Airbnb Superhost Analysis in Houston

MGMT 68700: AI for Business Decisions

Group 15

Hrishikesh Bhatt, Namra Shah

PURDUE UNIVERSITY

Introduction

The Superhost program is a big driver of revenue for Airbnb

- Airbnb listings in Houston hosted more than 1.8 million visitors(1) in 2023.
- Hosts in Houston generated more than $8.5 million(2) in local hotel occupancy taxes
- Globally Airbnb has more than 1 million(3) Superhosts
- In Q32022 the typical Superhost earned 64% more(4) than a regular host

---

Criteria for Superhost status (trailing 365 days)

>= 10	>= 90%	100%	>= 80%
trips hosted	**response rate** to guest requests	**confirmed reservations** completed	proportion of bookings leading to **5 star reviews**

(1), (2) https://news.airbnb.com/airbnb-generated-1-6-billion-in-economic-activity-in-houston-in-2023/
(3), (4) https://news.airbnb.com/airbnb-celebrates-1-million-superhosts/

PURDUE UNIVERSITY

Business Problem Overview

Predicting Superhost status changes to engage with existing hosts, provide incentives and reward potential superhosts

- Superhosts tend to make more income on average – This could mean greater revenue for Airbnb too
- Which are the key variables influencing Superhost status?
- How can Airbnb ensure that as many Superhosts as possible remain on Superhost status?
- How can Airbnb push non-Superhosts to Superhost status or influence hosts likely to lose Superhost status?

About the dataset

- Airbnb Evaluates Superhost status quarterly based on pre-defined criteria.
- Each row in the dataset represents a listing in one of 8 quarterly evaluation periods from July 2016 to April 2018.
- **118,677** rows (or listings)
- **111** columns (features)
- **11,797** unique hosts in the dataset

PURDUE UNIVERSITY

Approach (1/2)

We selected 19 variables that could help predict revenue and/or Superhost status

Host Performance Metrics

- prev_host_is_superhost
- ***rating_ave_pastYear***
- ***prop_5_StarReviews_pastYear***
- numReviews_pastYear
- num_5_star_Rev_pastYear
- ***numCancel_pastYear***

Historical Performance Metrics

- prev_rating_ave_pastYear
- prev_numReviews_pastYear
- prev_num_5_star_Rev_pastYear
- prev_numCancel_pastYear

Reservation and Booking Metrics

- ***numReservedDays_pastYear***
- prev_numReservedDays_pastYear
- booked_days
- available_days
- booked_days_avePrice
- available_days_aveListedPrice

Contextual and Environmental Metrics

- tract_superhosts_ratio
- listing_type
- superhost_change_lose_superhost

PURDUE UNIVERSITY

Approach (2/2)

We connected these variables to Superhost status and revenue to provide recommendations for Airbnb

1 Estimate revenue and connect Superhost status to revenue

- Calculating the relative revenue benefit from being a Superhost vs. non-Superhost
- Hypothesis testing to establish the connection between Superhost status and revenue

2 Predict Superhost status

- Use logistic regression and random forest to predict Superhost status in the next period
- Identify (a) hosts who are likely to become Superhosts in next period and (b) host who are likely to lose Superhost status in the next period

3 Identify important levers to influence Superhost status

- Select "most important" levers in predicting Superhost status based on the modelling
- Design interventions for Airbnb to take with those who are:
  - Likely to lose Superhost status
  - Likely to remain non-Superhosts

PURDUE UNIVERSITY

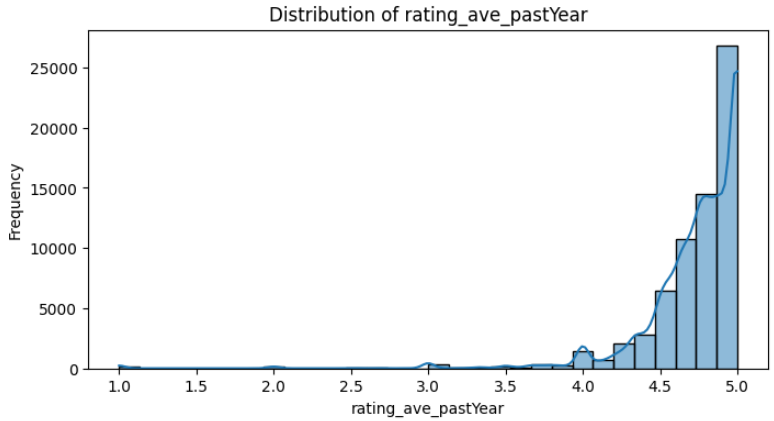
EDA and Pre-processing (1/3)

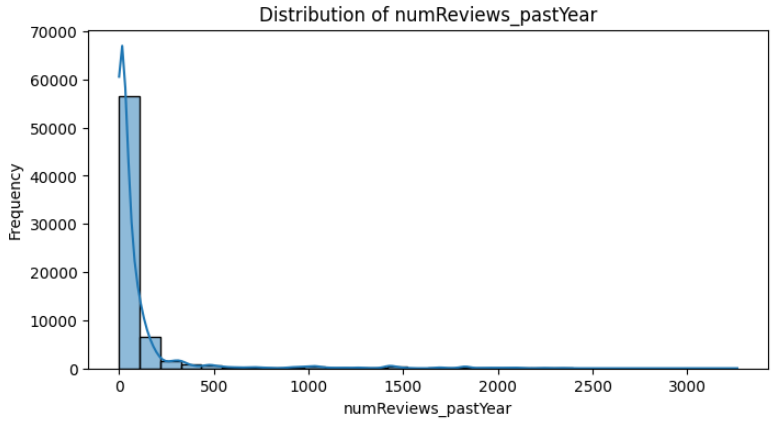
There were many missing values across the feature set of interest

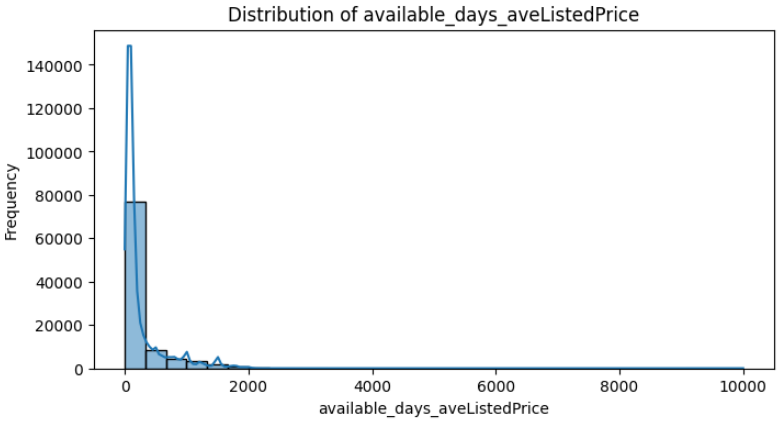
- We had between **0 – 71,733 missing values** across the features
- For interval variables, we imputed using **median**
- For class variables, we imputed using **mode**

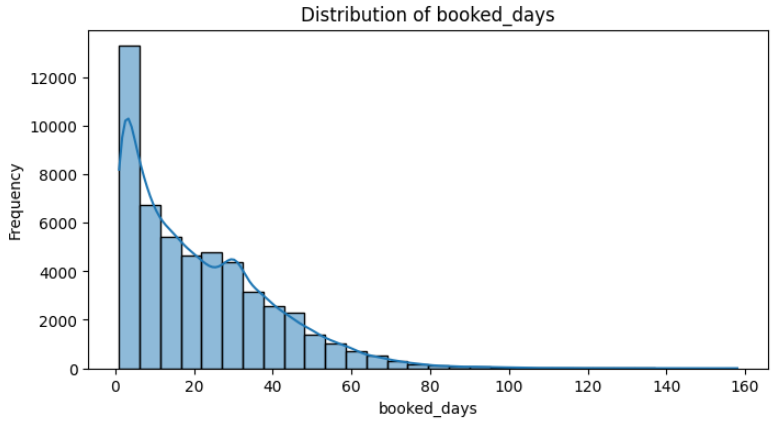
EDA and Pre-processing (2/3)

The variables of interest were highly skewed (a snapshot)







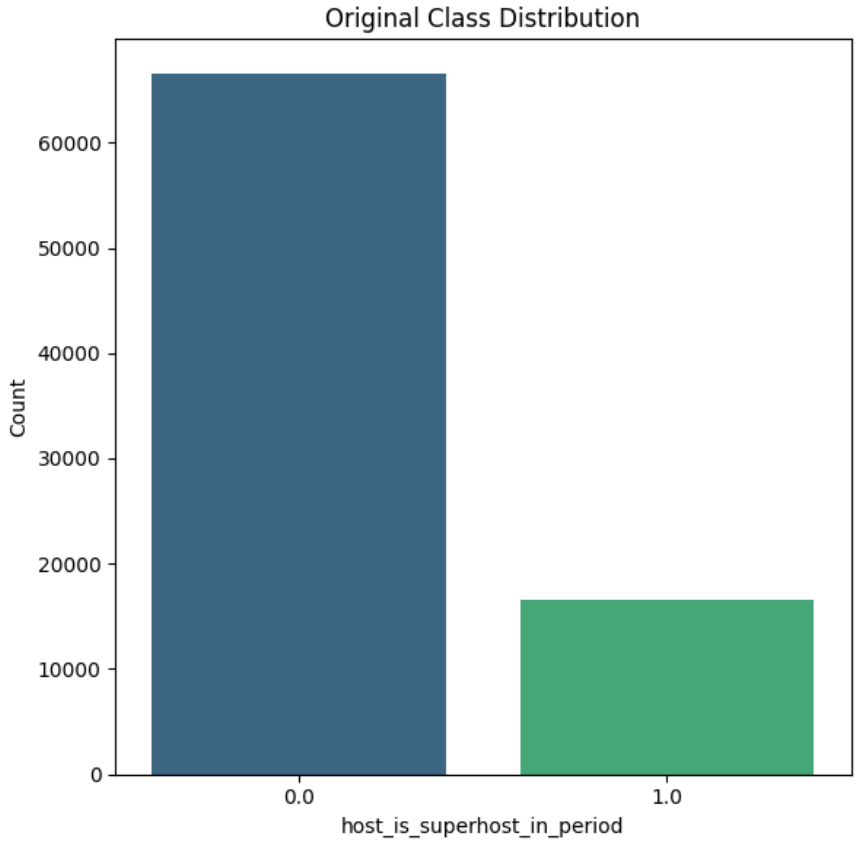


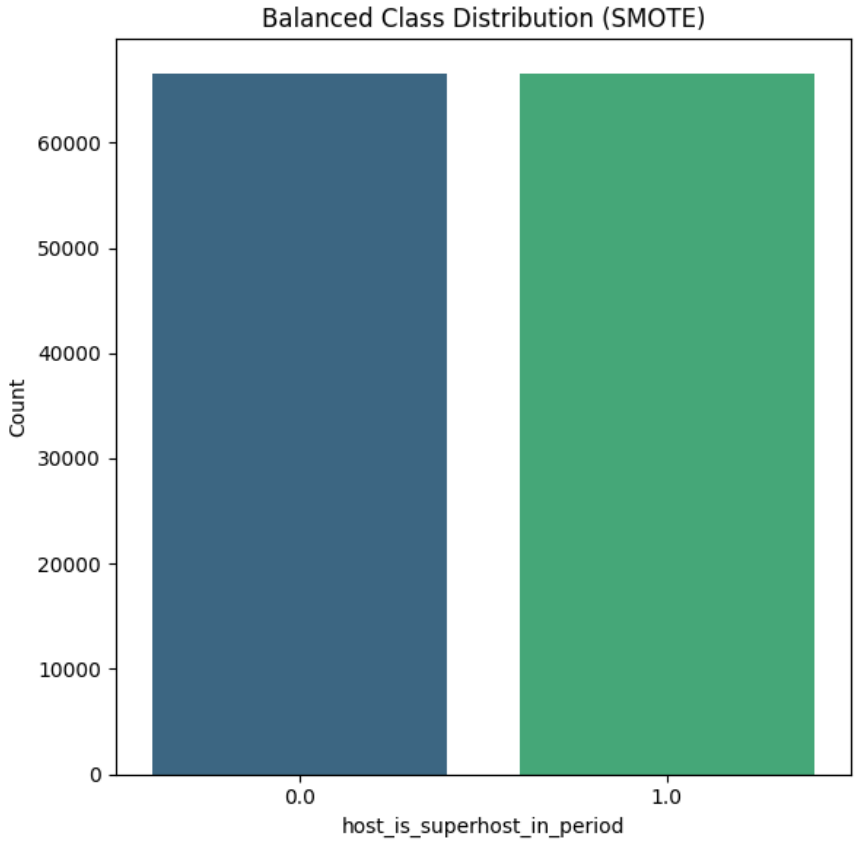
Solution: We performed a **logarithmic transformation** of the interval variables

PURDUE UNIVERSITY

EDA and Pre-processing (3/3)

The class target was highly unbalanced





Solution: We used the SMOTE algorithm to balance the class target

PURDUE UNIVERSITY

Modelling Revenue (1/2)

We ran a hypothesis test to check the impact of superhost status on revenue

Null Hypothesis
"There is no difference in mean revenue between Superhosts and non-Superhosts"

$$\mu_{Superhosts} = \mu_{Non\text{-}Superhosts}$$

Test Type
Independent samples t-test to compare the means of the two groups

---

Data used
We computed "estimated_revenue" for each host using the formula:

estimated_revenue = booked_days x booked_days_avePrice

We then used the mean of estimated_revenue for Superhosts vs. non-Superhosts

---

Results -

- **t-statistic** = 25.109
- **p-value** < 0.000

Conclusion: Reject the null hypothesis. Superhost status significantly impacts revenue

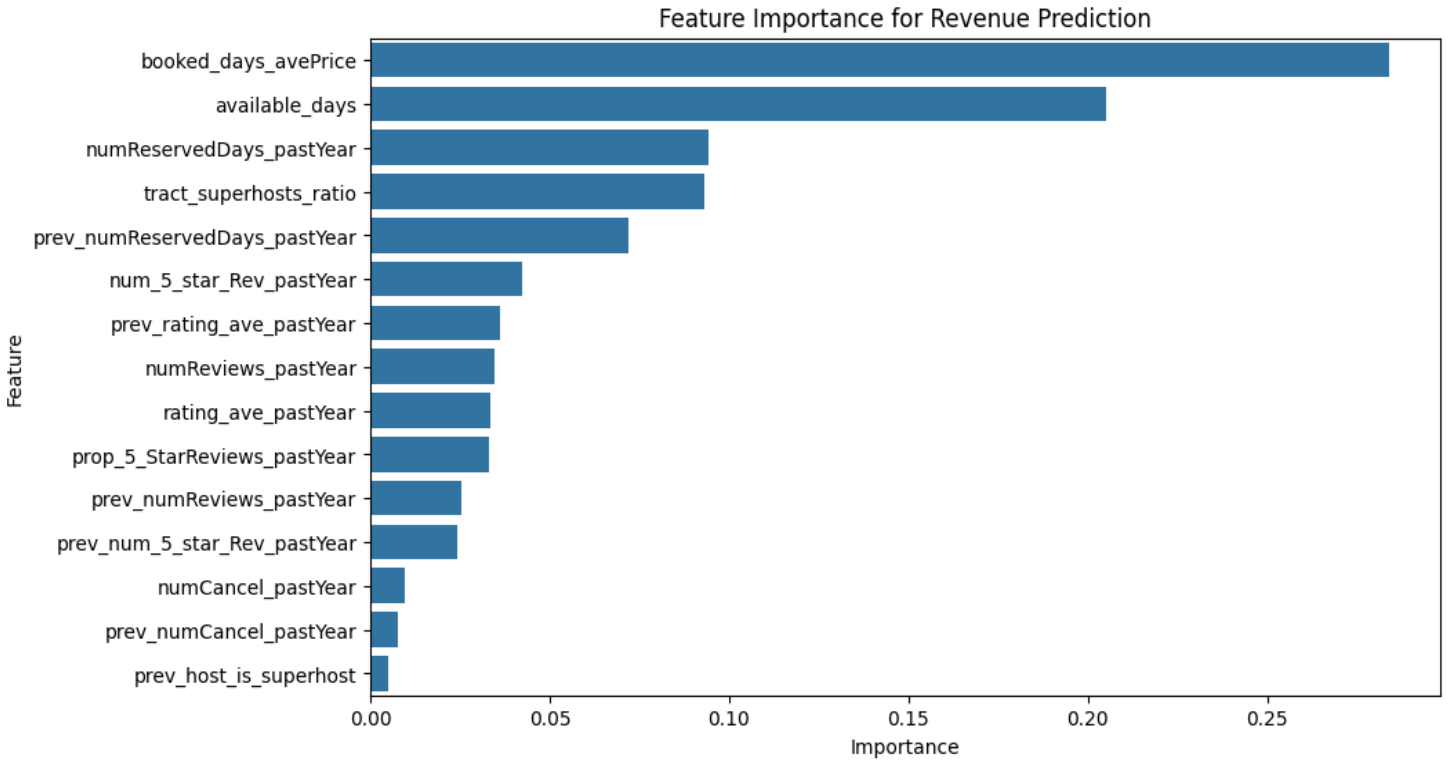
PURDUE UNIVERSITY

Modelling Revenue (2/2)

We trained a random forest model to predict revenue and analyzed feature importance

Model results:

- MAE = 1.262
- MSE = 5.941



PURDUE UNIVERSITY

Predicting Superhost Status (1/2)

We trained a logistic regression model to predict Superhost status

Approach

- We performed a 70-30 train-test split on the data and trained a logistic regression model using the statsmodels library
- If the resulting probability was >50%, we predicted that they would be a Superhost in the next period. For <= 50%, we predicted that they would not be a Superhost

Model results

- Pseudo R-squared = 0.7145
- ROC AUC score (test data) = 0.975

This indicates robustness in predicting Superhost status

Confusion Matrix (NS = Non-Superhost, S = Superhost)

(Predicted) \ (Actual)	NS	S
NS	27760	743
S	1268	5833

- The confusion matrix indicates high specificity (97%), showing a strong ability to identify non-Superhosts
- It has moderate sensitivity (82%), indicating that it misses about 18% of actual Superhosts

PURDUE UNIVERSITY

Predicting Superhost Status (2/2)

The model gave us important insights about host status in the next period

2,148	821	1,881	11,236
Predicted to maintain Superhost status	Predicted to gain Superhost status	Predicted to lose Superhost status	Predicted to maintain non-Superhost status

Factors positively associated with Superhost status

- Being Superhost in the previous period (prev_host_is_superhost) is very important, indicating reputational importance
- Average rating in the past year (rating_ave_pastYear) is very important
- Ratio of Superhosts in the census tract (tract_superhost_ratio) is important, indicating regional trend

Factors negatively associated with Superhost status

- Number of reviews in the past year (numReviews_pastYear), indicating reviews tend to be negative
- Cancellations in the past year (numCancel_pastYear), indicating that cancellation is a very strong red flag for visitors

PURDUE UNIVERSITY

Key variables and Insights

Alignment with Official Criteria:

- **Number of Stays (numReservedDays_pastYear):** Ensures hosts have adequate experience and engagement.
- **Response Rate (rating_ave_pastYear, indirectly through reviews):** Reflects the host's promptness and reliability in communication.
- **Cancellation Rate (numCancel_pastYear):** Maintains guest trust by minimizing disruptions.
- **Review Ratings (rating_ave_pastYear, prop_5_StarReviews_pastYear):** Directly correlates with guest satisfaction and overall experience
- **Average price (booked_days_avePrice):** The strongest predictor of revenue, highlighting the importance of competitive pricing.
- **Availability (available_days):** Indicates that ensuring high availability can maximize booking opportunities
- **Additional variables found in our analysis which also impact superhost status :**
  - Average listing price for available days - **available_days_aveListedPrice**
  - Ratio of superhosts in the census tract - **tract_superhosts_ratio:**

  Superhosts dominate in competitive areas; differentiation in services can mitigate the competition effect

PURDUE UNIVERSITY

Recommendations for Airbnb

Retaining Superhosts and Incentivizing non-Superhosts

Analyze local market and seasonal trends	Alternative recognition and training programs	Offer cancellation insurance and flexible booking options	Implement automation to boost interactions

Maximizing Impact:

- Focus on High-Impact Variables:
  - Prioritize improving **rating_ave_pastYear and prop_5_StarReviews_pastYear** as these have the strongest correlation with Superhost retention.
- Integrated Approach:
  - Combine efforts to enhance multiple variables simultaneously, such as improving response rates while maintaining low cancellation rates. Variables that we found also proved to be crucial: **available_days_aveListedPrice, tract_superhosts_ratio**
- Data-Driven Decision Making:
  - Use predictive models to identify which variables, when improved, will have the most significant impact on retaining or achieving Superhost status.

Increase Revenue as Superhosts:

Value based upselling and discounts	Improve conversion rates	Boost ratings and collaborate locally	Stand out in search and social media

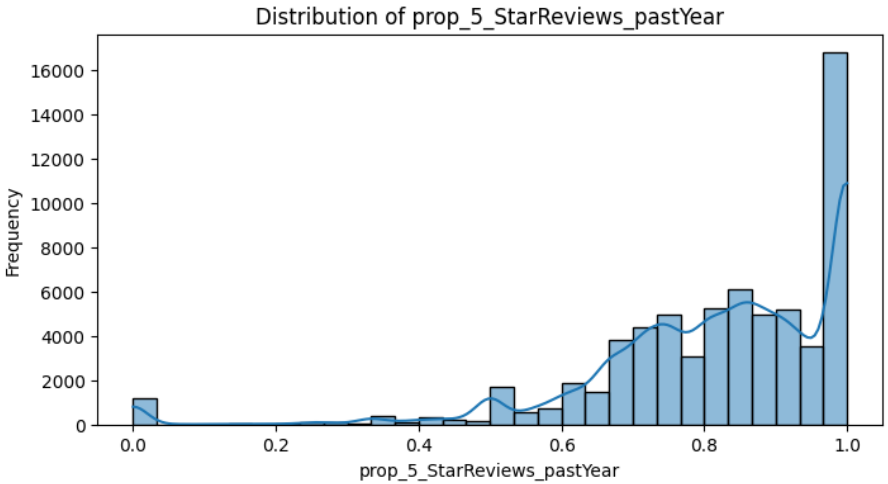
PURDUE UNIVERSITY

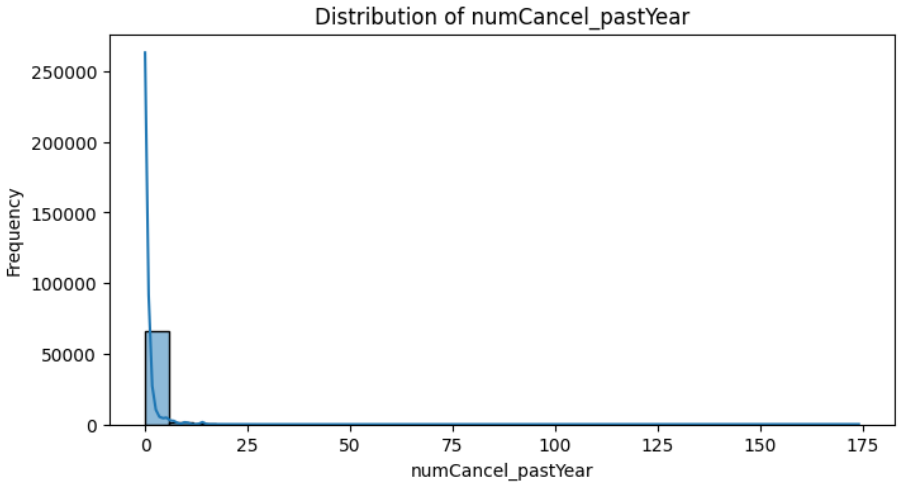
Appendix

Additional results from the EDA and modelling process

Exploratory Data Analysis

Other variables showing host performance were highly skewed as well





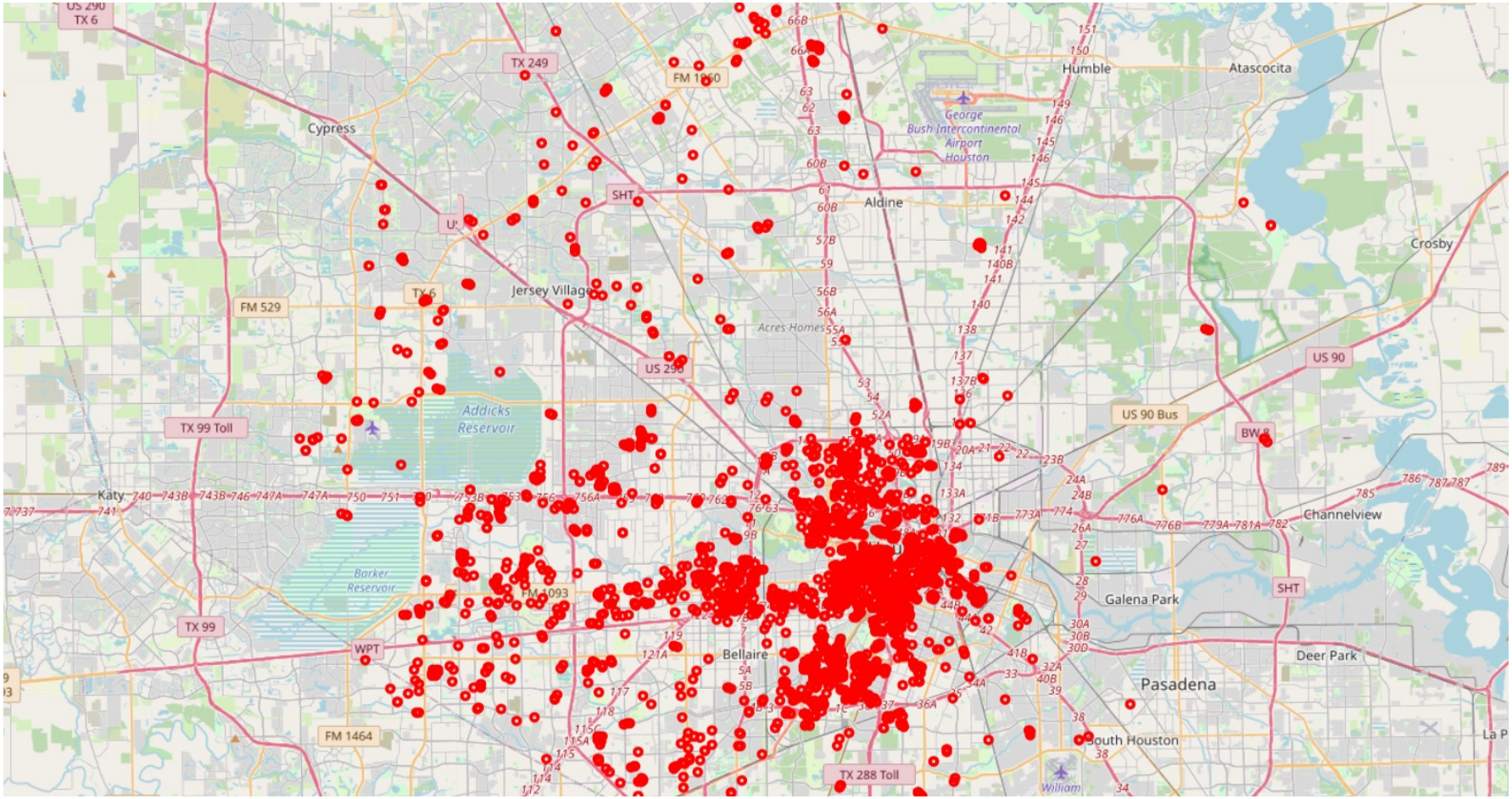
- Distribution of Proportion of 5-star reviews is highly skewed, showing that the Superhosts who crack the customer satisfaction puzzle achieve results consistently.
- Surprisingly, there are a lot of Superhosts who have all their reviews as 5 – star reviews (proportion = 1)

- A similar trend is reflected in number of cancellations (less is better).
- A significant number of capable hosts ensure that the number of cancellations is 0.

PURDUE UNIVERSITY

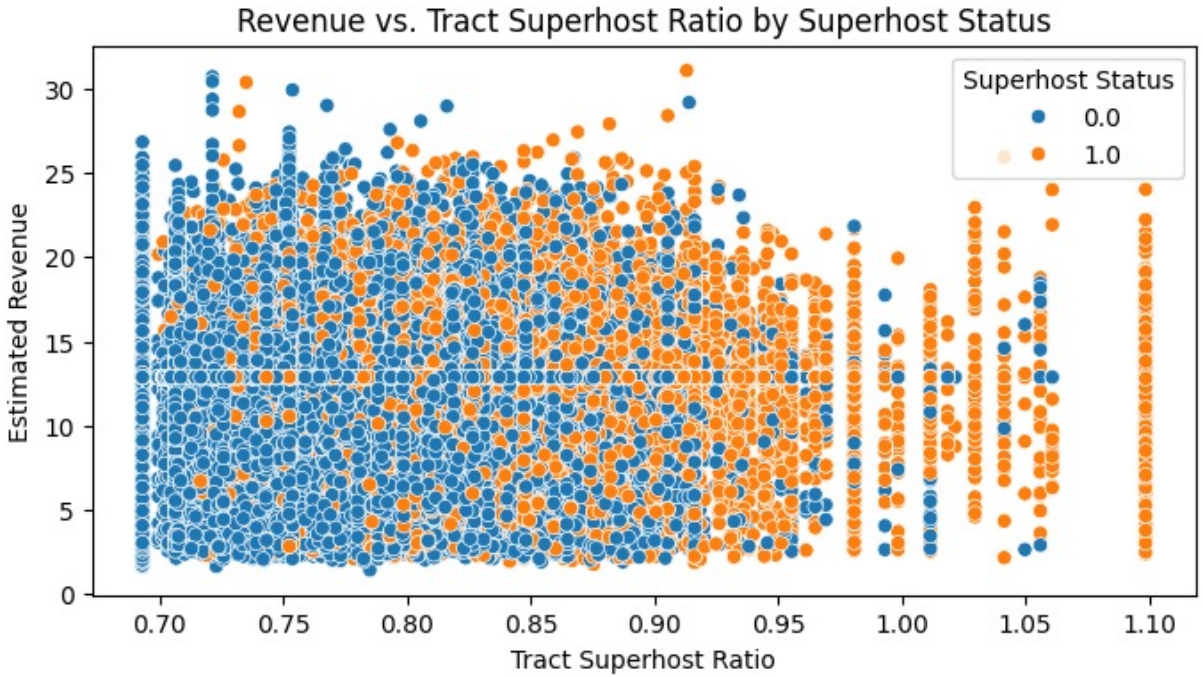
Exploratory Data Analysis

Superhost loss map for Houston shows losses concentrated in central and western region



Exploratory Data Analysis

Tract Superhost ratio is positively associated with Superhost status

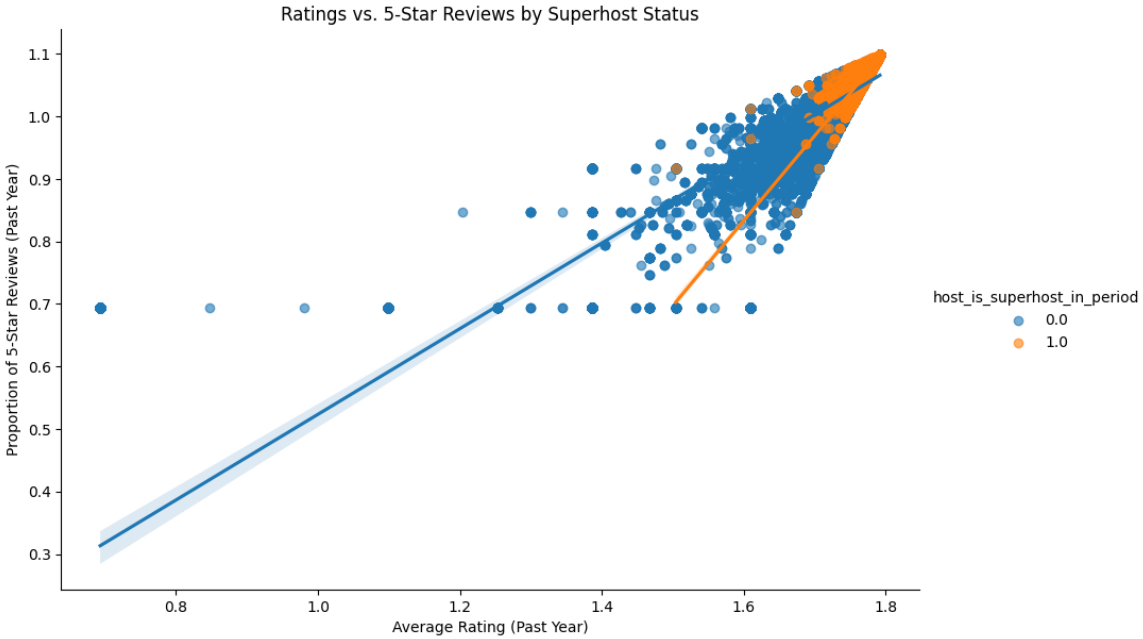


- While revenue varies widely for both groups, Superhosts are more likely to achieve higher revenue, even in highly competitive areas with a high tract superhost ratio.
- Non-Superhosts appear to clustered in less competitive tracts, indicating they may face challenges in thriving amidst intense competition.
- Overall, the data underscores the significant advantages of being a Superhost

PURDUE UNIVERSITY

Modelling Revenue

Ratings vs. 5 – Star reviews by host status

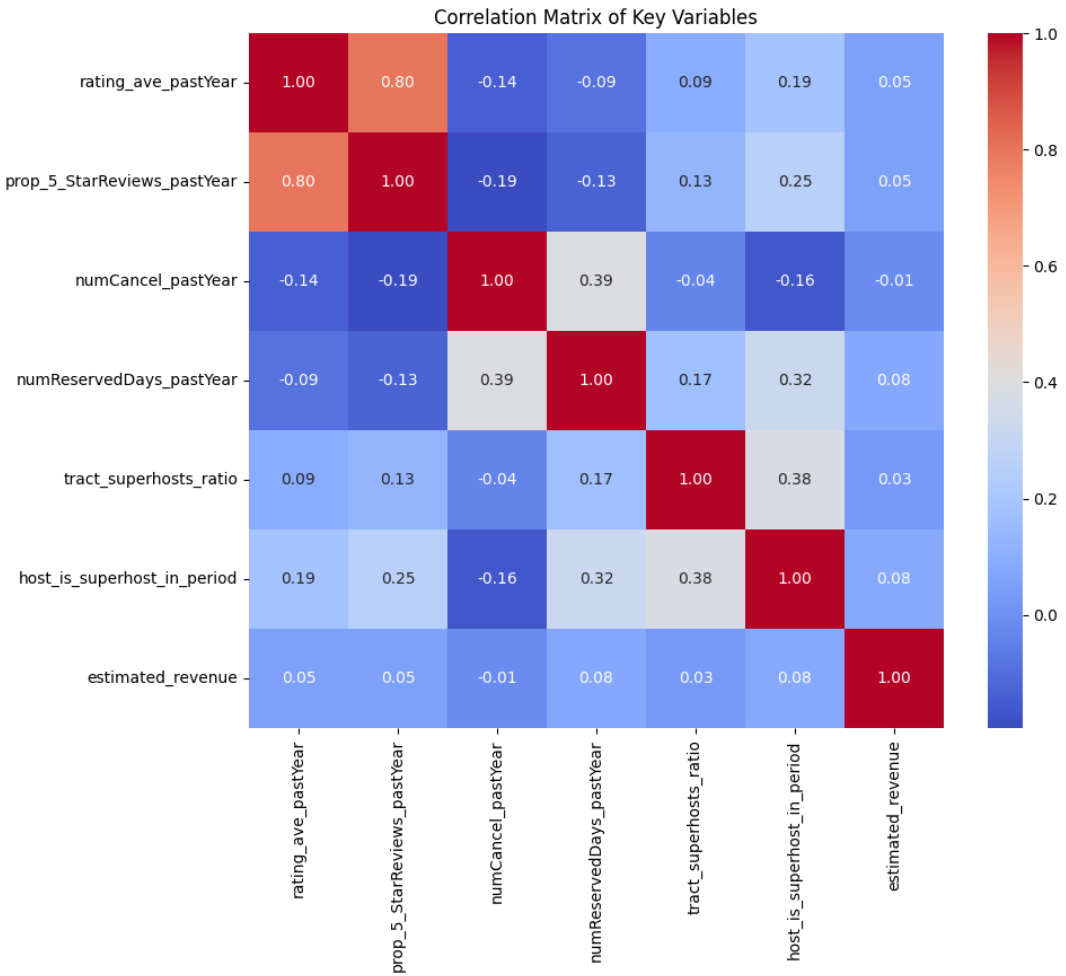


- Superhosts (orange points) consistently outperform non-Superhosts (blue points).
- Superhosts' proportion of 5-star reviews showcases their ability to deliver consistently exceptional experiences.
- The trendlines reinforce this advantage, as Superhosts show a higher baseline, highlighting their superior performance in maintaining customer satisfaction.

PURDUE UNIVERSITY

Modelling Revenue

Correlation matrix of key variables from revenue estimation



- Only 5-star reviews as proportion of total reviews is highly correlated
- Number of reserved days in the past year and tract superhost ration are mildly correlated

PURDUE UNIVERSITY

Predicting Superhost Status

Nearly all coefficients from the Logistic Regression were significant

	coef	std err	z	P>|z|	[0.025	0.975]
const	−116.4264	5.296	−21.983	0.000	−126.807	−106.046
prev_host_is_superhost	3.8298	0.045	85.585	0.000	3.742	3.917
rating_ave_pastYear	97.2107	3.712	26.189	0.000	89.936	104.486
prop_5_StarReviews_pastYear	−61.2670	1.826	−33.552	0.000	−64.846	−57.688
numReviews_pastYear	−19.9805	0.731	−27.334	0.000	−21.413	−18.548
num_5_star_Rev_pastYear	21.5575	0.748	28.803	0.000	20.091	23.024
numCancel_pastYear	−9.3583	0.258	−36.249	0.000	−9.864	−8.852
numReservedDays_pastYear	0.4696	0.021	22.165	0.000	0.428	0.511
prev_rating_ave_pastYear	1.2531	0.932	1.345	0.179	−0.573	3.079
prev_numReviews_pastYear	2.9952	0.397	7.543	0.000	2.217	3.773
prev_num_5_star_Rev_pastYear	−3.8837	0.418	−9.293	0.000	−4.703	−3.065
prev_numCancel_pastYear	2.1293	0.174	12.209	0.000	1.787	2.471
prev_numReservedDays_pastYear	−0.1787	0.021	−8.614	0.000	−0.219	−0.138
booked_days	0.1996	0.026	7.703	0.000	0.149	0.250
available_days	−0.1911	0.024	−7.925	0.000	−0.238	−0.144
booked_days_avePrice	0.2129	0.048	4.423	0.000	0.119	0.307
available_days_aveListedPrice	−0.5857	0.037	−15.996	0.000	−0.657	−0.514
tract_superhosts_ratio	13.2198	0.277	47.759	0.000	12.677	13.762

PURDUE UNIVERSITY

Thank You

PURDUE UNIVERSITY®