Prediction Average Rent in Manhattan

IBM Professional Certificate Program Final Project Report

1. Introduction

1.1 Business Problem

Choosing a house is a complicated process. You are looking for not only a nice house but also a familiarity of your old neighborhood. When you go to any rental website, you'll only see features specific to the houses, such as room numbers, bathroom numbers, etc. However, the surroundings of the house are also relevant when you judge the house itself and its quality. For example, sure, the house may be fine, but if there isn't a market nearby, is it worth the price?!

There are two problems to tackle here: 1) To find a place to meet some requirements (e.g.: If I have a child I want a primary school nearby my home), 2) After finding the house, make a judgement about its price (eg: Is it overvalued?) Because, if you specially don't know anything about the city (or even country) people may try to rip you off. Hence it's valuable to have information.

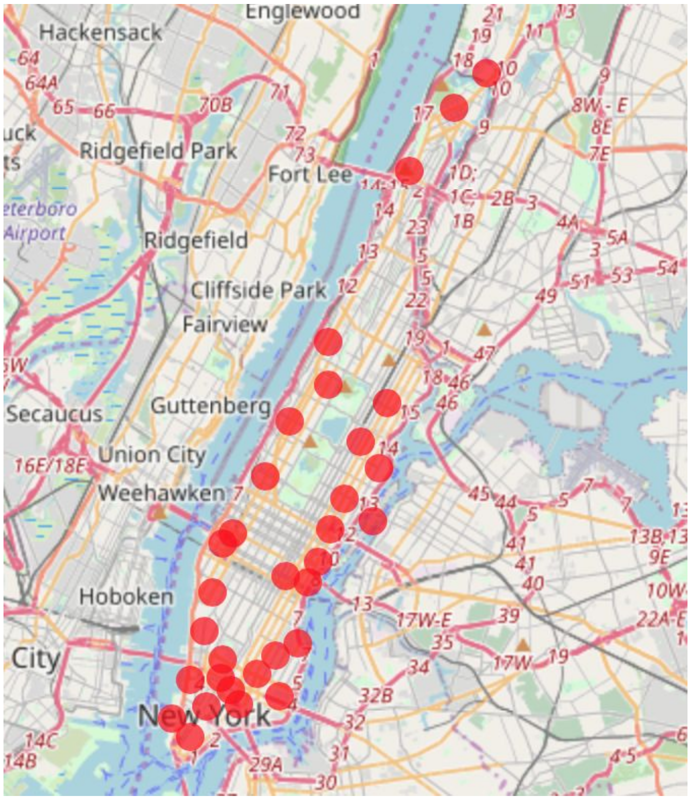
I'm going to focus on the 2nd problem. What I'm going to do in this project is to put all venue knowledge together and estimate the rent price of a house in Manhattan, NY. What I will do in this project is to predict the price of a house by using Foursquare venue data and house characteristics.

1.2 Targeted Audience

The results can be used by anyone who is looking for a house, especially people who don't know the area very well. Individuals will produce better results if they are able to make educated decisions.

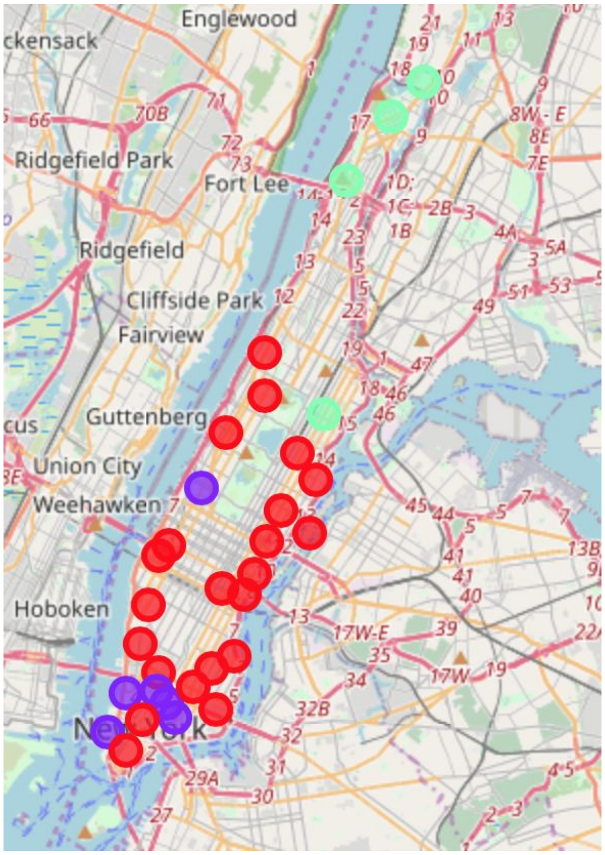
2. Data Preparation

Two separate datasets are used in this project. One is scrapped from rental websites which are *rentaljungle.com* and *rentcafe.com*. The main reason I chose these two is that they have the average rent information for each neighborhood in Manhattan. It is crucial that this information is provided in order to create a final data set in which information on locations is also available. Figure below depicts the neighborhoods in Manhattan.

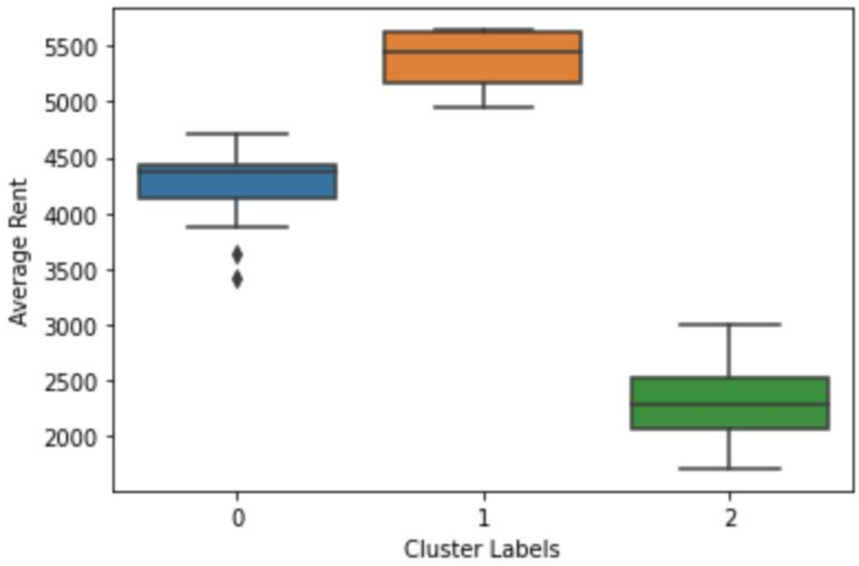


The other is Foursquare venue dataset. We utilize Foursquare API to get the location information for every neighborhood in Manhattan.

I create another feature *by clustering* the neighborhood average rent. Cluster number is only three which represents low-medium-high rents. Green dots are the neighborhoods with low rent range whereas red dots show the expensive ones.

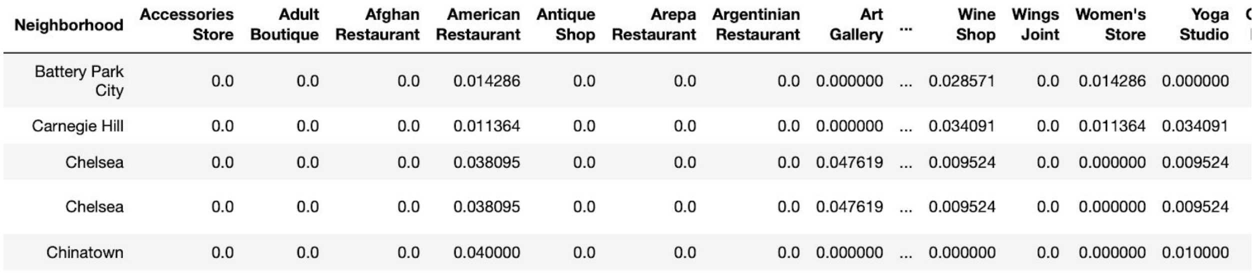


Box plot below shows the difference between average prices in each cluster.



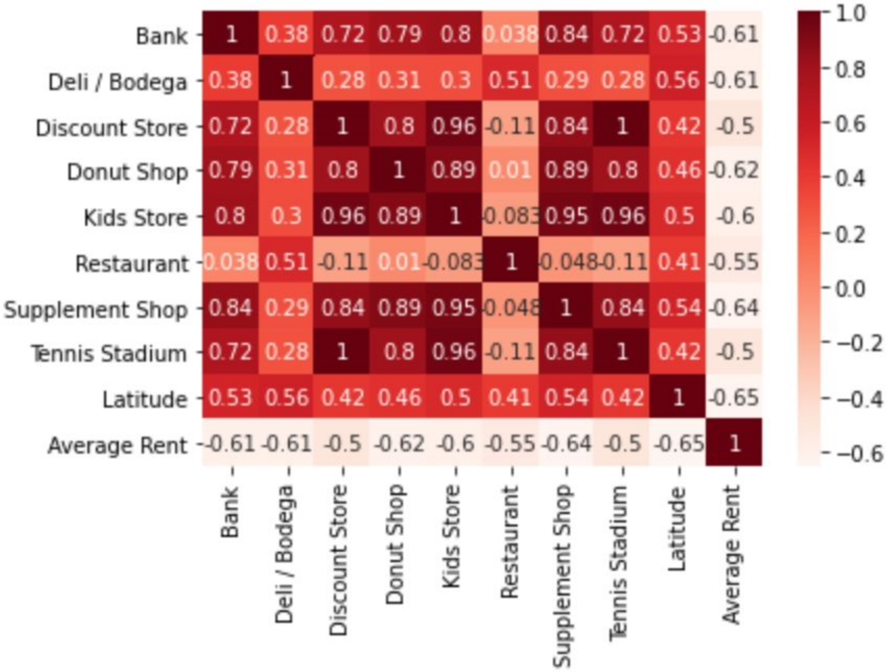
2.1 Feature Selection

In final data, there are 321 features which is way too much for this small data (observe the dots between column names).

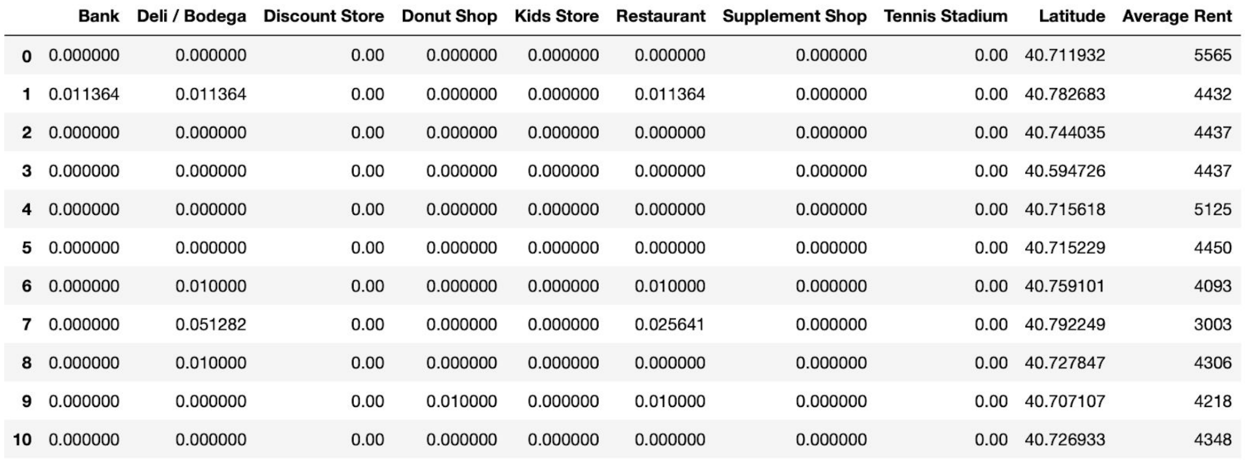


Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	...	Wine Shop	Wings Joint	Women's Store	Yoga Studio	C
Battery Park City	0.0	0.0	0.0	0.014286	0.0	0.0	0.0	0.000000	...	0.028571	0.0	0.014286	0.000000	
Carnegie Hill	0.0	0.0	0.0	0.011364	0.0	0.0	0.0	0.000000	...	0.034091	0.0	0.011364	0.034091	
Chelsea	0.0	0.0	0.0	0.038095	0.0	0.0	0.0	0.047619	...	0.009524	0.0	0.000000	0.009524	
Chelsea	0.0	0.0	0.0	0.038095	0.0	0.0	0.0	0.047619	...	0.009524	0.0	0.000000	0.009524	
Chinatown	0.0	0.0	0.0	0.040000	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.000000	0.010000	

I found “appropriate” columns by looking at the correlation of each variable with target variables and keeping the columns that have correlation coefficients greater than 0.5. (I assume that *>0.5* coefficient indicates a linear relationship).



After all preparations, final data looks like this:



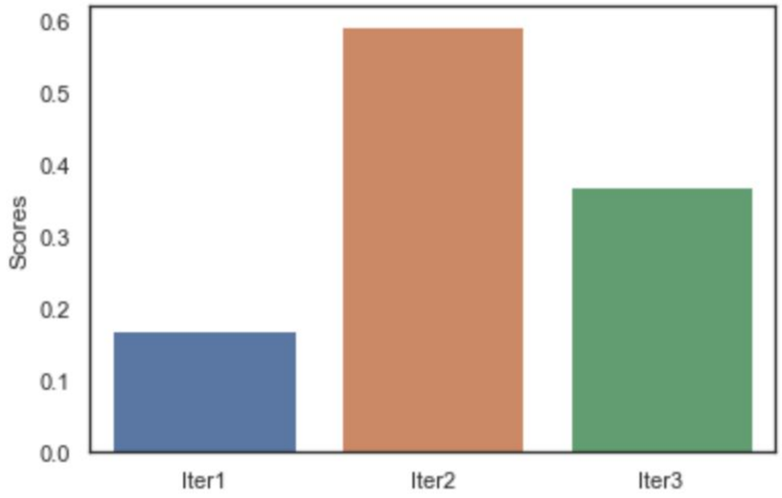
	Bank	Deli / Bodega	Discount Store	Donut Shop	Kids Store	Restaurant	Supplement Shop	Tennis Stadium	Latitude	Average Rent
0	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	40.711932	5565
1	0.011364	0.011364	0.00	0.000000	0.000000	0.011364	0.000000	0.00	40.782683	4432
2	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	40.744035	4437
3	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	40.594726	4437
4	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	40.715618	5125
5	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	40.715229	4450
6	0.000000	0.010000	0.00	0.000000	0.000000	0.010000	0.000000	0.00	40.759101	4093
7	0.000000	0.051282	0.00	0.000000	0.000000	0.025641	0.000000	0.00	40.792249	3003
8	0.000000	0.010000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	40.727847	4306
9	0.000000	0.000000	0.00	0.010000	0.000000	0.010000	0.000000	0.00	40.707107	4218
10	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	40.726933	4348

3. Results

The problem here is to predict a continuous number, i.e., the rent. Hence I'm going to use linear regression models. Since I have multiple features (i.e., columns), I will start with multiple regression. Scikit-learn uses the plain Ordinary Least Squares method to solve this problem. You can use different optimization approaches as well, but since I don't have a large dataset I'm going to stick with OLS. And again because of the small dataset, I used Kfold for training.

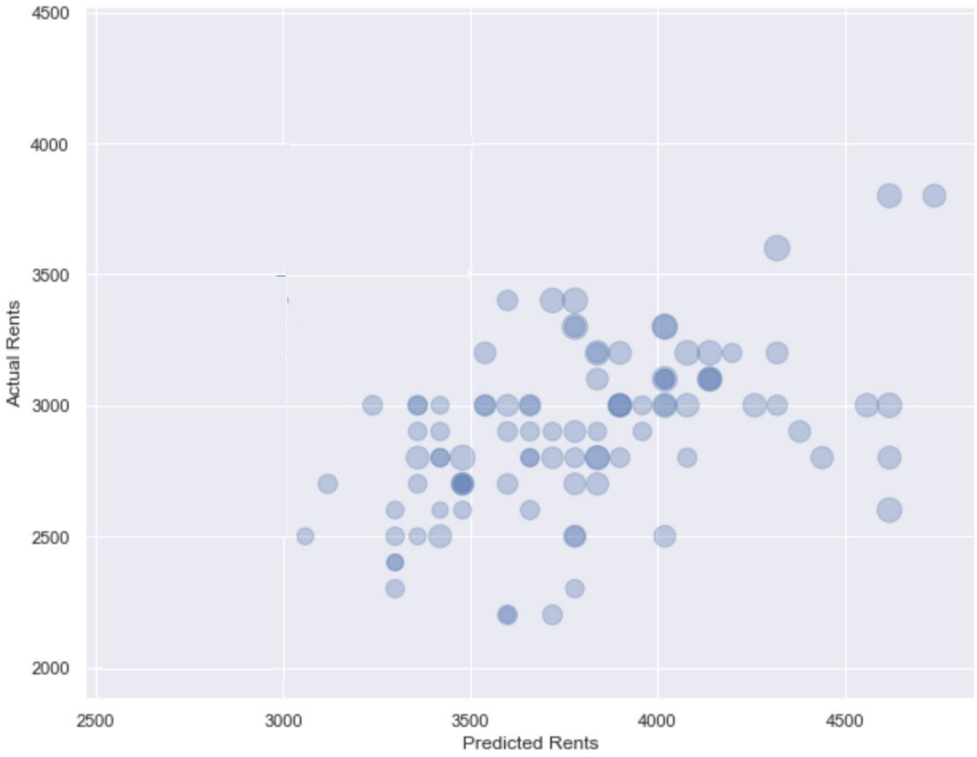
I've used R-squared for evaluation. R squared is not an error per se but is a popular metric for the accuracy of your model. It represents how close the data values are to the fitted regression line. The higher the R-squared, the better the model fits your data.

Below plot depicts the accuracy for each iteration.



So I chose the second iteration’s train indices to use in the actual prediction phase.

Accuracy (R_square) for the prediction is 0.61. The reason is that my dataset is not large enough. But still, linear relation can be captured through models.



4. Conclusion and Future Directions

- Different regression models' predictions can be merged for the final prediction.
- More rental data needed for a thorough analysis.
- The recommendation feature may be added. And the person's personal data would be included and the algorithm could recommend a location and then the person could look up different houses in the suggested area and check the rent.
- Different features should be included such as:

- Financial data of the person (for recommendation).
- Probability of Neighborhood Gentrification. The idea here is to take into account whether the neighborhood is going to be undergoing a process of gentrification and its effect on rents.