Coursera Capstone Project

for

IBM Data Science Professional Certificate

“The Choice for Housing”

Analysis of preferred apartments in New York City

By: Joseph Oisamoje

December, 2018

I. Description of the problem:

1. Introduction:

Relocating to a big city anywhere in the world is no small endeavor and for a middle aged family this could be very daunting. New York City the big apple as she's fondly called is one of the premier cities in the world and is known as the finance center and a tourist hot-spot in North America and as such the real estate cost are some of the priciest. A quick shows that real estate price can vary by a large margin from neighborhood to neighborhood. As an example, a 2-bedrooms condo in Central Park West, Upper West Side can cost up to $4.91 million on average; while in Inwood, Upper Manhattan, just 30 minutes away cost about $498 thousands.

What features of a neighborhood affects the price of real estate to such extend? One hypothesis is that the surrounding venues can be a decision factor.
Surely anyone, who has attempted to find an accommodation for rent or buy, has seen advertisements such as: This condo is located near the subway station, malls, supermarkets, dinners, etc. And it's likely that the price will be higher than others with locations not as "convenient".
Can the venues surrounding an accommodation effect its price? And what kind of venues can affect the most? And by what margin?

2. The question to answer:

This project will try to explore the neighborhoods of New York City to see:

- If the surrounding areas/landmarks can affect the price of real estate.
- The kind of surrounding areas/landmarks, and to what extent affects the price.
- Can we use the surrounding areas/landmarks to estimate the value of an accommodation over the average price of one area? And to what degree of accuracy?

The result can be useful for home buyers/renters, who can roughly estimate the value of a target house over the average in a given area. It can also be of help to realtors as valuable data that shows where projects can be cited for long-term revenue/profitability

II. Description of the data:

The main data used for this project will be from two sources:

- The average price by neighborhoods in New York City culled from a real estate portal. (CityRealty)
- The venues in each neighborhood from location services. (FourSquare API)

*Note: This project will only consider the average price of 2-bedrooms apartments, which is a common type of real estate among normal families.

1. Data collection process:

- The average price will be scrapped from the CityRealty website.
- For each neighborhood I'll use the Geocoder Python library to get its coordinate.
- For each neighborhood's coordinate, I'll use the FourSquare API to get the surrounding venues.
- Count the occurrences of each venue type and attach that information to each neighborhood.

The output of the data collecting process will be a 2 dimensional table:

- Each row represents a neighborhood.
- Each column will be the count of one type of venue in that neighborhood.
- The last column will be the average 2-bedroom apartment price of that neighborhood.

2. Using data to solve the question:

- First the correlation between price and surrounding areas (if any) would be analyzed.
- Second, if correlation exits, machine learning techniques (PCA, Regression, PCR) will be used to analyze the data. The output will be a list of venues types that effect the most on price, along with their weight on the result.