**CMSC 491 (Intro to Data Science)**

**Final Project –**

**“Relationships Between the Property Tax, Vacancy List, and Crime Rate of Regions of Baltimore”**

**Completed 5/5/2019**

Dagem Kebede

Dylan Zach

**ABSTRACT**

The primary goal of this research project was to predict trends in crime rates by region in Baltimore City through the analysis of initially seemingly unrelated factors. For this end, we analyzed the property vacancy frequency and property tax rate in relation to each other per region in Baltimore City (Northeast, central, etc), and the prevalence of victimized crime occurring in each of these regions.

The objective is to conclude some correlation between these data sets, and what could be inferred based on these results about the trends that divide communities at large. This report also seeks to examine the relationship that these variables have with each other in varying strengths, in order to use linear regression data training in order to predict the crime rates in the regions of Baltimore city using modeling.

In total, we have found strong evidence to support the correlation between property tax rate and property vacancy frequency in a given region, and that these properties can in fact be used to accurately predict the crime occurrences on a linear regression model.

**SECTION 1: AN IN-DEPTH DESCRIPTION OF DATA SETS USED**

For this project, a total of three major data sets were incorporated. All of the sets were obtained from the OpenBaltimore data resource site, which is a source of governmental data provided by the city in order to “promote transparency and promote innovative use” (to paraphrase the site).

The first data set is a listing of all currently vacant buildings within the city of Baltimore. The Housing Authority of Baltimore City maintains the data collection of over 16,000 current vacancies, and updates the current statuses every few months.

Each row in the set is representative of a singular building, with its address and relative location included. This includes its block number, the neighborhood and police district it is located in, as well as a special reference ID, which I can assume is used for the direct identification of a single building.

In a sense, the statistic of property vacancy is able to stand in for the economic health and prosperity of a region at a given point in time. Ideally, if there is a low vacancy of property, then businesses must be employing infrastructure to its fullest potential. So, as I would argue, vacancy is a strong indicator of periods of wealth or decline. To illustrate that, we can analyze data in terms of another variable of the set, or the “date of vacancy”: the date at which a given property was officially marked as vacant.

If we graph the total quantity of entries over a line graph by year, then we get a representation of what years of vacancy posting are most prevalent in our set. [See: Figure 1.A]

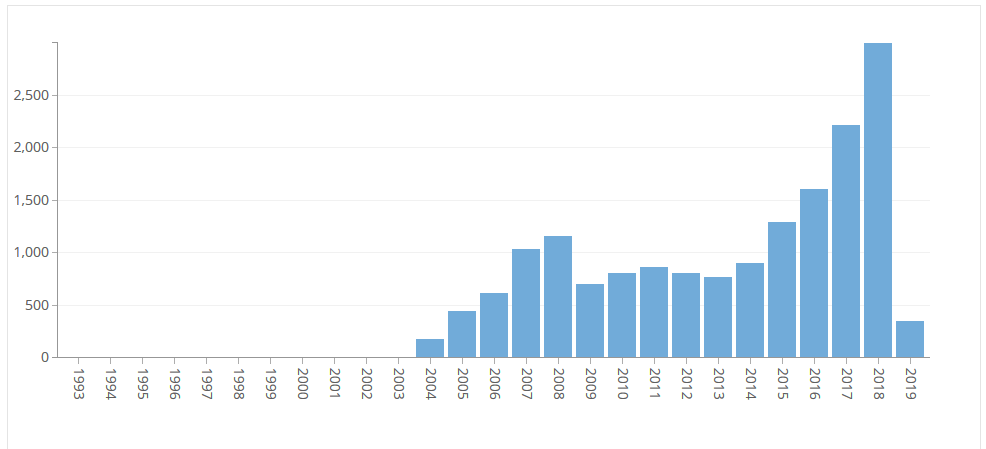


Figure .A: A depiction of # of buildings currently vacant in Balt. City by year originally made vacant.

As the graph shows, the year with the highest amount of vacancy dates is 2018, which makes sense. A vacant building is likely to become re-inhabited after a short period of time, and buildings which are less desirable to re-inhabit will remain vacant for long periods of time. So, there is drop-off over time. 2019 is not very well represented, but this is due to the data collection being updated at only February 15th, the most recent entry being from only the 7th of the month.

In addition, there seems to be an interesting spike in the data size for the years of 2007 and 2008, compared to the years that immediately follow it. This could be reflective of the economic recession of these years, a time period which would have resulted in a less stable economy, and a worse building and housing market. The increase in evictions and vacancies from that period could thus be still visible today.

This set’s importance thus lies in its probable correlation with economic health and stability. Buildings represent both infrastructural and corporate investments, as well as stable housing for the general population. An increase in vacancies could be closely correlated with negative economic impact, the results of which could impact other elements of people’s lives. This could include up to a correlation with crime rate; does crime rate increase when the infrastructure market sours? Even if it does not seem directly correlated, it would represent the portions of the bigger story.

[See: Addendum 1 – A representation of the vacancies in each region of the city.]

The second data set is the current real property tax on each registered building within Baltimore. This set is maintained by the Department of Finance, and is a truly massive set in its size, cataloguing over 238 thousand individual buildings and properties.

Each individual building has its location presented in multiple forms, including the address, block, neighborhood, and police district that was also recorded by the vacancies data set. (These similarities could be uniform across the departments that store the data, which also makes for convenient table cross-compilation.) Each property is also recorded as being either a primary residence or not (i.e. house or otherwise). This was a large difference not present in the vacancy set.

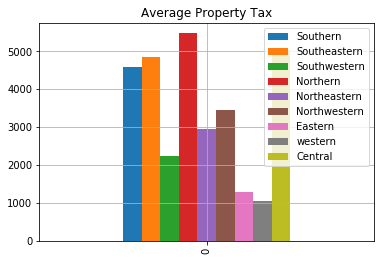
Also in the set was the recorded City and State tax which was calculated owed for each property, based on the size and usage of each property. Presumably, both city and state taxes on a property will increase in proportion with both the location of the property, the value of the property, and also the usage of the property. As an example, we can graph the city tax on a property across the various regions of Baltimore to get a rough estimate of property values by region. [See: Figure 1.B, to left].

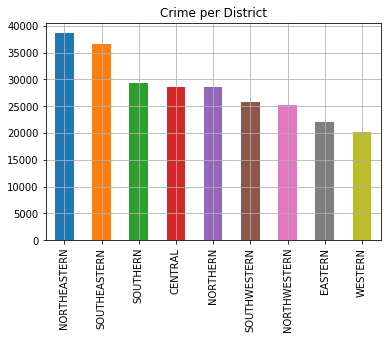
Figure 1.B: Average property tax by region (in USD), accounting only for principal residencies. Includes both city and state taxes.

If we graph each region by the average property tax on principal residencies, we see that the highest levels of property tax are in the Northern and Central regions of the city, wheras the lowest taxes tend to be in the Western region.

For the sake of simplicity, and because it is most likely to be directly relevant to a correlation with both the crime statistic and vacancies, we are excluding parts of the data to limit it to only principal residencies, or basically houses and apartments. This way, we can further our original hypothesis that the housing market specifically has a large correlation with the crime rate for a specific region.

One interesting thing about this is that each of us disagreed as to what correlation crime rate could have with mean property tax by region. It could be that places with lower average tax rates have a higher incidence of crime due to the lower costs of living in these regions. With uniform policies for housing and otherwise in the city, the major difference would come down to the property value and significance in each region. Because poorer areas are less significantly developed and have less demand for further development in the purely business model, I would argue that crime rates increase in areas where mean tax value is lower. However, this intuition could be incorrect. For the time being, I would say that any conclusions drawn from this data could still be significant in the analysis of crime contribution factors, and the components which lead to such.

Finally, both of these data sets are to be compared relative to a seemingly unrelated dataset – the Victim-based crime data provided by the Baltimore Police Department. This data set lists each victim-based crime in Baltimore, as well as a few key points for each incident. In total, there are around 255 thousand incidents, with the timeframe of data collection being from 2014 on New Year’s to April 27, 2019.

 The crime data set, much like the last two, also keeps track of location of incidents in a systematic manor, with the neighborhood, police region, and approximate address being recorded. (In the original data set, coordinates were also recorded, but I believe that this data set was formed as a derivative of the original, and this attribute is unincluded. I just thought this was an interesting statistic that was unique to the circumstance of the subject, where location is not always an exact address.)

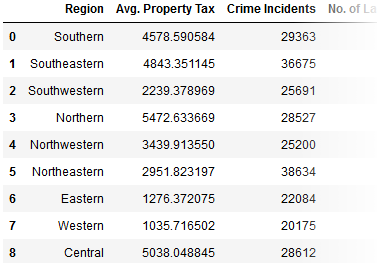
Finally, there is an attribute for a few different types of crimes, such as larceny, burglary, and “common assault”. As important as the amount of crime is, I hope to also compare results dealing with the particular offense at a later point. Although I could predict intuitively a jump in crime rates for when looking at places with poor housing infrastructure and economy, which crimes in particular would become more prevalent?

If we look at the cumulative crime prevalence per region in Baltimore [See: Figure 1.C], we see that it doesn’t immediately tell too much on its own about the region. The difference in number of incidents could just be tied to the population of the region. However, what we intend to explore ought to go beyond that. What does the crime rate mean, then, in the scope of the different regions as a whole? Our goal is to identify the correlative trends that separate well-off and struggling regions, so all aspects ought to be considered.

Figure 1.C: Total reported incidents per police district, 2014-Pres.

**SECTION 2: CONNECTIONS BETWEEN DIFFERING TABLES**

In order to effectively compare the data sets, we need to find some attribute to unite the sets by. This is so that we have a means to express the data. Since all tables share very similar ways in marking location of instances, we decided to use the Police Regions to express the other relationships between sets.

 The way that we incorporated data into each of these Region data points was, admittedly, mostly done by hand. Each column in a specially designed table was meant to incorporate a data instance taken from a given table, in terms of their relationship to the regions of Baltimore. For example, our first column tracks the average property tax for primary residencies of each different region of Baltimore. This is especially helpful, as it makes constructing scatter plots more practical and representative of their data. [See: Figure 2.A].

The driving factor for using this method primarily to represent the similarities between datasets comes from the fact that a lot of the major factors are numerical in data type. Scatter plots are easy to derive positive or negative correlations from as well, even if they lack the intrinsic values at each point.

Figure 2.A: Each region has its value for a particular stat recorded.

**A - The connection between Housing Vacancies and Local Property Tax Mean**

Housing vacancies and local property taxes are, as I surmised earlier, tied to the economic strength of a region. As unrelated as the two seemingly are, they both indicate a difference between poor and richer regions. For housing vacancies, a richer economy will waste less infrastructure and many people will be able to afford and live in homes. During a housing crash, this number will go up. Like I said before, this is still evident today with the increase in vacancies from the years 2008-2009.

Local property taxes are derived from the value that a certain property has, which is itself dependent on a number of other factors. This usually includes property location, size, etc. Assuming that tax laws are implemented uniformly, this can be a good indicator for where more expensive and expansive homes are located within the city if we find the mean for each of our regions.

For this comparison, I constructed a scatter plot comparing the mean (city+state) property taxes with the number of vacancies in given region, to get a sense of the property situation in the area. [See: Figure 2.B]

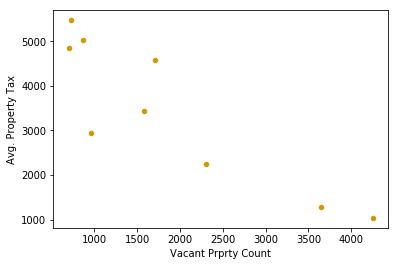
When we do plot the data sets against each other, we get a very strong negative correlation of the set, with a relative correlation of approx. -0.89 for the set. Although this may seem intuitive, it also contradicts another intuition about the set. For one, shouldn’t lower property taxes and property values mean more people are able to afford homes?

Figure 2.B: “Vacant Property Count vs Avg. Property Tax rate per region”. Corr = -0.89

In a way, this alone actually highlights a pretty striking detail about relative poverty in a region; the poor regions are far worse off than the rich ones. If these variables are so tightly correlated, then the income difference mean between these regions must be far greater than a linear correlation. If the latter was true, then the number of vacancies per region would probably stay more or less the same by region, just with the property values adjusted to poorer region’s incomes. In this model, the poor get poorer.

One thing that was really annoying about this set. For whatever reason, when I initially ran the scatterplot, the “Northeast” region was only returning about 8 instances of vacant buildings. Now, that could have some greater significance, such as possibly some astronomically better housing codes. I then went back to check the data set in-depth, and found that for every single instance of that region, the actual data set said “Notheast”. For every single one of them. This is a bit tragic, but it also illustrates that often the likelihood of an outlier in a set is just the result of some data error.

**B – The connection between Avg. Property Tax and Victim-based Crime**

The victim-based crime dataset we used catalogues all interpersonal-related crimes committed since 2014. Like before, we are going to group the incidents by their police region of occurrence, which is another statistic kept by this data set.

Now that we can be pretty sure of the differences between economically well-off and impoverished regions of Baltimore, it would make sense to see if this is indicative of any other trends, the most obvious to pick having been crime. Besides that, though, we picked this statistic specifically to see that could be the effects of poverty on crime levels. If there exists correlations, then we can be sure of the problems or lack thereof that come with the burden of poverty.

Again, we’re using a scatterplot to initially compare the mean property tax with the total number of victim-based crimes for each reason. Like I said before, we each disagreed on what the outcome would be. We agreed that it would likely be correlative, but whether places with higher taxes (and greater housing wealth) saw more crime or less was debateable. I argued that crime would likely stay within its region of the city, and that poverty in an area would also incur more people to commit a victim-based crime.

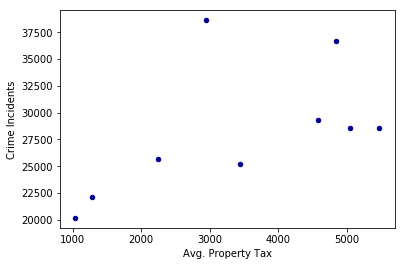
However, this was not necessarily the case. [See: Figure 2.C]. Overall, the prevalence of an amount of given violent crime increases as mean property tax increases for a region. The correlation is somewhat weak, at 0.55. However, I feel like this alone didn’t tell me enough about crime rates to be able to draw conclusions.

Figure 2.C: “Avg Property Tax vs Crime incidence rate per region.” Corr = 0.55

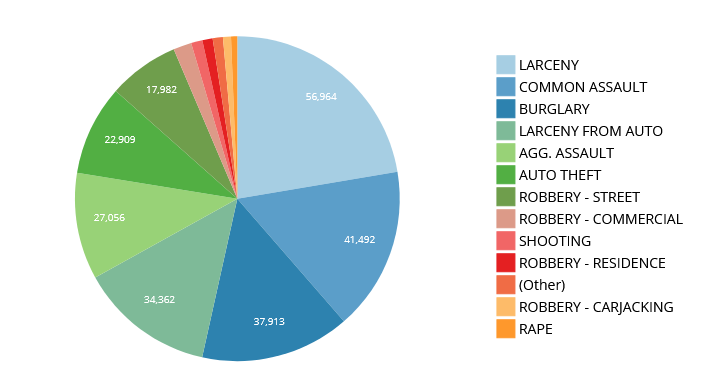
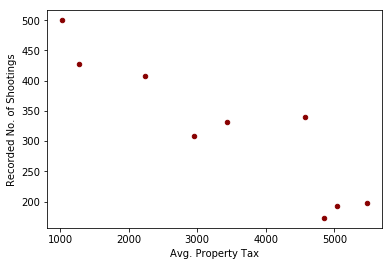
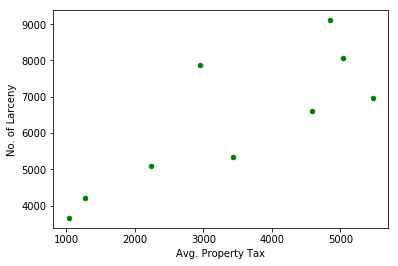


Figure 2.D: Amount of each recorded crime type from the Victim-based Crime data set.

If we break down the types of crime that the set reports [See: Figure 2.D], we see that the most common reported crime is larceny, which is a title which distinguishes what I presume to be petty theft. This is a crime which is far different from, say, shooting, both in general motive and severity. As a followup, I graphed the prevalence of these two crimes on more scatterplots to see if there was some difference that was hidden by the general case. [See: Figure 2.E-a and Figure 2.E-b].



Top Left- Figure 2.E-a: Scatterplot: “Avg. Property Tax compared to Larceny Incidents”. Corr = 0.89

Top Right- Figure 2.E-b: Scatterplot: “Avg. Property Tax compared to Shootings”. Corr = -0.92

One more, this was a fairly striking result. The graph of recorded shootings seemed to show a completely opposite trend from the general crime results. In that case, a lower property tax was also connected inversely to a larger number of reported shooting incidents for a given region, with a strong -0.92 correlation.

Additionally, the plot Figure 2.E-a is interesting, in that it shows a very similar looking set of points compared to the cumulative crime data set from Figure 2.C. However, the correlation is a lot stronger, which is quite strong evidence that larceny and its prevalence in the set over shootings were responsible for much of the positive correlation between property tax and incidents, as opposed to the inverse.

Interestingly, I noticed that the dot around the 5000 Avg. Property tax point moved the most upwards in comparison to the cumulative data graph. This is actually the data point for the Central Baltimore region, and if you think about it, it makes sense that larceny is a far more prevalent crime there. Central Baltimore is probably the most tourist-y region of the city, what with the inner harbor and its related attractions. It also has a lot of foot traffic and open areas, so I can definitely see it being a targeted place for pickpocketers and the like to prey on the tourists that visit the area.

All of these sets paint a very different picture of crime than I originally anticipated. It seems that while theft and larceny are committed more in highly taxed (and thus wealthier) areas, shootings and presumably other more violent crimes are committed more in lower taxed (poorer) neighborhoods. The significance of this is that crime, while stereotypically thought of as being restricted to poorer regions, actually just takes different forms in a dichotomy when looking at the differences between richer and poorer regions. This further emphasizes the difference between the poorer and wealthier parts of Baltimore city. And although I would not say that the residents in these places are necessarily more likely to commit one crime or the other, it is an important trend to notice in cases such as trying to predict where crimes of a certain type are likely to occur.

**C – Using linreg training to predict the better determination variable of Crime rates**

Finally, we’d like to be able to have some way to build a primitive model that could be used to predict likelihood of crime rates in an area based on the two attributes discussed previously: the tax rates, and the vacancy presence.

The reason for doing this has a variety of real-world applications: Even if just one incident is hard to predict the occurrence of, it is more than clear that trends emerge over time. Some areas are naturally more prone to crimes than others, for a variety of reasons. If a group were to develop an advanced training model, they could run simulations on areas that are likely to experience a certain level of some type of crime in the future. In theory, this could help in the field of crime prevention, by running scenarios with different variables.

In our case, though, the application is more for the sense of theory, and not for any actual applications. There are many reasons why our models fall very far of what could be used practically. Besides, there should realistically be better predictors of crime than just the housing tax rates and vacancy presence in an area.

But also realistically, one has to start somewhere. At the very least, we seek to answer which of these two attributes, given what we’ve come across so far, is able to best predict the crime data for that region.

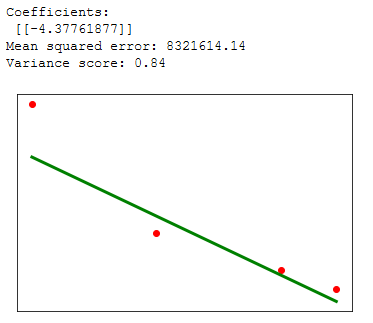
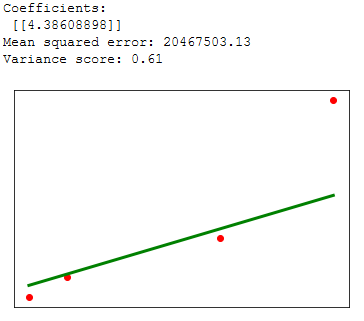
In a data science perspective, regardless of whether or not the correlation is positive or negative, the variance score is assigned to a linear regression set in terms of how well it can map the testing data. Here, our objective is to find the attribute which gives the model with a variance score closer to zero.

One of the first pitfalls of our training is that we don’t have a very large data set to train and test from. There are only nine regions of the city, so it’s very possible that the data ends up lopsided. In a more practical usage of this idea, we could have used neighborhoods or other smaller location data sets to agglutinate the data sets into more points.

As a result, the variance score can be wildly, well, variable, depending on which data points get picked for which purpose. Due to normal distribution, with enough points this doesn’t matter and the variance would be pretty consistent. But here it becomes a problem.

To be able to train, we created two different training data sets out of the comparison of property tax mean vs. crime rates per district, and another of the vacancy rates vs crime rates per district. By training each of these models with 5 out of the 9 regions for each, I obtained the following graphs showing the linear model in relation to the testing data. Note that these are not representative of what could be obtained by running this for yourself.

[See: Figure 2-F.a and Figure 2-F.b].



Top Left- Figure 2-F.a: “Linreg model of mean tax rates to predict crime rates in a region”. The x axis tracks mean property tax, while the y axis tracks crime rates. Top Right - Figure 2-F.b: “Linreg model of vacancy rates to predict crime rates in a region”. The x axis tracks vacancy rates, while the y axis tracks crime rates.

The data set, in the scope of the data analysis from earlier, comes back roughly as expected. Both property tax and residence vacancy rates share a fairly strong relationship with the crime rates in their area. It surprises me to find out that residence vacancy rates correlate so strongly with the crime rates in the area, given the variance score of 0.84.

Even though it was never a data column in the set, the number of instances in each region of Baltimore has returned a very promising result as to which variable to choose to accurately predict crime occurrences given a linear regression set.

Looking back on the previous sets, these models reaffirm their own usefulness through their consistency with the dichotomy presented from the data thus far. It is because the regions of Baltimore have such different levels of wealth and crime rate that these models are able to significantly predict the crime rate in any one region. And, like I said before, at first glance these statistics have little to do with one another. There is a definite trend that emerges, and a larger data set could almost certainly encompass the bigger picture.

**SECTION 3: CONCLUSIONS AND TAKE-AWAYS**

No doubt, there are significant correlations between the major three data points that we’ve tested. However, data is nothing without its effective applicability, so the ability to use these correlations to make predictions on something as overreaching as crime frequency is dramatic.

One important take-away from this is that it probably doesn’t help to be using such a limited data set as the nine districts of Baltimore, even if the properties of each are backed up by thousands and thousands of data entries for each cumulative attribute. For the purposes of training especially, it helps to either work the problem into such a way that the training happens before agglutinating the data, or start with a larger object data size, such as subdividing into the size of a neighborhood instead of 1/9th the entire size of Baltimore city.

What applications could these relationships have, given some more time and effort put into them? For one, the data itself could be important in enactment of government policy. If a region of the city has a higher-than-average presence of larceny and theft, then police officers might be deployed to patrol and scan the region more thoroughly. But, if we wanted to make smart decisions, then it’d definitely be in interest to analyze these trends that come through. Remember that first intuition isn’t always right. Crime such as larceny does happen more often in areas of higher income.

If the city government really wanted to reduce crime rates such as shootings, then they could take a start by improving programs that made housing more available for people who lack adequate dwelling. They could also enact programs to repurpose vacant, unused infrastructure for community projects. Although the matter could use more in-depth examination, there is evidence here to support a correlation between vacancies in a region and violent crime within that same region. Not all matters require one-dimensional brute force approaches, and it is the relationship between these data sets that strive to unearth the greater trends behind causality.

The goal of the training models should not necessarily be to reflect reality perfectly. Especially in the case of violent crime measurement, it should be in order to find the relationships between the variables, with the end goal of being able to find out what changes can be made to policy in order to eventually bring about positive change. Right now there seem to exist two different Baltimores. One has high crime incidence, the other has lower but more violent crime. One has low housing vacancy and higher property value and subsequent taxes, while the other is saddled with both high vacancy and low property values. In the practical application of these models, it would help to be able to run simulations on the data sets such that the effects of certain laws and policies that could be enacted could be estimated beforehand.

If we were to be able to have more time to go into the matter significantly, this would be the direction that I think we ought to take. Right now the linear regression sets really only reliably tell us the correlation that does exist between data sets, and by no means reliably or substantially. For one, the difference between shootings and larceny crimes was already distinguished in part 2B, so I think that running models differentiating the types of crime would be both more accurate and more conclusive in the scope of running full simulations. Now, that would definitely be difficult, but it could be more or less the direction that this project would go from here. Use multidimensional instances – possibly from data sets that more effectively represent the data sets than grouping by general region – to create simulations for crime events in the city of Baltimore. What we’ve accomplished with this report may not be fully conclusive, but it might accurately reflect the trends that could follow.

**ADDENDUMS**

This section includes any data sets or images placed “after-the-fact”, which would ruin formatting if included elsewhere.

* **1**

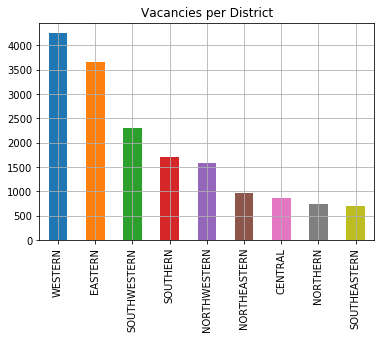


Figure ADD-1: Lists the total number of officially currently vacant homes within the city of Baltimore. Last updated Feb. 2019.