Housing Price Analysis

By James “Colton” Behannon

Table of Contents

Research Understanding

Objective

The objective of this project is to analyze the information about recently sold houses and to build a model that can be used to predict the prices of houses that have not yet been sold. Twenty recommendations of the most expensive homes will be estimated as these homes are deemed to have the highest potential investment growth. Model performance will be measured based on how many of these homes actually occur in the priciest homes list.

Data Understanding

Overview

There are four (4) raw data files containing information from an anonymous United States city assessor's office that is located in the North West region. The values in the data files are for individual residential real estate properties sold in that city over a time period of 4 years (year of sale not included in the scope of this project).

There is also one (1) Score data set that consists of many of the variables in the raw data files. There are a couple of caveats regarding the Score data set: a few variables that are in the raw data files are not available in the Score data set, the most critical of which is the Sale Price of the houses, along with a few others; there may be missing variables in your Score data set.

Data Files

1. Property Survey – 1 ◊ Contains 600 rows Variables: PID (Property Identification), Lot Area, Lot Shape, and Bldg Type

2. Property Survey – 2 ◊ Contains 1770 rows Variables: PID (Property Identification), Lot Area, Lot Shape, and Bldg Type

3. Quality Assessment Variables: PID (Property Identification), Overall Quality, Overall Condition, Exterior Quality, and Foundation

4. House Feature Variables: PID (Property Identification), Year Built, Year Remodel, Veneer Area of Exterior Wall , Bsmt Finish Type, Bsmt Finish Sqr ft, Bsmt Unfinish Sqr ft, Total Bsmt Sqr ft, Heating QC, 1st Flr Sqr ft, 2nd Flr Sqr ft, Above Ground Living Area, Number Full Bath Bsmt, Half Bath House, Number Full Bath House, Number Bedroom Above Ground, Number Room Above Ground, Fireplaces, Garage Type, Garage Cars, Garage Area, Wood Deck Sqr ft, Open Porch Sqr ft, Sale Price

5. Score Data - No Sale Price Variables: PID (Property Identification), Lot Area, Lot Shape, and Bldg Type, Overall Quality, Overall Condition, Year Built, Year Remodel, Veneer Area of Exterior Wall , Bsmt Finish Type, Bsmt Finish Sqr ft, Bsmt Unfinish Sqr ft, Heating QC, 1st Flr Sqr ft, 2nd Flr Sqr ft, Above Ground Living Area, Number Full Bath Bsmt, Half Bath House, Number Full Bath House, Number Bedroom Above Ground, Number Room Above Ground, Fireplaces, Garage Type, Garage Cars, Garage Area, Wood Deck Sqr ft, Open Porch Sqr ft

Variables

1. PID (Property Identification) – a unique number to identify each property.

2. Lot Area – The size of the lot, measured in square feet, on which the house is located.

3. Lot Shape – The general shape of the lot. A lot with a regular shape has a value of 1, and another with not a regular shape has a value of 0.

Data Understanding

4. Bldg Type (i.e., Building Type) – This describes the type of home in terms of its footprint. A single-family detached type of home is indicated by a value of 1, and a townhouse type of home is indicated by a value of 0.

5. Overall Quality – This is a rating of the overall material and finish of the house. The numeric scale of this rating is as follows. 10 - Very Excellent 9 - Excellent 8 - Very Good 7 -Good 6 - Above Average 5 - Average 4 - Below Average 3 - Fair 2 - Poor 1 - Very Poor

6. Overall Condition: This is a rating of the overall condition of the house. The numeric scale of this rating is as follows. 10 - Very Excellent 9 - Excellent 8 - Very Good 7 - Good 6 - Above Average 5 - Average 4 - Below Average 3 - Fair 2 - Poor 1 - Very Poor

7. Exterior Quality – This is a rating of the quality of the material on the exterior. A good quality is indicated by a 1, and an average quality is indicated by a 0. Predictive Analytics – Final Group Project 6

8. Foundation – This describes the type of foundation upon which the house is built. A concrete foundation is indicated by a value of 2; a cinder-block foundation by a value of 1; and brick foundation by a value of 0.

9. Year Built – This describes the year when the house was constructed.

10. Year Remodel – This describes the year when the house was remodeled. If the house was never remodeled, then the “year remodel” is the same as the “year built.”

11. Veneer Area of Exterior Wall – This describes the area in square feet of the exterior wall that is veneer.

12. Bsmt Finish Type (Basement Finished Type) – This indicates whether a home’s basement is finished or not in the sense that it can be lived in or not. When it is finished, it has a value of 1, and a value of 0 otherwise.

13. Basement Finished Sqr ft – This is the measure of the area of a finished basement.

14. Basement Unfinished Sqr ft – This is the measure of the area of an unfinished basement.

15. Total Bsmt Sqr ft – This is the measure of the total basement area.

16. Heating QC (Heating Quality Condition) – This is a measure of the rating of how well the heating unit is for a house. The rating scale is as follows. 3 - Excellent 2 - Good 1 - Average 0 – Fair

17. 1st Flr Sqr ft (First floor Sqt ft) – This is a measure of the living space on the first floor of a house.

18. 2nd Flr Sqr ft (Second floor Sqt ft) – This is a measure of the living space on the second floor of a house.

19. Above Ground Living Area – This is a measure of the living space of the entire house, excluding the basement. Predictive Analytics – Final Group Project 7

20. Number Full Bath Bsmt - This indicates the number of full bathrooms in the basement of a house. A value of 1 indicates that there is a full bathroom and a value of 0 indicates that there is not a full bathroom in the basement.

Data Understanding

21. Half Bath House - This indicates whether there is a half bathroom in the house (excluding the basement). A value of 1 indicates that there is a half bathroom and a value of 0 indicates that there is not a half bathroom in the house.

22. Number Full Bath House - This indicates the number of full bathrooms there are in the house, not including bathroom in the basement.

23. Bedroom Above Ground - This indicates the number of bedrooms there are in the house, not including the basement.

24. Number Room Above Ground - This indicates the number of rooms there are in the house, not including the basement.

25. Fireplaces – This indicates the number of fireplaces there are in the house, not including the basement.

26. Garage Type – Whether there is a garage of a given type is described and indicated as follows. 3 - Attached to house 2 - Built-In (Garage part of house - typically has room above garage) 1 - Detached from home 0 - No garage

27. Garage Cars – This indicates the number of cars that can be accommodated in the garage of the house.

28. Garage Area – This is the size of garage in square feet.

29. Wood Deck Sqr ft – This is the size of the wood deck area in square feet for a house.

30. Open Porch Sqr ft - This is the size of the open porch area in square feet for a house.

31. Sale Price – This is the sale price of a house (not included in the Score data set).

<u>Master Data Preparation: Importing Data into SAS EM</u>

In order to get the data files into SAS EM, SAS EG was first used to get the data in order. First, Property Survey - 1 and Property Survey - 2 were appended so as to have the survey information for every PID in one place. After appending these sets into one, it was inner joined with the House Features and Quality Assessment data sets. We now have all of our variables in one place and have a dataset with 31 columns and 2370 rows. This dataset was then exported into a file named Houses_1 and then imported into SAS EM.

Upon importing to SAS EM, the variables had to be given the correct configurations so that they would operate correctly in our project lifecycle. The resulting configurations are displayed in the following figure:

Data Understanding

Master Data Preparation: Data Configuration

Name	Role	Level
Above_Ground_Living_Area	Input	Interval
Bldg_Type	Input	Binary
Bsmt_Finish_Sqr_ft	Input	Interval
Bsmt_Finish_Type	Input	Interval
Bsmt_Unfinish_Sqr_ft	Input	Interval
Exterior_Quality	Input	Binary
Fireplaces	Input	Interval
Foundation	Input	Nominal
Garage_Area	Input	Interval
Garage_Cars	Input	Interval
Garage_Type	Input	Nominal
Half_Bath_House	Input	Binary
Heating_QC	Input	Ordinal
Lot_Area	Input	Interval
Lot_Shape	Input	Binary
Number_Bedroom_Above_Ground	Input	Interval
Number_Full_Bath_Bsmt	Input	Binary
Number_Full_Bath_House	Input	Interval
Number_Room_Above_Ground	Input	Interval
Open_Porch_Sqr_ft	Input	Interval
Overall_Condition	Input	Ordinal
Overall_Quality	Input	Ordinal
PID	ID	Nominal
Sale_Price	Target	Interval
Total_Bsmt_Sqr_ft	Input	Interval
Veneer_Area_of_Exterior_Wall	Input	Interval
Wood_Deck_Sqr_ft	Input	Interval
Year_Built	Input	Interval
Year_Remodel	Input	Interval
_1st_Flr_Sqr_ft	Input	Interval
_2nd_Flr_Sqr_ft	Input	Interval

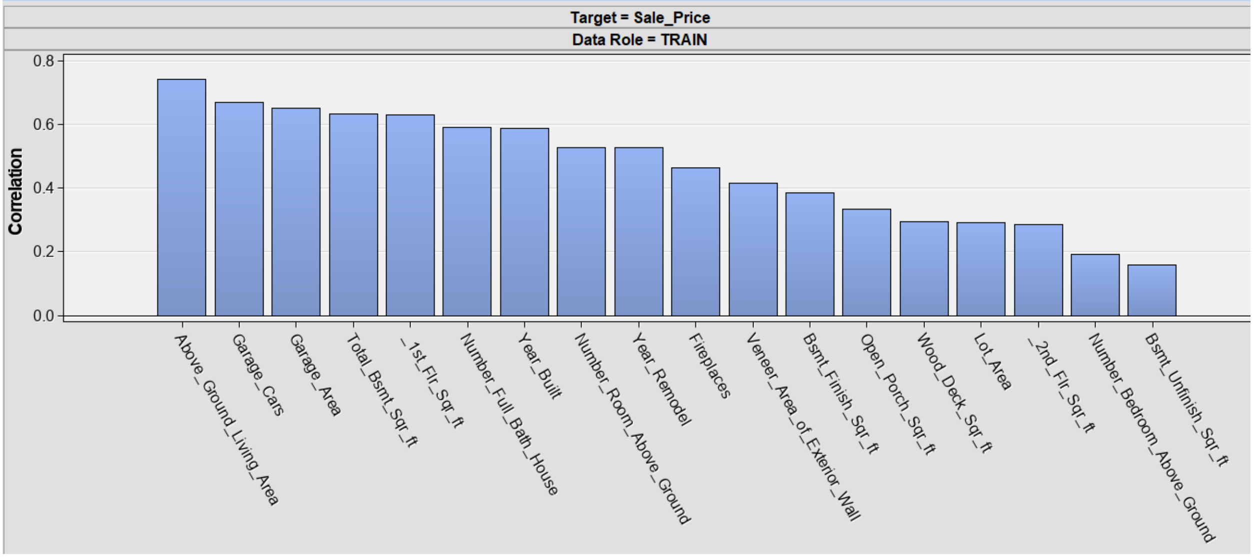
Role	Level	Count
ID	Nominal	1
Input	Binary	6
Input	Interval	18
Input	Nominal	2
Input	Ordinal	3
Target	Interval	1

Data Understanding

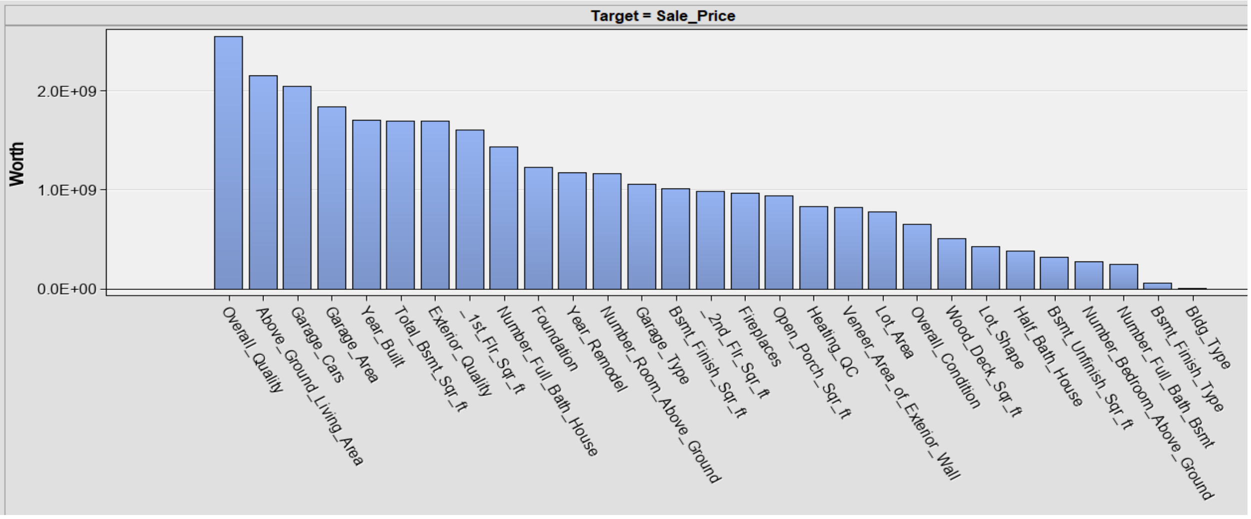
Master Data Preparation: EDA

The first step taken into exploratory data analysis for this project is to connect the Stat Explore node to our HOUSES_1 node. This will be the source of our initial insight.

Information about the contents of the numeric variables can be gained by examining the Pearson Correlation with Sale_Price. From the chart we can see that the 9 leftmost variables have a strong correlation with a value over 0.5. The next 7 have a moderate correlation below 0.5 but above 0.2. The final 2 variables have a weak correlation that is below 0.2.

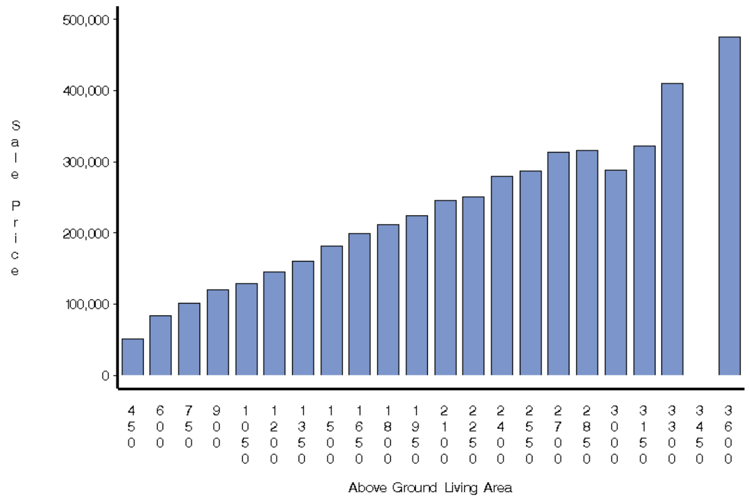


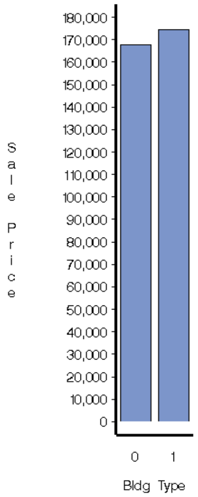
Next, we take a look at "Variable Worth". Variable worth plots the variables in order by their worth in predicting Sale_Price by using the Gini split worth statistic. From this metric we can see that the quality of the home (Overall_Quality), size of the living space (Above_Ground_Living_Area), size of the garage (Garage_Cars and Garage_Area), and how new the home is (Year_Built) are just a few of the more predictive variables. These attributes all make sense and would be things that every home buyer considers before making an offer.

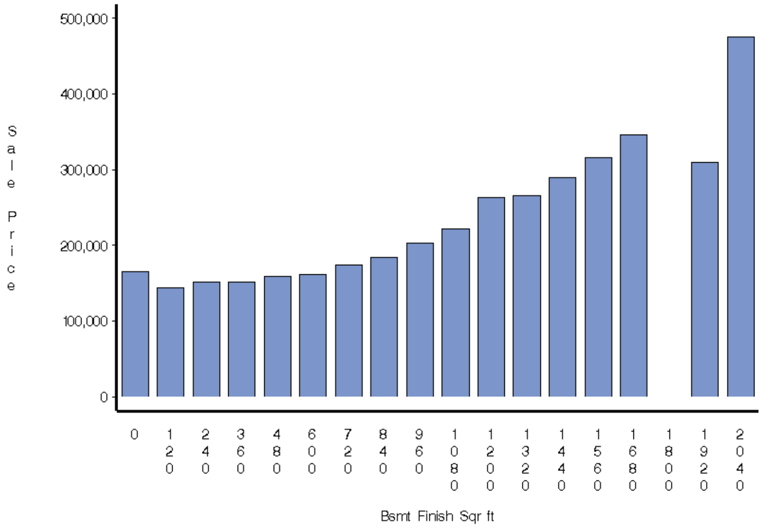


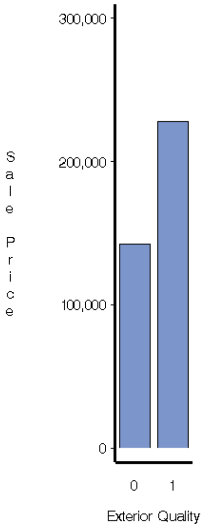
Data Understanding

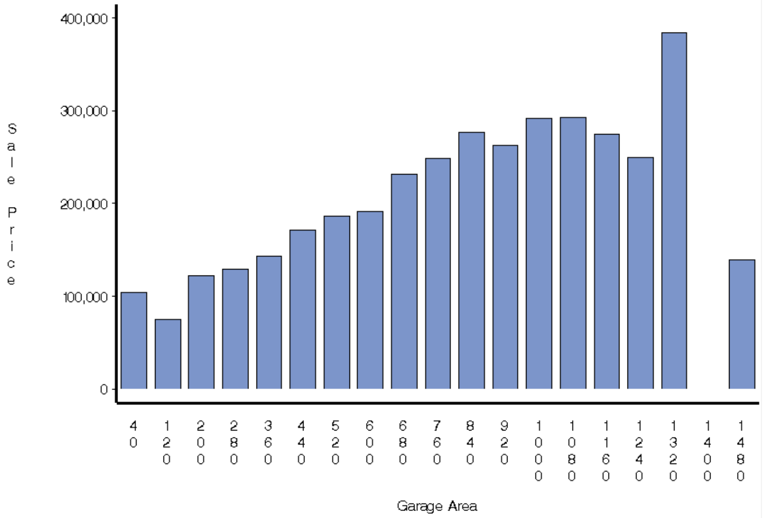
Interesting MultiPlot Results

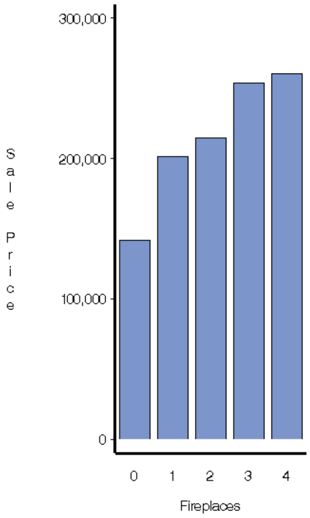

Above_Ground_Living_Area vs Sale_Price (mean)


Bldg_Type vs Sale_Price (mean)


Bsmt_Finish_Sqr_ft vs Sale_Price (mean)

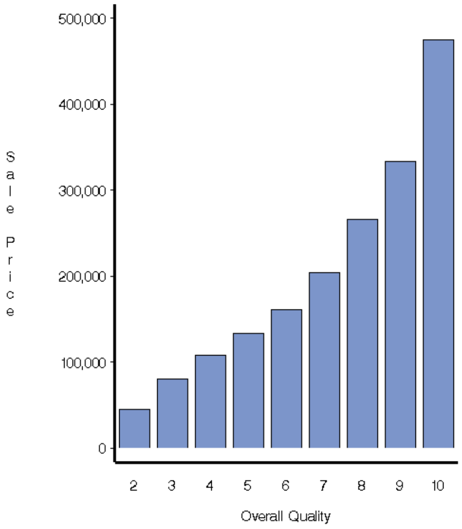

Exterior_Quality vs Sale_Price (mean)


Garage_Area vs Sale_Price (mean)

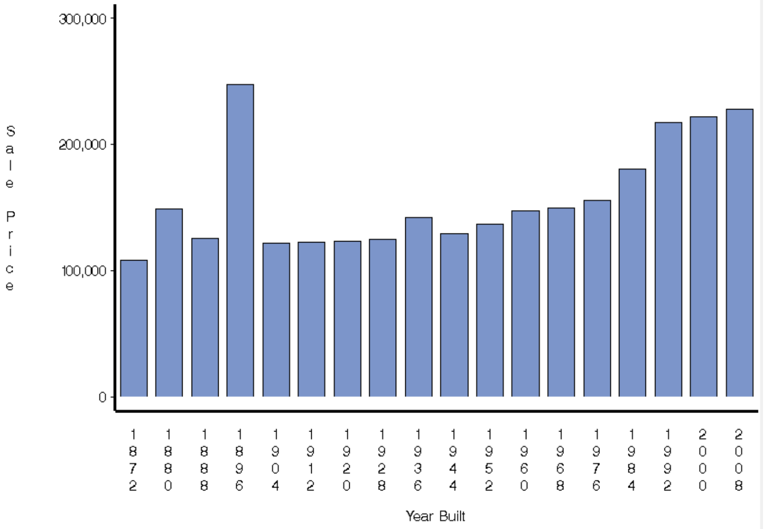

Fireplaces vs Sale_Price (mean)

Data Understanding

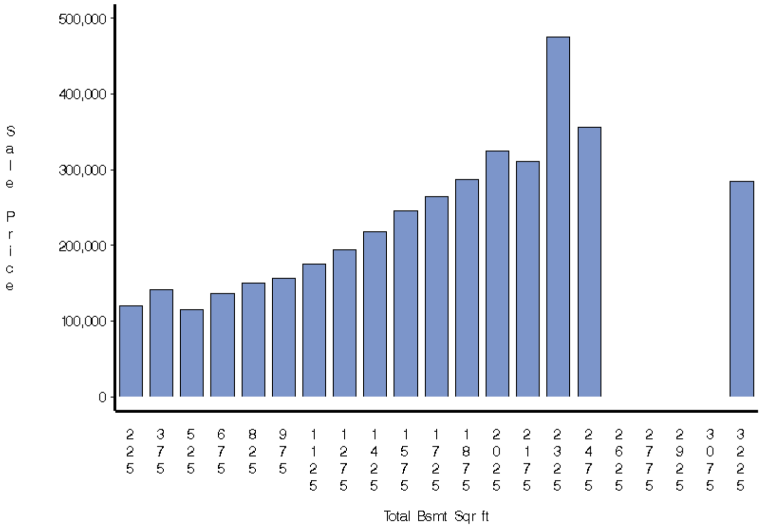
Overall_Quality vs Sale_Price (mean)



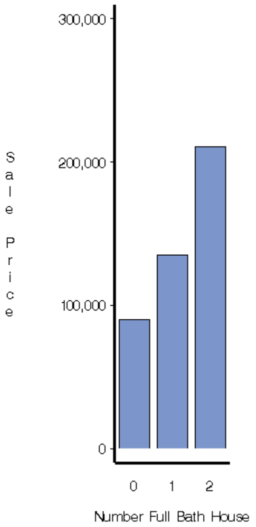
Year_Built vs Sale_Price (mean)



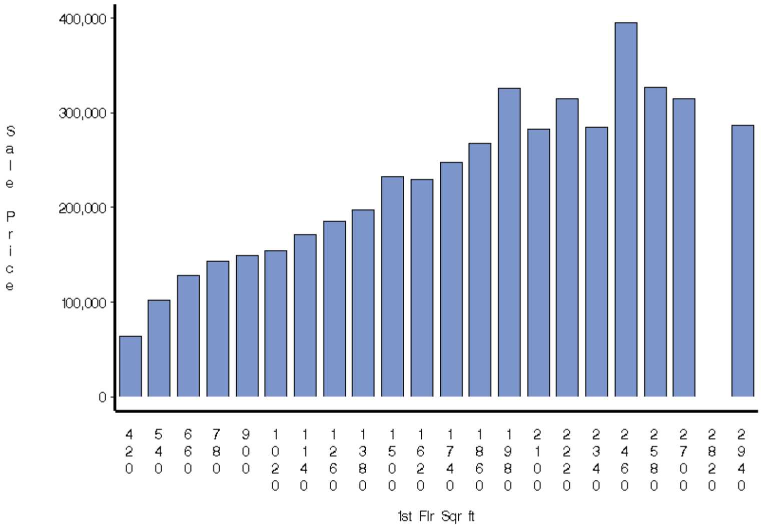
Total_Bsmt_Sqr_ft vs Sale_Price (mean)



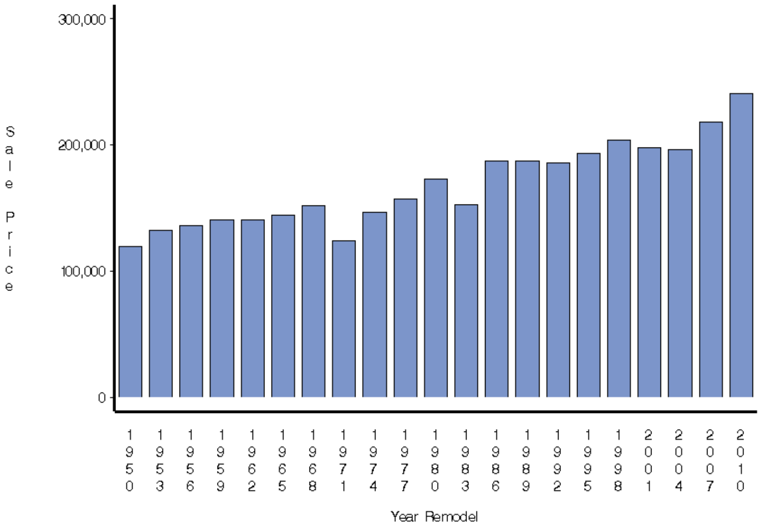
Number_Full_Bath_House vs Sale_Price (mean)



_1st_Flr_Sqr_ft vs Sale_Price (mean)



Year_Remodel vs Sale_Price (mean)



Data Understanding

There do not appear to be any missing values as shown by the running of the Impute node. The results yielded no changed values as there were none missing.

Score Data Preparation: Data Configuration

Name	Role	Level
Above_Ground_Living_Area	Input	Interval
Bldg_Type	Input	Binary
Bsmt_Finish_Sqr_ft	Input	Interval
Bsmt_Finish_Type	Input	Interval
Bsmt_Unfinish_Sqr_ft	Input	Interval
Fireplaces	Input	Interval
Garage_Area	Input	Interval
Garage_Cars	Input	Interval
Garage_Type	Input	Nominal
Half_Bath_House	Input	Binary
Heating_QC	Input	Ordinal
Lot_Area	Input	Interval
Lot_Shape	Input	Binary
Number_Bedroom_Above_Ground	Input	Interval
Number_Full_Bath_Bsmt	Input	Binary
Number_Full_Bath_House	Input	Interval
Number_Room_Above_Ground	Input	Interval
Open_Porch_Sqr_ft	Input	Interval
Overall_Condition	Input	Ordinal
Overall_Quality	Input	Ordinal
PID	ID	Nominal
Veneer_Area_of_Exterior_Wall	Input	Interval
Wood_Deck_Sqr_ft	Input	Interval
Year_Built	Input	Interval
Year_Remodel	Input	Interval
_1st_Flr_Sqr_ft	Input	Interval
_2nd_Flr_Sqr_ft	Input	Interval

Role	Level	Count
ID	Nominal	1
Input	Binary	4
Input	Interval	18
Input	Nominal	1
Input	Ordinal	3

Data Understanding

Score Data Preparation: EDA

When inspecting the summary statistics through StatExplore > View > Summary Statistics it informs us that there are no missing values. However, we can see that there are some missing variables in this dataset that are present in the other such as Foundation, Exterior_Quality and Total_Bsmt_Sqr_ft. These variables might have some relevance as they seem to have a decently high worth on the chart for predicting Sale_Price but we will simply ignore them as it would likely require more effort than the benefit is worth. Total_Bsmt_Sqr_ft could possibly be inferred from Bsmt_Finish_Sqr_ft and Bsmt_Unfinish_Sqr_ft. However, the variables Foundation and Exterior Quality do not seem possible to be realized.

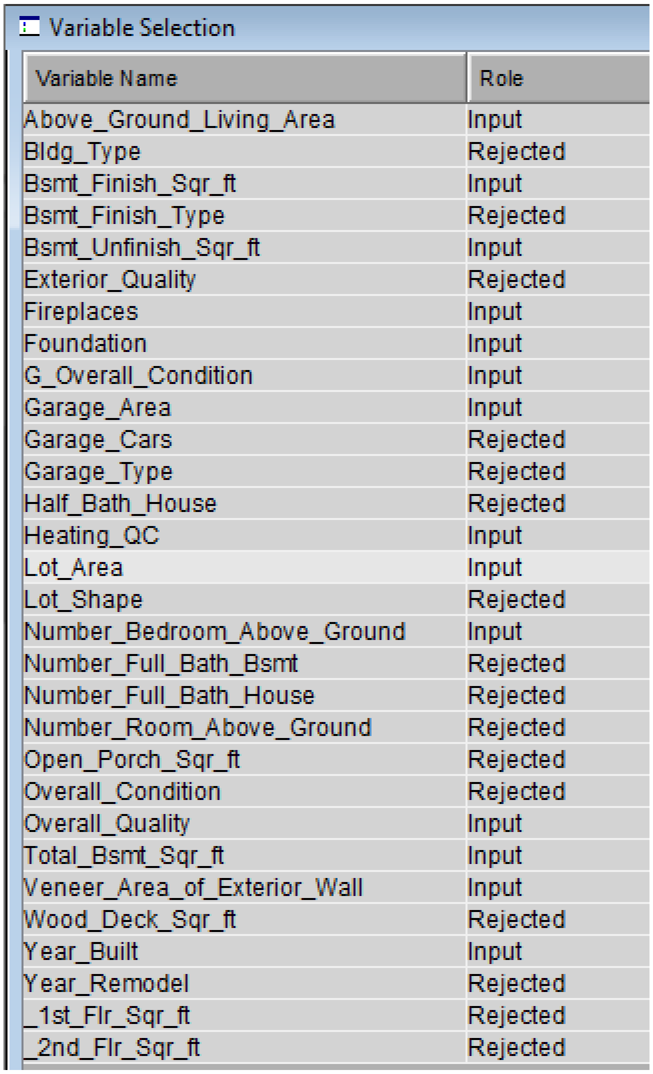
Something to note about the Score Data is that the distributions of the variables will not be the same as the Master Data as there is a much smaller sample. For example, the range of Overall_Quality in the Master Data spans from 2 to 10 whereas in the Score set it is 4 to 9. Another example is how Garage_Cars spans from 0 to 5 in the Master Data but from 1 to 3 on the Score Data.

Data Preparation

Data Preparation

We used the Impute node, set to median, on both the Master data and the Score data in order to check if there were any values that needed imputation. However, there do not appear to be any missing values.

Next we run the variable selection node which presents us with what variables SAS recommends rejecting based on small R-square values. This does a good job of selecting variables that were identified on the higher end within our “variable worth” selection but also rejects variables that could be superfluous such as Garage_Cars which is highly correlated with Garage_Area. It is interesting that it rejects Year_Remodel even thought it had relatively high worth but lacks the “R-square” to make the cut. It is also important to note that among the variables selected are Total_Bsmt_Sqr_ft and Foundation, which we have already decided to reject. This further leads us to conclude that having these variables available in the Score dataset could have given us better results.



Variable Selection

Variable Name	Role
Above_Ground_Living_Area	Input
Bldg_Type	Rejected
Bsmt_Finish_Sqr_ft	Input
Bsmt_Finish_Type	Rejected
Bsmt_Unfinish_Sqr_ft	Input
Exterior_Quality	Rejected
Fireplaces	Input
Foundation	Input
G_Overall_Condition	Input
Garage_Area	Input
Garage_Cars	Rejected
Garage_Type	Rejected
Half_Bath_House	Rejected
Heating_QC	Input
Lot_Area	Input
Lot_Shape	Rejected
Number_Bedroom_Above_Ground	Input
Number_Full_Bath_Bsmt	Rejected
Number_Full_Bath_House	Rejected
Number_Room_Above_Ground	Rejected
Open_Porch_Sqr_ft	Rejected
Overall_Condition	Rejected
Overall_Quality	Input
Total_Bsmt_Sqr_ft	Input
Veneer_Area_of_Exterior_Wall	Input
Wood_Deck_Sqr_ft	Rejected
Year_Built	Input
Year_Remodel	Rejected
_1st_Flr_Sqr_ft	Rejected
_2nd_Flr_Sqr_ft	Rejected

Next, we are finally ready to partition the data. We use a 50/50 split into Training and Validation data which leaves us with two samples of 1,185 each.

Data Modeling

Modeling

Regressions – The first two models we use are both linear regressions. We create one multiple regression that uses the variables identified within the variable selection node we added earlier. The second regression used is a forward regression that will select what it deems best from the entire variable catalog.

The regression with variable selection (minus Foundation and Total_Bsmt_Sqr_ft) has an average squared error of **4.6868E8** for the training set and **4.7737E8** for the validation set. These errors are relatively close which lets us know the model is not overfitting (at least not by a lot).

For the forward regression we manually select p-value of 0.10 for variable inclusion in order to include a larger set of variables. This model has an average squared error of **4.0193E8** for the training set and **4.3147E8** for the validation set. "The selected model, based on the error rate for the validation data, is the model trained in Step 19. It consists of the following effects:

[Intercept, Above_Ground_Living_Area, Bldg_Type, Bsmt_Finish_Sqr_ft, Fireplaces, Garage_Area, Garage_Cars, Heating_QC, Lot_Area, Lot_Shape, Number_Bedroom_Above_Ground, Number_Full_Bath_Bsmt, Open_Porch_Sqr_ft, Overall_Condition, Overall_Quality, Veneer_Area_of_Exterior_Wall, Year_Built, Year_Remodel, _1st_Flr_Sqr_ft , _2nd_Flr_Sqr_ft]"

Both of these models were trained without standardization because regressions do not require it as a precursor. However, that does not mean that we cannot still get better results under standardization. After adding the variable transformation node and using "standardization" for our interval inputs we obtain the following results.

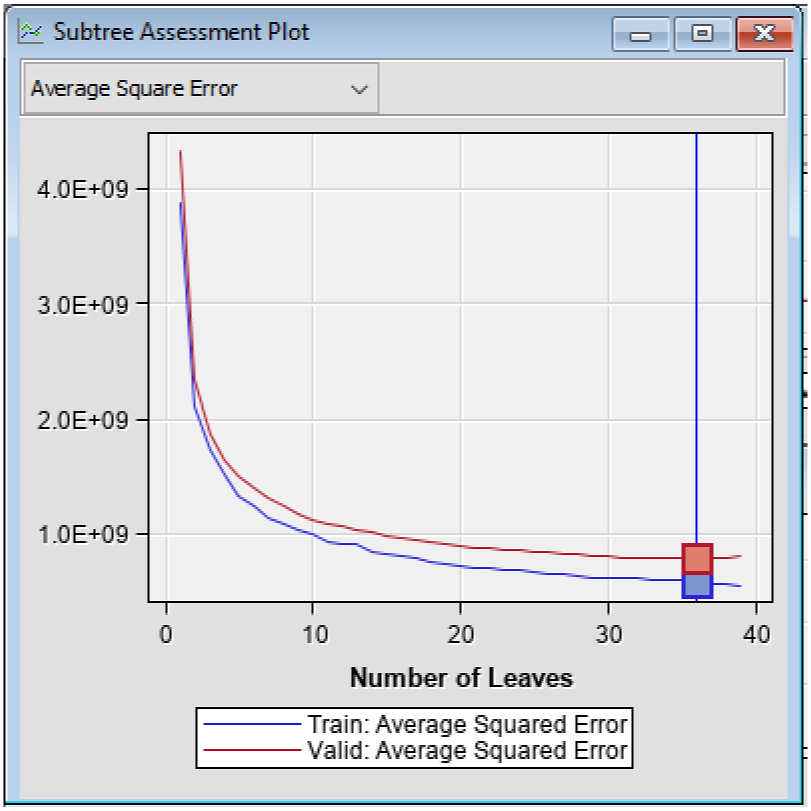
Regression (with variable selection): average squared error of **4.393E8** for the training set and **4.594E8** for the validation set.

Forward Regression: average squared error of **3.9172E8** for the training set and **4.2489E8** for the validation set.

Our results after standardization are significantly better for the model with manual feature selection and still slightly better for the forward regression, which has our best error rate so far.

Data Modeling

Decision Tree – The decision tree node is connected directly after the variable transformation node we added within the regression section. The decision tree uses the default settings and returns an average squared error of **5.9109E8** on the training set and **7.9501E8** on the validation set with 36 leaves. This large difference between the two errors likely suggests overfitting of the model although both the training and validation data continue to increase in accuracy until 36 leaves.

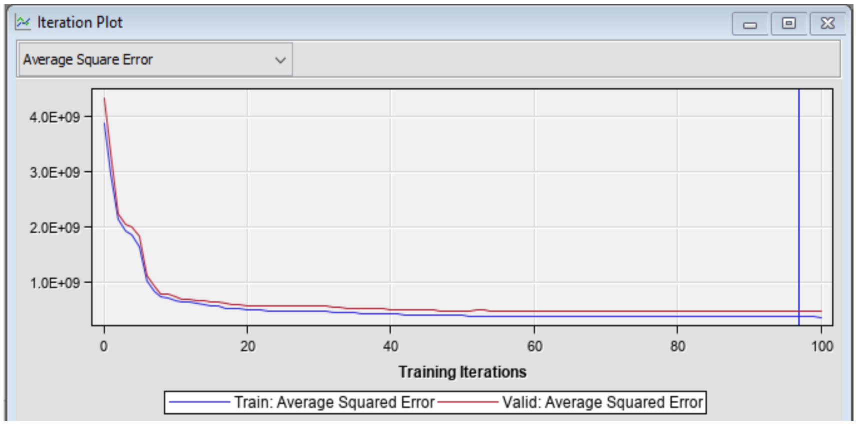


If we run the decision tree without using the standardization node we actually obtain slightly better results. The average squared error under this model is **5.8374E8** for the training set and **7.7828E8** for the validation set with 37 leaves. There is still a large gap between the training and validation average squared error.

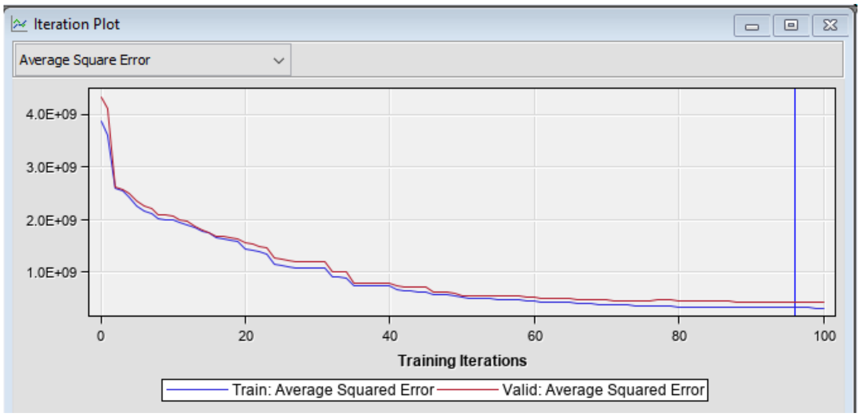
Data Modeling

Neural Network – The final model we will use to predict the SalePrice of a home is the neural network.

When using the neural network node with 3 hidden nodes attached directly to our data partition node, we obtain an average squared error of **3.6837E8** for the training set and **4.6335E8** for our validation set which has the largest gap between the two errors yet.

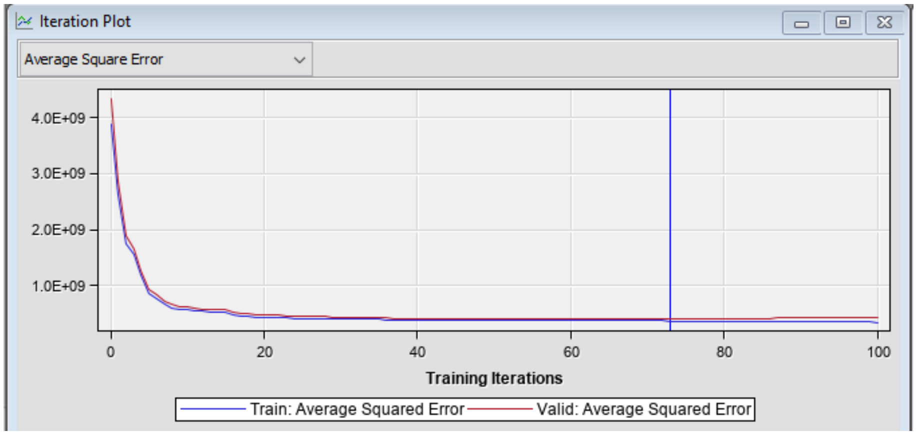


If we run the neural network with 4 hidden nodes, we obtain an average squared error of **3.1997E8** on the testing set and **4.3236E8** on the validation set.

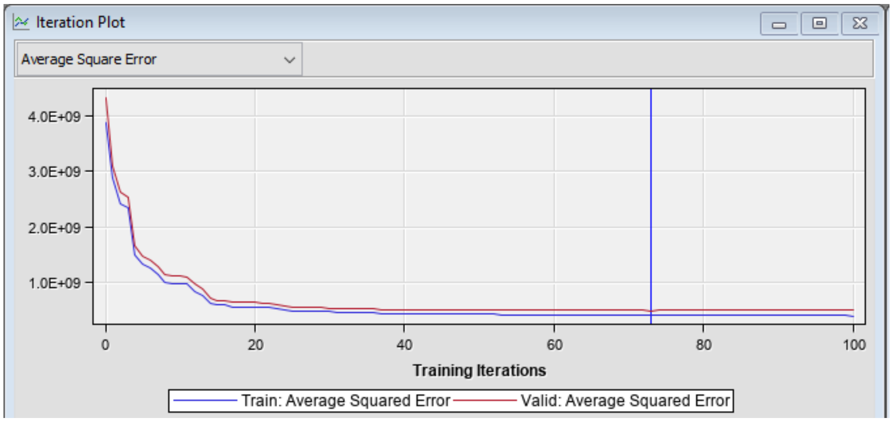


Data Modeling

If we run the neural network with 5 hidden nodes, we obtain an average squared error of **3.7762E8** on the testing set and **4.1356E8** on the validation set. This is now our best performing model by average squared error of the validation set.



If we run the neural network with 6 hidden nodes, we obtain an average squared error of **4.1207E8** on the training set and **4.9848E8** for the validation set. Here the performance of our validation set has decreased.



If we run these models after our variable transformation nodes, as was done for the other models, there will be no difference in results as the neural network node standardizes the variables by default.

Evaluation

Model Results and Best Model Selection

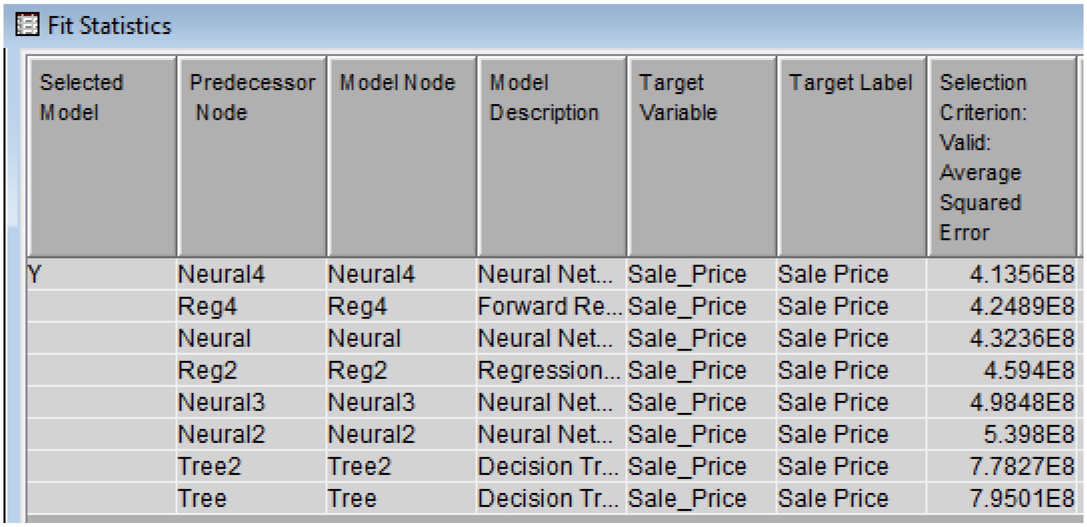
Model	Validation: Average Squared Error
Regression (with variable selection)	4.7737E8
Forward Regression	4.3147E8
Regression (with variable selection) - standardized	4.5940E8
Forward Regression - standardized	4.2489E8
Decision Tree	7.9501E8
Decision Tree - standardized	7.7828E8
Neural Network (3 Hidden Nodes)	4.6335E8
Neural Network (4 Hidden Nodes)	4.3236E8
Neural Network (5 Hidden Nodes)	**4.1356E8**
Neural Network (6 Hidden Nodes)	4.9848E8

*Best model of each type is underlined, best model overall is **bold**

Best Model Summary

After performing analysis with several models, the best performance came from a neural network model with one hidden layer and 5 hidden nodes with a validation average squared error of 4.1356E8. This model will be what is used on our Score dataset in order to predict the 20 most expensive homes from the list of 100.

This analysis is backed up by the model comparison node which selects "Neural4" (5 Hidden Nodes) as the best performing model.



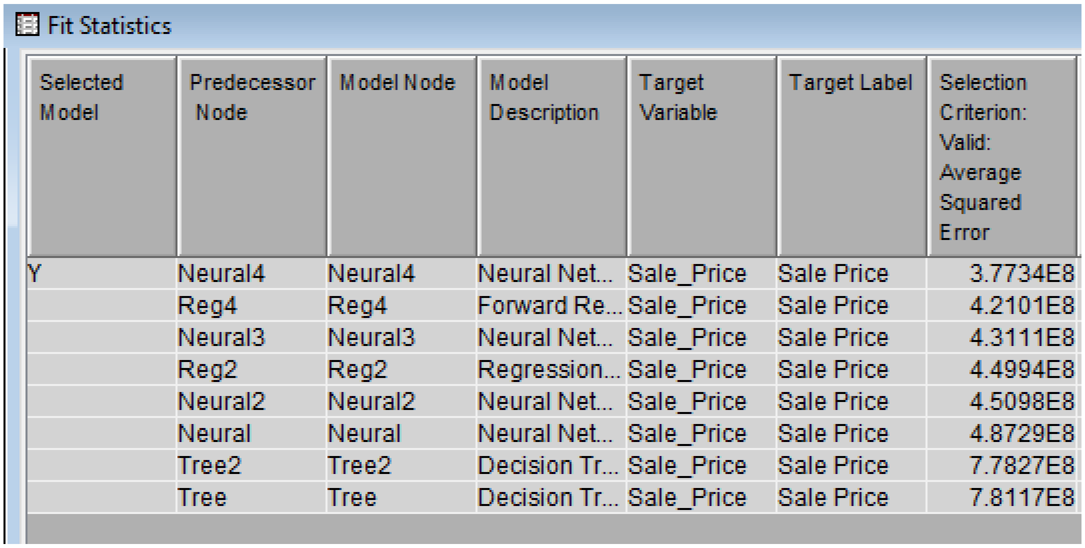
Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error
Y	Neural4	Neural4	Neural Net...	Sale_Price	Sale Price	4.1356E8
	Reg4	Reg4	Forward Re...	Sale_Price	Sale Price	4.2489E8
	Neural	Neural	Neural Net...	Sale_Price	Sale Price	4.3236E8
	Reg2	Reg2	Regression...	Sale_Price	Sale Price	4.594E8
	Neural3	Neural3	Neural Net...	Sale_Price	Sale Price	4.9848E8
	Neural2	Neural2	Neural Net...	Sale_Price	Sale Price	5.398E8
	Tree2	Tree2	Decision Tr...	Sale_Price	Sale Price	7.7827E8
	Tree	Tree	Decision Tr...	Sale_Price	Sale Price	7.9501E8

Evaluation

Could the Model Still be Improved?

The selected model based on best performance (lowest average squared error) is the neural network with 5 hidden nodes. The neural network SAS EM node automatically standardizes the variables and we therefore did not need to do so before creating our model. However, what if we were to include the replacement node beforehand? The replacement node eliminates the more intense values so that your model has less potential noise and is filled with more average data. When we do this, the results are quite interesting.



Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error
Y	Neural4	Neural4	Neural Net...	Sale_Price	Sale Price	3.7734E8
	Reg4	Reg4	Forward Re...	Sale_Price	Sale Price	4.2101E8
	Neural3	Neural3	Neural Net...	Sale_Price	Sale Price	4.3111E8
	Reg2	Reg2	Regression...	Sale_Price	Sale Price	4.4994E8
	Neural2	Neural2	Neural Net...	Sale_Price	Sale Price	4.5098E8
	Neural	Neural	Neural Net...	Sale_Price	Sale Price	4.8729E8
	Tree2	Tree2	Decision Tr...	Sale_Price	Sale Price	7.7827E8
	Tree	Tree	Decision Tr...	Sale_Price	Sale Price	7.8117E8

It still selects Neural4 as the best model but look at what has happened to the average squared error. IT has decreased by roughly 0.3622E8. For perspective, this much of a decrease on any of the other models would have made them the best performing. However, Neural (4 hidden nodes) increased its error by roughly 0.5E8.

When considering whether to keep this change it is important to think about what is actually happening. Yes, it is increasing performance, but it is doing so by reducing the more extreme values. These values are the ones that are most likely to characterize a home with a large price tag. In the end, our objective is to identify the 20 priciest homes and could therefore (possibly) run into trouble by using the replacement node in our model.

Also, when trying to predict our Score dataset it returns the same price for every home, which is clearly an error. We will stick with the original model.

Deployment

Deploying the Chosen Model

After connecting the Neural4 model and the Score dataset to the Score node we receive the house sale price predictions. Just glancing over the results, the criteria appears to make sense. The most expensive homes are characterized by the highest ratings of "Overall Quality" as well as generally larger living spaces and more bathrooms.

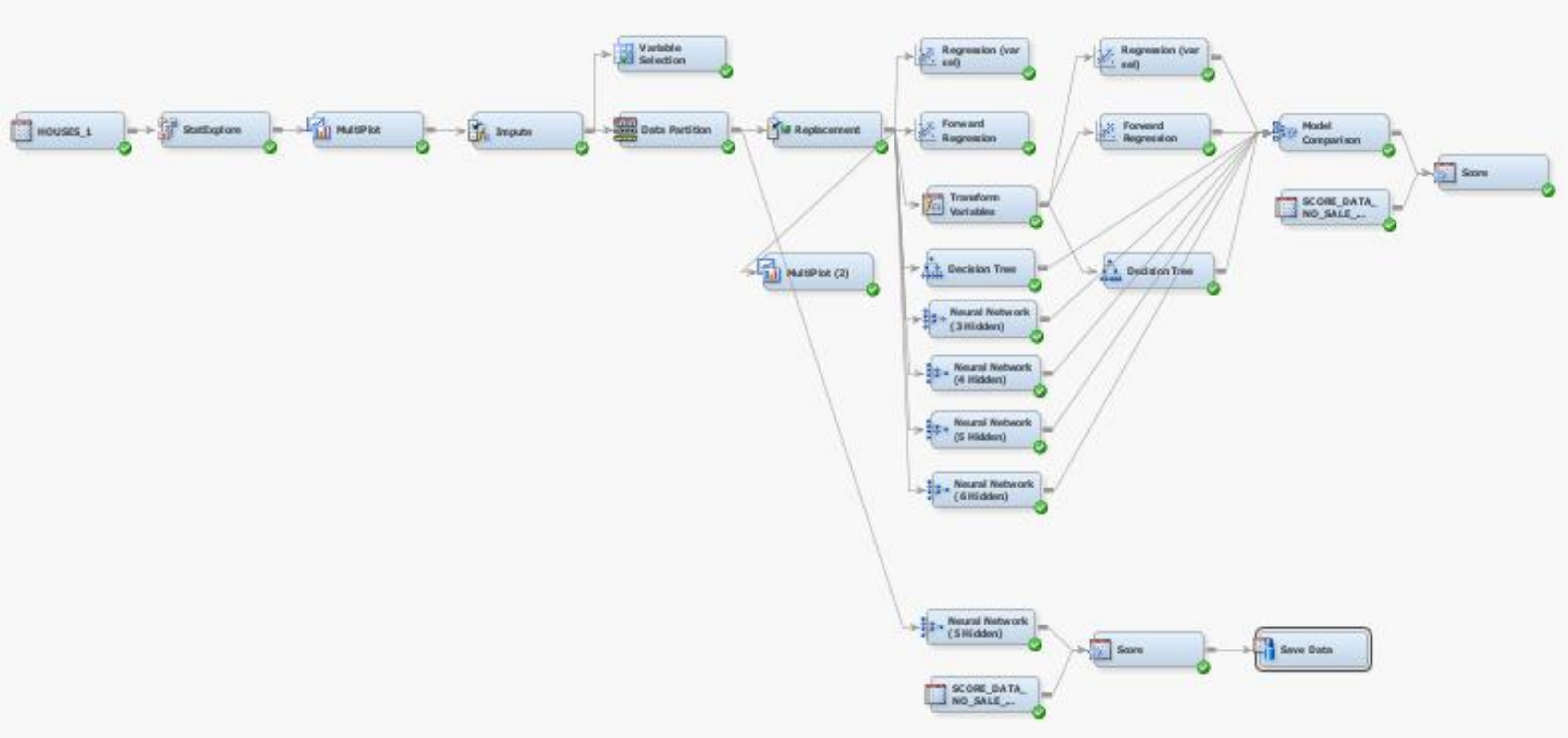
Results: The Top 20 Luxury Homes

PID	Sale_Price
0528166120	341620.3
0528178070	335175.1
0528118040	330693.5
0528114050	328280.8
0528172080	291011
0528170070	272143.6
0528445060	258633.1
0907253110	257072.4
0907254020	255819.5
0907285020	246938
0534128010	236643.6
0907254090	236441.9
0528280180	228485
0528253020	224542.7
0534128020	222096.5
0528228290	221941.5
0907181140	219890.4
0528458040	219446.9
0531382120	219290.7
0528235090	218983.6

Summary:

After testing numerous models and various iterations, we have arrived at our best model and the results of its sale price predictions. The predictions all seem reasonable and their attributes line up with what would make up a luxury home. If we were to try and create a better model in the future, we might try to capture the total basement square footage that is available within the Master set within our Score set. However, after all of the work put in, I am comfortable with the model arrived at.

Final Model Diagram


HOUSES_1
StatExplore
MultiPlot
Impute
Variable Selection
Data Partition
Replacement
MultiPlot (2)
Regression (var sel)
Forward Regression
Transform Variables
Decision Tree
Neural Network (3 Hidden)
Neural Network (4 Hidden)
Neural Network (5 Hidden)
Neural Network (6 Hidden)
Regression (var sel)
Forward Regression
Decision Tree
Model Comparison
SCORE_DATA_NO_SALE_...
Score
Neural Network (5 Hidden)
SCORE_DATA_NO_SALE_...
Score
Save Data