OPIM 5604: PREDICTIVE MODELING

King County House Pricing

Team 7

Submitted by:

Abhinav Suggula

Kunja Dutta

Jilei Dong

Zhuo Li

Table of Contents

1 Executive Summary

Problem Statement:

King County Housing Authority(KCHA) needed to analyze how to improve the experience of the buyer and the seller. They required a model to improve access of buyers to quality houses which would fulfill their requirements and suit their pockets, to make it easier for the sellers to get appropriate price and to train the real estate agents accordingly. Reaching a content buyer seller experience was the problem King County Housing Authority was facing.

Approach:

We downloaded the dataset from Kaggle and created a data dictionary for all the variables with their descriptions and types. We used SEMMA approach to analyze data. First, we extracted 25% of the total data set to concentrate on a reasonable sample size to achieve better results. Second, we performed data exploration to understand the basic patterns in the data and then identified the target variable for modelling which formed the basis for further data analysis. Third, we fine-tuned the data set with some modifications to handle the data better in terms of correlation and outliers. Fourth, we built multiple predictive models to meet KCHA mission. Finally, we evaluated all the models to decide upon the best model and made suitable recommendation to KCHA based on the business insights gathered.

Results:

- Identified “Price” as the target variable. This variable is of utmost importance to both buyer and seller.
- Initial data set had 21613 rows from which 5403 (25%) were extracted.
- Split date of sale to extract components such as day of sale, month of sale, year of sale etc.
- Binned variables such as number of bathrooms, bedrooms and floors to make them more intuitive
- Selected Decision Tree Model as the best model because of its good predictive power and less complexity

Recommendations:

- KCHA should find out why sales in certain areas (highlighted in) are low
- KCHA should emphasize buyers on renovating their houses prior to the sale
- KCHA should train specialized agents in line with the cluster characteristics
- KCHA can decide on the best possible starting price for a house with the help of model developed
- KCHA can decide the appropriate selling price or stop loss price based on the model developed

2 Dataset Fetching

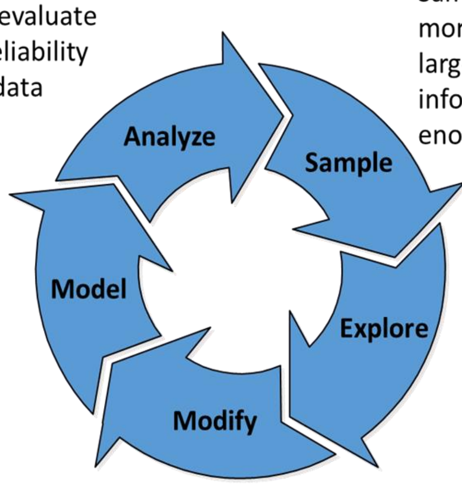
- King County is the most populous county in Washington, and the 13th-most populous county in the United States

- The data for these sales comes from the official public records of home sales in the King County area, Washington State. It contains 21,613 rows, each representing a home sold from May 2014 through May 2015.
- Our idea is to follow the process of statistical data analysis on the existing data

3 SEMMA – Approach

Assess the competing models by building charts to evaluate the usefulness and reliability of findings from the data modeling process

Sample the data by creating one or more tables. The samples should be large enough to contain significant information to retrieve, yet small enough to be used efficiently



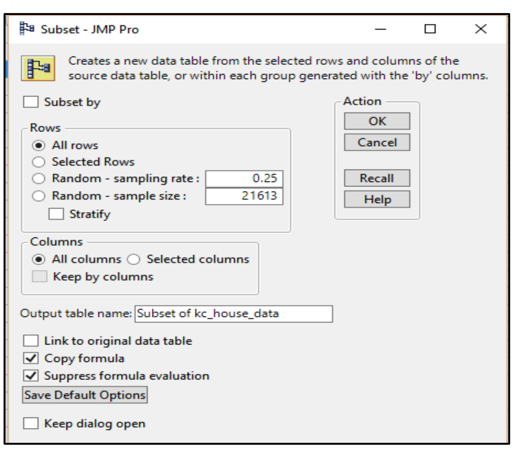
Model the data using various analytical techniques to search for a combination of the data that reliably predicts the desired outcome

Explore the data by searching for anticipated relationships, unanticipated trends and anomalies to gain understanding and ideas majorly through the process of visualization

Modify the data by creating, selecting and transforming the existing variables in preparation for data modeling

3.1 Sampling

The raw dataset consisted of data pertaining to 21,613 home sales in King County, out of which a random sample of 5,403 homes (25%) was extracted to analyze and make statistical inferences



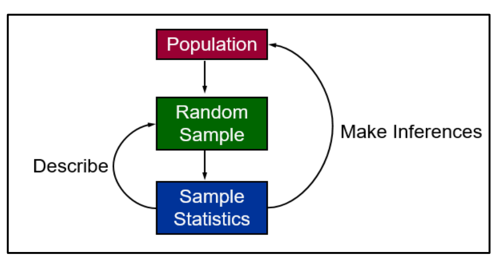


Figure 1 Sampling on SAS JMP

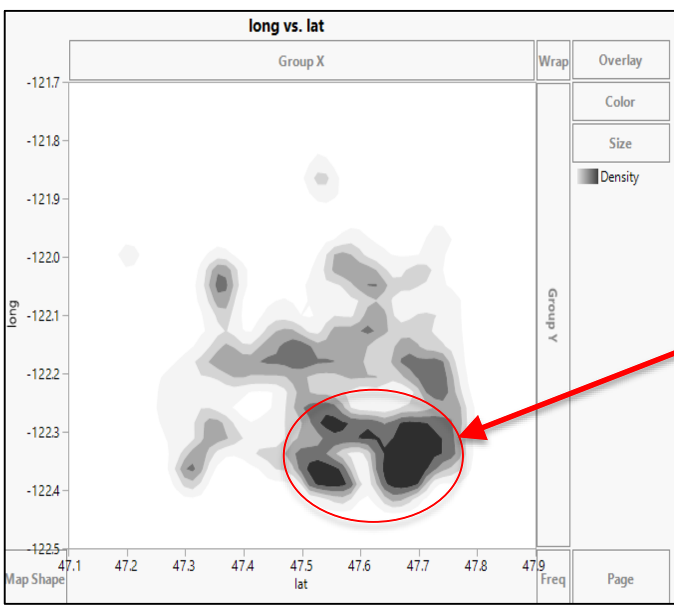
3.1.1 Data Statistics

Analysis Columns	Mean	Median	Std Dev	Min	Max	N Missing	Variance	CV	1 percentile	5 percentile	10 percentile	90 percentile	95 percentile	99 percentile
price	512,650	440,000	279,613	84,000	1,928,000	-	78,183,316,704	55	154,500	210,000	245,000	850,000	1,055,400	1,572,250
sqft_living	2,024	1,889	848	460	7,730	-	719,452	42	720	930	1,070	3,150	3,560	4,575
sqft_above	1,746	1,530	776	460	6,660	-	601,923	44	700	840	960	2,830	3,260	4,205
sqft_basement	279	-	429	-	2,600	-	184,138	154	-	-	-	940	1,160	1,630
yr_built	1,970	1,974	29	1,900	2,015	-	848	1	1,904	1,915	1,926	2,007	2,010	2,014
yr_renovated	82	-	397	-	2,015	-	157,401	483	-	-	-	-	-	2,009
lat	48	48	0	47	48	-	0	0	47	47	47	48	48	48
long	(122)	(122)	0	(123)	(121)	-	0	(0)	(122)	(122)	(122)	(122)	(122)	(122)
sqft_living15	1,960	1,830	661	620	5,790	-	437,342	34	960	1,130	1,240	2,850	3,210	3,965
sqft_lot15	12,267	7,620	25,079	750	560,617	-	628,935,953	204	1,191	2,002	3,708	16,808	36,355	127,359
Eff_area_per_floor	1,469	1,350	636	273	6,055	-	405,090	43	442	670	790	2,300	2,660	3,595
sqft_lot_capped	13,444	7,600	26,161	520	209,418	-	684,423,760	195	1,008	1,912	3,300	20,274	41,199	206,692
Normal 3 Mixture De	0.000065574	0.000077227	0.000037057	0.000000098	0.000107038	-	0.000000001	56.511977863	0.000000103	0.000001927	0.000003684	0.000105790	0.000106749	0.000107018
YOB_build	43.83398766	41	29.13249779	-1	115	0	848.7024273	66.46098003	0	4	8	89	99	110
YOB_renovation	0.688539914	0	4.467976044	-1	74	0	19.96280993	648.9058881	0	0	0	0	0	26

Figure 2 Data Statistics

- There are no missing values in the dataset
- Significant variance is present in independent variables ensuring that variables contain information and can be useful
- This step also assisted in getting a sense out of data and perform data treatment moving forward

3.2 Exploration



Majority of the houses sold are concentrated around latitude of 47.7 and longitude of -122.4.

Figure 3Longitude VS Latitude

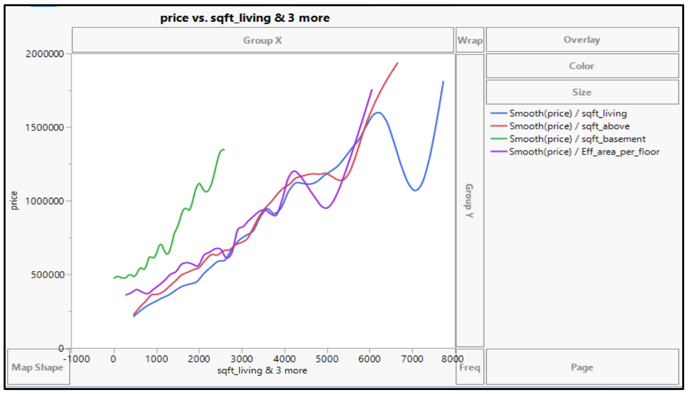


Figure 4 Price vs Area Related Variables

- All the area related variables are directly proportional to house price and the trends are fairly linear
- Basement area and effective area per floor have relatively steeper slope indicating higher correlation to price

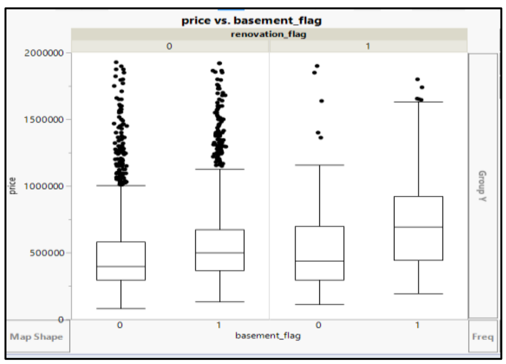


Figure 5 Price vs Basement and Renovation Flags

- Houses which underwent renovation or have a basement fetch higher prices

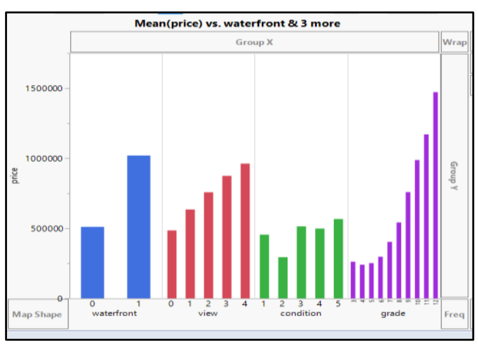


Figure 6 Price vs Quality Related Variables

- Price of the house increases with increase in quality related variables such as view, waterfront, condition and grade

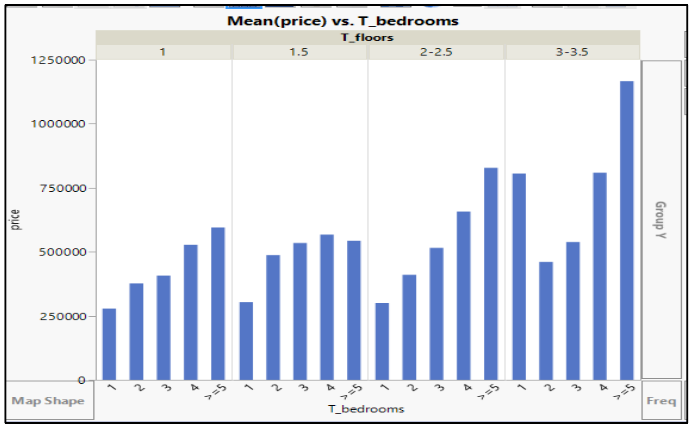


Figure 7 Price vs Bedroom and Floors

- The price of house increases with the increase in number of bedrooms

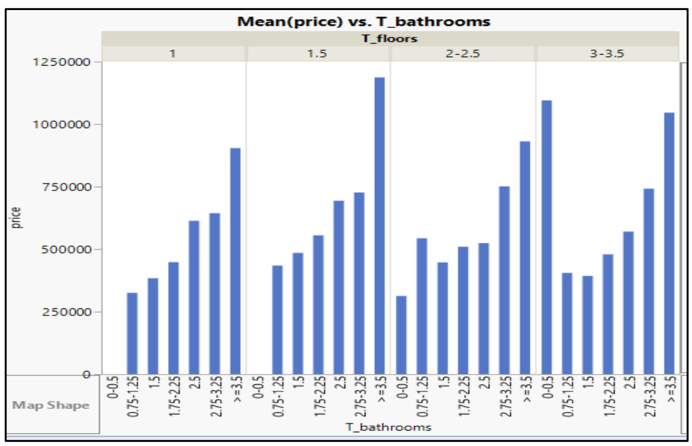


Figure 8 Price vs Bathrooms

- Increase in number of bedrooms and bathrooms also increase the overall price of house

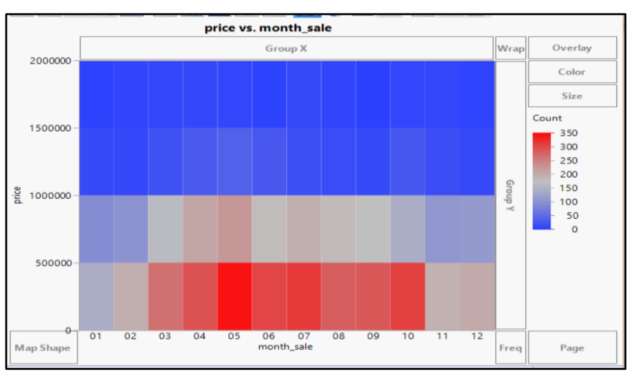


Figure 9 Price VS Month Of Sale

- The sales costly houses is fairly consistent across all months, but the sales of relatively less expensive houses peak between March and October, particularly peaking in May and July

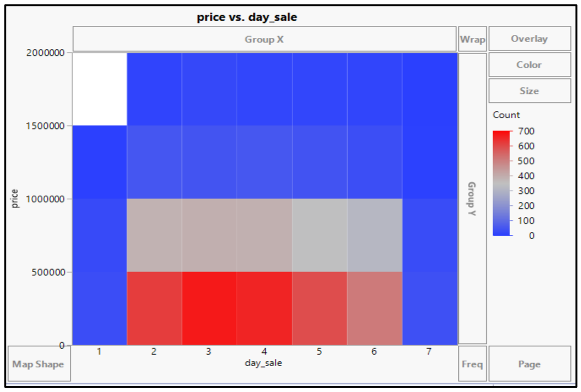


Figure 10 Price VS Day Of Sale

- The sales of lower and medium priced houses is relatively higher on weekdays when compared to weekends which is interesting

3.3 Modification

3.3.1 Data modification

- The overall data quality was very good
- The data did not have any missing values
- Number of outliers were very minimal
- Majority of the distributions were smooth and evenly spread
- However, some minor modifications had to carried out in order to finetune the dataset and analyze it better

3.3.2 Modifications done

- Some homes were sold for exorbitantly high prices, inclusion of these might skew the model. Hence, top 1% of the homes were deleted from dataset
- Variables such as bedrooms, bathrooms and floors showed disproportionate distributions. They were binned intuitively to make them even and more meaningful
- Area of parking lot had extreme values along with a highly-skewed distribution. It was upper capped at 99 percentiles and 'Normal 3 mixture' transformation was applied to smoothen the distribution
- A validation column was created with a split ratio of 70-20-10 for training, validation, and test datasets for future use in model creation

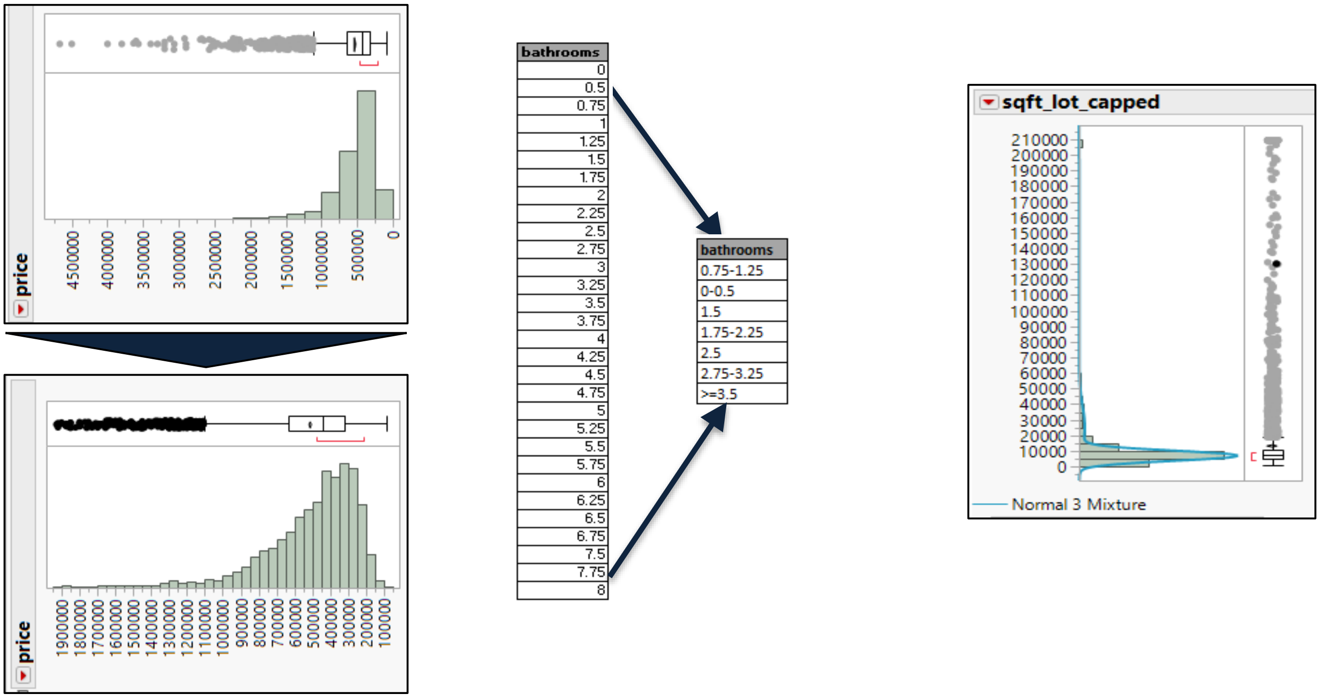
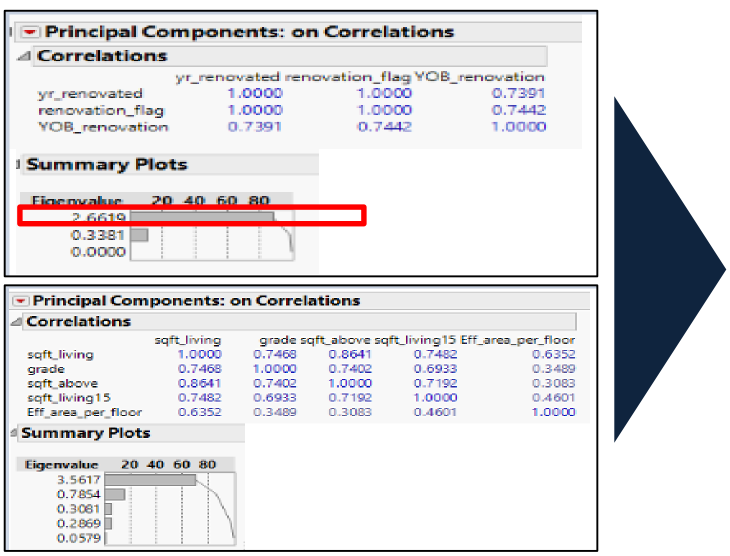


Figure 11 Illustration of Data Modifications

3.3.3 Correlation Analysis and PCA

Correlation analysis was performed to analyze how correlated the independent variables were, post which principal components were generated for each dimension where variables represented similar variables.



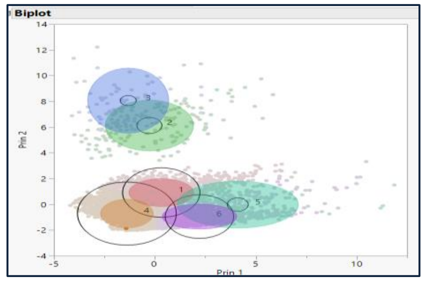
- Two principal components were generated for further use in model creation
- A cumulative variance cutoff of ~80% was used to narrow down on the number of components
- The variables used to create a principal component were highly intercorrelated

Figure 12 Principal Components

3.3.4 Segmentation – Kmeans Clustering

Cluster Comparison

Method	NCluster	CCC	Best
K-Means Clustering	3	9.12554	
K-Means Clustering	4	15.1554	
K-Means Clustering	5	4.43394	
K-Means Clustering	6	45.8981	Optimal CCC
K-Means Clustering	7	33.7708	
K-Means Clustering	8	26.7921	



Cluster Summary

Cluster	Count
1	1471
2	154
3	63
4	2402
5	109
6	1150

Figure 13 Clustering Results

- Clusters 1, 4 and 6 are the major clusters in the population and are nicely separated
- Clusters 2 and 3 can be clubbed because of very similar characteristics and less number of constituent homes
- Cluster 5 overlaps with other clusters, but had very few constituent houses so it could be merged with 6 which is its closest neighbor

3.4 Modeling

Model comparison was run by performing decision tree, bootstrap, boosted tree, regression, and neural network models. The below result was obtained.

Model Comparison Validation=Training

Predictors

Measures of Fit for price

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
price_decision_tree	Partition		0.8391	111875	74767	3751
price_bootstrap	Bootstrap Forest		0.8734	99223	62214	3751
price_boosted_tree	Boosted Tree		0.7804	130694	85947	3751
Price_gen_regression	Fit Least Squares		0.7012	152444	105871	3751
Price_neural	Neural		0.7538	138375	91212	3751

Model Comparison Validation=Validation

Predictors

Measures of Fit for price

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
price_decision_tree	Partition		0.7774	129515	84776	1063
price_bootstrap	Bootstrap Forest		0.7789	129095	83783	1063
price_boosted_tree	Boosted Tree		0.7675	132374	89349	1063
Price_gen_regression	Fit Least Squares		0.6981	150840	107070	1063
Price_neural	Neural		0.7430	139171	93639	1063

Model Comparison Validation=Test

Predictors

Measures of Fit for price

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
price_decision_tree	Partition		0.8221	123908	83898	535
price_bootstrap	Bootstrap Forest		0.7899	134637	84792	535
price_boosted_tree	Boosted Tree		0.7678	141572	94323	535
Price_gen_regression	Fit Least Squares		0.6927	162840	114231	535
Price_neural	Neural		0.7334	151667	99591	535

Figure 14 Model Comparison

Decision tree was the most suited model for training, validation, and test set.

3.4.1 Best Model – Decision Tree

On exploring the decision tree model, we got the following results.

	RSquare	RMSE	N	Number of Splits	AICc
Training	0.748	139924.42	3751	17	99573.2
Validation	0.731	142427.8	1063		
Test	0.784	136627.5	535		

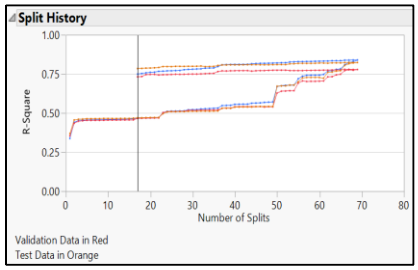


Figure 15 Decision Tree Split History

We see the Rsquare is good. Rsquare of test is higher than Rsquare of training and validation. Also, RMSE of test is lesser than RMSE for validation and training. The model will perform well on any new data.

From the split history, we saw that on clicking GO, it reached a split of 67. Om pruning the number of splits to 17, we see that Rsquare stabilizes after 17 and that the model is not very complex.

Next we took the plot of actual vs predicted values.

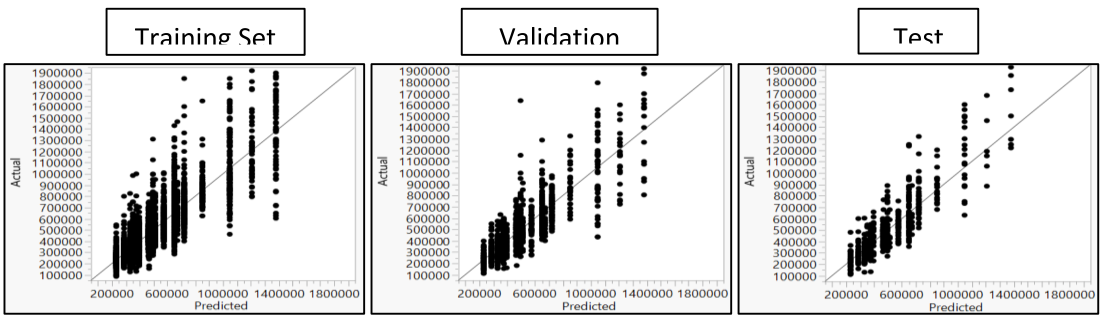


Figure 16 Predicted VS Actual Comparison

The actual and predicted values lie almost equally on either side of x=y line which means the error/residual is random in nature and normally distributed.

3.5 Assess

Based on our analysis, here are our recommendations to KCHA relevant to the business insights.

Insights and Recommendations

1. **Insight**: Majority of the houses are being sold around 47.7 latitude and -122.4 longitude.

   Recommendation: KCHA should evaluate possible causes and further investigate on why exactly this is happening. Is it because of lack of sales force/marketing efforts/quality of houses or some other issue? This analysis will help to increase the sales.

2. **Insight**: Houses with basement and renovation fetch higher prices.

   Recommendation: KCHA should try and emphasize the importance of renovation of a house prior to its sale because it fetches a higher price to the seller and increases the satisfaction of buyer thereby leading to a healthier deal.

3. **Insight**: There are three primary clusters majorly characterized by price, living area and year built.

   Recommendation: The sales agents can be trained keeping these key clusters characteristics in mind to ensure more customer satisfaction and better experience.

4. **Insight**: The decision tree model developed to predict the house price has good R-square and is quite cheap to implement and interpret.

   Recommendation: The model can be leveraged to determine the right price for a house prior to engaging in a deal and to provide the right starting price for the deal.

4 Conclusion

KCHA captured data pertaining to sale of all houses in its county. We leveraged this data and built a model to achieve the mission so that the buyer and seller will be satisfied after deal. During our project we put what we learned into practice including data preprocessing, visualization, correlation analysis, clustering and model comparison.

By exploring data visualizations, we gained deeper understanding of the relationship between Price and other variables. The overall data quality was very good, but we still carried out some minor modifications to finetune the dataset.

Correlation analysis was then performed to analyze the correlations among independent variables and also generate principal components out of correlated variables. After that we performed K-means Clustering and found that there were three primary clusters majorly characterized by price, living area and year of built. We set 0.7:0.2:0.1 as the portion of Training set, Validation set and Test set. We created multiple models for predicting the price of house and by comparison we found that Rsquare of test is greater than training and validation, RMSE for test is less than RMSE for training and validation. Finally we chose the best model-decision tree with good Rsquare and came up with recommendations according to our analysis.

Note of thanks to Professor Jose Cruz for the useful instructions, continuous support and feedback.

5 Appendix

5.1 Data Dictionary

Variable Name	Variable Description	Variable Type
id	primary identifier of the house	nominal
date	date of sale	continuous
price	price of sale	continuous
bedrooms	number of bedrooms	nominal
bathrooms	number of bathrooms	nominal
sqft_living	area of living area without basement in sqft	continuous
sqft_lot	area of parking lot in sqft	continuous
floors	number of floors in the house	nominal
waterfront	flag to indicate whether waterfront is there in the house	nominal
view	scenic view rating of house ranging from 1 to 5	ordinal
condition	building condition of the house, ranging from 1 to 5	ordinal
grade	building grade of the house, ranging from 1 to 13	ordinal
sqft_above	sqft_living — sqft_basement	continuous
sqft_basement	area of basement of the house in sqft	continuous
yr_built	year of build of house	continuous
yr_renovated	year of latest renovation	continuous
zipcode	zip code of house	nominal
lat	latitude	continuous
long	longitude	continuous
sqft_living15	the average house area of the 15 closest houses in sqft	continuous
sqft_lot15	the average parking lot area of the 15 closest houses in sqft	continuous
Validation	validation flag	nominal

Figure 17 Data Dictionary

Though decision tree model was chosen as the best model to suit our needs for this business problem, we created multiple models to compare and conclude that decision tree is the best.

5.2 Bootstrap Forest Model

The bootstrap forest model summary is shown below. The total number of trees in the forest is 19 and number of variables sampled per split is 5. The Rsquare for training dataset is very high at 0.88, but the Rsquare for validation and test fall down compared to training dataset and come out at 0.79 and 0.82. This indicates that the model was overfitting. So which we did not select this as the final model for case.

Bootstrap Forest for price

Specifications

Target Column:	price	Training rows:	3751
Validation Column:	Validation	Validation rows:	1063
		Test rows:	535
Number of trees in the forest:	19	Number of terms:	21
Number of terms sampled per split:	5	Bootstrap samples:	3751
		Minimum Splits Per Tree:	10
		Minimum Size Split:	5

Overall Statistics

Individual Trees	RMSE
In Bag	90143.7
Out of Bag	179325.4

	RSquare	RMSE	N
Training	0.882	95609.486	3751
Validation	0.786	127015.16	1063
Test	0.817	125753.48	535

Figure 18 Bootstrap Forest Performance

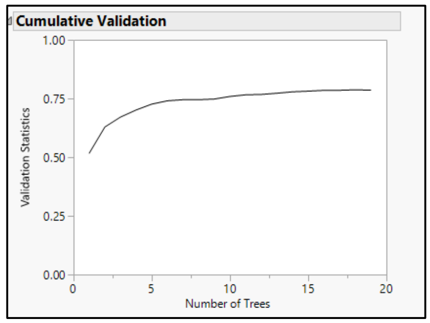


Figure 19 Bootstrap Forest Validation

5.3 Boosted Tree Model

The booted tree model summary is shown below. The total number of layers in the tree is 50 and number of splits per tree is 3. Also, the learning rate was set at 0.1. The Rsquare for training dataset is 0.78, but the Rsquare for validation and test fall down compared to training dataset and come out at 0.77 and 0.77. Since the decision tree model was giving a better predictive power compared to this and moreover the complexity of this is also more, we chose to reject this model.

Boosted Tree for price

Specifications

Target Column:	price	Number of training rows:	3751
Validation Column:	Validation	Number of validation rows:	1063
Number of Layers:	50	Number of test rows:	535
Splits Per Tree:	3		
Learning Rate:	0.1		

Overall Statistics

	RSquare	RMSE	N
Training	0.780	130693.96	3751
Validation	0.767	132373.5	1063
Test	0.768	141572.11	535

Figure 20 Boosted Tree Performance

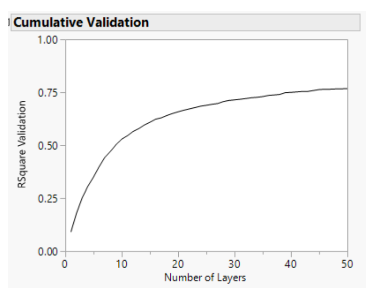
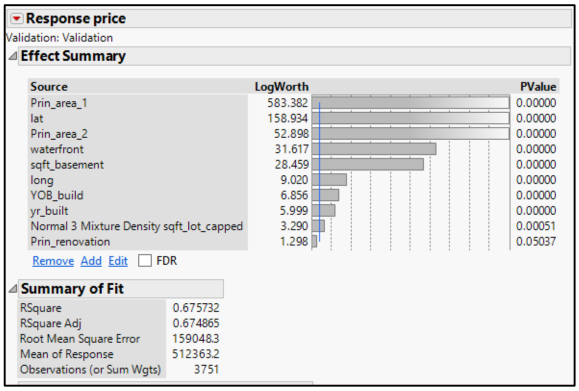


Figure 21 Boosted Tree Validation

5.4 Generalized Regression Model

The regression model summary is shown below. The total number of variables in the model was 10. The Rsquare for training dataset is 0.68, but the Rsquare for validation and test fall down compared to training dataset and come out at 0.68 and 0.68. Since the model Rsquare was very low especially when compared with decision tree model, we chose to reject this model.



Response price

Validation: Validation

Effect Summary

Source	LogWorth	PValue
Prin_area_1	583.382	0.00000
lat	158.934	0.00000
Prin_area_2	52.898	0.00000
waterfront	31.617	0.00000
sqft_basement	28.459	0.00000
long	9.020	0.00000
YOB_build	6.856	0.00000
yr_built	5.999	0.00000
Normal 3 Mixture Density sqft_lot_capped	3.290	0.00051
Prin_renovation	1.298	0.05037

Remove Add Edit FDR

Summary of Fit

RSquare	0.675732
RSquare Adj	0.674865
Root Mean Square Error	159048.3
Mean of Response	512363.2
Observations (or Sum Wgts)	3751

Figure 22 Regression Model Contribution

Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.6757	158815	3751
Validation Set	0.6790	155536	1063
Test Set	0.6766	167065	535

Figure 23 Regression Model Performance

5.5 Neural Network Model

The neural network model summary is shown below. The Rsquare for training dataset is 0.60, but the Rsquare for validation and test fall down compared to training dataset and come out at 0.56 and 0.54. Since the model Rsquare was very low especially when compared with decision tree model, we chose to reject this model.

Model NTanH(3)

Training: price Measures	Value	Validation: price Measures	Value	Test: price Measures	Value
RSquare	0.6013181	RSquare	0.5600822	RSquare	0.5368096
RMSE	176097.37	RMSE	182077.27	RMSE	199930.85
Mean Abs Dev	117256.49	Mean Abs Dev	127799.34	Mean Abs Dev	136339.65
-LogLikelihood	50629.989	-LogLikelihood	14383.586	-LogLikelihood	7289.196
SSE	1.163e+14	SSE	3.524e+13	SSE	2.139e+13
Sum Freq	3751	Sum Freq	1063	Sum Freq	535

Figure 24 Neural Network Model Performance

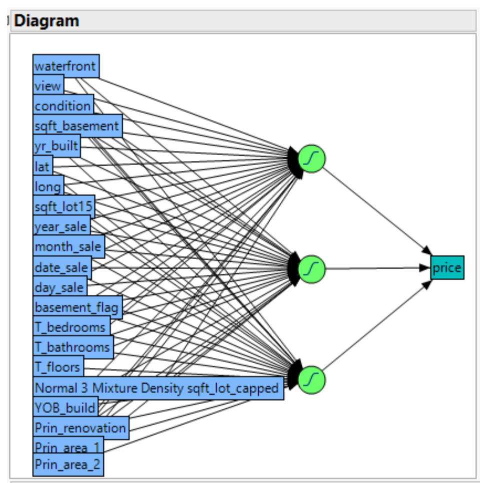


Figure 25 Neural Network Model Structure

6 References

1. https://www.kcha.org/
2. https://www.kaggle.com/harlfoxem/housesalesprediction
3. http://techtalk.seattle.gov/page/2/