**Module 7 Capstone Project**

Jorge Alvarez

Colorado State University Global

MIS581: Capstone – Business Intelligence and Data Analytics

Dr. Mills

September 8, 2024

**Module 7 Capstone Project**

**Abstract**

Since the 2008 Financial Crisis, housing demand has outpaced housing supply in the United States. One of the fastest growing states in the U.S. is North Carolina, which is also experiencing an unbalanced housing market. Most scholarly analysis of the housing market focuses on characteristics of the homes as predictors of a home’s value. This project explores forecasting housing prices from a supply and demand perspective that focuses on population as a predictor. Our findings will show that, in North Carolina, population alone is insufficient for predicting housing values with statistically significant accuracy. We recommend growing the model for demand of housing by accounting for external factors like interest rates, or adjusting for demographic breakdown such as number of working adults instead of total population.

**Introduction**

According to a recent report on National Association of Realtors (“NAR”) data, the housing inventory supply of homes for sale sits at about 4.1 months. This is a measure of how long current “for sale” homes would take to sell at the current pace, if no additional homes were to be put up for sale. It is believed that a “balanced” housing market would have 5-6 months’ worth of inventory. This means that the housing market is still “tight.” This reflects an imbalance between demand for housing and supply for housing. This project will explore a possible simple way to forecast future housing values based on population figures, which would help business leaders and policymakers to plan for housing demand at the local level. The hope is that such knowledge and predictive power would bring the housing market closer to balance.

**Objectives**

Forecasting future housing values, indeed future values of anything, is the purpose of many data analytics endeavors. Predicting future housing values with some degree of statistical significance would empower stakeholders to prepare for housing demand as population increases in the United States continue. North Carolina is forecasted to become the seventh most populous state in the union by the year 2030. This will come with population growth that could put a strain on existing housing. Existing housing supply would take less than six months to completely sell out if no additional homes were listed for sale ("North Carolina housing market," 2024). This six-month mark is typically associated as the balanced point between selling and buying home. Since the 2008 financial crisis, home inventory has been less than the six-month mark, which indicates pent-up housing demand. Providing sufficient housing helps to keep prices affordable and secures housing for all segments of the general population. The 2008 financial crisis exposed the dysfunction between housing demand and housing supply.

**Overview of Study**

In the state of North Carolina, the Office of State Budget and Management (“OSBM”) oversees the state’s budget process. One of the ancillary functions of this main objective is to analyze economic and demographic data to inform the budget process ("NC OSBM: About OSBM," 2024). The data is open license, and it is customizable. Since my goal is to fashion a predictive model of housing prices based on population projections, I began by exploring datasets at OSBM as well as from the U.S. Census.

There were several informative datasets at OSBM, but many provided information at the county level regarding housing units and population. However, a lot of the real estate valuation datasets I have seen provide information at the metro level or at zip code level. The resolution to getting these two geographic areas to be comparable was to get population data at the zip code level. The metro level would exclude many rural and/or remote counties and areas, making a predictive model less adaptable to all parts of North Carolina.

Fortunately, the U.S. Census provides a breakdown of its data into geographical area called Zip Code Tabulation Areas (“ZCTA”) (US Census Bureau, 2023). These areas have not changed since the 2000 Census. I found a dataset that provides population estimates for ZCTA’s in North Carolina from 2011 through 2022. I was able to join this dataset with data regarding median home values in zip codes throughout North Carolina for the same time period.

The combined dataset has six independent variables and one dependent variable. The six independent variables include zip code, municipality, metro area, county, population in the zip code, and year of the estimates. The dependent variable is the median value of homes in the zip code that year. Zip code, year, and population are integer variables, while municipality, metro, and county are text variables used to categorize the data. There is some covariance with these text variables and with the zip code integer variable since they all describe the geographic location.

In recent years in the United States there has been a lack of sufficient housing production. It is estimated that millions of more homes would be needed for housing demand from immigration and organic population growth ("2022 housing Underproduction in the U.S," 2022). From a supply and demand perspective, this means that supply is not keeping up with demand, and demand is exceeding supply. When this happens, prices are pushed up, all else being equal. This results in homes being less widely available and more competition for available homes. Another consequence of excessive demand / insufficient supply is fewer options for people looking to change cities or neighborhoods.

**Research Question and Hypotheses**

This project aims to provide a scalable model for predicting housing prices and price growth based on population forecasts below the county level, specifically at zip code level. This may provide an understandable basis for policymakers and stakeholders to plan for housing needs in the intermediate future and beyond. The aggregated data provides median housing prices and population counts in zip codes throughout the state of North Carolina in recent years. The first question to answer is whether population levels can be used as a proxy for housing demand to predict housing values. The null hypothesis would be:

Ho: Population levels do not correlate in a statistically significant way to housing prices.

And the alternative hypothesis to answer this question is:

Ha: Population levels do correlate to a statistically significant degree with housing prices.

For our tests, we will begin by testing against a level of significance of alpha = 0.05. By testing these hypotheses, this paper will develop a proxy for housing demand, and in particular, future housing demand based on population needs. We can test this hypothesis by running a simple linear regression on population and housing values in the data. A correlation analysis will also be useful here. A linear regression model is one approach. This will be used as part of a predictive model that will be usable by state, county and municipal leaders to plan for housing needs. The predictive model will provide results on housing prices, and year over year figures could provide insights into housing inflation. Inflation that is too high or an upward outlier would signal to stakeholders that they should plan for more housing if they want a balanced housing market in the future. The predictive model can be tested using the following hypotheses:

Ho: The predictive model is not statistically better than chance at predicting housing values.

Ha: The predictive model is better than chance at predicting housing values.

For our tests, we will begin by testing against a level of significance of alpha = 0.05.

This research question can be tested using several predictive modeling techniques. The data includes zip code location, county location, municipal location, population figures in a given year, and median housing value in the same given year. Different years are recorded in different records. This will allow use to weigh accuracy in the model on whether a housing market is in a municipality (versus some unincorporated area), and whether some counties have more accurate figures and thus have a higher level of confidence. The zip codes perhaps can be grouped by the first two, three, or four digits. For this, the paper will focus on testing several techniques, such as the general linear model, gradient boosted model, and a neural network. The general linear model can see how different variables affect a continuous dependent variable. For the purposes of this project, the independent variables each affect the continuous value of homes in each zip code. A gradient boosted model would use multiple decision trees to try to predict values.

A neural network with or without clustering could also be tested as a viable technique. The data will be divided into a training subset and a validation subset. A separate set of forecasted prediction values will be used to make actual predictions. The training subset will be saved after seeding so that the same subsets can be used to train the various models and to validate those models. With the same subsets having been seeded, the different models can be compared to one another to see which one or ones are the most accurate. If the subsets had been unique, we would not be able to rule out that the unique training data for each model contributed to its accuracy. In other words, we can compare the models’ accuracies directly when the training subsets are identical.

One concern with the data is the difference between zip code housing values and Zip Code Tabulation Areas (ZTCA) used by the US Census Bureau to count population in the area. Both zip codes and ZTCAs share five-digit unique identifiers. Also, ZTCA’s have not changed since the 2000 US Census, but zip codes have been modified since then (Krieger et al., 2002).

Using R, I explored the dataset in several ways. First, I looked at the first several lines to look at how the data is formatted. Using the code below, we can see that the data is grouped by zip code and by year, with a combination of these providing the unique entry of each row.

### Read in Data  
setwd("C:/Users/jorge/OneDrive/Desktop/581")  
  
valuesandpopulation.data <- read.csv("NC house values and population by zip.csv")  
  
### Run 5 Functions to Explore Data  
library(visdat)

## Warning: package 'visdat' was built under R version 4.3.1

dim(valuesandpopulation.data)

## [1] 8640 8

Sys.Date()

## [1] "2024-08-12"

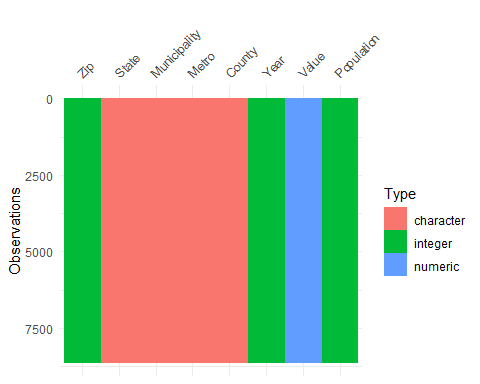
head(valuesandpopulation.data)

## Zip State Municipality Metro County  
## 1 27006 NC Advance Winston-Salem, NC Davie County  
## 2 27007 NC Ararat Mount Airy, NC Surry County  
## 3 27009 NC Belews Creek Winston-Salem, NC Forsyth County  
## 4 27011 NC Boonville Winston-Salem, NC Yadkin County  
## 5 27012 NC Clemmons Winston-Salem, NC Forsyth County  
## 6 27013 NC Cleveland Charlotte-Concord-Gastonia, NC-SC Rowan County

## Year Value Population  
## 1 2011 225382.7 13764  
## 2 2011 0.0 1874  
## 3 2011 167758.2 3038  
## 4 2011 0.0 5777  
## 5 2011 214939.5 26509  
## 6 2011 122065.0 6066

I then used the following code to visually see how the data variables are structured.

visdat::vis\_dat(valuesandpopulation.data, sort\_type = FALSE)



We can see from this graphic that there are well over 7,500 observations.

Finally, using R’s summary function, as below, we can see the ranges of the numerical variables and the means and medians.

summary(valuesandpopulation.data)

## Zip State Municipality Metro   
## Min. :27006 Length:8640 Length:8640 Length:8640   
## 1st Qu.:27605 Class :character Class :character Class :character   
## Median :28102 Mode :character Mode :character Mode :character   
## Mean :28058   
## 3rd Qu.:28517   
## Max. :28909   
## County Year Value Population   
## Length:8640 Min. :2011 Min. : 0 Min. : 37   
## Class :character 1st Qu.:2014 1st Qu.: 106016 1st Qu.: 2613   
## Mode :character Median :2016 Median : 151580 Median : 7292   
## Mean :2016 Mean : 172896 Mean :13760   
## 3rd Qu.:2019 3rd Qu.: 220052 3rd Qu.:20752   
## Max. :2022 Max. :1796977 Max. :81219

Notably, the value variable, which represents the median value of homes in that zip code in that year, has zero as its minimum value. Of course, this is really an absence of data for some entries, and our analysis will need to account for these.

This dataset was chosen because it provided population (demand for housing) as well as median value (supply for housing) at the zip code level. Most of us can recognize that housing values vary greatly within a county. Within a single county there can be several different neighborhoods with wildly different property values. For this reason, it seemed prudent to choose breakdowns sub-county, such as at the zip code level. The municipality level might also leave some remote and/or rural areas off the analysis. By using zip code data, we hope to be able to provide targeted predictive estimates that can be the basis of high-level planning or strategy within the housing market.

**Literature Review**

In reviewing literature for this project, I started by thinking about what it is I’m trying to do. In trying to give local leaders a tool to predict future housing needs based on population, I looked for issues specifically related to the State of North Carolina. Different states have different laws governing expansion, and different challenges. North Carolina specifically has tools at state, county, and local levels for encouraging housing development. For example, inclusionary housing, or making a few units of new housing as affordable, is one tool. Often, land use controls are the way local leaders can control new builds. Several programs throughout North Carolina’s cities provide tax rebates or investments or low-cost loans for encouraging housing growth. Matching these local-level controls with the future need for housing is the balancing act. In North Carolina, solutions should be tailored to the local real estate market and this means that housing supply initiatives should be addressed at the sub-county level (Amend, P., et al, 2019, p. 20). Indeed, many housing datasets center around county-level statistics, even while many counties contain multiple municipalities.

To create a predictive model of housing prices based on population figures, I had to look at the connection between the two. The first thing that came to mind was the measurement of Persons Per Household (PPH). I found that this statistic changes slowly over time, due to the interplay of various factors (Swanson & Hough, 2012, p. 244). My premise is that population is a proxy for housing demand. A relatively steady PPH which changes slowly over time means that higher populations will generally demand more housing.

Use of census data to try to predict median value of housing has been performed before with data in California and using Machine Learning (E. Simlai, 2021, p. 306-307). Prior attempts were noted to use multiple linear regression and other techniques. The data used here was also US Census data, as I am using. These authors also researched at the sub-county level, using Census tracts. One difference between our datasets can be seen in the number of subdivisions throughout the state – North Carolina has 100 counties, 1,079 zip codes, and 2,195 census tracts. This makes the census tracts the most granular level of the datasets.

I will continue to do more research, as the amount of data available at the census tract level would lend itself to machine learning methods if I can obtain housing estimates for census tracts throughout North Carolina. This research paper is very interesting as it sought objectives similar to what I seek, albeit in another state. Preliminarily, I know that data is readily available for several years from the American Community Survey regarding housing characteristics at the census tract level. Characteristics such as the year the structure was built, and how many bedrooms are in the structure, and other variables.

Another paper I read weighed three methods of modeling housing values a Boston suburb (Mu et al., 2014). In this paper, they determine that the Linear Regression of SVM and Linear Regression of LSSVM performed better than a partial least squares method. This paper also heralded the favorability of machine learning methods for predicting housing values.

From my research, it appears that using several variables that can contribute to a home’s value means that machine learning can be utilized to find connections in the data. In fact, the algorithm can learn from the data. Several different types of machine learning can prove effective, and trying different ones might be of use.

Most importantly though, my research has led me to question the dataset I currently have. I will research further and see whether I can obtain data that can inform about census tracts median values over time. This would serve my aims better, namely, to give granular level predictive capabilities to local planners in the state of North Carolina, based at least in part on population demands.

The limitations of this project include that these projections will be used to forecast median housing prices in geographical areas in North Carolina based on population estimates for those areas. Since there will not be data about characteristics of individual homes, such as number of bedrooms, number of bathrooms, nor any such housing characteristics, price prediction for a particular type of home will not be available. Simply, the median price point, where half the homes cost more and half cost less, will be made available.

The median does not usually speak to the spread of prices. However, using standard deviations or variances for a geographic area might give some ideas as to the spread of prices around a given median value. Again, some information about what sorts of homes are available around the median value, or how many bedrooms a median-priced home might have, will not be generated from this project.

This leads to the possibility that this information might be used (or misused) to create high-value homes that balance out low-value homes. Using only a median can contribute to a stark dichotomy of housing for very high-end and very low-end homes, instead of providing a continuum or spectrum of housing values that better serve a balanced community where most of the population is middle class. Given the history of the United States overall, such a situation could disproportionately affect minorities. So, the intent of this metric and predictive model is not to provide a “silver bullet” solution for predicting future housing values in North Carolina. Rather, the intent is to provide another metric for stakeholders to check their progress so they may adjust their policies as needed in this area.

**Findings**

To explore the data, we can see that there is a weak positive correlation between population and housing values. This is the code that was used in SAS Studio:

/\* Task code generated by SAS Studio 3.8

\* Generated on server 'ODAWS01-USW2-2.ODA.SAS.COM'

\* Generated on SAS version '9.04.01M7P08062020'

\* Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:130.0) Gecko/20100101 Firefox/130.0'

\* Generated on web client 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en\_US&zone=GMT-04%253A00&ticket=ST-100887-RUeXfvGocrhKYi5eTyex-cas'

\*

\*/

ods noproctitle;

ods graphics / imagemap=on;

proc corr data=WORK.'HOUSES AND PEOPLE'n pearson nosimple noprob

plots=matrix(histogram);

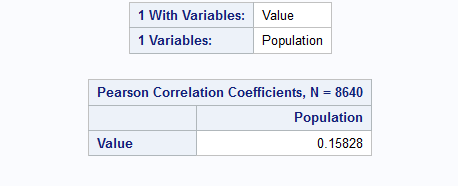
var Population;

with Value;

run;

Figure 1 on the following page shows the results of exploring the relationship between the predictor variable population and the dependent variable value, aka housing prices. The Pearson Correlation Coefficient for this relationship is 0.158 which is weak but positive correlation. Recall that the correlation coefficient is between -1 and 1 and it indicates the magnitude and direction of correlation between variables.

**Figure 1.**



To make predictions, our project requires a predictive model. Since we are seeking a simplified model to help in a broad range of contexts, we chose to use a linear regression model. The data we aggregated had population and median house values by zip code, with samples for different years. The years were a variable with its own column. The simple approach was to test to see how well a linear regression can receive as input population figures and then predict housing values in dollars. To begin, SAS Studio was used to import the data from csv file format into a usable SAS format. In order to facilitate the process, the data was randomly sampled to create subsets for training, validation, and testing. The following code was used to train, test, and validate a predictive regression model.

\* Task code generated by SAS Studio 3.8

\* Generated on web client 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en\_US&zone=GMT-04%253A00&ticket=ST-89174-7ljESwmHFrSwsdwcztT7-cas'

\*/

ods noproctitle;

ods graphics / imagemap=on;

proc glmselect data=WORK.'HOUSES AND PEOPLE'n

outdesign(addinputvars)=Work.Glmselect\_Design plots=(criterionpanel);

partition fraction(validate=0.3 test=0.3);

model Value=Population / selection=stepwise

(select=sbc) hierarchy=single;

run;

proc reg data=Work.Glmselect\_Design plots(only)=(partial observedbypredicted);

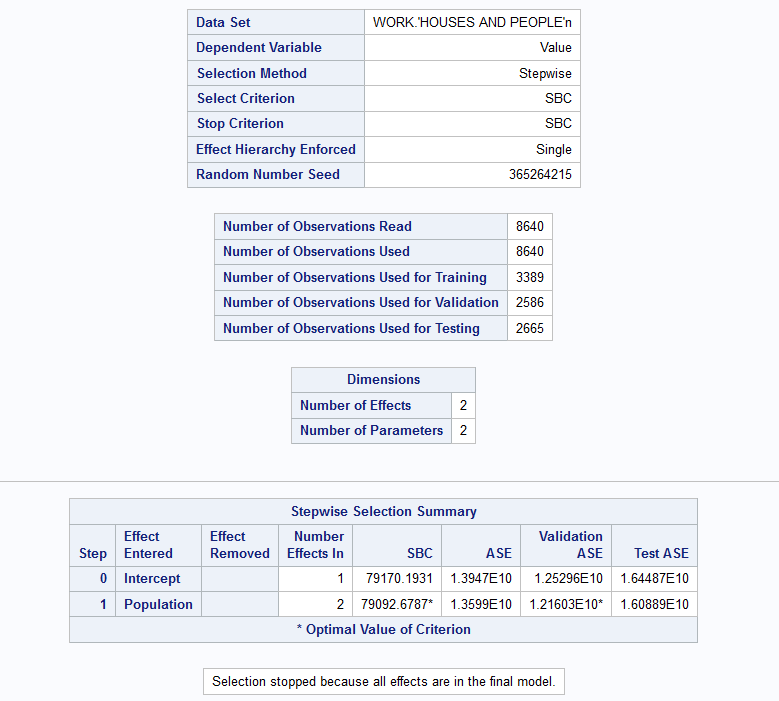
ods select PartialPlot ObservedByPredicted;

model Value=&\_GLSMOD / partial;

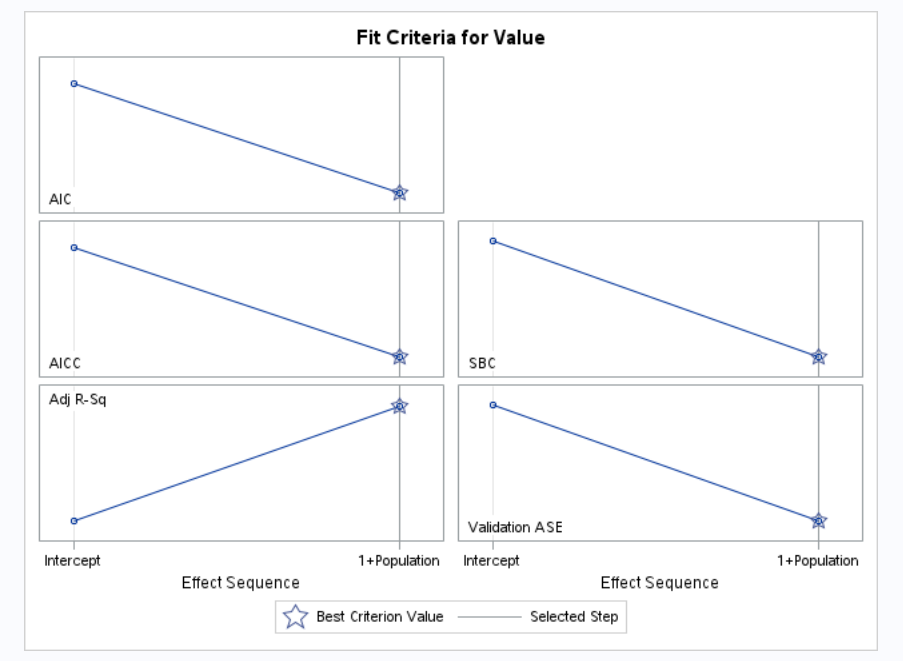
run;

quit;

Figure 2 below shows the results of this regression.

**Figure 2.**

The Average Square Error, aka ASE, is quite large, and is depicted in Figure 1 in scientific notation as a result. Figure 3 on the next page shows the Fit Criteria for the Value variable. This verifies our attempt

**Figure 3.**

to find a connection just between these two variables, the independent population and the dependent housing value. Let’s revisit the hypotheses for this project just a moment.

Ho: The predictive model is not statistically better than chance at predicting housing

values.

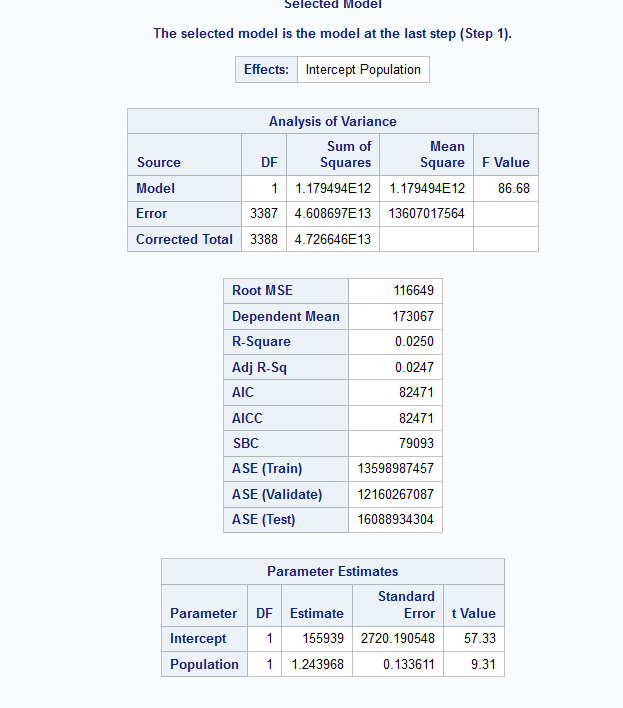
Ha: The predictive model is better than chance at predicting housing values.

We are using a level of significance of alpha = 0.05.

Figure 4 on the following page shows the Analysis of Variance and the model that this process generated. The T Value for the population variable was found to be 9.31 which is less than the critical value of 12.7. Therefore, we must fail to reject the null hypothesis and we conclude that we do not have enough information to say this is statistically significant. So, our model is not statistically better than chance with alpha = 0.05 at predicting housing values. Also, our findings show that Adjusted R-Squared is 0.0247 which indicates that only about 2% of the variance is explained by our model. Variance is a measure of the spread of the data. So, this Adjusted R Squared my speak to the model’s overall accuracy, but it does not mean necessarily that the model does a poor job of predicting the mean in an area. As discussed previously, a limitation when predicting the mean value is that it is possible to have the same mean with vastly different spreads of the data. Regardless, this simplified model was unable to link supply and demand for housing in a very easy way.

The next steps would include expanding the model to account for different demographics demands (families may generally have more persons per household than college towns or senior living areas), or to account for external economic factors such as prevailing interest rates at the time the measurements were taken.

**Figure 4.**



**Conclusion**

Our model shows that we could not predict housing values based on population figures alone. Throughout this project, I have realized that the existing literature on housing values has taken more complex approaches to the housing values issue because the issue is itself complex and oversimplification hurts predictive accuracy. I believe the main reason this experiment did not function as theorized is that this model does not appropriately use population as a proxy for demand for housing. Demand for housing fluctuates greatly with just interest rate changes, all else being equal. This fact demonstrates that population alone does not account for demand. There are external factors to consider.

**Recommendations**

The approach of looking at demand of housing instead of using features of the housing, such as number of bedrooms or proximity to good schools, may benefit from adjusting for demographic or geographic instances. One thing that stood out was that as populations increased in our model, the spread between minimum housing value and maximum housing value increased. This might reflect more “affordable housing projects” in urban centers, as well as outsized demand for high end housing in the same urban areas.

**References**

Amend, P., et al. (2019). IDENTIFYING AND ADDRESSING LOCAL AFFORDABLE HOUSING CHALLENGES IN NORTH CAROLINA. *Carolina Planning Journal*, *44*, 12-23.

E. Simlai, P. (2021). Predicting owner-occupied housing values using machine learning: An empirical investigation of California census tracts data. *Journal of Property Research*, *38*(4), 305-336. <https://doi.org/10.1080/09599916.2021.1890187>

Mu, J., Wu, F., & Zhang, A. (2014). Housing value forecasting based on machine learning methods. *Abstract and Applied Analysis*, *2014*, 1-7. <https://doi.org/10.1155/2014/648047>

*NC OSBM: About OSBM*. (n.d.). NC OSBM. <https://www.osbm.nc.gov/about-osbm>

Swanson, D. A., & Hough, G. C. (2012). An evaluation of persons per household (PPH) estimates generated by the American community survey: A demographic perspective. *Population Research and Policy Review*, *31*(2), 235-266. <https://doi.org/10.1007/s11113-012-9227-8>

US Census Bureau. (2023, August 10). *ZIP code tabulation areas (ZCTAs)*. Census.gov. <https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html>