Information Visualization Plan

Team Members

Zaharia Stefan-Tudor - 511

Introduction

The goal of this project is to find out what are the most important factors that affect the price of an Airbnb listing. By doing so, Airbnb hosts can have a better understanding of how to price their listings and how to stand out from the competition. The neighborhood, the amenities, the user reviews, the listing description are just a few important factors that affect the price of an Airbnb listing, but there are many more.

Better understanding of the data will help Airbnb hosts to stay competitive and maximize their profits, providing a better experience for the guests and offering a better value for money, which will ultimately lead to a better experience for everyone.

The main metrics that will be used to measure the success of an Airbnb listing are the popularity and the price.

Dataset Features

Using the New Orleans Airbnb Listings dataset (https://www.kaggle.com/datasets/ruthgn/new-orleans-airbnb-listings-and-reviews?select=new_orleans_airbnb_listings.csv), the goal is to investigate what are the most important factors that makes a listing popular and how to price it accordingly. The dataset contains 6028 listings and 49 columns. The features included in the dataset are the following:

- `id`: Airbnb's unique identifier for the listing
- `name`: Name of the listing
- `description`: Detailed description of the listing
- `neighborhood_overview`: Description of the neighborhood
- `host_id`: Airbnb's unique identifier for the host/user
- `host_since`: The date the host/user was created. For hosts that are Airbnb guests this could be the date they registered as a guest.
- `host_location`: The host's self reported location
- `host_response_time`: That average time it takes a host to respond to a message on the Airbnb platform.
- `host_response_rate`: That rate at which a host responds to a message on the Airbnb platform.
- `host_acceptance_rate`: That rate at which a host accepts booking requests.
- `host_is_superhost`: Whether host has an Airbnb 'Superhost' badge on the platform
- `host_listings_count`: The number of listings the host has (per Airbnb calculations)
- `host_verifications`: Types of verification the host has passed.
- `host_has_profile_pic`: Whether host included a profile picture on their profile on the Airbnb platform
- `host_neighbourhood`: The host's self reported neighborhood
- `host_identity_verified`: Whether the host's identity has been verified

- `neighbourhood_cleansed`: The neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.
- `latitude`: Uses the World Geodetic System (WGS84) projection for latitude and longitude.
- `longitude`: Uses the World Geodetic System (WGS84) projection for latitude and longitude.
- `property_type`: Self selected property type. Hotels and Bed and Breakfasts are described as such by their hosts in this field
- `room_type`: [Entire home/apt|Private room|Shared room|Hotel]
- `accommodates`: The maximum capacity of the listing
- `bathtrooms_text`: "The number of bathrooms in the listing. On the Airbnb web-site, the bathrooms field has evolved from a number to a textual description. For older scrapes, bathrooms is used."
- `bedrooms`: The number of bedrooms
- `beds`: The number of bed(s)
- `amenities`: List of amenities included with the rental
- `price`: Daily price in local currency
- `minimum_nights`: Minimum number of night stay for the listing
- `maximum_nights`: Maximum number of night stay for the listing
- `has_availability`: [t=true; f=false]
- `availability_30`: Availability over next 30 days. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.
- `availability_60`: Availability over next 60 days. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.
- `availability_90`: Availability over next 90 days. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.
- `availability_365`: Availability over next 365 days. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.
- `number_of_reviews`: The number of reviews the listing has
- `number_of_reviews_ltm`: The number of reviews the listing has in the last 12 months
- `number_of_reviews_l30d`: The number of reviews the listing has in the last 30 days
- `first_review`: The date of the first/oldest review
- `last_review`: The date of the last/newest review
- `review_scores_rating`: Overall ratings score
- `review_scores_accuracy`: Ratings score for description accuracy
- `review_scores_cleanliness`: Ratings score for rental property cleanliness
- `review_scores_checkin`: Ratings score for check-in convenience
- `review_scores_communication`: Ratings score for ease of communication with host
- `review_scores_location`: Ratings score for rental location
- `review_scores_value`: Ratings score for rental property value
- `license`: The licence/permit/registration number
- `instant_bookable`: [t=true; f=false]. Whether the guest can automatically book the listing without the host requiring to accept their booking request. An indicator of a commercial listing.
- `reviews_per_month`: The number of reviews the listing has over the lifetime of the listing

Data Analysis

The analysis will be focused on two main directions:

- Investigating the distribution of each feature individually, summarizing the main characteristics of the data and identifying any outliers. For this task, we can use visualizations such as histograms and boxplots to analyze the size and the shape of the distribution.
- Investigating the relationship between the features and the target variable (price/popularity). For this task, we can use visualizations such as scatter plots and heatmaps to analyze the correlation between the features and the target variable. The popularity metrics will be calculated based on the number of reviews in a given period of time and the availability of the listing.

Questions to Answer

These questions are a few examples of what can be answered using the data:

- How is the popularity affected by the reviews
- Are the reviews for communication correlated with the host response time, host response rate and host acceptance rate?
- How is the popularity affected by the location?
- Is the price correlated with the room type?
- What amenities drive the price up?
- What neighbourhoods have the best reviews for location?
- Is the room type and the property type correlated with the price?
- Will a profile picture, verified identity and the status of superhost increase the popularity of the listing?
- Will the number of bedrooms/bathrooms/beds increase the price?
- What description/name length will increase the popularity of the listing?
- What words are associated with the most popular listings?

Technologies

Python + PowerBI