652- Multivariate Data Analysis

Final Project

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.

```
Data columns (total 29 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   id                           24885 non-null  object
 1   host_name                    24765 non-null  object
 2   host_since                   24765 non-null  object
 3   host_response_time           19664 non-null  object
 4   host_response_rate           19664 non-null  object
 5   host_acceptance_rate         20072 non-null  object
 6   host_is_superhost            24764 non-null  object
 7   host_total_listings_count    24764 non-null  float64
 8   host_has_profile_pic         24764 non-null  object
 9   host_identity_verified       24764 non-null  object
 10  neighbourhood_cleansed       24881 non-null  object
 11  latitude                     24881 non-null  float64
 12  longitude                    24881 non-null  float64
 13  property_type                24881 non-null  object
 14  room_type                    24881 non-null  object
 15  accommodates                 24881 non-null  float64
 16  bathrooms_text               24828 non-null  object
 17  bedrooms                     23512 non-null  float64
 18  beds                         24629 non-null  float64
 19  price                        24880 non-null  object
 20  review_scores_rating         17213 non-null  float64
 21  review_scores_accuracy       16951 non-null  float64
 22  review_scores_cleanliness    16951 non-null  float64
 23  review_scores_checkin        16951 non-null  float64
 24  review_scores_communication  16951 non-null  float64
 25  review_scores_location       16950 non-null  float64
 26  review_scores_value          16951 non-null  float64
 27  instant_bookable             24880 non-null  object
 28  reviews_per_month            17213 non-null  float64
dtypes: float64(14), object(15)
```

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 –
    a. Columns defining property type – property_type, room_type, accommodates
    b. Price
    c. Review scores for – rating, accuracy, cleanliness, check-in, communication, location, value
    d. Reviews per month – reviews per month
3. Data cleaning steps –
    a. Handling missing or null values
    b. Updating records with null values
    c. Correcting datatypes
    d. Review columns required
    e. Categorical encoding
    f. Normalization or standardization

```
RangeIndex: 24886 entries, 0 to 24885
Data columns (total 29 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   id                           24885 non-null  object
 1   host_name                    24765 non-null  object
 2   host_since                   24765 non-null  object
 3   host_response_time           19664 non-null  object
 4   host_response_rate           19664 non-null  object
 5   host_acceptance_rate         20072 non-null  object
 6   host_is_superhost            24764 non-null  category
 7   host_total_listings_count    24764 non-null  float64
 8   host_has_profile_pic         24764 non-null  object
 9   host_identity_verified       24764 non-null  category
 10  neighbourhood_cleansed       24881 non-null  object
 11  latitude                     24881 non-null  float64
 12  longitude                    24881 non-null  float64
 13  property_type                24881 non-null  object
 14  room_type                    24881 non-null  category
 15  accommodates                 24886 non-null  int64
 16  bathrooms_text               24828 non-null  object
 17  bedrooms                     24886 non-null  int64
 18  beds                         24886 non-null  int64
 19  price                        24886 non-null  float64
 20  review_scores_rating         24886 non-null  float64
 21  review_scores_accuracy       24886 non-null  float64
 22  review_scores_cleanliness    24886 non-null  float64
 23  review_scores_checkin        24886 non-null  float64
 24  review_scores_communication  24886 non-null  float64
 25  review_scores_location       24886 non-null  float64
 26  review_scores_value          24886 non-null  float64
 27  instant_bookable             24880 non-null  category
 28  reviews_per_month            24886 non-null  float64
dtypes: category(4), float64(12), int64(3), object(10)
```

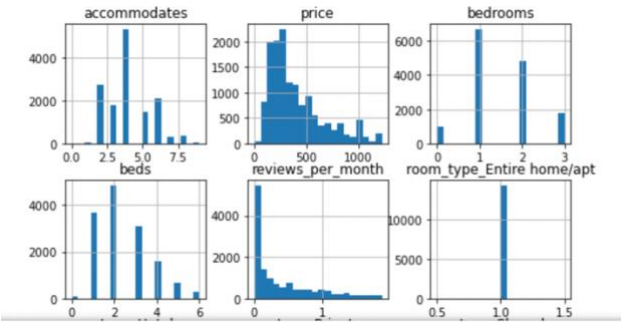
Handling and observations for special cases as per the dataset given –

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.

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.

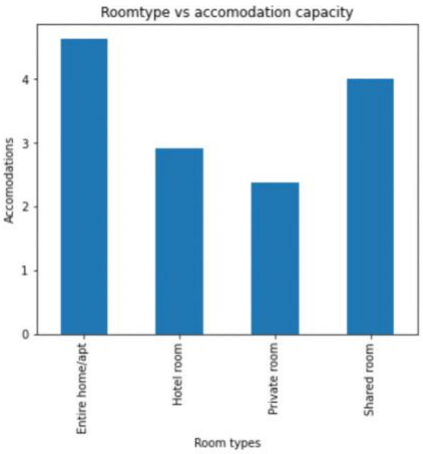


Final Project

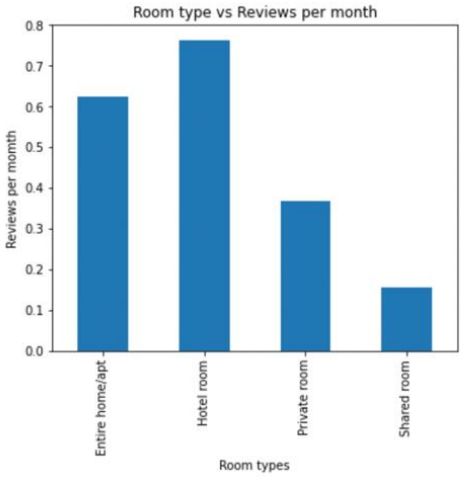
Problem statement 1 –

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.**



2. 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.**

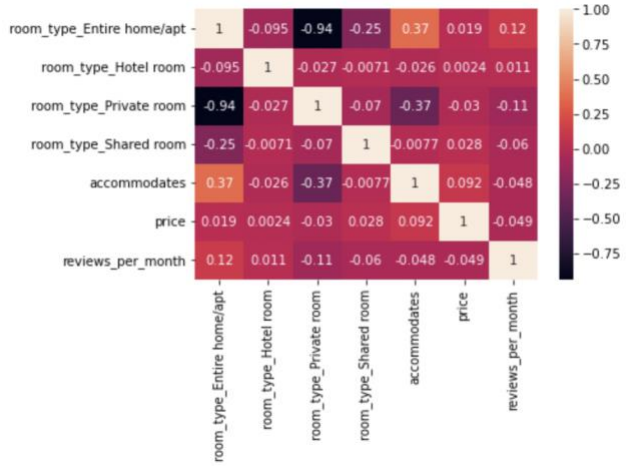
Here the probability of having greater accommodation for a larger house does not imply to have a greater number of reviews for that property. Therefore, they can be considered as independent events.

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.



652- Multivariate Data Analysis

Final Project

Building Prediction Model for given dataset –

1. **OLS Regression model** –

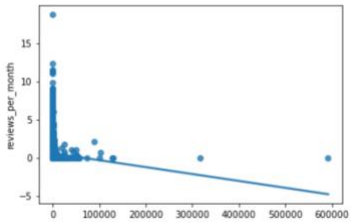
Using OLS regression model, since I chose to have reviews_per_months to be more inclined on combined effect of based on the correlation matrix derived with other variable values like – room_type, accommodates, price and reviews per month.
Resultant variables to observe –
R-squared – it is low as 0.085
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

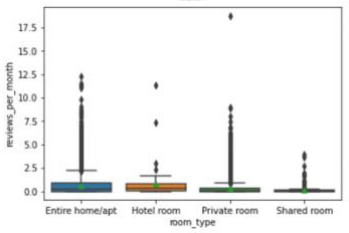
OLS Regression Results

Dep. Variable:	reviews_per_month	**R-squared:**	0.085
Model:	OLS	**Adj. R-squared:**	0.085
Method:	Least Squares	**F-statistic:**	441.2
Date:	Fri, 16 Dec 2022	**Prob (F-statistic):**	3.82e-274
Time:	18:56:46	**Log-Likelihood:**	-9718.4
No. Observations:	14253	**AIC:**	1.944e+04
Df Residuals:	14249	**BIC:**	1.948e+04
Df Model:	3		
Covariance Type:	nonrobust		

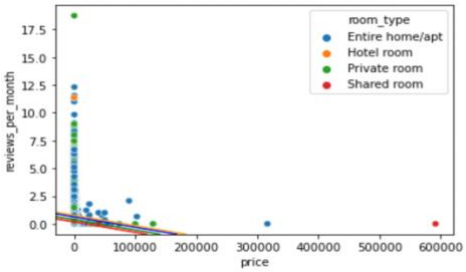
2. **Multivariable OLS Regression for explanatory variable**

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'.
I 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".





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 R^2. Hence Adjusted coefficient of determination help to solve this issue on having more explanatory variables.
The results are as follows -



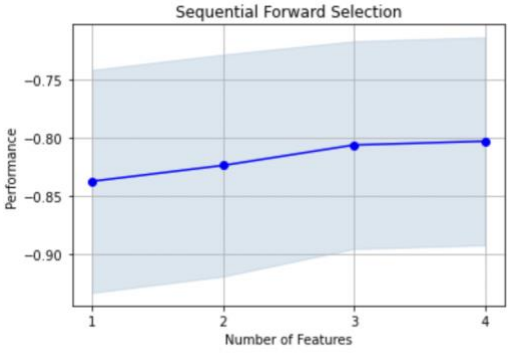

```
R-squared for Price vs reviews per month model is : 0.002403825770026069
R-squared for Category - roomtype vs reviews per monthmodel is : 0.016496073903080277
R-squared for Multiple variable model is : 0.019050208272687774

R-squared Adjusted for Price vs reviews per month model is : 0.0023637359060882934
R-squared Adjusted for Category - roomtype vs reviews per month model is : 0.016377469900254682
R-squared Adjusted for Multiple variable model is : 0.01889247394373994

RSE for Price vs reviews per month model is : 0.8462114606767809
RSE for Category - roomtype vs reviews per monthmodel is : 0.83442993851009
RSE for Multiple variable model is : 0.8322964019092711
```

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.

3. **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



Kanika Yadav

performance improved from -0.85 to -0.8 for the *'neg_mean_squared_error'* for Linear regression model.

4. 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.