Livrable n°3 : Statistiques

Dans le cadre de la partie Statistiques de la SAé 2.04, vous avez un livrable à fournir (livrable n°3), qui porte sur la Régression Linéaire Multiple.
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 TD2 sur les Sangliers et sur les données de la vue Parcoursup1.csv. Vous devrez :

- **proposer une problématique** (sérieuse ou absurde) pouvant être traitée à partir de données extraites de la Base Parcoursup,
- **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 5/06 au 9/06** (sauf pour les 1C : du , où vous disposez pour cela d'une séance d'1h de travail en autonomie et de la séance d'1h de projet encadré par votre enseignant.e de Base de Données.

Vous devrez déposer vos travaux sur Moodle avant le

vendredi 16/06 à 23h59 pour les A,B,D et E
lundi 19/06 à 23h59 pour les C

A rendre

Vous devrez déposer sur Moodle :

- Le rapport (entre 4 et 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.
- La qualité/exactitude du code Python,
- La qualité/exactitude des raisonnements/intérprétations statistiques,
- Les commentaires/explications du code (dans le code ou 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

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. La section 0 des TD2 sur les Sangliers et sur Parcoursup1 peut servir de modèle.

   Précisément :
   (a) Présentation des données contenues dans votre vue : la population et les variables choisies.
   (b) Enoncer une problématique en lien avec ces données.
   (c) Expliquer comment vous pensez appliquer la régression linéaire multiple : choix de la variable endogène, variables explicatives.
   (d) Expliquer en quoi la régression linéaire multiple avec ces choix de variables permettra de répondre à votre problématique.

1. **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.

2. **Choix des variables explicatives.** Comme dans la section 2 du TD2 sur Parcoursup1.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).
   (c) Conserver les 3 variables explicatives les plus pertinentes, indiquer lesquelles et pourquoi.

3. **Régression linéaire multiple.**
   (a) Expliquer et montrer les captures d'écran du code qui vous permet d'obtenir les paramètres de votre régression linéaire multiple.

   Bonus : si vous avez fait la partie *pour les plus rapides* du TD2 sur les Sangliers, faites la régression linéaire multiple matriciellement, expliquer et montrer les captures d'écran de votre code.
   (b) Donner les paramètres obtenus, et interprétez-les en détail.
   (c) Calculer le coefficient de corrélation multiple de votre régression. Interprétez-le.

4. **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).