
airbnb
Airbnb Analysis &
Price Prediction

Our Team

Dan Zhou

Urvi Vaidya

Sem Leontev

Soo Lee

Nishant Parate

Dan Zhou

Urvi Vaidya

Sem Leontev

Soo Lee

Nishant Parate

airbnb

Agenda

Problem Statement

The business problem which is going to be resolved by this project

Data Source

Web scrapped Airbnb listing data from opendatasoft.com

Data Cleaning

Removing Nulls and fabricating data.

Insights from Data

Exploratory Data Analysis to find the useful information from data.

Predicting Price

Using machine learning model to predict the price


airbnb
Problem Statement
“ Airbnb Host’s unable to competitively price their new listing ”

Data Source

Web scraped Data

The data is scraped from the Airbnb website and contain all the important attributes.

Location

This data has information about all the listings in Los Angeles

Volume

This dataset has 19427 Rows and 89 columns

airbnb

Data Cleaning

The original dataset contains lot of unwanted data & null. Hence, cleaning is required before deriving insights from data

Removing Extra's

Extra columns and nulls are removed

Populating Nulls

Filling the null values with appropriate data

Transforming Data

Creating new columns from existing data

Removing Extra's

Original Dataset has a lot of unwanted columns which are not useful, this steps removes all the unnecessary data.

Extra Columns

-47 Col's

Removing Columns Like URL, image etc.

Null Columns

-6 Cols

Removed columns with more than 50% null values

Null Rows

-6953 Row's

Removed Nulls with more than 6 null values

Filling Null

Even after removing most of the null values there are still a few null values that need to be handled, this step takes care of those null values

Geo Coordinates

Latitude	Longitude
34.1477565	-118.5913378
34.1645972	-118.6005123
34.1831496	-118.6450638
34.1742446	-118.605876

Zip Codes

Zipcode
91364
91364
91367
91367

```
# Getting zipcode from latitide and long
def get_zipcode(df, geolocator, lat_field, lon_field):
    location = geolocator.reverse((df[lat_field], df[lon_field]))
    return location.raw['address']['postcode']


geolocator = geopy.Nominatim(user_agent='http')


zipcodes = df_null_zip.apply(get_zipcode, axis=1, geolocator=geolocator, lat_field='Latitude', lon_field='Longitude'
```

Filling Null

Even after removing most of the null values there are still a few null values that need to be handled, this step takes care of those null values

Accommodates

Accommodates
6
3

Number of Bedrooms

Bedrooms
3
1

Review Score

Review Scores
95
96

Service Rating

Accuracy	Clean	Checkin	Comm	Location	Value
9	10	10	10	9	10
10	9	10	10	10	10

Mean Cleaning Fees

Cleaning Fees

```
# Replacing null 'Cleaning Fee' column with its mean value
df8['Cleaning Fee'] = np.where(df8['Cleaning Fee'].isna(),round(df8['Cleaning Fee'].mean()), df8['Cleaning Fee'])
```

Transforming Data

After Filling nulls, the dataset had all the required values but to get insights from the data we need to transform the data.

Geo Coordinates

Latitude	Longitude
34.1477565	-118.5913378
34.1645972	-118.6005123
34.1831496	-118.6450638
34.1742446	-118.605876

Distance to Airport

Dist_to_Airp
20
9
11
10

List of Amenities

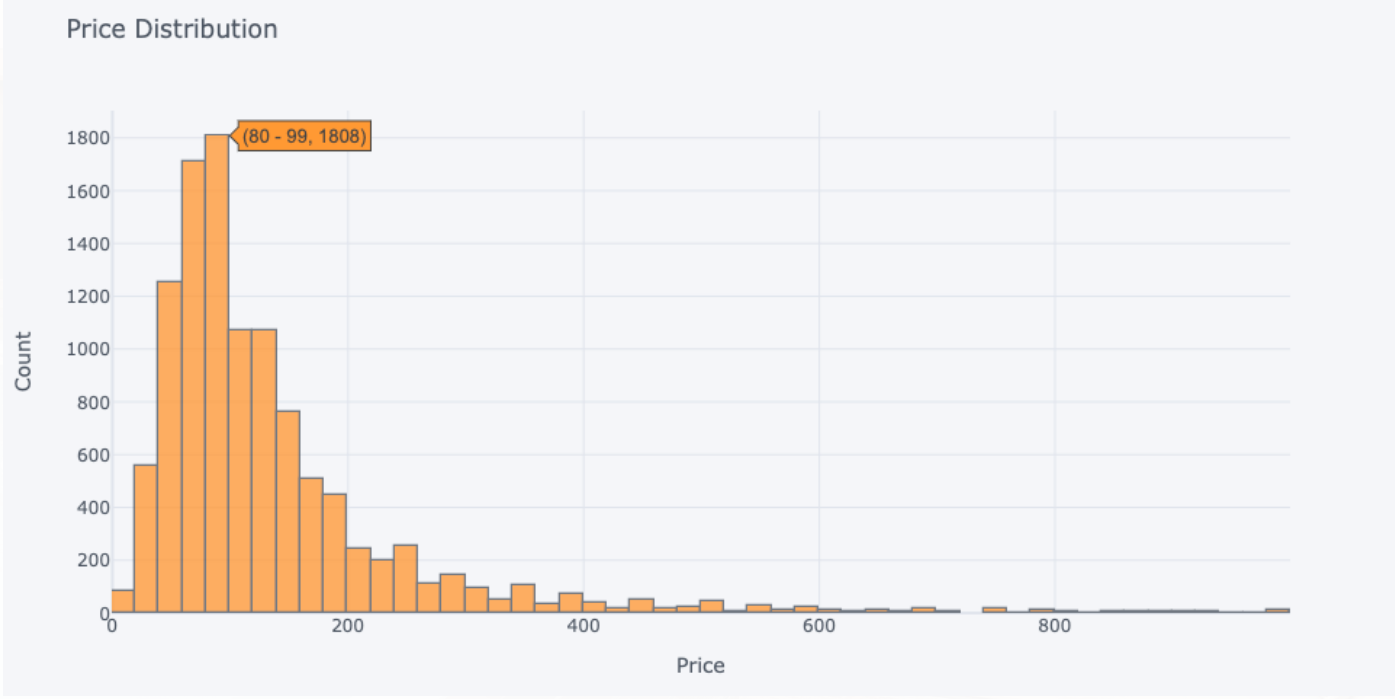
Amenities
TV,Wireless Internet,Air conditioning,Free p
Wireless Internet,Air conditioning,Pool,Kitcl
TV,Wireless Internet,Air conditioning,Kitche
TV,Cable TV,Internet,Wireless Internet,Air c

Number of Amenities

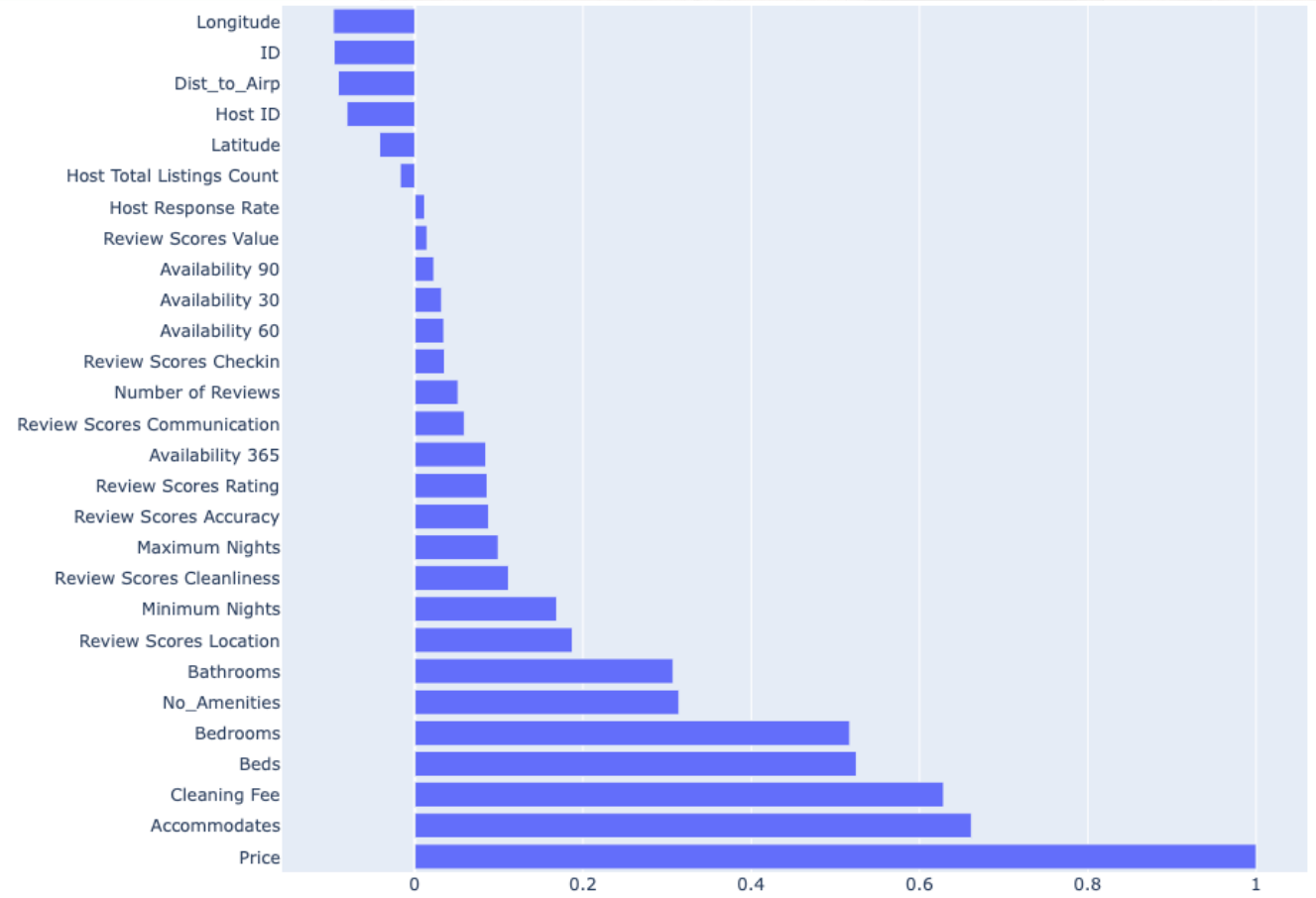
No_Amenities
21
19
22
22

Insights from Data

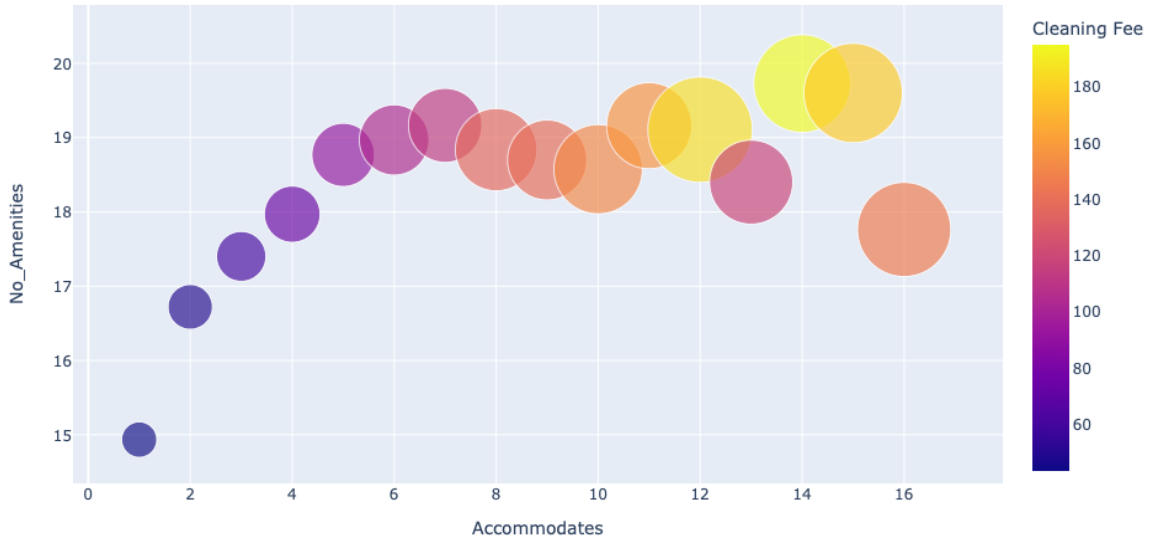
Price Distribution



Factors Influencing Price



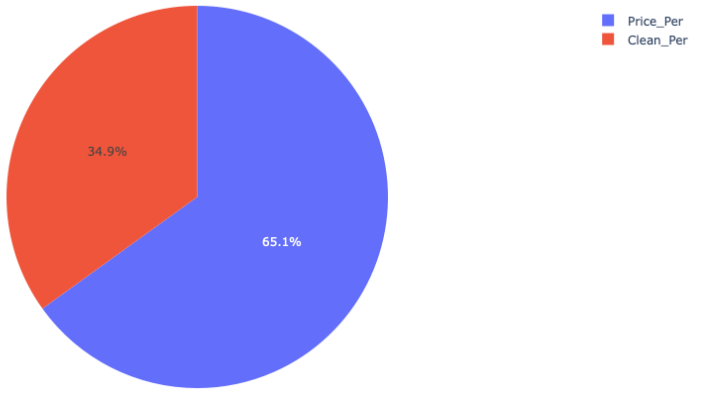
Relation of Price with important factors



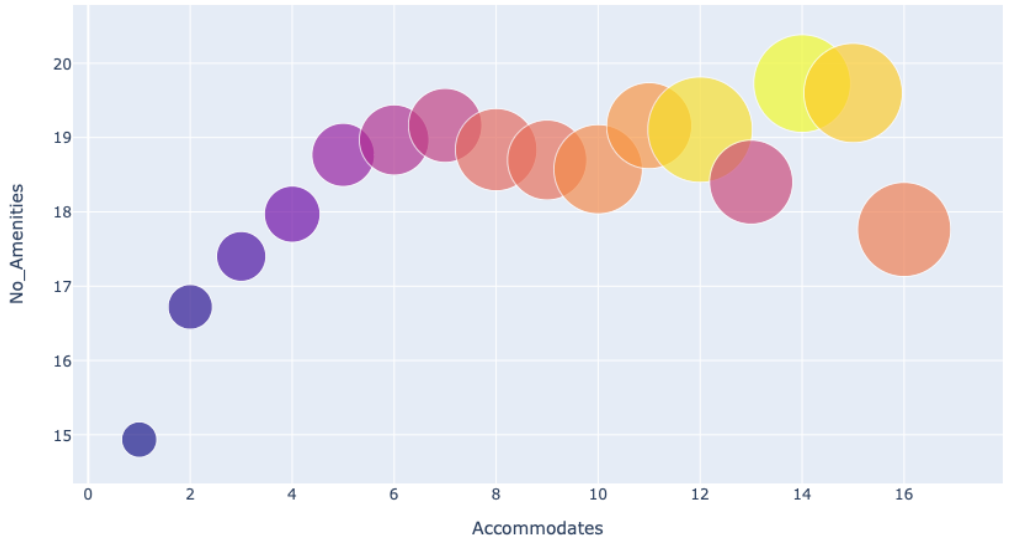
The Size of the bubble represents the Price

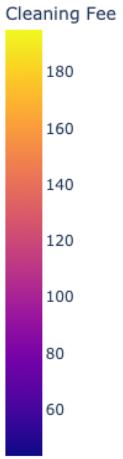
The color of the bubble represents Cleaning Fees

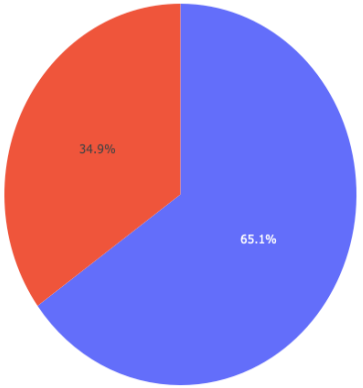
The Pie chart shows the average percentage of cleaning fees that is charged out of total cost.



Relation of Price with important factors

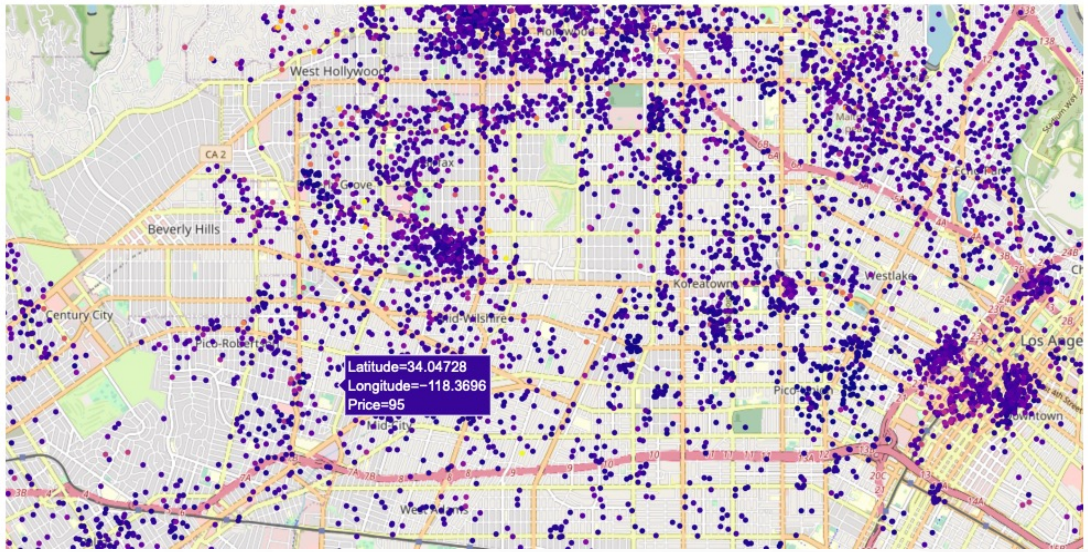






Relation of Price with Location

Air bnb Properties in LA

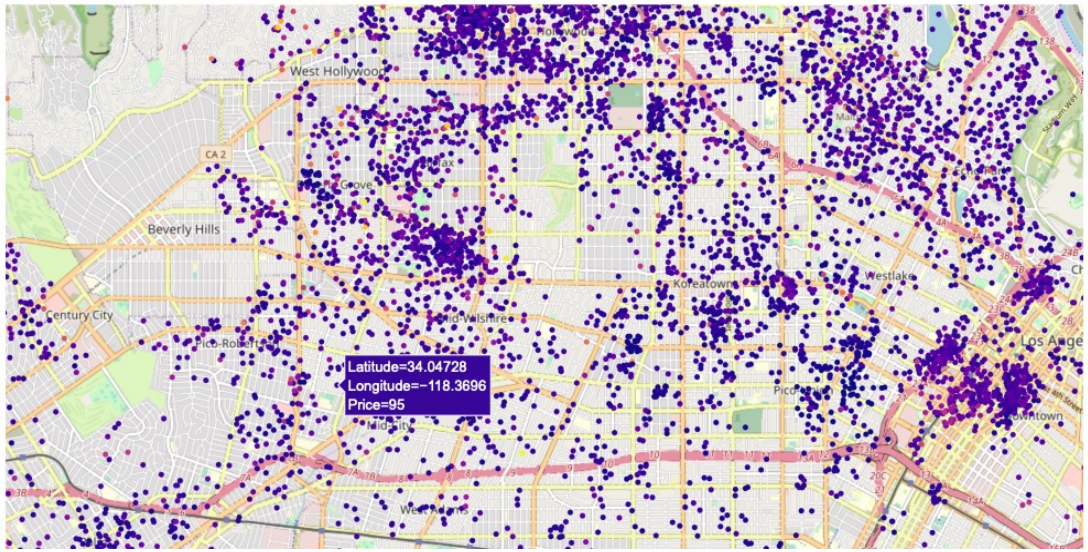


The map is interactive where you can zoom in to see price of a property

The average price of a property decreases with increase in distance from airport

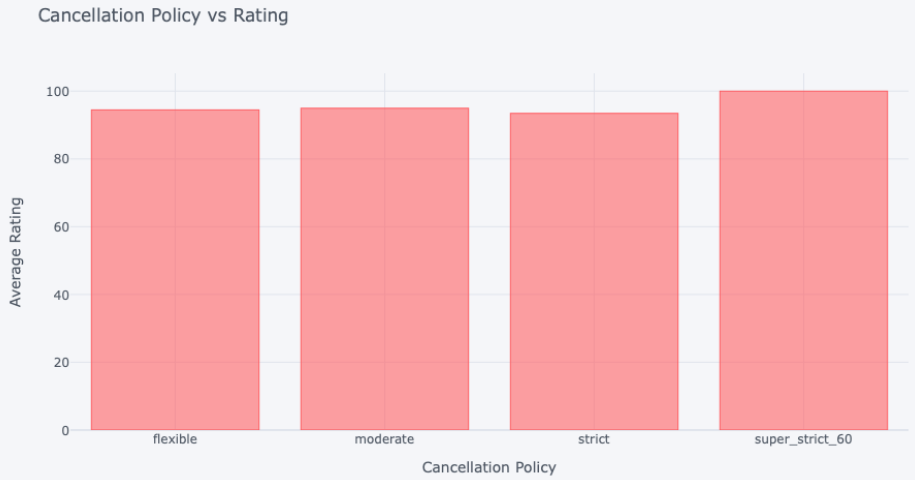
Relation of Price with Location

Air bnb Properties in LA



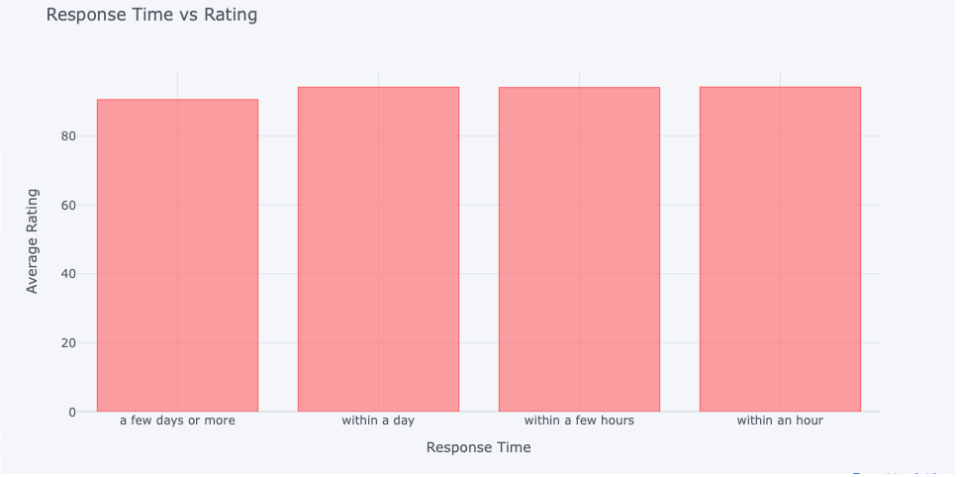
The average price of a property decreases with increase in distance from airport

Factors Affecting Rating

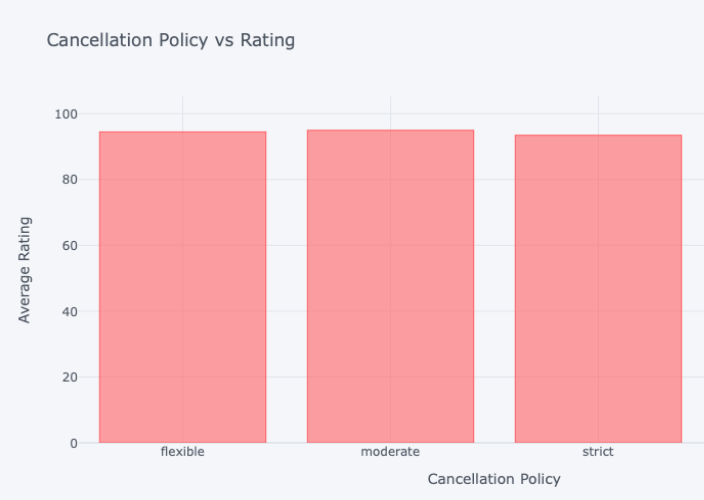


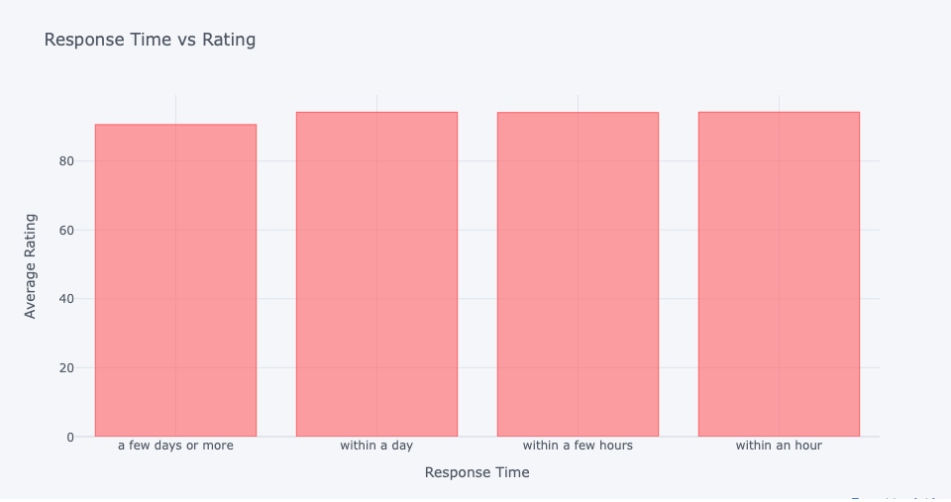
The average rating for a property is higher if the cancellation policy is super strict.

People prefer host who respond within a day rather than hosts that take a few days to reply

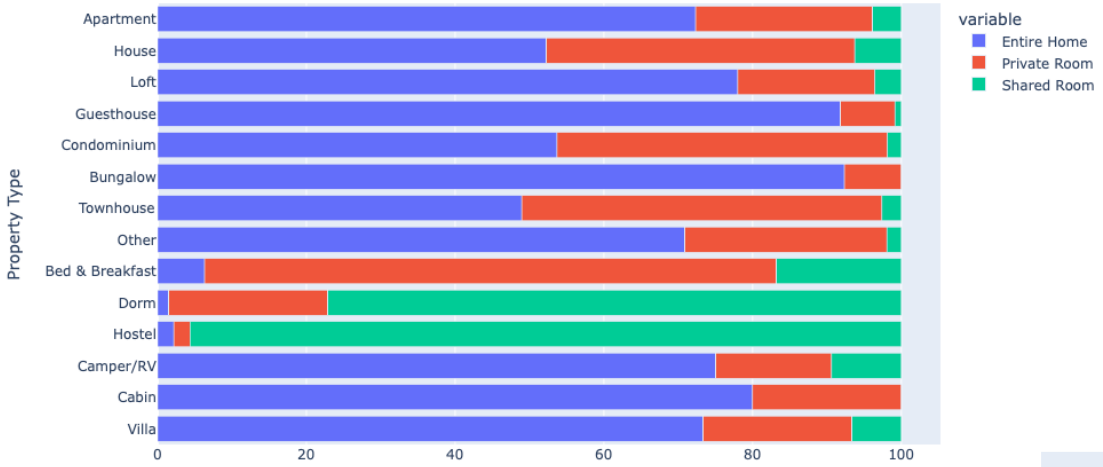


Factors Affecting Rating



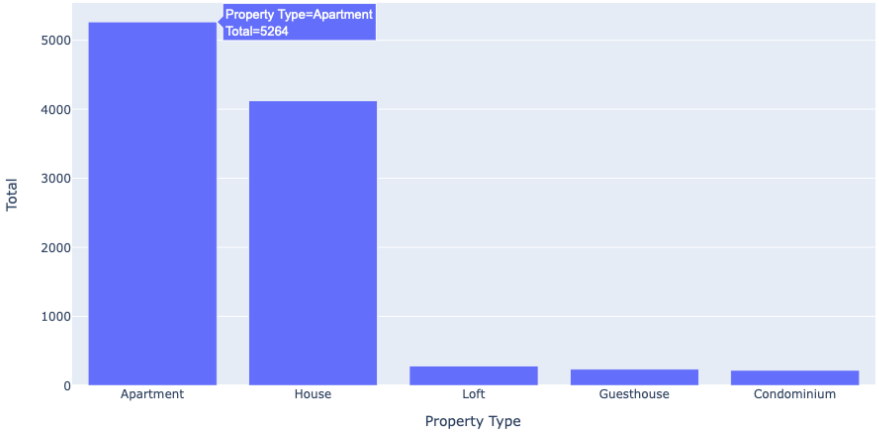


Market Analysis

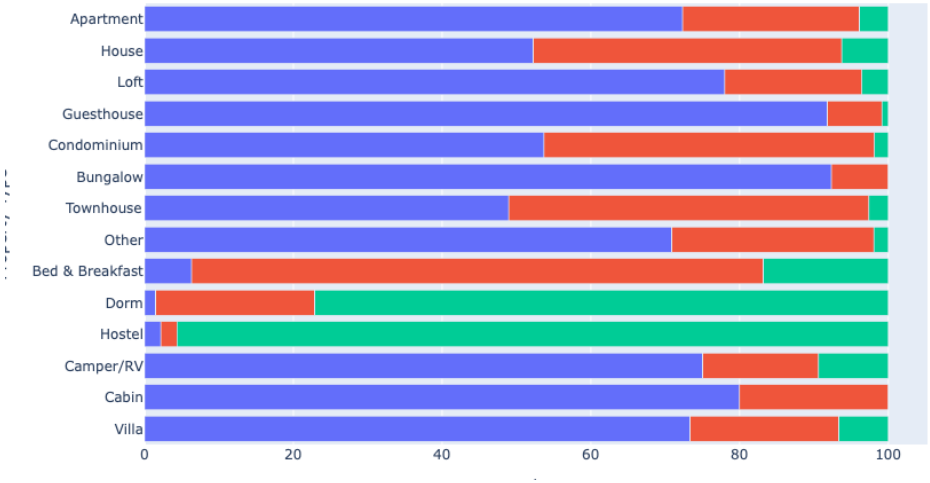


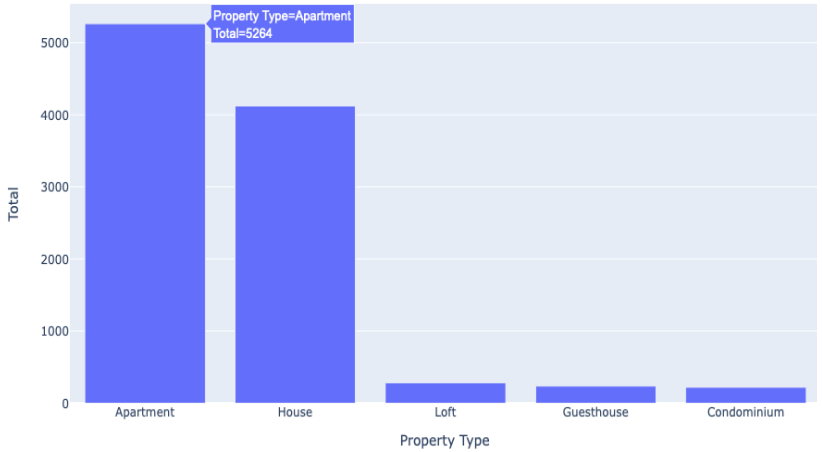
The graphs helps us understand distribution of different types of room by property type

Apartment and House are the most common type of properties that are listed on Airbnb



Market Analysis





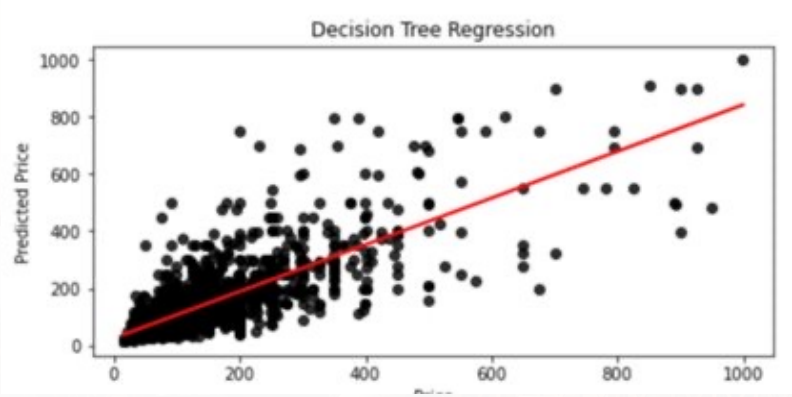
Machine Learning Algorithms

- Continuous output, ['Price']
- Decision Tree Regression
- Linear Regression

	Host Response Rate	Host Total Listings Count	Accommodates	Bathrooms	Bedrooms	Beds	Price	Cleaning Fee	Minimum Nights	Maximum Nights	Availability 30	Availability 60	Availability 90	Availability 365
0	100	36	6	2.5	3	4	450	200	3	1125	26	56	86	361
1	100	1	3	1.0	1	2	119	70	2	5	0	0	0	0
2	100	10	7	2.0	2	4	155	69	1	1125	15	25	51	138
3	90	1	3	1.0	1	1	140	50	2	10	8	8	8	8
4	100	10	6	2.0	2	3	149	95	2	90	9	23	33	33
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10937	96	16	6	1.0	2	4	147	90	3	1125	26	26	26	290
10938	96	23	1	0.0	1	1	59	70	1	365	0	22	52	327
10939	100	5	6	1.5	1	3	100	100	1	1125	29	59	81	349
10940	90	2	2	1.0	1	1	100	125	7	1125	30	60	90	365
10941	100	4	2	1.0	1	1	59	25	1	1125	21	28	28	302

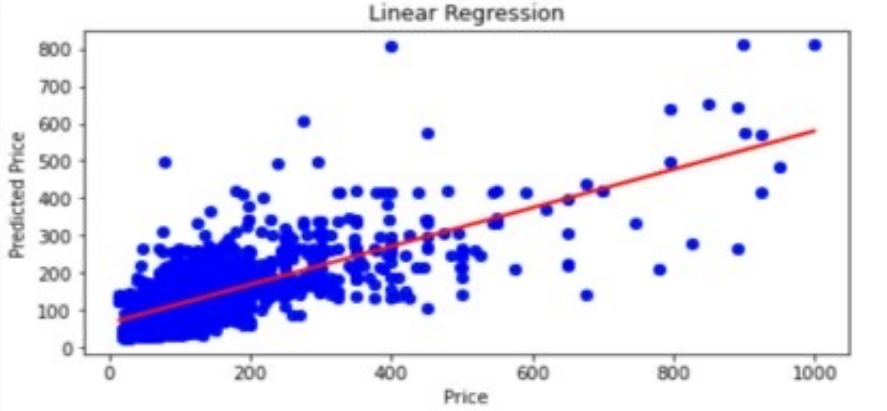
10942 rows × 58 columns

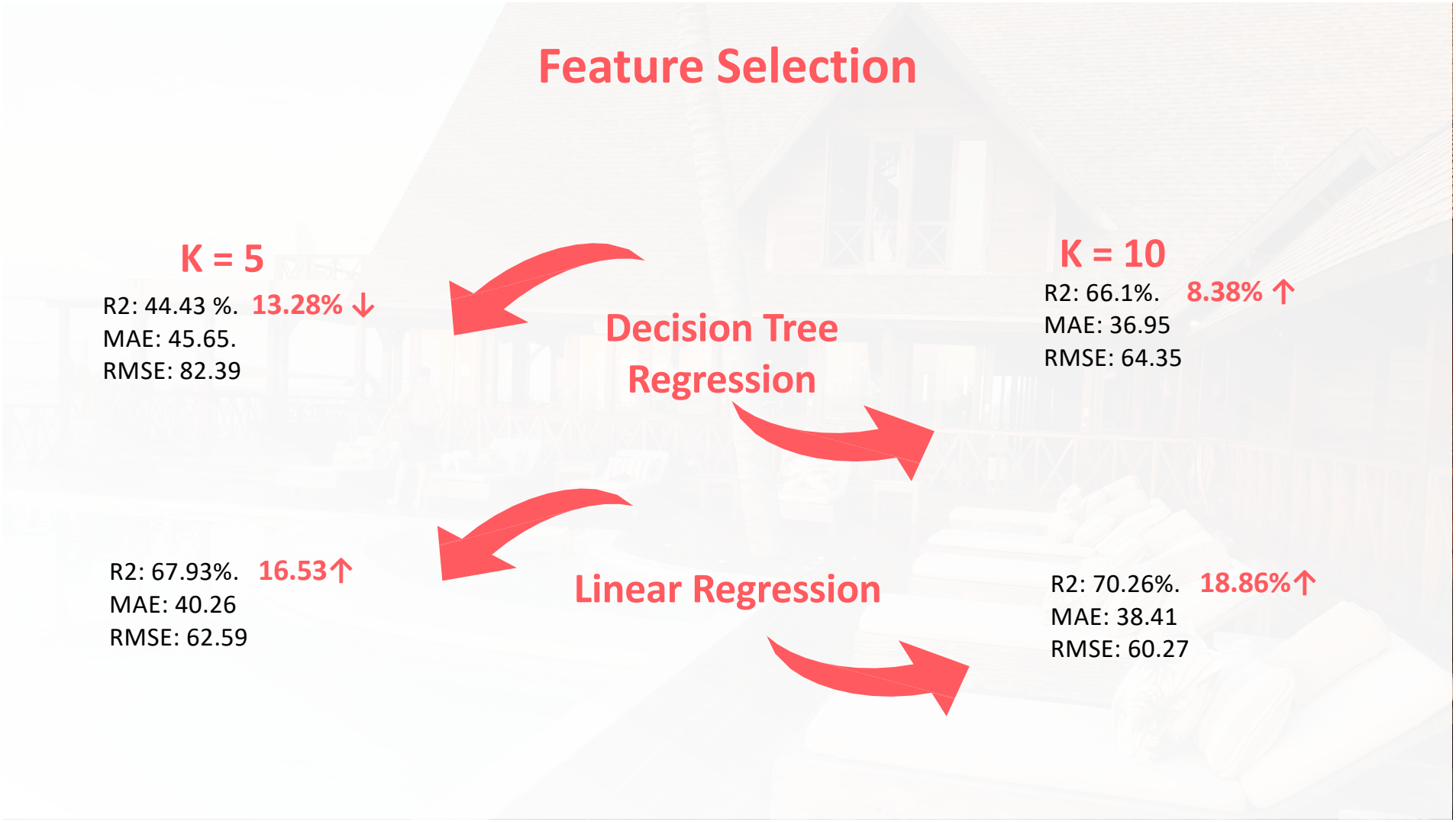
Benchmark with ALL Attributes



- R2: 57.71%.
- MAE: 41.3.
- RMSE: 71.88.

- R2: 51.40%
- MAE:50.48
- RMSE: 77.48




Feature Selection
K = 5
R2: 44.43 %. 13.28% ↓
MAE: 45.65.
RMSE: 82.39
Decision Tree Regression
K = 10
R2: 66.1%. 8.38% ↑
MAE: 36.95
RMSE: 64.35
R2: 67.93%. 16.53↑
MAE: 40.26
RMSE: 62.59
Linear Regression
R2: 70.26%. 18.86%↑
MAE: 38.41
RMSE: 60.27

Predicting Price

```
newInput = ['Accommodates':3,'Bathrooms':2,
            'Bedrooms':2,'Cleaning Fee':30,
            'Availability 90':1,'Review Scores Location':9,
            'Dist_to_Airp':5,'Property Type_House':1,
            'Room Type_Private room':0,'Room Type_Shared room':0]

pricePredict = regressor1.predict(newInput)
```

Predicted Price = $91


airbnb
“
Thank You,
Questions?
”