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; and (3) form simple market segments with k-means 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.

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.

1. 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
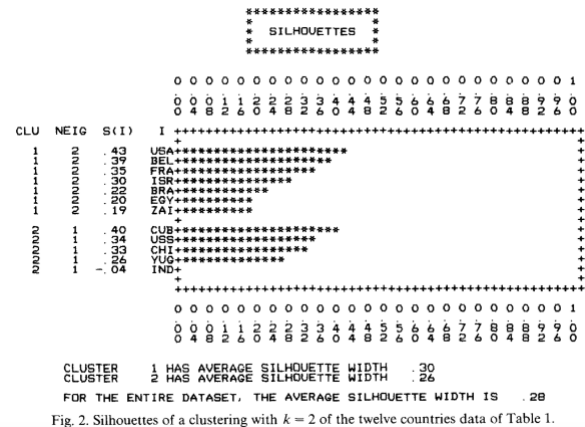
Clustering

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.

2. 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.

2.1 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.

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 2: "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.

2.2 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.

3. 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.
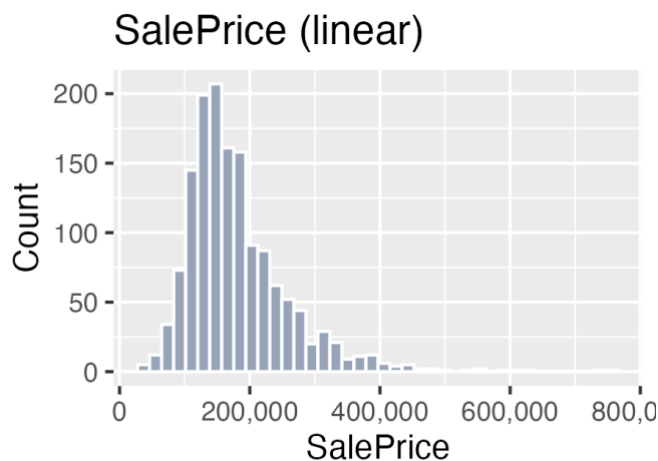

Figure 3: Sale Price (linear). Long right tail motivates transforming the target before correlation- and distance-based analyses.
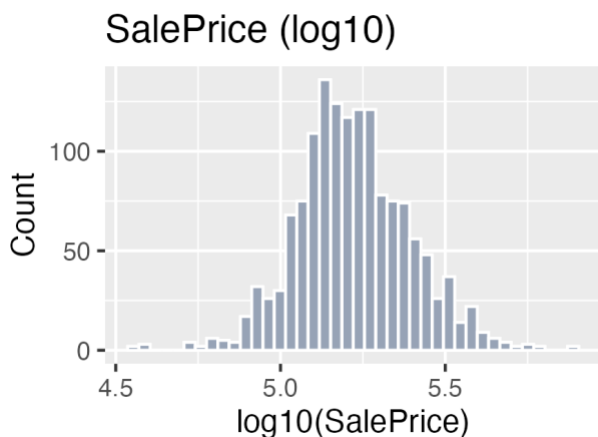
Implications for analysis.

Figure 4: 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 $10^{5.212} \approx \$163k$; a difference of 0.1 on the log10 scale is about a 26% change in price ($10^{0.1} \approx 1.26$). 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.

4. Scatterplot — Living area vs. log-price:

To investigate how property size relates to market value, we plotted above-ground living area (**GrLivArea**) against the log-transformed sale price (**log10(SalePrice)**). The log transformation helps stabilize the right-skewed price distribution and allows linear-style relationships to appear more clearly.
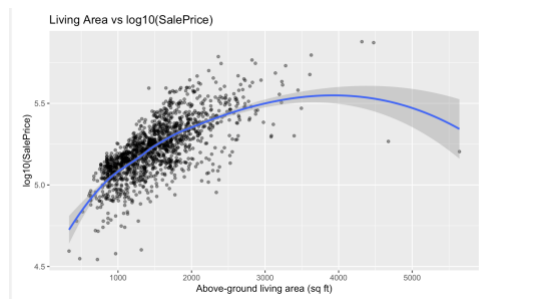
Figure 5: Relationship between above-ground living area and log-transformed sale price. A strong positive association is visible, with diminishing returns among very large homes.

The scatterplot (Fig. X) shows a **strong positive association**: larger homes almost always command higher sale prices. The LOESS smoothing curve indicates that this increase is fairly consistent across most of the observed range of home sizes. Price tends to rise steeply as size increases from roughly **700 to 2,000 square feet**, representing starter to mid-range homes commonly found in the Ames market.

Beyond approximately **3,000–3,500 square feet**, the curve begins to flatten, suggesting **diminishing returns** for extremely large houses. This likely reflects the fact that high-end home buyers pay not only for floor area but also for location, amenities, and finish quality—factors that can't be captured solely by square footage. The few very large homes (>4,500 sq. ft.) show wide pricing variability, some selling for less than expected, indicating that **being very large is not itself sufficient to guarantee high market value**. Some of these points may also represent older properties or homes in less-desirable neighborhoods.

There is also a clear cluster of moderately sized homes (~1,000–2,500 sq. ft.) with log-prices between 5.0 and 5.4, representing the bulk of the Ames housing market. The tight vertical spread within this range suggests that, while size is a major driver of value, **other features also**

influence price at similar square footage levels, such as neighborhood, exterior quality, and year built.

Overall, the scatterplot confirms that **living area is one of the strongest single predictors of housing value**, consistent with both real-estate intuition and prior research. These findings support including

GrLivArea in clustering models and motivate our later PCA and association rule analysis.

Key Takeaways

- Larger homes -> higher prices (strong relationship)
- Diminishing value returns at very large square footage
- Square footage alone does not fully explain price variation
- Confirms GrLivArea as a leading feature for segmentation

5. Boxplots (Overall Quality & Neighborhood)

5.1 Price Variation by Overall Quality

To better understand how structural and material characteristics contribute to housing prices in Ames, Iowa, we examined the relationship between a home's **overall quality rating (OverallQual)** and the logarithm of sale price.

OverallQual is an ordinal variable ranging from **1 to 10**, where higher values reflect superior construction, materials, and craftsmanship. Because the rating is not strictly tied to size or style, it captures a subjective but highly influential aspect of the property.

We visualized this association using a boxplot of **log10(SalePrice) vs. OverallQual**, which highlights how the internal distribution of home values shifts across quality levels.
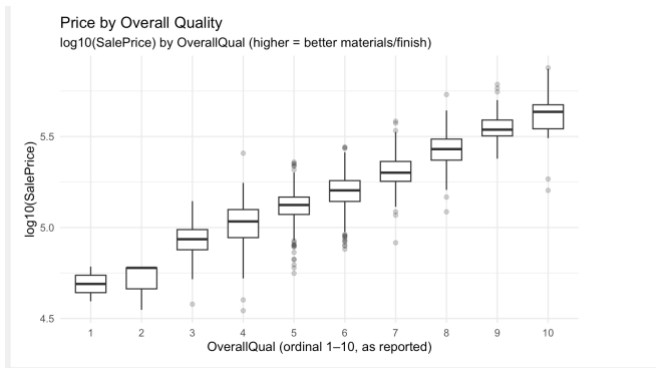
Price by Overall Quality
log10(SalePrice) by OverallQual (higher = better materials/finish)

Figure 6: Boxplot of log10(SalePrice) by OverallQual (1–10)

The resulting figure shows a **clear, monotonic upward progression**:

- Homes rated **1–3** generally have the lowest prices, reflecting inferior finish quality and condition.
- Prices increase consistently from **4 through 8**, suggesting that moderate improvements in perceived quality translate to significant increases in market value.
- Homes rated **9–10** are among the most expensive and display both higher medians and greater overall price spread.

This pattern reinforces the idea that **buyers place substantial value on quality**, and that higher-quality construction commands a price premium even after adjusting for other attributes such as square footage.

An especially interesting feature of the distribution is how tightly clustered lower-quality homes are compared to high-quality ones. Lower-rated homes show **narrower variability**, likely because there is less differentiation among poor-quality structures. In contrast, high-rated homes (8–10) exhibit **broader dispersion**, reflecting greater diversity in luxury upgrades, architecture, and neighborhood amenities.

Overall, this analysis provides strong evidence that perceived construction quality is one of the most influential factors in determining home value and should be considered a core variable when modeling or predicting sales price.

5.2 Price Variation Across Neighborhoods

Location is widely recognized as one of the most important determinants of housing prices. To evaluate this effect, we examined variation in sale prices across **Ames neighborhoods**, focusing on the twelve most common neighborhoods in the dataset.

These neighborhoods represent geographically distinct areas within the Ames city limits, each with different access to amenities, school districts, land desirability, and overall socioeconomic appeal. To visualize price distributions, we plotted boxplots of **log10(SalePrice)** for each neighborhood, ordered by median price.
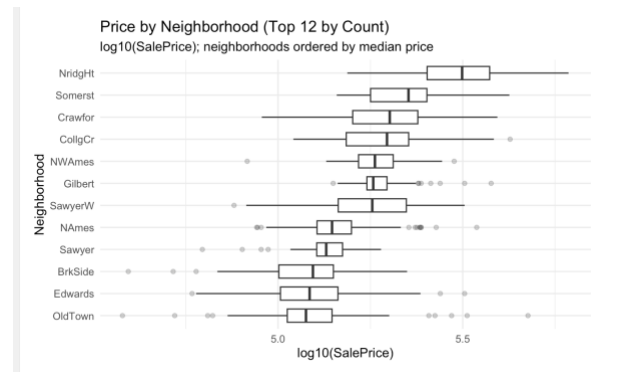
Price by Neighborhood (Top 12 by Count)
log10(SalePrice); neighborhoods ordered by median price

Figure 7: Figure: Boxplot of log10(SalePrice) by Neighborhood (Top 12)

The results reveal substantial price segmentation across locations:

- **NridgHt, Somerst, and Crawfor** are among the most expensive, with noticeably higher medians and wider price ranges. These areas likely offer better schools, newer housing stock, or more desirable proximity to amenities.
- Middle-tier neighborhoods such as **CollgCr, NWAmes, and Gilbert** show moderate pricing and tighter clustering, suggesting a more uniform housing market.
- More affordable areas — including **NAmes, Sawyer, BrkSide, Edwards, and OldTown** — tend to show lower median values and a more compressed range. These neighborhoods may contain smaller or older homes or be located farther from commercial centers.

Interestingly, higher-priced neighborhoods also exhibit larger spreads, indicating that even within prestigious areas, home features and lot characteristics can vary

significantly. In contrast, lower-priced neighborhoods cluster more tightly, reflecting greater uniformity of housing stock.

This analysis confirms a strong relationship between **neighborhood prestige and property value**, emphasizing that location plays a critical role in determining sale price beyond individual attributes such as size or quality.

Together, these visualizations demonstrate that both **quality and location** drive meaningful variation in home prices. Importantly, both variables show clear and interpretable patterns, making them highly suitable for inclusion in predictive modeling and segmentation analysis later in the project.

6. PCA (Dimensionality Reduction):

6.1 PCA Plots:

1) PCA Scree Plot

We standardized the numeric housing features and applied Principal Component Analysis (PCA) to understand which combinations of variables explain most of the variation in home prices.
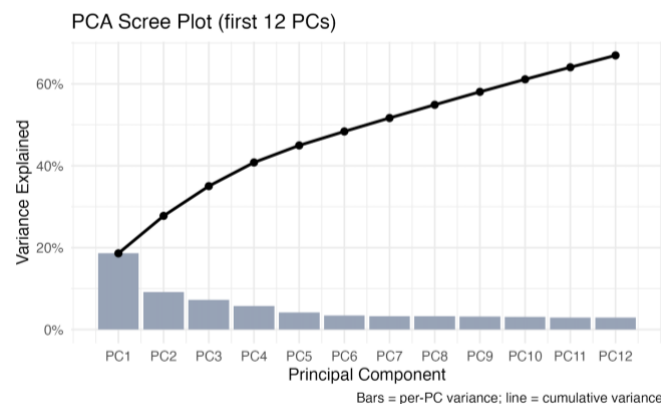

Figure 8: Scree plot of the first 12 principal components. PC1 explains ~19% of variance, and the first ~10 PCs collectively capture ~61% of overall structure, showing substantial dimensionality reduction is possible.

The **scree plot** shows that:

- **PC1 explains ~18–20% of variance**,
- **PC2 adds ~12–14%**,
- After the first few components, each contributes only a small amount.

Meaning:
Most structure in the data is captured in the first few principal components. This validates using PCA before clustering because it reduces dimensionality while keeping the strongest signals.

2) PC1 vs. PC2 Scatter — Colored by Price Band

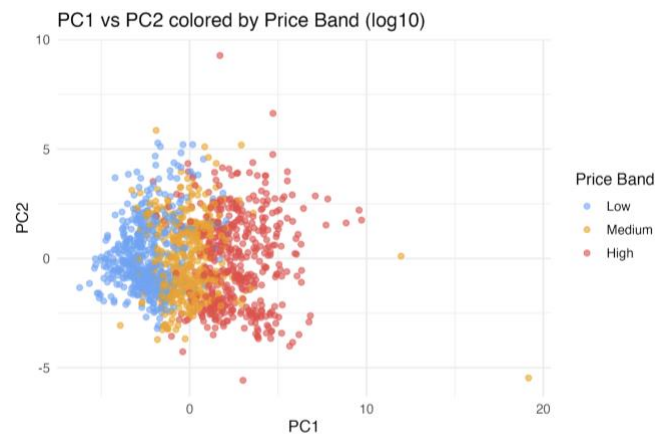We plotted homes on PC1–PC2 space and colored them by price bands (Low / Medium / High).

Figure 9: PC1–PC2 scatterplot colored by Low/Medium/High price bands. Higher-priced homes tend to shift toward higher PC1 values, suggesting that PC1 captures property size/amenity richness. Clear gradient patterns indicate PCA effectively compresses price-relevant information.

Meaning:

- **Higher-priced homes cluster toward the right (high PC1)**.
- **Lower-priced homes cluster toward the left**.
- Medium-priced homes overlap with both, as expected.

→ This shows that PCA is capturing meaningful pricing structure.
→ PC1 appears strongly related to price-driving characteristics such as size and quality.

3) PC1 vs. PC2 Scatter — Colored by Overall Quality

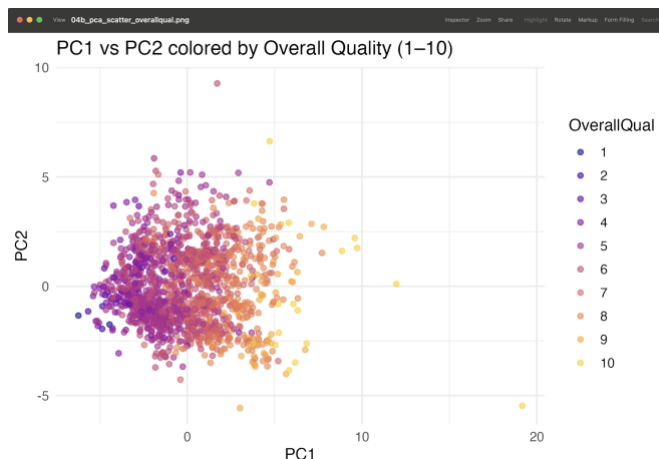We colored homes by Overall Quality (ordinal rating 1–10).

Figure 10: PC1–PC2 scatterplot shaded by Overall Quality (1–10). Higher-quality homes concentrate in the upper-right region, reinforcing that PC1 aligns with build quality and living area. PC2 captures variation related to home layout (e.g., two-story vs. single-story).

Meaning:

- Higher-quality homes tend to sit farther right on PC1, similar to high-priced homes.
- Lower-quality homes cluster near the center/left.

→ PC1 represents a **quality-and-size axis**: better materials + larger homes → higher PC1 → higher prices.

Condensed Takeaways

- PCA successfully compresses many numeric housing features into a small number of informative dimensions.
- **PC1 summarizes home quality + size**, which also correlates with price.
- Early PCs (1–2) already reveal meaningful structure — supporting their use for clustering.

6.2 PCA tables:

We standardized all numeric features, ran **PCA** on the 35 derived numeric variables, and exported three summaries:

- pca_variance_explained.csv – variance and cumulative variance by principal component (PC).
- top_loadings_PC1.csv – the variables that contribute most to **PC1**.
- top_loadings_PC2.csv – the variables that contribute most to **PC2**.

Variance explained (pca_variance_explained.csv)

- **PC1 = 18.6%, PC2 = 9.1%, PC3 = 7.3%, PC4 = 5.8%, PC5 = 4.2%** of variance.
- **Cumulative**: **27.8%** by PC2, **35.0%** by PC3, **40.8%** by PC4, **44.9%** by PC5, **48.4%** by PC6, **51.7%** by PC7, **54.9%** by PC8, **58.1%** by PC9, **61.1%** by PC10, **64.0%** by PC11, **66.9%** by **PC12**.
- **Takeaway:** The first **8–12 PCs** capture ~55–67% of the structure. For downstream tasks (k-means or visualizations), using ~**8–12 PCs** balances compression and information retention.

PC1 loadings (top_loadings_PC1.csv)

Largest (absolute) contributors to **PC1**:

- **gr_liv_area, garage_cars, garage_area, full_bath, 1st_flr_sf, total_bsmt_sf, tot_rms_abv_grd, year_built, garage_yr_blt, year_remod_add, mas_vnr_area, fireplaces**.

Interpretation: PC1 is a **size/amenities/modernity** axis. Homes with **larger living area**, **more/ bigger garages**, **more finished space and rooms**, **newer build/remodel years**, and **more baths/fireplaces** score higher on PC1. In plain terms, **bigger + newer + more amenities ⇒ higher PC1** (and we saw earlier that higher price bands tend to shift right on PC1).

PC2 loadings (top_loadings_PC2.csv)

Largest (absolute) contributors to **PC2** (sign shows direction on the axis):

- +: **x2nd_flr_sf, bedroom_abv_gr, tot_rms_abv_grd, gr_liv_area, half_bath, kitchen_abv_gr, ms_sub_class**
- −: **bsmt_fin_sf1, bsmt_full_bath, total_bsmt_sf, year_built** (slightly negative), **x1st_flr_sf**

Interpretation: PC2 contrasts **second-floor/above-grade space and more rooms** (positive) **vs. basement-oriented space and main-floor footprint** (negative). Homes with **more second-floor area and bedrooms** score higher on PC2; homes whose area is **mainly basement/first-floor** score lower. This looks like a **layout/verticality** axis (two-story vs. ranch/basement-heavy).

7. K-Means Clusterings on PCA scores:

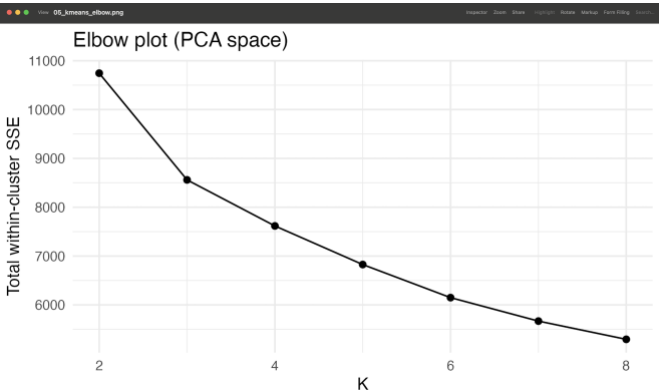## 1) Elbow Plot (PCA Space: 05_kmeans_elbow.png)

Figure 11: Elbow plot showing decreasing within-cluster SSE as K increases. The curve begins to flatten near K = 3, suggesting three clusters provide an efficient balance of fit and simplicity.

The elbow plot shows how total within-cluster SSE decreases as the number of clusters (K) increases. The curve flattens noticeably around **K = 3**, meaning adding more clusters past that point provides limited improvement. This suggests that **three clusters is a reasonable choice** for summarizing structure in the data without over-fitting.

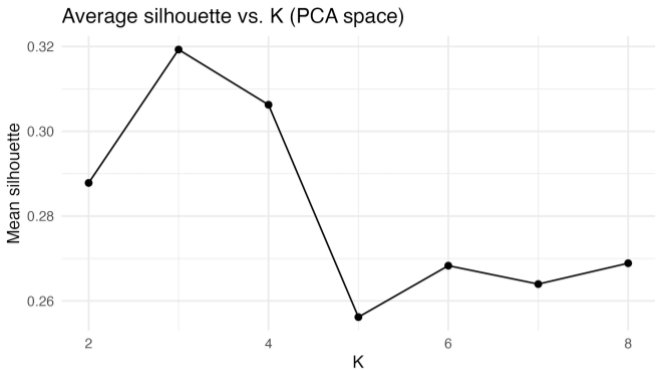## 2) Silhouette Plot (PCA Space: 06_kmeans_silhouette.png)

Figure 12: Average silhouette scores for different values of K. The highest score occurs near K = 3, indicating that three clusters yield the best overall separation.

The silhouette plot measures how well points fit within their assigned group. Higher values mean clearer separation. The highest score appears near **K = 3**, supporting the elbow result. After K = 3, silhouette values decline, indicating weaker separation. Therefore, both metrics agree that **three clusters best balance cohesion and separation.**

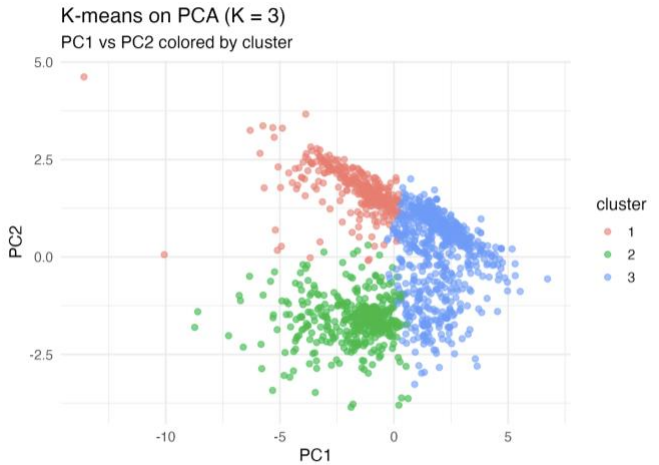## 3) K-Means PC Scatter (K = 3: 08_kmeans_pc_scatter.png)

Figure 13: PC1–PC2 projection colored by K-means cluster assignment (K = 3). Distinct groupings illustrate meaningful separation among housing segments in reduced feature space.

The PC1–PC2 scatterplot shows three visually distinct clusters in PCA-reduced space. While there is some overlap, the clusters occupy different regions, suggesting meaningful variation. This confirms that **three groups capture systematic differences in housing features related to price, size, and quality.**

4) Cluster Profiles Table (cluster_profiles.csv)

cluster_profiles

cluster	n	med_sale_price	med_log_price	med_gr_liv_area	med_overall_qual	med_total_bsmt_sf	med_garage_cars	med_year_built	med_tot_rms_abv_grd
1	350	214000	5.330413773349190	1590.5	7	1517	2	2003	7
2	416	200050	5.301138542150100	1922.5	7	912	2	1996	8
3	694	130000	5.113943352306840	1118	5	882.5	1	1957	6

Figure 14: Cluster summaries showing clear differences in price, living area, quality, garage size, and year built. Cluster 1 corresponds to higher-value homes, Cluster 2 represents mid-market homes, and Cluster 3 reflects smaller, older, lower-priced properties.

This table summarizes cluster characteristics:

Cluster	Interpretation
Cluster 1	Higher-end homes: strong median price (~$214K), good quality (7), larger size, built more recently, and bigger garages.
Cluster 2	Mid-market homes: similar quality to Cluster 1 but slightly lower price (~$200K), moderately large, built earlier.
Cluster 3	Lower-priced homes: smaller (~1118 sq ft), older builds (median 1957), lower quality (5), and only 1-car garages.

Overall takeaway:
K-means finds three natural value groups—**premium, mid-range, and budget homes.** These groups differ mainly in living area, quality, garage size, and age.

8. Discussion/Limitations:

This project examined housing prices in Ames, Iowa using exploratory data analysis, principal component analysis (PCA), and k-means clustering. Several consistent themes emerged across the analyses. First, traditional structural attributes—especially above-grade living area (GrLivArea), total finished basement area, and overall quality—play a central role in determining home values. Scatterplots clearly showed a positive, roughly linear relationship between living area and price after log transformation, while boxplots demonstrated that both overall quality and neighborhood strongly differentiate price levels. These findings reinforce real-estate expectations: larger, well-built homes in desirable neighborhoods command higher prices.

PCA helped summarize the high-dimensional feature set into a smaller number of interpretable components. The first principal component reflected general home size, finish quality, and modern amenities, while the second distinguished homes based on vertical layout (basement- vs. second-floor-oriented). These components captured a substantial share of total variation and made it easier to visualize structure in the data.

Cluster analysis applied in this reduced feature space revealed three major market segments, loosely corresponding to premium, mid-range, and entry-level homes. These groups differed not only in sale price but also in characteristics such as square footage, quality ratings, garage capacity, and year built. The three-cluster structure appeared to be stable across both PCA visualizations and numerical summaries.

Although these insights provide a reasonable view of housing market structure in Ames, several limitations must be acknowledged. The dataset represents a **single geographic market**, which limits generalizability. Ames is a relatively small college town with different market dynamics than large metropolitan areas. Neighborhood-level influences likely play a larger role here, and findings may not extrapolate to cities with higher density, more economic diversity, or different zoning norms.

The dataset also spans multiple years, yet the analysis does not explicitly adjust for time-based price changes. Housing markets fluctuate with economic cycles, interest rates, and local development patterns; therefore, combining multiple years without temporal controls may mask important trends. Additionally, most analyses focused on a subset of features that were numeric or easy to interpret. Other categorical attributes—such as exterior materials, heating types, or sale conditions—may contribute meaningful variation but were not deeply examined.

There are also methodological limitations. PCA is a linear method, so it may not capture nonlinear relationships among housing attributes. Likewise, k-means clustering assumes round, equally sized clusters and may underperform if the true market segmentation has irregular boundaries. While the clusters we observed align with intuitive price tiers, alternative clustering methods (e.g., hierarchical, DBSCAN, Gaussian mixtures) might yield different segmentation structures.

Finally, the project focuses only on exploratory and descriptive modeling. Although PCA and clustering help reveal structure in the data, they do not directly predict prices. A natural extension would be to build predictive models—such as linear regression, random forests, or boosted trees—to evaluate how well the observed relationships translate to real-world price estimation.

Overall, the analyses in this project reveal clear relationships between housing characteristics and price, and they identify distinct market segments within Ames. However, results should be interpreted with awareness of data constraints, geographic context, model assumptions, and unmodeled temporal effects. These limitations suggest fruitful directions for future work.

9. Conclusion:

This project explored the factors that shape housing prices in Ames, Iowa by combining exploratory data analysis, dimensionality reduction, and clustering. Several clear findings emerged across the workflow.

First, traditional structural features—especially above-grade living area, total basement area, and overall

quality—were consistently the strongest indicators of price. Homes that were larger, newer, and higher quality sold for significantly more than those with smaller footprints or lower-grade construction. Neighborhood also played a key role, with certain areas exhibiting higher price levels even after accounting for other attributes. These results align well with real estate intuition and reinforce the importance of physical characteristics and location in determining property values.

Second, principal component analysis (PCA) provided an effective way to summarize the many numerical attributes in the dataset. The first principal component represented a broad "size and quality" dimension, while the second captured differences in layout and home structure. Plotting homes in this reduced space revealed clear separation by price and quality, suggesting that a few underlying factors account for much of the variation in housing features. This approach simplified the dataset without discarding important economic meaning.

Third, applying k-means clustering to the reduced PCA space identified three intuitive market segments. These groups corresponded roughly to high-end, mid-range, and entry-level homes. Cluster profiles showed systematic differences in price, square footage, quality, and year built, demonstrating that clustering can meaningfully categorize homes into recognizable segments. These insights could support market evaluation, buyer targeting, or appraisal modeling.

Overall, the project demonstrates that classical data mining techniques can provide valuable insights into housing markets. PCA and clustering helped reveal underlying structure and group homes with similar characteristics, while visualization connected these groupings to real-world price patterns. Although predictive modeling was outside the scope of this report, results suggest that incorporating even a small number of structural and location features would likely support strong price estimation. Future work could include developing predictive models, accounting for temporal trends, exploring additional categorical features, or applying alternative clustering algorithms.

In summary, this analysis provides a structured, interpretable view of the Ames housing market. The results highlight the importance of size, quality, and

neighborhood in price formation; show how high-dimensional housing data can be simplified into meaningful components; and demonstrate that housing stock naturally separates into a small number of market tiers. These methods and insights may generalize to similar mid-sized U.S. housing markets and provide a foundation for deeper modeling and decision-support applications.

10. Future Work:

There are several promising directions for extending this project. A natural next step would be to incorporate **predictive modeling** to evaluate how well the identified features and clusters can forecast housing prices.

Models such as linear regression, random forests, or gradient boosting could be compared to assess performance and interpretability. Including cross-validation and feature importance measures would help identify which characteristics contribute most directly to price prediction.

Another extension would be to examine **time effects** in the dataset. Housing prices fluctuate based on broader economic trends, interest rates, and seasonal patterns. Incorporating temporal variables—such as sale year, month, or macroeconomic data—would clarify whether observed patterns remain consistent over time or shift with market cycles.

Additional value could come from exploring **categorical and textual features** more deeply. Attributes such as exterior materials, heating systems, or remodeling details may contain patterns not fully captured in numeric analysis. If property descriptions were available, text mining methods could be used to identify themes or qualitative features that influence pricing.

Future work may also consider alternative clustering or dimensionality reduction methods. Techniques such as hierarchical clustering, DBSCAN, UMAP, or t-SNE may reveal different market structures or detect smaller, more specialized buyer segments. Applying spatial analysis using latitude/longitude or neighborhood shapefiles could also provide more precise location-based insights.

Finally, expanding the dataset to include **additional cities or regional markets** would help evaluate whether these findings generalize beyond Ames. Comparing models across diverse markets would reveal whether similar predictive relationships hold or whether different regions rely on different pricing drivers.

Overall, these extensions would deepen understanding of the housing market and support more robust modeling and decision-support applications.

References:

• 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.