Developing an AirBnB Score: An Analysis on the Importance of Amenities and Location

Group Leader Name: Sarang Deshpande

Group members: Eric Sheehan; Yvanna Tchomnou

Introduction:

For the purposes of this project, we will be conducting an analysis of AirBnBs included in the Los Angeles dataset offered by AirBnB. With a population of 3.9 million people, it is the 2nd most populated city in America as of July 2022, which is when the last census data was collected. This places them only behind New York in terms of population. LA's public transportation is highly ranked amongst the urban centers of America, encompassing a majority of the city. They offer six lines serving 101 stations throughout LA. Their Metro Line A is the longest light rail in the world, spanning 50 miles from Azusa to Long Beach. In addition to trains, LA has two rapid bus transit lines that cover 29 stations dispersed throughout the city. On top of the rapid bus transit lines, LA has normal buses that operate along 117 more routes. As a result, traveling for tourists is not overly complicated.

Due to the expansive size of Los Angeles, there are many tourist destinations within LA, such as Venice Beach, Griffith Observatory, Disneyland & Universal Studios, the Walk of Fame, Rodeo Drive, and so many more. As a result, the demand for AirBnBs is high. In 2023, AirBnB released an article based on data from 2022, stating that "hosts in Los Angeles earned over $375 million, with the typical host earning over $24,000." Given that the demand has only grown since 2022, hosts are especially attracted to the possible passive income opportunities provided by hosting AirBnBs in the City of Angels.

Based on the background research conducted, as well as personal experience, we hypothesized that amenities such as WiFi, toiletries, cookware, and parking lead to higher customer satisfaction with the property. In addition, we also believe that the proximity of an AirBnB to major attractions in the city can yield higher customer satisfaction. From these hypotheses, we developed two research questions to guide us in the development of our methodology. The first question we developed was: *What should a prospective AirBnB owner consider investing in for a potentially higher-rated AirBnB property (amenities, communication, locations, etc)?* The second question we developed was: *How do certain amenity offerings affect the overall profitability of potential AirBnB listings (price, availability)?*

Methodology:

In order to answer our research question, we will need lots of information regarding listings in the Los Angeles area, as well as review data. Specifically, we want to look into general information like the number of bedrooms, bathrooms, types of residences, neighborhood or location, and even more specific attributes like amenities. In order to obtain this information, we will be using the detailed listing data and detailed review data for Los Angeles with the provided link. To collect the data properly to transfer to Colab, we will need to download the LA listings.csv.gz and LA reviews.csv.gz files and extract them into their own CSV files. Once they are properly formatted as CSV files, we are ready to upload the files to Google Drive, and pre-process the data in Colab.

Question 1: What should a prospective AirBnB owner consider investing in for a potentially higher-rated AirBnB property (amenities, communication, locations, etc)?

To answer Question 1, we conducted two analysis techniques: Clustering Analysis & Text Analytics.

Clustering Analytics

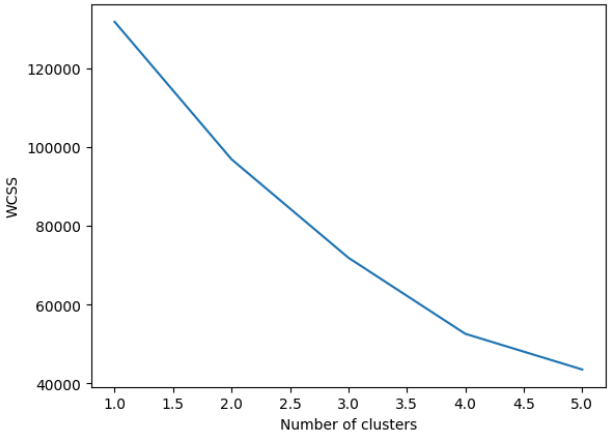
Pre-Processing:

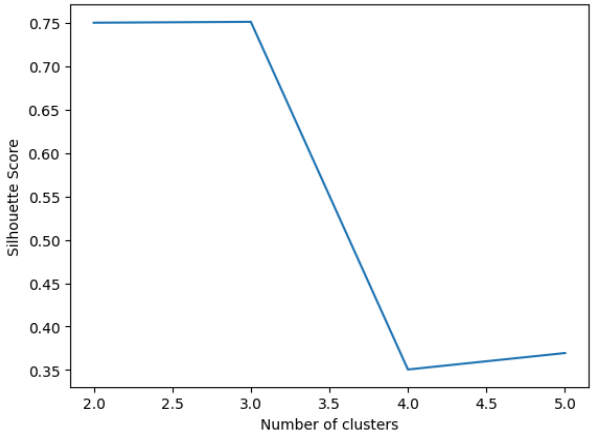
For the purposes of Clustering, we only used data from the *listings* dataset, which offers information such as price, review scores, and amenities. After reading the *listings* dataset, we wanted to use functions like df.head() and df.describe() to get a better idea of what kind of data and variables we are dealing with for the CSV file we created. The next step is addressing null values. Initially, we had approximately 45,000 different observations. We want to look at what columns have null values using the df.isna().sum() to narrow down what columns we need to look into when it comes to addressing null values. Once we know what null values we are dealing with, we can address the null values by either replacing certain null values, removing any rows or columns with null values, or maybe a combination of both. For the purposes of this project, we simply dropped the null values, bringing us to approximately 33,000 observations. From there, we also dropped all variables that held qualitative data, such as urls or descriptions. The only qualitative data we left in the dataset was amenities so that we could derive an amenities score later in the project.

To derive the amenities score, we first needed to determine what amenities the properties have. After converting the text in the amenities column to lowercase, we created new columns for each amenity, using True or False for each property for each amenity. We then replaced the True and False values to 1 & 0 to make them integers. From there, we iterated through the properties, averaging the values of each amenity column to come up with the amenity score for use in the analysis. Once that is done, we remove the dummy variables for each amenity to avoid repetitive data. The last step in the pre-processing step is to use the StandardScaler method to standardize the dataset.

Data Analysis:

For the purposes of clustering, we will be using the K-Means algorithm, as opposed to hierarchical or density-based, due to needs of the project and the need for an understanding as to the importance of variables. For our input variables, we will be using *price*, *review_scores_rating*, *review_scores_location*, and our previously derived variable *amenity_score*. We will be using the Elbow Method and the Silhouette Score methods to determine the optimal K-value.





Based on the WCSS score and the Silhouette score, we can determine that 2 or 3 clusters would be appropriate for the model. For the purposes of the study, we chose to use 2 clusters moving forward.

Clustering: Applying the K-means algorithm, which partitions the listings into K clusters by:

- Initializing K centroids randomly.
- Assigning each point (listing) to the nearest centroid.
- Recomputing the centroids as the mean of points assigned to each cluster.
- Repeating the assignment and centroid calculation steps until convergence or a set number of iterations is reached.

While developing our clustering model, we used a random_state of 100 and n_init of 25. After running the model, we found two clusters, with 32039 data points falling within one cluster and 902 data points falling within the other.

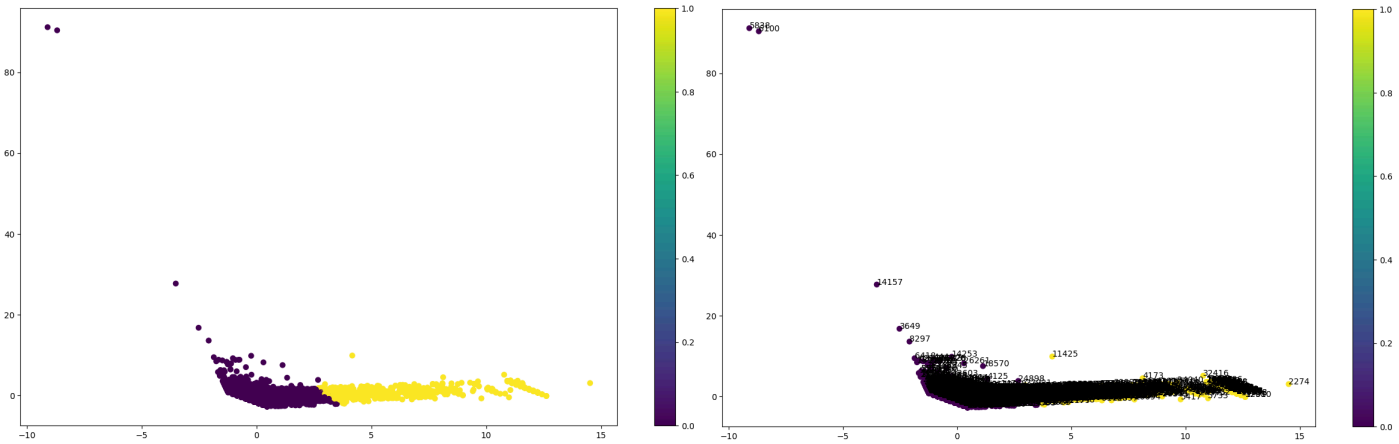
After assigning the data points to different clusters, we utilized PCA to reduce the dimensions of the clusters from four to two to aid in the visualization process.

Evaluation:

By utilizing the silhouette score method to select the optimal number of clusters, we guaranteed that our model had the highest possible silhouette score. For the model generated, we got a silhouette score of 0.7502224932758973, indicating that the clusters are clearly distinguishable and have a relatively high degree of separation from each other.

Result:

The visualizations below are identical, with the right one showcasing the indices of every data point on the visualization.



To gain a profile on the different clusters, we found the data of 2 indices from each cluster. The purple cluster generally had high ratings for the property and for location, as well as high prices. The yellow cluster generally had relatively lower pricing, but also had lower scores for property and location. The amenity scores varied drastically within both clusters and didn't play a significant role in the data distribution.

Text Analytics

Pre-Processing:

For our Text Analytics portion, we conducted extensive pre-processing steps on the *reviews.csv* dataset, which originally contained approximately 1.83 million reviews. The initial step involved removing null values to ensure data completeness. Utilizing Excel PowerQueries, we filtered the dataset to include only reviews from the year 2023, effectively reducing the dataset size to 108,129 entries. Following this, we converted the data type of the *'comments'* column from object to string and employed the 'langdetect' library, a Python adaptation of Google's language-detection tool, to identify and retain reviews written exclusively in English. This language filtering was executed using parallel processing techniques to increase efficiency and reduce overall runtimes, resulting in a new dataset ("*updated_reviews.csv*") that consisted solely of English reviews from 2023.

The next phase included several key preprocessing tasks tailored for text analytics. Initially, we checked for and removed any remaining null values. The text was then tokenized, converted to lowercase, and stripped of English stopwords to refine the dataset further and lastly, stemming was applied to reduce words to their root forms to optimize the data for the upcoming analysis processes. These preprocessed texts were pivotal in generating word clouds and conducting topic modeling. Lastly, utilizing the results from sentiment analysis, we also segmented the data into two distinct groups based on their polarity scores, indicating the sentiment expressed: positive and negative reviews, categorizing reviews with a polarity score below zero with negative reviews and vice versa. This division facilitated a more targeted analysis of the sentiments in the dataset to be used for the topic modeling analysis conducted, enabling nuanced insights into customer perceptions and experiences documented within the reviews.

Data Analysis:

For the purposes of text analytics, we performed Sentence-level Sentiment Analysis using vaderSentiment's *SentimentIntensityAnalyzer* on the pre-processed reviews in the *'comments'* column, yielding a polarity score of the review text (positive, negative, neutral), with scores ranging from -1 to 1. We then utilized the sentiment scores to segment the dataset into 2 data frames to separate positive and negative reviews that will be used in our word clouds and topic modeling.

By analyzing sentiment scores, we can classify reviews into positive and negative categories based on sentence-level polarity scores. This will allow us to understand the overall customer satisfaction and pinpoint the different aspects of the service that impact guest experiences positively or negatively. In addition, We will create word clouds with the 150 (i.e., max_words=150) most used tokens and topic modeling utilizing LDA on the text data from positive and negative reviews separately to reveal the key drivers behind these differing sentiments, allowing us to understand common patterns in customer feedback on areas where listings are lacking or thriving.

Evaluation:

To evaluate the predicted outcomes of the LDA topic modeling, we utilized the coherence score method, which measures how well a topic is 'supported' by a text set (called reference corpus). For the topic modeling results generated, the recommended number of topics for positive reviews was 20 topics as it yielded the highest coherence score of 0.5519954, and for the negative reviews, the recommended number of topics is 10 topics with a coherence score of 0.5186842.

Positive Sentiment Reviews

```
Coherence score 1 - 5 topics:  0.5095145425747948
Coherence score 2 - 10 topics:  0.5484604307959654
Coherence score 3 - 15 topics:  0.5450815014226736
Coherence score 4 - 20 topics:  0.5519954169162736
```

Negative Sentiment Reviews

```
Coherence score 1 - 5 topics:  0.5048531244656197
Coherence score 2 - 10 topics:  0.5186842362034787
Coherence score 3 - 15 topics:  0.5064721845551493
Coherence score 4 - 20 topics:  0.5008429398876051
```

Result:

Sentiment Analysis

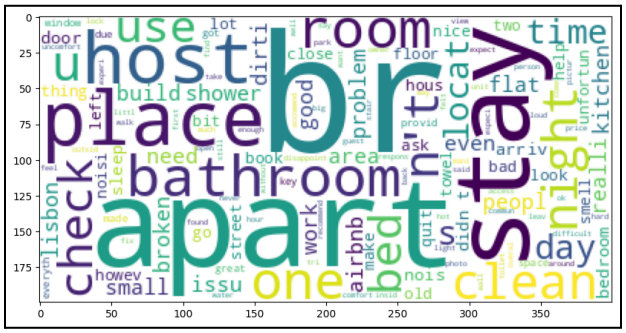
```
Top 5 Positive Sentiment Posts:
                                                 comments  sentiment_score_vader
16302  We loved our stay at this beautiful apartment....                  0.9987
25784  We had an amazing time staying in Miguels airb...                  0.9985
28541  Thanks Ana and Pedro, we really enjoyed our st...                  0.9985
69571  There is probably not Many lovely places Like ...                  0.9982
44504  This is my second time staying at one of João...                   0.9981

Bottom 5 Negative Sentiment Posts:
                                                 comments  sentiment_score_vader
1549    1/10 rating<br/>The BAD - smells bad, unclean ...                 -0.9933
104823  I regret after staying at 30+ Airbnbs and neve...                 -0.9927
4798    Unlike  advertised this is no luxury apartment...                 -0.9862
24155   There is promise here but it needs a LOT of at...                 -0.9857
96009   There is a bad black mold issue in this place ...                 -0.9851
```

After completing the sentiment analysis and segmentation of the reviews, based on the sentiment scores we found that 1,975 reviews were classified as “negative” reviews versus the 105,317 reviews that were classified as “positive”, which can indicate that for the most part, individuals enjoyed their AirBnB stays in L.A. about 98% of the time. Furthermore, the image above showcases the top & bottom 5 positive and negative sentiment review posts in which some of the common themes regard property cleanliness & accuracy, as well as experiences, had with the AirBnB hosts.

Word Clouds

Positive Reviews

Negative Reviews



The word clouds above were generated for the top 150 most used tokens in positive & negative reviews. Regarding positive reviews, the word clouds showcased phrases such as “great host”, “walk distanc”,

“perfect locat”, revealing the guests value properties in a good location with proximity to the attractions as well as their experiences with the AirBnB host. Concerning the negative reviews, some of the phrases showcased included “clean”, “small”, “uncomfort”, and “noisi”, revealing that some of the common issues guests faced in unfavorable reviews dealt with the overall cleanliness of the property as well as the size of the noise level of the area in which the listing was situated.

Topic Modeling

```
Topic: 0
Words: 0.017*"bar" + 0.013*"amaz" + 0.013*"thabk" + 0.012*"restaur" + 0.012*"park" + 0.010*"14" + 0.010*"," + 0.010*"drain" + 0.010*"great" + 0.010*"spend"
Topic: 1
Words: 0.030*"restaur" + 0.030*"grate" + 0.026*"authent" + 0.023*"communicative@check-in" + 0.019*"locat" + 0.018*"emili" + 0.017*"meant" + 0.016*"don@t" + 0.015*"citi" + 0.015*"arriv"
Topic: 2
Words: 0.046*"proa" + 0.033*"quick" + 0.032*"everi" + 0.021*"by-" + 0.020*"powder" + 0.017*"though" + 0.017*"walkabl" + 0.016*"parent" + 0.015*"everywher" + 0.014*")"
Topic: 3
Words: 0.040*"consid" + 0.021*"museum" + 0.017*"w/" + 0.015*"apart" + 0.014*"con" + 0.014*"annoy" + 0.012*"flat" + 0.010*"difficult" + 0.010*"amaz" + 0.008*"joao"
Topic: 4
Words: 0.065*"cover" + 0.029*"choic" + 0.024*"dusti" + 0.022*"somehow" + 0.020*"huge" + 0.017*"famou" + 0.014*"full" + 0.014*"responsive." + 0.013*"attent" + 0.013*"location-"
Topic: 5
Words: 0.037*"night" + 0.034*"famou" + 0.030*"," + 0.024*"descript" + 0.018*"contain" + 0.018*"cute" + 0.015*"lisbon" + 0.015*"need" + 0.013*"gorgeou" + 0.013*"well"
Topic: 6
Words: 0.046*"shelf" + 0.041*"also" + 0.036*"person" + 0.033*"christina" + 0.032*"routin" + 0.024*"full" + 0.023*"nearbi" + 0.018*"except" + 0.013*"con" + 0.013*","
Topic: 7
Words: 0.040*"still" + 0.029*"," + 0.025*"nevertheless" + 0.022*"full" + 0.018*"'ll" + 0.017*"3rd" + 0.016*"great" + 0.014*"sure" + 0.013*"airflow" + 0.012*"cute"
Topic: 8
Words: 0.048*"calm" + 0.024*"absolut" + 0.018*"afield" + 0.015*"plu" + 0.015*"," + 0.013*"spent" + 0.013*"tuktuk" + 0.012*"june" + 0.012*"you@r" + 0.011*"would"
Topic: 9
Words: 0.070*"plu" + 0.051*"need" + 0.035*"be.w" + 0.033*"," + 0.032*"check-in" + 0.031*"cute" + 0.031*"bit" + 0.028*"haven@t" + 0.024*"far" + 0.020*"don@t"
Topic: 10
Words: 0.101*"check-in" + 0.077*"cute" + 0.071*"authent" + 0.055*")" + 0.044*"great" + 0.039*"," + 0.036*"parent" + 0.036*"uber" + 0.035*"emili" + 0.029*"restaur"
Topic: 11
Words: 0.082*"cow" + 0.045*"zone" + 0.043*"short" + 0.039*"6" + 0.033*"immedi" + 0.031*"met" + 0.030*"includ" + 0.025*"head" + 0.023*"us" + 0.022*"side"
Topic: 12
Words: 0.087*"locat" + 0.069*"restaur" + 0.066*"park" + 0.050*"check-in" + 0.042*"cute" + 0.039*"," + 0.037*"great" + 0.034*"commun" + 0.033*"gave" + 0.030*"don@t"
Topic: 13
Words: 0.070*"restaur" + 0.065*"check-in" + 0.062*"great" + 0.047*"cute" + 0.042*"it@" + 0.037*"help" + 0.036*"love" + 0.034*"," + 0.032*"stun" + 0.028*"don@t"
Topic: 14
Words: 0.059*"afield" + 0.037*"etc" + 0.032*"restaur" + 0.032*"notch" + 0.029*"arriv" + 0.025*"wish" + 0.023*"great" + 0.022*"emili" + 0.022*"check-in" + 0.021*"closer"
Topic: 15
Words: 0.489*"apart" + 0.019*"," + 0.008*"host" + 0.008*"plu" + 0.008*")" + 0.006*"parent" + 0.006*"locat" + 0.005*"love" + 0.005*"joao" + 0.005*"walkabl"
Topic: 16
Words: 0.079*"coffe" + 0.037*"balconi" + 0.030*"cute" + 0.027*"exactli" + 0.026*"good" + 0.025*"check-in" + 0.023*"much" + 0.021*"parent" + 0.021*"restaur" + 0.020*"christina@"
Topic: 17
Words: 0.078*"host" + 0.059*"wonder" + 0.047*"though" + 0.031*"," + 0.026*"alley" + 0.026*"mani" + 0.024*"i@d" + 0.019*"joao" + 0.018*"quieter" + 0.017*"simpli"
Topic: 18
Words: 0.031*"help" + 0.027*"quiet" + 0.026*"pictur" + 0.019*"near" + 0.018*"restaur" + 0.014*"great" + 0.013*"clearli" + 0.012*"etc" + 0.011*"plu" + 0.009*"place."
Topic: 19
Words: 0.040*"tight" + 0.037*"nonsens" + 0.034*"drove" + 0.029*"kitchen" + 0.027*"nearest" + 0.023*"octob" + 0.017*"apart" + 0.016*"time" + 0.016*"mid" + 0.016*"uneventful."
```

Based on this output for positive reviews, we prepared potential interpretations of select topics below:

- Topic 0: Focuses on amenities and leisure activities near Airbnb locations. Words like "bar," "restaurant," "park," and "spend" suggest guests appreciated nearby dining and recreational options.
- Topic 1: Emphasizes the importance of authentic experiences and effective communication, particularly in the check-in process. Words like "authentic," "communicative," and "location" highlight the value guests place on these aspects.
- Topic 2: Suggests a preference for Airbnb locations with easy accessibility and quick responses from hosts. The terms "quick," "walkable," and "everywhere" indicate that guests value convenience and prompt service.
- Topic 9 to 11: Continue to emphasize the importance of check-in processes, cute decor, and the authenticity of the stay, reflecting a consistent theme across many topics of valuing genuine, hassle-free experiences.
- Topic 15: Stands out by heavily focusing on the "apart" aspect, perhaps indicating a high appreciation for the privacy or separation from typical hotel experiences that apartments offer.
- Topic 16 to 18: Illustrate preferences for specific amenities like balconies and coffee, the importance of quiet environments, and overall hospitality—highlighting varied but specific guest needs and desires.

Overall, these topics provide insights into what aspects of their stays are most impactful to guests, including the importance of location, authentic local experiences, responsive hosts, and comfortable, appealing accommodations. These insights can be particularly useful for Airbnb hosts looking to improve guest satisfaction and highlight features in their listings that align with these preferences.

```
Topic: 0
Words: 0.018*"night" + 0.014*"heavi" + 0.013*"open" + 0.011*"window" + 0.008*"." + 0.008*"happi" + 0.008*"cot" + 0.008*"avail" + 0.008*"attic" + 0.008*"reach"
Topic: 1
Words: 0.024*"quiet" + 0.014*"cot" + 0.013*"heavi" + 0.012*"open" + 0.009*"expens" + 0.008*"attic" + 0.007*"challeng" + 0.007*"window" + 0.007*"great" + 0.007*"small"
Topic: 2
Words: 0.014*"luggag" + 0.011*"avail" + 0.008*"heavi" + 0.007*"cot" + 0.007*"open" + 0.006*"." + 0.006*"start" + 0.006*"pull" + 0.005*"three" + 0.005*"paço"
Topic: 3
Words: 0.019*"." + 0.017*"open" + 0.016*"cot" + 0.014*"quiet" + 0.012*"'" + 0.010*"three" + 0.010*"water" + 0.009*"reach" + 0.009*"expens" + 0.008*"work"
Topic: 4
Words: 0.093*"quiet" + 0.017*"." + 0.009*"night" + 0.009*"window" + 0.008*"attic" + 0.008*"cot" + 0.007*"great" + 0.007*"luggag" + 0.006*"pull" + 0.006*"resolv"
Topic: 5
Words: 0.017*"quiet" + 0.017*"." + 0.014*"great" + 0.011*"heavi" + 0.008*"open" + 0.008*"conveni" + 0.007*"downtown" + 0.006*"decor" + 0.006*"window" + 0.006*"challeng"
Topic: 6
Words: 0.018*"quiet" + 0.011*"open" + 0.011*")" + 0.011*"." + 0.009*"cot" + 0.008*"help" + 0.007*"hot" + 0.007*"cute" + 0.007*"heavi" + 0.007*"airport"
Topic: 7
Words: 0.025*"." + 0.020*"great" + 0.013*"heavi" + 0.013*"often" + 0.011*"get" + 0.009*"key" + 0.009*")" + 0.007*"rais" + 0.007*"middl" + 0.007*"3rd"
Topic: 8
Words: 0.017*"." + 0.015*"hope" + 0.014*"great" + 0.012*"attic" + 0.012*"luggag" + 0.012*"heavi" + 0.009*"close" + 0.009*"open" + 0.009*"avail" + 0.008*")"
Topic: 9
Words: 0.018*"great" + 0.014*"heavi" + 0.013*"challeng" + 0.011*"open" + 0.009*")" + 0.009*"said" + 0.008*"downtown" + 0.008*"switch" + 0.007*"danger" + 0.007*"maria"
```

Based on this output for negative reviews, we prepared potential interpretations of select topics below:

- Topic 1: Discusses themes of quietness and challenges with accommodation, possibly referring to noise issues and the physical layout or accessibility of the property, such as attic spaces.
- Topic 2: Highlights issues related to luggage, accessibility, and the convenience of opening windows or doors. This might relate to guests having trouble with heavy luggage in places that require more effort to access.
- Topic 3: Suggests some dissatisfaction or notable mentions of features like "open," "quiet," "expensive," and "reach," indicating that while some guests appreciated the quiet, they might have found some aspects of the accommodation expensive or difficult to access.
- Topic 4: Strongly emphasizes the quietness of the location, which is a significant positive attribute for guests. However, there are mentions of challenges related to "night," "window," and "luggage," suggesting some issues with night-time comfort or noise.
- Topic 6: Again, focuses on the quiet environment but also mentions practical elements like "help," "hot," and "airport," suggesting that these aspects were significant in guests' experiences, possibly referring to the temperature control and proximity to the airport.
- Topic 9: Deals with great experiences juxtaposed with challenges ("heavy," "challenging," "danger"), possibly indicating some safety or accessibility concerns in the accommodation.

Overall, these topics suggest that while many guests found their stays satisfactory, particularly valuing quietness and great overall experiences, they also frequently encountered practical difficulties related to accommodation features like attic rooms, heavy items, and accessibility. These insights can be helpful for hosts to address specific guest needs and improve aspects of their properties for future stays.

Question 2: How do certain amenity offerings affect the overall profitability of potential AirBnB listings (price, availability)?

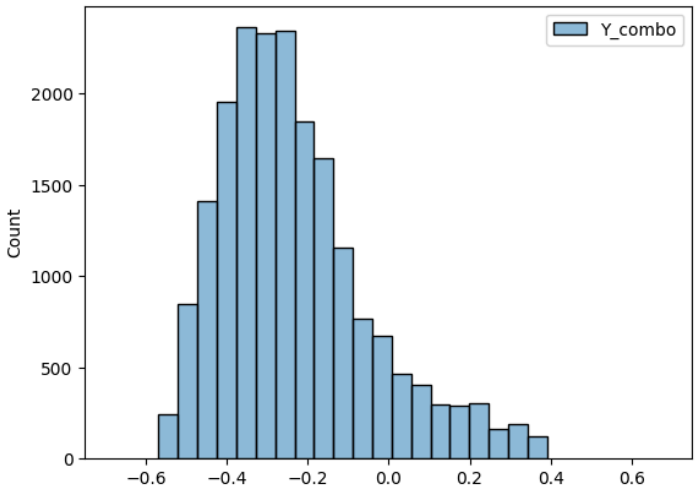
Pre-Processing:

For this question, we knew we planned to use linear regression with different listing attributes and amenities to predict a target variable that represents profitability from the AirBnB owner's perspective. Therefore, a lot of the pre-processing for this analysis was quantifying variables that were previously represented through text. To start, the "bedrooms" and "bathrooms" variables were both filled with only null values in the dataset for Los Angeles, so we had to be creative to get this information. Thankfully, there were two other variables, "beds" and "accommodates," that can be used to make up for "bedrooms" not being in the dataset. Additionally, "bathroom_text" contains a text description of how many bathrooms are in the unit, like "2 full bathrooms" or "3 and a half bathrooms." Using this, I was able to

create a numeric variable that represents how many bathrooms are in the unit from the text in that column within 0.5 of a bathroom. Lastly, I used str.contains() to measure if a certain amenity was included in the "amenities" column, which was a big string value with a list of amenities included for each listing. I created dummy variables for if the amenity was included in the list for Wi-Fi, Kitchen, Pool, Parking, Balcony, Gym, Laundry, TV, and Temperature Control. Once I was able to do this and create dummy variables for property type, the data just needed to have nulls removed, the data filtered down to the Los Angeles area, and create my target variable. The target variable I created was a reflection of what AirBnB property owners are looking for - price of the listing and how many nights of the next year it was booked for (this is just 365 - "availability_365"). I combined these two numbers using PCA and used this value as the target variable for the linear regression analysis. While this leads to some trouble for interpretation purposes, the goal is not to predict the price and availability of a listing, it is to provide recommendations on what characteristics can lead to increased profitability for an AirBnB listing. Lastly, I scaled the X variables and checked for multicollinearity and dropped "beds" which had a strong correlation with "accommodates." Now, after removing outliers within the target variable, the data is ready for linear regression analysis.

Data Analysis:

- The first linear regression run included 13 independent variables:
    - Accommodates, Bathrooms, Entire Home/Apt Room Type, Balcony/Patio, Shared Room Type, Laundry, Kitchen, Pool, Temperature Control, Parking, TV, Gym, and Wi-Fi
- However, one variable, Wi-Fi, had a p-value over 0.05, so I removed this variable and ran the analysis again. Below are the coefficients for each variable, along with the type of variable
    - Accommodates (0.087) - float
    - Bathrooms (0.057) - float
    - Entire Home/Apt Room Type (0.036) - binary
    - Balcony (0.021) - binary
    - Shared Room Type (-0.014) - binary
    - Laundry (-0.013) - binary
    - Kitchen (-0.013) - binary
    - Pool (0.012) - binary
    - Temperature Control (0.012) - binary
    - Parking (0.008) - binary
    - TV (0.006) - binary
    - Gym (0.003) - binary
- To put these coefficients in perspective, below is the distribution of the target variable



o

Evaluation:

The final model had an R-Squared value of 0.431 and an adjusted R-squared value of 0.430. This tells us a couple of things - the first is that 43.1% of the variation in the target variable can be explained by the independent variables in the model. While this is not perfect, it does show that these types of variables are a major part of AirBnB listing value. Location and proximity to attractions and restaurants and stores are factors that may make up the other 56.9% of the variation in the price and number of bookings over the next year. Additionally, with R-Squared and adjusted R-Squared being extremely close to each other, we can be confident that overfitting is not a concern with the model.

Additionally, we see a Mean Squared Error of 0.019, which means that there is very little variance in the residuals for the model (the difference between the predicted and actual values for the target variable). Likewise, we see a low standard deviation of the residuals in the model with a Root Mean Squared Error of 0.140. Lastly, the Mean Absolute Error is 0.106, meaning the residuals in the model have an average value of 0.106. All of these values are extremely low, but it is important to note that the target variable is very small, so this has a major impact on the values of the residuals. With that being said, the is still a range of around 1.0 for the target variable, so these numbers are still quite small and show that model performance is good when it comes to the variance and standard deviation of the residuals.

Result:

Now that we know the model has some validity to it, we can use the coefficients in the model as ratings of feature importance to find what attributes in AirBnB listings. According to the model, the two most helpful variables for raising profitability are the two non-binary variables, “accommodates” and “bathrooms.” This can also be interpreted as bedrooms and bathrooms, since “accommodates” reflects how many residents the AirBnB can hold. These two variables and their coefficients tell us what we would assume with the attributes of a listing - a higher number of bedrooms and bathrooms will increase the value of the listing. Additionally, entire home/appt room type is definitely a major variable as well, and it is significantly better to rent out an entire property than to have a shared space rented out via AirBnB. As for amenities, having a balcony (or patio), pool access, temperature control (A/C and Heating), are most advantageous for improving profitability in an AirBnB listing in Los Angeles. Laundry and Kitchen both had negative coefficients, so those variables may not be as important to include in a listing as we may have thought. Lastly, Parking, TV, and Gym all have positive coefficients, but they are all under 0.009 so they don’t have much significance in the model. With that being said, it doesn’t hurt to

include these amenities in a listing if they are available. In summary, when prospective AirBnB owners are looking for a property to list with AirBnB, they should be looking for properties with lots of bedrooms and bathrooms to accommodate a higher number of guests. Additionally, they should be looking to buy an apartment or home with the intention of listing the entire property being available to rent, and it should have access to a balcony/patio, pool, and have A/C and heating available within the listing.

Conclusion:

From the analysis conducted by the team, we have determined that topics such as location, authentic experiences, responsive hosts, and comfortable accommodations are especially likely to prompt AirBnB guests to leave a review, indicating that these topics are the most important features of a property to the general AirBnB guest. In addition, the findings from the text analytics highlights common issues like accessibility and room features, providing actionable insights for Airbnb hosts to enhance guest satisfaction. Conducting the linear regression analysis only backed up these findings, emphasizing that prospective AirBnB hosts should look for the following characteristics when purchasing a property in the Los Angeles area:

- Lots of beds/bedrooms & bathrooms to accommodate more guests
- Entire home/apartment rental
- Balcony/patio and pool access
- Heating and A/C available

The primary advantage of incorporating these amenities and features is that this can lead to a higher rating, which can allow hosts to charge higher prices, in turn increasing their passive income from the AirBnb.

References:

Airbnb. (2024, January 26). Chicago and Denver rank as top cities to start hosting in the US. Airbnb Newsroom. https://news.airbnb.com/top-cities-to-list-space-on-airbnb-and-some-best-new-hosts-in-those-cities/#:~:text=Chicago%2C%20IL&text=It's%20no%20surprise%20that%20people,Host%20earning%20approximately%20%2418%2C5004.

Fiorentino, F. (n.d.). 24 top tourist attractions in Los Angeles. Touropia. https://www.touropia.com/tourist-attractions-in-los-angeles/#:~:text=Disneyland,-What%20is%20this&text=It's%20been%20the%20star%20tourist,visitors%20access%20to%20both%20parks.

Hospitality & tourism. Los Angeles County Economic Development Corporation. (2022, June 9). https://laedc.org/industry-cluster-development/hospitality-tourism/

U.S. Census Bureau quickfacts: Los Angeles City, California. (n.d.). https://www.census.gov/quickfacts/fact/table/losangelescitycalifornia/PST045222