USC Viterbi
School of Engineering

# ISE 535 - FINAL PROJECT

## AIRBNB PRICE FORECASTING

Beulah Kannan

# AIRBNB PRICE FORECASTING
## CASE STUDY

- Business objective
- Exploratory Data Analysis (EDA) report
- Statistical analysis
- Predictive modeling for inference
- Cluster analysis
- Summary and conclusions

USC Viterbi
School of Engineering

# AIRBNB PRICE FORECASTING

## BUSINESS OBJECTIVE

- Understand the key drivers of nightly pricing for Airbnb listings.
- Identify natural groupings (clusters) of listings that align with consumer market segments (e.g., budget, family, luxury).
- Deliver actionable insights that could be used by hosts, platform operators, or consumers.

# AIRBNB PRICE FORECASTING

## DATA CHARACTERISTICS

DATA CHARACTERISTICS:

- ‘Airbnb_df.csv’ includes data on 3,982 Airbnb listings from a major metropolitan area, featuring 15 attributes per listing.
- Each row represents a unique listing with a combination of property, host, and pricing attributes.
- The dataset supports analysis of factors that influence nightly pricing and exploration of natural groupings of listings.

# EXPLORATORY DATA ANALYSIS REPORT

# EXPLORATORY DATA ANALYSIS REPORT

## OUTLINE

- Dataset summary
- Univariate analysis
- Bivariate/multivariate analysis
- Summary and next steps

# DATASET SUMMARY

## BASIC CHARACTERISTICS

- Intended use:
- **Price optimization and revenue forecasting** for Airbnb listings
- **Predictive modeling** of nightly rates based on listing features (e.g., location, amenities, property type)
- **Exploratory data analysis (EDA)** to uncover patterns in guest preferences, pricing trends, and listing performance across seasons and locations

USC Viterbi School of Engineering

# DATASET SUMMARY

## DATA DICTIONARY

| Variable Name | Data Type | Description |
|---|---|---|
| listing_id | int | Unique identifier for the listing |
| property_type | object | Type of property (Standard apartment, Tiny Studio, Luxury Home) |
| number_of_bedrooms | int | Number of bedrooms in the listing |
| guest_capacity | int | Maximum number of guests allowed |
| location_score | float | Numeric score representing location desirability |
| review_score | float | Average guest review score (1.0 to 5.0) |
| amenities_count | int | Count of amenities offered |
| host_response_time | object | How quickly the host typically responds |
| Season | object | Season of listing (peak, shoulder, off-peak) |
| minimum_stay_nights | int | Minimum number of nights required |
| years_as_host | float | Number of years the host has been active |
| cleaning_fee | float | Cleaning fee in USD |
| cancellation_policy | object | Cancellation policy (flexible, strict, moderate) |
| nightly_rate | float | Price per night in USD |

USC Viterbi
School of Engineering

# DATASET SUMMARY

## VARIABLE CLASSIFICATION

| Variable Name | Measure/Category | Scale Type | Continous/Discrete |
|---|---|---|---|
| listing_id | Categorical | Nominal | Discrete |
| property_type | Categorical | Nominal | Discrete |
| number_of_bedrooms | Quantitative | Ratio | Discrete |
| guest_capacity | Quantitative | Ratio | Discrete |
| location_score | Quantitative | Interval | Continuous |
| review_score | Quantitative | Interval | Continuous |
| amenities_count | Quantitative | Ratio | Discrete |
| host_response_time | Categorical | Ordinal | |
| Season | Categorical | Nominal | Discrete |
| minimum_stay_nights | Quantitative | Ratio | Discrete |
| years_as_host | Quantitative | Ratio | Continuous |
| cleaning_fee | Quantitative | Ratio | Continuous |
| cancellation_policy | Categorical | Ordinal | |
| nightly_rate | Quantitative | Ratio | Continuous |

# EXPLORATORY DATA ANALYSIS

## DATA QUALITY SUMMARY

› MISSING VALUES:

```
[25]: df.isnull().sum()
```

```
[25]: listing_id            0
      property_type         0
      number_of_bedrooms    0
      guest_capacity        0
      location_score        0
      review_score          0
      amenities_count       0
      host_response_time    0
      season                0
      minimum_stay_nights   0
      years_as_host         0
      cleaning_fee          0
      cancellation_policy   0
      nightly_rate          0
      dtype: int64
```

# EXPLORATORY DATA ANALYSIS

## DATA QUALITY SUMMARY

### › DATA TYPES

```
Data Types Summary:

listing_id              int64
property_type          object
number_of_bedrooms      int64
guest_capacity          int64
location_score        float64
review_score          float64
amenities_count         int64
host_response_time     object
season                 object
minimum_stay_nights     int64
years_as_host         float64
cleaning_fee          float64
cancellation_policy    object
nightly_rate          float64
dtype: object

Categorical Variables (to be encoded):
- property_type
- host_response_time
- season
- cancellation_policy
```

# EXPLORATORY DATA ANALYSIS

## STATISTICAL SUMMARY

› Statistical Summary of numerical features

Numerical Summary Table (with Skew & Kurtosis):

| | count | mean | std | min | 25% | 50% | 75% | max | skew | kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|
| **listing_id** | 3982.0 | 101990.500000 | 1149.648714 | 100000.00 | 100995.2500 | 101990.50 | 102985.7500 | 103981.00 | 0.000000 | -1.200000 |
| **number_of_bedrooms** | 3982.0 | 2.185836 | 1.373346 | 1.00 | 1.0000 | 2.00 | 3.0000 | 5.00 | 0.903153 | -0.522106 |
| **guest_capacity** | 3982.0 | 3.182320 | 1.604411 | 1.00 | 2.0000 | 3.00 | 4.0000 | 7.00 | 0.581191 | -0.412250 |
| **location_score** | 3982.0 | 70.315698 | 11.131437 | 33.33 | 62.7500 | 70.43 | 77.7800 | 107.26 | -0.020977 | -0.037316 |
| **review_score** | 3982.0 | 4.003890 | 0.559780 | 1.99 | 3.6300 | 4.01 | 4.3900 | 5.00 | -0.165083 | -0.344612 |
| **amenities_count** | 3982.0 | 7.943998 | 3.349540 | 3.00 | 5.0000 | 8.00 | 10.0000 | 20.00 | 0.504965 | -0.123790 |
| **minimum_stay_nights** | 3982.0 | 2.272476 | 1.236366 | 1.00 | 1.0000 | 2.00 | 3.0000 | 5.00 | 0.688846 | -0.464123 |
| **years_as_host** | 3982.0 | 3.003943 | 2.979756 | 0.00 | 0.9000 | 2.10 | 4.1000 | 22.90 | 1.943478 | 5.139720 |
| **cleaning_fee** | 3982.0 | 36.306449 | 19.792305 | -4.83 | 20.5625 | 32.67 | 48.7050 | 111.10 | 0.680534 | -0.182132 |
| **nightly_rate** | 3982.0 | 180.307187 | 120.990889 | 30.00 | 99.3000 | 147.25 | 226.0475 | 1000.00 | 2.116051 | 7.120920 |

Years as host and nightly rate show some skew – financial variables

# EXPLORATORY DATA ANALYSIS
## LOGICAL GROUPING OF ATTRIBUTES

### Identifiers & Listing Context
- listing_id
- property_type
- season

### Host Information
- host_response_time
- years_as_host
- cancellation_policy

### Guest Capacity & Property Features
- number_of_bedrooms
- guest_capacity
- amenities_count
- minimum_stay_nights

### Scores & Reviews
- review_score
- location_score

### Pricing
- nightly_rate
- cleaning_fee

# EXPLORATORY DATA ANALYSIS

## VARIABLE SELECTION

```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3982 entries, 0 to 3981
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   listing_id           3982 non-null   int64
 1   property_type        3982 non-null   object
 2   number_of_bedrooms   3982 non-null   int64
 3   guest_capacity       3982 non-null   int64
 4   location_score       3982 non-null   float64
 5   review_score         3982 non-null   float64
 6   amenities_count      3982 non-null   int64
 7   host_response_time   3982 non-null   object
 8   season               3982 non-null   object
 9   minimum_stay_nights  3982 non-null   int64
 10  years_as_host        3982 non-null   float64
 11  cleaning_fee         3982 non-null   float64
 12  cancellation_policy  3982 non-null   object
 13  nightly_rate         3982 non-null   float64
dtypes: float64(5), int64(5), object(4)
memory usage: 435.7+ KB
```

Listing_id is a unique identifier so we should remove that from our model. The target variable nighlty_price should also be removed

# EXPLORATORY DATA ANALYSIS

## STATISTICAL SUMMARY

› Statistical Summary of categorical features

```
Categorical Variables Summary:

property_type          | Count: 3982 | Unique:  3 | Top: Standard Apartment | Freq: 1988
host_response_time     | Count: 3982 | Unique:  3 | Top: within an hour     | Freq: 1979
season                 | Count: 3982 | Unique:  3 | Top: peak               | Freq: 1607
cancellation_policy    | Count: 3982 | Unique:  3 | Top: flexible           | Freq: 1998
```

Most Airbnb properties are standard properties and the cancellation policy is flexible

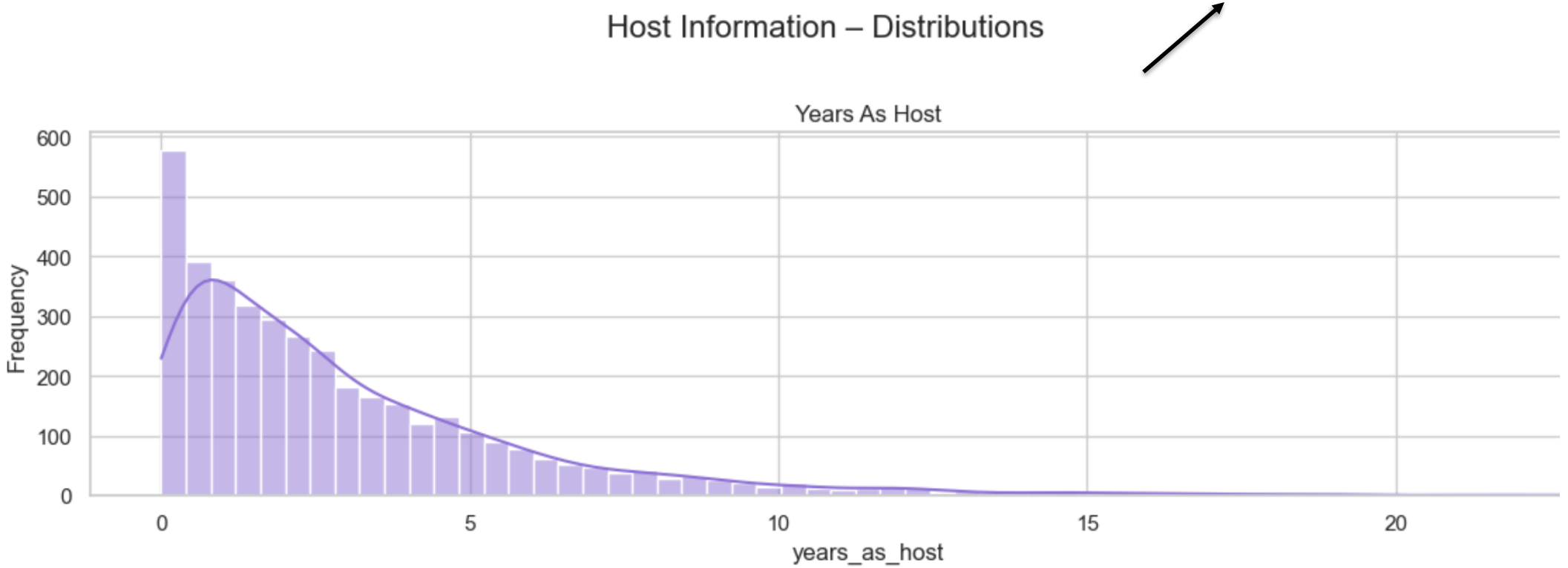
USC Viterbi School of Engineering

# UNIVARITE ANALYSIS
## HOST INFORMATION

📊 Host Information – Statistical Summary:

| | mean | std | min | max | skew | kurtosis |
|---|---|---|---|---|---|---|
| **years_as_host** | 3.0 | 2.98 | 0.0 | 22.9 | 1.94 | 5.14 |

The years_as_host variable is right-skewed (skew = 1.94), indicating that most hosts are relatively new, while a few have significantly longer experience




According to the domain knowledge, Airbnb started in 2008, so capping at 17 would be ideal here

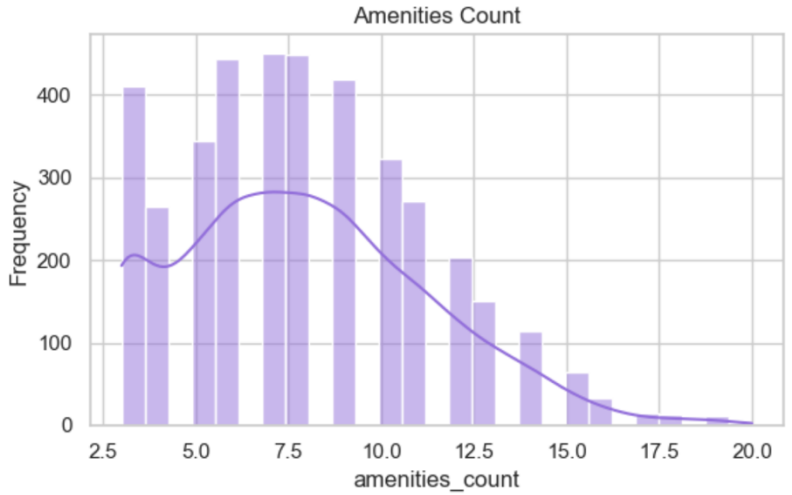
USC Viterbi School of Engineering

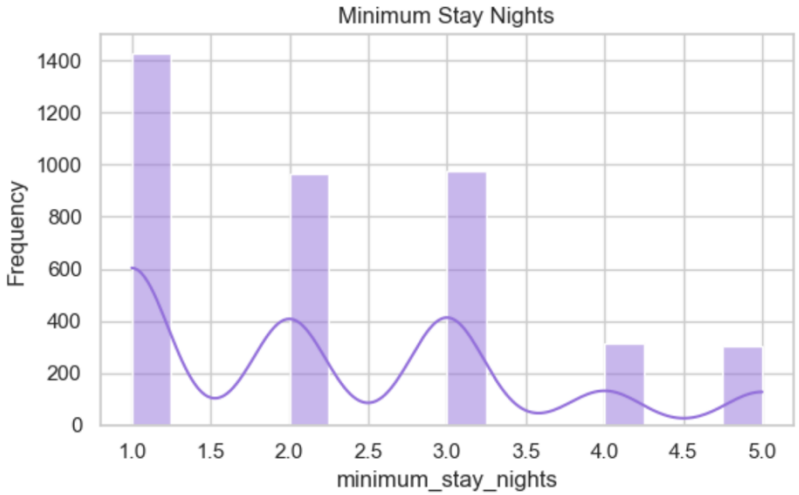
# UNIVARIATE ANALYIS

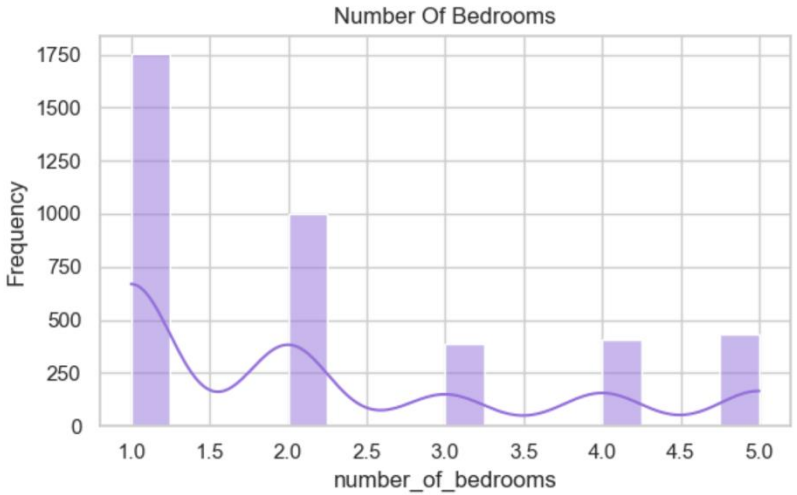
## GUEST CAPACITY AND PROPERTY FEATURES

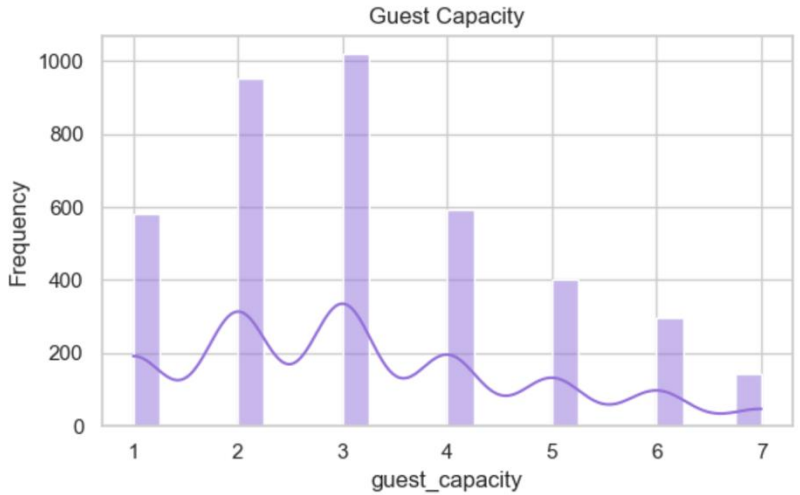
Guest Capacity & Property Features — Statistical Summa

| | mean | std | min | max | skew | kurtosis |
|---|---|---|---|---|---|---|
| **number_of_bedrooms** | 2.19 | 1.37 | 1.0 | 5.0 | 0.90 | -0.52 |
| **guest_capacity** | 3.18 | 1.60 | 1.0 | 7.0 | 0.58 | -0.41 |
| **amenities_count** | 7.94 | 3.35 | 3.0 | 20.0 | 0.50 | -0.12 |
| **minimum_stay_nights** | 2.27 | 1.24 | 1.0 | 5.0 | 0.69 | -0.46 |




The guest capacity and property-related features are mostly low-skewed or near-symmetric. Most listings have 1–2 bedrooms, accommodate 2–4 guests, and offer 5–10 amenities. A spike at 1-night minimum stay reflects short-term rental preferences

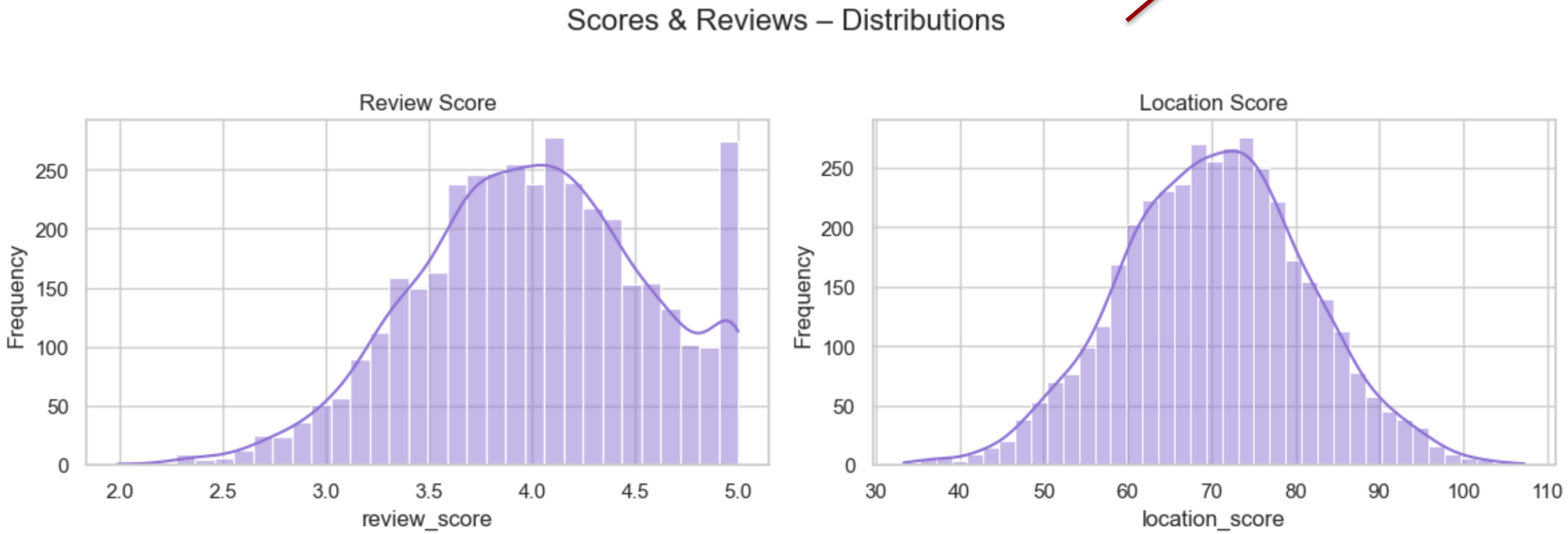
USC Viterbi School of Engineering

# UNIVARIATE ANALYSIS
## SCORES AND REVIEWS

Scores & Reviews – Statistical Summary:

| | mean | std | min | max | skew | kurtosis |
|---|---|---|---|---|---|---|
| **review_score** | 4.00 | 0.56 | 1.99 | 5.00 | -0.17 | -0.34 |
| **location_score** | 70.32 | 11.13 | 33.33 | 107.26 | -0.02 | -0.04 |

Both review and location scores are approximately symmetric and unimodal. Most listings have high review ratings (around 4–5) and location scores concentrated near 70, indicating generally positive guest experiences and well-rated neighborhoods



# UNIVARIATE ANALYSIS

## PRICING

Pricing – Statistical Summary:

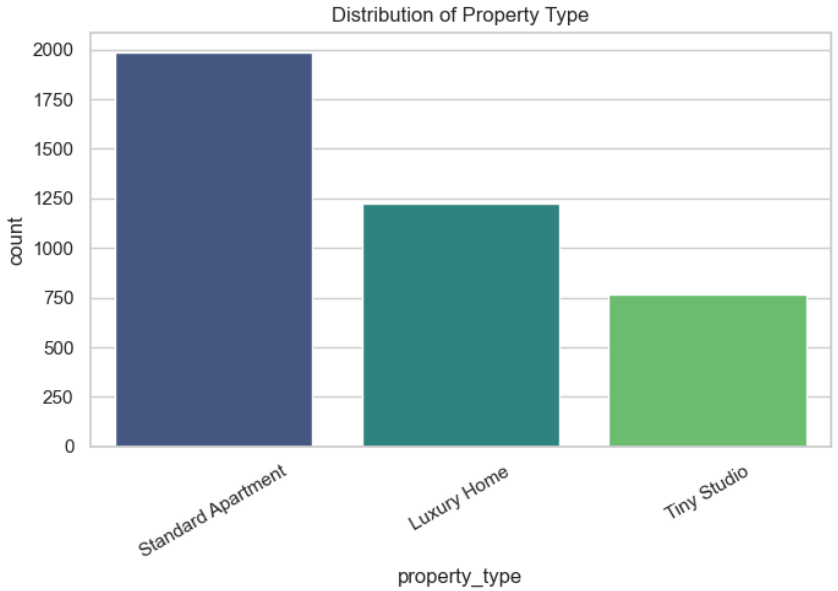
| | mean | std | min | max | skew | kurtosis |
|---|---|---|---|---|---|---|
| **nightly_rate** | 180.31 | 120.99 | 30.00 | 1000.0 | 2.12 | 7.12 |
| **cleaning_fee** | 36.31 | 19.79 | -4.83 | 111.1 | 0.68 | -0.18 |

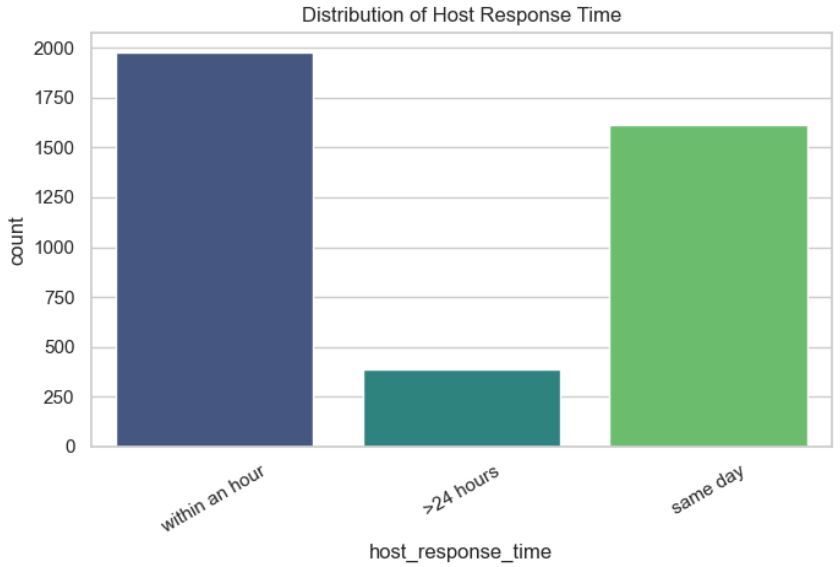
Both nightly rate and cleaning fee are right-skewed(as expected for financial variables), with most listings priced under $300 per night and cleaning fees concentrated below $50, indicating a few high-priced outliers influencing the distribution – need to log transform the target variable as well as cleaning fee.

USC Viterbi School of Engineering

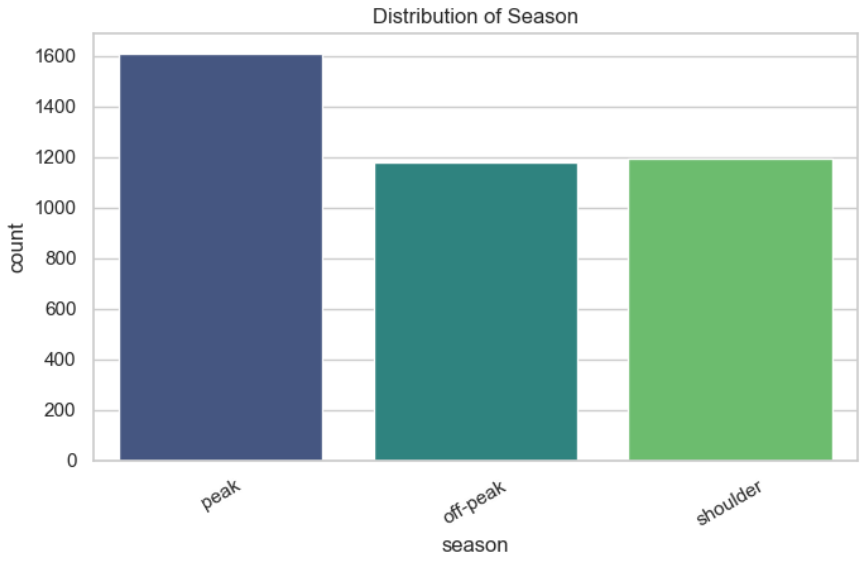
# UNIVARIATE ANALYIS

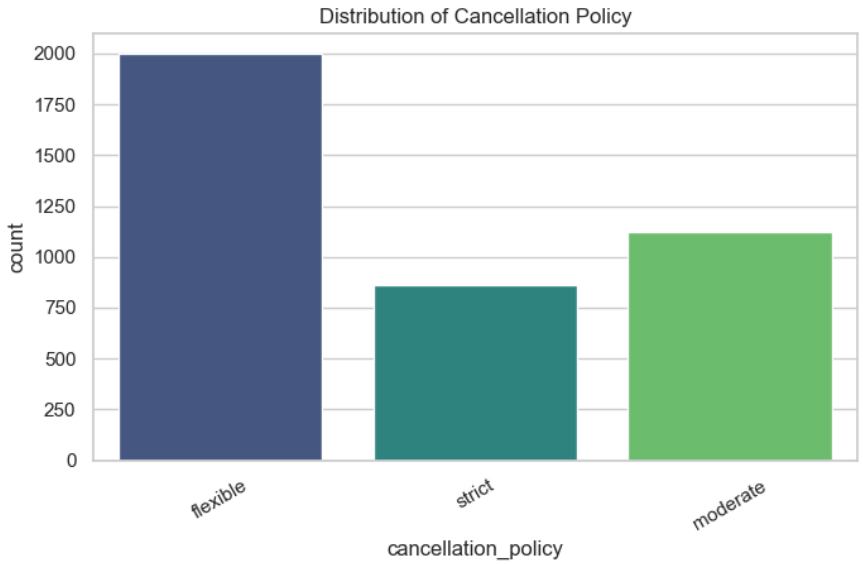
## CATEGORICAL VARIABLE DISTRIBUTIONS



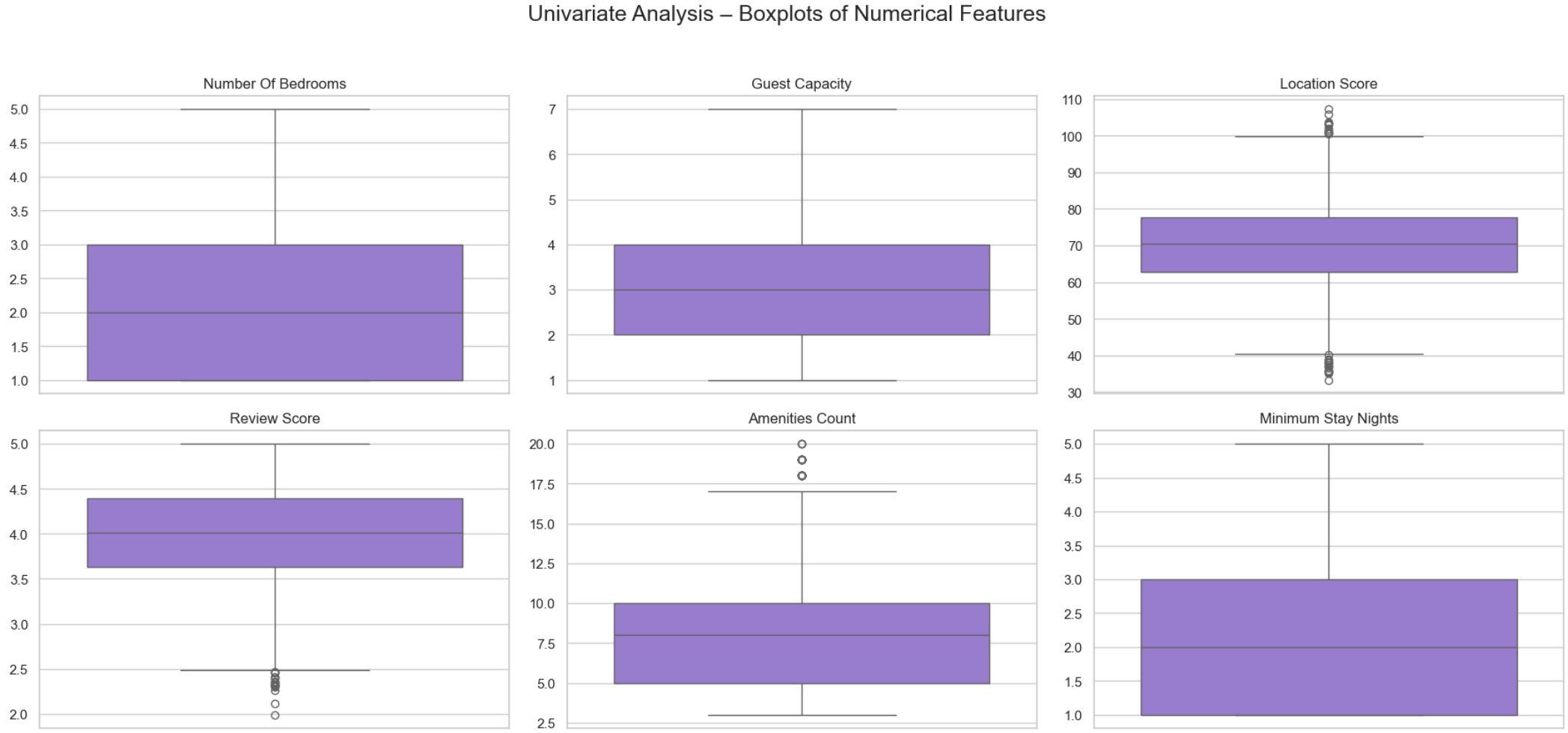

Most listings are standard apartments, hosted by quick responders (within an hour or same day), predominantly available in the peak season, and follow a flexible cancellation policy, indicating platform preferences for convenience and responsiveness



# UNIVARIATE ANALYSIS

UNIVARIATE ANALYSIS



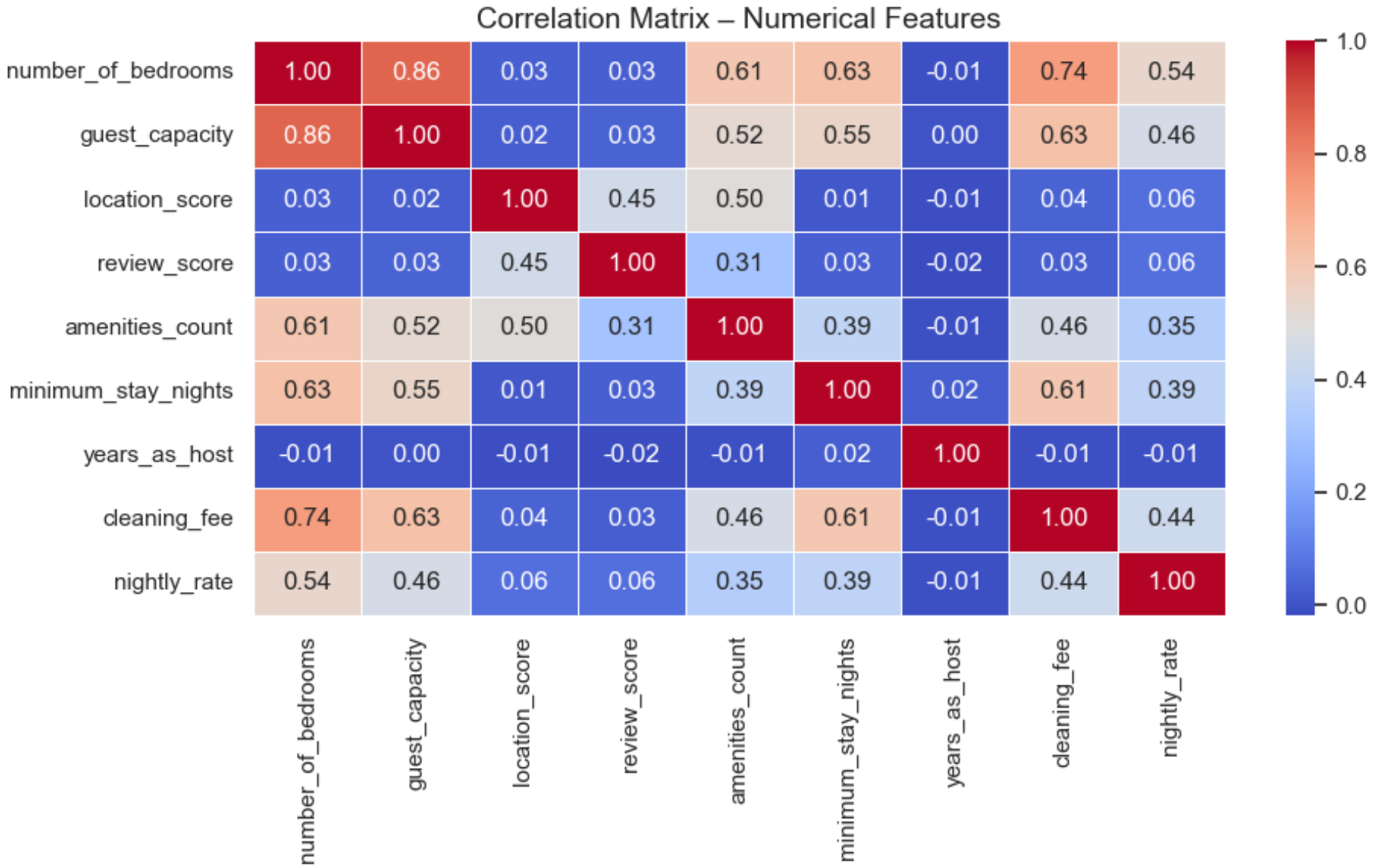
# UNIVARIATE ANALYSIS
## BOXPLOT OF MEASURES


The boxplots reveal the distribution and spread of each numeric feature. Most variables, like *guest_capacity* and *number_of_bedrooms*, show compact IQRs with some mild outliers. In contrast, *cleaning_fee*, *years_as_host*, and *nightly_rate* exhibit heavy right-skew with significant outliers, suggesting the need for transformation or capping in modeling.

USC Viterbi School of Engineering

# BIVARIATE ANALYSIS

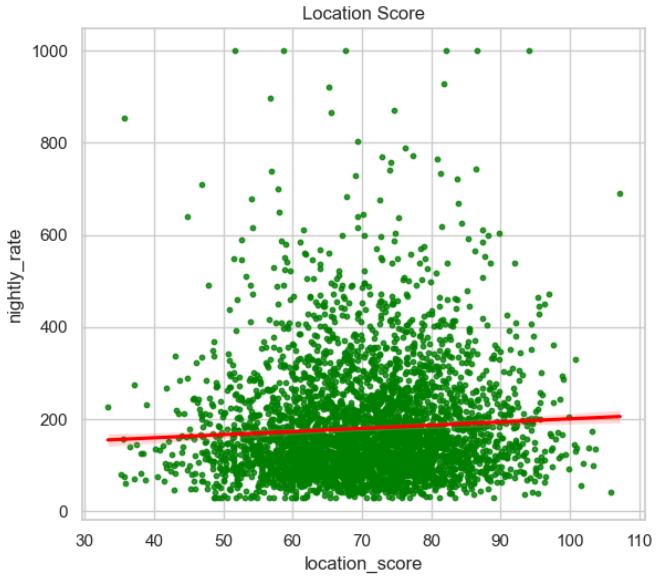
## CORRELATION MATRIX

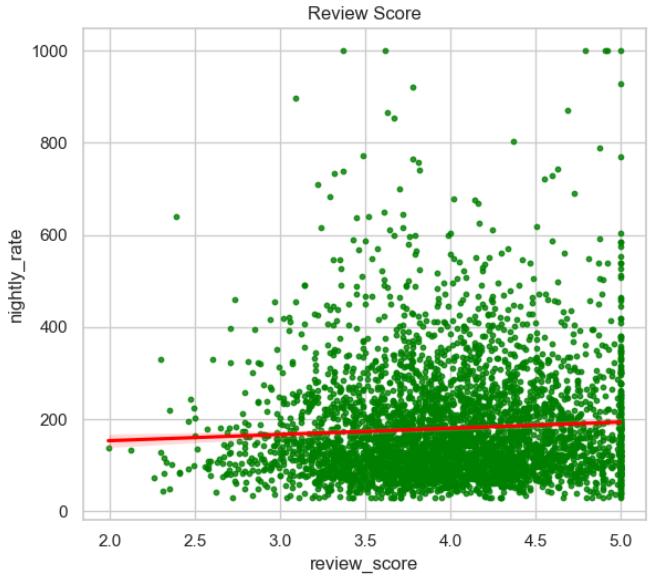



The matrix reveals strong positive correlations between *number_of_bedrooms*, *guest_capacity*, and *cleaning_fee*, indicating that larger listings tend to accommodate more guests and charge higher fees.

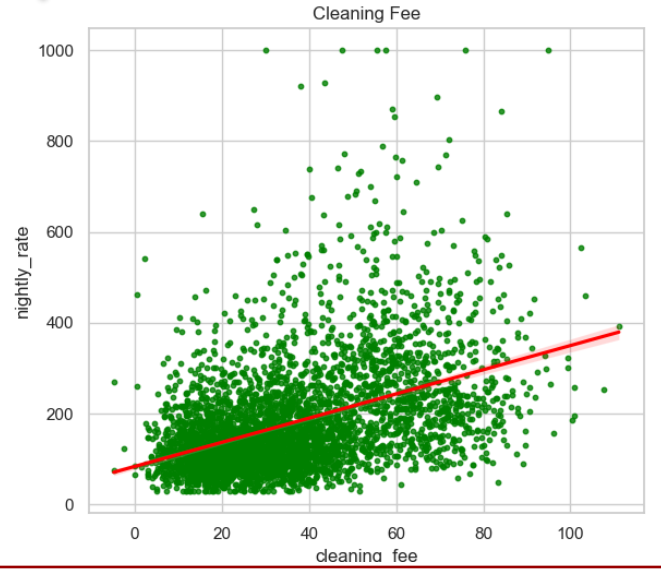
USC Viterbi
School of Engineering

# BIVARIATE ANALYSIS
# SCATTERPLOTS



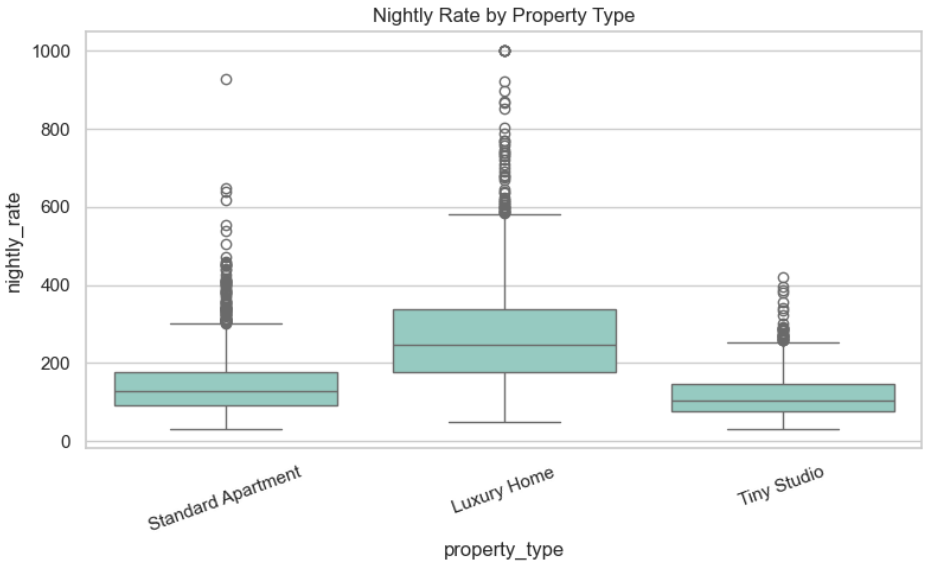

The scatterplots show weak or no linear relationship between *nightly_rate* and variables like *location_score*, *review_score*, and *years_as_host*, with the exception of *cleaning_fee*, which shows a clearer upward trend. This suggests non-linear effects may be better captured using tree-based models

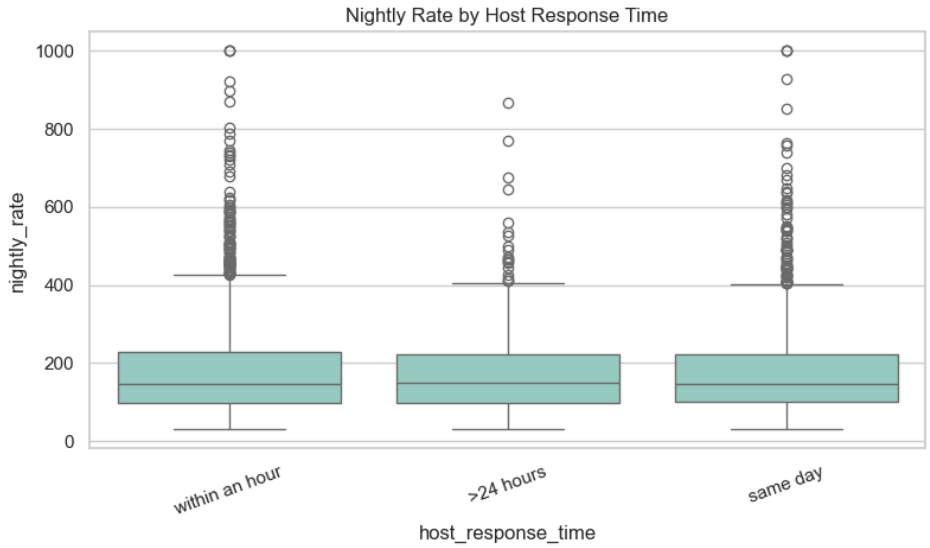



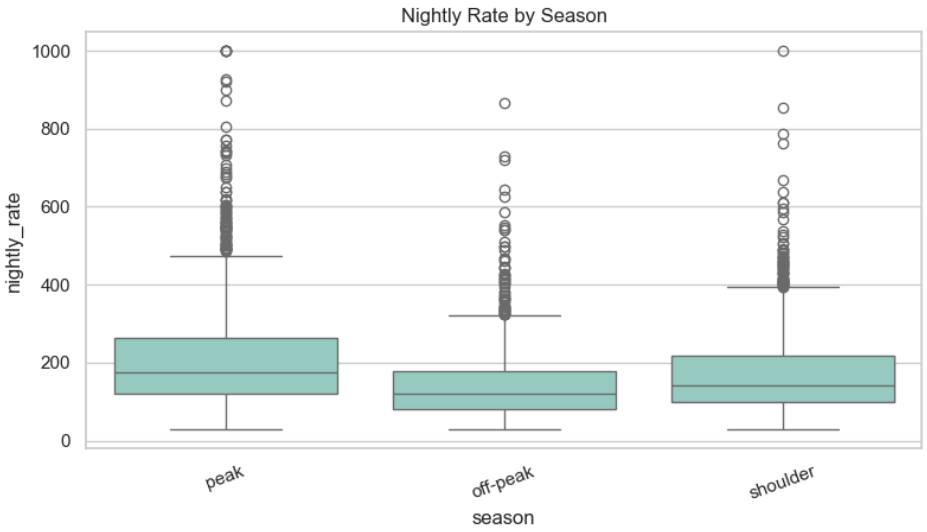
USC Viterbi School of Engineering

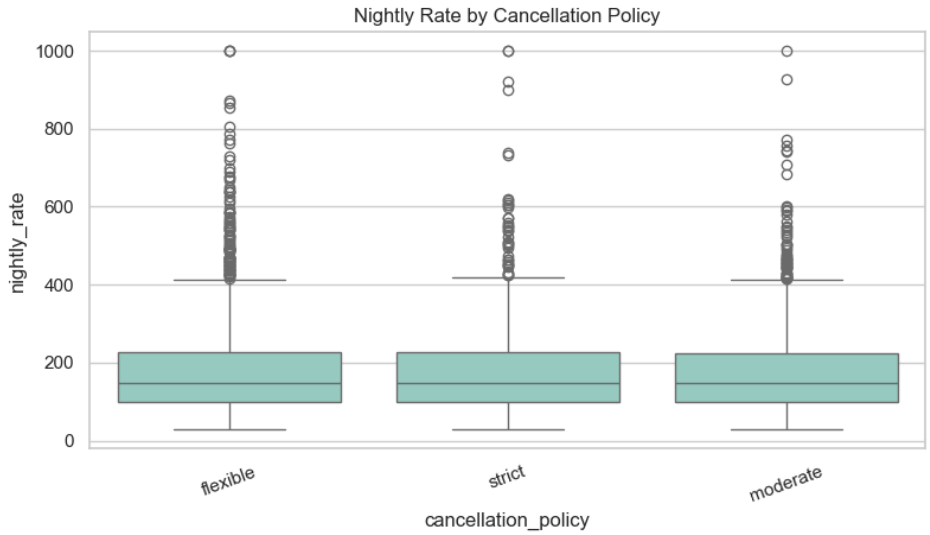
# BIVARIATE ANALYSIS

## BOXPLOTS – NIGHTLY RATE BY CATEGORICAL ATTRIBUTES



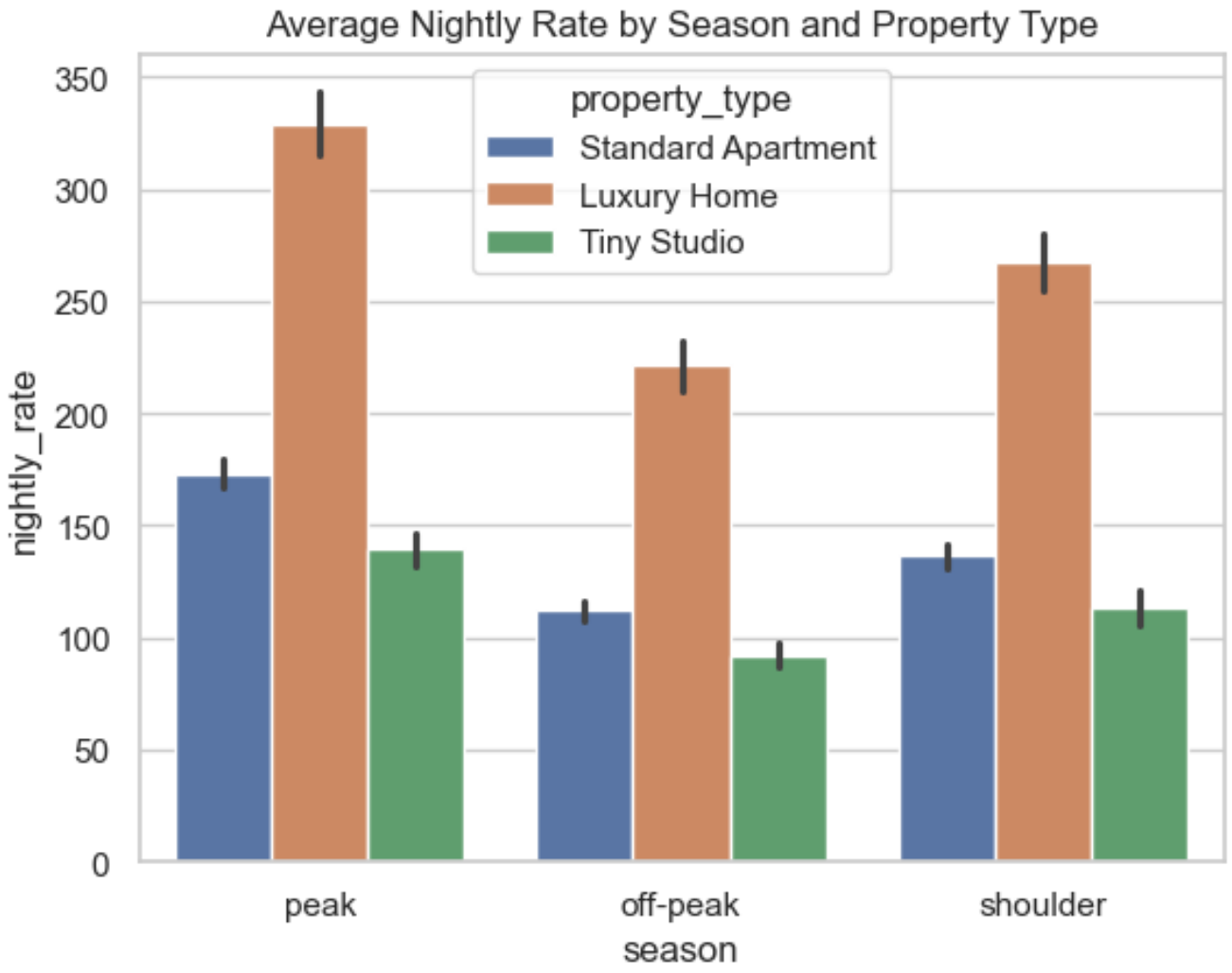




Nightly rates vary across categorical features. *Luxury Homes* and listings in *peak seasons* tend to have higher price ranges. *Cancellation policy* and *host response time* show minimal influence, while *property type* shows the most distinct price differentiation

# BIVARIATE ANALYSIS

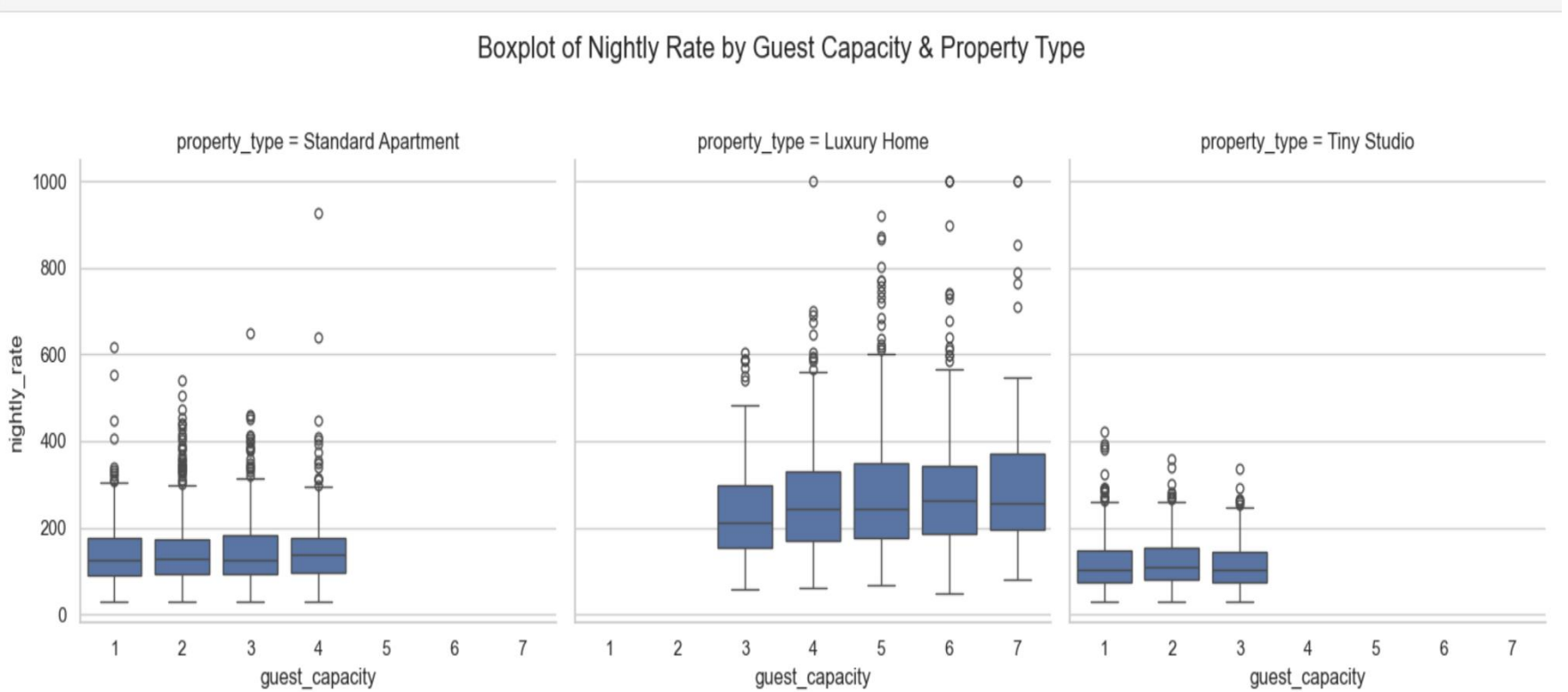
## AVERAGE NIGHLTY RATE BY SEASON & PROPERTY_TYPE

Seasonality impacts all listings, but Luxury Homes see the sharpest surge in peak periods, suggesting strong demand sensitivity.



# BIVARIATE ANALYSIS

## BOXPLOT OF NIGHTLY RATE BY GUEST CAPACITY & PROPERTY_TYPE

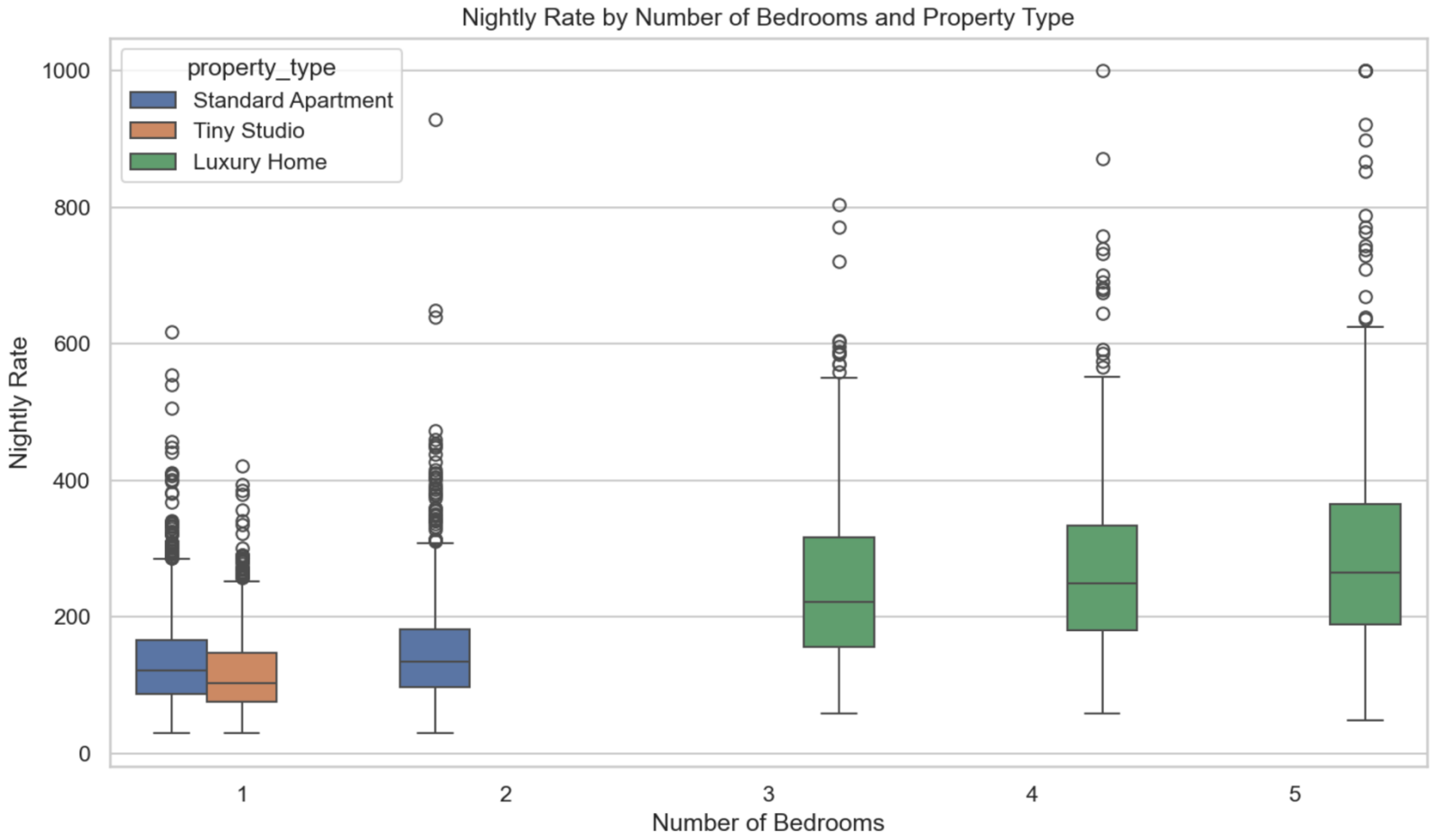



Nightly rate increases with guest capacity in Luxury Homes. Standard Apartments and Tiny Studios show flat trends across capacities

USC Viterbi School of Engineering

# BIVARIATE ANALYSIS

## BOXPLOT OF NIGHLY RATE BY BEDROOMS AND PROPERTY

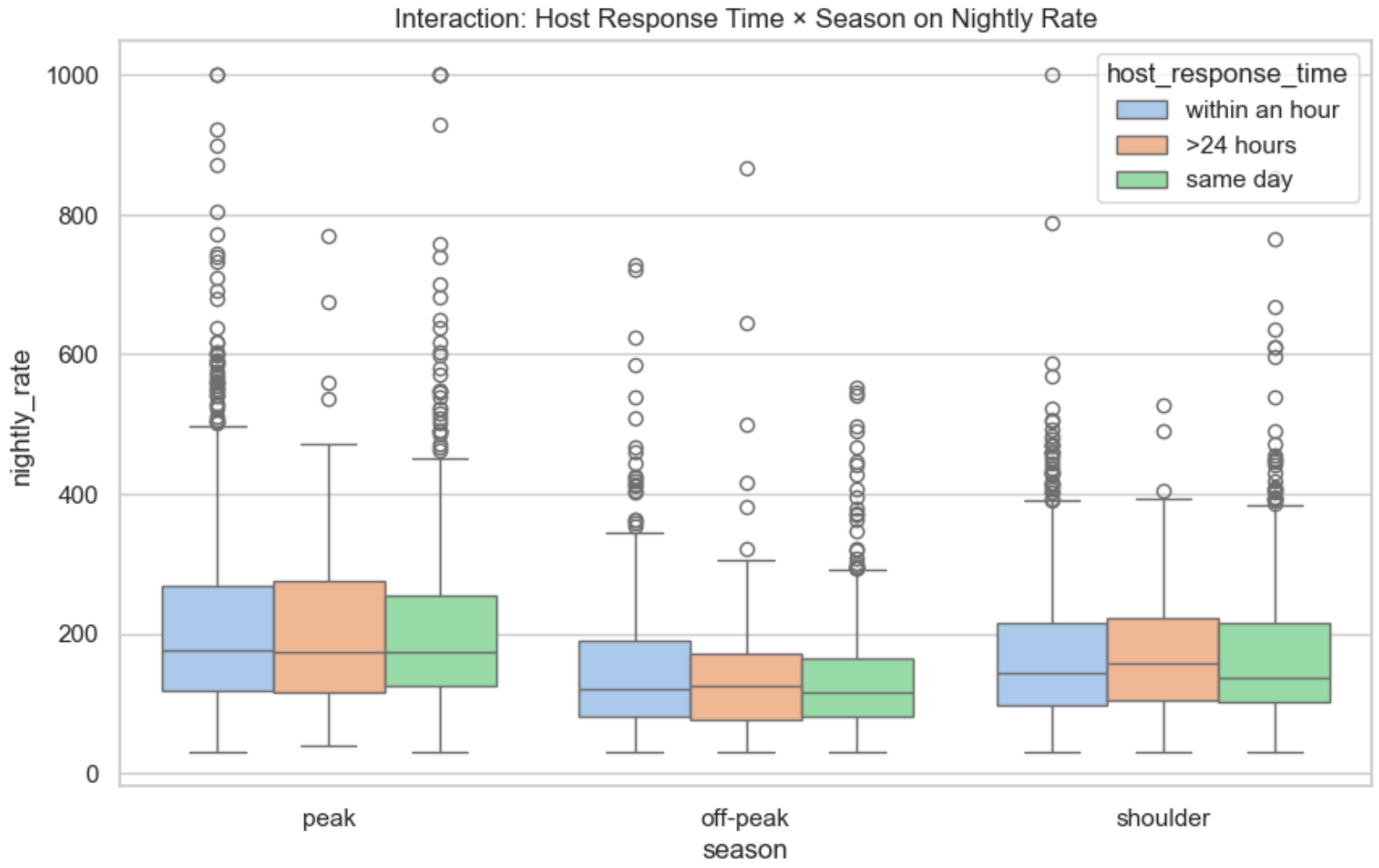



Luxury Homes with more bedrooms are priced higher, showing that space and amenities justify premium pricing

USC Viterbi School of Engineering

# BIVARIATE ANALYSIS

## BOXPLOT OF HOST RESPONSE TIME X SEASON

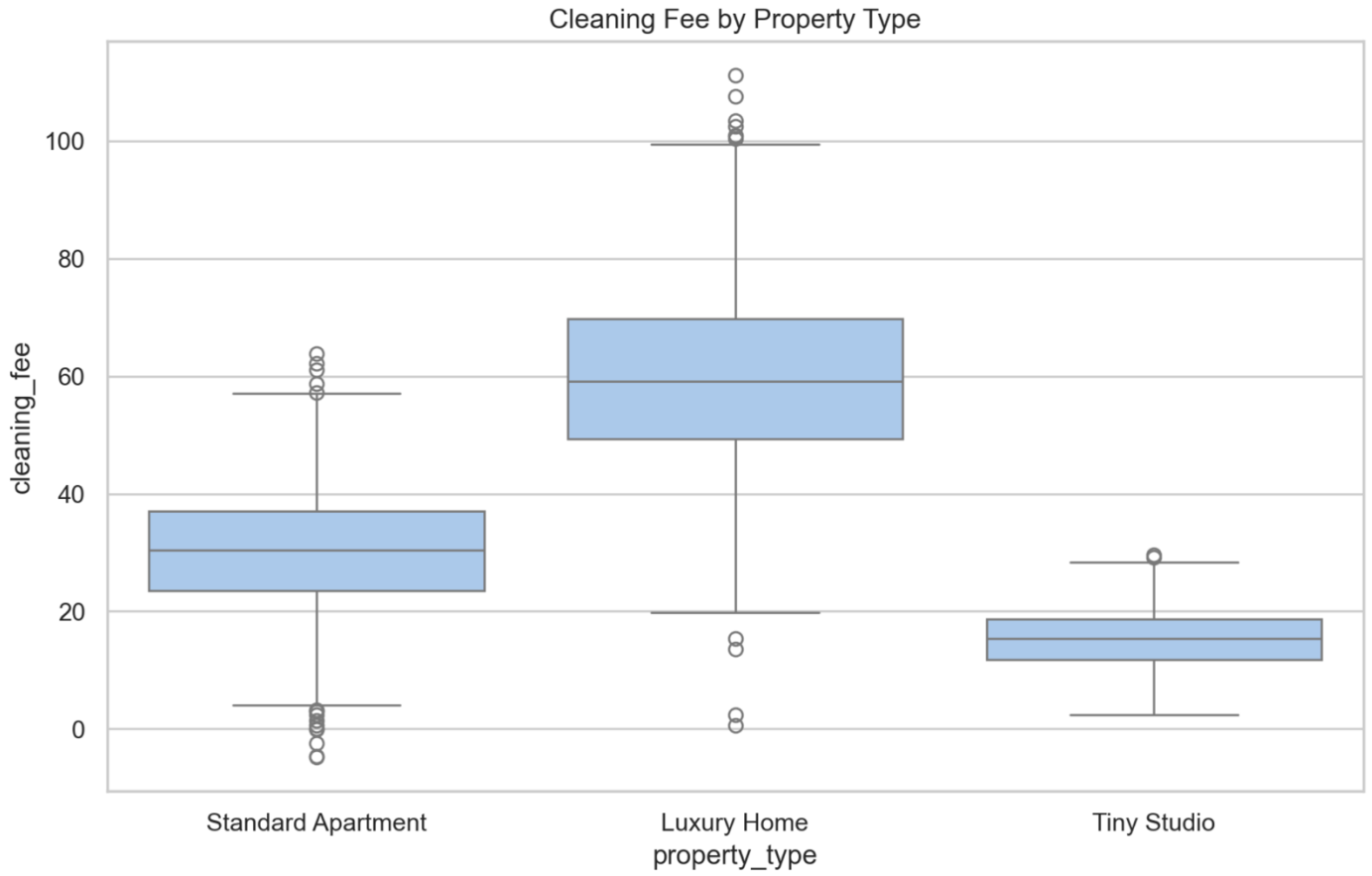



Median nightly rates differ across seasons and response times. Listings with quicker responses (e.g., “within an hour”) tend to show higher rates during peak seasons, with subtle shifts across off-peak and shoulder seasons. The spread within each group also suggests variation in pricing even within the same category

USC Viterbi School of Engineering

# BIVARIATE ANALYSIS

## BOXPLOT OF PROPERTY TYPE X CLEANING FEE

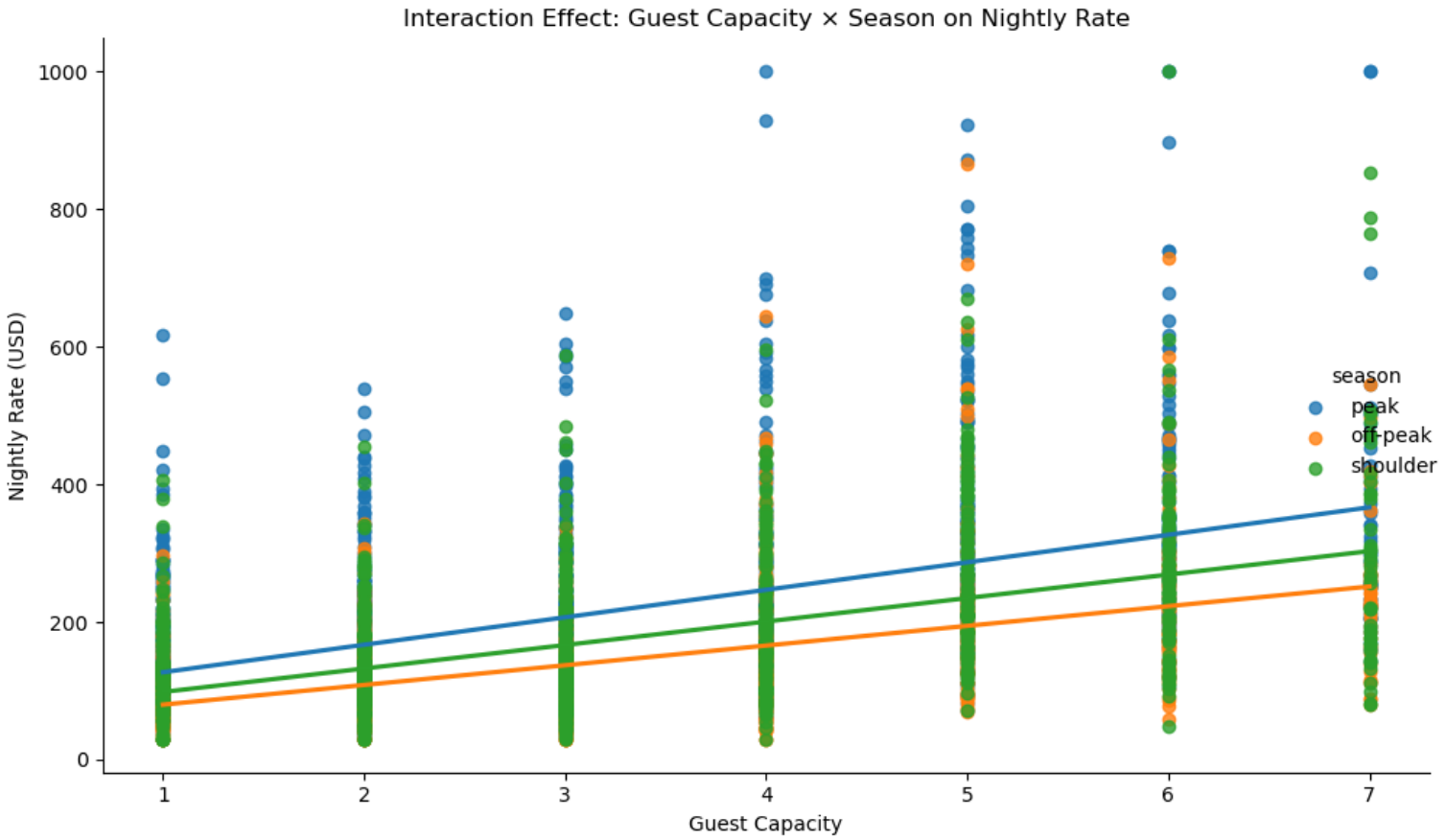



Luxury Homes generally have the highest cleaning fees, followed by Standard Apartments. Tiny Studios show the lowest fees with relatively little variation. The broader range for Luxury Homes indicates greater variability in how cleaning fees are priced within that category

USC Viterbi
School of Engineering

# BIVARIATE ANALYSIS

## SCATTERPLOT OF GUEST CAPACITY X SEASON




Nightly rates increase with guest capacity across all seasons, with peak season listings priced highest. The trend lines show a consistent upward pattern, confirming a positive interaction between guest capacity and season on pricing

# STATISTICAL ANALYSIS

# STATISTICAL ANALYSIS

## HYPOTHESIS

- 1. Does property type affect nightly rate?
  - $H_0$: Property type has no effect on nightly rate
  - $H_1$: Property type has a statistically significant effect on nightly rate.

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| C(property_type) | 1.704763e+07 | 2.0 | 822.623171 | 9.937069e-300 |
| Residual | 4.122941e+07 | 3979.0 | NaN | NaN |

  - Result : From the ANOVA table (PR(>F) = 9.937069e-300 < 0.05), we reject the null hypothesis.
  - **Conclusion**:
    We reject the null hypothesis and conclude that property type has a statistically significant effect on nightly rate

# STATISTICAL ANALYSIS

## HYPOTHESIS

- 2. Does guest capacity × season affect nightly rate?
  - $H_0$: There is no interaction effect between guest_capacity and property_type on nightly_rate
  - $H_1$: There is a statistically significant interaction effect between guest_capacity and property_type on nightly_rate

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| **C(property_type)** | 5.070438e+06 | 2.0 | 246.570446 | 1.132377e-101 |
| **guest_capacity** | 1.797526e+05 | 1.0 | 17.482384 | 2.962286e-05 |
| **C(property_type):guest_capacity** | 1.687177e+05 | 2.0 | 8.204578 | 2.780542e-04 |
| **Residual** | 4.088094e+07 | 3976.0 | NaN | NaN |

  - Result : There **is a statistically significant interaction effect** between guest_capacity and property_type on nightly_rate

# STATISTICAL ANALYSIS

## HYPOTHESIS

- 3. Does number of bedrooms × property type affect nightly rate?
  - $H_0$: The effect of number of bedrooms on nightly rate is the same across property types.
  - $H_1$: The effect of number of bedrooms on nightly rate depends on property type

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | nightly_rate | R-squared: | 0.305 |
| Model: | OLS | Adj. R-squared: | 0.304 |
| Method: | Least Squares | F-statistic: | 436.1 |
| Date: | Fri, 02 May 2025 | Prob (F-statistic): | 4.80e-312 |
| Time: | 17:03:00 | Log-Likelihood: | -24022. |
| No. Observations: | 3982 | AIC: | 4.805e+04 |
| Df Residuals: | 3977 | BIC: | 4.809e+04 |
| Df Model: | 4 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 165.4605 | 14.520 | 11.396 | 0.000 | 136.994 | 193.927 |
| C(property_type)[T.Standard Apartment] | -41.4693 | 16.196 | -2.560 | 0.010 | -73.222 | -9.716 |
| C(property_type)[T.Tiny Studio] | -37.5455 | 5.834 | -6.436 | 0.000 | -48.983 | -26.108 |
| number_of_bedrooms | 27.7193 | 3.527 | 7.858 | 0.000 | 20.803 | 34.635 |
| number_of_bedrooms:C(property_type)[T.Standard Apartment] | -14.0598 | 5.739 | -2.450 | 0.014 | -25.312 | -2.808 |
| number_of_bedrooms:C(property_type)[T.Tiny Studio] | -37.5455 | 5.834 | -6.436 | 0.000 | -48.983 | -26.108 |

| | | | |
|---|---|---|---|
| Omnibus: | 1744.818 | Durbin-Watson: | 1.970 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 12853.684 |
| Skew: | 1.930 | Prob(JB): | 0.00 |
| Kurtosis: | 10.910 | Cond. No. | 2.84e+15 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.89e-27. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Result: The interaction term with Standard Apartment is significant (p = 0.014), so there is some interaction.
Model $R^2$ = 0.305 suggests this model explains around 30.5% of the variance. We reject the null hypothesis and conclude that the effect of bedroom count on nightly rate significantly varies across property types.

# STATISTICAL ANALYSIS

## HYPOTHESIS

- 4. Does guest capacity affect nightly rate?
  - $H_0$: Guest capacity has no effect on nightly rate.
  - $H_1$: Guest capacity has a statistically significant effect on nightly rate.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | nightly_rate | R-squared: | 0.209 |
| Model: | OLS | Adj. R-squared: | 0.208 |
| Method: | Least Squares | F-statistic: | 1049. |
| Date: | Fri, 02 May 2025 | Prob (F-statistic): | 1.75e-204 |
| Time: | 17:03:22 | Log-Likelihood: | -24280. |
| No. Observations: | 3982 | AIC: | 4.856e+04 |
| Df Residuals: | 3980 | BIC: | 4.858e+04 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 70.6987 | 3.790 | 18.655 | 0.000 | 63.269 | 78.129 |
| guest_capacity | 34.4430 | 1.063 | 32.390 | 0.000 | 32.358 | 36.528 |

| | | | |
|---|---|---|---|
| Omnibus: | 1714.896 | Durbin-Watson: | 1.986 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 10857.854 |
| Skew: | 1.948 | Prob(JB): | 0.00 |
| Kurtosis: | 10.090 | Cond. No. | 8.42 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Result : The p-value for guest capacity is 0.000 and the coefficient is positive and large (coef = 34.44), so higher guest capacity increases nightly rate. We reject the null hypothesis and conclude that guest capacity has a statistically significant positive effect on nightly rate.

# STATISTICAL ANALYSIS

## HYPOTHESIS

- 5. Does season affect nightly rate?
  - $H_0$: Season has no effect on nightly rate.
  - $H_1$: Season has a statistically significant effect on nightly rate

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | nightly_rate | R-squared: | 0.209 |
| Model: | OLS | Adj. R-squared: | 0.208 |
| Method: | Least Squares | F-statistic: | 1049. |
| Date: | Fri, 02 May 2025 | Prob (F-statistic): | 1.75e-204 |
| Time: | 17:03:22 | Log-Likelihood: | -24280. |
| No. Observations: | 3982 | AIC: | 4.856e+04 |
| Df Residuals: | 3980 | BIC: | 4.858e+04 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 70.6987 | 3.790 | 18.655 | 0.000 | 63.269 | 78.129 |
| guest_capacity | 34.4430 | 1.063 | 32.390 | 0.000 | 32.358 | 36.528 |

| | | | |
|---|---|---|---|
| Omnibus: | 1714.896 | Durbin-Watson: | 1.986 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 10857.854 |
| Skew: | 1.948 | Prob(JB): | 0.00 |
| Kurtosis: | 10.090 | Cond. No. | 8.42 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Result: ANOVA p-value (PR(>F) = 8.002872e-52 < 0.05), so the effect of season is statistically significant. We reject the null hypothesis and conclude that nightly rate significantly varies by season

# STATISTICAL ANALYSIS

## HYPOTHESIS

- 6. Does property type × season affect nightly rate?
  - $H_0$: The effect of season on nightly rate is the same across property types.
  - $H_1$: The effect of season on nightly rate depends on property type

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | nightly_rate | R-squared: | 0.363 |
| Model: | OLS | Adj. R-squared: | 0.362 |
| Method: | Least Squares | F-statistic: | 282.8 |
| Date: | Fri, 02 May 2025 | Prob (F-statistic): | 0.00 |
| Time: | 17:04:05 | Log-Likelihood: | -23849. |
| No. Observations: | 3982 | AIC: | 4.772e+04 |
| Df Residuals: | 3973 | BIC: | 4.777e+04 |
| Df Model: | 8 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 221.3861 | 4.992 | 44.346 | 0.000 | 211.599 | 231.174 |
| C(property_type)[T.Standard Apartment] | -108.9963 | 6.408 | -17.009 | 0.000 | -121.560 | -96.433 |
| C(property_type)[T.Tiny Studio] | -129.2080 | 8.141 | -15.871 | 0.000 | -145.169 | -113.247 |
| C(season)[T.peak] | 107.8335 | 6.684 | 16.132 | 0.000 | 94.728 | 120.939 |
| C(season)[T.shoulder] | 46.3750 | 7.046 | 6.582 | 0.000 | 32.561 | 60.189 |
| C(property_type)[T.Standard Apartment]:C(season)[T.peak] | -46.9365 | 8.503 | -5.520 | 0.000 | -63.607 | -30.266 |
| C(property_type)[T.Tiny Studio]:C(season)[T.peak] | -60.2779 | 10.735 | -5.615 | 0.000 | -81.325 | -39.231 |
| C(property_type)[T.Standard Apartment]:C(season)[T.shoulder] | -22.2817 | 9.027 | -2.468 | 0.014 | -39.981 | -4.583 |
| C(property_type)[T.Tiny Studio]:C(season)[T.shoulder] | -25.2869 | 11.537 | -2.192 | 0.028 | -47.906 | -2.668 |

| | | | |
|---|---|---|---|
| Omnibus: | 1777.662 | Durbin-Watson: | 1.949 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 14591.990 |
| Skew: | 1.935 | Prob(JB): | 0.00 |
| Kurtosis: | 11.542 | Cond. No. | 16.1 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Result: Interaction terms (e.g., Standard Apartment × Peak, Tiny Studio × Shoulder) are all statistically significant (p-values < 0.05).
$R^2$ = 0.363 suggests the model explains ~36% of the variation. We reject the null hypothesis and conclude that the effect of season on nightly rate significantly differs across property types.

# STATISTICAL ANALYSIS
## HYPOTHESIS

- 7. Does host response time x season affect nighly price?
  - $H_0$: No interaction effect between host_response_time and season on nightly_rate
  - $H_1$: There is a statistically significant interaction between host_response_time and season on nightly_rate

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| **C(host_response_time)** | 2.779786e+04 | 2.0 | 1.006160 | 3.657134e-01 |
| **C(season)** | 3.347175e+06 | 2.0 | 121.152974 | 8.433731e-52 |
| **C(host_response_time):C(season)** | 2.040884e+04 | 4.0 | 0.369355 | 8.306165e-01 |
| **Residual** | 5.488238e+07 | 3973.0 | NaN | NaN |

Result: Only **season** significantly affects nightly_rate.. **Host response time** and its interaction with season are **not statistically significant**.

# STATISTICAL ANALYSIS

## HYPOTHESIS

- 8. Does Cleaning fee x property type affect nightly price?
  - $H_0$: There is no interaction effect between cleaning_fee and property_type on nightly_rate.
  - $H_1$: There is a statistically significant interaction between cleaning_fee and property_type on nightly_rate

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| C(property_type) | 5.998818e+06 | 2.0 | 289.317445 | 4.939699e-118 |
| cleaning_fee | 4.265689e+03 | 1.0 | 0.411461 | 5.212661e-01 |
| cleaning_fee:C(property_type) | 5.198837e+03 | 2.0 | 0.250735 | 7.782408e-01 |
| Residual | 4.121995e+07 | 3976.0 | NaN | NaN |

Result: Property type significantly affects nightly rate. Cleaning fee alone and its interaction with property type do not significantly affect nightly rate. This means while property_type is important, there's no statistical evidence that the effect of cleaning_fee varies across property types.

# STATISTICAL SIGNIFICANCE VS. PRACTICAL SIGNIFICANCE

## 1. Cleaning Fee × Property Type

- **Practical Significance** (EDA): Boxplots show visible variation in cleaning fees across property types — Luxury Homes have the highest and most variable fees.
- **Statistical Significance**: ANOVA results show **no statistically significant interaction** effect ($p > 0.05$), meaning observed differences may not be consistent across the population.

## 2. Guest Capacity × Season

- **Practical Significance** (EDA): Boxplots show visible variation in cleaning fees across property types — Luxury Homes have the highest and most variable fees.
- **Statistical Significance**: ANOVA results show **no statistically significant interaction** effect ($p > 0.05$), meaning observed differences may not be consistent across the population.

# STATISTICAL SIGNIFICANCE VS. PRACTICAL SIGNIFICANCE

## 3. Host Response Time × Season

- Practical Significance (EDA): Boxplots suggest listings with faster responses tend to charge more in peak season.
- Statistical Significance: The interaction effect is not statistically significant ($p > 0.05$), so differences may not hold consistently across the dataset.

## 4. Property Type × Season

- Practical Significance (EDA): Bar plots show Luxury Homes increase sharply in peak season, indicating strong demand sensitivity.
- Statistical Significance: OLS results confirm a statistically significant interaction ($p < 0.05$), validating that seasonal effects differ by property type.

# SUMMARY AND NEXT STEPS

- Listing_id (unique identifier) and nightly_rate (target variable) were excluded from the feature set. The target variable will be log-transformed due to skewness. Additionally, years_as_host will be capped at 17 based on domain knowledge, and cleaning_fee will be carefully evaluated due to its skewed distribution.
- Based on both practical and statistical significance, **property type**, **season**, and **number of bedrooms** are strong predictors of nightly rate. Some interaction effects (e.g., guest capacity × season) showed practical differences but lacked statistical support.
- The **correlation matrix** highlighted strong relationships between number of bedrooms and guest capacity.
- **Model Building**:
  - Apply regression or tree-based models using selected predictors.
  - Log-transform the nightly_rate to address skewness.
  - Encode categorical variables and include interaction terms with proven statistical significance.
- **Model Interpretation**:
  - Use SHAP values and other global interpretation techniques to explain model predictions.
  - Compare model performance with and without interaction terms.

# PREDICTIVE ANALYTICS FOR INFERENCE

# LINEAR MODELING

## OBJECTIVES

- Judging model fit, assumptions, and potential multicollinearity
- Understanding which predictors matter and estimating how much they matter
- Interpreting coefficients, interactions, and confidence intervals

USC Viterbi School of Engineering

# LINEAR MODELING

## INITIAL MODEL – BASELINE

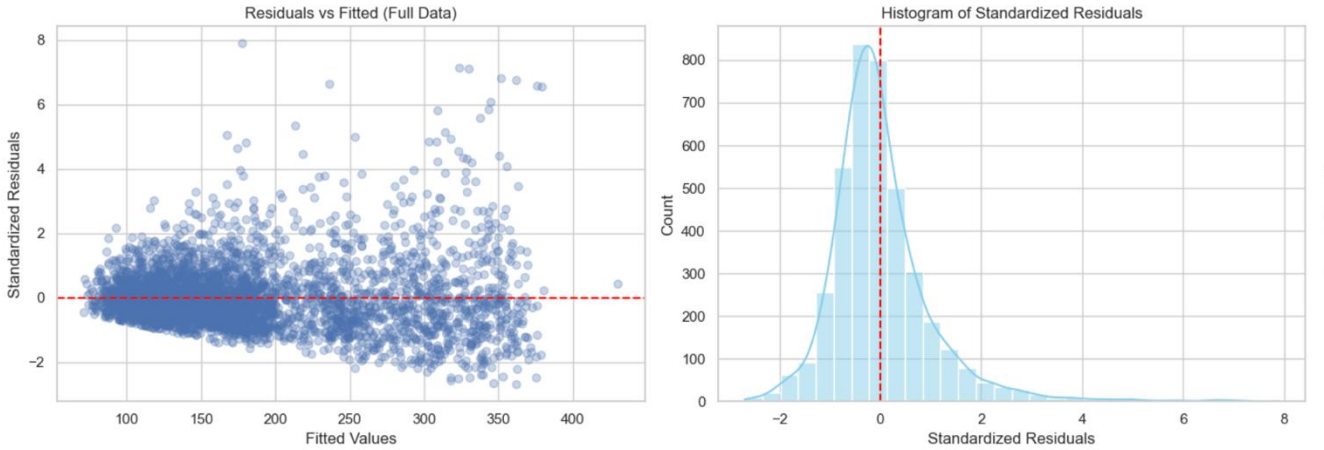
OLS Regression Results

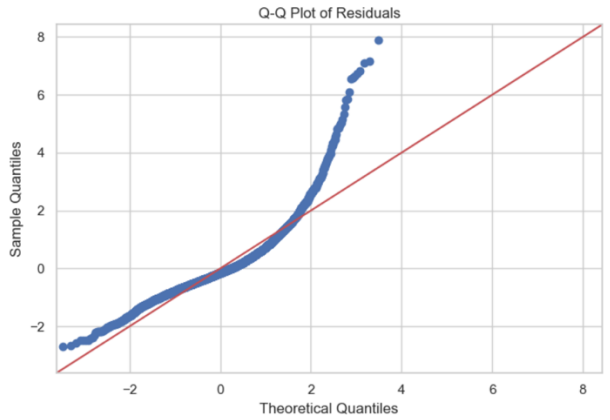
| | | | |
|---|---|---|---|
| Dep. Variable: | nightly_rate | R-squared: | 0.381 |
| Model: | OLS | Adj. R-squared: | 0.373 |
| Method: | Least Squares | F-statistic: | 47.40 |
| Date: | Fri, 09 May 2025 | Prob (F-statistic): | 0.00 |
| Time: | 18:13:08 | Log-Likelihood: | -23792. |
| No. Observations: | 3982 | AIC: | 4.769e+04 |
| Df Residuals: | 3930 | BIC: | 4.801e+04 |
| Df Model: | 51 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | -18.3397 | 196.599 | -0.093 | 0.926 | -403.786 | 367.106 |
| C(property_type)[T.Standard Apartment] | -81.0563 | 102.176 | -0.793 | 0.428 | -281.379 | 119.267 |
| C(property_type)[T.Tiny Studio] | -41.1393 | 124.498 | -0.330 | 0.741 | -285.226 | 202.947 |
| C(season)[T.peak] | 93.1050 | 35.434 | 2.628 | 0.009 | 23.635 | 162.575 |
| C(season)[T.shoulder] | 12.7873 | 37.864 | 0.338 | 0.736 | -61.448 | 87.023 |
| C(host_response_time)[T.same day] | 15.1909 | 36.013 | 0.422 | 0.673 | -55.416 | 85.798 |
| C(host_response_time)[T.within an hour] | 10.4240 | 35.315 | 0.295 | 0.768 | -58.814 | 79.662 |
| C(cancellation_policy)[T.moderate] | 10.9295 | 8.451 | 1.293 | 0.196 | -5.638 | 27.497 |
| C(cancellation_policy)[T.strict] | 17.3142 | 9.136 | 1.895 | 0.058 | -0.598 | 35.227 |
| C(property_type)[T.Standard Apartment]:C(season)[T.peak] | -45.2101 | 12.622 | -3.582 | 0.000 | -69.956 | -20.464 |
| C(property_type)[T.Tiny Studio]:C(season)[T.peak] | -58.9963 | 15.512 | -3.803 | 0.000 | -89.409 | -28.584 |
| C(property_type)[T.Standard Apartment]:C(season)[T.shoulder] | -17.8777 | 13.354 | -1.339 | 0.181 | -44.058 | 8.303 |
| C(property_type)[T.Tiny Studio]:C(season)[T.shoulder] | -18.7920 | 16.397 | -1.146 | 0.252 | -50.939 | 13.355 |
| number_of_bedrooms | -11.6339 | 30.179 | -0.385 | 0.700 | -70.801 | 47.533 |
| guest_capacity | 34.6637 | 25.617 | 1.353 | 0.176 | -15.560 | 84.887 |
| C(property_type)[T.Standard Apartment]:guest_capacity | -12.0104 | 8.745 | -1.373 | 0.170 | -29.156 | 5.135 |
| C(property_type)[T.Tiny Studio]:guest_capacity | -17.2361 | 11.989 | -1.438 | 0.151 | -40.741 | 6.269 |
| guest_capacity:C(season)[T.peak] | -0.1425 | 3.691 | -0.039 | 0.969 | -7.380 | 7.095 |
| guest_capacity:C(season)[T.shoulder] | 1.7460 | 3.900 | 0.448 | 0.654 | -5.899 | 9.391 |
| guest_capacity:C(cancellation_policy)[T.moderate] | -2.4374 | 2.854 | -0.854 | 0.393 | -8.033 | 3.159 |
| guest_capacity:C(cancellation_policy)[T.strict] | -5.4977 | 3.147 | -1.747 | 0.081 | -11.668 | 0.673 |
| location_score | 1.3740 | 1.332 | 1.032 | 0.302 | -1.237 | 3.985 |
| C(season)[T.peak]:location_score | 0.3114 | 0.378 | 0.823 | 0.410 | -0.430 | 1.053 |
| C(season)[T.shoulder]:location_score | -0.2130 | 0.399 | -0.533 | 0.594 | -0.996 | 0.570 |
| location_score:C(host_response_time)[T.same day] | -0.1698 | 0.505 | -0.336 | 0.737 | -1.160 | 0.820 |
| location_score:C(host_response_time)[T.within an hour] | -0.0562 | 0.495 | -0.113 | 0.910 | -1.028 | 0.915 |
| review_score | 39.1143 | 35.779 | 1.093 | 0.274 | -31.034 | 109.262 |
| C(property_type)[T.Standard Apartment]:review_score | -7.9496 | 9.710 | -0.819 | 0.413 | -26.986 | 11.087 |
| C(property_type)[T.Tiny Studio]:review_score | -16.6554 | 14.848 | -1.122 | 0.262 | -45.767 | 12.456 |
| review_score:C(season)[T.peak] | -2.0586 | 7.440 | -0.277 | 0.782 | -16.645 | 12.528 |
| review_score:C(season)[T.shoulder] | 9.7374 | 7.870 | 1.237 | 0.216 | -5.692 | 25.166 |
| amenities_count | 1.2233 | 5.074 | 0.241 | 0.809 | -8.724 | 11.170 |
| amenities_count:C(property_type)[T.Standard Apartment] | 0.2208 | 1.483 | 0.149 | 0.882 | -2.688 | 3.129 |
| amenities_count:C(property_type)[T.Tiny Studio] | -2.0599 | 2.015 | -1.022 | 0.307 | -6.010 | 1.891 |
| minimum_stay_nights | 8.9270 | 16.375 | 0.545 | 0.586 | -23.176 | 41.030 |
| years_as_host | 0.0492 | 0.519 | 0.095 | 0.925 | -0.969 | 1.068 |
| log_cleaning_fee | -11.4687 | 47.621 | -0.241 | 0.810 | -104.833 | 81.895 |
| C(property_type)[T.Standard Apartment]:log_cleaning_fee | 26.6981 | 24.494 | 1.090 | 0.276 | -21.325 | 74.721 |
| C(property_type)[T.Tiny Studio]:log_cleaning_fee | 29.0947 | 30.415 | 0.957 | 0.339 | -30.535 | 88.725 |
| guest_capacity:location_score | -0.0080 | 0.170 | -0.047 | 0.962 | -0.342 | 0.326 |
| review_score:location_score | -0.1266 | 0.293 | -0.431 | 0.666 | -0.702 | 0.449 |
| number_of_bedrooms:log_cleaning_fee | 7.4792 | 7.675 | 0.975 | 0.330 | -7.568 | 22.526 |
| guest_capacity:log_cleaning_fee | -6.4574 | 4.670 | -1.383 | 0.167 | -15.614 | 2.699 |
| log_cleaning_fee:review_score | -5.6884 | 7.623 | -0.746 | 0.456 | -20.635 | 9.258 |
| review_score:amenities_count | -0.1962 | 1.232 | -0.159 | 0.873 | -2.611 | 2.219 |
| guest_capacity:number_of_bedrooms | -0.8846 | 2.256 | -0.392 | 0.695 | -5.307 | 3.538 |
| minimum_stay_nights:location_score | -0.2493 | 0.148 | -1.688 | 0.092 | -0.539 | 0.040 |
| log_cleaning_fee:minimum_stay_nights | 2.7778 | 3.526 | 0.788 | 0.431 | -4.136 | 9.692 |
| cleaning_fee:C(cancellation_policy)[flexible] | 0.4095 | 0.379 | 1.080 | 0.280 | -0.334 | 1.153 |
| cleaning_fee:C(cancellation_policy)[moderate] | 0.2246 | 0.406 | 0.553 | 0.580 | -0.572 | 1.021 |
| cleaning_fee:C(cancellation_policy)[strict] | 0.4437 | 0.409 | 1.086 | 0.278 | -0.358 | 1.245 |
| location_score:number_of_bedrooms | 0.0969 | 0.222 | 0.436 | 0.663 | -0.338 | 0.532 |

| | | | |
|---|---|---|---|
| Omnibus: | 1700.779 | Durbin-Watson: | 1.952 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 12787.920 |
| Skew: | 1.863 | Prob(JB): | 0.00 |
| Kurtosis: | 10.949 | Cond. No. | 7.50e+04 |

Skewness of Standardized Residuals: 1.871




Trying the Linear model with all features and possible interaction terms

Evaluation Metrics (Full Data, Dollar Scale):
$R^2$ Score: 0.381
Mean Absolute Error (MAE): $65.44
Root Mean Squared Error (RMSE): $95.19

## Model Diagnostics Summary (No Transformation):

- **Residuals vs Fitted Plot**:
  A funnel-shaped pattern is visible, indicating **heteroscedasticity**—variance of errors increases with fitted values. This violates the constant variance assumption of linear regression.
- **Histogram of Standardized Residuals**:
  While roughly bell-shaped, there's **noticeable skewness** and deviation from perfect normality, suggesting residuals are not ideally distributed. The skew statistic of the residuals is 1.871
- **Q-Q Plot**:
  Residuals deviate from the red diagonal line, especially at the tails, confirming **non-normality** and potential **outliers**.

# LINEAR MODELING

## PROCESS DONE AFTER INITIAL MODELLING

- **Feature Inclusion**: Started with all features and possible interaction terms, then iteratively removed statistically insignificant ones based on p-values.
- **Multicollinearity Check**: Dropped guest_capacity due to high correlation with number_of_bedrooms as identified in the correlation matrix.
- **Feature Confirmation**: Used **backward elimination** to finalize significant predictors.
- **Outlier Removal**: Excluded data points with standardized residuals > 3.
- **Target Transformation**: Log-transformed nightly_rate to address right skewness.
- **Final Feature Set**: Selected property_type, season, number_of_bedrooms, and location_score.
- Scaled all numerical variables to improve interpretability of coefficients

# LINEAR MODELING

## FINAL LINEAR MODEL

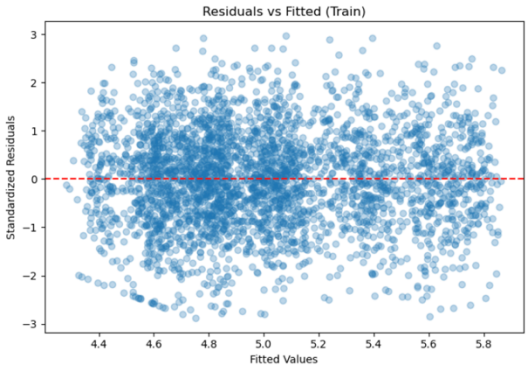
OLS Regression Results

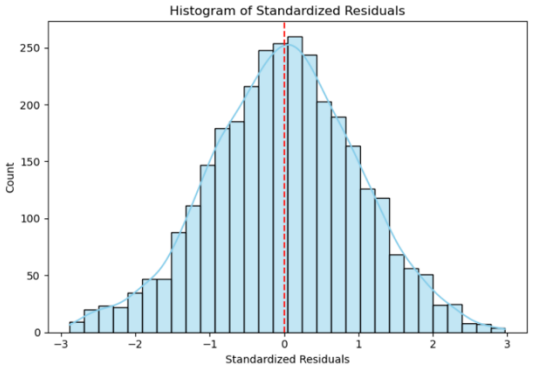
| | | | |
|---|---|---|---|
| Dep. Variable: | log_nightly_rate | R-squared: | 0.401 |
| Model: | OLS | Adj. R-squared: | 0.400 |
| Method: | Least Squares | F-statistic: | 353.5 |
| Date: | Tue, 06 May 2025 | Prob (F-statistic): | 0.00 |
| Time: | 22:06:02 | Log-Likelihood: | -2068.6 |
| No. Observations: | 3178 | AIC: | 4151. |
| Df Residuals: | 3171 | BIC: | 4194. |
| Df Model: | 6 | | |
| Covariance Type: | nonrobust | | |

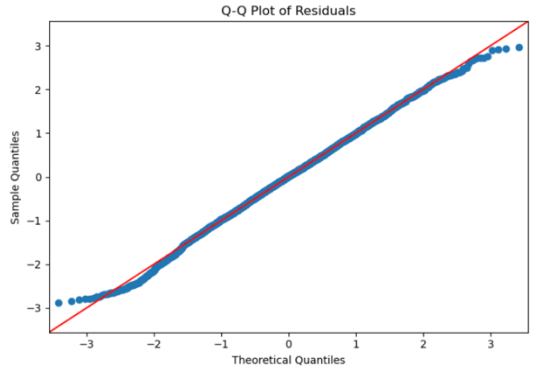
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 5.1103 | 0.033 | 156.989 | 0.000 | 5.047 | 5.174 |
| C(property_type)[T.Standard Apartment] | -0.4131 | 0.041 | -10.140 | 0.000 | -0.493 | -0.333 |
| C(property_type)[T.Tiny Studio] | -0.5912 | 0.049 | -11.977 | 0.000 | -0.688 | -0.494 |
| C(season)[T.peak] | 0.4038 | 0.020 | 20.231 | 0.000 | 0.365 | 0.443 |
| C(season)[T.shoulder] | 0.1749 | 0.021 | 8.185 | 0.000 | 0.133 | 0.217 |
| number_of_bedrooms | 0.1227 | 0.020 | 6.265 | 0.000 | 0.084 | 0.161 |
| location_score | 0.0461 | 0.008 | 5.601 | 0.000 | 0.030 | 0.062 |

| | | | |
|---|---|---|---|
| Omnibus: | 3.165 | Durbin-Watson: | 2.004 |
| Prob(Omnibus): | 0.205 | Jarque-Bera (JB): | 3.215 |
| Skew: | -0.072 | Prob(JB): | 0.200 |
| Kurtosis: | 2.943 | Cond. No. | 11.0 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.




Test Set Evaluation Metrics:
$R^2$ Score: 0.379
Mean Absolute Error (MAE): $62.89
Root Mean Squared Error (RMSE): $94.38

The model achieved an $R^2$ of 0.401 on the training set (log scale) and an $R^2$ of 0.379 on the test set (original price scale), indicating good fit and generalizability from the residual plot

# LINEAR MODELING

## MODEL DIAGNOSTICS OF FINAL LINEAR MODEL

- **Residuals vs Fitted Plot**: The residuals appear randomly scattered around zero with no clear pattern, indicating that the assumption of linearity and homoscedasticity (constant variance) is reasonably satisfied.
- **Histogram of Standardized Residuals**: The distribution of residuals closely resembles a normal bell curve, suggesting that the errors are approximately normally distributed.
- **Q-Q Plot**: Most points fall along the 45-degree reference line, with only minor deviations in the tails. This further supports that the residuals follow a near-normal distribution.
- Variance Inflation Factor (VIF) was computed to assess multicollinearity among the selected predictors. All VIF values were below 3, indicating no severe multicollinearity. This suggests that the predictors are sufficiently independent and stable for reliable coefficient estimation

| | Feature | VIF |
|---|---|---|
| 0 | C(property_type)[T.Standard Apartment] | 2.862030 |
| 1 | C(season)[T.peak] | 2.101974 |
| 2 | C(property_type)[T.Tiny Studio] | 2.073005 |
| 3 | number_of_bedrooms | 2.015174 |
| 4 | C(season)[T.shoulder] | 1.810669 |
| 5 | location_score | 1.001758 |

# LINEAR MODELING

## INFLUENTIAL FEATURES OF LINEAR MODEL

- Based on the log-linear regression analysis, the **most influential predictors** of Airbnb nightly pricing are **property type, season**, and **number of bedrooms**.
  - **Property type** has the strongest impact: listings categorized as **Tiny Studios** are priced approximately **$80 lower**, and **Standard Apartments** about **$61 lower** than **Luxury Homes**, the reference category. This shows that the type of accommodation significantly defines its market segment/
  - **Seasonality** also plays a major role. Listings in **Peak Season** are priced around **$89 higher** than those in the off-season, while **Shoulder Season** listings are **$34 higher**, reflecting demand-driven pricing adjustments aligned with travel patterns.
  - **Number of bedrooms**, though standardized, contributes meaningfully—each standard deviation increase in bedroom count raises the nightly rate by **approximately $24**, highlighting the importance of space in consumer willingness to pay.
  - **Location score**, while statistically significant, has a relatively modest impact (~$8 per SD increase), suggesting it may influence guest decision-making but not as strongly as the above factors.
- Overall, hosts and platforms should prioritize **property classification and seasonal pricing strategy** when estimating or optimizing listing prices.

# LINEAR MODELING

## COEFFICIENT INTERPRETATION SUMMARY

- Assuming an average nightly rate of $180.31, we reverse the log transformation using exp(β) - 1 to calculate percent change, then convert to dollar impact.
- **Luxury Homes** serve as the **reference category.** Their average predicted price is the **Intercept**, which is exp(5.1103) ≈ **$165.3**
- **When holding other variables constant:**
  - **Standard Apartments** are priced **~$61 lower** than entire homes/apartments.
  - **Tiny Studios** are priced **~$80 lower** than entire homes/apartments.
  - **Peak season** increases nightly rates by **~$89**, on average.
  - **Shoulder season** raises prices by **~$34**.
  - A **1 SD increase in number of bedrooms** is linked to a **~$24** increase.
  - A **1 SD increase in location score** results in a **~$8** rise.

# LINEAR MODELING

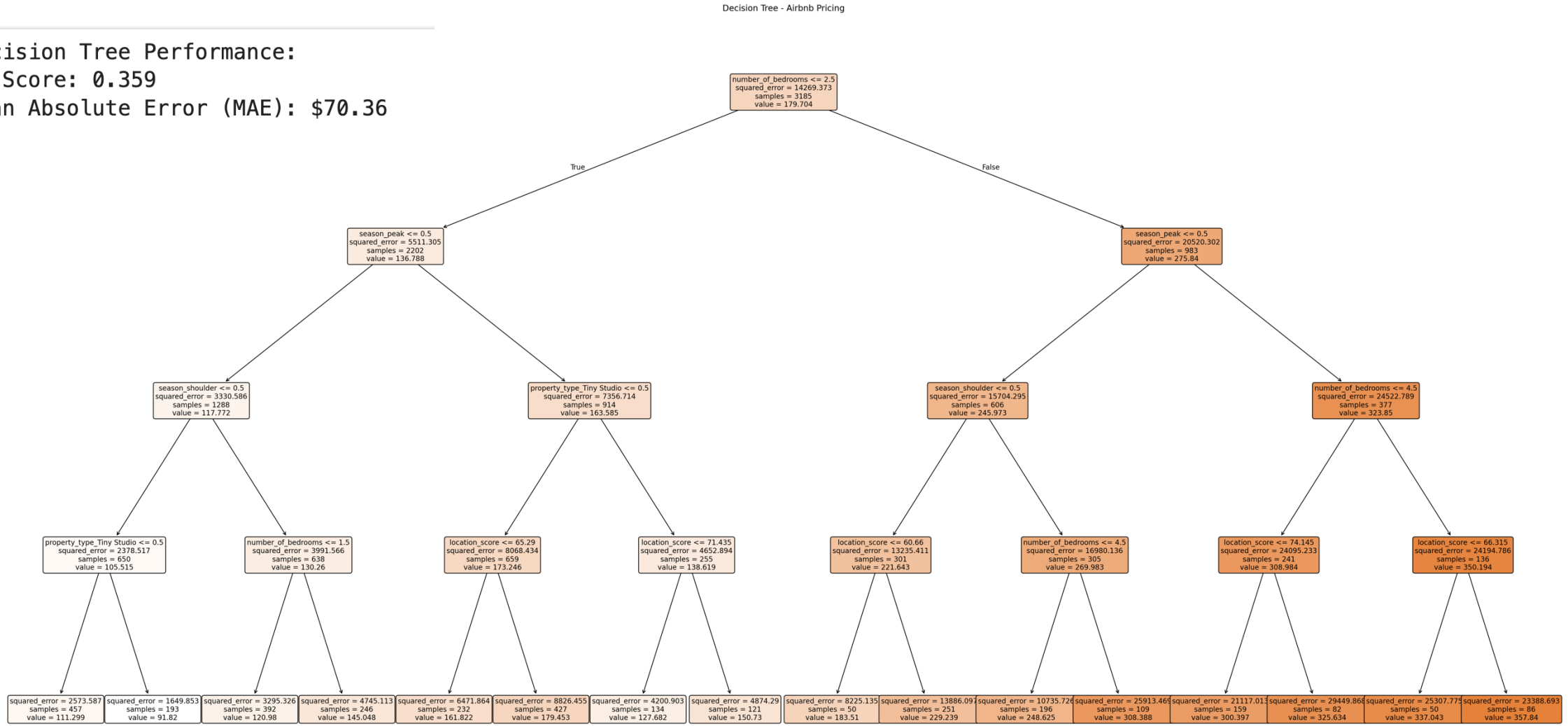
## INTERPRETING COEFFIENTS

| Predictor | Effect on Nightly Rate |
|---|---|
| Standard Apartment | –$61.31 |
| Tiny Studio | –$80.41 |
| Peak Season | +$89.60 |
| Shoulder Season | +$34.42 |
| Bedrooms (1 SD increase) | +$23.66 |
| Location Score (1 SD increase) | +$8.43 |

USC Viterbi
School of Engineering

# DECISION TREE MODELLING

Decision Tree Performance:
R² Score: 0.359
Mean Absolute Error (MAE): $70.36



# DECISION TREE MODELLING

## KEY DECISION RULES FROM THE TREE SPLITS

The tree uses a sequence of **if-else conditions** to segment listings into pricing categories:

1. **First Split – number of bedrooms <= 2.5**
   This is the most important feature in the tree. It separates listings with fewer bedrooms (typically smaller listings) from larger ones.
2. **Next Level (Left Branch)**
   - If it's **not peak season**, and not shoulder either, listings tend to be **low-priced** (value ≈ $117).
   - Then it checks **property type = Tiny Studio** and further splits by **number of bedrooms** again and **location_score** to assign prices.
3. **Next Level (Right Branch)**
   - If **season is peak** and **number of bedrooms > 2.5**, prices are higher (value ≈ $278–350).
   - It uses **location_score** and further refines pricing through more splits based on bedroom count and location score ranges.

# DECISION TREE MODELLING

## HOW THE MODEL SEGMENTS LISTINGS INTO PRICING TIERS

### Low Tier:

- Listings with ≤ 2.5 bedrooms
- Not in peak season
- Likely Tiny Studios or low location scores
- Price range: around **$90 – $145**

### Mid Tier:

- Listings in **shoulder season**, or mid-range location scores
- Price range: around **$150 – $250**

### High Tier:

- Listings with **> 2.5 bedrooms**
- During **peak season**
- High **location scores**
- Price range: up to **$350**

# DECISION TREE MODELLING

## PERFORMANCE COMPARISON WITH LINEAR MODEL

| Model Type | $R^2$ Score | MAE ($) |
|---|---|---|
| Linear Model | 0.379 | 62.89 |
| Decision Tree | 0.359 | 70.36 |

- The **linear model performs slightly better** in both $R^2$ and MAE.
- But the **decision tree is more interpretable** for identifying pricing rules and segments.
- The tree provides insight into **nonlinear interactions** (e.g., how season affects pricing differently for different property types and bedroom counts).

# PREDICTIVE MODELING DISCUSSION

Both the **linear model** and **decision tree** offer unique insights into Airbnb pricing dynamics.

- **Interpretability & Inference**:
  The linear model enables clear statistical inference with interpretable coefficients, making it suitable for quantifying the marginal effects of each predictor. It revealed that *property type*, *season*, *number of bedrooms*, and *location score* were significant drivers of nightly rates. It also helped translate these impacts into real-dollar estimates (e.g., peak season increases price by ~$89).

- **Nonlinear Relationships & Interactions**:
  The decision tree, though slightly lower in performance ($R^2$ = 0.359 vs. 0.379), effectively captured **nonlinearities** and **conditional interactions**. It uncovered pricing rules such as:
  - Higher prices during **peak season** combined with **>2.5 bedrooms** and high **location scores**.
  - Low-tier pricing for **tiny studios** in the **off-season** with fewer bedrooms.

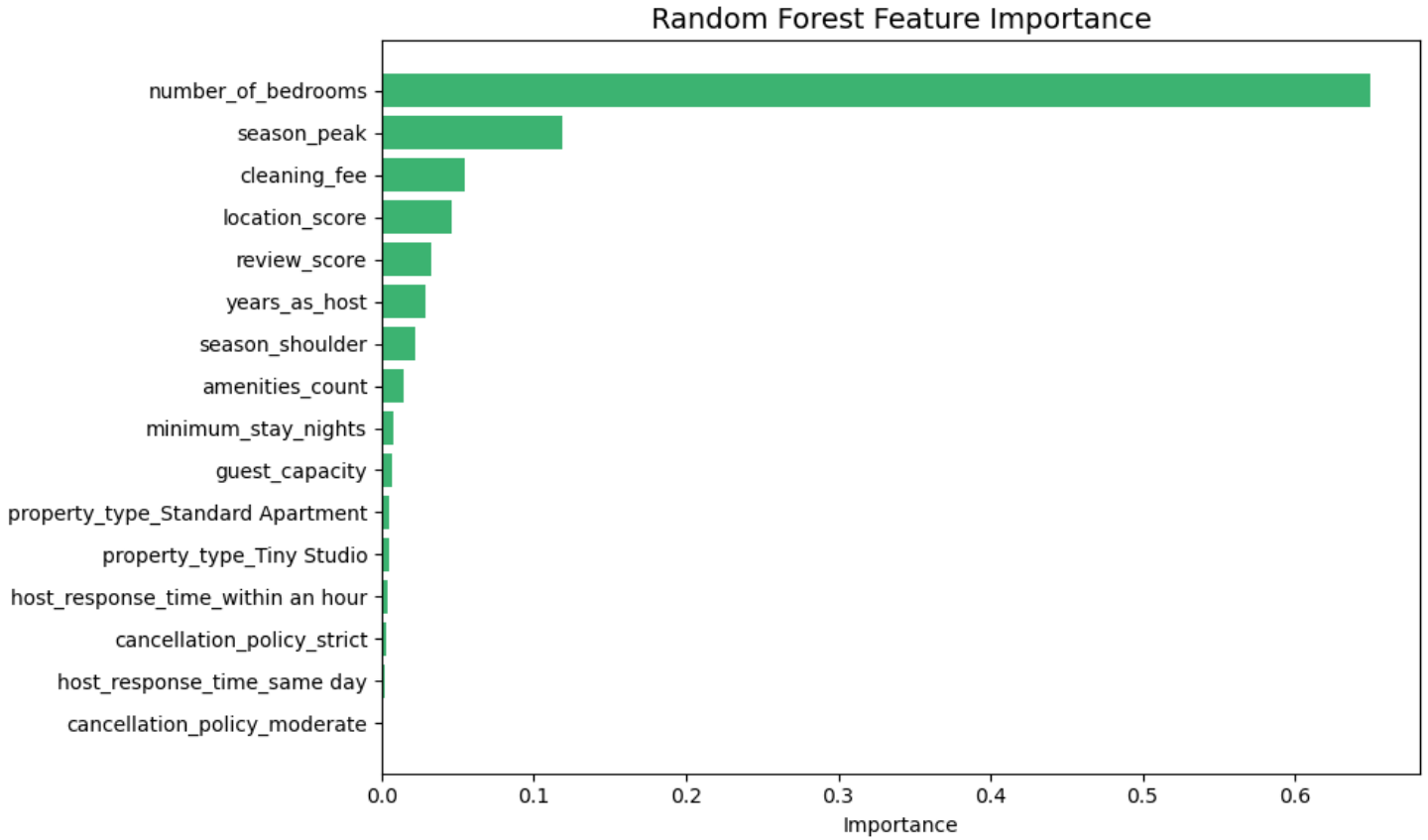
- **Complementary Value**:
  While the linear model supports hypothesis testing and effect estimation, the decision tree enhances understanding of **segment-based pricing rules** that are easier to act on. Together, they provide a fuller picture—**one explains how much** each factor contributes, the other **shows when and how** they interact to influence price

# MODEL INTERPRETATION AND EXPLAINABILTY TECHNIQUES

USC Viterbi School of Engineering

# RANDOM FOREST MODEL

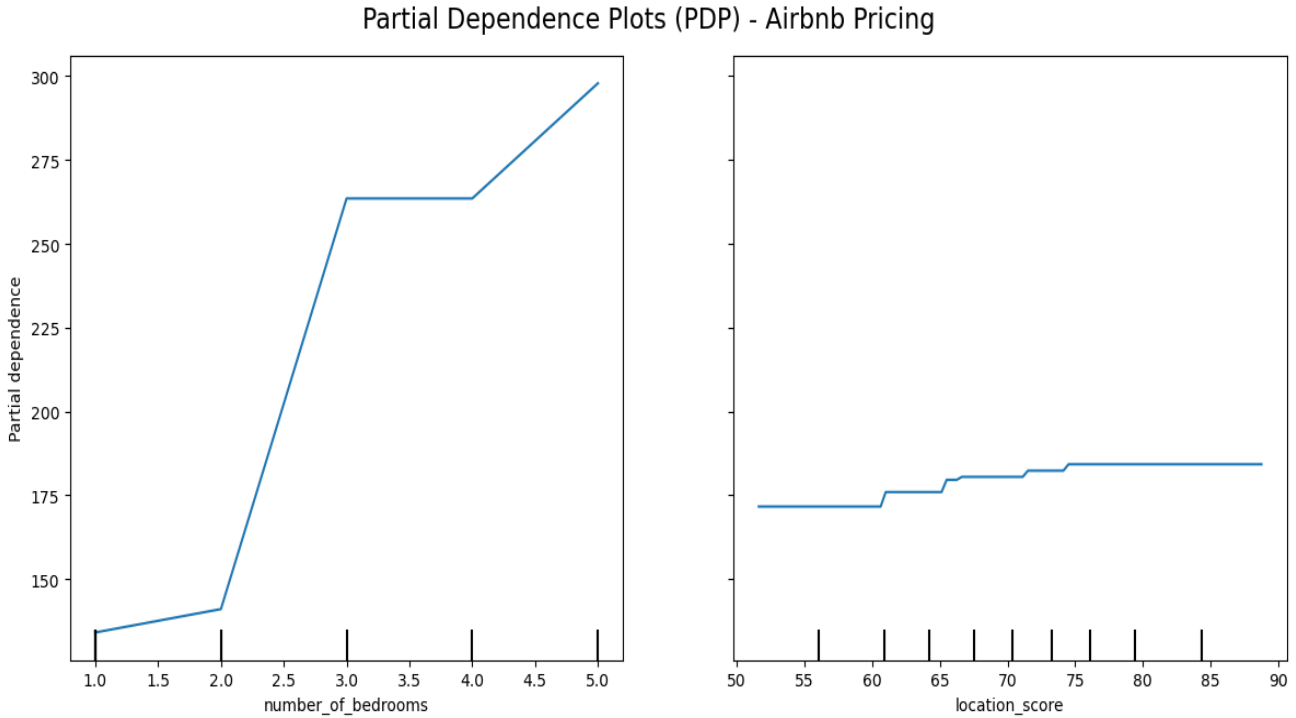
## MODEL FEATURE IMPORTANCE




**Random Forest Feature Importance:**
The Random Forest model identifies **number_of_bedrooms** as the most influential predictor of nightly rate, followed by **season_peak**, **cleaning_fee**, and **location_score**. These features contribute substantially more than others, indicating that room capacity, seasonal demand, and pricing elements like cleaning fee strongly drive price variation. Features like guest_capacity, host_response_time, and cancellation_policy show minimal impact, suggesting they are less critical in the ensemble's predictive performance.

# MODEL EXPLAINABILITY TECHNIQUES

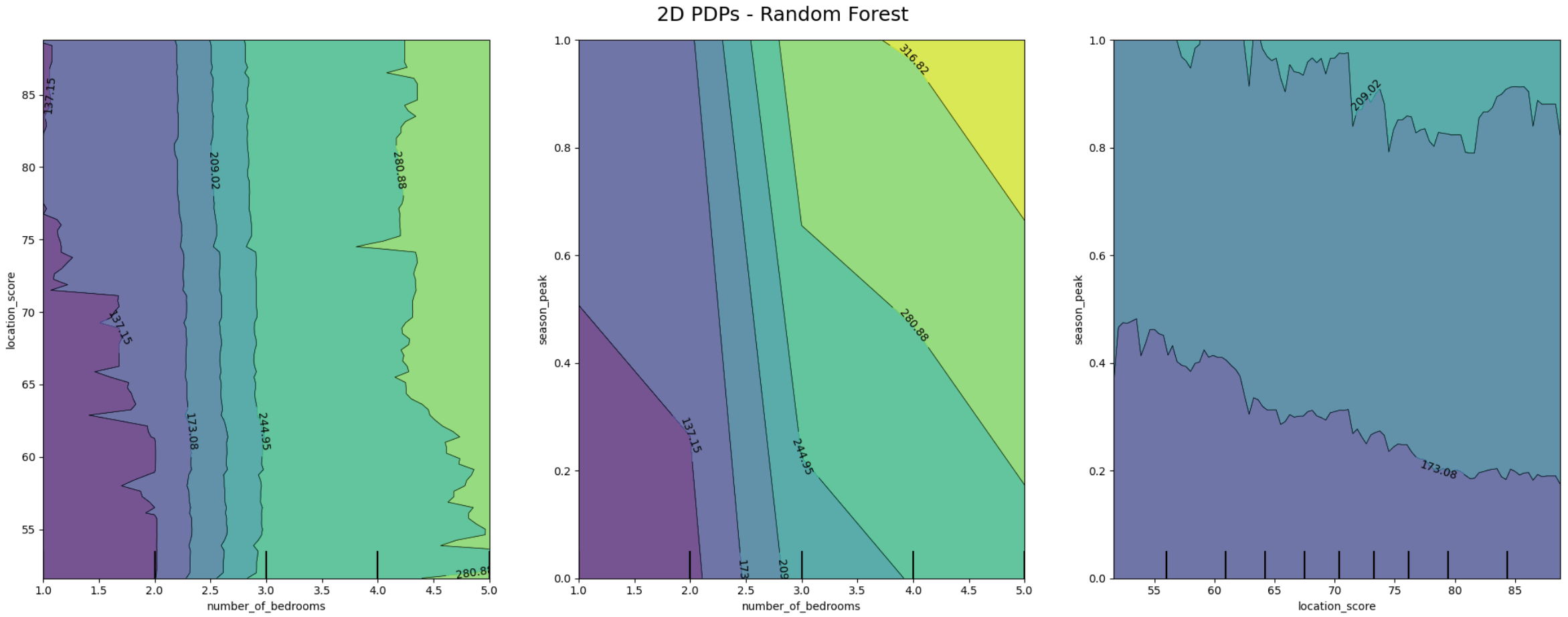
## 1D PDP PLOT – GLOBAL EXPLANATION TECHNIQUE




- **Number of Bedrooms:** Strong positive effect on price — especially sharp increase from 2 to 3 bedrooms, with prices rising steadily as bedrooms increase.
- **Location Score:** Slight positive impact — higher scores mildly increase predicted price, but the effect is limited

# MODEL EXPLAINABILITY TECHNIQUES

## 2D PDP PLOT



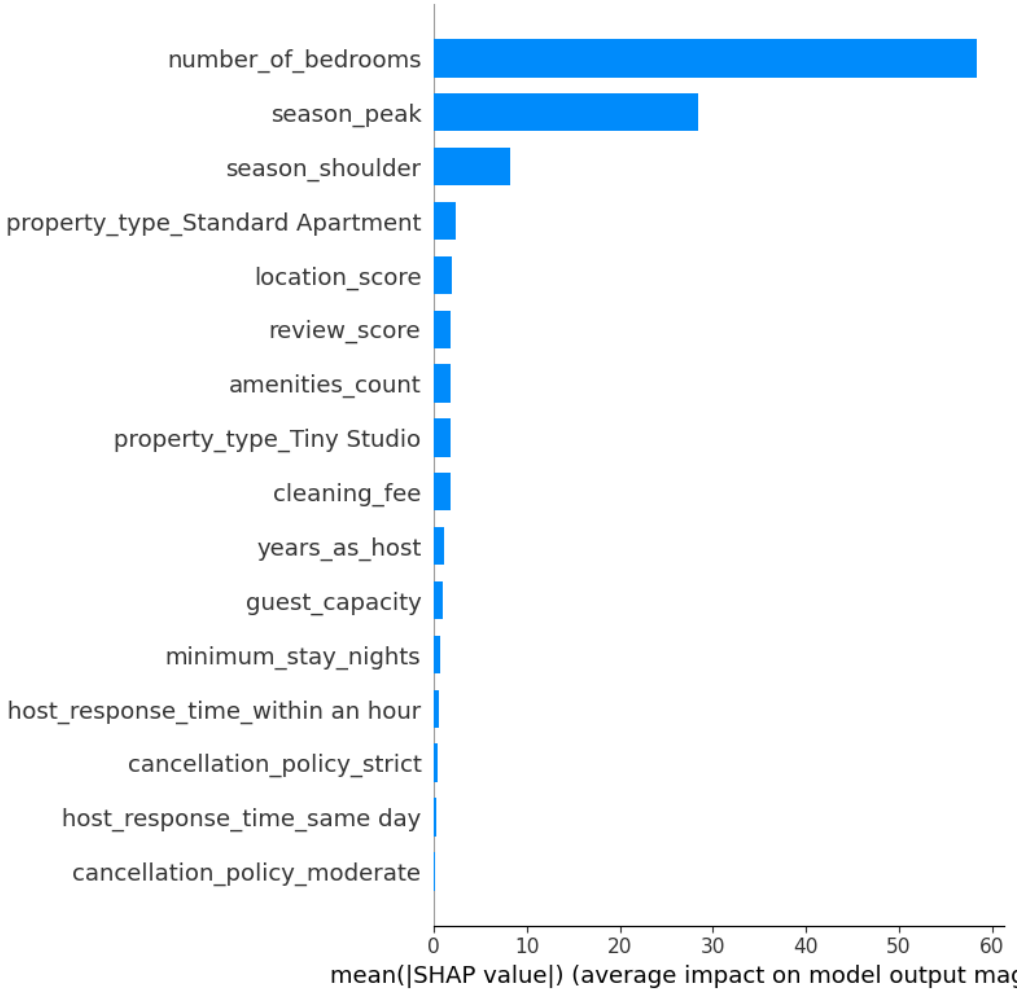
# MODEL EXPLAINABILITY TECHNIQUES

## INTERPRETATION OF 2D PDP PLOT

- **Number of Bedrooms × Location Score**
  → Nightly rate increases sharply with more bedrooms in high-scoring locations, showing a strong interaction between property size and desirability.
- **Number of Bedrooms × Season (Peak)**
  → Larger properties are priced significantly higher during peak season, indicating that demand for spacious listings rises with seasonal demand.
- **Location Score × Season (Peak)**
  → Peak season causes higher price surges in top-rated locations, suggesting seasonal demand amplifies location-based pricing effects.

# MODEL EXPLAINABILITY TECHNIQUES

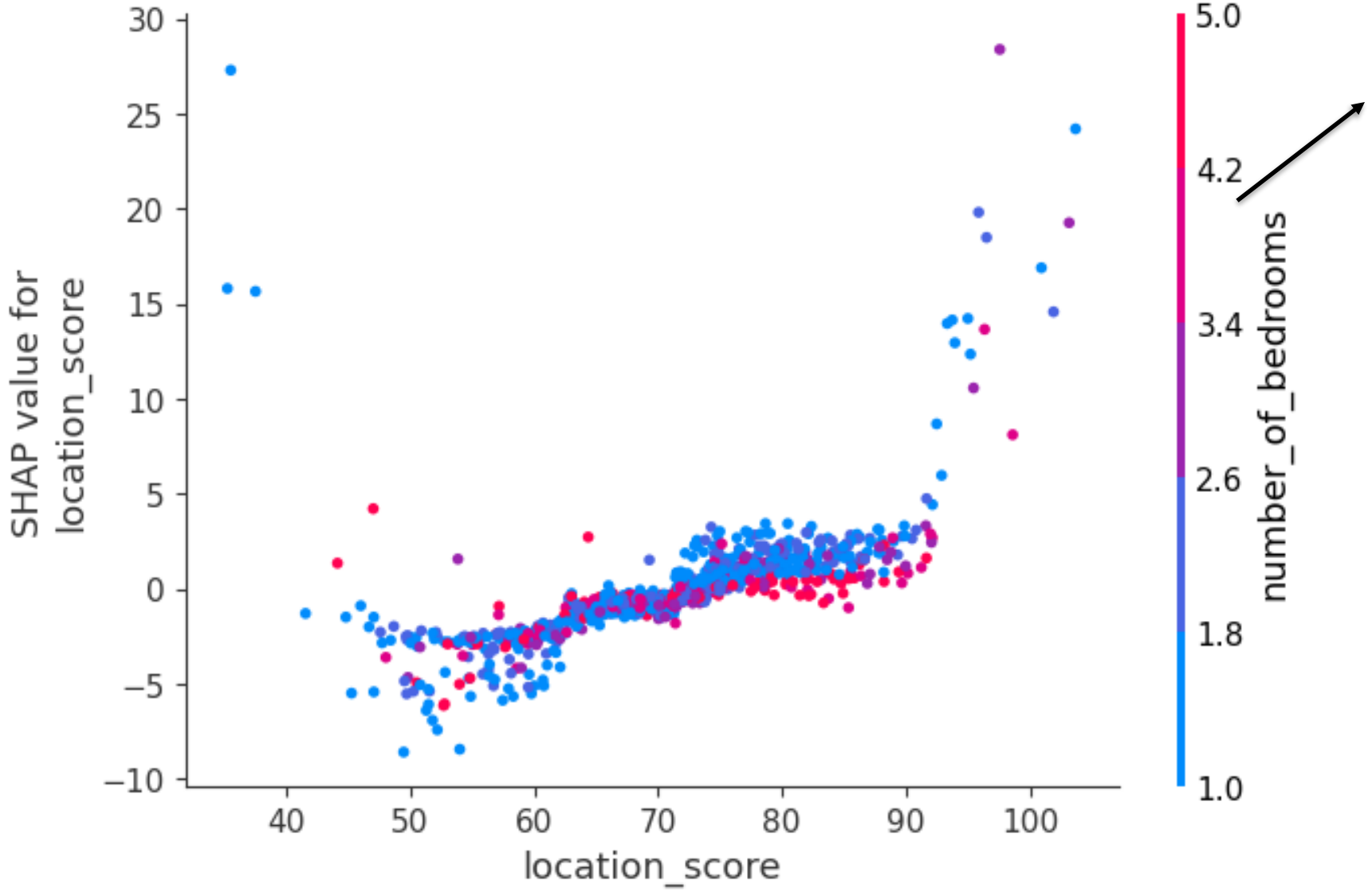
## SHAP SUMMARY PLOT




SHAP values confirm number_of_bedrooms and season_peak as the most influential features, with minor contributions from location, property type, and cleaning fee.

# MODEL EXPLAINABILITY TECHNIQUES

## SHAP DEPENDENCY PLOT

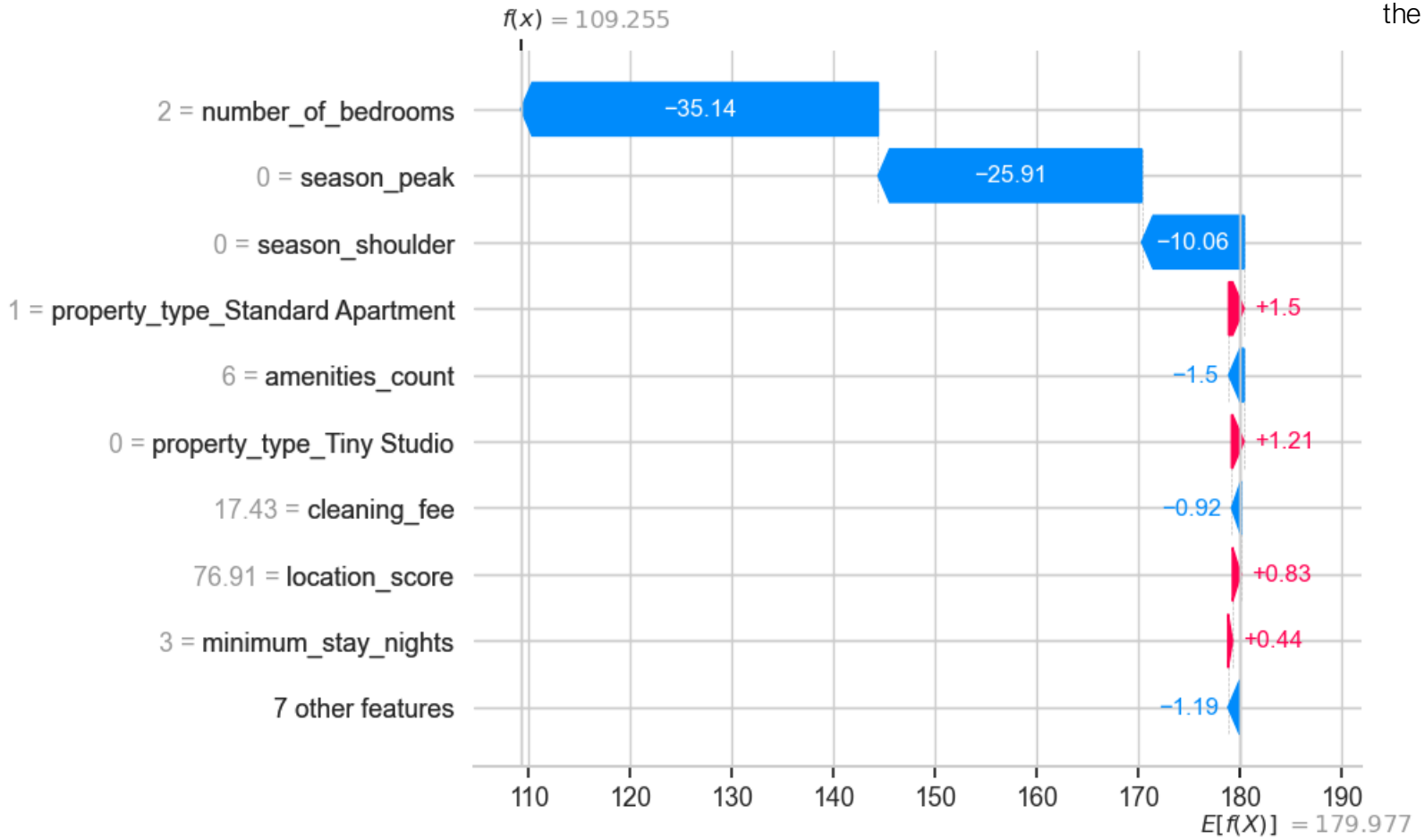



There no clear separation between the red and blue dots-signifies no clear interaction effect

# MODEL EXPLAINABILITY TECHNIQUES

## LOCAL EXPLAINATION TECHNIQUE - SHAP

### LOCAL SHAP WATERFALL PLOT




From the **local SHAP explanation**, you can conclude the following:

- The **individual prediction** for the selected Airbnb listing is primarily **driven downward** by the **number of bedrooms (2)** and the fact that it is **not in peak or shoulder season**.
- Despite having positive features like a **good location score**, **moderate cleaning fee**, and **reasonable amenities**, these were **not enough to raise the price substantially**.
- This illustrates that **bedroom count and seasonality** are **dominant factors** in determining price, consistent with global model insights.
- SHAP helps you see **how each feature contributed to the price for this specific listing**, supporting **personalized pricing recommendations** or **listing improvement advice**.

# MODEL EXPLAINABILITY TECHNIQUES

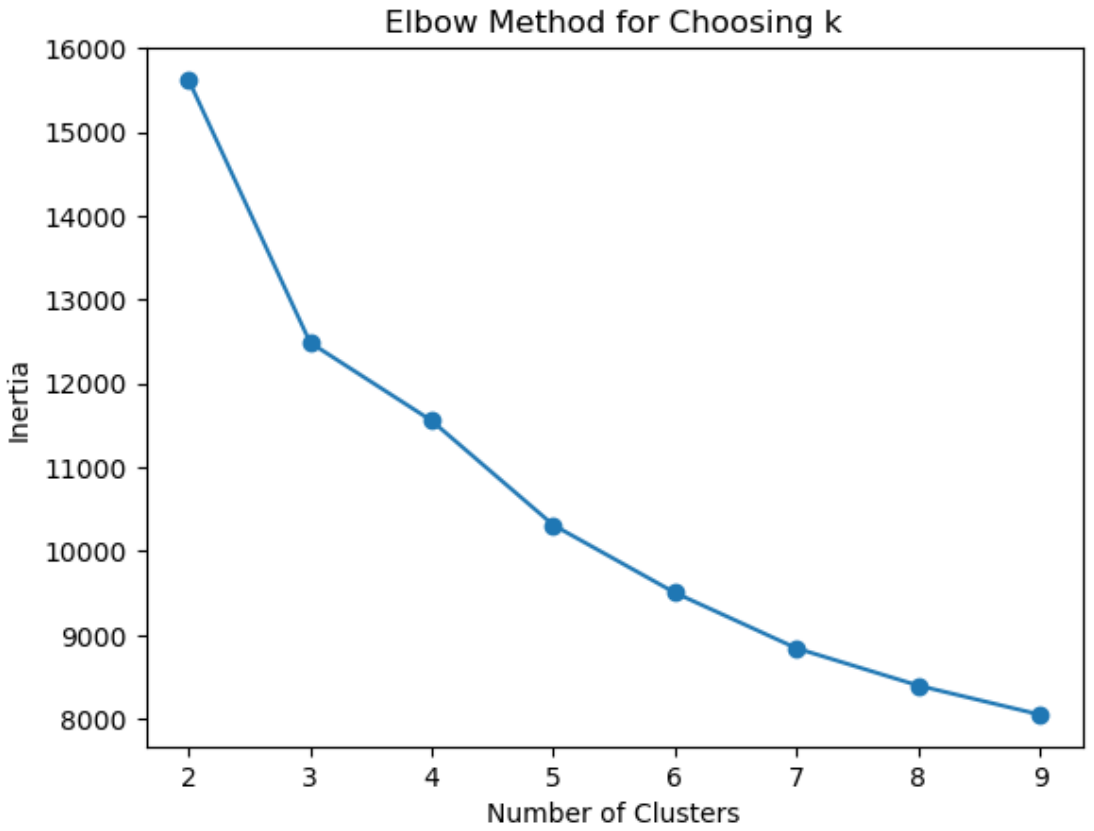
## INTERPRETATION OF SHAP DEPENDENCE

- **SHAP Dependence Plot Insight**: The SHAP dependence plot for location_score reveals a **non-linear relationship** with predicted price — SHAP values remain low for most listings but increase sharply when location_score exceeds ~90, highlighting a substantial impact in premium locations. The color mix across number_of_bedrooms suggests **no strong interaction** with that feature.
- **Difference from PDP and Linear Model**: Unlike SHAP, the **PDP plot shows smoothed average effects**, which can mask non-linear jumps and interactions. A **linear model** would assume a constant slope, missing the sharp rise SHAP captures. SHAP provides **individualized, local explanations** that expose subtleties PDP and linear models may overlook.

# CLUSTERING ANALYSIS

# CLUSTERING ANALYSIS

## ELBOW METHOD TO FIND THE OPTIMAL K VALUE

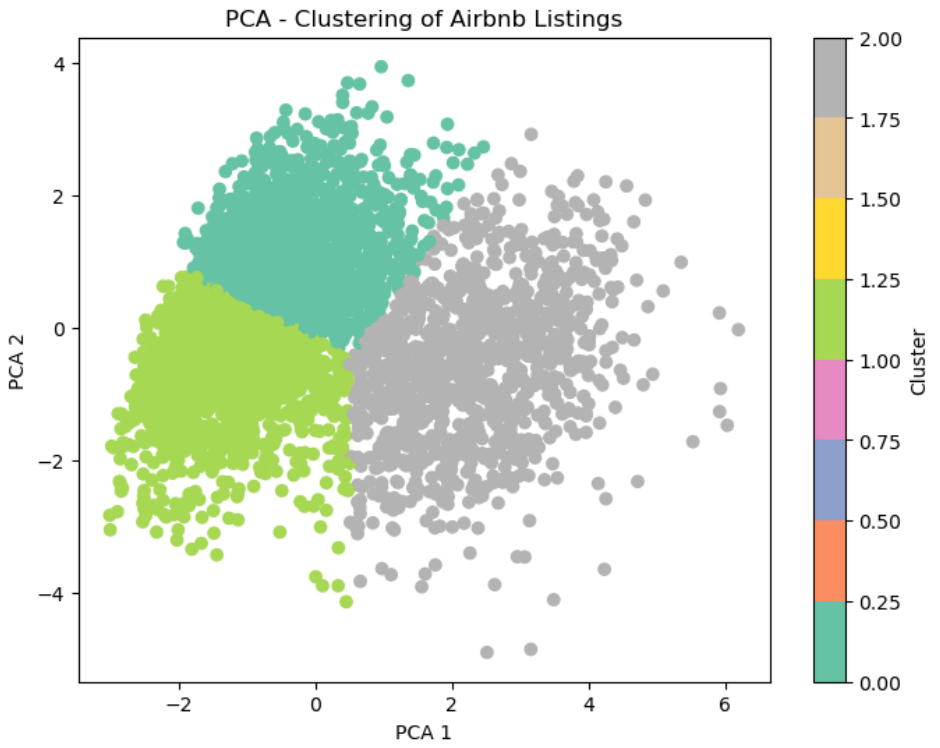



3 is the optimal number of clusters

USC Viterbi
School of Engineering

# CLUSTERING ANALYSIS
## INTERPRETATION OF CLUSTERS

- To identify meaningful consumer segments among Airbnb listings, I performed a clustering analysis using key consumer-facing features such as **nightly rate, number of bedrooms, guest capacity, amenities count, location score, and review score**.
- All features were **standardized** to ensure equal weighting in distance calculations. I applied the **KMeans algorithm**, and the **Elbow Method** was used to determine the optimal number of clusters. The plot indicated that **k=3** provided the best balance between cluster separation and within-cluster compactness.
- To interpret the clusters visually, I applied **PCA** to reduce dimensionality and plotted the clusters in 2D space. The **PCA plot revealed three well-separated groups**, validating the choice of k=3



USC Viterbi School of Engineering

# CLUSTERING ANALYSIS

## CLUSTER PROFILES

| cluster | nightly_rate | number_of_bedrooms | guest_capacity | amenities_count | location_score | review_score |
|---|---|---|---|---|---|---|
| 0 | 143.2 | 1.5 | 2.4 | 8.4 | 78.0 | 4.4 |
| 1 | 135.8 | 1.5 | 2.4 | 5.3 | 62.6 | 3.6 |
| 2 | 296.3 | 4.2 | 5.3 | 11.2 | 71.3 | 4.0 |

This table summarizes the mean values of key features for each cluster, offering a snapshot of typical listings per segment. Cluster 2 reflects Luxury listings (high price, size, and amenities), Cluster 1 aligns with Budget options (lower scores and smaller size), and Cluster 0 captures Mid-Tier listings (moderate price, good location, and strong reviews)

# CLUSTERING ANALYSIS

## CLUSTER SUMMARY ANALYSIS

Key observations:

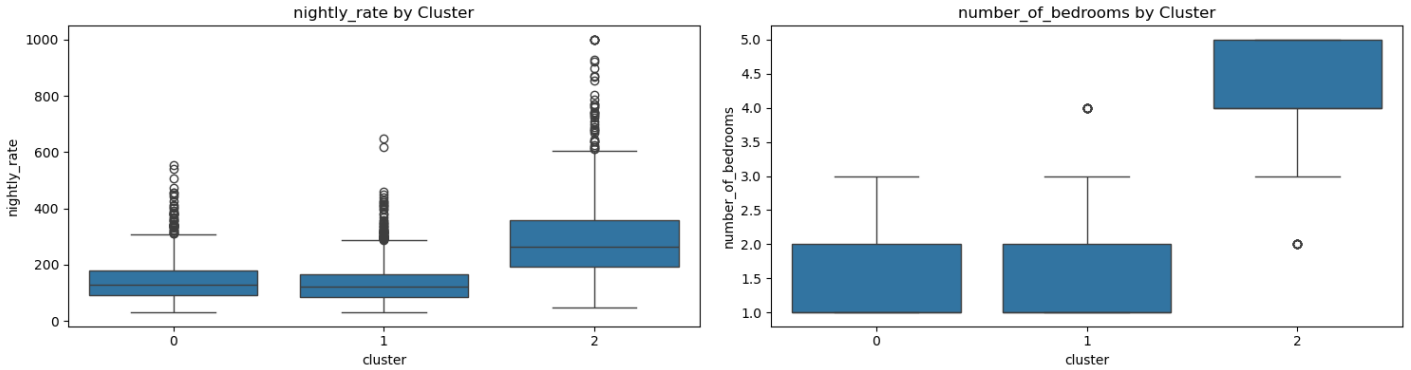
- **Cluster 2 (Luxury)**: Highest price, size, amenities — likely high-end listings.
- **Cluster 1 (Budget)**: Lowest price and location/review score — budget-friendly, lower-quality.
- **Cluster 0 (Mid-Tier)**: Moderate pricing with high review and location score but fewer amenities.

| Cluster | Nightly Rate | Bedrooms | Guest Capacity | Amenities | Location Score | Review Score | Label |
|---|---|---|---|---|---|---|---|
| 0 | 143.2 | 1.5 | 2.4 | 2.4 | 8.4 | 4.4 | Mid-Tier |
| 1 | 135.8 | 1.5 | 2.4 | 5.3 | 62.6 | 3.6 | Budget |
| 2 | 296.3 | 4.2 | 5.3 | 11.2 | 71.3 | 4.0 | Luxury |

USC Viterbi School of Engineering

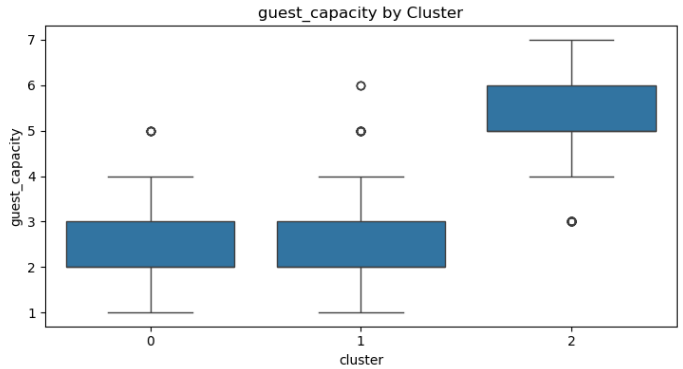
# CLUSTERING ANALYSIS

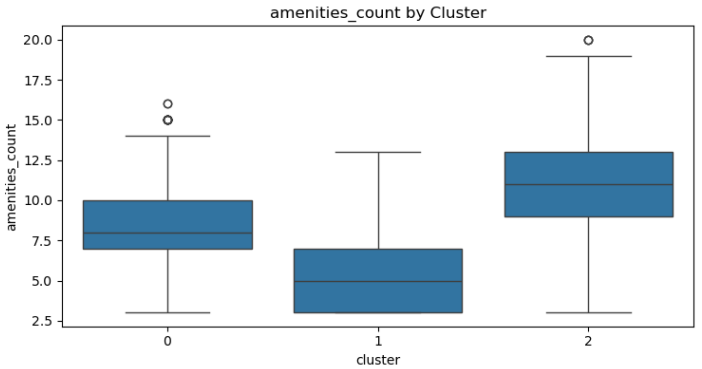
## BOX PLOTS

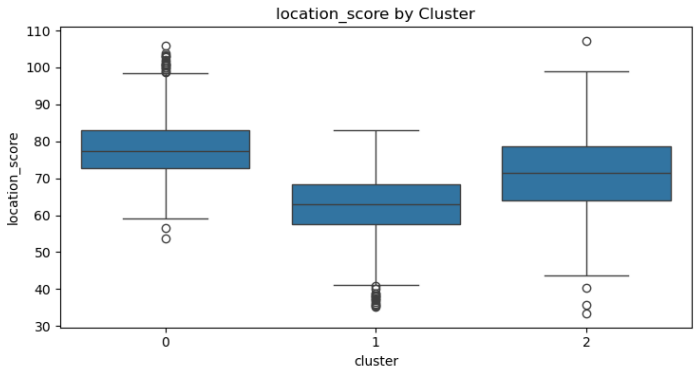








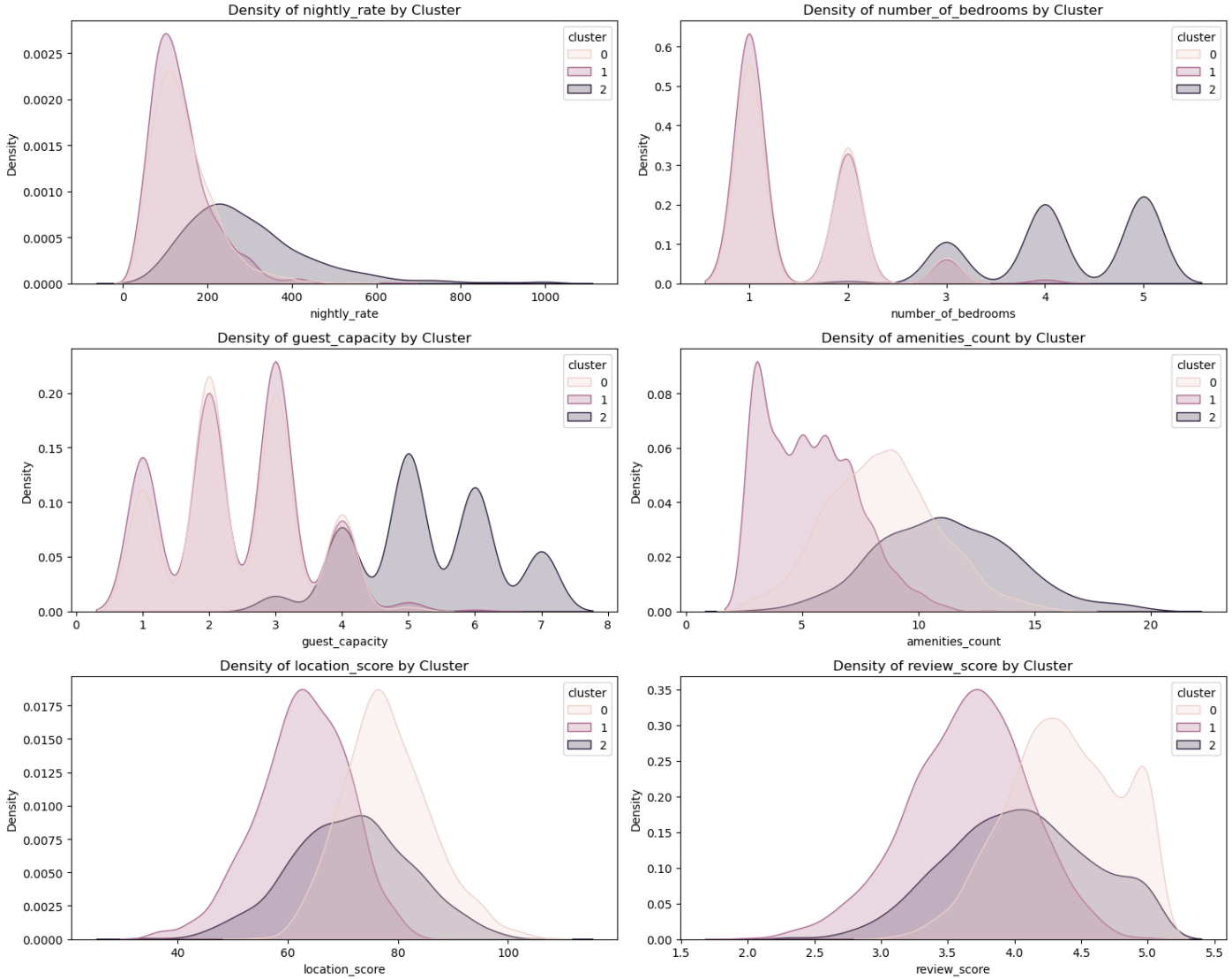


**Box plots reveal clear separation across clusters.** Cluster 2 shows higher medians for price, bedrooms, capacity, and amenities—suggesting luxury listings. Cluster 1 consistently has the lowest values, indicating budget listings, while Cluster 0 reflects mid-range properties with higher review scores but fewer amenities.

USC Viterbi
School of Engineering

# CLUSTERING ANALYSIS

## KDE PLOTS




1. Cluster 2 (Luxury) shows higher densities for nightly rate, number of bedrooms, guest capacity, and amenities—indicating larger, more premium listings.
2. Cluster 1 (Budget) has lower densities across price and quality-related features, including the lowest location and review scores.
3. Cluster 0 (Mid-Tier) lies between the two: moderate price and guest capacity, with relatively high review and location scores but fewer amenities.
4. The curves are **right-skewed** for most variables, particularly **nightly rate and amenities**, suggesting a few high-end outliers

# SUMMARY AND CONCLUSIONS

# SUMMARY AND CONCLUSIONS

Project Objective:

- Identify key drivers of Airbnb nightly pricing and segment listings into actionable consumer tiers to inform pricing strategies and target customer segments.

EDA Insights:

- Detected **skewed distributions** and **right-tailed** variables.
- Identified outliers and strong multicollinearity between guest_capacity and number_of_bedrooms.
- Guided feature selection and model design decisions.

Linear Regression Model:

- Built for **interpretability** and **inference**.
- Key predictors: **property_type, season, number_of_bedrooms, and location_score.**
- Applied log transformation on nightly_rate and removed outliers (|residual| > 3) to improve assumptions.
- Final model achieved:
  - **$R^2 = 0.379$ (test set)**
  - **MAE ≈ $63**
- Model satisfied assumptions of normality and homoscedasticity post-cleaning.
- Peak season and more bedrooms showed the strongest positive price impact.

Decision Tree Model:

- **$R^2 = 0.359$ and MAE ≈ $70** — slightly lower performance, but more interpretable for segmentation.
- **Captured nonlinear relationships** and thresholds (e.g., **price spikes when bedrooms > 2.5 and season = peak).**
- Effective in generating **conditional pricing rules.**

# SUMMARY AND CONCLUSIONS

Model Explainability (PDP & SHAP):

- PDPs:
  - Confirmed strong price increase with number_of_bedrooms.
  - Mild upward trend for location_score.
- 2D PDPs:
  - Revealed interactions between bedrooms × location, bedrooms × season, and season × location.
  - Highlighted amplified pricing in premium, high-demand areas.
- SHAP Values:
  - Reinforced the importance of number_of_bedrooms and season_peak.
  - SHAP dependence plot showed a **nonlinear jump in price** when location_score > 90.
  - No strong interaction between location_score and bedrooms was observed.

Clustering Analysis (KMeans + PCA):

- Features standardized; PCA used for 2D visualization.
- Elbow method suggested **k=3** as optimal cluster count.
- Cluster interpretations:
  - **Cluster 2 (Luxury):** High price, large size, many amenities — targets premium travelers.
  - **Cluster 1 (Budget):** Low price, fewer amenities, lower scores — budget-friendly.
  - **Cluster 0 (Mid-Tier):** Moderate pricing, high review & location scores — mainstream segment.
- Box and KDE plots confirmed distinct separation in pricing, size, and quality metrics.

# SUMMARY AND CONCLUSIONS

## ASSUMPTIONS, LIMITATIONS & NEXT STEPS

Assumptions:

- Linear model assumes normality, linearity, and constant variance.
- Clustering assumes spherical cluster shapes and standardized features.

Limitations:

- Temporal dynamics (e.g., holidays, event-driven demand) not modeled.
- External behaviors (host responsiveness, cancellations) excluded.
- KMeans may oversimplify due to reliance on Euclidean distances.

Next Steps:

- Add **time-based features**, **booking lead times**, and **seasonal curves**.
- Use **ensemble models (e.g., XGBoost)** for nonlinear modeling with better accuracy.
- Explore **text data (e.g., reviews) with NLP** to capture subjective quality signals.
- Enhance personalization by integrating user preferences and historical booking behavior.