Profil territorial des émissions de CO2 en France

Analyse descriptive des données

Tijou Louane & Gottschalk Eliott

2026-02-04

Table des matières

il faut un script R pour tout éxécuter en une seule fois (ou cellule quarto) code fold sur le pdf > code plié donc moins de place

Introduction

faire

Import des packages

Les packages suivants sont utilisés afin de faciliter l'importation des données, leur nettoyage, l'analyse exploratoire ainsi que la réalisation des analyses statistiques et graphiques nécessaires à l'étude.

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.2
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
Warning: le package 'naniar' a été compilé avec la version R 4.5.2
```

```
Attachement du package : 'EnvStats'

Les objets suivants sont masqués depuis 'package:stats':

    predict, predict.lm

L'objet suivant est masqué depuis 'package:base':

    print.default

Le chargement a nécessité le package : xts
Le chargement a nécessité le package : zoo

Attachement du package : 'zoo'

Les objets suivants sont masqués depuis 'package:base':

    as.Date, as.Date.numeric


########################## Warning from 'xts' package ##########################
#                                                                              #
# The dplyr lag() function breaks how base R's lag() function is supposed to  #
# work, which breaks lag(my_xts). Calls to lag(my_xts) that you type or       #
# source() into this session won't work correctly.                             #
#                                                                              #
# Use stats::lag() to make sure you're not using dplyr::lag(), or you can add #
# conflictRules('dplyr', exclude = 'lag') to your .Rprofile to stop           #
# dplyr from breaking base R's lag() function.                                 #
#                                                                              #
# Code in packages is not affected. It's protected by R's namespace mechanism #
# Set `options(xts.warn_dplyr_breaks_lag = FALSE)` to suppress this warning.  #
#                                                                              #
################################################################################

Attachement du package : 'xts'

Les objets suivants sont masqués depuis 'package:dplyr':

    first, last
```

```
Attachement du package : 'PerformanceAnalytics'

Les objets suivants sont masqués depuis 'package:EnvStats':

    kurtosis, skewness

L'objet suivant est masqué depuis 'package:graphics':

    legend

corrplot 0.95 loaded

Warning: le package 'reshape2' a été compilé avec la version R 4.5.2


Attachement du package : 'reshape2'

L'objet suivant est masqué depuis 'package:tidyr':

    smiths

Warning: le package 'FactoMineR' a été compilé avec la version R 4.5.2

Warning: le package 'factoextra' a été compilé avec la version R 4.5.2

Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

Warning: le package 'psych' a été compilé avec la version R 4.5.2


Attachement du package : 'psych'

L'objet suivant est masqué depuis 'package:outliers':

    outlier

Les objets suivants sont masqués depuis 'package:ggplot2':

    %+%, alpha
```

Import de la base de données

Cette base de données contient les émissions de CO des communes françaises, par secteur (agriculture, transport, énergie, industrie ...). Les données viennent du CITEPA et nous l'avons trouver sur la plateforme officielle *data.gouv.fr* (source)

Vérification des données importées

On regarde les premières lignes de la base de donnée avec pour s'assurer que les données ont été correctement importées et que les colonnes correspondent bien aux informations attendues.

```
# A tibble: 6 x 12
  `INSEE commune` Commune Agriculture `Autres transports` Autres transports in~1
  <chr>           <chr>         <dbl>               <dbl>                   <dbl>
1 01001           L'ABER~       3711.                 NA                       NA
2 01002           L'ABER~        475.                 NA                       NA
3 01004           AMBERI~        499.                213.                      NA
4 01005           AMBERI~       1859.                 NA                       NA
5 01006           AMBLEON        449.                 NA                       NA
6 01007           AMBRON~       4391.                133.                      NA
# i abbreviated name: 1: `Autres transports international`
# i 7 more variables: `CO2 biomasse hors-total` <dbl>, Déchets <dbl>,
#   Energie <dbl>, `Industrie hors-énergie` <dbl>, Résidentiel <dbl>,
#   Routier <dbl>, Tertiaire <dbl>
```

Vérification des noms des colonnes de la base de données et du nombre de ligne.

```
 [1] "INSEE commune"                  "Commune"
 [3] "Agriculture"                    "Autres transports"
 [5] "Autres transports international" "CO2 biomasse hors-total"
 [7] "Déchets"                        "Energie"
 [9] "Industrie hors-énergie"         "Résidentiel"
[11] "Routier"                        "Tertiaire"
```

```
[1] 35798
```

Description des Variabes

Explication de nos 12 variables :

- **Code_INSEE_commune** : (chr) Code unique à 5 chiffres attribué à chaque commune en France par l'INSEE. Les deux premiers chiffres correspondent au département.
- **Nom_commune** :(chr) Nom de la commune
- **Agriculture** :(num) Emissions de CO liées aux activités agricoles (en tonnes équivalent CO)
- **Autres_transports** : (num) Emissions liées aux transports non routiers et non internationaux,comme le transport fluvial, ferroviaire ou local, (en tonnes équivalent CO).
- **Autres_transports_international** : (num) Emissions liées aux transports internationaux (en tonnes équivalent CO).
- **CO2_biomasse** : (num) Emissions de CO provenant de la biomasse,c'est-à-dire la matière organique d'origine végétale ou animale utilisé pour produire de l'énergie, en tonnes équivalent CO .
- **Déchets** : (num) Emissions liées à la gestion des déchets (en tonnes équivalent CO).
- **Energie**: (num) Emissions liées à la production et consommation d'énergie (en tonnes équivalent CO).
- **Industrie** : (num) Emissions de l'industrie hors énergie, par exemple la fabrication de matériaux, de produits chimiques ou d'autres biens industriels, sans compter les émissions liées à la consommation d'énergie de ces industries), en tonnes équivalent CO .
- **Résidentiel** : (num) Emissions liées aux logements et aux habitations, incluant le chauffage, la consommation d'électricité et de gaz (secteur résidentiel), exprimées en tonnes équivalent CO .
- **Routier** : (num) Emissions liées au transport routier (en tonnes équivalent CO).
- **Tertiaire** : (num) Emissions liées au secteur tertiaire, c'est-à-dire les services comme les bureaux, commerces, administrations, écoles ou hôpitaux, (en tonnes équivalent CO).

Structure de la base de donnée

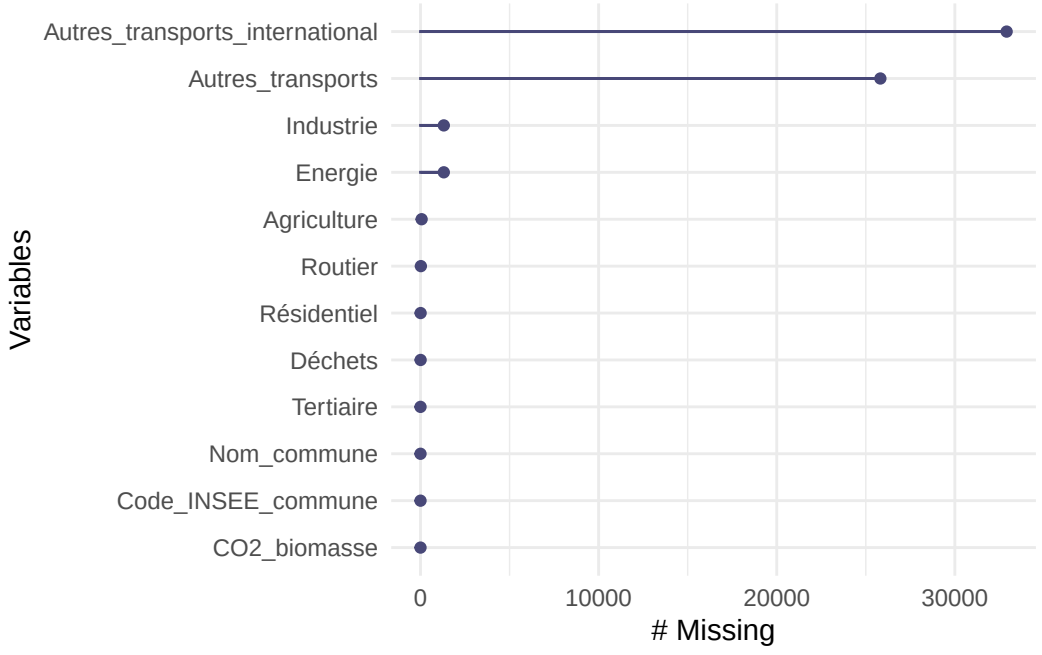
Vérification des types des variables. Cela permet de vérifier que les colonnes sont bien classées en numériques ou caractères avant de poursuivre l'analyse.

```
tibble [35,798 x 12] (S3: tbl_df/tbl/data.frame)
 $ Code_INSEE_commune             : chr [1:35798] "01001" "01002" "01004" "01005" ...
 $ Nom_commune                    : chr [1:35798] "L'ABERGEMENT-CLEMENCIAT" "L'ABERGEMENT-DE-
 $ Agriculture                    : num [1:35798] 3711 475 499 1859 449 ...
 $ Autres_transports              : num [1:35798] NA NA 213 NA NA ...
 $ Autres_transports_international: num [1:35798] NA NA NA NA NA NA NA NA NA NA ...
 $ CO2_biomasse                   : num [1:35798] 433 141 10313 1144 77 ...
 $ Déchets                        : num [1:35798] 101.4 140.7 5314.3 216.2 48.4 ...
 $ Energie                        : num [1:35798] 2.35 2.35 998.33 94.18 NA ...
 $ Industrie                      : num [1:35798] 6.91 6.91 2930.35 276.45 NA ...
 $ Résidentiel                    : num [1:35798] 309.4 104.9 16616.8 663.7 43.7 ...
 $ Routier                        : num [1:35798] 793 349 15642 1756 399 ...
 $ Tertiaire                      : num [1:35798] 367 112.9 10732.4 782.4 51.7 ...
```

Aucun problème : toutes nos variables quantitatives sont bien au format numérique, et nos deux variables qualitatives (Commune et Code_INSEE_commune) sont correctement au format caractère.

Vérification des données manquantes

Il est à présent temps de vérifier si il y a des données manquantes dans nos variables



On s'aperçoit que les variables *Autres_transports* et *Autres_transports_international* contiennent plus de 25 000 valeurs manquantes sur 35 798 lignes, ce qui représente environ 70 % des données. Étant donné cette proportion très élevée de données manquantes, il est préférable de supprimer ces deux colonnes pour éviter de fausser l'analyse.

```
[1] 32907
```

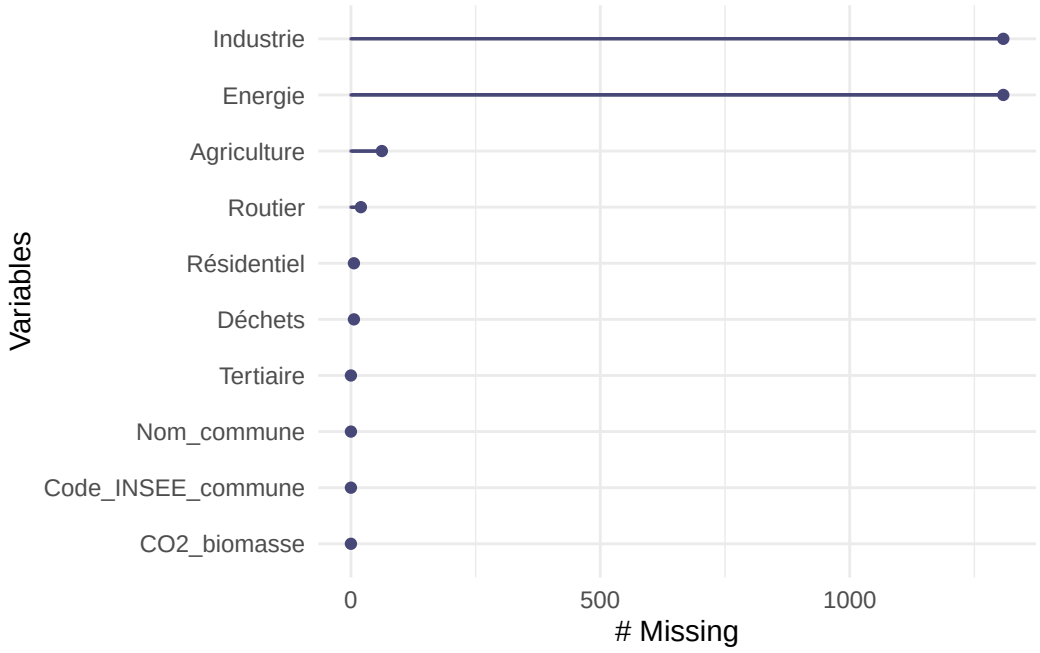
```
[1] 25819
```

```
[1] 1308
```

```
[1] 1308
```

Suppression des 2 variables avec plus de 80 % de NA

On s'aperçoit qu'il reste toujours quelques variables avec des NA



1. Analyse univarié

```
      Variable   Moyenne     Médiane     Minimum    Maximum Ecart_type
1  Agriculture 2459.9758 1559.381285 0.003431569   98949.32   2926.958
2 CO2_biomasse 1774.3815  424.849988 3.758087577  576394.18   7871.342
3      Déchets  410.8063   54.748653 0.132243124  275500.37   4122.473
4      Energie  662.5698    4.709115 2.354557741 2535857.56  26455.714
5    Industrie 2423.1278   13.822427 1.052998302 6765118.85  56703.738
6  Résidentiel 1783.6779  227.091193 1.027266053  410675.90   8915.902
7      Routier 3535.5012 1070.895593 0.555092164  586054.67   9663.157
8    Tertiaire 1105.1659  216.297718 0.000000000  288175.40   5164.183
```

Création de la variable département

Le fichier contient des données pour toutes les communes françaises, ce qui représente plusieurs milliers de lignes, soit 35 798 observations. Travailler à ce niveau de détail ralenti les calculs et rend l'analyse moins lisible, notamment pour l'ACP. Pour simplifier l'analyse et faciliter la synthèse des données, nous créons une nouvelle variable *departement* en récupérant les deux premiers chiffres du code INSEE de chaque commune, correspondant au code du département.

Cela nous permet par la suite de regrouper les communes par département et de passer à une analyse plus globale, plus facile à manipuler et à interpréter.

```
 01  02  03  04  05  06  07  08  09  10  11  12  13  14  15  16  17  18  19  21
410 805 318 199 168 163 339 457 332 431 436 286 134 621 255 394 469 290 285 705
 22  23  24  25  26  27  28  29  2A  2B  30  31  32  33  34  35  36  37  38  39
362 259 545 585 367 617 391 281 124 236 353 589 462 540 343 351 243 277 526 528
 40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59
331 283 326 258 212 327 326 319 176 250 516 617 429 258 594 501 256 729 310 648
 60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79
690 413 893 467 547 472 226 518 366 293 543 570 368 292 290  20 718 511 262 297
 80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95
782 320 195 153 151 269 280 200 510 434 102 196  36  40  47 185
```

On vérifie si il y a des données manquantes dans la nouvelle variable dans le cas où elle aurait été mal créer

```
[1] 0
```

Import d'une nouvelle base de donnée (INSEE)

```
New names:
* `` -> `...1`
```

```
Warning: One or more parsing issues, call `problems()` on your data frame for details,
e.g.:
  dat <- vroom(...)
  problems(dat)
```

```
Rows: 34935 Columns: 47
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (29): code_insee, nom_standard, nom_sans_pronom, nom_a, nom_de, nom_sans...
dbl (18): ...1, reg_code, epci_code, academie_code, taille_unite_urbaine, po...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

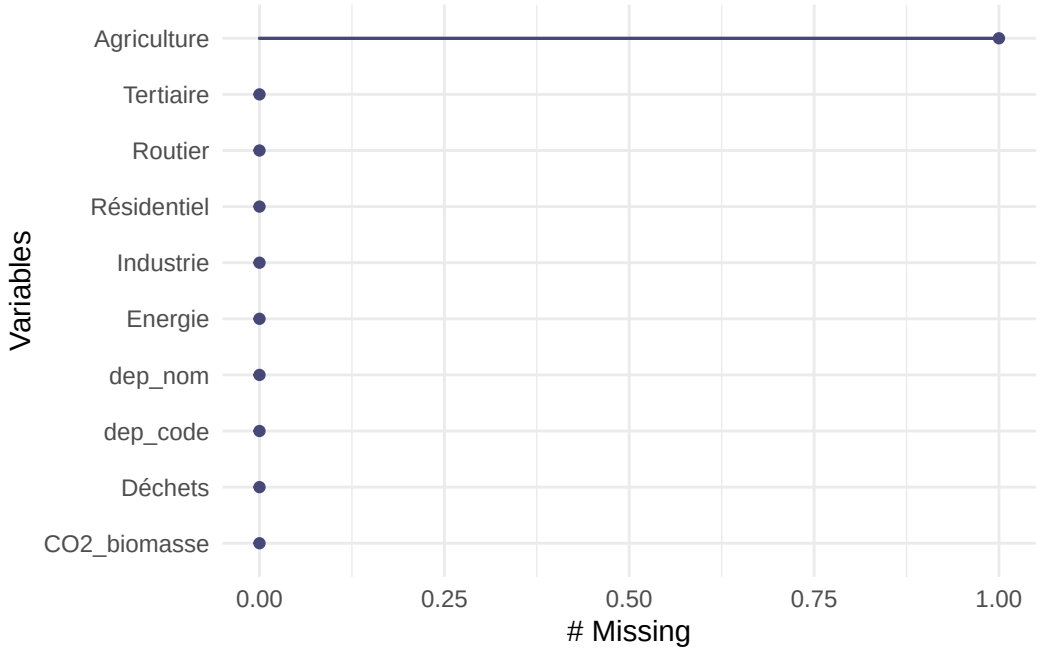
On garde seulement 2 variables sur les 47, le nom du département et le code de celui-ci

On garde seulement une ligne pour chaque numéros de département différent

Jointure gauche, on garde toutes les lignes de la base de donnée et on rajoute la colone dep_nom (nom du département) si il y a des correspondancse au niveau des codes de département

```
[1] 0
```

On supprime les colonnes dont nous avons plus besoin




```
tibble [96 x 10] (S3: tbl_df/tbl/data.frame)
 $ dep_code    : chr [1:96] "01" "02" "03" "04" ...
 $ dep_nom     : chr [1:96] "Ain" "Aisne" "Allier" "Alpes-de-Haute-Provence" ...
 $ Agriculture : num [1:96] 1975 1585 6132 1825 1848 ...
 $ CO2_biomasse: num [1:96] 1736 767 1780 583 502 ...
 $ Déchets     : num [1:96] 672 224 350 254 133 ...
 $ Energie     : num [1:96] 280.5 76.3 326.9 62.8 35 ...
 $ Industrie   : num [1:96] 1745 932 1452 314 103 ...
 $ Résidentiel : num [1:96] 1347 794 1402 587 729 ...
 $ Routier     : num [1:96] 3989 1722 3663 1963 2071 ...
 $ Tertiaire   : num [1:96] 1021 404 706 494 464 ...
```

Les données communales ont été agrégées au niveau départemental en calculant, pour chaque département, la moyenne des émissions de CO par secteur des communes qui le composent.

Les résultats doivent être interprétés comme des profils moyens de communes au sein des départements, et non comme des niveaux totaux d'émissions départementales.

Le passage d'une échelle communale à une échelle départementale permet de réduire la taille du jeu de données et de rendre par la suite l'analyse des corrélations et l'ACP plus lisibles et interprétables.

Analyse univarié

Variables quantitatives

	Variable	Moyenne	Médiane	Minimum	Maximum	Ecart_type
1	Agriculture	2622.9968	1967.2570	8.309835	7870.148	1933.838
2	CO2_biomasse	3079.7749	1503.9183	402.863341	59328.866	6910.402
3	Déchets	645.5479	281.3376	60.648921	7347.163	1157.415
4	Energie	1033.4024	184.8777	34.971220	39455.038	4227.943
5	Industrie	3260.7873	1329.7422	99.333536	86844.708	9353.330
6	Résidentiel	3803.3098	1228.1670	326.736594	96729.000	11277.329
7	Routier	5584.3430	3085.4544	966.116869	81279.136	9740.374
8	Tertiaire	2460.4902	797.2968	228.707087	66581.497	7531.315

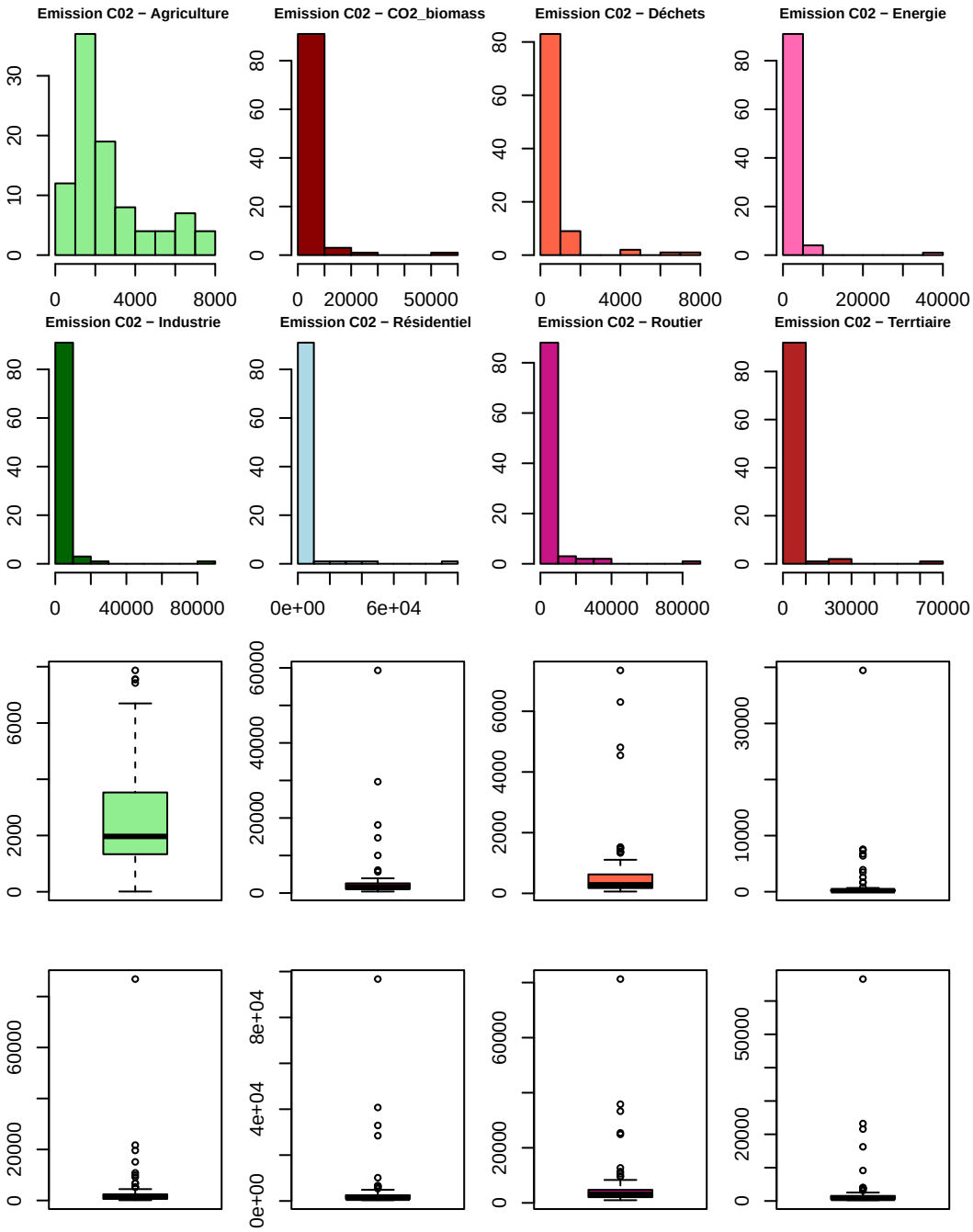
Pas d'agriculture à Paris c'est normal

On remarque que quelques départements très émetteurs de CO2 tirent fortement la moyenne vers le haut, tandis que la majorité des départements ont des niveaux d'émissions plus faibles.

On peut alors penser que par exemple pour les déchets il y a une forte concentration géographique des infrastructures de traitement des déchets ce qui explique une médiane plutot basse par rapport au maximum. De la même manière on remaque certains département beaucoup plus industrialisé que d'autre

En résumer, l'analyse descriptive met en évidence une forte hétérogénéité des émissions de CO2 selon les secteurs et les départements, avec des distributions très asymétriques et la présence de valeurs extrêmes.

Visualisation :



Les boxplots mettent en évidence la présence de nombreuses valeurs extrêmes pour la plupart des secteurs. Ces valeurs ne traduisent pas des erreurs de mesure, mais reflètent des spécificités territoriales fortes (départements très industrialisés, fortement urbanisés ou spécialisés dans la production d'énergie). Elles ont donc été conservées, car elles constituent une information

pertinente pour l'analyse des profils d'émissions.

Variables qualitative

La base de données contient deux variables qualitatives (dep_code et dep_nom) correspondant à l'identification des départements. Ces variables servent uniquement de repères pour l'interprétation des résultats. Nous vérifions seulement l'absence de doublons afin de nous assurer que chaque département est bien représenté une seule fois. La réalisation d'un graphique en camembert serait illisible compte tenu du nombre élevé de départements et n'apporterait aucune information supplémentaire à l'analyse.

```
[1] 0
```

```
[1] 0
```

Aucun doublon n'a été détecté.

2. Analyse bivariée

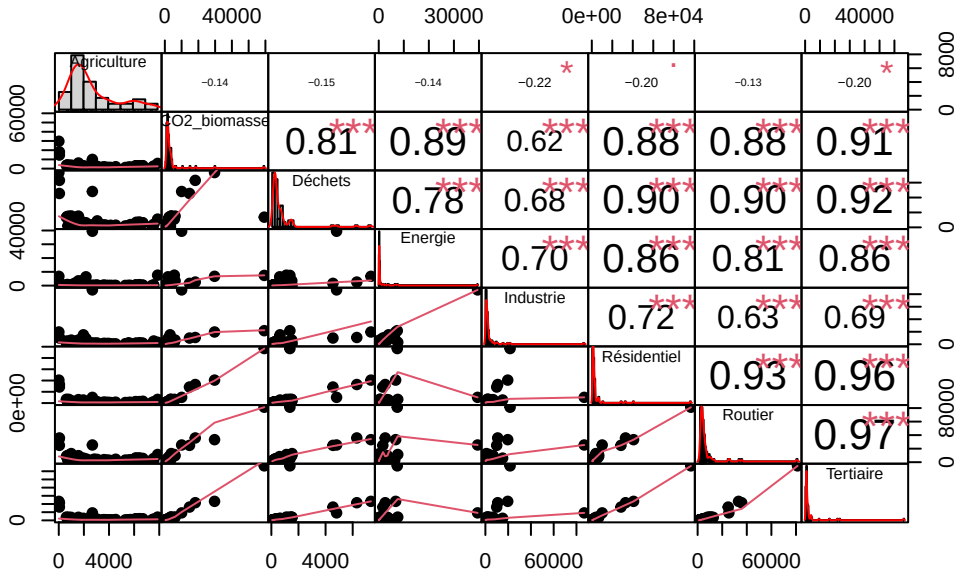
Variables quantitative - quantitative

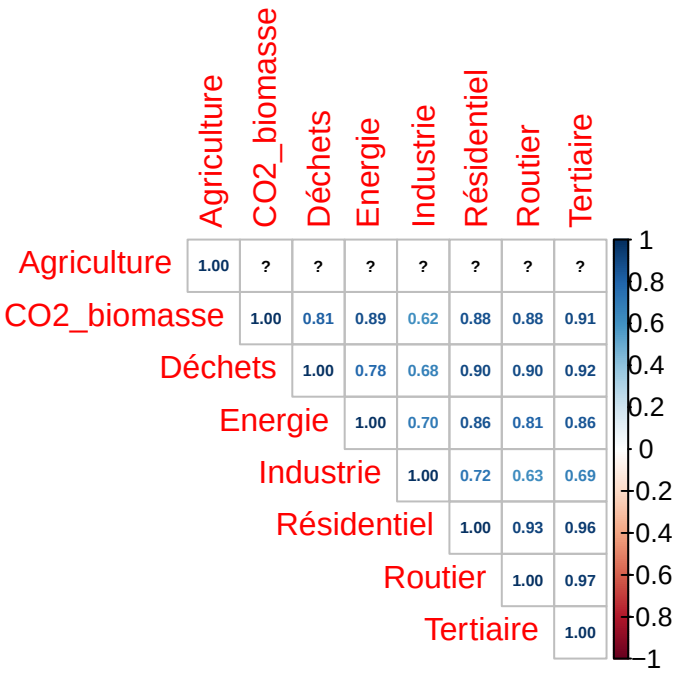
```
# A tibble: 20 x 3
   var1         var2          cor
   <chr>        <chr>       <dbl>
 1 Résidentiel  Tertiaire   0.996
 2 CO2_biomasse Résidentiel 0.992
 3 CO2_biomasse Tertiaire   0.989
 4 Routier      Tertiaire   0.980
 5 Résidentiel  Routier     0.975
 6 CO2_biomasse Routier     0.975
 7 Energie      Industrie   0.970
 8 Déchets      Routier     0.634
 9 Déchets      Industrie   0.586
10 CO2_biomasse Déchets     0.572
11 Déchets      Résidentiel 0.555
12 Déchets      Tertiaire   0.521
13 Déchets      Energie     0.517
14 Industrie    Routier     0.494
15 Energie      Routier     0.425
16 CO2_biomasse Industrie   0.408
```

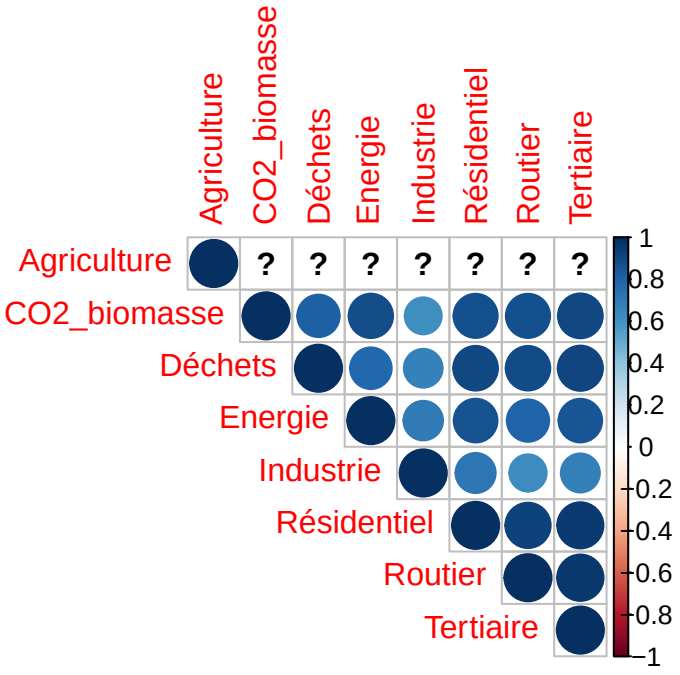
```
17 Industrie    Tertiaire   0.385
18 Industrie    Résidentiel 0.365
19 CO2_biomasse Energie     0.335
20 Energie      Tertiaire   0.310
```

	Agriculture	CO2_biomasse	Déchets	Energie	Industrie
Agriculture	1.0000000	-0.1395717	-0.1502380	-0.1358763	-0.2211506
CO2_biomasse	-0.1395717	1.0000000	0.8051792	0.8843085	0.6044933
Déchets	-0.1502380	0.8051792	1.0000000	0.7754059	0.6704087
Energie	-0.1358763	0.8843085	0.7754059	1.0000000	0.6955347
Industrie	-0.2211506	0.6044933	0.6704087	0.6955347	1.0000000
Résidentiel	-0.1958567	0.8751120	0.8983763	0.8578807	0.7157615
Routier	-0.1282895	0.8745521	0.8927772	0.8002100	0.6141517
Tertiaire	-0.2024356	0.9029675	0.9174552	0.8535274	0.6753779

	Résidentiel	Routier	Tertiaire
Agriculture	-0.1958567	-0.1282895	-0.2024356
CO2_biomasse	0.8751120	0.8745521	0.9029675
Déchets	0.8983763	0.8927772	0.9174552
Energie	0.8578807	0.8002100	0.8535274
Industrie	0.7157615	0.6141517	0.6753779
Résidentiel	1.0000000	0.9227464	0.9574748
Routier	0.9227464	1.0000000	0.9665873
Tertiaire	0.9574748	0.9665873	1.0000000







Pour analyser les relations entre les différents secteurs d'émissions de CO , nous utilisons la corrélation de Spearman. Bien que toutes les variables soient exprimées en tonnes de CO , leurs distributions sont très asymétriques et présentent des valeurs extrêmes. La corrélation de Spearman, basée sur les rangs des valeurs, est moins sensible aux outliers et permet d'identifier

les associations monotones entre les secteurs de manière robuste. Elle est donc plus appropriée que la corrélation de Pearson dans ce contexte.

Les variables sont quantitatives, exprimées dans une unité homogène (tonnes équivalent CO), et présentent de fortes corrélations, notamment entre les secteurs liés à l'urbanisation, aux transports et à l'activité économique. En particulier, les émissions résidentielle, tertiaire et routière sont très fortement corrélées entre elles (coefficients supérieurs à 0,9), traduisant une importante redondance informationnelle.

À l'inverse, le secteur agricole présente des corrélations faibles et négatives avec les autres variables, suggérant un profil d'émissions distinct. Cette structure justifie pleinement le recours à une analyse en composantes principales afin de synthétiser l'information et d'identifier des profils territoriaux d'émissions.

Variables qualitative - qualitative

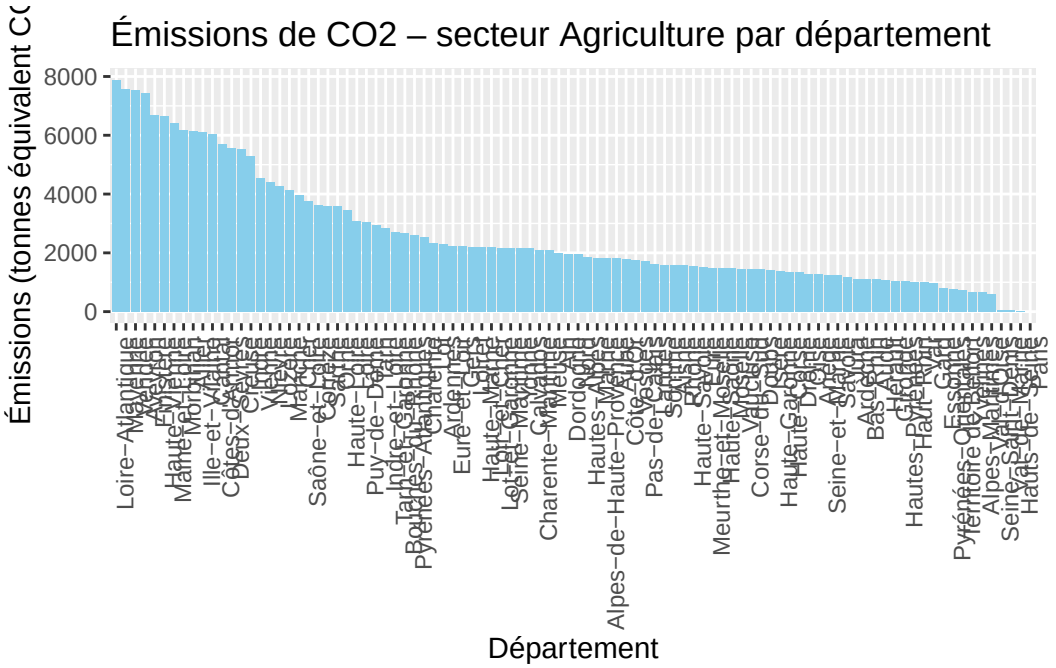
Les colonnes dep_code et dep_nom contiennent la même information sous des formats différents. Un tableau croisé ou un graphique n'apporterait aucune information supplémentaire et serait illisible.

Variables qualitative - quantitative

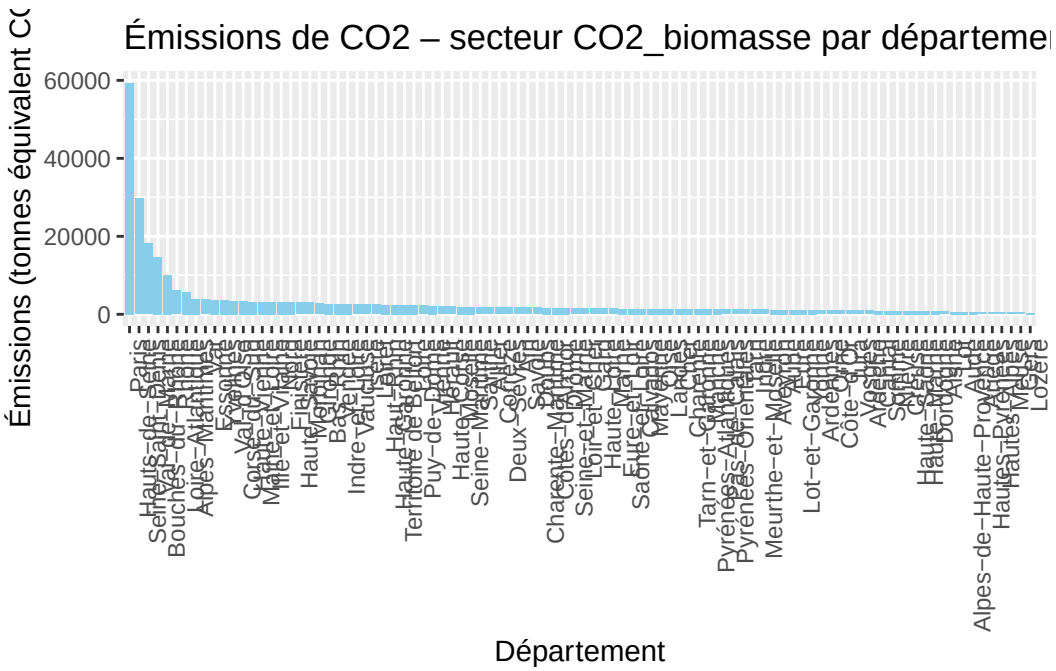
Répartition des émissions de CO par secteur et par département

```
$Agriculture
```

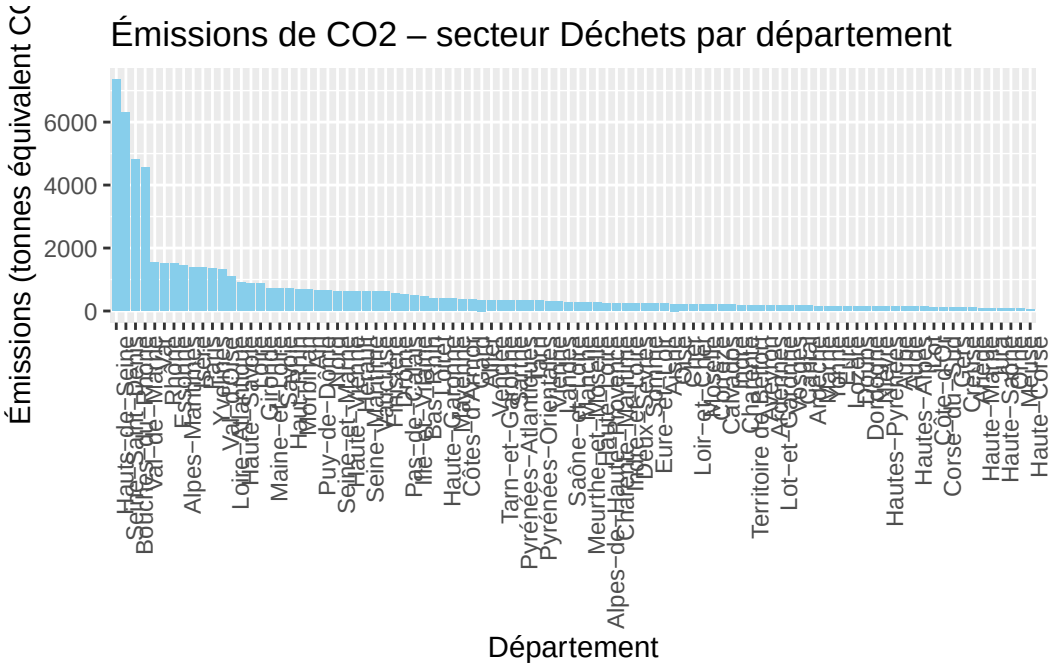
```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_bar()`).
```



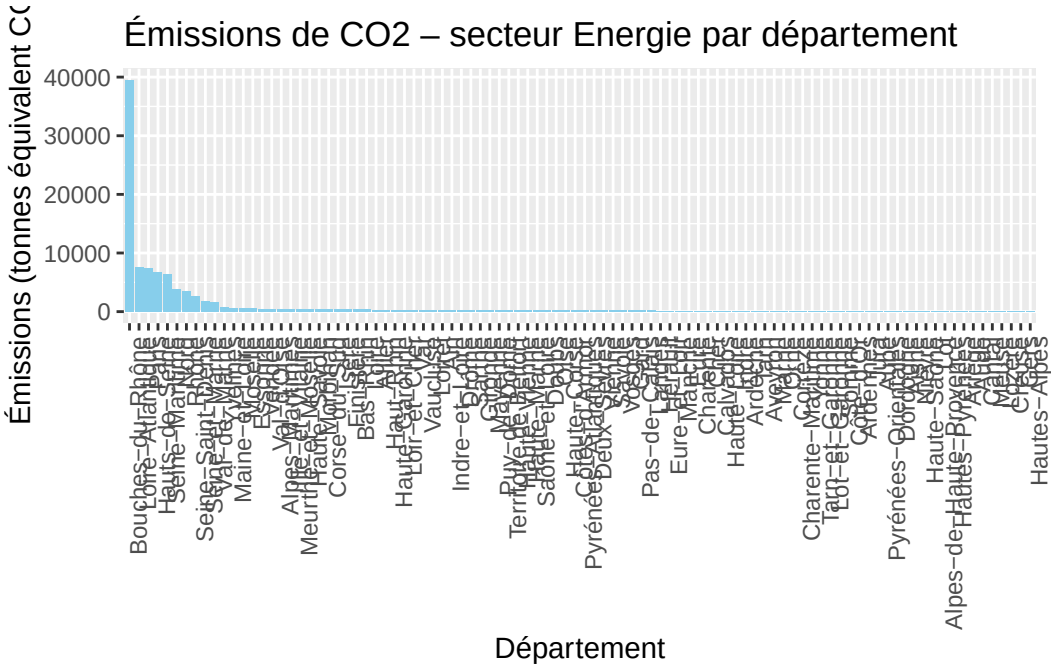

```
$CO2_biomasse
```



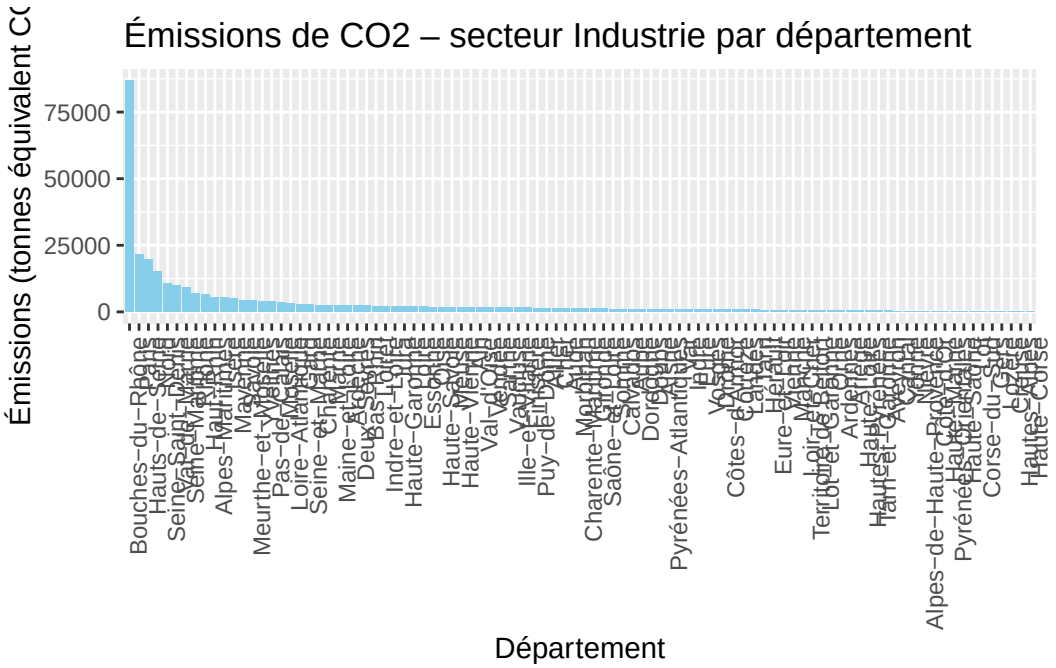
```
$Déchets
```



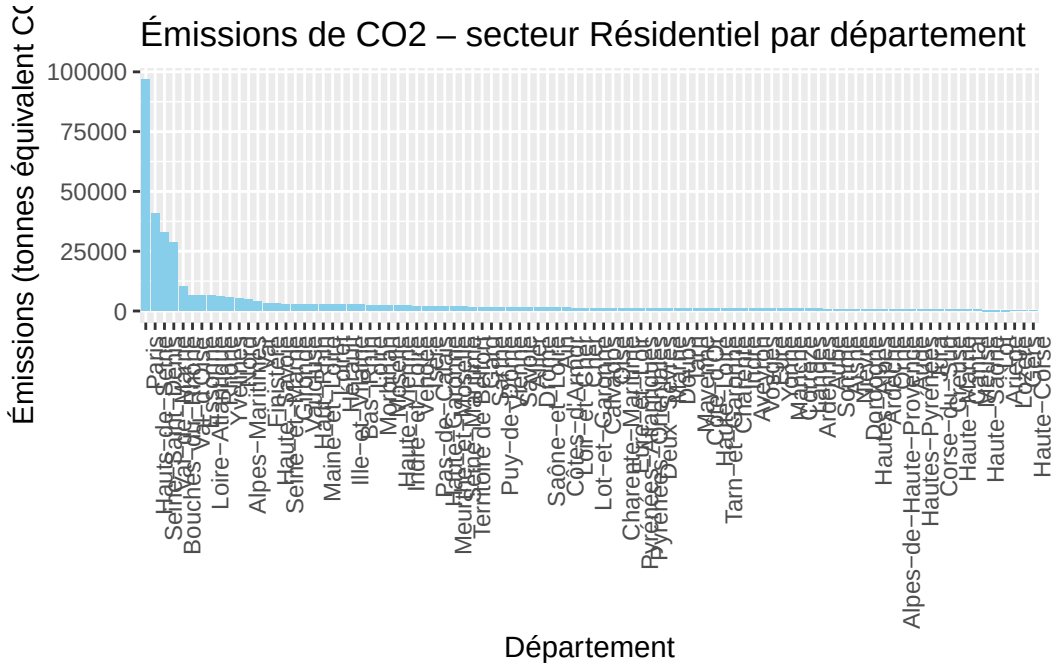

```
$Energie
```


Émissions de CO2 – secteur Energie par département
Émissions (tonnes équivalent CO
40000
30000
20000
10000
0
Département

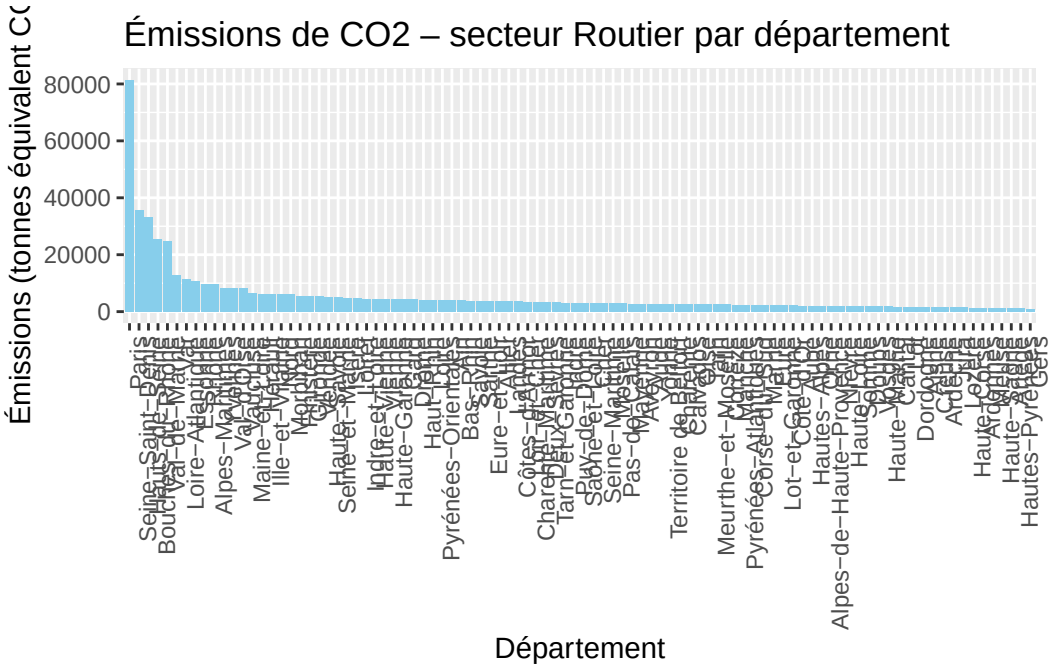
```
$Industrie
```


Émissions de CO2 – secteur Industrie par département
Émissions (tonnes équivalent CO
75000
50000
25000
0
Département

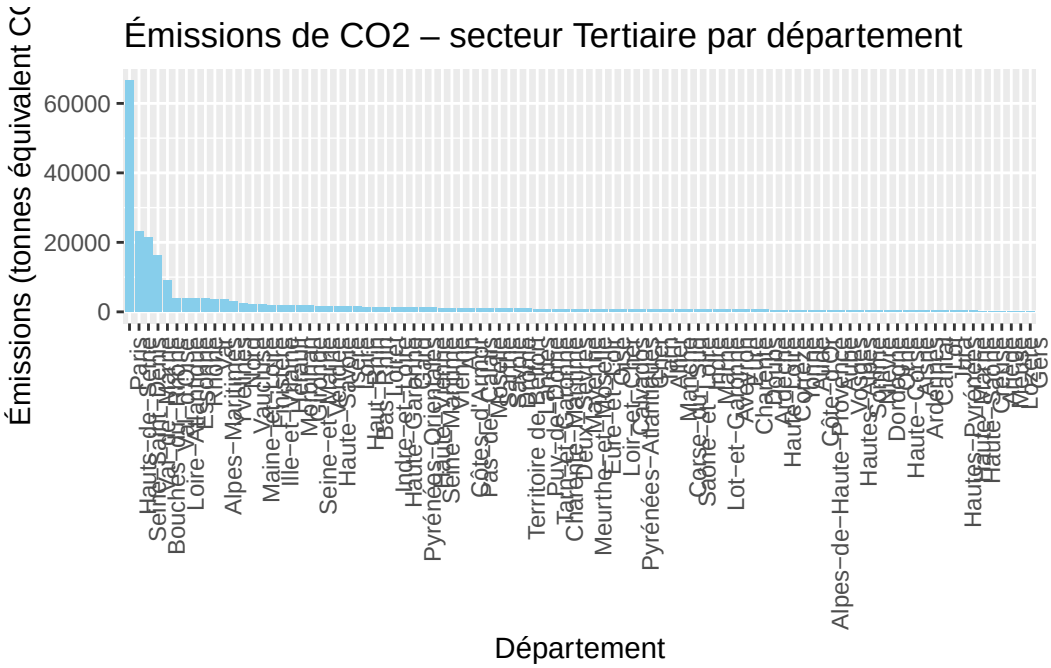
```
$Résidentiel
```




```
$Routier
```

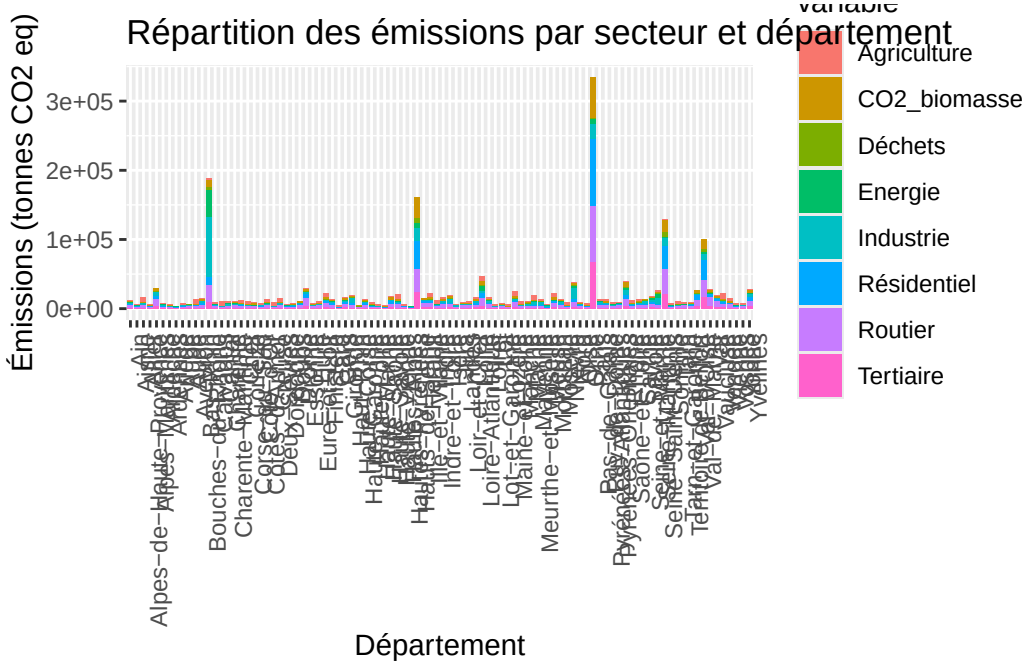

Émissions de CO2 – secteur Routier par département
Émissions (tonnes équivalent C
80000
60000
40000
20000
0
Département

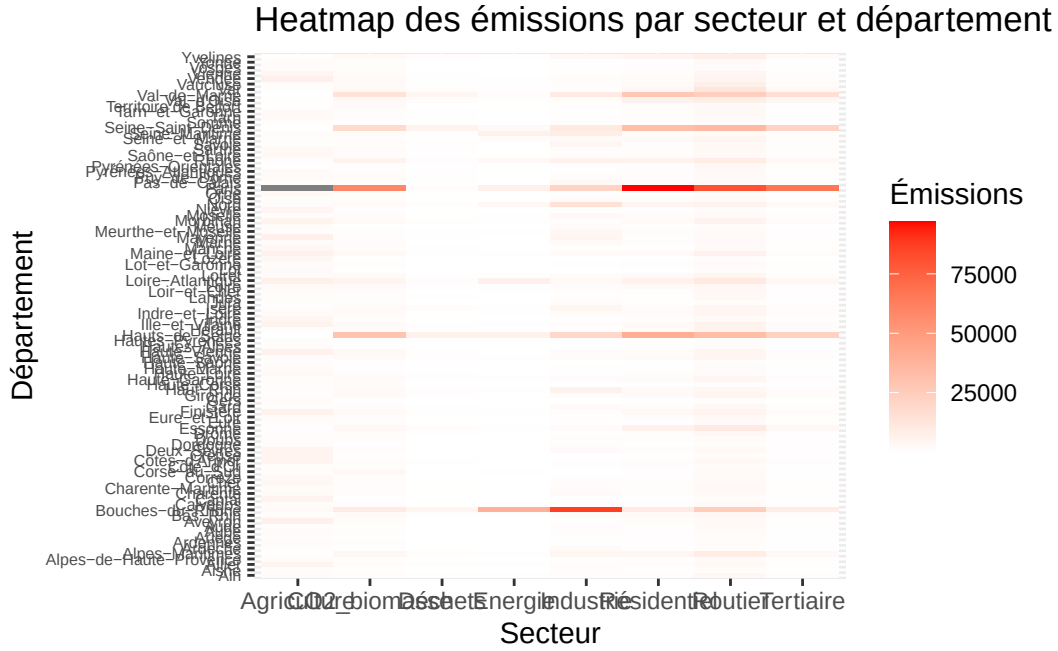
```
$Tertiaire
```


Émissions de CO2 – secteur Tertiaire par département
Émissions (tonnes équivalent C
60000
40000
20000
0
Département

Intensité des émissions de CO par secteur activité selon les départements

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_bar()`).
```





Comme observé, les émissions de CO par secteur sont fortement corrélées. Cette redondance complique l'interprétation directe. L'analyse en composantes principales (ACP) permettra de réduire la dimension du jeu de données, de synthétiser l'information et de mettre en évidence les profils départementaux similaires ou distincts.

3. Analyse factorielle

Justification de l'ACP

Nous choisissons de faire une analyse en composantes principales puisque nous avons principalement des variables quantitatives.

Indice KMO

Afin de vérifier l'adéquation de nos données à une Analyse en Composantes Principales, nous avons calculé l'indice KMO. Celui-ci permet d'évaluer si les corrélations entre variables sont suffisamment structurées pour être synthétisées par une analyse factorielle

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = Base_dep[, variables_quantitatives])
Overall MSA =  0.58
MSA for each item =
 Agriculture CO2_biomasse      Déchets      Energie    Industrie  Résidentiel
        0.13         0.79         0.41         0.62         0.66         0.68
     Routier    Tertiaire
        0.59         0.55
```

Le résultat de l'indice KMO est inférieur à 0,6, il est exactement de 0,58, nous décidons malgré tout de conserver l'analyse en composante principale car nous venons de démontrer de fortes corrélation entre certaines secteurs. Si l'on regarde plus précisément on voit que l'indice KMO ets notament tiré vers la bas à cause de la variable agriculture dont l'indice est seulement de 0,13 indiquant que ce secteur à un profil d'émission bien différent des autres secteurs. Par conséquent nous allons faire une ACP avec et sans la variable agriculture car ce secteur est malgré tout un secteur clé dans l'émission de CO2 en France et ces différences peuvent etre interessante à analyser.

Test de Bartlett

Afin de vérifié correctement et vérifié notre intuition si il existe des corrélations suffisante pour faire une ACP nous réalisons ce test de Bartlett.

```
$chisq
[1] 1792.397

$p.value
[1] 0

$df
[1] 28
```

Le test de Bartlett s'avère très significatif et permet de rejeté l'hypothèse nulle selon laquelle la matrice de corrélation est une matrice d'identité. Ainsi nous pouvons confirmer la présence de corrélations importante entre les secteurs d'émissions et justifié la pertinence de l'analyse en composantes principales.

ACP avec agriculture

Bien que toutes les variables soient exprimées dans la même unité (tonnes équivalent CO), leurs dispersions et ordres de grandeur diffèrent fortement selon les secteurs. Le centrage-réduction est donc nécessaire afin d'éviter que les secteurs les plus émetteurs ne dominent l'analyse et afin de mettre en évidence la structure relative des émissions entre départements.

```
[1] NA
```

Il n'est pas nul, donc il n'y a pas de corrélation parfaite.

```
Warning in PCA(Base_dep[, variables_quantitatives], scale.unit = TRUE, graph =
FALSE): Missing values are imputed by the mean of the variable: you should use
the imputePCA function of the missMDA package
```

Choix du nombre de dimenssion

Pourcentage d'inertie expliquée et règle de Kaiser

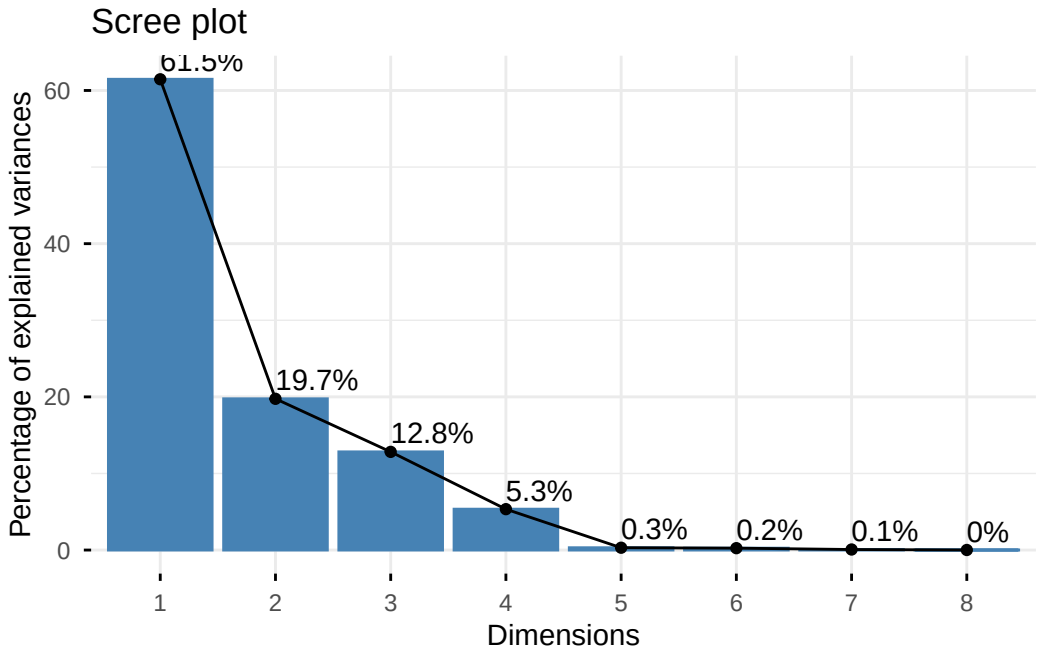
```
         eigenvalue percentage of variance cumulative percentage of variance
comp 1 4.9170836154           61.463545193                            61.46355
comp 2 1.5797584277           19.746980346                            81.21053
comp 3 1.0255247896           12.819059871                            94.02959
comp 4 0.4271011259            5.338764074                            99.36835
comp 5 0.0242978887            0.303723609                            99.67207
comp 6 0.0197720057            0.247150071                            99.91922
comp 7 0.0057397744            0.071747180                            99.99097
comp 8 0.0007223726            0.009029658                           100.00000
```

Après analyse des valeurs propres de l'ACP, on observe que les deux premières composantes principales expliquent 81,21% de la variance totale (Dim 1 : 61,46 %, Dim 2 : 19,74%). Ces deux axes contiennent donc l'essentiel de l'information et permettent de synthétiser efficacement les relations entre les secteurs et les départements. Nous avons donc choisi de retenir uniquement les deux premières composantes pour les analyses et visualisations ultérieures.

Selon la règle de Kaiser, on ne conserve que les composantes dont la valeur propre est supérieure à 1. Ici, les trois premières composantes ont une valeur propre > 1 (Comp 1 : 4,917, Comp 2 : 1,580, Comp 3 : 1,026). Cependant, la troisième composante n'apporte que 12,82 % d'inertie supplémentaire, donc nous choisissons de retenir les deux premières composantes.

Éboulis des valeurs propres

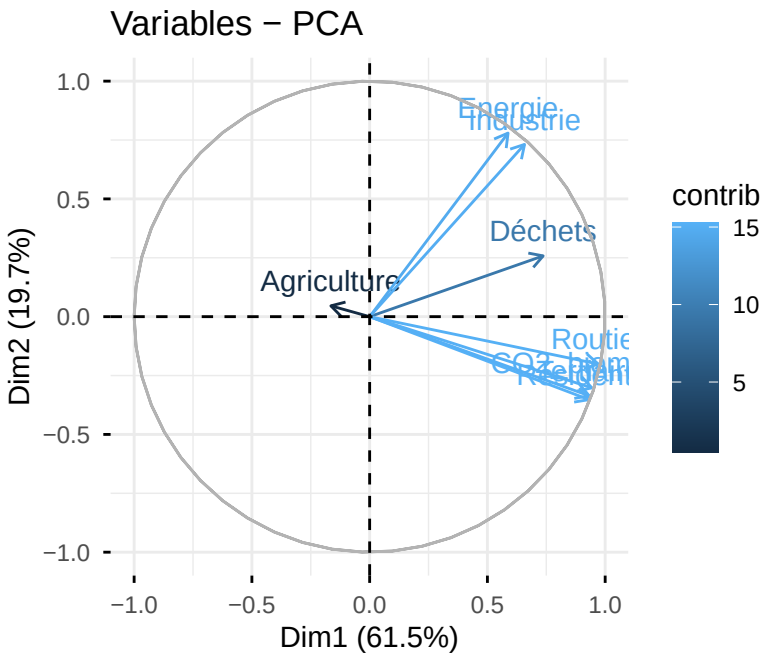
```
Warning in geom_bar(stat = "identity", fill = barfill, color = barcolor, :
Ignoring empty aesthetic: `width`.
```

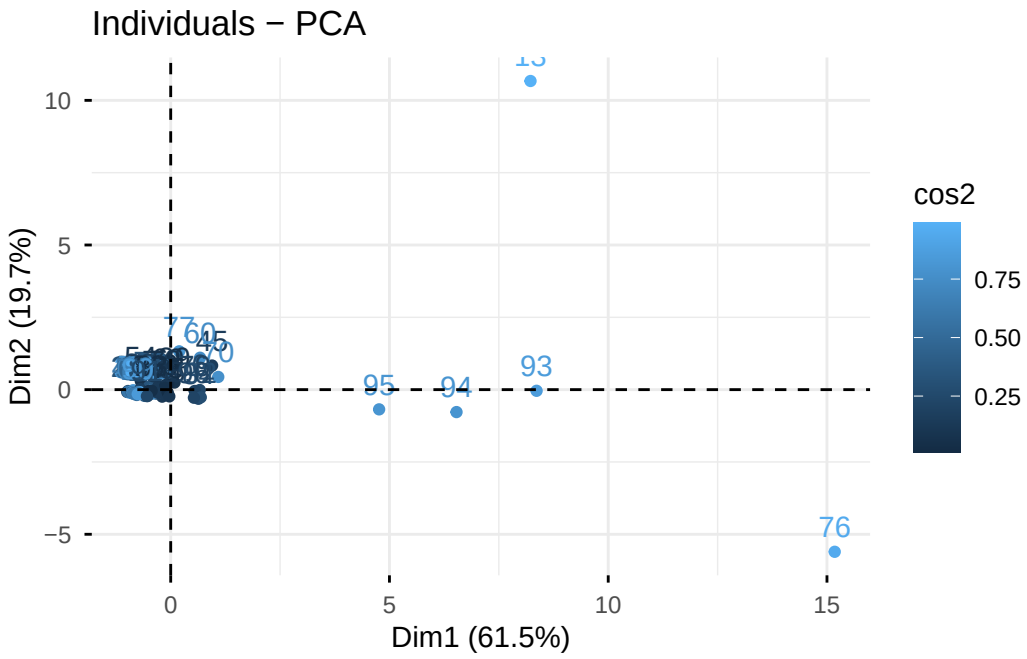


L'éboulis des valeurs propres montre une chute marquée après la deuxième composante, ce qui suggère que les deux premiers axes concentrent l'essentiel de l'information pertinente.

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
i The deprecated feature was likely used in the ggpubr package.
  Please report the issue at <https://github.com/kassambara/ggpubr/issues>.
```

```
Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
i Please use tidy evaluation idioms with `aes()`.
i See also `vignette("ggplot2-in-packages")` for more information.
i The deprecated feature was likely used in the factoextra package.
  Please report the issue at <https://github.com/kassambara/factoextra/issues>.
```

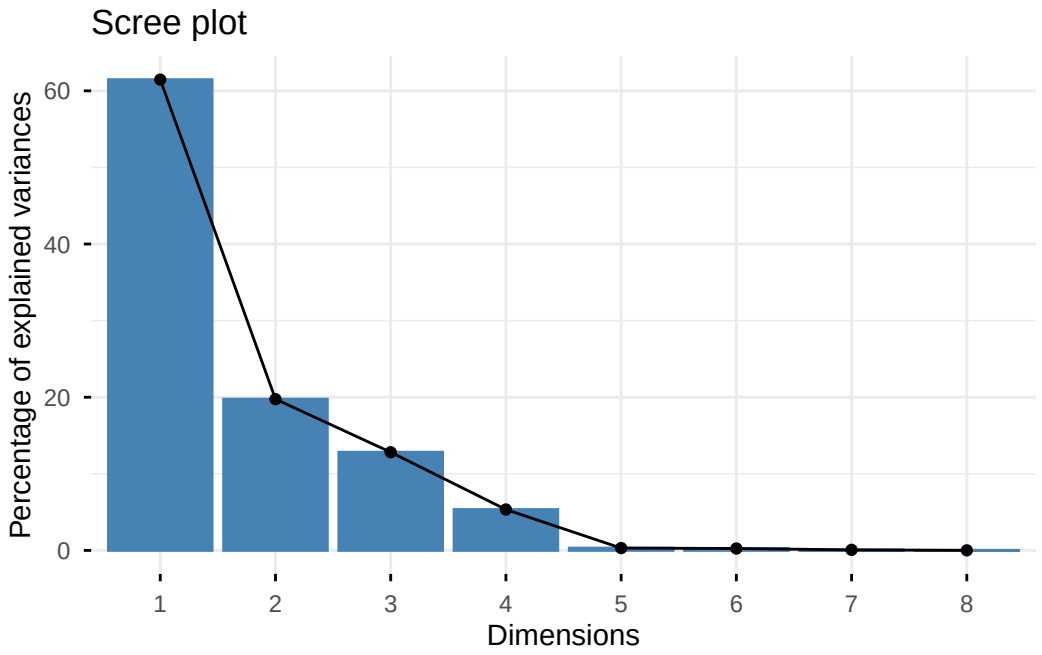




	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.9170836154	61.463545193	61.46355
comp 2	1.5797584277	19.746980346	81.21053

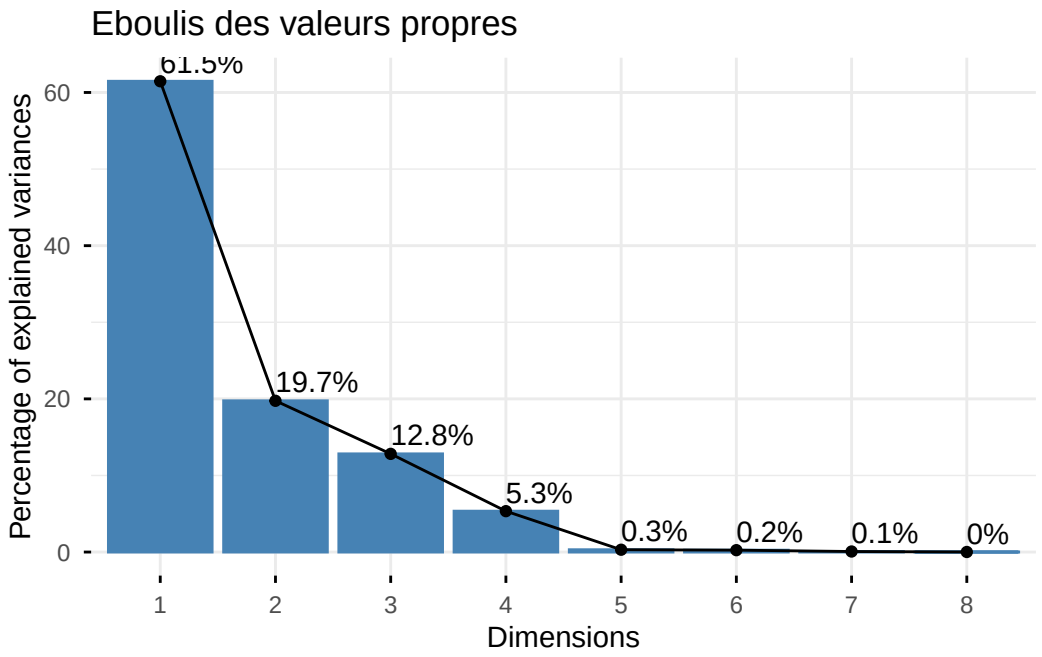
```
comp 3 1.0255247896          12.819059871                  94.02959
comp 4 0.4271011259           5.338764074                  99.36835
comp 5 0.0242978887           0.303723609                  99.67207
comp 6 0.0197720057           0.247150071                  99.91922
comp 7 0.0057397744           0.071747180                  99.99097
comp 8 0.0007223726           0.009029658                 100.00000
```

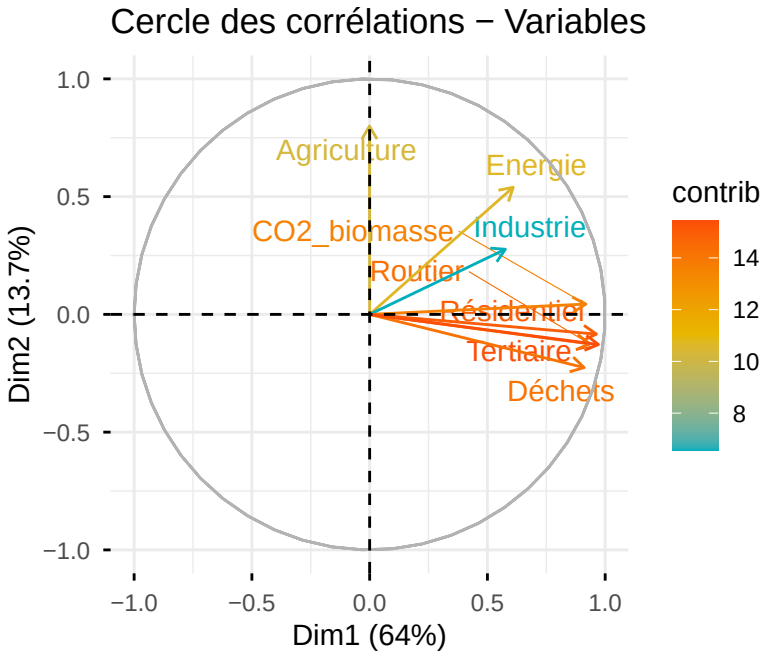
```
Warning in geom_bar(stat = "identity", fill = barfill, color = barcolor, :
Ignoring empty aesthetic: `width`.
```

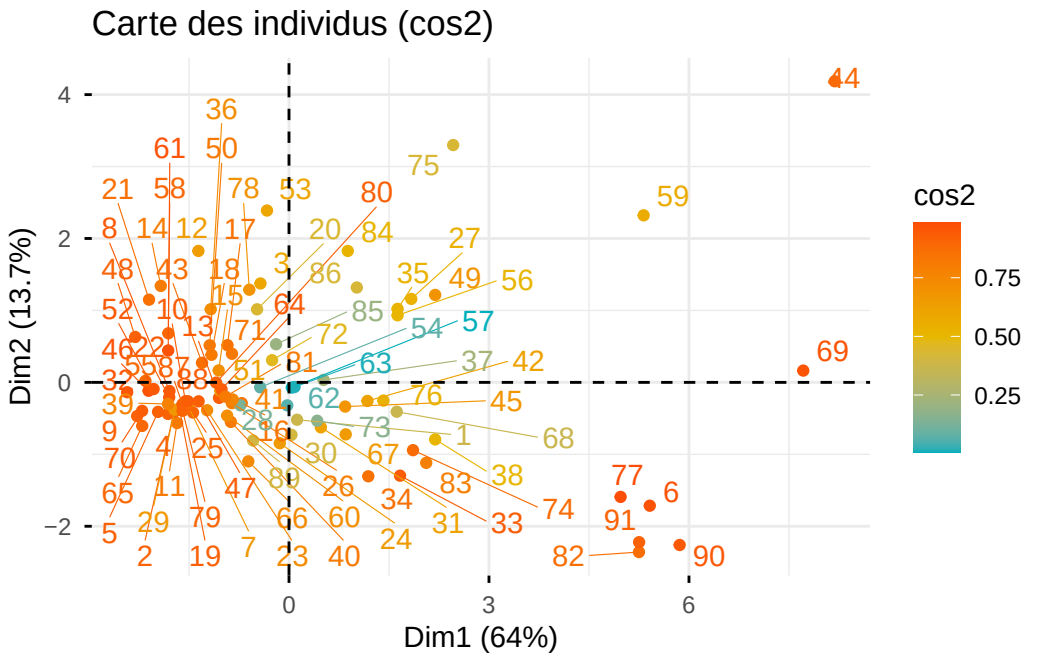


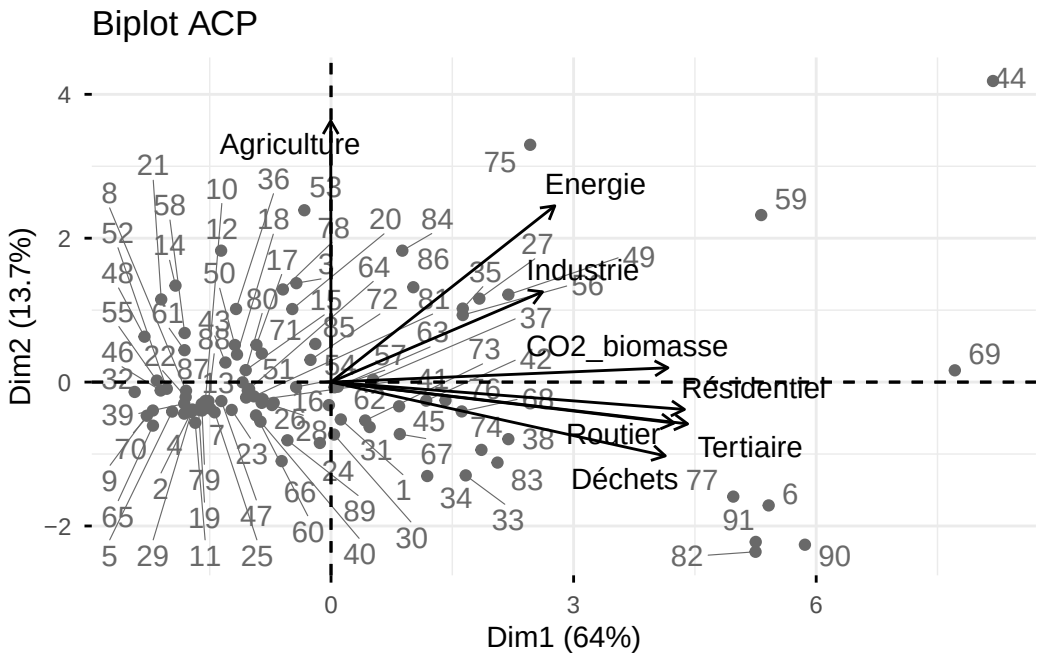
même procédé sans les valeurs extrêmes:

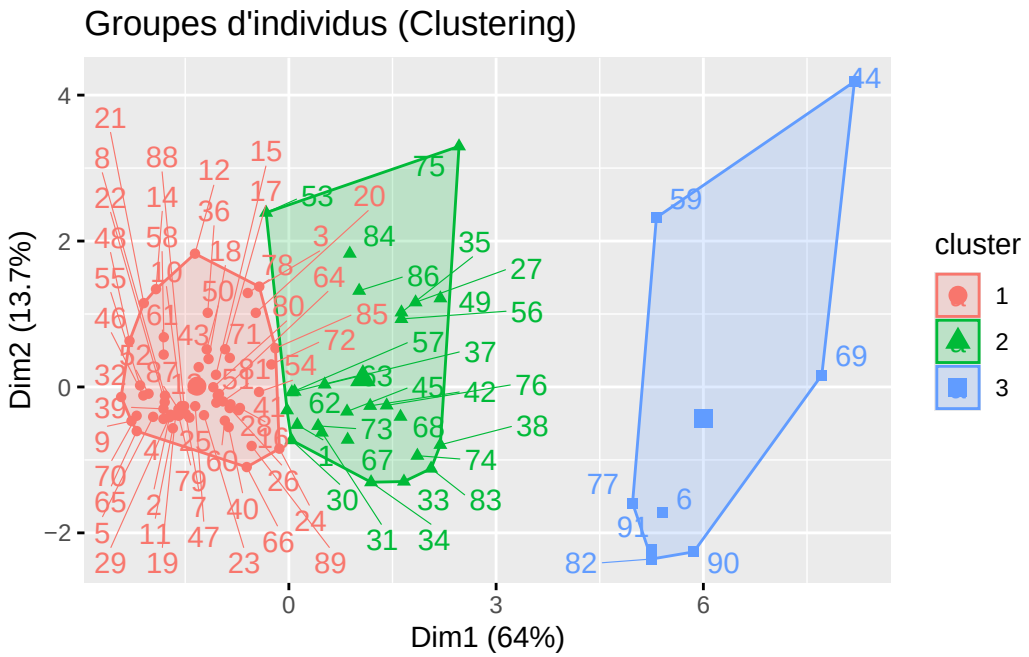
```
Warning in geom_bar(stat = "identity", fill = barfill, color = barcolor, :
Ignoring empty aesthetic: `width`.
```


Eboulis des valeurs propres
Percentage of explained variances
Dimensions
61.5%
19.7%
12.8%
5.3%
0.3%
0.2%
0.1%
0%


Cercle des corrélations − Variables
Dim2 (13.7%)
Dim1 (64%)
Agriculture
Energie
CO2_biomasse
Industrie
Routier
Résidentiel
Tertiaire
Déchets
contrib


Carte des individus (cos2)
Dim2 (13.7%)
Dim1 (64%)
cos2
0.75
0.50
0.25

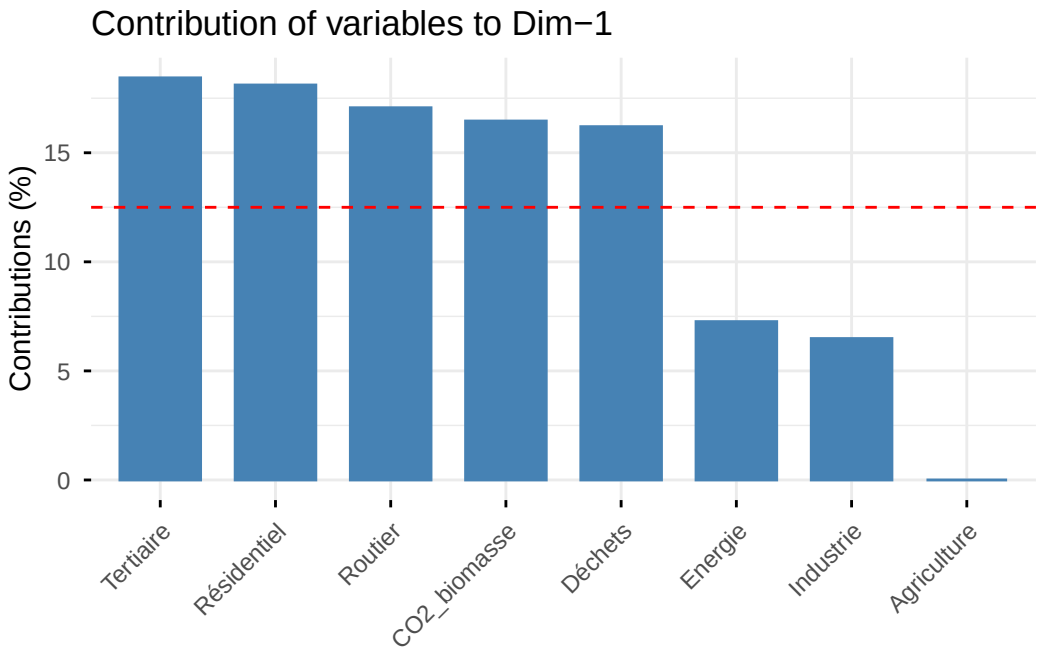

Biplot ACP
Dim2 (13.7%)
Dim1 (64%)
Agriculture
Energie
Industrie
CO2_biomasse
Résidentiel
Routier
Tertiaire
Déchets

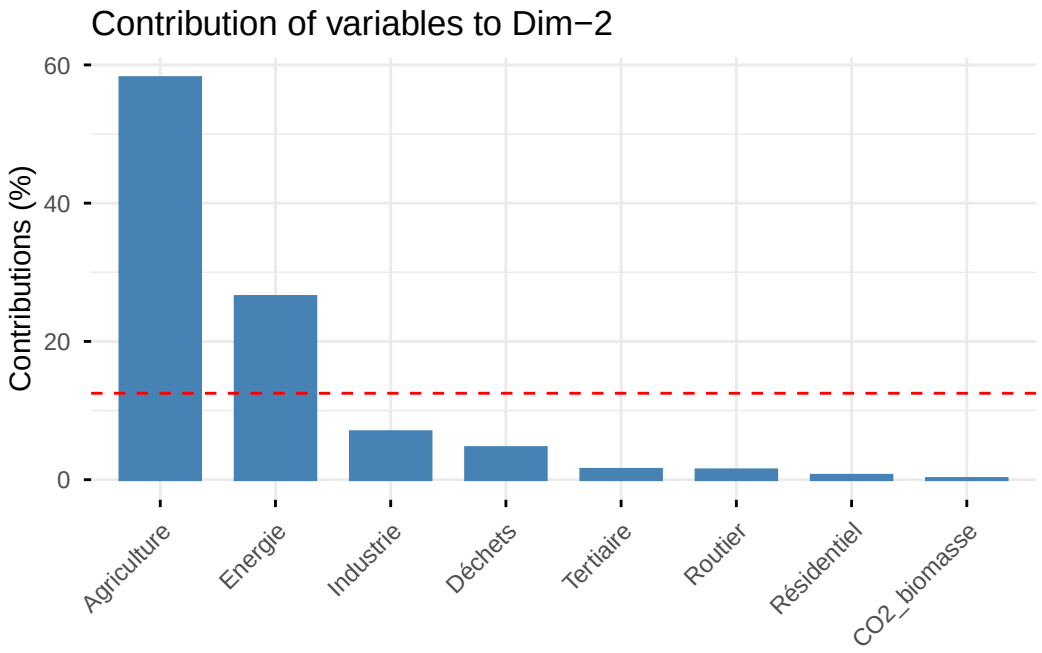


	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Tertiaire	18.434783	1.4913353	1.84467575	0.141897047	1.16976568
Résidentiel	18.106308	0.6378170	0.06868374	0.005414047	0.08671903
Routier	17.063667	1.4279680	5.68959553	0.227836288	2.58910713
CO2_biomasse	16.455521	0.1744146	2.82510831	0.377527714	69.97249929
Déchets	16.196887	4.6439950	0.70787414	3.825765101	20.97937384
Energie	7.259577	26.5070441	15.06140512	46.455631022	3.11595736

Link between the variable and the continuous variables (R-square)

	correlation	p.value
Tertiaire	0.9715441	1.852976e-57
Résidentiel	0.9628495	2.182588e-52
Routier	0.9347160	9.290446e-42
CO2_biomasse	0.9179083	1.715919e-37
Déchets	0.9106663	6.289265e-36
Energie	0.6096760	1.417360e-10
Industrie	0.5761555	2.291932e-09


Contribution of variables to Dim−1
Contributions (%)
0
5
10
15
Tertiaire
Résidentiel
Routier
CO2_biomasse
Déchets
Energie
Industrie
Agriculture


Contribution of variables to Dim−2
Contributions (%)
0
20
40
60
Agriculture
Energie
Industrie
Déchets
Tertiaire
Routier
Résidentiel
CO2_biomasse

ACP sans la variable agriculture

Afin de tester l'ACP sans la variable agriculture nous modifions ce qui est nécessaire.

Justification

KMO

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = base_dep_sans_Agriculture[, variables_quantitatives_sans_Agriculture])
Overall MSA =  0.66
MSA for each item =
CO2_biomasse      Déchets      Energie    Industrie  Résidentiel      Routier
        0.94         0.48         0.61         0.64         0.66         0.69
   Tertiaire
        0.60
```

Nous nous apperçevons que lorsque nous émettons la variable agriculture de notre modèle indice KMO passe de 0,58 à 0,66, ce qui est correct. Cette amélioration de l'indice montre le caratère atypique de la variable agriculture.

Bartlett

```
$chisq
[1] 1709.415

$p.value
[1] 0

$df
[1] 21
```

Concerant le test de bratlett celui est toujours significatif p-value < 0 ce qui confirme encore des corrélations suffisante entre les variables dont l'ACP syntétisera.

ACP

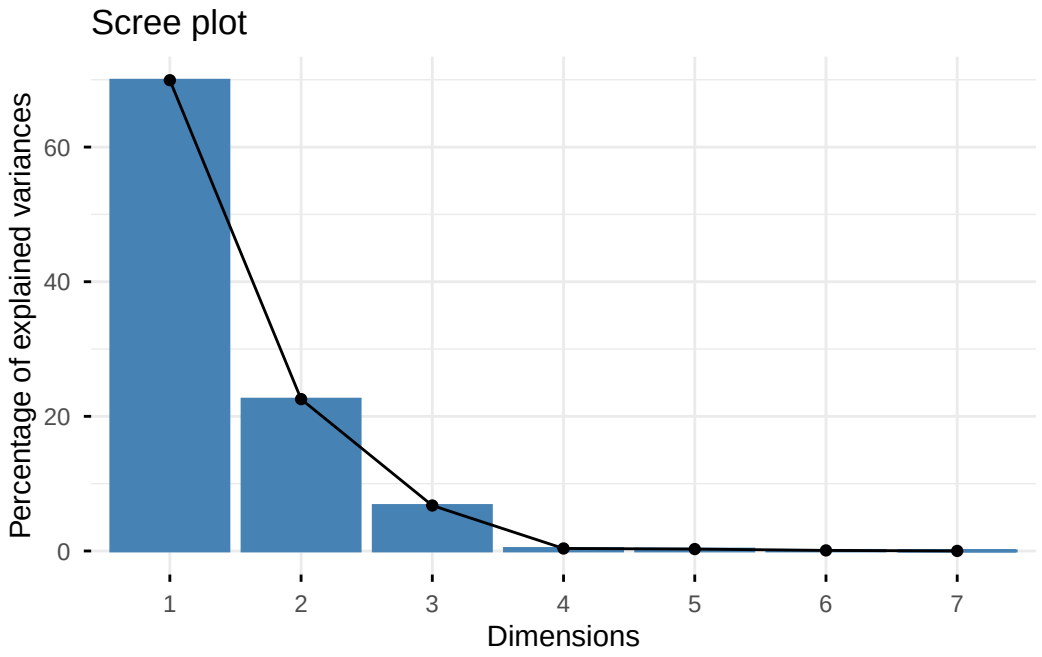
centré réduire

De la même manière que précédemment, bien que toutes les variables soient dans la même unité elles sont bien dispersé entre elle. Centré réduire est donc nécessaire afin d'éviter que les secteurs les plus émetteur dominent l'analyse.

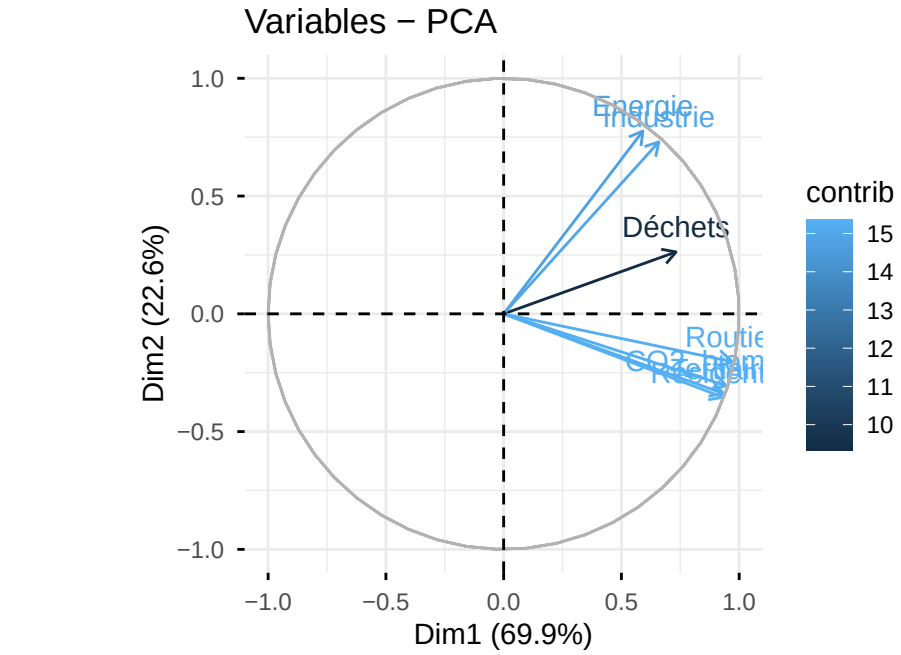
Nombre d'axes

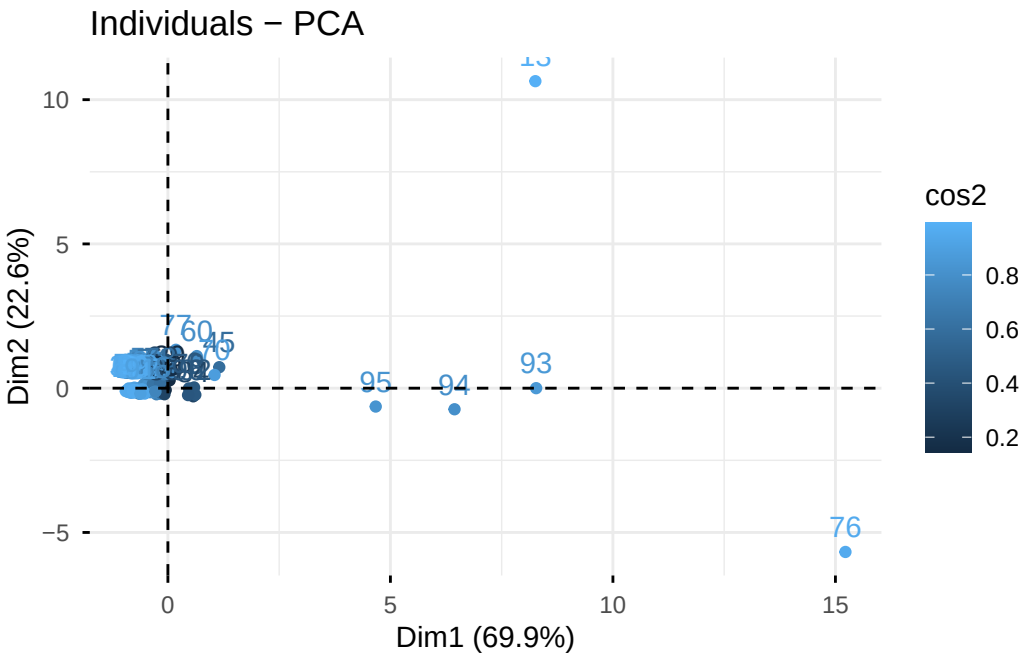
```
         eigenvalue percentage of variance cumulative percentage of variance
comp 1 4.8953928675            69.93418382                           69.93418
comp 2 1.5789811776            22.55687397                           92.49106
comp 3 0.4729437927             6.75633990                           99.24740
comp 4 0.0260614354             0.37230622                           99.61970
comp 5 0.0199015166             0.28430738                           99.90401
comp 6 0.0059942573             0.08563225                           99.98964
comp 7 0.0007249529             0.01035647                          100.00000
```

```
Warning in geom_bar(stat = "identity", fill = barfill, color = barcolor, :
Ignoring empty aesthetic: `width`.
```

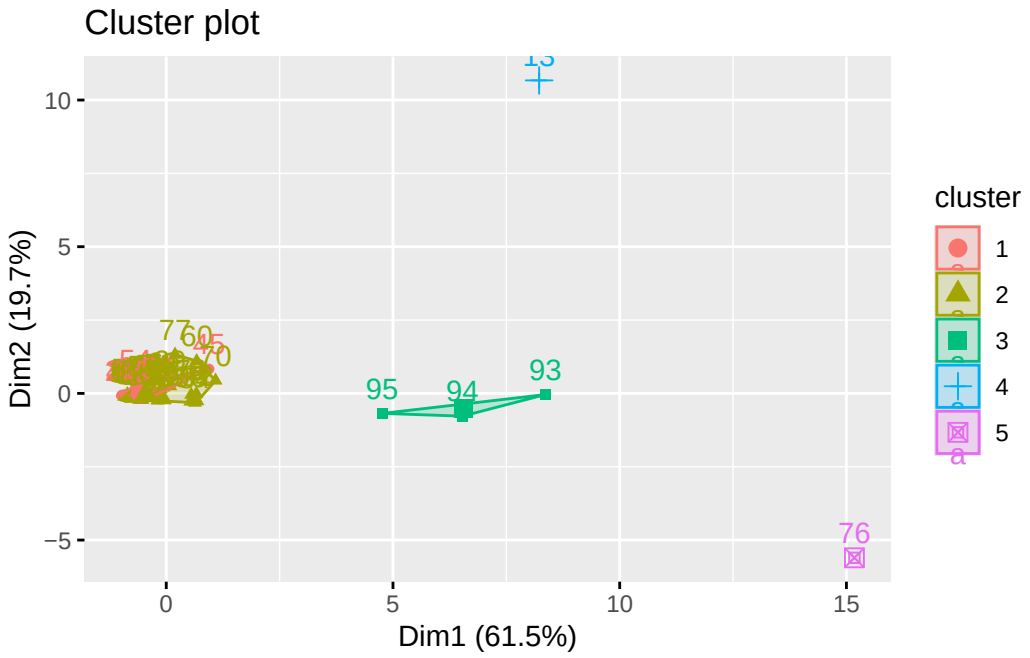


Ici encore nous décidons de garder 2 axes. Le cummul des 2 composantes principales expliquent 92,5 % de la variance totale (environ 70 % pour Dim1 et 23 % pour Dim 2) ce qui est très bon, ils synthétise l'essentiel de l'information présente dans le modèle.





Nous remarquons sur la carte des individus/départements que certains d'entre eux sont très éloigné du profil moyen 95, 94, 93, 76, 13, ils ont surment des psécificité teritorielle. Sur l'axe 1 ils sont tiré vers la droite POURQUOI ????? et sur l'axe 2 certains comme le 13 sont tiré vers le haut et certains comme le 76 tiré vers le bas quant au 93 94 et 95 ils sont proche de zéro sur l'axe des abcisses.



3 groupes identifiés

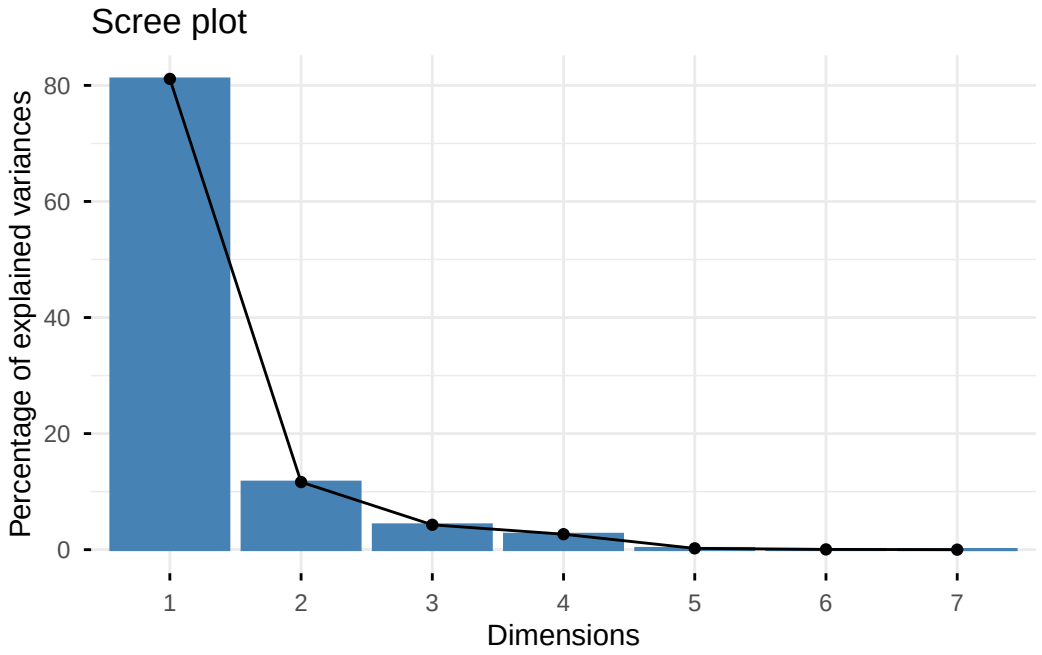
Mais la présence des départements atypiques masque un peu la distinction → il faudra envisager une ACP sans ces départements.

ACP sans les variables atypiques

Les 5 départements ont été retiré

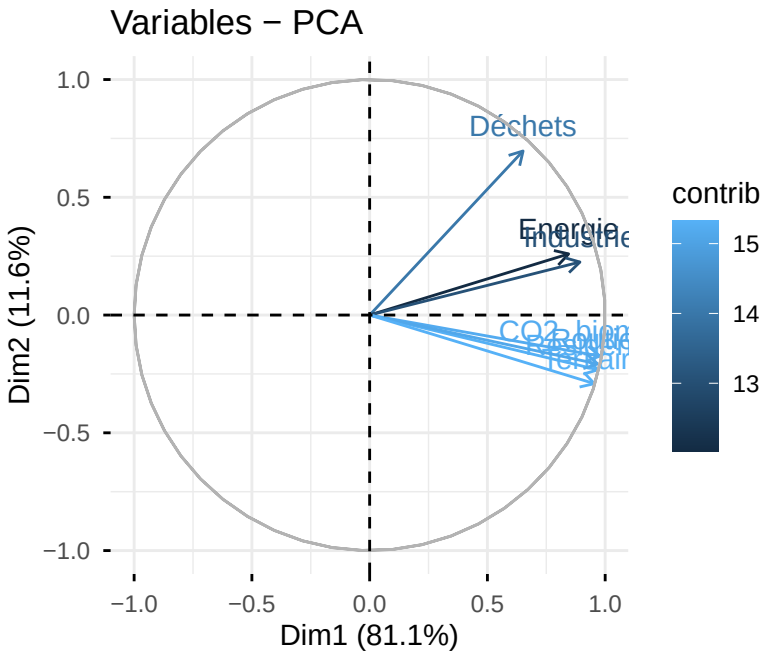
```
        eigenvalue percentage of variance cumulative percentage of variance
comp 1 5.678358469           81.119406693                            81.11941
comp 2 0.815498615           11.649980221                            92.76939
comp 3 0.300003657            4.285766527                            97.05515
comp 4 0.186702872            2.667183887                            99.72234
comp 5 0.015879557            0.226850816                            99.94919
comp 6 0.003204505            0.045778642                            99.99497
comp 7 0.000352325            0.005033214                           100.00000
```

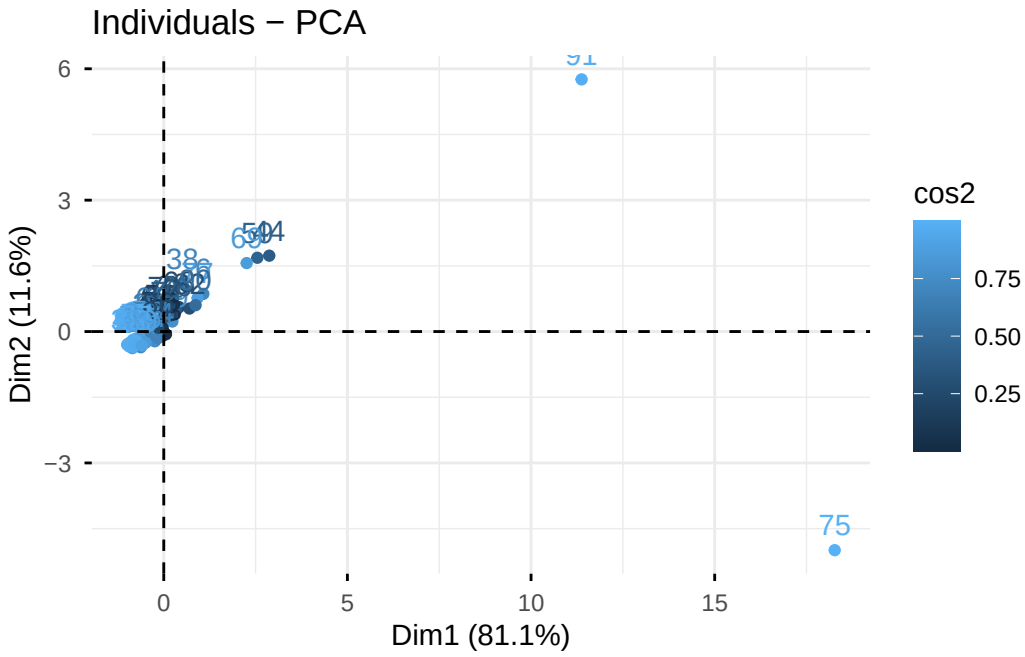
```
Warning in geom_bar(stat = "identity", fill = barfill, color = barcolor, :
Ignoring empty aesthetic: `width`.
```



avec 2 axes 92,77%

Graphique

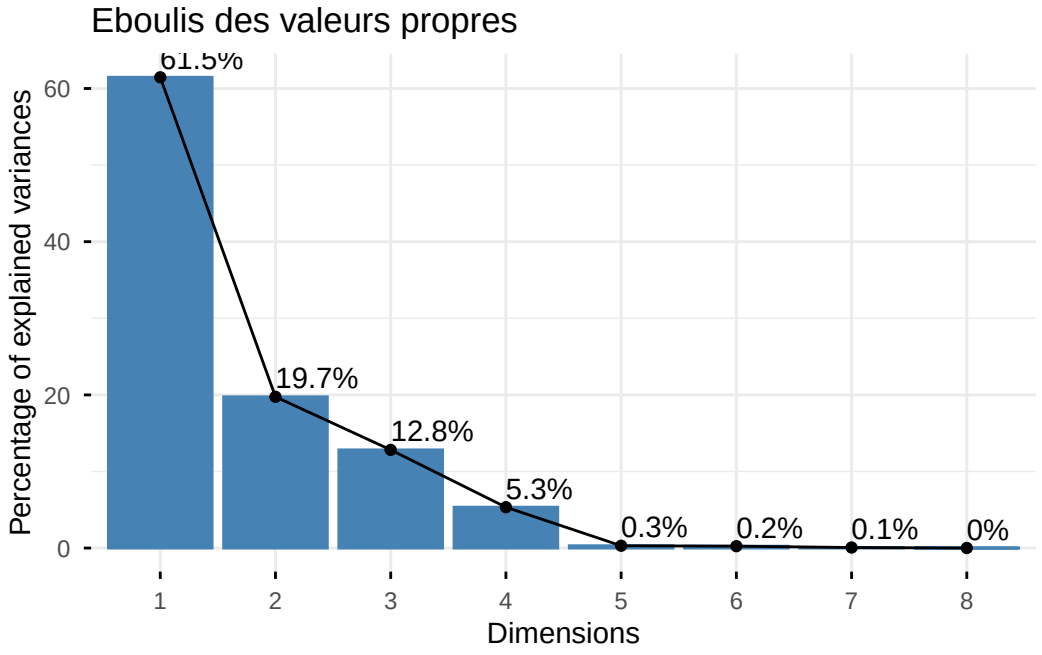


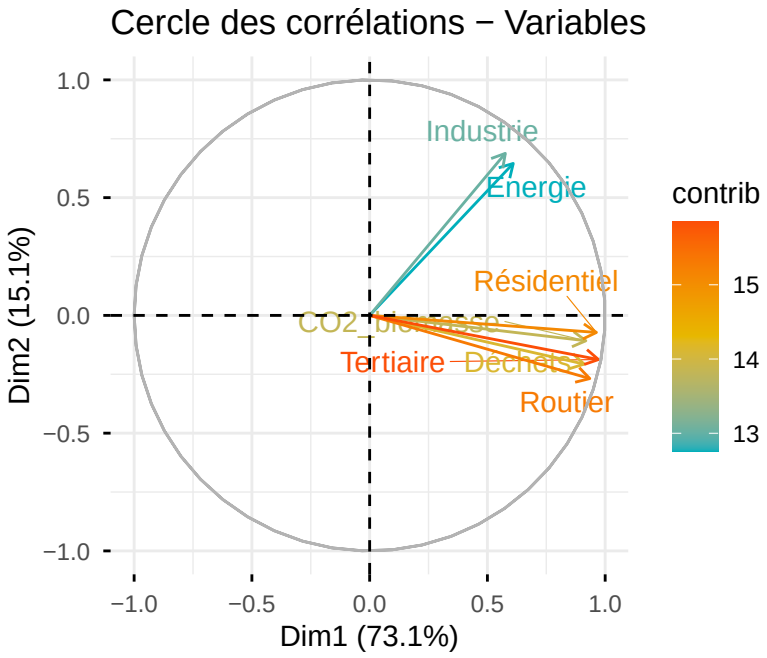


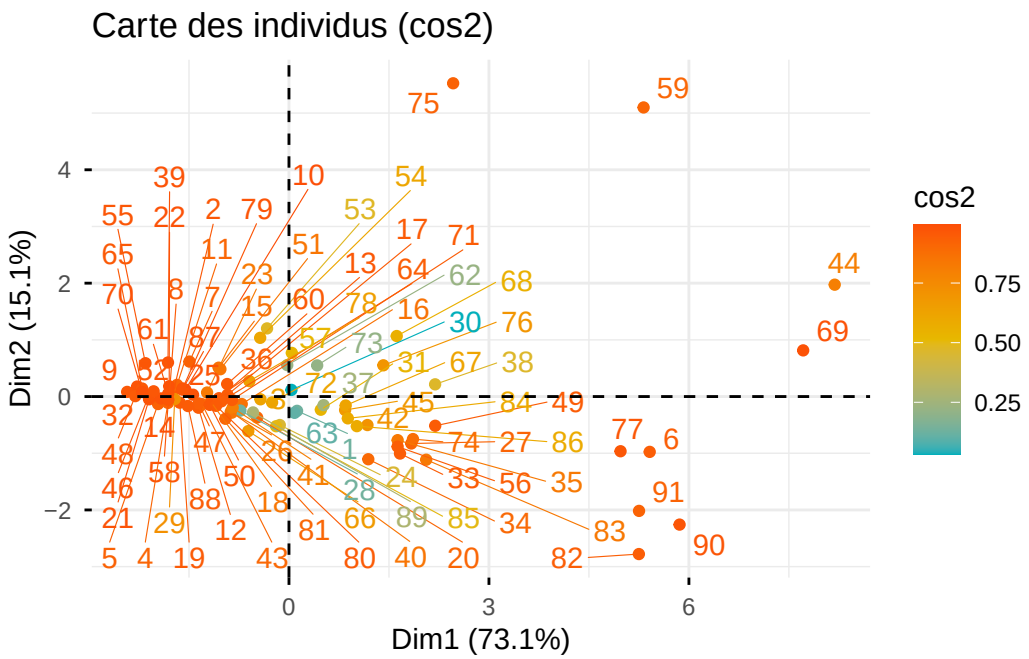
91 et 75 sont eux aussi dispersé loins du profil moyen BIZZARE

même procédé sans les valeurs extrêmes:

```
Warning in geom_bar(stat = "identity", fill = barfill, color = barcolor, :
Ignoring empty aesthetic: `width`.
```







on obtient pas pareil a cause de repel

Conclusion