Text

Description automatically generated**Housing Data Project Part 3**

ISYS 5843

Data Understanding

Stratton Carroll, Kelsey Arnold, Zach Denton

Group 02

Contents

[Description and Goal of The Housing Data 1](#_Toc99485314)

[Data Understanding – Initial Insights 1](#_Toc99485315)

[Initial Models 3](#_Toc99485316)

[Linear regression 3](#_Toc99485317)

[Data Preparation 3](#_Toc99485318)

[Modeling 3](#_Toc99485319)

[Decision Tree 3](#_Toc99485320)

[KNN 3](#_Toc99485321)

[Data Preparation 3](#_Toc99485322)

[Modeling 4](#_Toc99485323)

[Gradient Boosting – Best Model 4](#_Toc99485324)

[Appendix 4](#_Toc99485325)

[Data Dictionary 4](#_Toc99485326)

[Datatypes 4](#_Toc99485327)

[Visualizations 4](#_Toc99485328)

[Modeling 13](#_Toc99485329)

# Description and Goal of The Housing Data

The Housing Dataset contains 77 attributes of information pertaining to the houses purchased in Benton County dating back to the 1900’s. The dataset is rich in information including the city’s the houses are located, the date the houses were sold, the price the houses sold for, and even the different types of materials that were used to build the house, just to name a few. The objective of this project is to determine what attributes are the best predictors to determine sale price. To do this we will first start by looking into the housing dataset and performing descriptive statistics to better understand the data.

# Data Understanding – Initial Insights

When looking at the average sales price of homes in the Benton County we can see that the price varies dramatically depending on what town you are looking to buy. The average home price in Benton County is $495,444 with the average difference in price from the cheapest (Sulphur Spring) to the most expensive (Bentonville) being roughly $685,111. See Table 1 and Graph 1 for the average sale price for every town in Benton County along with the minimum and maximum sale price.

Looking at the sale price over the years we can see that houses in the Benton County area have change drastically. The average sale price back in 1961 was $17,500 and grew to as high as $1.4 million in 2006 right before the housing crash of 2008. The average sale price as of 2020 is $371,584. See Graph 2 for the average sale price trends from 1900 to 2021.

We can also see the trends of the number of homes sold over the past decades as they follow economic trends. The houses sold gradually increases all the way up to 2006, right before the 2008 housing crash. The Homes sold then again gradually trend upward until 2020 Covid-19 pandemic. See Graph 3 for the number of homes sold from 1987-2021

Looking at total living area, we can get an idea where the largest houses on average are in Benton County. Creating a calculated field adding both first floor and second floor living area we can see Elm Springs has the largest average living area followed by Bentonville and Cave Springs. The top 3 cities with the largest living area are also above the county’s average in total sale price. See Table 2 for the average largest living area.

When comparing Table 3 and 4, with the cost of land you can see that the majority of the houses sold have the highest land values in the year 2020. While most of the lowest land values occurred in 2016. 2016 also happened to be a year where the market had lower numbers for rates and prices. This could be a connection for the lower prices for land as well. As we have seen, prices have only increased since 2016 therefore land value has also.

This table is looking at the school districts that the houses are being sold in. It is listed from smallest to largest, but the data set also had it by district and the city which the house is located in. Bentonville City has the highest number of houses for sales in their school district within our data. Bentonville is also where the headquarters of Walmart is located so this may play a role in the number of houses being built in the area as well as sold to accommodate the employees of such a large company.

Over the past decade there have been noteworthy trends on the number of homes being sold in regard to their construction type. Std frame homes and masonry homes have seen sharp increases in numbers sold until 2019. Masonry homes have seen a steep drop off in 2020, and Std frame homes have seen slight drop off in 2020 (Graph 4).

When looking at school districts we want to know what school districts have the highest and lowest average home sale prices. School districts Bentonville City, Bentonville (Rogers City), and Springdale (Elm Springs) have the highest average home price. While Pea Ridge (Garfield), Gravette (Sulphur Springs), and Springdale (Lowell) have the lowest average home price (see Graph 5).

It is important to know the prices of types of homes and where they are selling. In the last 11 years we can see that mobile homes have the highest average price in Centerton, multifamily homes are highest in Bentonville, and single-family homes are highest in Elm Springs (Figure). With this being said we can start to profile where homes are valued the most.

Homes are seeing a rise in average acres. While average acres for a home peaked at 2.172 acres in 2012, 2013 saw a sharp decline. Since then homes have grown in acre size from 0.633 (2013) to 1.881 (2021). This tells us that home buys want homes with more land (see Graph 6).

While going through the variables we found that there are only two binary variables across the entire dataset as well as only one ID. When it comes to dates, there are very few and more of the numeric values deal with the square feet of rooms or supplies used in the homes. Some of the variables are simply left blank or have a “0” in for the value if it does not have that specific feature or supply.

# Modeling

## Initial Model

## Gradient Boosting – Best Model

Two Gradient Boosting models were conducted with the target variable sale price.

### Data Preparation

The data was transformed in several ways to reach a more useable data set for the Gradient Boosting Model. The nodes used in the model flow were drop column, transform variables, transform, impute.

### Modeling

Two Gradient Boosting models were created, N iterations of 50/Shrinkage 0.1, N iterations of 200/Shrinkage 0.2. The N iterations of 50 and Shrinkage of 0.2 gave us the lowest RASE of 328544.4 See Graph 14.

|  |  |
| --- | --- |
| **Model** | **RASE** |
| N iterations of 50/Shrinkage 0.1 | 328544.4 |
| N iterations of 200/Shrinkage 0.2 | 349711.2 |

## Final Models

## HP Neural Network and HP Forest

Several Hp neural networks and HP forests were created with the target variable sales price ([Graph 14](#_Modeling)).

#### Data Preparation

The data was prepared for the HP neural networks in several ways. The transformations are as follows:

1. Drop column node was placed to drop the grade column.
2. Metadata node was placed.
3. HP partition set to train 60, validate 40, changed “City” to stratification.
4. Interactive binning node was placed.
5. Variable transform node was placed to range interval inputs.
6. Another variable transform node was placed to optimize binning of interval inputs and target inputs.
7. HP impute node was placed.
8. Control point was placed.

#### Model Evaluation

Several flows of models with various components were compared. They are:

* HP Principal Components → HP Forest
* HP Linear Stepwise Regression → HP Forest
* HP Principal Components → HP Neural Network
* HP Linear Stepwise Regression → HP Neural Network
* HP Neural Network

The top performing models were the HP neural network and the linear stepwise regression with HP neural network. The Hp neural network slightly outperforming the HP regression with HP neural network with a root average squared error (RASE) of 509277.3 ([Table 15](#_Modeling)). The final model set out outperformed the initial model set, as seen in the table below.

|  |  |
| --- | --- |
| **Model** | **RASE** |
| Forward/Stepwise | 784689 |
| Backwards | 785456.5 |

|  |  |
| --- | --- |
| **Model** | **RASE** |
| HP Neural Network | 509277.3 |
| HP Linear Stepwise Regression → HP Neural Network | 596249.6 |

## Gradient Boosting

The Gradient Boosting node was used to create our best model with the target variable sales price ([Graph 15](#_Modeling)).

#### Data Preparation

The data was prepared for the Gradient Boosting in several ways. The transformations are as follows:

1. Metadata node was placed.
2. HP partition set to train 60, validate 40, changed “City” to stratification.
3. Variable transform node was placed to range interval inputs.
4. Another variable transform node was placed to optimize binning of interval inputs and target inputs
5. Gradient Boosting was placed with N-Iterations set to 300, Train Proportion set to 60, Max Branch set to 5, and Max Depth set to 5.

#### Model Evaluation

Several flows of models with various components were compared using different transforming techniques like Optimal Binning, Best, Quantile Binning. The overall best RASE is listed below:

|  |  |
| --- | --- |
| **Model** | **RASE** |
| Gradient Boost | 274846.1 |

## Decision Trees

#### Data Preparation

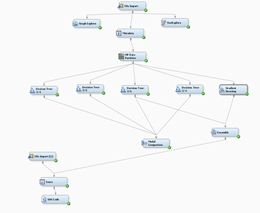
1. First the data was imported and then rejected all the variables that had “obyi” in them
2. Metadata node was used
3. Then put in a HP Data Partition with a 60/40
4. Then created Decision Trees, one ensemble with the best decision tree that included a gradient boosting node

#### Model Evaluation

* Decision Tree with 3 branches and depth 4
* Decision Tree with 3 branches and depth 6
* Decision Tree with 2 branches and depth 4
* Ensemble with
  + Decision Tree with 2 branches and depth 6
  + Gradient Boosting with default

The best model was the ensemble with the best decision tree connected to the gradient boosting node. This model had an RASE of 365362.3 on SAS E Miner and then 445302.2 on Kaggle. The Decision Tree alone, which was the best initial model was had an RASE of 383119.5 on SAS E Miner and scored a 460730.8 on Kaggle.

**Final Model Diagram:**



# Deployment

# Appendix

## Data Dictionary

|  |  |
| --- | --- |
| assessment\_year | Year of last assessment |
| basement\_finished\_w\_partitions | Basement area finished without partitions (sq ft) |
| basement\_finished\_wo\_partitions | Basement area finished with partitions (sq ft) |
| basement\_total\_sf | Basement area (sq ft) |
| basement\_unfinished | Basement area unfinished (sq ft) |
| baths\_full | Count of full bathrooms |
| baths\_half | Count of half bathrooms |
| City | City where the property is located |
| Construction Type | Construction type describes what material the building is made of. This does not include the foundation or finishing materials, but rather the structural components of the rest of the home. |
| County Name | County where the property is located |
| deed\_type | Type of deed |
| Effective Age | Age of property based on its current condition, rather than its actual age - assigned by appraisers |
| Fireplace | Description of fireplaces (if available) |
| Floor Type | Types of Flooring Materials |
| floor\_carpet | Carpet covering (sq ft) |
| floor\_ceramic | Ceramic covering (sq ft) |
| floor\_linoleum | Linoleum covering (sq ft) |
| floor\_none | None covering (sq ft) |
| floor\_parquet | Parquet covering (sq ft) |
| floor\_softwood | Softwood covering (sq ft) |
| floor\_stone | Stone covering (sq ft) |
| Foundation Type | Describes what type of foundation the property's building has been erected on |
| Grade | Building Grade means the type, brand and/or quality of materials Landlord designates from time to time to be the minimum quality to be used in the Building or the exclusive type, grade or quality of material to be used in the Building. |
| hardwood\_sheath/floor\_hardwood | Hardwood covering (sq ft) |
| heat\_ac | Type of HVAC system (heating, ventilation, and an air-conditioning unit). |
| id | a randomly generated unique identifier for the property |
| is\_homestead | all separately assessed real property parcels or parts thereof used as one-, two-, or three-family residential dwellings, including so much of the abutting land as is reasonably necessary for residential purposes |
| is\_over\_65 | Is the actual age of the property greater than 65 years old ? |
| land\_value | Ranges from Low, Typical, High, Vhigh - describes the value of the surrounding land, lot size, or similar desirable features |
| living\_area\_1st\_floor | First floor living area (sq ft) |
| living\_area\_2nd\_floor | Second floor living area (sq ft) |
| lot\_block |  |
| ne\_corner | longitude and latitude of north east corner of property |
| obyi\_adw | Asphalt Paving (sq ft) |
| obyi\_agdw | Aggregate Concrete Drive (sq ft) |
| obyi\_agps | Aggregate Concrete Patio (sq ft) |
| obyi\_agsw | Aggregate Concrete Walkway (sq ft) |
| obyi\_bsmp | BRICK/STONE MORTER PAVEMENT (Sq Ft) |
| obyi\_cbwx6 | Concrete Block 6" (Sq Ft) |
| obyi\_cbwx8 | Concrete Block 8" (Sq Ft) |
| obyi\_cdw | CONCRETE DRIVEWAY (Sq Ft) |
| obyi\_cp | Carport (Sq Ft) |
| obyi\_cpsf | Carport Storage Frame (Sq Ft) |
| obyi\_cpsm | Carport Storage Masonry (Sq Ft) |
| obyi\_csl | Light Concrete (Sq Ft) |
| obyi\_ffa | GARAGE - FRAME FIN ATTACHED |
| obyi\_ffb | GARAGE - FRAME FIN BUILT-IN |
| obyi\_fua | GARAGE - FRAME UNFIN |
| obyi\_gep | GLASS ENCLOSED PORCH (sq ft) |
| obyi\_mfa | GARAGE - MAS FIN ATTACHED |
| obyi\_mfb | GARAGE - MAS FIN BUILT-IN |
| obyi\_mua | GARAGE - MAS UNFIN ATTACHED |
| obyi\_mub | GARAGE - MASNRY UNFN BUILT IN |
| obyi\_mw | WALLS BRICK OR STONE (sq ft) |
| obyi\_op | OPEN PORCH (sq ft) |
| obyi\_op2 | 1/2 OPEN PORCH (sq ft) |
| obyi\_pca | PATIO COVER, ALUMINUM (sq ft) |
| obyi\_pcb | PATIO COVER, BUILTUP |
| obyi\_pcf | PATIO COVER, FIBERGLASS |
| obyi\_pcs | PATIO COVER, STEEL |
| obyi\_ps | PATIO SLAB |
| obyi\_sep | SCREEN ENCLOSED PORCH (SQ FT) |
| obyi\_wc | BRICK/STONE WAINSCOT (sq ft) |
| obyi\_wd | Wood decking (sq ft) |
| Occupancy Type | Building occupancy classifications refer to categorizing structures based on their usage and are primarily used for building and fire code enforcement. |
| parcel\_num | Unique Identifier for the property |
| Roof Type | A roof type (or roofing type) is the covering or material on the roof of a house, condo or building. The type of roof is usually stated on most real estate listings |
| SalePrice | Sales price of the property the last time it was sold |
| School District | A school district is an area which includes all the schools that are situated within that area and are governed by a particular authority. |
| Sec-Twp-Rng | Identifier of the "Section Township and Range" of the property |
| sold\_date | Date house was last sold |
| Story Height | Number of floors in the structure |
| Subdivision | City neighborhood |
| sw\_corner | longitude and latitude of south west corner of property |
| tax\_status | If property is taxable or not |
| Total Acres | Total Land Area (in acres) |
| Year Built | Year the property was erected |

## Datatypes

|  |  |
| --- | --- |
| assessment\_year | Nominal |
| basement\_finished\_w\_partitions | Interval |
| basement\_finished\_wo\_partitions | Interval |
| basement\_total\_sf | Interval |
| basement\_unfinished | Interval |
| baths\_full | Interval |
| baths\_half | Interval |
| City | Nominal |
| Construction Type | Nominal |
| County Name | Nominal |
| deed\_type | Nominal |
| Effective Age | Interval |
| Fireplace | Nominal |
| Floor Type | Nominal |
| floor\_carpet | Interval |
| floor\_ceramic | Interval |
| floor\_linoleum | Interval |
| floor\_none | Interval |
| floor\_parquet | Interval |
| floor\_softwood | Interval |
| floor\_stone | Interval |
| Foundation Type | Nominal |
| Grade | Nominal |
| hardwood\_sheath/floor\_hardwood | Integer |
| heat\_ac | Nominal |
| id | ID |
| is\_homestead | Binary |
| is\_over\_65 | Binary |
| land\_value | Interval |
| living\_area\_1st\_floor | Interval |
| living\_area\_2nd\_floor | Interval |
| lot\_block | Nominal |
| ne\_corner | Nominal |
| obyi\_adw | Interval |
| obyi\_agdw | Interval |
| obyi\_agps | Interval |
| obyi\_agsw | Interval |
| obyi\_bsmp | Interval |
| obyi\_cbwx6 | Interval |
| obyi\_cbwx8 | Interval |
| obyi\_cdw | Interval |
| obyi\_cp | Interval |
| obyi\_cpsf | Interval |
| obyi\_cpsm | Interval |
| obyi\_csl | Interval |
| obyi\_ffa | Interval |
| obyi\_ffb | Interval |
| obyi\_fua | Interval |
| obyi\_gep | Interval |
| obyi\_mfa | Interval |
| obyi\_mfb | Interval |
| obyi\_mua | Interval |
| obyi\_mub | Interval |
| obyi\_mw | Interval |
| obyi\_op | Interval |
| obyi\_op2 | Interval |
| obyi\_pca | Interval |
| obyi\_pcb | Interval |
| obyi\_pcf | Interval |
| obyi\_pcs | Interval |
| obyi\_ps | Interval |
| obyi\_sep | Interval |
| obyi\_wc | Interval |
| obyi\_wd | Interval |
| Occupancy Type | Nominal |
| parcel\_num | Interval |
| Roof Type | Nominal |
| SalePrice | Interval |
| School District | Nominal |
| Sec-Twp-Rng | Nominal |
| sold\_date | Nominal, Date |
| Story Height | Nominal |
| Subdivision | Nominal |
| sw\_corner | Decimal |
| tax\_status | Nominal |
| Total Acres | Interval |
| Year Built | Nominal |

## Visualizations

Table 1. Average sale price of homes between cities.

**Text

Description automatically generated with low confidence**

Table 2. Average Living Area by City (highlighted red indicates above average sale price)

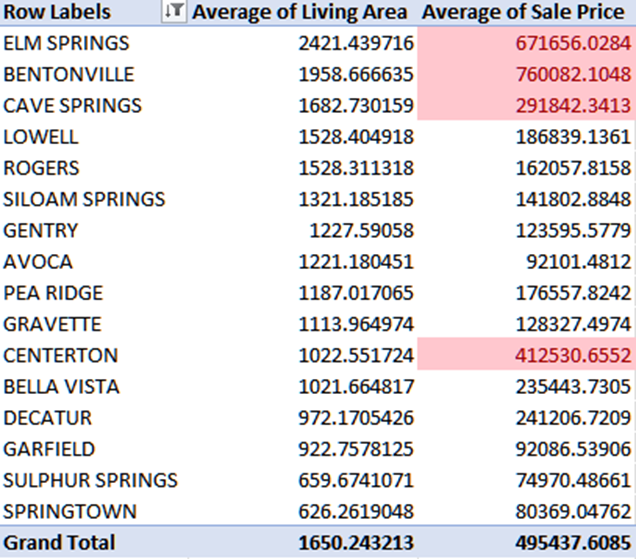


Table 3: Top 10 Highest Prices for Land

|  |  |  |
| --- | --- | --- |
| **City** | **Assessment Year** | **Land Value** |
| BENTONVILLE | 2020 | $1,195,700 |
| ROGERS | 2018 | $777,500 |
| BENTONVILLE | 2016 | $627,250 |
| BENTONVILLE | 2016 | $574,350 |
| ROGERS | 2016 | $493,500 |
| ROGERS | 2020 | $473,200 |
| BENTONVILLE | 2020 | $446,900 |
| ROGERS | 2020 | $444,500 |
| BENTONVILLE | 2020 | $441,050 |
| BENTONVILLE | 2017 | $435,150 |

Table 4: Top 10 Lowest Prices for Land

|  |  |  |
| --- | --- | --- |
| **City** | **Assessment Year** | **Land Value** |
| ROGERS | 2016 | 0 |
| ROGERS | 2016 | 0 |
| BENTONVILLE | 2016 | 0 |
| GENTRY | 2020 | 0 |
| BENTONVILLE | 2016 | 0 |
| LOWELL | 2017 | 0 |
| ROGERS | 2018 | 0 |
| SILOAM SPRINGS | 2016 | 0 |
| SILOAM SPRINGS | 2016 | 0 |
| BENTONVILLE | 2016 | 0 |

Table 5: Total Number of Houses Sold in Each School District

|  |  |
| --- | --- |
| **School Districts** | **Total Number** |
| 19 Gentry | 1 |
| 30 Rogers | 1 |
| CL50 SPRINGDALE (LOWELL CITY) | 1 |
| CPR30 Rogers (Pea Ridge City) | 1 |
| CBV109 Pea Ridge (Bella Vista City) | 7 |
| CG6 BENTONVILLE (GRAVETTE CITY) | 13 |
| CB30 Rogers (Bentonville City) | 15 |
| CGF109 Pea Ridge (Garfield City) | 15 |
| CCS30 ROGERS (CAVE SPGS CITY) | 19 |
| CS19 Gentry (Springtown City) | 42 |
| CR6 Bentonville (Rogers City) | 46 |
| CS6 BENTONVILLE (CAVE SPGS CITY | 107 |
| CGF30 Rogers (Garfield City) | 113 |
| C17 Decatur | 129 |
| CA30 Rogers (Avoca City) | 133 |
| CES50 Springdale (Elm Springs City) | 141 |
| CC6 Bentonville (Centerton City) | 145 |
| CS20 Gravette (Sulphur Springs City) | 224 |
| C21 Siloam Springs City | 486 |
| C19 Gentry City | 551 |
| C20 Gravette City | 558 |
| CBV20 Gravette (Bella Vista City) | 562 |
| C109 Pea Ridge City | 585 |
| CL30 Rogers (Lowell City) | 609 |
| CBV6 Bentonville (Bella Vista City) | 1,594 |
| C30 Rogers City | 2,968 |
| C6 Bentonville City | 10,532 |

Graph 1. Average sale price of homes between cities.

Chart, bar chart

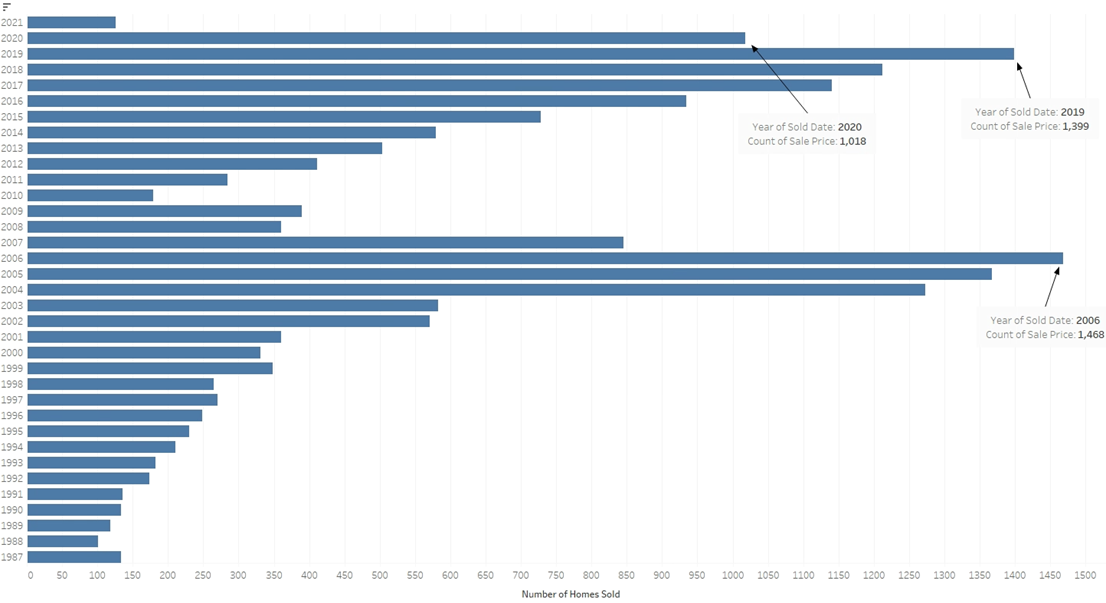
Description automatically generated

Graph 2. Average sale price trends from 1900-2021

Chart, histogram

Description automatically generated

Graph 3. Number of Homes Sold from 1987-2021



Graph 4.

Chart, line chart

Description automatically generated

Graph 5

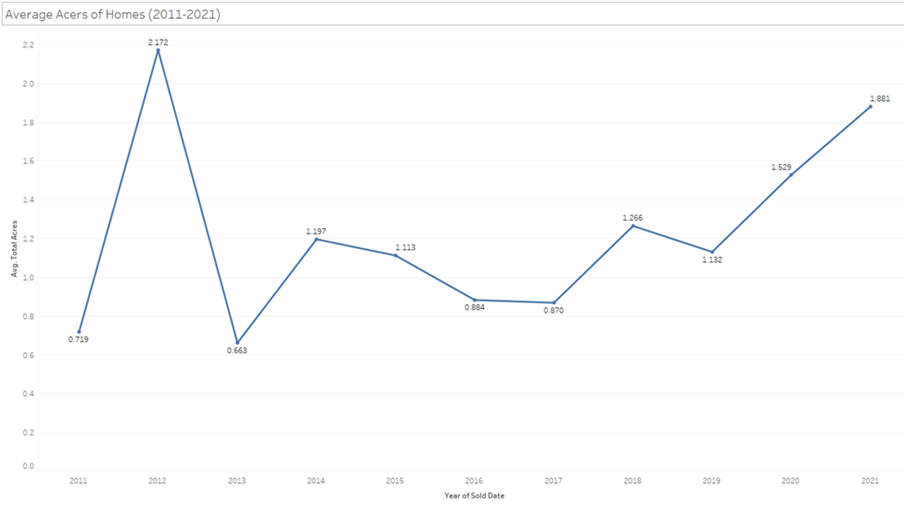
Chart, bar chart

Description automatically generated

Chart, bar chart, histogram

Description automatically generatedGraph 6.

Graph 7.



These are the correlations with the predicted variable sale price. However, these correlation variables have a lot of missing variables.

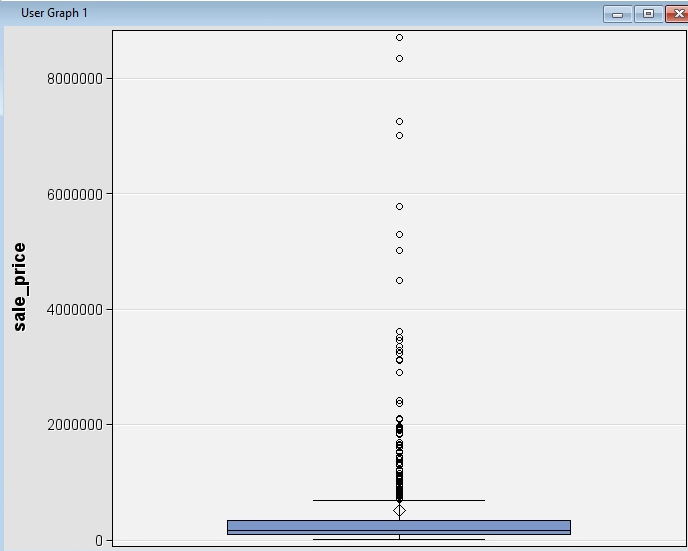
Chart

Description automatically generated

Table

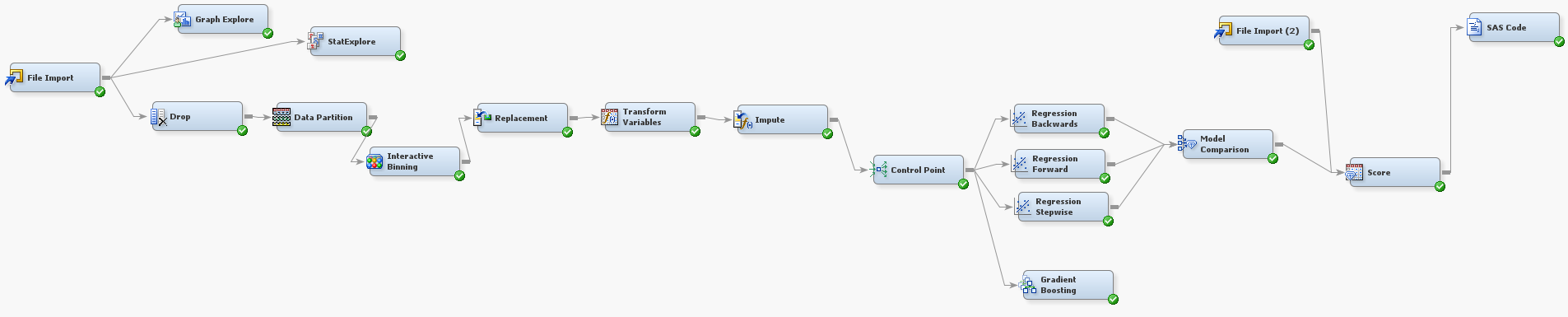
Description automatically generated

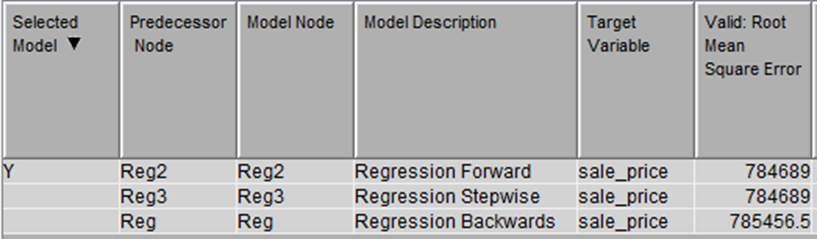
Looking at the missing variables there are a lot of attributes that have more missing values than non-missing values.



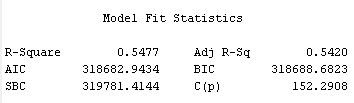
There are a lot of outliers when looking at sale price. Between the first and third quartile the range of sale price is $83000 and $330500.

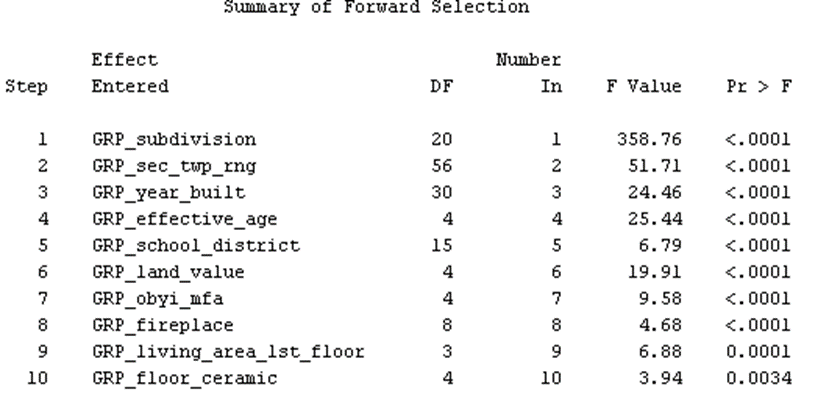
## Modeling

Graph 10: Linear Regression Pipeline

Graph 11: Linear Regression Model Comparison Output

Graph 12: Forward Linear Regression Output





Graph 13: KNN Model Diagram

A picture containing graphical user interface

Description automatically generated

Graph 14: HP Neural Network and HP Forests Model Flow

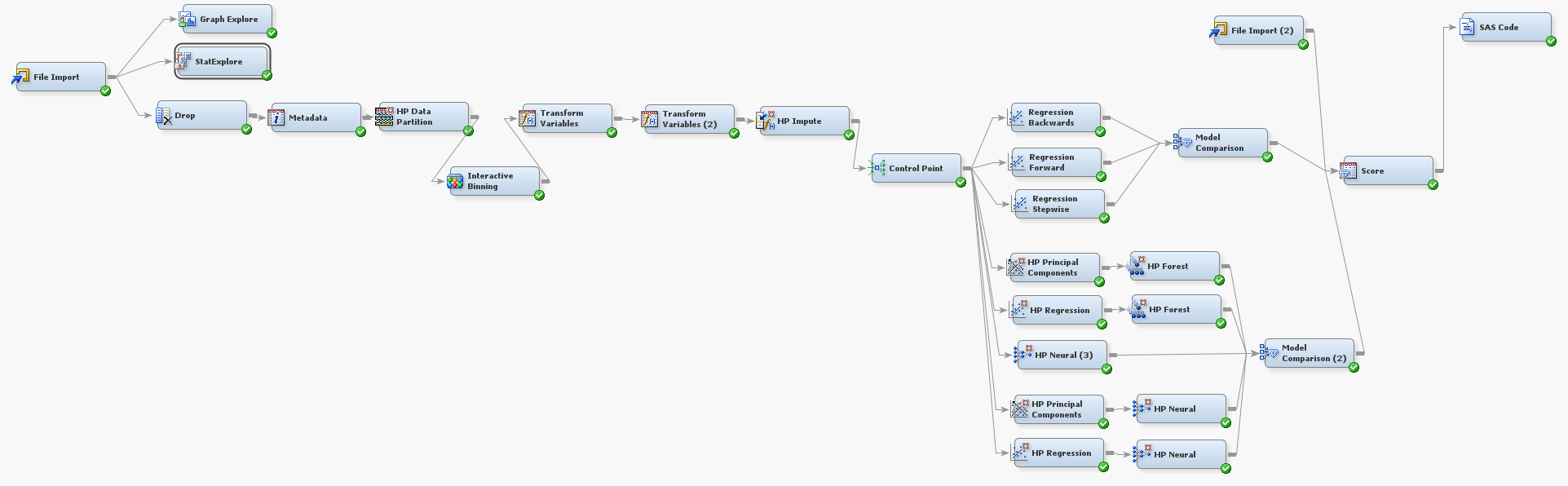


Table 15: HP Neural Network and HP Forests Comparison Output

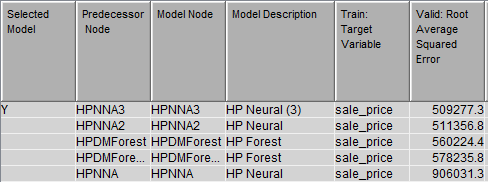


Table 16: HP Neural Network and HP Forests Comparison Output

Table

Description automatically generated

Graph 15: HP Neural Network and HP Forests Model Flow

Diagram

Description automatically generated