OPIM 5604

King County House Prices Prediction Model

SPRING SEMESTER 2017

INSTRUCTOR: IVA STRICEVIC

TEAM 6

SUMMARY

Team Members:

Abdallah Alsaqri
Sree Inturi
Pawan Shivhare
Sakshi Singhania
Karpagam Thamaya Vinayagam

This project was completed using a dataset acquired through Kaggle. The data for the dataset was provided by King County, Washington. The data includes homes sold between May 2014 and May 2015.

The Goal of this analysis is to predict the price of housing for 2016 in King County, based on the variables provided in the dataset. Any additional observation, unrelated to the goal of predicting house pricing will be recorded and summarized at the end of this paper.

Objectives

I. Overview of Data
II. Data pre-processing
III. Data visualization and pattern discovery
IV. Predictive Modeling
V. Model Implementation
VI. Plan for future upgrades

I. Overview of Data

The analysis dataset consists of Price of Houses in King County, Washington from sales between May 2014 and May 2015. Along, with house price it consists of information on 18 house features, Date of Sale and ID of sale. The table below describes the interpretation of the variables in the dataset.

Variable	Description
Id	Unique ID for each home sold
Date	Date of the home sale
Price	Price of each home sold
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms, where .5 accounts for a room with a toilet but no shower
Sqft_living	Square footage of the apartments interior living space
Sqft_lot	Square footage of the land space
Floors	Number of floors
Waterfront	A dummy variable for whether the apartment was overlooking the waterfront or not
View	An index from 0 to 4 of how good the view of the property was
Condition	An index from 1 to 5 on the condition of the apartment,
Grade	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design
Sqft_above	The square footage of the interior housing space that is above ground level
Sqft_basement	The square footage of the interior housing space that is below ground level
Yr_built	The year the house was initially built
Yr_renovated	The year of the house's last renovation
Zipcode	What zipcode area the house is in
Lat	Lattitude
Long	Longitude
Sqft_living15	The square footage of interior housing living space for the nearest 15 neighbors
Sqft_lot15	The square footage of the land lots of the nearest 15 neighbors

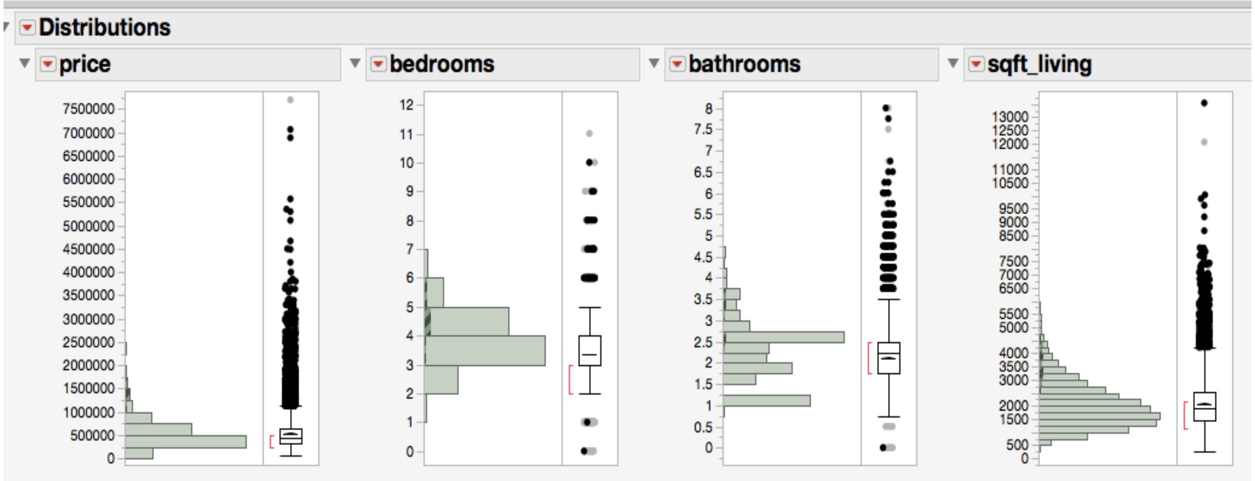
II. Data Preprocessing

A majority of the fields found in the King County housing dataset were deemed acceptable for performing our statistical analyses. However, while traversing the data we found that some of the columns needed to have their data types adjusted to meet our needs. Likewise, as we were discussing the data, we found that we needed to create new columns. New columns, essential for analysis, were created by applying new formulae on existing variables. In order to provide us with a better understanding of our data and to improve the accuracy of our model, we made the decision to retain the original columns along with the transformed data.

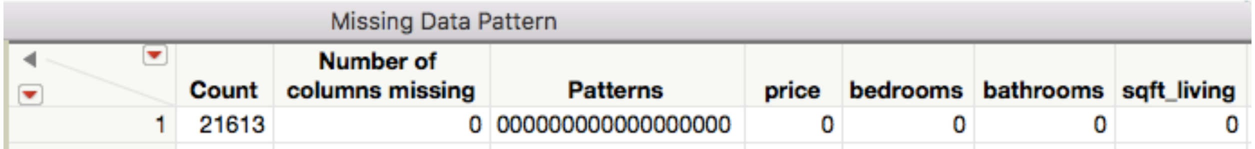
The columns transformed are listed below.

1. **Year**: New Column Data Type: Numeric Nominal
    a. Calculation: Extracted from Date Column
2. **Month**: New Column Data Type: Numeric Nominal
    a. Calculation: Extracted from Date Column
3. **Waterfront**: Changed Data Type from Numeric Continuous to Numeric Nominal
4. **View**: Changed Data Type from Numeric Continuous to Numeric Nominal
5. **Condition**: Changed Data Type from Numeric Continuous to Numeric Nominal
6. **Grade**: Changed Data Type from Numeric Continuous to Numeric Nominal
7. **Age**: New column Data Type: Numeric Continuous
    a. Calculation: 2017-Yr_Built
8. **Renovated_flg**: New Columns Data Type: Numeric Nominal
    a. Calculation: If Renovated_Year=0 then Renovated_flg =1 else Renovated_flg = 0

Outlier Detection: Outliers were detected and analyzed using the Outlier Boxplots. From the outliers boxplot we inferred that the data consists of many outliers for the target variable, Price. However, the outliers for price variable corresponded to outliers for Number of Bedrooms, Number of Bathrooms and Square Feet Living. On further investigation, we inferred that these outliers correspond to High values of Grade, Condition, and View. Hence, we concluded that these outliers are legitimate outliers and we decided to retain them in the data.



Missing Values Detection: Missing data pattern was used to identify the missing data in the dataset. From the table below it can be observed that the data does not consist of any missing data for any of the variables.



Missing Data Pattern

	Count	Number of columns missing	Patterns	price	bedrooms	bathrooms	sqft_living
1	21613	0	000000000000000000	0	0	0	0

Summary of other data inconsistencies: While exploring the data we found a few instances where the data between variables was inconsistent and didn't make logical sense. We chose to either make the values consistent by recoding or exclude those instances from the data.

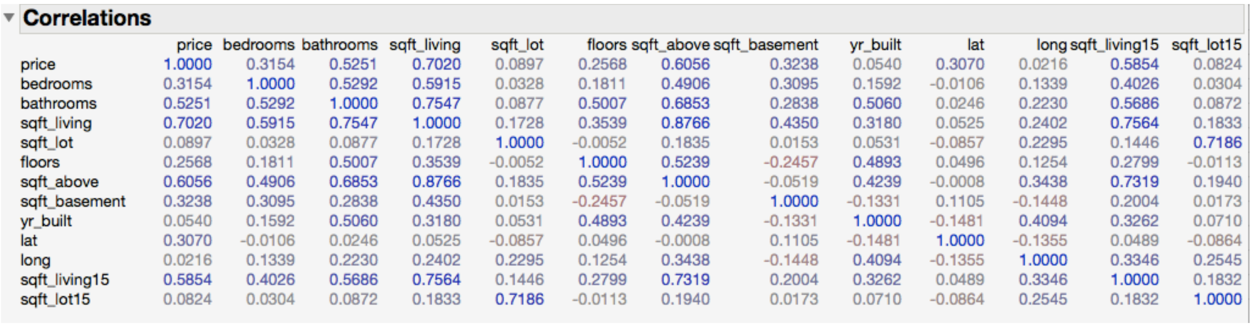
1. One observation with 33 bedrooms in 1620 Square feet with 1.75 bathrooms. The value of bedrooms was recoded to 3.
2. Ten observations with 0 bathrooms. Since it is not conventional to have houses without bathrooms, we decided to exclude these observations.

III. Data Visualization and Pattern Discovery

The objective of data visualization and pattern discovery was to reveal relationships between the house features and the response variable, price. We wanted to identify house features that affect price variable and could be potential predictors. Through visualization, we gathered the following information about the data.

1. Price increases with increase in Square Feet Living, Square Feet above, Number of Bathrooms and Number of Bedrooms.
2. Price increases as we move from South to North along the latitude and shows little variation along the longitude.
3. Price increase with increase in Grade and Condition.
4. Renovated houses are likely to have a higher price compared to Non-Renovated houses.
5. Houses with View 3 and 4 are likely to have a higher price compared to houses with View 0, 1, or 2.
6. Houses with waterfront are associated with high price compared to houses without waterfront.

Correlation Table: The below correlation table provides a summary of correlation between the continuous variables in the data. The objective behind analyzing correlation between the continuous variables in the data was to identify variables that have significant linear relationship with price and those who don't. Further, the table helps to identify relationship between potential predictors. If two predictors are highly correlated with each other they may explain the same variation in the price variable, leading to over fitting.



▾ Correlations

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	sqft_above	sqft_basement	yr_built	lat	long	sqft_living15	sqft_lot15
price	1.0000	0.3154	0.5251	0.7020	0.0897	0.2568	0.6056	0.3238	0.0540	0.3070	0.0216	0.5854	0.0824
bedrooms	0.3154	1.0000	0.5292	0.5915	0.0328	0.1811	0.4906	0.3095	0.1592	-0.0106	0.1339	0.4026	0.0304
bathrooms	0.5251	0.5292	1.0000	0.7547	0.0877	0.5007	0.6853	0.2838	0.5060	0.0246	0.2230	0.5686	0.0872
sqft_living	0.7020	0.5915	0.7547	1.0000	0.1728	0.3539	0.8766	0.4350	0.3180	0.0525	0.2402	0.7564	0.1833
sqft_lot	0.0897	0.0328	0.0877	0.1728	1.0000	-0.0052	0.1835	0.0153	0.0531	-0.0857	0.2295	0.1446	0.7186
floors	0.2568	0.1811	0.5007	0.3539	-0.0052	1.0000	0.5239	-0.2457	0.4893	0.0496	0.1254	0.2799	-0.0113
sqft_above	0.6056	0.4906	0.6853	0.8766	0.1835	0.5239	1.0000	-0.0519	0.4239	-0.0008	0.3438	0.7319	0.1940
sqft_basement	0.3238	0.3095	0.2838	0.4350	0.0153	-0.2457	-0.0519	1.0000	-0.1331	0.1105	-0.1448	0.2004	0.0173
yr_built	0.0540	0.1592	0.5060	0.3180	0.0531	0.4893	0.4239	-0.1331	1.0000	-0.1481	0.4094	0.3262	0.0710
lat	0.3070	-0.0106	0.0246	0.0525	-0.0857	0.0496	-0.0008	0.1105	-0.1481	1.0000	-0.1355	0.0489	-0.0864
long	0.0216	0.1339	0.2230	0.2402	0.2295	0.1254	0.3438	-0.1448	0.4094	-0.1355	1.0000	0.3346	0.2545
sqft_living15	0.5854	0.4026	0.5686	0.7564	0.1446	0.2799	0.7319	0.2004	0.3262	0.0489	0.3346	1.0000	0.1832
sqft_lot15	0.0824	0.0304	0.0872	0.1833	0.7186	-0.0113	0.1940	0.0173	0.0710	-0.0864	0.2545	0.1832	1.0000

From the correlation table the following can be inferred:

a) Price has a high positive correlation with Square Feet Living and moderate positive correlation with Number of Bathrooms, Square Feet Above, and Square Feet Living15.

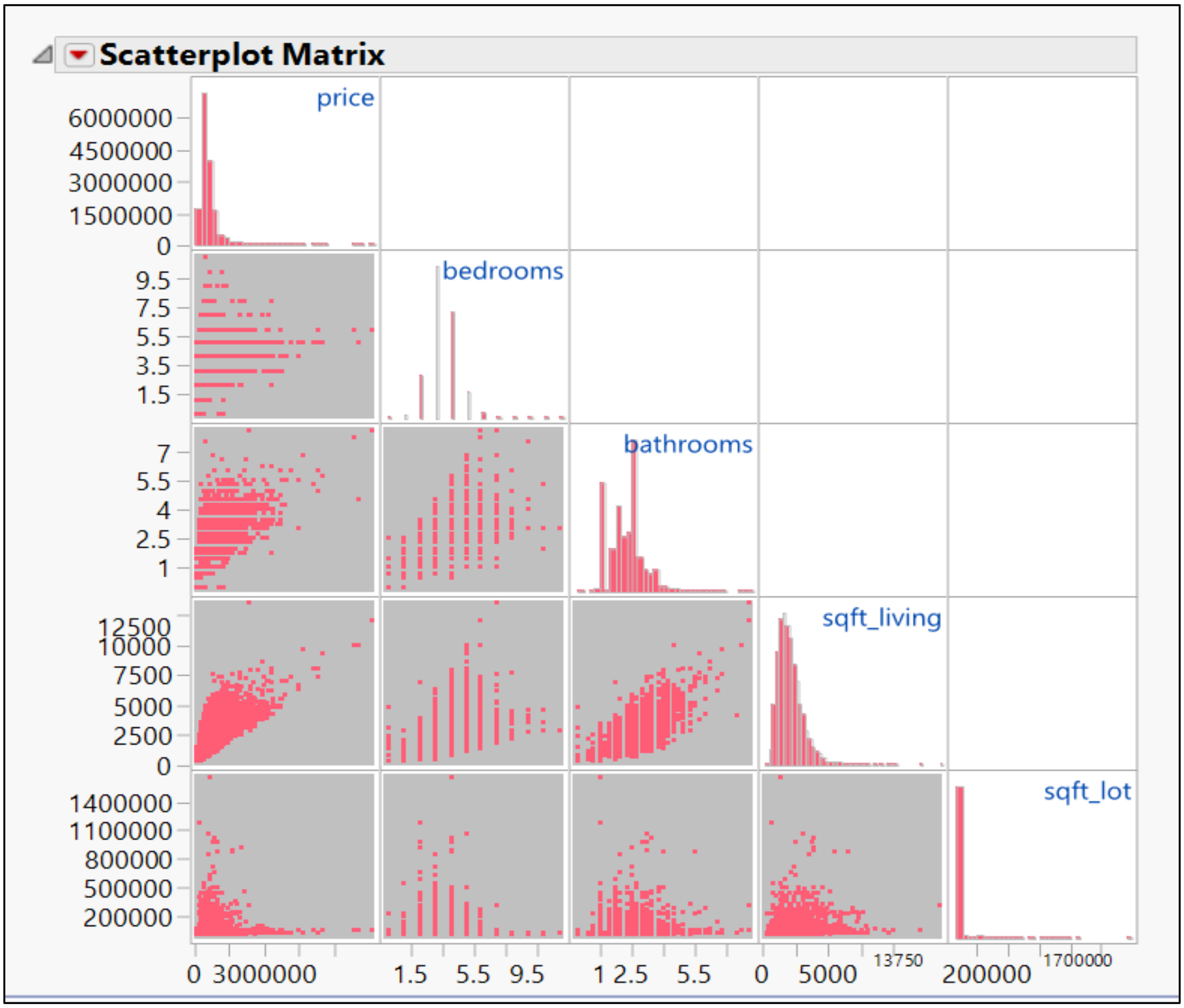
b) Price has low positive correlation with Number of Bedrooms, Floors, Square Feet Basement and Latitude.

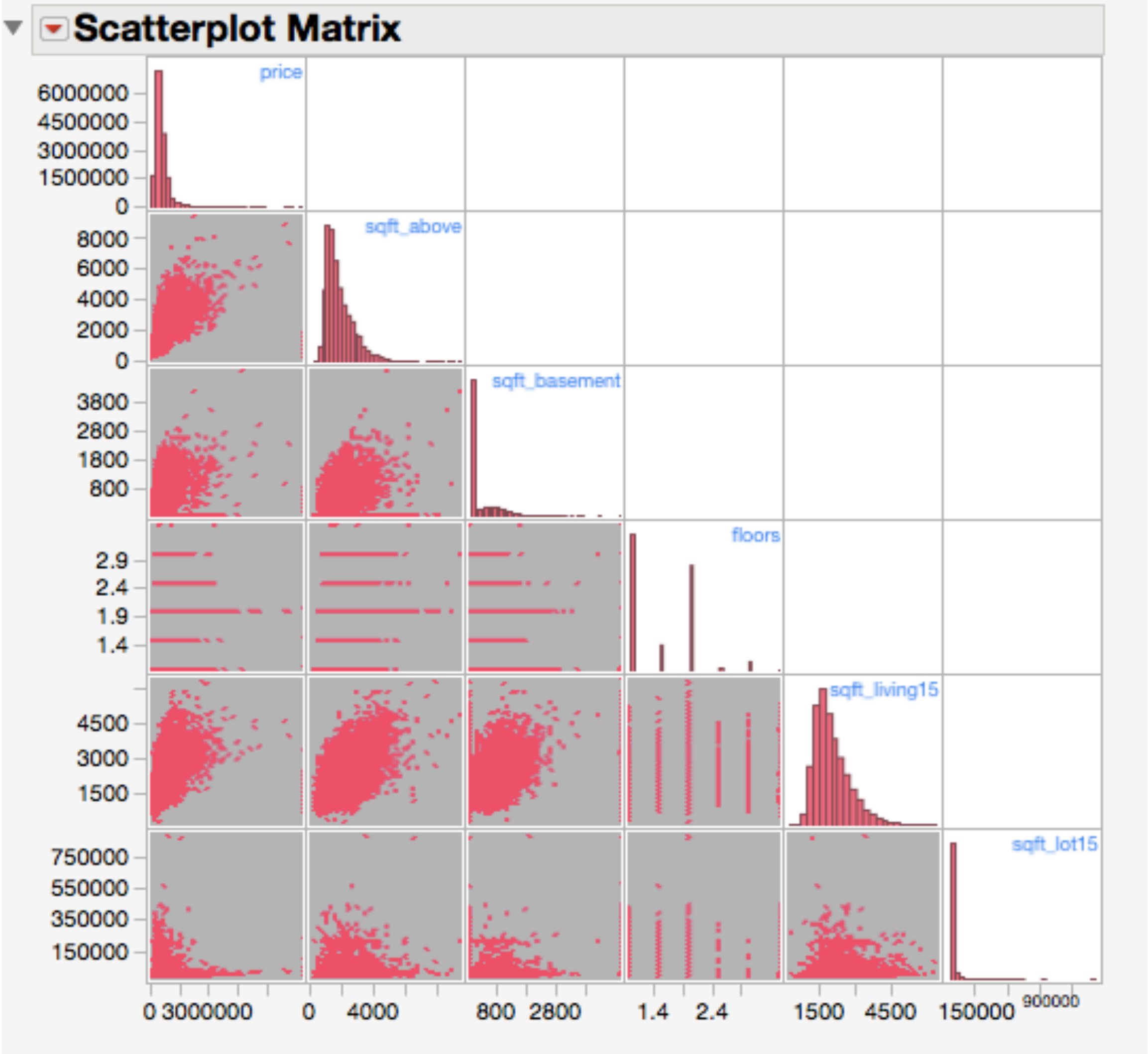
c) Price shows non-significant relationship with Square Feet Lot, Year Built, Longitude and Square Feet Lot15.

d) Square Feet Above, Square Feet Living15, Number of Bathrooms and Number of Bedrooms show high positive correlation with Square Feet Living and may explain the same variation in Price as Square Feet Living.

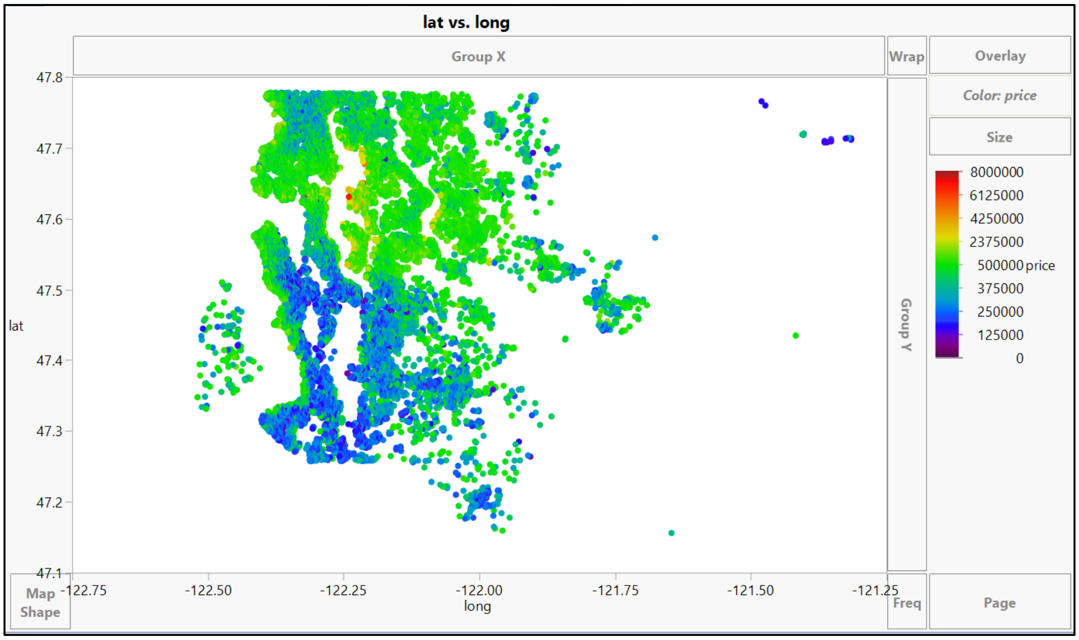
In addition to the correlation table the following charts were created:

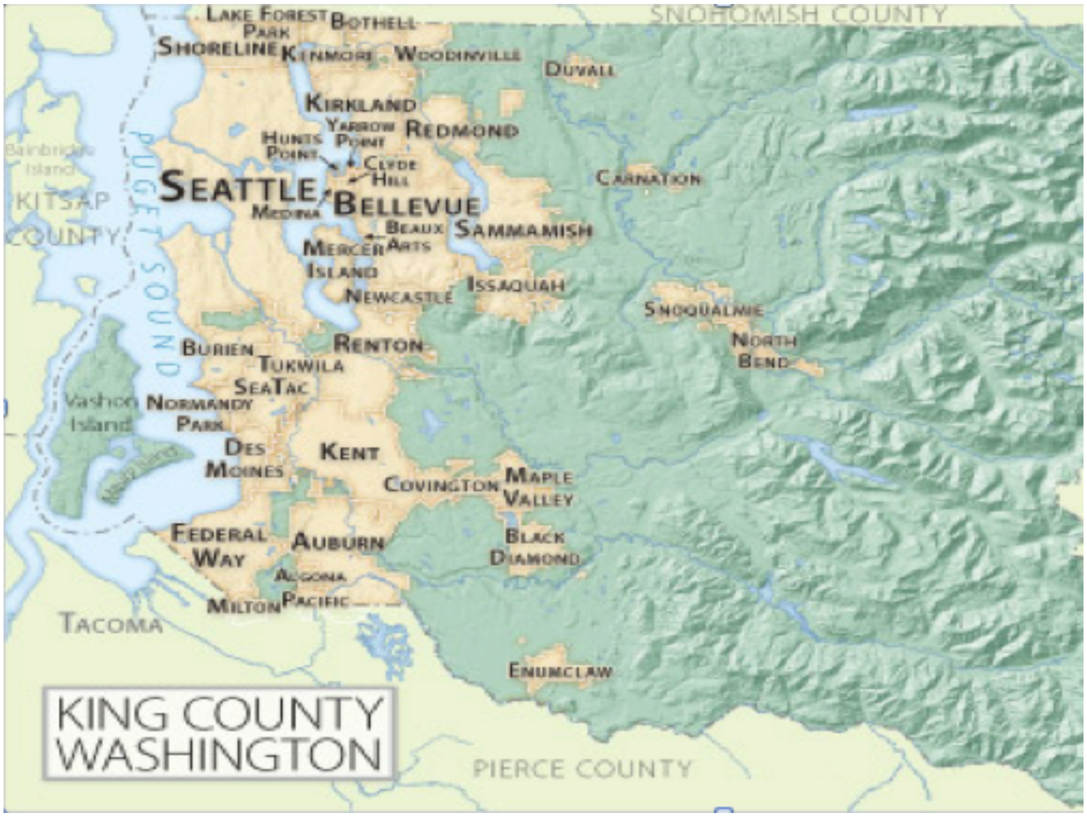
a. **Scatterplot Matrix:** The two, scatterplot matrixes confirm the inferences drawn from the correlation table.



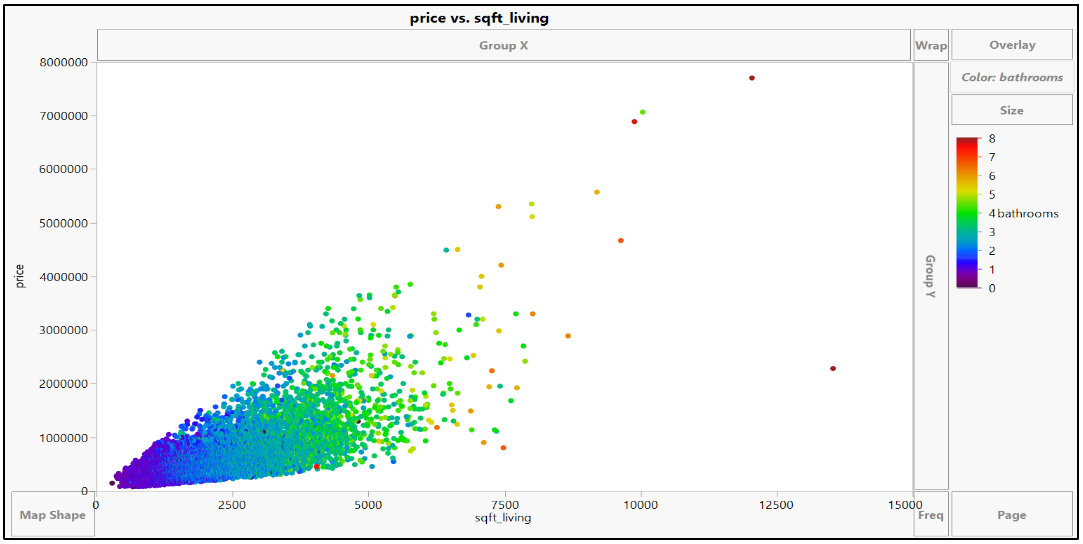


b. **Latitude Vs Longitude, Colored by Price:** The below graph is illustrative of King county region. As we juxtapose the map of king county with the below graph, we see that the areas without the color points are not incorporated cities. We observe that price increases as we move from South to North across the latitude but shows little variation as we move across the longitude. The houses near the Seattle and Bellevue along the coast in the Midwest region of King County have the highest home prices, indicating that a waterfront might mean high prices.

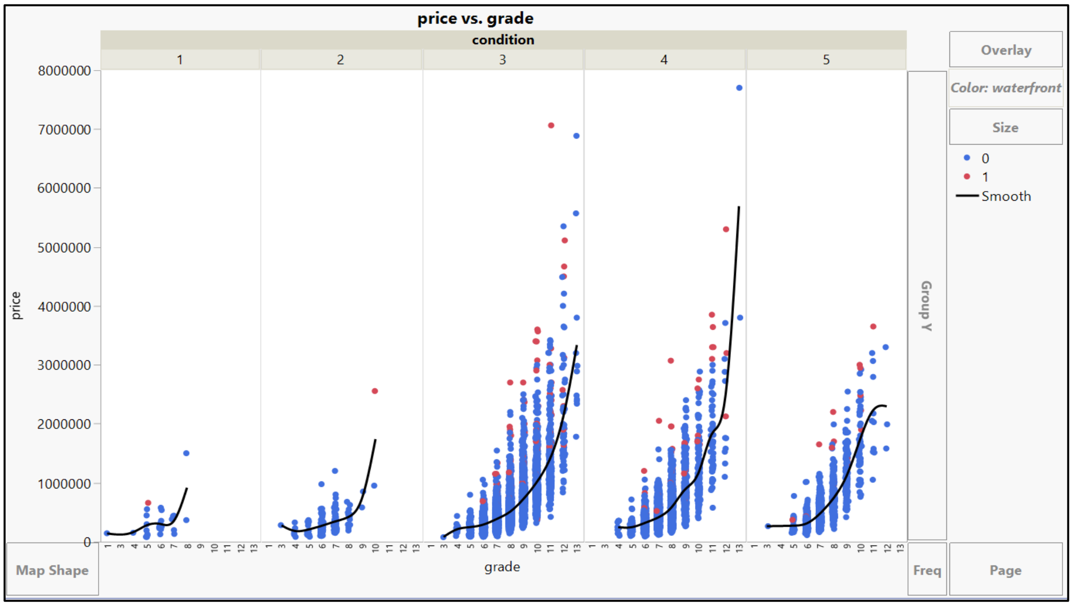

lat vs. long
Group X
Wrap
Overlay
Color: price
Size
8000000
6125000
4250000
2375000
500000 price
375000
250000
125000
0
47.8
47.7
47.6
47.5
47.4
47.3
47.2
47.1
lat
-122.75
-122.50
-122.25
-122.00
-121.75
-121.50
-121.25
long
Group Y
Map Shape
Freq
Page


Lake Forest Park
Bothell
Snohomish County
Shoreline
Kenmore
Woodinville
Duvall
Kirkland
Yarrow Point
Redmond
Hunts Point
Clyde Hill
Bainbridge Island
Kitsap County
Seattle
Medina
Bellevue
Puget Sound
Carnation
Beaux Arts
Sammamish
Mercer Island
Newcastle
Issaquah
Snoqualmie
North Bend
Burien
Renton
Tukwila
SeaTac
Vashon Island
Normandy Park
Des Moines
Kent
Covington
Maple Valley
Federal Way
Auburn
Black Diamond
Algona
Milton
Pacific
Tacoma
Enumclaw
King County Washington
Pierce County

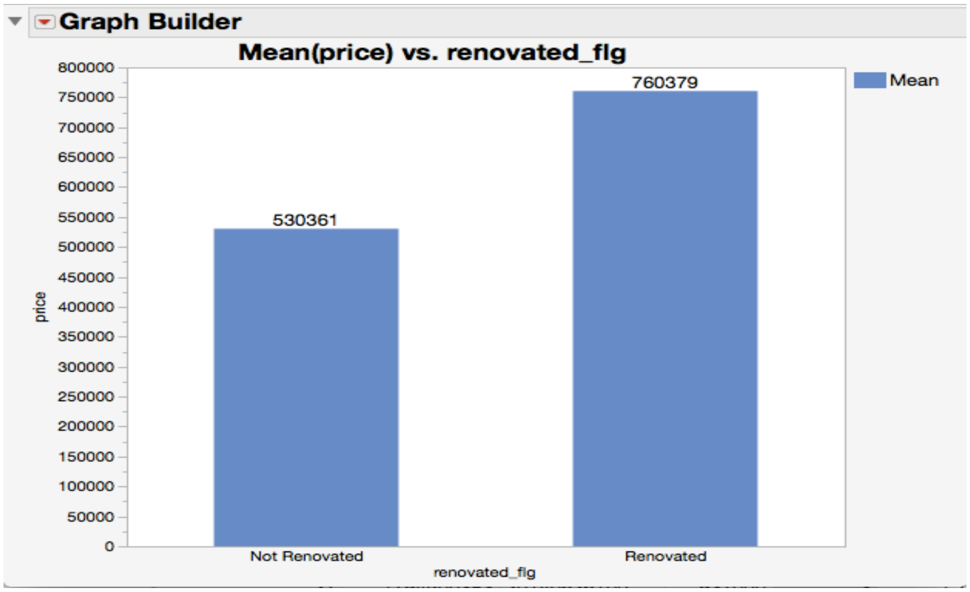
c. **Price Vs Square Feet Living, Colored by Number of Bathrooms:** The below scatterplot is indicative of house price increase with increase in square foot living and number of bathrooms. Home prices below 500,000 dollars have less than 2500 square foot living area and have less than 3 bathrooms.



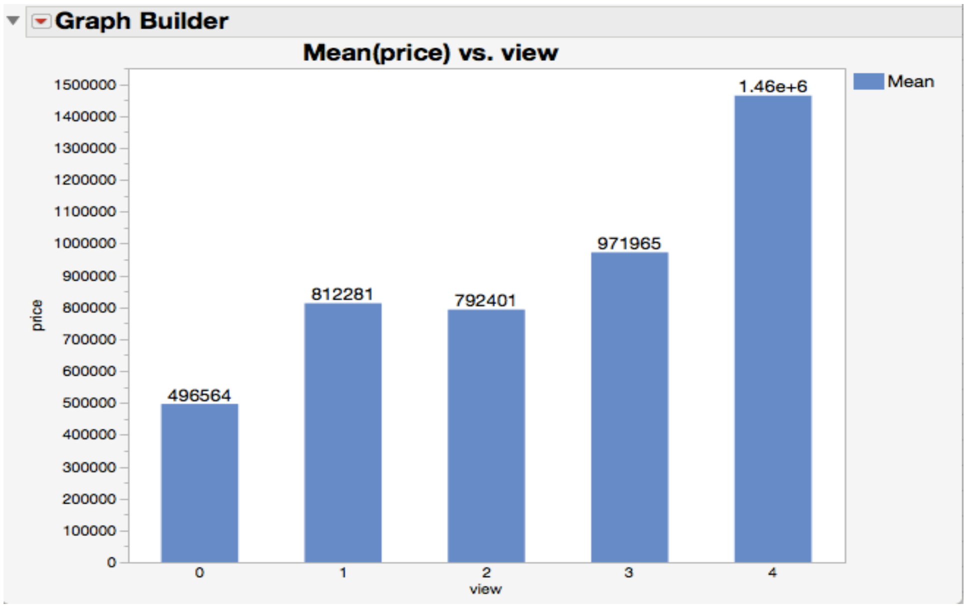
d. **Price Vs Grade, Colored by Waterfront for Different Values of Condition:** We see from the below graph that the price of the home rises with grade and condition. For a particular condition, the minimum home prices vary with grades. We also witness that there is an increase in price as the condition of the house increases from {1,2} to {3,4,5}. The presence of waterfront also improves the house prices as we observe that most of the houses with waterfront have higher home prices.



e. **Mean (Price) Vs Renovated_Flg:** From the below column chart it can be observed that the average house price associated with renovated houses is higher than the average house price associated with non-renovated house.



f. **Mean (Price) Vs View:** From the below column chart it can be observed that the average house price associated with View 3 and View 4 is much higher than the average house price associated with View 1 and 2, while the average house price associated with View 1 is the lowest.



IV. Predictive Modeling

The objective of this project is to predict house prices in King County by deploying a prediction model that accepts as inputs, variables that significantly influence the price. We believe that by solving this problem we will help prospective homeowners, developers, investors and other participants of real estate market, such as, mortgage lenders and insurers. Traditionally, the price of houses has been driven by availability of credit and sentiment and not shortage of houses. Therefore, the availability of a house price prediction model helps fill up an important information gap and improve the efficiency of the real estate market.

1. Models

We developed the following models:

1. Standard Least Squares
2. Stepwise Forward with Max Validation RSquare Stopping rule
3. Recursive Partitioning Model
4. Neural Network Model

The dataset was split into Training and Validation. All models were trained on the Training dataset and were evaluated on the basis of their performance on the validation dataset.

2. Model Results

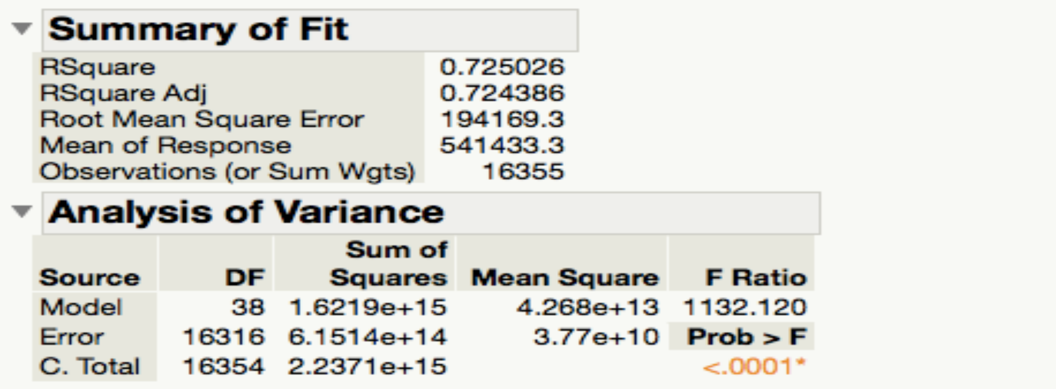
1. **Standard Least Squares:** Variable selection was done on the basis of the relationships between house features and price revealed during data visualization. Variables deemed even slightly important in predicting price were included in the model. After careful examination of change in RSquare and RMSE for training and validation datasets variables were manually removed or added.

   Summary:

   - From the summary of fit, it can be inferred that the model is a good fit. RSquare value indicates that 72.5% of the variation in price is explained by the predictors. Adjusted RSquare value is very close to RSquare, indicating that all predictors contribute in explaining the variation in price.

- F Statistic is much lesser than 0.05 indicating that there is statistical evidence that at least one of the predictors has a linear relationship with price at 95% confidence level.
- P Values corresponding to the predictors indicate that there is statistical evidence that predictors have a linear relationship with price at 95% confidence level.
- Crossvalidation suggests an improved RSquare and lower Root average square error for validation dataset, indicating that the model retains its predicting power when applied to validation dataset.

Summary of Fit:



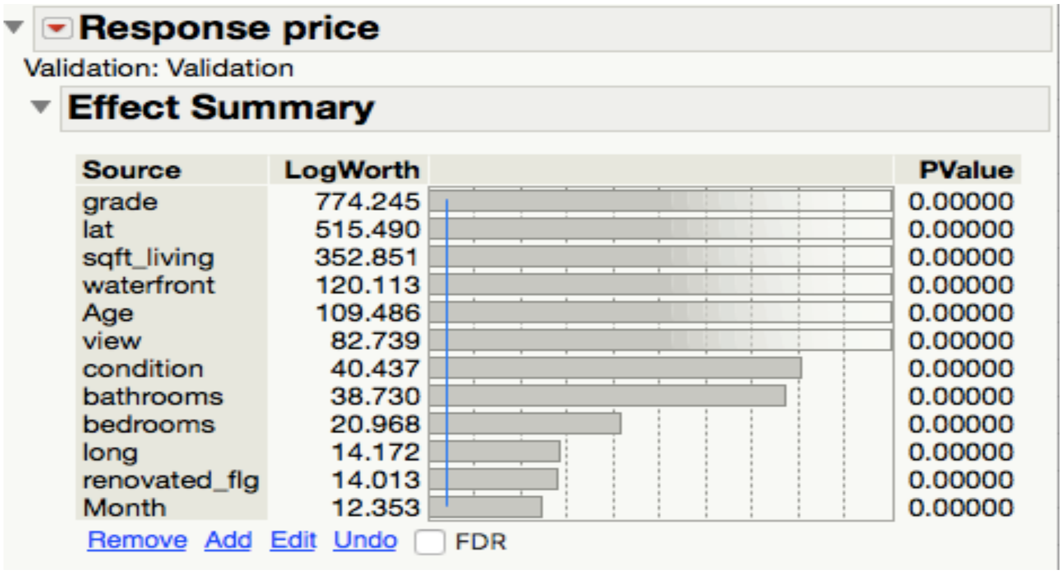
Summary of Fit

RSquare	0.725026
RSquare Adj	0.724386
Root Mean Square Error	194169.3
Mean of Response	541433.3
Observations (or Sum Wgts)	16355

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	38	1.6219e+15	4.268e+13	1132.120
Error	16316	6.1514e+14	3.77e+10	Prob > F
C. Total	16354	2.2371e+15		<.0001*

Effect Summary:



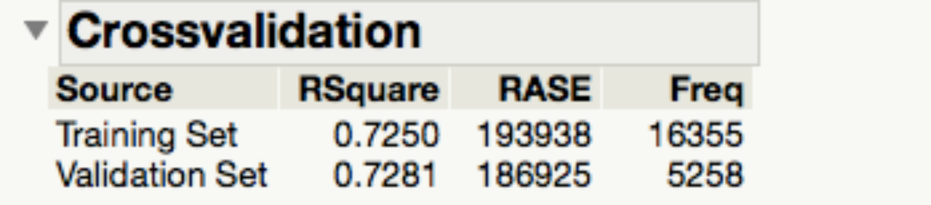
Response price

Validation: Validation

Effect Summary

Source	LogWorth	PValue
grade	774.245	0.00000
lat	515.490	0.00000
sqft_living	352.851	0.00000
waterfront	120.113	0.00000
Age	109.486	0.00000
view	82.739	0.00000
condition	40.437	0.00000
bathrooms	38.730	0.00000
bedrooms	20.968	0.00000
long	14.172	0.00000
renovated_flg	14.013	0.00000
Month	12.353	0.00000

Crossvalidation:



Crossvalidation

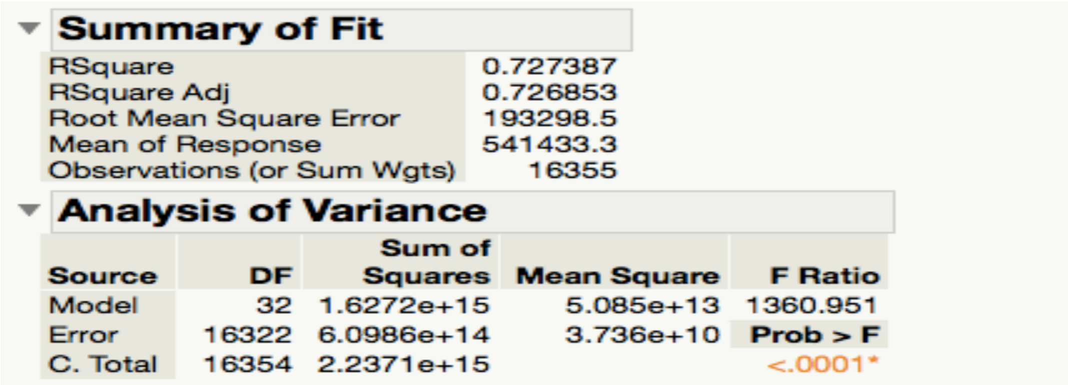
Source	RSquare	RASE	Freq
Training Set	0.7250	193938	16355
Validation Set	0.7281	186925	5258

2. **Stepwise Forward (Stopping Rule: Max Validation RSquare):** All the house features available in the data were defined as predictors. Stepwise model was run to identify model with Maximum Validation Square. The objective was to compare the best model from stepwise method with the standard least square model in terms of errors, explained variation in the outcome and the predictors.

Summary:

- From the summary of fit, it can be inferred that the model is a good fit. RSquare value indicates that 72.7% of the variation in price is explained by the predictors. Adjusted RSquare value is very close to RSquare, indicating that all predictors contribute in explaining the variation in price.
- F Statistic is much lesser than 0.05 indicating that there is statistical evidence that at least one of the predictors has a linear relationship with price at 95% confidence level.
- P Values corresponding to the predictors indicate that there is statistical evidence that most of the predictors have a linear relationship with price at 95% confidence level.
- Crossvalidation suggests an improved RSquare and lower Root average square error for validation dataset, indicating that the model retains its predicting power when applied to validation dataset.
- The stepwise model shows results very similar to the standard least square model in terms of fit, prediction accuracy and significance. The predictors identified by the stepwise model are almost same to the predictors used in Standard Least Square model.

Summary of Fit:



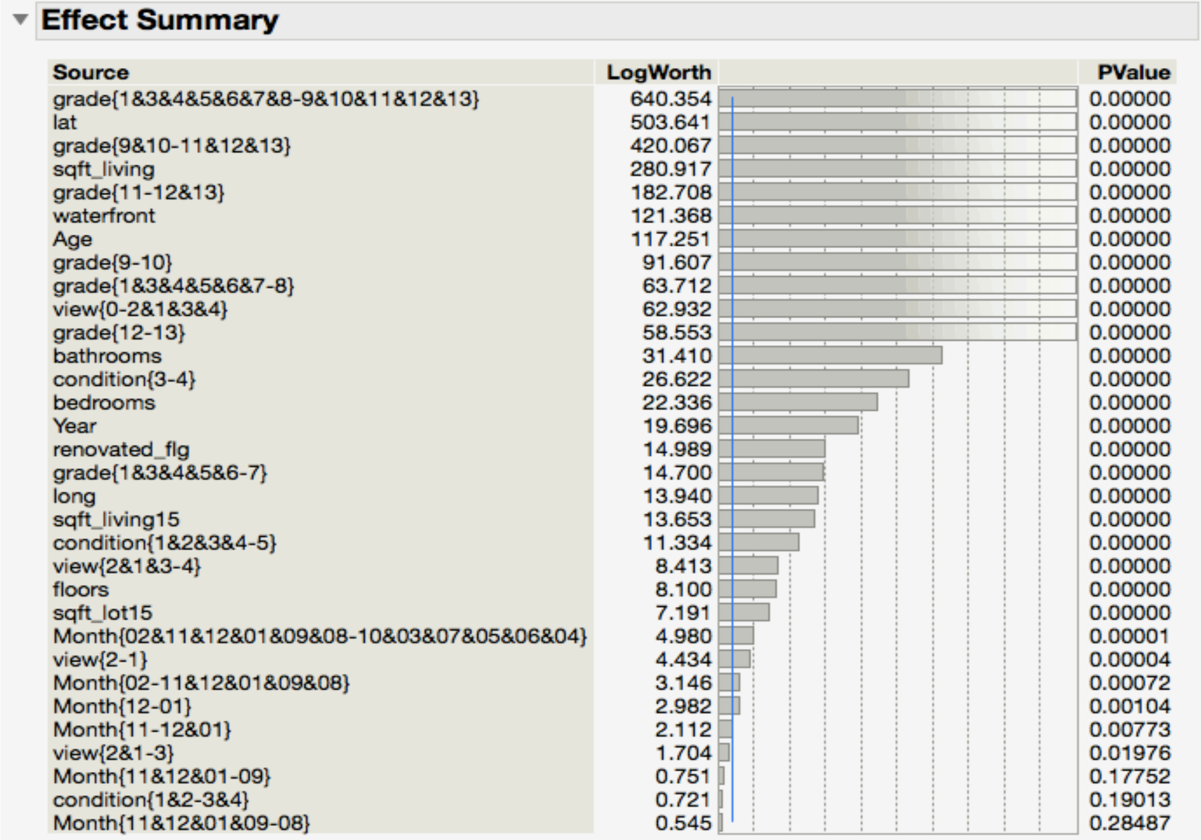
Summary of Fit

RSquare	0.727387
RSquare Adj	0.726853
Root Mean Square Error	193298.5
Mean of Response	541433.3
Observations (or Sum Wgts)	16355

Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	32	1.6272e+15	5.085e+13	1360.951
Error	16322	6.0986e+14	3.736e+10	Prob > F
C. Total	16354	2.2371e+15		<.0001*

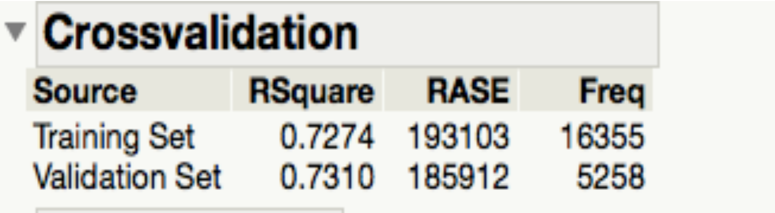
Effect Summary:



Effect Summary

Source	LogWorth	PValue
grade{1&3&4&5&6&7&8-9&10&11&12&13}	640.354	0.00000
lat	503.641	0.00000
grade{9&10-11&12&13}	420.067	0.00000
sqft_living	280.917	0.00000
grade{11-12&13}	182.708	0.00000
waterfront	121.368	0.00000
Age	117.251	0.00000
grade{9-10}	91.607	0.00000
grade{1&3&4&5&6&7-8}	63.712	0.00000
view{0-2&1&3&4}	62.932	0.00000
grade{12-13}	58.553	0.00000
bathrooms	31.410	0.00000
condition{3-4}	26.622	0.00000
bedrooms	22.336	0.00000
Year	19.696	0.00000
renovated_flg	14.989	0.00000
grade{1&3&4&5&6-7}	14.700	0.00000
long	13.940	0.00000
sqft_living15	13.653	0.00000
condition{1&2&3&4-5}	11.334	0.00000
view{2&1&3-4}	8.413	0.00000
floors	8.100	0.00000
sqft_lot15	7.191	0.00000
Month{02&11&12&01&09&08-10&03&07&05&06&04}	4.980	0.00001
view{2-1}	4.434	0.00004
Month{02-11&12&01&09&08}	3.146	0.00072
Month{12-01}	2.982	0.00104
Month{11-12&01}	2.112	0.00773
view{2&1-3}	1.704	0.01976
Month{11&12&01-09}	0.751	0.17752
condition{1&2-3&4}	0.721	0.19013
Month{11&12&01&09-08}	0.545	0.28487

Crossvalidation:



Crossvalidation

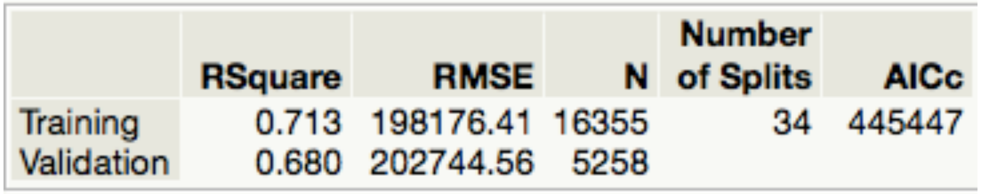
Source	RSquare	RASE	Freq
Training Set	0.7274	193103	16355
Validation Set	0.7310	185912	5258

3. **Recursive Partitioning Model:** A recursive partitioning model was run with all the house features. The objective of running a decision tree model was to identify the variables with most prediction power.

Summary:

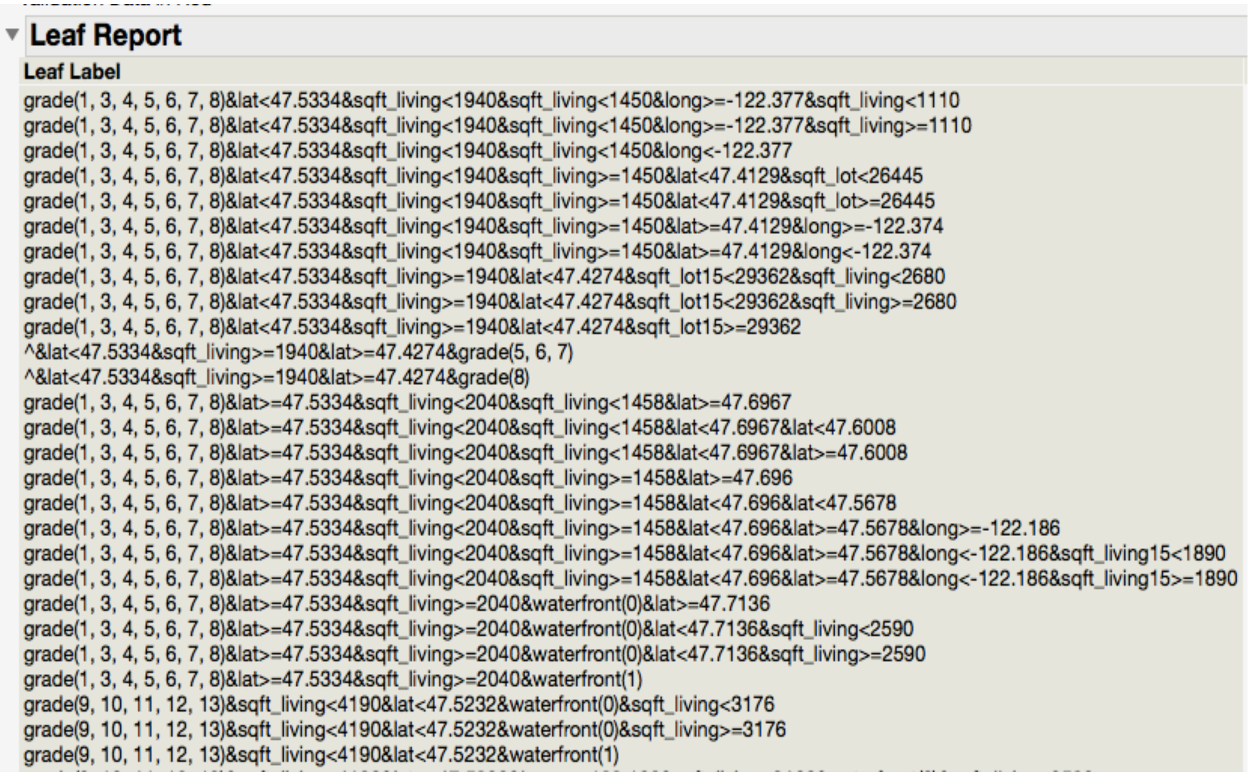
- The RSquare value for the training and validation dataset suggests that the recursive-partitioning model performs much worse on the validation data losing its prediction power from the training dataset.
- The rules in the leaf report use predictors similar to the stepwise model and standard least model confirming that these predictors have a strong prediction potential.

Crossvalidation:



	RSquare	RMSE	N	Number of Splits	AICc
Training	0.713	198176.41	16355	34	445447
Validation	0.680	202744.56	5258		

Leaf Report:



▼ Leaf Report

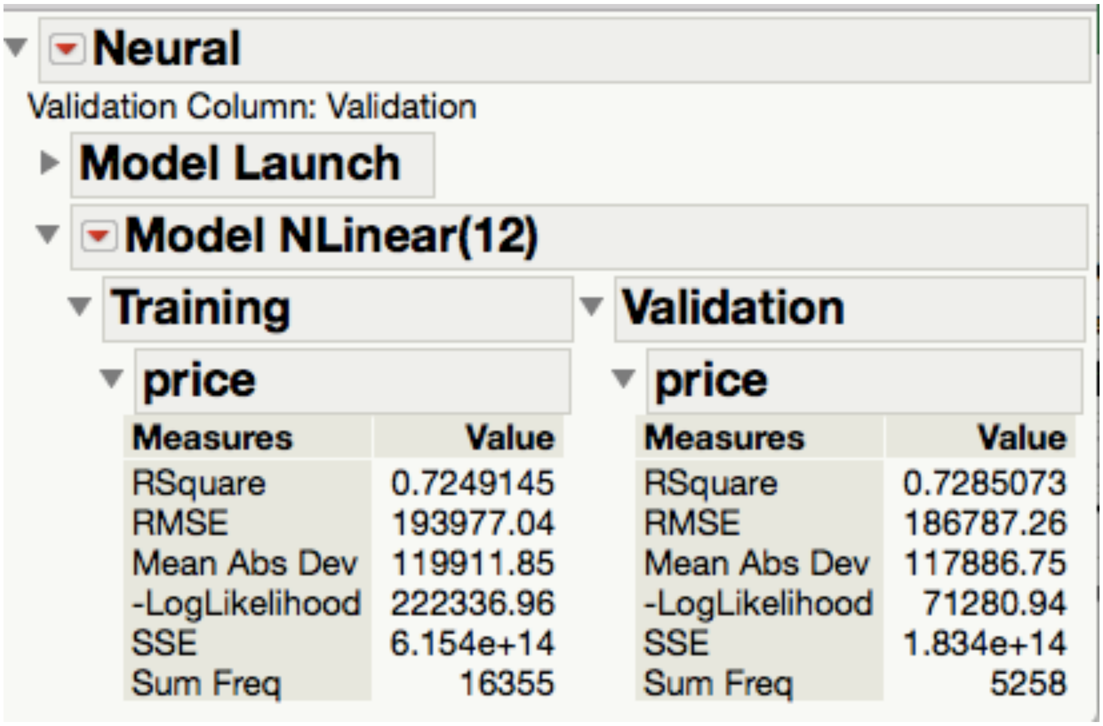
```
Leaf Label
grade(1, 3, 4, 5, 6, 7, 8)&lat<47.5334&sqft_living<1940&sqft_living<1450&long>=-122.377&sqft_living<1110
grade(1, 3, 4, 5, 6, 7, 8)&lat<47.5334&sqft_living<1940&sqft_living<1450&long>=-122.377&sqft_living>=1110
grade(1, 3, 4, 5, 6, 7, 8)&lat<47.5334&sqft_living<1940&sqft_living<1450&long<-122.377
grade(1, 3, 4, 5, 6, 7, 8)&lat<47.5334&sqft_living<1940&sqft_living>=1450&lat<47.4129&sqft_lot<26445
grade(1, 3, 4, 5, 6, 7, 8)&lat<47.5334&sqft_living<1940&sqft_living>=1450&lat<47.4129&sqft_lot>=26445
grade(1, 3, 4, 5, 6, 7, 8)&lat<47.5334&sqft_living<1940&sqft_living>=1450&lat>=47.4129&long>=-122.374
grade(1, 3, 4, 5, 6, 7, 8)&lat<47.5334&sqft_living<1940&sqft_living>=1450&lat>=47.4129&long<-122.374
grade(1, 3, 4, 5, 6, 7, 8)&lat<47.5334&sqft_living>=1940&lat<47.4274&sqft_lot15<29362&sqft_living<2680
grade(1, 3, 4, 5, 6, 7, 8)&lat<47.5334&sqft_living>=1940&lat<47.4274&sqft_lot15<29362&sqft_living>=2680
grade(1, 3, 4, 5, 6, 7, 8)&lat<47.5334&sqft_living>=1940&lat<47.4274&sqft_lot15>=29362
^&lat<47.5334&sqft_living>=1940&lat>=47.4274&grade(5, 6, 7)
^&lat<47.5334&sqft_living>=1940&lat>=47.4274&grade(8)
grade(1, 3, 4, 5, 6, 7, 8)&lat>=47.5334&sqft_living<2040&sqft_living<1458&lat>=47.6967
grade(1, 3, 4, 5, 6, 7, 8)&lat>=47.5334&sqft_living<2040&sqft_living<1458&lat<47.6967&lat<47.6008
grade(1, 3, 4, 5, 6, 7, 8)&lat>=47.5334&sqft_living<2040&sqft_living<1458&lat<47.6967&lat>=47.6008
grade(1, 3, 4, 5, 6, 7, 8)&lat>=47.5334&sqft_living<2040&sqft_living>=1458&lat>=47.696
grade(1, 3, 4, 5, 6, 7, 8)&lat>=47.5334&sqft_living<2040&sqft_living>=1458&lat<47.696&lat<47.5678
grade(1, 3, 4, 5, 6, 7, 8)&lat>=47.5334&sqft_living<2040&sqft_living>=1458&lat<47.696&lat>=47.5678&long>=-122.186
grade(1, 3, 4, 5, 6, 7, 8)&lat>=47.5334&sqft_living<2040&sqft_living>=1458&lat<47.696&lat>=47.5678&long<-122.186&sqft_living15<1890
grade(1, 3, 4, 5, 6, 7, 8)&lat>=47.5334&sqft_living<2040&sqft_living>=1458&lat<47.696&lat>=47.5678&long<-122.186&sqft_living15>=1890
grade(1, 3, 4, 5, 6, 7, 8)&lat>=47.5334&sqft_living>=2040&waterfront(0)&lat>=47.7136
grade(1, 3, 4, 5, 6, 7, 8)&lat>=47.5334&sqft_living>=2040&waterfront(0)&lat<47.7136&sqft_living<2590
grade(1, 3, 4, 5, 6, 7, 8)&lat>=47.5334&sqft_living>=2040&waterfront(0)&lat<47.7136&sqft_living>=2590
grade(1, 3, 4, 5, 6, 7, 8)&lat>=47.5334&sqft_living>=2040&waterfront(1)
grade(9, 10, 11, 12, 13)&sqft_living<4190&lat<47.5232&waterfront(0)&sqft_living<3176
grade(9, 10, 11, 12, 13)&sqft_living<4190&lat<47.5232&waterfront(0)&sqft_living>=3176
grade(9, 10, 11, 12, 13)&sqft_living<4190&lat<47.5232&waterfront(1)
```

4. **Neural Model:** A Linear neural model with one layer and 12 nodes was run using variables identified through visualization, stepwise method and recursive partitioning.

Summary:

- The RSquare value for training suggests that 72.4% of the variation in the outcome is explained by the predictors.
- RSquare for the validation dataset is 72.8%, slightly higher than the RSquare for training dataset. RMSE for validation is slightly lower than the RMSE for training dataset, indicating that the model retains its predicting accuracy for the validation dataset.

Model Measures:



Neural

Validation Column: Validation

Model Launch

Model NLinear(12)

Training

price

Measures	Value
RSquare	0.7249145
RMSE	193977.04
Mean Abs Dev	119911.85
-LogLikelihood	222336.96
SSE	6.154e+14
Sum Freq	16355

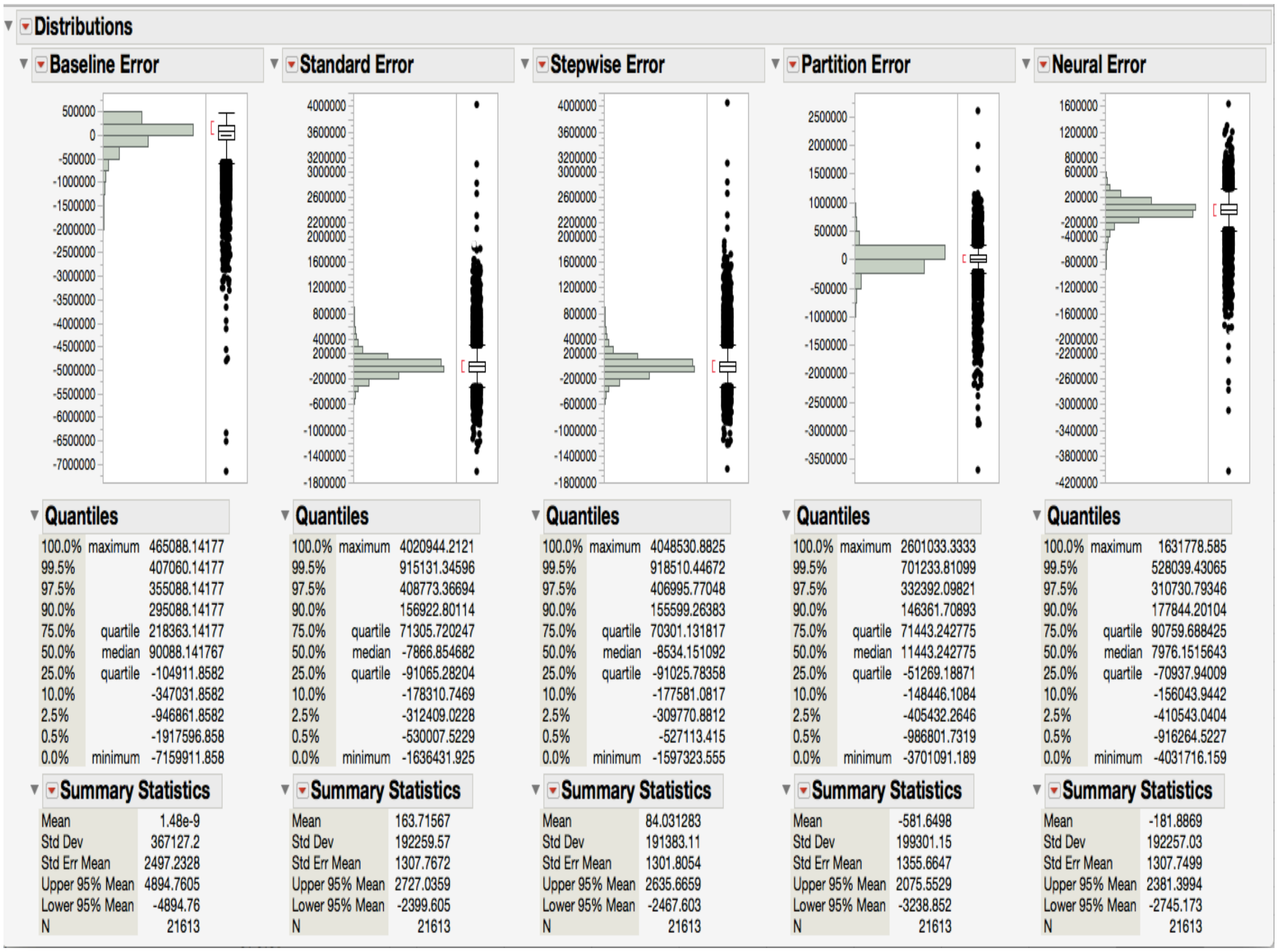
Validation

price

Measures	Value
RSquare	0.7285073
RMSE	186787.26
Mean Abs Dev	117886.75
-LogLikelihood	71280.94
SSE	1.834e+14
Sum Freq	5258

3. **Calculation of Errors:** After running the models, errors were calculated for all the versions and compared.

Absolute Error Comparison:



Quantiles

		Baseline Error	Standard Error	Stepwise Error	Partition Error	Neural Error
100.0%	maximum	465088.14177	4020944.2121	4048530.8825	2601033.3333	1631778.585
99.5%		407060.14177	915131.34596	918510.44672	701233.81099	528039.43065
97.5%		355088.14177	408773.36694	406995.77048	332392.09821	310730.79346
90.0%		295088.14177	156922.80114	155599.26383	146361.70893	177844.20104
75.0%	quartile	218363.14177	71305.720247	70301.131817	71443.242775	90759.688425
50.0%	median	90088.141767	-7866.854682	-8534.151092	11443.242775	7976.1515643
25.0%	quartile	-104911.8582	-91065.28204	-91025.78358	-51269.18871	-70937.94009
10.0%		-347031.8582	-178310.7469	-177581.0817	-148446.1084	-156043.9442
2.5%		-946861.8582	-312409.0228	-309770.8812	-405432.2646	-410543.0404
0.5%		-1917596.858	-530007.5229	-527113.415	-986801.7319	-916264.5227
0.0%	minimum	-7159911.858	-1636431.925	-1597323.555	-3701091.189	-4031716.159

Summary Statistics

	Baseline Error	Standard Error	Stepwise Error	Partition Error	Neural Error
Mean	1.48e-9	163.71567	84.031283	-581.6498	-181.8869
Std Dev	367127.2	192259.57	191383.11	199301.15	192257.03
Std Err Mean	2497.2328	1307.7672	1301.8054	1355.6647	1307.7499
Upper 95% Mean	4894.7605	2727.0359	2635.6659	2075.5529	2381.3994
Lower 95% Mean	-4894.76	-2399.605	-2467.603	-3238.852	-2745.173
N	21613	21613	21613	21613	21613

- Absolute Errors for all the versions of the model are normally distributed with mean very close to 0.
- Hence, it can be concluded that all versions qualify the test of residuals normality.

Relative Error Comparison:

Model Type	Relative Error				
	Baseline	Standard Least Square	Stepwise	Partition	Neural
Training	0.537	0.232	0.232	0.198	0.232
Validation	0.53	0.23	0.23	0.2	0.229

- All versions of the model reduce baseline error by 50% or more, indicating that selecting a model from these versions makes sense.
- Relative Error for Standard least square, Stepwise, and Neural is almost same for both training and validation.

- Relative Error is the smallest for the partition model since it predicts very few distinct values of price that are averages of groups, split based on decision tree rules.

V. Model Implementation

Model Selection: The final model from the available versions was selected based on the following.

1) The model should maintain its performance when applied to new data.

    - Looking at the cross validation reports for all the versions of the model, we can infer that Standard Least Square model, Stepwise model, and Neural model perform better on the validation dataset in terms of explaining variability in price and minimizing the error.
    - Recursive Partitioning model performs much worse on the validation dataset compared to the training dataset.
    - Hence, we can be confident that all the other versions except the recursive partitioning will maintain their performance when applied to new data.

2) The model should add value compared to “no model, baseline” scenario and have good predicting accuracy.

    - Looking at the error comparison report, we can infer that all the versions of the model perform better than the “no model, baseline” scenario reducing the relative error by almost 50%. In this case, it definitely makes sense to select and implement a model.
    - All the versions exhibit same predicting accuracy with relative error almost equal in all cases.

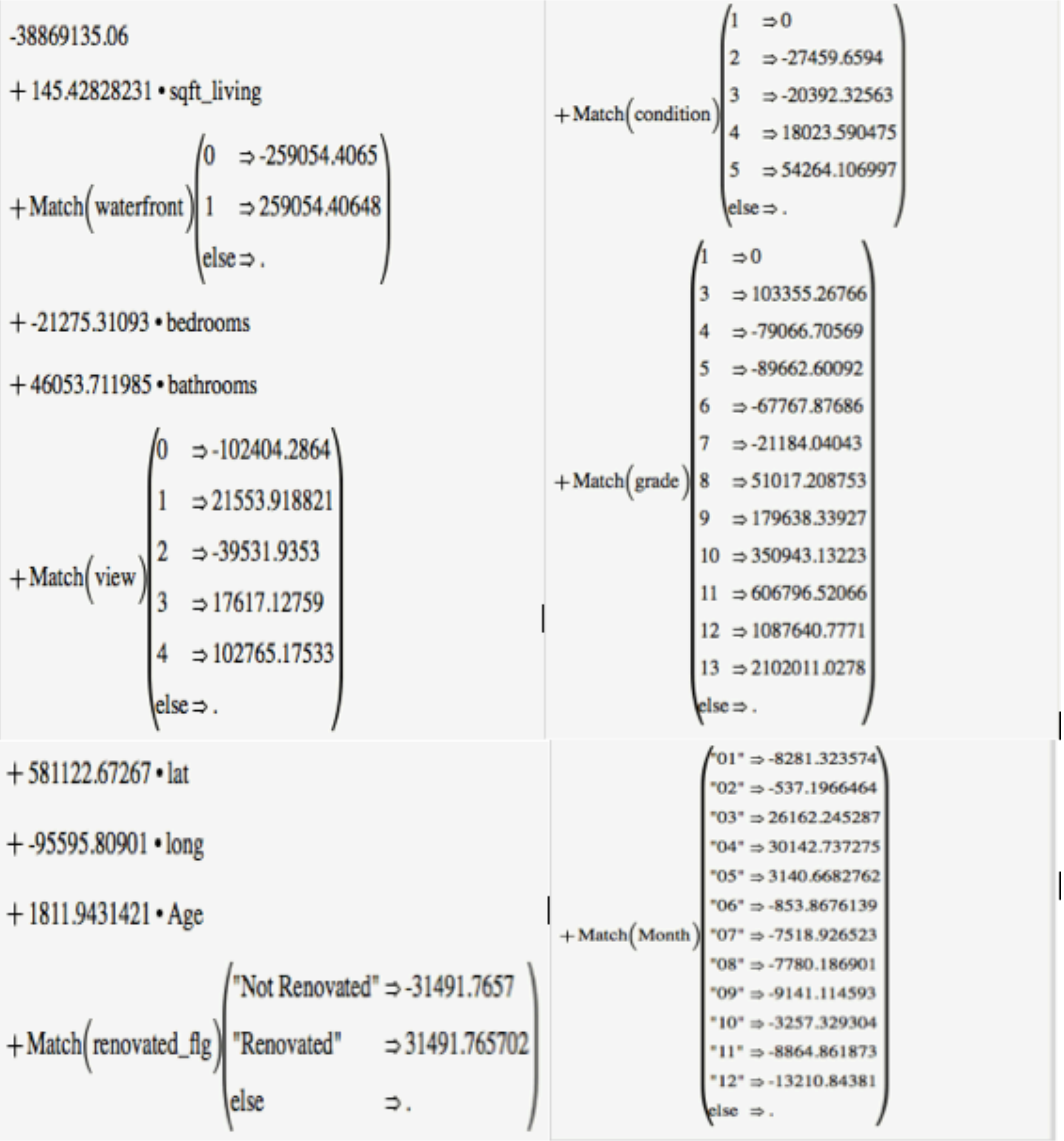
3) The model should be parsimonious and simple.

    - Standard least square model is the most simple model because of the least number of predictors it uses.

- Neural Model will not be a good selection in this case as it is expensive to implement.

On the basis of the model results and our criteria for model selection, we recommend the Standard Least Square Model to be implemented.

Model Description: The prediction formula for the Standard Least Square model is as follows:



$$-38869135.06$$

$$+145.42828231 \bullet \text{sqft\_living}$$

$$+\text{Match}\left(\text{waterfront}\right)\begin{pmatrix} 0 & \Rightarrow -259054.4065 \\ 1 & \Rightarrow 259054.40648 \\ \text{else} \Rightarrow . & \end{pmatrix}$$

$$+-21275.31093 \bullet \text{bedrooms}$$

$$+46053.711985 \bullet \text{bathrooms}$$

$$+\text{Match}\left(\text{view}\right)\begin{pmatrix} 0 & \Rightarrow -102404.2864 \\ 1 & \Rightarrow 21553.918821 \\ 2 & \Rightarrow -39531.9353 \\ 3 & \Rightarrow 17617.12759 \\ 4 & \Rightarrow 102765.17533 \\ \text{else} \Rightarrow . & \end{pmatrix}$$

$$+\text{Match}\left(\text{condition}\right)\begin{pmatrix} 1 & \Rightarrow 0 \\ 2 & \Rightarrow -27459.6594 \\ 3 & \Rightarrow -20392.32563 \\ 4 & \Rightarrow 18023.590475 \\ 5 & \Rightarrow 54264.106997 \\ \text{else} \Rightarrow . & \end{pmatrix}$$

$$+\text{Match}\left(\text{grade}\right)\begin{pmatrix} 1 & \Rightarrow 0 \\ 3 & \Rightarrow 103355.26766 \\ 4 & \Rightarrow -79066.70569 \\ 5 & \Rightarrow -89662.60092 \\ 6 & \Rightarrow -67767.87686 \\ 7 & \Rightarrow -21184.04043 \\ 8 & \Rightarrow 51017.208753 \\ 9 & \Rightarrow 179638.33927 \\ 10 & \Rightarrow 350943.13223 \\ 11 & \Rightarrow 606796.52066 \\ 12 & \Rightarrow 1087640.7771 \\ 13 & \Rightarrow 2102011.0278 \\ \text{else} \Rightarrow . & \end{pmatrix}$$

$$+581122.67267 \bullet \text{lat}$$

$$+-95595.80901 \bullet \text{long}$$

$$+1811.9431421 \bullet \text{Age}$$

$$+\text{Match}\left(\text{renovated\_flg}\right)\begin{pmatrix} \text{"Not Renovated"} & \Rightarrow -31491.7657 \\ \text{"Renovated"} & \Rightarrow 31491.765702 \\ \text{else} & \Rightarrow . \end{pmatrix}$$

$$+\text{Match}\left(\text{Month}\right)\begin{pmatrix} \text{"01"} & \Rightarrow -8281.323574 \\ \text{"02"} & \Rightarrow -537.1966464 \\ \text{"03"} & \Rightarrow 26162.245287 \\ \text{"04"} & \Rightarrow 30142.737275 \\ \text{"05"} & \Rightarrow 3140.6682762 \\ \text{"06"} & \Rightarrow -853.8676139 \\ \text{"07"} & \Rightarrow -7518.926523 \\ \text{"08"} & \Rightarrow -7780.186901 \\ \text{"09"} & \Rightarrow -9141.114593 \\ \text{"10"} & \Rightarrow -3257.329304 \\ \text{"11"} & \Rightarrow -8864.861873 \\ \text{"12"} & \Rightarrow -13210.84381 \\ \text{else} & \Rightarrow . \end{pmatrix}$$

Interpretation of Model formula: The following can be inferred from the model formula:

- Every unit increase in Square feet living, and Bathrooms will increase the predicted price. Predicted price decreases with unit increase in Bedrooms.
- The predicted price will increase with increase in latitude as the location moves from South to North and will decrease with increase in longitude as the location moves from West to East.
- Predicted price is high for houses with waterfront.
- Predicted price is high for houses with condition 4 or 5, grade greater than 8 and View 1, 3 or 4.
- Predicted price is high in the months of March, April and May because of seasonality impact.
- Predicted price increases with unit increase in age.

VI. Plan for future upgrades

- During exploratory data analysis, we inferred that the outcome variable has a lot of legitimate outliers. We believe that very high prices of houses are because of characteristics partially captured in the data. We would recommend identifying characteristics (eg: amenities, neighborhood) that undoubtedly make a house a High priced property through appropriate domain research. Based on a definition derived from these characteristics we would segment houses into luxury homes and ordinary homes. Further, we would recommend developing different models for luxury homes and ordinary homes.

- There exists evidence from the research in real estate industry that the price of the houses has relationship with other elements such as availability of credit (mortgage interest rates), consumer sentiment and other economic factors. We recommend collecting data on these elements and model should upgrade to factor the effect of these elements.

References:

https://resources.point.com/8-biggest-factors-affect-real-estate-prices/