PROJET DE DATA MINING

Les déterminants des prix des maisons à New-York

22 septembre

Université d'Angers

Réalisée par : LENSARI Yaakoub
Encadré par : Daniel Christophe

Introduction

Le marché de l'immobilier new-yorkais est réputé comme l'un des plus dynamique et dominants connu par sa capacité à s'adapter rapidement aux différents chocs économiques, caractérisée par une tendance perpétuelle de croissance, attire aussi tant les résidents que les investisseurs. Mettre en évidence les déterminants de prix de maisons dans cette métropole cosmopolite revient à étudier un milieu caractérisé par un environnement économique et sociales extrêmement complexes ou plusieurs facteurs tel que l'offre, la demande, la conjoncture et les politiques économiques s'entremêlent.

Dans ce rapport, à travers différentes méthodes d'analyse des données nous allons étudier une base de données en plus d'autres sources académiques sur le marché immobilier dans le but de dégager les principaux facteurs susceptibles de déterminer le niveau de prix des maisons au sein de la ville de New-York.

Ville cosmopolite, le marché immobiliers à New-York est très diversifié notamment chaque quartier est différentes ayants ces propres caractéristiques avec un environnement économiques, social et ou démographiques qui lui est propre entrainant une variation de prix de l'immobilier en fonction d'une multitude de facteurs.

L'objectif étant dans une première partie d'énoncé une revue de littérature d'un certain nombre des travaux qui ont été déjà réalisés sur le sujet principalement aux Etats-Unis et au canada. En second lieu, nous allons effectuer une présentation de la base de données ainsi qu'une analyse descriptive de ces données.

En outre, nous allons consacrer une troisième partie à l'analyse approfondie de la base à l'aide d'un ACP et je finirais cette partie par l'élaboration d'un modèle économétrique de régression linéaires multiple pour mettre en évidence les principaux déterminants des prix des maisons.

Finalement, nous allons abordés une dernière partie qui sera consacrée à l'application des méthodes PLS et PCR permettant d'approfondir notre analyse.

A travers cet méthodologie d'analyse, j'espère fournir des informations significatives et révéler la nature de l'influence exercé par chaque déterminant sur les prix des logements et ainsi fournir un rapport à la fin susceptible d'être un outil d'aide à la décision.

Revue de littérature

Les déterminants des prix des maisons ont fait l'objet de plusieurs études dans des nombreux régions du monde notamment aux états unis et plus particulièrement au sein de la ville de New-York disposant d'un des marchés de l'immobilier le plus adynamiques au monde, Créant un énorme attrait tant pour les chercheurs que les acteurs du marché immobilier.

1. Déterminants du prix réel des logements au Canada, Etude mené par **Mario FORTIN** et **André LECLERC** (Janvier 2002). L'objectif étant d'identifier quelles variables ont affecté l'équilibre du marché du logement

A travers un modèle sur des données annuelles qui ont été prélevés sur une période allant de 1956 à 2001 au Canada, Ils ont réussi à élaborés ce modèle permettant d'expliquer l'évolution du prix moyen et du nombre d'unités de logement. Les conclusions de leurs modélisations révèlent que Trois variables exercent une influence significative sur le prix réel, à savoir le revenu réel par personne adulte, le taux d'intérêt nominal sur les prêts hypothécaires à l'habitation à 5 ans et la croissance de la population de 25 à 54 ans. Donc cette étude met en lumière l'importance de l'emprunt et du revenu mais aussi il souligne le rôle que joue une composante socio-démographique particuliers dans la détermination du prix des maisons.

2.Les déterminants des prix de l'immobilier aux Etats-Unis après la Grande Récession une analyse des bornes extrêmes mené par **Achille Dargaud FOFACK** et **Serge Djoudji TEMKENG**.

Cette étude a pour objectif d'expliquer la hausse record des prix des logements aux Etats-Unis qui ont dépassé les niveaux record d'avant la crise de subprimes entrainant des inquiétudes des acteur du marché de l'immobilier. Ainsi à travers des données mensuelles sur la période allant de juillet 2009 – date à laquelle l'économie américaine est sortie de la récession – jusqu'en avril 2019, ont été étudiée via la méthode d'analyse des bornes extrêmes. A partir de cette étude ils ont dégagé 12 déterminants potentiels des prix de l'immobilier. Les résultats de l'étude ont montré les conclusions suivantes. Le crédit immobilier, le taux d'intérêt sur les prêts hypothécaires, les dépenses en constructions, la politique monétaire de l'assouplissement quantitatif qui a été mise en place à l'époque par la réserve fédérale américaine dans le but de stimuler les prés et relancer l'activité économique, constituent les principaux facteurs qui ont déterminés les prix des maisons aux Etats-Unis et expliquent notamment la hausse anormale du niveau des prix des loyers à l'époque.

Présentation des données

Les données utilisées dans ces rapports sont extraites sur le site Kagel et sont des données Public sur les déterminants des prix des logements aux États-Unis et plus précisément la ville de New-York. Cependant ces données sont mises à jour chaque année, sa qui signifie qu'ils sont bien adaptés comme outils à jour pour menées des études prospectives du marché de l'immobilier américains.

Ainsi, nous avons dans notre base de données 13 variables et 545 observations

Tableau 1 : nom, la signification et le codage des variables

Variables	**Signification**	**Codage**	
Prix	*Le prix d'une maison*		
Superficie	*La superficie de la maison*		
Chambres	*Nombres de chambre dont dispose la maison*		
Bains	*Nombres de Bains dont dispose la maison*		
Étages	*Nombres d'étages dont dispose la maison*		
Route	*Existence d'une route en face ou à côté de la maison*	*1*	*Existence d'une route*
		0	*Pas de Route*
Chamb_invit	*Existence d'une chambre pour l'invité dans la maison*	*1*	*Existence d'une chambre*
		0	*Pas de chambre*

Sous-sol	*Maison ayant la partie sous-sol*	*1*	*Existence d'une Sous-sol*
		0	*Pas de Sous-sol*
Chauffe_eau	*Maison ayant équipé ou non d'un dispositif permettant de chauffer l'eaux*	*1*	*Equipé du dispositif*
		0	*Non équipé du dispositif*
Climatisation	*Maison ayant équipé ou non d'un dispositif de climatisation*	*1*	*Equipé d'un climatiseur*
		0	*Non équipé d'un climatiseur*
Parking	*Nombres de parking dont dispose la maison*		
Prefarea			
Meublement	*Maison meublée ou non*	*1*	*Meublée*
		0	*Non meublée*

Analyse descriptive

Nous avons, ci-dessous les répartitions et les statistiques descriptives de la base des données. Cependant ces statistiques révèlent un certain nombre d'un insight significatif sur les fluctuations du prix de l'immobiliers aux États-Unis.

```
      prix             Superficie       chambres         bains           etages          Route         chamb_invit
 Min.   : 1750000   Min.   : 1650   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   :0.000
 1st Qu.: 3430000   1st Qu.: 3600   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.0000   1st Qu.:0.000
 Median : 4340000   Median : 4600   Median :3.000   Median :1.000   Median :2.000   Median :1.0000   Median :0.000
 Mean   : 4766729   Mean   : 5151   Mean   :2.965   Mean   :1.286   Mean   :1.806   Mean   :0.8587   Mean   :0.178
 3rd Qu.: 5740000   3rd Qu.: 6360   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:0.000
 Max.   :13300000   Max.   :16200   Max.   :6.000   Max.   :4.000   Max.   :4.000   Max.   :1.0000   Max.   :1.000
    sous-sol        chauffe_eau      Climatisation       parking          prefarea        meublement
 Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Length:545
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
 Median :0.0000   Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000   Mode  :character
 Mean   :0.3505   Mean   :0.04587   Mean   :0.3156   Mean   :0.6936   Mean   :0.2349
 3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :3.0000   Max.   :1.0000
```

En premier lieu, nous constatons une grande variabilité du prix des logements allant de 1 750 000 dollars à 13 300 000 dollars, mettant en évidence le caractère de diversité et cosmopolite de la ville. Quand a la médiane qui centre la distribution elle est de 4 340 000 dollars et proche du prix moyen des logements qui est de 4 766 729 dollars à New-York toute chose égal par ailleurs.

En outre la superficie des maisons est aussi très variée allant de 1650 16200 pieds carré d'où l'écart important entre la moyenne qui est de 5151 pieds carrés et la médian avec 4600 pieds carrés, Soulignant une grande hétérogénéité d'offre en matière de propriétés.

En matière de dotations en terme du nombre de chambres bains et route étages sont aussi variées allant 1 à 6 chambres au maximum et de 1 à 4 pour le nombre des bains et d'étages.

Cependant, en observant les différentes répartitions on aperçoit qu'en moyenne, nous avons 3 chambres 1 bains, environ 2 étages et 1 route principale par maison.

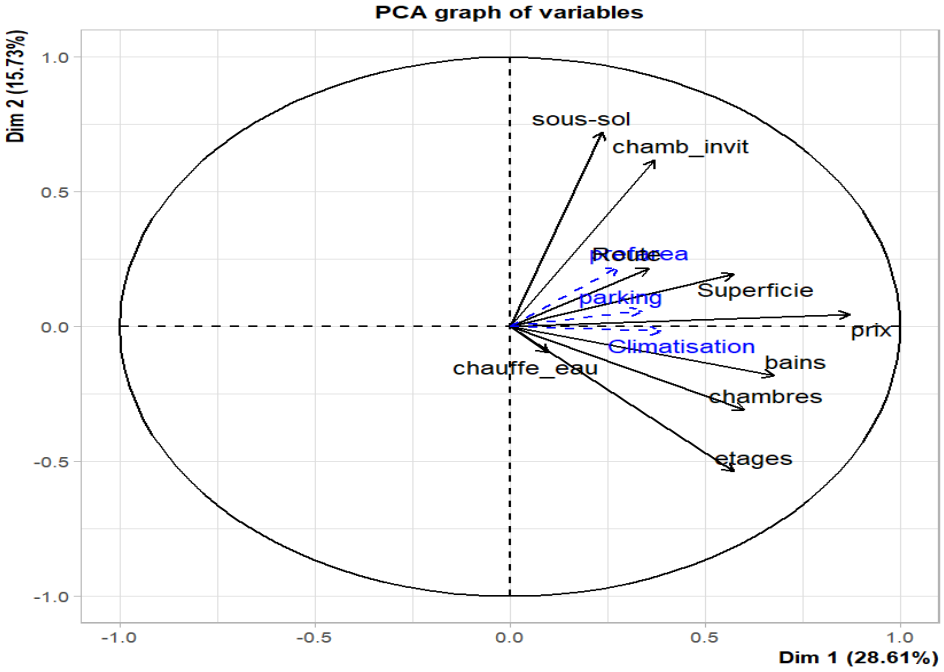
En outre, nous avons aussi d'autres caractéristiques supplémentaires qui peuvent expliquer la variabilité des prix comme le nombre de propriétés qui dispose d'un sous-sol représentant aux environs de 35%, quant aux parts des maison qui disposent d'un équipement permettant de chauffer l'eaux représentant à peine 5%, les maisons équipées d'un dispositif de climatisation représentent 32% et pour finir 69 % des maison ont au moins un parking. Ce chiffre met en évidence une plus grande importance des facteurs lieux à la structure et a l'aménagement dans la détermination du prix des logements comme la superficie, sous-sol, Parkings, nombres d'étages et chambres etc.

Analyse transversale

Le cercle de corrélations

Le cercle de corrélation permet d'analyser les liaisons entre les variables et Nous renseigner sur la qualité de la représentation de chaque variable. Plus la pointe de la flèche de la variable est proche du cercle, plus la variable est bien représentée. Si l'angle séparant les deux variables est petit, cela implique que la corrélation entre les deux variables est proche de 1. Et si les flèches sont opposées, cela implique que la corrélation est proche de -1.

Le graphique ci-dessous représente les variables ainsi que leurs positionnements par rapport à l'origine déterminant ainsi leurs contributions dans la construction des dimensions de l'ACP.



En premiers lieux, on constate un effet de taille. La majorité des variables sont fortement corrélés avec l'axe 1, notamment les niveaux des prix, la superficie, le nombre de chambres, d'étages et le nombres de bains dont dispose la maison sont les variables enregistrant les plus grandes contributions sur l'axe 1 et donc ayant l'effet de taille le plus important sur cette

dimension. Cependant on constate que le niveau des prix est fortement corrélé avec la dimension 1, ce qui signifie que cette dimension représente le niveau des prix des maisons.

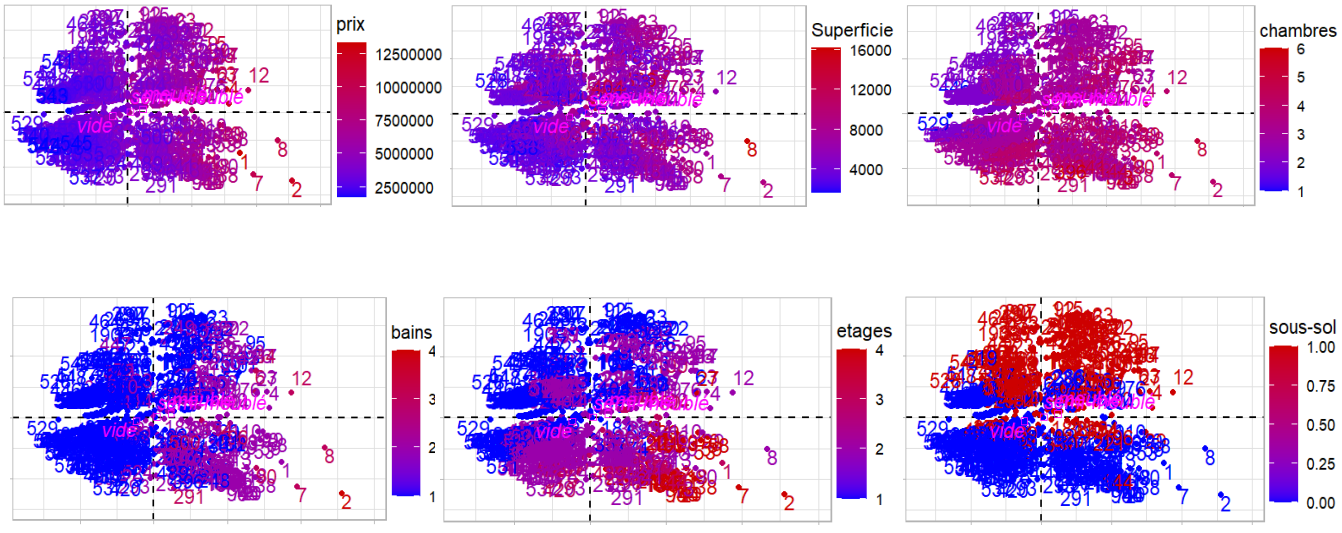
L'axe 1 oppose principalement les caractéristiques liées à la structure et a l'aménagement contre celles liées au confort et aux commodités.

Quant aux variables sous-sol et maison avec ou non une chambre d'invité enregistrent les plus grandes contributions sur l'axe 2 avec des effets de taille très élevé sur cette dimension.

En outre, la variable chouffe d'eau, parking, climatisation et prefarea enregistrent les plus faibles contributions à l'axe 1. Ce qui montre qu'ils existent d'autre dimensions sur lesquels ces variables sont mieux représenté. En effet Cet information montre que les déterminants tel que la climatisations l'existence ou non d'un parkings et d'équipement qui permet de réchauffer la maison ne constituent pas les déterminants principaux des prix des loyers au sein de la ville de New-York contrairement au reste des variables ayant des contributions respectivement très élevé sur les deux dimensions. Ces conclusions peuvent s'expliquer par le déplacement dans les grandes villes ou les transports en commun sont très développées, et ou les équipements tel que la climatisation et autres sont très peu cher par rapport au cout des loyers et ne constituent pas des facteurs capitaux de prise de décisions ou non de l'achat d'une maison dans une ville tel que New-York.

Graphiques des individus

Dans les graphiques ci- dessous nous avons la répartition des individus sur les graphes en fonction des principaux facteurs ayant l'influence le plus significatif dans l'achat d'une maison



Premièrement, nous avons les individus sous forme des nuages de points qui sont bien étalés sur les plans factoriels. Dans ce cas les points rouges représentent les maisons ayant la caractéristique qui est affichée sur la légende.

Quand on analyse le graphe avec les prix, On conste que l'axe 2 du graphe ci-dessus oppose les maisons les plus chères contre les maisons les moins chères. Notamment les maisons les plus chères sont localisées du cotés droit du graphe et les maisons avec des prix inférieur à la moyenne sont localisée du côté gauche du graphe. Autres constats en observant le reste des graphes on constate l'existence d'une forte corrélation positive entre le niveau de prix et les différentes caractéristiques des logements.

Régression linéaire multiple

Ci-dessous, nous avons les résultats du modèle qui est une régression du prix des maisons en fonctions d'un certains nombres de variables explicatifs a produits les résultats significatifs suivants dont dépend les variations des prix des loyers au sein de la ville de New-York.

En premier lieu, il est intéressant de vérifier la robustesse du modèle avant toute interprétation

Valeur vif : sert dans ce cas pour vérifier la présence d'une une éventuelle multi-colinéarité entre les variables du modèle

```
vif(MCO)
 Superficie       bains      etages       Route  chamb_invit Climatisation     parking     prefarea DM$`sous-sol`
   1.312957    1.213315    1.290135    1.146430    1.209870    1.173140    1.186446    1.143903    1.288307
```

Je constate dans ce cas que la valeur vif pour mes variables ne dépasse pas 1.5, ceux qui indique une absence de multi-colinéarité.

On 'aussi le P-value associé au test de ficher qui est inférieur à 2.2e-16, ceux signifient que le modèle est globalement significatif pour un seuil critique de 0,05 et un niveau de confiance de 95%.

```
Call:
lm(formula = prix ~ Superficie + bains + etages + Route + chamb_invit +
    Climatisation + parking + prefarea + DM$`sous-sol`, data = DM)

Residuals:
     Min       1Q   Median       3Q      Max
-2990719  -645740   -57532   496650  5263994

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -28228.49  195251.53  -0.145 0.885101
Superficie        251.07      24.85  10.105  < 2e-16 ***
bains          1073318.79  103154.77  10.405  < 2e-16 ***
etages          507787.45   61611.85   8.242 1.32e-15 ***
Route           431562.62  144515.88   2.986 0.002953 **
chamb_invit     312654.16  135193.09   2.313 0.021121 *
Climatisation   811934.76  109563.61   7.411 4.94e-13 ***
parking         325519.37   59489.16   5.472 6.85e-08 ***
prefarea        636729.40  118612.84   5.368 1.19e-07 ***
DM$`sous-sol`   418415.25  111841.08   3.741 0.000203 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1098000 on 535 degrees of freedom
Multiple R-squared:  0.6614,    Adjusted R-squared:  0.6557
F-statistic: 116.1 on 9 and 535 DF,  p-value: < 2.2e-16
```

En observant ce modèle, nous constatons la présence des relations significatives entre le niveau des prix et les variables explicatifs tel que : Surfaces nombres d’étages et salle de bains etc…

Parmi ces variables la superficie apparait comme une caractéristique majeure dans la détermi nation du prix d’une propriété sur le marché de l’immobilier. Notamment 1 hausse de la superfi cie d’une maison de 1 mètre carré entraine une hausse du loyer de 251,07 dollars toute chose égal par ailleurs. Nous avons aussi d’autres variables très significatives à savoir le nombre d’é tages et de salle de bains révélant une forte demande pour les maisons ayant un confort et un e plus grande capacité d’accueil. Notamment une extension de la maison de 1 étages supplém entaires entraine une hausse de son prix de 507787,45 dollars.

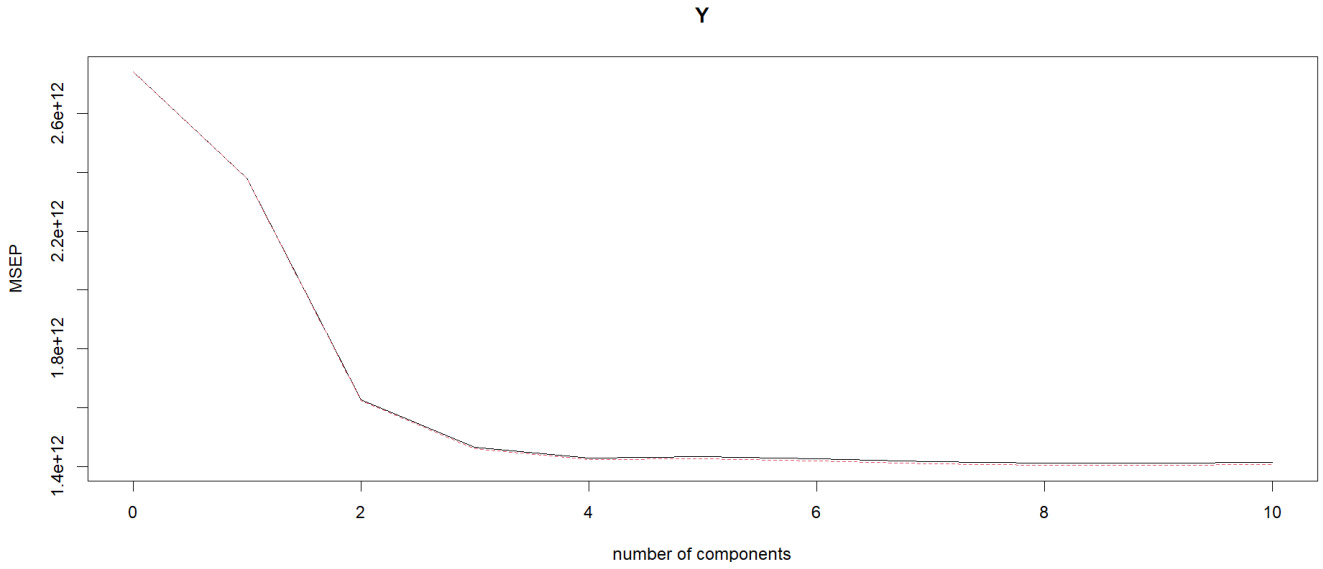
Finalement, nous avons les variables tel que proximité d’une route principale, climatisons, sou s-sol, chambre d’invité qui sont aussi significatif révélant néanmoins l’importance du rôle non n égligeable des commodités, des équipements ainsi d’un bon emplacement dans la déterminati on de la valeurs immobilières d’une maison.

Méthode PLS

Le choix de cette méthode se justifie par la nature de la variable d’intérêt et des données à ma dispositions et la régression PLS est le plus compatible pour mieux ajuster la variables prix en fonction des variables explicatifs qui sont numériques. néanmoins nous avons un grand nombres d’observation rendant difficile l’application de cette méthodes

Validation croisée

Cette méthode sert à déterminer le nombre de composante optimale dans une analyse par PL S et pour rendre plus compatible l'application d'un PLS à mes données j'ai réduit le nombre d' observation qui est très élevés 0a 300 observations.



Après l'observation on voit que 3 est le nombre de composantes correspondant au point ou l'erreur quadratique moyenne de prédiction est le plus minimale, correspondant aux nombres de composantes optimale pour ce modèle PLS.

Dans ce cas le test de validation croisée nous indique de garder que trois composantes qui minimise le RMSE.

Après avoir effectué un PLS en retenant que trois composantes, nous avons les résultats ci-dessous sur les pourcentages de la variance des données X et Y expliqué par chaque composante.

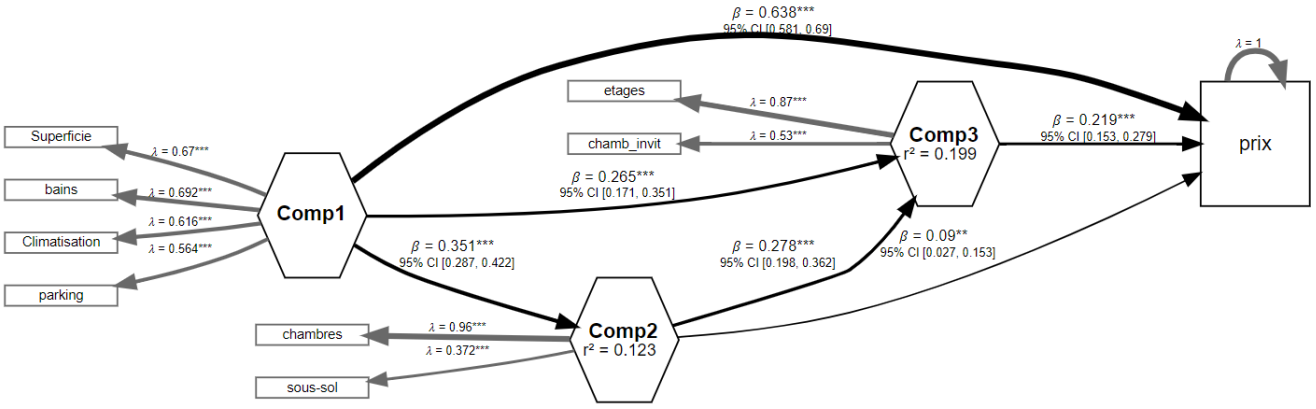
```
Data:    X dimension: 300 11
         Y dimension: 300 1
Fit method: kernelpls
Number of components considered: 3

VALIDATION: RMSEP
Cross-validated using 300 leave-one-out segments.
       (Intercept)  1 comps  2 comps  3 comps
CV          1655667  1193608  1184449  1185699
adjCV       1655667  1193552  1184382  1185610

TRAINING: % variance explained
   1 comps  2 comps  3 comps
X    15.31    27.95    37.41
Y    50.46    52.31    52.79
```

En effet, nous avons environs 37 % de la variance cumulé de la variable prédicteur X expliqué par le modèle. Cependant, nous avons 52.2% de la variance cumulé de la variable cible Y qui est expliqué par le modèle, ce qui signifie une bonne capacité d'ajustement et prédiction de notre modèle.

L'erreur de production baisse à mesure que le nombre de composante augmente et la proportion de la variance expliquée aussi augmente jusqu'au niveaux de la troisième composante qui constitue le point de retour ou l'erreur de prédiction devient constant malgré la hausse du nombre de composante.

Présentation graphique du modèle PLS

L'objectif étant d'expliquer les fluctuations du prix des maisons sur le marché de l'immobilier New-Yorkais à travers trois variables latentes On 'a les trois variable latente comp1, comp2 et comp3 qui ont des effets positifs sur le niveau de prix des maisons.



Premièrement, comp1 est la variable latente ayant le plus grand impacte directe dans la détermination du prix des maison mais aussi indirecte a travers la comp3 enregistrant le deuxième plus grand impact sur les prix, en outre comp3 enregistre le plus faible impacte dans

la détermination du valeur immobilières, néanmoins elle reste indispensable et met en évidence la diversité des facteurs susceptible d'influencer les prix des propriétés.

En outre, en observant les intervalles de confiance on conste ils ne contiennent pas de zéro ce qui montre que la qualité des relations entre les trois variables latentes.

Fiabilité du modèle

Le tableau ci-dessous permet de mesurer la fiabilité du modèle, l'objectifs étant de savoir le minimum de la variance expliqué par chaque variable latente par rapport à chaque variable contribuant à la construction des variable latentes.

	Comp1	Comp2	Comp3	prix
Superficie	0.449	0.000	0.000	0.000
bains	0.479	0.000	0.000	0.000
Climatisation	0.379	0.000	0.000	0.000
parking	0.318	0.000	0.000	0.000
chambres	0.000	0.922	0.000	0.000
sous-sol	0.000	0.138	0.000	0.000
etages	0.000	0.000	0.757	0.000
chamb_invit	0.000	0.000	0.281	0.000
prix	0.000	0.000	0.000	1.000

On voit que la variable latente comp1 explique un pourcentage non négligeable de la variance par rapport aux variables superficie bains climatisations et parking, la composante 2 explique une part très significative du pourcentage de la variance par rapport au variables nombres de chambres, et en fin la composante 3 quant à elle explique un pourcentage aussi très significati f de la variance par rapport au variable nombres d'étages.

Fiabilité interne du modèle

L'objectif est de connaitre la fiabilité interne du modèle suivantes un certain nombre de critères

	alpha	rhoC	AVE	rhoA
Comp1	0.523	0.731	0.406	0.524
Comp2	0.177	0.654	0.530	0.334
Comp3	0.083	0.671	0.519	0.097
prix	1.000	1.000	1.000	1.000

Alpha, rhoC, and rhoA should exceed 0.7 while AVE should exceed 0.5

Le constat est que suivant les premiers critères ***alpha rhoc*** et ***rhoA*** sont inférieur à 0.7 pour la composante 2 et 3 ce qui met en évidence une certaine faiblesse en termes de fiabilité du modèle. Mais néanmoins nous avons la variance moyenne extraite supérieur à 0.5 pour la deuxième et troisième composantes en plus du ***rhoC*** qui est supérieur 0.7 par rapport à la première composante, notamment tous ces éléments témoignent d'une certaine robustesse en termes de fiabilité interne du modèle.

Validation discriminante

Consistant à savoir si la variance expliqué est supérieure à la corrélation

	Comp1	Comp2	Comp3	prix
Comp1	0.637	.	.	.
Comp2	0.351	0.728	.	.
Comp3	0.363	0.371	0.721	.
prix	0.748	0.394	0.483	1.000

L'analyse des valeurs sur la diagonale représentants les variances expliquées, met évidence des corrélations par rapport à la deuxième et troisième variables latentes inférieur aux variances indiquant une bonne validité discriminante du modèle, cela montre que les différents variables latentes exercent chacune une influence distinct et qui lui est propres sur les déterminants des valeurs des propriétés au sein de l'agglomération de New-York et par ailleurs les prix présentent des corrélations élevés avec la première composante et faiblement avec les autres. Soulignant ainsi l'existence d'un lien significatif entre le niveau des prix des logements et cette variable latente

Conclusion

Comme nous l'avons précédemment le marché de l'immobilier new-yorkais est l'images d'une ville incarnant une diversité tant culturelle qu'économique qui se reflètent dans son paysage immobilier, Mettant en évidence une grande variabilité des prix des propriétés. Cependant ce grand écart s'explique par la combinaison d'une multitude des facteurs avec des niveaux de degré d'influences plus ou moins importants.

En outre parmi ces facteurs, j'ai réussi à dégager à travers l'ACP, le modèle de régression linéaire multiples et le modèles PLS les facteurs susceptibles de jouer le rôle le plus significatifs dans la détermination du prix des maison qui sont entre autre les facteurs lieux à la structure et a l'aménagement dans la détermination du prix des logements comme la superficie, sous-sol, Parkings, nombres d'étages et chambres se révélant comme les principaux déterminants des prix des maison et les facteurs liées au confort et aux commodités comme le climatiseur chambre d'invités proximité d'une route principale etc. Cependant les prix sont plus sensibles aux caractéristiques liées à la structure et a l'aménagement et moins sensibles pour ceux liées au confort et aux commodités. Mettant évidence le niveau de compétitivité sur ce marché immobilier qui pourrait s'expliquer notamment par rareté des terrains constructibles, une demande supérieure à l'offre l'existence des propriétés avec une histoire architectural, la difficulté d'être propriétaires dans un tel emplacements très prisés créant des préférences pour des investissements de long terme.

Autrement ces ensembles d'éléments explique la hausse permanant des prix de l'immobilier au sein de la ville.

Référence bibliographique

1. Mario Fortin et André Leclerc, Déterminants du prix réel des logements au Canada, L'Actualité économique, vol. 78, n° 3, 2002, p. 293-320.

2. Achille Dargaud Fofack et Serge Djoudji Temkeng, Les déterminants des prix de l'immobilier aux Etats-Unis après la Grande Récession : une analyse des bornes extrêmes, Volume 96, numéro 3, septembre 2020, Éditeur(s) HEC Montréal ISSN 0001-771X (imprimé) 1710-3991 (numérique)

3. https://frenchdistrict.com/new-york/articles/immobilier-investir-achat-appartement-condos-new-york-agent-francais/

Annexes

Projet-datamining

Lensari Yaakoub

2024-05-09

##

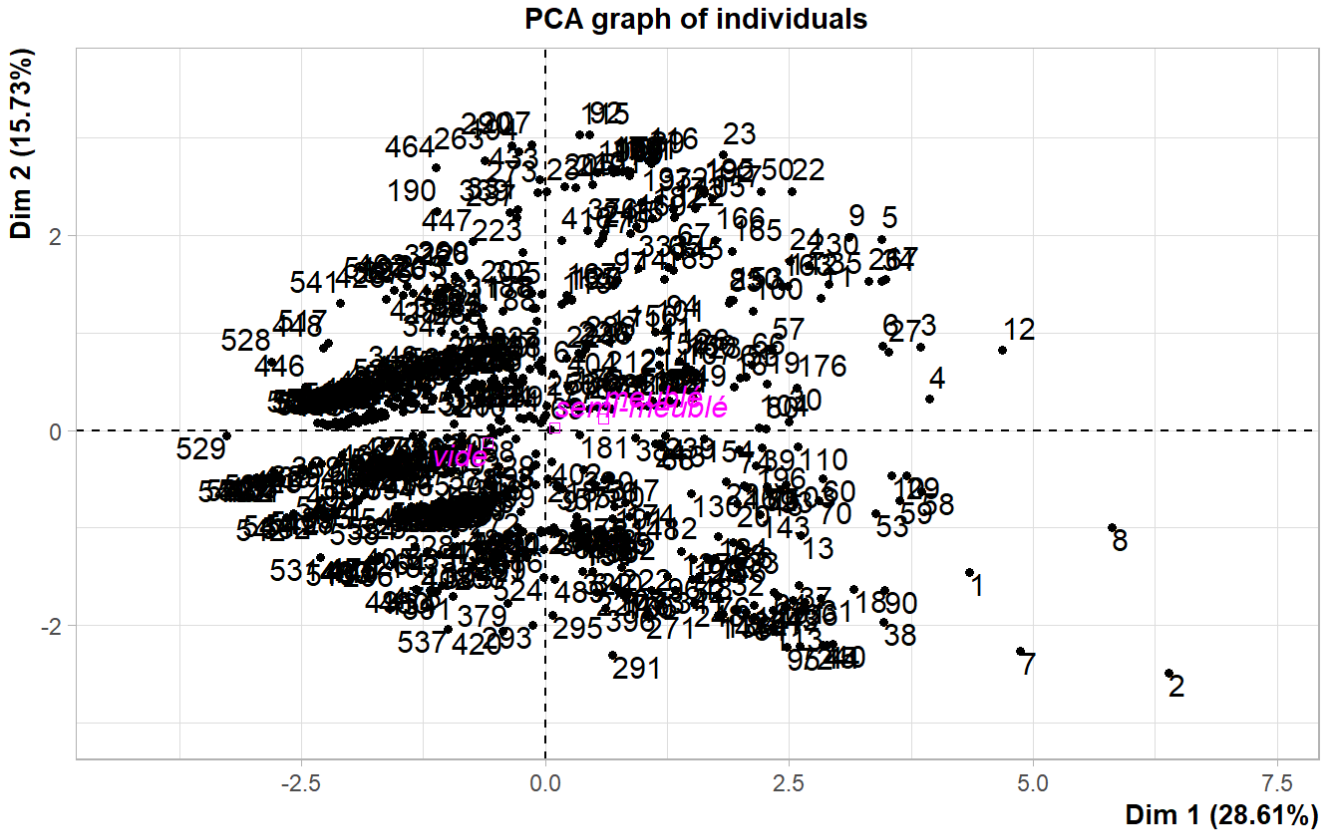
```
#################Nettoyage de Rstudio ####################################
rm(list = ls(all.names = TRUE)) # permet de nettoyer tout l'environnement avant de démarrer l'analyse
#############################################################################
####################Importer et préparer le jeu de données ###############
#############################################################################
library(readxl)
library(dplyr)
```

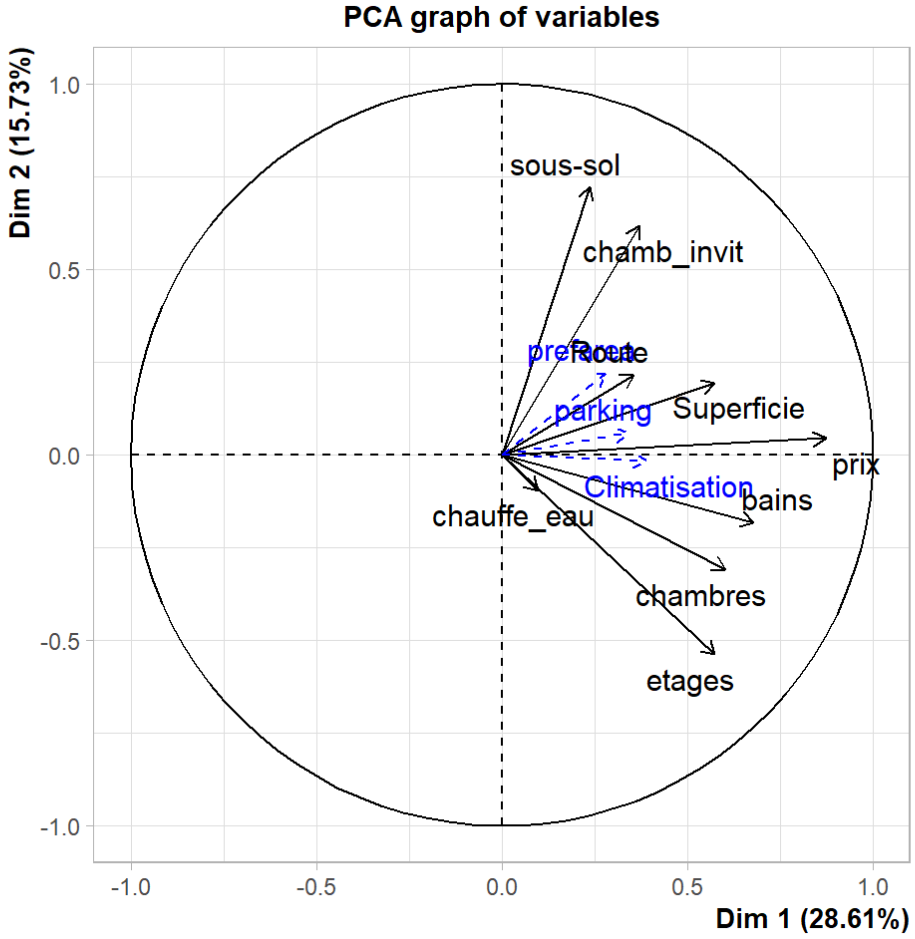
```
##
## Attachement du package : 'dplyr'
## Les objets suivants sont masqués depuis 'package:stats':
```

```
##
##     filter, lag
## Les objets suivants sont masqués depuis 'package:base':
##
##     intersect, setdiff, setequal, union
library(tidyr)
library(ggplot2)
getwd()
## [1] "C:/Users/33753/OneDrive/Documents/bases"
setwd("C:\\Users\\33753\\OneDrive\\Documents\\Modèles Office personnalisés")
DM<- read_excel("C:/Users/33753/OneDrive/Documents/Modèles Office personnalisés/base_datamining_5.xltx",
                        sheet = "F1",
                        col_types = c(  "numeric",
                                        "numeric", "numeric", "numeric",
                                        "numeric", "numeric", "numeric",
                                        "numeric", "numeric", "numeric",
                                        "numeric", "numeric", "text"



                                                                ))
DMM = DM[1:300,]
summary(DM)
##       prix             Superficie       chambres         bains
##  Min.   : 1750000   Min.   : 1650   Min.   :1.000   Min.   :1.000
##  1st Qu.: 3430000   1st Qu.: 3600   1st Qu.:2.000   1st Qu.:1.000
##  Median : 4340000   Median : 4600   Median :3.000   Median :1.000
##  Mean   : 4766729   Mean   : 5151   Mean   :2.965   Mean   :1.286
##  3rd Qu.: 5740000   3rd Qu.: 6360   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :13300000   Max.   :16200   Max.   :6.000   Max.   :4.000
##      etages           Route          chamb_invit       sous-sol
##  Min.   :1.000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:1.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :2.000   Median :1.0000   Median :0.000   Median :0.0000
##  Mean   :1.806   Mean   :0.8587   Mean   :0.178   Mean   :0.3505
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:0.000   3rd Qu.:1.0000
##  Max.   :4.000   Max.   :1.0000   Max.   :1.000   Max.   :1.0000
```

```
##   chauffe_eau      Climatisation       parking          prefarea
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.04587   Mean   :0.3156   Mean   :0.6936   Mean   :0.2349
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :3.0000   Max.   :1.0000
##   meublement
##  Length:545
##  Class :character
##  Mode  :character
##
##
##
```

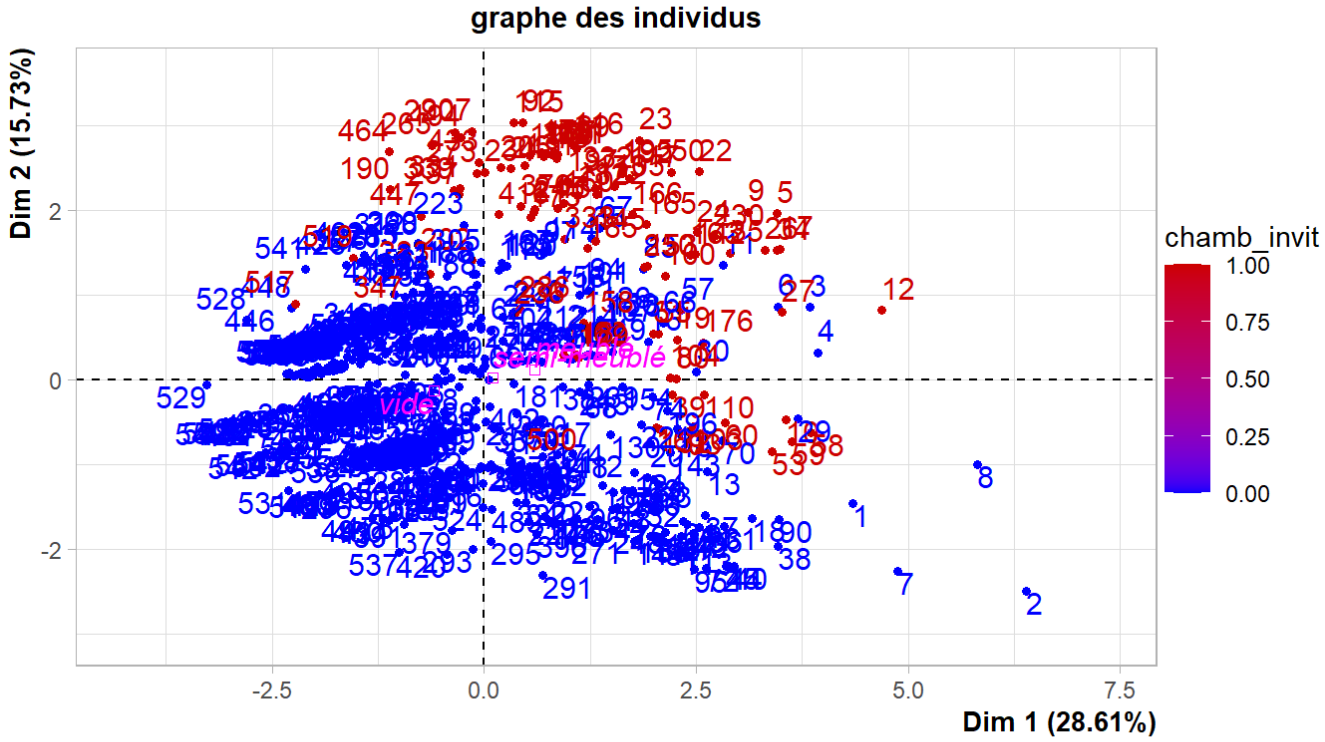
```
library(FactoMineR)
res.dm<-PCA(DM ,
         scale.unit=TRUE,
         ncp=5, quanti.sup=(10:12),
         quali.sup=c(13:13),
         graph = TRUE)
```


PCA graph of individuals
Dim 2 (15.73%)
Dim 1 (28.61%)
meublé
semi-meublé
vide
2
0
-2
-2.5
0.0
2.5
5.0
7.5



le graphique ci-dessous represente les variables ainsi que leurs positionnement par rappport a l'origine déterminant ainsi leurs cotribution dans la contructions des dimensions de l'ACP.

```
print(plot.PCA(res.dm, axes=c(1, 2),
               choix="ind",
               habillage = "chamb_invit",
               col.ind="black",
               col.ind.sup="blue",
               col.quali="magenta",
               label=c("ind", "ind.sup", "quali"),
               new.plot=TRUE,
               title="graphe des individus"))
```



En premier lieu, Nous remarquons la présence d'un effet de taille

tableaux de corrélation

```
correlation_matrix <- cor(DM[-13])
print(correlation_matrix)
```

```
##                      prix   Superficie    chambres       bains       etages
## prix           1.00000000  0.535997346  0.36649403 0.51754534  0.42071237
## Superficie     0.53599735  1.000000000  0.15185849 0.19381953  0.08399605
## chambres       0.36649403  0.151858486  1.00000000 0.37393024  0.40856424
## bains          0.51754534  0.193819531  0.37393024 1.00000000  0.32616471
## etages         0.42071237  0.083996051  0.40856424 0.32616471  1.00000000
## Route          0.29689849  0.288874114 -0.01203324 0.04239762  0.12170613
## chamb_invit    0.25551729  0.140296590  0.08054870 0.12646884  0.04353767
## sous-sol       0.18705660  0.047416989  0.09731242 0.10210571 -0.17239362
## chauffe_eau    0.09307284 -0.009229236  0.04604889 0.06715910  0.01884651
## Climatisation 0.45295408  0.222393104  0.16060326 0.18691503  0.29360200
## parking        0.38439365  0.352980481  0.13926990 0.17749582  0.04554709
```

```
## prefarea      0.32977705  0.234778798  0.07902306 0.06347174  0.04442487
##                    Route chamb_invit     sous-sol  chauffe_eau Climatisation
## prix           0.29689849  0.25551729  0.187056598  0.093072844    0.45295408
## Superficie     0.28887411  0.14029659  0.047416989 -0.009229236    0.22239310
## chambres      -0.01203324  0.08054870  0.097312424  0.046048887    0.16060326
## bains          0.04239762  0.12646884  0.102105706  0.067159096    0.18691503
## etages         0.12170613  0.04353767 -0.172393617  0.018846511    0.29360200
## Route          1.00000000  0.09233692  0.044002081 -0.011781490    0.10542300
## chamb_invit    0.09233692  1.00000000  0.372065708 -0.010307884    0.13817877
## sous-sol       0.04400208  0.37206571  1.000000000  0.004384836    0.04734119
## chauffe_eau   -0.01178149 -0.01030788  0.004384836  1.000000000   -0.13002283
## Climatisation  0.10542300  0.13817877  0.047341189 -0.130022833    1.00000000
## parking        0.20443255  0.03746575  0.051497175  0.067863888    0.15917268
## prefarea       0.19987578  0.16089694  0.228082853 -0.059411382    0.11738210
##                  parking    prefarea
## prix          0.38439365  0.32977705
## Superficie    0.35298048  0.23477880
## chambres      0.13926990  0.07902306
## bains         0.17749582  0.06347174
## etages        0.04554709  0.04442487
## Route         0.20443255  0.19987578
## chamb_invit   0.03746575  0.16089694
## sous-sol      0.05149718  0.22808285
## chauffe_eau   0.06786389 -0.05941138
## Climatisation 0.15917268  0.11738210
## parking       1.00000000  0.09162706
## prefarea      0.09162706  1.00000000
```

```
summary(DM$Superficie)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1650    3600    4600    5151    6360   16200
```

LES MCO

```
MCO = lm(prix ~ Superficie + chambres + bains + etages + Route +  chamb_invit + Cli
matisation + parking + prefarea + DM$`sous-sol` ,data = DM)

summary(MCO)
```

```
##
```

```
## Call:
## lm(formula = prix ~ Superficie + chambres + bains + etages +
##     Route + chamb_invit + Climatisation + parking + prefarea +
##     DM$`sous-sol`, data = DM)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2974326  -659446   -44918   484066  5212980
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -288470.78  242100.77  -1.192 0.233974
## Superficie        247.44      24.87   9.948  < 2e-16 ***
## chambres       134609.34   74331.30   1.811 0.070712 .
## bains         1029859.01  105696.21   9.744  < 2e-16 ***
## etages         465640.92   65738.59   7.083 4.47e-12 ***
## Route          463043.46  145252.97   3.188 0.001517 **
## chamb_invit    317881.82  134936.88   2.356 0.018844 *
## Climatisation  814490.67  109340.05   7.449 3.80e-13 ***
## parking        317238.01   59538.71   5.328 1.46e-07 ***
## prefarea       632426.16  118384.80   5.342 1.36e-07 ***
## DM$`sous-sol`  389377.61  112749.57   3.453 0.000597 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1095000 on 534 degrees of freedom
## Multiple R-squared:  0.6635, Adjusted R-squared:  0.6572
## F-statistic: 105.3 on 10 and 534 DF,  p-value: < 2.2e-16
```

Régression robuste

```
library(MASS)
```

```
##
## Attachement du package : 'MASS'
## L'objet suivant est masqué depuis 'package:dplyr':
##
##     select
```

```
robust_model <- rlm(prix ~ Superficie + chambres + bains + etages + Route +  chamb_
invit + Climatisation + parking + prefarea + DM$`sous-sol` ,data = DM)

summary(robust_model)
```

```
##
## Call: rlm(formula = prix ~ Superficie + chambres + bains + etages +
##     Route + chamb_invit + Climatisation + parking + prefarea +
##     DM$`sous-sol`, data = DM)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2761518  -564872    13552   560908  5534169
##
## Coefficients:
##                Value       Std. Error  t value
## (Intercept)    -8213.2783 209158.7918     -0.0393
## Superficie       244.5038     21.4895     11.3778
## chambres       62864.8108  64217.2452      0.9789
## bains         996072.5285  91314.4190     10.9082
## etages        498387.4531  56793.7223      8.7754
## Route         399131.2284 125488.8000      3.1806
## chamb_invit   368595.5476 116576.3916      3.1618
## Climatisation 735386.9186  94462.4542      7.7850
## parking       242107.9600  51437.4387      4.7068
## prefarea      602325.5652 102276.5100      5.8892
## DM$`sous-sol` 344931.6611  97408.0475      3.5411
##
## Residual standard error: 837500 on 534 degrees of freedom
```

Modèle GLM avec une distribution gaussienne

```
glm_model <- glm(prix ~ Superficie + bains + etages + Route +  chamb_invit + Climat
isation + parking + prefarea + DM$`sous-sol` ,data = DM, family = gaussian)

summary(glm_model)
```

```
##
## Call:
## glm(formula = prix ~ Superficie + bains + etages + Route + chamb_invit +
##     Climatisation + parking + prefarea + DM$`sous-sol`, family = gaussian,
```

```
##     data = DM)
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -28228.49  195251.53  -0.145 0.885101
## Superficie        251.07      24.85  10.105  < 2e-16 ***
## bains         1073318.79  103154.77  10.405  < 2e-16 ***
## etages         507787.45   61611.85   8.242 1.32e-15 ***
## Route          431562.62  144515.88   2.986 0.002953 **
## chamb_invit    312654.16  135193.09   2.313 0.021121 *
## Climatisation  811934.76  109563.61   7.411 4.94e-13 ***
## parking        325519.37   59489.16   5.472 6.85e-08 ***
## prefarea       636729.40  118612.84   5.368 1.19e-07 ***
## DM$`sous-sol`  418415.25  111841.08   3.741 0.000203 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.204547e+12)
##
##     Null deviance: 1.9032e+15  on 544  degrees of freedom
## Residual deviance: 6.4443e+14  on 535  degrees of freedom
## AIC: 16719
##
## Number of Fisher Scoring iterations: 2
```

Test de normalité des résidus

```
library(car)
## Le chargement a nécessité le package : carData
##
## Attachement du package : 'car'
## L'objet suivant est masqué depuis 'package:dplyr':
##
##     recode
residuals <- resid(MCO)  # Remplacez "votre_modele" par le nom de votre modèle
shapiro.test(residuals)
##
```

```
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.9528, p-value = 3.454e-12
```

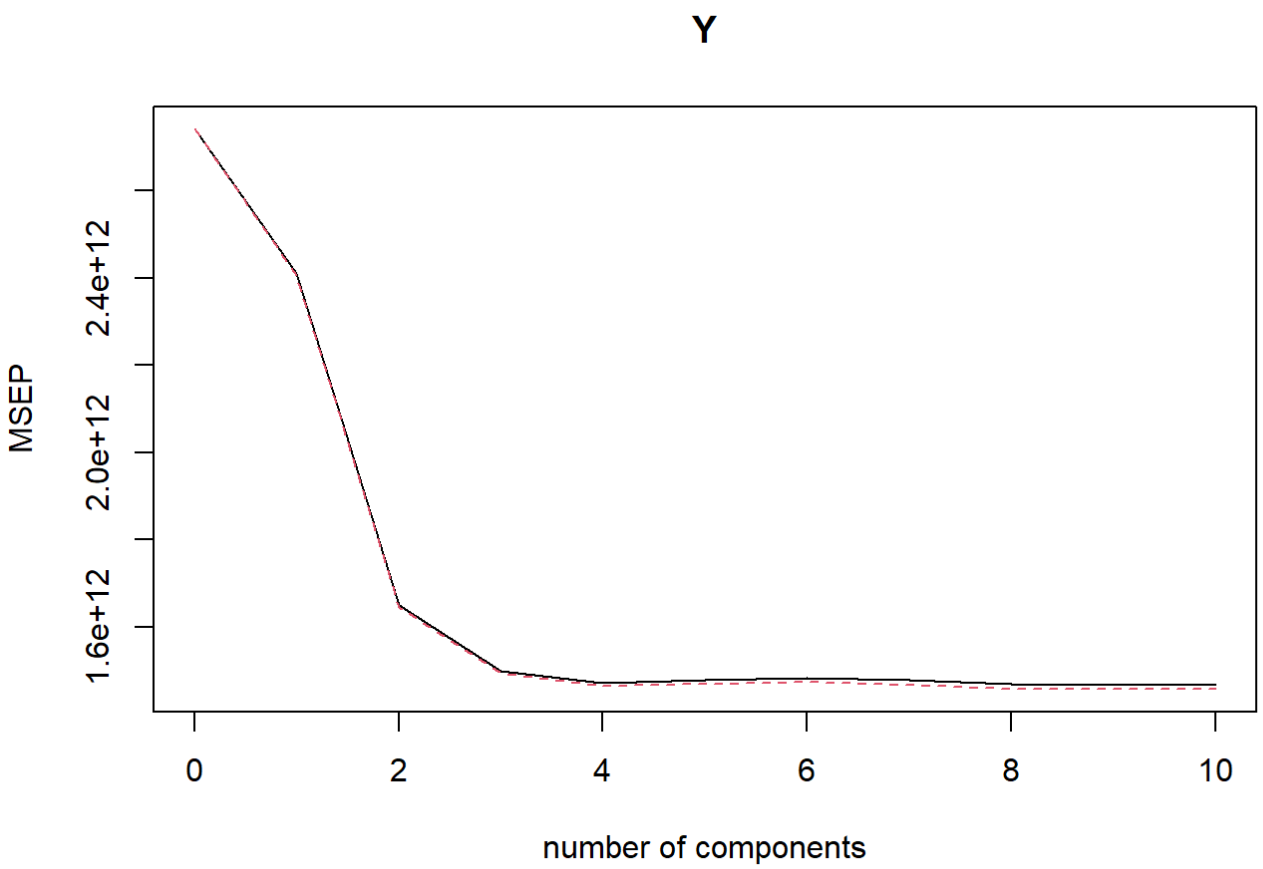
Test d'hétéroscédasticité (Test de Breusch-Pagan)

```
library(lmtest)
## Le chargement a nécessité le package : zoo
##
## Attachement du package : 'zoo'
## Les objets suivants sont masqués depuis 'package:base':
##
##     as.Date, as.Date.numeric
bptest(glm_model)
##
##  studentized Breusch-Pagan test
##
## data:  glm_model
## BP = 58.779, df = 9, p-value = 2.303e-09
```

la valeur vif

```
library(car)
vif(MCO)
##    Superficie      chambres         bains        etages         Route
##      1.321576      1.365064      1.279264      1.475007      1.163089
##   chamb_invit Climatisation       parking      prefarea DM$`sous-sol`
##      1.210424      1.173336      1.193487      1.144364      1.314900
library(pls)
##
## Attachement du package : 'pls'
## L'objet suivant est masqué depuis 'package:stats':
##
##     loadings
```

```
Y <- as.matrix(DMM[,1])
X <- as.matrix(DMM[,2:12])
# Spécifiez le nombre maximum de composantes à tester
n_max_components <- 10

# Créez le modèle PLS avec validation croisée
pls_model_cv <- plsr(Y ~ X, data = DMM, ncomp = n_max_components , validation = "CV
")

# Affichez le plot de validation croisée
validationplot(pls_model_cv, val.type = "MSEP")
```




```
data1 <- plsr(Y ~ X, ncomp=3, scale=TRUE, validation="LOO", data=DM)
summary(data1)
## Data:    X dimension: 300 11
##  Y dimension: 300 1
## Fit method: kernelpls
## Number of components considered: 3
```

```
##
## VALIDATION: RMSEP
## Cross-validated using 300 leave-one-out segments.
##        (Intercept)  1 comps  2 comps  3 comps
## CV         1655667  1193608  1184449  1185699
## adjCV      1655667  1193552  1184382  1185610
##
## TRAINING: % variance explained
##    1 comps  2 comps  3 comps
## X    15.31    27.95    37.41
## Y    50.46    52.31    52.79
```

```
data3 <- mvr(Y ~ X, ncomp=5, scale=TRUE, validation="LOO", data=DM)
summary(data3)
```

```
## Data:    X dimension: 300 11
##  Y dimension: 300 1
## Fit method: kernelpls
## Number of components considered: 5
##
## VALIDATION: RMSEP
## Cross-validated using 300 leave-one-out segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
## CV         1655667  1193608  1184449  1185699  1185330  1185016
## adjCV      1655667  1193552  1184382  1185610  1185243  1184930
##
## TRAINING: % variance explained
##    1 comps  2 comps  3 comps  4 comps  5 comps
## X    15.31    27.95    37.41    48.79    55.21
## Y    50.46    52.31    52.79    52.83    52.84
```

chargements pour X

```
loadings(data1)
```

```
##
## Loadings:
##               Comp 1 Comp 2 Comp 3
## Superficie     0.409        -0.338
## chambres       0.375 -0.270 -0.239
```

```
## bains           0.494          0.253
## etages          0.369 -0.362   0.510
## Route           0.226         -0.366
## chamb_invit             0.433 -0.394
## sous-sol       -0.103   0.755 -0.191
## chauffe_eau             0.303  0.323
## Climatisation   0.339          0.234
## parking         0.359         -0.158
## prefarea        0.165   0.392 -0.292
##
##                 Comp 1 Comp 2 Comp 3
## SS loadings      1.026  1.216  1.091
## Proportion Var   0.093  0.111  0.099
## Cumulative Var   0.093  0.204  0.303
```

#poids pour X

```
loading.weights(data1)
```

```
##
## Loadings:
##                Comp 1 Comp 2 Comp 3
## Superficie      0.405         -0.210
## chambres        0.305 -0.437 -0.358
## bains           0.510  0.102  0.162
## etages          0.354          0.570
## Route           0.200 -0.161 -0.284
## chamb_invit             0.202 -0.498
## sous-sol                0.690 -0.140
## chauffe_eau     0.121  0.395  0.198
## Climatisation   0.353          0.174
## parking         0.359
## prefarea        0.211  0.287 -0.225
##
##                 Comp 1 Comp 2 Comp 3
## SS loadings      1.000  1.000  1.000
## Proportion Var   0.091  0.091  0.091
## Cumulative Var   0.091  0.182  0.273
```

#poids pour Y

```
Yloadings(data1)
```

```
##
## Loadings:
##   Comp 1   Comp 2   Comp 3
## Y 916214.1 210433.4 117192.4
##
##                      Comp 1       Comp 2       Comp 3
## SS loadings    839448321514  44282231756  13734058450
## Proportion Var 839448321514  44282231756  13734058450
## Cumulative Var 839448321514 883730553270 897464611720
```

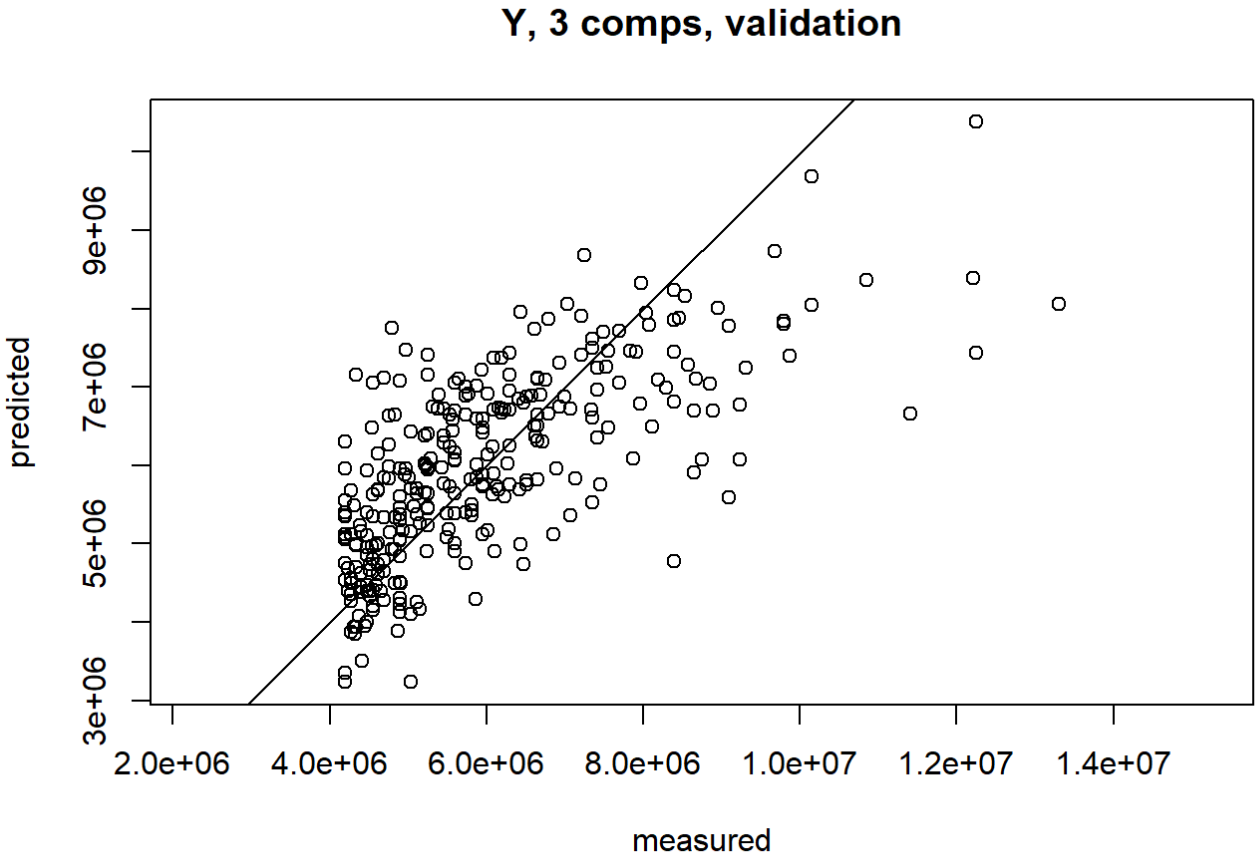
#coefficients de régression

```
coef(data1)
```

```
## , , 3 comps
##
##                        Y
## Superficie    356390.90
## chambres      134331.99
## bains         534861.68
## etages        380314.95
## Route         115514.05
## chamb_invit    66303.22
## sous-sol      173269.96
## chauffe_eau   243399.48
## Climatisation 380963.66
## parking       333823.68
## prefarea      251813.77
```

/ #graphique des prédictions

```
plot(data1, ncomp=3, asp=1, line=TRUE)
```



`

Modéle de mesure

```
library(seminr)
## Warning: le package 'seminr' a été compilé avec la version R 4.3.3
mm = constructs(
  composite("Comp1",multi_items("",c("Superficie","bains","Climatisation","parking"
))),
  composite("Comp2",multi_items("", c("chambres","sous-sol"))),
  composite("Comp3",multi_items("",c("etages", "chamb_invit"))),
  composite("prix",single_item("prix"))
  )
ms = relationships(
  paths("Comp1", "Comp2"),  # Relation entre Comp1 et Comp2
  paths("Comp1", "Comp3"),
  paths("Comp2", "Comp3"),  # Relation entre Comp2 et Comp3
  paths("Comp2", "prix"),
```

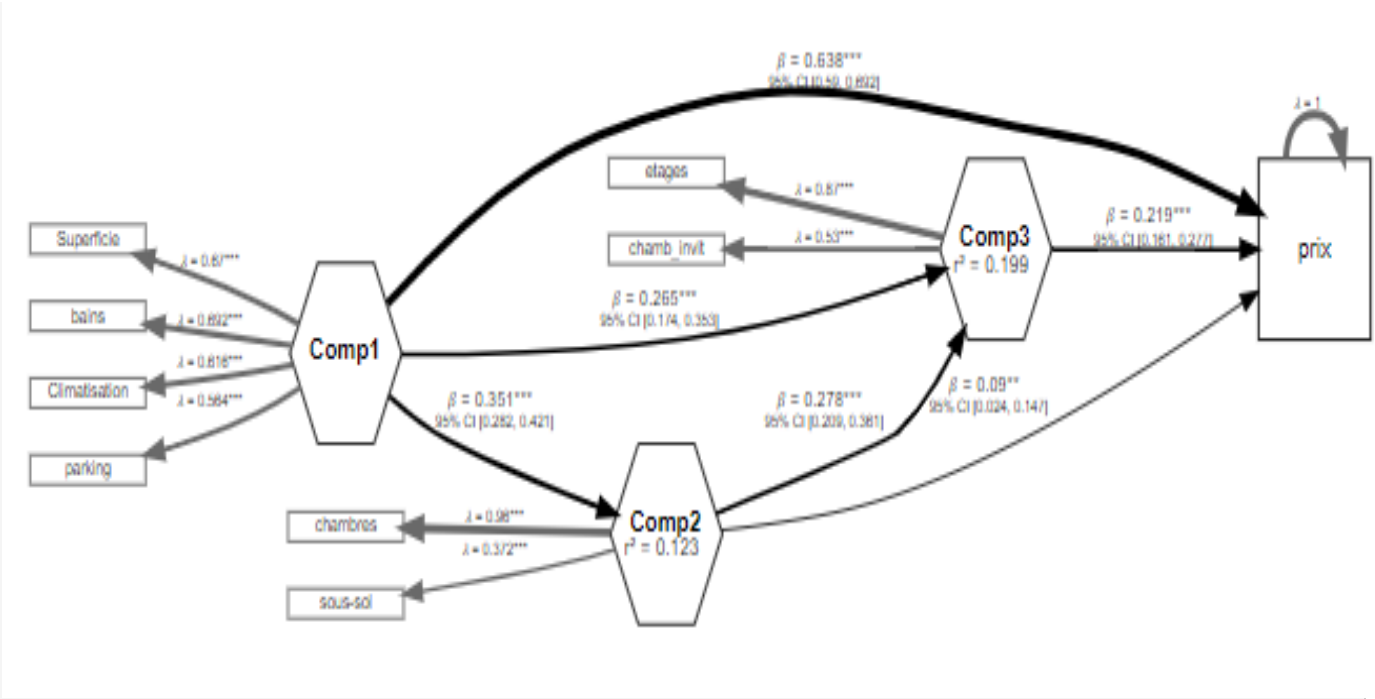
```
  paths("Comp3", "prix"),     # Relation entre Comp3 et la variable de réponse (prix
)
  paths("Comp1", "prix")      # Relation directe entre Comp1 et la variable de répon
se (prix)
)
```

estimation du modèle

```
satPLS.mod = estimate_pls(DM,mm,ms)
## Generating the seminr model
## All 545 observations are valid.
statpls.fit = summary(satPLS.mod)
plot(satPLS.mod)
```

bootstrop du modèle

```
boot.statpls = bootstrap_model(satPLS.mod,nboot =1000)
## Bootstrapping model using seminr...
## SEMinR Model successfully bootstrapped
sum.boot.statpls = summary(boot.statpls,alpha = 0.05)
                                plot(boot.statpls)
```



assement modèle de mesure

```
statpls.fit$loadings**2 # fiabilité de l'indicateur
```

```
##                Comp1 Comp2 Comp3  prix
## Superficie     0.449 0.000 0.000 0.000
## bains          0.479 0.000 0.000 0.000
## Climatisation 0.379 0.000 0.000 0.000
## parking        0.318 0.000 0.000 0.000
## chambres       0.000 0.922 0.000 0.000
## sous-sol       0.000 0.138 0.000 0.000
## etages         0.000 0.000 0.757 0.000
## chamb_invit    0.000 0.000 0.281 0.000
## prix           0.000 0.000 0.000 1.000
```

```
statpls.fit$reliability # fiabilité interne du modèle
```

```
##        alpha  rhoC   AVE  rhoA
## Comp1 0.523 0.731 0.406 0.524
## Comp2 0.177 0.654 0.530 0.334
## Comp3 0.083 0.671 0.519 0.097
## prix  1.000 1.000 1.000 1.000
##
## Alpha, rhoC, and rhoA should exceed 0.7 while AVE should exceed 0.5
```

```
statpls.fit$validity$fl_criteria
```

```
##       Comp1 Comp2 Comp3  prix
## Comp1 0.637     .     .     .
## Comp2 0.351 0.728     .     .
## Comp3 0.363 0.371 0.721     .
## prix  0.748 0.394 0.483 1.000
##
## FL Criteria table reports square root of AVE on the diagonal and construct corre
lations on the lower triangle.
```