Smith SCHOOL OF BUSINESS Queen's University

Global Master of Management Analytics

GMMA 867

Predictive Modeling

Dr. Tatsiana (Tanya) Levina

Assignment #1:

- **Kaggle Competition:** "House Prices: Advanced Regression Techniques"
- **Total Number of Team in this competition:** 5,094 teams
- **My Position on the leaderboard at the time of my submission:** 767

July 19, 2020

Nicole Hong

Order of files:

Filename	**Pages**	**Comments and/or Instructions**
GMMA 867 Individual Assignment 1	19 pages, including cover-page	
Individual Assignment 1_R Script File	n/a	Assignment 1 - R Codes v-FINAL (from R Studio)

Additional Comments:

Assignment 1 – Individual

Instructions

This assignment is to be completed and submitted individually.

Pre-Assignment
Complete the following to prepare yourself for the graded portion of this assignment:

1. If you haven't already done so, go to Kaggle's website and register an account.
2. Read up on what Kaggle is and reflect on how you may use it in your future jobs.

Report Requirements
In one comprehensive report, include the following:

1. From Kaggle, identify 3 competitions which are suitable for predictive modeling using regression analyses (i.e., where the goal is to predict a quantity) and have at least 200 entrants. Of the three, select the one in which you will participate; explain your choice. **[10 pts]**
   These competitions do not necessarily need to be active; if you can, submit to a competition that is already finished.
2. Join the competition and build a regression model for your selected competition. In your report, explain how you approached the task, the steps you took, how you revised your model as your analyses progressed, and comment on the quality of your predictions. Be sure to include the description of the data and the link to it on Kaggle. **[40pts]**
3. On the front page of your report, include your Kaggle name, the total number of teams on the leaderboard, and your position on the leaderboard at the time of your submission. **[10pts]**
4. In the Appendix of your report, include your model and a screenshot showing your highest position on the leaderboard.
5. Submit the PDF of your report as well as your model file(s) (R code, Excel spreadsheet, etc.) to this assignment.

Up to 40 points will be awarded for your rank in the Kaggle Competition you've chosen for your report. **[40pts]**

a. 20pts will be awarded for submissions that end up in the top 50% (i.e., those who beat the average)
b. An additional 20pts will be awarded for submissions that end up in top 33%
c. A bonus 20pts will be awarded for submissions in top 25%

When you are ready to make your submission, select "Add a File" to upload your work. When you have uploaded everything, select "Submit" to hand it in.

Keep an eye on the minibar at the top of the page for notifications of feedback on your work.

Report on Kaggle Competition:
"House Prices: Advanced Regression Techniques"
Predict sales prices and practice feature engineering, RFs, and gradient boosting

Link to the Kaggle Competition: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview

Identification & Selection of the Kaggle Competition

My selection of the Kaggle Competition was based on the complexity of datasets but also considering the ease of navigation of data – whether the data is understandable and continuous. Also, I focused on the competition that would more likely involve the quantitative models than the qualitative models. I also wanted to avoid the competition that would be subject to the classification type of models, and also with the datasets with big data, which would likely complicate my model development.

I rigorously searched for the competitions related to the prediction of prices or sales data, and also the competition that was still open to the public and allowed submission or late submission.

Identification – two competitions that I ended up not pursuing:

I looked at the following two Kaggle Competitions, which I decided not to pursue later:

TFI Restaurant Revenue Predictions:
Link to the competition: https://www.kaggle.com/c/restaurant-revenue-prediction

Datasets were much simpler, when compared to all the datasets from the competitions that I was reviewing, but they lacked the complexity that I was looking for.

Walmart Recruiting - Store Sales Forecasting:
Link to the competition: https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data

Datasets were 6 years ago. I felt that datasets were a bit outdated.

Selection – the one that I pursued:

The Kaggle Competition that I selected in this report included all the features that I was looing for. Also, this competition was educationally driven, which I liked the most.

"House Prices: Advanced Regression Techniques"
Predict sales prices and practice feature engineering, RFs, and gradient boosting

Link to the Kaggle Competition: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview

Model Building

The following steps are taken, which are related to the supervised learning for solving the machine learning problem:

- Data Definition
- Preprocessing
- Train/Test split
- Model / Algorithm Selection
- Training / Prediction
- Evaluate the Performance of Models

Data Definition

Step 1. Understanding Goals & Objectives

I started this Kaggle competition by understanding the goals and objective – the goal of this competition is to predict sale prices of houses by using various aspects of residential houses in Ames, Iowa in the US housing market. We can consider aspects such as years built, total area of houses in square feet, locations, the number of rooms and so on, which are provided in the datasets for this competition. Any insights obtained from this prediction can be used for the decision-making process by home buyers, which will also help home buyers plan and optimize the purchase of their dream homes within their budget.

The objective of this competition is to come up with the optimal, predictive model or methodologies to achieve the goal above.

Step 2. Understanding data in the datasets from the competition

From the Kaggle site, two dataset files were provided – train.csv and test.csv.
Link to the Kaggle Data: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

These datasets are comprised of 79 predictors or independent variables, one index ('Id') and one dependent variable. Data types in these datasets are either integers or strings, and can possibly be factors as well.

As for two data files provided from the Kaggle competition, train.csv, is used for building the model with the SalePrice values, and test.csv for predicting the SalePrice values, which are not provided to be calculated by using the model developed.

Pre-Loading Work
For consistency and simplicity of the overall process, I created another source csv file ("train_test.csv") by combining the rows of both, train.csv and test.csv. By doing so, I could upload one source file and consistently apply my algorithms to one data frame, instead of uploading and working on two separate data files.

Basic Set-up & Loading Packages
All my work was done in R Studio, using R codes.

I identified all the packages & libraries that need to be installed and imported, and loaded them in R.

I uploaded "train_test.csv" file to R Studio, and created the following two data frames:

- Dataframe, houseprice.data = train_test.csv, which is consolidated data of train.csv and test.csv
- Dataframe, no_salesprice.data = created from houseprice.data. I removed 'SalePrice' variable from houseprice.data and assigned the resulting datatframe to this dataframe.

I ensured the data structures, data types, the number of observations and type of variables (i.e. numerical or categorical variables) by applying str(), summary(), head() and tail() to dataframes above. For further details, please see 'Item 2. Data Structure of Datasets' in Appendix.

I also identified if these dataframes had any missing data and counted the number of missing data per each predictor by using colSums(is.na(dataframe)). For further details, please see 'Item 3. Summary of Missing Data per Variables' in Appendix.

Step 3. Gaining Brief Overview of the Data

I used the visualization approach in R to gain further insight on datasets.

For the numerical variables, I ran the scatter plot for each variable with the dependent variable, SalePrice, and observed if there were any linear or non-linear relationships (please see 'Item 4.1 – Scatter Plots for Numerical Variables' in Appendix). Based on the scatter plot, I found that the following predictors had the linear relationship with the 'SalePrice' variable:

- OverllQual
- YearBuilt
- GrLivArea
- GarageCars
- GarageArea
- TotalBstmtSF
- 1stFlrSF
- FullBath
- TotRmsAbvGrd

As for 'SalePrice' variable, I plotted the histogram to observe the skewness of datasets (please see 'Item 4.2 – Histogram for Dependent Variable, SalePrice' in Appendix). The histogram showed the positive skewness of data with the longer and fatter tail on the right. When applying the logarithm on the 'SalePrice' variable, the histogram showed the normal distribution.

From the visualization above, I decided to use both, the multiple linear regression and log transformation for building the model.

<u>Pre-Processing Data</u>

Step 4. Separating Id and SalePrice

Before moving on with pre-processing data, I separated 'Id', index variable, and 'SalePrice' dependent variable from the source datafames, and assigned them to a new dataframe, 'add.df' in R.

Step 5. Dealing with the Missing Data, 'NA'

In order to resolve the missing data issue, I approached the numerical and categorial variables differently.

First, I checked the summary of missing data for overall variables (please see ‘‘Item 3. Summary of Missing Data per Variables’ in Appendix) and the results of the visualization above. I also ran the md.pattern() in R to see the visual distribution of missing data in the dataframe, ‘no_salesprice.data’. As this competition is more quantitative in nature, the numerical variables would impact the performance of models more than the categorical variables, which are qualitative. Through this observation, I decided to take the different approach for the numerical variables and categorical variables.

This decision necessitated me to separate the numerical variables from the categorical variables in dataframes, which included both of these variables. For this task, I needed to assign my source dataframe to a new dataframe, which is named, ‘missing_no_salesprice.data’, and excluded the index variable, ‘Id’, and the dependent variable, ‘SalePrice’.

Before I separated these two types of variables, I identified any variables that should be removed from the source dataframe, ‘houseprice.data’, as those variables would be noises and interfere with the performance of the models. Based on the visualization results, I decided to remove ‘LotFrontage’, numerical variable from two source dataframes, ‘houseprice.data’ and ‘no_salesprice.data’ – this variable did not show any meaningful relationship with the ‘SalePrice’ variable in the scatter plot and yet, had 486 missing data, which was quite material.

For separating the numerical variables from the categorical variables, based on str() in R, the data type of all the numerical variables was integer and the data type of all the categorical variables was factor.
Using the different data types, I assigned the numerical variables and the categorical variables each to new dataframes - ‘fac.data’ dataframe for categorical variables and ‘int.data’ dataframe for numerical variables.

Missing Data in Categorical Variables

After taking the dataframe, ‘fac.data’ from above, I converted all the data in this dataframe from the factor data type to character data type, and replaced all the missing data, ‘NA’ with the string, “None”. This process filled all the ‘NA’, missing data with the string, “None” values in ‘fac.data’ dataframe.

Skipping to a few steps ahead in R, to avoid ‘new level not recognizable’ related errors in the categorical variables when running the multiple linear regression and log transformation, I selected the categorical variables that had the number of missing data less than 100. Using the PIVOT Table in the source csv file, I estimated most common data values for each variable. Then, in R, for each variable, I replaced missing data, ‘NA’, with each applicable, most common data value as follows:

Categorial Variable	**Most Common Data Value**
MSZoning	RL
Utilities	AllPub
Exterior1st	VinylSd
Exterior2nd	VinylSd
BsmtQual	TA
BsmtCond	TA

BsmtExposure	No
BsmtFinType1	Unf
KitchenQual	TA
Functional	Typ
SaleType	WD

After resolving the missing data issue with the categorical variables in 'fac.data' datatframe, I converted all the data values in 'fac.data' from factor data type to character data type, and assigned to a new dataframe, 'chr.data'.

Missing Data in Numerical Variables

After resolving the missing data issue in 'fac.data' dataframe, I combined this dataframe with 'int.data' dataframe that contained all the numerical variables, and assigned this combined dataset to a new dataframe, 'alldata.df'. I proceeded by using the Random Forest simulation in the Mice package to fill all the missing numerical data, 'NA' with the random, system assigned data values through this model.

The filled dataset was then re-assigned to te existing dataframe, 'alldata.df'.

Step 6. Wrapping up Pre-processing

I combined the following two dataframes and assigned the combined datasets to a new dataframe, 'houseprice.df':

- Dataframe, 'alldatadf' that contains both, numerical and categorical variables without missing data
- Dataframe, 'add.df' that contains 'Id', index variable, and 'SalePrice', dependent variable

As the order of the columns in the dataframe, 'houseprice.df' was not consistent with the source dataframe, 'houseprice.data', I re-ordered the columns in 'houseprice.df' to make the column order same as the order in the source dataframe, 'houseprice.data' in R.

Also, the following variables had the names starting with number, which could be problematic when running the regression models in the later stage, and thus, I changed their variable names within the dataframe, 'houseprice.df', in R:

Original Variable Name	Changed Variable Name
1stFlrSF	X1stFlrSF
2ndFlrSF	X2ndFlrSF
3SsnPorch	X3SsnPorch

Finally, I checked the distribution of missing data in the new dataframe, 'houseprice.df' to ensure that there is no 'NA', the order of the dataframe columns, and data structure.

Train/Test Split

For splitting the train dataset from the test dataset, the dataframe, 'houseprice.df', where data values from Id 1 to Id 1460 represented the train.csv, and from Id 1461 to Id 2919 represented the test.csv file.

I set up the training and testing datasets as follows, which was consistently used throughout the models developed:

- houseprice.df.training <-subset(houseprice.df, Id<=1460)
- houseprice.df.prediction <-subset(houseprice.df, Id>=1461)

Depending on the models used, the dataframe, 'houseprice.df.prediction', was assigned to a new dataframe, 'houseprice.df.testing', and used for testing before prediction.

There were a little variations in the names of the dataframes for different models developed, but the structure of the training, testing and predction datasets was the same.

Model / Algorithm Selection

I used the following regressions in order listed:

- Multiple Linear Regression
- Log Transformation -> Log-Linear
- Lasso Regression
- Ridge Regression
- SVM (Support Vector Machine)
- Hybrid of SVM, Log Transformation & Lasso Regression

Training / Prediction

Multiple Linear Regression

I included all the numerical and categorical variables in the training dataset and developed the model as follows:

fit.linear.reg <- lm(SalePrice ~ MSSubClass + MSZoning + LotArea + Street + Alley + LotShape + LandContour + Utilities + LotConfig + LandSlope + Neighborhood + Condition1 + Condition2 + BldgType + HouseStyle + OverallQual + OverallCond + YearBuilt + YearRemodAdd + RoofStyle + RoofMatl + Exterior1st + Exterior2nd + MasVnrType + MasVnrArea + ExterQual + ExterCond + Foundation + BsmtQual + BsmtCond + BsmtExposure + BsmtFinType1 + BsmtFinSF1 + BsmtFinType2 + BsmtFinSF2 + BsmtUnfSF + TotalBsmtSF + Heating + HeatingQC + CentralAir + Electrical + X1stFlrSF + X2ndFlrSF + LowQualFinSF + GrLivArea + BsmtFullBath + BsmtHalfBath + FullBath + HalfBath + BedroomAbvGr + KitchenAbvGr + KitchenQual + TotRmsAbvGrd + Functional + Fireplaces + FireplaceQu + GarageType + GarageYrBlt + GarageFinish + GarageCars + GarageArea + GarageQual + GarageCond + PavedDrive + WoodDeckSF + OpenPorchSF + EnclosedPorch + X3SsnPorch + ScreenPorch + PoolArea + PoolQC + Fence + MiscFeature + MiscVal + MoSold + YrSold + SaleType + SaleCondition, data=houseprice.df.training)

I ran the statistical test to review the p-values, R-squared and Adjusted-R squared.
I also assessed the regression by plotting the Residual vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage, and Residual Density plots.

Adjusted R-squared was 0.9192, which appeared to be very high in terms of overall fit of the model. The residuals plots did not show the sign of heteroskedasticity, but the normality of residuals was questionable, as the Residual Density plot was not normally distributed, but a little skewed to the left. There was also an issue of outliers.

Log Transformation -> Log-Linear

I applied log to 'SalePrice', dependent variable only, and included all the numerical and categorical variables in the training dataset and developed the model as follows:

fit.log <- lm(log(SalePrice) ~ MSSubClass + MSZoning + LotArea + Street + Alley + LotShape + LandContour + Utilities + LotConfig + LandSlope + Neighborhood + Condition1 + Condition2 + BldgType + HouseStyle + OverallQual + OverallCond + YearBuilt + YearRemodAdd + RoofStyle + RoofMatl + Exterior1st + Exterior2nd + MasVnrType + MasVnrArea + ExterQual + ExterCond + Foundation + BsmtQual + BsmtCond + BsmtExposure + BsmtFinType1 + BsmtFinSF1 + BsmtFinType2 + BsmtFinSF2 + BsmtUnfSF + TotalBsmtSF + Heating + HeatingQC + CentralAir + Electrical + X1stFlrSF + X2ndFlrSF + LowQualFinSF + GrLivArea + BsmtFullBath + BsmtHalfBath + FullBath + HalfBath + BedroomAbvGr + KitchenAbvGr + KitchenQual + TotRmsAbvGrd + Functional + Fireplaces + FireplaceQu + GarageType + GarageYrBlt + GarageFinish + GarageCars + GarageArea + GarageQual + GarageCond + PavedDrive + WoodDeckSF + OpenPorchSF + EnclosedPorch + X3SsnPorch + ScreenPorch + PoolArea + PoolQC + Fence + MiscFeature + MiscVal + MoSold + YrSold + SaleType + SaleCondition, data=houseprice.df.training)

I ran the statistical test to review the p-values, R-squared and Adjusted-R squared.
I also assessed the regression by plotting the Residual vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage, and Residual Density plots.

The results of the statistical test improved from the results of the statistical test in the Multiple Linear Regression.

Adjusted R-squared was 0.9321, which appeared to be much higher than the Multiple Linear Regression and with respect to the overall fit of the model. The residuals plots did not show the sign of heteroskedasticity. The normality of residuals also looked good per he Residual Density plot, as the plot was normally distributed. There was also an issue of outliers – I did not remove these outliers but kept in the datasets.

Lasso Regression

I applied the 'SalePrice' in training dataset to y, and model matrix of all the numerical and categorical variables in the dataframe, 'houseprice.df', to X.

Training and testing datasets were built from y and X, but consistently with the section, 'Train/Test Split' above. The optimal penalty parameter and lambda were calculated and applied to the Lasso Regression model, subsequently.

Ridge Regression

The application of dataset to y and X, and building the training and testing datasets were very similar to the Lasso Regression above.

The optimal penalty parameter and lambda were calculated and applied to the Ridge Regression model, subsequently.

SVM (Support Vector Machine)

I researched further to improve my RMSE score and position in this Kaggle Competition, and this model drastically improved both, my scores and position.

As for the categorical variables, this model only works with the factor data type and thus, did not take the character data type in my categorical variables. Thus, I had to convert the character data type of all the categorical variables to the factor data type.

Also, I conducted tests with the different tuning parameter cost, ranging from 0.1 to 10, and found that the tuning parameter of 3 produced the best result.

Hybrid of SVM, Log Transformation & Lasso Regression

In order to further improve my position in the Leaderboard, I tried the hybrid approach by mixing the three models (SVM, Lasso Regression and Log Transformation) and two models (SVM and Lasso Regression). The model performance improved but still, the performance of this methodology was slightly weaker than the performance from the SVM model.

<u>Evaluate the Performance of Models</u>

The evaluation of the performance of the models was done through the RMSE scores, which were computed in the Kaggle competition site, each time I uploaded my submission csv file. RMSE scores and the improvement of my position in the Leaderboard are summarized in the following table:

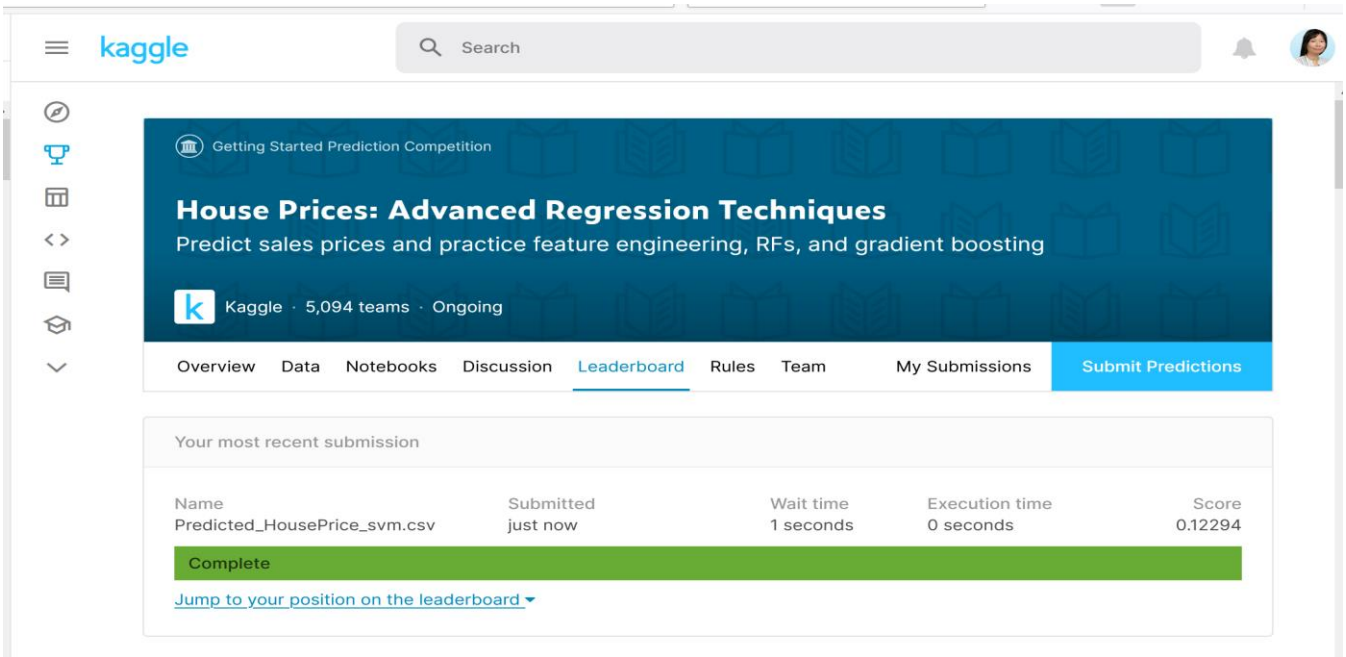
Model Used	RMSE Scores per Kaggle	My Position in the Leaderboard
Multiple Linear Regression	0.19732	4148
Log Transformation (Log-Linear)	0.15448	3411
Lasso Regression	0.13789	2269
Ridge Regression	0.14328	Did not improve my position
SVM	0.12294	767
Hybrid 1 – Three models mix	0.12785	Did not improve my position
Hybrid 2 – Two models mix	0.12378	Did not improve my position

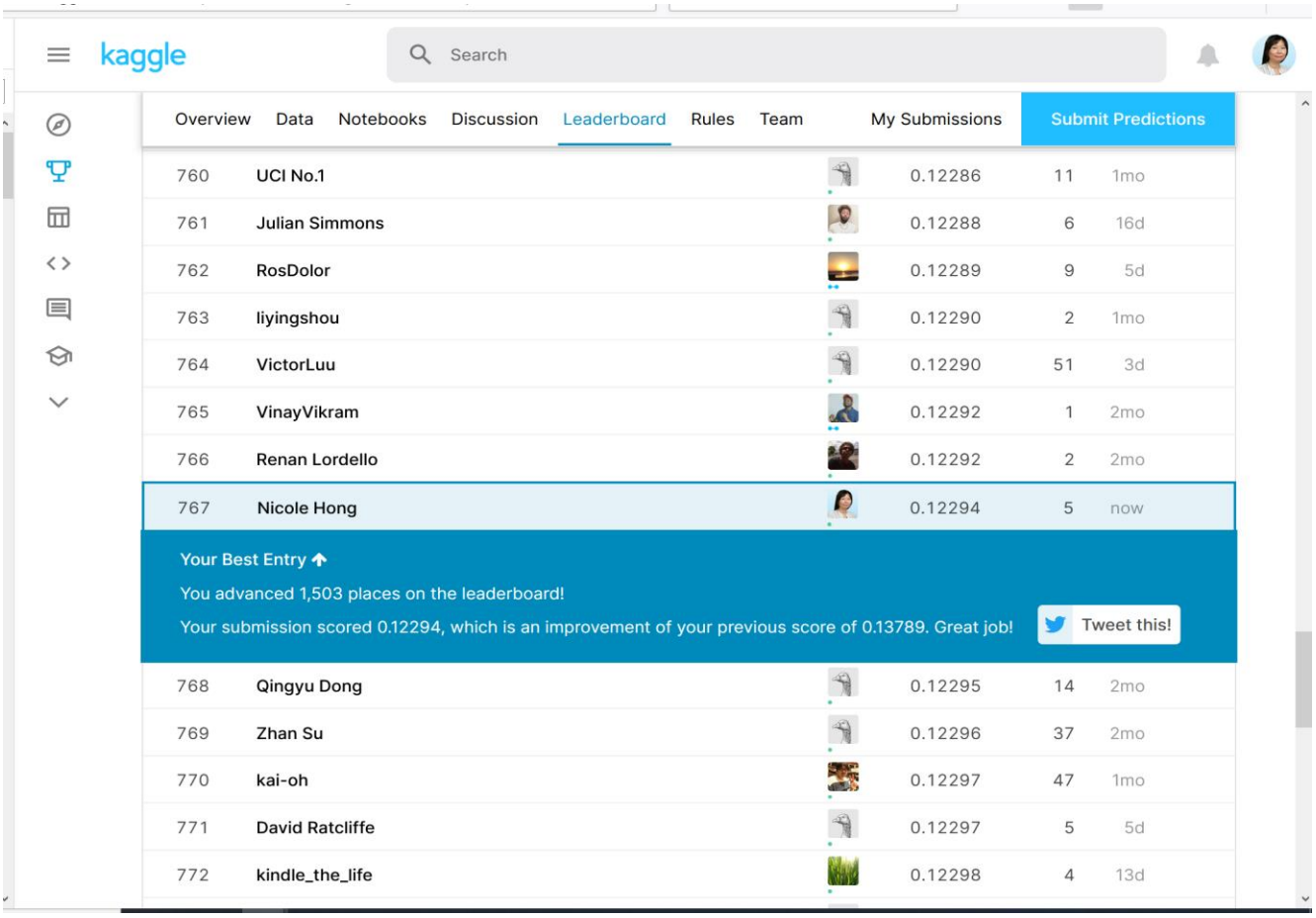
Conclusion

I successfully achieved top 15% on the Leaderboard – I initially started with the multiple linear regression to gain the better insight of the model performance with the given datasets and progressed further to the SVM model and then Hybrid approach.

Appendix

Item 1. Screenshot of my position in the Leaderboard in the Kaggle Competition





Item 2. Data Structure of Datasets

```
> str(houseprice.data)
'data.frame':   2919 obs. of  81 variables:
 $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
 $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
 $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
 $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
 $ Street       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
 $ Alley        : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA NA NA NA ...
 $ LotShape     : Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 1 1 1 1 4 1 4 4 ...
 $ LandContour  : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ Utilities    : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1 1 ...
 $ LotConfig    : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5 1 5 1 ...
 $ LandSlope    : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
 $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14 12 21 17 18 4 ...
 $ Condition1   : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5 1 1 ...
 $ Condition2   : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 1 ...
 $ BldgType     : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1 2 ...
 $ HouseStyle   : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6 1 2 ...
 $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
 $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
 $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
 $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
 $ RoofStyle    : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ RoofMatl     : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Exterior1st  : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 13 13 13 7 4 9 ...
 $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 14 14 14 7 16 9 ..
```

```
 $ MasVnrType   : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4 4 3 3 ...
 $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
 $ ExterQual    : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 4 3 4 3 4 4 4 ...
 $ ExterCond    : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
 $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...
 $ BsmtQual     : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 3 3 4 3 3 1 3 4 4 ...
 $ BsmtCond     : Factor w/ 4 levels "Fa","Gd","Po",..: 4 4 4 2 4 4 4 4 4 4 ...
 $ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4 4 ...
 $ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1 6 3 ...
 $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
 $ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 6 6 6 6 6 6 6 2 6 6 ...
 $ BsmtFinSF2   : int  0 0 0 0 0 0 0 32 0 0 ...
 $ BsmtUnfSF    : int  150 284 434 540 490 64 317 216 952 140 ...
 $ TotalBsmtSF  : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
 $ Heating      : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ HeatingQC    : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 1 3 1 1 1 1 3 1 ...
 $ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ Electrical   : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2 5 ...
 $ X1stFlrSF    : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
 $ X2ndFlrSF    : int  854 0 866 756 1053 566 0 983 752 0 ...
 $ LowQualFinSF : int  0 0 0 0 0 0 0 0 0 0 ...
 $ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
 $ BsmtFullBath : int  1 0 1 1 1 1 1 1 0 1 ...
 $ BsmtHalfBath : int  0 1 0 0 0 0 0 0 0 0 ...
 $ FullBath     : int  2 2 2 1 2 1 2 2 2 1 ...
 $ HalfBath     : int  1 0 1 0 1 1 0 1 0 0 ...
 $ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
 $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
 $ KitchenQual  : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4 4 ...
 $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
 $ Functional   : Factor w/ 7 levels "Maj1","Maj2",..: 7 7 7 7 7 7 7 7 3 7 ...
 $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
 $ FireplaceQu  : Factor w/ 5 levels "Ex","Fa","Gd",..: NA 5 5 3 5 NA 3 5 5 5 ...
 $ GarageType   : Factor w/ 6 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2 6 2 ...
 $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
 $ GarageFinish : Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3 2 ...
 $ GarageCars   : int  2 2 2 3 3 2 2 2 2 1 ...
 $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
 $ GarageQual   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 2 3 ...
 $ GarageCond   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
 $ PavedDrive   : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...
 $ WoodDeckSF   : int  0 298 0 0 192 40 255 235 90 0 ...
 $ OpenPorchSF  : int  61 0 42 35 84 30 57 204 0 4 ...
 $ EnclosedPorch: int  0 0 0 272 0 0 0 228 205 0 ...
 $ X3SsnPorch   : int  0 0 0 0 0 320 0 0 0 0 ...
 $ ScreenPorch  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PoolQC       : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA NA NA NA ...
 $ Fence        : Factor w/ 4 levels "GdPrv","GdWo",..: NA NA NA NA NA 3 NA NA NA NA ...
 $ MiscFeature  : Factor w/ 4 levels "Gar2","Othr",..: NA NA NA NA NA 3 NA 3 NA NA ...
 $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
 $ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...
 $ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
 $ SaleType     : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9 9 9 9 ...
 $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5 5 1 5 ...
 $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 11800
.

> summary(houseprice.data)

       Id           MSSubClass        MSZoning     LotFrontage
 Min.   :   1.0   Min.   : 20.00   C (all):  25   Min.   : 21.00
 1st Qu.: 730.5   1st Qu.: 20.00   FV     : 139   1st Qu.: 59.00
 Median :1460.0   Median : 50.00   RH     :  26   Median : 68.00
 Mean   :1460.0   Mean   : 57.14   RL     :2265   Mean   : 69.31
```

```
3rd Qu.:2189.5   3rd Qu.: 70.00   RM     : 460   3rd Qu.: 80.00
Max.   :2919.0   Max.   :190.00   NA's   :   4   Max.   :313.00
                                                NA's   :486

   LotArea        Street      Alley      LotShape   LandContour  Utilities
Min.   :  1300   Grvl:  12   Grvl: 120   IR1: 968   Bnk: 117    AllPub:2916
1st Qu.:  7478   Pave:2907   Pave:  78   IR2:  76   HLS: 120    NoSeWa:   1
Median :  9453               NA's:2721   IR3:  16   Low:  60    NA's  :   2
Mean   : 10168                           Reg:1859   Lvl:2622
3rd Qu.: 11570
Max.   :215245


  LotConfig    LandSlope   Neighborhood    Condition1     Condition2
Corner : 511   Gtl:2778   NAmes  : 443   Norm   :2511   Norm   :2889
CulDSac: 176   Mod: 125   CollgCr: 267   Feedr  : 164   Feedr  :  13
FR2    :  85   Sev:  16   OldTown: 239   Artery :  92   Artery :   5
FR3    :  14              Edwards: 194   RRAn   :  50   PosA   :   4
Inside :2133              Somerst: 182   PosN   :  39   PosN   :   4
                          NridgHt: 166   RRAe   :  28   RRNn   :   2
                          (Other):1428   (Other):  35   (Other):   2

  BldgType      HouseStyle    OverallQual      OverallCond      YearBuilt
1Fam  :2425   1Story :1471   Min.   : 1.000   Min.   :1.000   Min.   :1872
2fmCon:  62   2Story : 872   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954
Duplex: 109   1.5Fin : 314   Median : 6.000   Median :5.000   Median :1973
Twnhs :  96   SLvl   : 128   Mean   : 6.089   Mean   :5.565   Mean   :1971
TwnhsE: 227   SFoyer :  83   3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2001
              2.5Unf :  24   Max.   :10.000   Max.   :9.000   Max.   :2010
              (Other):  27

 YearRemodAdd    RoofStyle       RoofMatl     Exterior1st    Exterior2nd
Min.   :1950   Flat   :  20   CompShg:2876   VinylSd:1025   VinylSd:1014
1st Qu.:1965   Gable  :2310   Tar&Grv:  23   MetalSd: 450   MetalSd: 447
Median :1993   Gambrel:  22   WdShake:   9   HdBoard: 442   HdBoard: 406
Mean   :1984   Hip    : 551   WdShngl:   7   Wd Sdng: 411   Wd Sdng: 391
3rd Qu.:2004   Mansard:  11   ClyTile:   1   Plywood: 221   Plywood: 270
Max.   :2010   Shed   :   5   Membran:   1   (Other): 369   (Other): 390
                              (Other):   2   NA's   :   1   NA's   :   1
  MasVnrType     MasVnrArea     ExterQual ExterCond  Foundation   BsmtQual
BrkCmn :  25   Min.   :   0.0   Ex: 107   Ex:  12   BrkTil: 311   Ex  : 258
BrkFace: 879   1st Qu.:   0.0   Fa:  35   Fa:  67   CBlock:1235   Fa  :  88
None   :1742   Median :   0.0   Gd: 979   Gd: 299   PConc :1308   Gd  :1209
Stone  : 249   Mean   : 102.2   TA:1798   Po:   3   Slab  :  49   TA  :1283
NA's   :  24   3rd Qu.: 164.0             TA:2538   Stone :  11   NA's:  81
               Max.   :1600.0                       Wood  :   5
               NA's   :23

BsmtCond    BsmtExposure BsmtFinType1   BsmtFinSF1     BsmtFinType2
Fa  : 104   Av  : 418    ALQ :429     Min.   :   0.0   ALQ :  52
Gd  : 122   Gd  : 276    BLQ :269     1st Qu.:   0.0   BLQ :  68
Po  :   5   Mn  : 239    GLQ :849     Median : 368.5   GLQ :  34
TA  :2606   No  :1904    LwQ :154     Mean   : 441.4   LwQ :  87
NA's:  82   NA's:  82    Rec :288     3rd Qu.: 733.0   Rec : 105
                         Unf :851     Max.   :5644.0   Unf :2493
                         NA's: 79     NA's   :1        NA's:  80

  BsmtFinSF2        BsmtUnfSF       TotalBsmtSF      Heating     HeatingQC
Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Floor:   1   Ex:1493
1st Qu.:   0.00   1st Qu.: 220.0   1st Qu.: 793.0   GasA :2874   Fa:  92
Median :   0.00   Median : 467.0   Median : 989.5   GasW :  27   Gd: 474
Mean   :  49.58   Mean   : 560.8   Mean   :1051.8   Grav :   9   Po:   3
3rd Qu.:   0.00   3rd Qu.: 805.5   3rd Qu.:1302.0   OthW :   2   TA: 857
Max.   :1526.00   Max.   :2336.0   Max.   :6110.0   Wall :   6
NA's   :1         NA's   :1        NA's   :1
```

```
CentralAir Electrical     X1stFlrSF      X2ndFlrSF       LowQualFinSF
N: 196     FuseA: 188   Min.   : 334   Min.   :   0.0   Min.   :   0.000
Y:2723     FuseF:  50   1st Qu.: 876   1st Qu.:   0.0   1st Qu.:   0.000
           FuseP:   8   Median :1082   Median :   0.0   Median :   0.000
           Mix  :   1   Mean   :1160   Mean   : 336.5   Mean   :   4.694
           SBrkr:2671   3rd Qu.:1388   3rd Qu.: 704.0   3rd Qu.:   0.000
           NA's :   1   Max.   :5095   Max.   :2065.0   Max.   :1064.000


  GrLivArea     BsmtFullBath     BsmtHalfBath        FullBath
Min.   : 334   Min.   :0.0000   Min.   :0.00000   Min.   :0.000
1st Qu.:1126   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000
Median :1444   Median :0.0000   Median :0.00000   Median :2.000
Mean   :1501   Mean   :0.4299   Mean   :0.06136   Mean   :1.568
3rd Qu.:1744   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.000
Max.   :5642   Max.   :3.0000   Max.   :2.00000   Max.   :4.000
               NA's   :2        NA's   :2

   HalfBath       BedroomAbvGr   KitchenAbvGr   KitchenQual  TotRmsAbvGrd
Min.   :0.0000   Min.   :0.00   Min.   :0.000   Ex  : 205   Min.   : 2.000
1st Qu.:0.0000   1st Qu.:2.00   1st Qu.:1.000   Fa  :  70   1st Qu.: 5.000
Median :0.0000   Median :3.00   Median :1.000   Gd  :1151   Median : 6.000
Mean   :0.3803   Mean   :2.86   Mean   :1.045   TA  :1492   Mean   : 6.452
3rd Qu.:1.0000   3rd Qu.:3.00   3rd Qu.:1.000   NA's:   1   3rd Qu.: 7.000
Max.   :2.0000   Max.   :8.00   Max.   :3.000               Max.   :15.000


  Functional     Fireplaces     FireplaceQu   GarageType    GarageYrBlt
Typ    :2717   Min.   :0.0000   Ex  :  43   2Types :  23   Min.   :1895
Min2   :  70   1st Qu.:0.0000   Fa  :  74   Attchd :1723   1st Qu.:1960
Min1   :  65   Median :1.0000   Gd  : 744   Basment:  36   Median :1979
Mod    :  35   Mean   :0.5971   Po  :  46   BuiltIn: 186   Mean   :1978
Maj1   :  19   3rd Qu.:1.0000   TA  : 592   CarPort:  15   3rd Qu.:2002
(Other):  11   Max.   :4.0000   NA's:1420   Detchd : 779   Max.   :2207
NA's   :   2                                NA's   : 157   NA's   :159


GarageFinish   GarageCars      GarageArea     GarageQual  GarageCond  PavedDrive
Fin : 719    Min.   :0.000   Min.   :   0.0   Ex  :   3   Ex  :   3   N: 216
RFn : 811    1st Qu.:1.000   1st Qu.: 320.0   Fa  : 124   Fa  :  74   P:  62
Unf :1230    Median :2.000   Median : 480.0   Gd  :  24   Gd  :  15   Y:2641
NA's: 159    Mean   :1.767   Mean   : 472.9   Po  :   5   Po  :  14
             3rd Qu.:2.000   3rd Qu.: 576.0   TA  :2604   TA  :2654
             Max.   :5.000   Max.   :1488.0   NA's: 159   NA's: 159
             NA's   :1       NA's   :1


  WoodDeckSF       OpenPorchSF     EnclosedPorch      X3SsnPorch
Min.   :   0.00   Min.   :  0.00   Min.   :   0.0   Min.   :  0.000
1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:   0.0   1st Qu.:  0.000
Median :   0.00   Median : 26.00   Median :   0.0   Median :  0.000
Mean   :  93.71   Mean   : 47.49   Mean   :  23.1   Mean   :  2.602
3rd Qu.: 168.00   3rd Qu.: 70.00   3rd Qu.:   0.0   3rd Qu.:  0.000
Max.   :1424.00   Max.   :742.00   Max.   :1012.0   Max.   :508.000


 ScreenPorch        PoolArea        PoolQC       Fence      MiscFeature
Min.   :  0.00   Min.   :  0.000   Ex  :   4   GdPrv: 118   Gar2:   5
1st Qu.:  0.00   1st Qu.:  0.000   Fa  :   2   GdWo : 112   Othr:   4
Median :  0.00   Median :  0.000   Gd  :   4   MnPrv: 329   Shed:  95
Mean   : 16.06   Mean   :  2.252   NA's:2909   MnWw :  12   TenC:   1
3rd Qu.:  0.00   3rd Qu.:  0.000               NA's :2348   NA's:2814
Max.   :576.00   Max.   :800.000
```

```
     MiscVal             MoSold           YrSold        SaleType    SaleCondition
 Min.   :    0.00   Min.   : 1.000   Min.   :2006   WD     :2525   Abnorml: 190
 1st Qu.:    0.00   1st Qu.: 4.000   1st Qu.:2007   New    : 239   AdjLand:  12
 Median :    0.00   Median : 6.000   Median :2008   COD    :  87   Alloca :  24
 Mean   :   50.83   Mean   : 6.213   Mean   :2008   ConLD  :  26   Family :  46
 3rd Qu.:    0.00   3rd Qu.: 8.000   3rd Qu.:2009   CWD    :  12   Normal :2402
 Max.   :17000.00   Max.   :12.000   Max.   :2010   (Other):  29   Partial: 245
                                                    NA's   :   1
```

```
   SalePrice
 Min.   : 34900
 1st Qu.:129975
 Median :163000
 Mean   :180921
 3rd Qu.:214000
 Max.   :755000
 NA's   :1459
```

Item 3. Summary of Missing Data per Variables

Variable 1 = Categorical variables with missing data
Variable 2 = SalePrice, dependent variable

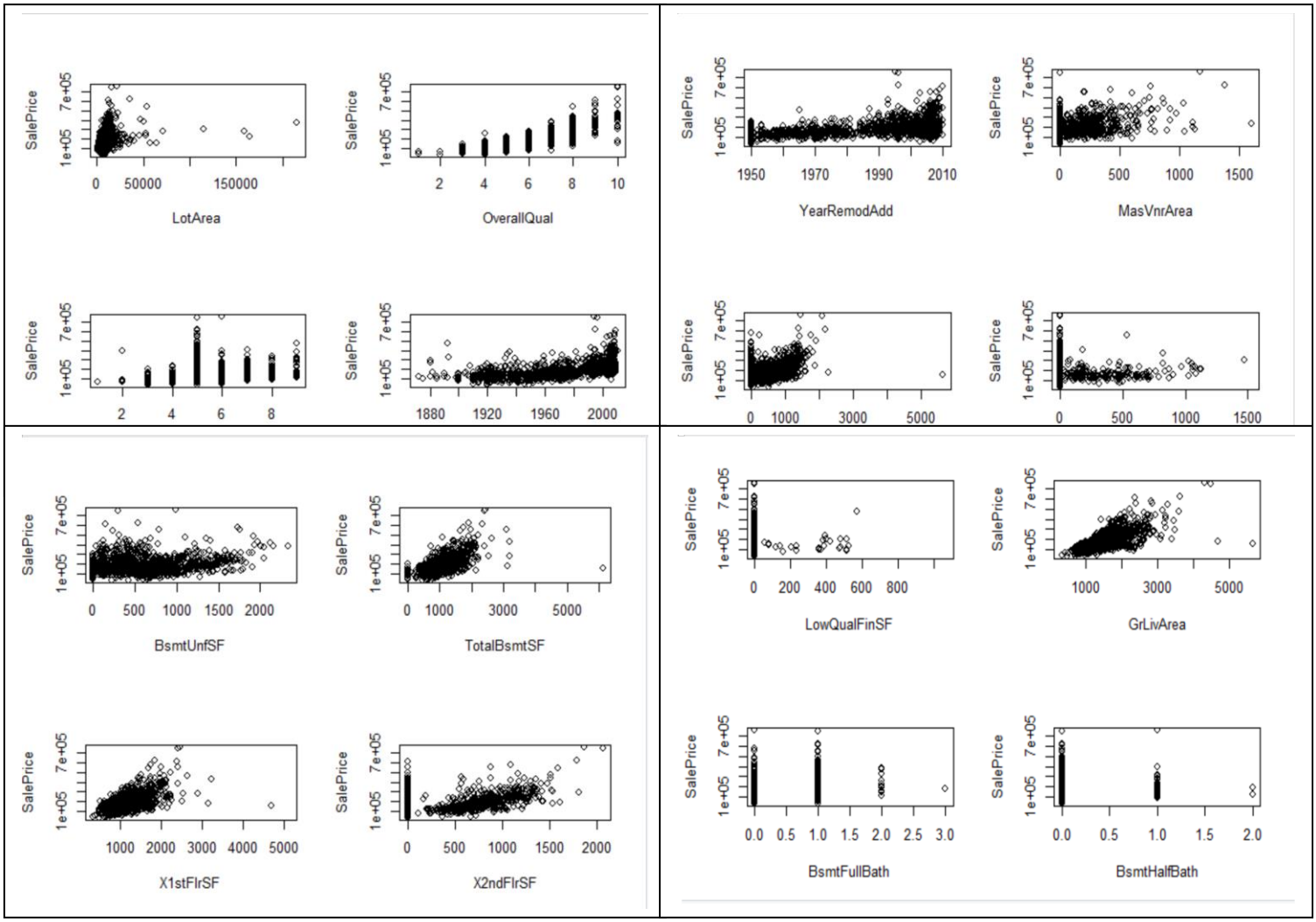
```
> colSums(is.na(houseprice.data))
           Id    MSSubClass      MSZoning   LotFrontage       LotArea
            0             0             4           486             0

       Street         Alley      LotShape   LandContour     Utilities
            0          2721             0             0             2

    LotConfig     LandSlope  Neighborhood    Condition1    Condition2
            0             0             0             0             0

     BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
            0             0             0             0             0

 YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
            0             0             0             1             1

   MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
           24            23             0             0             0

     BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
           81            82            82            79             1

 BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
           80             1             1             1             0

    HeatingQC    CentralAir     Electrical     X1stFlrSF     X2ndFlrSF
            0             0             1             0             0

 LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
            0             0             2             2             0

     HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
            0             0             0             1             0

   Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
            2             0          1420           157           159

 GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
          159             1             1           159           159
```

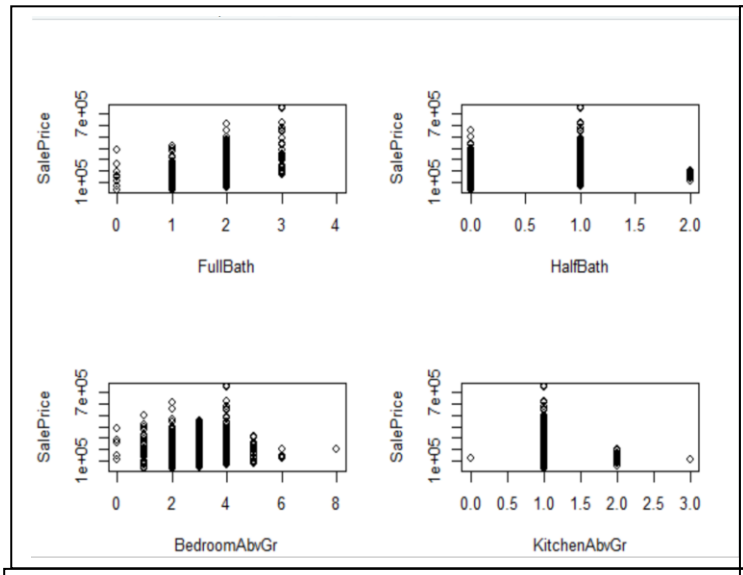
```
     PavedDrive     WoodDeckSF    OpenPorchSF EnclosedPorch     X3SsnPorch
              0              0              0             0              0

    ScreenPorch       PoolArea         PoolQC         Fence    MiscFeature
              0              0           2909          2348           2814

        MiscVal         MoSold         YrSold      SaleType  SaleCondition
              0              0              0             1              0

      SalePrice
           1459
```

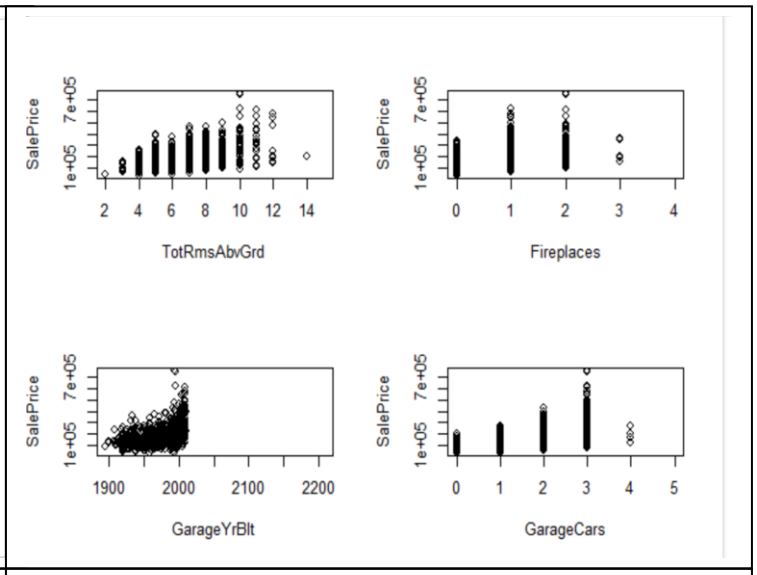
Item 4. Visualization - Plots of Numerical Variables

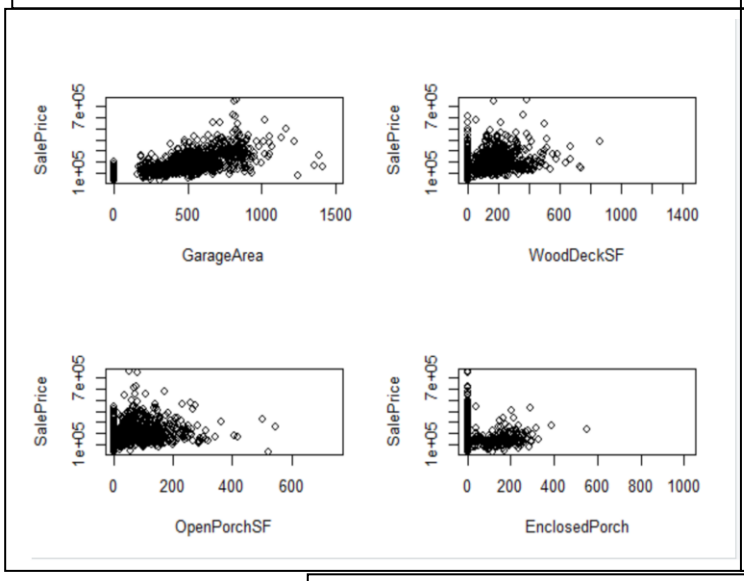
Item 4.1 – Scatter Plots for Numerical Variables

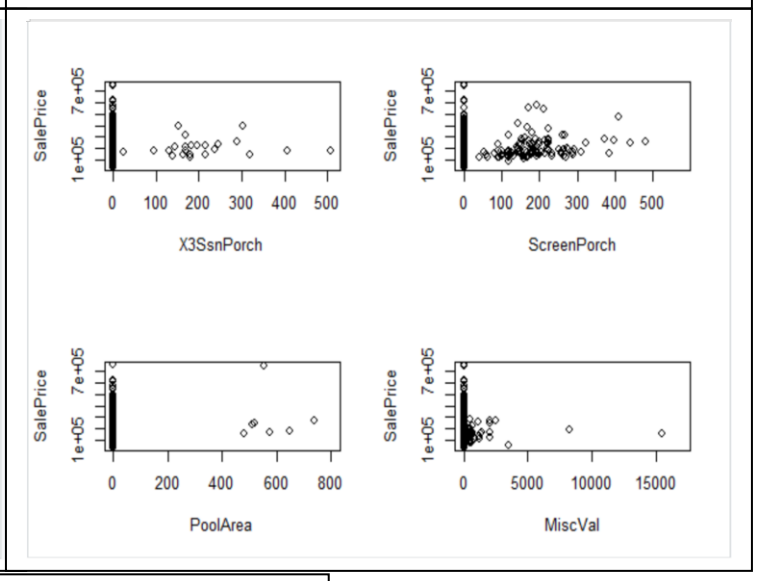
Note: the plot for 'LotFrontage' variable was not done, as this variable was removed, when resolving the issue for the missing data.

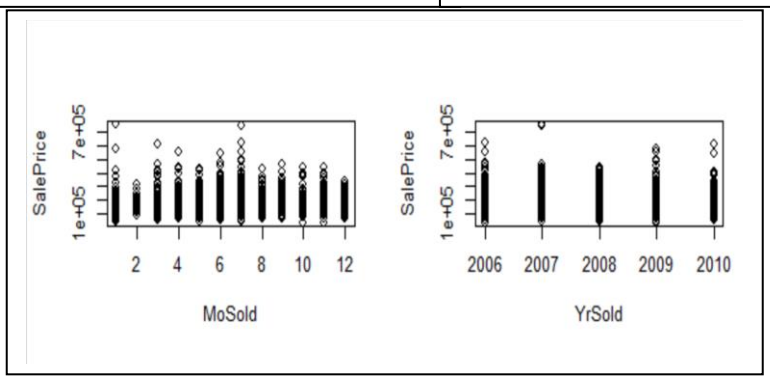



SalePrice
1e+05
7e+05
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr


SalePrice
1e+05
7e+05
TotRmsAbvGrd
Fireplaces
GarageYrBlt
GarageCars


SalePrice
1e+05
7e+05
GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorch


SalePrice
1e+05
7e+05
X3SsnPorch
ScreenPorch
PoolArea
MiscVal


SalePrice
1e+05
7e+05
MoSold
YrSold

Item 4.2 – Histogram for Dependent Variable, SalePrice