IBM CAPSTONE

REAL ESTATE PRICING - NEIGHBORHOOD

SEPTEMBER 2019

INTRODUCTION

The purpose of this project is to show the difference that the price of a property can have from one neighborhood to another neighborhood.

DATA

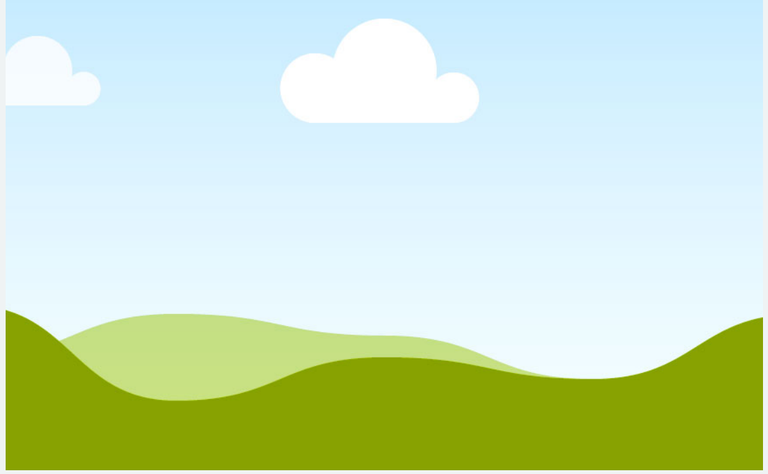
For this I used data from

- Foursquare API
- NYC Condos Average Price per Neighborhood from End of Last Year

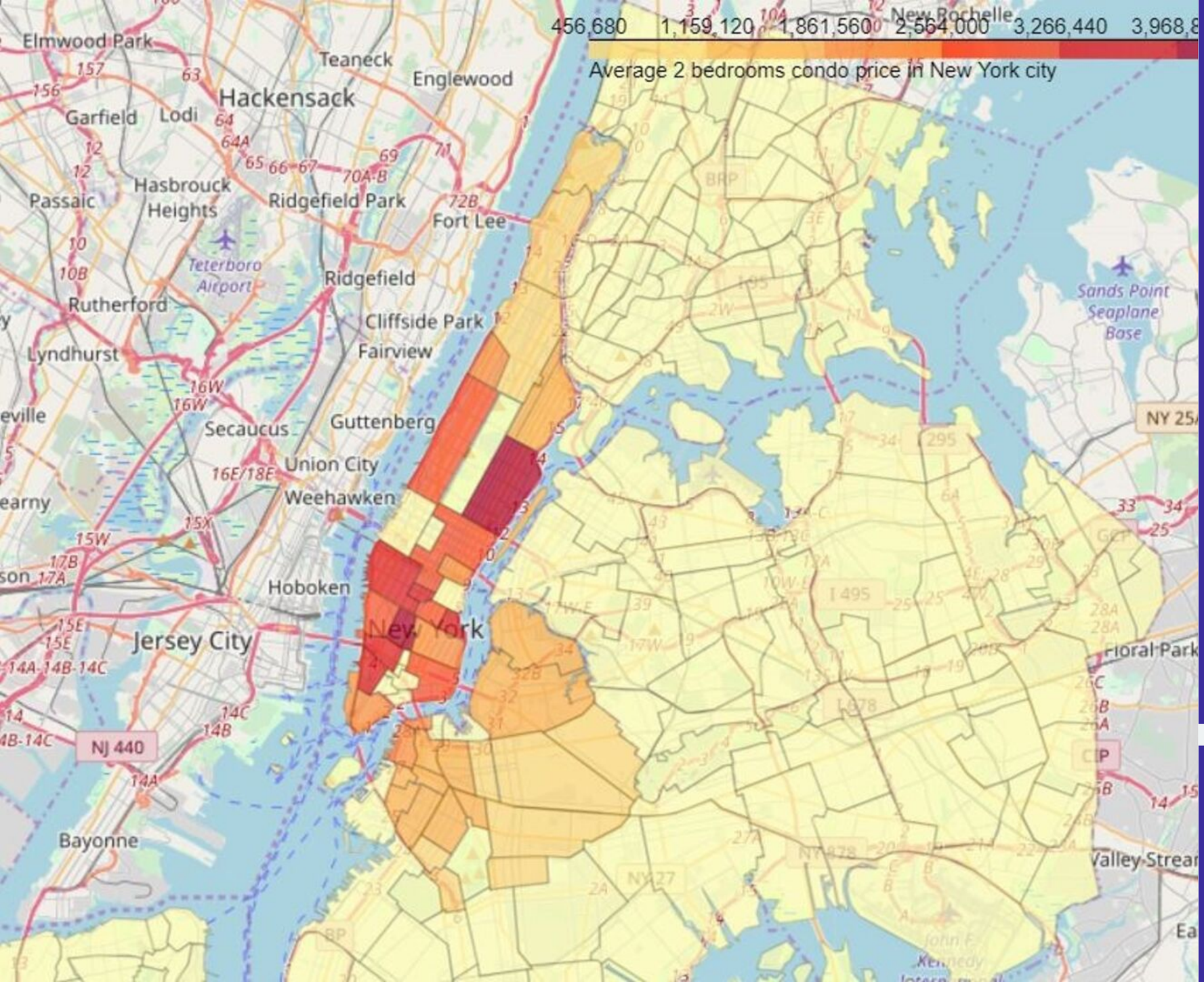
- For each neighborhood, I adopted Geocoder Python to get its coordinate.

- For each neighborhood's coordinate, call FourSquare API to get the surrounding venues.

- Count the occurrences of each venue type and attach that information to each neighborhood.



INSIGHTS



Choropleth map

The map (Figure 2) shows high price in neighborhoods that located around Central Park, Midtown and Lower Manhattan. The price reduces further toward North Manhattan or toward Brooklyn.

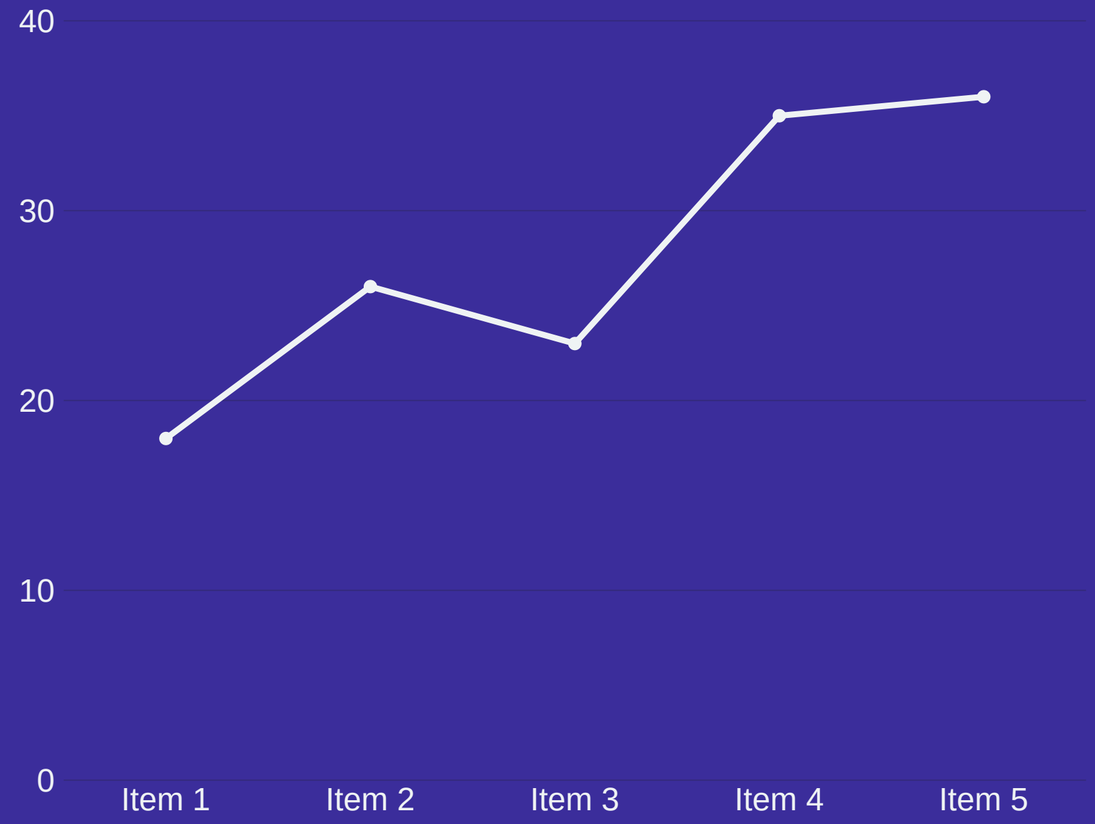
Manhattan can be considered the heart of New York city. It's where most businesses, tourist attractions and entertainments located. So, the venue types that can attract many people are expected to have the most positive coefficients in the regression model.

Linear Regression

```
R2-score: 0.23516904011781548
Mean Squared Error: 0.2676982235233264
Max positive coefs: [0.48676378 0.29146172 0.2680993  0.2457868  0.2457868  0.2229347
 0.2229347  0.21043096 0.21043096 0.21043096]
Venue types with most postive effect: ['Shanghai Restaurant' 'Colombian Restaurant' 'Kitchen Supply Store'
 'Cafeteria' 'Buffet' 'Library' 'School' 'Train Station'
 'Jewish Restaurant' 'Tennis Stadium']
Max negative coefs: [-0.19743204 -0.17726583 -0.17726583 -0.17726583 -0.17726583 -0.17691113
 -0.17691113 -0.1639553  -0.16333343 -0.16333343]
Venue types with most negative effect: ['Reservoir' 'Flea Market' 'Golf Driving Range' 'Physical Therapist'
 'Photography Studio' "Dentist's Office" 'Sports Club'
 'Argentinian Restaurant' 'Mini Golf' 'Print Shop']
Min coefs: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
Venue types with least effect: ['Hookah Bar' 'Food Stand' 'Gym Pool' 'Leather Goods Store'
 'Pakistani Restaurant' 'State / Provincial Park' 'Gas Station' 'Factory'
 'Video Store' 'TV Station']
```

R2 score is small, which means the model may not be suitable for the data. But on the bright side, the coefficient list shows some interest and logical information, for example:

- Bar” and “Market” sure are nice to visit sometimes but may not be a suitable neighborhood for family with kids. “Lighthouse” and “Golf” usually located in the rural areas. The demand for such locations is usually low.



Principal Component Regression

The result is promising as it shows improvement over the simple Linear Regression. As for the coefficient list, the size has been reduced after performing PCA.

```
R2 score: 0.45487512107642447
MSE: 0.1907989730288751
```

CONCLUSION

Looking at the correlations of the coefficients, it is noted that the proximity of places such as restaurants and supermarkets seem to increase the real estate value. Neighborhoods with many restaurants tend to have other businesses nearby as well, featuring a place with good circulation of people, businesses, and consequently high demand, causing a higher price in the region.

Thank you!