

Saudi Arabia Condo Price Predictions Report

Hypothesis:

In predicting Saudi Arabia condo prices, we hypothesize that the number of bedrooms, the number of beds, and the location will significantly influence the property's rental price. We expect that properties with more bedrooms and beds will have higher rental prices. Additionally, we anticipate that the location of the property will be a crucial factor, with prices varying across different cities in Saudi Arabia.

Saudi Arabia Condo Price Predictions Report

About the Data:

The dataset was obtained by scraping data from the Airbnb website in Saudi Arabia. It contains information on rental properties with the following columns:

name: The name of the property.

header: The location of the property.

beds: The number of beds in the property.

bedrooms: The number of bedrooms in the property.

date_range: The duration of stay in the property.

price: The price of staying in the property.

rating: The rating of the property.

Data Cleaning and Exploration:

Data Shape:

The dataset has a shape of (3600, 8).

Missing Values:

There are missing values in the 'price' and 'rating' columns.

Handling Missing Values:

Rows with missing values are dropped.

Feature Engineering:

Extracted features from the 'bedrooms' column to create 'num_beds' and 'num_bedrooms'.

Cleaned and converted 'price' values to float.

Extracted 'rating' and 'num_comments' from the original 'rating' column.

Drop Unnecessary Columns:

Dropped unnecessary columns: 'rating', 'bedrooms', 'date_range', 'Unnamed: 0'.

Location Mapping:

Mapped and renamed locations for consistency.

Data Analysis:**Location Distribution:**

Most locations are in Riyadh and Jeddah.

Price Distribution:

The mean price is \$156.54, with a minimum of \$48 and a maximum of \$448.

There are potential outliers in the data.

Number of Comments Distribution:

The mean number of comments is approximately 21.81, with a minimum of 3 and a maximum of 178.

There are potential outliers in the data.

Data Encoding and Cleaning:**One-Hot Encoding:**

Applied one-hot encoding to the 'location' column.

Replace Spaces:

Replaced spaces with NaN and dropped rows with null values.

Convert Boolean Columns:

Converted boolean columns to 1 and 0.

Convert Data Types:

Converted data types for 'num_beds', 'num_bedrooms', and other relevant columns.

Data Correlation:

A heatmap was created to visualize the correlation matrix of the cleaned dataset.

Model Building and Evaluation:

Trained and evaluated several regression models:

Linear Regression:

MSE: 635.47

R² Score: 0.37

K-Nearest Neighbors:

MSE: 158.35

R² Score: 0.84

Decision Tree:

MSE: 151.32

R² Score: 0.85

Random Forest:

MSE: 87.17

R² Score: 0.91

XGBoost:

MSE: 114.75

R² Score: 0.89

Conclusion:

The Random Forest model performed the best among the models evaluated, with the lowest MSE and highest R² score.

Further analysis and fine-tuning of the model could improve predictive performance.

Consideration of outliers in the data is recommended for more accurate predictions.

Hypothesis Validation:

The results of our regression models support our hypothesis. The number of bedrooms, the number of beds, and the location significantly influence condo prices in Saudi Arabia. The Random Forest model, which takes into account these features, provided the most accurate predictions. Adjustments and enhancements to the model, considering outliers and additional features, could further improve its performance.