Factors that affect the rent of apartments of Manhattan in New York City

1 Introduction

“Start spreading news, I am leaving today. I am gonna be a part of it. New York, New York”. No matter whether one has decided to move to the city of the big apple to start a new life or just do an internship lasting for a short period, the first thing that pops out of the to-do list is to find an apartment to move in. And as the borough with the highest population density,(27,812/km2 in 2015), the data regarding the rental market of apartments in Manhattan is believed to be representative. Therefore, studying factors that influence the rent in that region and coming up with a regression model not only helps people with different housing needs estimate the budget needed for housing but also helps developers determine which factors they should focus on the most to increase the rent.
To start with, the paper (Amenyah, I. D., & Fletcher, E. A. (2013).) suggested that the number of bedrooms present in the unit had the strongest correlation with rental prices. But since the article used data from Ghana, it would be interesting to see if the factor works differently in a region in a completely different continent.
Then the area available is also a critical factor to consider, as suggested by the article (Wickramaarachchi, N. (2015).). In New York City, construction style vary from building to building and a unit with more bedrooms does not mean more space available. For example, in New York City, there are townhouses with a whole floor being a single unit versus two to three-bedroom units scattered on a single floor in a high-density residential building. Given the independence of two factors in the context, both factors can be included in this study.
Other than the factors above, mentioned in this paper (Sirmans, S., Sirmans, C., & Benjamin, J. (1989).), access to public transportation also plays a significant role on rent. So, in this study, we will take the minutes needed to walk to the nearest MTA subway station as a measure to testify.
Based on the data of 3539 observations from StreetEasy- one of New York City’s largest real estate marketplace- on Kaggle, a deeper understanding of how the factors like above serve as a determinant of the rent price of an apartment in Manhattan can be done.

2 Methods

First, the data is imported into R, omitting any row with incomplete data. And then, data are randomly grouped into 2: train data with 2800 entries for analysis and the rest in test data.

2.1 Varaiable Selection

The first step is Exploratory Data Ananlysis. With histograms, boxplots, and scatter plots of the both predictor and response variable, check if the assumptions of linearity in parameters and normality in the response variable are satisfied and do the fixing(transformation) if necessary.

A first model is fitted using every numerical variable that could be a potential predictor variable based on common sense, graphs plotted during EDA, and related articles. Keep filtering out the least significant predictor variables until no more insignificant ones are left and the model is the final model. Next, compute the adjusted R square, SSRes AIC, AICc and BIC values of the first model and final model and make sure the final model is better(model with larger adjusted R square and smaller SSRes, AIC, AICc, BIC is a better model).

Then multicollinearity is checked to see if there’s any correlation between predictor variables in the final model using vif. If there is any predictor variable with vif > 5, it should be removed.

After that, random predictor variables are taken out of the final model to form a reduced model, and a partial F test is done to determine which model is better.

2.2 Model Violation Check and Diagnostics

First, 2 conditions are checked. Condition 1 is that conditional mean response is a single function of a linear combination of predictors, which can be checked by doing a plot of response against the fitted final model and observing if all the points are close or on the line. Condition 2 is that the conditional mean of each predictor is a linear function with another predictor. And this can be checked by selecting numerical variables and plotting them one against another and observing the pattern.

After the two conditions are met, 4 assumptions can be checked using residual plot and qqplot. First, if there is any sysmatic, cluster, or fanning pattern in residual plots mean that some of the first 3 assumptions are violated. Then, the non-linear pattern in the qqplot indicates the violation of the final assumption. And then fix the variables if necessary.

Further checks include problematic observations including outlier points, leverage points, and inferential points.

2.3 Model Validation

A test model is first fitted using the test data from the first step. Then the plots used to check conditions and assumptions are repeated on the test model. If the coefficients, adjusted R square,SSRes, AIC, AICc, BIC, and the plots are similar to both model, the final model fits well.

3 Result

3.1 Description of data:

	minimum	1stquantile	median	mean	3rdquantile	maximum
bedrooms	0	1	1	1.35	2	5
bathrooms	0	1	1	1.37	2	5
size_sqft	250	611.8	800	942.2	1150.2	3680
min_to_subway	0	2	4	5.01	6	43
rent	1300	3150	4000	5167	6000	20000

Table 1: Summary of the data of the numerical variables in train data

Rent in Manhattan is surely high, judging by the fact that the lowest rent in the data is 1300 USD per month and the average rent is 5167 USD per month. Even though Metropolitan Transportation Authority has an extensive subway network in Manhattan island, from the summary above, it seemed that there are units quite far from the nearest subway station or it could be an error in the data. Another interesting fact is that, in New York City, it’s common to find 0 bedroom or so-called Studio units in which tenant has to share the bathroom with other studio units on the same floor and that’s where the minimum of 0 bathroom observation came from. The minimum size of a unit in this dataset is 250 sqft, which is likely to be a 0-bedroom unit, and the maximum size of 3680 sqft, which is like the area of an entire house, could be a large unit on the top floors of some luxury building near central park.

	minimum	1stquantile	median	mean	3rdquantile	maximum
bedrooms	0	1	1	1.33	2	5
bathrooms	1	1	1	1.35	2	4
size_sqft	250	616.5	795	930.2	1100	4800
min_to_subway	0	2	3	4.82	6	43
rent	1443	3182	3990	5033	5995	20000

Table 2: Summary of the data of the numerical variables in test data

All variables of the test data have similar minimum, 1st quantile, median, mean, 3rd quantile, maximum compared to those of the train data.

3.2 Exploratory Data Analysis

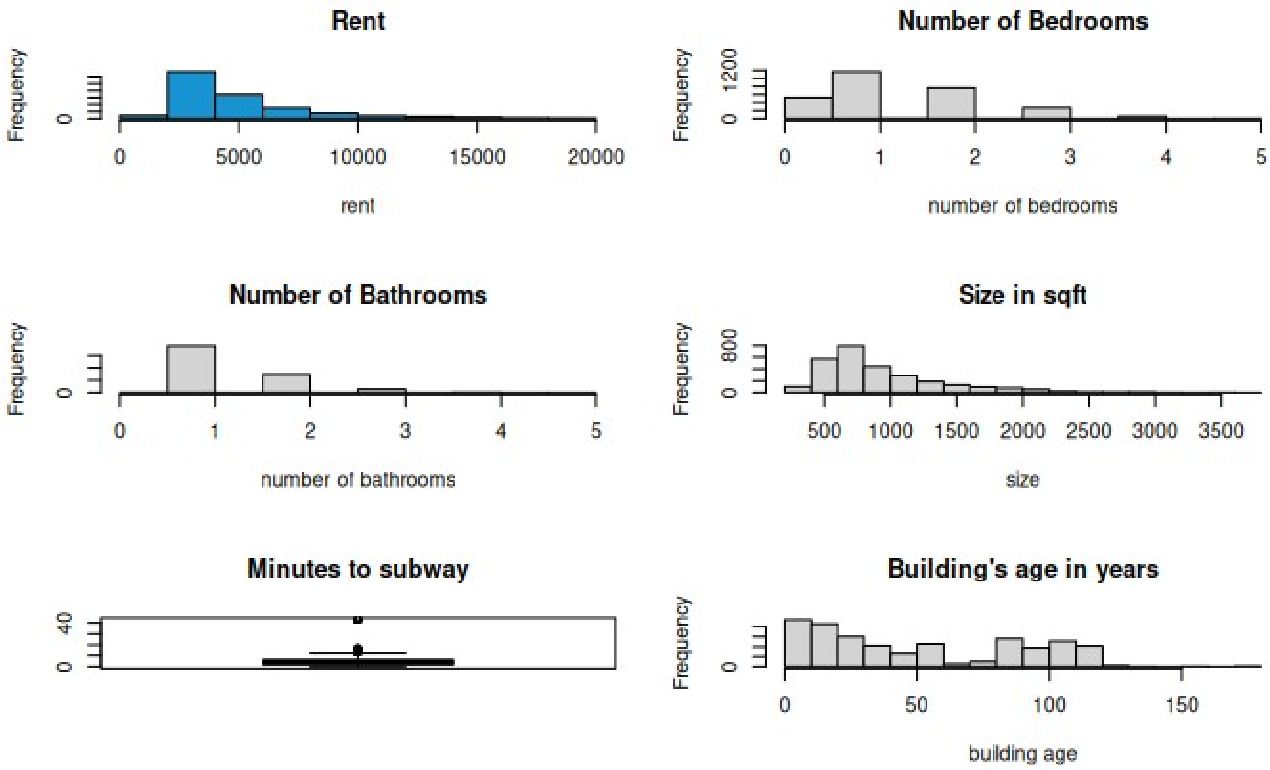


Figure 1: hisogram and box plot of numerical variables

The histograms and boxplots showed the data distribution of observations of numerical variables. And it becomes obvious that the proposed response variable rent is left-skewed and violates the normality assumption.

Fix normality violations by applying log to rent as transformation:

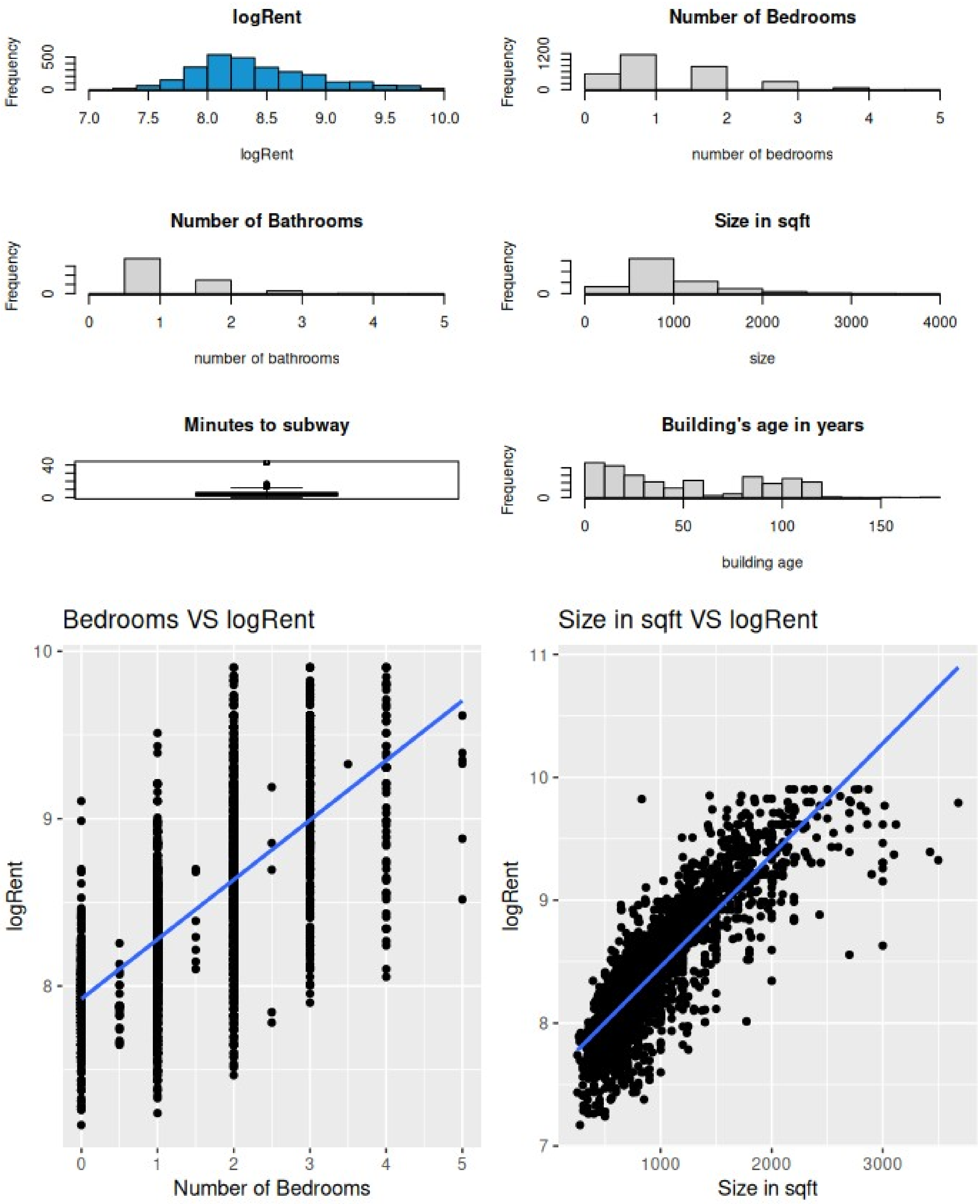


Figure 2: histograms, boxplot, and scatterplot for numerical variables

The histograms, boxplots and scatterplot showed the data distribution of observations of numerical variables where rent is now closer to normal distributed. The histograms of bedrooms, bathrooms, and size_sqft are left skewed, which indicated that most units have 1 bedroom, 1 bathroom, and area between 500 to 1000 sqft. And histogram of building_age_yrs showed another interesting fact about Manhattan-it has always been a construction site with three peaks in the recent 100 years, 1930s, 1970s and 2000s till now. Some of the most famous skyscrapers in Manhattan-Empire State Building, Crysler building, Rockefeller Center- were built in the 1930s, along with the trend of urbanization and more housing available. During the early 1970s, there was a trend lead by New York Port Authority to build a series of new skyscrapers with the most iconic twin towers of the World Trade Center. After 9/11, there were rebuilds of the new World Trade Center complex as well as some other new skyscrapers replacing buildings damaged beyond repairs in lower Manhattan. At the same time, more and more ultra tall skyscrapers-like 432 park ave, central park tower, etc. many being mixed usage towers, which explains why much more housing became available than previous peaks- have been built in midtown, dwarding the Empire State building. The Boxplot of min_to_subway showed that most units are relatively close to subway stations with the exception of the outlier- the 43 minutes walk distance one mentioned ahead. And also two numerical variables are selected to plot scatterplot with response variable, and it could be observed that both plots are going up, satisfying the linearity assumption.

3.3 Process of obtaining final model:

> Step 1: fitting a initial model.

First, as mentioned above, the reponse variable is $log\hat{Rent}$. Then, from common sense, graphs above as well as relevant articles about determinants of house rent, a series of numerical variable from the dataset is selected as predictor variables.

$log\hat{Rent}$ = 7.602 +0.020[number of bedrooms] +0.109[number of bathrooms] + 0.001[size in sqft]-0.005[minutes to subway]-0.002[building’s age in years]+0.004[floor]+0.015[has dishwasher]-0.007[has doorman]+0.023[has elevator]-0.021[has gym]+0.010[has roofdeck]

> Step 2: Filtering out insignificant variables from model1, model2 is formed.

Predictor variables [has dishwasher], [has doorman], [has elevator], [has gym], and [has roofdeck] have p-values greater than 0.05, thus they are insignificant. After removing them, a new model-model 2- is fitted using the remaining predictor variables.

$log\hat{Rent}$ = 7.608+0.019[number of bedrooms]+0.108[number of bathrooms]+0.001[size in sqft]-0.005[minutes to subway]-0.002[building’s age in years]+0.004[floor]

	SSres	Rsq	Rsq_adj	AIC	AIC_c	BIC
Model 2	159.859	0.780	0.780	-8004.625	-8004.573	-7953.126
Model 1	159.524	0.781	0.780	-8000.512	-8000.381	-7919.326

Table 3: Adjusted R^2, AIC, BIC of model 1 and model 2.

Because the model 2 has smaller AIC, AICc and BIC, model 2 is proved to be better.
And since there is no more insignificant variables left, no further filtering down is needed

> Step 3: multicollinearity check

	bedrooms	bathrooms	size_sqft	min_to_subway	building_age_yrs	floor
vif	2.778	3.231	3.769	1.047	1.247	1.198

Table 4: Summary of vif of predictor variables in the final model

Since all the predictor variables has vif < 5, they are not correlated.

> Step 4: Fit a reduced model 3 and do a partial F-test using anova

2 random predictor variables, bathrooms, floors, are removed and a reduced model model 3 is fitted using the remaining predictor variables.

$$log\hat{Rent} = b_0 + b_1[number\ of\ bedrooms] + b_2[size\ in\ sqft] + b_3[minutes\ to\ subway]$$
$$+b_4[building's\ age\ in\ years]$$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Model 2	2793	159.86				
Model 3	2795	168.92	-2	-9.06	79.18	2.20E-16

Table 5: ANOVA tables of full model and the reduced model

p-value = $2.2e^{-16}$ < 0.05, thus the full model model 2 should be selected as the final model.

3.4 Goodness of model check:

(1) check if the model satisfies condition 1 and condition 2.

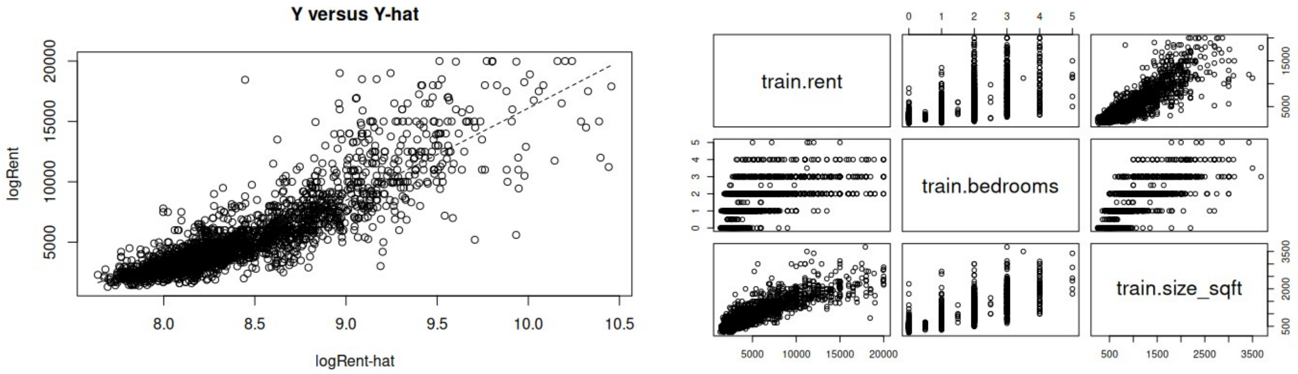
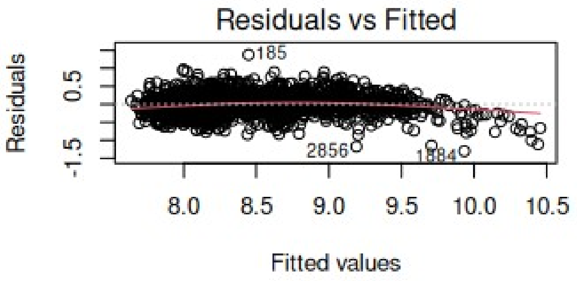
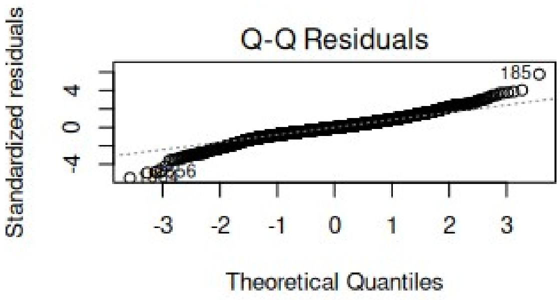


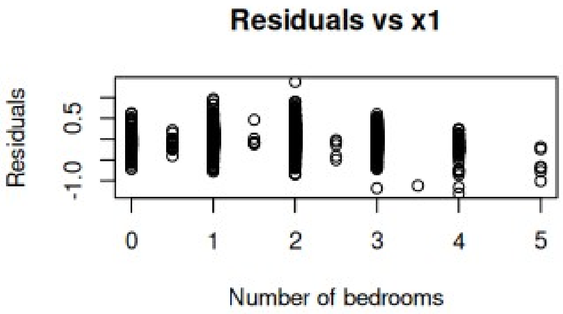
Figure 3: Plots used to check condition 1 and 2

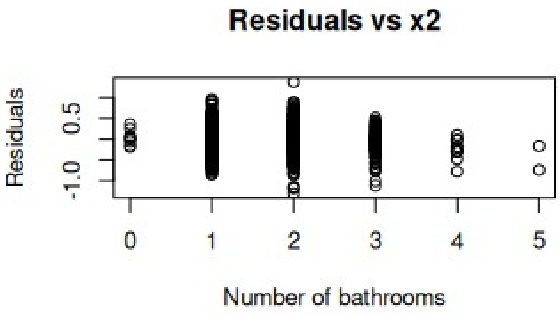
The plot on the left checks the relationship of y and $\hat{y}$, since the points are fitting the line, condition 1 is met. Then, for the numerical variables shown in the plot on the right, no non-linear relationship is observed, thus condition 2 is satisfied as well.

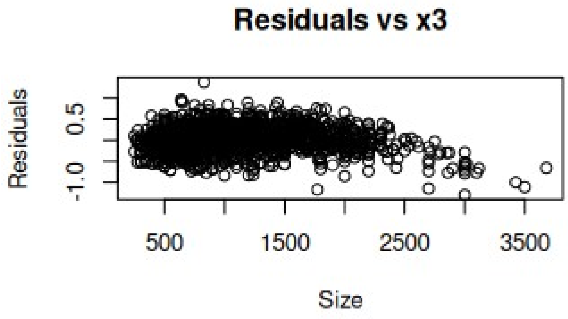
(2) check the assumptions using residual plot and qqplot

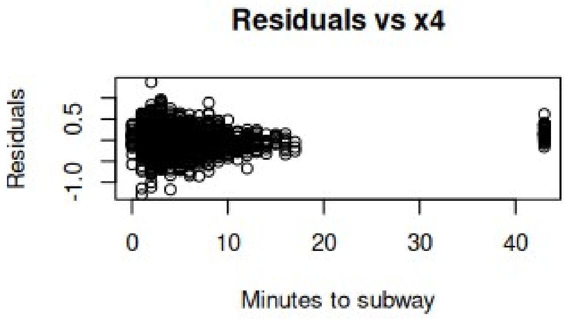

Residuals vs Fitted
185
2856
1884
Residuals
0.5
-1.5
8.0
8.5
9.0
9.5
10.0
10.5
Fitted values

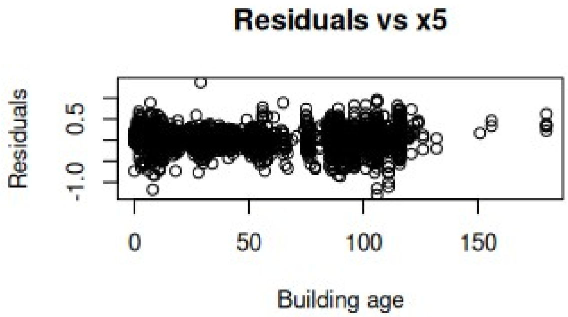

Q-Q Residuals
185
Standardized residuals
4
0
-4
-3
-2
-1
0
1
2
3
Theoretical Quantiles

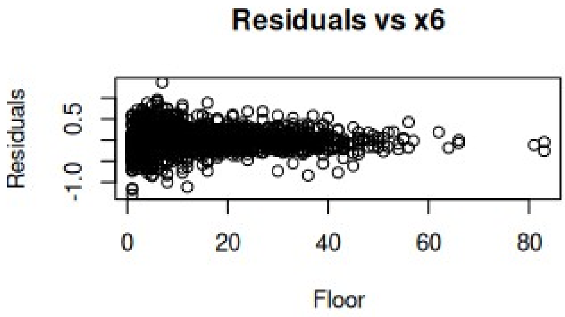

Residuals vs x1
Residuals
0.5
-1.0
0
1
2
3
4
5
Number of bedrooms


Residuals vs x2
Residuals
0.5
-1.0
0
1
2
3
4
5
Number of bathrooms


Residuals vs x3
Residuals
0.5
-1.0
500
1500
2500
3500
Size


Residuals vs x4
Residuals
0.5
-1.0
0
10
20
30
40
Minutes to subway


Residuals vs x5
Residuals
0.5
-1.0
0
50
100
150
Building age


Residuals vs x6
Residuals
0.5
-1.0
0
20
40
60
80
Floor

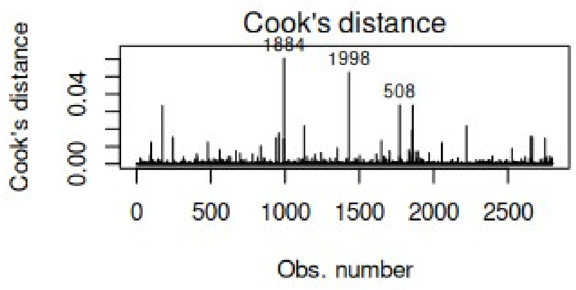


Figure 4: residual plots and qqplot of the train model

In the residual plots, first, no systematic, cluster, or fanning pattern is observed, thus linearity of coefficients, independence of errors, constant variance of error are all met. In the qqplot, nearly a straight line is observed, so the normality is met as well. Thus all 4 assumptions are satified in model 2. In the graph of cook distance, 3 points with the cook distance of 1884, 1998, and 508 might be influential points.

(3) problemetic observations, leverage points

There are 168 outlier points whose standard residuals are not between -2 and 2. There are also 188 leverage points in model 2 whose fitted values are larger than the values in the training set but there is no influential points in model 2.

3.5 model validation

	Intercept	bedrooms	bathrooms	size_sqft	min_to_subway	buidling_age_yrs	floor
Model 2 (Final Model)	7.608	0.019	0.108	0.001	-0.005	-0.002	0.004
Testing Model	7.533	0.035	0.235	0.001	-0.003	-0.002	0.010

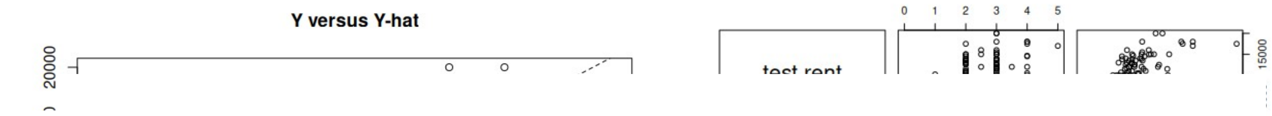
Table 6: Summary of coefficients of train vs test data generated models

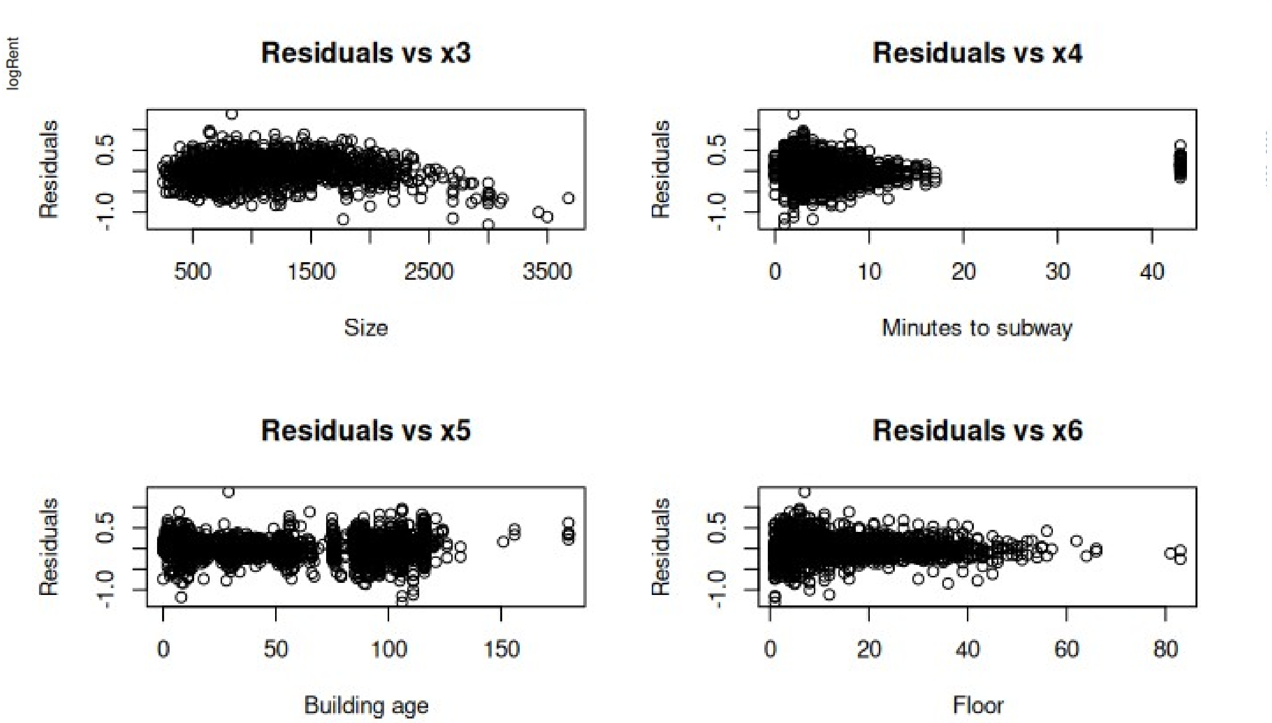
From the table, difference between coefficients of the variables between two models, which is minimum.

	SSres	Rsq	Rsq_adj	AIC	AIC_c	BIC
Model 2 (Final Model)	159.859	0.780	0.780	-8004.625	-8004.573	-7953.126
Testing Model	42.252	0.762	0.760	-11712.670	-11712.620	-11661.170

Table 7 : Summary of toal variacne, adjusted R^2, AIC, BIC of model generated by train data versus the one generated by test data.

From the table, difference in AIC as well as BIC is minimum.





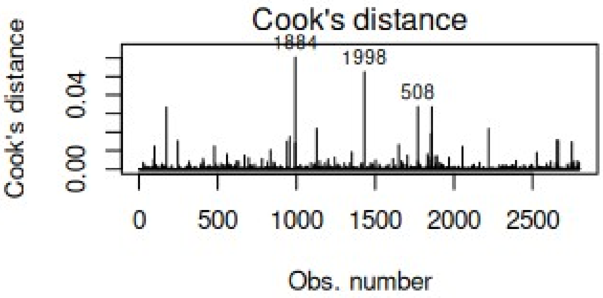


Figure 5: Plots used to check condition 1 and 2 , residual plot and qqplot of test model which are all similar to the ones of final model.

Since all the coefficients, values like total variance, adjusted R^2, and plots, the final model is considered ok.

4 Discussion

4.1 Final model Interpretation and Importance:
$log\hat{Rent}$ = 7.608+0.019[bedrooms]+0.108[bathrooms]+0.001[size_sqft]-0.005[min_to_subway]-0.002[building_age_yrs]+0.004[floor]

From the final model, some estimation of the monthly rent can be done with different requirements. For example, one prefers to live by themselves and requires 0 bedrooms, 1 bathroom, 300 sqft of area, 5 minutes walk from the subway, and 5th floor given the building is 10 years old. The estimated logRent will then be 7.991 and the estimated rent would be 2954.25 USD. However, if someone is planning to share a 3 bedroom apartment with 2 others with 1 bathroom, assuming size is 1000 sqft, 5 min to the subway station, 5th floor, and also 10 years old, the estimated total rent is 6298.8 USD and one's share became 2099.35 USD which is about 850 USD cheaper.

At the same time, the model is also important to developers of Manhattan as it shows how significant each determinant(predictor) is compared to others. For instance, bathrooms turn out to be the most significant predictor of all. One additional bathroom is followed by a 0.108 increase in the estimated logRent. Assume the original estimated rent is 3000 USD with an estimated logRent of 8.00, then an additional washroom is estimated to boost the estimated rent to 3320 USD. Thus developers could consider building units with more private washrooms in order to increase the potential rent.

Since the final model both helps new comers to New Yorker estimate their rent given their requirements and developers understand how to maximize their rent, the research fulfille its purposes.

4.2 Limitation of the Analysis

To start with, numerous outlier points and leverage points indicate that the final model cannot fully fit the data. Thus removing outlier points would be helpful. As for bad leverage points, either investigate the integrity of these points or fit a different model are the valid methods. Also, though it’s mentioned in introduction that the size of unit in square feet is independent from the number of bedroom in New York City, it’s probably not in case in many of the other regions and should be considered strongly correlated in general. Thus either one of them should be included without the other or a linear mixed model should be used instead. Finally, number of minutes to the subway is a relatively less significant predictor variable in this model, because there are extensive subway lines and even more intense stations across manhattan. But in other places less urbanized, access to public transportation may be a much more significant factor, especially given one doesn’t want to drive long distances daily.

Reference

Amenyah, I. D., & Fletcher, E. A. (2013). Factors determining residential rental prices. Asian Economic and Financial Review, 3(1), 39-50.

Wickramaarachchi, N. (2015). Determinants of the rental value for residential properties; A land owner's perspective for boarding homes.

Sirmans, S., Sirmans, C., & Benjamin, J. (1989). Determining apartment rent: the value of amenities, services, and external factors. Journal of Real Estate Research, 4(2), 33-43.

Sirmans, G., & John, B. (1991). Determinants of market rent. Journal of Real Estate Research, 6(3), 357-379.

Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (2019). Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. Land use policy, 82, 657-673.

Cui, N., Gu, H., Shen, T., & Feng, C. (2018). The impact of micro-level influencing factors on home value: A housing price-rent comparison. Sustainability, 10(12), 4343.

Singla, H. K., & Bendigiri, P. (2019). Factors affecting rentals of residential apartments in Pune, India: An empirical investigation. International Journal of Housing Markets and Analysis.