Tampa Bay Real Estate Valuation Predictor

Final Report

By Michael Ward

# Problem Statement

TB Real Estate Corporation is a real estate investment firm in the Tampa Bay, Florida area. The real estate market in the Tampa Bay area is very active. Single family homes are selling quickly. TB Real Estate Corporation needs to be able to assess the value of homes coming onto the market quickly and accurately so that they can beat the competition in making a competitive offer. They need to be able to evaluate the listing price against the predicted sale price in order to identify properties that may be priced below market value and would make good investments.

The Tampa Bay area is located within Hillsborough County, Florida. The Hillsborough County Property Appraiser (HCPA) provides real estate sales and parcel data on their website. The data includes all real estate parcels in the county including residential, commercial, industrial, farming, etc. The sales data is updated on a weekly basis and contains sales dating back to 1980.

Using the HCPA data from February 11, 2022, I was able to create a tool to predict the sales price of single family residential homes in Tampa Bay. Multiple regression models were evaluated using a combination of features and data. The XGBoost Regressor model proved to be the best predictor of sales price with an average accuracy of 92% and root mean squared error $42,000. Over the course of time, the accuracy of the model may begin to drift as the real estate market evolves. The model will likely need to be retrained on a regular basis.

# Data Wrangling

The HCPA provides new zip files each week on their website which contain the most current sales data, parcel data, supporting files, and explanatory documents. The sales and parcel data files are in database file format (\*.dbf). The parcel file included all 512,206 parcels in Hillsborough County with 47 features to describe each parcel. The sales file contained 2,166,282 property sale records dating back to 1980. Each sale is made up of 17 features.

The sales and the parcel data were merged on the Folio which is a unique identifier for each parcel. There were 0.066% of sales records that did not have a matching Folio ID in the parcel data. The reason for this was not clear and with such a small percentage, these sales records were dropped. There are several features that are in both the parcel and the sales record. In the cases where the data did not match, a logical judgement was made as to which value should be kept. The sales data contains values at the time of the sale while the parcel data is the current value at the time the parcel file was created.

The parcel data includes the Department of Revenue (DOR) code which indicates the type of property. The HCPA data includes a reference XLSX file which contains the valid DOR codes and descriptions. This was used to validate the DOR codes in the parcel data and to filter out all properties that were not residential since the problem statement is to predict sales prices for residential homes. The data was further filtered to only include properties that were post-construction (“Improved”) and where the sales were “Qualified” by the HCPA as verified, free-market sales.

The sales prices were analyzed to identify and research outliers. As a result of the analysis, outliers where the sale price was less than or equal to $1000 were dropped and sales prices greater than $30,000,000 were dropped.

Some features were dropped where they were duplicated between the sales and parcel data and some were dropped as they were deemed not useful for predicting sales prices. Also, some new features were derived from existing features in order to provide additional information regarding the location of the property. The Property Identification Number (PIN) is a combination of Municipality Code, Section, Township, Range, Subdivision, Block, and Lot. These were extracted into individual fields. The HCPA has defined 29 Market Areas of the county and the Market Area Code was extracted from the Neighborhood Code. A new “Region” feature was created to group the Market Areas into 8 geographic regions of the county.

The final dataset contained 799,380 sales records and 42 features.

# Exploratory Data Analysis

The sale price of the property is the target feature to predict. The current sales data includes all residential property sales in Hillsborough County Florida from 1980 through 2022. The sale amounts range from $1100 to over $22mm with a mean of $190k and a median of $148k.

Chart, histogram

Description automatically generated

TB Real Estate Corp is interested in certain property features that are favorable to most residential home buyers. The following categories were analyzed.

* Types of Sales
* Home Features
* Property Locations
* Taxable Values of Properties

## Types of Sales

The data included all types of residential homes as defined by the DOR codes. Most of the property sales are Single Family Residential (code 0100). Condominiums (0400) and Townhouses (0106) are the next highest categories, but they are just a small fraction of the total sales. TB Real Estate Corp is focused on Single Family Residential homes as Condominiums and Townhouses have more investment complications such as association fees and rules restricting sales and rentals. Therefore the remainder of the EDA only included DOR code 0100.

Graphical user interface, text, application, email

Description automatically generatedChart, bar chart

Description automatically generated

Property sales increased sharply from the mid-1990's to the mid-2000's when the U.S. mortgage crisis crashed the housing market. Over the 2010's the annual number of property sales increased to near the levels prior to the crash.

Chart, bar chart, histogram

Description automatically generated

Chart, line chart

Description automatically generated

Just prior to the 2006 market crash and following the recovery post-2011, the greatest growth in new home construction was in the Eastern, East Bay, and Southern regions of the county. The South Tampa region has seen the greatest growth in terms of sales prices.

Chart, scatter chart

Description automatically generated

During analysis, it was discovered that the 2021 sales data is inaccurate in some circumstances. There are sales that were marked as "Qualified" in the data and then later changed to "Unqualified" by the HCPA. This resulted in invalid outliers that negatively impacted the predictability of the model. In order to predict current sales prices / market values, the remainder of the analysis was focused on the sales data for the year of 2020. Although the number of sales each month were relatively consistent, the sales prices increased significantly.

Chart, bar chart

Description automatically generated

Chart, line chart

Description automatically generated

## Home Features

The number of bedrooms and bathrooms are important features for home sales. Most single family residential homes have 3 or 4 bedrooms and 2 bathrooms. TB Real Estate Corp is focused on investing in residential homes with 3-5 bedrooms and 1-4 bathrooms. Anything less or more would not be appealing to most home buyers.

Chart, bar chart, histogram

Description automatically generated

Most homes have just 1 building structure, 1 unit and 1 or 2 stories. However, there are a few outliers that have more. TB Real Estate is only interested in properties that have a single building/unit and 1-2 stories.

Chart, bar chart

Description automatically generated

The heated area is the square footage of the home that is heated. This is equivalent to the size of the home and is an important factor in home sales. Most single family residential homes are between 1500 and 2500 square feet.

Chart, histogram

Description automatically generated

The acreage is the size of the land and is an important factor in home sales. Most properties are between 0.14 and 0.24 acres.

Chart, histogram

Description automatically generated

The age and condition of the home can have a significant impact on the sales price. Old homes often require more investment in maintenance and repairs than newer homes. The mean home age is 22 years old and the median value is 16.

Chart, histogram

Description automatically generated

As the age of the home goes beyond 20 years, the condition of the homes and sales prices increasingly vary. In order to mitigate the risk of purchasing investments that require significant costs of repair and maintenance, TB Real Estate will restrict their investments to home that are less than 20 years old.

Chart

Description automatically generated

## Property Locations

Location, Location, Location!!

Location is a very important factor when evaluating market value of homes. The following is a map of Hillsborough County which is provided in the download files from the HCPA website.

Map

Description automatically generated

There are multiple ways of describing the property location based on the data.

* Postal Address includes the street address, city, and zip code.
* Property Identification Number (PIN) is defined by the County Property Appraiser to uniquely identify each property based on the above grid. It is made up of the following data:
  + Municipality Code (Tampa, Temple Terrace, Plant City, Unincorporated)
  + Township Code - Values 27-32 indicating the latitudinal rows on the grid map.
  + Range Code - Values 17-22 indicating the longitudinal columns on the grid map.
  + Section Code - Values 1-36 represented by the 6x6 sections of red squares within each Township/Range on the grid map.
  + Land Type ID / Subdivision Code - Identifies platted subdivisions (i.e. planned communities) which tend to have higher values due to the amenities provided.
  + Block Number - 6 digit code for the Block Number in platted Subdivisions. For unplatted land, it has a uniquely assigned number.
  + Lot Number - 6 digit code for the Lot Numbers in Platted Subdivisions. For all unplatted land, it has a uniquely assigned number.
* Market Area is defined by the HCPA for assessing values for tax purposes. Each Market Area contains multiple Neighborhoods. Each of the 29 Market Areas are represented by the bold, black division lines and large purple numbers in the grid map.
* Neighborhoods are represented by the colored areas in the grid map and have a unique 6 digit identifier.

Market Areas 02 and 03 (South Tampa West and East) had the highest average sales prices. This makes sense as these homes are in the heart of the city of Tampa and have water on both sides.

Market Areas 26 and 27 (Riverview and Southbay) had the most property sales over the year and the median sales prices that are slightly below the median for the county.

Chart, scatter chart

Description automatically generated

The city of Tampa (01-06) has the oldest homes and highest sales prices. The southeastern region (26-29) has the newest homes and the lowest sales prices.

Chart, scatter chart

Description automatically generated

These Southeastern Market Areas are attractive for investments due to the following features:

1. Most home sales over the past year
2. Lowest median age of homes sold
3. Median sales prices below the county median sales price

Of the top-10 Neighborhoods for number of sales, 7 of them are in Market Areas 26 and 27. The other 3 are in Market Areas 25, 28, and 29. This indicates that the southern portion of the county is the most active market for home sales.

Chart, funnel chart

Description automatically generated

The top market area is Riverview (“26”). However, Market Areas are quite large and can include thousands of home sales per year. Riverview is made up of 15 Neighborhoods with the top Neighborhoods having over 800 sales in 2020.

Chart, bar chart, histogram

Description automatically generated

Hillsborough County has over 11,000 defined subdivisions. Subdivisions are platted land also known as planned communities. Subdivisions are developed by a builder and the homes are typically similar as the builder will offer a few models for buyers to choose from. Therefore, it is logical that Subdivision would be a good granularity when predicting sale prices. The following map displays all of the Subdivisions in the county as the colored areas. The non-colored areas are unplatted and do not have a Subdivision Code.

Map

Description automatically generated

One of the most popular neighborhoods in the Riverview Market Area (26) is the Balm-Boyette Area Neighborhood (226001.0). This neighborhood has 29 subdivisions and also unplatted homes indicated by Subdivision Code ‘ZZZ’. The “BELMONT SOUTH PHASE 2D AND PASEO AL MAR BOULEVARD” (B88) Subdivision had the most sales in 2020 with over 150 with an average sales price just under $250,000.

Chart, bar chart, histogram

Description automatically generated

The following figure shows that the sales prices for the B88 Subdivision (BELMONT SOUTH PHASE 2D AND PASEO AL MAR BOULEVARD) are fairly consistent across all of the sales in 2020 for the subdivision. This indicates that the subdivision should be granular enough to be a good predictor of sales price.

Chart, histogram

Description automatically generated

## Taxable Values of Properties

The County Property Appraiser determines property values for tax purposes. These values are reassessed each year and is based on the following factors:

* The present cash value.
* The highest and best use of the property.
* The location of the property.
* The quantity or size of said property.
* The cost of said property.
* The condition of said property.
* The income of said property.
* The net proceeds of the sale of said property. Under Florida law, a residence must be reassessed at market value when it sells. https://www.hcpafl.org/Property-Info/Truth-In-Millage

Just Value is the just/market value of the parcel as determined by the County Property Appraiser. The following tax-related values are subcomponents or variations of the Just Value.

* Land Value - market value of the land for tax purposes
* Building Value - market value of all of the buildings on the property for tax purposes
* Extra Features Value - market value of all of the extra features on the property for tax purposes (i.e. fences, dock, pool, utility shed)
* Assessed Value - value of land & improvements, less any Save Our Homes 3% cap or non-homestead 10% cap
* Taxable Value – Assessed Value less any exemptions, such as homestead, widow, disability

Chart, histogram

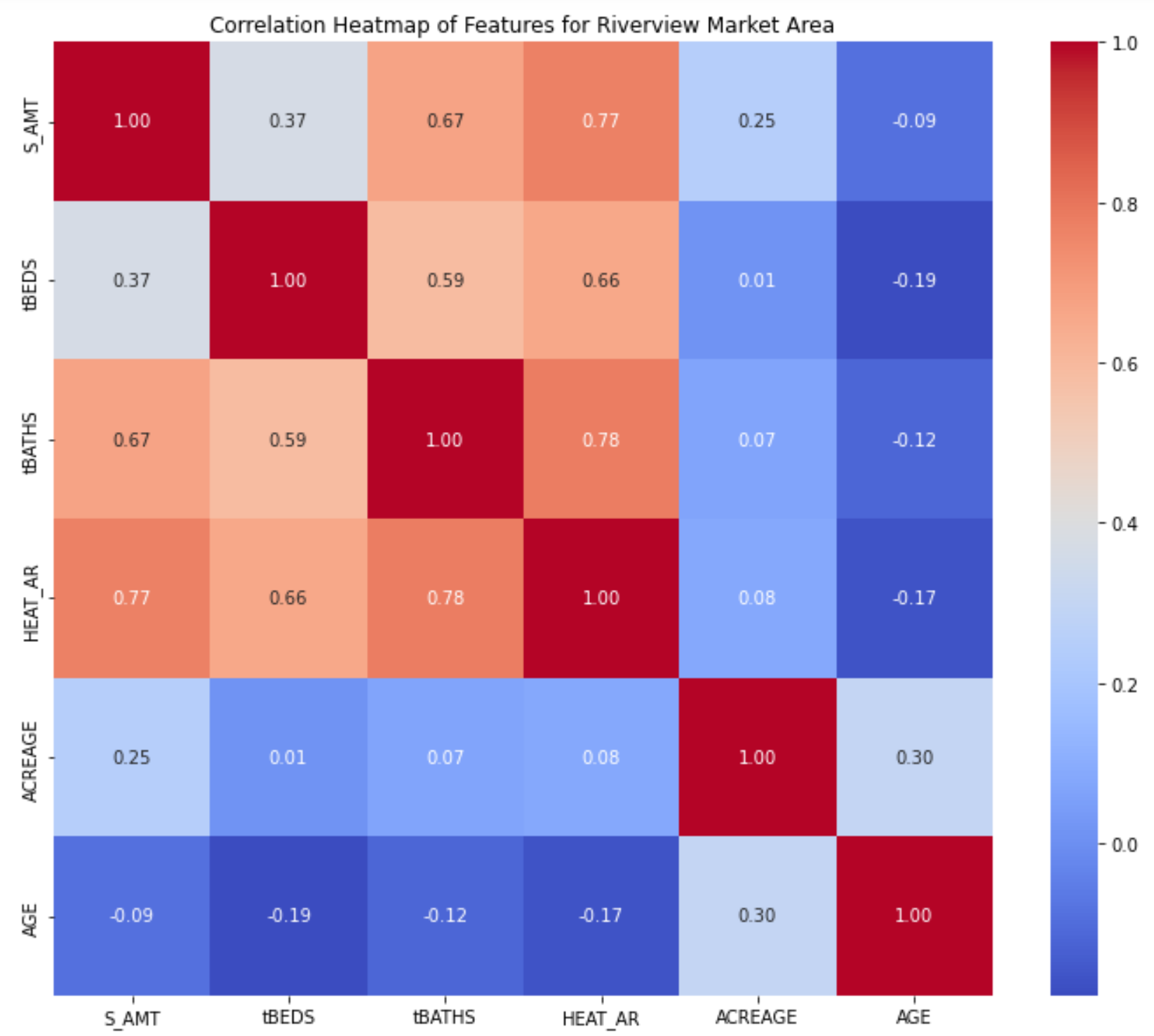
Description automatically generated

While the Just Value and other tax-related values are strongly correlated with Sales Prices, they should not be used in the model. These values are provided in the Parcel data which is a current snapshot and is not information that would have been available at the time of the sale. Therefore, it is a “look-forward” feature that allows leakage into the predictive model.

## Conclusion of Exploratory Data Analysis

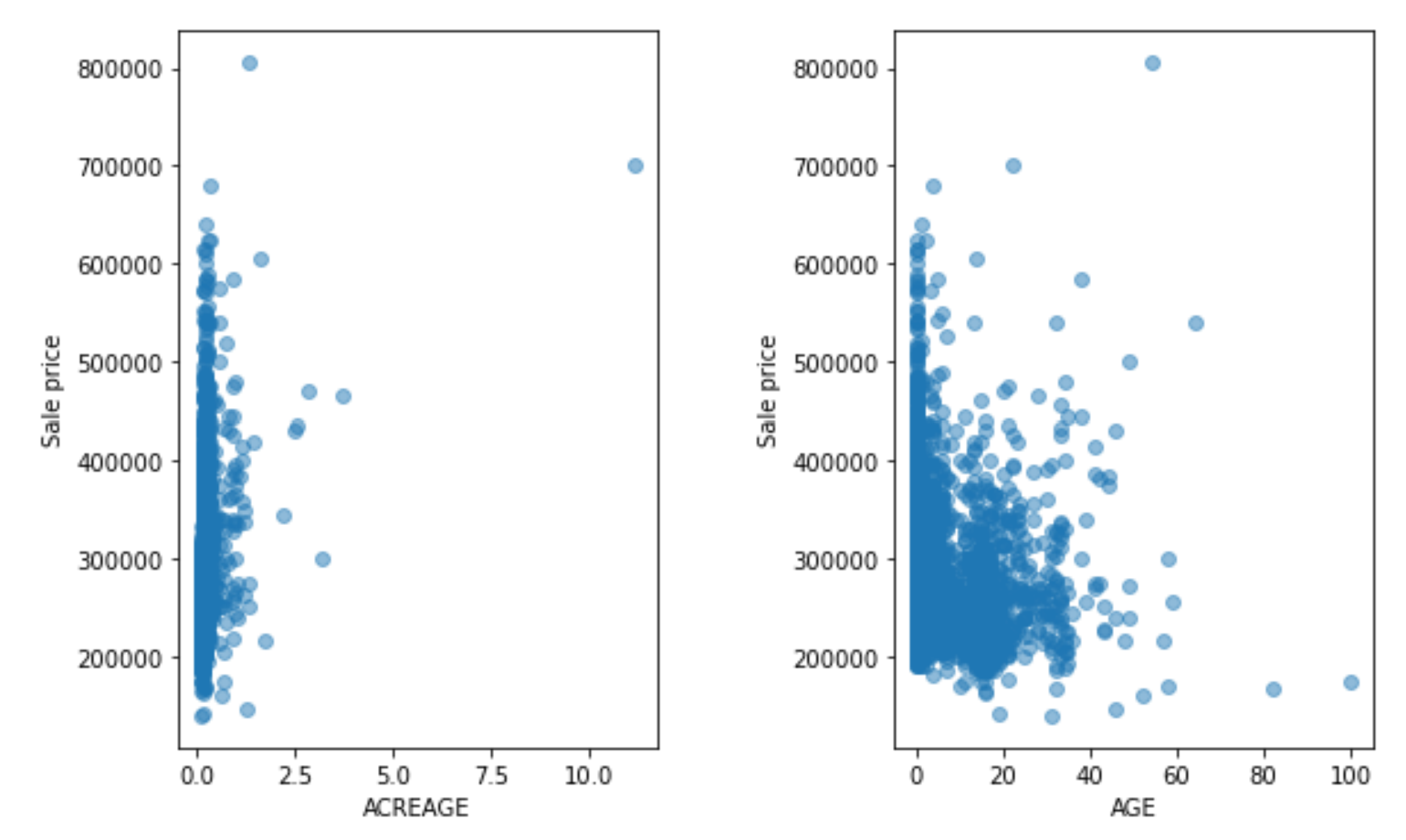
Based on the Exploratory Data Analysis, it appears that the best investment opportunities in the Tampa Bay area are in the Riverview Market Area (26).

There is a strong correlation of the Heated Area and number of Bathrooms to the Sales Price. This does not factor in location factors of Neighborhood and Subdivision as these are categorical features.

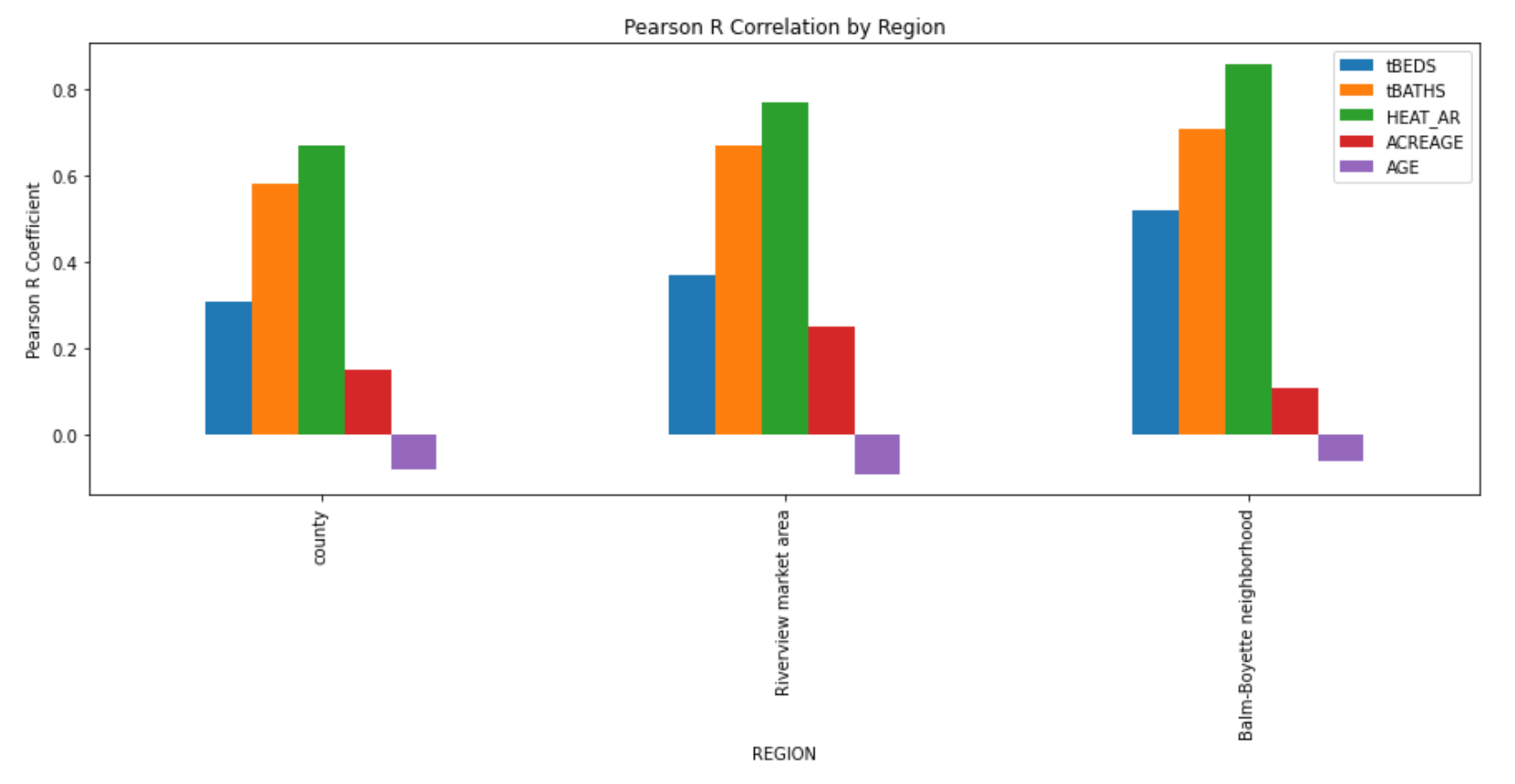


The following figures show the strong linear relationship between between the sales price and the heated area and number of bathrooms. The number of bedrooms shows a spike in sales price for 4 bedroom homes. Most homes have a similar number of acres





The Pearson R Correlations improve as we target the Riverview Market Area versus the entire county. As we compare the individual Balm-Boyette Area Neighborhood, the correlations of the number of bedrooms/bathrooms and heated area get stronger, but the acreage and age of the home are less correlated to the sales price.



# Feature Engineering

Based on the Exploratory Data Analysis, I determined that the following features would be good candidates for modeling.

* Types of Sales
  + Single Family Residential Homes (DOR Code 0100)
  + Qualified (verified as free-market sales)
  + Improved (post-construction sales – not vacant land)
  + Sales from the 2020 calendar year
* Home Features
  + Bedrooms (3 to 5)
  + Bathrooms (1 to 4)
  + Buildings / Units (1 only)
  + Stories (1 to 2 only)
  + Age of Home (less than 20 years old)
  + Heated Area
  + Acres of Land
* Property Locations
  + Market Area (26 – Riverview)
  + Neighborhood Code
  + Subdivision Code
* Taxable Values of Properties
  + None (invalid predictors)

As this problem required a regression model in order to predict the continuous variable of Sales Price, I needed to ensure that all features were numeric and I wanted to reduce the dimensionality of the features as much as possible.

I started by filtering the data by the Types of Sales so that I only had Qualified sales of Improved properties from the calendar year of 2020 for Single Family Residential Homes (DOR Code 0100).

I then filtered the Home Features to only 3-5 bedroom homes with 1-4 bathrooms and 1-2 stories that only had 1 building/unit and were less than 20 years old.

I also filtered the Property Locations to only include my target Market Area of Riverview (26).

# Model Selection

This project required a regression model as the target feature of Sales Price is a continuous variable. The statistician, George Box, was quoted as saying “Essentially all models are wrong, but some are useful.” I tested five different models in order to determine which was most useful for this problem. The five models that I tested were:

* K Nearest Neighbors (KNeighborsRegressor)
* Decision Tree (DecisionTreeRegressor)
* Random Forest (RandomForestRegressor)
* Linear Regression (LinearRegression)
* Extreme Gradient Boosting (XGBRegressor)

The metric that I focused on while training and testing the models was Root Mean Squared Error (RMSE). I chose this because I wanted to minimize the error in predicting the Sales Price and I wanted to use a metric in the same scale as the Sales Price.