FACULTÉ DES SCIENCES ECONOMIQUES

UNIVERSITÉ DE RENNES 1

Logiciel R et programmation
Ewen Gallic [1]
Tous documents autorisés

Master 1 Statistique & Économétrie
Date : 07 Décembre 2015
Durée : 2h

Vous avez reçu une archive nommée `cc_2_2015_2016.zip` [2]. Elle contient :

- `cc2.Rproj` : le fichier de projet ;
- `reponses.Rmd` : le fichier dans lequel vous reporterez les réponses aux questions des exercices.

Avant de commencer à rédiger :

1. renommez le fichier `reponses.Rmd` en `NOM_prenom.Rmd`, en remplaçant NOM et prenom par votre nom et votre prénom respectivement ;
2. ouvrez le fichier `cc2.Rproj` dans RStudio ;
3. ouvrez le fichier `NOM_prenom.Rmd` et complétez les premières lignes pour faire figurer votre nom, votre prénom et votre numéro d'étudiant. Rédigez dans ce fichier.

Au moment de remettre votre travail :

1. assurez-vous de bien avoir complété l'en-tête du fichier `NOM_prenom.Rmd` ;
2. envoyez le fichier `NOM_prenom.Rmd` à l'adresse e-mail suivante : ewen.gallic@gmail.com, avec en objet du message : "**[M1 R CC2] - NOM Prénom**", en remplaçant NOM et Prénom par votre nom et prénom respectivement.

Exercice 1 (6 points)

1. Charger en mémoire le jeu de données `climate_2014_nz` disponible à l'adresse suivante : http://egallic.fr/Enseignement/R/2015/cc2/climate_2014_nz.rda. Ce fichier contient des données journalières climatiques par régions en Nouvelle-Zélande pour l'année 2014. Les variables sont les suivantes :
   - `year` (chr) : année,
   - `month` (fctr) : abréviation du mois (trois lettres) en anglais,
   - `day` (chr) : jour du mois sur deux chiffres,
   - `precip` (dbl) : précipitations (millimètres),
   - `temp` (dbl) : température moyenne (degrés C),
   - `region` (chr) : région de Nouvelle-Zélande.
2. Créer un tableau de données indiquant uniquement pour le mois de janvier (donc se restreindre aux observations dont le mois est janvier), la moyenne, la valeur minimale et la valeur maximale des précipitations et des températures dans chaque région. Nommer le tableau comme suit : `df_tmp`. Les premières lignes du résultat doivent ressembler à la sortie ci-après :

```
## Source: local data frame [4 x 7]
##
##          region precip_mean precip_min precip_max temp_mean  temp_min
##           (chr)       (dbl)      (dbl)      (dbl)     (dbl)     (dbl)
## 1      Auckland   0.8717876          0   7.223698  19.19931 15.319401
## 2 Bay of Plenty   1.3495979          0  13.998614  18.96600 14.519504
## 3    Canterbury   1.9417494          0  11.670953  12.11989  7.750046
## 4      Gisborne   1.4001165          0  13.553460  18.91648 14.173536
## Variables not shown: temp_max (dbl)
```

1. ewen.gallic[at]univ-rennes1.fr
2. Elle est disponible à l'URL suivante : http://egallic.fr/Enseignement/R/2015/cc2/cc_2_2015_2016.zip

3. Exporter dans un fichier `HTML` nommé `climate_nz_2014_jan.html` le tableau de données `df_tmp`. Pour ce faire, utiliser les fonctions `xtable()` du *package* du même nom (consulter la page d'aide pour comprendre le fonctionnement) et `print.xtable()`. Ajouter le titre suivant en bas du tableau : "*New Zealand Climate - JAN*" ;
4. Reprendre le code précédent pour créer une fonction qui prendra un seul paramètre : le mois de l'année. Cette fonction doit, pour un mois donné, créer le tableau donnant les moyennes, valeurs minimales et maximales des précipitations et des températures pour chaque région (comme dans la question 2), puis exporter ce tableau au format `HTML` (comme dans la question 3). Le nom du fichier de sortie doit être `climate_nz_2014_*.html`, où `*` doit valoir `jan` pour janvier, `feb` pour février, etc. ;
5. Appliquer la fonction précédente à tous les mois de l'année, pour obtenir 12 fichiers de statistiques descriptives du climat néo-zélandais au format `HTML`.
   Note : il peut être pratique de s'appuyer sur l'objet ***`month.abb`*** *qui contient les abréviations des mois du calendrier.*

Exercice 2 (14 points)

Des listes des Musées de France pour les années 2011, 2012 et 2014 sont publiées en ligne par le Ministère de la Culture et de la Communication sur le site `data.gouv.fr` à l'adresse suivante : https://www.data.gouv.fr/fr/datasets/liste-et-localisation-des-musees-de-france/.
Les libellés des colonnes sont les suivants :

- `NOMREG` : régions,
- `NOMDEP` : départements,
- `DATEAPPELLATION` : date d'appellation,
- `FERME` : fermé (oui si le musé est fermé),
- `ANNREOUV` : date de réouverture,
- `ANNEXE` : annexe (si annexe au musée),
- `NOM DU MUSEE` : nom du musée,
- `ADR` : adresse,
- `CP` : code postal,
- `VILLE` : ville,
- `SITEWEB` : site web,
- `FERMETURE ANNUELLE` : périodes d'ouverture annuelle,
- `PERIODE OUVERTURE` : périodes de fermeture annuelle,
- `JOURS NOCTURNES` : jours nocturnes.

1. Télécharger à l'aide de la fonction appropriée le fichier contenant la liste des Musées de France de 2014 (le lien est le suivant :
   https://www.data.gouv.fr/s/resources/liste-et-localisation-des-musees-de-france/20150415-175525/Liste_musees_de_France.xls), et nommer le fichier `musees_2014.xls` ;
2. Importer le fichier Excel téléchargé dans `R`, nommer le tableau de données `musees_year` ;
3. Modifier le tableau `musees_year` pour ne conserver que les observations pour lesquelles le musée est ouvert ;
4. Créer le tableau de données `nb_musees_year` qui indique pour chaque région (variable `NOMREG`) le nombre de musées. Trier ensuite ce tableau par valeurs décroissantes du nombre de musées, puis ajouter la variable `annee` qui prendra la valeur `2014` pour toutes les observations.
   Note : la fonction ***`n()`*** *permet d'obtenir le nombre d'observations dans chaque groupe pour un tableau dans lequel les observations sont regroupées selon une ou plusieurs variables.*
   Les premières lignes du tableau `nb_musees_year` doivent ressembler à la sortie ci-après :

```
## Source: local data frame [6 x 3]
##
##                        NOMREG nb_musees annee
```

```
##                          (chr)     (int) (dbl)
## 1                ILE-DE-FRANCE       113  2014
## 2 PROVENCE-ALPES-CÔTE D'AZUR         96  2014
## 3                  RHÔNE-ALPES        86  2014
## 4                MIDI-PYRENEES        60  2014
## 5                    BOURGOGNE        55  2014
## 6                       CENTRE        51  2014
```

5. Créer une fonction qui télécharge la liste des Musées de France pour une année et un lien donnés (la fonction prend donc deux paramètres), charge les données dans `R` et retourne un tableau indiquant pour chaque année et chaque région, le nombre de Musées de France. Ce tableau doit être trié par valeurs des années décroissantes.

   Il s'agit ici de créer une fonction en s'appuyant sur le code déjà développé lors des questions précédentes. Il peut être pratique de créer, à l'intérieur de la fonction, la variable `nom_fichier`*, qui contiendra sous forme de chaînes de caractères, le nom du fichier Excel qui sera enregistré puis importé dans la session* `R`*;*

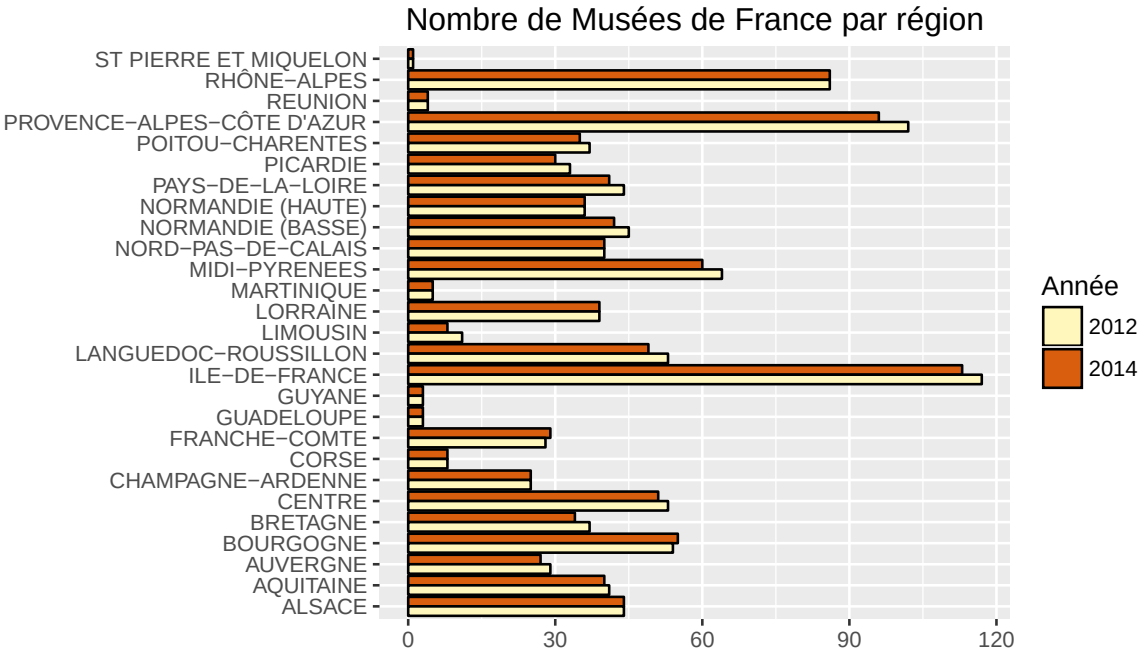
6. Appliquer la fonction créée à la question précédente pour récupérer le nombre de Musées de France par région pour les années 2012 et 2014, dont les liens sont respectivement :
   https://www.data.gouv.fr/storage/f/2013-12-05T14%3A58%3A34.851Z/liste-musees-de-france-2012.xls

   https://www.data.gouv.fr/s/resources/liste-et-localisation-des-musees-de-france/20150415-175525/Liste_musees_de_France.xls

   Ensuite, coller les deux tableaux l'un en dessous de l'autre, dans un nouveau tableau que l'on appellera `musees` ;

7. Transformer la variable `annee` du tableau `musees` en facteur ;

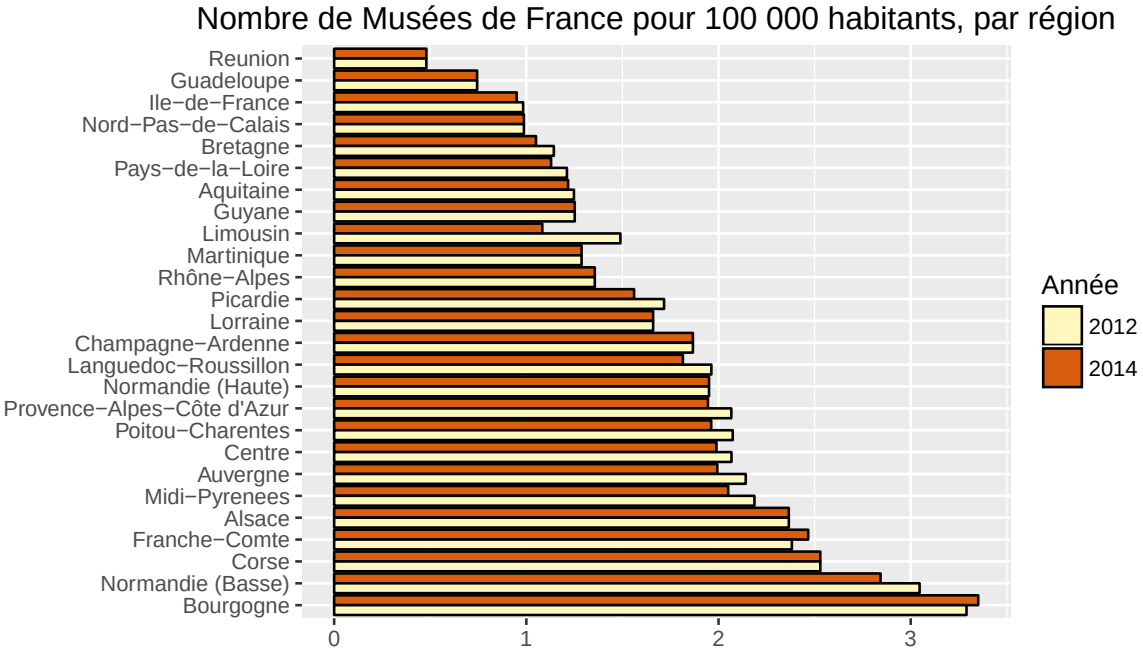
8. En utilisant les valeurs contenues dans le tableau `musees`[3], reproduire le graphique ci-après (faire appel à la fonction `coord_flip()` pour faire basculer les axes) ;



9. Charger dans la session `R` le tableau de données `pop` disponible à l'adresse suivante : http://egallic.fr/Enseignement/R/2015/cc2/insee_pop_2012.rda. Il s'agit de la population par région en France en 2012, telle que renseignée sur le site de l'INSEE. Ensuite, ajouter au tableau `pop` la variable `NOMREG` indiquant le nom de la région passé en majuscules ;

3. Disponible à l'adresse suivante pour les personnes n'ayant pas réussi les questions précédentes : http://egallic.fr/Enseignement/R/2015/cc2/musees.rda

10. Ajouter les informations contenues dans le tableau `pop` au tableau `musees` à l'aide d'une jointure, puis ajouter la variable `nb_musees_hab` qui donne le nombre de musées par région pour 100 000 habitants ;
11. Retirer les observations du tableau `musees` pour lesquelles la variable `nb_musees_hab` vaut `NA` ;
12. En utilisant les données du tableau `musees`, créer un tableau donnant le nombre moyen de musées par habitant et par région, puis trier ce tableau par valeurs décroissantes. Ensuite, en utilisant soit le symbole dollar ou la fonction `"["()`, extraire la colonne `region` de ce tableau, et stocker la chaîne de caractères retournée dans un objet que l'on appellera `noms_regions` ;
13. Modifier la variable `region` du tableau de données `musees`, de manière à ce qu'elle soit de type `factor`, et que ses niveaux soient définis par la chaîne de caractères `noms_regions` créée à la question précédente ;
14. Reproduire le graphique ci-après à partir des données contenues dans le tableau `musees`[4].



4. Disponible à l'adresse suivante pour les personnes n'ayant pas réussi les questions précédentes : `http://egallic.fr/Enseignement/R/2015/cc2/musees_2.rda`