EPFL

FIN-407 Machine Learning in Finance
Mini project 1 – Airbnb

Group 9

5, Mar 2024

Data Processing

Based on the results of the group presented last week, we performed the following data processing

- **Transfer *certain variables* to be numerical**
  E.g. price, n_bath
- **Remove extreme values using 3 sigma principle**
  E.g. price, minimum_nights
- **Calculate log transformation**
  E.g. price, accommodates, n_reviews
- **Winsorize extreme value**
  E.g. n_beds, n_bath

Data Processing

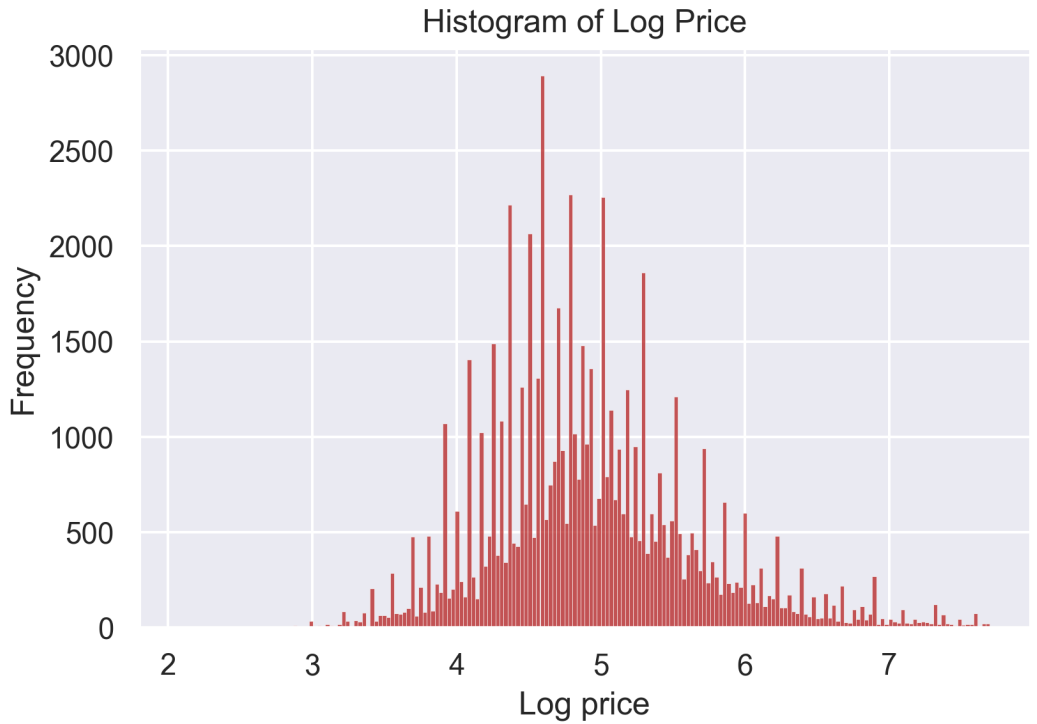
Based on the results of the group presented last week, we performed the following data processing

- To illustrate what we have done, here we plot the histogram of price after we remove extreme values using 3 sigma principle
- We can see that we still have some outliers but it is important to keep some those that are realistic (in the +/- 3 std range) because they are representative of the real estate market

Data Processing

Based on the results of the group presented last week, we performed the following data processing

- Then we calculate the log transformation
- Applying the log price allows us to have a distribution that is a bit more normal with thinner tails (more homoskedastic)



Feature Selection

We evaluate the importance of features and select those with the highest relevance to the target variable

➢ Summary of simple regression of features

	Model	Adj. R^2	AIC	BIC	
0	object	0.087039	140727.749007	140909.909411	neighborhoods
1	object	0.031049	144680.032570	144698.248610	entire_rooms
2	object	0.215758	129024.458582	129042.655734	n_beds
3	object	0.179847	130943.654774	130961.841213	n_baths
4	object	0.011778	103472.928933	103490.647623	review_score_rating
5	object	0.019416	103019.538212	103037.256325	review_score_location
6	object	0.047943	101475.225151	101546.096835	review_score
7	object	0.017016	93630.970323	93648.518838	min_nights
8	object	0.011548	93747.539994	93765.085410	host_superhost
9	object	0.308932	76675.273419	76692.818834	accommodates
10	object	0.317991	76045.809506	76063.354921	ln_accommodates
11	object	-0.000020	94302.546660	94320.092075	nbr_reviews
12	object	0.032539	92723.649791	92741.195206	availability

Simple Model with OLS

Now we regress *ln_price* on different predictors using OLS model

- Model 0: ln_price ~ Entire_room + accommodates + minimum_nights + review_scores_rating + availability_60
- Model 1: ln_price ~ Entire_room + accommodates + minimum_nights + review_scores_rating + availability_60 + beds + n_bath
- Model 2: ln_price ~ Entire_room + accommodates + minimum_nights + review_scores_rating + availability_60 + beds + n_bath + host_is_superhost
- Model 3: ln_price ~ Entire_room + accommodates + minimum_nights + review_scores_rating + availability_60 + beds + n_bath + host_is_superhost + review_scores_location + number_of_reviews

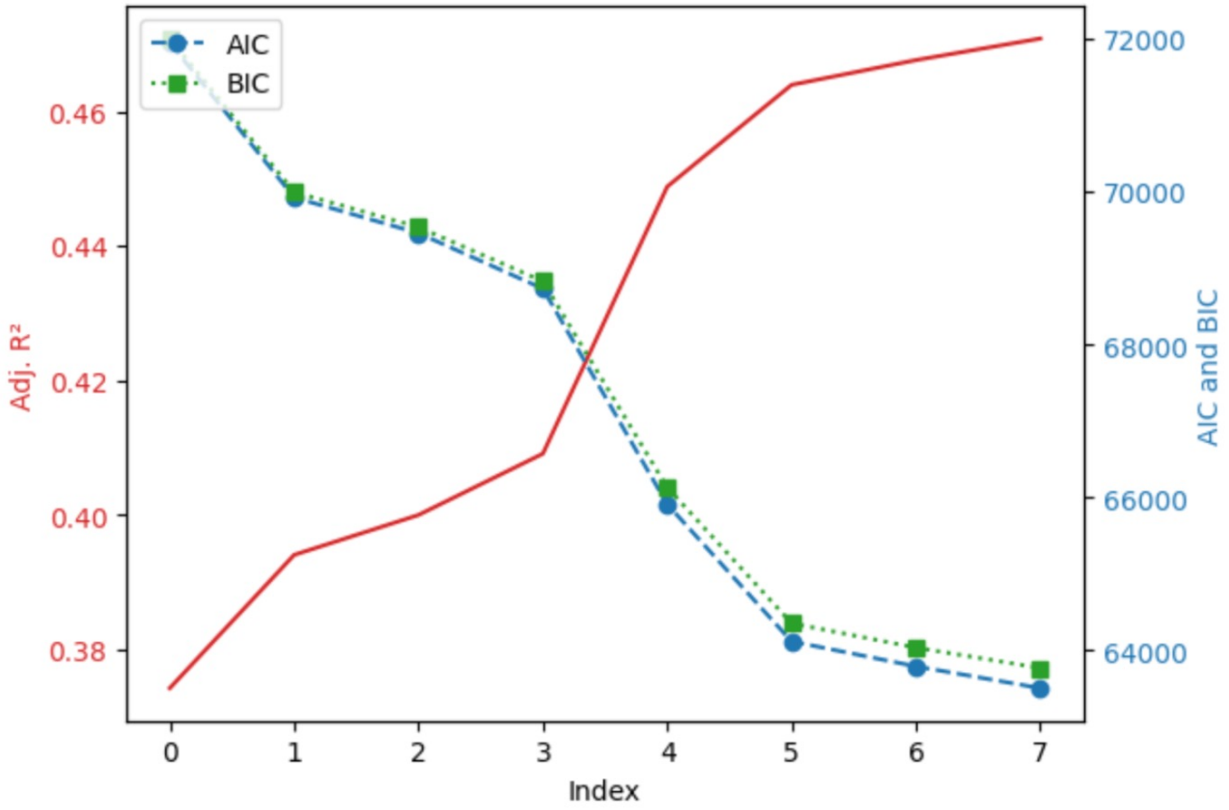
---

- Model 4: adding *neighborhoods* to Model 0
- Model 5: adding *neighborhoods* to Model 1
- Model 6: adding *neighborhoods* to Model 2
- Model 7: adding *neighborhoods* to Model 3

Simple Model with OLS

Summary of results

	Adj. R^2	AIC	BIC
0	0.372346	71589.007889	71641.618294
1	0.394003	69923.226717	69993.373923
2	0.399941	69456.538898	69535.454505
3	0.409060	68731.209738	68827.662146
4	0.447973	65509.921839	65729.131858
5	0.463931	64118.648708	64355.395529
6	0.467613	63792.293292	64037.808514
7	0.470820	63507.359505	63770.411528

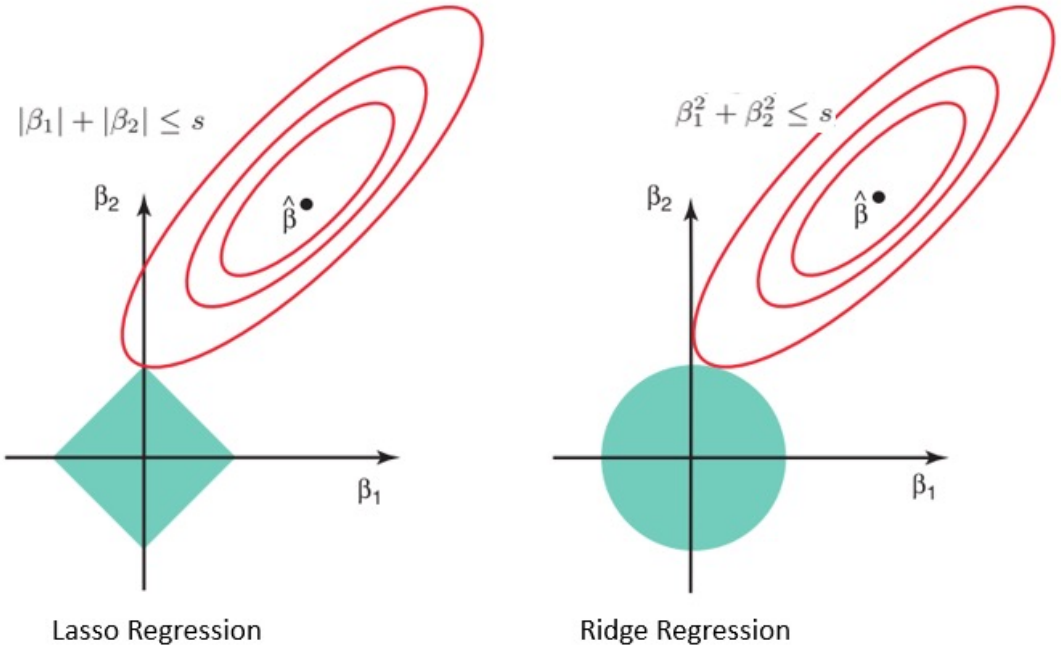


LASSO and RIDGE Model

We tried LASSO and RIDGE model to fit and evaluate them together with OLS models using cross-validation

- **LASSO** can result in some coefficients being shrunk to zero, effectively performing variable selection. It is useful when we have a large number of features, some of which might be irrelevant for the prediction.

- **RIDGE** is particularly useful when dealing with multicollinearity (when independent variables are highly correlated). It tends to shrink the coefficients of less important features but does not set them to zero, thus including all the features in the final model.

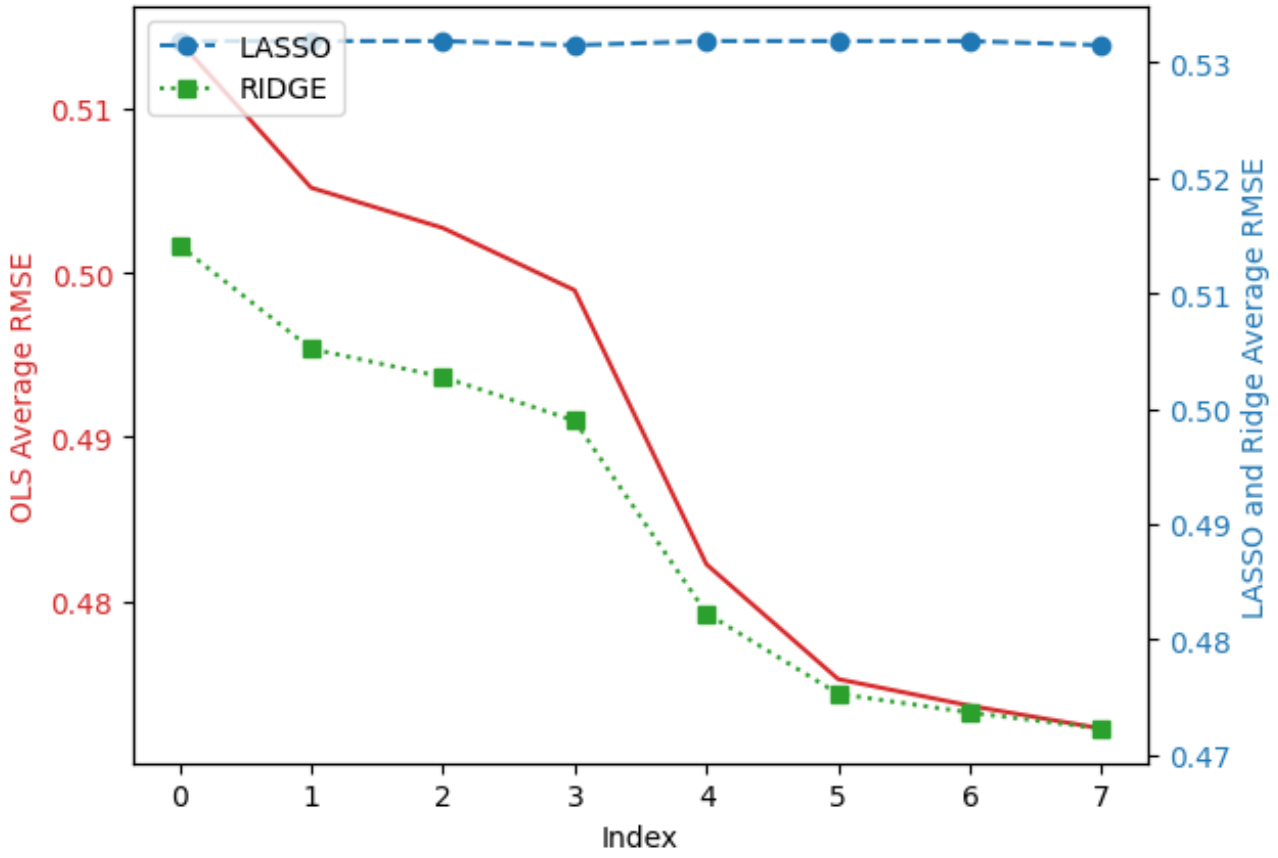


K-fold Cross-Validation

We divided the dataset into 5 subsets and then model is trained on 4 and tested on the remaining subset

- This process is repeated k = 5 times, with each subset used once as the test set. We calculated the average RMSE for OLS, LASSO, and RIDGE. Cross-validation helps in detecting overfitting, underfitting, and provides a more accurate measure of a model's predictive power on unseen data.

- Results are shown on the right.



Final Model Regression Result

Based on the previous explanation, we select Model 7 as the final model

No. Observations:	47496	R-squared:	0.471
Df Residuals:	47466	Adj. R-squared:	0.471
Df Model:	29	F-statistic:	1458
Covariance Type:	nonrobust	Prob (F-statistic):	0
AIC:	6.35E+04	Log-Likelihood:	-31724
BIC:	6.38E+04		

Final Model Regression Result

Based on the previous explanation, we select Model 7 as the final model

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.8410	0.0360	78.8770	0.0000	2.7700	2.9120
Observatoire[T.True]	0.0072	0.0130	0.5440	0.5860	-0.0190	0.0330
Hôtel_de_Ville[T.True]	0.3740	0.0130	27.8450	0.0000	0.3480	0.4000
Entrepôt[T.True]	0.0796	0.0100	7.7150	0.0000	0.0590	0.1000
Popincourt[T.True]	0.0432	0.0100	4.4230	0.0000	0.0240	0.0620
Louvre[T.True]	0.4523	0.0150	29.4580	0.0000	0.4220	0.4820
Bourse[T.True]	0.2284	0.0130	17.5470	0.0000	0.2030	0.2540
Gobelins[T.True]	-0.0775	0.0140	-5.6370	0.0000	-0.1040	-0.0510
Luxembourg[T.True]	0.3872	0.0150	26.6170	0.0000	0.3590	0.4160
Buttes_Chaumont[T.True]	-0.1659	0.0110	-14.9440	0.0000	-0.1880	-0.1440
Reuilly[T.True]	-0.0434	0.0130	-3.3810	0.0010	-0.0690	-0.0180
Élysée[T.True]	0.3581	0.0140	25.1770	0.0000	0.3300	0.3860
Panthéon[T.True]	0.2705	0.0140	19.7020	0.0000	0.2440	0.2970
Batignolles_Monceau[T.True]	0.0675	0.0110	6.1430	0.0000	0.0460	0.0890
Vaugirard[T.True]	0.1317	0.0100	12.5830	0.0000	0.1110	0.1520

Final Model Regression Result

Based on the previous explanation, we select Model 7 as the final model

	coef	std err	t	P>|t|	[0.025	0.975]
Opéra[T.True]	0.2116	0.0120	17.7360	0.0000	0.1880	0.2350
Palais_Bourbon[T.True]	0.4311	0.0150	29.0060	0.0000	0.4020	0.4600
Passy[T.True]	0.2630	0.0120	22.1600	0.0000	0.2400	0.2860
Temple[T.True]	0.2765	0.0120	23.0430	0.0000	0.2530	0.3000
Ménilmontant[T.True]	-0.1567	0.0110	-13.9290	0.0000	-0.1790	-0.1350
Entire_room	0.2844	0.0070	41.1640	0.0000	0.2710	0.2980
accommodates	0.1362	0.0030	52.1580	0.0000	0.1310	0.1410
beds	0.0436	0.0040	11.3750	0.0000	0.0360	0.0510
n_bath	0.2058	0.0060	33.7710	0.0000	0.1940	0.2180
minimum_nights	-0.0064	0.0000	-39.4370	0.0000	-0.0070	-0.0060
review_scores_rating	0.1567	0.0060	24.4730	0.0000	0.1440	0.1690
review_scores_location	0.0413	0.0080	4.8610	0.0000	0.0250	0.0580
number_of_reviews	-0.0006	0.0000	-16.3680	0.0000	-0.0010	-0.0010
availability_60	0.0042	0.0000	39.7230	0.0000	0.0040	0.0040
host_is_superhost	0.1146	0.0050	21.1300	0.0000	0.1040	0.1250

Final Model Regression Result

Based on the previous explanation, we select Model 7 as the final model

Omnibus:	6795.849	Durbin-Watson:	1.85
Prob(Omnibus):	0	Jarque-Bera (JB):	28691.928
Skew:	0.657	Prob(JB):	0
Kurtosis:	6.574	Cond. No.	1220

Breush Pagan Test

We conducted Breush-Pagan test and fitted with heteroscedasticity robust (HC3) standard errors

- **Breusch-Pagan test** is a statistical method used to test for heteroskedasticity in a linear regression model. Heteroskedasticity occurs when the variance of the errors in a regression model is not constant across observations, which can lead to inefficiencies in the estimation process and invalid standard errors, affecting hypothesis tests.

```
Breusch-Pagan Test:
Lagrange multiplier statistic: 1888.6787064394123
p-value: 0.0
f-value: 67.78102872080247
f p-value: 0.0
```

- The interpretation focuses on the p-value: if the p-value is less than a chosen significance level (commonly 0.05), there is evidence to reject the null hypothesis of homoskedasticity, indicating the presence of heteroskedasticity in the model.

Model Regression Result – HC3

HC3 adjusts the calculation of standard errors to account for heteroskedasticity without specifying its form

No. Observations:	47496	R-squared:	0.471
Df Residuals:	47466	Adj. R-squared:	0.471
Df Model:	29	F-statistic:	1044
Covariance Type:	HC3	Prob (F-statistic):	0
AIC:	6.35E+04	Log-Likelihood:	-31724
BIC:	6.38E+04		

Model Regression Result – HC3

HC3 adjusts the calculation of standard errors to account for heteroskedasticity without specifying its form

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.8410	0.0440	64.2330	0.0000	2.7540	2.9280
Observatoire[T.True]	0.0072	0.0130	0.5650	0.5720	-0.0180	0.0320
Hôtel_de_Ville[T.True]	0.3740	0.0140	27.2850	0.0000	0.3470	0.4010
Entrepôt[T.True]	0.0796	0.0100	7.9570	0.0000	0.0600	0.0990
Popincourt[T.True]	0.0432	0.0100	4.5350	0.0000	0.0250	0.0620
Louvre[T.True]	0.4523	0.0150	29.3380	0.0000	0.4220	0.4820
Bourse[T.True]	0.2284	0.0130	17.4020	0.0000	0.2030	0.2540
Gobelins[T.True]	-0.0775	0.0140	-5.7280	0.0000	-0.1040	-0.0510
Luxembourg[T.True]	0.3872	0.0170	23.4110	0.0000	0.3550	0.4200
Buttes_Chaumont[T.True]	-0.1659	0.0110	-15.7060	0.0000	-0.1870	-0.1450
Reuilly[T.True]	-0.0434	0.0120	-3.5250	0.0000	-0.0680	-0.0190
Élysée[T.True]	0.3581	0.0160	22.6310	0.0000	0.3270	0.3890
Panthéon[T.True]	0.2705	0.0140	19.2300	0.0000	0.2430	0.2980
Batignolles_Monceau[T.True]	0.0675	0.0110	6.1020	0.0000	0.0460	0.0890
Vaugirard[T.True]	0.1317	0.0110	12.2950	0.0000	0.1110	0.1530

Model Regression Result – HC3

HC3 adjusts the calculation of standard errors to account for heteroskedasticity without specifying its form

	coef	std err	z	P>|z|	[0.025	0.975]
Opéra[T.True]	0.2116	0.0120	17.8890	0.0000	0.1880	0.2350
Palais_Bourbon[T.True]	0.4311	0.0160	26.3730	0.0000	0.3990	0.4630
Passy[T.True]	0.2630	0.0130	20.8140	0.0000	0.2380	0.2880
Temple[T.True]	0.2765	0.0120	23.2000	0.0000	0.2530	0.3000
Ménilmontant[T.True]	-0.1567	0.0100	-15.3050	0.0000	-0.1770	-0.1370
Entire_room	0.2844	0.0080	33.9060	0.0000	0.2680	0.3010
accommodates	0.1362	0.0030	45.5720	0.0000	0.1300	0.1420
beds	0.0436	0.0040	9.9000	0.0000	0.0350	0.0520
n_bath	0.2058	0.0140	14.8050	0.0000	0.1790	0.2330
minimum_nights	-0.0064	0.0000	-22.4350	0.0000	-0.0070	-0.0060
review_scores_rating	0.1567	0.0070	21.2340	0.0000	0.1420	0.1710
review_scores_location	0.0413	0.0100	4.2350	0.0000	0.0220	0.0600
number_of_reviews	-0.0006	0.0000	-12.4890	0.0000	-0.0010	0.0000
availability_60	0.0042	0.0000	36.9080	0.0000	0.0040	0.0040
host_is_superhost	0.1146	0.0050	21.7720	0.0000	0.1040	0.1250

Model Regression Result – HC3

HC3 adjusts the calculation of standard errors to account for heteroskedasticity without specifying its form

Omnibus:	6795.849	Durbin-Watson:	1.85
Prob(Omnibus):	0	Jarque-Bera (JB):	28691.928
Skew:	0.657	Prob(JB):	0
Kurtosis:	6.574	Cond. No.	1220

VIF Test

We consider VIFs of 5 and higher to represent problematic amounts of multicollinearity

- **Variance Inflation Factor test** quantifies how much the variance of an estimated regression coefficient increases if your predictors are correlated. A VIF of 1 indicates no multicollinearity. Conversely, a VIF exceeding 5 suggests potentially problematic levels of multicollinearity.

	VIF		VIF
Intercept	275.0464	Opéra[T.True]	1.4062
Observatoire[T.True]	1.2940	Palais_Bourbon[T.True]	1.2504
Hôtel_de_Ville[T.True]	1.3102	Passy[T.True]	1.4298
Entrepôt[T.True]	1.5537	Temple[T.True]	1.4227
Popincourt[T.True]	1.6976	Ménilmontant[T.True]	1.4251
Louvre[T.True]	1.2303	Entire_room	1.1237
Bourse[T.True]	1.3425	accommodates	3.9825
Gobelins[T.True]	1.2602	beds	3.6061
Luxembourg[T.True]	1.2654	n_bath	1.5137
Buttes_Chaumont[T.True]	1.4387	minimum_nights	1.0449
Reuilly[T.True]	1.3126	review_scores_rating	1.4451
Élysée[T.True]	1.2818	review_scores_location	1.5332
Panthéon[T.True]	1.2882	number_of_reviews	1.0373
Batignolles_Monceau[T.True]	1.4827	availability_60	1.0432
Vaugirard[T.True]	1.5674	host_is_superhost	NaN

RESET Test

We used RESET test as a diagnostic tool to detect specification errors in the linear regression model

- **Regression Equation Specification Error Test** involves adding higher-order terms of the model's predictions (e.g., squared and cubed terms) to the original model and then reassessing the model. If these additional terms are statistically significant, it suggests that the model may be mis-specified.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.489
Model:	OLS	Adj. R-squared:	0.489
Method:	Least Squares	F-statistic:	1.122e+04
Date:	Tue, 05 Mar 2024	Prob (F-statistic):	0.00
Time:	18:46:00	Log-Likelihood:	-30517.
No. Observations:	46960	AIC:	6.104e+04
Df Residuals:	46955	BIC:	6.109e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	21.4669	2.216	9.688	0.000	17.124	25.810
0	-13.5140	1.588	-8.509	0.000	-16.627	-10.401
1	3.4933	0.421	8.304	0.000	2.669	4.318
2	-0.3469	0.049	-7.120	0.000	-0.442	-0.251
3	0.0114	0.002	5.499	0.000	0.007	0.015

Omnibus:	7032.585	Durbin-Watson:	1.833
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21468.146
Skew:	0.786	Prob(JB):	0.00
Kurtosis:	5.916	Cond. No.	8.90e+05