Prédire le prix des maisons présentant plusieurs caractéristiques

Une analyse par les méthodes de régression sur variables latentes

Axel-Cleris Gailloty

Table of Contents

# Introduction générale

L’estimation de la valeur d’un bien immobilier est une étape cruciale dans la vente du bien. L’économie de la décision a mis en exergue une opposition entre le comportement des acheteurs et celui des vendeurs de biens immobiliers. Dans un premier temps, les vendeurs de biens immobiliers ont tendance à surévaluer leurs propriétés pour des raisons personnelles (il peut s’agir d’une propriété familiale à laquelle le vendeur attache une valeur sentimentale) tandis que l’attitude des acheteurs est toute autre : les acheteurs ont tendance à fixer un prix de réserve bien souvent inférieur à la valeur à laquelle les propriétaires vendent le bien immobilier. Cette situation peut conduire à une inefficience du marché des biens immobiliers.

Ce problème d’inefficience est principalement dû au différentiel d’estimation de la valeur que font les parties prenantes sur le marché. C’est pour cette raison que dans le cadre de cette étude, la question qui nous occupe est celle de comprendre les facteurs entrant dans la valorisation des biens immobiliers. Plus précisément il s’agit de se pencher sur les caractéristiques observables du bien qui entre positivement ou négativement dans la valeur du bien.

Pour comprendre donc ces facteurs cette étude se base sur une historique des des ventes de propriétés résidentielles effectuées dans la ville de Ames dans l’Etat de l’Iowa aux Etats-Unis sur la période de 2006 à 2010.

Mots clés : Econométrie, analyse exploratoire des données, analyses en composantes principales, régression

## Présentation des données

Le jeu de données sur lequel ce travail se base est décrit la vente de propriétés résidentielles individuelles à Ames, une ville de l’Etat de l’Iowa aux USA entre 2006 et 2010. L’ensemble de données contient 2930 observations et 80 variables explicatives (23 nominales, 23 ordinales, 14 discrètes et 20 continues) impliquées dans l’évaluation de la valeur des propriétés résidentielles.

## Objectifs de l’étude

Bien que la motivation première de cette étude est celle de déterminer les facteurs entrant dans l’évaluation d’un bien immobilier, cette étude porte un grand intérêt à l’utilisation des méthodes économétriques d’estimation. Pour cette étude nous adoptons la démarche économétrique qui consiste à mettre en évidence les relations de causalité entre le prix auquel le bien immobilier est vendu et les caractéristiques de ce bien. Il ne s’agit pas seulement d’une analyse des corrélations, mais de déterminer aussi les causalités en hiérarchisant les effets des caractéristiques observables d’un bien immobilier sur son prix de vente et par cela identifier également les potentiels effets non observables.

## Plan de l’étude

Nous commencerons par faire des statistiques descriptives univariées des données. Dans cette partie nous cherchons à comprendre la distribution des variables que nous avons sélectionnées. Puis nous nous intéresserons aux statistiques bivariées en étudiant notamment les corrélations entre chaque paire de variables.

Après avoir fait les statistiques descriptives et bivariées nous chercherons à modéliser le prix de vente des biens immobiliers en passant en revue les différentes méthodes d’estimations économétriques. Nous commencerons par utiliser la méthode des moindre carrés ordinaires. Nous ferons les différents tests et déterminer les éventuelles limites de cette méthode. Ensuite nous allons utiliser la méthode de régression sur composantes principales puis la méthode des moindres carrés partiels pour adresser les limites des moindres carrés ordinaires.

# Analyse exploratoire des données

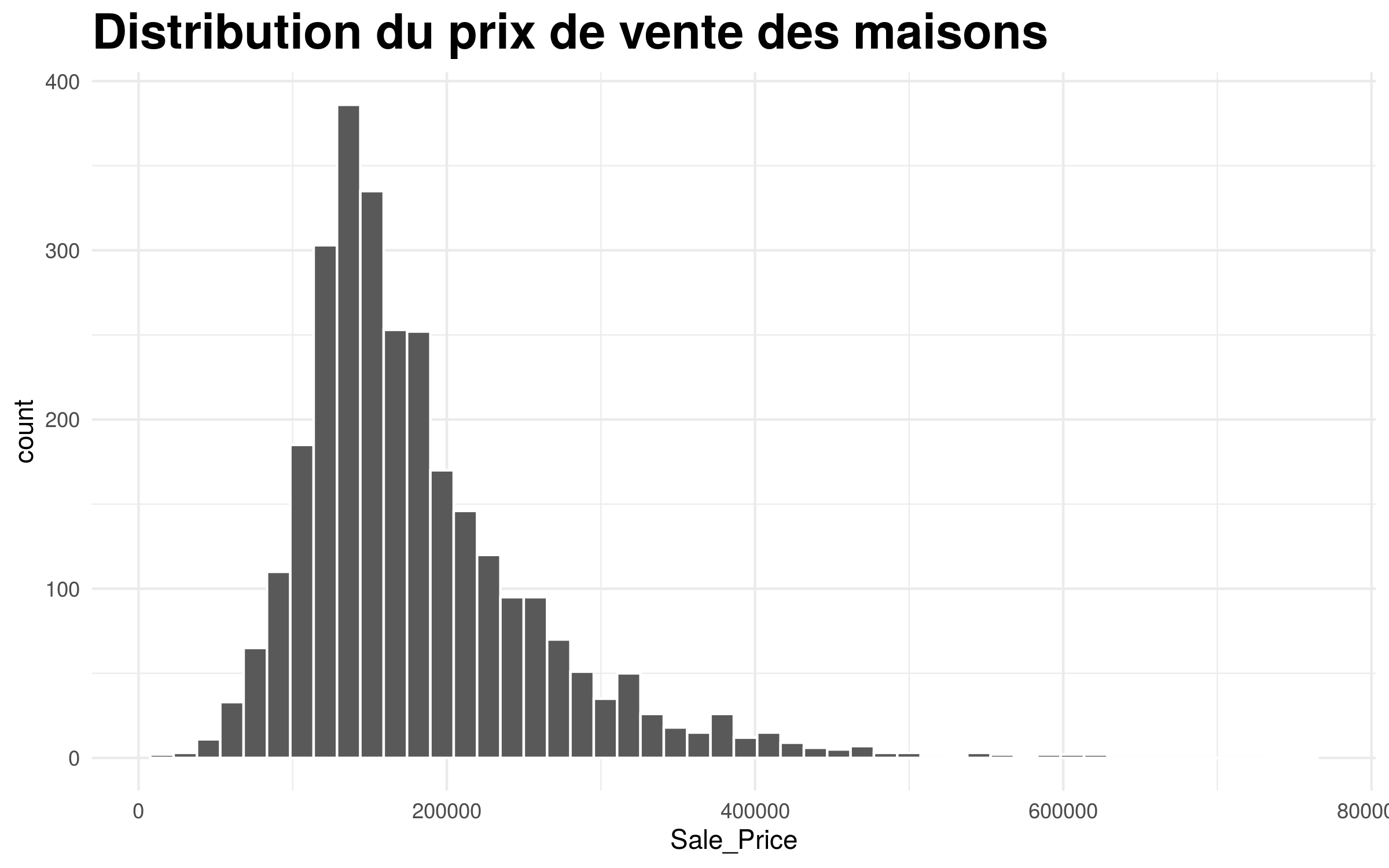
## Statistiques descriptives univariées

Nous allons afficher les statistiques descriptives des variables quantitatives continues présentes dans le jeu de données. Ces statistiques descriptives seront présentées sous formes d’un tableau à 5 colonnes qui affichent respectivement le minimum, la moyenne, l’écart-type, le maximum et la médiane.

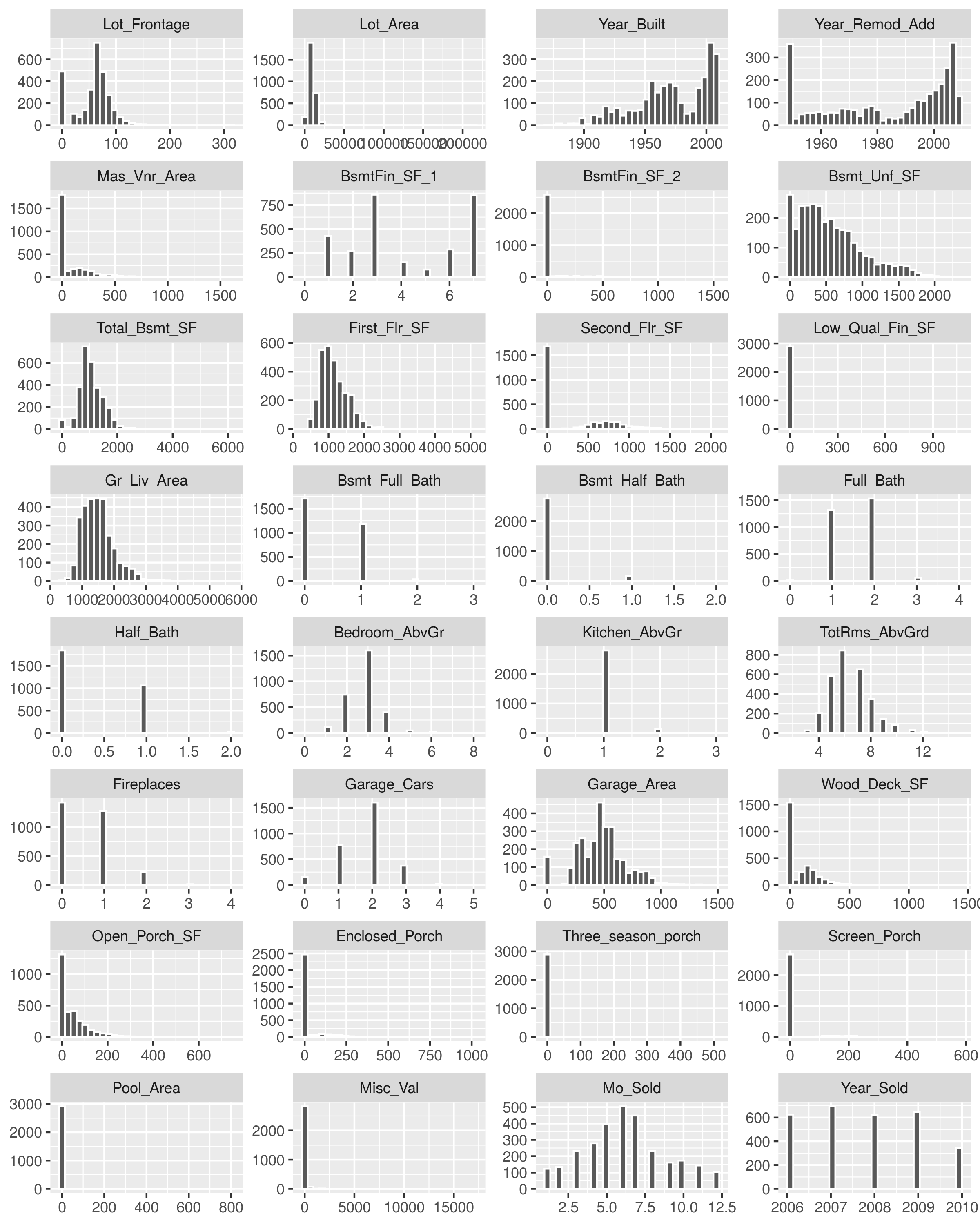
Tableau 1 : Statistiques descriptives sur les colonnes numériques

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Min | Moyenne | Ecart\_type | Max | Médiane |
| Lot\_Frontage | 0 | 57.64778157 | 33.49944079 | 313 | 63.0 |
| Lot\_Area | 1300 | 10147.92184300 | 7880.01775944 | 215245 | 9436.5 |
| Year\_Built | 1872 | 1971.35631399 | 30.24536063 | 2010 | 1973.0 |
| Year\_Remod\_Add | 1950 | 1984.26655290 | 20.86028588 | 2010 | 1993.0 |
| Mas\_Vnr\_Area | 0 | 101.09692833 | 178.63454483 | 1600 | 0.0 |
| BsmtFin\_SF\_1 | 0 | 4.17747440 | 2.23337248 | 7 | 3.0 |
| BsmtFin\_SF\_2 | 0 | 49.70546075 | 169.14208929 | 1526 | 0.0 |
| Bsmt\_Unf\_SF | 0 | 559.07167235 | 439.54057106 | 2336 | 465.5 |
| Total\_Bsmt\_SF | 0 | 1051.25563140 | 440.96801766 | 6110 | 990.0 |
| First\_Flr\_SF | 334 | 1159.55767918 | 391.89088525 | 5095 | 1084.0 |
| Second\_Flr\_SF | 0 | 335.45597270 | 428.39571501 | 2065 | 0.0 |
| Low\_Qual\_Fin\_SF | 0 | 4.67679181 | 46.31051003 | 1064 | 0.0 |
| Gr\_Liv\_Area | 334 | 1499.69044369 | 505.50888747 | 5642 | 1442.0 |
| Bsmt\_Full\_Bath | 0 | 0.43105802 | 0.52476196 | 3 | 0.0 |
| Bsmt\_Half\_Bath | 0 | 0.06109215 | 0.24517502 | 2 | 0.0 |
| Full\_Bath | 0 | 1.56655290 | 0.55294061 | 4 | 2.0 |
| Half\_Bath | 0 | 0.37952218 | 0.50262925 | 2 | 0.0 |
| Bedroom\_AbvGr | 0 | 2.85426621 | 0.82773114 | 8 | 3.0 |
| Kitchen\_AbvGr | 0 | 1.04436860 | 0.21407624 | 3 | 1.0 |
| TotRms\_AbvGrd | 2 | 6.44300341 | 1.57296440 | 15 | 6.0 |
| Fireplaces | 0 | 0.59931741 | 0.64792092 | 4 | 1.0 |
| Garage\_Cars | 0 | 1.76621160 | 0.76113672 | 5 | 2.0 |
| Garage\_Area | 0 | 472.65836177 | 215.18719571 | 1488 | 480.0 |
| Wood\_Deck\_SF | 0 | 93.75187713 | 126.36156188 | 1424 | 0.0 |
| Open\_Porch\_SF | 0 | 47.53344710 | 67.48340014 | 742 | 27.0 |
| Enclosed\_Porch | 0 | 23.01160410 | 64.13905921 | 1012 | 0.0 |
| Three\_season\_porch | 0 | 2.59249147 | 25.14133103 | 508 | 0.0 |
| Screen\_Porch | 0 | 16.00204778 | 56.08737023 | 576 | 0.0 |
| Pool\_Area | 0 | 2.24334471 | 35.59718062 | 800 | 0.0 |
| Misc\_Val | 0 | 50.63515358 | 566.34428826 | 17000 | 0.0 |
| Mo\_Sold | 1 | 6.21604096 | 2.71449243 | 12 | 6.0 |
| Year\_Sold | 2006 | 2007.79044369 | 1.31661292 | 2010 | 2008.0 |
| Sale\_Price | 12789 | 180796.06006826 | 79886.69235666 | 755000 | 160000.0 |

En moyenne les maisons vendues coûtent un peu plus de 180 mille dollars. Les prix varient fortement d’une maison à l’autre puisque le prix minimum de vente est de 12789 dollars tandis que le prix maximum est de plus de 755 mille dollars. Pour avoir une meilleure vision de la distribution des prix, nous pouvons représenter un histogramme des prix.



Nous pouvons également avoir un aperçu de l’ensemble des colonnes numériques du jeu de données en faisant une représentation graphique combinée de ces colonnes.

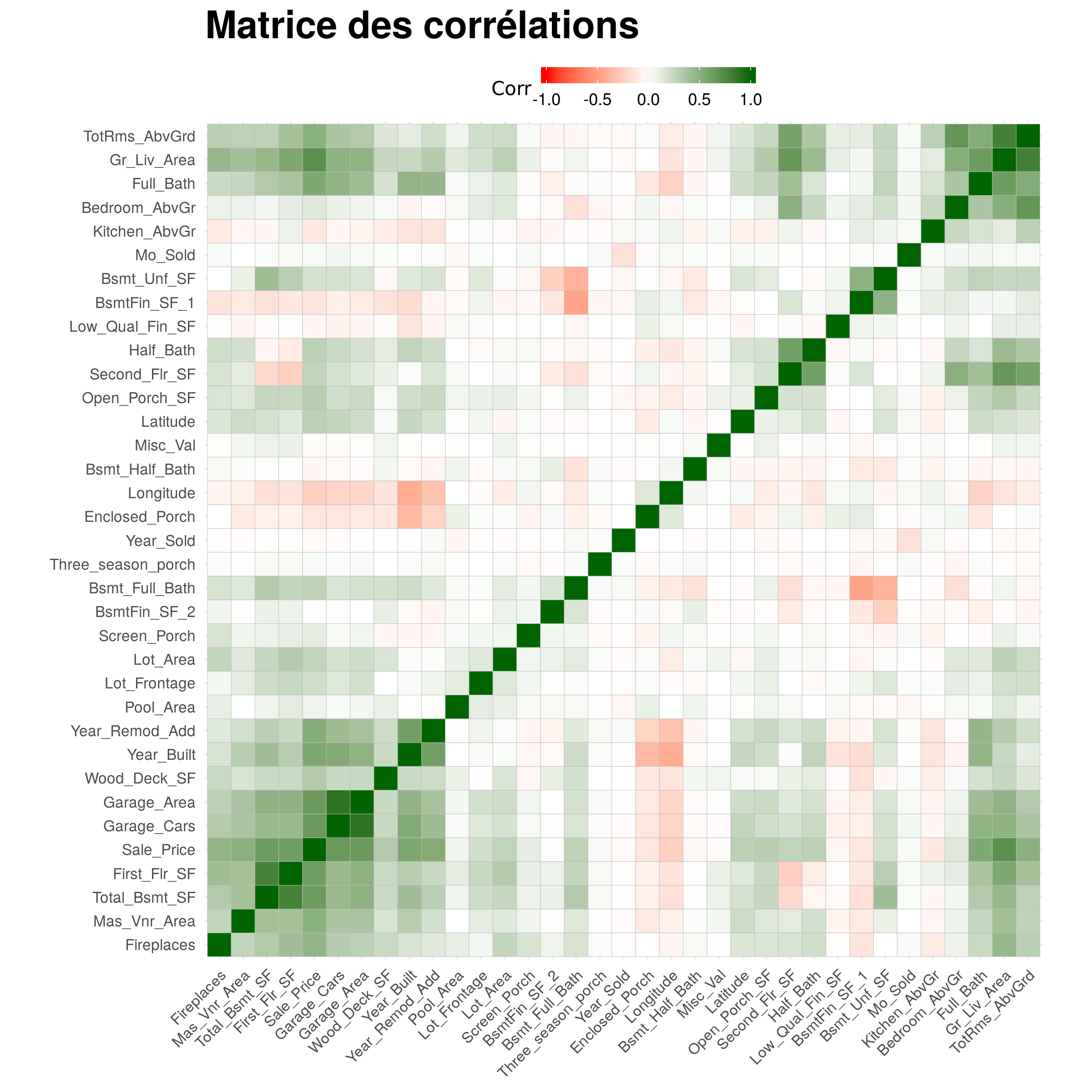


Il est important que nous explorions les données pour découvrir la structure des données. Le but de cette analyse exploratoire est d’étudier la variance associée au jeu de données. En effet une base de donnée comporte en elle une variance ou information et le but de tout travail économétrique consistant à modéliser la variable dépendante est de trouver un modèle qui arrive à expliquer la variance de cette variable dépendante en se basant sur des variables qu’on appelle variable explicatives.

## Statistiques descriptives bivariées

Les statistiques bivariées nous permettent d’observer le degré de liaisons entre les différentes variables. Nous allons nous intéresser au coefficient de corrélation pour chaque paire de variables afin de voir le degré de liaisons pour les variables prises deux à deux.

Pour simplifier la lecture de ces paires de corrélation, nous allons représenter la matrice des paires de corrélation sous la forme d’un graphique. Cette façon nous permet de voir à la fois dans la globalité et pour chaque variable le degré de corrélation.



Nous voyons que sur un plan global les corrélations entre les variables sont davantage positives que négatives. Etant donné le grand nombre des variables que nous étudions, nous n’allons commenter que les coefficients de corrélation les plus fortes. Pour cela nous affichons dans les tableaux suivants les respectivement les plus fortes corrélations positives et les plus fortes corrélations négatives.

Les 10 premières fortes corrélations entre les variables

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | cor.coef |
| 2 | Garage\_Cars | Garage\_Area | 0.88986599 |
| 5 | First\_Flr\_SF | Total\_Bsmt\_SF | 0.80042869 |
| 8 | Gr\_Liv\_Area | Sale\_Price | 0.70677992 |
| 11 | Gr\_Liv\_Area | Second\_Flr\_SF | 0.65525118 |
| 14 | Garage\_Cars | Sale\_Price | 0.64756161 |
| 17 | Sale\_Price | Total\_Bsmt\_SF | 0.63252885 |
| 20 | Gr\_Liv\_Area | Full\_Bath | 0.63032081 |
| 23 | Year\_Remod\_Add | Year\_Built | 0.61209525 |
| 26 | Second\_Flr\_SF | Half\_Bath | 0.61163367 |
| 29 | Gr\_Liv\_Area | First\_Flr\_SF | 0.56216584 |

La variable Garage\_Cars mesure la taille du garage en capacité de voiture et la variable Garage\_Area mesure la taille du garage en pieds carrés (mesure américaine). La corrélation entre ces deux variables est positive et très forte. La superficie totale en pieds carrés du sous-sol (Total\_Bsmt\_SF) est fortement corrélée avec la superficie du premier étage (First\_Flr\_SF). Nous voyons également que la corrélation entre la superficie totale (Gr\_Liv\_Area) de la maison et son prix (Sale\_Price) est également positive et forte. Nous voyons également qu’il y a une corrélation positive entre le prix de la maison et la superficie totale du sous-sol. La superficie totale corrèle également positivement et fortement avec les salles de bain complètes au-dessus du niveau du sol (Full\_Bath).

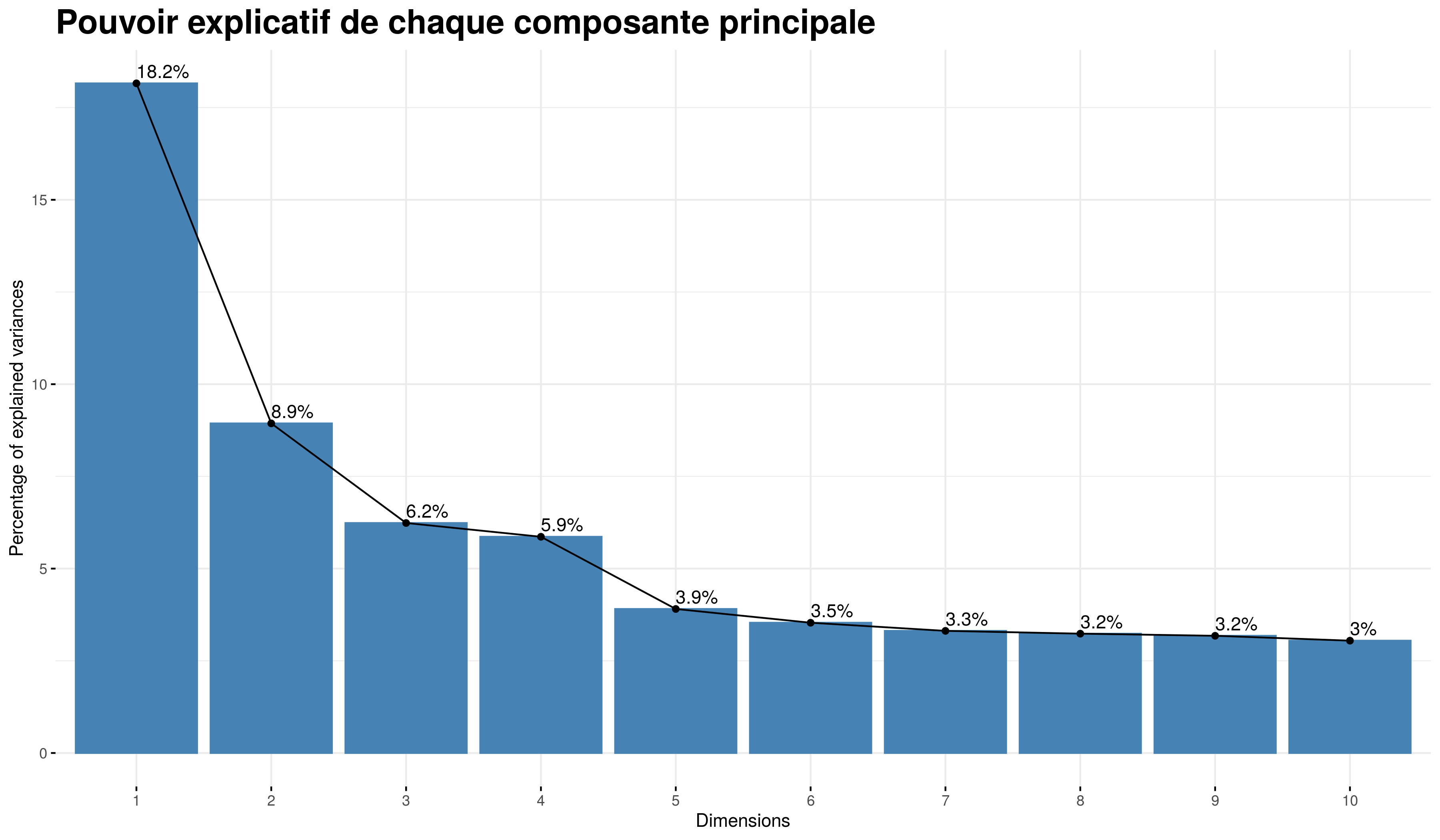
Les 10 premières fortes corrélations négatives entre les variables

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | cor.coef |
| 2 | BsmtFin\_SF\_1 | Bsmt\_Full\_Bath | -0.47342438 |
| 5 | Bsmt\_Full\_Bath | Bsmt\_Unf\_SF | -0.39862948 |
| 8 | Year\_Built | Enclosed\_Porch | -0.37436441 |
| 11 | Longitude | Sale\_Price | -0.25139725 |
| 14 | First\_Flr\_SF | Second\_Flr\_SF | -0.25005720 |
| 17 | Longitude | Full\_Bath | -0.22877349 |
| 20 | Year\_Remod\_Add | Enclosed\_Porch | -0.22038327 |
| 23 | Longitude | Garage\_Area | -0.20820322 |
| 26 | Total\_Bsmt\_SF | Second\_Flr\_SF | -0.20448703 |
| 29 | Total\_Bsmt\_SF | BsmtFin\_SF\_1 | -0.16331128 |

La lecture de ces corrélations nous montre certaines choses qu’on connaissait déjà à savoir par exemple que le prix d’une maison corrèle positivement avec sa superficie. Toutefois, étudier les paires de corrélations une à une peut être redondant du fait que plusieurs variables peuvent évoluer dans le même sens. Il serait donc plus intéressant de synthétiser les axes de variabilités qui existent entre les variables.  
Pour ce faire nous pouvons utiliser l’analyse en composante principale qui est une méthode factorielle permettant de résumer la matrice des variances-covariances.

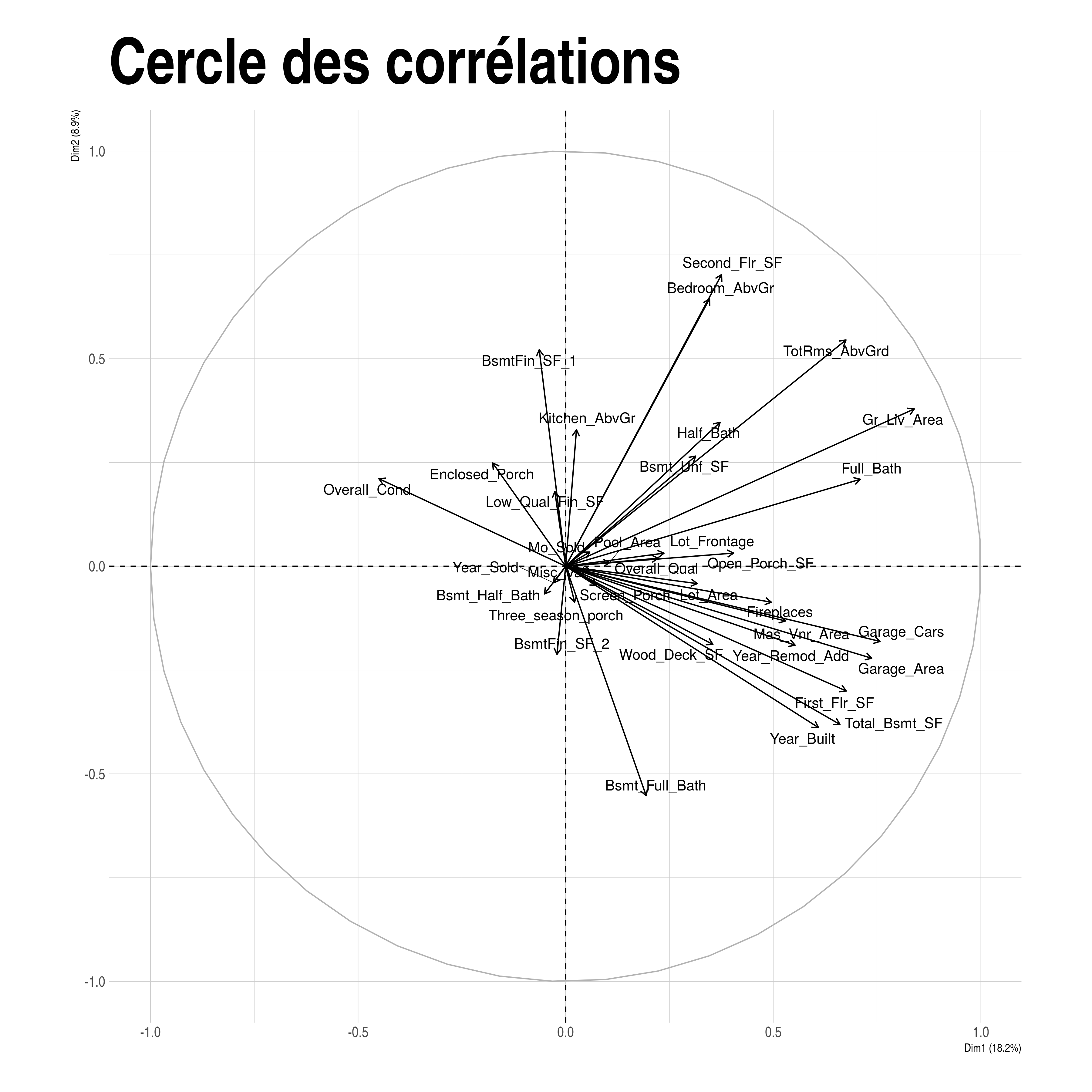
## L’analyse en composantes principales

Le cercle de corrélation que nous construisons ici est le résultat d’une analyse en composantes principales sur les données quantitatives. Par souci de simplicité nous ne représenterons que le premier plan factoriel.  
Ce premier plan factoriel comporte affiche la direction et la force de corrélation de chacune des variables sur les deux premiers axes factoriels. En ACP le pouvoir explicatif des axes est descendant, les deux premiers axes expliquent souvent une grande partie de la variance. L’intérêt de cette représentation est d’observer dans quel sens vont les corrélations et comment les variables peuvent être groupées sur les axes factoriels.



Le pouvoir explicatif marginal d’une composante est négative. Les deux premières composantes arrivent à expliquer 27,1% de la variance des caractéristiques des maisons.

Nous pouvons représenter sur un plan factoriel les deux premières composantes principales pour observer le sens des variances.



Le cercle des corrélations montre qu’une grande partie des variable est fortement corrélée avec l’axe 1. Ce qui signifie que ces variables évoluent dans le même sens.

Pour avoir une meilleure idée de ce que représente chacun des axes, nous pouvons lister les contributions respectives des variables à la définition des axes.

Les dix premières contributions à la première composante

|  |  |
| --- | --- |
| Variable | Dim.1 |
| Gr\_Liv\_Area | 11.4212735 |
| Garage\_Cars | 9.2884089 |
| Garage\_Area | 8.8030918 |
| Full\_Bath | 8.1580769 |
| First\_Flr\_SF | 7.3976080 |
| TotRms\_AbvGrd | 7.3824529 |
| Total\_Bsmt\_SF | 7.0686579 |
| Year\_Built | 6.0076902 |
| Year\_Remod\_Add | 4.9471814 |
| Mas\_Vnr\_Area | 4.5410862 |

Les variables qui ont les plus fortes contributions sur le premier axe factoriel sont principalement des variables qui touchent à la superficie de la maison ainsi que la superficie des pièces (Garages, salles de bain) que contient la maison. Cette composante désigne donc globalement la taille de la maison.

Les dix premières contributions à la deuxième composante

|  |  |
| --- | --- |
| Variable | Dim.2 |
| Second\_Flr\_SF | 16.2427384 |
| Bedroom\_AbvGr | 13.6543823 |
| Bsmt\_Full\_Bath | 10.0511123 |
| TotRms\_AbvGrd | 9.7774602 |
| BsmtFin\_SF\_1 | 8.9508482 |
| Year\_Built | 4.9721088 |
| Total\_Bsmt\_SF | 4.7761850 |
| Gr\_Liv\_Area | 4.7257344 |
| Half\_Bath | 3.9566234 |
| Kitchen\_AbvGr | 3.5545581 |

Sur la deuxième composante ce sont les caractéristiques relatives au nombre d’étages et de pièces situés dans les étages des maisons.

# Modélisations

## Transformations des données

Comme nous l’avions présenté dans la présentation des données, 46 des colonnes présentes dans le jeu de données ne sont pas quantitatives et continues. En effet 23 sont qualitatives ordinales et 23 autres sont qualitatives simples. Il est donc important que nous trouvions une stratégie pour inclure ces colonnes dans le modèle car ces informations qui ne sont pas nécessairement quantitatives sont importantes pour comprendre l’inertie de la variable que nous cherchons à modéliser qui est en l’occurrence le prix du bien immobilier (Sale\_Price).

Pour nous assurer de l’importance d’inclure les variables non quantitatives dans le modèle, nous allons construire deux modèles de régression multiples simples. Le premier modèle inclut uniquement les variables quantitatives continues et le second modèle va inclure en plus des variables quantitatives les variables qualitatives transformées.

## Modèle de régression multiple

Call:  
lm(formula = Sale\_Price ~ ., data = base\_export)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-611577 -18168 -1914 15215 291823   
  
Coefficients: (1 not defined because of singularities)  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 173248.379363 1034265.831414 0.1675 0.8669816   
Lot\_Frontage 87.803069 21.071267 4.1670 3.177e-05 \*\*\*  
Lot\_Area 0.276464 0.092621 2.9849 0.0028603 \*\*   
Year\_Built 366.052233 40.899707 8.9500 < 2.2e-16 \*\*\*  
Year\_Remod\_Add 520.804027 43.831432 11.8820 < 2.2e-16 \*\*\*  
Mas\_Vnr\_Area 39.274696 4.390740 8.9449 < 2.2e-16 \*\*\*  
BsmtFin\_SF\_1 120.551239 399.798684 0.3015 0.7630321   
BsmtFin\_SF\_2 -11.081489 4.231329 -2.6189 0.0088670 \*\*   
Bsmt\_Unf\_SF -12.311265 2.668471 -4.6136 4.129e-06 \*\*\*  
Total\_Bsmt\_SF 37.970421 3.310015 11.4714 < 2.2e-16 \*\*\*  
First\_Flr\_SF 61.330334 4.008845 15.2988 < 2.2e-16 \*\*\*  
Second\_Flr\_SF 60.450368 3.540552 17.0737 < 2.2e-16 \*\*\*  
Low\_Qual\_Fin\_SF 12.832757 14.791305 0.8676 0.3856919   
Gr\_Liv\_Area NA NA NA NA   
Bsmt\_Full\_Bath 6293.288284 1856.188624 3.3904 0.0007072 \*\*\*  
Bsmt\_Half\_Bath -1858.613118 2914.083872 -0.6378 0.5236520   
Full\_Bath 1920.740936 1998.496886 0.9611 0.3365858   
Half\_Bath -2438.265685 1935.228354 -1.2599 0.2077936   
Bedroom\_AbvGr -9685.349040 1216.270979 -7.9632 2.387e-15 \*\*\*  
Kitchen\_AbvGr -34847.281701 3621.088385 -9.6234 < 2.2e-16 \*\*\*  
TotRms\_AbvGrd 4500.668956 881.131470 5.1078 3.470e-07 \*\*\*  
Fireplaces 7560.968660 1265.264411 5.9758 2.569e-09 \*\*\*  
Garage\_Cars 8144.488352 2093.470742 3.8904 0.0001023 \*\*\*  
Garage\_Area 20.914087 7.215244 2.8986 0.0037765 \*\*   
Wood\_Deck\_SF 23.006374 5.795860 3.9694 7.379e-05 \*\*\*  
Open\_Porch\_SF -6.849731 10.854623 -0.6310 0.5280624   
Enclosed\_Porch 27.890298 11.454418 2.4349 0.0149561 \*   
Three\_season\_porch 10.081139 26.568812 0.3794 0.7043926   
Screen\_Porch 62.867980 12.296415 5.1127 3.382e-07 \*\*\*  
Pool\_Area -67.534786 19.396372 -3.4818 0.0005054 \*\*\*  
Misc\_Val -9.683926 1.196666 -8.0924 8.517e-16 \*\*\*  
Mo\_Sold 87.976438 249.179130 0.3531 0.7240654   
Year\_Sold -941.543008 514.104536 -1.8314 0.0671401 .   
Overall\_Qual 4832.147729 1133.167888 4.2643 2.070e-05 \*\*\*  
Overall\_Cond -2726.736579 1385.223170 -1.9684 0.0491120 \*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 35862 on 2896 degrees of freedom  
Multiple R-squared: 0.80075, Adjusted R-squared: 0.79848   
F-statistic: 352.68 on 33 and 2896 DF, p-value: < 2.22e-16

La significativité globale de ce modèle de régression multiple est très bonne comme l’indique le test de Fisher. Le modèle explique 80% de la variance des prix de vente, ce qui est très bien pour des données transversales. La grande majorité des variables que nous avions sélectionnées pour inclure dans le modèle se sont montrés significatifs. Le test de Student effectué sur chacun des coefficients montre les différents dégrés de significativités des coefficients.

A cause de multicolinéarité, la variable Gr\_Liv\_Area qui représente la surface habitable hors étage de la maison a été enlevée du modèle.

Les variables sont toutes en niveau, nous pouvons donc interpréter les coefficients dans l’unité de la variable. Par exemple l’augmentation de la superficie du terrain (Lot\_Area) d’un pied carré augmente le prix de la maison de 87.8 dollars, toute chose égale par ailleurs.

### Le test d’inflation de la variance (VIF)

Il n’est pas possible de réaliser le test d’inflation de la variance (*Variance Inflation Factor*) sur le modèle car le coefficient de la variable Gr\_Liv\_Area n’a pas été calculé à cause de la multicolinéarité. Pour pouvoir quantifier la multicolinéarité entre les variables, nous devons donc estimer un nouveau modèle dans lequel la variable Gr\_Liv\_Area est absente.

Voici donc les résultats du test VIF que trouvons lorsque la variable Gr\_Liv\_Area est omise.

Test d’inflation de la variance (VIF)

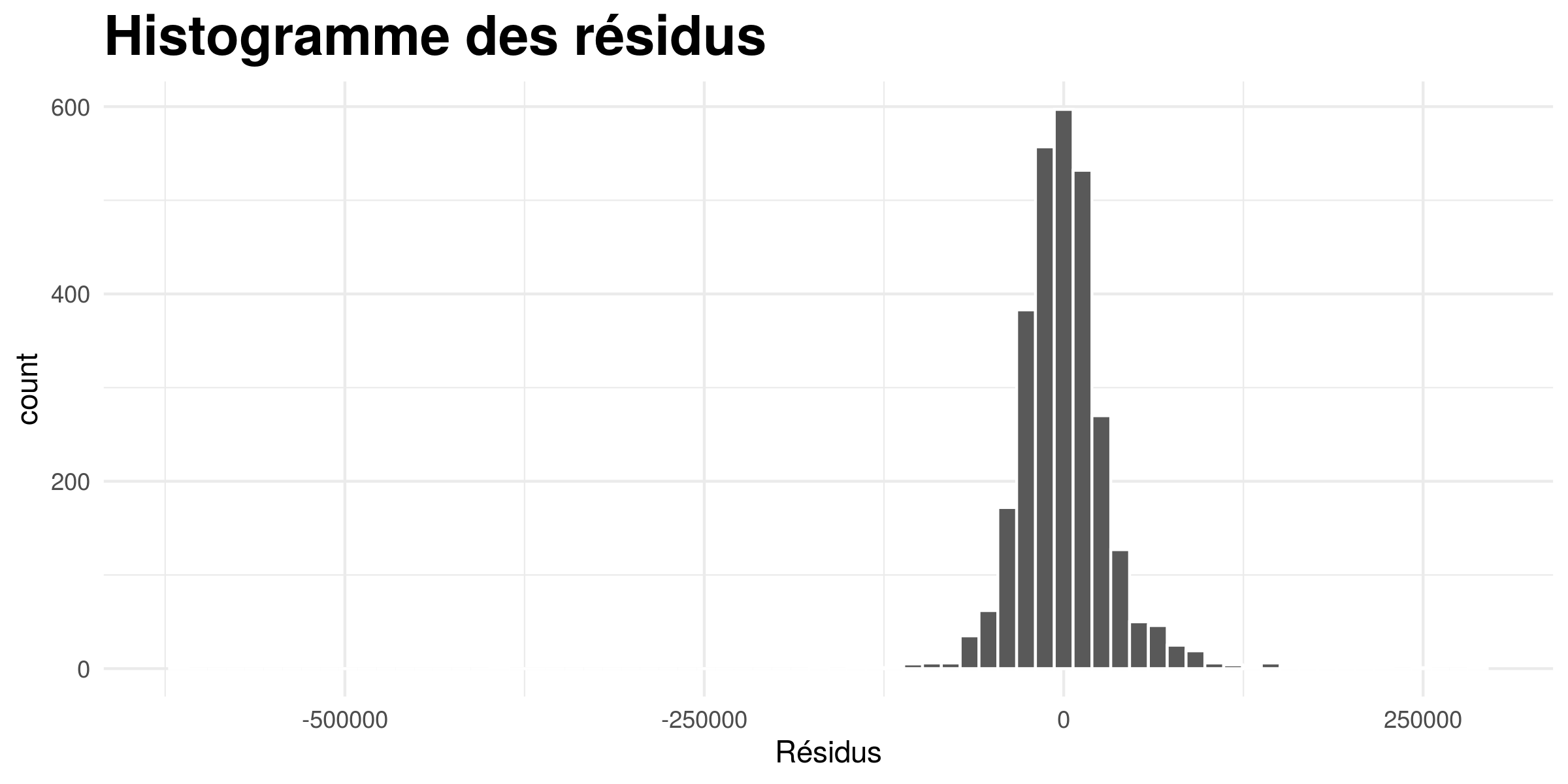
|  |  |
| --- | --- |
|  | VIF |
| Garage\_Cars | 5.7824481 |
| First\_Flr\_SF | 5.6211017 |
| Garage\_Area | 5.4901912 |
| Second\_Flr\_SF | 5.2394381 |
| Total\_Bsmt\_SF | 4.8520670 |
| TotRms\_AbvGrd | 4.3749371 |
| Year\_Built | 3.4850622 |
| Bsmt\_Unf\_SF | 3.1331090 |
| Full\_Bath | 2.7810995 |
| Bedroom\_AbvGr | 2.3082980 |
| Bsmt\_Full\_Bath | 2.1608346 |
| Half\_Bath | 2.1548278 |
| Year\_Remod\_Add | 1.9039890 |
| BsmtFin\_SF\_1 | 1.8157567 |
| Overall\_Cond | 1.5939918 |
| Fireplaces | 1.5305911 |
| Mas\_Vnr\_Area | 1.4010673 |
| Kitchen\_AbvGr | 1.3685725 |
| Enclosed\_Porch | 1.2292602 |
| Open\_Porch\_SF | 1.2220136 |
| Wood\_Deck\_SF | 1.2215701 |
| Lot\_Area | 1.2131699 |
| BsmtFin\_SF\_2 | 1.1665670 |
| Bsmt\_Half\_Bath | 1.1625433 |
| Overall\_Qual | 1.1515654 |
| Lot\_Frontage | 1.1347730 |
| Pool\_Area | 1.0857374 |
| Screen\_Porch | 1.0832776 |
| Low\_Qual\_Fin\_SF | 1.0686235 |
| Misc\_Val | 1.0460663 |
| Year\_Sold | 1.0434501 |
| Mo\_Sold | 1.0419652 |
| Three\_season\_porch | 1.0161864 |

Puisqu’il n’y a aucune valeur VIF supérieure à 10, il est tentant d’affirmer qu’il n’y a pas de multicolinéarité. Car nous avons vu dans l’analyse des corrélations que certaines variables étaient fortement corrélées. Ainsi bien qu’il n’y a pas de relation linéaire exacte entre les variables, il existe des pseudo-colinéarités relativement fortes entre les variables.

Nous pouvons vérifier les autres hypothèses des OLS pour déduire la colinéarité.

### Vérifions l’hypothèse de normalité des résidus

Nous pouvons représenter un histogramme de la distribution des résidus du modèle que nous venons d’estimer. L’une des hypothèses importantes des MCO est la normalité des résidus. Nous pouvons visuellement à l’aide de l’histogramme des résidus dire si les résidus suivent une loi normale.

 A l’aide de l’histogramme nous observons que la distribution ressemble à une loi normale, mais elle est fortement dirigée vers la droite. Nous pouvons faire le test de Shapiro-Wilk pour tester plus formellement la normalité des résidus. L’hypothèse nulle de ce test est que la distribution est normalement distribuée.

Shapiro-Wilk normality test  
  
data: mod$residuals  
W = 0.811785, p-value < 2.22e-16

Ici la p-value est inférieure à 5% donc, nous rejettons cette hypothèse nulle, la distribution ne suit pas une loi normale.

Dès lors que les résidus ne suivent pas une loi normale, les estimations faites par les OLS sont biaisées. Il importe donc d’utiliser d’autres méthodes d’estimations pour construire le modèle.

## Régressions sur variables latentes

### Pourquoi faire une régression sur les variables latentes ?

Les méthodes de régressions sur les latentes nous permettent d’adresser certaines limites que posent les méthodes de régression par les moindres carrés. En effet pour qu’une régression par les OLS soit sans biais il faut que toutes les hypothèses posées par les OLS soient vérifiées. Parmi ces hypothèses nous pouvons citer l’hypothèse de l’absence de corrélations entre les variables ou encore l’hypothèse d’absence de colinéarité entre les variables.

Or en général les phénomènes que nous cherchons à expliquer, dans notre travail nous cherchons à expliquer le prix de vente d’un bien immobilier, ne vérifient pas toujours ces hypothèses à cause entre autres de l’interdépendance entre les variables. Dans le cas d’un bien immobilier, la taille des pièces du bien dépend de la superficie totale de la maison. Cette situation nous amène à des résultats potentiellement biasés et non généralisables pour expliquer le phénomène.

Les méthodes de régression sur les variables nous permettent donc de corriger ces limites des régression standards en créant des variables synthétiques sur lesquelles nous allons faire les estimations.

Les variables synthétiques créées lors de l’estimation présentent une propriété très intéressante qui est celle de l’orthogonalité. En effet les variables synthétiques (ou les composantes) ne sont pas corrélées entre elles-mêmes, ce faisant nous pouvons donc estimer les paramètres du modèle que nous construisons sans courir le risque de violer les hypothèses posées par les moindres carrés ordinaires.

Pour mettre en place ces méthodes, nous allons utiliser la procédure PLS (Partial Least Squares) implémentée dans le logiciel SAS.

Les techniques mises en œuvre par la procédure PLS sont les suivantes:

* La régression sur composantes principales, qui extrait des facteurs qui expliquent le plus la variance des variables explicatives.
* La régression de rang réduit, qui extrait des facteurs qui expliquent le plus la variation de la variable réponse (y). Cette technique, également appelée analyse de redondance (maximale), diffère de la régression linéaire multiple uniquement lorsqu’il y a plusieurs variables endogènes.
* La régression sur les moindres carrés partiels, qui équilibre les deux objectifs d’explication de la variation de l’a réponse’endogène et expliquant la variation des prédicteurs. Deux formulations différentes pour les moindres carrés partiels sont disponibles: la méthode prédictive originale de Wold (1966) et la méthode SIMPLS de de Jong (1993).

En raison du fait que nous ne cherchons ici qu’à expliquer une seule variable (y), à savoir le prix des maisons, nous n’utiliserons pas la régression sur rang réduit car cela revient à faire une régression multiple, ce que nous avons déjà fait dans la section 3.2.

# Le principe de la méthode de la régression sur les composantes principales

Dans le cadre d’une régression multiple, nous cherchons à expliquer l’endogène y à l’aide de variables explicatives X. Nous cherchons donc

La régression sur les composantes principales consiste à remplacer la matrice des (n lignes et m colonnes) par une nouvelle matrice T dérivée de X, de n lignes et k colonnes, avec k très inférieur à m.

Chaque colonne k doit être une combinaison linéaire des variables d’origine.

Avec la matrice (m,k) des coefficients définissant les combinaisons linéaires, T une matrice dont les colonnes forment des variables « artificielles » obtenues par combinaison linéaire des variables d’origine.  
Une régression multiple est ensuite appliquée avec à la place de :

Il faut donc déterminer de manière à avoir une matrice de variables explicatives T plus adaptée au calcul de la régression que la matrice X d’origine.

Deux méthodes pour construire (et donc ) : RCP et RPLS

## Régression sur composantes principales (RCP ou PCR)

ACP puis régression multiple. L’ACP nécessite des variables quantitatives. A partir des variables centrées, l’ACP normée donne la matrice T dont les colonnes sont les composantes principales.

Elles sont orthogonales entre elles (plus de problème de colinéarité). La matrice est la matrice des coefficients des combinaisons linéaires (coordonnées sur les composantes principales).

Pour cette méthode, nous allons utiliser conjointement SAS et R afin de ressortir le maximum d’informations utiles à l’interprétation des résultats.

## Résultats d’estimations avec SAS

Nous avons spécifié la méthode “PCR” dans le procédure PLS afin de réaliser une régression sur les composantes principales. Il y a en total 34 variables explicatives pour 2930 observations.  
Nous choisissons également de faire une validation croisée sur le jeu de données.

Pour pouvoir déterminer avec confiance le nombre de composantes principales (ou facteurs) à retenir dans la PCR, l’algorithme utilise une technique de validation croisée. Une technique de validation croisée consiste à diviser l’échantillon en plusieurs groupes sur lesquels des ajustements successifs sont effectués. L’ajustement du modèle est effectué sur le groupe d’apprentissage et les résultats sont comparés aux observations du groupe test. L’efficaité prédictive est évaluée à l’aide l’indicateur PRESS (Predicted REsidual Sum of Squares).

Le tableau suivant nous montre que la valeur minimale de la validation croisée est atteinte si le nombre de composantes principales est 8. Toutefois il est préférable de se fier au test qui indique si oui ou non il existe une différence significative entre les 8 composantes.

Le test nous indique qu’à partir de la 4e composante il n’existe plus de différences significatives entre les composantes. Nous allons donc retenir 5 composantes pour la suite de l’analyse.

Il convient maintenant d’interpréter les résultats du modèle construit.  
Le tableau suivant affiche la variation de pourcentage expliquée par composantes principales.

Le tableau affiche deux catégories de résultats : la variation de pourcentage expliquée par composantes principales sur les effets du modèle (les variables explicatives) et la variation de pourcentage expliquée de la variable dépendante pour chaque facteur (ou composante) ajouté.

Nous voyons donc que pour les 5 composantes retenues, la régression sur les composantes principales explique 75.93% de la variable dépendante et environ 43% des variables explicatives.

Cela signifie que les 5 composantes retenues qui arrivent à expliquer 43% de la variance des variables explicatives sont capables d’expliquer 75.93% de la variance des prix.

Les résultats de SAS ne nous donnent pas les coefficients associés à chaque composante. Nous pouvons donc les estimer avec R en utilisant le même nombre de composantes que la validation croisée avec SAS a indiqué.

Call:  
lm(formula = Sale\_Price ~ ., data = coord\_pc)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-535355 -20542 -3320 17245 304779   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 180796.06 724.62 249.5033 < 2.2e-16 \*\*\*  
Dim.1 27395.65 291.64 93.9374 < 2.2e-16 \*\*\*  
Dim.2 -7046.03 415.69 -16.9502 < 2.2e-16 \*\*\*  
Dim.3 478.57 497.63 0.9617 0.3363   
Dim.4 2778.86 513.24 5.4144 0.00000006649 \*\*\*  
Dim.5 5769.77 628.86 9.1750 < 2.2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 39224 on 2924 degrees of freedom  
Multiple R-squared: 0.75934, Adjusted R-squared: 0.75893   
F-statistic: 1845.2 on 5 and 2924 DF, p-value: < 2.22e-16

Les résultats globaux du modèle que nous trouvons sont les mêmes avec R. Les 5 composantes retenues expliquent bien 75,93% de la variance des prix de vente de la maison.

Nous pouvons écrire ce résultat sous forme d’une équation :

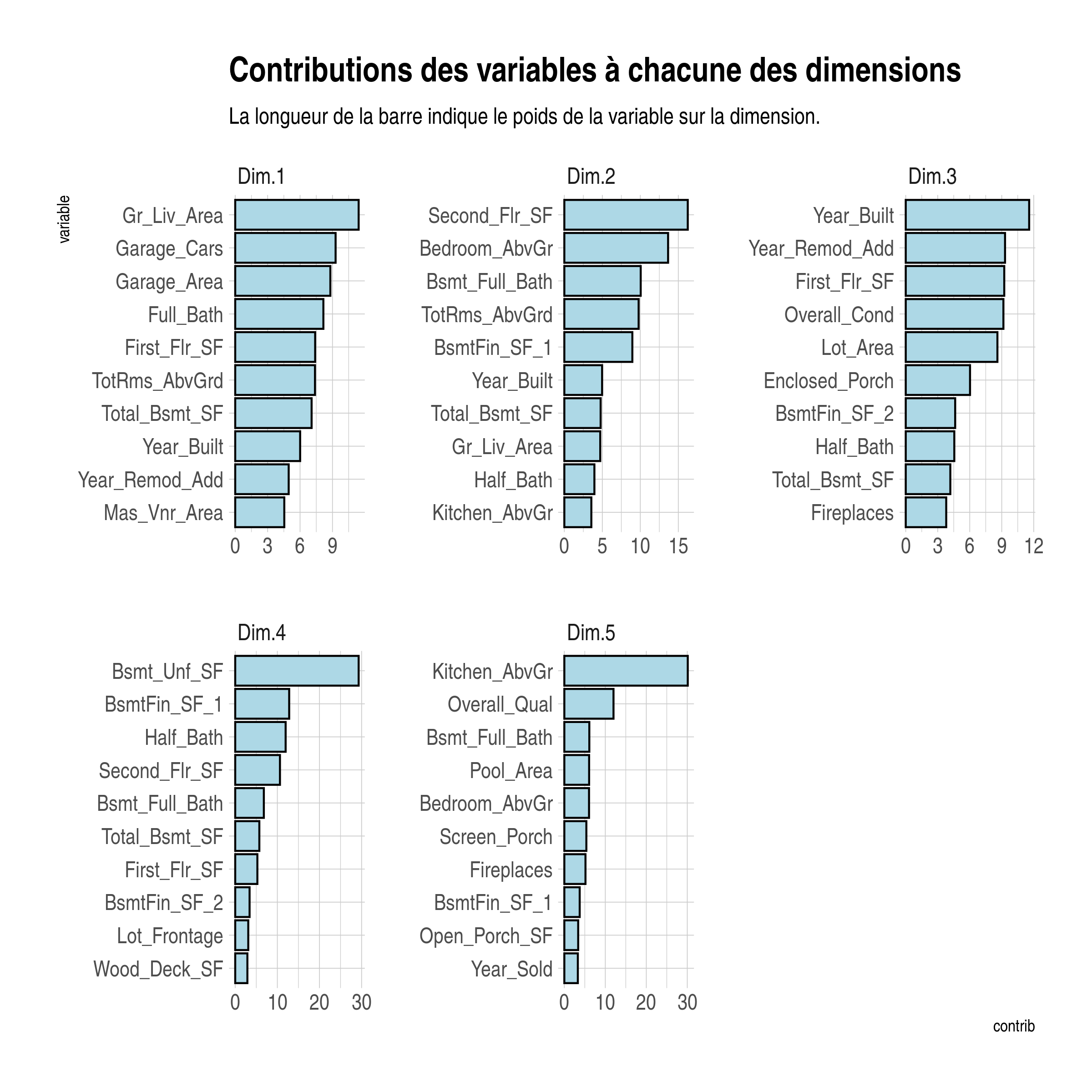
Tous les coefficients sont positifs, à l’exception du coefficient de la dimension 2.

Le coefficient de la dimension 3 est le plus faible et est non significatif (p-value > 5%).

Pour pouvoir interpréter les coefficients, il importe de savoir ce que mesure chaque composante. A l’aide de R nous pouvons, comme nous l’avons fait dans la section 2.3, afficher ce que mesure chaque composante en affichant les 10 premières contributions à chaque composante.

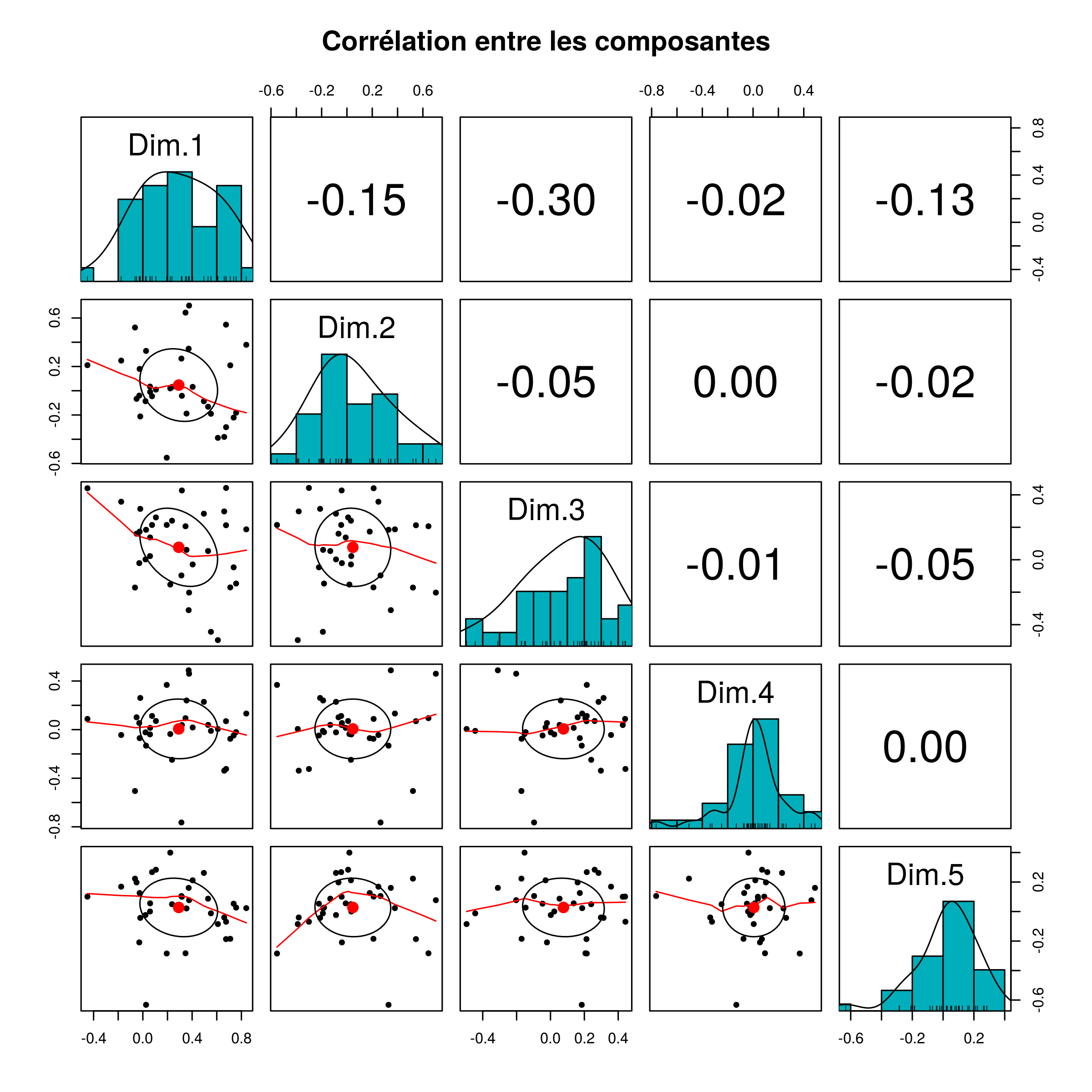
Nous décidons de représenter ces composantes sur un graphique.

Les contributions des variables aux différents axes factoriels sont :



Le poids des variables sur les trois premières dimensions est assez équilibrée. Les deux dernières dimensions sont chacune fortement dominées par une variable.

Les variables synthétiques que produit l’ACP sont orthogonales entre elles. Pour le vérifier nous pouvons représenter les corrélations entre les dimensions.



### Dimension 1

La dimension 1 fait référence à la superficie de la maison, que ce soit la taille de la résidence et des pièces qui composent la maison. Cette composante est également dans une certaine mesure influencée par l’âge de la maison (l’année où la maison a été construite et l’année où elle a été renovée.)

### Dimension 2

La deuxième dimension se refère aux caractéristiques des maisons telles que la superficie du deuxième étage, et le nombres de pièces (salles de bain, chambres, cuisines) se trouvant au deuxième étage. Cette composante détermine donc le prestige de la maison.

### Dimension 3

La troisième composante mesure l’âge de la maison et la condition de la maison au moment de la vente.

### Dimension 4

Cette composante est principalement dominée par le fait que la maison contient des parties inachevées.

### Dimension 5

Cette composante est fortement dominée par le fait s’il y a une cuisine à l’étage et par la condition générale de la maison.

## Limite de la régression sur les composantes principales

Nous venons de voir l’intérêt de la régression sur les composantes principales. Nous avons vu notamment que les composantes issues de l’ACP sont orthogonales en ce sens qu’il cette méthode permet d’obtenir des composantes orthogonales entre elles. Toutefois, l’analyse en composante principale que nous avons effectuée maximise les variances des variables indépendamment des variations de Y : la variance de Y n’est pas maximisée. Ceci peut surestimer le poids de certaines explicatives et conduire de ce fait à de mauvaises interprétations des résultats.

Nous pouvons donc utiliser la deuxième méthode de cette famille.

## Régéressions sur les moindres carrés partiels

Comme pour la régression PCR, calcul de T dont les éléments sont les « scores » et les colonnes les « composantes » : avec matrice des poids (ou loadings) et matrice des variables explicatives centrées (n,m).

Mais, contrairement à la PCR, le calcul de se fait en tenant compte de la variable à prédire . Double modélisation :

(1)

(2)

Avec : R matrice des résidus associées à la prédiction de ; vecteur des résidus associé à la prédiction de .

Première étape : calculer t1 la première composante principale, puis estimer (1) et (2) à une seule composante :

$$X = t\_1 . p\_1 + R1$ et $Y = t\_1 . q\_1 + F\_1$$

Avec de dimension (n,1) de dim (1,m) loadings de dim (1,1) loadings

On introduit une deuxième composante . Ainsi, une ligne de est égale à :

L’introduction de nouvelles composantes se fait selon la même procédure : partant d’un modèle à k composantes, on créé un nouveau modèle à k+1 composantes en calculant une nouvelle composante tk, puis les paramètres des deux modèles couplés :

$ X = t\_1p\_1+t\_2 p\_2 +…+t\_k p\_k + t\_k\_+1 p\_k+1 + R\_k+1 $

### Les loadings et les weights

# Comparaison des performances des différentes méthodes d’estimation

# Bibliographie

Alboukadel Kassambara and Fabian Mundt (2019). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.6. <http://www.sthda.com/english/rpkgs/factoextra>

Bjørn-Helge Mevik, Ron Wehrens and Kristian Hovde Liland (2019). pls: Partial Least Squares and Principal Component Regression. <http://mevik.net/work/software/pls.html>, <https://github.com/bhmevik/pls>.

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

Jan Niklas Schneider and Timothy Raymond Brick (2019). plsr: Pleasure - Partial Least Squares Analysis with Permutation Testing. R package version 0.0.1.

Max Kuhn (2017). AmesHousing: The Ames Iowa Housing Data. R package version 0.0.3.

The PLS Procedure, SAS/STAT® 13.1 User’s Guide <https://support.sas.com/documentation/onlinedoc/stat/131/pls.pdf>

# Annexes

## Annexe 1 : Liste exhaustive des colonnes

Le jeu de données contient initialement 81 colonnes

|  |  |  |  |
| --- | --- | --- | --- |
| Colonnes | .. | … | …. |
| MS\_SubClass | Roof\_Style | Central\_Air | Garage\_Area |
| MS\_Zoning | Roof\_Matl | Electrical | Garage\_Qual |
| Lot\_Frontage | Exterior\_1st | First\_Flr\_SF | Garage\_Cond |
| Lot\_Area | Exterior\_2nd | Second\_Flr\_SF | Paved\_Drive |
| Street | Mas\_Vnr\_Type | Low\_Qual\_Fin\_SF | Wood\_Deck\_SF |
| Alley | Mas\_Vnr\_Area | Gr\_Liv\_Area | Open\_Porch\_SF |
| Lot\_Shape | Exter\_Qual | Bsmt\_Full\_Bath | Enclosed\_Porch |
| Land\_Contour | Exter\_Cond | Bsmt\_Half\_Bath | Three\_season\_porch |
| Utilities | Foundation | Full\_Bath | Screen\_Porch |
| Lot\_Config | Bsmt\_Qual | Half\_Bath | Pool\_Area |
| Land\_Slope | Bsmt\_Cond | Bedroom\_AbvGr | Pool\_QC |
| Neighborhood | Bsmt\_Exposure | Kitchen\_AbvGr | Fence |
| Condition\_1 | BsmtFin\_Type\_1 | Kitchen\_Qual | Misc\_Feature |
| Condition\_2 | BsmtFin\_SF\_1 | TotRms\_AbvGrd | Misc\_Val |
| Bldg\_Type | BsmtFin\_Type\_2 | Functional | Mo\_Sold |
| House\_Style | BsmtFin\_SF\_2 | Fireplaces | Year\_Sold |
| Overall\_Qual | Bsmt\_Unf\_SF | Fireplace\_Qu | Sale\_Type |
| Overall\_Cond | Total\_Bsmt\_SF | Garage\_Type | Sale\_Condition |
| Year\_Built | Heating | Garage\_Finish | Sale\_Price |
| Year\_Remod\_Add | Heating\_QC | Garage\_Cars | Longitude |

Description des colonnes :

SalePrice - the property’s sale price in dollars. This is the target variable that you’re trying to predict.

MSSubClass: The building class  
MSZoning: The general zoning classification  
LotFrontage: Linear feet of street connected to property  
LotArea: Lot size in square feet  
Street: Type of road access  
Alley: Type of alley access  
LotShape: General shape of property  
LandContour: Flatness of the property  
Utilities: Type of utilities available  
LotConfig: Lot configuration  
LandSlope: Slope of property  
Neighborhood: Physical locations within Ames city limits  
Condition1: Proximity to main road or railroad  
Condition2: Proximity to main road or railroad (if a second is present)  
BldgType: Type of dwelling  
HouseStyle: Style of dwelling  
OverallQual: Overall material and finish quality  
OverallCond: Overall condition rating  
YearBuilt: Original construction date  
YearRemodAdd: Remodel date  
RoofStyle: Type of roof  
RoofMatl: Roof material  
Exterior1st: Exterior covering on house  
Exterior2nd: Exterior covering on house (if more than one material)  
MasVnrType: Masonry veneer type  
MasVnrArea: Masonry veneer area in square feet  
ExterQual: Exterior material quality  
ExterCond: Present condition of the material on the exterior  
Foundation: Type of foundation  
BsmtQual: Height of the basement  
BsmtCond: General condition of the basement  
BsmtExposure: Walkout or garden level basement walls  
BsmtFinType1: Quality of basement finished area  
BsmtFinSF1: Type 1 finished square feet  
BsmtFinType2: Quality of second finished area (if present)  
BsmtFinSF2: Type 2 finished square feet  
BsmtUnfSF: Unfinished square feet of basement area  
TotalBsmtSF: Total square feet of basement area  
Heating: Type of heating  
HeatingQC: Heating quality and condition  
CentralAir: Central air conditioning  
Electrical: Electrical system  
1stFlrSF: First Floor square feet  
2ndFlrSF: Second floor square feet  
LowQualFinSF: Low quality finished square feet (all floors)  
GrLivArea: Above grade (ground) living area square feet  
BsmtFullBath: Basement full bathrooms  
BsmtHalfBath: Basement half bathrooms  
FullBath: Full bathrooms above grade  
HalfBath: Half baths above grade  
Bedroom: Number of bedrooms above basement level  
Kitchen: Number of kitchens  
KitchenQual: Kitchen quality  
TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)  
Functional: Home functionality rating  
Fireplaces: Number of fireplaces  
FireplaceQu: Fireplace quality  
GarageType: Garage location  
GarageYrBlt: Year garage was built  
GarageFinish: Interior finish of the garage  
GarageCars: Size of garage in car capacity  
GarageArea: Size of garage in square feet  
GarageQual: Garage quality  
GarageCond: Garage condition  
PavedDrive: Paved driveway  
WoodDeckSF: Wood deck area in square feet  
OpenPorchSF: Open porch area in square feet  
EnclosedPorch: Enclosed porch area in square feet  
3SsnPorch: Three season porch area in square feet  
ScreenPorch: Screen porch area in square feet  
PoolArea: Pool area in square feet  
PoolQC: Pool quality  
Fence: Fence quality  
MiscFeature: Miscellaneous feature not covered in other categories  
MiscVal: $Value of miscellaneous feature  
MoSold: Month Sold  
YrSold: Year Sold  
SaleType: Type of sale  
SaleCondition: Condition of sale

Présenter le sujet : - Statistiques descriptives univariées (variable par variable) - Répondre à une question ou une série de questions - Présenter les statistiques bivariées : corrélation , test du khi-deux - ACP au préalable Comparer les résultats des différentes méthodes : - dire pourquoi utiliser chacune des méthode : pourquoi une OLS, pourquoi une PCR, PLS. - Dire quel résultat auquel on s’attend

Présenter uniquement les tableaux de résultats principaux et les graphique. Mettre ce qui est moins pertinent en annexes. Commenter les résultats

Conclure : analyser les résultats et les comparer à ce qu’on avait dit au début Quels résultats ces méthodes apportent? -