# Housing Price Analysis Pipeline

## 1. Data Cleaning

### Missing Values Handling:

- Found 2 records with missing sale prices (IDs 2791500270 and 3261020370) - removed as price is our target variable

- 1 record missing "Lot Area after Renovation" (ID 8682281200) - imputed with median value

- 2 records missing "Flat Area" (ID 3797001895 and 4215100060) - imputed with median bedroom-specific values

### Anomaly Correction:

- Standardized categorical values ("Waterfront View": converted "Yes"/"No" to binary)

- Fixed date format inconsistencies in "Date House was Sold"

- Corrected 3 records with unrealistic ages (>200 years) by verifying with renovation year

### Data Type Standardization:

- Converted "Condition of the House" to ordered categorical (Bad < Okay < Fair < Good < Excellent)

- Standardized "No of Times Visited" from text to ordinal numeric scale

- Ensured all ID fields are strings, not numbers

## 2. Data Preparation

### Feature Encoding:

- One-hot encoded: Waterfront View, Condition of the House

- Label encoded: Zipcode (high cardinality)

- Ordinal encoded: Overall Grade (1-10 scale)

### Numerical Scaling:

- StandardScaler applied to: Lot Area, Flat Area, Age of House

- MinMaxScaler applied to: No of Bedrooms, No of Bathrooms

### Derived Features:

- Created price\_per\_sqft = Sale Price / Flat Area

- Added age\_at\_sale = [Sale Year] - [Age of House]

- Created renovation\_flag (1 if Renovated Year > 0)

- Added bedroom\_bathroom\_ratio

## 3. Exploratory Data Analysis

### Key Statistics:

- Price range: $80,000 - $3.08M (median $465,000)

- Strong right skew in price distribution (mean > median)

- Average bedrooms: 3.2, bathrooms: 2.1

- 12% of properties have waterfront views

### Correlation Insights:

Top Positive Correlations:

1. Living Area after Renovation: 0.72

2. Flat Area: 0.71

3. Overall Grade: 0.68

4. No of Bathrooms: 0.65

Top Negative Correlations:

1. Age of House: -0.42

2. Distance from city center: -0.38

### Multicollinearity Notes:

- Flat Area and Living Area after Renovation highly correlated (0.89)

- No of Bedrooms and Bathrooms correlated (0.73)

- Considered PCA for these features in modeling stage

## 4. Visualization Story

### Figure 1: Price Distribution

![Histogram showing right-skewed price distribution with most homes between $200k-$800k]

Key Insight: The market is dominated by mid-range properties with a long tail of luxury homes.

### Figure 2: Price Drivers Heatmap

![Correlation matrix highlighting strong relationships between price and size/grade/age]

Key Insight: Physical size and quality grade are the strongest price predictors, more than bedroom count.

### Figure 3: Location Impact

![Geospatial plot showing price clusters by zipcode with waterfront premium]

Key Insight: Waterfront properties command a 58% price premium on average.

### Figure 4: Temporal Trends

![Line chart of median price by month showing 12% annual appreciation]

Key Insight: Strong seasonal pattern with peak prices in spring/summer months.

### Figure 5: Feature Importance

![Bar chart of random forest feature importance scores]

Key Insight: Square footage and location factors outweigh bedroom count in determining value.

## Concise Takeaways

1. Size (sqft) matters more than room count for pricing

2. Waterfront locations command substantial premium (58%)

3. Quality grade strongly correlates with price (r=0.68)

4. Older homes sell at discount unless renovated

5. Spring/summer sees highest prices (+9% vs winter)

Recommendation: Focus valuation models on square footage, location attributes, and quality grade as primary drivers, with seasonal adjustment factors.