Blighted Data property

Blight can be considered any abandoned or vacant property, either a building or home in an unacceptable condition of disrepair. It is inhabitable and has little to no economic value. These dwellings usually consist of overgrown grass and trash dumping. Being an eye-soar to the community, they reduce the neighborhood's property value. However, who can we say is ultimately responsible for blight? We can't blame this problem on individual owners. All Memphians, from top positions to bottom, are accountable for this situation. Now, where does Memphis stand when it comes to blight rating? "The last city-wide survey of residential properties was collected from 2008-2010 through a partnership with the University of Memphis (CBANA) and the City of Memphis. This survey, which included approximately 200,000 parcels, indicated that the City had a blight rate of 22%, approximately 40,000 residential parcels."

The US is facing the problem of empty neighborhoods across the country. We can see similar problems in Memphis, TN, and the causes for this are: poor planning, poverty, and sub-urbanization are the causes of the urban blight. The problems resulted from unemployment, depopulation problems, promoted crimes, and loss of deindustrialization. Blight-affected areas are often featured with dirty, cloistered, tumbledown buildings and inhospitable living.

Business question:

As we progressed through the assigned data research on the effects of blighted properties, one question that kept coming to mind was why these properties reached such a depleted state to cause these effects during the time we have chosen to research. We began to consider the period after the great recession and the massive amount of foreclosures during the events leading up through that time. We conclude that much of the data information we explored came about either directly or indirectly from the results of foreclosures from an inability to make mortgage payments or pay the tax liens upon the property. Even though the homes were in foreclosure, they were, for a time, still occupied during proceedings and, for the most part, maintained and kept up until the properties became vacant and eventually abandoned. At that time, the downward spiral toward blight seems to begin. Our theory from the data research on these properties offers some ideas from the within the observed categories, crime, blight ratings, sales, and demolition, for insight into possible solutions and precursors for further consideration of prevention.

Even though foreclosure is costly to the owners in default, the financial losses spiral downward immediately, beginning with a reduction in the appreciative values of the surrounding homes even before the foreclosed homes become vacant. However, once vacant, these homes often become unmaintained, unkempt, and left unsecured, leading to eventual abandonment. Once a property is empty or abandoned, its property loss effect is more than half of foreclosure's effect, with significant increases in crime turning into a significant contributor.

One of the prevention considerations we considered is demolition. The dummy variable of the demolition fees can serve as an indicator of blighted property. However, other data or information doesn't appear directly related to the blighted property except for the demolition fees. Therefore, it is reasonable to expect that predicting this target and characterizing the relationships between the target and predictors could give better results.

Business problem statement:

By exploring relevant data (i.e., crime, sales, blight rating, demolishing fees, etc.) for blighted properties considering the demolition fees, we will attempt to identify properties that are possible to become blighted. This information can guide the local community and government for a long-term development plan and help minimize the harmful effects of a blighted property/area.

Target Variables: demolition dummy variable (Binary data type)

Data Source and description:
https://drive.google.com/drive/folders/0B5AOGoio00aTNDNORE1oVDVfVU0

- Crime (Data from 2010-2015)
  For crime data, 'crime.data.csv' is the raw file
- Tax (Data from 2010-2016)
  For crime data, 'Tax.data.csv' is the raw file,
- MLGW (Data from 2010-2016)
  For crime data, 'mlgw.data.csv' is the raw file,
- Sales (Data from 2010-2016)
  For crime data, 'sales.data.txt' is the raw file,
- Windshield Data (2010)
  The data from a survey conducted in 2010 about the structural issues of parcels in Memphis as 'windshield.data.csv'

Datasets	Content of each data set
Windshield	**Thorough property information includes house condition details of all the Memphis and Shelby county property, Memphis and Shelby county regulations with code enforcement, and ~150,000 parcel ID. The dataset was not clean and carried numerous errors.**
Tax aggregation	**There were thirty-two statement records of aggregated property tax information from 2010 to 2016 and 46,751 parcel ID data collected from 2010 to 2016.**
MLGW aggregation	**The total counts of MLGW cut off the supply to one property. Interval (discrete) data with data range [0, 2], two types of data recorded (gas cut and electricity cut) per year for 15,678 parcel ID**
Sales aggregation	**The total counts of sales happened each year. Interval (discrete) data with data range [0, 27], one sale record of each year for 127,609 parcel ID**
Crime aggregation	**The total counts of crimes happen on each property. Interval (discrete) data with data range [0, 447], three types of crimes (violent, nonviolent, other) document for each year; 98,366 data point**

Data Preprocessing

- Initially, we had files with missing data, and the data files were messy
- After cleaning data files, we merged the datasets into one single file using the SAS enterprise guide, and the final merged file is named "injoint_all_wdemolitiondummy."
- We created some dummy variables for demolition fees and paid demolition fees.

Detailed processing steps for data files are as follows:

1. Randomly sampled parcel ID in the tax aggregation data and compared the tax value with the information on the Shelby county trustee website
2. Selected the tax aggregation data from a year with the most data points documented.
3. Cleaned up all types of errors in Windshield data.
4. Converted Parcel IDs in the above five datasets into the same format.
5. Inner joined all the datasets and created a dummy variable for demolition fees.

Visualization using tableau (descriptive analysis)

Explored dataset using tableau:

Examining 2010-year Violent Crime.

Of 98,365 properties, 6% were documented as violent crimes, with the maximal 44 crime incidents and 6113 total crimes. Further, exploring the areas where has spotted most of the crime. Parcel IDs starting with "07" was among the most affected, having a total crime rate of 1,872 and 7% of crime and the maximum number of crimes equaling 44.

The area starting with "05" showed a high percentage of crime in 2010, resulting in 637 crime incidents and a 6% crime rate, with the maximum number equaling 17.

The area starting with "06" has shown a high percentage of crime, which equaled 6% and had a total crime number of 588 and a maximum of 33 for one parcel.

The crime area beginning with the code "04" showed 578 total violent crimes equaling a 5% crime rate from that area, with the maximum number of violent crime incidents reaching 20.

Parcel ID starting with "09" had 695 crime incidents with a maximum of 33 crimes. Only 5% of those crimes in this area were recorded.

Compared to 2011, the crime in the top crime areas for 2010 has been slightly reduced, particularly in earlier affected regions. The percentage change (2011 vs. 2010) spreadsheet is attached:

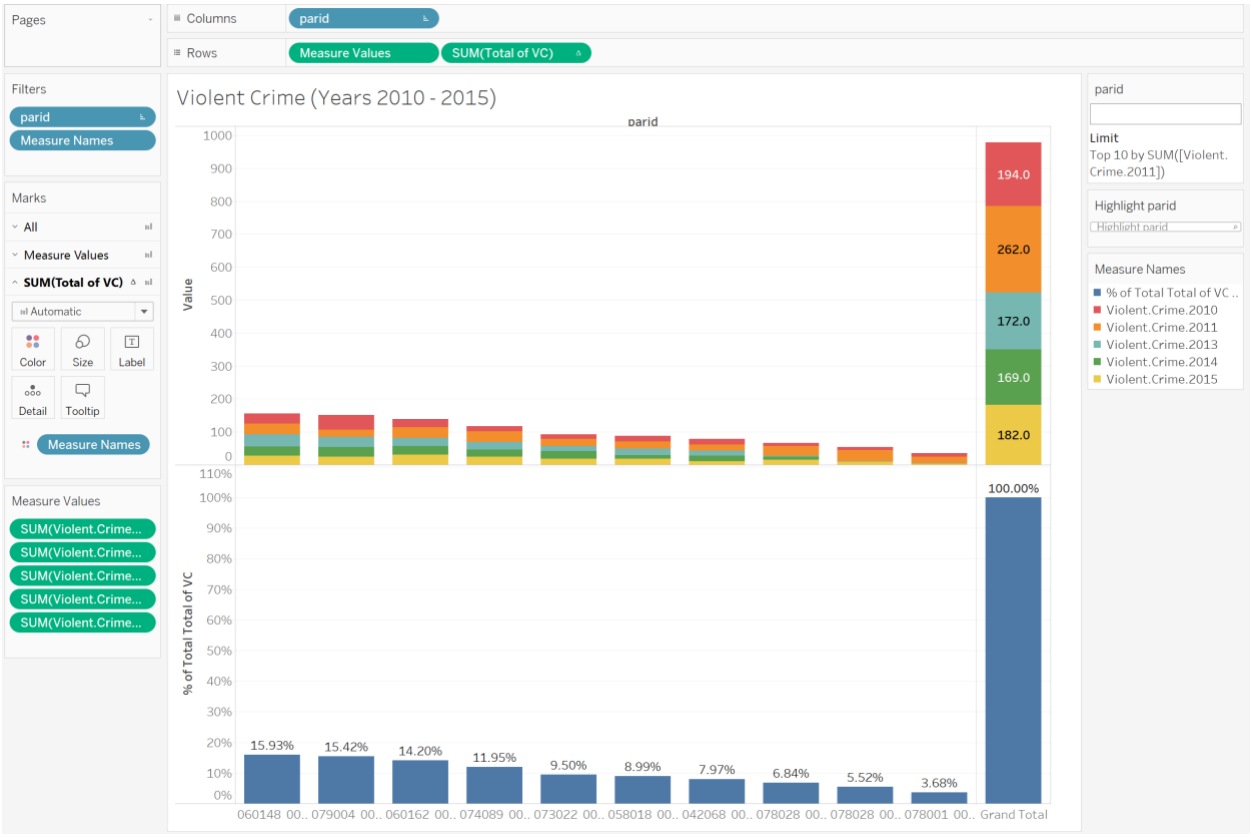
Columns: Measure Names

Rows: VC % Change 2011 vs.. | parid

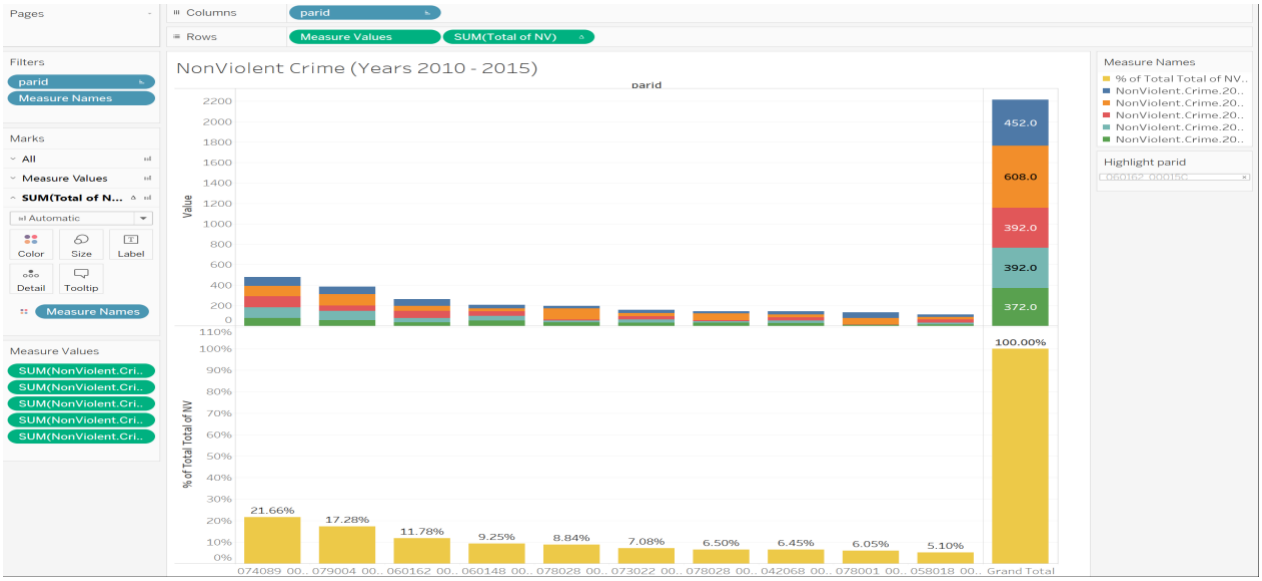
VC %..	parid	PercentageCal	Violent.Crime.2010	Violent.Crime.2011
Down	079004 00019	0.50	44.00	22.00
	093100 00124	0.39	33.00	13.00
	073017 00195	0.05	21.00	1.00
	041034 00127	0.55	20.00	11.00
	072047 00274	0.70	20.00	14.00
	079087 00059C	0.55	20.00	11.00
UP	060148 00135	1.03	31.00	32.00
	060162 00015C	1.28	25.00	32.00

The following chart depicts violent crime results comparing top crime happening parcels over the years, showing that the most crime happens in parcel # 060148 00135.

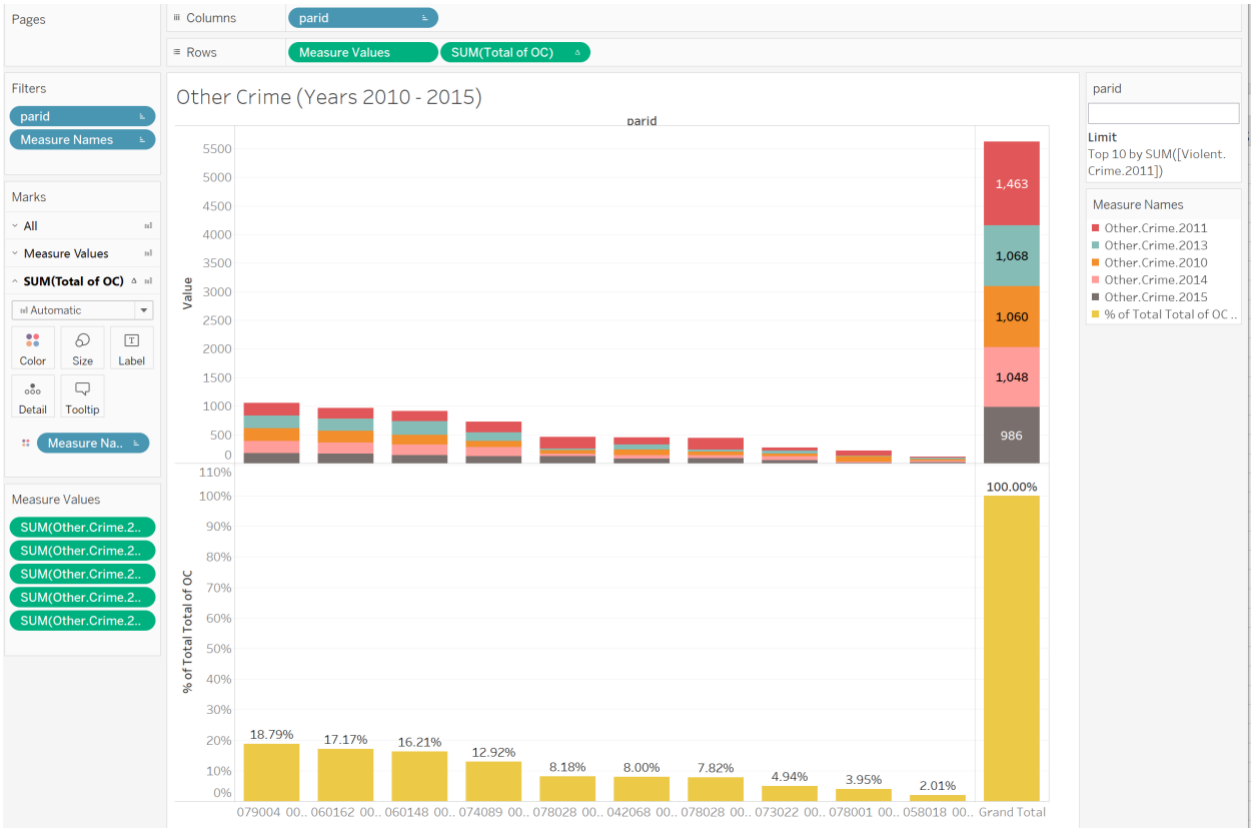
Yearly, the following table shows that in 2010, there was a total of 194 violent crimes: 262 crimes in 2011, 172 in 2012, 169 in 2014, and 182 in 2015. If focusing on the top ten parcels, the most affected by crime is parcel ID 060148 00135, showing 15.93% of the total offense for the time from 2010 to 2015, followed by parcel #079004 00019, which demonstrated an aggregated crime rate of 15.42% and then by the parcel ID 060162 00015C depicting a 14.20%.



Nonviolent crime showed the following results for the same top 10 parcels showing property ID 074089 00048 the most affected (21.66% of total nonviolent crime) by nonviolent crime; followed by parcel ID 079004 00019 with a percentage of 17.28%. Property ID 060148 00135 with the most violent crimes revealed 9.25% of the nonviolent crime during the five years.

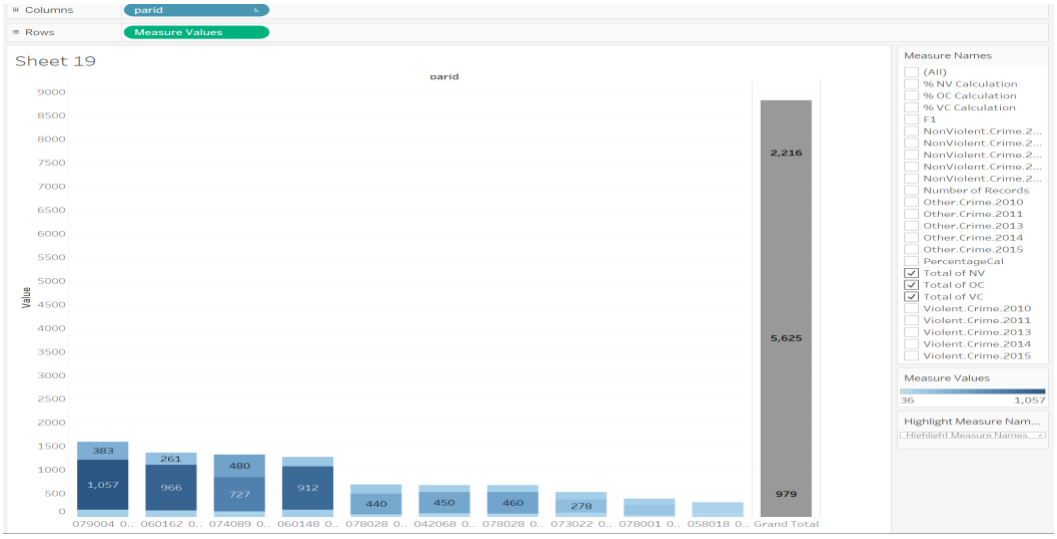


Investigate other crimes for the same properties, and the following results are drawn:

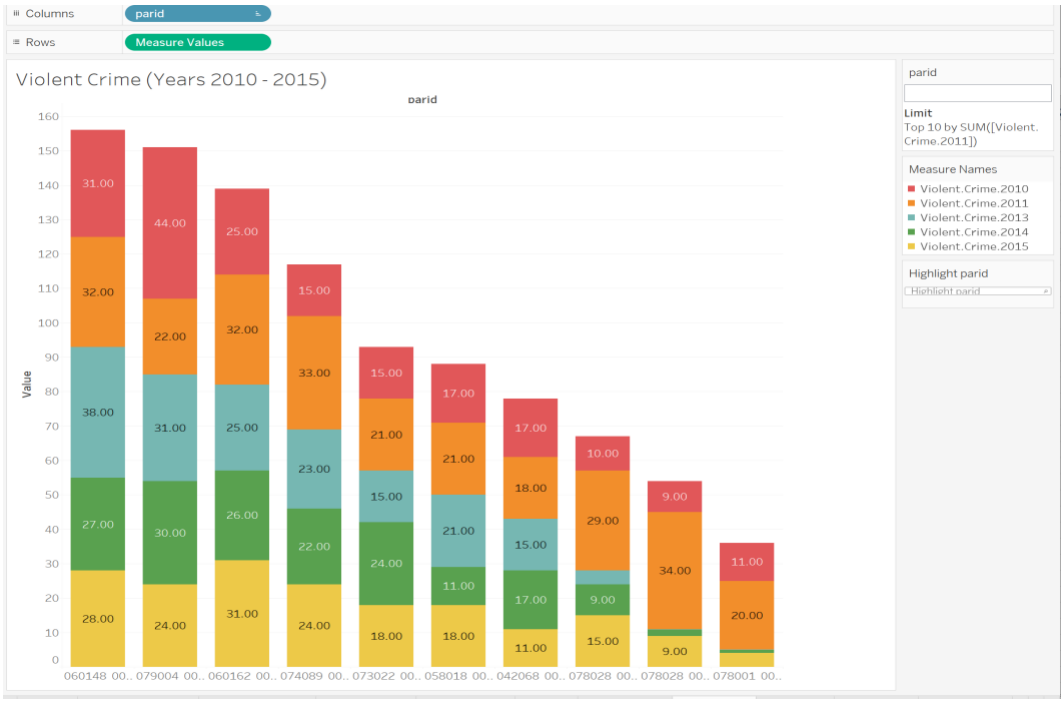


The property ID 060148 00135 showed 16.21% other crimes, yet it ranked in the top 4th place by total crime counts.

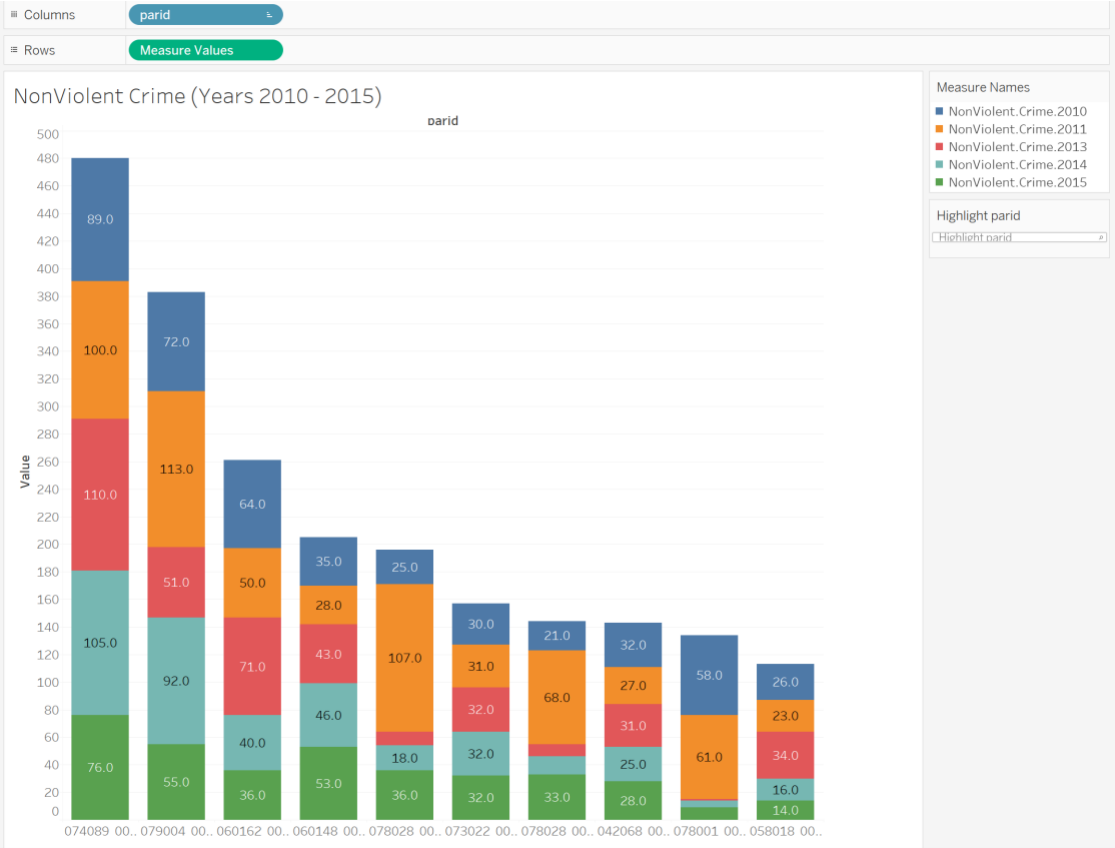
The following chart shows the top 10 properties ranked by violent, nonviolent, and other crimes. The property ID 079004 00019 is the most affected by other crimes occurring in the area.



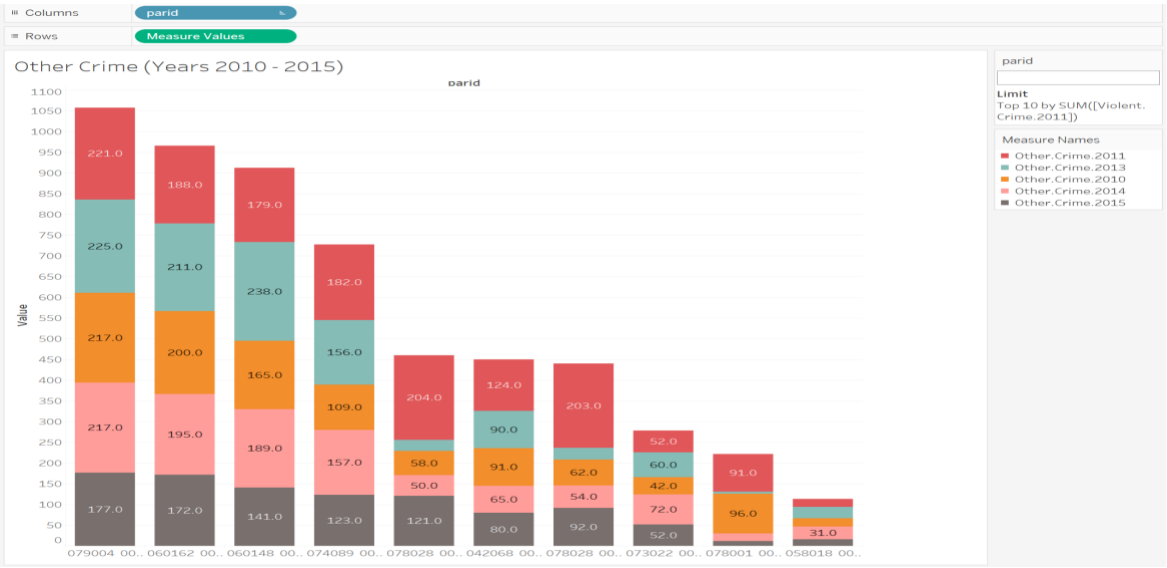
Violent crime results:



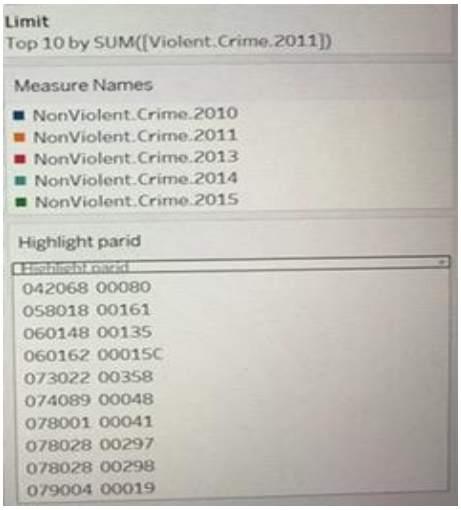
Nonviolent crime results:



Other crime results:

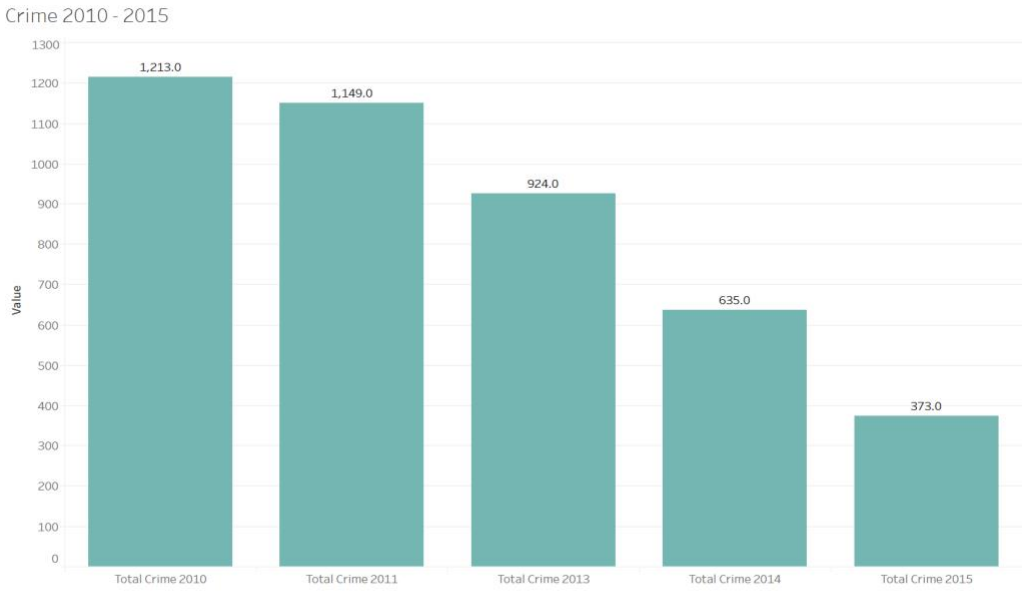


A comparison of all the crimes of all time revealed the pattern displaying most of the crimes are going in the areas starting with codes 06 and 07; therefore, further, we would like to pay attention to these parcels to investigate the causes.



Looking at the bigger picture, we have explored and identified the following information about our variables:

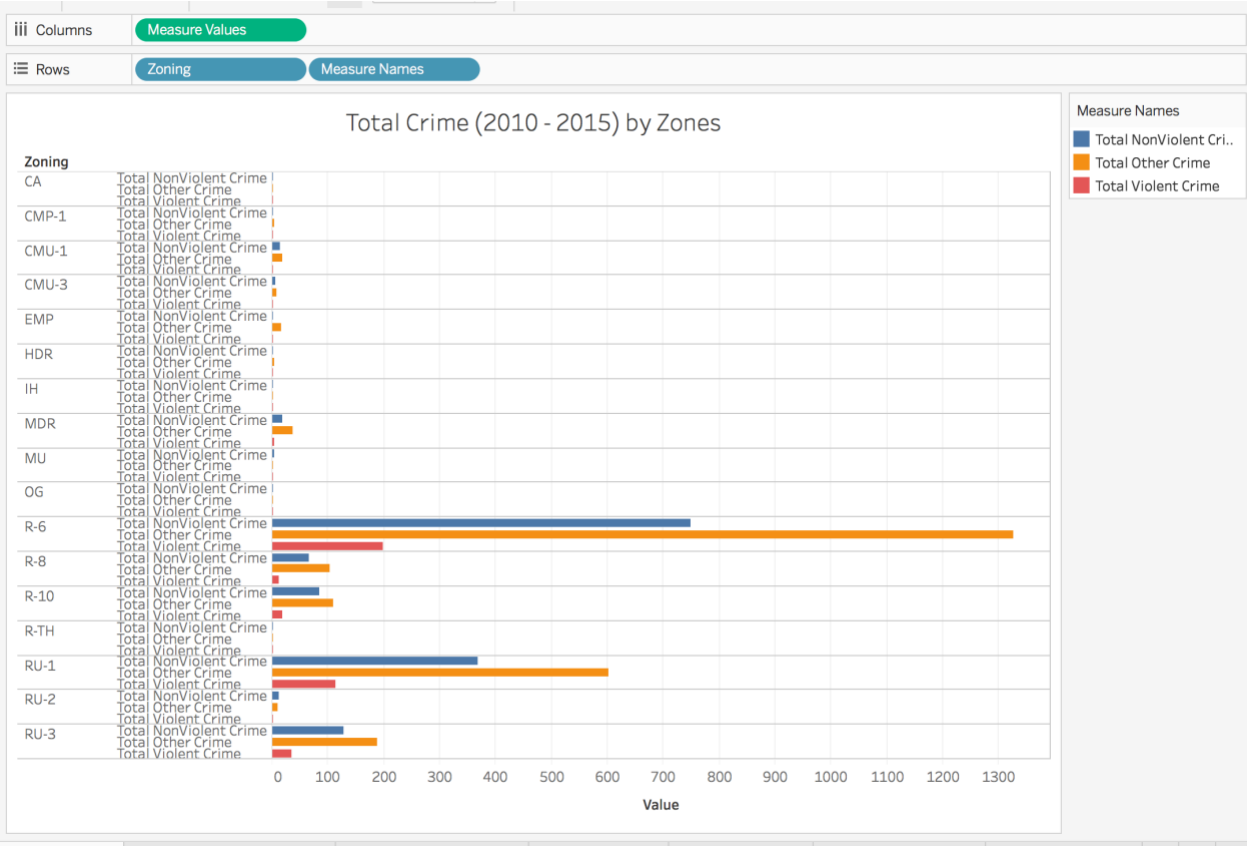
During the period from 2010 to 2015, total crime decreases. The following chart shows the crime decrease.



The majority of the crime occurs mainly in Zones R-6 and RU-1. Zones R-6, a residential single-family district, and RU-1, a residential urban district, show the most "activity." Residential

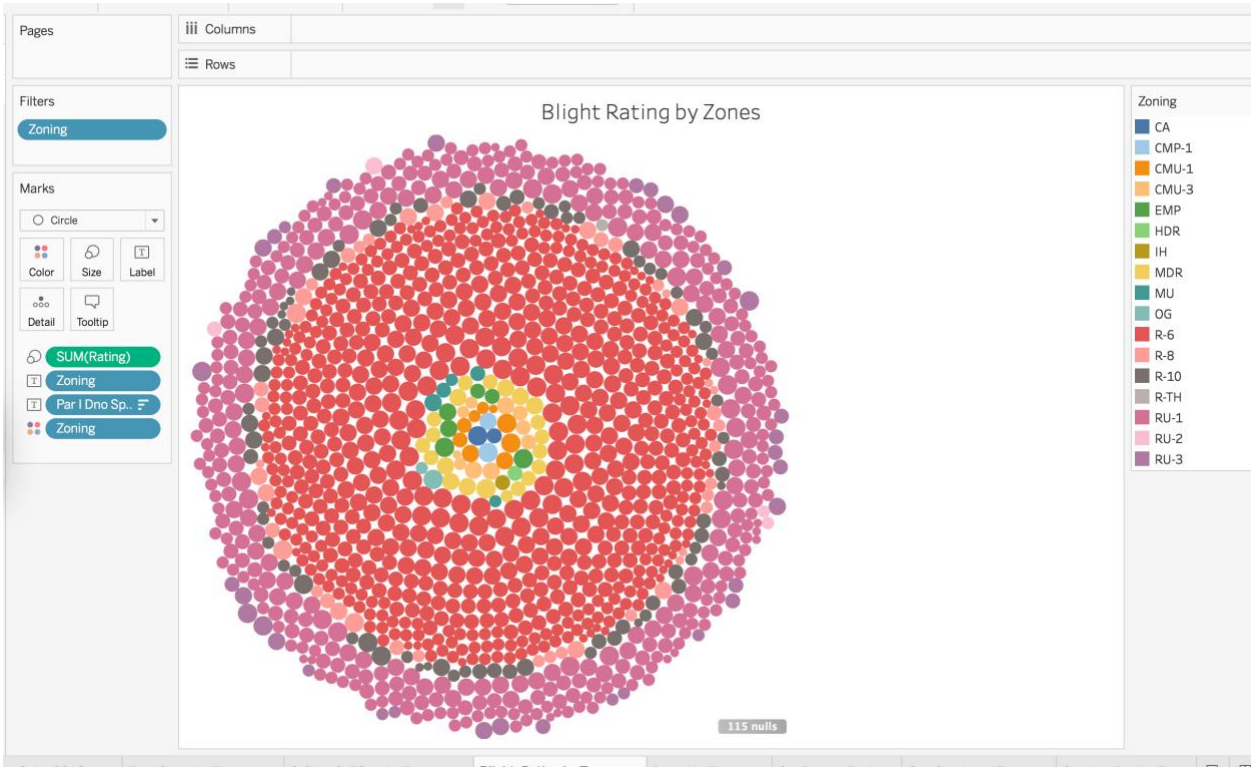
development in the RU-1 District allows various housing types, including single-family detached (conventional, side yard house, cottage) and single-family attached (semi-attached, two-family). New RU-1 districts should have a shared street network and are generally located at least 500 to 1,000 feet from a CMU-1, CMU-2, CMU-3, or CBD district or at least 500 to 1,000 feet from an arterial.

It was a fascinating observation that the most frequent Sales occur in the most affected by all types of crime areas.



It was an exciting observation that the most frequent Sales occur in the most affected by all types of crime areas.

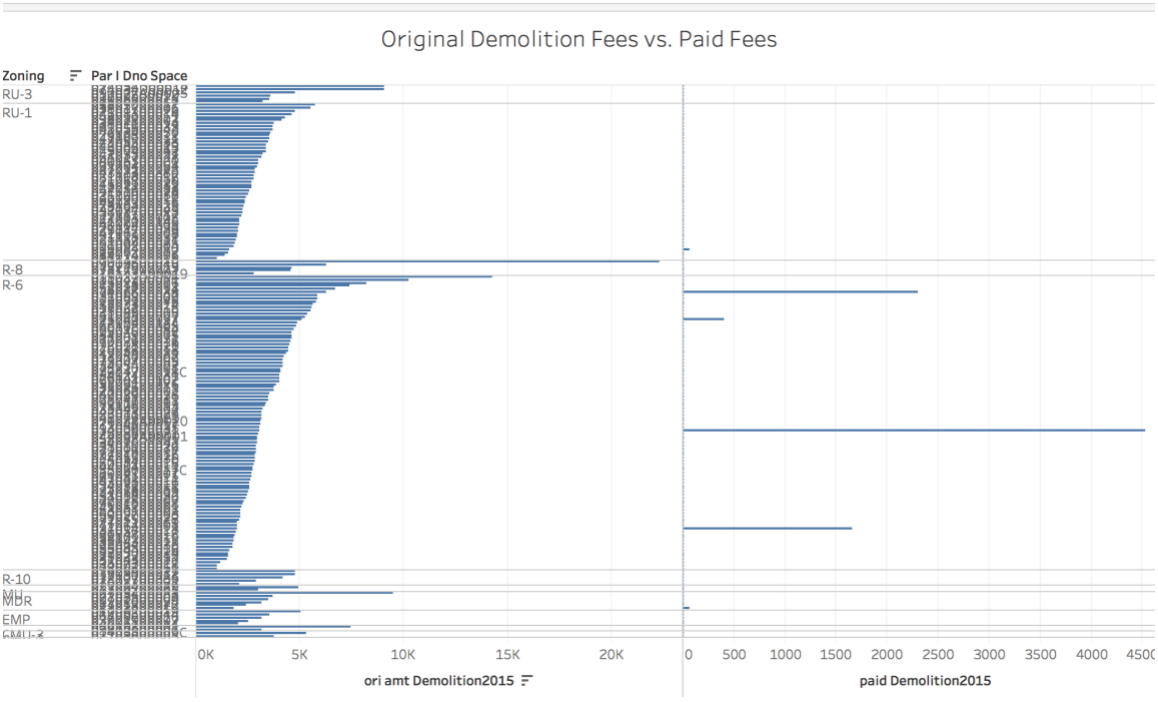
We defined blighted properties by rating them on a scale from 1 to 5, where the rating of 5 identifies the most affected by the blight areas. Besides, the same zones, R-6 and RU-1, also show the highest possible blighted rating.



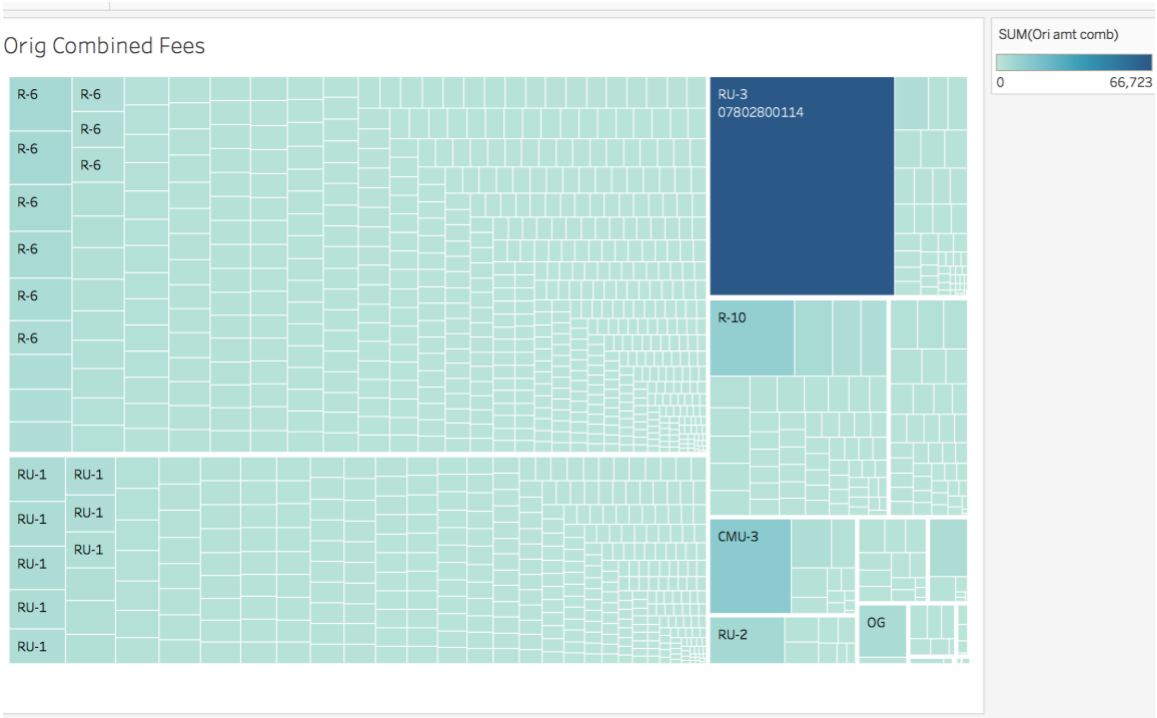
Since we have decided on the target variable as demolition fees, we want to use a visual presentation to examine the pattern potentially carried on by the demolition fees and the areas with the highest amount of original demolition fees. In addition, if those fees were collected on time, how this information relates to the zones affected the most by crime and sales areas.

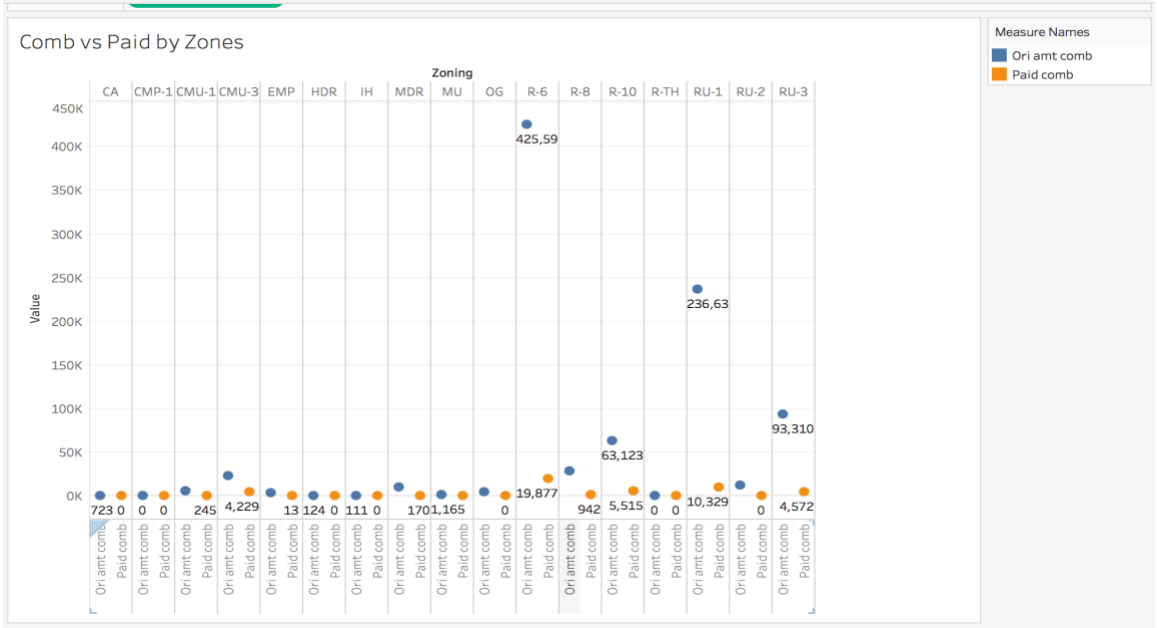
The following picture shows the original demolition fees, depicting the same affected by Crime and Sales areas.

Original demolition fees were paid inconsistently, and the percentage of fees collected is insignificant compared to the actual amount issued.



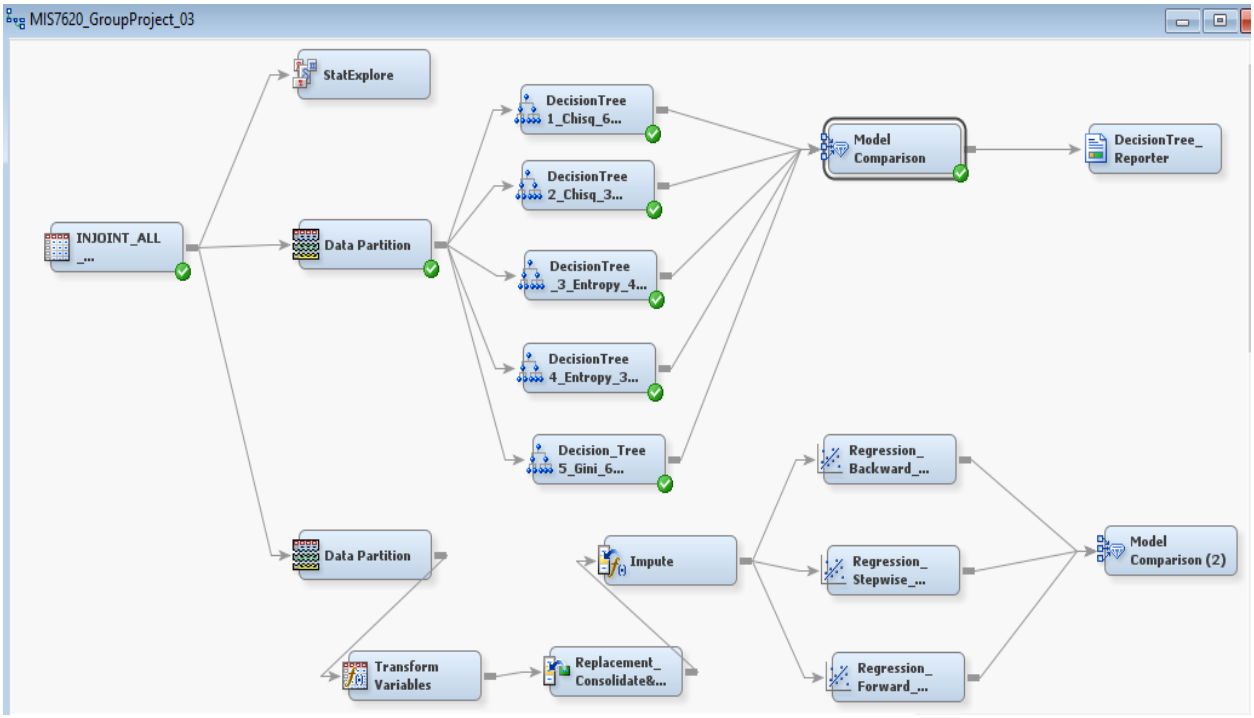
Combined original tax fees were recorded for the same areas, R-6 and RU-1. Combined fees, just like demolition fees, were not getting paid or, even if they were, the payments were insignificant compared to the outstanding fees issued due to the overall low-income residents.





Using descriptive statistics through simple summaries, we visualized our datasets to determine if there are any general trends and patterns worth exploring.

Predictive Analysis Using SAS Enterprise Miner



We used different predictive analytics techniques like Decision trees, Regressions, and Neural Networks to predict our target variable: Dummy_variable_demolition

Variables - Ids

(none) not Equal to

Columns: Label Mining Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Dummy_Variable_Demolition	Target	Binary	No		No	.	.
Dummy_Variable_PaidDemolition	Rejected	Binary	No		No	.	.
Dummy_Variable_PaidDemolition1	Rejected	Binary	No		No	.	.
Late_fees_comb	Input	Interval	No		No	.	.
Net_due_comb	Rejected	Interval	No		No	.	.
NonViolent_Crime_2010	Input	Nominal	No		No	.	.
NonViolent_Crime_2011	Input	Nominal	No		No	.	.
NonViolent_Crime_2013	Input	Nominal	No		No	.	.
NonViolent_Crime_2014	Input	Nominal	No		No	.	.
NonViolent_Crime_2015	Input	Nominal	No		No	.	.
Ori_amt_comb	Input	Interval	No		No	.	.
Other_Crime_2010	Input	Nominal	No		No	.	.
Other_Crime_2011	Input	Nominal	No		No	.	.
Other_Crime_2013	Input	Nominal	No		No	.	.
Other_Crime_2014	Input	Nominal	No		No	.	.
Other_Crime_2015	Input	Nominal	No		No	.	.
Paid_comb	Input	Interval	No		No	.	.
ParIDnoSpace	Rejected	Nominal	No		No	.	.
Sale_2010	Input	Nominal	No		No	.	.
Sale_2011	Input	Nominal	No		No	.	.
Sale_2012	Input	Nominal	No		No	.	.
Sale_2013	Input	Nominal	No		No	.	.
Sale_2014	Input	Nominal	No		No	.	.
Sale_2015	Input	Nominal	No		No	.	.
Sale_2016	Input	Nominal	No		No	.	.

Nodes used before the decision tree, regression, or neural network model are:

- **Data partition**: available data was divided into 50% used as training data, and the other 50% served as validation data.
- **Impute**: for handling missing values. Property settings are as follows:

Property	Value
General	
Node ID	Impt
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Count
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Mean
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	.
Method Options	
Random Seed	12345
Tuning Parameters	...
Tree Imputation	

- **Transform variables**: This node is added to reduce the skewing effect of extreme values on the target prediction. The skewness of these variables dropped during this process.
-

Following screenshot showing the transformed variables

Late_fees_comb	**Log**	**4**	Input	Interval
Net_due_comb	**Default**	**4**	Rejected	Interval
NonViolent_Crime_2010	**Default**	**4**	Input	Nominal
NonViolent_Crime_2011	**Default**	**4**	Input	Nominal
NonViolent_Crime_2013	**Default**	**4**	Input	Nominal
NonViolent_Crime_2014	**Default**	**4**	Input	Nominal
NonViolent_Crime_2015	**Default**	**4**	Input	Nominal
Ori_amt_comb	**Log**	**4**	Input	Interval
Other_Crime_2010	**Default**	**4**	Input	Nominal
Other_Crime_2011	**Default**	**4**	Input	Nominal
Other_Crime_2013	**Default**	**4**	Input	Nominal
Other_Crime_2014	**Default**	**4**	Input	Nominal
Other_Crime_2015	**Default**	**4**	Input	Nominal
Paid_comb	**Log**	**4**	Input	Interval

late_fees_Sanitation2015	**Log**	**4**	Input	Interval
late_fees_Weed_Cutting20	**Log**	**4**	Input	Interval
litter	**Default**	**4**	Input	Nominal
net_due_CBID2015	**Default**	**4**	Rejected	Binary
net_due_Demolition2015	**Default**	**4**	Rejected	Interval
net_due_N2015	**Default**	**4**	Rejected	Binary
net_due_Sanitation2015	**Default**	**4**	Rejected	Interval
net_due_Weed_Cutting201	**Default**	**4**	Rejected	Interval
occupancy	**Default**	**4**	Input	Nominal
ori_amt_CBID2015	**Default**	**4**	Input	Binary
ori_amt_Demolition2015	**Default**	**4**	Rejected	Interval
ori_amt_N2015	**Default**	**4**	Input	Binary
ori_amt_Sanitation2015	**Log**	**4**	Input	Interval
ori_amt_Weed_Cutting2015	**Log**	**4**	Input	Interval
other	**Default**	**4**	Input	Binary
paid_CBID2015	**Default**	**4**	Rejected	Unary
paid_Demolition2015	**Default**	**4**	Input	Nominal
paid_N2015	**Default**	**4**	Rejected	Unary
paid_Sanitation2015	**Default**	**4**	Input	Nominal
paid_Weed_Cutting2015	**Log**	**4**	Input	Interval

Decision tree: we created decision trees considering the different number of maximum branches, maximum depth, and nominal target criteria as Entropy or Gini. Properties snap for one of the trees is as shown below.

Property	Value
General	
Node ID	Tree3
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	Entropy
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	4
Maximum Depth	10
Minimum Categorical Size	5
Node	

After comparing different decision tree models, decision tree 1 with six maximum branches is chosen as the best model. The minimum validation misclassification for the same is 0.128

Fit Statistics

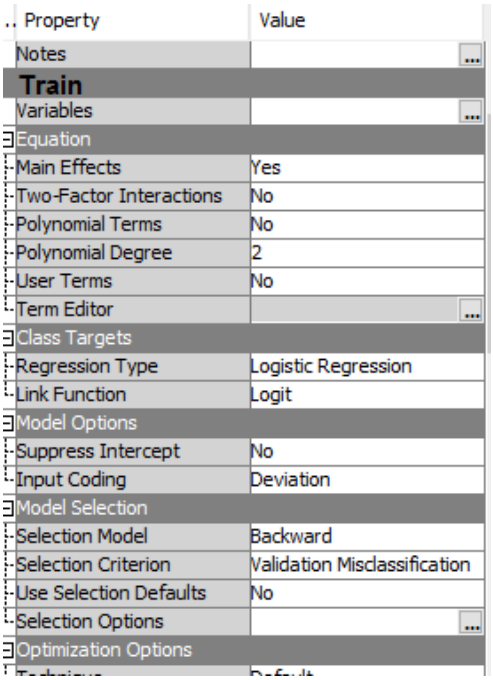
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error
Y	Tree	Tree	DecisionTr...	Dummy_Va...	DVDA	0.128686	748	0.13369	0.974359
	Tree2	Tree2	DecisionTr...	Dummy_Va...	DVDA	0.128686	748	0.141711	0.869173
	Tree3	Tree3	DecisionTr...	Dummy_Va...	DVDA	0.131367	748	0.131016	0.951691
	Tree5	Tree5	Decision_T...	Dummy_Va...	DVDA	0.132708	748	0.11631	0.967213
	Tree4	Tree4	DecisionTr...	Dummy_Va...	DVDA	0.140751	748	0.120321	0.962547

Regression: we did the regression modeling with the following methods,

1. Forward Regression

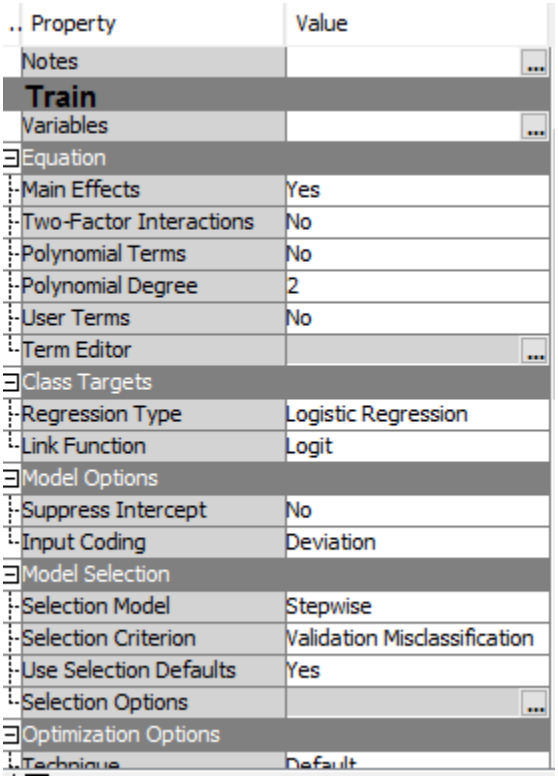
Property	Value
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Validation Misclassification
Use Selection Defaults	No
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	No

2. Backward Regression



Property	Value
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Backward
Selection Criterion	Validation Misclassification
Use Selection Defaults	No
Selection Options	
Optimization Options	

3. Stepwise Regression



Property	Value
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Misclassification
Use Selection Defaults	Yes
Selection Options	
Optimization Options	
Technique	Default

Fit statistics for regression are listed below. The reg model is the best, and its validation misclassification rate is 0.117. it is the minimum rate among all models we created.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function
Y	Reg6	Reg6	Regression...	Dummy_Va...	DVDA	0.117962	513.7846	0.104089	0.322049	732	16	748	1496	481.7846
	Reg7	Reg7	Regression...	Dummy_Va...	DVDA	0.117962	513.7846	0.104089	0.322049	732	16	748	1496	481.7846
	Reg5	Reg5	Regression...	Dummy_Va...	DVDA	0.131367	503.3824	0.093017	0.293705	716	32	748	1496	439.3824
	Reg4	Reg4	Regression...	Dummy_Va...	DVDA	0.147453	499.9952	0.083413	0.264703	696	52	748	1496	395.9952

Neural Network:

The fit statistics for the neural network is as shown in the figure below:

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Total Degrees of Freedom	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Number of Estimated Weights	Train: Akaike's Information Criterion	Train: Schwarz's Bayesian Criterion	Train: Average Squared Error	Train: Maximum Absolute Error	Train: Divisor for ASE	Train: Sum of Frequencies	Tra... Av... Sq... Err...
Y	Neural3	Neural3	Neural Net...	Dummy_Va...	DVDA	0.119303	748	-261	1009	1009	.	.	0.081204	0.979464	1496	748	
	Neural2	Neural2	Neural Net...	Dummy_Va...	DVDA	0.123324	748	645	103	103	739.0663	1214.659	0.113483	0.980973	1496	748	
	Neural	Neural	Neural Net...	Dummy_Va...	DVDA	0.126005	748	243	505	505	1471.511	3803.3	0.096403	0.987235	1496	748	

The degree of freedom is 748, and the validation rate is 0.119. The model named "network 3" was the best-optimized model. Following re the properties for model comparison

Property	Value
Train	
Variables	...
Assessment Reports	
Number of Bins	20
ROC Chart	Yes
Recompute	No
Model Selection	
Selection Data	Default
Selection Statistic	Default
Grid Selection Statistic	Default
Selection Table	Train
Selection Depth	10
Score	
Selection Editor	...
Report	
Selected Model	
Target	Dummy_Variable_Demolitio
Model Node	Neural3
Model Description	Neural Network_HiddenU6
Selection Criteria	Valid: Misclassification Rate
Status	
Create Time	5/4/17 4:45 AM

Using Model comparator and setting show all property to yes, we compared all the models and generated a PDF report using reporter node.

Key Learnings:

- Never assume the datasets are accurate or clean from any source without examining them.
- Identifying the related variables and reducing the unnecessary data complexity will help the SAS EM perform better---- it is not a good idea to shovel the data into EM and rely on it to perform well.
- The output log file provides thorough documentation with statistic calculation, model parameters, and iteration steps, which are helpful for troubleshooting model optimization. Therefore, it is worth learning how to read the file and understand a certain number of those statistics**.**