Analyse exploratoire - ML Non Supérvisé

Data et librairies

“db_imp” est un dataset avec 100k observations imputées afin d’éliminer les données manquantes grâce à la librairie “miceranger”, la quelle se base sur de random forest pour imputer les NA.

“db_job” est un dataset avec 39754 observations, il s’agit du périmètre de salaries (personnes actives).

“learn_code” est un datasets avec 100k observations avec des données manquantes.

“db_job_clust” est une datasets avec les clusters k-means réalisé avec l’aide de la library (kamila) (périmètre des salaries)

```
library(reshape2)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(viridis)
```

```
## Warning: package 'viridis' was built under R version 4.0.5
```

```
## Loading required package: viridisLite
```

```
## Warning: package 'viridisLite' was built under R version 4.0.5
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.0.4
```

```
## corrplot 0.84 loaded
```

```
library(multcompView)
```

```
## Warning: package 'multcompView' was built under R version 4.0.5
```

```
db_imp = read.csv("datasets/db_imp_code.csv")
db_job = read.csv("datasets/learn_job_pred.csv")
learn_code = read.csv("datasets/learn_code.csv",encoding = "UTF-8")
```

1. Caractérisation des données

L'objectif de ce chapitre est de caractériser la data en fonction des departements. Dans cette partie aucune imputation a été faite. le choix de la médiane a été fait car les données n'ont pas une distribution symétrique. En prenand la médiane de l'emolument, de l'age, du nombre d'heurs travaillé et la répartition des étudiants par département des clusters sont constatés. En revanche, la répartition des hommes et des fammes ne permet pas la construction des clusters par département. Pour les variables catégoriques deux tableau sont fait en fonction des départements, a) Croissement entre le département et les modalitées de la variable en fonction de la médiane de l'emolument; b) même croissement en fontion du nombre des observations. Visuelment des differences sont constatées entre les départements, un exemple est la difference entre Hauts-des-seine et la Seine-saint-Dennis (deux départements proche geographiquement). Dans le chapitre suivant la différence entre départements sera observé statistiquement.

Ici les plots.

```
#plot_generator()
```

2. Différence statistique entre départements

L'objectif de cette partie est de pouvoir confirmer statistiquement les résultats de la data visualisation. Deux analyses sont réalisés, Anova pour les variables numériques et chi^2 pour les variables catégoriques. La statistique confirme ce qui a été observé dans la première partie. Par exemple, pour la variable "sex" le test de chi^2 montre qu'il n'y a pas de dépendance entre le sex et les départements. En revanche, il existe une dépendance entre les départements et la densité des étudiants. La "lm" permet d'identifier des différence entre les départements et le departement de réference (intercept) "Ain" en fonction de la variable utilisé pour la régression. Par exemple, les departements comme "Alpes-de-haute-provence", "Allier" et tous les autre pour lesquels la p-value est grande sont des départements sans différence significative de salaire avec "Ain".

```
testGeographicNum  = function(variable, db) {

                    mod = lm(data = db,
                            formula(paste0(variable, "~ Nom.du.département")))
                    s = summary(mod)
                    print(s)
}
testGeographicCat = function(variable1, db) {
                    tab_cont = table(db[,variable1], db[,"Nom.du.département"])
                    preuve = chisq.test(tab_cont, simulate.p.value = TRUE)
                    print(preuve)
}
testGeographicNum("EMOLUMENT", db = learn_code)# Il y a des différences significative entre d
epartement
```

```
##
## Call:
## lm(formula = formula(paste0(variable, "~ Nom.du.département")),
##     data = db)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32075  -8216  -1923   5518 181977
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          19655.63    1040.84  18.884  < 2e-16
## Nom.du.départementAisne                610.57    2138.15   0.286 0.775218
## Nom.du.départementAllier              -296.70    1863.50  -0.159 0.873498
## Nom.du.départementAlpes-de-Haute-Provence -4218.05    3951.98  -1.067 0.285832
## Nom.du.départementAlpes-Maritimes      3566.52    1156.38   3.084 0.002042
## Nom.du.départementArdèche              1123.22    1973.96   0.569 0.569347
## Nom.du.départementArdennes             -952.42    1997.58  -0.477 0.633517
## Nom.du.départementAriège                674.67    2523.64   0.267 0.789209
## Nom.du.départementAube                 -537.74    2076.85  -0.259 0.795696
## Nom.du.départementAude                 -399.23    1838.02  -0.217 0.828050
## Nom.du.départementAveyron                25.89    2171.88   0.012 0.990490
## Nom.du.départementBas-Rhin             1702.31    1121.58   1.518 0.129077
## Nom.du.départementBouches-du-Rhône     2331.46    1097.64   2.124 0.033670
## Nom.du.départementCalvados             2929.37    1332.47   2.198 0.027923
## Nom.du.départementCantal               4879.72    2492.65   1.958 0.050278
## Nom.du.départementCharente              580.79    1520.41   0.382 0.702467
## Nom.du.départementCharente-Maritime     419.95    1340.96   0.313 0.754153
## Nom.du.départementCher                  427.35    1469.69   0.291 0.771223
## Nom.du.départementCorrèze              -635.06    1821.94  -0.349 0.727421
## Nom.du.départementCorse-du-Sud        -1825.78    1910.06  -0.956 0.339140
## Nom.du.départementCôte-d'Or             176.64    1293.29   0.137 0.891362
## Nom.du.départementCôtes-d'Armor         -31.53    1920.05  -0.016 0.986899
## Nom.du.départementCreuse              -3186.83    2791.37  -1.142 0.253597
## Nom.du.départementDeux-Sèvres          2247.07    1506.60   1.491 0.135844
## Nom.du.départementDordogne             -993.50    1460.84  -0.680 0.496452
## Nom.du.départementDoubs                 899.52    1298.79   0.693 0.488577
## Nom.du.départementDrôme                1122.28    1529.10   0.734 0.462985
## Nom.du.départementEssonne              5239.15    1122.44   4.668 3.06e-06
## Nom.du.départementEure                 2059.69    1371.05   1.502 0.133034
## Nom.du.départementEure-et-Loir         2802.44    1364.55   2.054 0.040006
## Nom.du.départementFinistère            1011.98    1179.16   0.858 0.390774
## Nom.du.départementGard                  284.13    1177.57   0.241 0.809335
## Nom.du.départementGers                  718.81    2492.65   0.288 0.773064
## Nom.du.départementGironde              2379.01    1091.42   2.180 0.029283
## Nom.du.départementHaut-Rhin            1187.17    1167.03   1.017 0.309039
## Nom.du.départementHaute-Corse         -1322.42    1900.30  -0.696 0.486496
## Nom.du.départementHaute-Garonne        5033.49    1102.90   4.564 5.04e-06
## Nom.du.départementHaute-Loire          1532.94    2463.07   0.622 0.533703
## Nom.du.départementHaute-Marne           -13.08    3282.26  -0.004 0.996821
## Nom.du.départementHaute-Saône           271.00    1666.35   0.163 0.870811
## Nom.du.départementHaute-Savoie         1383.41    1142.72   1.211 0.226044
## Nom.du.départementHaute-Vienne          480.10    1496.14   0.321 0.748292
## Nom.du.départementHautes-Alpes         -566.33    3807.90  -0.149 0.881772
## Nom.du.départementHautes-Pyrénées       -14.04    2357.03  -0.006 0.995246
## Nom.du.départementHauts-de-Seine       11159.53    1108.08  10.071  < 2e-16
## Nom.du.départementHérault                864.21    1131.56   0.764 0.445032
```

```
## Nom.du.départementIlle-et-Vilaine          3320.14    1142.10   2.907 0.003651
## Nom.du.départementIndre                    3012.58    1633.70   1.844 0.065186
## Nom.du.départementIndre-et-Loire           1462.17    1262.27   1.158 0.246722
## Nom.du.départementIsère                    2819.66    1110.09   2.540 0.011088
## Nom.du.départementJura                     1235.68    1604.40   0.770 0.441195
## Nom.du.départementLandes                  -1234.49    1483.71  -0.832 0.405398
## Nom.du.départementLoir-et-Cher             3478.71    1411.30   2.465 0.013710
## Nom.du.départementLoire                     672.29    1167.41   0.576 0.564698
## Nom.du.départementLoire-Atlantique         2479.35    1104.45   2.245 0.024782
## Nom.du.départementLoiret                   1730.19    1232.19   1.404 0.160279
## Nom.du.départementLot                     -1108.28    2704.17  -0.410 0.681926
## Nom.du.départementLot-et-Garonne            833.05    1523.27   0.547 0.584463
## Nom.du.départementLozère                   2814.92    4523.62   0.622 0.533766
## Nom.du.départementMaine-et-Loire           1047.29    1151.78   0.909 0.363209
## Nom.du.départementManche                   2114.51    1409.63   1.500 0.133611
## Nom.du.départementMarne                     482.65    1557.25   0.310 0.756609
## Nom.du.départementMayenne                   603.77    1763.93   0.342 0.732136
## Nom.du.départementMeurthe-et-Moselle       -294.68    1174.94  -0.251 0.801968
## Nom.du.départementMeuse                   -3717.83    2791.37  -1.332 0.182900
## Nom.du.départementMorbihan                 2342.40    1200.62   1.951 0.051065
## Nom.du.départementMoselle                  1398.44    1130.78   1.237 0.216203
## Nom.du.départementNièvre                   -171.93    1596.57  -0.108 0.914246
## Nom.du.départementNord                     1857.50    1082.66   1.716 0.086228
## Nom.du.départementOise                     2541.49    1169.96   2.172 0.029839
## Nom.du.départementOrne                     -164.16    1676.38  -0.098 0.921993
## Nom.du.départementParis                   12759.26    1090.70  11.698  < 2e-16
## Nom.du.départementPas-de-Calais             143.57    1120.96   0.128 0.898086
## Nom.du.départementPuy-de-Dôme               734.90    1478.94   0.497 0.619255
## Nom.du.départementPyrénées-Atlantiques      987.72    1302.05   0.759 0.448105
## Nom.du.départementPyrénées-Orientales      -356.10    1681.52  -0.212 0.832284
## Nom.du.départementRhône                    3638.35    1088.93   3.341 0.000835
## Nom.du.départementSaône-et-Loire           -424.92    1281.55  -0.332 0.740217
## Nom.du.départementSarthe                   2571.84    1517.58   1.695 0.090140
## Nom.du.départementSavoie                   1014.76    1732.06   0.586 0.557966
## Nom.du.départementSeine-et-Marne           4808.05    1120.27   4.292 1.78e-05
## Nom.du.départementSeine-Maritime           1862.25    1109.01   1.679 0.093123
## Nom.du.départementSeine-Saint-Denis        2485.85    1113.26   2.233 0.025558
## Nom.du.départementSomme                   -1922.03    2091.47  -0.919 0.358109
## Nom.du.départementTarn                     -740.50    1726.03  -0.429 0.667911
## Nom.du.départementTarn-et-Garonne          1938.97    2226.95   0.871 0.383932
## Nom.du.départementTerritoire de Belfort    3357.24    1973.96   1.701 0.088996
## Nom.du.départementVal-d'Oise               4842.05    1133.47   4.272 1.94e-05
## Nom.du.départementVal-de-Marne             5929.31    1122.76   5.281 1.29e-07
## Nom.du.départementVar                      1226.75    1160.87   1.057 0.290630
## Nom.du.départementVaucluse                -2011.78    2333.19  -0.862 0.388558
## Nom.du.départementVendée                    760.62    1517.58   0.501 0.616230
## Nom.du.départementVienne                   -698.89    1463.02  -0.478 0.632864
## Nom.du.départementVosges                     85.25    1890.77   0.045 0.964037
## Nom.du.départementYonne                    -515.04    1454.44  -0.354 0.723257
## Nom.du.départementYvelines                 9502.37    1118.36   8.497  < 2e-16
##
## (Intercept)                               ***
## Nom.du.départementAisne
## Nom.du.départementAllier
## Nom.du.départementAlpes-de-Haute-Provence
## Nom.du.départementAlpes-Maritimes         **
## Nom.du.départementArdèche
## Nom.du.départementArdennes
```

```
## Nom.du.départementAriège
## Nom.du.départementAube
## Nom.du.départementAude
## Nom.du.départementAveyron
## Nom.du.départementBas-Rhin
## Nom.du.départementBouches-du-Rhône        *
## Nom.du.départementCalvados                *
## Nom.du.départementCantal                  .
## Nom.du.départementCharente
## Nom.du.départementCharente-Maritime
## Nom.du.départementCher
## Nom.du.départementCorrèze
## Nom.du.départementCorse-du-Sud
## Nom.du.départementCôte-d'Or
## Nom.du.départementCôtes-d'Armor
## Nom.du.départementCreuse
## Nom.du.départementDeux-Sèvres
## Nom.du.départementDordogne
## Nom.du.départementDoubs
## Nom.du.départementDrôme
## Nom.du.départementEssonne                 ***
## Nom.du.départementEure
## Nom.du.départementEure-et-Loir            *
## Nom.du.départementFinistère
## Nom.du.départementGard
## Nom.du.départementGers
## Nom.du.départementGironde                 *
## Nom.du.départementHaut-Rhin
## Nom.du.départementHaute-Corse
## Nom.du.départementHaute-Garonne           ***
## Nom.du.départementHaute-Loire
## Nom.du.départementHaute-Marne
## Nom.du.départementHaute-Saône
## Nom.du.départementHaute-Savoie
## Nom.du.départementHaute-Vienne
## Nom.du.départementHautes-Alpes
## Nom.du.départementHautes-Pyrénées
## Nom.du.départementHauts-de-Seine          ***
## Nom.du.départementHérault
## Nom.du.départementIlle-et-Vilaine         **
## Nom.du.départementIndre                   .
## Nom.du.départementIndre-et-Loire
## Nom.du.départementIsère                   *
## Nom.du.départementJura
## Nom.du.départementLandes
## Nom.du.départementLoir-et-Cher            *
## Nom.du.départementLoire
## Nom.du.départementLoire-Atlantique        *
## Nom.du.départementLoiret
## Nom.du.départementLot
## Nom.du.départementLot-et-Garonne
## Nom.du.départementLozère
## Nom.du.départementMaine-et-Loire
## Nom.du.départementManche
## Nom.du.départementMarne
## Nom.du.départementMayenne
## Nom.du.départementMeurthe-et-Moselle
## Nom.du.départementMeuse
```

```
## Nom.du.départementMorbihan                .
## Nom.du.départementMoselle
## Nom.du.départementNièvre
## Nom.du.départementNord                    .
## Nom.du.départementOise                    *
## Nom.du.départementOrne
## Nom.du.départementParis                   ***
## Nom.du.départementPas-de-Calais
## Nom.du.départementPuy-de-Dôme
## Nom.du.départementPyrénées-Atlantiques
## Nom.du.départementPyrénées-Orientales
## Nom.du.départementRhône                   ***
## Nom.du.départementSaône-et-Loire
## Nom.du.départementSarthe                  .
## Nom.du.départementSavoie
## Nom.du.départementSeine-et-Marne          ***
## Nom.du.départementSeine-Maritime          .
## Nom.du.départementSeine-Saint-Denis       *
## Nom.du.départementSomme
## Nom.du.départementTarn
## Nom.du.départementTarn-et-Garonne
## Nom.du.départementTerritoire de Belfort   .
## Nom.du.départementVal-d'Oise              ***
## Nom.du.départementVal-de-Marne            ***
## Nom.du.départementVar
## Nom.du.départementVaucluse
## Nom.du.départementVendée
## Nom.du.départementVienne
## Nom.du.départementVosges
## Nom.du.départementYonne
## Nom.du.départementYvelines                ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13210 on 39658 degrees of freedom
##   (60246 observations deleted due to missingness)
## Multiple R-squared:  0.05555,    Adjusted R-squared:  0.05329
## F-statistic: 24.55 on 95 and 39658 DF,  p-value: < 2.2e-16
```

```
testGeographicNum("Age_2019", db = learn_code)# Il y a des différences significative entre de
partement
```

```
##
## Call:
## lm(formula = formula(paste0(variable, "~ Nom.du.département")),
##     data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.892 -16.600  -0.534  15.278  65.400
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         49.91727    0.99695  50.070  < 2e-16
## Nom.du.départementAisne             -0.15865    1.95221  -0.081 0.935228
## Nom.du.départementAllier            -0.74917    1.65972  -0.451 0.651714
## Nom.du.départementAlpes-de-Haute-Provence  4.36844    3.27415   1.334 0.182134
## Nom.du.départementAlpes-Maritimes    2.67238    1.09503   2.440 0.014670
## Nom.du.départementArdèche            0.97746    1.71710   0.569 0.569187
## Nom.du.départementArdennes           0.83273    1.88335   0.442 0.658381
## Nom.du.départementAriège             1.45209    2.16196   0.672 0.501804
## Nom.du.départementAube               1.18576    1.86270   0.637 0.524401
## Nom.du.départementAude               1.19551    1.59048   0.752 0.452254
## Nom.du.départementAveyron            1.80230    1.77631   1.015 0.310283
## Nom.du.départementBas-Rhin          -1.73786    1.07961  -1.610 0.107462
## Nom.du.départementBouches-du-Rhône  -0.82797    1.05207  -0.787 0.431286
## Nom.du.départementCalvados           0.52437    1.27993   0.410 0.682040
## Nom.du.départementCantal             0.44309    2.16196   0.205 0.837615
## Nom.du.départementCharente           1.57141    1.38496   1.135 0.256534
## Nom.du.départementCharente-Maritime  1.43806    1.25206   1.149 0.250740
## Nom.du.départementCher               1.07446    1.35570   0.793 0.428041
## Nom.du.départementCorrèze            2.30721    1.54381   1.494 0.135050
## Nom.du.départementCorse-du-Sud      -0.86637    1.57460  -0.550 0.582174
## Nom.du.départementCôte-d'Or         -0.29322    1.22923  -0.239 0.811460
## Nom.du.départementCôtes-d'Armor      3.97462    1.78941   2.221 0.026341
## Nom.du.départementCreuse             0.50645    2.11087   0.240 0.810387
## Nom.du.départementDeux-Sèvres        2.44594    1.39032   1.759 0.078535
## Nom.du.départementDordogne           0.58076    1.34091   0.433 0.664938
## Nom.du.départementDoubs              1.78594    1.21688   1.468 0.142205
## Nom.du.départementDrôme             -1.98828    1.48408  -1.340 0.180332
## Nom.du.départementEssonne           -3.65209    1.09574  -3.333 0.000859
## Nom.du.départementEure              -1.73288    1.31080  -1.322 0.186171
## Nom.du.départementEure-et-Loir       0.50304    1.27757   0.394 0.693770
## Nom.du.départementFinistère          1.66433    1.11191   1.497 0.134443
## Nom.du.départementGard               2.30527    1.10963   2.078 0.037756
## Nom.du.départementGers              -0.97768    1.93274  -0.506 0.612964
## Nom.du.départementGironde           -0.19545    1.04453  -0.187 0.851569
## Nom.du.départementHaut-Rhin         -1.02750    1.11538  -0.921 0.356940
## Nom.du.départementHaute-Corse        1.59536    1.57119   1.015 0.309925
## Nom.du.départementHaute-Garonne     -1.38329    1.06118  -1.304 0.192394
## Nom.du.départementHaute-Loire        2.48128    1.98848   1.248 0.212097
## Nom.du.départementHaute-Marne        0.86844    2.27206   0.382 0.702295
## Nom.du.départementHaute-Saône        1.41328    1.45898   0.969 0.332709
## Nom.du.départementHaute-Savoie      -1.58546    1.10779  -1.431 0.152378
## Nom.du.départementHaute-Vienne       2.53490    1.37184   1.848 0.064633
## Nom.du.départementHautes-Alpes      -3.34152    3.65686  -0.914 0.360841
## Nom.du.départementHautes-Pyrénées    0.70610    1.90958   0.370 0.711556
## Nom.du.départementHauts-de-Seine    -3.97233    1.07784  -3.685 0.000228
## Nom.du.départementHérault            0.20561    1.07297   0.192 0.848038
```

```
## Nom.du.départementIlle-et-Vilaine         -1.65346    1.10236  -1.500 0.133636
## Nom.du.départementIndre                    2.24743    1.48168   1.517 0.129317
## Nom.du.départementIndre-et-Loire          -0.07523    1.19954  -0.063 0.949995
## Nom.du.départementIsère                   -1.58271    1.07017  -1.479 0.139162
## Nom.du.départementJura                    -1.31256    1.43641  -0.914 0.360835
## Nom.du.départementLandes                   2.14738    1.36905   1.569 0.116763
## Nom.du.départementLoir-et-Cher             1.21963    1.34329   0.908 0.363910
## Nom.du.départementLoire                    0.39024    1.11152   0.351 0.725528
## Nom.du.départementLoire-Atlantique        -1.58007    1.06284  -1.487 0.137111
## Nom.du.départementLoiret                  -0.58734    1.18739  -0.495 0.620852
## Nom.du.départementLot                      2.44039    1.99390   1.224 0.220983
## Nom.du.départementLot-et-Garonne           2.81386    1.39905   2.011 0.044301
## Nom.du.départementLozère                   6.43458    2.92552   2.199 0.027847
## Nom.du.départementMaine-et-Loire          -0.29735    1.10509  -0.269 0.787873
## Nom.du.départementManche                   1.22616    1.34876   0.909 0.363299
## Nom.du.départementMarne                   -0.07162    1.51901  -0.047 0.962397
## Nom.du.départementMayenne                 -1.25341    1.64629  -0.761 0.446449
## Nom.du.départementMeurthe-et-Moselle      -0.68086    1.12313  -0.606 0.544371
## Nom.du.départementMeuse                   -0.49969    2.34156  -0.213 0.831014
## Nom.du.départementMorbihan                 2.65342    1.13786   2.332 0.019706
## Nom.du.départementMoselle                 -0.42852    1.08356  -0.395 0.692492
## Nom.du.départementNièvre                  -0.64849    1.47463  -0.440 0.660110
## Nom.du.départementNord                    -2.74094    1.03941  -2.637 0.008365
## Nom.du.départementOise                    -2.11194    1.12527  -1.877 0.060545
## Nom.du.départementOrne                    -0.26061    1.53479  -0.170 0.865167
## Nom.du.départementParis                   -4.30851    1.05296  -4.092 4.28e-05
## Nom.du.départementPas-de-Calais           -1.04899    1.07067  -0.980 0.327211
## Nom.du.départementPuy-de-Dôme             -1.85652    1.42411  -1.304 0.192361
## Nom.du.départementPyrénées-Atlantiques     2.10428    1.21784   1.728 0.084012
## Nom.du.départementPyrénées-Orientales      1.44959    1.48408   0.977 0.328689
## Nom.du.départementRhône                   -2.31748    1.04700  -2.213 0.026870
## Nom.du.départementSaône-et-Loire           0.69563    1.21736   0.571 0.567713
## Nom.du.départementSarthe                   0.84450    1.45790   0.579 0.562418
## Nom.du.départementSavoie                  -1.06989    1.62310  -0.659 0.509794
## Nom.du.départementSeine-et-Marne          -4.11613    1.09049  -3.775 0.000160
## Nom.du.départementSeine-Maritime          -0.97292    1.06080  -0.917 0.359064
## Nom.du.départementSeine-Saint-Denis       -5.85094    1.07855  -5.425 5.81e-08
## Nom.du.départementSomme                   -1.88583    1.88761  -0.999 0.317770
## Nom.du.départementTarn                     3.03118    1.54844   1.958 0.050284
## Nom.du.départementTarn-et-Garonne          3.22730    1.85869   1.736 0.082509
## Nom.du.départementTerritoire de Belfort    3.53042    1.64411   2.147 0.031771
## Nom.du.départementVal-d'Oise              -3.94955    1.10335  -3.580 0.000344
## Nom.du.départementVal-de-Marne            -4.78172    1.08881  -4.392 1.13e-05
## Nom.du.départementVar                      1.24048    1.09813   1.130 0.258635
## Nom.du.départementVaucluse                 0.80808    1.96740   0.411 0.681268
## Nom.du.départementVendée                   0.61093    1.42875   0.428 0.668946
## Nom.du.départementVienne                   0.50783    1.34938   0.376 0.706665
## Nom.du.départementVosges                  -1.04708    1.71984  -0.609 0.542641
## Nom.du.départementYonne                   -0.09563    1.34631  -0.071 0.943372
## Nom.du.départementYvelines                -3.18659    1.08559  -2.935 0.003332
##
## (Intercept)                               ***
## Nom.du.départementAisne
## Nom.du.départementAllier
## Nom.du.départementAlpes-de-Haute-Provence
## Nom.du.départementAlpes-Maritimes         *
## Nom.du.départementArdèche
## Nom.du.départementArdennes
```

```
## Nom.du.départementAriège
## Nom.du.départementAube
## Nom.du.départementAude
## Nom.du.départementAveyron
## Nom.du.départementBas-Rhin
## Nom.du.départementBouches-du-Rhône
## Nom.du.départementCalvados
## Nom.du.départementCantal
## Nom.du.départementCharente
## Nom.du.départementCharente-Maritime
## Nom.du.départementCher
## Nom.du.départementCorrèze
## Nom.du.départementCorse-du-Sud
## Nom.du.départementCôte-d'Or
## Nom.du.départementCôtes-d'Armor           *
## Nom.du.départementCreuse
## Nom.du.départementDeux-Sèvres             .
## Nom.du.départementDordogne
## Nom.du.départementDoubs
## Nom.du.départementDrôme
## Nom.du.départementEssonne                 ***
## Nom.du.départementEure
## Nom.du.départementEure-et-Loir
## Nom.du.départementFinistère
## Nom.du.départementGard                    *
## Nom.du.départementGers
## Nom.du.départementGironde
## Nom.du.départementHaut-Rhin
## Nom.du.départementHaute-Corse
## Nom.du.départementHaute-Garonne
## Nom.du.départementHaute-Loire
## Nom.du.départementHaute-Marne
## Nom.du.départementHaute-Saône
## Nom.du.départementHaute-Savoie
## Nom.du.départementHaute-Vienne            .
## Nom.du.départementHautes-Alpes
## Nom.du.départementHautes-Pyrénées
## Nom.du.départementHauts-de-Seine          ***
## Nom.du.départementHérault
## Nom.du.départementIlle-et-Vilaine
## Nom.du.départementIndre
## Nom.du.départementIndre-et-Loire
## Nom.du.départementIsère
## Nom.du.départementJura
## Nom.du.départementLandes
## Nom.du.départementLoir-et-Cher
## Nom.du.départementLoire
## Nom.du.départementLoire-Atlantique
## Nom.du.départementLoiret
## Nom.du.départementLot
## Nom.du.départementLot-et-Garonne          *
## Nom.du.départementLozère                  *
## Nom.du.départementMaine-et-Loire
## Nom.du.départementManche
## Nom.du.départementMarne
## Nom.du.départementMayenne
## Nom.du.départementMeurthe-et-Moselle
## Nom.du.départementMeuse
```

```
## Nom.du.départementMorbihan                *
## Nom.du.départementMoselle
## Nom.du.départementNièvre
## Nom.du.départementNord                    **
## Nom.du.départementOise                    .
## Nom.du.départementOrne
## Nom.du.départementParis                   ***
## Nom.du.départementPas-de-Calais
## Nom.du.départementPuy-de-Dôme
## Nom.du.départementPyrénées-Atlantiques    .
## Nom.du.départementPyrénées-Orientales
## Nom.du.départementRhône                   *
## Nom.du.départementSaône-et-Loire
## Nom.du.départementSarthe
## Nom.du.départementSavoie
## Nom.du.départementSeine-et-Marne          ***
## Nom.du.départementSeine-Maritime
## Nom.du.départementSeine-Saint-Denis       ***
## Nom.du.départementSomme
## Nom.du.départementTarn                    .
## Nom.du.départementTarn-et-Garonne         .
## Nom.du.départementTerritoire de Belfort   *
## Nom.du.départementVal-d'Oise              ***
## Nom.du.départementVal-de-Marne            ***
## Nom.du.départementVar
## Nom.du.départementVaucluse
## Nom.du.départementVendée
## Nom.du.départementVienne
## Nom.du.départementVosges
## Nom.du.départementYonne
## Nom.du.départementYvelines                **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.21 on 99904 degrees of freedom
## Multiple R-squared:  0.01073,    Adjusted R-squared:  0.009785
## F-statistic:  11.4 on 95 and 99904 DF,  p-value: < 2.2e-16
```

```
testGeographicNum("WORKING_HOURS", db = learn_code)# Il y a des différences significative ent
re departement
```

```
##
## Call:
## lm(formula = formula(paste0(variable, "~ Nom.du.département")),
##     data = db)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1721.4  -213.2   198.7   277.7  1510.0
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          1503.1688    42.1505  35.662  < 2e-16
## Nom.du.départementAisne               164.4112    86.3829   1.903 0.057011
## Nom.du.départementAllier               73.6943    75.3043   0.979 0.327774
## Nom.du.départementAlpes-de-Haute-Provence -186.9187   159.5794  -1.171 0.241477
## Nom.du.départementAlpes-Maritimes      91.6450    46.8346   1.957 0.050380
## Nom.du.départementArdèche              91.7001    80.2296   1.143 0.253057
## Nom.du.départementArdennes             -8.6354    80.7121  -0.107 0.914797
## Nom.du.départementAriège               58.7706   101.9354   0.577 0.564248
## Nom.du.départementAube                 57.6090    83.9099   0.687 0.492365
## Nom.du.départementAude                 90.6734    74.2767   1.221 0.222187
## Nom.du.départementAveyron              88.7271    87.7434   1.011 0.311921
## Nom.du.départementBas-Rhin             81.9725    45.4038   1.805 0.071018
## Nom.du.départementBouches-du-Rhône     56.1655    44.4461   1.264 0.206353
## Nom.du.départementCalvados            155.8154    53.8954   2.891 0.003841
## Nom.du.départementCantal              153.5960   100.6850   1.526 0.127140
## Nom.du.départementCharente            118.7887    61.5853   1.929 0.053757
## Nom.du.départementCharente-Maritime   116.4650    54.2816   2.146 0.031913
## Nom.du.départementCher                124.6139    59.5172   2.094 0.036289
## Nom.du.départementCorrèze             224.7274    73.9489   3.039 0.002376
## Nom.du.départementCorse-du-Sud        -20.1982    77.1820  -0.262 0.793558
## Nom.du.départementCôte-d'Or            79.3144    52.3166   1.516 0.129516
## Nom.du.départementCôtes-d'Armor       140.2641    77.5852   1.808 0.070634
## Nom.du.départementCreuse              -75.5918   112.7387  -0.671 0.502540
## Nom.du.départementDeux-Sèvres         159.4843    60.9135   2.618 0.008843
## Nom.du.départementDordogne            -13.1989    59.0688  -0.223 0.823187
## Nom.du.départementDoubs               128.0909    52.6697   2.432 0.015022
## Nom.du.départementDrôme               111.2370    61.9401   1.796 0.072521
## Nom.du.départementEssonne              97.4601    45.4420   2.145 0.031982
## Nom.du.départementEure                153.3701    55.4499   2.766 0.005679
## Nom.du.départementEure-et-Loir        168.4696    55.1880   3.053 0.002270
## Nom.du.départementFinistère            47.7111    47.7379   0.999 0.317588
## Nom.du.départementGard                 16.0159    47.6645   0.336 0.736863
## Nom.du.départementGers                106.5665   100.6850   1.058 0.289872
## Nom.du.départementGironde              90.6833    44.1966   2.052 0.040194
## Nom.du.départementHaut-Rhin           109.6319    47.2619   2.320 0.020364
## Nom.du.départementHaute-Corse         116.0665    77.1820   1.504 0.132640
## Nom.du.départementHaute-Garonne        64.5442    44.6635   1.445 0.148432
## Nom.du.départementHaute-Loire         100.4312    99.4916   1.009 0.312768
## Nom.du.départementHaute-Marne         271.2201   132.5491   2.046 0.040745
## Nom.du.départementHaute-Saône         115.6565    67.3539   1.717 0.085960
## Nom.du.départementHaute-Savoie         97.0279    46.2571   2.098 0.035949
## Nom.du.départementHaute-Vienne        156.3809    60.4916   2.585 0.009737
## Nom.du.départementHautes-Alpes       -159.7841   153.7640  -1.039 0.298740
## Nom.du.départementHautes-Pyrénées      -0.8354    95.2133  -0.009 0.992999
## Nom.du.départementHauts-de-Seine      192.1771    44.8793   4.282 1.86e-05
## Nom.du.départementHérault              39.1022    45.8145   0.853 0.393392
```

```
## Nom.du.départementIlle-et-Vilaine          127.1892    46.2471   2.750 0.005958
## Nom.du.départementIndre                    111.7670    66.2165   1.688 0.091438
## Nom.du.départementIndre-et-Loire           109.3312    51.0672   2.141 0.032286
## Nom.du.départementIsère                     95.6979    44.9403   2.129 0.033224
## Nom.du.départementJura                     194.0468    65.0173   2.985 0.002842
## Nom.du.départementLandes                    51.3280    60.0887   0.854 0.392997
## Nom.du.départementLoir-et-Cher             148.5642    57.1400   2.600 0.009326
## Nom.du.départementLoire                     91.6287    47.2619   1.939 0.052540
## Nom.du.départementLoire-Atlantique          71.2896    44.7191   1.594 0.110907
## Nom.du.départementLoiret                   123.5455    49.8910   2.476 0.013279
## Nom.du.départementLot                      -35.0259   109.2202  -0.321 0.748447
## Nom.du.départementLot-et-Garonne           151.1027    61.7021   2.449 0.014333
## Nom.du.départementLozère                   320.8312   182.6524   1.757 0.079009
## Nom.du.départementMaine-et-Loire            75.0634    46.6288   1.610 0.107448
## Nom.du.départementManche                   111.4667    57.0721   1.953 0.050817
## Nom.du.départementMarne                    152.0620    62.9550   2.415 0.015722
## Nom.du.départementMayenne                  198.4359    71.2888   2.784 0.005379
## Nom.du.départementMeurthe-et-Moselle        54.8483    47.5580   1.153 0.248797
## Nom.du.départementMeuse                     10.0620   112.7387   0.089 0.928883
## Nom.du.départementMorbihan                  94.4493    48.5837   1.944 0.051896
## Nom.du.départementMoselle                   92.5512    45.7868   2.021 0.043250
## Nom.du.départementNièvre                   100.9668    64.6971   1.561 0.118624
## Nom.du.départementNord                      82.9487    43.8389   1.892 0.058482
## Nom.du.départementOise                      80.8444    47.3569   1.707 0.087806
## Nom.du.départementOrne                     100.3612    67.9657   1.477 0.139779
## Nom.du.départementParis                    129.2805    44.1660   2.927 0.003423
## Nom.du.départementPas-de-Calais             79.0307    45.3882   1.741 0.081653
## Nom.du.départementPuy-de-Dôme               99.7293    59.8940   1.665 0.095901
## Nom.du.départementPyrénées-Atlantiques      68.9122    52.7030   1.308 0.191031
## Nom.du.départementPyrénées-Orientales       24.4070    68.1766   0.358 0.720348
## Nom.du.départementRhône                     99.2055    44.0962   2.250 0.024470
## Nom.du.départementSaône-et-Loire            55.0268    51.8438   1.061 0.288517
## Nom.du.départementSarthe                   181.7473    61.3559   2.962 0.003057
## Nom.du.départementSavoie                     9.2049    70.0034   0.131 0.895387
## Nom.du.départementSeine-et-Marne           118.0980    45.3604   2.604 0.009230
## Nom.du.départementSeine-Maritime            82.5145    44.9016   1.838 0.066118
## Nom.du.départementSeine-Saint-Denis         87.9398    45.0735   1.951 0.051060
## Nom.du.départementSomme                    -25.5084    84.4997  -0.302 0.762749
## Nom.du.départementTarn                      26.8530    69.7605   0.385 0.700290
## Nom.du.départementTarn-et-Garonne          266.3201    89.9651   2.960 0.003076
## Nom.du.départementTerritoire de Belfort    215.9442    79.7598   2.707 0.006784
## Nom.du.départementVal-d'Oise               110.5616    45.8877   2.409 0.015984
## Nom.du.départementVal-de-Marne              89.8007    45.4581   1.975 0.048223
## Nom.du.départementVar                       33.0185    47.0036   0.702 0.482393
## Nom.du.départementVaucluse                 -14.9688    94.2515  -0.159 0.873814
## Nom.du.départementVendée                    51.8592    61.3559   0.845 0.397993
## Nom.du.départementVienne                    80.2858    59.1566   1.357 0.174733
## Nom.du.départementVosges                   126.0170    76.4043   1.649 0.099085
## Nom.du.départementYonne                     67.0932    58.8960   1.139 0.254635
## Nom.du.départementYvelines                 158.5557    45.2828   3.501 0.000463
##
## (Intercept)                               ***
## Nom.du.départementAisne                   .
## Nom.du.départementAllier
## Nom.du.départementAlpes-de-Haute-Provence
## Nom.du.départementAlpes-Maritimes         .
## Nom.du.départementArdèche
## Nom.du.départementArdennes
```

```
## Nom.du.départementAriège
## Nom.du.départementAube
## Nom.du.départementAude
## Nom.du.départementAveyron
## Nom.du.départementBas-Rhin               .
## Nom.du.départementBouches-du-Rhône
## Nom.du.départementCalvados               **
## Nom.du.départementCantal
## Nom.du.départementCharente               .
## Nom.du.départementCharente-Maritime      *
## Nom.du.départementCher                   *
## Nom.du.départementCorrèze                **
## Nom.du.départementCorse-du-Sud
## Nom.du.départementCôte-d'Or
## Nom.du.départementCôtes-d'Armor          .
## Nom.du.départementCreuse
## Nom.du.départementDeux-Sèvres            **
## Nom.du.départementDordogne
## Nom.du.départementDoubs                  *
## Nom.du.départementDrôme                  .
## Nom.du.départementEssonne                *
## Nom.du.départementEure                   **
## Nom.du.départementEure-et-Loir           **
## Nom.du.départementFinistère
## Nom.du.départementGard
## Nom.du.départementGers
## Nom.du.départementGironde                *
## Nom.du.départementHaut-Rhin              *
## Nom.du.départementHaute-Corse
## Nom.du.départementHaute-Garonne
## Nom.du.départementHaute-Loire
## Nom.du.départementHaute-Marne            *
## Nom.du.départementHaute-Saône            .
## Nom.du.départementHaute-Savoie           *
## Nom.du.départementHaute-Vienne           **
## Nom.du.départementHautes-Alpes
## Nom.du.départementHautes-Pyrénées
## Nom.du.départementHauts-de-Seine         ***
## Nom.du.départementHérault
## Nom.du.départementIlle-et-Vilaine        **
## Nom.du.départementIndre                  .
## Nom.du.départementIndre-et-Loire         *
## Nom.du.départementIsère                  *
## Nom.du.départementJura                   **
## Nom.du.départementLandes
## Nom.du.départementLoir-et-Cher           **
## Nom.du.départementLoire                  .
## Nom.du.départementLoire-Atlantique
## Nom.du.départementLoiret                 *
## Nom.du.départementLot
## Nom.du.départementLot-et-Garonne         *
## Nom.du.départementLozère                 .
## Nom.du.départementMaine-et-Loire
## Nom.du.départementManche                 .
## Nom.du.départementMarne                  *
## Nom.du.départementMayenne                **
## Nom.du.départementMeurthe-et-Moselle
## Nom.du.départementMeuse
```

```
## Nom.du.départementMorbihan               .
## Nom.du.départementMoselle                *
## Nom.du.départementNièvre
## Nom.du.départementNord                   .
## Nom.du.départementOise                   .
## Nom.du.départementOrne
## Nom.du.départementParis                  **
## Nom.du.départementPas-de-Calais          .
## Nom.du.départementPuy-de-Dôme            .
## Nom.du.départementPyrénées-Atlantiques
## Nom.du.départementPyrénées-Orientales
## Nom.du.départementRhône                  *
## Nom.du.départementSaône-et-Loire
## Nom.du.départementSarthe                 **
## Nom.du.départementSavoie
## Nom.du.départementSeine-et-Marne         **
## Nom.du.départementSeine-Maritime         .
## Nom.du.départementSeine-Saint-Denis      .
## Nom.du.départementSomme
## Nom.du.départementTarn
## Nom.du.départementTarn-et-Garonne        **
## Nom.du.départementTerritoire de Belfort  **
## Nom.du.départementVal-d'Oise             *
## Nom.du.départementVal-de-Marne           *
## Nom.du.départementVar
## Nom.du.départementVaucluse
## Nom.du.départementVendée
## Nom.du.départementVienne
## Nom.du.départementVosges                 .
## Nom.du.départementYonne
## Nom.du.départementYvelines               ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 533.2 on 39514 degrees of freedom
##   (60390 observations deleted due to missingness)
## Multiple R-squared:  0.005636,   Adjusted R-squared:  0.003245
## F-statistic: 2.357 on 95 and 39514 DF,  p-value: 2.261e-12
```

```
testGeographicCat("Is_student", db = learn_code)# Dépendance entre la variable et les départe
ments
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  tab_cont
## X-squared = 447.52, df = NA, p-value = 0.0004998
```

```
testGeographicCat("SEX", db = learn_code)# il n'y a pas de Dépendance entre la variable et le
s départements
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  tab_cont
## X-squared = 104.22, df = NA, p-value = 0.2579
```

```
testGeographicCat("Occupation_42", db = learn_code)# Dépendance entre la variable et les dépa
rtements
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  tab_cont
## X-squared = 12586, df = NA, p-value = 0.0004998
```

```
testGeographicCat("N2", db = learn_code)# Dépendance entre la variable et les départements
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  tab_cont
## X-squared = 6369.9, df = NA, p-value = 0.0004998
```

```
testGeographicCat("Terms_of_emp", db = learn_code)# Dépendance entre la variable et les dépar
tements
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  tab_cont
## X-squared = 623.19, df = NA, p-value = 0.0004998
```

```
testGeographicCat("JOB_CONDITION", db = learn_code)# Dépendance entre la variable et les dépa
rtements
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  tab_cont
## X-squared = 787.45, df = NA, p-value = 0.0004998
```

```
testGeographicCat("highest_degree", db = learn_code)# Dépendance entre la variable et les dép
artements
```

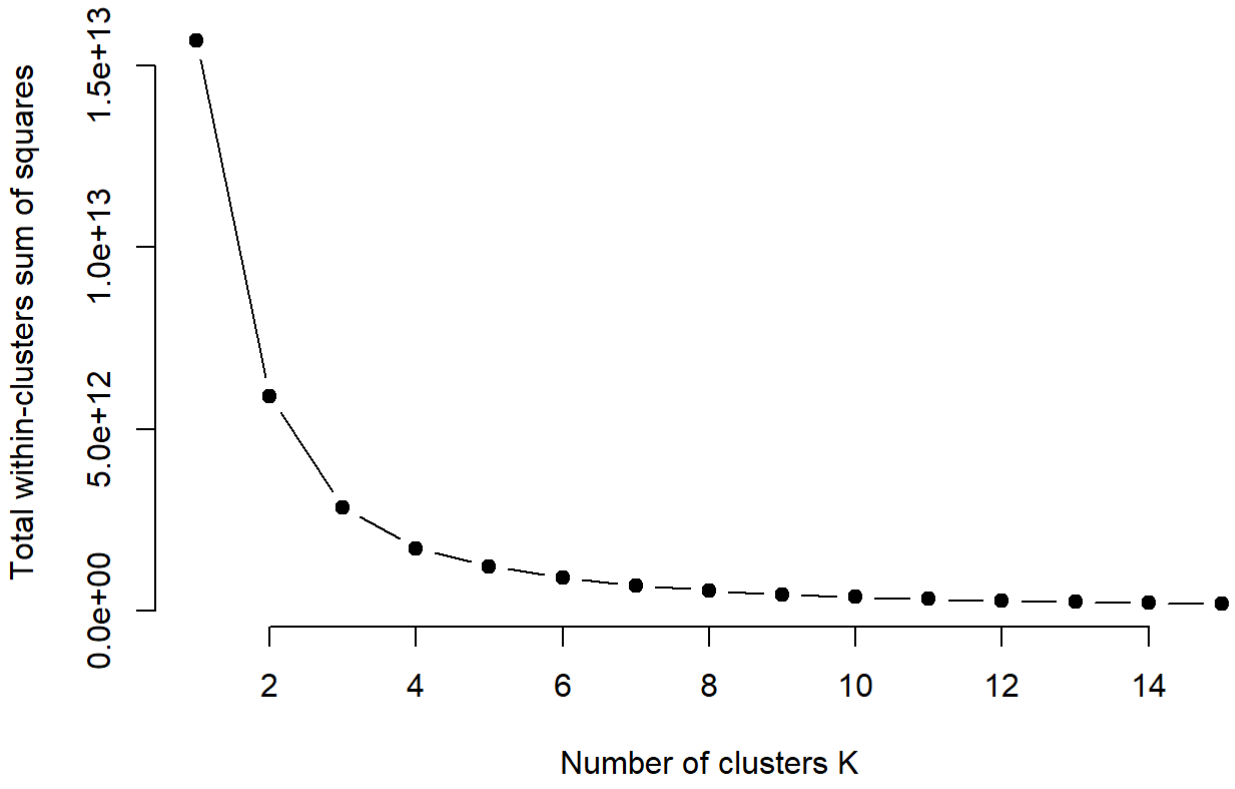
```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  tab_cont
## X-squared = 8368.6, df = NA, p-value = 0.0004998
```

3. Clusters en utilisant le dataset imputé sur des variables numéiques

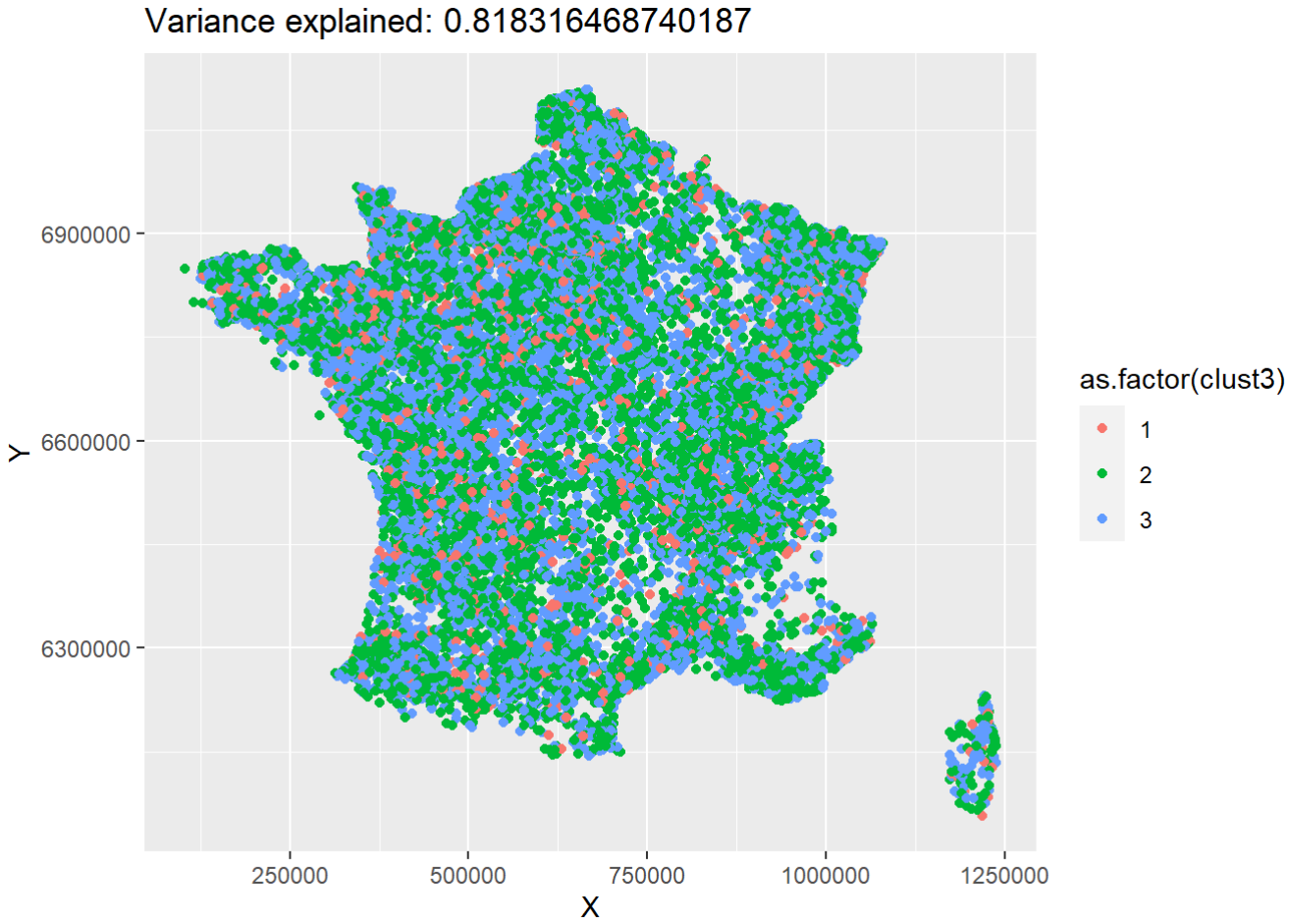
3.1 K-means :

le K optimal après le digrame du coude est de 4 clusters. En utilisant uniquement les variables numériques, les clusters ne sont pas observablés.

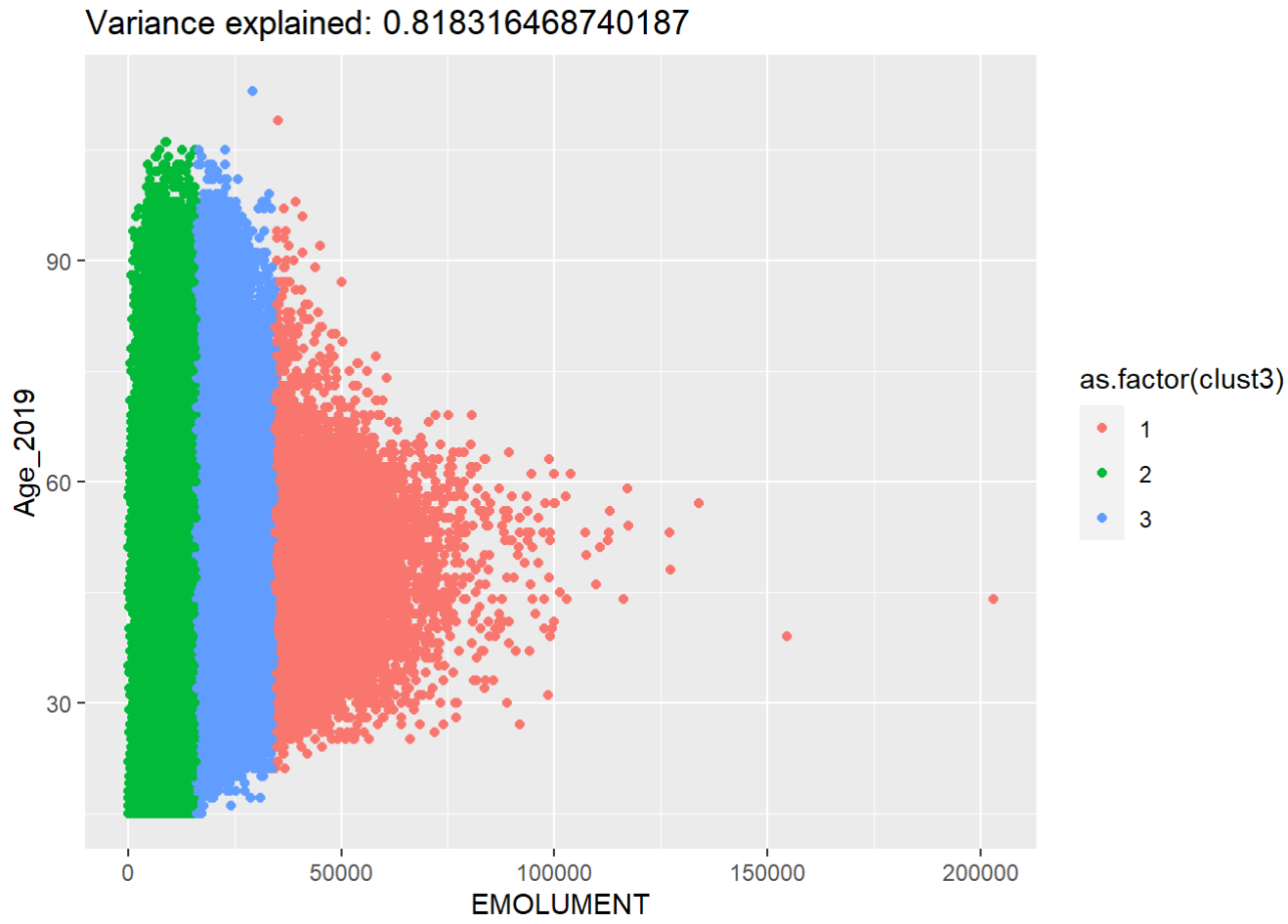
```
set.seed(123)
# Compute and plot wss for k = 2 to k = 15.
k.max = 15
data = db_imp[,c("Age_2019", "EMOLUMENT", "WORKING_HOURS")]
# wss = sapply(1:k.max,
#              function(k){kmeans(data, k, nstart = 50, iter.max = 15 )$tot.withinss})
# save(wss, file = "wss.RData")
load("wss.RData")
plot(1:k.max, wss,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```



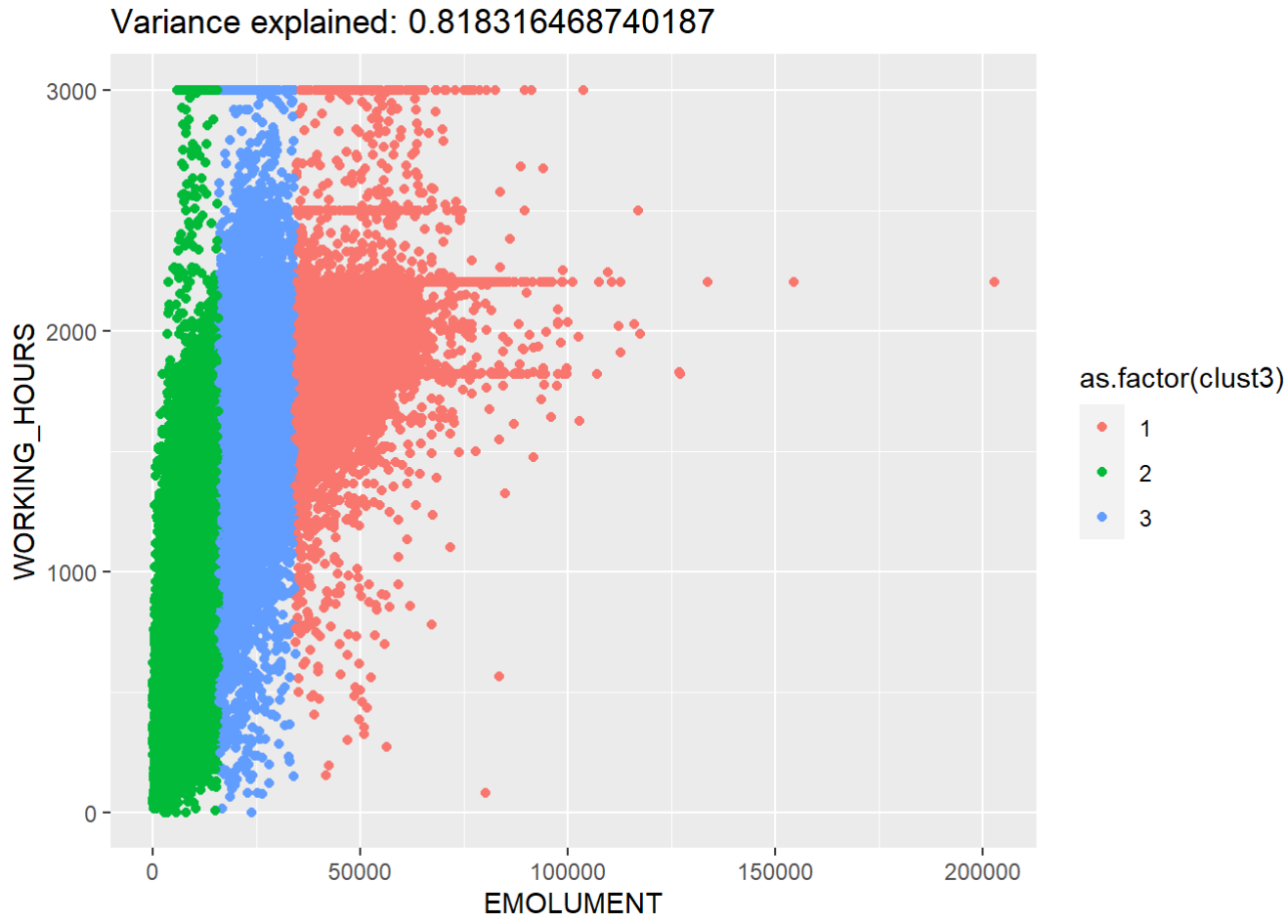
```
db_imp_num = db_imp[,c("Age_2019", "EMOLUMENT", "WORKING_HOURS")]
clust3 = kmeans(db_imp_num, 3)
db_imp_num$clust3 = clust3$cluster
var_per3 = clust3$betweenss / clust3$totss
clust4 = kmeans(db_imp_num, 4)
db_imp_num$clust4 = clust4$cluster
var_per4 = clust4$betweenss / clust4$totss
clust5 = kmeans(db_imp_num, 10)
db_imp_num$clust5 = clust5$cluster
var_per5 = clust5$betweenss / clust5$totss
db_imp_num$X = db_imp$X
db_imp_num$Y = db_imp$Y

ggplot(db_imp_num, aes(X, Y, color = as.factor(clust3))) +
    geom_point() + ggtitle(paste0("Variance explained: ", var_per3))
```



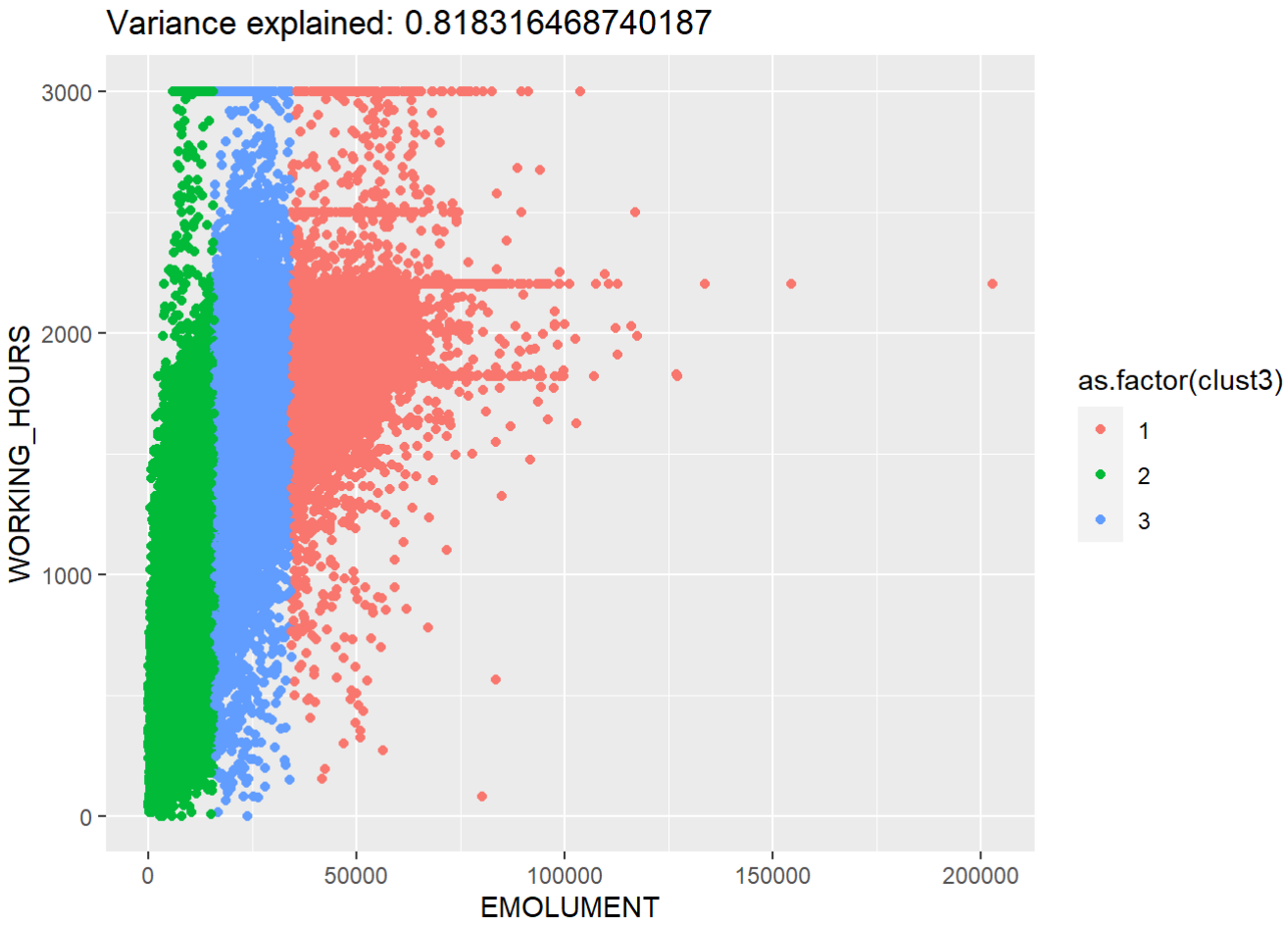

```
ggplot(db_imp_num, aes(EMOLUMENT, Age_2019, color = as.factor(clust3))) +
    geom_point() + ggtitle(paste0("Variance explained: ", var_per3))
```



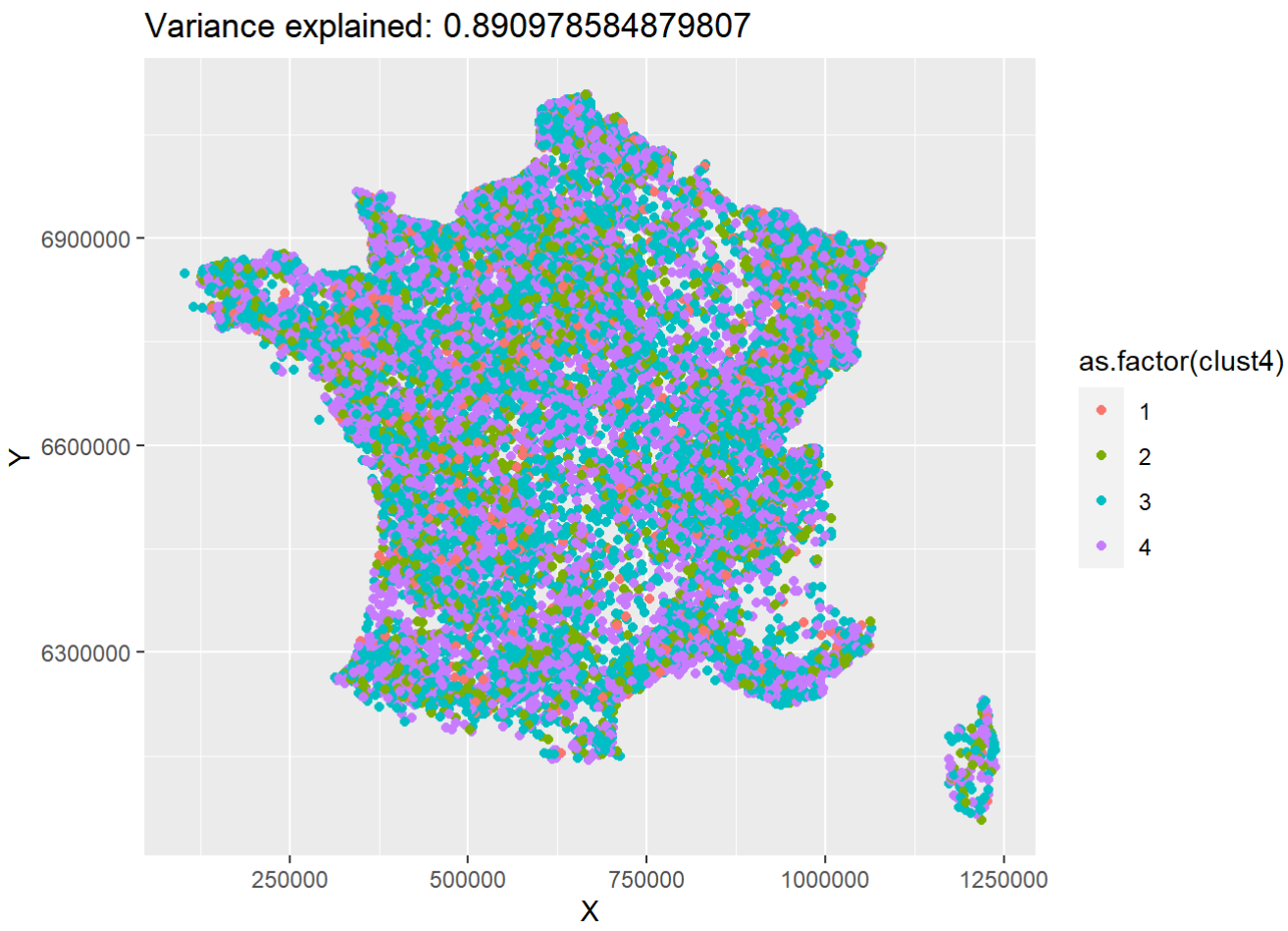

```
ggplot(db_imp_num, aes(EMOLUMENT, WORKING_HOURS, color = as.factor(clust3))) +
    geom_point() + ggtitle(paste0("Variance explained: ", var_per3))
```



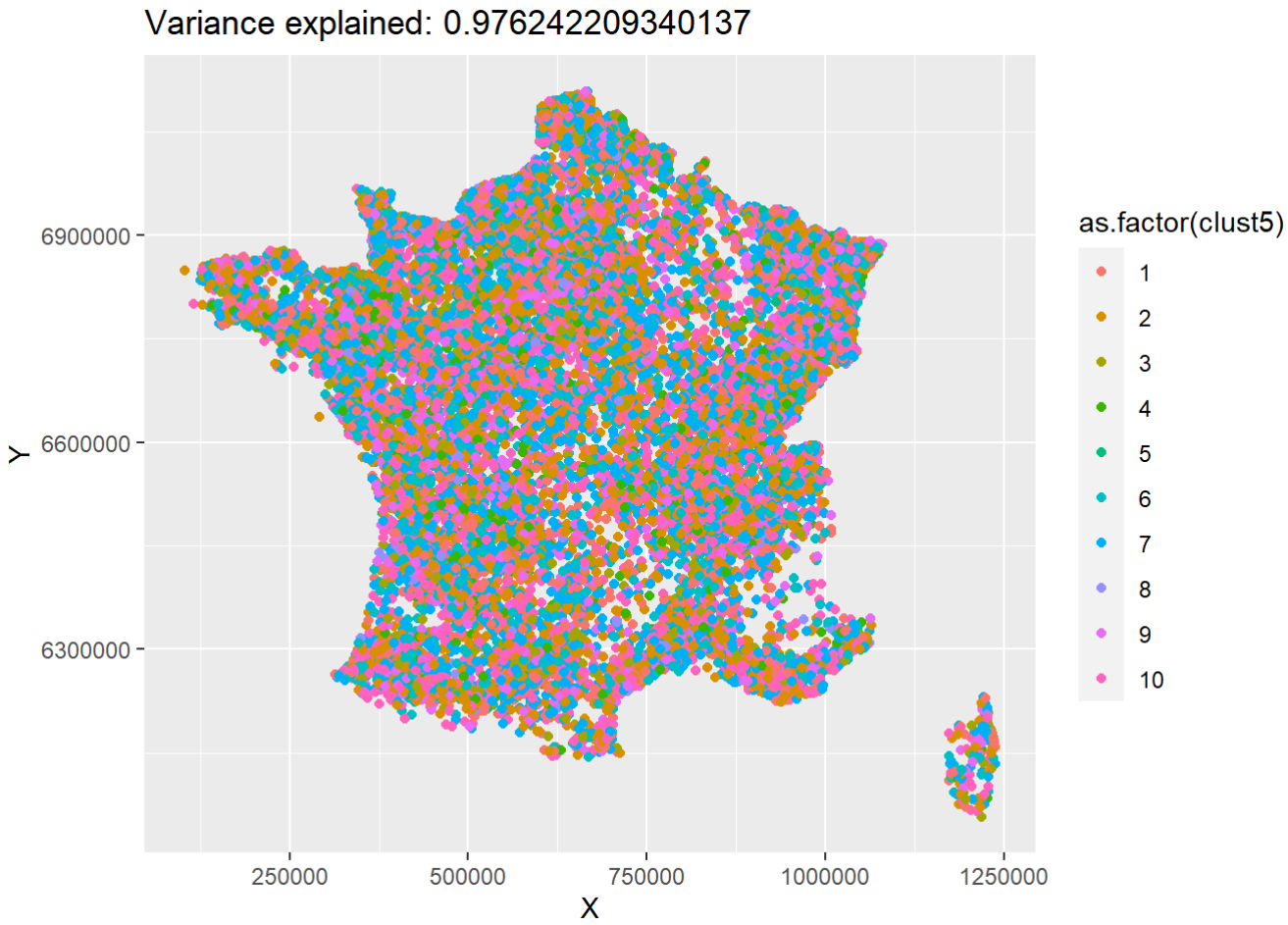
```
ggplot(db_imp_num, aes(EMOLUMENT, WORKING_HOURS, color = as.factor(clust3))) +
    geom_point() + ggtitle(paste0("Variance explained: ", var_per3))
```




```
ggplot(db_imp_num, aes(X, Y, color = as.factor(clust4))) +
    geom_point() + ggtitle(paste0("Variance explained: ", var_per4))
```




```
ggplot(db_imp_num, aes(X, Y, color = as.factor(clust5))) +
    geom_point() + ggtitle(paste0("Variance explained: ", var_per5))
```



#3.2 Hierarchical clustering : Grâce au graphe du coude le nombre de cluster optimal est de 5 en utilisant cette méthode. Afin de pouvoir conturné les problèmes de RAM l'algoritme est tester avec 1% du dataset.

```
library(dendextend)
```

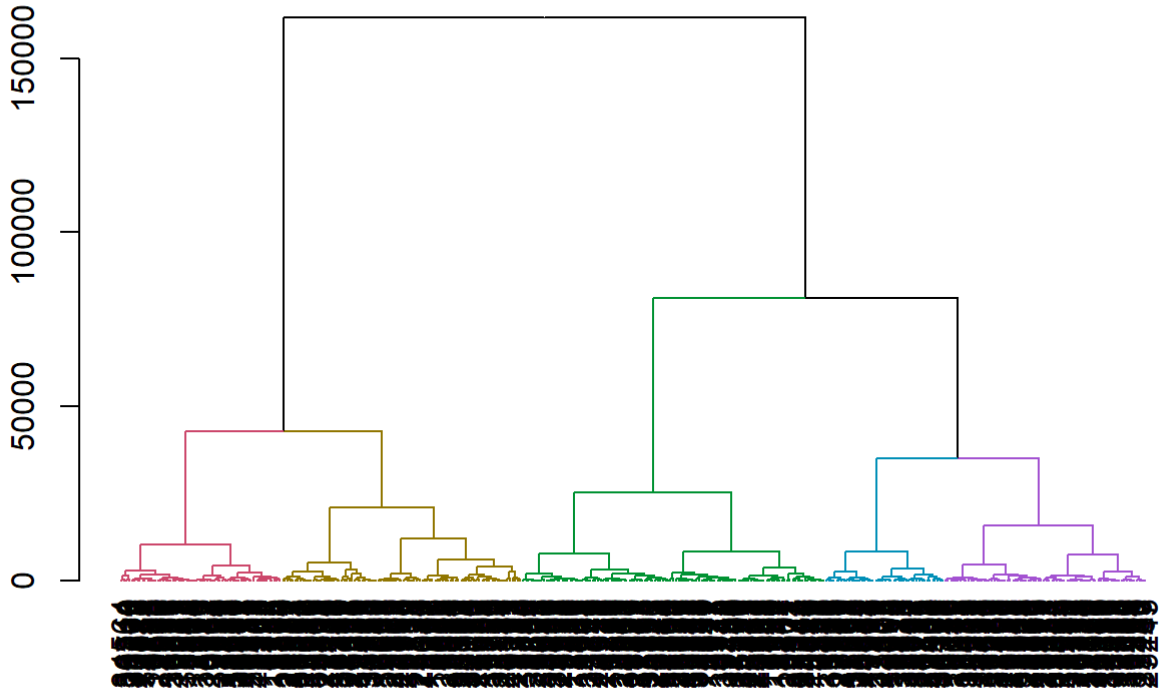
```
## Warning: package 'dendextend' was built under R version 4.0.5
```

```
##
## ---------------------
## Welcome to dendextend version 1.15.1
## Type citation('dendextend') for how to cite the package.
##
## Type browseVignettes(package = 'dendextend') for the package vignette.
## The github page is: https://github.com/talgalili/dendextend/
##
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/i
ssues
## Or contact: <tal.galili@gmail.com>
##
##  To suppress this message use:  suppressPackageStartupMessages(library(dendextend))
## ---------------------
```

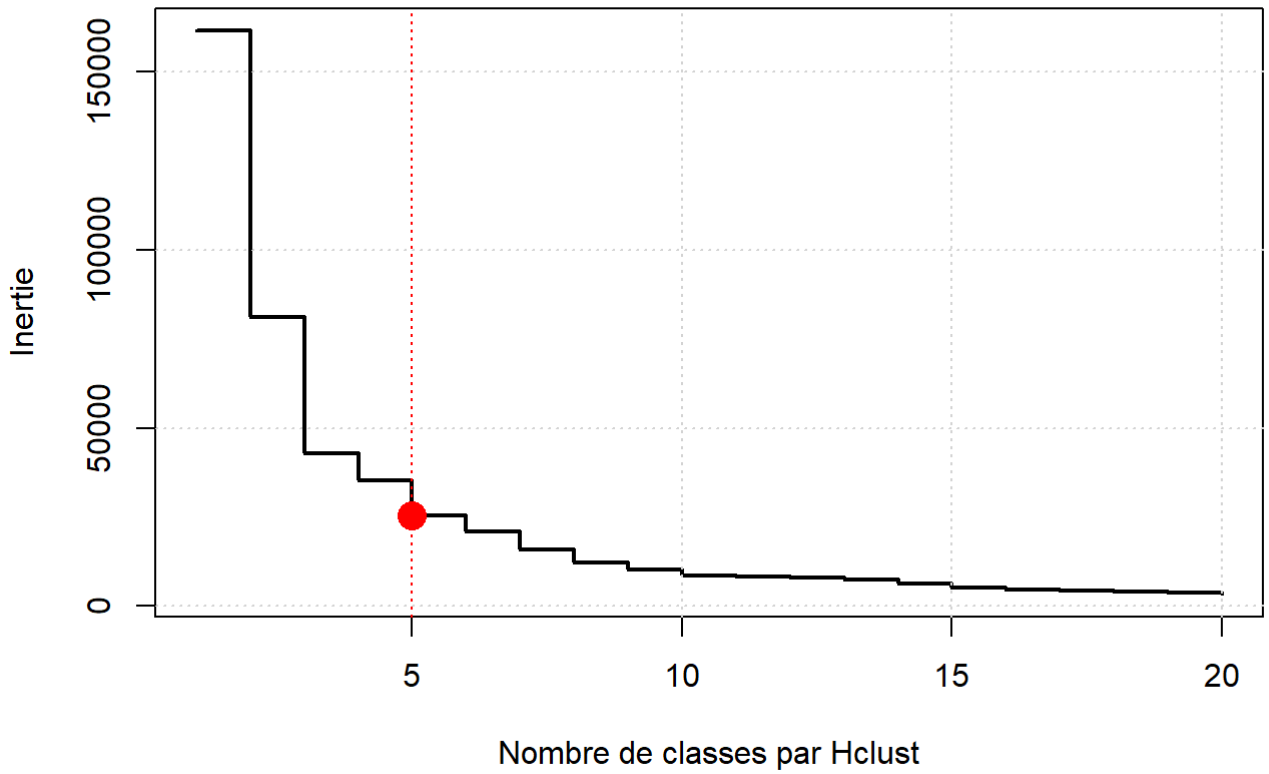
```
##
## Attaching package: 'dendextend'
```

```
## The following object is masked from 'package:stats':
##
##     cutree
```

```
set.seed(123)
db_imp_num = db_job[, c("Age_2019", "EMOLUMENT", "WORKING_HOURS")]
ind = sample(1:nrow(db_imp_num),
             size = floor(0.01*nrow(db_imp_num)),
             replace = FALSE)
db_clust = db_imp_num[ind,]
d = dist(db_clust, method = "euclidean")# Distance matrix
hc = hclust(d, method = "ward.D2")
dend = as.dendrogram(hc) %>% color_branches(k = 5)
plot(dend)
```




```
#db_clust$hc = labels
inertie <- sort(hc$height, decreasing = TRUE)
plot(inertie[1:20],
     type = "s", xlab = "Nombre de classes par Hclust", ylab = "Inertie",lwd=2)
grid()
k = 5
abline(v=k,col="red",lty=3)
points(k,inertie[k],pch=16,cex=2,col="red")
```



#3.3 Performing DBSCAN : Afin de pouvoir conturné les problèmes de RAM l'algoritme est tester avec 10% du dataset -Avec un epsilon de 100 (en couluers) il est possible d'utiliser cette méthode pour la construction de clusters mais il restent beaucoup des observation sans clusteriser (point en noir) _Avec un epsilon de 1000 (en rouge) il n'est pas possible d'utiliser cette méthode car un suele clusters est réalisé par l'altgortime.

Ce résult permet de penser qu'il conviendra diviser le dataset afin d'avoir deux dataset avec des caractérisque similaire donc à l'intérieur de chaque data l'information sera plus comparablé.

```
library(fpc)
```

```
## Warning: package 'fpc' was built under R version 4.0.5
```

```
library(dbscan)
```

```
## Warning: package 'dbscan' was built under R version 4.0.5
```

```
##
## Attaching package: 'dbscan'
```

```
## The following object is masked from 'package:fpc':
##
##     dbscan
```

```
set.seed(123)
ind = sample(1:nrow(db_imp_num),
             size = floor(0.1*nrow(db_imp_num)),
             replace = FALSE)
db_dbscan = db_imp_num[ind,]
dbs = fpc::dbscan(db_dbscan, 100, MinPts = 5)
plot(dbs, db_dbscan, main = "DBSCAN", frame = FALSE)
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "frame" n'est pas
## un paramètre graphique
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" n'est pas un
## paramètre graphique
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "frame" n'est pas
## un paramètre graphique
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" n'est pas un
## paramètre graphique
```

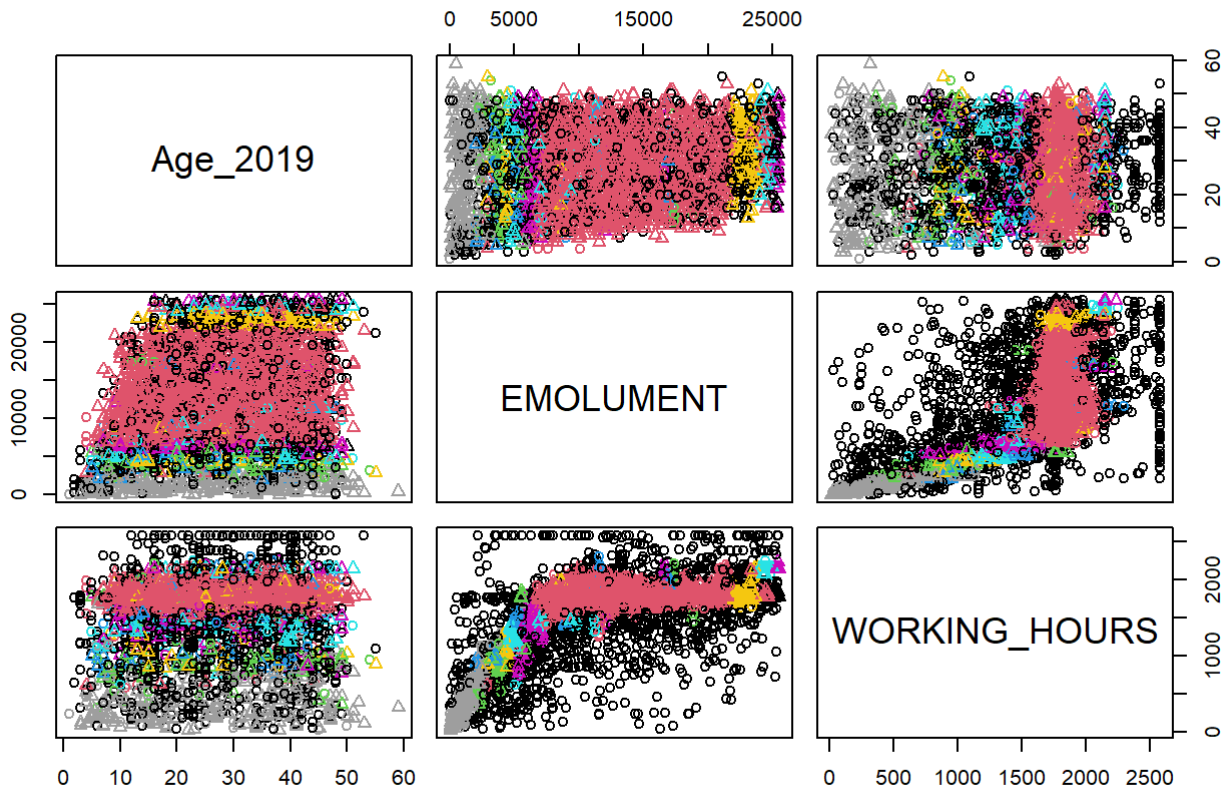
```
## Warning in axis(side = side, at = at, labels = labels, ...): "frame" n'est pas
## un paramètre graphique
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" n'est pas un
## paramètre graphique

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" n'est pas un
## paramètre graphique
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "frame" n'est pas
## un paramètre graphique
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" n'est pas un
## paramètre graphique

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" n'est pas un
## paramètre graphique
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "frame" n'est pas
## un paramètre graphique

## Warning in axis(side = side, at = at, labels = labels, ...): "frame" n'est pas
## un paramètre graphique
```



```
dbs = fpc::dbscan(db_dbscan, 1000, MinPts = 5)
plot(dbs, db_dbscan, main = "DBSCAN", frame = FALSE)
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "frame" n'est pas
## un paramètre graphique
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" n'est pas un
## paramètre graphique
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "frame" n'est pas
## un paramètre graphique
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" n'est pas un
## paramètre graphique
```

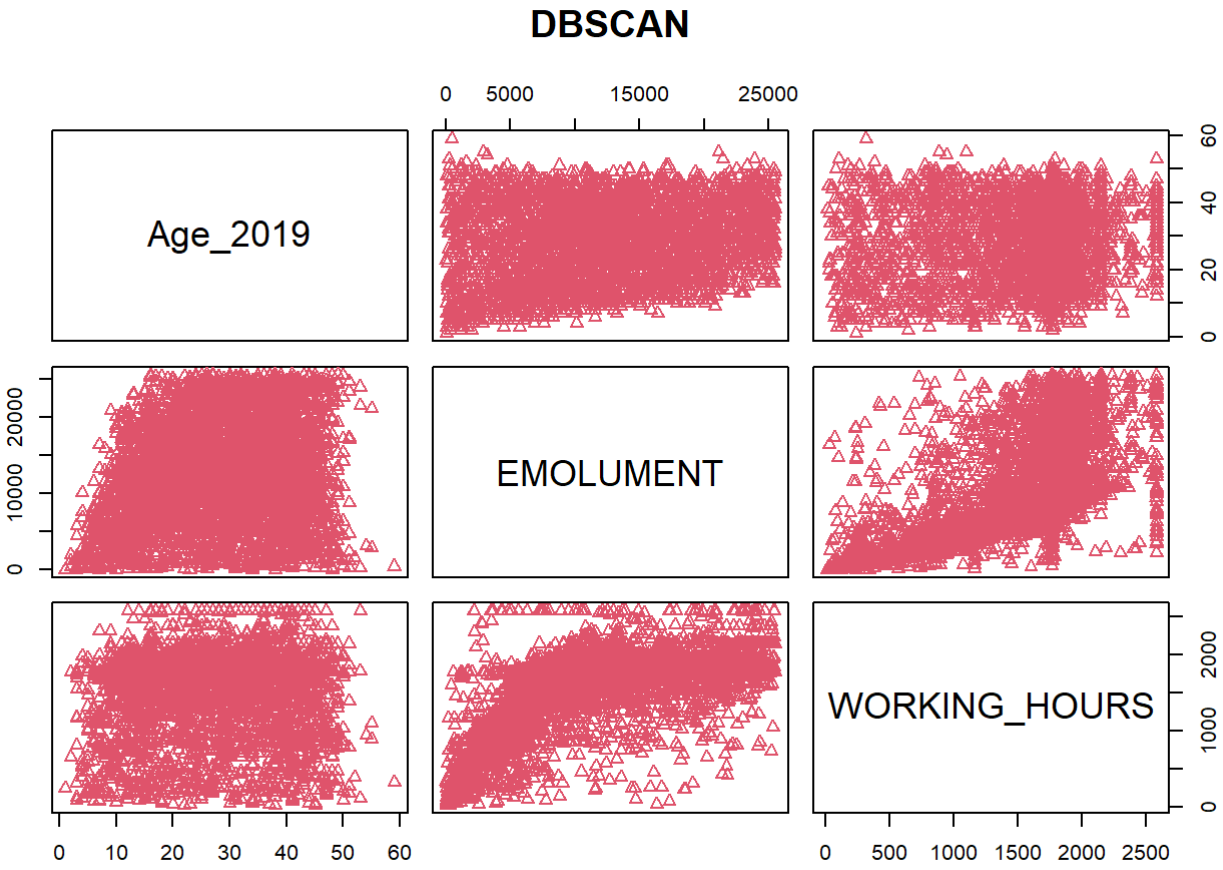
```
## Warning in axis(side = side, at = at, labels = labels, ...): "frame" n'est pas
## un paramètre graphique
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" n'est pas un
## paramètre graphique

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" n'est pas un
## paramètre graphique
```

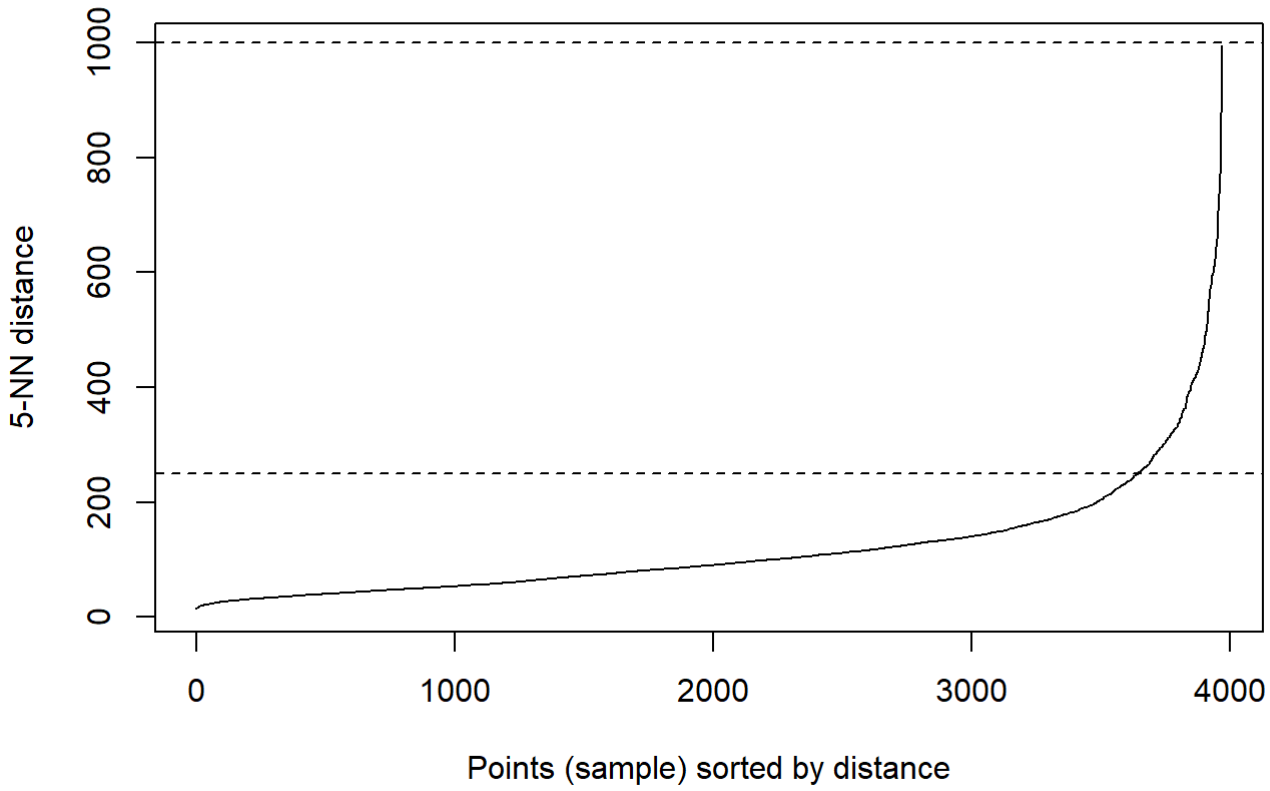
```
## Warning in axis(side = side, at = at, labels = labels, ...): "frame" n'est pas
## un paramètre graphique
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" n'est pas un
## paramètre graphique

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" n'est pas un
## paramètre graphique
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "frame" n'est pas
## un paramètre graphique

## Warning in axis(side = side, at = at, labels = labels, ...): "frame" n'est pas
## un paramètre graphique
```




```
dbscan::kNNdistplot(db_dbscan, k =  5)
abline(h = 250, lty = 2)
abline(h = 1000, lty = 2)
```



4 K-means sur l'ensemble des variables En utilisant que les variables numériques les clusters ne sont pas identifiables et ce constat est la même indépendamment de l'algoritme utilisé pour cette raison il est nécessaire d'utiliser l'ensemble de variables donc le K-means pour data mixte est utilisée à l'aide de la librairy "kamila"
Pour des raisons de resources (RAM) les clusters seront faits sur le périmètre de salaries (personnes actives).

```
library (kamila)
```

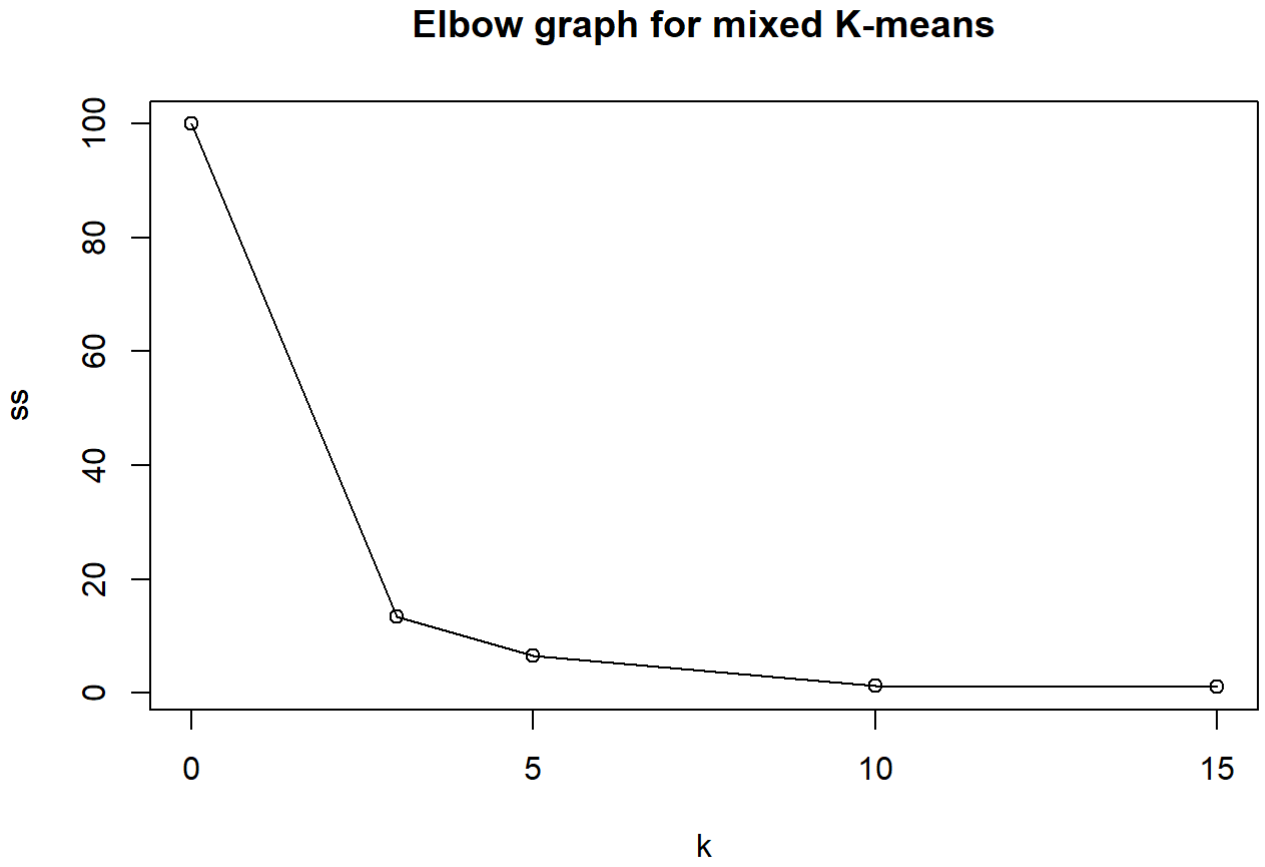
```
## Warning: package 'kamila' was built under R version 4.0.5
```

```
# kmeans_job5 = mixedDataClusteringKmeans(db_job, 5, importance = 0.5)
# kmeans_job10 = mixedDataClusteringKmeans(db_job, 10, importance = 0.5)
# kmeans_job15 = mixedDataClusteringKmeans(db_job, 15, importance = 0.5)
# db_job$K5 = as.factor(kmeans_job5$cluster)
# db_job$K10 = as.factor(kmeans_job10$cluster)
# db_job$K15 = as.factor(kmeans_job15$cluster)
# write.csv(db_job, "datasets/db_job_cluster.csv", row.names=F)
db_job_clust = read.csv("datasets/db_job_cluster.csv")
#db_job_clust=db_imp
```

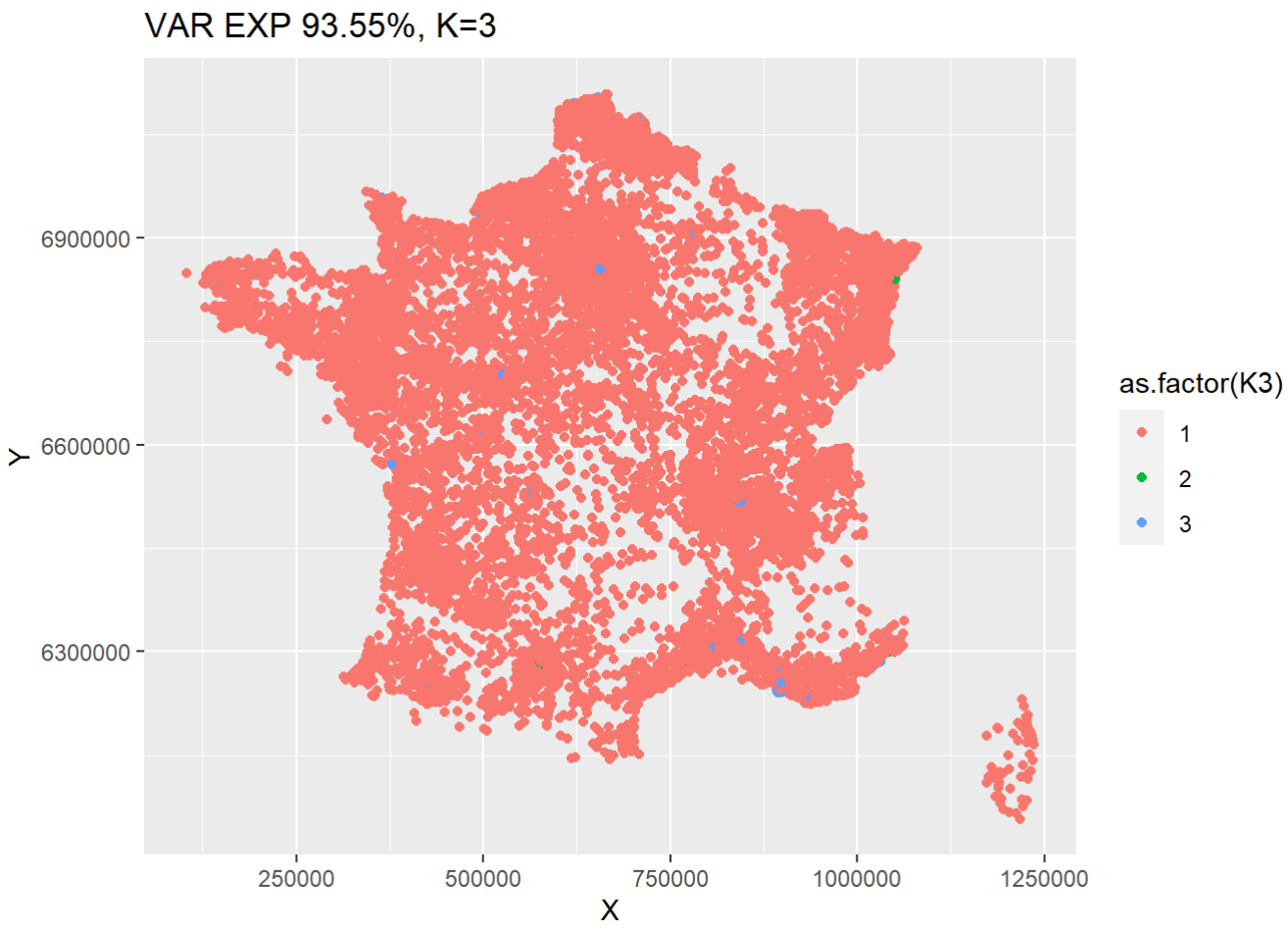
#Geographique clusters

Ci-dessous les clusters par département en utilisant toutes les variables sur le perimètre de perssonnes acttives (salaires). La variance expliqué avec 3 ou 5 clusters est trés proche (94%). En revanche, le diagrame du coude permet de constater que le meilleur niveau de k est 5. il est aussi observablé une réaprtition des clusters plus proche à la réalité du pays en fonction du salaire, il est constaté dans la carte ci-dessous des clusters : -en Ille de France _dans la côte sud -

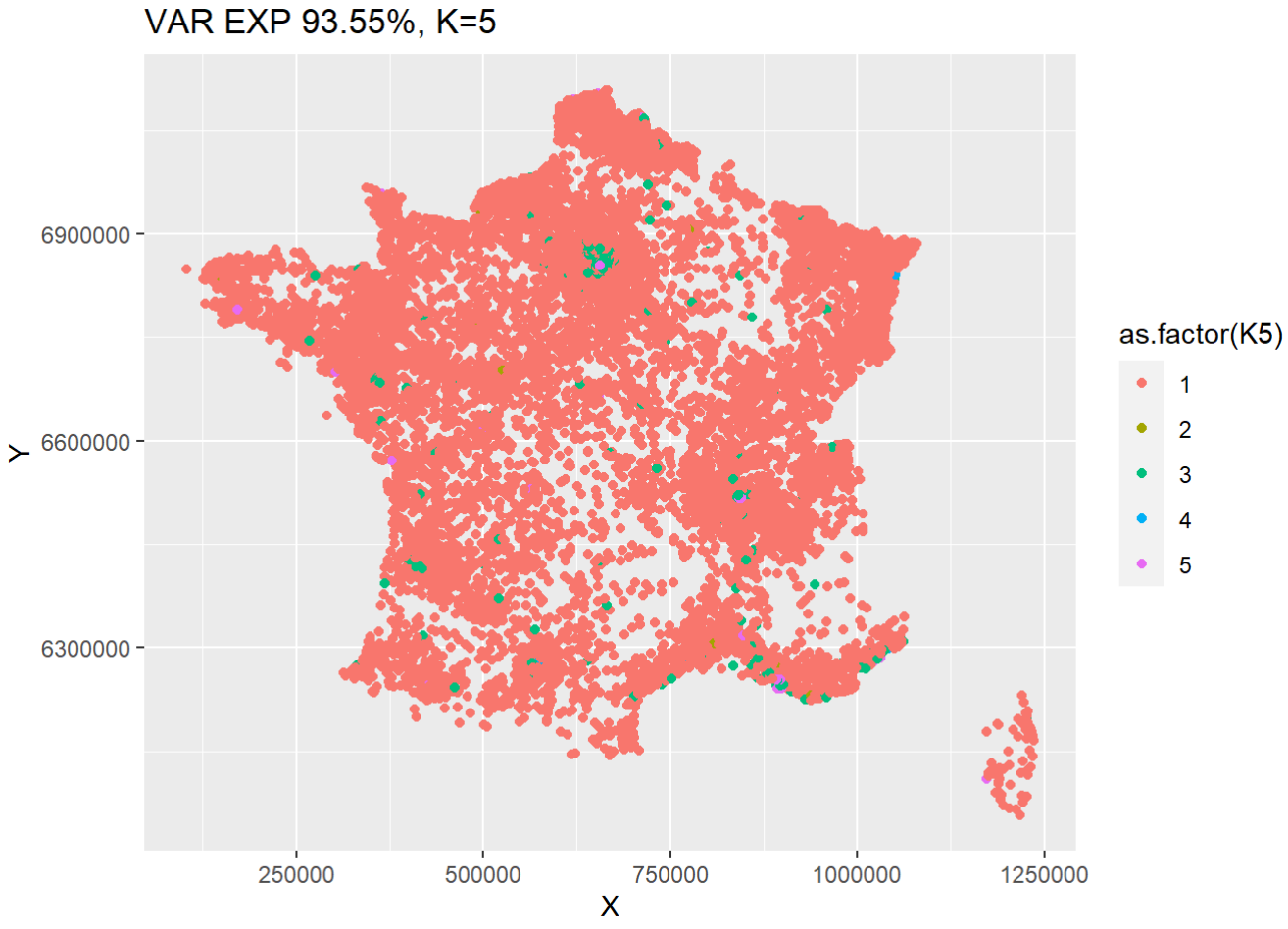
```
library(ggplot2)

plot(c(0,3,5,10,15),c(100,100-db_job_clust[1,c(35,37,39,41)]), type = "l", ylab = "ss", xlab
 = "k", main = "Elbow graph for mixed K-means")
points(c(0,3,5,10,15),c(100,100-db_job_clust[1,c(35,37,39,41)]))
```



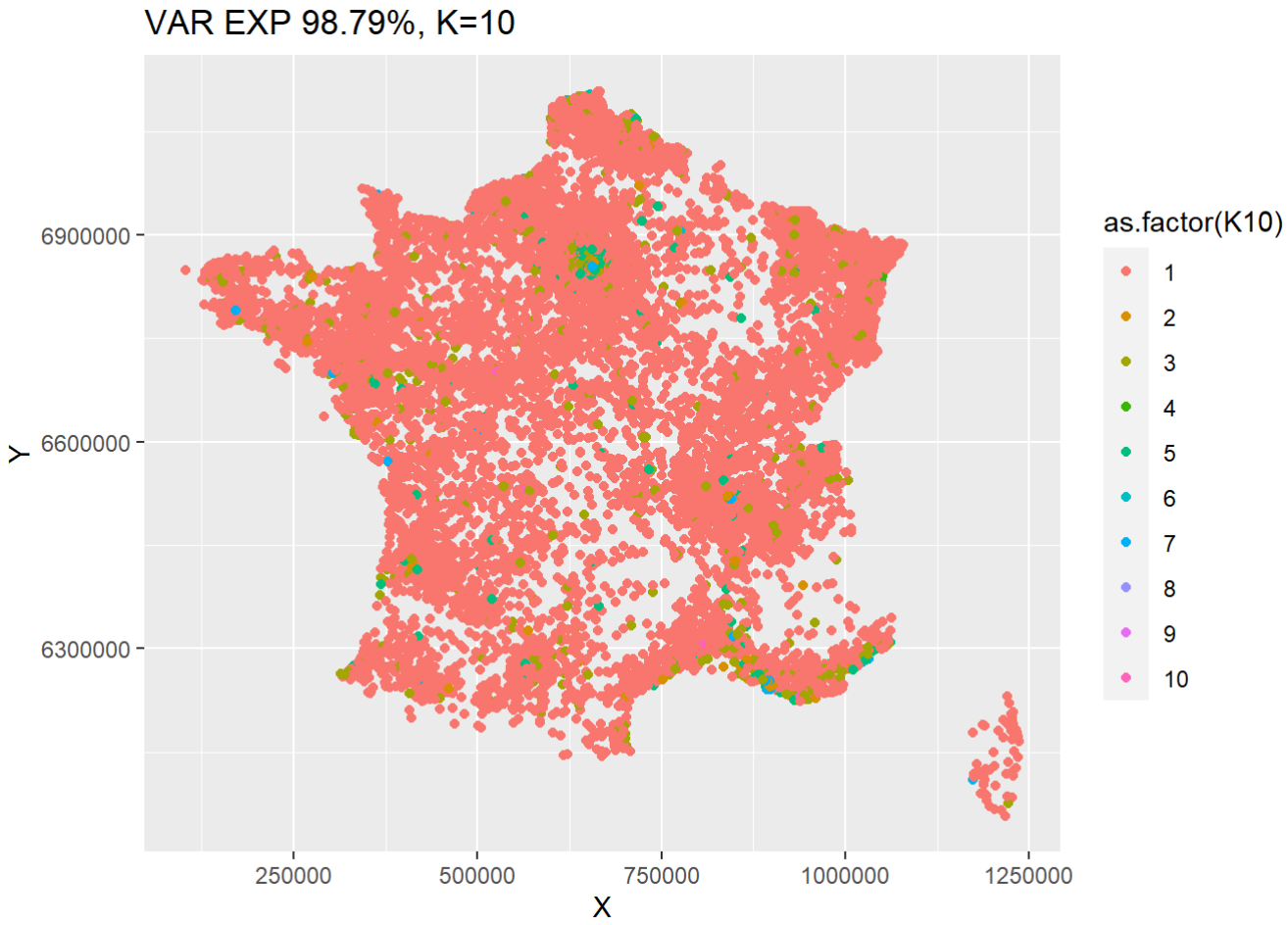

```
# result_map = map_data("france")
# map = ggplot() + geom_polygon(result_map, mapping = aes(long, lat, group = group, fill = gr
oup)) + coord_map()
# map + geom_point(data = db_job_clust,aes(x=long, y=Lat, colour = (K10)))# + ggtitle(paste0
("VAR EXP ",as.character(unique(db_job_clust$var_exp3)),"%, K=3"))
ggplot(data = db_job_clust) + geom_point(aes(x = X, y = Y, col = as.factor(K3))) + ggtitle(pa
ste0("VAR EXP ",as.character(unique(db_job_clust$var_exp5)),"%, K=3"))
```



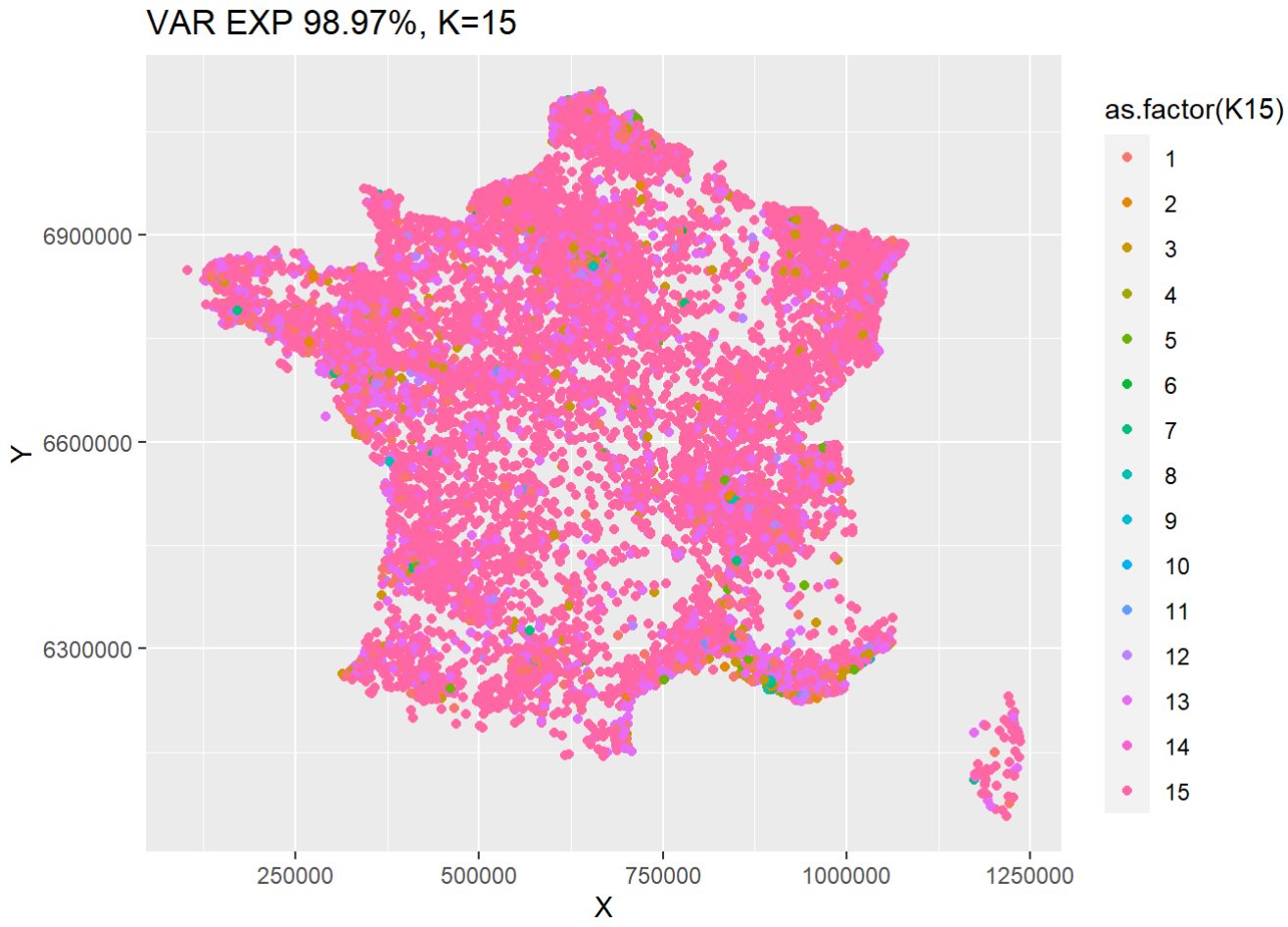

```
ggplot(data = db_job_clust) + geom_point(aes(x = X, y = Y, col = as.factor(K5))) + ggtitle(pa
ste0("VAR EXP ",as.character(unique(db_job_clust$var_exp5)),"%, K=5"))
```



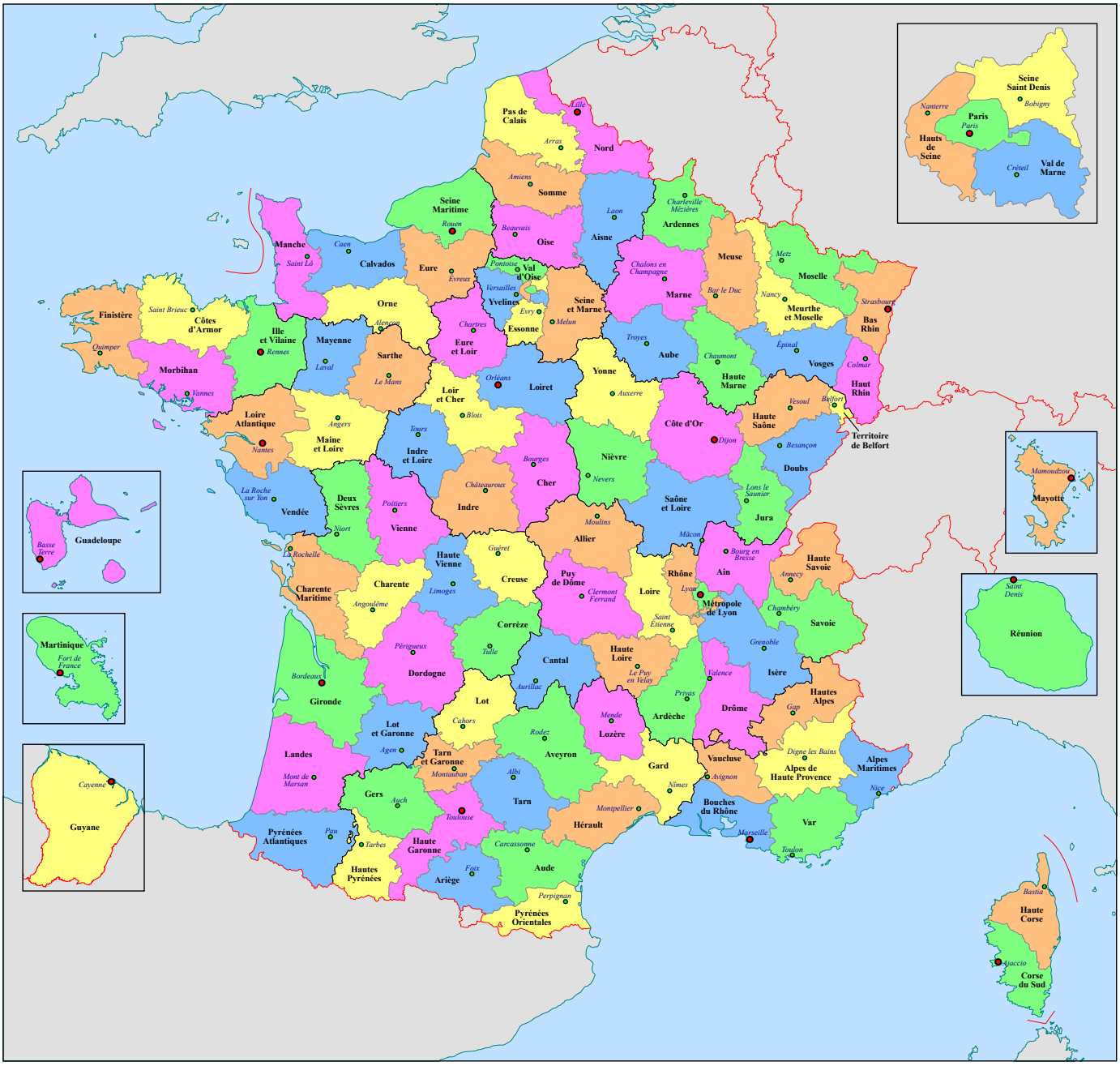
```
ggplot(data = db_job_clust) + geom_point(aes(x = X, y = Y, col = as.factor(K10))) + ggtitle(p
aste0("VAR EXP ",as.character(unique(db_job_clust$var_exp10)),"%, K=10"))
```




```
ggplot(data = db_job_clust) + geom_point(aes(x = X, y = Y, col = as.factor(K15))) + ggtitle(p
aste0("VAR EXP ",as.character(unique(db_job_clust$var_exp15)),"%, K=15"))
```


VAR EXP 98.97%, K=15
6900000
6600000
6300000
Y
250000
500000
750000
1000000
1250000
X
as.factor(K15)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

####ggh



Map of France.