Our Video Recording Link - Click Here!
(Please select 1080p Quality)

WQD7005 DATA MINING – GROUP PROJECT

HOUSE PRICE PREDICTION

GROUP 5

Name	Matric Number
Jasmeen Kah Ying Bong	S2142739
Hii Yew Han	S2037987
Dinesh V M Ramachandran	S2119167
Aw Yeong Fung Mun	17197465
Lee Ziteng Nicholas	S2132376

METHODOLOGY

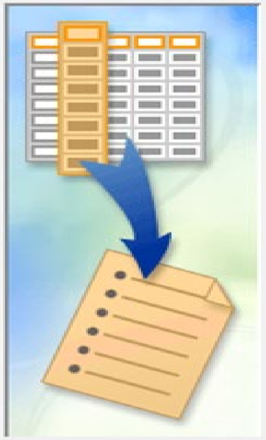
Sample	Identification of variables or factors (both dependent and independent) impacting the process is the aim of the first stage of the process.
Explore	Univariate and multivariate analysis are carried out in order to investigate interrelated relationships between data items and to find data gaps.
Modify	Business logic is used to derive the lessons discovered during the exploration phase from the data gathered during the sample phase.
Model	Employs a variety of data mining techniques to create a projected model of how this data achieves the process's final, desired result.
Assess	The model's applicability and dependability to the subject under study are assessed.

MODIFY – DATA MODIFICATION

To perform classification, we have created a new attributes named price_range
If price <= median value of price, price_range = 0
If price > median value of price, price_range = 1

price_range
0
0
1
0
0
1
0
1
0
0
0
0
1
1
0
1
1
1
1
1
1
1
1
1
1

Data Source Wizard -- Step 5 of 8 Column Metadata



(none) | not | Equal to | Apply | Reset

Columns: Label | Mining | Basic | Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper L
bathrooms	Input	Interval	No		No	.	
bedrooms	Input	Interval	No		No	.	
condition	Input	Ordinal	No		No	.	
date	Time ID	Nominal	No		No	.	
floors	Input	Interval	No		No	.	
grade	Input	Ordinal	No		No	.	
id	ID	Interval	No		No	.	
lat	Input	Interval	No		No	.	
long	Input	Interval	No		No	.	
price	Input	Interval	No		No	.	
price_range	Target	Binary	No		No	.	
sqft_above	Input	Interval	No		No	.	
sqft_basement	Input	Interval	No		No	.	
sqft_living	Input	Interval	No		No	.	
sqft_living15	Input	Interval	No		No	.	
sqft_lot	Input	Interval	No		No	.	
sqft_lot15	Input	Interval	No		No	.	
view	Input	Ordinal	No		No	.	
waterfront	Input	Binary	No		No	.	
yr_built	Input	Interval	No		No	.	
yr_renovated	Input	Interval	No		No	.	
zipcode	Input	Interval	No		No	.	

Show code | Explore | Compute Summary | < Back | Next > | Cancel

MODIFY – INCONSISTENT DATA

Replace Inconsistent Data with Correct Value using Talend

Inconsistent yr_built value	After replacement
192102	1921
19570522	1957
190810	1908
19310401	1931
192703	1927
19590731	1959
191006	1910

MODIFY – NOISY DATA & INCOMPLETE DATA

Use the Limits Method to replace Noisy Data and Incomplete Data with missing value

Variable	Error Type	Limits method
bathrooms bedrooms sqft_above price lat long sqft_living15 sqft_lot15 sqft_living sqft_lot sqft_basement	Noisy	Extreme Percentiles
bathrooms	Incomplete	Mean

MODIFY – NOISY DATA & INCOMPLETE DATA

We padded incomplete and missing data with mean.
Imputation summary showing imputed variable and impute value

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
REP_bathrooms	MEAN	IMP_REP_bathrooms	2.107488	INPUT	INTERVAL	Replacement: bathr...	178
REP_bedrooms	MEAN	IMP_REP_bedrooms	3.359458	INPUT	INTERVAL	Replacement: bedro...	75
REP_lat	MEAN	IMP_REP_lat	47.56079	INPUT	INTERVAL	Replacement: lat	215
REP_long	MEAN	IMP_REP_long	-122.215	INPUT	INTERVAL	Replacement: long	214
REP_price	MEAN	IMP_REP_price	528856.8	INPUT	INTERVAL	Replacement: price	216
REP_sqft_above	MEAN	IMP_REP_sqft_above	1774.22	INPUT	INTERVAL	Replacement: sqft_...	201
REP_sqft_basement	MEAN	IMP_REP_sqft_bas...	282.1509	INPUT	INTERVAL	Replacement: sqft_...	103
REP_sqft_living	MEAN	IMP_REP_sqft_living	2063.344	INPUT	INTERVAL	Replacement: sqft_li...	200
REP_sqft_living15	MEAN	IMP_REP_sqft_livin...	1978.331	INPUT	INTERVAL	Replacement: sqft_li...	213
REP_sqft_lot	MEAN	IMP_REP_sqft_lot	13092.99	INPUT	INTERVAL	Replacement: sqft_lot	214
REP_sqft_lot15	MEAN	IMP_REP_sqft_lot15	11534.16	INPUT	INTERVAL	Replacement: sqft_l...	213

MODIFY – DATA TRANSFORMATION

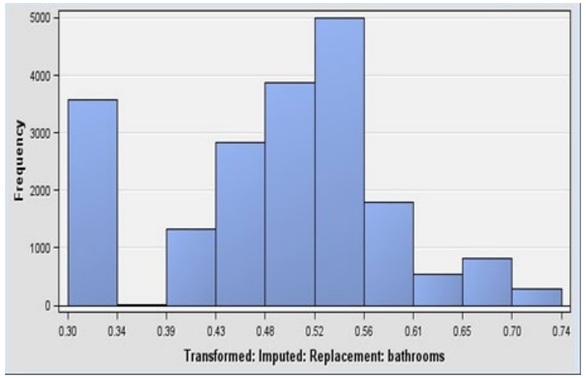
Apply normalisation

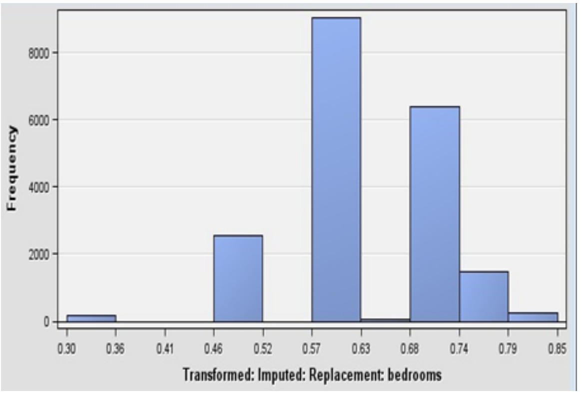
- it lessens skewness
- it is advantageous for machine learning algorithms that assume the feature variable has a normal distribution, lowering the level of measurement while keeping the ratio constant
- it enhances the model's training efficiency

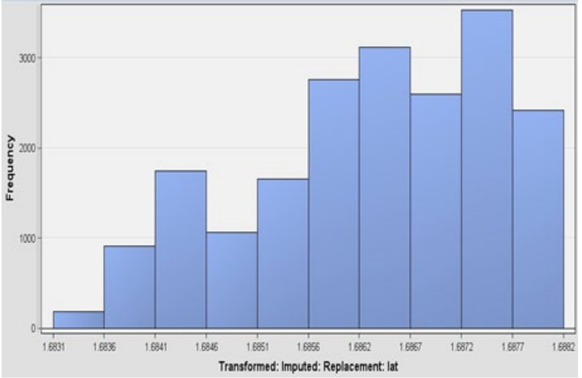
Log 10 Transformation

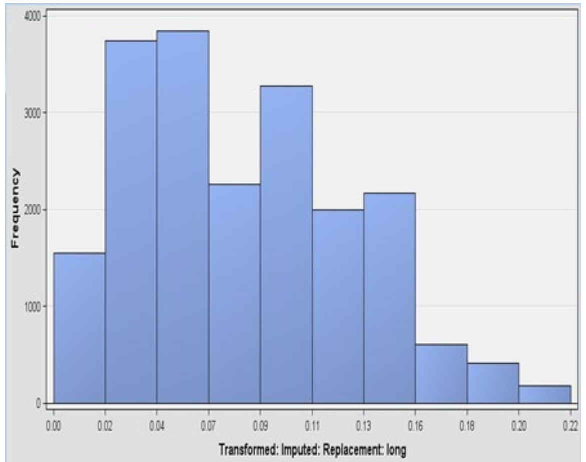
MODIFY – DATA TRANSFORMATION

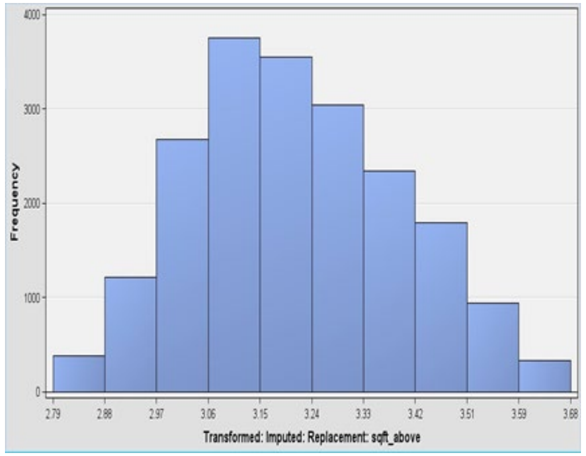
After Log10 Transformation

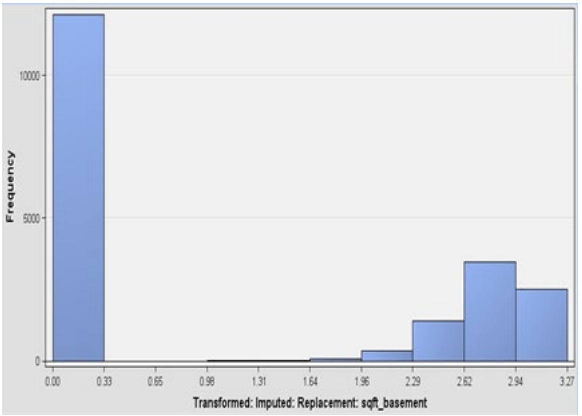






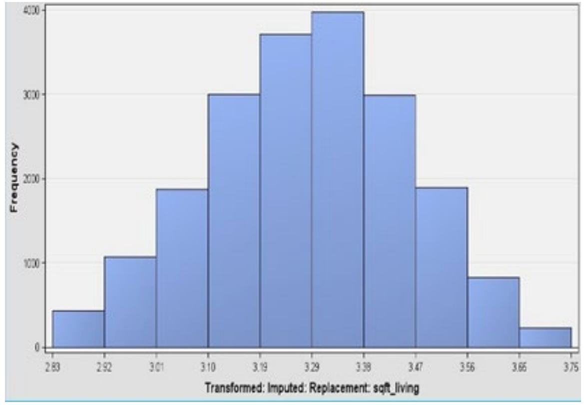


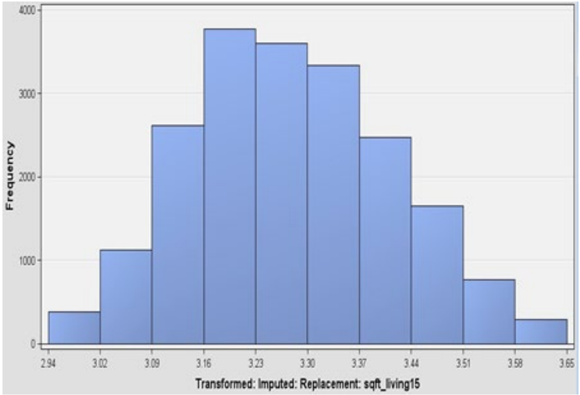


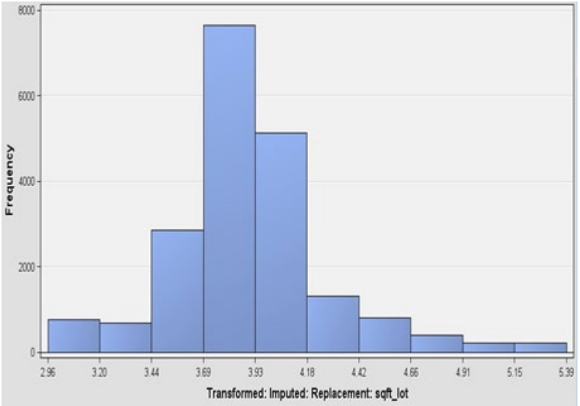


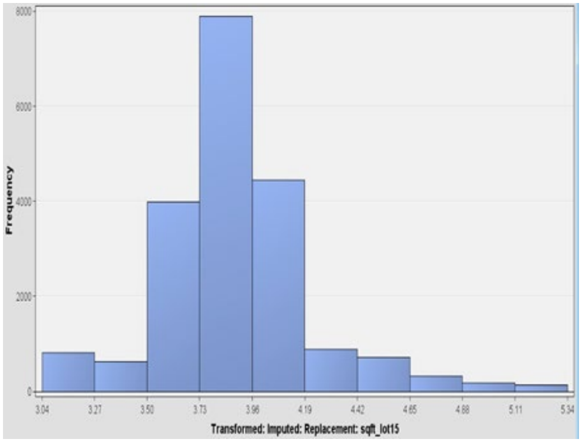
MODIFY – DATA TRANSFORMATION

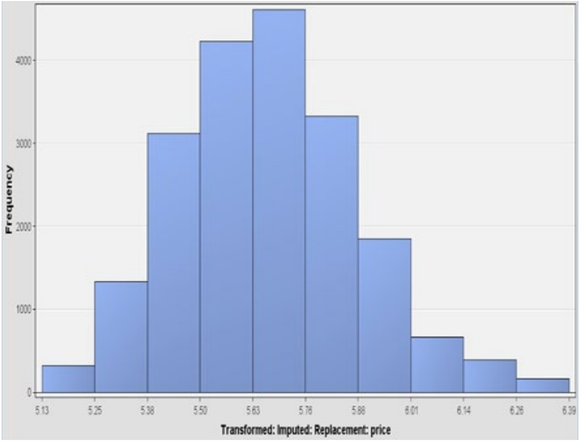
After Log10 Transformation





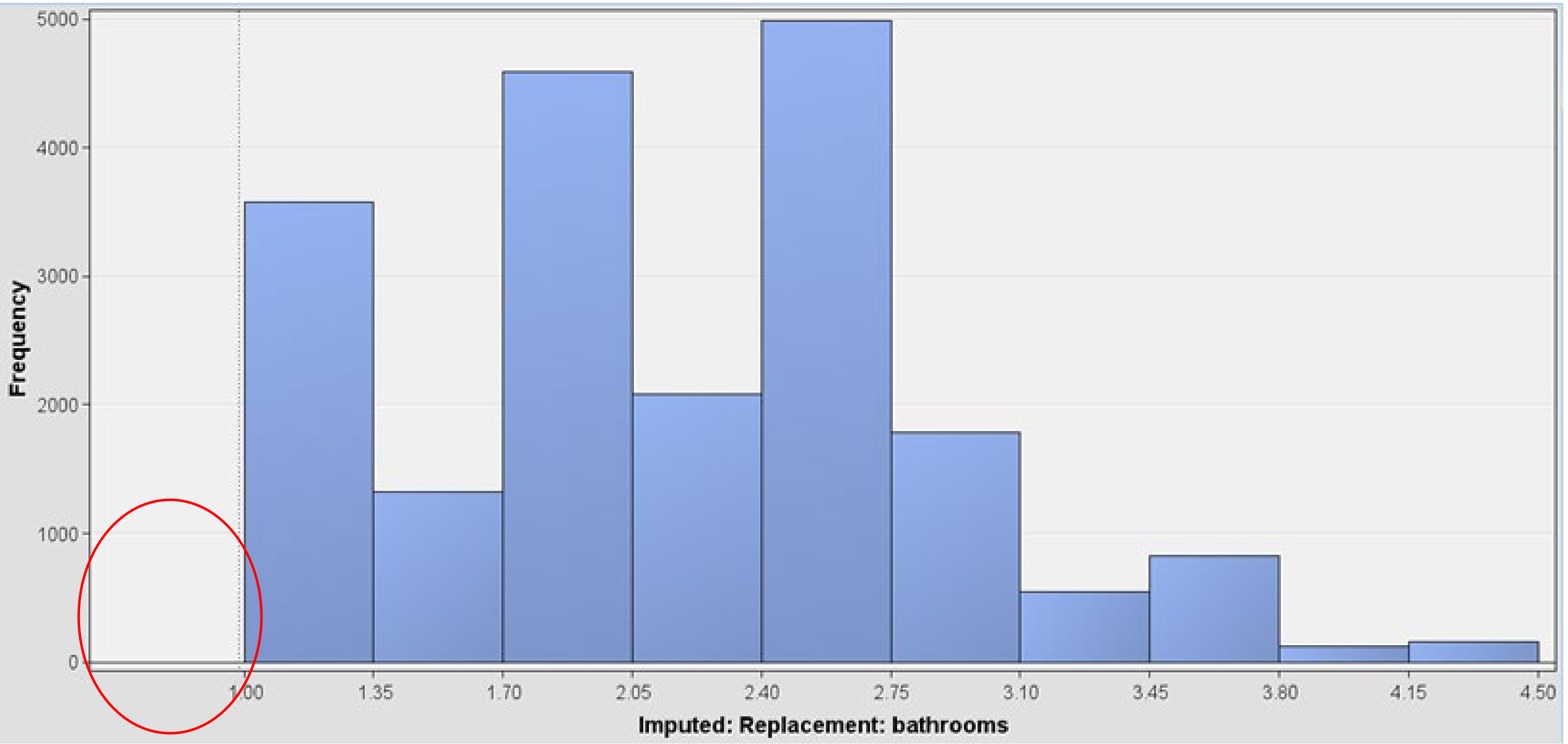






MODIFY – EXAMINING EXPORTED DATA

Bathrooms variables – no more missing data



MODIFY – EXAMINING EXPORTED DATA

yr_built variable -> no more inconsistent data

yr_built (integer)	yr_built_substring (integer)
1970	1970
1948	1948
2003	2003
192102	1921
2007	2007
1958	1958
1916	1916
1950	1950
2007	2007

yr_built (integer)	yr_built_substring (integer)
1968	1968
1978	1978
1994	1994
1996	1996
1912	1912
191006	1910
1976	1976

yr_built (integer)	yr_built_substring (integer)
1952	1952
1954	1954
2005	2005
1983	1983
1976	1976
19590731	1959
1908	1908
1955	1955

MODIFY – EXAMINING EXPORTED DATA

yr_built variable -> no more inconsistent data

yr_built integer	yr_built_substring integer
1968	1968
1918	1918
192703	1927
1960	1960
1959	1959
1928	1928
1997	1997

yr_built integer	yr_built_substring integer
1964	1964
1945	1945
1996	1996
19310401	1931
1968	1968
1995	1995
1981	1981

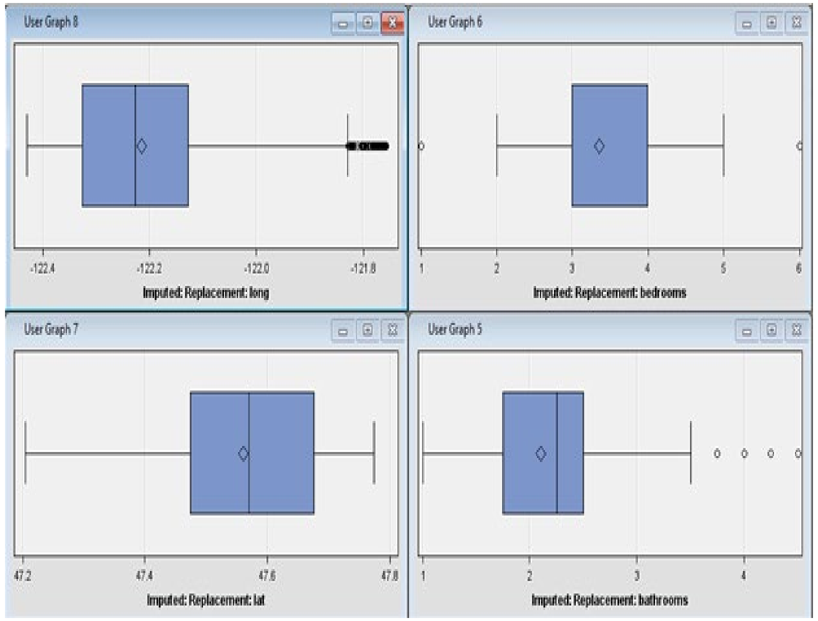
yr_built integer	yr_built_substring integer
2004	2004
1976	1976
190810	1908
2005	2005
1966	1966
1904	1904
1926	1926

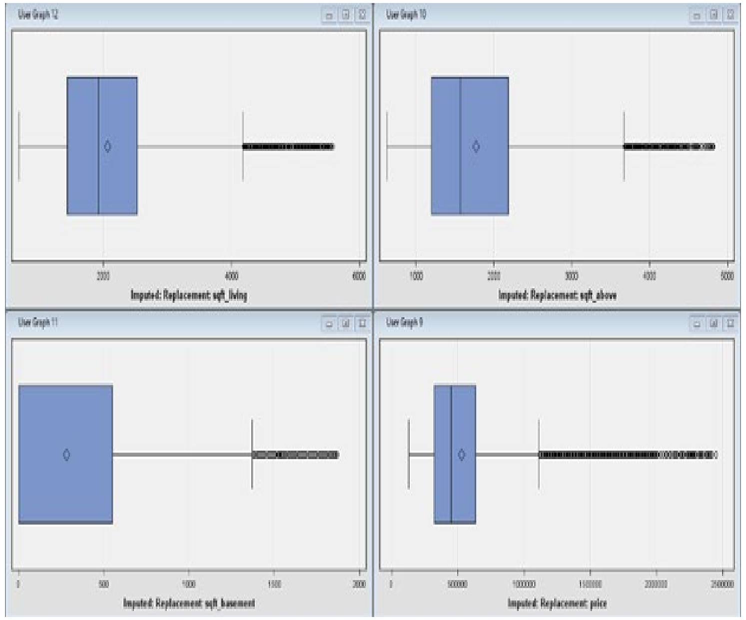
yr_built integer	yr_built_substring integer
1981	1981
1990	1990
19570522	1957
1980	1980
2005	2005
1952	1952
1967	1967

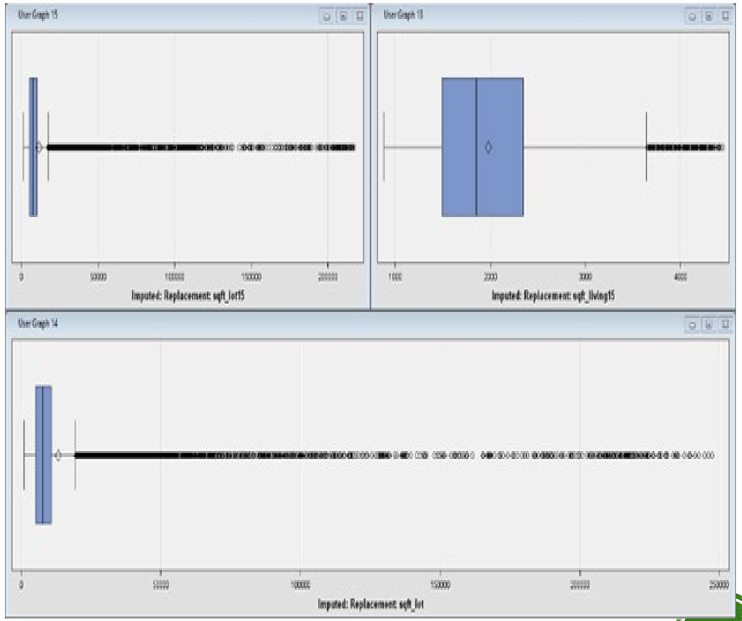
MODIFY – EXAMINING EXPORTED DATA

bathrooms variable, bedrooms variable, sqft_above variable, price variable, lat variable, long variable, sqft_living15 variable, sqft_lot15 variable, sqft_living variable, sqft_lot variable and sqft_basement variable -> some of them still have outliers

To maintain the originality of the dataset to prevent overfitting, we decide to keep the remaining outliers.







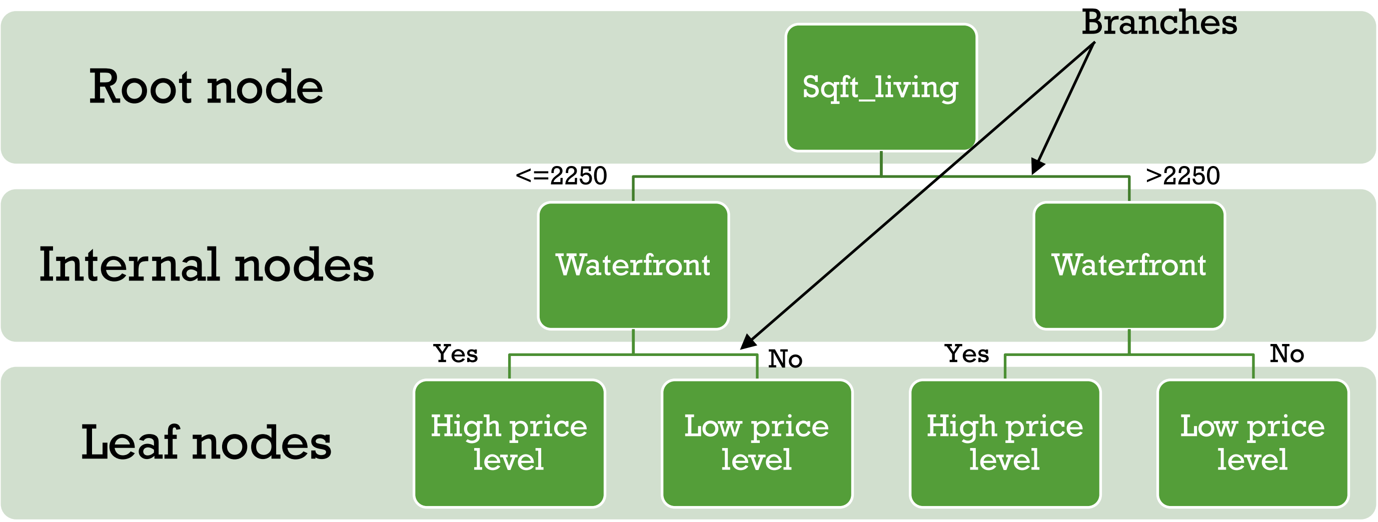
MODIFY –TRAINING & VALIDATION DATA

Name	Drop	Role	Level
LG10_IMP_REP_bathrooms	Yes	Input	Interval
LG10_IMP_REP_bedrooms	No	Input	Interval
LG10_IMP_REP_lat	No	Input	Interval
LG10_IMP_REP_long	Yes	Input	Interval
LG10_IMP_REP_price	Yes	Input	Interval
LG10_IMP_REP_sqft_above	Yes	Input	Interval
LG10_IMP_REP_sqft_basement	Yes	Input	Interval
LG10_IMP_REP_sqft_living	No	Input	Interval
LG10_IMP_REP_sqft_living15	Yes	Input	Interval
LG10_IMP_REP_sqft_lot	Yes	Input	Interval
LG10_IMP_REP_sqft_lot15	No	Input	Interval
condition	No	Input	Ordinal
date	Yes	Time ID	Nominal
floors	Yes	Input	Interval
grade	No	Input	Ordinal
id	Yes	ID	Interval
price_range	No	Target	Binary
view	No	Input	Ordinal
waterfront	No	Input	Binary
yr_built	Yes	Input	Interval
yr_renovated	No	Input	Interval
zipcode	Yes	Input	Interval

Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0

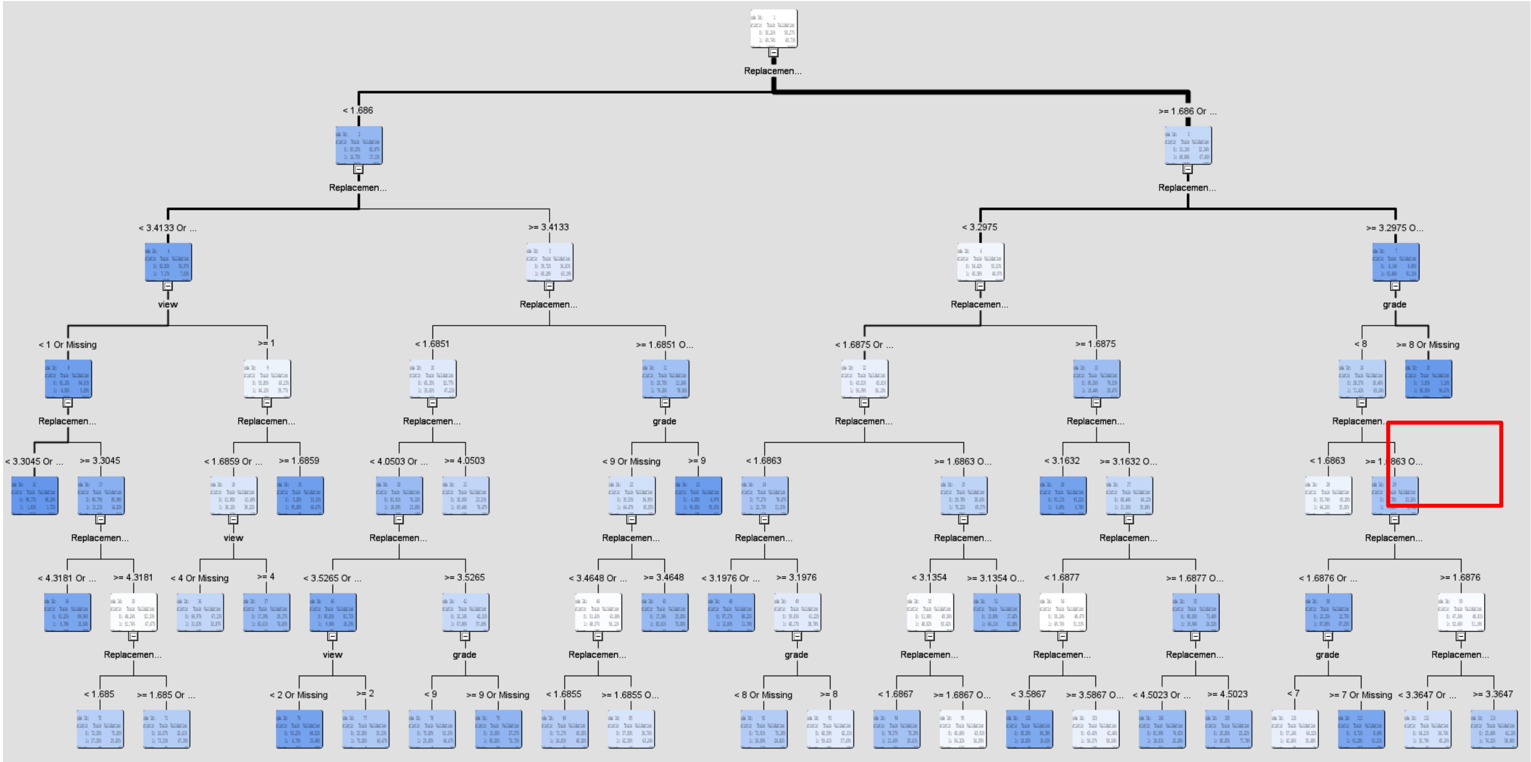
- The modified dataset exported from Talend
- Imported to SAS Enterprise Miner with the specified roles and levels.
- The modified dataset is partitioned into 50:50 training and validation data.

MODEL: DECISION TREE



- A supervised model
- Tree-based structure, consists of a root node, internal nodes, branches, leaf nodes
- Simple and can manage a high volume of data (21,613 rows in this study).

MODEL: DECISION TREE



Interesting findings:

1. About 33 decision rules.
2. Latitude has a high level of information gain, selected as the root node.
3. Shortest split is after 3 layers.
4. When the grade is high, the price range is more likely to be high.

MODEL: DECISION TREE

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
price_range		_NOBS_	Sum of Frequencies	10806	10807	
price_range		_MISC_	Misclassification Rate	0.117897	0.12751	
price_range		_MAX_	Maximum Absolute Error	0.981655	0.981655	
price_range		_SSE_	Sum of Squared Errors	1884.061	2050.38	
price_range		_ASE_	Average Squared Error	0.087177	0.094863	
price_range		_RASE_	Root Average Squared Error	0.295257	0.307999	
price_range		_DIV_	Divisor for ASE	21612	21614	
price_range		_DFT_	Total Degrees of Freedom	10806	.	

Statistical Output:

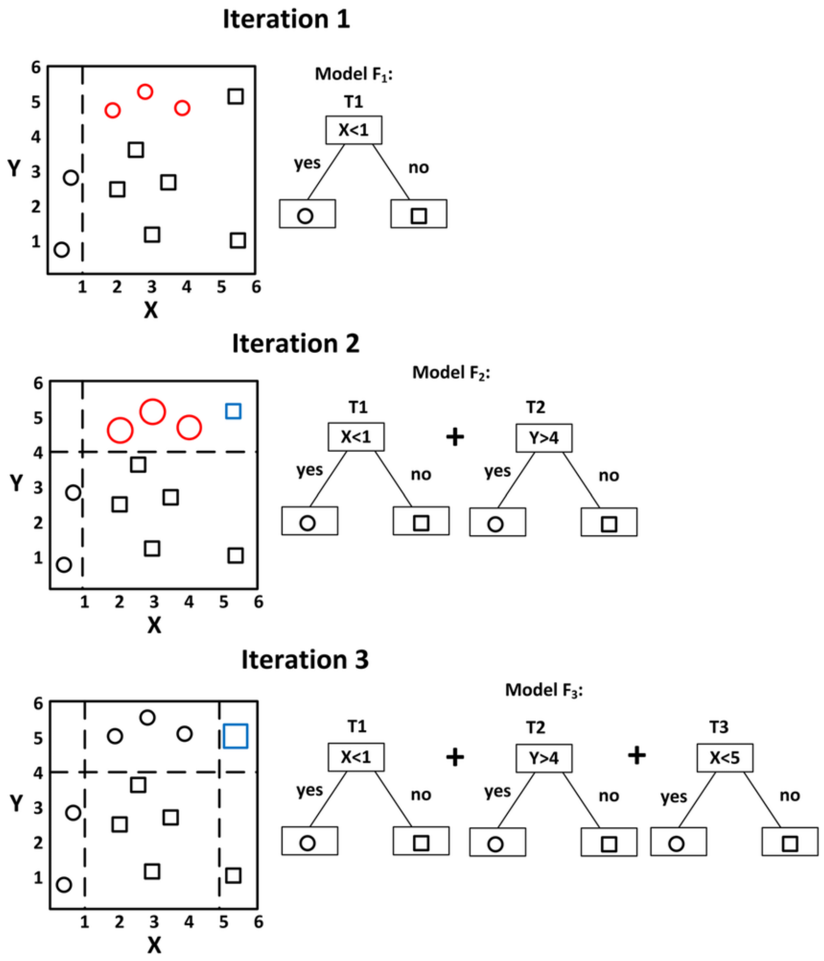
- The misclassification rate= 0.12751
- Accuracy= 0.87249

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
REP_LG10_IMP_REP_lat	Replacement: Transformed: Imputed: Replac...	11	1.0000	1.0000	1.0000
REP_LG10_IMP_REP_sqft_living	Replacement: Transformed: Imputed: Replac...	9	0.8081	0.8390	1.0383
grade		5	0.2468	0.2648	1.0733
REP_LG10_IMP_REP_sqft_lot15	Replacement: Transformed: Imputed: Replac...	4	0.2153	0.2150	0.9984
view		3	0.1921	0.1637	0.8525
REP_LG10_IMP_REP_bedrooms	Replacement: Transformed: Imputed: Replac...	0	0.0000	0.0000	.
REP_yr_renovated	Replacement: yr_renovated	0	0.0000	0.0000	.
condition		0	0.0000	0.0000	.
waterfront		0	0.0000	0.0000	.

Variable importance:

- Top 3 variables are latitude, sqft_living, grade.

MODEL: GRADIENT BOOSTING



- Used for both regression and classification tasks.
- An ensemble of weak predictors, which are usually decision trees.
- Each new tree is built to improve on the deficiencies of the previous trees and this concept is called *boosting*.
- Helps in reducing bias error in the model.

MODEL: GRADIENT BOOSTING

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
price_range		_NOBS_	Sum of Frequencies	10806	10807	.
price_range		_SUMW_	Sum of Case Weig...	21612	21614	.
price_range		_MISC_	Misclassification R...	0.1191	0.122421	.
price_range		_MAX_	Maximum Absolute...	0.962634	0.971617	.
price_range		_SSE_	Sum of Squared Er...	1924.193	1989.676	.
price_range		_ASE_	Average Squared ...	0.089034	0.092055	.
price_range		_RASE_	Root Average Squ...	0.298385	0.303406	.
price_range		_DIV_	Divisor for ASE	21612	21614	.
price_range		_DFT_	Total Degrees of F...	10806	.	.

Statistical Output:

- The misclassification rate= 0.1191
- Accuracy= 0.8809

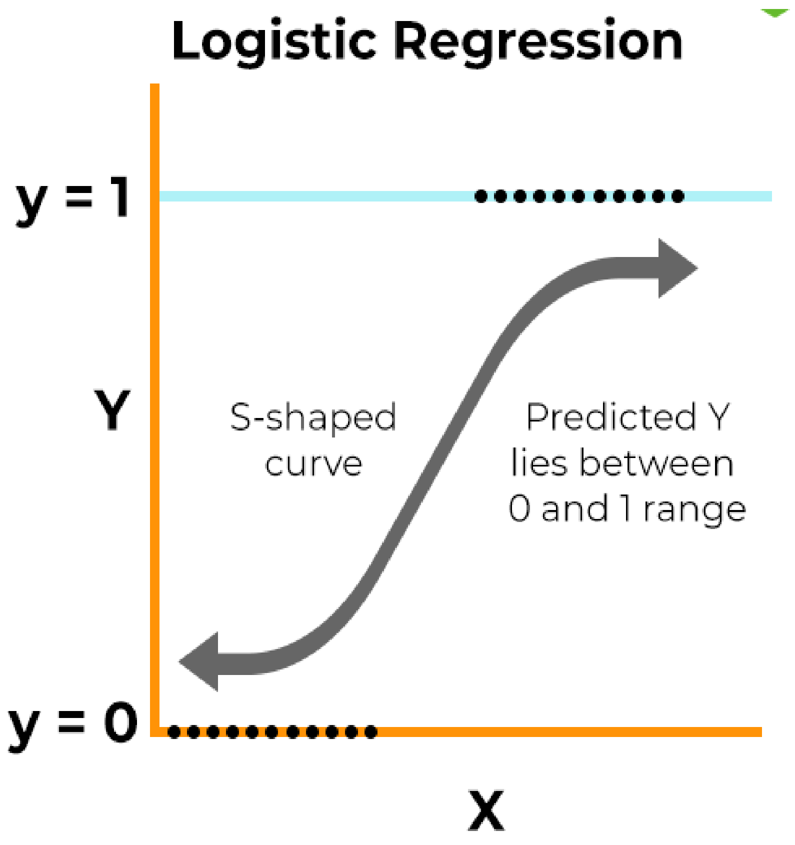
Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LG10_IMP_REP_lat	Transformed: Imputed:...	82	1	1	1
LG10_IMP_REP_sqft_living	Transformed: Imputed:...	35	0.751581	0.785669	1.045356
grade		14	0.601595	0.668711	1.111564
view		7	0.154703	0.151371	0.978463
LG10_IMP_REP_sqft_lot15	Transformed: Imputed:...	12	0.141388	0.149133	1.054778
LG10_IMP_REP_bedrooms	Transformed: Imputed:...	0	0	0	.
yr_renovated		0	0	0	.
waterfront		0	0	0	.
condition		0	0	0	.

Variable importance:

- Top 3 variables are latitude, sqft_living, grade.

MODEL: LOGISTIC REGRESSION



For **classification and predictive** analytics.

Commonly used algorithm for solving **Binary Classification** problems.

Predicts a dependent variable by analyzing the **relationship** between one or more existing independent variables.

The advantage is the ability to use more than one continuous attribute **simultaneously.**

MODEL: LOGISTIC REGRESSION

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
price_range		_AIC_	Akaike's Information ...	9272.608	.	.
price_range		_ASE_	Average Squared Error	0.138747	0.138657	.
price_range		_AVERR_	Average Error Function	0.425995	0.427032	.
price_range		_DFE_	Degrees of Freedom ...	10773	.	.
price_range		_DFM_	Model Degrees of Fr...	33	.	.
price_range		_DFT_	Total Degrees of Fre...	10806	.	.
price_range		_DIV_	Divisor for ASE	21612	21614	.
price_range		_ERR_	Error Function	9206.608	9229.877	.
price_range		_FPE_	Final Prediction Error	0.139597	.	.
price_range		_MAX_	Maximum Absolute E...	0.99586	0.999712	.
price_range		_MSE_	Mean Square Error	0.139172	0.138657	.
price_range		_NOBS_	Sum of Frequencies	10806	10807	.
price_range		_NW_	Number of Estimate ...	33	.	.
price_range		_RASE_	Root Average Sum of...	0.372487	0.372367	.
price_range		_RFPE_	Root Final Prediction ...	0.373627	.	.
price_range		_RMSE_	Root Mean Squared ...	0.373057	0.372367	.
price_range		_SBC_	Schwarz's Bayesian ...	9513.107	.	.
price_range		_SSE_	Sum of Squared Errors	2998.597	2996.932	.
price_range		_SUMW_	Sum of Case Weight	21612	21614	.
price_range		_MISC_	Misclassification Rate	0.201832	0.202461	.

Statistical Output:

- The misclassification rate= 0.2025
- **Accuracy= 79.76%**

MODEL: LOGISTIC REGRESSION

Odds Ratio Estimates

Effect		price_ range	Point Estimate
LG10_IMP_REP_bathrooms		1	4.283
LG10_IMP_REP_bedrooms		1	0.103
LG10_IMP_REP_lat		1	.
LG10_IMP_REP_long		1	29.239
LG10_IMP_REP_sqft_above		1	9.406
LG10_IMP_REP_sqft_basement		1	1.464
LG10_IMP_REP_sqft_living		1	3.435
LG10_IMP_REP_sqft_living15		1	57.098
LG10_IMP_REP_sqft_lot		1	0.873
LG10_IMP_REP_sqft_lot15		1	0.452
condition	1 vs 5	1	0.412
condition	2 vs 5	1	0.357
condition	3 vs 5	1	0.583
condition	4 vs 5	1	0.629
floors		1	1.962

Odds ratio measures how strong is the **association** of an event with exposure.

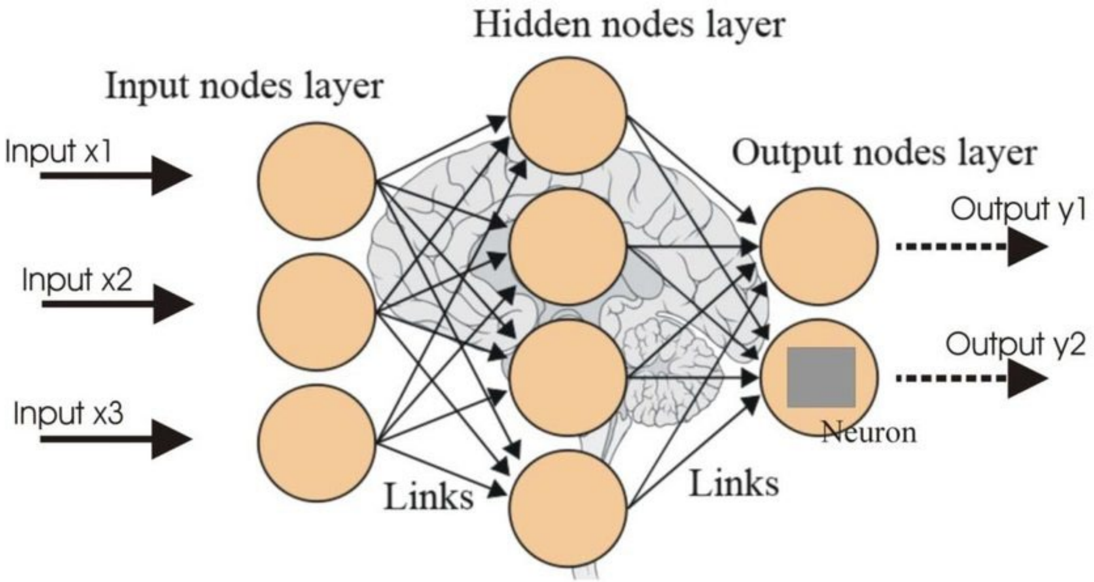
Based on the output of Odds Ratio Estimates we found that:

1) Sqft_living15 has **57 times** the odds of having a higher price range than a lower price range level.

2) Long(Longitude) has **29 times** the odds of having a higher price range than a lower price range level.

3) Sqft_above has **9 times** the odds of having a higher price range than lower price range level.

MODEL: NEURAL NETWORK



A neural network

- machine learning process (deep learning)
- collection of algorithms that employ linked neurons in a layered framework.

MODEL: NEURAL NETWORK

Hidden layers	Misclassification rate (Accuracy in percentage)
3	0.11363 (88.64%)
4	0.113723 (88.63%)
5	**0.111872 (88.81%)**
10	0.122791 (87.72%)
15	0.115758 (88.42%)
20	0.138244 (86.18%)
25	0.152679 (84.73%)
30	0.143888 (85.61%)
35	0.13445 (86.56%)
40	0.150828 (84.92%)

Generate neural network:

3- layer
4- layer
5- layer
10-layer
15-layer
20-layer

20-layer
25-layer
30-layer
35-layer
40-layer

MODEL: LOGISTIC REGRESSION

Odds Ratio Estimates

Effect		price_ range	Point Estimate
LG10_IMP_REP_bathrooms		1	4.283
LG10_IMP_REP_bedrooms		1	0.103
LG10_IMP_REP_lat		1	.
LG10_IMP_REP_long		1	29.239
LG10_IMP_REP_sqft_above		1	9.406
LG10_IMP_REP_sqft_basement		1	1.464
LG10_IMP_REP_sqft_living		1	3.435
LG10_IMP_REP_sqft_living15		1	57.098
LG10_IMP_REP_sqft_lot		1	0.873
LG10_IMP_REP_sqft_lot15		1	0.452
condition	1 vs 5	1	0.412
condition	2 vs 5	1	0.357
condition	3 vs 5	1	0.583
condition	4 vs 5	1	0.629
floors		1	1.962

Odds ratio measures on how strongly an event is associated with exposure in this scenario would be the price range.

Based on the output of Odds Ratio Estimates we found that:

1) Sqft_living15 has **57 times** the odds of having higher price range than lower price range level.

2) Long(Longitude) has **29 times** the odds of having higher price range than lower price range level.

3) Sqft_above has **9 times** the odd of having higher price range than lower price range level.

MODEL: NEURAL NETWORK

Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Neural3	5 layer N...	price_ran...		0.111872
Neural	3 layer N...	price_ran...		0.11363
Neural2	4 layer N...	price_ran...		0.113723
Neural5	15 layer ...	price_ran...		0.115758
Neural9	10 layer ...	price_ran...		0.122791
Neural7	35 layer ...	price_ran...		0.13445
Neural10	20 layer ...	price_ran...		0.138244
Neural8	30 layer ...	price_ran...		0.143888
Neural6	40 layer ...	price_ran...		0.150828
Neural4	25 layer ...	price_ran...		0.152679

Statistical Output:

- Top 3 of neural network in house price prediction:

1. 5- layer
   Misclassification rate=0.111872
   Accuracy=88.81%
2. 3- layer
   Misclassification rate=0.11363
   Accuracy= 88.64%
3. 4- layer
   Misclassification rate=0.113723
   Accuracy= 88.63%

ASSESS

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Train: Sum of Frequencies	Train: Misclassification Rate	Selection Criterion: Valid: Misclassification Rate
Y	Boost	Boost	Gradient Bo...	price_range	10806	0.1191	0.122421
	Tree	Tree	Decision Tr...	price_range	10806	0.117897	0.12751
	Reg	Reg	Regression	price_range	10806	0.201832	0.202461

Comparison of models:

Gradient Boosting:

 Misclassification rate=0.122421 (accuracy=88%)

Decision Tree:

 Misclassification rate=0.12751 (accuracy=87.42%)

Logistic Regression:

 Misclassification rate= 0.202461 (accuracy=79.75%)

ASSESS

Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Neural3	5 layer N...	price ran...		0.111872
Neural	3 layer N...	price ran...		0.11363
Neural2	4 layer N...	price ran...		0.113723
Neural5	15 layer ...	price ran...		0.115758
Neural9	10 layer ...	price ran...		0.122791
Neural7	35 layer ...	price ran...		0.13445
Neural10	20 layer ...	price ran...		0.138244
Neural8	30 layer ...	price ran...		0.143888
Neural6	40 layer ...	price ran...		0.150828
Neural4	25 layer ...	price ran...		0.152679

Comparison of neural network models:

1. 5- layer
   Misclassification rate=0.111872
   Accuracy=88.81%
2. 3- layer
   Misclassification rate=0.11363
   Accuracy= 88.64%
3. 4- layer
   Misclassification rate=0.113723
   Accuracy= 88.63%

ASSESS

Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Neural11	5 layer N...	price_ran...		0.111872
Boost	Gradient ...	price_ran...		0.122421

Comparison of models:

1. 5- layer
   Misclassification rate=0.111872
   Accuracy=88.81%
2. Gradient Boosting
   Misclassification rate= 0.122421
   Accuracy= 87.75%

CONCLUSION

Number	Attributes
1	grade
2	lat
3	sqft_living
4	view
5	waterfront
6	condition
7	yr_renovated
8	sqft_lot15
9	bedrooms

SAS Enterprise Miner -> variable selection tool -> relevant variables

CONCLUSION

- Interesting Patterns

Phase	Interesting Patterns
Exploration	○ Higher grade level shows a high **positive relationship** with price. Longer boxplot body length and higher price were observed as the grade increased. ○ Sqft_living shows a **positive relationship** with price. ○ Sqft_living and sqft_above show a **strong correlation** coefficient, 0.8766.
Modeling Phase – Decision Tree	○ Latitude is selected as the **root node** as it has a high level of information gain. ○ Shortest split is **after 3 layers**. ○ When the grade is high, the price range is more likely to be **high**.