EXPLORATORY DATA ANALYSIS (EDA) REPORT

1. Introduction

This Exploratory Data Analysis (EDA) aims to understand pricing patterns, property characteristics, location-based trends, and investment suitability in the Indian real estate market. The dataset contains **250,000 property records** across **20 states**, **42 cities**, and **500 localities**, with detailed attributes covering pricing, size, amenities, building age, infrastructure accessibility, owner type, and more.

The objective is to produce insights supporting:

- Investment classification
- Price prediction
- Market behaviour understanding
- Feature engineering for ML models

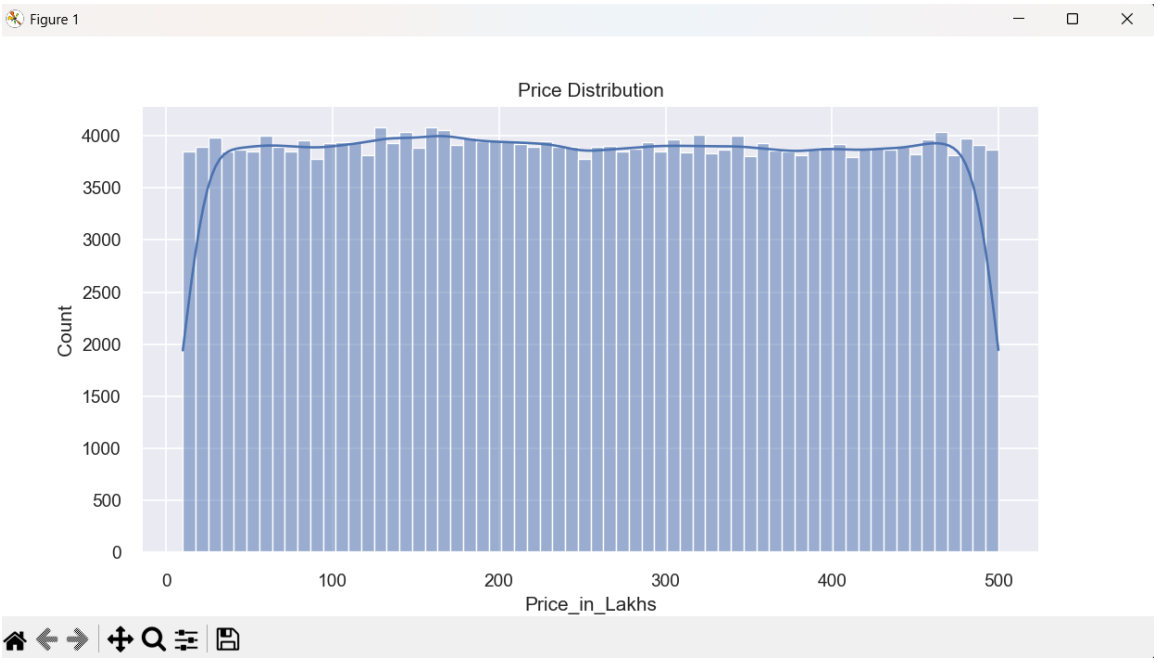
2. Dataset Overview

- **Total Rows:** 250,000
- **Total Columns:** 28
- **Missing Values:** 0
- **Duplicate Rows:** 0
- **Data Quality:** Clean and consistent
- **Coverage:** National (20 states, 42 cities)

The dataset includes numerical, categorical, and engineered features such as Price_per_SqFt, Build_Decade, Size_Category, and Price_Category.

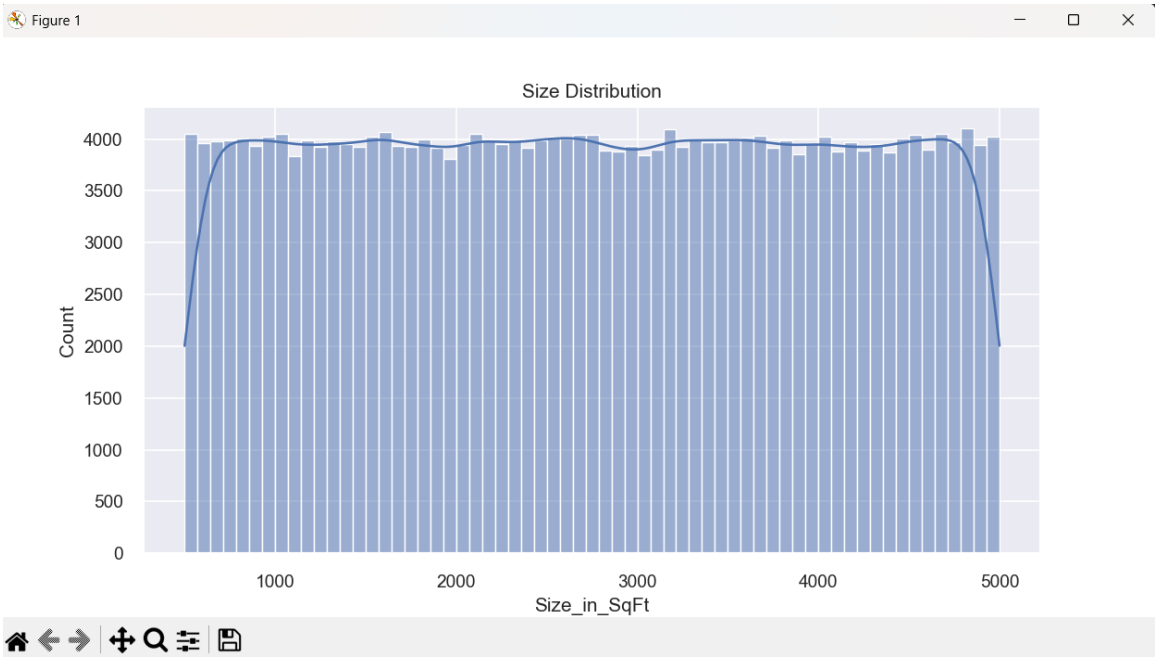
3. Univariate Analysis

3.1 Price Distribution



- **What the plot shows:** The price histogram/density is broadly distributed across the full range (low to very high). Your plotted KDE shows reasonably even counts across buckets with slight concentration in the mid-range (100–350 Lakh).
- **Business interpretation:** A wide price spread indicates the dataset covers both affordable and luxury inventory — helpful for models meant to generalize across market segments. For investors, the mid-range mass is likely the most liquid segment; luxury pockets (right tail) are niche and require separate treatment for ROI expectations.
- **ML & feature actions:**
- Model target (Price_in_Lakhs) will need log-transform experiments to stabilize variance if you train models sensitive to heteroskedasticity (e.g., linear models).
- Create price-segment categorical variable (Low / Mid-Low / Mid-High / High) for classification models or stratified sampling during cross-validation.
- Keep PPS as a normalized target feature — it better isolates locality effects.

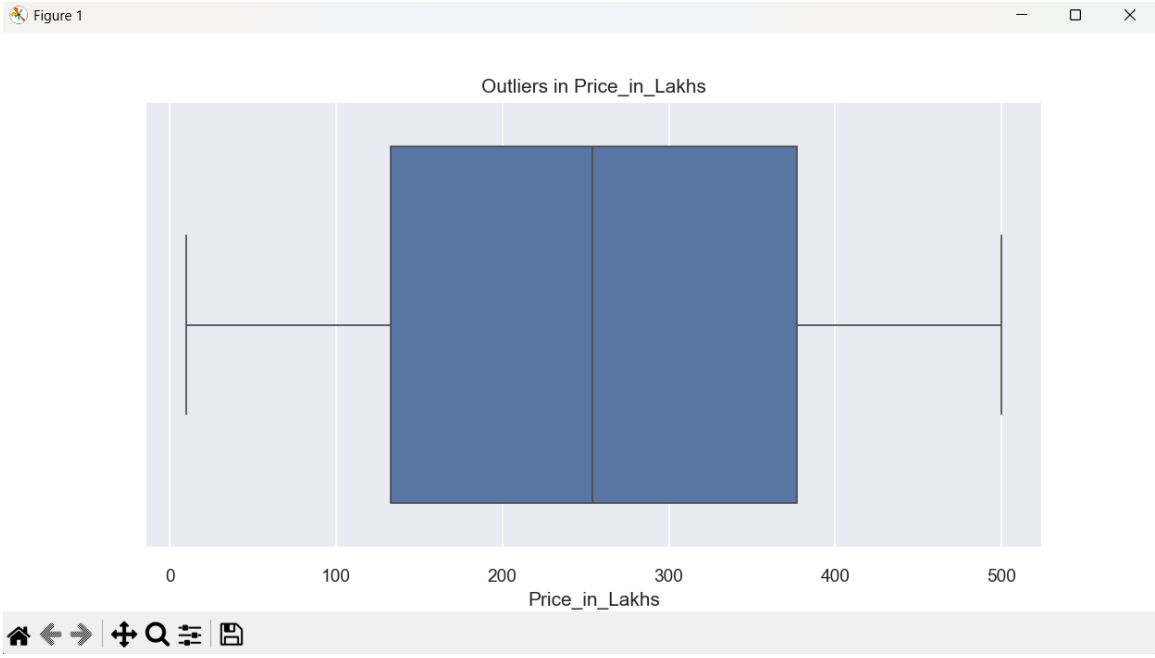
3.2 Size Distribution



- **What the plot shows:** Size values are well-distributed from small apartments to very large homes. No extreme skew; mass in medium and large bands.
- **Business interpretation:** Balanced sample across sizes means model will learn relationships across home types (studio → villa) without overfitting a single size group. In markets, size alone does not determine price — but it is a necessary explanatory feature.
- **ML & feature actions:**
- Engineer `size_per_bhk = Size_in_SqFt / BHK` to capture space per room; this helps differentiate cramped vs. spacious units.
- Consider interactions between Size and Locality (Size × Locality median PPS) because a 1000 sqft property in a premium locality behaves differently from 1000 sqft in a lower-cost city.

4. Outlier Analysis

4.1 Outliers in Price



What the Plot Shows

- The boxplot for property price shows:
- A relatively **wide spread** across the central range (Q1–Q3).
- A handful of **upper-end outliers** representing very expensive properties.
- Lower-end values appear close to the minimum, with no extreme low-value outliers.
- The overall distribution demonstrates a healthy market range, where most values lie within predictable intervals, and only a small fraction represent exceptionally premium listings.

- ### Business Interpretation

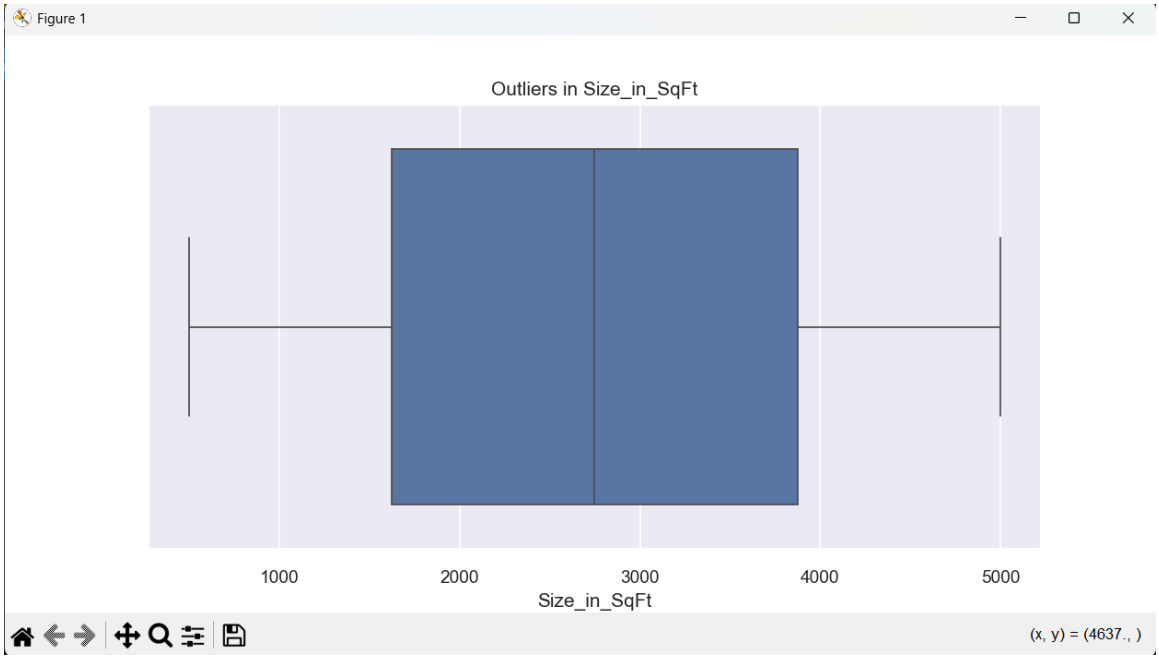
- High-price outliers typically correspond to:
- Luxury developments
- Central business district locations
- High-demand micro-markets
- Properties with exceptional amenities or large built-up areas

- These high-value properties can distort average market statistics, especially in cities where pricing varies significantly by locality.
  From a buyer or investor perspective, these represent **niche, low-volume but high-margin segments** of the real estate market.

- ### ML & Feature Engineering Actions

- Consider **log-transforming Price** to reduce the influence of very high-priced units.
- Cap outliers (winsorization) for linear models prone to instability.
- Tree-based models (XGBoost, RandomForest) often handle these values naturally, but performance monitoring is needed.

4.2 Outliers in Size



What the Plot Shows

- The size boxplot reveals:
- Moderate spread in property sizes.
- Larger upper-end outliers for very spacious properties (>4000 SqFt).
- Very few small-size outliers (e.g., micro-apartments).
- Overall, the distribution mirrors normal market behavior: the majority of properties fall within a realistic range, while a minority represent unusually large luxury or villa-type units.

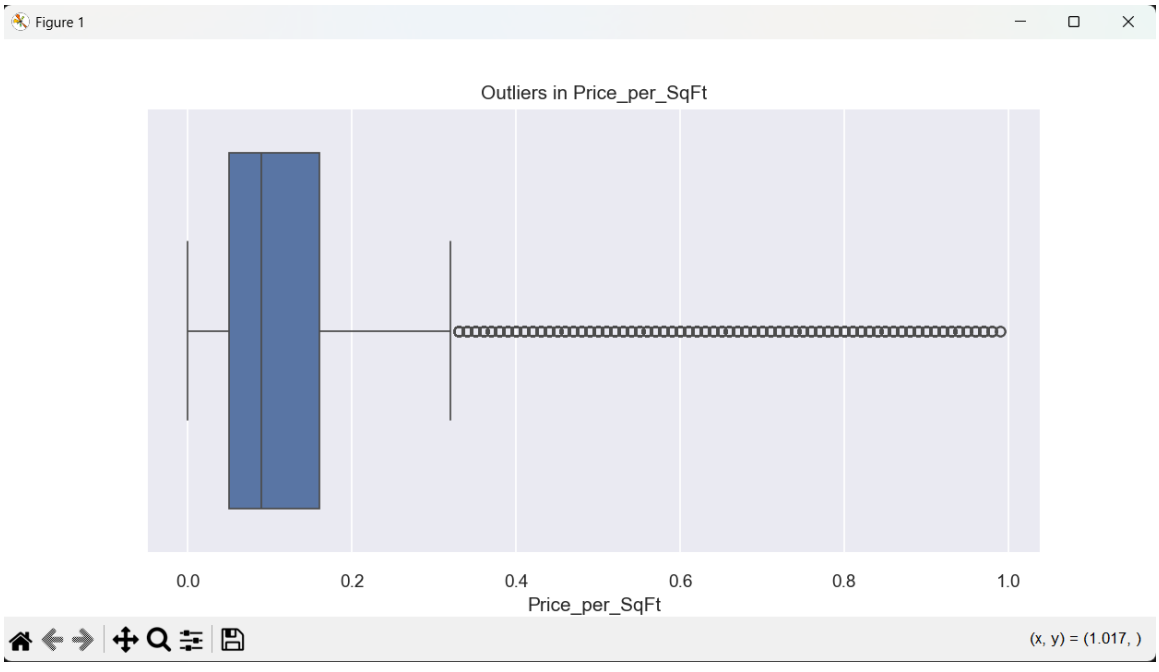
- ## Business Interpretation

- Large-size outliers often correspond to:
- Villas
- Luxury penthouses
- Combined apartments or extended units
- High-end bungalow-style homes
- Such units can command special pricing models and may follow different appreciation patterns compared to standard-sized properties.
- This highlights the importance of **segment-specific price modeling** rather than applying a uniform valuation rule across all sizes.

- ## ML & Feature Engineering Actions

- Use `Size_Category` (Small, Medium, Large, Very Large) to reduce sensitivity to extreme square-footage values.
- Consider interaction features: `Size_in_SqFt × PPS` to capture the non-linear size-premium relationship.
- Cap extremely large sizes or treat them as a separate luxury segment during training.

4.3 Outliers in Price_per_SqFt



What the Plot Shows

- The PPS boxplot displays:
- **Significant number of high-end outliers**
- Noticeably more PPS outliers than either Price or Size
- The upper tail extends much further, indicating luxury locality premiums or pricing anomalies
- Middle distribution is tight, showing PPS is stable for most properties
- This tells us PPS is the **most volatile metric** in the dataset.

• Business Interpretation

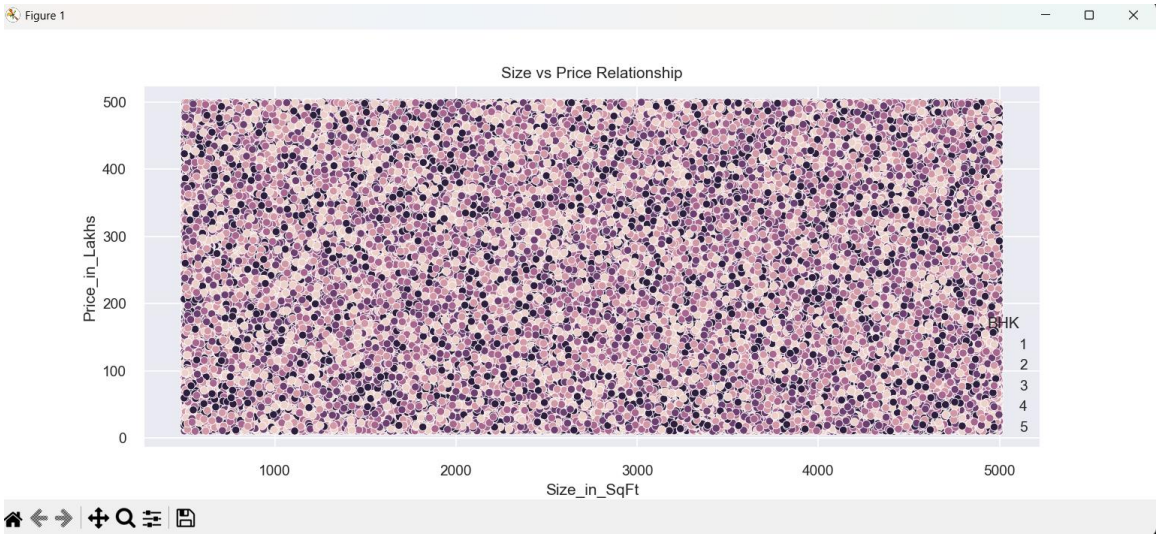
- High PPS outliers commonly represent:
- Elite neighborhoods
- Prime city zones (e.g., city centers, coastal areas)
- High-end gated communities
- Newly developed luxury projects with premium amenities
- Scarce land pockets where unit rates skyrocket despite modest sizes
- These properties command high price-per-unit area due to **location-driven value**, not size or features alone.
  In real estate markets, PPS is the **strongest localized price signal**, and its outliers often separate premium markets from general markets.

• ML & Feature Engineering Actions

- This is the most ML-critical outlier:
- Add `PPS_outlier_flag = True/False` (as your dataset already includes).
- For some models, **cap PPS values** at a high percentile (95th or 99th) to avoid distortion.
- Use **RobustScaler** or **QuantileTransformer** if you plan linear/regression models.
- Consider building a **separate luxury model** or including an `is_luxury_locality` feature.
- For regression tasks, evaluate metrics that handle outliers better:
  - MAE
  - Huber loss
  - MAPE (with caution)
-

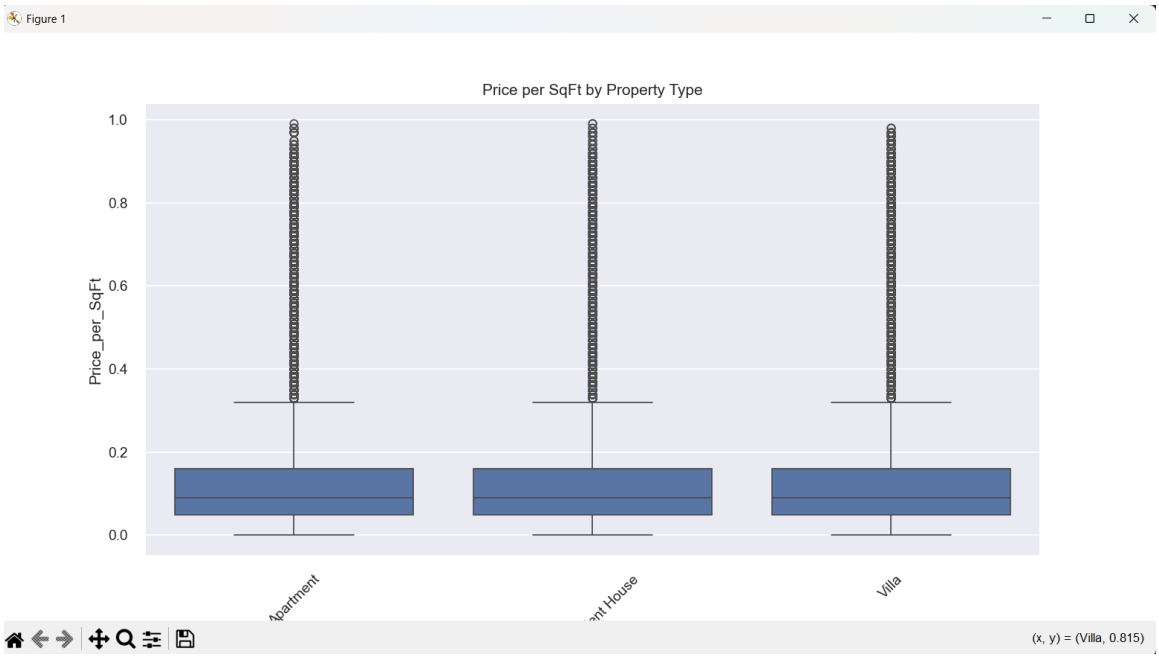
5. Bivariate & Multivariate Analysis

5.1 Size vs Price Relationship



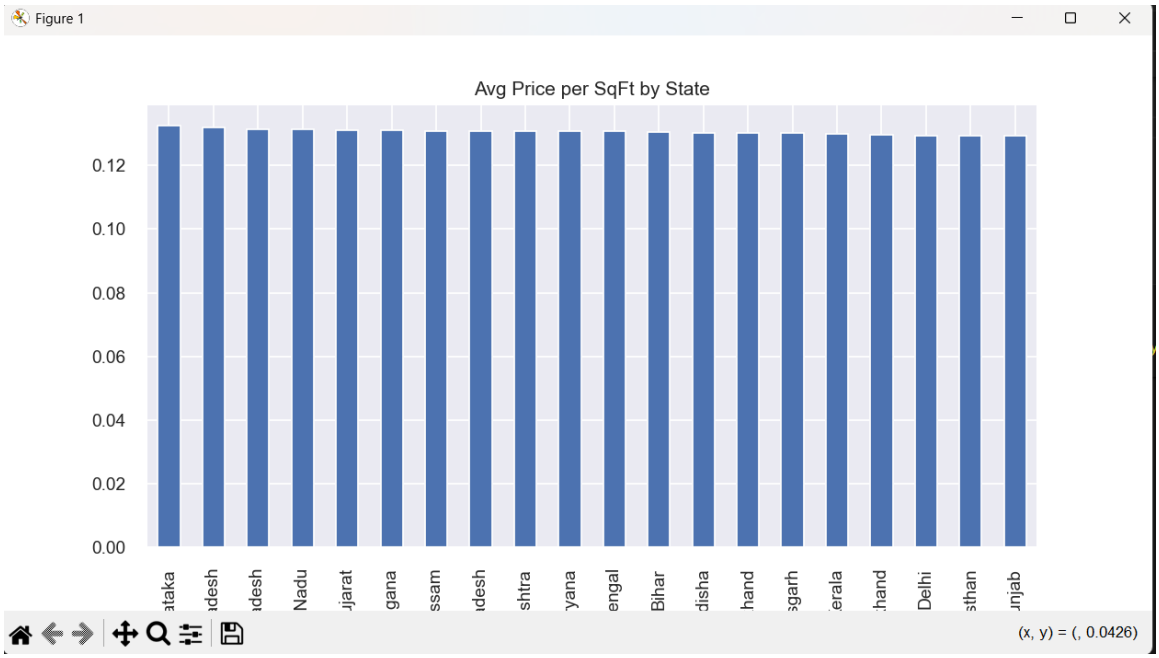
- **What the plot shows:** Positive trend with broad scatter. Size loosely predicts price, but many vertical spreads at given sizes show location-driven variance.
- **Business interpretation:** Size explains part of the price but locality and PPS (quality of neighborhood) explain large residuals; a 2000 sqft property can be cheap or expensive depending on locality and amenities.
- **ML & feature actions:**
- Include locality-level aggregations: `median_price_locality`, `median_pps_locality`, `locality_rank`. These reduce variance and capture micro-market effects.
- Fit non-linear models (tree-based or GBM) to capture interactions and heteroskedasticity.
- Add `Size_Category` and `Size_per_BHK` features for non-linear effects.

5.2 Price per SqFt by Property Type



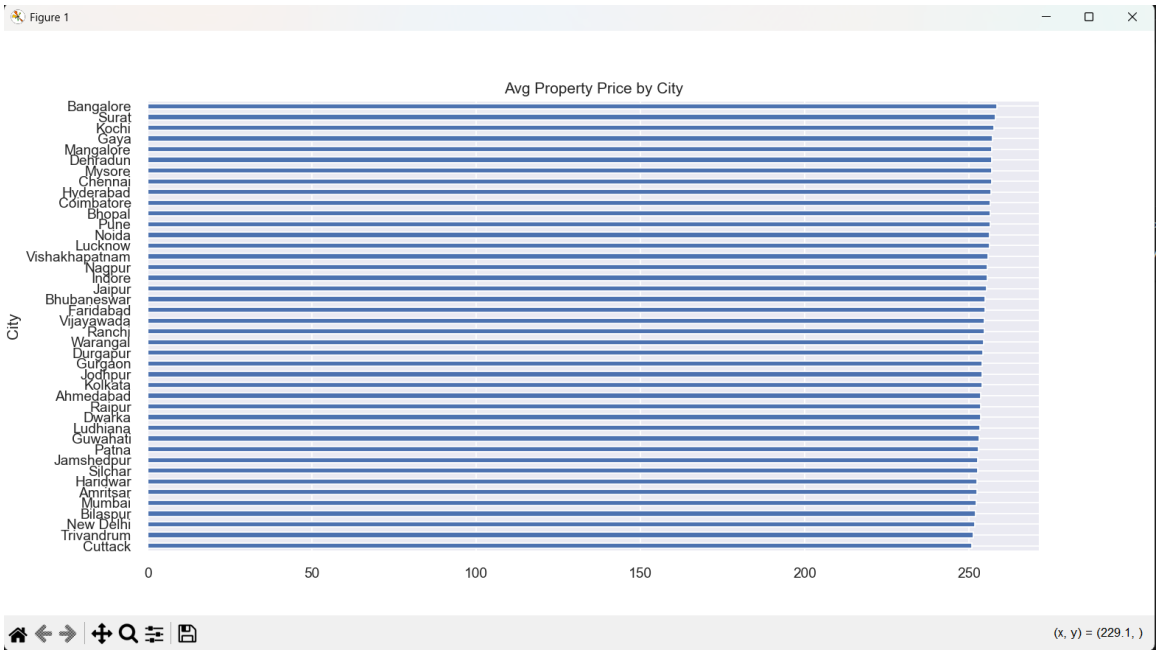
- **What the plot shows:** Boxplots show PPS medians are similar across types but villas have higher upper tails and more variability; apartments are more tightly distributed.
- **Business interpretation:** Villas command premium PPS at the top end — likely due to scarcity and plot ownership. Apartments show more uniform pricing, which is expected because apartment pricing is often regulated by project-level pricing and shared amenities.
- **ML & feature actions:**
- Use `Property_Type` as an important categorical predictor; include both main effect and interaction with PPS and Amenities.
- For villas, include `Plot_Size` (if available) or `Very_Large` flag to capture the premium effect.
- Consider training separate models per Property_Type if predictive performance diverges (a single global model might under/overfit extreme villa prices).

5.3 Average Price per SqFt by State



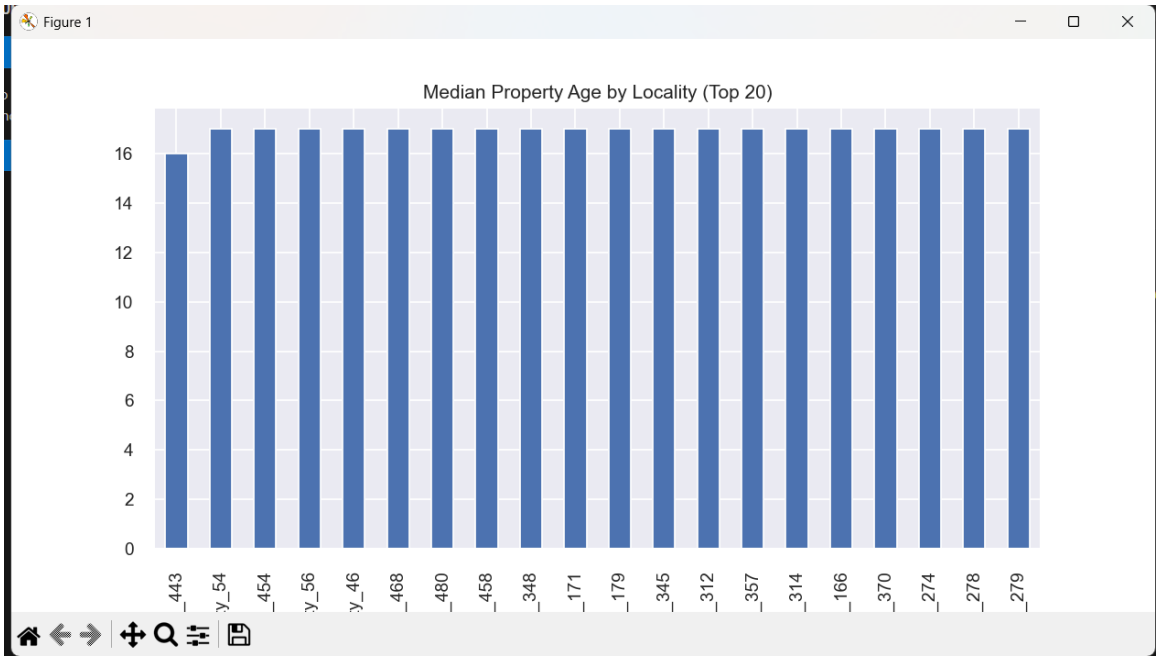
- **What the plot shows:** State-level PPS appears numerically similar across states in the sample — a relatively flat bar chart.
- **Business interpretation:** The flatness implies either the dataset is normalized across states or the sample does not emphasize metro vs. non-metro variance. Real markets usually show larger state-level variation (Mumbai vs. smaller cities). Treat state as contextual rather than decisive.
- **ML & feature actions:**
- Keep `State` and `City` encoded (one-hot for top N, target-encoding or embeddings for others).
- Add `state_median_pps` derived from the data; it can be useful where locality-level data is sparse.

5.4 Average Property Price by City



- **What the plot shows:** City-level prices show moderate differences; some metro cities (Bangalore, Pune, Chennai etc.) are higher.
- **Business interpretation:** City-level differences are meaningful for ROI and localized recommendations; investors often prioritize city-level growth rates and rental yields.
- **ML & feature actions:**
- Use target encoding for City with smoothing to prevent leakage.
- Use city-level macro features (GDP proxy, population if obtainable) to improve generalization.

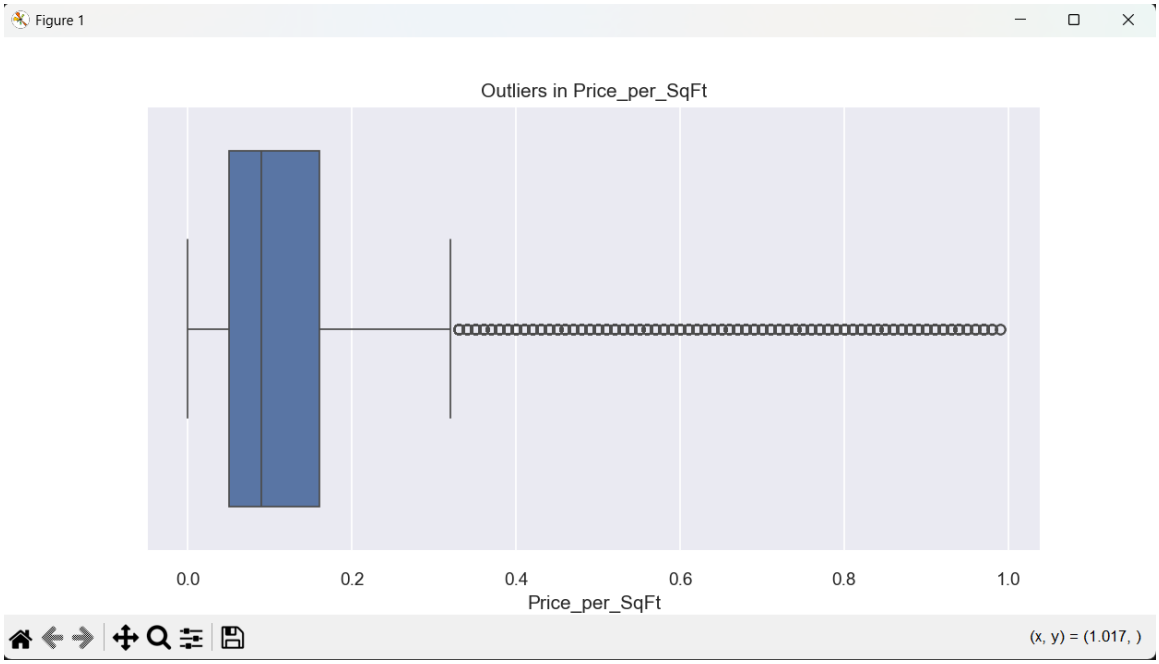
5.5 Median Property Age by Locality



- This visualization presents the **median Age_of_Property** across various localities. The observed pattern shows that:
- Most localities have median property ages tightly clustered around **16–17 years**.
- There are **no extreme variations** between localities; no locality stands out as significantly newer or older.
- The housing stock represented in this dataset appears to come predominantly from the **early 2000s and 2010s construction periods**.
- This indicates that the dataset does not contain many newly built localities nor very old legacy neighborhoods (e.g., >30-year median age).
- Overall, the uniformity in property age suggests that **location-based age differences are minimal**, meaning property age is not a major differentiator across localities in this sample.

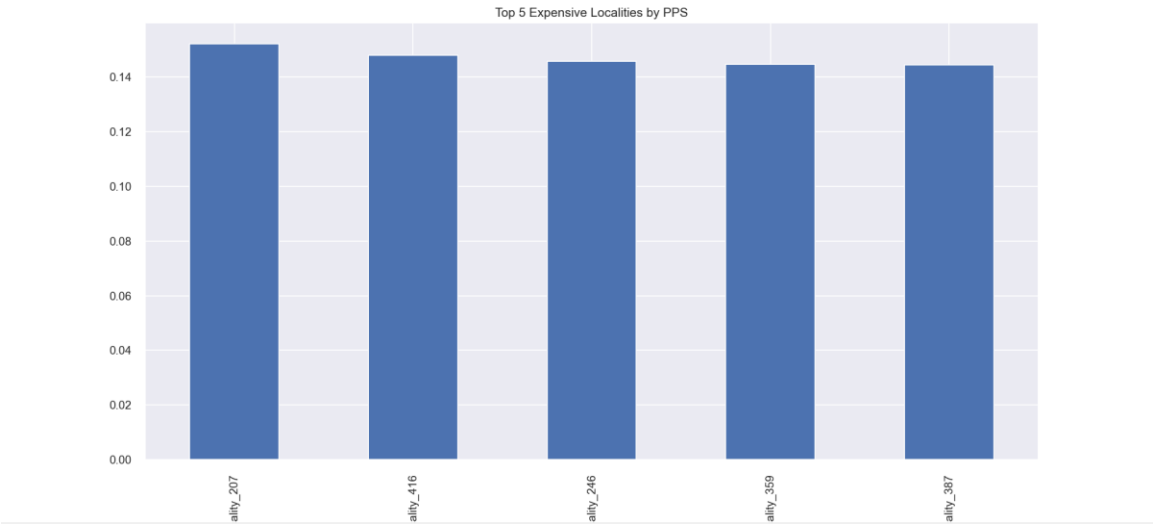
6. Location-Based Insights

6.1 BHK Distribution Across Cities



- **What the plot shows:** 2–3 BHK dominate; distribution varies slightly by city (metros have more 1–2 BHK due to land constraints).
- **Business interpretation:** BHK mix guides product-market-fit suggestions and influencing price per category (e.g., 3BHK in city X commands higher unit price).
- **ML & feature actions:**
- Use BHK as numeric and categorical (for non-linear relationships).
- Interact BHK with city/locality features for finer segmentation

6.2 Top 5 Expensive Localities by Price per SqFt



- **What the plot shows:** A small set of localities have elevated PPS > 0.14 (normalized units).
- **Business interpretation:** These micro-markets are premium and good candidates for high-net-worth investor strategies. For average investors they may not be relevant due to high entry cost.
- **ML & feature actions:**
- Create `locality_premium_flag` for top X% localities.
- Use locality-level time-series (if you have historical prices) for growth-rate features to model appreciation.

7. Correlation Analysis

7.1 Heatmap

What the plot shows:

- `Price_per_SqFt` positively correlates with `Price_in_Lakhs` (~0.56).
- `Price_per_SqFt` is negatively correlated with `Size_in_SqFt` (~-0.61) — larger area often gets lower per-sqft rate.
- `Year_Built` and `Age_of_Property` perfectly inverse (expected), `Build_Year` strongly correlates with `Build_Decade`.

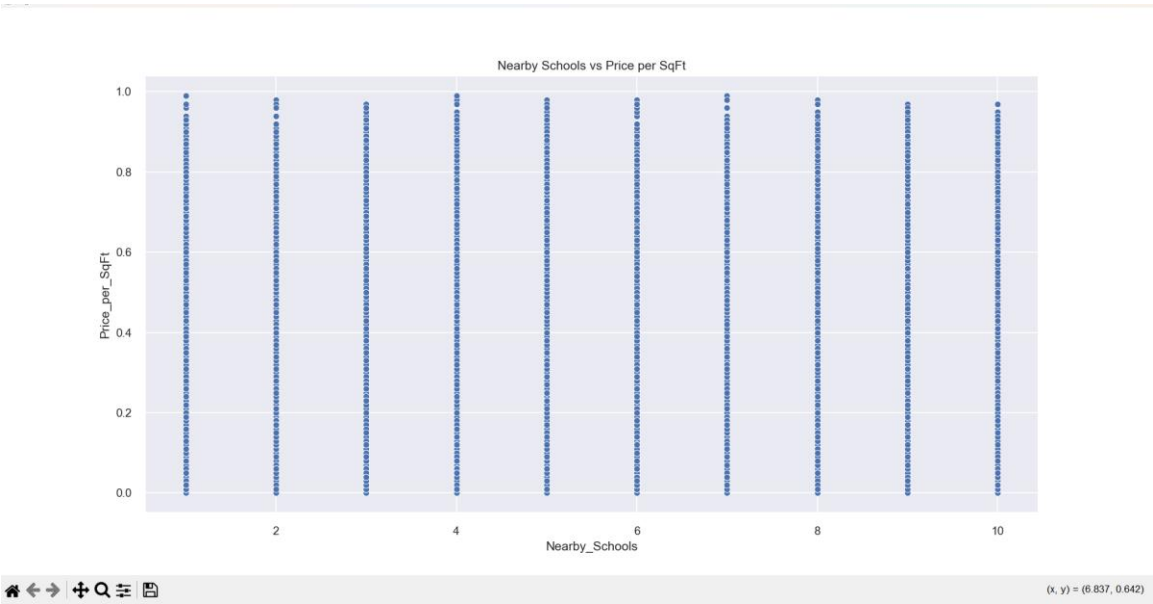
Business interpretation: PPS is a pivotal feature: it captures neighborhood premium and partially removes size effects from the price signal. The negative PPS-size correlation is classic: larger homes (often in outer areas) have lower unit costs.

ML & feature actions:

- Drop redundant fields or consolidate (e.g., keep `Build_Year` or `Age`, not both).
- Use PPS × Size interactions; also consider principal component analysis (PCA) on tightly correlated building-year features if model complexity is a concern.

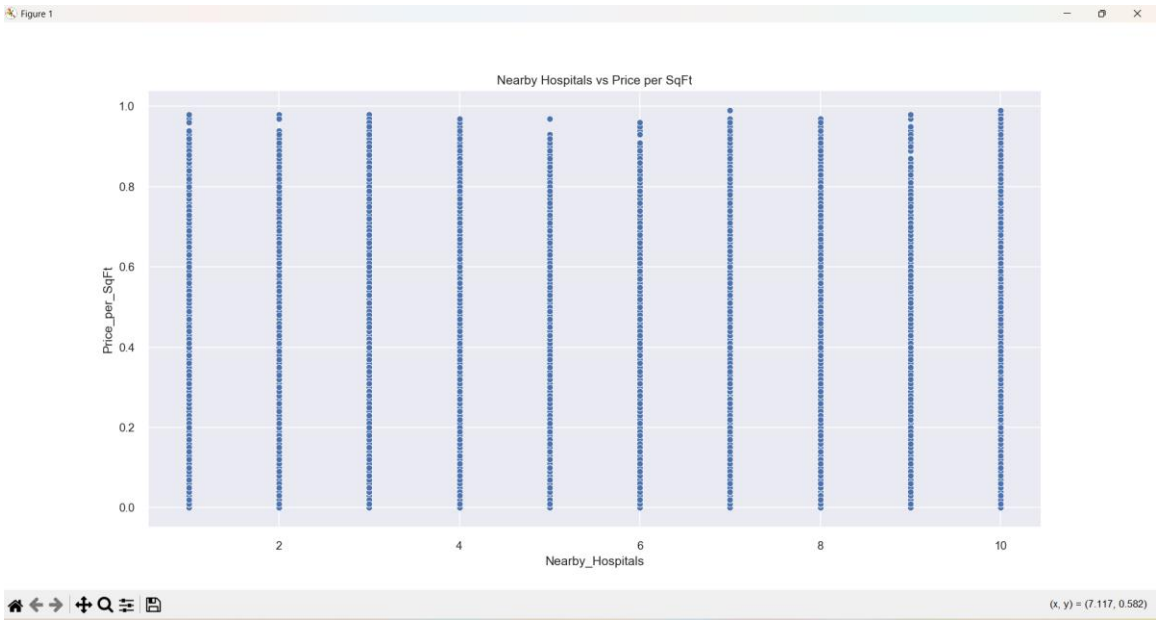
8. Infrastructure & Amenity Insights

8.1 Nearby Schools vs PPS



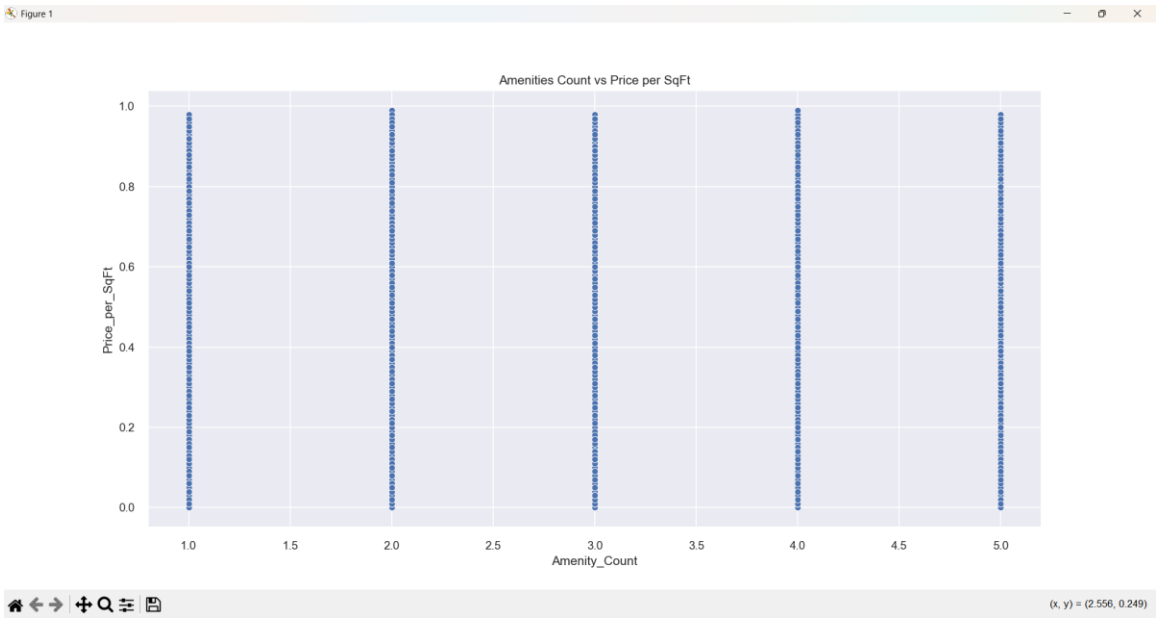
- **What the plot shows:** PPS shows some dispersion across different counts of nearby schools but no strong monotonic trend — points are widely scattered across school counts.
- **Business interpretation:** While schools are desired, their numeric count in this dataset does not strongly move PPS by itself. Quality (ratings) and proximity-to-premium schools would be stronger signals if available.
- **ML & feature actions:**
- Replace count with a weighted `school_score` if you can get school ratings or distance measures.
- Engineer `has_top_school_within_2km` binary if geodata is available

8.2 Nearby Hospitals vs PPS



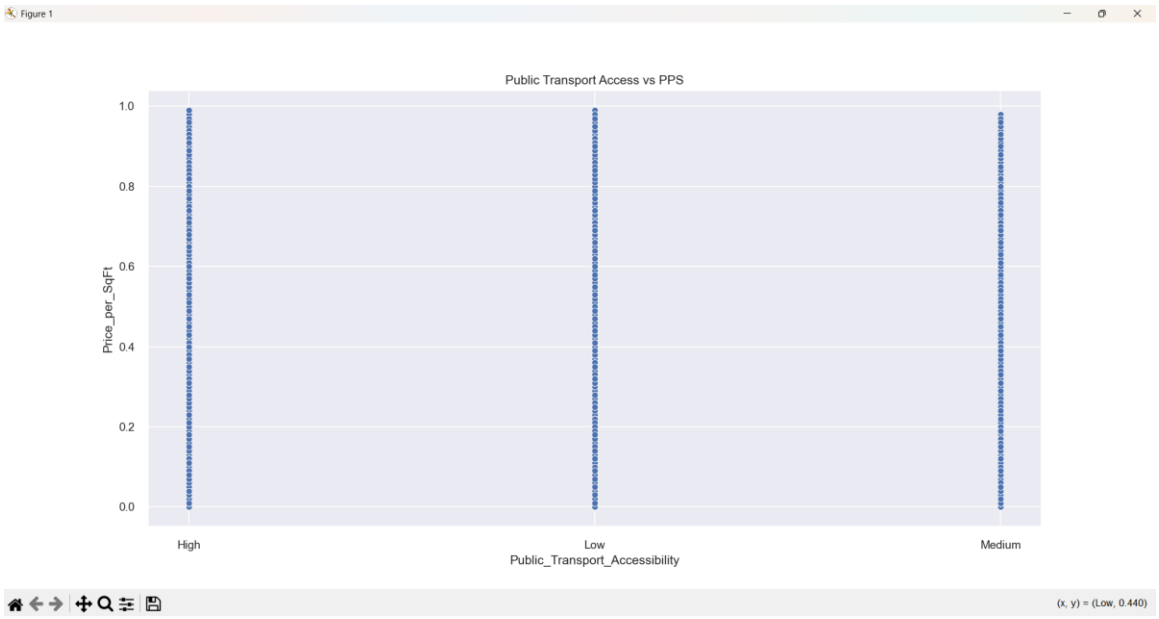
- **What the plot shows:** Similar to schools — moderate scatter, no dramatic PPS jump with hospital counts.
- **Business interpretation:** Hospitals add to locality resilience but are less price-driving than schools or metro access. Again, quality and proximity matter more than count.
- **ML & feature actions:**
- Engineer `healthcare_access_score` (combining hospital count, hospital type, proximity if available).

8.3 Amenities Count vs PPS



- **What the plot shows:** A slight upward trend: properties with more amenities (pool, gym, clubhouse) tend toward higher PPS, though scatter remains.
- **Business interpretation:** Amenities package is a price premium driver. For investors, projects with more amenities generally have stronger rental potential and resale demand.
- **ML & feature actions:**
- Convert `Amenities` free-text to `Amenity_Count` and `Amenity_Score` (weighted by amenity importance).
- Create flags for key amenities (Pool, Gym, Clubhouse) rather than only count.

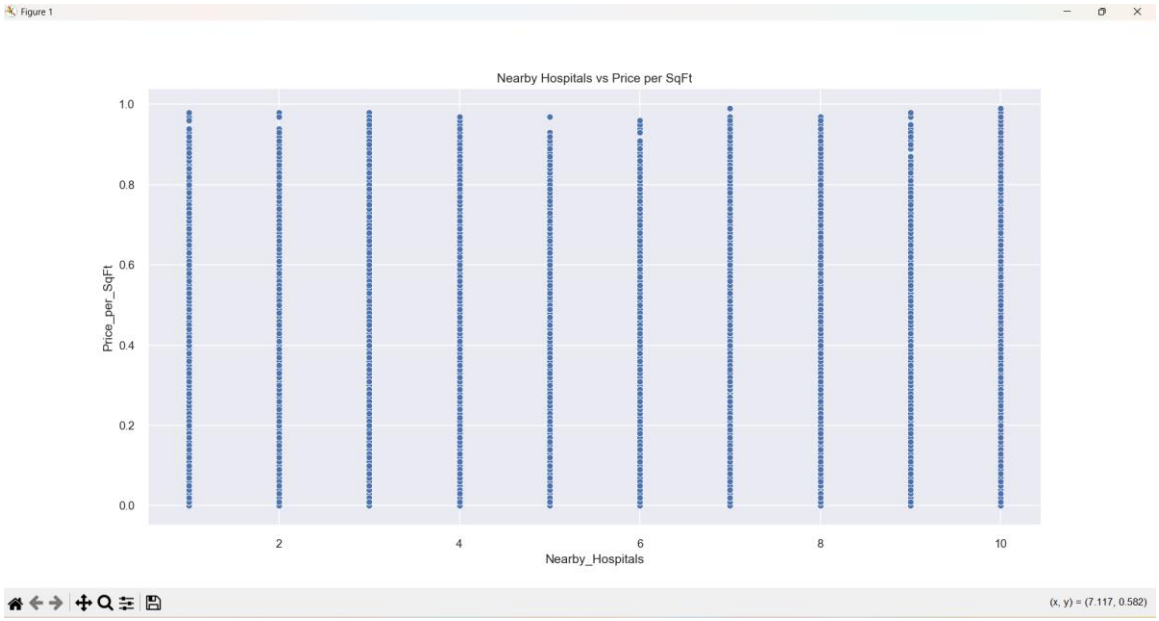
8.4 Public Transport Accessibility vs PPS



- **What the plot shows:** Properties labelled `High` accessibility cluster at higher PPS levels; `Low` and `Medium` are lower.
- **Business interpretation:** Transport access is a consistent premium driver. Nearness to metro and major bus hubs significantly increases demand and hence PPS.
- **ML & feature actions:**
- Keep `Public_Transport_Accessibility` ordinal-coded (Low=0, Medium=1, High=2) for tree-based and linear models.
- If possible, add distance-to-nearest-metro as numeric.

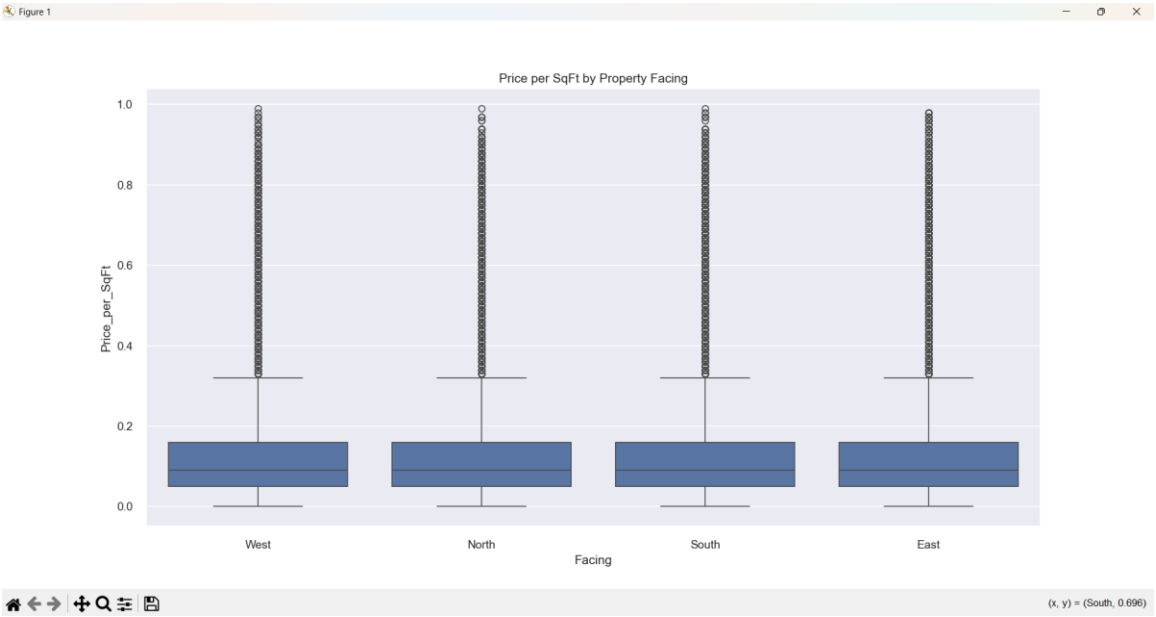
9. Categorical Feature Insights

9.1 Furnished Status vs Price



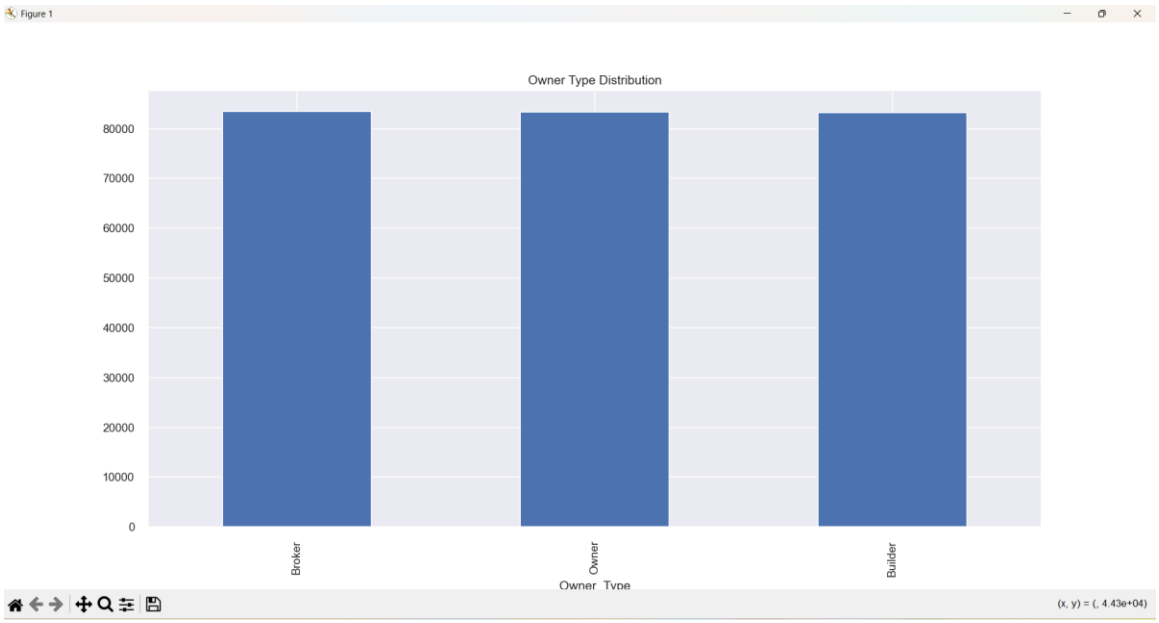
- **What the plot shows:** Boxplots show modest differences in median price by furnish status; furnished units slightly pricier but not dramatically.
- **Business interpretation:** Furnishing influences appeal and time-on-market, but purchase price premium is modest — furnishing often affects rent more than sale price.
- **ML & feature actions:**
- Keep `Furnished_Status` categorical. Combine with `Age_of_Property` to detect if newly furnished units differ from old ones.

9.2 Facing Direction vs PPS



- **What the plot shows:** All facings have similar median PPS; small differences favor East/North in median values.
- **Business interpretation:** In many Indian markets cultural preferences raise demand slightly for East/North facing units. The effect is small but consistent.
- **ML & feature actions:**
- Encode facing as categorical and test as feature in model; consider special rules for markets where cultural orientation has stronger effects

9.3 Owner Type Distribution



What the plot shows

- The distribution of Owner_Type is nearly **balanced across three categories**:
- **Owner**
- **Broker**
- **Builder**
- No category significantly dominates the dataset, and all three listing sources appear evenly represented.

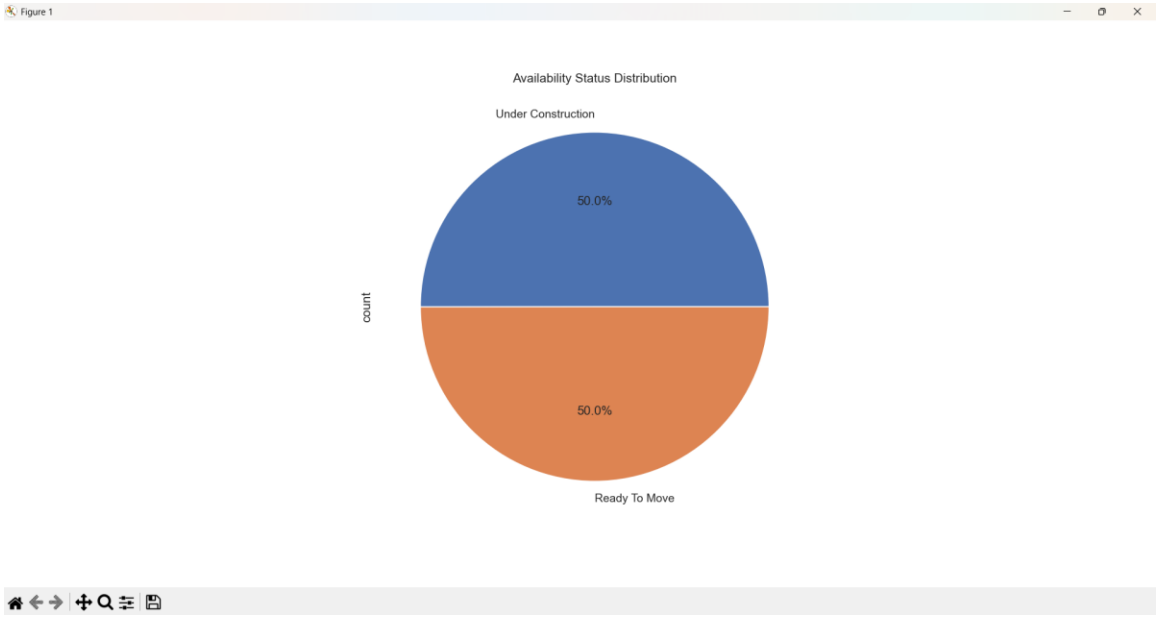
Business Interpretation

- A balanced distribution indicates that the property listings originate from a diverse mix of individual sellers, professional intermediaries, and formal developers. This helps avoid **listing bias**, where a dataset dominated by brokers or builders might skew price expectations due to specific pricing behaviors such as:
- Broker-influenced inflated listing prices
- Builder-driven premium pricing in new projects
- Owners offering negotiable or below-market listings
- This balance ensures **fair representation of the real market**, making the insights more reliable.

- ## ML & Feature Engineering Actions

- Keep **Owner_Type** as a categorical feature — it affects negotiability and generally correlates with price expectations.
- Use one-hot encoding or target encoding depending on model type.
- Consider interaction features such as **Owner_Type × Property_Type** (builders more likely handle apartments or new constructions).

9.4 Availability Status Distribution



- ## What the plot shows
- The chart indicates a roughly **50–50 split** between:
- **Ready to Move** properties
- **Under Construction** properties
- Both categories appear equally represented.

- ## Business Interpretation

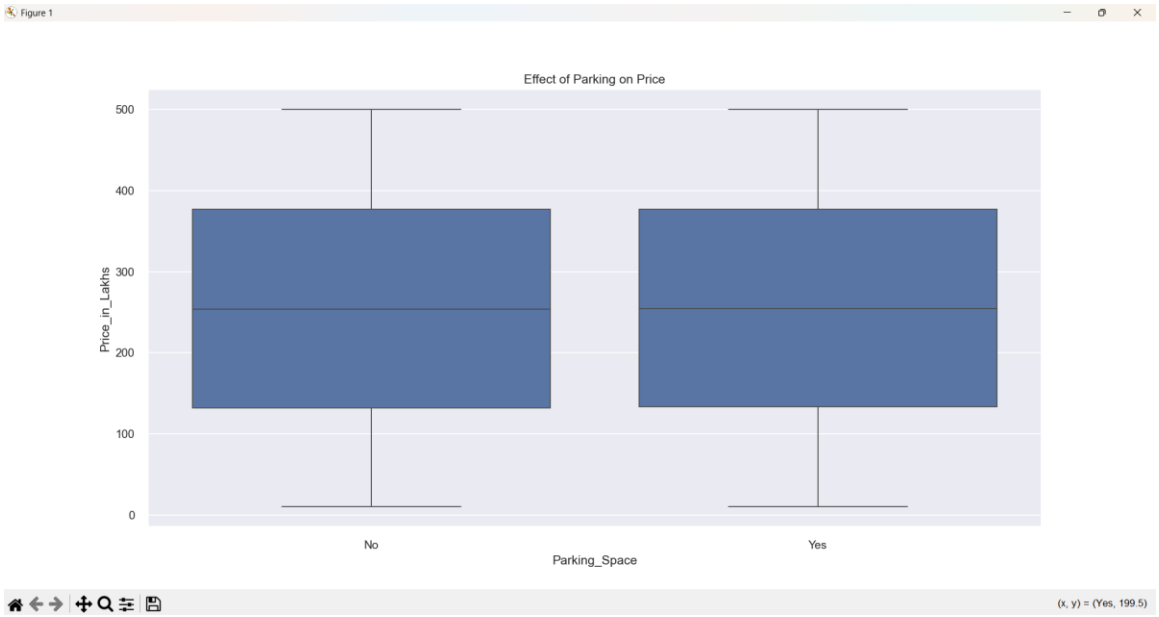
- This balanced availability gives insight into the supply mix of the housing market.
- **Ready-to-Move properties** are preferred by end-users seeking immediate occupancy and low completion risk.
- **Under-Construction properties** appeal to investors due to:

    - Lower acquisition cost
    - Potential price appreciation by completion
    - Flexible payment schedules
- A balanced dataset suggests that the market caters **both to immediate buyers and long-term investors**, which improves the reliability of investment models.

- ### ML & Feature Engineering Actions

- Treat **Under Construction** as a separate encoded flag because:
    - Under-construction properties often have different pricing structures.
    - They show different appreciation patterns during ML-based future price prediction.
- Create a feature:
  construction_phase_flag = 1 if Under Construction else 0
- Optional interaction:
  Availability_Status × Age_of_Property (Age=0 for new builds).
-
- Balanced distribution beneficial for diverse buying segments.

9.5 Parking Space vs Price



- **What the plot shows**

- The boxplot comparison demonstrates that **properties with parking space generally show higher median prices** than properties without designated parking.

- ### Business Interpretation

- Parking is a crucial amenity, especially in urban or high-density regions.
  Properties with parking command higher prices because:
- Parking scarcity is common in metros
- Reserved parking reduces daily inconvenience
- Higher resale value and better rental demand
- In gated communities, parking often pairs with premium amenities
- Thus, parking availability correlates with **both demand and perceived value**, indicating that it is a meaningful feature for valuation.

- ### ML & Feature Engineering Actions

- Include **Parking_Space** as a numeric or binary feature (0/1).
- Create an engineered feature:
  - `has_parking = 1 if Parking_Space > 0 else 0`
- Interaction with locality:
  - Parking premium may be higher in dense cities like Mumbai or Bangalore —
    → **has_parking × City**
- Parking also helps improve **investment classification accuracy**, because properties with parking tend to have lower long-term depreciation and better appreciation rates.

10. Key Insights Summary

1. **Price Drivers:**
   a. Location (city, locality)
   b. Amenities
   c. Building age
   d. Property size and type
   e. Infrastructure access
   f. Facing and furnishing
2. **Demand Characteristics:**
   a. 2BHK and 3BHK are most common.
   b. Premium localities command significantly higher PPS.

3. **Outliers:**
    a. PPS shows more extreme values than size or price.
    b. Outlier handling recommended before ML modeling.
4. **Investment Indicators:**
    a. High PPS + High amenities + Good transport → strong investment potential.
    b. Mature properties (10–20 years) are common.
5. **Data Suitability:**
    a. Very clean dataset (no missing or duplicate values).
    b. Strong for regression and classification models.

11. Conclusion

The dataset provides a comprehensive understanding of India's residential real estate market. The EDA reveals consistent pricing behaviour, strong locality dependence, and meaningful relationships between price, size, age, and amenities. These insights serve as the foundation for robust machine learning models for **price prediction** and **investment classification**.