ISYS 5843 – Spring 2023
Group Project

UNIVERSITY OF ARKANSAS
SAM M. WALTON COLLEGE OF BUSINESS
Department of Information Systems

ARIZONA HOUSING DATA PROJECT

Olubunmi Mercy Ojudun

Priyanka Bejgam

Table of Contents

PART 1: EXPLORATORY DATA ANALYSIS

The Arizona Housing dataset has 12,297 observations that contain details of houses listed for sale as well as their respective agents/agencies, features, and other relevant information. There are a total of 106 variables in the dataset comprising categorical data, numerical data, and dates. Of these 106 variables, 28 are quantitative (Measures) variables, 66 are qualitative (Categorical) variables and 12 are dates [***Refer to Table 1 for comprehensive data dictionary]***. The combination of these different data types makes it ideal for different exploratory analysis such as visual analytics, descriptive statistics, and data distribution/ relationship analysis.

1.1: DATA QUALITY

Missing data:

The data for several predictor variables including Legal, End_Date, Temp_Off_Market_Date, Cancel_Date, UCB_or_CCBS, Unit__, St_Dir_Sfx, VAR34, Assessor_Parcel_Ltr, Out_of_Area_Schl_Dst, Week_Avail_Timeshare, Comp_to_Subagent, Other_Compensation, Guest_House_SqFt, and Lead_Based_Hazard_Disclosure have blank values, meaning there is no information recorded for these variables.

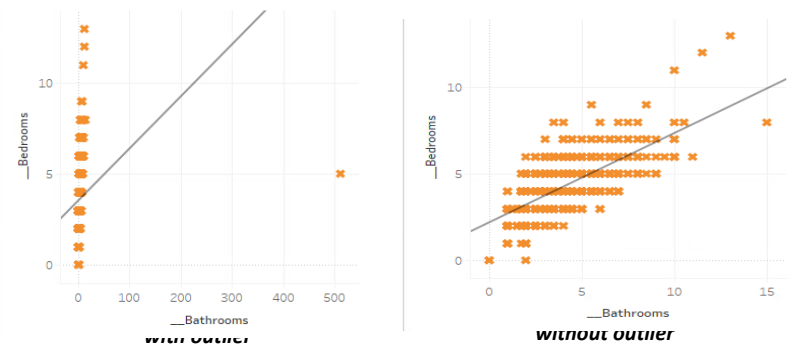
In addition, Flood_Zone, Hndrd_Blk_Directionl, Co_Selling_Agent, and Model are also missing data to a greater extent. It is also important to note that some variables, such as "Legal", "Directions", "Subdivision", "Agency_Name", "Agency_Phone", "Listing_Agent", "Co_Listing_Agent", "Selling_Agency", "Selling_Agent", and "Co_Selling_Agent", may not be useful as predictors for our target variable.

As a result, we would be dropping all the variables with more than 50% of the data missing from our models. For the other variables with less than 50% missing data, we would be making use of the Replacement/Impute method to handle these missing values.

Outliers & Incorrect Data:

Using Histograms and Box Plots which are useful at detecting outliers in continuous variables, we have been able to determine that there exist several variables with extreme outliers in the dataset. Further exploration of the dataset revealed that some of these outliers might have been due to data entry errors and variability in measurement. An example of such incorrect data

entry which is likely a mistake rather than a legitimate outlier is for the variable bathrooms which has a data point for the number of bathrooms as 512. Another instance of an incorrect data entry resulting in an outlier is for the variable Price SqFt which has a datapoint of 320,000. Inclusion of these outliers could lead to incorrect predictions in our model and as have even observed to also cause problems during our statistical analysis [***Refer to Figure B5(a), B5(b), B8(a) and B8(b)***]. These outliers may need to be excluded from the dataset during model building.



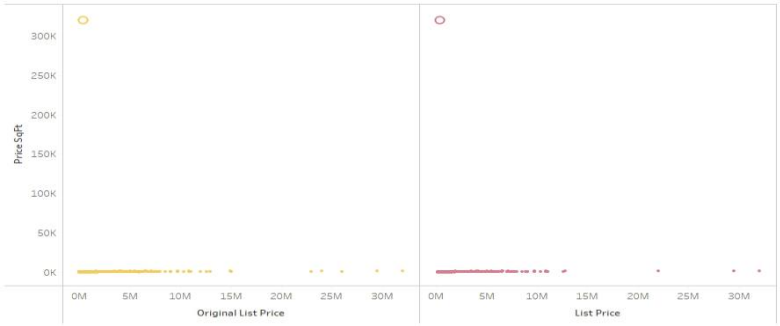


Figure B8(a): Price Square Ft vs List Price and Original List Price with outliers

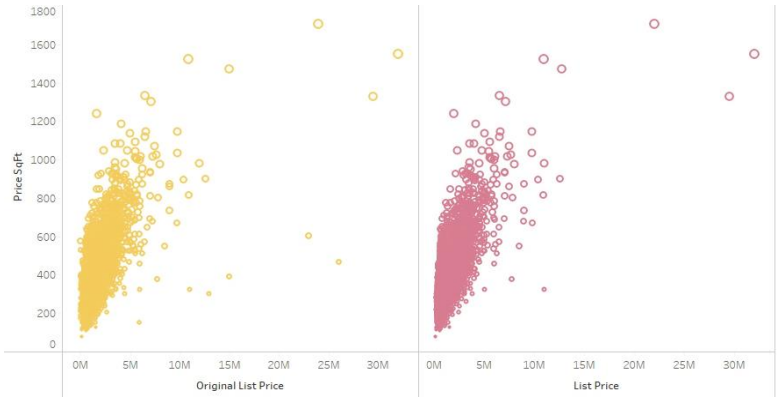


Figure B8(b): Price Square Ft vs List Price and Original List Price without outliers

Negative Values:

There were no negative values for all numeric data except Longitude which is to be expected and Days on Market. This is consistent with our expectations as prices, taxes, square feet, number of rooms etc., are unlikely to be negative.

Upon further investigation of the presence of negative values for Days on Market, we concluded that it might also be some form of data entry error due to the following reasons:

- Using both the On_Market_Date and Off_Market_Date to calculate the days on market we realized that those data observations for both fields were the same indicating that the houses effectively spent zero days on the market.
- It is highly unlikely that a property will be listed on the market for negative days.

1.2: DATA DISTRIBUTION

After checking the quality of the data, we then decided to evaluate the distribution of the continuous variables to determine their degree of symmetry/skewness. From our observations the majority of the variables were right skewed indicating that most of the observations are on the small/medium sides with a few much larger than the rest. The variables observed to be right

skewed include Approx Lot Acres, Approx Lot SqFt, Compensation to Buyer Broker, Days on Market, List Price, Original List Price, Sold Price, Taxes, Bathrooms, Approx SqFt, Interior Levels, Exterior Levels, Bedrooms Plus, Bedrooms, and Price SqFt [***Refer to Figures A1, A2, A3, A4, A7, A8, A9, A10, A11, A13, A14, A15, A16, A17, A18***]. Further analysis on the relationship between these variables in the next section will add credibility to our hypothesis.

On the other hand, Year Built is left skewed which indicates that most of the houses are recently built houses with a few old houses being listed on the market. These old houses can also be considered as outliers in the distribution [***Refer to Figure A12].***

1.3 DATA RELATIONSHIPS & PATTERNS

Relationships between the predicting factors

Overall, most of the predictor variables have a positive relationship with another which gives credence to our hypothesis made above that as one variable increases, the others are likely to increase. There were, however, some predictor variables that appeared to be almost perfectly correlated with each other. This means that using both variables would lead to high multicollinearity and as such we would be dropping some of these highly correlated predictor variables. An instance of such perfect correlation is between Approx Lot Acres and Approx Lot SqFt with correlation coefficient of 99.99%. This tells us that both variables are referencing the same figures albeit in different units of measurements. Another instance of almost perfect correlation is between List Price and Original List Price. Below is a summary of the relationships and correlations of the predictor variables.

1. The relationship between list price and original list price is positive and highly correlated with a correlation coefficient of 94.03%. [***Refer to Figure B6***]. We might need to consider using only one of the two.

Figure B6: Original List Price vs List Price

2. The relationship between Approx_Lot_Acres and Approx_Lot_SqFt is positive and highly correlated with a correlation coefficient of 99.99%. [***Refer to Figure B2 above***]. We would be removing one of these variables from the model.

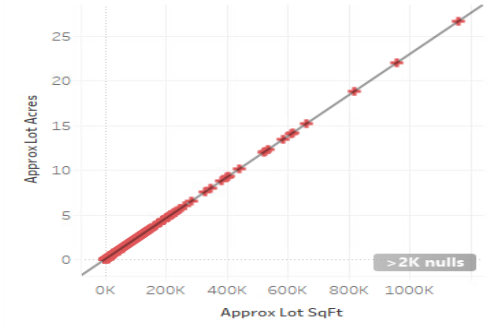


Figure B2: Approx Lot Acres and Approx Lot SqFt

3. The relationship between Approx_Lot_SqFt and Approx_SQFT is positive and has a moderate correlation with a correlation coefficient of 53.44%. [***Refer to Figure B1***].

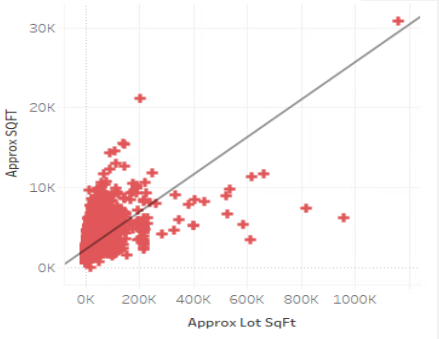


Figure B1: Approx SqFt and Approx Lot SqFt

4. The relationship between Bedrooms_Plus and __Bedrooms is positively correlated with a correlation coefficient of 82.87%. [***Refer to Figure B4].***

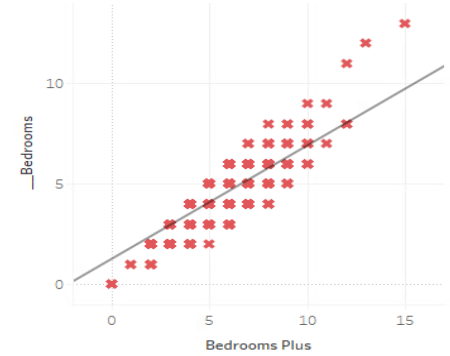


Figure B4: Bedroom vs Bedrooms Plus

5. The relationship between Exterior_Stories and __of_Interior_Levels is positively correlated with a correlation coefficient of 83.77%. [***Refer to Figure B3*]**.

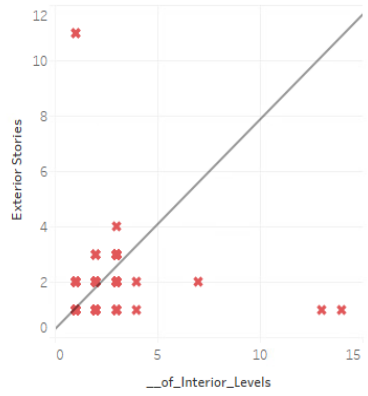


Figure B3: Exterior Stories vs Interior Levels

6. The relationship between __Bathrooms and __Bedrooms has a low correlation with a correlation coefficient of 15.20%. However, after removing the bathroom outlier, the correlation increased to 61.56%. [***Refer to Figure B5(a) and B5(b)*]**. The relationship, however, remains positive indicating that the more bedrooms a house has, the more bathroom it contains.

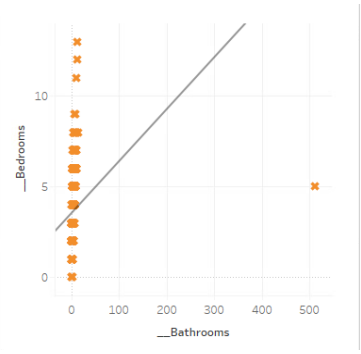


Figure B5(a): Bedrooms vs Bathrooms with outlier

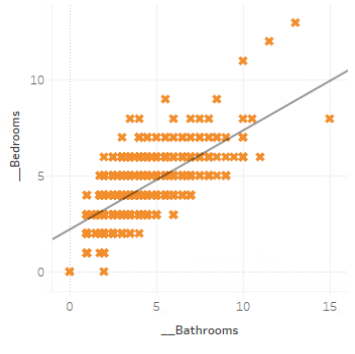


Figure B5(b): Bedrooms vs Bathrooms without outlier

7. Initial correlation between Price_SqFt and Original List price was 2.19% and between Price_SqFt and List price was 2.35% but after removing the outlier, correlation between Price_SqFt and Original List price is now 65.24% while the correlation between Price_SqFt and List price is 69.74%. [***Refer to Figure B8(a) and 85(b)***].

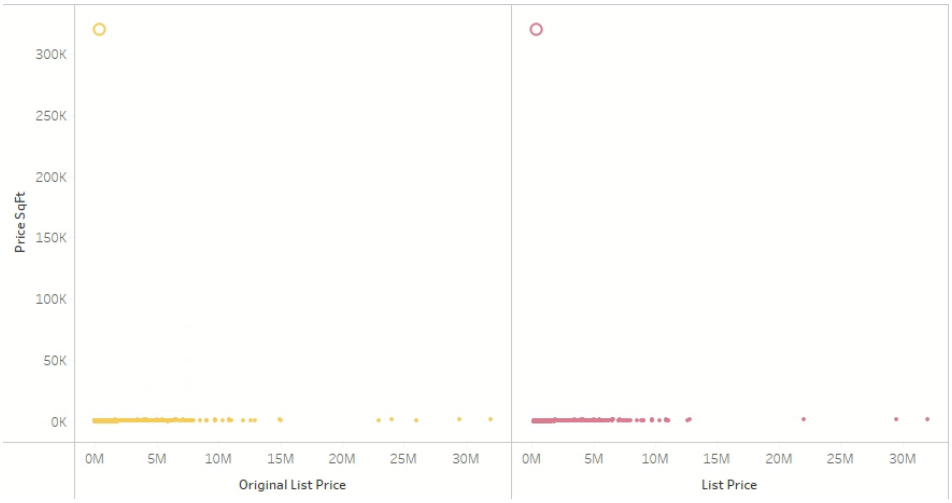


Figure B8(a): Price Square Ft vs List Price and Original List Price with outliers

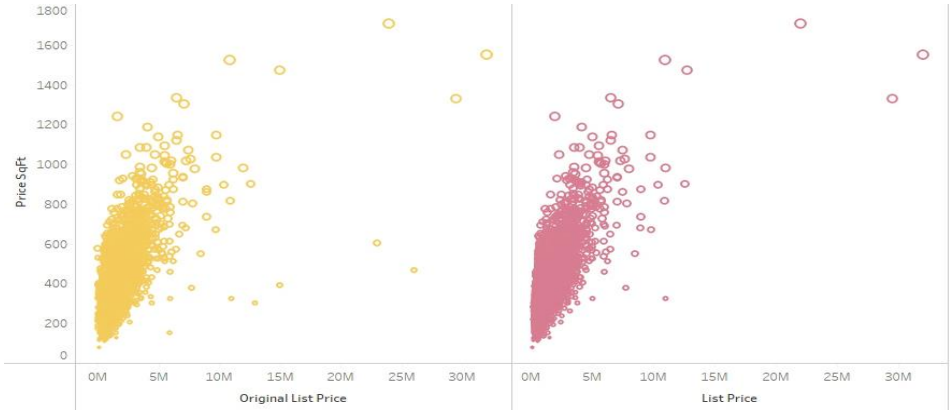


Figure B8(b): Price Square Ft vs List Price and Original List Price without outliers

8. The relationship between On_Market_Date, Off_Market_Date, and Days_on_Market is highly correlated with a correlation coefficient of 93.19%. Insight derived from this is that Days_on_Market appears to be a derived field from the first two, as indicated by the high correlation. The correlation is, however, not perfect due to the high number of missing observations for the On Market Date. [***Refer to Figure B7***]. We would be removing both Dates from our model and using the Days on market instead.

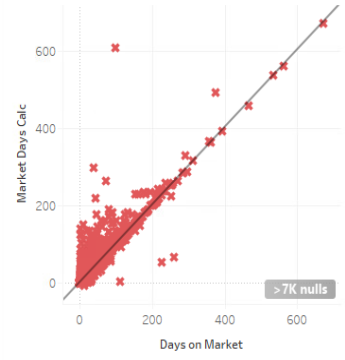


Figure B7: Days on Market vs Calc Market Days

Relationship between the target variable "Sold Price" and various predictor variables:

[***Table 2]*** shows that most of the predictor variables have a positive relationship with the target variable "Sold Price". The R-Squared value measures the proportion of variation in the target variable that can be explained by the predictor variable, and the correlation is a measure of the strength of the relationship between the two variables.

From the table, it can be concluded that the predictor variable "List Price" has the highest R-Squared (98.7%) and almost perfect correlation (99%) among all the predictor variables, meaning that it has the strongest relationship with the target variable "Sold Price". We can thus infer from this that the List Price seems to be the most important factor in determining the Sold Price of a property. The correlation between the Original List Proce and the Sold Price is also high at 94%. However, since it is almost perfectly correlated with the List Price, we would be removing it from the model. Other predictor variables with high R-squared and correlation with the target include "Approx SQFT", "Taxes", and "Guest House SqFt" with correlations of 78%, 51%, and 61% respectively. On the other hand, the predictor variable "Price SqFt" has a very weak correlation with "Sold Price" (2%).

PART 2: DATA PREPARATION

2.1: OUTLIERS

The result from our data exploration revealed a few extreme outliers in the dataset that might have been due to data entry errors. Two notable examples include the variable bathrooms which has the largest number of bathrooms in a house as 512 with the next largest number after this being 15 and the variable Price SqFt which highest price of $320,000 with the next highest price being $1,704. Due to these large gaps, we have reason to believe that these numbers are outliers caused by data entry errors and as a result we have decided to exclude them from the model. To achieve this, we have used the '*Filter*' node in SAS EM to set the upper limit of the bathroom observations to 20 and Price SqFt to $2,000. The exclusion of these two data points led to a significant improvement in the performance of our models as seen by the drastic reduction in the Root Average Squared Error (RASE). [***Refer to Figure D1(a) and D1(b)*]**.

2.2: MISSING RECORDS

Of the total 106 variables in the dataset, 35 of them contained missing observations. Variables which were observed to contain more than 50% missing data points were outrightly rejected from being used in the model. Twenty-five (25) of the thirty-six (36) variables identified above as including missing records fell into this category [**Refer to Table 3**]. Of the remaining eleven (11) variables which include missing records, seven of them were rejected and not used in the model due to them either containing too much free text that would not be relevant for our model building, being highly correlated with another variable, or simply being an identifier [**Refer to Table 4**].

Rejecting these variables left us with only three (3) variables, two class and one interval, with missing records that we would be using in the model. For the class variables – Compass and Payment Type – which had 1 and 716 missing records respectively, we imputed the missing records with the mode of the observation. For the interval variable – Comp_to_Buyer_Broker – which had 21 missing records, imputation was done with the mean.

2.3: ELIMINATION OF UNARY RECORDS

Six of the variables in the model were observed to have the same data for all the records and as such were eliminated from the model. These variables include Card_Format, Country, Dwelling_Type, Property_Type, State_Province, and Status.

We have also gone ahead to classify County_Code currently with binary records as unary. The County variable consists of Maricopa and Mohave; however, Mohave occurs only once across the twelve thousand plus recorded observations which gives us reason to believe that this might have been a data entry error. Thus, we have replaced it with Maricopa, classified County as unary and rejected it in the model.

2.4: COLLINEARITY

The data exploration performed in part 1 revealed that high degree of multicollinearity exists between some of the predictor variables. As a result, we have proceeded to exclude these variables from the model.

- List price and original list price were highly correlated with a correlation coefficient of 94.03% therefore original list price has been dropped from the model. The decision to drop original list price was due to the numerous data entry errors noted as opposed to List price which had less errors.
- Approx_Lot_Acres and Approx_Lot_SqFt are also highly correlated with a correlation coefficient of 99.99% therefore Approx_Lot_Acres has been dropped from the model. The decision to drop Approx_Lot_Acres was due to it containing missing data (16%) as opposed to to Approx_Lot_SqFt which had 0% missing data.
- Bedrooms_Plus and __Bedrooms are positively correlated with a correlation coefficient of 82.87% therefore __Bedrooms has been dropped from the model.
- Exterior_Stories and __of_Interior_Levels are positively correlated with a correlation coefficient of 83.77% therefore Exterior_Stories has been dropped from the model.
- On_Market_Date, Off_Market_Date, and Days_on_Market are highly correlated with a correlation coefficient of 93.19%. Insight derived from this is that Days_on_Market appears to be a derived field from the first two. As a result, we have dropped On_Market_Date and Off_Market_Date from the model due to missing data.

2.5: NORMALIZATION AND STANDARDIZATION

To ensure that our model worked best, especially for the KNN, we had to standardize the dataset using the transform node in SAS EM. After comparing the RASE values for both the 'range' and 'standardize' methods of transforming our variables, we found that 'range' returned higher RASE values than 'standardize' did. As a result, we chose to standardize our variables to improve the accuracy of our model.

We also tried to see if normalizing the data for all the interval variables would make a difference for the regression and decision tree models. Our findings on this led us to conclude that

normalizing the data had no impact on the model for both the regression and decision tree as there wasn't a change in the validation RASE. On the other hand, our decision tree models seemed to perform better without imputation/replacement and transformation of the data.

2.6: CATEGORY REDUCTION

From our in-depth analysis of the dataset in the exploratory phase, we discovered that the variable Builder_Name had a lot of naming inconsistencies that were referencing the same builder. For instance, the builder T.W. Lewis was recorded with four different formats as seen below. As a result, we had to combine those builder names that based on our judgement and understanding of the dataset represented the same builder. This led to a reduction in the number of levels for Builder_Name from 1819 to 1506.

Builder_Name
T W Lewis
T. W. Lewis
T.W. Lewis
TW Lewis

PART 3: DATA MODELLING AND EVALUATION

Now that our data is prepped and ready to be used in the model, we would be conducting different types of supervised data mining tasks to help us determine the best model to predict the sales price of the house. Our target variable as identified is the ***Sold_Price*** which is an interval variable. Therefore, we would be carrying out Linear regression, Decision Tree, and K Nearest Neighbor tasks to help with the prediction.

For all the datamining tasks, we have used partition of 60/40/0 training/validation/test and a random seed of 12345.

3.1: LINEAR REGRESSION

We created four different types – default, forward, backward, and stepwise – of linear regression on the data using different combinations of variables for each type. The result of each model is summarized in the table below:

Linear Regression: Target = Sold_Price

Software	Model	Settings	Variable Selection	Variables Used in Model	Adj R^2	Validation RASE
SAS EM	Linear Regression	Default	None	Approx_Lot_SqFt, Approx_SQFT, Assessor_s_Map__ , Assessor_s_Parcel__ , Bedrooms_Plus, Buyer_Concession, City_Town_Code, Days_on_Market, Elem_School_Dist__, High_School_Dist__, Horses, IMP_Comp_to_Buyer_Broker, Jr__High_School, List_Price, Loan_Type, Map_Code_Grid, Ownership, Pool, Price_SqFt, REP_Compass, REP_Payment_Type, REP_St_Suffix,	0.9902	102,141.5

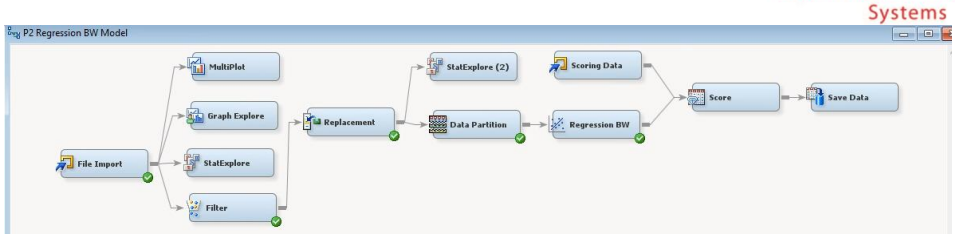
				Seller_Concession, Tax_Municipality, Tax_Year, Taxes, Type, Variable_Commission, Year_Built, __Bathrooms, __Bedrooms, __of_Interior_Levels, mod_timestamp		
SAS EM	Linear Regression	Default	Backward	Approx_Lot_SqFt, Approx_SQFT, Days_on_Market, Horses, List_Price, Loan_Type, Map_Code_Grid, Ownership, Price_SqFt, Year_Built, __Bathrooms, mod_timestamp	0.9903	90,617.6
SAS EM	Linear Regression	Default	Forward	Approx_Lot_SqFt, Approx_SQFT, Days_on_Market, Horses, List_Price, Loan_Type, Map_Code_Grid, Ownership, Price_SqFt,	0.9903	90,617.6

				Year_Built, __Bathrooms, mod_timestamp		
SAS EM	Linear Regression	Default	Stepwise	Approx_Lot_SqFt, Approx_SQFT, Days_on_Market, Horses, List_Price, Loan_Type, Map_Code_Grid, Ownership, Price_SqFt, Year_Built, __Bathrooms, mod_timestamp	0.9903	90,617.6
SAS EM	Linear Regression	Default	Backward, Forward	Approx_Lot_SqFt, Days_on_Market, Horses, List_Price, Pool, Price_SqFt, Year_Built, __Bathrooms, __Bedrooms	0.9919	51622.33
SAS EM	Linear Regression	Default	None	Approx_Lot_SqFt, Buyer_Broker, Days_on_Market, Dwelling_Styles, Elem_School_Dist__, Exterior_Stories, High_School, High_School_Dist__,	0.9919	70432.26

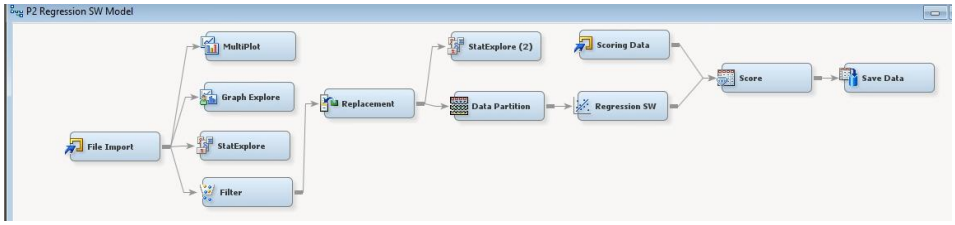
				Horses, Jr__High_School, List_Date, List_Price, Ownership, Pool, Price_SqFt, REP_Compass, Type, Year_Built, __Bathrooms, __Bedrooms, __of_Interior_Levels		
SAS EM	Linear Regression	Default	Stepwise	Approx_Lot_SqFt, Days_on_Market, High_School, Horses, List_Price, Pool,Price_SqFt, Type, Year_Built, __Bathrooms, __Bedrooms	0.9919	58015.35
SAS EM	Linear Regression	Default	Backward			

Commented [PB1]: Add new regression values heer

The best models for our regression analysis were for the Forward and Backward models with adjusted R squared of 99% meaning that 99% of the variability in the model can be explained by the following independent variables Approx_Lot_SqFt, Days_on_Market, Horses, List_Price, Pool, Price_SqFt, Year_Built, __Bathrooms, __Bedrooms. We chose these two as our best models from all the different combinations we carried out as they gave us the lowest validation Root Average Square Error (RASE) of $51,622.33.



Regression: Backward



Regression: Stepwise

3.2: DECISION TREE

We created different decision tree models using a combination of different parameters for creating the trees as well as the algorithm used to create the splitting rules. The result of each model is summarized in the table below:

Decision Tree: Target = Sold Price

Software	Model	Parameter Settings	Variables Used	Validation RASE
SAS EM	Decision Tree	Branches=2 Depth=6 Splitting: ProbF	List_Price, Price_SqFt, Days_on_Market	101,277
SAS EM	Decision Tree	Branches=2 Depth=11	Approx_SQFT, List_Price, Price_SqFt,	98,046.1

		Splitting: ProbF	Days_on_Market, Pool, Tax_Year	
SAS EM	Decision Tree	Branches=2 Depth=9 Splitting: Variance	List_Price, Approx_SQFT , Price_SqFt, Days_on_Market, Assessor_s_Map__, Pool, Year_Built, Payment_type, Tax_Year, Map_Code_Grid, Taxes, __Bathrooms, Assessor_s_Parcel__, Approx_Lot_SqFt, Comp_to_Buyer_Broker, Jr__High_School	96,475.6
SAS EM	Decision Tree	Branches=2 Depth=11 Splitting: Variance	List_Price, Approx_SQFT , Price_SqFt, Map_Code_Grid, Taxes, Tax_Year, Days_on_Market, Loan_Type, Pool, __Bathrooms, Year_Built, mod_timestamp, Assessor_s_Parcel__, Payment_Type, Elem_School_Dist__, Assessor_s_Map__, Comp_to_Buyer_Broker, Approx_Lot_SqFt,	95,958.2

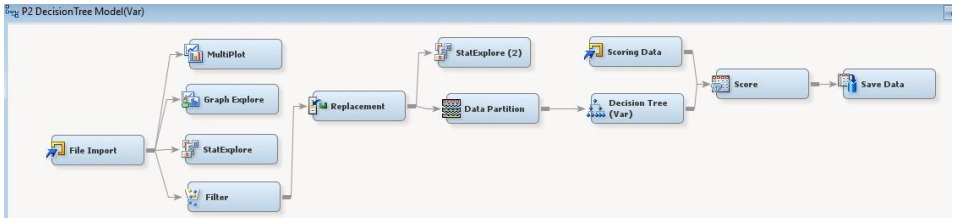
			Bedrooms_Plus, Jr__High_School, Compass, Tax_Municipality, City_Town_Code	
SAS EM	Decision Tree	Branches=2 Depth=6 Splitting: ProbF	Days_on_Market Price_SqFt List_Price	57,573.66
SAS EM	Decision Tree	Branches=2 Depth=6 Splitting: Variance	Days_on_Market Price_SqFt List_Price	57,276.62
SAS EM	Decision Tree	Branches=2 Depth=3 Splitting: ProbF	List_Price	102,310.5

For the first model highlighted, the ProbF splitting criterion was used, and the tree splits on Days_on_Market, Price_SqFt, and List_Price. The RASE on the validation data for this model was $57,573.66.

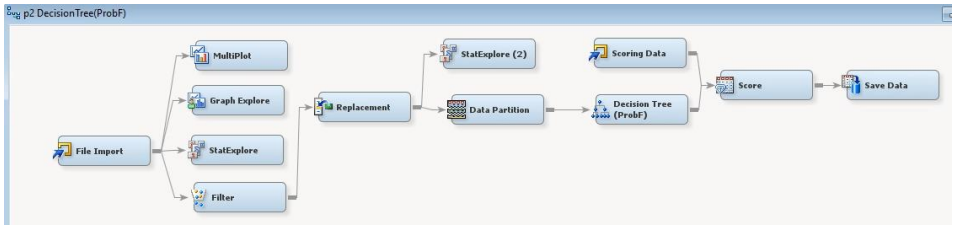
For the second model highlighted, the Variance splitting criterion was used, and the tree also splits on Days_on_Market, Price_SqFt, and List_Price. The RASE on the validation data for this model was $57,276.62.

Comparing the two decision tree models, we can see that they have similar validation RASEs: $57,573.66 for the model with the ProbF splitting criterion and 57,276.62 for the model with the Variance splitting criterion. This suggests that both models perform similarly in predicting Sold_price based on the three independent variables.

Days_on_Market, Price_SqFt, and List_Price are important predictors of Sold_price in the Arizona housing market. Days_on_Market reflects the property's desirability, Price_SqFt measures its size, and List_Price reveals the seller's expectations, all of which can influence Sold_prices.



Splitting: Variance



Splitting: ProbF

3.3: K-NEAREST NEIGHBOUR

While creating our kNN model, the main parameter we manipulated in trying to establish the best model was k (number of neighbors). The result of each model is summarized in the table below:

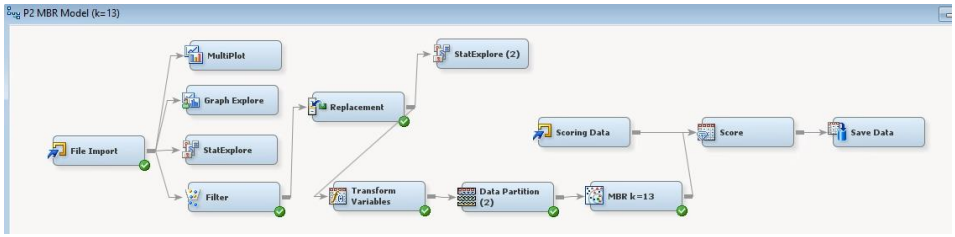
k-Nearest Neighbors: Target = Sold Price

Model	Settings	Variables Used in Model	Validation RASE ($)
kNN	K=16	STD_Approx_Lot_SqFt STD_Days_on_Market STD_Elem_School_Dist__ STD_Exterior_Stories	79573.11
kNN	K=15		79057.84
kNN	K=13		78512.33

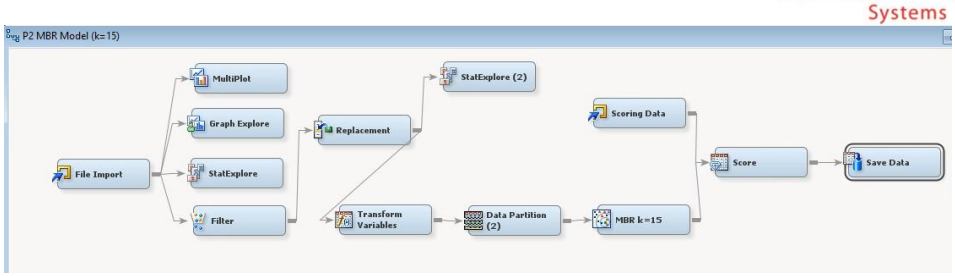
kNN	K=5	STD_High_School_Dist__ STD_List_Price STD_Price_SqFt STD_Year_Built STD___Bathrooms STD___Bedrooms STD___of_Interior_Level	79770.35

The kNN model with k=13 produced the best performance in predicting Sold_price in the Arizona housing market, as it had a lower validation RASE value compared to the kNN model with k=15. The reason for this is that the smaller k value reduced the impact of noise and outliers in the data, making the model more accurate.

Overall, we can infer from our models that the major determinants of the Sold_Price as seen to be consistent across all the six models tested are Days_on_Market, Price_SqFt, and List_Price.



kNN: k=13



kNN: k=15

3.4: FEATURE ENGINEERING

We performed feature engineering by creating several new variables as shown in the table below. Of all the newly created variables, the variable called TotalPrice, which represents the forecasted/estimated price of the property based on the product of the Approx_Sqft and Price_SqFt significantly improved the predictive power of the model. This was evidenced by the reduction in the root average squared error (RASE) values to a great extent, indicating a better fit between the predicted and actual Sold_Price values. The addition of this variable also provided better results in predicting the Sold_Price, suggesting that it captures additional information that was not present in the original variables.

TotalPrice (Approx_Sqft * Price_SqFt)

Changes were made to the roles of the variables tabulated below to simplify and enhance the predictive power of the model.

Old Variable		New Variable	
Name	**Role**	**Name**	**Role**
Guest_House_SqFt	Interval	New_Guest_House_SqF	Binary
Building_Number	Interval	Is_in_a_Building	Binary
Unit__	Interval	Has_Unit_Number	Binary
Other_Compensation	Interval	New_Other_Compensation	Binary
Co_Listing_Agent	Nominal	REP_ Co_Listing_Agent	Binary
Co_Selling_Agent	Nominal	REP_ Co_Selling_Agent	Binary

1. Guest_House_SqFt -> New_Guest_House_SqF:
   The original variable was an interval variable representing the square footage of a guest house on the property. The new variable is a binary variable that indicates whether the property has a guest house. This change is useful because it simplifies the variable, eliminates the missing data points, and makes it easier to use in a predictive model.

Additionally, the binary variable captures the information that the property has a guest house, which is an important feature for many real estate buyers.

2. Building_Number -> Is_in_a_Building:
   The original variable represented the number of the building that the property is located in. The new variable is a binary variable that indicates whether the property is in a building. This change simplifies the variable and captures the important information that the property is in a building or not. This information can be useful in predicting the price of the property, as properties in buildings may have different prices and values than those that are not.
3. Unit__ -> Has_Unit_Number:
   The original variable represented the unit number of the property. The new variable is a binary variable that indicates whether the property has a unit number. This change simplifies the variable and captures the important information that the property has a unit number or not. This information can be useful in predicting the price of the property, as properties with unit numbers may have different prices and values than those that do not.
4. Other_Compensation -> New_Other_Compensation:
   The original variable was an interval variable representing other compensation associated with the sale of the property. The new variable is a binary variable that indicates whether there is other compensation associated with the sale. This change simplifies the variable and captures the important information regarding if there is other compensation associated with the sale or not. This information can be useful in predicting the price of the property, as properties with additional compensation may have different prices and values than those that do not.
5. Co_Listing_Agent -> REP_ Co_Listing_Agent:
   The original variable was a nominal variable representing the name of the co-listing agent. The new variable is a binary variable that indicates whether there is a co-listing agent involved in the sale. This change simplifies the variable and captures the important information that there is or is not a co-listing agent involved in the sale. This information can be useful in predicting the price of the property, as properties with co-listing agents may have different prices and values than those that do not.
6. Co_Selling_Agent -> REP_ Co_Selling_Agent:
   The original variable was a nominal variable representing the name of the co-selling agent. The new variable is a binary variable that indicates whether there is a co-selling agent involved in the sale. This change simplifies the variable and captures the important information that there is or is not a co-selling agent involved in the sale. This information can be useful in predicting the price of the property, as properties with co-selling agents may have different prices and values than those that do not.

3.5: ADVANCED MODELS

Model	Settings/Parameters	Validation RASE
Sequential (Gradient Boosting + HP Regression)	Gradient Boosting: Default HP Regression: LAR	12.06
Sequential (Range NN + DT + Reg)	Decision Tree: Split = Var, Max Branch = 2, Max Depth = 11 Transform Inputs: Interval=Range, Class=Dummy Neural Network: Default Regression: Backwards	11.32
Ensemble (Clustering, DT, Regression + HP Reg)	Clustering: Standardization, Automatic Clusters Decision Tree: Default Regression: Backwards HP Regression: LAR	10.91
Sequential (Clustering + DT + Reg)	Clustering: Standardization, Automatic Clusters Decision Tree: Default Regression: Backwards	11.35

We used four advanced models for our data mining project, which are Sequential (Gradient Boosting + HP Regression), Sequential (Range NN + DT + Reg), Ensemble (Standardize Clustering, DT, Regression + HP Reg) and Sequential (Clustering + DT + Reg).

The first model we used is the Sequential (Gradient Boosting + HP Regression). Gradient boosting is a popular technique that combines multiple weak models to create a stronger prediction model. This model also uses hyperparameter regression (HP Regression) with the LAR algorithm, which is a powerful technique for feature selection. The model achieved a Sold_price prediction error of 12.06. The results suggest that the model might need some tuning to improve its accuracy.

The second model we used is the Sequential (Range NN + DT + Reg). This model uses a decision tree with a split criterion of variance, a maximum branch of 2, and a maximum depth of 11. The inputs were transformed using the range interval and dummy classification. The neural network used the default settings, and the regression method used was backwards. The model achieved a Sold_price prediction error of 11.32. The results suggest that the model has improved compared to the previous one, but it might still need some tuning.

The third model we used is the Ensemble (Standardize Clustering, DT, Regression + HP Reg). This model uses a clustering technique that standardizes the data and automatically determines the clusters. It also uses a decision tree with default settings and backward regression. The clustering, decision tree and regression models were combined sequentially and then ensembled with the hyperparameter (HP) regression with the LAR algorithm. The model achieved a Sold_price prediction error of 10.91, which is the lowest among the three models. The results suggest that this model is the most accurate and reliable one for predicting Sold_price.

The fourth model is a variation of the third model, but without creating an ensemble with the hyperparameter (HP) regression. The model is a Sequential model that uses standardization clustering, a decision tree with default settings, and backward regression. The model achieved a Sold_price prediction error of 11.35, which is higher than the third model's error of 10.91 but lower than the errors of the first two models. The results suggest that this model is relatively accurate, but not as accurate as the third model.

In conclusion, the Ensemble model has a more robust performance compared to the other two models. Its accuracy is higher, which suggests that it is more reliable in predicting Sold_price.

3.6: MODEL COMPARISON: SAS EM vs KAGGLE

Model	Settings/Parameters	SAS EM Validation RASE	Kaggle Validation RASE
Linear Regression	Backward	11.37	11.00
Sequential (Gradient Boosting + HP Regression)	Gradient Boosting: Default HP Regression: LAR	12.06	11.43
Sequential (Range NN + DT + Reg)	Decision Tree: Split = Var, Max Branch = 2, Max Depth = 11 Transform Inputs: Interval=Range, Class=Dummy Neural Network: Default Regression: Backwards	11.32	11.25
Ensemble (Standardize Clustering, DT, Regression + HP Reg)	Clustering: Standardization, Automatic Clusters Decision Tree: Default	10.91	11.04

	Regression: Backwards HP Regression: LAR		
Sequential (Clustering + DT + Reg)	Clustering: Standardization, Automatic Clusters Decision Tree: Default Regression: Backwards	11.35	11.49

PART 4: DEPLOYMENT

4.1: OUTCOME

The data set used for this project is related to real estate sales in Arizona. In the data preparation phase, we have explored the data and cleaned it by removing missing values and unnecessary variables. The data has been split into training and validation sets to evaluate the models.

We developed multiple models in the previous phases of the project, including Linear Regression, kNN, and decision trees. The best model chosen is Linear Regression with Backward and Forward feature selection, which achieved a validation RASE of 51622.33 on the SAS EM validation set.

During the advanced modeling phase, multiple models were developed, and the best four were identified. These models were evaluated based on their ability to predict with accuracy and provide value to the business. The top models were chosen based on their SAS EM validation RASE and Kaggle validation RASE scores. The Ensemble model achieved the best results with a SAS EM validation RASE of 10.91 and a Kaggle validation RASE of 11.04. This model used Standardization Clustering, a Decision Tree with default settings, Backwards Regression, and LAR HP Regression. The other models, including Sequential models with Gradient Boosting and HP Regression and a Range NN with DT and Reg, also provided value with RASE scores ranging from 11.25 to 12.06, but the Ensemble model stood out as the best performer.

Based on the models developed and evaluated, we would recommend that real estate companies and property investors can use these models to accurately predict the sale prices of houses in Arizona. They can use these models to make informed decisions about buying and selling properties, as well as to assess the value of their current real estate holdings. The models could also be used by individuals looking to sell their property to get an idea of its market value.

It is recommended to use these models in conjunction with other factors such as property condition, location, and amenities, as well as market trends and economic indicators to make informed decisions.


UNIVERSITY OF
ARKANSAS
SAM M. WALTON
COLLEGE OF BUSINESS
Department of Information
Systems

APPENDIX

TABLES

Table 1: Data Dictionary

Variable Name	*Variable Type*	*Variable Description*	*Example Values*
dataobs	id		
Status		Status of house	C
Compass			N, E, W
Country	Nominal	Country where house is located	USA
Subagents	Binary	If the house has subagents	N, Y
Pool	Nominal	The type of pool the house has	Private, Community, None
Taxes	Interval	Amount of tax paid	1213, 2362, 616
Legal			Blank
Directions	Text	Directions for locating the house	From Dobson and Queen Creek, head north on Dobson. East on Mockingbird.
mod_timestamp			0.037166667, 0.004025463
Auction	Binary	States if a house is up for auction	N, Y
Horses	Binary	States if a house has horses	N, Y
Ownership	Binary	The type of house ownership	Fee Simple, Condominium
Type	Binary	Used to indicate the type of listing. Exclusive right-to-sell (ER) and exclusive agency (EA) listing	EA, ER

Model			101, 740, Coralina, Sereno
Subdivision			COLLEGE PARK 14, TWELVE OAKS 4 LOT 601-844 TR A-E
List_Number	ID	unique ID number assigned to the property when listed	6149617, 6260501
Agency_Name	Nominal	Name of agency representing the seller	Realty ONE Group (reog03), Daniel D. Smith & Associates, LLC (ddsm01)
Agency_Phone	Nominal	Phone number of agency representing the seller	(480) 967-1333
Listing_Agent	Nominal	Name of agent who listed the house representing the seller	Renee' Merritt
Co_Listing_Agent	Nominal	Name of agent who co-listed the house	Michael D Smith
Property_Type	Binary	Indicates the type of property	Residential
Card_Format	Nominal		Residential
Selling_Agency	Nominal	Name of selling agency representing the buyer	Daniel D. Smith & Associates, LLC (ddsm01)
Selling_Agent	Nominal	Name of selling agent representing the buyer	Sharma Glenn
Co_Selling_Agent	Nominal	Name of co-selling agent	Jefferey Salazar
End_Date	Date		12/10/1961, 5/4/1962
Dwelling_Type	Nominal	Type of dwelling	Single Family - Detached
List_Date	Date	Date house was listed	5/22/1961, 10/16/1960

Close_of_Escrow_Date	Date	Date both buyer and seller meet the conditions in the homebuying contract. This is when the funds are dispersed to the seller and the buyer officially has the home title in their name.	3/3/1961, 8/8/1961
Under_Contract_Date	Date	Date the buyer and the seller agree on the offer	7/11/1961, 10/16/1960
Fallthrough_Date	Date	Date the agreed deal between the buyer and seller collapses	7/22/1961
Status_Change_Date	Date	The date the status of the property is changed.	8/5/1961
Temp_Off_Market_Date	Date	Date the property listing is temporarily taken off the market	Blank
Cancel_Date	Date	Date when there is no longer an active brokerage agreement/listing contract with the seller.	Blank
UCB_or_CCBS	Binary	UCB (Under Contract Accepting Back-Up Offers) – This means an offer has been accepted, however the seller has chosen to accept back-up offers. CCBS (Contract Contingent on Buyer Sale) – This status is reserved for when the buyer's	Blank

		offer to purchase the property is contingent on them first selling their own property	
Original_List_Price	Interval	Original listing price of property in $	265,000, 469,000
List_Price	Interval	Current listing price of property in $	265,000, 469,000
Sold_Price	Interval	Sale price of property in $	265,000, 469,000
Price_SqFt	Interval	Price per square feet of the property in $	304.42, 250.84
Map_Code_Grid	Nominal		S38, V39
House_Number	Interval	Number assigned to the property on a street	98, 1934
Building_Number	Interval	Unique number given to each building on a street	174, 5873
Street_Name	Nominal	identifying name given to a street where property is located	Apollo, Riviera
Unit__	Interval	Unit number of the property within a building	1, 8
St_Dir_Sfx	Ordinal	Indicates the direction of the street location	NE, W, S
St_Suffix	Nominal	Part of a street name that denotes its type of roadway	St., Dr., Ave.
VAR34			%, $
City_Town_Code	Nominal	City in which the property is located	Chandler, Tempe

State_Province	Nominal	State in which the property is located	AZ
County_Code	Nominal	County in which the property is located	Maricopa, Mohave
Geo_Lat	Interval	Latitude coordinates of property	33.33673
Geo_Lon	Interval	Longitude coordinates of the property	-111.822666, -111.889725
Approx_SQFT	Interval	The approximate square footage of a property.	1863,1196
Bedrooms_Plus	Interval	Additional rooms in a property that are used as bedrooms.	4,3
Year_Built	Date	The year a property was built.	1995,1983
VAR45	Nominal (Continuous)		%,%
Buyer_Broker	Binary	Indicate if the buyer of a property was ed by a broker.	Y, Y
Variable_Commission	Binary	Refer to the commission paid to a real estate agent that is not a fixed amount.	Y, Y
__Bedrooms	Interval	The number of bedrooms in a property.	4,3
__Bathrooms	Interval	The number of bathrooms in a property.	2.5,1.75
__of_Interior_Levels	Nominal	The number of interior levels in a property.	2,2

Exterior_Stories	Nominal	The number of stories in a property's exterior.	2,2
Source_Apx_Lot_SqFt	Interval	Indicate the source of the estimated square footage of a property's lot.	T, T
Tax_Year	Interval	The year for which a property's taxes were assessed.	2020, 2020
Legal_Description__Abbrev_	Nominal	Abbreviated form of the legal description of a property.	LOT 541 SPRINGS UNIT THREE MCR 037415, LOT 42 COVE AT TIBURON MCR 024926
Public_Remarks	Nominal	Contains public comments or remarks about a property.	This Chandler two-story cul-de-sac home offers a patio, and a two-car garage.,Beautiful 3 bed 2bathrooms located in very clean area, new interior paint,
Private_Rmks___DND2	Nominal	Contains private remarks about a property that should not be disclosed to the public.	For financing options and to qualify for 2% back (details in listing documents), please see www.opendoorhomeloans.com.
Assessor_Number	ID	The assessor's identification number for a property.	303-69-692,302-80-524

Ownr_Occ_Name___DND2	Nominal	Name of the owner of a property, but is marked "DND2" which suggests that it should not be disclosed.	OPENDOOR PROPERTY TRUST I, MIKE & BEATRIZ TOVAR
Owner_Occ_Phn___DND2	Nominal	Phone number of the owner of a property, but is marked "DND2" which suggests that it should not be disclosed.	-
Marketing_Name	Nominal	Marketing name of a property.	Lake Community
Builder_Name	Nominal	Name of the builder of a property.	Hacienda,Coventry Homes
Assessor_Parcel_Ltr	Nominal	Parcel letter assigned to a property by the assessor.	-
Out_of_Area_Schl_Dst	Binary	Indicate if a property is located outside of the area served by a school district.	-
Source_of_Sq Ft	Nominal	Indicate the source of the square footage information for a property.	T, O
Tax_Municipality	Nominal	The municipality responsible for collecting taxes on a property.	Chandler, Maricopa
Hundred_Block	Interval	The 100-block number associated with a property's address.	600, 300

Elementary_School	Nominal	Name of the elementary school associated with a property.	Rudy G. Bologna, Challenger
Jr__High_School	Nominal	Name of the junior high school associated with a property.	Willis, Chandler High
High_School	Nominal	Name of the high school associated with a property.	Perry, Chandler High
Elem_School_Dist__	Nominal	The school district responsible for an elementary school associated with a property.	80, 80
High_School_Dist__	Nominal	The school district responsible for a high school associated with a property	80, 80
Hndrd_Blk_Directionl	Nominal	A direction indicator for the hundred block of the property's street address.	N, E, S, W
Week_Avail_Timeshare	Nominal	Availability of the property as a timeshare property on a weekly basis.	0
Comp_to_Sub agent	Interval	Commission paid to a sub-agent involved in the sale of the property.	2.5
Comp_to_Buyer_Broker	Interval	Commission paid to the buyer's broker.	2.25, 2.25

Other_Compensation	Interval	Any additional compensation received in the sale of the property.	0
Guest_House_SqFt	Interval	The square footage of any guest houses on the property.	0
Approx_Lot_SqFt	Interval	The approximate square footage of the lot on which the property is located.	5358, 4434
Assessor_s_Book__	Nominal	The book number assigned by the assessor for property tax purposes.	303, 302
Assessor_s_Map__	Nominal	The map number assigned by the assessor for property tax purposes.	69, 80
Assessor_s_Parcel__	Nominal	The parcel number assigned by the assessor for property tax purposes.	692, 524
Off_Market_Date	Date	The date the property was taken off the market.	22481, 22310(format)
Cross_Street	Nominal	The street that intersects with the street on which the property is located.	101 & Warner
Dwelling_Styles	Binary	The style of dwelling, such as a single-family home, townhome, or condominium.	Attached, Detached

Flood_Zone	Nominal	Is property located in flood zone?	No, yes, TBD
Approx_Lot_Acres	Interval	The approximate size of the lot in acres.	0.123, 0.102
On_Market_Date	Date	The date the property was put on the market for sale.	22484
Lead_Based_Hazard_Disclosure	Binary	A disclosure indicating whether the property may contain lead-based paint hazards.	-
Days_on_Market	Interval	The number of days the property has been on the market.	12,6
Loan_Type	Nominal	The type of loan used to purchase the property.	Cash, FHA
Payment_Type	Nominal	The type of payment, such as cash	Fixed, Fixed
Buyer_Concession	Interval	Any concessions made by the buyer, such as closing cost assistance.	$, %
Seller_Concession	Interval	Any concessions made by the seller, such as offering a home warranty.	$, %
Features		Other features of the property	-

Table 2: Relationship between Target variable and other numerical predictor variables.

Target	Predictor	Relationship	Correlation
Sold_Price	Days_on_Market	Positive	23%
Sold_Price	__Bathrooms	Positive	17%
Sold_Price	__Bedrooms	Positive	38%
Sold_Price	__of_Interior_Levels	Positive	11%
Sold_Price	Approx_Lot_Acres	Positive	48%
Sold_Price	Approx_SQFT	Positive	78%
Sold_Price	Bedrooms_Plus	Positive	48%
Sold_Price	Days_on_Market	Positive	23%
Sold_Price	Exterior_Stories	Positive	9%
Sold_Price	Guest_House_SqFt	Positive	61%
Sold_Price	List_Price	Positive	99%
Sold_Price	Original_List_Price	Positive	94%
Sold_Price	Price_SqFt	Positive	2%
Sold_Price	Taxes	Positive	51%
Sold_Price	High_School_Dist	Negative	10%
Sold_Price	Elem_School_Dist	Positive	14%

Table 3: Variables with More Than 50% of Missing Data

Variable	Amount Missing	Missing (%)	Role
Building_Number	12,267	100%	Rejected
Cancel_Date	12,297	100%	Rejected
Co_Listing_Agent	8,289	67%	Rejected
Co_Selling_Agent	10,335	84%	Rejected
Comp_to_Subagent	12,239	100%	Rejected
End_Date	12,294	100%	Rejected
Fallthrough_Date	12,041	98%	Rejected

Flood_Zone	10,173	83%	Rejected
Guest_House_SqFt	11,749	96%	Rejected
Hndrd_Blk_Directionl	8,945	73%	Rejected
Hundred_Block	9,069	74%	Rejected
Lead_Based_Hazard_Disclosure	12,297	100%	Rejected
Legal	12,297	100%	Rejected
Marketing_Name	9,260	75%	Rejected
Model	10,021	81%	Rejected
On_Market_Date	6,841	56%	Rejected
Other_Compensation	12,116	99%	Rejected
Out_of_Area_Schl_Dst	12,283	100%	Rejected
Owner_Occ_Phn___DND2	11,400	93%	Rejected
St_Dir_Sfx	12,191	99%	Rejected
Temp_Off_Market_Date	12,297	100%	Rejected
UCB_or_CCBS	12,297	100%	Rejected
Unit__	12,005	98%	Rejected
VAR34	12,239	100%	Rejected
Week_Avail_Timeshare	12,006	98%	Rejected

Table 4: Variables with Less Than 50% of Missing Data but Rejected in Model

Variable	Missing	Missing (%)	Reason for rejection
Agency_Phone	21	0%	Not useful for predicting target
Cross_Street	152	1%	Contains too much free text and is not meaningful for our model
House_Number	7	0%	ID number of houses. Not useful for predicting target

Legal_Description__Abbrev_	216	2%	Contains too much free text and is thus not meaningful for our model
Ownr_Occ_Name___DND2	2	0%	Not useful for predicting target
VAR45	21	0%	Not useful for predicting target
Approx_Lot_Acres	1,925	16%	High correlation with Approx_Lot_SqFt

Table 5: Variables used for Modelling.

VARIABLE	ROLE
__Bathrooms	Input
__Bedrooms	Input
__of_Interior_Levels	Input
dataobs	Id
Approx_Lot_SqFt	Input
Approx_SQFT	Input
Assessor_s_Map__	Input
Assessor_s_Parcel__	Input
Auction	Input
Bedrooms_Plus	Input
Buyer_Concession	Input
City_Town_Code	Input
Comp_to_Buyer_Broker	Input
Compass	Input
Days_on_Market	Input
Elem_School_Dist__	Input
Exterior_Stories	Input

High_School	Input
High_School_Dist__	Input
Horses	Input
Jr__High_School	Input
List_Price	Input
Loan_Type	Input
Map_Code_Grid	Input
mod_timestamp	Input
Ownership	Input
Payment_Type	Input
Pool	Input
Price_SqFt	Input
Seller_Concession	Input
Sold_Price	Target
St_Suffix	Input
Subagents	Input
Tax_Municipality	Input
Tax_Year	Input
Taxes	Input
Type	Input
Variable_Commission	Input
Year_Built	Input

Table 6: Best Advanced Model Parameters.

				Cluster + DT	Cluster + DT + Reg	HP Reg	Ensemble
	Data Transformation (Variable creation)	**Missing Values**	**Outliers**	**Valid RASE**	**Valid RASE**	**Valid RASE**	**Valid RASE**
Model 1	Yes	Yes	Yes	36,364	11.35		
Model 2	Yes	No	No	257,654	11.62		
Model 3	Yes	No	Yes	36,364	11.35		10.91
Model 4	Yes	No	No			10.87	

CHARTS AND VISUALIZATIONS

A: Histogram & Box Plots Showing Distribution of Numerical Variables

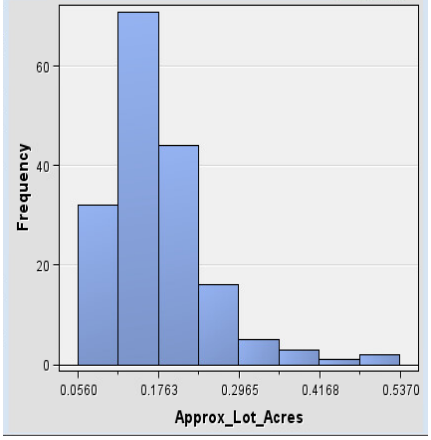


Figure A1: Distribution of Approx Lot in Acres

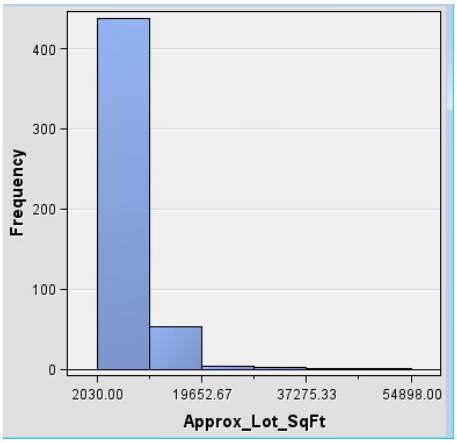


Figure A2: Distribution of Approx Lot in Square Foot

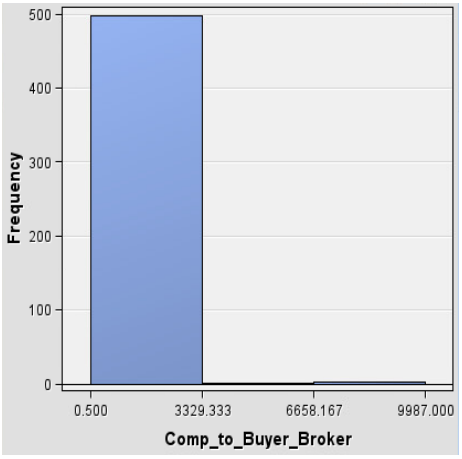


Figure A3: Distribution of Comp to Buyer Broker

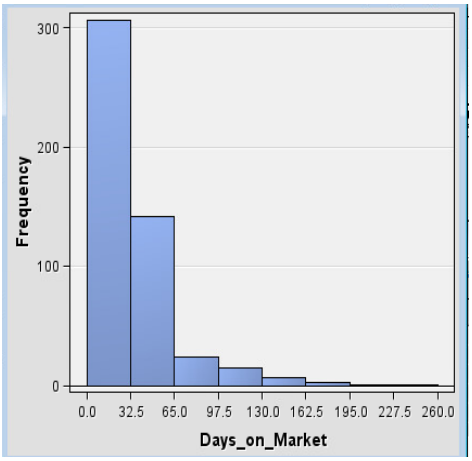


Figure A4: Distribution of Days on Market

Figure A5(a): Distribution of List Price

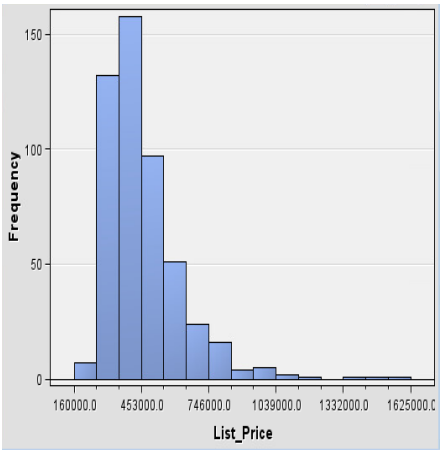
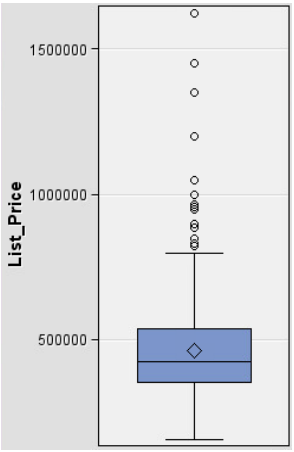


Figure A5(b): Box Plot of List Price



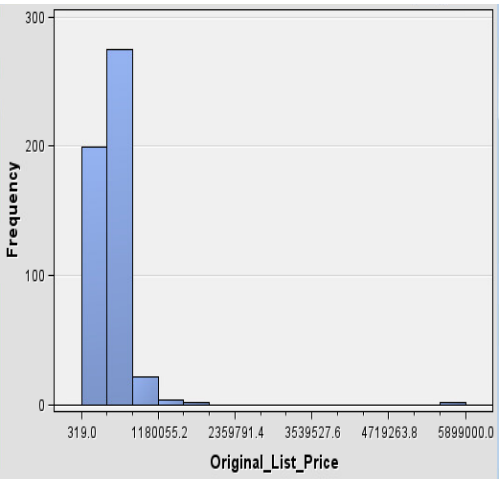


Figure A6(a): Distribution of Original List Price

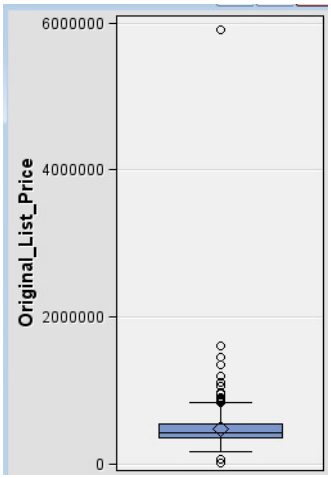


Figure A6(b): Box Plot of Original List Price

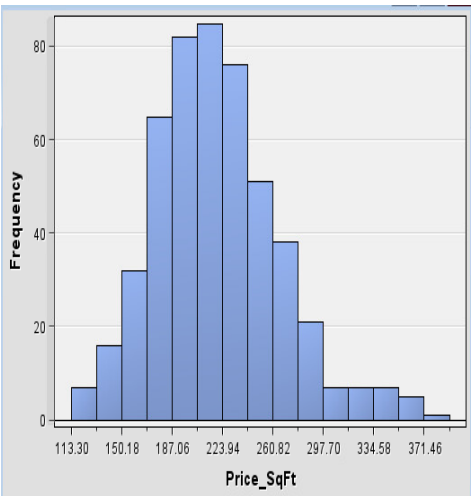


Figure A7(a): Distribution of Price Square Ft

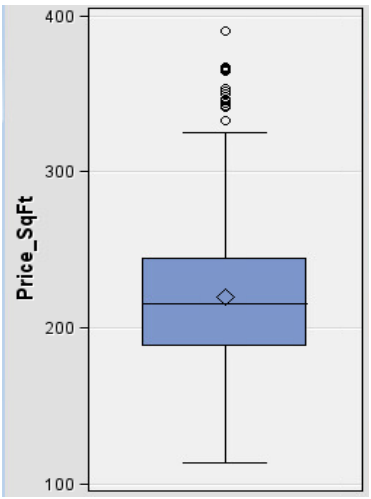


Figure A7(b): Box Plot of Price Square Ft

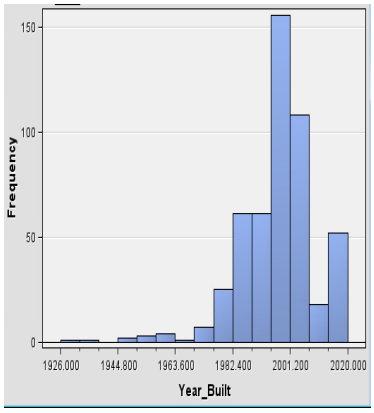


Figure A8: Distribution of Year Built

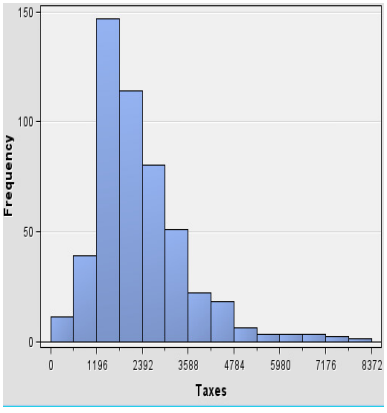


Figure A9: Distribution of Taxes

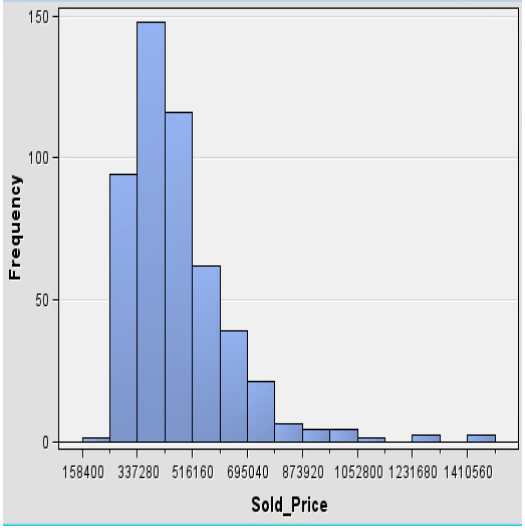
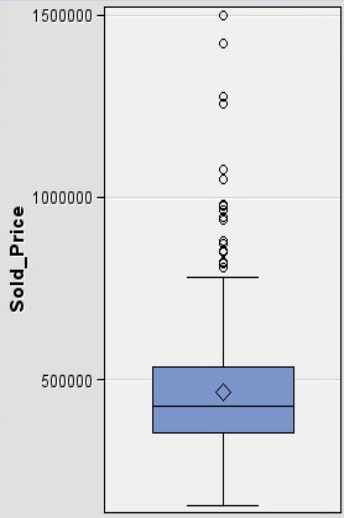


Figure A10(a): Distribution of Sold Price

Figure A10(b): Box Plot of Sold Price



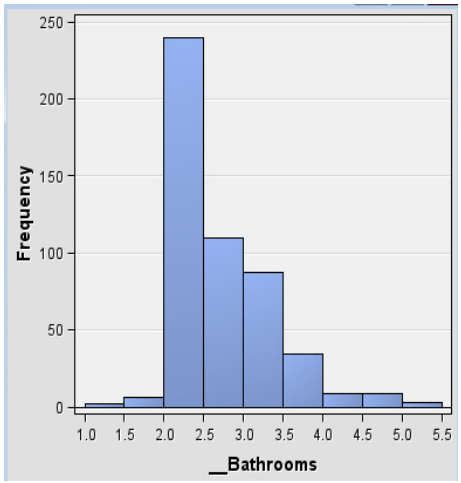


Figure A11(b): Box Plot of Bathrooms

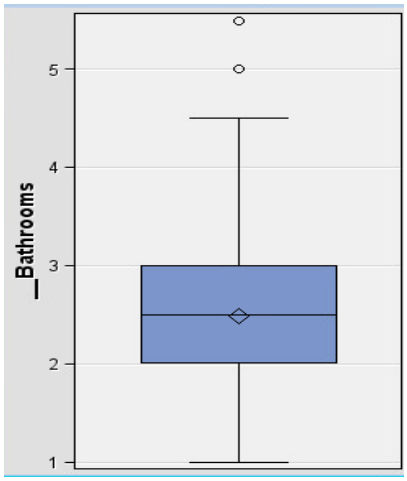


Figure A11(a): Distribution of Bathrooms

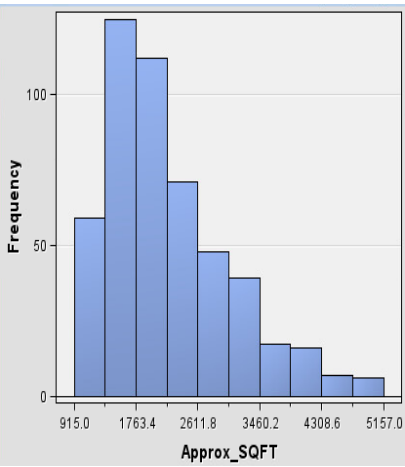


Figure A12: Distribution of Approx Square Ft

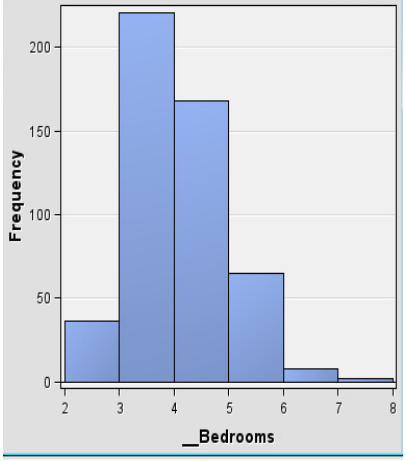


Figure A13: Distribution of Bedrooms

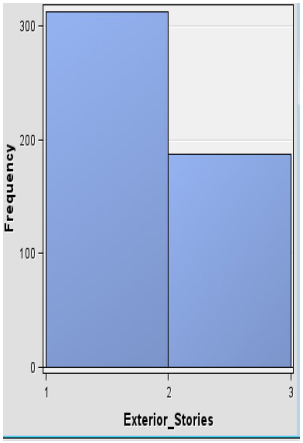


Figure A14: Distribution of Exterior Stories

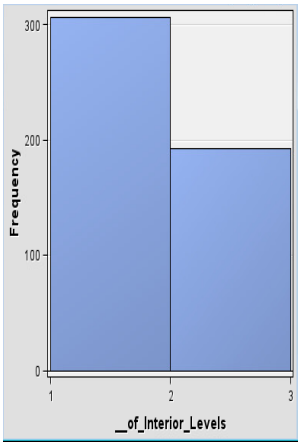


Figure A15: Distribution of Interior Levels

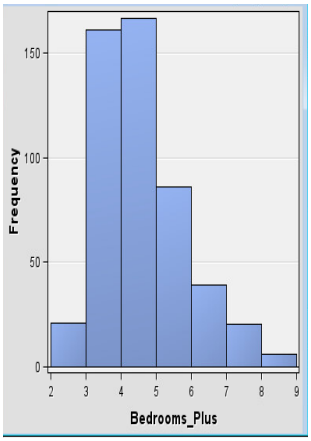


Figure A16: Distribution of Bedroom Plus

B: Charts Showing Relationship Between Target & Categorical Variables

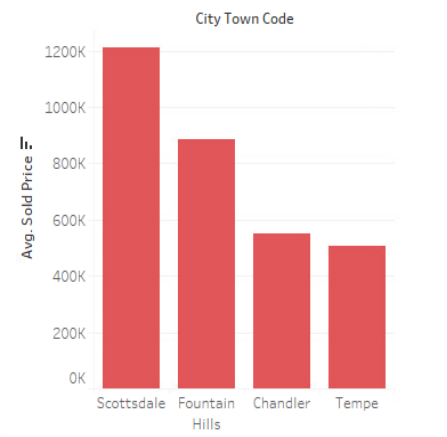


Figure C1: Avg Sold Price by City

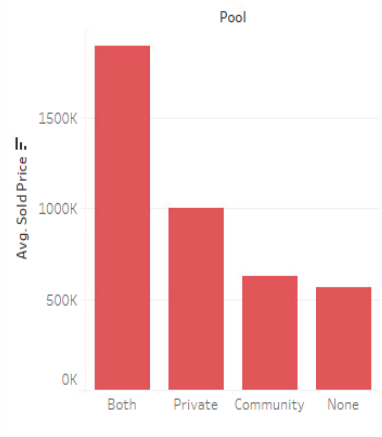


Figure C2: Avg Sold Price by Pool Type

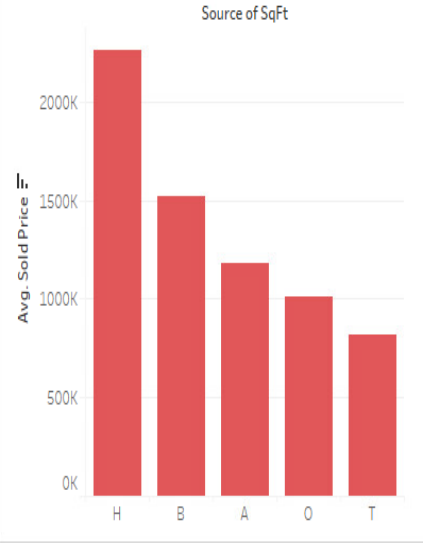


Figure C3: Avg Sold Price by Source of Square Ft

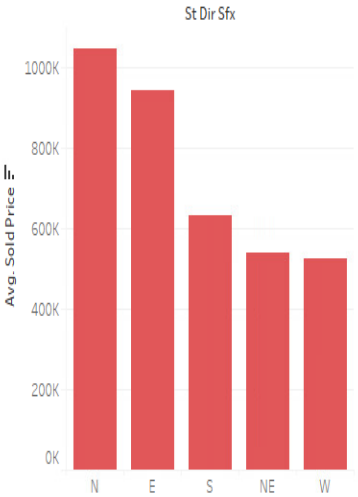


Figure C4: Avg Sold Price by St Direction Suffix

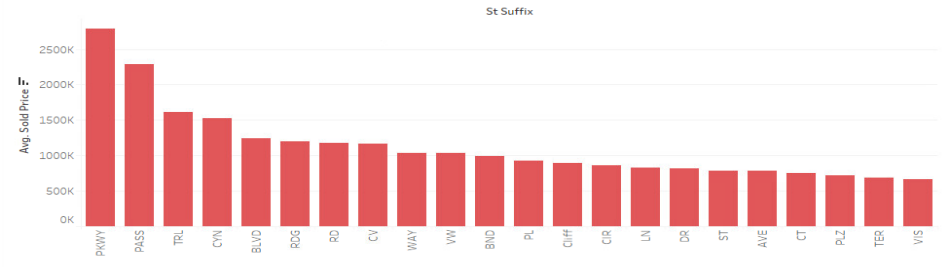


Figure C5: Avg Sold Price by St Suffix

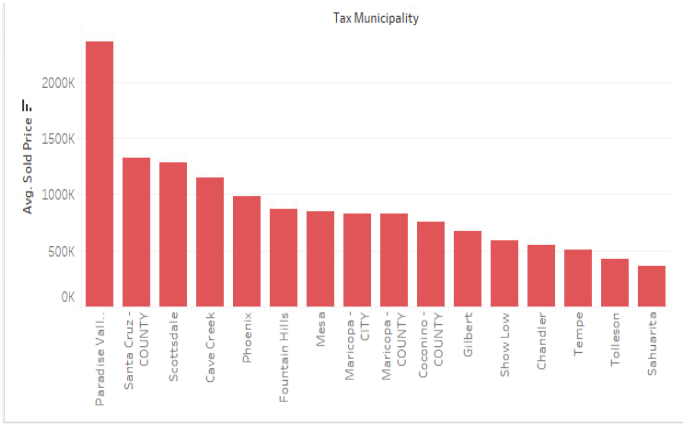


Figure C6: Avg Sold Price by Tax Municipality

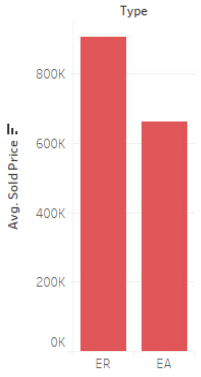


Figure C7: Avg Sold Price by Listing Type

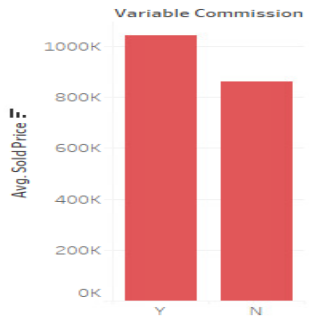


Figure C8: Avg Sold Price by Var Commission

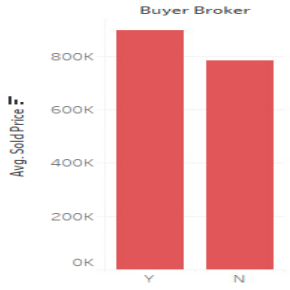


Figure C9: Avg Sold Price by Buyer Broker

Figure C10: Avg Sold Price by Loan Type

Figure C11: Avg Sold Price by Payment Type

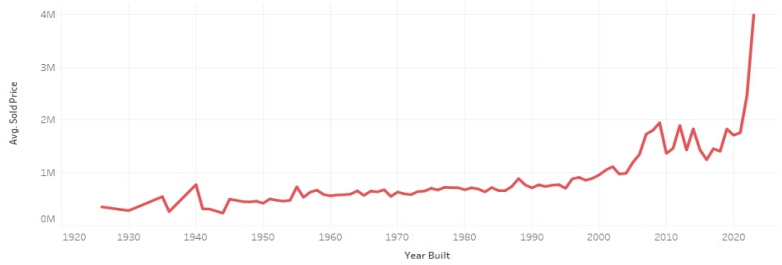


Figure C12: Avg Sold Price by Year Built

C: Model Building Snapshots

Selected Model	Model Description	Valid: Root Average Squared Error	Target Variable
Y	Regression - Fwrd	93892.22	Sold_Price
	Regression - SW	93892.22	Sold_Price
	Decision Tree B:2, D:11, S:Var	96087.96	Sold_Price
	Decision Tree B:2, D:9, S:Var	96516.33	Sold_Price
	Decision Tree B:2, D:10, S:ProbF	98046.12	Sold_Price
	Decision Tree B:2, D:11, S:ProbF	98046.12	Sold_Price
	Decision Tree B:2, D:7, S:ProbF	98283.86	Sold_Price
	Decision Tree B: 2, D: 6, S:ProbF	101277	Sold_Price
	Regression - Bkwrd	101294.4	Sold_Price
	Decision Tree B:2, D: 4, S: ProbF	128023.7	Sold_Price
	Regression - Default	137652.1	Sold_Price
	MBR k = 5	182350.7	Sold_Price
	Decision Tree B:2, D:3, S:ProbF	197004.8	Sold_Price
	MBR k = 11	197752.3	Sold_Price
	MBR k = 13	198371.4	Sold_Price
	MBR k = 16	200643.4	Sold_Price

Figure D1(a): Model Results before removing outliers.

Selected Model	Model Description	Valid: Root Average Squared Error	Target Variable
Y	Decision Tree B:2, D:11, S:Var	51660.25	Sold_Price
	Decision Tree B:2, D:9, S:Var	51906.66	Sold_Price
	Decision Tree B:2, D:10, S:ProbF	53510.26	Sold_Price
	Decision Tree B:2, D:11, S:ProbF	53510.26	Sold_Price
	Decision Tree B:2, D:7, S:ProbF	53739.39	Sold_Price
	Decision Tree B: 2, D: 6, S:ProbF	54827.92	Sold_Price
	Regression - Bkwrd	58280.68	Sold_Price
	Decision Tree B:2, D: 4, S: ProbF	65607.31	Sold_Price
	Regression - SW	69087.36	Sold_Price
	Regression - Fwrd	69092.06	Sold_Price
	MBR k = 11	75532.39	Sold_Price
	MBR k = 13	76287.43	Sold_Price
	MBR k = 16	76927.82	Sold_Price
	MBR k = 5	79190.58	Sold_Price
	Regression - Default	82802.3	Sold_Price
	Decision Tree B:2, D:3, S:ProbF	95723.52	Sold_Price

Figure D1(b): Model Results after removing outliers.

D: Model Flow Diagrams

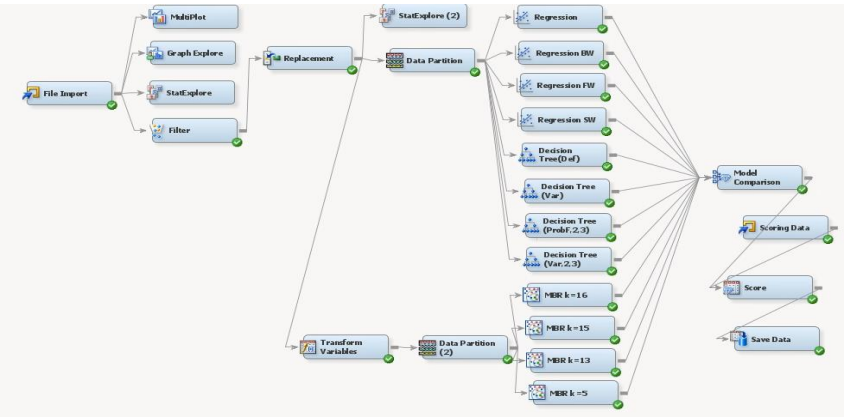


Figure E1: Process Flow 1

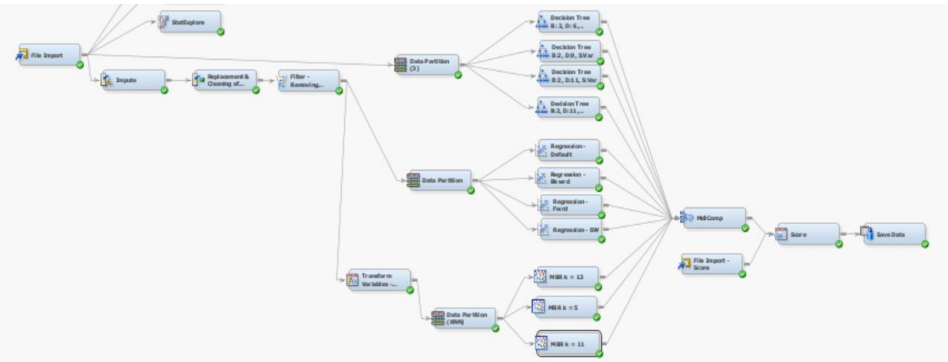


Figure E2: Process Flow 2