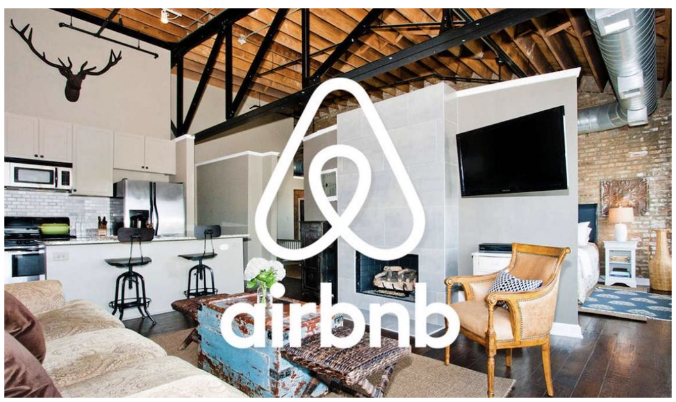
# Linear Nonlinear Models

## Airbnb Analysis - Predicting Sales Prices for Property Listings



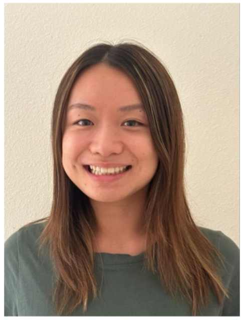

**Madeline Huynh**
**Natalie Kim**
**Juan Bautista**
**Luyao Xie**
**March 6, 2024**

# Agenda

1) Team Introduction
2) Executive Summary
3) Source Data Overview and Data Cleaning
4) Feature Engineering
5) Model Implementation
6) Conclusions
7) Limitations and Recommendations

# *Team Introduction*

**Juan Bautista**
**ML Engineer**



**Madeline Huynh**
**Researcher**

**Natalie Kim**
**Data Scientist**

**Luyao Xie**
**Data Engineer**

# *Executive Summary*

- **Industry Overview:** Airbnb is a platform for individuals to rent out their properties or book accommodations in diverse locations, facilitating unique travel experiences and challenging traditional hospitality models.
- **Problem Statement:** The **dynamic pricing** observed across Airbnb listings nationwide presents a challenge in comprehensively understanding the underlying factors influencing market dynamics, necessitating research to elucidate the complexities and optimize strategies for pricing and market analysis in the hospitality sector.
- **Key Objective:** To **analyze dynamic pricing patterns** across various cities and urban landscapes within the Airbnb marketplace, discern correlations between pricing variations and factors such as **reviews, competitive pricing strategies, and listing popularity**, and ultimately provide actionable insights to hosts to optimize their pricing strategies and enhance the overall user experience. Through extensive analysis, we aim to contribute to a better understanding of the Airbnb marketplace dynamics and facilitate **informed decision-making among hosts**.

# Data Overview

- **Source Data:** Over 250,000 listings and 32 potential covariates
  - Host information, Location information, Property features, and Review scores
  - Response Variable: Listing's Price per Night

**Host**

- Date host joined Airbnb
- How long a host takes to respond
- Acceptance rate host has for their listings
- Superhost
- Verified identify
- Total listings count

**Location**

- Neighborhood
- District
- City
- Latitude
- Longitude

**Property**

- Type of property
- Type of room
- Number of people it accommodates
- Number of bedrooms
- Minimum number of acceptable nights
- Maximum number of acceptable nights
- Bookable instantly

**Reviews**

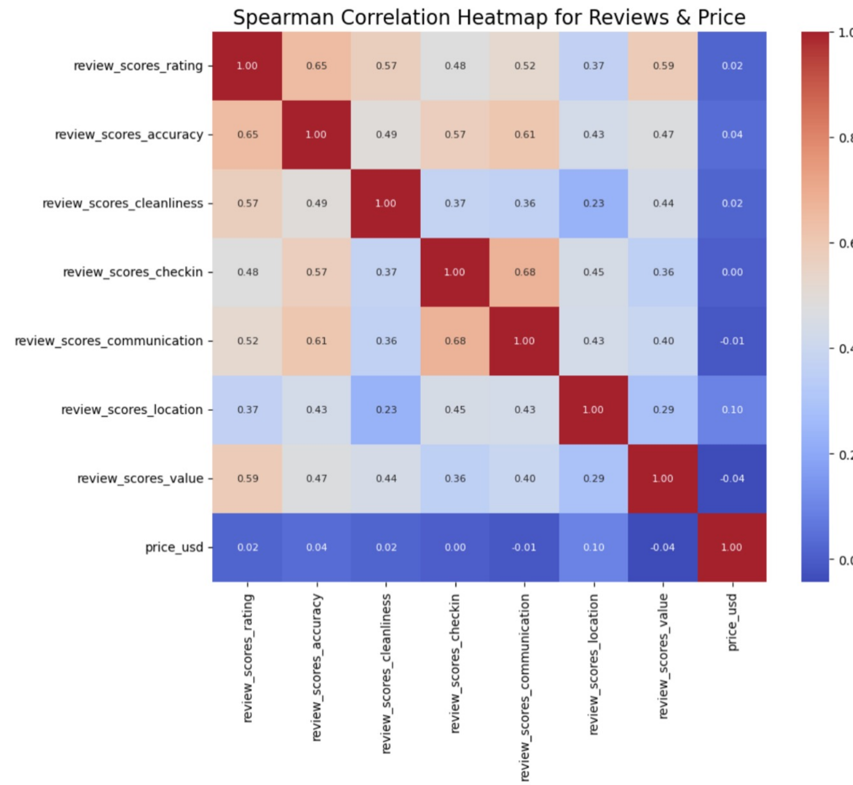
- Overall listing's rating
- Accuracy of unit compared to listing
- Cleanliness score
- Check-in experience score
- Communication score
- Location score
- Value of listing score

# *Data Cleaning*

- **Missing Value Imputation:**
  - 18 predictors missing values
  - Mean & Mode imputation for numerical & categorical predictors
- **Data Type Conversions**
  - One-hot encode categorical predictors (up to 4 categories)
  - Date to duration conversion
- **Response Cleaning & Transformation**
  - Converted local currencies to US dollar
  - Outlier handling
  - Apply log transformation
- **Filtering for Relevant Features:**
  - Removal of District, Host's location, Listing's name, etc.

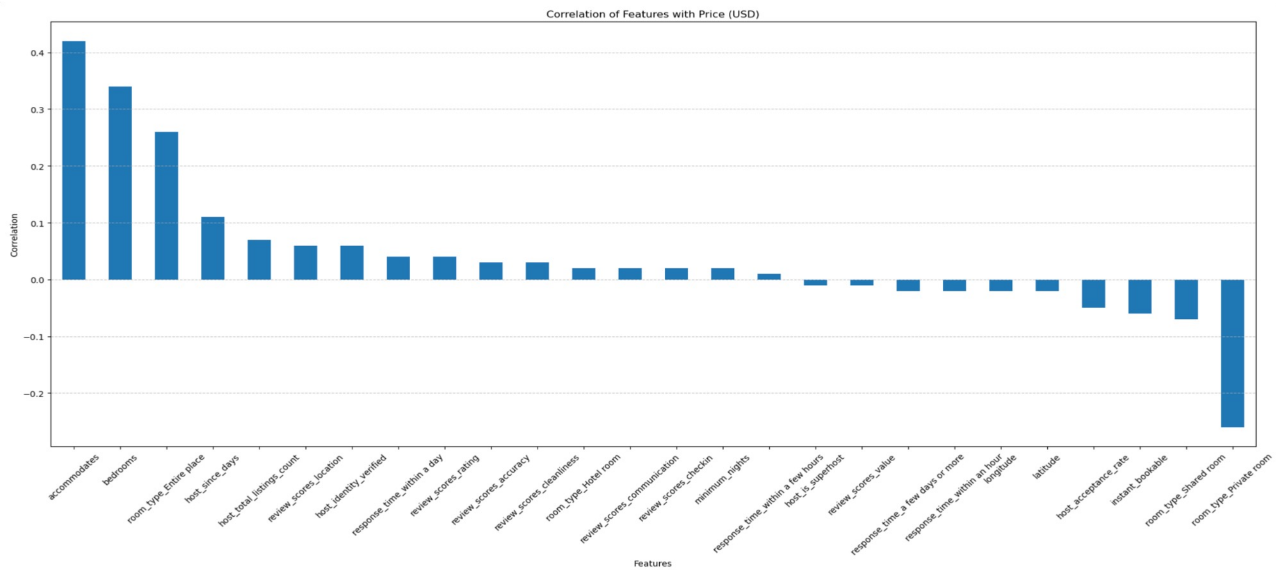
# *Correlation and Principal Component Analysis (PCA)*

- PC1 and PC2 only explain 25.2% of the variance
- PC1 captured all of the review predictors
  - they all had moderate negative influence
- Observed high correlation values between the review predictors



# Feature Selection

- ➢ First used plotted correlation of features with just listing's nightly
- ➢ Applied Recursive Feature Elimination (RFE) with linear regression to identify the top features



# *Mixed Effects Model*

- **A Mixed Effects Model** is a framework that integrate both fixed effects, which represent stable and quantifiable influences present across all data points, and random effects, which capture fluctuations within specific data segments
- For Airbnb listings, fixed effects are the **predictors** and random effects are the **group variables that represent hierarchical data** (i.e. neighborhood, city, etc) nested in different groups to predict the **dependent variable price**
- **Neighborhood** as the Group Variable:
  - Each predictor is statistically significant
  - Standard errors are relatively low
  - **Private rooms, shared rooms, host total listings, and host verified are associated with lower prices**
  - **P- value of the neighborhood variable is 6.288e-22,** suggesting that the neighborhood significantly contributes to the explanation of variation in Airbnb prices
  - We cannot interpret the coefficient of the group variable the same way as we interpret predictor coefficients of a linear regression model due to its hierarchical nature
    - **The coefficient represents the average effect of the neighborhood variable on Airbnb prices, but it doesn't directly quantify the difference between all 650 groups.**

```
                 Mixed Linear Model Regression Results
=========================================================================
Model:                    MixedLM     Dependent Variable:     price_usd
No. Observations:         191921      Method:                 REML
No. Groups:               650         Scale:                  0.3170
Min. group size:          1           Log-Likelihood:         -163499.4670
Max. group size:          10301       Converged:              Yes
Mean group size:          295.3
-------------------------------------------------------------------------
                              Coef.  Std.Err.    z    P>|z| [0.025 0.975]
-------------------------------------------------------------------------
Intercept                     3.995    0.027 146.619 0.000  3.941  4.048
host_response_rate           -0.022    0.001 -15.725 0.000 -0.025 -0.020
host_identity_verified       -0.031    0.001 -22.677 0.000 -0.033 -0.028
host_total_listings_count     0.051    0.001  39.193 0.000  0.048  0.054
accommodates                  0.275    0.002 157.342 0.000  0.271  0.278
bedrooms                      0.063    0.002  38.733 0.000  0.060  0.066
review_scores_rating          0.034    0.001  25.890 0.000  0.031  0.036
host_since_days               0.031    0.001  22.557 0.000  0.028  0.034
response_time_within_an_hour  0.026    0.001  18.109 0.000  0.023  0.029
room_type_Entire_place       -0.088    0.004 -19.705 0.000 -0.097 -0.079
room_type_Private_room       -0.236    0.004 -53.653 0.000 -0.245 -0.228
room_type_Shared_room        -0.122    0.002 -69.960 0.000 -0.125 -0.118
Group Var                     0.449    0.047
=========================================================================
```

# *Mixed Effects Model*

- **City** as the group variable
  - Each predictor is statistically significant
  - Standard errors are relatively low
  - Higher negative log-likelihood value compared to the neighborhood model results
  - Similar to the results with neighborhood as the group variable, the same predictors have a downward influence on Airbnb listing prices with an added predictor of "entire place"
  - **P-value of the city variable is 0.065:** indicates that there may be some evidence of variability in the price between cities, but does not reach the conventional threshold for statistical significance
- **Hosts can optimize their pricing strategies by considering the important features identified in the models**

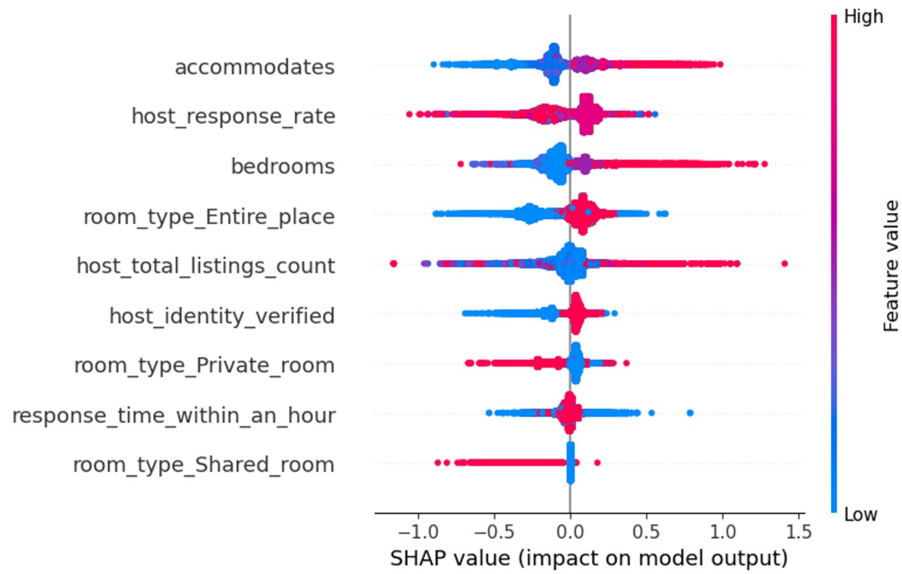
```
                 Mixed Linear Model Regression Results
=====================================================================
Model:                  MixedLM    Dependent Variable:    price_usd
No. Observations:       191921     Method:                REML
No. Groups:             10         Scale:                 0.3540
Min. group size:        4847       Log-Likelihood:        -172764.0618
Max. group size:        44945      Converged:             Yes
Mean group size:        19192.1
---------------------------------------------------------------------
                             Coef.  Std.Err.    z    P>|z| [0.025 0.975]
---------------------------------------------------------------------
Intercept                    3.978    0.191  20.804 0.000  3.604  4.353
host_response_rate          -0.018    0.001 -12.166 0.000 -0.021 -0.015
host_identity_verified      -0.021    0.001 -15.253 0.000 -0.024 -0.019
host_total_listings_count    0.053    0.001  38.874 0.000  0.050  0.056
accommodates                 0.277    0.002 151.967 0.000  0.273  0.280
bedrooms                     0.069    0.002  40.451 0.000  0.066  0.073
review_scores_rating         0.035    0.001  25.332 0.000  0.032  0.037
host_since_days              0.041    0.001  28.986 0.000  0.038  0.044
response_time_within_an_hour 0.019    0.002  12.537 0.000  0.016  0.022
room_type_Entire_place      -0.114    0.005 -24.355 0.000 -0.123 -0.105
room_type_Private_room      -0.284    0.005 -61.771 0.000 -0.293 -0.275
room_type_Shared_room       -0.135    0.002 -73.957 0.000 -0.139 -0.131
Group Var                    0.366    0.242
=====================================================================
```

# Mixed Effects Model Results and Limitations

- **Model Comparisons:** Unlike “city”, “neighborhood” exhibited statistical significance
  - **Granularity of Groups**
    - With over 600 different neighborhoods, “neighborhood” represents the smaller, more localized area compared to “city”, resulting in greater variability in the Airbnb prices
  - **Heterogeneity**
    - Neighborhoods exhibit greater heterogeneity in respects to socioeconomic factors, leading to pronounced differences in the prices
  - **Sample Size**
    - With a  larger number of observations within each neighborhood group, the model may be able to have more precision in estimating neighborhood-level effects
- **Model Precautions:**
  - **Log-Likelihood:** High, negative log-likelihood values across both models
    - Although a model with a higher log-likelihood value (or a more negative value) indicates a better fit for the data, we need to exercise caution due to potential overfitting with model complexity
  - **Multicollinearity:** Potential correlation between predictors
    - Although we are not observing high standard errors, it is important to note that because of potential overfitting, the relationship between predictors and price may be obscured

# *SHapley Additive exPlanations (SHAP)*

- Serves as a unified prediction interpretation framework designed to explain the predictions made by machine learning models.
- SHAP assigns an importance value (SHAP value) to each feature for a particular prediction, calculating the marginal contribution of each feature to the model's prediction and providing detailed explanations for each prediction made by the model.
- From our results:
    - ***accommodates*** seems to have a high and varying positive impact on the price, with higher values generally leading to higher predicted prices. This also seems to be the case with ***bedrooms***.
    - ***host_response_rate*** has a mostly negative impact on the price, meaning that as the host response rate increases, the price decreases.
    - ***room_type*** is important as well, whether the listing is for the entire place is one of the top variables impacting the model.



# *Conclusions*

- How many people a listing ***accommodates*** is more impactful to price than the number of ***bedrooms***.
  - This tells us that customers care more about how many people fit in the space rather than how many rooms are available.
- ***Room type*** (or more accurately, listing type) has a negative impact on the price – shared and private room listings having a much larger negative impact.
  - Customers prefer to have access to the entire place and are willing to pay more for that privilege.
- ***Host response rate*** has a inverse relationship with price.
  - This may warrant further investigation as with the current data we do not have a way of definitively understanding it.

# Constraints of the Analysis

- **Limited Model Availability:**
  - In attempting to use PCA to limit dimensionality, the first two components only explain 25.2% of the variance.
  - With fewer suitable models available, our ability to generate accurate insights may be compromised.
- **Model Assumptions:**
  - **Skewness of Data:**
    - Although we took the log transformation of price to mitigate skewness and improve normality, this does not guarantee that the transformation will adhere to a normal distribution.
  - **Multicollinearity:**
    - May lead to potential model instability and unreliability of predictive insights.
- **Data Quality and Availability:**
  - Due to the complexity of the Airbnb listings data (i.e. incompatible data types, nonsensical values), some predictors cannot be easily used for XAI models.
  - Property attributes exhibit a diverse set of values (i.e. many property types with some not as common) compromising predictive accuracy and generalizability.

# *Recommendations*

- **Incorporate a dynamic pricing tool**
  - Shown as a suggestion to hosts when creating listings
  - Sent as a pricing adjustment recommendation to hosts that are receiving less bookings due to overpricing.
- **Share findings with hosts in the form of recommendations**
  - Recommend hosts capitalize on the available space by purchasing/offering air mattresses, sofa beds, bunk beds etc. which will help increase the number of people a location can accommodate.
- **Further data collection and analysis**
  - In order to better understand the complex relationship between variables such as ***host_response_rate*** and price, more research may be necessary.