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, performs some brief exploratory data analysis, then outlines our approach to fitting regression models to predict a final sale price of a house given its physical attributes.

1. Introduction to our chosen dataset. Our project aims to assess the feasibility of utilizing regression techniques for interpolating and modeling housing prices. We have selected a dataset of housing prices in Ames, Iowa, along with their features from the *Journal of Statistics Education* (Cock, 2011). This dataset was prepared by Dean De Cock for use as an end-of-semester regression project. His intent was to provide data of substantial size ($n = 2930$) with easy-to-understand variables that are known to affect the final sale price such as build date, lot size, and living space square footage. Applying this to today's real estate market, our goal is to improve the accuracy and reliability of housing price predictions.

2. Exploratory Data Analysis. This dataset contains 2930 rows and 82 columns. There are 80 explanatory variables, consisting of 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables. Several columns contain many missing values and will be dropped before we begin fitting models. As highlighted in De Cock's paper, several unusual outlier house sales exist in the data; these will also be removed. There are no duplicate rows.

As seen in table 1, 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, as seen below:

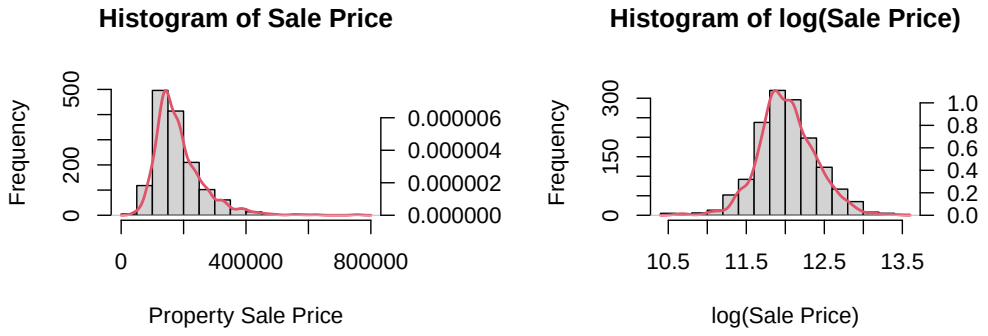

FIGURE 1. *Histograms of the response variable, Sale Price*

Keywords and phrases: housing, advanced regression.

Some variables of interest are `Neighbourhood` and `LotArea`, where we see in table 2 to have significant differences in the average property sale price. One way in which these neighborhoods could differ is in the size of the lots of the houses that reside there. Our team would need to consider variable associations such as these to deal with collinearity.

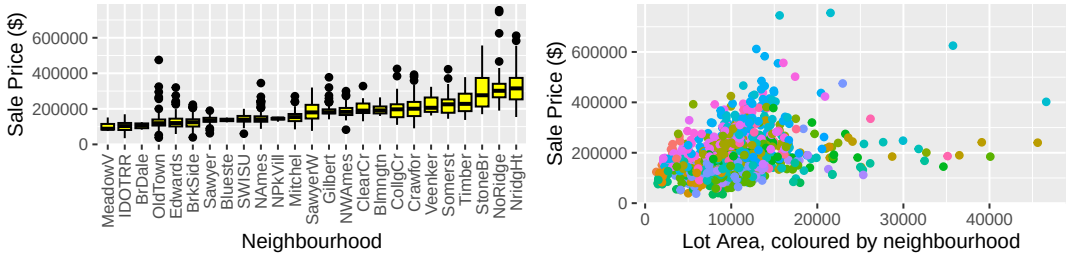

FIGURE 2. *Plots showing the distribution of final Sale Price across the different Neighborhoods*

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.

REFERENCES

COCK, D. D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. *Journal of Statistics Education* **19**.