Technical Report
Preprocessing and Exploratory Analysis of Airbnb NY Listings

This report documents the preprocessing pipeline and initial exploratory analysis conducted on the Airbnb New York City listings dataset. The dataset includes host attributes, property characteristics, booking details, pricing, and review-related information. The preprocessing aimed to clean, transform, and engineer features suitable for later pattern discovery and divergence analysis.

Data Preprocessing

We use the **Airbnb NYC listings dataset** (listings.csv.gz), which includes host, property, location, and review-related features. We select a subset of relevant columns covering host information, property attributes, reviews, and pricing.

Column Selection. We select subset of relevant features was extracted, covering:

- **Host attributes**: hosting start date, response/acceptance rates, superhost status, verified identity.
- **Property details**: property type, room type, number of bedrooms, beds, and bathrooms.
- **Listing characteristics**: amenities, accommodates, price, booking policy.
- **Review-related variables**: review counts, dates of first/last reviews, review scores.

The **review_scores_rating** is the target outcome of interest. We will investigate changes in the review scores.

Handling Rare Categories and Missing Values.

- **host_neighbourhood and property_type**: To avoid sparsity from categorical variables with many unique values, we use function aggregate_not_frequent to replace infrequent values with frequency <1% with "other". We applied it to the columns host_neighbourhood and property_type. We also mapped missing values in these columns to "other".
- **host_response_time**: We set missing entries of the field host_response_time to "NA" (string category).

Feature Engineering.

- **Amenities**: We transformed the list of amenities into a new feature #amenities = number of amenities.
- **Review dates (first_review, last_review) and host start date (host_since)**: We converted the fields first_review and last_review into numeric features measuring the years since the event respect to the current time Missing dates were left as NaN after transformation.

- **bathrooms**: We used the field bathrooms_text to derive a binary variable shared_bathroom (1 if “shared” mentioned, else 0). We dropped bathrooms_text afterward since the field bathrooms already contained numeric counts.
- **Price**: We converted from string to numeric (removed $ and cast to float).
- **Host rates**: We converted host_response_rate and host_acceptance_rate transformed from string percentages ("85%") into floats (85.0).

Discretization. To prepare the dataset for subgroup discovery, we discretized the data as follows.

- **General discretization**: We binned the continuous variables (e.g., price, #amenities, review counts) into 3 categories by frequency.
- **Discretization for number_of_reviews_l30d**: The automatic discretization yielded only one bin, so it was manually grouped into: "0" (no reviews), "1" (exactly one review), ">1" (more than one review).

Data Filtering

- We removed the listings with missing review_scores_rating (the target).
- We removed features strongly correlated with the target were excluded to prevent leakage, such as other fields related to review scores ('review_scores_accuracy','review_scores_cleanliness', 'review_scores_checkin','review_scores_communication', 'review_scores_location','review_scores_value') and calculated fields ('calculated_host_listings_count', 'calculated_host_listings_count_entire_homes', 'calculated_host_listings_count_private_rooms', 'calculated_host_listings_count_shared_rooms').

- **host_response_time**: The time it takes for the host to respond to a booking request, with missing values replaced by "NA".
- **host_response_rate:** Percentage of booking requests the host responds to/
- **host_acceptance_rate**: Percentage of booking requests the host accepts, converted to a float.
- **host_is_superhost**: Indicates if the host is a Superhost or not.
- **host_neighbourhood**: Neighborhood of the host; low-frequency neighborhoods are replaced with "other".
- **host_identity_verified**: Whether the host’s identity has been verified.
- **neighbourhood_group_cleansed:** District of the listing (e.g., Manhattan, Brooklyn).
- **property_type**: type of property (e.g., Private room in rental unit, Entire rental unit); low-frequency types are grouped as "other".
- **room_type**: Type of room offered (e.g., Entire home/apt, Private room,).
- **accommodates**: Number of guests the listing can accommodate.
- **bathrooms**: Number of bathrooms.
- **bedrooms**: Number of bedrooms.
- **beds**: Number of beds.

- **price**: Listing price in USD (currency symbols removed).
- **number_of_reviews**: Total number of reviews the listing has received.
- **number_of_reviews_ltm**: Number of reviews in the last 12 months.
- **number_of_reviews_l30d**: Number of reviews in the last 30 days, discretized into '0', '1', or '>1'.
- **instant_bookable**: Indicates if the listing can be booked instantly without host approval.
- **reviews_per_month**: Average number of reviews per month for the listing.
- **#amenities**: Count of amenities available for the listing.
- **first_review_year**: Years since the first review.
- **last_review_year**: Years since the most recent review.
- **host_since_year** :Years since the host joined Airbnb.
- **shared_bathroom**: 1 if the bathroom is shared, 0 otherwise.
- **review_scores_rating** – Overall review score (target variable), typically from 1 to 100.