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

Daniel Phelps — 12 December 2025

Email: dphelps9693@floridapoly.edu | 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):

We produced a set of exploratory plots and summary tables to understand the distribution of sale prices and their relationships with key predictors. The distribution of SalePrice was examined on both the raw and log scales to justify the use of a log transformation and to identify extreme values that may require special handling. A scatter plot of GrLivArea versus SalePrice on a log scale highlighted the strong positive relationship between living area and price, while also revealing a small number of unusually large homes that behave as outliers. Box plots of SalePrice by Neighborhood and by OverallQual were used to visualize systematic shifts in the center and spread of prices across categories, making clear where large differences in typical sale price occur. A correlation heatmap of numeric predictors helped identify redundancy among variables and guided later PCA and feature selection decisions. Finally, a compact summary table reporting N, mean, median, and IQR for SalePrice and several top predictors provided a concise numeric snapshot of the dataset.

Each figure is accompanied by a short narrative of two to three sentences describing what the data show, such as the monotonic increase in price with higher OverallQual or neighborhoods that shift the median sale price by a large margin. These descriptive insights are not ends in themselves; they directly informed the choice of variables used in PCA and shaped the subsequent clustering analysis.

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:

We compute principal components analysis on the correlation matrix, which is equivalent to centering and scaling each feature before applying PCA. A scree plot and cumulative variance plot are reported, and we retain the first kkcomponents that together explain roughly 70–85% of the total variance, with kk expected to be between two and four. Component loadings are then inspected to interpret the meaning of each axis, such as distinguishing a "size and amenities" dimension from an "age and condition" dimension. Finally, we produce a two-dimensional PCA scatter plot colored by price band (Low, Medium, High), where bands are defined by tertiles of SalePrice. Optionally, ellipses are added to summarize the spread of each price group in the PCA space.

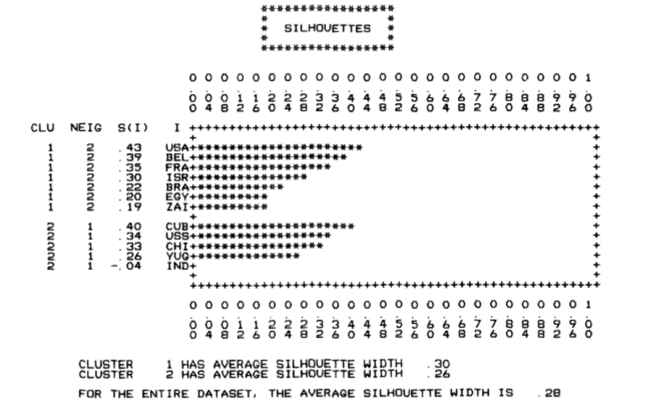
What I will conclude:

This analysis identifies which combinations of measurements form the dominant axes of variation in the housing data. It also shows whether low-, medium-, and high-price homes separate in the PCA plane, which helps assess whether clustering on these standardized variables is reasonable. In addition, the loadings reveal which original features contribute most strongly to the retained components, guiding the selection of variables to carry forward into k-means clustering.

Deliverables:

The outputs from this step include a scree plot, a loading table reporting the largest absolute loadings for each retained component, and a two-dimensional PCA scatter plot.

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 clustering uses the same standardized numeric variables that fed into PCA, with the option to include one or two binary indicators (such as CentralAir) if they provide clear separation. All included variables are standardized so that no single measurement dominates the distance calculation simply because it is measured on a larger scale.

Choosing K:

We run k-means for K=2 through 8 using nstart = 25 to reduce sensitivity to random initialization. For each K, we compute the average silhouette width using Euclidean distance on the standardized feature set. We then select the KK with the highest silhouette score, breaking ties (or near-ties) by preferring the simpler solution with fewer clusters. We reference Fig. 1 when explaining and justifying this choice in the write-up.

Cluster assignment and profiling:

After choosing K, we fit the final k-means model and assign each home a cluster label. We then build a cluster profile table that summarizes each segment using practical, report-friendly statistics: cluster size (N), median SalePrice (and median log price), typical GrLivArea, median OverallQual, and the top three neighborhoods by share. Each cluster is also given a short descriptive sentence (for example, "Large, higher-quality homes concentrated in Neighborhoods A/B"). To communicate the differences quickly, we include a bar chart comparing key medians across clusters.

Quality and stability checks:

We inspect within-cluster variation in price and living area to confirm the clusters are not being driven by a handful of extreme outliers. If time allows, we also run a light stability check by re-fitting k-means on several 80% subsamples and verifying that the cluster profiles remain broadly consistent. This step is optional, but it helps confirm that the segments represent real structure in the data rather than artifacts of sampling noise.

Outcome:

The end result is a small set of clear, interpretable housing segments that can be discussed directly in the final presentation, such as "Entry-level small homes," "Mid-tier family homes," and "Large premium homes." These segments also provide useful context for interpreting the association rules by connecting patterns in features to meaningful housing market groups.

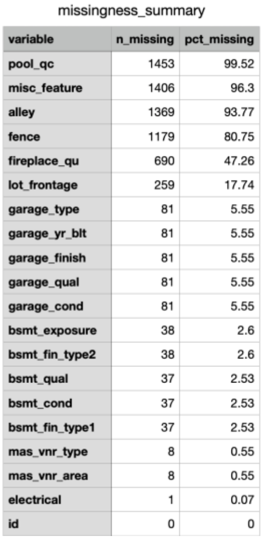
2. Data Setup:

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.



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 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.

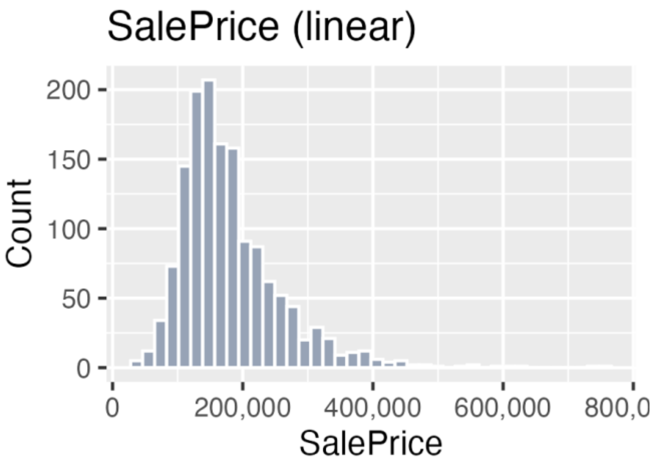


Figure 3: Sale Price (linear). Long right tail motivates transforming the target before correlation- and distance-based analyses.

Implications for analysis.

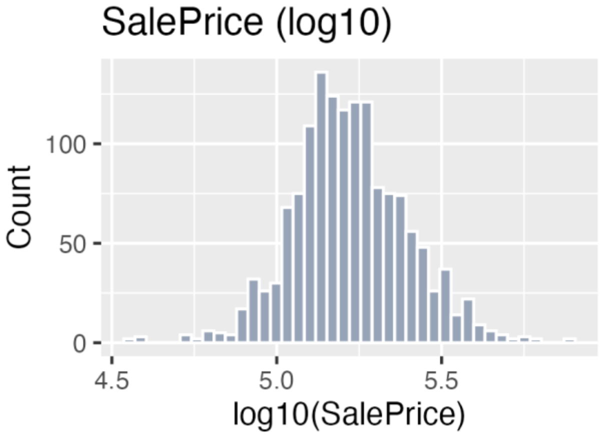


Figure 4: Sale Price (log10). Distribution is substantially closer to symmetric, making effect sizes and cluster separation easier to interpret.

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 105.212 ≈ $163k; a difference of 0.1 on the log10 scale is about a 26% change in price. 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.

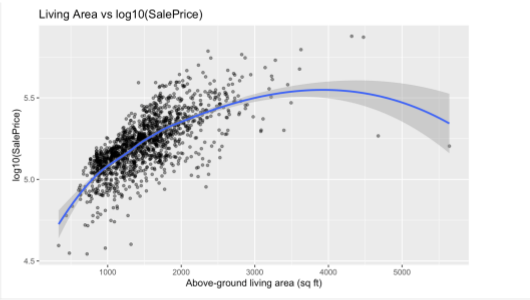


Figure 5: Relationship between above-ground living area and log-transformed sale price. A strong positive association is visible, with diminishing returns among very large homes.

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, which aligns with real estate intuition and prior research. The clear positive relationship between GrLivArea and SalePrice, along with visible nonlinearity at very large sizes, justifies emphasizing living area in the clustering models and motivates its inclusion in the later PCA and association rule analysis.

Key Takeaways:
Larger homes are associated with higher prices, indicating a strong and consistent relationship between square footage and value. However, the plot also shows diminishing returns at very large living areas, where additional square footage adds less incremental value. At the same time, substantial price variation remains at similar square footage levels, indicating that size alone does not fully explain housing prices. Together, these patterns confirm GrLivArea as a leading feature for segmentation, but not a sufficient one on its own.

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.

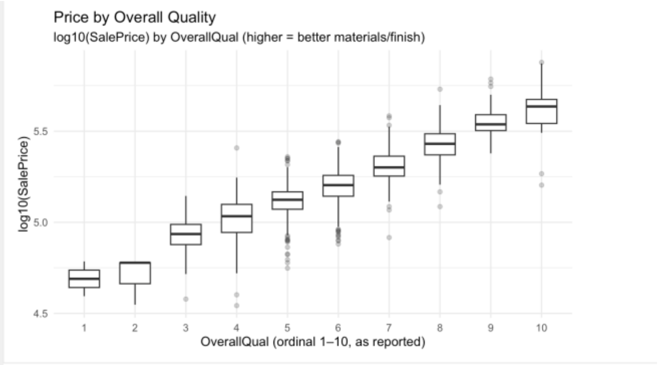


Figure 6: Boxplot of log10(SalePrice) by OverallQual (1–10)

The resulting figure shows a clear, monotonic upward progression in housing prices as overall quality increases. Homes with ratings of 1–3 have the lowest prices, reflecting inferior construction quality and condition. Prices rise steadily from ratings 4 through 8, indicating that even moderate improvements in perceived quality translate into substantial gains in market value. Homes rated 9–10 sit at the top of the price distribution and exhibit both higher median prices and greater overall spread, suggesting that premium quality homes command consistently higher values but also show more variation at the high end of the market.

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.

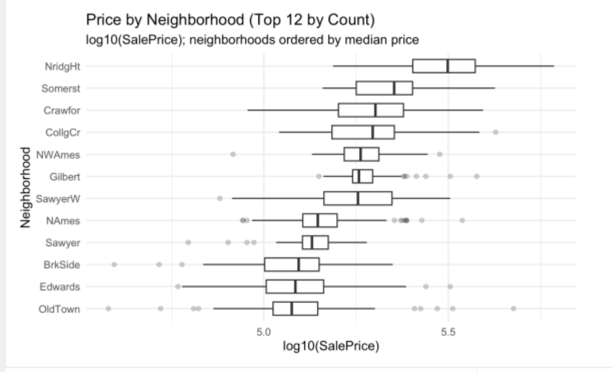


Figure 7: Figure: Boxplot of log10(SalePrice) by Neighborhood (Top 12)

The results show clear price segmentation across neighborhoods. Areas such as NridgHt, Somerst, and Crawfor are among the most expensive, with noticeably higher median prices and wider spreads, likely reflecting newer housing stock, stronger school districts, or more desirable access to amenities. Middle-tier neighborhoods like CollgCr, NWAmes, and Gilbert exhibit moderate prices with tighter clustering, suggesting a relatively homogeneous housing market. In contrast, more affordable neighborhoods, including NAmes, Sawyer, BrkSide, Edwards, and OldTown, tend to have lower median prices and more compressed ranges, which may reflect smaller or older homes or locations farther from major commercial and employment 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.

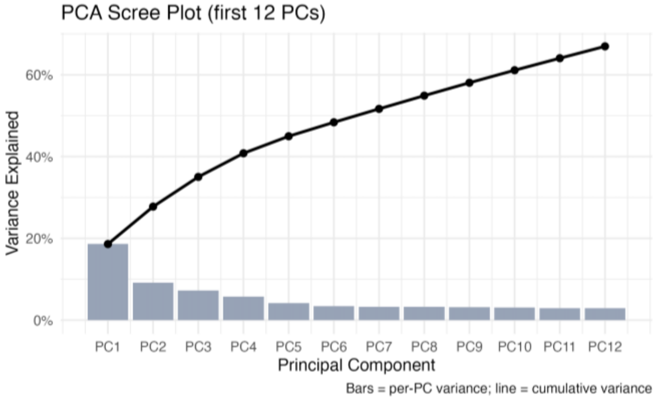


Figure 8: Scree plot of the first 12 principal components. PC1 explains ~19% of variance, and the first ~10 PCs collectively capture ~61% of overall structure, showing substantial dimensionality reduction is possible.

The scree plot shows that the first principal component explains roughly 18–20% of the total variance, while the second adds an additional 12–14%. Beyond the first few components, each subsequent component contributes only a relatively small amount of explained variance.

This pattern indicates that most of the meaningful structure in the data is concentrated in the early principal components. As a result, using PCA before clustering is well justified, since it reduces dimensionality while preserving the strongest and most informative signals in the dataset.

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).

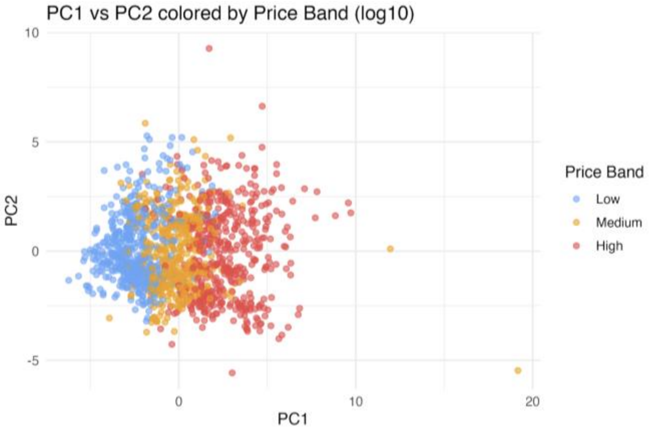


Figure 9: PC1–PC2 scatterplot colored by Low/Medium/High price bands. Higher-priced homes tend to shift toward higher PC1 values, suggesting that PC1 captures property size/amenity richness. Clear gradient patterns indicate PCA effectively compresses price-relevant information.

The PCA scatter indicates that higher-priced homes tend to cluster toward the right side of the plot, corresponding to high values of PC1, while lower-priced homes cluster toward the left. Medium-priced homes overlap with both groups, which is expected given their intermediate characteristics.

This pattern shows that PCA is capturing meaningful structure related to housing prices. PC1 appears to be strongly associated with price-driving characteristics such as home size and overall quality, reinforcing its usefulness for downstream clustering and segmentation.

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

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

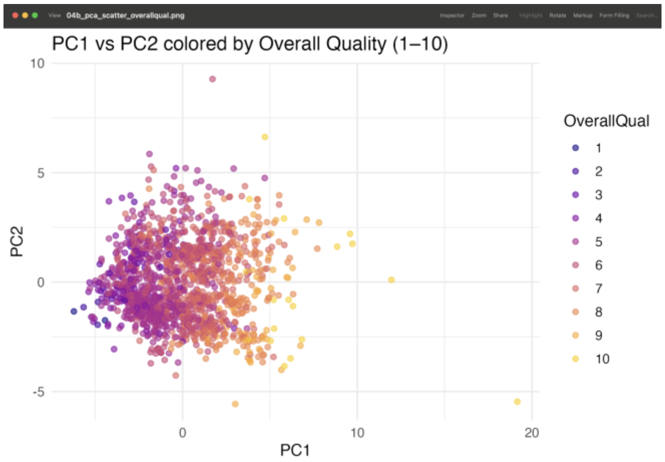


Figure 10: PC1–PC2 scatterplot shaded by Overall Quality (1–10). Higher-quality homes concentrate in the upper-right region, reinforcing that PC1 aligns with build quality and living area. PC2 captures variation related to home layout (e.g., two-story vs. single-story).

Meaning:

Higher-quality homes tend to appear farther to the right along PC1, mirroring the pattern observed for high-priced homes, while lower-quality homes cluster closer to the center or left of the axis. This alignment suggests that PC1 represents a combined quality-and-size dimension, where better construction and larger living areas translate into higher PC1 values and, in turn, higher prices.

Condensed Takeaways

PCA effectively compresses a large set of numeric housing features into a small number of informative dimensions. The first principal component captures a blend of home quality and size that is strongly correlated with price. Even the early components, particularly PC1 and PC2, reveal meaningful structure in the data, supporting their use as inputs for clustering and segmentation.

6.2 PCA tables:

We standardized all numeric features, ran PCA on the 35 derived numeric variables, and exported three summary outputs. The file pca_variance_explained.csv reports the variance explained and cumulative variance for each principal component, while top_loadings_PC1.csv and top_loadings_PC2.csv list the variables with the largest contributions to the first two components.

Variance explained (pca_variance_explained.csv):

The first principal component explains 18.6% of the total variance, followed by 9.1% for PC2, 7.3% for PC3, 5.8% for PC4, and 4.2% for PC5. Cumulatively, the first two components explain 27.8% of the variance, increasing to 35.0% by PC3, 40.8% by PC4, and 44.9% by PC5. The cumulative share reaches 51.7% by PC7, 54.9% by PC8, 58.1% by PC9, 61.1% by PC10, 64.0% by PC11, and 66.9% by PC12.

The key takeaway is that the first 8–12 principal components capture roughly 55–67% of the overall structure in the data. For downstream tasks such as k-means clustering or visualization, retaining around 8–12 components provides a practical balance between dimensionality reduction and information retention.

PC1 loadings (top_loadings_PC1.csv):

The largest absolute contributors to PC1 include 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, and fireplaces. These variables collectively describe the overall size, amenities, and relative modernity of a home. Homes with larger living areas, bigger or multiple garages, more finished space and rooms, newer construction or renovation years, and more bathrooms and fireplaces receive higher PC1 scores. In plain terms, PC1 represents a size–amenities–modernity axis: bigger, newer homes with more features tend to score higher on PC1, which is consistent with the earlier observation that higher-priced homes shift to the right along this dimension.

PC2 loadings (top_loadings_PC2.csv):

The strongest contributors to PC2 split clearly by direction. Positive loadings include x2nd_flr_sf, bedroom_abv_gr, tot_rms_abv_grd, gr_liv_area, half_bath, kitchen_abv_gr, and ms_sub_class, while negative loadings include bsmt_fin_sf1, bsmt_full_bath, total_bsmt_sf, year_built (slightly negative), and x1st_flr_sf. This pattern indicates that PC2 contrasts homes with more above-grade, second-floor living space and bedrooms against homes whose square footage is concentrated in the basement or on the main floor. As a result, PC2 can be interpreted as a layout or verticality axis, separating multi-story homes from more ranch-style or basement-heavy layouts.

7. K-Means Clusterings on PCA scores:

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

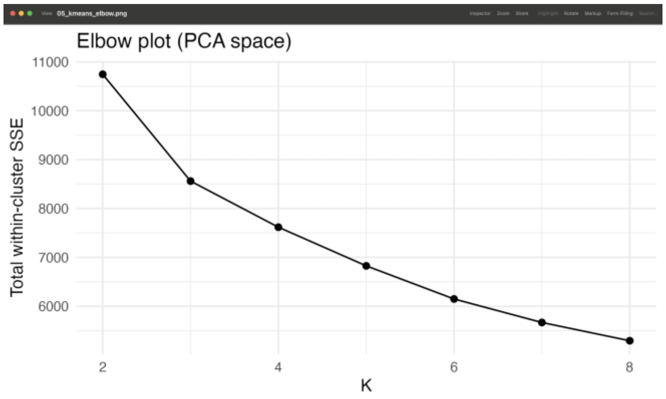


Figure 11: Elbow plot showing decreasing within-cluster SSE as K increases. The curve begins to flatten near K = 3, suggesting three clusters provide an efficient balance of fit and simplicity.

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)

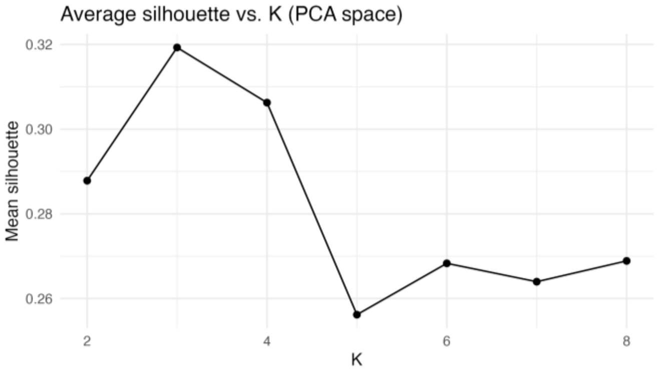


Figure 12: Average silhouette scores for different values of K. The highest score occurs near K = 3, indicating that three clusters yield the best overall separation.

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)

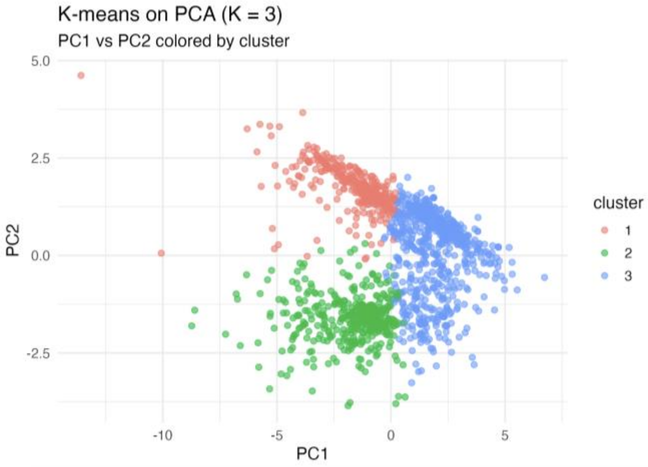
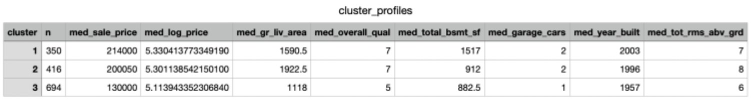


Figure 13: PC1–PC2 projection colored by K-means cluster assignment (K = 3). Distinct groupings illustrate meaningful separation among housing segments in reduced feature space.

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 the key characteristics of each housing cluster. Cluster 1 represents higher-end homes, with a strong median sale price of roughly $214K, higher overall quality ratings (around 7), larger living areas, more recent construction, and larger garages. Cluster 2 captures mid-market homes that have similar quality levels to Cluster 1 but slightly lower median prices (around $200K). These homes are moderately large and tend to have been built earlier on average. Cluster 3 consists of lower-priced homes, characterized by much smaller living areas (around 1,118 square feet), older construction with a median build year near 1957, lower quality ratings (around 5), and typically only one-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.

There are several promising directions for extending this project. A natural next step is to incorporate predictive modeling to evaluate how well the identified features and clusters can forecast housing prices. Comparing models such as linear regression, random forests, and gradient boosting would allow for a trade-off between interpretability and predictive performance. Using cross-validation and feature importance measures would further clarify which characteristics contribute most directly to price formation.

Another valuable extension is to examine time effects in the housing data. Prices are influenced by broader economic conditions, interest rates, and seasonality, none of which are captured in a purely cross-sectional analysis. Adding temporal variables such as sale year or month, or linking the data to macroeconomic indicators, would help determine whether the observed relationships are stable over time or vary across different market cycles.

Additional insights could be gained by exploring categorical and textual features in more depth. Variables such as exterior materials, heating systems, and remodeling details may encode information not fully reflected in numeric measures. If property descriptions were available, text mining techniques could be applied to extract qualitative themes that influence pricing and buyer perceptions.

Future work may also explore alternative clustering and dimensionality reduction methods. Approaches such as hierarchical clustering, DBSCAN, UMAP, or t-SNE could uncover different market structures or reveal smaller, more specialized housing segments. Incorporating spatial analysis using geographic coordinates or neighborhood shapefiles could further enhance location-based understanding of price patterns.

Finally, expanding the dataset to include additional cities or regional markets would allow for an assessment of how well these findings generalize beyond Ames. Comparing results across diverse housing markets would reveal whether similar pricing drivers persist nationwide or whether regional differences dominate. Together, these extensions would deepen insight into housing market dynamics 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.