HOME VALUE PATROL

Relationship between crime occurrence and home values in Seattle

Team 3

Akashdeep Jaswal | Aparna Ramasamy Dharmakkan | Atinderpal Singh | Varuna Damodaran

ABSTRACT

Our project is an attempt to understand how property crime affects the home value or sale price of houses in a particular neighborhood over time. In order to understand this, we have retrieved data from two sources - Seattle 911 incident response data and Zillow Home Value Index data. Using data from these two sources, we performed exploratory data analysis to spot any initial anomalies that would drive our analysis process at the later stages. We then continued with an extensive data cleaning process in order to define a common point of comparison between the two datasets i.e. neighborhoods.

After data cleaning, we proceeded to fit a linear multiple regression model and used it to understand the relationship between Zillow Home Value Index (ZHVI) and five different types of property crimes in the city of Seattle from 2010 to 2015. The results of multiple linear regression model suggest that home value or property value has a negative linear relationship with crimes such as arrests, burglaries robberies in the neighborhood. Our results also indicate that the number of burglaries in a neighborhood does not have a significant negative relationship with the property values. Our model can be perfected by using several other variables which could be significantly altering the relationship between property values and crime; such as proximity to public transportation, proximity to schools and offices and access to other amenities like departmental stores, etc.

Introduction

Crime affects the society in several ways. It poses a large threat to the society. Crime imposes indirect costs to the society (Maximino, M." The impact of crime on property values: Research roundup", March, 2014). According to Pope (2011) "crime can be viewed as a neighborhood disamenity. One market that captures some of these neighborhood crime disamenities is the housing market." There have been several studies done in order to understand if there really exists a relationship between crimes and property values. In a 2012 report by Center for American Progress, police data from eight cities in the United States - Seattle, Milwaukee, Houston, Dallas, Boston, Philadelphia, Chicago and Jacksonville were analyzed to determine the impact of violent crimes on the property values in these cities. The results of the indicate that housing values increased by 0.83% for a 10% reduction in homicides. (Shapiro, R,J. Hassett, K,A, "The Economic Benefits of Reducing Violent Crime", June,2012). Federal Reserve Bank of New York released a report titled "Crime, House prices, and Inequality" which finds the relationship between crimes and house values in Latin America. It was found that decrease in crime was followed by a rise in the values of houses in Rio de Janeiro. (Seckan, B. "Crime, house prices and inequality: Examining Rio de Janeiro's favelas", February,2013).

Based on FBI's data, In 2014, there were an estimated 8,277,829 property crime offenses in the nation which resulted in losses estimated at $14.3 billion (Federal Bureau of Investigation, Uniform Crime Reports, 2014). The motivation for our study was to understand if sale prices and values of houses could reduce, if the area was known to be prone to criminal activities or on the contrary could the values of homes increase if the area was known to be safe from criminal activities. Our project is to explore the relationship between crime in Seattle and the property values in the city. In this paper the word "crime" includes violent as well as non violent crimes. Specifically, property crime comprises of Arrests, Assaults, Burglary, Disturbances and Robbery that have occurred in the city of Seattle from 2010 to 2015.

Research Question:

Our primary research questions are :

1. Is there any significant linear relationship between number of crimes in the neighborhoods of Seattle and the property values (home values) in Seattle?
2. Among the type of crimes (Arrests, Assaults, Burglary, Disturbances and Robbery) which had the most significant relationship with the home values in Seattle?

Datasets

We have used two datasets for our project. The first data we used was Seattle Police 911 incident response data set retrieved from https://data.seattle.gov/view/mzrk-e8qt and the second data set we used was the Zillow Home Value dataset retrieved from http://www.zillow.com/research/data/#median-home-value.

Seattle Police 911 response data gives information about the 911 calls received by the Seattle Police department, the crime classification i.e. burglary, murder or violent acts and the response times. This dataset contains information about 46 different types of 911 incident calls received from 2010 to 2015. It has a total of 1.14 millions records. This dataset comprises of 19 variables, but the most important ones for the consideration of this project are mentioned below:

- Date of the incident/event
- Incident classification
- General offense number of the event
- Incident location in terms of address
- Time taken to arrive at the scene
- Time taken to clear the scene
- Incident location in terms of latitude and longitude

Zillow home value data contains information about property values calculated by an index called the Zillow Home Value Index for a particular neighborhood for a specific date. This dataset is a time series data which has 5692 records with 88 columns. This dataset contains the following information about the following fields:

- Neighborhood
- City
- State
- Postal code
- Date
- Zillow Home Value Index (ZHVI)
- ZHVI yearly trends (82 columns)

Zillow Home Value Index (ZHVI) is calculated by Zillow by taking a set of home prices in a particular neighborhood in consideration with many other factors to finally arrive at the value benchmark. ZHVI is calculated based on another estimate of Zillow called the Zestimate which calculates the estimated sale price of home in a neighborhood. ZHVI is the median of the Zestimate values for a particular geographical location.

Data cleaning

Data cleaning was one of the most crucial steps in the progress of this project and, undoubtedly, the most challenging one as well. The two datasets being used were completely different in terms of their formats, so there was no way to initially align them for a comparative study. In order to reduce these datasets to a format that complements the requirement of our analysis, the following steps were followed:

1) Removing unwanted records and columns:

Functions used: read.csv(), subset()

Seattle's 911 Incident Response Dataset: The dataset, which was in comma separated values (.csv) format had to be imported into R. Our study required only a few properties (columns) of each call for analysis hence we retained only 4 columns (Event Clearance Group, Event Clearance Date, Longitude and Latitude) from the original 19 columns. The Date column also split to two columns one for year and one for month for ease of access. Each call is associated with the occurrence of an event which can be one of 46 different types (health related, violence, disturbances, etc.) Out of the different types we identified 5 events to be most relevant with our study: Arrests, Assaults, Disturbances, Burglary and Robbery and subsetted the dataset accordingly. NULL and NA records were also removed from the dataset.

Zillow Home Value Dataset: This dataset had data of the home values of different neighbourhoods from several cities in America which was expressed through columns such as Region name, City, Metro, State, County followed by several columns depicting the home value in each neighbourhood by year and month. Once the dataset (also in the .csv format) was converted into a data frame, a subset was created containing only Seattle's data because of the scope of our analysis. This made columns such as Metro, City, County and State unnecessary and hence they were removed.

2) Identifying Neighbourhoods for Seattle's 911 Incident Response Dataset:

Functions used: revgeocode(), readOGR(), proj4string(), spTransform(), point.in.poly()
Packages used: ggmap, rgdal. spatialEco, rgeos, maptools

The location from where each 911 call was placed is expressed as latitudes and longitudes in the 911 dataset, while the home values in the Zillow dataset were represented neighbourhood-wise. For our research experiment, we needed to find common grounds to merge the two datasets - Neighbourhood. The identification of neighborhoods from the latitudes and longitudes was one of the trickiest problems we faced We did extensive research to find different efficient ways to do this. To outline a few of them:

1) Reverse Geocoding - The ggmap library in R has a function: revgeocode() which finds the zip code by making API calls to google maps. However, there is an upper limit of 2500 API calls can be made in a day. Considering our dataset has over a million records, this wasn't feasible.

2) Using a 3rd dataset that maps zip codes to latitude longitudes - We were able to find a dataset that mapped latitude and longitudes to a zip code. However the latitude and longitude in our dataset didn't completely match with the new dataset and it seemed unreasonable to make approximations while dealing with latitudes and longitudes and hence we weren't able to proceed in this route.

3) Using Shapefiles - Shapefiles are often used by Geospatial information systems to describe geographic features such as rivers, elevations, waterfalls, etc through vectors such as points lines and polygons. Shapefiles representing Seattle's neighborhoods is publicly available at https://data.seattle.gov/dataset/Neighborhoods/2mbt-aqqx and can be imported into R using the readOGR() function. In this file, each neighborhood in Seattle is a represented as a polygon on the XY plane and map of Seattle can be plotted very accurately using these files. In order to map the latitudes and longitudes from the 911 calls dataset to these shape files, they had to be converted to a format that was consistent with that of the shapefiles which can be done using the functions proj4string() and spTransform() from the rgdal package. Once the 911 call dataset is in the same projection plane as the shapefile, the neighborhood from which the call is made was found by identifying the polygon (which represents the neighbourhood) in which the call location lies in performed using the function point.in.poly(). This merges the 911 call dataset with the shapefile by linking each call with a neighbourhood (L_HOOD) and a more accurate sub neighbourhood (S_HOOD).

3) Matching Neighbourhoods between datasets:

Functions used: which(), merge()

The neighbourhood data for the 911 dataset was obtained by using the shapefiles retrieved from Seattle's open repository of government data and didn't match perfectly with Zillow's data. Hence a lot of research had to be done in order to equate the neighbourhoods in both datasets in order to perform the most optimal inner join between them. This was a challenge because the neighborhood information in the 911 dataset was staggered across two different columns L_HOOD and S_HOOD where the former was a broader neighborhood and the latter represented a sub neighborhood. Hence an exhaustive study of the type of parameter best suited to join with the zillow data to obtain the best match was done. From our research, we deduced that S_HOOD column could be used effectively to make the join provided that the following changes were made:

1) Briarcliff, Lawton Park, Southeast Magnolia in the S_HOOD column of the 911 dataset was renamed/mapped to the bigger Neighborhood Magnolia.
2) Central Business District, International District, Pike-Market, Pioneer Square, Yesler Terrace, Interbay in S_HOOD column of the 911 dataset were renamed/mapped to the parent neighborhood Downtown.
3) Atlantic, Harrison/Denny-Blaine, Mann in the S_HOOD column of the 911 dataset were renamed Central since they belonged to that neighborhood.

Additionally, Mt Baker in the Zillow dataset was renamed Mount Baker to match the neighborhood retrieved in the 911 dataset.

4) Tidying Datasets:

Functions used: t(), gsub(), substr(),gather()
Packages used: stringr, dplyr, tidyr

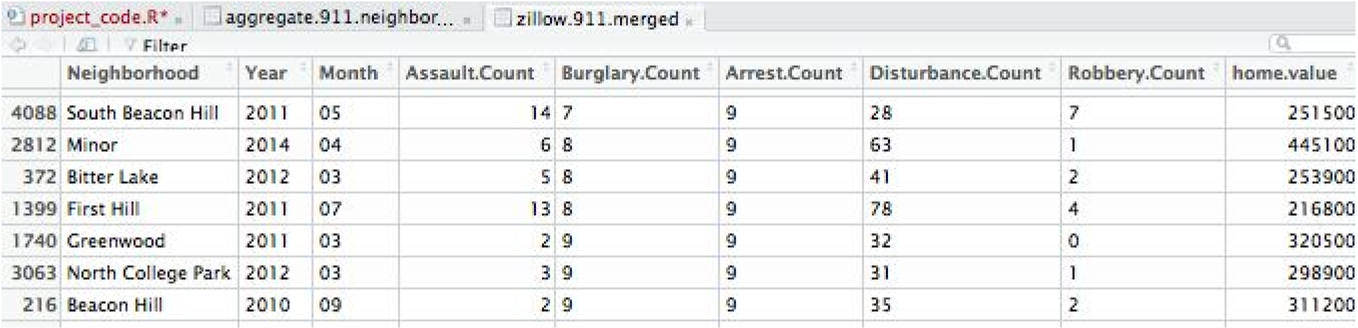
Seattle 911 Incident Response Dataset: The dataset now has Event Type, Year, Month, Neighborhood columns. However, in order to find out how each type of crime affects the home value, we need to run regression models on the different type of events (Burglary, Robbery, Assault, Disturbances and Arrests) implying that we need data which describes the home value against the different types of events. In order to transform our dataset to such a format, we used the technique outlined in the paper "Tidy Data" by Hadley Wickham (2014). We first created separate columns for each of the event types set to a default value of 0. And then based on the event type column we set the corresponding column value to 1. For instance if a particular call was to report a burglary then the burglary column was set to 1 and the remaining columns retained the default value of 0. By using this method, the resultant dataset was segregated based on event type using different columns and not just one event type column. These values were aggregated at a later stage by year, month and neighborhood.

Zillow Home Value Dataset: The Zillow data described the monthly home values in different columns and had data for each month starting 1996 to 2015. For our proposed analysis, we needed to represent home values of each neighbourhood for each month as different rows. Hence we once again used techniques outlined by Hadley Wickham in his paper. We transposed the data so that the data was now represented with neighborhoods as columns using the function t(). To create separate records for each month-year combination for each neighborhood, the function gather() was used. These transformations helped us convert the zillow dataset into a format that completed our proposed analysis methods. The resulting dataset was transformed from the dimensions 5692 x 88 to 20592 x 4.

5) Data Aggregation and Merging clean datasets

Functions: aggregate() , xtabs(), merge()

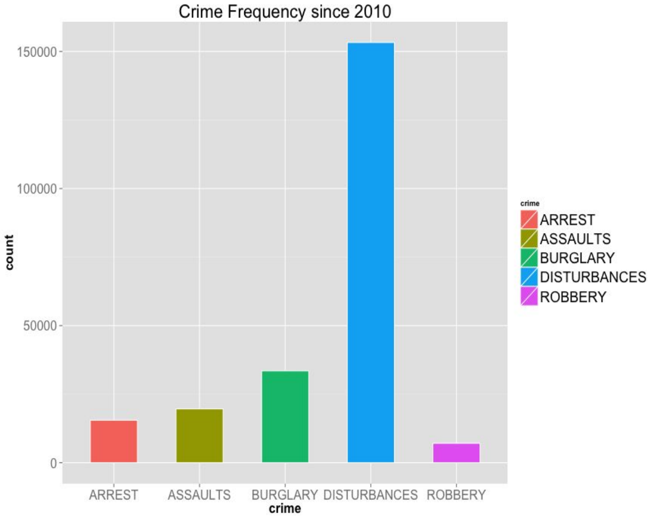
The next step was to aggregate the Seattle 911 incident response dataset by month, year and neighbourhood and we utilized the aggregate() and xtabs() functions.. Once aggregated, the data describe the total assaults, burglaries, disturbances, robberies and arrests in each neighbourhood for each month between years 2010 and 2015. Now that a common column existed between the Zillow and 911 incident response datasets, the were merged based on the common property: Neighbourhood using the merge() function. Here is a brief snapshot of our final dataset.

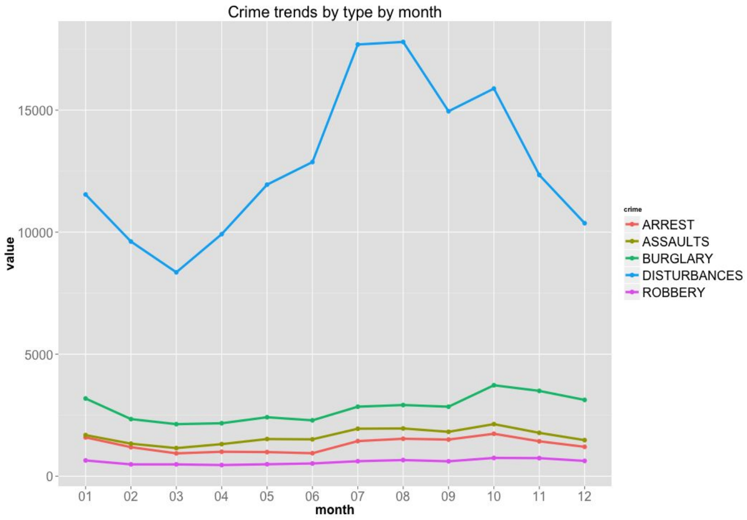


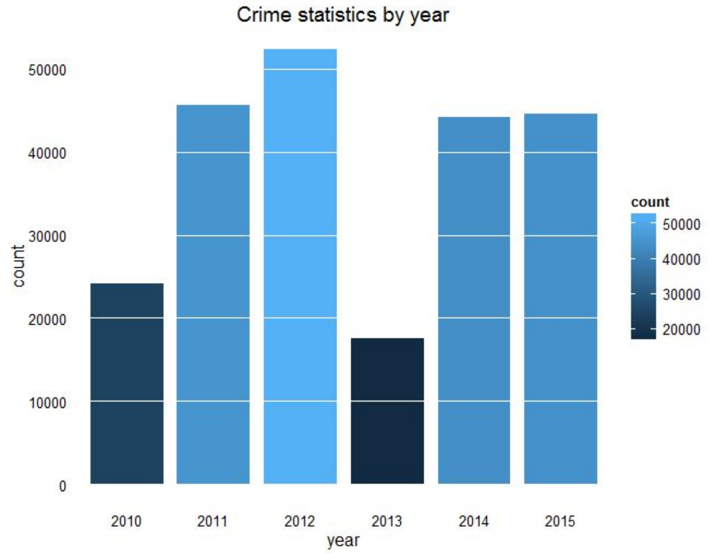
	Neighborhood	Year	Month	Assault.Count	Burglary.Count	Arrest.Count	Disturbance.Count	Robbery.Count	home.value
4088	South Beacon Hill	2011	05	14	7	9	28	7	251500
2812	Minor	2014	04	6	8	9	63	1	445100
372	Bitter Lake	2012	03	5	8	9	41	2	253900
1399	First Hill	2011	07	13	8	9	78	4	216800
1740	Greenwood	2011	03	2	9	9	32	0	320500
3063	North College Park	2012	03	3	9	9	31	1	298900
216	Beacon Hill	2010	09	2	9	9	35	2	311200

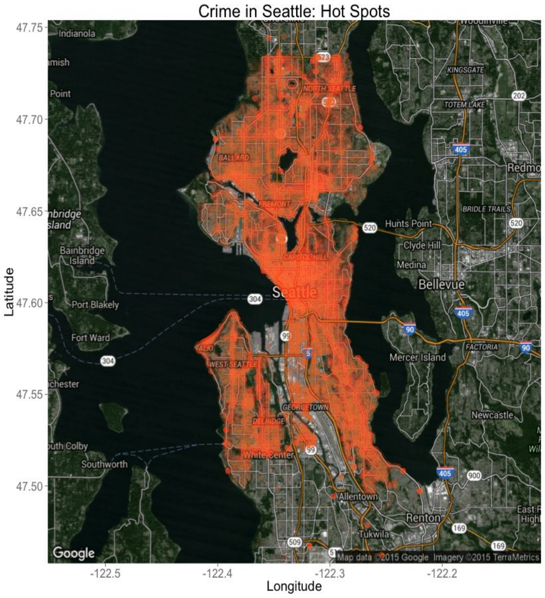
Exploratory Data Analysis

We initiated our data exploration process by studying the Seattle 911 incident response data for crime frequency, year / monthly trends, and crime hot spots by location.









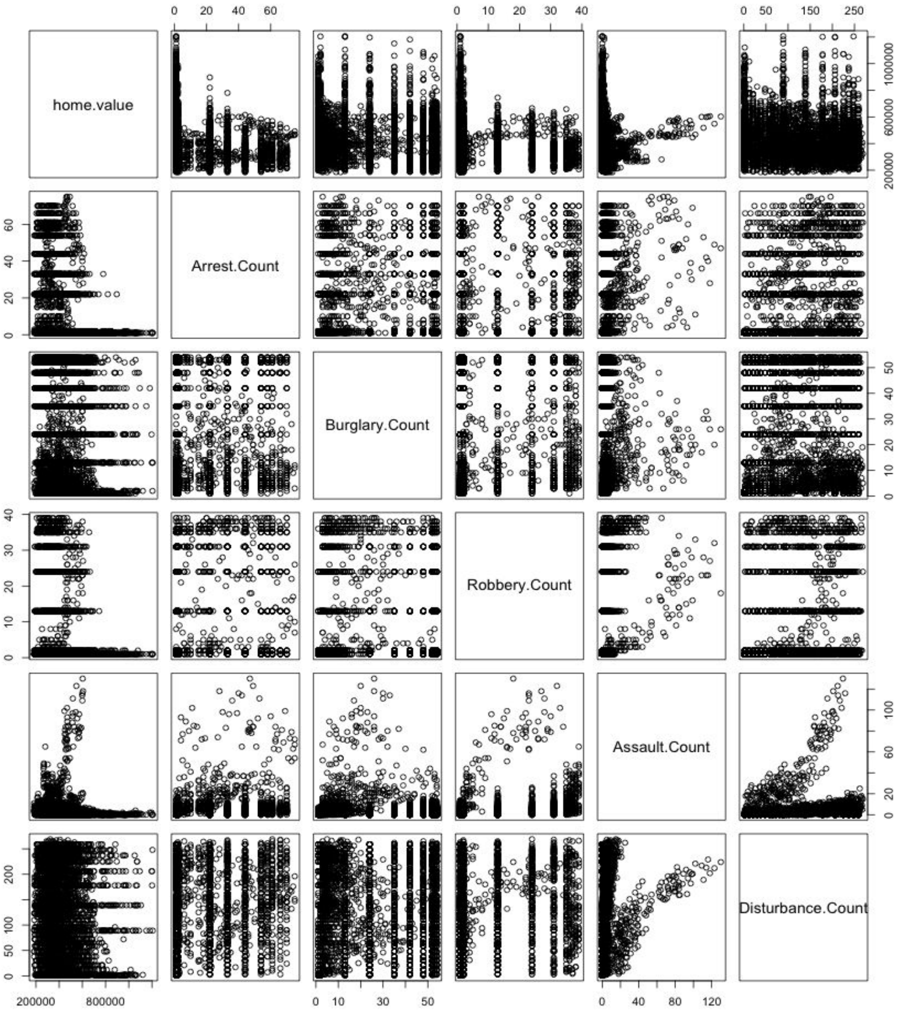
The above graphs show numbers for the types of property crimes (Arrests, Assaults, Burglary, Disturbances, Robbery) and their frequency in Seattle from the year 2010 to 2015. As can be seen disturbances are the most frequent type of crime,followed by burglary,assaults.arrests and robbery. From the second figure, we try to see if there is a monthly trend in any specific crime over the past few years. Again, we see most disturbances in the months of July and August (Summer vacation) -- it can be inferred that there are more disturbance complaints owing to parties being hosted in the holidays. There also seem to be lesser property related crimes in the years 2010 and 2013 while the other years do not have extreme variations. The crime hot spots from the map represent heavy volumes of crimes occurring in the central part of Seattle, while the other areas have evenly distributed values.

Methods

The final dataset after cleaning and aggregation contains the following columns :

- Neighborhood
- Year
- Month
- Assault count (Number of assaults in a particular neighborhood in a particular month for a given year)
- Burglary count (Number of burglaries in a particular neighborhood in a particular month for a given year)
- Arrest count (Number of arrests in a particular neighborhood in a particular month for a given year)
- Disturbance count (Number of disturbances in a neighborhood in a particular month for a given year)
- Robbery count (Number of robberies in a particular neighborhood in a particular month for a given year)
- Home value in a particular neighborhood in a particular month in $$.

To test our variables for collinearity we plotted them together; it satisfied all 4 assumptions for linear modelling.



We then created a multiple linear regression model in R using Home value as the outcome for a function of all the five types of crime (Assault,Burglary,Arrest,Disturbance and Robbery count) as the predictors.

$$home.value = \beta_0 + \beta_1 * Assault.Count + \beta_2 * Arrest.Count + \beta_3 * Burglary.Count + \beta_4 * Disturbance.Count + \beta_5 * Robbery.Count$$

We determined from our results that there are two variables deviating from the expected negative effect - Disturbances count and Assaults count had a positive relationship with the home value index. We ran a multiple linear regression again without these confounding variables.

Results

Here are our results from multiple linear regression:

Multiple R-squared: 0.1106

	Estimate	**Pr (>|t|)**	**95% Confidence Interval**
(Intercept)	469028.28	< 2e-16	461599.3, 476457.2
Assault.Count	1401.19	5.18e-09	931.7, 1870.6
Arrest.Count	-2278.03	< 2e-16	-2531.6, -2024.4
Burglary.Count	-114.53	0.2973	-329.9, 100.87
Disturbance.Count	56.38	0.0273	6.3, 106.4
Robbery.Count	-2413.51	< 2e-16	-2876.40, -1950.6

Model Interpretation:

Intercept:

- The intercept is 469028.28 which indicates that when all the types of crimes have 0 count, the home value in Seattle expected to be $469,028.28.
- The adjusted multiple R squared value for the regression is 0.1106 which indicates how well the linear model explains the variations in the data. The closer the value of adjusted R squared to 1,the better the

model fits the data. Our adjusted R squared value is less, which indicates that there may be other variables which have a significant role in explaining the variations in the data.

Assaults:

- The slope value is 1401.19 which indicates that every 1 count increase in the number of assaults the home value increases by $1401.19, holding the values of other types of crimes constant.
- Suppose we took many samples and built a confidence interval from each sample then 95% of these intervals would contain the actual value of the slope in them and would lie between 931.7 and 1870.6.
- The p value is < 0.05 which indicates that the results are statistically significant.
- Though the results are statistically significant, we believe they are not practically significant since they seem to indicate that home values increase when there is increase in the number of assaults in the city. To further explain and understand the cause behind this, we found that 4471 of total 17523 assault cases reported in Downtown Seattle (25% of the total) -- Since the volume is large, this was affecting the overall model largely. And we assume that the value in Seattle Downtown are always on the rise, assaults do not affect them.

Arrests:

- The slope value is -2278.03 which indicates that every 1 count increase in the number of arrests the home value decreases by $2278.03, holding the values of other types of crimes constant.
- Suppose we took many samples and built a confidence interval from each sample then 95% of these intervals would contain the actual value of the slope in them and would lie between the -2531.6 and -2024.4.
- The p value is < 0.05 which indicates that the results are statistically significant.

Burglary:

- The slope value is -114.53 which indicates that every 1 count increase in the number of burglaries the home value decreases by $114.53, holding the values of other types of crimes constant.
- Suppose we took many samples and built a confidence interval from each sample then 95% of these intervals would contain the actual value of the slope in them and would lie between 329.9 and -100.87.
- The p value is 0.2973 which is greater than 0.05 which indicates that the results are not statistically significant. This can be explained by the fact that a burglary is one of the most common crime types -- The victim does not have to be present during the crime.This makes it less egregious and hence it affects the home value with lesser magnitude.

Disturbance:

- The slope value is 56.38 which indicates that every 1 count increase in the number of disturbances the home value increases by $56.38, holding the values of other types of crimes constant.
- Suppose we took many samples and built a confidence interval from each sample then 95% of these intervals would contain the actual value of the slope in them and would lie between 6.3 and 106.4.
- The p value is < 0.05 which indicates that the results are statistically significant.

★ Though the results are statistically significant,we believe they are not practically significant since they seem to indicate that home values increase when there is increase in the number of disturbances in the city. This can be explained by the fact that the values are evenly distributed through all neighborhoods.

Robbery:

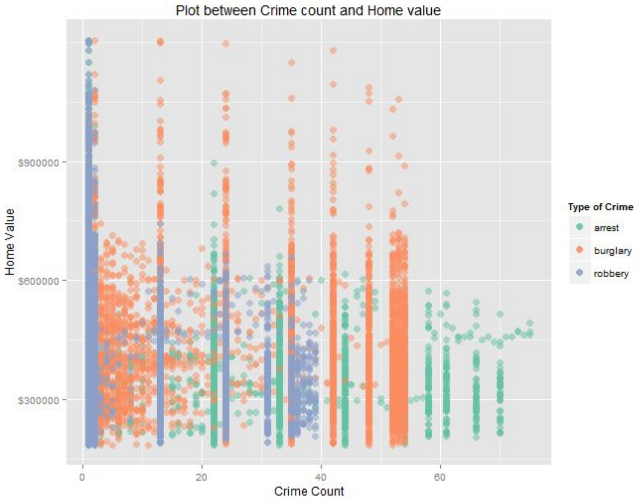
★ The slope value is -2413.51 which indicates that every 1 count increase in the number of robberies the home value decreases by $2413.51, holding the values of other types of crimes constant.

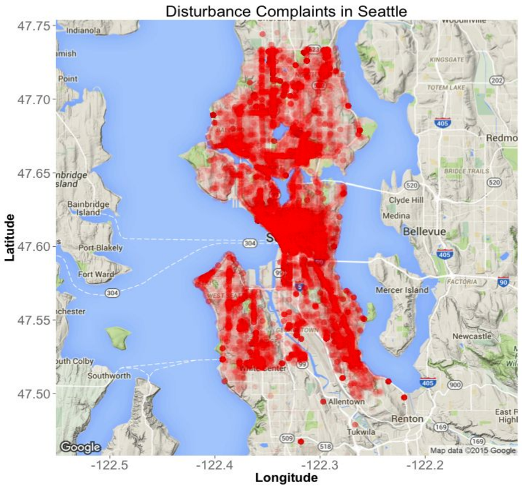
★ Suppose we took many samples and built a confidence interval from each sample then 95% of these intervals would contain the actual value of the slope in them and would lie between -2876.40 and -1950.6.

★ The p value is < 0.05 which indicates that the results are statistically significant.

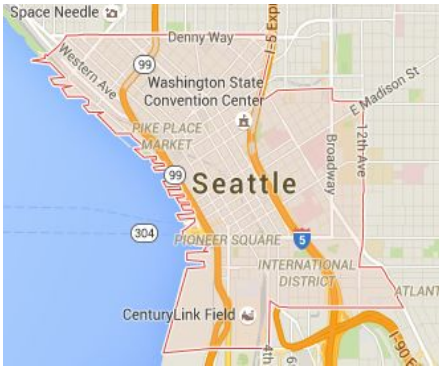
Crime Type	**Home Value Effect**
For every 1 Assault Count	**$ 1401.19**
For every 1 Arrest Count	**$ 2278.03**
For every 1 Burglary Count	**$ 114.53**
For every 1 Disturbance Count	**$ 56.38**
For every 1 Robbery Count	**$ 2413.51**

The above table summarises the slope value results from the multiple regression. The green colored slope values for disturbances and assaults indicate that there is a positive relationship between the type of crime and the home value ie, they cause an increase in the values of homes in a given neighborhood based on our study. The red colored slope values for arrests, burglary and robbery indicate that there is a negative relationship between the type of crime and the home value, ie the increased number of these crimes cause the home values to go down.





The above plots show the relationship between home value and the three types of crimes which have a negative relationship with home value - arrests,burglary and robbery. Also, the heat map indicates that disturbance counts are more or less evenly distributed. They do not have a significant relationship with the neighborhood. This could be a possible explanation for why disturbance count had a slightly positive relationship with home value since they are common. We also assumed that some regions have higher disturbances because there are more parties organized there like Capitol Hill, Lower Queen Anne -- People want to live in these areas, so they have higher home values.



To explain why assaults have a positive relationship with crimes, we found that 4471 of total 17523 assault cases reported in Downtown Seattle (25% of the total) -- Since the volume is large, this was affecting the overall model largely. And we assume that the value in Seattle Downtown are always on the rise, assaults do not affect them.

Conclusions

The results of our project confirms to some extent, with our belief, that home values or property values do have a negative relationship with crimes such as the number of arrests, burglaries and robberies in the city of Seattle. Through the course of the project we also discovered some interesting facts: You are less likely to be arrested if you stay in an expensive house and that Seattlers like to Party! We believe that the scope of our project can be scaled and our linear model can be made more accurate by including other crucial variables for which have practically significant relationship with the home values in the city. Such crucial variables include:

- ➔ Proximity to public transportation
- ➔ Proximity to workplace location
- ➔ Proximity to schools and educational institutions such as University of Washington
- ➔ Proximity to departmental stores,shopping malls, hospitals.
- ➔ History of the neighborhood
- ➔ Past data about the property values in the city.

References

1. Maximino,M. (March 12, 2014). The impact of crime on property values: Research roundup - Retrieved from http://journalistsresource.org/studies/economics/real-estate/the-impact-of-crime-on-property-values-research-roundup.

2. Pope, J,C. Pope, D,G. (August 26,2011) .Crime and property values: Evidence from the 1990s crime drop- Retrieved from http://faculty.chicagobooth.edu/devin.pope/research/pdf/Website_Crime_Property.pdf

3. Shapiro, R,J. Hassett, K,A. (June 19,2012). The Economic Benefits of Reducing Violent Crime - A Case Study of 8 American Cities - Retrieved from https://www.americanprogress.org/issues/economy/report/2012/06/19/11755/the-economic-benefits-of-reducing-violent-crime/

4. Seckan,B. (February 1,2013). Crime, house prices and inequality: Examining Rio de Janeiro's favelas - Retrieved from http://journalistsresource.org/studies/international/development/crime-house-prices-inequality-upps-rio-favela

5. Hadley Wickham (December 9, 2014) Journal of Statistical Software, vol. 59, 2014. Retrieved from http://www.jstatsoft.org/article/view/v059i10

Appendix:

The project was done using R language and the code can be found at in our Github Repository - https://github.com/akashjaswal/home-value-patrol