République du Sénégal

Un peuple - Un but - Une foi

****

Ministère de l'Economie, des Finances et du Plan

****

Agence Nationale de la Statistique et de la Démographie (ANSD)

****

Ecole Nationale de la Statistique et de l'Analyse Economique (ENSAE)

PROJET STATISTIQUE AVEC R

SYNTHÈSE DES EXPOSÉS

Redigé par :

IBRAHIM SOULEYMANE AMADOU

YANNICK BRYAN

AÏSSATA GUEYE

Elèves Ingénieurs Statisticiens Economistes

Sous la supervision de :

M. HEMA ABOUBACAR

Enseignant ENSAE

Juillet 2023

I GtSummary

• Description et importance

GtSummary est un package R qui permet de créer des tableaux statistiques. Son importance réside dans sa capacité à combiner plusieurs tableaux et à créer des tableaux personnalisés. Cela permet d'obtenir des résumés attrayants et simplifiés.

Pour agréger des tableaux, GtSummary utilise les fonctions **tbl_stack** (pour empiler les tableaux les uns au-dessus des autres) et **tbl_merge()** (pour les placer côte à côte).

La personnalisation des tableaux est possible grâce aux fonctions préfixées **theme_gtsummary_*()**, qui permettent de modifier l'apparence par défaut des tableaux.

À la fin d'un travail, il est possible d'exporter les tableaux créés à l'aide des packages gt ou flextable.

• Quelques autres fonctions utiles

- ☞ La fonction **'tbl_summary()'** permet de créer des résumés simples, principalement pour les variables qualitatives. En modifiant les paramètres de cette fonction, il est possible d'obtenir une grande flexibilité et d'afficher davantage d'informations. Par exemple, en utilisant le paramètre 'by()', il est possible de définir des groupes. Le paramètre 'percent' permet de calculer les profils par ligne ou par colonne. Le paramètre 'include' est utilisé pour sélectionner les variables à inclure dans le tableau, et le paramètre 'statistic()' permet de spécifier les statistiques à calculer.
- ☞ 'add_overall()' : Permet d'inclure un résumé global ou une statistique agrégée dans un tableau.
- ☞ 'add_difference()' : Utilisée pour comparer les moyennes entre les groupes dans un tableau.
- ☞ 'add_label()' : Permet d'ajouter des labels personnalisés à une variable dans un tableau.
- ☞ 'tbl_cross()' : Permet de créer des tableaux croisés.
- ☞ 'tbl_regression()' : Génère un tableau récapitulatif des résultats d'un modèle de régression.
- ☞ 'add_p()' : Ajoute les valeurs de p (p-values) dans un tableau.

II R_Shiny

• Fonctionnement et importance

C'est un package qui permet de créer des Dashboards et des applications Web interactives.
Il fonctionne selon un modèle ui-server.

- ☞ UI = User Interface : Il s'agit de l'interface utilisateur, qui correspond à la partie visible et interactive d'une application. Elle permet aux utilisateurs d'interagir avec l'application de manière visuelle et pratique.
- ☞ Server : Le serveur est responsable de l'exécution du code R et de la génération des sorties dynamiques. Il traite les demandes de l'interface utilisateur et renvoie les résultats appropriés en fonction des actions effectuées par l'utilisateur.

- **Les packages nécessaires**

- ☞ Le package shiny qui est indispensable pour programmer une application shiny.
- ☞ Le package shinydashborad : Il permet de créer des tableaux de bord interactifs avec une mise en page structurée.
- ☞ Le package DT : il permet de générer des tableaux personnalisés et interactifs.

- **Quelques fonctions de l'UI**

- ☞ fluidPage : Crée une page Shiny fluide qui s'adapte à la taille de la fenêtre du navigateur.
- ☞ titlePanel : Affiche un titre en haut de la page Shiny.
- ☞ selectInput : Crée un menu déroulant permettant à l'utilisateur de sélectionner une option parmi une liste.
- ☞ verbatimTextOutput : Affiche du texte généré par le serveur dans l'interface utilisateur
- ☞ Etc.

- **Quelques fonctions du server**

- ☞ renderText() : Cette fonction est utilisée pour générer et afficher du texte dynamique ;
- ☞ renderPlot() : Elle permet de générer et d'afficher des graphiques interactifs ;
- ☞ renderTable() : Cette fonction génère et affiche des tableaux interactifs.

III Cartographie sur R

- **Description et quelques concepts de bases**

Pour faire de la cartographie avec R, il est indispensable de connaître certains concepts de base :

- ☞ Le système de coordonnées de référence : il définit comment les coordonnées spatiales sont représentées dans un système de référence donné ;

- ☞ Types de fichier de donnée spatial : Les données spatiales sont généralement stockées sous forme de fichiers, qui peuvent être de deux types : les fichiers raster et les fichiers vecteur. Dans la catégorie des fichiers vecteur, on distingue trois types : les points, les lignes et les polygones. Les points peuvent représenter des mosquées, les lignes peuvent représenter des routes, et les polygones peuvent représenter des habitations. Les fichiers raster, quant à eux, sont généralement présentés sous forme d'images.
- ☞ Les données spatiales sont stockées sous forme de fichiers shapefile ou GeoJSON. Les fichier shapeFiles portent l'extension .shp

- **Les principaux packages**

  - ☞ Cartography : Permet de réaliser des cartes
  - ☞ Sp : pour l'ancienne classe d'objets spatiaux
  - ☞ Sf : Pour la nouvelle classe d'objets spatiaux
  - ☞ GISTools

- **Quelques fonctions importantes**

  - ☞ Pour créer un CRS on utilise la fonction CRS() du package sp.
  - ☞ coordinates() : Permet de passer d'un objet de classe data.frame, à un objet de classe "Spatial" sp.
  - ☞ readOGR() : Elle permet d'importer les types de données spatiales.
  - ☞ st_write() permet d'exporter de nombreux types de fichiers.
  - ☞ Etc.

IV R_Excel

- **Fonctionnement et importance**

Le package r2excel est un package R qui permet de créer, lire et formater facilement des fichiers Excel avec R. Il est très utile pour les utilisateurs de R qui ont besoin d'importer ou d'exporter des données à partir de fichiers Excel. Le package fournit des fonctions pour ajouter des tableaux, des graphiques, des hyperliens et des en-têtes aux fichiers Excel. Il est également possible de formater les cellules et les feuilles de calcul Excel à l'aide de ce package.

- **Quelques fonctions utiles**

  - ☞ xlsx.addHeader pour ajouter des titres ;
  - ☞ xlsx.addPlot pour ajouter des graphiques

- ☞ xlsx.addParagraph pour ajouter des paragraphes de textes
- ☞ xlsx.addTable pour ajouter des data.frames
- ☞ xlsx.addLineBreak pour ajouter un saut de ligne
- ☞ Etc.

V Text Mining

• Description et importance

Le textmining consiste en l'utilisation des techniques statistiques dans le cadre de la programmation, pour le traitement d'un corpus ou texte, afin d'en ressortir des informations significatives. Il peut être utilisé pour ressortir un profilage de discussion sur les réseaux ou de centre d'intérêt lors de la navigation sur internet, des enquêtes de satisfactions, etc.
En guise d'exemple de ce qu'on peut faire avec du textmining : après un retraitement préalable du corpus ou du jeu de mots, on peut ressortir la fréquence des mots ou des bigrammes, la matrice terme des mots, le nuage de mots, le réseau des mots, etc. on peut également l'utilisé sur les messages et discussions des plateformes réseaux pour ressortir la chronologie des messages, le classement des identifiants(ID) selon le nombre de message envoyé, les mots les plus fréquents utilisés par chaque ID , comparaison des mots les plus fréquents dans l'ensemble. On peut faire des comparaisons entre groupes, et plus encore.

• Les principaux packages

Les packages nécessaires avec quelques fonctions essentielles pour faire du textmining sont :

- ☞ wordcloud : permet de faire le nuage de mots ;
- ☞ ggraph : permet de faire le réseau des mots.
- ☞ tidytext :

VI Résolution de système d'équations non linéaires avec R

Les systèmes d'équations nous permettent, à partir de plusieurs expressions mettant en relation un certain nombre de variables, de déterminer les valeurs qu'elles doivent prendre dans un domaine précis, pour les vérifier simultanément. Pour un nombre relativement réduit d'équations linéaires, l'Homme peut s'essayer à résoudre un tel système à la main. Cependant, pour un nombre d'équations relativement grand, à plus forte raison lorsqu'elles sont non linéaires, l'Homme peut faire recours à l'outil informatique ; et s'il se trouve dans

le cadre de la programmation sur R, il existe des packages, contenant une multitude de fonctions, qui lui permettent d'obtenir les résultats très rapidement.

- **Méthode directe**

  ☞ rootSolve : :multiroot() : Cette fonction principale du package est utilisée pour résoudre les systèmes d'équations non linéaires multivariées. Elle prend en entrée le vecteur des équations multivariée qui constitue le système, ainsi qu'un vecteur initial de valeurs approchées pour les variables, et retourne une approximation des valeurs des variables qui satisfont le système d'équations.

  ☞ Library(nleqslv) : C'est une bibliothèque de fonctions disponible dans le logiciel R qui permet de résoudre des systèmes d'équations non linéaires multivariées en utilisant la méthode itérative de Newton et broyden . La principale fonction utilisé est nleqslv : :nleqslv(), qui prend en argument les équations multivariée qui constitue le système, un vecteur initial et retourne une approximation des valeurs des variables qui satisfont le système d'équations. Dans la syntaxe de cette fonction, on précise Méthode= 'Newton'.

- **Méthode indirecte**

On peut aussi résoudre un système d'équations non linéaires à l'aide d'un problème d'optimisation. Pour cela, on transforme le système d'équations en un problème d'optimisation en définissant une fonction objective à minimiser ou à maximiser. Cette fonction objective est généralement construite en utilisant une mesure de l'écart entre les valeurs réelles des équations et les valeurs calculées à partir des variables inconnues. R propose pour cela, les fonctions comme optim () dans le package stats, optimx () dans le package optimx.un cas pratique a été effectué pour appliquer ces fonctions afin de résoudre le système d'équation proposé par ce groupe.

VII Packages Janitor

Lorsque l'on commence à travailler avec une base de données, il est essentiel de s'assurer de sa qualité et de la nettoyer si nécessaire. Pour accomplir cette tâche, le package "janitor" est utilisé, offrant deux fonctions principales : l'exploration et le nettoyage de la base de données.

L'exploration permet de parcourir la base de données afin de détecter d'éventuels doublons, valeurs manquantes ou autres anomalies. Janitor joue ainsi un rôle crucial dans l'analyse approfondie des données, en vérifiant leur organisation et la cohérence des variables selon une nomenclature appropriée, par exemple.

Si l'exploration révèle la présence de l'un de ces défauts ou plusieurs, la fonction de nettoyage de Janitor intervient pour les corriger.

- **Quelques fonctions**

Parmi les fonctions de nettoyage disponibles dans le package Janitor, nous avons sélectionné les suivantes :

- ☞ 'clean_names()' : Cette fonction analyse les noms des variables et les nettoie des caractères indésirés. Vous pouvez également spécifier les noms des variables à ne pas modifier en utilisant 'clean_names(., 'var1', 'var2', ...)'. Le symbole '%>%' permet de composer des fonctions.
- ☞ 'make_clean_names()' : Cette fonction permet une utilisation plus générale de la fonctionnalité de 'clean_names()' en l'appliquant aux vecteurs.
- ☞ 'compare_df_cols()' : Lorsque les colonnes diffèrent ou que les classes de colonnes ne correspondent pas entre deux dataframes, cette fonction de Janitor permet de les comparer là où dplyr échoue.
- ☞ 'return_mismatch()' : Cette fonction permet de nettoyer les données en supprimant les lignes ou les variables vides ou comportant des valeurs manquantes (NA). Si moins de 10% des valeurs dans une variable sont des NA, la fonction 'remove_empty()' ne supprime pas cette variable. Il convient de noter que tous les NA ne sont pas des valeurs manquantes, car il peut y avoir des cas où un individu n'est pas censé avoir une réponse pour une variable donnée. Janitor ne traite pas ces cas, mais d'autres packages permettent de prendre en compte ces NA dans l'analyse en leur attribuant un poids.
- ☞ 'remove_constant()' : Cette fonction supprime les colonnes d'un dataframe qui ne contiennent que des constantes.
- ☞ 'get_dupes()' : Cette fonction permet de repérer les enregistrements ou les identifiants en double lors du nettoyage des données, ce qui est généralement une situation indésirable.

- **Limites**

Il faut faire appel à d'autres packages pour la mise en forme des tableaux (GT. Ou flextable).

VIII RMarkDown

RMarkdown est un langage de balisage qui permet de créer des documents dynamiques intégrant du code R, des résultats d'analyses et des éléments de texte. Voici quelques-unes de ses descriptions et de son importance :

- ☞ Facilité d'utilisation : RMarkdown utilise une syntaxe simple et intuitive, ce qui facilite la création de documents. Les utilisateurs peuvent combiner du texte narratif avec des morceaux de code R, des résultats d'exécution et des graphiques, le tout dans un seul document.
- ☞ Reproductibilité : RMarkdown permet de créer des documents reproductibles, car le code et les résultats sont directement intégrés dans le document final. Cela signifie que toute personne disposant du fichier RMarkdown peut reproduire les analyses, les résultats et les visualisations sans avoir à exécuter manuellement le code source.
- ☞ RMarkdown offre une grande flexibilité dans la création de documents. Il permet d'inclure différents types de contenu, tels que des équations mathématiques, des tableaux, des images et des liens hypertexte. De plus, il prend en charge plusieurs formats de sortie, notamment HTML, PDF, Word, présentations PowerPoint, ce qui permet de partager facilement les résultats de l'analyse.
- ☞ Automatisation : RMarkdown peut être utilisé pour automatiser la génération de rapports et de documents. En utilisant des scripts R, les analyses et les visualisations peuvent être mises à jour automatiquement lorsque de nouvelles données sont disponibles, ce qui facilite la production régulière de rapports.

En résumé, RMarkdown est un outil puissant pour la création de documents dynamiques, reproductibles et conviviaux, ce qui en fait un choix privilégié pour la communication et la diffusion des résultats d'analyses statistiques et de données.

IX R-Python

Il est possible de transcrire un travail réalisé entre R et Python en utilisant plusieurs packages tels que Reticulate, rPython, PythonInR et Jupyter Notebooks.

Reticulate est un package de R qui établit une connexion entre l'environnement R et l'environnement Python. Il offre les fonctionnalités suivantes :

- ☞ Importation de modules Python dans R.
- ☞ Appel de fonctions Python.
- ☞ Accès aux objets de R ou de Python.
- ☞ Conversion d'objets entre R et Python.

Pour exécuter du code Python, vous avez trois options :

- ☞ Dans un chunk R Markdown.
- ☞ À l'aide de la fonction 'py_run_string'.
- ☞ À l'aide de la fonction 'rpy2.robjects.r'.

Ces options vous permettent d'intégrer et d'exécuter du code Python dans votre environnement R, facilitant ainsi la combinaison des deux langages dans votre travail.

X QARTO

Le package Quarto est une nouvelle création pour R qui offre la possibilité de créer des fichiers en incluant des images, des bibliographies, des liens, etc. Il peut être considéré comme une version améliorée de Rmarkdown.
Il permet de générer une variété de formats de sortie tels que HTML, PDF, EPUB et Word, tout en offrant une flexibilité et une facilité d'utilisation accrues par rapport à d'autres packages similaires comme Rmarkdown.