BAX 452 Machine Learning Final Project

Airbnb Listings in San Diego

Master of Science in Business Analytics

University of California - Davis

Huixin ‘Blessia’ Li
Hanish Singla
Jialu ‘Carol’ Wang

Table of Contents

Executive Summary

This report aims to address the business problem of enhancing Airbnb's competitiveness in the travel and hospitality industry in San Diego. Based on the dataset Airbnb listings in San Diego as of August 2019, our analysis focuses on EDA and model buildings, which involves CART, random forest, and lasso regression etc. We are able to achieve 70% accuracy in predicting the price variation using various factors involving property location, customer feedback, hotel policies and amenities. We also identified important factors in determining these price fluctuations out of these which can be used for strategic decision making. We recommend managers to evaluate these factors to seek an advantage in the market. Some insights center around pricing strategies, property management, customer experience, and marketing. Although these approaches are a new way of proceeding into the future, these should be adopted organically over a planned period and have safeguards which can detect any inconsistencies in early stages.

Background & Domain Knowledge

Airbnb is a global online marketplace that connects travelers with hosts who offer unique accommodations and experiences. Founded in 2008, the platform has grown exponentially and now operates in over 220 countries and regions. Airbnb allows property owners, known as hosts, to list their accommodations for short-term rentals, providing guests with an alternative to traditional hotels and other lodging options. The company's primary product is its platform, which allows hosts to list their accommodations and guests to search for and book suitable stays. The platform offers various types of accommodations, including entire homes, private rooms,

and shared rooms. In addition to accommodations, Airbnb also offers "Experiences," which are unique activities or events hosted by locals. The platform has various features and services for both hosts and guests, such as listing management, booking management, reviews and ratings, and secure payments, and customer support.

San Diego, with its beautiful beaches, vibrant culture, and numerous tourist attractions, has become a popular destination for both domestic and international travelers. The short-term rental market in San Diego has experienced significant growth in recent years, largely driven by the popularity of Airbnb and other similar platforms. The growth of short-term rentals has sparked increased competition among hosts and property managers, who are looking for ways to stand out and attract more bookings. In San Diego, Airbnb faces competition from other short-term rental platforms, such as VRBO and Booking.com, as well as traditional lodging options like hotels, motels, and bed & breakfasts.

Traditional Approaches

In order to stay ahead of the competition, hospitality companies in the industry have employed a variety of approaches:

- For pricing, traditional strategies to tackle this problem include researching comparable properties in the area and adjusting prices based on seasonal demand, local events, or day of the week.
- For property management, traditional strategies involve analyzing booking patterns, demand and pricing to optimize room occupancy rates and maximize revenue. Hotels and

other hospitality providers often use revenue management software or employ revenue managers to implement these approaches.

- For marketing, traditional methods to improve visibility include crafting compelling listing titles and descriptions, showcasing high-quality photos, and promoting listings on social media or through partnerships with local tourism organizations.
- For customer experience, strategies to enhance guest experience include providing detailed check-in instructions, offering a clean and well-maintained property, and being responsive to guest inquiries and issues.

By analyzing the dataset of all active Airbnb listings in San Diego as of August 2019, hosts and property managers can gain valuable insights to inform their strategies and improve their competitiveness in the market. This includes optimizing property features, pricing, and guest experience to maximize bookings, revenue, and customer satisfaction.

Analysis

EDA

The data set we used was Airbnb listings for San Diego in 2019. It contains 75 columns of variables and 13,051 observations, which record in detail the information of each post (such as its link, introduction, space, geographical location, etc.), information of each host (such as the time of joining Airbnb, response rate, etc.), and information related to the reviews (such as the total number of comments, the time of the last evaluation, etc.). The data was cleaned first by removing the "$" and "," from the price class, such as "price_per_stay", and converting it to floating format. Then we replace the "NA" value in each column. For instance, we find that the NA value in "reviews_per_month" is because the post has not been commented on, so we assign

it a value of 0. For the rest of the numerical data such as 'cleaning_fee', 'security_deposit', we think it is reasonable to assign an average value. For other categorical data such as 'host_is_superhost', we use the mode to replace NA.

We checked the correlation between variables (Appendix 1.1) and found almost no strong correlation between them, except for data related to reviews, such as "review_scores_rating," "review_scores_accuracy," "review_scores_cleanliness," "review_scores_checkin," "review_scores_communication," "review_scores_location," and "review_scores_value."

Additionally, we created a visualization of Airbnb prices in different geographical locations (Appendix 1.2). The figure showed no clear relationship between prices and latitude and longitude, but it revealed a connection between prices and the region and neighborhood (Appendix 1.3). For instance, we noticed dense dark blue dots around the San Diego Zoo, indicating relatively cheaper housing prices in that area. Another graph (Appendix 1.4) showed that Airbnb rentals in San Diego tended to be entire rooms/apartments rather than shared rooms.

Finally, we observed from the scatter chart (Appendix 1.5, 1.6) that the housing price of Airbnb was inversely proportional to the number of comments and cleaning costs. The reason for this could be that cheaper Airbnb attracts more guests, and as a result, more comments are made. Besides, some hosts may lower the housing price listed in the post by raising the cleaning cost.

Model Development, Estimation and Result

We are trying to leverage historic data for two aspects here: firstly, based on the given parameters, we wish to predict the optimum price for properties accurately; secondly, understand which factors contribute to higher price and see if property managers can access these factors as levers for profit maximization.

We have leveraged different machine learning algorithms to evaluate which one would be able to better model the variance in price using the below approach:

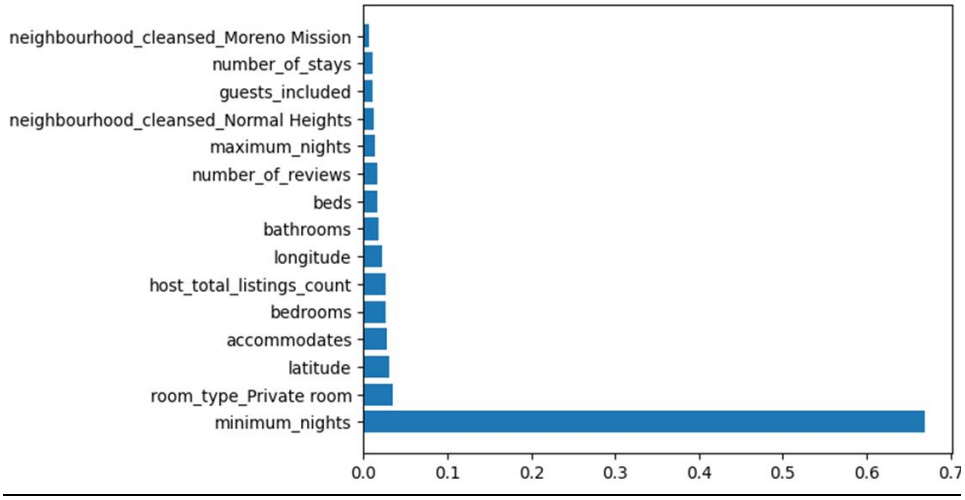
1. Split our dataset into training and test datasets in the ratio of 80:20.
2. Ran linear regression as a proof of concept using all variables to see the percentage of variation which can be explained via these parameters.
3. Standardizing the dataset and performing cross validation on the training dataset to find a suitable lambda value for lasso regression and then using that lambda value (.01) for lasso regression. This step-in addition to prediction will also assist us in feature selection as non-important variables will shrink to zero. A total of 58 variables were found to be non-zero. (Appendix 2.2)
4. Ran CART, Random Forest, XGBoost algorithms which would be able to also see if there are any interactions between variables and have higher accuracy. (Appendix 2.3,2.4)
5. Given a lot of these variables are correlated with each other for example - higher no of beds will have higher no of bathrooms as well. To account for these correlations, we also ran PCA to lower the number of predictors. The total number of predictors were selected based on the amount of explained variation (75% cutoff chosen) and then the amount of decrease in MSE for each increase in predictor (based on the elbow rule).
6. After finding these appropriate predictors (20), we rerun the steps from 2-4 with those limited predictors to see if they have comparable performance with original models. (Appendix 2.5)

Results:

Model	Test Rsquare	Test MSE	Comments
Linear Regression	20.12	861	
Lasso Regression	21.18	1101	Lamda - 0.10
CART	56.27	636.27	
Random Forest	70.34	519	Total Estimators - 100
XGBoost	68.85	538.65	Total Estimators - 10
PCA Linear Regression	19.36	866.12	Total NC = 25
PCA CART	32.87	790	Total NC = 25
PCA Random Forest	61.08	601	Total NC = 25
PCA XGBoost	42.65	730	Total NC = 25

R Square < 30
R Square > 30 and < 60
R Square > 60

Prediction Model Summary



Top 15 features based on Random Forest

More detailed results are available in the results section, but most important takeaways based on our results are mentioned below:

A) Random Forest Model is performing the best out of all the other models and one of the reasons which can be attributed for this is that random forest uses the approach of bagging, through which it is able to generalize any biases.

B) After re-running all models with PCA selected components, we see that although we are just using around just 25 predictors instead of original (>138 predictors), we still get

comparable model performance. These can be used for businesses where they can't spend much on computational power and model efficiency.

C) All the important factors mentioned above can be categorized as:

a) Location Parameters: Latitude, Longitude, Neighborhood

b) Room Parameters: Room Type, Bedrooms, Accommodates, Beds, Bathrooms

c) Booking Parameters: Minimum No of Night stays required, Max No of Nights, Guests Included

d) Review/Feedback Parameters: Number of reviews, Total host listing counts

All these features intuitively make sense on how they can have an important role to play in deciding the overall price for the property.

Limitations

We can build onto this analysis by adding data from other properties apart from Airbnb to get a more general sense of the market. We can also try to look at other locations and see how this model compares to those locations. We also understand, this data did not cover factors such as 'Brand', 'Yearly/Seasonally booking trend' which can make the analysis more comprehensive.

Recommendations

Based on the four important categories and independent variables identified in our analysis, we can devise a pricing framework for the Airbnb properties in San Diego mentioned below. These can also translate into dos and don'ts thorough which property owners can optimize their profitability and beat off their competitors:

- Non - Controllable Factors: Location and Room Parameters - Both these parameters are non-controllable in the sense that property owners don't have the flexibility to tweak them or work towards changing them on a regular basis. So instead, they can use these factors to evaluate their competitiveness in the market and then be smart about deciding the price. For example: Offering discounts if they lie in a low footfall zone and charge a premium if they lie in a high footfall or near shoreline areas.
- Controllable Factors: Managers can exercise control over areas such as ‘Minimum nights’ required to book a hotel, ‘Enhancing customer experience’ where they can find innovative ways to make up for any of the non-controllable factors. For example: number of bedrooms can be made up by providing more amenities such as Cab service or an activity room with games. They can be more proactive in these arenas which would help them go a long way in penetrating the market and getting a higher control over prices.

Some actionable insights for the managers are mentioned below:

- Pricing strategies:
    - Analyze booking patterns, such as seasonality or events that may influence demand. This can help hosts adjust their prices and availability to maximize occupancy rates and revenue. Time series models or other forecasting techniques can be employed for this purpose.
    - Implement price differentiation between room types, as private rooms are often more sought after and can command higher prices.

    - Encourage customers to book more nights or stay longer by offering a loyalty program if they stay a minimum no of nights or provide discounts for extended stays (more nights mean more discounts).

- Property management:
    - Optimize property availability by analyzing booking patterns, such as peak seasons, off-peak seasons, and popular local events or attractions, to identify periods with higher demand. Strategically block or unblock dates on the calendar based on these patterns, ensuring that the property is available during high-demand periods to maximize occupancy rates and revenue. Set minimum and maximum stay requirements for peak seasons or events to optimize turnover and reduce vacancy gaps between bookings.
    - Use dynamic pricing that automatically adjusts prices based on real-time market data and booking patterns. Regularly review and fine-tune pricing rules to ensure they remain competitive and reflect the property's unique offerings and target market. Additionally, monitor performance metrics, such as average daily rate and revenue per available room, to evaluate the effectiveness of the dynamic pricing strategy and make adjustments as needed.
- Customer experience:
    - Offer flexibility in adding more beds to accommodate larger groups or families.
    - Incentivizing customers to leave reviews for their hotels.
    - Continuously improve property amenities and services based on guest feedback.
- Marketing and branding:

- Advertise properties based on the important variables identified above and their unique features and offerings.
- Encourage hosts to engage in online and offline advertising, public relations, social media marketing, and partnerships with local businesses or tourism organizations.

Lastly, these recommendations are more generally true and have a solid rationale, but as said ‘One size doesn't fit all’. So before implementing these, we also recommend users to do a small A/B test before a full-fledged launch just to test the waters and see if there are any factors specific to that situation or their property where a given recommendation doesn't work.

Conclusion

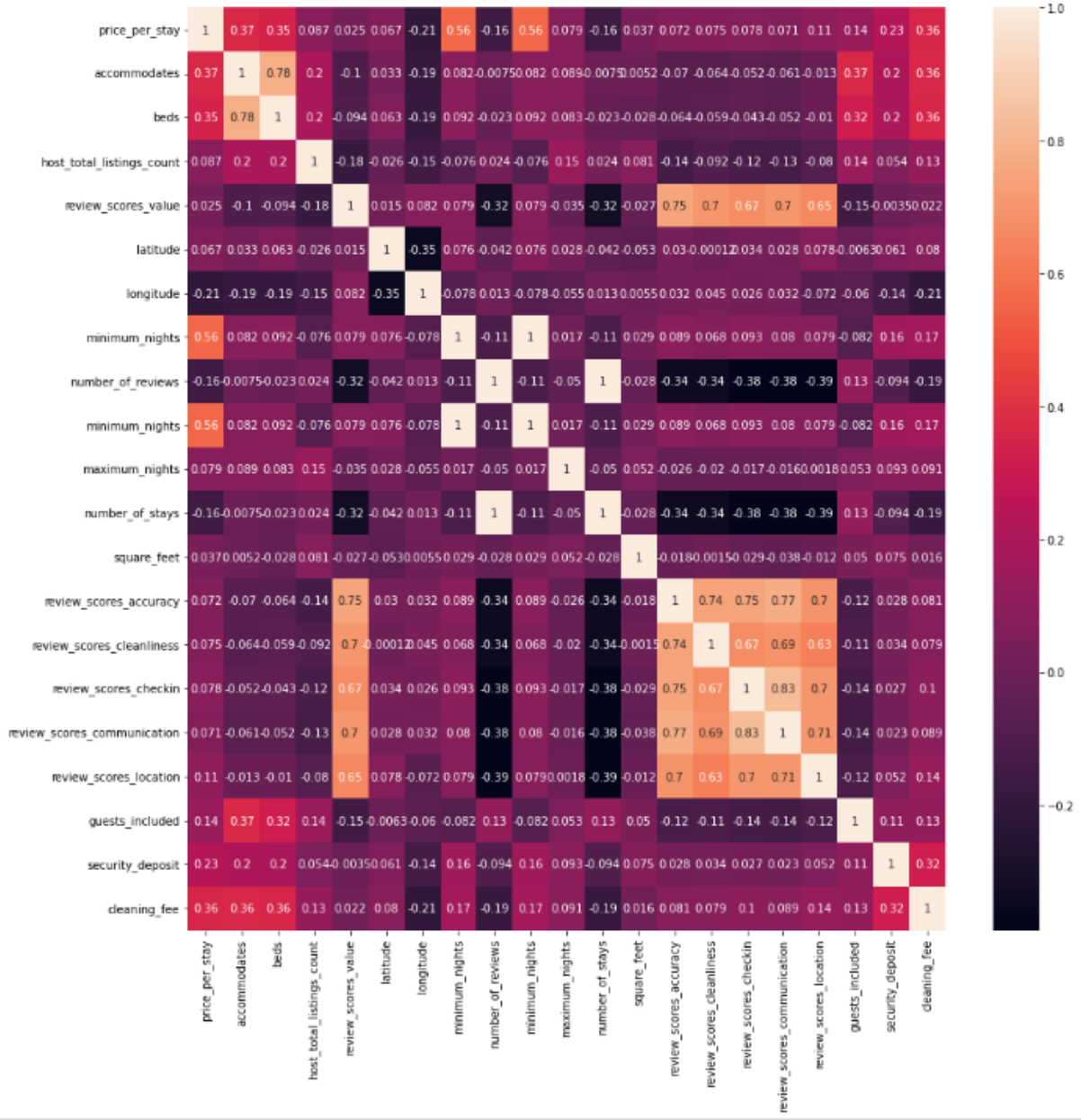
Pricing has always been a slippery slope therefore most companies tend to be very cautious in making even slight updates to it. As depicted in this analysis, a data driven approach provides managers a good benchmark to compare prices with their contemporaries and can also act as a basis for any changes to it. But still in early stages, we recommend users to deploy these data driven approaches in conjunction with their traditional methods as a safeguard which would help catch any inconsistencies.

References

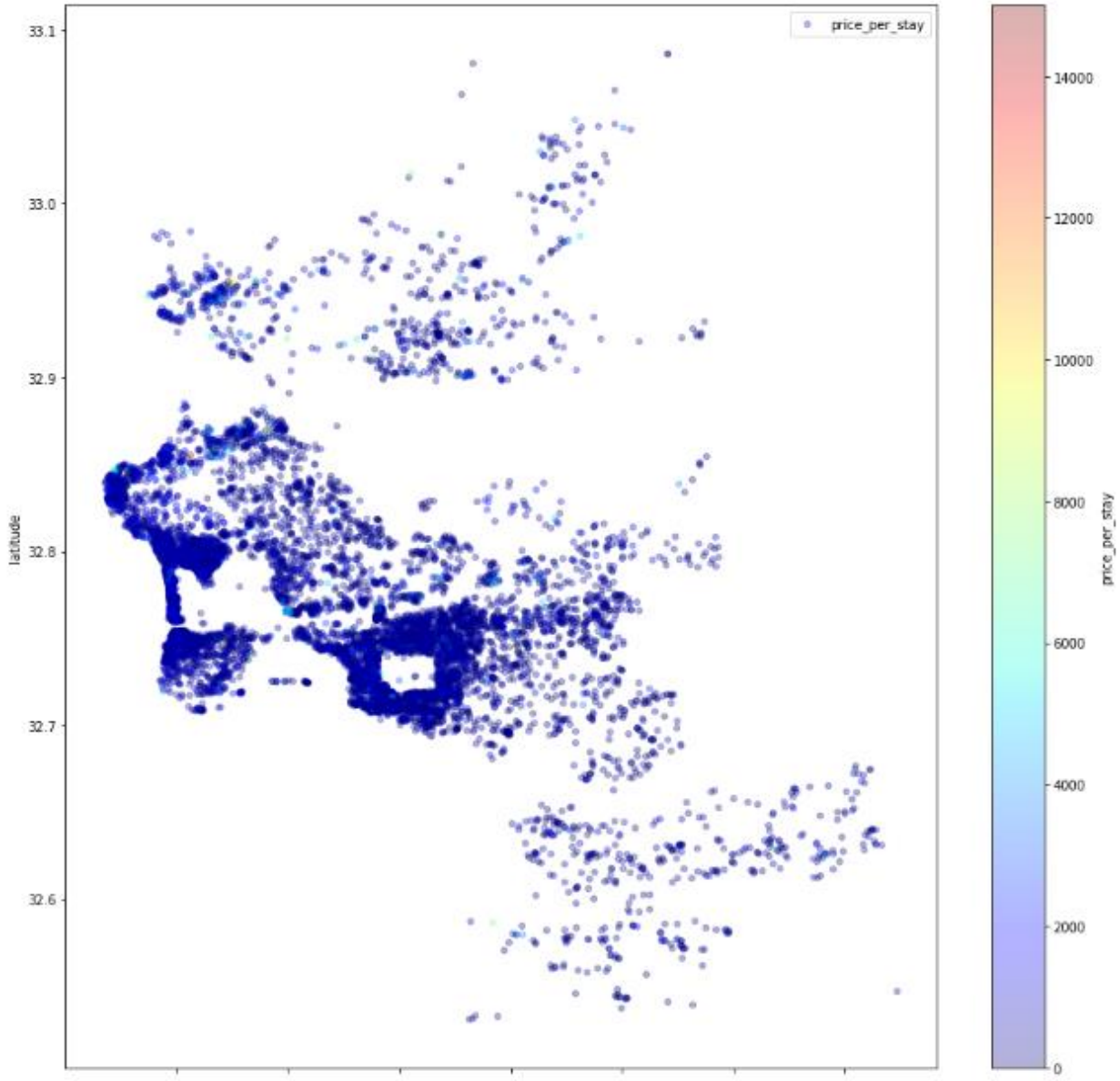
[1] Airbnb-Price-Prediction/Airbnb Price Prediction_2020.ipynb at main · ShaoniMukherjee/Airbnb-Price-Prediction · GitHub
[2] Data Exploration on NYC Airbnb | Kaggle
[3] San Diego 2019 Airbnb dataset

Appendix

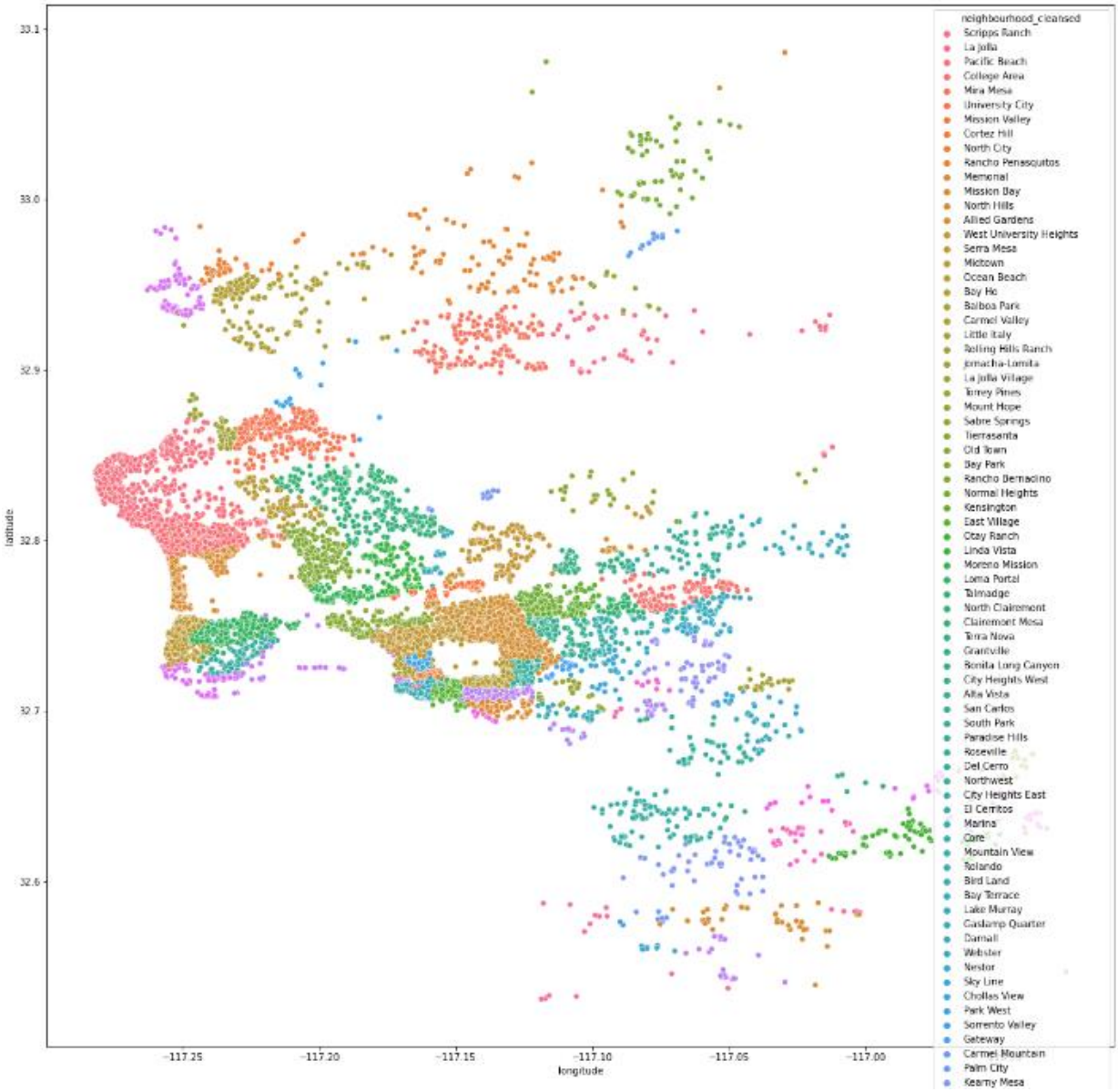
Appendix 1.1 The correlation between variables



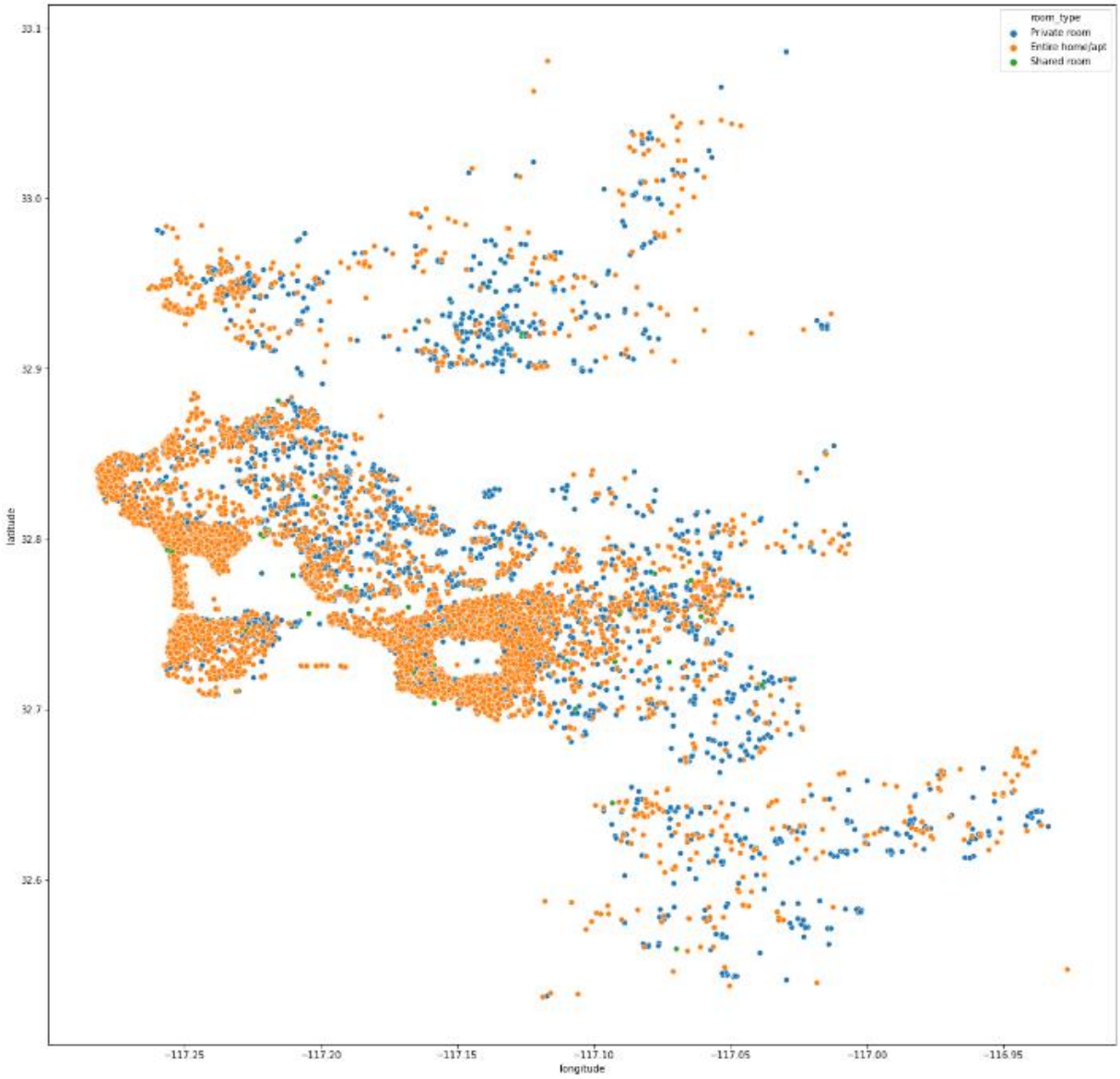
Appendix 1.2 Airbnb prices in San Diego reflected on the map.



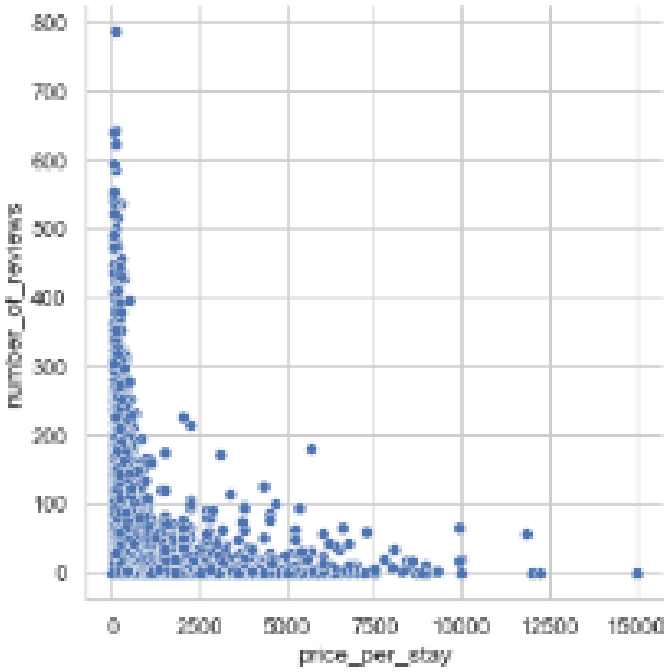
Appendix 1.3 Distribution of neighborhood in San Diego



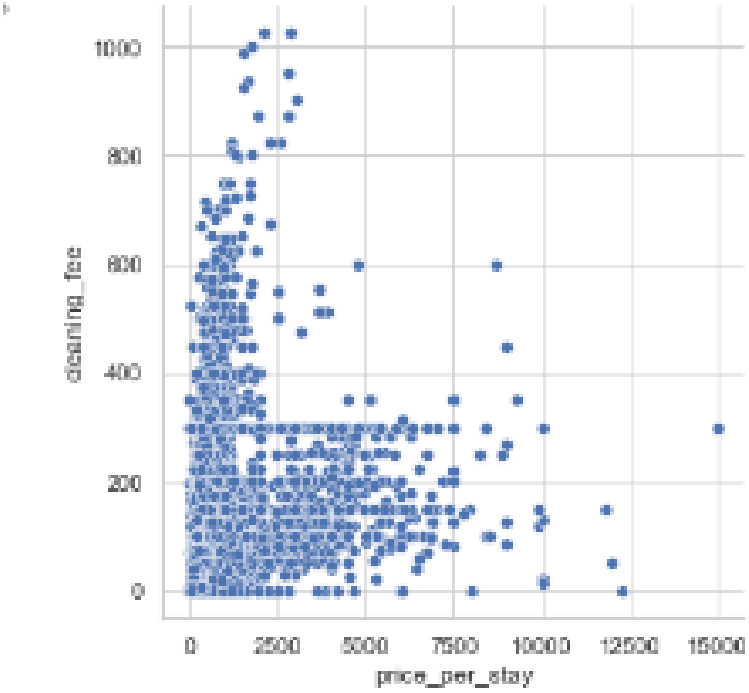
Appendix 1.4 The map of different airbnb room types in San Diego



Appendix 1.5 The relationship between the price and the number of reviews



Appendix 1.6 The relationship between the price and cleaning fee



Appendix 2.1 The Results of Linear Regression with Intercept

OLS Regression Results

Dep. Variable:	price_per_stay	R-squared:	0.691
Model:	OLS	Adj. R-squared:	0.682
Method:	Least Squares	F-statistic:	75.23
Date:	Thu, 23 Mar 2023	Prob (F-statistic):	0.00
Time:	12:56:38	Log-Likelihood:	-26307.
No. Observations:	4322	AIC:	5.287e+04
Df Residuals:	4196	BIC:	5.367e+04
Df Model:	125		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
accommodates	16.1374	1.767	9.134	0.000	12.674	19.601
beds	2.4949	2.771	0.900	0.368	-2.937	7.927
host_total_listings_count	-1.6558	0.443	-3.739	0.000	-2.524	-0.788
review_scores_value	-14.0645	4.312	-3.262	0.001	-22.519	-5.610
latitude	-79.4248	120.087	-0.661	0.508	-314.859	156.010
longitude	-598.8537	122.951	-4.871	0.000	-839.902	-357.805
number_of_reviews	-0.0344	0.005	-6.562	0.000	-0.045	-0.024
minimum_nights	107.6345	1.950	55.211	0.000	103.812	111.457
maximum_nights	-0.0005	0.001	-0.396	0.692	-0.003	0.002
number_of_stays	-0.0688	0.010	-6.562	0.000	-0.089	-0.048
square_feet	-226.0448	44.232	-5.110	0.000	-312.763	-139.327

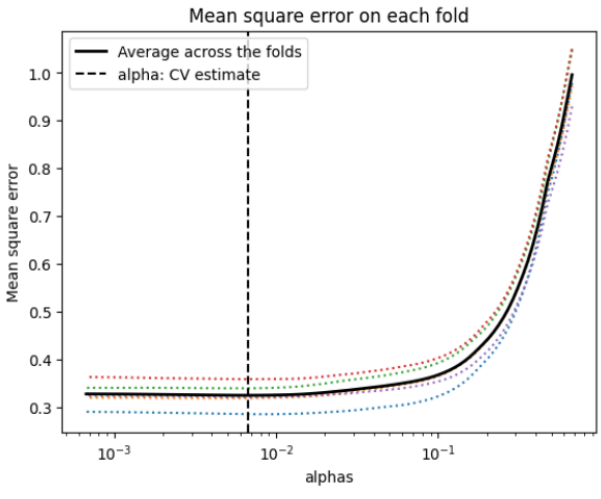
...

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The smallest eigenvalue is 7.08e-23. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

Appendix 2.2 Lasso Model Selection via CV – Mean square error on each fold and the top features chose.



Weight[?]	Feature
+0.525	minimum_nights
+0.181	accommodates
+0.114	bedrooms
+0.038	bathrooms
+0.027	neighbourhood_cleansed_Gaslamp Quarter
+0.025	neighbourhood_cleansed_East Village
+0.022	review_scores_cleanliness
+0.022	neighbourhood_cleansed_La Jolla
+0.021	neighbourhood_cleansed_Mission Bay
+0.020	neighbourhood_cleansed_Marina
+0.016	neighbourhood_cleansed_Core
+0.014	neighbourhood_cleansed_Columbia
+0.013	cancellation_policy_super_strict_60
+0.011	neighbourhood_cleansed_Del Mar Heights
+0.007	neighbourhood_cleansed_Cortez Hill
+0.007	neighbourhood_cleansed_Eastlake Vistas
+0.006	beds
+0.005	cancellation_policy_strict_14_with_grace_period
+0.004	neighbourhood_cleansed_Eastlake Woods
+0.002	neighbourhood_cleansed_Horton Plaza
+0.001	cancellation_policy_super_strict_30
+0.000	neighbourhood_cleansed_Midtown District
+0.000	neighbourhood_cleansed_Little Italy

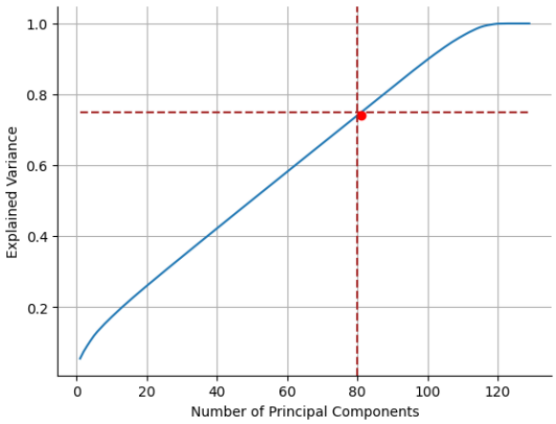
Appendix 2.3 Decision Trees – Features and their importance

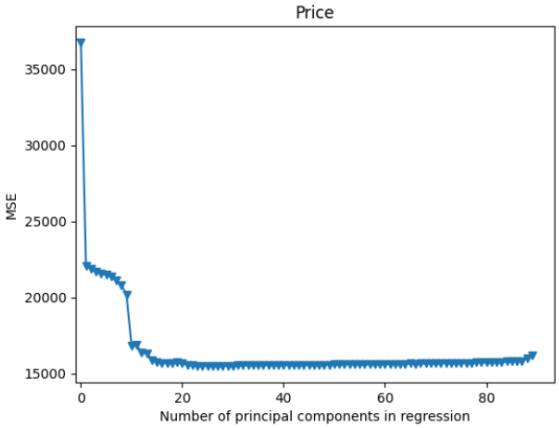
```
Feature Ranking
1. feature minimum_nights (0.476433)
2. feature accommodates (0.099322)
3. feature bedrooms (0.080349)
4. feature longitude (0.064872)
5. feature latitude (0.045475)
6. feature room_type_Private room (0.029992)
7. feature host_total_listings_count (0.020486)
8. feature maximum_nights (0.020223)
9. feature number_of_stays (0.020195)
10. feature number_of_reviews (0.019922)
11. feature guests_included (0.018187)
12. feature bathrooms (0.012022)
13. feature host_is_superhost_t (0.006978)
14. feature beds (0.006275)
15. feature cancellation_policy_moderate (0.005474)
16. feature review_scores_cleanliness (0.004839)
17. feature review_scores_value (0.004556)
18. feature neighbourhood_cleansed_Bay Park (0.004394)
19. feature instant_bookable_t (0.004375)
20. feature cancellation_policy_strict_14_with_grace_period (0.003808)
21. feature neighbourhood_cleansed_Mission Bay (0.003320)
22. feature host_identity_verified_t (0.003231)
...
130. feature neighbourhood_cleansed_East Lake (0.000000)
131. feature neighbourhood_cleansed_Bird Land (0.000000)
132. feature neighbourhood_cleansed_Carmel Mountain (0.000000)
133. feature neighbourhood_cleansed_Egger Highlands (0.000000)
```

Appendix 2.4 Random Forest – Features and their importance

```
Feature Ranking
1. feature minimum_nights (0.469219)
2. feature accommodates (0.095179)
3. feature bedrooms (0.092999)
4. feature longitude (0.069848)
5. feature latitude (0.040566)
6. feature room_type_Private room (0.026895)
7. feature maximum_nights (0.021418)
8. feature number_of_reviews (0.019513)
9. feature number_of_stays (0.019291)
10. feature host_total_listings_count (0.017016)
11. feature bathrooms (0.016180)
12. feature guests_included (0.013658)
13. feature beds (0.013103)
14. feature review_scores_value (0.004934)
15. feature cancellation_policy_strict_14_with_grace_period (0.004779)
16. feature instant_bookable_t (0.004696)
17. feature host_is_superhost_t (0.004586)
18. feature host_identity_verified_t (0.004425)
19. feature review_scores_cleanliness (0.004300)
20. feature neighbourhood_cleansed_Mission Bay (0.003742)
21. feature neighbourhood_cleansed_Bay Park (0.003708)
22. feature cancellation_policy_moderate (0.003053)
23. feature host_response_time_within a few hours (0.002783)
...
130. feature review_scores_communication (0.000000)
131. feature review_scores_location (0.000000)
132. feature neighbourhood_cleansed_Yosemite Dr (0.000000)
133. feature neighbourhood_cleansed_Eastlake Trails (0.000000)
```

Appendix 2.5 PCA – The relationship between the number of principal components, explained variance and MSE.


Explained Variance
Number of Principal Components
1.0
0.8
0.6
0.4
0.2
0
20
40
60
80
100
120


Price
MSE
Number of principal components in regression
35000
30000
25000
20000
15000
0
20
40
60
80