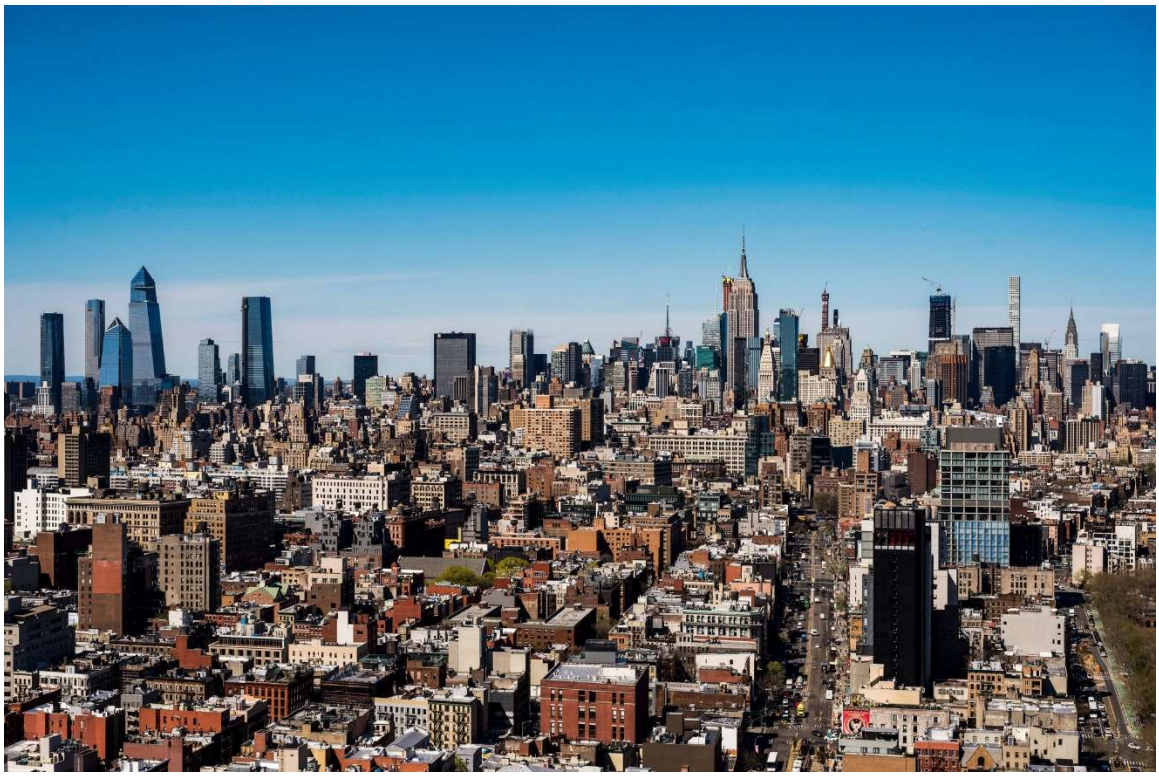

New York City Real State Analysis

COURSERA CAPSTONE PROJECT

Javier A. Jaime-Serrano | IBM Applied Data Science Capstone | August 20, 2020

Part 1: Introduction

BACKGROUND

New York City (NYC) is the most populated city int the United States (US), with an estimated population of more than 8 million distributed over more than 300 square miles.

It is also the largest metropolitan area in the US with more than 20 million people, and is composed of five boroughs: Brooklyn, Queens, Manhattan, the Bronx, and Staten Island.

Real estate is a major part in the city economy, as the total value of all NYU passed the $1 trillion mark in thew 2017 fiscal year with an increase of more than 10% over the previous year. NYC is home to some of the world's most valuable real state. [1]

BUSINESS PROBLEM

An investment on real state in NYC will come with some challenges, and any potential investor will need to be assured of the value of the investments. This value will come from selecting the best neighborhood for a target amount to be invested.

The best neighborhood will depend mainly of the proximity to the venues preferred by the potential investor. Other factors like crime statistics were not considered, assumed not relevant to the higher market segments analyzed in this report.

The question is: How to select the best neighborhood for a given investment amount and the venues preferences of a potential investor.

[1] New York City (https://en.wikipedia.org/wiki/New_York_City)

Part 2: Data Acquisition

DATA SOURCES

In order to segment the neighborhoods of NYC and explore them, we need the first dataset. This dataset was extracted from the NYU Spatial Data Repository [2] and contains the 5 boroughs and the 306 neighborhoods as week as the latitude and longitude coordinates of each neighborhoods. This dataset was previously used in the course week 3 lab Segmenting and Clustering Neighborhoods in New York City.

The second dataset was extracted from Zillow Research Data [3] and contains a time series of the Zillow Home Value Index (ZHVI), this is a smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type.

The dataset selected from Zillow Research is for the mid-tier condo/coops (typical value in US dollars for homes that fall within 33rd to 67th percentile range for a given region). The mid-tier condo/coop ZHVI is assumed to be representative of the average home values of a highly dense region with a low percentage of single-family homes.

The third data set was extracted with the Foursquare Developer Application Programming Interface (API) [4] and includes the venues, categories and location data of the neighborhoods requested by the API within a given radius. This dataset was previously used in the curse week 2 lab Foursquare API.

DATA CLEANING

Dataset 1 (from NYU) was extracted first in a json file, and filtered to the features key that contains all relevant data, this data was transformed to a pandas dataframe using loop script in Python (Programming language), see figure 1 for the first five rows of the dataframe.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Figure 1

Dataset 2 (from Zillow) was extracted first in a Comma Separated Values (CSV) file and transferred directly to a dataframe (read csv method) and filtered to show NYC only, not required columns were dropped and remaining columns were renamed keeping the

neighborhood name as the key index, see figure 2 for the first 5 rows and the last 10 columns of the dataframe.

Neighborhood	2019-09	2019-10	2019-11	2019-12	2020-01	2020-02	2020-03	2020-04	2020-05	2020-06
Upper West Side	1233923.0	1226329.0	1228089.0	1234873.0	1232891.0	1228788.0	1218444.0	1218807.0	1214683.0	1210707.0
Upper East Side	929593.0	927350.0	926257.0	927430.0	927683.0	929749.0	925260.0	925438.0	925215.0	933518.0
East New York	341233.0	340846.0	340971.0	340998.0	342076.0	343255.0	344385.0	345332.0	345130.0	345802.0
Washington Heights	566566.0	560402.0	556665.0	552960.0	548861.0	544782.0	541350.0	538371.0	533715.0	531615.0
Astoria	513852.0	514333.0	513360.0	514066.0	514298.0	514380.0	513222.0	513961.0	514368.0	515600.0

Figure 2

Dataset 3 (from Foursquare) was extracted first in a json file, flattened (json normalize method), filtered to the venues and their categories, and keeping the location data, see figure 3 for the first five rows of the dataframe.

	name	categories	lat	lng
0	The Bar Room at Temple Court	Hotel Bar	40.711448	-74.006802
1	The Beekman, A Thompson Hotel	Hotel	40.711173	-74.006702
2	Alba Dry Cleaner & Tailor	Laundry Service	40.711434	-74.006272
3	Gibney Dance Center Downtown	Dance Studio	40.713923	-74.005661
4	The Class by Taryn Toomey	Gym / Fitness Center	40.712753	-74.008734

Figure 3

FEATURE SELECTION

After cleaning the data, we ended with 3 datasets: the first with 306 neighborhoods and the required location data, the second with 190 neighborhoods and ZHVI 296 months' time series and the third with 10,058 venues, neighborhoods, categories and location data.

The field in common in the 3 datasets is the neighborhood and was selected as the key index to link the 3 datasets. The number of venues and neighborhoods will need to be reduced after the exploratory data analysis.

[2] NYU Spatial Data Repository (https://geo.nyu.edu/catalog/nyu_2451_34572)

[3] Zillow Research Data (https://www.zillow.com/research/data/)

[4] Foursquare Developer (https://developer.foursquare.com/)]