Final Report - Analysis of Airbnb Data

Haskayne School of Business, University of Calgary
BTMA 531 W2024 Data Analytics Tools for Business
Presented to Dr. Hooman Hidaji
April 3, 2024

Group 8

Aneesha Bapat (30094054)
Cooper Chung (30061289)
Justin Yu (30093886)
James Zhou (30065110)

Table of Contents

Problem Formulation

Due to COVID-19, many traveling plans were put on hold. As of late, we are starting to see loosened traveling measures against COVID-19, and many countries are opening back up to accepting travelers and vacationers. For hosts with listings on the rental property website Airbnb, this also opens up profit opportunities. However, due to the change in the real-estate climate and climbing costs, how can hosts ensure their pricing strategy is equitable yet competitive? How can hosts get the most out of their hard-earned property? This leads us to our problem statement - **what factors significantly impact the pricing of Airbnb listings, and how can these insights inform hosts to optimally and strategically price their listings?** We will generate a model to help gauge trends in pricing and listings and help hosts determine a reasonable market price to list their property. While the focus of this business problem is informing a pricing strategy for the Airbnb host, the models and analysis done to predict price have numerous use cases, including internally gauging trends in pricing and listings, for Airbnb guests tracking the Airbnb market for their next vacations, or for investors determining which properties to invest in for Airbnb ventures.

Data

The dataset of interest for this analytics project is the "Airbnb Price Dataset" derived from Kaggle. It contains 29 columns and over 74,000 observations, with each unit of observation being a property listing on Airbnb. The primary target variable is 'log_price' which we have exponentiated in the initial data processing stage to instead work with non-logarithm price values. We also focus the majority of analysis on factors that directly describe either the listing or the property. The nontrivial variables are categorized as follows:

- Variables describing the property in the listing: *property_type, room_type, amenities, accommodates, bathrooms, bed_type, bedrooms, beds*
    - Geographical data: *city, latitude, longitude, neighbourhood, zipcode*
- Variables detailing the listing: *cancellation_policy, cleaning_fee, instant_bookable, last_review, name, number_of_reviews, review_scores_rating, thumbnail_url, description*
- Variables describing the host profile: *host_has_profile_pic, host_identity_verified, host_response_rate, host_since*

Code

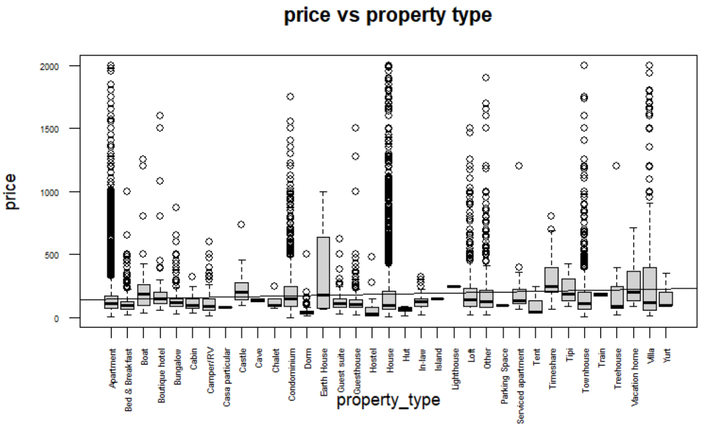
- *EDA.Rmd:* This file consists of data processing for initial exploratory data analysis and regression models. It features boxplots, regression models, Q-Q plots and residual plots.
- *TEXTANALYTICS.R*: This file consists of text-mining operations to gather information on keywords. It features analysis on listing titles, descriptions, and the creation of wordclouds.
- *KMEANS.R:* This file consists of some data manipulation, creation of k-means models, and plots of the results. Since k-means is usually done with numerical data, some data preparation is required to ensure that the data is suitable for k-means analysis. Models were created for k values of 2, 3, and 4, and each of these were plotted in the end to give a visual indication of the results of clustering.

- *REGRESSIONTREE.R:* This file consists of the creation of a regression tree. It features a test and training dataset, along with a regression tree model and prediction with plot.
- *CLASSIFICATION_TREE.R:* This file consists of the creation of a classification tree model. It features some minor data manipulation, the creation of the test and training dataset, along with a classification tree model, and prediction with plot.

Analysis

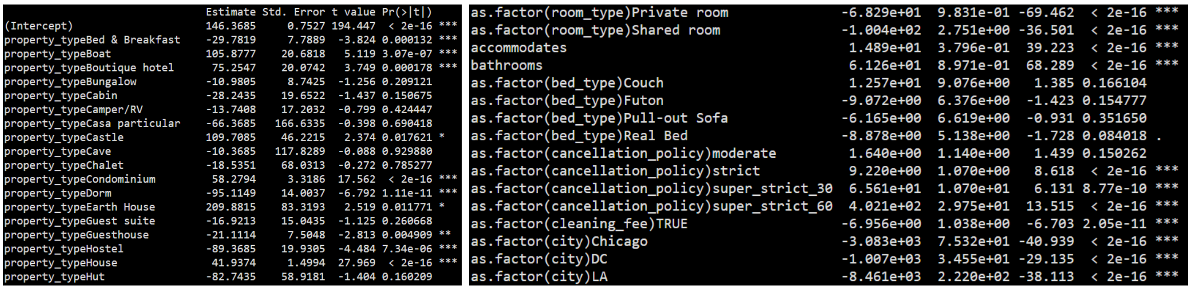
Linear Regression and EDA

As part of the initial exploratory analysis, several linear regressions, box plots, and quantile-quantile plots were performed to promote a basic understanding of the dataset. A basic linear regression and boxplot to analyze the relationship between listing price and the 35 different property types reveals the following statistics: *multiple R-squared=0.023; RMSE=166.63; MAE=97.18*. This indicates property type has low explanatory power and high error margin, and there are other variables that contribute to the target variable. Despite this, 13 of the coefficients are deemed statistically significant based on p-values that are less than 0.001. These coefficients estimate the price change for the property type compared to the intercept; for example, villas were estimated to be $237.31 higher than the reference category, and dorms were estimated to be $95.11 lower.



A multiple linear regression was also performed to predict the price based on 14 of the 28 categories deemed most relevant, including city, cancellation policy, room type, and review score, with the following statistics: *multiple R-squared = 0.515; RMSE=93.80; MAE=54.34.* This model has a higher explanatory power and lower error compared to the simple regression, and over half of the coefficients with statistically significant p values. A limitation of this model is multicollinearity, as redundant variables may correlate closely with each other. To assess the severity of multicollinearity, we computed the variance inflation factor (VIF) among predictor variables. This verified our concern, most notably with city, latitude, and longitude having moderate to severe multicollinearity levels. While these models give a good initial insight into the data, we rely on the other analysis in this paper to combat these limitations.

variable	*Variance Inflation Factor*
as.factor(city)	*2.961614e+07*
latitude	*2.151886e+03*
longitude	*6.254250e+04*

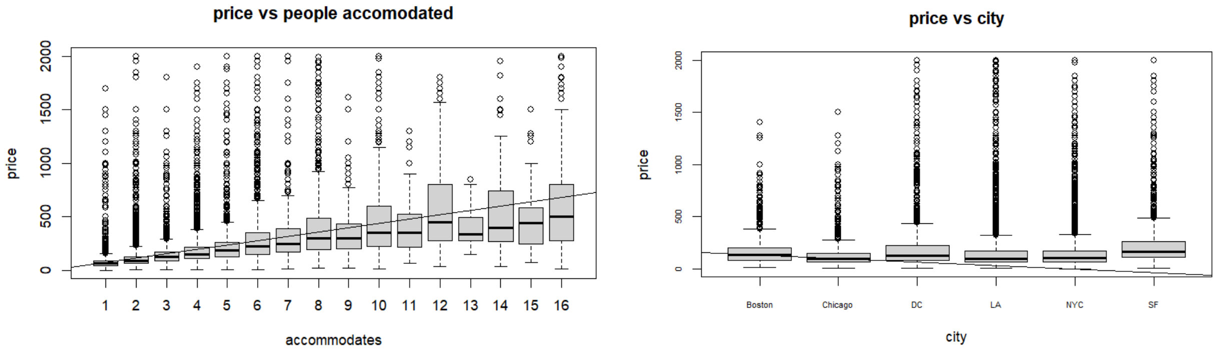


```
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 146.3685     0.7527 194.447  < 2e-16 ***
property_typeBed & Breakfast -29.7819    7.7889  -3.824 0.000132 ***
property_typeBoat           105.8777    20.6818   5.119 3.07e-07 ***
property_typeBoutique hotel  75.2547    20.0742   3.749 0.000178 ***
property_typeBungalow       -10.9805     8.7425  -1.256 0.209121    
property_typeCabin          -28.2435    19.6522  -1.437 0.150675    
property_typeCamper/RV      -13.7408    17.2032  -0.799 0.424447    
property_typeCasa particular -66.3685  166.6335  -0.398 0.690418    
property_typeCastle         109.7085    46.2215   2.374 0.017621 *  
property_typeCave           -10.3685   117.8289  -0.088 0.929880    
property_typeChalet         -18.5351    68.0313  -0.272 0.785277    
property_typeCondominium     58.2794     3.3186  17.562  < 2e-16 ***
property_typeDorm           -95.1149    14.0037  -6.792 1.11e-11 ***
property_typeEarth House    209.8815    83.3193   2.519 0.011771 *  
property_typeGuest suite    -16.9213    15.0435  -1.125 0.260668    
property_typeGuesthouse     -21.1114     7.5048  -2.813 0.004909 ** 
property_typeHostel         -89.3685    19.9305  -4.484 7.34e-06 ***
property_typeHouse           41.9374     1.4994  27.969  < 2e-16 ***
property_typeHut            -82.7435    58.9181  -1.404 0.160209    
```

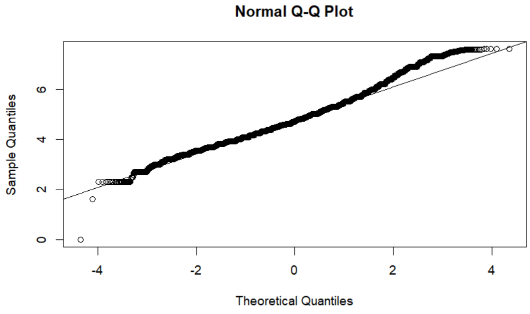
```
as.factor(room_type)Private room                 -6.829e+01  9.831e-01 -69.462  < 2e-16 ***
as.factor(room_type)Shared room                  -1.004e+02  2.751e+00 -36.501  < 2e-16 ***
accommodates                                      1.489e+01  3.796e-01  39.223  < 2e-16 ***
bathrooms                                         6.126e+01  8.971e-01  68.289  < 2e-16 ***
as.factor(bed_type)Couch                          1.257e+01  9.076e+00   1.385 0.166104    
as.factor(bed_type)Futon                         -9.072e+00  6.376e+00  -1.423 0.154777    
as.factor(bed_type)Pull-out Sofa                 -6.165e+00  6.619e+00  -0.931 0.351650    
as.factor(bed_type)Real Bed                      -8.878e+00  5.138e+00  -1.728 0.084018 .  
as.factor(cancellation_policy)moderate            1.640e+00  1.140e+00   1.439 0.150262    
as.factor(cancellation_policy)strict              9.220e+00  1.070e+00   8.618  < 2e-16 ***
as.factor(cancellation_policy)super_strict_30     6.561e+01  1.070e+01   6.131 8.77e-10 ***
as.factor(cancellation_policy)super_strict_60     4.021e+02  2.975e+01  13.515  < 2e-16 ***
as.factor(cleaning_fee)TRUE                      -6.956e+00  1.038e+00  -6.703 2.05e-11 ***
as.factor(city)Chicago                           -3.083e+03  7.532e+01 -40.939  < 2e-16 ***
as.factor(city)DC                                -1.007e+03  3.455e+01 -29.135  < 2e-16 ***
as.factor(city)LA                                -8.461e+03  2.220e+02 -38.113  < 2e-16 ***
```

Snippet of simple linear regression results (left) and multiple linear regression results (right)

A Q-Q plot reveals the logarithm of price roughly follows a normal distribution. Further examination of the room type, city, accommodations, bedrooms, and bathrooms variables and their effect on price is done using boxplots. Intuitively, price and bedrooms/bathrooms follow a general upward positive correlation. Because of the large size of the dataset (74,000+ observations), city and room types contain a high number of outliers that fall outside the interquartile range.



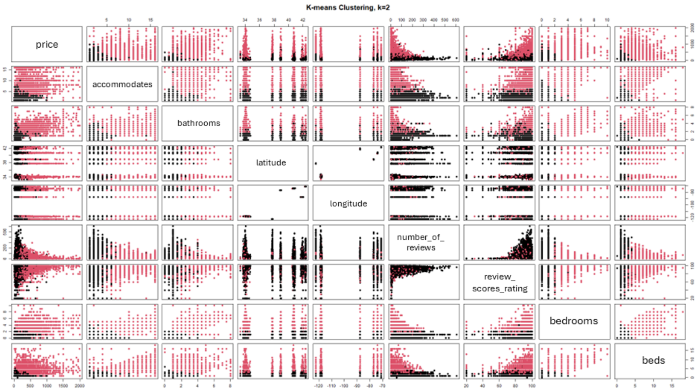
Boxplots showing relationship between price and people accommodated (left); and city (right).



Q-Q plot of the logarithm of price variable. Roughly following the slope of the 45 degree line confirms assumptions of normal distribution.

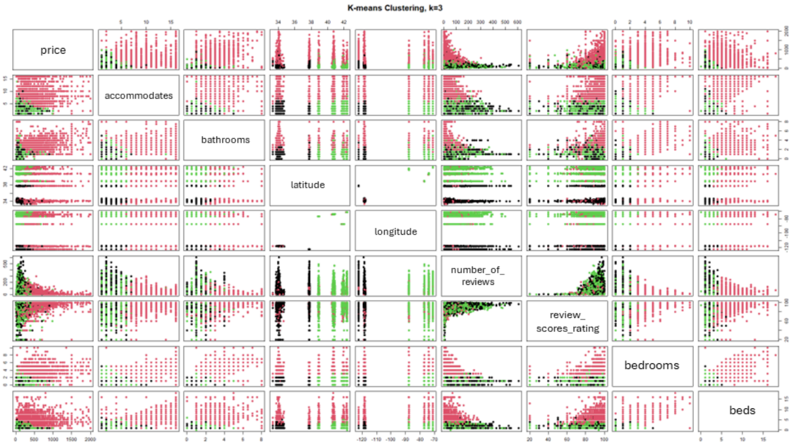
Clustering

For unsupervised learning, we performed k-means clustering on our data to gain some additional insight into the data. Since k-means mostly applies to numerical data, the data was filtered to include just the numeric fields, and excluded id since this is an arbitrary value. The result was 9 numeric variables: price, accommodates, bathrooms, latitude, longitude, number_of_reviews, review_scores_rating, bedrooms, and beds. For the number of clusters, we used k values of 2, 3, and 4 to see what clusters could be created.



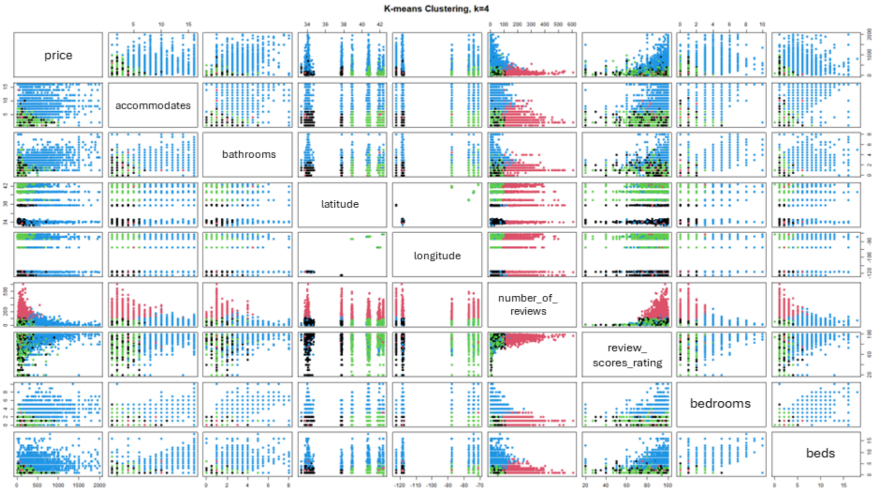
K-means clustering for k=2

From the result for k=2, we can see that the two clusters primarily correspond with larger listings and smaller listings physically. Those in the pink cluster all have high values of accommodates, bedrooms, and beds, which are indicative of larger listings. This makes sense because a listing that accommodates more people should have more bedrooms and beds.



K-means clustering for k=3

When k is 3, the pink cluster still encompasses the larger listings, while the smaller listings have been divided between the green and black clusters. The main distinction for the black cluster can be spotted in latitude and longitude; when both latitude and longitude are lower (meaning further south and west), points in the black cluster are gathered. This location corresponds with cities in California (SF and LA), so this cluster encompasses smaller places in California.



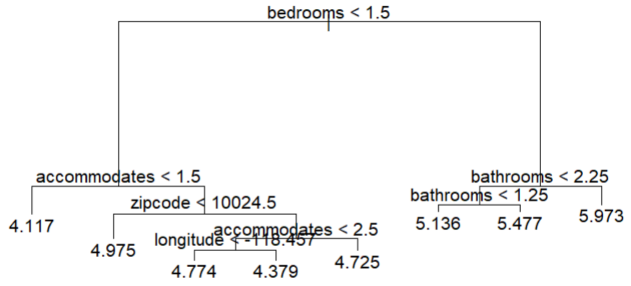
K-means clustering for k=4

Finally when k=4, the blue cluster still represents the larger listings. The pink cluster now captures (smaller) places that have a lot of reviews. Finally, the distinction between the green and the black ones appears to be the latitude and longitude, with the black cluster corresponding to places in the southwest, and the green cluster in the northeast.

From performing clustering, we can see different ways to differentiate between Airbnbs. In particular, all 3 values of k grouped together larger listings, suggesting that these have many factors in common. Beyond size, listings can also be clustered by location and number of reviews,, which are both important factors when someone is deciding where to stay.

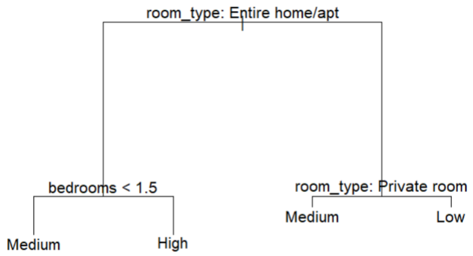
Regression Tree Analysis

The regression tree model predicts Airbnb listing prices by first considering the number of bedrooms. Listings with fewer bedrooms split into categories based on guest capacity and location, yielding lower price predictions. Listings with more bedrooms are differentiated by the number of bathrooms, with a greater number of bathrooms corresponding to higher price predictions. The model provides log price estimates that range from approximately 4.117 to 5.973, which can be converted to actual prices through exponentiation.



Classification Tree Analysis

The classification tree model simplifies the prediction of Airbnb listing prices into 'Low', 'Medium', and 'High' categories based on room type and the number of bedrooms. Listings are initially split into 'Entire home/apt' and 'Private room', with the latter consistently falling into the 'Low' price category. 'Entire home/apt' listings are further divided by bedroom count, with those having fewer than 1.5 bedrooms classified as 'Medium' and those with more as 'High'. This model prioritizes interpretability by using a minimal number of variables, which could limit its predictive accuracy but enhances its simplicity and ease of use.



Keyword Analysis

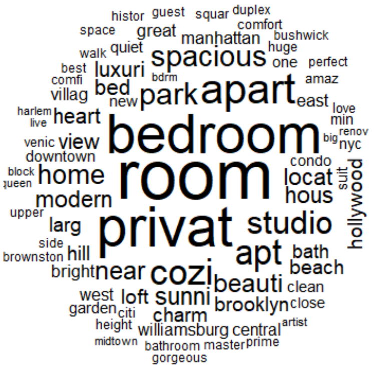
When looking at Airbnb listings, each consists of a title, description, pictures, and a price. These are the four components of each listing on Airbnb. However, we realize that these four components can generally be applied when selling just about anything. When everyone has the same access to these components - what sets one listing apart from another? What can hosts do to stand out amongst the rest? Which factor(s) is the most important? As hosts try to make the most out of their hard-earned property, these are questions that they must know the answer to, as they can be the difference between a sale or none at all.

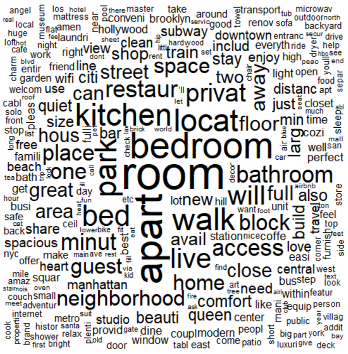
If we take a closer look at the four core components, two of them are entirely text-based, and are what provide customers with the most information about what they are potentially renting. The title has to be attractive and informative enough at a glance so customers can mentally filter out listings when skimming through. When they choose to look at the listing, the description needs to accurately and gracefully describe the property - sell the listing and make it attractive to customers while providing as many details as possible. Compared to pictures, pictures are supplementary information and only provide a visual glimpse into the state of the property, while words can most accurately describe the setting with detail. In essence, what words do hosts most commonly use to describe their properties in hopes of making a sale?

To perform text-mining analysis on the listing titles and descriptions, we first need to break down all of the words from all listings into their most basic forms. To do this, after we read the listing titles and descriptions into R, we turned all letters lowercase, removed punctuations/white space/numbers/stop words, and then stemmed the document into its most basic English words. From here, we can start doing some analysis.

For the titles, we found that most of them can be broken into three parts - the first being adjectives such as private, cozy, spacious, beautiful, and modern. These words attempt to give customers a feel for the property, and if that style interests them or not. The second part is an indicator of what type of property it is such as a bedroom, basement, apartment, home, etc…The third part is a toss-up between describing where in the city it is located, and the word "near". Describing where it is located can help customers decide if that listing is worth pursuing since most travelers have an idea of WHAT they want to do, not where they want to stay. Customers might choose to find listings close to where their planned activities are. The word "near" being a common occurrence suggests that the location at a glance is one of the most important things customers consider. Customers often visit during events, and they may not be familiar with the local area. As a result, when customers look at the title and see that the property is within walking distance of an event or transit, this makes the listing much more attractive. An example listing title for New York would be "Cozy apartment in East Village".

For descriptions, hosts have the freedom to add more to their listing and describe their property as best they can. The description often starts with a copy-paste of the title and instantly goes into some deeper elaboration. From the common words, again, the description often mentions what is around the area, and what is near the location. The next thing that descriptions most commonly include is the amenities, as the next most common words are kitchen, laundry, wifi, bed, and bath. The kitchen, laundry, and wifi seem to be a popular feature for travelers, as it most importantly offers options for them. Kitchens and on-suite laundry help put travelers at ease as they know that they don't have to actively think about food, clothing, etc. Things like wifi and other comforts that we have grown to have at home are always useful to have at convenience while traveling, making listings more attractive. Beds and bathrooms are simply just accommodations and are something that travelers expect to have out of a listing. On top of some more adjectives to physically describe the place, the next most used words are space/spacious, square, and large. This suggests that travelers are looking to get the most space for the price they pay. Some other words such as backyard, closet, and rooms (presumably from the master bedroom) can also be helpful to customers to understand how the square footage is broken down. The following are the word clouds of the most commonly used words in the titles (left) and descriptions (right):





Overall, we can see that of the four components, the title and description are by far the most important factors in a listing. While that may not definitively answer what puts one listing above another, this information can help hosts understand how they need to describe their property to make their listing unique, and how they can potentially attract more customers. With all things considered, this can also help hosts realize the true value of their property, and help them price accordingly.

Results and Discussion

Returning to the business problem and our recommendation backed by analysis, there are evidently many factors that must be considered in setting a competitive price for an Airbnb property. Variables relating to the physical size of the property such as 'accommodates', 'bathrooms', 'bed_type', 'bedrooms', and 'beds', followed a consistent linear positive correlation with the caveat being numerous outliers. Apartment, condo, and house are the most common property types. Based on the dataset and linear regression model, villas, timeshares, and earth houses ranked as the most expensive on average, while dorms, hostels, and huts were the most affordable. We recommend that hosts focus on these factors describing the property itself, as it has the largest impact on determining the price of the property. The city and location also played a factor, as each property is subject to its local real estate climate - we recommend that hosts price accordingly. When it comes to listing attractiveness, we recommend that hosts make their listing look as attractive as possible - sell the property as a place that travelers would want to call home for some time. This would include paying attention to how you word the title and description and paying attention to the proximity to local transit, seasonal demand, and any special events happening during their stay. All of these factors can contribute to a higher price.

There are several limitations to the dataset and methods used in this project. One limitation of the dataset is geographical coverage, which is confined to only six major cities in the US. This overlooks trends specific to other regions and countries and limits global applicability. Additionally, useful details such as the number of individuals renting each booking and the absence of time series data limit a more complete understanding of demand dynamics and seasonal trends within the Airbnb rental market. Additional information such as local regulations and economic conditions would help to facilitate a more comprehensive understanding of the Airbnb rental property market and provide better insight into the business challenge. Furthermore, some inconsistencies in the listings and subjective text-based variables can introduce bias and inaccuracies into our analysis. These limitations reinforce the importance of critically interpreting our findings and considering the ethical implications of our data analysis and results, as they are used in the context of the business problem.

Future work for this analysis project includes gathering more data to apply our findings to different contexts. A more granular analysis of each predictor variable and the impact on price through regression and more advanced machine learning models will promote a better understanding of the dataset. The ability to analyze Airbnb listings from different countries or even regions in the US will be informative for hosts outside of the major metropolitan areas covered in the current dataset. Future work to collect and analyze time series data for the listings, including times when the listing is vacant and occupied, can potentially inform hosts of seasonal trends in demand and adjust their pricing and expectations accordingly. Lastly, as the rental property market is closely tied to economic data and legislation, further research on different policies and bylaws surrounding Airbnb property rentals, and how they can affect the supply of listings and in turn, the price, is another future direction.

Reference

RUPINDER SINGH RANA. "Airbnb Price Dataset." www.kaggle.com, 2 Feb. 2024, www.kaggle.com/datasets/rupindersinghrana/airbnb-price-dataset. Accessed 2 Apr. 2024.