PREDICTIVE MODEL FOR HOUSE PRICING IN KING COUNTY

BY

SURENDRA POTUPUREDDY

PROBLEM STATEMENT:

The main aim of this project is to estimate the price of houses in king county, USA based on various factors entailed to it. With the help of such analysis people and agencies can know beforehand as to how much a house with a defined set of specifications and features would cost in an area. Through this project, I will be analyzing the following:

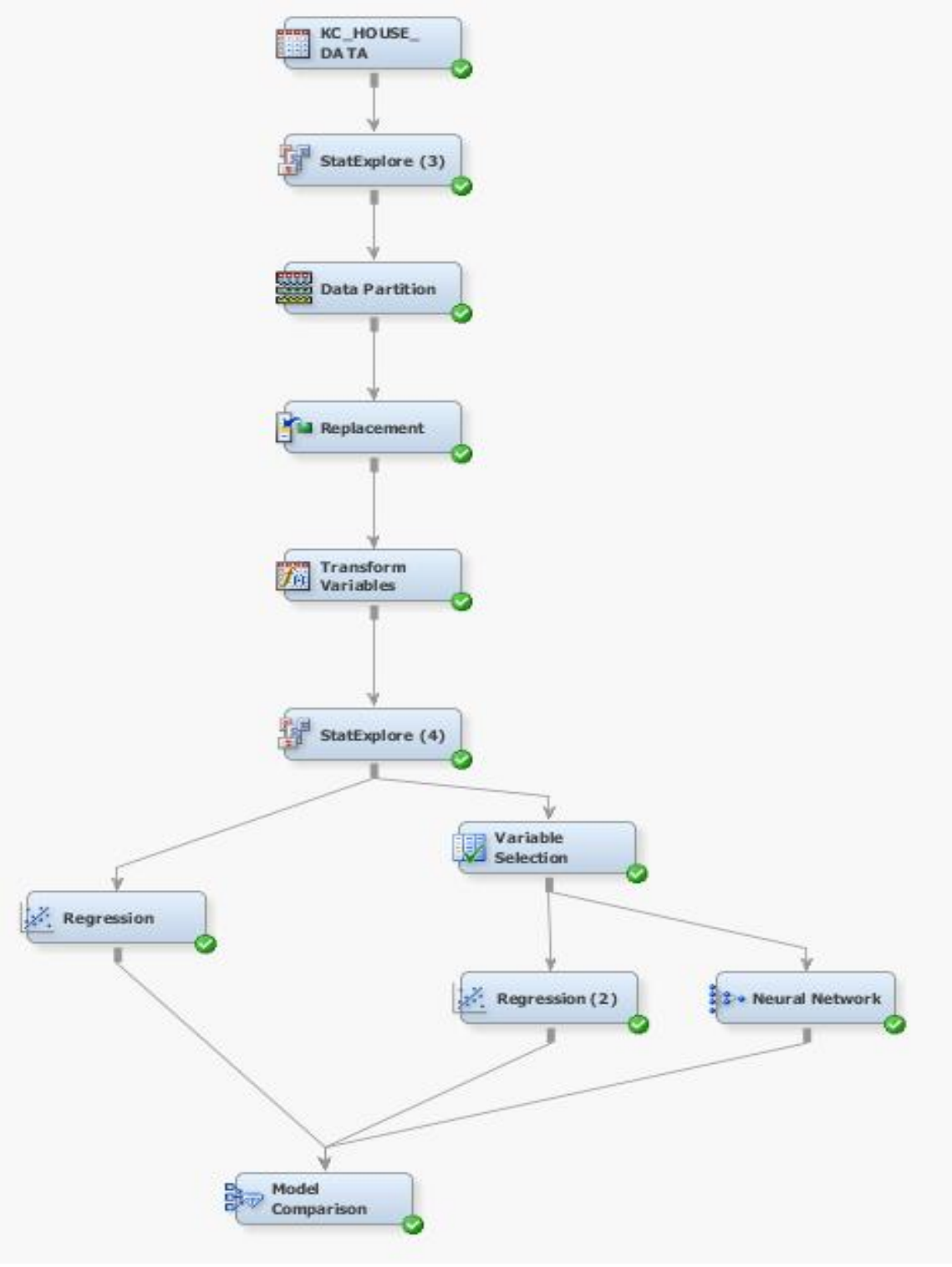
1. How much a house costs with certain features in king county, USA?
2. What are the core main factors that are influencing the house price?

DATA DESCRIPTION:

I used second hand dataset retrieved from **Kaggle** Website. This dataset contains house sale prices for King County homes sold between May 2015 and May 2016.
This dataset consists of **21613** rows **and 21** features including the unique Id column and price.

1) **Id** – Id column
2) **Date** – Date on which the house was sold
3) **Price** – Price at which the house was sold
4) **Bedrooms** – Number of bedrooms
5) **Bathrooms** – Number of bathrooms
6) **Sqft_living** – Size of the living space in Square foot
7) **Sqft_lot** – Size of the lot in Square foot
8) **Floors** – Number of floors
9) **Waterfront** – Whether the house has a waterfront or not
10) **View** - No.of views
11) **Condition** – Level of condition
12) **Grade** – Grade of the house
13) **Sqft_above** – (Sqftliving – sqft basement)
14) **Sqft_basement** – size of the basement
15) **Year_built** – Year the house was built
16) **Year_Renovated** – year the house was renovated
17) **Zipcode** – zipcode of the location of house
18) **Lat** – latitude of the house
19) **Long** – longitude of the house
20) **Sqft_lot15** – average lot size of the 15 closest houses
21) **Sqft_Living15** – average size of living space of the 15 closest houses

MODEL DIAGRAM:



DATA PRE-PROCESSING:

DATA CLEANING:

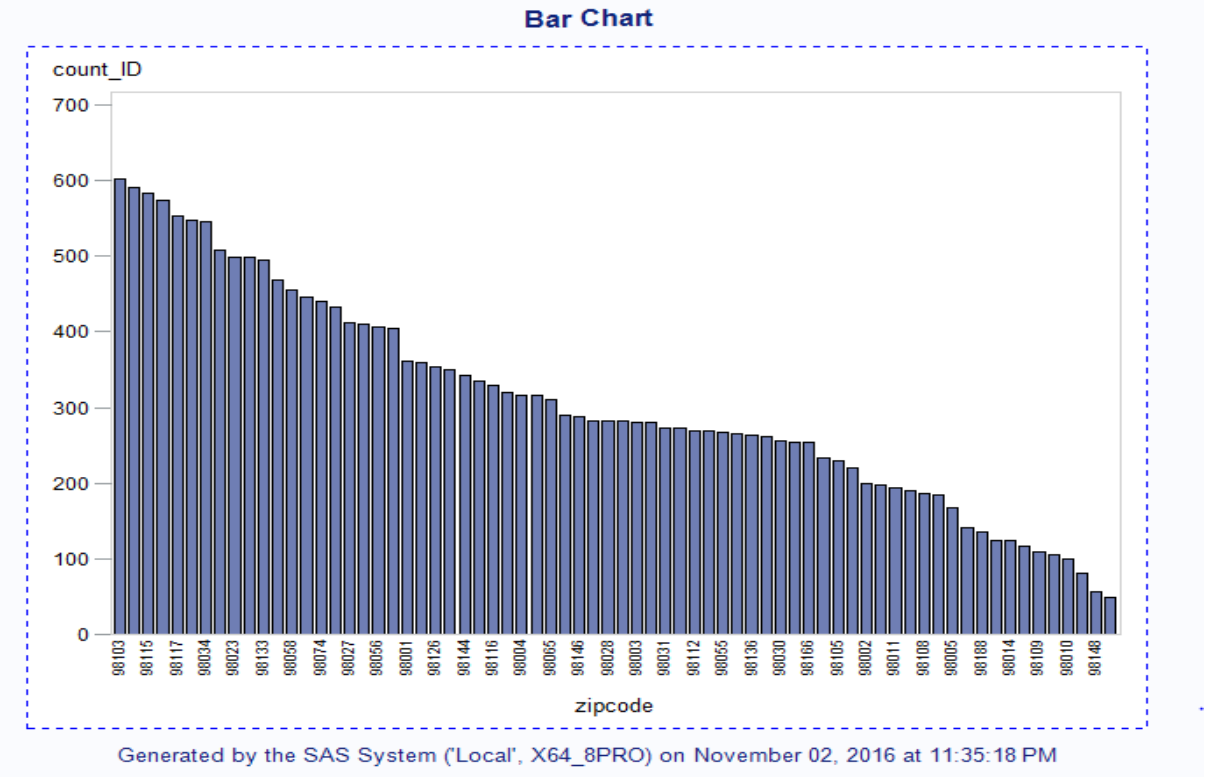
Eliminated few variables at the start of analysis as they don't play a significant role in predicting the house prices like the unique Id, latitude, longitude and Date the house was sold.

Name	Use	Report	Role	Level
bathrooms	Default	No	Input	Nominal
bedrooms	Default	No	Input	Nominal
condition	Default	No	Input	Nominal
date	Default	No	Rejected	Nominal
floors	Default	No	Input	Nominal
grade	Default	No	Input	Nominal
id	Default	No	Rejected	Nominal
lat	Default	No	Rejected	Interval
long	Default	No	Rejected	Interval
price	Default	No	Target	Interval
sqft_above	Default	No	Input	Interval
sqft_basemer	Default	No	Input	Binary
sqft_living	Default	No	Input	Interval
sqft_living15	Default	No	Input	Interval
sqft_lot	Default	No	Input	Interval
sqft_lot15	Default	No	Input	Interval
view	Default	No	Input	Nominal
waterfront	Default	No	Input	Binary
yr_built	Default	No	Input	Nominal
yr_renovated	Default	No	Input	Nominal
zipcode	Default	No	Input	Nominal

DATA EXPLORATION:

Context of the analysis:

Firstly, checked the house prices based on their Zip codes to find out insights for the problem being explored.



Have pasted one of the graphs that was used for the below analysis. I have observed that some zip codes have a lot more house sales recorded than others. The number of observations range from ~50 to ~600. We can see that on average, the houses of some zip codes are more expensive and are also bigger in sqft. The houses close to some zip codes are relatively old compared to the houses in the rural area. This observation helped in identifying the zip codes that are to be used in building my regression model.

After analyzing the houses based on location I moved on to exploring the output variable i.e., house price and then the relation between the output variable and the rest of the variables.

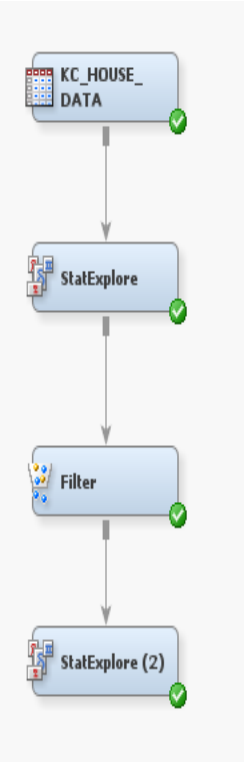
Analysis on the output variable:

I have observed that there are lot of outliers at the top of the distribution, with a few houses above the 5000000$ value and the distribution is slightly skewed to the right.

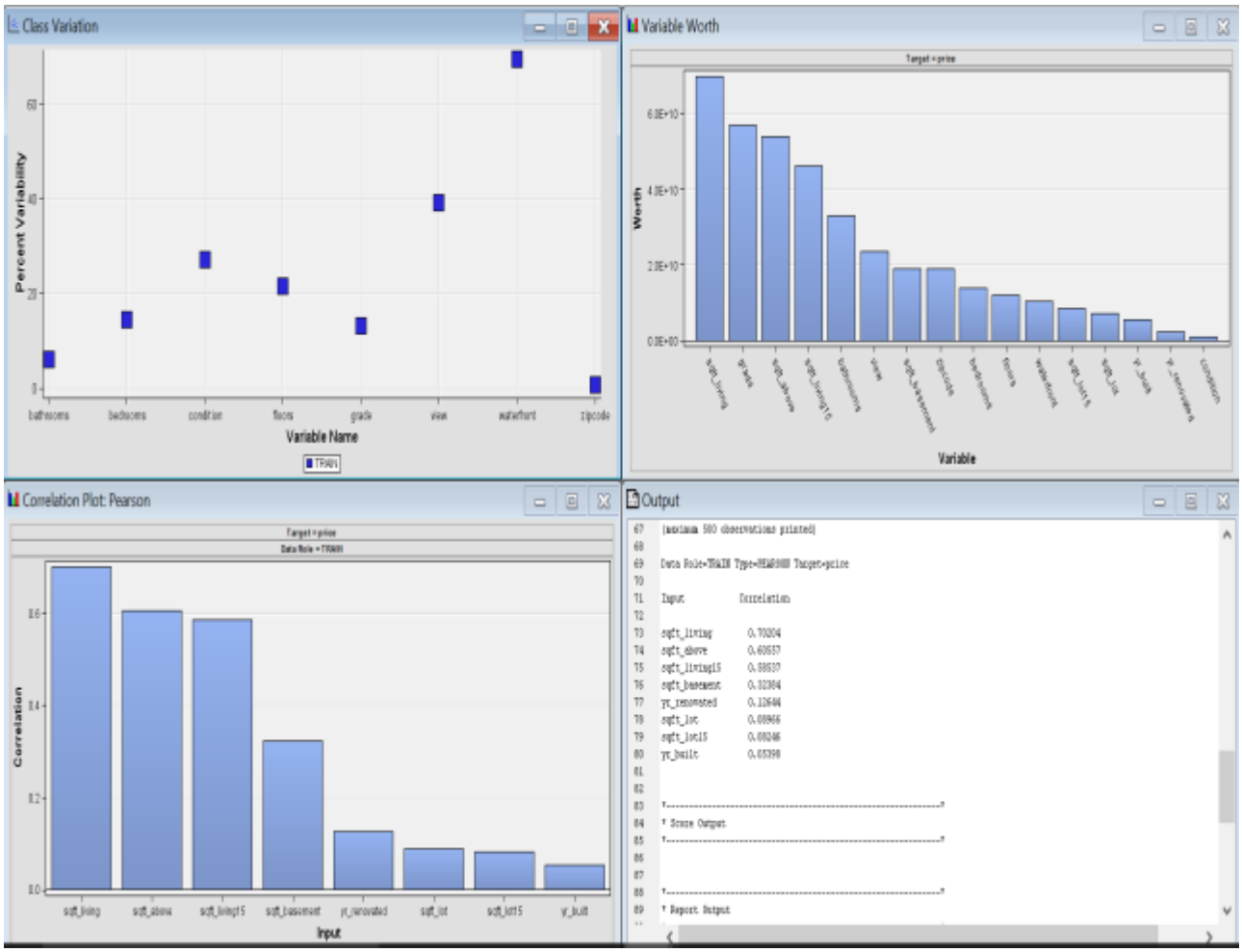
Association and Correlation between price and continuous variables:

Now analyzed the relationship between the continuous variables available in the dataset and the output variable (i.e., house price). I have used correlation coefficients to explore potential associations between the variables.

Diagram for exploring the data:



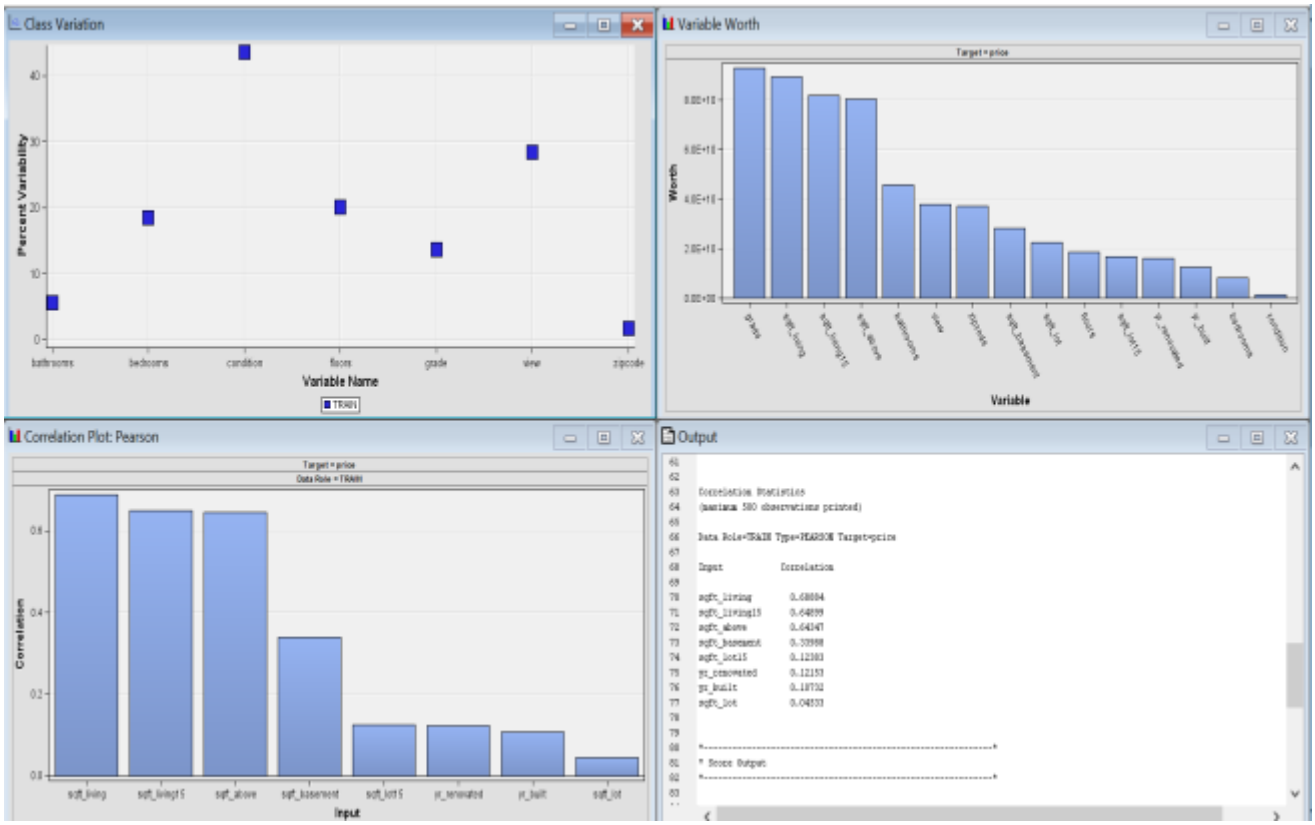
StatExplore:



There is a clear linear association between sqft_living and price (r = 0.7), indicating a strong positive relationship and is a good predicator of house price. Similarly, found the correlation coefficients for all the continuous variables.

The correlation between price and sqft_lot is 0.08966
The correlation between price and sqft_above (i.e., sqft_above = sqft_living - sqft_basement) is 0.60557
The correlation between price and sqft_basement is 0.32384
The correlation between price and sqft_living15 is 0.58537
The correlation between price and sqft_lot15 is 0.08246
The correlation between price and yr_built is 0.05398
The correlation between price and yr_renovated is 0.12644

Based on the correlation plot I concluded that sqft_lot, sqft_lot15 and yr_built seem to be poorly related to price. We can also see in our dataset that there are a lot of zeros in the sqft_basement distribution (i.e., no basement) and the yr_renovated variable.
So, I calculated the correlation coefficients for these two variables excluding the rows where their values are zeros.



Based on the analysis it was evident that the house price is moderately correlated with the size of the basement (if basement present) and there is also a small correlation with the year of the renovation (if renovated).

It might be more interesting for my analysis to classify basement and renovation as dichotomous variables (e.g., 0 for no basement, 1 for basement present).

Multicollinearity Test:

Using the scatterplots, I observed that sqft_above and sqft_living15 are strongly related to price. So, analyzed their associations (along with sqft_living) and as envisaged, there is a strong positive relationship between the 3 variables (r>0.7). It was obvious for sqft_above as it is equal to sqft_livng - sqft_basement and both have an impact on price.

For sqft_living15 however, I was not sure if the relationship with house price is due to the average square footage of the 15 closest houses because of its high correlation with sqft_living. I found the impact of sqft_living15 using the variance inflation factor(VIF) and included the column along with sqft_living in our final prediction model.

Association and Correlation between price and categorical variables:

I have observed the relation between the output variable and the categorical variables using the Variable worth and found that bedrooms, bathrooms, floors, views, grade are moderate to strongly correlated with price.

Association Conclusion:

Continuous Variables:

- sqft_living, sqft_above, sqft_basement -- moderately/strongly associated with price.
- There is a strong correlation between these 3 variables as sqft_living = sqft_above and sqft_basement.
- sqft_lot, sqft_lot15, yr_built -- poorly associated with price.

Dichotomous variables:

- waterfront, basement_present, renovated -- slightly associated with price.

Categorical variables:

- bedrooms, bathrooms, floors, views, grade -- moderate to strongly associated with price.

DATA TRANSFORMATION:

Replacement:

Year_Renovated and Sqft_basement had a lot of zeros, so transformed these variables into binary variables to derive better results.

Replaced the year_renovated which is a nominal variable into a binary variable that would designate whether the house was renovated or not.

yr_renovated	0		13449
yr_renovated	2014	1	64
yr_renovated	2000	1	28
yr_renovated	2005	1	26
yr_renovated	2007	1	25
yr_renovated	2003	1	24
yr_renovated	2013	1	23
yr_renovated	2004	1	21

Transformed the variable sqft_basement into a binary variable as well so that it would designate whether there is a basement or not.

sqft_basement	0		8547 N
sqft_basement	700	1	153 N
sqft_basement	500	1	140 N
sqft_basement	600	1	138 N
sqft_basement	400	1	127 N
sqft_basement	800	1	124 N
sqft_basement	900	1	105 N
sqft_basement	1000	1	100 N
sqft_basement	300	1	92 N
sqft_basement	480	1	82 N
sqft_basement	200	1	70 N
sqft_basement	530	1	70 N
sqft_basement	750	1	70 N

I even found few typo errors in the dataset by checking the outliers and have replaced accordingly. For ex: For one of the houses the no. of bedrooms were given as 33 where as the no. of floors is 1 and sqft_living is 1640 sqft.

Transform Variables:

Name	Method	Number of Bins	Role	Level
REP_sqft_basement	**Default**	**4**	Input	Binary
REP_yr_renovated	**Default**	**4**	Input	Nominal
bathrooms	**Default**	**4**	Input	Nominal
bedrooms	**Default**	**4**	Input	Nominal
condition	**Dummy Indicator**	**4**	Input	Nominal
floors	**Dummy Indicator**	**4**	Input	Nominal
grade	**Dummy Indicator**	**4**	Input	Nominal
price	**Default**	**4**	Target	Interval
sqft_above	**Log 10**	**4**	Input	Interval
sqft_basement	**Default**	**4**	Rejected	Binary
sqft_living	**Log 10**	**4**	Input	Interval
sqft_living15	**Log 10**	**4**	Input	Interval
sqft_lot	**Default**	**4**	Input	Interval
sqft_lot15	**Default**	**4**	Input	Interval
view	**Dummy Indicator**	**4**	Input	Nominal
waterfront	**Default**	**4**	Input	Binary
yr_built	**Default**	**4**	Input	Nominal
yr_renovated	**Default**	**4**	Rejected	Nominal
zipcode	**Dummy Indicator**	**4**	Input	Nominal

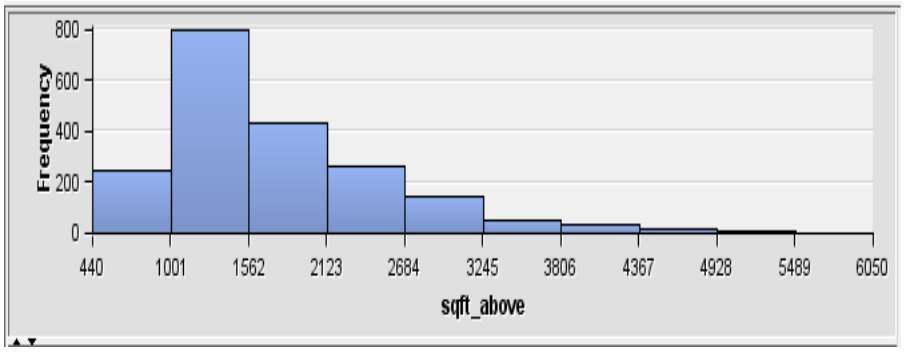
Dummy Indicator method:

A dummy indicator is used to create dummy variables(columns) to represent an attribute (class variables) with two or more distinct categories/levels. Here I was going to create dummy variables for condition, floors, grade, view and zip code.

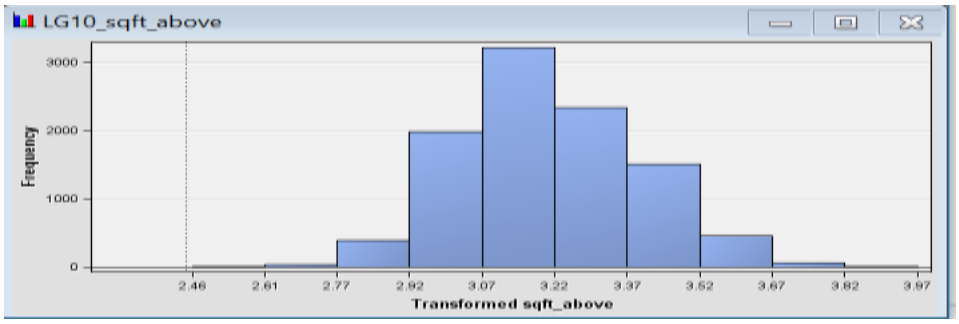
Log10 method for reducing skewness:

I have explored the data and removed the skewness of a few attributes such as Sqft_living, Sqft_living15 and sqft_above.

Skewness before transformation:



Skewness after transformation:



The below table indicates the way the dummy indicator has created columns for all the class variables, continuous variables after skewness was reduced and dichotomous variables which are replaced.

Name	Use	Report	Role	Level
LG10_sqft_above	**Default**	**No**	Input	Interval
LG10_sqft_living	**Default**	**No**	Input	Interval
LG10_sqft_living15	**Default**	**No**	Input	Interval
REP_sqft_basemen	**Default**	**No**	Input	Binary
REP_yr_renovated	**Default**	**No**	Input	Nominal
TI_condition1	**Default**	**No**	Input	Binary
TI_condition2	**Default**	**No**	Input	Binary
TI_condition3	**Default**	**No**	Input	Binary
TI_condition4	**Default**	**No**	Input	Binary
TI_condition5	**Default**	**No**	Input	Binary
TI_floors1	**Default**	**No**	Input	Binary
TI_floors2	**Default**	**No**	Input	Binary
TI_floors3	**Default**	**No**	Input	Binary
TI_floors4	**Default**	**No**	Input	Binary
TI_floors5	**Default**	**No**	Input	Binary
TI_floors6	**Default**	**No**	Input	Binary
TI_grade1	**Default**	**No**	Input	Binary
TI_grade10	**Default**	**No**	Input	Binary
TI_grade11	**Default**	**No**	Input	Binary
TI_grade12	**Default**	**No**	Input	Binary
TI_grade2	**Default**	**No**	Input	Binary
TI_grade3	**Default**	**No**	Input	Binary
TI_grade4	**Default**	**No**	Input	Binary
TI_grade5	**Default**	**No**	Input	Binary
TI_grade6	**Default**	**No**	Input	Binary
TI_grade7	**Default**	**No**	Input	Binary
TI_grade8	**Default**	**No**	Input	Binary
TI_grade9	**Default**	**No**	Input	Binary
TI_view1	**Default**	**No**	Input	Binary
TI_view2	**Default**	**No**	Input	Binary
TI_view3	**Default**	**No**	Input	Binary
TI_view4	**Default**	**No**	Input	Binary
TI_view5	**Default**	**No**	Input	Binary
TI_zipcode1	**Default**	**No**	Input	Binary
TI_zipcode10	**Default**	**No**	Input	Binary
TI_zipcode11	**Default**	**No**	Input	Binary
TI_zipcode12	**Default**	**No**	Input	Binary

I tried to fit a higher-order polynomial on the input or to include interaction variables. For example, I tried to fit a quadratic function on sqft_living and other features.

Bedrooms*bedrooms: I was doing this so that houses with more bedrooms 'get more weight' so to speak. This means houses with many bedrooms will be affected by this feature.

Bedrooms*bathrooms: This value is large when both values are large. So, houses with many bedrooms and bathrooms will also 'get more weight'.

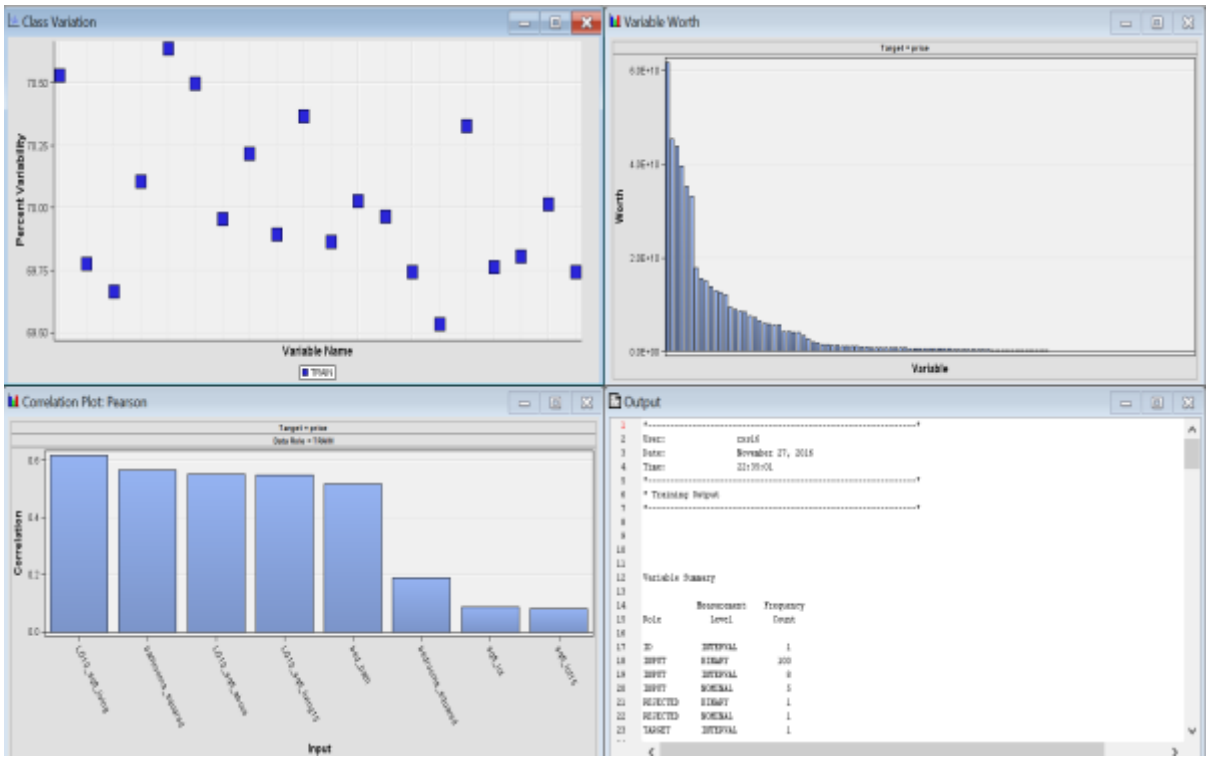
```
label title ='bed_bath';
length title $10;
bed_bath=bedrooms*bathrooms;

label title = 'bathrooms_squared';
length title $30;
bathrooms_squared=bathrooms*bathrooms;

label title = 'bedrooms_squared';
length title $30;
bedrooms_squared=bedrooms*bedrooms;
```

Stat explorer on the transformed variables:

The stat explorer has provided the correlation between price and the newly created, transformed and replaced variables. Based on the results obtained I tested multiple features combinations to find the best fit for our model.



Data Partition (Splitting the data):

Divided the data into training (60%), validation (20%) and test (20%).

Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	60.0
Validation	20.0
Test	20.0
Report	
Interval Targets	Yes
Class Targets	Yes

REGRESSION MODEL:

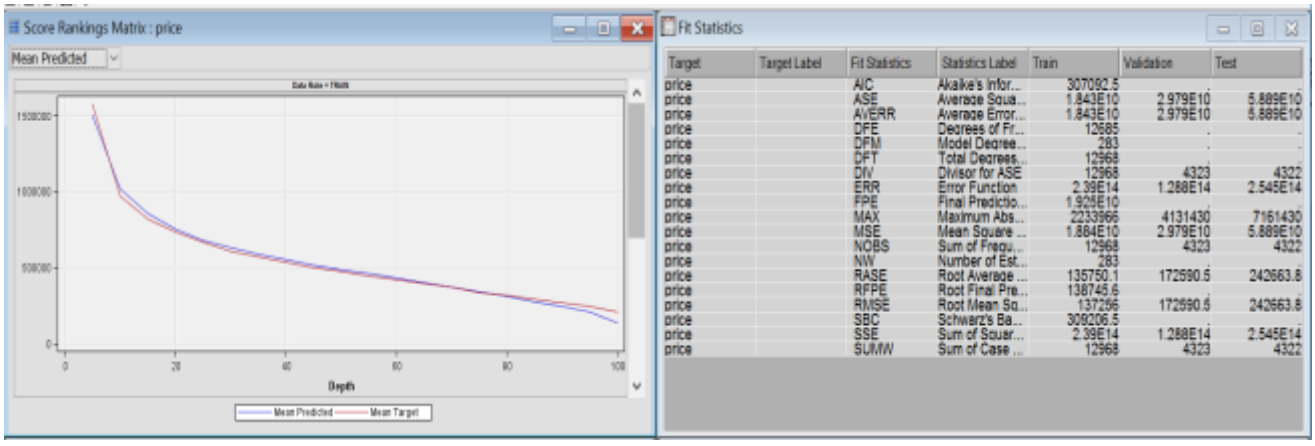
Simple Linear Regression:

I first tried to predict house prices using simple linear regression by using sqft_living as input and calculated the Root Mean Square Error on the test data. Similarly, I ran the same test on all the features in the dataset and compared the RMSE to assess the best estimator of house price. Sqft_living provided the smallest test error (RMSE: 268279.643883) indicating the best estimator of house price for the dataset considered.

Multiple Regression:

Now I tried to predict price using multiple features. Based on the results obtained using simple linear regression, used the best single estimator sqft_living along with the remaining features.
Tested all the remaining features one by one in combinations with sqft_living (e.g., sqft_living and bedrooms_squared etc) and selected the best combination using training error. At the end, selected the model complexity (number of features) using the validation error and the test error.

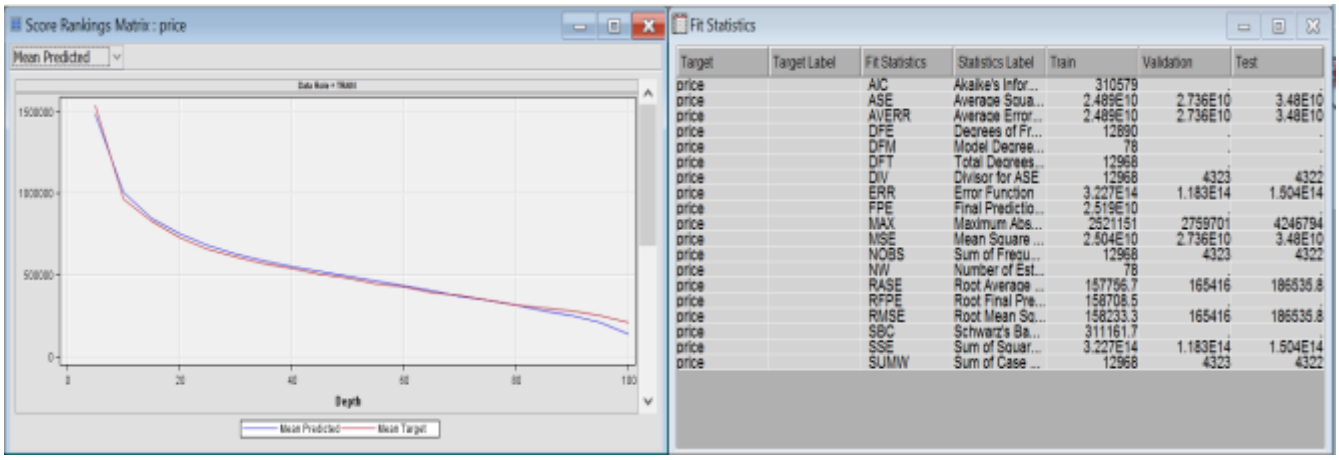
The regression result obtained when all the features were selected:



I tested multiple features combinations to find the best fit for my model and selected the below features:

Name	Use	Role	Level
LG10_sqft_living	Yes	Input	Interval
LG10_sqft_living15	Yes	Input	Interval
TI_condition4	Yes	Input	Binary
TI_condition5	Yes	Input	Binary
TI_floors1	Yes	Input	Binary
TI_grade10	Yes	Input	Binary
TI_grade11	Yes	Input	Binary
TI_grade12	Yes	Input	Binary
TI_grade7	Yes	Input	Binary
TI_grade8	Yes	Input	Binary
TI_grade9	Yes	Input	Binary
TI_view1	Yes	Input	Binary
TI_view4	Yes	Input	Binary
TI_view5	Yes	Input	Binary
TI_zipcode25	Yes	Input	Binary
TI_zipcode26	Yes	Input	Binary
TI_zipcode4	Yes	Input	Binary
TI_zipcode42	Yes	Input	Binary
TI_zipcode48	Yes	Input	Binary
TI_zipcode49	Yes	Input	Binary
bathrooms_squared	Yes	Input	Interval
price	Yes	Target	Interval
waterfront	Yes	Input	Binary
yr_built	Yes	Input	Nominal

The regression result obtained after the above-mentioned features were selected:



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
price		AIC	Akaike's Infor...	310579	.	.
price		ASE	Average Squa...	2.489E10	2.736E10	3.48E10
price		AVERR	Average Error...	2.489E10	2.736E10	3.48E10
price		DFE	Degrees of Fr...	12890	.	.
price		DFM	Model Degree...	78	.	.
price		DFT	Total Degrees...	12968	.	.
price		DIV	Divisor for ASE	12968	4323	4322
price		ERR	Error Function	3.227E14	1.183E14	1.504E14
price		FPE	Final Predictio...	2.519E10	.	.
price		MAX	Maximum Abs...	2821151	2759701	4246794
price		MSE	Mean Square ...	2.504E10	2.736E10	3.48E10
price		NOBS	Sum of Frequ...	12968	4323	4322
price		NW	Number of Est...	78	.	.
price		RASE	Root Average ...	157756.7	165416	186535.8
price		RFPE	Root Final Pre...	158708.5	.	.
price		RMSE	Root Mean Sq...	158233.3	165416	186535.8
price		SBC	Schwarz's Ba...	311161.7	.	.
price		SSE	Sum of Squar...	3.227E14	1.183E14	1.504E14
price		SUMW	Sum of Case ...	12968	4323	4322

Model comparison result:

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error	Train: Mean Square Error	Train: Sum of Frequencies	Train: Number of Estimate Weights	Train: Root Average Sum of Squares	Train: Root Final Prediction Error	Train: Root Mean Squared Error
Y	Reg4	Reg4	Regression (2)	price		2.736E...	310579...	2.489E...	2.489E...	12890	78	12968	12968	3.227E...	2.519E...	26211...	2.504E...	12968	78	15775...	15870...	158233.3
	Reg3	Reg3	Regression	price		2.979E...	30709...	1.843E...	1.843E...	12685	283	12968	12968	2.39E14	1.925E...	22339...	1.884E...	12968	283	13575...	13874...	137256

CONCLUSION:

- we can see from the above results that the RMSE on training dataset was low in the scenario when **all the features** were used, but the RMSE on test data was **high** and this can be due to overfitting.

- When a set of features were selected, the errors were balanced and the RMSE on Training, Validation and Test Datasets were close to each other. This model was also selected as the best model by the model comparison node.

- I also ran the Neural Networks after the selection of features and the results of Neural Networks and Regression Model were close to each other.