HOME PRICING STATISTICAL ANALYSIS

IN KANSAS CITY, MO

for KC Residential Investment Group

by Zack Strathe
STAT 705 Final Project

INTRODUCTION

- The home pricing market is notoriously difficult to predict
- Real estate investment companies utilizing large and complex price prediction models have incurred huge losses when those models failed
- For KC Residential Investment Group, I propose instead utilizing a simple linear model to decide where to focus their efforts on purchasing, improving, and re-selling residential homes in Kansas City, MO
  - Utilize linear model to identify best postal code by growth in home sales price, identified by the LM slope parameter
  - Evaluate the linear model predictive performance versus a more-complex model

DATA SET

- Data from Realtor.com's Real Estate Data Library
- Contains home pricing data aggregated by postal code, every month from July 2016 through February 2022
- Data is often incomplete (i.e., many postal codes are missing entries for some months)
- Response variable: "median_listing_price_per_square_foot"
- Predictor variables: "postal_code", "month_date_yyyymm"
- Transformed "month_date_yyyymm" data from "yyyymm" format into an integer representation for the linear model to correctly utilize it

LINEAR MODEL DEFINITION

- Linear model equation:

$$y = \beta_{0,i}'postal\ code'_i + \beta_1 month + \beta_{2,i} month.'postal\ code'_i + \epsilon_{month.postal\ code_i}$$

  - $\beta_{0,i}$ represents the model intercept, where for each postal code:

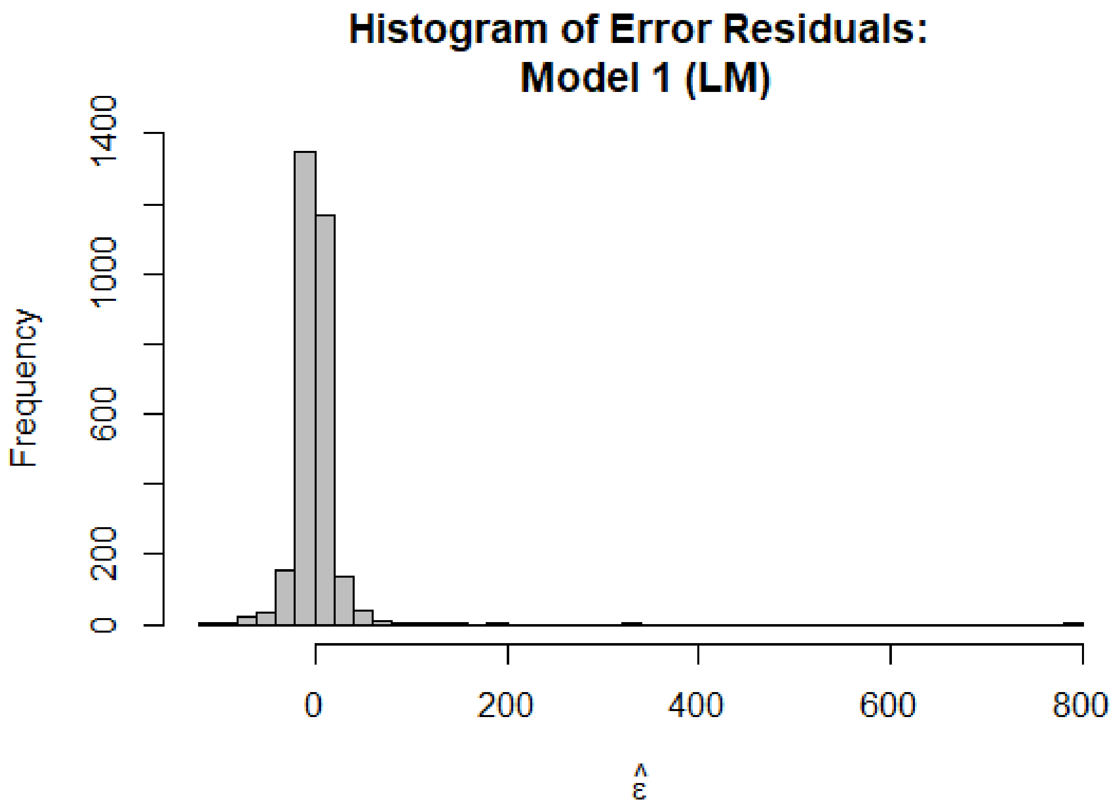
$$\beta_{0,i}'postal\ code'_i = \begin{cases} 1\ if\ i = 'postal\_code' \\ else\ 0 \end{cases}$$

  - $\beta_1 month$ represents the model's "base slope" (for the first postal code in the data set)
  - $\beta_{2,i}$ represents the slope added onto the base slope, where for each postal code:

$$\beta_{2,i} month.'postal\ code'_i = \begin{cases} 1\ if\ i = 'postal\_code' \\ else\ 0 \end{cases}$$

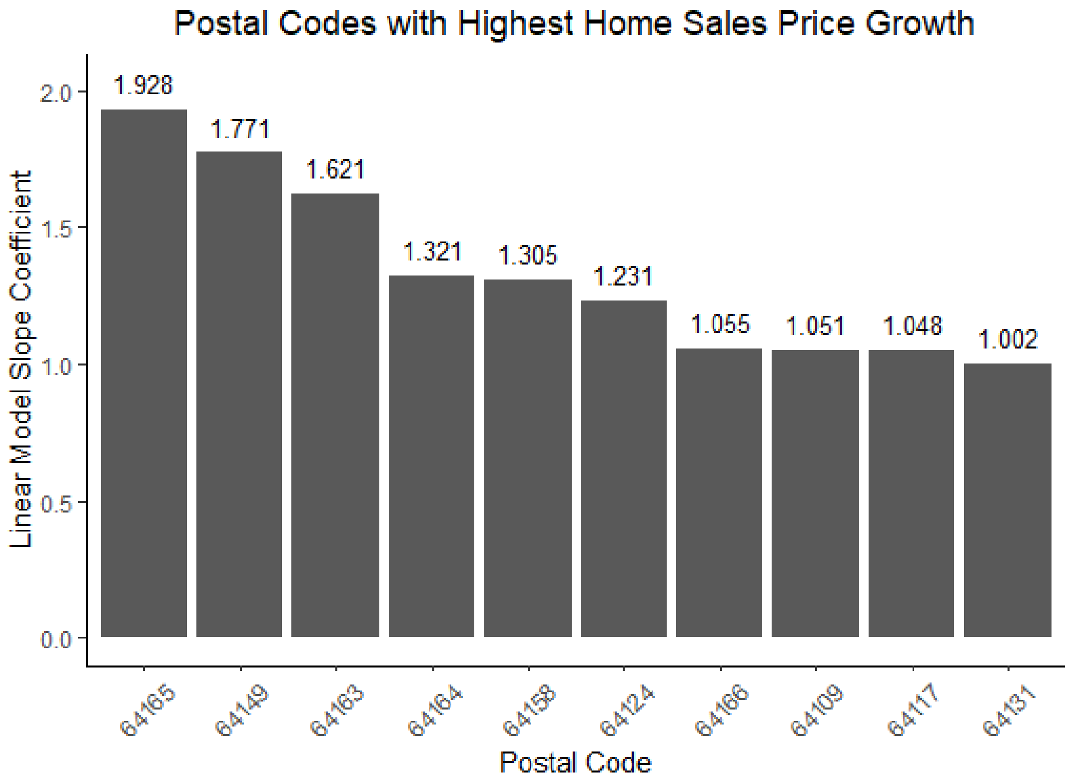
  - $\epsilon_{month.'postal\ code_i'}$ represents the error term, the discrepancy between model prediction and actual value for each combination of month and postal code

LINEAR MODEL EVALUATION – RESIDUALS ANALYSIS



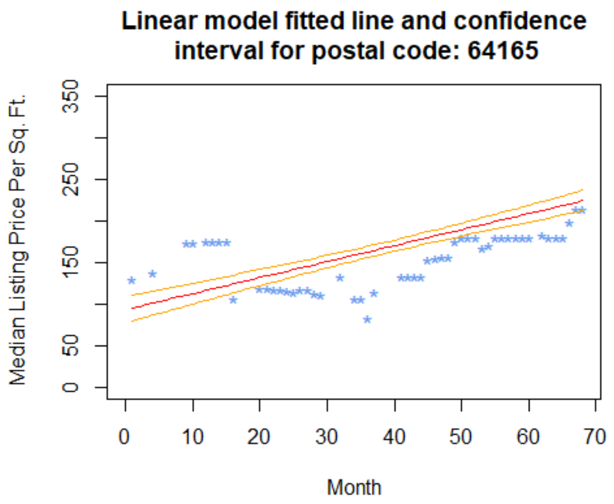
- Histogram plot of linear model error residuals
- Appears to be approximately normally distributed (except for some outliers in the right-tail)
  - Indicates that model is a good fit to the data

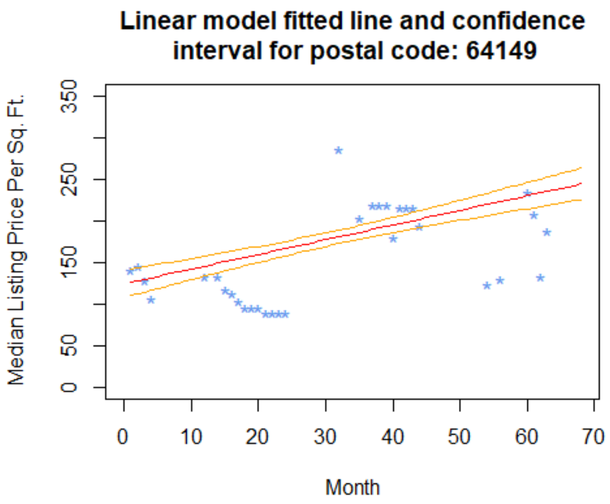
LINEAR MODEL EVALUATION – IDENTIFICATION OF BEST POSTAL CODE BY PRICE GROWTH (1/2)

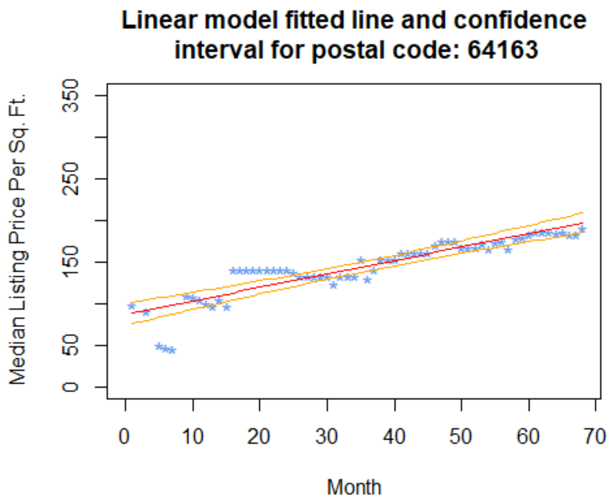


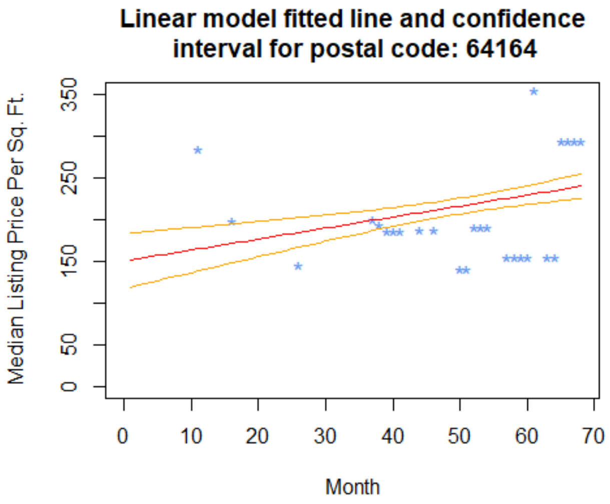
- Slope coefficients extracted from linear model and processed to combine base slope coefficient (β_1) with the individual slope ($\beta_{2,i} month.'postal\ code'_i$) for each postal code
- Extracted coefficients sorted in descending order to identify those with highest growth in home sales listing prices

LINEAR MODEL EVALUATION – IDENTIFICATION OF BEST POSTAL CODE BY PRICE GROWTH (2/2)









- Plots of data, fitted line, and slope coefficient confidence interval for each postal code
- Some non-linear patterns in the data are visible

LINEAR MODEL EVALUATION – IDENTIFICATION OF BEST POSTAL CODE BY PRICE GROWTH (3/3)

Postal Code	ß > 0 at 95% Confidence	ß > 0 at 99% Confidence
64165	✔ Yes	✖ No
64149	✔ Yes	✖ No
64163	✔ Yes	✔ Yes
64164	✖ No	✖ No

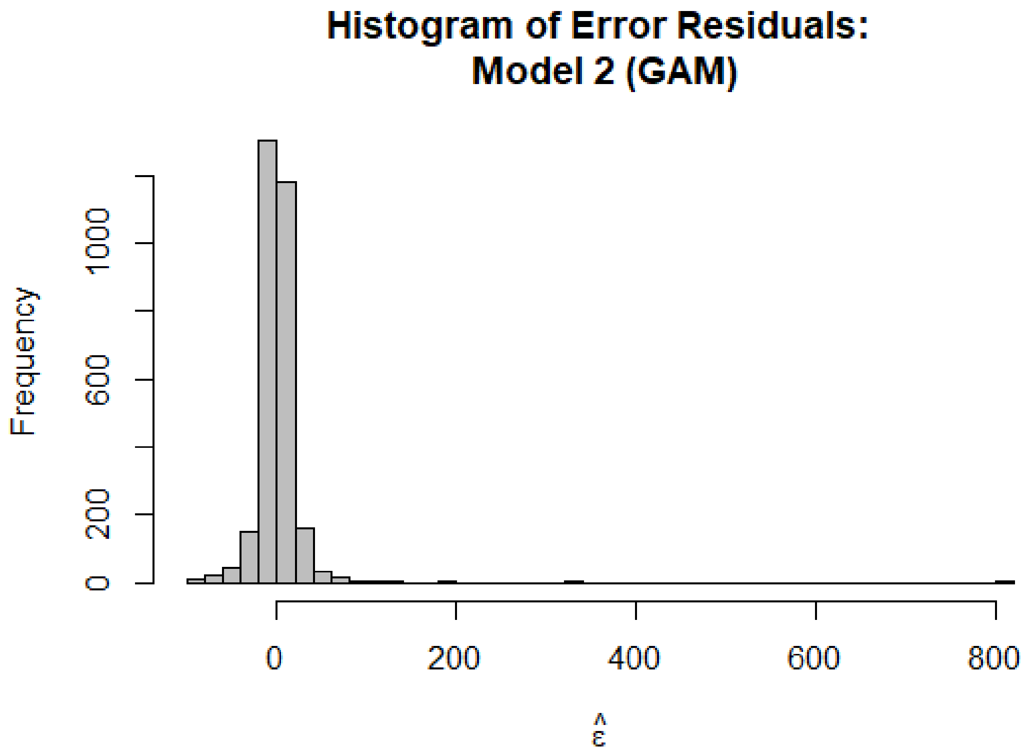
- Table of hypothesis testing results for each postal code, testing the null and alternative hypothesis:

$$H_0\colon \beta_2 = 0$$

$$H_a\colon \beta_2 > 0$$

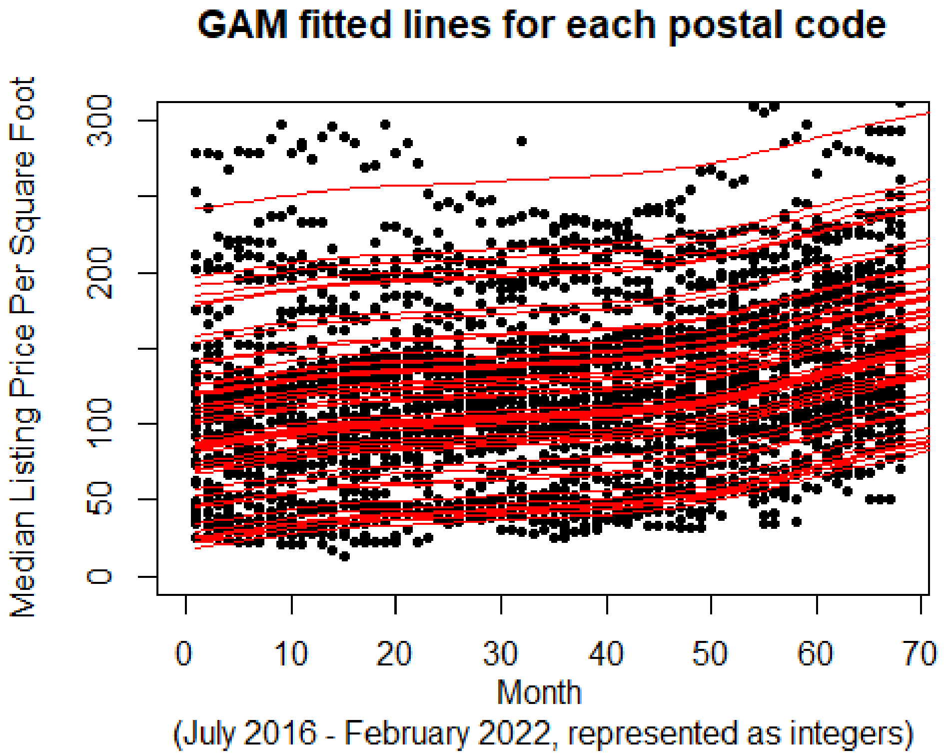
- Postal code 64163 is statistically the best, with the test statistic indicating that the slope in the linear model for this postal code is greater than zero at 99% confidence.

GENERALIZED ADDITIVE MODEL EVALUATION – RESIDUAL ANALYSIS



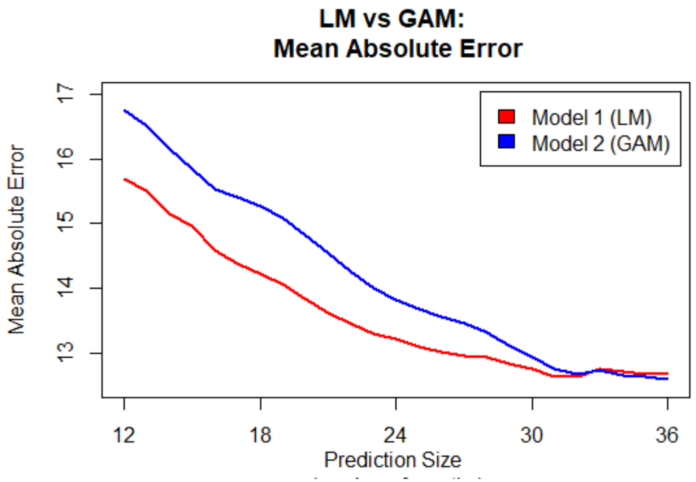
- Histogram plot of GAM model error residuals
- Appears to be approximately normally distributed (except for some outliers in the right-tail)
  - Indicates that model is a good fit to the data

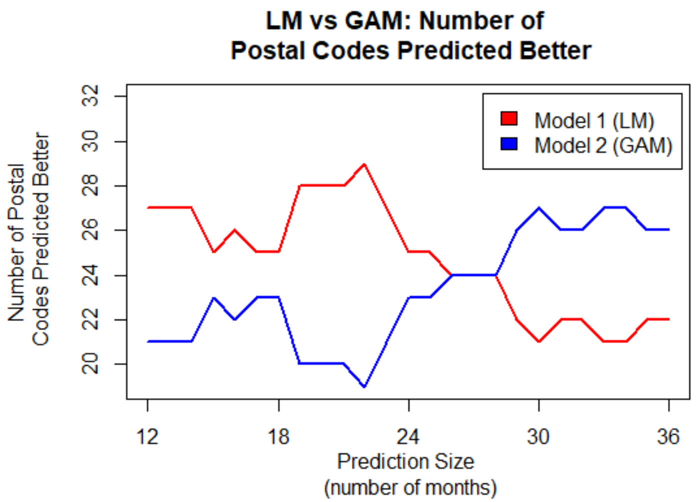
GENERALIZED ADDITIVE MODEL EVALUATION – ANALYSIS



- GAM is able to use spline curves to model non-linear relationships in the data
- Cannot differentiate slope between postal codes
  - Only the intercepts vary in this model

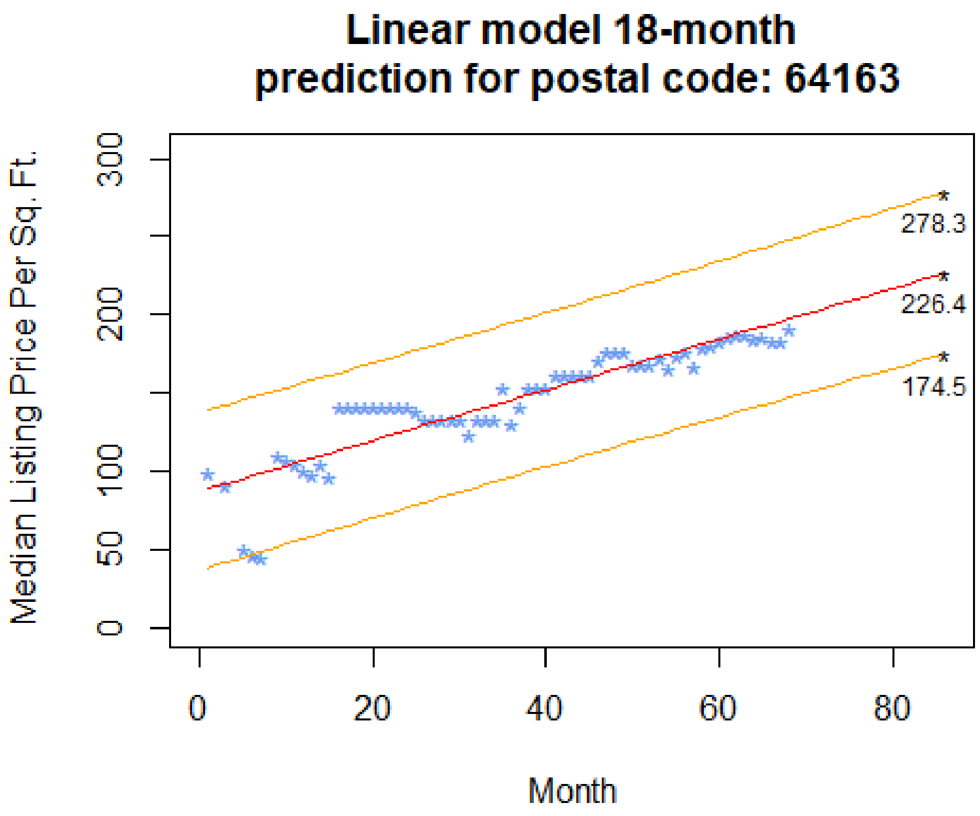
COMPARISON OF LINEAR MODEL VS. GENERALIZED ADDITIVE MODEL IN PREDICTIVE PERFORMANCE





- Comparison of LM versus GAM for prediction ranges between 12 and 36 months (by varying split between training and evaluation data sets)
  - At each prediction range, calculated absolute error and how many postal codes were better predicted by each model
- Linear model appears to have lower absolute error in general
- Around 28 months and longer prediction ranges, GAM appears to perform better (based on count of individual postal code performance)

RECOMMENDATIONS



- For KC Residential Investment Group 18-months at maximum seems like a reasonable estimate for the time to purchase, make improvements to, and sell a home
  - Utilizing linear model for predictions since it performed better at this prediction range
- With the best postal code identified (64163), the price predictions from the linear model are plotted
  - Showing predicted median price per square foot at +18 months (August 2023), with 95% prediction interval