Problem

Existing host’s/ new property listers need to have a profitable price.

Why?

Small differences in prices will effect a change customer decisions and host's profitability thus resources can be prioritized.

airbnb

How?

Based on Neighbourhood, Property type, Amenities provided and review ratings.

Additional Benefits

- ❏ Understanding location wise customer demand.
- ❏ Customer preferences while booking the Airbnb.

Data Source: Inside Airbnb website.

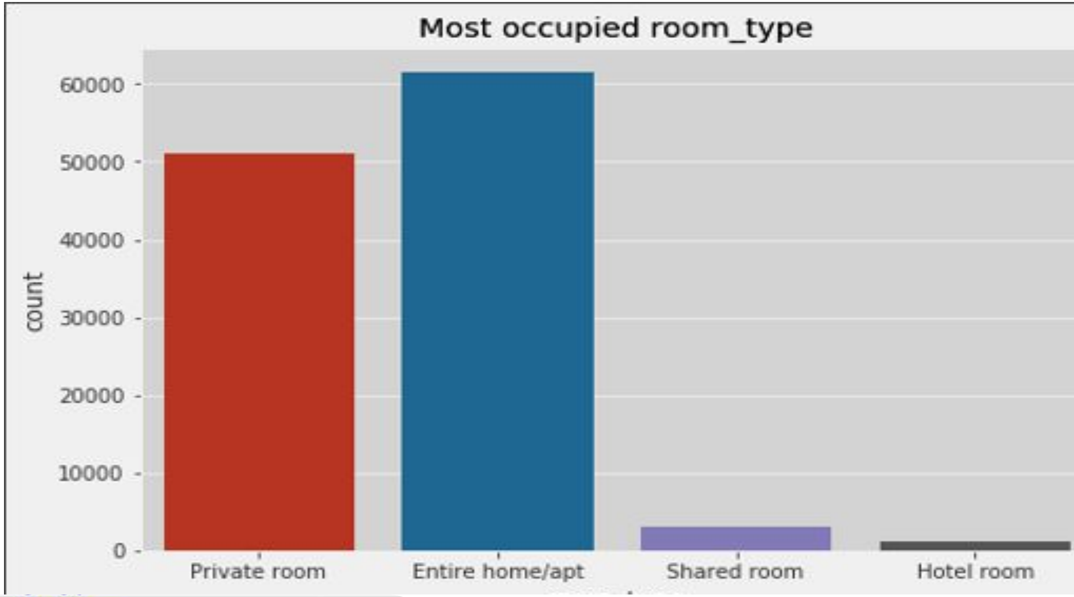
- ❑ We are trying to predict Price using Regression Analysis.
- ❑ Dataset contains 1,53,254 rows and 106 Features where, Categorical-64 & Numerical- 42.
- ❑ 20% of the data have NaN values.
- ❑ After EDA and feature engineering left with 45 features on those performed Statistical analysis,feature selection techniques and Assumptions of Linear Regression

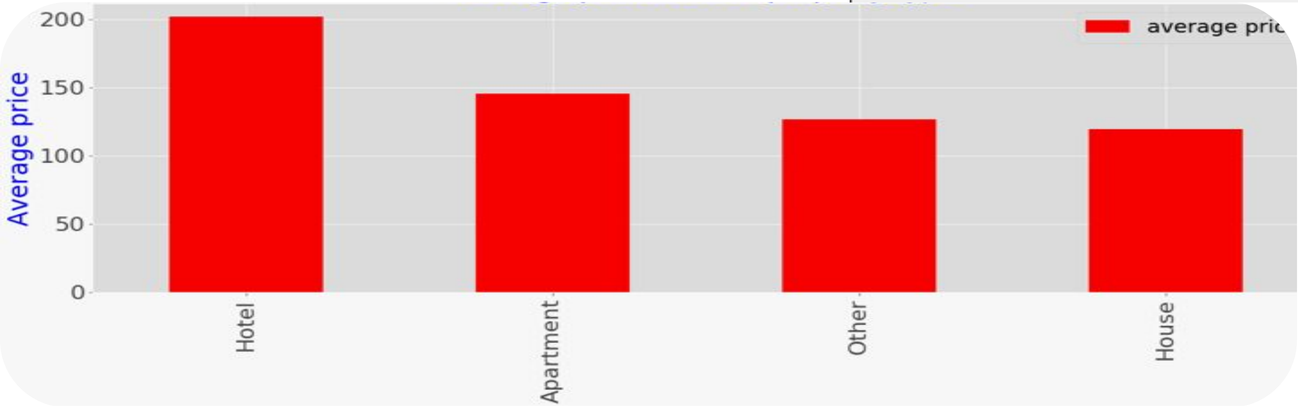
Challenges faced:

- ❑ Null values and imputation
- ❑ Data Cleaning
- ❑ Size of the data
- ❑ Feature Reduction
- ❑ Overfitting of the model

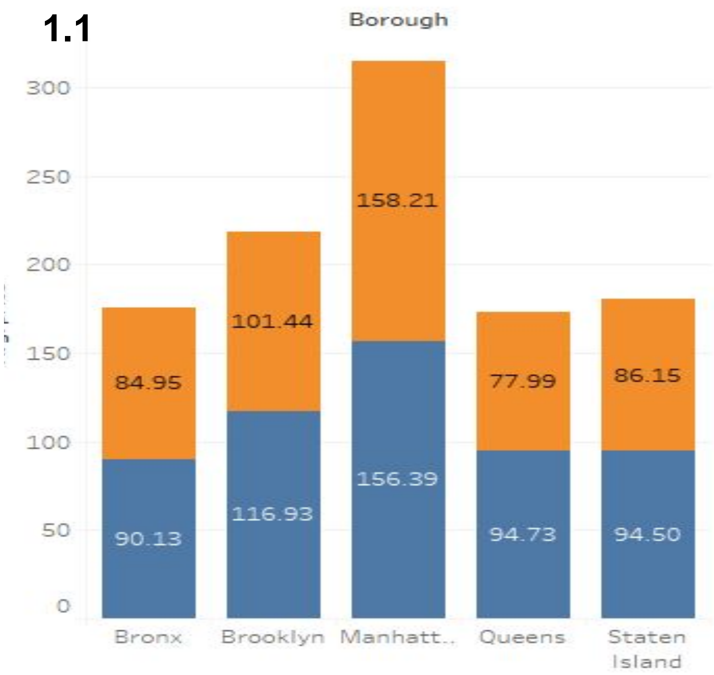
EDA

In Avg.Price by property type we can see hotels are highly priced and also by most occupied room we can see that hotels are least occupied.



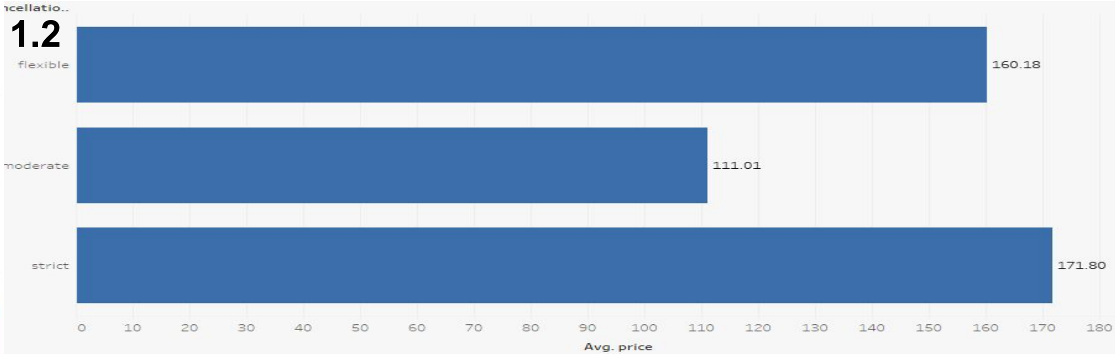


1.1



1.1 Avg price vs super host:
1.Orange colour represents the notsuperhost
2. Blue colour represent the supperhost.
3.This bar plot gives idea that superhost average prices are more than not superhost.
4.Because superhost gets some extra facility .

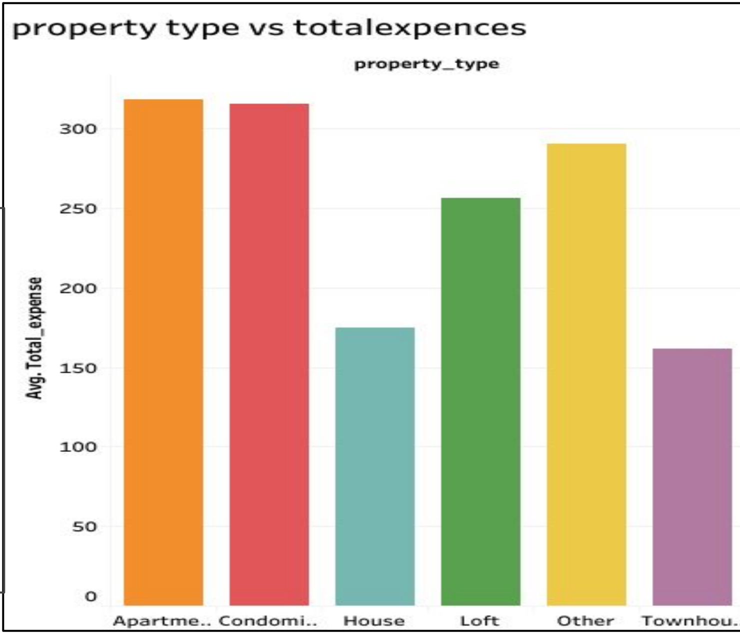
1.2

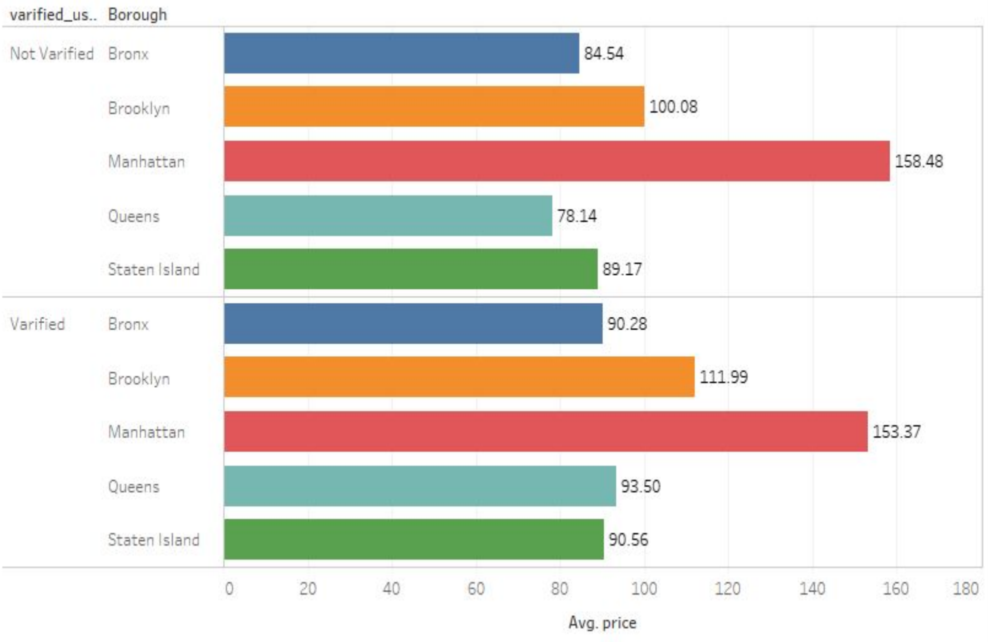


1.2 Cancellation_policy vs Avg price :For strict the average price is high.

1.3

1.3
Property type vs total expenses:
1.for apartment
Total expences are more.
2.Total _expences=
Cancellationfee+
security deposit





Verified host vs borough

1.Verified host are charging higher price as compared to non verified host because verified host strictly follow some rules and most people prefer that.

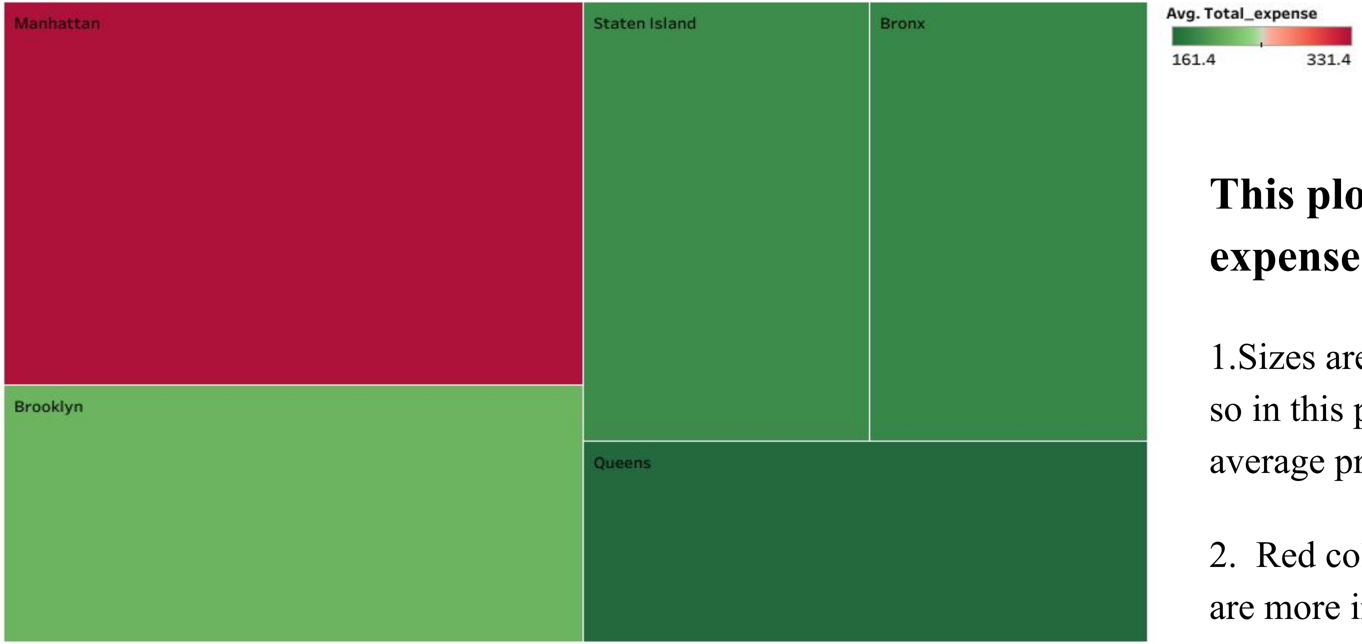
2.From this plot we can say verified host are affecting price .

Price per amenities available

1.This plots gives us idea that which property has some amenities present costs more compared to non amenities property.

2.Amenities like barbeque, swimming pool,nature views are affecting the price.

total expences



This plot describes :(Total expenses by borough)

1.Sizes are describing the average prices so in this plot Manhattan size is more so average price is more for this place.

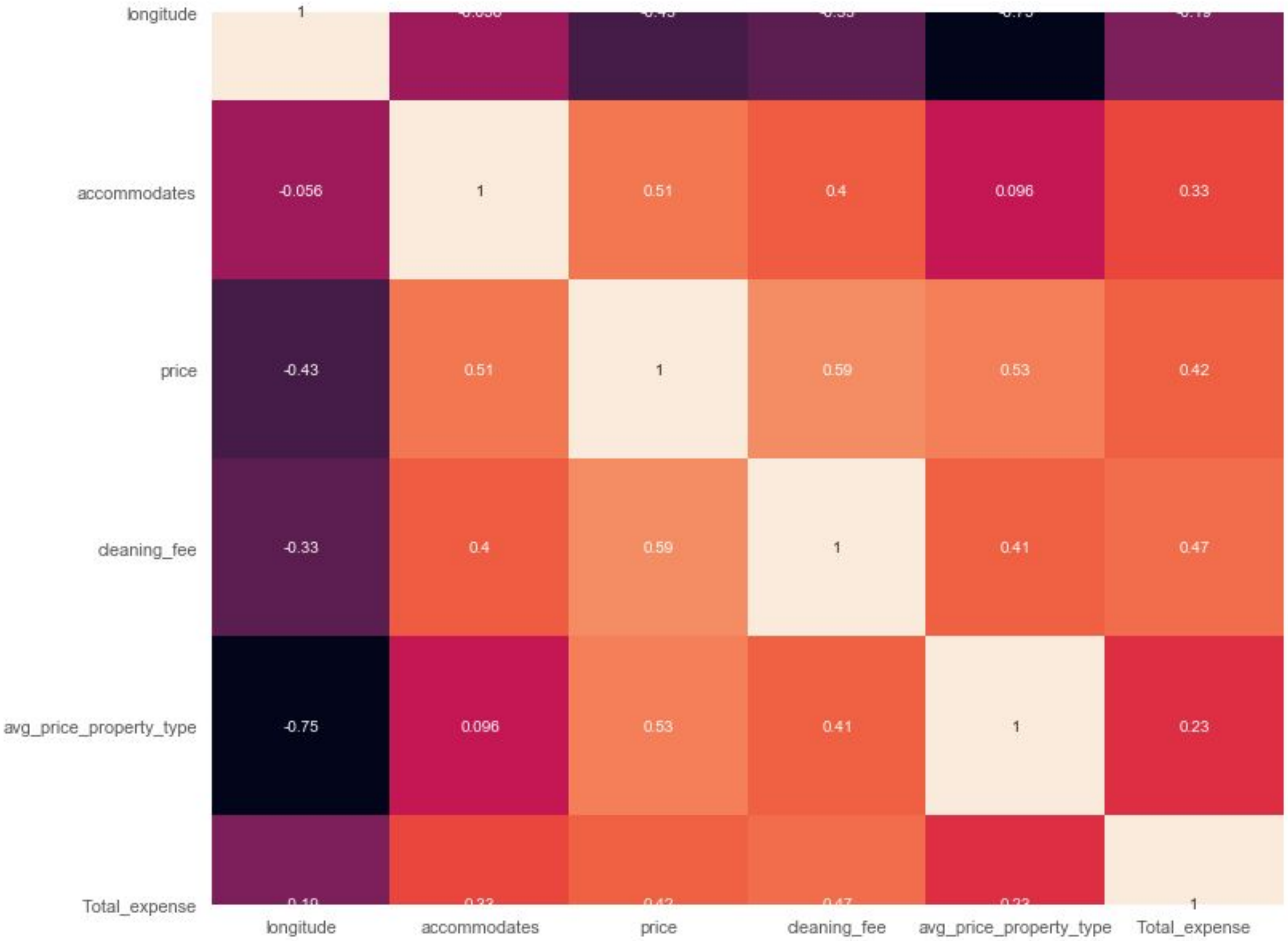
2. Red colour shows the total expenses are more in this place where green color shows that the total expenses are less.

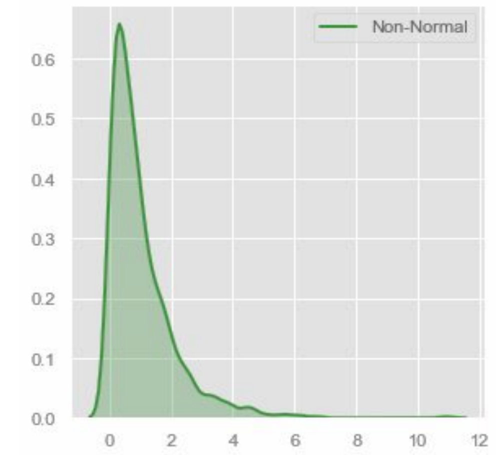
3. For Manhattan prices are more also total expences are more and for Queens total expences are less also average prices for property is less.

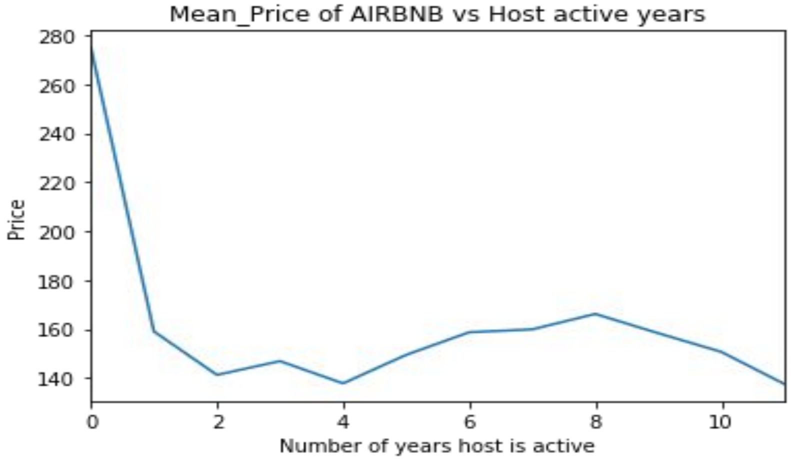
Correlation Analysis

Avg price per property type has strong negative correlation to Price.

Cleaning Fee also has 0.33 correlation with target variable



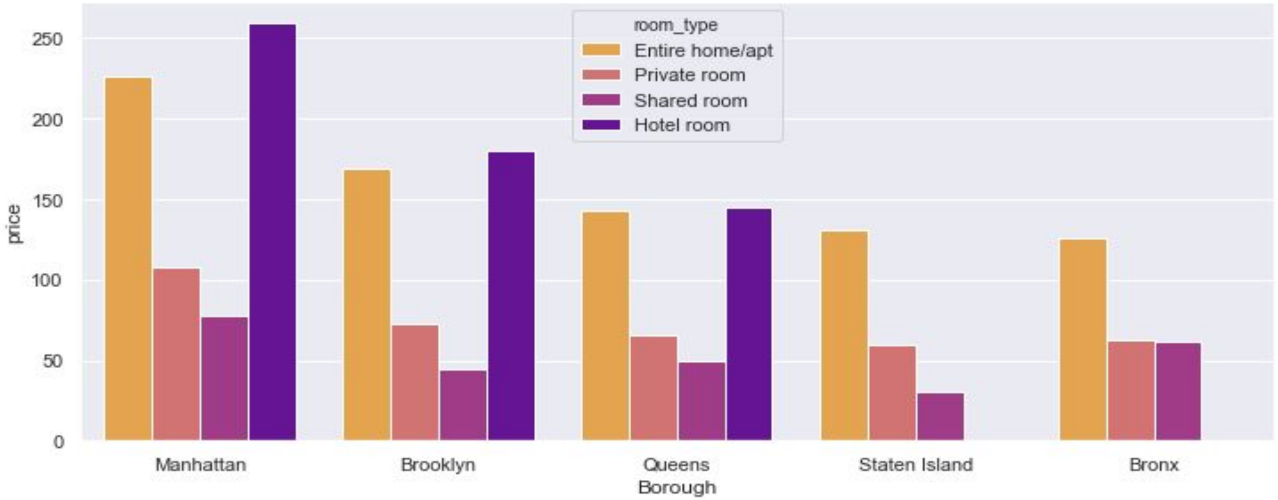




Above plot shows Box Cox transformation to reduce the skewness in the Data.

Above plot: As the age of Airbnb increases price decreases

Explains distribution of room type across boroughs



Statistical Analysis

1. To check significance of cancellation policy on price
2. To check significance of review scores on price
3. Variation of mean price across borough
4. Variation of mean price with different room types

Assumptions:

1. The level of significance is assumed to be 5 percent (i.e. alpha = 0.05)
2. Parametric tests(assumes already distribution is present(ANOVA) can be performed
3. Non-parametric tests(do not rely on any distributions(CHI-SQUARE) have to be performed.

- The variable 'price' contains large outliers. Therefore, to improve the normality of data, we will take the data between 25th percentile and 75th percentile, thereby eliminating the effects of outliers.

Testing the assumptions:

- Normality Test (shapiro test)
- Variance Test(levenge test)

Both the test failed so going with non parametric test -krushkal wallis test(alternative to one way anova)

Null hypothesis considered:

Price vs Neighbourhood (anova and kruskal wallis test) P value= 0.00

H0 (null hypothesis): mean_price(Brooklyn) = mean_price(Manhattan) = ... = mean_price(Bronx)

Prive vs Room_type

H0 (null hypothesis): mean_price(private_room) = mean_price(shared_room) = mean_price(entire_home/apt)

Room Type vs Neighbourhood Group , P value= 0.00

Chi Squared Test

H0 (null hypothesis): There is no association between Room Type and Neighbourhood Group.

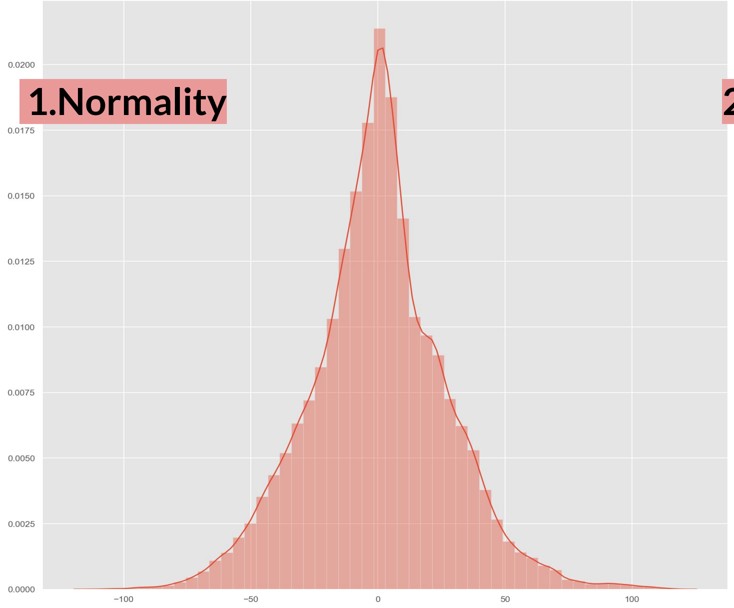
room_type vs property_type

H0 (null hypothesis): There is no association between Room Type and Neighbourhood Group.

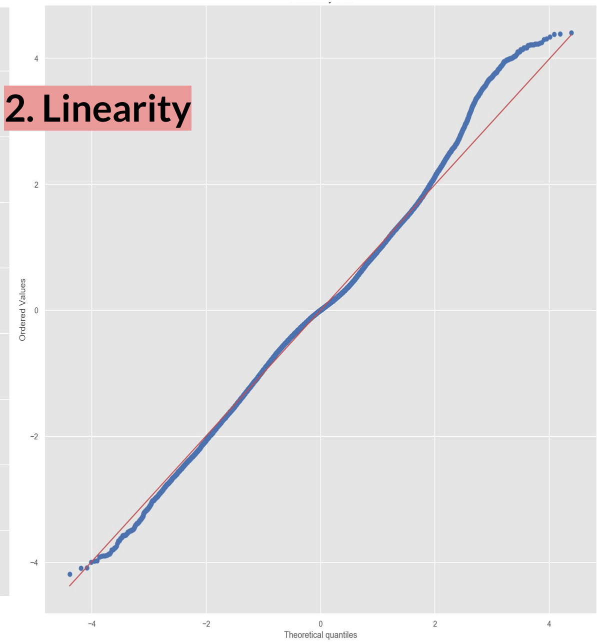
Concluding Remarks

- mean price has relation and dependence on room type and borough
- room type has association with borough and property type
- Cancellation policy has significant effect on price
- Review policy also has significant effect on the avg price

1.Normality



2. Linearity



3. Homoscedasticity -

Breuschpagan Test:
p-value:0.0,
P-value greater than .05 indicates homoscedasticity.

4. No Autocorrelation -

Durbin-Watson: 1.9038

Values of 1.5 < d < 2.5 generally show that there is no autocorrelation in the data

5. NO MULTI COLLINEARITY: after doing 4 iterations and eliminating features systematically below are some of the features left with

	room_type	cancellation_policy_strict	host_response_rate_bins_100%	bed_type_Airbed	host_total_listings_count	number_of_review	review_scores_accuracy	instant_bookable	host_days_active_years	host_listing_since	special_amenities	common_amenities	avg_price_property_type	avg_review_score
vif	2.008159	1.610802	0.0	1.502187	1.238783	1.490423	0.0	1.120396	1.31275	1.337671	1.149674	1.471343	1.170521	1.101603

One hot encoder with feature selection using RFE and applied Standard Scaler.

Model / Metrics	Linear Regression		Lasso Reg		Gradient Boosting	
	Train Scores	Test Scores	Train Scores	Test Scores	Train Scores	Test Scores
RMSE	27.0638	27.12201	27.065	27.13	25.270	25.4
R2	0.6369	0.6354	0.64	0.636	0.69	0.68
MAE	20.436	20.5123	20.43	20.511	18.6	18.75
MSE	732.45	735.603	732.52	735.52	643.6	649.88

One hot encoder with feature Selection using LassoCV and Standard Scaler.

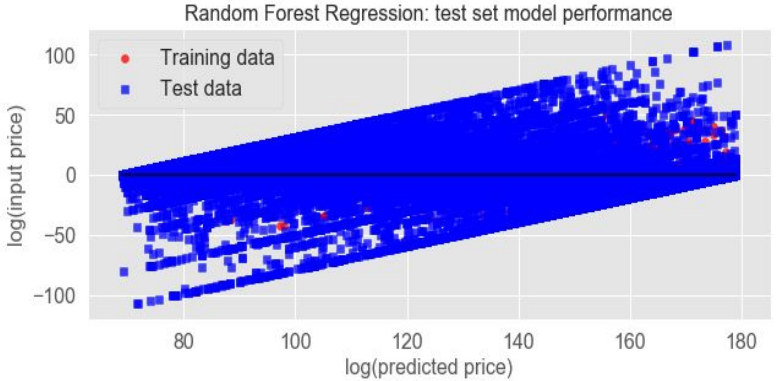
Model / Metrics	Linear Regression		Lasso Reg		Gradient Boosting	
	Train Scores	Test Scores	Train Scores	Test Scores	Train Scores	Test Scores
RMSE	27.06	27.122	27.12	27.06	25.84	25.98
R2	0.636	0.635	0.64	0.63	0.668	0.665
MAE	20.43	20.51	20.4	20.5	19.04	19.14

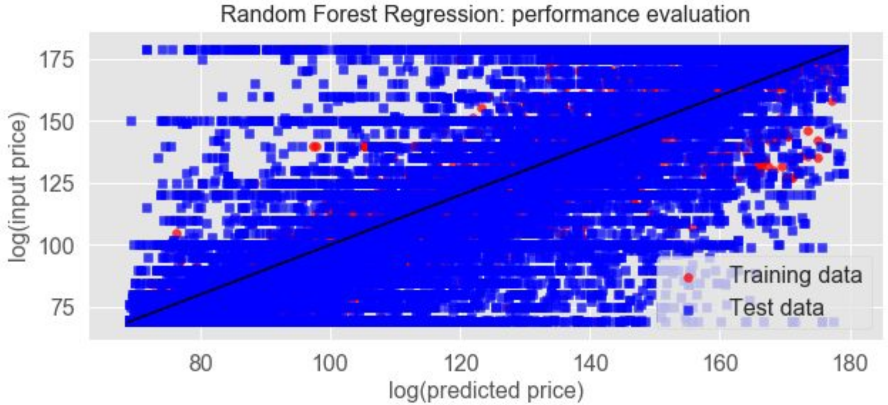
HyperParameter Tuning

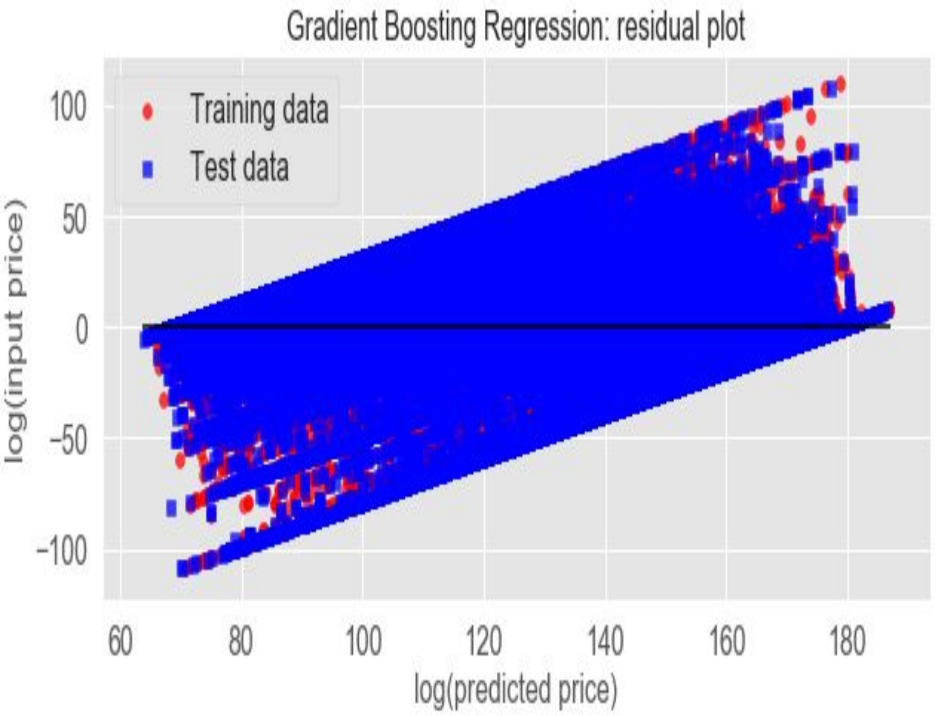
Model / Metrics	Random Forest regressor		Gradient Boosting Regressor	
	Train Scores	Test Scores	Train Scores	Test Scores
RMSE	6.95	17.5	22.24	23.45
R2	0.976	0.848	0.75	0.72

Our best accuracy model is Hypertuned Gradient Boosting Regressor.

Residual plots showing the Fit of Train and Test data







Further scope: Can apply ensemble techniques and can try advance ML techniques

Thank You