



Final Project

Predicting Property Prices

Group 2 members: Diego Correa, Jagdish Chhabria,
Orli Khaimova, Richard Zheng, Stephen Haslett



Introduction

The aim of this study is to explore the factors that influence home buyers when buying homes.

Examples:

- Location
- Property size
- Age
- Number of rooms

In order to explore this topic, we used the "Ames Housing" dataset.

In our investigation to predict house prices, we are given eighty-one categorical and numerical independent variables to use in a multiple linear regression model



Key Words

- real estate
- house prices
- multiple regression
- assessed value
- home buyers
- linear models



Literature Review

- Understanding Recent Trends in House Prices and Home Ownership
(https://www.kansascityfed.org/documents/3224/pdf-Shiller_0415.pdf) by Robert J. Shiller
-
- Cracking the Ames Housing Dataset with Linear Regression
(<https://towardsdatascience.com/wrangling-through-dataland-modeling-house-prices-in-ames-iowa-75b9b4086c96>) by Alvin T. Tan



Methodology

Exploratory data analysis includes

- Familiarize with data values
- Identify the categorical and numerical values
- Imputation of missing values
- Visualize distribution and correlations

Model creation and selection

- Produce 4 multiple linear regression models
- Analyze residuals
- Prioritize the adjusted R-square and Root Mean Square Error




Dataset

The Ames Housing dataset consists of 81 variables describing the characteristics of 1,460 homes in Ames, Iowa sold between 2006 and 2010.

Target variable is SalePrice. Of the predictor variables, 38 are numeric and 42 are categorical. 18 of the variables have NA values which in some cases denotes that that feature is missing from that home, and not an empty value.

The dataset is available for download via [Kaggle.com](https://www.kaggle.com/datasets/colinmccormac/ames-housing-dataset).

The Ames Housing dataset is feature rich, and contains many of the features that home buyers consider when buying a house such as overall condition, location, number of rooms, etc.



Exploratory - Correlations

After performing transformations on the dataset to impute the missing data, we can see the correlations of the numerical variables in the dataset

The chart indicates the correlations of numerical independent variables to both the independent and dependent variables

Key takeaway:

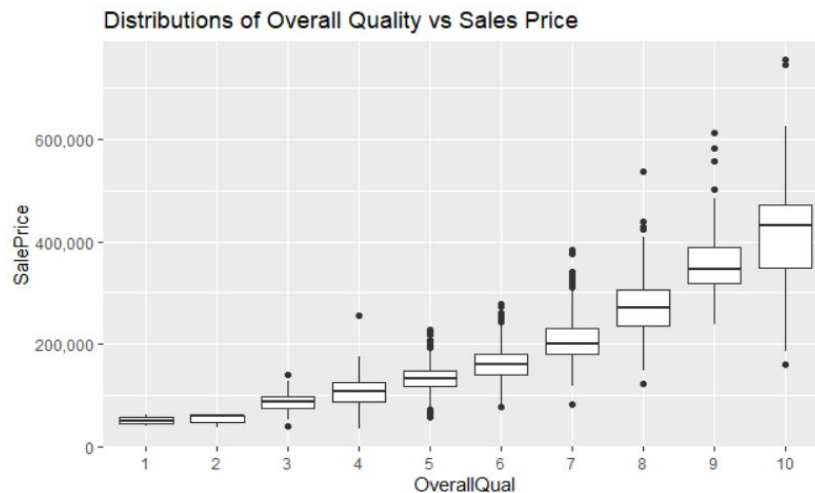
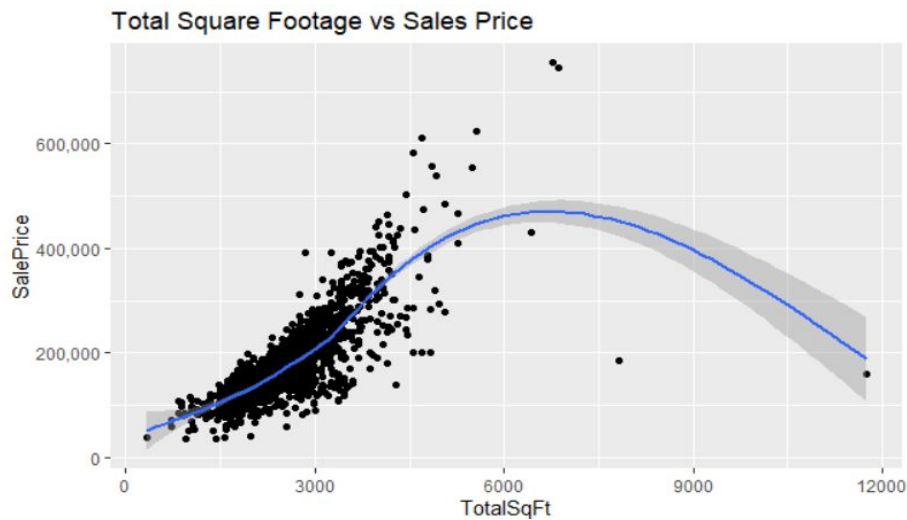
- Overall Quality, Square footage, and Number of Cars in Garage have the highest correlation to Sale Price

row	column	cor
GarageCars	GarageArea	0.8824754
YearBuilt	GarageYrBltd	0.8368630
GrLivArea	TotRmsAbvGrd	0.8254894
TotalBsmtSF	X1stFlrSF	0.8195300
OverallQual	SalePrice	0.7909816
GrLivArea	SalePrice	0.7086245
X2ndFlrSF	GrLivArea	0.6875011
BedroomAbvGr	TotRmsAbvGrd	0.6766199
BsmtFinSF1	BsmtFullBath	0.6492118
GarageCars	SalePrice	0.6404092

Exploratory - Relationships relative to Sale Price

Key Takeaways:

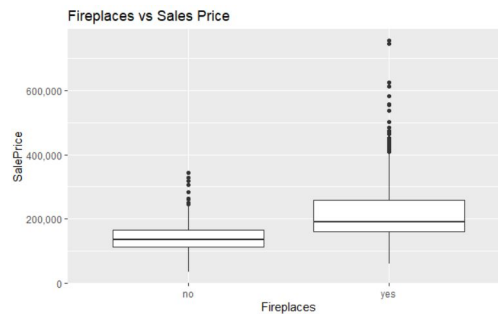
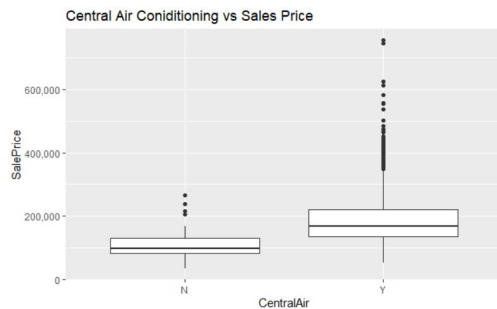
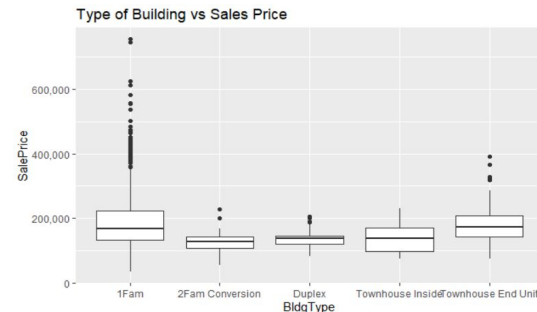
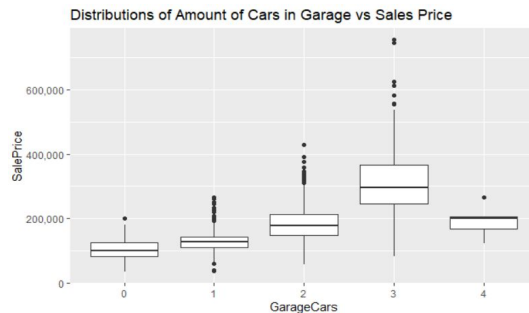
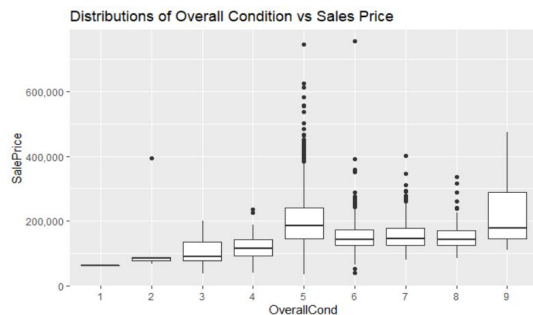
- Livable Sq. Ft. & Overall Quality have outliers on the right side of distribution - expected as the left is bounded by 0, but not the right
- Overall Quality is discrete
- Livable Sq. Ft. is continuous



Exploratory - Continued

Key Takeaways:

- Heavily skewed right when:
 - Overall Quality equal to five
 - Cars in Garage equal to three
- Building Type equal to 1 family
- Central Air equal to Yes
- Fireplaces equal to Yes

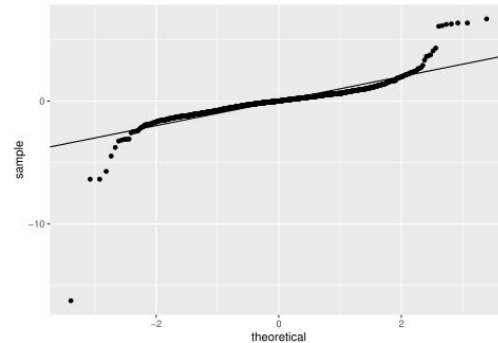
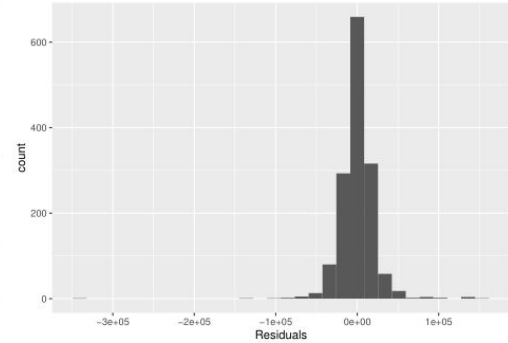
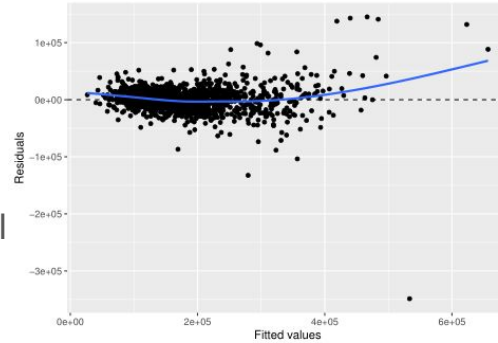


Model 1

Model 1 is a multiple linear regression using all variables in the data set

Residual graphs

Adjusted R^2 : 0.9063

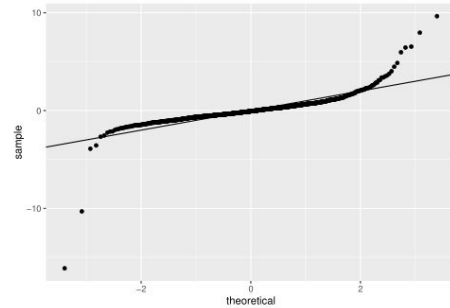
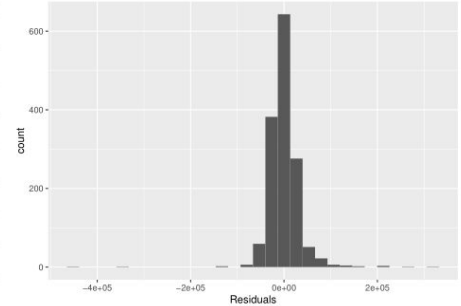
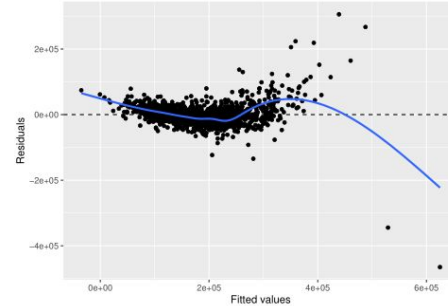


Model 2

Model 2 is multiple linear regression that filters out the categorical variables, and only uses the numerical values

Residual graphs

Adjusted R^2 : 0.8086

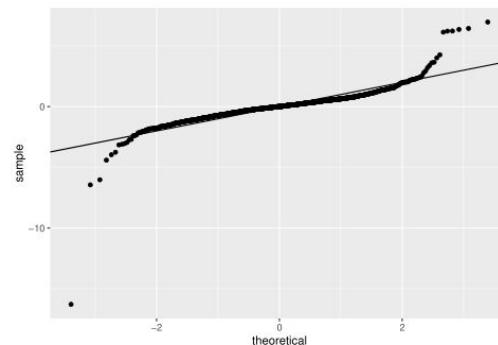
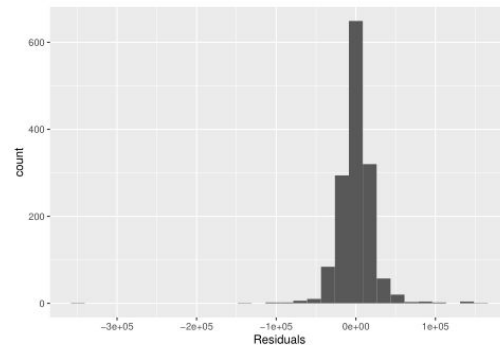
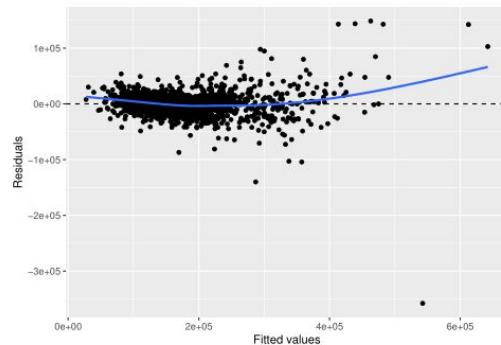


Model 3

Model 3 uses model 1 and perform a backward stepwise elimination to optimize feature engineering

Residual graphs

Adjusted R^2 : 0.9076



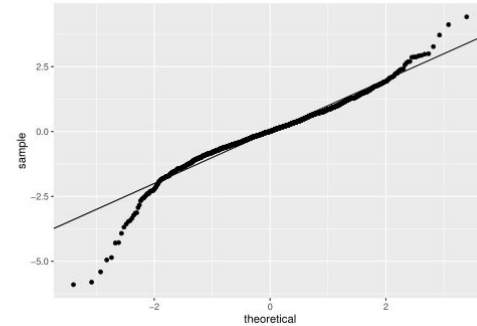
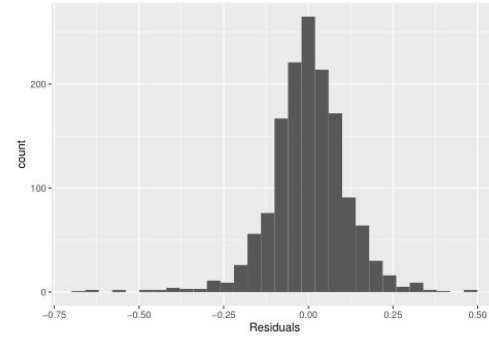
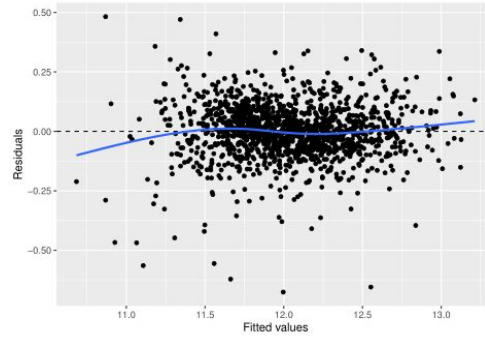
Model 4

Model 4 is a multiple linear regression that uses the suggestion of the author that includes:

- Combining the Bath variables into one
- Calculating age of house
- Grouping Sale Conditions
- Identify if house is sold new

Residual graphs

Adjusted R^2 : 0.9143





Model Selection

We determined that model 4 is the best selection for a multiple linear regression model based on:

- Model 4 fit our data the best (highest R^2 value)
- Reasonable computational expenditure - fast performance speed
- Residuals fit the assumptions for linear regression:
 - normal distribution and constant variability



Conclusion

- Applying transformations to variables improved model performance
- Square feet of living area and neighborhood were the most impactful attributes when determining housing price
- Limitations
 - Dataset only contained houses in Iowa
 - Many variables were heavily unbalanced towards one level



References

De Cock, D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. Journal of Statistics Education, 19(3). <https://doi.org/10.1080/10691898.2011.11889627>

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management, 5(1), 81–102. [https://doi.org/10.1016/0095-0696\(78\)90006-2](https://doi.org/10.1016/0095-0696(78)90006-2)

Kaggle. (2012). House Prices - Advanced Regression Techniques, Version 1. <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

OECD Statistics. (2005, July 8). Hedonic Method. OECD Glossary of Statistical Terms. <https://stats.oecd.org/glossary/detail.asp?ID=1225>.

Shiller, R. J. (2007). Understanding Recent Trends in House Prices and Homeownership. In Proceedings - Economic Policy Symposium - Jackson Hole (pp. 89–123), Federal Reserve Bank of Kansas City. https://www.kansascityfed.org/documents/3224/pdf-Shiller_0415.pdf

Tan, A. T. (2021, May 6). Cracking the Ames Housing Dataset with Linear Regression. Medium. <https://towardsdatascience.com/wrangling-through-dataland-modeling-house-prices-in-ames-iowa-75b9b4086c96>.

Weinstock, L. R. (2021, May 3). Introduction to U.S. Economy: Housing Market. Congressional Research Service. <https://fas.org/sgp/crs/misc/IF11327.pdf>.