**Blighted Data property**

Blight can be considered as any abandoned and/or vacant property, either building or home, that is 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 bring down the neighborhood by reducing its property value. However, who can we say is ultimately responsible for blight? We can’t blame this problem on one individual person. All Memphians from top positions to bottom, are responsible for this situation. Now where does Memphis stand when it comes the rating of blight? “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”

US is facing the problems 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 are caused because of – Unemployment, depopulation problems, promote crime, loss of deindustrialization. Blight affected areas are characterized by 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, which kept coming to mind, was why these properties, during the time we have chosen to research, reached such a depleted state to cause these affects. We began to consider the period after The Great Recession and the huge amount of foreclosures during the events leading up through this time period, we conclude that much of the data information we explored came about either directly or indirectly from the results of foreclosures from 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 at such time as the properties became vacant and eventually abandoned, at which 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, demolition, for insight into possible solutions and precursors for further consideration of prevention.

Even though foreclosure is costly to the owners who are in default, the financial losses spiral downward from there immediately beginning with 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 vacant or abandoned, its property loss effect is more than half foreclosure’s affect with large increases in crime becoming a major contributor.

One of the prevention considerations we considered is demolition. The dummy variable of the demolition fees would be used as an indicator of blight property. Except the demolition fees, other data or information doesn’t appear to be directly related to the blight property. 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 fess, we will attempt to identify properties that are possible to become blighted. This information can provide guidance to local community and government for long-term development plan and help minimize the bad 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 survey conducted in 2010 about the structural issues of parcels in Memphis as ‘Windshield.Data.csv’

|  |  |
| --- | --- |
| **Datasets** | **Content of each data set** |
| **Windshield** | **The thorough housing information , condition details of all the Memphis and Shebly property, and Memphis and Shelby county regulation and code enforcement, ~150,000 parcel ID. The data set was not cleaned** |
| **Tax aggregation** | **32 statement records of aggregated (sum) of property tax information over the period of 2010-2016, 46,751 parcel ID, data collected from 2010 to 2016.** |
| **MLGW aggregation** | **count how many times 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** | **count how many times sale happens each year. Interval (discrete) data with data range [0, 27], one sale record of each year for 127,609 parcel ID** |
| **Crime aggregation** | **count how many 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 merged the datasets into one single file using SAS enterprise guide. The final merged file is injoint\_all\_wdemolitiondummy
* We created some dummy variables for demolition fees and paid demolition fees.

**Detailed processing steps for data files is as follows:**

* 1. Randomly sampling parcel ID in the tax aggregation data and compare the tax value with the information on the Shelby county trustee website
  2. Select the tax aggregation data from a year that has the most data points documented
  3. Clean up all types of errors in Windshield data
  4. Convert Parcel IDs in the above five datasets into the same format.
  5. Inner joined all the dataset and create a dummy variable for demolition fees

**Visualization using tableau (descriptive analysis)**

Exploring dataset using tableau:

Examining 2010-year Violent Crime.

Out of 98,365 properties, on average, 6% were recorded as violent crimes with a maximum of 44 crime incidents and a total of 6113 crimes. Further, exploring the areas where most of the crime had been spotted, the parcel ID starting with 07 was the most affected with a total crime rate of 1,872 and 7% of crime in the area, with the maximum number of crimes which equals 44.

The Area starting with 05 had been showing a high percentage of crime in 2010 resulting in a total of 637 crime incidents and a 6% crime rate with the maximum number equaling 17.

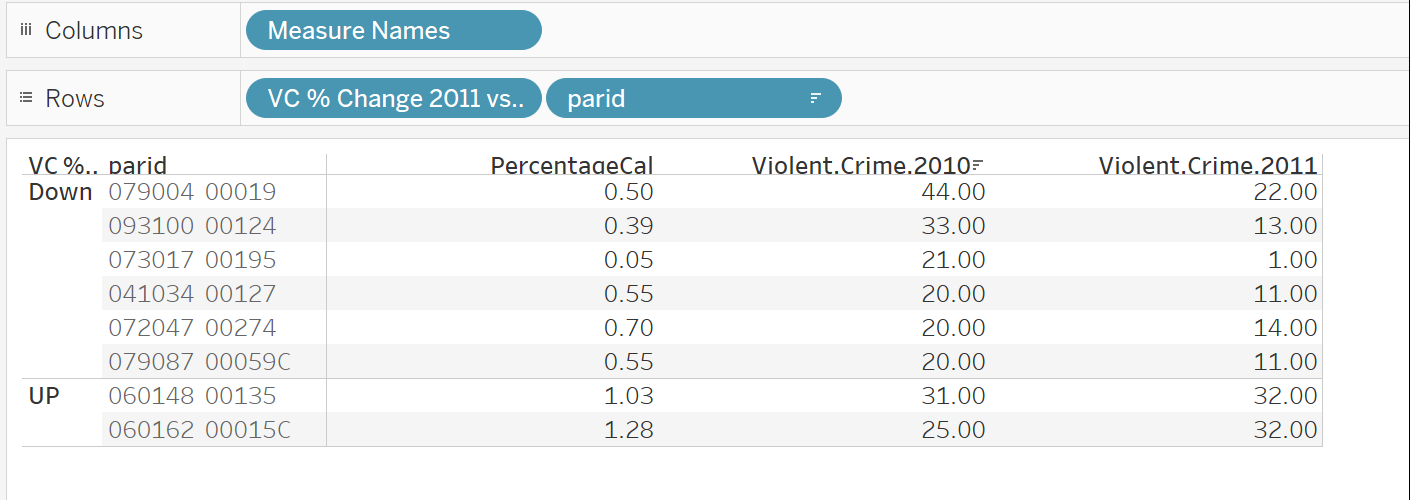
The area starting with 06 has been showing a high percentage of crime as well which equals to 6% with the total number of crime of 588 and maximum of 33 for one parcel.

The crime area starting with the code of 04 is showing a total of 578 violent crime equaling a 5% crime rate in the area with the maximum number of violent crime incidents at 20.

Parcel ID starting with 09 have had 695 crime incidents with the maximum of 33 crime. 5% of crime in this area have been recorded.

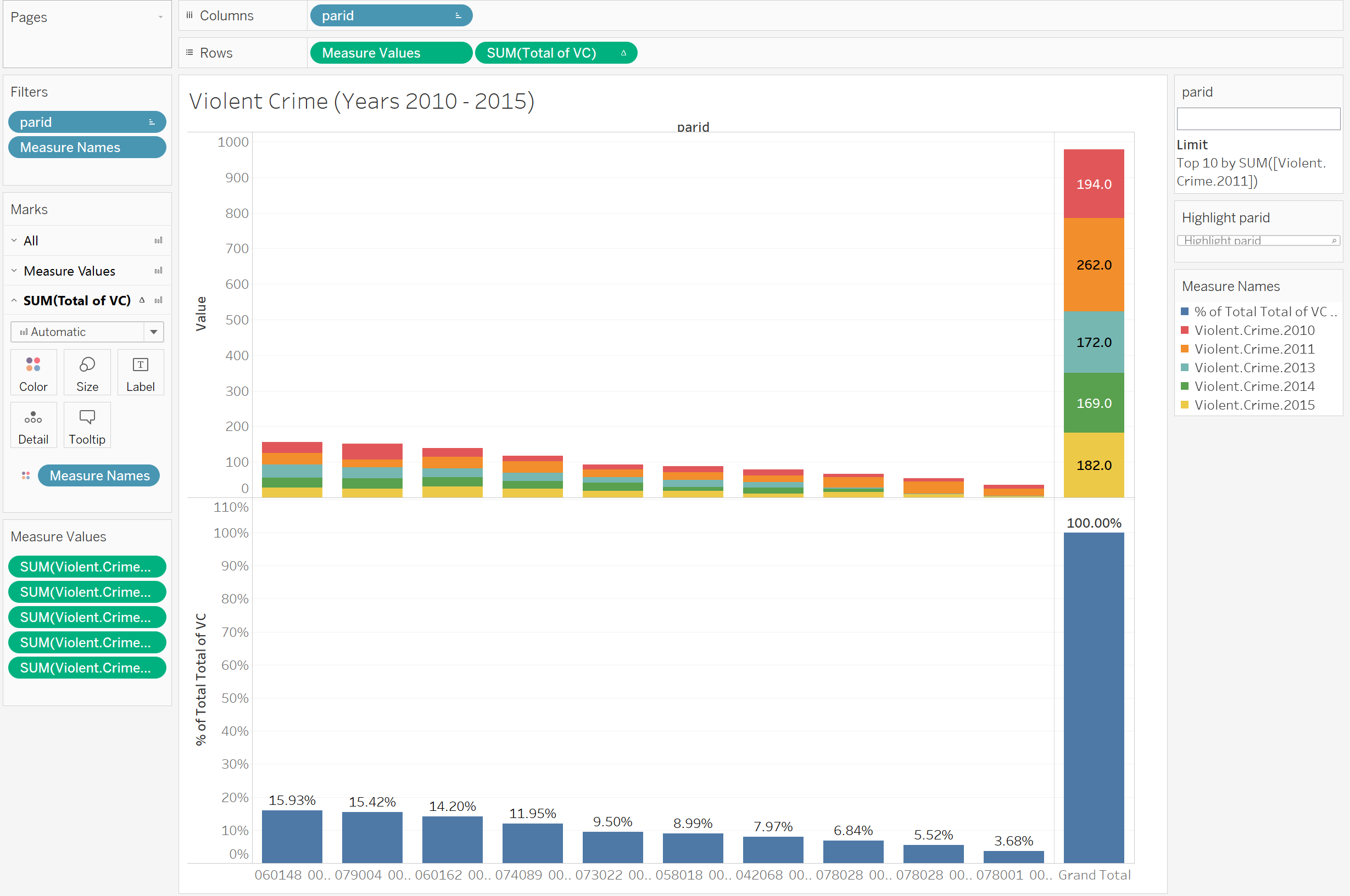
Compared to 2011 year, the crime in the top crime areas for 2010 has been slightly reduced, particularly in affected earlier areas.

The percentage change (2011 vs 2010) spreadsheet has been attached:

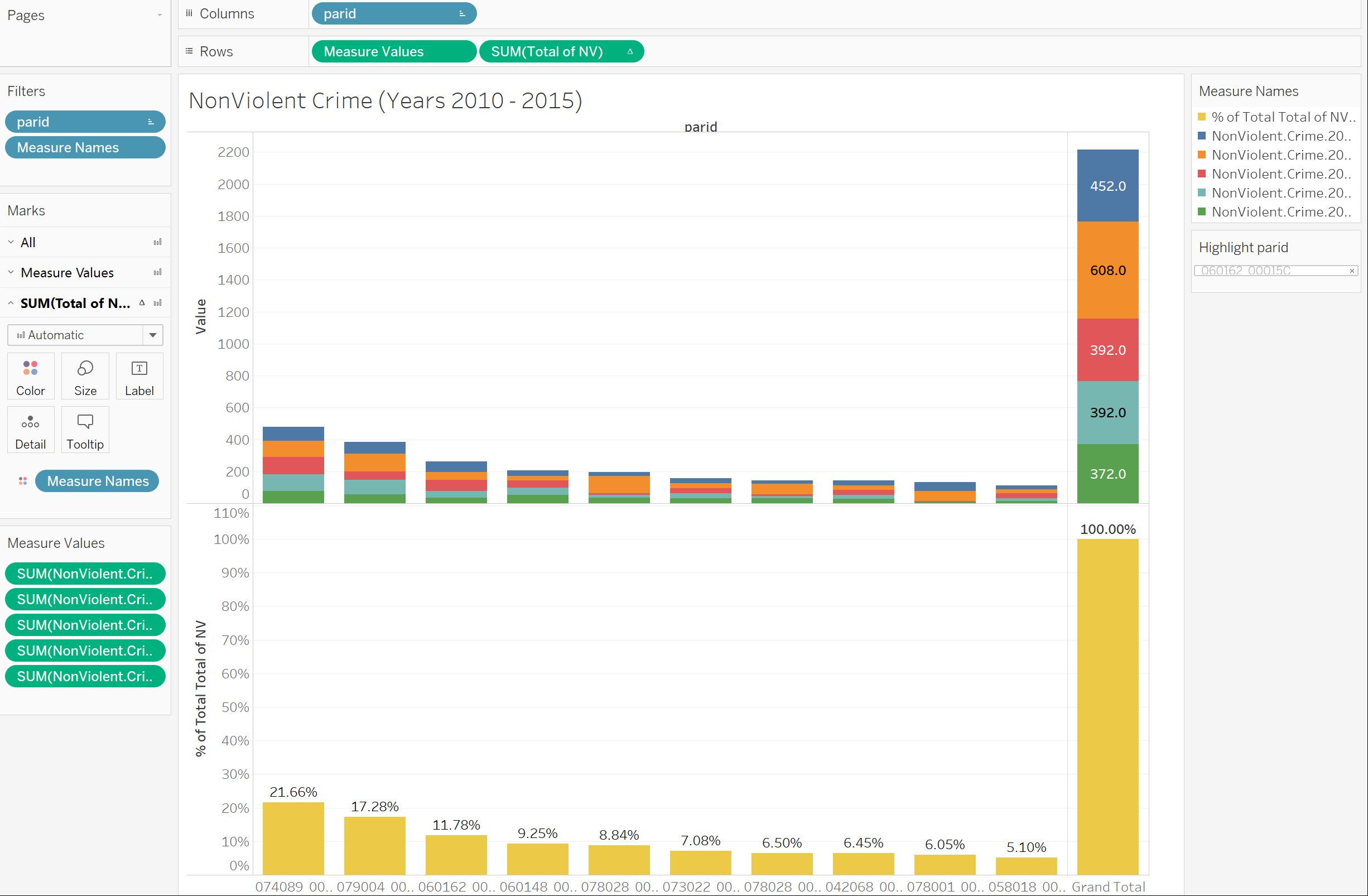


The following chart depicts Violent Crime results for top crime happening parcels compared over years; showing that the most crime happens in parcel # 060148 00135.

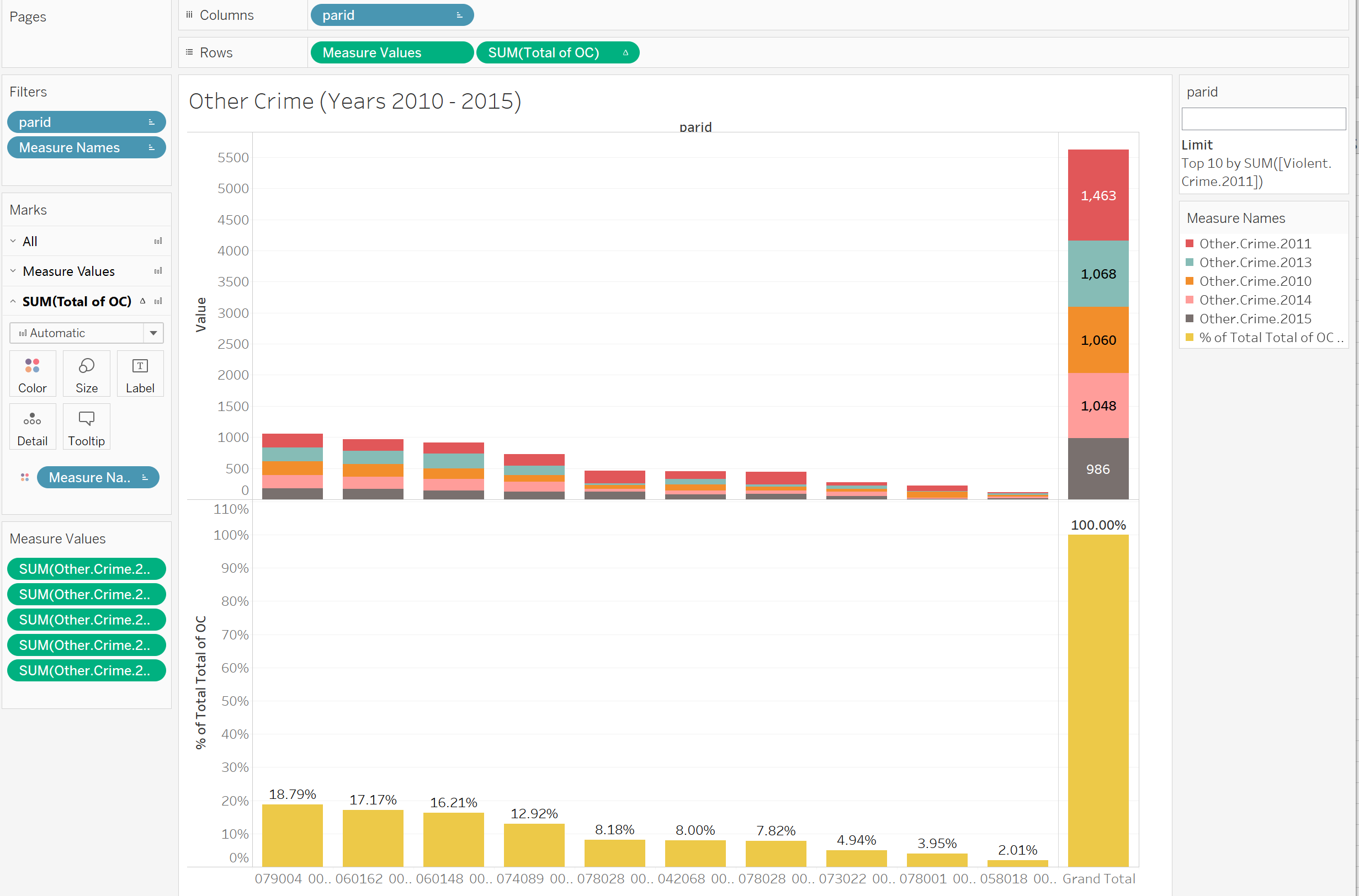
Yearly, the following table shows that in 2010, there was a total of 194 violent crime, in 2011 – 262 violent crimes, in 2012 – 172, in 2014 – 169, and in 2015 – 182. Looking at the top ten parcels, there the most affected by crime in is parcel # 060148 00135 showing 15.93% of the total crime for the period from 2010 – 2015, followed by parcel # 079004 00019 showing 15.42% of the total crime, followed by the parcel # 060162 00015C depicting 14.20% of the total crime.



Nonviolent crime is showing the following results for the same top 10 parcels showing property #074089 00048 the most affected (21.66% of total nonviolent crime) by nonviolent crime; followed by the parcel # -079004 00019 with percentage of 17.28%. Property # 060148 00135 with the most # of violent crime is showing 9.25% of the nonviolent crime during five year period.

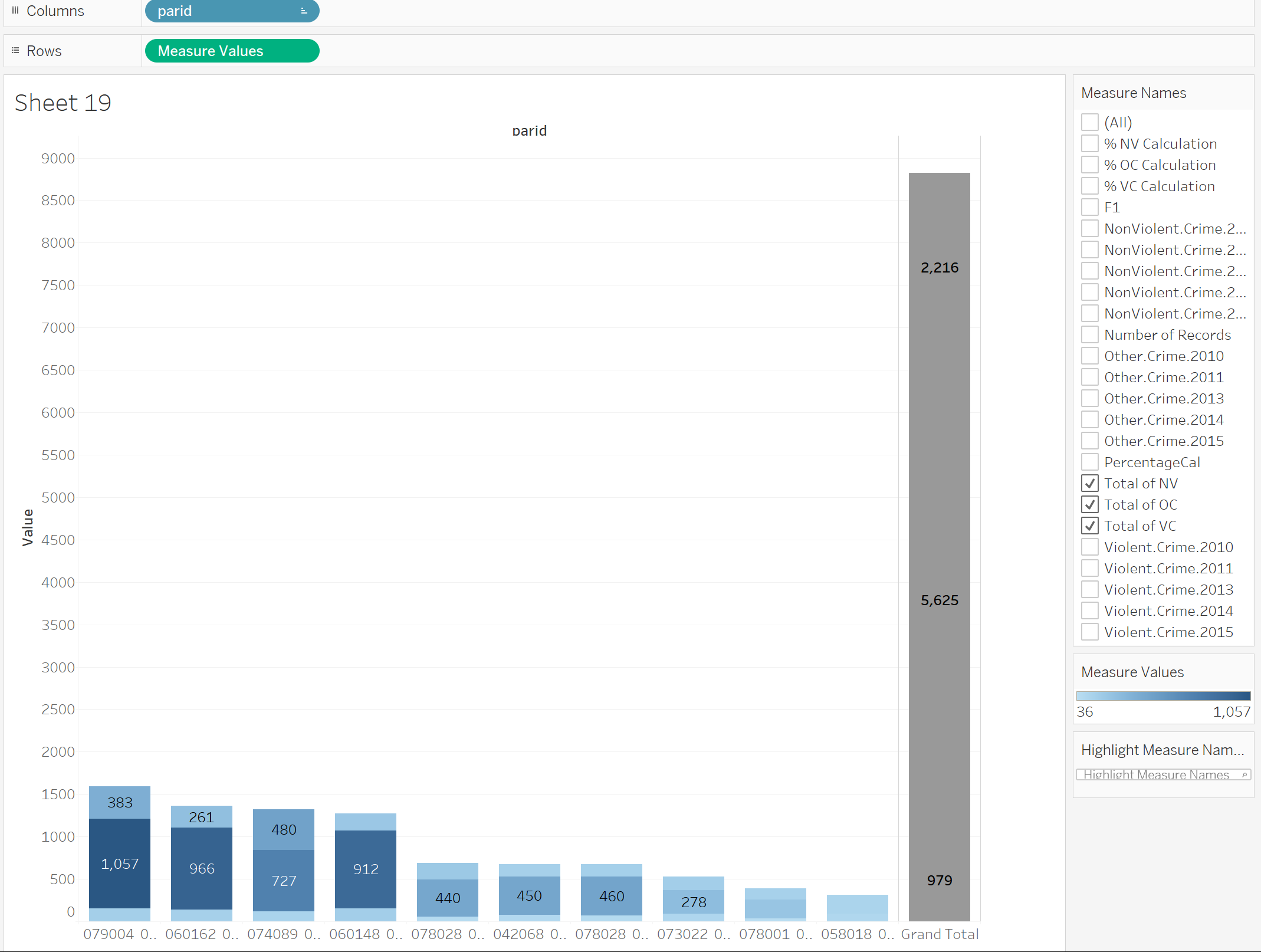


Investigatine Other crime for the same properties, the following results have been drawn:

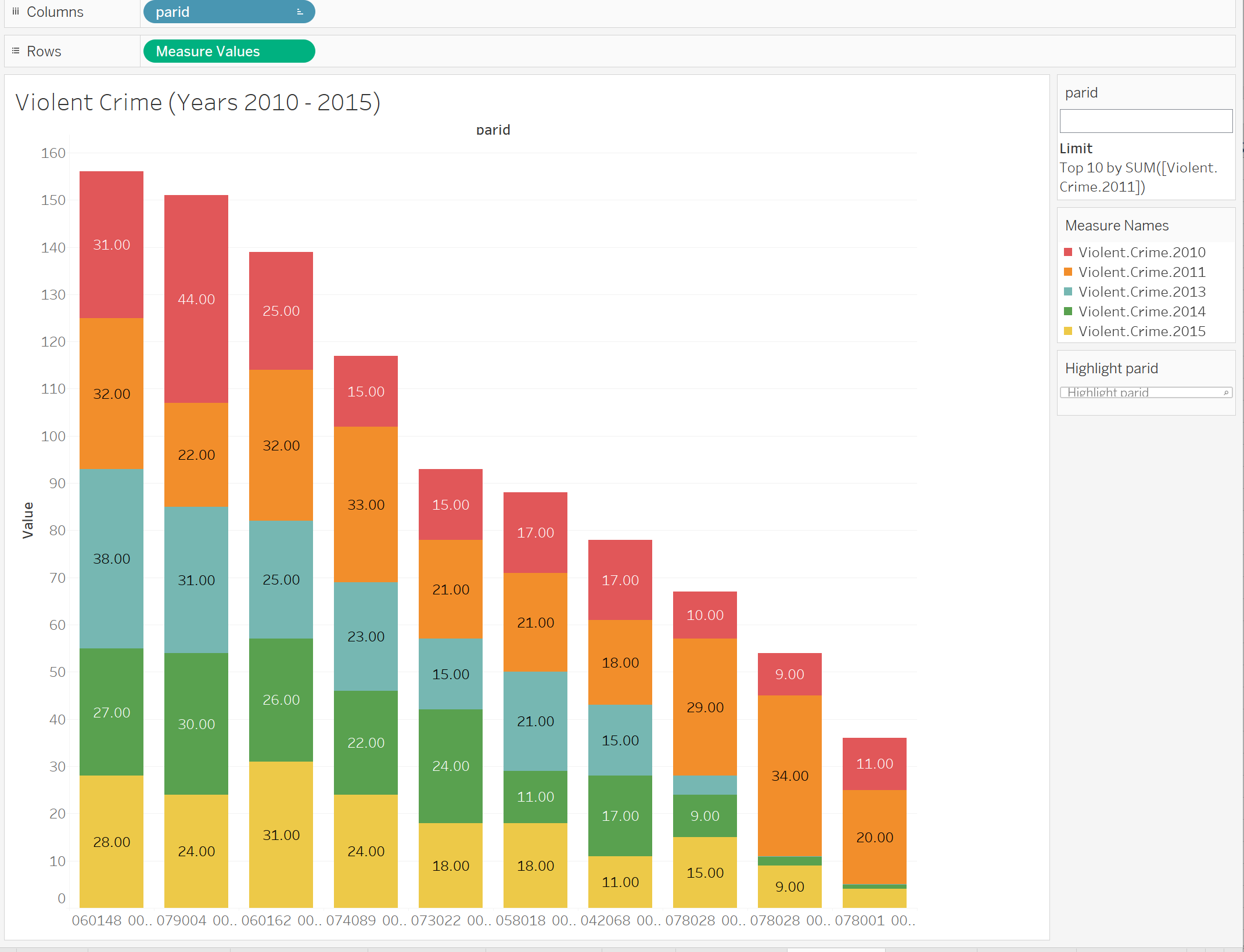


The property #060148 00135 is showing 16.21% other crime; yet it is in 4th place by total crime.

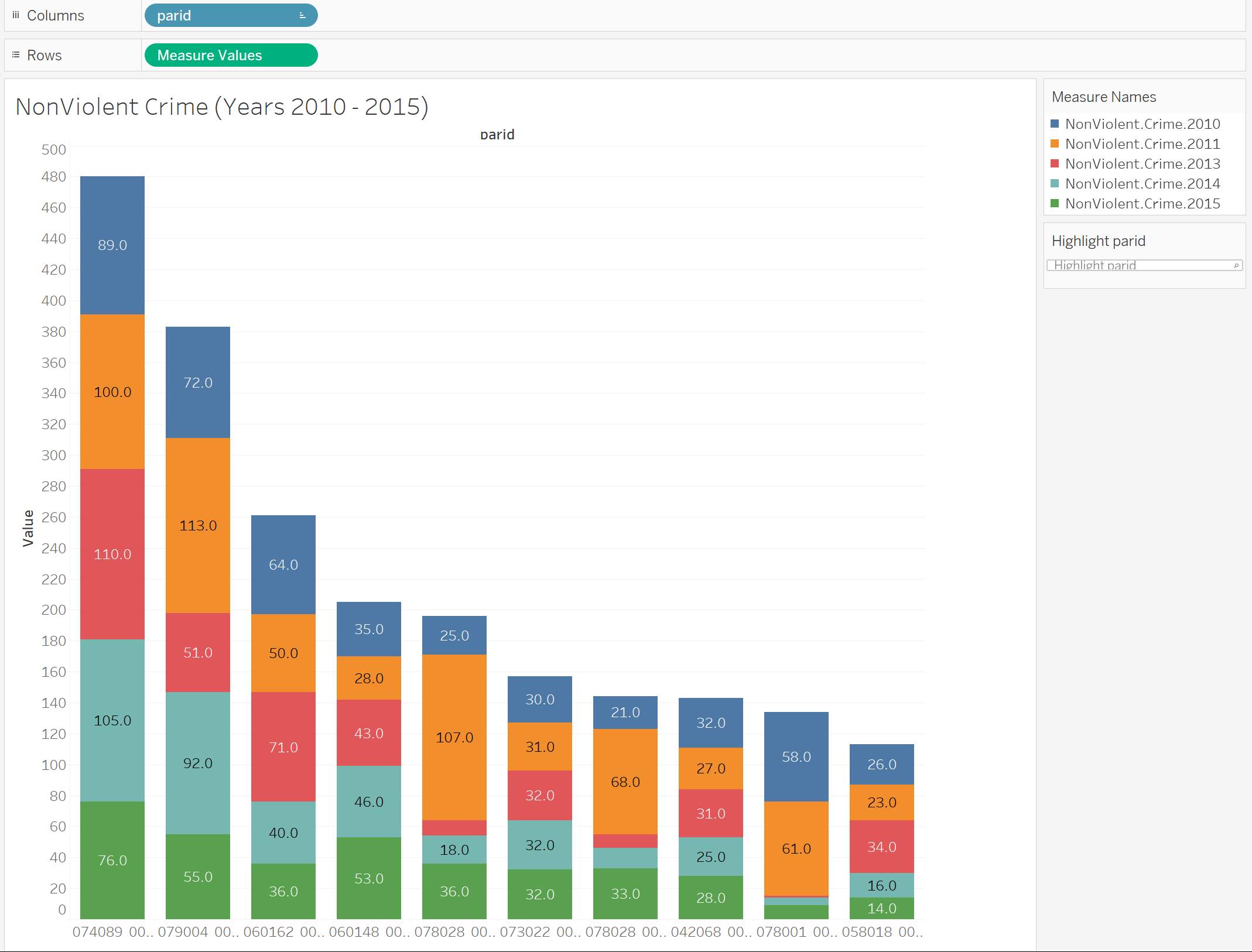
The following chart is showing the top 10 properties rank by violent, nonviolent, other crime showing the property 079004 00019 the most affected by crime with the majority of other crime going on in the area.



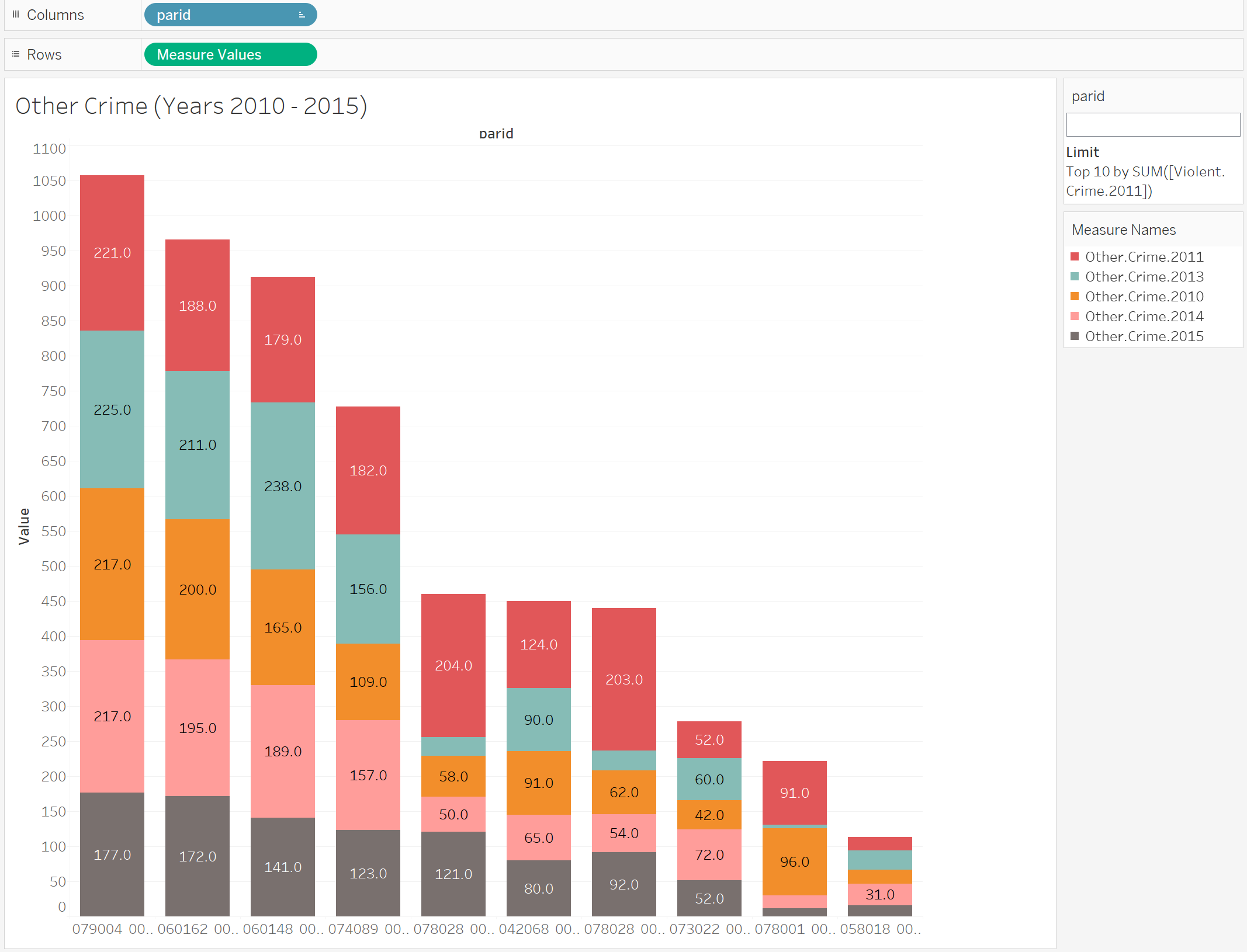
Violent crime results:



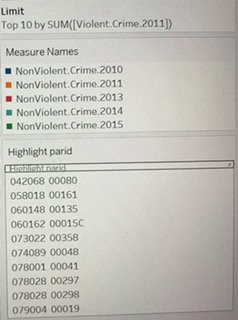
Nonviolent crime results:



Other crime results:



Comparing all the years violent, nonviolent and other crimes, the patter is showing that much of crime is going in the areas starting with codes 06 and 07; therefore, further we would like to pay attention to these parcels in order to investigate the reason of the crime.



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

During the period from 2010 – 2015, total crime decreases. The following chart is showing the crime decrease.

****

Majority of the crime occurs mostly in Zones R-6 and RU-1. Zones R-6, which is residential Single-Family and RU-1, which is Residential Urban District are showing the majority of “activity”. Residential development in the RU-1 District allows a variety of 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 with and are generally located at least 500 to 1,000 feet from a CMU-1, CMU-2, CMU-3, or CBD district or at are least 500 to 1,000 feet from an arterial.

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

****

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

We defined Blight by rating on a scale from 1 to 5, where rating of 5 identifies the most affected by Blight areas. Same zones R-6 and RU-1 are showing the highest possible Blight rating. ****

Since we have determined our target variable as Demolition fees, we were seeking for a visual picture to see what is happening with demolition fees, which areas have the highest amount of original demolition fees, if the fees are getting paid on time and how this information corresponds with the most affected by crime and sales areas.

The following picture is showing the original demolition fees; depicting the same affected by Crime and Sales areas.



Original Demolition Fees are getting paid randomly and the percentage of fees paid is not significant compared to the original amount issued.

**DemolPaid.tiff**

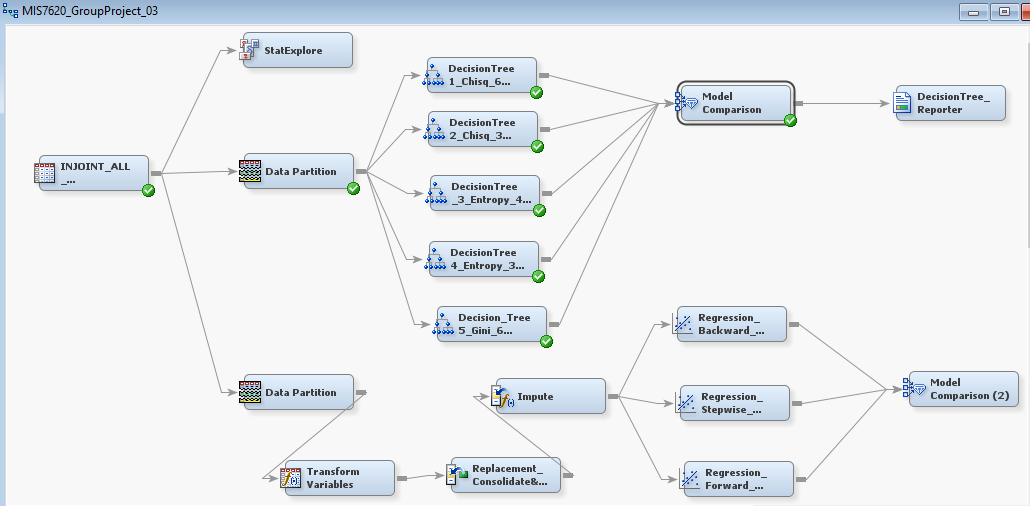
Combined original tax fees are distributed for the same areas R-6 and RU-1. Combined fees, just like demolition fees are not getting paid or, if they are, the payment amount is not that significant compared to the outstanding fees due to low income level.

**Demo.tiff**

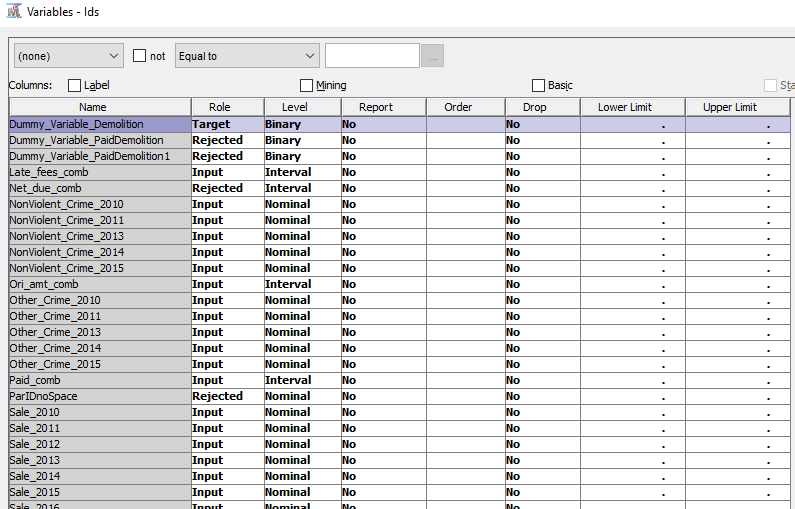
**ComboFees.tiff**

By using descriptive statistics, through simple summaries, we studied our dataset visually in order to determine if there any general trends and patterns exist which need to be further investigated.

**Predictive Analysis Using SAS Enterprise Miner**

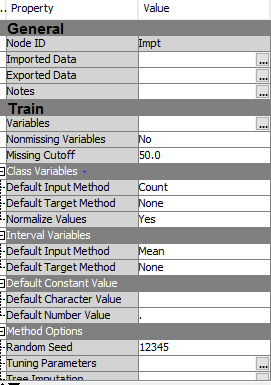


We used different predictive analytics techniques like Decision trees, Regressions, Neural Networks to predict our target variable –: Dummy\_variable\_demolition



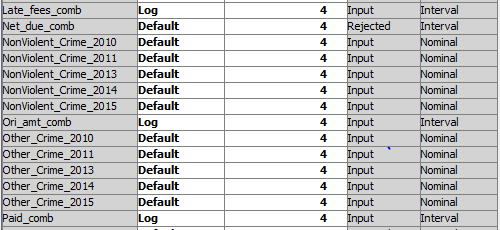
**Nodes used before decision tree, regression or neural network model are:**

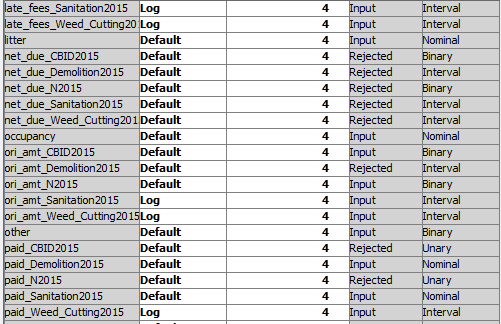
* **Data partition**: Available Data is divided into 50% of Train Data and 50% of Validation Data.
* **Impute**: for handling, missing values. Property settings are as follows:



* **Transform Variables**: This node is included to reduce the effect of extreme values on the Target Prediction. Skew-ness of these variables are reduced by this process.

**Following screenshot showing the transformed variables**

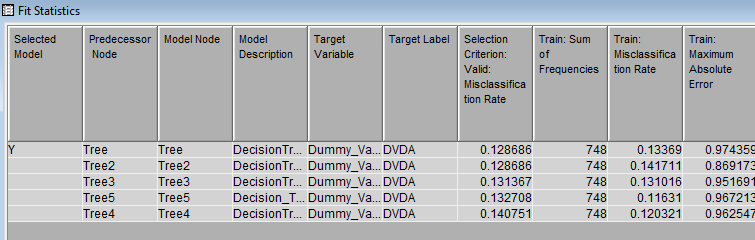




**Decision tree:** we created the decision tree considering different number of maximum branch, maximum depth, nominal target criterion as Entropy or Gini. Properties snap for one of the trees is as shown below

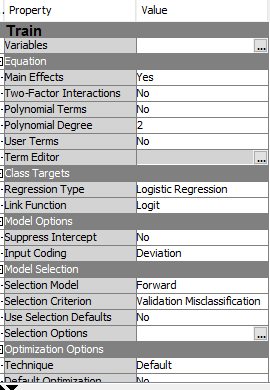


After comparing different decision tree models, decision tree 1 with 6 maximum branches is considered as a best model among all. The minimum validation misclassification for same is 0.128

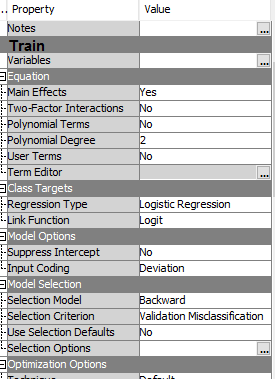


**Regression**: we did the regression model with following methods,

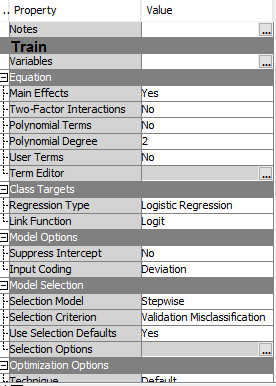
1. Forward Regression



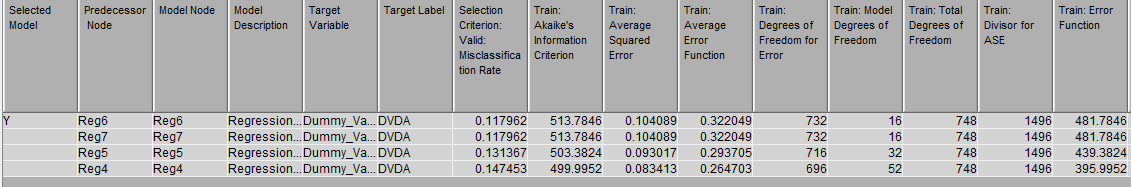
1. Backward Regression



1. Stepwise Regression

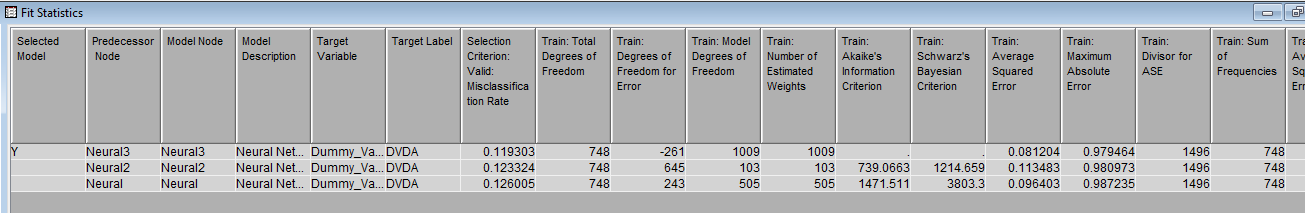


Fit statistics for regression is as shown below. The Reg model is chosen as a best model among all and the validation misclassification rate for it is: 0.117. it is the minimum rate among all models we created.

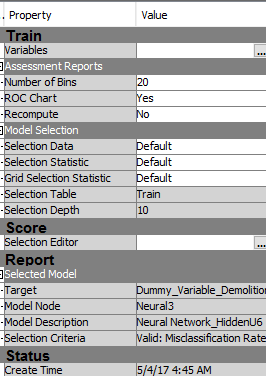


**Neural Network**:

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



The degree of freedom is 748 and the validation rate is 0.119. The nework3 is selected as a best optimized model. Following re the properties for model comparison



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.
* The Identification of the related variables and reduction to 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 a thorough documentation with statistic calculation, model parameters and iteration steps, which are helpful to trouble-shooting model optimization. Therefore, it is worth to learn how to read the file and understand certain amount of those statistics