%

Figure 2: Accuracies for other modeling attempts

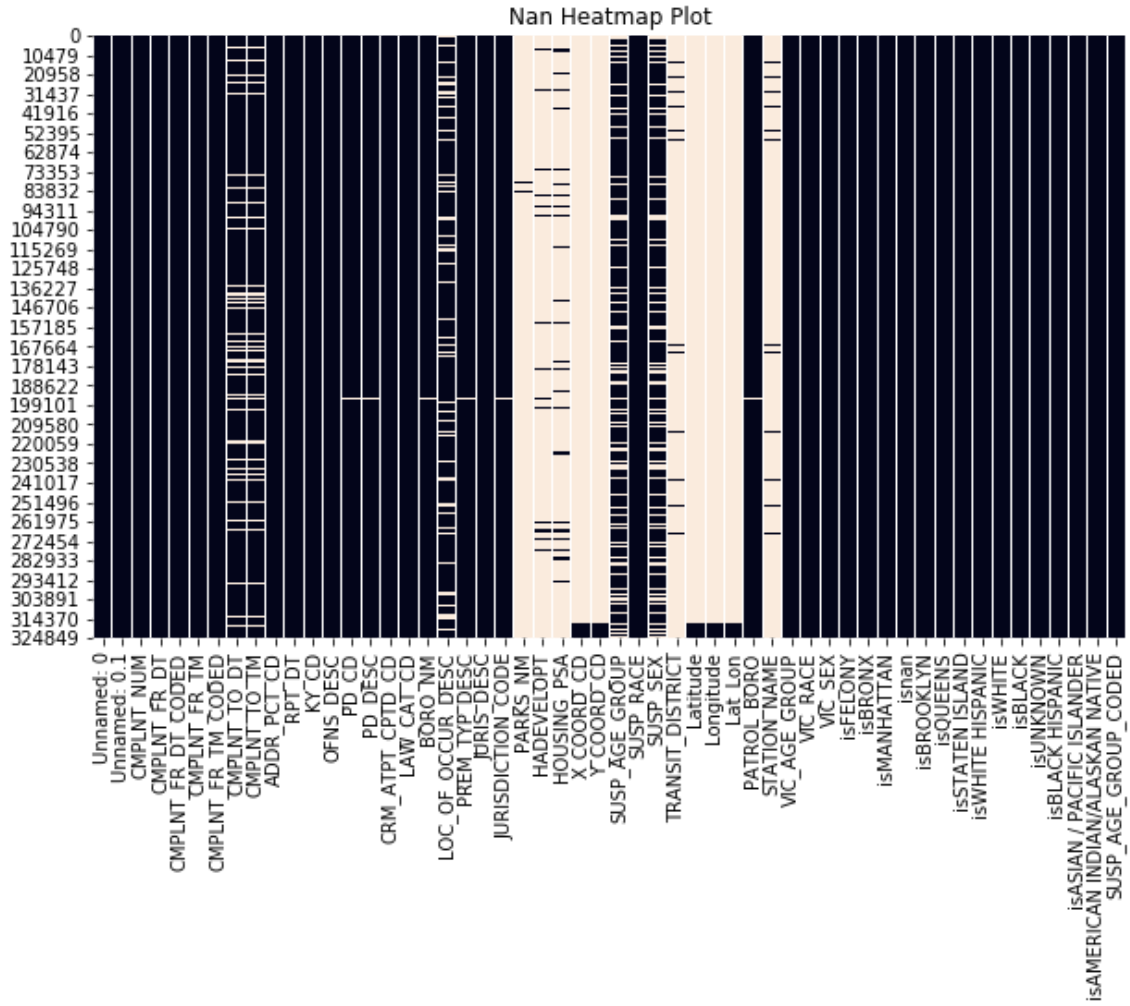
Conclusion

Ultimately, we were able to generate some valuable models for predicting crime classification, but there is clear room for improvement. Our model is able to predict with reasonable accuracy what kind of crime a particular report could have been, which was the initial question we set out to answer. However, if we had more time, we would look into better ways of encoding our features and being more discriminate with our feature selection. Another challenge was the imbalance in the data, especially between the type of crimes. A solution to this would be adding (or subtracting data) until there is relatively balanced data and using models that penalize for imperfect data.

Appendix:

Exploratory Data Analysis

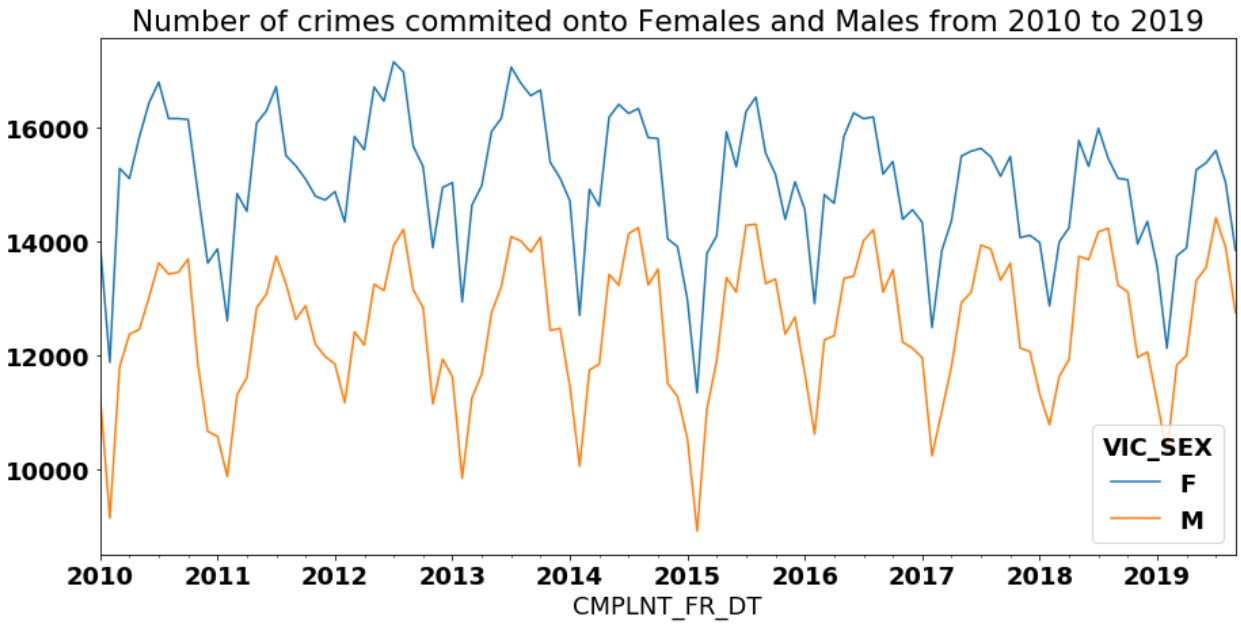
The plots below were utilized to explore the data along.



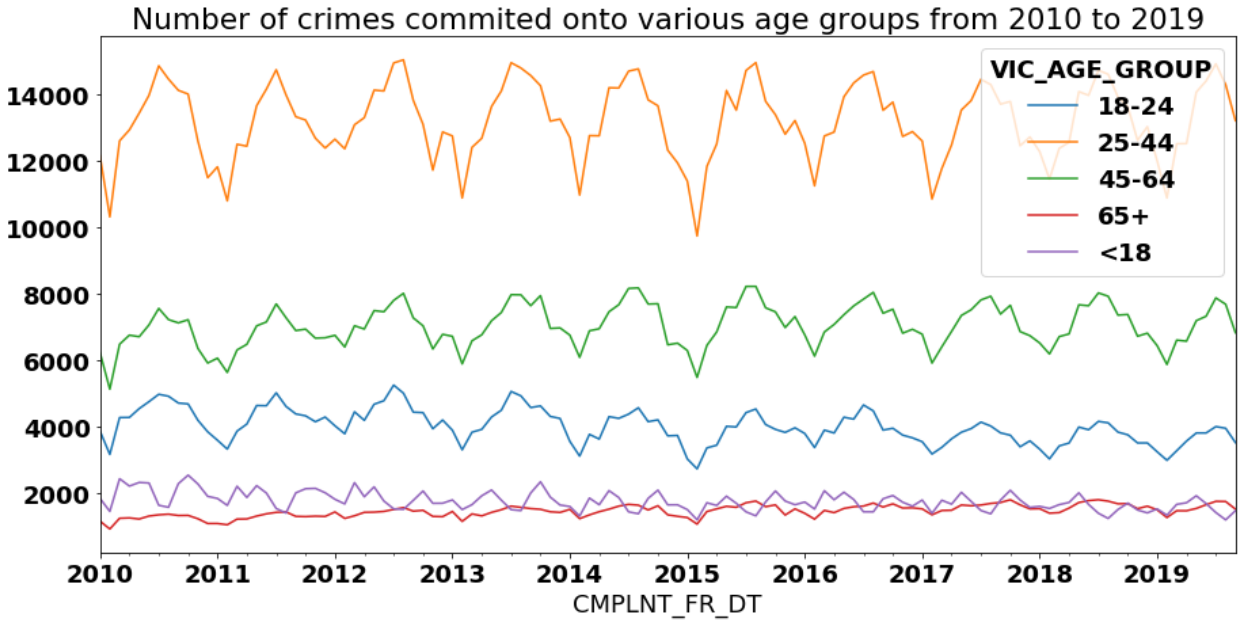
Appendix 1: Heatmap of Nan values in the dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 335300 entries, 0 to 335299
Data columns (total 54 columns):
Unnamed: 0                          335300 non-null int64
Unnamed: 0.1                        335300 non-null int64
CMPLNT_NUM                          335300 non-null int64
CMPLNT_FR_DT                        335300 non-null object
CMPLNT_FR_DT_CODED                  335300 non-null float64
CMPLNT_FR_TM                        335300 non-null object
CMPLNT_FR_TM_CODED                  335300 non-null float64
CMPLNT_TO_DT                        292479 non-null object
CMPLNT_TO_TM                        292653 non-null object
ADDR_PCT_CD                         335300 non-null int64
RPT_DT                              335300 non-null object
KY_CD                               335300 non-null int64
OFNS_DESC                           335294 non-null object
PD_CD                               335088 non-null float64
PD_DESC                             335088 non-null object
CRM_ATPT_CPTD_CD                    335300 non-null object
LAW_CAT_CD                          335300 non-null object
BORO_NM                             335085 non-null object
LOC_OF_OCCUR_DESC                   273470 non-null object
PREM_TYP_DESC                       333975 non-null object
JURIS_DESC                          335300 non-null object
JURISDICTION_CODE                   335088 non-null float64
PARKS_NM                            3270 non-null object
HADEVELOPT                          15896 non-null object
HOUSING_PSA                         24491 non-null object
X_COORD_CD                          20498 non-null float64
Y_COORD_CD                          20498 non-null float64
SUSP_AGE_GROUP                      257228 non-null object
SUSP_RACE                           335300 non-null object
SUSP_SEX                            257228 non-null object
TRANSIT_DISTRICT                    8577 non-null float64
Latitude                            20498 non-null float64
Longitude                           20498 non-null float64
Lat_Lon                             20498 non-null object
PATROL_BORO                         335088 non-null object
STATION_NAME                        8577 non-null object
VIC_AGE_GROUP                       335300 non-null object
VIC_RACE                            335300 non-null object
VIC_SEX                             335300 non-null object
isFELONY                            335300 non-null float64
isBRONX                             335300 non-null float64
isMANHATTAN                         335300 non-null float64
isnan                               335300 non-null float64
isBROOKLYN                          335300 non-null float64
isQUEENS                            335300 non-null float64
isSTATEN ISLAND                     335300 non-null float64
isWHITE HISPANIC                    335300 non-null float64
isWHITE                             335300 non-null float64
isBLACK                             335300 non-null float64
isUNKNOWN                           335300 non-null float64
isBLACK HISPANIC                    335300 non-null float64
isASIAN / PACIFIC ISLANDER          335300 non-null float64
isAMERICAN INDIAN/ALASKAN NATIVE    335300 non-null float64
SUSP_AGE_GROUP_CODED                335300 non-null float64
dtypes: float64(24), int64(5), object(25)
memory usage: 138.1+ MB
```

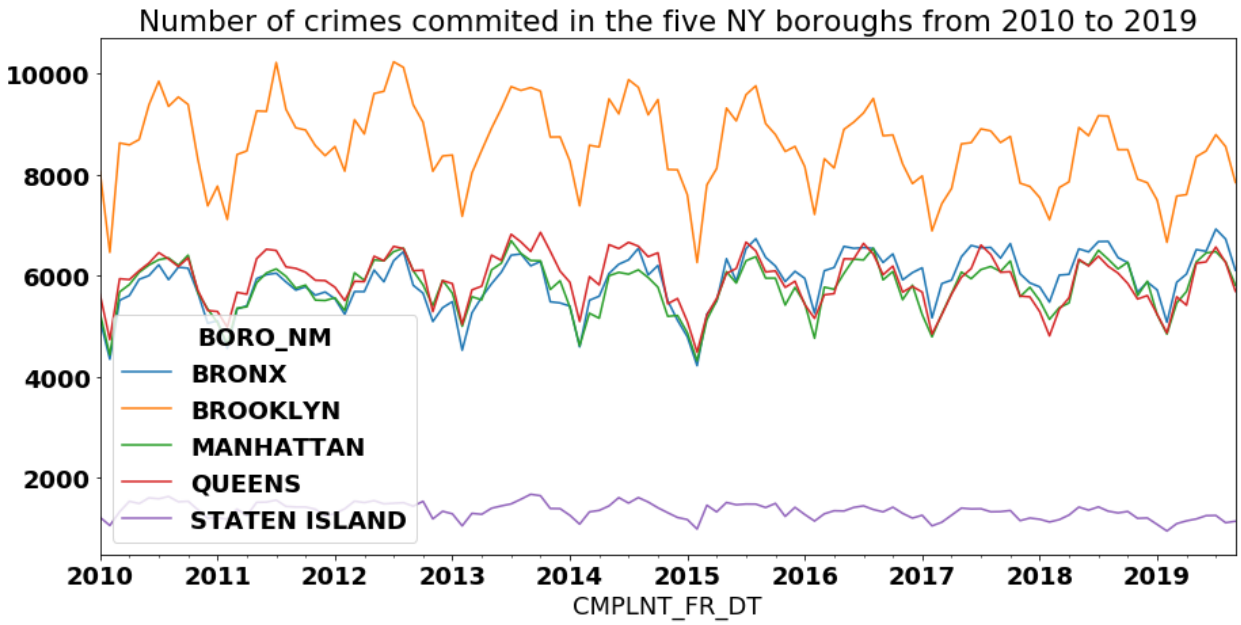
Appendix 2: Output from df.info()



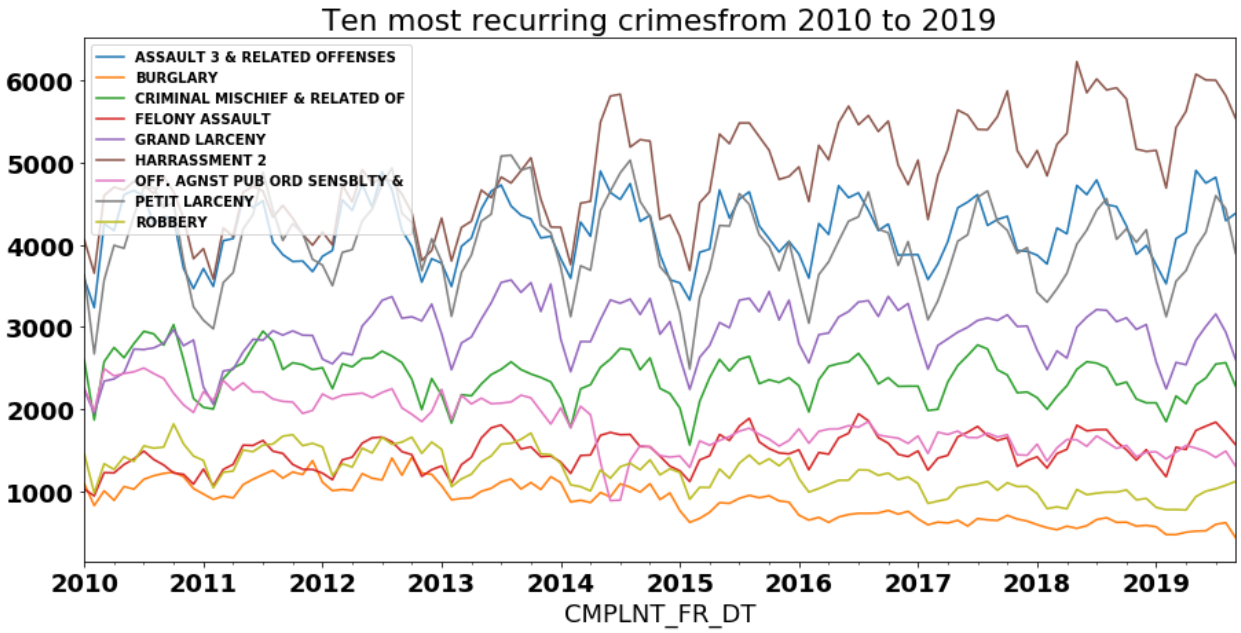
Appendix 3: Plot of crimes over time committed by men versus women, before the dataset was truncated to 2019 only.



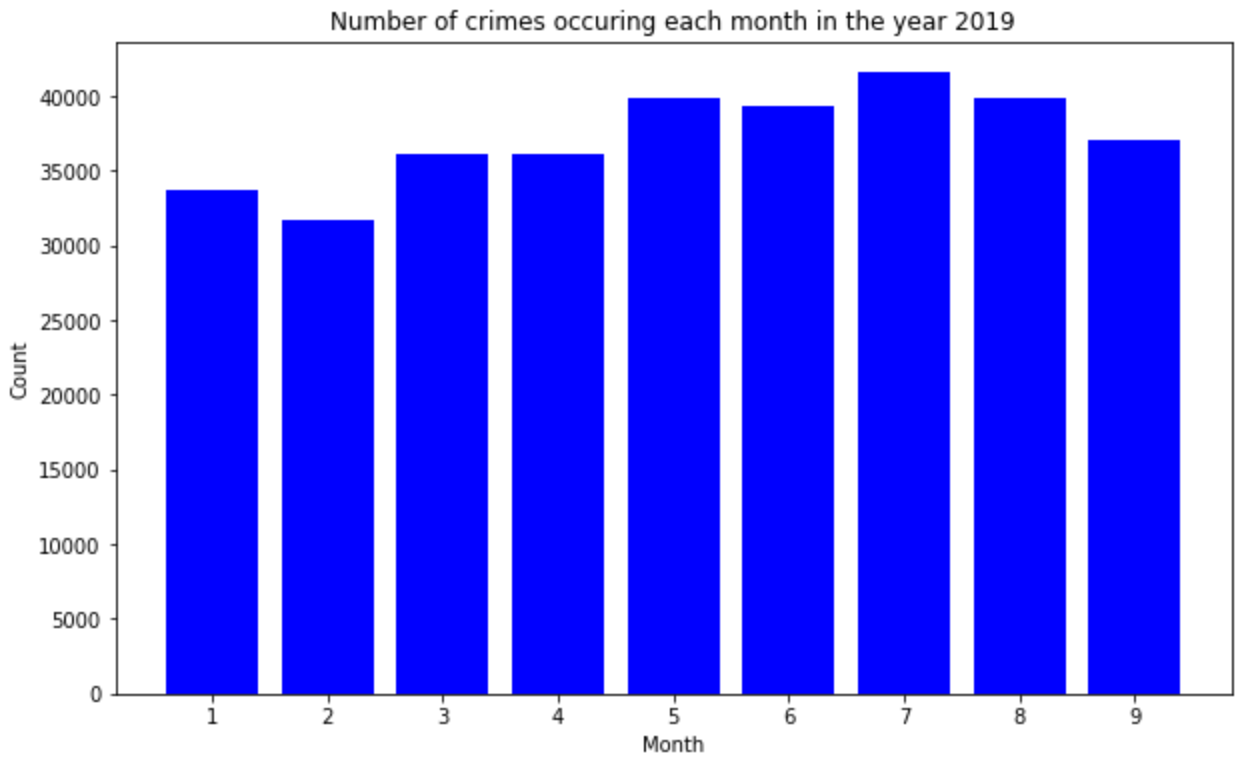
Appendix 4: Plot of crimes over time committed by the different age group categories, before the dataset was truncated to 2019 only.



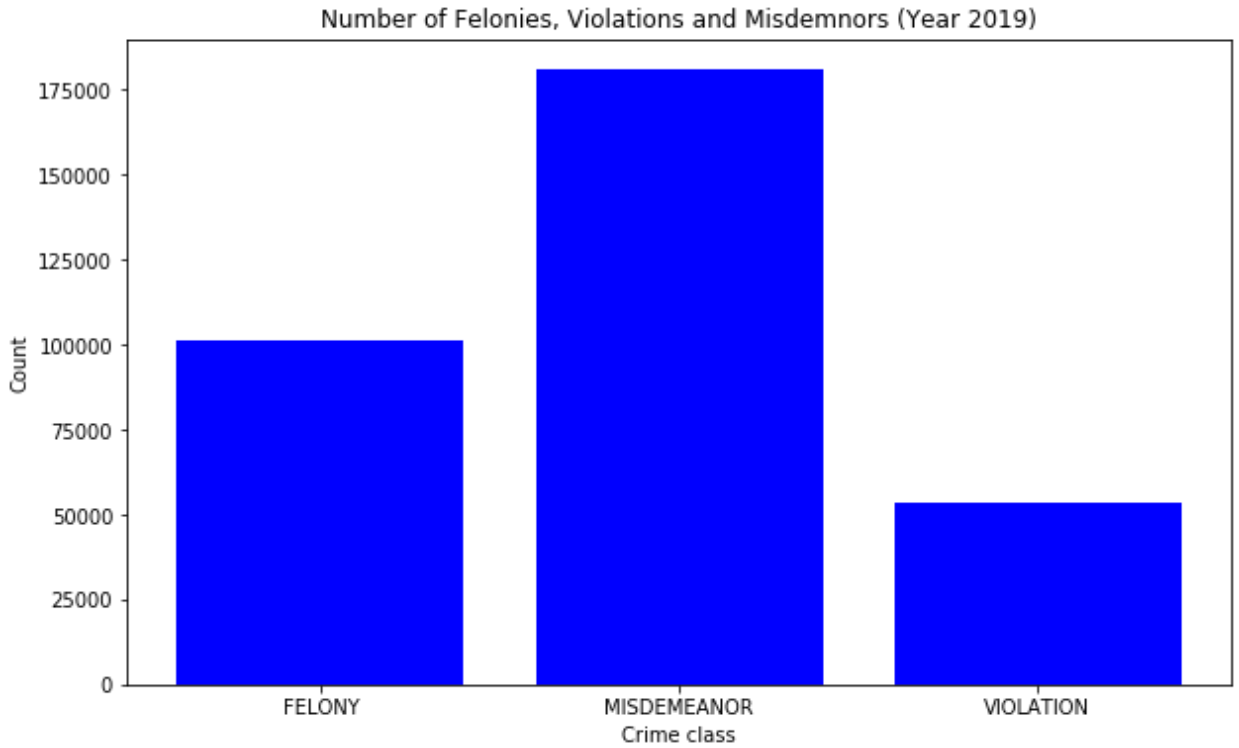
Appendix 5: Plot of crimes over time committed in the different boroughs, before the dataset was truncated to 2019 only.



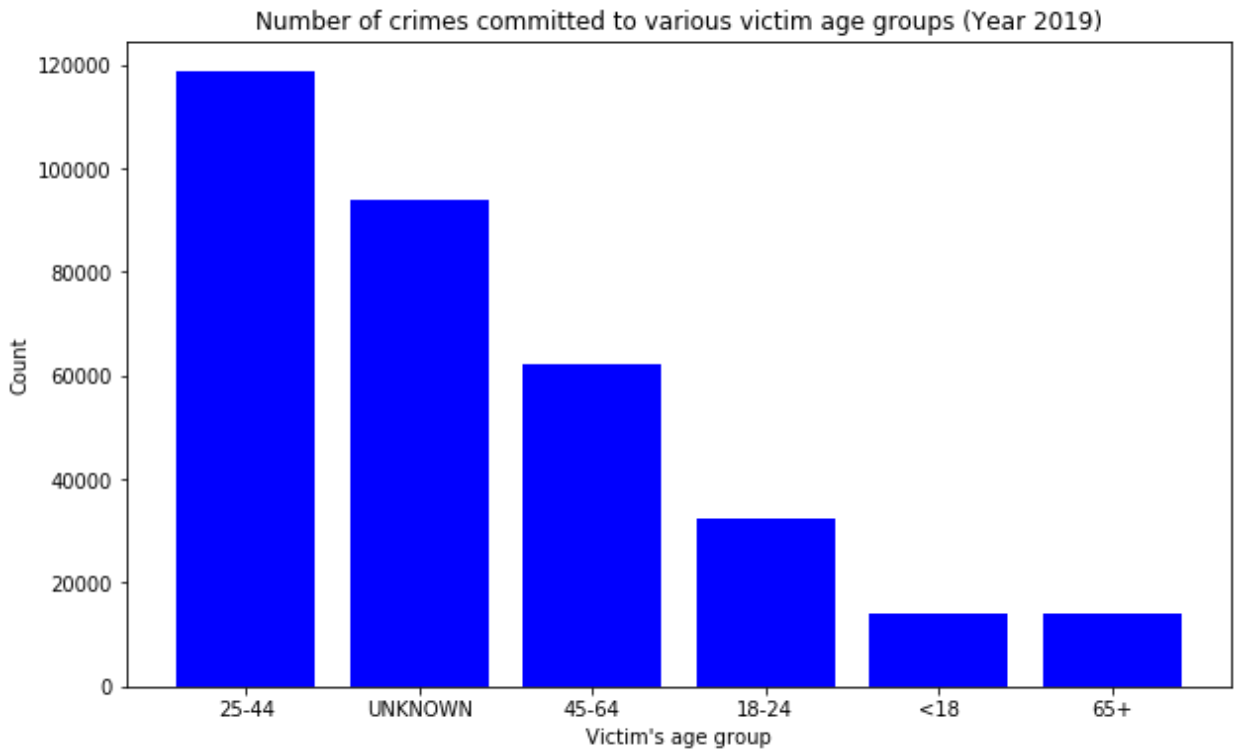
Appendix 6: Plot of ten post recurring crime types, before the dataset was truncated to 2019 only.



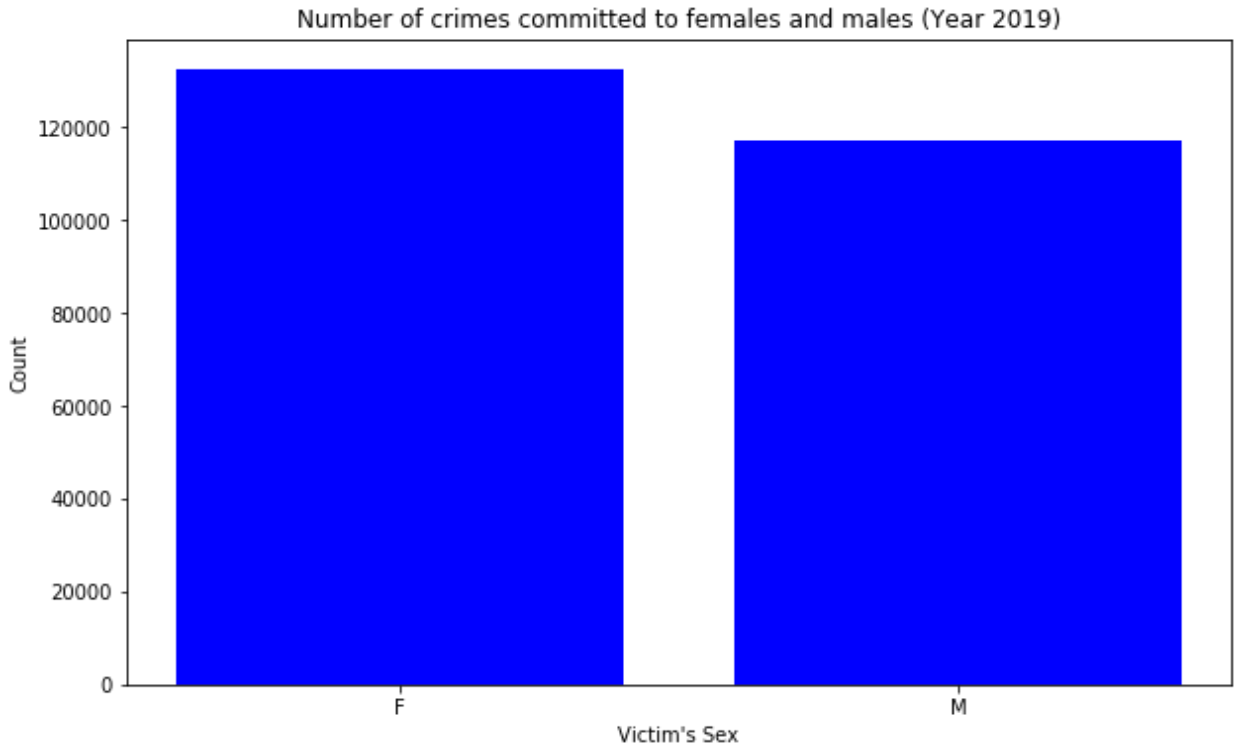
Appendix 7: Plot of crimes committed during each month in 2019



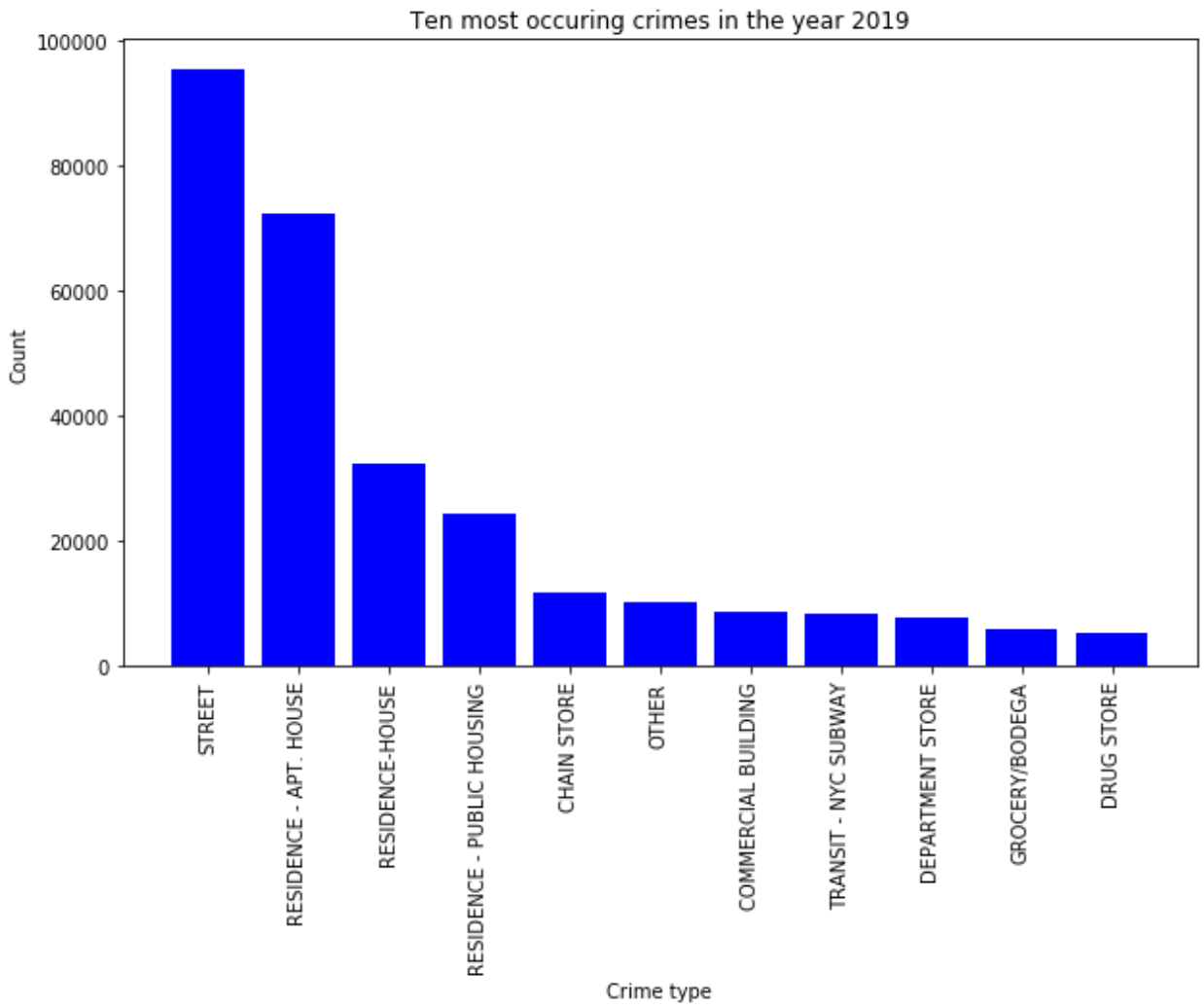
Appendix 8: Plot of the counts for each crime class in 2019



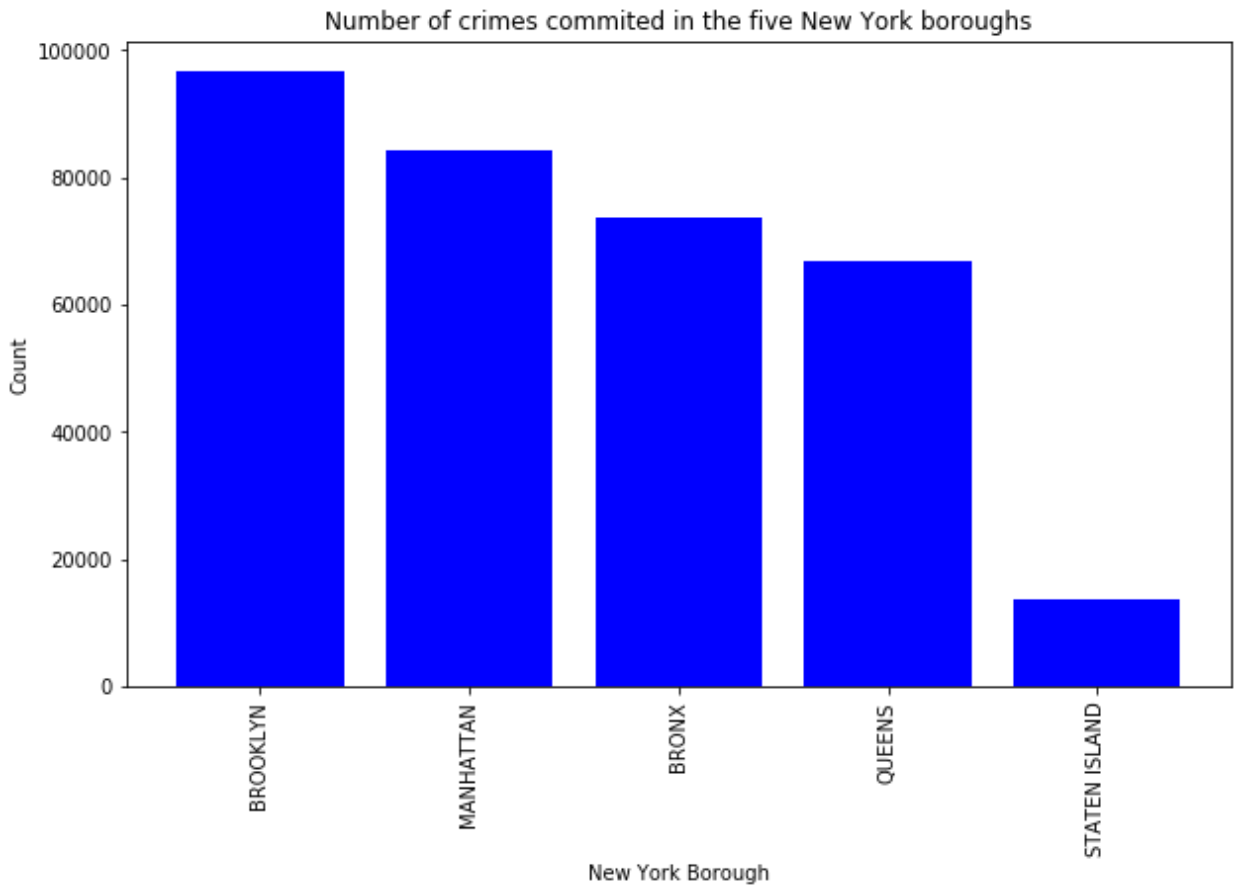
Appendix 9: Plot of the counts for each age group in 2019



Appendix 10: Plot of the counts for victim sex in 2019



Appendix 11: Plot of the counts for each crime type in 2019



Appendix 12: Plot of the counts for crimes committed in each borough of New York in 2019