SENTIMENT ANALYSIS OF AIRBNB REVIEWS

TEAM : SENTIMENT SYNTHESIZERS

Avantika Gargya, Basava Satish Velagapudi, Krisha Gandhi, Kenneth Fung, Sai Bharadwaj

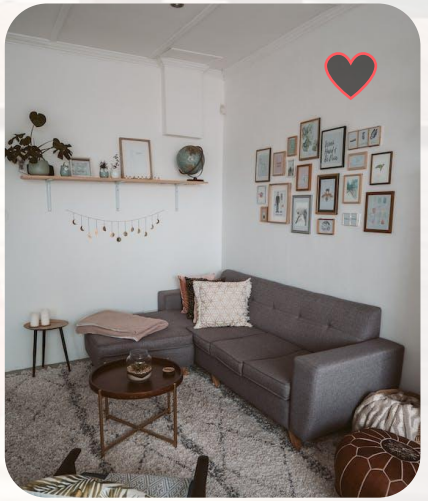
Colab: https://colab.research.google.com/drive/13aAb9QW8bqrzDqp8i5yhze-P2frbOMkh?authuser=4#scrollTo=efzQ7KLOQfv4

Problem Statement

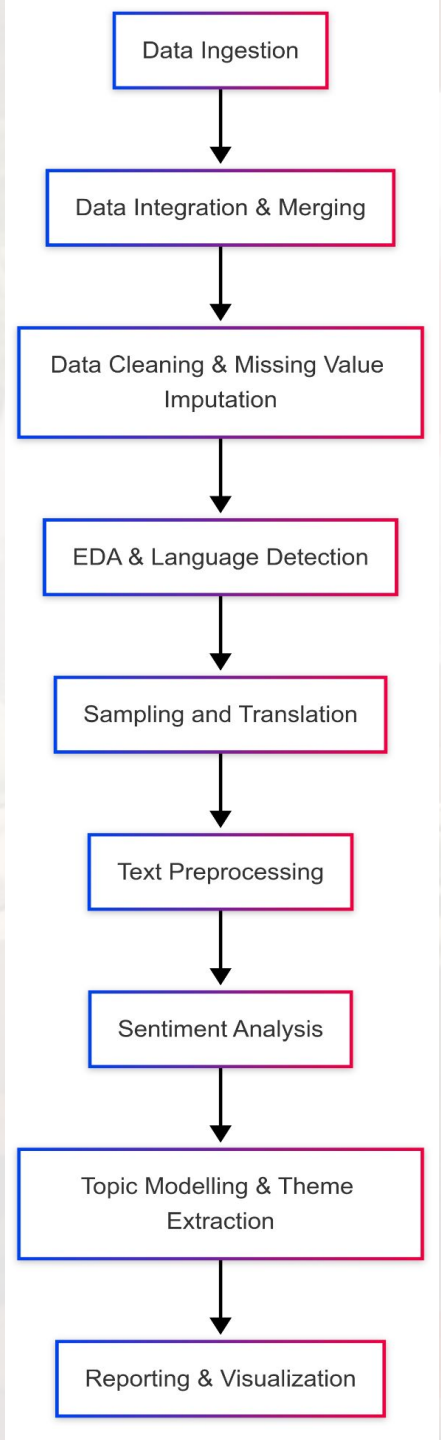
Advanced NLP-Driven Insights from Airbnb Reviews for Neighborhood and Listing Centric Recommendations

Goals

- Help Hosts understand the **positive and negative aspects of the property, fix issues** and **boost their bookings**
- Help Customers **quickly assess listings** & make an i**nformed booking**
- Help Airbnb **identify trends, highlight popular areas** and **optimize pricing** and **market strategy**



Workflow


Data Ingestion
Data Integration & Merging
Data Cleaning & Missing Value Imputation
EDA & Language Detection
Sampling and Translation
Text Preprocessing
Sentiment Analysis
Topic Modelling & Theme Extraction
Reporting & Visualization

Dataset Overview

Source

We got access to the Airbnb listings dataset from their Inside Initiative - https://insideairbnb.com/get-the-data/

Scope

We used New York City's data as of January 3, 2025- focusing on Manhattan and Brooklyn neighbourhoods.

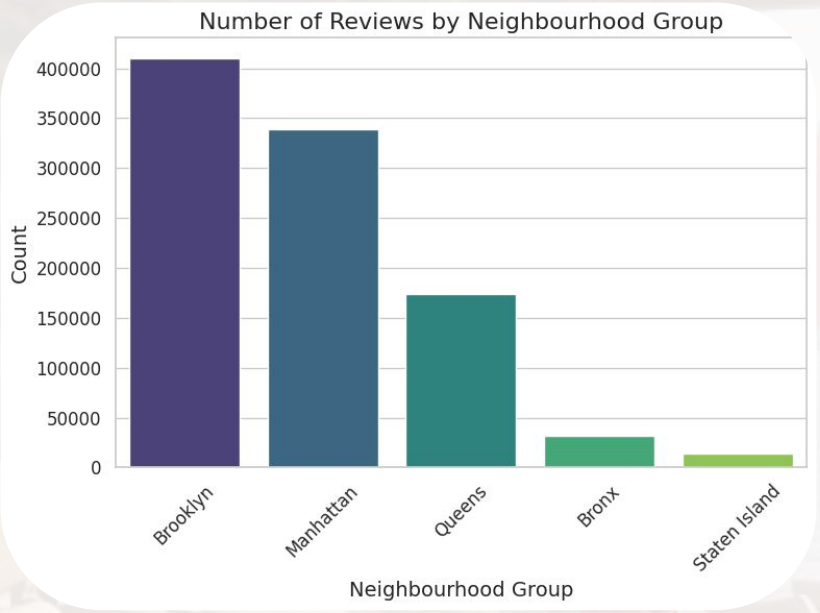
Listings

There are about **37.7K** listings across 5 boroughs with information across 75 columns with information such as listing info (description, photos, price), host info, property type, neighbourhood info, availability

Reviews

We have over **969K** reviews across all the listings. It contains the listing id, reviewer info (name, id, date of posting)in addition to the review itself.

Additionally, we can access information on the **neighbourhoods**, as well as the availability **calendar**.



Pre-processing

1

Merged Dataset

Merged reviews data with listings data on listing id; added property type, price and neighbourhood and removed reviewer details.

2

Missing Values

- Reviews - 235 missing - dropped.
- Prices - 19K missing, imputed by median by neighbourhood

3

Neighbourhood Division

To generate more granular insights, we split NY city into 5 boroughs - Manhattan, Brooklyn, Staten Island, The Bronx and Queens.

4

Sampling

Focus neighbourhoods
: Manhattan - 340K
: Brooklyn - 410K
Since the data was still sizeable leading to slow analysis speed
we resorted to a sampling approach.

Pre-processing

Language Detection

- We used the **langdetect library** to identify languages in Airbnb reviews, finding over **30 languages,** with English, Spanish & French being the most common.
- We focused on **Brooklyn** and **Manhattan**, which had the highest number of listings & reviews.

Language Distribution & Sampling

- We **filtered reviews** to include only languages with over **5,000** reviews, refining our dataset to English, Spanish, and French.
- A **subsample** (**~51k reviews**)
  - High-review listings(>100 reviews)
  - Sampling an equal number of reviews per listing
  - **46** reviews per listing-Brooklyn
  - **65** reviews per listing-Manhattan

Language Translation

- We translated non-English reviews (**French** and **Spanish**) into English using the **Google Translator API**
- This **ensured consistency** in sentiment analysis while leveraging automation

Text Pre-processing for Sentiment Analysis

1.

Setup & Initialization

- ***Lemmatization:*** WordNetLemmatizer
- ***Dictionary Validation:*** US, UK, Canadian, Australian, New Zealand, and South African
- ***Stopword Management:*** Retain essential stopwords like "not," "no," and "against" to preserve sentiment meaning

2.

Text Cleaning & Normalization

- ***Convert text to lowercase***
- ***Tokenization:*** Split text into individual words
- ***Stopword Removal:*** Remove irrelevant words
- ***Apostrophe Normalization:*** Standardize apostrophes (e.g., "dog”s" → "dog’s")
- ***Punctuation Removal:*** Retain only letters, digits, spaces & apostrophes

3.

Token Filtering & Lemmatization

- ***Validate words:*** Retain only those recognized in at least one of the English dictionaries
- ***Apply Lemmatization:*** Convert words to their base forms, reducing dimensionality

4.

Reconstructing the Processed Text

- ***Rejoin tokens:*** After filtering & lemmatization, tokens are rejoined into structured text
- ***Implementation on the dataframes:*** We now apply preprocessing separately to the Brooklyn & Manhattan datasets

Sentiment Analysis

Library Used: TextBlob

Polarity Score: a float value that is in the range of [-1,1] where 1 means a positive statement and -1 means a negative statement

Subjectivity Score: a float value that is in the range of [0,1]. 0 means a personal opinion and 1 means factual information

Score Labelling: <0, tagged it as a negative review; if >=0, it was labelled as positive
<0.3, labelled as objective, if <=0.7, mixed, else subjective.

Analysis Approaches:
Sentiment analysis across all Manhattan & Brooklyn listing reviews

Sentiment scores averaged by listing_id to help hosts gauge guest feedback

cleaned_reviews	polarity	subjectivity	sentiment_label	subjectivity_label
neighborhood quiet people peaceful amenity eas...	0.208333	0.402778	Positive	Mixed
great communicating home great stay highly rec...	0.586667	0.680000	Positive	Mixed
stayed again enjoyed again	0.500000	0.700000	Positive	Mixed
check in process absolutely easy responsive qu...	0.266667	0.622222	Positive	Mixed
not sure cleaning fee charged cleaner do visib...	-0.263333	0.466111	Negative	Mixed

Theme Extraction

Idea:

- **Go beyond sentiment analysis to extract key themes and features from reviews.**
- **Provide actionable insights for both hosts and guests.**

Key Takeaways

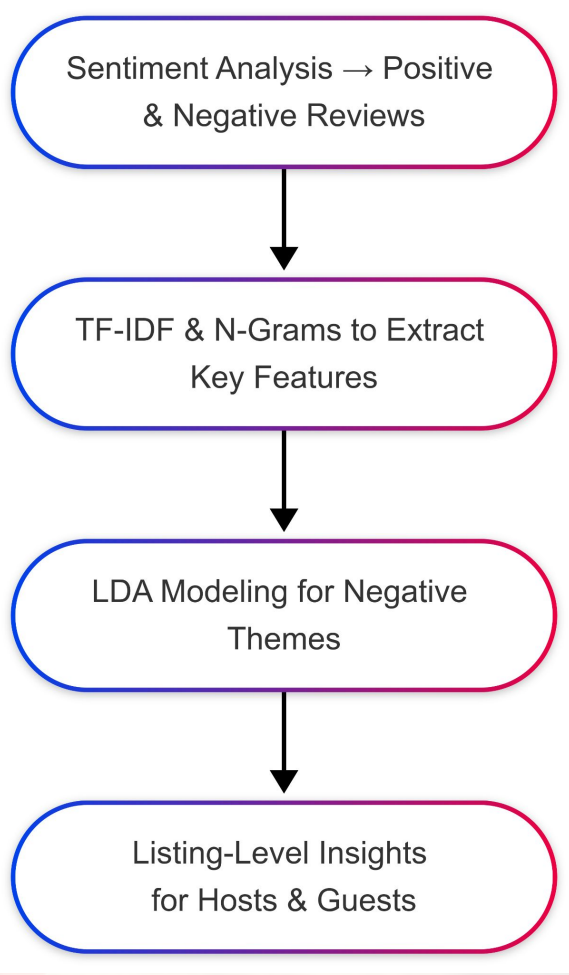
For Hosts 1

- Improve listing descriptions
- Adjust pricing for competitiveness
- Address recurring issues for better ratings

For Guests 2

- Summarized review insights for quick decisions
- Avoid common pitfalls (e.g., noisy streets)
- Find stays that match their needs

Data-driven insights → Higher ratings, better reviews, increased bookings!



Results - Neighborhood

	Brooklyn	Manhattan
Sentiment_Score	0.4141	0.4086

Sentiment Label	Count in Brooklyn	Count in Manhattan
Positive	50548	50502
Negative	558	848

Word Cloud for Positive Reviews - Brooklyn

Word Cloud for Positive Reviews - Manhattan

Results - Neighborhood

Negative Reviews for Brooklyn listings		Negative Reviews for Manhattan listings	
Feature	**Mean TF-IDF Score**	**Feature**	**Mean TF-IDF Score**
within walking distance	0.009175	worst experience ever	0.010547
keep in mind	0.005817	not worth price	0.006691
in sunset park	0.005376	within walking distance	0.005896
bathroom not clean	0.005376	few block away	0.004717
not clean enough	0.005376	front desk staff	0.004427

Results - LDA

1

Comfort

Complaints about **broken furnitures** (e.g., "not work") and comfort in certain areas (e.g., "ceiling fan", "air conditioning").

2

Cleanliness

Dissatisfaction with **tidiness** and **hygiene** of the listing, especially when shared with others (“flat mate”, “common area”).

3

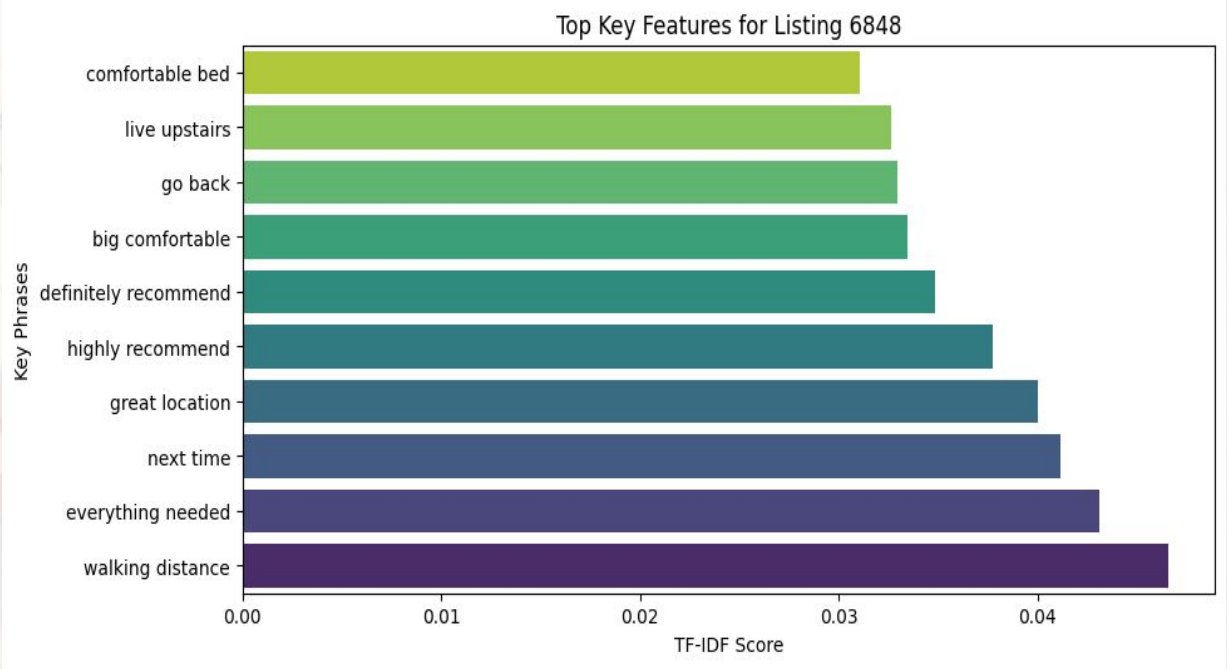
Expectations vs Reality

Concerns with **unmet expectations** ("in city", "in photo"), **noise** ("lot noise"), and comfort (e.g., "not comfortable", "air conditioning").

Results – Listing

Listing ID	Average Sentiment Score	Sentiment Label
6848	0.3623	Positive

Feature	Relevance	Feature	Relevance
in great location	0.013591	within walking distance	0.009175
within walking distance	0.010521	keep in mind	0.005817
home away home	0.007245	in sunset park	0.005376



Limitations

1. Use entire dataset
2. Dissect needs and reviews by cultural backgrounds
3. Identify patterns over the year
4. Filter out fake reviews

Conclusion

Reveal Customer Feelings

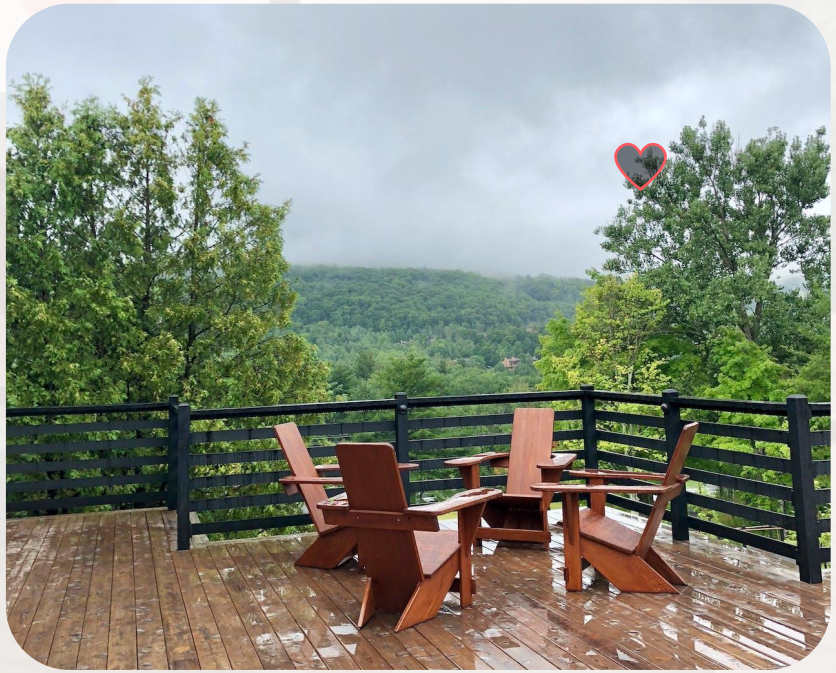
How do Airbnb guests feel about their stays?

Drivers for Sentiment

What factors drive good or bad experiences?

Benefits

Improve listing, pricing, host services, listing-search



Thank You

Questions?