Investigation of house sale price and neighborhood venues¶

Zhi Yuan
December 16, 2019

1. INTRODUCTION

1.1 Background

New York City (NYC) is known as one of the most expensive city in the world in terms of real estate prices. It is also one of the most active real estate market in the world. Not only New Yorkers are trying to buy their dream houses, NYC also attracts investors all over the world to their real estate market. In a year from September 2016 to September 2017, there are over 80 thousand real estate sale transactions with a total amount of over 30 billion US Dollars. Predicting house sale price is an important problem for not only each property owner and potential property buyers, but also for investors and investment companies, as real estate has long become an investment instrument. For example, real estate investment trust (REIT) has currently attracted over 1.7 trillion US Dollars around the world.

1.2 Problem

There are many factors that are related to the sold house price, such as building age, number of units in the building, etc. This project will be focused on other relating factors from the neighborhood venue categories. More specifically, we will investigate what kinds of venues in a NYC neighborhood have correlation with the neighborhood property price. In order to explore the neighborhood venues, we will use the database of FourSquare. Based on the most common venue categories that appears in a FourSquare search of a neighborhood, we will investigate into which venue categories may be correlated with the real estate property price in the neighborhood. Our study will be focused on Manhattan, the most expensive borough of NYC. However, we believe the studied insight and methodologies used could also be used in other part of the world.

2. Data

In this project, we will use two sources of data: one for house sale prices, and one for the neighborhood venue exploration.

2.1 House sale price dataset from Kaggle

The first data set is the Kaggle dataset of NYC Property Sales: https://www.kaggle.com/new-york-city/nyc-property-sales. This dataset contains properties sold in New York City over a 12-month period from September 2016 to September 2017, together with their borough and neighborhood information, among others. We will use this dataset to gain insights into the house sale price in each neighborhood. Our focus is on one of the five boroughs of NYC, Manhattan. We extract all the Manhattan house sale records from the data set. There are over 18 thousand house sale transactions in Manhattan with a total transaction amount of 48 billion US Dollars.

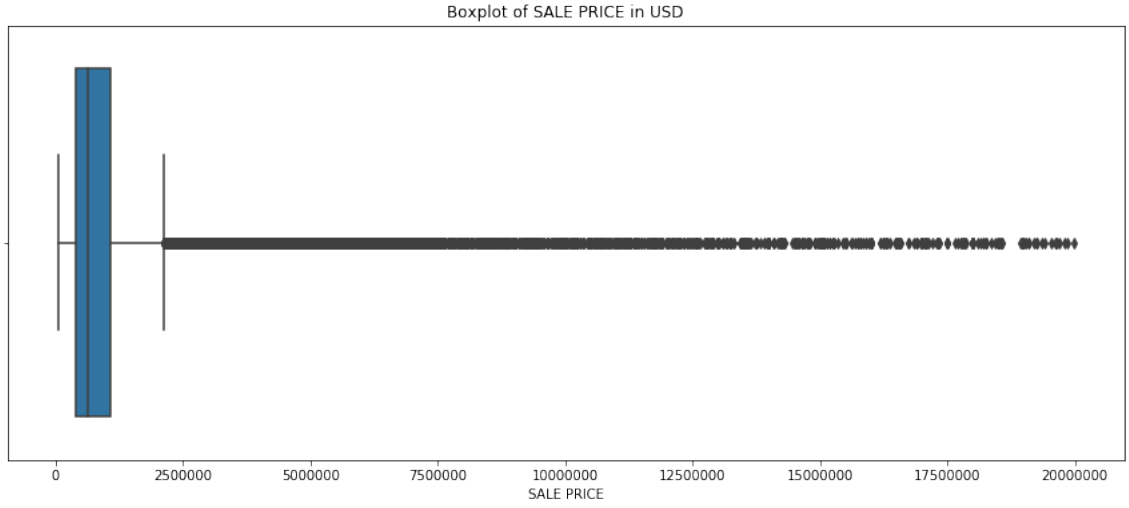
2.2 Neighborhood venue data from Foursquare

To explore the venues in a neighborhood in Manhattan, we have used the Foursquare API to search its database. As Foursquare requires a latitude and longitude to explore, and the Kaggle dataset does not provide the latitude and longitude, we also use the geocoder package with the Nominatim API to get the latitude and longitude of each listed neighborhood in Manhattan. As the Geocoding API does not recognize many of the Manhattan neighborhood in the Kaggle dataset, a lot of data cleansing and wrangling work has to be done before the data can be used.

3. Methodology

The Kaggle dataset of NYC Property Sales requires a lot of data cleansing work. Firstly, the records with missing sale price value need to be discarded.

Many sales occur with a nonsensically small dollar amount: $0 most commonly. These sales are actually transfers of deeds between parties: for example, parents transferring ownership to their home to a child after moving out for retirement. Such extreme low prices should be discarded. We have considered in our project only the house sale prices between 50,000 to 20,000,000 USD.



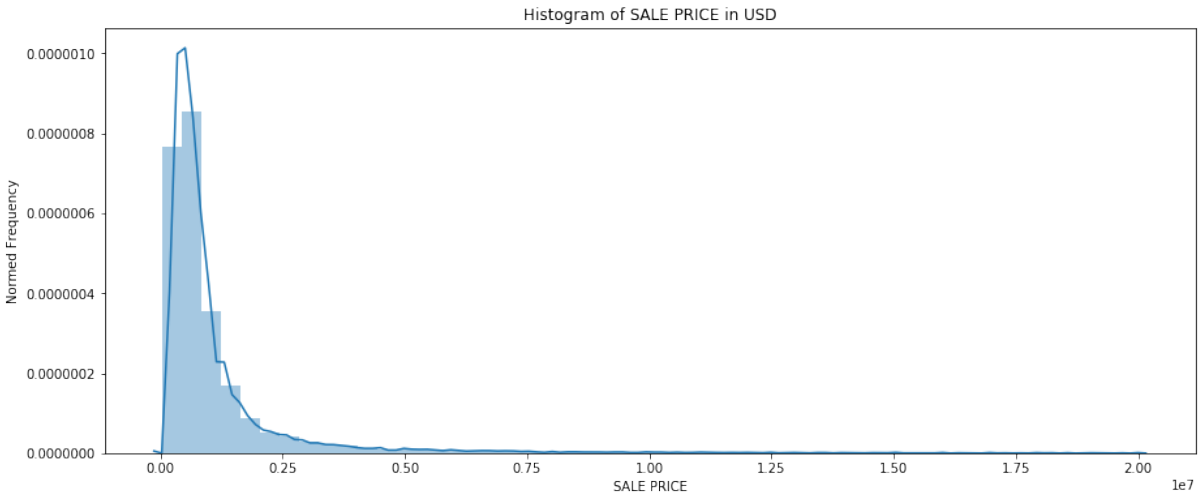


Figure 1: Distribution of NYC house sale price in USD

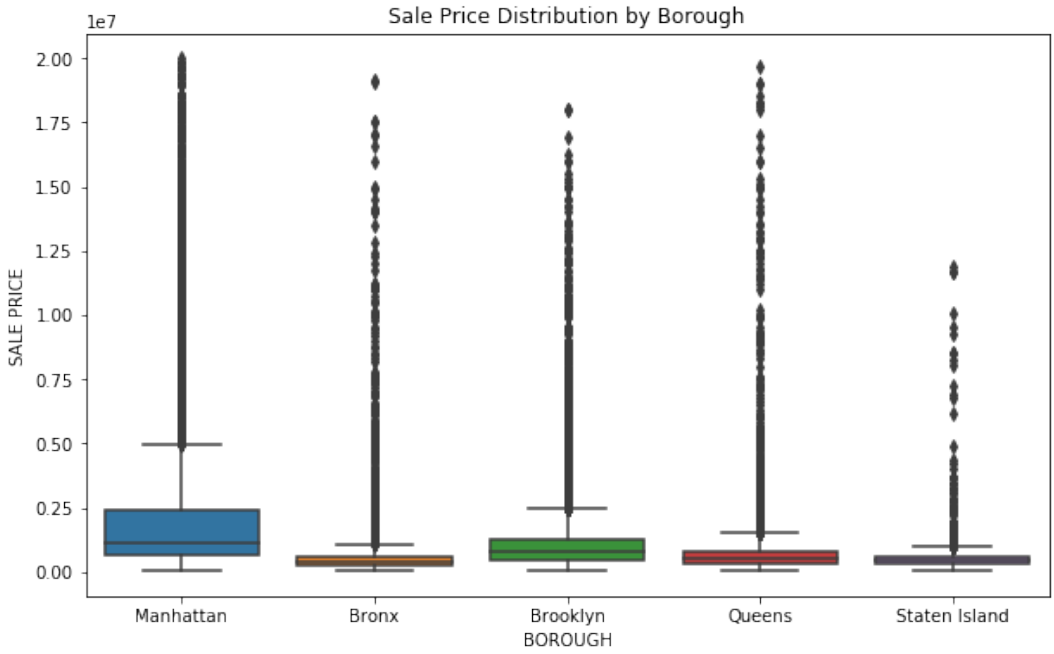


Figure 2: NYC house sale price distribution by boroughs.

Figure 2 shows the NYC house sale price distribution by boroughs. During the year from September 2016 to September 2017, there are 84,548 house sale transactions with a total amount of 89,335,360,909 USD in NYC, where there are 18,306 house sale transactions with a total amount of 48,196,678,399 USD in Manhattan. We will focus this study in the Manhattan borough.

There are a total of 39 neighborhoods in Manhattan listed in the Kaggle dataset. However, some of the neighborhood names are not recognized by the Nominatim geocoder, and we have to merge some unrecognized neighborhood, such as Harlem-East and Harlem-West into Harlem, etc. It reduces the number of neighborhoods to 30. The median sale price in each of the 30 neighborhoods are listed in Table 1. The distribution of the sale price by neighborhoods are shown in the boxplot in Figure 3.

NEIGHBORHOOD	SALE PRICE
ALPHABET CITY	1014940
CHELSEA	1382500
CHINATOWN	1751425
CIVIC CENTER	5083613
CLINTON	1140000
EAST VILLAGE	1050000
FASHION	1692500
FINANCIAL	1461188
FLATIRON	2435000
GRAMERCY	880000
GREENWICH VILLAGE	1360000
HARLEM	875000
INWOOD	428160.5
JAVITS	830000
KIPS BAY	1112643.5
LITTLE ITALY	3750000
LOWER EAST SIDE	822500
MANHATTAN VALLEY	951000
MIDTOWN	1321343
MIDTOWN EAST	865500
MIDTOWN WEST	1100000
MORNINGSIDE HEIGHTS	585000
MURRAY HILL	835000
ROOSEVELT ISLAND	862837
SOHO	2800000
SOUTHBRIDGE	999000
TRIBECA	2950000
UPPER EAST SIDE	1153372
UPPER WEST SIDE	1200000
WASHINGTON HEIGHTS	550000

Table 1: Median house sale price by 30 neighborhoods in Manhattan

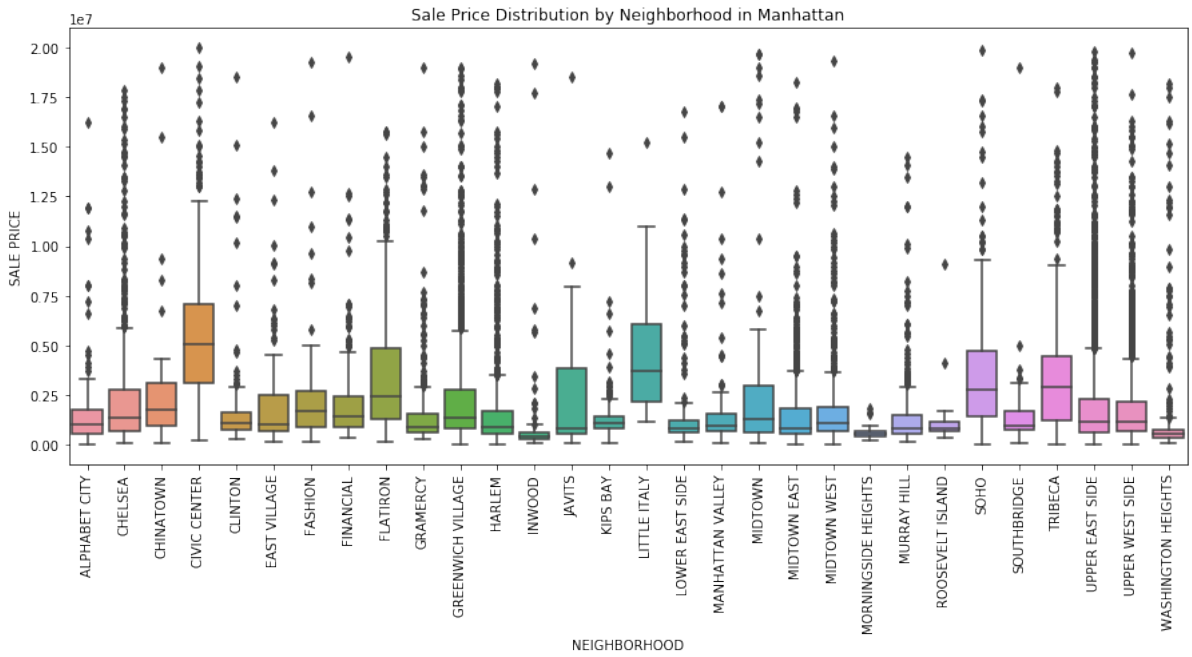


Figure 3: Distribution of property sale price in Manhattan by neighborhoods.

We have chosen the median instead of the mean value, as median value is more robust in terms of high price values, and thus more relevant to larger proportion of property buyers. We have also displayed the 30 neighborhoods of Manhattan in the interactive map by folium as follows. The name and the median property price can be shown once clicked.

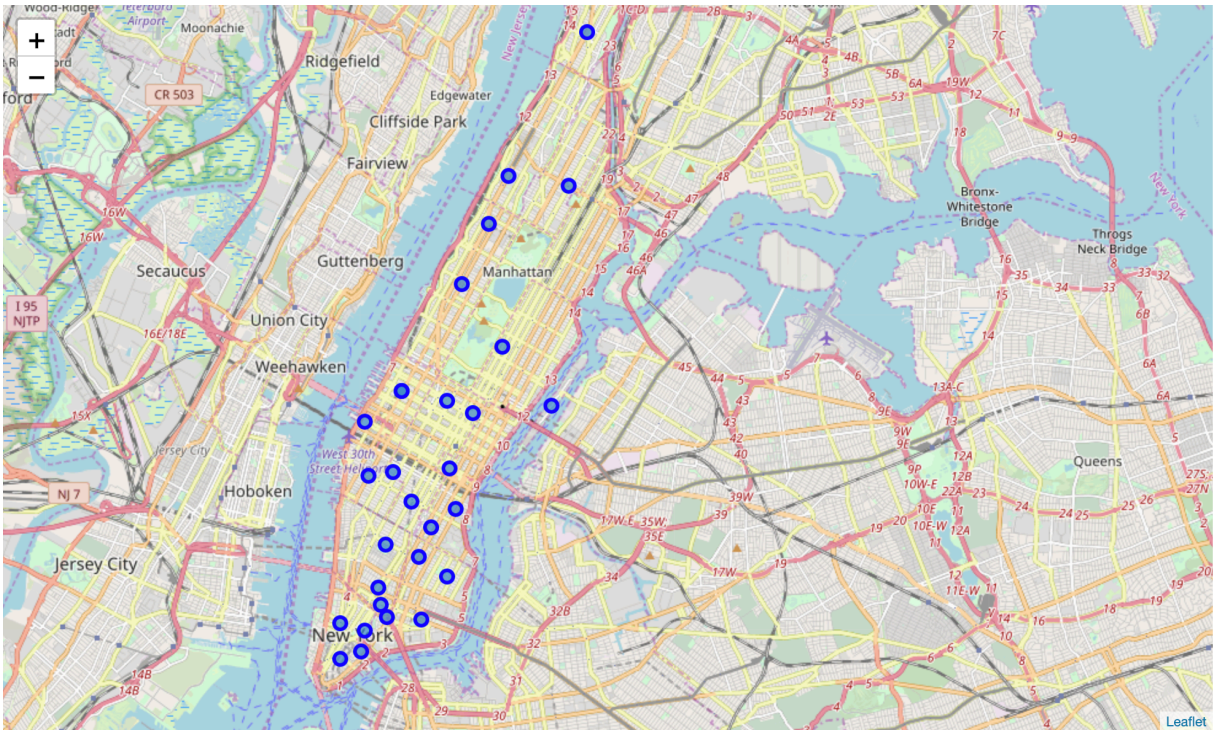


Figure 4: Map of Manhattan with neighborhoods.

In order to explore each neighborhood of Manhattan, we use the API provided by Foursquare. The URL to access the Foursquare API can be the following:

https://api.foursquare.com/v2/venues/**explore**?client_id=**CLIENT_ID**&client_secret=**CLIENT_SECRET**&ll=**LATITUDE**,**LONGITUDE**&radius= **RADIUS**&v=**VERSION**&limit=**LIMIT**

We need to register as developer at Foursquare to get a client ID and client secret as well as a version number of the API. Then we need to tell Foursquare

- The geographic location to explore in terms of latitude and longitude;
- The radius to explore;
- The limit of returned vanues.

A free user is limited to 100 search results per query.

For each found venue, we will calculate the percentage of each venue category in an neighborhood. There are over 300 different venue categories classified by Foursquare. Examples include gym, monument, museums, various types of restaurants, bars, entertainment facilities, etc.

We will investigate the correlation between the density of each venue category with the house sale price. To do so, we will use the pearson's correlation coefficient. The Pearson Correlation measures the linear dependence between two variables X and Y, in our case, the density of a venue category with the house sale price.

4. Results

We investigate the correlation between the density of each venue category with the house sale price using the Pearson's correlation coefficient. The resulting coefficient is a value between -1 and 1 inclusive, where:

- 1: Total positive linear correlation.
- 0: No linear correlation, the two variables most likely do not affect each other.
- -1: Total negative linear correlation.

The P-value is the probability value that the correlation between these two variables is statistically significant. Normally, we choose a significance level of 0.05, which means that we are 95% confident that the correlation between the variables is significant.

By convention, when the

- p-value is < 0.001: we say there is strong evidence that the correlation is significant.
- the p-value is < 0.05: there is moderate evidence that the correlation is significant.
- the p-value is < 0.1: there is weak evidence that the correlation is significant.
- the p-value is > 0.1: there is no evidence that the correlation is significant.

We have found by the Pearson's statistics the following 14 venue categories that have significant correlation (with p-value <0.05) to the neighborhood house sale price.

venue_category	pearson_coef	p_value
Dim Sum Restaurant	0.603476	0.000415
Optical Shop	0.568231	0.001054
Salon / Barbershop	0.511324	0.003879
Martial Arts Dojo	0.453967	0.01174
Falafel Restaurant	0.446707	0.013336
Deli / Bodega	-0.43502	0.016284

Furniture / Home Store	0.420752	0.020598
Arts & Crafts Store	0.41243	0.023521
Dance Studio	0.410312	0.024317
Dessert Shop	0.407086	0.025571
Chinese Restaurant	0.39062	0.032825
Women's Store	0.379428	0.038644
Dog Run	-0.376383	0.040363
Pizza Place	-0.369133	0.044703

11 of them are positively correlated, including dim sum restaurant, optical shop, salon / barbershop, martial arts dojo, falafel restaurant, furniture / home store, arts & crafts store, dance studio, dessert shops, Chinese restaurant, women's store. 3 of them are negatively correlated, including Deli / Bodega, dog run, and pizza place.

5. Discussion

There are 14 venue categories identified in the previous section to be correlated with the neighborhood house sale prices. These venue categories can be useful features for building predictive models for house sale price prediction.

In this project, we use venue category density features from Foursquare to investigate their correlations with the neighborhood house sale price. In fact, these features can also be used to predict sale price of individual real estate property, given the address of the individual property, instead of the price of the neighborhood. However, we did not pursue that idea in this work, because the connection to geocoder with Nominatim is very unstable, and we got too often a service timeout. This technical difficulty has unfortunately hindered the collection of venue category features for each individual house price prediction.

6. Conclusion

In this project, we have analyzed the Kaggle dataset of one year's house sale data, used venue category density features from Foursquare, and investigated their correlations with the neighborhood house sale price. We have identified 14 venue categories that have significant correlation with the house sale price.