Zhuoying Cai, Weicheng Huang, Ka Shing Wai
4/29/2019

Interim Project Update

"One thing that would be great to see more of is the larger context of why you want to model housing prices. What is it you hope to find? Why is that important? The better the case you can make for the purpose of your analysis, the more compelling it will be, and the better able you will be to stand apart from all of the other examples of similar analysis that exist out there using the same data set."

Feedback Reflection

From our housing price analysis, we hope to find if a house is worth to purchase, either for living or investing purposes. Nowadays, housing price in some boroughs of New York tends to be overpriced. It is significant for people to see if a house price is reasonable since it usually takes a lot of money to buy a house. Moreover, we are going to use a combination of more recent and thorough datasets, which other analysis does not contain, to analyses and build our prediction model. Our goal for this project is to analyze and predict the housing price of a given neighborhood. In addition, instead of predicting the housing price based on the basic criteria, such as size, the number of bedroom and bathroom, our analysis will use factors such as transportation and neighborhood information that may affect the housing price to make the prediction.

Data Exploration Summary

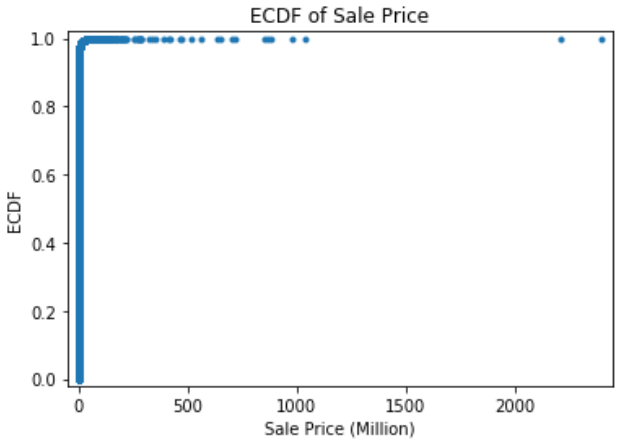
For now, we have explored *Annualized Sales Update* and *Rolling Sales Data* datasets. *Annualized Sales Updata* dataset contain yearly sales information of properties sold in New York City from 2003 to 2017. Rolling sales Data contain the same type of data from April 2018 to March 2019. We decided to use data from the last 24 months because the gap in housing price range between 2003 and 2019 is relatively large. These data can provide more relevant and applicable information for us to analyze how the recent housing price changed and predict the

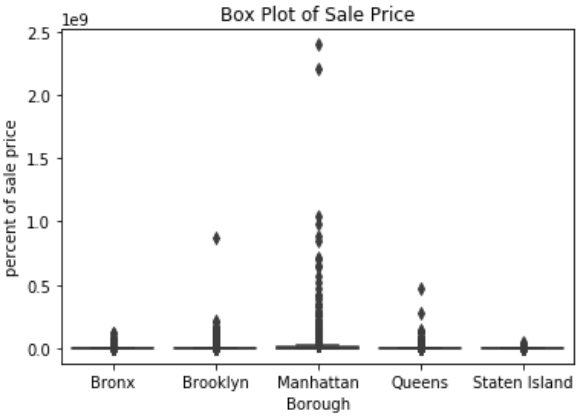
housing price. We have also explored the PLUTO (The Primary Land Use Tax Lot Output) dataset and New York City Neighborhood Data Profiles. Both of them contain land use, community information and geographic data. We discovered that these data are not coherent. Such data will not have much effect on the analysis of the housing price fluctuations over time. Since these two datasets contain relatively similar data, we decide to use PLUTO dataset after we finishing analyzes the housing price fluctuations. We will put the housing price data on the NYC map and add PLUTO dataset on it as supplementary information of a neighborhood.

After loading the last 24 months of data from *Annualized Sales Update* and *Rolling Sales Data* datasets, we have discovered below:

1. We inspected that there are many missing data in the data frame. We see that the data frame has 170407 entries, but *TAX CLASS AT PRESENT, EASE-MENT, BUILDING CLASS AT PRESENT, APARTMENT NUMBER, ZIP CODE, RESIDENTIAL UNITS, COMMERCIAL UNITS, TOTAL UNITS, LAND SQUARE FEET, GROSS SQUARE FEET* and *YEAR BUILT* has fewer entries than that.
2. *BOROUGH* is represented as numerical values rather than categorical. We changed the *BOROUGH* values to the actual borough name.
3. Many data for *SALE PRICE* are zero or missing. As stated in the dataset documentation, a $0 sale indicates that there was a transfer of ownership without a cash consideration [1][2]. Therefore, we decided to remove all data that contains *SALE PRICE* is zero. At the same time, we removed all data that contains missing *GROSS SQUARE FEET* and *YEAR BUILT* to ensure every data has features that we need for prediction.
4. After removing data that contains *SALE PRICE* is zero, there are still many *SALE PRICE* at unreasonably low values, which also likely to be transferred properties.
5. Whether in Manhattan or the Bronx, there exist extremely expensive housing prices. The reason is that either a unit itself is expensive, or a whole building is being traded. We do not have interests in such transactions. Therefore, we decided to limit the housing price between one hundred thousand to three million, which fall within the price range that most people interested in, in the data frame.

Chart





We first plot the ECDF and box plot of the uncleaned data. The ECDF graph on the left shows that some housing prices are extremely expensive. The highest housing price in the ECDF graph reaches up to two billion dollars. From the box plot on the right, we notice that there are many outliers occur in Manhattan. However, outliers also exist in other boroughs.

The three graphs below show the data after cleaning. From the graphs, the cleaning result is desired and satisfiable. We can conclude that:

1. Most of the housing prices are below one million.
2. The housing price in Manhattan is the most diverse. However, the Bronx and Staten Island rarely have housing prices for over one million.
3. Ordering by turnover rate, Manhattan >> Brooklyn > Queens > Bronx > Staten Island

Above graphs conclude our preliminary data cleanup. Our ultimate goal is to analyze the housing price in a neighborhood, but not in a whole borough. However, this will depend on the amount of data we have. Since these data will be divided into all neighborhoods, the data in each neighborhood may not be sufficient to support our analysis.

Reference

1. NYC Department of Finance*, Annualized Sales Update,* Data and Lot information, source: https://www1.nyc.gov/site/finance/taxes/property-annualized-sales-update.page
2. NYC Department of Finance, *Rolling Sales Data,* Data and Lot information, source: https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page
3. PLUTO, the City's Department of Buildings database, source: https://www1.nyc.gov/site/planning/data-maps/open-data.page