COMP309 Assignment 2

Ben Allan

300476366

Part 1: Business and Data Understanding [40 marks]

1. Perform an initial EDA

a.

1460 instances / rows

81 features/ columns

43 Categorical Variables

38 Numeric Variables

b.

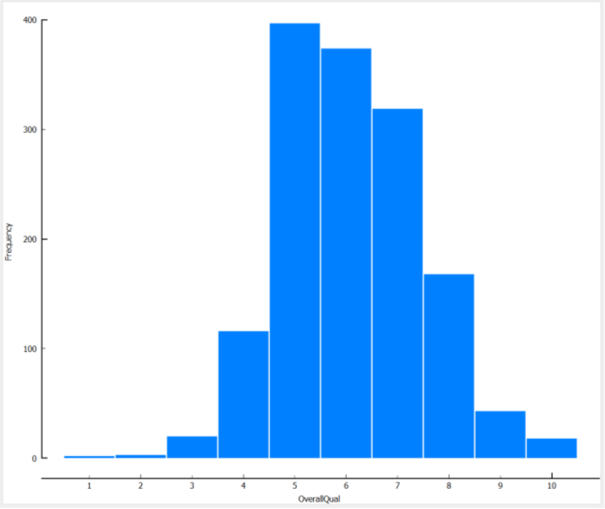
The top 5 numerical features highly correlated with the target variable (“SalePrice”)

1) +0.791, Overall Quality (OverallQual) Numerical
2) +0.709, Above grade (ground) living area square feet (GrLivArea) Numerical
3) +0.604, Garage space in number of cars (GarageCars) Numerical
4) +0.623, Garage space in area (GarageArea) Numerical
5) +0.614, Total basement area in square feet, (TotalBsmtSF) Numerical

C.

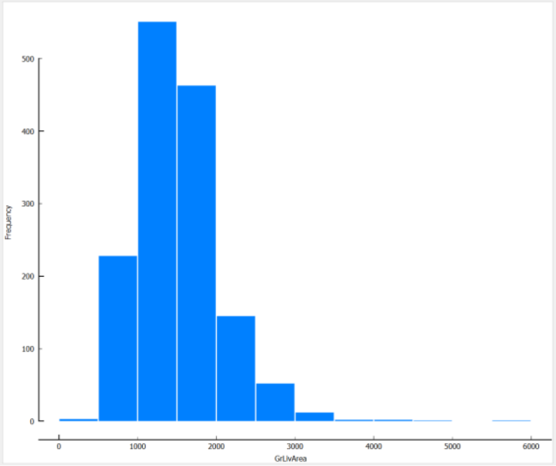
Overall quality: Skewness: 0.2167 & Kurtosis: 0.0919

The Distribution of overall quality is skewed right and is only slightly too peaked.

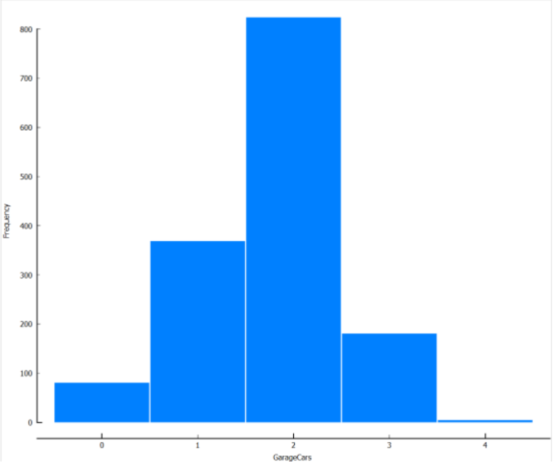


Above grade (ground) living area square feet: Skewness: 1.3652 & Kurtosis: 4.8743

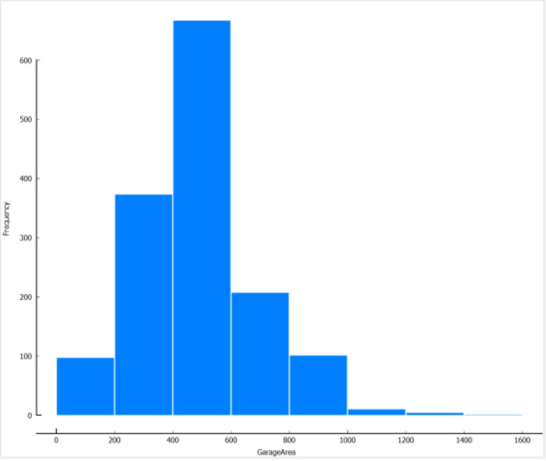
The Distribution of living area above ground is skewed right and too peaked.



Garage Cars: Skewness: -0.3422 & Kurtosis: 0.2161
The distribution of garage cars is skewed slightly left and is too peaked.

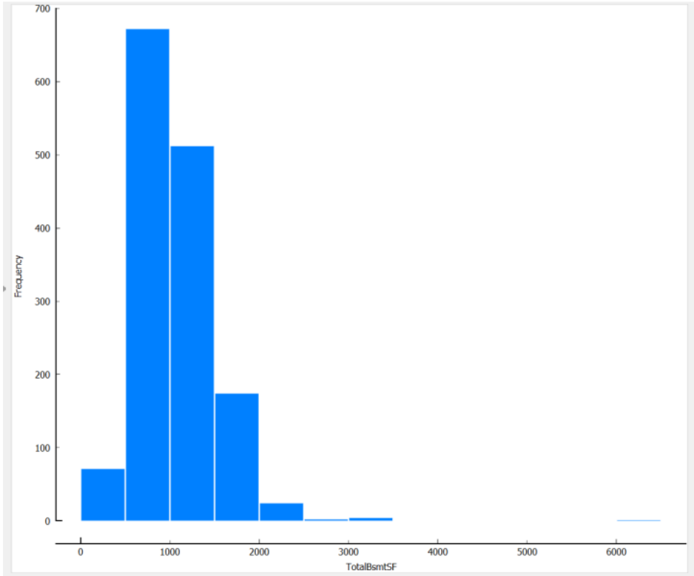


Garage Area: Skewness: 0.1798 & Kurtosis: 0.9098
The distribution of garage areas is skewed to the right and is too peaked.



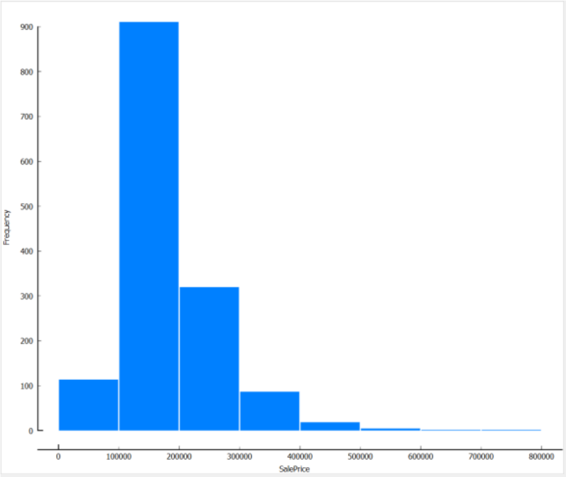
Total basement area: Skewness: 1.5227 & Kurtosis: 13.2010

The distribution of Total basement area is skewed to the right and is extremely peaked.



Sale Price: Skewness: 1.8809 & Kurtosis: 6.5098

The distribution of sale price is skewed to the right and is extremely peaked.



D.

There are a lot of missing data within the dataset (5.9% of the entire dataset). With 19 features having at least one instance missing. 4 features have over 80% of their instances missing these are PoolQC, MiscFeature, Alley and Fence, with PoolQC having 99% missing. The remaining 15 features have 47% or less of their instances missing. See appendix for exact values and percentages for individual features.

2. Investigate the business understanding questions

a. Translate the two business questions into two data mining goals;

Business Questions:

1.What factors affect the house price?

2.How do these factors affect the house price?”/“in which way do the factors affect the house price?

Data Mining Goals:

1.Use existing features within the dataset and correlation metrics to determine which factors have the biggest influence on house price weather positive or negative.

2.Use machine learning paradigms to explore how these factors have an effect on house prices and even predict pricing in the future.

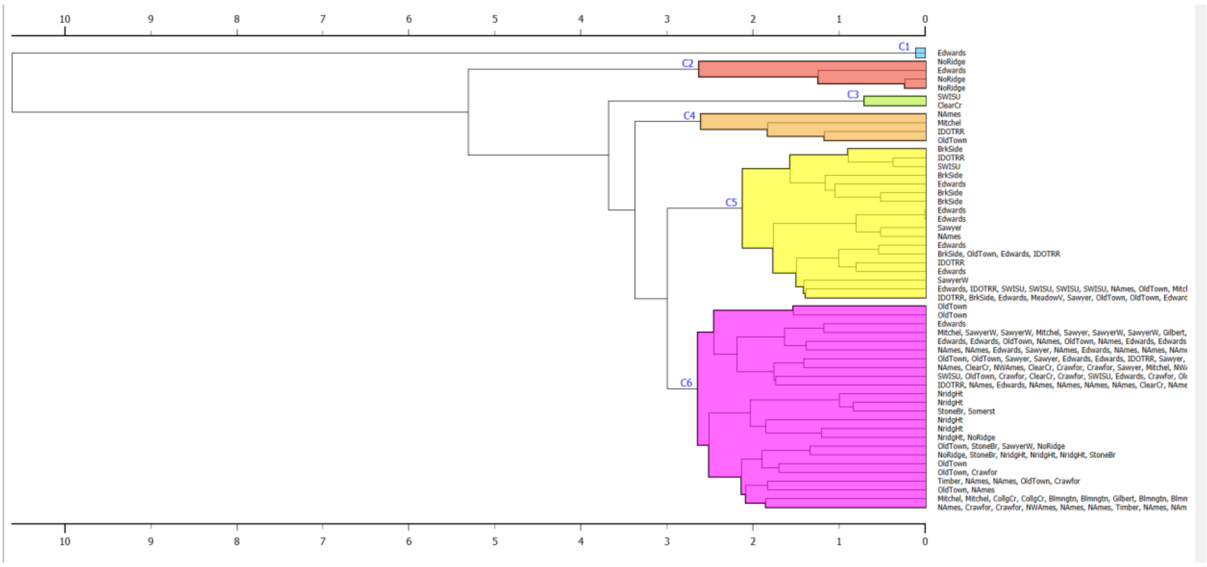
b. Select two machine learning paradigms, e.g. classification, regression, dimensionality reduction and so forth, that can help you achieve these goals. Provide justifications of your decision.

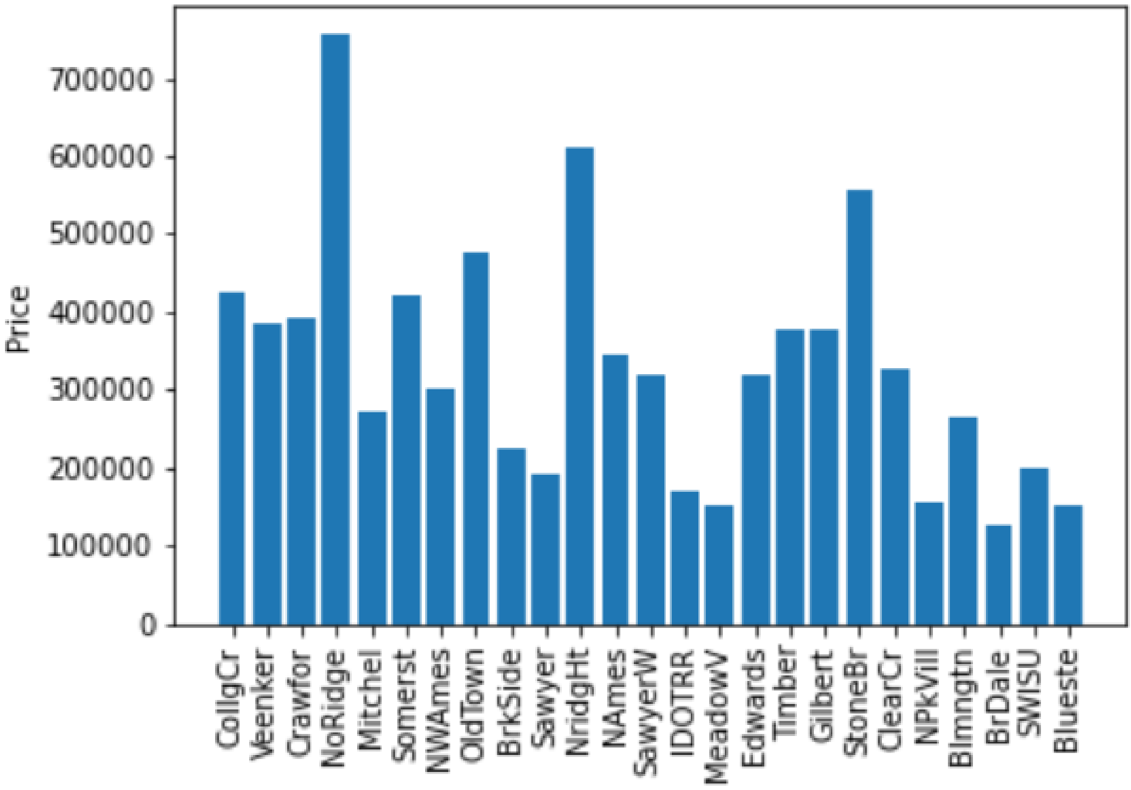
Since all the highest correlating features are numerical as well as the target feature it would be wise to use a regression approach first to gain insights regarding the above goals. Multiple linear regression could work best at first using the first 5 of the most highly correlated variables, although we should assess the explained variance from these first 5 features and re-iterate if needed.

We could also use dimensionality reduction such as Feature selection or PCA to reduce the high number of 81 features down to a lower dimension space while still retaining some of the meaningful properties from the original data.

3. Provide a further EDA on the dataset using Hierarchical clustering on the 5 numerical features

Yes, the house price does vary by Neighbourhood as can be seen within the diagrams below. For example, within Crawford the average price of a home is substantially higher than any other neighbourhood within the dataset. This paired with the variation in average house price per neighbourhood is enough evidence to suggest that yes, house price does vary by neighbourhood.


10
9
8
7
6
5
4
3
2
1
0
C1
C2
C3
C4
C5
C6
Edwards
NoRidge
Edwards
NoRidge
NoRidge
SWISU
ClearCr
NAmes
Mitchel
IDOTRR
OldTown
BrkSide
IDOTRR
SWISU
BrkSide
Edwards
BrkSide
BrkSide
Edwards
Edwards
Sawyer
NAmes
Edwards
BrkSide, OldTown, Edwards, IDOTRR
IDOTRR
Edwards
SawyerW
Edwards, IDOTRR, SWISU, SWISU, SWISU, SWISU, NAmes, OldTown, Mitcl
IDOTRR, BrkSide, Edwards, MeadowV, Sawyer, OldTown, OldTown, Edward
OldTown
OldTown
Edwards
Mitchel, SawyerW, SawyerW, Mitchel, Sawyer, SawyerW, SawyerW, Gilbert,
Edwards, Edwards, OldTown, NAmes, OldTown, NAmes, Edwards, Edwards
NAmes, NAmes, Edwards, Sawyer, NAmes, Edwards, NAmes, NAmes, NAme
OldTown, OldTown, Sawyer, Sawyer, Edwards, Edwards, IDOTRR, Sawyer,
NAmes, ClearCr, NWAmes, ClearCr, Crawfor, Crawfor, Sawyer, Mitchel, NW
SWISU, OldTown, Crawfor, ClearCr, Crawfor, SWISU, Edwards, Crawfor, Ol
IDOTRR, NAmes, Edwards, NAmes, NAmes, NAmes, NAmes, ClearCr, NAme
NridgHt
NridgHt
StoneBr, Somerst
NridgHt
NridgHt
NridgHt, NoRidge
OldTown, StoneBr, SawyerW, NoRidge
NoRidge, StoneBr, NridgHt, NridgHt, NridgHt, StoneBr
OldTown
OldTown, Crawfor
Timber, NAmes, NAmes, OldTown, Crawfor
OldTown, NAmes
Mitchel, Mitchel, CollgCr, CollgCr, Blmngtn, Blmngtn, Gilbert, Blmngtn, Blmn
NAmes, Crawfor, Crawfor, NWAmes, NAmes, NAmes, Timber, NAmes, NAm


Price
0
100000
200000
300000
400000
500000
600000
700000
CollgCr
Veenker
Crawfor
NoRidge
Mitchel
Somerst
NWAmes
OldTown
BrkSide
Sawyer
NridgHt
NAmes
SawyerW
IDOTRR
MeadowV
Edwards
Timber
Gilbert
StoneBr
ClearCr
NPkVill
Blmngtn
BrDale
SWISU
Blueste

Part 2: Data Preparation and Machine Learning [60 marks]

1 (20 marks) Determine and describe the data pre-processing steps applied to the provided dataset

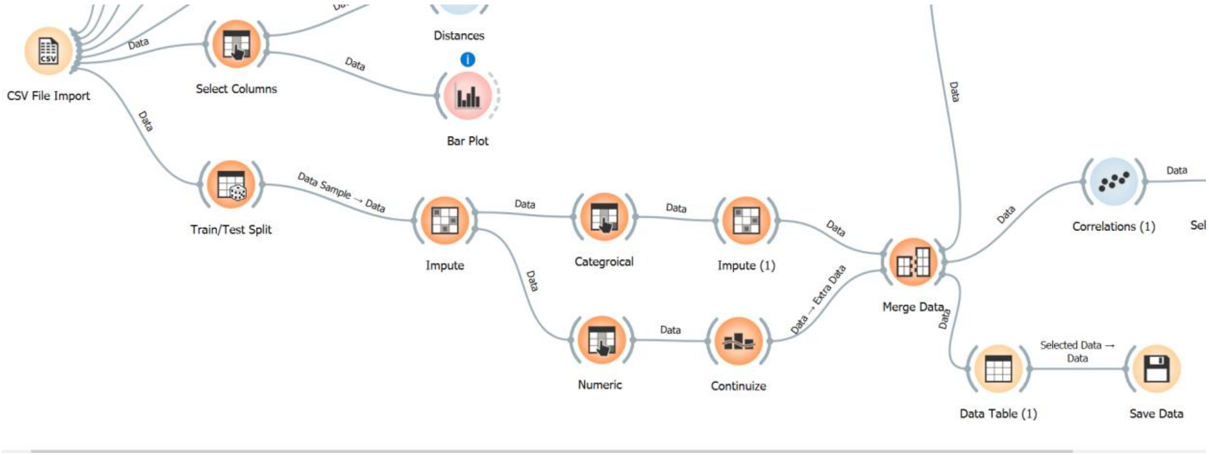
First thing first is splitting the data into a training set and a test set with a 70-30 split.

Change all NaN values within categorical features that recorded no entry as nothing. For example, PoolQC has 99% missing values, but the dataset has been recorded so that these missing values = No pool at that house. These will be changed to NP or No Pool for example.

The NaN values within the numerical variables and categories that don't use NAN as a category will be imputed.

Any categorical data will be encoded using the imputer within orange, and numerical data will be normalised through the containerize feature within orange.

All redundant instances, outliers or non-effective instances were removed by orange through the imputer feature.



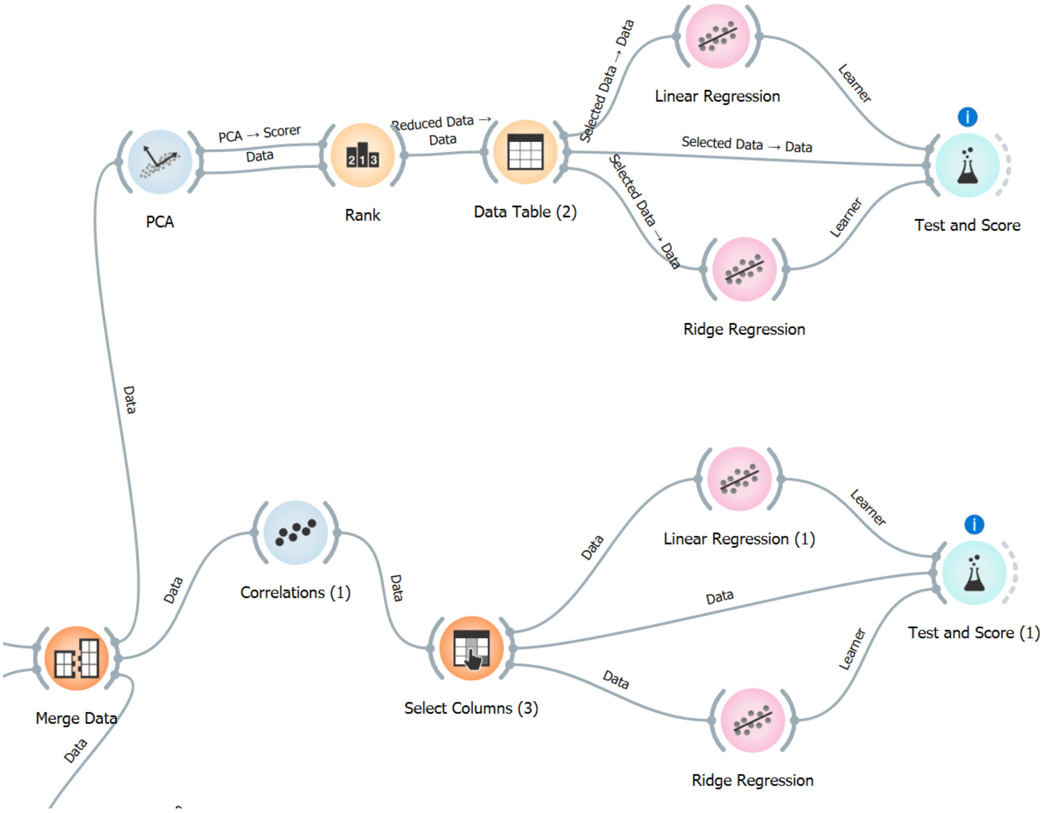
Pre-processing workflow.

(Full process diagram will be shown within the appendix below)

2. Utilise two different dimensionality reduction techniques to identify which features are irrelevant and/or redundant to predicting the house price

I used PCA to identify which features are irrelevant when predicting house price and retuned with 9 features that explain the most variance. Within Orange I applied PCA to my processed dataset and found 9 most relevant features to continue onto linear and ridge regression with, removing all other features. This can be seen within the workflow below

I also used feature selection to select the 5 highest positive correlating features with house price and applied them to linear and ridge regression, removing all other features. This can be seen within the workflow below



PCA analysis and feature selection processes.

3. Approach data mining goals on your pre-processed data using machine learning methods

Principle Component Analysis:

Training Set:

Linear Regression MSE: 2425844262.3

Ridge Regression MSE: 2401342362.9

Test Set:

Linear Regression MSE: 2479703851.5

Ridge Regression MSE: 2468176121.3

Feature Selection:

Training Set:

Linear Regression MSE: 1493135263.8

Ridge Regression MSE: 1493235976.9

Test Set:

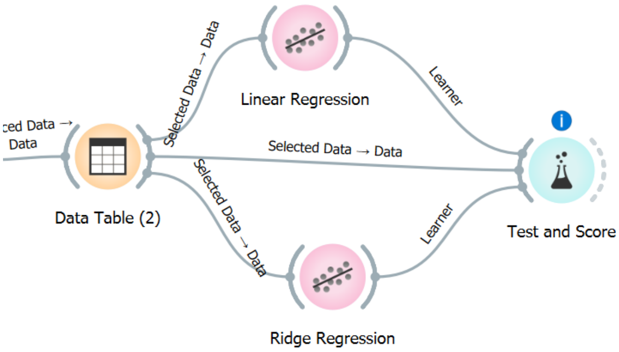
Linear Regression MSE: 1576014342.7

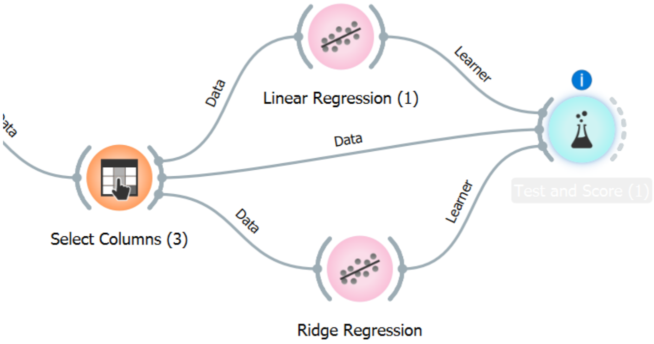
Ridge Regression MSE: 1572583394.4

As can be seem within the MSE results Ridge Regression performed better on both feature groups with Feature Selection producing the lowest MSE value on training set. This could be due to the degree of bias ridge regression estimates, therefore reducing the standard errors.

The MSE values are extremely large suggesting the model is not a good fit to the data.

As a result, I cannot conclude accurate answers to business questions as the data mining goals have not been completely achieved.

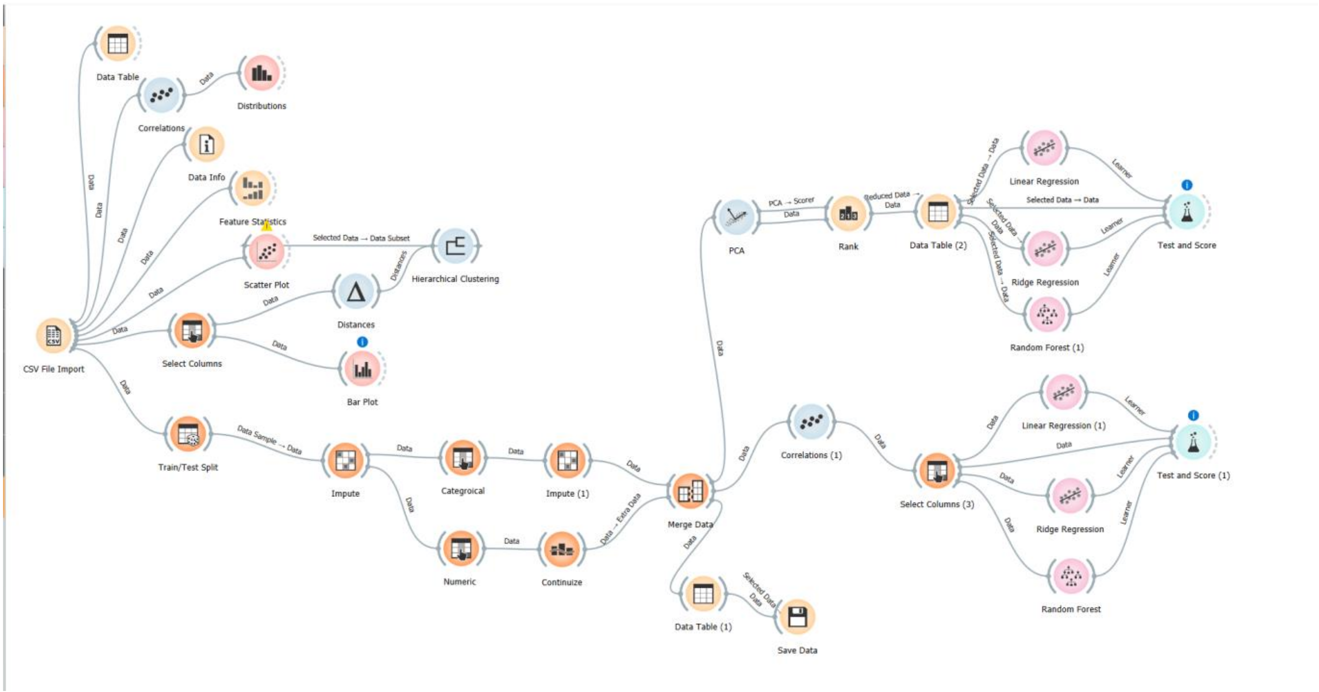




3B) Random Forest regression method to predict the house price

The random forest produced better MSE scores when compared to the other regression methods. Within the PCA group, MSE was 2044675125.4 on the test set, which is considerably lower than both linear and ridge regression. Within the feature selection group, MSE was 1220890300.1 on the test set, which is also considerably lower than both linear and ridge regression. This implies that Random Forest is a better fit to the data producing more accurate results than the other methods as evaluated by the MSE value.

Appendix:



Data Table
Distributions
Correlations
Data Info
Feature Statistics
Scatter Plot
Selected Data → Data Subset
Hierarchical Clustering
Distances
Distances
CSV File Import
Select Columns
Bar Plot
Train/Test Split
Data Sample → Data
Impute
Categroical
Impute (1)
Numeric
Continuize
Data → Extra Data
Merge Data
Data Table (1)
Selected Data → Data
Save Data
PCA
PCA → Scorer
Rank
Reduced Data → Data
Data Table (2)
Selected Data → Data
Linear Regression
Ridge Regression
Random Forest (1)
Learner
Test and Score
Correlations (1)
Select Columns (3)
Linear Regression (1)
Ridge Regression
Random Forest
Test and Score (1)
Data