SAé 2.04 - 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).
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 Vue.csv.

Conditions

Ce travail est à réaliser en binôme, **le même binôme que pour les précédents livrables de cette SAé**.
Vous réaliserez ce travail sur la **semaine du 16/06 au 20/06**.
Vous devrez déposer vos travaux sur Moodle avant le

vendredi 20/06 à 23h59

Etapes

1h de SAé encadrée par votre enseignant.e de Statistique	**Proposer une problématique** (sérieuse ou absurde) pouvant être traitée à partir de données extraites de la Base de données de la SAé. **Choix de la vue** à concevoir. Elle devra comprendre **au moins 5 variables**, et un **effectif total** ≥ 20.
1h de SAé encadrée par votre enseignant.e de BD	**Réalisation de la Vue choisie**
2h de SAé en autonomie	**Utiliser la matrice des corrélations et la régression linéaire multiple** pour répondre à la problématique choisie. Voir consignes ci-après.

A rendre

Vous devrez déposer sur Moodle :

- Le rapport en .pdf : voir les consignes ci-dessous.
- Le fichier .csv contenant votre 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 4 points (vus comme les points au-dessus de 16).
- 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),

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 pour le rapport

Consignes sur la forme du rapport

- Votre rapport doit contenir les captures d'écran de votre code en Python avec des explications de ce code (soit en commentaires dans le code, soit à côté des captures d'écran). Le.a correcteurice ne doit avoir à ouvrir votre fichier .py que en cas de doutes/problèmes.
- 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.
- Les variables normalisées devront être nommées avec un suffixe N.

Consignes sur le fond/contenu du rapport

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

0. **Les données - Problématique.** Vous commencez par une partie qui présente votre vue et votre problématique. Vous pourrez utiliser comme modèle de présentation : les 2 premières pages du TP3 sur les notes des étudiant.es de DUT Info 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).

1. **Import des données, mises en forme, normalisation.** 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 normaliser vos données.

2. **Représentations graphiques.** Dans cette partie, vous devez illustrer les corrélations que vous voyez apparaitre (ou non) dans la suite, en montrant quelques représentations graphiques de vos données. Cela peut être :
   - des diagrammes en batons,
   - ou des boites à moustache.

   Vous devrez commenter/interpréter les graphiques présentés.

3. **Exploration des données avec la matrice de coefficients de corrélation.** Comme dans la partie 2 du TP3 des Statistiques pour la SAé, affiner le choix de vos variables explicatives grâce à la matrice des coefficients de corrélation :
   (a) Expliquer la démarche et montrer des captures d'écran du code qui permet de calculer la matrice des corrélations 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). Cibler les lignes ou colonnes qui montrent les corrélations qui vous intéressent.

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 des coefficients de corrélation.
   (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`.

   + *difficile.* Calculer le coefficient de corrélation multiple avec la formule vue dans le Cours 1 des Statistiques pour la SAé (dans sa partie 6), et montrer les captures d'écran de votre code.

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 et/ou absurdes).