Predicting Property Prices with Core Data/Text Mining Techniques

Daniel Phelps — 8 September 2025

Email: dphelps9693@floridapoly.edu | Project Webpage | Repository

Abstract

Goal: explain what drives higher vs. lower home prices using only tools from the syllabus. I will clean the data, explore with clear visuals, reduce to a few main directions (PCA), find natural groups (clustering), and mine simple patterns that connect feature combinations to price bands (association rules). If a dataset with short listing descriptions is approved, I will add basic text features (TF-IDF and sentiment) to see whether words add signal. All steps will be reproducible with code, figures, and a short write-up on a GitHub webpage. Deliverables include a final report and a 12-minute presentation.

1. Motivation

Accurate prices support buyers, sellers, and planners. Numeric features (living area, quality, year built) and area information (neighborhood) are known drivers. Short descriptions such as "renovated kitchen" or "near downtown" may add useful clues. This project uses class-approved methods to answer: (i) which features and combinations relate most to price; and (ii) whether short listing text adds value beyond numbers.

2. Related Work (brief)

Hedonic analysis treats price as a function of characteristics. Clustering summarizes markets into segments. Association rules reveal frequent combinations tied to outcomes (e.g., high price). Basic NLP (TF-IDF, sentiment) turns words into numeric features. This project combines these ideas in a compact, reproducible study.

3. Data and Preprocessing

Datasets: (1) Ames Housing (Kaggle) for structured features; and (2) optionally, an approved listing dataset with short descriptions. Cleaning and preparation steps:

• Drop rows with invalid/missing price; cap extreme outliers so they do not dominate plots.

• Impute missing numeric values (median) and categories (mode).

• Group rare categories into "Other" to simplify rules and clustering; standardize numeric columns for PCA/k-means.

• Create clear price bands (Low / Medium / High) for summaries and rule mining.

• For rule mining, discretize key numeric variables (e.g., living area → small/medium/large).

• If text is present: lowercase, remove punctuation/stopwords; make TF-IDF features; compute simple sentiment (VADER/AFINN).

• Train/validation/test split: 70/15/15 with a fixed random seed.

4. Methods

4.1 EDA & Visualization

Plot distributions (price, living area), box plots by neighborhood and quality, and simple correlation heatmaps. Each figure will have a one-sentence caption explaining the main takeaway.

4.2 Dimensionality Reduction (PCA, t-SNE)

Use PCA to summarize the main directions of variation and as a helper for clustering. Use t-SNE to create a 2-D map that makes groups easier to see (with a fixed seed for consistency).

4.3 Clustering

Run k-means and hierarchical clustering on standardized numeric features and selected encoded categories. Choose the number of clusters with the silhouette score. Profile each cluster with simple tables: size, typical living area, quality, neighborhood mix, and share of each price band.

4.4 Association Rules (Frequent Pattern Mining)

Apply Apriori or FP-Growth on categorical + discretized features. Report rules for High and Low price bands with support, confidence, and lift. Translate each top rule into plain English (e.g., "Large living area + High quality appears with the High price band more often than expected").

4.5 Text Features (optional)

Create TF-IDF vectors and a sentiment score. Check whether certain terms are enriched in High-price listings. Keep the text analysis small and focused.

4.6 Anomaly Checks

Flag unusual points using z-scores on residuals from a simple price vs living-area line or distance from cluster centers. Show results with and without outliers to test robustness.

4.7 Planned Figures

F1: Price distribution by neighborhood/quality; F2: Price vs living area (log scale); F3: PCA scree and 2-D scores; F4: t-SNE map colored by price band; F5: Cluster profiles; F6: Top rules table and (optional) one text wordcloud.

4.8 Tools

Python (pandas, matplotlib/seaborn), scikit-learn for PCA and clustering, mlxtend for Apriori/FP-Growth, NLTK/VADER for TF-IDF and sentiment. Work in Colab or VS Code; version control in GitHub.

5. Planned Experiments

E1 — EDA & Price Bands

• Clean data; create Low/Medium/High bands; summary plots with captions.

E2 — PCA / t-SNE

• Show structure in 2-D; describe visible groups; explain how many PCs matter.

E3 — Clustering

• Run k-means & hierarchical; choose K by silhouette; make easy cluster profiles.

E4 — Association Rules

• Mine rules for High & Low bands; report top rules with support, confidence, lift; 1–2 lines of interpretation each.

E5 — Optional Text

• Add TF-IDF/sentiment; re-check clusters or rules; note changes.

E6 — Outliers & Sensitivity

• Re-run key plots/rules with and without flagged outliers; note differences.

Success criteria: clear visuals + rules that make sense; concise explanations a non-expert can follow.

6. Risks & Mitigations

• Weak/no text → proceed structured-only; note text as future work.

• Sparse categories → group rare levels; discretize continuous features for rules.

• Outliers/noise → cap extremes; provide both "with" and "without" versions of key results.

7. Ethics, Privacy, and Licensing

Use only public, license-compliant data; avoid personal identifiers; cite sources on the project webpage. Discuss limits (e.g., results may not transfer to other markets without adjustment).

8. Reproducibility & Deliverables

• Public GitHub repo with code and a short README to run everything end-to-end.

• GitHub Pages website with the proposal PDF, results, figures, and plain-language explanations.

• Final report and a 12-minute presentation (10% of the project grade).

9. Timeline (matches course milestones)

• Proposal (Mon 8 Sep): EDA plan; publish proposal.pdf on the webpage.

• Checkpoint I (Mon 6 Oct): 5-page update—EDA complete; PCA/t-SNE; first clustering results.

• Checkpoint II (Mon 3 Nov): 8-page update—association rules; refined clusters; optional text; outlier analysis.

• Final (Fri 5 Dec): 9–11 page report + slides; post figures and a short summary on the webpage.

10. Expected Contributions

• A clear, visual explanation of what drives price using only course tools.

• A small set of clusters and rules that summarize the market in plain English.

• A clean, reproducible repo and webpage that classmates can follow.

References

• Jolliffe, I. T. (2002). Principal Component Analysis. Springer.

• van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. JMLR.

• Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation of Cluster Analysis. J. Comput. Appl. Math.

• Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. SODA.

• Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD.

• Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. VLDB.

• Han, J., Pei, J., & Yin, Y. (2000). Mining Frequent Patterns without Candidate Generation (FP-Growth). SIGMOD.

• Tan, P.-N., Steinbach, M., & Kumar, V. (2018). Introduction to Data Mining. Pearson.

• Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.

• Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. JMLR.

• Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. ICWSM.

• Nielsen, F. Å. (2011). AFINN: A New Word List for Sentiment Analysis.

• Ames Housing dataset (Kaggle). Public benchmark for structured housing data.