EDA

1. **Data Overview**

   For our project, we will be using “listings” and “reviews” tables from the Boston data section on the website http://insideairbnb.com/get-the-data/, providing a thorough overview of Boston's Airbnb listings.

   The "listings" dataset features 75 attributes across 4204 entries, including 25 quantitative variables suitable for numerical analysis, and the rest being text, categories, and other non-numeric types. The "reviews" dataset offers insights through 6 attributes over user experiences on Airbnb. (see Appendix 1&2)

2. **Data Peaking and preparation**

   For our review dataset, we eliminate rows with N/A values. For listings dataset, we categorized the data, discarded columns with all nulls, non-analytical URLs, and redundant information. We also transformed 'host_since' into cumulative months, filtered out unavailable listings, and standardized 'price'. Missing values were filled with medians for numerical and modes for categorical columns, while text columns with 'Unknown' were dropped for analysis purposes.

   We set a high-rating benchmark based on the 70th percentile of 'review_scores_rating' at 4.88 and eliminated extreme outliers. Percentage columns were converted to decimals, and location coordinates were dropped. After normalizing numerical data and encoding categorical data, we combined them into a single cleaned DataFrame for analysis.

   We also employ Principal Component Analysis(PCA) in listing dataset, using 5 components as it captures 95% of the variance (Appendix 4). Meanwhile, these five components have been renamed to reflect their underlying factors based on their correlation with original attributes, which are ‘Review Activity Score’, ‘Host Experience & Quality Score’, ‘Host Duration Flexibility’, ‘Pricing Duration’ and ‘Quality Premium’. These revised component names sharpen the interpretability of our PCA, offering a lucid narrative of the dataset's characteristics.

   However, after attempting to form clusters using PCA, we found the results not ideal and challenging to interpret effectively, indicating limitations in PCA's analytic capability. Therefore, we decide to use this as our data visualization tools only during the preparation stage, and choose the top three components to present the numerical information more intuitively.

3. **Text preprocessing**

   In this part, we deal with text columns in listing dataset and review dataset respectively. We change all upper cases into lower cases, and remove all trailing spaces as well as punctuation. We also implement tokenization on all text columns as well as WordNetLemmatizer.

Analysis Process:

T-SNE

After EDA as well as PCA, we performed t-SNE with categorical attributes in listing dataset for dimension reduction. We selected ‘highly_rated’ column as target variable, and generated 3D visualization using 3 components of t-SNE and ‘highly_rated’ as color legend. The graph presents a spiral shape and seems to contain some correlation with ‘highly_rated’. Hence, we performed a hierarchical clustering due to the shape.

After having a dendrogram, we selected 2 clusters and employed the WARD technique to classify the data. After putting the cluster label into the original dataset, we can determine which cluster has a higher percentage of being highly rated and the key attributes within clusters.

Market Basket Analysis

Under this circumstance, we decided to further implement market basket analysis in the listing dataset. First, we further encoded the data into true and false value. Then, we performed a market basket analysis, setting a single antecedent variable, and 'highly_rated' as our target consequent. Our objective was to find attributes that are highly correlated with the high review rate by comparing support and lift.

Documentation clustering and Correlation Analysis

For text columns in the listing dataset, after dealing with certain text preprocessing steps, we decide to focus more on the column 'neighborhood_overview'. We utilized Word2Vec model to create word embeddings and defined a function to average the word vectors for each listing's overview to create a single embedding vector representing the text. After that, we applied the function to the column and implemented corresponding id at the same time. In this case, we decided to do KMean clusterings on embedding vectors and used Elbow Method to decide the optimal number of clusterings, which is 3, and fitted the algorithms (Appendix 5). Then, we merged the cluster labels into the original dataset. Based on the heatmap in the data preprocessing part (Appendix 3), we plan to observe the relationship between clustering and numerical attributes of number of reviews, price as well as review score rating. As a result, we performed a 3D scatter plot to visualize the correlation (Appendix 6).

Sentiment Analysis

When it comes to text columns in the review dataset, we initially employed SentiWordNet for sentiment analysis, which revealed some inconsistencies, such as positive comments receiving negative scores. Upon further investigation through t-SNE visualization, we noticed unclear boundaries between sentiment categories. WordCloud also shows that there are complimentary words in negative word clouds. This prompted a shift to the VADER model, which provided more distinct and reasonable sentiment categorization, evident in both histogram and word cloud analyses. However, t-SNE has similar ambiguous results which might be caused by the difficulty of representing sentiment in a binary classification. Additionally, we combined two dataframes based on listing_id and compared sentiment scores to user ratings, VADER demonstrated a stronger correlation and statistical significance. We also conducted a Linear Regression model to further confirm VADER's predictive power.

***Result and Insights*:**

TSNE:

As a result of TSNE, we can conclude that Cluster 1 in hierarchical clustering has a lower percentage in high rate, while Cluster 2 has a relatively high percentage. Therefore, we consider Cluster 1 as "low rated cluster" while Cluster 2 being "high rated cluster". Then, we dive deeper into seeing the relationship between clusters with 'super_host', and concluded that Cluster 2 tends to have a higher percentage of being as a super host.

In this case, we can provide suggestions for hosts that being a super host will help their listings to achieve better scores. Also, hosts in Cluster 1 may need to review and enhance their listings to improve guest experiences.

Market Basket Analysis:

By performing market basket analysis on listing dataset, the analysis highlights the critical importance of a host's status, with 'superhost_prefix_1'(host is super host) emerging as the most

influential attribute having the highest support and lift, which perfectly aligns with the result from TSNE. Additionally, “Room type” being Entire home/apt and the number of beds being 1 are also crucial, pointing towards a preference for certain accommodations that likely align with guests' comfort and space. The analysis goes further to reveal more specific attributes as the list goes on.

Under this circumstance, insights derived from this analysis offer an understanding in the rental market, suggesting that achieving high ratings is influenced by a combination of host quality, property characteristics, and the capacity to meet diverse guest needs. This investigation highlights the significance of host reliability, property type, and fit for visitor demands as crucial factors to earning top guest ratings. This could potentially benefit AirBnb on the host level as it gives a clear recommendation structure for listing.

Documentation clustering and Correlation Analysis

Based on the 3D visualization as well as the statistical results , it is noticeable that for price segmentation, cluster 1 has relatively high mean and median price suggesting these listings could be more premium. Cluster 2 has the lowest, indicating these could be budget listings. Talking about review score ratings, the mean review scores are consistent across all clusters, hovering around 4.72 to 4.75. This high score indicates that across various price points, guests are generally satisfied with their stays. Besides, Cluster 0 is considered as a high activity and moderate price listing and we consider this cluster might include terms related to accessibility, convenience, or popular tourist attractions.

The analysis suggests that word embeddings effectively capture descriptive features in listings, indicating that neighborhood overviews provided by host are reliable and enable guests to have clear expectations, meaning that hosts are accurately describing their listings, ensuring consistency between descriptions and guest experiences.

Sentiment Analysis

The analysis reveals that VADER significantly outperforms SentiWordNet, as it more accurately identifies nuances of positive and negative sentiments. VADER's ability to precisely identify negative words in word clouds, along with its strong correlation with user rating scores, indicates that it reflects more users' emotional tendencies. The robust predictive power of VADER demonstrated in linear regression analyses further validates its effectiveness.

Thus, accurate sentiment analysis provided by VADER can help customers better understand the evaluations of other users, enabling them to make more informed accommodation choices, rather than solely relying on user ratings. Using efficient sentiment analysis tools like VADER can reveal key emotions and trends within customer feedback, enabling hosts to identify and address potential issues, thereby improving guest satisfaction and return rates.

Challenges:

The original dataset is relatively messy and large, which took us the majority of times to do data cleaning and process it. Meanwhile, we lack time and knowledge in deep learning, which results in some difficulties in performing more accurate text analysis.

Future Work:

Firstly, we will deal with the data imbalance. It is also potential for our team to use PCA and TSNE results to perform supervised models to predict highly rated listings. Meanwhile, it’s likely that we perform neural networks to enhance performance and accuracy in text analysis. Also, since t-SNE faces challenges in distinguishing between different sentiment categories. We plan to explore the optimization of t-SNE parameters, such as adjusting the learning rate and perplexity, in hopes of achieving a more accurate distribution of sentiment categories.

Contributions

Name	Workload	Total percentage
Yingming Ma	Load and peek at the data, Check Null value, Determine the threshold for highly-rated listings, Categorical data manipulation, and Sentiment Analysis.	20%
Yile Xu	Check Null value for listing CSV, Creating numerical, categorical, and text columns into dataframe, Standardization, TSNE	20%
Yiwen Zhu	Dropped unknowns for text columns, Text processing, Vectorization, Documentation Clustering, and Check Null value for listing csv	20%
Shizuka Takahashi	Check Null value for listing csv, Heatmap, PCA variance, (performed clustering however was not ideal so is not included)	20%
Yanqi Su	Check Null value for listing csv, Check outliers for listing csv, Normalize the data, Pairplot, Market Basket Analysis	20%

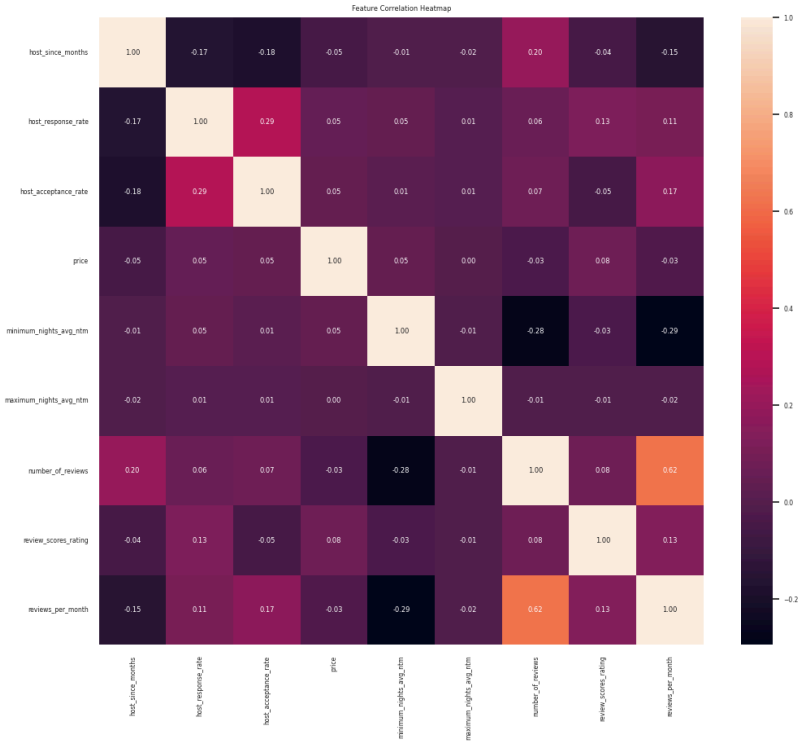
Appendix 1

#	Column	Non-Null Count	Dtype
0	id	4204 non-null	int64
1	listing_url	4204 non-null	object
2	scrape_id	4204 non-null	int64
3	last_scraped	4204 non-null	object
4	source	4204 non-null	object
5	name	4204 non-null	object
6	description	0 non-null	float64
7	neighborhood_overview	2742 non-null	object
8	picture_url	4204 non-null	object
9	host_id	4204 non-null	int64
10	host_url	4204 non-null	object
11	host_name	4204 non-null	object
12	host_since	4204 non-null	object
13	host_location	3342 non-null	object
14	host_about	2957 non-null	object
15	host_response_time	3656 non-null	object
16	host_response_rate	3656 non-null	object
17	host_acceptance_rate	3711 non-null	object
18	host_is_superhost	4168 non-null	object
19	host_thumbnail_url	4204 non-null	object
20	host_picture_url	4204 non-null	object
21	host_neighbourhood	4081 non-null	object
22	host_listings_count	4204 non-null	int64
23	host_total_listings_count	4204 non-null	int64
24	host_verifications	4204 non-null	object
25	host_has_profile_pic	4204 non-null	object
26	host_identity_verified	4204 non-null	object
27	neighbourhood	2742 non-null	object
28	neighbourhood_cleansed	4204 non-null	object
29	neighbourhood_group_cleansed	0 non-null	float64
30	latitude	4204 non-null	float64
31	longitude	4204 non-null	float64
32	property_type	4204 non-null	object
33	room_type	4204 non-null	object
34	accommodates	4204 non-null	int64
35	bathrooms	0 non-null	float64
36	bathrooms_text	4203 non-null	object
37	bedrooms	0 non-null	float64
38	beds	4150 non-null	float64
39	amenities	4204 non-null	object
40	price	3854 non-null	object
41	minimum_nights	4204 non-null	int64
42	maximum_nights	4204 non-null	int64
43	minimum_minimum_nights	4204 non-null	int64
44	maximum_minimum_nights	4204 non-null	int64
45	minimum_maximum_nights	4204 non-null	int64
46	maximum_maximum_nights	4204 non-null	int64
47	minimum_nights_avg_ntm	4204 non-null	float64
48	maximum_nights_avg_ntm	4204 non-null	float64
49	calendar_updated	0 non-null	float64
50	has_availability	3854 non-null	object
51	availability_30	4204 non-null	int64
52	availability_60	4204 non-null	int64
53	availability_90	4204 non-null	int64
54	availability_365	4204 non-null	int64
55	calendar_last_scraped	4204 non-null	object
56	number_of_reviews	4204 non-null	int64
57	number_of_reviews_ltm	4204 non-null	int64
58	number_of_reviews_l30d	4204 non-null	int64
59	first_review	3086 non-null	object
60	last_review	3086 non-null	object
61	review_scores_rating	3089 non-null	float64
62	review_scores_accuracy	3088 non-null	float64
63	review_scores_cleanliness	3089 non-null	float64
64	review_scores_checkin	3087 non-null	float64
65	review_scores_communication	3089 non-null	float64
66	review_scores_location	3087 non-null	float64
67	review_scores_value	3087 non-null	float64
68	license	2695 non-null	object
69	instant_bookable	4204 non-null	object
70	calculated_host_listings_count	4204 non-null	int64
71	calculated_host_listings_count_entire_homes	4204 non-null	int64
72	calculated_host_listings_count_private_rooms	4204 non-null	int64
73	calculated_host_listings_count_shared_rooms	4204 non-null	int64
74	reviews_per_month	3086 non-null	float64

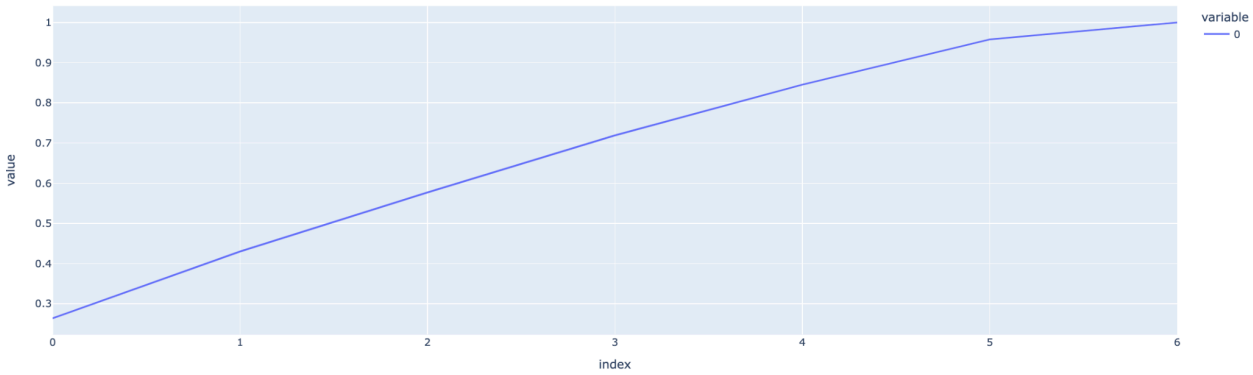
Appendix 2

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 182482 entries, 0 to 182481
Data columns (total 6 columns):
 #   Column          Non-Null Count   Dtype 
---  ------          --------------   ----- 
 0   listing_id      182482 non-null  int64 
 1   id              182482 non-null  int64 
 2   date            182482 non-null  object
 3   reviewer_id     182482 non-null  int64 
 4   reviewer_name   182481 non-null  object
 5   comments        182430 non-null  object
dtypes: int64(3), object(3)
memory usage: 8.4+ MB
```

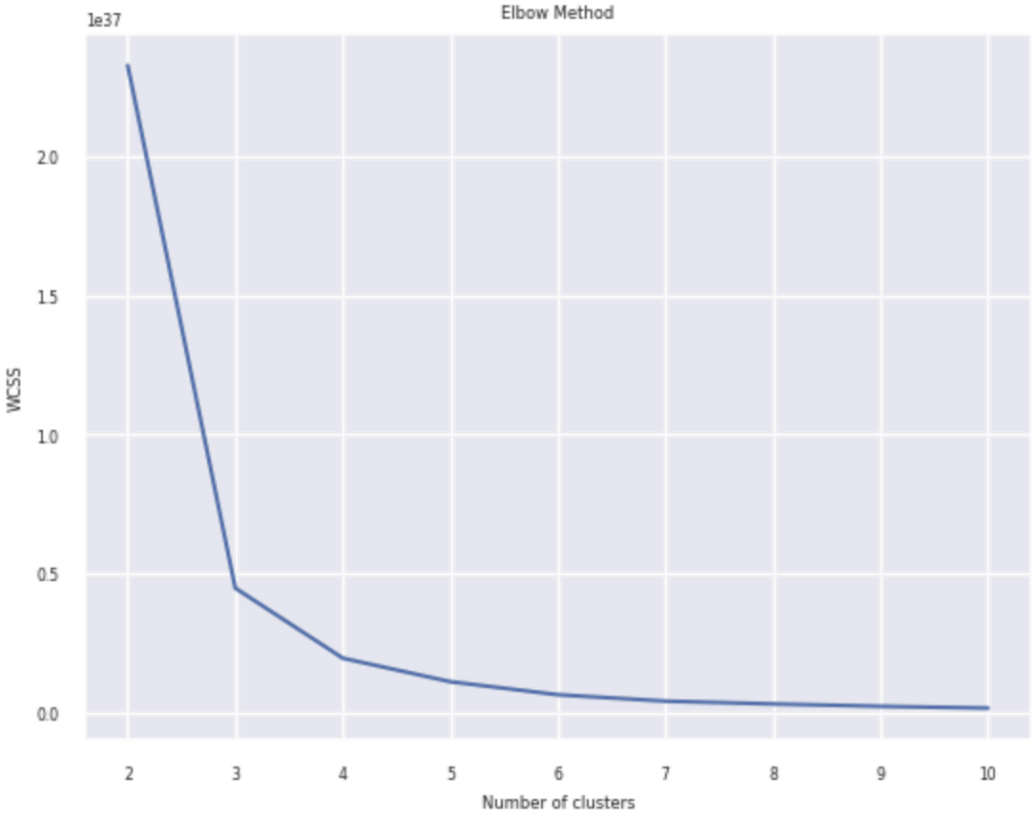
Appendix 3



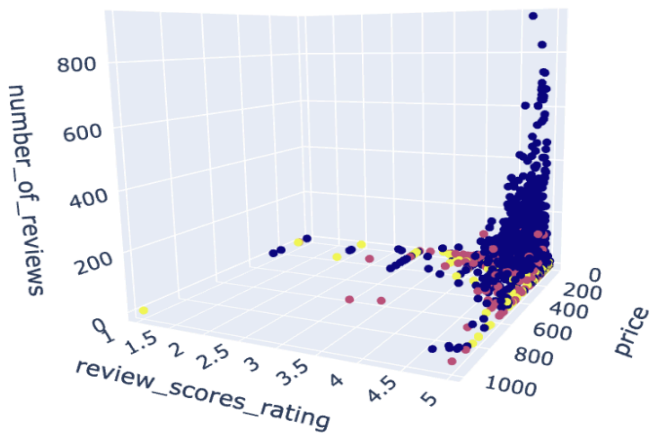
Appendix 4



Appendix 5



Appendix 6


number_of_reviews
800
600
400
200
0
review_scores_rating
0
1.5
2
2.5
3
3.5
4
4.5
5
price
0
200
400
600
800
1000