TP Noté
A déposer aux formats .ipynb et .pdf sur moodle jusqu'au mardi 28 mars 17h

Le rapport de TP sous forme d'un fichier .ipynb et d'un fichier .pdf sera déposé sur moodle. Ces fichiers doivent contenir tous vos codes, vos sorties graphiques et/ou numériques si pertinentes et les commentaires de vos sorties. Une trame pré-remplie est disponible sur moodle. Vous pouvez l'utiliser et rajouter autant de cellules que vous le souhaitez, que ce soient des cellules de codes ou de texte (Markdown).

La salle informatique 36.209 est à votre libre disposition pour travailler sur ordinateur tous les soirs de la semaine, à partir de 16h45 ou 18h30 selon les jours.

Le but de ce TP est de faire une analyse descriptive d'un jeu de données des résultats de l'épreuve d'heptathlon féminin aux Jeux Olympiques de Tokyo en 2020. L'heptathlon est une combinaison de 7 épreuves d'athlétisme : 100m haies, saut en hauteur, lancer de poids, 200m, saut en longueur, lancer de javelot et 800m. Pour calculer le score et le classement final, les temps ou longueurs obtenus sont convertis en points. Le fichier hepta.csv contient les points obtenus à chaque épreuve pour chaque participante ainsi que le total.

Récupérez le fichier `hepta.csv` sur la page moodle du cours et sauvegardez-le dans votre répertoire de travail. Vous aurez besoin des librairies suivantes.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
```

Importez le jeu de données via la commande

```
hepta = pd.read_csv('hepta.csv', sep=';',decimal=',')
```

Description du jeu de données

1. Examiner le jeu de données. Combien y a-t-il d'individus ? Combien de variables quantitatives ?
2. Représenter graphiquement dans la même fenêtre les boîtes à moustaches des 7 épreuves de l'heptathlon.
3. Quelle épreuve a l'étendue maximale ?

4. Quelle épreuve a la plus grande moyenne ? la plus grande variance ?

Analyse de la liaison entre les épreuves de Lancer de Javelot et du 800m

5. Extraire du tableau de données les variables Javelot et 800m
6. Calculer les coefficients de la droite de régression du score au 800m en fonction du score au Javelot.
7. Tracer le nuage des points du score au 800m en fonction du score au Javelot et superposer la droite de régression
8. Evaluer la qualité de l'ajustement et commenter.

Analyse en composante principale

9. Extraire le tableau des score des 7 épreuves (sans le score total), centrer et réduire les données et lancer l'ACP.
10. Examiner les valeurs propres. Quel est le pourcentage de variance expliquée par les deux premiers axes ? Combien d'axes faudrait-il garder pour expliquer 90% de la variabilité des données ?

Dans toute la suite on ne gardera que les deux premiers axes.

11. Extraire les facteurs.
12. Tracez le cercle des corrélation dans le premier plan factoriel pour les variables.

Examinez le cercle des corrélations dans le premier plan factoriel des variables.

13. Commentez les positions respectives des variables 800m et Javelot dans ce plan.
14. Commentez les positions respectives des variables 800m et 100mhaies dans ce plan.
15. Les variables Poids et Hauteur sont-elles bien représentées dans ce plan ?
16. Extraire les composantes principales.
17. Tracer le nuage de point des individus projetés dans le premier plan factoriel.

Examiner la projection des individus dans le premier plan factoriel.

18. Commentez la position de Nafissatou Thiam dans le premier plan factori
19. Commentez la position d'Ekaterina Voronina, Xenia Krizsan et Vanessa Rimm dans le premier plan factoriel.
20. Commentez la position d'Anouk Vetter dans le premier plan factoriel.

Quand vous avez terminé, sauvegardez votre fichier au format .ipynb et exportez-le au format .pdf. Déposez les deux fichiers sur moodle.