DM Mise à niveau en R

Votre rendu doit être au format Rmarkdown (document Rmd qui doit absolument compiler + version pdf ou html) et doit être déposé sur Moodle au plus tard le 18 septembre. Si vous avez un problème d'accès à Moodle, vous pouvez m'envoyer votre rapport par mail : anna.bonnet@upmc.fr

Exercice 1 : Alimentation et changement climatique

Contexte

Provenance des données

Les données proviennent de l'étude Agribalyse qui référence l'impact environnemental des différents produits agricoles.

La récolte des données a été élaborée et validée dans le cadre de partenariats notamment avec l'ADEME et l'INRAE.

Les données sont publiques et téléchargeables à l'adresse suivante : https://doc.agribalyse.fr/

Description du jeu de données

La base de données contient des informations sur le type de produit (groupe d'aliment, sous-groupe d'aliment, composition etc) ainsi que différentes mesures de l'impact environnemental (émissions de $C0_2$, particules fines...). La description complète du jeu de données se trouve ici :

Dans cet exercice, nous proposons d'étudier un sous-ensemble de variables décrites ici : https://doc.agribalyse.fr/documentation/

1. Code CIQUAL (identifie un aliment)
2. Groupe d'aliments
3. Sous-groupe d'aliments
4. Nom du produit en français
5. Changement climatique provenant de l'agriculture (en kg C02/kg de produit)
6. Changement climatique provenant de la transformation (en kg C02/kg de produit)
7. Changement climatique provenant de l'emballage (en kg C02/kg de produit)
8. Changement climatique provenant du transport (en kg C02/kg de produit)
9. Changement climatique provenant de la distribution (en kg C02/kg de produit)
10. Changement climatique provenant de la consommation (en kg C02/kg de produit)
11. Changement climatique total (somme des 6 précédents) (en kg C02/kg de produit)

Remarques sur la construction du jeu de données

- Le calcul de l'impact d'un produit est effectué en moyenne sur tous les produits du même type. La notice nous donne l'exemple suivant : " L'impact d'une pizza Margherita « standard », constituée de tomates « standards » conventionnelles, de gruyère et de jambon standards « conventionnels », issus des systèmes de production majoritaires aujourd'hui, et d'emballages majoritaires observés pour ce type de produit. Les impacts de la « tomate standard conventionnelle » de la pizza représentent la moyenne pondérée des impacts de tomates majoritairement utilisés pour les produits transformés (c'est-à- dire 18 % des tomates issues de la production française, 46 % de tomates italiennes et 36 % de tomates espagnoles). "

- La notice pdf fournie sur le site https://doc.agribalyse.fr/ est très complète et renseigne sur le détail précis de la constitution de la base de données

Objectifs

Questions

On propose d'étudier :

- le lien entre le type de produit et les émissions de $C0_2$
- les parts d'émission de C02 dues à chaque étape du processus de commercialisation d'un aliment, en différenciant par type d'aliment

Outils utilisés

- tri et manipulation des données avec tidyverse
- Statistique descriptive et visualisation avec ggplot

packages R à installer

- tidyverse
- ggplot2
- readxl

Pré-traitement des données pour faciliter leur utilisation

Lancer le code suivant, qui est une étape de pré-traitement des données qui consiste à :

- renommer certaines colonnes spour améliorer la lisibilité
- sélectionner certaines variables contenues dans différentes feuilles du fichier excel global afin de créer un tableau **agri_cc** qui contient uniquement les variables décrites plus haut
- modifier une entrée dont code CIQUAL apparaissait en doublon

```
library(tidyverse) ## pour la mise en forme des données
library(readxl) ## lecture des données au format excel



agri <- read_excel("AGRIBALYSE3.1_produits alimentaires_2.xlsm", sheet = 2, col_names = TRUE, na = "", s
dim(agri)
```

```
## [1] 2517   29
```

```
head(agri)
```

```
## # A tibble: 6 x 29
##   `Code\r\nAGB` `Code\r\nCIQUAL` `Groupe d'aliment`       Sous-groupe d'alimen~1
##   <chr>         <chr>            <chr>                   <chr>
## 1 11084         11084            aides culinaires et ing~ algues
## 2 11023         11023            aides culinaires et ing~ herbes
## 3 11000         11000            aides culinaires et ing~ herbes
## 4 11093         11093            aides culinaires et ing~ herbes
## 5 20995         20995            aides culinaires et ing~ algues
## 6 20998         20998            aides culinaires et ing~ algues
## # i abbreviated name: 1: `Sous-groupe d'aliment`
## # i 25 more variables: `Nom du Produit en Français` <chr>, `LCI Name` <chr>,
```

```
## #   `code saison (0 : hors saison ; 1 : de saison ; 2 : mix de consommation FR)` <dbl>,
## #   `code avion (1 : par avion)` <dbl>, Livraison <chr>,
## #   `Matériau d'emballage` <chr>, Préparation <chr>,
## #   `DQR - Note de qualité de la donnée (1 excellente ; 5 très faible)` <dbl>,
## #   `mPt/kg de produit` <dbl>, `kg CO2 eq/kg de produit` <dbl>, ...
```

```
pour_nom <- read_excel("AGRIBALYSE3.1_produits alimentaires_2.xlsm", sheet = 2, col_names = TRUE, na = "
colnames(agri)[!grepl("\\.\\.\\.",colnames(pour_nom))] <- colnames(pour_nom)[!grepl("\\.\\.\\.",colnames
dim(agri)
```

```
## [1] 2517   29
```

```
agri_detail <- read_excel("AGRIBALYSE3.1_produits alimentaires_2.xlsm", sheet = 3, col_names = TRUE, na
colnames(agri)[1] <- "Code AGB"

agri_all_indic <- agri[,13:29]

## agri changement climatique :
## probleme de sauce au poivre 11212 !

agri_detail$`Code AGB`[agri_detail$`Code CIQUAL`=="11212"] <- "11212"

agri_cc <- full_join(agri[,-c(13:29)], agri_detail[,c(1,16:22)])

colnames(agri_cc) <- gsub("\\.\\.\\.[0-9][0-9]","", colnames(agri_cc))

dim(agri_cc)
```

```
## [1] 2517   19
```

```
head(agri_cc)
```

```
## # A tibble: 6 x 19
##   `Code AGB` `Code\r\nCIQUAL` `Groupe d'aliment`          Sous-groupe d'alimen~1
##   <chr>      <chr>            <chr>                       <chr>
## 1 11084      11084            aides culinaires et ingréd~ algues
## 2 11023      11023            aides culinaires et ingréd~ herbes
## 3 11000      11000            aides culinaires et ingréd~ herbes
## 4 11093      11093            aides culinaires et ingréd~ herbes
## 5 20995      20995            aides culinaires et ingréd~ algues
## 6 20998      20998            aides culinaires et ingréd~ algues
## # i abbreviated name: 1: `Sous-groupe d'aliment`
## # i 15 more variables: `Nom du Produit en Français` <chr>, `LCI Name` <chr>,
## #   `code saison (0 : hors saison ; 1 : de saison ; 2 : mix de consommation FR)` <dbl>,
## #   `code avion (1 : par avion)` <dbl>, Livraison <chr>,
## #   `Matériau d'emballage` <chr>, Préparation <chr>,
## #   `DQR - Note de qualité de la donnée (1 excellente ; 5 très faible)` <dbl>,
## #   Agriculture <dbl>, Transformation <dbl>, Emballage <dbl>, ...
```

```
### changement des noms de quelques modalités pour des raisons cosmétiques
agri_cc$`Groupe d'aliment`<-plyr::revalue(agri_cc$`Groupe d'aliment`,c("viandes, œufs, poissons"="viand
agri_cc$`Groupe d'aliment`<-plyr::revalue(agri_cc$`Groupe d'aliment`,c("fruits, légumes, légumineuses e
agri_cc$`Groupe d'aliment`<-plyr::revalue(agri_cc$`Groupe d'aliment`,c("aides culinaires et ingrédients
agri_cc$`Sous-groupe d'aliment`<-plyr::revalue(agri_cc$`Sous-groupe d'aliment`,c("œufs"="oeufs"))
agri_cc$`Sous-groupe d'aliment`<-plyr::revalue(agri_cc$`Sous-groupe d'aliment`,c("produits à base de po
```

Début de l'exercice

Type d'aliment et émissions de $C0_2$

1. Proposer une visualisation graphique permettant de comparer l'émission de $C0_2$ (en kg par kg de produit) liée aux différents aliments. Commenter.
2. Quels sont les deux groupes d'aliments au sein desquels il y a le plus de variabilité ? Pour ces deux groupes, refaire la même analyse qu'à la question précédente au niveau du sous-groupe d'aliments. Commenter.
3. Afficher le top 20 des produits qui émettent le plus de $C0_2$. Commenter.
4. (Bonus) Concentrons nous sur les émissions de $C0_2$ associées à la viande. Créer une nouvelle variable "type_viande" qui associe à chaque aliment (des sous-groupes "viandes crues" et "viandes cuites" uniquement) le type de viande (agneau, boeuf, poulet...). Pour cela, on pourra faire une recherche sur les principaux noms de viande et créer une modalité "other" pour les viandes qui apparaissent peu souvent. Représenter les émissions de $C0_2$ en fonction du type de viande. Commenter.

Emissions de $C0_2$ par étape du processus de commercialisation d'un produit

5. Représenter les émissions associées à chaque étape (agriculture, transport...), tous produits confondus. Commenter.
6. Refaire la même chose en séparant par groupe de produits. Commenter.
7. Afficher le top 20 des produits dont le transport produit le plus de C02. Faire pareil pour la transformation et l'agriculture. Commenter.
8. Parmi les produits identifiés à la question précédente, afficher pour chacun la part d'émission due à chaque étape. Commenter.
9. (Bonus) Cette base de données est très riche : si vous avez des idées d'analyse complémentaire, vous pouvez les faire ici !

Exercice 2 : Statistique et génétique

Etude de l'influence du genre sur le BMI

Charger le jeu de données à partir de l'URL suivante :

```
load(url("https://www.biostatistics.dk/teaching/bioinformatics/data/gwaspt.rda"))
head(phenotypes)
```

```
##     BMI gender      age
## 1 23.39 Female 33.02133
## 2 22.72 Female 28.32208
## 3 23.52   Male 25.73901
## 4 25.03   Male 30.80271
## 5 21.64 Female 34.95409
## 6 22.16   Male 28.13544
```

Ce jeu de données contient le BMI (body mass index ou indice de masse corporelle en français), le genre et l'âge de 1324 individus. On se demande s'il y a une différence de BMI entre les hommes et les femmes.

1. Proposer une visualisation des données qui permet d'observer une éventuelle différence. Commenter.
2. Proposer un test pour répondre à cette question, en prenant soin d'écrire le modèle et les hypothèses testées. Conclure.

Etude du lien entre la génétique et le BMI

On dispose maintenant d'information génétique sur les individus, contenu dans le data.frame **genotypes** à charger grâce au lien suivant :

```
load(url("https://www.biostatistics.dk/teaching/bioinformatics/data/gwasgt.rda"))
dim(genotypes)
```

```
## [1]  1324 32019
```

Chaque colonne correspond à un variant du génome, pour lequel il y a 3 modalités possibles "0", "1" ou "2". Il s'agit ici d'une variable qualitative et non quantitative (les modalités des variants pourraient s'appeler A, B et C au lieu de 0, 1 et 2.) L'entrée

```
genotypes[2,5]
```

```
## [1] 0
```

nous dit donc que le 5e variant de l'individu 2 est de type "0". On s'intéresse à savoir si un ou plusieurs variants génétiques sont liés au BMI.

1. Créer une nouvelle variable **BMI_cat** qui prend les modalités : "normal" si le BMI est inférieur à 25 et "overweight" si le BMI est supérieur à 25.
2. En utilisant cette nouvelle variable discrète, proposer une visualisation qui permet d'observer un éventuel lien entre le premier variant (celui donné dans la première colonne de **genotypes**) et le BMI. On pourra utiliser la fonction **table** pour créer une table de contingence. Commenter.
3. Faire un test pour savoir si ce premier variant est indépendant du BMI. Ecrire le modèle, les hypothèses et conclure. Remarque : on fera attention au choix du test à cause de la taille de certains effectifs.
4. Tester si chaque variant est indépendant du BMI. Combien de variants trouvez-vous significativement liés au BMI ? Commenter le résultat.
5. Générer une variable binaire qui vaut 0 ou 1 avec probabilité 1/2, de même longueur que le nombre d'individus. Faire le test pour voir si chaque variant est indépendant de cette nouvelle variable. Commenter.