AirBnB: San Francisco

Andrew Sportsman

Abstract

Looking at AirBnB rentals in San Francisco, what factors contribute the most to how much a place costs? Neighborhood? Bedrooms?

Using a decision tree and correlation coefficients, I've found that price doesn't heavily relate to many features and proves to be a difficult measure to predict.

Motivation

The housing market in San Francisco is notoriously expensive, but the city is still a tourist attraction for people all over the world. Guests and owners alike would be interested in what makes certain AirBnB’s more expensive than others.

Dataset

My dataset contains listings on AirBnB for different properties in San Francisco

Source: https://www.kaggle.com/jeploretizo/san-francisco-airbnb-listings

Variables such as: Bedrooms, Bathrooms, Several Rating/Reviews, Property type, Neighborhood, amenities, etc.

Response Variable: Price (per night)

Data Cleaning

The field used as my response, price, was a string and had ‘$’ as the first character of each entry

- Using list comprehensions took care of this cleaning quickly

```
listings['price'] = [x[1:] for x in listings.price]
```

```
listings['price'] = [x.strip().replace(',','') for x in listings.price]
```

Observations with a price over $350 were identified as outliers and removed from the analysis

Data Preparation

Fields such as Neighborhood and Property_type needed conversion to dummy (indicator) variables

Dummy Variable Creation Snapshot

```
neighborhood_dummy = pd.get_dummies(listings['neighbourhood'])
neighborhood_dummy.head()
```

	Alamo Square	Balboa Terrace	Bayview	Bernal Heights	Chinatown	Civic Center	Cole Valley	Cow Hollow	Crocker Amazon	Daly City
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0	0
3	0	0	0	0	0	0	1	0	0	0
4	0	0	0	0	0	0	1	0	0	0

Research Question

Can limited features such as neighborhood, bedrooms, bathrooms, and guest reviews successfully predict the price of AirBnB rentals in San Francisco?

Methods

Correlation Coefficients were used to investigate individual relationships and initial exploration

A regression decision tree was used to predict the AirBnB price

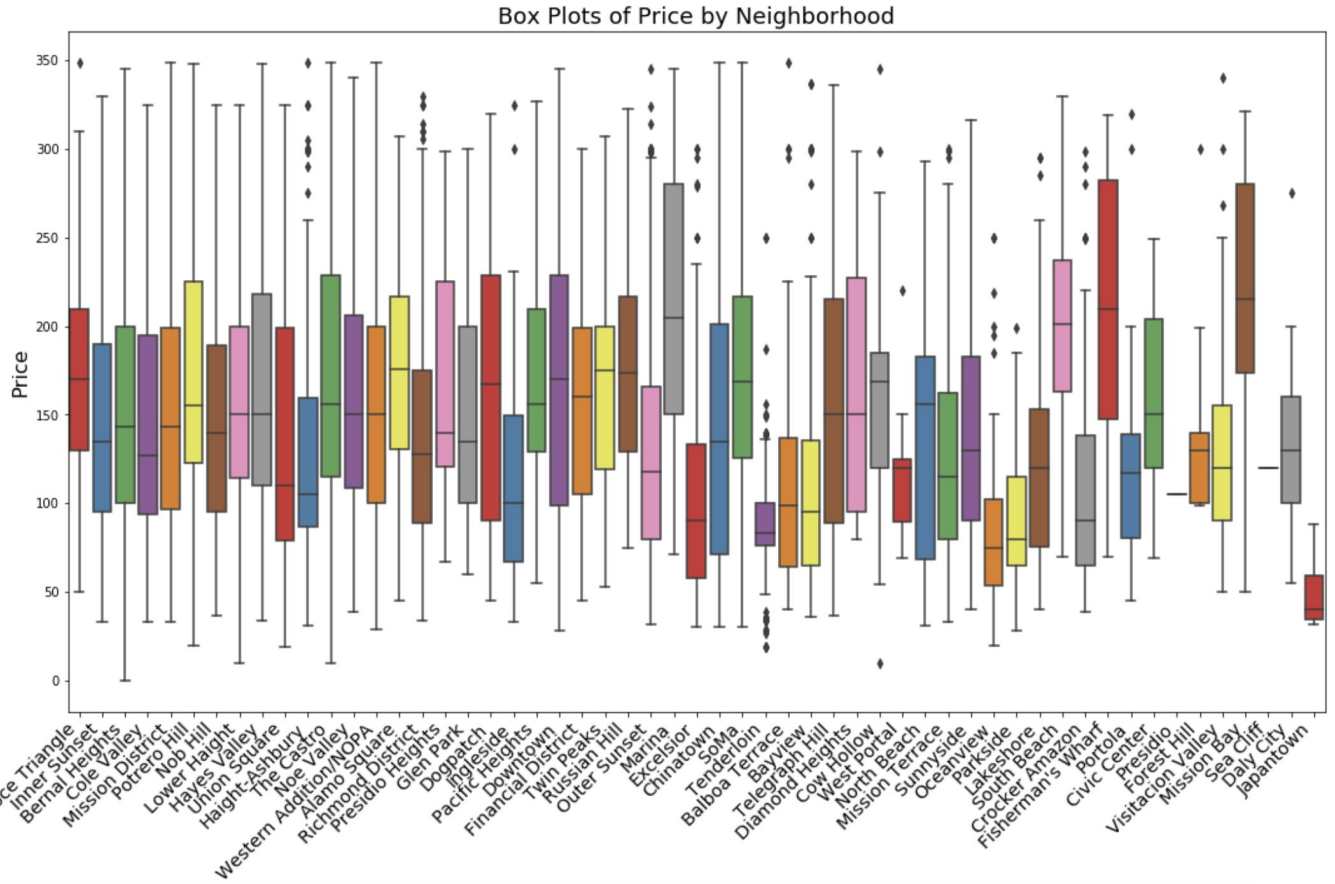
- Bedrooms, Bathrooms, Neighborhood, Property Type, and several review scores were all used to predict Price
- Note regression for continuous response not classification

Findings

Location, Location, Location?

Initially, I believed neighborhood would be a huge factor in determining price.

However, many neighborhoods span the same price ranges

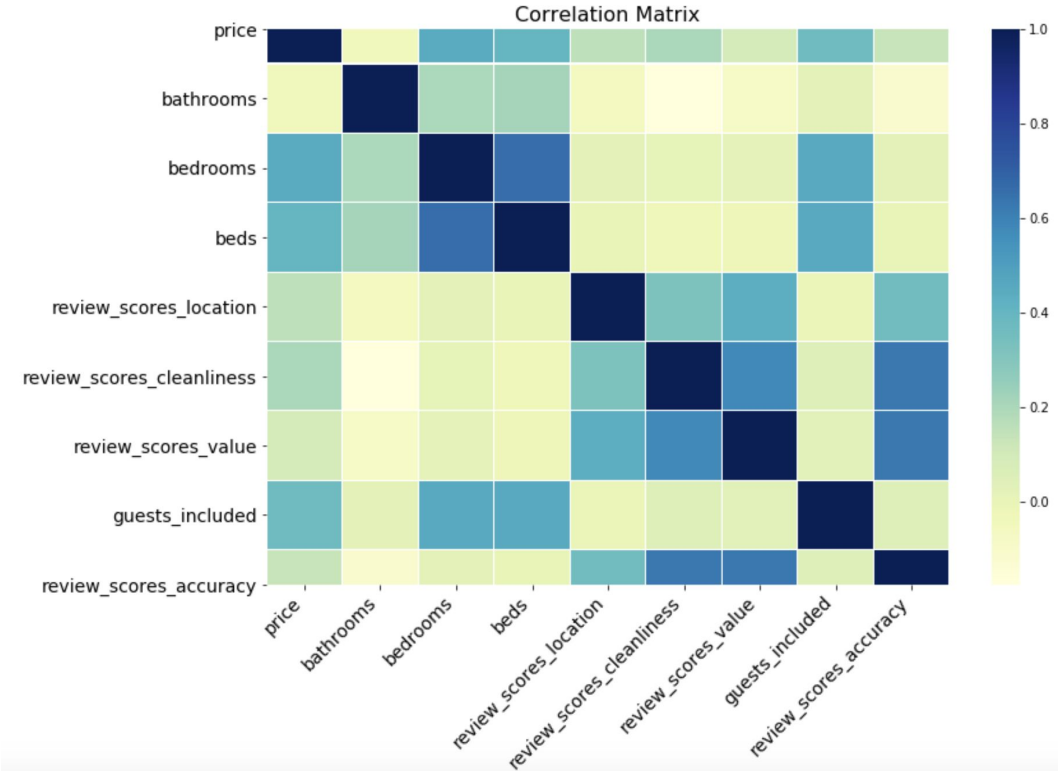


Findings

The variables seen to the right provide the highest correlation with price.

Yet, none reach above ~0.4, which is a weak to moderate positive relationship.

However, it is interesting to see bedrooms and beds have the strongest relationship with price



Findings - Decision Tree Results

While less than the mean of the test set, the RMSE is still relatively high for this data.

On average the errors of prediction are about $63, which could make a difference to a property owner or guest

RMSE

62.70

Test Set Stats

	price
count	1883.000000
mean	152.917153
std	76.857885
min	10.000000
25%	95.000000
50%	139.000000
75%	200.000000
max	350.000000

Limitations

There were many observations that had extreme prices and threw off predictions when left in the model. These had to be removed.

The ‘square ft’ variable had almost all missing values. I believe this would have been a useful feature for the model

More experience on my end with refining and fine tuning a model would have been beneficial for improving the prediction power

Conclusions

While there is some relationship between factors provided in this dataset, nothing could predict price with high level accuracy. The low level correlations and lack of a strong pattern in pricing data made analysis difficult, and ultimately provided a poor decision tree. Perhaps with more detailed location information or factors strongly related to price, an accurate prediction could be made.

Acknowledgements

Data was scraped by a kaggle user

Source: https://www.kaggle.com/jeploretizo/san-francisco-airbnb-listings

References

All analysis was done on my own