DSCI 510 Final Project

Kuan Hsien Lin

USC ID: 7837198015

DSCI 510: Principles of Programming for Data Science

May 8, 2023

1. Abstract

This project aims to explore the relationship between housing prices and regional characteristics, such as location, nearby facilities, and population census data, in New York City real estate market. In addition, we use some machine learning models to make predictions about housing prices by learning and identifying patterns from our datasets.

In the project, we use three datasets (house data, facilities count data, and census data) obtained from web scraping and API. Our analysis focuses on all housing types and 2b2b (two bedrooms and two bathrooms) housing. For all types of housing price prediction, we employed three machine learning models, such as Decision Tree, Lasso linear regression, Random Forest, and OLS, to predict housing prices and compare the performance of the three models to find the best one. There are four main findings from our analysis result. One is that the Random Forest model is the most suitable for predicting housing prices in our dataset. Second, the factors such as housing space, bachelor's degree percentage, the number of restaurants and supermarkets, and the number of bathrooms affect housing prices. This result differs from our previous assumption that the number of public transportation facilities around a house is most highly related to housing prices. On the other hand, for the prediction of 2b2b median housing prices, bachelor's and associate degree percentages are the most important key factors for predicting 2b2b median house. Geographically, higher housing prices are concentrated in the lower side of Manhattan. This analysis can help home buyers, investors, and policymakers better understand the real estate market, make informed decisions, and adjust urban planning and housing policies accordingly.

2. Motivation

In previous research, we analyzed the relationship between affordable housing and subway stations in New York City. As a result, we found that the accessibility of subway stations significantly impacts the evaluation of affordable housing site selection. This discovery sparked my interest in further exploring the relationship between prices and regional characteristics in the New York City real estate market. Based on the above, we hypothesize that the number of mass transit facilities nearby may most highly affect New York City housing prices. To test this hypothesis, we will analyze all types of housing and 2b2b housing in New York City in this project and examine the relationship between these factors and housing prices. In addition, I choose 2b2b housing because two bedrooms and two-bathroom types are essentially the "bread and butter" of the housing sales market in NYC and are the most in-demand properties.

3. Data Source

We fetch our datasets from below three sources. The house and census data sets are obtained by web scraping, and the number of surrounding facilities is caught from API.

Source 1:
Real Estate website (Web scraping)
https://www.trulia.com/NY/New_York/{zip_code}/

We web-scraped the housing data from the above URL for each zip code in five boroughs. The zip code data is from source 3. This scraping dataset includes 28492 property sales (buy) in New York City. Each instance in the housing dataset has the address, zip code, price, house space (sqft), the number of bedrooms and bathrooms, latitude, longitude, and property type. Source 1 data set is saved as source1_web_scaping_full.csv. Below Figure (1) are a few examples from the source 1 dataset.

source1_web_scraping_full

address	zipcode	price	sqft	bedroom	bathroom	latitude	longitude	Property Type
515 18th St #1904, New York, NY 10011	10011	7475000	2167	3	3	40.745083	-74.00653	condo
119 Waverly Pl #3-4, New York, NY 10011	10011	11500000	3600	4	4	40.73279	-73.999275	coop
147 W 15th St #9S, New York, NY 10011	10011	5600000	2654	2	3	40.739025	-73.99806	coop
80 W Washington Plce, New York, NY 10011	10011	22500000	8757	6	8	40.731827	-74.00004	single-family home
500 22nd St #A, New York, NY 10011	10011	9985000	3005	3	4	40.74701	-74.00503	condo

Fig. (1) Source 1 data set (example)

Source 2.
Overpass API

```
req = f"""
    [out:json];
    node(around:{radius},{lat},{lon});
    out;
    """
    result = api.query(req)
```

API documentation: https://wiki.openstreetmap.org/wiki/Tag:railway%3Dsubway_entrance

The Overpass API is to obtain all facility nodes within a specified radius centered on the latitude and longitude of the location. For example, we could write code to calculate the number of facilities, such as schools, subway, parks, metro, buses, supermarkets, banks, parking, cinema, mall, restaurant, gyms, boutiques, museums, arts, and theatres, near the property address. The search radius for our analysis is set to 1km (about 10 minute's walk). Source 2 data set is saved as source2_api_full.csv. Below Figure (2) are a few examples from the source 2 datasets.

source2_api_full

address	latitude	longitude	Subway_count	School_count	Park_count	Metro_count	Bus_count	Supermarket_count	Bank_count	Parking_count	Cinema_count	Mall_count	Restaurant_count	Gym_count	Boutique_count	Museum_count	Arts_count	Theatre_count
80 Gold St #14B, New York, NY 10038	40.71015	-74.00403	120	10	5	22	71	18	57	8	1	2	290	0	0	5	0	1
111 Fulton St #424, New York, NY 10038	40.70974	-74.00662	122	9	7	21	78	14	44	8	2	0	259	0	0	6	0	1
28 Cliff St #4, New York, NY 10038	40.70806	-74.005104	126	9	4	22	73	7	42	8	1	0	220	0	0	5	0	0
80 Gold St #3B, New York, NY 10038	40.70925	-74.00361	116	10	5	19	68	16	49	8	1	2	270	0	0	5	0	1
100 South #C-P, New York, NY 10038	40.7656	-73.9767	109	9	0	16	104	6	51	16	2	1	219	0	0	3	1	11

Fig. (2) Source 2 data set (example)

Source 3.
Unitedstateszipcode website (Web scraping)
https://www.unitedstateszipcodes.org/
(1). https://www.unitedstateszipcodes.org/ny/#zips-list
(2). https://www.unitedstateszipcodes.org/{zipcode}/

We web-scraped all the zip codes in the state of New York (from (1) website) and then selected all the zip codes in the five boroughs of New York City (Richmond County, Bronx County, New York County, Kings County, and Queens County) for source one house data web scraping. From the (2) website, we could get the census tract statistics by searching different zip codes from (1). The information includes median household income, population density (people per sq mi), unemployment rate, and education level. Source 3 data set is saved as source3_web_scraping_full.csv. Below, Figures (3) and (4) are a few examples from the source 3 datasets.

(1).

new_york_state_zipcode

ZIP Code	County
501	Suffolk County
544	Suffolk County
6390	Suffolk County
10001	New York County
10002	New York County

Fig. (3) Source 3 data set from URL (1) (example)

(2).

source3_web_scraping_full

zipcode	Median income	Median home value	Less than High School Diploma	High School Graduate	Associate degree	Bachelor degree	Master degree	Professional school degree	Doctorate degree	Unemployment rate	Population Density
10001	81671	650200	0.08	0.19	0.03	0.38	0.21	0.07	0.03	0.26	33959
10002	33218	535600	0.36	0.31	0.04	0.21	0.06	0.02	0.01	0.45	92573
10003	92540	817700	0.04	0.16	0.02	0.43	0.21	0.09	0.04	0.26	97188
10004	129313	894200	0.02	0.16	0.04	0.42	0.22	0.13	0.02	0.12	5519
10005	124670	1000001	0.023	0.083	0.007	0.478	0.239	0.129	0.041	0.1	97048
10006	119274	714300	0.01	0.11	0.06	0.41	0.24	0.11	0.06	0.06	32796
10007	216037	1000001	0.06	0.15	0.02	0.35	0.24	0.12	0.06	0.22	42751
10009	59929	672800	0.16	0.23	0.03	0.36	0.14	0.05	0.02	0.33	99492
10010	97955	746200	0.03	0.18	0.05	0.4	0.2	0.11	0.03	0.3	81487
10011	104238	914500	0.05	0.15	0.03	0.4	0.22	0.1	0.04	0.22	77436
10012	86594	1000001	0.07	0.16	0.02	0.42	0.18	0.08	0.08	0.21	74517
10013	83725	1000001	0.19	0.18	0.01	0.34	0.18	0.07	0.03	0.32	50154
10014	108483	947300	0.02	0.14	0.02	0.47	0.21	0.1	0.04	0.18	56119

Fig. (4) Source 3 data set from URL (2) (example)

The order of getting the data set is very important. Before grasping the data from source 1, we need the data from source 3, and then we need the data from source 1 to grasp the data from source 2. Below Figure (5) is the flow chart for generating three datasets. We first grab the data from source 3, then from source 1, and finally from source 2.

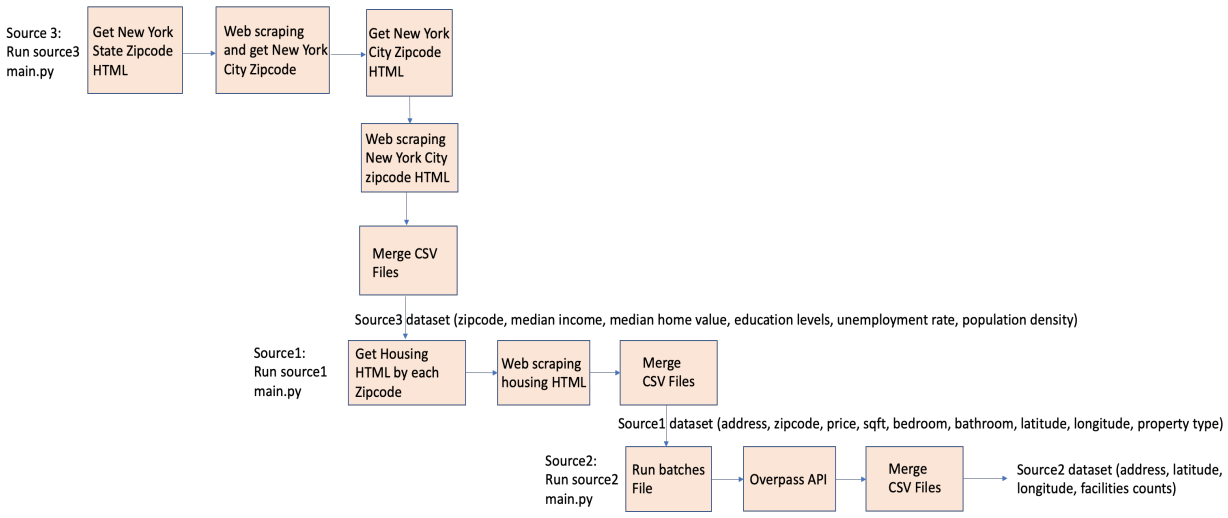


Fig. (5) The flow of generating three datasets

As for data combinations, we first create a read CSV function to read three data sets and use Pandas to create three data boxes each. We then combined data from the real estate website with data from the Overpass API. Specifically, we use pandas' merge() function and the left join operation to match and combine raw real estate data based on house address information as the primary key with facility data obtained from the Overpass API as the foreign key. Next, we'll combine it with census data from the U.S. Zip code website. We combine the census data with the existing real estate and facility data by matching the home's zip code (as a foreign key). Finally, data cleaning and de-weighting are essential. We use the dropna function to clean up missing values and ensure data integrity. In addition, we use the drop_duplicates function in the Pandas library to remove duplicate property data based on address information (primary key), ensuring data exclusivity. After data cleaning, transformation, and normalization, we will use these integrated data to conduct a regression analysis, study the relationship between housing prices and different factors, and visualize the analysis results.

4. Technical Solution

The algorithm flow for our technical solutions is shown in Figure (6). There are three major parts of technical solutions in our algorithm. The first part focuses on data cleaning and transformation. In data cleaning, we dealt with missing and duplicate data by using the dropna function to clear missing values and the drop_duplicates function to remove duplicate property data based on address information to ensure data exclusivity. About the data transformation, we use the function to replace "Studio" in the "bedroom" column with 0. In the data combination of the second part, we used the Pandas library to create a data frame and combined three data sets (real estate data, Overpass API facility data, and census data) into the data frame. We use the

merge() function and the left join operation to combine these data sets based on house address information and zip code. In the third part, the technical solutions are feature derivation and data encoding. We create a new “total amenities” feature by summing all facilities from the original data. In addition, the housing type is one hot encoded. This step transforms the categorical variables into numerical variables that the machine learning model can use. We convert the Property Type field’s Boolean values (True/False) to 0 and 1. True is represented as 1, and False is described as 0. In the fourth part, the technical solutions are feature selection and utilizing machine learning models. We extract meaningful features from the original data to facilitate model processing. Before doing model analysis, we performed a feature analysis (correlation) for all types of homes and 2b2b homes to understand the relationship between home prices and different factors. For all housing types, we apply various machine learning models (from the sci-kit-learn library), such as Lasso linear regression, Decision Tree, and Random Forest regression, to house price prediction. We compare the MSE and R-squared scores of the three models and select the one with the best performance among the three models as our model to predict housing prices. Then, we visualized the predicted results. For 2b2b houses, we used ordinary least squares (OLS) to forecast 2b2b median home prices because the data set has 155 instances, a relatively small data set for a machine learning model. In this case, we try to use OLS linear regression, a relatively simple model for regression analysis, to prevent serious overfitting. Finally, we visualize the price distribution of 2b2b housing on an interactive web map (using folium and branca library) of New York City.

There are some technical challenges that we encounter. When we apply machine learning models, we need to adjust the model parameters to determine the best model performance for our data set, avoiding overfit and underfit. In addition, some machine learning models cannot use plus or minus signs and the value of coefficients to judge the influence of features. For example, Random Forest cannot directly obtain the coefficients of the fitting equation, so we need to use the importance method (rf.feature_importances_) to get the scores of important features. The last technical challenge is creating a web map to show the 2b2b housing price level. First, we create a map object by using folium.Map and use quantiles to assign colors, adding a circular marker with a corresponding color to the map. After that, we customize a legend to represent the relationship between different colors and price ranges and add the legend using branca.element method to the map.

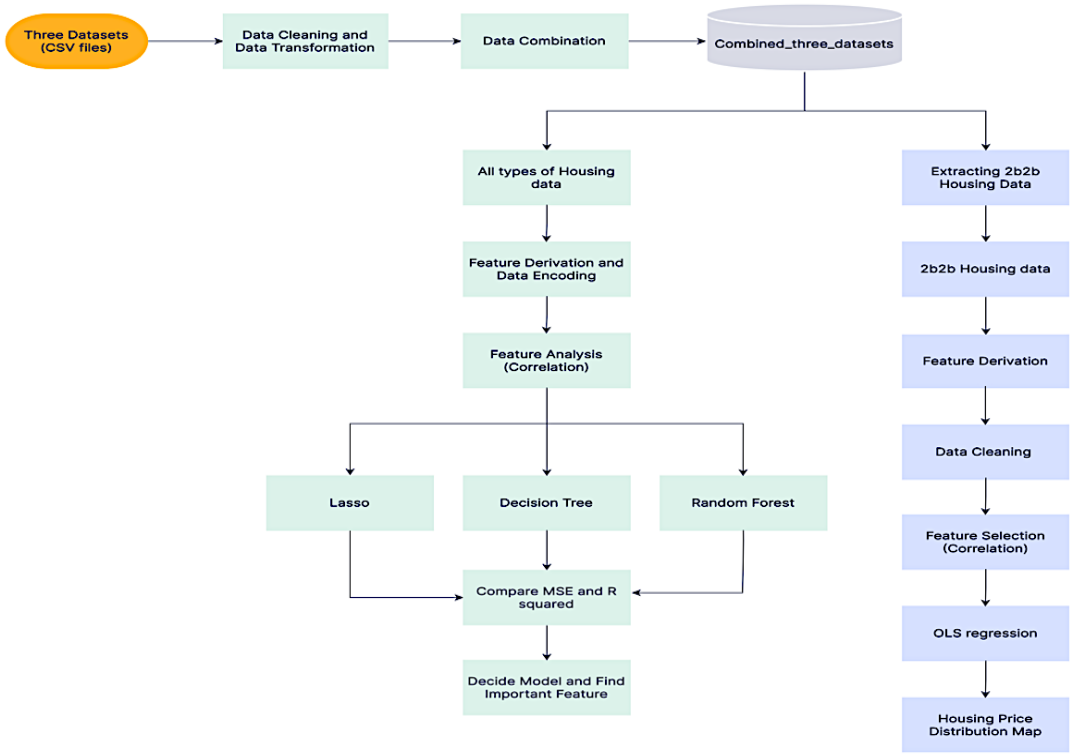
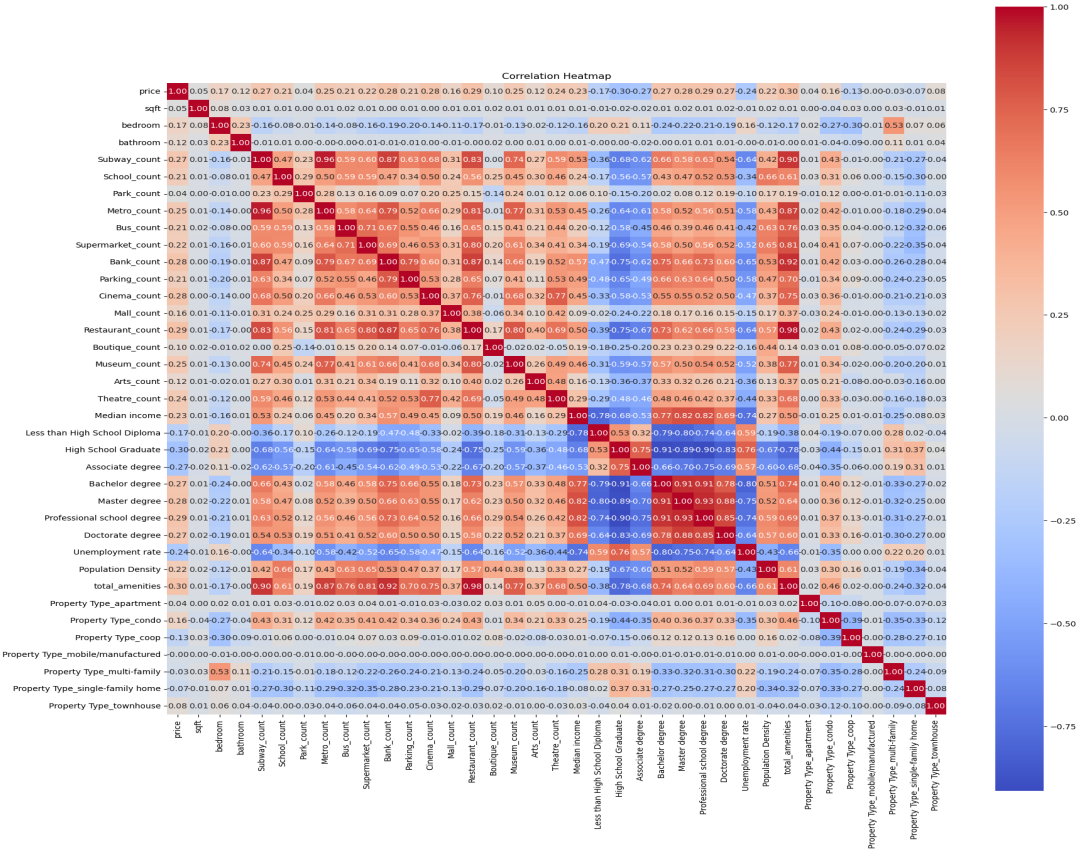


Fig. (6) A algorithm flow chart for our technical solutions

5. Analysis result

In this project, there are two analysis parts. One part is for all housing types, and the other is for 2b2b housing. All detailed analysis process is written in my Jupyter notebook file. Here I summarized the important analysis result in this section. In all types of housing analysis part, we found that the linear relationship between the features and the price is weak in Figure (7) because the top five correlation coefficients of both positive and negative correlations are all less than 0.5. This may be because of the complexity of the data; some of the features may have a more complex nonlinear relationship with the housing price, or multiple features may interact to affect the target variable. In this case, simple linear regression may not be enough to capture all the information and patterns. Therefore, we consider using Lasso regression because it can handle certain nonlinear relations and has the property of feature selection.




```
Top 5 Positively correlated features:
                                Positive Correlation
Feature
total_amenities                             0.298844
Restaurant_count                            0.292281
Professional school degree                  0.289400
Cinema_count                                0.281617
Master degree                               0.277562

Top 5 Negatively correlated features (Absolute Value):
                                  Negative Correlation
Feature
High School Graduate                          0.297494
Associate degree                              0.271405
Unemployment rate                             0.238329
Less than High School Diploma                 0.171795
Property Type_coop                            0.133631
```

Fig. (7) Feature correlation analysis for all type of housing data

After Lasso analysis, Lasso regression cannot perform well on our data set. The result indicated that our data set might have complex non-linear or feature interactions that can cause linear models not to capture the underlying structure of the data well. Therefore, we try to use more complex models, such as Decision Tree and Random Forests, to find the best combination of models and features. In addition, we adjust the parameters to avoid overfitting or underfitting the model. According to all models' analysis results in Figure (8), we found that the Random Forest model performs better than other models in predicting all housing prices, which may be due to the existence of a nonlinear relationship in our data, while Lasso regression have a poor fitting effect on the nonlinear relationship. Furthermore, the Random Forest performs better than Decision Tree because Random Forests combine multiple Decision Tree, reducing the model's

variance and improving the prediction accuracy. Overall, Random Forest performs best and is the best prediction of house prices on our data set.

Evaluation Index	Lasso	Decision Tree	Random Forest
Mean Squared Error (Training)	0.0002858490776420378	0.00012662190614502457	6.156194738492702e-05
R-squared (Training)	0.15852641718462412	0.6272543892543361	0.8187758629189715
Mean Squared Error (Test)	0.0001799353222560663	7.307388824734906e-05	4.013481599307881e-05
R-squared (Test)	0.14743690680444932	0.6537639224202654	0.8098346537029645

Fig. (8) Models comparison results table for all types of housing data

The prediction effect and residual chart of the Random Forest model are shown in Figure (9). The Random Forest model is a nonlinear model that combines many Decision Tree to make predictions. Therefore, a linear regression line representing the random forest model cannot be drawn directly. In this case, we can approximate the predicted results of the model by adding a simulated regression line (y = 0.89x) but note that this line represents only a general trend and may not fully capture the non-linear behavior of the model. From the prediction results and residual illustrations, most of the points in the data were captured by the model. However, there were still some points with significant differences in residual compared with other data points. These points could be data errors or phenomena the model failed to capture. We may need to further study these points later to determine if we need to revise the data or improve the model.

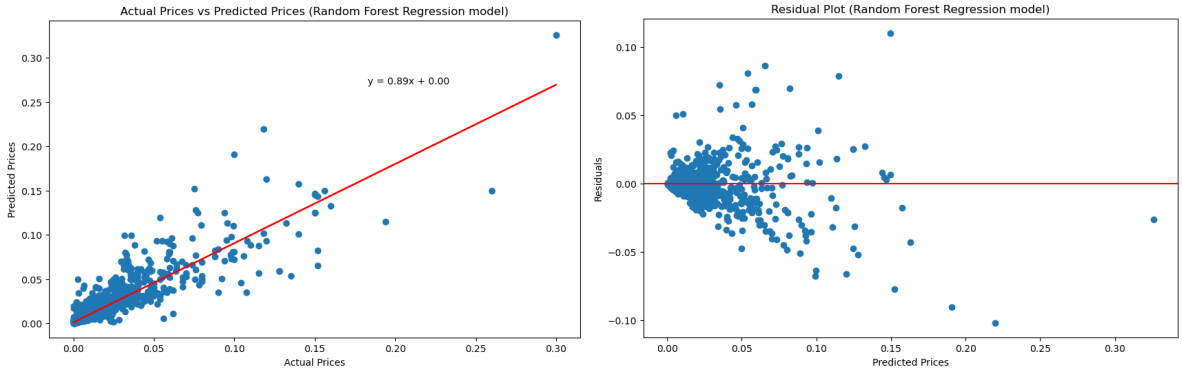


Fig. (9) Prediction effect of Random Forest model for all type of housing normalized data

We would like to know what essential features affect the housing price. Hence, we can use the Random Forest regression model to find the important features. First, we plot the importance of each feature from the Random Forest model. Then, we sorted them in descending order by the feature importance. Finally, the top 10 most important features are listed, and the feature importance map is drawn to visualize the feature importance in Figure (10). For example, the importance of the sqft feature is the highest, about 0.3108. This could mean that a home's number of square feet is crucial in predicting prices. Then the importance of a bachelor's degree is 0.1439, ranking second. This could mean that the number of people with bachelor's degrees, perhaps in a particular region or neighborhood, is also essential in predicting house prices. In the

third ranking, the importance of the count of restaurants, supermarkets, and bathrooms are 0.0901, 0.0768, and 0.0769, respectively, which also play a specific role in affecting the housing price. Finally, some features such as Property Type Condo, Theatre count, High School Graduate, cinema count, and subway count are less critical but still contribute to the predicted results.

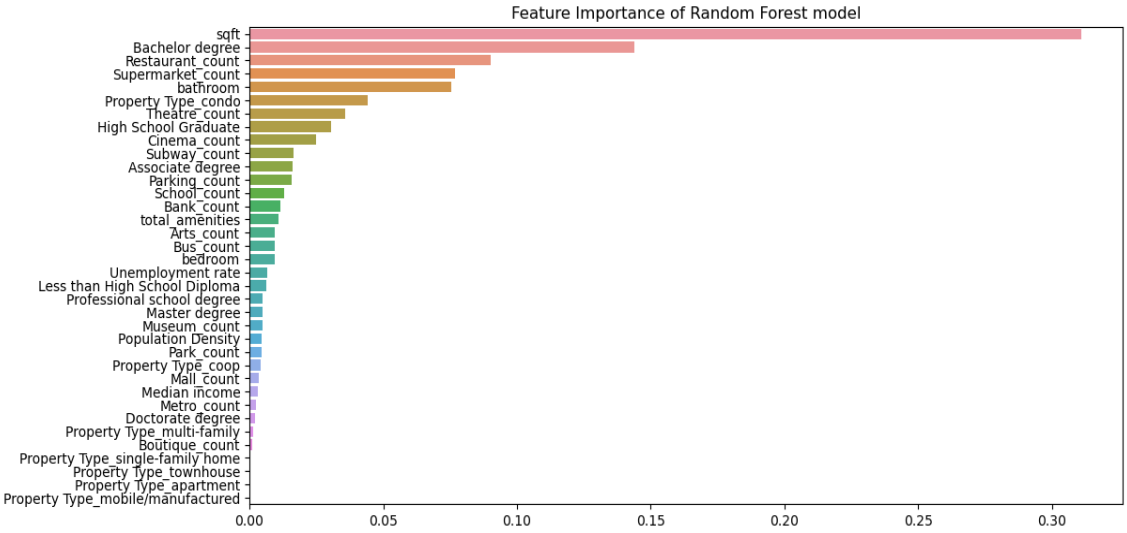


Fig. (10) Important features for all types of housing price prediction

As for the 2b2b housing analysis result, we found some important features that affect the 2b2b median housing price from the correlation analysis. According to the results of the correlation analysis, we found that the absolute value of correlation coefficients of some features is greater than 0.5. When the absolute value of correlation coefficients of some features is greater than 0.5, it indicates that there is a strong linear relationship between these features and the target variable (2b2b median housing price). In other words, these features have a larger effect on the target variable and may be more explanatory in predicting 2b2b median house prices. From Figure (11), education level is strongly positively correlated with the 2b2b median housing price, among which the correlation coefficients of bachelor's degree (0.78), master's degree (0.70), professional school degree (0.72) and doctor's degree (0.66) all show that higher education is usually correlated with higher housing price. This may be because higher education is often linked to higher incomes, allowing people to afford more expensive property. There is also a strong positive correlation between median income (0.56) and median home price, meaning that residents of higher-income areas are more likely to buy properties at higher prices. In addition, population density (0.45) has a moderately positive correlation with median home prices, likely due to a better location and job opportunities that attract more people and thus drive-up home prices. On the other hand, those who have not completed a high school diploma (-0.45) or only a high school diploma (-0.79) are negatively correlated with the median housing price, this might indicate that both groups generally have lower incomes and result in relatively lower home prices

in the areas where they live. Finally, there is a strong negative correlation between the unemployment rate (-0.64) and the median home price, possibly because a high unemployment rate indicates poor economic conditions in the area, which decrease home prices.

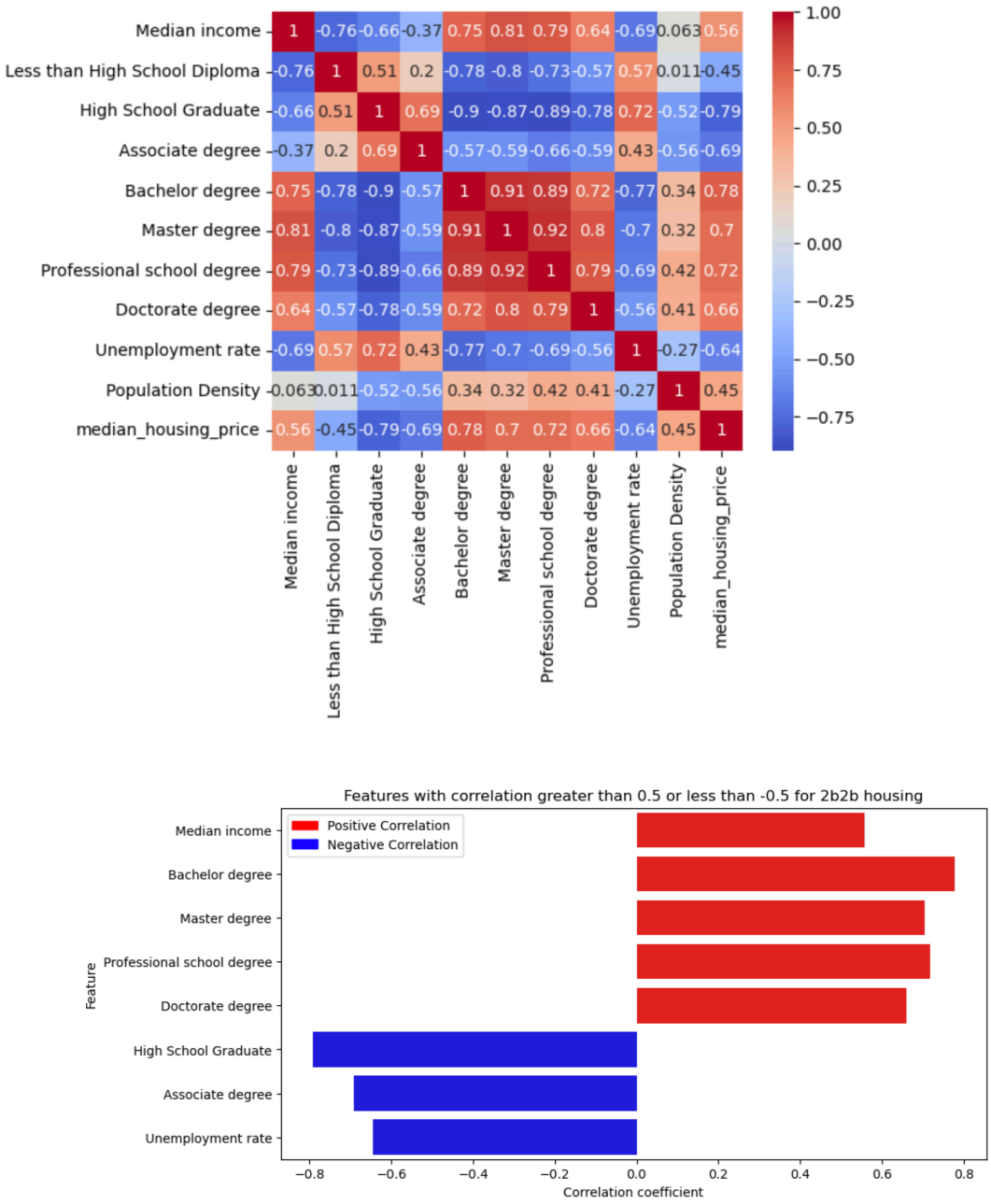


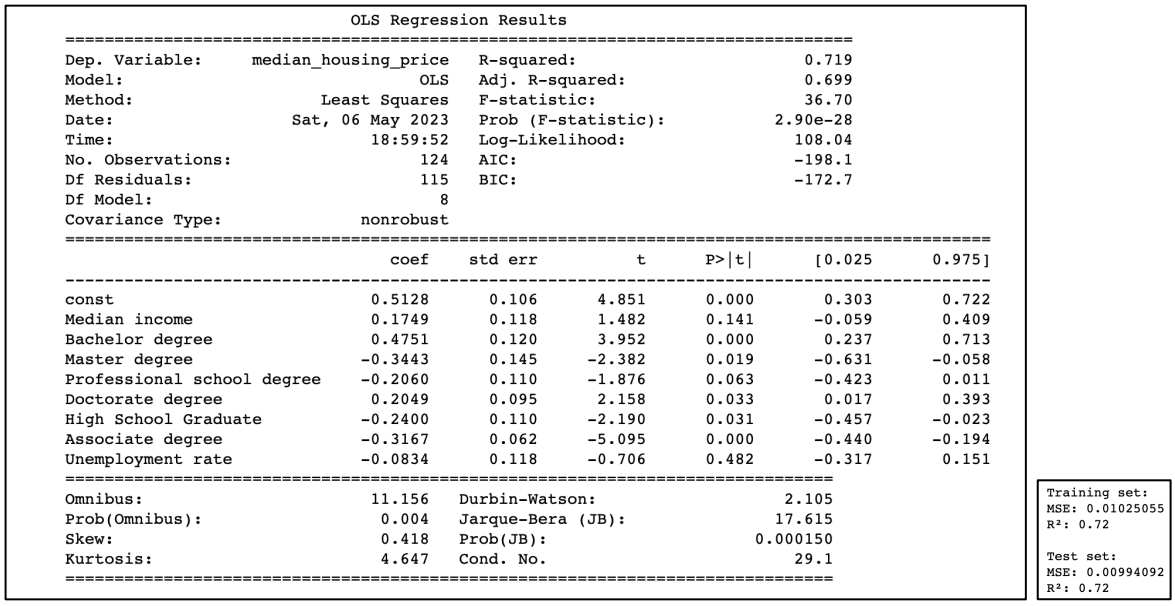
Fig. (11) Important features for 2b2b median housing price

In predicting 2b2b median housing price, because the 2b2b median price data set is a relatively small data set for a machine learning model, we use OLS linear regression and adjust

model parameters to prevent serious overfitting, a relatively simple model for 2b2b median price prediction. As for the 2b2b housing feature correlation analysis result, we found some critical features whose absolute coefficients are more significant than 0.5. We use those important features from correlation analysis to do OLS regression. Based on the OLS regression results in Figure (12), I select the most important features with large coefficients and a p-value less than 0.05 in the OLS model. In our OLS model, these selected features are bachelor's and associate degrees. The coefficient of the bachelor's degree is 0.4751. If all other variables are held constant, we expect the 2b2b median housing price to increase by 0.4751 units for every additional unit when there is an additional person with a bachelor's degree. On the other hand, the 2b2b median housing price is expected to decline by 0.3167 units per associate degree increase, likely because people with associate degrees live in less affordable areas. The equation of the prediction model is as follows:

```
OLS Regression Equation:
y = 0.5128 + (0.1749 * Median income) + (0.4751 * Bachelor degree) + (-0.3443 * Master degree) + (-0.2060 * Professio
nal school degree) + (0.2049 * Doctorate degree) + (-0.2400 * High School Graduate) + (-0.3167 * Associate degree) +
(-0.0834 * Unemployment rate)
```

The above equation shows the relationship between predicted house prices and various characteristics (median income, the proportion of different degree holders, and the unemployment rate). The coefficient of each feature represents the predicted change in house prices when the characteristic value is increased by one unit while the other features are held constant.



OLS Regression Results

Dep. Variable:	median_housing_price	R-squared:	0.719
Model:	OLS	Adj. R-squared:	0.699
Method:	Least Squares	F-statistic:	36.70
Date:	Sat, 06 May 2023	Prob (F-statistic):	2.90e-28
Time:	18:59:52	Log-Likelihood:	108.04
No. Observations:	124	AIC:	-198.1
Df Residuals:	115	BIC:	-172.7
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5128	0.106	4.851	0.000	0.303	0.722
Median income	0.1749	0.118	1.482	0.141	-0.059	0.409
Bachelor degree	0.4751	0.120	3.952	0.000	0.237	0.713
Master degree	-0.3443	0.145	-2.382	0.019	-0.631	-0.058
Professional school degree	-0.2060	0.110	-1.876	0.063	-0.423	0.011
Doctorate degree	0.2049	0.095	2.158	0.033	0.017	0.393
High School Graduate	-0.2400	0.110	-2.190	0.031	-0.457	-0.023
Associate degree	-0.3167	0.062	-5.095	0.000	-0.440	-0.194
Unemployment rate	-0.0834	0.118	-0.706	0.482	-0.317	0.151

Omnibus:	11.156	Durbin-Watson:	2.105
Prob(Omnibus):	0.004	Jarque-Bera (JB):	17.615
Skew:	0.418	Prob(JB):	0.000150
Kurtosis:	4.647	Cond. No.	29.1

Training set:
MSE: 0.01025055
R^2: 0.72

Test set:
MSE: 0.00994092
R^2: 0.72

Fig. (12) OLS regression result

The last analysis for 2b2b housing is that we create an interactive map based on latitude and longitude to show the price distribution of 2B2B properties. First, we calculate the price value per square foot and then divide the price per square foot value into different colors based on the quantile (0.2, 0.4, 0.6, 0.8). For example, 0.8 quantile means that eighty percent of home prices are worth less than or equal to this quantile per square foot value. Each color indicates a different 2b2b median housing price level. From the map in Figure (13), we can see each region's price level. The higher 2b2b housing prices concentrate on Manhattan's lower side.

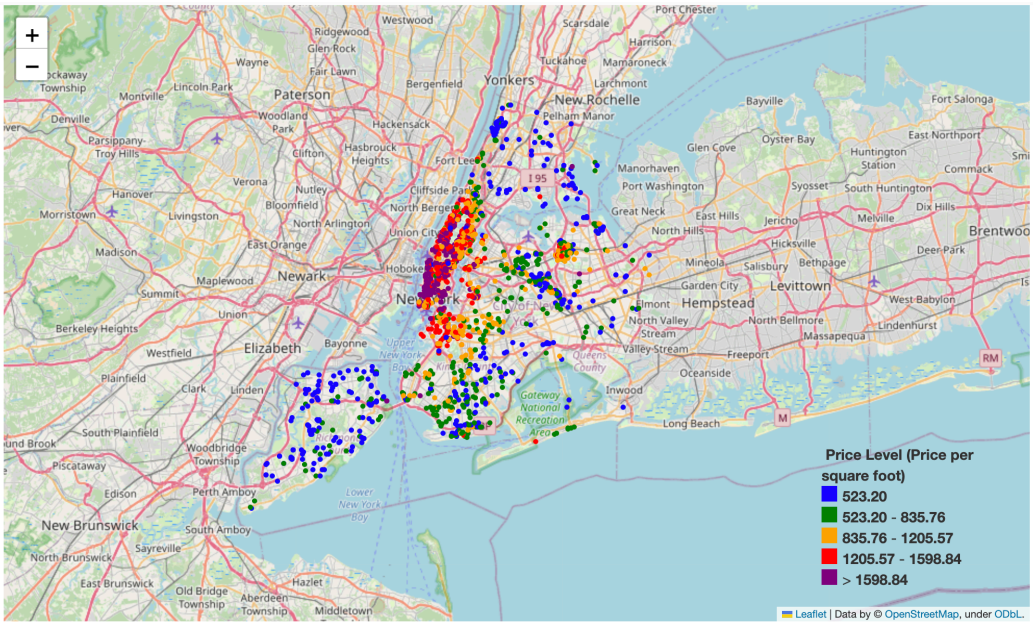


Fig. (13) 2b2b Housing Price Distribution Map

6. Insights

This analysis project is divided into two parts: one for all housing types and the other for 2b2b housing. Based on our analysis, we found some interesting insights. First, in all types of housing analysis, linear regression could not capture non-linear information and patterns in the data. So, we use the Random Forest model to predict house prices. The random forest model has the best performance on our data set, which may be because there are nonlinear relations and too many features in our data, and such relations cannot be well-fitted in Lasso regression. In addition, we find that square feet (sqft), the number of people with a bachelor's degree, and other essential features, including the number of restaurants, supermarkets, and bathrooms, play a crucial role in predicting house prices. These results differ from our previous hypothesis that the number of transportation facilities significantly affects housing prices.

On the other hand, in our 2b2b housing analysis, we find a strong positive correlation between high education level and 2b2b median housing price. This may be because higher education is often associated with higher income levels, enabling people to buy more expensive property. Combining all types of housing and 2b2b housing analysis results, we found that the percentage of bachelor's degree is a common factor affecting housing prices. Furthermore, the

2b2b home price distribution map shows the higher prices in Lower Manhattan. In terms of location, the Lower Manhattan area is the core business district of New York City, with many financial, commercial, and recreational facilities. That makes a living and working on the site more accessible, pushing up property prices. In addition, lower Manhattan is home to high-income groups, who are more likely to buy property in the area, driving up prices. And then there's the limited land in lower Manhattan, the limited space available for construction. Housing prices naturally rise as the population grows and the demand for better locations increases. Finally, the lower Manhattan area has many famous tourist attractions, such as the Statue of Liberty, the World Trade Center site, the Wall Street Bull, etc. These attractions attract many tourists and bring a unique historical and cultural atmosphere to the area. That makes the place extra attractive to home buyers. Buyers who value historical and cultural values may be more willing to pay a higher price.

7. Conclusions

After analyzing all types of housing price data, we found that the Random Forest prediction model is more suitable than the decision tree and Lasso linear regression. The Random Forest has the highest performance in the whole data set. The housing space, the percentage of bachelor's degrees, the number of restaurants and supermarkets, and the number of bathrooms are important features for predicting housing prices. In predicting 2b2b median housing price, the proportion of bachelor's degrees and the proportion of associate degrees are the most important features in predicting 2b2b median housing price. Notably, 2b2b housing prices are higher in lower Manhattan.

8. Limitations

There are some limitations in this project. One limitation is that this project currently focuses on the New York City real estate market. We need to determine whether our method and technique can be applied to real estate markets in other regions. Second, we only collect information about current selling real estate and may need to expand the data volume further to improve the model's generalization ability. The last limitation is that we only selected a certain number of features for analysis. However, there may be other important features that should be considered.

9. Challenges

In the prediction results and residual graphs of the Random Forest model, there are still some data points whose residual is significantly different from other data points. These points could be data errors or phenomena the model failed to capture. Dealing with these outliers and potential data errors poses a challenge because they can affect the accuracy and predictive ability of the model. Furthermore, the timeliness of housing data means that our analysis and predictions may change over time. Moreover, although the Random Forest model performs well on our data set, there is still room for optimization.

10. Future work

In future work, we can also collect data on properties that have been sold to increase our data volume. Second, we can explore more characteristics related to housing prices (such as whether the housing price is located on the main road) to improve the model's performance further. In addition, we also consider using more complex nonlinear models, such as support vector machines, to capture nonlinear relationships in the data better. The advanced model is better suited to handle complex data sets and provide more accurate predictions. Moreover, to ensure the timeliness and accuracy of the analysis results, future work needs to update the data regularly and retrain the model to adapt to the real estate market changes. At the same time, we can conduct more in-depth cleaning and preprocessing of the data to eliminate potential outliers and noise to improve the stability and reliability of the model. Furthermore, we also collect time-series home price data, allowing us to consider historical trends and seasonal changes in home prices to capture the dynamics of the housing market. Finally, we would like to use cross-validation methods to evaluate the stability of the model.