TP 6 DATA MINING : Analyse des Correspondances Simples

SHANMUGAVADIVEL Sophitha- DINAR Rayan - AIT-BAHA Nabil

2018-05-13

1 Contexte

L' analyse factorielle des correspondances simples est une extension de l'analyse en composantes principales permettant d'analyser l'association **entre deux variables qualitatives (ou catégorielles)** .
L'AFC permet de résumer et de visualiser l'information contenue dans le tableau de contingence formé par les deux variables catégorielles. Le tableau de contingence contient les fréquences formées par les deux variables.

2 Problématique

Existe-t-il un lien entre la couleur des cheveux et la couleur des yeux ?

3 Importation des données

Ci-dessous nos données brut, en colonne nous avons les couleurs de cheveux et en lignes les couleurs des yeux. Ce tableau de contingence contient donc les fréquences couleurs des "cheveux" et couleurs des "yeux".

	Brun	Châtain	Roux	Blond	Total	Chinoises	Suédoises
Marron	68	119	26	7	220	75	5
Noisette	15	54	14	10	93	20	10
Vert	5	29	14	16	64	5	20
Bleu	20	84	17	94	215	0	65
Total	108	286	71	127	592	100	100

4 Analyse du tableau

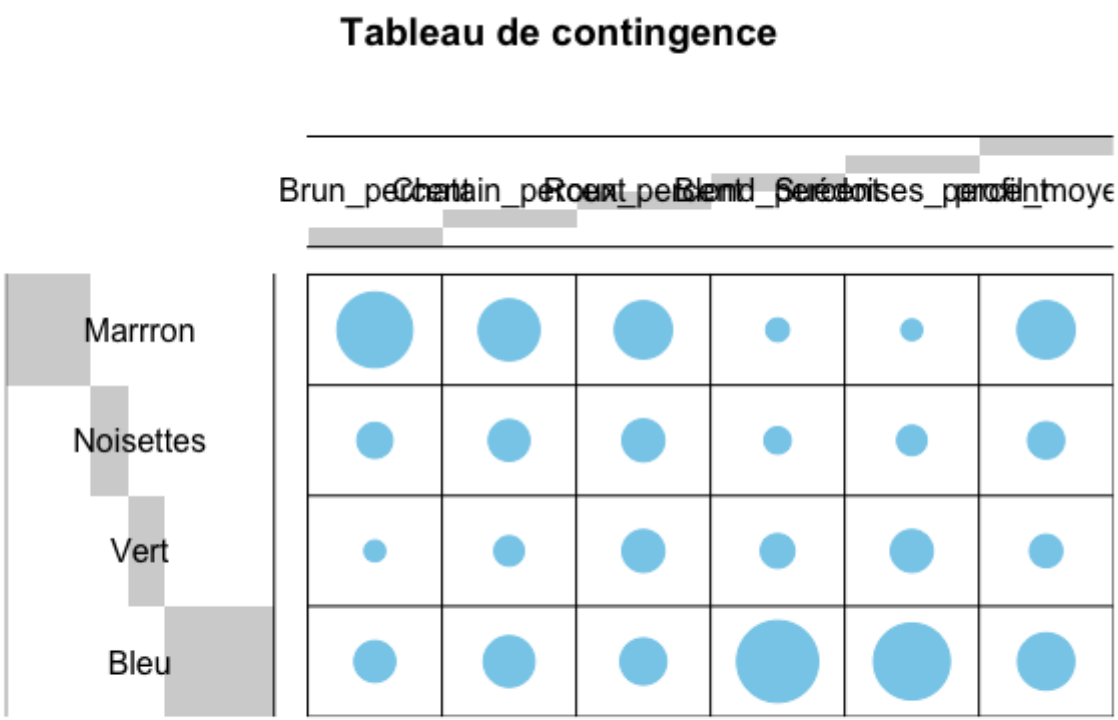
	Brun	Chatain	Roux	Blond	Chinoises	Suédoises	profil_moyen
Marrron	63%	42%	37%	6%	75%	5%	37%
Noisettes	14%	19%	20%	8%	20%	10%	15%
Vert	5%	10%	20%	13%	5%	20%	12%
Bleu	19%	29%	24%	74%	0%	65%	36%
Total	100%	100%	100%	100%	100%	100%	100%

On peut lire le tableau de la manière suivante :

- **63% des anglaises aux cheveux bruns ont les yeux marrons**
- **37% des anglaises au global ont les yeux marrons**
- **Une grande majorité de personnes ayant les yeux bleu (profil_moyen = 36%) sont des Blond (74%). Une personne ayant les yeux bleu a donc plus de chances d'avoir les bleu.**

A première vue on remarque une opposition radical entre les couleurs "marron" et "bleu" en termes de proportions. En effet, pour chaque couleur de cheveux, lorsque l'une de ces proportions est élevée l'autre est plus faible et inversement.

La couleur de cheveux "chatain" est proche du profil moyen.



On retrouve sur ce tableau de contingence, en colonne le même ordre que les variables situés dans le tableau vu précédement, le tableau des profils colonnes.

5 Selection des individus et des variables actives

Les couleurs de cheveux, en colonnes, sont **déclarées actives** et déterminent les ressemblances et les dissemblances entre les couleurs des yeux. Ce sont elles qui vont intéragir dans le modèle et nous permettre

de répondre à la problèmatique.

Les populations des chinoises et suédoises sont déclarés comme des **fréquences illustratives.**

6 Significativité statistique

Pour interpréter l'AFC, la première étape consiste à évaluer s'il existe une dépendance significative entre les lignes et les colonnes.

Une méthode rigoureuse consiste à utiliser la **p-value** pour examiner l'association entre les modalités des lignes et celles des colonnes

```
## [1] 5.623837e-27
```

Une p-value élevée signifie un lien fort entre les lignes et les colonnes, comme nous pouvons le voir ici.

7 Visualisation & Interprétation

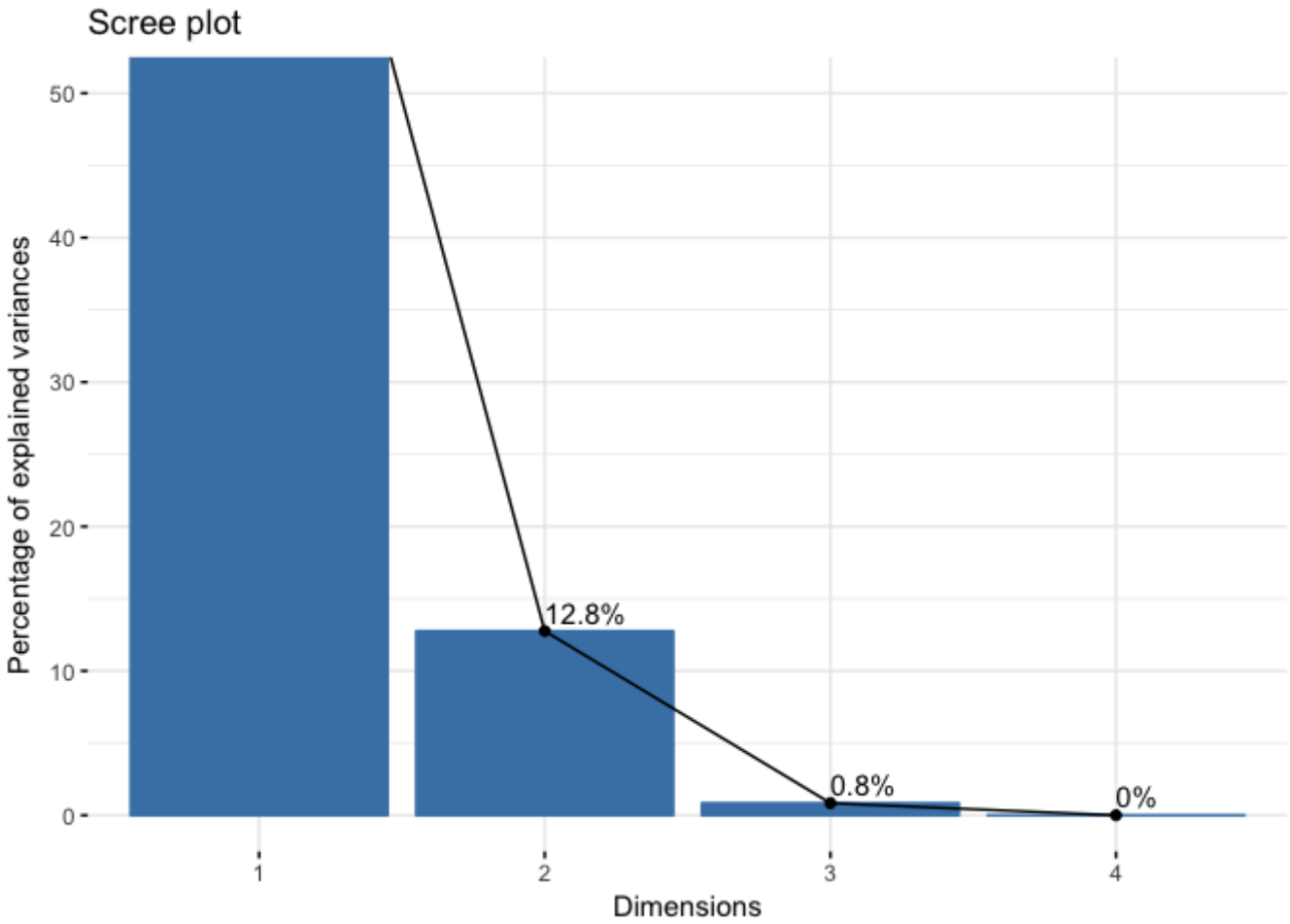
7.1 Détermination des axes : Analyse des valeurs propres

L'examination des valeurs propres permet de déterminer le nombre d'axes principaux à considérer. Les valeurs propres correspondent à la quantité d'informations (l'inertie) retenue par chacun axe. Contrairement à l'ACP, vue précédemment, **aucune valeur propre n'est supérieure à 1** . Cela est dû à un nombre de modalités actives plus élevé.

Les valeurs propres peuvent être utilisées pour déterminer le nombre d'axes à retenir. Il n'y a pas de «règle générale» pour choisir le nombre de dimensions à conserver pour l'interprétation des données. Par exemple en médecine nous retenons des variance autour de 90% alors qu'en marketing une variance 70% peut être acceptable.

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.0976814	86.3962254	86.39623
Dim.2	0.0144254	12.7588205	99.15505
Dim.3	0.0009528	0.8427415	99.99779
Dim.4	0.0000025	0.0022126	100.00000

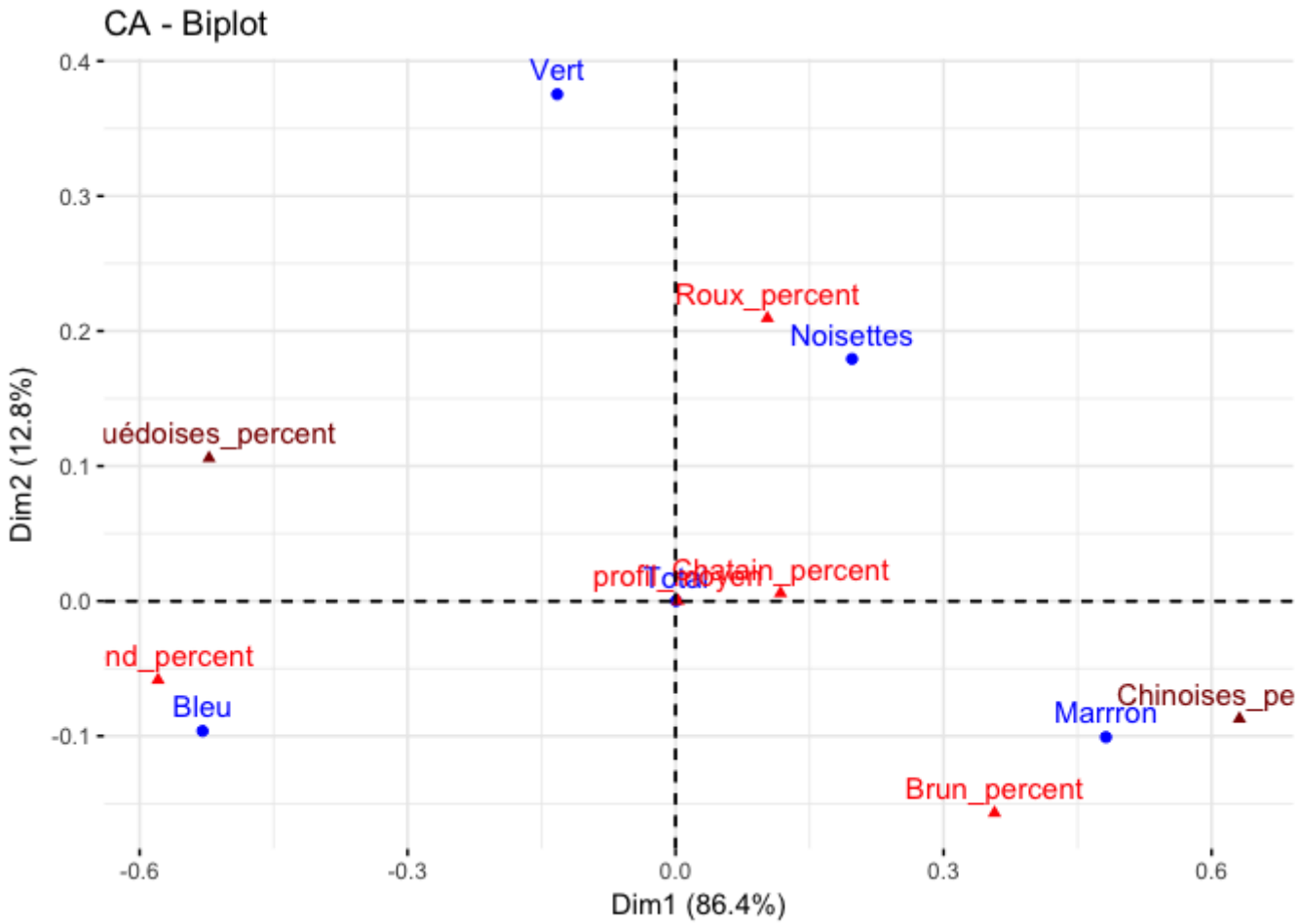
Les deux premiers axes restituent près de 99% de l'inertie total. Soit près de 99% de la quantité d'informations contenue dans le tableau de contigence.



Une autre méthode pour déterminer le nombre de dimensions est de regarder le graphique des valeurs propres (critère de coude), ordonnée de la plus grande à la plus petite valeur. Le nombre d'axes est déterminé par le point, au-delà duquel les valeurs propres restantes sont toutes relativement petites et de tailles comparables.

Dans notre cas nous choisirons donc **les deux premiers axes**

7.2 Visualiation



Dans le graphique ci-dessus, les lignes sont représentées par des points bleus, les colonnes actives par des triangles rouges et les colonnes illustratives par des triangles bordeaux.

La distance entre les points lignes ou entre les points colonnes donne une mesure de leur similitude. Les points lignes avec un profil similaire sont proches sur le graphique. Il en va de même pour les points colonnes.

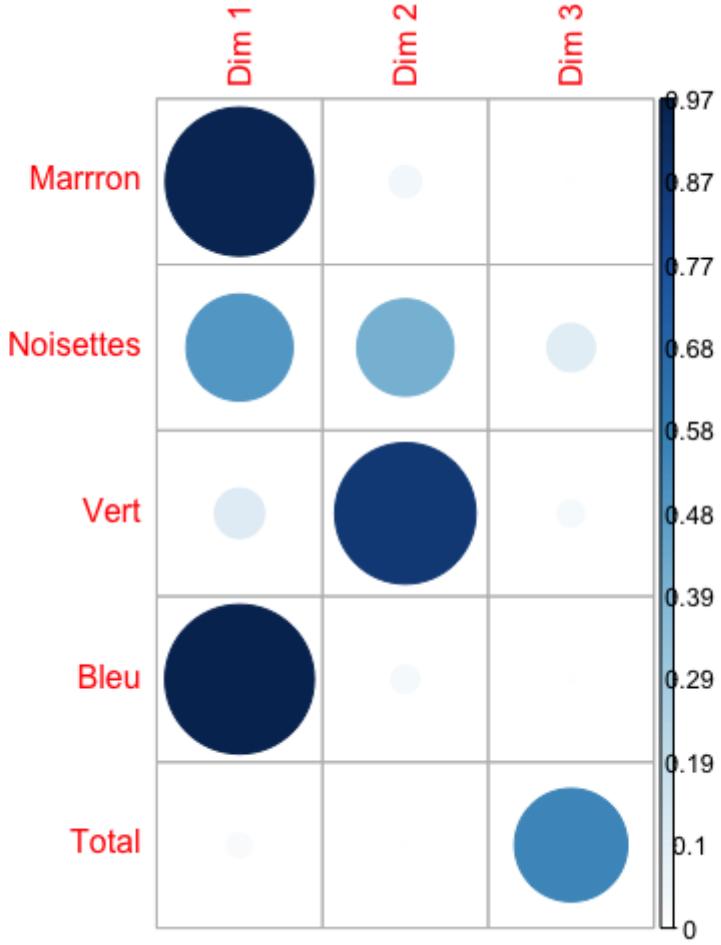
La couleur **“Châtain”** est la couleur la plus proche du profil moyen, celle dont la répartition des couleurs de yeux est la plus proche de l’ensemble.

La couleur des yeux qui correspond le plus aux Suédoises est davantage le Bleu **(65%)** et le Vert **(20%)** que les autres couleurs **(5% pour le Marron et 10% pour Noisettes)** .

Pour les Chinoises, la proportion ayant les yeux de couleur Bleu est nulle **(0%)** contrairement à celle ayant les yeux Marron **(75%)** .

8 Qualité de représentation

Le résultat de l’analyse montre que le tableau de contingence est bien représenté dans un espace à faibles dimensions en utilisant l’AFC. Les deux dimensions 1 et 2 sont suffisantes pour conserver 99% de l’inertie totale (variation) contenue dans les données.



Les cosinus carrés les plus élevés correspondent aux lignes **bien représentées sur chacun des axes** .

Les valeurs de cos2 sont comprises entre 0 et 1. La somme des cos2 pour les lignes sur toutes les dimensions de l'AFC est égale à 1.

Si un point ligne est bien représenté par deux dimensions, **la somme des cos2 est proche de 1** .

Pour certains éléments lignes, plus de 2 dimensions sont nécessaires pour représenter parfaitement les données.

On remarque que les yeux marrons et bleu sont bien répresentés par l'axe 1, la couleur verte est quant à elle bien representée par l'axe 2. Cette axe permet principalement la visualisation des préférences de couleur des yeux pour la couleur Roux. La couleur noisette est representée par les deux premiers axes.

8.1 Conclusion

Il existe bien un lien entre la couleur des cheveux et la couleur des yeux **sauf pour la couleur Châtain** dont les proportions sont vraiment proche du profil moyen. En effet les Anglaises blondes ont plus de chances d'avoir les yeux bleus **(74%)** , les brunes ont plus de chances d'avoir des yeux marrons **(63%)** et les rousses ont plus de chances d'avoir des yeux verts. En outre, les Chinoises ont une probabilité élévée d'avoir des yeux marrons **(75%)** alors que les Suédoises ont plus de chances d'avoir des yeux bleus **(25%)** ou verts **(65%)** .