# FINAL PROJECT

# ISE-535

Zuxin Xie

# BUSINESS OBJECTIVE

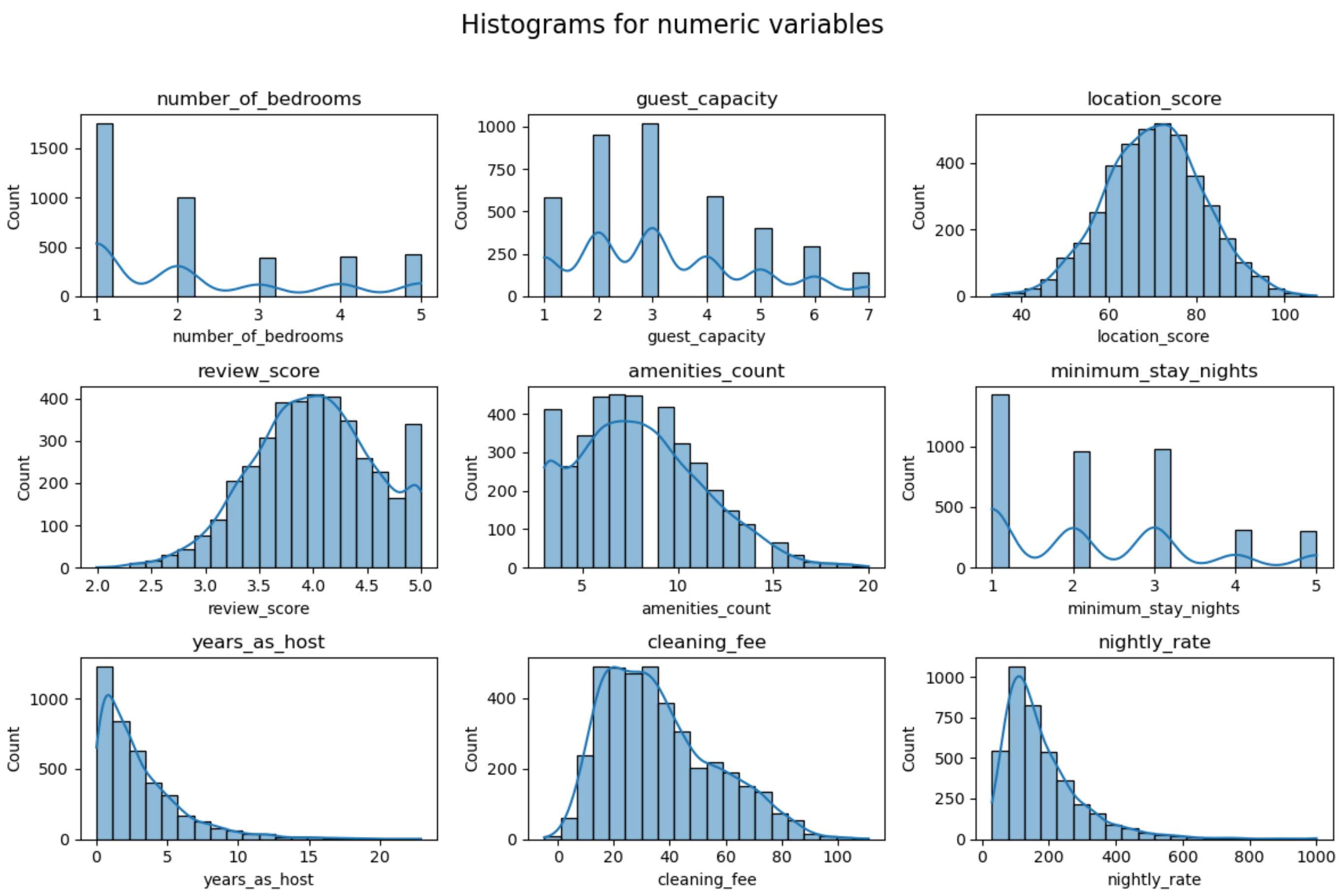
- Understand the key drivers of nightly pricing for Airbnb listings.
- Identify natural groupings (clusters) of listings that align with consumer market segments (e.g., budget, family, luxury).
- Deliver actionable insights that could be used by hosts, platform operators, or consumers.

# EDA REPORT
## DATA OVERVIEW

- 1. Data structure
  - Contains 3982 rows and 14 columns
  - Mixed types: numerical (e.g., number_of_bedrooms, nightly_rate) and categorial (e.g., property_type ,season)
- 2. Missing values
  - No missing values

# VISUALIZATION FOR NUMERIC VARIABLES

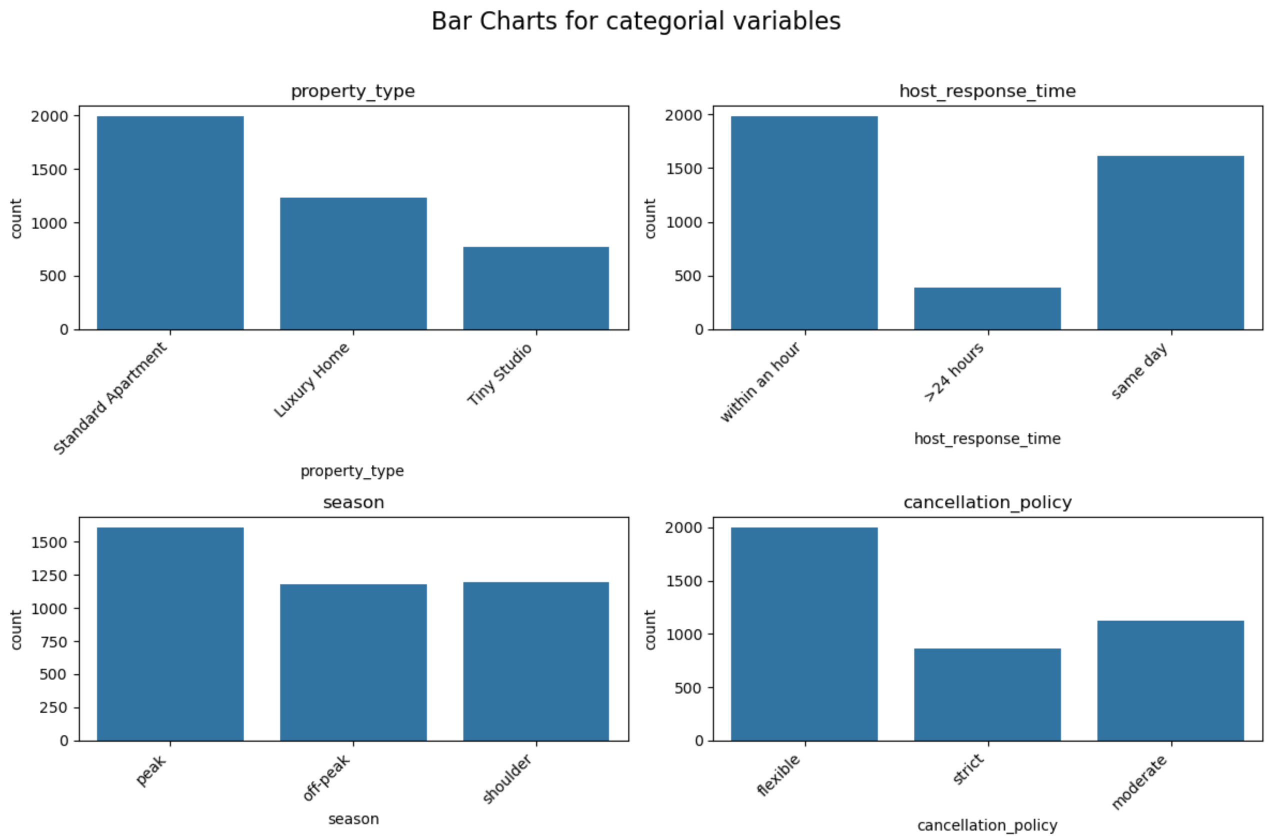
- The target nightly_rate is right-skewed and has lots of outliers, suggesting the need for  log transformation.
- The location score displays a normal distribution, which is suitable for direct use.



# EDA REPORT

## VISUALIZATION OF CATEGORIAL VARIABLES

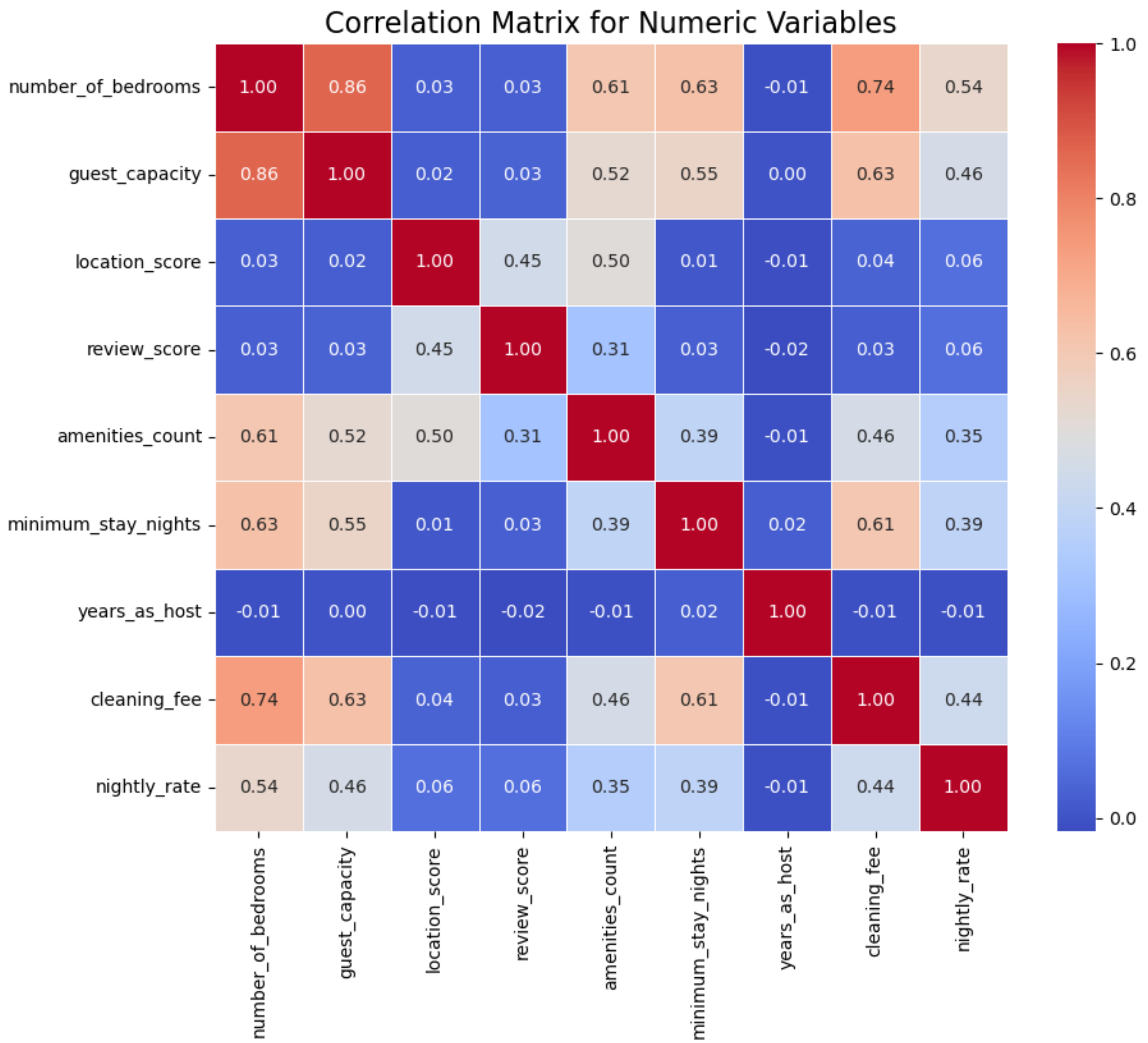
- Most listings in property_type are Standard Apartments, followed by Luxury Homes and Tiny Studios.
- Distribution in season is relatively balanced, though peak season has the highest count.



# EDA REPORT

## VISUALIZATION OF CORRELATION MATRIX

- Number_of_bedrooms is highly correlated to several features (e.g., guest_capacity, cleaning_fee), suggesting potential multicollinearity.
- Location_score and review_score show little linear correlation with nightly_rate, indicating that nonlinear relationship may exist.



USC Viterbi
School of Engineering

# STATISTICAL ANALYSIS
## HYPOTHESIS TEST

- The P-value is 0, so we can reject the H0 hypothesis, which means there is a statistically significant difference between different property type.
- Practically, the mean nightly rate for Luxury Home($277) is almost double than that of Standard Apartment($144), suggesting that property classification is an important pricing factor for hosts and guests.

### 3.1 Compare `nightly_rate` between different `property_types`

$H_0$: All property types have the same average nightly rate.
$H_1$: At least one property type has a different average nightly rate.

```python
from scipy.stats import f_oneway

df_h1=airbnb_df[['property_type', 'nightly_rate']].dropna()
print(df_h1.groupby('property_type')['nightly_rate'].mean().sort_values())
grouped_data = [group['nightly_rate'].values for name, group in df_h1.groupby('property_type')]
f_stat, p_val = f_oneway(*grouped_data)
print(f"\nF-statistic: {f_stat:.4f}, P-value: {p_val:.4f}")
```

✓ 0.0s

```
property_type
Tiny Studio           118.088750
Standard Apartment    144.535543
Luxury Home           277.287439
Name: nightly_rate, dtype: float64

F-statistic: 822.6232, P-value: 0.0000
```

# STATISTICAL ANALYSIS
# HYPOTHESIS TEST

- The p-value is 0.3966, which is greater than 0.05, so we fail to reject the H0 hypothesis. This means there is no statistically significant difference in nightly rates among listings with different host response times.
- Practically, the average nightly rate only ranges from $176.88 (for >24 hours) to $182.91 (for within an hour), a difference of less than $7. This small variation is likely not meaningful from a business perspective, suggesting that host response time may not be a key factor influencing price.

### 3.2 Compare `nightly_rate` between different `host_response_time`

$H_0$: All host_response_time have the same average nightly rate.
$H_1$: At least one host_response_time has a different average nightly rate.

```
df_h2=airbnb_df[['host_response_time', 'nightly_rate']].dropna()
print(df_h2.groupby('host_response_time')['nightly_rate'].mean().sort_values())
grouped_data = [group['nightly_rate'].values for name, group in df_h2.groupby('host_response_time')]
f_stat, p_val = f_oneway(*grouped_data)
print(f"\nF-statistic: {f_stat:.4f}, P-value: {p_val:.4f}")
```

✓ 0.0s

```
host_response_time
>24 hours         176.880615
same day          177.937688
within an hour    182.913739
Name: nightly_rate, dtype: float64

F-statistic: 0.9249, P-value: 0.3966
```

# STATISTICAL ANALYSIS
## HYPOTHESIS TEST

- The p-value is 0.0024, which is less than 0.05, so we can reject the H0 hypothesis. There is a statistically significant difference in nightly rates between high and low location score groups.
- Practically, the average difference is $12, suggesting that location score may not be a key factor influencing price.

3.3 Compare nightly_rate between two groups based on location_score.

$H_0$: There is no difference in average nightly rates between high and low location score listings.
$H_1$: There is a difference in average nightly rates between high and low location score listings.

```
from scipy.stats import ttest_ind

airbnb_df['location_group'] = airbnb_df['location_score'] > airbnb_df['location_score'].median()

df_h3=airbnb_df[['location_group', 'nightly_rate']].dropna()
high_loc=df_h3[df_h3['location_group']==True]['nightly_rate']
low_loc=df_h3[df_h3['location_group']==False]['nightly_rate']

print("Mean nightly_rate (high location):", high_loc.mean())
print("Mean nightly_rate (low location):", low_loc.mean())

t_stat, p_val = ttest_ind(high_loc, low_loc, equal_var=False)
print(f"\nT-statistic: {t_stat:.4f}, P-value: {p_val:.4f}")
```

✓ 0.0s

```
Mean nightly_rate (high location): 186.12470351758793
Mean nightly_rate (low location): 174.49551204819278

T-statistic: 3.0357, P-value: 0.0024
```

# STATISTICAL ANALYSIS
## HYPOTHESIS TEST

- The P-value is 0, so we can reject the H0 hypothesis, which means there is a statistically significant difference between different seasons.
- Practically, the mean nightly rate for peak($277) is much higher than that of off-peak($143), suggesting that season is an important pricing factor.

### 3.4 Compare `nightly_rate` between different `season`

$H_0$: All seasons have the same average nightly rate.
$H_1$: At least one season has a different average nightly rate.

```
df_h4=airbnb_df[['season', 'nightly_rate']].dropna()
print(df_h4.groupby('season')['nightly_rate'].mean().sort_values())
grouped_data = [group['nightly_rate'].values for name, group in df_h4.groupby('season')]
f_stat, p_val = f_oneway(*grouped_data)
print(f"\nF-statistic: {f_stat:.4f}, P-value: {p_val:.4f}")
```

✓ 0.0s

```
season
off-peak    143.157441
shoulder    173.695573
peak        212.502321
Name: nightly_rate, dtype: float64

F-statistic: 121.2035, P-value: 0.0000
```

# STATISTICAL ANALYSIS
## HYPOTHESIS TEST

- The P-value is 0, so we can reject the H0 hypothesis, which means there is a statistically significant difference between different seasons.
- Practically, the mean nightly rate increased from $129 for one-bedroom group to $305 for 5-bedroom group, suggesting that number of bedroom is an important pricing factor.

### 3.5 Compare `nightly_rate` between different `number_of_bedrooms`

$H_0$: All number of bedrooms have the same average nightly rate.
$H_1$: At least one group has a different average nightly rate.

```
df_h5=airbnb_df[['number_of_bedrooms', 'nightly_rate']].dropna()
print(df_h5.groupby('number_of_bedrooms')['nightly_rate'].mean().sort_values())
grouped_data = [group['nightly_rate'].values for name, group in df_h5.groupby('number_of_bedrooms')]
f_stat, p_val = f_oneway(*grouped_data)
print(f"\nF-statistic: {f_stat:.4f}, P-value: {p_val:.4f}")
```

✓ 0.0s

```
number_of_bedrooms
1    129.085428
2    151.310359
3    249.738252
4    274.192118
5    305.067796
Name: nightly_rate, dtype: float64

F-statistic: 430.4302, P-value: 0.0000
```

# LINEAR MODEL

## ORIGINAL MODEL

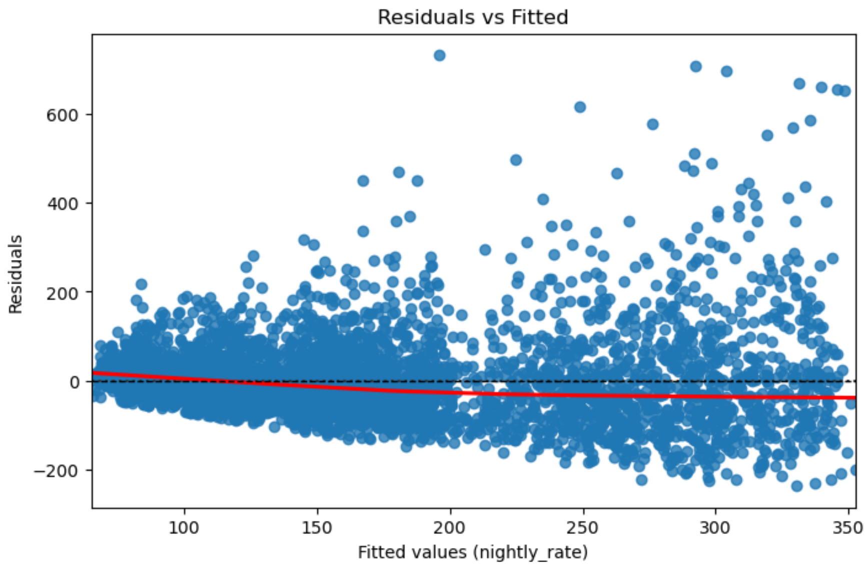
- Converted all categorial variables to one-hot encoded binary variables.
- No transformation on numerical variables.
- There is a strong multicollinearity problem in this model, which can be inferred from the correlation matrix earlier.
- Some factors show no statistical significance.

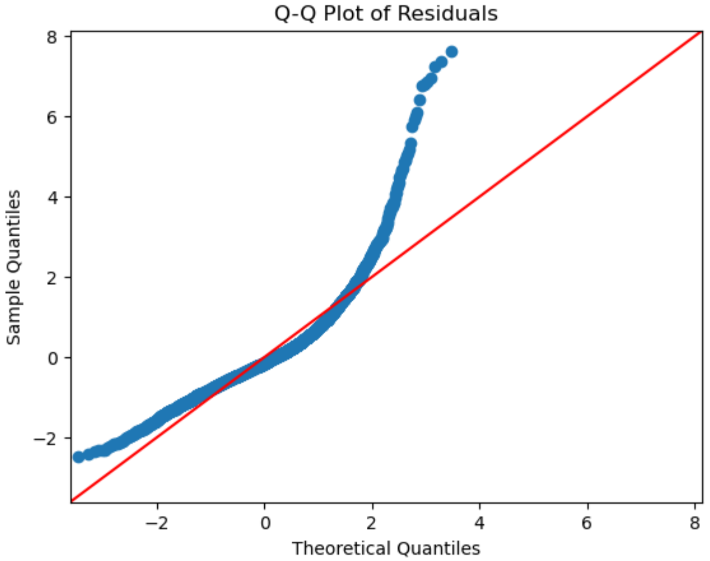
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            nightly_rate   R-squared:                       0.371
Model:                             OLS   Adj. R-squared:                  0.368
Method:                  Least Squares   F-statistic:                     137.3
Date:                 Sat, 10 May 2025   Prob (F-statistic):               0.00
Time:                         16:33:45   Log-Likelihood:                -23825.
No. Observations:                 3982   AIC:                         4.769e+04
Df Residuals:                     3964   BIC:                         4.780e+04
Df Model:                           17
Covariance Type:             nonrobust
=============================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
const                          103.3828     23.500      4.399      0.000      57.310     149.456
number_of_bedrooms              22.4123      3.425      6.544      0.000      15.697      29.127
guest_capacity                  -0.8843      1.872     -0.472      0.637      -4.554       2.786
location_score                   0.3200      0.256      1.252      0.211      -0.181       0.821
review_score                     4.4313      3.077      1.440      0.150      -1.601      10.464
amenities_count                  0.2264      0.729      0.310      0.756      -1.204       1.656
minimum_stay_nights              2.0573      1.797      1.145      0.252      -1.467       5.581
years_as_host                   -0.0139      0.513     -0.027      0.978      -1.019       0.991
cleaning_fee                    -0.0693      0.138     -0.501      0.617      -0.341       0.202
location_group                   2.9849      5.075      0.588      0.556      -6.965      12.935
property_type_Standard Apartment -77.3458    9.035     -8.561      0.000     -95.059     -59.633
property_type_Tiny Studio      -92.0623     11.985     -7.681      0.000    -115.560     -68.565
host_response_time_same day      2.8512      5.445      0.524      0.601      -7.824      13.527
host_response_time_within an hour 6.1827     5.343      1.157      0.247      -4.294      16.659
season_peak                     72.2113      3.697     19.532      0.000      64.963      79.460
season_shoulder                 29.6601      3.954      7.502      0.000      21.909      37.412
cancellation_policy_moderate    -3.4985      3.594     -0.974      0.330     -10.544       3.547
cancellation_policy_strict       0.8349      3.925      0.213      0.832      -6.860       8.530
==============================================================================
Omnibus:                      1754.776   Durbin-Watson:                   1.954
Prob(Omnibus):                   0.000   Jarque-Bera (JB):            13779.283
Skew:                            1.920   Prob(JB):                         0.00
Kurtosis:                       11.265   Cond. No.                     1.38e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.38e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

# LINEAR MODEL

## VISUALIZATIONS FOR ORIGINAL MODEL

- Residual vs Fitted plot
  - Displayed a funnel shape, indicating heteroscedasticity problem
  - A log transformation of nightly_rate is needed.
- Q-Q plot
  - This nonlinear curve suggests non-normal residuals, which can affect inference accuracy.



# LINEAR MODEL

## IMPROVED MODEL

- Converted all categorial variables to one-hot encoded binary variables.
- Applied log transformation on nightly_rate.
- Removed the factors that show no statistical significance.
- Compared to the original model, R square increased, which explains the variations better. The skewness and kurtosis are within normal range, suggesting normality.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            nightly_rate   R-squared:                       0.403
Model:                             OLS   Adj. R-squared:                  0.402
Method:                  Least Squares   F-statistic:                     446.7
Date:                 Sat, 10 May 2025   Prob (F-statistic):               0.00
Time:                         20:39:06   Log-Likelihood:                -2596.8
No. Observations:                 3982   AIC:                             5208.
Df Residuals:                     3975   BIC:                             5252.
Df Model:                            6
Covariance Type:             nonrobust
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                              4.6596      0.072     64.888      0.000       4.519       4.800
number_of_bedrooms                 0.0939      0.013      7.326      0.000       0.069       0.119
location_score                     0.0035      0.001      5.301      0.000       0.002       0.005
property_type_Standard Apartment  -0.4162      0.037    -11.379      0.000      -0.488      -0.344
property_type_Tiny Studio         -0.5723      0.044    -12.891      0.000      -0.659      -0.485
season_peak                        0.4123      0.018     23.124      0.000       0.377       0.447
season_shoulder                    0.1910      0.019     10.008      0.000       0.154       0.228
==============================================================================
Omnibus:                         3.559   Durbin-Watson:                   1.975
Prob(Omnibus):                   0.169   Jarque-Bera (JB):                3.496
Skew:                           -0.065   Prob(JB):                        0.174
Kurtosis:                        3.064   Cond. No.                         827.
==============================================================================

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

# LINEAR MODEL

## INTERPRETATION FOR IMPROVED MODEL

- Since a log transformation is applied on the target, the way we interpret the coefficients changed.

$$\log(nightly\ rate) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

$$nightly\ rate = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)$$

Take location score as an example. For each unit increase in location score, it will lead to an increase of $(e^{0.0035} - 1)\% \approx 0.35\%$ on nightly_rate.

For the property type, if the property is Standard Apartment, it will lead to a decrease of $(1 - e^{0.4162})\% \approx 51.6\%$ on the rate.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | nightly_rate | R-squared: | 0.403 |
| Model: | OLS | Adj. R-squared: | 0.402 |
| Method: | Least Squares | F-statistic: | 446.7 |
| Date: | Sat, 10 May 2025 | Prob (F-statistic): | 0.00 |
| Time: | 20:39:06 | Log-Likelihood: | -2596.8 |
| No. Observations: | 3982 | AIC: | 5208. |
| Df Residuals: | 3975 | BIC: | 5252. |
| Df Model: | 6 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 4.6596 | 0.072 | 64.888 | 0.000 | 4.519 | 4.800 |
| number_of_bedrooms | 0.0939 | 0.013 | 7.326 | 0.000 | 0.069 | 0.119 |
| location_score | 0.0035 | 0.001 | 5.301 | 0.000 | 0.002 | 0.005 |
| property_type_Standard Apartment | -0.4162 | 0.037 | -11.379 | 0.000 | -0.488 | -0.344 |
| property_type_Tiny Studio | -0.5723 | 0.044 | -12.891 | 0.000 | -0.659 | -0.485 |
| season_peak | 0.4123 | 0.018 | 23.124 | 0.000 | 0.377 | 0.447 |
| season_shoulder | 0.1910 | 0.019 | 10.008 | 0.000 | 0.154 | 0.228 |

| | | | |
|---|---|---|---|
| Omnibus: | 3.559 | Durbin-Watson: | 1.975 |
| Prob(Omnibus): | 0.169 | Jarque-Bera (JB): | 3.496 |
| Skew: | -0.065 | Prob(JB): | 0.174 |
| Kurtosis: | 3.064 | Cond. No. | 827. |

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

# LINEAR MODEL
## INTERPRETATION FOR IMPROVED MODEL

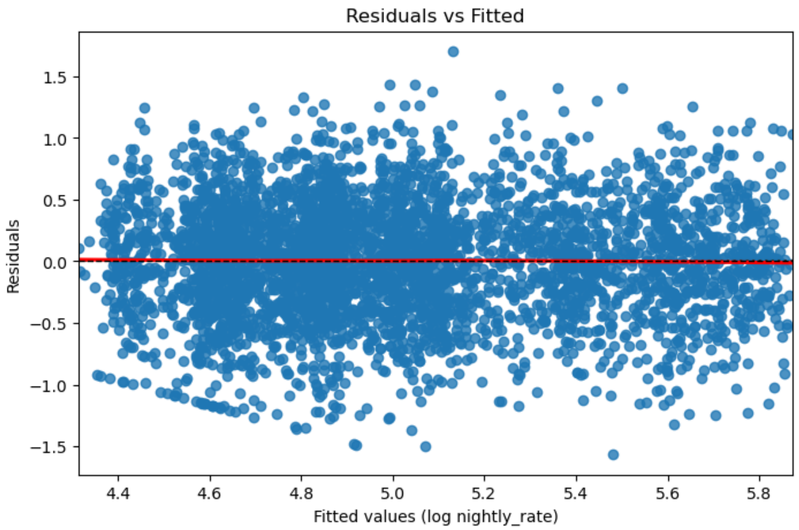
- Const Note:
  - I used drop_first = True when encoding categorial variables to reduce multicollinearity. Thus, coefficients for Tiny Studio and Standard Apartment are compared to Luxury Home. So do the coefficients for season. The coefficients for peak and shoulder are compared to off-peak

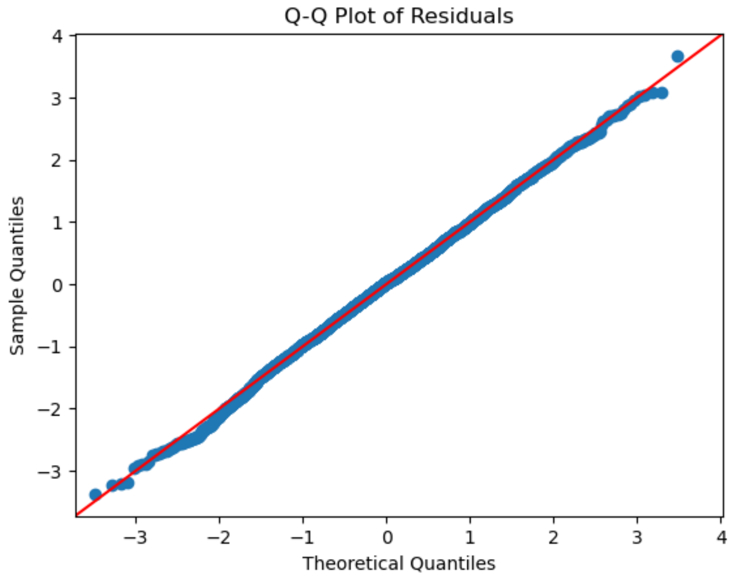
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            nightly_rate   R-squared:                       0.403
Model:                             OLS   Adj. R-squared:                  0.402
Method:                  Least Squares   F-statistic:                     446.7
Date:                 Sat, 10 May 2025   Prob (F-statistic):               0.00
Time:                         20:39:06   Log-Likelihood:                -2596.8
No. Observations:                 3982   AIC:                             5208.
Df Residuals:                     3975   BIC:                             5252.
Df Model:                            6
Covariance Type:             nonrobust
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                              4.6596      0.072     64.888      0.000       4.519       4.800
number_of_bedrooms                 0.0939      0.013      7.326      0.000       0.069       0.119
location_score                     0.0035      0.001      5.301      0.000       0.002       0.005
property_type_Standard Apartment  -0.4162      0.037    -11.379      0.000      -0.488      -0.344
property_type_Tiny Studio         -0.5723      0.044    -12.891      0.000      -0.659      -0.485
season_peak                        0.4123      0.018     23.124      0.000       0.377       0.447
season_shoulder                    0.1910      0.019     10.008      0.000       0.154       0.228
==============================================================================
Omnibus:                        3.559   Durbin-Watson:                   1.975
Prob(Omnibus):                  0.169   Jarque-Bera (JB):                3.496
Skew:                          -0.065   Prob(JB):                        0.174
Kurtosis:                       3.064   Cond. No.                         827.
==============================================================================

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

# LINEAR MODEL

## IMPROVED MODEL

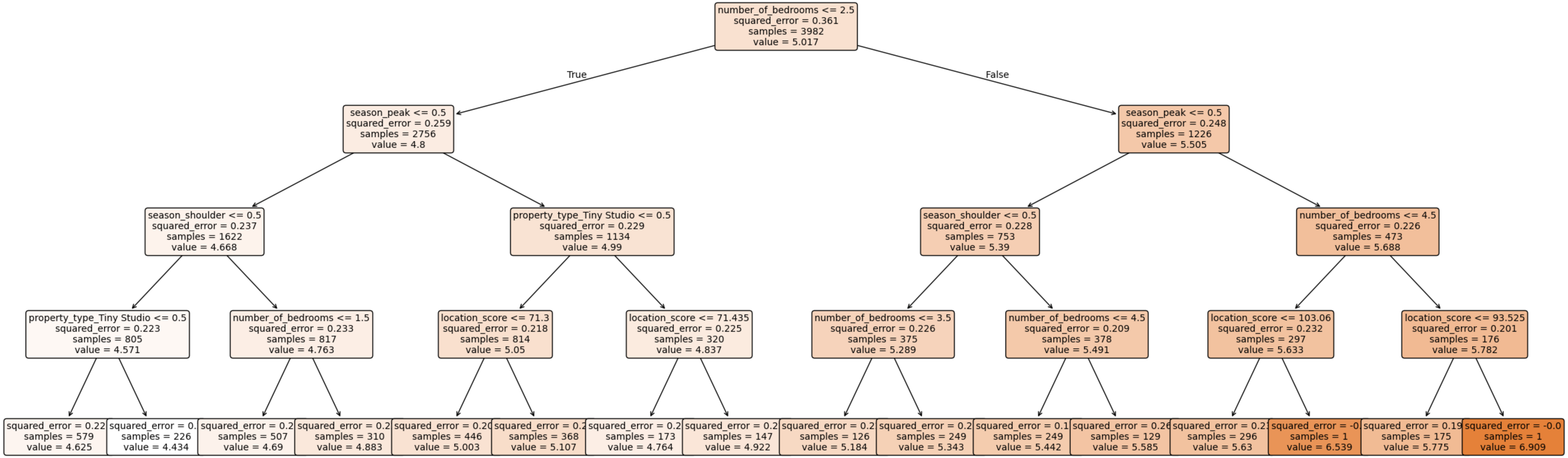
- Residual vs Fitted plot
  - Displayed no special patterns, suggesting good model fit
  - Indicates reduction in heteroscedasticity, satisfying linear model assumptions more closely
- Q-Q plot
  - Residuals align well with the 45-degree line
  - Confirms residuals are approximately normally distributed, improving reliability of statistical inference



# TREE MODEL

## VISUALIZATION OF FULL TREE



# TREE MODEL

## INTERPRETATION FOR TREE MODEL

- Goal
  - The tree model is predicting log(nightly_rate) using decision rules based on key features.
- Root
  - The first split is on number_of_bedrooms, suggesting that room size is the strongest predictor for price.
- Sub-tree left (bedroom<=2.5)
  - If not in peak season, prices drop slightly further if it's shoulder season
  - If not a Tiny Studio, nightly rate is higher
- Sub-tree right (bedroom>=2.5)
  - If not peak season and not shoulder season, prices are highest
  - Peak season and large property leads to higher price

# TREE MODEL

## COMPARISON BETWEEN LINEAR AND TREE MODEL

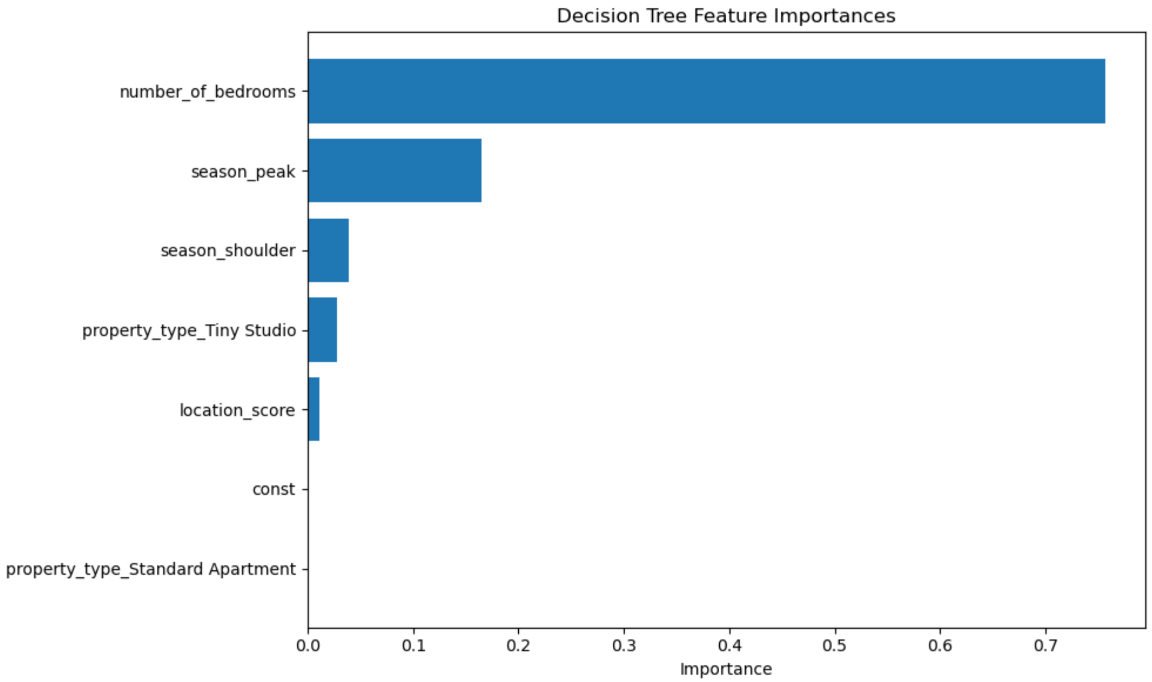
- Using same dataset, the performance between two models are almost the same.
- Linear Model is more interpretable and statistically robust
- Decision Tree captures non-linear interactions and rules more naturally

```
Linear Regression:
  R²: 0.37885064997462947
  MAE: 0.39171134277518804

Decision Tree:
  R²: 0.37376302459665656
  MAE: 0.39207600056686276
```

# MODEL INTERPRETATION AND EXPLANATION

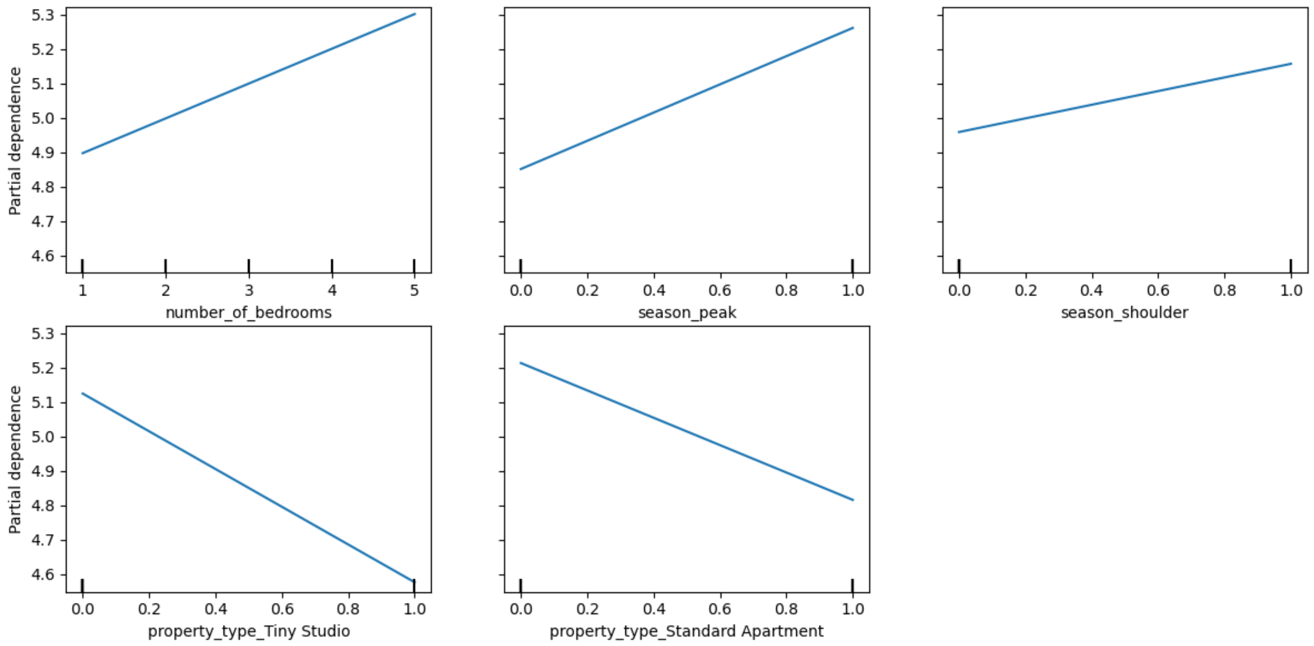
## FEATURE IMPORTANCE

- Number_of_bedrooms
  - This feature contributes more than 70% of total importance.
- Season
  - Season_peak and season_shoulder follow, affecting the price strongly.



# MODEL INTERPRETATION AND EXPLANATION

## PDPS

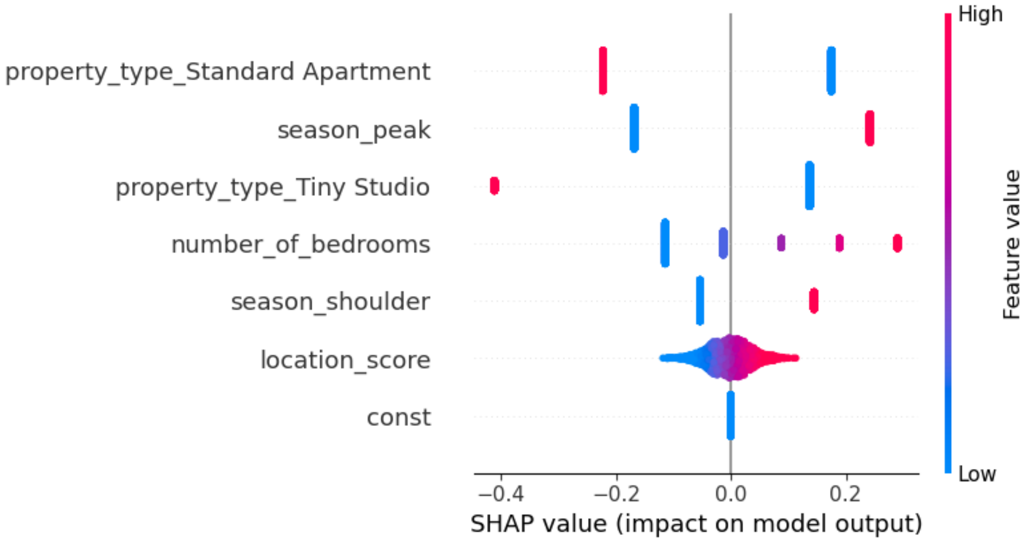


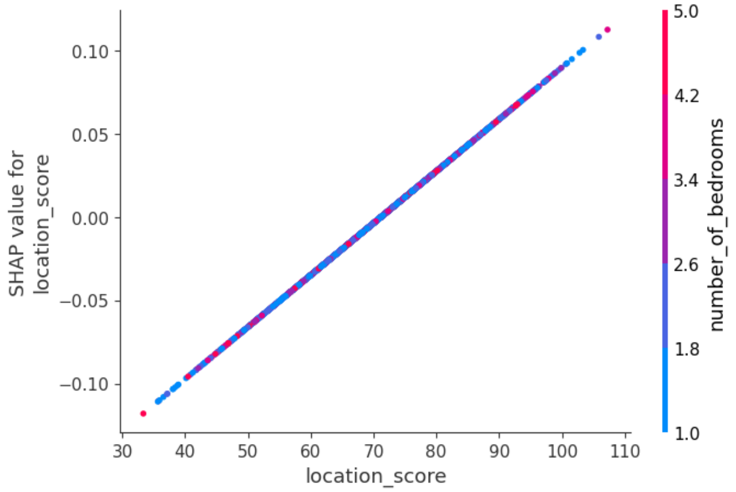

- Number of bedrooms: A positive relationship that more bedrooms lead to higher nightly rate
- Season: Switching from off-peak to peak or shoulder both increase the nightly rate. But the effects from shoulder is weaker than that from peak.
- Property type: Switching from Luxury Home to Tiny Sudio or Standard Apartment displays a negative relationship to the nightly price.

# MODEL INTERPRETATION AND EXPLANATION

## SHAP

- The most influential feature is property_type_standard Apartment, due to its widest shap value spread.
- The shap value for location score is perfect linear, indicating a simple monotonic effect on target.
- The colors are mixed together which means that there is no visible interaction with number_of_bedrooms.

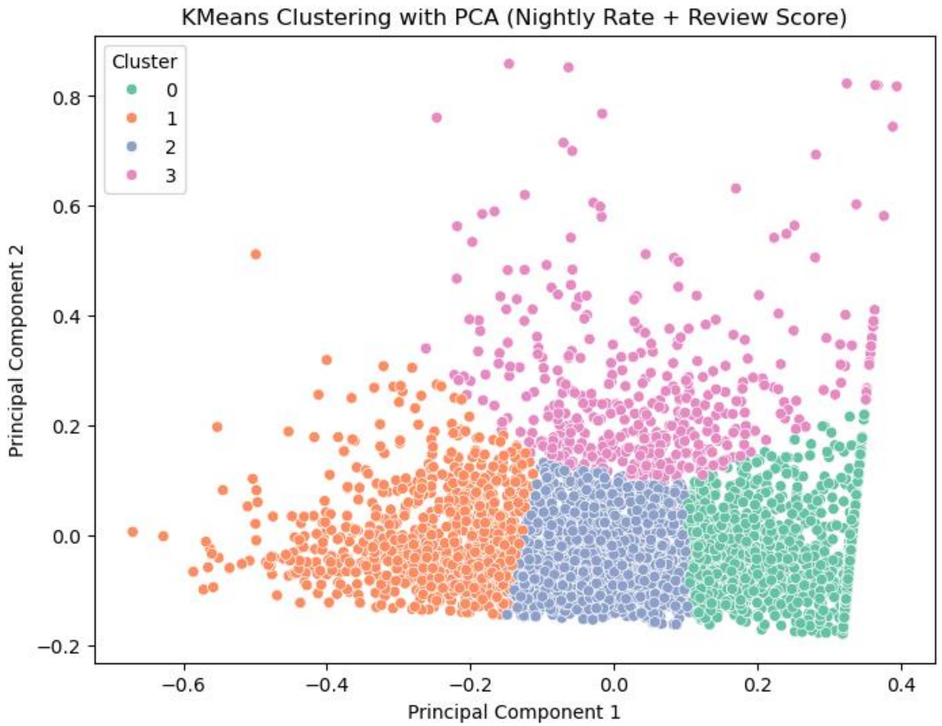


# CLUSTER ANALYSIS

## K-MEANS ON REVIEW_SCORE

- Using k-means, I clustered the nightly_rate and review_score into 4 clusters.
- Cluster 0 (green dots)
  - Label: top-value
  - The house with moderate price and high reviews which means good quality at a reasonable price
- Cluster 1 (orange dots)
  - Label: risky mid-tier
  - They have similar price to top value group but the remarks are significantly worse, suggesting poor qualities

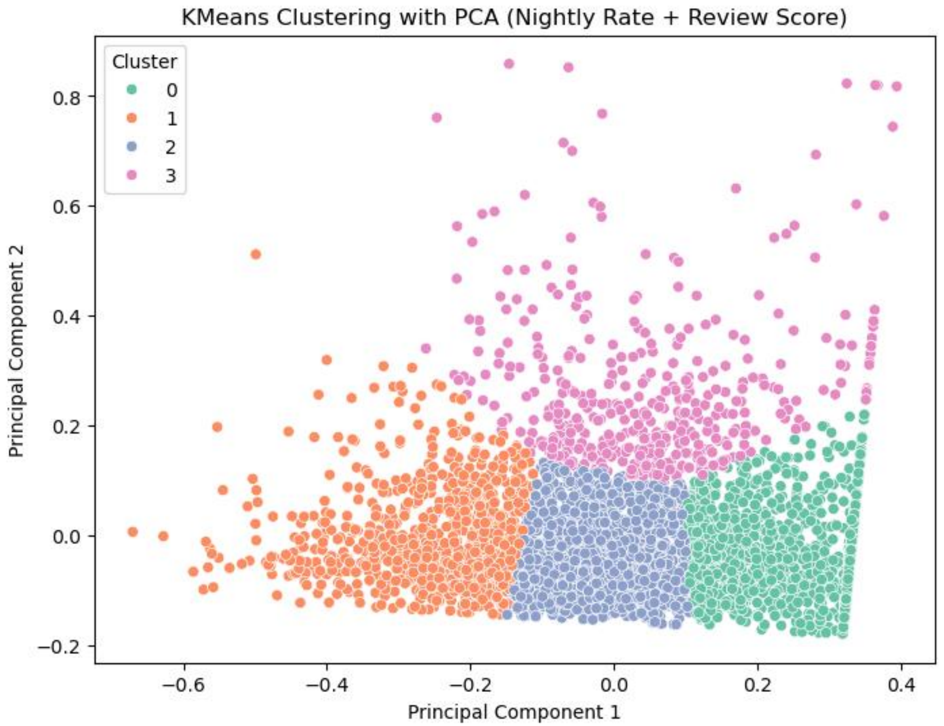



| | nightly_rate | review_score |
|---|---|---|
| cluster | | |
| 0 | 162.250725 | 4.672298 |
| 1 | 156.336554 | 3.264250 |
| 2 | 138.873562 | 3.964760 |
| 3 | 438.622826 | 4.082826 |

# CLUSTER ANALYSIS

## K-MEANS ON REVIEW_SCORE

- Cluster 2 (blue dots)
  - Label: budget friendly
  - The houses with low price and not-bad reviews which means you can enjoy the okay quality at a low cost
- Cluster 3 (pink dots)
  - Label: luxury tier
  - The price is extremely high and the reviews are decent, suggesting expensive listings, good but not elite reviews




| cluster | nightly_rate | review_score |
|---|---|---|
| 0 | 162.250725 | 4.672298 |
| 1 | 156.336554 | 3.264250 |
| 2 | 138.873562 | 3.964760 |
| 3 | 438.622826 | 4.082826 |

# CONCLUSION

- EDA
  - EDA revealed right-skewed distributions, seasonal trends, and strong correlations.
- Linear Regression
  - Prices increase with bedrooms, location score, and during peak seasons
  - Tiny Studios and Standard Apartments are priced lower than Luxury Homes
  - Log transformation improved model assumptions and interpretability
- Decision Tree
  - Large homes in peak season fetch highest prices
  - Bedrooms and seasonality are dominant split criteria

# CONCLUSION

- SHAP/PDP
  - Bedroom count and season are most influential
  - Location has a consistent, linear effect
  - Property type has strong categorical shifts in pricing
- Cluster Analysis
  - Prioritize pricing strategy by cluster: emphasize value, watch out for risky mid-tier listings
  - Invest in features that drive price: larger size, peak availability, high location score
  - Use insights to refine listing strategy, optimize revenue, and identify outlier properties