Rapport sur la base de données : Boston Housing Dataset

Master 1 Data Sciences – Data Mining

RICHARD Paul-Antoine – TERNAT Clément

Année 2022-2023

0. INTRODUCTION

L'objectif de ce rapport est d'analyser une base de données à l'aide de plusieurs méthodes d'analyse statistiques. La base de données que nous avons sélectionnée est une base de données qui porte sur le prix de l'immobilier dans la ville de Boston. Ci-dessous le lien pour la télécharger : https://www.kaggle.com/datasets/altavish/boston-housing-dataset

Cette base de données contient donc logiquement un grand nombre d'informations sur les habitants et habitations de la région qui proviennent d'un recensement réel de la ville de Boston. Pour étudier le modèle nous pourrons nous aider de 14 variables :

- **CRIM** : Taux de criminalité par quartier.
- **ZN** : Proportion de terrains résidentiels pour des lots de plus de 25000m².
- **INDUS** : Proportion d'acres de terrains non-commerciaux par quartier.
- **CHAS** : Variable binaire prenant la valeur 1 si le secteur est bordé par le fleuve Charles, 0 sinon.
- **NOX** : Concentration d'oxyde d'azote dans l'air (parties par millions).
- **RM** : Nombre moyen de pièces par logement.
- **AGE** : Proportion d'unités occupées par leur propriétaire construites avant 1940.
- **DIS** : Distances pondérées vers cinq centres d'emploi de Boston.
- **RAD** : Indice d'accessibilité aux autoroutes.
- **TAX** : Taux d'imposition foncière de pleine valeur pour 10000dollars
- **PTRATIO** : Ratio élèves-enseignants par ville.
- **B** : $1000(Bk-0.63)^2$, avec Bk la proportion de personnes d'origine afro-américaine par quartier.
- **LSTAT** : Pourcentage de la population à statut socio-économique inférieur.
- **MEDV** : Valeur médiane des logements occupés par leur propriétaire en milliers de dollars.

Tout au long du rapport nous allons analyser les données afin de répondre à deux problématiques :

- Problématique 1 : Quelle est l'influence des caractéristiques socio-économiques sur la valeur médiane des logements ?
- Problématique 2 : Existe-t-il une corrélation entre la concentration de crime et les autres caractéristiques des quartiers ?

Pour répondre à cette première problématique nous allons procéder à une régression linéaire et une Analyse en Composantes Principales en fonction de la variable cible MEDV qui correspond à la valeur médiane des logements. Pour répondre à la seconde problématique nous utiliserons aussi l'ACP et une régression linéaire multiple (cette fois en fonction de la variable CRIM), puis nous compléterons l'analyse avec une régression PLS.

1. ANALYSE DES DONNEES

Pour commencer, nous allons regarder quelques statistiques descriptives :

```
      CRIM                ZN              INDUS            CHAS              NOX              RM              AGE              DIS        
 Min.   : 0.00632   Min.   :  0.0   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
 1st Qu.: 0.08221   1st Qu.:  0.0   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.880   1st Qu.: 45.08   1st Qu.: 2.101  
 Median : 0.26169   Median :  0.0   Median : 9.69   Median :0.00000   Median :0.5380   Median :6.202   Median : 77.15   Median : 3.183  
 Mean   : 3.70752   Mean   : 11.3   Mean   :11.20   Mean   :0.06876   Mean   :0.5552   Mean   :6.280   Mean   : 68.58   Mean   : 3.788  
 3rd Qu.: 3.69311   3rd Qu.: 12.5   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.619   3rd Qu.: 94.10   3rd Qu.: 5.118  
 Max.   :88.97620   Max.   :100.0   Max.   :27.74   Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
                                    NA's   :3                         NA's   :2                        NA's   :1                        
      RAD              TAX            PTRATIO             B              LSTAT            MEDV      
 Min.   : 1.00   Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.730   Min.   : 5.0  
 1st Qu.: 4.00   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.33   1st Qu.: 7.093   1st Qu.:17.0  
 Median : 5.00   Median :330.0   Median :19.10   Median :391.45   Median :11.430   Median :21.2  
 Mean   : 9.61   Mean   :409.2   Mean   :18.46   Mean   :356.66   Mean   :12.705   Mean   :22.5  
 3rd Qu.:24.00   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.24   3rd Qu.:16.992   3rd Qu.:25.0  
 Max.   :24.00   Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.970   Max.   :50.0  
 NA's   :1                                                        NA's   :1                      
```

On peut tout d'abord constater que la valeur minimale du taux de criminalité par quartier est proche de zéro, ce qui suggère que certains quartiers ont un taux de criminalité très bas, cependant la moyenne est elle plus élevée ce qui suggère une forte variation de présence de la criminalité dans la ville.

On observe pour la variable binaire CHAS que la moyenne est égale 0.0688, ce qui signifie que seulement 7% des habitations sont bordées par le fleuve.

D'après la variable PTRATIO, il y aurait en moyenne 18 élèves par professeurs.

La médiane de la variable LSTAT est de 11%, ce qui signifie qu'une grande partie de la population semble vivre sous un niveau de statut socio-économique inférieur.

Enfin, la médiane de MEDV est de 21.2 milliers de dollars, il s'agit de la valeur médiane des habitations occupées par leurs propriétaires.

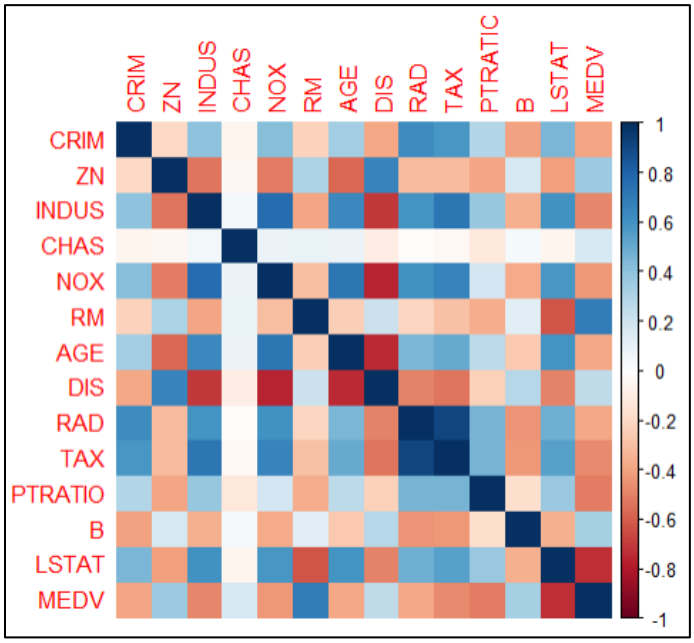
Nous allons représenter graphiquement la corrélation entre les 14 variables quantitatives mais pour cela nous devons « laver » les variables, c'est-à-dire supprimer les lignes dans lesquelles il pourrait manquer des informations. D'après le '*summary*' exprimé ci-dessus on peut voir que nous avons 409 observations et que nous avons 8 informations manquantes au total. Ce manque d'informations concerne donc 2% de données ce qui est minime, nous allons donc pouvoir supprimer ces lignes.

Notre matrice de corrélation est donc la suivante :

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS
CRIM	1.00000000	-0.20331596	0.40875045	-0.05965646	0.42059351	-0.22269113	0.34930434	-0.38123988
ZN	-0.20331596	1.00000000	-0.53491821	-0.04111069	-0.51861135	0.31580560	-0.57411890	0.66809978
INDUS	0.40875045	-0.53491821	1.00000000	0.05952956	0.76225204	-0.39781993	0.64418315	-0.70802919
CHAS	-0.05965646	-0.04111069	0.05952956	1.00000000	0.08806774	0.09157728	0.08290254	-0.09702062
NOX	0.42059351	-0.51861135	0.76225204	0.08806774	1.00000000	-0.30907113	0.72939573	-0.76881060
RM	-0.22269113	0.31580560	-0.39781993	0.09157728	-0.30907113	1.00000000	-0.24844058	0.21084597
AGE	0.34930434	-0.57411890	0.64418315	0.08290254	0.72939573	-0.24844058	1.00000000	-0.74816208
DIS	-0.38123988	0.66809978	-0.70802919	-0.09702062	-0.76881060	0.21084597	-0.74816208	1.00000000
RAD	0.62912176	-0.31490980	0.59529934	-0.01164625	0.60773094	-0.21843035	0.45138612	-0.49426212
TAX	0.58690139	-0.31775621	0.72109409	-0.03951085	0.66502550	-0.29977895	0.50123287	-0.53342714
PTRATIO	0.29393239	-0.39334903	0.38569493	-0.12307546	0.18920131	-0.36121647	0.26804272	-0.23463259
B	-0.40072712	0.17396481	-0.35359530	0.04974673	-0.37722908	0.12687466	-0.26963306	0.28827366
LSTAT	0.45134944	-0.41391466	0.60232189	-0.05698174	0.58827091	-0.62075050	0.59954324	-0.49511471
MEDV	-0.39404261	0.36226616	-0.48607569	0.17643636	-0.43026854	0.69518466	-0.38034089	0.25170580

	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM	0.62912176	0.58690139	0.2939324	-0.40072712	0.45134944	-0.3940426
ZN	-0.31490980	-0.31775621	-0.3933490	0.17396481	-0.41391466	0.3622662
INDUS	0.59529934	0.72109409	0.3856949	-0.35359530	0.60232189	-0.4860757
CHAS	-0.01164625	-0.03951085	-0.1230755	0.04974673	-0.05698174	0.1764364
NOX	0.60773094	0.66502550	0.1892013	-0.37722908	0.58827091	-0.4302685
RM	-0.21843035	-0.29977895	-0.3612165	0.12687466	-0.62075050	0.6951847
AGE	0.45138612	0.50123287	0.2680427	-0.26963306	0.59954324	-0.3803409
DIS	-0.49426212	-0.53342714	-0.2346326	0.28827366	-0.49511471	0.2517058
RAD	1.00000000	0.91117190	0.4689268	-0.44005859	0.48581208	-0.3848283
TAX	0.91117190	1.00000000	0.4663984	-0.43705312	0.54104228	-0.4709928
PTRATIO	0.46892675	0.46639836	1.0000000	-0.17726137	0.37966014	-0.5127405
B	-0.44005859	-0.43705312	-0.1772614	1.00000000	-0.35538409	0.3331801
LSTAT	0.48581208	0.54104228	0.3796601	-0.35538409	1.00000000	-0.7374946
MEDV	-0.38482826	-0.47099276	-0.5127405	0.33318007	-0.73749463	1.0000000

Pour qu'elle soit plus lisible nous allons directement étudier sa représentation graphique :



Sur ce graphique, plus la couleur tend vers un bleu foncé (respectivement rouge) et plus les corrélations sont élevées positivement (rspct négativement).

La variable CRIM a une corrélation positive avec les variables INDUS, NOX, AGE et LSTAT et une corrélation négative avec ZN, RM et DIS. Autrement dit le taux de criminalité pourrait être en lien avec des logements plus anciens et des statuts sociaux inférieurs, ainsi qu'avec la distance entre les logements et des centres d'emplois.

On remarque aussi que la variable MEDV a une corrélation négative avec CRIM, INDUS, NOX, PRATIO et LSTAT, et une corrélation positive avec RM. Autrement dit des logements avec plus de pièces couteraient donc plus cher, et les habitations se situant dans les quartiers avec une forte criminalité auront tendance à couter moins cher.

Nous avons donc déjà plusieurs hypothèses pour répondre à nos problématiques. Afin de les confirmer nous devons réaliser différents tests pour compléter les recherches.

2. REGRESSION LINEAIRE

2.1 En fonction de MEDV

Notre objectif par cette première régression va être d'obtenir les premiers éléments pour répondre à notre première problématique, c'est-à-dire « Quelle est l'influence des caractéristiques socio-économiques sur la valeur médiane des logements ? ». Nous avons préalablement lavé nos données.

On réalise alors une première régression linéaire en fonction de la variable MEDV pour répondre à notre première problématique :

```
Call:
lm(formula = MEDV ~ ., data = housing_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3943  -2.7366  -0.5257   1.7465  26.1385

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.737e+01  5.152e+00   7.253 1.62e-12 ***
CRIM        -1.145e-01  3.265e-02  -3.506 0.000497 ***
ZN           4.773e-02  1.407e-02   3.393 0.000749 ***
INDUS        1.644e-02  6.190e-02   0.266 0.790697
CHAS         2.671e+00  8.649e-01   3.088 0.002130 **
NOX         -1.830e+01  3.847e+00  -4.758 2.58e-06 ***
RM           3.753e+00  4.211e-01   8.913  < 2e-16 ***
AGE         -8.412e-05  1.341e-02  -0.006 0.994999
DIS         -1.504e+00  2.016e-01  -7.460 4.00e-13 ***
RAD          3.072e-01  6.669e-02   4.607 5.22e-06 ***
TAX         -1.204e-02  3.796e-03  -3.173 0.001605 **
PTRATIO     -9.677e-01  1.322e-01  -7.318 1.05e-12 ***
B            9.451e-03  2.682e-03   3.524 0.000466 ***
LSTAT       -5.171e-01  5.079e-02 -10.180  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.763 on 487 degrees of freedom
Multiple R-squared:  0.7409,    Adjusted R-squared:  0.734
F-statistic: 107.1 on 13 and 487 DF,  p-value: < 2.2e-16
```

La valeur du R^2 ajusté est 0.734 ce qui signifie que les variables indépendantes du modèle expliquent environ 73.4% de la variance de la variable MEDV (valeur médiane des logements occupés par leur propriétaire en milliers de dollars). Le R^2 est donc relativement élevé ce qui indique que le modèle ajuste bien les données.

La valeur de la F-statistic est de 107.1. Cette statistique est utile pour tester l'hypothèse nulle comme quoi tous les coefficients de lé régression sont nuls. Dans ce cas, la valeur de la statistique est élevée, ce qui suggère que le modèle de régression dans son ensemble est significatif. En effet, on constate que les p-values associées aux coefficients des variables sont quasiment toutes significatives, les variables INDUS et AGE étant les deux seules exceptions.

Nous pouvons aussi déduire de la sortie R que les valeurs qui pour une unité semblent le plus augmenter le prix d'une habitation sont les variables B, ZN et RAD, et celles qui semblent le réduire sont les variables LSTAT et PTRATIO.

Nous allons donc maintenant effectuer des tests pour valider notre modèle :

- Test de Shapiro :

```
        Shapiro-Wilk normality test

data:  residualsMEDV
W = 0.90268, p-value < 2.2e-16
```

L'hypothèse nulle de ce test est : « Les résidus suivent une loi Normale ». Nous avons obtenu une p-value très faible inférieure à 5%. Nous rejetons alors l'hypothèse nulle, nos résidus ne suivent pas une loi Normale.

- Test de Brush-Pagan :

```
        studentized Breusch-Pagan test

data:  modelMEDV
BP = 63.932, df = 13, p-value = 1.03e-08
```

Il s'agit du test d'homoscédasticité. L'hypothèse nulle est la suivante « La variance des résidus est constante ». Notre p-value très faible suggère donc la présence d'hétéroscédasticité dans les résidus.

- Test de Durbin-Watson :

```
        Durbin-Watson test

data:  modelMEDV
DW = 1.0785, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Il s'agit d'un test portant sur l'autocorrélation des résidus. Dans notre cas, la p-value est encore une fois très faible, ce qui suggère un rejet de l'hypothèse nulle « Aucune corrélation entre les résidus ».

- Test VIF :

```
     CRIM         ZN      INDUS       CHAS        NOX         RM        AGE        DIS
RAD        TAX    PTRATIO          B      LSTAT
 1.881589  2.357765  4.190391  1.104009  4.487682  2.136551  3.196781  3.970938
9.670730 12.062595  1.867524  1.318706  3.368033
```

Le test VIF est utilisé pour évaluer la multicolinéarité entre les variables indépendantes. Les résultats d'un test VIF supérieur à 5 est considéré comme préoccupant. On constate donc dans notre exemple qu'il y a une certaine corrélation entre les variables RAD et TAX. Autrement dit ces deux variables contribuent de manière redondante à l'explication de MEDV. Il pourrait donc être intéressant de supprimer l'une des deux variables et de comparer les résultats avec nos résultats actuels.

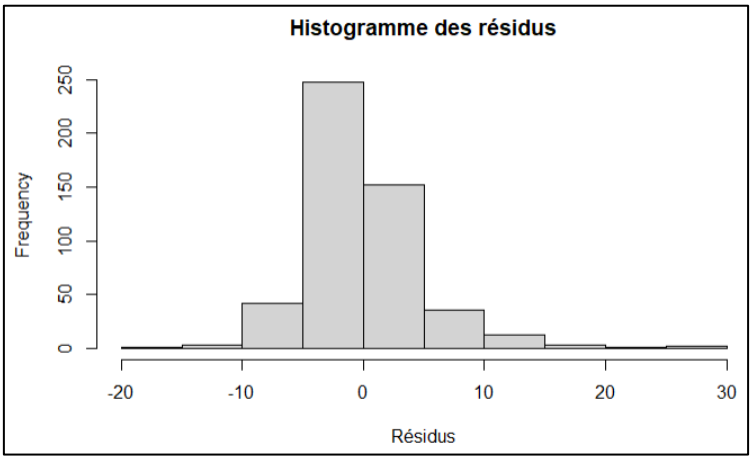
- Test de Ramsey :

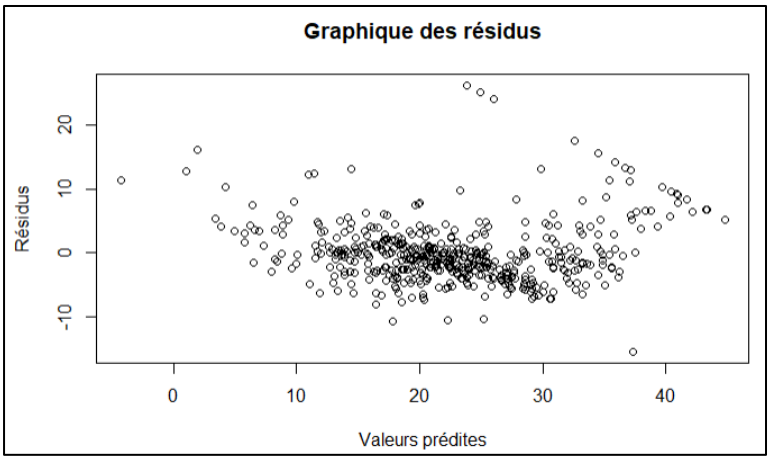
```
        RESET test

data:  modelMEDV
RESET = 95.344, df1 = 2, df2 = 485, p-value < 2.2e-16
```

L'hypothèse nulle de ce test est la suivante : « Le modèle est bien spécifié ». Notre p-value étant très faible, nous rejetons cette hypothèse ce qui signifie que le modèle n'est spécifié (non-linéaire).

Nous pouvons donc conclure de l'intégralité de ces tests qu'il y a un certain nombre de limitations identifiées dans notre modèle avec la non-normalité des résidus, l'hétéroscédasticité, l'autocorrélation positive et la possible mauvaise spécification du modèle.

On trace alors l'histogramme et le graphe des résidus :





On peut observer que la distribution n'est pas similaire à une distribution normale. Il y a donc une non-normalité des résidus. Les résidus n'étant pas bien centrés, les observations semblent donc biaisées.

Pour ce qui concerne la distribution, elle semble inégale, ce qui confirme un problème d'homoscédasticité des résidus, de même qu'il pourrait y avoir des valeurs aberrantes sur la droite.

Ce graphe nous permet d'analyser la relation entre les résidus du modèle et les valeurs prédites. On se rend compte assez facilement qu'il y a un problème d'homoscédasticité dans le modèle étant donné la dispersion des résidus. Les valeurs ne semblent pas non plus linéaires. Nous devons donc bien trouver un moyen d'améliorer notre modèle. Nous avons la possibilité d'attribuer des poids à chacune des variables, ou encore de modifier directement nos données comme le suggérait le test VIF. Poser un logarithme pourrait aussi permettre d'améliorer la linéarité.

2.2 En fonction de CRIM

Dans cette deuxième sous-partie nous allons donc réaliser les mêmes tests mais avec CRIM la variable à expliquer. Le modèle est donc le suivant :

```
Call:
lm(formula = CRIM ~ ., data = housing_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-10.769  -2.137  -0.335   0.999  74.856

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.931e+01  7.381e+00   2.617 0.009157 **
ZN           4.669e-02  1.940e-02   2.407 0.016455 *
INDUS       -6.675e-02  8.480e-02  -0.787 0.431583
CHAS        -7.334e-01  1.197e+00  -0.613 0.540208
NOX         -1.108e+01  5.370e+00  -2.064 0.039588 *
RM           4.483e-01  6.221e-01   0.721 0.471528
AGE          9.292e-04  1.839e-02   0.051 0.959712
DIS         -1.045e+00  2.879e-01  -3.629 0.000314 ***
RAD          5.938e-01  8.942e-02   6.641 8.35e-11 ***
TAX         -3.632e-03  5.254e-03  -0.691 0.489691
PTRATIO     -2.927e-01  1.905e-01  -1.537 0.125050
B           -9.693e-03  3.697e-03  -2.622 0.009014 **
LSTAT        1.096e-01  7.651e-02   1.432 0.152690
MEDV        -2.151e-01  6.134e-02  -3.506 0.000497 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.529 on 487 degrees of freedom
Multiple R-squared:  0.4627,    Adjusted R-squared:  0.4484
F-statistic: 32.26 on 13 and 487 DF,  p-value: < 2.2e-16
```

La valeur du R^2 ajusté est 0.4484 ce qui signifie que les variables indépendantes du modèle expliquent environ 44.84% de la variance de la variable CRIM. Le R^2 est donc relativement faible ce qui indique que le modèle n'ajuste pas spécialement bien les données.

La valeur de la F-statistic est de 32.26. Cette statistique est utile pour tester l'hypothèse nulle comme quoi tous les coefficients de lé régression sont nuls. Dans notre étude, la valeur de la statistique n'est pas très élevée, cependant le modèle de régression dans son ensemble est significatif. En effet on constate que 7 des 14 variables ne sont pas significatives au seuil de 10%.

Nous pouvons aussi déduire de la sortie R que les valeurs qui pour une unité semblent le plus augmenter le prix d'une habitation sont les variables RAD ou ZN et celles qui semblent le diminuer sont les variables MEDV, B et NOX.

Nous allons donc maintenant effectuer des tests pour valider notre modèle :

- Test de Shapiro :

```
        Shapiro-Wilk normality test

data:  residualsCRIM
W = 0.52316, p-value < 2.2e-16
```

L'hypothèse nulle de ce test est : « Les résidus suivent une loi Normale ». Nous avons obtenu une p-value très faible inférieure à 5%. Nous rejetons alors l'hypothèse nulle, nos résidus ne suivent pas une loi Normale.

- Test de Brush-Pagan :

```
        studentized Breusch-Pagan test

data:  modelCRIM
BP = 33.715, df = 13, p-value = 0.001329
```

Il s'agit du test d'homoscédasticité. L'hypothèse nulle est la suivante « La variance des résidus est constante ». Notre p-value très faible suggère donc la présence d'hétéroscédasticité dans les résidus.

- Test de Durbin-Watson :

```
        Durbin-Watson test

data:  modelCRIM
DW = 1.4897, p-value = 4.178e-10
alternative hypothesis: true autocorrelation is greater than 0
```

Il s'agit d'un test portant sur l'autocorrélation des résidus. Dans notre cas, la p-value est encore une fois très faible, ce qui suggère un rejet de l'hypothèse nulle « Aucune corrélation entre les résidus ».

- Test VIF :

```
     CRIM         ZN      INDUS       CHAS        NOX         RM        AGE        DIS
RAD        TAX    PTRATIO          B
 1.881589  2.357765  4.190391  1.104009  4.487682  2.136551  3.196781  3.970938
9.670730 12.062595  1.867524  1.318706
```

Le test VIF nous donne le même résultat que pour la régression précédente ce qui semble parfaitement cohérent. Comme dans le point précédent il serait donc intéressant de regarder si nous obtenons un meilleur modèle en retirant l'une des deux variables.

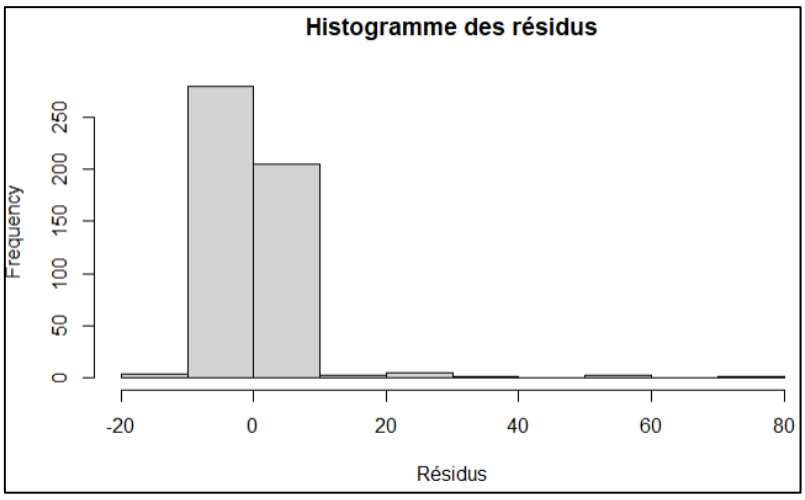
- Test de Ramsey :

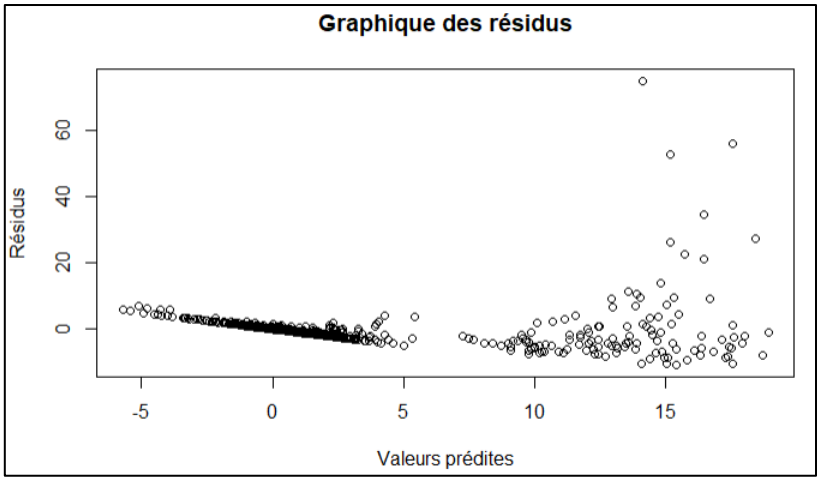
```
        RESET test

data:  modelCRIM
RESET = 48.157, df1 = 2, df2 = 485, p-value < 2.2e-16
```

L'hypothèse nulle de ce test est la suivante : « Le modèle est bien spécifié ». Notre p-value étant très faible, nous rejetons cette hypothèse ce qui signifie que le modèle n'est spécifié (non-linéaire).

Nous arrivons donc à la même conclusion qu'à la régression précédente, notre modèle a ses problèmes qu'il faudra améliorer.

Comme pour le modèle basé sur la variable MEDV, nous traçons donc l'histogramme et le graphe des résidus :





On constate que la distribution de nos résidus ne suit toujours pas une distribution normale, qui de plus est biaisée. On observe aussi qu'il y a toujours ce problème d'homoscédasticité et des valeurs aberrantes sur la droite.

Afin de parfaire ce modèle, il nous faudra recommencer cette démarche en commençant par supprimer la variable qui apporte le moins d'informations entre TAX et RAD. Comme expliqué précédemment nous pourrions aussi linéariser à l'aide d'un logarithme ou encore attribuer des poids à chacune des variables. L'objectif final étant d'améliorer le modèle en supprimant les différents soucis qu'il peut y avoir basés sur l'hétéroscédasticité, la non-linéarité du modèle ou encore la corrélation entre les individus.

On réalise à nouveau les tests avec une base de données sans la variable RAD puis sans la variable TAX :

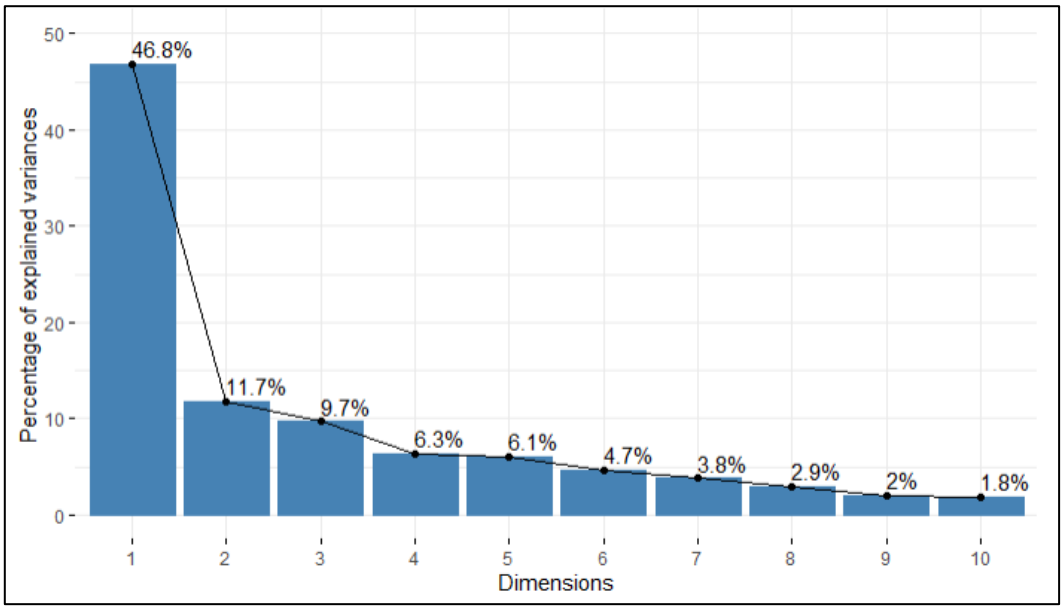
```
        Shapiro-Wilk normality test

data:  residualsMEDV2
W = 0.89676, p-value < 2.2e-16


        studentized Breusch-Pagan test

data:  modelMEDV2
BP = 61.695, df = 12, p-value = 1.106e-08


        Durbin-Watson test

data:  modelMEDV2
DW = 1.0579, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0


        RESET test

data:  modelMEDV2
RESET = 92.996, df1 = 2, df2 = 487, p-value < 2.2e-16


        Durbin-Watson test

data:  modelMEDV2
DW = 1.0579, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0


        Shapiro-Wilk normality test

data:  residuals(modelMEDV2)
W = 0.89676, p-value < 2.2e-16

    CRIM       ZN    INDUS     CHAS      NOX       RM      AGE      DIS      TAX  PTRATIO        B
1.688145 2.291279 3.683058 1.061298 4.268784 1.905030 3.063213 3.986875 3.421242 1.751079 1.342044
   LSTAT
2.915111
```

```
        Shapiro-Wilk normality test

data:  residualsMEDV2
W = 0.90872, p-value < 2.2e-16


        studentized Breusch-Pagan test

data:  modelMEDV2
BP = 61.091, df = 12, p-value = 1.427e-08


        Durbin-Watson test

data:  modelMEDV2
DW = 1.0902, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0


        RESET test

data:  modelMEDV2
RESET = 96.707, df1 = 2, df2 = 486, p-value < 2.2e-16


        Durbin-Watson test

data:  modelMEDV2
DW = 1.0902, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0


        Shapiro-Wilk normality test

data:  residuals(modelMEDV2)
W = 0.90872, p-value < 2.2e-16

    CRIM       ZN    INDUS     CHAS      NOX       RM      AGE      DIS      RAD  PTRATIO        B
1.815210 2.204167 3.221059 1.057770 4.342821 1.941360 3.086730 3.986168 2.830816 1.806977 1.346701
   LSTAT
2.918161
```

On constate qu'en dehors du test VIF qui a été modifié car on a retiré une variable, il n'y a aucun changement dans les p-values, donc tous les problèmes liés aux résidus restent inchangés. Il faudrait donc utiliser d'autres moyens comme la transformation des variables à l'aide de fonctions mathématiques pour résoudre ces problèmes. Nous pourrions aussi ajouter des variables explicatives, utiliser des poids robustes ou alors utiliser des modèles non-linéaires.

3. ACP

Nous allons procéder à un ACP normé sur notre base de données. On commence par déterminer les axes que nous retiendront pour la suite. Ci-dessous l'histogramme du pourcentage de variance expliquée par chacun des axes factoriels :



D'après le critère de Benzecri, nous décidons de garder tous les axes qui ont un pourcentage de variance expliquée supérieur à 1/14=7.14%. C'est donc pour cette raison que nous allons étudier les trois premiers axes factoriels. La somme des trois axes comprend donc 68.2% de l'information des 14 variables. Nous allons donc étudier pour chacun des plans (1-2, 1-3, 2-3) le cercle des corrélations, pour voir quelles variables donc les plus corrélées à chacun des axes.

Le premier graphe est donc le suivant :

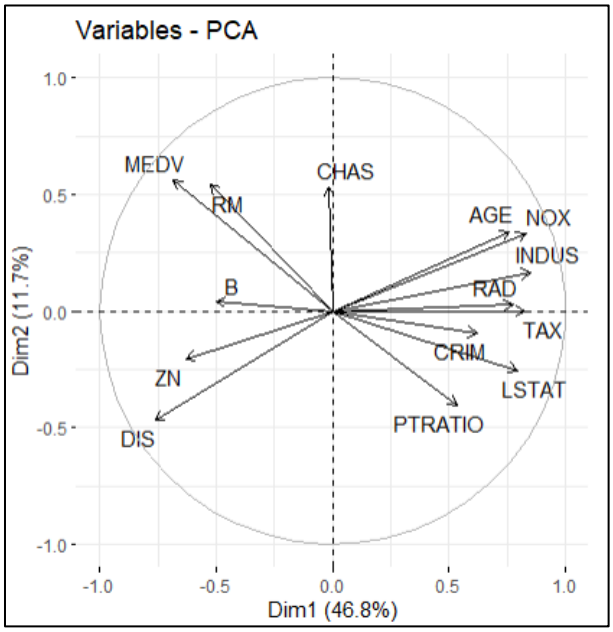
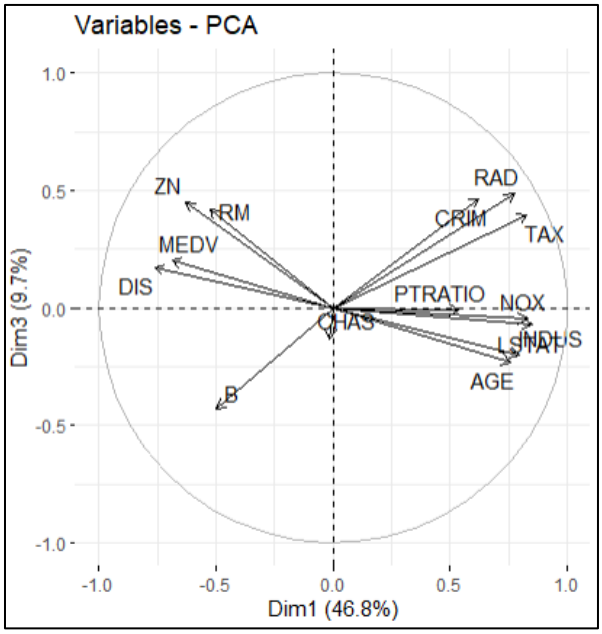


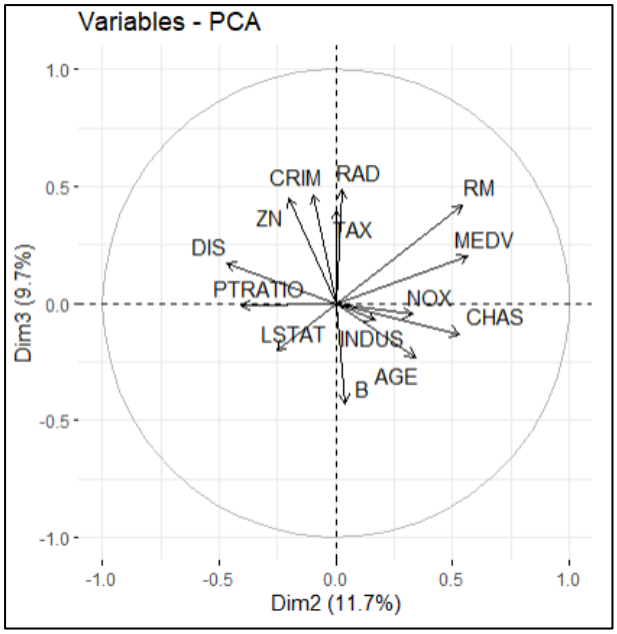
Table des contributions

	Dim.1	Dim.2	Dim.3
CRIM	5.912695598	5.578506e-01	15.917876064
ZN	6.067797783	2.567456e+00	14.697939285
INDUS	10.989975564	1.686639e+00	0.376013109
CHAS	0.004739237	1.710409e+01	1.245570999
NOX	10.507128432	6.623682e+00	0.168668954
RM	4.235331315	1.781007e+01	13.055528985
AGE	8.769297677	6.994181e+00	3.963376204
DIS	8.865542411	1.336827e+01	2.116595678
RAD	9.181556057	4.139299e-02	17.282428076
TAX	10.471762901	3.839038e-07	11.599046835
PTRATIO	4.394278799	9.956291e+00	0.008486602
B	3.802060185	9.879919e-02	13.658694174
LSTAT	9.626768781	3.960715e+00	2.967025506
MEDV	7.171065259	1.923056e+01	2.942749529

A l'aide du graphe de l'ACP sur les deux premiers axes et de la table des contributions, on se rend compte que les variables INDUS, NOX, TAX et AGE sont celles qui semblent apporter le plus d'informations à l'axe 1, et PTRATIO, MEDV et LSTAT à l'axe 2.

On observe alors les graphes des plans 1-3 et 2-3 :





Nos observations sur le premier plan semblent se confirmer. On pourra alors ajouter que les variables CRIM, ZN, B et LSTAT sont les principales contributrices de la troisième dimension. Nous allons donc voir si les variables sont bien représentées en observants les cos2.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
CRIM	0.3876322977	9.171292e-03	0.2159833170	8.720815e-03	0.0003407445
ZN	0.3978006914	4.221002e-02	0.1994304810	9.763043e-02	0.0132047410
INDUS	0.7204953154	2.772904e-02	0.0051019720	4.918911e-05	0.0004944410
CHAS	0.0003107011	2.811982e-01	0.0169006565	4.680347e-01	0.2116034825
NOX	0.6888401861	1.088960e-01	0.0022886018	1.878507e-03	0.0337205896
RM	0.2776654374	2.928048e-01	0.1771452701	7.570097e-02	0.0004438611
AGE	0.5749091851	1.149872e-01	0.0537774723	6.905259e-03	0.0175850339
DIS	0.5812189243	2.197799e-01	0.0287192433	3.158521e-02	0.0080121633
RAD	0.6019365638	6.805177e-04	0.2344983794	1.739351e-03	0.0487112448
TAX	0.6865216460	6.311536e-09	0.1573828442	4.765745e-04	0.0256437779
PTRATIO	0.2880859262	1.636855e-01	0.0001151513	8.327263e-02	0.3320180243
B	0.2492604771	1.624299e-03	0.1853293782	3.238866e-02	0.1254136892
LSTAT	0.6311244068	6.511578e-02	0.0402583867	4.817294e-02	0.0294400741
MEDV	0.4701301559	3.161583e-01	0.0399289957	2.842629e-02	0.0033231862

Les variables INDUS, NOX, AGE, DIS, RAD et TAX ont des cos2 élevés pour la première dimension, ce qui signifie qu'elles sont donc bien représentées dans cette dimension. Il en est de même pour les variables RM et PTRATIO et CRIM, ZN et B respectivement pour les dimensions 2 et 3.

Nous nous pencherons dans la conclusion sur les détails de l'analyse de cette ACP.

4. Régression PLS

Avec cette régression PLS on cherche à expliquer la variable CRIM à l'aide des autres variables explicatives de notre modèle. Nous retiendrons donc 3 axes factoriels pour la suite des analyses. Nous obtenons la sortie R suivante :

```
Data:   X dimension: 501 14
        Y dimension: 501 1
Fit method: kernelpls
Number of components considered: 3
TRAINING: % variance explained
   1 comps  2 comps  3 comps
X    45.65    55.11    60.62
y    54.18    90.25    99.43
```

Pour les variables d'entrées X le modèle est capable d'expliquer une partie de la variance des variables d'entrée. La variance expliquée avec 3 axes principaux atteint donc 60.62% d'explication. Cela suggère que le modèle est capable de capturer une partie significative de la structure des variables X.

Pour la variable Y (CRIM) le modèle explique quasiment la totalité de la variance de la variable de réponse.

Ces résultats nous montrent donc que le modèle PLS à une certaine capacité à interpréter la variance des variables d'entrées et de la variable de réponse. Elle semble donc très bonne pour la variable réponse Y et plutôt satisfaisante pour X.

On s'intéresse aux Loadings :

```
Loadings:
        Comp 1 Comp 2 Comp 3
CRIM     0.310  0.563  0.353
ZN      -0.230  0.475 -0.387
INDUS    0.340 -0.296
CHAS           -0.125  0.304
NOX      0.334 -0.265
RM      -0.203  0.275 -0.109
AGE      0.295 -0.350  0.267
DIS     -0.299  0.321 -0.307
RAD      0.349  0.197 -0.341
TAX      0.363         -0.378
PTRATIO  0.222 -0.103 -0.218
B       -0.229 -0.168  0.426
LSTAT    0.324 -0.197
MEDV    -0.282  0.152
```

Les valeurs obtenues indiquent les poids qui ont été attribués à chaque variable pour chaque composante principale. On constate par exemple que les variables NOX, INDUS ou encore LSTAT ont des chargements positifs forts et sont donc plus influentes contrairement à des variables comme AGE, CHAS ou B. Les variables qui influencent le plus le second axe sont CRIM, ZN, RM et DIS avec des chargements positifs relativement élevés. Les variables les plus influentes sont CRIM, CHAS, AGE, DIS et RAD sont les plus influentes pour l'axe 3.

Enfin nous pouvons analyser les poids associés aux variables X, ils permettent de donner une mesure d'importance à chacune des variables pour prédire CRIM

.

	y
CRIM	0.963133228
ZN	0.010030436
INDUS	-0.063607536
CHAS	0.023247502
NOX	-0.032159708
RM	0.013980502
AGE	0.007860057
DIS	-0.055261642
RAD	0.081066667
TAX	-0.004175808
PTRATIO	-0.047373308
B	0.013024865
LSTAT	0.036115695
MEDV	-0.019649583

On comprend ici que la variable ZN exerce une influence limitée sur la prédiction de CRIM. On peut ajouter que des variables comme INDUS, NOX, DIS, TAX, PTRATIO sont associés à des valeurs plus faibles de Y, contrairement à LSTAT, RAD, AGE ou encore RM.

Nous allons donc pouvoir passer à la conclusion, dans laquelle nous allons essayer de répondre à nos deux problématiques en nous aidant de tous les éléments obtenus via nos analyses.

5. Conclusion – Réponse aux problématiques

A l'aide de tous les résultats que nous avons obtenus nous allons donc pouvoir répondre à nos deux problématiques.

5.1 Problématique 1 – Valeur médiane des logements

La problématique exacte étant : « Quelle est l'influence des caractéristiques socio-économiques sur la valeur médiane des logements ? », pour répondre à cette question nous allons pouvoir utiliser les résultats obtenus grâce à l'ACP et à la régression linéaire effectuée en fonction du paramètre MEDV.

Tout d'abord dans la régression linéaire nous avons pu constater que le taux de criminalité (variable CRIM) présentait une corrélation négative qui était significative, ce qui laissait sous-entendre qu'un taux de criminalité plus élevé pouvait faire baisser la valeur médiane des logements. Une autre variable (NOX) correspondant à la concentration d'oxyde d'azote dans l'air est corrélée négativement avec MEDV, cela sous-entend que les zones moins polluées semblent être plus prisées, c'est une des raisons pour lesquelles ces habitations coutent plus cher.

On observe le même genre de relation entre la variable MEDV et les variables correspondant aux grands terrains résidentiels (ZN) et au nombre de pièces par logement (RM). En effet il y a une corrélation positive entre ces variables ce qui signifie que des terrains plus spacieux et un nombre de pièces plus élevé va faire augmenter les prix des logements. De même pour la variable CHAS, qui nous montre que le fait d'habiter au plus proche du fleuve augmente le prix des habitations.

Autrement dit, plusieurs caractéristiques socio-économiques peuvent influencer la valeur médiane des logements, comme le taux de criminalité, la taille des terrains résidentiels, la proximité des cours d'eau, la qualité de l'air et le nombre de pièces par logement (sous-entendu une plus grande maison).

En parallèle on apprend grâce aux corrélations obtenues avec l'ACP que les variables RM (nombre moyen de pièces par habitation) et PTRATIO (ratio élèves-enseignants par quartier de Boston) sont fortement corrélées positivement avec la première composante principale. Une augmentation de ces variables est donc associée à l'augmentation de la variable MEDV. Ces résultats semblent cohérents vis-à-vis de la régression, et pour ce qui concerne le nombre d'élèves par professeurs, il semblerait que dans les zones ou les logements coûtent le cher il y ait plus de professeurs.

On retrouve encore une fois la variable RM correspondant au nombre de pièces dans une maison et donc à la taille de la maison comme une variable positivement corrélée à la variable MEDV avec la deuxième composante principale.

Nous avons donc pu répondre à notre problématique sur l'influence des caractéristiques socio-économiques sur la valeur médiane des logements.

5.2 Problématique 2 – Taux de criminalité dans la ville

Nous allons donc pouvoir répondre à la dernière problématique « Existe-t-il une corrélation entre la concentration de crime et les autres caractéristiques des quartiers ? » en utilisant les mêmes méthodes que pour la problématique précédente, le tout combiné à une régression PLS.

Tout d'abord la régression linéaire va nous indiquer quelles variables sont positivement ou négativement corrélées à la variable que nous étudions CRIM. Deux variables sont associées positivement à l'augmentation du taux de criminalité : ZN (proportion de terrains résidentiels pour des lots de plus de 25000m²) et RAD (Indice d'accessibilité aux autoroutes). De ces variables nous pourrions déduire que les quartiers les plus proches des autoroutes auraient des taux de criminalité plus élevés car les autoroutes se situant à l'extérieur des villes, les crimes pourraient avoir plus tendances à être commis dans les banlieues.

Cette même régression nous indique que 3 variables sont corrélées négativement à la variable CRIM : DIS (distance pondérée vers les centres d'emploi de Boston), B (la proportion de personnes d'origine afro-américaines par quartier) et MEDV (valeur médiane des logements). On apprend ici donc que les centres d'emploi sont situés dans les zones les plus à risque, ce qui rejoint la variable MEDV car on pourrait se dire que le fait de ne pas avoir de travail fait que l'on a moins de moyens, et donc que l'on doit vivre dans des endroits qui coutent moins chers et ou l'on peut être tenté plus facilement par le crime.

Il est important de noter qu'il n'y avait pas de corrélation significative pour les autres variables.

On utilise maintenant les résultats de la PLS. L'utilisation de ce modèle est pertinente car elle nous permet d'analyser les poids attribués à la variable CRIM. Les variables RAD et PTRATIO ont des poids relativement importants ce qui nous montre que le taux de criminalité peut être influencé par une plus grande proximité aux autoroutes et un ratio professeur-élève plus élevé, ce qui rejoint ce qui était évoqué dans le point précédent.

Il y aussi des variables qui ont des poids négatifs et pour lesquels une augmentation de la variable signifie une diminution du taux de criminalité. On retrouve ici comme dans les points précédents la proportion de personnes noires (variable B).

Nous avons donc pu répondre à cette deuxième problématique à l'aide de la régression linéaire effectuée en fonction de la variable CRIM et de la régression PLS.