Programmation Statistique sous R
EXAMEN
Documents autorisés

Antoine ROLLAND et Anthony SARDELLITTI

7 Avril 2022

Table des matières

CONSIGNES

Vous avez 1h30 pour réaliser l'ensemble du devoir. Vous devez vous connecter sur les sessions EXAM.

Voici les recommandations pour que votre travail soit pris en compte et pour ne pas avoir de pénalités :

- Le script R est à rendre sur la session EXAM
- Votre fichier doit se nommer **NOM_PRENOM** et doit être un script R
- Le rendu est individuel
- Pour chaque question n'oubliez pas :
  - d'écrire le code R permettant de répondre à la question
  - de préciser le numéro de la question et de l'exercice en commentaire.

Exemple :

```
# Ex 1-b
Data<-read.csv()
```

Présentation

Pour cet examen, on utilise le fichier `fastfood.csv` qui décrit les caractéristiques des produits que proposent des enseignes de restauration rapide. Les données sont issues du site OpenFoodFact. Dans le fichier, une ligne correspond à un produit.

Voici une description des données :

- `code` : le code du produt
- `product_name_fr` : le nom du produit
- `generic_name_fr` : le nom générique
- `quantity` : la quantité
- `packaging` : la liste des emballages
- `brands` : la marque de restauration rapide
- `categories` : la catégorie du produit
- `allergens` : la liste des allergènes contenus dans le produit
- `energy.kcal_value` : le nombre de calories
- `fat_value` : le niveau de matière grasse en g
- `saturated.fat_value` : le niveau de matière grasse saturée en g
- `sugars_val` : le niveau de sucre en g
- `proteins_value` : le niveau de protéine en g
- `salt_value` : le niveau de sel en gr
- `off.nutriscore_score` : le nutriscore du produit
- `Sandwich` : si le type du produit est un sandwich ou non.

Conseils

- Pensez à copier le dataset sur votre bureau
- N'ouvrez pas le fichier avec excel mais privilégiez le bloc note ou Notepad+
- Pensez à vous autocorriger avant de vous lancer dans la question suivante
- Pensez à sauvegarder votre script régulièrement
- Pour certaine question, la réponse est affichée pour vous aider

1 Exercice 1 : Importer les données

a. Importez le jeu de données `fastfood.csv` avec la fonction `read.csv( )` **uniquement** dans un objet appelé `df`.

b. Combien de produits sont présents dans ce dataset ?

```
## [1] 331
```

c. Affichez le nom des colonnes.

d. Affichez un résumé des données avec la fonction adaptée. On remarque qu'aucun tri à plat n'est effectué à cause du type des données des variables qualitatives. D'après le résultat de la fonction, il semble aussi que ce jeu de données présente quelques valeurs manquantes.

e. Il y a beaucoup de variables qualitatives. Construire une boucle qui parcourt chaque colonne du dataframe et qui transforme les colonnes en type `factor` si elles ne sont pas de type `numeric.`

f. Affichez le type des variables avec la fonction adaptée pour vérifier que la commande précédente est correcte.

2 Exercice 2 : Statistiques descriptives

Pour chaque question (sauf la e) de cet exercice, il est demandé d'ajouter une phrase d'interprétation en commentaire. Si des graphiques sont demandés, pensez à renseigner un titre et ajouter une phrase d'interprétation.

a. Combien y a-t-il de marques de fastfood différentes (`brands`)?

b. Calculez le nombre de produit par marque et représentez graphiquement le résultat dans un diagramme en barre.

c. Calculez la médiane et l'écart-type du nombre de calories pour l'ensemble des produits (`energy.kcal_value`)

d. Représentez cette dispersion dans un diagramme en boîte aussi appelé "boîte à moustache". Qu'observez-vous de particulier ?

e. Affichez les 2 produits avec le plus grand nombre de calories. Qu'est-ce qui vous saute aux yeux ?

	product_name_fr	brands	energy.kcal_value
93	Salade - Caesar	KFC	1018
74	BOXMASTER® MAXX ORIGINAL	KFC	910

e. En réalité, le nombre de calories de cette salade semble être faux. D'après d'autres sources, on observe qu'en réalité le nombre de calories de cette salade est de 268 kcal. Effectuez cette correction avec une commande R. Un filtre sur le nom du produit `Salade - Caesar` pourrait vous aider.

f. Calculez les **centiles** du nombre de calories. Quelle est le pourcentage de produit avec strictement moins de 200 kcal ?

3 Exercice 3 : Tris, filtres et agrégations

Pour chaque question si des graphiques sont demandés, pensez à renseigner un titre et ajouter une phrase d'interprétation.

a. Construisez un objet `requete_a` avec uniquement les produits correspondants à des sandwichs (`Sandwich`). Construire un graphique pour visualiser la dispersion du nombre de calories des produits pour comparer les marques entre-elles.

b. Construisez un objet `requete_b` avec la moyenne du nombre de calories des produits agrégée par marque. Quelle est en moyenne la marque avec les produits les plus caloriques ?

brands	energy.kcal_value
Burger King	418.1304
KFC	388.2105
McDonald's	311.0455
Quick	247.5000

c. Construisez un objet `requete_c` avec les 5 sandwichs les moins caloriques.

	product_name_fr	brands	energy.kcal_value
329	Ptit Italien	McDonald's	232
280	280 recette originale	McDonald's	239
239	Casse Croûte Jambon Fromage	McDonald's	240
240	Casse Croûte Bœuf Moutarde **	McDonald's	250
241	Casse Croûte Poulet Curry ***	McDonald's	250

d. Quelle est la part de valeurs manquantes sur le nutriscore (`off.nutriscore_score`) ?

```
##
##     FALSE      TRUE
## 0.5317221 0.4682779
```

4 Exercice 4 : Corrélation

a. Calculez le coefficient de corrélation entre le nutriscore et le nombre de calories pour les sandwichs. Quelle conclusion pouvez-vous en tirer ?

```
## [1] 0.1439669
```

b. Calculez la matrice de corrélation sur l'ensemble des informations sur l'apport nutritionnel (`energy.kcal_value`, `fat_value`, `saturated.fat_value`, `sugars_value`, `proteins_value`, `salt_value`)

	energy.kcal_value	fat_value	saturated.fat_value	sugars_value	proteins_value	salt_value
energy.kcal_value	1.00	0.89	0.73	0.05	0.80	0.75
fat_value	0.89	1.00	0.83	-0.11	0.65	0.59
saturated.fat_value	0.73	0.83	1.00	0.04	0.59	0.38
sugars_value	0.05	-0.11	0.04	1.00	-0.11	-0.04
proteins_value	0.80	0.65	0.59	-0.11	1.00	0.78
salt_value	0.75	0.59	0.38	-0.04	0.78	1.00

c. Après avoir identifié la corrélation la plus forte, construisez un nuage de points avec la représentation de ces deux variables. Personnalisez le graphique en affichant une couleur pour chaque marque.

5 Exercice 5 : Application

Un ami n'est pas très fan des frites et des sodas. Il ne prend donc jamais de menu dans les fastfoods. Pour se remplir la panse, il préfère prendre deux sandwichs. Il affirme que la moyenne du nombre de calories par sandwhich lui permet d'en prendre 2 sans dépasser le nombre de calories recommandées pour un repas de 900 kcal. Vous devez vérifier avec un échantillonnage si ce qu'il affirme est vrai.

a. Créer une fonction appelée `SimulSandwhichs()` qui prend en entrée le choix de la marque de restauration et qui retourne un vecteur avec les valeurs de nombre de calories de 5 sandwhichs au hasard. Voici un exemple du résultat attendu :

```
SimulSandwhichs('KFC')
```

```
## [1] 268 268 587 600 284
```

b. Répliquez cette expérience aléatoire 1000 fois avec des marques au hasard et stockez les échantillons dans un dataframe. Voici un exemple du résultat attendu :

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
543	119	268	298	189	341	297	284	256	256
270	239	284	543	290	245	825	910	268	268
543	25	274	270	431	258	395	200	180	557
23	431	587	298	45	250	266	436	268	256
543	893	910	304	266	189	297	53	180	698

c. Calculez la moyenne de chaque échantillon avec la fonction adaptée et représentez toutes ces moyennes dans un histogramme. Voici un exemple du résultat attendu :

```
##  [1] 384.4 341.4 464.6 342.6 244.2 256.6 416.0 376.6 230.4 407.0
```