Capstone Project - The Battle of Neighborhoods - Week 2

EXPLORING THE NEW YORK CITY NEIGHBORHOODS AND RENT CORRELATION ANALYSIS

INTRODUCTION:

- This project is the capstone for the 4-courses IBM Applied Data Science Specialization on Coursera. The requirement is 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 chosen topic is the correlation between the real estate value and its surrounding venues.

- The idea comes from the process of, any family searching a home to stay after moving to another city. It is common that the owners or agents advertise their properties are closed to some kinds of venues like supermarkets, restaurants, public transport, hospital, school or coffee shops, etc.; showing the “convenience” of the location in order to raise their house’s value to sale or rent.

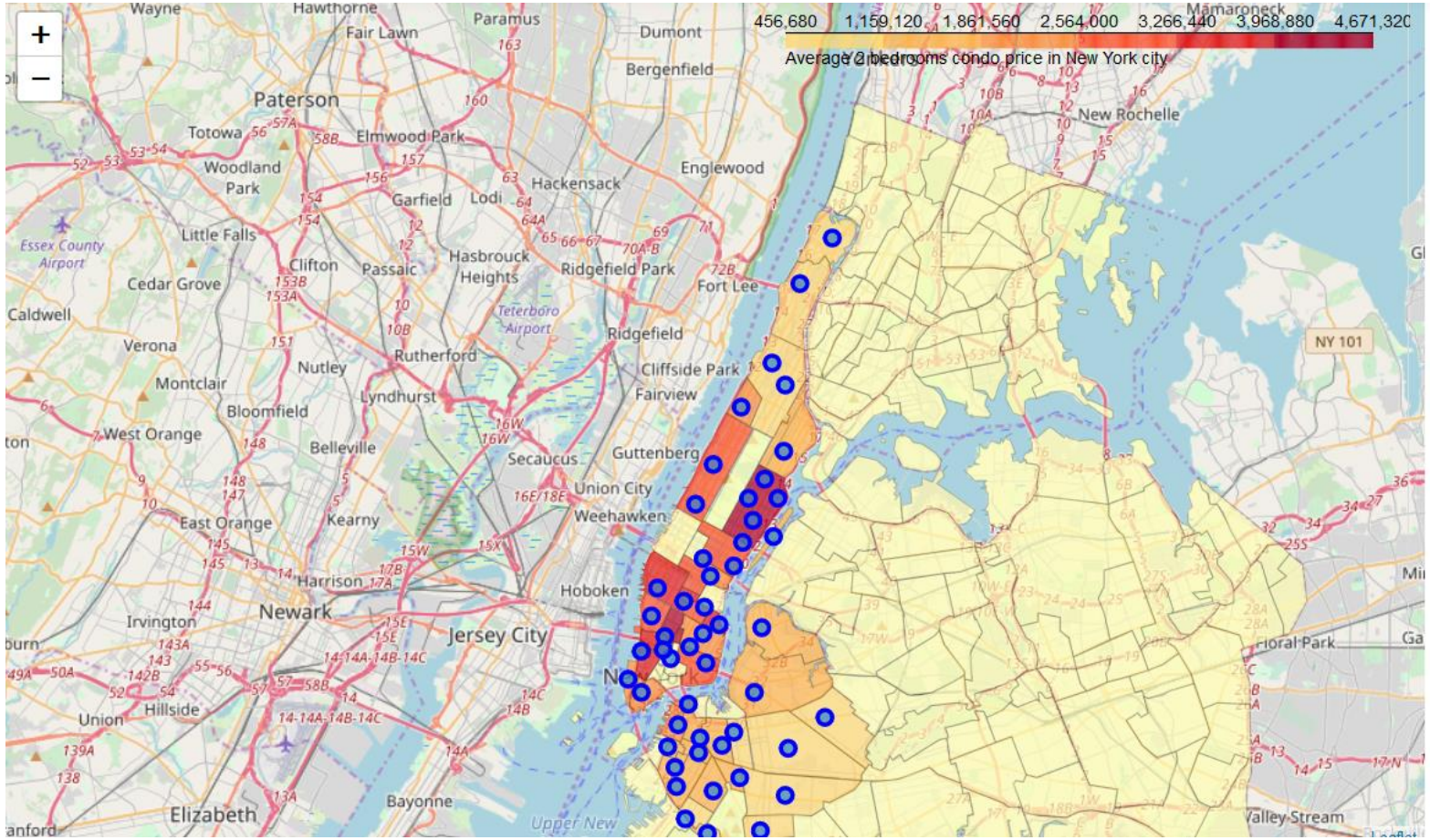
DATA ACQUISITION AND CLEANING:

Borough	Neighborhood	Latitude	Longitude
Manhattan	Marble Hill	40.876551	-73.910660
Brooklyn	Bay Ridge	40.625801	-74.030621
Brooklyn	Bensonhurst	40.611009	-73.995180
Brooklyn	Sunset Park	40.645103	-74.010316
Brooklyn	Greenpoint	40.730201	-73.954241

1. **Scrap CityRealty website for neighborhoods average prices:**
• URL: https://www.cityrealty.com/nyc/market-insight/features/get-to-know/average-nyc-condo-prices-neighborhood-june-2018/18804

2. Get the neighborhoods coordinate:
• Free geodata is available free:
https://geo.nyu.edu/catalog/nyu_2451_34572

DATA ACQUISITION AND CLEANING:



New York City Neighbourhoods In Manhattan and Brooklyn were chosen as the observing Target.

ANALYSIS:

```
R2-score: -0.09267315291287392
Mean Squared Error: 0.3824461577385236
Max positive coefs: [0.43725563 0.29969864 0.25684479 0.25684479 0.25129688 0.21334786
 0.20890573 0.20890573 0.20890573 0.20890573]
Venue types with most postive effect: ['Shanghai Restaurant' 'Colombian Restaurant' 'Public Art' 'Cafeteria'
 'Daycare' 'School' 'Tennis Stadium' 'Train Station' 'Resort'
 'Jewish Restaurant']
Max negative coefs: [-0.22352035 -0.21625058 -0.16691904 -0.16691904 -0.16284558 -0.16284558
 -0.16284558 -0.16128374 -0.16057373 -0.15866141]
Venue types with most negative effect: ['Reservoir' 'Newsstand' 'Lighthouse' 'Rest Area' 'General Entertainment'
 'Camera Store' 'Sports Club' 'Food' 'Beach' 'Indie Theater']
Min coefs: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
Venue types with least effect: ['Gas Station' 'Pakistani Restaurant' 'Cemetery' 'Gym Pool' 'TV Station'
 'Volleyball Court' 'Music Store' 'Indoor Play Area' 'Hookah Bar'
 'Video Store']
```

The Result Doesn't Look Promising:

- The R2 Score Is Small.
- There Are No Really Strong Coefficient Correlations.

APPLYING PCR FOR BETTER RESULT:

```
r2_max = scores_df['R2'].idxmax()
print("Best n:", r2_max, "R2 score:", scores_df['R2'][r2_max])

mse_min = scores_df['MSE'].idxmin()
print("Best n:", mse_min, "MSE:", scores_df['MSE'][mse_min])
```

```
Best n: 1 R2 score: 0.2282642741532619
Best n: 1 MSE: 0.27011495830461724
```

```
Max positive coefs: [0.00354808 0.00333641 0.00323187 0.00316966 0.00316587 0.00304529
 0.00299787 0.0028402  0.00276214 0.00267814]
Venue types with most positive effect: ['Clothing Store' "Women's Store" 'Hotel' 'Cycle Studio' 'Cosmetics Shop'
 'French Restaurant' 'Mediterranean Restaurant'
 'Paper / Office Supplies Store' 'Perfume Shop' 'Noodle House']
Max negative coefs: [-0.00353    -0.00351123 -0.00270299 -0.00252015 -0.00247271 -0.0024354
 -0.00242926 -0.00239611 -0.00231234 -0.00225048]
Venue types with most negative effect: ['Deli / Bodega' 'Pizza Place' 'Bank' 'Mobile Phone Shop'
 'Caribbean Restaurant' 'Convenience Store' 'Playground'
 'Latin American Restaurant' 'Bar' 'Mexican Restaurant']
Min coefs: [-3.98315858e-07 -3.98315858e-07  1.24795798e-05 -2.11981857e-05
 -2.12403779e-05 -3.14793129e-05 -3.26419457e-05 -3.26419457e-05
 -3.36592062e-05 -3.49673940e-05]
Venue types with least effect: ['Ukrainian Restaurant' 'Taiwanese Restaurant' 'German Restaurant'
 'Greek Restaurant' 'Bistro' 'Whisky Bar' 'Stationery Store'
 'Basketball Stadium' 'Coworking Space' 'Street Art']
```

The result is promising as it shows improvement over the simple Linear Regression.

RESULT:

- Based on the assumption that the price of a real estate is dependent on its surrounding venues. Regression techniques were used to get the coefficient correlation between each venue type and the price. And at the end, producing a model to predict how higher or lower a neighborhoods price compared to the mean, based on the occurrence of its surrounding venue types.
- First, Simple Linear Regression was used to see how the approach would perform. Then a more sophisticated method, Principal Component Regression (PCR), was applied to improve the result.
- Unfortunately, the end result isn't very promising. With the R2 score (or Coefficient of determination) of 0.22, the model isn't really fit to the data set; and thus, can't be used for further predicting the real estate price.

Conclusion and Future Scope:

- First, about the data. Real estate prices are not usually available to the public. So, collecting a large set is impossible without connections with some real estate agencies. In this project, there are only 50 samples, but with more than 300 features. Since collecting more samples is not possible at the moment, PCR was chosen to solve the problem by reducing the features size before applying regression.

- Second, about the analysis process and conclusion. With no formal academic background in statistics and mathematics, the tools and methods might not be used with their optimal configuration. And the insight might not be drawn out fully, or even worse not correct at all. Further study in statistical inference and multivariate statistical analysis after this program is a must.