1. **Abstract**

This project analyzes the short-term rental market in Portland city, Oregon. The data we will be working with is the Airbnb listings dataset. The analysis uses complex statistical approaches to investigate the distribution of listings across neighborhoods, detect pricing and availability patterns over time, and forecast the popularity of specific kinds of properties or places. So, in this project we have performed explanatory data analysis, Linear regression, GLM’s and GAM’s to find out the best fit for our model to gain some insights. This project aims at the aspects that influence the demand for short-term rentals, such as the number of beds and bathrooms, location, and the quality of previous visitors' reviews. The results of the analysis provide insights into the short-term rental market in Portland city.

***Keywords:******EDA, generalized linear models, generalized additive models, linear regression***

1. **Introduction**

Airbnb is a company operating an online marketplace for short-term home stays and experiences. The company acts as a broker and charges a commission from each booking. Airbnb opens the door to interesting homes and experiences. It's almost like a Home away from a Home. Hotels can be expensive. Especially, when traveling in large groups or with family. Most people choose Airbnb because they are cheaper, you also get the local experience, see your destination through the eyes of a local, and get a small glimpse of how locals live. The Room type also varies in each listing, be it private or shared room.

The dataset that we will be using includes the bulk of the data — listing name, neighborhood, host, room type, price, minimum nights a listing must be booked for, number of reviews for that listing, date of last review, average reviews per month, availability (how many days out of 365 the listing is available for booking), and number of listings per host, number of bedrooms and bathrooms and more.

So, this analysis is useful for the people who travel more and for the people who looking to invest in short-term rental properties, and this helps to enhance their business strategies. Also, when a traveler is looking for places to stay, it takes time to go through each listing and determine if it matches his needs. The cost of a listing varies depending on the neighborhood, amenities, number of rooms, and kind of property. Advanced statistical analysis, on the other hand, can give useful insights into the factors that influence price, occupancy rates, and other crucial metrics in the short-term rental market. we can detect patterns and trends in data on rental properties, bookings, and customer reviews, allowing property owners, investors, and policymakers to make better decisions.

1. **Data**

This dataset is taken from insideairbnb.com website. The website contains information regarding Airbnb listings in all the major cities across the world.

This data is completely observational. The data on the website is collected through the observations of Airbnb users who list their properties and the guests who book them. since the data is collected because of natural interactions between hosts and guests, without any manipulation or interference from researchers, we can say that the data is purely observational.

Some of the vital attributes in our dataset used for analyzing short-term rental market:

Neighborhood, latitude, longitude, property type, room type, accommodates(number of listing could accommodate) ,number of bedrooms, number of beds, amenities , price , minimum number of nights that listing could be book for , maximum no.of nights that listing could be book for, no.of reviews and total no.of host listings in Portland . travelers could choose from a wide variety of property types ranging from bungalows, apartments to tree houses, boat houses , mansions and villas . the room types available are shared room, private room and hotel rooms.

* 1. **Data Cleaning:**

when comes to data cleaning, there are a lot of unnecessary variables and outliers in our dataset which are not useful for analyzing and predicting the prices and removing the features which do not impact the price . Some of the features that were removed are scraped id, url of listing, host url, location of host, host responses, host picture url’s and more. In the dataset there were many NA values and missing values in many columns. For instance: we had a column named “ Number of reviews “ where we have replaced those values with the mean of the respective columns.

A picture containing text, screenshot, diagram, line

Description automatically generated

The density plot of price looks to be positively skewed, which may be due to outliers. Another peak may be seen around the price range of $10,000.

There is no way a listing could cost more than $1,000 a night, thus we can simply rule out the possibility that some listings might cost $10,000 per night.

For this project we are only considering listing which have price less than 300 $. Since there are not much listings having price more than 300 $ per night .

Chart, histogram

Description automatically generated

Before cleaning:

A screenshot of a computer

Description automatically generated with low confidence

After cleaning:

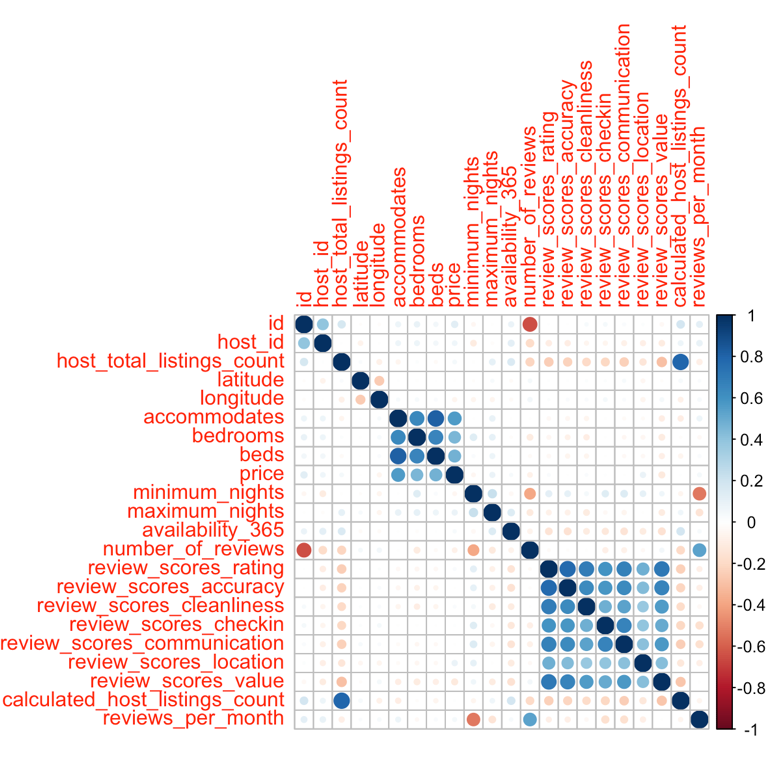
A screenshot of a computer

Description automatically generated with low confidence

1. **Exploratory Data Analysis:**

Exploratory Data Analysis (EDA) is an important phase in the statistical analysis process since it aids in understanding the features of the data and identifying patterns or correlations that can be used to influence further studies.

Correlation plot for all the attributes to find if any correlation occurs between the variables:

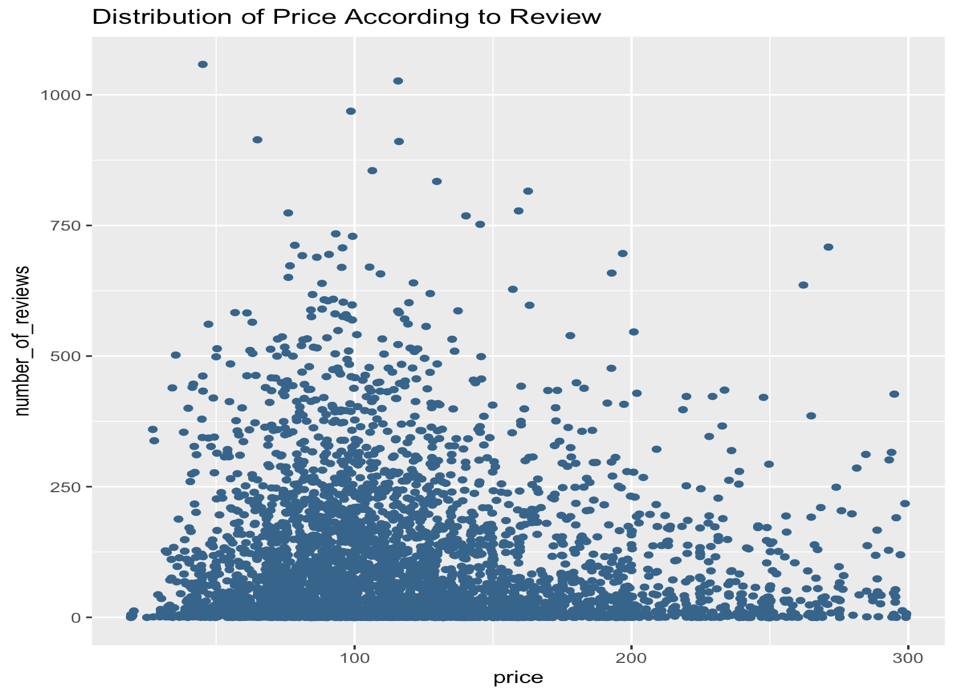


This correlation shows that the price is closely correlated with the number of beds, bedrooms, bathrooms, and accommodations. It is evident from this that prices rise as the number of rooms increases. Furthermore, the reviews are strongly co-related with each other.

A picture containing text, screenshot, design

Description automatically generated

From this plot, we could see that more number of listings, have the room type of entire room, rather than shared rooms.



From this scatter plot of price v/s number of reviews, we could see that listings which have price less 200 have more number of reviews. Though the difference is not huge, there may be some effect of number of reviews on price.

Chart

Description automatically generated

From this density plot we could infer that, room types having entire shared room cost less price than private rooms in shared space or entire space.

Chart, histogram

Description automatically generated

From this density plot of price v/s accommodate, we could infer that, listings which could accommodate a greater number of people are high priced.

1. **Methods:**

Our main goal is to model the relationship between the independent variables and the dependent variable while controlling for the effects of other variable and to see which factors affect the price variables for that we first fit a linear model.

**5.1 ANOVA:**

ANOVA is a generalization of the t-test, which is used to compare the means of two groups.

Here in this project, we perform one-way ANOVA where we model price as a function of room type, number of reviews, accommodates, minimum nights, beds, bedrooms, property type and neighborhood individually.

Findings from this ANOVA test:

Through the summary generated by the ANOVA test. we observed that the independent variables (accommodates, bedrooms, beds) had higher F- score value and lower AIC than other variables. From this we can say that there is more correlation between

(i)price vs accommodates

(ii) price vs bedrooms

(iii) price vs beds

A screenshot of a computer code

Description automatically generated with low confidence

**5.2 Linear Regression:**

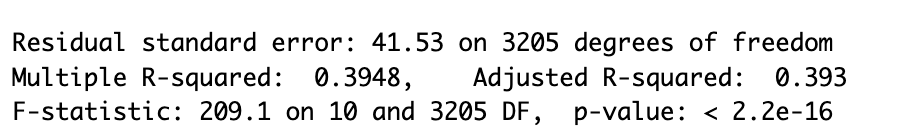
**Price Regression analysis:**

In this analysis we tried to predict the price and we are training the dataset with the ratio of 80:20

**Model 1:**

price as a dependent variable and the independent variables are accommodates, bedrooms ,minimum nights , number\_of\_reviews , availability\_365 , beds ,maximum\_nights , room\_type.

We obtain the value of Adjusted R-Square as: 0.393



The adjusted R-squared value is 0.39, which indicates that the independent variables explain 39% of the variation in the dependent variable "price".

**Model 2 :**

Independent variables : accommodates , bedrooms , minimum\_nights , number\_of\_reviews ,availability\_365 , beds , room\_type , calculated\_host\_listings\_count, latitude, longitude, reviews\_per\_month.

Dependent variable : price

We obtain the value of Adjusted R-Square as: 0.4039

**Text

Description automatically generated**

The adjusted R-squared value is 0.4039, which is slightly higher than the previous model. This suggests that including latitude and longitude and removing availability\_365 and maximum nights in the model improves the goodness-of-fit.

**Model 3:**

The modified independent variables for this model are accommodates, bedrooms, number\_of\_reviews, reviews\_per\_month,minimum\_nights , latitude ,longitude , beds , room\_type, calculated\_host\_listings\_count.

So now we obtained the value of Adjusted R-Square: 0.4071

Text, letter

Description automatically generated

The adjusted R-squared value is 0.4071, which is slightly higher than the previous model. This suggests that including host\_total\_listings\_count in the model further improves the goodness-of-fit.

Overall, the adjusted R-squared values suggest that the independent variables in Model 2 and Model 3 explain more of the variation in the dependent variable "price" than Model 1. However, it's important to consider other factors such as the statistical significance of the coefficients and the assumptions underlying the regression models before drawing any conclusions.

Text

Description automatically generated

From the summary of model 3, we could say that “accommodates”, “Private room type”, “bedrooms”, “minimum nights”, “number\_of\_reviews”, and “host\_total\_listings\_count” are the most important to the price. Some are negative correlation, and some are positive correlation ones.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model1 | Model2 | Model3 |
| AIC | 33108 | 33052 | 33035 |
| BIC | 33181 | 33143 | 33126 |
| Adjusted R2 | 0.393 | 0.4039 | 0.4071 |

**5.3 Prediction:**

Based on the information provided, Model 3 has the highest Adjusted R-Square value (0.407) and the lowest AIC and BIC values (33035 and 33126, respectively), indicating that it is the best-fitting model. Therefore, Model 3 would likely be the best model to use for prediction. However, it's important to also consider other factors such as the assumptions of the model and the validity of the results before making any predictions.

**A picture containing text, receipt

Description automatically generated**

From the Actual v/s Predicted values, we can see that the model did a relatively good job at predicting the price.

A picture containing text, screenshot, diagram, map

Description automatically generated

**5.4 Generalized Linear Model:**

To run the same models with GLMs, we would first need to specify the appropriate probability distribution and link function for the dependent variable.

For example: Our dependent variable(price) is continuous and has a relatively normal distribution, we can use the Gaussian distribution and the identity link function.

If the dependent variable is binary, we can use the Bernoulli distribution and the logit link function.

We fit generalized linear models (GLMs) to our training data using price as response variable and other combinations of predictor variables used in the earlier mentioned linear models. Then we generate predicted values for the response variable using the model fit on the testing data. Later, we calculated mean squared prediction error (MSPE), Root Mean square error, R2 value between the actual and predicted price values on the test data. The AIC, BIC values were also calculated on the fitted model.

**Model 1:**

Price as a dependent variable and the independent variables are accommodates, bedrooms, minimum nights , number\_of\_reviews , availability\_365 , beds ,maximum\_nights , room\_type.

Text

Description automatically generated

**Model 2:**

Price as a dependent variable and the Independent variables are accommodates , bedrooms , minimum\_nights , number\_of\_reviews ,availability\_365 , beds , room\_type , calculated\_host\_listings\_count, latitude, longitude, reviews\_per\_month.

A screenshot of a computer code

Description automatically generated with low confidence

**Model 3:**

Price as a dependent variable and the Independent variables for this model are accommodates, bedrooms, number\_of\_reviews, reviews\_per\_month, minimum\_nights , latitude ,longitude , beds , room\_type, calculated\_host\_listings\_count.

Text

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model1 | Model2 | Model3 |
| AIC | 33110 | 33059 | 33059 |
| BIC | 33176 | 33138 | 33138 |
| Adjusted R2 | 0.388 | 0.4000 | 0.4028 |
| MSPE | 1843 | 1767 | 1750 |

**5.5 Multicollinearity:**

From the above models we could see that AIC and BIC scores chose model Model 2 and R2 and MSPE chose model Model 3. Then, we check for multicollinearity to see if there is any collinearity among predictor variables.

Table

Description automatically generated

From the above output we could say that there is only relatively little multicollinearity among the predictor variables. Only accommodates variable had a bit higher VIF of 3.6. This could lead to reduced interpretability of the model. We could perform ridge regression to help reduce the impact of multicollinearity.

A picture containing text, screenshot, font, design

Description automatically generated

From the correlation plot, we could see that there are pairwise correlations between bedrooms and accommodates, bedrooms and beds, accommodates and bedrooms, accommodates and beds.

**5.6 Ridge Regression:**

Ridge regression is a regularization technique that helps to reduce the impact of multicollinearity by shrinking the regression coefficients towards zero.

Table

Description automatically generated with medium confidence

From the above output*,* the coefficients estimated by the ridge regression model represent the change in the response variable for a unit change in the corresponding predictor variable, holding all other variables constant.

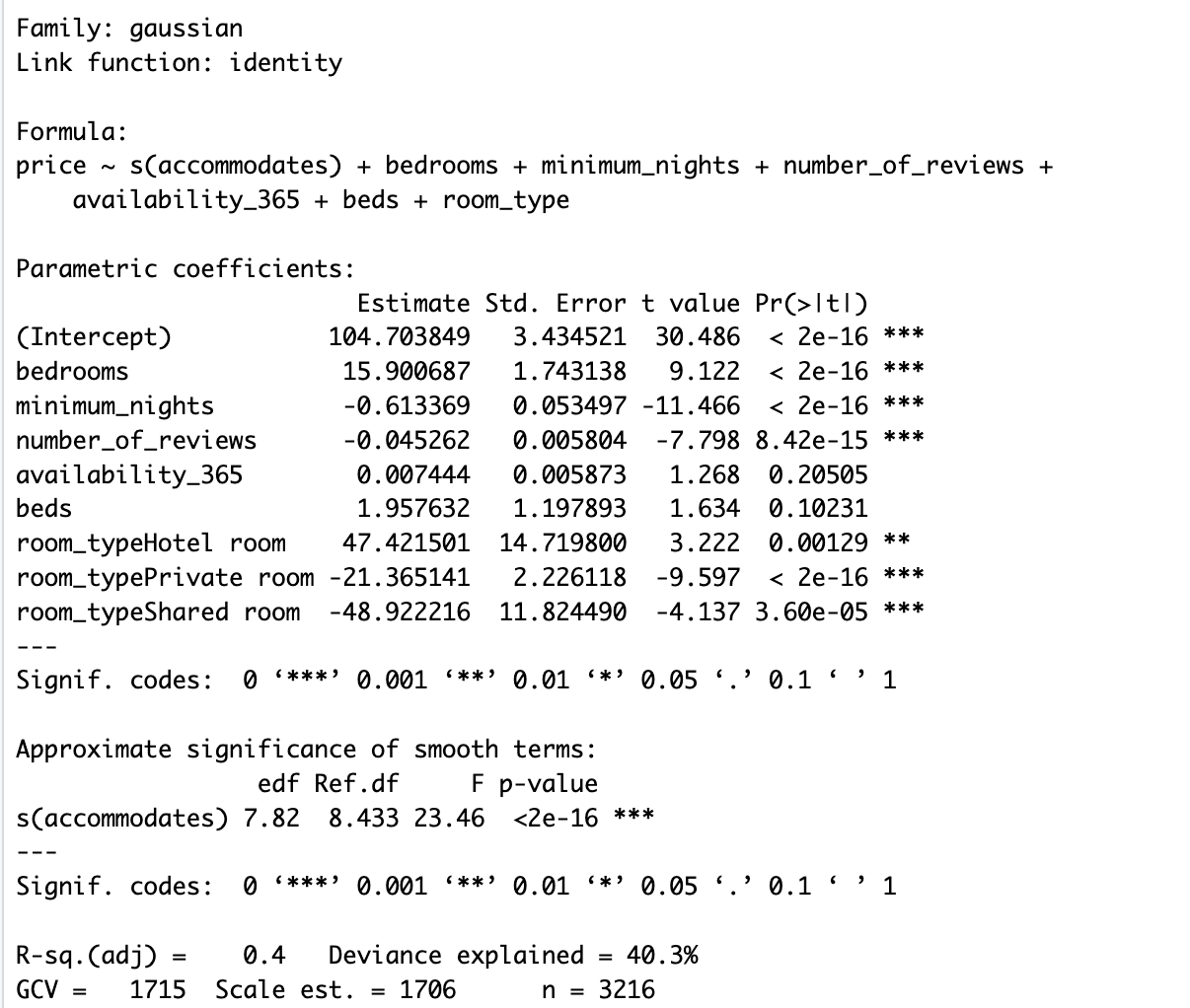
The accommodates variable had a coefficient of ~9.3. It infers that for every additional unit increase in the number of accommodates, the predicted price of listings increases by approximately $9.3, holding all variables constant.

Also, the predicted price of shared room is 59$ less than the predicted price of entire home/apartment type of rooms, holding all the variables constant.

**5.4 Generalized Additive Model:**

GAMs are like linear regression models, but instead of using a linear function to model the relationship between the dependent variable and independent variables, GAMs use non-parametric smoothing functions, such as splines or smoothing splines. The smoothing function allows for more flexible modeling of the relationship between the dependent variable and independent variables, which can improve the model's ability to capture non-linear relationships.

**We have implemented model 3 attributes using GAM’s :**



The model's goodness of fit is reported, with an adjusted R-squared of 0.416 and a deviance explained of 42.1%. The Generalized Cross Validation (GCV) and the estimated scale of the model are also shown, along with the number of observations used to fit the model.

**Findings:** Private rooms and shared rooms have negative coefficients compared to the reference category of complete home/apartment, whereas hotel rooms have a positive coefficient. Smooth parameters are also major price predictors, with accommodates, minimum nights, and quantity of reviews having a positive influence on price, while beds had no effect.

Overall, our model reveals that the number of bedrooms, kind of accommodation, accommodates, minimum nights, and number of reviews are significant pricing determinants for GAM listings in GAM’s.

**5.4.1 semi parametric:**

Generally, semi parametric model is a type of regression model that combine both parametric and non-parametric approaches to model the relationship between the dependent variable and the independent variables.

A screenshot of a computer

Description automatically generated with medium confidence

**Findings :** According to the model output, the number of bedrooms, minimum nights, and room type (particularly, whether the rental is a private room or a shared room) appear to be important predictors of rental pricing. The impact sizes for the number of reviews and availability are less but still statistically significant. The number of beds and the smoothed term for the number of accommodations do not appear to be important pricing predictors.

The model explains 40.3% of the variance in rental price, according to the adjusted R-squared value of 0.4. The model's prediction accuracy is measured by the Generalized Cross Validation (GCV) value of 1715. A lower GCV suggests higher predictability.

So now we are comparing the best model for all the methods we have used by considering the MSPE metric:

**Output shown as below:**

A screenshot of a computer code

Description automatically generated with medium confidence

**Findings** : The MSPE for the additive model is 1755.007, for the semiparametric model is 1808.013, for the linear regression model is 1750.517, and for the extended linear model is 1843.14 based on the supplied output.

When these numbers are compared, the linear regression model has the lowest MSPE of the four models, suggesting that it performs the best in terms of test set prediction accuracy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear regression | Generalized linear models | Additive model | Semi parametric model |
| AIC | 33035 | 33059 | 32527 | 32854 |
| BIC | 33126 | 33138 | 32872 | 32992 |
| Adjusted R2 | 0.4071 | 0.4028 | 0.5 | 0.44 |
| MSPE | 1781 | 1750 | 1540 | 1700 |

**Conclusion:**