Impact of neighborhood characteristics on housing prices

1. Introduction

Housing price prediction is an interesting topic for many of us, whether you are in the real estate industry trying to make money buying and selling houses, or you just want to purchase a decent house for yourself with a reasonable price. However, prediction is not an easy job since there are many factors affecting the house prices, such as living area, year of construction, location, number of bedrooms, number of bathrooms.

In this project we are going to investigate the impact of the neighborhood characteristics on housing prices. We are going to answer two main questions:

1. Does information on the neighborhood characteristics help to improve housing prices prediction?
2. If it does, what types of neighborhoods are likely to have higher housing prices?

2. Methodology

We will look for the answers to the above questions using a data science approach. Two datasets are explored in this project: a New York housing price dataset from Realtor, which contains information on houses sold recently in New York, and a New York neighborhood information dataset from Foursquare, which contains neighborhoods' recommended venues. Based on the datasets, two prediction models will be generated. The first model uses only the housing price dataset, while the second model uses the housing price dataset in combination with the neighborhood information dataset. The models are built and trained using the multi linear regression method. The performances of the two models are then compared to see if the neighborhood information improves the prediction. If it does, we will further investigate how the characteristics of the neighborhoods affect the housing prices.

3. Data

For this analysis we will use the following 2 datasets:

1. New York housing prices dataset from Realtor: this dataset contains a list of houses sold recently in New York, together with their information such as lot size, location, year built, number of bedrooms and so on.

2. New York neighborhood information dataset from Foursquare: this dataset contains a list of New York's neighborhood and their recommended venues. In order to import this dataset, we also need the coordinate data of the neighborhoods which is imported from https://geo.nyu.edu/catalog/nyu_2451_34572.

3.1. New York housing prices data

3.1.1. Data importing

The New York housing prices dataset from Realtor is imported using the Realtor API. The dataset contains a list of houses sold recently in New York together with their information. The downloaded data is a json file. I converted it to a pandas dataframe which is shown in Figure 1.

	property_id	listing_id	prop_type	list_date	last_update	year_built	beds	baths_full	baths_half	prop_status	...	of
0	O3324614978	2870154772	condo	2019-12-23T21:36:02Z	2020-05-04T12:20:03Z	1964.0	2.0	1.0	NaN	not_for_sale	...	1c570f54a97857fa51eb9a423b2
1	O9949620288	2512980267	condo	2019-10-29T14:06:43Z	2020-05-04T15:26:41Z	2019.0	2.0	2.0	NaN	not_for_sale	...	33fb537ee27bfbf0303f8cfd324
2	O4933762654	2465636145	condo	2019-07-09T14:25:59Z	2020-05-04T14:17:55Z	1960.0	2.0	1.0	1.0	not_for_sale	...	d0300f260f315de7c6eb8c5cec
3	O9359213003	2756391634	condo	2019-10-25T17:43:21Z	2020-05-01T19:58:54Z	1957.0	1.0	1.0	NaN	not_for_sale	...	f61053eabb4db1531dcf98b4473
4	O4635149680	2854255611	multi_family	2019-12-12T20:25:51Z	2020-05-04T13:49:51Z	1920.0	3.0	3.0	NaN	not_for_sale	...	99df1bf8266bd14faf4f4346e77

Figure 1 Original housing prices dataframe

The dataframe contains 72 columns and 6364 data points. However, not all data points are useful. We first have to clean up the data.

3.1.2. Data cleaning

First, we will drop all the duplicated rows in the dataframe. Furthermore, we can see that there are columns which are not informative for our housing price prediction such as 'listing_id', 'prop_status', 'office.phones'. Therefore we will create a new dataframe which contains only the features which are relevant to the house price. Some long column names are shorten for convenience. All the rows with NaN values are dropped. The resulting dataframe is shown in Figure 2.

	property_id	prop_type	beds	baths_full	baths	year_built	lot_size	building_size	neighborhood	postal_code	lat	lon	price
0	O3324614978	condo	2.0	1.0	1.0	1964.0	NaN	1000.0	Coney Island - Brooklyn,NY	11224	40.578851	-73.969859	440000
1	O9949620288	condo	2.0	2.0	2.0	2019.0	NaN	1133.0	Bath Beach - Brooklyn,NY	11214	40.603675	-74.002613	728000
2	O4933762654	condo	2.0	1.0	2.0	1960.0	12222.0	NaN	NaN	11414	NaN	NaN	360000
3	O9359213003	condo	1.0	1.0	1.0	1957.0	NaN	775.0	Riverdale - Bronx,NY	10463	40.883872	-73.911068	190500
4	O4635149680	multi_family	3.0	3.0	3.0	1920.0	1500.0	1175.0	Sheepshead Bay - Brooklyn,NY	11229	40.596440	-73.955704	799000

Figure 2 Selective housing prices dataframe

We notice that the 'neighborhood' column also contains the borough part, e.g. 'Riverdale - Bronx,NY'. We will remove the borough part in the 'neighborhood' column. Next, we see that there are 4 different property types in the dataset:

```
single_family    641
condo            493
multi_family     461
mobile             1
Name: prop_type, dtype: int64
```

We will drop the only one 'mobile' property and convert other property types to integer values for regression purpose:

- 'condo' = 0
- 'single_family' = 1
- 'multi_family' = 2

Furthermore, let us investigate the lot_size and building_size:

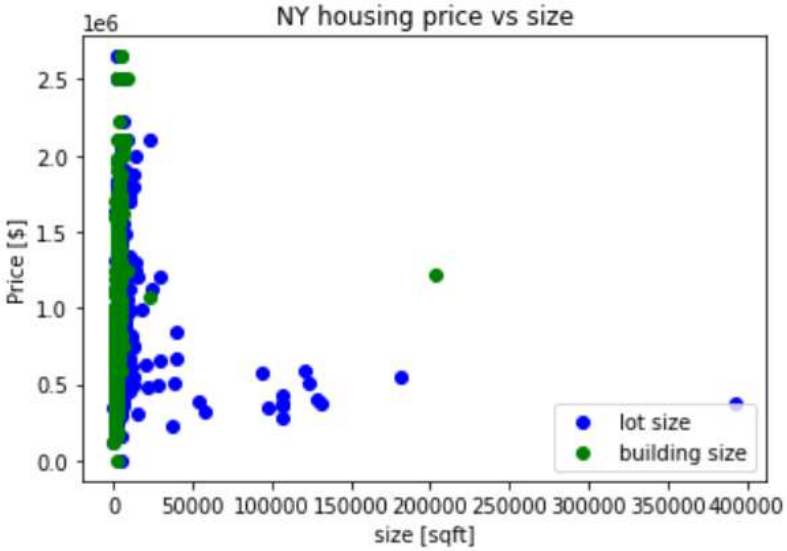

Figure 3 House price vs lot_size and building_size

We observe that there are some suspicious data points where the size is very large but the price is relatively low. For this reason, we will drop all the data points with lot_size and building size

larger than 20000 square feet. Finally, we will reset the index of the dataframe. The cleaned dataframe is shown in Figure 4.

	property_id	prop_type	beds	baths_full	baths	year_built	lot_size	building_size	neighborhood	postal_code	lat	lon	price
0	O4635149680	2	3.0	3.0	3.0	1920.0	1500.0	1175.0	Sheepshead Bay	11229	40.596440	-73.955704	799000
1	O3100293204	2	2.0	1.0	1.0	1899.0	4000.0	1752.0	South Shore	10307	40.507548	-74.251220	627900
2	O3154267434	2	5.0	3.0	3.0	1960.0	2200.0	2200.0	Canarsie	11236	40.644589	-73.888601	750000
3	O3971533857	1	3.0	1.0	2.0	1930.0	2500.0	1404.0	Auburndale	11360	40.771425	-73.787866	800000
4	O3981442835	2	4.0	3.0	4.0	2018.0	5781.0	3057.0	Pleasant Plains	10309	40.522243	-74.221881	1135000

Figure 4 Cleaned housing prices dataframe

3.2. New York neighborhoods data

3.2.1. Data importing

The New York neighborhoods dataset contains the neighborhoods' recommended venues from Foursquare. In order to import neighborhoods' recommended venues from Foursquare, we first need the latitudes and longitudes of the neighborhoods, which can be imported from https://geo.nyu.edu/catalog/nyu_2451_34572. The imported json data contains several information on the neighborhoods in New York. We extract the neighborhoods' names and their coordinates to a dataframe. Note that many neighborhoods in this dataset do not appear in our housing prices dataset, which is therefore not interesting for us. We will keep only the neighborhoods that appear in both datasets. The dataframe is shown in Figure 5.

```
There are 104 neighborhoods that appear in both datasets.
```

	neighborhood	neighborhood_lat	neighborhood_lon
0	Wakefield	40.894705	-73.847201
1	Eastchester	40.887556	-73.827806
2	Morris Heights	40.847898	-73.919672
3	Unionport	40.829774	-73.850535
4	Bay Ridge	40.625801	-74.030621

Figure 5 Neighborhood coordinates dataframe

With the neighborhoods' coordinates, we import recommended venues for each neighborhood from Foursquare using Foursquare API *explore* call.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

Figure 6 Neighborhoods' recommended venues

The most popular venues categories are:

```
Pizza Place           102
Deli / Bodega          83
Chinese Restaurant     71
Pharmacy               68
Bakery                 60
Name: Venue Category, dtype: int64
```

3.2.2. Data cleaning

First we will use one hot encoding to convert each venue category to a column. Then we group rows by neighborhood and take the mean of the frequency of occurrence of each category. The resulting dataframe is shown in Figure 7.

	neighborhood	Accessories Store	Afghan Restaurant	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	...	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnam Restaur
0	Arden Heights	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.00000	0.0	
1	Arverne	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.00000	0.0	
2	Astoria	0.0	0.0	0.010204	0.0	0.0	0.0	0.0	0.0	0.0	...	0.010204	0.00000	0.0	
3	Auburndale	0.0	0.0	0.055556	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.00000	0.0	
4	Bath Beach	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.02381	0.0	

Figure 7 Neighborhoods' recommended venues one hot encoded

Finally we create a dataframe that combines both housing data and neighborhoods' venues data. The data is now ready.

4. Prediction models

The datasets have been imported and cleaned. We are now ready to build multi linear regression models to predict housing price in New York. We will generate 2 models, one with housing data only, and another with both housing data and neighborhood information. The performances of the 2 models will be compared to see how the neighborhood information affects the prediction.

4.1. Prediction model using housing data only

4.1.1. Build and train model

We first create a features dataframe with housing features only as shown in Figure 8. A multi linear regression model is built using the features in this dataframe. We then split this dataset in to train set and test set. The regression model is trained using the train set. Its performance is then verified using the test set

	prop_type	beds	baths_full	baths	year_built	lot_size	building_size	postal_code	lat	lon
0	1	3.0	2.0	2.0	1970.0	1488.0	1376.0	10312	40.555243	-74.196336
1	1	3.0	1.0	1.0	1925.0	4000.0	1201.0	10312	40.547307	-74.192839
2	0	2.0	1.0	2.0	1986.0	640.0	980.0	10312	40.552280	-74.199922
3	0	2.0	1.0	2.0	1979.0	1330.0	1441.0	10312	40.550653	-74.191594
4	0	3.0	2.0	3.0	1981.0	2340.0	1800.0	10312	40.549002	-74.199929

Figure 8 Features dataframe with housing features only

4.1.2. Results

The multi linear regression model has a prediction score of 0.58. This is not a great score, but at least it gives us some ideas on the relations between the house's features and its price. We will take a look at some of these relations.

The relation between lot_size and house price is shown in Figure 9. As expected, our model predicts that the price increases with the lot size. Next, we investigate the impact of number of bedrooms on house price which is shown in Figure 10. The model predicts that houses with more bedrooms are more expensive as we expected.

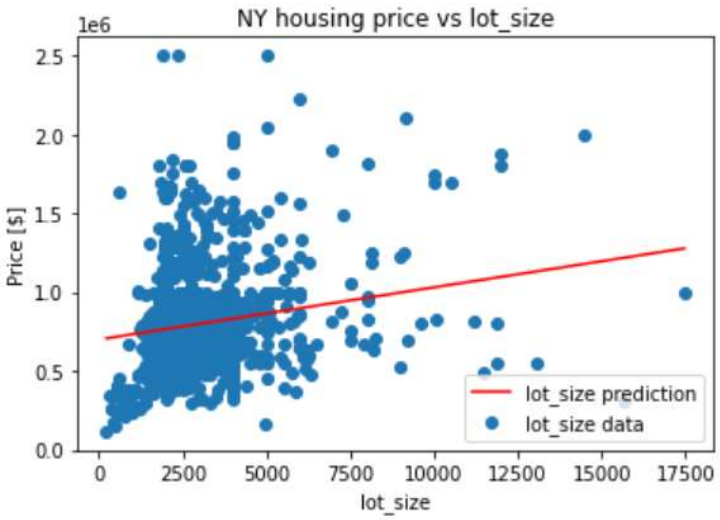

Figure 9 House price vs lot_size

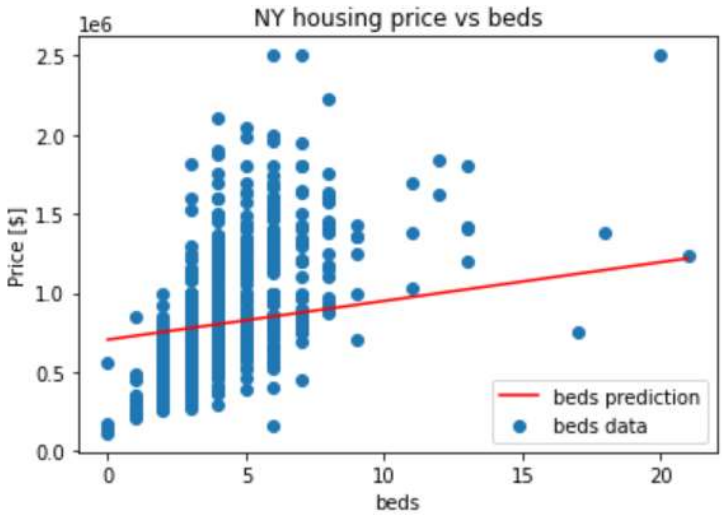

Figure 10 House price vs beds

4.2. Prediction model using housing data and neighborhood data

4.2.1. Build and train model

Now let us incorporate the neighborhood data into our prediction model to see if it helps to improve the prediction accuracy. We will not take all the venue categories into account because many of them have very low number of appearances and show up in only 1 or 2 neighborhoods, which is not very informative and may even lower our prediction accuracy. Therefore we will consider only the top 50 most popular venue categories. The top 50 venue categories are added to our housing features dataframe to create a new dataframe. The new dataframe contains both housing features and neighborhoods' venues features:

	prop_type	beds	baths_full	baths	year_built	lot_size	building_size	postal_code	lat	lon	...	Discount Store	Spanish Restaurant	Dessert Shop	Greek Restaurant
0	1	3.0	2.0	2.0	1970.0	1488.0	1376.0	10312	40.555243	-74.196336	...	0.0	0.0	0.0	0.0
1	1	3.0	1.0	1.0	1925.0	4000.0	1201.0	10312	40.547307	-74.192839	...	0.0	0.0	0.0	0.0
2	0	2.0	1.0	2.0	1986.0	640.0	980.0	10312	40.552280	-74.199922	...	0.0	0.0	0.0	0.0
3	0	2.0	1.0	2.0	1979.0	1330.0	1441.0	10312	40.550653	-74.191594	...	0.0	0.0	0.0	0.0
4	0	3.0	2.0	3.0	1981.0	2340.0	1800.0	10312	40.549002	-74.199929	...	0.0	0.0	0.0	0.0

Figure 11 Features dataframe with both housing features and neighborhood features

A multi linear regression model is built using the features in this dataframe. We then split this dataset in to train set and test set. The regression model is trained using the train set. Its performance is then verified using the test set

4.2.2. Results

With the addition of neighborhood information, the prediction score has increases considerably from 0.58 to 0.71. This proves that the neighborhood information does help to increase prediction accuracy. Our first research question has been answered.

5. Discussion: neighborhood characteristics and housing prices

We will now investigate how the characteristics of the neighborhoods impact the housing prices. Let us first investigate the most popular venue category which is pizza places. From Figure 12 we can see our model predict that the more pizza places are in the neighborhood, the lower the house prices will be.

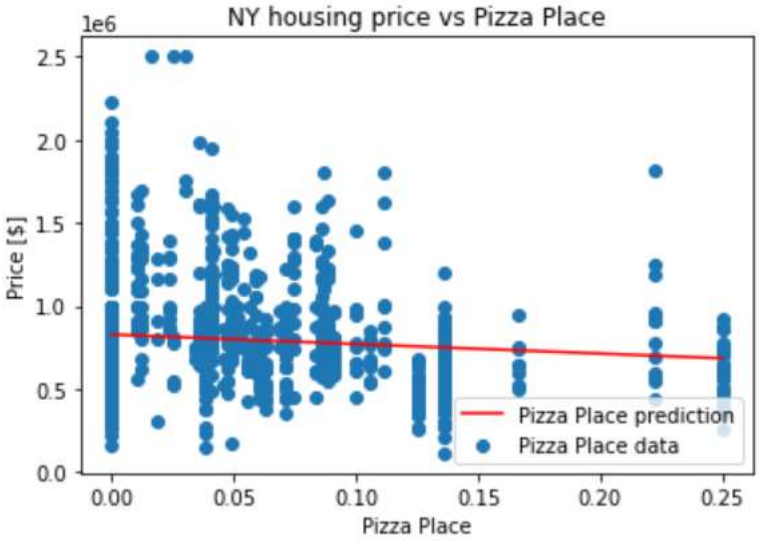

Figure 12 Housing price vs Pizza place

Next, let us take a look at the second most popular one, which is delis / bodegas. In Figure 13 our model also predicts an inverse linear relationship between the number of delis / bodegas and the house prices.

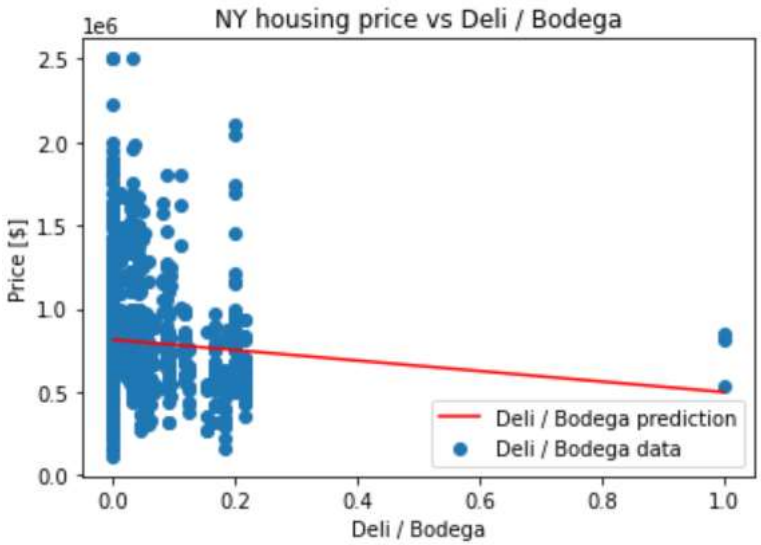

Figure 13 House price vs Deli / Bodega

Perhaps people prefer quieter neighborhoods? Let us investigate the relation between parks and housing prices in Figure 14. Although not very significant, it looks like houses in neighborhoods with more parks are more expensive.

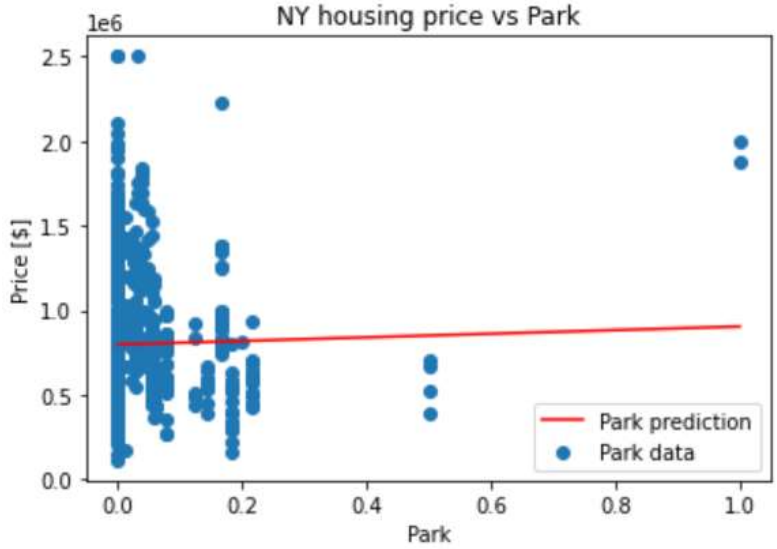

Figure 14 House price vs Park

We will now investigate the relation between several types of restaurants and housing prices shown in Figure 15. It is interesting to see that the house prices go up with the number of Italian and Mexican restaurants, but go down with the number of Chinese and Spanish restaurants.

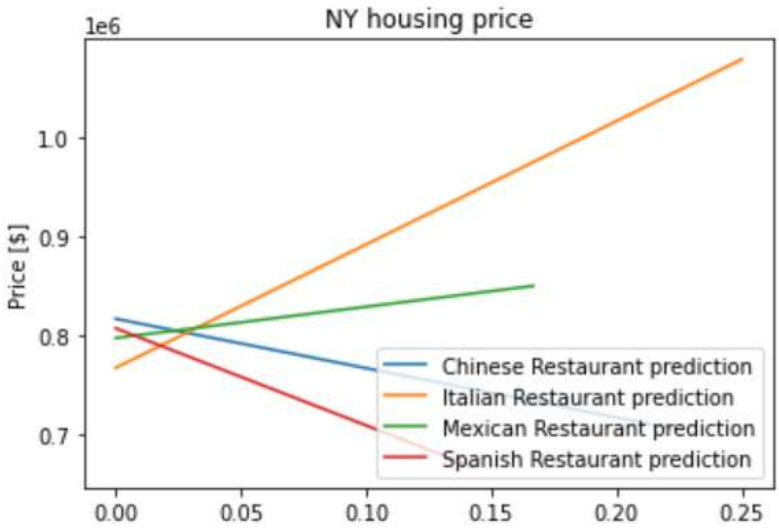

Figure 15 House price vs Restaurants

Another interesting venue category is bars. The relation between bars and housing price is shown in Figure 16. Although there is a positive linear relation, the impact of bars on housing price seems to be small.

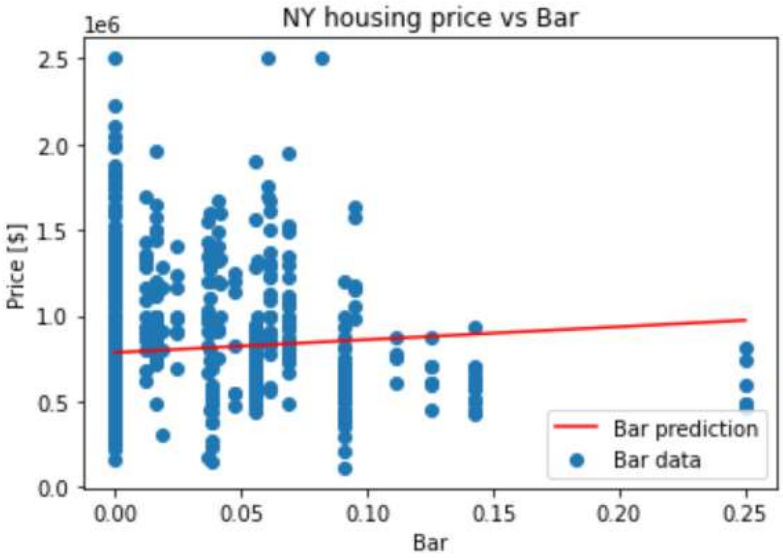

Figure 16 House price vs Bar

Next, according to Figure 17, the housing prices seems to decrease with the number of playgrounds.

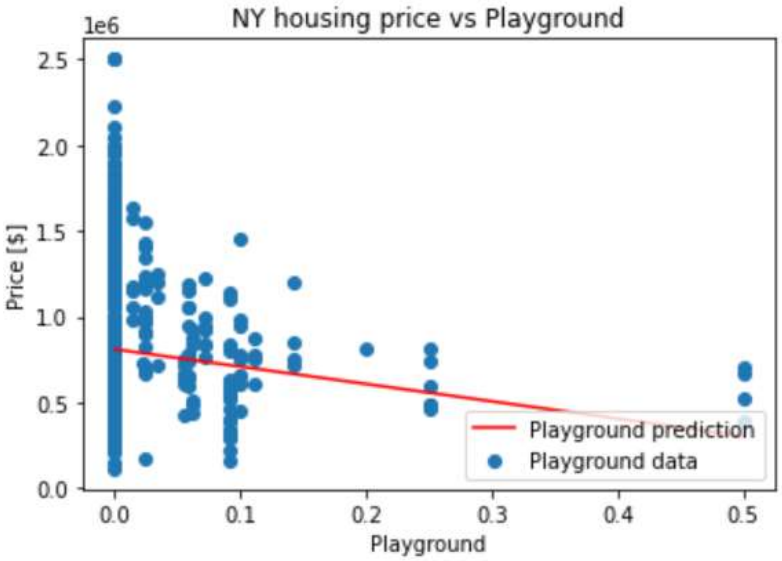

Figure 17 House price vs Playground

The above-mentioned relations are not easy to explain. This opens up some interesting questions for further investigation.

6. Conclusions

In this project we have used a data science approach to investigate the impact of neighborhood characteristics on housing prices. We have found the answers to our 2 research questions:

1. **Does information on the neighborhood characteristics help to improve housing prices prediction?**
   The answer is that information on neighborhood characteristics indeed does help to improve the prediction accuracy. By incorporating the neighborhood information into the regression model, we increased the prediction score from 0.58 to 0.71.

2. **If it does, what types of neighborhoods are likely to have higher housing prices?** With the trained regression model, we have found some interesting relations between neighborhoods and housing prices. It looks like the more pizza places and delis / bodegas are in the neighborhood, the lower the house prices will be. The price increases if there are more parks nearby. Another interesting observation is that the housing prices seems to go up with the number of Italian and Mexican restaurants, but go down with the number of Chinese and Spanish restaurants. These relations are not easy to explain, but this opens up interesting questions for further investigation.

For future investigation it is also interesting to apply other advanced regression methods to this problem to see how the prediction is improved.