Predicting Average Rent

IBM Capstone Project Presentation

Introduction

- Choosing a house is a complicated process. You are looking for not only a nice house but also a familiarity of your old neighborhood.

- There are two problems to tackle here: 1) finding a place while meeting some of the user's requirements (e.g.: If I have a child I want a primary school nearby my home), 2) After finding the house, making a judgement about its price (eg: Is it overvalued?)

- I focused on the 2nd problem.

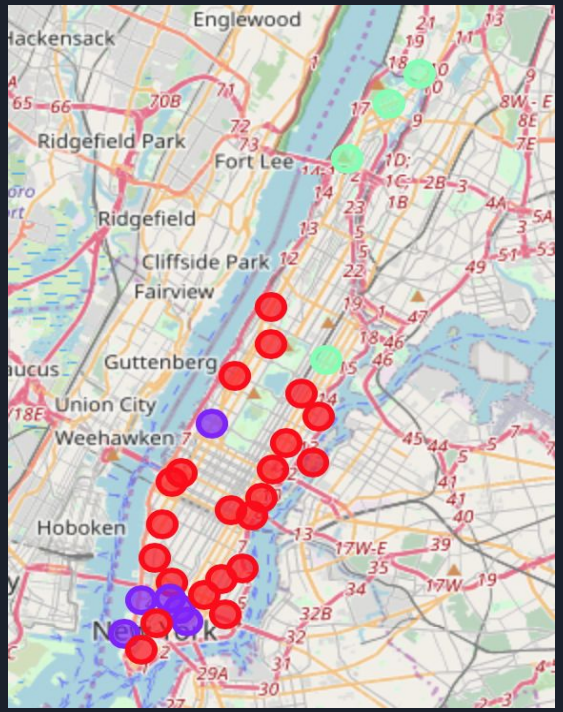
- **Targeted Audience:** The results can be used by anyone who is looking for a house, especially people who don't know the area very well.

Data Acquisition

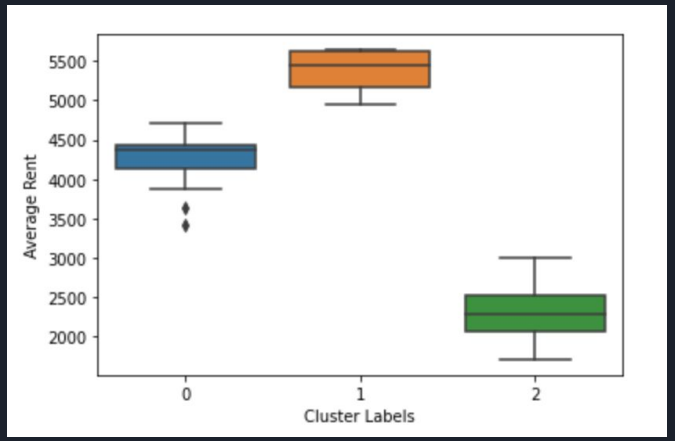
- Two separate datasets are used in this project:

    1. 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.
    2. The other is Foursquare venue dataset. We utilize Foursquare API to get the location information for every neighborhood in Manhattan.

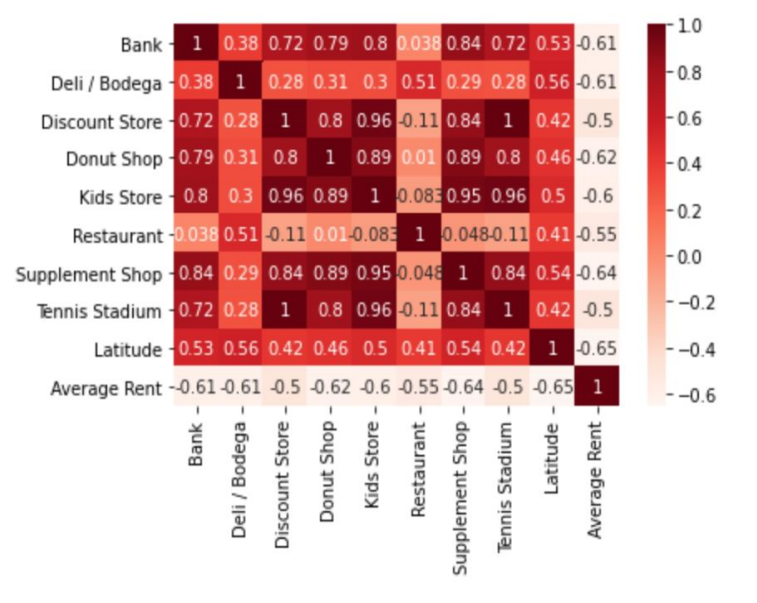
Data Preparation



- 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.



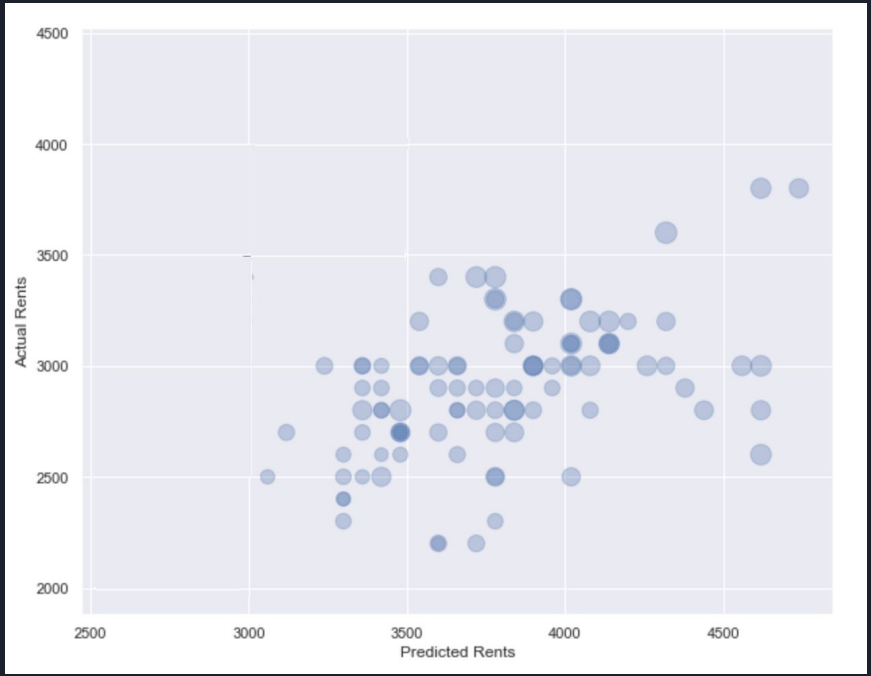
Feature Selection



- In the final data, there are 321 features which is way too much for this small data.

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

Results



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 model.

Conclusion and Future Direction

- 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).
    - Neighborhood Gentrification Probability. The idea here is to look at whether the neighborhood will go under a gentrification process and its effect on the rent.