Airbnb Price Prediction

[Final Report]

11/28/2024

Tina Nosrati

Fall 2024
CMPS 240
Professor Lawrence D'Antonio

Problem Definition

This project analyzes Airbnb weekend prices in Athens to identify key factors influencing pricing using visualizations, descriptive statistics, hypothesis testing, and regression analysis on the Athens Weekends dataset. The goal is to provide insights for hosts to set competitive prices and for travelers to understand cost influences, focusing on variables like room type, location, and Superhost status.

Data [Link]

This dataset is from Kaggle and focuses on weekend Airbnb prices in Athens.

Features in the dataset:

RealSum [price]	Guest_satisfaction_overall [guest rating]
Room_type [private, shared]	Bedrooms [number of bedrooms]
Room_shared [shared or not]	Dist [distance from the city center]
Room_private [private or not]	Metro_dist [distance from the nearest metro]
Person_capacity [max number of people]	Lng [longitude]
Host_is_superhost [superhost or not]	Lat [latitude]
Multi [multiple rooms or not]	Biz [business purposes or not]
Cleanliness_rating [cleanliness rating]	Dist [distance from the city center]
Attr_index_norm [Attractiveness index]	Rest_index_norm [Resturant index]

Python libraries used:

1. Pandas
2. Numpy
3. Seaborn
4. Matplotlib
5. Scipy
6. Statsmodels
7. folium

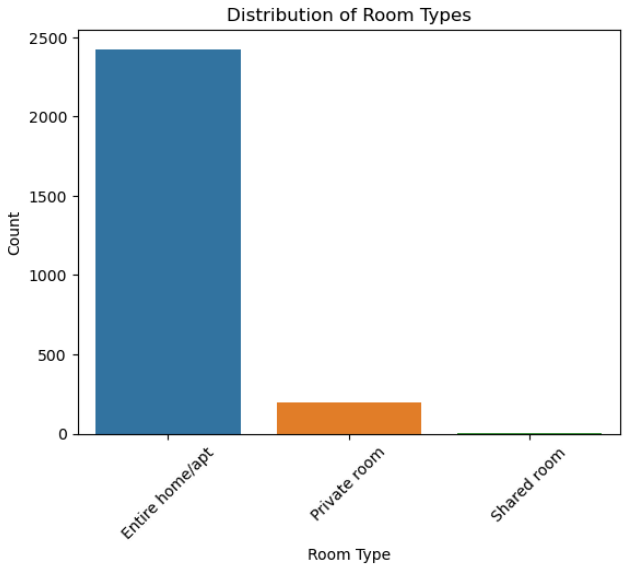
Research Questions

1. What insights can be gained by visualizing data from different perspectives?
2. Is there a correlation between different variables in this dataset?
3. Is the mean price (realSum) equal to 100?
4. Is the mean price for superhost listings the same as for non-superhost listings?
5. Is the mean price the same across all bedroom groups?
6. Is the mean guest satisfaction equal to 95?
7. Is the cleanliness rating the same for business and non-business listings?
8. Is the mean price the same across all distance quartiles?
9. Does the price distribution (realSum) follow a normal distribution?
10. Does the data (realSum) follow an exponential distribution?
11. How do the features in the dataset explain the price (realSum)?

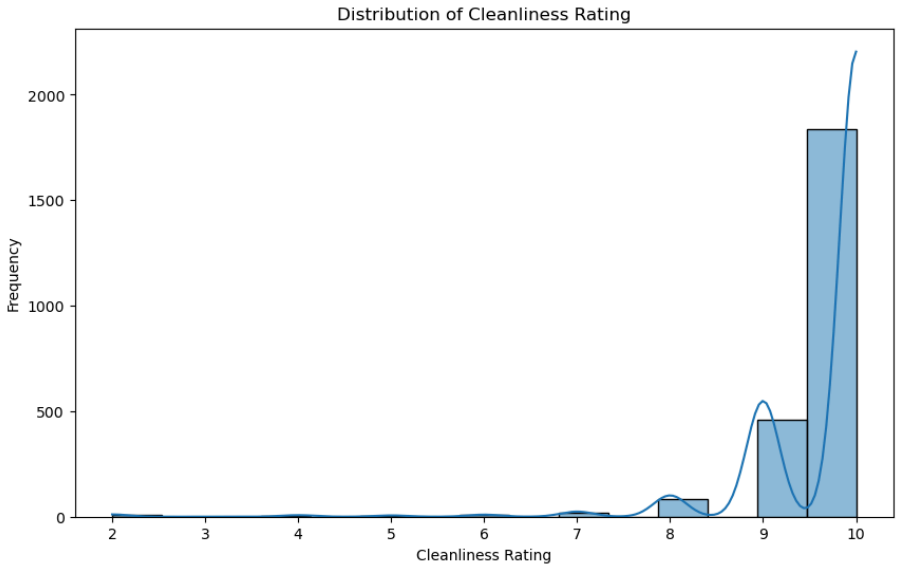
Data Manipulation

1. Removing unnecessary columns
2. Checking for missing values
3. Removing duplicated rows
4. Make sure geographical information belongs to Athens
5. Convert boolean variables to integer
6. Removing 'Private room' and 'Shared room' because of few observations

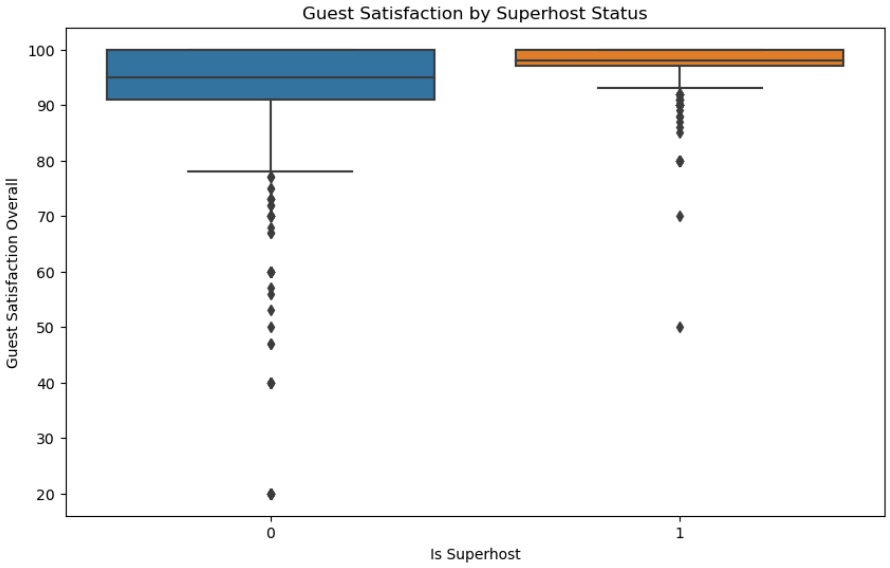
Data Visualization



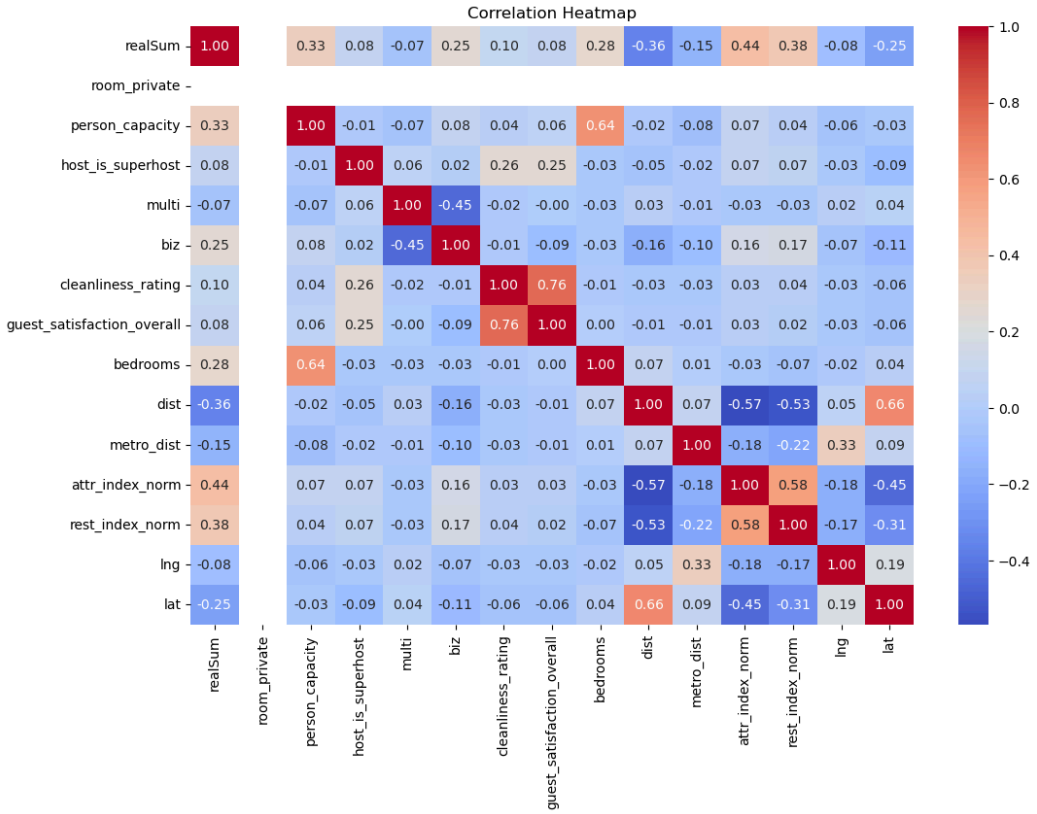
1- From visualizing the distribution of Room Types, I decided to only keep 'Entire home/apt' observations



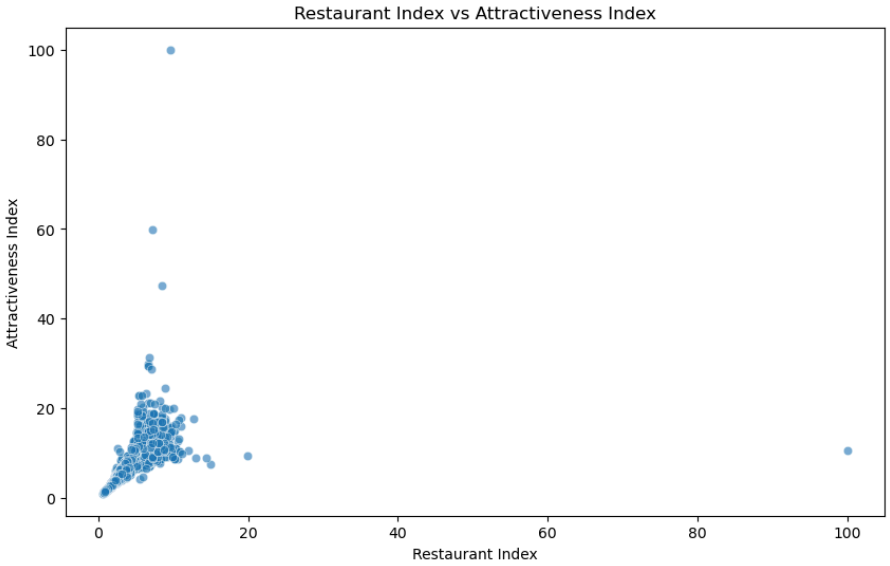
2- From Visualizing the distribution of cleanliness ratings, I can conclude that most of the cleanliness ratings are high



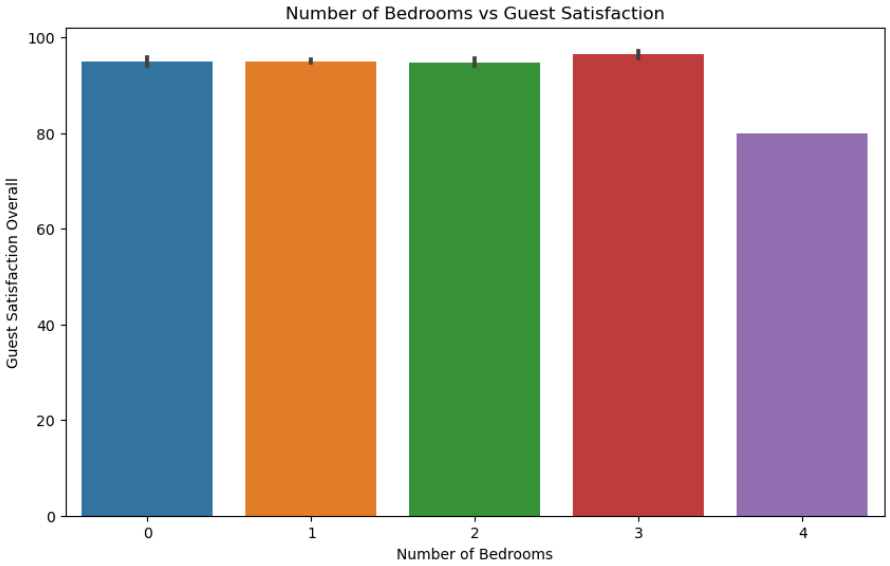
3- From Visualizing the satisfaction rating vs. superhost status, Superhosts tend to have higher and more consistent guest satisfaction scores than non-superhosts.



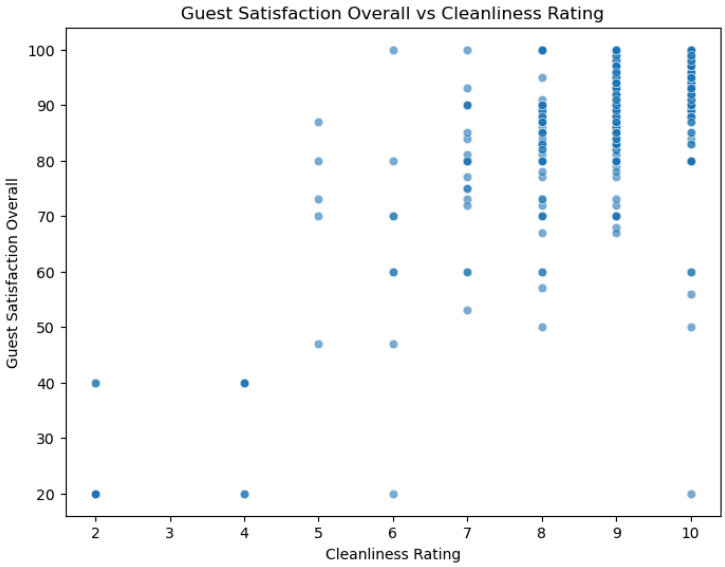
4- There is not much significant correlation between variables, but Guest satisfaction overall has the strongest positive correlation with cleanliness rating, while attributes like distance (dist) and metro distance show weak or negative correlations with most variables.



5- A weak relationship exists between the Restaurant Index and Attractiveness Index



6- Guest satisfaction is consistently high with locations with less than four bedrooms but drops for locations with four bedrooms.



7- Higher cleanliness ratings are generally associated with higher guest satisfaction

Hypothesis Testing

1. **Is the mean price (realSum) equal to 100?**
    - Null Hypothesis: The mean price (realSum) is 100.
    - Alternative Hypothesis: The mean price (realSum) is not 100.
    - Null Hypothesis rejected

2. **Is the mean price for superhost listings the same as for non-superhost listings?**
    - Null Hypothesis: The mean price for superhost and non-superhost listings is the same.
    - Alternative Hypothesis: The mean price for superhost and non-superhost listings is different.
    - Null Hypothesis rejected

3. **Is the mean price the same across all bedroom groups?**
    - Null Hypothesis: The mean price is the same across all bedroom groups.
    - Alternative Hypothesis: The mean price is different across at least one bedroom group.
    - Null Hypothesis rejected

4. **Is the mean guest satisfaction equal to 95?**
    - Null Hypothesis: The mean guest satisfaction is 95.
    - Alternative Hypothesis: The mean guest satisfaction is not 95.
    - Null Hypothesis not rejected

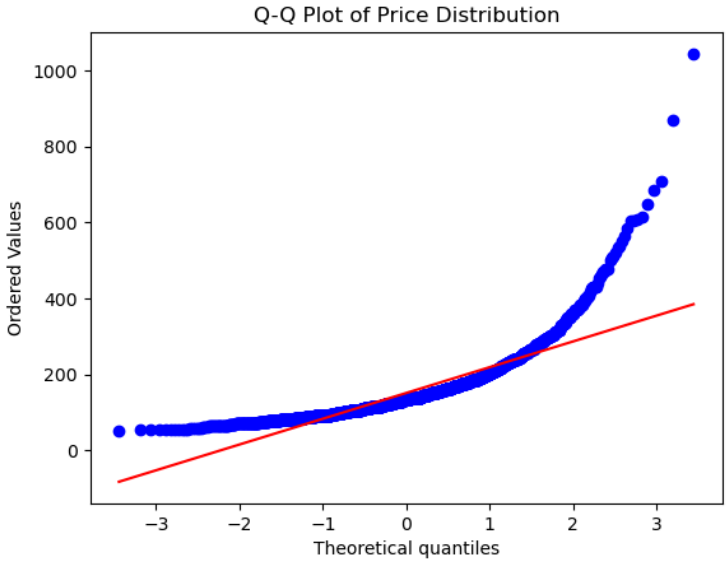
5. **Is the cleanliness rating the same for business and non-business listings?**
    - Null Hypothesis: The cleanliness rating is the same for business and non-business listings.
    - Alternative Hypothesis: The cleanliness rating is different for business and non-business listings.
    - Null Hypothesis not rejected

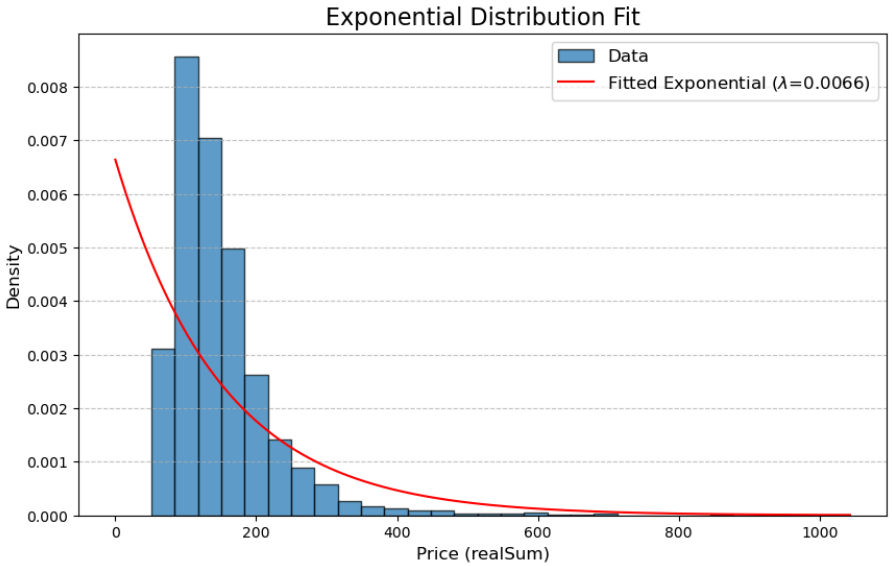
6. **Is the mean price the same across all distance quartiles?**
    - Null Hypothesis: The mean price is the same across all distance quartiles.
    - Alternative Hypothesis: The mean price is different across at least one distance quartile.
    - Null Hypothesis rejected

7. **Does the price distribution (realSum) follow a normal distribution?**
    - Null Hypothesis: The price distribution (realSum) follows a normal distribution.
    - Alternative Hypothesis: The price distribution (realSum) does not follow a normal distribution.
    - Null Hypothesis rejected

8. **Does the data (realSum) follow an exponential distribution?**
    - Null Hypothesis: The data (realSum) follows an exponential distribution.
    - Alternative Hypothesis: The data (realSum) does not follow an exponential distribution.
    - Null Hypothesis rejected

To visually validate the 7th and 8th hypotheses, I plotted these visualizations and concluded that the distribution is closer to exponential than normal:





Linear Regression

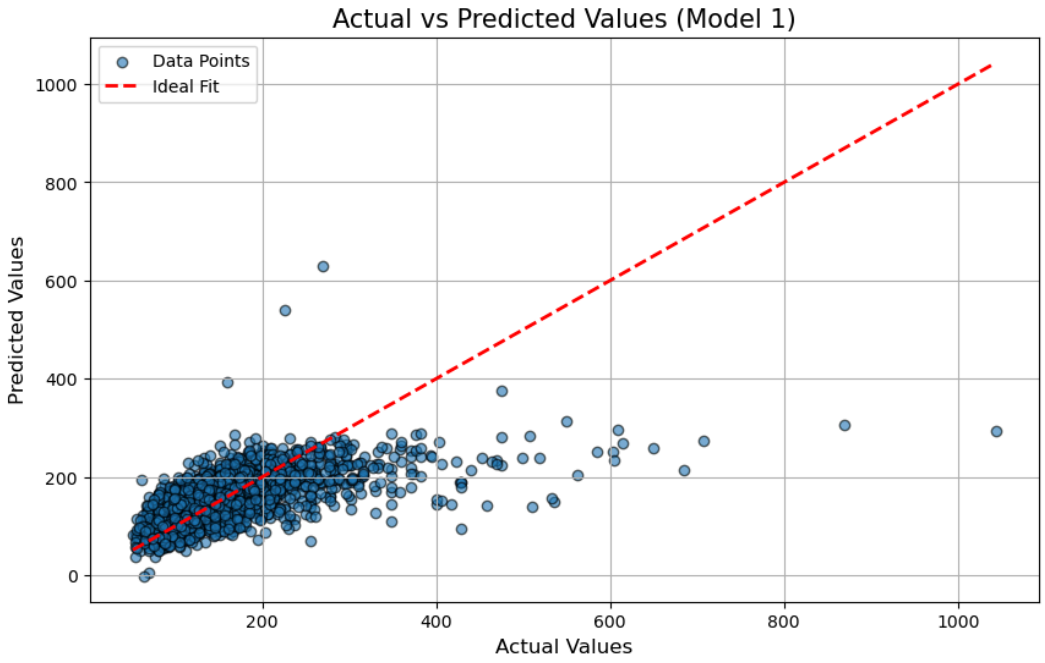
The regression model explains 37.3% of the variation in the target variable, showing that factors like person_capacity, biz, and bedrooms positively influence the outcome, while dist has a negative effect. However, the scatter plot reveals the model struggles to accurately predict higher values, indicating room for improvement in its performance.

Model with All Variables:

OLS Regression Results

Dep. Variable:	realSum	R-squared:	0.373
Model:	OLS	Adj. R-squared:	0.370
Method:	Least Squares	F-statistic:	130.4
Date:	Thu, 28 Nov 2024	Prob (F-statistic):	5.19e-235
Time:	00:00:41	Log-Likelihood:	-13422.
No. Observations:	2425	AIC:	2.687e+04
Df Residuals:	2413	BIC:	2.694e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-14.6586	17.016	-0.861	0.389	-48.026	18.709
room_private	1.884e-13	2.56e-13	0.737	0.461	-3.13e-13	6.9e-13
person_capacity	9.5383	1.317	7.245	0.000	6.957	12.120
host_is_superhost	4.9195	2.632	1.869	0.062	-0.241	10.080
multi	7.9590	3.247	2.452	0.014	1.593	14.325
biz	29.1349	3.012	9.673	0.000	23.228	35.042
cleanliness_rating	4.0414	2.316	1.745	0.081	-0.501	8.584
guest_satisfaction_overall	0.3348	0.233	1.437	0.151	-0.122	0.791
bedrooms	25.2072	2.490	10.122	0.000	20.324	30.091
dist	-10.1738	1.654	-6.150	0.000	-13.418	-6.930



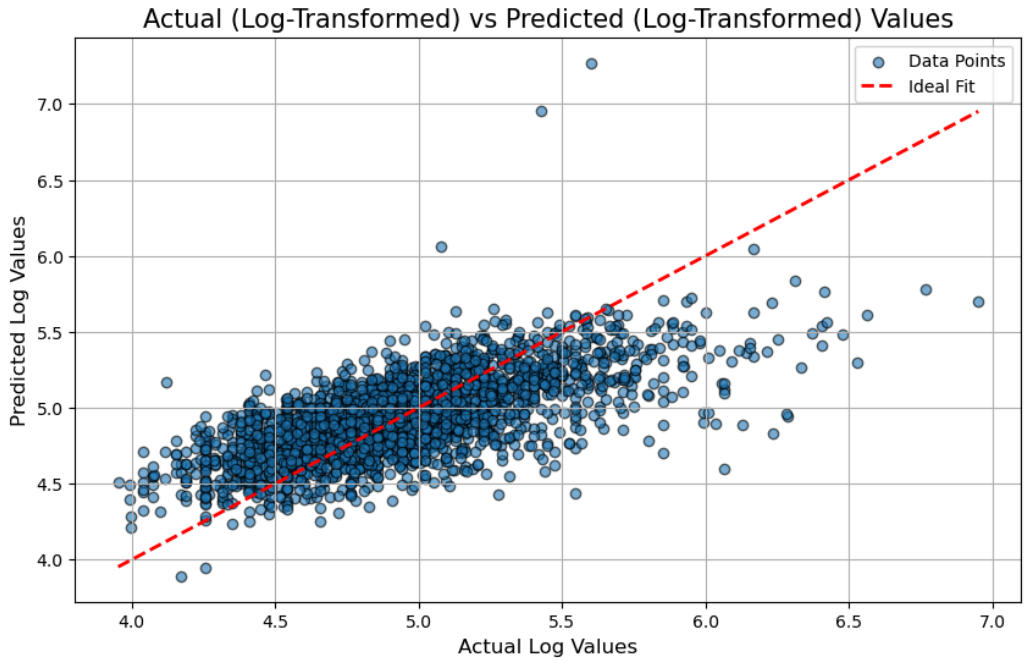
log-transformed regression

The updated regression model, using log-transformed values, explains 45.6% of the variation in the dependent variable. Predictors like person_capacity, host_is_superhost, and bedrooms show significant positive effects, while dist and metro_dist negatively impact the outcome.

OLS Regression Results

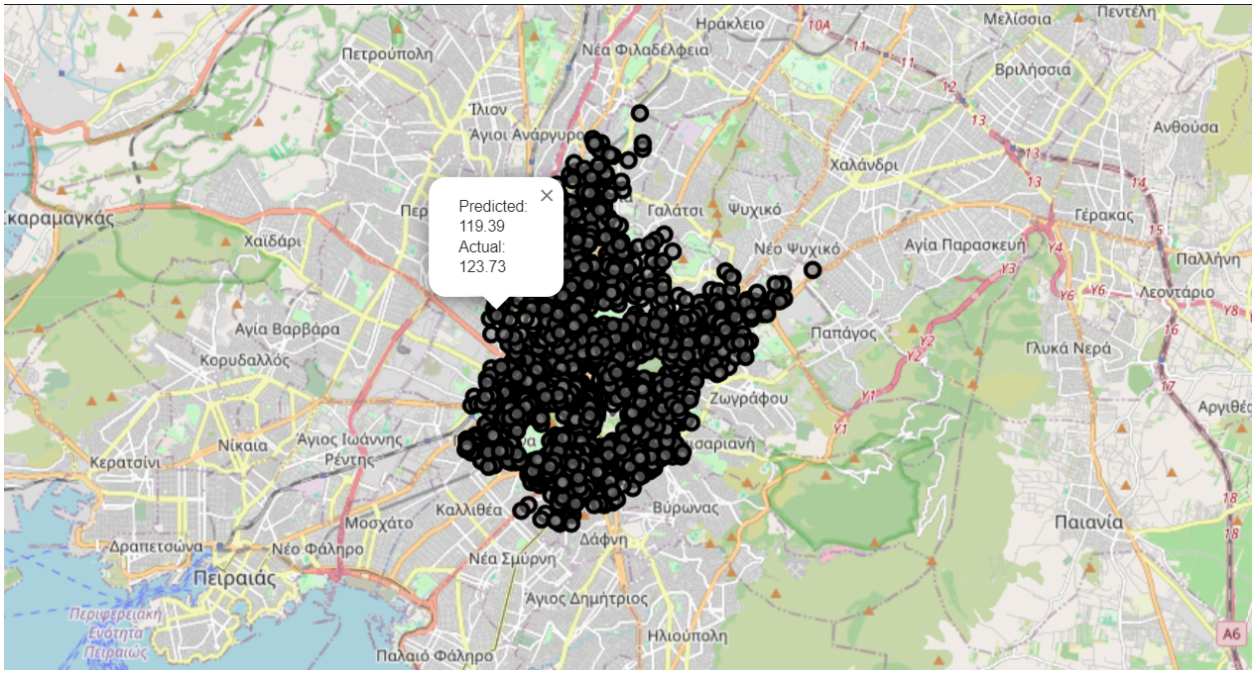
Dep. Variable:	realSum	R-squared:	0.456
Model:	OLS	Adj. R-squared:	0.453
Method:	Least Squares	F-statistic:	183.7
Date:	Thu, 28 Nov 2024	Prob (F-statistic):	7.84e-309
Time:	00:17:15	Log-Likelihood:	-534.81
No. Observations:	2425	AIC:	1094.
Df Residuals:	2413	BIC:	1163.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.9559	0.084	47.243	0.000	3.792	4.120
room_private	6.117e-14	1.26e-15	48.597	0.000	5.87e-14	6.36e-14
person_capacity	0.0679	0.006	10.485	0.000	0.055	0.081
host_is_superhost	0.0377	0.013	2.914	0.004	0.012	0.063
multi	0.0400	0.016	2.504	0.012	0.009	0.071
biz	0.1760	0.015	11.877	0.000	0.147	0.205
cleanliness_rating	0.0383	0.011	3.361	0.001	0.016	0.061
guest_satisfaction_overall	0.0014	0.001	1.244	0.214	-0.001	0.004
bedrooms	0.1204	0.012	9.826	0.000	0.096	0.144
dist	-0.0858	0.008	-10.535	0.000	-0.102	-0.070
metro_dist	-0.0891	0.022	-3.986	0.000	-0.133	-0.045



Map

The final result of prediction and actual value has been visualized on a map and saved as an HTML file



What results did you obtain, and what conclusions did you make?

This project looks at what affects Airbnb weekend prices in Athens, using charts, statistics, and analysis. It found that prices change a lot based on things like room type, distance, and Superhost status, while guest satisfaction and cleanliness stayed steady. Bigger spaces and more bedrooms tend to cost more, but being farther from the center usually lowers the price. A log-transformed regression model explained nearly half of the price differences, giving useful tips for hosts to set fair prices and for travelers to understand what impacts costs.