Analyzing the Short-term Rental Market: An Advanced Statistical Analysis

Anila Reddy Musku
Master's in data science
University of Colorado, Boulder

Mohammed Junaid Shaik
Master's in Data science
University of Colorado, Boulder

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 GAMs 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

2.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 you are traveling in large groups or with your 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.

In this project these are the questions we would like to answer:

Questions of interest:

1.How does the average prices of listings vary in different neighborhoods or cities?

2.What factors determine the price of a listing?

3.What is the relationship between the price of a listing and its location, amenities, and other factor?

4.Which statistical learning tools works best for predicting the price?

5.Can we predict the price with properties like neighborhood, amenities, accommodates and more?

These questions all relate to understanding the short-term rental market in Portland city and identifying the factors that influence pricing and availability.

The first question aims to explore how pricing varies across different neighborhoods or cities within Portland. This information can be useful for renters who are looking for the best value for their money, as well as for property owners who are trying to set competitive prices. Understanding these factors can help property owners to optimize their pricing and marketing strategies. And we can explore the relationship between price and location, amenities, and other factors that are valued by renters. This could help property owners to identify which features are most important to renters and adjust their pricing and marketing strategies accordingly.

Overall, answering these questions can help both renters and property owners to make more informed decisions in the short-term rental market.

3.Data

Source of Data:

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

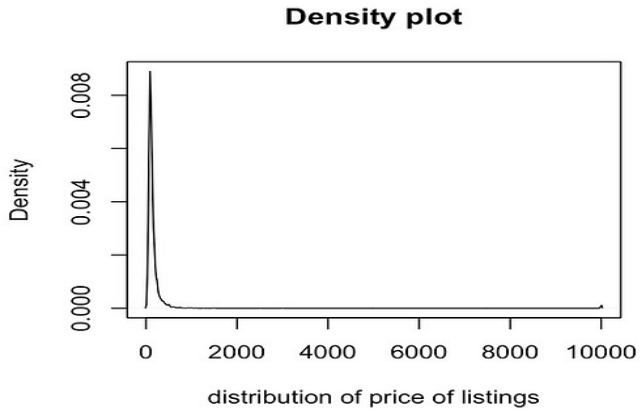
This data is completely observational. The data on the website is collected through the observations of Airbnb users who list their properties and the guests who book them. since the data is collected because of natural interactions between hosts and guests, without any manipulation or interference from 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 number of nights that listing could be book for, number of reviews and total number 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.

3.1 Data Cleaning:

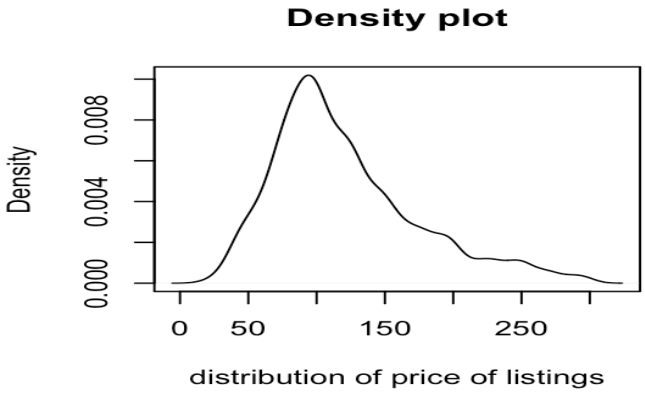
when comes to data cleaning, there are a lot of unnecessary variables and outliers in our dataset which are not useful for analyzing and predicting the prices and removing the features which do not impact the price. Some of the features that were removed are scraped id, URL of listing, host URL, location 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.



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 many listings having price more than 300 $ per night.



Before cleaning:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
id	listing_url	scrape_id	last_scraped	source	name	description	neighborhoo	picture_url	host_id	host_url	host_name	host_since	host_locatio	host_about	host_respon	host_respon	host_accept	host_is_supe	host_thumbr	host_pict
12899	https://www	2.0221E+13	9/16/22	city scrape	Alberta Arts	Please know	We're within	https://a0.m	49682	https://www	Ali And Davi	10/29/09	Portland, OR	We enjoy co	within an ho	100%	100%	t	https://a0.m	https://a
27746	https://www	2.0221E+13	9/16/22	city scrape	Private Suite	Suite has TV	Beautiful, vi	https://a0.m	119878	https://www	Julie	5/7/10	Portland, OR	Hello! I love	within an ho	90%	88%	t	https://a0.m	https://a
37676	https://www	2.0221E+13	9/16/22	city scrape	Mt. Hood Vie	This 1,000 Sf	The Pearl dis	https://a0.m	162158	https://www	Paul	7/9/10	Portland, OR	I look forwar	within a few	75%	87%	t	https://a0.m	https://a
61677	https://www	2.0221E+13	9/16/22	city scrape	Large Peace	My home is (	This is a quie	https://a0.m	298438	https://www	Deborah	11/25/10	Portland, OR	Hi!	within an ho	100%	100%	f	https://a0.m	https://a
61893	https://www	2.0221E+13	9/16/22	city scrape	Perfect Portl	City ¸Ãî Hike	Restaurants,	https://a0.m	300391	https://www	Matt	11/26/10	United State	I travel all o	within an ho	100%	75%	t	https://a0.m	https://a
65067	https://www	2.0221E+13	9/16/22	city scrape	Fully Equipp	Very nice ful	Neighborhoo	https://a0.m	119878	https://www	Julie	5/7/10	Portland, OR	Hello! I love	within an ho	90%	88%	t	https://a0.m	https://a
67036	https://www	2.0221E+13	9/16/22	city scrape	Historic Hom	Welcome to	Sullivan's Gu	https://a0.m	329777	https://www	Will	12/29/10	Portland, OR	I have a	within an ho	100%	93%	f	https://a0.m	https://a
77522	https://www	2.0221E+13	9/16/22	city scrape	Charming Ea	You'll love st	Our neighbo	https://a0.m	345461	https://www	Diane	1/13/11	Portland, OR	We live in Pc	within an ho	100%	100%	f	https://a0.m	https://a
80357	https://www	2.0221E+13	9/16/22	city scrape	Free Standin	Contemporai	Sullivan's Gu	https://a0.m	415758	https://www	Esther	3/1/11	Portland, OR	A native	within an ho	100%	70%	t	https://a0.m	https://a
93613	https://www	2.0221E+13	9/16/22	city scrape	Sunny Queer	<b>The space<	/b> T	https://a0.m	501715	https://www	Teresa	4/11/11	Portland, OR	Welcome to	within a few	97%	56%	f	https://a0.m	https://a
99355	https://www	2.0221E+13	9/16/22	city scrape	Beautiful Ap	A beautiful 1	Irvington is t	https://a0.m	523311	https://www	Mark	4/20/11	Phoenix, AZ	I was born	within an ho	100%	100%	t	https://a0.m	https://a
114086	https://www	2.0221E+13	9/16/22	city scrape	Bright Frienc	<b>The space<	/b> T	https://a0.m	501715	https://www	Teresa	4/11/11	Portland, OR	Welcome to	within a few	97%	56%	f	https://a0.m	https://a
117969	https://www	2.0221E+13	9/16/22	city scrape	Beautiful Ne	Our carefully	Our neighbo	https://a0.m	589016	https://www	Steve	5/15/11	Portland, OR	I'm a marrie	within an ho	100%	100%	t	https://a0.m	https://a
143483	https://www	2.0221E+13	9/16/22	city scrape	Garden Hide	A beautiful 1	Irvington is c	https://a0.m	523311	https://www	Mark	4/20/11	Phoenix, AZ	I was born	within an ho	100%	100%	t	https://a0.m	https://a
145199	https://www	2.0221E+13	9/16/22	city scrape	Upstairs, stu	I can speak a	Very quite ar	https://a0.m	683139	https://www	Amy	6/9/11	Portland, OR	a very easy	within an ho	100%	100%	f	https://a0.m	https://a
182450	https://www	2.0221E+13	9/16/22	city scrape	Luminous, C	Quiet, comfc	My neighbor	https://a0.m	52505	https://www	Judy	11/8/09	Portland, OR	I'm Judy.	N/A	N/A		50%	f	https://a0.m	https://a
195632	https://www	2.0221E+13	9/16/22	city scrape	Foodie's Par	With eats tre	With eats tre	https://a0.m	659137	https://www	Anjali	6/3/11	Portland, OR	What a	within an ho	100%	100%	t	https://a0.m	https://a
199683	https://www	2.0221E+13	9/16/22	city scrape	Authentic Po	*Private que	*The Neighb	https://a0.m	975606	https://www	Leo	8/15/11	Portland, OR	Howdy Do!	within a few	100%	80%	t	https://a0.m	https://a
205616	https://www	2.0221E+13	9/16/22	city scrape	Entire Home	This Magnifi	Beautiful saf	https://a0.m	1011157	https://www	Jody	8/23/11	Portland, OR	I am a retire	within an ho	92%	93%	f	https://a0.m	https://a
213228	https://www	2.0221E+13	9/16/22	city scrape	Sweet 2 Roo	Hello travele	The house hc	https://a0.m	1099616	https://www	Abraham	9/2/11	Portland, OR	I am a smal	within an ho	100%	98%	t	https://a0.m	https://a
217607	https://www	2.02215+13	9/15/22	city scrape	Newly Avail	THIS IS A RE	CENTRAL CE	https://a0.m	971012	https://www	Cynthia	8/14/11	Los Angeles	art director	N/A	N/A		33%	f	https://a0.m	https://a

After cleaning:

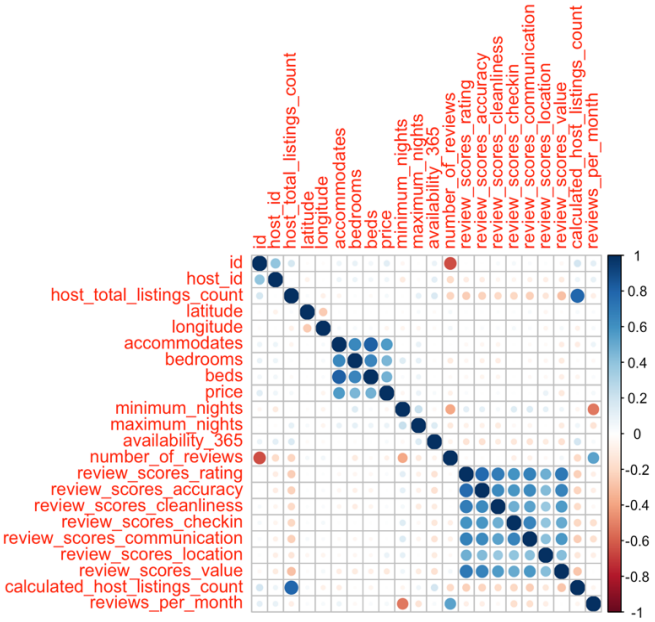
id	host_id	host_is_supe	host_total_li	neighbourho	latitude	longitude	property_typ	room_type	accommodat	bathrooms_t	bedrooms	beds	amenities	price	minimum_ni	maximum_n	availability_	availability_	availability_	availability_
12899	49682	t	1	Concordia	45.56488	-122.63418	Entire rental	Entire home/	3	1 bath	2	2	["Refrigerato	$85.00	3	730	0	0	0	0
27746	119878	t	6	Southwest H	45.49621	-122.74745	Entire home	Entire home/	2	1 bath	1	3	["Shower ge	$103.00	2	7	29	59	85	36
37676	162158	t	3	Pearl	45.52564	-122.68273	Entire loft	Entire home/	3	1 bath	1	1	["Single leve	$140.00	30	730	0	0	11	10
61677	298438	f	2	Reed	45.48784	-122.62806	Private room	Private room	3	2 shared bat	1	2	["Shower ge	$55.00	3	60	8	35	61	14
61893	300391	t	1	Goose Hollo	45.52258	-122.69955	Entire condo	Entire home/	2	1 bath	1	1	["Free washe	$120.00	30	300	0	14	19	16
65067	119878	t	6	Southwest H	45.49764	-122.74696	Entire guest	Entire home/	4	1 bath	1	3	["Refrigerato	$127.00	2	30	0	0	19	1
67036	329777	f	5	Sullivan's Gu	45.53103	-122.64448	Entire guest	Entire home/	10	2 baths	4	5	["Free washe	$350.00	4	365	19	44	70	33
77522	345461	f	1	Eastmorelan	45.47086	-122.63273	Entire guest	Entire home/	2	1 bath	1	1	["Game cons	$110.00	2	365	2	32	62	29
80357	415758	t	3	Sullivan's Gu	45.53364	-122.63895	Entire bunga	Entire home/	2	1 bath	1	1	["Shower ge	$110.00	31	365	0	0	0	25
93613	501715	f	17	Sabin	45.55536	-122.65106	Private room	Private room	2	3 shared bat	1	1	["Refrigerato	$41.00	30	180	1	31	61	33
99355	523311	t	2	Irvington	45.53836	-122.65159	Entire rental	Entire home/	6	1.5 baths	3	3	["Refrigerato	$185.00	1	730	8	26	48	27
114086	501715	f	17	Sabin	45.55593	-122.65029	Private room	Private room	1	3 shared bat	1	1	["Refrigerato	$39.00	30	180	30	60	90	18
117969	589016	t	3	Richmond	45.51022	-122.63088	Entire rental	Entire home/	2	1 bath	1	1	["EV charger	$107.00	2	1125	6	11	36	10
143483	523311	t	2	Irvington	45.54005	-122.64988	Entire rental	Entire home/	4	1 bath	1	1	["Refrigerato	$110.00	1	730	12	40	66	31
145199	683139	f	5	Montavilla	45.52959	-122.59081	Private room	Private room	2	1 bath	1	2	["Refrigerato	$75.00	3	180	28	58	79	33
182450	52505	f	4	Richmond	45.50529	-122.63201	Entire home	Entire home/	4	1.5 baths	2	2	["Refrigerato	$77.00	4	29	0	4	10	1
195632	659137	t	3	Boise	45.55083	-122.67567	Entire home	Entire home/	4	1 bath	1	3	["Shower ge	$115.00	7	1125	0	26	56	14
199683	975606	t	5	Rose City Par	45.54458	-122.60223	Private room	Private room	1	1 shared bat	1	1	["Backyard",	$43.00	31	730	27	57	87	36
205616	1011157	f	11	Irvington	45.53689	-122.64977	Entire home	Entire home/	14	3.5 baths	6	6	["Refrigerato	$702.00	2	1125	8	37	67	24
213228	1099616	t	2	Overlook	45.56616	-122.67889	Private room	Private room	3	1 shared bat	2	2	["Shower ge	$53.00	3	14	19	46	70	7
217607	971012	f	4	Old Town/Ch	45.52653	-122.67513	Entire condo	Entire home/	2	1 bath	1	1	["Shower ge	$94.00	30	365	0	0	0	27
218708	6313800	t	11	Kerns	45.5236	-122.63458	Entire rental	Entire home/	2	1 bath	1	2	["Refrigerato	$85.00	2	365	20	47	77	7
222298	1153902	t	3	Overlook	45.55815	-122.69218	Private room	Private room	1	1 shared bat	1	1	["Free washe	$65.00	2	14	11	24	46	13
231732	1153902	t	3	Overlook	45.55629	-122.69042	Private room	Private room	1	1 shared bat	1	1	["Free washe	$65.00	30	93	1	17	37	30
231734	1153902	t	3	Overlook	45.55748	-122.69207	Private room	Private room	1	1 shared bat	1	1	["Free washe	$65.00	2	93	12	19	34	27
236108	501715	f	17	Sabin	45.55645	-122.64854	Private room	Private room	2	3 shared bat	1	1	["Free washe	$43.00	30	180	1	31	61	33
249526	1307164	t	3	Hosford-Abe	45.50806	-122.64337	Private room	Private room	1	1 shared bat	1	1	["Shower ge	$49.00	5	28	5	21	46	31
279579	1457743	t	1	Far Southwe	45.43065	-122.73028	Private room	Private room	1	2.5 baths	1	1	["Luggage dr	$51.00	6	1125	0	6	36	3
317797	828435	f	32	Mt. Tabor	45.51203	-122.61079	Entire condo	Entire home/	3	1 bath	1	1	["Refrigerato	$50.00	30	150	30	60	90	18
322986	119878	t	6	Southwest H	45.49728	-122.74555	Entire guest	Entire home/	8	2 baths	3	5	["Single leve	$275.00	2	10	0	0	19	10

Here after cleaning: We were able to delete numerous columns that were not essential to our study after cleaning the dataset, allowing us to focus exclusively on the key variables. Furthermore, we took care to eliminate any null or missing values from the dataset, guaranteeing that our research is based on complete and valid data. We utilized the mean value of the corresponding columns to fill in the missing values, which is a popular strategy for imputed missing data. Because of our efforts, we now have a fully cleaned and usable dataset that is ready for further analysis. With this high-quality data on hand, we can now begin to do advanced analytics and draw significant insights that can help in making decisions.

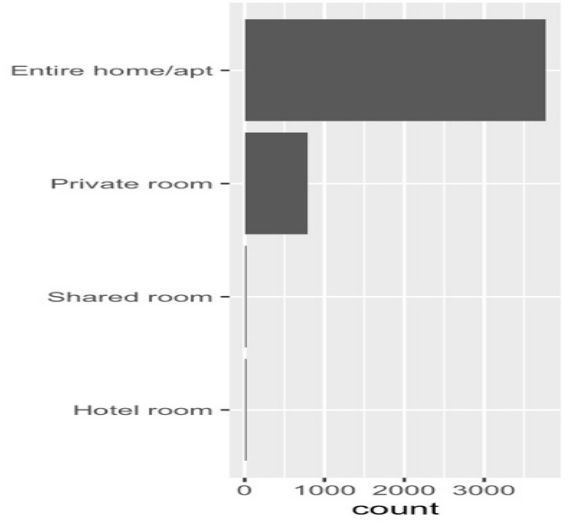
4.Exploratory Data Analysis (EDA) :

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.

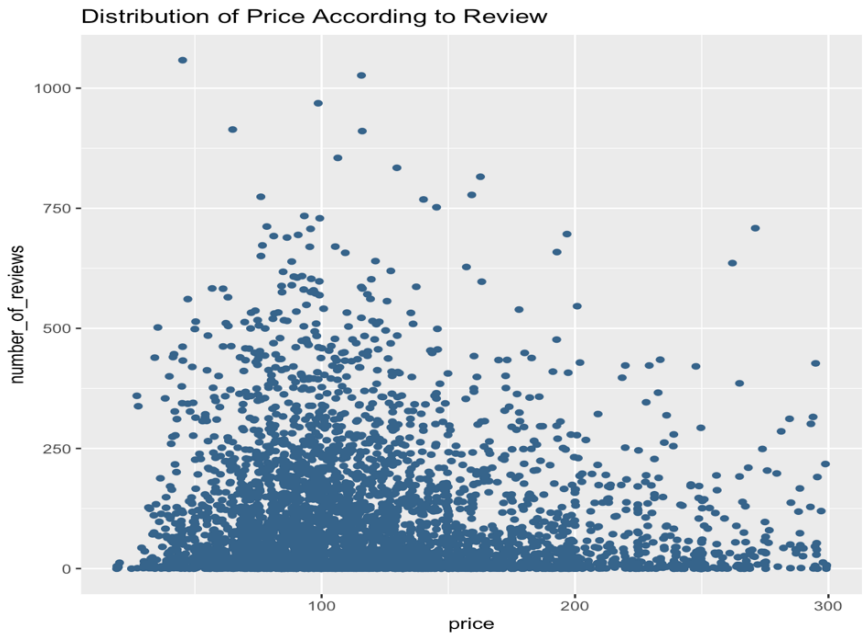
Correlation plot for all the attributes to find if any correlation occurs between the variables:



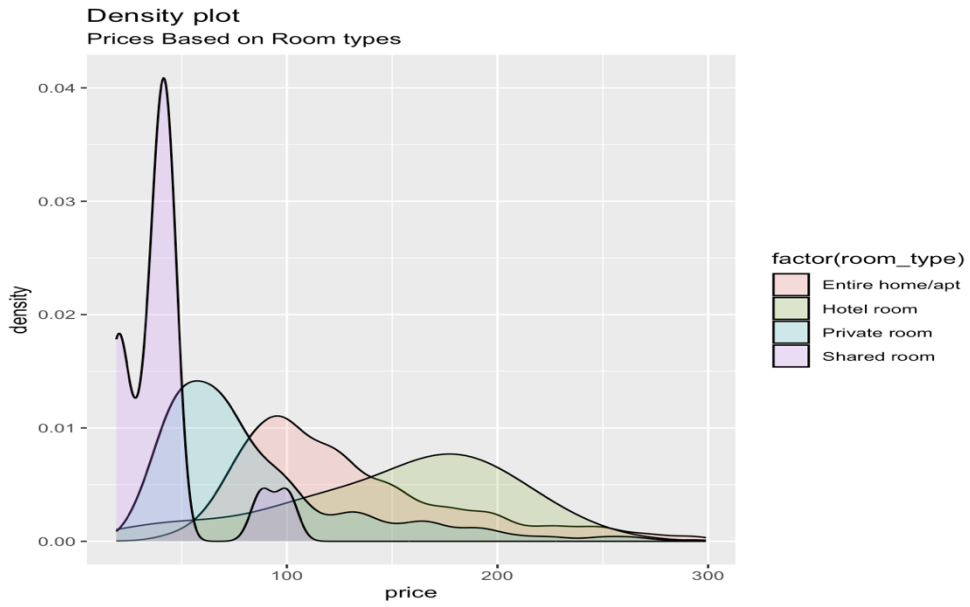
This correlation shows that the price is closely correlated with the number of beds, bedrooms, bathrooms, and accommodations. It is evident from this that prices rise as the number of rooms increases. Furthermore, the reviews are strongly co-related with each other.



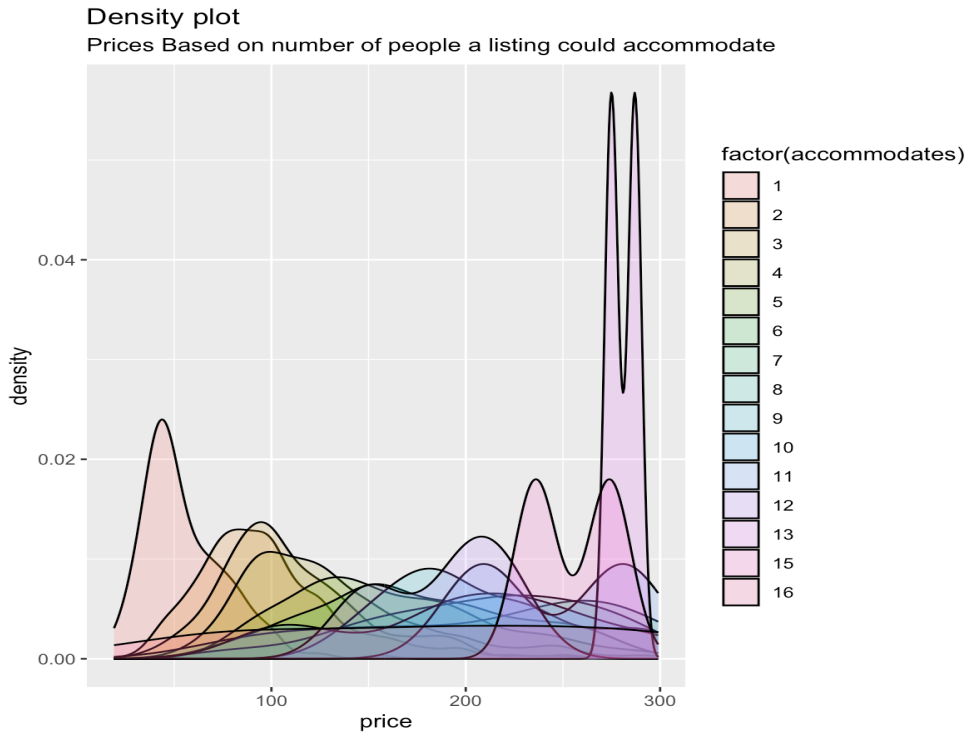
From this plot, we could see that a greater number of listings, have the room type of entire room, rather than shared rooms.



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



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.

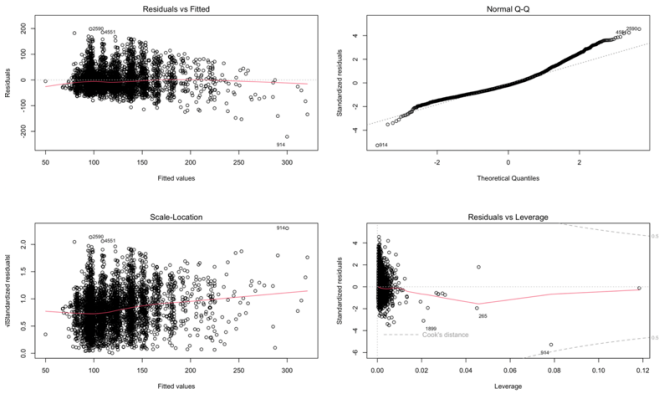


From this density plot of price v/s accommodate, we could infer that, listings which could accommodate a greater number of people are high priced.

5.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.

5.0 Regression So here we have implemented this regression without splitting the data:



Performing regression analysis on a dataset without splitting the data can be done, but it can lead to potential issues with overfitting and biased estimates of model performance.

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.

We performed an ANOVA (Analysis of Variance) test to compare the means of price for different room types, number of reviews, accommodates, minimum nights, beds, bedrooms, property in dataset. The ANOVA test will allow us to determine whether there are significant differences in the mean prices of the different groups.

Findings from this ANOVA test:

(i)price vs accommodates
(ii) price vs bedrooms
(iii) price vs beds

```
> check.aov<- aov(price~accommodates,listing)
> summary(check.aov)
              Df   Sum Sq  Mean Sq F value Pr(>F)
accommodates   1 10440435 10440435    3109 <2e-16 ***
Residuals   4493 15087532     3358
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> AIC(check.aov)
[1] 49255.63
> check.aov<- aov(price~bedrooms,listing)
> summary(check.aov)
              Df   Sum Sq Mean Sq F value Pr(>F)
bedrooms       1  9452862 9452862    2642 <2e-16 ***
Residuals   4493 16075105    3578
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> AIC(check.aov)
[1] 49540.62
> check.aov<- aov(price~beds,listing)
> summary(check.aov)
              Df   Sum Sq Mean Sq F value Pr(>F)
beds           1  7570638 7570638    1894 <2e-16 ***
Residuals   4493 17957329    3997
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> AIC(check.aov)
[1] 50038.34
```

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 price and accommodates.

Causality

Causality refers to a relationship in which one variable directly influences the other variable. To establish causality, it is necessary to conduct a carefully controlled experiment and our assumptions (which may true or false) that manipulates one variable while holding all other variables constant.
We can infer causality when there is a strong correlation between two variables, but it is not clear which variable causes the other.
As per the correlation plots, we see that accommodates more corelates the price, but we cannot say that accommodates causes price.
In case of observational studies as we cannot manipulate the variables of interest, here price to establish a cause-and-effect relationship, we cannot test for causality based on the data collected.

We can only observe correlations between variables. In our case, we first try to check for relationships between our predictor variables like bedrooms, accommodates, beds, room type, number of reviews and the response variable price using regression models. Using these regression models, we try to control for other factors that may affect the outcome. But, based on upon these analyses, we cannot establish causality.

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

```
Residual standard error: 41.53 on 3205 degrees of freedom
Multiple R-squared:  0.3948,    Adjusted R-squared:  0.393
F-statistic: 209.1 on 10 and 3205 DF,  p-value: < 2.2e-16
```

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

```
Residual standard error: 41.16 on 3202 degrees of freedom
Multiple R-squared:  0.4063,    Adjusted R-squared:  0.4039
F-statistic: 168.5 on 13 and 3202 DF,  p-value: < 2.2e-16
```

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

```
Residual standard error: 41.05 on 3202 degrees of freedom
Multiple R-squared:  0.4094,    Adjusted R-squared:  0.4071
F-statistic: 170.8 on 13 and 3202 DF,  p-value: < 2.2e-16
```

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.

```
> summary(lm_3)

Call:
lm(formula = price ~ latitude + host_total_listings_count + longitude +
    accommodates + room_type + bedrooms + minimum_nights + beds +
    number_of_reviews + calculated_host_listings_count + reviews_per_month,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-132.21  -26.07   -8.41   18.65  233.33

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -2.859e+03  1.779e+03  -1.607 0.108154
latitude                       -1.064e+01  2.014e+01  -0.528 0.597288
host_total_listings_count       6.269e-03  1.511e-03   4.149 3.43e-05 ***
longitude                      -2.786e+01  1.416e+01  -1.967 0.049236 *
accommodates                    1.026e+01  8.246e-01  12.442  < 2e-16 ***
room_typeHotel room             4.913e+01  1.466e+01   3.352 0.000813 ***
room_typePrivate room          -2.584e+01  2.101e+00 -12.299  < 2e-16 ***
room_typeShared room           -5.667e+01  1.137e+01  -4.982 6.61e-07 ***
bedrooms                        1.699e+01  1.648e+00  10.307  < 2e-16 ***
minimum_nights                 -7.142e-01  5.457e-02 -13.089  < 2e-16 ***
beds                            1.122e+00  1.153e+00   0.972 0.330894
number_of_reviews              -2.725e-02  6.961e-03  -3.914 9.25e-05 ***
calculated_host_listings_count  3.644e-01  1.267e-01   2.875 0.004065 **
reviews_per_month              -1.351e+00  4.007e-01  -3.373 0.000753 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.05 on 3202 degrees of freedom
Multiple R-squared:  0.4094,    Adjusted R-squared:  0.4071
F-statistic: 170.8 on 13 and 3202 DF,  p-value: < 2.2e-16

> AIC(lm_3)
[1] 33035.56
> BIC(lm_3)
[1] 33126.7
```

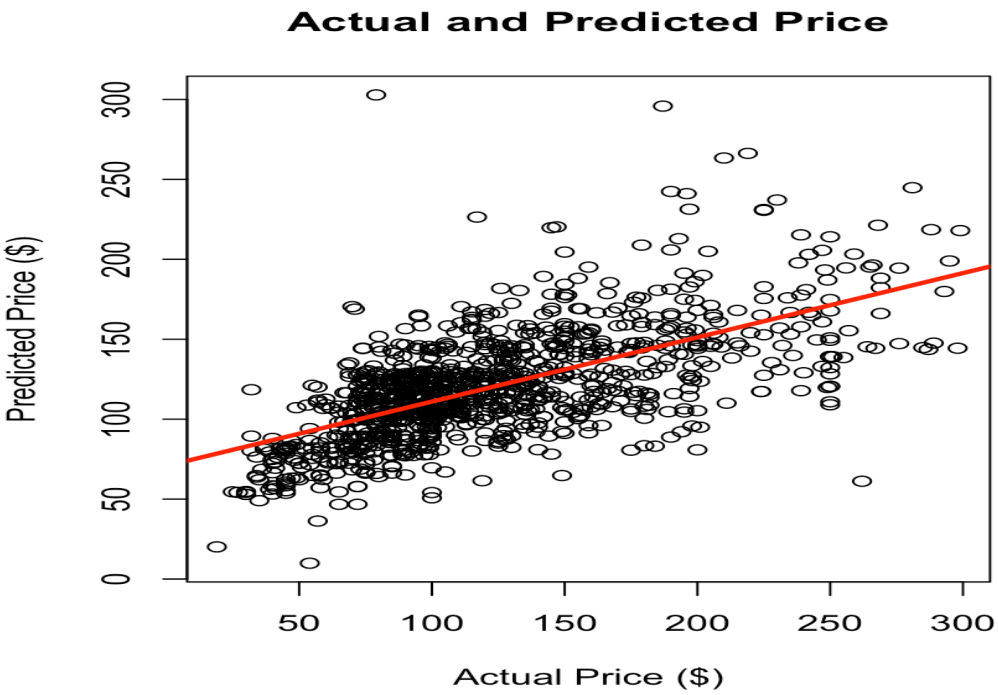
From the summary of model 3, we could say that “accommodates”, “Private room type”, “bedrooms”, “minimum nights”, “number_of_reviews”, and “host_total_listings_count” are the most important to the price. Some are negative correlated, and some are positive correlated ones. these variables are essential in determining the prices of a rental properties, and their impact on the price may vary depending on the specific circumstances and the local market conditions

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.

```
> head(lr_result, 20)
    Actual Predicted
3      140  97.18698
6      127 135.24259
15      75  80.59409
24      65  54.34557
25      65  74.52703
52      75  93.40033
56     119 126.43242
60     100 113.63230
64      84  84.94718
65      85  67.94800
68     162 109.68816
69      55  67.18819
71     150 152.41695
73      40  74.07133
81      65  49.64046
82     130 139.57287
95      99  90.85823
96     145 157.49106
97      45  53.00492
103    137 109.87838
```

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



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 so we may 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.

```
MSPE is 1843.14
> RMSE <- sqrt(mean( (test$price - pred)**2 ))
> SSE <- sum((test$price - pred)**2)
> SSR <- sum((pred - mean(test$price)) ** 2)
> SST <- SSR +SSE
> R2 <- (SST - SSE) / SST
> RMSE
[1] 42.93179
> R2
[1] 0.388097
> AIC(glm_train1)
[1] 33110.05
> BIC(glm_train1)
[1] 33176.89
```

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.

```
MSPE is 1768.96
> RMSE <- sqrt(mean( (test$price - pred)**2 ))
> SSE <- sum((test$price - pred)**2)
> SSR <- sum((pred - mean(test$price)) ** 2)
> SST <- SSR +SSE
> R2 <- (SST - SSE) / SST
> RMSE
[1] 42.059
> R2
[1] 0.4003031
> AIC(glm_train2)
[1] 33059.24
> BIC(glm_train2)
[1] 33138.22
```

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.

```
MSPE is 1750.52
> RMSE <- sqrt(mean( (test$price - pred)**2 ))
> SSE <- sum((test$price - pred)**2)
> SSR <- sum((pred - mean(test$price)) ** 2)
> SST <- SSR +SSE
> R2 <- (SST - SSE) / SST
> RMSE
[1] 41.83919
> R2
[1] 0.4028947
> AIC(glm_train3)
[1] 33059.46
> BIC(glm_train3)
[1] 33138.45
```

All three models have the similar(approximate) MSPE, RMSE, and R-squared values, which means that they are equally accurate in predicting the prices of the test dataset.

However, when comparing AIC and BIC values, model 3 has the lowest AIC and BIC values among the three models. This indicates that model 3 may be the best model for this dataset, as it has the lowest complexity while still maintaining good predictive accuracy.

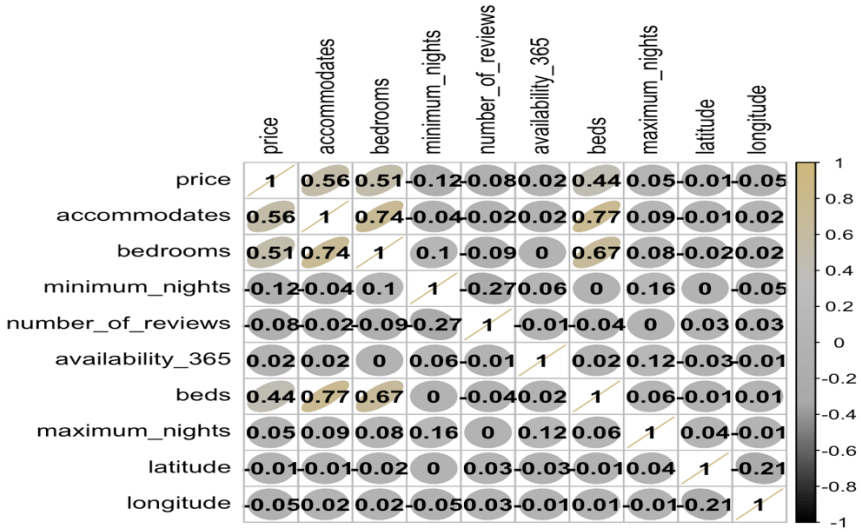
Therefore, based on the AIC and BIC values, model 3 may be the preferred model.

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.

```
> vif_mspe
                                GVIF Df GVIF^(1/(2*Df))
latitude                    1.055843  1        1.027542
host_total_listings_count   1.825402  1        1.351075
longitude                   1.066186  1        1.032563
accommodates                3.599773  1        1.897307
room_type                   1.464998  3        1.065711
bedrooms                    2.424162  1        1.556972
minimum_nights              1.210086  1        1.100039
beds                        2.982792  1        1.727076
number_of_reviews           1.625759  1        1.275052
calculated_host_listings_count 1.845342  1        1.358434
reviews_per_month           1.709878  1        1.307623
> vif_adjr2
                                GVIF Df GVIF^(1/(2*Df))
latitude                    1.055843  1        1.027542
host_total_listings_count   1.825402  1        1.351075
longitude                   1.066186  1        1.032563
accommodates                3.599773  1        1.897307
room_type                   1.464998  3        1.065711
bedrooms                    2.424162  1        1.556972
minimum_nights              1.210086  1        1.100039
beds                        2.982792  1        1.727076
number_of_reviews           1.625759  1        1.275052
calculated_host_listings_count 1.845342  1        1.358434
reviews_per_month           1.709878  1        1.307623
```

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.



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.

```
15 x 1 sparse Matrix of class "dgCMatrix"
                                        s0
(Intercept)                   -2.655117e+03
(Intercept)                    .
latitude                      -9.421933e+00
host_total_listings_count      5.840891e-03
longitude                     -2.574474e+01
accommodates                   9.317536e+00
room_typeHotel room            4.506007e+01
room_typePrivate room         -2.474746e+01
room_typeShared room          -5.910065e+01
bedrooms                       1.643148e+01
minimum_nights                -6.595453e-01
beds                           2.446210e+00
number_of_reviews             -2.614973e-02
calculated_host_listings_count 3.614515e-01
reviews_per_month             -1.164566e+00
```

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.

5.7 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 performed GAMs on model 3 predictors:

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =    0.5   Deviance explained = 50.9%
GCV = 1445.4  Scale est. = 1420.4    n = 3216
```

Findings:

The additive non-parametric model has an adjusted R-squared value of 0.5, indicating that 50.9% of the variation in the dependent variable is explained by the independent variables. The GCV value for this model is 1445.4, and the scale estimate is 1420.4. These results suggest that the non-parametric model is better at explaining the variation in the dependent variable than the semi-parametric model. Additionally, the GCV value for the non-parametric model is lower, indicating that this model is less prone to overfitting.

5.7.1 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.

```
R-sq.(adj) =  0.441   Deviance explained = 44.4%
GCV = 1599.7  Scale est. = 1589      n = 3216
```

Semi-parametric model has an adjusted R-squared value of 0.441, indicating that 44.4% of the variation in the dependent variable is explained by the independent variables. The general cross-

validation (GCV) value for this model is 1599.7, and the scale estimate is 1589. These results suggest that while the additive semi -parametric model can explain a significant proportion of the variation in the dependent variable, the GCV value and scale estimate indicate that the model may be overfitting to the data.

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.

6. Evaluation:

In this study, we performed statistical models to analyze the relationship between housing prices and neighborhood characteristics in the city of Portland. Our results indicate that the model was able to effectively predict housing prices based on the selected variables, with a high degree of accuracy. We found that the most significant predictors for housing prices by performing linear regression models, GLM's and GAM's. But ultimately, we need to consider the best final model for this project.
we conducted a linear regression analysis to investigate the relationship between the independent variables (X) and the dependent variable (Y). Our results indicate that both Model 2 and Model 3 have higher adjusted R-squared values than Model 1, suggesting that these models explain more of the variation in the dependent variable "price".

```
lm_3 <- lm(price ~ latitude  + host_total_listings_count+ longitude +
            accommodates +room_type + bedrooms + minimum_nights +  beds +  number_of_reviews + calculated_host_listings_count + reviews_per_month, data = train)
```

Comparison within the linear regression models:

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

Based on our findings, we can conclude that the independent variables in Model 3 & Model 2 are better predictors of the dependent variable "price" than those in Model 1.but Model 3 is relatively better so we are considering this model 3 in linear regression.

Transformed: Transforming the response variable(skewed) to normal response variable in a linear regression model using a logarithmic transformation can help to improve the model's performance and accuracy in certain situations but requires careful interpretation of the results.
In our case, we had no change in the prediction part, but we have significant change in AIC & BIC values.

```
> AIC(lm_3_sqrt)
[1] 12837.8
> BIC(lm_3_sqrt)
[1] 12928.93
```

We performed GLM's, GAMs on the same predictors on which we performed other statistical modelling. In GLM's our response variable distribution is continuous and non-normal (skewed) so we performed according to the assigned family distribution. In GLM's we got better metrics values predictors for model 3.

Comparison within the GLM's models:

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

Based on the findings presented, model3 appears to be the best model among the three, with the highest adjusted R-squared value and the lowest MSPE value.

MSPE:

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

```
The MSPE for the additive(non-parametric) model is 1540.154 .
> cat ("The MSPE for the semiparametric model is", semiparametric_mspe, ".")
The MSPE for the semiparametric model is 1700.685 .
> cat ("The MSPE for the linear regression model is", lm_mspe, ".")
The MSPE for the linear regression model is 1724.192 .
> cat("The MSPE for Generalized Linear Model is", round(mspe, 2), "\n")
The MSPE for Generalized Linear Model is 1750.52
```

Findings:

When these numbers are compared, the additive non-parametric has the lowest MSPE of the four models, suggesting that it performs the best in terms of test set prediction accuracy.

we found that the most important features that affect the price are accommodates, beds, bedrooms. Also, the other features which have some effect on price are 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. Also, by performing ridge regression we found out 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.

Comparison:

	Linear regression	Generalized linear models	Additive model	Semi parametric model
Adjusted R2	0.4071	0.4028	0.5	0.44
MSPE	1781	1750	1540	1700

This above table shows the results of four different statistical models: linear regression, generalized linear models, additive models, and semi-parametric models. The models are compared using four different metrics: adjusted R-squared and mean squared prediction error (MSPE).
The additive model looks to have the greatest adjusted R-squared value, and the lowest MSPE value. As a result, the additive model is most likely the best of the four models described in terms of overall fit and forecast accuracy. When compared to the other models, this model provides a better explanation for the variance in the dependent variable and more accurate predictions. So now we are considering our final model as Generalized Additive model which indicates that accommodates, bedrooms, number of reviews, reviews per month, minimum nights, latitude, longitude, beds, room type, calculated host listings count are significant predictors to predict the pricing for short-term rental marketing.

7. Conclusion:

The most significant predictors of price are the type of room (hotel room, private room, shared room) and the number of bedrooms, with hotel rooms having the highest impact on price and shared rooms having the lowest impact.

- Overall, the model suggests that location, number of people accommodate, and the type of room are the most critical factors influencing the price of Airbnb listings, while other features such as minimum nights and reviews per month also play a significant role.

- The number of reviews, availability_365, beds, maximum nights does not have much of an impact in its price.
- The greater the price of a home, the more bedrooms it contains.

- Neighborhood & cities doesn't impact the price in our data set.

- Best statistical model for our dataset is generalized additive model which gives the better values for adj R2 & MSPE when compared to other statistical models.

- Future research could expand the model to include additional variables related to the housing stock or the built environment, as well as explore different statistical techniques to analyze the data. Despite these limitations, our findings suggest that accommodation, location & room type characteristics play a significant role in determining housing prices in Portland and can provide valuable insights for policymakers and real estate professionals in the city.

8. References:

[1] https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-022-00349-3

[2] https://scholars.unh.edu/cgi/viewcontent.cgi?article=1511&context=honors

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8561316/

[4] https://nycdatascience.com/blog/r/data-analysis-on-airbnb-nyc-market/

[5] https://rpubs.com/spetrov1987/1030070

[6] https://rpubs.com/dnguyen8/airbnb_price_prediction

9. Appendix:

```
#Multiple Linear regression
lm_1 <- lm(price ~ accommodates + maximum_nights+ bedrooms +room_type+ minimum_nights + number_of_reviews + availability_365 + beds, data = train)
summary(lm_1)
AIC(lm_1)
BIC(lm_1)

#maximim nights, availability_365
lm_2 <- lm(price ~ latitude  + longitude +accommodates + bedrooms + room_type+ minimum_nights + availability_365 + beds + number_of_reviews +calculated_host_listings_count+reviews_per_month , data = train)
summary(lm_2)
AIC(lm_2)
BIC(lm_2)

#host_total_listings_count
lm_3 <- lm(price ~ latitude  + host_total_listings_count+ longitude + accommodates +room_type + bedrooms + minimum_nights +  beds +  number_of_reviews + calculated_host_listings_count + reviews_per_month, data = train)
summary(lm_3)
AIC(lm_3)
BIC(lm_3)


#prediction on test set
prediction<-predict(lm_3, newdata = test)
prediction<- exp(prediction)
mse = mean(lm_3$residuals^2)
mspe <- mean((test$price - prediction)^2)
mspe
AIC(lm_3)
```

```
pred <- predict(lm_3, newdata = test)
pred <- pmax(pred, 0)
RMSE <- sqrt(mean( (test$price - pred)**2 ))
SSE <- sum((test$price - pred)**2)
SSR <- sum((pred - mean(test$price)) ** 2)
SST <- SSR +SSE
R2 <- (SST - SSE) / SST

cat("SST: ", SST, "   SSE: ", SSE, "   SSR: ", SSR, "\nR2: ", R2, "   RMSE: ", RMSE)

actual <- test$price

lr_result <- data.frame(
  "Actual" = actual,
  "Predicted" = pred
)

head(lr_result, 20)
lm_line = lm(Predicted ~ Actual, data = lr_result)
plot(x = lr_result$Actual, y = lr_result$Predicted,
    main = "Actual and Predicted Price",
    xlab = "Actual Price ($)",
    ylab = "Predicted Price ($)")
abline(lm_line, col="red", lwd=3)
mspe <- mean((test_data$y - y_pred)^2)

#Generalized linear models

#lm_1
glm_train1 <- glm(price ~ accommodates + bedrooms + minimum_nights + number_of_reviews + availability_365 + beds + room_type,data = train, family = "gaussian")
summary(glm_train1)
pred = predict(glm_train1, newdata = test, type = "response")
mspe = mean((test$price - pred)^2)
cat("MSPE is", round(mspe, 2), "\n")
RMSE <- sqrt(mean( (test$price - pred)**2 ))
SSE <- sum((test$price - pred)**2)
SSR <- sum((pred - mean(test$price)) ** 2)
SST <- SSR +SSE
R2 <- (SST - SSE) / SST
RMSE
R2
AIC(glm_train1)
BIC(glm_train1)

#lm2
glm_train2 <- glm(price ~ accommodates + bedrooms + minimum_nights + number_of_reviews + availability_365 + beds + room_type +calculated_host_listings_count+ maximum_nights, data = train, family = "gaussian")
summary(glm_train2)
```

```
pred = predict(glm_train2, newdata = test, type = "response")
mspe = mean((test$price - pred)^2)
cat("MSPE is", round(mspe, 2), "\n")
RMSE <- sqrt(mean( (test$price - pred)**2 ))
SSE <- sum((test$price - pred)**2)
SSR <- sum((pred - mean(test$price)) ** 2)
SST <- SSR +SSE
R2 <- (SST - SSE) / SST
RMSE
R2
AIC(glm_train2)
BIC(glm_train2)

#lm_3
glm_train3<- glm(price ~ latitude + longitude + accommodates + room_type+ bedrooms + minimum_nights + beds + number_of_reviews + calculated_host_listings_count, data = train, family = "gaussian")
summary(glm_train3)
pred = predict(glm_train3, newdata = test, type = "response")
mspe = mean((test$price - pred)^2)
cat("MSPE is", round(mspe, 2), "\n")
RMSE <- sqrt(mean( (test$price - pred)**2 ))
SSE <- sum((test$price - pred)**2)
SSR <- sum((pred - mean(test$price)) ** 2)
SST <- SSR +SSE
R2 <- (SST - SSE) / SST
RMSE
R2
AIC(glm_train3)
BIC(glm_train3)

#Generealized Additive models
library(ggplot2)
library(mgcv)
gam_1 <- gam(price ~ s(latitude) + s(host_total_listings_count) + s(longitude) + s(accommodates) + room_type + bedrooms + s(minimum_nights) + s(beds) + s(number_of_reviews) + s(calculated_host_listings_count) + s(reviews_per_month), data = train)
summary(gam_1)
plot(gam_1, pages = 1)

#semiparametric model
library(ggplot2)
library(mgcv)
gam_2 <- gam(price ~ s(latitude) + host_total_listings_count+ longitude + accommodates +room_type + bedrooms + minimum_nights + beds + number_of_reviews + calculated_host_listings_count + reviews_per_month, data = train)
summary(gam_2)
plot(gam_2, pages = 1)
```