Institut National Supérieur des Sciences et Techniques d'Abéché

Département : Génie Informatique

Option : Génie Logiciel

Filière : Informatique Industrielle et de Gestion

Niveau : Licence III

Matière : Analyse des données

PROJET D'ANALYSE DES DONNÉES:

VENTE DES VOITURES

Présenté par :

1- BEKAIYOGOTO VANGILA

Téléphone: +235 63 91 24 80 et Email: bekaiyogotovangila@gmail.com

2- RONDOUBA MBAÏADEM ROCHTO

Téléphone: +235 65 75 81 37 et Email: rondoubarochto49@gmail.com

Encadré par : Dr Farikou Ousman

Année académique 2019-2020

TABLE DES MATIÈRES ET DES FIGURES

INTRODUCTION

L'analyse en composantes principales est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables liées entre elles (dites « corrélées » en statistique) en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées « composantes principales », ou axes principaux. Elle permet au praticien de réduire le nombre de variables et de rendre l'information moins redondante.

Avant tout, ce travail est une collecte des données faite pour construire un jeu de données, à partir de différents sites d'**Automobiles** et qui nous amène à la fin de cette étude "analyse en composantes principales" à conclure que le prix de la voiture est fortement corrélé surtout, avec les caractéristiques qui décrivent la puissance de la voiture.

Il est parfois difficile de décider quel véhicule acheter, avec quelles caractéristiques et avec quelle énergie car il arrive que l'on ait une mauvaise surprise en se retrouvant avec un véhicule qui ne correspond pas vraiment à ce que l'on attendait .Ce qui rend intéressant à étudier ce sujet à l'aide de l'analyse en composantes principales.

I. IMPORTATION DES DONNES

Dans un premier temps, nous importons le tableau des individus et variables actifs X (xij ; i =1,…, n, nombre d'observations ; j = 1,…,p, nombre de variables) pour la construction des axes factoriels. Nous utilisons la librairie **Pandas**. Certaines options de **read_excel()** sont susceptibles de modifications.
Nous remarquons que:
• Le fichier est un classeur Excel nommé «Jeu_de_donnees_voiture.xlsx »;
• Les données actives sont situées dans la première feuille (sheet_name = 0);
• La première ligne correspond aux noms des variables (header = 0);
• La première colonne aux identifiants des observations (index_col = 0).
Nous affichons la dimension de la matrice, nous récupérons le nombre d'observations (n=59) et de variables (p = 17), enfin nous affichons les valeurs mêmes.

II. PRÉPARATION DES DONNÉES

Nous devons explicitement centrer et réduire les variables pour réaliser une ACP normée avec PCA. Nous utilisons la classe StandardScaler pour ce faire.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

Où $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ est la moyenne de la variable X_j, $\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$ son écart-type.

#transformation – centrage-réduction

```
[58 rows x 16 columns]
0.23.1
[[-2.08323818e-01  9.78506460e-02 -1.30969099e+00 -1.16933431e+00
  -5.72049857e-01  8.52783546e-02  1.47489413e+00  1.41461294e+00
  -5.24037198e-01 -4.11416391e-01  3.37210092e-01  1.46446582e+00
   2.56973112e-01  1.21745477e-01 -1.33270262e+00 -5.72049857e-01]
 [-1.46316746e-01  2.74286008e-01  5.61296137e-01 -7.87913491e-01
  -4.66172449e-01 -1.48030351e-01  6.44371221e-01 -9.54646770e-02
  -5.24037198e-01 -1.65345262e+00  7.65327053e-01  1.46446582e+00
  -3.18487911e-01 -1.77935696e+00 -1.03937466e+00 -4.66172449e-01]
 [-2.22913717e-01 -2.25614184e-01 -5.36349642e-01 -1.00246270e+00
  -4.92641801e-01 -5.21324281e-01  6.44371221e-01 -9.54646770e-02
  -8.35421621e-02  7.55344919e-01  2.47080206e-01 -1.61159673e+00
  -3.18487911e-01 -1.77935696e+00 -9.97470666e-01 -4.92641801e-01]
 [-2.41672159e-01 -4.90267226e-01 -8.35707582e-01 -1.16933431e+00
  -5.45580505e-01 -5.67986022e-01  6.44371221e-01 -9.54646770e-02
   5.77200392e-01 -2.44383932e+00  4.19829155e-01  1.74504104e-01
   5.44703623e-01 -1.23618484e+00  1.70694863e+00 -5.45580505e-01]
 [-1.47098347e-01  6.56562625e-01 -1.62152217e-01 -1.09781791e+00
  -4.68378228e-01  8.52783546e-02 -1.01667459e+00 -5.98823883e-01
   7.97447911e-01  4.55672854e+00  2.32058558e-01  1.06755452e+00
  -1.46940996e+00 -9.64598776e-01  5.81987565e-01 -4.68378228e-01]
```


Vérifions, par acquit de conscience, les propriétés du nouvel ensemble de données. Les moyennes sont maintenant nulles (aux erreurs de troncature près) :

#moyenne

```
In [2]: print(numpy.mean(Z,axis=0))
[ 6.50820394e-17  1.26335723e-16  0.00000000e+00 -2.86169555e-16
 -6.31678617e-17 -4.93857828e-16  1.01451414e-16  1.99074473e-16
 -2.18216250e-16  5.88609621e-17  1.53134210e-17  2.05774095e-16
  5.55111512e-17  2.08645362e-16 -3.19667664e-16  1.62705098e-17]
```

Et les écarts-type unitaires.

#écart-type

```
In [3]: print(numpy.std(Z,axis=0,ddof=0))
[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
```

Nous sommes maintenant parés pour lancer l'ACP.

III. ANALYSE EN COMPOSANTES PRINCIPALES AVEC PCA DE ''SCIKIT-LEARN''

Le paramètre (svd_solver = 'full') indique l'algorithme utilisé pour la décomposition en valeurs singulières. Nous choisissons la méthode ''exacte'', sélectionnée de toute manière par défaut pour l'appréhension des bases de taille réduite.

1. Valeurs propres et scree plot

La propriété **.explained_variance_** semble faire l'affaire pour obtenir les variances (valeurs propres, λk) associées aux axes factoriels.

#variance expliquée

```
In [12]: print(acp.explained_variance_)
[3.03855589e+00 2.13712146e+00 1.63914352e+00 1.57215853e+00
 1.42637444e+00 1.20035430e+00 1.01976855e+00 9.63492064e-01
 8.70387806e-01 7.15433196e-01 5.18241570e-01 4.78228713e-01
 4.04119141e-01 1.79682007e-01 1.17640580e-01 3.66676246e-32]
```

PCA fournit également les proportions de variance associées aux axes.

#proportion de variance expliquée

```
In [20]: print(acp.explained_variance_ratio_)
[1.86635437e-01 1.31267158e-01 1.00680152e-01 9.65657715e-02
 8.76113609e-02 7.37286583e-02 6.26366461e-02 5.91800082e-02
 5.34613200e-02 4.39436338e-02 3.18316481e-02 2.93739619e-02
 2.48219731e-02 1.10365026e-02 7.22576840e-03 2.25221401e-33]
```

Nous disposons des éléments permettant de construire le graphique ''Scree plot'' (éboulis des valeurs propres) de la figure 1.

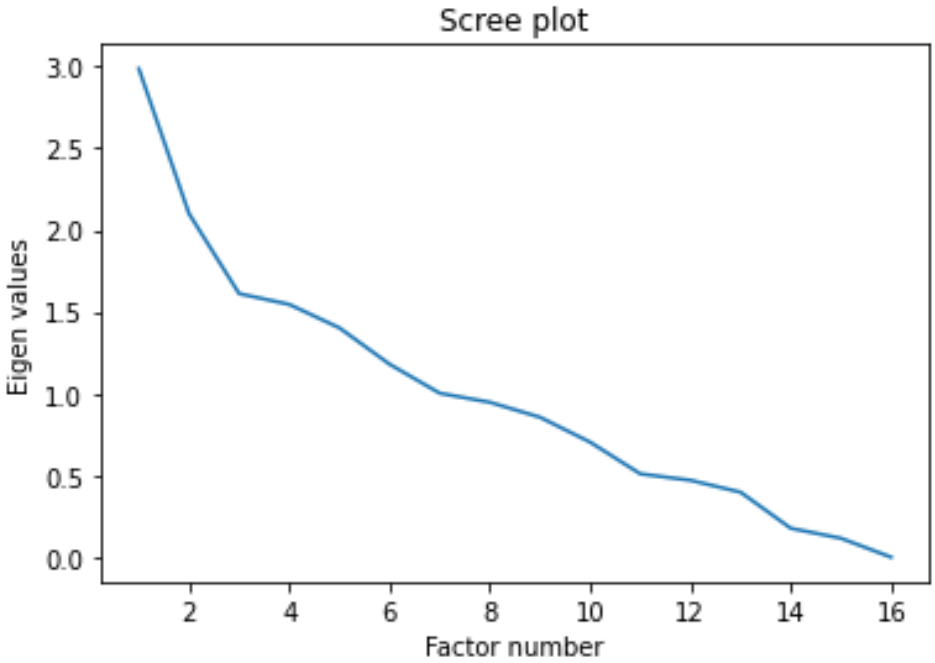

Figure 1 : Scree plot

Le graphique du cumul de variance restituée selon le nombre de facteurs peut être intéressant également (Figure 2).

#cumul de variance expliquée

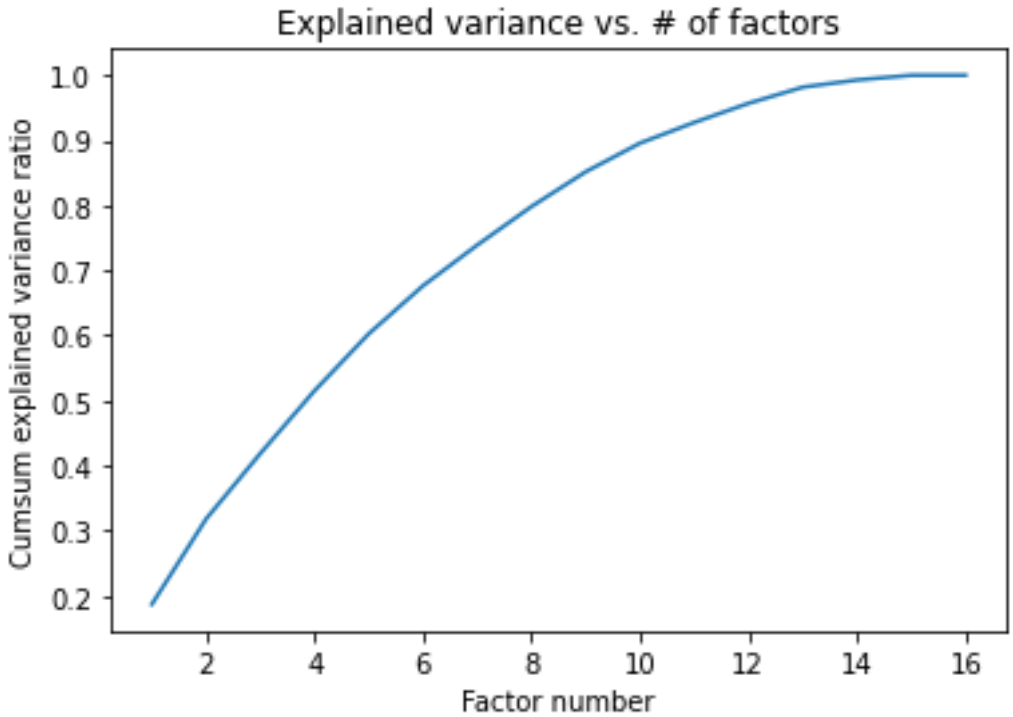

Figure 2 - Variance expliquée vs. Nombre de facteurs

2. Représentation des individus et des Outils pour l'interprétation

Les coordonnées factorielles (Fik) des individus sont positionnées dans le premier plan factoriel avec leurs labels pour situer et comprendre les proximités entre les véhicules.

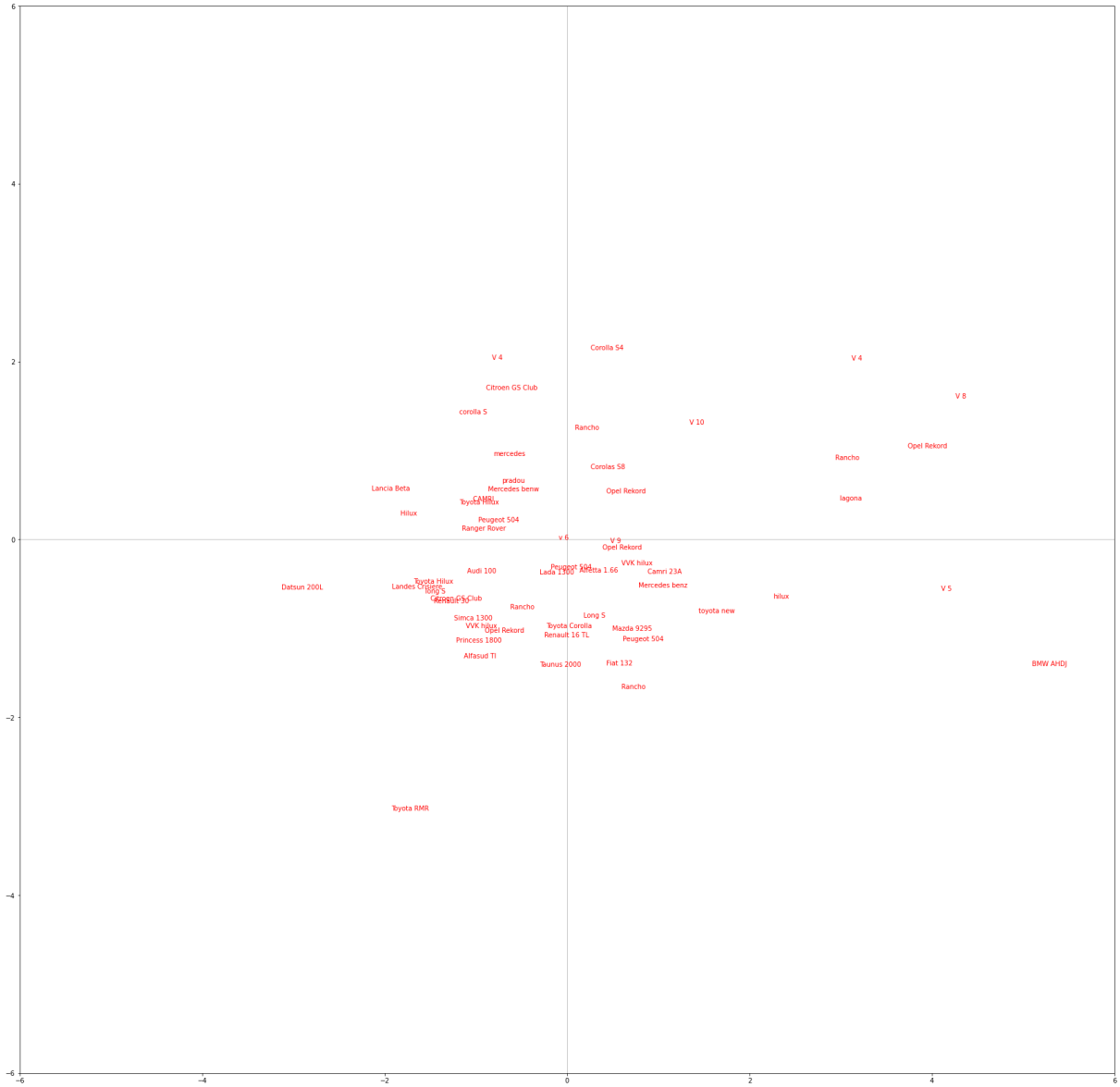

Figure 3 : Représentation des individus dans le premier plan factoriel

3. Contribution des individus aux axes (CTR)

Elles permettent de déterminer les individus qui pèsent le plus dans la définition de chaque facteur.

$$CTR_{ik} = \frac{F_{ik}^2}{n \times \lambda_k}$$

```
In [27]: print(pandas.DataFrame({'id':X.index,'CTR_1':ctr[:,0],'CTR_2':ctr[:,
1]}))
                id      CTR_1      CTR_2
0        Alfasud TI   1.286865   1.779955
1          Audi 100   1.200340   0.141145
2        Simca 1300   1.544896   0.823919
3   Citroen GS Club   2.241393   0.468688
4          Fiat 132   0.184296   1.998406
5       Lancia Beta   4.586104   0.306570
6       Peugeot 504   0.371891   1.299694
7     Renault 16 TL   0.061224   1.197736
8        Renault 30   2.138264   0.509675
9    Toyota Corolla   0.052627   0.987555
10     Alfetta 1.66   0.018336   0.134524
11    Princess 1800   1.476119   1.330750
12      Datsun 200L   9.793302   0.308823
13      Taunus 2000   0.089509   2.047584
14           Rancho   0.390797   0.607797
15       Mazda 9295   0.243271   1.043212
16      Opel Rekord   0.149997   0.012028
17        Lada 1300   0.089728   0.150784
18            RAV 4   5.879954  58.429000
19     Ranger Rover   1.332839   0.010311
20  Landes Crisiere   3.681137   0.306299
21     Toyota Hilux   2.840513   0.245028
22          BMW AHDJ  26.005992   2.018177
23      Toyota RMR    3.717657   9.266663
24        Camri 23A   0.779919   0.143784
```

4. Représentation des variables – Outils pour l'aide à l'interprétation

Nous avons besoin des vecteurs propres pour l'analyse des variables. Ils sont fournis par le champ **.components_**

#le champ components_ de l'objet ACP

```
In [29]: print(acp.components_)
[[-7.12285228e-02 -1.28256551e-01 -6.56419183e-02  5.67054799e-04
   4.72680734e-01  1.56152075e-01 -3.66032509e-01 -4.21629299e-01
  -5.13989497e-02  1.51471573e-01 -1.76432472e-01  2.06106318e-01
   1.66988500e-01  2.20830999e-01 -1.23829221e-01  4.72680734e-01]
 [-1.97862437e-01  8.99483493e-02  3.84943633e-01  5.65869330e-01
   1.10592954e-01 -2.27049360e-01  1.11358630e-02 -6.14329575e-02
   3.64122798e-01 -1.42785276e-01  4.13956692e-01 -1.18633299e-01
   3.13485089e-02  2.36492483e-01 -9.56062151e-02  1.10592954e-01]
```

Les variables sont maintenant en ligne, les facteurs en colonne :

#corrélation des variables avec les axes

```
In [40]: print(corvar)
[[-1.23086659e-01 -2.86748653e-01  8.15905850e-02  6.37555594e-01
   5.62962465e-02 -1.87673576e-01 -1.08218172e-01 -2.25925248e-01
   3.04329668e-01  4.39212739e-01  3.18489378e-01 -8.59718593e-03
  -1.13770807e-02 -2.36850880e-02 -2.49500034e-02  4.68734106e-33]
 [-2.21634112e-01  1.30356062e-01  5.46813581e-01 -2.62282344e-01
  -1.54602509e-01  1.16768836e-01 -3.57283830e-01  5.14518440e-01
   1.20925705e-01  1.30105835e-02  2.74915010e-01  6.25380050e-02
   1.91250116e-01  2.80134855e-02  6.72882748e-02 -1.30555946e-32]
 [-1.13432711e-01  5.57872783e-01  5.89800502e-01  3.58368534e-02
  -8.46533740e-02 -3.94737341e-01  1.20171611e-01 -4.74266225e-02
  -1.20566008e-01 -4.21638515e-05 -5.59761848e-02 -2.86343841e-01
   4.67555139e-02 -1.98489216e-01 -7.64294656e-02  3.03890952e-32]
```

Si l'on s'en tient spécifiquement aux deux premiers facteurs :

#on affiche pour les deux premiers axes

```
In [41]: print(pandas.DataFrame({'id':X.columns,'COR_1':corvar[:,
0],'COR_2':corvar[:,1]}))
                              id     COR_1     COR_2
0                  CYLINDRE (cm3) -0.123087 -0.286749
1                    PUISSANCE kW -0.221634  0.130356
2                       LONGUEUR  -0.113433  0.557873
3                        LARGEUR   0.000980  0.820076
4                           POIDS  0.816817  0.160275
5                           V.MAX  0.269839 -0.329047
6                           PORTE -0.632524  0.016138
7                           PLACE -0.728598 -0.089031
8      CONSOMMATION URBAINE(km/l) -0.088820  0.527699
9                    RÉSERVOIR(l)   0.261751 -0.206929
10              NIVEAU CO2(g/kg) -0.304885  0.599919
11          PUISSANCE EN CHEVAUX   0.356163 -0.171927
12         PUISSANCE MAXI(tr/min)  0.288565  0.045431
13  ACECELERATION  0-100 km/h (s)  0.381608  0.342733
14                    HAUTEUR(cm) -0.213983 -0.138556
15                      PRIX FCFA  0.816817  0.160275
```

Nous pouvons dessiner maintenant le cercle des corrélations (Figure 4)

#cercle des corrélations

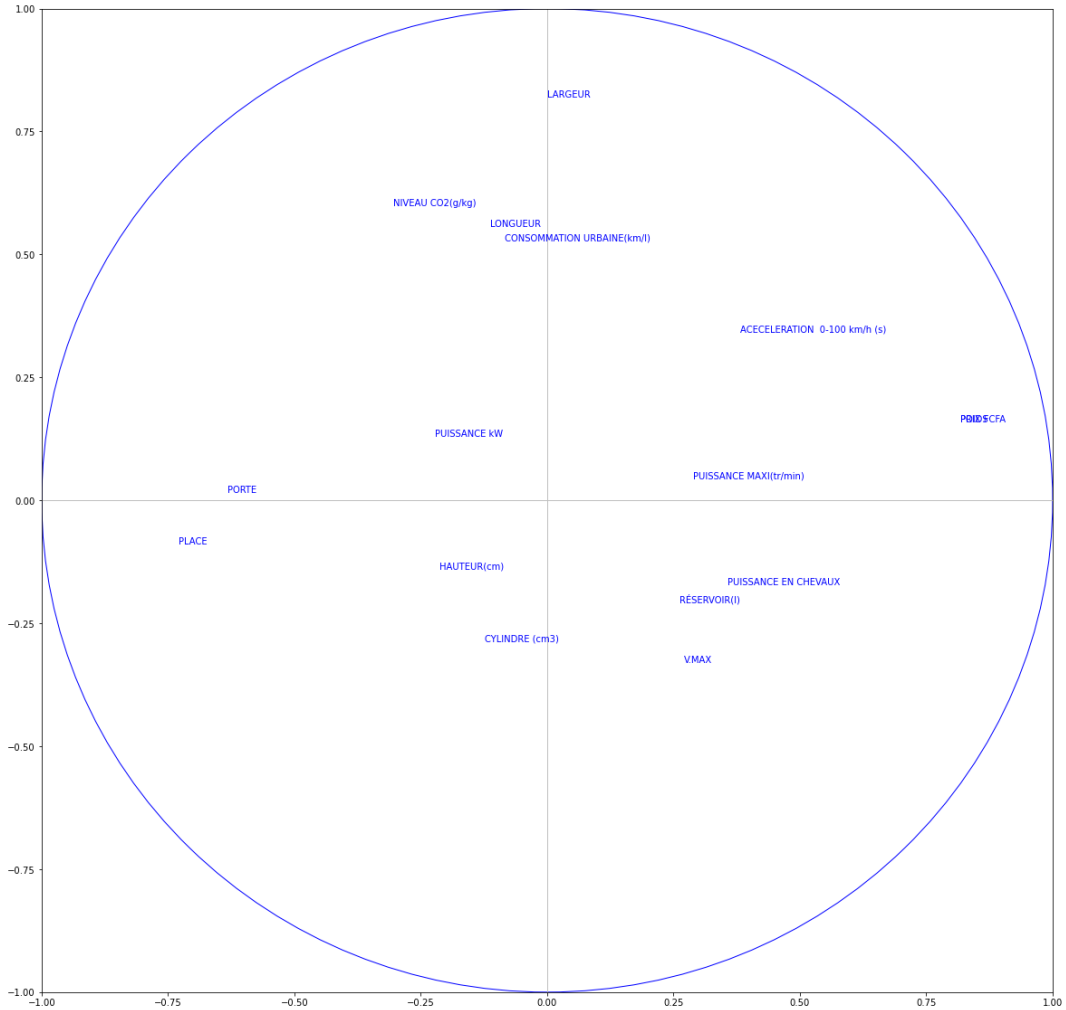

Figure 4 : Cercle des corrélations

On perçoit clairement l'effet taille sur le premier axe : les voitures en puissance max et en accélérations de 0-100km/s la relation globale entre les variables est en réalité déterminée par le PRUX FCFA et POIDS.

CONCLUSION

Les champs d'application sont aujourd'hui multiples, allant de la biologie à la recherche économique et sociale, et plus récemment le traitement d'images et l'Apprentissage Automatique. L'ACP est majoritairement utilisée pour:

- décrire et visualiser des données ;
- les décorréler ;
- la nouvelle base est constituée d'axes qui ne sont pas corrélés entre eux ;
- les débruiter, en considérant que les axes que l'on décide d'*oublier* sont des axes *bruités*.
- effectuer une réduction de dimension des données d'entrainement en Apprentissage Automatique

On a pu réduire tout le problème de la synthèse de l'information à partir de 59 voitures et 16 variables en deux dimensions avec une précision de 84.4% de l'information totale. À partir de l'ACP faite on a pu répondre à la problématique et on peut dégager les résultats suivants : le prix de la voiture est totalement lié à sa puissance beaucoup plus que son design.

RÉFÉRENCE

Certaines données sont prises dans :

1. http://tutoriels-datamining.blogspot.com/2018/06/acp-avec-python.html
2. Sur Wikipedia, Analyse en composantes principales
3. PROJET *ACP de* CHARNI KHOULOUD **ÉTUDE SUR LES PROFILS VOITURES** *en* 30 avril 2016