IBM – Coursera

Data Science Specialization

Capstone project - Final report

# Correlation between a neighborhood real estate price and its surrounding venues

Gilles Clausse – 2021

# Table of content:

# I. Introduction:

This report is for the final course of the Data Science Specialization. A 9-courses series created by IBM, hosted on Coursera platform. The problem and the analysis approach are left for the learner to decide, with a requirement of leveraging the Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve.

The main goal will be exploring the neighborhoods of New York city in order to extract the correlation between the real estate value and its surrounding venues.

The idea comes from the process of a normal family finding a place to stay after moving to another city. It's common that the owners or agents advertise their properties are closed to some kinds of venues like supermarkets, restaurants or coffee shops, etc.; showing the "convenience" of the location in order to raise their house's value.

So, can the surrounding venues affect the price of a house? If so, what types of venues have the most affect, both positively and negatively?

The target audience for this report are:

- Potential buyers who can roughly estimate the value of a house based on the surrounding venues and the average price.
- Real estate makers and planners who can decide what kind of venues to put around their products to maximize selling price.
- Houses sellers who can optimize their advertisements.

## II. Data description:

New York city neighborhoods were chosen as the observation target due to the following reasons:

- The availability of real estate prices. Though very limited.
- The diversity of prices between neighborhoods. For example, a 2-bedrooms condo in Central Park West, Upper West Side can cost $4.91 million on average; while in Inwood, Upper Manhattan, just 30 minutes away, it's only $498 thousands.
- The availability of geo data which can be used to visualize the dataset onto a map.

The type of real estate to be considered is 2-bedroom condo, which is common for most normal nuclear families.

The dataset will be composed from the following two main sources:

- CityRealty which provides the neighborhoods average prices. https://www.cityrealty.com/nyc/market-insight/features/get-to-know/average-nyc-condo-prices-neighborhood-june-2018/18804
- FourSquare API which provides the surrounding venues of a given coordinates.

The process of collecting and clean data:

- Scrap the CityRealty webpage for a list of New York city neighborhoods and their corresponding 2-bedroom condo average price.
- Find the geographic data of the neighborhoods. Both their center coordinates and their border.
- For each neighborhood, pass the obtained coordinates to FourSquare API. The “explore” endpoint will return a list of surrounding venues in a pre-defined radius.
- Count the occurrence of each venue type in a neighborhood. Then apply one hot encoding to turn each venue type into a column with their occurrence as the value.
- Standardize the average price by removing the mean and scaling to unit variance.

The result dataset is a 2 dimensions data frame (Figure 1):

- Each row represents a neighborhood.
- Each column, except the last one, is the occurrence of a venue type. The last column will be the standardized average price.

| | Neighborhood | Accessories Store | Adult Boutique | African Restaurant | American Restaurant | Animal Shelter | Antiq Shop | ... | Whisky Bar | Wine Bar | Wine Shop | Wings Joint | Women's Store | Yoga Studio | StandardizedAvgPrice |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Battery Park City | 0 | 0 | 0 | 3 | 0 | 0 | | 0 | 1 | 4 | 0 | 1 | 0 | -1.303912 |
| 1 | Bedford-Stuyvesant | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 1 | 6 | 0 | 0 | 1 | -0.418350 |
| 2 | Boerum Hill | 0 | 0 | 0 | 1 | 0 | 0 | | 0 | 0 | 2 | 0 | 0 | 2 | 0.015011 |
| 3 | Brooklyn Heights | 0 | 0 | 0 | 2 | 0 | 0 | | 0 | 1 | 4 | 0 | 0 | 5 | -1.099479 |
| 4 | Bushwick | 0 | 0 | 0 | 1 | 0 | 0 | | 0 | 0 | 1 | 0 | 0 | 2 | -0.587926 |

*Figure 1 - Final dataset*

The dataset has 50 samples and more than 300 features. The number of features may vary for different runs due to FourSquare API may returns different recommended venues at different points in time.

The number of features is much bigger than the number of samples. This will cause problem for the analysis process. Detail and counter-measurement will be discussed further in the next section.