Forecasting Crime Around the Georgia Tech Campus

Interim Report

Phillip Spratling

Introduction

Despite having a large police force, Atlanta has some of the highest crime rates in the country today. Forbes recently rated it the 6th most dangerous U.S. city, with a violent crime rate of 1,433 per 100,000 residents. Apart from the high crime rate, Atlanta is also home to some of the nation's best universities. In particular, the Georgia Institute of Technology is located right in the heart of Atlanta. Students come from across the world in pursuit of a higher education, and to do so they must learn to keep themselves safe. Students are advised to follow certain rules such as "Don't walk alone after dark!" or "Stay away from Home Park!". Even when following these rules, sometimes being the victim of a crime is inevitable. Knowing this, the Georgia Tech Police Department (GTPD) do all they can to keep their students safe.

In an effort to make crime around campus more transparent, the GTPD have since 2010 posted crime logs on their website available for public access. Every time an incident is reported, it is added to the log with various attributes such as the time of the incident, the type, and its location. My goal with this project is to analyze this data over the years and determine which factors play a role in predicting crime. I will then build a model to predict future crime. The analysis and predictions will enable the GTPD to better place their officers around campus, as well as help them decide which specific types of crimes to work towards reducing. It will also help Georgia Tech students know how to better keep themselves safe.

Data Gathering

The first and major dataset I used in my analysis is the crime logs posted by the Georgia Tech Police Department. In an effort to make crime around campus more transparent, the GTPD have since 2010 posted crime logs on their website available for public access. Every time an incident is reported, it is added to the log with various attributes such as the time of the incident, the type, and its location. The data can be located at the GTPD public website.

In addition to this dataset, I merged demographics data from the city of Atlanta and the surrounding areas of Georgia Tech. This data includes information such as population levels and income levels for various age segments, with the idea being that these statistics are usually a strong predictor of crime in any given area. This demographics data was obtained at the census.gov website.

Finally, I merged Georgia Tech enrollment data. Student enrollment has increased over time, and is less in the summer semesters. As the number of people on campus increases, the opportunities for crime rise, so this data could be a strong predictor of crime. This data was obtained from the Georgia Tech website.

Data Cleaning

The uncleaned data after first being merged had 19292 rows and 44 columns, ranging from location information, time information, status of the crime, and supplemental enrollment and demographics information at the time of the crime. The first step after merging the data was to clean and process it to be used for analysis.

Many of the columns were not the right data type. For example, the latitude and longitude columns were objects instead of floats. After analysis, this was determined to be because of a few incorrectly entered points. After converting incorrectly entered points to 0 or null values, the columns were able to be converted to the right type.

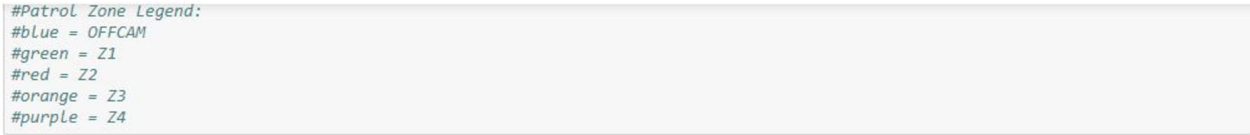
There were also some naming inconsistencies throughout the years. In some documents, many of the crime descriptions and location descriptions are written with a en dash (–) and many are written with a hyphen (-). I replaced all the en dashes with hyphens to merge duplicate values. Similarly, on some of the years, the landmark column had additional zone information appended to the end. This led to many of the same landmarks being treated as different ones. I was able to strip this information from the end to merge these duplicate values.

There was a large amount of missing data as well. This had to be dealt with on a point by point basis. For some less important columns, this information was left as blank. However, others were imputed or removed. For example, some missing latitude and longitude coordinates were able to be filled in from other points at the same landmark location. A similar process was used for missing crime codes and their corresponding crime descriptions.

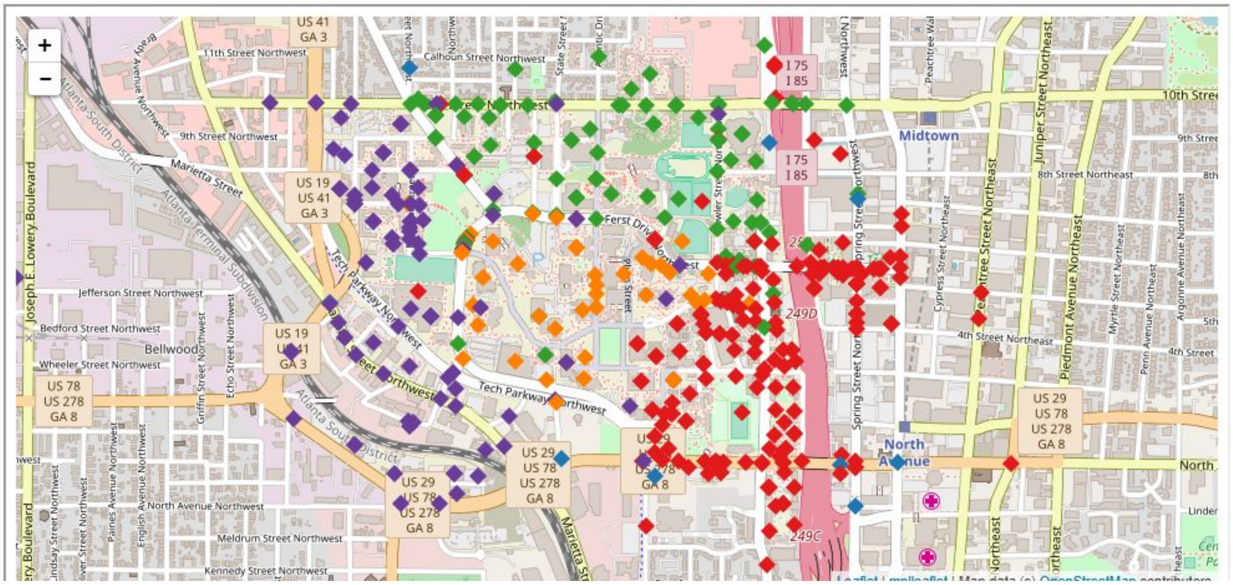
There were some miscellaneous, column specific errors that needed to be cleaned as well. For example, since the data is input manually, some latitude and longitude columns were entered in reverse. These were found and swapped accordingly.

Finally, after cleaning all the columns, bad and unnecessary rows of data were dropped. Every time someone calls the GTPD has to be logged, regardless of whether a crime occured. This led to many rows of non-crimes that are irrelevant to our analysis. These rows were dropped. In addition, crimes that didn’t have any location or description data were dropped. I also dropped points from the most current month, as these numbers are incomplete and would lead to false statistics in the following sections of the project.

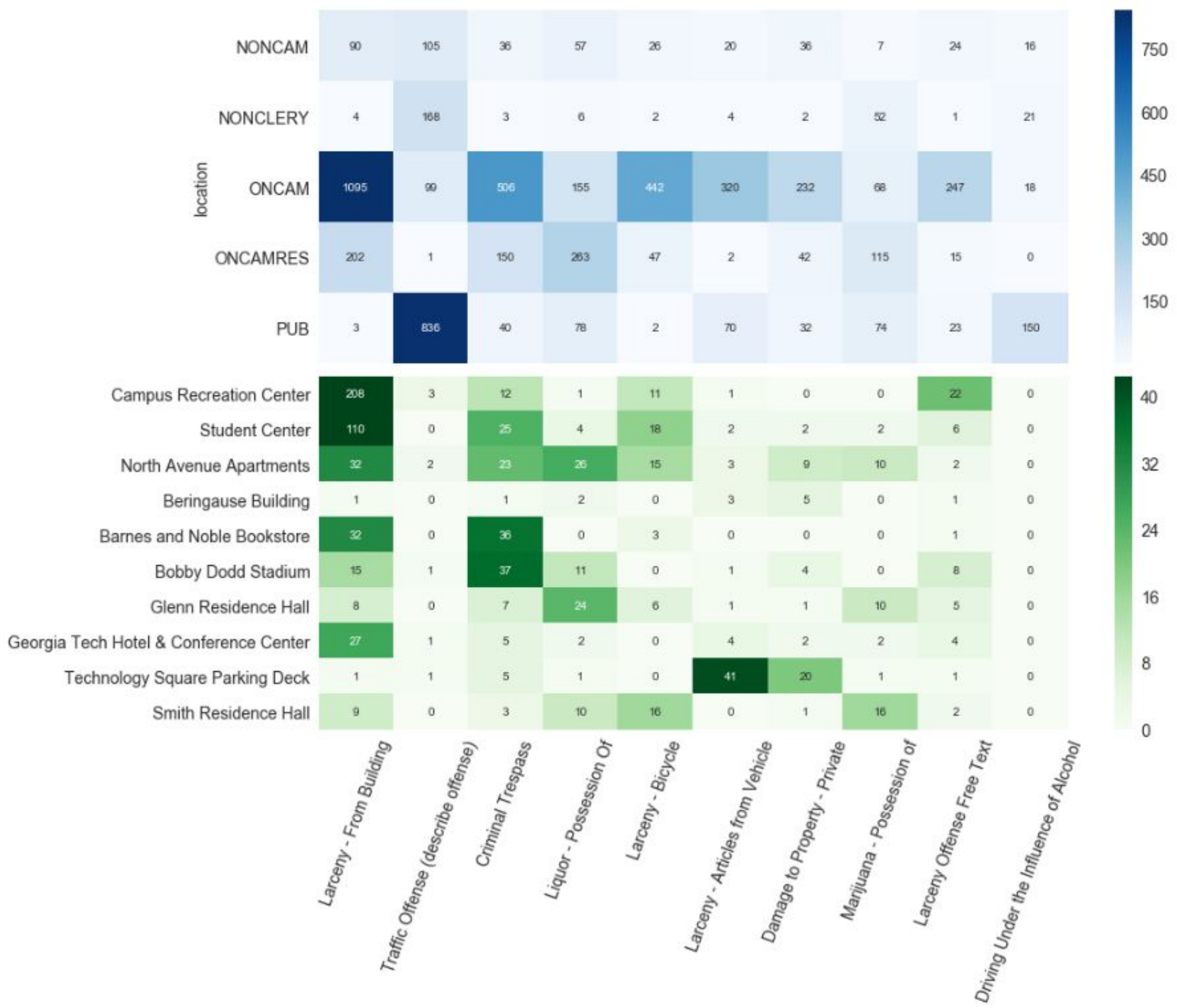
Exploratory Data Analysis



```
#Patrol Zone Legend:
#blue = OFFCAM
#green = Z1
#red = Z2
#orange = Z3
#purple = Z4
```



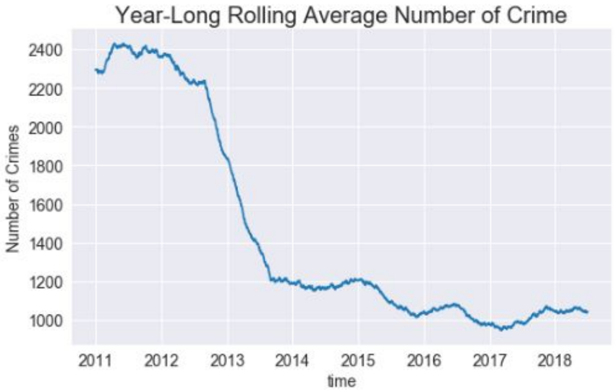
First, location based crime trends were analyzed. Above is a map showing all the unique locations crimes have been reported since 2010. It appears that crimes have happened in almost every location on campus. They seem to be distributed the densest in the residence halls and fraternity houses on the eastern portion of campus. Additionally, crimes appear to be more sparse near the center of campus. The closer you get to the edges of campus, the more crime locations there are. Further analysis was done on which types of crime are more frequent in different locations.

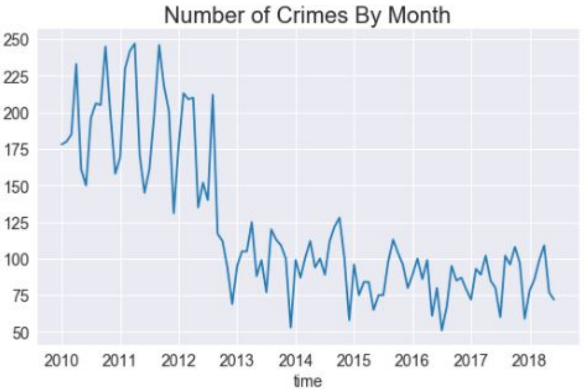


location	Larceny - From Building	Traffic Offense (describe offense)	Criminal Trespass	Liquor - Possession Of	Larceny - Bicycle	Larceny - Articles from Vehicle	Damage to Property - Private	Marijuana - Possession of	Larceny Offense Free Text	Driving Under the Influence of Alcohol
NONCAM	90	105	36	57	26	20	36	7	24	16
NONCLERY	4	168	3	6	2	4	2	52	1	21
ONCAM	1095	99	506	155	442	320	232	68	247	18
ONCAMRES	202	1	150	263	47	2	42	115	15	0
PUB	3	836	40	78	2	70	32	74	23	150
Campus Recreation Center	208	3	12	1	11	1	0	0	22	0
Student Center	110	0	25	4	18	2	2	2	6	0
North Avenue Apartments	32	2	23	26	15	3	9	10	2	0
Beringause Building	1	0	1	2	0	3	5	0	1	0
Barnes and Noble Bookstore	32	0	36	0	3	0	0	0	1	0
Bobby Dodd Stadium	15	1	37	11	0	1	4	0	8	0
Glenn Residence Hall	8	0	7	24	6	1	1	10	5	0
Georgia Tech Hotel & Conference Center	27	1	5	2	0	4	2	2	4	0
Technology Square Parking Deck	1	1	5	1	0	41	20	1	1	0
Smith Residence Hall	9	0	3	10	16	0	1	16	2	0

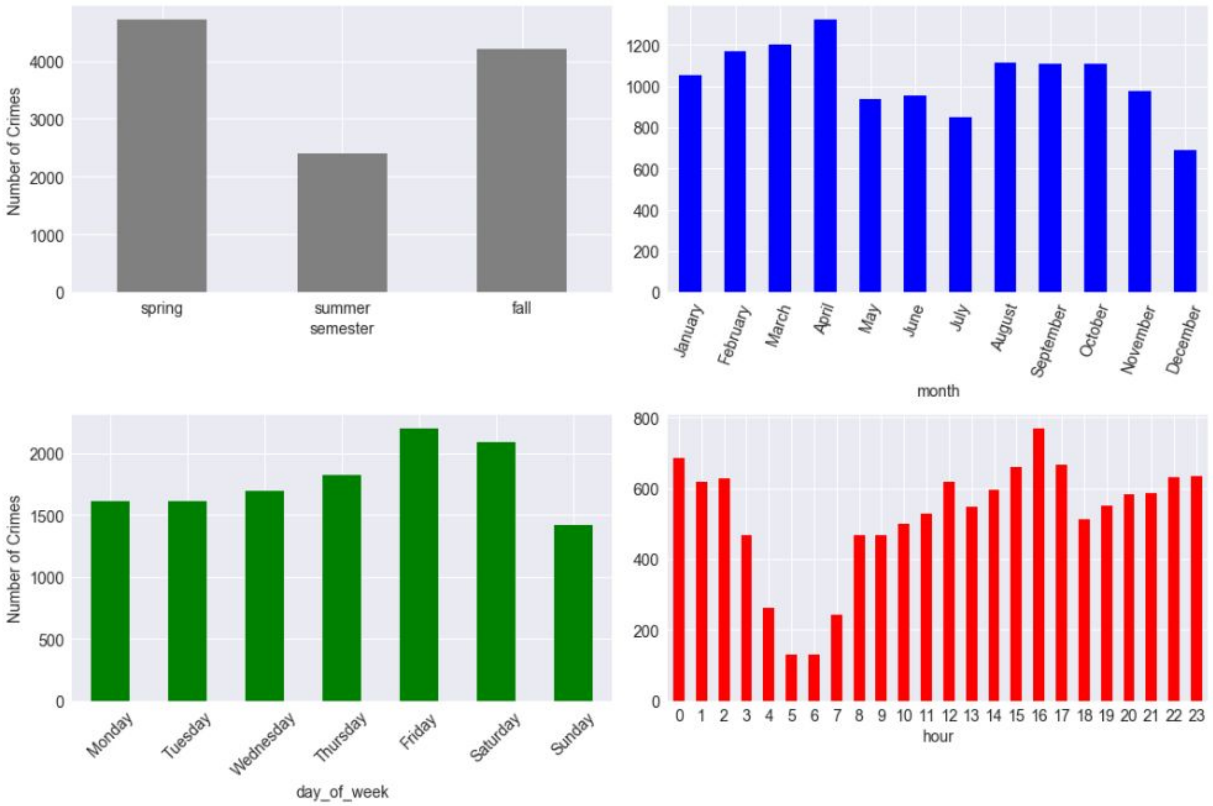
This figure shows the frequencies of the 10 most frequent crimes in both the 5 location zones and the 10 most frequent landmarks around campus. It seems that most crimes are being reported on campus as opposed to in the surrounding areas or in residential areas. Larceny is very prevalent throughout campus, happening frequently in the Student Center and the CRC in particular. The absence of the CULC is interesting to note - despite being one of the most popular locations for students to be on campus, it is not even in the top 10 highest locations for crime. This could be because of the active security guards always on location.

Larceny from vehicle and damage to property appears to be overwhelmingly in the Tech Square Parking Deck as opposed to the many other parking locations on campus. This might be because of the public access to this parking deck. It may be a good idea to choose another location to park on campus, even if this one is conveniently located. Residential areas seem to be fairly safe. Larceny and criminal trespass seem to occur occasionally, but the most frequent crime occurring in these areas appears to be illegal possession of drugs (alcohol and marijuana).

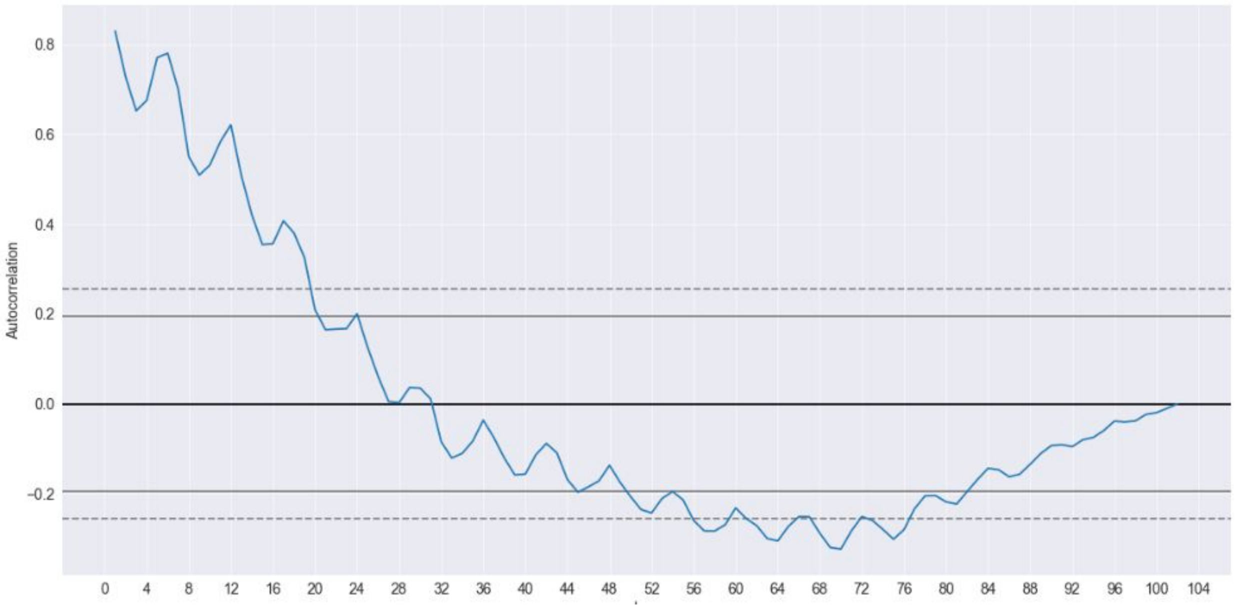

Year-Long Rolling Average Number of Crime

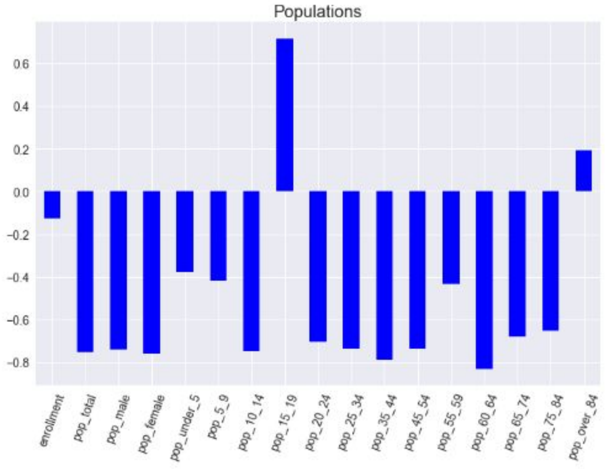

Number of Crimes By Month

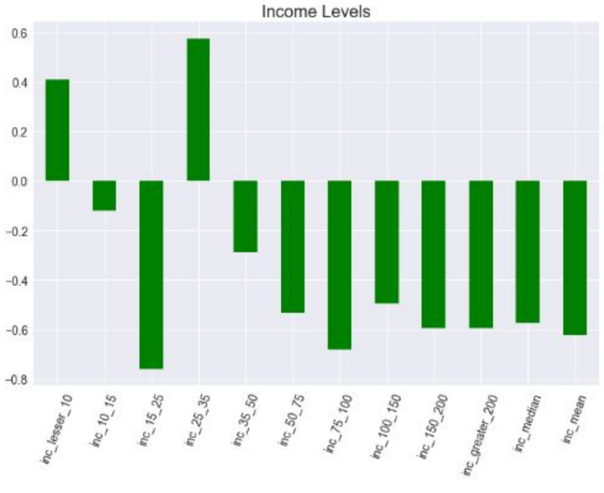
Temporal crime trends were also analyzed. The figure above shows that crimes have overall decreased in frequency since 2010. 2012 is particularly interesting - the spring semester months have a similar number of crimes to the years prior, but an increased proportion due to the drop in crimes overall. The overall drop for 2012 is attributed to the fall semester having a major decrease in crimes for this year. This is the first semester where I believe the Georgia Tech Police Department changed their reporting methods or drastically increased their policing presence, corresponding to the large dip in the crimes in the following years, shown in the aggregated crime over time charts above.



As shown in the figure above, crimes tend to dip significantly in the summer semester months. Crimes also peak on Friday and Saturday, and in general occur most frequently in the evening at late night. As we will be predicting crimes per month, correlations between certain statistics and number of crimes in a given month were also analyzed.





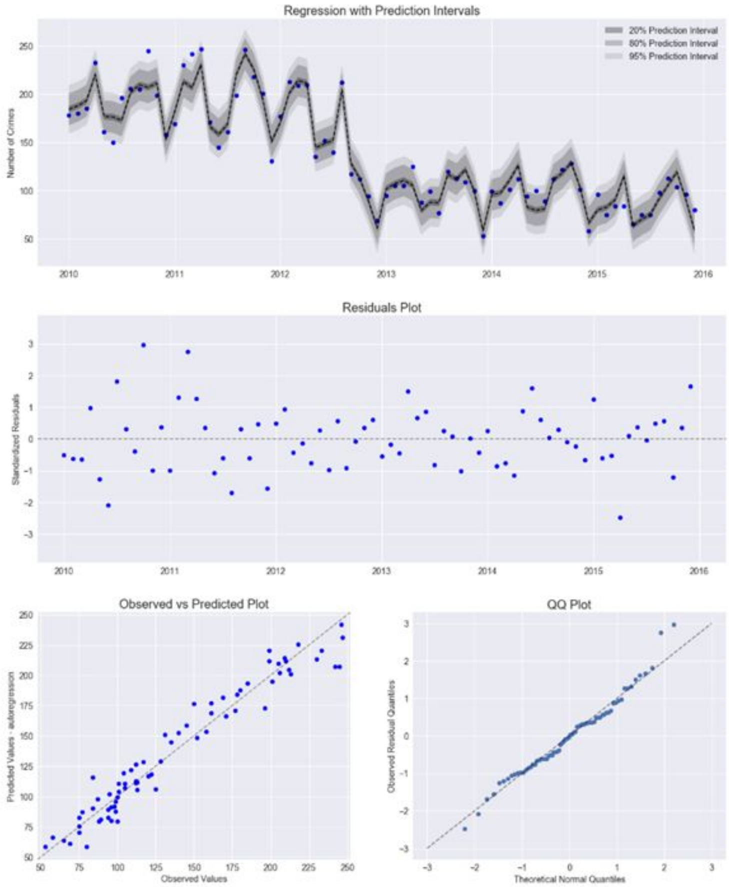


The top chart shows the autocorrelation of the number of crimes per month with the number of crimes for various months in the past. It appears that the autocorrelation chart has periodic spikes in intervals of 6. Since many of these correlations are significant beyond the 99% level (designated by the dashed line), we will want to use these as predictors in our final moel. Finally, correlations with other variables were found. Many of these correlations may be misleading - for instance the enrollment correlation is negative where we might expect a positive one. This is because of the large decrease in crimes reported after the fall semester of 2012 discussed in the analysis above. Still, this tells us what we may want to use as predictors in our model. Further exploratory analysis and statistical analysis can be found in the Exploratory Data Analysis iPython notebook.

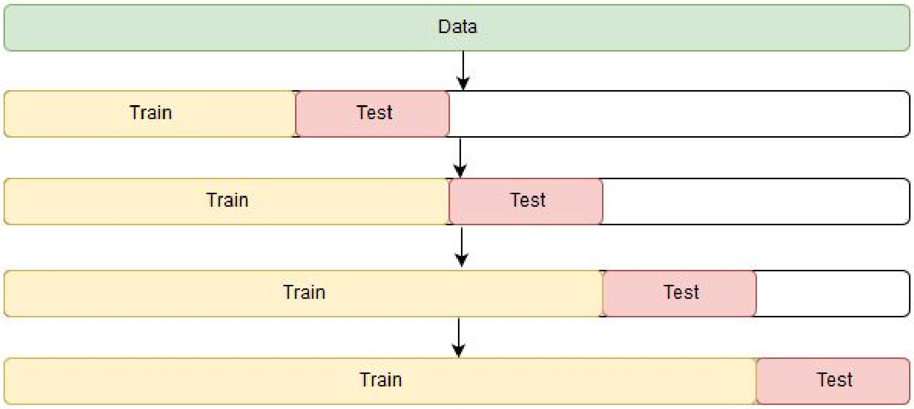
Regression

Before any regression models were trained, feature engineering was performed based on the exploratory data analysis done earlier. Various features were added, including 12 variables of lagged number of crimes, dummy variables indicating the month, and a variable indicating whether the month was before or after the fall semester of 2012 (the large dip in crime), among others. These features were added based on their predictive capability. Next, the data was split into training and test sets to ensure the chosen model would generalize well to unseen data. The training set consisted of the years 2010-2015, and the testing set consisted of 2016-2018.

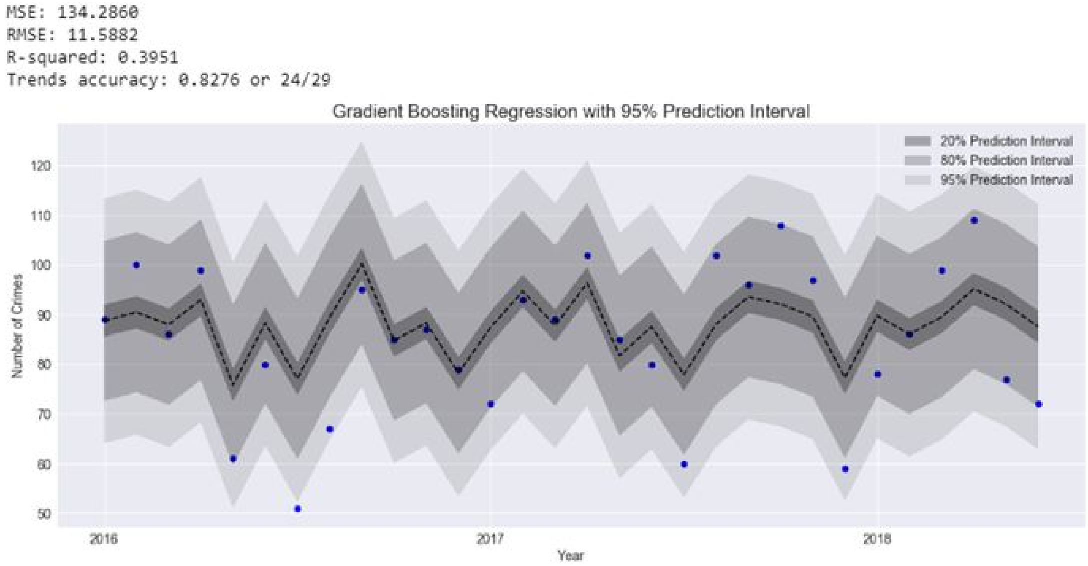
Various regression models were fit to the training data. In total, a linear regression, an elastic net regression, a support vector machine regression, a random forest regression, and a gradient boosting regression were fit. Each of these models was tested to fit various assumptions, including normality and homoscedasticity of residuals, and testing if the model predicted equally well for high and low levels of crime. For models sensitive to overfitting with too many features, recursive feature evaluation was used to maximize adjusted R^2.



Time series grid search cross-validation was used in order to find the best hyperparameters for the models. They were trained on the first 4 years of the training set, and validated on the last 2 years. The combination of hyperparameters that yielded the lowest MSE for each model was returned, and these were used in the final evaluation on the test set.

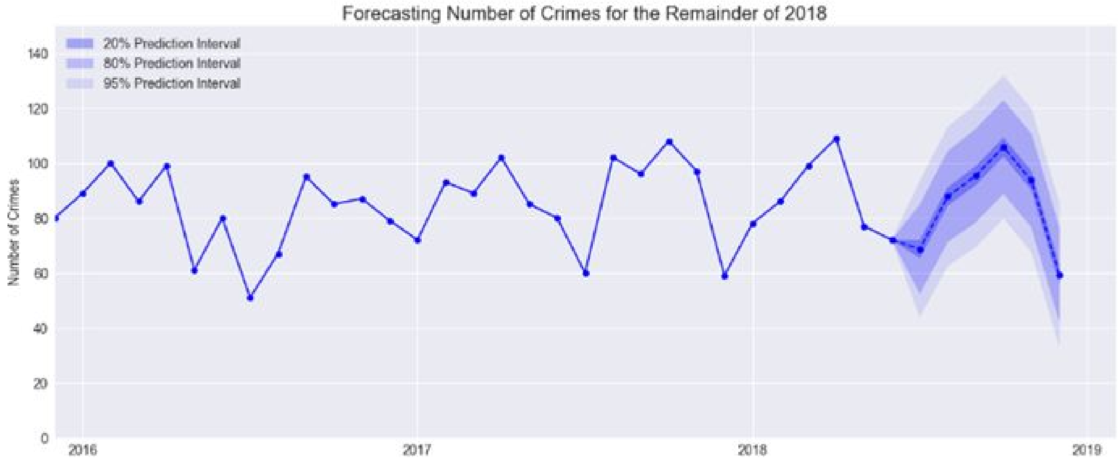


Finally, the models were all evaluated on the test set, using RMSE and trends accuracy (how many times the model correctly predicted whether crime would go up or down in a given month). Gradient boosting performed the best among all the models, with a RMSE of 11.6 and a trends accuracy of 24/29 or 82.8%. In addition, only 1 of the 30 points fell outside the 95% prediction interval, and 9 out of 30 were within the 20% prediction interval.



```
MSE: 134.2860
RMSE: 11.5882
R-squared: 0.3951
Trends accuracy: 0.8276 or 24/29
```

Finally, this gradient boosting model was trained on the entire dataset and then used to predict crime levels for the remainder of 2018.



Conclusion

In this project, we explored the Georgia Tech crime logs, Atlanta demographics statistics, and Georgia Tech enrollment information, and used them to forecast crimes around the Georgia Tech campus. We found that a gradient boosting regression performed the best among all the models tested, with the ability to predict the number of crimes per month with a RMSE of 11.6 and a trends accuracy of 24/29. We also forecasted 6 months of future data.

Although this project chose to answer the question of predicting crime levels for a given month, this data could be used to answer other similar questions. Some of these include building a model to predict when the next crime would occur and what type, or using the location information to plan better police patrolling zones. Finally, this analysis and methodology could also be extended to other areas, provided that they keep a log of crime similar to Georgia Tech's. Analysis like this will be instrumental to helping police and keeping citizens safe around the world.