UNIVERSITY OF
SOUTH DAKOTA

PROJECT ANALYSIS REPORT

Predicting New York Housing Price

Riti Dahal

Course Code : DSCI 724

Course Name : Data Mining for Managers

Instructor : Dr. Tiahrt

INTRODUCTION

The main objective of the project is to predict property prices in New York City. , Thus, the predictive model helps to identify the risks and help buyers make informed decisions. The data set used in this project contains 17 variables that include broker title, type, price, beds, bath, property square feet, address, state, main address, administrative area, locality, sub-locality, street name, long name, latitude, and longitude. However, for better results only 6 predictor variables are used. There are 4799 observations in the data set.

BUSINESS PROBLEM DESCRIPTION

Due to the fluctuating market, it is difficult to analyze the housing prices in bigger cities like New York. Thus, the predictive model helps to provide accurate pricing and helps in understanding the market trends. In addition, it allows buyers to identify the risk and improve their investment plans.

DATA SOURCE

The data set used in the project is obtained from Kaggle which focuses on New York Housing Sales data. The dataset can be extracted from the given link:
https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market

The above link is the original data set used in the project proposal. However, moving forward with the project, the outliers were detected which impacted the analysis part. To make the analysis and prediction models easier to understand, the outlier has been eliminated. Below is a link to the updated dataset that was used for the project analysis:
https://drive.google.com/file/d/1qF84RIkw0gUNqPc8HdeCKGqQ2QF4-zF1/view?usp=sharing

VARIABLE DESCRIPTIONS

Initially, the data set includes 17 variables. However, for convenience, the project only uses 6 variables: Price, Beds, Baths, Property, Latitude, and Longitude.

Response Variable

Price

In this project Price is the response variable that needs to be predicted. Here, the Price represents the monetary value of the properties.

Predictor Variable

Bedrooms

This indicates the number of bedrooms in a property and has a significant influence on housing prices. If the property has more bedrooms, the housing prices become more expensive.

Bathrooms

This indicates the number of bathrooms in a property and has a significant influence on housing prices. Similar to bedrooms, if the property has more bathrooms, the housing prices become more expensive.

Property Size (in Square Feets)

This indicates the property size in square feet. The property can be found in different types such as condos, townhouses, Co-Op, Multi-Family Homes, etc. These types of property have different sizes and can influence the housing price.

Latitude and Longitude

These variables indicate the geographic coordinates of the property. These factors can be used in finding a location, which has a direct effect on its perceived value and price.

MODEL SPECIFICATION

A multiple regression model with Price as the response variable and Bedrooms (Bed), Bathrooms (Bath), Property Size (Size), Latitude, and Longitude as predictor variables can be expressed mathematically as:

$y=\boldsymbol{\beta}0 + \boldsymbol{\beta}1x1 + \boldsymbol{\beta}2x2 +....+\boldsymbol{\beta}pxp + \epsilon$

For model specification, it is calculated as:

$\text{Price} =\boldsymbol{\beta}0 + \boldsymbol{\beta}1(\text{Bed}) + \boldsymbol{\beta}2(\text{Bath}) + \boldsymbol{\beta}3(\text{Size}) + \boldsymbol{\beta}4(\text{Latitude}) + \boldsymbol{\beta}5(\text{Longitude}) + \epsilon$

Here,

y = response variable

β0=y-intercept

β1, **β**2, **β**3, **β**4, **β**5 are the slope coefficients of each independent variable

x1, x2,x3,x4, and x5 are predictors variables. Where,

x1= Bed

x2= Bath

x3= Size

x4= Latitude

x5= Longitude

Residuals:

Min	1Q	Median	3Q	Max
-4.733	-1.273	-0.974	-0.348	61.069

Coefficients:

Variable	Estimate	Std.Error	t_value	Pr_greater_t	Significance
(Intercept)	-8.1E+07	69260000	-1.164	0.24456	
BEDS	-585.3	23960	-0.024	0.98051	
BATH	-656600	32730	-20.061	< 2e-16	***
PROPERTYSQFT	69.99	26.3	2.661	0.00783	**
LATITUDE	-762000	754100	-1.01	0.31234	
LONGITUDE	-1554000	651300	-2.386	0.01709	*
Residual standard error: 3896000 on 4793 degrees of freedom Multiple R-squared: 0.08283, Adjusted R-squared: 0.08187 F-statistic: 86.57 on 5 and 4793 DF, p-value: < 2.2e-16					

Table 1. Regression Coefficient Table for Model 1

The first model was created using target variable i.e. Price and all the predictor variables i.e. Bath, Bed, Property Sqrt, Latitude and Longitude. In this model output, we can see that Beds and Latitude are not statistically significant as their p-value is higher than 0.05. Thus, beds and latitude variable has been excluded because they do not add any value to the model.

Residuals:

Min	1Q	Median	3Q	Max
-4.681	-1.280	-0.978	-0.353	61.107

Coefficients:

Variable	Estimate	Std.Error	t_value	Pr_greater_t	Significance
(Intercept)	-1.4E+08	41150000	-3.325	0.00089	***
BATH	-659000	32620	-20.206	< 2e-16	***
PROPERTYSQFT	68.49	23.78	2.88	0.003996	**
LONGITUDE	-1895000	556600	-3.404	0.000668	***
Residual standard error: 3895000 on 4795 degrees of freedom Multiple R-squared: 0.08263, Adjusted R-squared: 0.08206 F-statistic: 144 on 3 and 4795 DF, p-value: < 2.2e-16					

Table 2. Regression Coefficient Table for Model 2

As we can see, the p-value of Bath, PopertySqft and Longitude is less than 0.05 and are statistically significant. Therefore, these variables are included to create a second model.

The equation for the multiple regression model is represented as:

PRICE = -136.802 + (-6.590e-01 * BATH) + (6.849e-05 * PROPERTYSQFT) + (-1.895 * LONGITUDE) + ε

PREDICTIVE RESULTS

a) Interaction Terms

The interaction term is a variable created after multiplying two independent variables together. With this, we can analyze the relationship between two dependent variables and their impact on one another. Here, the interaction terms can be found between Property Square Feet and Longitude.

Residuals:

Min	1Q	Median	3Q	Max
-4.681	-1.280	-0.978	-0.353	61.107

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	83647498	60742721	1.377079	0.168552
BATH	-656228	32541.59	-20.1658	7.16E-87
PROPERTYSQFT	-106219	21579.94	-4.92211	8.85E-07
LONGITUDE	1087308	821561.4	1.323465	0.185744
PROPERTYSQFT:LONGITUDE	-1437.47	291.8551	-4.92528	8.71E-07
Residual standard error: 3886000 on 4794 degrees of freedom Multiple R-squared: 0.08725, Adjusted R-squared: 0.08649 F-statistic: 114.6 on 4 and 4794 DF, p-value: < 2.2e-16				

Table 3. Interaction Terms

Here, we have an interaction term between Property Sqft and Longitude where the coefficient is (-1437.5) and p-value is (8.71e-07). The model indicates that property sqft has a negative impact on pricing as longitude increases. Further, looking at the R-squared (0.08725), adding an interactive term has not made any significant changes in the model, thus, the interaction terms are not included in the project.

b) Non- Linear Predictor Transformation

In this project, the non-linear transformation is used to improve the model.

Residuals:

Min	1Q	Median	3Q	Max
-4631181	-1288172	-997285	364550	61082346

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-140825352.8	41145569.8	-3.422612774	0.000625413
BATH	-661763.653	32616.31615	-20.28934384	7.11E-88
I(PROPERTYSQFT^2)	0.001178	0.000565	2.08477	0.037143
LONGITUDE	-1950752	556463.5	-3.50562	0.00046
Residual standard error: 3897000 on 4795 degrees of freedom Multiple R-squared: 0.08188, Adjusted R-squared: 0.0813 F-statistic: 142.5 on 3 and 4795 DF, p-value: < 2.2e-16				

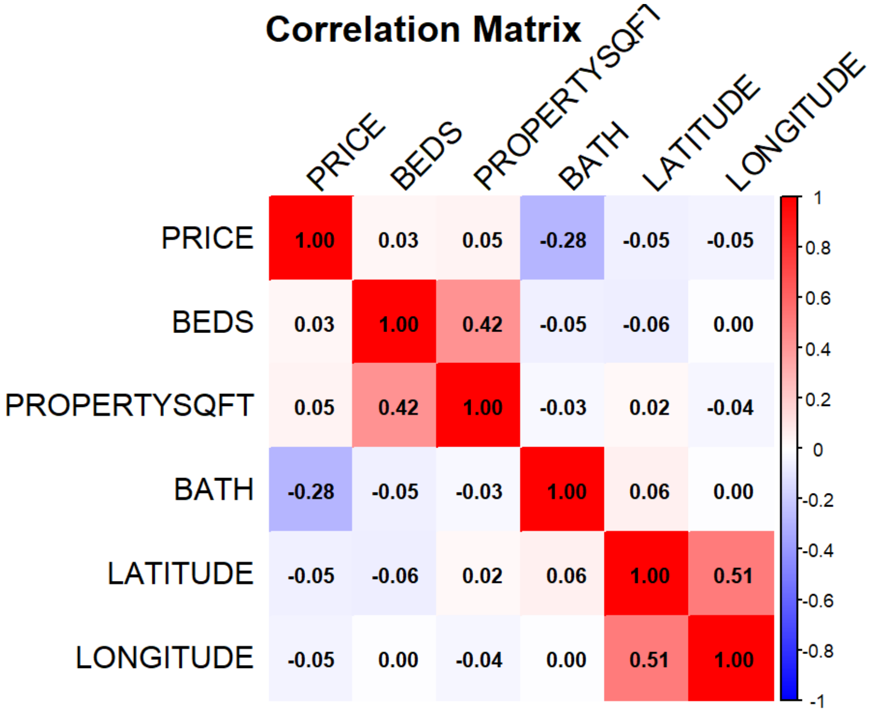
Table 4. Non-Linear Transformation

As per the model output, the price of properties is influenced by all the three variables. The inclusion of squared variables indicates the relationship between price and property size is non-linear. However, there is only a small changes in a R-squared which does not add much in the improvement of the model.

c) Descriptive Statistics

i. <u>Correlation Matrix</u>

A correlation matrix indicates the degree of relationship between two variables. The range of the correlation number is +1 to -1. A positive correlation between two variables is shown by a positive number, whereas a negative correlation is shown by a negative number. On the other hand, two variables are independent of one another and are not correlated if the correlation number is 0. A correlation matrix also provides a detailed understanding of the relationship between multiple variables.



	PRICE	BEDS	PROPERTYSQFT	BATH	LATITUDE	LONGITUDE
PRICE	1.00	0.03	0.05	-0.28	-0.05	-0.05
BEDS	0.03	1.00	0.42	-0.05	-0.06	0.00
PROPERTYSQFT	0.05	0.42	1.00	-0.03	0.02	-0.04
BATH	-0.28	-0.05	-0.03	1.00	0.06	0.00
LATITUDE	-0.05	-0.06	0.02	0.06	1.00	0.51
LONGITUDE	-0.05	0.00	-0.04	0.00	0.51	1.00

Fig 1. Heat Map Showing Correlation Matrix

The correlation matrix as shown in Fig 1 shows the relationships between variables in the data.

As we can see, baths and beds have a weak negative correlation i.e. -0.05, while Property Size (in Square Feet) has a moderate correlation with beds (0.42) and a weak correlation with bath (-0.03). Further, the target variables (Price) and Beds (0.05), Bath (0.08), and Property Sqft (0.11) range between 0 to 0.3, which shows that there is a weak positive correlation between them. Therefore, the connection between these variables is not very strong. Likewise, most of the variables, including the target variable, show weak negative relationships with latitude (-0.05) and longitude (-0.05). This implies that neither the price nor other variables have a strong linear relationship with geographic location (based on latitude and longitude).

	PRICE	BEDS	BATH	PROPERTYSQFT	LATITUDE	LONGITUDE
PRICE	1	0.032771	-0.2805	0.049341505	-0.051885094	-0.047455449
BEDS	0.032771	1	-0.05325	0.420518053	-0.061066528	-0.004451374
BATH	-0.2805	-0.05325	1	-0.027623135	0.05936032	-0.004082684
PROPERTYSQFT	0.049342	0.420518	-0.02762	1	0.022749297	-0.036968031
LATITUDE	-0.05189	-0.06107	0.05936	0.022749297	1	0.514102016
LONGITUDE	-0.04746	-0.00445	-0.00408	-0.036968031	0.514102016	1

Table 5. Correlation Matrix

ii. Means and Standard Deviations

Variables	**Means**	**Standard Deviation**
Price	1869803.32	4.065642e+06
Beds	3.36	2.602252e+00
Bath	2.36	1.724865e+00
PropertySqft	2181.17	2.367243e+03
Latitude	40.71	8.768797e-02
Longitude	-73.94	1.011008e-01

Table 6. Summary of Mean and Standard Deviation

iii. Box Plots

Box Plot also known as box and whisker plot is a graphical method for visualizing data distribution. It indicates a group of numerical data using its quartiles. A box plot summarizes a set of data using five numbers, which are

Minimum: It is the dataset with lowest value.

First Quartile (Q1): It has 25% of the data in the First Quartile (Q1).

Median (Q2): The dataset's midpoint is known as the median (Q2).

Third Quartile (Q3): It has 75% of the data in the Third Quartile (Q3).

Maximum: It is the dataset's highest value, excluding outliers.

The box plot for each variable is shown as below:

a. **Price Box Plots**

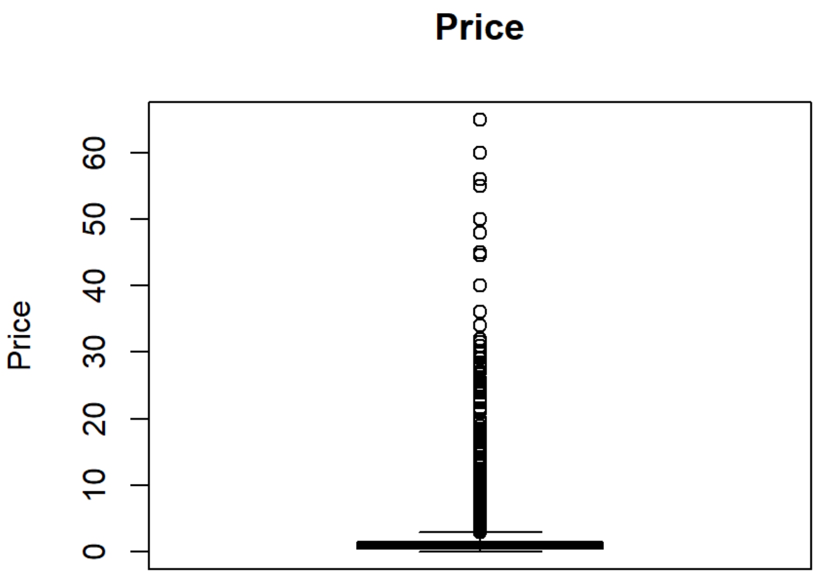


Fig 2. Box Plot for Price

The above box plot for the target variable (Price) shows a high concentration of data towards the lower end and has large number of outlier. Thus, an outlier shows a positively skewed distribution. Further, the box is tightly packed which means the housing prices are within a small range.

b. Beds Box Plots

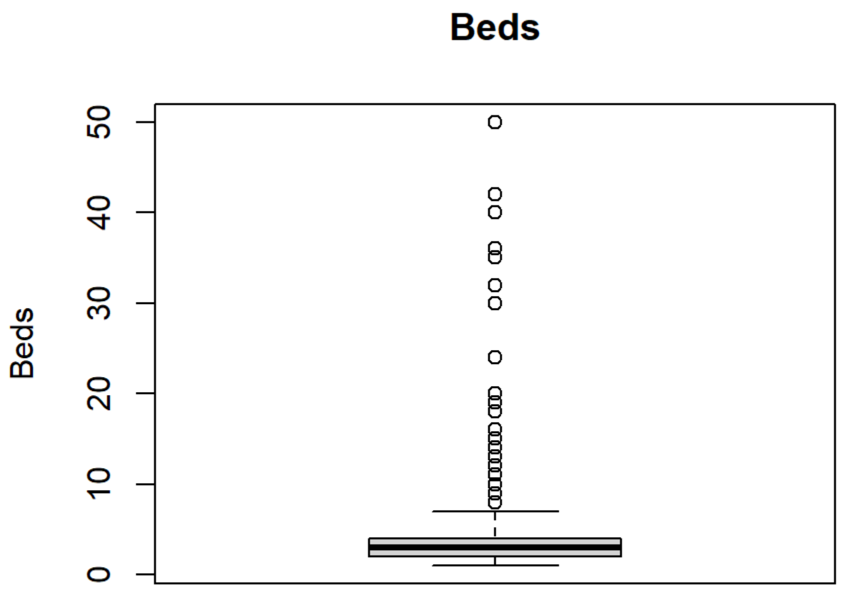


Fig 3. Box Plot for Beds

The above box plot for beds shows that the median is close to 2 or 3 which means most of the properties in New York City have two or three bedrooms. Also, whiskers show an average of six beds. However multiple outliers above this range indicate large or unusual properties. This shows a positively skewed distribution with most of the properties centered at the lower end.

c. Bath Box Plots

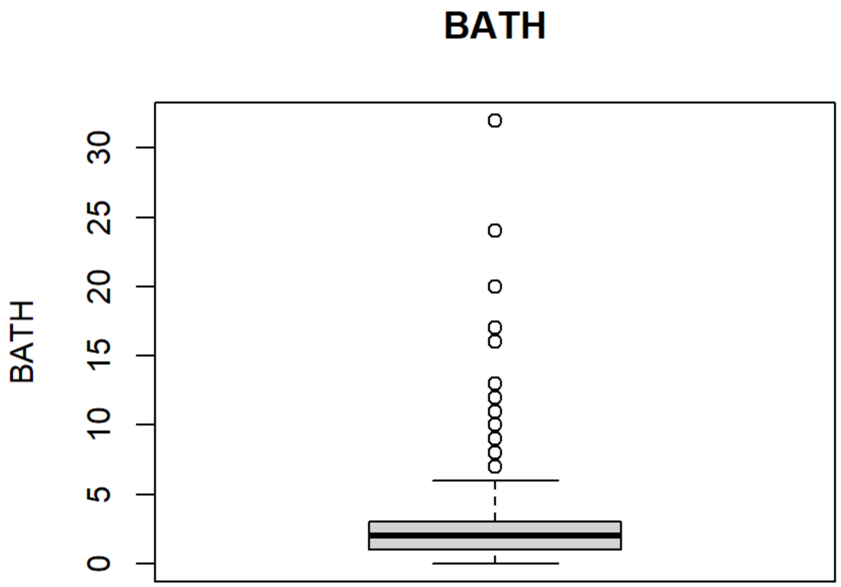


Fig 4. Box Plot for Bath

The above box plot for Baths shows that there are multiple outliers and the majority of the Bath is located at the lower end. Here, the whisker stretches upward and the median is closer to lower quartile which indicates it is right-skewed.

d. Property Sqft Box Plot

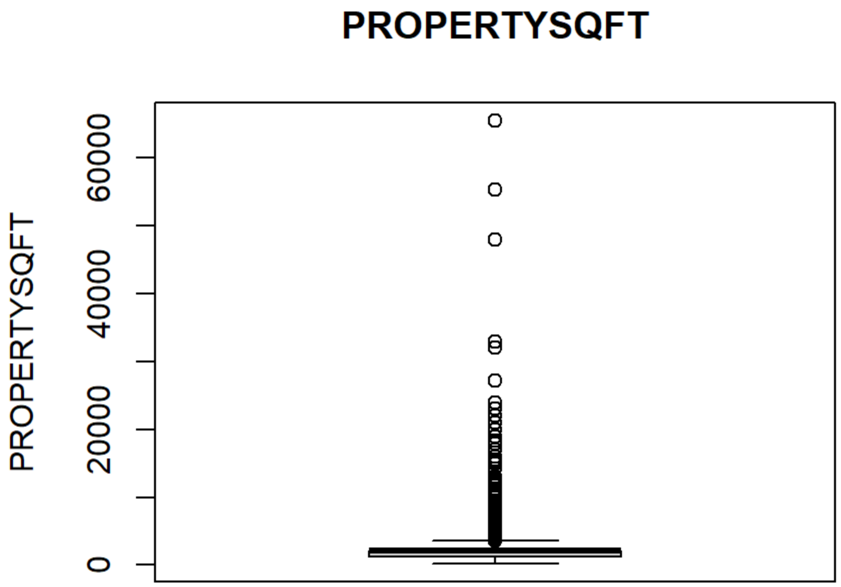


Fig 4. Box Plot for Property Size (in Square Feet)

The above box plot for Property Size (in Square Feet) shows a high concentration of values towards the lower end and a high number of outliers. These outliers show that, there are some properties with very large size (in Square Feet).

e. Latitude Box Plot

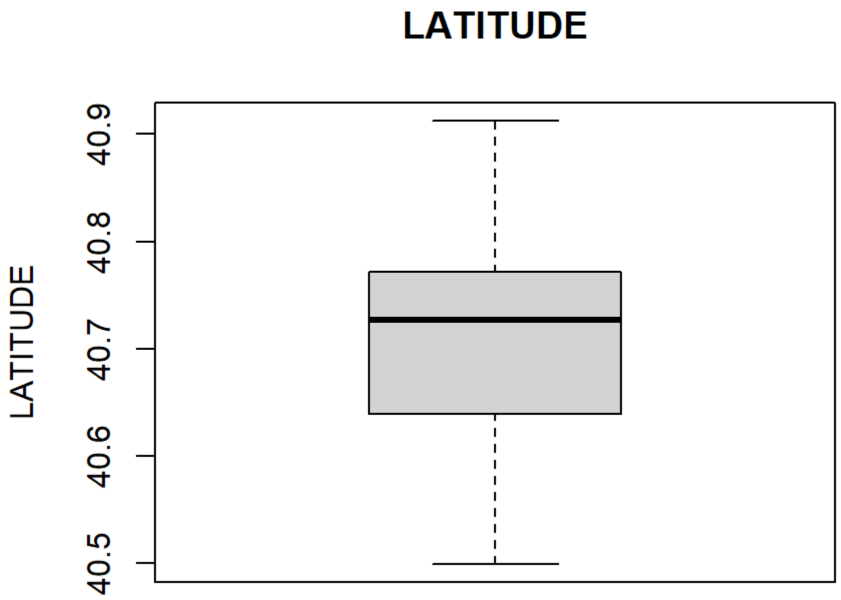


Fig 5. Box Plot for Latitude

The box plot for Latitude shows median value situated within the interquartile range. This indicates a symmetrical distribution. Also, there is no outliers in this case and the whiskers is seen to be evenly distributed.

f. Longitude Box Plot

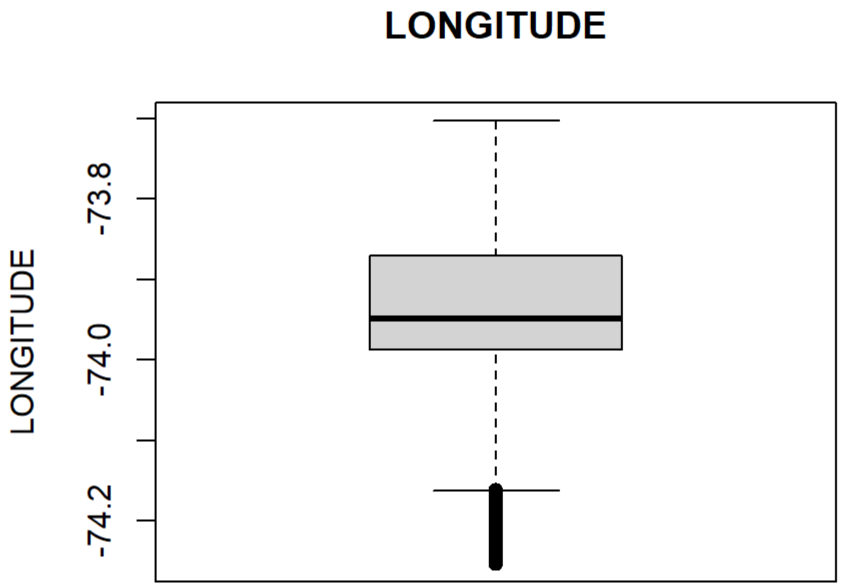


Fig 6. Box Plot for Longitude

The above box plot shows that the median is around -74.0 which indicate to be slightly right-skewed. In the lower end of the whisker, the outliers are present which means it has some extreme longitude values.

iv. Tukey's Five Number Summary

	Min	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Max**
Price	2494	499000	825000	1869803	1495000	65000000
Beds	1.000	2.000	3.000	3.356	4.000	50.000
Bath	0.000	1.000	2.000	2.355	3.000	32.000
Property Sqft.	230	1200	2184	2181	2184	65535
Latitude	40.50	40.64	40.73	40.71	40.77	40.91
Longitude	-74.25	-73.99	-73.95	-73.94	-73.87	-73.70
Price_Conc	0.00249	0.49900	0.82500	1.86980	1.49500	65.00000

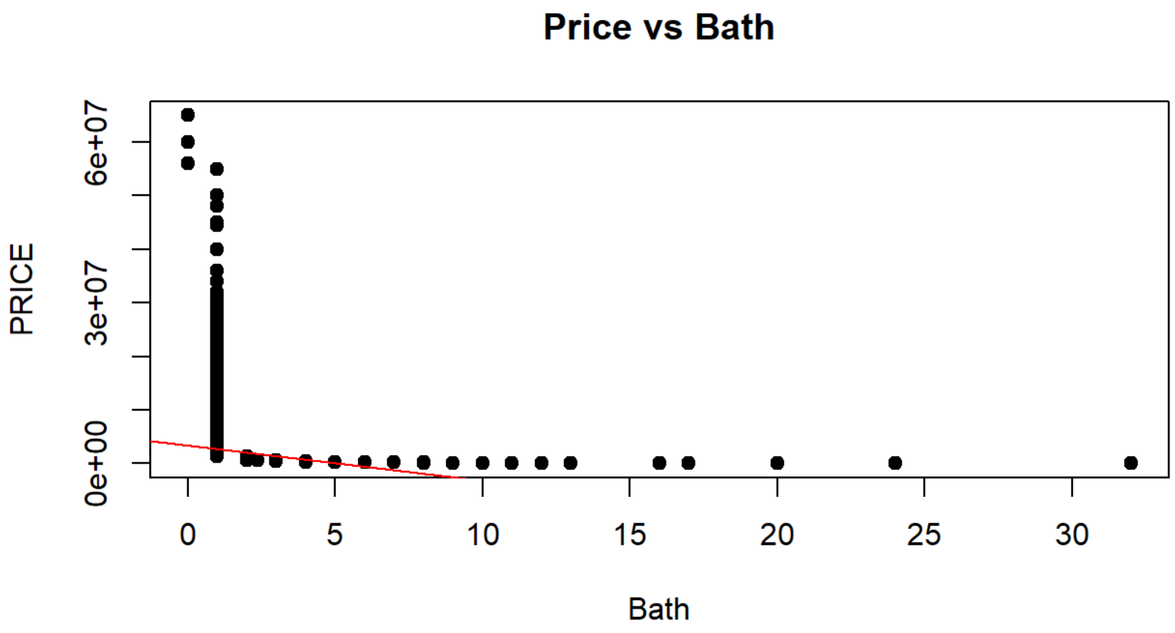
Tukey's Five Number Summaries includes Minimum, first, third and last quartile, mean and maximum.

v. Summary of Evaluation Metrics

R-squared	0.08187596
Adjusted R-squared	0.08130153

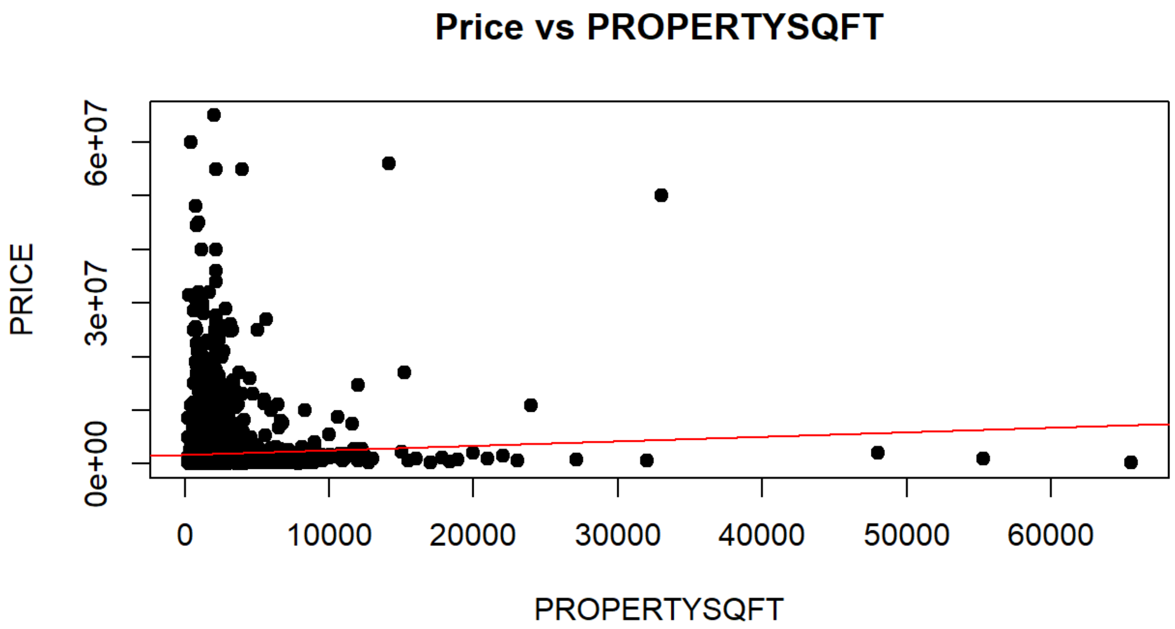
For the evaluation part, this project uses R-squared metrics, also known as coefficient of determination. Here, R-Squared is 0.08187, which is a low value. This shows that only 8.18% of the variance in the dependent variable can be explained by the model. Additionally, to address the over fitting in the model, the adjusted R-squared has also been used. The adjusted R-squared is 0.0813, which means there is no significant difference between R-squared and adjusted R-squared. Thus, both the value shows a weak relationship between predictors and dependent variables.

vi. Plots of individual relationships between response and each predictor



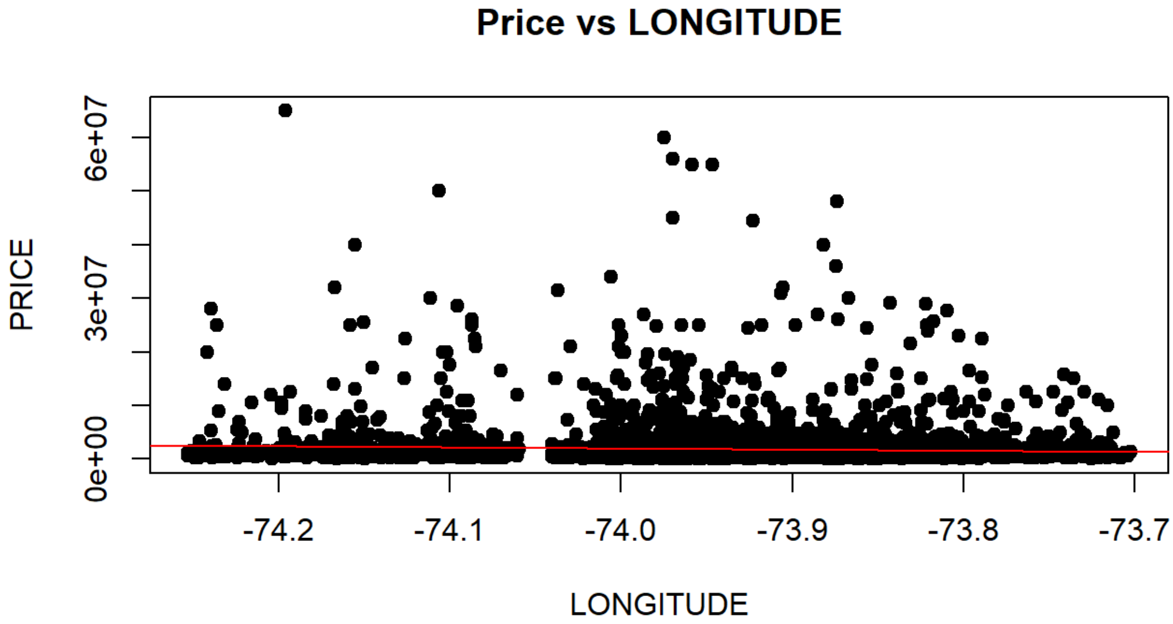
Analysis:

The above plot shows a relationship between a number of bathrooms and price, which indicates a negative linear pattern. Since the numbers of bathrooms are concentrated around lower bound, it shows that a property might have few numbers of bathrooms. Also, when the number of bathroom increases the price decreases. Thus, it shows a negative relationship between number of bathrooms and housing price in New York City.



Analysis:

The above plot shows a relationship between Price and Property Square Foot, which has a slightly positive trend. The plot shows the majority of the properties are smaller in size. We can see that the larger properties have slightly higher price whereas the property with similar size shows a large degree of variability.



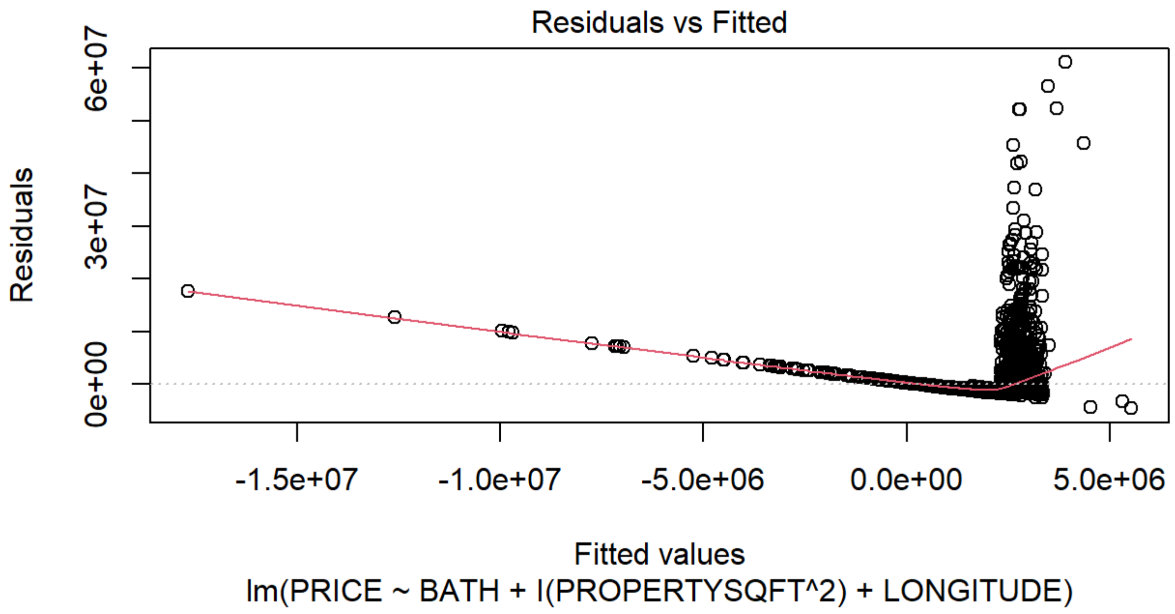
<u>Analysis:</u>

The plot shows a relationship between price and longitude. As the longitude increases, the property price tends to decrease which shows a negative trend. The plot also depicts large number of property are geographically concentrated.

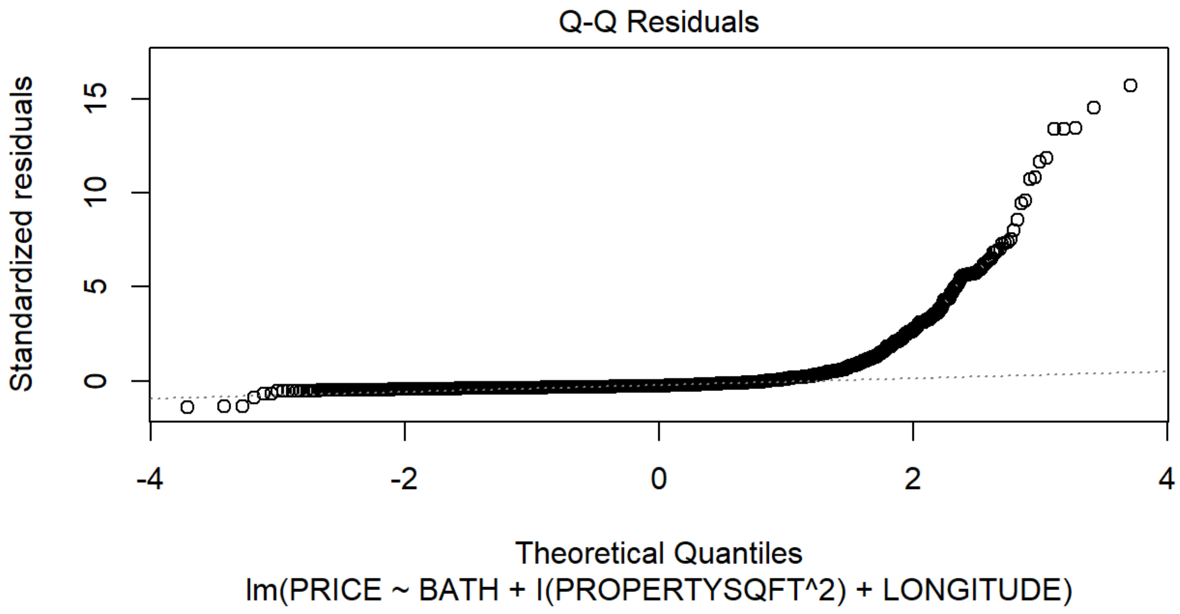
DIAGNOSTIC PLOTS

a. <u>Residuals Vs. Fitted</u>

The plot below shows residual errors versus the predicted values that have a curved pattern. The non-linear relationship between dependent and independent variables shows that at least one of the variables (Bath, Property Sqft., and Longitude) is not linear in the model. Thus, it fails to observe the variability in the price. Further, the residuals should have a constant variance, however, from the plot we can see that the residuals increase as the fitted value increases which violates the homoscedasticity assumption.

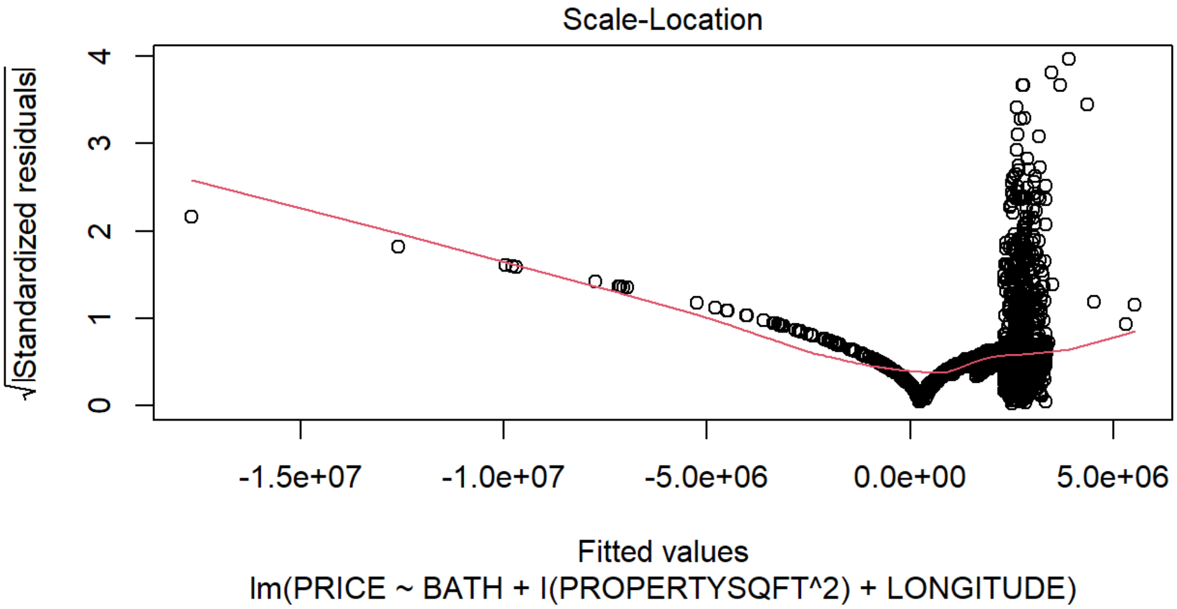


b. Normal Q-Q



The Q-Q plot is used to show if the residuals of a regression model follow a normal distribution or not. From the above plot, we can see that the residuals are separated from the diagonal line, which indicates it is not normally distributed. The upper tail curves upward which indicates the positive outliers and the lower tail curves slightly downward which indicate the presence of negative outliers.

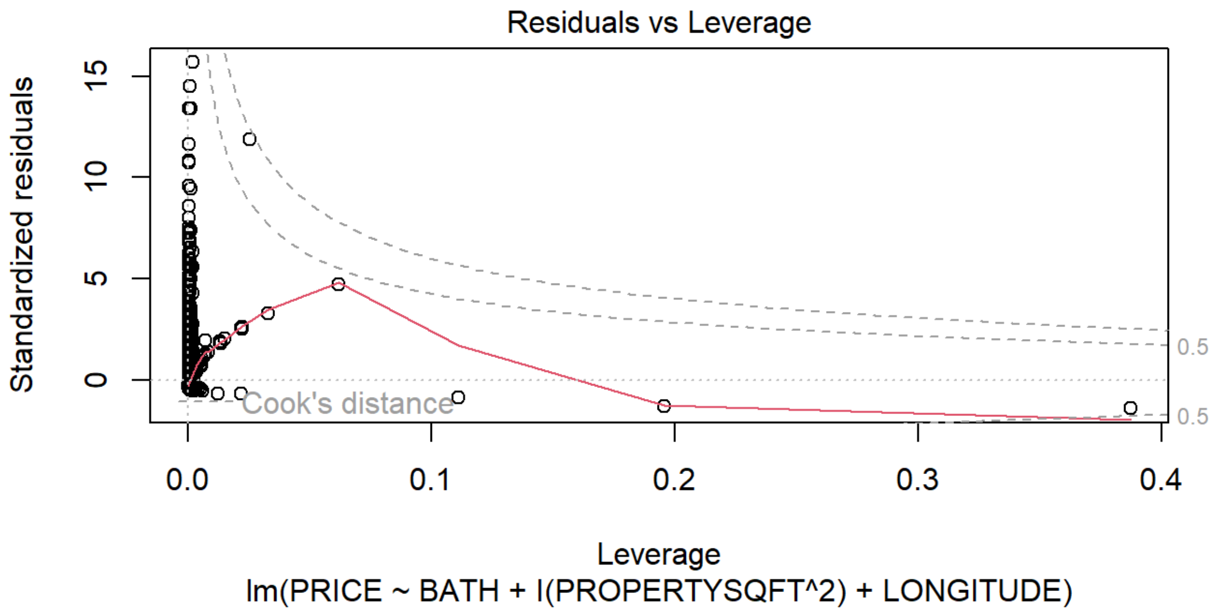
c. Scale Location



The above plot shows the square root of standardized residuals and the fitted values of a regression model. It shows the increase in residuals as the fitted value increases. Thus, the plot shows a non-constant variance also known as Heteroscedasticity.

d. Residuals Vs. Leverage

From the plot below, we can observe that most of the points are clustered around zero which indicates there are no significant changes in the regression model. The points have large standardized residuals and show a significant deviation from the predicted model.



VARIABLE INFLATION FACTOR VALUES

The variance inflation factor is used to check the multicollinearity of the regression model. If the VIF is above 4 and below 0.25, it indicates there is multicollinearity.

Variance Inflation Factor Values

Beds	1.229
Bath	1.008
Property Square Foot	1.226
Longitude	1.371
Latitude	1.382

From the table we can see that, all the variables are close to 1. This shows there is no significant multicollinearity among the predictor variable.

SUMMARY

- In this first model output, Beds and Latitude are not statistically significant as their p-value is higher than 0.05 and are excluded from the model.
- Only three variable Bath, Property Square Foot and Longitude are included in the final

model as their p-value is less than 0.05 and are statistically significant.

- The interaction terms shows a negative impact on a target variable. Thus, adding an interaction terms doesn't make the model better.
- The non-linear transformation does not make significant difference to the model.
- The R-squared shows that only 8.18% of the variance in the dependent variable can be explained by the model.
- It shows a non-linear relationship between dependent and independent variables.
- The normal distribution does not exist in the model.
- Residuals are clustered.
- The outliers have impacted the predictive model performance and statistical tests.
- Lastly, there is no significant multicollinearity among the predictor variable.

APPENDIX

TASK	Estimated Duration	Actual Time	START	END
Initiation				
Data Set Research	1	1	10/19/24	10/20/24
Data Source	1	1	10/21/24	10/22/24
Box Plots of Tukey’s Five Number Summaries	2	2	10/22/24	10/24/24
Introduction	4	4	10/27/24	10/31/24
Business Problem Description	1	1	10/27/24	10/28/24
Variable Descriptions	2	3	10/28/24	10/30/24
Model Specification	2	3	10/30/24	11/1/24
Predictive Results				
Interaction Terms	3	4	11/1/24	11/4/24
Non-Linear Predictors	1	3	11/1/24	11/2/24
Descriptive Statistics	5	5	11/2/24	11/7/24
Summary of Evaluation Metrics	2	2	11/7/24	11/9/24
Plots of Individual	4	4	11/9/24	11/13/24
Diagnostic Plots of a Final Model				
Residuals Vs Fitted	2	2	11/15/24	11/17/24
Noraml Q-Q	3	3	11/18/24	11/21/24
Scale Location	3	3	11/20/24	11/23/24
Residuals Vs Leverage	3	3	11/24/24	11/27/24
Reviewing Project	1	1	11/27/24	11/28/24
Final				
Variable Inflation Factor Values	1	2	11/29/24	11/30/24
Summary	1	2	11/30/24	12/1/24
Evaluate progress	1	2	12/2/24	12/3/24
FeedBack from Professor	1	1	12/3/24	12/4/24
Changes	1	2	12/4/24	12/5/24
Appendix	2	1	12/5/24	12/7/24
Project Submission	0	0	12/7/24	12/7/24