Pricing - San Francisco

Gibran Cornejo

Problem Introduction

Pricing

Pricing is an important concern for customers booking travel online. Also, as a host, having the correct price for the house will have a direct impact on the level of reserves, as well as on profits.
Investors are always trying to maximize their rental income so they want to understand how to get a better ROI. They would also want to understand the current market trends so that they can make better decisions. Also they would want to know what makes a property popular and could work on helping their listings.

Objective

As an objective we decided to work on a model that can predict the price that will surely be very useful for current hosts as well as for those who are considering putting their home up for temporary rent. Also we will try to figure out some of the key factors that effect the price of a property.

Dataset Variables

Property type

Apt / House / Condo…

Room type

Entire Apt/ Room

Accommodates

1 / 2 / 3 / 4 / …

Bathrooms

1 / 2 / 3 / 4 …

Bedrooms

1 / 2 / 3 / 4 …

Availability

1 / 2 / 3 / 4 / 5 …

Location

Neighborhood

Amenities

Apt / House / Condo…

Score Review

0 to 100

Beds

Matress/airbed/sofa

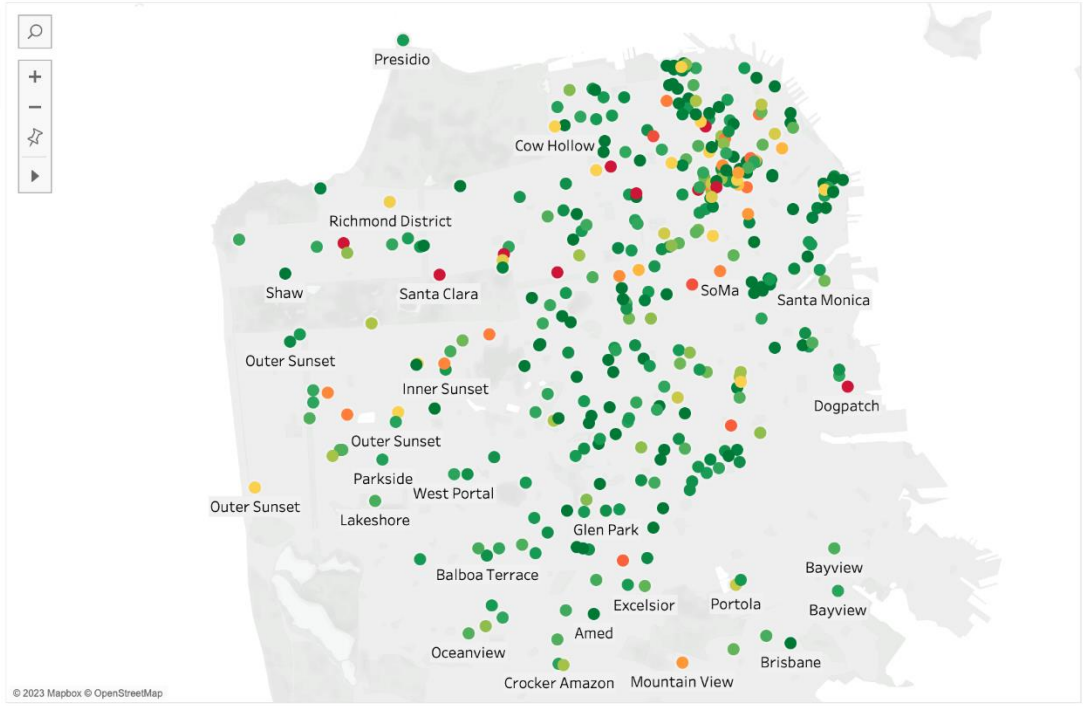
Price

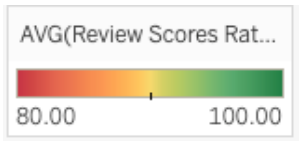
Target Variable

Exploratory Analysis

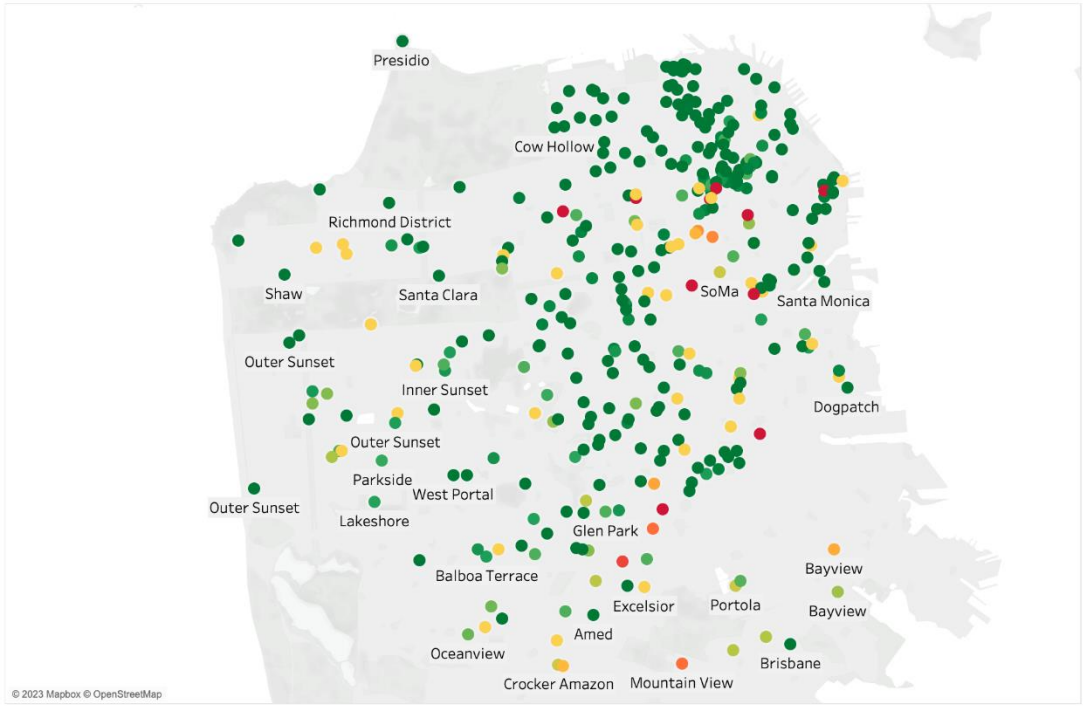
Review Scores Vs Review Location

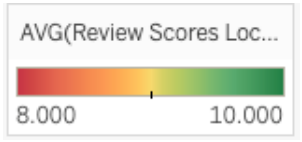
SF Scores Rating





SF Review scores location



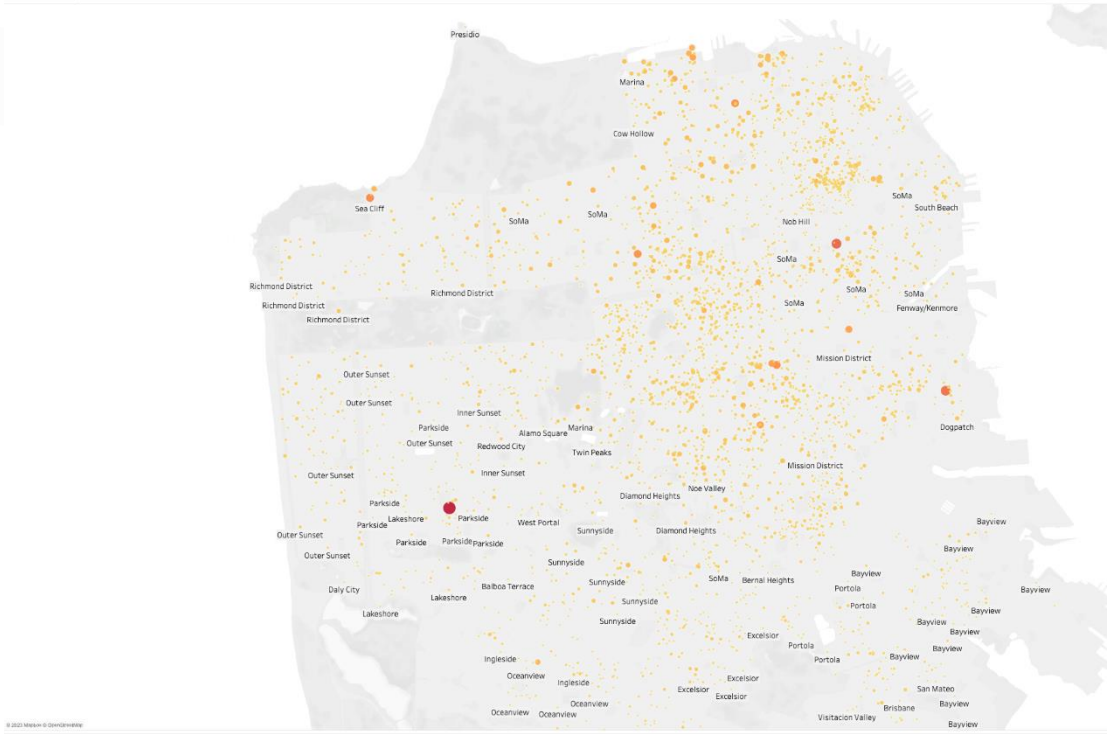


Exploratory Analysis

Pricing Map Vs Occupancy rate

*Occupancy Rate = 1-(Yearly availability / 365)

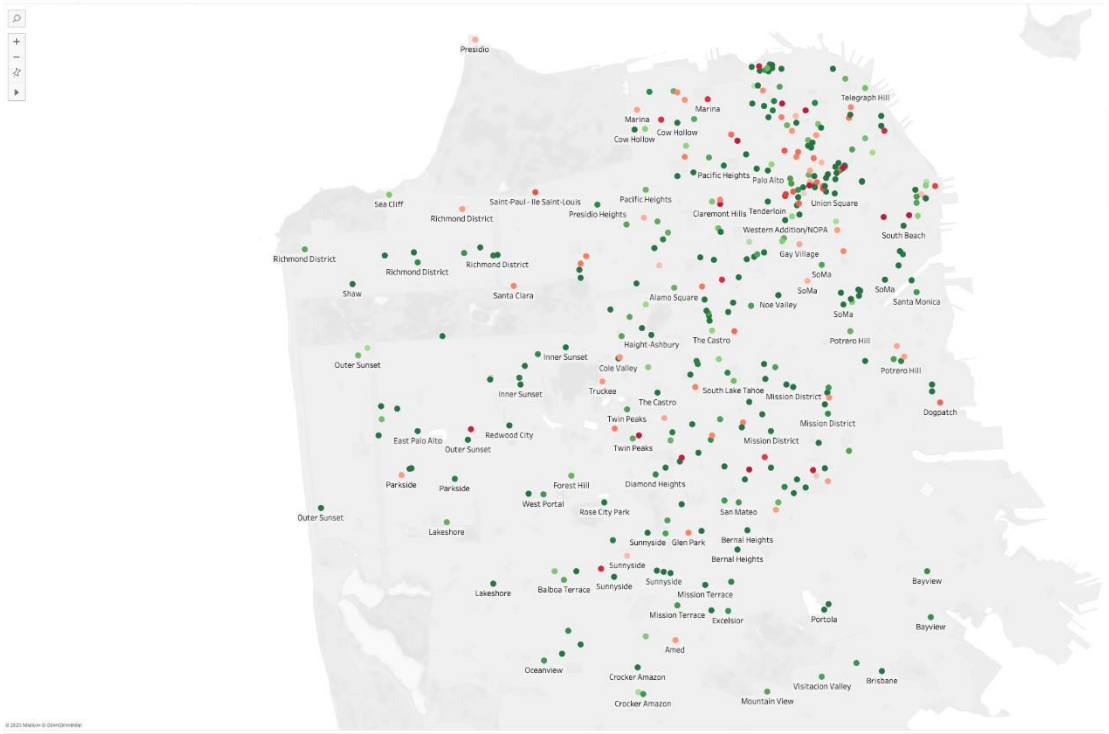
SF Price Map



AVG(Price)

0 8,000

SF Occupancy Rate 0 to 70%

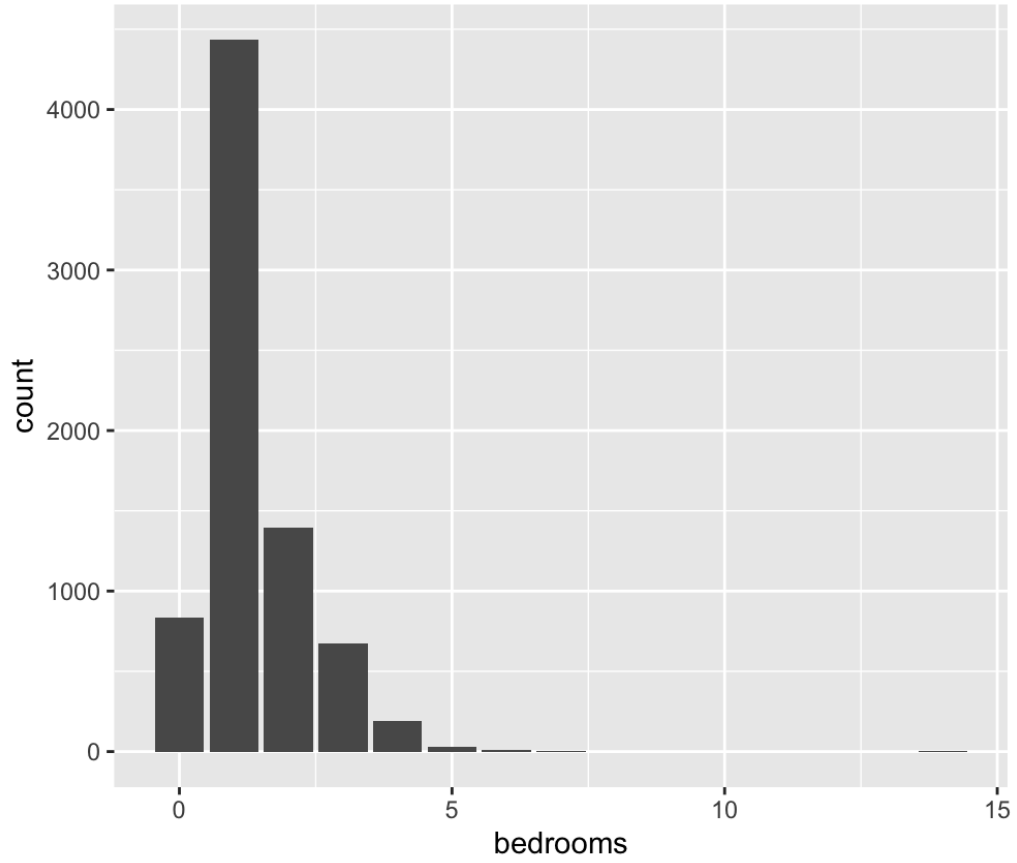


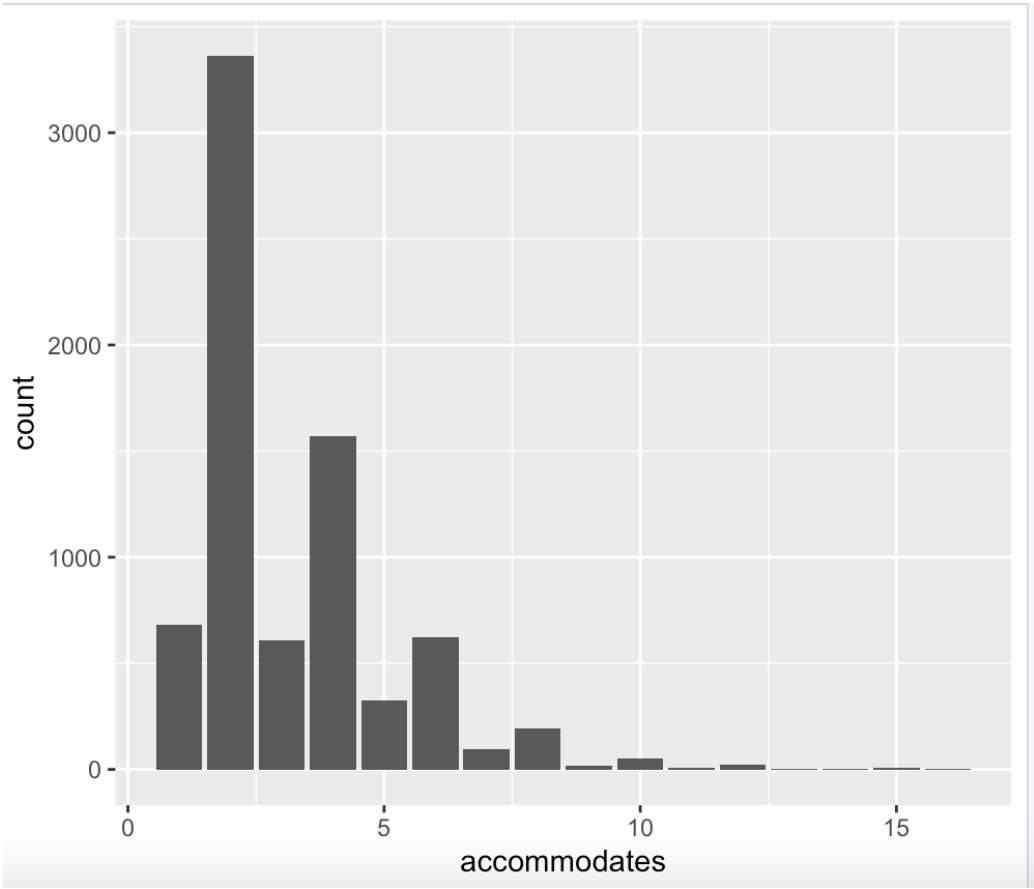
AVG(Occupacy Rate)

0.0000 0.7000

Exploratory Analysis

Bedrooms Vs Accommodates

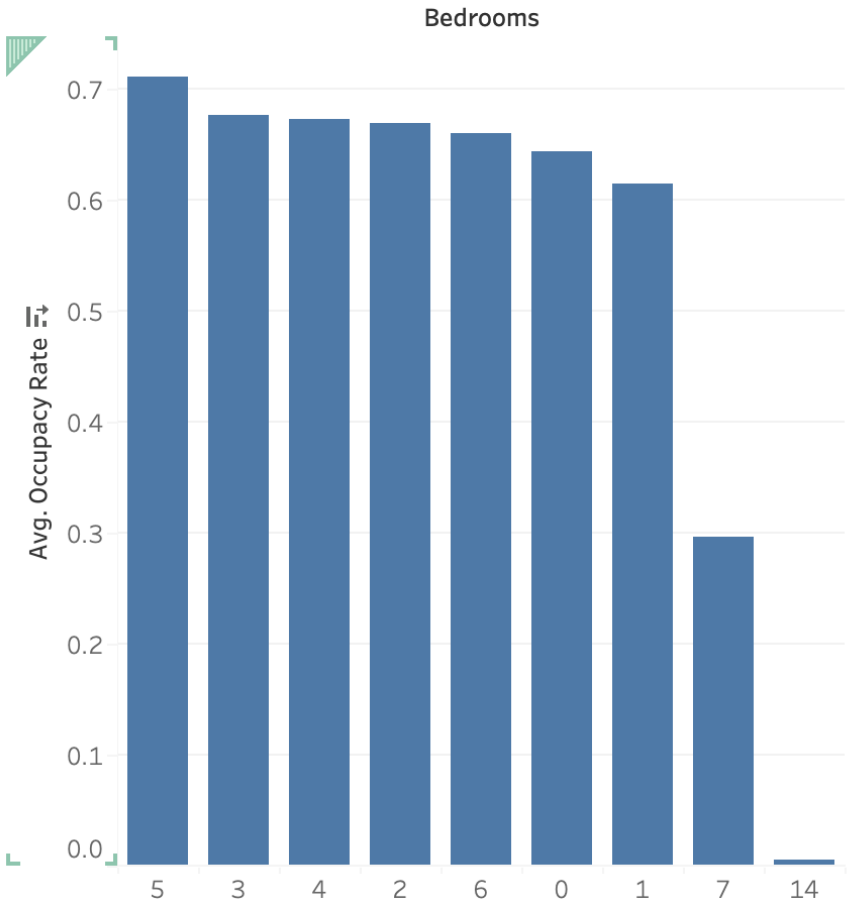




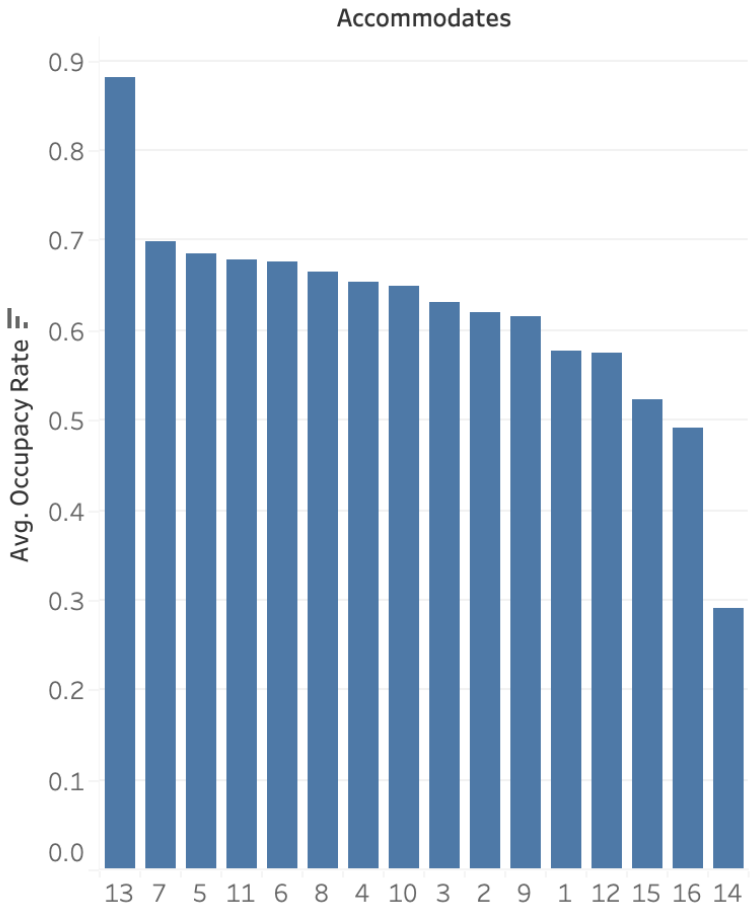
Exploratory Analysis

Bedrooms & accommodates Vs Occupancy rate

rooms vs occ rate



Accomodates - occupancy rate



Assumptions from exploratory analysis

Most of Airbnb offer are focused are 1 bedroom ideal for couples or single travelers. This can be confirmed with the accommodates, plot, most popular 2 persons

Most popular and locations with best review are the neighborhoods that are close to the Bay Area and Mission District

Even though, houses with the highest occupancy rate are the ones with 5 rooms and this is confirmed with the 13 persons accommodates. So people are demanding more space.

Data Split

65%

Training data

35%

Test Data

Models Used

Linear Regression

We performed linear regression on our dataset and checked for the R-squared values.

As the values were not satisfactory hence we decided to go with other models.

```
> summary(airbnb_model)

Call:
lm(formula = price ~ neighbourhood_encoded + accommodates + bathrooms +
    bedrooms + beds + amenities_count + guests_included + availability_365 +
    number_of_reviews + cancellation_policy_encoded, data = airbnb)

Residuals:
    Min      1Q  Median      3Q     Max
-1153.8   -76.2   -28.0    30.3  7879.1

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -32.618926  11.322395  -2.881 0.003976 **
neighbourhood_encoded      0.497073   0.165429   3.005 0.002667 **
accommodates              39.793864   2.815958  14.132  < 2e-16 ***
bathrooms                 12.206115   3.602131   3.389 0.000706 ***
bedrooms                  55.005233   4.681099  11.750  < 2e-16 ***
beds                     -22.729167   4.348796  -5.227 1.77e-07 ***
amenities_count            1.478085   0.246274   6.002 2.04e-09 ***
guests_included            7.149994   2.378445   3.006 0.002654 **
availability_365          -0.003338   0.021129  -0.158 0.874471
number_of_reviews         -0.294006   0.036336  -8.091 6.83e-16 ***
cancellation_policy_encoded  5.702399   2.143469   2.660 0.007822 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 227.8 on 7538 degrees of freedom
  (26 observations deleted due to missingness)
Multiple R-squared:  0.2129,    Adjusted R-squared:  0.2118
F-statistic: 203.8 on 10 and 7538 DF,  p-value: < 2.2e-16
```

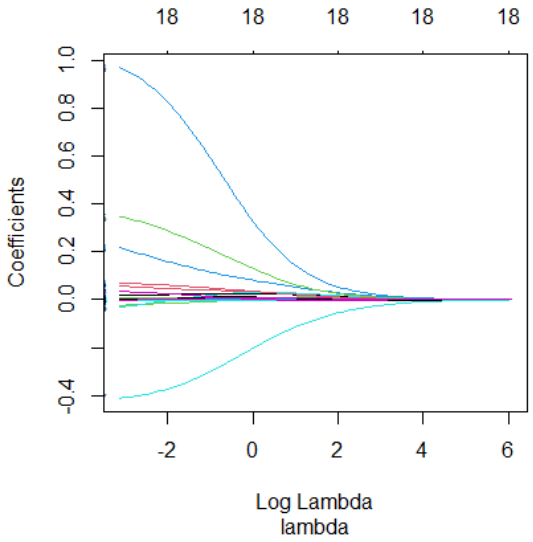
Models Used

Ridge Regression

Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization.

Why did we choose Ridge Regression?

We are using ridge regression so that the coefficients are regularized or shrunk. This helps in pushing the estimated coefficients towards 0, so that they work well for new data.



We can see form the graph that as the value of lambda increases, the value of the coefficients decreases.

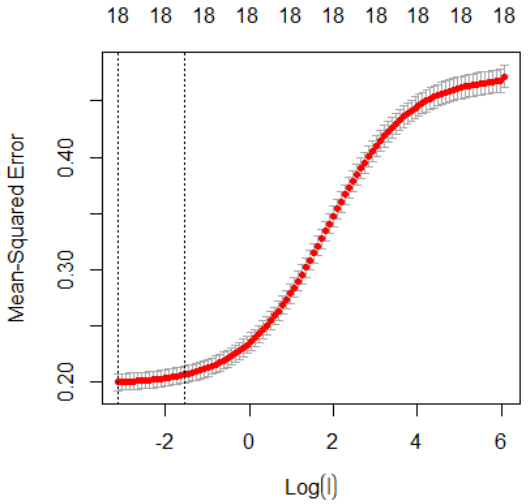
If the lambda value is too small, the model may overfit the data and conversely, if the vale is too high the model may underfit the data.

Model Results

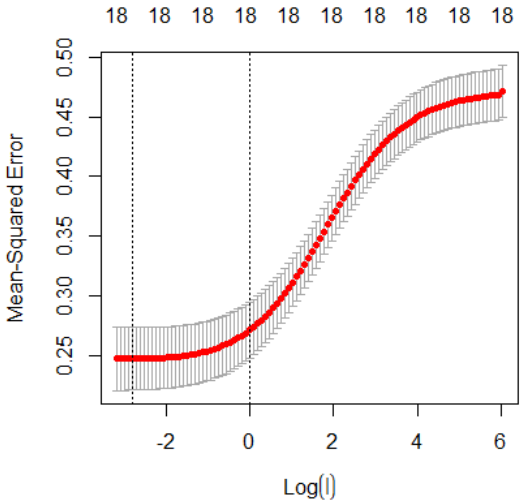
Our model accuracy was about 59%

Below we can see the difference between the Cross Validation plot graph of the training data vs testing data.

The minimum MSE is achieved when Lambda is 0.04 on the testing data, the minimum MSE is achieved when Lambda is 0.06 on the testing data. So we can see that the model performed very similar with the testing data when compared with the training sample.



Training Data



Testing Data

Insights

Our recommendations for new users is that they should invest in properties that has more number of bedrooms and accommodates.

On addition they should consider to offer more amenities as higher count in amenities resulted in more occupancy rate.

Apart from this it is also evident that more lenient cancellation policy results in lesser availability.

Conclusion

Below are some of the key factors that investors should take into consideration before investing

Accommodates

Bedrooms

Amenities

Cancellation Policy

Thank You!