REVENUE ESTIMATION FOR PROPERTIES

# Airbnb

Group members: Piyush Kakde, Karan Karnik, Monika Madugula, Timothy Samuel

Get started

# Executive Summary

In our presentation 'Revenue Estimation for Properties,' we introduce a predictive model designed to estimate revenue for Airbnb properties in Dallas. This model leverages historical data by conducting exploratory data analysis (EDA), building the model, and identifying key features. Our comprehensive analysis focuses on identifying factors that significantly influence revenue, such as property characteristics, socio-economic elements, and seasonal trends. By applying machine learning techniques, we aim to empower Airbnb hosts with data-driven insights for optimized revenue decisions in the dynamic short-term rental market.

# Introduction and Problem Statement

01

Airbnb

# Company Overview

Founded in 2007 by Joe Gebbia and Brian Chesky, Airbnb is a digital platform that facilitates the rental of private homes, rooms, and unique accommodations, connecting hosts with guests. It offers a diverse range of lodging options, including apartments, beach houses, cabins, and more

**>7M** active listings worldwide as of June 30, 2023

**220+** countries and regions with Airbnb listings as of December 31, 2022

**$180B** earned by Hosts, all-time as of December 31, 2022

**100K** cities and towns with active Airbnb listings as of December 31, 2022

*See References*

Problem Statement

# Developing a Revenue Estimation Model for Properties

This involves creating a predictive model that estimates potential revenue 3 from now based on a combination of property characteristics, socio-economic factors, and geographic features.

Examining from the perspective of key stakeholders

# Why is this problem important?

## Airbnb

As per the Q3 2023 shareholder letter, Airbnb's strategic priorities include, "Making hosting mainstream" and 'Perfecting core services"

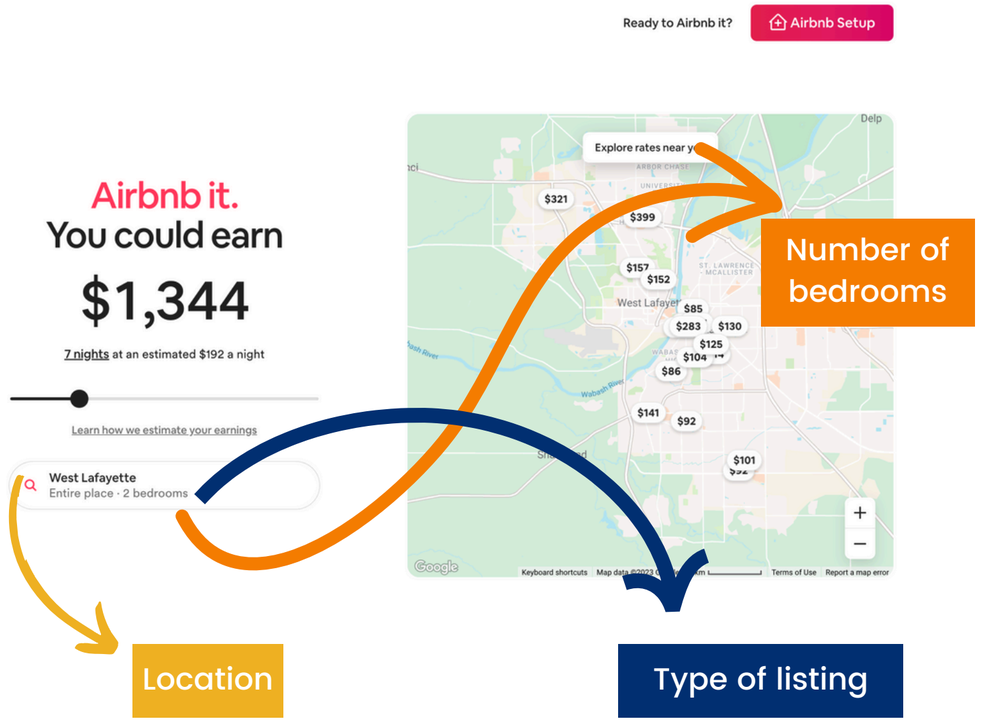
## Hosts

Hosts are interested in predicting the potential revenue they can earn from their property listings

## Guests

As demand for short-term rentals will continue to show strong growth in 2023 with a 5.5% increase in booked stays, guests will benefit from new listings and price variety

*See References*




Interactive tool offered by Airbnb

# What's My Place Worth?

While Airbnb offers a tool that assists prospective hosts in understanding potential earnings from sharing their space (nightly rate), it only takes into account factors such as location, type of listing, and number of bedrooms

*See References*

# 02

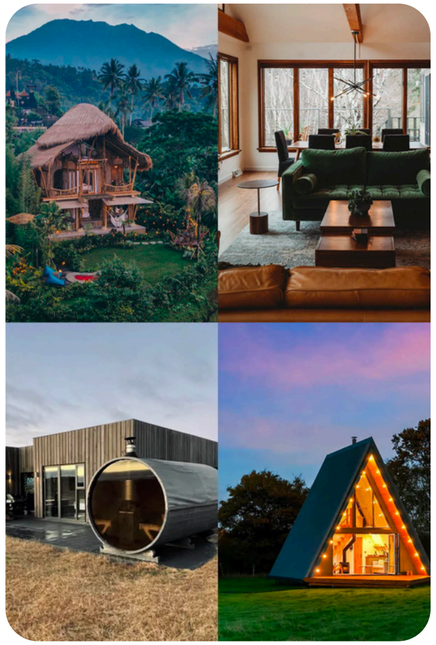
# Data Investigation

Dataset

# Airbnb: Dallas

Each row represents an Airbnb listing's data for a specific Superhost evaluation period, capturing both current and historical performance metrics

- **Number of rows:** 48,711
- **Number of columns**: 95 (few columns are duplicated)
- **Missing values in the 'revenue' column**: 16,861
- **Evaluation periods (superhost_period_all):** 5 to 20
- **Property IDs that do not have information across all evaluation periods:** 9,348
- **prev_xyz variables:** Capture information about previous evaluation period
- **xyz_year variables:** Capture information about previous year



Superhost Evaluation Periods and Corresponding Data Time Frames

| superhost_period_all | Evaluation months | Time period of data checked |
|---|---|---|
| 5 | Jul 1st - Sept 30th, 2016 | Jul 1st - Jun 30th, 2015 |
| 6 | Oct 1st - Dec 31st, 2016 | Oct 1st - Sep 30th, 2015 |
| 7 | Jan 1st - Mar 31st, 2017 | Jan 1st - Dec 31st, 2016 |

# Exploratory Data Analysis

Target variable: ‘revenue’

## 01 Removed Irrelevant Records

- Dropped rows with missing values for ‘revenue’
- Dropped variables that weren’t useful in our prediction or were duplicated
- Dropped variables from the current evaluation period and retained those from the previous period. For instance, we kept 'prev_occupancy_rate' and dropped 'occupancy_rate'

## 02 Handled Missing values

- Imputed missing values for numerical variables using the Mean Absolute Deviation method and for categorical variables using the Modal Absolute Deviation method, after grouping the variables by 'Property Type', 'Zipcode', 'Bedrooms', 'Bathrooms', and 'Max Guests'

## 03 Outlier Treatment

- Used the flooring and capping method for outlier treatment
- Formula used to set upper and lower bounds:
  - lwr_bound = q1-(1.5*IQR)
    upr_bound = q3+(1.5*IQR)

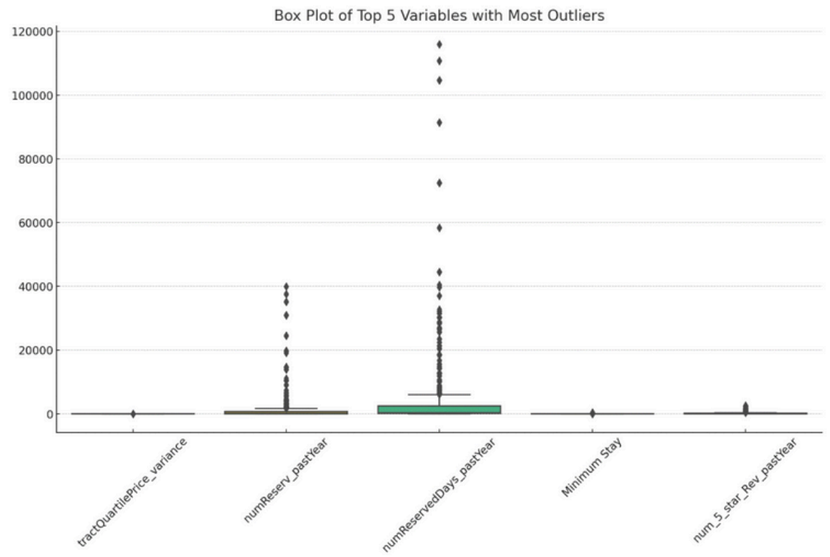
# Exploratory Data Analysis

## Missing values

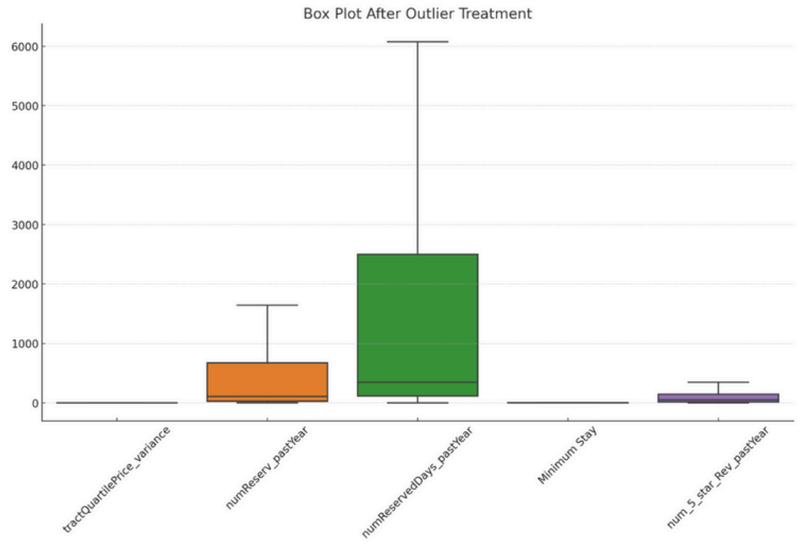
| Column | Null Percentage |
|---|---|
| prev_revenue | 38.90 |
| prev_occupancy_rate | 38.90 |
| prev_booked_days_avePrice | 38.90 |
| prev_booked_days | 38.90 |
| revenue | 34.61 |
| occupancy_rate | 34.61 |
| booked_days | 34.61 |
| booked_days_avePrice | 34.61 |
| prev_prop_5_StarReviews_pastYear | 31.00 |
| prev_rating_ave_pastYear | 31.00 |
| prev_num_5_star_Rev_pastYear | 30.31 |
| prev_numCancel_pastYear | 30.31 |
| prev_numReviews_pastYear | 30.31 |
| prev_Rating Overall | 29.64 |
| prop_5_StarReviews_pastYear | 23.87 |
| rating_ave_pastYear | 23.87 |
| num_5_star_Rev_pastYear | 23.03 |
| numCancel_pastYear | 23.03 |
| numReviews_pastYear | 23.03 |

# Exploratory Data Analysis

## Outlier treatment




Before outlier treatment




After outlier treatment

Steps in Data Transformation

# Feature Creation

- Added a **'quarter' column** to facilitate the analysis of quarterly data
- Used **one-hot encoding to process categorical variables** such as, 'superhost_period_all,' 'Property Type,' 'Listing Type,' ‘quarter,’ 'Neighborhood,' and 'Pets Allowed
- Created a **'date_diff' column** which calculates the number of days from the listing date on Airbnb to January 1st, 2023
- **Converted percentages to actual numbers,** resulting in the creation of two new columns: 'tract_white_count' and 'tract_black_coun

| quarter_Q1 | quarter_Q2 | quarter_Q3 | quarter_Q4 |
|---|---|---|---|
| 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 |

One-hot encoded quarter variable

| date_diff |
|---|
| 4965 |
| 4791 |
| 4471 |
| 4421 |
| 4234 |

date_diff

Lagged variables

# Feature Creation

- For certain features lacking data from the previous evaluation period, we introduced a one-period lag and dropped the original variables.
- **The variables for which we created lagged columns include:**
  - 'Max Guests', 'Cleaning Fee (USD)', 'Minimum Stay', 'Number of Photos', 'tract_superhosts', 'tract_superhosts_ratio', 'tract_price_variance', 'tractQuartilePrice_variance', 'booked_days_period_city', 'revenue_period_city', 'booked_days_period_tract', 'revenue_period_tract', 'tract_booking_share', 'tract_revenue_share'
- Choosing not to include these variables from the same evaluation period in the revenue prediction model is a strategic decision aimed at ensuring the model's insights are both predictive and robust. This also ensures consistency in the dataset.

| | | Dropped | New column |
|---|---|---|---|
| | Airbnb Host ID | tract_price_variance | tract_price_variance_lagged |
| 0 | 18837.0 | 938.562500 | NaN |
| 1 | 18837.0 | 1166.666667 | 938.562500 |
| 2 | 18837.0 | 15774.018750 | 1166.666667 |
| 3 | 18837.0 | 12649.249074 | 15774.018750 |
| 4 | 18837.0 | 93693.260417 | 12649.249074 |

**Illustration for evaluation period 7:**

| timeframe of 'revenue' (target variable) | timeframe of 'tract_price_variance_lagged' |
|---|---|
| Jan 1st, 2017 - Mar 31st, 2017 | Oct 1st, 2016 - Sept 30th, 2017 |

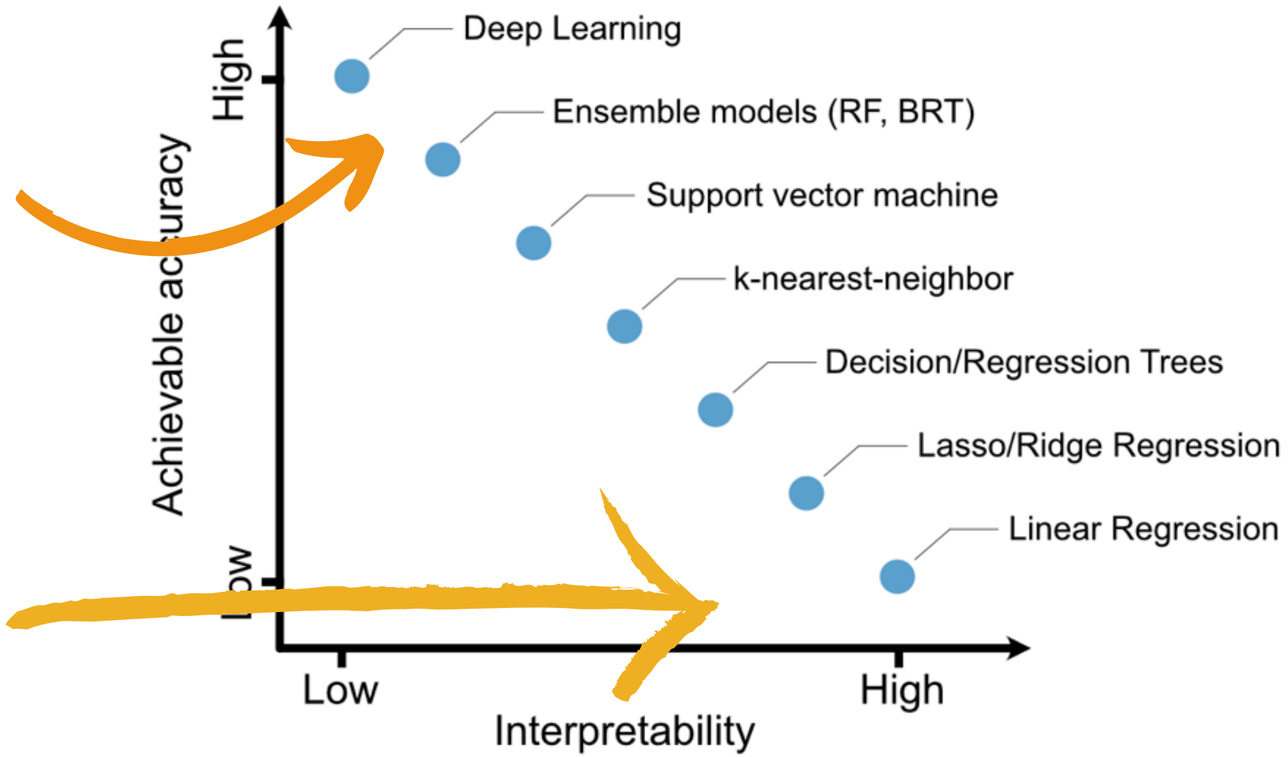
# Model Implementation

03

Balancing Accuracy and Interpretability in Model Development

# Our Approach

1. Developed a high-accuracy model with low interpretability
2. Identified the important variables influencing the model
3. Created another model with the important variables which provides low accuracy and high interpretability




Accuracy v/s Interpretability

# Model Summary

## 01 XG Boost

**R-squared: 0.7124**
Excels in predictive accuracy due to its advanced regularization which prevents overfitting, making it ideal for complex datasets with many features, as it can effectively capture intricate patterns

## 02 Optimised XG Boost

**R-squared: 0.7201**
The grid-search optimization process fine-tunes the model parameters, increasing the R-squared value.

## 03 Variable Identification

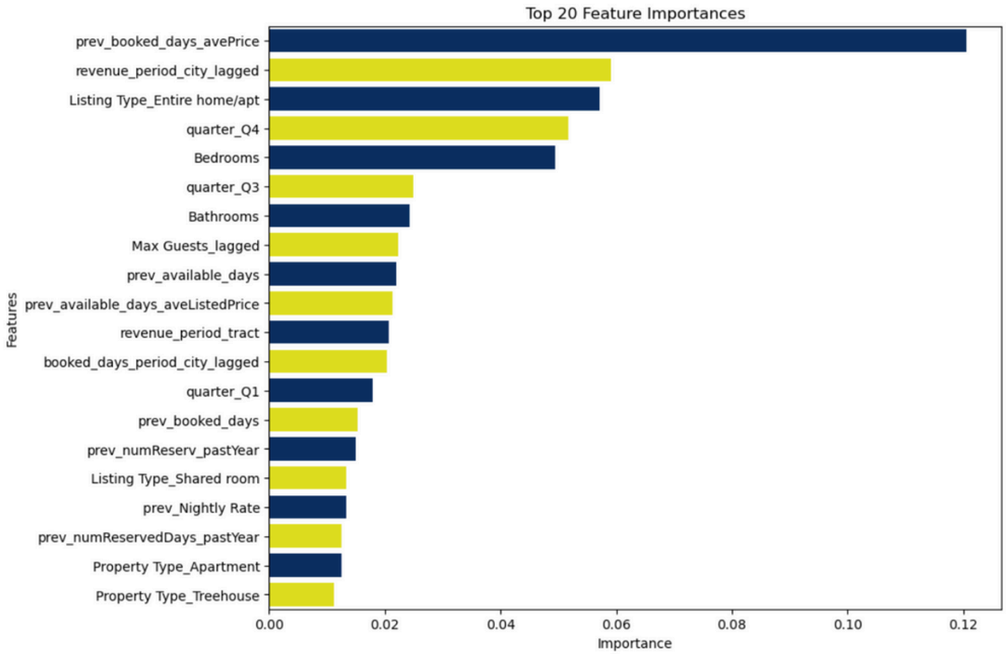
Identifying key variables streamlines the model, this not only improves model efficiency but also aids in understanding the driving forces behind revenue trends.

## 04 Backward Linear Regression

**R-squared: 0.6718**
Provided a simpler, yet effective model that can be particularly useful for stakeholders who require a straightforward model that is easy to interpret and communicate

# Model Summary: XG Boost

## XG Boost

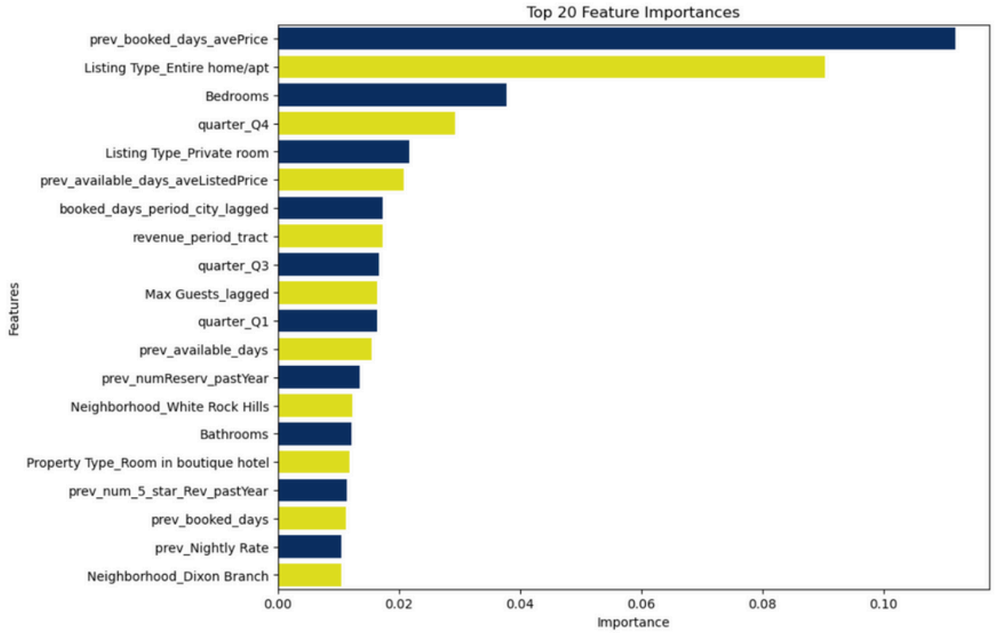



Variable importance

## Optimised XG Boost

```
Fitting 3 folds for each of 81 candidates, totalling 243 fits
Best Parameters: {'learning_rate': 0.05, 'max_depth': 6, 'min_child_weight': 1, 'n_estimators': 300}
```

Gridsearch results




Variable importance

# Model Summary: Backward Regression

Calculator for revenue estimation

Predicted Revenue = **-801.6** + (270.6 X Bathrooms) + (113.0 X Bedrooms) + (385.0 X Listing_Type_Entire_home_apt) - (482.4 X Listing_Type_Shared_room) + (140.3 X Max_Guests_lagged) + (3315.1 X Property_Type_Treehouse) - (0.0350 X booked_days_period_city_lagged) + (1.9543 X prev_Nightly_Rate) + (7.8075 X prev_available_days) - (3.1037 X prev_available_days_aveListedPri) + (17.2615 X prev_booked_days) + (13.7430 X prev_booked_days_avePrice) + (0.5187 X prev_numReserv_pastYear) - (0.0729 X prev_numReservedDays_pastYear) + (228.6 X quarter_Q3) - (464.5 X quarter_Q4) + (0.000197 X revenue_period_city_lagged) + (0.000339 X revenue_period_tract)

**Illustration:**

- Increasing the number of bedrooms by 1 results in a $113 rise in total revenue, all else being equal.
- Adding 1 bathroom leads to a $270 increase in total revenue, with all other factors constant

Property characteristics

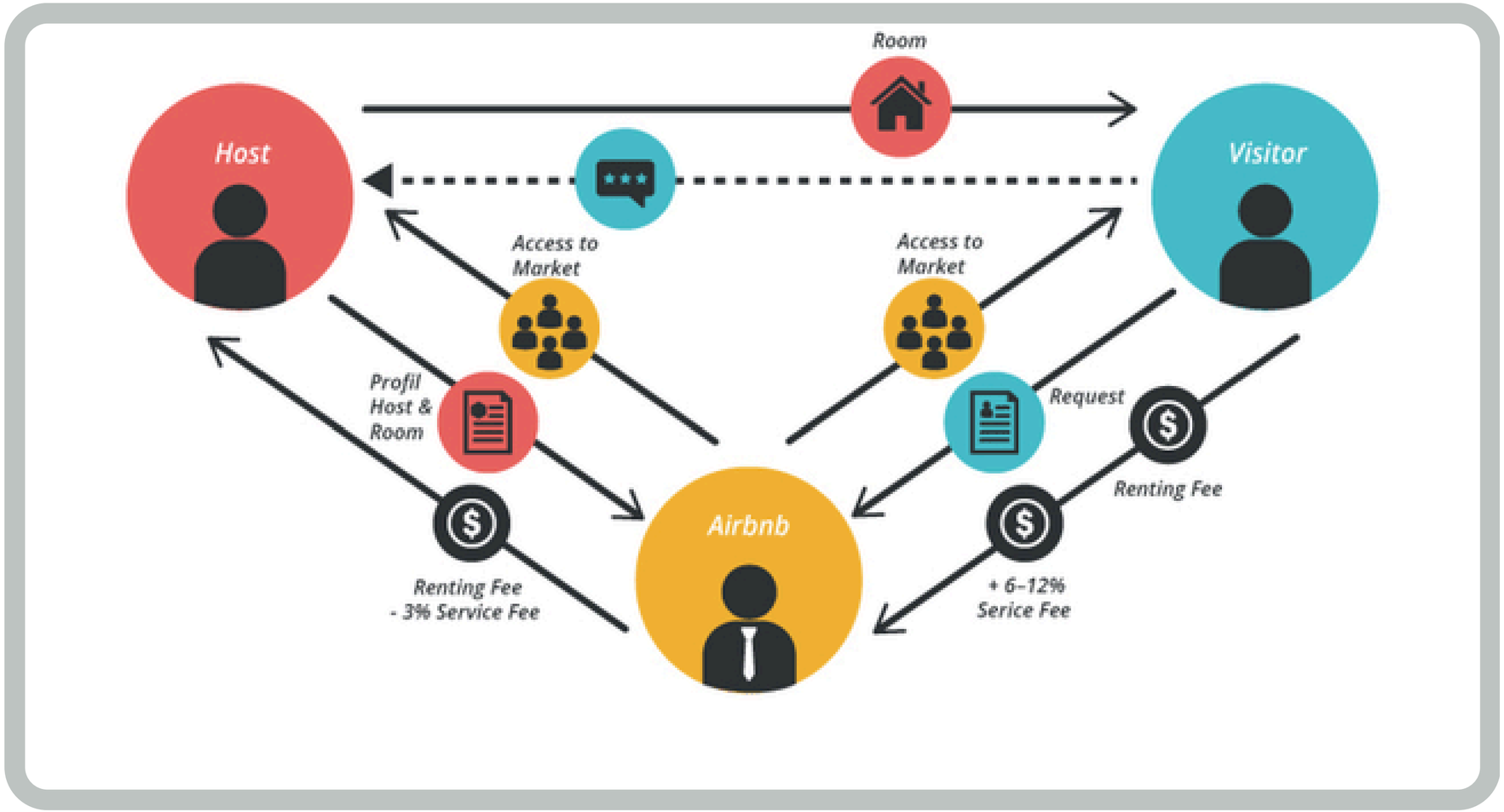
Socio-economic factors

Seasonality

# Business Application

04

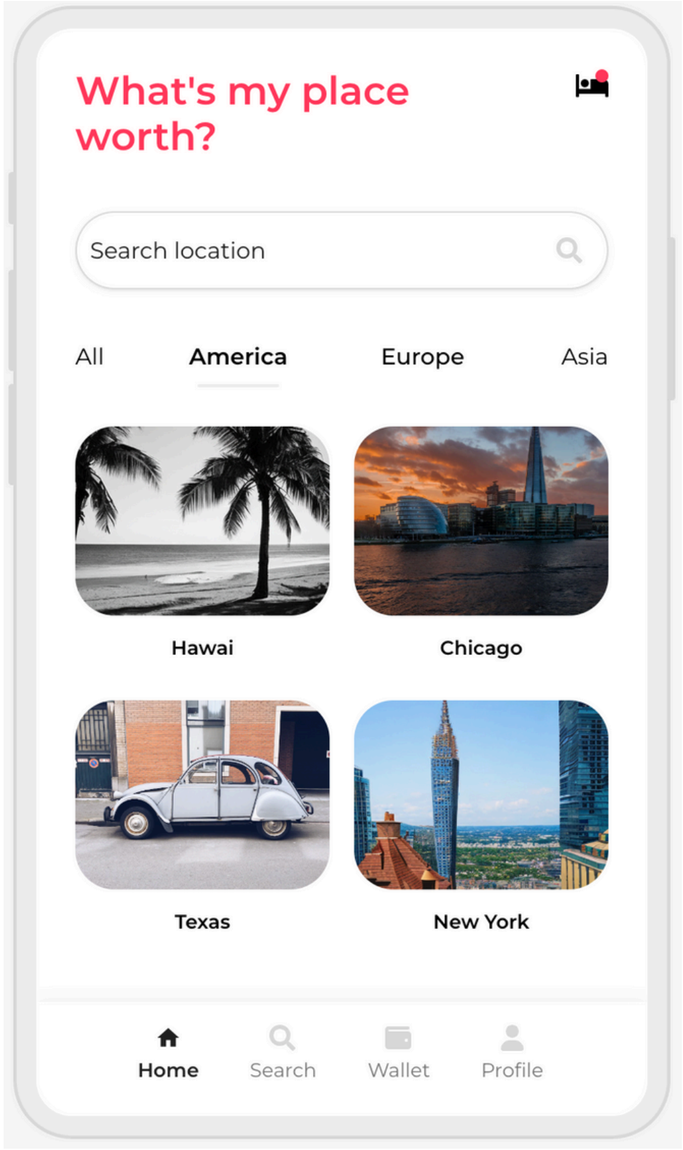
# Network Effects

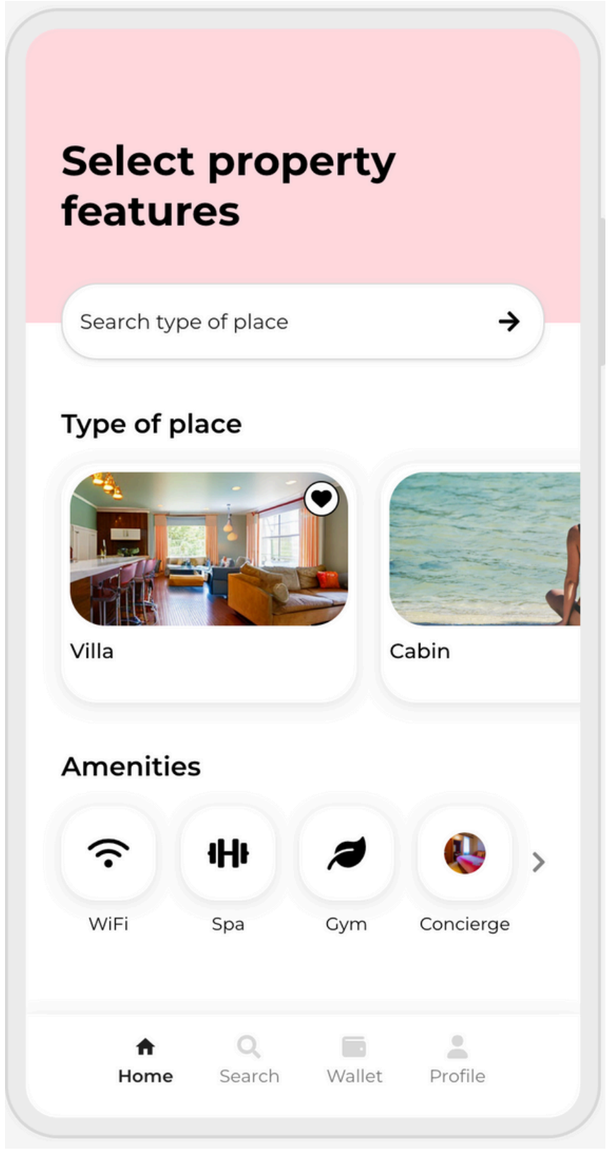


New interactive tool

# App Mockup

We aim to reimagine the existing interactive tool 'What's My Place Worth' to provide hosts or potential hosts with a more accurate estimate of their revenue, based on a broader range of features



Product Portfolio

# Current Host Revenue Forecast

| Variable | Value | Coefficient | Calculated Values |
|---|---|---|---|
| Bathrooms | 2 | 270.6 | 541.2 |
| Bedrooms | 2 | 113 | 226 |
| Listing_Type_Entire_home_apt | 1 | 385 | 385 |
| Listing_Type_Shared_room | 0 | -482.4 | 0 |
| Max_Guests_lagged | 5 | 140.3 | 701.5 |
| Property_Type_Treehouse | 0 | 3315.1 | 0 |
| booked_days_period_city_lagged | 57500 | -0.035 | -2012.5 |
| prev_Nightly_Rate | 160 | 1.9543 | 312.688 |
| prev_available_days | 148 | 7.8075 | 1155.51 |
| prev_available_days_aveListedPri | 139 | -3.1037 | -431.4143 |
| prev_booked_days | 26 | 17.2615 | 448.799 |
| prev_booked_days_avePrice | 126 | 13.743 | 1731.618 |
| prev_numReserv_pastYear | 1286 | 0.5187 | 667.0482 |
| prev_numReservedDays_pastYear | 4748 | -0.0729 | -346.1292 |
| quarter_Q3 | 0 | 228.6 | 0 |
| quarter_Q4 | 1 | -464.5 | -464.5 |
| revenue_period_city_lagged | 48712 | 0.000197 | 9.596264 |
| revenue_period_tract | 201139 | 0.000339 | 68.186121 |
| Intercept | | | -801.6 |

## Dua Lipa

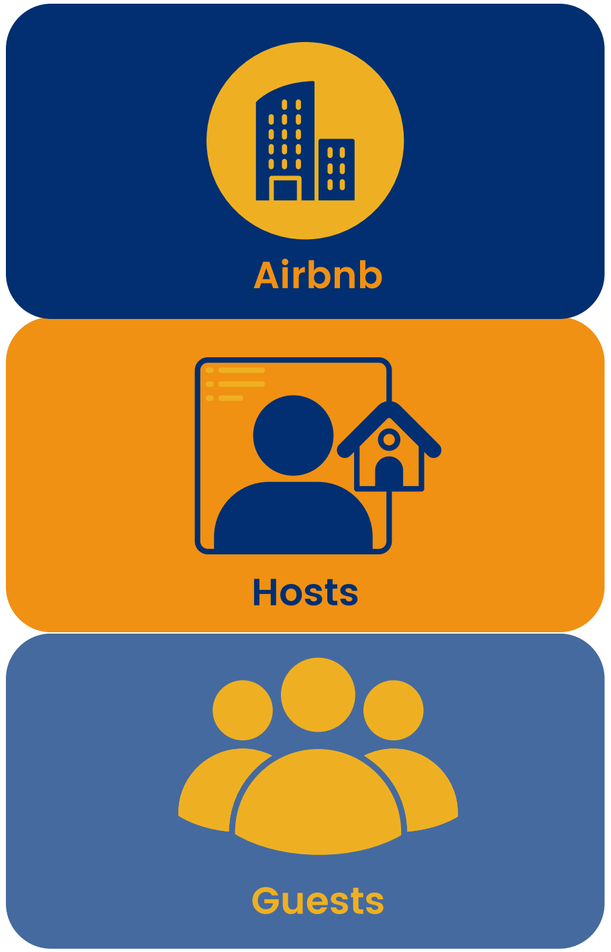
**Airbnb host since 2016**

Dua is interested in maximizing her income using data-driven insights to make informed decisions about property management, pricing strategies, and potential upgrades that could increase her property's appeal and profitability

**Predicted revenue: 2191.002 USD**

# Conclusion

## Examining from the perspective of key stakeholders




- **Supply and Pricing Optimization:** Airbnb's revenue estimates enable hosts to price properties effectively, aligning with market supply and demand
- **Market Intelligence:** Making data-driven decisions in terms of platform updates, policy changes, or new feature rollouts. Airbnb can identify and promote properties with high revenue potential, thereby increasing its commission earnings.

- **Revenue Prediction**: Current and future hosts can benefit by understanding the potential income from their property listings
- **Investment Decisions**: Provides data-driven insights for hosts considering property investments or enhancements to maximize future revenue

- **Improved Listings**: Revenue predictions could motivate hosts to enhance property features, improving guest experience.
- **Availability**: Confidence in revenue generation may encourage hosts to increase the number of available listings, providing guests with more options and possibly lower prices

# Thank You

# References/ Appendix


- https://news.airbnb.com/about-us/
- https://news.airbnb.com/airbnb-q3-2023-financial-results/
- https://community.withairbnb.com/t5/Ask-about-your-listing/Estimating-Revenue/m-p/1704423
- https://www.airdna.co/industry-report/2023-us-airdna-outlook-report
- Interactive tool: https://www.airbnb.com/host/homes

# SAS EM diagram

| Property | Value |
|---|---|
| **General** | |
| Node ID | Part |
| Imported Data | |
| Exported Data | |
| Notes | |
| **Train** | |
| Variables | |
| Output Type | Data |
| Partitioning Method | Default |
| Random Seed | 12345 |
| Data Set Allocations | |
| Training | 60.0 |
| Validation | 40.0 |
| Test | 0.0 |
| **Report** | |
| Interval Targets | Yes |
| Class Targets | Yes |
| **Status** | |
| Create Time | 12/7/23 11:05 PM |
| Run ID | c6c769f7-b520-4f48-ad9b-f3 |
| Last Error | |
| Last Status | Complete |
| Last Run Time | 12/7/23 11:09 PM |
| Run Duration | 0 Hr. 0 Min. 3.92 Sec. |
| Grid Host | |
| User-Added Node | No |