Factors Affecting Rent Cost in Manhattan

Hongkai Wu, Ruogu Li, Hansen Ding, Heqing Huang, Yuexuan Li, Yaqing Wang

1. Introduction

New York City, one of the global centers of economy, finance, education, and tourism, has experienced a huge influx of foreigners in the past year. And Manhattan, as the center of New York, is perceived most deeply. Therefore, choosing the right apartment is an issue that has to be faced by the incomers living in New York City.

In the report, we will analyze and discuss the rental data of Manhattan in the past year with different models to get the impact of different factors on apartment rent. In addition to that, we hope that the conclusions drawn from the analysis of the data will provide some guidance to potential renters, which will enable them to choose the right apartment according to their needs and budget.

2. Exploratory Data Analysis

There are 18 different columns in our dataset including rental price, apartment size, number of bedrooms, floor number, age of buildings, minutes to transit, whether there is a doorman, gym, and some other binary data. We want to know how each piece of data is related to one other.

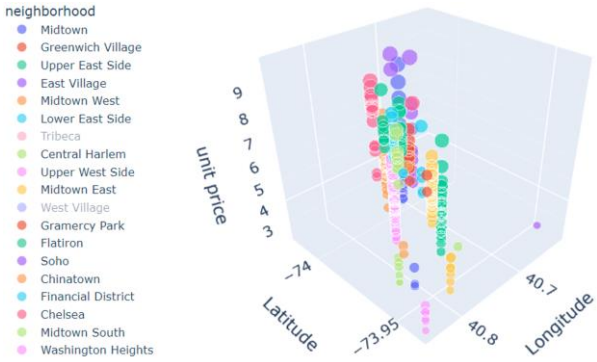


Fig. 1: Apartment District Distribution Model

Before heading into actual analysis, we need to explore the data. Our dataset has 18 variables, which is not a small number. Therefore, we had to first process the dataset to check for missing values, outliers, and bad data. In Figure 1, We also partitioned Manhattan into 5 different sections to limit the impact of the variable of geographic location on our subsequent calculations.

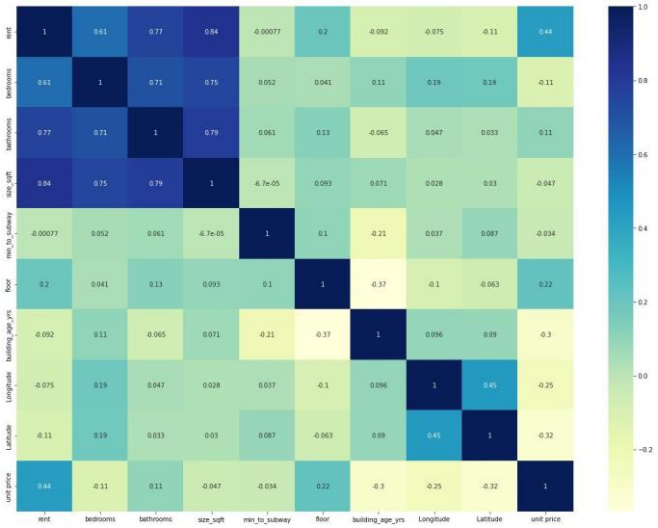


Fig. 2: Parameter Correlation Heat Map

A preliminary exploration of our data correlation is shown in Figure 2. We can see that the variables "Number of Bedrooms", "Number of Bathrooms", and "The Size of Apartment" are positively correlated with rent cost.

3. Methodology

After preprocessing the data, we brought up five different methods that could potentially be used in analyzing and processing the data, and eventually used four of them, including linear regression, nonlinear regression, logistic regression, and decision trees. The linear regression model and the nonlinear regression model will be used to analyze the non-binary parameter variables in the model[1]. We will compare the accuracy of the two models by a reference standard, so as to adopt the model that performs better of two as a reference. In the logistic regression section, we will obtain some coefficients related to the variables and determine their correlation through the logistic regression model[2]. Finally, we will analyze other binary data using decision trees. The above method will cover all the variables in the data set. We compare and complement these four methods to make our conclusions more precise.

3.1 Linear & Non-Linear Regression

Both of them are used to pursue a most fitted line in the dataset hence to determine the relationship between any two variables, and from there, predict the future value from the calculated function.

In the project we used linear regression to analyze the effect of the number of bedrooms, the number of bathrooms, and room size each has on the rental price, and the result showed that all of them have positive correlations, which means that the more bedroom or bathroom there are in an apartment, or the bigger the size of the apartment is, the higher rental prices are. To test how accurate our regression models are, we introduced a value called R-squared, which is used to determine how well the data fit the regression model[3]. It's usually a decimal number between 0 and 1, the closer to 1 the more accurate a model is. Our R-squared values of each linear regression model turned out to be 0.407, 0.592, and 0.736, which were all decent ratios, especially for room sizes vs. rental prices.

When we were trying to find other correlations we found out that the impact from room size was too big so we decided to use unit price as our new dependent variable to offset the differences that the room size had brought to us by using rental price divided by room size. After that, we partitioned Manhattan into over 30 neighborhoods and assigned coordinates in terms of longitude and latitude to each of them, tempting to see the relationship between location and price. We also used linear regression and found it had negative correlations, which stated the closer a neighborhood is to downtown, Manhattan, the higher unit price is.

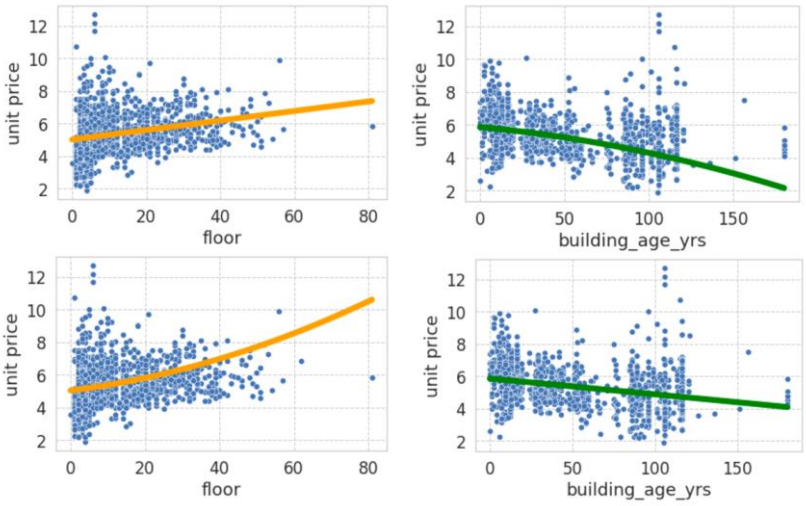


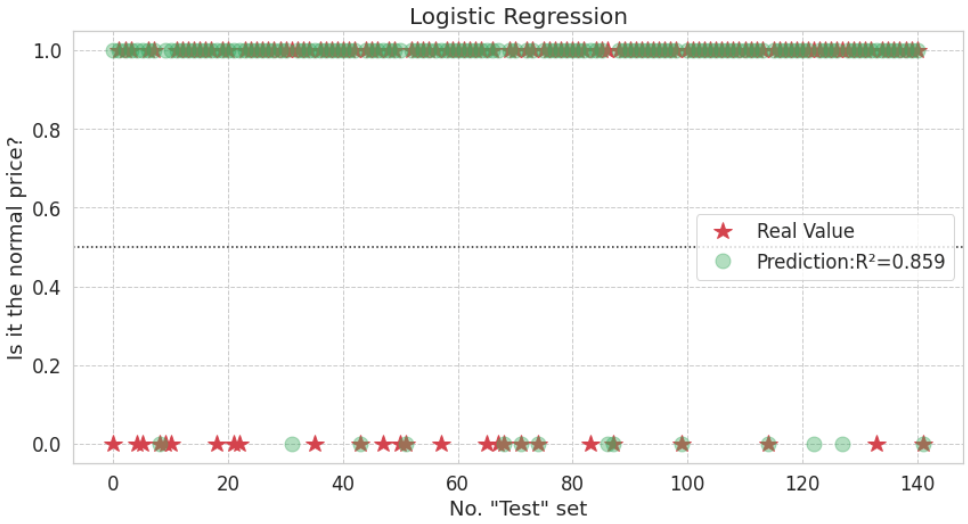
Fig. 3: Linear vs. non-linear regression

For factors like floor number and building ages, we found that their linear correlations to prices weren't strong enough. So the group tried both linear and nonlinear regression. Since there isn't any R-squared in nonlinear regression, we convert our R-squared values from unit price vs. floor number and building ages into RMSE (Root-Mean-Square Deviation) to compare with nonlinear regression[4]. It turned out that nonlinear regression outperformed in both situations. Despite both graphs showing the higher floor the more expensive and the newer buildings come with higher prices (Fig 3).

We then plotted the prediction lines of the model to see how well the model predictions fit the observed data. We need a better way to assess how well the model fits the data than visual observation. The metric used is the coefficient of determination, which indicates the percentage of the total variation in the dependent variable captured by the model.

3.2 Logistic Regression

For the logistic regression method, we defined all prices below the upper quartile of the unit price as normal prices and set it to the value of 1. Then set all prices above 6.048 $/ft2 to the value of 0. Next, divide the whole data set into training and test set, and use logistic regression and fitting functions to train model parameters to obtain coefficients.



Accuracy of LR Classifier: 0.8591549295774648

	precision	recall	f1-score	support
nomal	0.71	0.38	0.50	26
expensive	0.88	0.97	0.92	116

Fig. 4: Logistic regression model accuracy criterion

The coefficient represents the correlation with the variable. In the results Similar to a heat map, the closer the value is to 0, the weaker the correlation is [5]. Figure 4 shows the distribution of the predicted and actual values. We calculated the value of R-square as 0.859. because the R-square represents the accuracy of this model. So here also the accuracy of the linear regression model is proved.

3.3 Decision Tree

Decision trees are another way to classify the data. Use Gini and Entropy index as the error metric and split on the features-"size_sqft", "min_to_subway", "floor", "building_age_yrs", "has_doorman", "has_elevator", "has_dishwasher", "has_patio", "has_gym", "Longitude", "Latitude" to get the cut-point of each factor. Figure 6 shows the specific results of the decision tree.

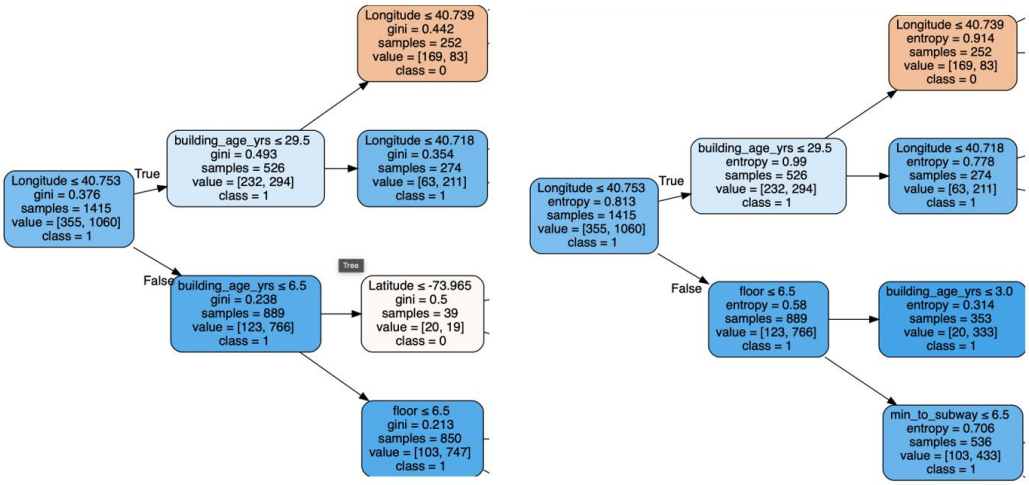


Fig. 5: Decision tree by using Gini index & Entropy index

According to the characteristics of the features, the logistic regression was easier to make decisions for all data. The tree would get too big if we don't define the depth. So the decision tree is better for local analysis[5].

4. Conclusion & Future Extension

Our results from each chart show that room size has the greatest impact on rental prices, followed by the number of bedrooms, and then the number of bathrooms. Typically, the larger the apartment, the higher the price will be, and the number of bedrooms and bathrooms requires a larger room size. Other factors, such as the number of floors, age of the building, dishwasher, gym, and doorman, will also have a smaller impact on the rent. Of note is the "location" variable. Both graphs negatively correlate with unit price, meaning that living in lower Manhattan will cost more than living in an uptown apartment.

So far, every variable collected has had some effect on rental prices. Some variables are statistically significant and economically significant, some are neither.

- From the previous results, it appears that the size of the apartment has the largest effect on the rent.
- From the linear regression results, the lower the longitude at which the apartment is located, the higher the price.
- From the results of the logistic regression, the presence of a doorman, elevator, dishwasher, and gym in the apartment affects the rent. If an apartment includes all of these amenities and the unit price is less than $6.048 per square foot, we would consider it a cost-effective apartment.

In this Manhattan apartment rent data processing and analysis, we did not use a neural network due to the limited amount of data in the dataset we were given. And this amount of data is not good enough to train a more accurate neural network model for a multivariate problem. Therefore our next goal is to collect

more data to build a neural network that explains apartment rents and sales prices, adding more binary data for tenants to give them more options when using logistic regression to find out if an apartment has a high performance-to-price ratio.

Reference

[1] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[2] Kleinbaum, David G., et al. *Logistic regression*. New York: Springer-Verlag, 2002.

[3] Seber, George AF, and Alan J. Lee. *Linear regression analysis*. John Wiley & Sons, 2012.

[4] Amemiya, Takeshi. "Non-linear regression models." *Handbook of econometrics* 1 (1983): 333-389.

[5] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

[6] Song, Yan-Yan, and L. U. Ying. "Decision tree methods: applications for classification and prediction." *Shanghai archives of psychiatry* 27.2 (2015): 130.