**Finding a suitable property base on the venue of neighborhood and real estate values**

Marrugo Nicolas, April 16 of 2019

**1. Introduction**

**1.1 Background**

Nowadays, according with the 2018 homebuyer report driven by Nerdwallet, discovered that 75% of North Americans still say that have a house is a top priority, but approximately just 15% of the people reported had purchased a home in the past five years, and 32% intend to do so in the next half-decade.

Despite the fact this percentage seems very low, according with CBRE Econometric Advisors since 2017, the U.S homeownership rate is stable around 63.7%, after some years of recovering from the drop of the global financial crisis of 2007-2008, that caused a decreasing in the rate of at least 10 %, having reports of homeownership below of 60%.

Taking into account, that the number of people buying a house per year is quite stable, lead to suppose that the number of renter-occupied households will be stable, but actually is continuing to grow in most markets, driven by population growth and household formation. Where cities like Chicago and Detroit will be actually stable, while cities like New York City and Boston will be weak growth in the renter-occupied households.

The reasons why the market have this behavior is basically due to the lack of inventory of houses, limited construction of new houses, construction of luxury properties, having zones with rules that restrict a higher density development or simple that investors keep the new acquisitions as rentals. This lead to have a real challenge in order to find a suitable house that fulfill the desires of the client in terms of location, price and neighborhood.

**1.2 Problem**

Taking into account the previous statements, is possible determine that from the point of view of buyers a new client with the intention of buy a house, the first challenge to affront is actually finding a suitable home to move to, in terms of price per square foot and the location (venues near of the house). This project aims to help the client to find a suitable house for buying, analyzing the venues of each neighborhood and the price per square foot for the city of New York, giving as a result a list of neighborhoods with its corresponding square foot value and promising price behavior which accomplish the client preferences.

**1.3 Interest**

In that order of ideas, this project will fulfill the necessities of mid-range and late-range aged persons that are looking for a new or used house to buy in New York city, considering the relation between square foot price and location, and its corresponding behavior in the real estate market.

**2. Data acquisition and cleaning**

**2.1 Data sources**

For developing this project is necessary real state data for all the neighborhoods in New York, this information is provided by the New York Times listings database and the New York City government records, which provide data from the last year, the last 3 months and the current day squared foot price for 286 neighborhoods. The information gather is given by an analysis of price and sales trends as well as many other metrics to give readers an idea of current conditions for this market and nearby markets, as well as historical market trends.

On the other hand, the venues information is obtained by the foursquare database will give the 200 venues near to the location of each neighborhood in a radius of 500 meters. Finally, the location information is given by the New York (City), Department of City Planning, where has a total of 5 boroughs and 306 neighborhoods, each neighborhood that exist in each borough have the latitude and longitude coordinates.

**2.2 Data cleaning**

The location data gathered the New York (City), Department of City Planning is in json format, from which is necessary take the parameters of borough, neighborhood, latitude, longitude, and then this data is converted into a pandas dataframe.

The real estate data had to be collected by neighborhood from the New York Times listings database and the New York City government records, due to some of the neighborhoods in the real estate data are separated in the location data, was necessary a rearranged of the data in order to have a price per square foot for each neighborhood in the 306 of the location dataset.

Then the real estate data was cleaning from Not a Number values and then was normalize using the min-max function. Finally, the venues data obtained from the foursquare database was gather in Json format, from where the name, latitude, longitude and category of the venue were extracted and converted into a pandas dataframe.

Once this data was extracted, the venues for each neighborhood was grouped in order to find the proportion that each category is present in each neighborhood. In this way is possible determine the 10 most common venues categories per each neighborhood.

**2.3 Feature selection**

After data cleaning, there were 10,722 venues for 306 neighborhoods, were 432 unique categories were identified, from these categories just 10 were chosen for each neighborhood, which represent the 10 most repetitive category on each neighborhood. Additional to this data the normalized value of the price per square foot today was added for the cluster features. Given the following result:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Borough** | **Neighborhood** | **price nowNormal** | **1st Most Common Venue** | **2nd Most Common Venue** | **3rd Most Common Venue** |
| Bronx | Allerton | 0.012022 | Pizza Place | Supermarket | Deli / Bodega |
| Bronx | Baychester | 0.042995 | Pizza Place | Breakfast Spot | Mexican Restaurant |
| Bronx | Bedford Park | 0.017294 | Pizza Place | Mexican Restaurant | Chinese Restaurant |
| Bronx | Belmont | 0.046172 | Italian Restaurant | Pizza Place | Deli / Bodega |
| Bronx | Bronxdale | 0.0000001 | Italian Restaurant | Chinese Restaurant | Performing Arts Venue |

Table 1. Features Selected for the clustering algorithm.

The features matrix for the clustering algorithm was set as 11 columns that represent the 10 most common categories for each neighborhood and the 306 rows represent each neighborhood in New York