Detecting Fraud In Hawaii Airbnb Listings

Ber Bakermans, Sophia Jaskoski, Marjan Abedini, Tharuni Tekula

Problem & Stakeholder

Airbnb is facing the challenge of fraudulent listings that manage to slip through their security checks. These fake listings can deceive users into booking places that don't exist or that are not accurately described, leading to issues like financial loss and inconvenience for travelers.

Stakeholders:

- **Airbnb Guests:** feel safe with booking genuine properties
- **Airbnb Managers:** reliable tools to find fake listings to keep the platform trustworthy
- **Airbnb Hosts:** protections for their listings to have a fair marketplace

Envisioned Solution

Harness anomaly detection and clustering methods to flag listings as fraudulent

Outliers/anomalies are flagged for further investigation, allowing Airbnb to take a closer look and verify whether the listing is legitimate.

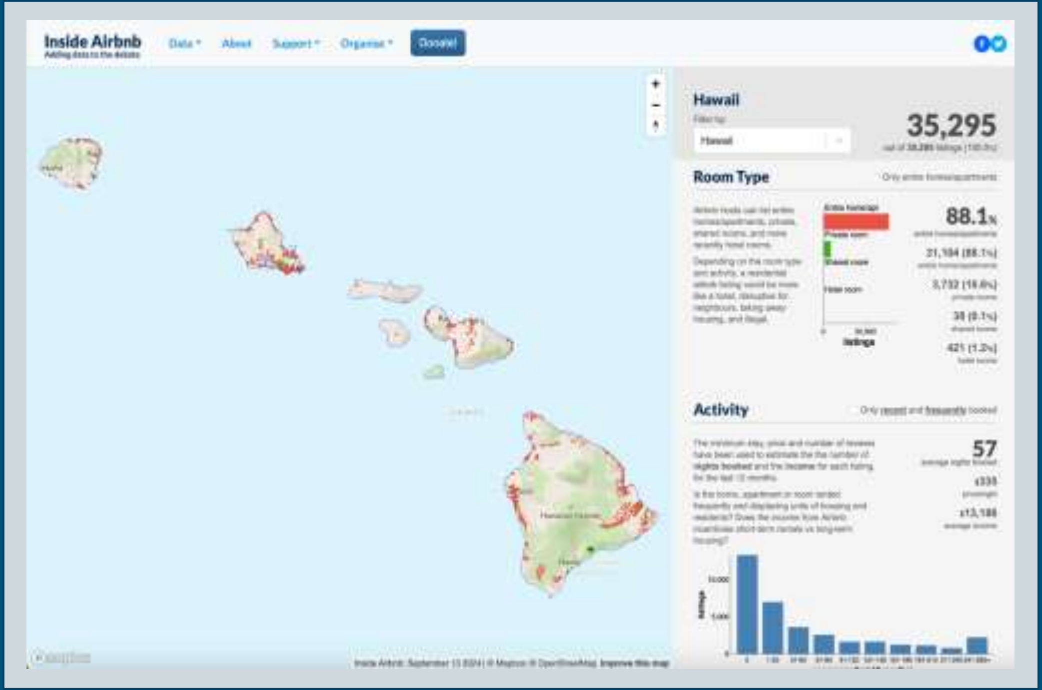
Additionally, listing descriptions are subject to frequency and similarity analysis to detect duplicate listings and/or spam

Listings which are flagged by all models will be manually inspected for fraud or duplication

Data

Our data comes from *Inside Airbnb*, a public web scraping platform which provides comprehensive information on Airbnb listings in major cities throughout the world. We elected to use the listings for Hawaii, which is an ideal vacation destination for many travelers.

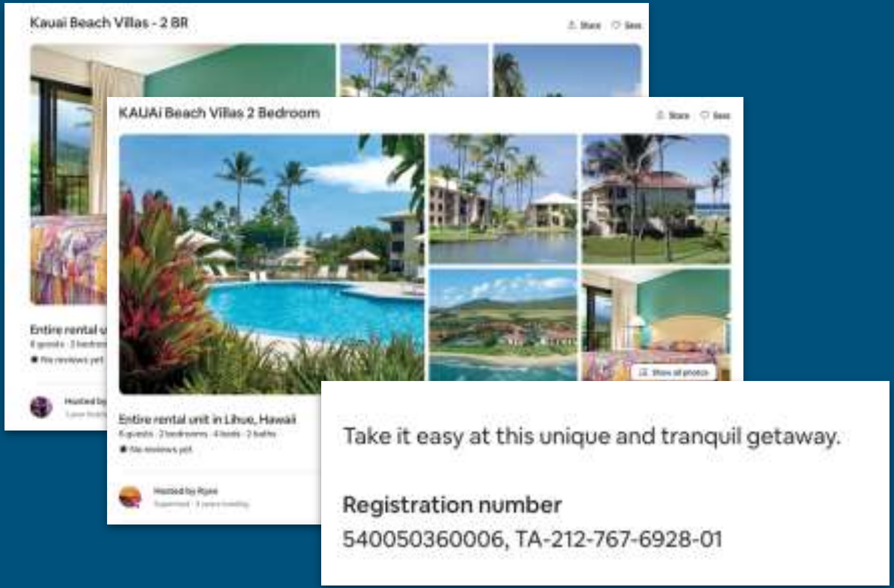
This dataset includes **35,295 rows** and **75 columns**, capturing a wide range of features that describe the properties and their characteristics. The data includes details such as listing prices, host information, location data, reviews, and other metrics that can help us identify patterns and detect anomalies.

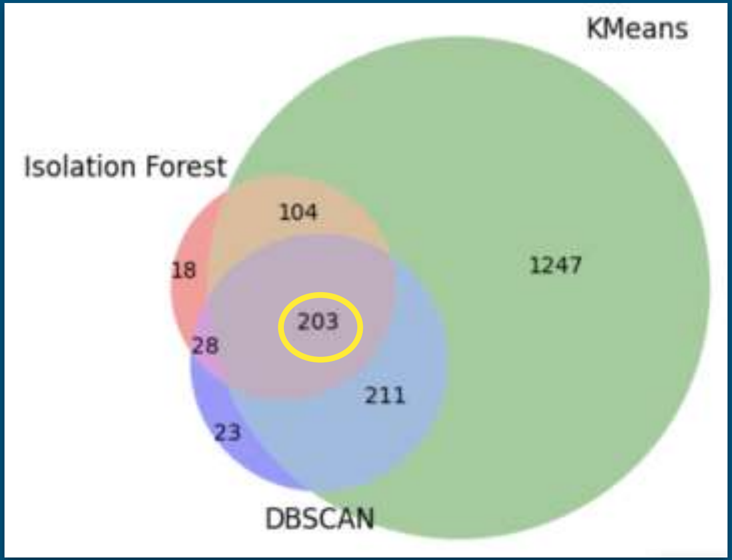


Initial Results

Cluster Analysis

Text Mining Analysis





Challenges

Skewed Variables	Large Feature Space	Lack of Ground Truth
Many of our variables including price, the maximum and minimum number of nights, and the number of listings per host were significantly right skewed, which affected how PCA collapsed our data into principal components.	While we started with 75 columns, the encoding of all of our categorical columns created too large of a feature space for us to analyze and crashed our code space. Consequently, we had to consciously drop and feature engineer certain columns.	We do not know which, if any, of the listings are fraudulent. We defined a fraudulent property as one which was detected by all three of our models in our clustering analysis and was noted as having a high similarity score in our text mining analysis.

Plan and Goals

Dec 5

Improve Clustering Visualizations and Select Best Model Combo

Apply TSNE or UMAP over PCA transformed data to create better visualization of distinct fraudulent/non-fraudulent clusters

Dec 10

Harmonize Text Mining and Clustering Analyses

Identify listings which flagged as outliers and anomalies by best models for both text mining and clustering to reduce number of flagged listings

Dec 14

Validate Fraudulent Listings

Manually inspect listings which were identified in both analyses for duplication, false advertising, or suspicious hosts

Dec 17

Submit Final Report