LAHostInsight: Unveiling Airbnb Superhosts and Rating Influencers

in Great Los Angeles Area

COMM 557 Final Presentation

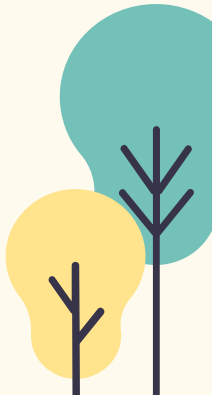
Kelly Choy
Shih-Min (Julia) Huang
Tongxin (Shirley) Ye

Project Introduction

- Analyze Airbnb listings data in LA
  - Exploratory data analysis & visualization
  - Data modeling for superhost prediction & exploring review score influencers
- Dataset:
  - Provided by Inside Airbnb
  - Dataset time frame: 9/4/2023

Key Questions

- **Superhost Prediction**
  - Identify and investigate on the features that make a homeowner a Superhost
  - Build & experiment ML models to accurately predict Superhost
- **What Features Influences an Airbnb Homeowner Review Score**
  - Use regression modeling to find significant variables that influences review score, and capture additional insights or patterns



Goals: Provide Directions & Actionable Insights

Airbnb Hosts

- Obtain Superhost verification
- Improve overall review scores

Airbnb (Company)

- Refine current Superhost criteria
- Enhance the recommendation system by identifying key features influencing review scores

Exploratory Data Analysis and Visualizations

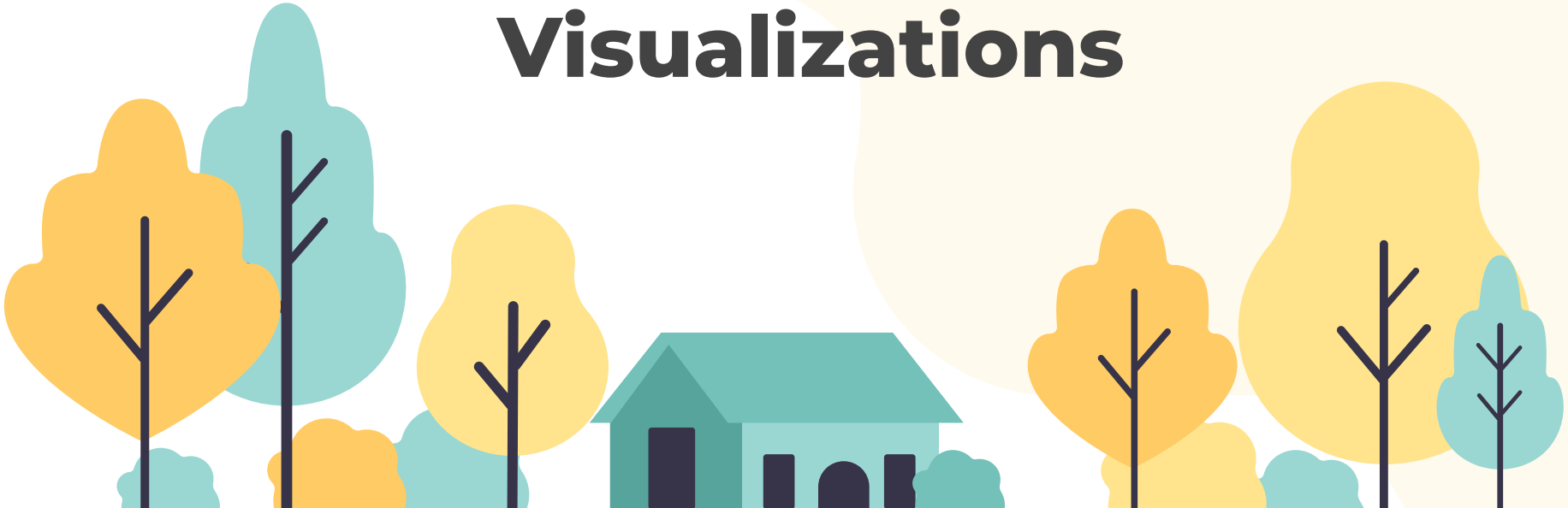


Tableau Server

Added visualizations on Tableau Server where users can select their own filters

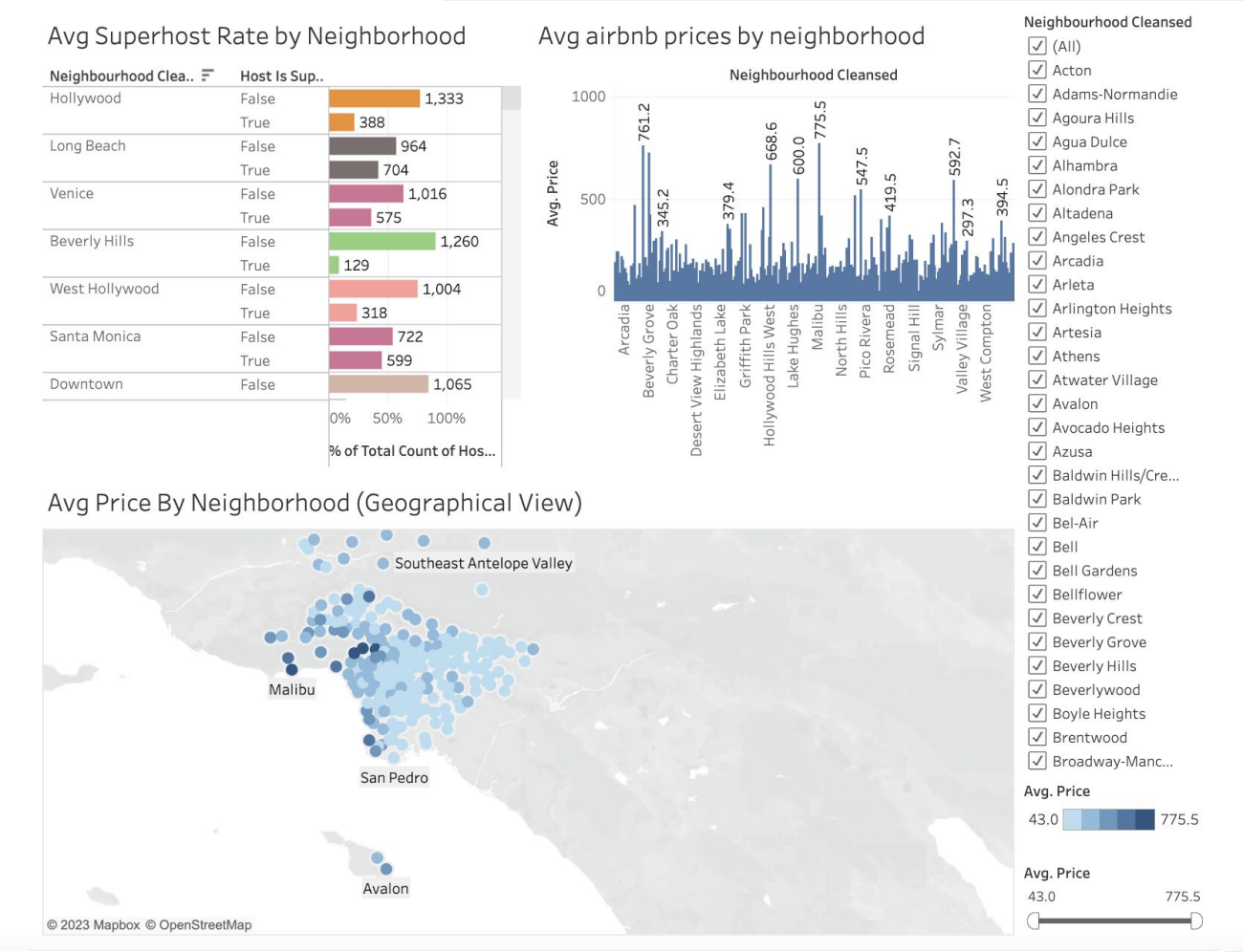
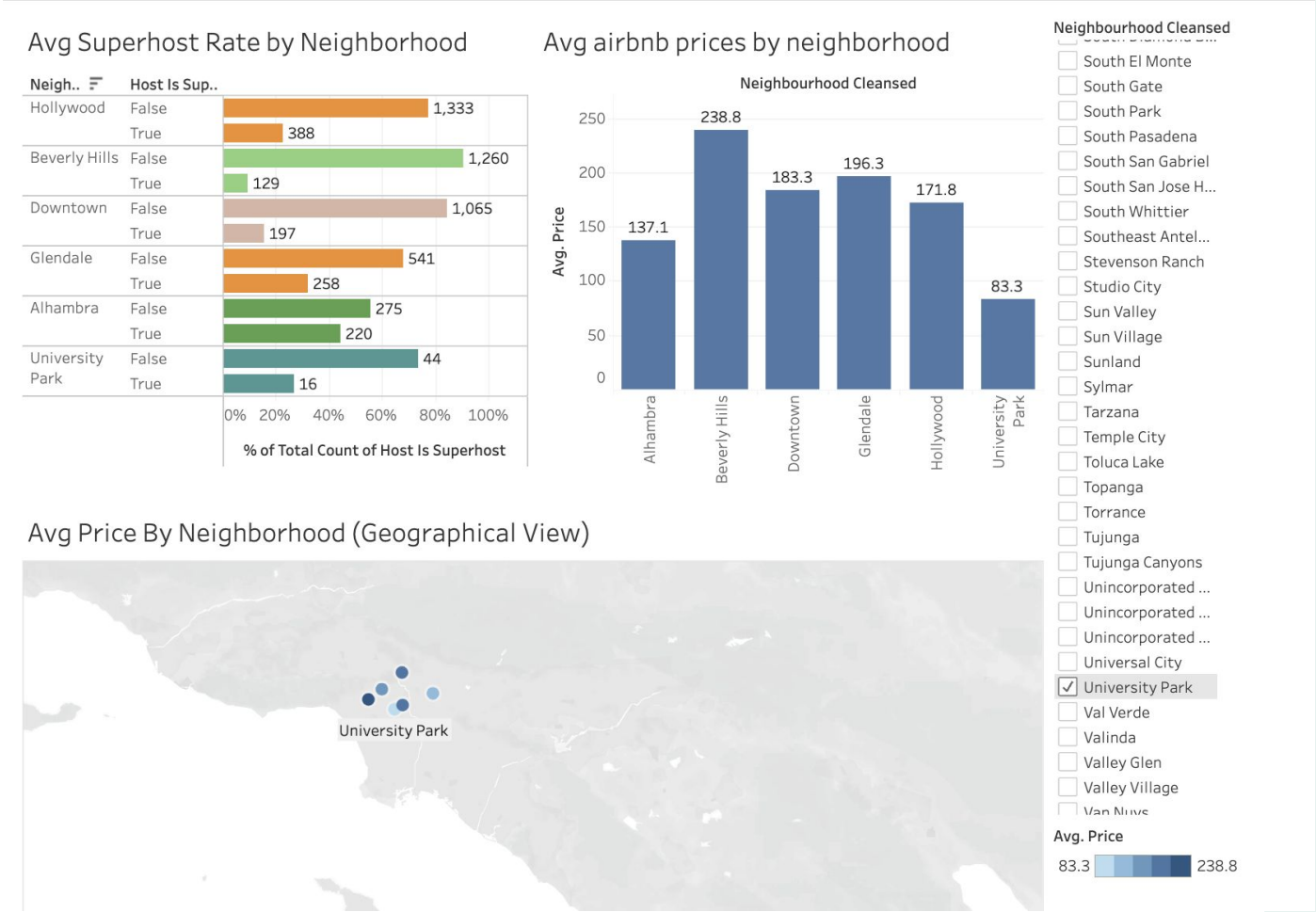


Tableau Server

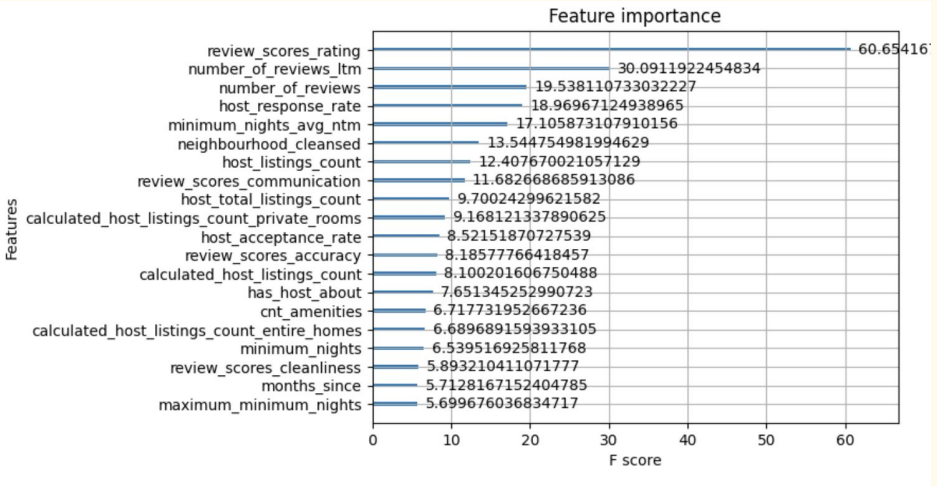
Example of selecting certain neighborhoods



Superhost Prediction

XGBoost Classifier

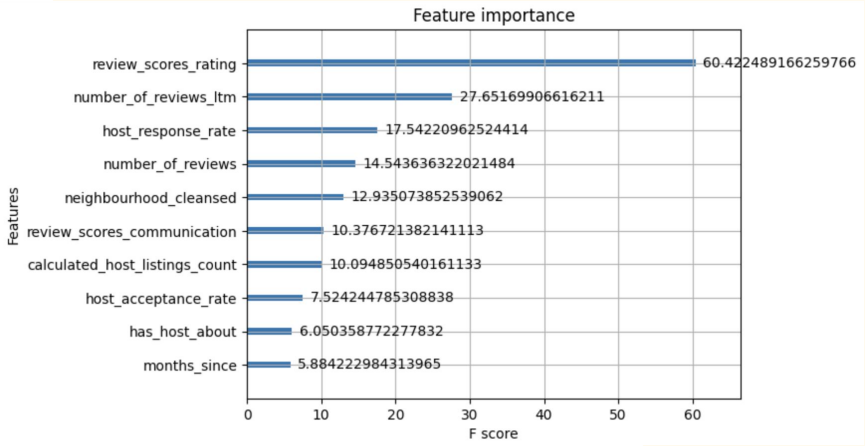
- XGBoost Classifier
  - Select Top 20 features based on the average gain of splits which use the feature.
  - Use the 20 features to train XGBClassifier with 5-fold cross validation, mean accuracy = 0.8627, std = 0.0039



Random Forest

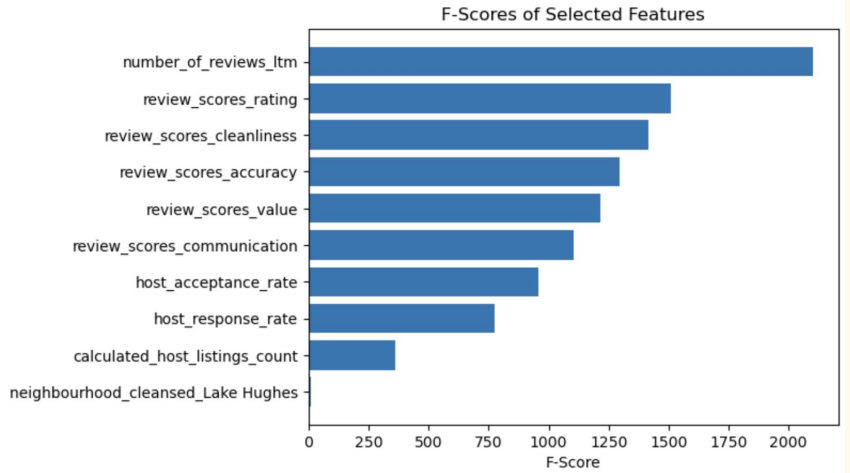
- Used Python sklearn's RandomForestClassifier with 5-fold cross validation
- Number of features selected: 10
- Accuracy: 83.59%

```
Random Forest Accuracy: 0.8359829391236914

Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.86      0.85      2749
           1       0.83      0.81      0.82      2409

    accuracy                           0.84      5158
   macro avg       0.84      0.83      0.84      5158
weighted avg       0.84      0.84      0.84      5158
```



Logistic Regression

- Used Python sklearn's Logistic Regression
- Accuracy: 72.6%



Confusion Matrix		
	Positive	**Negative**
Positive	1780	786
Negative	629	1963

Classification Report:

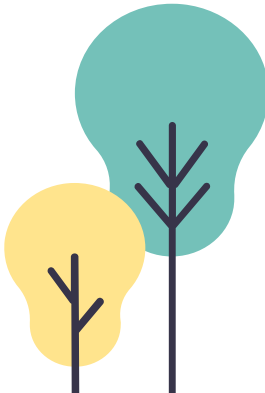
	precision	recall	f1-score	support
0	0.76	0.71	0.74	2749
1	0.69	0.74	0.72	2409
accuracy			0.73	5158
macro avg	0.73	0.73	0.73	5158
weighted avg	0.73	0.73	0.73	5158

Comparing the different models

Accuracy		
XGBoost Regressor	**Random Forest**	**Logistic Regression**
86.3%	83.6%	72.6%

Top Features:

- review_scores_rating
- number_of_reviews_ltm
- host_response_rate
- host_acceptance_rate
- review_scores_communication
- calculated_host_listings_count



Unveiling Key Factors for Becoming an Exceptional Airbnb Host

Comparative analysis of best performers and the rest

Method 1 - K-Means Clustering

- # of clusters = 3, silhouette score = 0.88
- Mathematically distinct but not necessarily meaningful or interpretable in real-world terms;

Method 2 - Statistical Tests

- Divide listings into 2 groups (rating>=4.9, rating<4.9)
- Independent Samples T-test
  - Two groups are distinct on numerical dependent variables.
- Chi-Square Tests
  - Examine the relationship between categorical variables.

Chi-Square Tests

Listings with rating >= 4.9 are more likely:

1. A Superhost
2. Have self-introduction

champion * host_is_superhost Crosstabulation

			host_is_superhost		
			.00	1.00	Total
champion	.00	Count	8269	4931	13200
		% within champion	62.6%	37.4%	100.0%
		% within host_is_superhost	61.0%	40.3%	51.2%
		% of Total	32.1%	19.1%	51.2%
	1.00	Count	5276	7311	12587
		% within champion	41.9%	58.1%	100.0%
		% within host_is_superhost	39.0%	59.7%	48.8%
		% of Total	20.5%	28.4%	48.8%
Total		Count	13545	12242	25787
		% within champion	52.5%	47.5%	100.0%
		% within host_is_superhost	100.0%	100.0%	100.0%
		% of Total	52.5%	47.5%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1110.112[a]	1	.000		
Continuity Correction[b]	1109.281	1	.000		
Likelihood Ratio	1117.933	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	1110.069	1	.000		
N of Valid Cases	25787				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5975.49.

b. Computed only for a 2x2 table

champion * has_host_about

Crosstab

			has_host_about		
			.00	1.00	Total
champion	.00	Count	5234	7966	13200
		% within champion	39.7%	60.3%	100.0%
		% within has_host_about	52.1%	50.6%	51.2%
		% of Total	20.3%	30.9%	51.2%
	1.00	Count	4803	7784	12587
		% within champion	38.2%	61.8%	100.0%
		% within has_host_about	47.9%	49.4%	48.8%
		% of Total	18.6%	30.2%	48.8%
Total		Count	10037	15750	25787
		% within champion	38.9%	61.1%	100.0%
		% within has_host_about	100.0%	100.0%	100.0%
		% of Total	38.9%	61.1%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.042[a]	1	.014		
Continuity Correction[b]	5.979	1	.014		
Likelihood Ratio	6.043	1	.014		
Fisher's Exact Test				.014	.007
Linear-by-Linear Association	6.042	1	.014		
N of Valid Cases	25787				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 4899.20.

b. Computed only for a 2x2 table

T-Tests

Group Statistics

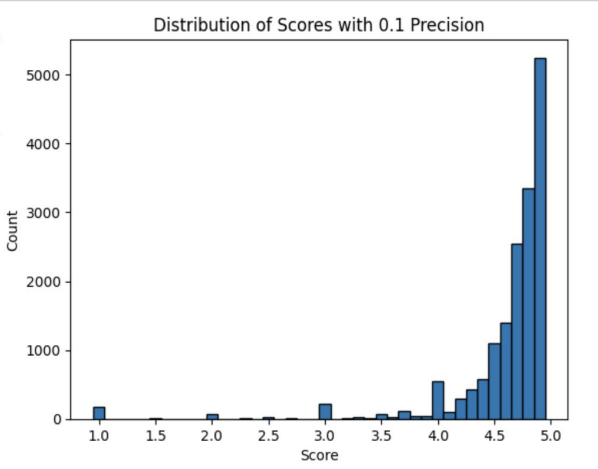
	champion	N	Mean	Std. Deviation	Std. Error Mean
calculated_host_listings_count	1.00	12587	17.9847	77.96296	.69491
	.00	13200	18.5539	60.37917	.52553
months_since	1.00	12587	79.9666	39.89264	.35558
	.00	13200	76.5337	39.17194	.34095
price	1.00	12587	272.3025	1014.92220	9.04631
	.00	13200	205.8863	296.69068	2.58236
host_response_rate	1.00	12587	.9731	.11130	.00099
	.00	13200	.9636	.12323	.00107
number_of_reviews	1.00	12587	44.7047	87.57869	.78062
	.00	13200	56.9367	91.77537	.79880

Listings with rating >= 4.9 have significantly:

1. Higher price
   a. Possible reason: better quality
2. Faster host response rate
3. Fewer reviews
   a. Possible reason:
      i. Newer Listings
      ii. Volume vs. Quality Trade-Off
4. Longer duration
   a. Possible reason: more experienced

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
calculated_host_listings_count	Equal variances assumed	14.191	.000	-.657	25785	.511	-.56920	.86609	-2.26678	1.12838
	Equal variances not assumed			-.653	23705.554	.514	-.56920	.87125	-2.27691	1.13852
months_since	Equal variances assumed	4.141	.042	6.971	25785	.000	3.43284	.49241	2.46769	4.39799
	Equal variances not assumed			6.968	25673.927	.000	3.43284	.49262	2.46727	4.39841
price	Equal variances assumed	58.524	.000	7.203	25785	.000	66.41617	9.22110	48.34231	84.49003
	Equal variances not assumed			7.060	14628.152	.000	66.41617	9.40768	47.97593	84.85640
host_response_rate	Equal variances assumed	78.808	.000	6.505	25785	.000	.00953	.00146	.00666	.01240
	Equal variances not assumed			6.521	25709.613	.000	.00953	.00146	.00666	.01239
number_of_reviews	Equal variances assumed	18.946	.000	-10.940	25785	.000	-12.23205	1.11813	-14.42365	-10.04044
	Equal variances not assumed			-10.952	25784.986	.000	-12.23205	1.11689	-14.42122	-10.04288



Multiple Regression Modeling

- **Goal**
  - Quantify the impact of various features on our target variable, **'review_scores_rating'**
- **Process**
  - Transformed **categorical** variables into **dummy** variables for compatibility with the regression model
  - Investigate the **inter-correlations** among variables to prevent collinearity
  - Experiment different modelings and variables combinations

Multiple Regression Modeling

Significant variables (y = review_scores_rating)

Positive	Negative
review_scores_communication: (+) *** review_scores_location: (+) *** review_scores_cleanliness: (+) *** review_scores_checkin: (+) *** **host_response_rate: (+) ***** **reviews_per_month: (+) ***** **months_since: (+) **	**host_acceptance_rate: (-) ***** **calculated_host_listings_count: (-) **

Conclusion

Impact and Future Vision

Project Outcome & Impact

- **Better Hosting Experience**: Identified some key contributors to review scores and guide hosts to enhance listings for improved visibility, trust, and booking rates.
- **Platform Improvement**: Airbnb can use model insights to guide homeowners, refine superhost criteria, and improve overall customer experience.

Possible Directions for Future Work

- **Advanced NLP for reviews:** Dive deeper into textual reviews to capture subtle patterns in the reviews.
- **Interaction effects:** Investigate interaction effects between variables to understand if certain combinations of factors have a more significant impact on review scores.

THANK YOU