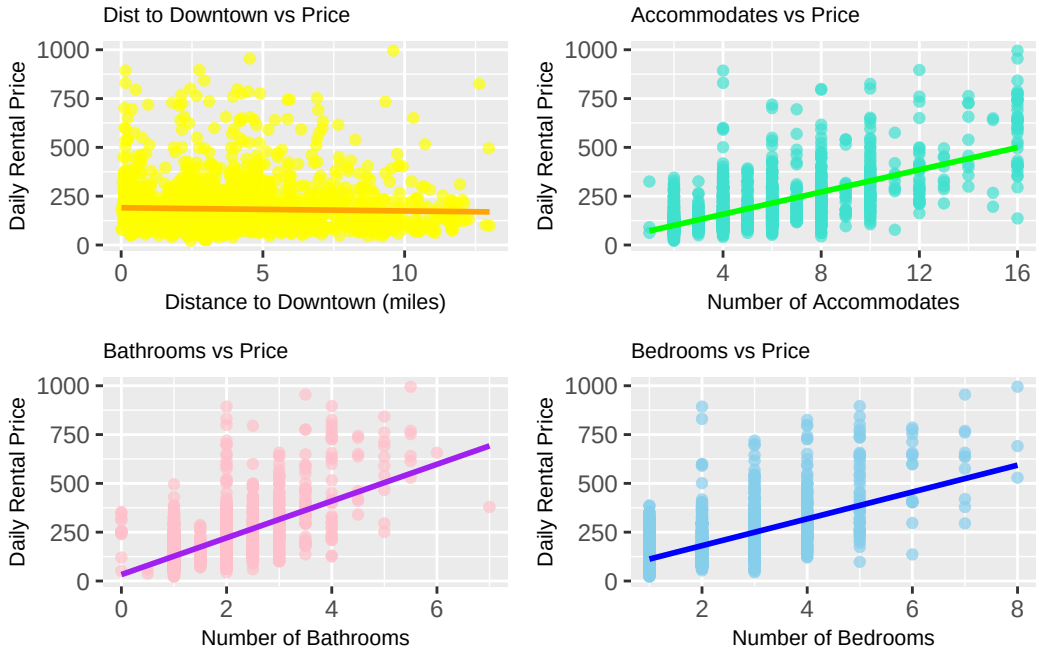
Airbnb Price Regression Modeling - Report for Airbnb executives

Introduction

The objective of this project is to generate a model that allows setting prices for Airbnb ads in Asheville, North Carolina. To achieve this goal, we have a database of various Airbnb rental listings in the city, which contains detailed information about the listings, such as price, number of rooms, amenities, number of bathrooms, property location, etc. This database was extracted in June 2023 by the company Inside Airbnb and contains information about 3,239 rental listings in Asheville, NC. Therefore, the conclusions drawn are based on this time period.

To develop a model that enables us to set prices for Airbnb ads in Asheville, we will use a linear regression approach. Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. In this context, the dependent variable would be the price of Airbnb ads, while independent variables may include features such as the number of rooms, amenities, number of bathrooms, and property location, among others.

In the following scatter plots, we observe positive linear relationships between '*Accommodates*', '*Bathrooms*' and '*Bedrooms*' and the dependent variable '*Price*'. In contrast, '*Dist to Downtown*' presents a negative relationship, indicating that homes closer to downtown tend to be more expensive. These patterns illustrate the goal of a regression model: to identify a line that minimizes the sum of squared errors between predicted and actual values



Methods

After data cleaning, variable consideration, and iterative model refinement, we developed a linear regression model for Airbnb pricing. Missing data were addressed through decision rules, resulting in a final database of 2,745 rows from the original 3,239.

Results

The final model obtained is the following:

ln(price) = 3.743 + 0.129(is entire home apt) + 0.082*(n° bedrooms) + 0.142*(n° bathrooms) - 0.048*(dist to downtown) + 0.192*(has entertainment amenities) + 0.443*(has climate control) + 0.07*(n°accommodates)*

The coefficient accompanying a variable in a linear regression is the amount by which the response variable (dependent variable) changes for each one-unit change in the predictor variable (independent variable), while keeping all other variables constant. For example, one would expect that, on average, for each additional room, the logarithm of the price would increase by 0.082.

If we were in the situation of defining the price of a new property, we should evaluate the variables of that property in the model. For example, let's assume we have a rental in Asheville, which is an entire home apartment (coded as 1), has 3 bedrooms, 1 bathroom, is located 3 miles from downtown, lacks a TV, Netflix, or any entertainment amenities (codes as 0), has air conditioning (coded as 1), and can accommodate 5 guests. Finally, we could predict a price as follows:

ln(price) = 3.743 + 0.129(1) + 0.082*(3) + 0.142*(1) - 0.048*(3) + 0.192*(0) + 0.443*(1) + 0.07*(5) = 4.909*

price = e^4.909 = 135.5

So, a suggested price for this rental is $136 per night

A measure obtained from our model is an R^2 of 0.52, which means that approximately 52% of the variability in the rental prices in Assville can be explained by the variables included in our model.

Conclusion

With the database, we were able to create a model to determine the rental price in Assville, as seen in the previous example. However, it is important to consider that other factors may be explaining the rental price, and here we are bound by our original database. With our model, we were only able to explain 52% of the variance, However, it can still be useful as a guide if our goal is to determine the listing price for a new rental on Airbnb.

Airbnb Price Regression Modeling - Report for Data Science Team

Introduction

The objective of this project is to generate a linear regression model that allows setting prices for Airbnb ads in Asheville, North Carolina.

Dataset

We have a database of various Airbnb rental listings in the city, obtained through web scraping, which contains detailed information about the listing. The original database includes **75 variables**. This dataset was extracted in June 2023 by the company Inside Airbnb and contains information about **3239** rental listings in Asheville, North Carolina. Therefore, the conclusions drawn are based on this time period. The database and dictionary can be downloaded from listings.csv and Inside Airbnb Data Dictionary

Data cleaning

First, the database was cleaned to be used for a linear regression model. Variables that did not contribute information to the model were excluded, such as: *"id", "listing_url", "scrape_id", "last_scraped", "source", "name", "description", "neighborhood_overview", "picture_url," host_id",* etc.

Secondly, certain transformations were performed on some variables to handle them within the model, such as *"host_response_time", "host_response_rate", "host_acceptance_rate", "host_is_superhost", "neighbourhood", "latitude", "longitude" ,"property_type", "room_type", "bathrooms_text", "amenities"* and *"price".*

It is worth mentioning that the ***'amenities'*** variable considered combinations of 2115 different features, so the following transformation was performed: new binary variables were generated to group the most frequent amenities. For example, ***'security'*** considers *'smoke and fire alarms', 'fire extinguisher', and 'first aid kit'.* ***'kitchen_amenities'*** *includes 'dishes and silverware', 'microwave', 'kitchen', 'refrigerator', 'cooking basics', 'coffee', 'freezer', 'wine glasses', 'oven', 'toaster', and 'dining table'."*

Some variables underwent treatment for missing values, for instance, in the 'beds' variable. For records with NA in the 'beds' field, we observed that the majority represent campsites; therefore, we will consider 'beds = 0' for these instances. In other cases, rows with missing values were simply removed. The final model, as a result, incorporated 2,745 out of the original 3,239 rows.

Methods

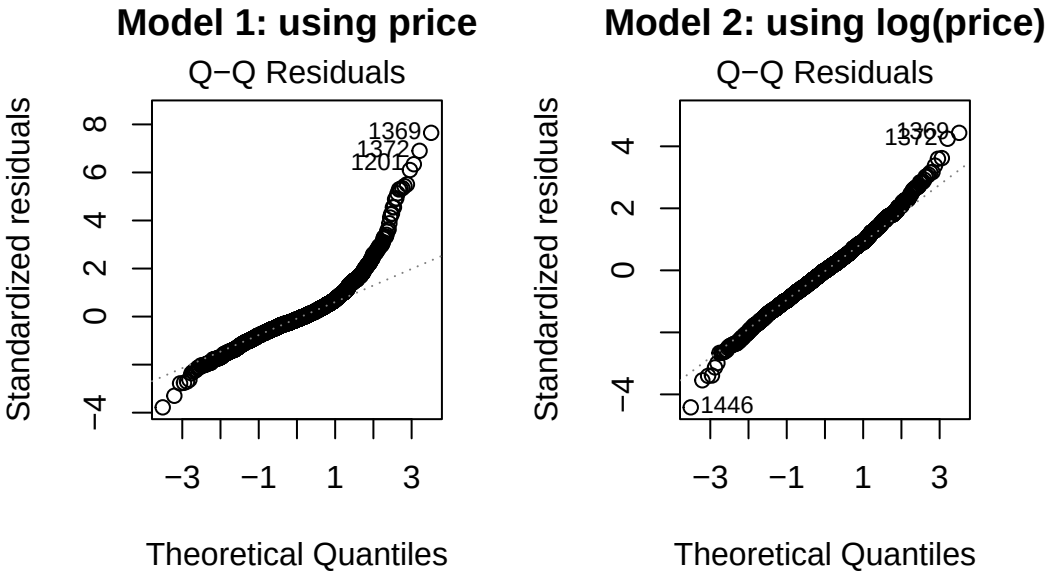
After data cleaning, various models were iteratively built based on different performance metrics. In particular, the following explanatory variables were considered as a foundation:

- Room type (which was transformed into the binary variable '**entire_home_apt**')
- Number of bedrooms.
- Distance to downtown.
- Amenities, which, as mentioned earlier, were categorized into the following binary variables: *("security", "essentials", "kitchen amenities", "entertainment", "bedroom amenities", "work", "parking", "toiletries", "climate control", "convenience features")*

Additionally, other variables were taken into consideration to include in the model. Factors appropriate for a new host to set a price were considered. In this context, and thinking about new hosts, variables such as *"review scores"* or *"number of reviews"* will not be considered, as we are focusing on a new listing that has not yet been evaluated by customers.

Results

	price_numeric		
Predictors	Estimates	CI	p
(Intercept)	21.30	-17.73 – 60.34	0.285
entire home aptTRUE	-31.24	-57.55 – -4.93	**0.020**
bedrooms	11.34	3.95 – 18.73	**0.003**
bathrooms numeric	39.56	31.98 – 47.13	**<0.001**
dist to dt	-9.38	-10.69 – -8.07	**<0.001**
entertainmentTRUE	12.08	-3.00 – 27.16	0.116
climate controlTRUE	45.36	14.45 – 76.27	**0.004**
accommodates	17.43	14.73 – 20.13	**<0.001**
Observations	2243		
R^2 / R^2 adjusted	0.532 / 0.530		


Model 1: using price
Q−Q Residuals
Standardized residuals
−4
0
2
4
6
8
1369
1372
1201
−3
−1
1
3
Theoretical Quantiles
Model 2: using log(price)
Q−Q Residuals
Standardized residuals
−4
0
2
4
1369
1372
1446
−3
−1
1
3
Theoretical Quantiles

- The binary variables *"security", "essentials", "bedroom amenities", "convenience features", "kitchen amenities"* and *"toiletries"* proved to be less significant in predicting the price, and thus, they were removed
- Price transformation using logarithm is performed, as the QQ plot exhibited a systematic deviation between the theoretical distribution and the actual distribution of the data. This could stem from both the non-normality of the residuals and heteroscedasticity.

After several iterations, the finally selected model was:

ln(price) = 3.743 + 0.129(is entire home apt) + 0.082*(n° bedrooms) + 0.142*(n° bathrooms) - 0.048*(dist to downtown) + 0.192*(has entertainment amenities) + 0.443*(has climate control) + 0.07*(n°accommodates)*

The details of the coefficients and statistics obtained with this model are as follows:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.74331908	0.090379291	41.417885	9.425626e-279
entire_home_aptTRUE	0.12854368	0.060909097	2.110418	3.493317e-02
bedrooms	0.08229573	0.017118045	4.807542	1.629618e-06
bathrooms_numeric	0.14243040	0.017542693	8.119073	7.676692e-16
dist_to_dt	-0.04762966	0.003023946	-15.750827	4.360741e-53
entertainmentTRUE	0.19240505	0.034916792	5.510388	3.992241e-08
climate_controlTRUE	0.44343276	0.071558455	6.196791	6.840710e-10
accommodates	0.06937661	0.006253087	11.094777	7.007055e-28

```
R-squared: 0.5240476
```

```
Mean Squared Error: 0.1494199
```

In general, the model explains approximately 47.28% of the variability in the logarithm of the price. The F-statistic and p-value suggest that the model is statistically significant. Additionally, the coefficients have small p-values, indicating that they are statistically significant.

The Mean Squared Error (MSE) of 0.1557079 suggests that, on average, the predictions of the regression model exhibit a relatively low mean squared error, indicating a reasonable level of accuracy. In the context of model evaluation, this MSE is considered relatively good, signifying that the model performs well in minimizing prediction errors.

If we calculate the Variance Inflation Factor (VIF) for our model, we obtain the following:

entire_home_apt	bedrooms	bathrooms_numeric	dist_to_dt
1.019634	5.926032	3.363982	1.080871
entertainment	climate_control	accommodates	
1.067202	1.043949	5.006517	

Most variables in the model exhibit VIF values near 1, indicating no major multicollinearity concerns. Although variables like "bedrooms" and "accommodates" have slightly higher VIF values (around 1.016), they are within an acceptable range. Overall, low VIF values enhance the stability and interpretability of the regression model by suggesting weak correlations between variables.

Conclusion

- Considering the task of helping new hosts set prices for Airbnb listings in Asheville, NC, falls more into the category of a **Prediction Problem.** The primary goal is to build a model that can accurately predict or generate prices for Airbnb listings based on various factors. We want to provide hosts with a tool that can predict the optimal price for their listings rather than drawing in-depth inferences about the underlying relationships between variables.
- The price is transformed using a logarithm due to a systematic deviation observed in the QQ plot between the theoretical distribution and the actual distribution of the data.
- The VIF values assess multicollinearity among predictor variables in a regression model. Values close to 1 indicate low correlation with other predictors, which is acceptable. Most variables in the model exhibit VIF values near 1, indicating no major multicollinearity concerns.
- A key metric for evaluating the model's performance is the Mean Squared Error (MSE). In summary, the model demonstrates a satisfactory level of accuracy, evidenced by a relatively low MSE. The successful minimization of prediction errors highlights commendable performance, aligning with expectations.
- Generally, we can consider this to be a valid model for predicting or determining the price of a home in Asheville, given its low multicollinearity, statistical significance of the model, significance of its variables, and the associated MSE (mean squared error).