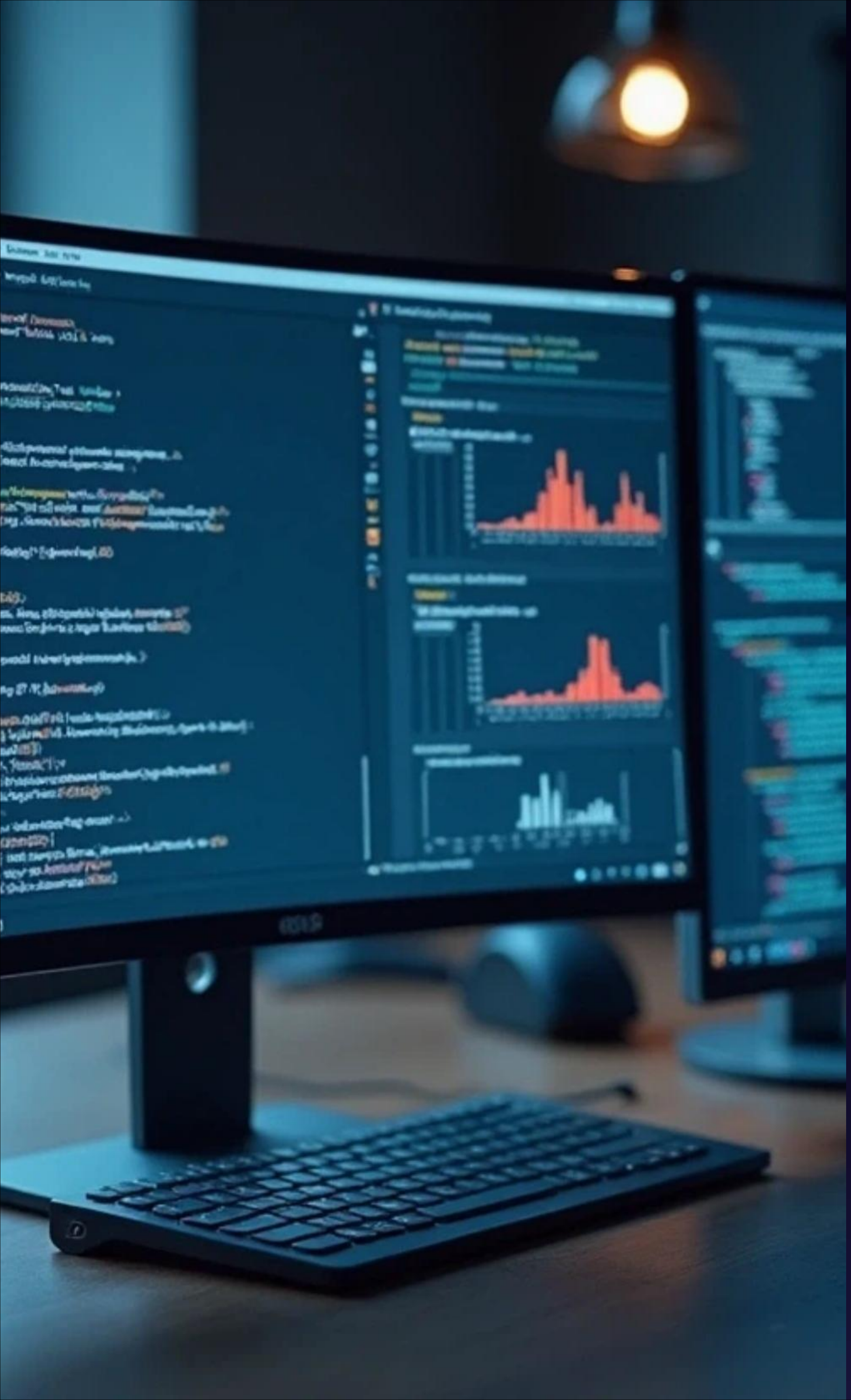
Real Estate Price Prediction — Data Preprocessing & Feature Engineering

Created by / *Mohamed Waleed Elmasry*

CONTENTS



1. Import Libraries

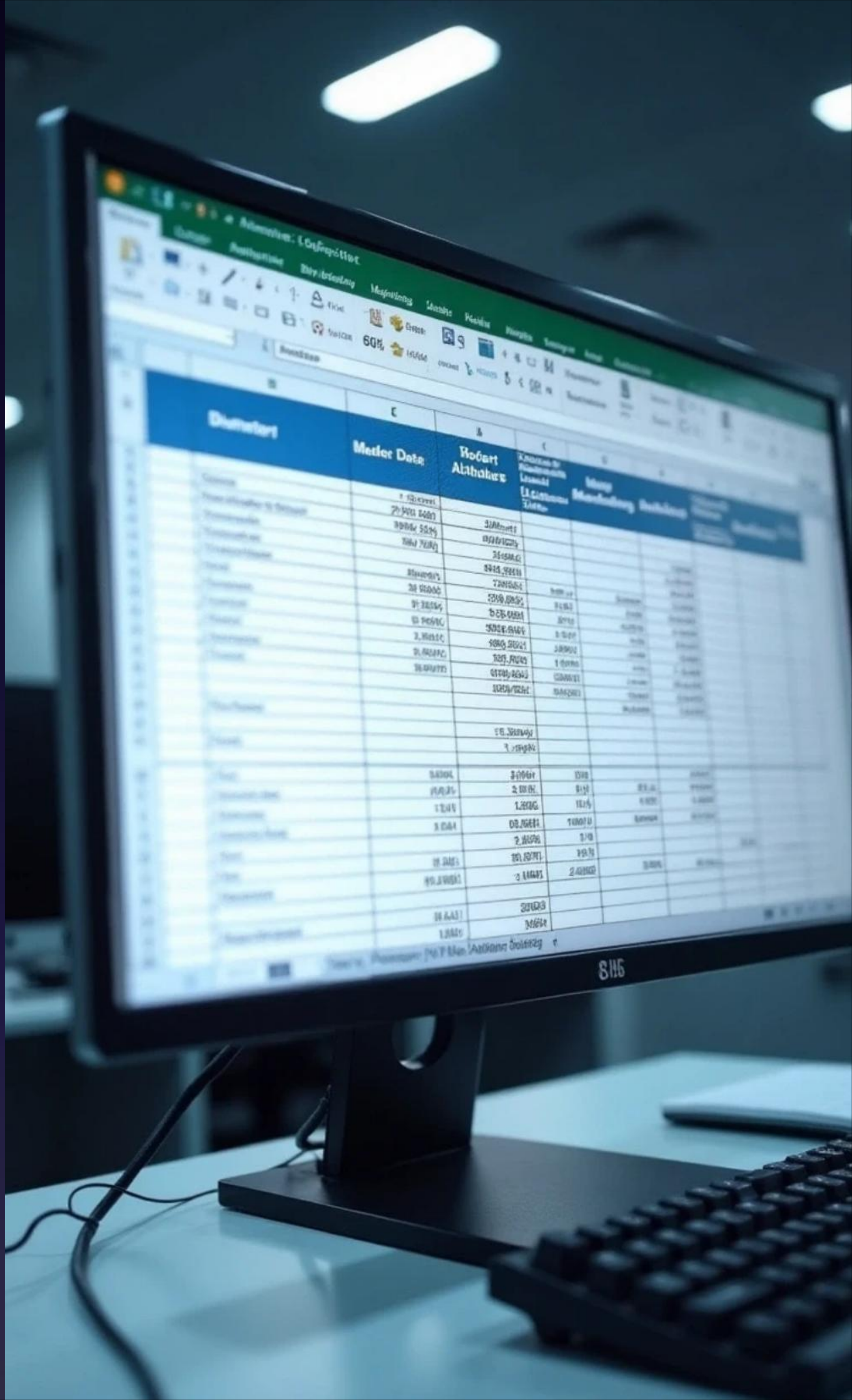
Main Libraries

- NumPy: Essential for numerical calculations.
- Pandas: Key for data handling and analysis.
- Matplotlib: Ideal for generating visual charts.
- Seaborn: Specialized in statistical data visuals.
- category_encoders: Designed for converting categorical variables.
- Scikit-learn: Essential tools for model selection and preparation.

2. Loading the Dataset

Dataset details

- The dataset is loaded from an Excel file containing various property attributes.



3. Renaming Columns

Purpose

Standardize column names for easy data manipulation.

Column names were standardized to English for consistency and easier processing.

id	price	type	n_bedrooms	n_bathrooms	area	sub_location	location
1	9500000	شقة	3	2	140	أخرى مناطق	الجيزة
2	5800000	شقة	3	2	145	أكتوبر 6	الجيزة
3	15150000	فيلا	3	3	225	أكتوبر 6	الجيزة
4	11200000	شقة	3	2	150	أكتوبر 6	الجيزة
5	21918000	فيلا	4	4	265	الخامس التجمع	الجديد القاهرةm

4. Category Value Encoding

1

Convert category values to English for easier reading.

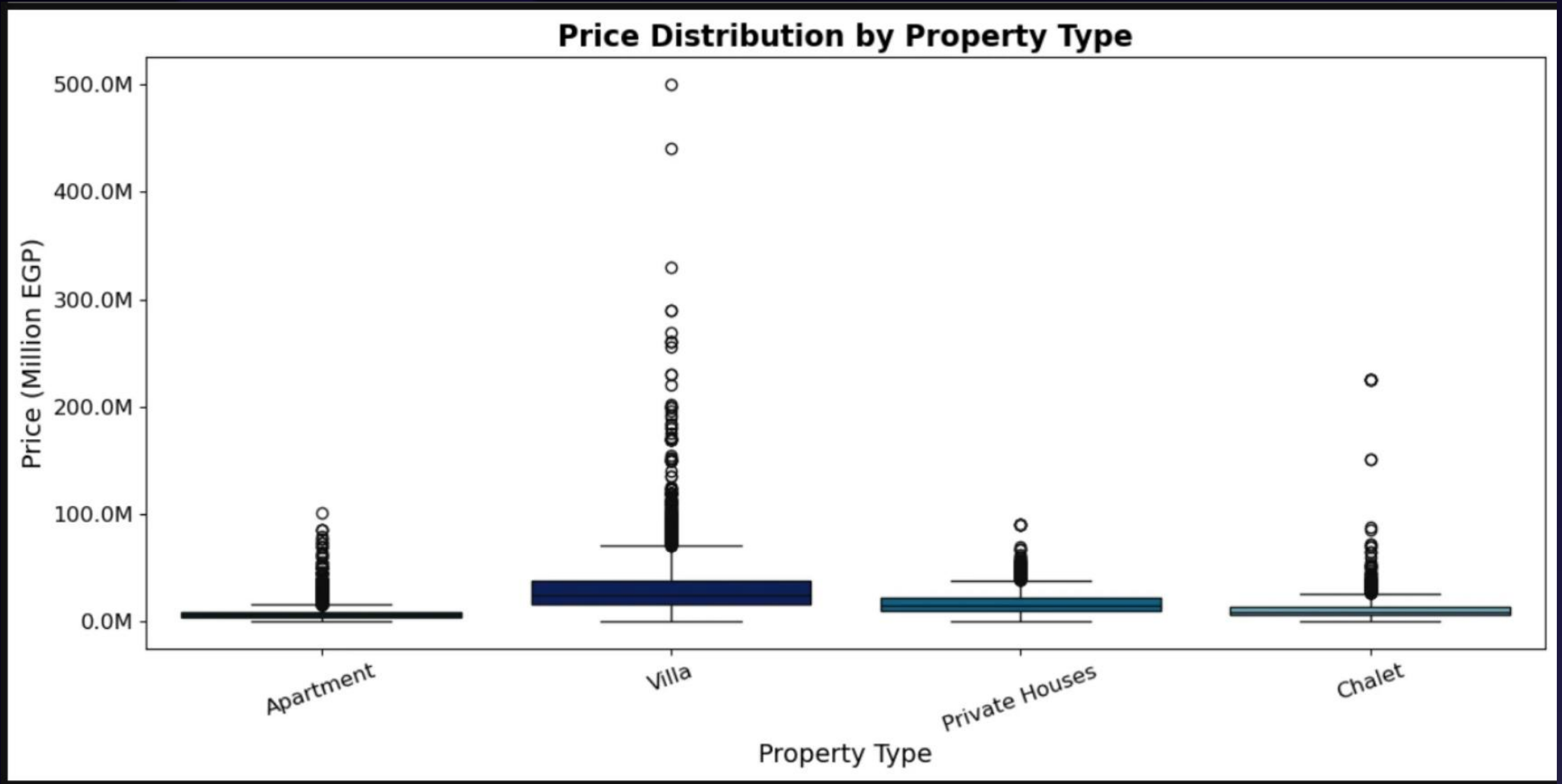
2

Encoding strategy: Replace Arabic property types with their English counterparts.

Number	Price	Type	Number of Bedrooms	Number of Bathrooms	Area	Sub-location	Location
1	9500000	Apartment	3	2	140	Other Areas	Giza
2	5800000	Apartment	3	2	145	October 6	Giza
3	15150000	Villa	3	3	225	October 6	Giza
4	11200000	Apartment	3	2	150	October 6	Giza
5	21918000	Villa	4	4	265	Fifth Residential Area	New Cairo

5. Data Analysis & Visualization

Visualizing Price Distribution by Property Type



numerous outliers

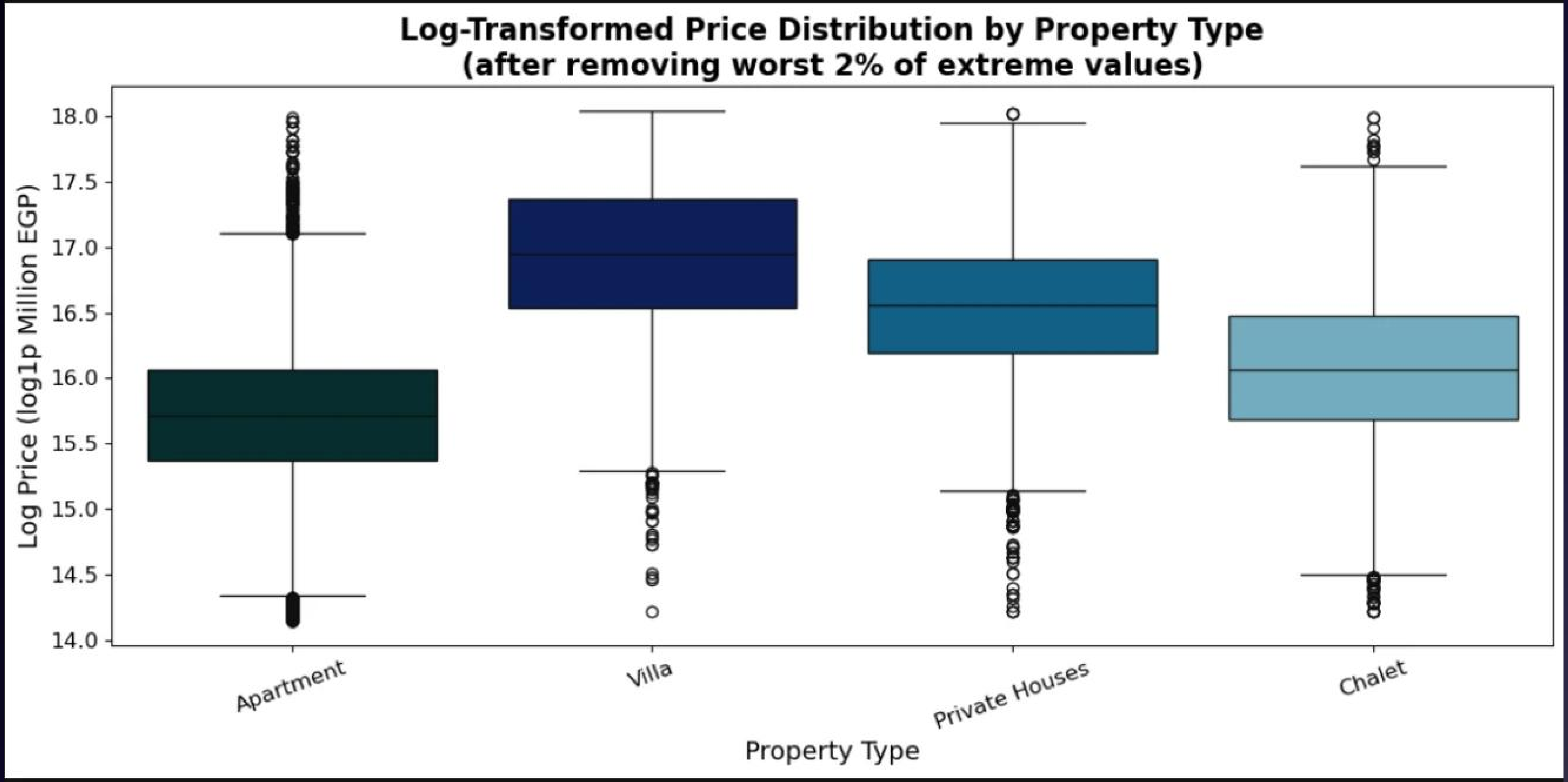
6. Outlier Removal & Log Transformation — Price

Outlier Removal

Remove the top and bottom 1% of extreme values.

Log Transformation

Apply log transformation to the price.

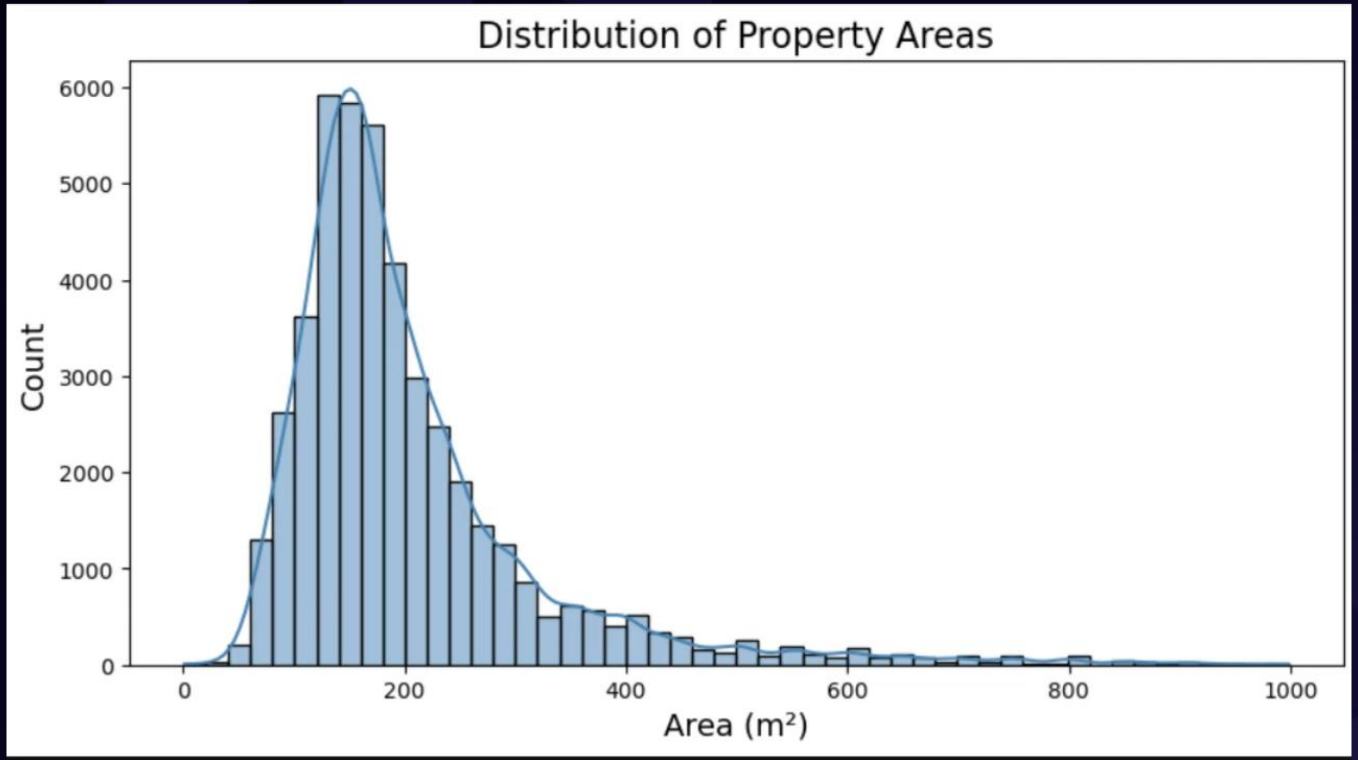


7. Outlier Removal & Log Transformation — Area

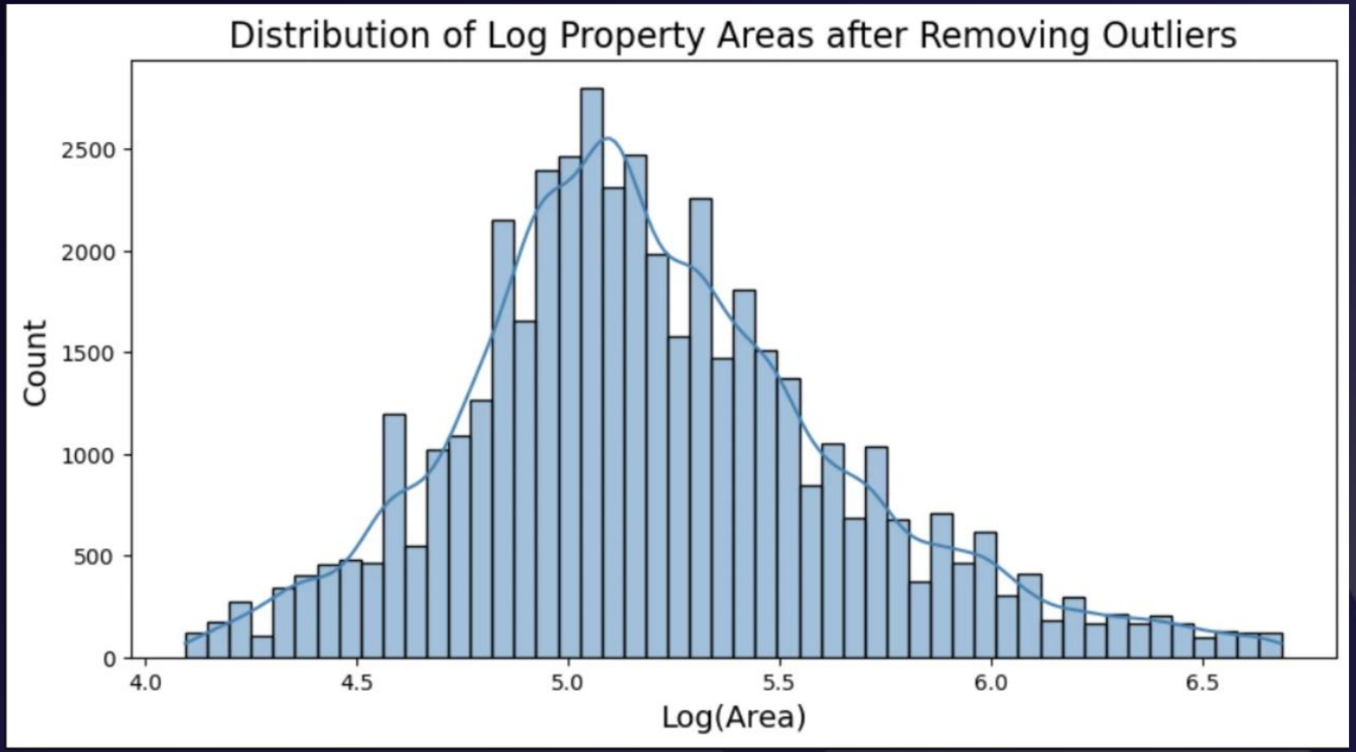
Outlier Removal&Transformation - Area

Extreme values in the area column were removed using quantile filtering (top 0.5% and bottom 0.5%).

Logarithmic transformation was applied to area (and previously to price) to reduce skewness and stabilize variance.



Before Transformation



After Transformation

8. Correlation Analysis

Visualizing Correlation with Price and Log Price

🔍Insights from Correlation Analysis

Log_price offers greater stability as a target variable compared to raw price — the log transformation enhances linearity and minimizes the impact of outliers.

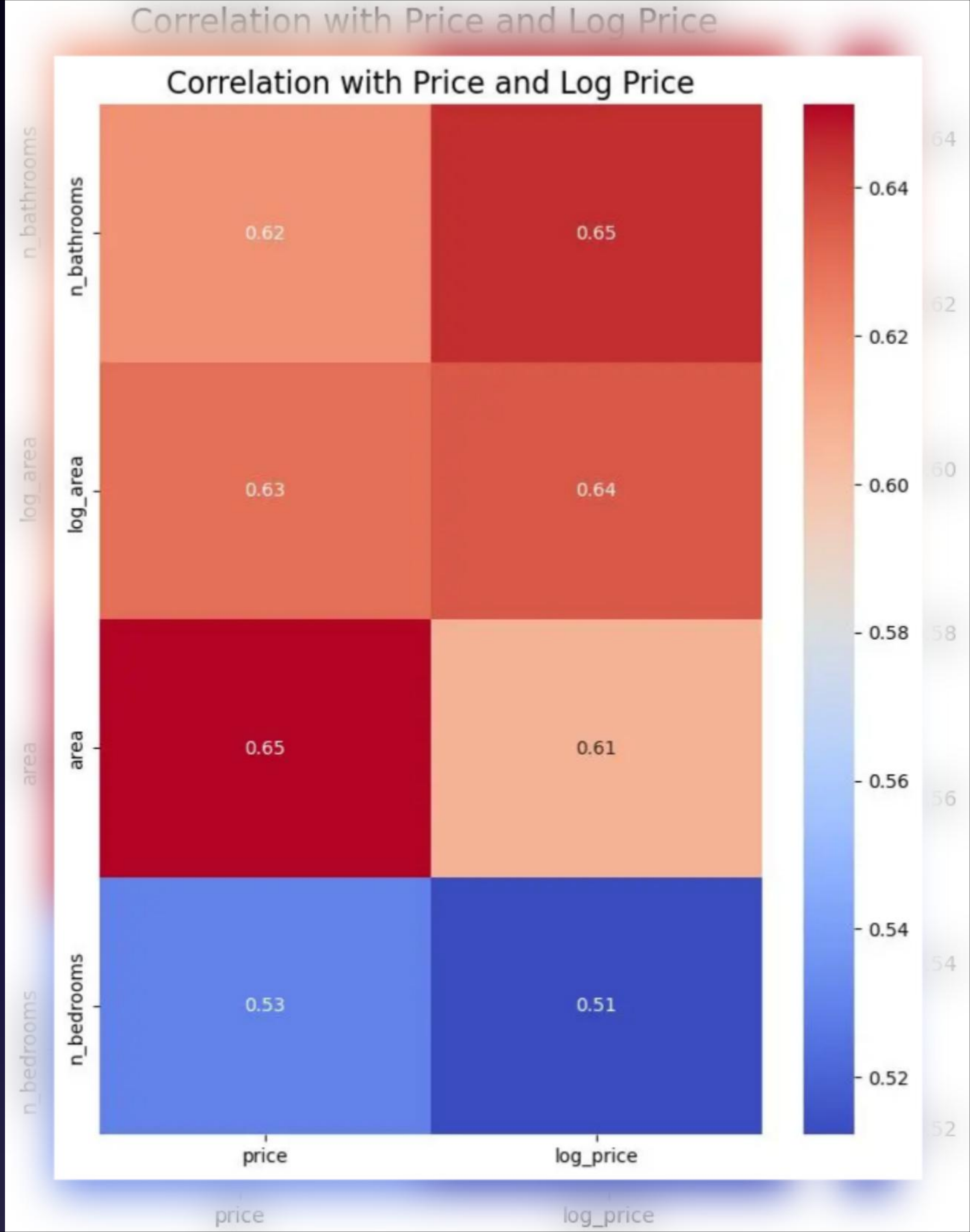
The variable n_bedrooms exhibits the lowest correlation (~0.51) with both price and log_price.

Log_area demonstrates a robust and consistent correlation with log_price (0.64), validating that log transformation mitigates skewness.

The variable area has a stronger correlation with raw price (0.65) compared to a slightly lower correlation with log_price (0.61).

N_bathrooms shows a higher correlation with log_price (0.65) than with raw price (0.62).

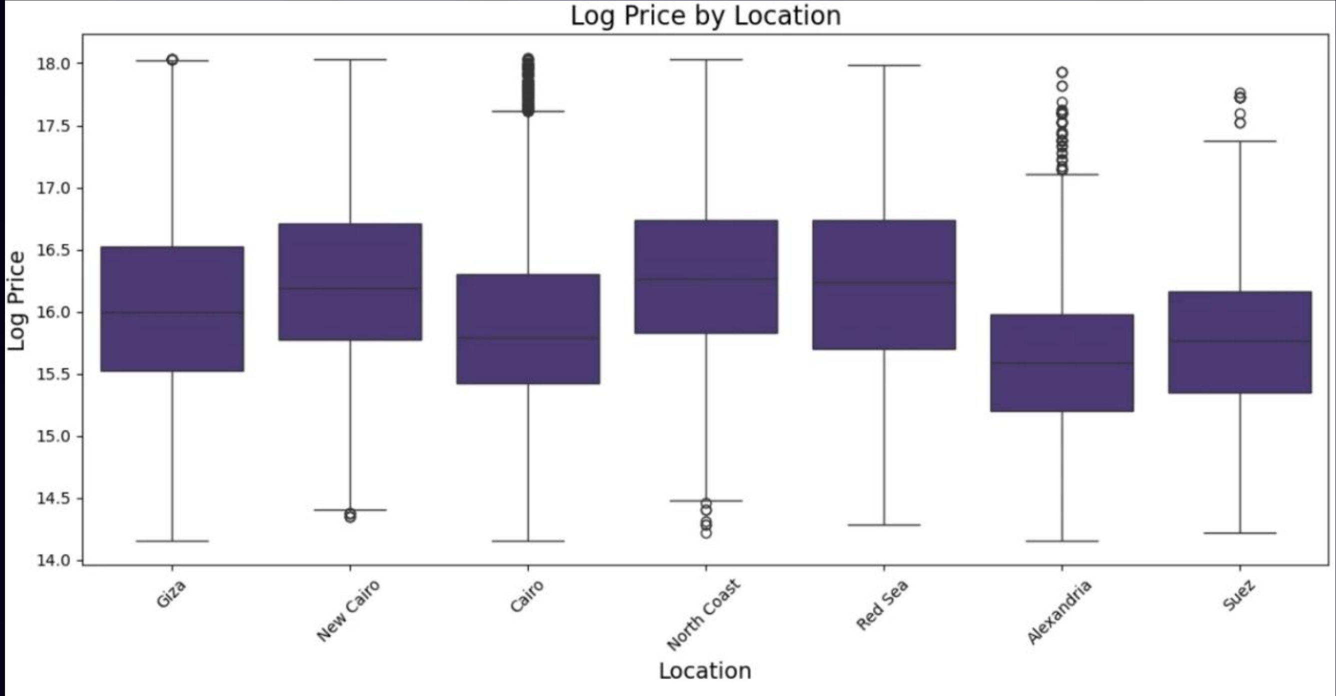
✅ Final Conclusion:
Log transformation boosts the stability of the model while strengthening the connection between essential numeric features and the target variable.

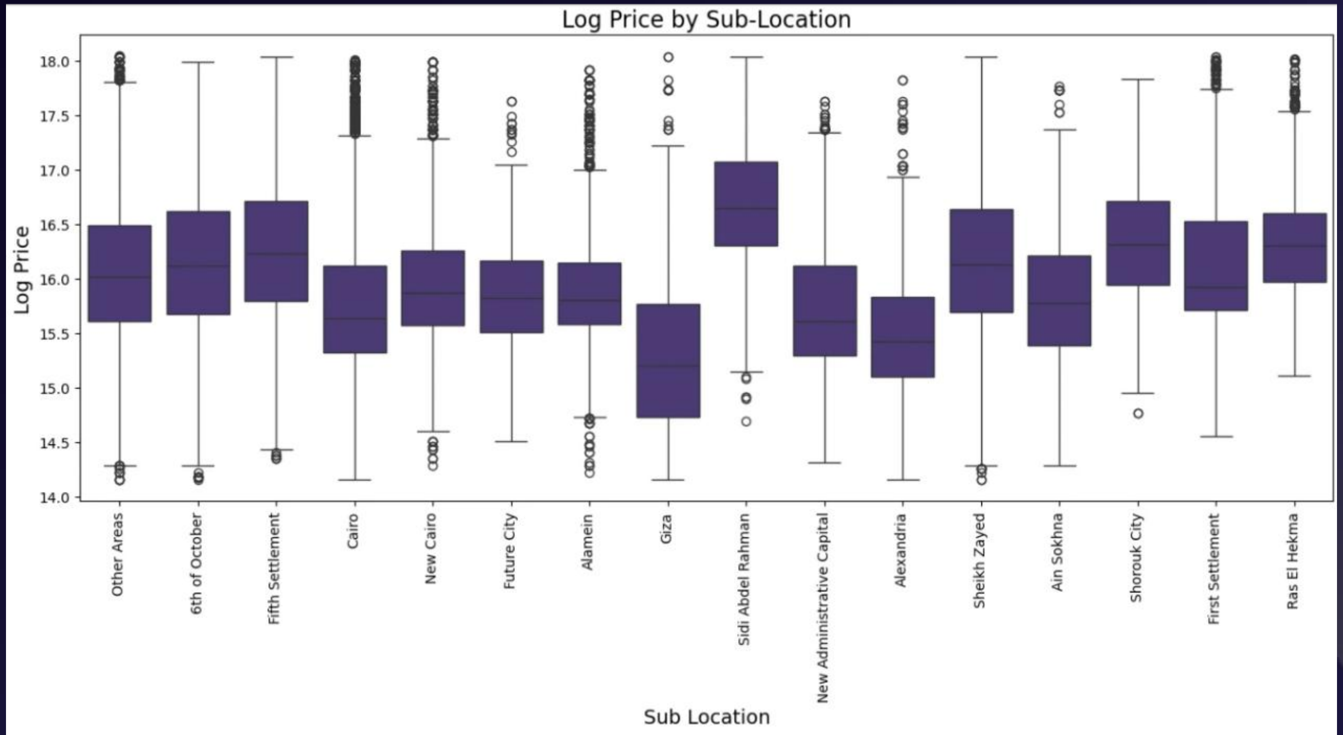


9. Aggregated Analysis by Property Type & Location

Analysis

Calculate average and median log prices by property type and location.

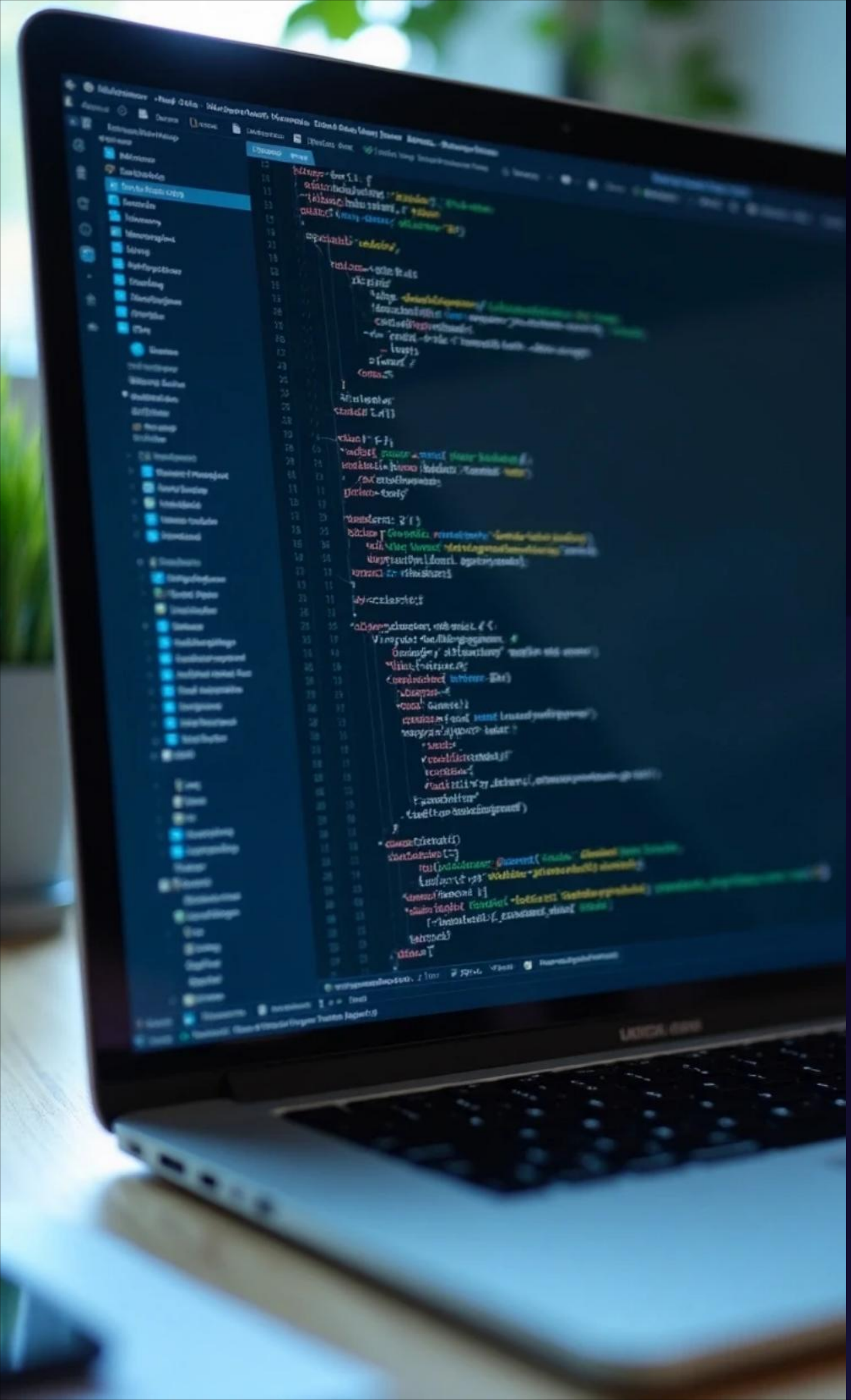




There is a clear imbalance in the dataset: most listings are Apartments in Cairo.

Both property type and location are strong predictors of housing prices.

There is a clear imbalance in the datase



10. Feature Engineering & Data Export

Additional Feature Creation and Saving Process

Data Preparation Overview

1 **Dataset Copy**

Create a copy to maintain the original data.

2 **Drop Area Column**

Remove the area column for log_area usage.

3 **Define Target & Features**

Set log_price as target and define feature matrix X.

4 **Train/Test Split**

Split data with 80% training and 20% testing.

5 **Ensure Reproducibility**

Utilize fixed random state for shuffling.

Target Encoding

Purpose

Prepares categorical features to improve model accuracy.

Technique

Uses target averages in a cross-validation setup.

Features

Applies to location, type, and sub_location.

Cross-Validation

Utilizes 6-fold to prevent data leakage.

Benefits

Captures target relationships while ensuring generalization.

DATA After Encoding

type	n_bedrooms	n_bathrooms	log_area	sub_location	location
15.724519	3	3	5.298317	16.279147	16.258963
15.723623	3	2	5.135798	16.273965	16.256207
15.723623	3	3	5.135798	16.342638	15.902682
16.927353	4	4	6.152733	16.139945	16.258745
15.72204	3	1	5.010635	15.809305	15.907546

Adding

Adding New Features Overview

1. **Bathroom-to-Bedroom Ratio**
   Indicates comfort level by comparing bathrooms to bedrooms.

2. **Total Number of Rooms**
   Sum of bedrooms and bathrooms for overall property size.

3. **Area per Room**
   Average space available for each room shows spaciousness.

4. **Area per Bedroom**
   Total area dedicated to bedrooms indicating living quality.

5. **Location–Area Interaction**
   Engages location data to reflect size impact on property price.

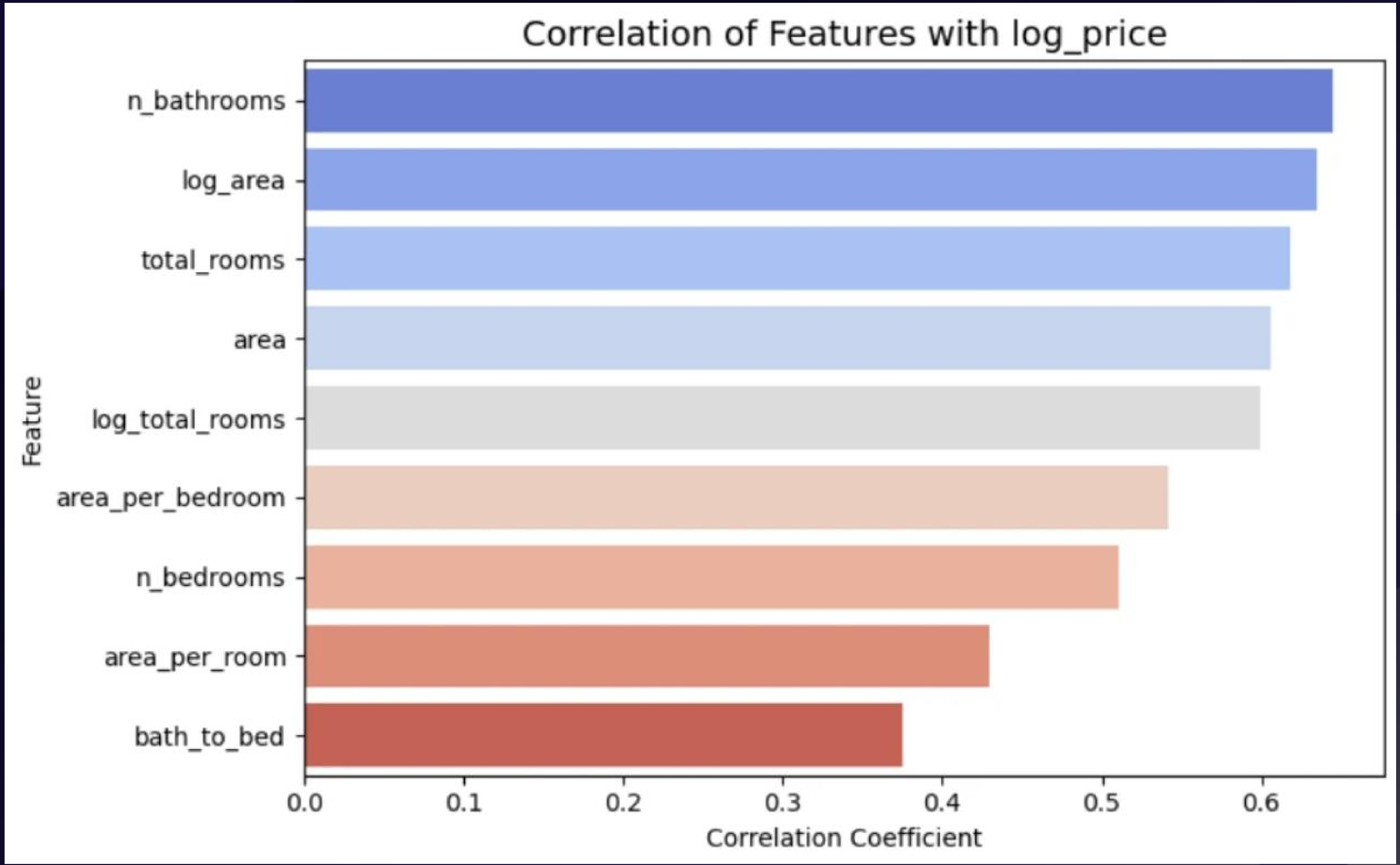
6. **Location–Sub-Location Interaction**
   Captures the hierarchical relationship between a property's main location and its sub-location.

7. **Log of Total Rooms**
   Applies a logarithmic transformation to the total number of rooms.

type	n_bedrooms	n_bathrooms	log_area	sub_location	location	bath_to_bed	total_rooms	area	area_per_room	area_per_bedroom	loc_area_interaction	loc_cross	log_total_rooms
15.724519	3	3	5.298317	16.279147	16.258963	1.000000	6	200.0	28.571429	50.0	86.145144	264.682048	1.945910
15.723623	3	2	5.135798	16.273965	16.256207	0.666666	5	170.0	28.333333	42.5	83.488601	264.552934	1.791759
15.723623	3	3	5.135798	16.342638	15.902682	1.000000	6	170.0	24.285714	42.5	81.672971	259.891776	1.945910
16.927353	4	4	6.152733	16.139945	16.258745	1.000000	8	470.0	52.222222	94.0	100.03571	262.415244	2.197225
15.72204	3	1	5.010635	15.809305	15.907546	0.333333	4	150.0	30.000000	37.5	79.706912	251.487241	1.609438

Objective

These engineered features enhance the dataset by offering structural, spatial, and relational insights.
Collectively, they empower the regression model to better capture non-linear relationships and boost overall predictive accuracy.

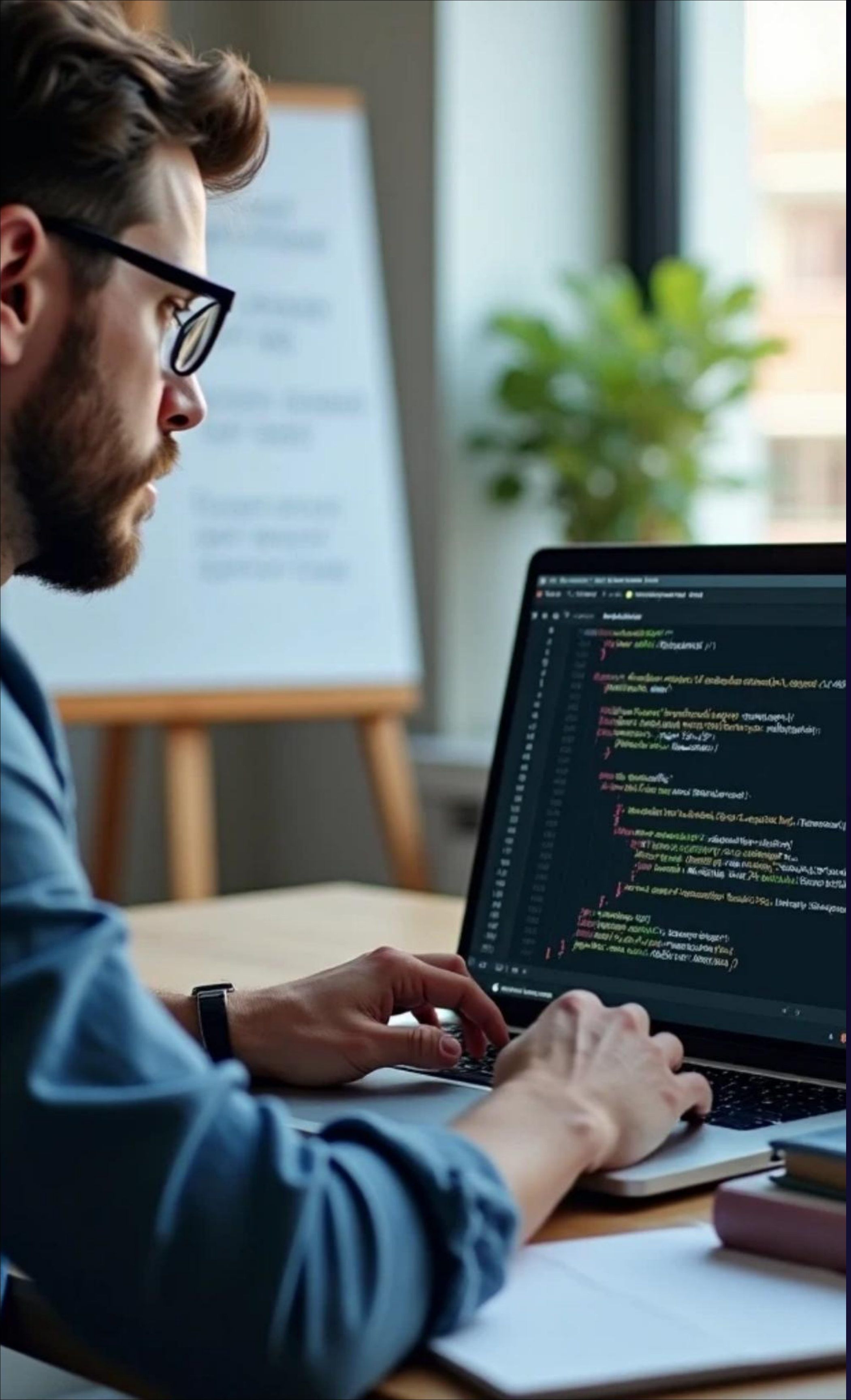


11. Features Summary

1

Summary of key features for the final model inputs, showcasing their intended purpose and insights derived from them.

Feature Name	Type	Source	Description	Purpose / Insight
type	Encoded (Target)	Original	Target-encoded representation of the property type.	Captures architectural and functional differences between property types.
n_bedrooms	Numeric	Original	Number of bedrooms in the property.	Direct measure of property size; strongly influences price.
n_bathrooms	Numeric	Original	Number of bathrooms in the property.	Indicates comfort level and modern facilities.
log_area	Numeric (log-transformed)	Original (Transformed)	Logarithm of total built-up area.	Stabilizes skewed data; models proportional area effects.
location	Encoded (Target)	Original	Target-encoded main city or district.	Captures macro-level price variation due to geography.
sub_location	Encoded (Target)	Original	Target-encoded sub-region or neighborhood.	Adds micro-level geographical differentiation.
bath_to_bed	Numeric (Derived)	Engineered	Bathrooms-to-bedrooms ratio.	Reflects property luxury and comfort balance.
total_rooms	Numeric (Derived)	Engineered	Total count of rooms (bed + bath).	Simple proxy for overall property size.
log_total_rooms	Numeric (Transformed)	Engineered	Logarithm of total room count.	Reduces impact of large room counts on model stability.
area	Numeric (Derived)	Engineered	Actual property area (exponentiated from log_area).	Represents physical space and scale.
area_per_room	Numeric (Derived)	Engineered	Average area per room.	Indicates spaciousness and internal density.
area_per_bedroom	Numeric (Derived)	Engineered	Area divided by number of bedrooms.	Captures the space dedicated to private use.
loc_area_interaction	Numeric (Interaction)	Engineered	Interaction between log_area and location.	Models how the impact of area differs by location.
loc_cross	Numeric (Interaction)	Engineered	Product of sub_location and location.	Captures cross-geographic effects within regions.
log_price	Numeric (Target Variable)	Original (Target)	Logarithm of property price	Target variable; improves learning on wide price



Saving Train & Test Data

1

Prepare CSV Files

Create separate CSV files for training and testing data.

2

Save Encoded Data

Export the processed encoded data to CSV format.

3

Data Processing Summary

Finalizes and saves the processed data sets for model training.

Thank you!

Questions or feedback? Feel free to ask!

e-mail : mwezzat16@gmail.com

Thank You