**Problem Statement – Based on dataset from**

1. Larger properties should receive more reviews because larger properties can accommodate more guests and therefore generate more traffic.
2. A property is overpriced is one of the most important factors in determining how many reviews it will receive. In general, the listings that are priced higher than listings of similar sizes and/or locations will receive fewer reviews than those that are priced lower. This is because people are more likely to leave a review if they feel like they got good value for their money.
3. Build a predictive model for reviews\_per\_month and compare different models or conduct variable selections.

Table

Description automatically generated

**Given Data Analysis-**

Understanding and preprocessing of dataset –

1. Dataset of 24886 entries with 29 columns
2. Filtering columns based on problem statement we need the following –
   1. Columns defining property type – property\_type, room\_type, accommodates
   2. Price
   3. Review scores for – rating, accuracy, cleanliness, check-in, communication, location, value
   4. Reviews per month – reviews per month
3. Data cleaning steps –
   1. Handling missing or null values
   2. Updating records with null values
   3. A close-up of a document

      Description automatically generated with medium confidenceCorrecting datatypes
   4. Review columns required
   5. Categorical encoding
   6. Normalization or standardization

**Handling and observations for special cases as per the dataset given –**

1. Handling null values for columns and set them to zero - 'reviews\_per\_month', 'accommodates', 'beds', 'bedrooms'
2. Change to ‘integer’ type for column – ‘accommodates’
3. Replacing values in column ‘room\_type’– ‘nan’, ‘6/21/22’ to the most frequent value - ‘Entire home/apt’, changing the type to ‘category’ since this is our categorical variable based on which most of the evaluation and analysis is done
4. Changing ‘price’ to float values removing the $ sign for more clear numbers, also replace nulls with 0
5. Data imputation of missing values with mean for the columns - 'review\_scores\_rating', 'review\_scores\_accuracy', 'review\_scores\_cleanliness', 'review\_scores\_checkin', 'review\_scores\_communication', 'review\_scores\_location', 'review\_scores\_value'
6. It is observed that “property\_type” has as many as 84 types and these can be combined under category of “room\_type” which correctly identifies the type of Property. The Column name is misleading in this case. Hence, we choose “room\_type” over “property\_type for further evaluation of the problem.

Chart

Description automatically generated**Initial data Analysis based on graphs –**

To understand the distribution of values for each of the filtered columns to understand their values and proportion for each of the values.

1. *Accommodates* distribution shows there are a greater number of availability for 4 number of people
2. Most of the review scores are right skewed and have similar distributions
3. The *reviews\_per\_month* dataset is right skewed which indicates there are majority of properties with are having no or very less review rates for a month, which we analyze further in the report.

Chart, bar chart, histogram

Description automatically generated

**Problem statement 1 –**

Larger properties should receive more reviews since they have higher accommodation.   
1. Firstly, on an average there is a correlation observed between the property types and the accommodation capacity as seen in adjacent graph. The properties rented out as an **entire house or apartments and shared room** were observed to have the **highest accommodation capacity.**

Chart, bar chart

Description automatically generated2. On Further analyzing the relation between the property types and their respective reviews received per month, **the Hotel rooms are receiving a greater number of reviews** as compared to entire home apartment or even for a shared rooms which were having higher accommodation capacity than the Hotel rooms.

Hence, **the assumption for larger properties to receive a greater number of reviews per month stands false or inapplicable in this case.**

Here the probability of having greater accommodation for a larger house does not imply to have a greater number of reviews for that property. Therefore, they can be considered as independent events.

Chart, bar chart, histogram

Description automatically generated**Problem statement 2 –**

To understand the correlation between the prices and the number of reviews received for each property. Interestingly, the **Shared room has highest pricing**. Viewing their relation for the number of reviews received for each property we can observed the **Shared room has lowest reviews.** Hence, the assumption of having high priced property types does not deduce that they will be reviewed highly. There looks like a negative correlation between the two.

Hence again the assumption of having high or overprices property to be highly reviewed is not applicable in this case.

Also, if we see the correlation matrix for the room\_types, accommodation capacity, price, vs reviews per month only larger room type and hotels have positive correlation whereas all other are holding a negative values which indicate the reviews are not really dependent on the price or the property size.

A picture containing timeline

Description automatically generated

Table

Description automatically generated**Building Prediction Model for given dataset –**

1. **OLS Regression model** –

Using OLS regression model, since I chose to have reviews\_per\_months to be more inclined on combined effect of based on the correlation matrix derived with other variable values like – room\_type, accommodates, price and reviews per month.   
Resultant variables to observe –

R-squared – it is low as 0.085

AIC is lower than BIC which sounds unacceptable

LogLikelihood of this model is negative

F-stats are too high to make these acceptable

Hence, I would like to explore more other model options

1. Chart, scatter chart

   Description automatically generated**Multivariable OLS Regression for explanatory variable**

For the Prediction model I used Multiple regression model where we use the multiple explanatory variables like – ‘price’, 'accommodates' and target variable here in this case is ‘reviews\_per\_month’. We also have one categorical variable as- room\_type with values – 'Entire home/apt', 'Hotel room','Private room','Shared’.

Box and whisker chart

Description automatically generatedI used the formula-based OLS model where response variable – “review\_per\_month’ is on the left and the explanatory variable are on the right. So our formula turns out as - “reviews\_per\_month ~ price’, ‘bedrooms’, 'accommodates'. The params shows model coefficients

1 intercept and 1 slope coefficient. The result graph looks non-linear.

Here we have chosen two variables for model i.**e multiple explanatory variables** to better fit the model.

We can clearly see the slopes of all the trend lines are parallel to each other that mean they have same slopes and equal in all plot dot axline calls. Hence it is also known as – “parallel sloped regression”.

Chart

Description automatically generated with medium confidence

We further evaluate the model Performance based **on Coefficient of determination (R-Squared) value** that determines how well the linear regression line fits the observed values. Larger **R-squared value the better.**

The residual standard error RSE is the typical size of the residuals of the errors identified from the expected values and actual values. Hence the smaller these values the better the model performance is. More explanatory variable increases R2. Hence Adjusted coefficient of determination help to solve this issue on having more explanatory variables.

The results are as follows - Text

Description automatically generated

The difference is minimal in the adjusted coefficient for all the variables. Hence the model with multiple variables has the best adjusted coefficient of all the other models, evaluating for Residual Standard Error for each model as above.

It is again observed that as we include more variables the value of RSE is lowest in case of combined variable model than the single linear model. Hence all metrics indicate models with two explanatory variables is better in performance than a single explanatory model.

It looks not a satisfactory solution to a prediction model. Hence, we will be trying to evaluate with other different model and their result measure its performance parameters.

Chart, line chart

Description automatically generated

1. **Linear Regression with Sequential Feature Selector**

On using Linear regression model with Sequential feature selector and there is a clear evidence of having more number of features to predict “*reviews\_per\_month*” the model performs better with 4 variables and having performance improved from -0.85 to -0.8 for the *‘neg\_mean\_squared\_error’* for Linear regression model.

1. Lasso CV with mse results –

Lastly using Lasso with StandardScaler function with Cross Validation has resulted into a *mean\_squared\_error* result of 0.847. Steps related to same can be viewed in the coding section.

**Conclusion-**

1. Data set needs to be properly validated before using in the model to build a robust prediction or estimation or a classification model
2. Data needs to be cleaned and clearly signify the important relationship interpreted through graphs, in this case it was a bit tough since there was no linearity in the datasets observed but the multivariable models. It is necessary to choose models based on the importance of the features that could possibly impact *reviews\_per\_month*.
3. Prediction models behaves the way we train without dataset, if there is no correlation between the variables selected the results will be no good. Hence, to have a better performing model it is necessary to have a dataset that will bring the best prediction results.