**Réponse questions e-commerce (Partie statistiques)**

1 - ACP : Expliquer comment et pourquoi on passe d'un tableau de Big data que l'on souhaite explorer à la matrice de variance-covariance (ou matrice de dispersion), puis à celle de corrélation.

Expliquer comment on en arrive à considérer les vecteurs propres de cette matrice de dispersion comme les directions des axes les plus propices à la conservation de la plus grande inertie et à une exploration pratique de données projetées de grandes dimensions.

Au départ, on a un tableau de données contenant de nombreuses variables (dimensions) et des observations (cas) dans un espace de grande dimension. Chaque variable représente une caractéristique différente des observations, et cela peut être difficile à visualiser ou à analyser de manière significative lorsque le nombre de variables est élevé.

La première étape de l'ACP consiste à calculer la matrice de variance-covariance à partir du tableau de données. Cette matrice mesure les relations statistiques entre toutes les paires de variables dans les données. Elle indique à quel point les variables varient conjointement les unes par rapport aux autres.

La matrice de variance-covariance est essentielle car elle révèle les dépendances et les relations entre les variables, ce qui est important pour comprendre la structure des données.

Ensuite, on peut calculer la matrice de corrélation à partir de la matrice de variance-covariance. La matrice de corrélation normalise les valeurs de la matrice de variance-covariance, en les divisant par les écarts-types des variables.

La matrice de corrélation exprime les relations linéaires entre les variables sous une forme standardisée, en indiquant les coefficients de corrélation entre chaque paire de variables. Cela permet de comparer plus facilement l'importance des relations entre les variables.

La prochaine étape est de décomposer la matrice de corrélation en vecteurs propres et valeurs propres. Les vecteurs propres sont les directions dans lesquelles les données varient le plus, tandis que les valeurs propres indiquent la quantité de variance expliquée par chaque vecteur propre.

En considérant les vecteurs propres correspondant aux valeurs propres les plus élevées, on identifie les directions ou les axes dans lesquels les données sont les plus dispersées ou les plus "inertes". Ces axes sont les plus informatifs pour réduire la dimensionnalité tout en préservant au maximum les informations contenues dans les données d'origine.

Enfin, on peut projeter les données d'origine dans l'espace défini par les vecteurs propres sélectionnés. Cette projection permet de transformer les données initiales en un nouvel ensemble de variables (composantes principales) qui capturent l'essentiel de la variation des données d'origine tout en réduisant la dimensionnalité.

Cette transformation simplifie l'exploration des données, car on peut désormais visualiser et analyser les données dans un espace de dimensions réduites tout en conservant l'essentiel de leur structure et de leurs relations.

2 - ACP : Expliquer comment on interprète la représentation des individus et l'origine de cette manière de faire.

Comment peut-on alors déterminer la signification concrète des axes principaux F1 et F2 ?

Qu'apportent dans ce contexte les résultats fournis par la fonction PCA de FactoMineR ?

Dans l'ACP, les individus ou les cas d’un ensemble de données sont représentés comme des points dans l'espace des composantes principales, qui est un espace de dimensions réduites. Chaque individu est projeté sur ces axes principaux, et sa position dans cet espace est déterminée par les valeurs des composantes principales correspondantes.

L'interprétation des individus dans cet espace dépend de la proximité des individus les uns par rapport aux autres. Des individus proches dans l'espace des composantes principales partagent des caractéristiques similaires, tandis que des individus éloignés diffèrent davantage.

Les axes principaux F1, F2, sont des vecteurs propres issus de la décomposition de la matrice de corrélation ou de covariance. F1 est le vecteur propre qui capture la plus grande variance des données, F2 la deuxième plus grande.

La signification concrète des axes principaux dépend de la nature des données d'origine. Pour comprendre ce que représentent F1 et F2, on doit examiner les coefficients de chaque variable sur ces axes. Les variables qui ont des coefficients élevées ou négatives sur un axe particulier sont celles qui contribuent le plus à cette composante.

Par exemple, si les variables liées à la taille et au poids des individus ont des coefficients élevés sur F1, alors F1 peut être interprété comme un "axe de taille" ou un "axe de poids". De même, si F2 est fortement influencé par les variables liées à l'âge et au revenu, il peut être interprété comme un "axe d'âge" ou un "axe de revenu".

La fonction PCA permet d'effectuer une ACP sur nos données et fournit des résultats détaillés pour aider à l'interprétation.

Les résultats fournis par PCA comprennent notamment les coordonnées des individus dans l'espace des composantes principales, les coefficients des variables sur chaque composante, les valeurs propres expliquant la variance, les contributions individuelles des individus à chaque composante, les graphiques etc.

Ces résultats nous aident à comprendre la signification des axes principaux F1 et F2, ainsi que la contribution de chaque individu et de chaque variable à ces axes. On peut également utiliser des graphiques, pour visualiser les relations entre les individus, les variables et les axes principaux.

3 - ACP : Expliquer comment on interprète la représentation des variables et l'origine de cette manière de faire.

Que sont en fait les coordonnées des points qui représentent chaque variable ?

Quel rapport entre la qualité de la projection et le cosinus ?

Qu'apportent dans ce contexte les résultats fournis par la fonction PCA de FactoMineR ?

Expliquer pourquoi et comment on peut voir que des variables sont corrélées d'une manière ou d'une autre ou, au contraire, ne le sont pas du tout.

Dans l'ACP, les variables (ou caractéristiques) sont également représentées dans l'espace des composantes principales. Chaque variable est représentée comme un vecteur à partir de l'origine de cet espace jusqu'à son point correspondant.

L'orientation de ces vecteurs par rapport aux axes principaux indique la relation de chaque variable avec les composantes principales. Plus la variable est alignée avec un axe principal, plus elle est corrélée (positivement ou négativement) avec cette composante principale.

Les coordonnées des points représentant chaque variable dans l'espace des composantes principales sont les charges (ou poids) de chaque variable sur les composantes principales. Ces charges indiquent la contribution de chaque variable à chaque composante principale.

Si une variable a une charge élevée sur une composante principale particulière, cela signifie que cette variable contribue fortement à cette composante principale et qu'elle est importante pour expliquer la variance des données dans cette direction.

La qualité de la projection d'une variable sur une composante principale est souvent évaluée en utilisant le cosinus de l'angle entre le vecteur de la variable et le vecteur de la composante principale. Plus le cosinus est proche de 1, meilleure est la projection de la variable sur cette composante principale.

Un cosinus proche de 1 indique que la variable est bien alignée avec la composante principale et qu'elle contribue significativement à expliquer la variance dans cette direction. Un cosinus proche de 0 signifie que la variable a peu d'impact sur cette composante principale.

La fonction PCA de FactoMineR, comme d'autres outils similaires, fournit des informations sur les coordonnées des variables, les charges sur les composantes principales, les cosinus, et d'autres statistiques utiles pour l'interprétation.

Ces résultats aident à comprendre comment les variables d'origine sont liées aux composantes principales, quelles variables contribuent le plus à expliquer la variance, et dans quelle mesure les variables sont alignées ou non avec les axes principaux.

Pour déterminer si des variables sont corrélées, on peut examiner les charges des variables sur les composantes principales. Si deux variables ont des charges similaires sur une même composante principale, cela indique une corrélation positive entre elles.

Au contraire, si deux variables ont des charges opposées sur une composante principale (l'une positive et l'autre négative), cela indique une corrélation négative entre elles.

Si une variable a des charges faibles sur toutes les composantes principales, elle est moins corrélée avec les autres variables et a moins d'influence sur la structure des données.