Iowa Housing Market Price Regression

Bernardo D'Agostino (III ECO), Mattia Pasqualini (III ECO)

Introduction

Our project's main goal is to develop a model to use to predict at what price the houses in our dataset were sold for and to discover if a data-driven selection of predictive variables can improve models. Other goals of our project are testing different regressions and feature selection methods for price prediction and learning how to apply unsupervised learning algorithms to categorical variables and manage to discover possible ways to manage missing values.

In order to do so, we used a dataset containing information about approximately 1460 houses sold from 2006 to 2010, we first pre-processed our data, then, we apply a series of classification techniques.

This report is organized as follows: in section 1 we describe our original dataset. In section 2 we illustrate the pre-processing procedures we had to implement before undertaking any regression technique in order to make the data tractable.

In section 3 we present the FAMD technique that works as a principal components analysis (PCA) for quantitative variables and as a multiple correspondence analysis (MCA) for qualitative variables

In section 4 we present the results of an attempt to find out whether there was a way to subgroup our observations into natural clusters.

In section 5 we display the final results of k-fold cross-validation on the classification techniques we employed to measure the stability of the models tested.

1The Dataset

The original dataset we adopted for our investigations was published in 2011. It contains 79 variables,46 of which are categorical, related to 1460 houses sold in the city of Ames Iowa in the United States from 2006 to 2010 (Dean De Cock 2011). The dependent variable is SalePrice and it is expressed in USD.

We choose this dataset because it was challenging for its number of features both categorical and continuous, and for its missing values problem.

2 Pre-Processing

Our dataset needed to be processed before operating any kind of operations since some features are very similar in nature to others and this could cause problems for multicollinearity, and also we found a significant amount of missing values in our dataset.

2.1 Missing values

First, we dealt with the high level of missing values within the dataset; in total, we found 6965 missing values distributed across 19 variables. We decided to eliminate the 5 variables with a missing value ratio greater than 15%, and then used a random forest algorithm to impute the remaining 609 missing values.

The imputation algorithm used is part of the missForest package (Stekhoven 2022) and is structured as follows: first, the columns containing missing values are identified, then the missing values are replaced by the mean values for continuous variables, and the mode values for categorical variables. Then in descending order by the percentage of missing values, the columns containing missing values are used as the dependent variable of a random forest model, which uses the observations without NA in the dependent variable as training data and uses the model thus created to predict the values of the dependent variable for the other observations.

This process is carried out repeatedly for a total of five iterations. In this project, we have left the tuning parameters of the model with default values, and we set as a possible future goal an optimization of these parameters to achieve better imputation of the missing variables.

2.2 Scaling and Encoding

Next, we standardized all continuous variables except the dependent variable SalePrice by subtracting the mean from the observations and dividing by the standard deviation (both estimated on the sample):

$$x_{ij}^{(st)} = \frac{x_{ij} - \underline{x}_{ij}}{\hat{\sigma}_j}$$

Regarding the categorical variables we operated a one-hot encoding, whereby each recorded value of each categorical variable is assigned its own new variable, in which the value 1 indicates the presence of the specific value of the categorical variable associated with the new variable. This encoding procedure resulted in an increase in the number of variables in our dataset from 79 to 295.

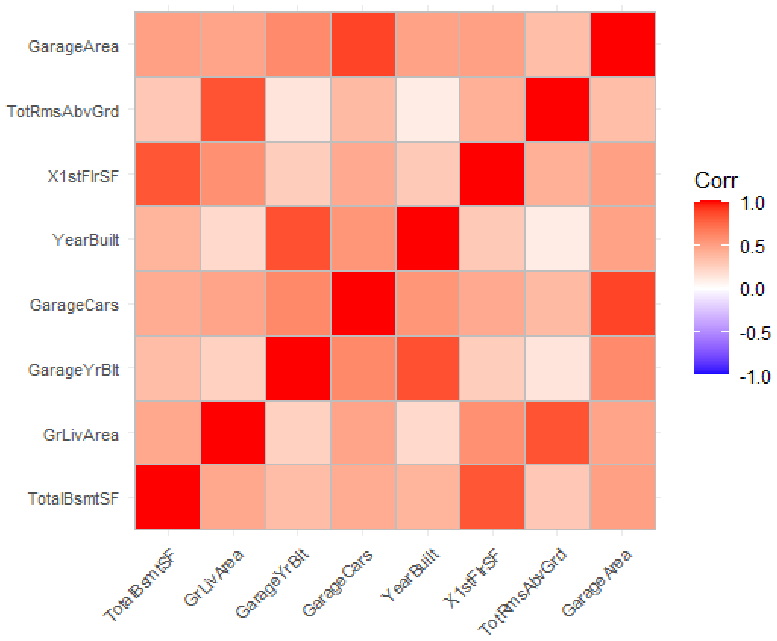
2.3 Multicollinearity

Next, we analysed the continuous variables in the dataset with respect to multicollinearity.

Multicollinearity between variables occurs when, within a regression model, one predictor variable can itself be predicted with some degree of accuracy by the other predictor variables, due to linear associations between the predictors. Multicollinearity can cause problems in parameter estimation within a regression model (Alin 2010), specifically causing an increase in the sample variance of the estimators.

A useful index for detecting the presence of multicollinearity between variables is The Bravais-Pearson linear correlation coefficient, which can be calculated for each pair of variables.

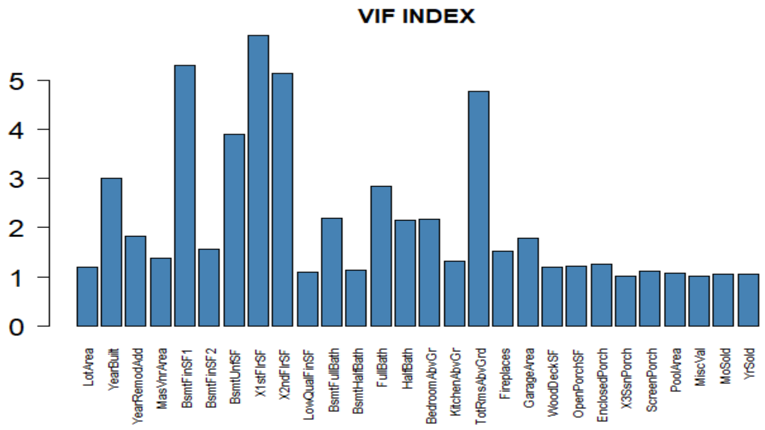
Figure 2.1 Correlation matrix between the values that had a correlation higher than 0.8.



As a result of this analysis, we found 4 correlations between variables greater than 0.8 Figure 2.1, between the variables GarageYrBlt and YearBuilt, TotalBsmtSF and X1stFlrSF, TotRmsAbvGrd and GrLivArea, GarageCars and GarageArea. We then decided to eliminate 1 variable for each of these pairs.

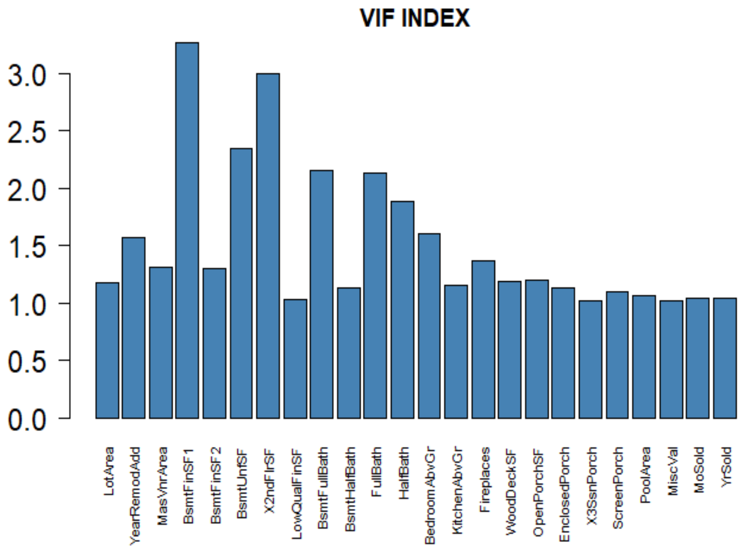
Another useful index to detect multicollinearity within a dataset is the Variance Inflation Factor (VIF) index, which provides an estimate of multicollinearity based on a linear regression of each variable on all the others. It is calculated as the reciprocal of the so-called tolerance (Kim 2019): $VIF(x) = \frac{1}{TOLLERANCE}$; $TOLLERANCE(x) = 1 - R^2(x)$. $R^2(x)$ in this case is the coefficient of determination for the regression of a specific predictor variable x on all other predictor variables.

Figura 2.2 Values of the VIF index for all continuous predictor variables.



It can be seen from Figure 2.2 how the variable X1stFlrSF has the highest value of the VIF index, to reduce multicollinearity, we decided to eliminate this variable, and subsequently it can be seen from Figure 2.3 how there are no variables, among those retained, with a VIF index close to or greater than 5.

Figure 2.3 Values of the VIF index for all continuous predictor variables after the elimination of the variable X1stFlrSF.



3 FAMD

After the pre-processing operations, we move to the FAMD analysis.

Factor Analysis of Mixed Data (FAMD) is the Factorial method devoted to data tables in which a group of individuals is described both by quantitative and qualitative variables (Agès, J. 2004).

FAMD works as a principal components analysis (PCA) for quantitative variables and as a multiple correspondence analysis (MCA) for qualitative variables.

Essentially it is a principal component method dedicated to analysing a data set containing both quantitative and qualitative variables. It makes it possible to analyse the similarity between individuals by taking into account mixed types of variables. Additionally, one can explore the association between all variables, both quantitative and qualitative variables.

Roughly speaking, the FAMD algorithm can be seen as a mix between principal component analysis (PCA) and multiple correspondence analysis (MCA)

Figure 3.1 Results of the Factor Analysis of Mixed Data for Quantitative variables, and for both qualitative and quantitative variables over the first 2 principal components.

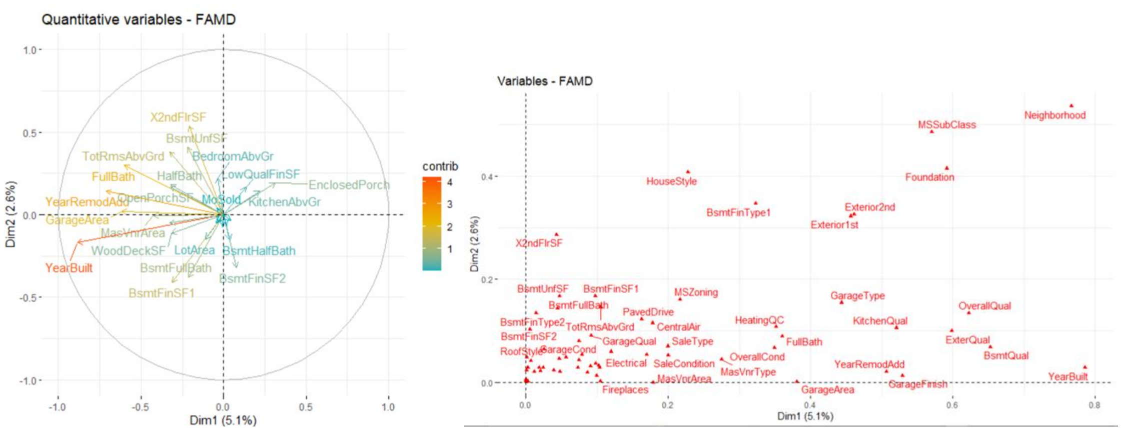
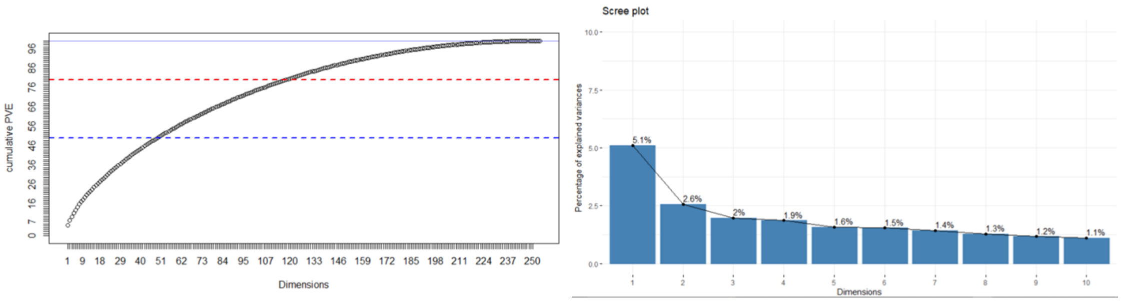


Figure 3.2 Percentage of variability explained by FAMD and scree plot of the first 10 components.



From the analysis in figure 3.1, we can see the variables Neighbourhood and MSS_sub_class, which both refer to the neighbourhood where the house is situated, are the ones that influence both the first 2 principal components, and they are both categorical variables.

We can then see from Figure 3.2 that the first 50 components explain 50% of the variability and that the first 120 components explain 80% of it. So, we choose to compare the performance of an OLS model on the 2 sets in section 5.

4 Clustering

We choose to do our cluster analysis only on the 28 continuous variables that we kept after our analysis on multicollinearity.

First of all, we looked for the optimal number of clusters for the K-means method, we reinitialized the centroids multiple times and we consistently found that 3 was the optimal number of clusters for the 'elbow method'. Then we did a silhouette analysis on the 3 clusters found with this method and we got some promising results Figure 4.1.

Figure 4.1 Results of the Silhouette widths method, and of the optimal number of clusters analysis for k-means clustering.

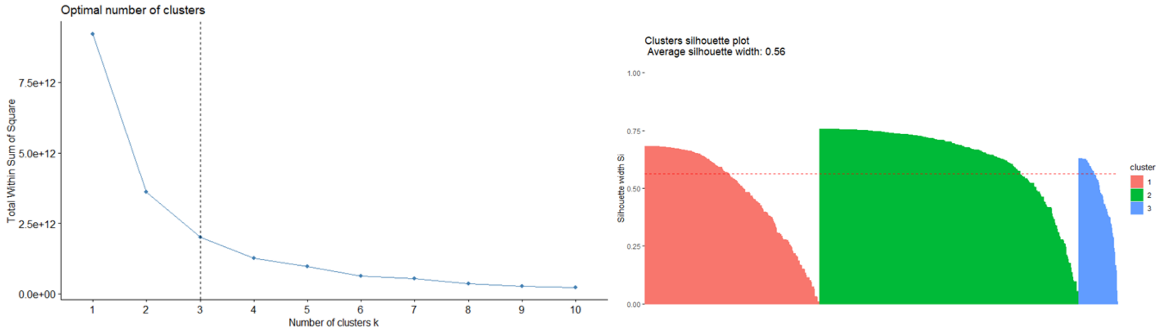
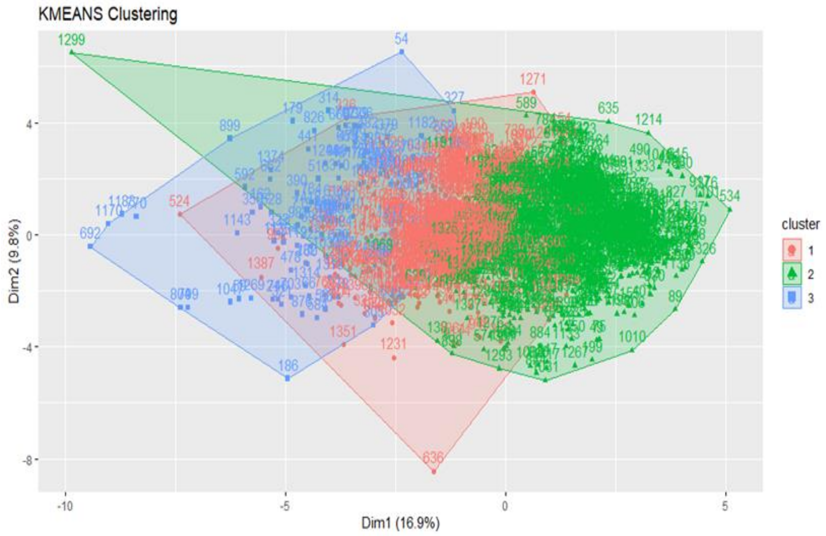


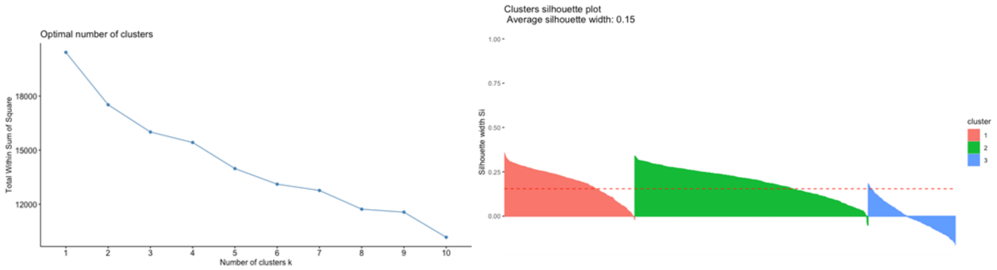
Figure 4.2 Results of the K-means clustering method with 3 clusters.



We then plotted the clusters on the first 2 principal components and found them overlapping but still distinguishable Figure 4.2.

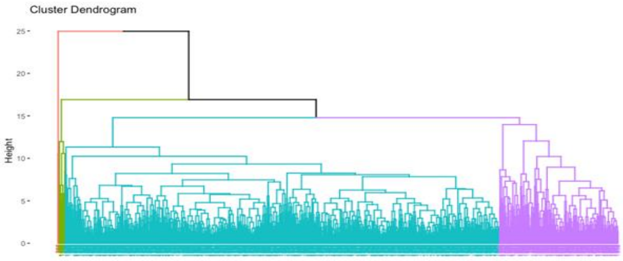
After that of all we tried to find the optimal number of clusters in the dataset for the hierarchical clustering algorithm, but we couldn't find any optimal number of clusters with the elbow method Figure 4.3.

Figure 4.3 Results of the Silhouette widths method, and of the optimal number of clusters analysis for hierarchical clustering.



So, we chose to try a silhouette analysis on3 clusters since that was the optimal amount for the k-means algorithm. But we found less promising results than those achieved by the previous clustering algorithm.

Figure 4.2 Results of the Hierarchical Clustering



5 Supervised Analyses

After that we finished with the un-supervised analyses, we moved to Supervised Analysis, we performed and compared several prediction models. We start presenting Ridge and Lasso Models then we move to the Forward Selection Method and 2 linear regressions performed on the features selected by the Lasso model and on those computed by the FAMD algorithm.

5.1 Method to calculate model stability.

As part of this work, we compared 6 different prediction models, measuring their stability with regard to predictive ability (RMSE), and the number of selected variables.

To compare the stability of each model we divided the dataset into 10 folds of equal size, then we used each of these folds in turn as a test set and the remaining 9 folds as a train set for each of the 6 models compared in this paper. For each of the 10 iterations in this process we estimated the optimal parameters of each model on the current train folds, resulting in 10 different sets of parameters for each model, once we estimated the optimal parameters for each iteration, we trained the model on the 9 train folds and calculated the RMSE of the predictions on test folds.

In this way we compared the box plots and averages of the 10 out-of-sample RMSE values for each model, also comparing the number of variables selected for each iteration.

5.2 Ridge and Lasso Models

The first models we used were the Ridge and Lasso penalty regression models, both of these models attempt to reduce the sampling variability of OLS estimators and overfitting, to do this they introduce a bias in the estimators via a penalty parameter λ, this bias reduces the variance of the estimators trying to minimize the sum of the bias and variance and the out of sample MSE by tuning the penalty parameter (James G., Witten D., Hastie T. and R. Tibshirani 2013).

For both models we performed a tuning process for the penalty parameter λ, estimating the MSE of the models for different levels of λ by cross validation. Following the cross validation, we selected the value of λ that minimized the MSE, and the value of λ within one standard error from the minimum.

Figure 5.1 Results of the cross-validated tuning of the λ for the Ridge and Lasso penalized regression models

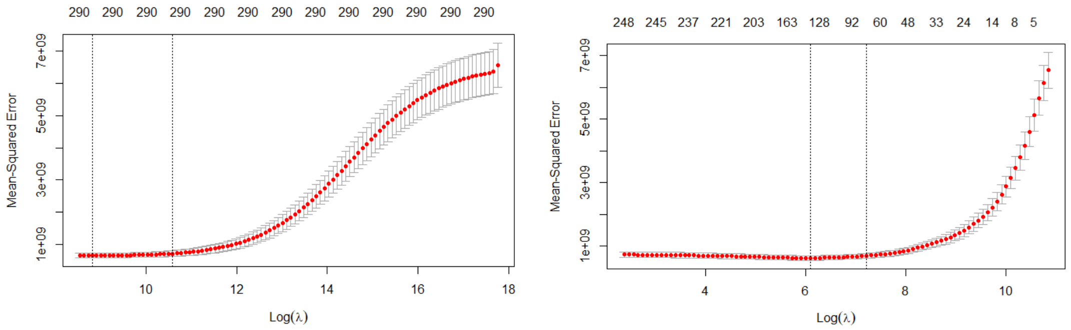
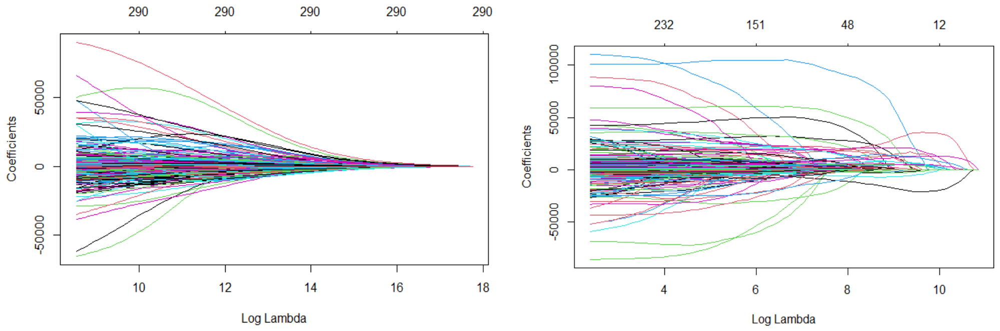


Figure 5.2 Variation of the values of the coefficients for the Ridge and Lasso models with respect to the λ parameter.



In Figure 5.1 it's possible to observe the results of estimation by cross-validation of λ minimum and λ within one standard error from the minimum for the Ridge and Lasso Models. In Figure 5.2, on the other hand, it is possible to observe the values of the coefficients of the models as the penalty parameter λ increases. It can be observed that the Lasso model has fewer non-zero coefficients than the Ridge model for the minimum and within one standard error value of λ, 130, and 80 variables with non-zero coefficients for the respective values of λ. In both models, the highest coefficient is for the categorical variable that represents the highest possible quality for a home.

5.3 Forward Selection

Next, we used the forward selection method, this method sorts the independent variables starting with the variable that creates the model with the lowest RSS, and then iteratively adding the variable that, with the previously selected variables, forms the linear regression model with the lowest RSS. Then we compared through the indices C_P, R^2_{adj} *and* BIC the possible sets of different sizes of the variables sorted through the forward selection method.

It can be seen in Figure 5.3 how the three indices do not provide an unambiguous indication for the optimal number of variables to be selected. Therefore, we compared the stability (see 5.1) of the predictive capability of the models built with the variables selected from the optimal values of the 3 indices cited above. We then found the model with the variables selected by the BIC index to be the one with the lowest average out-of-sample RMSE Figure 5.4.

Figure 5.3 Number of features ordered by forward selection that have the optimal value for the C_P, R^2_{adj} *and* BIC indeces.

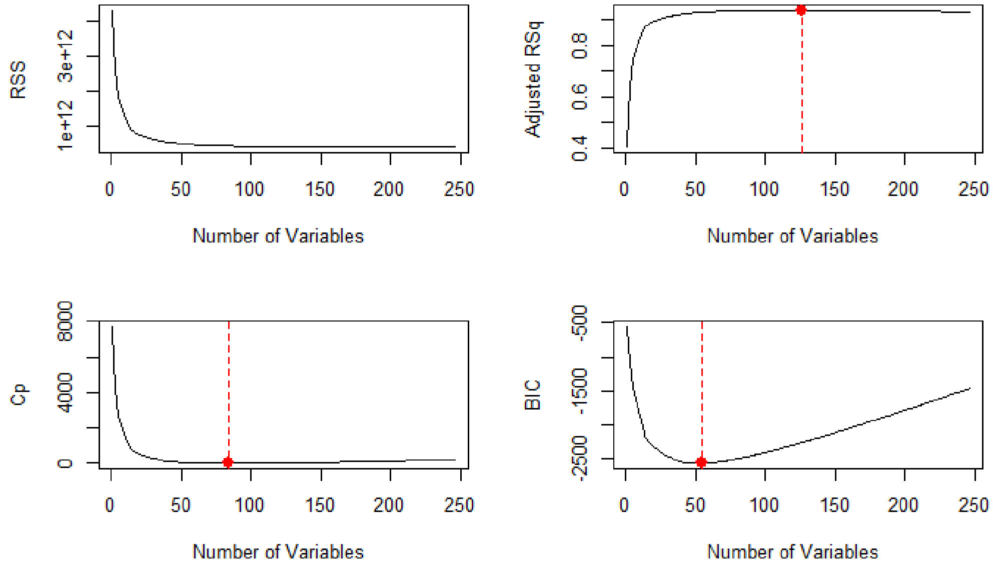
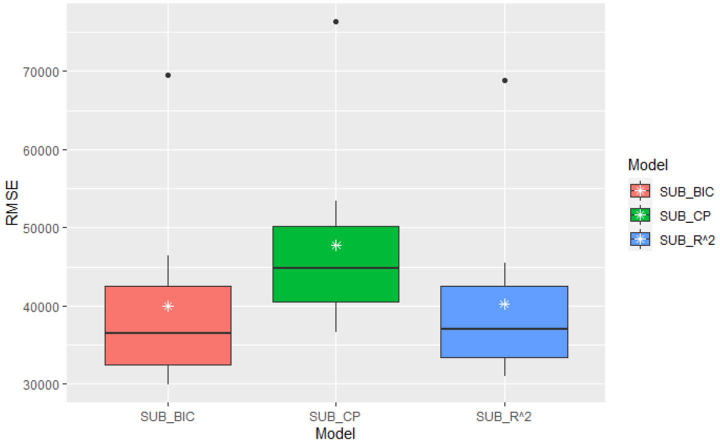


Figure 5.4 Boxplot and average value of the out of sample RMSE over ten folds, of a OLS model that uses the number of feature, ordered by the forward selection methods, for the optimal values of the C_P, R^2_{adj} *and* BIC indeces.

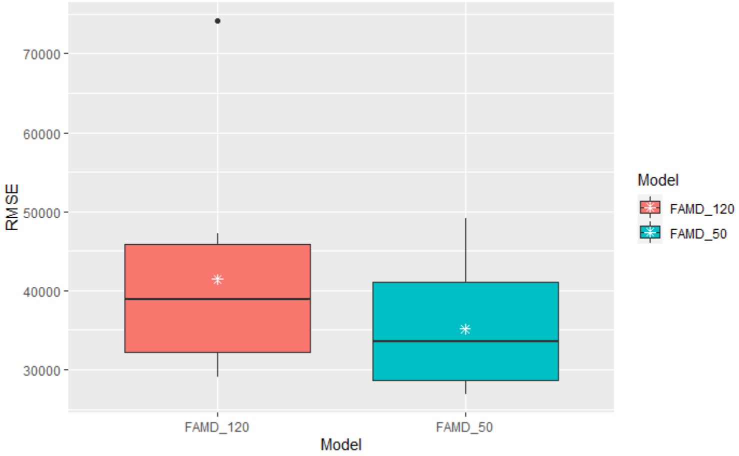


5.3 OLS post FAMD and OLS post Lasso regressions

Finally, we applied a linear regression to the datasets composed of the first 50 and 120 components computed by FAMD (see 3) It can be seen from Figure 5.5, that the model with the first 50 principal components has a lower mean RMSE over the ten folds calculated for model stability.

We also estimated an OLS regression on the dataset composed of the variables with non-zero coefficients estimated by Lasso model with the λ min tuned with cross validation. The regression operated on the variables selected by the Lasso model aims to reduce the Bias introduced by the penalty parameter in the regression (see 5.2) to reduce the out of sample MSE.

Figure 5.5 Boxplot and average values over 10 folds of 2 OLS models that use the first 120 and 50 principal components calculated by the FAMD algorithm (see 2.1).



5.5 Analysis of the results for model stability and performance.

Figure 5.7 shows the box plot of the out of sample RMSE computed for each model over the 10 folds (see 5.1) while Figure 5.6 shows the plot of the individual folds by RMSE and number of variables selected in the model.

It is immediately visible that the model with the greatest variability in performance on the 10 folds is the linear model, and that all models except the forward selection model and the simple OLS model have similar average values for RMSE, between 30,000 and 35,000, it is possible to observe that 3 models have one-fold in particular with an RMSE of over 55,000. These include the forward selection model, which has a fold with an RMSE of 70,000, which raises the average RMSE considerably and brings it over 35,000.

Figure 5.6 Plot of the RMSE and number of features selected for each fold computed for each model.

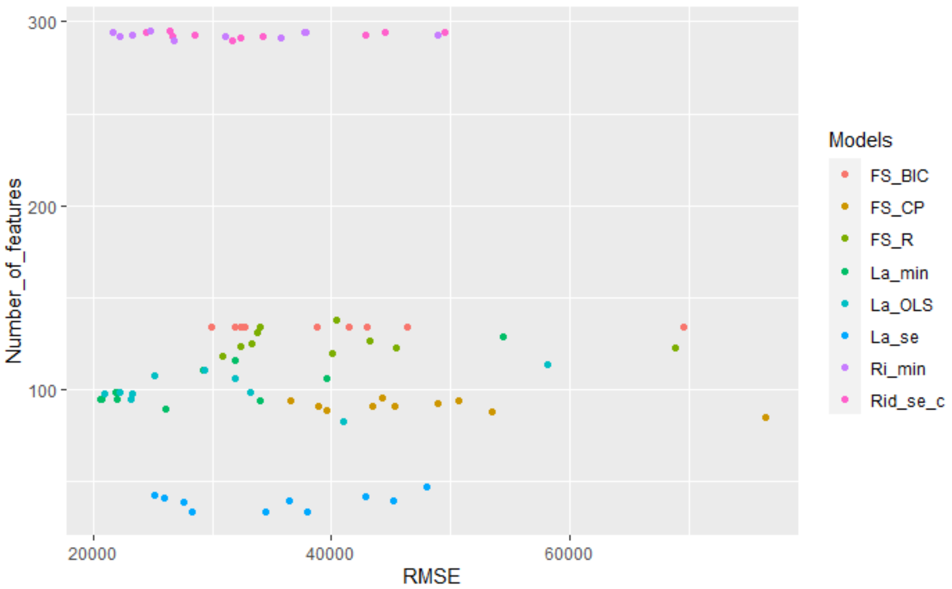
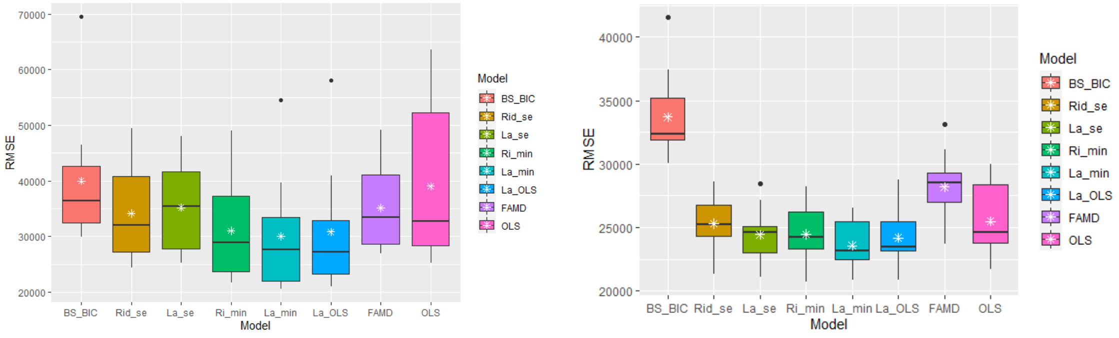


Figure 5.7 Boxplot and average values over 10 folds, of the previously explained modes with and without outliers.



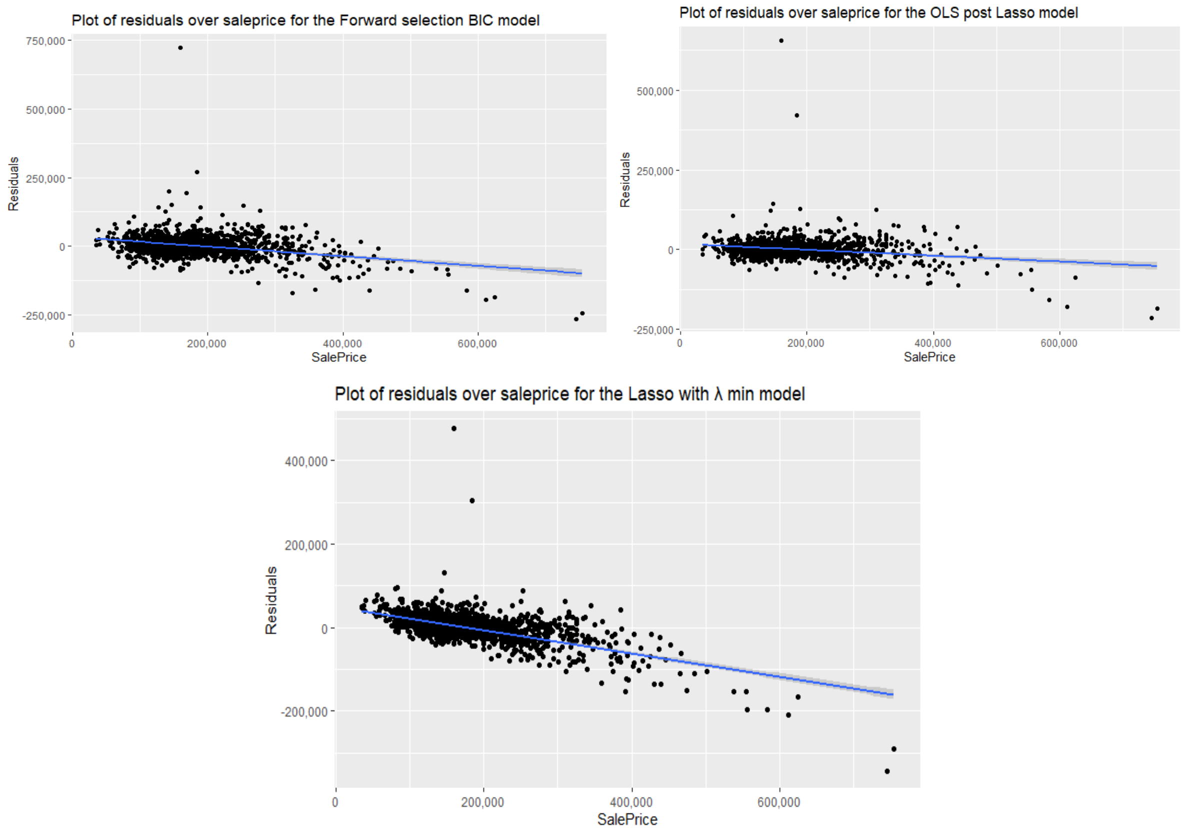
It can also be seen in Figure 5.6, how the models are consistent in the selection of the number of variables, there isn't much variability over the number of variables picked by a model over the 10 folds, and that there is no strong correlation between the number of variables and model performance. But we can also see that our models select a very wide amount of variables between themselves, the Lasso model with λ Standard error selects only about 50 variables, while the Ridge model basically keeps all 290 of them. But all the rest of the models fall into a specific range of features selected between 100, and 150 features.

5.8 Ouliers and residuals analysis

Continuing the analysis of the previous paragraph, we analysed the residuals of the models whose performance on one of the folds was particularly sub-optimal (Lasso min λ, OLS post Lasso, and Forward selection BIC models), from all three graphs it is possible to identify that there are some observations in particular that considerably increase the value of the RMSE in some folds. These observations are 2 houses belonging to the medium price range that are very overpriced, and 2 houses with very high prices that are considerably under-priced.

After removing these 4 observations there is a clear improvement on the variability and predictive capability of all models, whit the majority of them now having a RMSE lower than 25.000.

Figure 5.7 Plot of residuals over sale price for the Forward BIC, OLS post Lasso and min λ Lasso.


Plot of residuals over saleprice for the Forward selection BIC model

Plot of residuals over saleprice for the OLS post Lasso model

Plot of residuals over saleprice for the Lasso with λ min model

Conclusions

Through our work, we have faced numerous challenges, both of a theoretical nature and practical/implementation nature.

The first big challenge was the pre-processing work that we carried out in order to develop methodological strategies.

The unsupervised analyses has shown us to observe how different strategies allow to obtain different results to the different theoretical imprint.

FAMD allowed us to explore further static techniques in the context of component analysis.

There are numerous ways we can think of to improve our research; we should compare the results obtained by our regressions to some nonparametric models like Decision tree and random forest. We could optimize our missing value imputation method by tuning the parameters of our random forest method and implementing the cluster algorithms on the FAMD dataset.

References

Agès, J. (2004). “Analyse Factorielle de Donnees Mixtes.” *Revue Statistique Appliquee* 4: 93–111.

Daniel J. Stekhoven, (2022), *Nonparametric Missing Value Imputation using Random Forest, (https://www.r-project.org).*

Dean De Cock (2011), *Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project*, (Journal of Statistics Education, Volume 19, Number 3).

James G., Witten D., Hastie T. and R. Tibshirani (2013), *An Introduction to Statistical Learning with Applications in R, (*New York: Springer Science+Business Media).