Treyce Kangas-Simon

Jean-Marc Romain

Georgette Powley

Elias Pennell

NTEC365 46182

Team Research Document

For this group project, Team Brainy Bunch set out to use the core concepts learned in Big Data and Analytics to analyze data involving house prices in Portland, OR and reported crimes in the surrounding neighborhoods. The ultimate outcome is to determine if the price of a listed property is at all related to the amount of reported crime near the listed property. The project will employ the techniques of Python automated web scraping in conjunction with traditional data analysis and correlation calculation.

Team Brainy Bunch’s web scraping project is proceeding quite quickly and smoothly with only a few minor technical issues. The program, as currently designed, has been successful at gathering information from realtor.com and writing it into a .csv file. The main technical difficult experienced thus far was getting the program to loop to read multiple web pages worth of data in a single running. At the time of writing, realtor.com is hosting 85 pages worth of listing with an average of 42 listings per page. When first written, the program was able to retrieve data from realtor.com and write it into a .csv file, but only with a single page at a time. The acquired data is in the following format:

Address,Price,Beds,Baths,Sizes

"1150 NW Quimby St Unit 906, Portland, OR 97209","$1,039,900",2,3,"1,612sqft"

"2939 SE Beyee Blvd, Portland, OR 97202","$1,675,000",5,3,"4,592sqft"

"8420 NW Whitney St, Portland, OR 97231","$659,000",3,2,"1,680sqft"

After tweaking, the scraping program can read through all 85 pages of available listings. This translates to a total of 3,570 listings. The program took approximately 15 minutes to run while scraping realtor.com and these listings were stored in a .csv file.

Crime data was gathered from a publicly posted resource hosted by the Portland Police Bureau. Initially, we tried to locate crime data for Portland, OR from https://data.gov/ but this resource did not host crime data for Portland, OR. Luckily, the Portland Police Bureau has publicly accessible crime data that can downloaded as a .csv file. This allowed Team Brainy Bunch to access crime data quickly and easily for YTD in 2022. The data officially covers January 2022 through September 2022, as PPB states their posted data is always a month behind from when it is accessed. The most salient data in the collected .csv file displays the neighborhood the crime is committed in, the date of the crime, and the type of offense.

After data had been collected and stored, the monumental task of cleaning said data was undertaken. The main issue with the data from realtor.com was how much extraneous information was gathered. For example, the web scraping program downloaded the number of bedrooms and bathrooms the listed house had, which isn’t necessary information to create a comparison to levels of crime in the area. Some of the data cleaning had to be done manually in order to assign neighborhoods to the listed houses. An example is the Buckman neighborhood. Technically, the Buckman neighborhood is divided into both an East and a West Buckman, however, realtor.com listed the majority of the houses in this area simply as the Buckman neighborhood. Entries such as these had to be cleaned and clarified due to how Portland Police Bureau lists their crime data. Due to privacy concerns, the data doesn’t list an exact address but simply broken down into the neighborhood the crime is committed in. Using a combination of the Portland Maps website and a located Portland Police Bureau article that shows the breakdown of their patrolled neighborhoods, it was determined which neighborhood several of these addresses belonged in.

Some difficult decisions had to be made with the data, as well. There were a fair number of listings that represented condominiums, apartments, or some form of sub home. It was determined that including any data entries that had some form of “sub-address,” e.g. 123 Fake St. Apt A, should be removed from the data used for analysis. The rational behind this decision is that counting every sub home artificially inflated the amount of homes listed in the area and could skew the data. If there was an entire condominium worth of sub homes for sale in a neighborhood the collected data showed every sub home listed at that address. Furthermore, any entries for mobile homes for sale or any land only sales were removed from the list. The prices for these entries were far below other entries in the same neighborhood and it was believing that these outliers would skew the mean house pricing too far down. Finally, any address that had no confirmed street numbers or no assigned address, such as in cases were the buildings are too new to be officially mapped by the city yet, were removed.

Analysis of the data has begun and information such as count of listed houses, mean price, standard deviation, minimum price, 25% of price, 50% of price, 75% of price, and maximum price has been calculated for each neighborhood. The next steps to take will be to calculate the number of crimes per neighborhood and do a Pearson correlation calculation to determine if the number of crimes per neighborhood impacts the price of housing in any significant manner.

The Brainy Bunch had also considered mapping the housing and crime data but due to the interest of time and not having readily available geographical coordinates for the crimes committed, it was determined that it would be best to shelve that endeavor for the time being.

Text

Description automatically generated

References:

<https://www.realtor.com/realestateandhomes-search/Portland_OR>

<https://www.portlandoregon.gov/police/71978>

<https://www.portlandoregon.gov/police/article/32642>

<https://portlandmaps.com/>