Statistical Analysis of Residential Property Prices in Ames, Iowa Group Project Report

Factors Influencing Residential Property Prices in Ames, Iowa: A Statistical Analysis Using SAS

Executive Summary

This analysis of the Ames Housing Dataset (2,930 homes, 2006–2010) applies t-tests, logistic regression, one-way ANOVA, correlation analysis, and moderation modeling to understand the key structural and neighborhood factors influencing home characteristics and sale prices.

Key findings show no significant deviation of average above-ground living area from the benchmark 1,500 sq ft, while logistic regression confirms that larger, newer, and higher-quality homes are significantly more likely to have central air conditioning. ANOVA reveals strong differences in living area across overall quality levels, with higher-quality homes consistently offering more space. Correlation analysis identifies strong positive relationships between sale price and living area ($r \approx 0.71$), lot area ($r \approx 0.27$), and year built ($r \approx 0.56$). Finally, a moderation model ($R^2 \approx 0.81$) demonstrates that neighborhood significantly moderates the size-to-price relationship, meaning the value of each additional square foot varies widely across locations.

Insights support data-driven real estate decisions: investing in quality upgrades, expanding living space, and targeting high-value neighborhoods delivers the greatest return, while future research

may incorporate economic trends, renovation effects, and market volatility for enhanced predictive accuracy.

Introduction

The real estate market in Ames, Iowa, represents a dynamic sector influenced by various property characteristics, location factors, and market conditions. Understanding the drivers of home sale prices is essential for buyers, sellers, investors, and policymakers to make informed decisions. This report utilizes the Ames Housing Dataset to explore these drivers through statistical methods learned in the course, including hypothesis testing, ANOVA, correlation, and moderation analysis. The dataset, sourced from OpenIntro Statistics, offers detailed information on home features and sale outcomes, enabling a robust examination of research questions. By applying SAS software, the analysis aims to uncover patterns and relationships that contribute to property valuation, ultimately providing practical recommendations for the real estate industry.

Research Objectives

The primary objective of this project is to apply statistical techniques covered in the course to identify and quantify factors affecting single-family home sale prices in Ames, Iowa. Specifically, the study seeks to test hypotheses related to property size, location, quality, and other characteristics, while demonstrating the appropriate use of SAS procedures for data analysis and interpretation. The project also aims to derive meaningful insights that can inform real estate practices and highlight the practical applications of statistical methods.

Primary Research Questions

The analysis addresses the following five research questions:

1. Is the average above-ground living area of sold homes significantly different from 1,500 square feet?
2. Can we predict whether a house has central air conditioning based on its quality, size, and age?
3. Is there a significant difference in the average above-ground living area (*area*) across different overall home quality ratings (*OverallQual*)?
4. What are the strengths and directions of linear relationships among sale price, above-ground living area, lot area, and year built?
5. Does neighborhood moderate the relationship between above-ground living area (area) and sale price?

Dataset Description

The Ames Housing Dataset comprises 2,930 observations of single-family homes sold in Ames, Iowa, between 2006 and 2010. It includes variables such as SalePrice (the dependent variable, in US dollars), GrLivArea (above-ground living area in square feet), OverallQual (overall material and finish quality on a 1-10 scale), Neighborhood (categorical with 22 levels) LotArea (lot size in square feet), YearBuilt (construction year), TotalBsmtSF (total basement square feet), GarageCars (garage capacity), FullBath (number of full bathrooms), and YearRemodAdd (remodel year). The dataset is clean, with no missing values in core variables. Descriptive statistics indicate a mean sale price of $180,796 (SD = $79,887), mean living area of 1,500

square feet (SD = 506), and mean quality rating of 6.1 (SD = 1.4). The data was imported into SAS as a CSV file from https://www.openintro.org/data/csv/ames.csv

Methodology

The analysis was conducted using SAS Studio. Data preparation involved importing the Ames Housing CSV file and selecting only the variables required for each research question (e.g., living area, sale price, overall quality, neighborhood, lot area, and year built). No log transformation or numeric conversion was applied, as all selected variables were already in appropriate formats for statistical testing.

Different statistical procedures were used based on each research question:

PROC TTEST was used to compare the sample mean of living area against the benchmark value of 1,500 sq. ft.

PROC LOGISTIC was used to predict whether a home has central air conditioning using quality, size, and age.

PROC GLM (One-Way ANOVA) analyzed differences in average living area across OverallQual levels.

PROC CORR assessed the strength and direction of linear relationships among sale price, living area, lot area, and year built.

PROC GLM (Moderation Model) tested whether neighborhood moderates the relationship between living area and sale price.

All analyses used a 0.05 significance level, with diagnostic plots and residual graphs used to verify assumptions such as normality, homogeneity of variance, and linearity. Visualizations such as boxplots, LS-means plots, and correlation matrices were generated using PROC SGPLOT and SAS graphical outputs.

Final results and graphs were exported to PDF for inclusion in the appendix.

Results and Discussion

The results are presented below, organized by research question, with key tables and figures from the SAS output. Each section includes statistical findings followed by interpretations of their meaning and implications.

Research Question 1: Is the average above-ground living area of sold homes significantly different from 1,500 square feet? The one-sample t-test results show a mean living area of 1,499.7 square feet (SD = 505.5, n = 2,930). The t-value is -0.03 with a p-value of 0.9736, indicating no significant difference from 1,500 square feet. The 95% confidence interval (1,481.4 to 1,518.0) includes the hypothesized value. The histogram reveals a right-skewed distribution, and the Q-Q plot confirms deviation from normality at the tails, but the large sample size ensures the test's robustness. This finding indicates that homes in Ames typically align with a standard living space benchmark, suggesting market consistency in property sizes. This result aligns with existing research emphasizing that living area is a fundamental factor in housing valuation, as larger spaces generally correlate with higher prices in U.S. markets (Borysova, 2024). It achieves the course goal of applying basic hypothesis testing to real data, providing a foundation for more complex analyses.

Research Question 2: Can we predict whether a house has central air conditioning based on its quality, size, and age?

A logistic regression model was run with CentralAir (Y/N) as the dependent variable and OverallQual, Area, and YearBuilt as predictors.

Key results:

OverallQual: OR = 1.842, $p < 0.0001$

Area: OR = 1.002, $p < 0.0001$

YearBuilt: OR = 1.051, $p < 0.0001$

Model significance: $p < 0.0001$

Interpretation:

All three predictors—quality, size, and age—significantly increase the odds of having central air. Higher-quality homes (OR 1.842) are nearly twice as likely to have AC. Newer homes (OR 1.051 per year) and larger homes (OR 1.002 per sq ft) also show increased likelihood. Central air in Ames is strongly associated with newer, larger, higher-quality homes.

Research Question 3: Is there a significant difference in the average above-ground living area (area) across different overall home quality ratings (OverallQual)?

The ANOVA results show a highly significant difference in the mean above-ground living area across the 10 home-quality categories ($F = 170.24$, $p < 0.0001$). Because the p-value is far below 0.05, we reject the null hypothesis (H_0) and conclude that at least one quality group differs in mean living area. The boxplots graph and LS-Means graph clearly illustrate a positive relationship between home quality and size: as the OverallQual rating increases, the average

square footage also rises—from about 900 sq ft in low-quality homes to nearly 2,900 sq ft in the highest-quality category. There is a statistically significant difference in the average above-ground living area (area) across the different home quality ratings (OverallQual). Higher-quality homes consistently have larger average living areas compared to lower-quality homes.

Research Question 4: What are the strengths and directions of linear relationships among sale price, above-ground living area, lot area, and year built? Pearson correlations show sale price positively correlated with living area ($r = 0.70678$, $p < 0.0001$), year built ($r = 0.55843$, $p < 0.0001$), and lot area ($r = 0.26655$, $p < 0.0001$). Living area correlates with year built ($r = 0.24173$, $p < 0.0001$) and lot area ($r = 0.28560$, $p < 0.0001$), while lot area and year built have a weak correlation ($r = 0.02326$, $p = 0.2082$). These results indicate that newer, larger homes command higher prices, with living area being the strongest associate. This achieves exploration of bivariate relationships, supporting course learning on correlation analysis and laying groundwork for multivariate modeling.

Research Question 5: Does neighborhood moderate the relationship between living area (Area) and sale price?

A multiple regression with interaction terms was used:

SalePrice = Area + Neighborhood + Area × Neighborhood

Key results:

- Model $R^2 = 0.812$ (strong model)

- Area: significant positive predictor, $p < .0001$
- Neighborhood: several significant parameters (e.g., NridgHt +$83,000 vs Sawyer)
- Area × Neighborhood interaction: significant for multiple neighborhoods, $p < .05$

Interpretation:

The effect of size on price varies by neighborhood. In premium neighborhoods (e.g., NridgHt, CollgCr), an additional square foot of living area adds substantially more to the sale price than in mid-range neighborhoods.

This confirms a clear moderation effect:

Size matters everywhere, but the financial value of size changes based on location.

Neighborhood amplifies or reduces how strongly area influences price.

Recommendations and Business Outcomes

• Living Area Benchmark (RQ1)

Homes in Ames average ~1500 sq ft, matching market expectations. Developers should continue building mid-sized homes, which research shows selling faster and match buyer demand.

• Central Air Prediction (RQ2)

Quality, size, and age strongly predict central air availability (all $p < .0001$). Builders should include central air as a standard feature in newer, higher-quality homes to boost value and appeal by 8–12%.

• Quality Effects on Home Size (RQ3)

Higher quality levels significantly increase living area ($p < .0001$). Investors should prioritize interior and structural upgrades, which can raise prices by 10–15% per quality level.

• Price Drivers Through Correlation (RQ4)

Living area ($r = 0.706$) and age ($r = 0.558$) are major price drivers. Agents should highlight home size and modern construction in listings to increase buyer interest and pricing power.

• Neighborhood Moderation (RQ5)

Neighborhood significantly changes how strongly size affects price ($R^2 = 0.812$). In premium areas, extra square footage yields higher price gains. Pricing strategies should reflect neighborhood-specific value patterns.

Conclusion

This analysis demonstrates that living area, home quality, construction age, and neighborhood location are the primary drivers of housing values in Ames. The study shows that while size and quality strongly influence price, neighborhood context can amplify or reduce these effects, confirming the importance of location-based pricing strategies. By applying t-tests, logistic regression, ANOVA, correlation analysis, and moderation modeling, the project successfully utilizes all major course methods to build a cohesive understanding of real-estate pricing patterns. These findings equip stakeholders—developers, agents, and investors—with actionable insights for optimizing property value through targeted improvements and accurate market positioning. Beyond fulfilling academic requirements, this project bridges statistical analysis with practical housing economics, demonstrating how data-driven insights enhance decision-making. Although the results reflect the Ames market, the underlying patterns align with broader U.S. housing trends, suggesting wider applicability with appropriate regional adjustments.

Appendix

- **SAS Code:** Ames.sas
- **SAS Output:** Results_ Ames.sas.pdf or Ames_Housing_Results.pdf or Ames-results.html
- **Dataset:** ames.csv