# Project Report

Trinath Sai Subhash Reddy Pittala, Uma Maheswara R Meleti, Hemanth Vasireddy

2023-04-04

## Introduction

### Airbnb Price Determinants in Europe

We want to work on Airbnb's dataset from kaggle.com. It provides information about hotel rooms in Europe.

Each major city has its dataset for weekends and weekdays Variables included in the dataset: Host ID (Id) The total price of listing (realSum) Room type: private, shared, entire home, apt (room_type) Whether or not a room is shared (room_shared) Max number of people allowed in property (person_capacity) Whether or not the host is superhost (host_is_superhost) Whether or not it is multiple rooms (multi) Whether for business or family use (biz) Distance from the city center (dist) Distance from nearest metro (metro_dist) Latitude and longitude (lat long) Guest satisfaction (guest_satisfaction_overall) Cleanliness (cleanliness_rating) The total quantity of bedrooms available among all properties for a single host (bedrooms)

Questions we can answer with the dataset: Price Forecasting: use pricing, room type, and amenities to predict potential rental prices in the future. Hotspots: use listing location in relation to business and tourism centers and correlate this with pricing to determine where Airbnb rentals would be most profitable Customer Sentiment Analysis: analyze customer comments and satisfaction ratings to evaluate listing on overall customer experience and use it to optimize hosts' services to improve user satisfaction ratings.

How can this information be used: Data can help travelers find accommodation that meets their needs without exceeding budget. Can help hosts set competitive pricing and optimize listings to get more bookings. Help investors evaluate the value of investing in real estate in different European cities based on pricing trends.

### Variable Description

realSum: The total price of the Airbnb listing

roomtype: The type of room offered (e.g. private room, shared room, entire home/apt).

room_shared: Whether the room is shared or not.

room_private: Whether the room is private or not.

person_capacity: The maximum number of people that can be accommodated in a single listing.

host_is_superhost: Whether or not a particular host is identified as a superhost on Airbnb.

multi: Whether multiple rooms are provided in one individual listing or not.

biz: Whether a particular listing offers business facilities like meeting area/conference rooms in addition to

cleanless_rating: The rating associated with how clean an individual property was after guests stayed at it.

guest_satisfaction_overall: The rating associated with how clean an individual property was after guests stayed at it.

dist: The total quantity of bedrooms available among all properties against a single hosting id.

metro_dist: Distance from metro station associated with every rental property.

attr_index: attraction index

attr_index_norm: attraction index, normalized

rest_index: restaurant index

rest_index_norm: restaurant index, normalized

lng Longitude measurement corresponding to each rental unit.

lat: Latitude measurement corresponding to each rental unit

# Pre Processing and Cleaning the Data

Before we could begin the analysis, we preprocessed the data to ensure that it was consistent. This involved combining the 20 files to a single table with a additional column of city_day.

We also removed certain redundant columns such as room_shared and room_private whose information is present in room_type completely ans exhaustively.

Next we have separated the outliers using IQR ranges.

We have also Dropped attr_index and rest_index because they were already normalized and given as seperate attributes.

We have split the training and testing data here itself with 7:3 split on constant seed.

# Exploratory Data Analysis

We performed several analyses to identify the factors that affect Airbnb pricing. Firstly, we used descriptive statistics to analyze the distribution of the variables and examine any patterns or trends.

## Outlier Analysis

### Metro Dist vs Real Sum

We started the EDA with Outlier analysis, We have planned to analyse the filtered data along with outlier data. Here outlier data represents the hotel rooms with high prices.

In general the rooms that are closer to metro have comparatively higher prices. But, in Rome city the distance to metro is almost same for both categories of price.

### Real Sum vs Distance

In general the pricey rooms are near to the centre of the city according to the Scatterplot

### Real Sum vs Attraction Index Normal

The range of values falling b/w outliers and normal data is almost same . So there isn't a relationship b/w attr_index and realSum.

### Real Sum vs Restaurant Index Normal

There is no relationship between outliers and rest_index

### Room Type Vs Real Sum

The price of entire home/apt tend to be higher compared to other two categories. And the count of entire home /apt is also more.

### Room Type Vs Person Capacity

The overall price is distributed similarly across the spectrum irrespective of person_capacity. But for some cities like london, london_weekdays, lisbon the price is higher with person capacity. So, person capacity along with city will be an important variable for determining price.

### Real Sum Vs host_is_superhost

The prices are spread across all the spectrum irrespective of super_host or not.

### Real Sum Vs multi

The prices are similar irrespective of multi or not.

### Real Sum vs biz

The prices are similar irrespective of biz or not.

### Real Sum vs Cleanliness

The cleanliness rating doesn't really have an effect on price

### Scatterplot of Price vs Guest Satisfaction filtered by city

The plot depicts that there is no correlation of price with guest satisfaction, good satisfaction rate is found across all the prices. In some cities like lonon, we can see a group of reviews with low guest satisfaction.

### Real Sum Vs Bedroom Count

No observable relation.

## Non - Outlier Analysis

### Boxplot of Price Vs City

The highest prices in Europe are found in Amsterdam.

**Density plot of Price vs Room type**

The prices of entire home are high comparatively

**Scatterplot of Prices in Rome w.r.t Latitude and Longitude during weekdays**

This plot is within expectations of general trends, which suggests similar types of establishments (price and hospitality) tend be in clusters.

## Different Model Selection and Training

We conducted a regression analysis to determine the variables that had the most significant impact on Airbnb pricing.

Linear regression has given a R^2 value of 0.21 which is really low.

The situation is same with Lasso step regression.

## Conclusion at this Point in Time

Even though EDA has given us good insights in price determinants, both Linear and Lasso Step Regression are not good for this case which is to be expected since all common data tends to be generally skewed Normal or Guassian(to be tested).

Further modelling is required and will be conducted which includes trying of different linear techniques and also models from different family.