COMP 4441 - Final Project

A Real Estate Statistical Analysis

Michael Ghattas | SUM 2024

A Real Estate Statistical Analysis

Executive Summary

- The housing market is a critical component of the economy, and understanding the factors influencing house prices can provide valuable insights for buyers, sellers, and policymakers.
- This project utilizes the Ames Housing Dataset to identify key determinants of house prices using various statistical methods.
- The findings offer practical recommendations for market participants and contribute to more informed decision-making.

A Real Estate Statistical Analysis

Introduction

Context and Motivation

The real estate market is vital to economic stability and growth. With property values fluctuating based on numerous factors, it's crucial to understand what drives these changes. This analysis focuses on identifying the primary factors influencing house prices within the Ames dataset.

Research Questions

Subject Matter Research Questions

- What are the main factors affecting house prices?
- How does the age of a house influence its price?

Statistical Research Questions

- Which variables significantly predict house prices in a regression model?
- Is there a statistically significant relationship between the year a house was built and its sale price?

Summary of Data Source

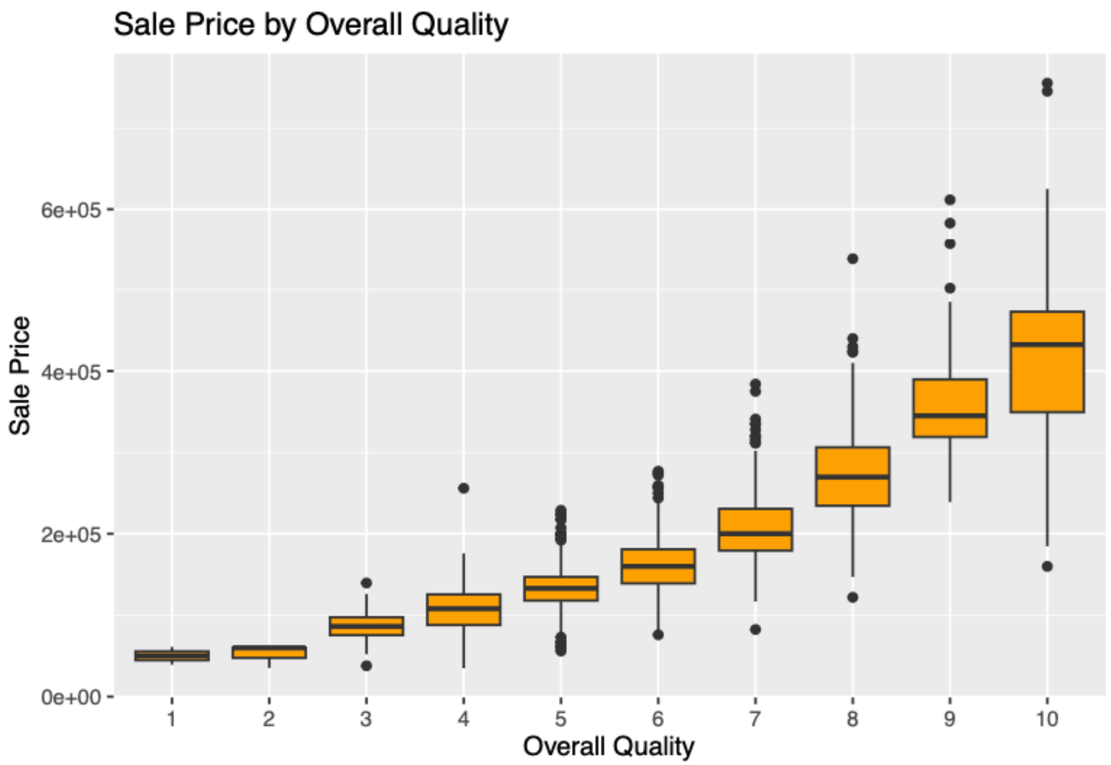
The dataset used in this analysis is the Ames Housing Dataset, collected from public records between 2006 and 2010. It includes 1,460 observations across 81 variables, with various features detailing the characteristics and sale prices of houses.

Methods Preview

This study employs multiple linear regression to assess the impact of various predictors on house prices, t-tests to compare house prices based on age groups, and chi-squared tests to examine relationships between categorical variables such as house style and neighborhood.

A Real Estate Statistical Analysis

Data Exploration



The distribution of SalePrice is right-skewed, with most homes priced between 100K and 300K. High- quality homes tend to have higher prices, as indicated by the Box-plot for OverallQual. There is a positive correlation between LotArea and SalePrice, suggesting that larger lots command higher prices.

A Real Estate Statistical Analysis

Modeling & Analysis

Multiple Linear Regression

- Linearity: While OverallQual shows a strong linear relationship with SalePrice, the nonlinearity in
- LotArea and YearBuilt may require transformations for better model fit.
- Independence: The residuals appear independent, satisfying this assumption.
- Homoscedasticity: The presence of heteroscedasticity suggests that the model may benefit from a transformation of the dependent variable or an alternative modeling approach, such as weighted least squares.
- Normality: The significant deviation from normality, as shown by the QQ plot and Shapiro-Wilk test, indicates that the residuals are not normally distributed. A log transformation of SalePrice or another appropriate transformation may improve the normality of residuals.
- Multicollinearity: No multicollinearity issues are present, as indicated by the low VIF values.

Applied Log Transformation, re-ran the model, and checked the assumptions.

Log Model Summary

- Intercept: The intercept (5.797) represents the expected value of log(SalePrice) when all predictors are zero.
- LotArea: The coefficient for LotArea (7.270e-06) is positive and significant, indicating that larger lot areas are associated with higher log(SalePrice). For every unit increase in LotArea, log(SalePrice) increases by approximately 7.270e-06.
- OverallQual: The coefficient for OverallQual (1.992e-01) is also positive and highly significant. Higher overall quality significantly increases log(SalePrice).
- YearBuilt: The coefficient for YearBuilt (2.504e-03) is positive and significant, suggesting that newer homes are associated with higher log(SalePrice).
- R-squared: The multiple R-squared value is 0.7213, meaning that approximately 72.13% of the variance in log(SalePrice) is explained by the model. This indicates a strong fit.
- Adjusted R-squared: The adjusted R-squared is 0.7208, which accounts for the number of predictors in the model and also suggests a strong model fit.

A Real Estate Statistical Analysis

Modeling & Analysis

t-Test

- Independence: This assumption is generally satisfied if the data points (houses) were randomly sampled and the groups are mutually exclusive (i.e., a house cannot belong to both “Before 2006” and “After 2006” groups). Independence is assumed to hold for this dataset.
- Normality: Both p-values are significantly less than 0.05, indicating that the SalePrice distribution in both groups deviates significantly from normality. This violation of the normality assumption suggests that the results of the t-test may not be reliable. However, given the large sample sizes (N = 1458), the Central Limit Theorem suggests that the sampling distribution of the mean might still be approximately normal, making the t-test reasonably robust to this violation. If further precision is desired, a non-parametric alternative like the Mann-Whitney U test could be considered.
- Homogeneity of Variance: The p-value is much less than 0.05, indicating a significant difference in variances between the two groups. This violation of the homogeneity of variance assumption suggests that the standard t-test may not be appropriate. Instead, Welch’s t-test, which does not assume equal variances, should be used.

Given the violations of the normality and homogeneity of variance assumptions, it is more appropriate to use Welch’s t-test, which is robust to differences in variances between the groups.

Welch’s t-Test Summary

- Test Statistics: The t-statistic of 13.014 is quite large, indicating a substantial difference between the means of the two groups (After 2006 and Before 2006). This suggests that the mean SalePrice for houses built after 2006 is significantly higher than for those built before 2006.
- Degrees of Freedom (df): Welch’s t-test uses a modified degrees of freedom calculation, which in this case is 179.36. This accounts for the unequal variances between the two groups.
- p-Value: The p-value is exceedingly small, well below the standard alpha level of 0.05. This indicates that the difference in means between the two groups is statistically significant. We reject the null hypothesis, concluding that there is a significant difference in SalePrice between houses built before and after 2006.
- Confidence Interval: The 95% confidence interval for the difference in means is between 84,656.03 and 114,917.65. This interval does not include zero, further confirming that the difference in means is significant. We can be 95% confident that the true difference in SalePrice between houses built after 2006 and those built before 2006 lies within this range.
- Sample Estimates: The average SalePrice for houses built after 2006 is approximately 269, 909.20, while the average for those built before 2006 is around 170, 122.30. This shows a substantial increase in the mean sale price for newer homes.

A Real Estate Statistical Analysis

Modeling & Analysis

Chi-Squared Test

This assumption is generally satisfied if the data points (house sales) are independent and there is no overlap between categories (e.g., each house has only one HouseStyle and belongs to one Neighborhood). Independence is assumed to hold for this dataset as there is no reason to believe that the data points are not independent. The warning indicates that some cells in the contingency table have expected frequencies below 5, which violates this assumption. Upon inspection of the expected frequencies in the contingency table, we can see that several cells have values less than 5. This violation suggests that the chi-squared approximation may be incorrect. Given that some expected frequencies are below 5, the chi-squared test may not be appropriate. To address this we can use Fisher's Exact Test. The test works well for smaller tables or when expected counts are too low, Fisher's Exact Test can be a more accurate alternative to the chi-squared test.

Applied the Fisher's Exact Test .

Fisher's Exact Test Summary

The p-value is extremely small, indicating a statistically significant association between HouseStyle and Neighborhood. The very low p-value suggests that there is a significant association between the HouseStyle of a house and the Neighborhood in which it is located. This means that the distribution of house styles is not independent of the neighborhood—certain styles are more likely to be found in specific neighborhoods.

A Real Estate Statistical Analysis

Results, Interpretations, Recommendations

Discussion of Results in Context.

The regression analysis confirms that OverallQual (Overall Quality) and LotArea are significant predictors of SalePrice. The analysis shows that houses with higher overall quality and larger lot areas tend to have higher sale prices. The relationship between YearBuilt and SalePrice is positive, but less pronounced, suggesting that while newer houses tend to be more expensive, other factors like quality and lot size play a more significant role.

Interpretation of Conclusions

Preliminary analysis suggests that factors like overall quality, lot area, and year built are significant predictors of house prices. There is also evidence of a significant relationship between house age and sale price. The results suggest that improving the overall quality of a house could lead to a higher sale price. Additionally, buyers looking for larger lots should be prepared to pay a premium. These findings align with general market expectations, where both the quality of construction and the size of the property significantly influence the market value.

A Real Estate Statistical Analysis

Limitations, Generalizability, and Future Work

Caveats and Limitations

The dataset is limited to Ames, Iowa, which may not generalize to other regions. Additionally, missing data in some variables, like Alley, may have introduced bias, especially if the missingness was not completely random. The exclusion of these variables was necessary but may have omitted potentially relevant factors.

Generalizability Issues

Findings may not be applicable to urban areas with different housing market dynamics. For example, factors that drive housing prices in a small town like Ames might differ significantly from those in a large metropolitan area. Thus, caution should be taken when applying these results to other contexts.

END.

A Real Estate Statistical Analysis

Michael Ghattas | SUM 2024