SLLD FINAL PRESENTATION

Iowa house price regression

Mattia Pasqualini

Bernardo D’Agostino

The dataset

79 variables, 46 of which are categorical.

They describe a lot of aspects related to 1460 houses sold in the city of Ames Iowa in the United States from 2006 to 2010.

The dependent variable is SalePrice 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.

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Our aim for this project

Test different regression and feature selection methods for price prediction.

Learn how to apply unsupervised learing algorithms to categorical variables.

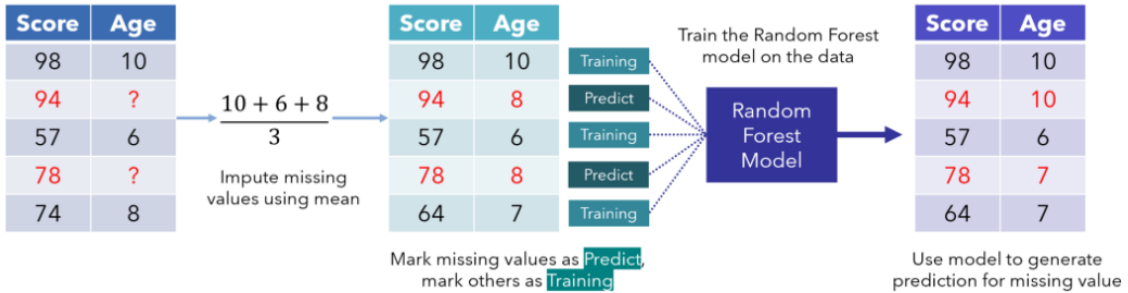
Lear possible ways to manage missing values.

Missing Values

The dataset had 6965 missing values.

We eliminated 5 columns which had more than 15% missing values.

We then used a Random Forest imputation method (missForest Package) for the remaining 609 missing values.



Pre-Processing

First, we one-hot encoded all 46 categorical variables getting a dataset with a total of 295 predictor variables.

Then we normalized all our continuous features.

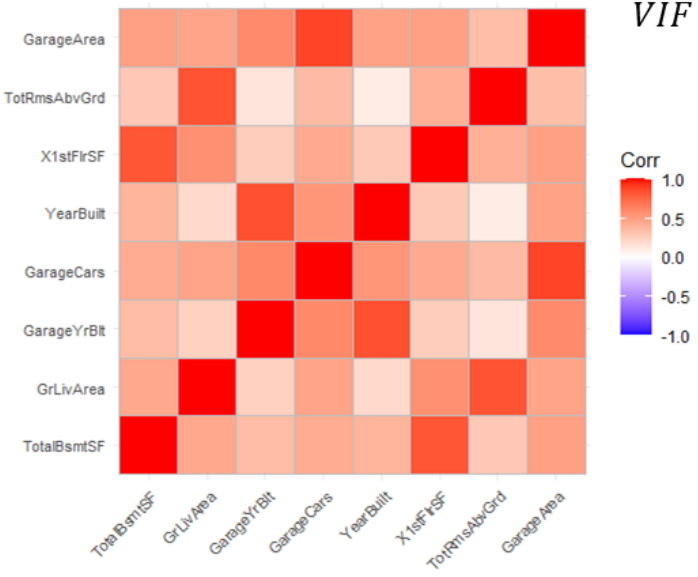
We then checked for collinearity and eliminated 5 variables.

Multicollinearity in the dataset

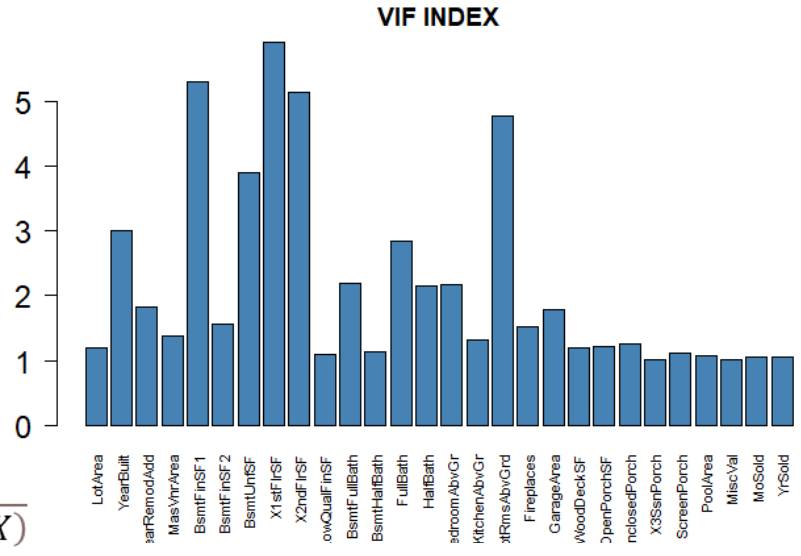
4 correlations higher than 0.8:

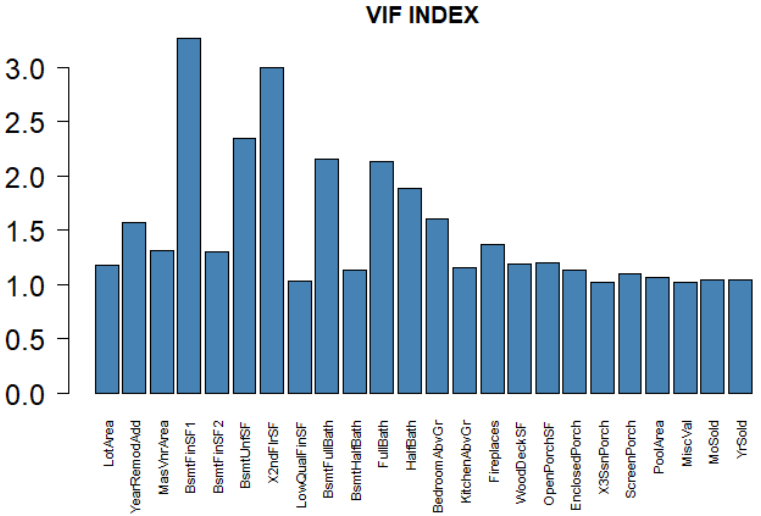
- GarageYrBlt and YearBuilt
- TotalBsmtSF and X1stFlrSF
- TotRmsAbvGrd and GrLivArea
- GarageCars and GarageArea

X1stFlrSF Had the highest VIF index



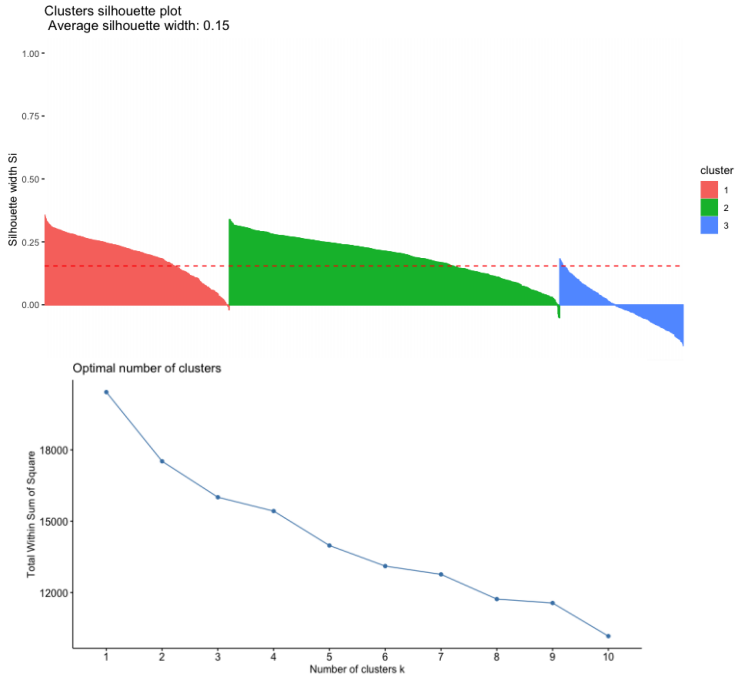
$$VIF = \frac{1}{1 - R^2(X)}$$





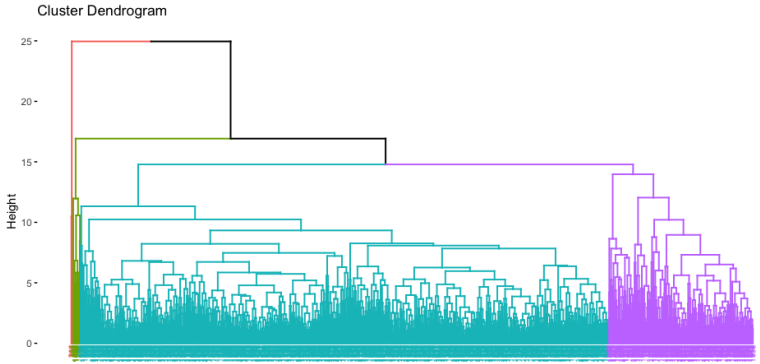
Cluster Analysis hierarchical

- We performed *Silhouette widths* method to pre-approach the cluster Analysis

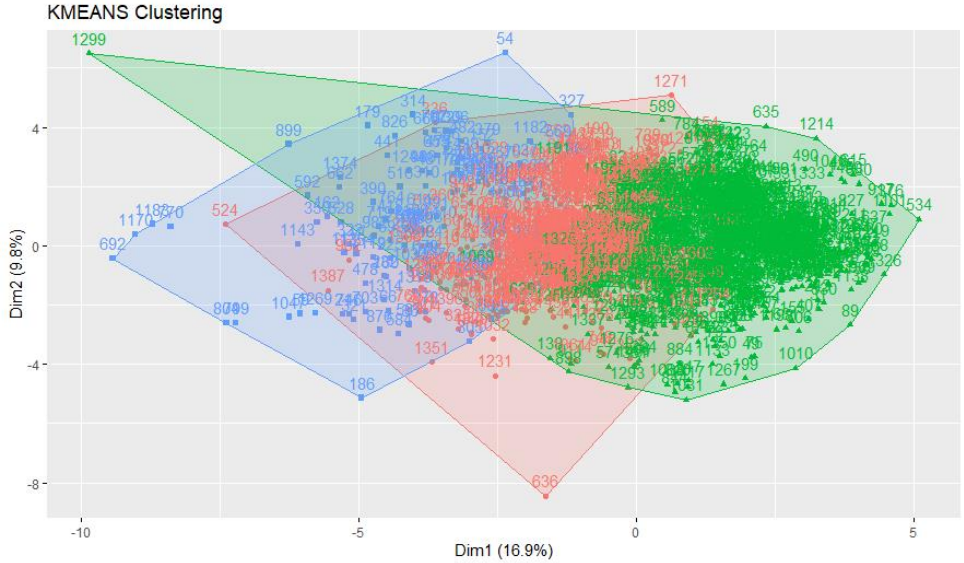


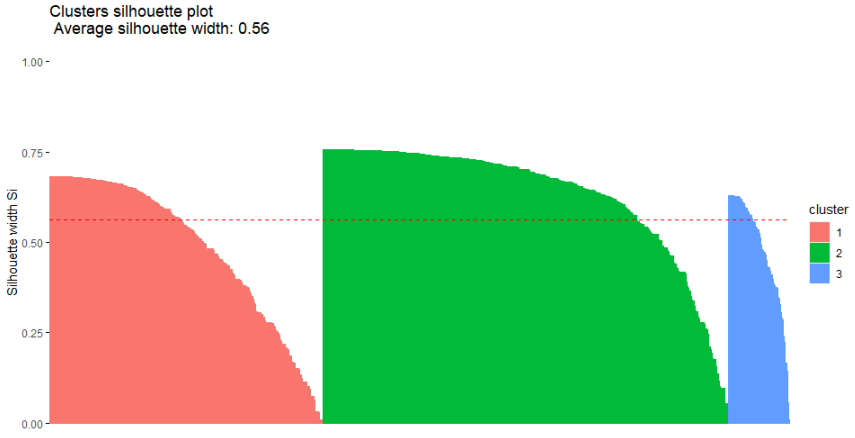
- Fundamental aspect given the complexity of the Dataset and its components

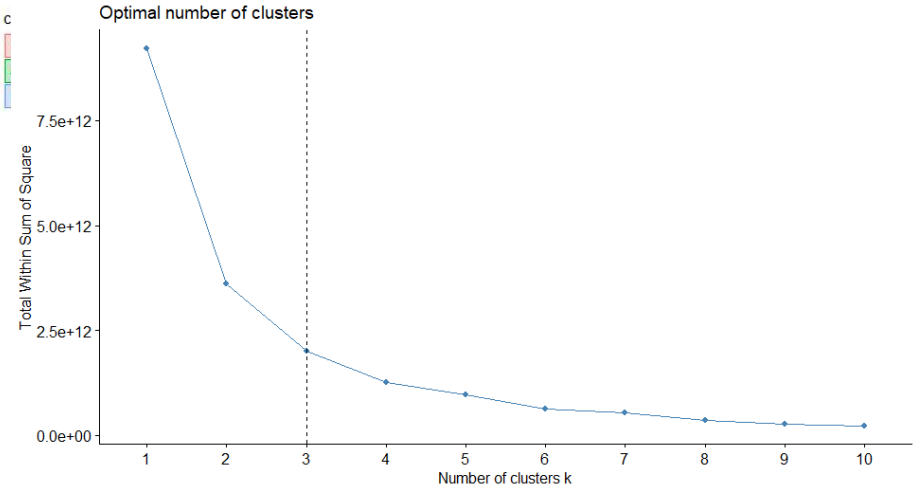
- Aimed at understanding the analysis approach to be used in the actual CA



Cluster Analysis K-means





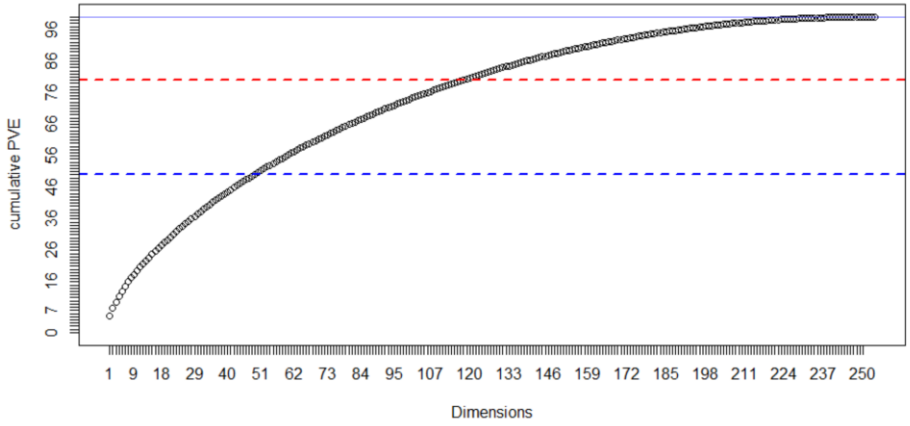


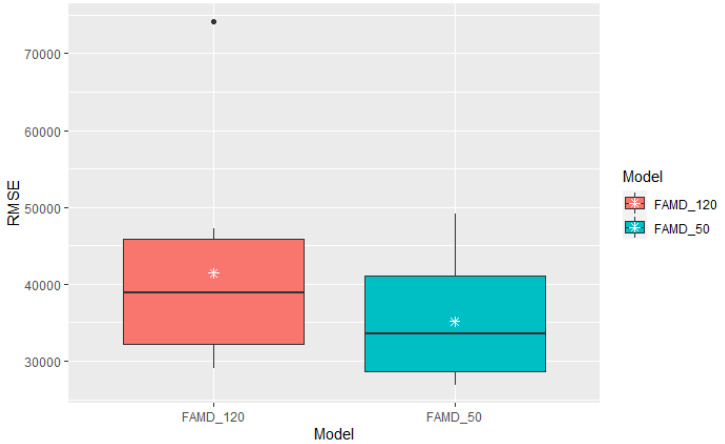
Factor Analysis of Mixed Data (FAMD)

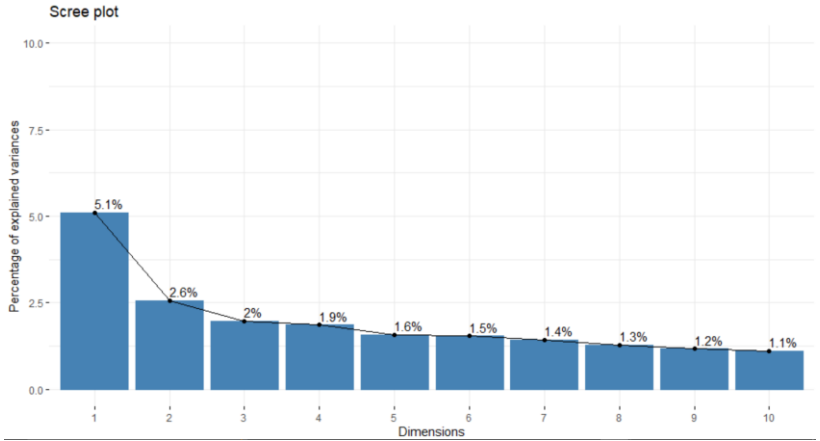
- It’s the Factorial method devoted to data tables in which a group of individuals is described both by quantitative and qualitative variables.
- FAMD works as a principal components analysis (PCA) for quantitative variables and as a multiple correspondence analysis (MCA) for qualitative variables
- It requires further work of processing the dataset or because the method does not work with some particular types of data

Number of principal components chosen

The first 50 components explain 50% of the variability.
The first 120 components explain 80% of the variability.
We cross validated the performance of an OLS model with both amounts of features and the first 50 components had a better RMSE, 35132 against 41450.

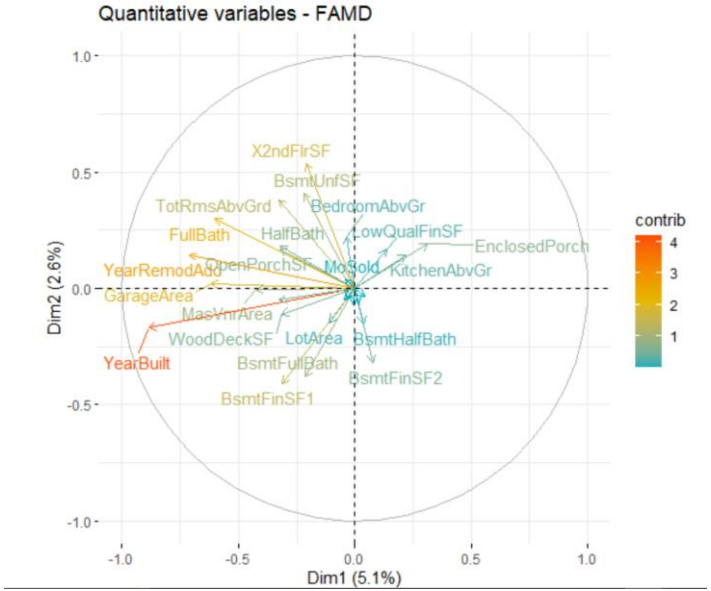


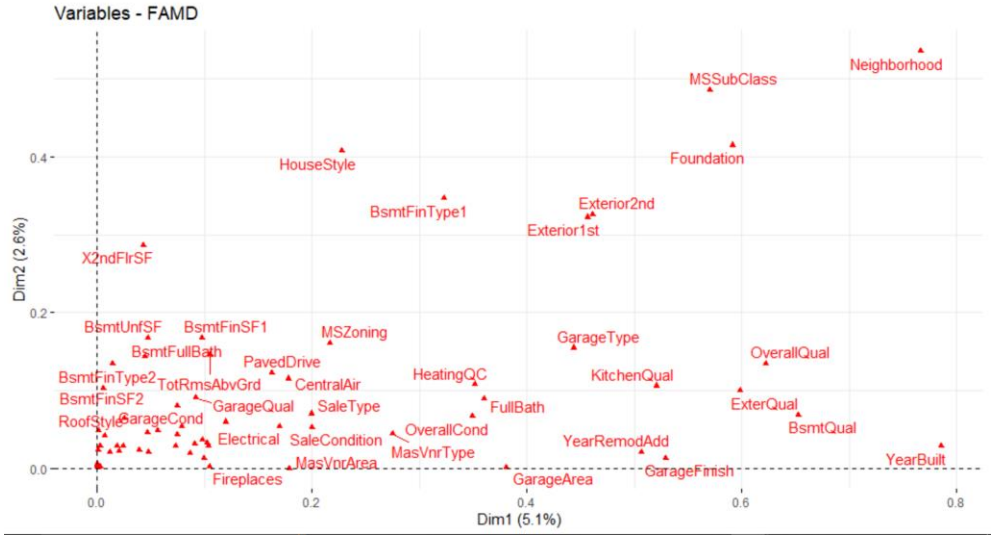


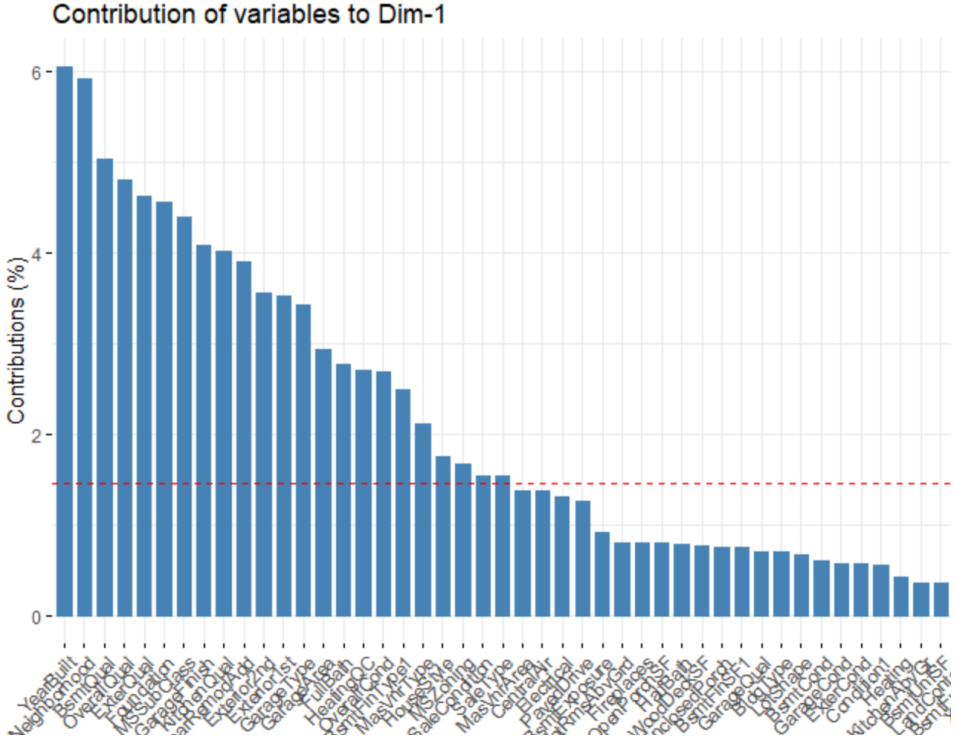


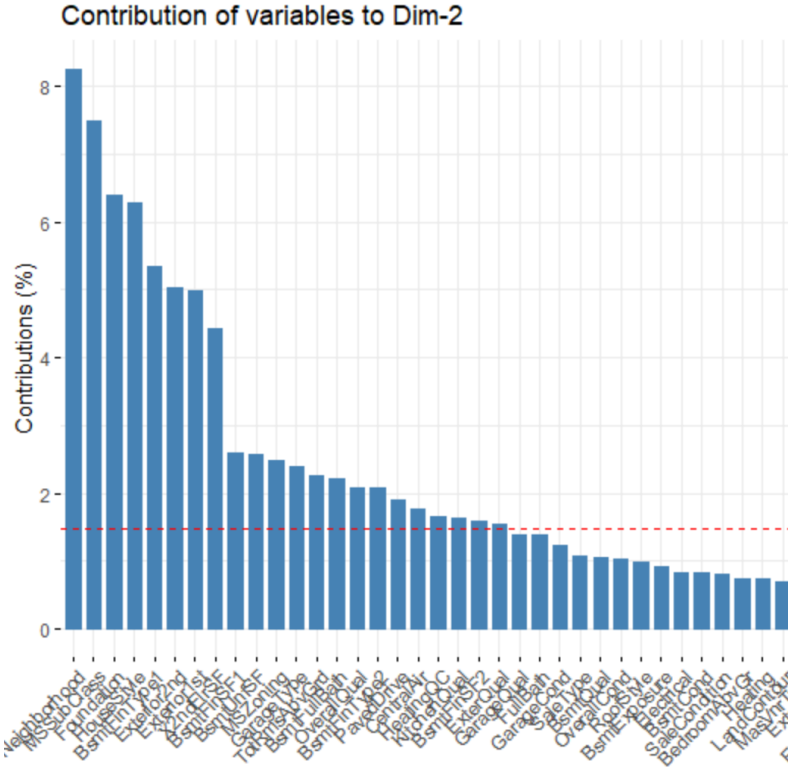
Variable contributions to 2 main principal components

- Analysis of the quantitative variables within the dataset
  Contribution to the main components






Contribution of variables to Dim-1
Contributions (%)
0
2
4
6

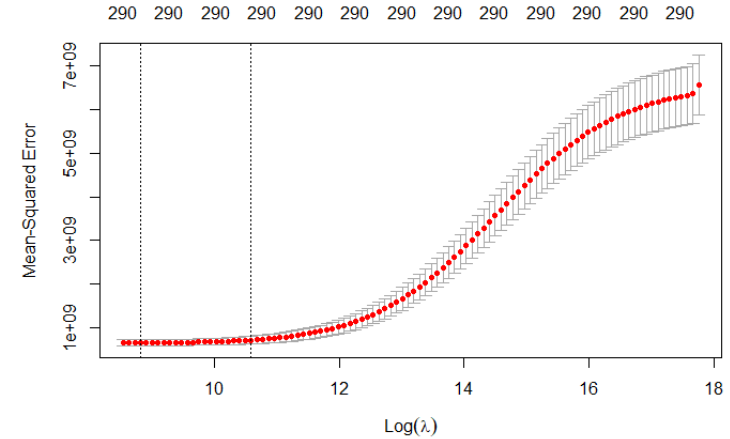

Contribution of variables to Dim-2
Contributions (%)
0
2
4
6
8

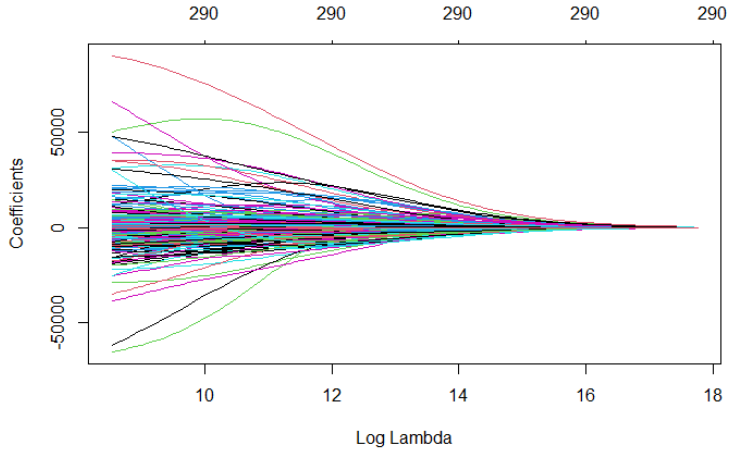
Supervised analysis

We compared the 10-fold cross validated performance of different models and feature selection algorithms.

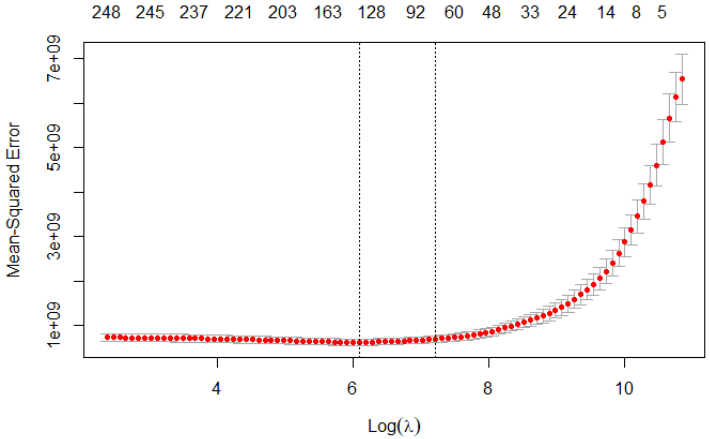
- OLS
- Lasso
- Ridge
- OLS on FAMD
- Best-Subset
  - We applied a forward best subset due to the high number of variables
- Lasso OLS
  - We fitted an OLS model on the features selected by the Lasso model with **λ_min**

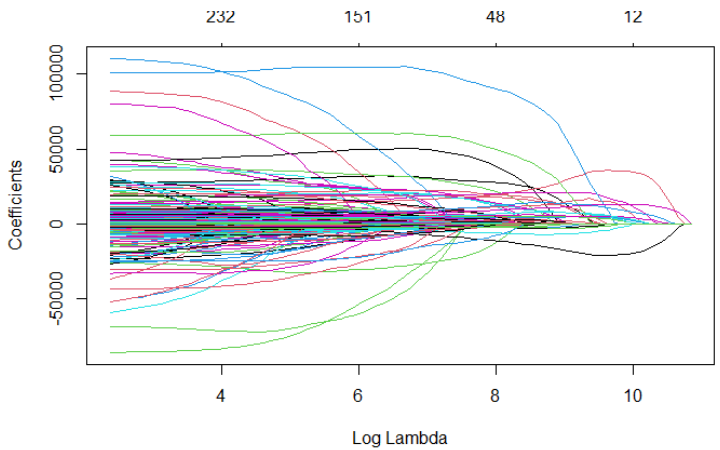
Ridge


290 290 290 290 290 290 290 290 290 290 290 290 290
Mean-Squared Error
1e+09
3e+09
5e+09
7e+09
10
12
14
16
18
Log(λ)

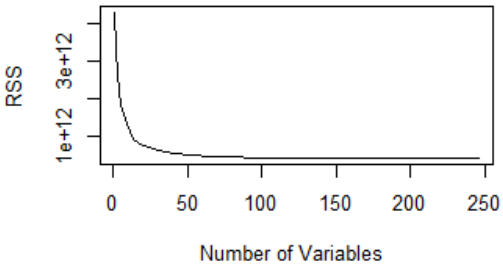

290 290 290 290 290
Coefficients
-50000
0
50000
10
12
14
16
18
Log Lambda

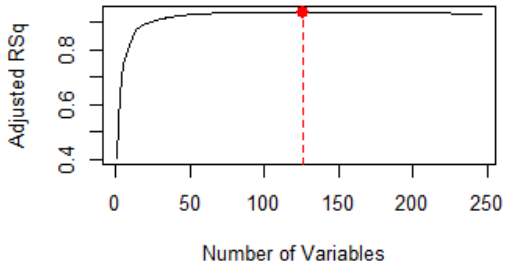
Lasso

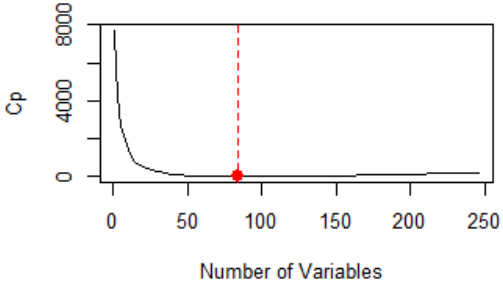

248 245 237 221 203 163 128 92 60 48 33 24 14 8 5
Mean-Squared Error
1e+09
3e+09
5e+09
7e+09
4
6
8
10
Log(λ)

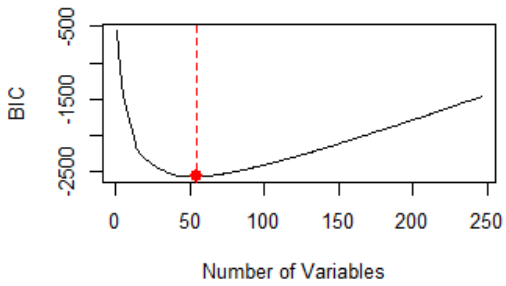

232 151 48 12
Coefficients
-50000
0
50000
100000
4
6
8
10
Log Lambda

Best Subset



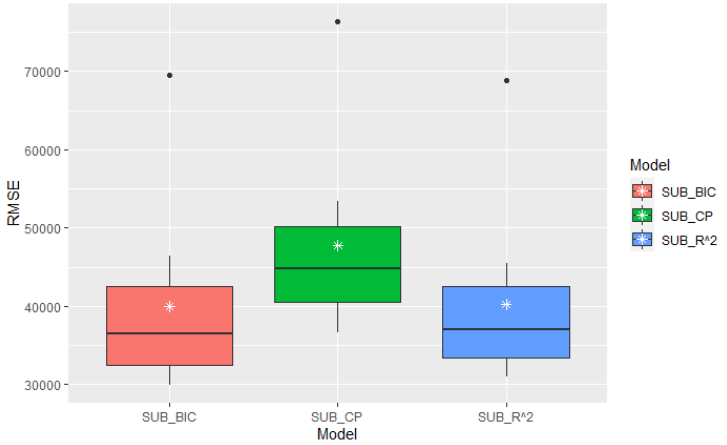




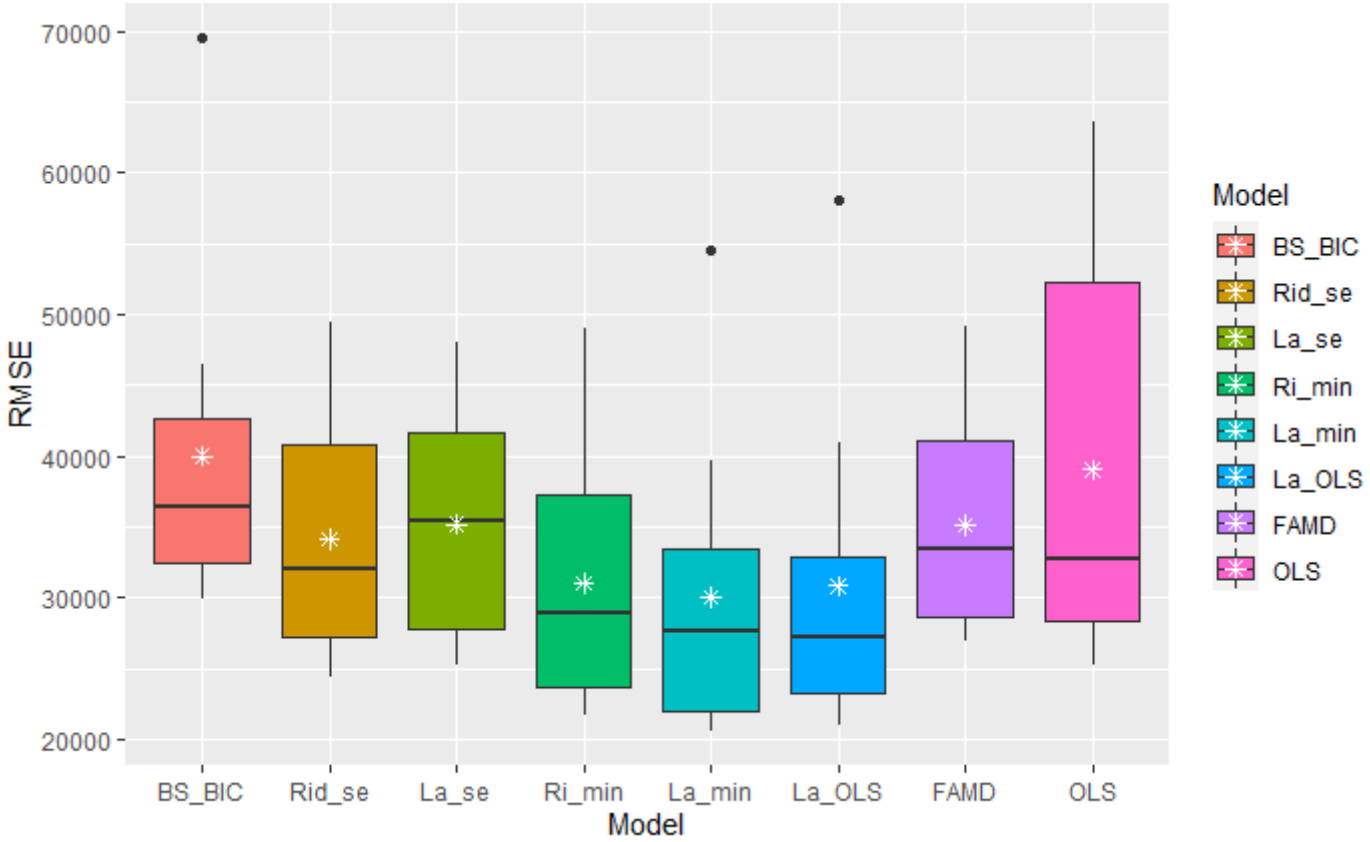


The 3 solutions that we found where unanimous, so we operated a 10-fold cross validation on all 3 subsets.

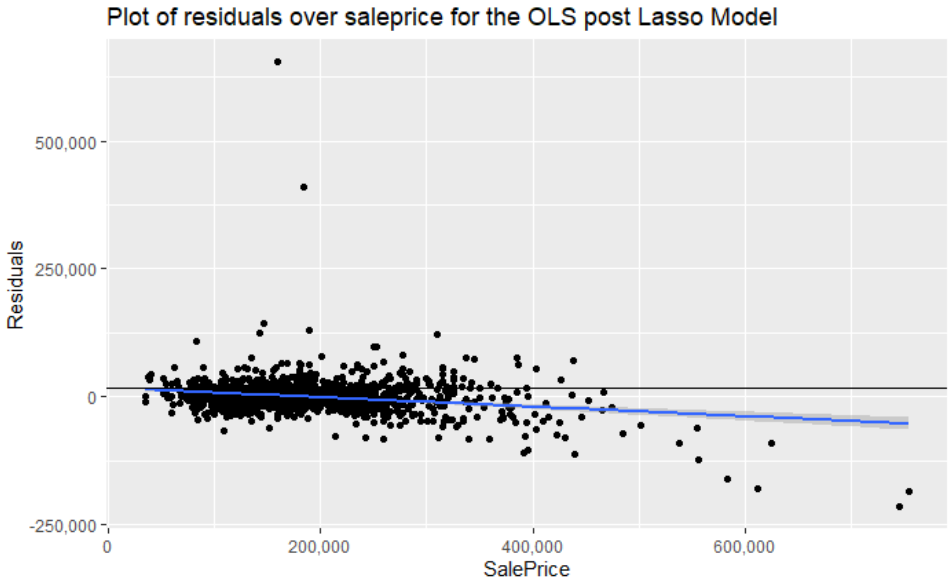
And we found the BIC to be the one with the least cross validated RMSE, but by a very small margin. Which was also the model with the least features selected.

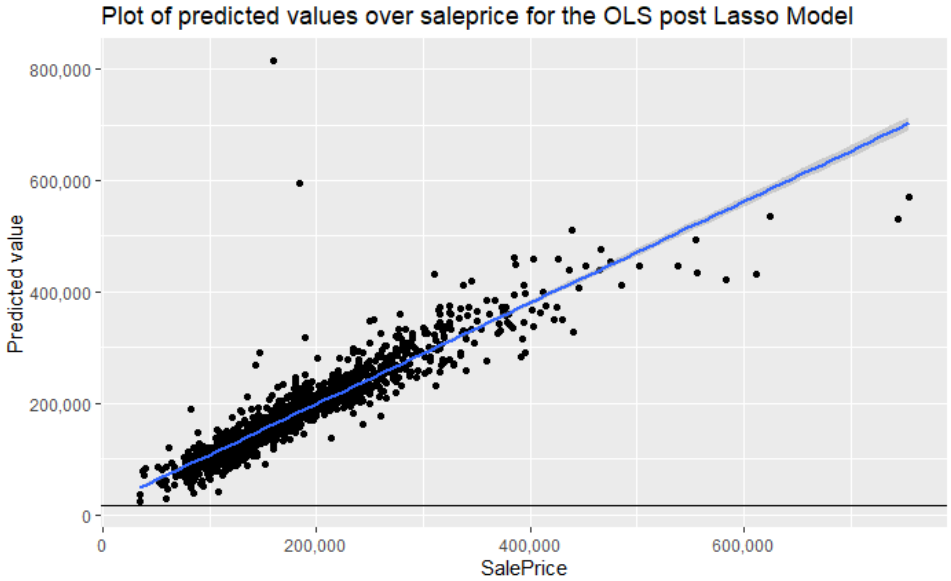


Comparison of Cross validated models



Analysis of predictions for OLS post Lasso





Possible developments

- Implement a grid search optimisation on Random Forest imputation of missing values.
- Test other prediction methods on the dataset like random forest or decision tree.
- Do an analysis on wich features are the most important over the different models.
- Find and manage outliers in the dataset.
- Clusterization with FAMD model.

Citations

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