**Exploring the Relationship between 911 Calls, Property Tax, and Demographic Info in Baltimore**

**Introduction**

For this project we chose to use 3 datasets, found on the Open Baltimore website: Real Property Tax, 911 Calls for Service, and Census Demographics (2010 - 2014). The datasets weighed in at 239k, 6.22mil, and 55, respectively. The plan was to clean each dataset and link them using the common attribute, “Neighborhood”. Although the Census Demographics dataset is comparatively small, we planned to use it as the anchor for our project, generating statistics from the other two and drawing conclusions about how they relate to the demographics in each neighborhood. We were then tasked with cleaning and whittling down the other two very large data sets in a meaningful way. We decided to average the tax information per neighborhood in the Property Tax dataset, sum the calls for various crimes in the 911 Calls dataset per neighborhood, and then join the three afterwards. This meant sorting through unnecessary attributes, labels, messy or non homogenous neighborhood data, and more. We did this in separate Jupyter notebooks using the pandas, seaborn, and matplotlib libraries, expanding on the process in markdown as we went along. We found some interesting correlations, but couldn’t conclude anything groundbreaking without further data and much more research.

**Dataset Details**

Real Property Taxinitially contained 16 attributes and about 239,000 entries. Real property tax is calculated according to the fair market value of the real estate. This dataset looks at various properties in Baltimore with information about property location as well as size and what the state and local taxes are for that specific property.

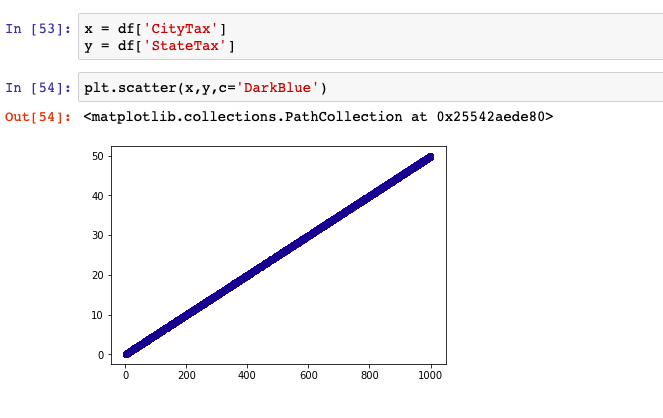
911 Police Calls for Serviceinitially contained 17 attributes and 6.22 million entries. This looks at calls to emergency and non-emergency calls to 911 and categorizes them based on priority. This ranges from “non-emergency” which includes business checks to “high” which includes calls for aggravated assault or common assault. Some of the attributes give information about location, call data and time, and police district.

Census Demographics 2010initially contained 27 attributes and 55 entries. It essentially grouped Neighborhoods and categorized the populations there by race, age, gender and other socioeconomic classifiers like household size and household income. This dataset was particularly interesting because it allowed us to look at how socio economics can affect property taxes and the types of 911 calls.

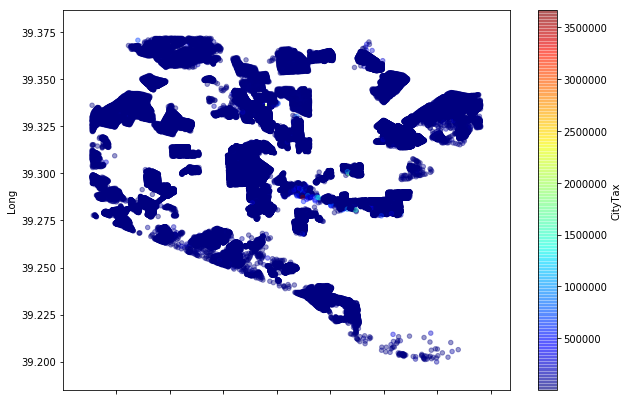
# 

# **Discussion and Description: Working Through the Data**

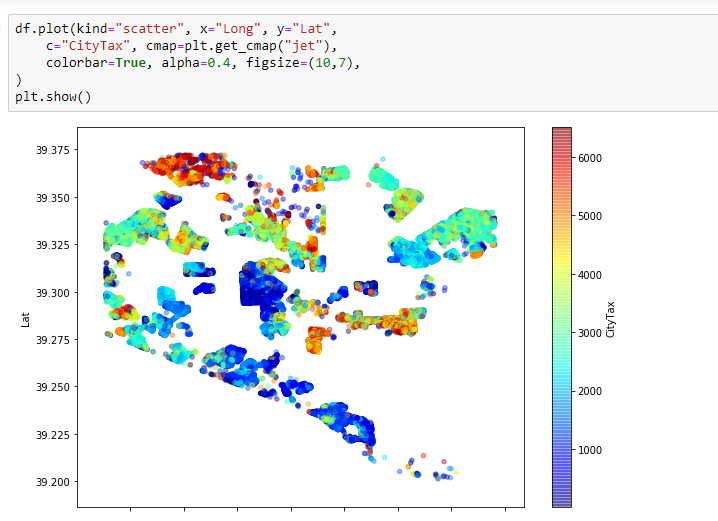
We began with the Real Property Tax data set. While trimming down the data we found that the attributes “CityTax” and “StateTax” were positively correlated. This tells us that they are probably calculated the same but “CityTax” is always larger than “StateTax”.



Since the Real Property Tax dataset had longitude and latitude information, we created a heatmap of CityTax. Our first attempt at a heatmap for CityTax produced almost entirely low heat because of extremely high outliers.

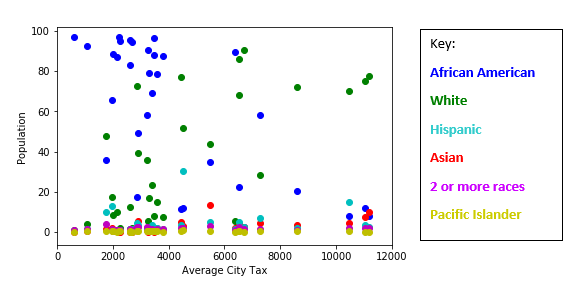


Based on this heat map, almost all of Baltimore falls under $500,000 city tax. There are huge properties skewing the data with very high city tax. So, we filtered out those properties with huge tax. We researched what the average household property tax was in Baltimore and doubled to allow for more expensive properties. This gave us a heat map with more variation.

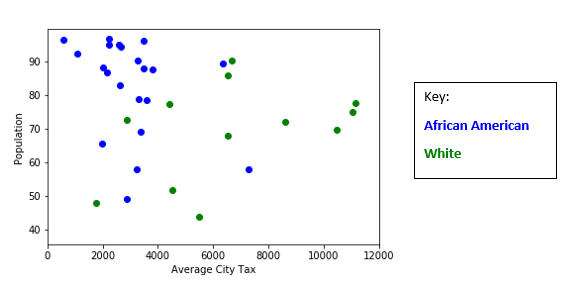


It was interesting that these heat maps did not display “Long” in the x-axis even though we included that information in the code. We had this issue with all of our heat maps that use longitude and latitude.

For Census Demographic 2010, we used immediately started looking for relationships between race and average city tax, and combined the two datasets by merging on Neighborhood and averaging the tax values.



First we looked at all races in Baltimore, the population and average city tax. We quickly discovered that Baltimore is predominately white and black, so we decided to just look at the percentage of White Americans and the percentage of African Americans at specific longitudes and latitudes since we had that information. We developed heat maps and found and inverse relationship in many parts of Baltimore between the population of African Americans and the population of white Americans.

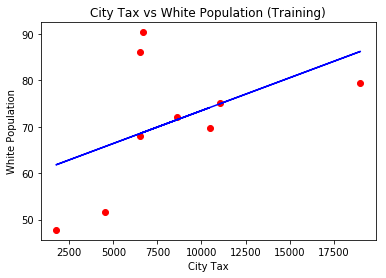
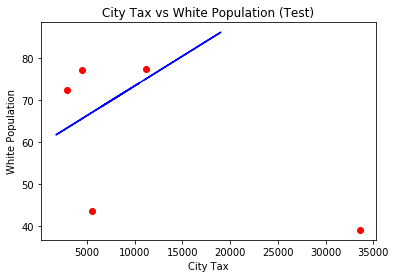


If you recall the image in the City Tax section that displayed the city taxes at various latitudes and longitudes and refer to the images below, you can see that predominantly African American areas tend to have lower city taxes (first image) while predominantly white areas generally have higher city taxes.

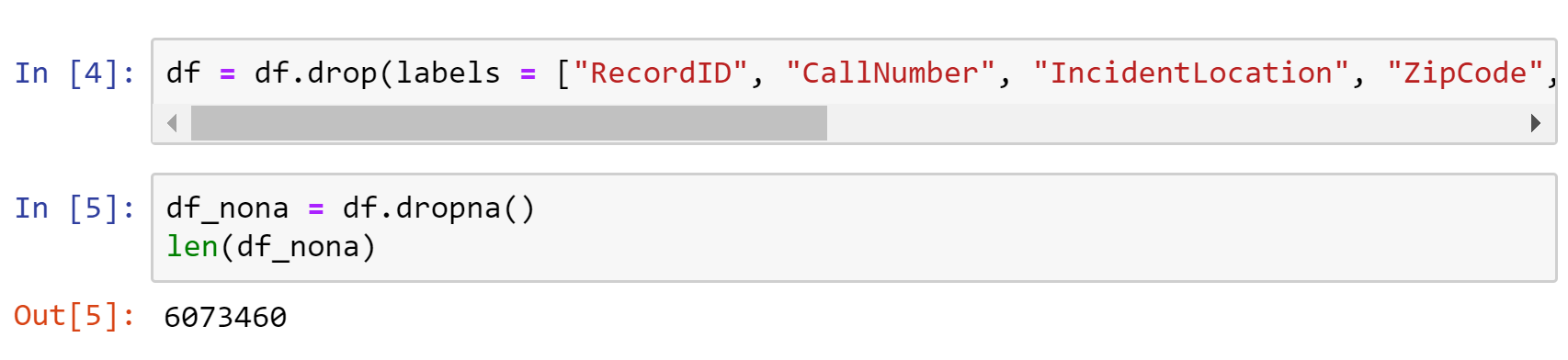




This led us to create a model using linear regression on the white population.

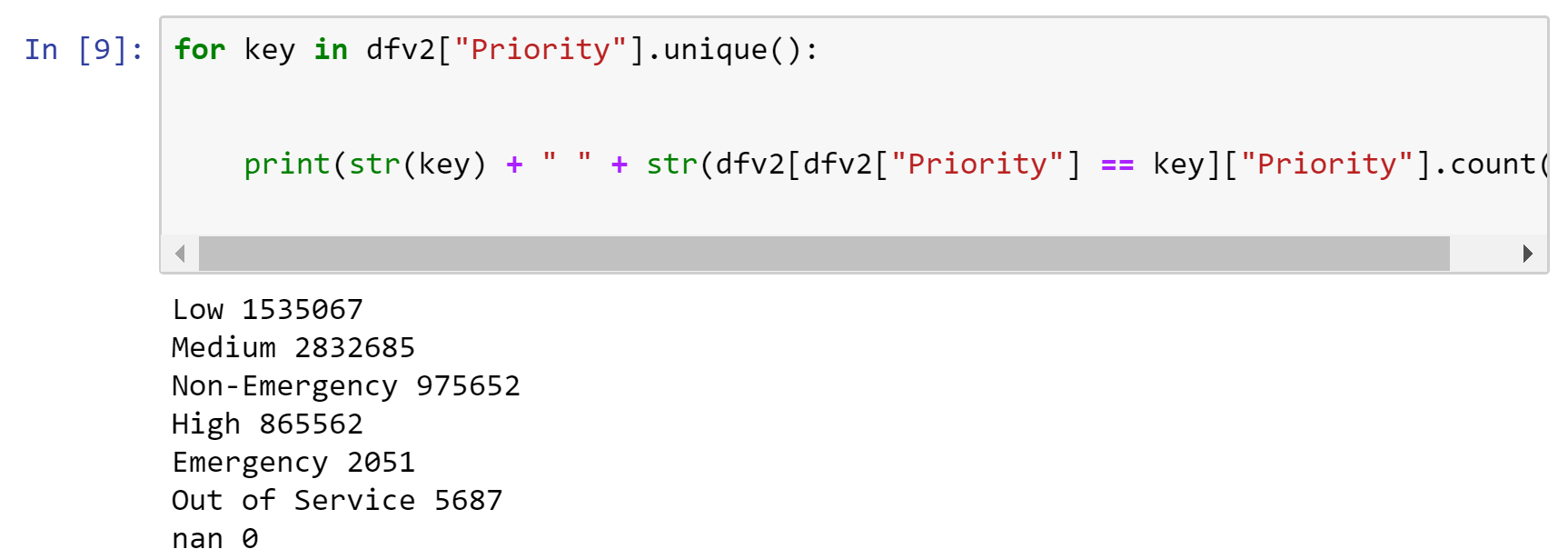
This shows an overall positive correlation between the white population and city tax. The model we used was far from perfect considering we had a correlation coefficient of [0.0014126], mean squared error of 1068.87 and a variance score of -2.73

We then moved forward to the 911 calls for service dataset. With 6 million data points and over 10 columns, we had to experiment with different ways of cutting the data down, without blindly chopping away at the information. Towards this, the first thing we did was eliminate the redundant and less useful attributes: RecordID, CallNumber, PolicePost, Census\_Tracts, and a few others. These attributes didn’t bring much value to what we were trying to do, so they were the first to go. The most surprising thing is that when I dropped the NaN values, before removing these columns, the dataset dropped from 6 million to 830k, but after I removed these unnecessary columns, dropping NaN values barely affected the dataset. It seems that most of the incorrect or incomplete data was entered in these fields, supporting the idea that they weren’t necessary or important for our analysis:



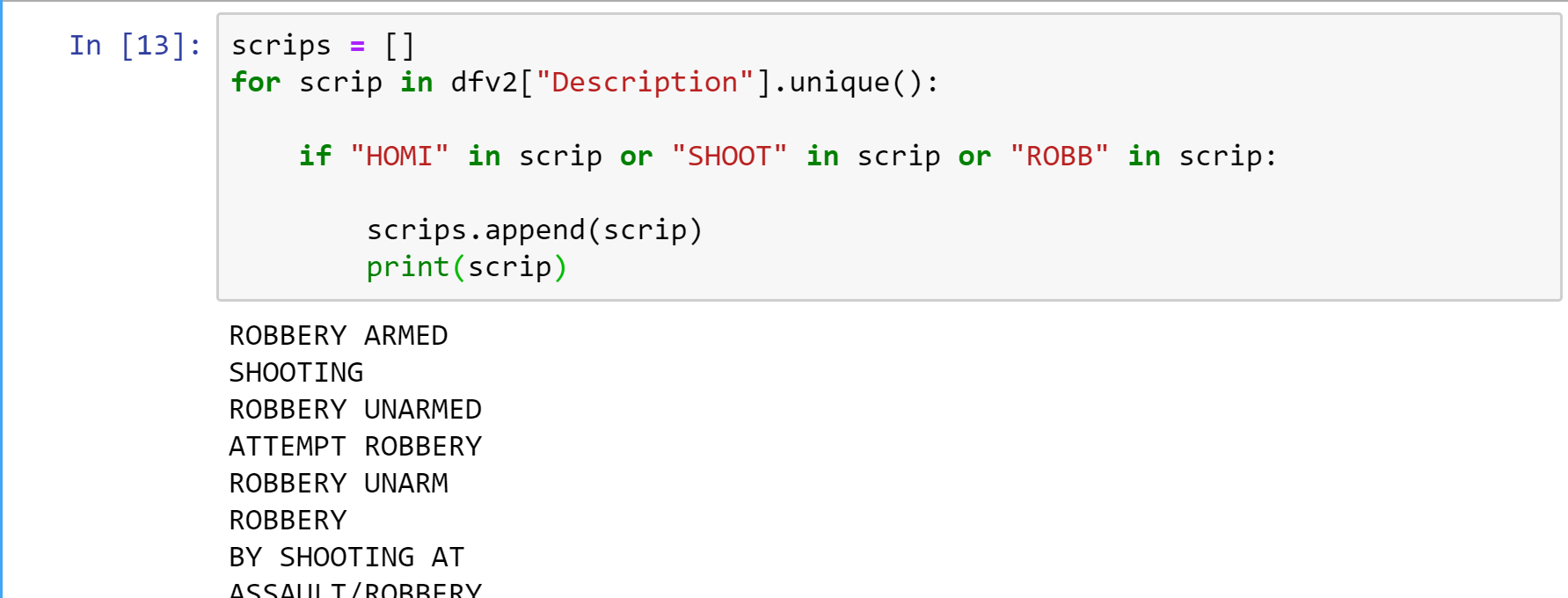
We concluded that dropping the NaN values later wouldn’t affect the dataset, and did so after exploring the columns and other ways to cut the dataset down.

As we took a deeper look into the columns, we decided on four attributes, District, Description, Neighborhood, and Priority, as the columns we would focus on (while still holding on to CallDateTime). Intuitively, visualizing the distributions with graphs was the first step, but we soon realized that with 6 million data points, that would take much too long. Instead we examined each column individually and explored the unique values and their frequency distribution. For Priority, we found that that Emergency, Out of Service, and NaN Priority calls accounted for less than 10k of the calls, and that Non-Emergency accounted for 975k. We dropped the first three for their inconsequence, and the last one because it was irrelevant and we were focused on crimes (which would be higher than non-emergency).

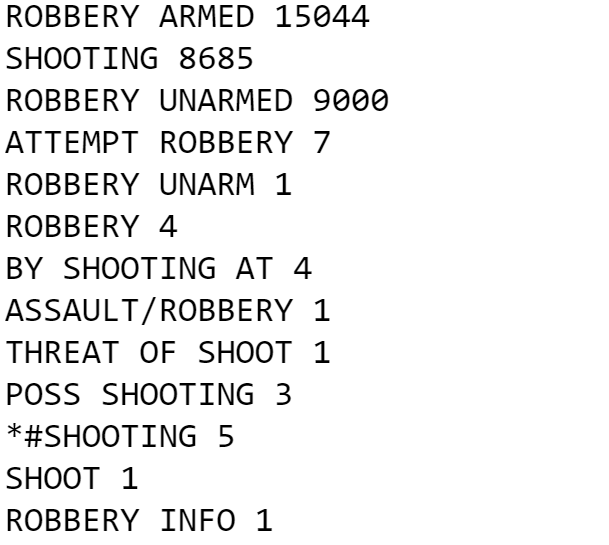


We continued similarly for Districts, but ultimately decided that there wasn’t anything meaningful we wanted to do with that column. When we got to Descriptions, we found that there were 14k unique descriptions in the dataset, and we had the intuition that we’d be able significantly cull the numbers by exploring this attribute.

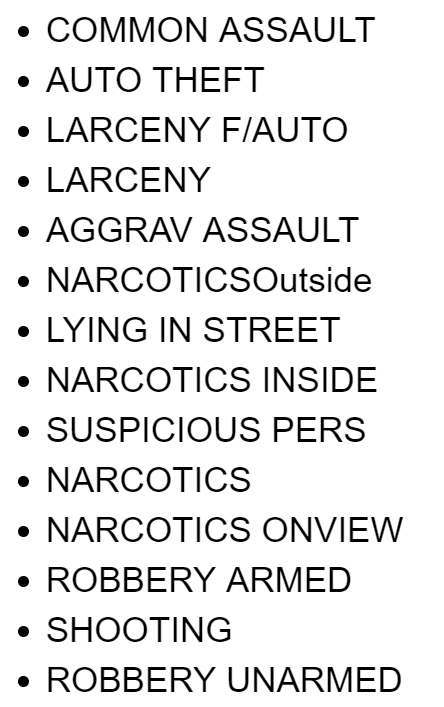
The descriptions field was a collection of strings that different people must’ve written or entered to say what kind of phone call it was. Targeting entries that had crime-related substrings, we searched through in groups to find what descriptions had the highest frequency and could be useful. For example, for starters, we began with shooting homicide, and robbery:



We got the count for each identified description, and chose the ones with high frequency:



We proceeded similarly until we settled on the list below:



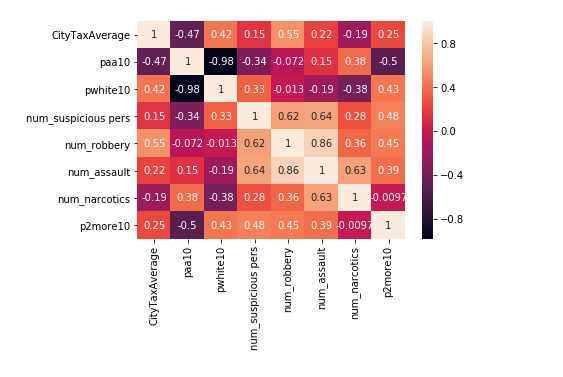
Clearly, there are several descriptions that refer to the same crime, so we ran some code to group those (robbery armed -> robbery, narcotics inside -> narcotics) , and then trimmed the dataset by selecting only calls with the chosen descriptions. This left us with a dataset of only 689k rows, but they all had meaningful, hand picked data. From this we learned that a relatively small number of the 911 calls, about 10 percent, are crime-related calls.

After that, we decided it was time to group the datasets. We loaded in the merged Census and Tax datasets in order to user the Neighborhood groupings to do the last prep for this dataset. Upon relabeling the neighborhoods in the 911 Calls dataset, we discovered that there were many neighborhoods in this data that weren’t included in the census data. We settled on the groups in the census, but further research can be done in the future to understand the census groupings. After relabeling the neighborhoods to fit in the correct census groups, we summed the calls for each crime per neighborhood and merged the datasets all together.

# **Results**

Once the datasets were connected, we focused on specific attributes of interest from each dataset. From Real Property Tax, we just looked at the City Tax Average. From Census Demographics, we found that areas tend to either be predominantly white or predominately African American. So we included the percentage of African Americans and the percentage of White Americans. From 911 calls, we decided to look specifically at the number of suspicious persons, number robberies, number of assaults and number of narcotics. We felt that these calls about these crimes/issues were less frequently considered in favor of shootings and homicides. We wanted to see if there were any correlations between narcotics, a type of crime that we thought might be correlated with the population of whites given the recent rise in opioid abuse cases, and the number of whites. We sought similar correlations with the other crimes and whites, aiming to break the stereotype that black neighborhoods have generally more crime in all areas. The results were disappointing, but we did identify some interesting relationships.

Although the correlations involved weren’t always necessarily very strong, we found the complete inverse relationship in most cases when comparing crime related calls and the percent of blacks vs whites to be insightful nonetheless. We also hoped to show that the city tax average would be correlated to the crime call most highly correlated to the percentage of blacks. This was also not the case. You can see our results in the correlation matrix, mapped onto a heatmap, below:



We observed that:

* The percentage of whites is almost perfectly, negatively correlated with the percentage of blacks, at -.98, supporting what we saw in the heat maps using location.
* The number of suspicious person calls is negatively correlated with the percentage of African Americans, and opposite for the percentage of whites.
* The percentage households with two or more races is negatively correlated with the percentage of African Americans, and opposite for whites.
* The number of robberies is positively correlated with the CityTaxAverage, with .55 correlation. Implying that more expensive properties are more likely to report robberies. And interestingly, there is almost no correlation between the percentage of African Americans and Whites and the number of 911-reported robberies. However, there is another positive/negative correlation between tax average and blacks and whites, respectively. Thus, the crime call with the least correlation to any race had the highest correlation with Tax. The implication may be that the increase in tax is somewhat independent of crimes and more related to race.

# 

# **Insight**

From the results of analysis, we now have specific information about the kind of city Baltimore is. Many of this information supports things people already knew. Based on the negative correlation of the percentage of blacks vs. the percentage of white in a given neighborhood, we can see that Baltimore is very segregated. This gives insight into the effects of gentrification and white flight. Based on the heat maps we produced, we saw that areas that are predominately white tend to have higher City Tax Averages.

# 

# **Future Work**

There were more neighborhoods in the calls dataset than there were in the census dataset, so we left a lot more data out. In the future, we would do some research to see why that is, and try to match some of those ungrouped neighborhoods and create new groupings. In addition, regarding the results, further work may explore causation behind these observations, and experiment with other crime calls and their correlations. If we could find a dataset that explains the relationship between certain committed crimes and race, we may be able to link these and further explore the meaning behind the tax-crime call relationships observed.

**Resources**

1. Real Property Tax Dataset: <https://data.baltimorecity.gov/Financial/Real-Property-Taxes/27w9-urtv>
2. 911 Calls for Service Dataset: [https://data.baltimorecity.gov/Public-Safety/911-Police-Calls-for-ervice/xviu-ezkt](https://data.baltimorecity.gov/Public-Safety/911-Police-Calls-for-Service/xviu-ezkt))
3. Census Demographics (2010 - 2014): <https://data.baltimorecity.gov/Neighborhoods/Census-Demographics-2010/cix3-h4cy>
4. Average Property Tax in Baltimore:

<https://lakesidetitle.com/property-taxes/>