IBM Data Science Capstone project

Battle of the Neighborhoods

1. Introduction

The purpose of this project is to investigate the possible correlation between housing prices and surrounding venues.

Real estate prices in New York are some of the highest in the world. We know for a fact that the neighborhood has an effect on housing prices, but do surrounding venues also have an effect on these prices?

It is this question that we will try to define an answer for.

The target audience for this report:

- Buyers ,
  Can determine how high the price of real estate will be in certain neighborhoods based on surrounding venues.
- Sellers,
  Can determine at what price they want to market their real estate, based on the surrounding venues.
- Real estate developers,
  Can determine in what neighborhoods it would be interesting to develop new real estate, based on existing venues.

2. Description of data

As mentioned in the intro, we will analyze real estate prices in New York City. Therefor, the most important dataset that we will use are the average prices of real estate in New York City per Neighborhood.

Furthermore, the Foursquare API will be used to collect data concerning the surrounding venues.

The tables below display the dataframes which contain the used data.

Figure 1: average real estate price per neighborhood

	Area	Neighborhood	AvgPrice
0	Brooklyn	Bedford-Stuyvesant	750000
1	Brooklyn	Boerum Hill	1.69e+06
2	Brooklyn	Brooklyn Heights	2.15e+06
3	Brooklyn	Bushwick	967000
4	Brooklyn	Carroll Gardens	1.51e+06

Figure 2: Latitude and Longitude per neighborhood

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

Figure 3: Dataframes joined

	Neighborhood	AvgPrice	Latitude	Longitude
0	Bedford-Stuyvesant	750000	40.687232	-73.941785
1	Boerum Hill	1.69e+06	40.685683	-73.983748
2	Brooklyn Heights	2.15e+06	40.695864	-73.993782
3	Bushwick	967000	40.698116	-73.925258
4	Carroll Gardens	1.51e+06	40.680540	-73.994654

3. Methodology

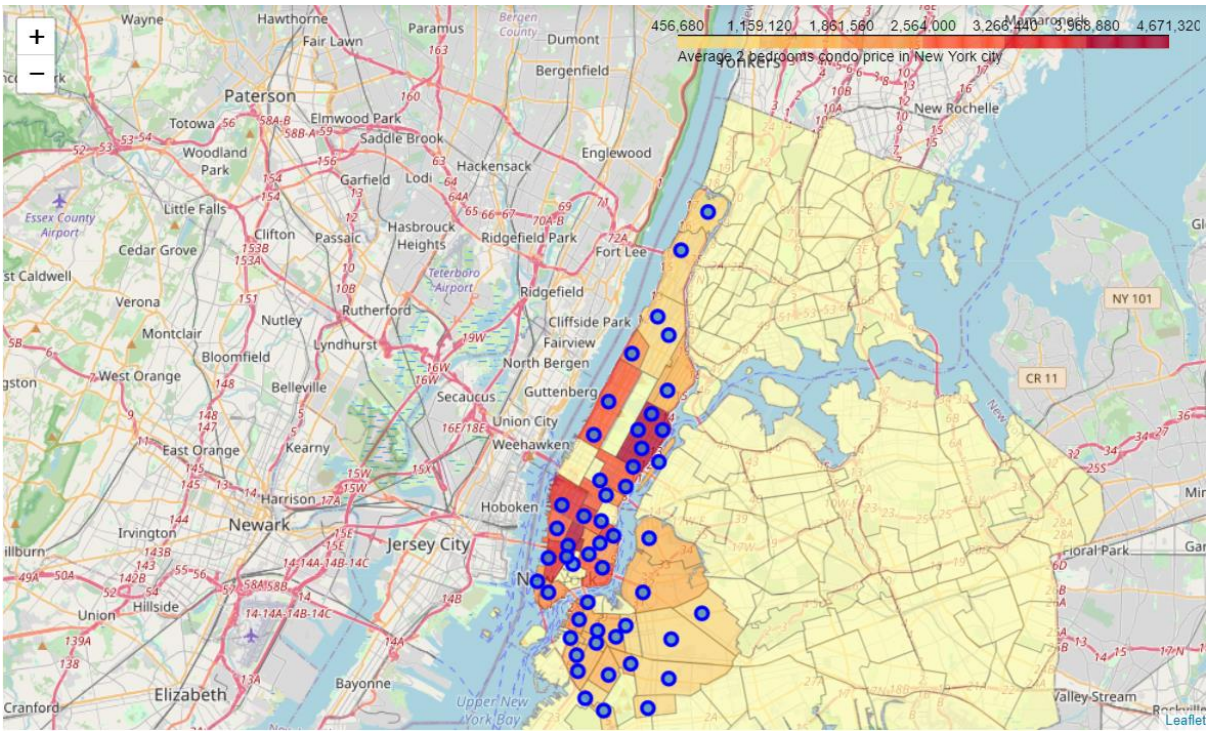
3.1 Techniques

To determine if venues have any effect on housing price, the correlation has te be determined. Therefor a regression analysis will be applied to this business problem.

3.2 Exploratory data analysis

To create a first insight in in real estate prices, the most basic visualization was used. Figures 1, 2 and 3 show the main characteristics of the data in a standard table.

After this initial insight, I have chosen to develop a choropleth map of the area that I investigate. This map uses differences in shading and coloring to indicate a property's price. This map give a quick overview of the price differences in a neighborhood.



3.3 Regression analysis

Linear regression should give us an insight in de correlation between venues and real estate prices. The price would be the dependent variable in this model. We will use de MSE and R2 score to determine how strong the correlation between de variables is.

The result of this analysis is shown in Figure 4

Figure 4: Regression analysis

```
# let's see how well Linear Regression fit the problem
y_pred = lreg.predict(X_test)

print('R2-score:', r2_score(y_test, y_pred)) # r2 score
print('Mean Squared Error:', mean_squared_error(y_test, y_pred)) # mse

print('Max positive coefs:', lreg.coef_[np.argsort(-lreg.coef_)[:10]])
print('Venue types with most postive effect:', X.columns[np.argsort(-lreg.coef_)[:10]].values)
print('Max negative coefs:', lreg.coef_[np.argsort(lreg.coef_)[:10]])
print('Venue types with most negative effect:', X.columns[np.argsort(lreg.coef_)[:10]].values)
coef_abs = abs(lreg.coef_)
print('Min coefs:', lreg.coef_[np.argsort(coef_abs)[:10]])
print('Venue types with least effect:', X.columns[np.argsort(coef_abs)[:10]].values)
```

```
R2-score: -0.0035500373879067126
Mean Squared Error: 0.35125220645735017
Max positive coefs: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
Venue types with most postive effect: ['Athletics & Sports' 'Insurance Office' 'Kitchen Supply Store' 'Lawyer'
 'Liquor Store' 'Miscellaneous Shop' 'Moving Target' 'Music Venue'
 'Other Repair Shop' 'Outdoors & Recreation']
Max negative coefs: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
Venue types with most negative effect: ['Athletics & Sports' 'Insurance Office' 'Kitchen Supply Store' 'Lawyer'
 'Liquor Store' 'Miscellaneous Shop' 'Moving Target' 'Music Venue'
 'Other Repair Shop' 'Outdoors & Recreation']
Min coefs: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
Venue types with least effect: ['Athletics & Sports' 'Insurance Office' 'Kitchen Supply Store' 'Lawyer'
 'Liquor Store' 'Miscellaneous Shop' 'Moving Target' 'Music Venue'
 'Other Repair Shop' 'Outdoors & Recreation']
```

This result show that the correlation between house prices is weak. Other information that we derive is as follows:

The problem with the data set might be the fact that there are only 50 rows and 300 features. To get a better understanding of the correlation, a more comprehensive dataset would be best.

4. Results

As mentioned above, the results of this analysis imply that this might not be the best model to analyses the effect of venues on housing prices. If we want to accurately predict price of real estate, a multiple linear regression model might be more suitable, but the data we used does not support this.

5. Discussion

The biggest challenge and also recommendation that I would make is to define a comprehensive dataset for this problem. The first challenge is finding the right dataset.

The second recommendation that I would make is to reserve a lot of time for combining different datasets that consist of data.

Finally it might be good to apply different kinds of Machine Learning techniques tot determine the correlation of the variables. For the sake of time and effort I chose to use just one, but different techniques might give different insights.

6. Conclusion

We started this report with our main question: Do surrounding venues have any kind of effect on the housing prices?

The result of this small investigation using a regression analysis show that there is little effect that venues have on the housing prices in New York City.

Please note that this analysis has been done as part of a course by a novice to python and data science. It would be to premature to conclude that there is no real effect that venues have on the real estate prices.