Predicting Housing Prices with Course Methods: A Trimmed, Reproducible Study

Daniel Phelps — 12 October 2025

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

Abstract

This proposal narrows scope to fit course timelines while keeping the project meaningful. Using the Ames Housing dataset, I will: (1) clean and explore the data; (2) reduce dimensions with PCA; (3) form simple market segments with k-means; and (4) mine association rules that link feature combinations to Low/Medium/High price bands. Deliverables include code, figures, and a concise write-up on a GitHub Pages site.

Motivation

Accurate pricing helps buyers, sellers, and planners. Numbers such as living area, overall quality, and neighborhood are known drivers. This project uses only techniques covered in the syllabus to explain patterns in a plain, visual way—no heavy modeling—so results are reproducible and easy to grade. The focus is on clarity and alignment with course topics (EDA/visualization, PCA, clustering, frequent pattern mining, and simple anomaly checks).

Related Work / Literature Review

Hedonic pricing models express home prices as a function of characteristics (Rosen, 1974). The Ames dataset is a widely used alternative to the Boston Housing data for benchmarking feature effects and predictive workflows (De Cock, 2011). For unsupervised structure, clustering with k-means and silhouette analysis is common in real-estate segmentation studies (Rousseeuw, 1987). Association rule mining (Agrawal & Srikant, 1994; Han et al., 2000) summarizes frequent co-occurring attributes; several housing papers use rules to describe high/low value patterns. For basic text, TF-IDF and lexicon-based sentiment (VADER/AFINN) are standard tools to turn short descriptions into features. This project combines these well-established methods in a compact, transparent pipeline.

Key references (short list)

- Rosen, S. (1974). Hedonic prices and implicit markets. Journal of Political Economy.
- De Cock, D. (2011). Ames, Iowa: Alternative to the Boston Housing Data Set. Journal of Statistics Education.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation of cluster analysis. J. Comput. Appl. Math.
- 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.
- Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based sentiment model. ICWSM.
- Nielsen, F. Å. (2011). AFINN word list for sentiment analysis.

Data & Preprocessing (Ames Housing)

Data split: 70/15/15 (train/val/test) with a fixed seed. Cleaning includes: removing invalid prices; imputing numeric medians and categorical modes; grouping rare categories into "Other"; capping extreme outliers for stable visuals. For rules, discretize key numerics (e.g., living area → small/medium/large) and define price bands (Low/Medium/High) using quantiles. Standardize numeric features for PCA and k-means. Keep transformations simple and well-documented.

Methods (from the course)

Exploratory Data Analysis (EDA) & Visualization

I will begin with a structured EDA pass to understand distributions, spot quality issues, and identify variables that are plausibly related to price.

Variables in scope: Numerics: SalePrice, GrLivArea (above-ground living area), TotalBsmtSF, GarageArea, GarageCars, YearBuilt, OverallQual, OverallCond, FullBath, LotArea.

Categoricals: Neighborhood, HouseStyle, BldgType, MSZoning, Exterior1st, KitchenQual, CentralAir.

Cleaning/transform hints: Because SalePrice is right-skewed, I will inspect both the raw scale and log10(SalePrice). I will flag extreme outliers (e.g., top/bottom 0.5–1%) and show results **with** and **without** them to make patterns robust. Missing values in numerics will be median-imputed; rare categories will be merged into "Other" when appropriate.

Plots and tables (with one-line takeaways).

- Distribution of SalePrice (linear and log) to justify log scale and outlier handling.
- Scatter of GrLivArea vs SalePrice (log-y), highlighting large-area outliers; report a simple correlation.
- Box plots of SalePrice by Neighborhood and by OverallQual to visualize major shifts in central tendency.
- Heatmap of correlations among numeric predictors to spot redundancy and guide PCA/feature selection.
- A compact summary table (N, mean, median, IQR) for SalePrice and 3–5 top predictors.

Outcome. A short narrative (2–3 sentences per figure) that says what the data show (e.g., "Price rises monotonically with OverallQual; some neighborhoods shift the median price by >$X."). These insights feed the PCA feature set and the clustering choices.

Dimensionality Reduction (PCA)

PCA is used to summarize structure and reduce redundancy among the standardized numeric variables; it is *not* a predictive model in this project.

Inputs.Standardized versions of key numerics from EDA: (GrLivArea, TotalBsmtSF, GarageArea, GarageCars, YearBuilt, OverallQual, FullBath, possibly LotArea). Categorical effects (e.g., Neighborhood) will be kept for later profiling rather than included in PCA.

Procedure.

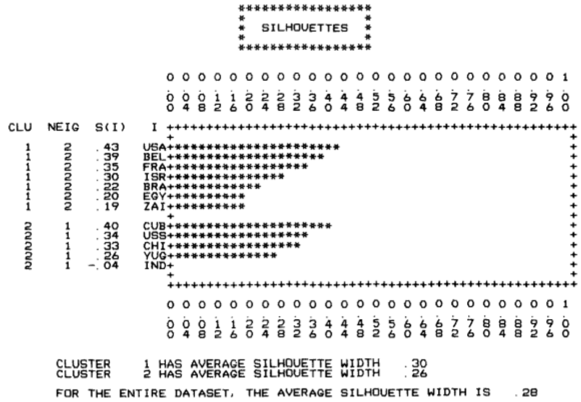
- Compute PCA on the correlation matrix (i.e., after centering and scaling each feature).
- Report a scree plot and cumulative variance. Retain the first k components that together explain ~70–85% of the variance (anticipated k ≈ 2–4).
- Inspect component loadings to interpret axes (e.g., "size/amenities axis" vs. "age/condition axis").
- Produce a 2-D PCA scatter colored by the price band (Low/Medium/High, defined by tertiles of SalePrice). Optionally mark ellipses for each band.

What I will conclude.

- Which combinations of measurements form the dominant axes of variation.
- Whether price bands separate in the PCA plane (and therefore whether clustering on these standardized variables is sensible).
- Which original features load strongly on the kept components; these will guide the subset used for k-means.

Deliverables: Scree plot, loading table (top absolute loadings per component), and a PCA scatter

Clustering



```
                              *****************
                              *               *
                              *  SILHOUETTES  *
                              *               *
                              *****************

                     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
                     . . . . . . . . . . . . . . . . . . . . . . . . . .
                     0 0 0 1 1 2 2 2 3 3 4 4 4 5 5 6 6 6 7 7 8 8 8 9 9 0
                     0 4 8 2 6 0 4 8 2 6 0 4 8 2 6 0 4 8 2 6 0 4 8 2 6 0
CLU  NEIG  S(I)    I +++++++++++++++++++++++++++++++++++++++++++++++++++
                     +                                                 +
 1    2     .43  USA+*********************                             +
 1    2     .39  BEL+*******************                               +
 1    2     .35  FRA+*****************                                 +
 1    2     .30  ISR+***************                                   +
 1    2     .22  BRA+***********                                       +
 1    2     .20  EGY+**********                                        +
 1    2     .19  ZAI+**********                                        +
                     +                                                 +
 2    1     .40  CUB+********************                              +
 2    1     .34  USS+*****************                                 +
 2    1     .33  CHI+****************                                  +
 2    1     .26  YUG+*************                                     +
 2    1    -.04  IND+                                                  +
                     +                                                 +
                     +++++++++++++++++++++++++++++++++++++++++++++++++++
                     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
                     . . . . . . . . . . . . . . . . . . . . . . . . . .
                     0 0 0 1 1 2 2 2 3 3 4 4 4 5 5 6 6 6 7 7 8 8 8 9 9 0
                     0 4 8 2 6 0 4 8 2 6 0 4 8 2 6 0 4 8 2 6 0 4 8 2 6 0

     CLUSTER     1 HAS AVERAGE SILHOUETTE WIDTH   .30
     CLUSTER     2 HAS AVERAGE SILHOUETTE WIDTH   .26

     FOR THE ENTIRE DATASET,  THE AVERAGE SILHOUETTE WIDTH IS    .28
```

Fig. 2. Silhouettes of a clustering with $k = 2$ of the twelve countries data of Table 1.

Figure 1: De Cock, D. (2011). Ames, Iowa: Alternative to the Boston Housing Data Set. Journal of Statistics Education.

Goal: find market segments in the standardized feature space and describe them in plain English.

Feature set.
The standardized numerics used in PCA (and possibly 1–2 binary indicators such as CentralAir if they add clear separation). Standardization ensures variables contribute equally.

Choosing K.

- Compute k-means for K = 2…8 with nstart = 25 random initializations.
- For each K, compute average silhouette width (Euclidean distance on standardized features).
- Select the K with the highest silhouette, preferring the simplest K if scores are tied or nearly tied. (We will reference Fig. 1 when discussing this choice.)

Cluster assignment and profiling.

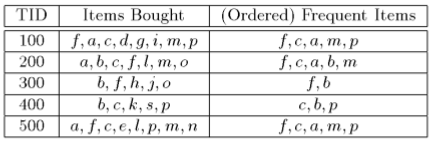
- Fit final k-means at the chosen K; record cluster labels.
- Produce a profile table per cluster: size (N), median SalePrice (and log), typical GrLivArea, median OverallQual, top 3 neighborhoods by share, and a short sentence describing the segment (e.g., "Large, higher-quality homes, mostly in Neighborhoods A/B").
- Create a bar plot comparing medians across clusters to communicate differences quickly.

Quality and stability checks.

- Inspect within-cluster variation of price and area to ensure clusters are not dominated by a few outliers.
- If time permits, run a quick bootstrap re-fit on 80% subsamples to see if profiles are stable (not required but informative).

Outcome.
Clear, interpretable segments that can be discussed in the final presentation (e.g., "Entry-level small homes," "Mid-tier family homes," "Large premium homes"). These segments also contextualize the association rules.

Association Rules (Apriori / FP-Growth)



TID	Items Bought	(Ordered) Frequent Items
100	f, a, c, d, g, i, m, p	f, c, a, m, p
200	a, b, c, f, l, m, o	f, c, a, b, m
300	b, f, h, j, o	f, b
400	b, c, k, s, p	c, b, p
500	a, f, c, e, l, p, m, n	f, c, a, m, p

Table 1: A transaction database as running example.

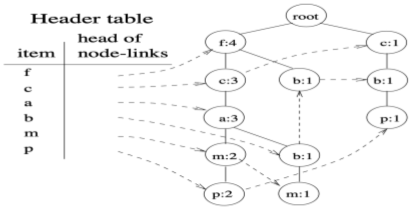


Figure 1: The FP-tree in Example 1.

Figure 2: FP-tree structure used by FP-Growth (adapted from Han, Pei, & Yin, 2000).

Objective: describe **combinations of attributes** that frequently occur with **Low** or **High** price bands.

Preparation.

- Define price_band as tertiles of SalePrice (Low/Medium/High).
- Discretize selected numerics using quantiles:

GrLivArea → area_band ∈ {small, medium, large}

TotalBsmtSF → {small, medium, large}

YearBuilt → {oldest, mid, newest}

OverallQual → {low, mid, high} (based on practical cut points)

- Keep a small set of cleaned categoricals: Neighborhood (top 6–8 levels + Other), CentralAir (Y/N), KitchenQual (grouped).

Mining strategy.

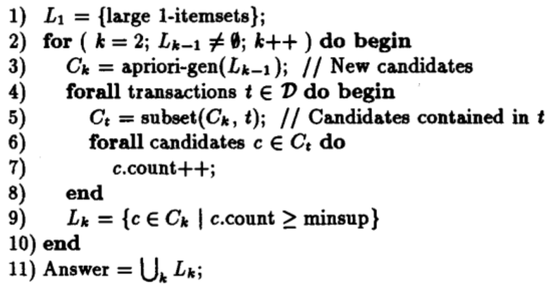
- Use **Apriori** (and/or **FP-Growth**) on a transaction dataset built from the discretized variables.
- **Appearance constraint**: RHS must be price_band=high or price_band=low (we are only interested in rules that imply price band).
- Start with **min support ≈ 2%** and **min confidence ≈ 60%**; adjust to get a manageable set (<50 rules).
- Rank rules by **lift** (how much more likely the RHS is given the LHS); report the **top 10** by lift for each price band.
- Remove **redundant** or **subsumed** rules (where a shorter LHS explains the same RHS with nearly equal metrics).

- Sanity-check for spurious rules due to tiny categories; if needed, raise min support or merge categories.

Reporting.

- A compact table with columns: LHS (conditions), RHS (price band), support, confidence, lift.
- Short, human-readable translations, e.g.,“area_band=large + OverallQual=high ⇒ price_band=high (support 9%, conf 78%, lift 1.35).”
- A 2–3 sentence discussion on *why* the strongest rules make sense (linking back to EDA/cluster profiles).

Outcome: Descriptive takeaways that connect **combinations** (not just single variables) to observed price levels—useful for storytelling in the final talk. We will reference **Fig. 2** (Apriori workflow) and **Fig. 3** (FP-tree) as small method insets; the results table itself uses our data.



```
$L_1$ = {large 1-itemsets};
for ( $k = 2$; $L_{k-1} \neq \emptyset$; $k$++ ) do begin
    $C_k$ = apriori-gen($L_{k-1}$); // New candidates
    forall transactions $t \in \mathcal{D}$ do begin
        $C_t$ = subset($C_k$, $t$); // Candidates contained in $t$
        forall candidates $c \in C_t$ do
            $c$.count++;
    end
    $L_k = \{c \in C_k \mid c.\text{count} \geq \text{minsup}\}$
end
Answer = $\bigcup_k L_k$;
```

Figure 1: Algorithm Apriori

Figure 3: Apriori workflow (adapted from Agrawal & Srikant, 1994).

Planned Figures

F1: Price distribution; F2: Price vs. living area (log scale); F3: PCA scree + 2-D PCA scatter; F4: Cluster profiles (bars/table); F5: Top association rules table.

Milestones & What “Done” Looks Like

• Proposal (this document) posted on the site.
• Checkpoint I: EDA complete; PCA plots; initial K selection; draft cluster profiles.
• Checkpoint II: association rules complete; refined clusters.
• Final: short report (9–11 pages), slides, repo with notebooks and figures.

Risks & Mitigations

• Outliers dominating visuals → cap extremes and show with/without views.
• Sparse categories → merge into “Other”; discretize key numerics.
• Time constraints → prioritize EDA, PCA, clustering, and rules.

Reproducibility & Artifacts

Public GitHub repo with notebooks (`01_eda.ipynb` → `04_association_rules.ipynb’), figures folder for exported plots, and a GitHub Pages site with links. A README describes how to run the notebooks.

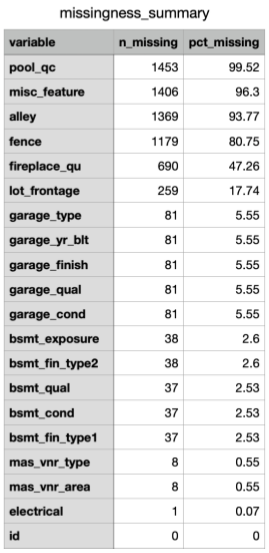
Data Setup:

Data Loading & Structure

We imported the Ames Housing training file (train.csv) into R (tidyverse + janitor) and standardized column names using clean_names(). After loading, the dataset contained 1460 rows × 81 columns. The target SalePrice is present and numeric. For later analysis we created log_price = log10(SalePrice) to reduce right skew.

3.2 Missingness Overview

We computed missing counts and percentages for every variable and saved the table as figures/missingness_summary.csv. This artifact documents data quality and guides imputation: numeric NAs will be imputed with the median; ordered quality fields that are NA due to absence (e.g., basement/kitchen quality) will map to a lowest/“None” level for descriptive plots but are excluded from PCA/k-means numerics. No rows were removed at this stage.



missingness_summary

variable	n_missing	pct_missing
pool_qc	1453	99.52
misc_feature	1406	96.3
alley	1369	93.77
fence	1179	80.75
fireplace_qu	690	47.26
lot_frontage	259	17.74
garage_type	81	5.55
garage_yr_blt	81	5.55
garage_finish	81	5.55
garage_qual	81	5.55
garage_cond	81	5.55
bsmt_exposure	38	2.6
bsmt_fin_type2	38	2.6
bsmt_qual	37	2.53
bsmt_cond	37	2.53
bsmt_fin_type1	37	2.53
mas_vnr_type	8	0.55
mas_vnr_area	8	0.55
electrical	1	0.07
id	0	0

Figure 4: “Missingness summary (variables by percent missing). Full table available as figures/missingness_summary.csv.”

Reproducibility: All outputs from this step (including the missingness CSV) are written to the project figures/ directory.

3.3 Target Distribution (Quick Summary)

SalePrice exhibits the expected right skew: min $34,900; Q1 $129,975; median $163,000; mean $180,921; Q3 $214,000; max $755,000. After log transformation, log10(SalePrice) is much closer to symmetric (min 4.543; Q1 5.114; median 5.212; mean 5.222; Q3 5.330; max 5.878). We will report relationships using log-price where appropriate because distances and correlations are more interpretable on the log scale.

EDA Price histograms (linear & log):

Overview:
Before building any models or computing distances, it is essential to understand the basic shape of the response variable. Figure 5 displays the raw distribution of SalePrice across 1,460 Ames homes. The mass of the distribution sits roughly between $120k and $220k, but the histogram exhibits a long right tail extending beyond $500k and up to $755,000. This is consistent with housing markets in which a relatively small number of high-end properties transact at prices that are several standard deviations above the median. The sample summaries confirm this skew: min $34,900; Q1 $129,975; median $163,000; mean $180,921; Q3 $214,000; max $755,000. The fact that the mean > median and that the upper tail is much longer than the lower tail indicates positive skew and the presence of influential observations.

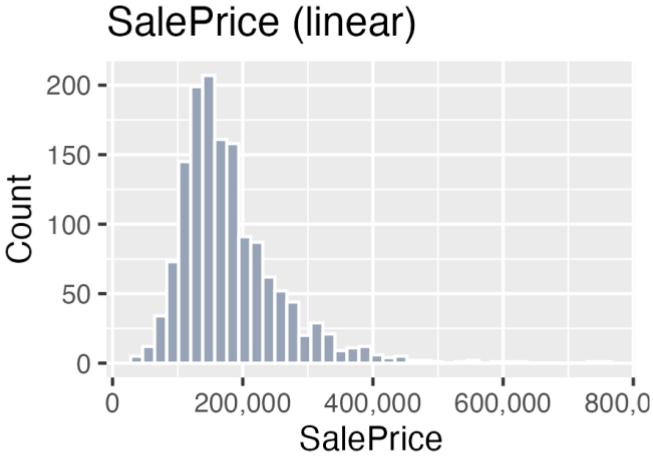


Figure 5: Sale Price (linear). Long right tail motivates transforming the target before correlation- and distance-based analyses.

Implications for analysis.

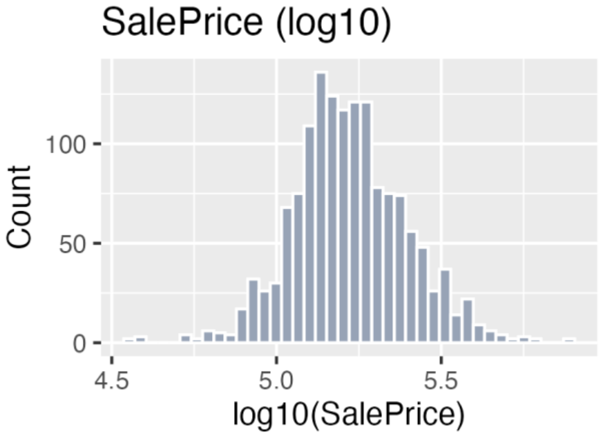


Figure 6: Sale Price (log10). Distribution is substantially closer to symmetric, making effect sizes and cluster separation easier to interpret.

Positive skew matters for two reasons. First, many of the techniques used later in this project—correlation heatmaps, PCA, and k-means clustering—rely either explicitly or implicitly on Euclidean geometry and variance. When the response is highly skewed, differences among high-priced homes dominate distances, while differences among mid-priced homes are relatively compressed. Second, relationships between price and size/quality tend to be multiplicative rather than additive: a 10% increase in living area is associated with a roughly proportional (percentage) change in price, not a constant dollar change. Modeling on the original dollar scale obscures that structure.

Effect of the transform:
Figure 6 shows the distribution after applying a base-10 logarithm to SalePrice, log10(SalePrice). The visual change is substantial: the distribution is now approximately symmetric with a single mode. The corresponding summaries—min 4.543; Q1 5.114; median 5.212; mean 5.222; Q3 5.330; max 5.878—indicate that location and spread are now much more balanced (the mean and median are nearly equal). Interpreting the log scale is straightforward: the median of 5.212 corresponds to roughly 105.212 ≈ $163k; a difference of 0.1 on the log10 scale is about a 26% change in price (100.1≈1.26). This interpretation aligns with elasticity-style reasoning in hedonic pricing.

Robustness and comparability:
Working on the log scale reduces the influence of extreme luxury homes on summaries, correlations, and distances. For example, when computing PCA on standardized features, the component loadings will not be driven

primarily by a handful of very expensive properties. Likewise, in k-means, cluster centroids measured in log-price space represent typical multiplicative differences among market segments rather than being pulled toward outliers. This choice also improves comparability across figures: scatterplots of living area vs. log-price will show a more linear trend, and boxplots of log-price by OverallQual will display cleaner separation with fewer extreme whiskers.

Practical takeaway for the project:
All downstream analyses that involve relationships with price—scatterplots, correlation summaries, PCA, and clustering—will use log10(SalePrice) unless stated otherwise. Final narrative results (e.g., cluster profiles) will still be reported back in dollar terms for readability, but the analytical pipeline operates on the transformed target to improve stability and interpretability. The raw-scale histogram (Figure 5) is retained for context: it conveys the actual monetary range that stakeholders care about, while Figure 6 justifies the statistical treatment used to extract structure from the data.

Limitations:
Log transformation assumes positive prices (satisfied here) and interprets differences multiplicatively; in markets with price floors/ceilings or strong discontinuities, additional transforms (e.g., Box-Cox with estimated λ) could be considered. However, the standard log transform is widely adopted for housing data and adequately addresses the skewness observed in Ames.