PROJET  **et SAS**
ETUDE DE STATIONS DE SKI

Mot clé : Outils graphiques, Analyse en composante principale, Classification, discrimination, programmation légère

On a relevé pour 43 stations de ski en France 11 variables qualitatives décrites dans le fichier de données et les 19 variables quantitatives ci dessous :

altitude de la station, altitude du sommet des pistes,

kilométrage total des pistes de ski alpin,

nombre de pistes vertes (faciles),

nombre de pistes bleues (moyennes),

nombre de pistes rouges (difficiles),

nombre de pistes noires (très difficiles),

nombre de téléphériques,

nombre de télécabines,

nombre de télésièges,

nombre de téléskis,

prix du forfait journée de ski alpin adultes,

kilométrage total des pistes de ski de fond,

prix du forfait journée de ski de fond adultes,

nombre de lits disponibles pour l'hébergement,

prix hôtel 2 étoiles par personne et par jour basse saison,

prix hôtel 2 étoiles par personne et par jour haute saison,

prix appartement 4 personnes par semaine basse saison,

prix appartement 4 personnes par semaine haute saison.

L'objectif de ce projet est de faire une étude descriptive de ce jeu de données et en particulier de construire une typologie des stations de ski en classes. On utilisera uniquement les variables quantitatives et la variable qualitative label ski France.

Partie R

Analyse exploratoire des données et analyse en composantes principales.

L'objectif de l'analyse exploratoire des données est de visualiser l'information contenue dans chacune des variables.

1) Construire un tableau X avec les 19 variables quantitatives centrées et réduites. Dans toute la suite, on travaillera sur les données centrées réduites.

2) Représenter, à l'aide d'outils graphiques disponibles sous l'information contenue dans les 10 premières variables. Faîtes travailler votre imagination pour exploiter au mieux les richesses graphiques du logiciel choisi. L'utilisation des graphes conditionnés disponibles dans le package lattice sera appréciée.

3) Effectuer une ACP sur les 10 premières variables quantitatives. Donner le pourcentage de variance expliquée par les différents axes de l'ACP.

4) Construire un tableau Z contenant les coordonnées des observations sur les deux premiers axes principaux.

5) Effectuer une classification en 3 classes avec la méthode des K-means à partir du tableau Z.

6) Visualiser les individus sur le plan factoriel. Colorez-les selon leur classe obtenue avec la méthode des K-means.

7) Calculer la matrice D des distances entre les observations à partir du tableau Z.

8) Effectuer une classification ascendante hiérarchique à partir de la matrice D.

9) Représenter le dendrogramme.

10) Ecrire une fonction permettant d'effectuer les étapes 7 à 9 précédentes à partir d'un tableau individus-variables.

Partie SAS

II. Modèle de classement via l'analyse discriminante

L'objectif de cette deuxième partie est construire un modèle pour discriminer les stations selon le label ski France.

On considère maintenant le tableau contenant toutes les composantes principales.

1) Partitionner les données en deux échantillons :
   - Un échantillon d'apprentissage sur lequel vous construirez le modèle (117 observations) : vous sélectionnerez les 117 observations aléatoirement parmi les 167 mais en prenant soin de conserver la proportion dans chaque classe.
   - Un échantillon de test sur lequel vous testerez le modèle (50 observations) : les 50 observations restantes sont le complémentaire de l'échantillon d'apprentissage.

2) Construire un modèle issu d'une analyse discriminante à partir de la base d'apprentissage
3) Calculer le pourcentage de bonne classification de la base d'apprentissage ainsi que de la base de test.
1) Facultatif : Reprendre les questions 1 à 3 en utilisant les variables initiales. Que constatez-vous ?