

**2017**

Forecasting King County House prices

Using SAS E-Miner

# Introduction:

The three basic items that human being concern themselves are: Food, Clothing and Shelter. Considering the everyday struggles of finding a right shelter for a family at the right price is been a struggle. So, to reduce the painstaking process of determining the right house at the right place and at the right time, this project enlightens the consumer to plan from ahead of time for the purchase of the house. In this project, we aim to perform data analysis on King County, USA data set. We plan to perform predictive analysis using SAS Enterprise Miner. King County is a county located in Washington. It is the 13th most populous county in the United States. The county seat is Seattle which is the state's largest city. As of the 2010 US Census report, there were 851,261 housing units at an average density of 402.4 per square mile (155.4/Km2). The goal of the project is to enable us to study and evaluate the variation of house prices in King County based on different house attributes.

# Problem Statement:

Our main aim is to predict the prices of a house in king county, for which the data set has been collected. The following information can be derived once we analyze the data

* Pricing of a specific house
* Factors affecting the prices of a house in King County

This analysis can be helpful for an individual consumer as well as real estate agents increase profits.

* Individual consumer can determine the price of a house depending on the floor size, number of rooms and other factors.
* Similarly, an Individual consumer can predict the appropriate price for selling his house.
* Real estate agents can determine the right price and secure good profits in their business
* Additionally, King County government can keep track of the demographic population using the data and housing prices.

# Data Set Description:

Our project uses second-hand data retrieved from [Kaggle](https://www.kaggle.com/harlfoxem/housesalesprediction)**.** This dataset encapsulates a varied set of variables which play a major role in defining the actual price of the house. This dataset contains house sale prices for homes sold between May 2014 and May 2015 in King County, USA. In all, it contains 19 house features plus the “price” and the “id columns”, along with 21613 observations.

# High Level Data summary:

|  |  |
| --- | --- |
| Total observations in the dataset | 21613 |
| Total number of binary variables | 1 |
| Total number of nominal variables | 3 |
| Total number of interval variables | 17 |
| Outcome / target variable | “price” “grade” |
| Level of the target variable (nominal, binary or interval) | Interval, ordinal |

Below is the description about each variable ‘column’ used in the dataset.

id - a unique notation for a house Year – Year the house was sold

price - price is prediction target (price of the house) bedrooms - number of Bedrooms/House

bathrooms - number of bathrooms/bedrooms sqft\_living - square footage of the home sqft\_lot - square footage of the lot

floors - total floors (levels) in house

waterfront - house which has a view to a waterfront view

condition - how good the condition is (Overall rating [1 - 5] - 5 being the best condition) grade - overall grade given to the housing unit, based on King County grading system (Rating [1 - 13] - 13 being the best)

sqft\_above - square footage of house apart from basement sqft\_basement - square footage of the basement

yr\_built – Year in which the house was built yr\_renovated - Year when house was renovated zipcode – zip code of the area

lat - Latitude coordinate long - Longitude coordinate

sqft\_living15 - Living room area in 2015(implies-- some renovations) This might or might not have affected the lot size area

sqft\_lot15 - lot Size area in 2015(implies-- some renovations)

# Data Mining Technique:

In our dataset, initially data cleaning will be performed to replace/remove missing values. We tried to classify the target variable using Decision trees, to determine the category of output variable. In later stages, we would use neural network, linear regression and gradient boosting and compare the models to interpret different models. Finally, we would also like to use techniques such as Neural networks and Random forests to build an ensemble model.

# Project Diagram:

**Data Pre-Processing:**

Feature Selection:

* We are rejecting “lat” and “long” variables as spatial data analysis is out of scope.
* “Id” and “year” is rejected as it doesn’t contribute to the actual prediction and

do not have any predictive power.

# File Import:

Our dataset consists of one csv file. The csv file was imported using the File import node. The local path was provided to import file field to successfully import the dataset.

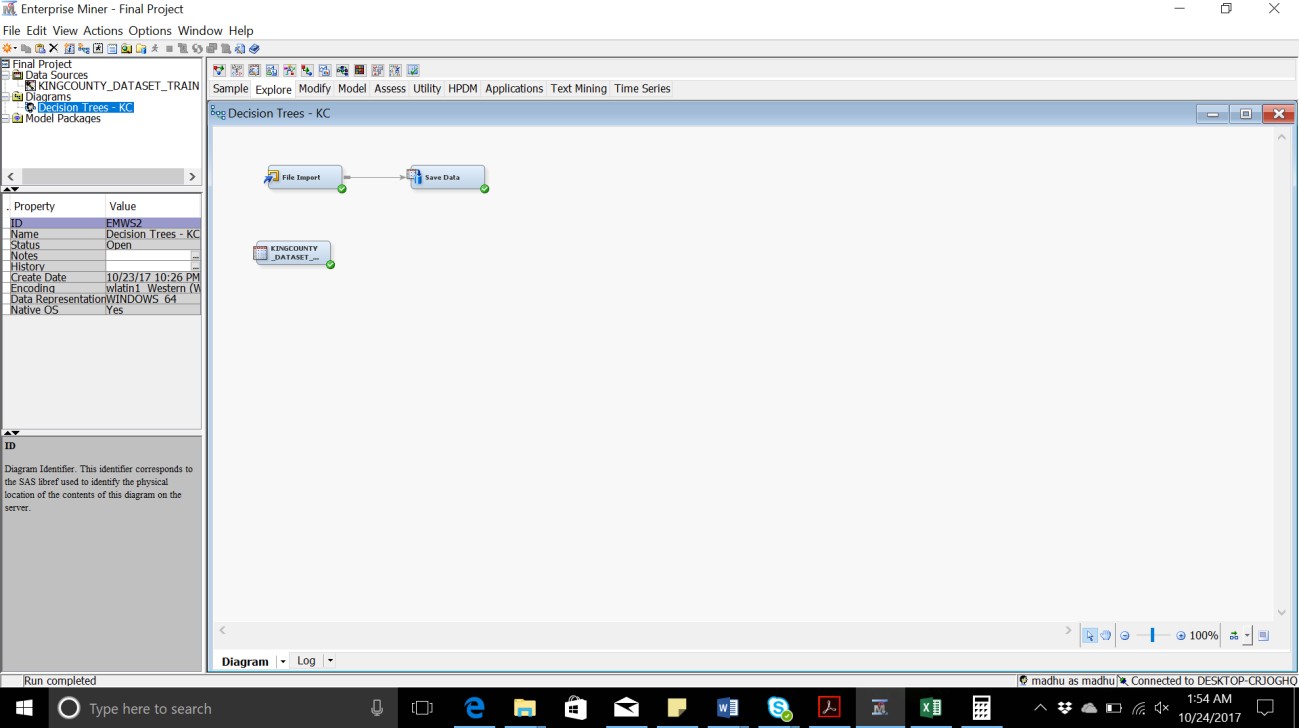
# Save Data:

The node was used to convert the dataset to sas7bdat file. The file is saved as KingCounty\_Dataset by passing the name to the field “Filename Prefix” and the library in which the file should be stored was provided to the field “SAS Library

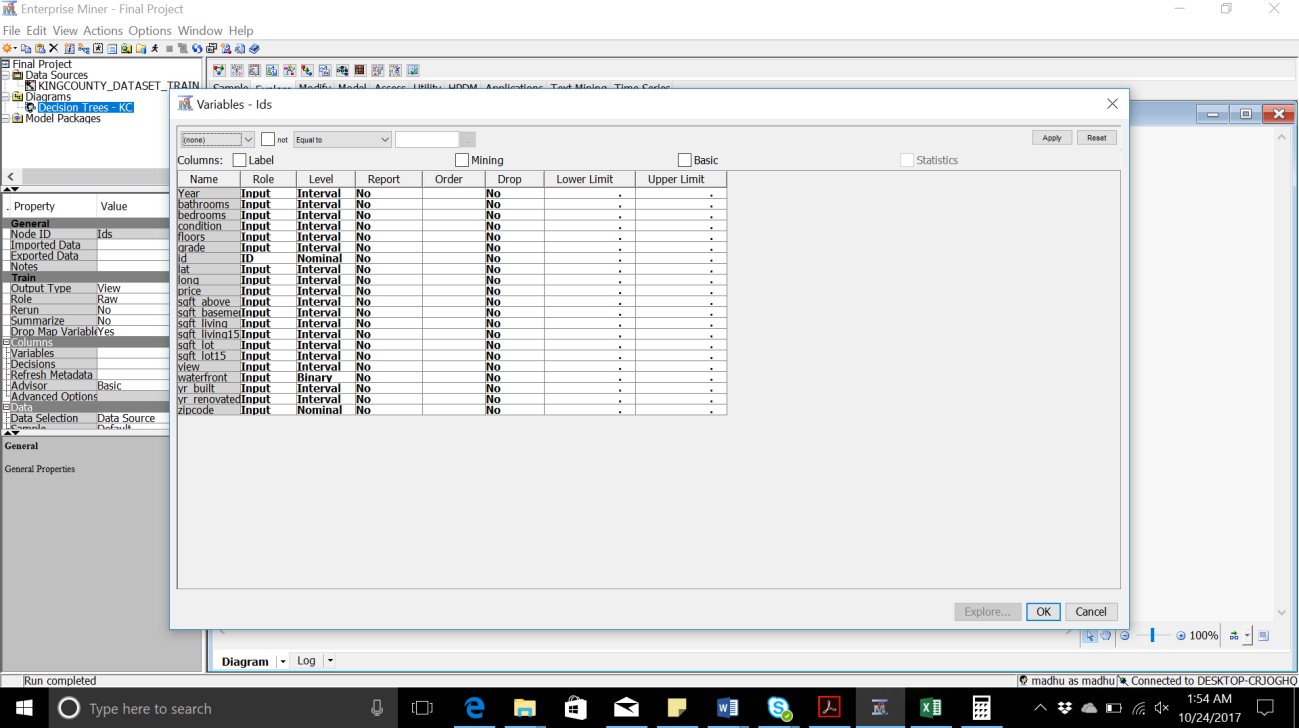
name”.

# Data Source:

The above saved sas7bdat file was then created as a new data source.

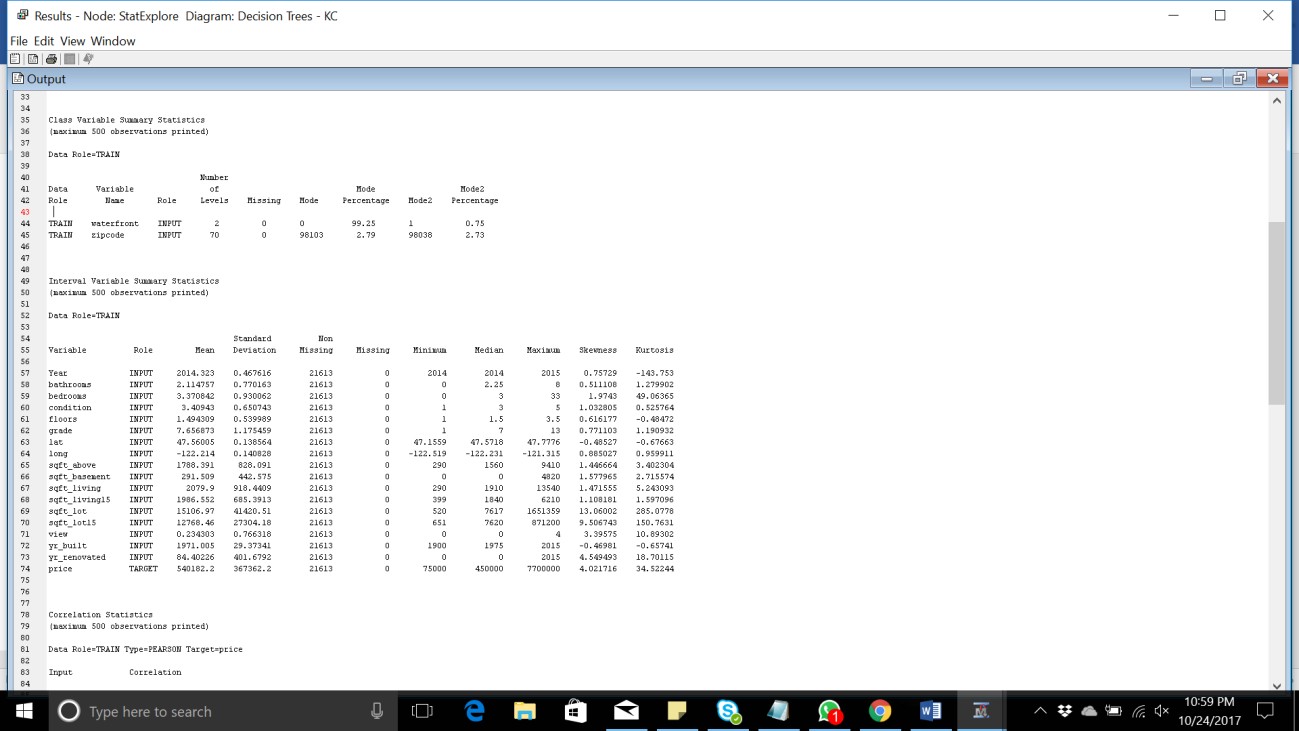
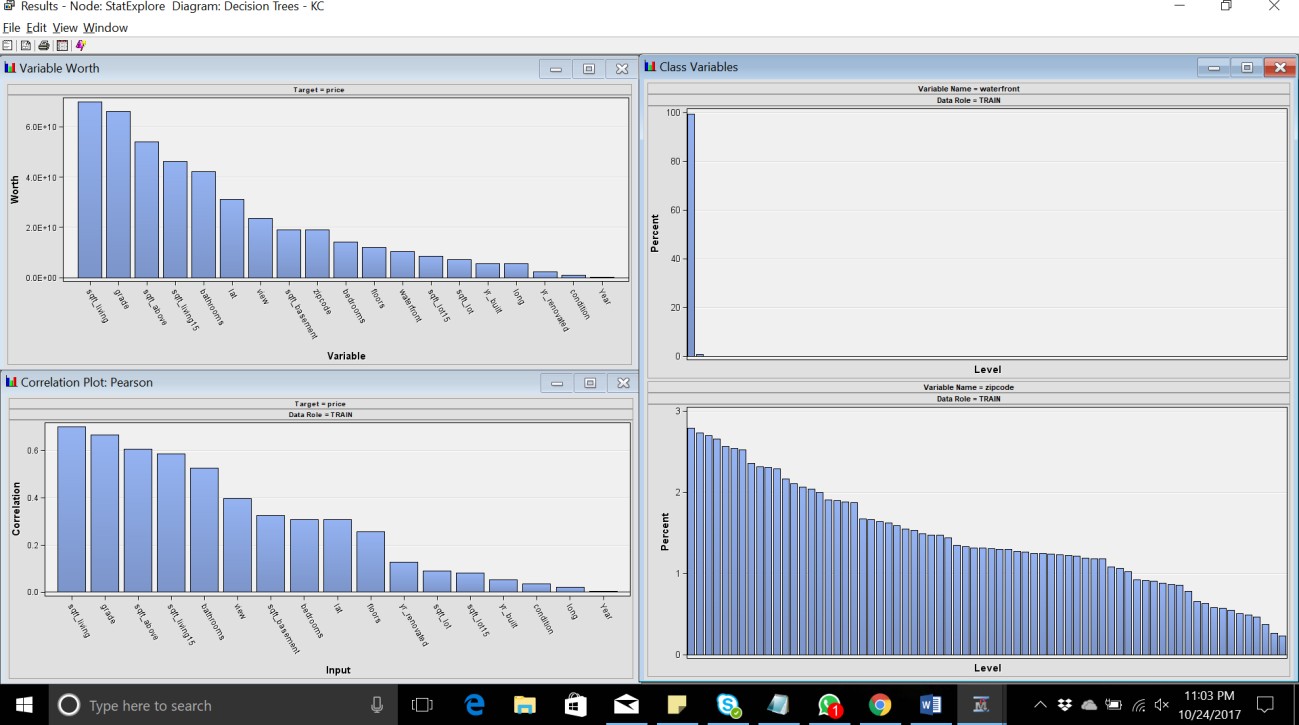


Below is the list showing all the 21 variables:



# Stat Explore:

This node is used to generate a summary about the input variables. From the results, we found that there are no incorrect and missing values in the dataset. Below is the summary.



From the above screenshot, we find that the variable “yr\_renovated” has a minimum value of 0, which is incorrect. The zero indicates that the house is not renovated from the time it was built. Such correction will be done in the later stage of the flow, where we will replace the zero with the year in which the house was built. Additionally, the number of bathroom variable represents numbers in decimals, which is unrealistic. So, we can replace the decimal values to a whole number or a

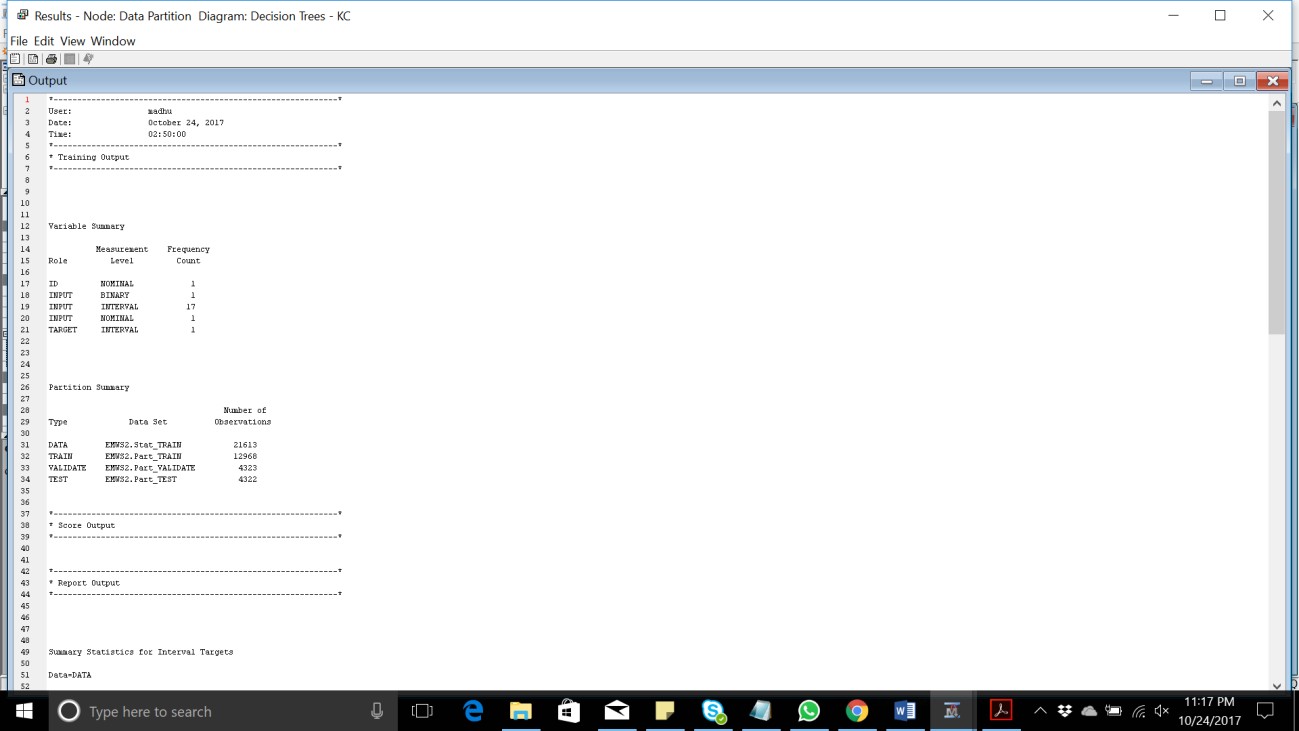
0.5 indicating a different bathroom and shower tub. For the above modifications, we will use Replacement and Impute nodes.

# Data Partition:

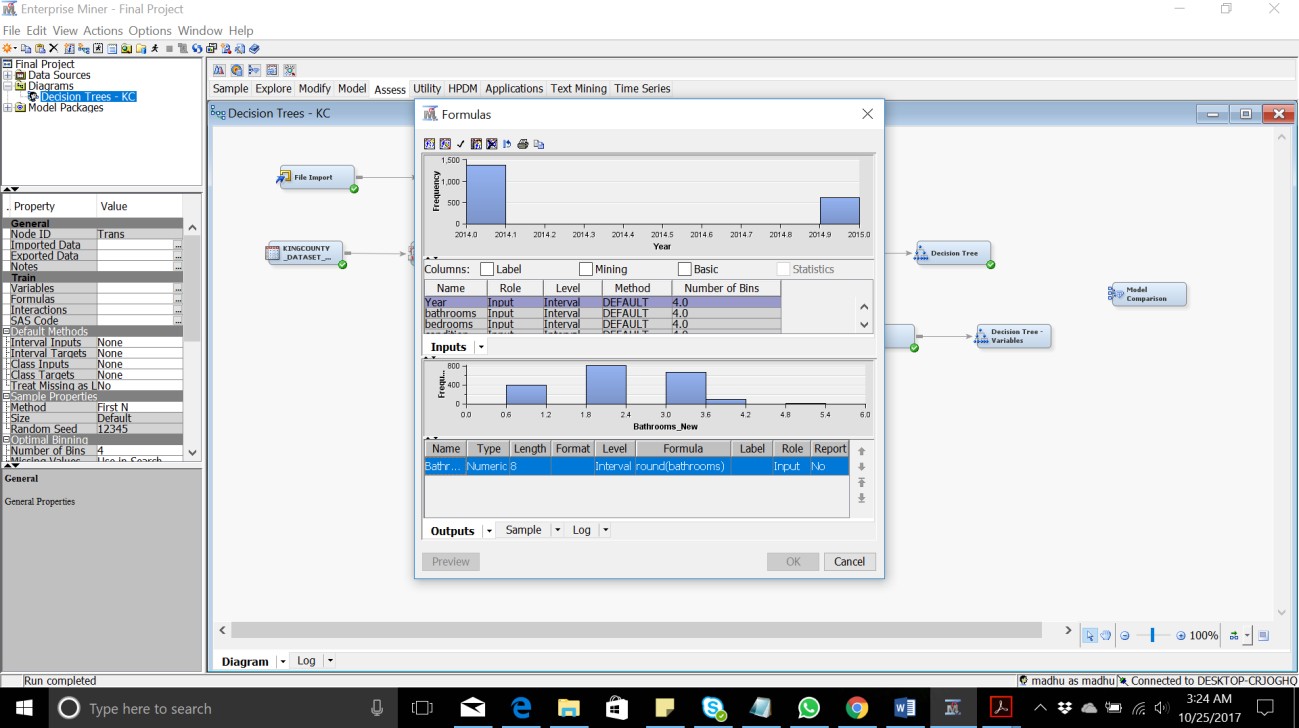
This node is used to partition the dataset into three parts namely, Training set: 60%

Validation set: 20%

Test set: 20%

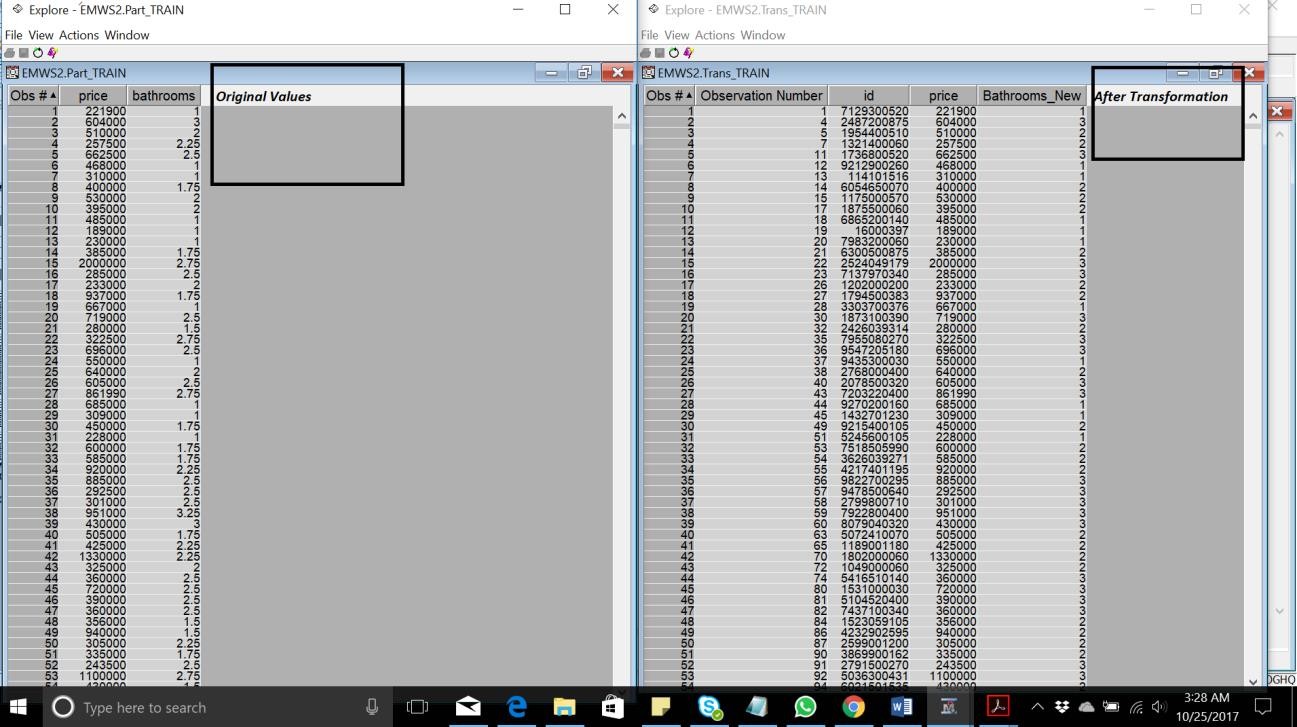


# Transform Variables:

This node is used to create a new column “Bathroom\_New” in which the batrooms are rounded off. As we have decimal values in Bathroom such as 1.25,1.5 and 1.75, these values are replaced with the closest integer value.

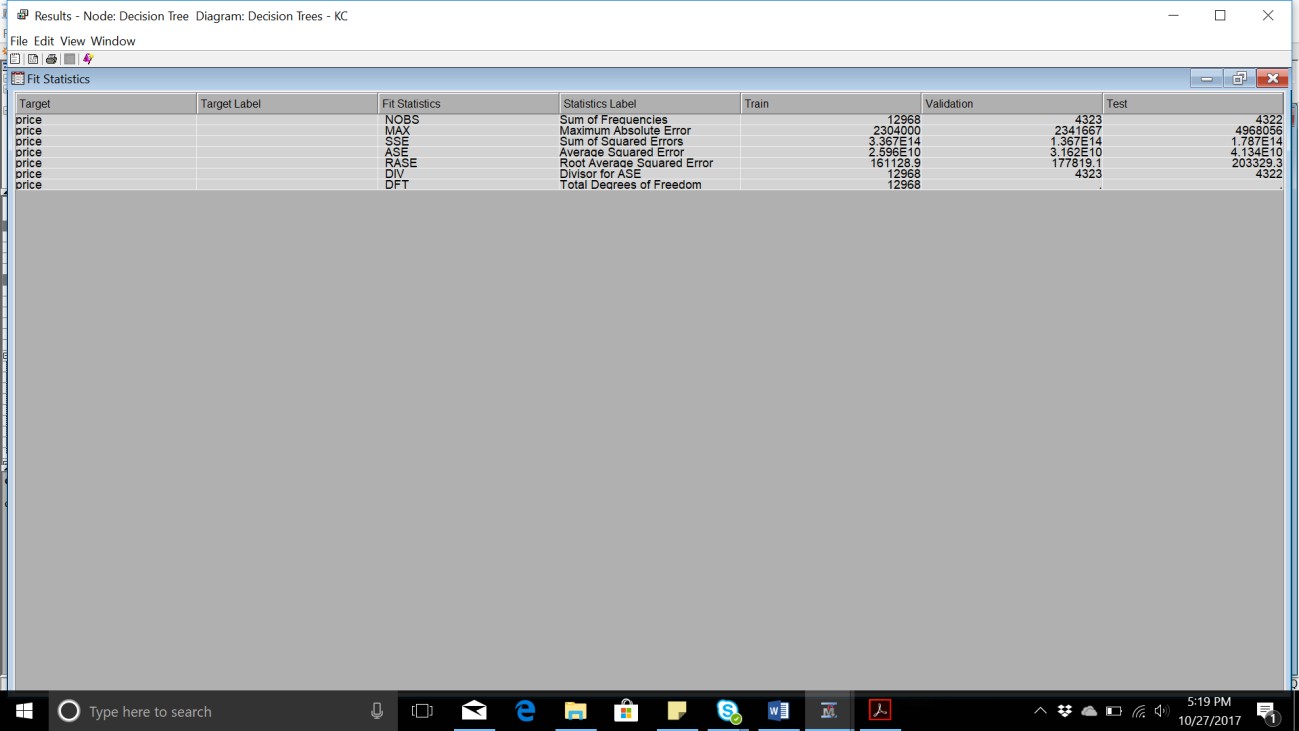
# Stat Explore Transformed:

This node is generated to generate the summary of the transformed variable.



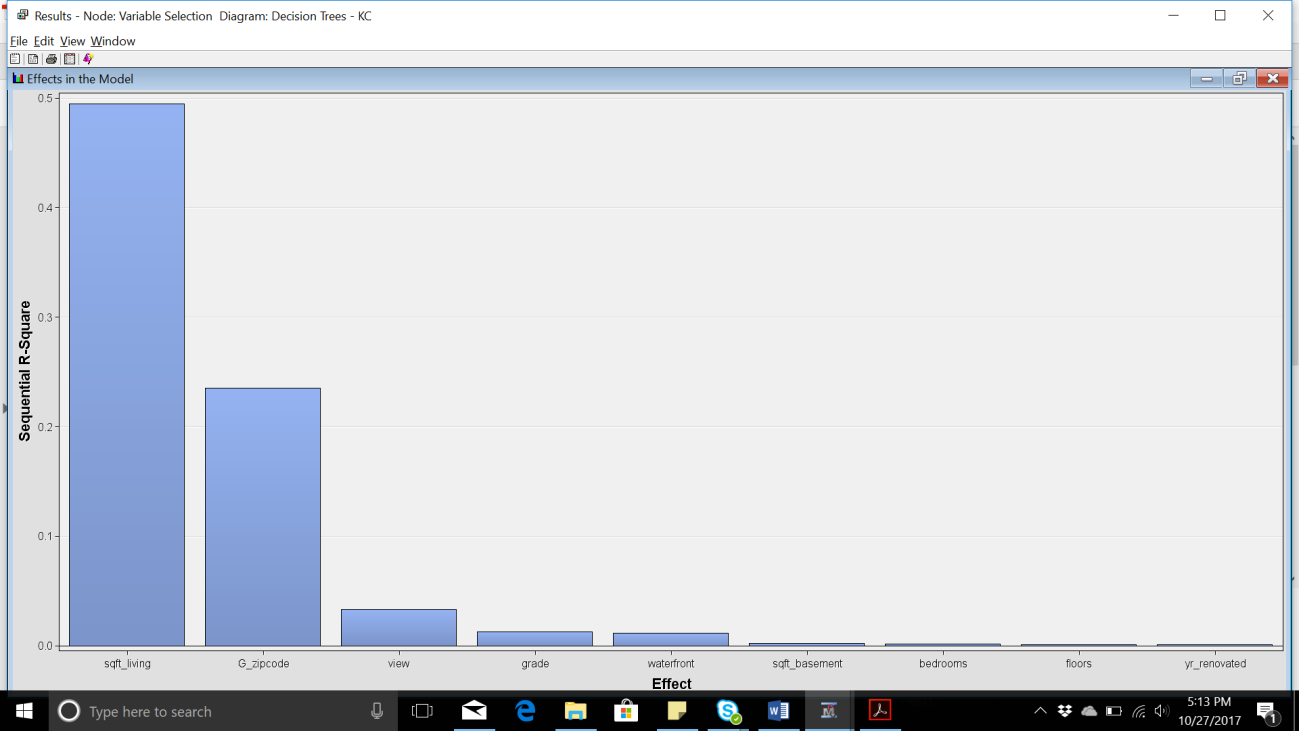
# Decision Tree:

This tree is used to predict the “price” target variable from the transformed data. For the time being we don’t select the “zipcode” and “id” as they are nominal.

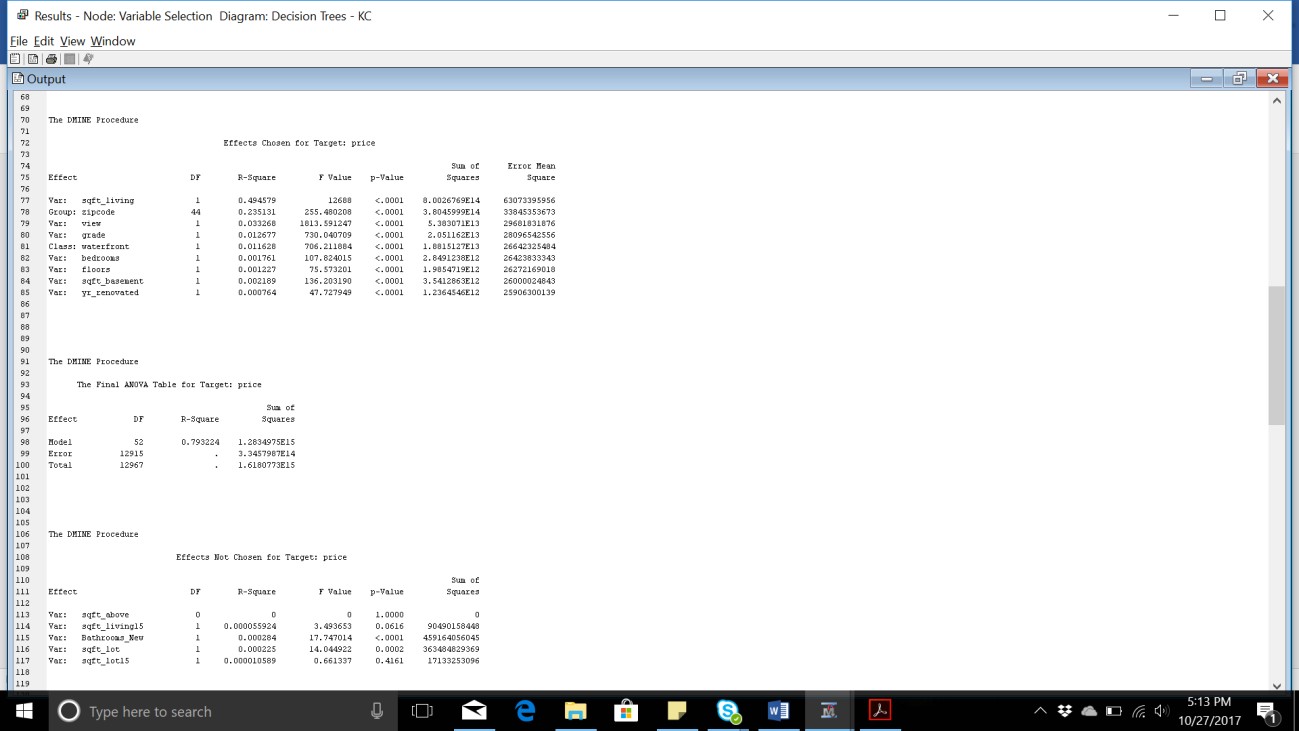


# Variable Selection:

This node is used to filter the variables based on their R-square values.

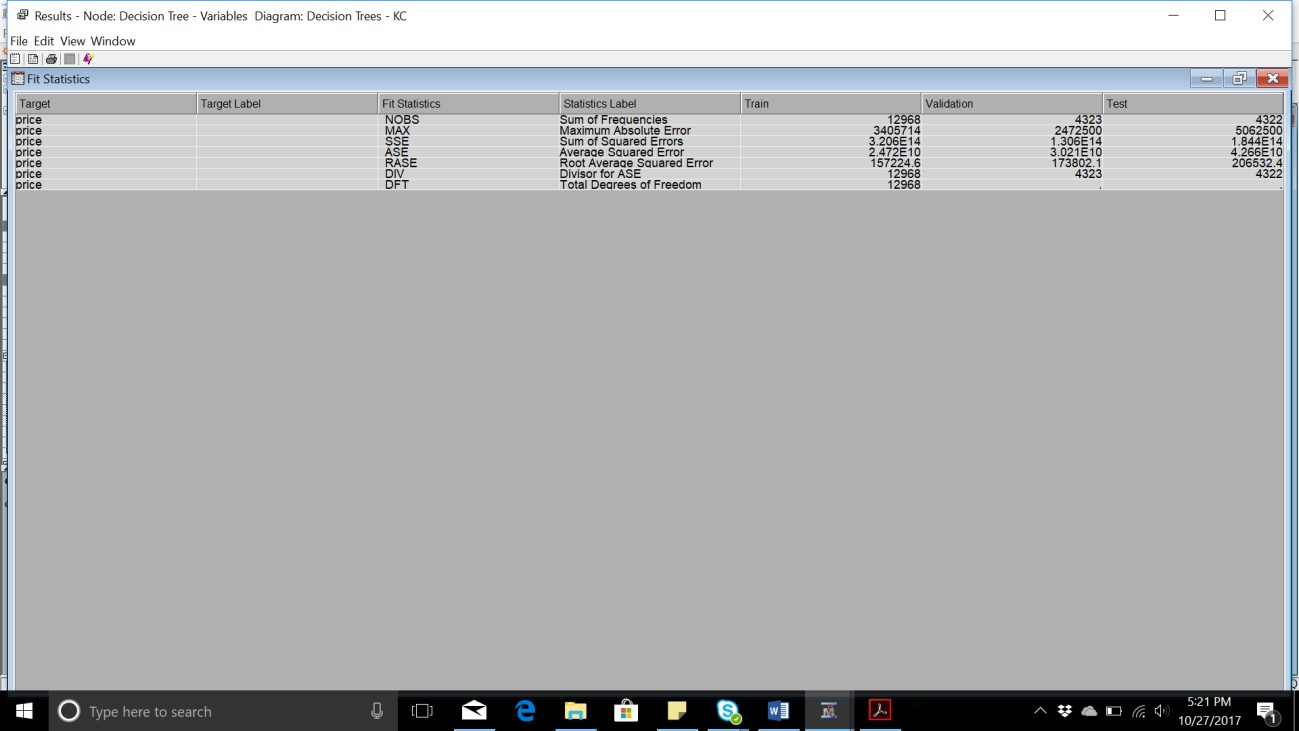


The below screenshot describes which variables are chosen (selected for analysis) and rejected as per the R-square analysis.



# Decision Tree – Variables:

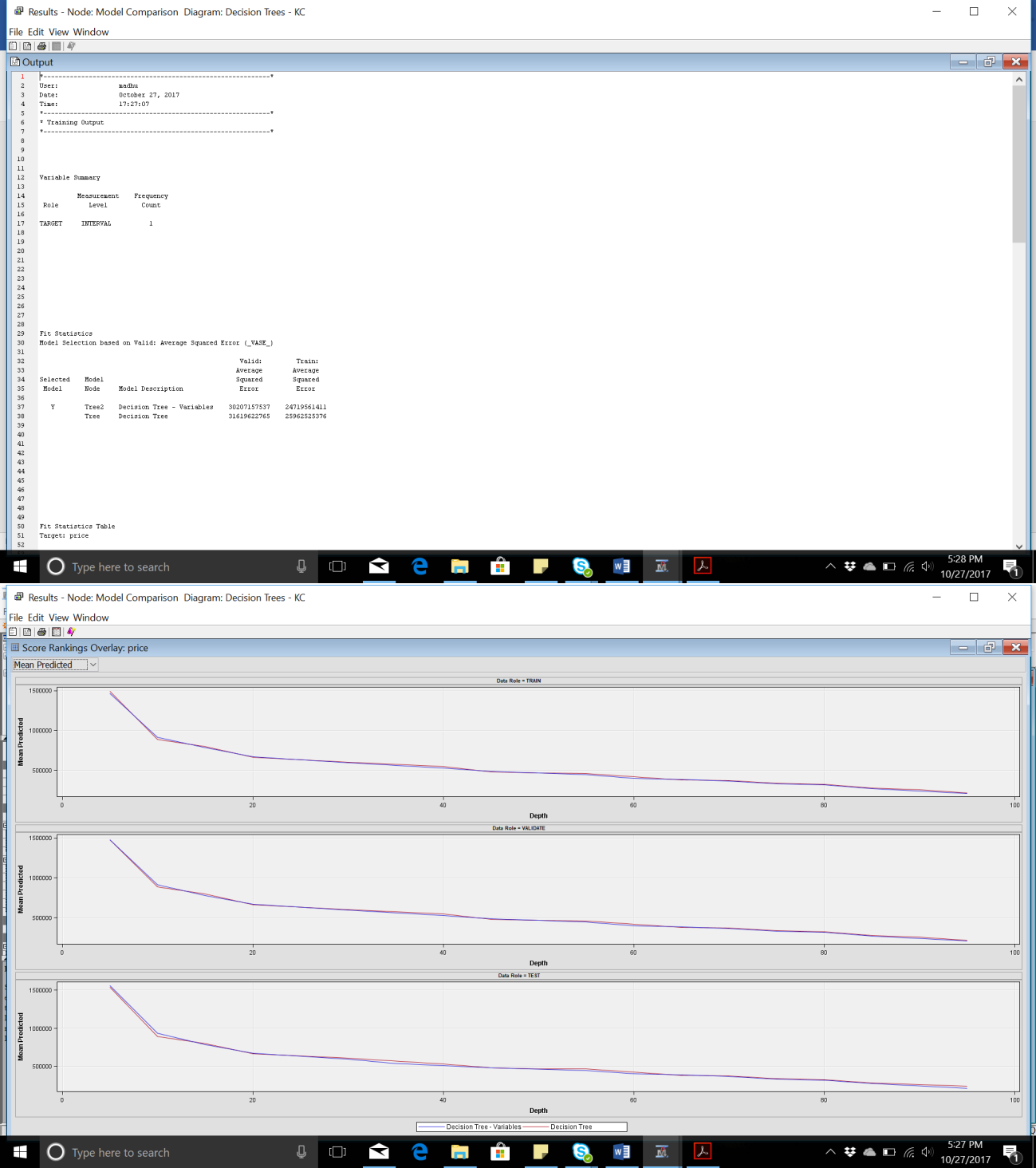
This node is used to predict the “price” for the selected variables.



# Model Comparison:

This node is used to compare both the models. The main factor distinguishing these models is their Mean Squared Error. The following are the mean squared error for both models:

Without variable selection: 25962525376 With variable selection: 24719561411



# Conclusion:

* + The model words best without selecting the variables
  + Even after variable selection and rejecting many input variables there isn’t much

difference in the mean squared error of the two models.

* + Using more robust predictors might reduce the mean squared error further.

# Future Progress Plan:

* + To implement different values of maximum depth and maximum number of trees using random forest.
  + To create a model of gradient boosting and compare the performance.
  + To make use of various other classification technique like Neural network and linear regression to compare and identify the best predictive model.