Group 12's Investor Report for AirBnB's in San Diego

Chinmay Gupta, Shruti Sharma,Siddhita Bhagwe, Harsh Sharma, Mansi Kosamkar, Anuj Doshi

Table of Contents

01 Kaggle Overview

1. Cleaning

Handling of missing values. Numeric values were replaced with median of the particular column.

2. Feature Selection

We applied domain knowledge,used feature selection models and visualizations to understand the significant variables and dropped 37 redundant variables.

3. Modeling

We started with basic linear and logistic models and slowly moved onto ensemble methods.

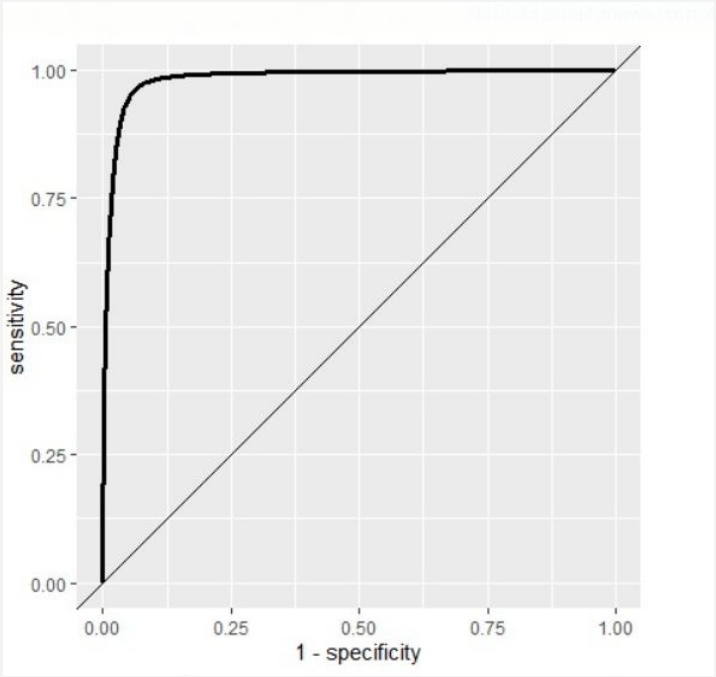
4. Conclusion

Our final model is the XGBoost model with 29 variables which gave us an accuracy of 84% and roc estimate of 0.91

.metric <chr>	.estimator <chr>	.estimate <dbl>
roc_auc	binary	0.913636

1 row

```
Confusion Matrix and Statistics

             high_booking_rate
predictedClass     0     1
             0 19517  2483
             1  2350  5968

               Accuracy : 0.8406
                 95% CI : (0.8364, 0.8447)
    No Information Rate : 0.7213
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6016

 Mcnemar's Test P-Value : 0.0576

            Sensitivity : 0.7062
            Specificity : 0.8925
         Pos Pred Value : 0.7175
         Neg Pred Value : 0.8871
             Prevalence : 0.2787
         Detection Rate : 0.1968
   Detection Prevalence : 0.2744
      Balanced Accuracy : 0.7994

       'Positive' Class : 1
```



02 Business Requirements

Develop a business case for an investor who is interested in acquiring homes in San Diego to put them up as AirBnB rentals.

Project Statement

Requirement 1

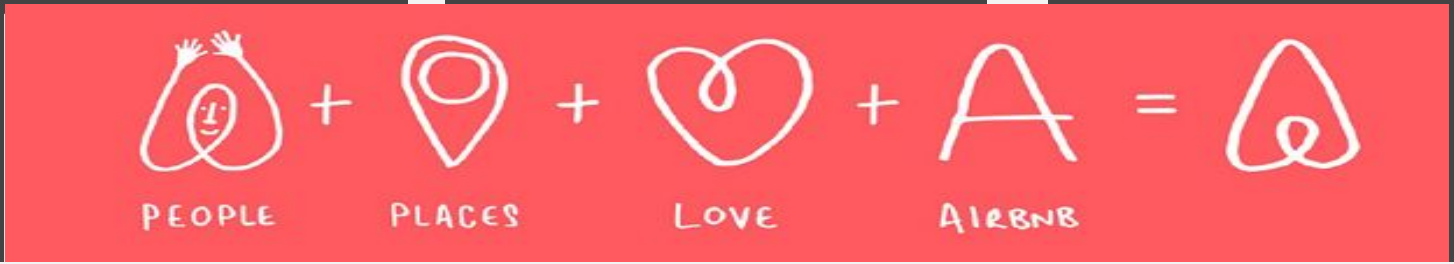
Investor aims to purchase those properties which yield high airbnb traffic.

Requirement 2

Investor needs to strategize how to advertise the property for airbnb rentals based on different variables.

Requirement 3

Investor needs to provide the amenities which result high airbnb traffic.



Market Analysis: San Diego

03

Tourism

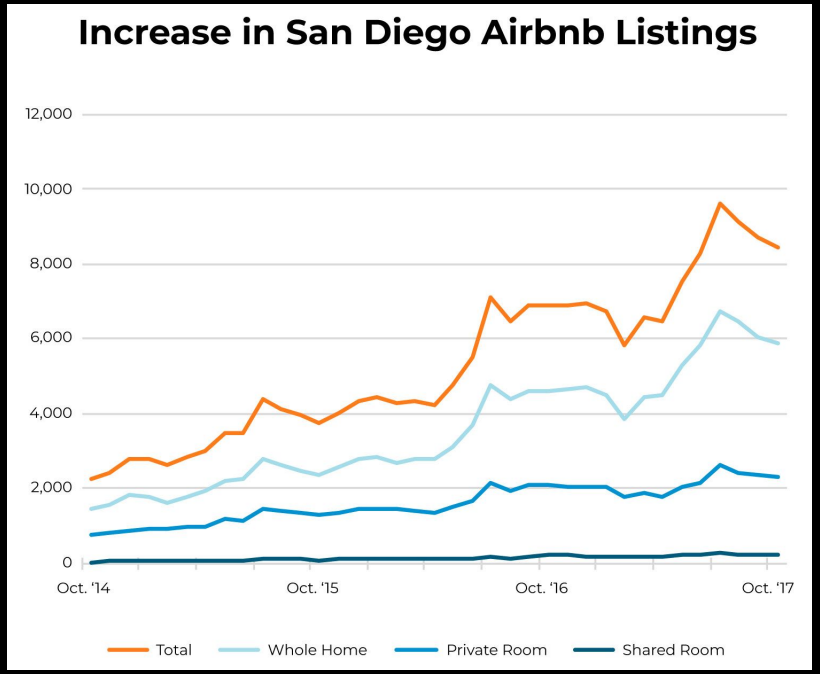
Popular Events

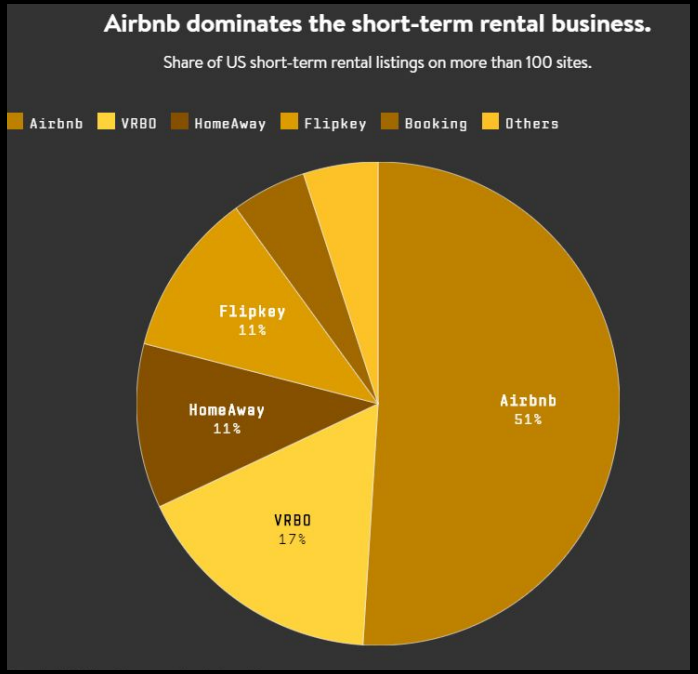
- SAN DIEGO COMIC-CON
- SAN DIEGO COUNTY FAIR
- MIRAMAR AIR SHOW
- KAABOO DEL MAR

Popular Attractions

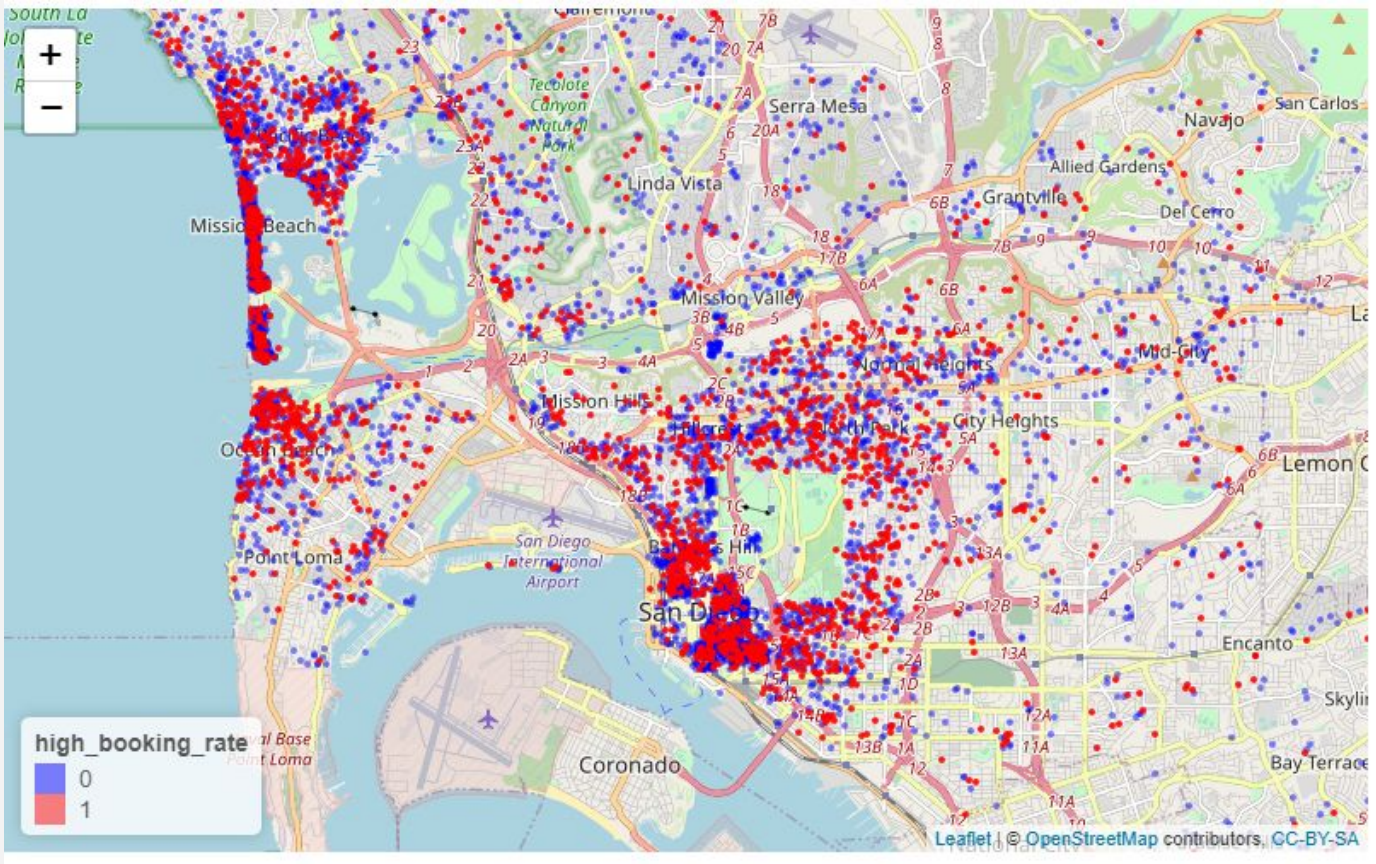
- LEGOLAND
- San Diego Zoo
- Balboa Park
- Sea World

Market Data Analysis

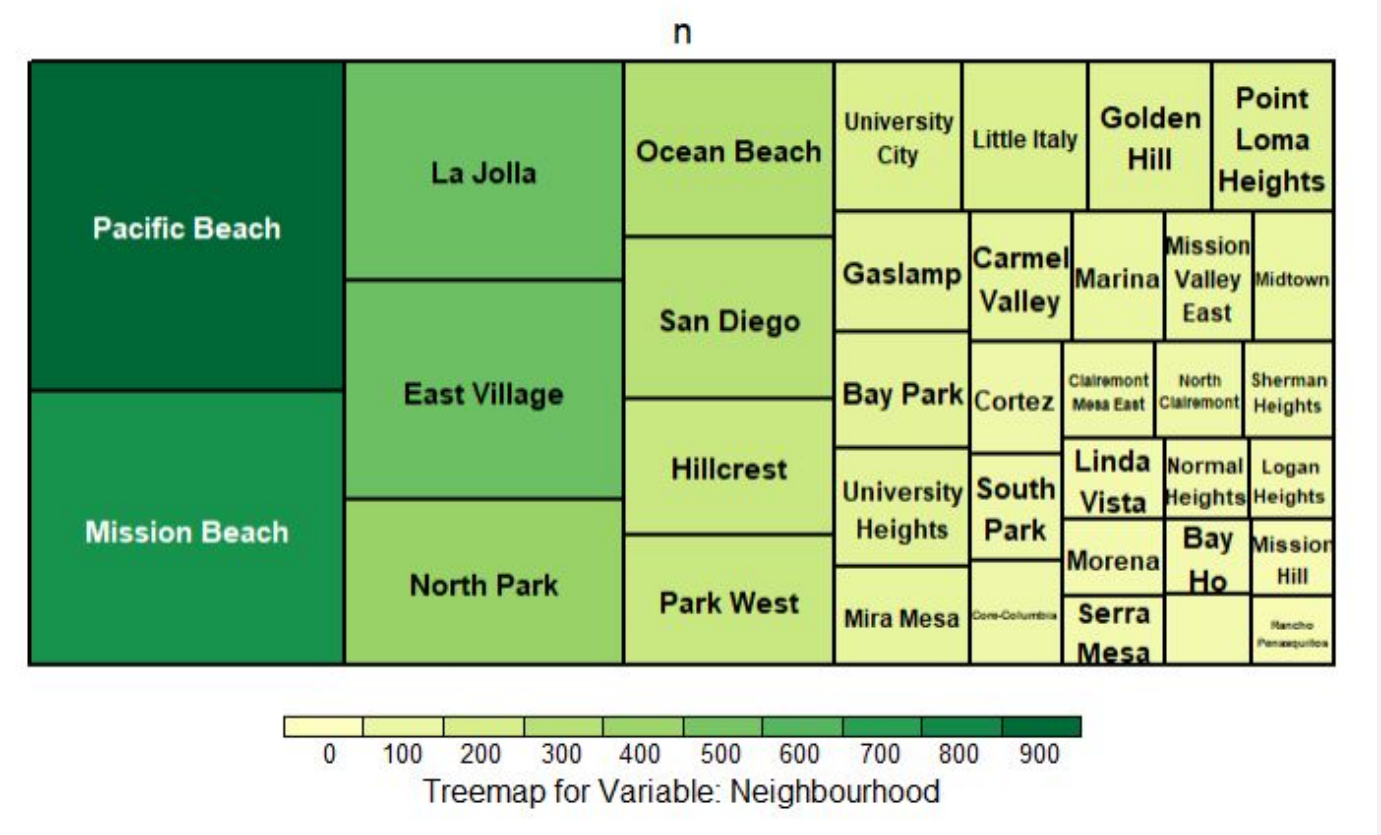




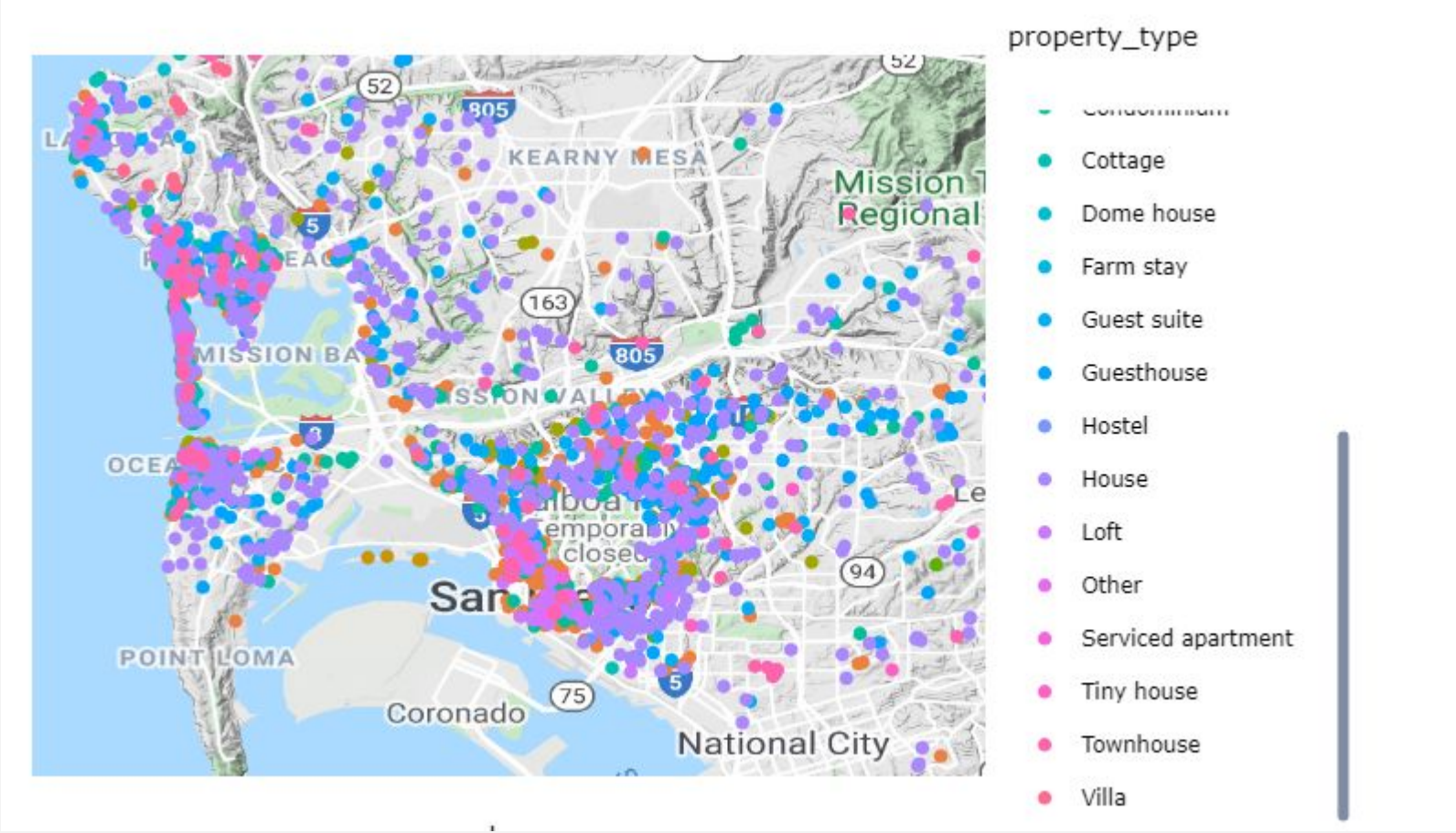
Distribution of properties across San Diego



A Tree Map showing the popular neighbourhoods:

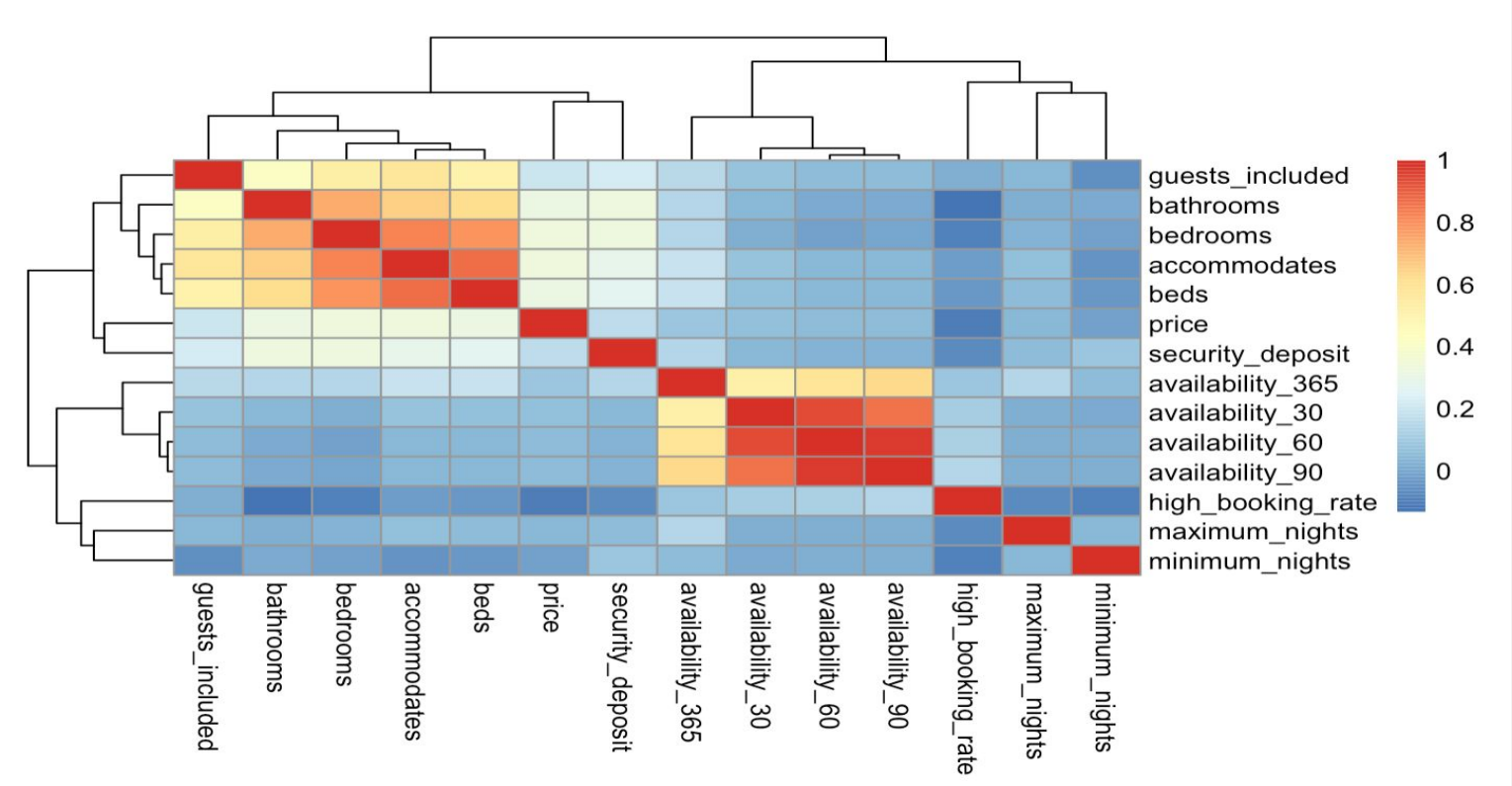


Popular property types:



04 Business Cases

To Focus on Marketing and Advertising an AirBnB property - We establish correlation among variables



Business Case 1

Price vs. High Booking Rate

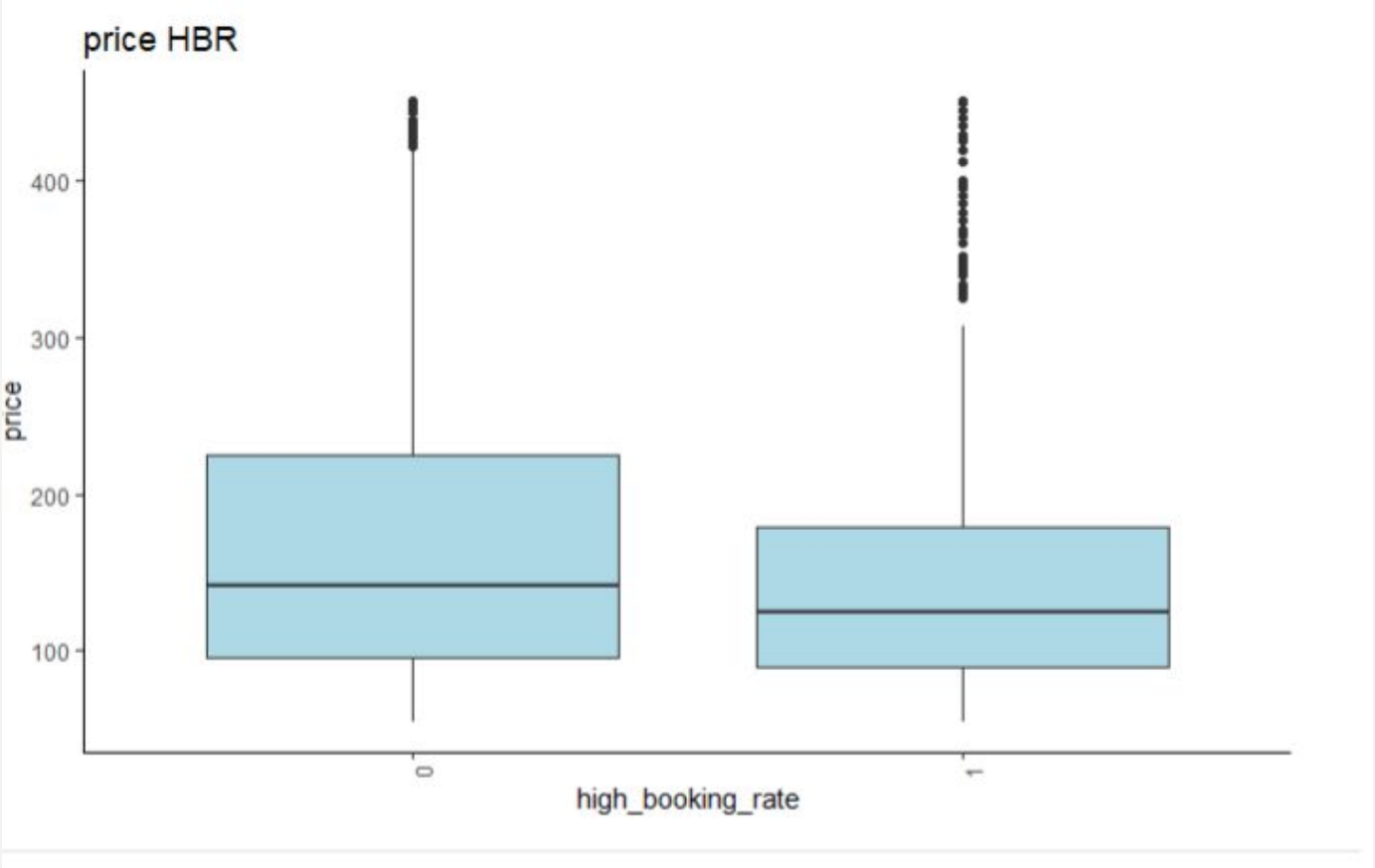
1. Can the price alone determine the High Booking rates?

2. How Much Should you charge for the property?

There does not exist positive correlation between Prices and High Booking Rate. So no direct inferences can be made

	high_booking_rate	price
high_booking_rate	1.0000000	-0.1011064
price	-0.1011064	1.0000000

Boxplot of prices with High booking rate



Conclusion:

1. The price shows no direct effect on high_booking_rate.
2. The properties with higher booking rates have a median price of $120.

Business Case 2

Factors affecting the high booking rate

Is Booking Rate of a property dependent on how “accommodative” an Airbnb property is ?

Variables considered

High_booking_rate,accommodates,availability_30,availability_365,availability_60,availability_90,bathrooms,bedrooms,maximum_nights,minimum_nights,guests_included,beds

Model Applied

Logistic Regression Model

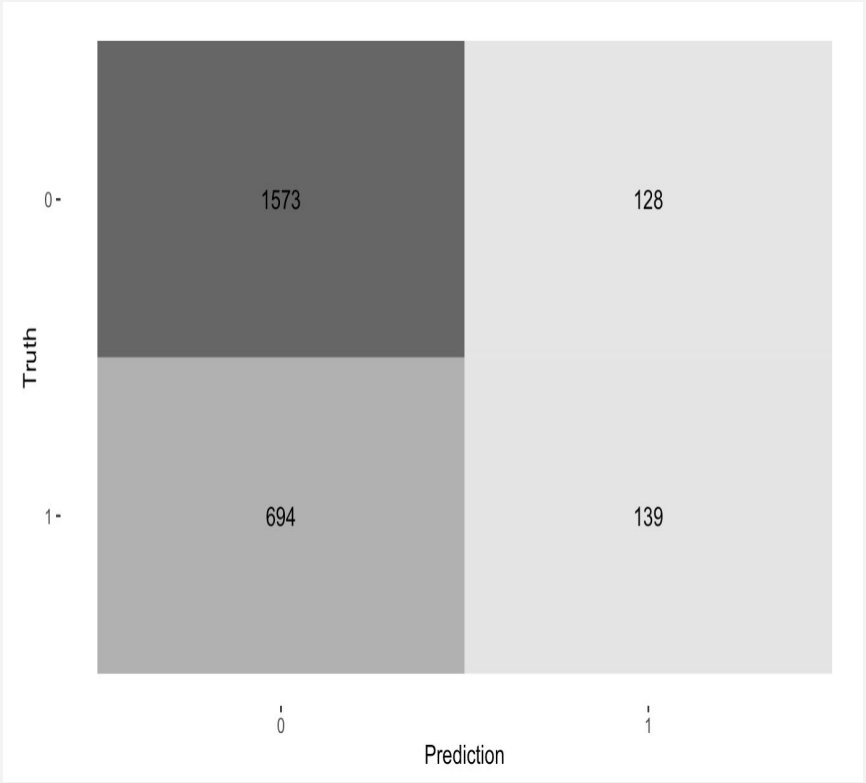
```
Call:
glm(formula = high_booking_rate ~ ., family = "binomial", data = dfcTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4886  -0.8532  -0.6254   1.1839   3.0063

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -4.521e-01  9.449e-02  -4.784 1.71e-06 ***
accommodates     7.379e-02  2.518e-02   2.931 0.003378 **
availability_30  3.212e-02  9.123e-03   3.521 0.000431 ***
availability_365 7.256e-04  3.320e-04   2.186 0.028830 *
availability_60 -6.723e-02  9.400e-03  -7.153 8.52e-13 ***
availability_90  4.182e-02  4.603e-03   9.085  < 2e-16 ***
bathrooms       -6.147e-01  7.434e-02  -8.268  < 2e-16 ***
bedrooms        -1.700e-01  5.764e-02  -2.949 0.003189 **
maximum_nights  -3.586e-04  6.202e-05  -5.781 7.40e-09 ***
minimum_nights  -7.546e-02  9.032e-03  -8.354  < 2e-16 ***
guests_included  5.316e-02  1.801e-02   2.951 0.003168 **
beds             3.632e-02  3.711e-02   0.979 0.327802
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6288.9  on 5293  degrees of freedom
Residual deviance: 5736.9  on 5282  degrees of freedom
AIC: 5760.9

Number of Fisher Scoring iterations: 6
```

p-values indicates that statistically important variables are

accommodates,availability_30, availability_365,availability_60, availability_90,bathrooms, bedrooms,maximum_nights, minimum_nights, guests_included

Accuracy of the model

```
Confusion Matrix and Statistics

                high_booking_rate
predictedClass     0    1
             0  1573  694
             1   128  139

               Accuracy : 0.6756
                 95% CI : (0.657, 0.6938)
    No Information Rate : 0.6713
    P-Value [Acc > NIR] : 0.3292
```



Confusion matrix is used to find the Accuracy as the dataset was highly skewed.

Conclusion:

For an Airbnb property to have higher booking rates, it must be flexible with respect to its booking duration and be able to provide accommodation to the guests.

Business Case 2

Factors affecting the high booking rate

Is the high booking rate affected by various charges included with price?

Variables considered

- Cleaning_fee
- Extra_people
- Price
- Security_deposit

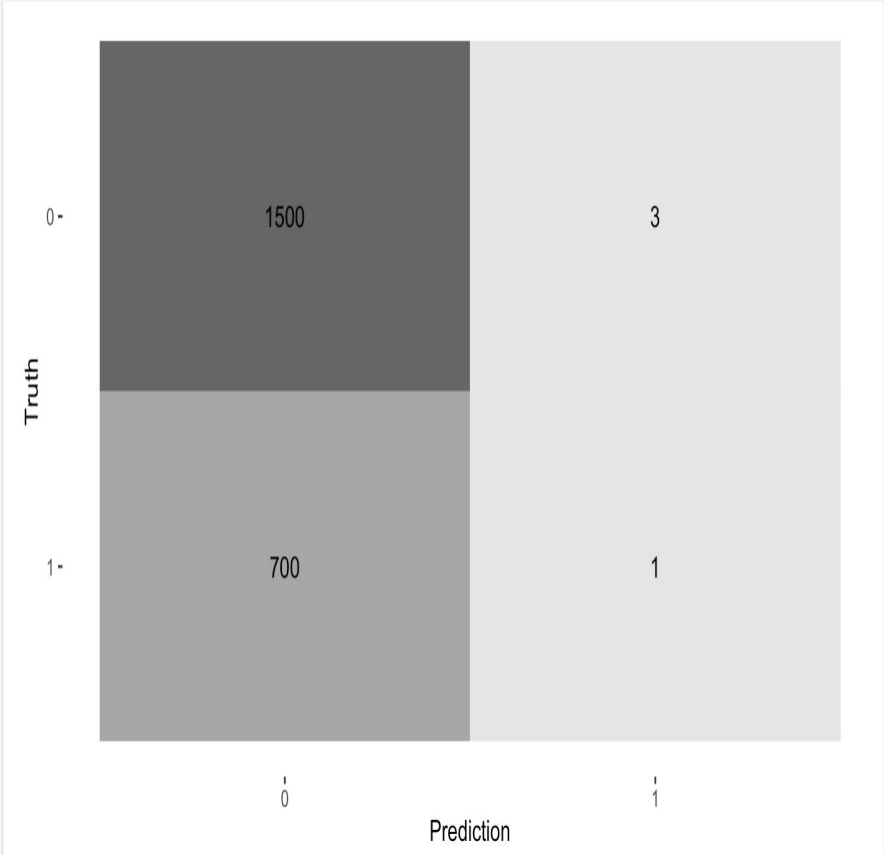
Model Applied

Logistic Regression Model

```
Call:
glm(formula = high_booking_rate ~ ., family = "binomial", data = dfcTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4240  -0.8881  -0.7884   1.4429   3.7715

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -5.492e-01  4.877e-02 -11.261  < 2e-16 ***
cleaning_fee     -8.307e-04  4.865e-04  -1.707 0.087733 .
extra_people      4.464e-03  1.190e-03   3.752 0.000176 ***
price            -1.630e-03  2.554e-04  -6.384 1.72e-10 ***
security_deposit -5.069e-05  8.361e-05  -0.606 0.544305
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6398.2  on 5293  degrees of freedom
Residual deviance: 6240.1  on 5289  degrees of freedom
AIC: 6250.1

Number of Fisher Scoring iterations: 6
```

Only extra-people and price are statistically important variables

```
Confusion Matrix and Statistics

                high_booking_rate
predictedClass    0    1
             0 1500  700
             1    3    1

               Accuracy : 0.681
                 95% CI : (0.6611, 0.7005)
    No Information Rate : 0.6819
    P-Value [Acc > NIR] : 0.5466
```



Conclusion

- If a property provides for extra_people despite charging for the same - its booking rate improves.
- Cleaning fee and security deposits do not have any impact on the higher booking rates.
- If the price of a property is high, its booking rate decreases.

Business Case 2

Factors affecting the high booking rate

Does being a superhost affect booking rate?

Variable Selection

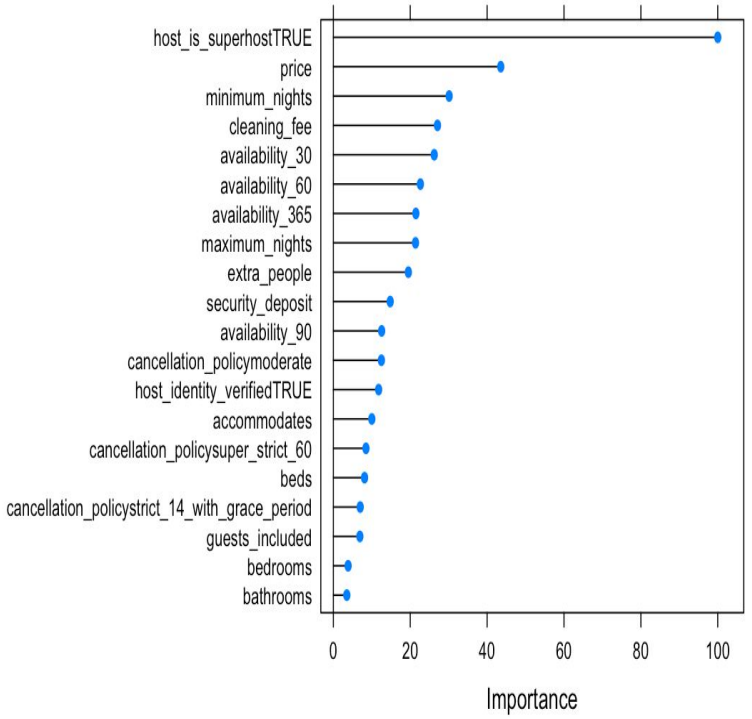
- Factors that contribute towards the location setting, generalized pricing strategy and in house services provided, were considered while structuring this business case
- Initially, a logistic model was run to understand the importance and significance of each variable
- Accuracy achieved for this model was 77.49%

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2849  -0.7233  -0.4239   0.7975   4.2468

Coefficients:
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                -2.334e+00  1.328e-01 -17.581  < 2e-16 ***
host_is_superhostTRUE                       1.438e+00  6.872e-02  20.926  < 2e-16 ***
price                                      -1.774e-03  3.194e-04  -5.555 2.77e-08 ***
minimum_nights                             -5.928e-02  7.562e-03  -7.839 4.54e-15 ***
cleaning_fee                               -3.518e-03  7.180e-04  -4.900 9.57e-07 ***
availability_365                            1.786e-03  3.567e-04   5.006 5.55e-07 ***
availability_60                            -4.997e-02  9.835e-03  -5.081 3.76e-07 ***
availability_30                             3.538e-02  9.378e-03   3.773 0.000162 ***
maximum_nights                             -2.563e-04  6.321e-05  -4.055 5.02e-05 ***
extra_people                               -3.194e-03  1.514e-03  -2.110 0.034901 *
security_deposit                            8.439e-05  9.221e-05   0.915 0.360116
availability_90                             2.800e-02  4.778e-03   5.860 4.62e-09 ***
cancellation_policymoderate                 1.114e+00  1.016e-01  10.970  < 2e-16 ***
host_identity_verifiedTRUE                  5.520e-01  6.833e-02   8.078 6.56e-16 ***
accommodates                                1.299e-01  2.666e-02   4.873 1.10e-06 ***
cancellation_policysuper_strict_60         -1.051e+00  3.066e-01  -3.429 0.000607 ***
beds                                        6.893e-05  3.709e-02   0.002 0.998517
cancellation_policystrict_14_with_grace_period 8.345e-01  9.989e-02   8.354  < 2e-16 ***
guests_included                             9.308e-02  2.111e-02   4.409 1.04e-05 ***
bedrooms                                   -7.141e-02  6.063e-02  -1.178 0.238868
bathrooms                                  -3.396e-01  7.541e-02  -4.504 6.68e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6870.6  on 5700  degrees of freedom
Residual deviance: 5340.6  on 5680  degrees of freedom
AIC: 5382.6

Number of Fisher Scoring iterations: 6
```

XG Boost as a reference model for variable selection

The Xgboost model showed an accuracy of 79.38% and Sensitivity of 52.32%. The sensitivity is more important in this particular case as we don't want to overestimate the High_booking_rate factor.

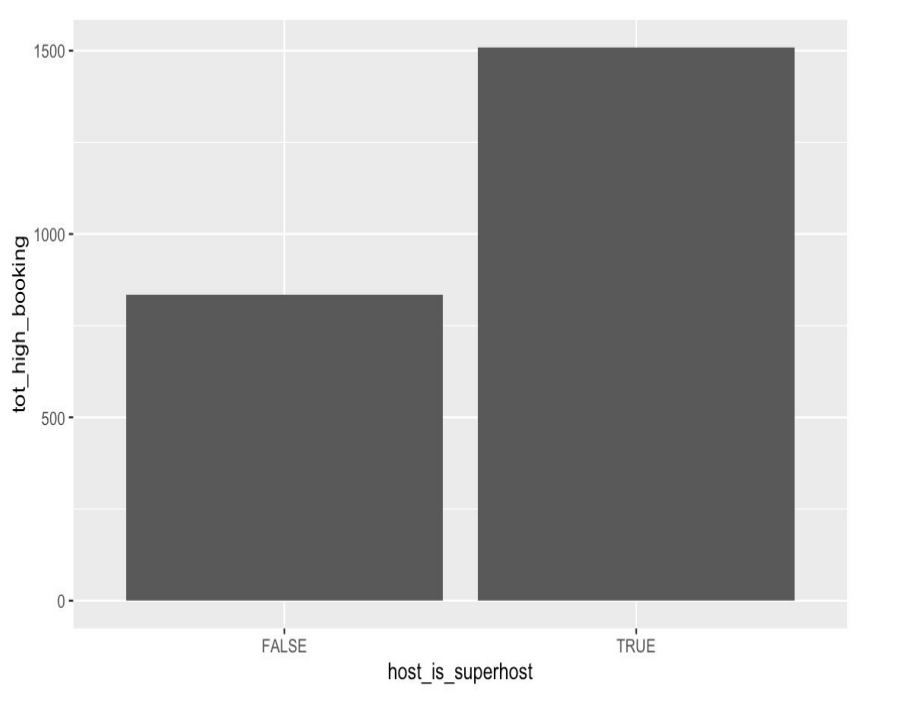
```
Confusion Matrix and Statistics

                high_booking_rate
predictedClass    0    1
             0 1811  391
             1  191  429

               Accuracy : 0.7938
                 95% CI : (0.7784, 0.8086)
    No Information Rate : 0.7094
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.461

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5232
            Specificity : 0.9046
         Pos Pred Value : 0.6919
         Neg Pred Value : 0.8224
             Prevalence : 0.2906
         Detection Rate : 0.1520
   Detection Prevalence : 0.2197
      Balanced Accuracy : 0.7139

       'Positive' Class : 1
```



How does a host being a superhost affect booking rate?

The most important variable from the XGBoost model is the host_is_superhost = True.

From the bar graph, we can see that the booking rate for host_is_superhost = True is about 55% more than where host is not a superhost



Conclusion

- Variables, excluding security deposits, beds and bedrooms are statistically important.

- While the host holding a superhost status improves the booking rate the most.

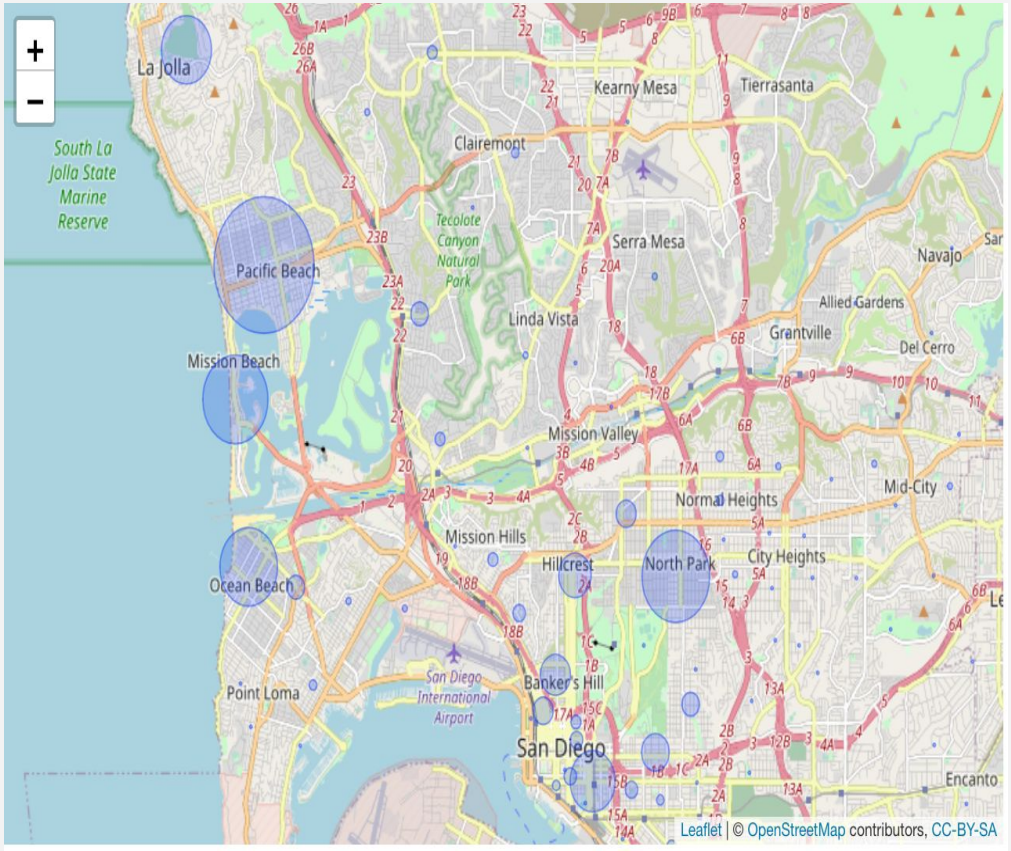
Business Case 2

Factors affecting the high booking rate

Which neighbourhoods result in a higher booking rate?

The regions and the number of high booking rates are displayed on the map. The size of the circle also correlates with the number of high booking rates. These regions can be used as a factor to decide on initial acquisition as well as pricing.

Due to various events like the Comic Con, Surfing tournaments, and La Jolla festival at La Jolla and general attraction for beaches, we can safely say that these regions have a higher booking rate.



Business Case 3

Amenities vs. High Booking Rate

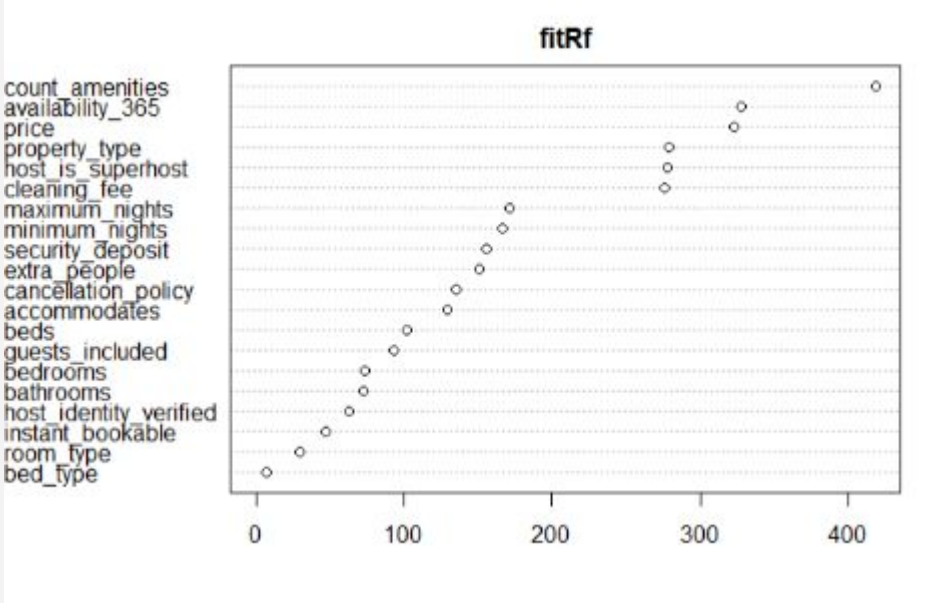
How do the provided amenities affect the booking rate of an AirBnB Property?

Which are the significant variables according to random forest model?

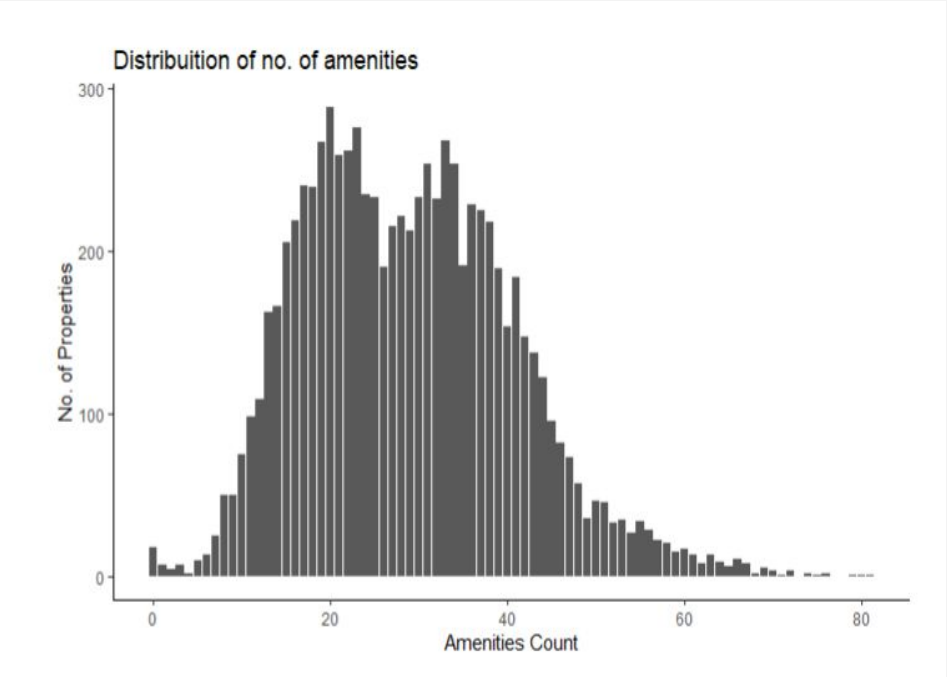
A new column, **“count_amenities”** was created which counted the number of amenities provided at each AirBnB property in San Diego.

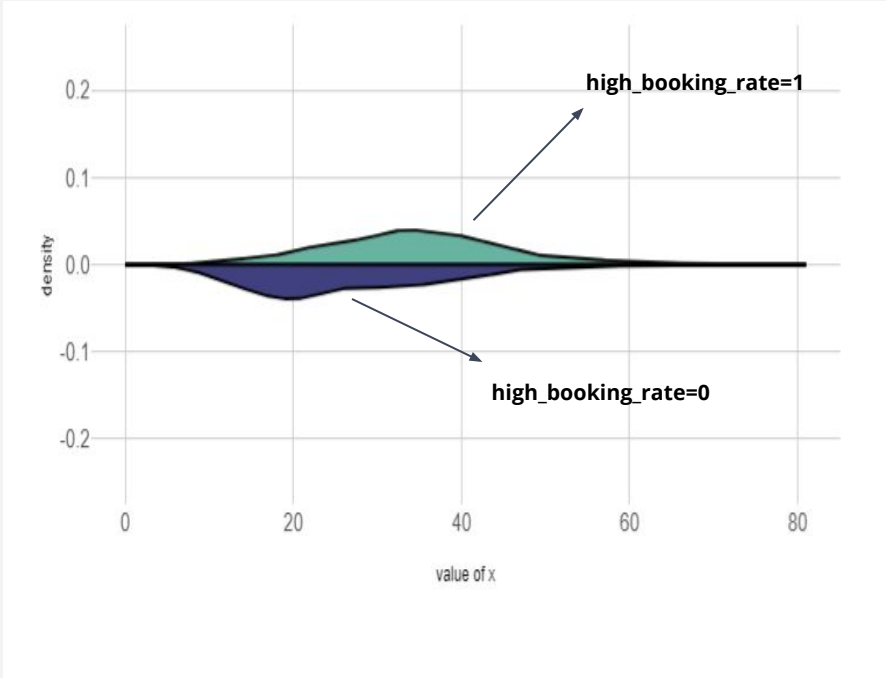
According to variable selection from the random forest model, the newly created variable is the most significant variable.

The other significant variables have already been discussed in the previous slides.



Distribution graphs of count_amenities





Word Cloud of Amenities where high_booking_rate=1

Word Cloud of Amenities where high_booking_rate=0

Conclusion

1. The number of provided amenities affects the booking rate.

2. The most frequently provided amenities are Free Parking, Smoke Detectors, WiFi, Laptop friendly workspace.

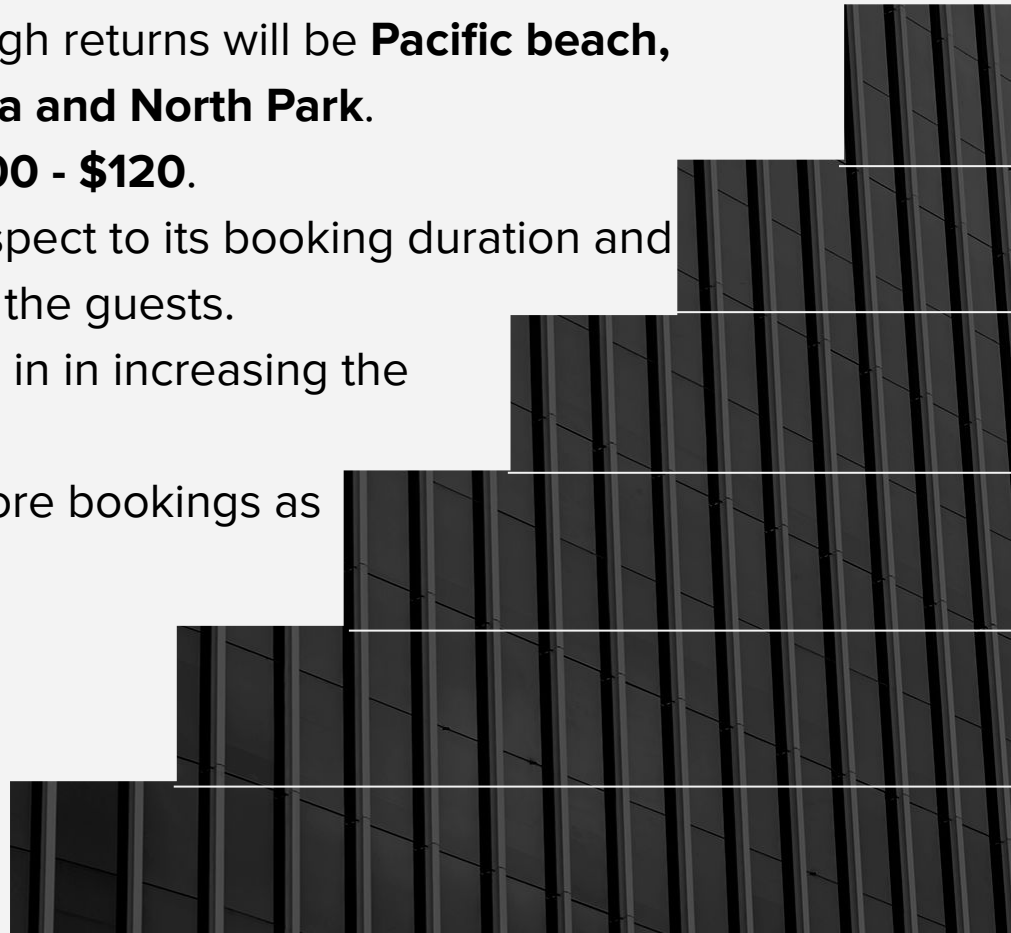
3. More or less, the provided amenities are same for all properties and thus does not affect the booking rate.

Final Summary

05

For an investor considering to invest in Airbnb's in the San Diego Market:

1. Neighbourhoods which would yield high returns will be **Pacific beach, Mission Beach, Ocean Beach, La Jolla and North Park**.
2. An appropriate price range will be **$100 - $120**.
3. The property must be **flexible** with respect to its booking duration and be able to provide accommodation to the guests.
4. Achieving a **superhost** status will help in in increasing the booking rates.
5. A **house** or an **apartment** will have more bookings as compared to other property types.
6. Amenities which matter the most are **Free Parking, Smoke detectors, Wifi, and Laptop friendly workspace.**
7. Providing **more number of amenities** will affect the booking rate.



Any Questions?