Datamining tp2

David Tabarie

May 17, 2019

Contents

1 Sujet

Le TP consiste à préparer les données d'un fichier afin de pouvoir effectuer des régressions ou fouilles dessus. A partir du fichier de Consommation électrique annuelle à la maille département disponible à l'adresse `https://www.data.gouv.fr/fr/datasets/r/7f293530-354f-4721-aa4e-ae578a377180` et décrit sur `https://www.data.gouv.fr/fr/datasets/consommation-electrique-annuelle-a-la-mai` :

2 Charger le fichier csv avec la commande read.csv

```
data <- read.csv(file="consommation-electrique-par-secteur-dactivite-departement.csv",
```

3 Vérifier le bon chargement avec un sommaire (summary)

```
data <- read.csv(file="consommation-electrique-par-secteur-dactivite-departement.csv",
names(data) # Summary ne s'affiche pas bien dans mon éditeur, je met donc le commande
```

Année
Nom.département
Code.département
Nom.région
Code.région
Nb.sites.Résidentiel
Conso.totale.Résidentiel..MWh.
Conso.moyenne.Résidentiel..MWh.
Nb.sites.Professionnel
Conso.totale.Professionnel..MWh.
Conso.moyenne.Professionnel..MWh.
Nb.sites.Agriculture
Conso.totale.Agriculture..MWh.
Nb.sites.Industrie
Conso.totale.Industrie..MWh.
Nb.sites.Tertiaire
Conso.totale.Tertiaire..MWh.
Nb.sites.Secteur.non.affecté
Conso.totale.Secteur.non.affecté..MWh.
Nombre.d.habitants
Taux.de.logements.collectifs
Taux.de.résidences.principales
Superficie.des.logements. . . 30.m2
Superficie.des.logements.30.à.40.m2
Superficie.des.logements.40.à.60.m2
Superficie.des.logements.60.à.80.m2
Superficie.des.logements.80.à.100.m2
Superficie.des.logements. . . 100.m2
Résidences.principales.avant.1919
Résidences.principales.de.1919.à.1945
Résidences.principales.de.1946.à.1970
Résidences.principales.de.1971.à.1990
Résidences.principales.de.1991.à.2005
Résidences.principales.de.2006.à.2010
Résidences.principales.après.2011
Taux.de.chauffage.électrique
Geo.Shape
Geo.Point.2D

4 Supprimer les colonnes de type texte et la colonne geoshape

```
data <- read.csv(file="consommation-electrique-par-secteur-dactivite-departement.csv",
data$Geo.Shape <- Nom.département <- Nom.région <- NULL
summary(data)
```

Min. :2011	Ain : 7	Min. : 1.0	Occitanie : 91	Min. :
1st Qu.:2012	Aisne : 7	1st Qu.:25.0	Auvergne-Rhône-Alpes : 84	1st Qu
Median :2014	Allier : 7	Median :48.5	Nouvelle Aquitaine : 84	Media
Mean :2014	Alpes-de-Haute-Provence: 7	Mean :48.3	Grand-Est : 70	Mean
3rd Qu.:2016	Alpes-Maritimes : 7	3rd Qu.:72.0	Bourgogne-Franche-Comté: 56	3rd Qu
Max. :2017	Ardèche : 7	Max. :95.0	Île-de-France : 56	Max.
nil	(Other) :616	nil	(Other) :217	nil

5 Pour chaque intervalle de superficie, ajouter une colonne de consommation en MWh (utiliser la conso globale résidentielle)

```
data <- read.csv(file="consommation-electrique-par-secteur-dactivite-departement.csv",
data <- data$Geo.Shape <- Nom.département <- Nom.région <- NULL
data <- cbind(data, "Consommation <30m2(MWh)" = (data$Conso.totale.RÃ.sidentiel..MWh./
summary(data)
```

Min. : NA	Min. : NA	Min. : NA	Min. : NA	Min. : NA	Min. : NA	Min. :
1st Qu.: NA	1st Qu.: NA	1st Qu.: NA	1st Qu.: NA	1st Qu.: NA	1st Qu.: NA	1st Qu
Median : NA	Median : NA	Median : NA	Median : NA	Median : NA	Median : NA	Media
Mean :NaN	Mean :NaN	Mean :NaN	Mean :NaN	Mean :NaN	Mean :NaN	Mean
3rd Qu.: NA	3rd Qu.: NA	3rd Qu.: NA	3rd Qu.: NA	3rd Qu.: NA	3rd Qu.: NA	3rd Qu
Max. : NA	Max. : NA	Max. : NA	Max. : NA	Max. : NA	Max. : NA	Max. :

6 Pour chaque catégorie (Résidentiel, Professionnel, Agriculture, Industrie, Tertiaire), ajouter une colonne de pourcentage pour calculer le % de consommation de la catégorie

```
data <- read.csv(file="consommation-electrique-par-secteur-dactivite-departement.csv",
```

```
data <- data$Geo.Shape <- Nom.département <- Nom.région <- NULL
data <- cbind(data,"pourcentage Consommation Résidentiel" = ( 100 * data$Conso.totale.
summary(data)
```

Min. : NA	Min. : NA	Min. : NA	Min. : NA	Min. : NA	Min. : NA
1st Qu.: NA	1st Qu.: NA	1st Qu.: NA	1st Qu.: NA	1st Qu.: NA	1st Qu.: NA
Median : NA	Median : NA	Median : NA	Median : NA	Median : NA	Median : NA
Mean :NaN	Mean :NaN	Mean :NaN	Mean :NaN	Mean :NaN	Mean :NaN
3rd Qu.: NA	3rd Qu.: NA	3rd Qu.: NA	3rd Qu.: NA	3rd Qu.: NA	3rd Qu.: NA
Max. : NA	Max. : NA	Max. : NA	Max. : NA	Max. : NA	Max. : NA