UNIVERSITÉ SORBONNE PARIS NORD

DM : Projet Initiation à R

Étude des déterminants de la performance environnementale des entreprises de EUROSTOXX 600

Fait par : BALTAGI Reina

Master 1- Parcours BIDABI

Le 10/11/2024

Introduction

Pour commencer, la première étape était de définir le répertoire de travail, avec la commande stewd(), puis, afin de permettre à R de lire le fichier Excel, il était primordiale d'installer le package read.xl. Après avoir donné à R les outils pour pouvoir lire le fichier « variables_EuroStoxx600.xlsx », et après avoir nommé la data frame exporté d'Excel « perf_envi », il fallait redéfinir les variables que nous allons utiliser tout au long de notre afin de rendre la tâche plus simple. Comme figurant dans le script, les variables ont été redéfinie par affectation, la prochaine étape de transformer les objets, à savoir ici les variables en de valeurs numériques à l'aide de la commande as.numeric. Puis il était important d'éliminer les NA afin d'éviter toute confusion pendant les calculs, c'est ainsi où il y a eu une détermination des NA en premier avec la commande any(is.na(perf_envi)), puis l'élimination avec la commande omit.na(perf_envi). C'est ainsi qu'il était possible d'aborder les questions, notamment les questions 3 et 4.

Question 1

D'après Senadheera et al. (2021), le pilier environnemental (EPS) constitue une approche clé pour assurer une gestion durable de l'entreprise, car elle permet de repérer les problématiques écologiques largement influencées par les activités industrielles. Pour mieux expliquer, ce pilier prend en compte une variété d'indicateurs, tels que les déchets, les émissions, le changement climatique et la gestion des risques. D'après Senadheera et al. (2021), les scores environnementaux fournis par les fournisseurs de données correspondent souvent aux niveaux d'émissions environnementales. Ce qui fait que la variable **Carbon Intensity per Energy Produced,** fait déjà partie de l'EPS. De plus, il est plus simple de recueillir des données sur l'EPS que sur la **Carbon Intensity per Energy Produced.** Sans oublier le fait que l'EPS fait directement partie de la norme ESG (environnement, social, gouvernance), et il s'agit d'une mesure globale de la performance environnementale d'une entreprise, qui couvre différents aspects environnementaux

Question 2

Partie 1 : Justification de l'utilisation de l'ACP

L'ACP, analyse en composantes principales est une méthode permettant de réduire la dimensionalité des données étudiées, dans le cas de cette étude, il s'agit d'une synthèse des facteurs influant la variation des prix. Les variables en question représentent la variation de prix sur différentes périodes, à savoir (1 jour, 52 semaines, 4semaines, 5 jours, etc.), elles contiennent également des variables en rapport avec le pourcentage de variation du prix depuis le début de l'année ou même le prix le plus élevé obtenu au bout de 52 semaines. Donc en tout, il y a 8 variables financières, l'application de l'ACP permettra de conserver l'essentiel de l'information à travers des composantes principales extraites à partir des variables étudiées.

Généralement parlant, **l'ACP ne s'applique que sur les variables quantitatives corrélées entre-elles,** de ce fait, afin de justifier l'utilisation d'une telle méthode statistique dans cette étude il est important de jeter un coup d'œil sur la corrélation entre les variables. Pour faire cela, cette étude présentera une analyse en paire de la matrice de corrélation des variables et de son corrplot correspondant, ainsi que

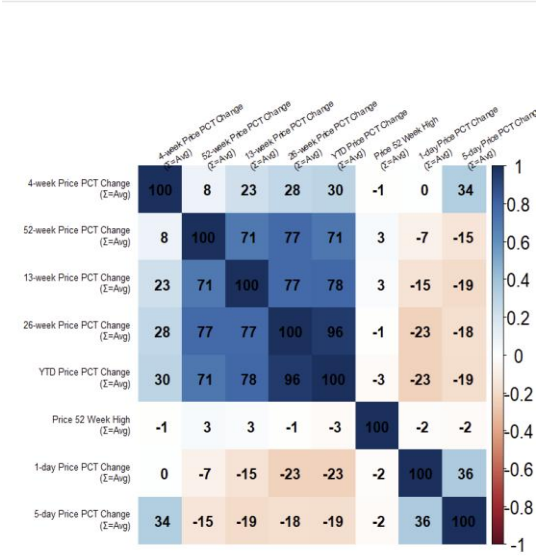
les résultats du test de Bartlett qui attestera de la présence ou pas de corrélation permettant l'application de l'ACP.

La matrice de corrélation a été obtenue par le code suivant :

matrice_correlation <- cor(var_acp, use = "complete.obs")

print(matrice_correlation)

Et la meilleure façon pour visualiser ses résultats est à partir du corrplot suivant :



Dans ce Corrplot montrant les pourcentages de corrélation entre les différentes variables, il est détectable que les variables montrant une variation de prix sur des périodes longues (13, 26 et 52 semaines) sont fortement corrélés entre-elles (avec des coefficients de corrélation supérieurs à 70%, et des cases en bleu foncé)

Les variations à court terme (1 jour, 5 jours) montrent des corrélations faibles voire négative avec le reste des variables (à travers les cases à couleurs plus vert l'orange et des pourcentages négatifs ou très faibles)

La variable Price 52 Week High semble moins corrélée avec les autres, ce qui suggère qu'elle apporte une information différente.

Pour interpréter ces résultats et savoir en quoi l'ACP est applicable dans ce cas, il est important de savoir que la forte corrélation entre les variables montrant la variance des prix à de longues durées, signifie une redondance dans l'information communiquée par ces variables, d'où l'importance de l'utilisation de l'ACP pour trouver les facteurs synthétiques réduisant la redondance. Toutefois, les variations à court termes (1 day et 5 days) ainsi que la variable Price 52 week high montrant des corrélations faibles montrent l'apport d'informations différentes. Leur rôle dans l'ACP pourra être étudié afin de voir si elles se regroupent dans une autre composante ou si elles restent isolées. Cette redondance de l'information avec les variables à long termes, et l'apport de nouvelles information avec celles à court termes et la variable 52 weeks high montrent en quoi l'ACP est essentiel à mettre en place pour réduire cette redondance, donc cette dimensionalité, rendre l'analyse plus simple en se focalisant uniquement sur les composantes principales qui expliquent le plus l'information sans rentrer dans la complexité d'une analyse directe.

De plus, le test de Bartlett a donné les résultats suivants :

```
> bartlett_test <- bartlett.test(var_acp)
> print(bartlett_test)

        Bartlett test of homogeneity of variances

data:  var_acp
Bartlett's K-squared = 75044, df = 7, p-value < 2.2e-16
```

Avec une P-value aussi faible, l'hypothèse H0 d'absence de corrélation doit être rejeter, dans ce cas, ce test montre en quoi les variables sont suffisamment corrélées pour pouvoir réaliser une analyse en ACP.

L'ACP est donc la méthode adaptée pour résumer les variations de prix sur différentes périodes. Elle réduit la dimensionnalité des données, élimine la redondance entre les variables à long terme et extrait des facteurs explicatifs clés. Les résultats de la matrice de corrélation, du corrplot et du test de Bartlett confirment que l'utilisation de l'ACP est pertinente pour cette analyse. Dans le but d'identifier le facteur important qui peut synthétiser les variations des variables, l'ACP sera réalisé dans ce qui suit.

Partie 2 : A faire l'analyse des composantes principales et présenter les principaux résultats

Pour la réalisation de l'ACP, plusieurs étapes ont été suivie, les étapes ainsi que les résultats, seront présentées au fur et à mesure de cette analyse.

Premièrement, il est important de savoir qu'une étape essentielle avant de mettre en place l'ACP est de standardisé les variables à une moyenne 0 et un écart-type 1 pour éviter que les échelles différentes n'influent les résultats. 2 méthodes pour faire cela soit dans le code de l'ACP directement (avec scale.unit=TRUE), soit en code avant celui de l'ACP avec la fonction scale (les deux méthodes sont détaillés dans le code, ici, il n'y aura que celle qui a été utilisé).

```
acp_var_centrees_reduites<-scale(var_acp,center = TRUE, scale= TRUE)
```

Puis l'ACP a été réalisé pour extraire les composantes principales expliquant la variation des variables :

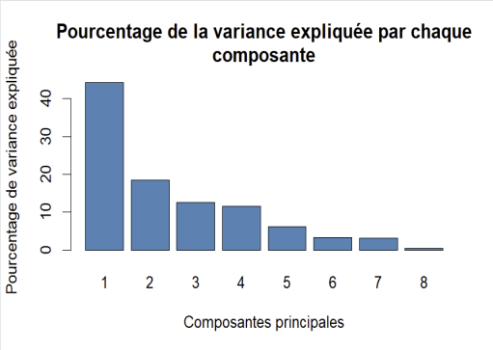
```
resultats_acp<-PCA(acp_var_centrees_reduites, graph = F)
```

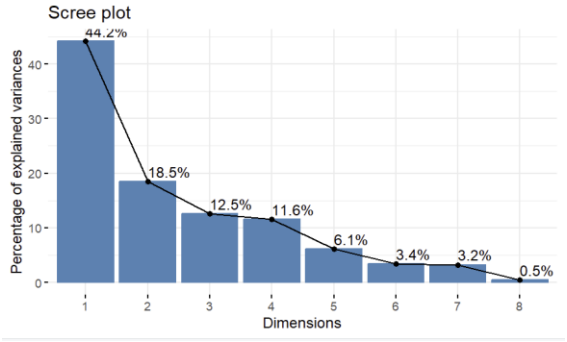
L'étape suivante est celle de la détermination du nombre de composante à retenir, pour faire cela, les valeurs propres doivent être consultées, ainsi que le pourcentage cumulatif de variance.

```
> print(resultats_acp$eig)
       eigenvalue percentage of variance cumulative percentage of variance
comp 1 3.53481611             44.1852014                            44.18520
comp 2 1.47832951             18.4791188                            62.66432
comp 3 1.00259659             12.5324574                            75.19678
comp 4 0.92748349             11.5935436                            86.79032
comp 5 0.48892215              6.1115269                            92.90185
comp 6 0.27113961              3.3892451                            96.29109
comp 7 0.25939379              3.2424223                            99.53352
comp 8 0.03731875              0.4664844                           100.00000
>
```

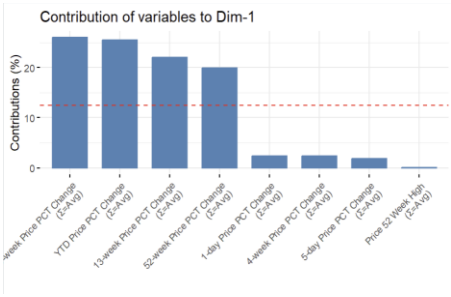
En se basant sur le percentage of variance et le cumulative percentage of variance, nous pouvons voir que les composantes 1 et 2 expliquent 62.66% de la variance des variables, avec 44.19% expliqué par la composante 1 et 18.48% expliquée par la composante 2. Ces composantes sont suivies par la composante 3 expliquant 12.53% de la variance des variables.

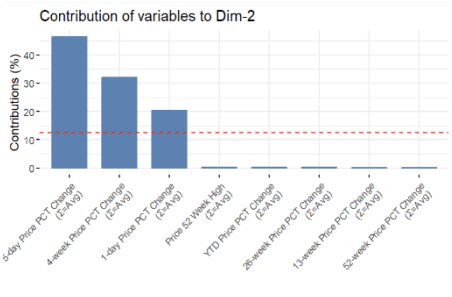
De plus, en se basant sur les graphiques suivants, particulièrement le deuxième graphique :





Nous avons choisi de nous limiter aux deux premières composantes car elles expliquent ensemble 62.66% de la variance totale (44.2% pour la première composante et 18.5% pour la deuxième). Cette part majoritaire de l'information justifie le fait que nous les avons choisis pour l'analyse. De plus, selon la règle du coude, et en se basant sur le graphe 2, il y a une forte décroissance de la variance expliquée pour les deux premières dimensions, suivie d'un aplatissement à partir de la troisième dimension (12.5%), indiquant que les dimensions supplémentaires n'apportent qu'une contribution marginale. Enfin, une analyse de deux dimensions est plus simple et plus pratique que celle de 3 dimensions, et vu la part marginale non significative expliquée par la composante 3, seules les 2 premières composantes seront retenues. Ainsi, travailler en deux dimensions facilite la lecture des graphiques et l'interprétation des résultats, sans compromettre significativement la qualité de l'analyse. Ce choix représente un bon compromis entre la simplicité et la pertinence des données.

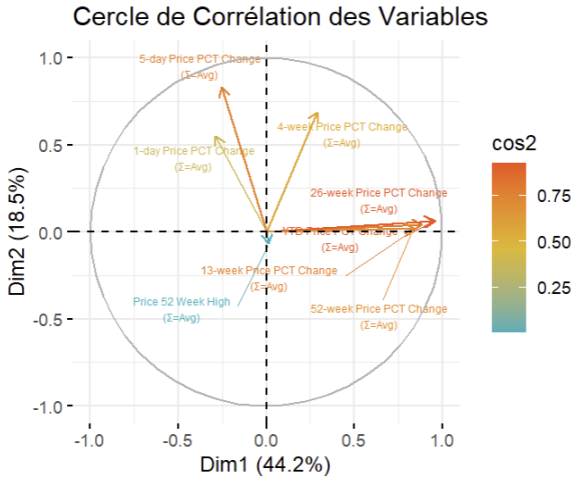




Concernant la contribution des variables aux axes principaux (1 et 2), il est clair d'après les graphiques que les variables comme 26 weeks, YTD price, 13 weeks et 52 weeks, ont le plus contribué à la construction de l'axe 1, avec des pourcentages dépassant les 20% pour les 26 semaines et YTD. En revanche, pour l'axe 2, les variables comme 5 day Price change, 4 weeks price change, et 1 day price change, sont les seules à contribuer significativement à la construction de cet axe, avec plus de 45% de contribution de la part de la variable 5 days change. C'est ainsi qu'une conclusion primaire, qui sera par la suite validée par l'analyse du cercle de corrélation, peut avoir lieu, l'axe 1 synthétise principalement les variables de variation de prix à long terme, alors que l'axe 2 synthétise celles de la variation de prix à court terme. Il est important de mentionner plus la variable contribue à un axe, plus elle influence la variation expliqué par cet axe.

```
$cos2
                                          Dim.1        Dim.2        Dim.3
1-day Price PCT Change\r\n(Σ=Avg)    0.0839235012 0.3014916125 6.996887e-03
5-day Price PCT Change\r\n(Σ=Avg)    0.0652138632 0.6882759310 1.230584e-03
4-week Price PCT Change\r\n(Σ=Avg)   0.0831185968 0.4740593931 7.800149e-05
13-week Price PCT Change\r\n(Σ=Avg)  0.7775745985 0.0018150453 1.533568e-03
26-week Price PCT Change\r\n(Σ=Avg)  0.9202023903 0.0036569189 4.228152e-04
YTD Price PCT Change\r\n(Σ=Avg)      0.9006442079 0.0038547770 2.141549e-03
52-week Price PCT Change\r\n(Σ=Avg)  0.7040048274 0.0004977064 4.042401e-03
Price 52 Week High\r\n(Σ=Avg)        0.0001341276 0.0046781220 9.861508e-01
                                          Dim.4        Dim.5
1-day Price PCT Change\r\n(Σ=Avg)    0.5033943654 9.884476e-02
5-day Price PCT Change\r\n(Σ=Avg)    0.0011357279 2.361034e-01
4-week Price PCT Change\r\n(Σ=Avg)   0.3030870386 1.156671e-01
13-week Price PCT Change\r\n(Σ=Avg)  0.0104503169 3.792620e-03
26-week Price PCT Change\r\n(Σ=Avg)  0.0002787208 1.770420e-03
YTD Price PCT Change\r\n(Σ=Avg)      0.0002914041 4.946005e-05
52-week Price PCT Change\r\n(Σ=Avg)  0.1010485850 3.268912e-02
Price 52 Week High\r\n(Σ=Avg)        0.0077973331 5.339065e-06
```

Dans cette sortie de code montrant les valeurs des cos2 associés à chaque variable, des informations sur la qualité de représentation des variables sur les axes correspondants peut être retirée, un cos2 proche de 1 signifie une très bonne qualité de représentation de la variable sur l'axe et un faible cos2 c'est l'inverse. De ce fait, nous pouvons observer que les variables 1 day, 5-days, 4 weeks change, sont les mieux représentés sur l'axe 2 avec les cosinus carré les plus élevé à savoir respectivement, 0.3, 0.5 et 0.7 (valeurs arrondies) en précisant que la variable 4 weeks est la mieux représentée sur l'axe 2 parmi ces 3 puisqu'elle a le cos2 le plus proche de 1. Sur l'axe 1, les variables les mieux représentées sont 13 weeks, 26 weeks, YTD et 52-weeks avec des cos2 conséquent entre 0.7 et 0.9 tous proches de 1 donc, ces variables sont toutes bien représentés sur la composante 1.

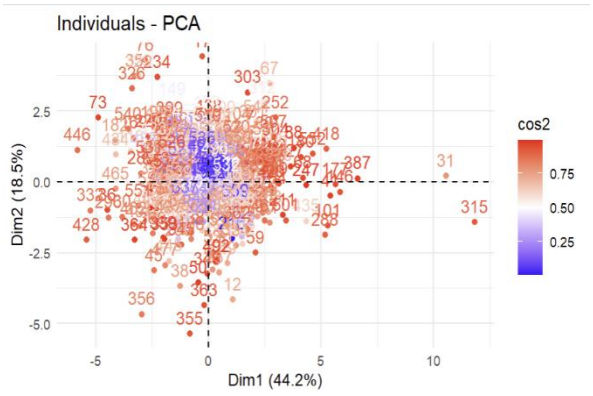


Dans ce cercle de corrélation montrant la corrélation des variables avec les axes 1 et 2, plusieurs éléments seront abordés, notamment qu'il s'agit du graphique principal de l'ACP avec des informations supplémentaires, à savoir le cos2 donc la qualité de représentation de chaque variable dans le plan factoriel. Plus le cos2 est élevé (proche de 1) mieux la variable est représentée dans le plan, dans le cas de ce cercle plus la couleur de la flèche tend vers l'orange foncé, mieux la variable est représentée dans le plan, alors que les variables en bleu ne le sont pas. C'est ainsi que nous pourrons détecter que la plupart des variables sont bien représentée dans le plan, à l'exception de Price 52 Week High qui ne l'est pas, puisqu'elle a une couleur bleu donc un cosinus carré faible aux alentours de 0.25. Les variables 1 day price change et 4 week price change ont une qualité de représentation intermédiaire puisque leurs flèches sont de couleur jaune et jaune foncé respectivement avec des cosinus carré aux alentours des 0.5-0.6.

Quant à l'analyse des variables elles-mêmes dans ce cercle de corrélation, il est important de noter que plus une flèche est longue, mieux la variable est représentée dans le plan factoriel, dans ce cercle, la variable 26 weeks semble être bien représentée dans le plan. La direction des flèches et leurs proximités

donnent des indications sur la corrélation entre les variables, à savoir les variables comme celles de longues durée, proches l'une de l'autre dans le cercle, sont fortement corrélées entre elles. Et si une flèche est proche d'un des axes elle contribue principalement à cette dimension, à savoir 26 weeks avec la composante 1.

Pour l'interprétation générale du cercle, la dimension 1 semble capturer une dynamique de long terme dans le sens où les variables de variation de prix sur les longues périodes sont celles qui contribuent le plus à cette dimension. Tandis que, la dimension 2 semble capturer une dynamique différente plutôt celle liée aux variations à court terme. Cela revient au fait que les variables à long terme comme 13 weeks, 26 weeks, YTD et 52 weeks sont fortement corrélés à F1 (très proche de l'axe 1), alors que les variables comme 1 day, 5 days et 4 weeks, sont plutôt corrélés avec F2 (proche de l'axe 2).



Ce graphe montre le nuage des individus dans le plan factoriel issu de l'ACP, montrant la position des individus selon leurs contributions à la variance ainsi que leur qualité de représentation cos2. Chaque point représente un individu de l'échantillon, dans le cas de ces données, il s'agit d'entreprises. Il est important de savoir que les points proches du centre ont une contribution plus faible à la variance des dimensions principales, alors que les points éloignés ont une contribution plus importantes et peuvent être considéré comme ayant des caractéristiques distinctes, il s'agit ici du cas des points 31, 315 et 355. Les couleurs des points montrent leur qualité de représentation dans le plan factoriel, à savoir, les points rouge sont ceux ayant un cos2 proche de 1 donc sont les mieux représentés dans le plan factoriel, alors que ceux de couleurs mauve, proche de l'origine, ces points sont les moins bien représentés dans le plan factoriel. Pour la distribution globale, nous pouvons distinguer une distribution relativement homogène des entreprises, avec 31, 315 et 355 éloignées du centre ce qui peut signifier qu'ils peuvent potentiellement nécessiter une analyse approfondie, puisqu'elles ont une forte influence sur la variance des variables.

Quant au croisement du graphique des individus et du cercle de corrélation, il est important de savoir que, les entreprises qui tendent vers la droite selon la dim1 auront des fortes valeurs pour les variables 13 weeks, 26 weeks, 52 weeks et YTD (des variations à long termes), et celles qui tendent vers la gauche pour cette même dimension, auront des valeurs faibles pour ces mêmes variables. Toutefois, les entreprises qui tendent vers le haut suivant la dim2 auront des fortes valeurs pour les variables 1 day, 5 days et 4 weeks, alors que celles qui tendent vers le bas selon cette même composante, auront des valeurs plutôt faibles pour ces même variables correspondantes. Cela permettra de créer des typologies selon les ressemblances des entreprises en question.

En conclusion, cette ACP nous a permis de voir en quoi la variance des données est fortement structurée autour de deux axes principaux qui synthétisent le mieux les variables en question. Le premier axe est la composante 1 capturant les variations de prix à long terme, et la composante 2 capturant les variations

de prix à court terme. Ces composantes synthétisent très bien la variation des variables, comme ils expliquent 62.66% de la variation des variables en question.

Partie 3 : Extraire la première composante Fi comme une mesure financière de la variation des prix.

Dans cette partie, le code pour extraire la Fi est le suivant :

pc1 <- resultats_acpindcoord[, 1]

Après avoir fait l'extraction de Fi, il est intéressant de l'intégrer la régression linéaire afin de voir comment ça va influencer le modèle : (Voir question 4, Régression linéaire modèle 1)

mod_updated <- lm(EPS ~ BOARD + COD + CSR + EMP + MARGIN + MCAP + RCAP + SALES + pc1, data = perf_envi)

summary(mod_updated)

```
CODPO       -7.197e+00  7.730e+00  -0.931  0.35222
CODPT        1.353e+01  1.131e+01   1.196  0.23224
CODSD       -4.189e+00  3.783e+00  -1.107  0.26869
CODSW       -1.153e+00  3.591e+00  -0.321  0.74823
CODUK       -3.274e+00  2.988e+00  -1.096  0.27377
CSRNo       -1.992e+01  2.308e+00  -8.632  < 2e-16 ***
EMP          2.376e-05  1.285e-05   1.849  0.06499 .
MARGIN      -3.004e-02  1.573e-02  -1.910  0.05668 .
MCAP         1.939e-11  1.188e-11   1.633  0.10315
RCAP        -4.723e-02  4.059e-02  -1.163  0.24517
SALES        6.733e-08  2.441e-08   2.759  0.00601 **
pc1          1.480e+00  4.618e-01   3.205  0.00143 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.04 on 524 degrees of freedom
Multiple R-squared:  0.3705,    Adjusted R-squared:  0.3368
F-statistic: 11.01 on 28 and 524 DF,  p-value: < 2.2e-16
```

En comparaison avec la régression linéaire faite à la fin de ce devoir, à savoir dans la question 4, il est intéressant de voir en quoi le Adjusted R-squarred a évolué. En effet, le adjusted R-squarred dans le modèle 1 de la régression linéaire était de 0.3251, alors que dans la nouvelle régression linéaire prenant en compte le pc1, il est de 0.3368, nous observons donc une amélioration faible mais non nul de l'explication de la variance de la variable dépendante par rapport aux variables indépendantes dans le modèle. Donc dans ce nouveau modèle, 33,68 % de la variance de la variable dépendante est expliquée par les variables indépendantes, face à 32,51% dans le modèle initial.

Question 3

Les variables qualitatives

```
> table(perf_envi$`COUNTRY OF DOMICIL`)

 BD  BG  BM  DK  ES  FA  FN  FR  IM  IR  IT  LX  MA  NL  NW  OE  PO  PT
 63  16   1  23  22   1  16  69   1  11  27   7   1  28  16   7   7   3
 SD  SW  UK
 56  53 125
> table(perf_envi$`CSR Sustainability Committee`)

  N   Y
106 447
>
```

Commençons en premier avec les variables qualitatives, toutefois, afin de pouvoir manipuler ces variables, il est essentiel de savoir les modalités de ces variables ainsi que leurs fréquences. C'est pour cela nous avons débuter par la commande table () (cf. Script) qui nous permet de faire cela. Nous avons observé que les 2 variables qualitatives, ont les modalités suivantes :

A savoir, les pays sont les suivants :

BD : Bangladesh, BG : Bulgarie, BM : Bermudes, DK : Danemark, ES : Espagne, FA : Falkland Islands, FN : Finlande, FR : France, IM : Île de Man, IR : Irlande, IT : Italie, LX : Luxembourg, MA : Maroc, NL : Pays-Bas, NW : Norvège, OE : Oman, PO : Pologne, PT : Portugal, SD : Soudan, SW : Suède et UK : Royaume-Uni.[1]

Et les modalités pour CSR, Sustainability Committee sont tout simple : Y : Yes, N : No.

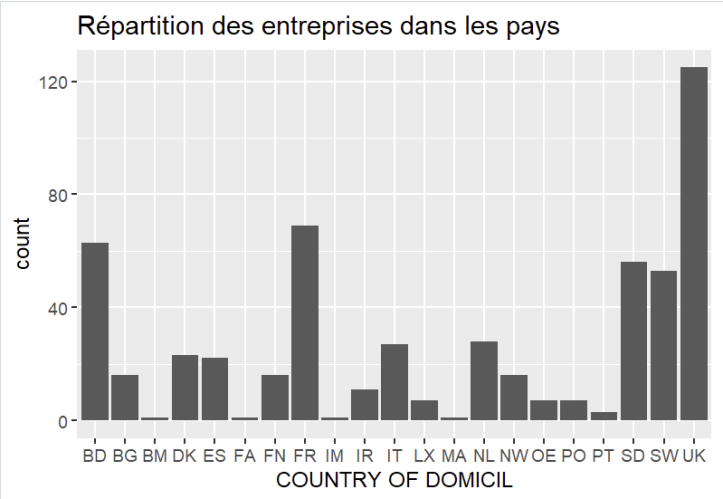
Il est également possible de visualiser les fréquences des 2 variables qualitatives, le Pays de l'entreprise et la présence de comité de durabilité, en pourcentage.

```
> prop.table(table(perf_envi$`COUNTRY OF DOMICIL`))

         BD          BG          BM          DK          ES          FA
0.113924051 0.028933092 0.001808318 0.041591320 0.039783002 0.001808318
         FN          FR          IM          IR          IT          LX
0.028933092 0.124773960 0.001808318 0.019891501 0.048824593 0.012658228
         MA          NL          NW          OE          PO          PT
0.001808318 0.050632911 0.028933092 0.012658228 0.012658228 0.005424955
         SD          SW          UK
0.101265823 0.095840868 0.226039783
> prop.table(table(perf_envi$`CSR Sustainability Committee`))

        N         Y
0.1916817 0.8083183
```

[1] https://www.sirius-upvm.net/doc/usuels/iso3166.html

C’est ainsi qu’il a été intéressant de visualiser les graphiques de ces variables, avec l’utilisation de plusieurs commande de graphiques du package « ggplot2 », (cf. Script)

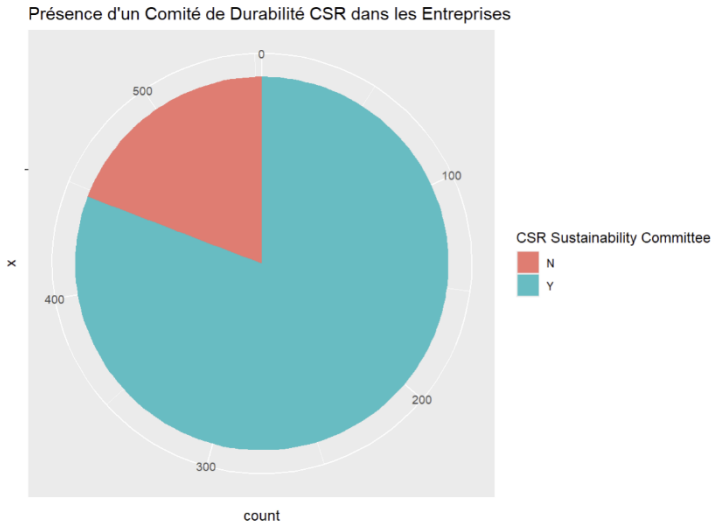
En premier, un histogramme montrant la répartition des entreprises selon les pays, puisque que ce graphique est basé sur la variable « Country of Domicile » :



Interprétation de l’histogramme :

Dans cet histogramme, il y a une répartition des nombres d’entreprises selon les pays, alors qu’il s’agit uniquement d’un graphique de fréquence, il est intéressant de mentionner que la plus grand nombre d’entreprise existe au Royaume-Uni, suivit par la France et des pays comme Burundi, Île de Man, Falkland Islands et le Maroc qui n’ont qu’une seule entreprise.

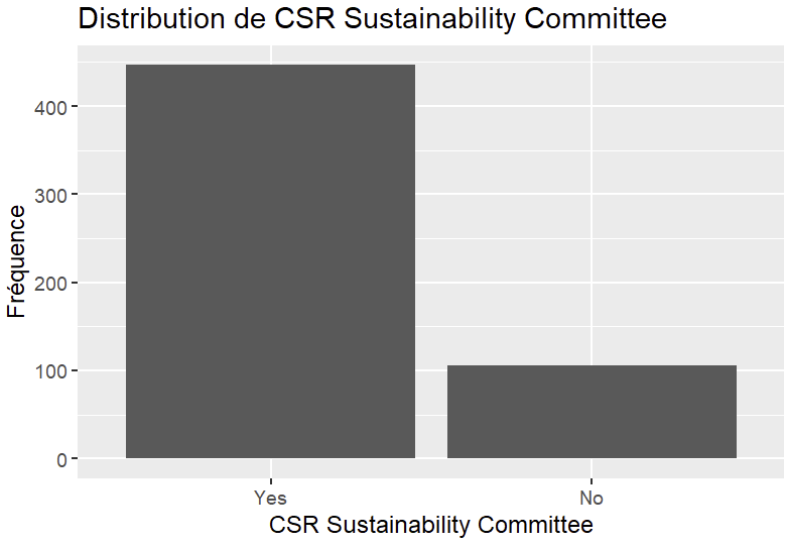
Le deuxième graphique, toujours fait avec la commande ggplot (cf.Script), est un camembert montrant le nombre des entreprises ayant un comité de durabilité, et ceux n’ayant pas ce comité.



Interprétation du camembert :

Ce camembert montre le nombre d’entreprises qui ont un comité de durabilité, et ceux qui n’en ont pas. Aux alentours de 450 entreprises ont un comité, et presque 105 n’ont pas de comité. Les NA n’ont pas été pris en compte.

Diagramme en barres de la variance de la variable Sustainability Committee



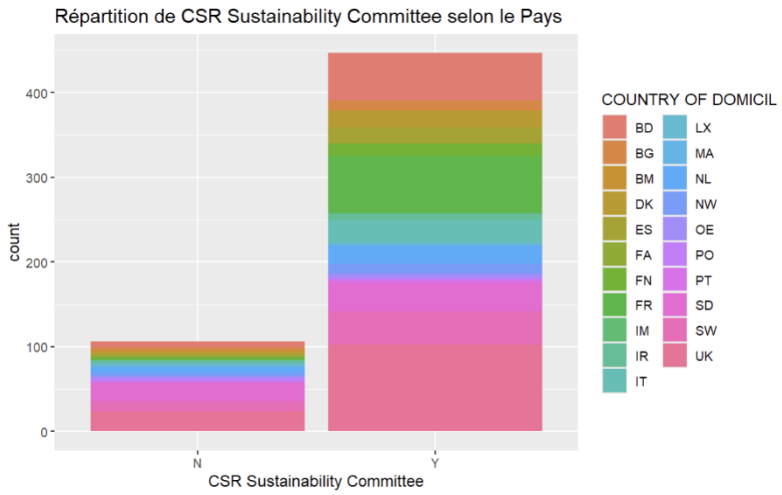
Interprétation digramme en barre :

Tout comme le graphique camembert juste avant, cet histogramme en barre sert a montré les fréquences de la présence d'une comité de durabilité dans les entreprises ou pas, aux alentours de 450 entreprises en possèdent, alors que presque 105 entreprises n'en ont pas. Il faut quand même noter que nous n'avons pas pris en compte les NA.

Pour avancer dans les analyses, notamment en arrivant à la régression linéaire dans la question 4, il est essentiel de transformer ces variables qualitatives en factors, avec la commande factors(), (cf.Script).

Après avoir transformer les variables qualitatives en factors pour pouvoir mieux les manipuler statistiquement, il est essentiel de commencer à croiser ces deux variables entre-elles, afin de voir comment les deux interagissent graphiquement.

Premièrement, il s'agit d'un histogramme à barre empilé, qui montre la répartition des comité de durabilité selon les pays.

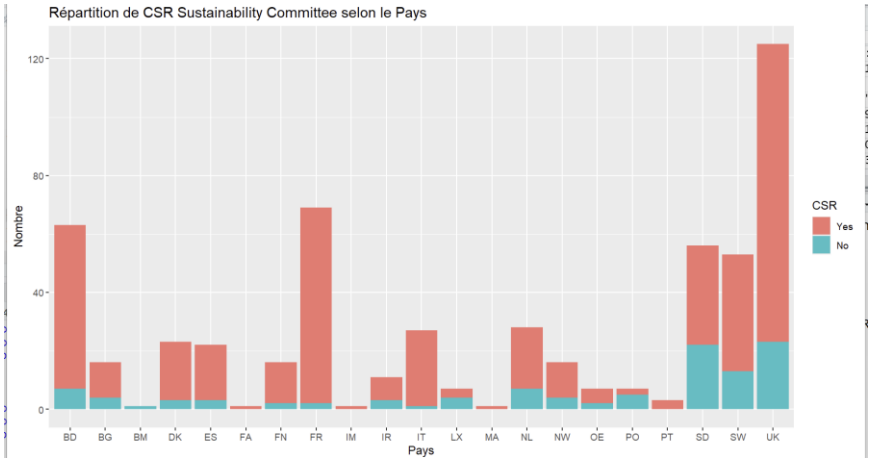


Interprétation de l'histogramme à barres empilés :

Cet histogramme montre la présence ou l'absence des comité de durabilité, mais également leur répartition selon le pays. Il s'agit d'un croisement des deux variables Comité et Pays. On aura ainsi, la Grande Bretagne avec le plus grand nombre d'entreprises ayant une comité de durabilité, suivit par la France puis le Bangladesh. Ces 3 pays possèdent également respectivement, 7, 4 et 2 entreprises n'ayant pas de CSR.

Ces graphiques montrent la répartition de la fréquence de la présence de comité de durabilité dans l'entreprise selon les pays, cela s'agit ainsi d'un croisement entre ces deux variables qualitatives, CSR et COD.[2]

Ce même croisement statistique indiquant une certaine relation entre ces deux variables qualitatives peut être visualiser par des tests statistiques, à savoir par des tests de corrélations ou des tableaux de contingence. Néanmoins, vu que les variables sont qualitatives, nous ne pouvons pas calculer une corrélation entre ces deux variables, mais il est possible de réaliser un tableau de contingence.



<u>Interprétation de l'histogramme à barres empliés</u>:

Il s'agit du même principe que le graphique juste avant, mais celui là est un peu plus clair, parce qu'il montre chaque pays avec le nombre d'entreprises ayant un CSR et ceux n'ayant pas un CSR dans ce même pays. La lecture de cet histogramme est plus simple, ayant les Pays comme X. Pour la France par exemple, elle possède un peu près 69 entreprises dont 2 qui n'ont pas de CSR.

De ce fait, dans ce cas de figure, nous avons calculé des tableaux de contingence et avons appliqué le test d'indépendance du Chi-deux

La commande table() génère un tableau de contingence (effectifs) pour nos deux variables qualitatives. Par exemple, table(CSR, COD) a pour but d'afficher la répartition des différentes combinaisons de chaque modalités de deux variables qualitatives, CSR et COD, en d'autres mots, c'est un tableau de répartition d'une variable qualitative par rapport à une autre. En stockant ce tableau dans une variable (par exemple CSR_COD=table(CSR,COD)), il est possible par la suite d'appliquer le prop.table(CSR_COD) pour obtenir des **fréquences relatives** de chaque combinaison, en rapportant chaque effectif au total des observations. Cela donne les pourcentages ou proportions de chaque catégorie par rapport à l'ensemble. De plus, avec la fonction round(), il est possible d'arrondir les valeurs pour qu'elles soient plus lisibles.

```
     COD
CSR      BD   BG   BM   DK   ES   FA   FN   FR   IM   IR   IT   LX   MA
  Yes 0.89 0.75 0.00 0.87 0.86 1.00 0.88 0.97 1.00 0.73 0.96 0.43 1.00
  No  0.11 0.25 1.00 0.13 0.14 0.00 0.12 0.03 0.00 0.27 0.04 0.57 0.00
     COD
CSR      NL   NW   OE   PO   PT   SD   SW   UK
  Yes 0.75 0.75 0.71 0.29 1.00 0.60 0.75 0.82
  No  0.25 0.25 0.29 0.71 0.00 0.40 0.25 0.18
```

Ce tableau ne montre que des fréquences de répartition, ce qui nous pousse à nous demander s'il existe une relation entre ces deux variables.

Il serait intéressant ici de tester s'il existe une relation significative entre ces 2 variables qualitatives, puisque ces deux variables seront par la suite utilisées dans notre régression

[2] CSR: Sustainability committee et COD: Country of Domicile (NB, ces deux variables son ten Factors)

linéaire notamment en tant que variables explicatives. D'où le tableau de contingence entre les 2 variables CSD Sustainability Committee et COD Country of Domicil, toutefois, le but n'est pas de s'arrêter là, parce qu'après la perception d'une certaine relation dans les tableaux de contingence, il est primordial de mesurer l'intensité d'une telle relation, ainsi que les valeurs qui peuvent être potentiellement problématiques.

Pour cela, il existe 2 méthodes, le test du Khi2 qui permettra de tester l'indépendance entre deux variables qualitatives. La H0 (hypothèse nulle) stipule qu'il n'y a pas de relation (les variables sont indépendantes), et la H1 (hypothèse alternative) stipule qu'il y a une relation (les variables sont dépendantes).

```
> chisq.test(CSR_COD)

        Pearson's Chi-squared test

data:  CSR_COD
X-squared = 63.882, df = 20, p-value = 1.756e-06
```

Dans ces résultats du test du Khi2, la P-value est inférieur au seuil de confiance de 5%, à savoir, 1.756^{e}-06<0,05, signifiant que H0 est à rejeter, donc l'hypothèse de l'indépendance des deux variables est à rejeter, ce qui fait qu'il existe une relation entre CSR et COD.

De plus, afin de mesurer l'intensité globale de cette relation, on utilise le V de Cramer :

```
> assocstats(CSR_COD)
                    X^2 df   P(> X^2)
Likelihood Ratio 63.770 20 1.8296e-06
Pearson          63.882 20 1.7565e-06

Phi-Coefficient   : NA
Contingency Coeff.: 0.322
Cramer's V        : 0.34
```

VALEUR DU V DE CRAMER	INTENSITÉ DE LA RELATION ENTRE LES VARIABLES
< 0,10	RELATION NULLE OU TRÈS FAIBLE
>= 0,10 ET <0,20	RELATION FAIBLE
>= 0,20 ET <0,30	RELATION MOYENNE
>= 0,30	RELATION FORTE

L'analyse du V de cramer nous permet de savoir que l'intensité de la relation entre ces deux variables qualitatives explicatives est forte, à savoir 0,34>0,30, ce qui peut poser un problème de multi-colinéarité en effectuant la régression linéaire. Pour s'assurer de la présence ou pas de la multi-colinéarité, le test de VIF de Variance a été réalisé, les résultats pour les deux variables CSR et COD étaient plus petit que 5, ce qui fait qu'il n'y a pas multi-colinéarité (une multi-colinéarité existe si le 5<VIF<10).

```
        GVIF Df GVIF^(1/(2*Df))
CSR 1.127146  1        1.061671
COD 1.127146 20        1.002997
>
```

Il est important de noter, toutefois, que la relation entre ces variables qualitatives explicatives et la variable expliquée EPS sera abordée par le test ANOVA et les graphiques correspondant, dans une partie ultérieure du rapport.

Les variables quantitatives

On commence par l'analyse des données de summary

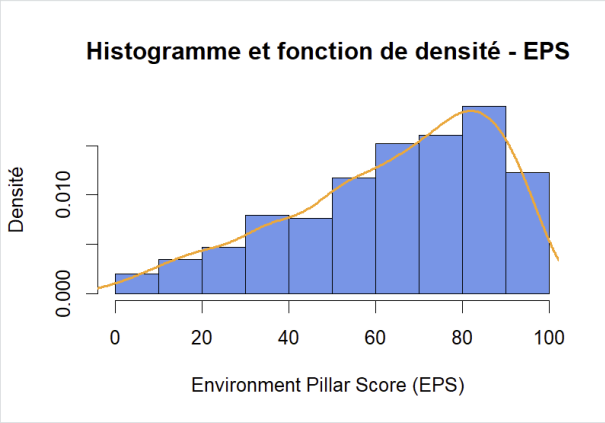
```
                                                (Other):159
      EPS               EMP              MARGIN             SALES
 Min.   : 0.00   Min.   :     0   Min.   :-275.78   Min.   :    69783
 1st Qu.:49.20   1st Qu.:  4285   1st Qu.:   7.47   1st Qu.:  2541000
 Median :69.01   Median : 14687   Median :  12.93   Median :  7353900
 Mean   :64.03   Mean   : 41599   Mean   :  20.67   Mean   : 23869937
 3rd Qu.:83.00   3rd Qu.: 45980   3rd Qu.:  23.05   3rd Qu.: 24081000
 Max.   :98.16   Max.   :671205   Max.   : 731.44   Max.   :554184257

      MCAP                RCAP             BOARD
 Min.   :1.480e+09   Min.   :-38.14   Min.   :  0.00
 1st Qu.:5.554e+09   1st Qu.:  4.61   1st Qu.: 50.00
 Median :1.357e+10   Median :  8.24   Median : 65.00
 Mean   :4.132e+10   Mean   : 11.02   Mean   : 64.11
 3rd Qu.:3.672e+10   3rd Qu.: 13.77   3rd Qu.: 80.00
 Max.   :1.322e+12   Max.   :437.74   Max.   :100.00
```

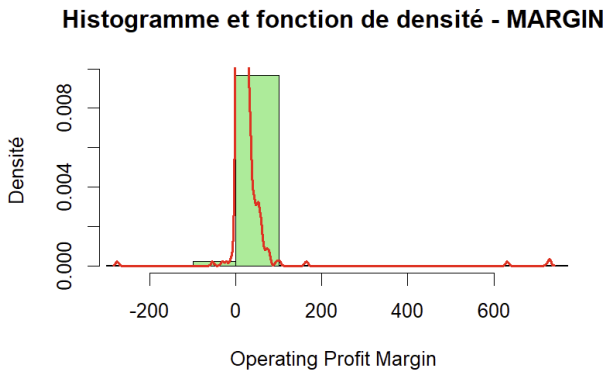
L'analyse de telle données est assez simple, c'est pour cela nous allons détailler uniquement une seule variable, EPS, qui servira de référence pour les autres, premièrement, le score environnementale a un minimum de 0.00 et un score maximal de 98.16. L'EPS a également une médiane de 69.01, signifiant que 50% des entreprises ont un score supérieur ou inférieur à cette valeur. La moyenne de EPS est de 64.03, ce qui est attirant est que cette valeur est proche de la médiane ce qui peut signifier que la distribution des scores se focalise autour de cette valeur de la moyenne. Le Premier quartile est de 49.20, montrant que 25% des scores sont inférieur ou égales à cette valeur, et le troisième quartile de 83.00 montre que 75% des entreprises ont un score inférieur ou égale à 83 et les 25% restant sont supérieur.

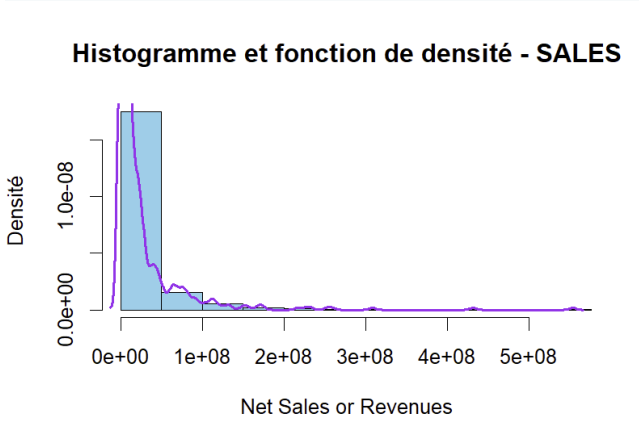
Quant aux variables quantitative il est primordial de comprendre si ces dernières adhèrent ou pas à une distribution normale, comme cela sera une condition importante pour la régression linéaire qui sera introduite dans la question 4.

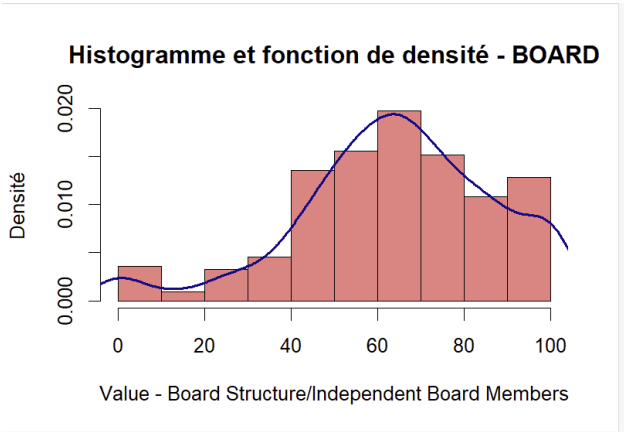
L'histogramme en fonction de la densité est une approche visuelle pour évaluer la normalité d'une variable. Le graphique QQ-plot de même, ainsi que les tests statistiques de Jarque-Bera et de Kolmogorov-Smirnov. C'est pour cela, nous allons utiliser les 4 méthodes sur les différentes variables pour savoir si les variables explicatives suivent une distribution normale ou pas.

Pour l'histogramme et sa fonction de densité, ce graphique permet de voir la répartition des données et de vérifier si elle suit une forme en cloche, typique d'une distribution normale.







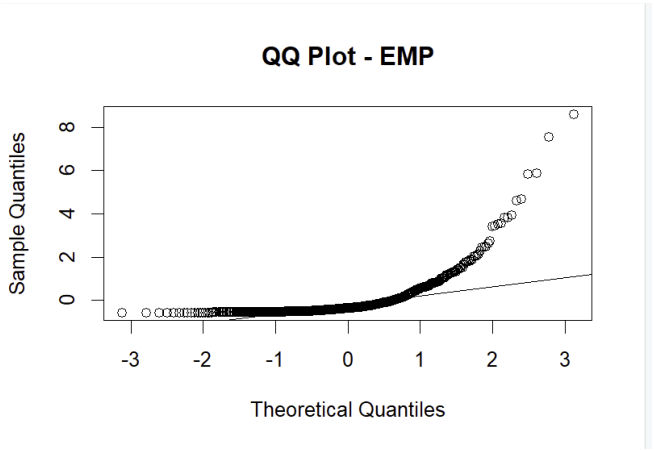


Pour l'histogramme de l'EPS, nous pouvons détecter une courbe de densité sous forme de cloche, signifiant que la variable suit généralement une loi normale, toutefois, vu que cette courbe n'est pas totalement symétrique, cela signifie qu'il existe des modalités extrêmes qui n'adhèrent pas à la loi normale notamment vers les extrémités des modalités.

Pour l'histogramme de MARGIN, la courbe de densité est en forme de cloche est symétrique à un certains degrés, signifiant que cette variable suit la loi normale avec quelques valeurs extrêmes potentielles.

Pour l'histogramme des SALES, il est en forme de courbe mais pas symétrique du tout, ce qui fait qu'elle adhère à une loi normale mais avec un nombre non-négligeable de valeurs extrêmes qui peuvent influencer non seulement la distribution mais également la régression par la suite.

Finalement, pour l'histogramme du BOARD, il s'agit d'une courbe en cloche, pas totalement symétrique, mais une tendance de symétrie peut être détecter ce qui fait que cette variable suit la loi normale avec quelques variables qui s'écartent de la normalité.



Pour la variable Employés, afin d'essayer d'autres graphique, nous avons opté pour un QQ-plot, un graphique permettant de visualiser la distribution de la variable EMP et de vérifier sa normalité. Nous pouvons voir dans ce graphique une ligne droite, une partie des modalités la suivent mais d'autres s'écartent de cette ligne de normalité, donc cette variable ne suit pas totalement la loi normale, elle l'a suit que partiellement avec des valeurs extrêmes qui s'écartent de la normalité.

Pour les tests statistiques, ces derniers permettent de vérifier si la variable suit une loi normale de manière formelle. Il existe deux tests principaux, celui de Jarque-Bera et celui de Kolmogorov-Smirnov.

```
> library(normtest)
> jb.norm.test(perf_envi$RCAP)

        Jarque-Bera test for normality

data:  perf_envi$RCAP
JB = 2291077, p-value < 2.2e-16
```

Pour la variable RCAP, le résultat du test de Jarque-Bera est une p-value de $<2.2^{e}$-16<0.05 le seuil de confiance, ce qui fait que l'hypothèse nulle H0 sera rejetée, cette dernière était que les données suivent une distribution normale, une fois cette hypothèse rejetée, H1 est acceptée donc les données ne suivent pas une distribution normale.

Pour la variable MCAP, le résultat du test de Kolmogorov-Smirnov qui compare une distribution observée à une distribution théorique, était une p-value de $<2.2^{e}$-16 ce qui est à son tour<0.05, donc H0 rejetée donc les données ne suivent pas une loi normale.

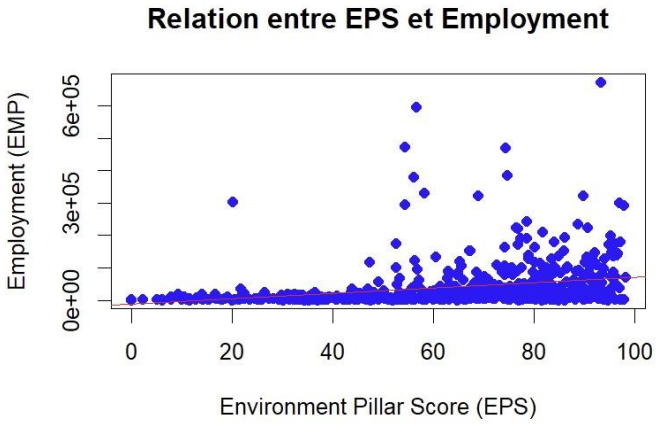
```
> # Test de Kolmogorov-Smirnov
> ks.test(perf_envi$MCAP, "pnorm", mean = 1, sd = 1)

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  perf_envi$MCAP
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Croisement des variables explicatives avec la variable expliquée : Une vision avant la régression

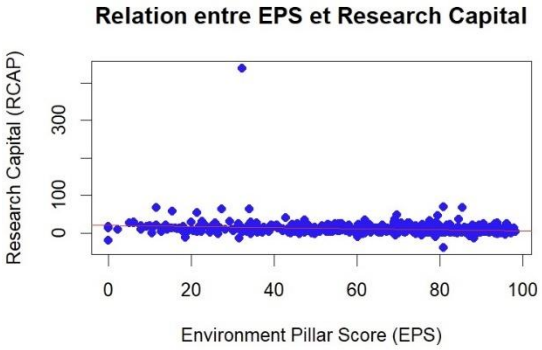
Commençons par des éléments simples, les graphiques plots, ces graphiques te permettront de visualiser la relation entre EPS et chaque variable explicative quantitative. Ainsi que la matrice de corrélation (cf. script). Pour pouvoir interpréter la matrice de données il est important de savoir qu'un coefficient de corrélation proche de 1 indique une forte corrélation positive, celui proche de -1 indique une forte corrélation négative et celui proche de 0 signifie qu'il n'y a pas de corrélation linéaire significative entre les variables.

	EPS	EMP	MARGIN	SALES	MCAP
EPS	1.00000000	0.271039371	-0.15689682	0.26361671	0.129453349
EMP	0.27103937	1.000000000	-0.10309195	0.35734273	0.089419784
MARGIN	-0.15689682	-0.103091946	1.00000000	-0.07232533	0.059099831
SALES	0.26361671	0.357342729	-0.07232533	1.00000000	0.606741558
MCAP	0.12945335	0.089419784	0.05909983	0.60674156	1.000000000
RCAP	-0.15620897	-0.091339577	0.13263904	-0.07355857	0.077508951
BOARD	0.07383135	-0.002675382	-0.01024218	0.02830672	0.006452966

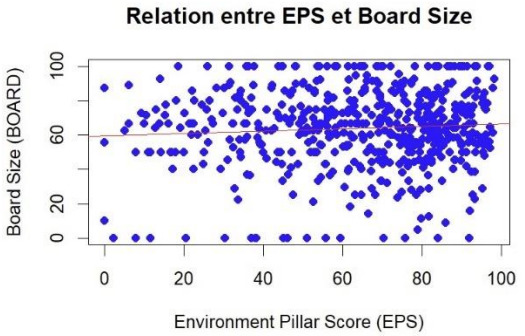
	RCAP	BOARD
EPS	-0.1562089733	0.0738313540
EMP	-0.0913395774	-0.0026753819
MARGIN	0.1326390370	-0.0102421800
SALES	-0.0735585652	0.0283067209
MCAP	0.0775089514	0.0064529656
RCAP	1.0000000000	-0.0001349917
BOARD	-0.0001349917	1.0000000000



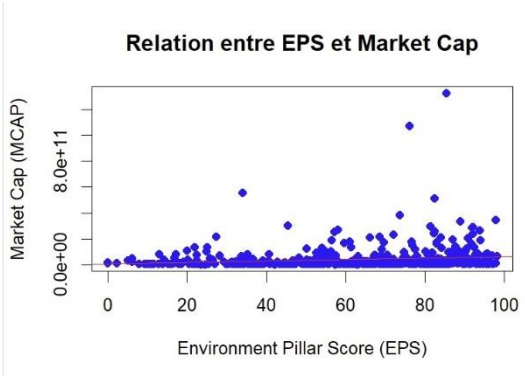
D'après ce graphique plot et la matrice de corrélation (0.2 proche de 0), nous pouvons déduire qu'il n'existe pas une correlation directe entre l'EPS et l'employment puisque dans le graphique toutes les valeurs sont dans le bas du graphique, malgré la variation dans les valeurs d'EPS.



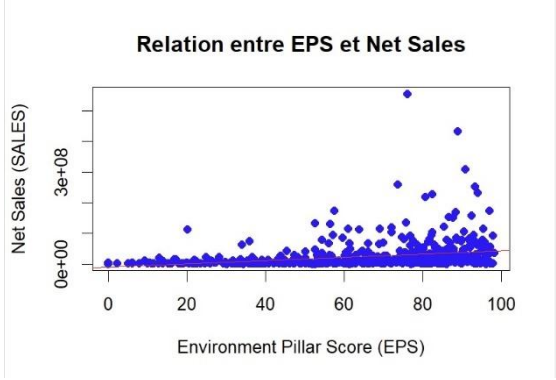
Ce scatter plot montre la relation entre l'EPS et le RCAP. La plupart des points de données se regroupent en bas mais cela indique aucune correlation entre ces deux variables. D'après la matrice de corrélation, la valeur -0.15 est proche de 0 donc il n'y a pas de corrélation avec EPS



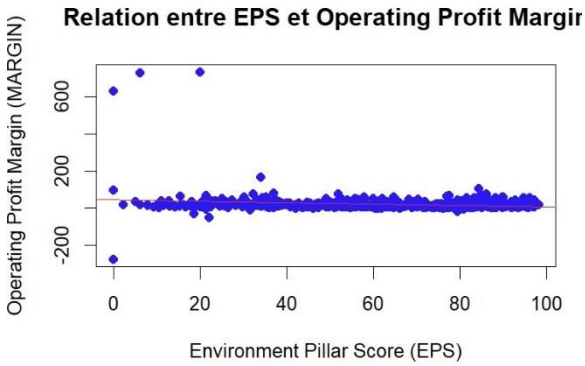
EPS et Board size: dans ce scatter plot, les valeurs sont étalées tout au long du plot, montrant aucune tendance entre les deux variables, et le coefficient de corrélation est de 0.07 proche de 0



EPS et market cap: un EPS plus élevé n'indique pas dans ce graphique une tendance croissante du MCAP, de plus, le coefficient de corrélation est de 0.12 proche de 0, donc il n'y a pas de corrélation.



Pour EPS et NET sales, à mesure que l'EPS augment, il y a une faible augmentation dans le Net Sales. Toutefois, elle non significative. Avec un coefficient de 0.2 proche de 0, la corrélation n'existe pas.



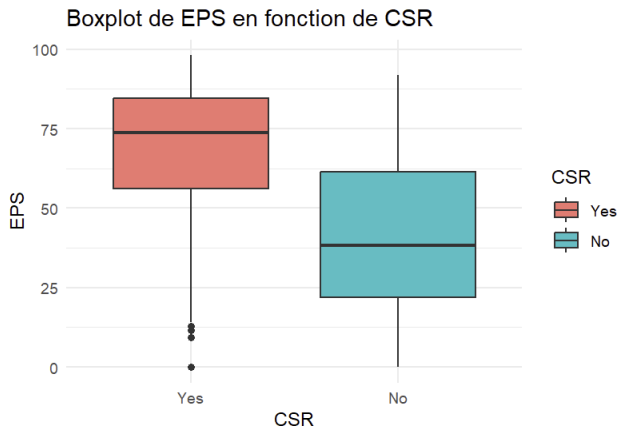
Pour EPS et Operating Profit, similairement à l'operating profit margin, il n'y a pas une claire corrélation entre l'EPS et cette variable. Le coefficient de corrélation est de -0.15 proche de 0 donc il n'y a pas de corrélation.

Pour les variables qualitatives et leurs croisement avec l'EPS, nous ne pouvons pas faire un test de corrélation, néanmoins, avec le test ANOVA, nous pouvons tester si les moyennes d'une variable continue (EPS) diffèrent en fonction des catégories d'une variable qualitative (comme CSR ou COD)

```
> summary(anova_result)
             Df Sum Sq Mean Sq F value Pr(>F)
CSR           1  67103   67103   156.4 <2e-16 ***
Residuals   552 236823     429
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
46 observations deleted due to missingness
> summary(anova_result2)
             Df Sum Sq Mean Sq F value   Pr(>F)
COD          20  38315  1915.7   3.871 4.88e-08 ***
Residuals   532 263304   494.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

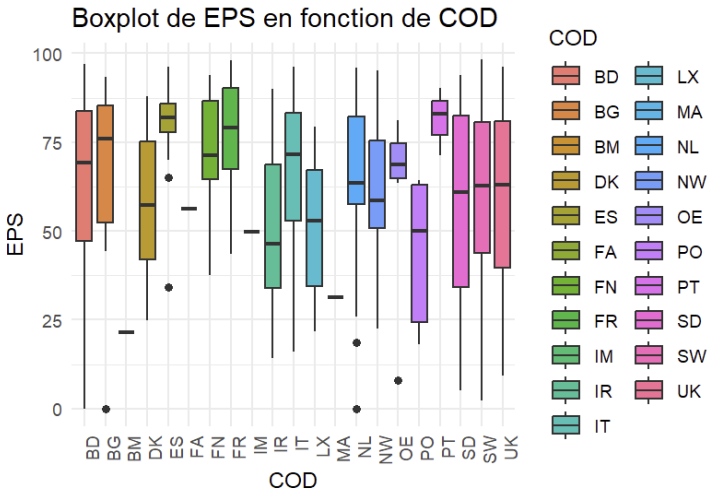
Les résultats sont des P-value faibles inférieures au seuil de confiance de 0.05 pour les deux variables indiquant que **les groupes** de la variable qualitative influencent significativement l'EPS, mais cela ne signifie pas qu'il y a une relation de **corrélation**.

Quelques graphiques montrant la relation :



Interprétation Boîte moustache:

Ce diagramme en boîte permet d'évaluer l'EPS en fonction de la présence ou de l'absence d'une comité de durabilité. Il est possible de détecter que la médiane des EPS est plus élevée en présence de la CSR, donc la présence de ce comité donne un meilleur score environnemental. De plus, l'intervalle interquartile du diagramme de la modalité oui est plus compact que celui de la modalité non, indiquant plus de cohérence en présence de la CSR.



Interprétation Boîte moustache:

Cette boîte moustache montre que chaque pays diffère en termes d'EPS. De plus, la médiane varie de manière considérable entre les pays figurant. Par exemple, dans des pays comme l'Espagne, les valeurs de la médiane sont élevées, contrairement à la Pologne. Concernant les points noirs, ce sont des valeurs aberrantes, bien plus faibles que la distribution principale, comme le cas Espagne et aux Pays-Bas.

Pour mieux comprendre cette relation entre les variables explicatives et celle expliquée, il est primordial d'établir une régression linéaire.

Question 4

L'équation introduite par la question 4, est celle d'une régression linéaire multiple, nous pouvons estimer cette dernière et l'appliquer sur R, à travers la fonction lm, en premier nous allons inclure toutes les variables explicatives dans le modèle suivant :

mod<-lm(EPS~BOARD+COD+CSR+EMP+MARGIN+MCAP+RCAP+SALES , data=perf_envi)

summary(mod)

Il y aura les résultats suivant :

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.957e+01  3.593e+00  16.577  < 2e-16 ***
BOARD        8.639e-02  3.919e-02   2.204 0.027925 *
CODBG        9.804e+00  5.493e+00   1.785 0.074843 .
CODBM       -2.368e+01  1.948e+01  -1.215 0.224742
CODDK       -1.022e+01  4.897e+00  -2.088 0.037324 *
CODES        1.636e+01  4.776e+00   3.424 0.000665 ***
CODFA       -1.027e+01  1.939e+01  -0.530 0.596543
CODFN        6.567e+00  5.497e+00   1.195 0.232734
CODFR        1.101e+01  3.398e+00   3.240 0.001271 **
CODIM       -1.684e+01  1.937e+01  -0.869 0.385122
CODIR       -1.195e+01  6.313e+00  -1.894 0.058836 .
CODIT        2.735e+00  4.452e+00   0.614 0.539236
CODLX       -3.063e+00  7.728e+00  -0.396 0.692059
CODMA       -3.558e+01  1.940e+01  -1.834 0.067251 .
CODNL        2.198e-01  4.474e+00   0.049 0.960838
CODNW       -7.389e+00  5.679e+00  -1.301 0.193824
CODOE        7.992e-01  7.721e+00   0.104 0.917598
CODPO       -8.241e+00  7.791e+00  -1.058 0.290633
CODPT        1.615e+01  1.138e+01   1.419 0.156495
CODSD       -5.101e+00  3.806e+00  -1.340 0.180699
CODSW       -6.459e-01  3.619e+00  -0.178 0.858422
CODUK       -2.487e+00  3.004e+00  -0.828 0.408182
CSRNo       -2.059e+01  2.318e+00  -8.881  < 2e-16 ***
EMP          2.333e-05  1.296e-05   1.800 0.072438 .
MARGIN      -3.183e-02  1.586e-02  -2.007 0.045244 *
MCAP         2.357e-11  1.191e-11   1.979 0.048369 *
RCAP        -5.360e-02  4.090e-02  -1.310 0.190656
SALES        7.353e-08  2.455e-08   2.996 0.002869 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.2 on 525 degrees of freedom
Multiple R-squared:  0.3581,    Adjusted R-squared:  0.3251
F-statistic: 10.85 on 27 and 525 DF,  p-value: < 2.2e-16
```

Dans un tel summary, il y a plusieurs éléments à detecter, en premier, il est essentiel de voir la significativité du modèle, ainsi que le R2. Le la stat de Ficher, avoir une P-value de <2.2^{e}-16<0.05 signifie que le modèle est globalement significatif. Toutefois, le R2 de 0.3581 proche de 0 donc faible signifie que le modèle n'explique que 35.8% de la variance de l'EPS.

Concernant les variables explicatives il est intéressant de savoir s'ils sont significatifs ou pas, à savoir, un Pr(<|t|) inférieur à 0.05 le seuil de confiance, signifie que la variable est significative et influence l'EPS. Il est important de préciser que pour les variables qualitatives en factor, leur significativité se fait par rapport à la variable de référence, qui est celle manquante du tableau, en cas de CSR c'est le Yes et en cas de COD c'est le BD, donc Bangladesh. Pour savoir si une variable qualitative est significative dans sa globalité sur le modèle nous devons utiliser la fonction drop1, qui teste cela, ça sera une étape ultérieure.

Revenons à l'interprétation ; les variables quantitatives explicatives significative sont ; BOARD, MARGIN, MCAP, SALES puisqu'elles ont toutes des p-value inférieures à 0.05. Ce qui signifie que pour chaque augmentation d'une unité de la variable BOARD, l'EPS augmente de 8.639^{e}-02 et pour chaque augmentation d'une unité de la variable MARGIN, l'EPS diminue de -3.183^{e}-2. Pour chaque augmentation d'une unité de MCAP, l'EPS augmente de 2.357^{e}-11 et pour chaque augmentation d'une unité de SALES, l'EPS augmente de 7.353^{e}-8. Malgré l'existence de ces variations, il s'agit de variation très faibles.

Pour les variables qualitative nous allons uniquement interpréter une significative puisqu'il ne s'agit pas d'une significativité par rapport à l'EPS mais par rapport à la variable de référence. Donc, pour une absence de CSR, le coefficient est de -2.059^{e}-11 avec une P-value faible <0.05, donc cette modalité a un effet plus faible sur l'EPS qu'une présence de CSR. Et pour le COD en France, avec une P-value faible et un coefficient de 1.01^{e}+01, donc cette modalité a un effet plus élevé que Bangladesh sur l'EPS.

Pour savoir si une variable qualitative est significative dans sa globalité sur le modèle nous devons utiliser la fonction drop1

Tester la significativité de la CSR

```
> drop1(mod_quali1, .~., test = "F")
Single term deletions

Model:
EPS ~ CSR
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              236167 3353.5
CSR     1     65452 301619 3486.8  152.71 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Tester la significativité de la COD

```
> drop1(mod_quali2,.~.,test="F")
Single term deletions

Model:
EPS ~ COD
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              263304 3451.6
COD    20     38315 301619 3486.8  3.8707 4.876e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Les deux variables sont significatives, car la P-value est très faible<0.05 faisant que H0 de non significativité a été rejeté.

Modèle de Régression 2 :

Puisque il y a eu plusieurs variables non significatives ayant un P-Value>0.05, nous avons décider de refaire le modèle de régression avec que les variables significatives du premier modèle, nous avons donc enlever les variables RCAP et EMP, et cela peut être expliqué par la matrice de corrélation déjà évoqué qui a montré qu'il n'existe aucune corrélation entre ces variables et EPS (déjà expliqué), de même ces variables ne suivent pas totalement une loi de distribution normale comme vu avant (Analyse Quantitatif).

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.021e+01  3.535e+00  17.032  < 2e-16 ***
BOARD        8.793e-02  3.929e-02   2.238 0.025622 *
CODBG        9.398e+00  5.489e+00   1.712 0.087500 .
CODBM       -2.379e+01  1.952e+01  -1.219 0.223478
CODDK       -1.120e+01  4.891e+00  -2.289 0.022470 *
CODES        1.633e+01  4.783e+00   3.415 0.000687 ***
CODFA       -1.144e+01  1.944e+01  -0.589 0.556302
CODFN        5.469e+00  5.487e+00   0.997 0.319373
CODFR        1.164e+01  3.394e+00   3.430 0.000651 ***
CODIM       -1.699e+01  1.942e+01  -0.875 0.382086
CODIR       -1.249e+01  6.318e+00  -1.977 0.048600 *
CODIT        1.959e+00  4.442e+00   0.441 0.659449
CODLX       -2.713e+00  7.742e+00  -0.350 0.726169
CODMA       -3.689e+01  1.945e+01  -1.897 0.058416 .
CODNL        6.627e-02  4.477e+00   0.015 0.988194
CODNW       -9.044e+00  5.592e+00  -1.617 0.106404
CODOE        5.451e-01  7.731e+00   0.071 0.943815
CODPO       -8.884e+00  7.796e+00  -1.140 0.254979
CODPT        1.584e+01  1.141e+01   1.388 0.165626
CODSD       -6.241e+00  3.761e+00  -1.660 0.097596 .
CODSW       -1.249e+00  3.615e+00  -0.346 0.729759
CODUK       -3.281e+00  2.992e+00  -1.097 0.273217
CSRNo       -2.134e+01  2.294e+00  -9.301  < 2e-16 ***
MARGIN      -3.418e-02  1.584e-02  -2.158 0.031388 *
MCAP         1.963e-11  1.181e-11   1.663 0.096998 .
SALES        9.343e-08  2.266e-08   4.123 4.35e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.26 on 527 degrees of freedom
Multiple R-squared:  0.352,     Adjusted R-squared:  0.3212
F-statistic: 11.45 on 25 and 527 DF,  p-value: < 2.2e-16
```

Dans ce nouveau modèle, l'interprétation sera la même d'où nous n'allons pas l'évoquer de nouveau, mais la P-value de MCAP a augmenté et elle est maintenant supérieur à 5%, avec un R2 très faible montrant que le modèle n'explique pas vraiment la variance d'EPS, ce qui fait qu'un troisième modèle sera présenté, sans MCAP

Modèle 3 de régression

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.002e+01  3.539e+00  16.959  < 2e-16 ***
BOARD        8.850e-02  3.935e-02   2.249 0.024921 *
CODBG        9.477e+00  5.498e+00   1.724 0.085355 .
CODBM       -2.377e+01  1.955e+01  -1.216 0.224653
CODDK       -9.073e+00  4.729e+00  -1.919 0.055573 .
CODES        1.646e+01  4.790e+00   3.437 0.000635 ***
CODFA       -1.078e+01  1.947e+01  -0.554 0.579946
CODFN        5.507e+00  5.496e+00   1.002 0.316824
CODFR        1.192e+01  3.395e+00   3.510 0.000487 ***
CODIM       -1.678e+01  1.945e+01  -0.863 0.388759
CODIR       -1.232e+01  6.328e+00  -1.947 0.052088 .
CODIT        2.038e+00  4.449e+00   0.458 0.647058
CODLX       -2.667e+00  7.755e+00  -0.344 0.731022
CODMA       -3.633e+01  1.948e+01  -1.866 0.062665 .
CODNL        3.556e-01  4.481e+00   0.079 0.936783
CODNW       -7.511e+00  5.525e+00  -1.360 0.174540
CODOE        5.365e-01  7.744e+00   0.069 0.944791
CODPO       -9.004e+00  7.809e+00  -1.153 0.249393
CODPT        1.589e+01  1.143e+01   1.390 0.165173
CODSD       -5.000e+00  3.692e+00  -1.354 0.176232
CODSW       -8.608e-01  3.613e+00  -0.238 0.811790
CODUK       -3.180e+00  2.996e+00  -1.061 0.289013
CSRNo       -2.118e+01  2.296e+00  -9.223  < 2e-16 ***
MARGIN      -3.192e-02  1.581e-02  -2.020 0.043917 *
SALES        1.143e-07  1.888e-08   6.055 2.67e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.29 on 528 degrees of freedom
Multiple R-squared:  0.3486,    Adjusted R-squared:  0.319
```

Ce troisième modèle de régression est toujours faible quant à l'explication de la variance, avec un R2 faible de 0.319, toutefois, toutes les variables sont significatives, à savoir avec l'augmentation d'une unité de BOARD, et SALES, l'EPS augmente respectivement de 8.8^{e}-2 et de 1.14^{e}-07 et avec l'augmentation d'une unité de MARGIN, l'EPS diminue de -3.19^{e}-02. Toutefois ces variations sont toujours très marginales voire presque inexistantes.

En conclusion, avec un R2 faible et des coefficient de variations très bas, ce modèle n'est pas vraiment adéquat à l'explication de la variation de l'EPS, ni les variables. En effet, dans l'études des variables qualitatives et quantitatives, plusieurs violations des conditions de la régression linéaire ont eu lieu, notamment la non normalité des variables et l'absence de corrélation avec la variable expliquée, ce qui peut poser des questions quant à la pertinence de ce modèle. Il est peut être envisageable de considérer les variables en tant que non paramétriques et les traiter selon cette conditions. Toutefois, il ne s'agit que d'une hypothèse provisoire à tester.