Projet R & Dataviz

Analyse de la performance environnementale

Partie 2 : Analyse en composantes principales

Réalisé par:

CHIAB MARIEM

EL ABSODI SALMA

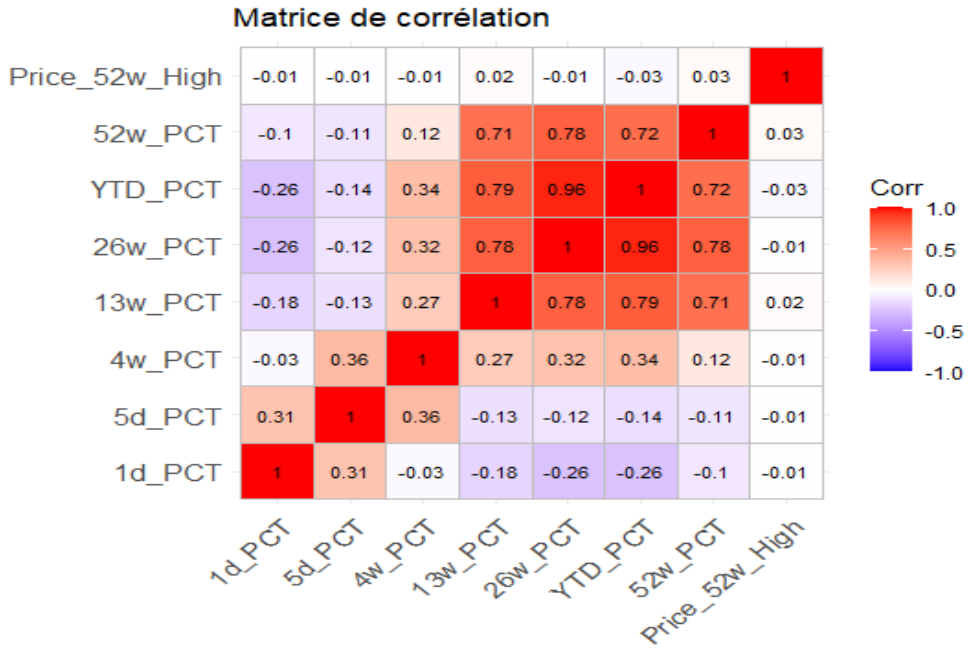
Master: M1 EEM parcours BIDABI

Année universitaire: 2024/2025

Analyse en composantes principales

Justification de l'utilisation de l'ACP :

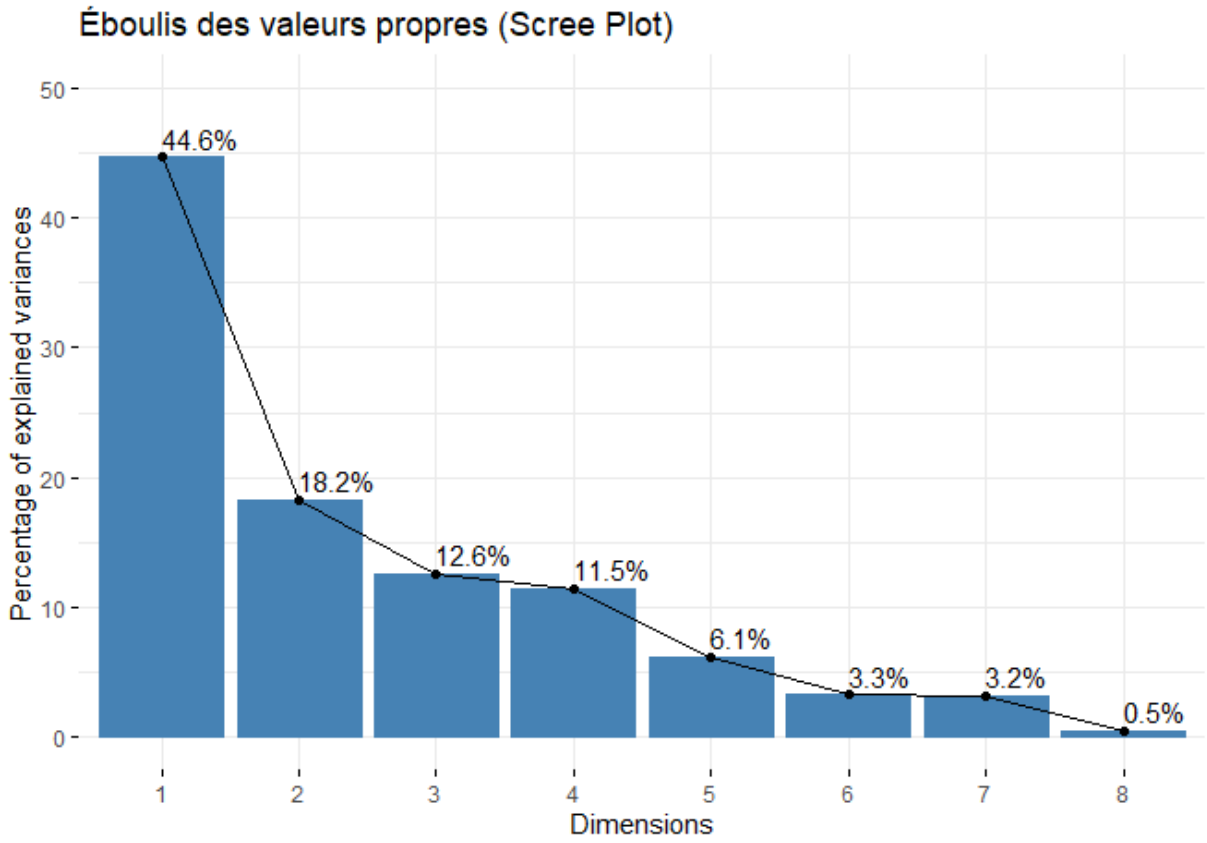
L'analyse en composantes principales est bien évidemment une technique statistique forte du fait qu'elle permet de garder l'essentiel des données malgré la réduction de leur dimensionnalité. L'utilisation de cette technique se justifie par plusieurs raisons. En premier lieu, L'ACP fait un éclairage sur les relations linéaires qui existent entre les variables et qui peuvent être corrélés en les partageant dans des groupes dont chaque groupe comprend les variables qui évoluent d'une façon identique. Dans notre cas, les variables sont de nature financière en représentant les évolutions de prix pendant des moments différents d'une entreprise ce qui permet de supposer qu'il existe une forte corrélation entre ces variables. D'où, on a calculé la matrice de corrélation. Cette étape est si nécessaire de la faire avant l'ACP pour savoir si les variables sont parfaitement corrélées dans le but de savoir si l'ACP pourrait être faite ou pas du fait qu'elle se base sur ce principe.



En se basant sur le graphique, il parait clairement qu'il y une certaine redondance qui caractérise les variables en étant fortement corrélés ce qui justifie bien évidemment l'utilisation de l'ACP afin de mettre en place des composantes principales qui permettent de capturer l'essentiel de la variance en baissant la dimensionalité de ces variables. A titre d'exemple, la variable 52w_PCT se caractérise par une corrélation forte avec YTD_PCT,26w_PCT et 13w_PCT en ayant respectivement une valeur de (0.78), (0.72) et (0.71). Ainsi que la variable YTD_PCT qui est fortement corrélée avec 26w_PCT, 13w_PCT et 52w_PCT en ayant une valeur de 0.78, 0.76 et 0.78. Comme, il y'a certaines variables qui se caractérises par des corrélations faibles ou modérées en gardant leur significativité dans l'ensemble. A titre d'exemple, la variable 1d_PCT et 5d_PCT qui se caractérisent par une corrélation modérée d'une valeur de 0.31.

Face à cette corrélation, l'ACP vise à faire face à la multi colinéarité en générant des composantes principales indépendantes en résumant les huit variables tout en préservant l'essentiel de l'information, ce qui permet également de faciliter l'interprétation et la visualisation des résultats.

Représentations graphiques des valeurs propores :



Valeurs propres et pourcentage d'inertie expliqué :

```
> print(acp$eig)
       eigenvalue percentage of variance cumulative percentage of variance
comp 1 3.57176195              44.6470244                           44.64702
comp 2 1.45772467              18.2215584                           62.86858
comp 3 1.00488317              12.5610396                           75.42962
comp 4 0.91983911              11.4979889                           86.92761
comp 5 0.49186568               6.1483211                           93.07593
comp 6 0.26244168               3.2805210                           96.35645
comp 7 0.25302016               3.1627521                           99.51921
comp 8 0.03846356               0.4807945                          100.00000
```

Commentaire :

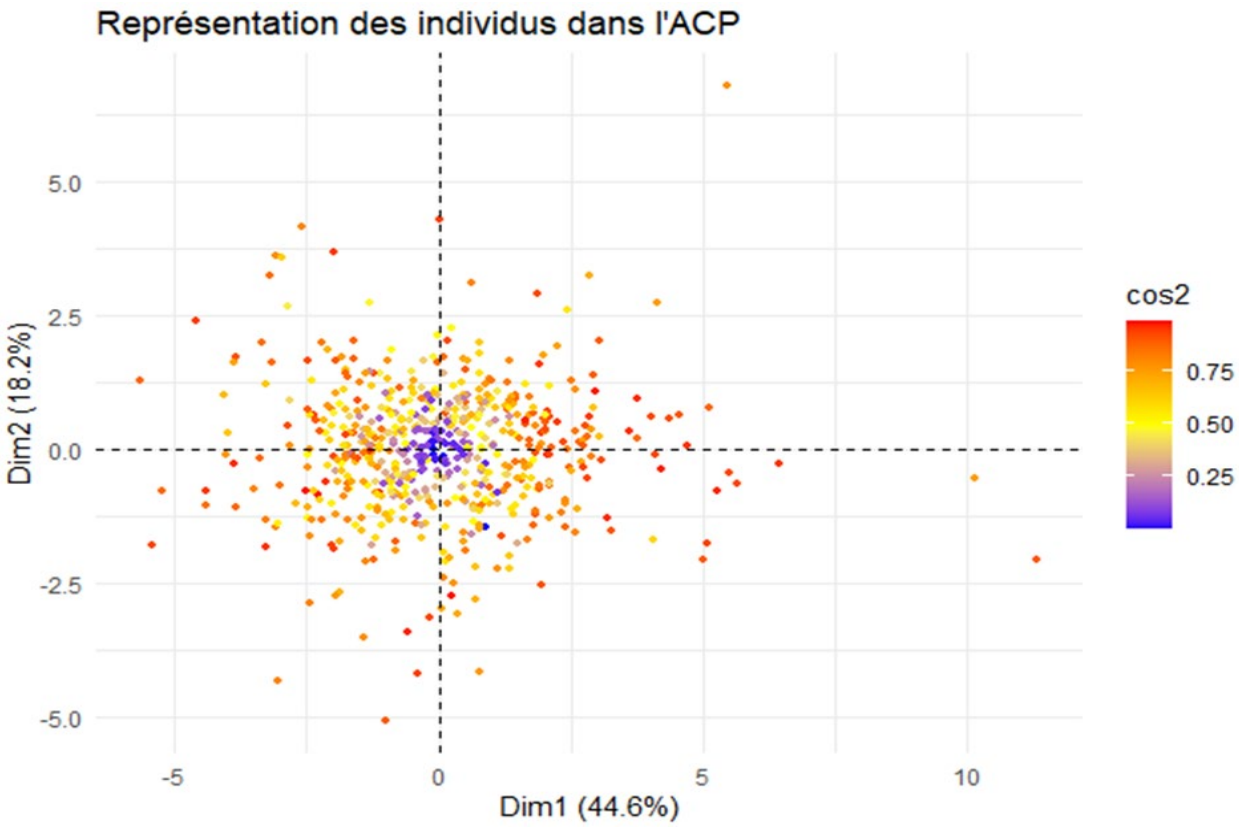
Les valeurs propres montrent la quantité de la variance expliquée par chaque composante principale. La première composante capture 44,64% de la variance avec une valeur propre

de 3,57 c'est la composante qui capture le plus d'information. Tandis que la deuxième capture que 18,22 % le cumul des deux fait 62,86% de la variance.

La troisième et la quatrième composantes ont des pourcentages approximatifs : 12,6% et 11,5% respectivement tandis que le cumul des 4 dernières est de 13,1% (une représentation faible de la variance).

Afin de simplifier l'analyse, il est suffisant de garder les deux premières en conservant plus de 62% de l'information.

Représentation graphique des individus sur les deux axes :

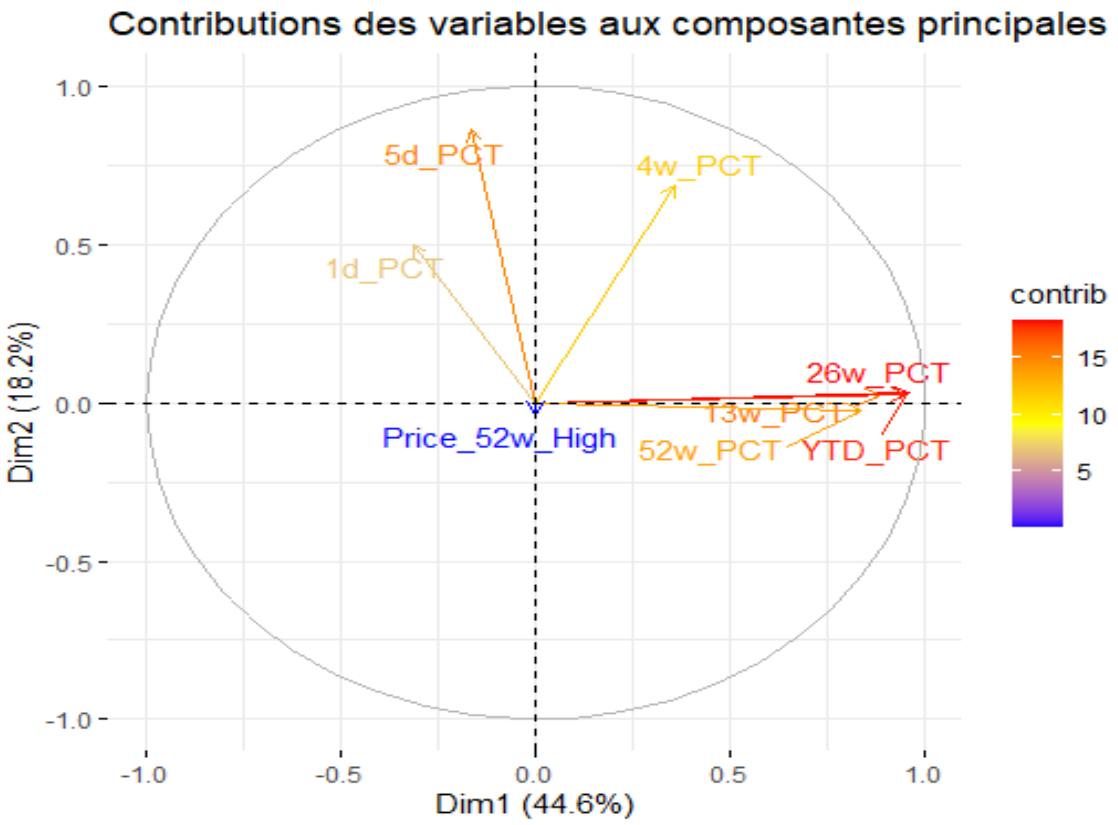


Commentaire :

Le graphique ci-dessus est une représentation des individus (entreprises dans notre cas) sur le plan des deux premiers axes. Ce graph permet d'identifier les individus avec des profils similaires.

Chaque point représente un individu (une entreprise), sa position (distance par rapport à l'origine) indique sa contribution à la réalisation des deux axes tandis que la couleur des points précise la qualité de la représentation (cos^2) : les individus avec point rouge sont les mieux représentés, contrairement aux individus du point bleu qui sont très proches de l'origine.

Le grand nombre d'individus et leur concentration rend difficile de tirer des conclusions significatives ou des résultats clairs pour les commenter.



Avant d'interpréter ce graphique, il serait pertinent de signaler que l'utilité de ce dernier est de faire un éclairage sur les variables pertinentes dans la construction des premières composantes. En se basant sur la couleur et la longueur des vecteurs, on peut facilement déterminer l'importance des variables du fait que le vecteur le plus long en ayant une couleur rouge par exemple représente la variable qui contribue plus à la création de la dimension qui lui correspond. Dans notre cas, il parait clairement que la construction de la dimension 1 est fortement due aux variables 26w_PCT, YTD_PCT et 13w_PCT en expliquant 44.6% de la variance totale alors que la création de la dimension 2 est fortement liée aux 1d_PCT, 5d_PCT et 4w_PCT en représentant 18.2% de la variance totale.

	Dim.1	Dim.2
1d_PCT	2.785947e+00	16.97388062
5d_PCT	7.585983e-01	50.60666142
4w_PCT	3.600629e+00	32.13052939
13w_PCT	2.200779e+01	0.03924258
26w_PCT	2.586244e+01	0.05946404
YTD_PCT	2.536577e+01	0.05212270
52w_PCT	1.961882e+01	0.05114207
Price_52w_High	9.581980e-06	0.08695717

Pour avoir la valeur de la contribution des variables pertinentes dans la construction de chaque dimension, on se base sur ces résultats numériques qui montrent que les variables 26w_PCT, YTD_PCT et 13w_PCT engendrant la création de DIM1 représentent respectivement les pourcentages suivants (25,86 %), (25,36 %) et (22,01 %), ce qui permet

d’affirmer que les tendances et le cumul des évolutions des prix sur des durées moyennes et longues sont principalement captés par cette dimension. En revanche, les variables 5d_PCT, 4w_PCT et 1d_PCT construisant principalement la dimension 2 représentent respectivement 50,61 % ,32,13 % et 16.97% ce qui permet de constater que les variations et les fluctuations à court terme sont principalement captées par cette dimension.

	Variable	Dim1_Ctr	Dim1_Cos2	Dim2_Ctr	Dim2_Cos2
1d_PCT	1d_PCT	2.785947e+00	9.950740e-02	16.97388062	0.2474324459
5d_PCT	5d_PCT	7.585983e-01	2.709533e-02	50.60666142	0.7377057901
4w_PCT	4w_PCT	3.600629e+00	1.286059e-01	32.13052939	0.4683746547
13w_PCT	13w_PCT	2.200779e+01	7.860660e-01	0.03924258	0.0005720488
26w_PCT	26w_PCT	2.586244e+01	9.237447e-01	0.05946404	0.0008668220
YTD_PCT	YTD_PCT	2.536577e+01	9.060048e-01	0.05212270	0.0007598054
52w_PCT	52w_PCT	1.961882e+01	7.007374e-01	0.05114207	0.0007455106
Price_52w_High	Price_52w_High	9.581980e-06	3.422455e-07	0.08695717	0.0012675962

Commentaire :

D'après les chiffres, on constate que dans la rubrique Dim1_cos² pour la dimension une, les variables 13w_PCT,26w_PCT et YTD-PCT représentent des valeurs proches de 1 ce qui permet de conclure que ces variables construisent principalement la dimension 1. Tandis que, les variables 1d_PCT, 5d_PCT et 4w_PCT représentent des valeurs importantes par rapport aux autres variables dans la rubrique Dim2_cos², à titre d’exemple la variable 5d_PCT, elle représente une valeur de 0.7377 qui proche de 1, ce qui permet d’affirmer que la dimension 2 est principalement construite par ces variables. A nouveau, on arrive aux mêmes interprétations mentionnées avant, les tendances et le cumul des fluctuations des prix sur des durées longues et moyennes sont captés par la dimension 1 alors que les évolutions du prix sur le court terme sont captées par la dimension 2.

L’extraction des données :

Dans cette partie, nous avons décidé d’extraire les deux premières composantes : F1 qui représente le premier axe avec 44,6% et F2 qui représente le deuxième axe avec 18,2% afin d’effectuer plusieurs tests et tirer des conclusions sur l’apport de l’ACP à l’estimation de la relation entre EPS Environment Pillar Score et les autres variables explicatives.

La régression linéaire :

Dans l'étude de la relation entre EPS et les autres variables nous avons décidé de tester plusieurs modèles :

1. Premier Modèle :

Dans ce modèle nous avons pris dans notre équation, en plus des variables données, les deux premières composantes de notre ACP qui représentent 62,86% de la variance cumulée.

```
F1           2.552e-01  4.740e-01    0.539  0.59044
F2          -4.960e-01  7.889e-01   -0.629  0.52987
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.25 on 517 degrees of freedom
Multiple R-squared:  0.3585,    Adjusted R-squared:  0.3225
F-statistic: 9.963 on 29 and 517 DF,  p-value: < 2.2e-16
```

Commentaire :

Les p-values des deux coefficients du F1 et F2 montrent que ces dernières ne sont pas significatives.

Le R^2 est de 0,3585 et le R^2 ajusté est de 0,3225.

Ce modèle explique environ 35,85% de la variance de la variable EPS.

2. Deuxième modèle :

Dans ce modèle nous avons pris que la première composante de notre ACP qui représente 44,6% de la variance :

```
F1           2.647e-01  4.734e-01    0.559 0.576348
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.24 on 518 degrees of freedom
Multiple R-squared:  0.358,     Adjusted R-squared:  0.3233
F-statistic: 10.32 on 28 and 518 DF,  p-value: < 2.2e-16
```

Commentaire :

La p-value du coefficient du F1 montre que cette dernière n'est pas significative.

Le R^2 est de 0,358 et le R^2 ajusté est de 0,3233.

Presque comme le modèle précedent, celui là explique environ 35,8% de la variance de la variable EPS.

3. Troisième modèle :

Dans ce modèle nous avons pris que la deuxième composante de notre ACP :

```
F2          -5.094e-01  7.880e-01  -0.646 0.518260
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.24 on 518 degrees of freedom
Multiple R-squared:  0.3581,    Adjusted R-squared:  0.3234
F-statistic: 10.32 on 28 and 518 DF,  p-value: < 2.2e-16
```

Commentaire :

Comme pour F1 : La p-value du coefficient du F2 montre que cette dernière n’est pas significative.

Le R^2 est de 0,3581 et le R^2 ajusté est de 0,3234.

Presque comme les modèle précedents, celui là explique environ 35,81% de la variance de la variable EPS.

4. Quatrième modèle :

Dans ce modèle nous avons décidé de ne pas prendre une variable financière :

```
Board        8.914e-02  3.927e-02   2.270 0.023617 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.23 on 519 degrees of freedom
Multiple R-squared:  0.3576,    Adjusted R-squared:  0.3242
F-statistic:  10.7 on 27 and 519 DF,  p-value: < 2.2e-16
```

Commentaire :

Le R^2 est de 0,3576 et le R^2 ajusté est de 0,3242.

Légèrement inférieus aux autres modèle, celui là explique environ 35,76% de la variance de la variable EPS. En terme de R^2 ajusté ce modèle est légèrement mieux.

Conclusion :

Les différents modèles testés montrent que les deux premières composantes issues de notre ACP qui représentent des variables financières, n’ont pas de contribution significative dans l’explication de la variable EPS Environment Pillar Score.

Dans notre précèdent rapport, sans passé par l’ACP, nous avons pris comme variable financière dans notre équation de régression linéaire : YTD Price PCT Change (Year-to-Date Price Percentage Change)

```
YTD                      1.560e+01  4.293e+00   3.635 0.000306 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.98 on 524 degrees of freedom
Multiple R-squared:  0.3739,    Adjusted R-squared:  0.3405
F-statistic: 11.18 on 28 and 524 DF,  p-value: < 2.2e-16
```

Contrairement aux résultats obtenus avec l'ACP, YTD a montré un effet statistiquement significatif (p-value de 0,000306) pour un coefficient de 15,6. Mais le R^2 ajusté reste légèrement supérieur par rapport aux autres modèles.

Pour conclure, Les mesures financières de notre base de données ne sont pas des déterminants significatifs de la performance environnementale des entreprises.