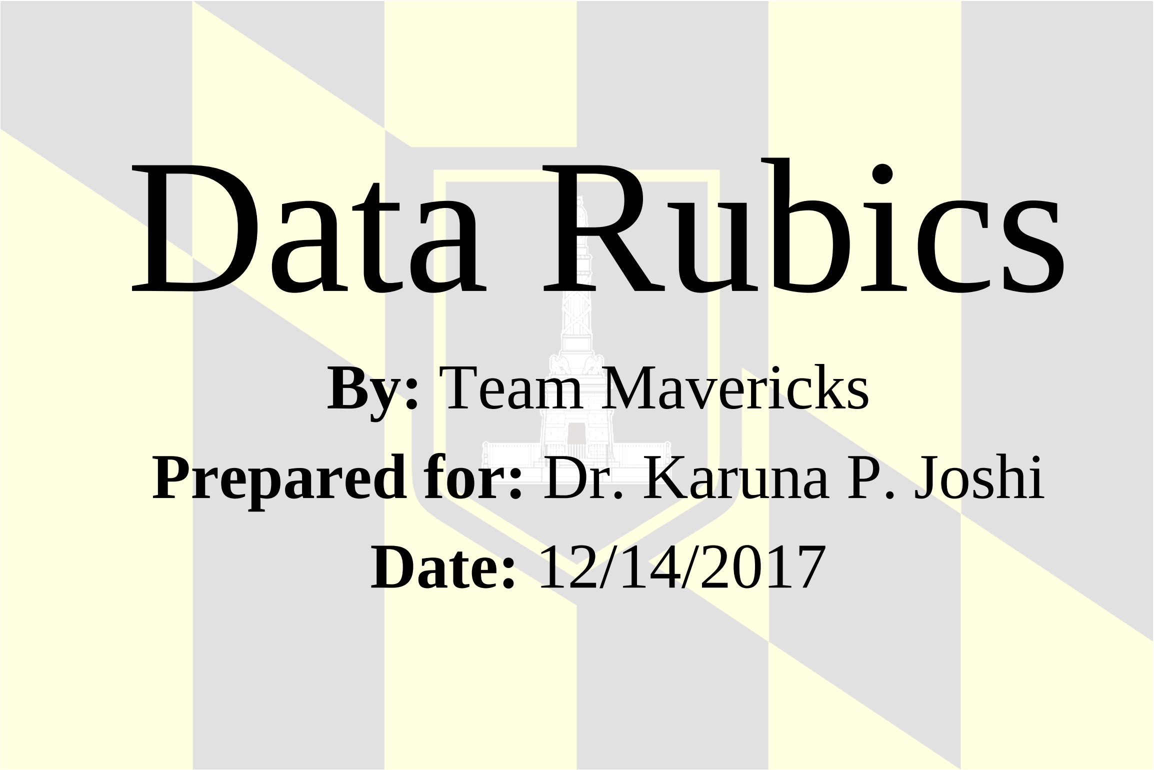


Data Rubics

By: Team Mavericks

Prepared for: Dr. Karuna P. Joshi

Date: 12/14/2017

Data Rubics – Correlation between Vacant Buildings and Crimes in Baltimore City

Kris Singh
Information Systems Department of University of Maryland, Baltimore County (UMBC)
Baltimore, U.S.A
Kris15@umbc.edu

Harika Parakala
Information Systems Department of University of Maryland, Baltimore County (UMBC)
Baltimore, U.S.A
Pharika1@umbc.edu

Ashish
Information Systems Department of University of Maryland, Baltimore County (UMBC)
Baltimore, U.S.A

Divya
Information Systems Department of University of Maryland, Baltimore County (UMBC)
Baltimore, U.S.A

***Abstract*—This paper features our product Data Rubics, in which we analyze and visualize the correlation between crime and vacant buildings in Baltimore County.**

Keywords—crime, vacant buildings, Baltimore City, neighborhoods, locations, districts,

I. Introduction

There is a preconceived notion that Baltimore city has a very high crime rate, and people also assume that vacant and abandoned buildings attract more crime. Baltimore city has a total population of about 626,848 and has a violent crime rate of “1,417 per 100,000 residents” according to a recent study by Forbes Magazine. The same study also ranked Baltimore city as the “seventh most dangerous U.S cities” [1].

According to the data provided by the Baltimore City Open Data Source, there are about 16577 officially confirmed vacant buildings in the city of Baltimore [3]. A building is considered vacant in the city of Baltimore if the building has been unoccupied, marked as unsafe or unfit for people to live or work inside, has two code violations that have not been fixed or has six code violations in the past year.

II. Hypothesis

Our hypothesis for our research is that: Crime rate is higher in areas with vacant buildings in Baltimore City.

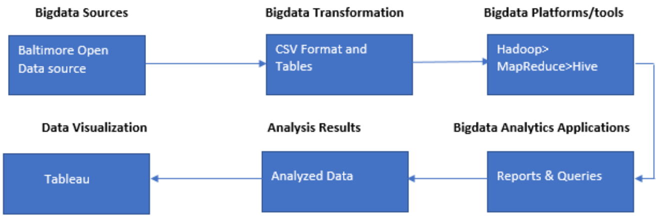
A. Motivation

There is a preconceived notion that Baltimore city has a high crime rate and people assume that vacant and abandoned buildings attract more crimes. To test this notion, we would like to analyze whether vacant buildings and crimes do actually have any correlation. From this research, we want to help the city of Baltimore reduce its crime rate by suggesting ways to protect its citizens in a more cost-efficient way so that they do not waste their resources while trying to fix something that might not even be an issue.

III. Related Work

There have been several studies done in the area of vacant buildings and its relation to crime around it. One of the paper that was analyzed was, “An Intelligent Analysis of a City Crime Data Using Data Mining.” [2] In this research, the dataset was made available by the department of Police and the range of years available and utilized was between the year 2000 and 2009. The missing data values were replaced with the medians of the attributes (variables) computed from pertinent clusters. The end result data was visualized using Weka after preprocessing, filtering, and clustering techniques. The results of the analysis were portrayed in a graphical way by plotting the year on the X axis and the number of crimes that occurred in that year on the Y axis.

IV. System Architecture



A. Data Cleaning

For our research, we used two data sets from the Baltimore City Open Data Source. The two data sets that were used were: The Vacant Building dataset and the Crime dataset.

We took several steps to clean our two data sets. Some minor initial changes included standardizing, capitalization to make our data more compatible with various types of visualization tools and that of Hive. Much more cleaning work went into the largest dataset, Crime in Baltimore, MD from 2013 to present. This data had many missing values under various attributes. To handle missing values, we decided to remove any instance that contained a missing value, rather than worry about treating missing values specially. This was not a great loss, as the Crime data is large, with 80,796 instances after cleaning. The date format was changed to meet Hive standards, also removed special characters which were not standard for Hive.

B. *Data Transformation*

The Cleaned data is then converted to CSV file and then loaded in Hive by creating tables and inserting data into it.

C. *Bigdata Platforms and Tools*

The Cleaned data is then converted to CSV file and then loaded in Hive by creating tables and inserting data into it.

D. *Map Function*

We are going to select the rows of Neighborhood/district/location (Latitude Longitude) from a particular Date. The system automatically applies the above criteria to each data node. The result will be the count of that data node.

E. *Reduce Function*

We aggregated the count of crime and Vacant buildings obtained from the Map Function.

F. *Bigdata Analytics Application*

The analysis is done by querying the data of counts on each of the neighborhood/district/location or on the date or year and by looking at the results generated after querying for the counts of crime against the vacant buildings and comparing them to find the pattern of the data flow.

G. *Data Visualization*

Tableau is used to visualize the data fetched after passing through the MapReduce function. The hot spots are shown on the Map where the count of the crime and vacant buildings are shown with two different colors.

V. Design Criteria Approach

In our design, we looked at the Baltimore City's Crime data starting from 2013 to 2017 and data from Baltimore City's vacant building up to 2017. We loaded the two data sets to the database using Hive to do our comparisons.

We also looked at the data from three different perspectives. First, we looked at the data between the two tables at the District level which gave us a broad view of crime and vacant buildings in a district in a given year.

Second, we drilled down into the district and look at the Neighborhood level to compare crimes and vacant buildings in a neighborhood in a given year.

Finally, we will drill even deeper into a specific location using their longitude and latitude information of different vacant buildings to compare crime and vacant buildings. Drilling down to the location level will help us find a direct correlation between vacant buildings and crimes in the neighborhood. This will help us understand more specific solutions for reducing crime in the area.

We generated a count of crimes, and vacant buildings in particular area (district, neighborhood, and location) in a given year. We will also find the ratio of crime vs vacant buildings in that area to help us suggest different suggestions to what should be done of the vacant building.

Based on our results we can suggest the city different ideas on possibly help reduce crime in that area. Some of our suggestion include:

1. Fencing the vacant area
2. Maintaining the vacant building
3. Rent/utilize the vacant building
4. Demolish the vacant building
5. Securitize the buildings

A. *Design Focus*

There is a preconceived notion that Baltimore has high crime rate and we assume that vacant and abandoned buildings attract more crime. We had also assumed that the crime rate would increase in an area where there are vacant buildings. In this research, we are also treating all of the crime as equal.

For our analysis, we have marked the following fields as out of scope from our Crime data sets:

- Age
- Sex
- Race
- Arrest location
- Incident offense
- Charge
- Charge description
- post

We have marked the following fields as out of scope from our Vacant Buildings data sets:

- Block
- Lot
- Council District ID

We have excluded the Age, Sex, Race, Arrest Location, Incident Offense, Charge, Charge Description, Post, Block, Lot, and Council District ID because they are irrelevant to our research. Since we will be using the longitude and latitude for location, the block, lot number and council district ID became irrelevant. The Arrest location is irrelevant because we assumed it did not have any relevancy with where the crime had occurred. Since we are not distinguishing the individual crimes we decided not to use Incident offense, charge, charge description, age, sex, and race data.

VI. Analysis of Crime vs. Vacant Building Data

A. Map Function

Our Map Function maps the Crime data with the Vacant Building data with respect to the District/Neighborhood/Location (Latitude, Longitude) data which binds both datasets together which helped us write the reduce function. We selected the rows of District/Neighborhood/Location from a date. The system automatically applies the above criteria to each data node. The result will be the count of that data node. The Map operator tree for the below query includes tables scan, filter operators, select operators, and map join operator.

Example:

Map - 1 Sample Output:

Baltimore	23	4
Arundel	14	8

Map - 2 Sample Output:

Baltimore	67	9
Arundel	78	2

B. Reduce Function

The reduce function aggregates the count of Crime and Vacant buildings obtained from the Map function. The Reduce Operator Tree includes Group by operator, and Aggregations (count).

Reducer Sample Output:

Baltimore	90	13
Arundel	82	10

District Level:

```
“Select a.policedistrict, count(distinct b.arrestID), b.year
from buildings a, crime b
where b.district = a.policedistrict
group by a.policedistrict, b.year
order by b.year, a.policedistrict”
```

Neighborhood Level:

```
“Select a.neighborhood, count(DISTINCT b.arrestID),
count(DISTINCT a.referenceid)
from buildings a, crime b
where b.neighborhood = a.neighborhood and b.arrestdate
< '2017-01-01' and b.arrestdate > '2016-01-01' and
a.noticedate < '2017-01-01' group by a.neighborhood;”
```

Location Level:

```
“Select a.neighborhood, b.location, round(b.latitude,
4),round(b.longitude,4),
count(distinctb.arrestID),count(DISTINCT
A.REFERENCEID), 2017
from buildings a, crime b
where A.NOTICEDATE <= '2017-12-31' --update the
year and a.neighborhood = b.neighborhood
and a.longitude between b.longitude-0.0009 and
b.longitude+0.0009 and a.latitude between b.latitude-
0.0009 and b.latitude+0.0009 group by a.neighborhood,
b.location, b.latitude, b.longitude
order by a.neighborhood, b.location, b.year;”
```

C. Analytics Tools Used

Excel is used to convert reports generated in the form of csv and used to analyze the results. We have used both Excel and R to calculate Pearson Correlation. Tableau is also used to graphically analyze the data by adding various filters.

D. Pearson Correlation

The Pearson product-moment correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases [4].

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}}$$

VII. Analysis Results

Our analysis initially started with looking at the raw data collected from open Baltimore data source. We analyzed the columns present for each data set and period of the data. Found that we have 5years (2013-2017) crime data and (1996-2017) vacant building data. Based on our design we planned to compare the count of crime and vacant buildings for the various years and Neighborhood/District/Location. As per our hypothesis the count of crime should be increasing as the vacancy increases. Here as the important variables are Crime_Date, Crime_Count, Neighborhood/District/Location, Vacant_Count, Notice_Date, ArrestId, Refid_Vacant the other columns which are of low importance are ignored for now. The vacant and crime count is calculated using the MapReduce function and a CSV of the resulting data is generated and is placed on HDFS.

The CSV file is opened in an Excel sheet and chunk of data is taken for analyses graphically. The count of crime Vs. Count of Vacant buildings is plotted on a histogram and tried to analyze the pattern of the flow of count of crime and vacant. Most of the data showed that initially there is lot of crime count in the area but gradually it started decreasing. It did not show any kind of relationship with the vacant building count.

A. Extra Data Cleaning

Went back into the raw crime data and realized that there are various kinds of crime which may not have any relationship with vacant building. Filtered out all the crimes which would be appropriate in our analyses like burglary, drugs, rape, robbery, Murder, Unnatural Death, auto thefts etc.

Thus, repeated the steps to find the count with newly cleansed data. The resulting data consisted of some of the negligible values like 0 or 1 in both vacant and crime field i.e the data was sparsely distributed. By ignoring such negligible rows, we found 61 neighborhoods with considerable special distribution of data. This filtration is done at Location level also. The PEARSON Correlation of the resulting data is calculated and the following results are obtained for different datasets.

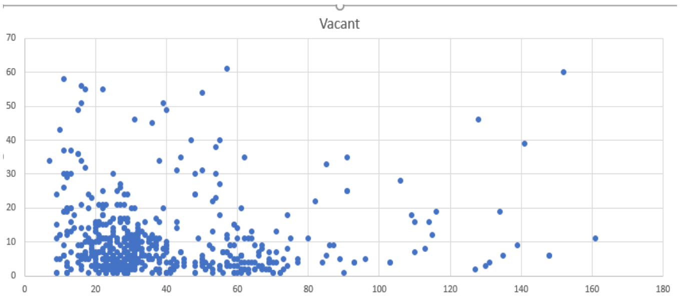
Neighborhood Level for Half Yearly:

r Coefficient - Pearson Correlation

The average r value for 61 Neighborhoods	**0.23767**
Number of Neighborhoods whose r values is greater than 0.7	**7**
Number of neighborhoods whose r values is greater than 0.5	**16**
Number of neighborhoods whose r values is greater than 0.23	**28**

Percentage of Neighborhoods whose r values is greater than 0.7	**11.47%**
Percentage of neighborhoods whose r values is greater than 0.5	**26.22%**
Percentage of neighborhoods whose r values is greater than 0.23	**45.90%**

Scatterplot:



Neighborhood Level for Yearly:

The average r for 96 Neighborhoods: **0.061417,** this shows that there is some correlation at Neighborhood Level for Yearly data.

District Level:

The average r for 9 Neighborhoods is **-0.94018,** this shows they are not correlated at District Level.

Location Level Yearly:

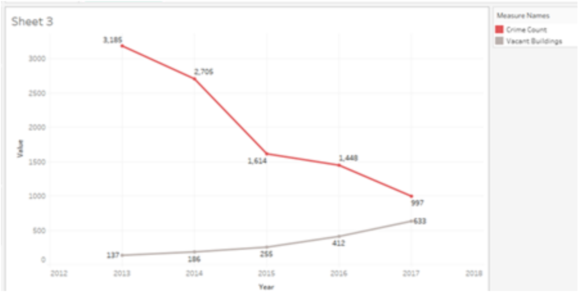
The average r value for 8196 locations	**0.248446**
Number of Locations whose r values is greater than 0.24	**2757**
Percentage of Locations whose r values is greater than 0.24	**33.63%**

As per the above analysis Considering Neighborhood Level Half Yearly the data supports our hypothesis with a confidence of 45% as the percentage of neighborhoods falling in the range of average r value is 45%. This the data gives support to our analysis when compared to other level Data. Even at location level the data supports our hypothesis with a confidence level 33.6%

VIII. DATA VISUALIZATION

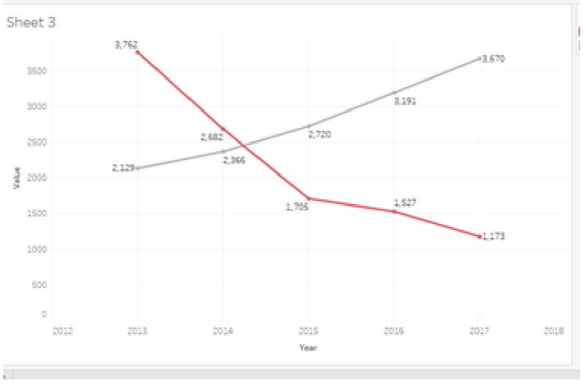
We used Tableau to visualize the crime and vacant buildings data from four perspectives which are district, neighborhood, and location (latitude-longitude). The following visualizations hence prove our analysis results. In the graphs mentioned below, Crime count is shown in red color and Vacant building count is shown in grey color.

*A. Perspective 1: **District Level***



District: **Southeastern**

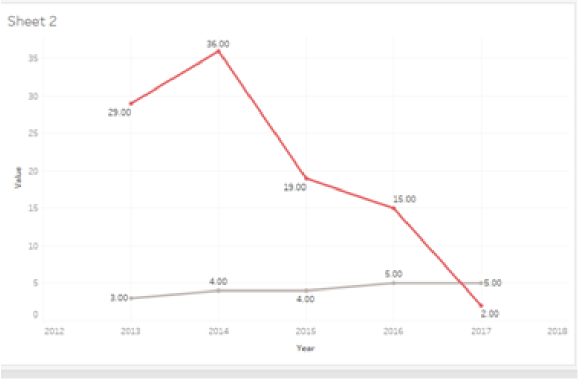
The graph above for the Southeastern District shows that the count of the vacant buildings and crimes at the District level. The pearson coefficient for this district is -0.94018, which shows that there is an no correlation between crime and vacant buildings.



District: **Eastern**

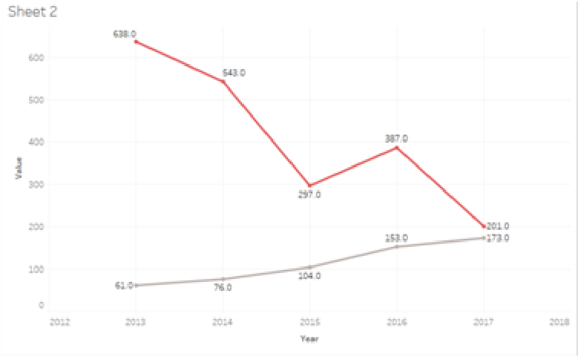
The graph above for the Eastern District shows that as the crime count decreased, the number of vacant buildings also increased which can be seen from negative pearson coefficient. Tablue can be used to change the filter foe each individual district to show our respective visualization for crime and vacant buildings.

*B. Perspective 2: **Neighborhood Level (Yearly)***



Neighborhood: **Franklin Town**

The graph above for the Franklin Town neighborhood shows that the count of vacant buildings and crimes in Baltimore city. As per our analysis, the average pearson coefficient for the yearly data for this particular neighborhood is 0.061417. This shows that there is almost no correlation between the crime count and the vacant buildings.



Neighborhood: **Brooklyn**

From the graph above, we can notice that the crime did increase in some cases from the previous year to next year, but that increase is comparatively less than the crime count that is decreasing from the previous years and we can also notice that the vacant buildings count is increasing. This proves that the result of the analysis which relates to a low pearson coefficient.

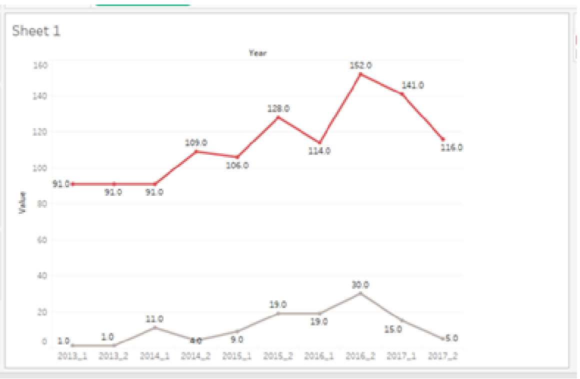
Tableau can be used used to change the filters for every neighborhood to show our respective visualizations for crime and vacan building fro that particular neighborhood.

*C. Perspective 3: **Neighborhood Level (Semi-Yearly Data)***

The average pearson coefficiency for semi yearly data of the neighborhoods is 0.2376. We can say that there is some correlation between the Crime count and Vacant Building counts for some of the neighborhoods. However, there were some negative pearson coefficeint for some of the neighborhoods.

1) Neighborhood Proving our Hypothesis

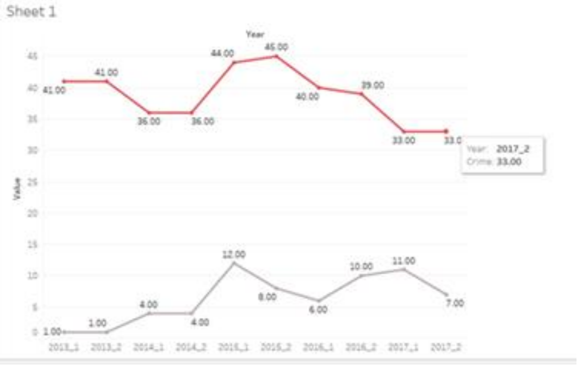
Brookly is one of the neighborhood that had proved our hypothesis. The pearson coefficient for this neighborhood is 0.8047. There is an increasing trend in Crive vs Vacant Buildings from 2016 first quarter to 2016 second quarter. There is a deacreasing trend in Crime vs Vacant Buildings from 2017 first quarter to 2017 second quarter. In the graph above we can see the flat trend in Crime vs Vacant Buildings from 2013 first quarter to 2013 second quarter. The screenshot below proves this statement.



Neighborhood: **Brooklyn**

2) Neighborhood Disproving our Hypothesis

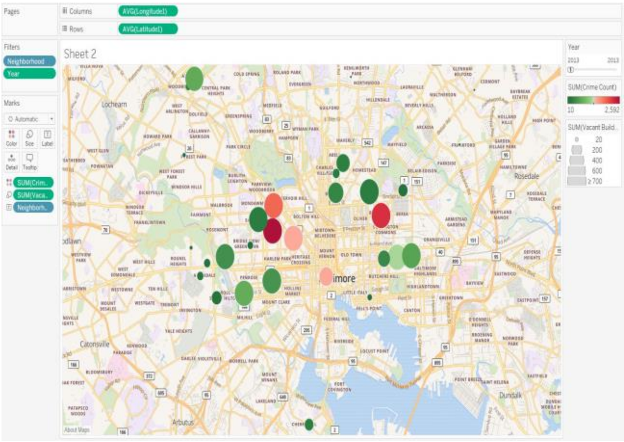
Barclay is one of the neighborhoods that disproves our hypothesis. The pearson coefficient for this neighborhood is -0.0080. Example quarters for disproof from Visualizations are from 2013 2nd quarter to 2014 1st quarter, 2015 1st quarter to 2015 2nd quarter and 2016 2nd quarter to 2017 1st quarter. The below screenshot proves the result.



Neighborhood: **Barclay**

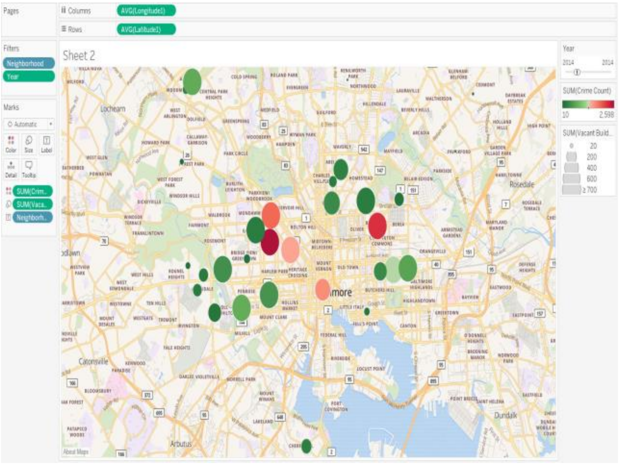
D. *Perspective 4: **Location Level***

The crime count is taken as a measure (sum) of the crime for each year from 2013 to 2017 at selected neighbourhoods, high and low crime count is represented with red and green colours respectively. The size of the vacant building is taken as a measure (sum) of the vacant buildings of each year from 2013 to 2017 in selected neighbourhoods and is represented with a bubble. The size of the bubble gives good view about the number of vacant buildings at that locations. With the analysis results obtained at the location level, we are visualizing the results at a higher level showing the neighborhoods with the sum of crime buildings and sum of the crime in that neighborhood are shown.



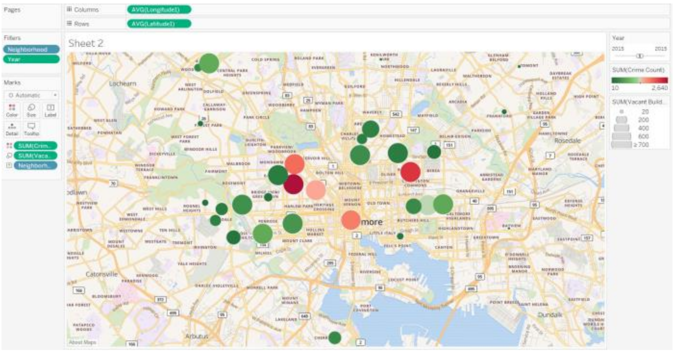
Year: **2013**

Here we are showing the trend analysis for each year from 2013 to 2017 of the results we achieved. In the year 2013, we found that there is a correlation between the vacant buildings and the crime at a neighborhood.



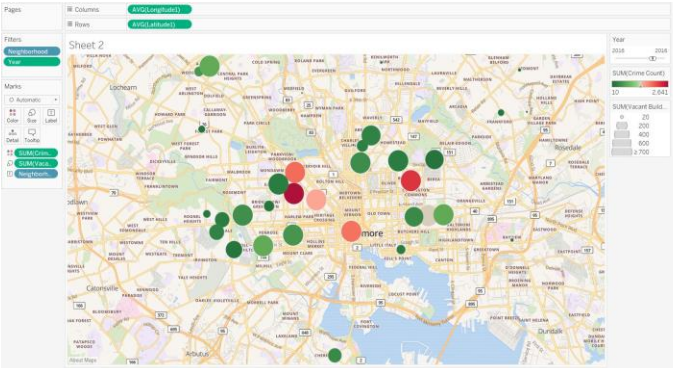
Year: **2014**

In the year 2014, we observed that the crime is being the same whereas the vacant buildings have been increasing. This says that the ratio of vacant buildings and crime is not as expected.



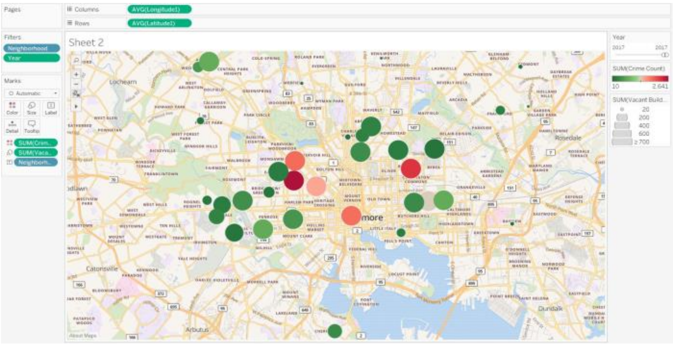
Year: **2015**

In 2015, we have observed there is a little increase in crime where as some of the vacant buildings are also increasing. But this does not support our hypothesis.



Year: **2016**

Some of the neighborhoods remains same with the crime count and the number of vacant buildings were still increasing irrespective of the crime, which shows a little correlation between the crime and the vacant buildings.



Year: **2017**

Similarly, like previous years, here the sum of vacant buildings is increasing more than the crime rate in a neighborhood. As shown above, we can see that the difference of the bubble increasing whereas the color of that neighborhood does not have a great change. This clearly says that there is very little correlation between the crime and vacant buildings.

As we are visualizing a location level, for showing the hotspots we are pointing only one point in a neighborhood. For that, we are adding the entire crime count and vacant buildings in that neighborhood. With the help of filter, we have chosen the neighbourhoods which were supporting our hypothesis and some of them which were not supporting our hypothesis and visualizing the hotspots.

When looking from higher level the neighborhoods with the count of crime and vacant buildings doesn't yield a good visualization aspect of the results we have obtained in the analysis phase. As this is the total sum of crime and total sum of vacant buildings of a neighborhood. So, let us further drill drown to the latitude and longitude level i.e. looking at each location in a neighbourhood so that we can get a crisp visualization results.

E. Perspective 5: *Location Level – Each Location*

For better visualization of our analysis at location level, we are drilling down deeper to look at each location in a neighborhood to better understand the scenario of crime in and around vacant buildings.

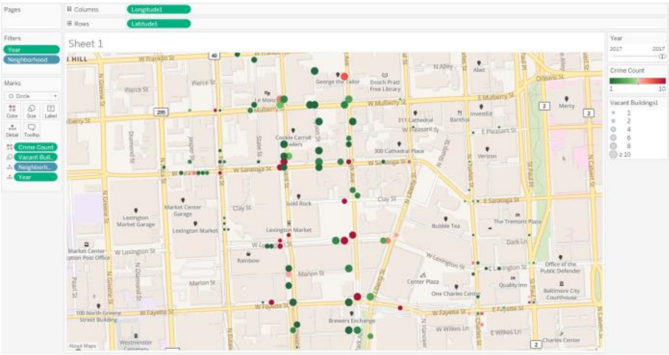
Every location in the neighborhood has a correlation, but some locations in a neighborhood shows positive correlation of 0.7 and above whereas some locations in that same neighborhood shows correlation below 0.7 and negative correlation between crime count and no. of vacant buildings. In every neighborhood, there are both locations with positive and negative. So, neighborhoods with most of the positive correlation locations in a neighborhood is taken as a hotspot.

Some neighborhoods supporting our hypothesis at location level are Downtown, Arlington, Upton, Elwood Park/Monument, Penn North. And some neighborhoods not supporting our hypothesis at location level are Better Waverly, Darley Park, Easter Wood, Broadway East, Sandtown-Winchester.

Here we have limited the crime count as 10, which means that the crime-count nearby 10 and above 10 shows high crime rate at that location and is represented with red color and less crime count is represented with green color. And the size of the vacant buildings is taken as 10, which means vacant buildings nearby 10 and above 10 at a location is represented with size.

1) *Neighborhood supporting our hypothesis*

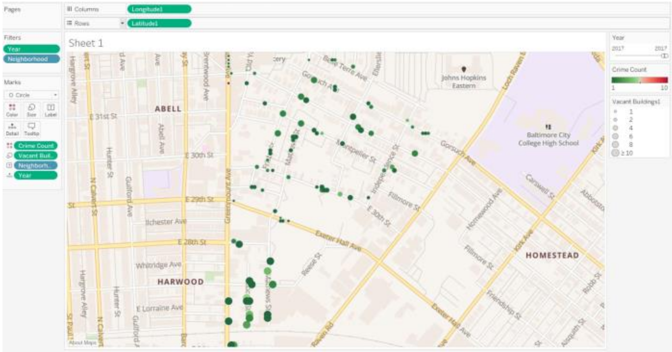
Baltimore Downtown is one of the neighborhood hotspot which supports our hypothesis.



Baltimore Downtown is one of the neighborhood hotspot showing most of the locations with positive correlation. This neighborhood has some locations with the vacant buildings count as 3 and the crime count as 116 from year 2013 to 2017, which is indeed a very high crime rate at a location. As we don't have any supporting documents we cannot strictly say that crime has occurred due to the vacant building, but we can strongly say that there is crime happening in and around that vacant buildings and should be minimized.

2) *Neighborhood not supporting our hypothesis*

One of the neighborhood that does not support our hypothesis is Better Waverly.



Columns: Longitude1
Rows: Latitude1
Filters: Year, Neighborhood
Marks: Circle
Crime Count, Vacant Buil.., Neighborh.., Year
Sheet 1
Year: 2017 – 2017
Crime Count: 1 – 10
Vacant Buildings1: 1, 2, 4, 6, 8, ≥ 10

As shown below you can see that the crime count is less than that of vacant buildings. Which proves our analysis results with little correlation between crime and vacant buildings in a neighborhood. If you look at the map you can see that the count of vacant buildings is more, and the crime count is minimum at every location. Here the locations with most of the correlations below 0.7 and negative correlations exists than locations with positive correlation. The ratio of vacant buildings is more in this neighborhood than crime count and this shows a very little correlation with crime count and vacant buildings.

IX. Conclusion

As per the above analysis Considering Neighborhood semiyearly level, the data supports our hypothesis with a confidence of 45% as the percentage of neighborhoods falling in the range of average r value is 45%.

The data supports our analysis when compared to other level data. At location level, the data supports our hypothesis with a confidence level of 33.6% for the following locations: Brooklyn, Care, and Franklin Square as these neighborhoods had the highest correlation are ought to be taken immediate action to control the crime rate by demolishing the vacant buildings or increasing the security around or near the buildings.

Arlington, Better Waverly, Ellwood Park/Monument, Franklintown Road are neighborhoods with considerably string correlation which also should be taken action to reduce the crime count by fencing the vacant buildings or renting the vacant buildings.

Acknowledgment

We are very grateful to the City of Baltimore for allowing us to use the Baltimore City Open Data Source for our research. We are also grateful to Dr. Joshi for guiding us over the course of our project.

References

[1] "The 10 Most Dangerous U.S. Cities." Forbes, 2017, https://www.forbes.com/pictures/mlj45jggj/7-baltimore/#56c6fbca5487. Accessed 10 Dec. 2017.

[2] A, Malathi, Dr. Santhosh Baboo, and Anbarasi A. "An intelligent Analysis of a City Crime Data Using Data Mining." *2011 International Conference on Information and Electronics Engineering*, vol. 6, 2011, pp. 130-34, www.ipcsit.com/vol6/26-E049.pdf. Accessed 8 Dec. 2017.

[3] *Open Baltimore*, City of Baltimore, 2017, https://data.baltimorecity.gov/. Accessed 10 Dec. 2017.

[4] Pearson Product-Moment Correlation. (n.d.). Retrieved December 15, 2017, from https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

[5] Computing the Pearson Correlation Coefficient. (n.d.). Retrieved from http://www.stat.wmich.edu/s216/book/node122.html

[6] Tableau Training & Tutorials. (n.d.). Retrieved December 15, 2017, from https://www.tableau.com/learn/training?qt-training_tabs=1#qt-training_tabs

[7] I. (n.d.). Tableau Tutorial for Beginners - Tableau Tutorial PDF. Retrieved December 15, 2017, from https://intellipaat.com/tutorial/tableau-tutorial/

[8] HUD USER. (n.d.). Retrieved December 15, 2017, from https://www.huduser.gov/portal/periodicals/em/winter14/highlight1.html