Projet en Statistique descriptive

Deadline : 17 mai 2024

Description du projet

Une ONG environnementale cherche à évaluer l'implication de plusieurs pays du monde dans la lutte contre le réchauffement climatique, et en particulier l'utilisation des énergies renouvelables dans leur production d'électricité. Elle dispose du jeu de données "renew.txt" contenant les variables suivantes :

- Pays
- Année (de 2000 à 2019)
- Pourcentage de la population qui a accès à l'électricité
- Pourcentage d'énergies renouvelables dans la consommation totale d'énergie
- Électricité produite à l'aide d'énergie fossile (en TeraWatt-heures)
- Électricité produite à l'aide des énergies renouvelables (en TeraWatt-heures)
- Consommation totale d'énergie par personne (en KiloWatt-heures)
- Émissions de CO2 par personne (en Tonnes par personne)
- PIB par habitant ($ US)
- Continent

Pour expliquer et décortiquer ces données, l'ONG a décidé de faire appel à des statisticiens (en herbe). C'est vous, étudiants en 3MIC à l'INSA Toulouse, qui avez été choisis pour résumer et décrire ces données à l'ONG.

Consignes

I) Commencez par décrire l'ensemble du jeu de données, en précisant bien la nature de chaque variable.

II) Ensuite, menez des analyses uni- et bi-variées du jeu de données. Observez-vous des anomalies ? Peut-on supprimer certains individus ? Certaines variables sont-elles liées ? Une attention particulière sera portée sur le choix des représentations, et sur l'interprétation des résultats présentés.

III) Menez une analyse en composantes principales (ACP) sur les variables qui vous semblent pertinentes. En particulier, précisez bien, en argumentant, le type d'ACP que vous faîtes et définissez la matrice de travail correspondante (en posant bien toutes vos notations). Précisez, en justifiant, combien de composantes principales vous décidez de garder. Enfin, donnez une interprétation de chacune des composantes que vous gardez en vous basant sur des représentations graphiques que vous aurez bien définies.

IV) Enfin, proposez une classification non supervisée (clustering) des données afin de les regrouper en plusieurs classes homogènes. Précisez le nombre de classes choisi qui vous semble pertinent et décrivez chaque classe. Vous pouvez comparer différentes méthodes de clustering, en justifiant le choix des paramètres utilisés.

Vous rendrez un rapport **par binôme** ou **trinôme** (du même groupe) au format **pdf** obtenu grâce à R Markdown, de 20 pages **maximum** (avec les graphes). Ce rapport doit être intitulé **gpX-Nom1-Nom2-Nom3.pdf** où **X** est à remplacer par votre groupe (A ou B). Pensez à laisser toutes les commandes R visibles dans votre rapport (option `echo = TRUE` dans les balises R).
Le rapport est à déposer **sur Moodle** au plus tard le **17 mai 2024** (aucun retour mail ne sera accepté).

Remarques : Gardez en tête qu'un des objectifs principaux de la statistique descriptive est de synthétiser l'information. Ne mettez en aucun cas des sorties R sans commentaire : si vous n'interprétez pas les résultats, autant ne pas les afficher. Enfin, le nombre de pages étant limité, n'utilisez que les outils ou méthodes qui vous semblent les plus pertinents.

Modalités d'évaluation

Vous serez évalués sur la présentation et la rédaction du rapport, sur la pertinence des choix des représentations (à argumenter) ainsi que sur l'interprétation des différentes sorties obtenues (graphiques ou autres). Vous serez également évalués sur la manipulation de R et de RMarkdown (pensez à laisser le code visible dans le rapport). Plus précisément, vous serez évalués sur les compétences suivantes.

Compétences transversales

- *Rédaction :* Savoir mener un argumentaire clair et concis. Savoir justifier un raisonnement. Penser à définir toutes notations utilisées.
- *Modélisation :* Savoir modéliser une situation :
  - Identifier la nature des variables (qualitative nominale/ordinale, quantitative discrète/continue).
  - Définir un modèle statistique.
- *Calcul :* Maîtriser les outils de calcul (algébrique, matriciel, intégral, différentiel, etc).
- *Logiciel R :* Savoir mener l'étude d'un jeu de données grâce à R et écrire un rapport en RMarkdown.

Statistique descriptive

Partie 1 : Statistiques descriptives unidimensionnelle et bidimensionnelle

- Maîtriser les définitions des indicateurs usuels de statistique descriptive (moyenne, mode, variance, quantiles, fonction de répartition empirique, covariance, corrélation...).
- Savoir choisir les indicateurs et représentations adaptés aux données.
- Savoir mener une interprétation des graphiques usuels de statistique descriptive (histogrammes, boxplots, barplots, diagramme en secteur, matrice de corrélation, mosaicplot,...).

Partie 2 : Analyse en composantes principales (ACP)

- Maîtriser le vocabulaire de l'ACP : inertie, inertie axiale, axes principaux, composantes principales, plan factoriel.
- Maîtriser les spécificités de l'ACP centrée et l'ACP centrée réduite.
- Maîtriser le principe de l'ACP :
  - Diagonalisation de la matrice ΓM.
  - Lien entre les valeurs propres et inerties axiales.
  - Lien entre les vecteurs propres et les axes factoriels.
- Maîtriser la définition des graphiques issus de l'ACP :
  - Projection des individus sur un plan factoriel.
  - Corrélations des variables avec les composantes principales.
- Savoir mener une interprétation des graphiques issus de l'ACP :
  - Interprétation individuelle de chaque graphique.
  - Interprétation croisée des différents graphiques.

Partie 3 : Classification non supervisée (clustering)

- Connaître et savoir appliquer les différentes méthodes de clustering (kmeans, DBSCAN, CAH) et leurs variantes.
- Savoir calibrer les paramètres et choisir le nombre de classes d'une méthode de clustering, à l'aide de différents critères (R^2, Calinski-Harabasz, Silhouette...).
- Savoir interpréter les classes données par une méthode de clustering.