Home Pricing Statistical Analysis

in Kansas City, MO

for KC Residential Investment Group

Zack Strathe
STAT 705
5-13-2022

Table of Contents

Executive Summary

Home property values can be difficult to predict, which is abundantly clear after the real-estate company Zillow lost $321 million in just the 3rd quarter of 2021 from their home-buying business segment, which utilized complex artificial intelligence-based models combined with their massive collection of real-estate data to estimate residential home values. With complex valuation models still being in their infancy and clearly in need of further refinement, I propose utilizing a simple linear model instead for KC Residential Investment Group to decide where to focus their efforts on purchasing, improving, and re-selling residential homes.

With a data set from Realtor.com, I developed a linear model (LM) that utilizes what is called the interaction effect between time and different postal codes within Kansas City, MO. With the interaction effect, the linear model outputs a different slope value for each unique postal code, making it easy to identify where home values are increasing the most. With further statistical analysis and hypothesis testing of the top identified postal codes, I determined that postal code 64163 should be the focus of KC Residential Investment Group's efforts due to a higher statistical likelihood of home values increasing in that area.

In addition to a linear model, since I observed some non-linear variation in the home pricing data over time, I also developed and evaluated a generalized additive model (GAM) to see if it provides more accurate predictions. In evaluations of the LM versus GAM, the LM was more accurate for short to medium predictions (≤ 28 months), while the GAM appears to be a better fit for longer term predictions (> 28 months). Since the maximum timeframe for buying a home, completing improvements, and selling a home would likely be less than 18 months, I concluded that the linear model would be best to use for the future pricing analysis of this report. In the most likely future scenario, a home purchased today would re-sell in 18-months for a 17.9% return on investment. And in the case that home values increase much more than expected, there could be a return on investment of 44.9%. However, it is possible that home values will decrease, and in that case, there could be a 9.1% loss on the investment.

To conclude this report, I also consider the future aspects of predictive modeling that KC Residential Investment Group should focus on. To build upon the linear model proposed in this report, it could be beneficial to add additional binary predictor variables which could help focus on which aspects of a property should be prioritized when making improvements in order to maximize returns. And while AI-based modeling for real estate valuation may still need refinement, it is likely that these types of models will continue to become more dominant in the market. Therefore, it would be wise for KC Residential Investment Group to focus more resources on collecting and storing as much data as possible, and to expand their data science and analytics staff to enhance data collection and storage practices, and to begin development of more advanced valuation models.

Introduction

The home pricing market is notoriously difficult to predict, but there is a large potential financial upside if one can manage to do it. While the popular methods among large real-estate firms are trending toward the utilization of complex models developed by artificial intelligence (AI) and machine learning (ML) methods, there are some inherent flaws in these models. Just recently in the 3rd quarter of 2021, the real-estate company Zillow incurred a net loss of $328 million attributable to its home-buying operations where they had purchased homes but incorrectly estimated a higher sales-price than they could obtain in reality. This blunder led Zillow to exit the home-flipping business entirely, due to having insufficient trust in the complex pricing model that they had developed (Levy, 2021). And while Zillow has resigned from the home-buying market, many other firms continue to push forward with AI and ML-based home pricing models, despite continuing losses from their operations. For example, another large competitor in this space, Opendoor, posted losses of $662 million in 2021 (Ponsford, 2022).

The problem inherent with attempting to model home values is that there are very many variables that can impact the overall price of a home. And while a complex ML pricing model may take into account thousands, or even millions, of variables, these models can be incomprehensible to understand how the pricing model actually works; they become a "black box" that we expect to predict accurately, which Zillow especially has demonstrated isn't always the case. And while developments in AI and ML in general show that there is still huge potential for these methods to become dominant in the real estate market, the biggest barrier to their development is a limit of useful data available for each individual home in the US. While these complex pricing models take into account school districts, tax history, or even Yelp reviews of local bars, among countless other factors, many such as Opendoor's pricing model still rely on human-annotation of data, especially individual home inspection data (Ponsford, 2022).

So, while the current trend may be to utilize complex modeling for home price prediction, I believe that the underlying technology and data-collection practices still need considerable refinement before it becomes a reliable method. Instead, I propose utilizing a simple linear model to identify real-estate opportunities. The benefit of using a linear model is that it is simple to understand how the model works and whether predictor variables have statistical significance to the response variable.

Specifically, for KC Residential Investment Group, who are seeking to develop statistical analysis of home property values in Kansas City, MO, I propose utilizing a linear model as a method for comparing different postal codes in the city. By identifying which postal code which has the greatest slope parameter for estimating the price, you will be able to narrow your search to a specific area of the city. And from there, you can utilize trained-experts to inspect homes within that postal code to identify opportunities for investment. Based on the analysis in this report, you will be able to identify—statistically—the area with the greatest potential for home price growth. And further, you will benefit from cost efficiencies with this approach, as you will require a smaller staff of home-inspectors than would be necessary if you were to consider all homes that are for sale within the city.

In addition, I consider the application of a more-complex model, a generalized additive model, versus the linear model for predictive use. Using the best model identified for prediction, this report deliers deliver an analysis of projected home prices in the postal code that is the most statistically probable for greatest home price growth. And finally, I consider the future development of further models, and how we might prepare to utilize AI-based models once the technology has been shown to be more reliable.

Methods

Data Set

The data set utilized for this analysis is from Realtor.com's Real Estate Data Library (Realtor.com, 2022). It is a .csv file containing monthly home pricing data for each postal code in the United States, and ranges from July 2016 through February 2022. It is based on a database of MLS-listed for-sale homes, with each row entry representing aggregate monthly data for each postal code. Each entry contains 37 data points; however, for this analysis I will only be utilizing the *"median_listing_price_per_square_foot",* "*postal_code*", and "*month_date_yyyymm*" columns. An important note regarding this data set is that it does not contain a value for every month for each postal code, and some postal codes contain very few data points at all. Therefore, it is expected that there will be an increased uncertainty in model fitting and prediction for those postal codes where data is sparse. So, while those postal codes with sparse data are unlikely to be in consideration for the purpose of this report; however, they will still be used to construct the models as they still provide useful information for comparative purposes.

One important assumption in this analysis is that the observations of the response variable *"median_listing_price_per_square_foot"* are normally distributed. Because each observation is an aggregate by postal code, the variance in observation values could be too erratic to provide a good generalization if data quality is low or lacking in quantity. Since this assumption is impossible to test, for simplicity of this analysis I will assume that the assumption is true. Though potential issues regarding the fit of a model could be attributable to this assumption being violated.

Data Pre-Processing

Since the data set contains information about every postal code in the United States, I first filtered the data set to only Kansas City, MO zip codes by utilizing the *subset* function in R. Next, to simplify analysis I removed any rows that contained no data for the response variable of interest ("*median_listing_price_per_square_foot*") by using the *drop_na()* function from the *tidyr* R library, and this reduced the number of zip codes in the data set from 55 to 50. Next, I reformatted the dates, since they are in "mmmmyy" format by default, which doesn't allow a linear model to recognize the change in time from year to year (for example, going from "202012" (December 2020) to "202101" (January 2021) looks like a change of 89 months to the model unless remedied). To yield a monotonically increasing value to represent the change in time, I remapped each unique month value to an integer instead. This was done by first extracting the unique month values from the data set in R, then utilizing a dictionary in R to map each of those unique values to an integer in a new column, starting with the integer 1 for "201607" and continuing to 68 to represent "202202". The last necessary data preparation step is converting the *"postal_code"* column from an integer into a character data-type. This is necessary for utilizing the postal code as a category, for utilizing in a linear model within R to get the interaction effect between time and each unique zip code.

Linear Model

I developed a linear model utilizing the *lm()* function in R to estimate the median price per square foot with predictor variables *"month"* and *"postal_code"* for response variable *"median_listing_price_per_square_foot":*

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

Where $y = (13, \dots, 1000)'$ is the median price per square foot from observations in the data set. The $x = (1, 2, \dots 67, 68)'$ is an integer encoding of the month, where 1 represents July 2016 and 68 represents February 2022. The parameters of the model include an individual intercept for each postal code ($\beta_{0,i}'postalcode'_i$), where for each postal code:

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

, which represents the modeled median price per square foot at month 1 (July 2016) for each postal code i. For the parameters of $\beta_1$and $\beta_{2,i}$, the model utilizes what is known as the interaction effect between the categorical variable "postal_code" and the month. This interaction effect slope allows the model to estimate an independent slope for each postal code (Faraway, 2014). The parameters include a base slope (β_1) which represents the slope for the first postal code in the data set: "64102" in this case. The second slope parameter ($\beta_{2,i}$) in the model represents the addition or subtraction to the base slope for all other postal codes where:

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

The $\epsilon = (\epsilon_{1.'64102'}, \epsilon_{2.'64102'}, \dots, \epsilon_{68.'64199'})'$ is the error term, representing the discrepancy between the model prediction for each combination of postal code and month versus the actual value. The parameters of the model were estimated using a maximum likelihood estimation approach. Due to a lack of data, the slope parameters $\beta_{2,64141}month.postal\ code_{64141}$ and $\beta_{2,64144}month.postal\ code_{64144}$ were unable to be estimated, as those postal codes each contain only 1 row of data.

To utilize the linear model for identification of the postal code with the largest increase in median price per square foot, I extracted the coefficients as a matrix from the model in R. I then performed some manipulation of that data: removed unnecessary columns, removed the intercept values, and added each postal code's coefficient value to the base value to get the true slope value for each, and then converted to a data frame for ease of further visualization.

Generalized Additive Model

By using the *gam()* function from the *mgcv* library in R, I developed a generalized additive model (GAM) that is better able to encode non-linear effects in the data . The model, utilizing the predictor variables *"month"* and *"postal_code"* is defined as:

$$y = \beta_{0,i}'postal\ code_i' + f_s(x) + \epsilon_{month.postal\ code_i}$$

Where $y = (13, \dots, 1000)'$ is the median price per square foot from observations in the data set. The $X = (1, 2, \dots 67, 68)'$ is an integer encoding of the month, where 1 represents July 2016 and 68 represents February 2022. The parameters of the model include an intercept (β_0) which represents the modeled median price per square foot at month 1 (July 2016) for each postal code, where:

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

The slope for the GAM model is represented by a smoothing spline function (f_s) on the month predictor variable (x). The $\epsilon = (\epsilon_{1.'64102'}, \epsilon_{2.'64102'}, \dots, \epsilon_{68.'64199'})'$ is the error term, representing the discrepancy between the model prediction for each combination of postal code and month versus the actual value. The parameters of this model were estimated using an approach to minimize the least squares fit to the data (Hastie & Tibshirani, 2014).

Results

Linear Model Evaluation

Calculating the error residuals of the model to test if they are normally distributed (that $\epsilon \sim N(0, \sigma^2 I)$) can be used for model checking (Faraway, 2014). Figure 1 displays a histogram of the error residuals for the linear model.

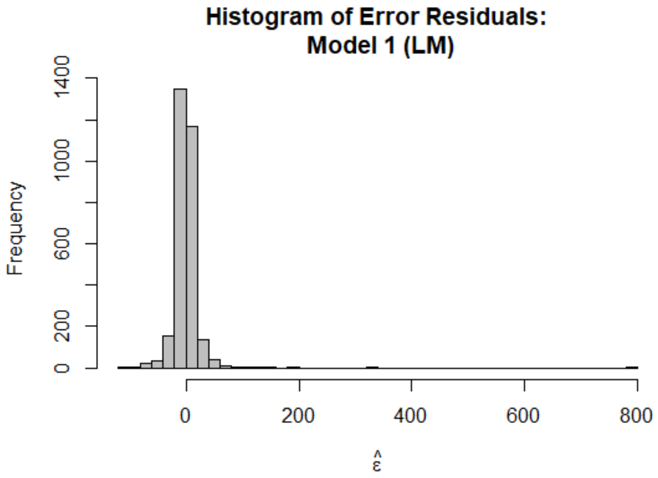


Figure 1: Histogram of Linear Model Error Residuals

While there are some extreme residual values seen in the histogram plot, the error residuals from the model appear to be approximately normally distributed with a mean of 0 and a constant variance. This indicates that overall, the linear model is a good fit to the data.

Identification of Postal Code with Highest Listing Price Growth

The benefit of utilizing a linear model for this analysis is that it allows for different growth rates in home listing prices to be extracted and compared. With this model, I transformed and sorted the slope coefficients to yield a table for the estimated linear growth rate for each postal code. Figure 2 displays the ten postal codes with the largest median listing price per square foot growth rate. These coefficients are interpreted as the increase in the median listing price per square foot per month. From viewing this figure, it’s apparent that postal code 64165 has the highest growth rate, while postal code 64149 and 64163 follow close behind.

To examine how well the model fits to the data for the highest growth postal codes, Figure 3 displays plots with fitted lines and the 95% confidence interval of the slope parameter coefficient for each. While the fitted lines for postal codes 64165 and 64163 appear to be a good fit to the data, the fit to postal codes 64149 and 64164 appear to be weaker due to a clear non-linear pattern of the actual median listing price for these postal codes. With these non-linear patterns showing in the data, it signals that maybe another model than the traditional linear model would be a better fit to this dataset. Accordingly, I am also evaluating a Generalized Additive Model (GAM), which can better represent non-linear variations from the underlying data.

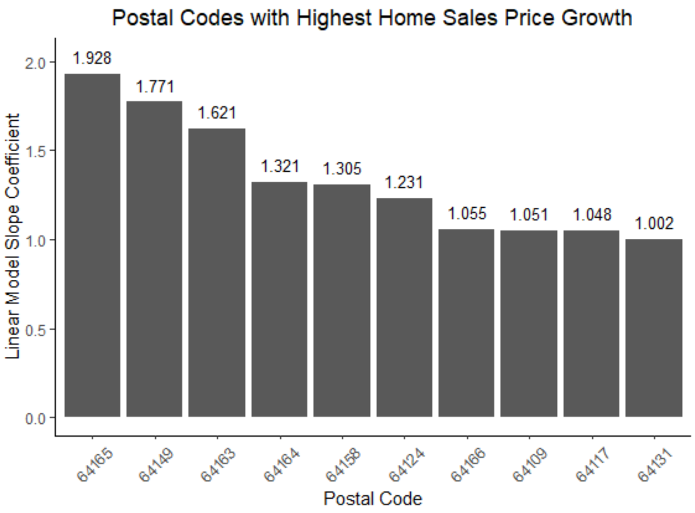


Figure 2: Plot showing the top postal codes by slope coefficient estimations from the linear model

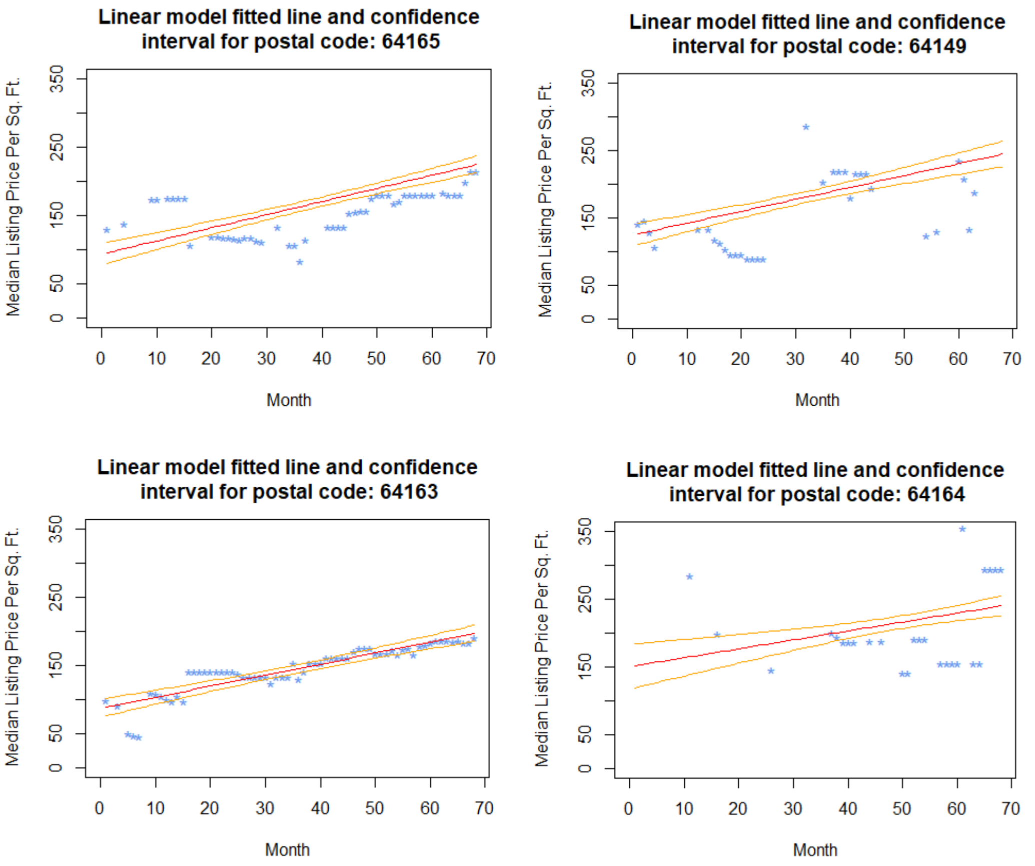


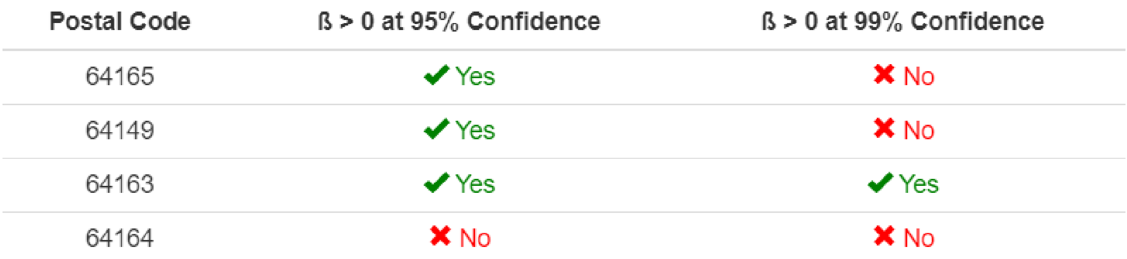
Figure 3: Plots displaying fitted line and 95% confidence intervals for top four postal codes identified from the extracted linear model slope coefficients

While Figure 2 shows that postal code 64165 clearly has the highest growth in home listing price, I performed hypothesis testing on each of the top four postal codes (64165, 64149, 64163, and 64164) to test the null and alternative hypothesis:

$$H_0: \beta_2 = 0$$

$$H_a: \beta_2 > 0$$

, where β_2 is the slope parameter for the interaction effect of postal code and month in the linear model. In order to individually test each postal code, a linear model for each was generated and tested using a subset of only the data pertaining to that postal code. If the test-statistic returned was less than 0.05, then it was concluded that the null hypothesis can be rejected, and that $\beta_2 > 0$ at 95% confidence. Similarly, if the test statistic returned was less than 0.01, then it was concluded that the null hypothesis can be rejected at 99% confidence. Figure 4 shows the results of this hypothesis testing.



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

Figure 4: Table of one-sided hypothesis testing results that the slope coefficient is greater than zero for top four postal codes

The results shown in Figure 4 indicate that postal code 64163 is a safer bet to focus real estate investments into homes, because hypothesis testing reveals that the slope coefficient is greater than zero at 99% confidence. Meanwhile, hypothesis testing for the other top four postal codes indicates that the slope coefficient is positive at 95% confidence but not 99% confidence. And testing for postal code 64164 indicates that we cannot estimate a positive slope with either confidence level.

Generalized Additive Model Evaluation

Similar to the evaluation of the linear model, the fit of the GAM can be evaluated with a plot of the error residuals, shown in Figure 5. And similarly to the linear model, the histogram plot of error residuals for the GAM shows an approximately normal distribution (that $\epsilon \sim N(0, \sigma^2 I)$), with a small number of extreme outliers. Therefore, the GAM also appears to be a good fit to the data.

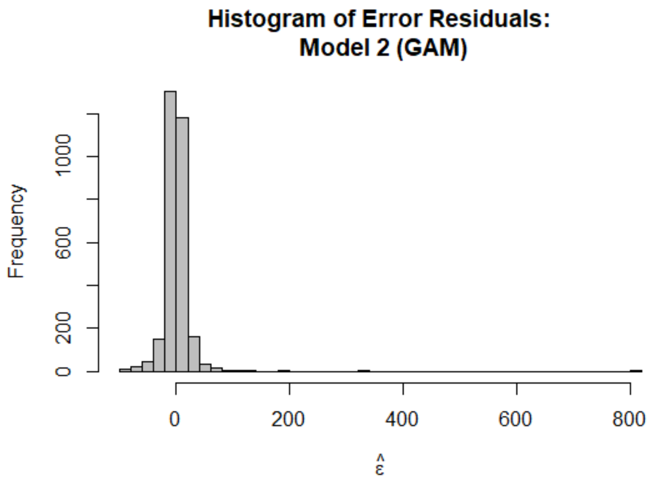


Figure 5: Histogram of Generalized Additive Model Error Residuals

One important consideration for the GAM is that it does not allow for the option of an interaction effect between the postal code and month, so it cannot be used to differentiate the slope coefficients for postal codes, as was done in the previous subsection of this analysis with the linear model. With the GAM, each postal code is assumed to have the same growth rate of home median listing prices (which in reality we do not believe to be true). This effect is visualized in Figure 6, showing the GAM fitted line plotted for each postal code individually. This plot shows each fitted line having the same curve, but with varying intercepts. And while the GAM is not too useful for finding the highest growing postal code, it can still be beneficial in prediction, to estimate the median home listing values in the future. So, the next subsection compares the predictive performance of the LM versus GAM.

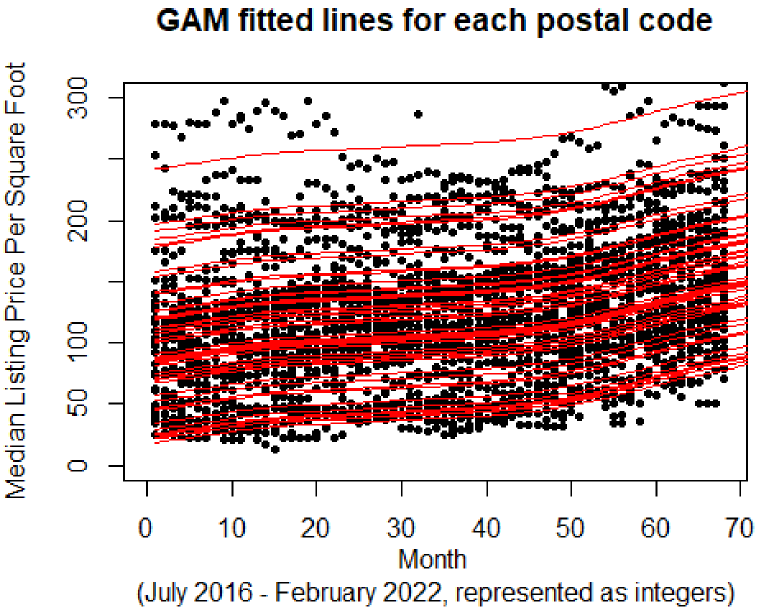


Figure 6: Plot of Generalized Additive Model fitted lines for each postal code

Comparison of LM and GAM Predictive Performance

In addition to identifying the postal code with the highest-growth rate in median home listing prices, I'm also seeking to utilize a model for future prediction of listing prices to determine the potential return on investment that could be expected for a home that we buy to fix and sell. Therefore, both the LM and GAM were evaluated on predictive accuracy from 12-month through 36-month prediction sizes, with the data being split for model fitting and evaluation. In the case of the 12-month prediction size, the model was fit with the first 56 months of data, then predictions form that model evaluated against the last 12-months of data. And similarly, the 36-month prediction size uses the first 32 months of data for model fitting and the last 36 months for model evaluation. For evaluation at each prediction size, each postal code was predicted with both the LM and GAM models, with the mean absolute error for each model calculated over all observations, as well as a count of postal codes that were better predicted for each model. The results of the model comparisons are displayed in Figures 7 and 8 respectively.

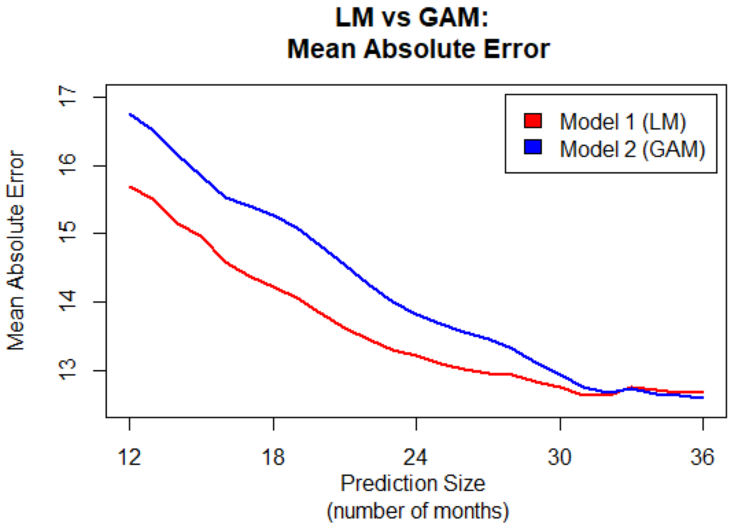


Figure 7: Plots displaying the mean absolute error for the LM and GAM models at each prediction size

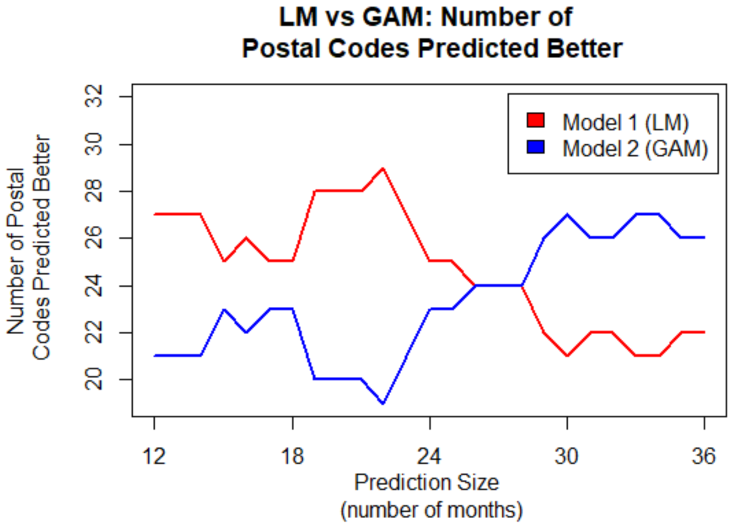


Figure 8: Plots displaying the number of postal codes that each model predicted better, at each prediction size

For short and medium-term predictions (≤ 28 months), it appears that the linear model has slightly better predictive performance than the generalize additive model. With a 12-month prediction range, the linear model is able to better predict listing price values for 27 postal codes, while the GAM predicts better for 21 postal codes, and the LM has a 1.07 lower mean absolute error at 15.7 versus 16.77 for the GAM. With a 24-month prediction range, the linear model is able to better predict listing price values for 25 postal codes, while the GAM predicts better for 23 postal codes, and the LM has a slightly lower mean absolute error at 13.21 versus 13.81 for the GAM.

For longer-term predictions (> 28 months), it looks like the GAM is likely superior to the LM. At the 36-month prediction size, the linear model is better able to predict home values for 22 postal codes, while the GAM is better at prediction for 26 postal codes. Though, the difference in mean absolute error between the models is only 0.07, with the LM achieving a mean error rate of 12.67 while the GAM achieved a slightly lower mean error rate of 12.60.

Discussion & Recommendations

Prediction of Future Home Listing Prices

As I concluded in the previous analysis of the slope coefficients for each postal code, postal code 64163 is a safer bet for residential real estate investments due to a higher statistical likelihood of the median price per square foot to be increasing over time. And to predict the future price of a home, the preferred model to use seems to differ based on the prediction range considered. In the case of KC Residential Investment Group, I believe that 18-months is a reasonable estimate for the maximum time-span that it would require to purchase, make improvements, and sell a home. Therefore, since the linear model seems to predict better at intervals of 18-months and less I will use it for price predictions in this section.

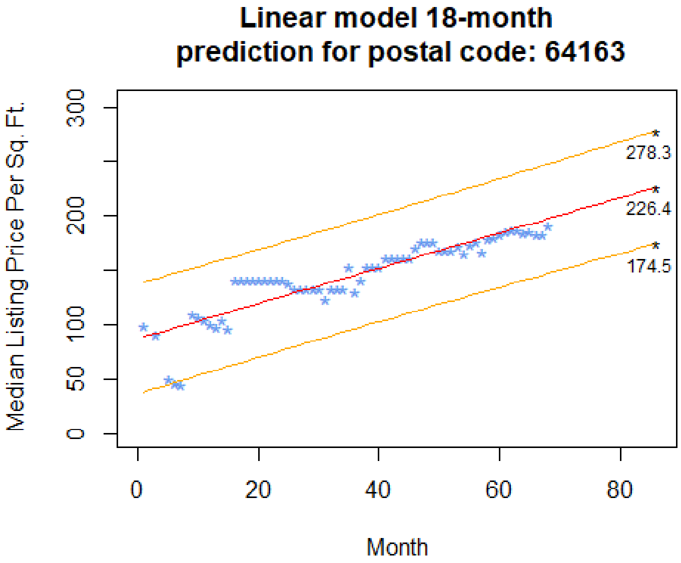
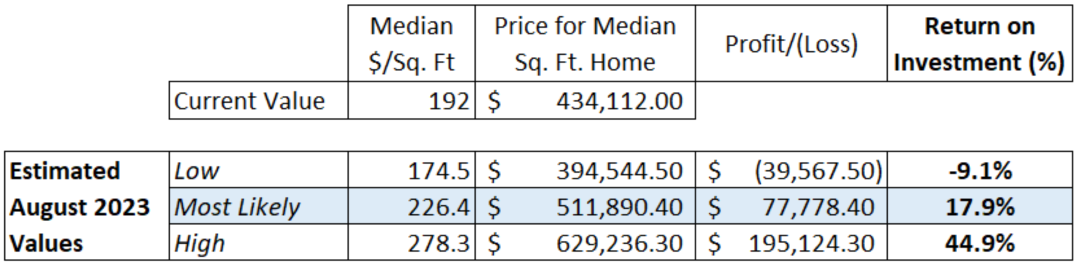


Figure 9: Plot displaying the linear model 18-month (August 2023) median price per square foot estimate with a 95% prediction interval for postal code 64163

To estimate the potential return-on-investment potential in this analysis, for calculating an approximate dollar value I am using the median square footage of a single-family home in the US, which is 2,261 square feet as of 2020 (Statista, n.d.). The estimated investment returns as of August 2023 for a median square footage house in postal code 64163, are summarized in Figure 10. These estimated values are utilizing the 95% confidence prediction intervals from the linear model. In the most likely scenario, a home purchased today would re-sell in 18-months for a 17.9% return on investment. And in the case that home values increase much more than expected, there could be a return on investment of 44.9%. However, it is possible that home values will decrease, and in that case, there could be a 9.1% loss on the investment.



		Median $/Sq. Ft	Price for Median Sq. Ft. Home	Profit/(Loss)	**Return on Investment (%)**
	Current Value	192	$ 434,112.00		
Estimated August 2023 Values	*Low*	174.5	$ 394,544.50	$ (39,567.50)	**-9.1%**
	Most Likely	226.4	$ 511,890.40	$ 77,778.40	**17.9%**
	High	278.3	$ 629,236.30	$ 195,124.30	**44.9%**

Figure 10: Table of Estimated Return on Investment values for a median square footage home in postal code 64163 using 95% confidence interval predictions from the linear model

Future Models to Consider

Though the results of this analysis seem promising, there are undoubtedly improvements that can build upon it. For instance, another linear model could be considered with the addition of one or more binary predictor variables, such as where there is a swimming pool on the property, whether the driveway is cracked, or whether the laundry room is on the main living floor, etc. Such additions to this model would help KC Residential Investment Group to have a better understanding of where residential property renovation efforts should be directed in order to maximize the potential return on investment from a property.

While the linear model is well-understood, and the proposed linear model seems to be a safe bet for comparing and predicting future home values, it seems likely that artificial intelligence (AI) based valuation methods will continue to become more refined and dominant in the real estate investment market as time goes on. Therefore, the best way to compete in the future is to begin developing AI valuation methods of your own. Begin dedicating resources to collecting and storing as much data as can possibly be collected from property inspections, especially photos which could be computer analyzed to collect facts about a property which may be too numerous and cumbersome for someone to manually record (i.e., whether there is crown moulding in a room, whether the ceilings are textured, what type of insulation is present in the attic, what type of joists are used, etc.). In addition, it would be wise to expand your data science and analytics team so that you are maximizing your ability to maintain stored data and develop improved models.

References

Faraway, J. J. (2014). *Linear Models with R: Second Edition.* CRC Press.

Hastie, T., & Tibshirani, R. (2014). *Generalized Additive Models.* Retrieved from Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels): https://doi.org/10.1002/9781118445112.stat03141

Levy, A. (2021, November 3). *Zillow plunges 25% to lowest since July 2020, after company exits home-buying business*. (CNBC) Retrieved from https://www.cnbc.com/2021/11/03/zillow-stock-plunges-24percent-after-company-exits-home-buying-business.html

Ponsford, M. (2022, April 13). *House-flipping algorithms are coming to your neighborhood*. (Technology Review) Retrieved from https://www.technologyreview.com/2022/04/13/1049227/house-flipping-algorithms-are-coming-to-your-neighborhood/

Realtor.com. (2022, February). *Monthly Historical Data by Zip.* Retrieved from Realtor.com Real Estate Data Library: https://www.realtor.com/research/data/

Statista. (n.d.). Retrieved from Median size of single family housing unit in the United States from 2000 to 2020 : https://www.statista.com/statistics/456925/median-size-of-single-family-home-usa/