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.

1.1 House sale price dataset from Kaggle

The first data set is the Laggle 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.

1.2 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 google geocoding 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.