DRIDI SELIM
LEFEBVRE ELIE
DALIBARD MATTEO

Rapport statistique : Analyse de la relation entre le revenu médian, le taux de pauvreté et le rapport inter décile selon les départements

2020-2021 TSE

En prenant comme population les départements français, Nous allons Analyser une éventuelle corrélation entre le niveau de vie médian (en euros), le rapport inter décile et le taux de pauvreté pour les Départements. En vulgarisant, si la richesse d'un département à un impact sur la répartition De son revenu et sur son taux de pauvreté. Pour cela nous allons analyser avec le logiciel R différent outils statistiques permettant de nous éclairer sur une potentiel corrélation. Tout d'abord nous allons résumer chaque variable individuellement puis nous allons analyser deux relations entres elles. Enfin nous interprèterons les résultats obtenus.

Nous allons nous demander s'il existe une relation entre le revenu médian, le rapport Inter décile et le taux de pauvreté ?

Origine des donnés : https://www.insee.fr/fr/statistiques/4190004
Base de données de l'INSEE Revenu et Pauvreté

3 variables :
-Médiane du revenu disponible par unité de consommation (en euros)
-Taux de pauvreté de l'ensemble de la population transformé en variable qualitative ordinale Via des classes.
-Rapport inter-décile D9/D1 du revenu disponible par unité de consommation.

Deux études de liaison :
-Revenu Médian * Rapport inter-décile = Quantitative * Quantitative
-Revenu Médian* Taux de pauvreté= Quantitative* Qualitative

Tout d'abord nous avons transformer la variable quantitative taux de pauvrete en variable qualitative selon 4 classes: inferieur à 12, entre 12 et 18, entre 18 et 24 et superieur à 24.
Respectivement pour: faible, intermediaire, eleve et tres-eleve.

Voici un résumé globale des différentes variables:
6 premières valeurs:

Departements	Libelle_commune	`Mediane_niveau_de_vie	taux_de_pauvrete	Rapport_interdécile
1 01	Ain	22272	faible	3.48
2 02	Aisne	18818	eleve	3.12
3 03	Allier	19476.	intermediaire	3.00
4 04	Alpes-de-Haute-Prov…	19719.	intermediaire	3.25
5 05	Hautes-Alpes	19949.	intermediaire	3.04
6 06	Alpes-Maritimes	21246	intermediaire	3.81

6 dernières valeurs:

Departements	Libelle_commune	`Mediane_niveau_de_vie_(en eu…	taux_de_pauvrete	Rapport_interdeci…
1 92	Hauts-de-Seine	26571.	intermediaire	4.93
2 93	Seine-Saint-Denis	16996.	tres-eleve	3.86
3 94	Val-de-Marne	21958	intermediaire	4.20
4 95	Val-d'Oise	21259.	intermediaire	3.69
5 972	Martinique	17057.	tres-eleve	4.37
6 974	La Reunion	14733.	tres-eleve	4.55

Résumé de chaque variable:

Departements	Libelle_commune	Mediane_niveau_de_vie_	taux_de_pauvrete
Length:98	Length:98	Min. :14733	faible :16
Class :character	Class :character	1st Qu.:19571	intermediaire:66
Mode :character	Mode :character.	Median :20093	eleve :13

Mean :20392 tres-eleve : 3
3rd Qu.:20999
Max. :26808

Rapport_interdecile
Min. :2.568
1st Qu.:2.964
Median :3.096
Mean :3.244
3rd Qu.:3.378
Max. :6.347

Le nombre d'individus des variables departements et libellé est le même, length= 98.
Mean signifie moyenne et length signifie longeur. 1st Qu. et 3st Qu.sont les premiers et troisièmes quartils.

1) Description univarié des données

1)Etude de la variable Niveau de vie Median

Résumé:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
14733	19571	20093	20392	20999	26808

Variance selon R du niveau de vie Median en euros:
2927825

On définit n comme la longeur de l'effectif:
N=98

Pour retrouver la variance du cours, on divise (n-1) par la variance multiplié à l'effectif ce qui donne :
2897950

Ecart type selon R:
1711.089

Ecart type du cours:
1702.337

Coefficiant de variation du cours:
0.08348183

Minimum et Maxium:
14733.45 26808.00

Etendue:
12074.55

Intervalle inter-quartile:
1427.673

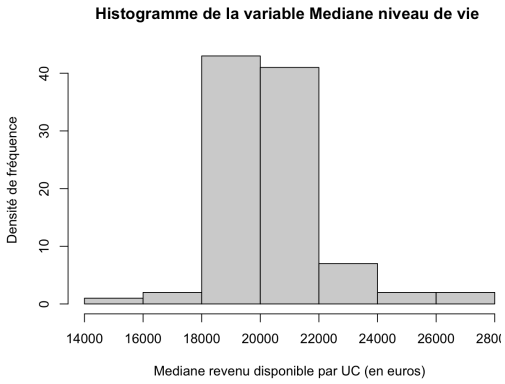
Quartiles:

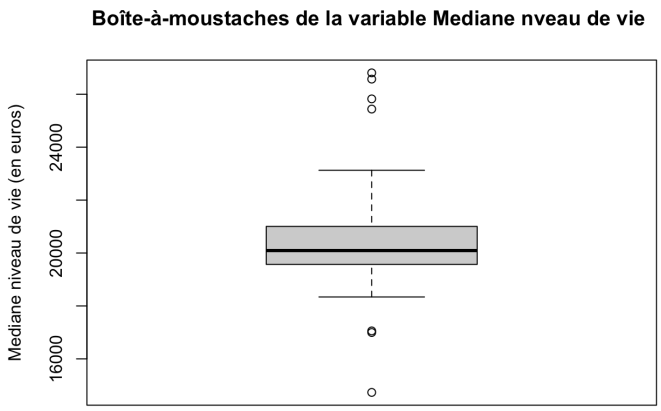
0%	25%	50%	75%	100%
14733.45.	19570.95.	20092.71.	20998.62	26808.00

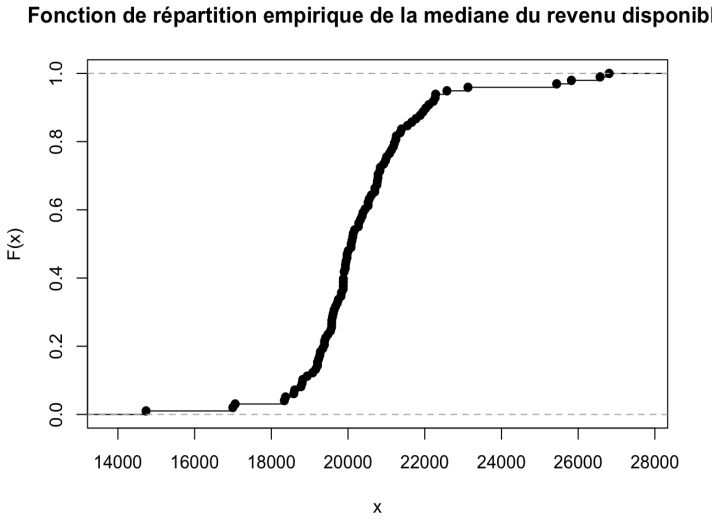
Deciles:

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
18899.20	19384.20	19639.44	19894.02	20092.71	20459.79	20783.00	21231.76	22049.97	26808.00

Médiane: 2009.71







Maintenant nous allons étudier la variable Rapport interdécile:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.568	2.964	3.096	3.244	3.378	6.347

Variance selon R:
0.2687923

Variance du cours:
0.2660496

Ecart type:
0.5184519

Ecart type du cours:
0.5157999

Coefficiant de variation:
0.1590062

Minimum et Maxium:
2.567641 6.346678

Etendue:
3.779038

Intervalle inter-quartile:
0.4147796

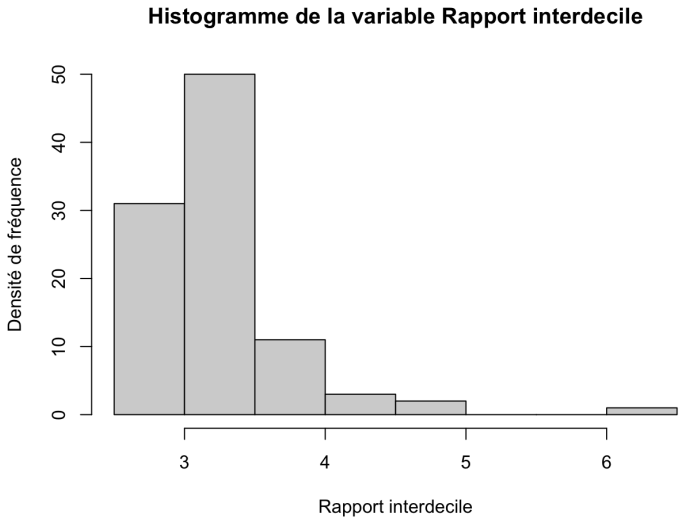
Quartiles:

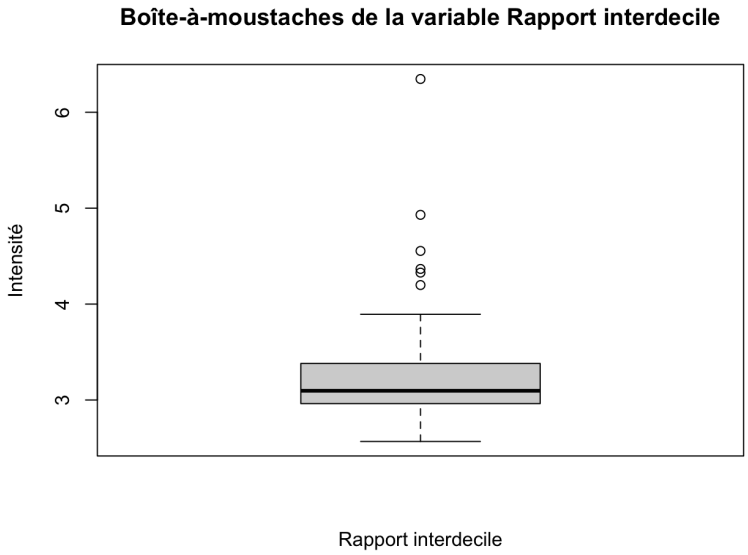
0%	25%	50%	75%	100%
2.567641	2.963663.	3.096145.	3.378442	6.346678

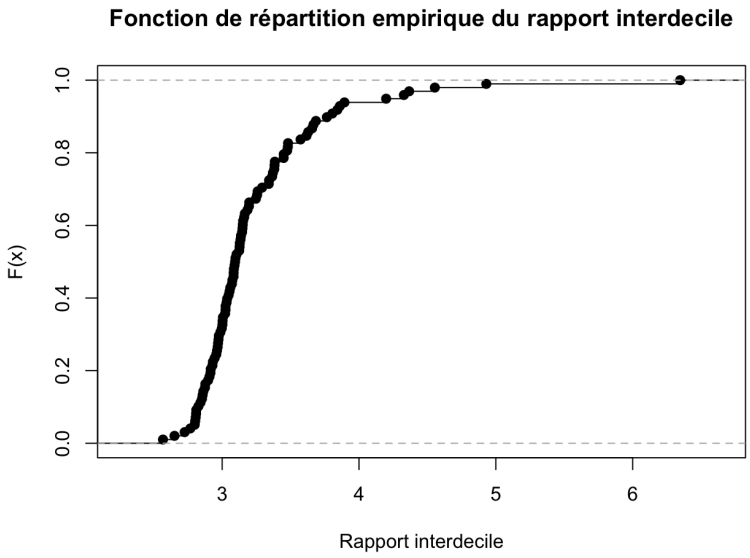
Deciles:

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
2.837333	2.921435	2.988337	3.045487	3.096145	3.151144	3.288606	3.464919	3.776917	6.346678

Médiane:
3.096145


Histogramme de la variable Rapport interdecile
Densité de fréquence
0
10
20
30
40
50
3
4
5
6
Rapport interdecile


Boîte-à-moustaches de la variable Rapport interdecile
Intensité
3
4
5
6
Rapport interdecile


Fonction de répartition empirique du rapport interdecile
F(x)
0.0
0.2
0.4
0.6
0.8
1.0
3
4
5
6
Rapport interdecile

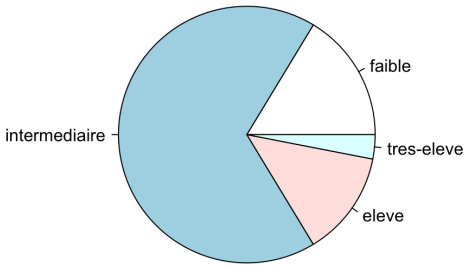
Analyse de la variable taux de pauvreté

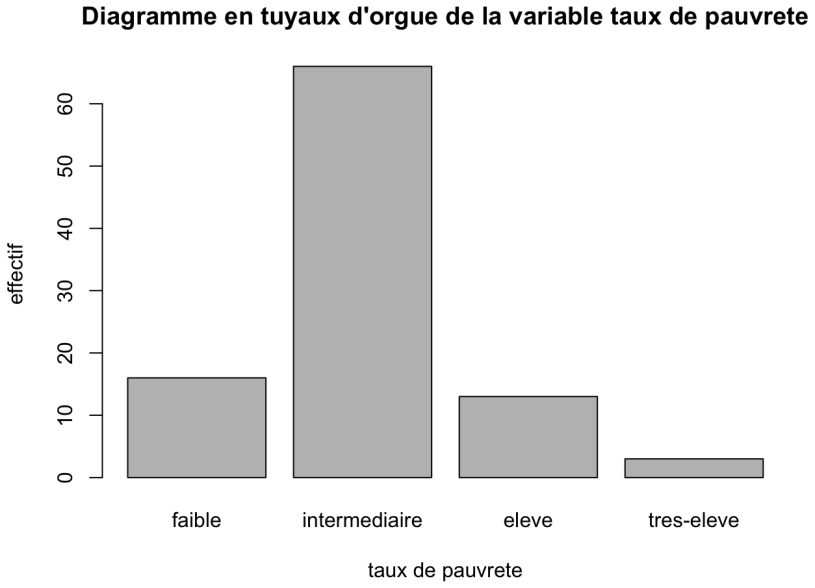
Tableau en effectif :

faible	intermediaire	eleve	tres-eleve
16	66	13	3

Tableau en fréquence:

faible	intermediaire	eleve	tres-eleve
0.16326531	0.67346939	0.13265306	0.03061224





Liaison entre la variable niveau de vie median et rapport interdecile

Niveau de vie median:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.

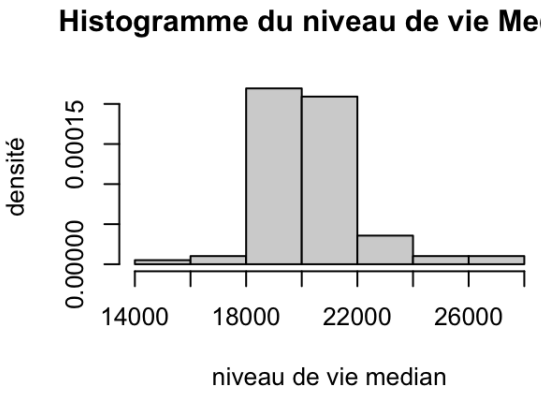
14733 19571 20093 20392 20999 26808

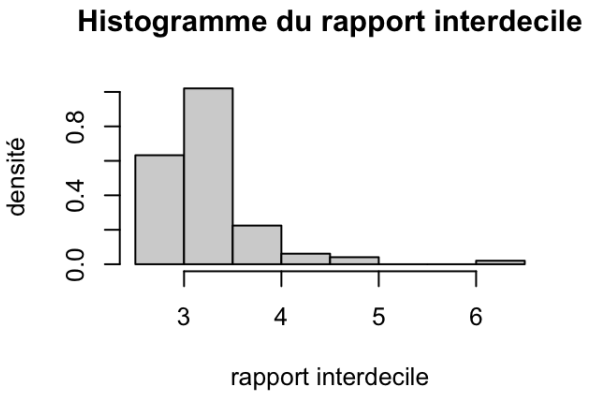
Rapport inter décile :
Min. 1st Qu. Median Mean 3rd Qu. Max.
2.568 2.964 3.096 3.244 3.378 6.347

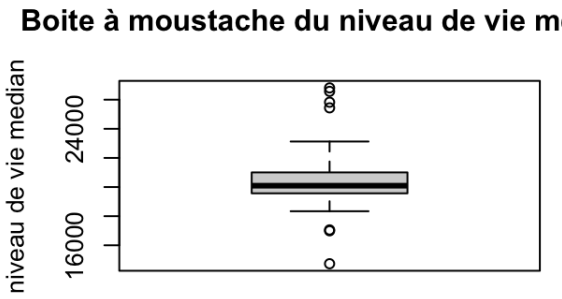
la médiane des niveaux de vie médian est de 20093€ tandis que sa moyenne est de 20392€
la médiane du rapport inter décile est de 3.096 tandis que sa moyenne est de 3.244
La médiane < moyenne pour les deux séries donc il y a une surreprésentation des petites valeurs.

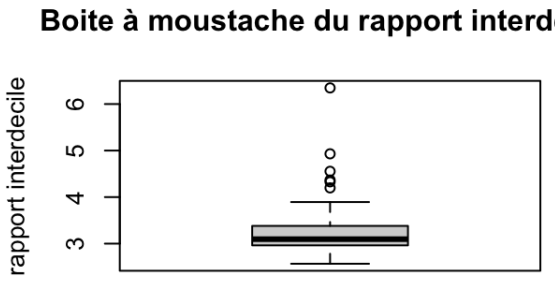
Variance niveau de vie médian :
2897950

Variance rapport inter décile :
0.2660496

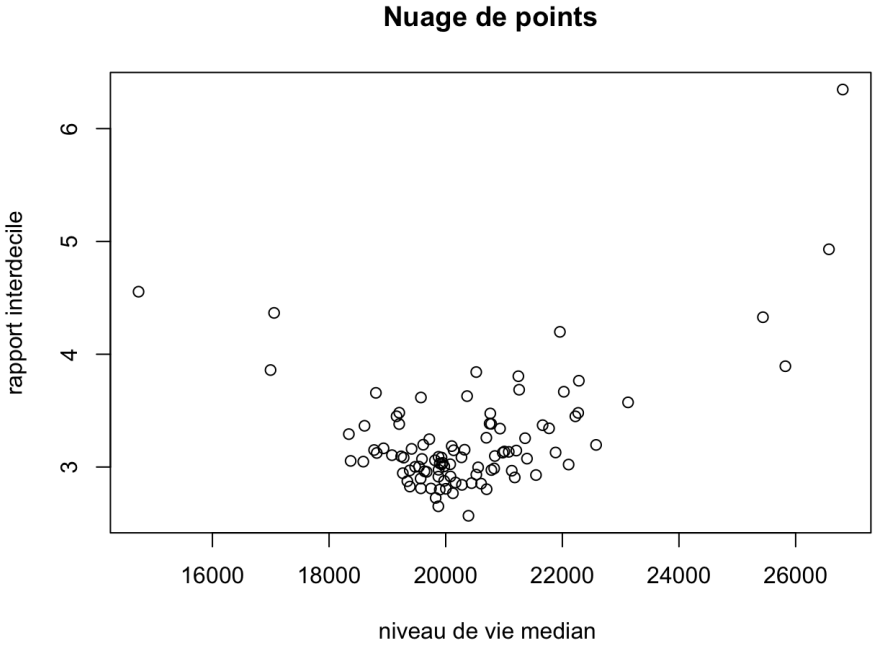








on choisit le rapport inter décile comme variable à expliquer (=Y, en ordonnées)
et le niveau de vie médian comme variable explicative (=X, en abscisses)
le nuage de points est concentré avec une légère corrélation positive: quand le niveau de vie médian augmente, le rapport inter décile augmente légèrement)
on remarque une dispersion pour les valeurs des niveau de vie médian inferieur à 18000 ou supérieur à 24000
on remarque un point isolé avec un niveau de vie médian très élevé qui peut être un point influent



Covariance:
367.9503
Covariance>0 donc liaison positive entre les deux variables

Coefficient de corrélation linéaire :
0.4147707

il est positif mais proche de 0.4, il y a donc une faible corrélation linéaire positive entre les deux variables.

equation de la droite de regession (D): rapport_interdecile
= 0,000126*mediane_niveau_de_vie+0.681201

Coefficiants précis:

(Intercept) base$`Mediane_niveau_de_vie_(en euros)`
0.6812012436 0.0001256736

Valeurs ajustées:

1	2	3	4	5	6	7	8	9	10	11
3.480203	3.046127	3.128834	3.159333	3.188221	3.351262	3.179352	3.040387	3.060705	3.146163	2.985958
12	13	14	15	16	17	18	19	20	21	22
3.178936	3.260472	3.264547	3.141112	3.171557	3.228605	3.186190	3.204988	3.369862	3.229736	3.017070
23	24	25	26	27	28	29	30	31	32	33
3.104125	3.403542	3.207659	3.293117	3.337019	3.282812	3.240753	3.044283	3.088758	3.474087	3.179205
34	35	36	37	38	39	40	41	42	43	44
3.365740	3.141261	3.343470	3.117446	3.300553	3.431498	3.299066	3.250452	3.260606	3.181854	3.173069
45	46	47	48	49	50	51	52	53	54	55
3.389215	3.330526	3.185624	3.078873	3.140633	3.210172	3.195059	3.292698	3.112042	3.178784	3.282242
56	57	58	59	60	61	62	63	64	65	66
3.148892	3.270998	3.289347	3.137114	3.094469	3.346788	3.102261	2.989825	3.321037	3.317581	3.153577
67	68	69	70	71	72	73	74	75	76	77
3.019751	3.417366	3.481627	3.449192	3.162919	3.191321	3.215870	3.459554	3.878277	4.050258	3.235391

```
      78       79       80       81       82       83       84       85       86       87       88
3.518731 3.926596 3.181854 3.121154 3.143524. 3.098637 3.290729 3.094613 3.243685 3.204391 3.211764
      89       90       91       92       93       94     95       96       97       98
3.117174 3.191579 3.311717 3.587629 4.020516 2.817181 3.440742 3.352938 2.824868 2.532806
```

Résidus:

```
                          1          2          3          4          5          6          7
-0.0001637257  0.0786544221 -0.1269448058  0.0873115566 -0.1517076630  0.4538324322 -0.2036958719
                          8          9         10         11         12         13         14
 0.1102574069  0.1050525230  0.0517315532  0.3058833396 -0.2641711118  0.5815106415 -0.2672517553
                         15         16         17         18         19         20         21
-0.3300344863 -0.1119974706 -0.1428769208 -0.1610915471 -0.2885733951 -0.2953090771 -0.3878995733
                         22         23         24         25         26         27         28
 0.0306602337 -0.0208942821 -0.0325671105 -0.0223044789 -0.3198115101 -0.3696981371 -0.4796952918
                         29         30         31         32         33         34         35
 0.3879377023  0.6132273643  0.3610877745 -0.0247194144 -0.0886343180 -0.1103041455  0.4751335668
                         36         37         38         39         40         41         42
-0.4359091059 -0.2906189128 -0.2029231819 -0.3025660962 -0.3117244727 -0.3931890507 -0.3281705030
                         43         44         45         46         47         48         49
-0.1494098943 -0.4469667737 -0.4602509073 -0.1961797708 -0.1015635867  0.0263938670 -0.2432386244
                         50         51         52         53         54         55         56
-0.4425266691 -0.3882342713  0.0900249495 -0.2375640684 -0.5270346835 -0.0227511022 -0.1864483566
                         57         58         59         60         61         62         63
-0.4182882056  0.0956077325 -0.1338675757  0.2864623272 -0.2009351454 -0.1569606950  0.0643618996
                         64         65         66         67         68         69         70
-0.1834052120 -0.1902149468 -0.1961941249  0.3456115253 -0.0742811148  0.2832141882  0.2181637439
                         71         72         73         74         75         76         77
-0.3532320541 -0.3142377009 -0.3534998866 -0.4377205515  0.4497866753  2.2964199054 -0.0822484911
                         78         79         80         81         82         83         84
-0.3226840392 -0.0329942670 -0.3823989911  0.0396643429 -0.0722092463 -0.0039773059  0.1842383538
                         85         86         87         88         89         90         91
 0.3869439578 -0.6760448810 -0.1797332934 -0.0631118924 -0.1470432526 -0.1857946401  0.0288893107
                         92         93         94         95         96         97         98
-0.0145028751  0.9098411159  1.0425612810  0.7575762175  0.3322642018  1.5415253019  2.0213650980
```

Test pour le departement 14 Calvados
3.264547

Résumé de la regression :
Call:
lm(formula = base$Rapport_interdecile ~ base$`Mediane_niveau_de_vie_(en euros)`)

Residuals:
```
    Min      1Q  Median      3Q     Max
-0.67604 -0.29011 -0.11115  0.08935  2.29642
```

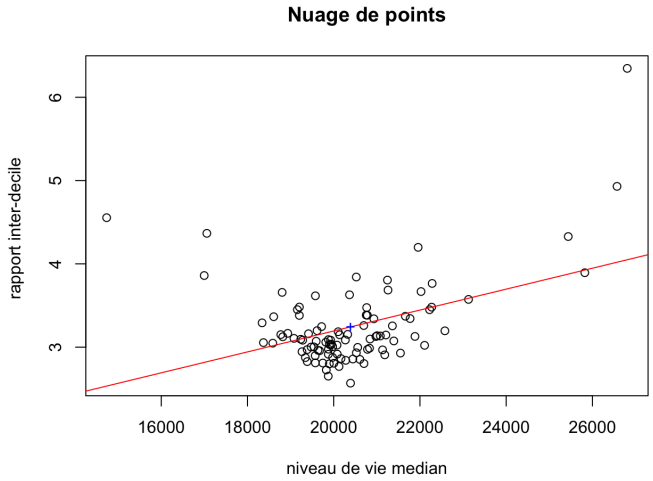
Coefficients:
```
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     6.812e-01  5.758e-01   1.183     0.24
base$`Mediane_niveau_de_vie_(en euros)` 1.257e-04  2.814e-05   4.466 2.17e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 0.4742 on 96 degrees of freedom
Multiple R-squared: 0.172,	Adjusted R-squared: 0.1634
F-statistic: 19.95 on 1 and 96 DF, p-value: 2.174e-05

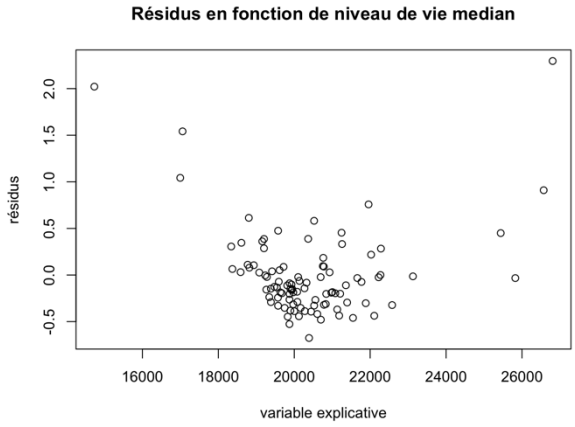
Coefficiant de correlation R2:
0.1720348
17% de la variation de la variable rapport interdecile est expliqué par la regression donc, de la variable niveau de vie median.

le R2 est assez proche de 0 donc le modèle est plutot de mauvaise qualité.

on ajoute la droite de régression (en rouge)



Résidus en fonction de SE



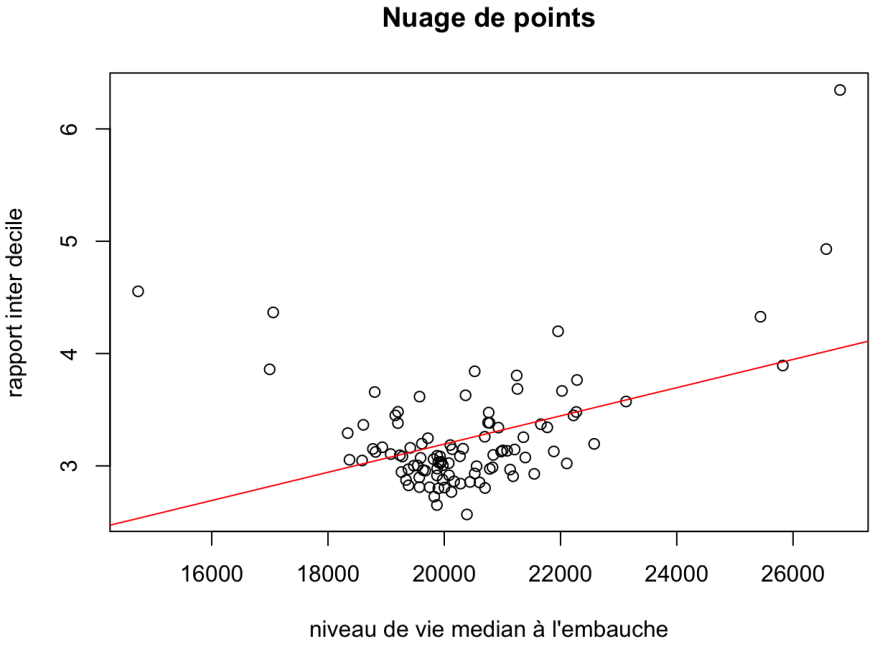
Moyenne des résidus:
-1.875448e-17

en théorie la somme des résidus est également nulle, sa valeur est aussi très proche de 0

Sommes des résidus:
-1.83447e-15

Il est important de repérer les departements ayant de forts résidus car leur rapport interdecile a été mal prédit par la régression (erreur de saisie ? comportement particulier ? point influent = qui a un fort impact dans l'estimation des coef. de la régression ?)

en rouge ce sont les valeurs prédites par la regression



graphiquement, le departement ayant le plus fort résidu est le 76:
2.29642

4 valeurs avec le résidu le plus élevé:
98 76 97 94
2.021365 2.296420 1.541525 1.042561

Valeur du niveau de vie median maximum:
26808

On crée de nouveaux vecteurs de données sans ce departement:

Le coefficiant de correlation est nettement inferieur sans la valeur extrême:
0.2474673

Résumé sans la valeur extrême:

Call:
lm(formula = SA2 ~ SE2)

Residuals:
Min 1Q Median 3Q Max
-0.6484 -0.2660 -0.1180 0.1416 1.7016

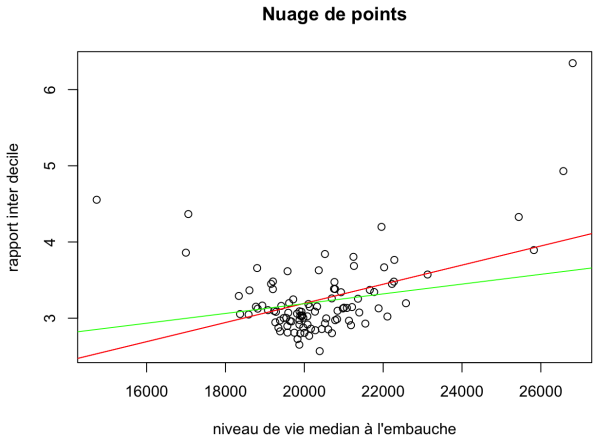
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.906e+00 5.263e-01 3.621 0.000473 ***
SE2 6.426e-05 2.581e-05 2.489 0.014533 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4019 on 95 degrees of freedom
Multiple R-squared: 0.06124,	Adjusted R-squared: 0.05136
F-statistic: 6.197 on 1 and 95 DF, p-value: 0.01453

le R2 vaut 0,06124, il est lui aussi inférieur

ajout de cette nouvelle droite de régression sur le nuage de points
La droite verte est assez éloigné de la rouge : on peut donc considérer que le point 76, Paris est un point influent.



Étude sans le point de plus fort résidu (numéro 76)

on pourrait faire cette analyse en excluant tous les individus que l'on a repérés comme ayant de forts résidus (numéros : 98,76,97,94)

Corrélation sans le département au plus fort résidu, (toujours Paris) :
0.2474673

Même résumé:

Call:
lm(formula = SA3 ~ SE3)

Residuals:
    Min   1Q Median   3Q   Max
-0.6484 -0.2660 -0.1180 0.1416 1.7016

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.906e+00 5.263e-01   3.621 0.000473 ***
SE3   6.426e-05 2.581e-05   2.489 0.014533 *

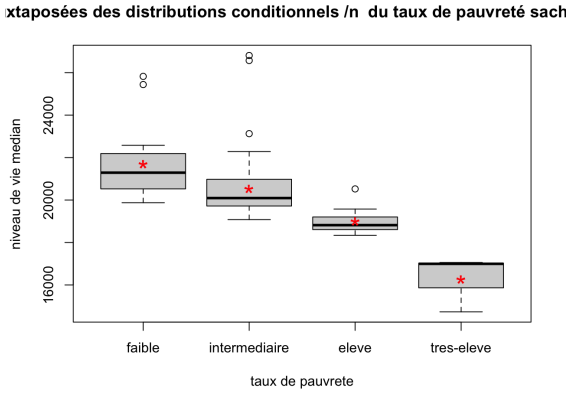
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4019 on 95 degrees of freedom
Multiple R-squared: 0.06124,	Adjusted R-squared: 0.05136
F-statistic: 6.197 on 1 and 95 DF, p-value: 0.01453

on voit que le R2 a diminué : 6% au lieu de 17%
les coefficients de régression ont évolué.

Ici, on va cherche à étudier la liaison entre le niveau de vie median et le taux de pauvrete
est-ce que le niveau de vie median a une influence sur le taux de pauvrete ? en particulier,
est-ce que le taux de pauvrete de PARIS 75 est plus faible que celui des autres departements?

3. Etude de la liaison entre le salaire et le statut professionnel
3.1 Représentation graphique des distributions conditionnelles



on remarque de grandes différences de niveau de vie median selon le taux de pauvrete
les departements au faible taux de pauvrete ont des revenues median plus élevés que les autres departements
les departements au taux de pauvreté elevé ont des niveau de vie median intermédiaires, avec une très faible dispersion
alors que les niveau de vie median des trois autres catégories sont assez dispersés

3.2 Résumés numériques des distributions conditionnelles

Moyenne conditionnelle sachant le taux de pauvrete

faible	intermediaire	eleve	tres-eleve
21700.14	20537.93	18991.85	16262.37

Tableau effectif selon le taux de pauvrete:

faible	intermediaire	eleve	tres-eleve
16	66	13	3

Variance conditionelle:

faible	intermediaire	eleve	tres-eleve
2796028.3	1913303.4	311090.3	1169428.3

Résumé selon le taux de pauvreté:
$faible

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19874	20566	21289	21700	22149	25824

$intermediaire

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19079	19726	20093	20538	20966	26808

$eleve

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18339	18608	18818	18992	19203	20524

$`tres-eleve`
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
14733	15865	16996	16262	17027	17057

3.3 Calcul du rapport de corrélation

Variance intra:
1822112

Variance totale:
2927825

Variance inter:
1105714

Calcul direct variance inter:

faible	intermediaire	eleve	tres-eleve
42376324296	42424673623	42489034325	42602777596

calcul du rapport de corrélation eta2:
0.377657

eta2=0,377657 est assez proche de 0.4. Il existe donc une liaison faible mais quand même notable entre le taux de pauvrete et le niveau de vie median, avec la moyenne des niveau de vie median des departements au taux de pauvrete faible qui est environ 50% plus élevée que de la moyenne de ceux au taux de pauvreté tres elevé.

Conclusion:

Nous avons donc analyser et observer une legère corélation entre le revenu médian, le rapport interdécile et le taux de pauvreté, 17% de la variation de la variable rapport interdecile est expliqué par la regression donc, de la variable niveau de vie median ce qui est relativement faible mais Cela est quand même notable. Pour le rapport de correlation entre le taux de pauvreté est le niveau de vie médian cela est plus significue car le R2 est de 0.4. Cependant cette analyse statistique a des limites, l'études à l'échelle des départements montre une version déja harmonisé des differentes valeurs. De plus, le niveau vie est ici la médiane au sein de chaque département. Enfin, les outils utilisés offre également uniquement un aperçu de la relation entre les différentes variables et non une analyse réélement poussé.