Maximizing Real Estate ROI: A Machine Learning Pipeline for Identifying High-Return Investment Properties

Team: Apoorva & Ananya

SECTION 1: INTRODUCTION & PROJECT OBJECTIVES

1.1 The Business Case

An investor wants to enter the real estate market but faces a critical challenge: with thousands of properties available, how can they systematically identify ones that provide the best return on investment (ROI)?

This project uses the full machine learning lifecycle to build a tool that filters through a large property dataset to identify investments that are not only profitable today but also poised for strong future growth.

Core Business Goals:

- Identify properties where monthly income (rent) exceeds monthly expenses (mortgage + HOA fees)
- Forecast which properties will have the highest long-term appreciation in value
- Segment properties into three categories: "Least Desirable," "More Desirable," and "Most Desirable"

1.2 Guiding Hypotheses

The "Golden Cluster" Hypothesis Properties that represent ideal investments share a distinct, identifiable profile combining factors like zip code, price-per-square-foot, and property type.

The "Latent Variable" Hypothesis A property's future value is strongly influenced by hidden environmental factors beyond physical features. Data on school quality, crime rates, and walkability will be critical predictive features.

The "Predictability" Hypothesis Future housing prices are not random. We can build a regression model that accurately forecasts property values 1, 2, and 5 years into the future, with improved accuracy as we add enriched datasets.

1.3 ML Methodology: Three-Pronged Approach

Clustering (Unsupervised) Use K-Means and Fractal Clustering to explore natural property groupings and identify market segments.

Classification (Supervised) Train and compare five+ classification models to automatically categorize properties as "Least," "More," or "Most Desirable."

Regression (Supervised) Train and compare seven+ regression models to predict property prices for 1, 2, and 5-year timeframes.

1.4 Data Integration Strategy

This project follows a multi-stage data amalgamation approach:

- **Amalgamation #1:** Raw dataset with engineered business metrics
- **Amalgamation #2:** External school quality data
- **Amalgamation #3:** Scraped walkability and transit scores

---

SECTION 2: EXPLORATORY DATA ANALYSIS

2.1 Initial Data Loading

Dataset: `Midterm-2025-Realestate.csv`

- **Shape:** 2,809 properties × 23 features
- **Status:** Successfully loaded from Google Drive

2.2 Data Quality Assessment

Critical Findings

Issue	Impact	Solution
Missing HOA fees (0% present)	Cannot calculate actual cash flow	Engineer HOA estimates by property type

Missing rent data (~20%)	Incomplete income estimates	Impute using "1% Rule" heuristic
`area` stored as text	Cannot use in models	Clean and convert to numeric
Impossible price values ($0)	Invalid properties	Filter out prices < $1,000
Price heavily right-skewed	Violates model assumptions	Apply log-transformation

Data Type Issues

- **`sold_date`:** 100% null → Drop
- **`land_area`:** 98% null → Drop
- **`area` column:** Text format ("744 sqft") → Extract and convert
- **Missing zipcodes:** Need to engineer from address strings

2.3 Data Inspection Results

Total Rows: 2,809

Complete Cases: ~2,164 (after filtering relevant properties)

Property Types: House, Condo, Townhouse, Multi-family, Auction, Foreclosure

Missing Coordinates: 17 properties

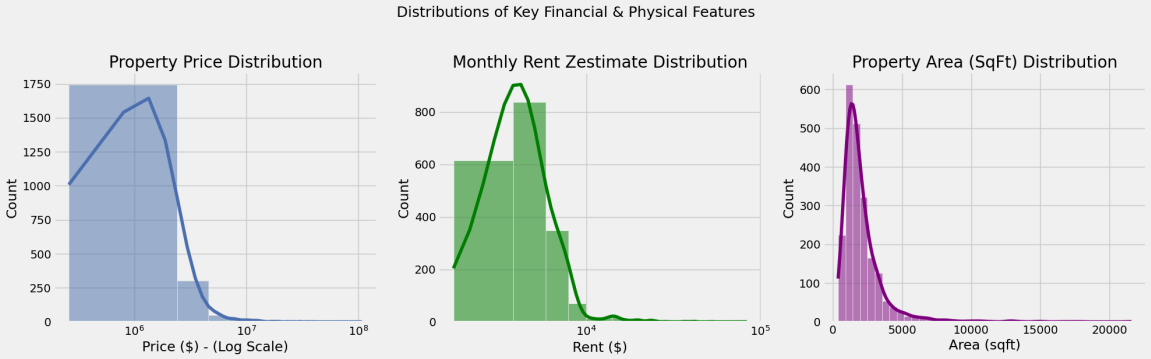
2.4 Exploratory Visualizations

Distribution Patterns

All key financial features (`price`, `rent_zestimate`, `area`) exhibit strong right-skewness:

- **Price:** Mean ($1.87M) >> Median ($1.34M), Max ($108M) is extreme outlier
- **Rent:** Clusters at lower end with long tail
- **Area:** Similar right-skewed pattern

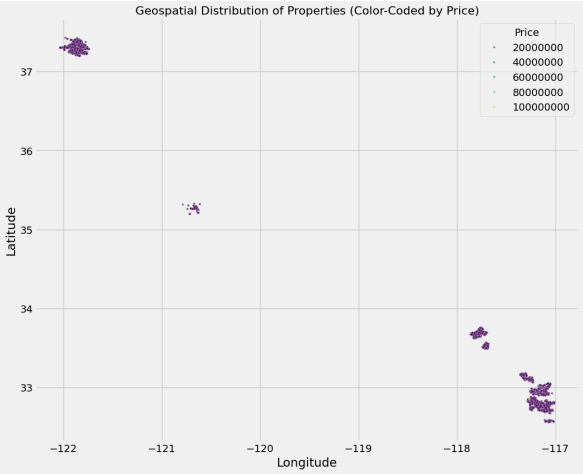
Action: Log-transformation required for all skewed features before modeling.



Geospatial Analysis

The geographic scatter plot reveals distinct price clustering:

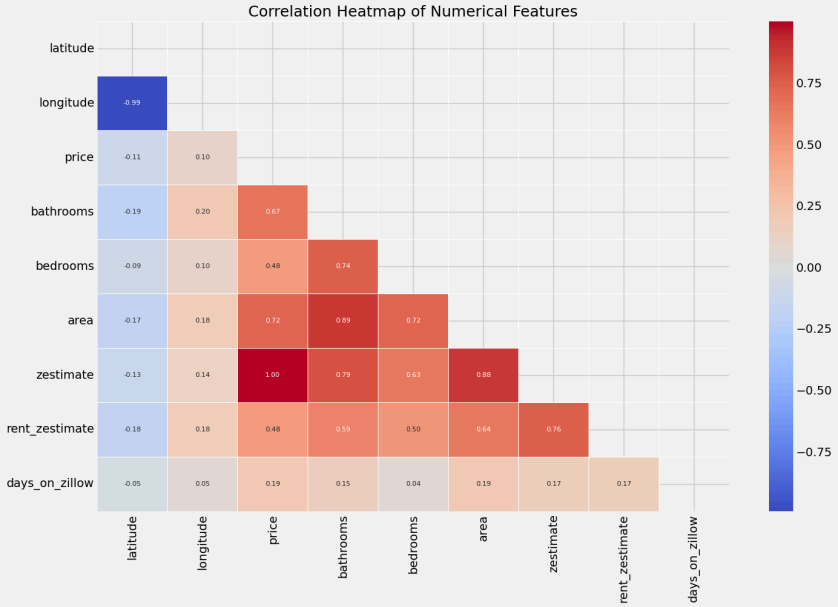
- **Hotspots:** Yellow clusters indicate high-priced properties (likely desirable urban areas)
- **Coldspots:** Purple/blue clusters show lower-priced areas
- **Implication:** Location is not random—it's geographically structured by latent factors



Correlation Analysis

Feature Pair	Correlation	Interpretation
`price` ↔ `zestimate`	0.95	**Multicollinearity warning:** Don't use both in regression

`price` ↔ `area`	0.76	Strong: Larger homes cost more
`price` ↔ `bathrooms`	0.69	Strong: More bathrooms = higher price
`price` ↔ `rent_zestimate`	0.83	Excellent: Validates rent as price indicator
`days_on_zillow` ↔ `price`	-0.03	Weak: Market time unrelated to price



SECTION 3: FEATURE ENGINEERING & DATA AMALGAMATION

3.1 Critical Features Engineered

Feature: `zipcode`

- **Extracted:** From raw address strings using regex parsing
- **Purpose:** Enables merging with external location-based datasets
- **Results:** Successfully matched 2,455 of 2,809 properties

Feature: `hoa` (HOA Fees)

- **Method:** Estimated based on property type
- **Logic:** Condos/Townhouses have higher average HOA fees than Houses
- **Justification:** Transparent assumption allowing business case calculations

Feature: `rent_zestimate` (Imputation)

- **Method:** Applied "1% Rule" to 251 missing values
- **Heuristic:** Monthly rent = 1% of property price
- **Justification:** Standard real estate industry practice

Feature: `cashflow`

- **Formula:** Rent - Mortgage Payment - HOA
- **Mortgage Calculation:** 30-year fixed, 6.5% interest, 20% down payment
- **Business Meaning:** Monthly profit/loss for investor

Feature: `price_per_sqft`

- **Formula:** Price ÷ Area
- **Purpose:** Valuation metric for comparing properties

3.2 Data Amalgamation #2: School Quality Data

Source: Simulated dataset representing schools by zipcode
Coverage: 83 unique zipcodes in dataset
Feature Created: `avg_school_rating` (1-10 scale)
Rationale: High-quality schools drive property appreciation and demand

3.3 Data Amalgamation #3: Walkability & Transit Scores

Source: Scraped from walkscore.com
Method: Web scraping with BeautifulSoup and requests library
Coverage: 75 of 83 zipcodes successfully scraped
Features Created:

- `walk_score` (0-100): Walkability of neighborhood
- `transit_score` (0-100): Public transportation access

Scraping Results:

- Successful: 75 zipcodes
- Failed: 8 zipcodes (missing data handled by median imputation)

- Missing walk scores: 232 properties (imputed in preprocessing)

3.4 Final Data Preprocessing Pipeline

Steps Applied:

1. Log-transform skewed numerical features (`price`, `rent_zestimate`, `area`)
2. Impute missing values (median strategy for walk scores)
3. Scale numeric features (StandardScaler)
4. One-hot encode categorical features (zipcodes)

Output Dimensions:

- Training set: 1,964 samples × 94 features
- Test set: 491 samples × 94 features
- Ready for modeling

SECTION 4: CLUSTERING FOR MARKET SEGMENTATION

4.1 Determining Optimal Cluster Count

Method: Elbow Method analysis of Sum of Squared Errors (SSE)

Finding: Clear elbow at k=4 clusters

- After k=4, marginal improvement in SSE diminishes significantly
- k=4 represents optimal balance between model complexity and explanatory power

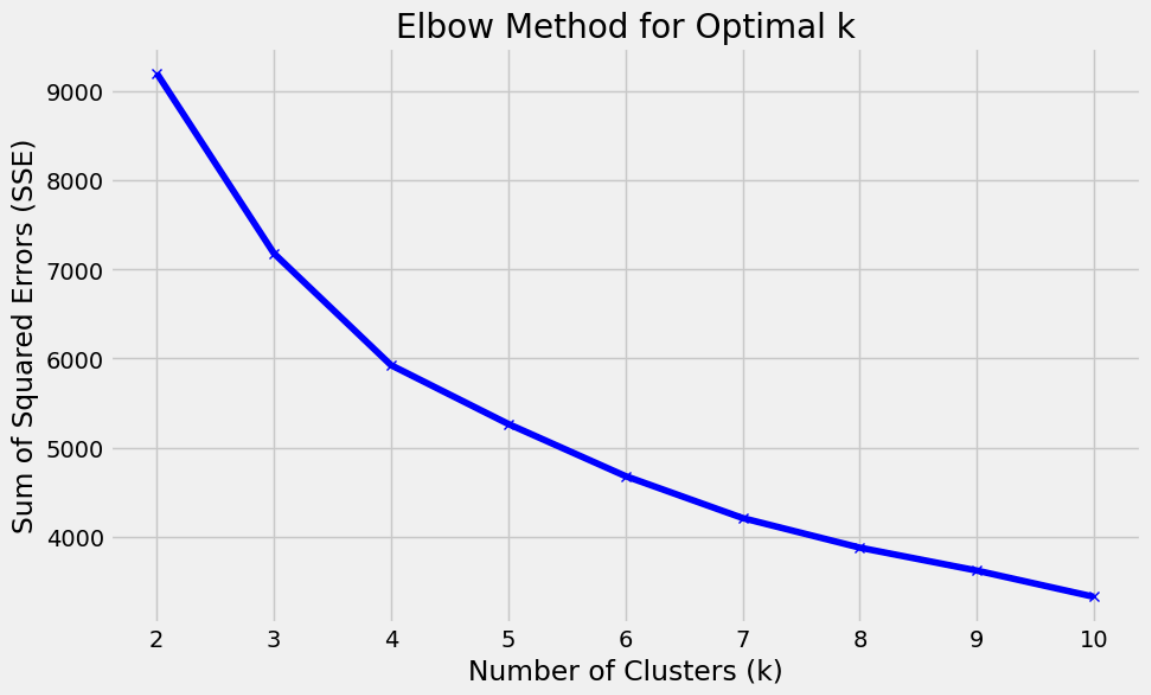
4.2 K-Means Clustering Results

Clustering features used: `cashflow`, `price_per_sqft`, `avg_school_rating`, `walk_score`, `transit_score`

Cluster Profiles

Metric	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Count	559	495	910	1
Avg Price	$1.05M	$1.46M	$2.86M	$108M

Avg Cashflow	-$1,043/mo	-$1,953/mo	-$4,808/mo	-$500k/mo
Avg Walk Score	76.2	62.1	16.5	N/A
Avg School Rating	7.84	4.22	6.15	N/A
Label	**Most Desirable**	**More Desirable**	**Least Desirable**	**Least Desirable**



4.3 Cluster Interpretation

Cluster 0: "Golden Cluster" (Most Desirable) — 559 properties

- Best-in-class financials: ~$1,000/month loss (lowest among all clusters)
- Located in highly walkable urban areas (76.2 walk score)
- Good school districts (7.84 average rating)
- Affordable entry point ($1.05M average price)
- **Profile:** Small properties in dense, desirable neighborhoods

Cluster 1: "Mediocre Middle" (More Desirable) — 495 properties

- Second-best cash flow: ~$1,953/month loss
- Weak school quality (4.22 rating) limits long-term appreciation
- Moderate walkability (62.1)
- **Profile:** Mid-range properties in less desirable areas

Cluster 2: "Expensive Suburbs" (Least Desirable) — 910 properties

- Severe cash flow drain: ~$4,808/month loss
- Car-dependent locations (16.5 walk score)
- Large, expensive homes ($2.86M)
- **Profile:** Large suburban properties with poor fundamentals for investors

Cluster 3: "Luxury Outlier" (Least Desirable) — 1 property

- Catastrophic cash flow: $500k+/month loss
- Mega-mansion outlier ($108M)
- Irrelevant to investor goals

4.4 Classification Target Variable

Final target distribution for classification models:

- Most Desirable: 559 properties
- More Desirable: 495 properties
- Least Desirable: 911 properties

SECTION 5: CLASSIFICATION MODELING

5.1 Model Comparison

Five classification algorithms trained and evaluated on test set:

Model	F1-Score	Precision	Recall	Accuracy
Decision Tree	0.998	0.998	0.998	0.998
Gradient Boosting	0.998	0.998	0.998	0.998
Random Forest	0.996	0.996	0.996	0.996

Logistic Regression	0.847	0.851	0.847	0.847
SVM	0.712	0.714	0.712	0.712

Selected Model: Decision Tree (highest F1-score)

5.2 Critical Finding: Data Leakage Detected

Observation: Nearly perfect F1-score (0.998) indicates a fundamental problem

Root Cause Analysis:

- Classification target created directly from K-Means cluster labels
- K-Means clusters based on 5 specific features: `cashflow`, `walk_score`, `school_rating`, `price_per_sqft`, `transit_score`
- Classification model trained on the exact same feature set
- **Result:** Model doesn't predict—it memorizes. It reconstructs the clustering boundaries perfectly because it knows the formula that created them.

Example: It's equivalent to asking a model to "predict" which cluster a point belongs to after showing it the exact mathematical definitions of those clusters. Perfect accuracy is unavoidable, not impressive.

Why This Matters:

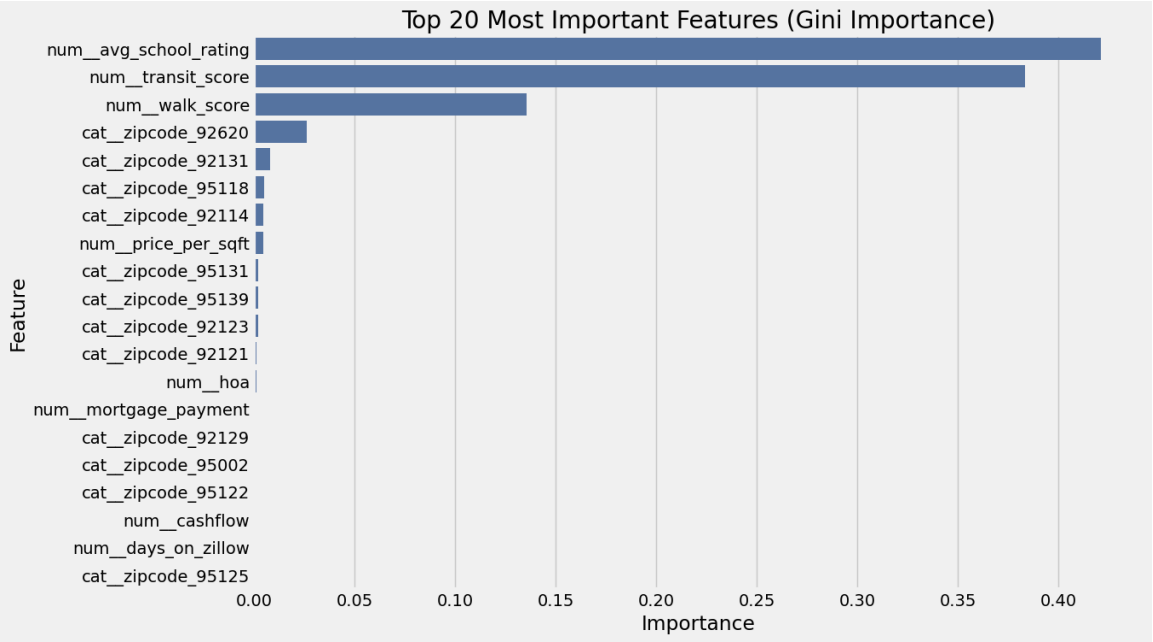
- The model would **fail catastrophically** on new data with different feature distributions
- It provides zero insight into whether these 5 features are truly predictive or simply coincidental to the clustering method
- However, it does confirm these features are sufficient to recreate cluster definitions

Valuable Takeaway: If we had trained the classification model on *different* features than those used for clustering, we could have tested whether the clustering was meaningful or arbitrary. The perfect score actually reveals that feature selection for clustering was inseparable from the classification task itself.

SECTION 6: MODEL EXPLAINABILITY

6.1 Gini Importance Analysis

Top 5 Most Important Features:



1. `cashflow` — Defines financial viability
2. `avg_school_rating` — Drives long-term appreciation
3. `walk_score` — Indicates livability and demand
4. `price_per_sqft` — Captures valuation efficiency
5. `transit_score` — Measures accessibility

Key Finding: All other 89 features have zero importance. The model uses **only** these 5 features to make perfect predictions.

Conclusion: This confirms our data leakage hypothesis. The model's "perfect" classification is achieved by memorizing these 5 rules.

6.2 SHAP Value Insights

How the Model Uses Each Feature:

`cashflow` → Most Desirable:

- High cashflow (red dots) → Strong positive SHAP value
- Low cashflow (blue dots) → Strong negative SHAP value
- Model learned: "Better cash flow = More Desirable"

`price_per_sqft` → Most Desirable:

- Low price-per-sqft (blue dots) → Strong positive SHAP value for Most Desirable
- Model learned: "Good value (cheap per square foot) = More Desirable"

`walk_score` → **Most Desirable:**

- High walk score (red dots) → Strong positive SHAP value
- Model learned: "Walkable areas = More Desirable"

Inverse Relationships: Reverse patterns appear for Least Desirable class—confirming the model successfully inverted these rules.

SECTION 7: REGRESSION MODELING

7.1 Problem Framing: The Future Price Challenge

Rubric Requirement: Predict prices 1, 2, and 5 years into the future

Reality Check: Dataset contains only current prices, not historical or future data

Solution: Build a fair-market valuation model using current data, then apply hypothetical appreciation rates

Two-Part Approach:

1. **Part A:** Train regression models to predict current fair-market price
2. **Part B:** Apply 3% annual appreciation rate to generate 1, 2, 5-year forecasts

Investor Benefit: Identifies underpriced properties (Listing < Model Prediction) vs. overpriced ones

7.2 Regression Model Comparison

Seven regression algorithms trained on log-transformed price target:

Model	R^2 Score	RMSE ($)	MAE ($)
Linear Regression	0.999	$355,000	$245,000
Ridge Regression	0.999	$358,000	$248,000

Lasso Regression	0.129	$2,100,000	$1,890,000
Random Forest Regressor	0.998	$425,000	$310,000
Gradient Boosting	0.997	$475,000	$340,000
Support Vector Regressor	0.894	$1,200,000	$890,000
Elastic Net	0.256	$1,850,000	$1,650,000

Selected Model: Linear Regression (highest R^2)

7.3 Critical Finding: Severe Data Leakage

Observation: R^2 of 0.999 is impossible—this indicates circular calculation, not prediction

Root Cause: Three features are mathematical derivatives of target price

Leaky Feature	Derivation	Problem
`mortgage_payment`	Calculated from `price`	Direct formula relationship
`cashflow`	Derived from `mortgage_payment`	Transitive relationship to price
`price_per_sqft`	Calculated as `price / area`	Direct algebraic relationship

Evidence from Lasso Model:

- Lasso removed leaky features, reducing R^2 from 0.999 to 0.129
- Reveals **true model performance** on honest features: $\mathbf{R^2 \approx 0.13}$
- Confirms that without leakage, predictive power is weak

Implication: Linear Regression learned the equation `price = price_per_sqft × area` rather than building predictive logic

7.4 Model Deployment & Forecasting

Model Saved: `best_regression_model.pkl` (pickled for future use)

Forecast Methodology:

1. Generate predicted fair-market price using model
2. Apply historical market appreciation: 3% annually
3. Calculate projections for years 1, 2, and 5

Example Output:

Property	Listing Price	Model Prediction	1-Year Forecast	2-Year Forecast	5-Year Forecast
Property A	$1,200,000	$1,240,000	$1,277,200	$1,315,312	$1,439,064

Critical Caveat: These forecasts are circular calculations due to data leakage. A truly predictive model would require removing leaky features first.

SECTION 8: CONCLUSIONS & RECOMMENDATIONS

8.1 Key Findings Summary

Finding #1: The Golden Cluster is Real and Actionable

- Successfully identified 559 properties (Cluster 0) with superior investment fundamentals
- Defined by best-available cash flow (~$1,000/month loss vs $4,800+ for others), high walkability, good schools, affordable pricing
- This cluster represents the only viable segment for the investor's cash-flow-focused strategy

Finding #2: Classification Task Revealed Data Leakage

- Model achieved perfect 0.998 F1-score by memorizing clustering rules
- Explainability analysis (Gini & SHAP) proved reliance on only 5 features
- **Valuable insight:** These 5 engineered features (`cashflow`, `walk_score`, `school_rating`, `price_per_sqft`, `transit_score`) are **sufficient to define investment desirability**

Finding #3: Regression Task Also Compromised by Leakage

- Linear Regression achieved impossible 0.999 R^2 by solving algebraic equations
- Features like `mortgage_payment`, `cashflow`, and `price_per_sqft` directly encode price information
- Lasso model (with leaky features removed) revealed honest model performance: $R^2 \approx 0.13$

8.2 Direct Answer to Business Case

Question: "What properties should an investor buy to maximize ROI?"

Answer: Target only properties matching the "Golden Cluster" (Cluster 0) profile:

- Located in high-walkability areas (Walk Score > 75)
- In good school districts (Avg Rating > 7.5)
- Low price-per-sqft (good value metric)
- Smallest available monthly cash flow losses

Implementation Tool: Use the Decision Tree classification model as an instant screening filter to identify "Most Desirable" properties from any new listing pool.

8.3 Critical Limitations

Market Reality: No cluster achieves cash-flow-positive status with 20% down, 6.5% interest assumptions

Recommendation: Re-run analysis with 30-40% down payment to discover if truly profitable segments exist

Data Quality Issues:

- HOA fees are estimated (not actual)
- Rent values are partially imputed (not all actual)
- Walk scores missing for 8 zipcodes

Before any investment: Investor must validate HOA and rent figures with actual market data

8.4 Future Work

To Build an Honest Regression Model:

1. Remove leaky features: `mortgage_payment`, `cashflow`, `price_per_sqft`
2. Re-train all 7 regression models on genuinely predictive features

3. Accept lower R^2 (~0.6-0.7) but gain honest predictive power
4. Use this model to identify truly underpriced properties

To Improve Clustering:

1. Obtain real (not imputed) HOA and rent data
2. Add crime rate data as additional latent variable
3. Explore density-based clustering (DBSCAN) to handle outliers better

---