Projet final - Logiciel statistique R

2023-07-09

Le projet est individuel. Vous devrez créer un dossier nomme **Nom\_Prenom\_projet R\_ ENSAE 2023** dans votre compte github. Seul le lien de votre dossier sur le github sera envoyé par email à [aboubacarhema94@gmail.com](mailto:aboubacarhema94@gmail.com) avant le 24 juillet 2023. L’objet email doit être **Nom Prenom Projet R ENSAE 2023**.

# Partie 1

L’objectif de ce projet est que vous appliquiez les outils que vous avez étudiés dans le cours du logiciel statistique R, dans le cas d’une étude de cas réelle. Le devoir est à faire seul et à rendre au format **.docx** ou **.pdf**. Les codes que vous utiliserez pour répondre aux questions seront à intégrer dans le corps de votre rapport. C’est pourquoi l’utilisation de **R Markdown/Quarto** est à privilégier, mais vous pouvez également utiliser d’autres éditeurs de textes. . . En ce qui concerne l’organisation du travail à rendre, vous vous inspirerez de la façon dont est organisé le sujet du projet (vous n’êtes pas obliger de recopier l’énoncé). Le barème est indiqué en face de chaque question. Il tient compte du code proposé et des commentaires sur les résultats obtenus, nous tiendrons compte des commentaires que vous aurez fait dans la notation.

**IMPORTANT** : il est demandé de commenter vos lignes de code. Vous pourriez être pénalisés si vous ne commentez pas vos codes.

**Objectif de l’étude**

Cette enquête vise à identifier et à caractériser des bioénergies durables pour les petites et moyennes entreprises (PME) agroalimentaires d’Afrique de l’Ouest.

## 1 Préparation des données

### 1.1 Description

Le fichier **Base\_Partie1.xlsx** contient 250 observations et 33 variables.  
La première colonne **key** correspond à l’identifiant de la PME. Les variables sont les suivantes :

* *q1*: Région
* *q2*: Département
* *q23*: Sexe du dirigeant/responsable de la PME
* *q24*: Age du dirigeant/responsable de la PME
* *q24a\_1*=1 si le dirigeant/propriétaire de la PME parle le français
* *q24a\_2*=1 si le dirigeant/propriétaire de la PME parle le wolof
* *q24a\_3*=1 si le dirigeant/propriétaire de la PME parle le Diola
* *q24a\_4*=1 si le dirigeant/propriétaire de la PME parle le Serere
* *q24a\_5*=1 si le dirigeant/propriétaire de la PME parle le Peul
* *q24a\_6*=1 si le dirigeant/propriétaire de la PME parle le Mandingue
* *q24a\_7*=1 si le dirigeant/propriétaire de la PME parle le Balante
* *q24a\_9*=1 si le dirigeant/propriétaire de la PME parle le Bambara
* *q24a\_10*=1 si le dirigeant/propriétaire de la PME parle une autre langue
* *q25*: Niveau d’instruction du dirigeant/responsable de la PME
* *q26*: Nombre d’années d’experience professionnelle du dirigeant/responsable de la PME dans l’entreprise
* *q12*: Statut juridique (SARL, SA, SUARL, GIE, Association, Groupement,)
* *q14b*: Autorisation de fabrication et de mise en vente (FRA)
* *q16*: L’entreprise est-elle désservie par une route bitumée ?
* *q17*: Etat de la route bitumée
* *q19*: l’état de la piste qui mène à l’entreprise
* *q20*: Avez-vous des associés dans l’entreprise ?
* *filiere\_1*=1 si la PME est dans la filière **arachide**
* *filiere\_2*=1 si la PME est dans la filière **anacarde**
* *filiere\_3*=1 si la PME est dans la filière **mangue**
* *filiere\_4*=1 si la PME est dans la filière **riz**
* *q8*: Activité principale de l’entreprise :
* *q81*: propriétaire ou locataire
* *gps\_menlatitude*: coordonnées géographiques de la PME (latitude)
* *gps\_menlongitude*: coordonnées géographiques de la PME (longitude)
* *submissiondate*: la date de soumission des informations de la PME
* *start*: la date de début de l’enrégistrement des informations de la PME par l’enquêteur
* *today*: la date de l’enquête

## 1.2 Importation et mise en forme

* Importer la base de données dans un objet de type **data.frame** nommé **projet**
* Selection les variables mentionnees dans la section description.
* Faites un tableau qui resume les valeurs manquantes par variable
* Vérifier s’il y a des valeurs manquantes pour la variable **key** dans la base **projet**. Si oui, identifier la (ou les) PME concernée(s).

# 1.3 Création de variables

* Rénommer la variable **q1** en **region**
* Rénommer la variable **q2** en **departement**
* Rénommer la variable **q23** en **sexe**
* Créer la variable **sexe\_2** qui vaut **1** si **sexe égale à Femme** et **0** sinon.
* Créer un **data.frame** nommé **langues** qui prend les variables **key** et les variables correspondantes décrites plus haut.

**Indication**: Vous remarquerez que ces variables commencent par **q24a\_**.

* Créer une variable **parle** qui est égale au nombre de langue parlée par le dirigeant de la PME.
* Sélectionnez uniquement les variables **key** et **parle**, l’objet de retour sera **langues**.
* Merger les **data.frame** **projet** et **langues**:

# 2 Analyses descriptives (10 pts)

Quelle est la répartion des PME suivant:

* le sexe?
* le niveau d’instruction?
* le statut juridique?
* le propriétaire/locataire?
* le statut juridique et le sexe?
* le niveau d’instruction et le sexe?
* Propriétaire/locataire suivant le sexe?

**Faite les statistiques descriptives de votre choix sur les autres variables**

**Recommandations** :

-1 nous tiendrons compte de la façon dont vous présenterez les résultats en sortie : la qualité du graphique, l’affichage des résultats, etc.

-2 nous privilegions l’utilisation du package gtsummary et la sortie doit etre un seul tableau.

**Indication**:

* Priorisez une analyse par **filière**
* Utilisez également des variables sur les dates(lubridate)

# 3 Un peu de cartographie

* Transformer le **data.frame** en données géographiques dont l’objet sera nommé **projet\_map**.
* Faites une réprésentation spatiale des PME suivant le sexe
* Faites une réprésentation spatiale des PME suivant le niveau d’instruction
* Faites une analyse spatiale de votre choix

# Partie 2

Le fichier excel **Base\_Partie 2.xlsx** contient un ensemble de données artificielles créé dans le cadre de ce projet. La première feuille contient des micro-données au niveau individuel des répondants à l’enquête, la deuxième feuille contient des données pour les districts dans lesquels les répondants de la première feuille ont été interrogés. La troisième feuille contient des explications sur les variables incluses dans toutes les feuilles précédentes.

## Nettoyage et gestion des données

* Renommer la variable “country\_destination” en “destination” et définir les valeurs négatives comme manquantes.
* Créer une nouvelle variable contenant des tranches d’âge de 5 ans en utilisant la variable “age”.
* Créer une nouvelle variable contenant le nombre d’entretiens réalisés par chaque agent recenseur.
* Créer une nouvelle variable qui affecte aléatoirement chaque répondant à un groupe de traitement (1) ou de controle (0).(fixer l’aléa)
* Fusionner la taille de la population de chaque district (feuille 2) avec l’ensemble de données (feuille 1) afin que toutes les personnes interrogées aient une valeur correspondante représentant la taille de la population du district dans lequel elles vivent.
* Calculer la durée de l’entretien et indiquer la durée moyenne de l’entretien par enquêteur.
* Renommez toutes les variables de l’ensemble de données en ajoutant le préfixe “endline\_” à l’aide d’une boucle. (ne pas faire une boucle for)

## Analyse et visualisation des données

* Créez un tableau récapitulatif contenant l’âge moyen et le nombre moyen d’enfants par district.
* Testez si la différence d’âge entre les sexes est statistiquement significative au niveau de 5 %.
* Créer un nuage de points de l’âge en fonction du nombre d’enfants(ggplot)
* La variable “intention” indique si les migrants potentiels ont l’intention de migrer sur une échelle de 1 à 7. Estimez l’effet de l’appartenance au groupe de traitement sur l’intention de migrer.
* Créez un tableau de régression avec 3 modèles. La variable de résultat est toujours “intention”. Modèle A : Modèle vide - Effet du traitement sur les intentions. Modèle B : Effet du traitement sur les intentions en tenant compte de l’âge et du sexe. Modèle C : Identique au modèle B mais en contrôlant le district. Les résultats des trois modèles doivent être affichés dans un seul tableau.

# Partie 3

Dans cette partie vous utiliserez la base **ACLED-Western\_Africa.csv**. C’est une base sur les evenements (violences politiques, marches, …. ) en Afrique de l’Ouest.

Faites une application r shiny qui permet:

* de visualiser les evenements par pays (le nombre d’evenement par pays dans une carte)
* de visualiser les evenements par pays, type, annee et la localisation.

Important: Seul le lien de l’application sera partage avec le prof. Les codes doivent etre dans votre compte github.