Baltimore, Maryland:

Neighborhoods Explored by Number of Crimes and of Vacant Buildings

Introduction:

1. Background:

Baltimore, Maryland is a troubled city. Known as "Charm City" Baltimore has lost its charm with a shrinking population, high crime rate, and bleak outlook.

Baltimore has many problems, but also many amenities. Home of both the Baltimore Orioles and Baltimore Ravens it has two large sporting venues near the Inner Harbor. The Inner Harbor is a popular tourist venue with many sites such as the National Aquarium in Baltimore and the USS Constellation.

2. Problem:

The problem explored in this project is the relationship between the number of vacant buildings in the city of Baltimore and the highest crime areas of Baltimore. Looking at neighborhood data we will view the neighborhoods that have the highest reported number of crimes and the neighborhoods that have highest number of vacant buildings. We will also explore the relationship between the two variables.

Then we will use the Foursquare location data to explore the venues in selected neighborhoods to see what the venue data can tell us about those neighborhoods.

3. Stakeholders:

For anyone who loves the city of Baltimore or cares about the health of cities and their developments it is important to understand the relationships between different variables that can affect quality of life. The safer a neighborhood the better for everyone who lives and works there.

Data:

1. Data Sources:

The data sources for this project are the open city data for Baltimore City from the following location:

https://data.baltimorecity.gov/

The source for our crime data is victim-based crime report which holds all individual incidents dating back to 2012. Each line is a criminal incident and includes location, type of crime, date and other details.

https://data.baltimorecity.gov/Public-Safety/BPD-Part-1-Victim-Based-Crime-Data/wsfq-mvij/data

The vacant building data is at the following location. It shows all the locations in Baltimore of vacant buildings with their addresses and neighborhoods and other identifying information.

https://data.baltimorecity.gov/Housing-Development/Vacant-Buildings/qqcv-ihn5/data

The GIS file for the Baltimore neighborhoods themselves. This had the Geojson file that provided the shapefiles for the neighborhoods.

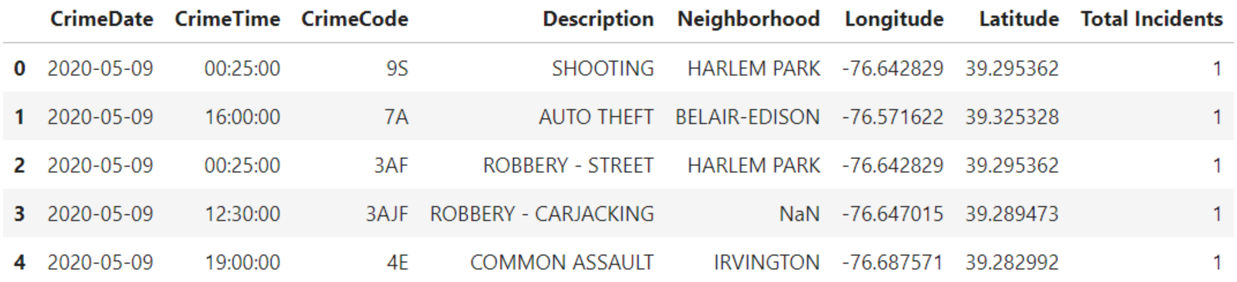
https://data.baltimorecity.gov/Geographic/Neighborhoods/h8i5-gvdz

Foursquare API. This is where one accesses the Foursquare location data.

https://api.foursquare.com

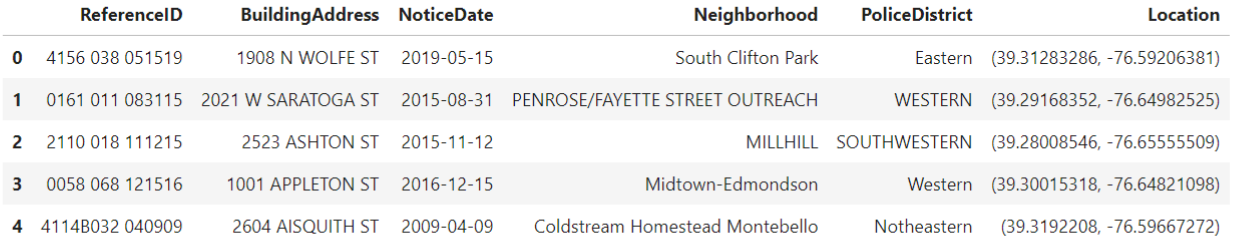
2. Data Selection and Clean-Up:

To reflect the current crime levels of a neighborhood we pulled the crime data for the past calendar year (May 09 2019 - May 09 2020). In order to keep the data saved we saved it as an excel. We then pulled the excel into pandas to create a data frame.



	CrimeDate	CrimeTime	CrimeCode	Description	Neighborhood	Longitude	Latitude	Total Incidents
0	2020-05-09	00:25:00	9S	SHOOTING	HARLEM PARK	-76.642829	39.295362	1
1	2020-05-09	16:00:00	7A	AUTO THEFT	BELAIR-EDISON	-76.571622	39.325328	1
2	2020-05-09	00:25:00	3AF	ROBBERY - STREET	HARLEM PARK	-76.642829	39.295362	1
3	2020-05-09	12:30:00	3AJF	ROBBERY - CARJACKING	NaN	-76.647015	39.289473	1
4	2020-05-09	19:00:00	4E	COMMON ASSAULT	IRVINGTON	-76.687571	39.282992	1

The same method was used to pull in the vacant buildings data.

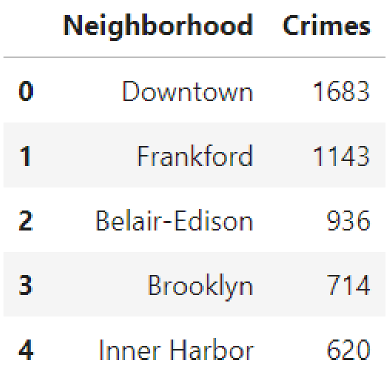


	ReferenceID	BuildingAddress	NoticeDate	Neighborhood	PoliceDistrict	Location
0	4156 038 051519	1908 N WOLFE ST	2019-05-15	South Clifton Park	Eastern	(39.31283286, -76.59206381)
1	0161 011 083115	2021 W SARATOGA ST	2015-08-31	PENROSE/FAYETTE STREET OUTREACH	WESTERN	(39.29168352, -76.64982525)
2	2110 018 111215	2523 ASHTON ST	2015-11-12	MILLHILL	SOUTHWESTERN	(39.28008546, -76.65555509)
3	0058 068 121516	1001 APPLETON ST	2016-12-15	Midtown-Edmondson	Western	(39.30015318, -76.64821098)
4	4114B032 040909	2604 AISQUITH ST	2009-04-09	Coldstream Homestead Montebello	Notheastern	(39.3192208, -76.59667272)

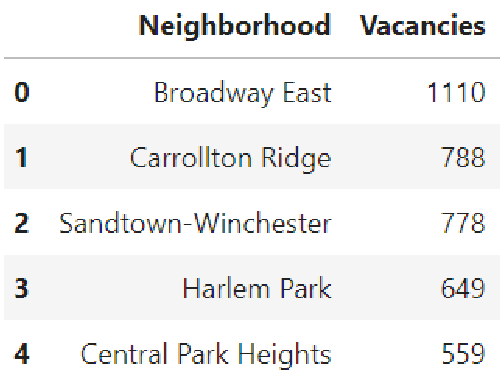
As we can see from the sample of the vacant building data the neighborhoods are formatted differently within this single dataset. We ran stripping code and code to change the format of the words in order to standardize the neighborhoods in the data set.

Since the focus is on neighborhoods, we had to ensure that the neighborhoods are in the same format across all three data sources. Since the Geojson is the hardest file to change the formatting on, the neighborhoods will be in title format.

We then group both the crime and vacancy data by the neighborhoods to get the count of individual items for each neighborhood.



	Neighborhood	Crimes
0	Downtown	1683
1	Frankford	1143
2	Belair-Edison	936
3	Brooklyn	714
4	Inner Harbor	620



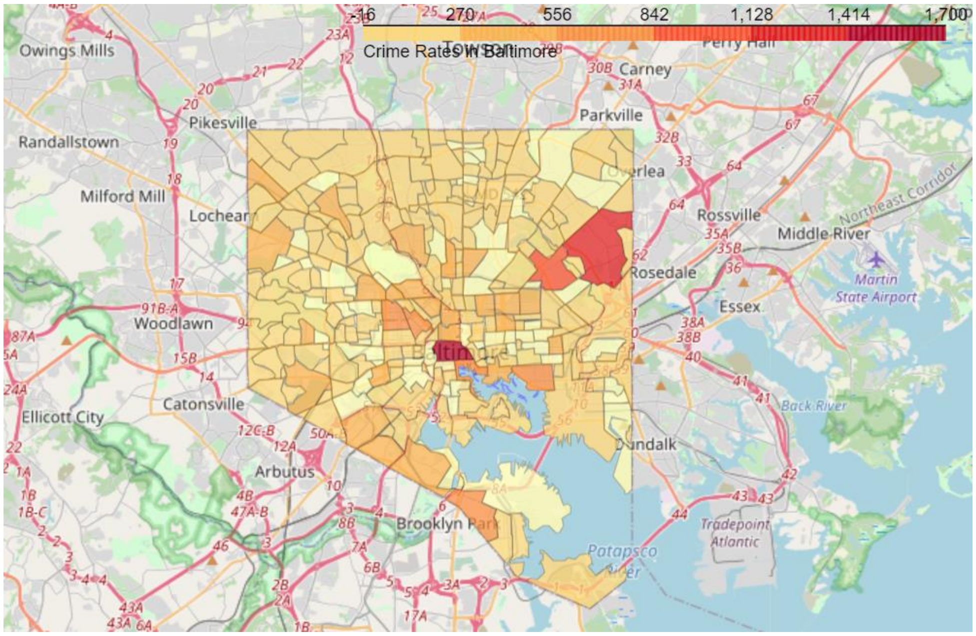
	Neighborhood	Vacancies
0	Broadway East	1110
1	Carrollton Ridge	788
2	Sandtown-Winchester	778
3	Harlem Park	649
4	Central Park Heights	559

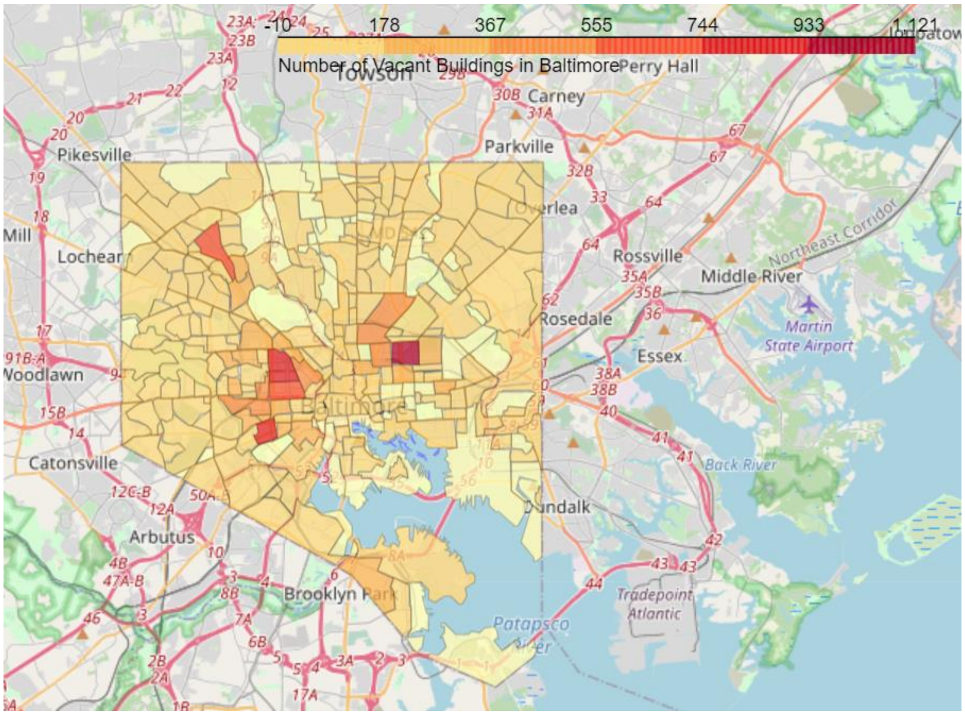
Methodology:

1. Compare the Different Data Sets:

The first thing is to compare the two datasets. Crime lists 270 neighborhoods while the vacancies data only shows 220 neighborhoods. This is understandable as not every neighborhood will have vacancies.

In order to understand our problem set better we visualize where our neighborhoods are located. We create two choropleth maps to show the number of crimes by neighborhood and the number of vacancies by neighborhood.

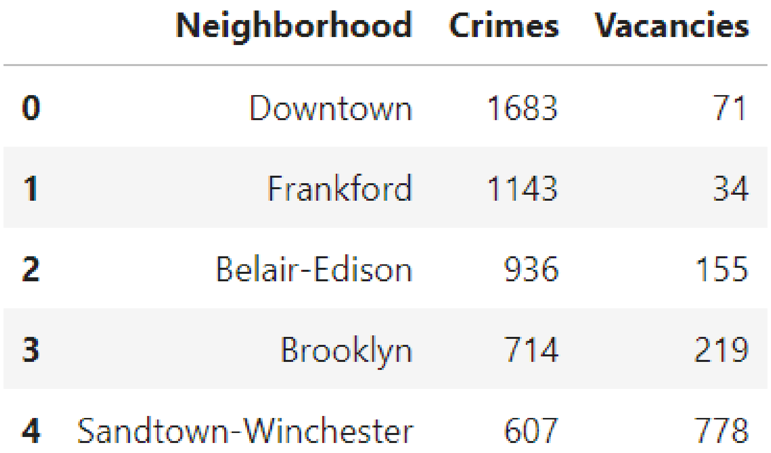




Upon seeing visually that there is not a strong overlap in crime and vacancies, i.e. that the highest crime numbers are not in the highest vacancy areas, we perform some regression analysts to confirm our eye-test.

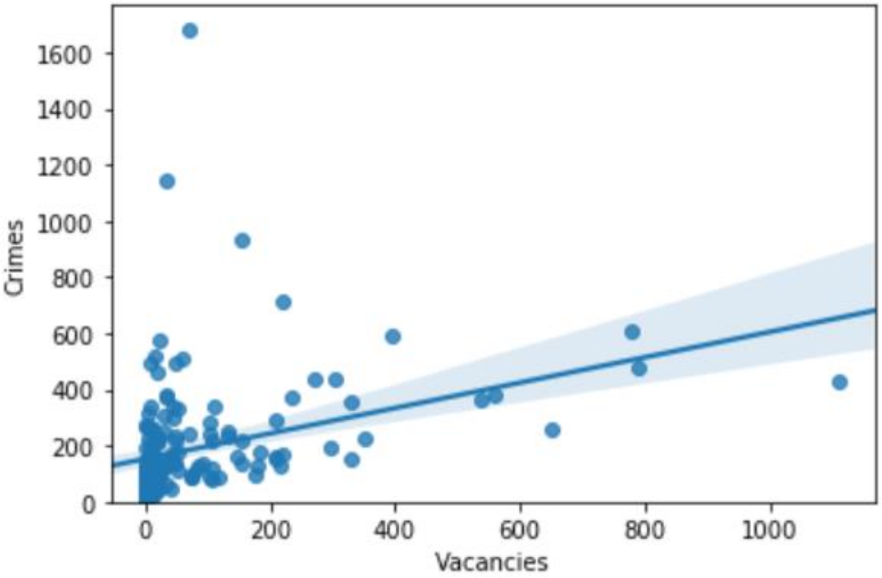
2. Combine Data Sets:

We combine the two data sets into one data in order to do comparisons and run some models.



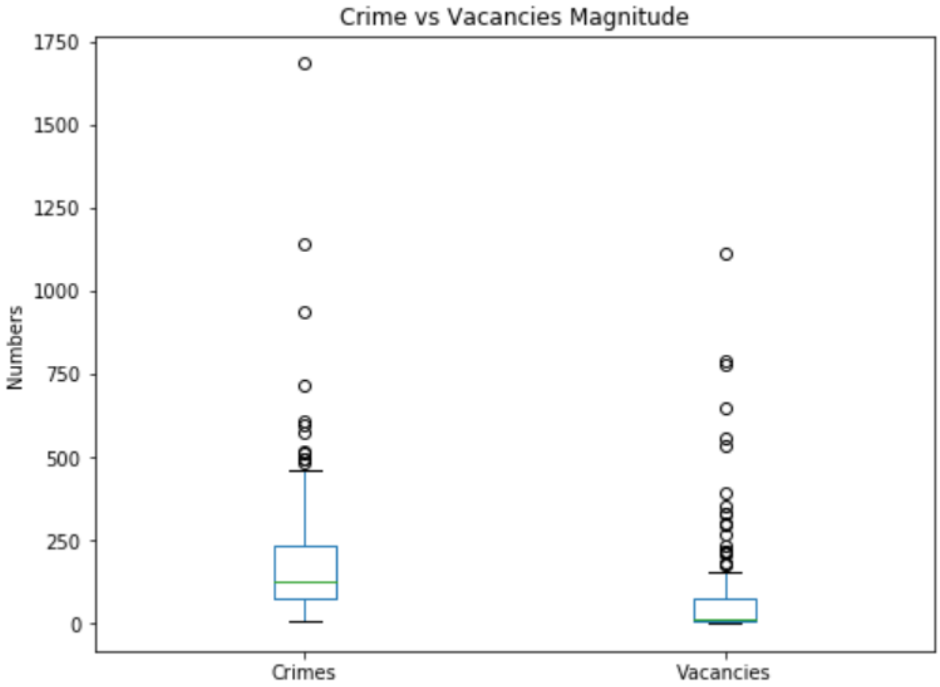
	Neighborhood	Crimes	Vacancies
0	Downtown	1683	71
1	Frankford	1143	34
2	Belair-Edison	936	155
3	Brooklyn	714	219
4	Sandtown-Winchester	607	778

First we do a regression model to confirm our eye-test of weak correlation.



We also pull the correlation index and get a 0.344. showing that there is not a strong correlation between the two variables.

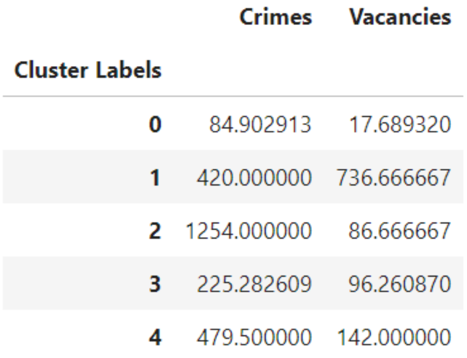
Even though there is not a strong correlation we explore deeper into the neighborhoods to see what insight we can still gain with our data. Next we compare the rates of each occurrence by creating a box plot with our combined data.



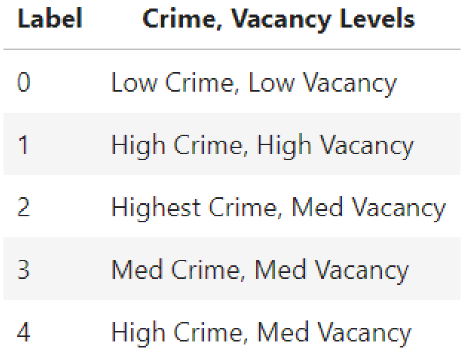
The box plot shows us that crime incidents are much higher than the number of vacancies. This helps us determine what numbers are within a reasonable range and what numbers are exceedingly high for both variables.

3. K-Clustering:

To explore the neighborhoods based on these variables further we create clusters using k-clustering. This was used to find any relationships that we were not seeing based off of pure correlation. Finding the mean number of the clusters and using the information we glean from the box plot we are able to determine five separate types of clusters.



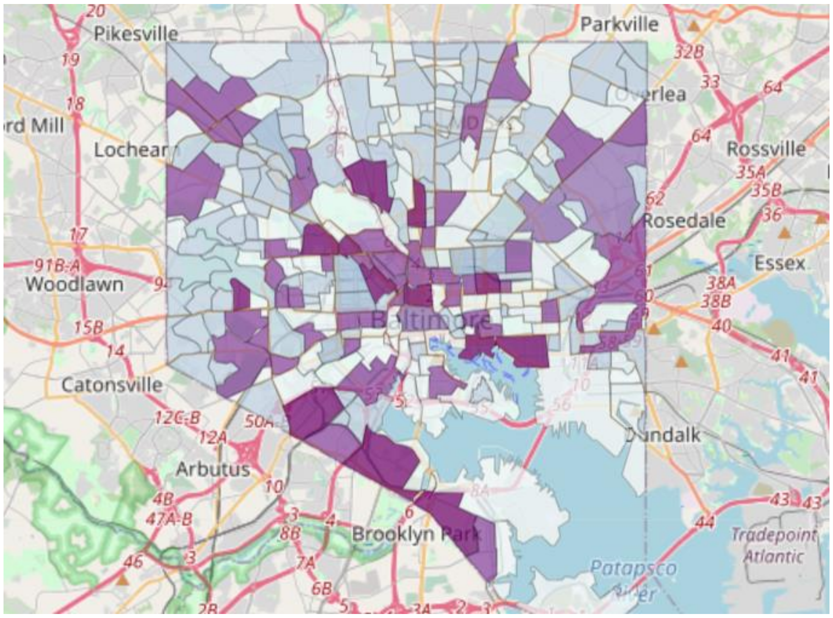
	Crimes	Vacancies
Cluster Labels		
0	84.902913	17.689320
1	420.000000	736.666667
2	1254.000000	86.666667
3	225.282609	96.260870
4	479.500000	142.000000



Label	Crime, Vacancy Levels
0	Low Crime, Low Vacancy
1	High Crime, High Vacancy
2	Highest Crime, Med Vacancy
3	Med Crime, Med Vacancy
4	High Crime, Med Vacancy

With this information we see that the weak correlation between the variables lead to a wide range or crime numbers within the vacancy numbers.

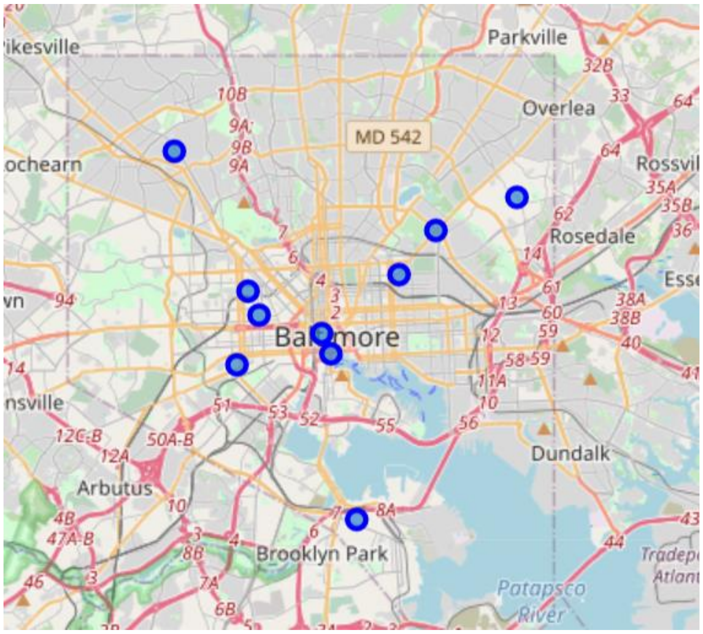
We mapped our clusters to again show the variance within the city.



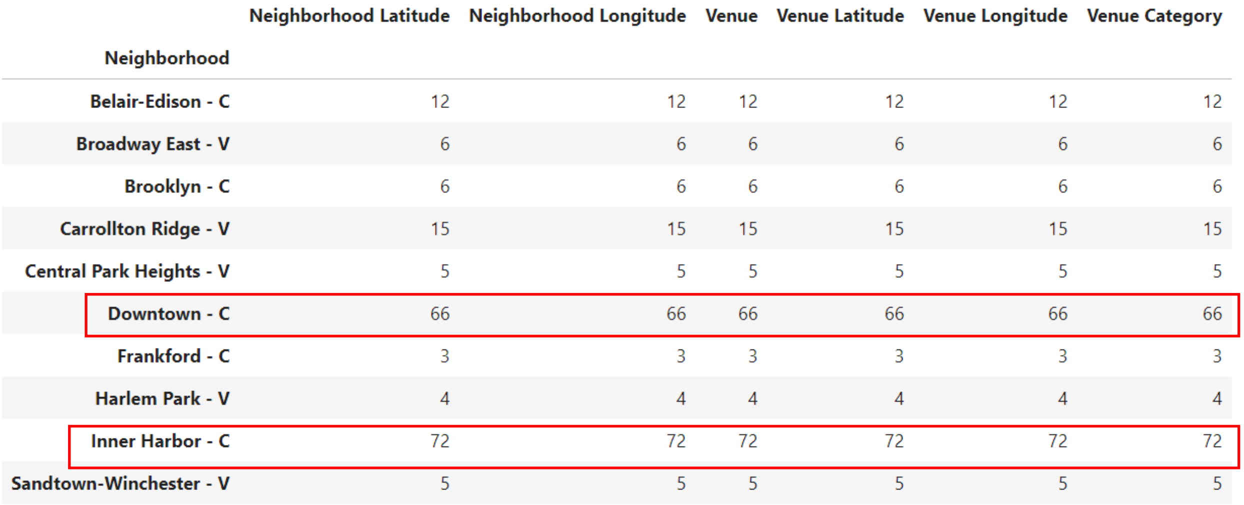
4. Foursquare Data:

To further explore the neighborhoods, we decide to look at the neighborhoods with the top five vacancy numbers and the neighborhoods with the top five crime numbers and add them into one list.

We created a top ten neighborhoods of interest list by combining the two top five lists. We labeled each neighborhood with a ‘C’ or a ‘V’ to show which list it came in on. Based on our previous research we are not surprised that there are no overlapping neighborhoods in this data set. The longitude and latitude of each neighborhood was gathered and we created a map of their locations.



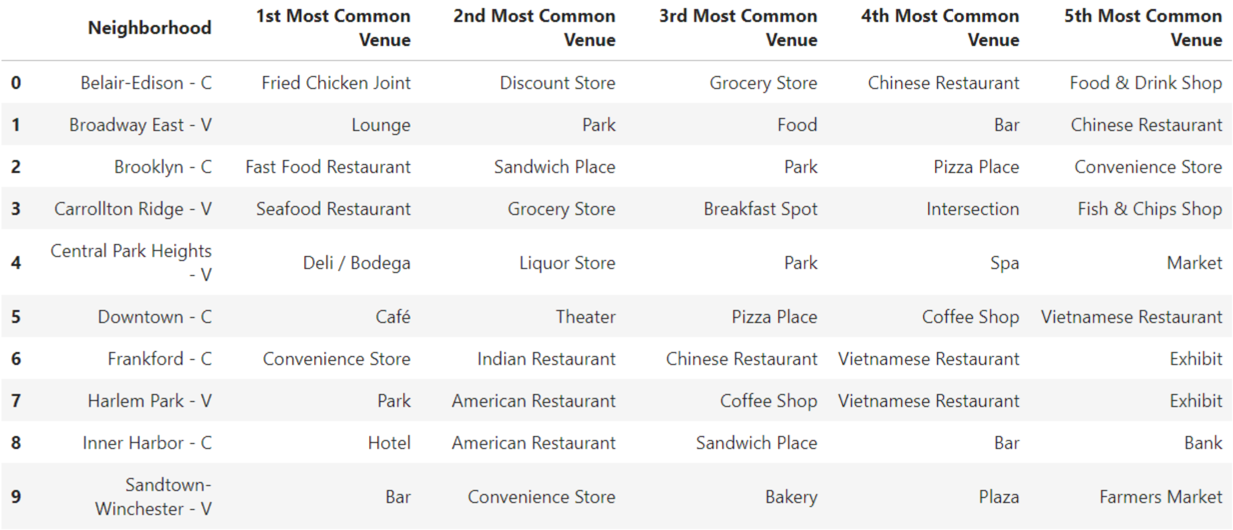
Using the foursquare data we pulled the number of venues for each neighborhood.



	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Belair-Edison - C	12	12	12	12	12	12
Broadway East - V	6	6	6	6	6	6
Brooklyn - C	6	6	6	6	6	6
Carrollton Ridge - V	15	15	15	15	15	15
Central Park Heights - V	5	5	5	5	5	5
Downtown - C	66	66	66	66	66	66
Frankford - C	3	3	3	3	3	3
Harlem Park - V	4	4	4	4	4	4
Inner Harbor - C	72	72	72	72	72	72
Sandtown-Winchester - V	5	5	5	5	5	5

We see that all but two of the neighborhoods have fewer than 15 venues. The two neighborhoods with more than 15 venues are the Inner Harbor and the Downtown neighborhoods with 72 and 66 venues respectively which is magnitudes higher than the other neighborhoods.

Since the other neighborhoods have such few venues, we do a top five venue search and see the following.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Belair-Edison - C	Fried Chicken Joint	Discount Store	Grocery Store	Chinese Restaurant	Food & Drink Shop
1	Broadway East - V	Lounge	Park	Food	Bar	Chinese Restaurant
2	Brooklyn - C	Fast Food Restaurant	Sandwich Place	Park	Pizza Place	Convenience Store
3	Carrollton Ridge - V	Seafood Restaurant	Grocery Store	Breakfast Spot	Intersection	Fish & Chips Shop
4	Central Park Heights - V	Deli / Bodega	Liquor Store	Park	Spa	Market
5	Downtown - C	Café	Theater	Pizza Place	Coffee Shop	Vietnamese Restaurant
6	Frankford - C	Convenience Store	Indian Restaurant	Chinese Restaurant	Vietnamese Restaurant	Exhibit
7	Harlem Park - V	Park	American Restaurant	Coffee Shop	Vietnamese Restaurant	Exhibit
8	Inner Harbor - C	Hotel	American Restaurant	Sandwich Place	Bar	Bank
9	Sandtown-Winchester - V	Bar	Convenience Store	Bakery	Plaza	Farmers Market

Results:

The results were disappointing, where we expected to find correlation there was a very weak one. In using the k-clustering to see what relationships we could find in the clusters we did discover one surprising result which was that the highest crime area only had medium vacancy numbers, while other medium vacancy areas ranged from low crime to high crime. Those clusters were not surprising based on the weak correlation previously discovered.

Looking at the venue data we see that most of the high vacancy/high crime neighborhoods have very little venues and places of interest. The exceptions are the Inner Harbor and Downtown neighborhoods, which were high crime neighborhoods. These results indicate that criminal incidents may be high due to the number or individuals in that neighborhood, primarily visitors and tourists. Outside of those two neighborhoods, however, we can see that the highest crime and highest vacancy neighborhoods do suffer from a lack of venues.

Discussion:

Part of data modeling is getting the results one does not expect. If our assumptions were always correct we would not need to model them. In exploring the neighborhoods we see that some affluent areas have a high number of crimes. Further study could be made into the different types of crimes and whether some crimes are higher in some neighborhoods verses others. Does the type of crime track more closely to vacancy rates?

Furthermore, the scope of this study only looked at crime numbers, not rates. Due to the nature of visitors into a city, it would be hard to calculate the number of crimes per people at any given time in a neighborhood (plus citizens of Baltimore go to other neighborhoods throughout their day). Therefore, the data is not looking at actual crime rates, just reported incidents. This is same for vacancies. The data was not normalized where the full number of available buildings was calculated for each neighborhood. This means we are just counting vacant buildings and not the percentage of vacant units in a neighborhood. Since the neighborhoods are not the same size these distinctions could easily create outliers.

The above are all avenues of approach that could be explored in future studies.

Conclusion:

The correlation between the number of crimes and the number of vacant buildings in a neighborhood is weak. The types of crimes and the actual percentage of vacant buildings were not explored in this study. It is recommended future studies explore those issues before deciding that there is no correlation between the two variables.

Neighborhoods with high vacancies and high crime tend to have few venues. The exception are the major tourist hubs of Baltimore, which have high crime but a large amount of venues. This indicates that economic development is still critical for improving neighborhoods, but that crime is a more complex issue that will not be solved just through economic development.