Livrable n°3 : Statistiques

Dans le cadre de la partie Statistique de la SAé 2.04, vous avez un livrable à fournir (livrable n°3).
Ce travail est à réaliser en binôme, **le même binôme que pour les précédents livrables de cette SAé**.

Objectifs

Le but de ce livrable est de réaliser une analyse du même type que celle faite dans le TP3 sur les données de la vue Colleges.csv. Vous devrez :

- **proposer une problématique** (sérieuse ou absurde) pouvant être traitée à partir de données extraites de la Base Colleges,
- **créer une vue** pouvant répondre à la problématique. Cette vue devra comprendre **au moins 5 variables**, et un **effectif total d'au moins 20**.
- puis **utiliser la régression linéaire multiple** pour répondre à la problématique choisie.

Dates

Vous réaliserez ce travail sur la **semaine du 10/06 au 14/06** où vous disposez pour cela de

- une séance d'1h de SAé encadrée par votre enseignant.e de BDD (avancée au 7/06 pour les 1E),
- 1h de SAé encadrée par votre enseignant.e de Statistique,
- et une séance de 2h de travail en autonomie.

Vous devrez déposer vos travaux sur Moodle avant le

vendredi 14/06 à 23h59

A rendre

Vous devrez déposer sur Moodle :

- Le rapport (environ 6 pages) en .pdf : voir les consignes ci-dessous.
- Le fichier .csv contenant la Vue extraite de la Base de Données.
- Le fichier Python des commandes qui vous ont permis d'obtenir vos résultats.

Critères d'évaluation

- Le suivi des consignes (voir ci-dessous) pour le rapport : il y a des points attribués pour chaque partie et sous-partie.
- Les sous-parties indiquées comme *+ difficile* seront au total sur 3 points (vus comme les points au-dessus de 17).
- La qualité/exactitude des raisonnements/interprétations statistiques,
- La qualité/exactitude du code Python,
- Les commentaires/explications du code (dans le code ou dans le rapport),
- Dans votre code Python, les noms des variables désignant des np.array, des DataFrame ou des listes devront se terminer par un suffixe qui indique le type de la variable : Ar pour les numpy.array, DF pour les DataFrame, Li pour les listes.

Il ne sera pas pris en compte dans l'évaluation le fait que les variables choisies soient ou non bien corrélées (vous ne pouvez pas le savoir à l'avance), mais seulement l'analyse des résultats.

Consignes : contenu du rapport

Voici le détail de ce qui devra apparaitre dans votre rapport :

1. **Les données - Problématique.** Vous commencez par une partie qui présente votre vue et votre problématique. Vous pourrez utiliser comme modèle de présentation : la première page du TP2 sur les Sangliers (à l'exception de l'application de la régression linéaire multiple, plus loin dans le rapport) ou la première page du TP3 sur les données Colleges. Précisément :
   (a) Présentation des données contenues dans votre vue :
      - la population,
      - description de chacune des variables choisies,
      - capture d'écran des premières lignes de votre fichier (avec les noms des colonnes)
   (b) Enoncer une problématique en lien avec ces données, en précisant quelle variable vous souhaitez expliquer (la future variable endogène).

2. **Import des données, mises en forme, centrage-réduction.** Dans cette partie, expliquer et montrer les captures d'écran du code qui vous a permis :
   (a) d'importer vos données .csv en Python,
   (b) de régler d'éventuels problèmes de mise en forme (problèmes de type, cases vides, etc.),
   (c) de centrer et réduire vos données.

3. **a. Exploration des données par représentations graphiques.** Dans cette partie, vous devez présenter vos premières approches des données que vous avez choisies, qui permettent de s'en faire une première idée. Cela peut être :
      - des diagrammes en batons (voir questions 9 et 10 du TP3),
      - ou des boites à moustache (voir questions 11 à 15 du TP3),

- *+ difficile.* de "simples" calculs de moyennes et de variance, pour des sous-groupes de votre population comme dans les questions 26 à 28 du TP3.

Vous devrez commenter/interpréter les graphiques/résultats présentés.

3. **b Exploration des données avec la matrice de covariance.** Comme dans la partie 3b du TP3 sur Colleges.csv, affiner le choix de vos variables explicatives grâce à la matrice de covariance :
   (a) Expliquer la démarche et montrer des captures d'écran du code qui permet de calculer la matrice de covariance correspondant à vos données.
   (b) Montrer une capture d'écran de votre matrice de covariance (de préférence grâce à l'onglet Variable Explorer de Spyder).

4. **Régression linéaire multiple.**
   (a) Expliquer comment vous pensez appliquer la régression linéaire multiple : choix de la variable endogène et des variables explicatives.
   (b) Préciser les raisons du choix des variables explicatives, à partir de la matrice de covariance.
   (c) Expliquer en quoi la régression linéaire multiple avec ces choix de variables permettra de répondre à la problématique que vous avez énoncée dans la partie 1(b).
   (d) Expliquer et montrer les captures d'écran du code pour calculer les paramètres de votre régression linéaire multiple.
   + difficile. Si vous avez fait la partie *pour les plus rapides* du TP2 sur les Sangliers, faites la régression linéaire multiple matriciellement, expliquer et montrer les captures de votre code.
   (e) Donner les paramètres obtenus, et interprétez-les en détail.
   (f) Calculer le coefficient de corrélation multiple de votre régression avec la fonction de `sklearn`. Interprétez-le.
   + difficile. Calculer le coefficient de corrélation multiple avec la formule du Cours-TP2, et montrer les captures d'écran de votre code.

5. **Conclusion.** Dans cette partie, vous faites le lien avec la problématique initiale :
   (a) Rappeler la problématique et proposer une réponse.
   (b) Justifier votre réponse à l'aide des paramètres/coefficient de corrélation obtenus. Donner toutes les informations que vos calculs permettent d'obtenir en lien avec la problématique.
   (c) Proposer des interprétations personnelles de vos résultats (sérieuses ou absurdes).