Data Science Problem

This project attempts to predict crime rates in areas of Washington D.C based on proximity to certain landmarks and public spaces in the city and various demographic factors. The study is important not only to law enforcement, but also to the general population for safety reasons. Previous studies have examined how demographics, housing projects, and alcohol establishments influence crime rates. This particular project will look at these variables that have been studied and those that have little previous research. The study looks at colleges and universities, churches, police stations, bus stations, metro stations, public housing projects, alcohol establishments, and census data. The goal of the project is to determine how proximity to these various places and demographics affect crime rates in Washington D.C. If the study can find how these influence crime rates in the surrounding neighborhoods, then it will be possible to predict crime rates of any location in Washington D.C based on the proximity to the studied variables.

Thomas L. McNulty and Steven R. Holloway from the University of Georgia examined the relationship among race, crime, and public housing in Atlanta. They studied crime rates in predominately black neighborhoods, and correctly hypothesized that high crime rates in such areas are contingent more upon proximity to housing projects, than other socio-economic factors. This relates to our work as this study may be able to use proximity to housing projects as a factor in predicting crime.

A study done at New Mexico Statistical Analysis Center found that while controlling for outside factors, alcohol establishments and certain type of schools increase the amount of crime in surrounding areas. Therefore, alcohol establishments were shown to directly influence the

level of crime in neighborhoods. The study also looked at the relationship between church locations and crime. The researchers hypothesized that churches would have a negative relationship with crime – as they promoted a safe environment. The study found, however, no correlation between the two. (Willits, Dale, Broidy, and Denman). These results provide a basis for constructing hypotheses for this particular study, which examines similar factors. Since the purpose of this project is to predict crime specific to Washington D.C, previous results may not agree.

Data

This study includes a comprehensive dataset of crimes that have occurred in Washington D.C. The dataset includes many variables which are described in the table below:

Variable	Description	Variable Type	Python Variable Type
Crime Data			
BLOCK	Block of Crime	string	string
CENSUS TRACT	Geographic Area defined for purpose of census	numeric	string
END_DATE	End Date of crime	string	string
LATITUDE	Latitude of location of crime	numeric	float
LONGITUDE	Longitude of location of crime	numeric	float
OBJECT_ID	ID Number	numeric	int
OFFENSE	Criminal Offense	string	string
REPORT_DAT	Exact time crime was reported	string	string
START_DATE	Start date of crime	string	string
XBLOCK	X Block of location of crime	numeric	float
XCOORD	X Coordinate of location of crime	numeric	float
YBLOCK	Y Block of location of crime	numeric	float
YCOORD	Y Coordinate of crime	numeric	float
Church Data			

ADDRESS	Address of Church	string	string
NAME	Name of Church	string	string
RELIGION	Religious Affiliation	categorical	string
XCOORD	X Coordinate	numeric	float
YCOORD	Y Coordinate	numeric	float
College Data			
ADDRESS	Address of School	string	string
NAME	Name of School	string	string
XCOORD	X Coordinate	numeric	float
YCOORD	Y Coordinate	numeric	float
Police Station Data			
ADDRESS	Address of Station	string	string
ADDRESS_ID	Identity of Address	numeric	int
GIS_ID	GIS Identification	string	string
Metro Station Data			
ADDRESS	Address of Station	string	string
GIS_ID	GIS Identification	string	string
LINE	Metro Line	string	string
NAME	Name of Station	string	string
OBJECTID	Object Identification	numeric	int
Bus Station Data			
LATITUDE	Latitude of Station	numeric	float
LONGITUDE	Longitude of Station	numeric	float
ON_STR	On Street	string	string
REG_ID	Identification Number	string	string
Alcohol Establishments			
ADDRESS	Address of Building	string	string
ADDRESS_ID	Address Identification	numeric	int
BREW_PUB	Type of Establishment	categorical	string
DANCING	Type of Establishment	categorical	string
ENTERTAINMENT	Type of Establishment	categorical	string
SIDEWALK_CAFE	Type of Establishment	categorical	string
TASTING	Type of Establishment	categorical	string
TRADE_NAME	Type of Establishment	categorical	string

WINE_PUB	Type of Establishment	categorical	string
Public Housing			
ADDRESS	Address of Building	string	string
ADDRESS_ID	Address Identification	numeric	int
FULLADDRESS	Full Address	string	string
LATITUDE	Latitude of Building	numeric	float
LONGITUDE	Longitude of Building	numeric	float
Census Data			
AGE0_17	Population 0-17	numeric	int
AGE18PLUS	Population 18+	numeric	int
AMERIND	American Indian Population	numeric	int
AREA	Size of Census tract	numeric	int
AREASQMI	Square miles in census tract	numeric	float
ASIAN	Asian Population	numeric	int
BLACK	African American Population	numeric	int
CHAMERIND	Change in American Indian Population	numeric	int
CHASIAN	Change in Asian Population	numeric	int
CHBLACK	Change in African American Population	numeric	int
CHHISP	Change in Hispanic Population	numeric	int
CHOTHER	Change in 'Other' Population	numeric	int
CHTOTAL	Change Total Population	numeric	int
CHWITE	Change White Population	numeric	int
CONSTRCOST	Typical home construction cost	numeric	int
FAGI_MEAN_2005	Federally Adjusted Gross Income 2005 (Mean)	numeric	float
FAGI_MEDIAN_2005	Federally Adjusted Gross Income 2005 (Median)	numeric	float
FAGI_TOTAL_2006	Federally Adjusted Gross Income 2006(Total)	numeric	float
FAGI_MEDIAN_2006	Federally Adjusted Gross Income 2006 (Median)	numeric	float
FAGI_TOTAL_2005	Federally Adjusted Gross Income 2005(Total)	numeric	float
FEDTRACTNO	Federal Tracking NUmber	numeric	float
FMHHCHREL	Total Female Household % Change	numeric	int
FMHHCHRELN	Total Female Household Total CHange	numeric	int
FULLTOTAL	Total Population	numeric	int
GIS_ID	GIS ID Number	numeric	int

HAWAIIAN	Hawaiian Population	numeric	int
HISPANIC	Hispanic Population	numeric	int
LEN		numeric	int
MARRIED	Total Married	numeric	int
MHHCHREL	Total Male Household % Change	numeric	int
MHHCHRELNO	Total Male Household Total CHange	numeric	int
MRDCHREL	Total Married % Change	numeric	int
MRDCHRELNO	Total Married Change Total	numeric	int
MULTICOST	Multiple Housing Costs	numeric	int
MULTIPERSO	Multiple Housing Persons	numeric	int
MULTIUNIT	Multiple Housing Units	numeric	int
NAME	Census Tract Name	numeric	string
OBJECTID_1	Object ID	numeric	int
OCCUPIEDHO	Occupied housing units	numeric	int
OTHER	Other	numeric	int
PCTOWNEROC	Percent housing owned by occupant	numeric	int
POPDENSITY	Population Density	numeric	float
SHAPE_Area	Census Tract Shape Are	numeric	float
SHAPE_Length	Census Tract Shape Length	numeric	float
SINGLECOST	Single	numeric	int
SINGLEFAMI	Single Parent Family	numeric	int
SINGLEPERS	Single Parent Total	numeric	int
TOTAL	Total Population 2010	numeric	int
TOTAL00	Total Population 2000	numeric	float
TOTAL90	Total Population 1990	numeric	float
TOTALHOUSEUN	Total Households	numeric	int
TRACT90	Census Tract Number 1990	numeric	float
TRACTNO	Census Tract Number	numeric	float
WHITE	Total White Population	numeric	int

Analysis

Basic Statistical Analysis

Before beginning more complex statistical analyses, it is good practice to examine common descriptive statistics of the data being used. Looking at basic statistics can often guide the hypotheses in the study and help determine the proper techniques to be used. To begin the analysis, the following variables are analyzed:

	Variable	Description	Mean	Median	Mode	Standard Deviation
1	CENSUS_TRACT	Census tract			5800	
2	OFFENSE	Type of Crime			THEFT\OTHER	
3	AGE0_17	Number of people age 17 or younger	612	570		419
4	AGE18PLUS	Number of people age 18 or older	2431	2308		1255
5	MARRIED	Number of married people	301	252		241
6	MONTH	Month of Crime (1 to 12)			5	
7	DATE	Day of Crime (1 to 31)			24	
8	WHITE	Number of white people	847	136		1323
9	POPDENSITY	Population density in area	14415	11762		10593
10	CONSTRCOST	Construction cost in area	570220	0		2644131

1. 5800 is the most common census tract.

2. OFFENSE is categorical so only mode is examined. The most frequent type of crime committed is THEFT/OTHER.
3. The mean number of people under age 17 is 612. The median is 570, with a high standard deviation of 419.
4. The mean number of people over the age of 18 is 2,431. The median is 2,308 with a standard deviation of 1,255.
5. The mean number of married people is 301. The median is 252, with a high standard deviation of 241.
6. The most frequent month in which crimes occur is May.
7. The most frequent date of the month on which crimes occur is the 24th.
8. The mean number of white people in an area is 847. The mean is 136, with a standard deviation is 1,323. Obviously, this data is extremely skewed.
9. The mean population density is 14,415. The median is 11,762 with a standard deviation of 10,593.
10. The mean construct cost in an area is 570,220, while the median is 0. So clearly the distribution is skewed and the standard deviation is 2,644,131.

Data Cleaning

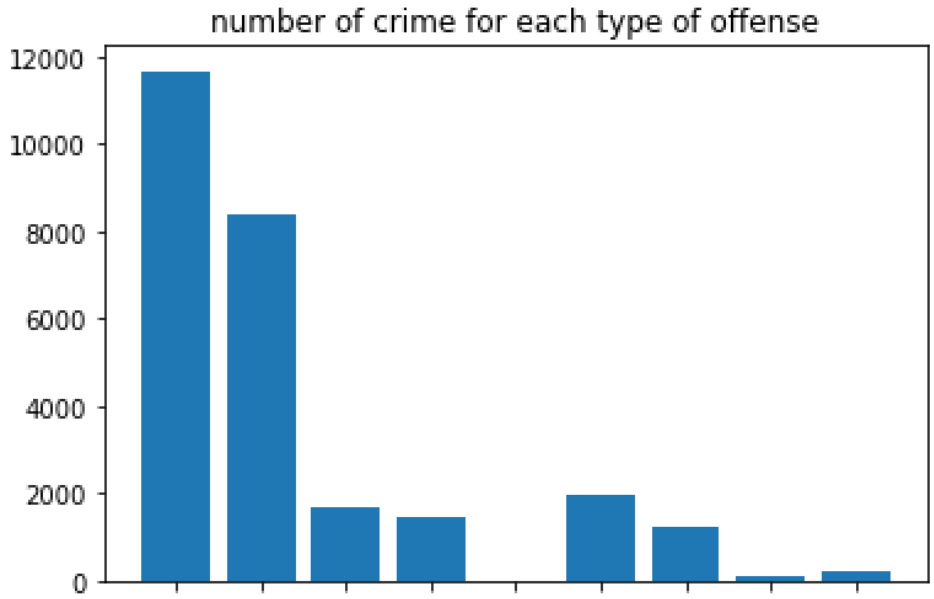
One important part of statistical analysis is the handling of outliers. A key element in this study is that only Washington, D.C. is being studied; no other areas are being examined. This can be checked by looking at the attributes for latitude and longitude. To clean the data, the code checked these two attributes to ensure that all values were in the appropriate range for the city of Washington, D.C. All values in the dataset fell in the correct range for the city. The dataset did have 1,243 rows of data without an end date to the crime. Since this is only 3% of the total data, removing these rows in not an issue.

Another key element of preliminary statistical analysis involves data manipulation. One method of modifying data involves binning continuous data into categorical data. For this study,

the variable TOTAL, representing the total population for each area, was binned. This done in the code via min and max method with an even space of 1,000. This makes sense for the data because population numbers can vary widely and the precise number is not important.

Histograms

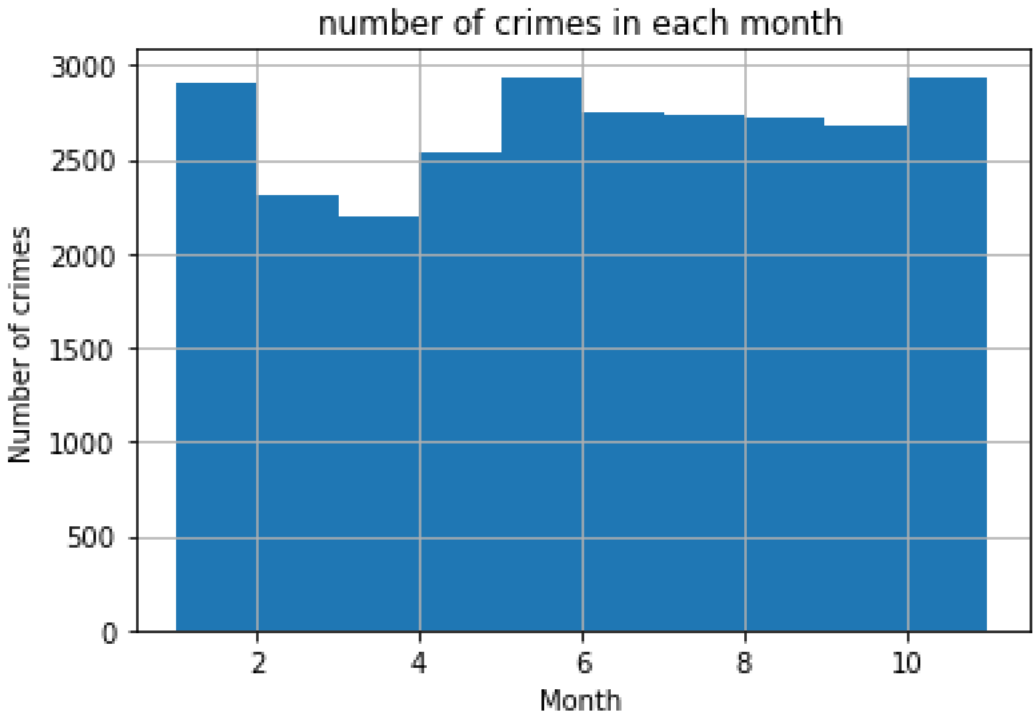
The variable OFFENSE can be analyzed via a histogram:




```
['THEFT/OTHER', 'THEFT F/AUTO', 'ROBBERY', 'ASSAULT W/DANGEROUS
WEAPON', 'ARSON', 'MOTOR VEHICLE THEFT', 'BURGLARY', 'HOMICIDE', 'SEX
ABUSE']
```

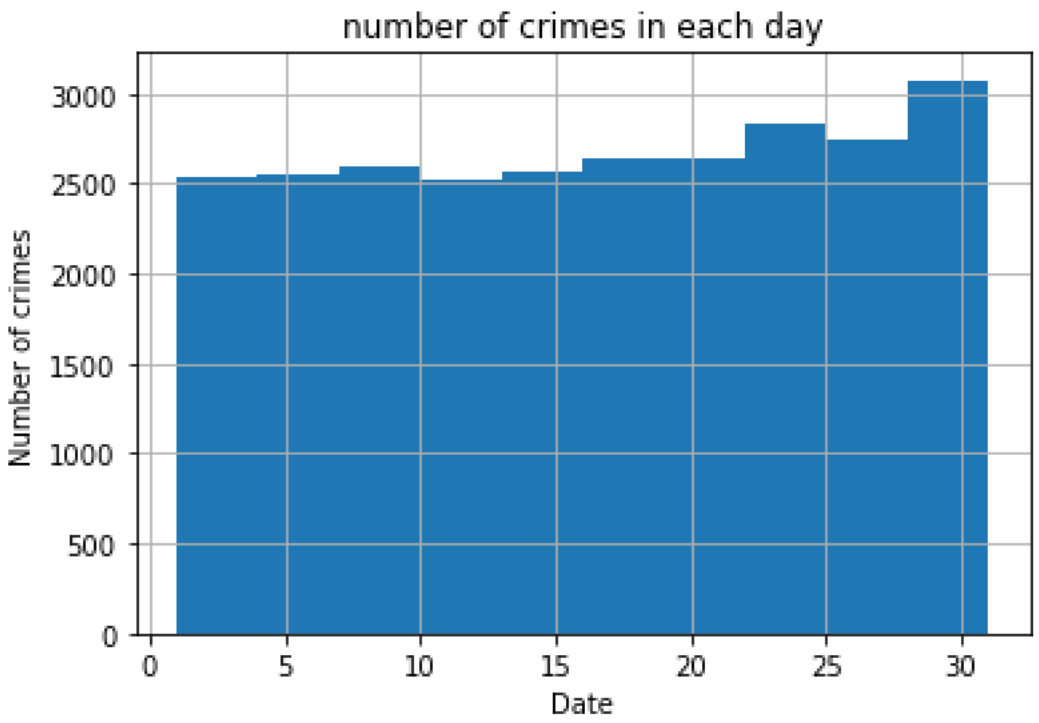
The values of ‘THEFT/OTHER’ and ‘THEFT F/AUTO’ have the highest values. This makes sense that such crimes would be more common in D.C. than other more serious or more violent crimes. It should be noted that are nearly no crimes in the dataset for ‘ARSON.’

The next variable analyzed via histogram is ‘MONTH.’ This is rather interesting as it provides insight into when crimes take place during the year.



This shows that the months with the most crime include January, May, and December. The exact reason is not a vital part of this study. This could be because of holiday seasons, changes in weather, or the beginning of summer.

Next, date is examined:



The histogram suggests that crimes happen more towards the end of a given month.

The following 3 variables from the census data will examined for correlation (The variables are labeled V1, V2, V3 accordingly):

OCCUPIEDHO	Occupied housing units	numeric	int
MARRIED	Total Married	numeric	int
WHITE	Total White Population	numeric	int

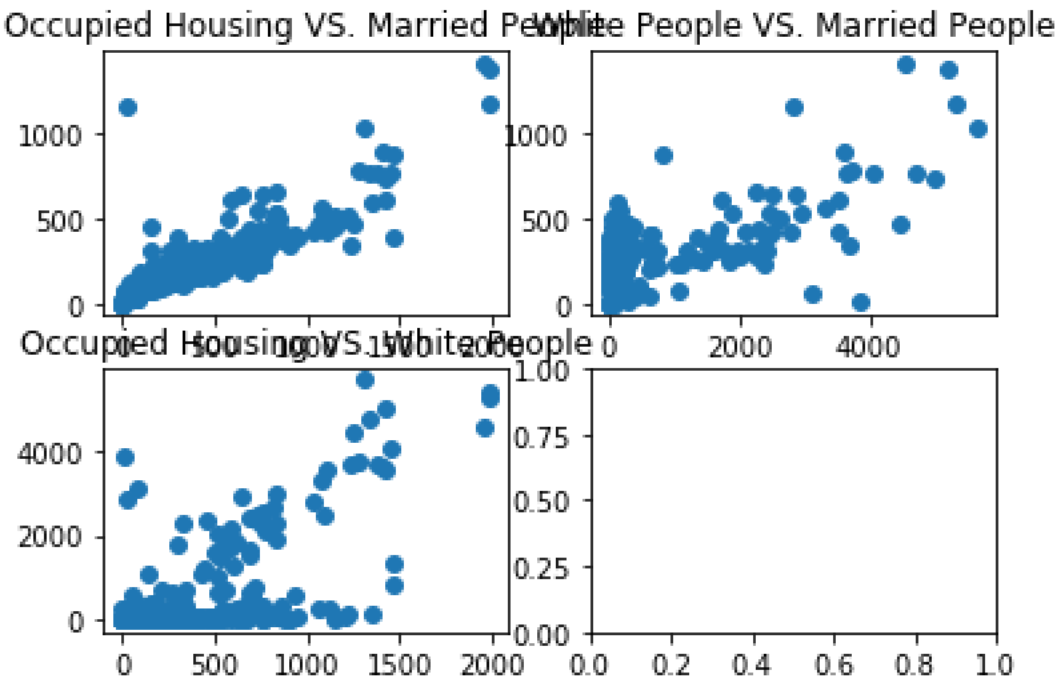
The following correlation table was obtained from the code in this analysis:

<Table length=1>

Cor betw V1 and V2	Cor betw V2 and V3	Cor betw V1 and V3
float64	float64	float64
0.827742007147	0.69555525982	0.610488055777

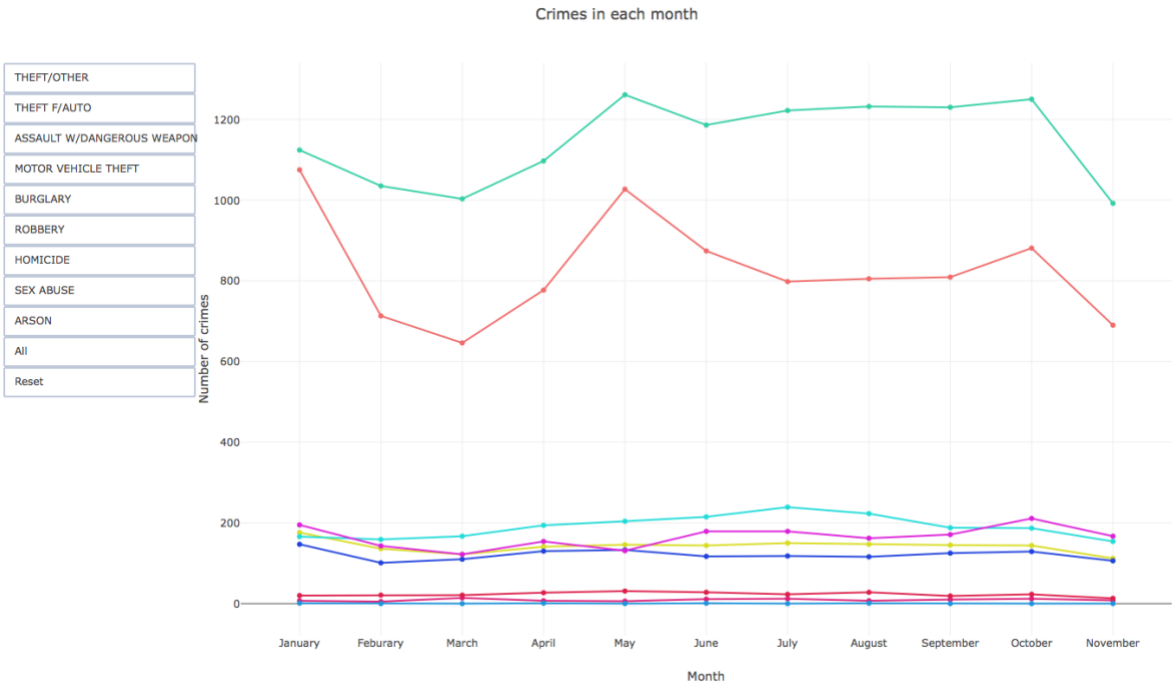
Occupied housing and married have pretty high positive correlation of .828. This is not surprising that these two would be correlated. Married and White have a correlation of .696, which is moderately positive. This suggests that there is a positive relationship between the number of white people and the number of married people in an area. This could just be due to differences in population numbers. Occupied housing and White have a fairly positive correlation of .610. This is not too strong but indicates that there may be a positive relationship between the amount of occupied housing and number of white people in an area.

These relationships are further shown in the plots below:



Visuals and Interactive Plots

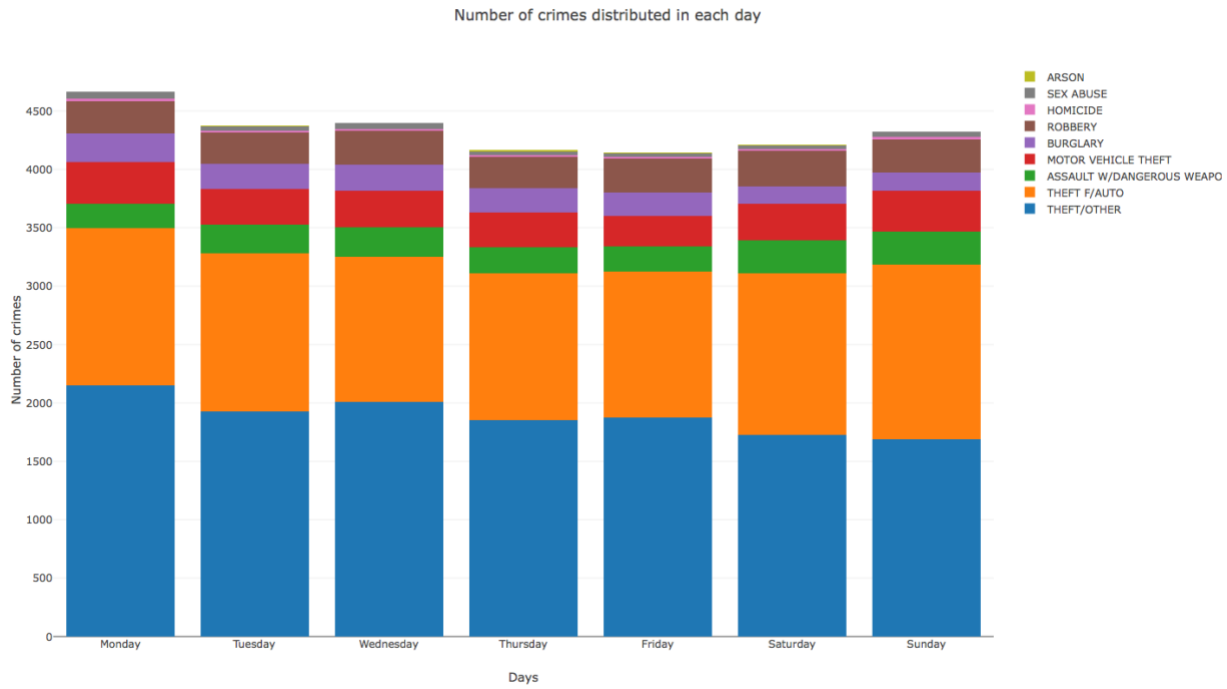
Type of crimes committed in each month



Obviously, theft is the most committed crime. The visual also shows again that May seems to have more crimes than other months.

https://plot.ly/~jiaweyu/25.embed

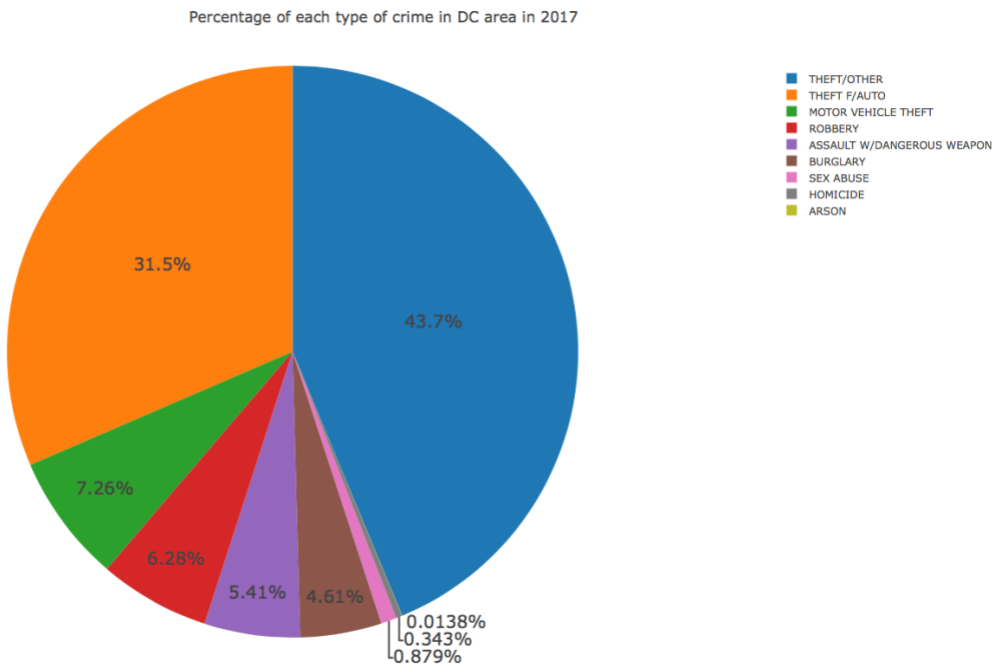
Type of crimes committed each day of the week



Monday appears to have the highest crime rate. The type of crimes appears to be consistent for each day of the week.

https://plot.ly/~jiaweyu/23.embed

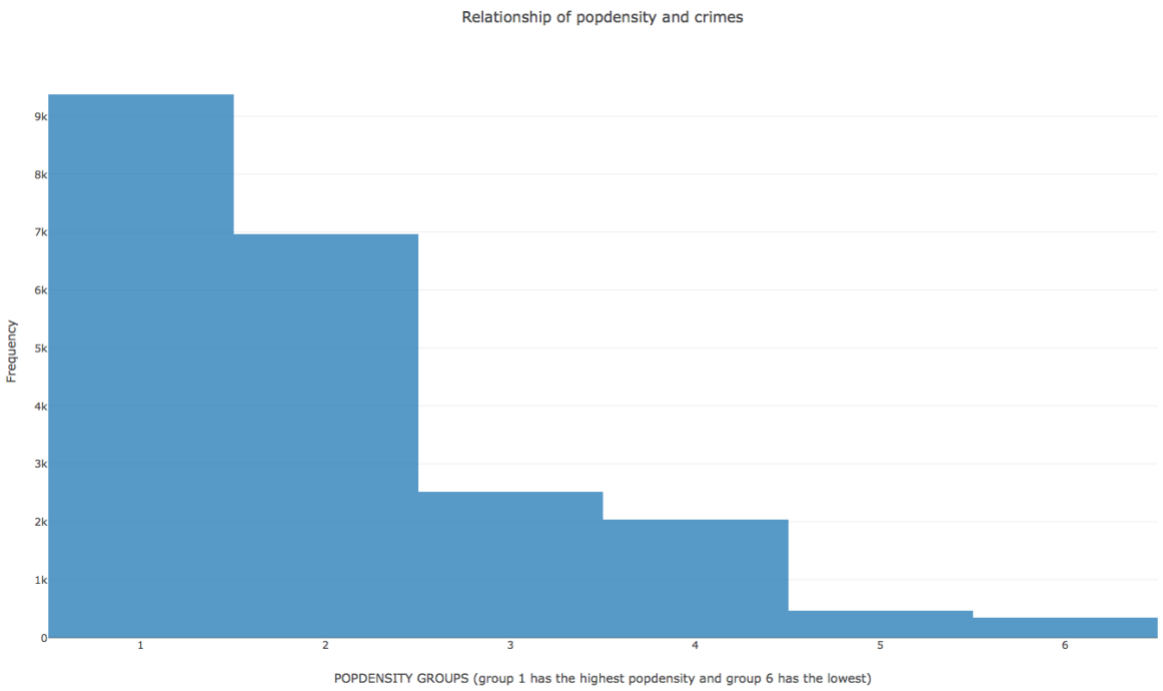
Pie chart of crime types



Theft is the most common crime, while less than 1% of crimes are sex abuse, homicide, or arson.

https://plot.ly/~jiaweyu/19.embed

Relationship between population density and crime



This graph simply shows that higher population density correlates with higher crime rates, which is not surprising.

https://plot.ly/~jiaweyu/21.embed

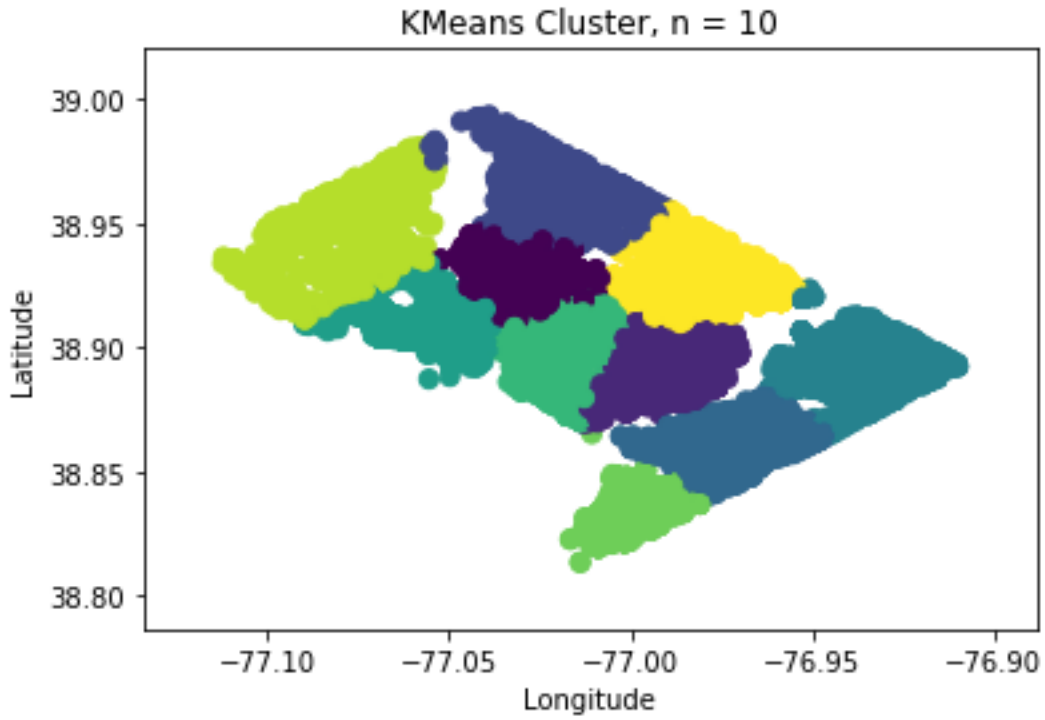
The following link provides access to interactive plots in Tableau:

https://public.tableau.com/profile/jiawei.yu#!/vizhome/Project3-Group3/Story1?publish=yes

One of the most interesting facts taken away from the Tableau story is that certain crimes are more common in specific areas of D.C. For example, there is only one homicide anywhere close to Georgetown University, while other parts of D.C. have several homicides. Other crimes like burglary are more common in areas of the city with high population density. In general, it seems that Northwest D.C. is the safest part of the city.

Cluster Analysis

Next, cluster analysis will be performed on the data. First, k-means clustering will be performed with 10 clusters. The analysis will attempt to cluster crimes based on location via latitude and longitude.



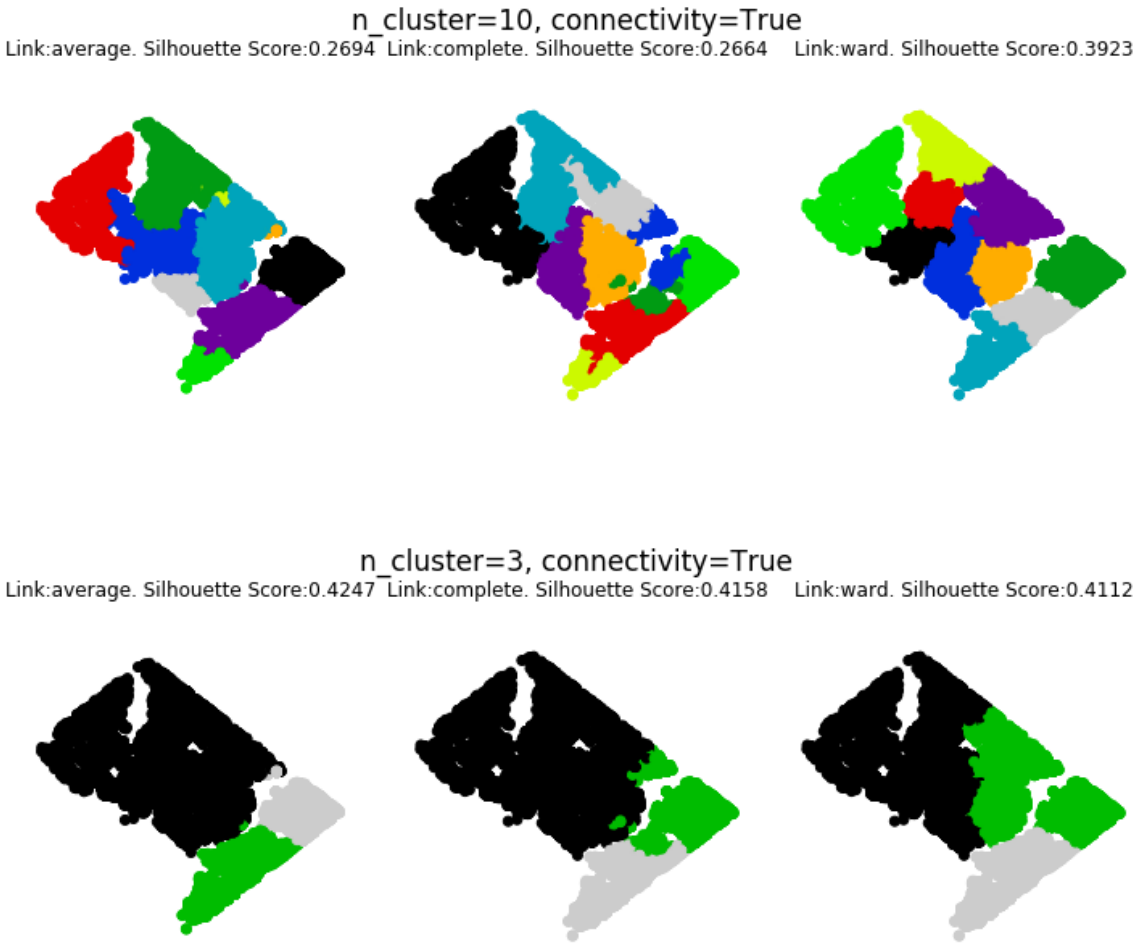
Silhouette Coefficient: 0.426

This Silhouette Coefficient is not high enough to draw any major conclusions.

When clustering crimes based on Latitude and Longitude, there is not much that can be taken away from the results. With more than 30,000 points clustered around DC, the data is just too dense to make any real analysis about it. As expected, Anacostia is separated from DC proper

thanks to the Anacostia River (creates a geographic barrier). Interesting note is that no crimes are reported within Rock Creek Park, or the US Naval Observatory.

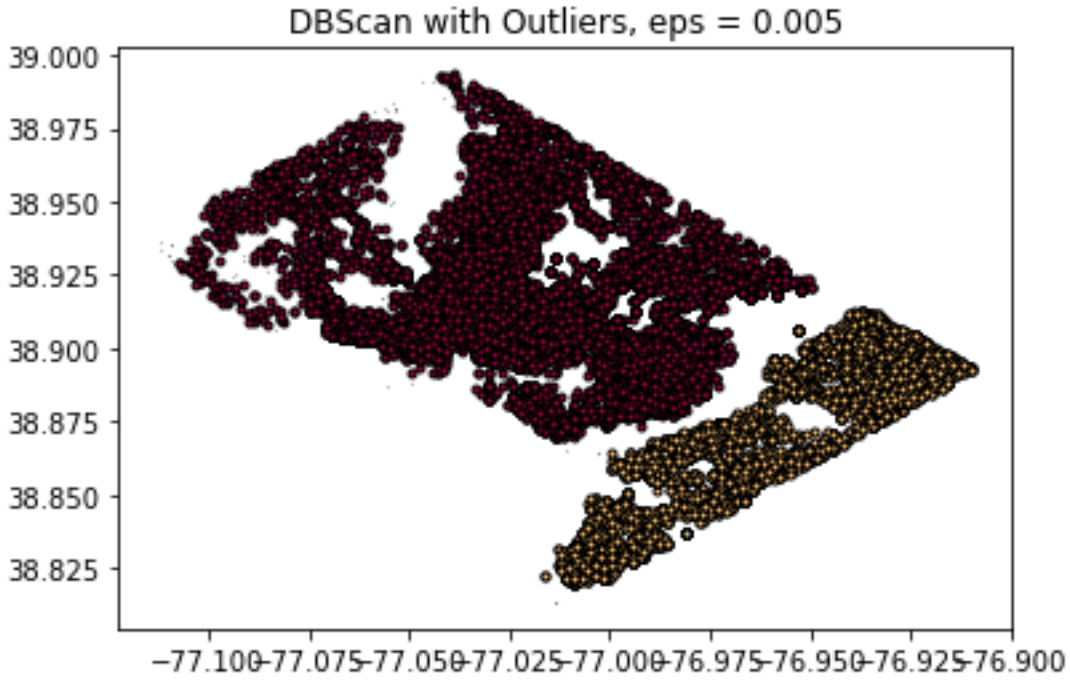
Next, this study considers Hierarchical clustering with 10 and 3 clusters.



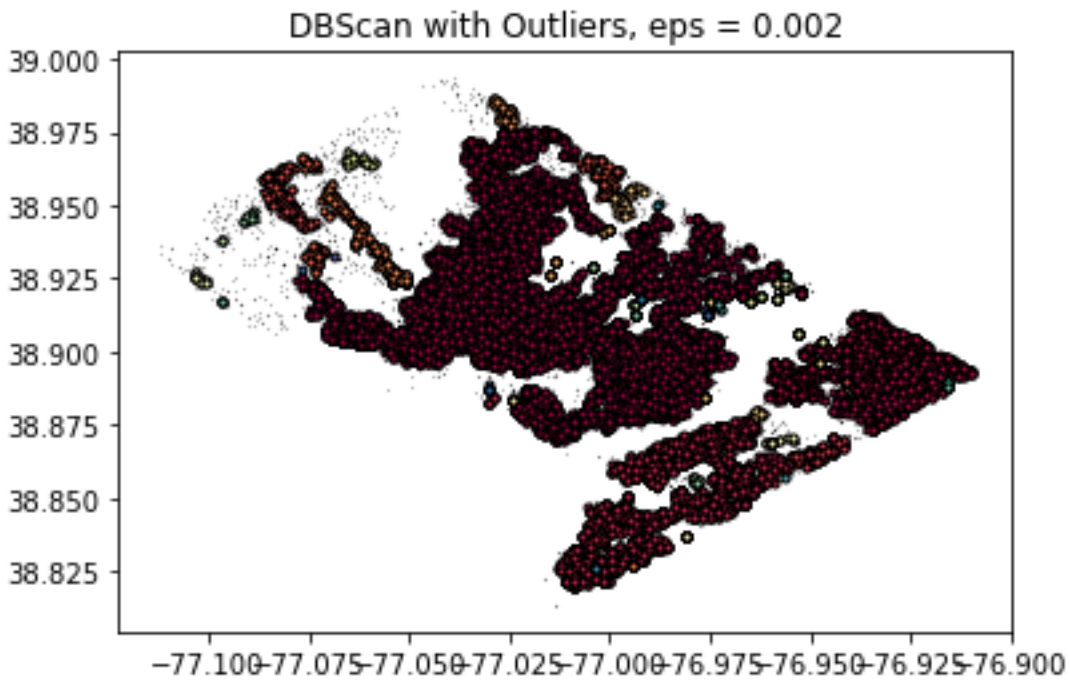
The results turn out like those of the k-means clustering. The data is a bit too dense for Hierarchical clustering to produce useful results. With 10 clusters, no meaningful conclusions are

apparent. At 3 clusters, average and ward linkage break up the crimes more interestingly. Average considers DC as one big mass, and ward considers Anacostia as one big mass.

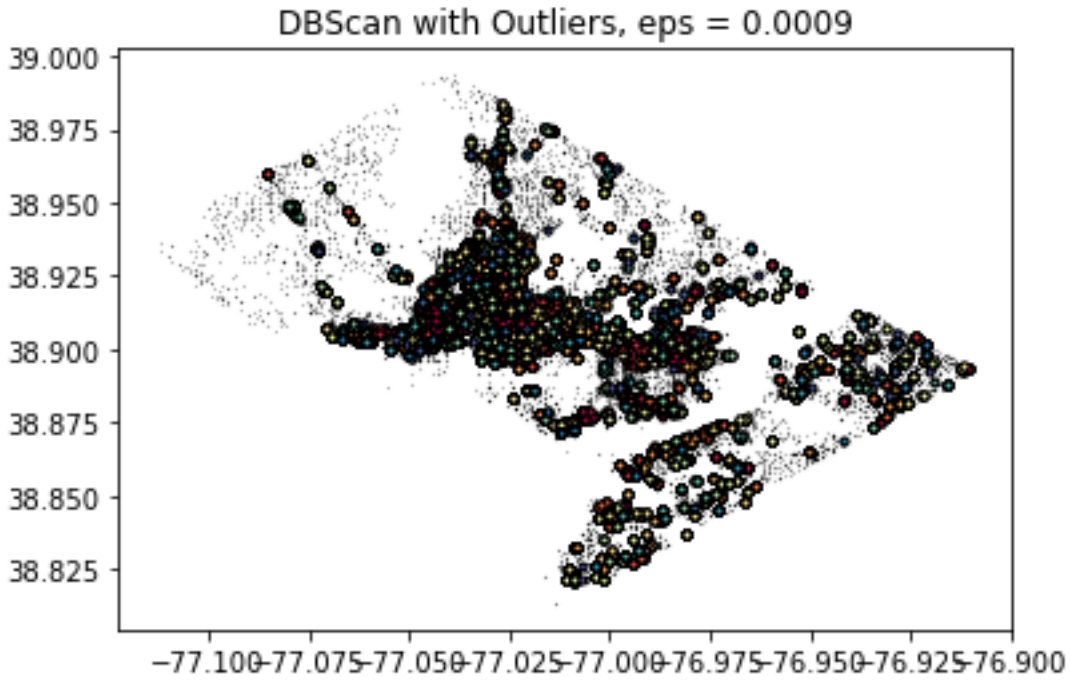
The next method used is DBScan:



Silhouette Coefficient: 0.158



Silhouette Coefficient: -0.479



Silhouette Coefficient: 0.042

Density clustering finally shows us how dense the points in the dataset are. It took an ep = 0.005 to find any different cluster (which of course is DC versus Anacostia again). Even though the silhouette score drops when we use an ep = 0.0009, it starts to remove some "outliers". Some interesting visuals begin to form here. It is noteworthy that at around (38.95, -77.075) there are two distinct lines leaving the city (Wisconsin and Connecticut Av.) and another line at (38.95, -77.025, Georgia Ave.). This is the first evidence to suggest that information about transportation could be relevant to predicting to crimes.

Association Rules / Frequent Itemset Mining Analysis

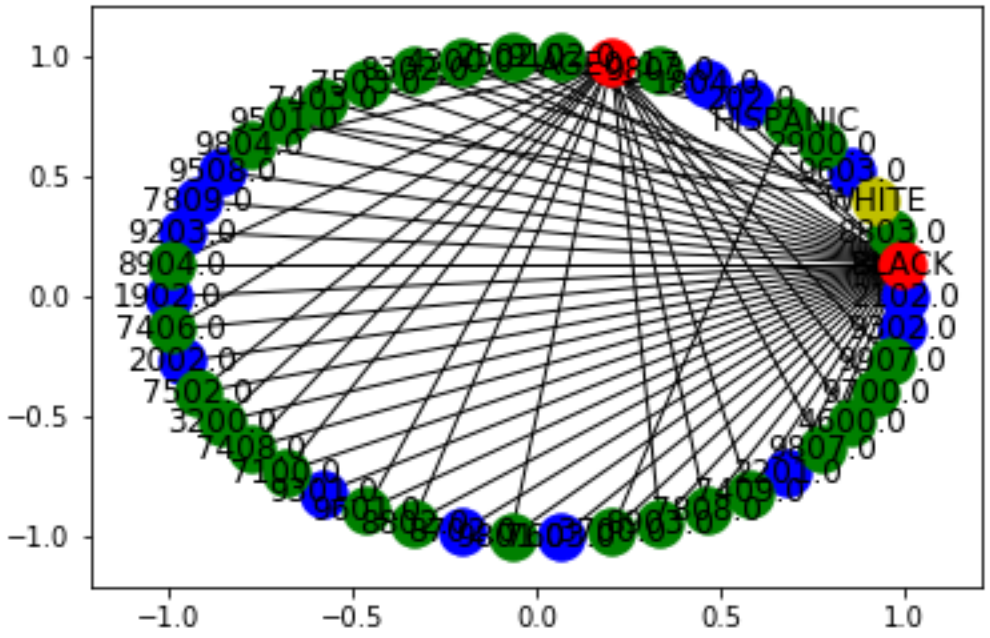
Next the study will consider the possibility of association rule and frequent itemset mining analysis. Given the size of the data and the vast array of attributes and factors that may contribute to crime, this analysis is a bit trickier than the previous methods considered.

Min Support = 30%	
THEFT F/AUTO	0.30795
THEFT/OTHER	0.4394
Min Support = 20%	
Same as minimum support 30%	
Min Support = 10%	
Friday	0.13562
Monday	0.15252
Saturday	0.13712
Sunday	0.14247
Thursday	0.13755
Tuesday	0.14699
Wednesday	0.14771

Apriori did not yield any significant results from the dataset. When given the set {Day of week (binned from Date), Crime, Census Tract} (Census Tract was even sub-stringed from our digits to 2, to halve the number of total tracts), it never found a set of more than one items which had minimum support higher than 10%. Thus, this method did not produce helpful results for the study. Again, this is not incredibly surprising given the nature of the dataset.

Network Analysis

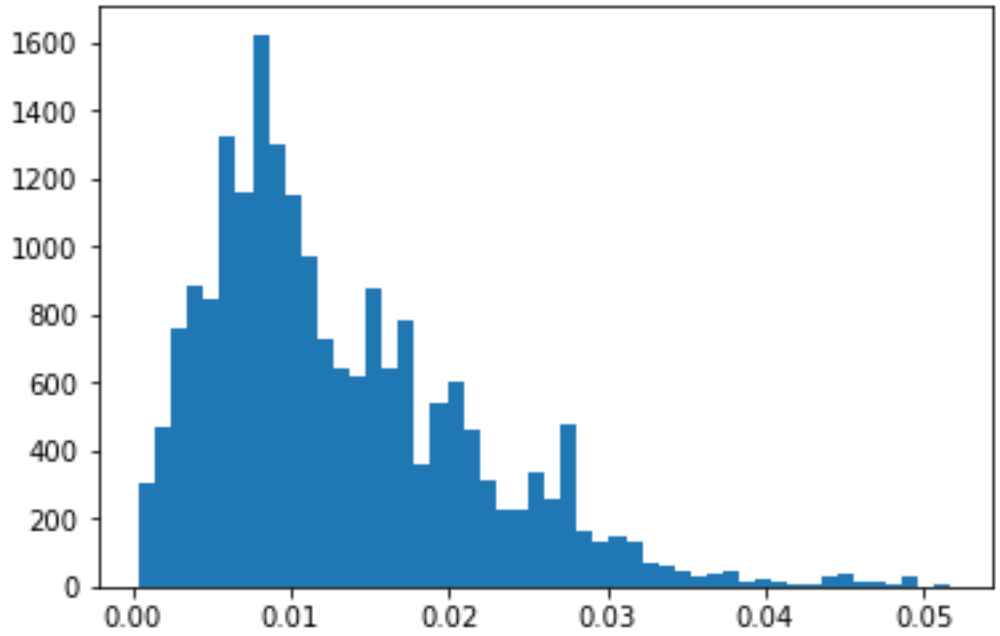
Next, network analysis was considered on the data. Because of the nature of the data and its size, some filtering was done so that a useful network could be created. The network looked at census tracts and considered races in each tract that comprised at least 25% of the population. These percentages were used as weights in the network. Then the number of each type of crime committed were counted. The only mild conclusion we can get from this is that there are a higher number of homicides in census tracts that have a higher black population and a large population of people under 17. This group is followed by census tracts that are largely white.



Hypothesis Testing and Results

1) There are fewer crimes in close proximity to police stations and more crimes close to housing projects.

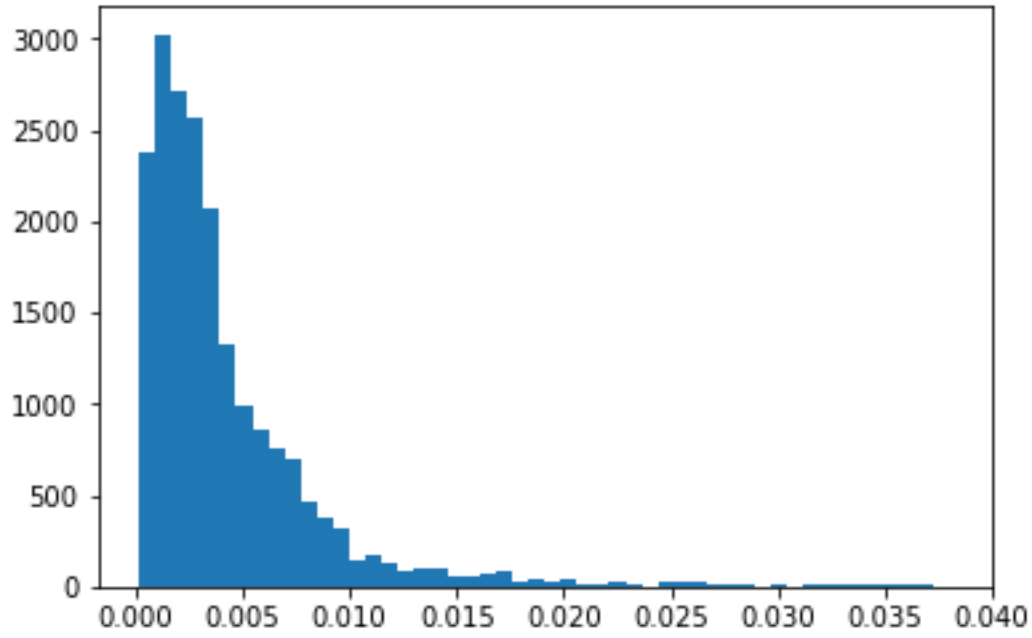
Here we take the Euclidean distance from crimes to police stations. The closest police stations for each crime is used. The following histogram shows the distance from each crime to the nearest police station:



(Note x-axis is distance, y-axis is number of crimes)

This produced an interesting result. The number of crimes in the immediate vicinity of police stations is somewhat low, but as the distance increases, then so does the number of crimes. After the peak, however, there is a significant downward trend. In fact, the histogram shows that the majority of crime happens fairly close to police stations. This result is not quite what was anticipated in the hypothesis. Further research is likely needed to discover the exact cause of this result.

Next, the same technique was used on housing projects. Recall that prior research done in Atlanta showed that crime was more prevalent in areas near housing projects. The methods previously described produced the following results for housing projects:



(Note x-axis is distance, y-axis is number of crimes)

Here the results are quite obvious. Areas in close proximity to housing projects have higher crime rates. This result is consistent with both previous research and the hypothesis of this study.

2) Given data on demographics such as racial makeup, population density and various cost measures in an area, a model can be constructed to predict type of crime in that area.

Given: {'AGE0_17','AMERIND','ASIAN','BLACK','FAGI_MEAN_2005', 'HAWAIIAN','HISPANIC','MARRIED','POPDENSITY','SINGLECOST', 'MULTICOST','WHITE'}

Predict: {'THEFT','THEFT F/AUTO','MOTOR VEHICLE THEFT','ROBBERY','BURGLARY','ASSAULT W/DANGEROUS WEAPON','SEX ABUSE','HOMICIDE','ARSON'}

K nearest neighbor did a decent job, with an accuracy of 44% on the test set. Note, however that it could not correctly predict a single homicide or arson out of training:

```
Training Set:
KNN: Number of mislabeled points out of a total 16824 points : 9079.
46.0354% accuracy
[[5527 1658  112  118   28   61    0   59    0]
 [2879 2092  102   58   20   44    0   22    0]
 [ 670  369   40   24   12   51    0    5    0]
 [ 636  317   37   26    6   19    0   21    0]
 [ 467  198   10   14   12    6    0    5    0]
 [ 523  265   18   19   14   45    0   24    0]
 [  96   34    2    1    2    0    0    2    0]
 [  27   11    2    4    0    4    0    3    0]
 [   1    2    0    0    0    0    0    0    0]]
Testing Set:
KNN: Number of mislabeled points out of a total 4206 points : 2355.
43.9848% accuracy
[[1333  421   27   28    5   17    0   10]
 [ 736  481   36   22   12   17    0    8]
 [ 166  111   16    5    5   12    0    5]
 [ 172   87   12   11    1    6    0    1]
 [ 113   64    5    5    1    4    0    4]
 [ 107   66    6    4    4    8    0    4]
 [  19   12    0    4    1    1    0    0]
 [   8    2    0    1    0    0    0    0]]
```

Next, Decision tree had high accuracy, but realistically only predicted theft, auto theft, burglary, and assault with a deadly weapon.

```
Training Set:
CART: Number of mislabeled points out of a total 16824 points : 8310.
50.6063% accuracy
[[6078 1436    0    0   16   33    0    0    0]
 [2803 2379    0    0   13   22    0    0    0]
 [ 753  383    0    0    4   31    0    0    0]
 [ 707  329    0    0   14   12    0    0    0]
 [ 470  215    0    0   22    5    0    0    0]
 [ 633  235    0    0    5   35    0    0    0]
 [ 101   36    0    0    0    0    0    0    0]
 [  36    9    0    0    2    4    0    0    0]
 [   1    2    0    0    0    0    0    0    0]]
Testing Set:
CART: Number of mislabeled points out of a total 4206 points : 2214.
47.3609% accuracy
[[1447  379    0    0    5   10    0    0]
 [ 759  536    0    0    4   13    0    0]
 [ 184  129    0    0    1    6    0    0]
 [ 191   91    0    0    4    4    0    0]
 [ 121   68    0    0    3    4    0    0]
 [ 139   53    0    0    1    6    0    0]
 [  20   16    0    0    0    1    0    0]
 [   8    3    0    0    0    0    0    0]]
```

Naïve Bayes obviously will not work well since the variables are highly unlikely to be independent. It is interesting that this method attempts to (incorrectly) predict most crimes as arson, which is not heavily represented in the dataset.

```
Training Set:
NB: Number of mislabeled points out of a total 16824 points : 14877.
11.5728% accuracy
[[1082  973    0    0    0    0    0  365 5143]
 [ 596  855    0    0    0    0    0  369 3397]
 [  73  146    0    0    0    0    0  146  806]
 [ 106  117    0    0    0    0    0   85  754]
 [  56  108    0    0    0    0    0   78  470]
 [  48   83    0    0    0    0    0  122  655]
 [   9   24    0    0    0    0    0    9   95]
 [   0    1    0    0    0    0    0    7   43]
 [   0    0    0    0    0    0    0    0    3]]
Testing Set:
NB: Number of mislabeled points out of a total 4206 points : 3693.
12.1969% accuracy
[[ 281  237    0    0    0    0    0   84 1239]
 [ 165  232    0    0    0    0    0   87  828]
 [  23   35    0    0    0    0    0   52  210]
 [  19   31    0    0    0    0    0   27  213]
 [  22   29    0    0    0    0    0   15  130]
 [   7   17    0    0    0    0    0   22  153]
 [   2    3    0    0    0    0    0    5   27]
 [   0    1    0    0    0    0    0    0   10]
 [   0    0    0    0    0    0    0    0    0]]
```

SVM had a decent accuracy, but it placed crimes into either theft, or auto theft.

Recall from our initial analysis that theft was the most well-represented in the dataset.

```
Training Set:
SVM: Number of mislabeled points out of a total 16824 points : 9198.
45.3222% accuracy
[[6967  596    0    0    0    0    0    0    0]
 [4559  658    0    0    0    0    0    0    0]
 [ 967  204    0    0    0    0    0    0    0]
 [ 952  110    0    0    0    0    0    0    0]
 [ 585  127    0    0    0    0    0    0    0]
 [ 775  133    0    0    0    0    0    0    0]
 [ 121   16    0    0    0    0    0    0    0]
 [  43    8    0    0    0    0    0    0    0]
 [   3    0    0    0    0    0    0    0    0]]
Testing Set:
SVM: Number of mislabeled points out of a total 4206 points : 2357.
43.961% accuracy
[[1696  145    0    0    0    0    0    0]
 [1159  153    0    0    0    0    0    0]
 [ 254   66    0    0    0    0    0    0]
 [ 263   27    0    0    0    0    0    0]
 [ 168   28    0    0    0    0    0    0]
 [ 174   25    0    0    0    0    0    0]
 [  32    5    0    0    0    0    0    0]
 [  10    1    0    0    0    0    0    0]]
```

Random Forrest produced similar results to SVM with fairly high accuracy, but mostly just predicts theft and auto theft.

```
Training Set:
RF: Number of mislabeled points out of a total 16824 points : 8329.
50.4933% accuracy
[[6045 1502    0    0   16    0    0    0    0]
 [2776 2428    0    0   13    0    0    0    0]
 [ 771  396    0    0    4    0    0    0    0]
 [ 703  345    0    0   14    0    0    0    0]
 [ 474  216    0    0   22    0    0    0    0]
 [ 662  241    0    0    5    0    0    0    0]
 [ 103   34    0    0    0    0    0    0    0]
 [  41    8    0    0    2    0    0    0    0]
 [   2    1    0    0    0    0    0    0    0]]
Testing Set:
RF: Number of mislabeled points out of a total 4206 points : 2213.
47.3847% accuracy
[[1444  384    7    0    5    1    0    0]
 [ 763  542    2    0    4    1    0    0]
 [ 193  121    3    0    1    2    0    0]
 [ 191   91    3    0    4    1    0    0]
 [ 121   70    1    0    3    1    0    0]
 [ 142   53    2    0    1    1    0    0]
 [  22   15    0    0    0    0    0    0]
 [   7    4    0    0    0    0    0    0]]
```

A linear regression was also used using the same data. The results not meaningful. The regression only produced an r^2 value of .05, which is abysmal. This is somewhat surprising. Since our earlier Naïve Bayes analysis performed poorly, we could assume the attributes were dependent. We could therefore assume our predictors are correlated, and this should inflate our r^2, which of course, it does not.

Finally, the distance metrics previously used were also included with the demographic data to test if a model could predict crime type. Crime type was expressed as binary (either that type or not). For example, a model was trained and tested on whether or not each crime was theft. Theft had the highest accuracy using k nearest neighbor and random forest.

```
Training Set:
KNN: Number of mislabeled points out of a total 11962 points : 2549. 78.6825% accuracy
[[7221  942]
 [1608 2191]]
Testing Set:
KNN: Number of mislabeled points out of a total 7976 points : 2274. 71.4769% accuracy
[[4520  926]
 [1349 1181]]
Training Set:
RF: Number of mislabeled points out of a total 11962 points : 1861. 84.4424% accuracy
[[7470  693]
 [1168 2631]]
Testing Set:
RF: Number of mislabeled points out of a total 7976 points : 2321. 70.9002% accuracy
[[4433 1013]
 [1308 1222]]
```

Next, burglary was predicted well using decision trees.

```
Training Set:
CART: Number of mislabeled points out of a total 7975 points : 222. 97.2163% accuracy
[[7615    6]
 [ 216  138]]
Testing Set:
CART: Number of mislabeled points out of a total 11963 points : 696. 94.1821% accuracy
[[11235   229]
 [  467    32]]
```

These results suggest that it may be possible to predict certain types of crime using machine learning techniques.

3) Demographic data, such as racial makeup, affects average income in Washington D.C.

Income is often used to predict crime. Here a t-test was performed to see if demographics in D.C. affect income. This test could have important implications because varying levels of income may also see varying levels of crime. This test compared the income of areas that were more than 50% white and areas that were more than 50% black. The statistic is 13.54, with a p-value of

1.014e-28. Thus, we have evidence to reject our null hypothesis (of the incomes being the same) and see that the incomes are statistically different.

4) Theft crimes depend on location in Washington D.C., as measured by latitude and longitude and census tract

Here we perform several machine learning tests, in order to examine the relationship. From the code, the following results are obtained:

Train Data:

KNN: 0.815197 (0.023657)

CART: 0.755523 (0.025743)

NB: 0.823530 (0.026250)

SVM: 0.836044 (0.032844)

RF: 0.792251 (0.023768)

Test Data:

KNN 0.803284807764

CART: 0.768010451661

NB: 0.818029115342

SVM 0.826427771557

RF 0.786487495334

The accuracy for each algorithm is high, with each being at least 75%. It is important to note here that crimes are being predicted as either theft or non-theft crimes. Because most crimes are of this type, some algorithms may achieve high accuracy based on a high probability of being a theft crime. But based on these results, D.C. theft crime may be linked to location. This makes our Tableau visual of theft based on location more insightful.

Conclusion

Our analysis shows that predicting crime is a very difficult task. The data on crime utilized in this study was so dense that clustering algorithms had little success. Our results did show that numerous crimes occur along the busiest streets in the city. Some machine learning algorithms had some moderate success in predicting type of crime, but it could be due to predicting the most frequent cases nearly every time. Association Rules proved to be of little use in this study, as no real results amounted from its implementation. Using regression, the study was not able to predict type of crime based on demographic factors in an area, suggesting that there are many other factors contributing to crime. The success in this study came when examining how proximity to police stations and housing projects affects crime. The study showed there is indeed correlation between these factors and crime rate. Crime rates are higher near police stations but then decrease with high distance from the stations. Locations in close vicinity to housing projects demonstrated higher rates of crime. Additionally, utilizing these

distances along with demographic factors, some machine learning techniques achieved high accuracy.

Further study is likely needed in order to establish a model for predicting crime. Given the number of variables that affect crime, this is not an easy task. This study, however, showed that distance to certain locations affects crime rates. This study was not able to consider the many factors that were originally thought of when collecting data. Further research that included these factors, along with others (such as churches, schools, transportation stops) would aid in prediction. One major drawback of this study was that it examined crime only in one major city. The reality is that there are many crimes that take place in Washington D.C., which is a small area with high population density. Thus, the data was extremely dense. The clustering methods showed that some form of crime happens almost everywhere. Gaining a better understanding of how many variables affect the thousands of crimes that happen every year would require very extensive research with careful (and likely very slow) algorithms. Fortunately, this study points to successful modeling being a possibility, just an extremely difficult one.

Works Cited

Mcnulty, Thomas L., and Steven R. Holloway. "Race, Crime, and Public Housing in Atlanta: Testing a Conditional Effect Hypothesis." *Social Forces* 79.2 (2000): 707. Web.

Willits, Dale, Lisa Broidy, and Kristine Denman. "Schools, Neighborhood Risk Factors, and Crime." *Crime & Delinquency* 59.2 (2013): 292-315. Web.

This is a group project, and the members list below:

Alex Archer
Alek Traczyk
Hanlong Peng
Jiawei Yu