Submitted to the Annals of Applied Statistics

STAT 444 FINAL PROJECT PROPOSAL

BY ANGELO CARREON [1,*] , HOSEOK LEE [1,†]
JOY CHEN [1,‡] AND STEVEN SHEN [1,§]

[1] *Department of Statistics and Actuarial Science, University of Waterloo,* [*]*jaccarre@uwaterloo.ca;* [†]*h349lee@uwaterloo.ca;* [‡]*z635chen@uwaterloo.ca;* [§]*s58shen@uwaterloo.ca*

This paper contains our proposal for the STAT 444 final project. It outlines our dataset,

1. Introduction to our chosen dataset.

Our project aims to assess the feasibility of utilizing regression techniques for interpolating and modeling housing prices. From the Journal of Statistics, we have selected a dataset of housing prices in Ames, Iowa, along with other pertinent features. In today's dynamic real estate market, precise and dependable housing price predictions hold immense significance for homeowners, buyers, and real estate professionals alike. By tapping into the potential of advanced modeling techniques, our goal is to improve the accuracy and reliability of housing price predictions, contributing to more informed decision-making in the industry.

2. Exploratory Data Analysis.

2.1. *Summary Statistics.*

- The dataset contains 2930 rows and 82 columns.
- There are 80 explanatory variables, including 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables.

2.2. *Missing/duplicate Values.*

- Four columns contain many missing values and are thus dropped, as they are not critical to our analysis.
  - Pool.QC has only 13 observations.
  - Misc.Feature has only 106 observations.
  - Pool.QC has only 198 observations.
  - Fence has only 572 observations.
- There are no duplicate rows.

2.3. *Distribution of dependent variable:* `Sale Price`.

- The distribution of `Sale Price` is significantly right-skewed.
- The sale prices range from $12,789 to $755,000 with a mean of $180,796 and a standard deviation of $79,886.69.
- To achieve a more normal distribution, we can apply a log transformation on the dependent variable.

Keywords and phrases: We love regression.

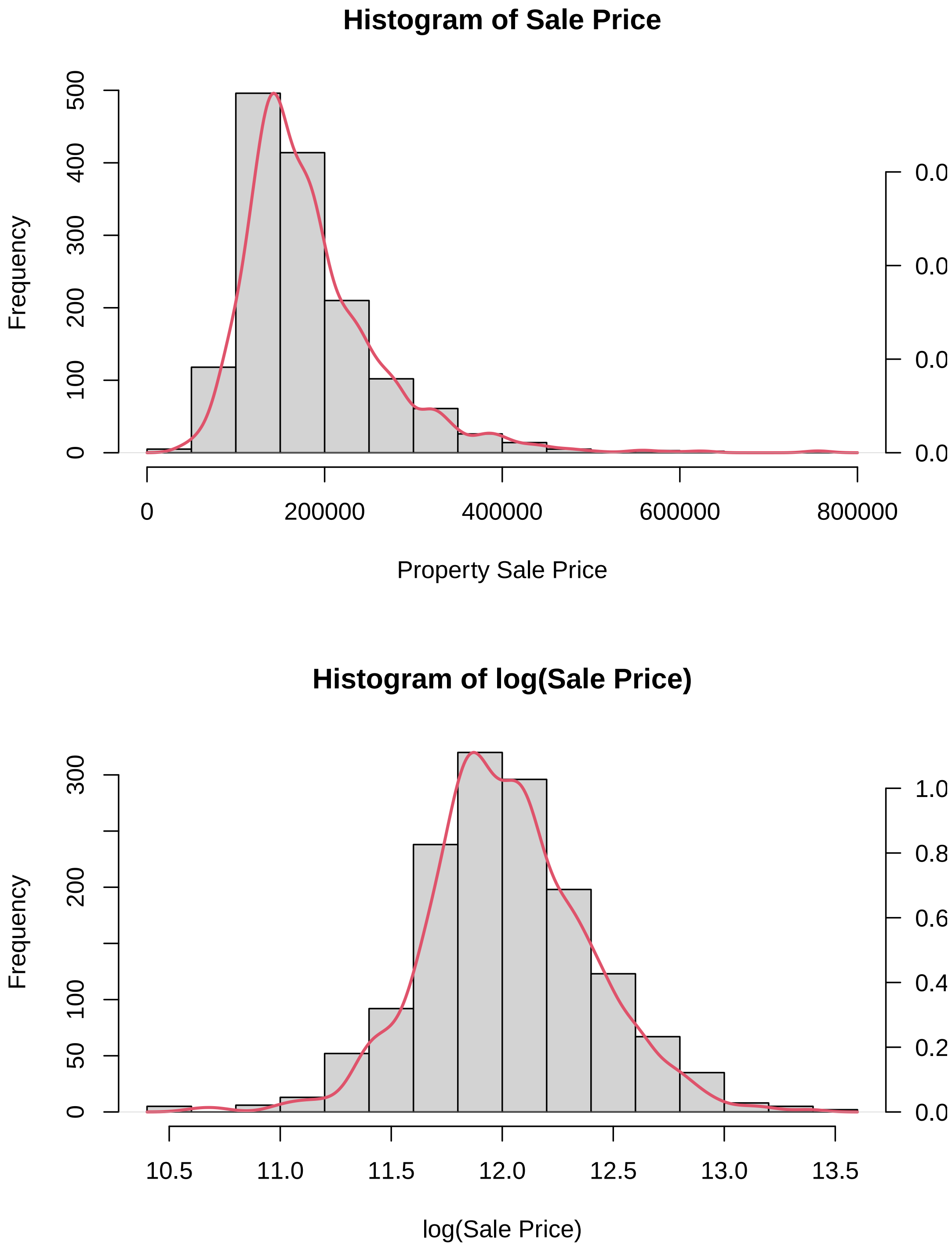

FIGURE 1. *Histogram of Sale Price (top) and Log Sale Price (bottom)*

2.4. *Correlations with the dependent variable.*

- Based on our correlation analysis, the five numeric variables with the highest correlations with our dependent variable are:
  - Overall.Qual: Rates the overall material and finish of the house.
  - Gr.Liv.Area: Above grade (ground) living area square feet.
  - Garage.Cars: Size of garage in car capacity.
  - Garage.Area: Size of garage in square feet.
  - Total.Bsmt.SF: Total square feet of basement area.

2.5. *Check if there is a high degree of correlation or linear association among independent variables.*

- Some independent variables are strongly correlated; this behavior is to be expected, as some covariates provide similar information.
  - For example, the high correlation between the variable **Garage Area** (size of garage in square feet) and **Garage Cars** (size of garage in car capacity) is not surprising.
  - To address such redundant pairs, we will select one covariate from each pair, excluding the other from the analysis.

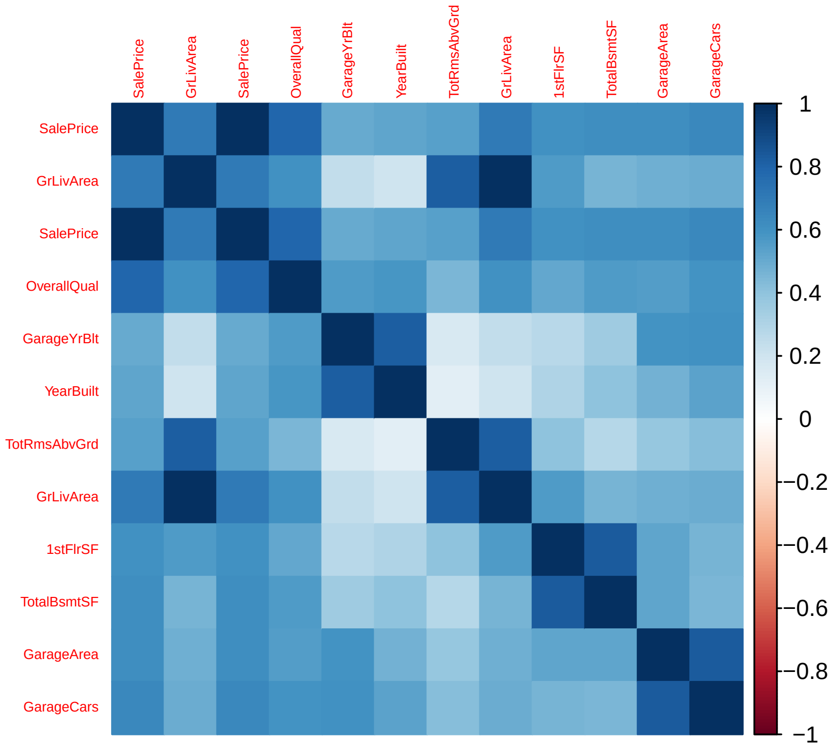

FIGURE 2. *Correlation matrix of independent variables*

2.6. *Other interesting but non-numeric variables.*

We note some non-numeric variables of interest below.

- **Neighborhood.** There appears to be a significant difference in the average property sale price by neighborhood.
- The dataset contains several categorical variables with low variance (i.e. near-constant), which likely hold negligible predictive power, including:
  - `Street` - type of road access to property
  - `Utilities` - type of utilities available
  - `Roof.Matl` - roof material

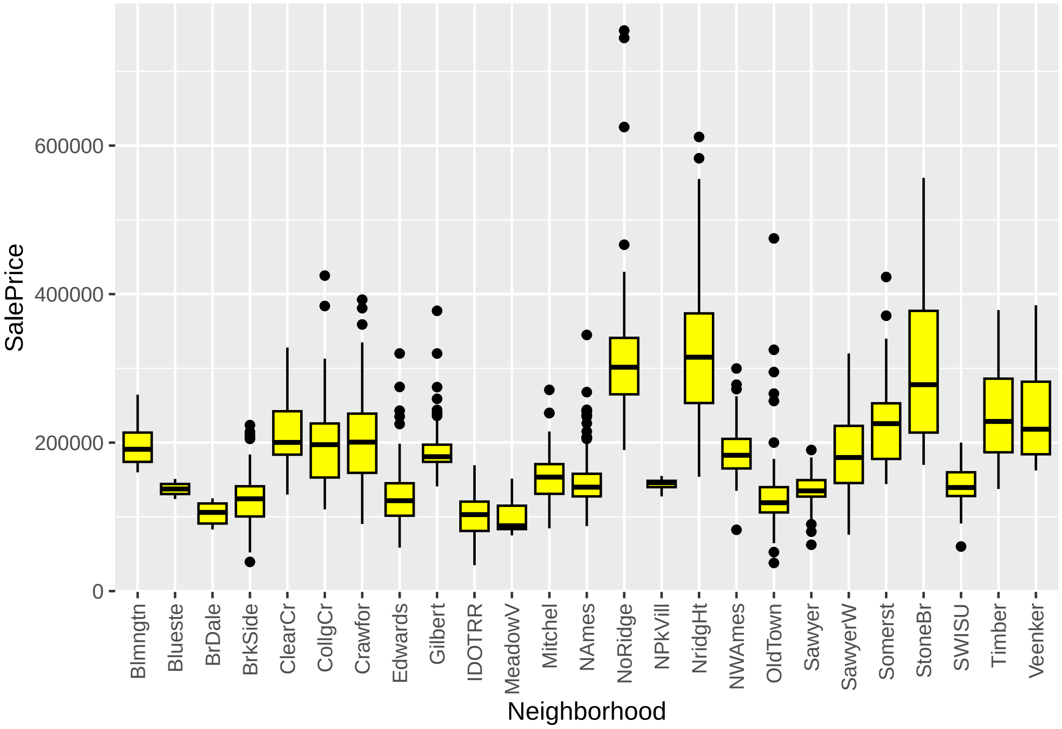

FIGURE 3. *Boxplot showing the distribution of SalePrice across Neighborhood*

3. Plan.

The objective of this study is to determine the viability and effectiveness of employing additive models, splines, and polynomial regression for predicting house prices. We focus our attention on the following:

- **Accuracy**: Assess the predictive accuracy of additive models, splines, and polynomial regression in comparison to traditional regression models commonly used in the real estate domain.
- **Flexibility**: Analyze the ability of these techniques to capture complex relationships between house price predictors, such as square footage, location, number of bedrooms, and other relevant features.
- **Interpretability**: Evaluate the interpretability and explainability of the models, ensuring that the predictions can be easily understood and justified by stakeholders.

To achieve our objective, we propose the following methodology as an outline:

1. **Initial Benchmark**: Use the performance of linear regression, with simple categorical encoding, on the full feature set as an initial benchmark.
2. **Data Preprocessing**: Handle missing/invalid values and outliers and perform feature engineering to enhance the models' predictions. To mitigate the "curse of dimensionality" and reduce collinearity in the dataset, we will use a low variance filter and clustering techniques such as Principal Component Analysis (PCA).
3. **Model Implementation**: Develop additive models, splines, and polynomial regression models using appropriate algorithms and frameworks, such as generalized additive models (GAMs) and polynomial regression libraries in R. We will also apply penalization techniques to prevent overfitting (e.g. forward/backward selection, LASSO/Ridge regression) while including interactions between covariates to model any non-linear relationships.
4. **Model Evaluation**: Assess the performance of the models using appropriate evaluation metrics, such as mean squared error (MSE), root mean squared error (RMSE), and R-squared values and compare the results with benchmark models.
5. **Interpretation and Explainability**: Examine the contributions of each feature and the underlying relationships identified by the models. Visualize the results in an intuitive yet comprehensive manner.