Projet en Statistique descriptive

Pour le vendredi 19 mars 2021

Description du projet

Un client vous demande de faire une analyse descriptive du jeu de données `spotify-3MIC.txt` disponible sur la page Moodle de l'UF. Ce jeu de données comprend des caractéristiques de 10 000 morceaux provenant de l'API Web Spotify [1]. Il a été extrait du jeu de données *Spotify Dataset 1921-2020, 160+ Tracks* disponible sur kaggle. Chaque morceau est décrit par les variables suivantes :

- `year` : année de sortie du morceau,
- `acousticness` : métrique relative interne de l'acoustique morceau,
- `duration` : durée du morceau en millisecondes (ms),
- `energy` : métrique relative interne de l'intensité, des rythmes du morceau (rapide, fort, bruyant),
- `explicit` : vaut 1 si le morceau contient des vulgarités, et 0 sinon,
- `key` : tonalité en début de morceau (en notation anglaise, "A" correspondant à La, "Bb" à La ♯ ou Si ♭, "B" à Si, etc),
- `liveness` : proportion du morceau où l'on entend un public (en live),
- `loudness` : mesure relative du volume du morceau (en décibels, dB),
- `mode` : mode du morceau (0 si la tonalité est mineure, et 1 si la tonalité est majeure),
- `tempo` : le tempo du morceau, en battement par minute (bpm),
- `pop.class` : la popularité du morceau ("A" pour très populaire, à "D" pour "pas populaire").

Dans un premier temps, commencez par décrire l'ensemble du jeu de données, en précisant bien la nature de chaque variable.

Dans un deuxième temps, menez des analyses uni- et bi-dimensionnelles du jeu de données. Observez-vous des anomalies ? Une attention particulière sera portée sur le choix des représentations, et sur l'interprétation des résultats présentés.

Enfin, menez un 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).

Expliquez brièvement au client le principe de l'ACP, et précisez combien de composantes principales vous lui conseillez de garder.

Enfin, pour chaque graphe, précisez ce qui est représenté et interprétez les résultats.

Remarques : Gardez en tête qu'un des objectifs principaux de la statistique descriptive est de synthétiser l'information. De plus, ne mettez en aucun cas des sorties R sans commentaire : si vous n'interprétez pas les résultats, autant ne pas les afficher.

1. Vous pourrez trouver plus d'informations sur les variables ici : http://developer.spotify.com/documentation/web-api/reference/#endpoint-get-track

Consignes

Vous rendrez un rapport **par trinôme** (du même groupe) au format **pdf** obtenu grâce à R Markdown, de 15 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** le **vendredi 19 mars 2021 à 12h00** au plus tard (aucun retour mail ne sera accepté).

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). Plus précisément, vous serez évalués sur les compétences suivantes.

Compétences transversales

- Savoir mener un argumentaire clair et concis.
- Savoir mener l'étude d'un jeu de données grâce au logiciel R.

Partie 1 : Statistiques descriptives unidimensionnelle et bidimensionnelle

- Savoir identifier la nature des variables.
- Maitriser les définitions des indicateurs usuels de statistique descriptive.
- Savoir choisir les indicateurs et représentations adaptés aux données.
- Savoir mener une interprétation des graphiques usuels.

Partie 2 : Analyse en composantes principales (ACP)

- Maitriser le vocabulaire de l'ACP.
- Maitriser les spécificités de l'ACP centrée et l'ACP centrée réduite.
- Maitriser le principe de l'ACP.
- Maitriser la définition des graphiques issus de l'ACP.
- Savoir mener une interprétation des graphiques issus de l'ACP.