Optimizing an Airbnb Listing

Laura Brooks 501146410

Supervised by: Ceni Babaoglu

Submitted: July 17, 2023

Ryerson University

Table of Contents

Optimizing an Airbnb Listing
Laura Brooks 501146410

Abstract

Airbnb is an online platform that allows users to book lodging rentals and/or experiences in any destination around the world. (About Airbnb: What It Is and How It Works, n.d.). As of December 2022, There are over 6.6 million active listings hosted on Airbnb (Airbnb, 2023) making Airbnb a database full of both unstructured and structured data. This data has the potential to be used in classification and regression analysis as well as used for data mining.

The data collected by Airbnb can be used to predict the optimal price a new listing should be listed for by analyzing many of the features shown in a listing such as location, number of beds, number of guests, etc. Data mining on the dataset may also be able to provide valuable insights into which features may be the most important to users looking to host their property in hopes that the user can increase the popularity of their listing. Using the existing data, it may be possible to help existing users decide if their property is under priced, over priced, or competitive with other listings on Airbnb.

The data that will be used for this analysis consists of 494,954 records made publicly available by Inside Airbnb (Airbnb - Listings, 2017). Due to the large size of the entire data set, the data will be restricted to records located in Toronto which reduces the dataset down to approximately 12,500+ records (Airbnb - Listings, 2017). The data consists of categorical and numeric data. The whole dataset is available for download at Airbnb - Listings — Opendatasoft.

Exploratory data analysis plays a key role as an initial phase of implementation of the algorithms as there are initially 89 features in the dataset. As the dataset consists of both structured and unstructured data, restructuring or removal of unstructured data occurred during the preprocessing phase. The data was preprocessed extracting the most significant features and removing outliers and insignificant values. The dataset included missing values which required treatment by replacing the value or removing the feature or record.

Regression algorithms were performed on the dataset with Python as the sole tool for implementation. Performance of the regression model will be evaluated with Mean Squared Error (Wu, 2021) as the scoring metric. Clustering was performed to reduce the number of classifications of one of the features.

In conclusion, the regression analysis performed on the Airbnb data for Toronto listings will help a Toronto Airbnb host make their listing more profitable by pricing accurately for the listings features and strategically for the Toronto market.

Literature Review

Introduction

Airbnb is changing the hospitality industry by allowing individuals to use their homes, guest houses, investment properties, etc. as an alternative to renting in a hotel. This literature review critically examines existing research completed on Airbnb and analyzes different elements making an Airbnb listing successful. By reviewing six key articles, this study will identify attributes that contribute to or diminishes the success of an Airbnb listing. This research will allow existing and future hosts of Airbnb listings in Toronto evaluate their listing and make adjustments where required to make their listing more successful.

Article Summaries

Article 1: Price and RevPAR determinants of Airbnb listings: Convergent and divergent evidence

Details
Authors: Arvanitidis, P., Economou, A., Grigoriou, G., & Kollias, C.

Published: 2022

Summary
This article looked at factors that determine price and revenue across Airbnb listings in Milan. It is important to look at time as a factor as local events can increase the demand for Airbnb in an area. Size of the listing (beds and bathrooms) and listing type explain a third of price variance. Rental policies, reviews, and host attributes do not have a large effect on price. It is important to look at listing price and understand the difference between price and revenue per available room (RevPAR). Managerial skills and pricing strategies can be used to improve price.

Article 2: What do Airbnb users care about? An analysis of online review comments

Details
Authors: Cheng, M., & Jin, X.

Published: 2019

Summary
A study was performed using text mining and sentiment analysis on a big data set to determine attributes that influence an Airbnb's user's experience. The key attributes that users typically evaluate their experience is based on location, amenities and host. The results found surprising

results that price is not indicated as a factor that influenced users. Hotel attributes were used as a frame of reference to compare to the Airbnb reviews.

Article 3: Motivations and constraints of Airbnb consumers: Findings from a mixed-methods approach

Details
Authors: Fung, S., Kevin Kam, So, K. K. F., Oh, H., & Min, S.

Published: 2018

Summary
Many studies have been conducted looking at the behaviours and attitudes of people who choose to use Airbnb. Qualitative and quantitative methods were used to perform the research. Focus groups were used to conduct the research combining both users and nonusers of Airbnb to form the qualitative research. An online survey was created to collect results for the quantitative research. Results of the research indicate that “price value, community, home atmosphere, sustainability” are strong motivators for people to choose Airbnb and distrust and lack of value are motivators against the use of Airbnb.

Article 4: Why Tourists Choose Airbnb: A Motivation-Based Segmentation Study

Details
Authors: Guttentag, D., Smith, S., Potwarka, L., & Havitz, M

Published: 2018

Summary
This study aims to understand the motivations of Airbnb consumers and categorize the different types of consumers based on their motivations. Segmentation can provide insight into the types of users and prove valuable to marketers. Data was collected via an online survey. As the number of people who use Airbnb is relatively low to the entire population, many non random sampling techniques were used to reduce bias. Clustering analysis was performed on the dataset to help interpret the motivation data. Characteristics included in the survey indicated type of travel and type of accommodation used. A strong majority of respondents (80%) booked their Airbnb for leisure. Over 70% of respondents rented an Airbnb where they had the entire place, while almost 28% had a private room and just over 2% had booked a shared space. Clusters were named based on their motivation and include “Money Savers, Home Seekers, Collaborative Consumers, Pragmatic Novelty Seekers, and Interactive Novelty Seekers” and the clusters were compared on their similarities and differences. Results indicated that the practical advantages of using Airbnb are the main draws and experience is secondary.

Article 5: Dynamic pricing in Airbnb: Individual versus professional hosts

Details
Authors: Abrate, G., Sainaghi, R., & Mauri, A. G

Published: 2022

Summary
Research was conducted on Airbnb listings in Rome and Milan. This article looks at three gaps that determine a listing's performance: the ability to apply dynamic pricing, degree of professionalism, and the effectiveness of dynamic pricing. Dynamic pricing which is pricing that changes based on demand and other factors. Professional hosts are hosts that have more than one listing and have the superhost qualification. This study shows that dynamic pricing and professionalism are contributing factors to increasing revenue. Average booking lead time also positively impacts revenue. Non professional hosts can employ dynamic pricing, however they typically do not.

Article 6: Use Python Data Analysis to Gain Insights from Airbnb Hosts

Details
Authors: Tian, Z.

Published: 2021

Summary
There are many existing studies that are researching the same topic but coming up with different conclusions. This study conducts experiments using Python programming to perform analysis. The author uses Python to preprocess the data. The author uses sentiment analysis to study the behaviour of customers as it can replace surveys and focus groups. It appears that being a super host helps bring more traffic to a listing and therefore more bookings. The author recommends that hosts use the reviews on their listings to find what makes their listing unique and market towards that. Pricing strategies are extremely important and should be adjusted throughout the year to meet the demand. Prices also should remain competitive with local hotels in the areas. Accuracy, cleanliness, and perceived value are high value areas that all hosts should work to improve.

Analysis and Synthesis

Given the various articles listed above, there is a lot of research done on why consumers choose to use Airbnb over traditional lodging options such as a hotel. The literature review uncovered that people who are hesitant to use Airbnb tend to be hesitant due to distrust with hosts ((Cheng & Jin, 2019), (Fung et al., 2018), (Guttentag et al., 2018)). This was a common

theme that came up in many articles looking at consumer motivation articles. Additionally, the other research looked more at how hosts can improve their listing via pricing strategies but do not give recommendations for the price. The research was performed on major cities, such as Rome and Milan, but not directly on Toronto listings. The research that will be performed will be specific to Toronto and will look at the seasonality and pricing strategies specifically for Toronto. When looking at which features are most important for pricing a Toronto Airbnb, the results will be compared to research done by Arvanitidis et al. to help guide feature selection for the regression models as the goal of the research is similar.

The research that will be completed will focus on the following questions:

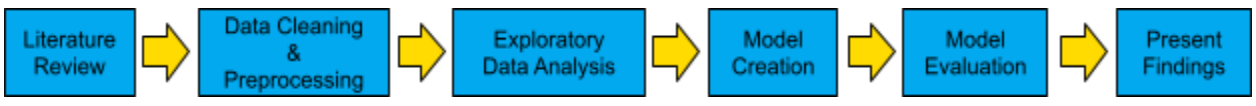
1. What attributes of an Airbnb rental are more likely to make a rental successful in Toronto?
2. Can the optimal price of a Toronto Airbnb be determined by the AirBnbs attributes?
3. How many Airbnb properties are accurately priced vs being over or underpriced?

This research is important because it will be specific to the Toronto market. The previous research done provides general observations and not specific recommendations.

Methodology

To conduct this literature review, research was performed on scholarly articles found via https://library.torontomu.ca/. The research focuses on why a person would select an Airbnb so the search keywords used were selected to reflect that. Keywords for the search included words such as ‘Airbnb’, ‘consumer’, ‘perception’, ‘pricing’. In addition, the keyword ‘data’ was used to identify if there were studies done relating to the collection and usage of Airbnb data.

Following the literature review, research continued with further data cleaning and preprocessing. Steps completed are outlined in the Descriptive Statistics section below walking through the Exploratory Data Analysis. Three different models (Linear Regression, Random Forest Regressor, and Decision Tree) were created and refined, optimizing the features selected (using Recursive Feature Elimination) and the number of folds as part of k-fold cross validation. A complete evaluation of the performance across the various models was completed. The findings are reported in this document and will be formally presented and discussed.



Data Preprocessing

The dataset used to conduct the research is a large dataset that is made publicly available and available for download at Airbnb - Listings — Opendatasoft. The dataset contains almost 500,000 records but has been reduced down to only include Toronto Airbnb listings. Records were only kept if they had more than one review to keep predictions relating to active or previously active records and not listings that were created and never used. This brings the dataset down to 9831 records to begin preprocessing.

The dataset consists of 89 columns including a mix of numerical, string, list and dictionary data types. After preprocessing, 33 columns remained for feature selection.

Remove Columns Not Contextually Relevant

ID Columns

There are 3 numeric columns that consist of IDs (id, scrape_id, and host_id). An ID does not contribute to the prediction of the record. All 3 ID columns have been removed.

URL Columns

There are 8 columns containing text that provide URLs linking to the listing, photos, the host, etc which have all been removed. The URL columns removed are listing_url, thumbnail_url, medium_url, picture_url, xl_picture_url, host_url, host_thumbnail_url, and host_picture_url.

Review Columns

As the research looks to provide information to Airbnb hosts as they create listings, columns containing review information would not be available so the columns were removed. The removed columns are number_of_reviews, first_review, last_review, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, and reviews_per_month.

Treat Null Values

Completely Null Columns

There are 4 columns that consist of only Null values (neighbourhood_group_cleansed, has_availability, license, jurisdiction_names). These Columns were removed.

Majority Null Columns (>60%)

Columns with a large percentage of null values have been removed: square_feet - 98%, weekly_price - 81%, and monthly_price - 80%.

Notes is null 58% of the time, however this will be explored more throughout the research as notes could be useful to people looking to book an Airbnb.

Other Columns with Null Values

Fields like security_deposit and cleaning_fee have many Null values (4021 and 2203 respectively) however the Null values were replaced with 0s to represent that there was no fee.

Other columns did not have many records with Null values. The records with Null values were removed. Host_listings_count had 1 Null value. host_total_listings_count had 1 Null value. Bathrooms had 15 Null values. Bedrooms had 6 Null values. Beds had 10 Null values. Amenities had 32 Null values. Price had 13 Null values. Features had 2 Null values. daysAsHost had 1 Null value.

Treat Columns that Do Not Add Information

Remove Columns with Single Value

The 6 columns that only had 1 unique value were last_scraped, experiences_offered, city, country_code, country, calendar_last_scraped.

Additionally state, market, smart_location have also been removed as their values were very similar to each other (for example state had various values that all referred to Ontario).

Removed Columns that Contain Information Similar to Other Columns

Geolocation was removed as it is a combination of longitude and latitude columns.

Convert Columns

Group Latitude and Longitude via Classification

Latitude and Longitude were grouped into 10 clusters. The aim of this was to reduce the number of neighbourhoods from 139 to 10. As such zipcode, neighbourhood, and neighbourhood_cleansed were removed as they represent similar information.

Long Text Fields

There are 11 fields that contain unstructured free text. Since the regression models require numeric fields, these long text fields were converted into integers representing their length. If a field was Null, it was given the value 0. The reasoning behind this is the text provides a user with information. For example a listing with a long, beautiful description could make someone more likely to book while a listing with a long list of house rules may make it less likely for a user to book.

The fields converted to their length are space, neighbourhood_overview, summary, description, notes, transit, access, interaction, house_rules, host_about, and name.

Date to Number of Days Since

host_since is a date field that is now converted to an integer called daysAsHost. This will help evaluate if a more experienced host would be better at accurately pricing their listing.

Categorical Text to Integers

Categorical text fields have been converted to integers in order to be used in the regression algorithms. The text field was then removed.

The columns host_verifications, amenities, and features contain lists. There are 22 unique host_verification values within the lists, 7 within features lists, and 108 within amenities lists. In total, 23 columns have been dropped which leaves 63 columns remaining. Of those 60 remaining columns, 30 columns are numeric.

Remove Outliers

The Interquartile Range was determined for every column. Any value that was >= the Upper Bound or <= to the Lower Bound was removed, except in cases where the Lower Bound was equal to the Upper Bound. 6276 records containing outliers were removed, leaving 3486 records remaining.

Table 1: Attributes

Attribute	Type	Status	Rationale
id	Numeric ID	Removed	Not relevant information
listing_url	URL	Removed	Not relevant information
scrape_id	String	Removed	Not relevant information
last_scraped	Date	Removed	Only 1 value in column
name	Unstructured Text	Converted to nameLen	See if length of free text field is valuable
summary	Unstructured Text	Converted to summaryLen	See if length of free text field is valuable
space	Unstructured Text	Converted to spaceLen	See if length of free text field is valuable
description	Unstructured	Converted to	See if length of free text field is valuable

	Text	descriptionLen	
experiences_offered	Unstructured Text	Removed	Only 1 value in column
neighborhood_overview	Unstructured Text	Converted to neighourbood_overviewLen	See if length of free text field is valuable
notes	Unstructured Text	Converted to notesLen	See if length of free text field is valuable
transit	Unstructured Text	Converted to transitLen	See if length of free text field is valuable
access	Unstructured Text	Converted to accessLen	See if length of free text field is valuable
interaction	Unstructured Text	Converted to interactionLen	See if length of free text field is valuable
house_rules	Unstructured Text	Converted to house_rulesLen	See if length of free text field is valuable
thumbnail_url	URL	Removed	Not relevant information
medium_url	URL	Removed	Not relevant information
picture_url	URL	Removed	Not relevant information
xl_picture_url	URL	Removed	Not relevant information
host_id	Numeric ID	Removed	Not relevant information
host_url	URL	Removed	Not relevant information
host_name	Unstructured Text	Removed	Not relevant information
host_since	Date	converted to Days as Host	Created to see if people who are hosts longer have more accuracy with their listings
host_location	Unstructured	Removed	Not relevant information

	Text		
host_about	Unstructured Text	Converted to hostAboutLen	
host_response_time	Categorical	Converted to host_response_timeCode	
host_response_rate	Numeric	Removed	Not available for new hosts
host_acceptance_rate	Numeric	Removed	All Null values
host_thumbnail_url	URL	Removed	Not relevant information
host_picture_url	URL	Removed	Not relevant information
host_neighbourhood	Categorical	Converted to host_neighbourhoodCode	160 potential values
host_listings_count	Numeric	Removed	Highly correlated with host_total_listings and calculated_host_listings
host_total_listings_count	Numeric		
host_verifications	List	Removed	22 Potential Values in each list
street	Unstructured Text	Removed	Captured under NeighbourhoodNew
neighbourhood	Categorical	Removed	Captured under NeighbourhoodNew
neighbourhood_cleansed	Categorical	Converted to neighbourhood_cleansedCode	138 potential values
neighbourhood_group_cleansed	Null	Removed	All Null Values
city	Unstructured Text	Filtered & Removed	Keep only values equal to ‘Toronto’ so information is relevant to research, then

			drop as only 1 value in column
state	Unstructured Text	Removed	Only 4 values in column - all different ways of saying Ontario (Ontario, ON, Ont, On)
zipcode	Unstructured Text	Removed	Captured under NeighbourhoodNew
market	Unstructured Text	Removed	Only 7 values in column, don't appear accurate/relevant
smart_location	Unstructured Text	Removed	Only 2 values in column - saying the same things ('Toronto, Canada' and 'Toronto , Canada'
country_code	Categorical	Removed	Only 1 value in column
country	Categorical	Removed	Only 1 value in column
latitude	Numeric	Converted to neighbourhoodNew and dropped	Used to create 10 new clusters to help reduce number of neighbourhoods
longitude	Numeric	Converted to neighbourhoodNew and dropped	Used to create 10 new clusters to help reduce number of neighbourhoods
property_type	Categorical	Converted to property_typeCode	20 potential values
room_type	Categorical	Converted to room_typeCode	3 potential values
accommodates	Numeric		
bathrooms	Numeric		
bedrooms	Numeric		
beds	Numeric		
bed_type	Categorical	Converted to	5 potential values

		bed_typeCode	
amenities	List	Removed	108 Potential Values in each list
square_feet	Numeric	Removed	Was Null for >60% of records
price	Numeric	Target Attribute	
weekly_price	Numeric	Removed	Was Null for >60% of records
monthly_price	Numeric	Removed	Was Null for >60% of records
security_deposit	Numeric		Replaced null with 0 to represent a $0 fee
cleaning_fee	Numeric		Replaced null with 0 to represent a $0 fee
guests_included	Numeric		
extra_people	Numeric		
minimum_nights	Numeric		
maximum_nights	Numeric		
calendar_updated	Unstructured Text	Removed	Not valuable for new listings
has_availability	Null	Removed	All Null Values
availability_30	Numeric		
availability_60	Numeric		
availability_90	Numeric	Removed	Strongly correlated with availability_60
availability_365	Numeric		
calendar_last_scraped	Date	Removed	Only 1 value in column
number_of_reviews	Numeric	Removed	Would not be available for a new listing
first_review	Date	Removed	Would not be available for a new listing

last_review	Date	Removed	Would not be available for a new listing
review_scores_rating	Numeric	Removed	Would not be available for a new listing
review_scores_accuracy	Numeric	Removed	Would not be available for a new listing
review_scores_cleanliness	Numeric	Removed	Would not be available for a new listing
review_scores_checkin	Numeric	Removed	Would not be available for a new listing
review_scores_communication	Numeric	Removed	Would not be available for a new listing
review_scores_location	Numeric	Removed	Would not be available for a new listing
review_scores_value	Numeric	Removed	Would not be available for a new listing
license	Null	Removed	All Null Values
jurisdiction_names	Null	Removed	All Null Values
cancellation_policy	Categorical	Converted to cancellation_policyCode	3 potential values
calculated_host_listings_count	Numeric	Removed	Strongly correlated with host_listings_count and host_total_listings_count
reviews_per_month	Numeric	Removed	Would not be available for a new listing
geolocation	Dictionary	Removed	Combination of latitude and longitude
features	List	Removed	7 Potential Values in each list

Descriptive Statistics

Numerical Attributes

Below are tables of the statistics describing the remaining columns.

Table 2a: Numerical Statistics on Columns

	host_listings_count	host_total_listings_count	accommodates	bathrooms	bedrooms	beds
count	3486.000	3486.000	3486.000	3486.000	3486.000	3486.000
mean	1.351	1.351	2.501	1.107	1.054	1.325
std	0.637	0.637	1.092	0.314	0.523	0.560
min	0.000	0.000	1.000	0.000	0.000	1.000
25%	1.000	1.000	2.000	1.000	1.000	1.000
50%	1.000	1.000	2.000	1.000	1.000	1.000
75%	2.000	2.000	3.000	1.000	1.000	2.000
max	3.000	3.000	6.000	5.000	3.000	3.000

Table 2b: Numerical Statistics on Columns

	price	security_deposit	cleaning_fee	guests_included	extra_people	minimum_nights	maximum_nights
count	3486.000	3486.000	3486.000	3486.000	3486.000	3486.000	3486.000
mean	94.376	134.091	28.238	1.259	7.283	1.904	737.357
std	44.038	168.364	26.982	0.490	10.461	1.041	519.883
min	18.000	0.000	0.000	1.000	0.000	1.000	1.000
25%	59.000	0.000	0.000	1.000	0.000	1.000	30.000
50%	89.000	0.000	25.000	1.000	0.000	2.000	1125.000
75%	120.000	250.000	50.000	1.000	15.000	2.000	1125.000
max	228.000	700.000	122.000	3.000	45.000	5.000	1200.000

Table 2c: Numerical Statistics on Columns

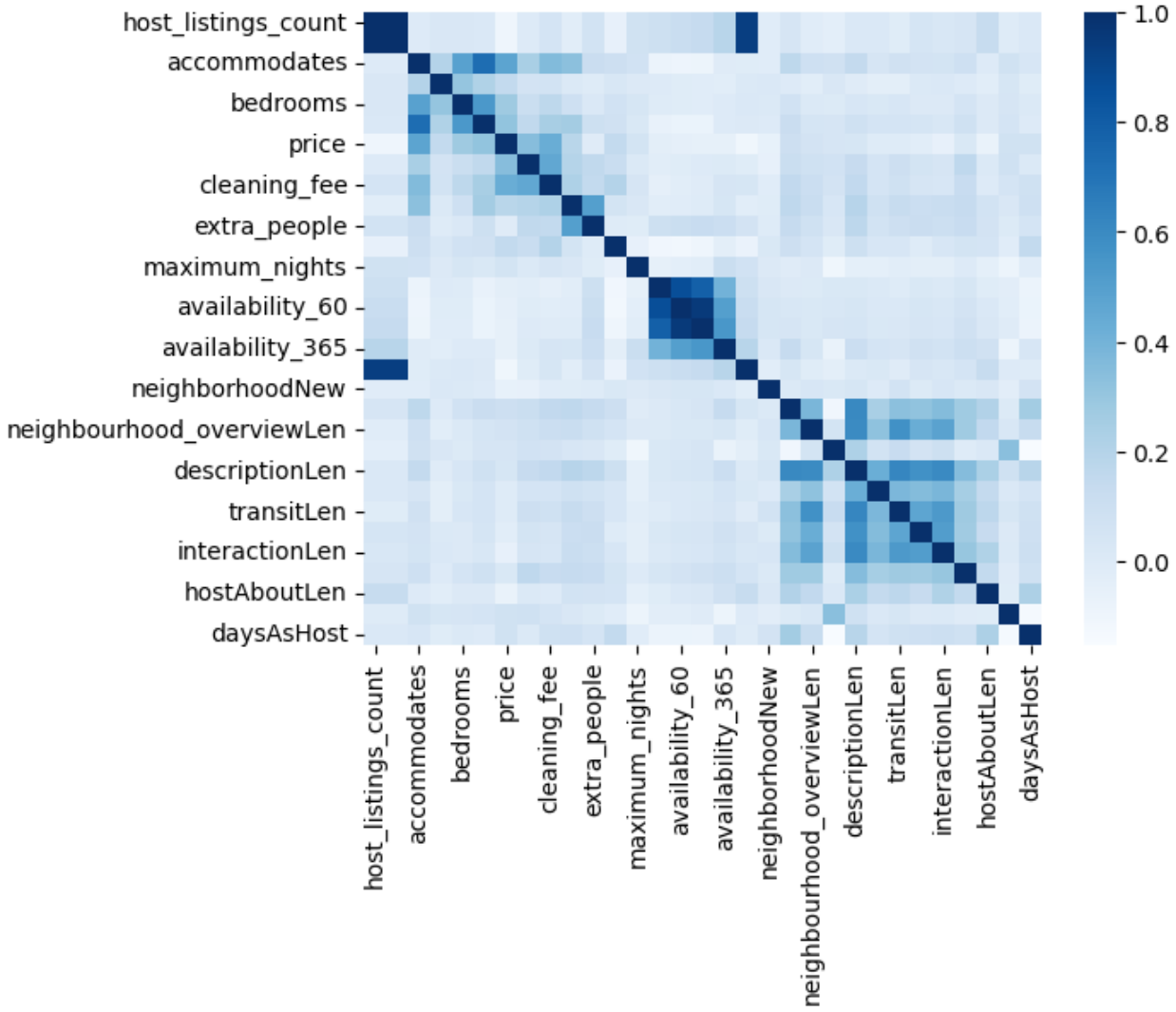
	availability_30	availability_60	availability_90	availability_365	calculated_host_listings_count
count	3486.000	3486.000	3486.000	3486.000	3486.000
mean	5.293	14.594	26.431	129.824	1.305
std	7.348	17.168	27.755	129.865	0.594
min	0.000	0.000	0.000	0.000	1.000
25%	0.000	0.000	0.000	1.000	1.000
50%	2.000	7.000	17.000	78.000	1.000
75%	9.000	26.000	48.000	272.000	1.000
max	27.000	57.000	87.000	362.000	3.000

Table 2d: Numerical Statistics on Columns

	spaceLen	neighbourhood_overviewLen	summaryLen	descriptionLen	notesLen	transitLen
count	3486.000	3486.000	3486.000	3486.000	3486.000	3486.000
mean	230.705	148.775	302.336	723.068	44.135	111.274
std	292.444	189.467	124.603	311.702	77.554	129.951
min	0.000	0.000	0.000	0.000	0.000	0.000
25%	0.000	0.000	231.000	444.000	0.000	0.000
50%	115.000	73.500	267.000	883.000	0.000	74.000
75%	351.750	250.000	419.000	1000.000	70.000	183.750
max	1000.000	930.000	635.000	1000.000	343.000	579.000

Correlated Attributes

Chart 1: Numerical Attribute Correlation Heat Map



Some interesting correlations are observed below. While the heatmap above shows only absolute values, the true values observed may indicate negative correlations, though all are weak.

Table 7: Numerical Data Correlation

Column 1	**Column 2**	**Correlation**	**Notes**
host_listings_count	host_total_listings_count	1.00	Drop column host_total_listings_count
accommodates	beds	0.726488	Strong positive correlation, which makes sense given the context of the data. Keep both attributes for now
availability_30	availability_60	0.8665514	Strong positive correlation, contextually very similar Drop availability_60 and availability_90
availability_90	availability_60	0.9521526	
availability_30	availability_90	0.7850194	

Categorical Data

There are 8 columns with categorical data.

Table 8: Categorical Data

Column Name	**Number of Possible Values**	**Possible Values**
host_response_time	4 (See Chart 2)	within an hour - 0 within a day - 1 within a few hours - 2 None - 3 a few days or more - 4
neighbourhoodNew	10	0-9

property_type	21 (See Chart 6)	Had to combine many options into category 1 due to an unbalanced dataset. Condominium - 0 Apartment - 2 House - 3 Loft - 1 Guest suite - 1 Guesthouse - 1 Townhouse - 1 Bed & Breakfast - 1 Other - 1 Bungalow - 1 Cabin - 1 Serviced apartment - 1 Boutique hotel - 1 Dorm - 1 Boat - 1 Hostel - 1 Tent - 1 Villa - 1 Camper/RV - 1 In-law - 1
room_type	3 (See Chart 7)	Entire home/apt - 0 Private room - 1 Shared room - 2
bed_type	5 (See Chart 8)	Had to combine many options into category 1 due to an extremely unbalanced dataset. Real Bed - 0 Pull-out Sofa - 1 Futon - 1 Couch - 1 Airbed - 1
cancellation_policy	4 (See Chart 9)	strict - 0 flexible - 1 moderate - 2

Chart 2: neighbourhoodNew Histogram

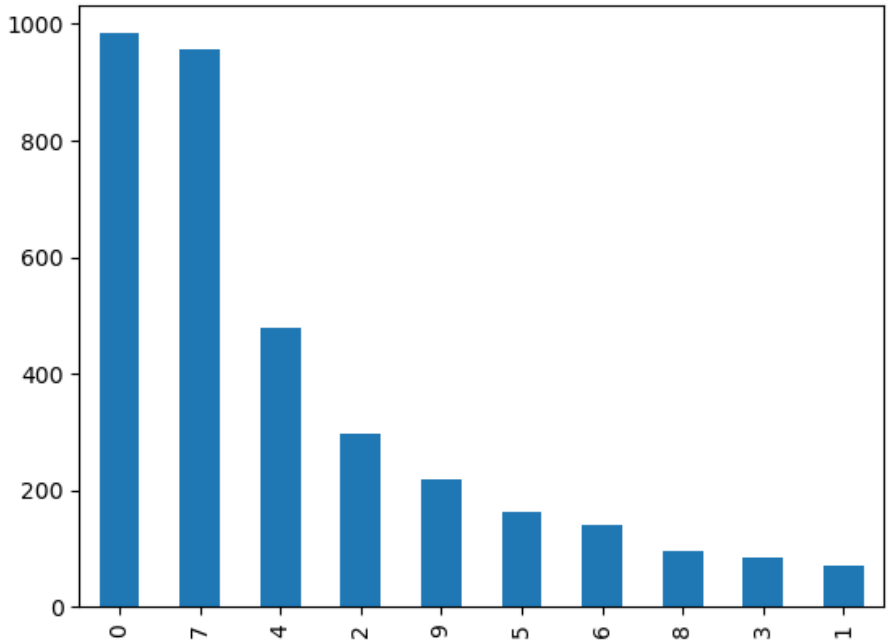


Chart 3: property_typeCode Histogram

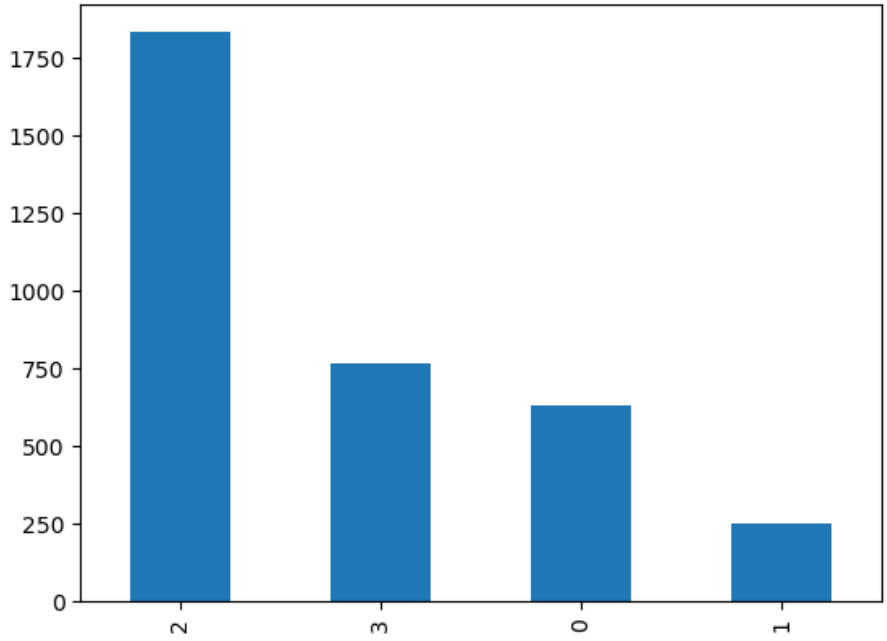


Chart 4: room_typeCode Histogram

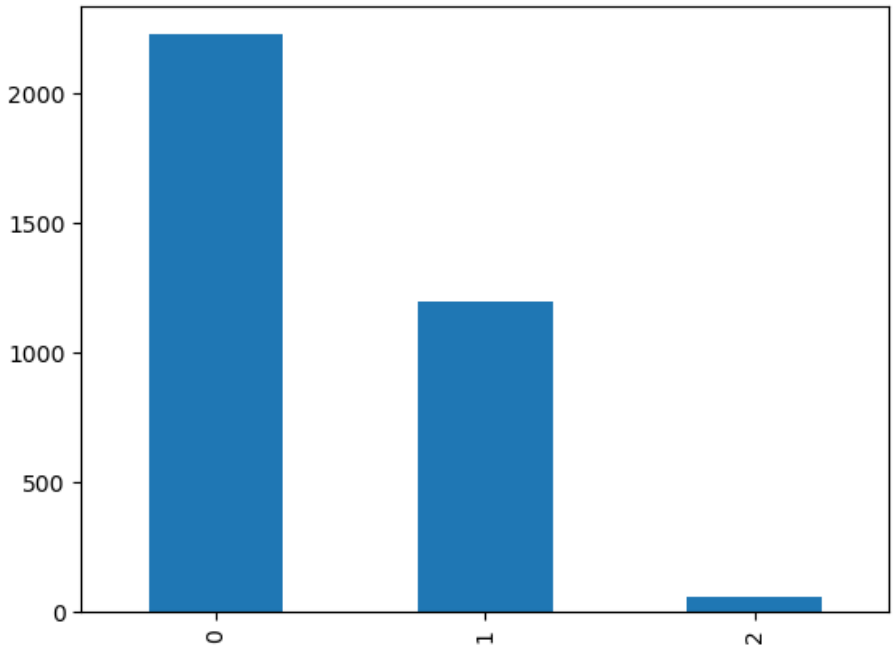


Chart 5: bed_typeCode Histogram

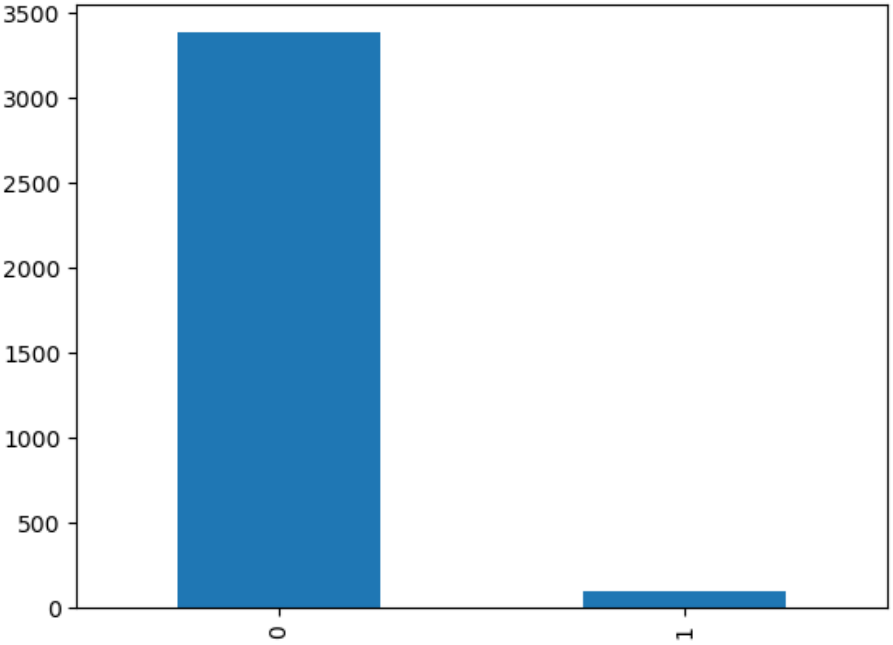
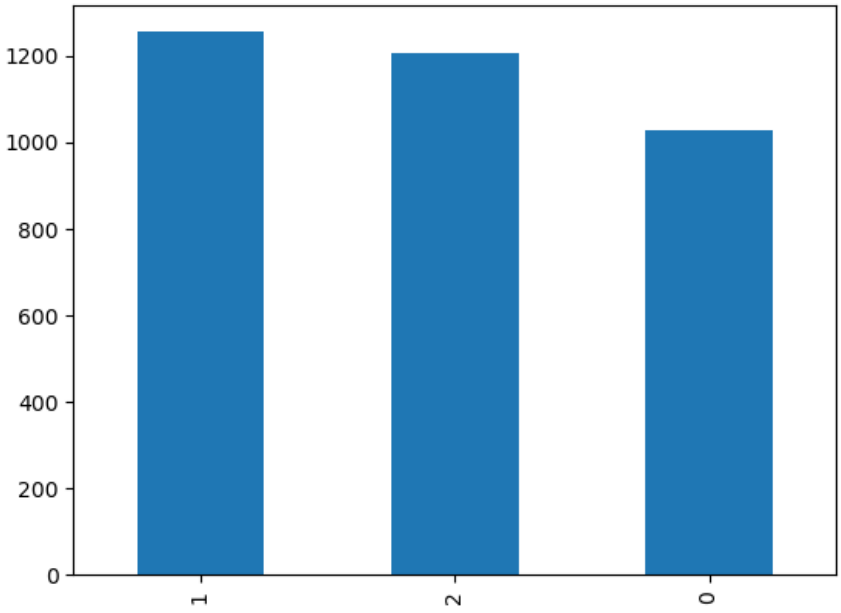


Chart 9: cancellation_policy Histogram



Model Evaluation and Feature Selection

As the research is looking to predict the price for an Airbnb listing, regression algorithms are best suited. The research evaluates Linear Regression, Random Forest Regressor, and Decision Trees.

First models were created with different numbers of folds for k-fold cross validation. Folds ranged from 3-10 with all features included. All models were compared using Mean Square Error (MSE) and Standard Deviation (SD). A combined score was created with a 70% weight on MSE. Decision Tree performed the worst and the MSEs were almost double what the MSEs were showing for Linear Regression and Random Forest.

Once the models were evaluated with all features, feature selection could begin. The RFE package was used to complete the feature selection. The best performing model with all features was the Random Forest Regressor. The MSE was 886.2972512 with a standard deviation of 29.79839485.

After performing RFE, 24 features were selected.

The selected features are:

- host_total_listings_count
- accommodates
- bathrooms

- bedrooms
- beds
- security_deposit
- cleaning_fee
- guests_included
- extra_people
- minimum_nights
- maximum_nights
- availability_30
- availability_60
- availability_365
- neighborhoodNew
- spaceLen
- neighbourhood_overviewLen
- summaryLen
- descriptionLen
- notesLen
- transitLen
- accessLen
- interactionLen
- houseRulesLen
- hostAboutLen
- nameLen
- daysAsHost
- property_typeCode
- room_typeCode
- bed_typeCode
- cancellation_policyCode

The selected model had a MSE of 893.5748428288821 and SD of 23.446059785814146. This means that if an Airbnb was listed for $100 and the model predicts it to be $129.82, the squared difference would be (100-129.88)^2 = 892. A smaller SD indicates that the model is more stable because the variability is less.

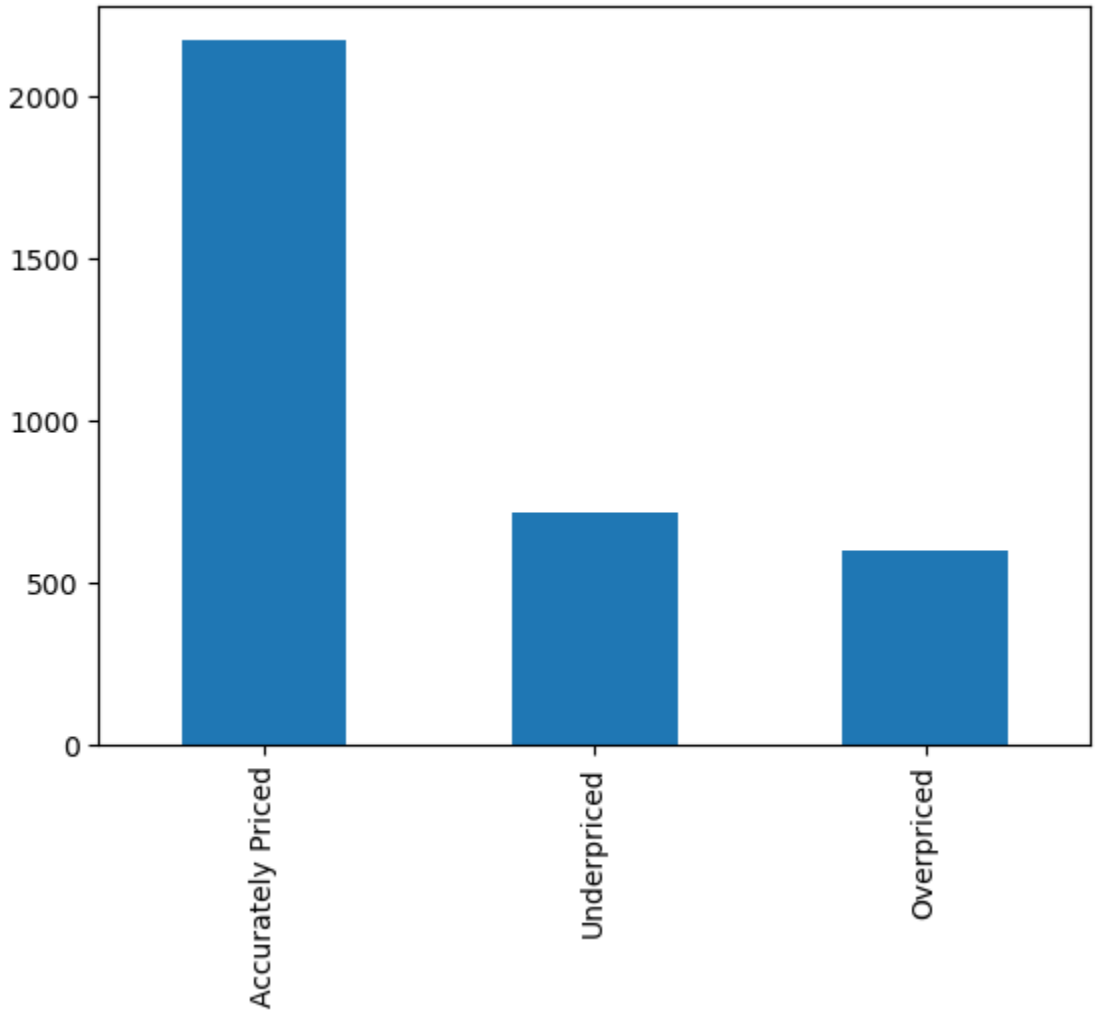
When using the model to predict the following metrics were given:

- Mean Absolute Error (MAE): 22.618955823293174
- Mean Squared Error (MSE): 904.6209608146874
- Root Mean Squared Error (RMSE): 30.076917408781895

In terms of looking at the actual price vs the predicted price, the difference in price vs predicted was compared to the MAE and RMSE to determine how many listings were accurately priced vs

under or over priced as shown in Chart 10. Over 2000 are accurately priced, while approximately 600 are under priced and approximately 500 are overpriced.

Chart 10: Comparing the Predicted Price to Actual Price



Limitations

There are some limitations with this model and future areas for research.

Limitation 1: Changing World Conditions

Inflation
This model uses data that was extracted from Airbnb in 2017. Since then inflation has increased and so has the cost of living. This is something that can change how much a host would need to charge, and may need to increase fees like cleaning fees.

Pandemics
In 2020 the world was hit with a global pandemic called COVID-19. Many countries imposed lockdowns and people were unable to travel. Based on supply and demand, it is likely that Airbnb would not have many customers and would need to consider lowering their prices.

Solution
It is important to continue to monitor the performance of the model. If values such as the MSE increased significantly, the model would need to be retrained with data that more accurately reflects the real world conditions.

Limitation 2: More Data Available

Short Timelines
Columns in the original data sets contained lists were dropped as there was so much data to use and evaluate already. Due to the length of the course, more research could be performed on the values present in those lists. For example, Amenities contained information such as whether an Airbnb listing has laundry, air conditioning, television, etc. and these are all factors that a user may consider when booking a rental.

Recommendation
Separate the list into more columns with binary values. Continuing the example above, create columns to represent laundry, air conditioning, television, etc. and have the value as 1 if the listing has that amenity, and 0 if it does not.

Limitation 3: Computationally Heavy

Some functions created during the research take a long time to compute. Specifically when evaluating Random Forest and how many folds and features are recommended can take up to 30 minutes to run.

Recommendation
Look for more built in functions to complete this task more efficiently.

Conclusion

In conclusion, the price for an Airbnb can be predicted using a Random Forest regression model. There are many features that contribute to the decision on the price selected. Using the metrics of the model, you can observe how often Airbnb listings are accurately priced vs over or under priced. There is room for future work as more data becomes available and features that have not been explored.

GitHub Repository

All files will be stored using the following GitHub repository: https://github.com/11lmb23/cind820

References

About Airbnb: What it is and how it works. (n.d.). Airbnb Help Centre. Retrieved May 14, 2023, from https://www.airbnb.ca/help/article/2503

Abrate, G., Sainaghi, R., & Mauri, A. G. (2022). *Dynamic pricing in airbnb: Individual versus professional hosts* Elsevier. doi:10.1016/j.jbusres.2021.12.012

Airbnb. (2023, April 12). About us - Airbnb Newsroom. Airbnb Newsroom. https://news.airbnb.com/about-us/

Airbnb - Listings. (2017, July 18). Opendatasoft. Retrieved May 14, 2023, from https://public.opendatasoft.com/explore/dataset/airbnb-listings/information/?disjunctive.host_verifications&disjunctive.amenities&disjunctive.features&dataChart=eyJxdWVyaWVzIjpbeyJjaGFydHMiOlt7InR5cGUiOiJjb2x1bW4iLCJmdW5jIjoiQ09VTlQiLCJ5QXhpcyI6Imhvc3RfbGlzdGluZ3NfY291bnQiLCJzY2llbnRpZmljRGlzcGxheSI6dHJ1ZSwiY29sb3IiOiJyYW5nZS1jdXN0b20ifV0sInhBeGlzIjoiY291bnRyeSIsIm1heHBvaW50cyI6IiIsInRpbWVzY2FsZSI6IiIsInNvcnQiOiIiLCJjb25maWciOnsiZGF0YXNldCI6ImFpcmJuYi1saXN0aW5ncyIsIm9wdGlvbnMiOnsiZGlzanVuY3RpdmUuaG9zdF92ZXJpZmljYXRpb25zIjp0cnVlLCJkaXNqdW5jdGl2ZS5hbWVuaXRpZXMiOnRydWUsImRpc2p1bmN0aXZlLmZlYXR1cmVzIjp0cnVlfX0sInNlcmllc0JyZWFrZG93biI6Imhvc3RfcmVzcG9uc2VfdGltZSJ9XSwidGltZXNjYWxlIjoiIiwiZGlzcGxheUxlZ2VuZCI6dHJ1ZSwiYWxpZ25Nb250aCI6dHJ1ZX0%3D&location=2,13.11707,-0.08341&basemap=jawg.light

Arvanitidis, P., Economou, A., Grigoriou, G., & Kollias, C. (2022). *Trust in peers or in the institution? A decomposition analysis of airbnb listings' pricing* Channel View publications. doi:10.1080/13683500.2020.1806794

Cheng, M., & Jin, X. (2019). *What do airbnb users care about? an analysis of online review comments* Elsevier. doi:10.1016/j.ijhm.2018.04.004

Fung, S., Kevin Kam, So, K. K. F., Oh, H., & Min, S. (2018). *Motivations and constraints of airbnb consumers: Findings from a mixed-methods approach* Elsevier. doi:10.1016/j.tourman.2018.01.009

Guttentag, D., Smith, S., Potwarka, L., & Havitz, M. (2018). *Why tourists choose airbnb: A*

motivation-based segmentation study Sage Publications. doi:10.1177/0047287517696980

Practical Guide to Clustering Algorithms & Evaluation in R. (2017, April 13). HackerEarth.

https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/clustering-algorithms-evaluation-r/tutorial/

Tao, C. (2021, December 13). How to Evaluate a Classification Machine Learning Model. Medium.

https://towardsdatascience.com/how-to-evaluate-a-classification-machine-learning-model-d81901d491b1

Tian, Z. (2021). Use Python Data Analysis to Gain Insights from Airbnb Hosts. *Advances in Mathematical Physics, 2021*https://doi.org/10.1155/2021/1079850

Wu, S. (2021, December 14). 3 Best metrics to evaluate Regression Model? - Towards Data Science. Medium.

https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b