Paper 2022-2017

CRIME IN PHILADELPHIA

Edwin Baidoo
Christina Jones
Muniza Naqvi
Kennesaw State University

Introduction and Problem

Crime is a serious problem in cities around the country. Our project focused on crime in the city of Philadelphia. The data we looked at came from opendataphilly.org. It is the open data repository for the city of Philadelphia and has additional data from other sources in the region. In doing this project we hoped to find patterns in crime occurrences that help law enforcement official decrease the number of crimes that the City of Philadelphia experiences. More specifically, we wanted to determine when, where, and what types of crimes were most prevalent as well as other factors that affected crime occurrences, such as traffic patterns, surrounding property values, and surrounding commercial areas. To complete this task we had to merging four data sets into one, and do a lot of data cleaning and new variable creation. We also wanted to see if we could predict whether a crime would be violent or nonviolent, and what variables were most important in predicting whether a crime was violent or nonviolent. This required additional new variable creation and data cleaning. In exploring our data we found that some crimes happened more than others, specifically thefts, so additionally we wanted to look at factors that affect thefts in more detail.

Introduction to Data

All our data sets came from www.opendataphilly.org, which is the official open data repository for Philadelphia, PA. We knew we wanted to model crime in Philadelphia, but the crime data set did not include any potential predictors except date and time. Since we wanted to explore the relationship between crime and other city attributes we searched their website for relevant data. We had four data sets in total; crime data, property data, commercial data, and traffic data. Fig 1.1 to 1.4 show their basic properties such as how many observations they each had, how many variables and file size which was an important limiting factor in this analysis.

The crime data set was our main data set. Figure 1.5 shows the metadata for the Crime data. Some of the variables were redundant, for example the original variable called "Dispatch_Date_Time" was used to create four more variables in the data set; Date, time, hour and month. So instead we only kept the original variable and extracted needed information from it using SAS® Code. This helped us conserve processing resources when uploading the data. For the supplementary data sets, we kept the variables which made sense to be related to crime rate. Metadata for the supplementary variables is shown in Figure 1.6 through 1.8. Several of these data sets had multiple ID numbers, or multiple location identifiers. We only used relevant variables, which are listed in Figures 1.6 through 1.8.

Merging and Cleaning Data

The variable "Market_Value" seemed to have a lot of erroneous data. It varied from a few hundred dollars to a maximum of $690M data value. We researched online to see the range of house values in Philadelphia, and found out that the average minimum house sells for $35K, and the most expensive houses are for under $20M. So we decided to delete anything valued at less than $10K, and more than $20M. This was only 3.7% of our original data, so it did not impact our analysis greatly or warrant imputation. Figure 1.13 shows how skewed the data was before, and figure 1.14 shows the data afterwards.

We faced a few challenges when merging the data. Our main data set had crime events by time and location, whereas our supplementary data sets described the locations. Thus we decided to merge based on location. The challenge in doing this was that the original data for longitude and latitude had up to 13 decimal places. We needed to create a more general longitude and latitude with less decimal places, so the merge node could find matching values. Next to identify different spots we had to concatenate longitude and latitude to identify each different location. This is how we created the variable "GenLonLat". We left the date out since we were only going to merge based on location. The next challenge we faced was that our supplementary data sets now contained more than one value for each general location; multiple house values and multiple building conditions. So we needed to find the overall trend in each general location. We did this by using the Means Procedure for all our data sets. Using the Means Procedure on our Crime data also gave us a crime count by location and time, so we could tell which locations were likely to be hotspots. We also created a binary variable called “High_Crime” which was “yes” if the crime count for a specific location on a specific date was more than one and “no” otherwise. Once we had prepared our data sets we were able to merge them by "GenLonLat".

To see if we could predict whether a crime would be violent or nonviolent we had to create a new variable. We looked at the violent vs. nonviolent crimes using the uniform crime reporting (UCR) index. Figure 1.9 through 1.10 show the separation of violent vs. nonviolent crimes. Violent crimes are categorized as crimes against other humans where force or threats are used. Non-violent crimes are generally crimes related to property or other lesser offenses. We created a binary variable where 0 denoted nonviolent crime, and 1 denoted violent crimes. Figure 1.11 shows the distribution of our target variable, called "ViolentCr". Nonviolent crime is a lot more prevalent in Philadelphia then violent crime. Figure 1.12 shows all the new variables we created with descriptions for each.

While we were exploring the commercial data set’s variables we found some interesting information. When we merged the crime and commercial data sets, we encountered many missing values in the commercial variables (Figure 2.9). We attributed this to the fact that not all locations are commercial areas. One possible solutions to this problem was to create two different data sets, one with nonmissing commercial data and one with missing commercial data. Another solution was to create a binary variable to indicate whether a location is commercial or noncommercial.

Another challenge we faced was the size of our data set. When we decided to look only at the theft data, we had to split the large data set into 4 separate data sets to import the into SAS Enterprise Miner® and then go back and append them once they were in Enterprise Miner. The Enterprise Miner diagram that we used to rebuild and analyze the theft data, figure 4.1, can be found in appendix D.

Analysis

Some things we were particularly interested in were determining is where crime happened most and what type of crime was most common. In exploring the data we discovered that the most common crime in Philadelphia was theft which accounted for about 20% of all crime (Figure 2.1). We also discovered that the police district that experienced the most crimes was district 11 (Figure 2.2). The municipalities that experienced the most crimes were the South, West, River Ward, and Lower North. The North and Lower South municipalities had the least amount of crime (Figure 2.3). The zip codes that experienced the most crime were 19134 and 19102 (Figure 2.16). We also found that crime tended to take place in areas with low annual average daily traffic, and that this variable was very right skewed indicating that crimes usually happened in more isolated areas (Figure 2.10).

We were also interested in when crimes occurred. We determined that the least amount of crime were reported between 5AM and 7AM while the most crimes were reported between 4PM and 5PM (Figure 2.4). We also found that if there was a crime committed at a particular location on a particular date, there was a very small chance of another crime being committed the same date (Figure 2.5). Additionally we wanted to see what weekday crime was most prevalent. We found that Tuesdays were the days with the highest crime rate, and Sundays were the days that the least amount of crime occurred (Figure 2.6). Additionally, we looked at monthly and yearly trends. We found that crime tends to increase in the summer months and that over the years crime has been decreasing (Figure 2.7). Using time series analysis and exponential smoothing we predict crime will continue to decrease into 2017 (Figure 2.7). Lastly we looked at monthly trends for each different type of crime. We found that the most common types of crimes committed, thefts, vandalism/criminal mischief, and miscellaneous crimes peaked in August, and that other assaults (assaults not classified as aggravated assaults) peaked in May (Figure 2.11).

Since thefts were the most prevalent crimes, we decided to look at them in more detail. We found that thefts from vehicles accounted for about 40% of all thefts (Figure 2.13). Thefts and thefts from vehicles dropped to their lowest point for the year in February, and peak in summer months (Figure 2.15). There has also been an overall downward trend for thefts from vehicles since 2010 and an overall upward trend of regular thefts since 2010 (Figure 2.12). Perhaps increased car security is leading thieves to target other areas. Lastly, we discovered that both thefts from cars and regular thefts occurred least in the morning and peaked between 4 PM and 5 PM (Figure 2.14). This may be because people do not realize a theft has occurred until they get off work to find something valuable missing. We found that thefts from vehicles occurred most in the zip codes 19130, 19106, 19123, and 19102 (Figure 2.17), and that regular thefts occurred most in the zip codes 19137 and 19102 (Figure 2.18).

Lastly, we wanted to predict violent crimes. We found that violent crimes only accounted for about 7.7% of total crimes (Figure 2.8). This created an issue when we tried to create a model to predict whether a crime was a violent crime or a nonviolent crime. We did however discover that violent crimes happened at about the same frequency every day, while nonviolent crimes happened most on Tuesdays and least on Sundays (Figure 2.6). We also found that the municipalities with the most crime were also the municipalities with the most violent crime (Figure 2.3). To actually predict whether a crime was violent or nonviolent we used 5 classification algorithms (logistic regression – with and without transformation, decision tree, neural Network, k-nearest neighbor). Our Enterprise Miner diagram, figure 4.2 can be found in appendix D.

In all models, the decision tree showed the least miss-classification rate, with the neural network showing the highest miss-classification rate of almost 0.26 (Figure 3.1). The average squared error (ASE) was identical for all models, varying by a few decimal points (Figure 3.1). Our tree model can be seen in figure 3.2. The variables that were deemed important by the logistic regression model where our variables were standardized were the municipality, stage of development of commercial area, the condition of public and private buildings, the amount of traffic in the area, the number of houses in the area, the average market value of houses in the area, number of occupied commercial spaces in the area, and the quality of commercial establishments (Figure 3.3).

In general, the prediction power of the various models were not convincing. However, there were other metrics that suggested otherwise. For example, using the cumulative lift as a plausible metric, the k-nearest neighbor model showed the most significant lift of all the other models (Figure 3.4). The ROC curves for all the k-nearest neighbor model is also the best out of all of the models (Figure 3.5).

Further Research

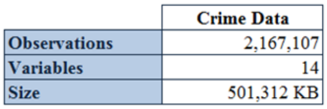
While this project focused on describing crime occurrences and predicting whether a crime was violent or nonviolent, we would like to do more in depth research that would help identify hotspots for crime, when future crime will take place, and what types of crimes will take place. We are also interested in looking at different variable transformations and creating subsets of our original data (data sets with 50% violent and 50% nonviolent or data sets dealing with only commercial areas or noncommercial areas) to improve our current model. To identify hotspots we hope to use geocode data when it is available to us. Finding more demographic data to add to our merged data set is something else we would like to look into.

Conclusion

In this study, we found that crimes generally occurred in more isolated areas where there was less traffic, and that there were certain locations that had higher crime counts than others. We also discovered there was less crime in the morning and more in the afternoon as well as less crime on Sunday and more crime on Tuesday. During the summer months total crime occurrences as well as the most prevalent types of crime occurrences (thefts, vandalism/criminal mischief, and miscellaneous crimes, and other assault) peaked. Theft occurrences, the most prevalent crime occurrences, showed many of the same trends as overall crime occurrences. We found that thefts and vehicle thefts as well as overall crime occurrences were most prevalent in the zip code 19102. Law enforcement officials could possibly bring down their overall crime by cracking down in this areas. The models we built to try to classify a crime as violent or nonviolent were not very fruitful, but the tree model was the best in terms of validation misclassification error rate. We hope to explore this data set further to determine where there are crime hotspots, and add demographic data to explore the relationships between demographics and crime occurrence.

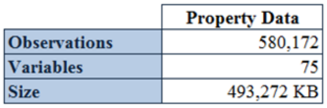
Appendix A: Introduction to Data and Data Cleanup

Figure 1.1



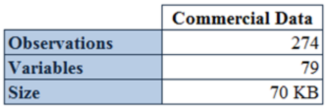
	Crime Data
Observations	2,167,107
Variables	14
Size	501,312 KB

Figure 1.2



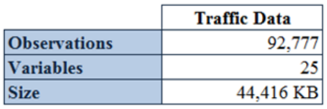
	Property Data
Observations	580,172
Variables	75
Size	493,272 KB

Figure 1.3



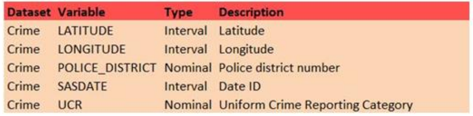
	Commercial Data
Observations	274
Variables	79
Size	70 KB

Figure 1.4



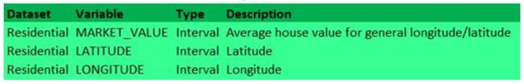
	Traffic Data
Observations	92,777
Variables	25
Size	44,416 KB

Figure 1.5



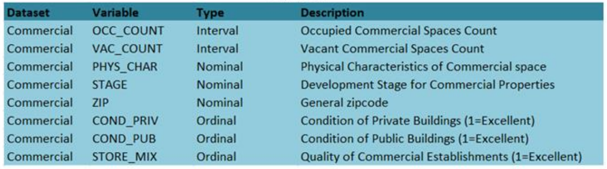
Dataset	Variable	Type	Description
Crime	LATITUDE	Interval	Latitude
Crime	LONGITUDE	Interval	Longitude
Crime	POLICE_DISTRICT	Nominal	Police district number
Crime	SASDATE	Interval	Date ID
Crime	UCR	Nominal	Uniform Crime Reporting Category

Figure 1.6



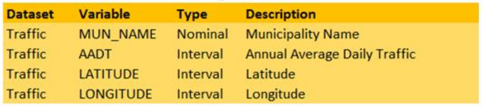
Dataset	Variable	Type	Description
Residential	MARKET_VALUE	Interval	Average house value for general longitude/latitude
Residential	LATITUDE	Interval	Latitude
Residential	LONGITUDE	Interval	Longitude

Figure 1.7



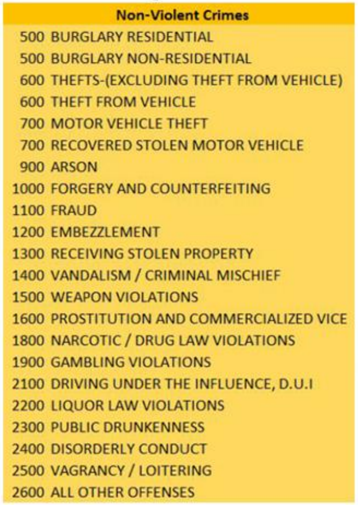
Dataset	Variable	Type	Description
Commercial	OCC_COUNT	Interval	Occupied Commercial Spaces Count
Commercial	VAC_COUNT	Interval	Vacant Commercial Spaces Count
Commercial	PHYS_CHAR	Nominal	Physical Characteristics of Commercial space
Commercial	STAGE	Nominal	Development Stage for Commercial Properties
Commercial	ZIP	Nominal	General zipcode
Commercial	COND_PRIV	Ordinal	Condition of Private Buildings (1=Excellent)
Commercial	COND_PUB	Ordinal	Condition of Public Buildings (1=Excellent)
Commercial	STORE_MIX	Ordinal	Quality of Commercial Establishments (1=Excellent)

Figure 1.8



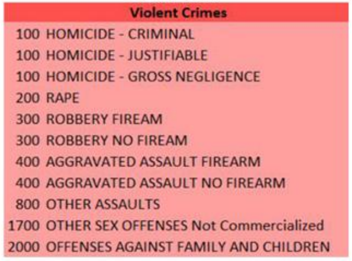
Dataset	Variable	Type	Description
Traffic	MUN_NAME	Nominal	Municipality Name
Traffic	AADT	Interval	Annual Average Daily Traffic
Traffic	LATITUDE	Interval	Latitude
Traffic	LONGITUDE	Interval	Longitude

Figure 1.9



Non-Violent Crimes	
500	BURGLARY RESIDENTIAL
500	BURGLARY NON-RESIDENTIAL
600	THEFTS-(EXCLUDING THEFT FROM VEHICLE)
600	THEFT FROM VEHICLE
700	MOTOR VEHICLE THEFT
700	RECOVERED STOLEN MOTOR VEHICLE
900	ARSON
1000	FORGERY AND COUNTERFEITING
1100	FRAUD
1200	EMBEZZLEMENT
1300	RECEIVING STOLEN PROPERTY
1400	VANDALISM / CRIMINAL MISCHIEF
1500	WEAPON VIOLATIONS
1600	PROSTITUTION AND COMMERCIALIZED VICE
1800	NARCOTIC / DRUG LAW VIOLATIONS
1900	GAMBLING VIOLATIONS
2100	DRIVING UNDER THE INFLUENCE, D.U.I
2200	LIQUOR LAW VIOLATIONS
2300	PUBLIC DRUNKENNESS
2400	DISORDERLY CONDUCT
2500	VAGRANCY / LOITERING
2600	ALL OTHER OFFENSES

Figure 1.10



Violent Crimes	
100	HOMICIDE - CRIMINAL
100	HOMICIDE - JUSTIFIABLE
100	HOMICIDE - GROSS NEGLIGENCE
200	RAPE
300	ROBBERY FIREAM
300	ROBBERY NO FIREAM
400	AGGRAVATED ASSAULT FIREARM
400	AGGRAVATED ASSAULT NO FIREARM
800	OTHER ASSAULTS
1700	OTHER SEX OFFENSES Not Commercialized
2000	OFFENSES AGAINST FAMILY AND CHILDREN

Figure 1.11

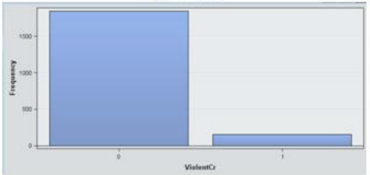
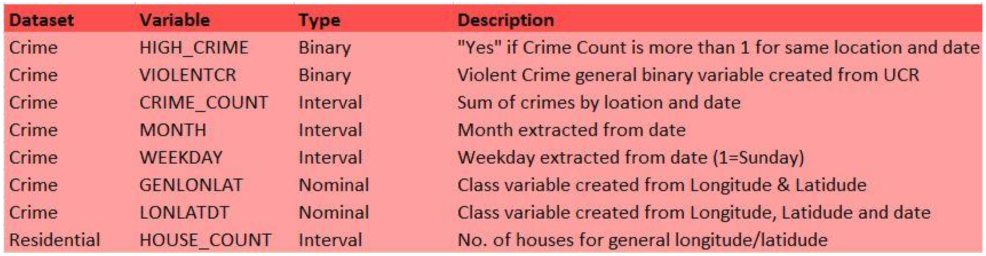


Figure 1.12



Dataset	Variable	Type	Description
Crime	HIGH_CRIME	Binary	"Yes" if Crime Count is more than 1 for same location and date
Crime	VIOLENTCR	Binary	Violent Crime general binary variable created from UCR
Crime	CRIME_COUNT	Interval	Sum of crimes by loation and date
Crime	MONTH	Interval	Month extracted from date
Crime	WEEKDAY	Interval	Weekday extracted from date (1=Sunday)
Crime	GENLONLAT	Nominal	Class variable created from Longitude & Latidude
Crime	LONLATDT	Nominal	Class variable created from Longitude, Latidude and date
Residential	HOUSE_COUNT	Interval	No. of houses for general longitude/latidude

Figure 1.13

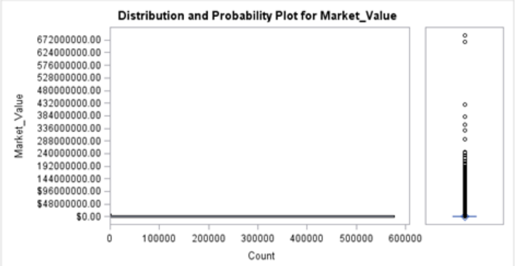
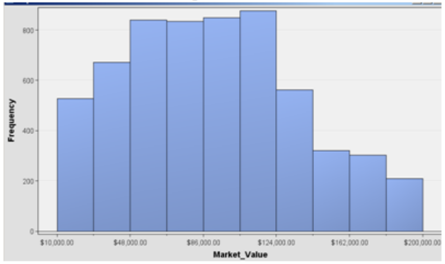


Figure 1.14



Appendix B: Data Exploration

Figure2.1: Bar Chart of Types of Crime

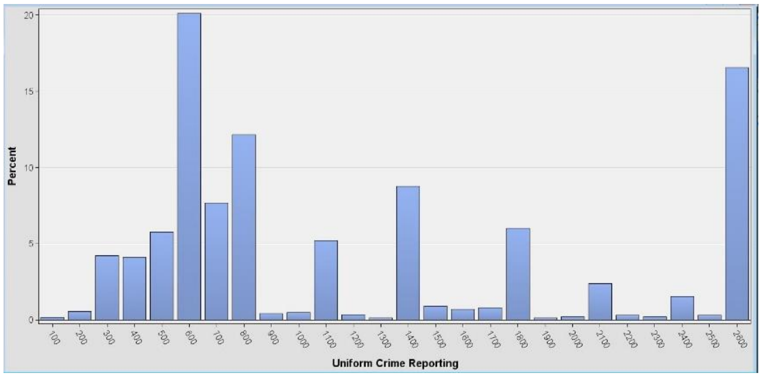


Figure 2.2: Bar Chart of Police District with Violent Crime (0 = nonviolent, 1 = violent)

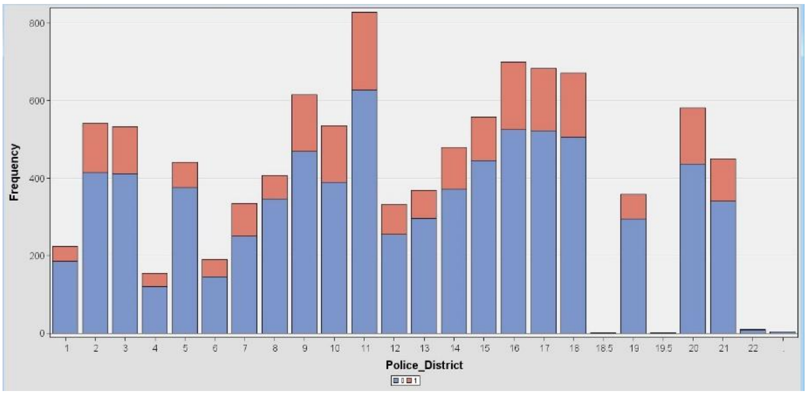


Figure 2.3: Violent and Nonviolent Crimes within Municipalities

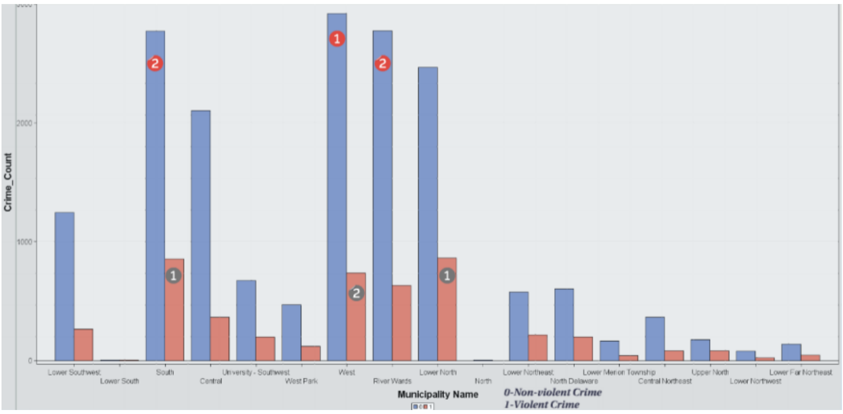


Figure 2.4: Frequency of Crime by Hour

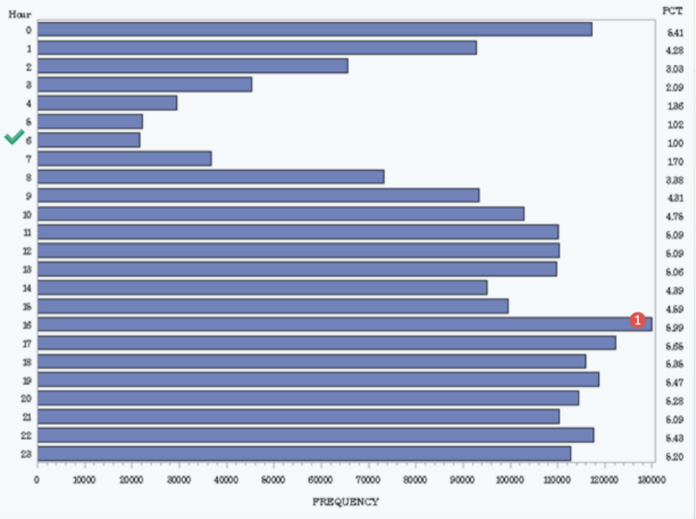


Figure 2.5: Frequency of Crime Count

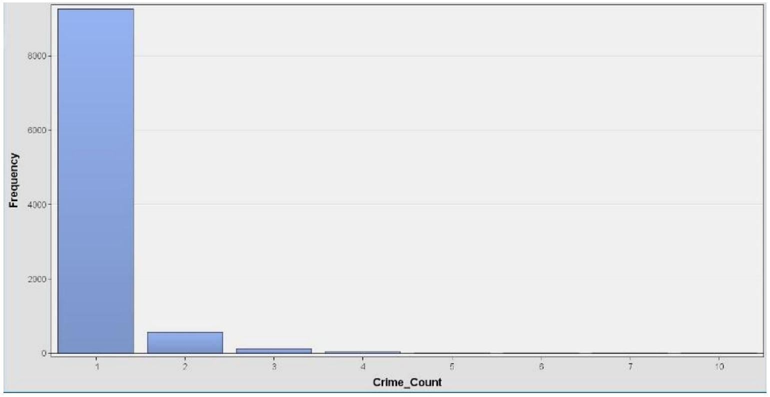


Figure 2.6: Bar Chart of Weekday by Violent Crime

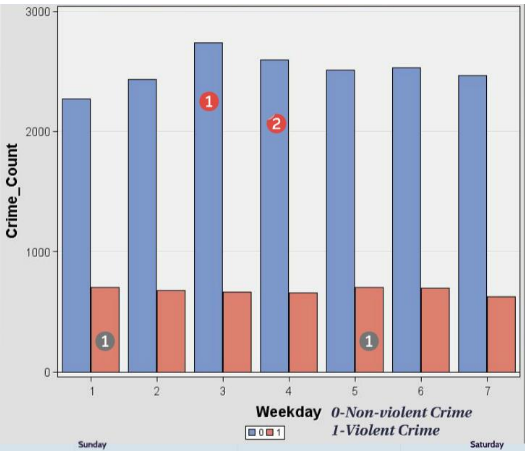


Figure 2.7: Exponential Smoothing Monthly Time Series Model

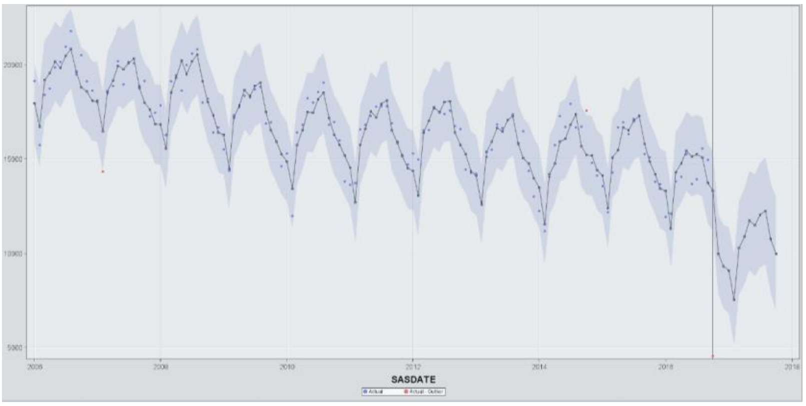


Figure 2.8: Bar Chart of Violent Crime (0 = nonviolent, 1 = violent)

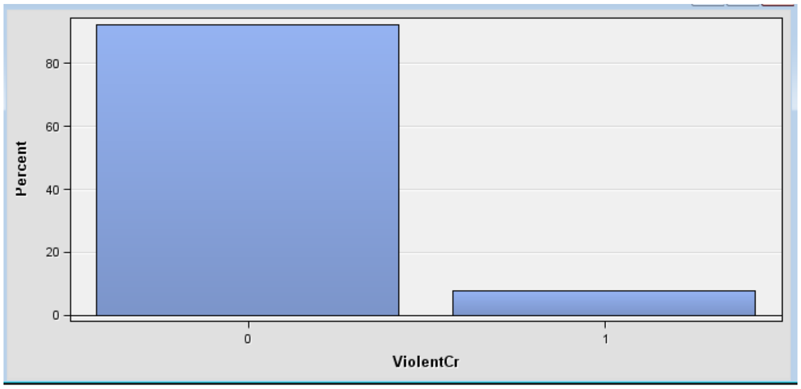


Figure 2.9: Bar Charts and Histograms for Commercial Data Set Variables

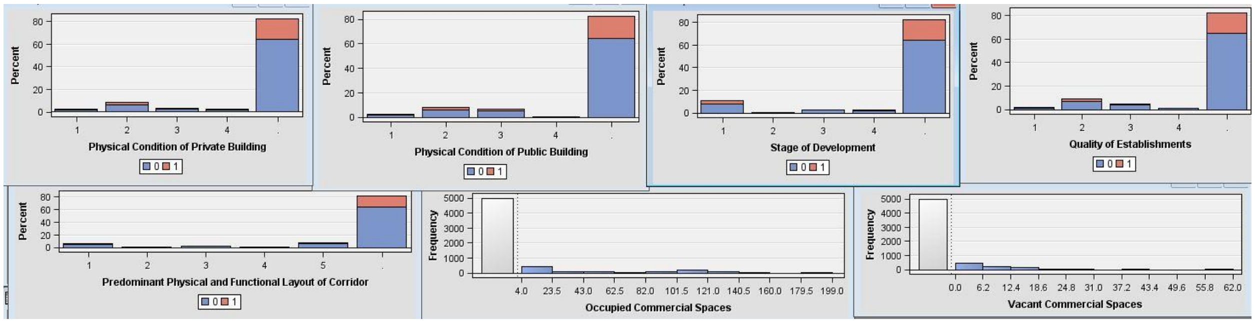


Figure 2.10: Histogram of Annual Average Daily Traffic

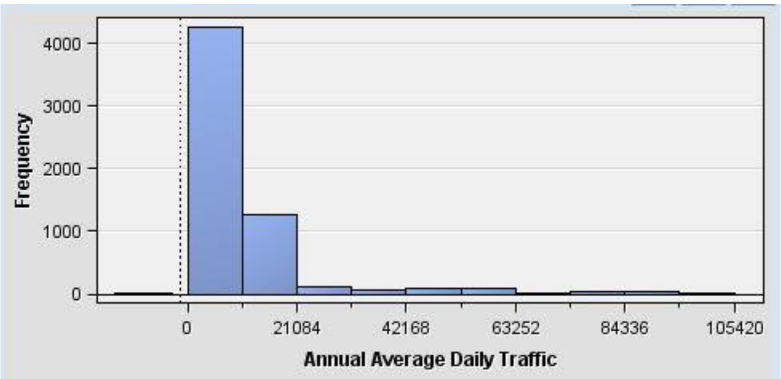


Figure2.11: Crime type by month

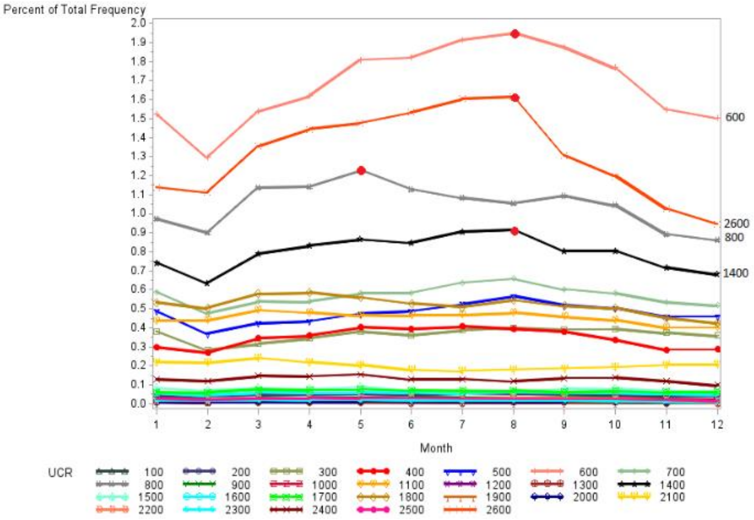


Figure 2.12: Theft Type over time

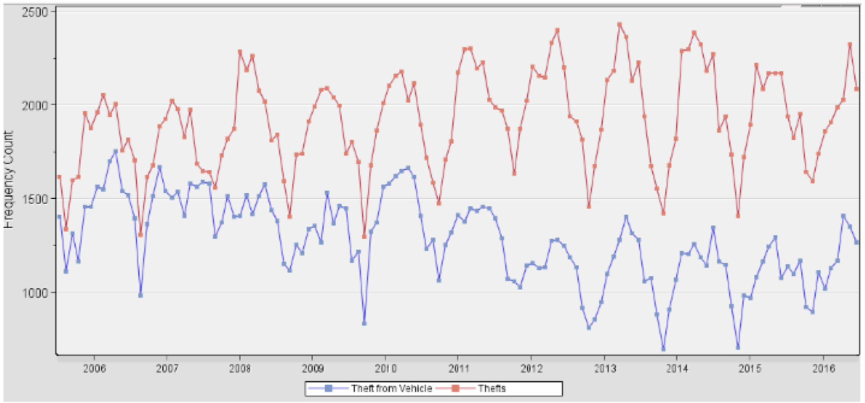


Figure 2.13: Bar Chart of Theft Type

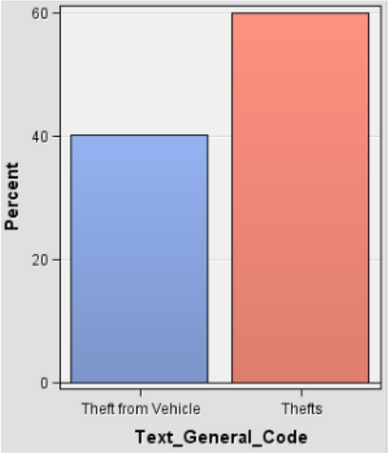


Figure 2.14: Theft Type by Hour

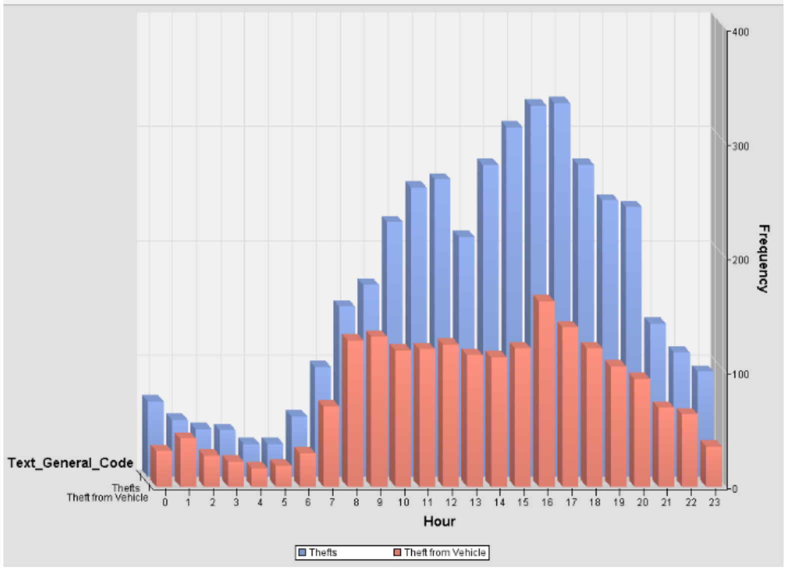


Figure 2.15: Theft Type by Month

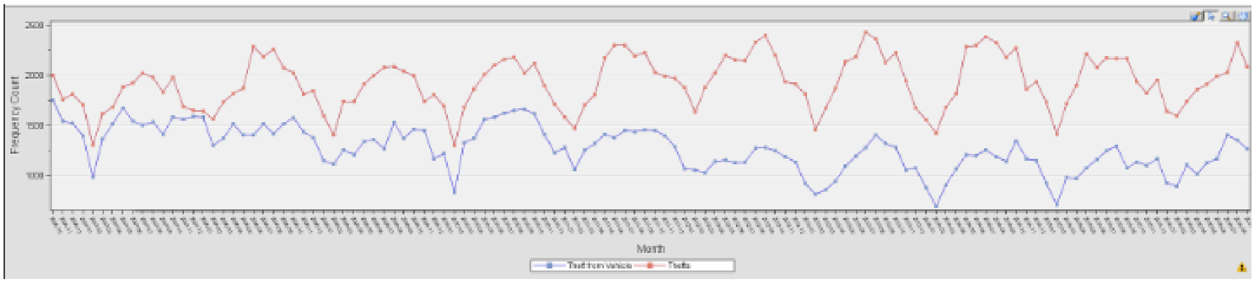


Figure 2.16: Heat Map for Crime Counts by Zip Code

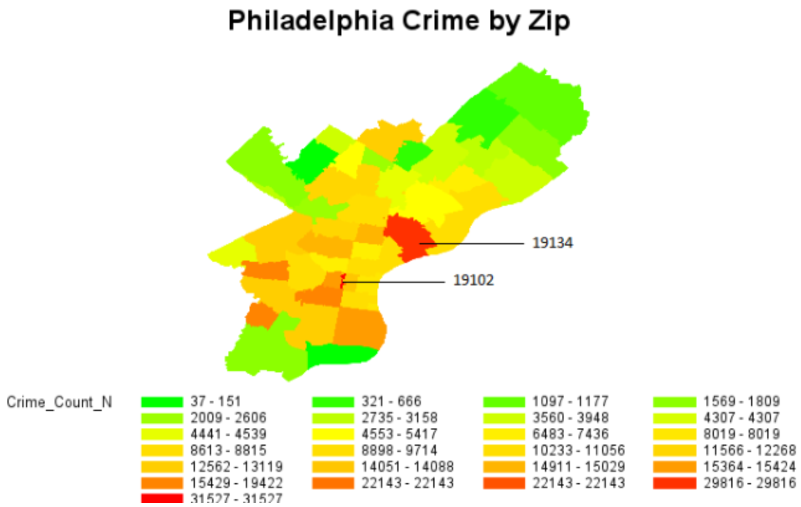


Figure 2.17: Heat Map of Vehicle Thefts by Zip Code

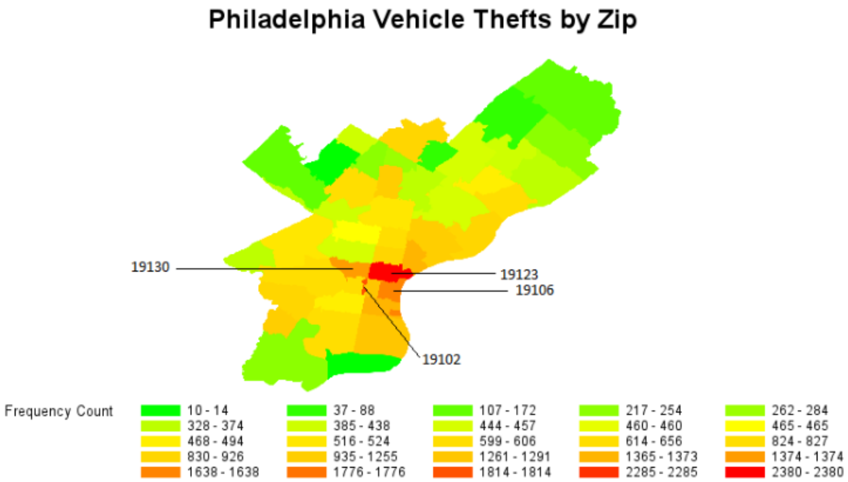
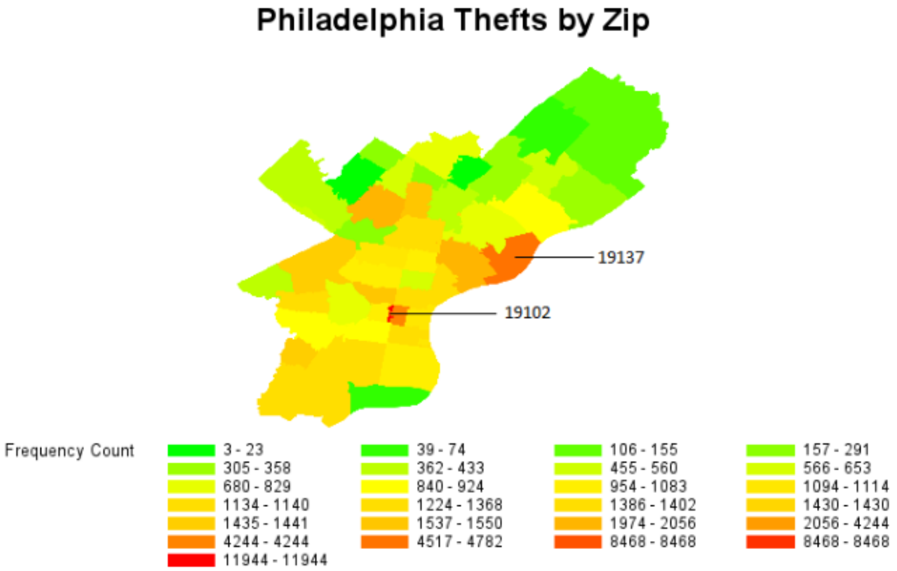
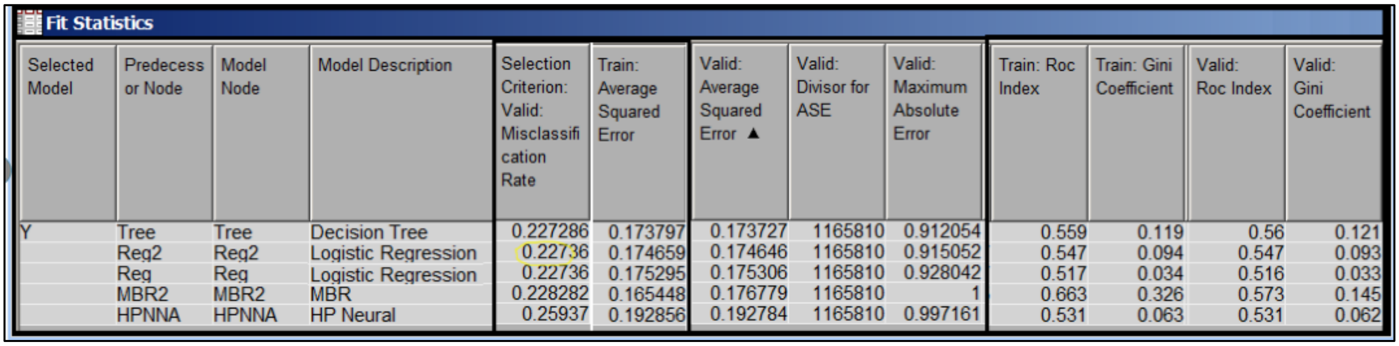


Figure 2.18: Heat Map of Thefts by Zip Code



Appendix C: Predictive Modeling

Figure 3.1: Model Comparison


Fit Statistics

Selected Model	Predecess or Node	Model Node	Model Description	Selection Criterion: Valid: Misclassification Rate	Train: Average Squared Error	Valid: Average Squared Error ▲	Valid: Divisor for ASE	Valid: Maximum Absolute Error	Train: Roc Index	Train: Gini Coefficient	Valid: Roc Index	Valid: Gini Coefficient
Y	Tree	Tree	Decision Tree	0.227286	0.173797	0.173727	1165810	0.912054	0.559	0.119	0.56	0.121
	Reg2	Reg2	Logistic Regression	0.22736	0.174659	0.174646	1165810	0.915052	0.547	0.094	0.547	0.093
	Reg	Reg	Logistic Regression	0.22736	0.175295	0.175306	1165810	0.928042	0.517	0.034	0.516	0.033
	MBR2	MBR2	MBR	0.228282	0.165448	0.176779	1165810	1	0.663	0.326	0.573	0.145
	HPNNA	HPNNA	HP Neural	0.25937	0.192856	0.192784	1165810	0.997161	0.531	0.063	0.531	0.062

Figure 3.2: Tree Model

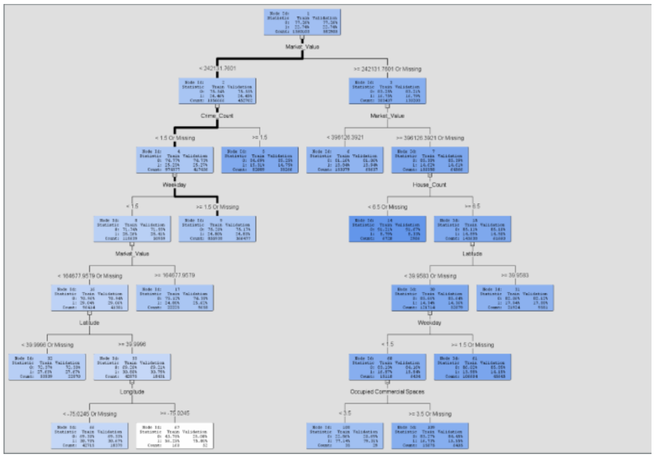
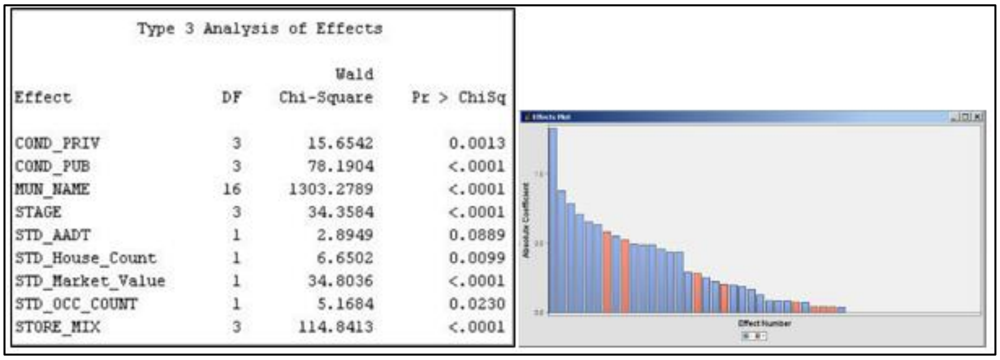


Figure 3.3: Analysis of effect for Regression after Standardization of Variables



```
                Type 3 Analysis of Effects

                                  Wald
Effect               DF    Chi-Square    Pr > ChiSq

COND_PRIV             3       15.6542        0.0013
COND_PUB              3       78.1904        <.0001
MUN_NAME             16     1303.2789        <.0001
STAGE                 3       34.3584        <.0001
STD_AADT              1        2.8949        0.0889
STD_House_Count       1        6.6502        0.0099
STD_Market_Value      1       34.8036        <.0001
STD_OCC_COUNT         1        5.1684        0.0230
STORE_MIX             3      114.8413        <.0001
```

Figure 3.4: k-nn cumulative lift

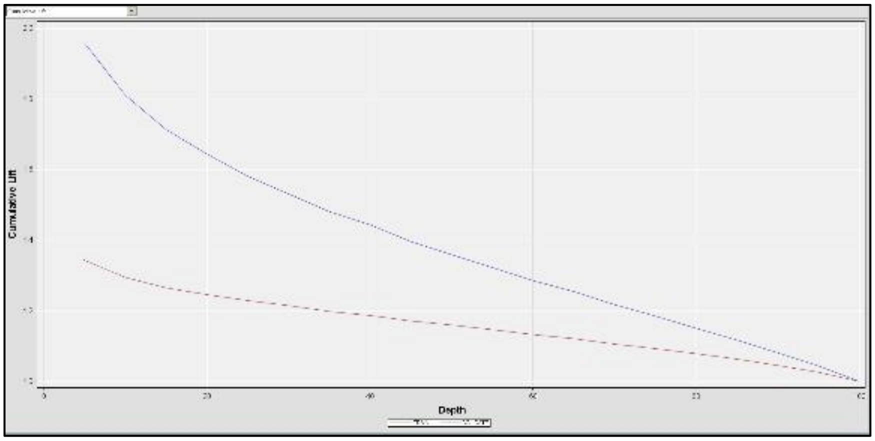
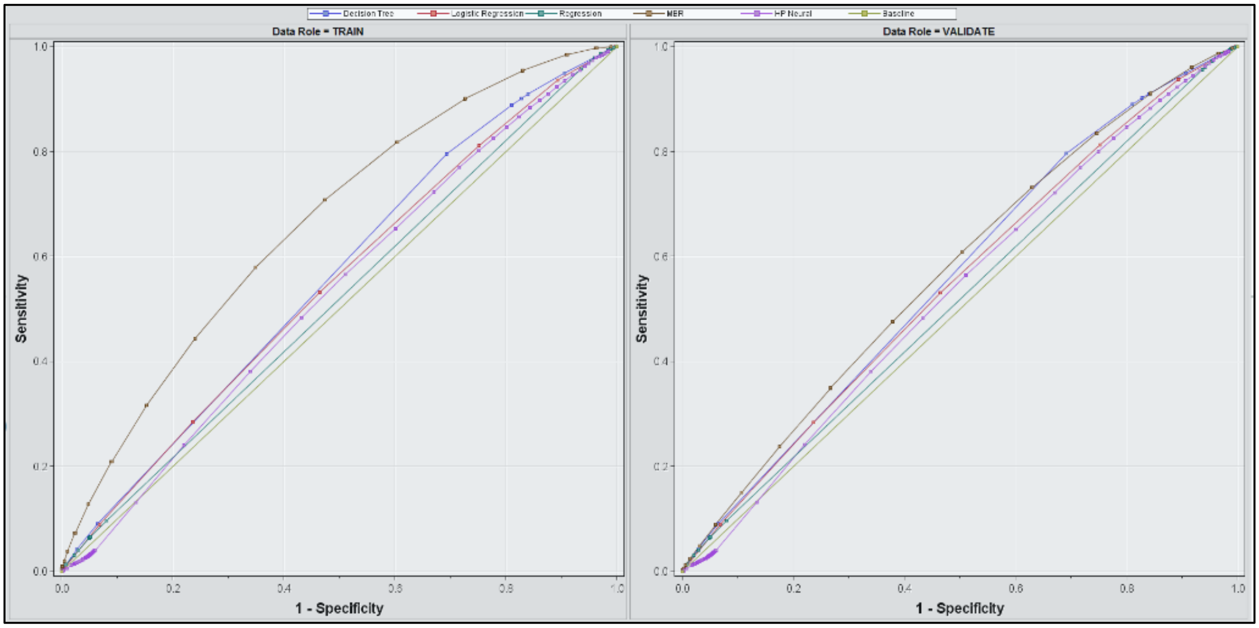


figure3.5: ROC curves



Appendix D: SAS Code

```
*Means procedure for Crime Data;
Proc means data=CrimeTemp1 N Mean Median Mode noprint;
 var UCR_General Police_Districts Lon Lat;
 class LonLatDT;
output out=new (Drop=_TYPE_ _FREQ_)
                                 N(Lon)=
                                 MEAN(UCR_General Police_Districts Lon Lat)=
                                 Mode(UCR_General)= /autoname;
run;

*Residential data clean up;
Data Crime.Residential;
       set Crime.Property (keep =Lon Lat Market_Value);
if Lat=. then delete;
if Lon=. then delete;
if Market_Value=. then delete;
if market_value < 10000 then delete;
if market_value > 20000000 then delete;
run;
data ResiTemp;
 set Crime.Residential;
run;
data ResiTemp1 (drop= Lon Lat LonTemp LonTemp1 LonTemp2 LatTemp LatTemp1 LatTemp2) ;
 set ResiTemp ;
 Longitude=input(Lon,best13.);
 Latitude=input(Lat,best12.);
 LonTemp=INT(Longitude*10000);
 LatTemp=INT(Latitude*10000);
 LonTemp1=put(LonTemp,8.);
 LatTemp1=put(LatTemp,8.);
 LonTemp2=substr(LonTemp1,2,5);
 LatTemp2=substr(LatTemp1,2,5);
 GENLonLat=CATX("-",LonTemp2,LatTemp2);
run;

*Running the Means procedure on Market Value;
Proc means data=ResiTemp1 N Mean noprint;
 var Market_Value;
 class GENLonLat;
output out=new1 (Drop=_TYPE_ _FREQ_)
                                 N(Market_Value)=
                                 MEAN(Market_Value)= /autoname;
run;
Data Crime.ResiEM;
       set New1 (firstobs=2);
run;

*SAS Enterprise Miner Code;
*Code to create variables for Crime data;
data &EM_EXPORT_TRAIN;
set &EM_IMPORT_DATA;
Weekday=weekday(SASDate);
Month=month(SASDate);
GENLonLat=substr(LonLatDT,1,4)||'-'||substr(LonLatDT,8,5);
run;

*Code to create GenLonLat for Traffic data so we can merge it;
data &EM_EXPORT_TRAIN;
set &EM_IMPORT_DATA;
LonTemp=INT(Longitude*10000);
```

```
LatTemp=INT(Latitude*10000);
LonLat=CATX("-",put(LatTemp,8.),put(LonTemp,8.));
GENLonLat=substr(LonLat,1,4)||'-'||substr(LonLat,8,5);
If Type ne "Volume" then delete; *deleted 15min volume etc.;
run;

*Creating new variables such as target variable ViolentCR;
data &EM_EXPORT_TRAIN (rename=(Lon_N=Crime_Count Market_Value_N=House_Count
     Lon_Mean=Longitude Lat_Mean=Latitude Market_Value_Mean=Market_Value
        Police_Districts_Mean=Police_District));
set &EM_IMPORT_DATA;
select (UCR);
when (100,200,300,400,800,1700,2000) ViolentCr=1;
otherwise ViolentCr=0;
end;
if Lon_N>1 then HighCrime=1;
       else HighCrime=0;
if UCR=. then delete;
label AADT="Annual Average Daily Traffic"
       COMM_COUNT="Total Commercial Spaces"
       MUN_Name="Municipality Name"
       OCC_Count="Occupied Commercial Spaces"
       Type="Type of Traffic Recording"
       UCR="Uniform Crime Reporting"
       VAC_Count="Vacant Commercial Spaces"
       P_Dist="Planing District";
run;
```

```
*Heat Map - Data prepartion;
data Crime.Theftmaps;
set Theftmaps;
 LonTemp=INT(Lon*100);
 LatTemp=INT(Lat*100);
 length LonLat $14;
 LonLat=CATX("-",put(LatTemp,8.),put(LonTemp,8.));
run;
*Data with zipcodes;
data SZip;
set Sashelp.Zipcode;
where STATECODE="PA";
rename zip=ZCTA5CE10;
 LonTemp=INT(X*100);
 LatTemp=INT(Y*100);
 length LonLat $14;
 LonLat=CATX("-",put(LatTemp,8.),put(LonTemp,8.));
run;
*Merge;
Proc SQL;
create table crime.Theftheatmap as
Select C1.TEXT_GENERAL_CODE, C1.HOUR, C1.MONTH, C1.LONLAT, C2.ZCTA5CE10
from Crime.Theftmaps as C1 left join Work.Szip as C2
       on C1.LonLat=C2.LonLat;
run;
*Delete missing;
data crime.Theftheatmap;
set crime.Theftheatmap;
if ZCTA5CE10=. then delete;
run;
*Count crime by zip;
Proc freq data=crime.Theftheatmap ;
tables TEXT_GENERAL_CODE*hour*ZCTA5CE10 /out=crime.theftmapfreq;
run;
Proc freq data=crime.Theftheatmap2 ;
```

```
tables TEXT_GENERAL_CODE*ZCTA5CE10 /out=crime.theftmapfreq2;
run;

*Reformatting zip for heat map;
data crime.theftmapfreq2 ;
set crime.theftmapfreq2;
Zip=put(ZCTA5CE10,5.);
run;
data crime.theftmapfreq2 ;
set crime.theftmapfreq2 (drop=ZCTA5CE10);
rename Zip=ZCTA5CE10;
run;

*splitting data into thefts vs thefts from vehicle;
data crime.GTheatmap crime.VTheatmap;
set crime.theftmapfreq2;
IF TEXT_GENERAL_CODE="Thefts" then output crime.GTheatmap;
      else output crime.VTheatmap;
run;

*Heat Map for Thefts General;
Proc MapImport
datafile="\\Client\C$\MSAS\Kaggle\PA Crime\tl_2010_42_zcta510\tl_2010_42_zcta510.shp"
out=crime.pa_zshp;
run;

*Map data;
ODS Graphics on;
ods html style=statistical;
goptions ftitle='Arial';
title1 " "; title10 h=2.2 "Philadelphia Thefts by Zip"; footnote " ";
/* colors from green->red */
pattern1 v=s c=cx00ff00;
pattern2 v=s c=cx35ff00;
pattern3 v=s c=cx65ff00;
pattern4 v=s c=cx88ff00;
pattern5 v=s c=cx9aff00;
pattern6 v=s c=cxbaff00;
pattern7 v=s c=cxccff00;
pattern8 v=s c=cxd0ff00;
pattern9 v=s c=cxe0ff00;
pattern10 v=s c=cxffff00;
pattern11 v=s c=cxffee00;
pattern12 v=s c=cxffe000;
pattern13 v=s c=cxffdd00;
pattern14 v=s c=cxffdc00;
pattern15 v=s c=cxffd800;
pattern16 v=s c=cxffd100;
pattern17 v=s c=cxffcd00;
pattern18 v=s c=cxffc000;
pattern19 v=s c=cxffb700;
pattern20 v=s c=cxff9a00;
pattern21 v=s c=cxff8700;
pattern22 v=s c=cxff7700;
pattern23 v=s c=cxff5400;
pattern24 v=s c=cxff3400;
pattern25 v=s c=cxff0000;

proc gmap data=crime.GTheatmap
map=crime.pa_zshp;
id ZCTA5CE10;
choro Count/ levels=25 coutline=same;
run;
```

```
quit;

ods html close;
ods graphics off;

*Heat Map for Vehicle Thefts General;
*Map data;
ODS Graphics on;
ods html style=statistical;
goptions ftitle='Arial';
title1 " "; title10 h=2.2 "Philadelphia Vehicle Thefts by Zip"; footnote " ";
/* colors from green->red */
pattern1 v=s c=cx00ff00;
pattern2 v=s c=cx35ff00;
pattern3 v=s c=cx65ff00;
pattern4 v=s c=cx88ff00;
pattern5 v=s c=cx9aff00;
pattern6 v=s c=cxbaff00;
pattern7 v=s c=cxccff00;
pattern8 v=s c=cxd0ff00;
pattern9 v=s c=cxe0ff00;
pattern10 v=s c=cxffff00;
pattern11 v=s c=cxffee00;
pattern12 v=s c=cxffe000;
pattern13 v=s c=cxffdd00;
pattern14 v=s c=cxffdc00;
pattern15 v=s c=cxffd800;
pattern16 v=s c=cxffd100;
pattern17 v=s c=cxffcd00;
pattern18 v=s c=cxffc000;
pattern19 v=s c=cxffb700;
pattern20 v=s c=cxff9a00;
pattern21 v=s c=cxff8700;
pattern22 v=s c=cxff7700;
pattern23 v=s c=cxff5400;
pattern24 v=s c=cxff3400;
pattern25 v=s c=cxff0000;

proc gmap data=crime.VTheatmap
map=crime.pa_zshp;
id ZCTA5CE10;
choro Count/ levels=25 coutline=same;
run;
quit;

ods html close;
ods graphics off;




/* Create plot of crime type by month */
proc gplot data= crime.ucrfreq;
   plot percent*month=UCR / hminor=0;
run;

/* Define symbol characteristics */
symbol1 color=dabg interpol=join value=triangle height=1 width=2;
symbol2 color=mob interpol=join value=circle height=1 width=2;
symbol3 color=day interpol=join value=square height=1 width=2;
symbol4 color=red interpol=join value=dot height=1 width=2;
symbol5 color=blue interpol=join value=V height=1 width=2;
symbol6 color=lime interpol=join value=_ height=1 width=2;
```

```
symbol7 color=darkseagreen interpol=join value=diamond height=1 width=2;
symbol8 color=gray interpol=join value=hash height=1 width=2;
symbol9 color=green interpol=join value=Y height=1 width=2;
symbol10 color=crimson interpol=join value=Z height=1 width=2;
symbol11 color=orange interpol=join value=# height=1 width=2;
symbol12 color=purple interpol=join value=: height=1 width=2;
symbol13 color=brown interpol=join value=+ height=1 width=2;
symbol14 color=black interpol=join value== height=1 width=2;
symbol15 color=aquamarine interpol=join value=% height=1 width=2;
symbol16 color=aqua interpol=join value=- height=1 width=2;
symbol17 color=salmon interpol=join value=plus height=1 width=2;
symbol18 color=darkgoldenrod interpol=join value=$ height=1 width=2;
symbol19 color=chocolate interpol=join value=traingle height=1 width=2;
symbol20 color=navy interpol=join value=* height=1 width=2;
symbol21 color=gold interpol=join value=/ height=1 width=2;
symbol22 color=coral interpol=join value=? height=1 width=2;
symbol23 color=cyan interpol=join value=( height=1 width=2;
symbol24 color=maroon interpol=join value=X height=1 width=2;
symbol25 color=deeppink interpol=join value=point height=1 width=2;
symbol26 color=orangered interpol=join value=paw height=1 width=2;

 /* Define axis characteristics */
axis1 label=none
      value=('JAN' 'FEB' 'MAR' 'APR' 'MAY' 'JUN'
             'JUL' 'AUG' 'SEP' 'OCT' 'NOV' 'DEC')
      order = 1 to 12 by 1
      offset=(2)
      width=1;
*axis2 label=('Degrees' justify=right 'Fahrenheit')
      order=(0 to 100 by 10)
      width=1;

 /* Enhance the legend */
*legend1 label=none value=(tick=1 'Minneapolis');
goptions reset gsfname=;
   plot percent*month=UCR /
        haxis=axis1 hminor=0
        vaxis=axis2 vminor=1
        legend=legend1;
run;
quit;
```

Figure 4.1: Enterprise Miner Theft Data Diagram

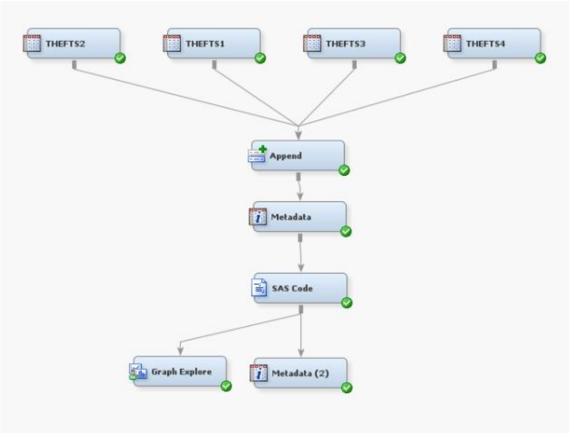
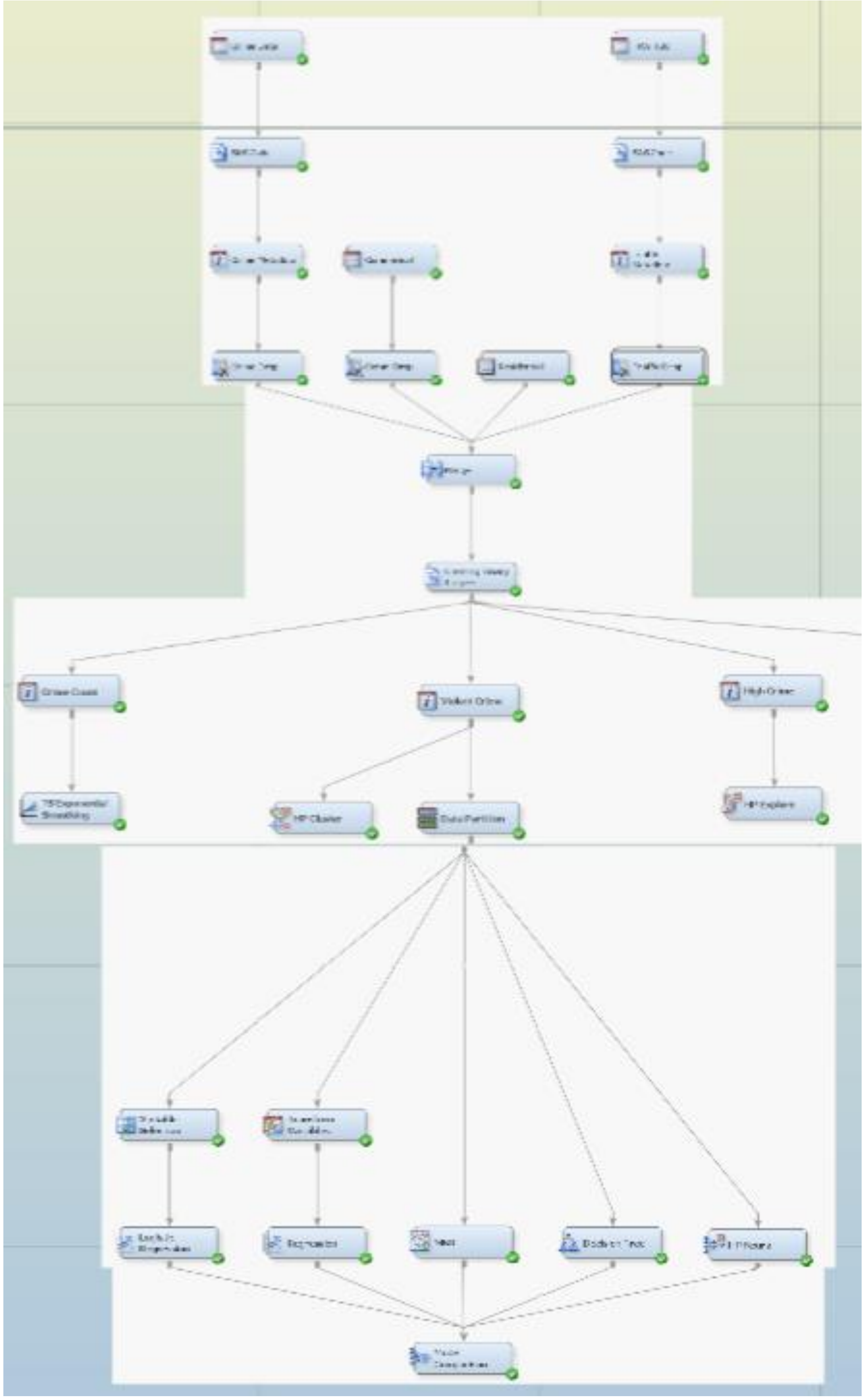


Figure 4.2: Enterprise Miner Diagram For Merging Data and Predictive Modeling


Merge
Crime Count
Violent Crime
High Crime
TS Exponential Smoothing
HP Cluster
Data Partition
HP Explore
Regression
Decision Tree