Projet R avancé

Margaux Bailleul - Marie Guibert

2023-04-25

```
library(rvest)
library(tidyverse)
library(jsonlite)
```

Présentation de notre projet

Lors de cette étude, nous avons abordé un sujet qui nous tient particulièrement à coeur : les festivals de musique. En effet, nous avons choisi d'aborder ce sujet et de nous focaliser les festivals de musique amplifiées et électroniques. Ces évènements ont une portée culturelle très importante et un impact économique dans le monde entier.

Données provenant de l'API

Extraction des données via une API

```
url_festival <- "https://data.culture.gouv.fr/api/records/1.0/search/?dataset=panorama-des-festivals&q=
```

Nous utilisons le package **jsonlite** pour extraire les données de l'API.

```
contenu <- fromJSON(url_festival)
df_festivals <- contenu$records$fields
glimpse(df_festivals)
```

Tout d'abord, visualisons nos données de façon globale sans effectuer de modifications :

```
head(df_festivals,3)
```

```
##   dept_sk    coordonnees_insee date_de_fin_ancien
## 1      29 48.400500, -4.502791         2019-02-10
## 2      34  43.571628, 3.832218         2019-03-09
## 3      85 46.866613, -1.022161         2019-03-09
##   mois_indicatif_en_chiffre_y_compris_double_mois nom_departement departement
## 1                                               2       Finistère          29
## 2                                               3         Hérault          34
## 3                                               3          Vendée          85
##   periodicite mois_habituel_de_debut code_postal
```

```
## 1    Annuelle           02 (février)       29200
## 2    Annuelle              03 (mars)       34430
## 3    Annuelle              03 (mars)       85500
##                     complement_domaine libelle_commune_pour_calcul_cp_insee
## 1 Musiques amplifiées ou électroniques                                BREST
## 2 Musiques amplifiées ou électroniques                    ST JEAN DE VEDAS
## 3 Musiques amplifiées ou électroniques                        LES HERBIERS
##              domaine date_debut_ancien code_insee commune_principale
## 1 Musiques actuelles        2019-02-05      29019              BREST
## 2 Musiques actuelles        2019-03-08      34270   ST JEAN DE VEDAS
## 3 Musiques actuelles        2019-03-08      85109       LES HERBIERS
##             region nom_de_la_manifestation ndeg_identification date_de_creation
## 1         Bretagne      ASTROPOLIS L'HIVER               CD023       2012-01-01
## 2        Occitanie         THIS IS ENGLAND               KD257       2012-01-01
## 3 Pays de la Loire        FREE SONS DIVERS               LD113       2003-01-01
##   check_edition ndeg_de_l_edition_2018 ndeg_de_l_edition_2019
## 1             0                      7                      8
## 2             0                      7                      8
## 3             0                     16                     17
##
## 1
## 2 8ème édition. L'événement qui tient son nom d'une chanson des CLASH propose un voyage au pays natal
## 3
##                     site_web autres_communes
## 1                       <NA>            <NA>
## 2 tafproduction.blogspot.fr/            <NA>
## 3     www.freesonsdivers.com            <NA>
##   soutenu_en_2017_par_le_ministere_de_la_culture
## 1                                           <NA>
## 2                                           <NA>
## 3                                           <NA>
##   soutenu_en_2017_par_le_centre_national_des_varietes mois_indicatif
## 1                                                <NA>           <NA>
## 2                                                <NA>           <NA>
## 3                                                <NA>           <NA>
##   soutenu_en_2018_par_le_centre_national_des_varietes
## 1                                                <NA>
## 2                                                <NA>
## 3                                                <NA>
##   soutenu_en_2017_par_le_centre_national_du_cinema
## 1                                             <NA>
## 2                                             <NA>
## 3                                             <NA>
```

Nettoyage de la base de données

Afin de faciliter notre étude, nous avons choisi de supprimer certaines colonnes de la base de données. De plus, certaines informations sont redondantes, nous avons donc choisi de les omettre aussi. Par exemple, le domaine correspond aux musiques actuelles et plus spécialement aux musiques amplifiées ou électroniques (complement_domaine). Ces deux colonnes n'étaient donc pas pertinentes pour la suite de notre analyse.

```
df_festivals <-df_festivals |>
  select(coordonnees_insee,date_de_fin_ancien,nom_departement,departement,periodicite,mois_habituel_de_
```

```
head(df_festivals,3)
```

```
##      coordonnees_insee date_de_fin_ancien nom_departement departement
## 1 48.400500, -4.502791         2019-02-10       Finistère          29
## 2  43.571628, 3.832218         2019-03-09         Hérault          34
## 3 46.866613, -1.022161         2019-03-09          Vendée          85
##   periodicite mois_habituel_de_debut code_postal
## 1    Annuelle           02 (février)       29200
## 2    Annuelle              03 (mars)       34430
## 3    Annuelle              03 (mars)       85500
##   libelle_commune_pour_calcul_cp_insee date_debut_ancien           region
## 1                                BREST        2019-02-05         Bretagne
## 2                    ST JEAN DE VEDAS        2019-03-08        Occitanie
## 3                        LES HERBIERS        2019-03-08 Pays de la Loire
##   nom_de_la_manifestation                   site_web
## 1      ASTROPOLIS L'HIVER                       <NA>
## 2         THIS IS ENGLAND tafproduction.blogspot.fr/
## 3        FREE SONS DIVERS     www.freesonsdivers.com
```

Deuxièmement, nous allons transformer les variables caractères en facteurs pour effectuer des traitements de données et des graphiques plus facilement.

```
# str(df_festivals) # permet de connaître le type de chaque variable du dataframe
df_festivals[c("nom_departement",
               "periodicite",
               "code_postal",
               "libelle_commune_pour_calcul_cp_insee",
               "region")] <- lapply(df_festivals[c("nom_departement",
               "periodicite",
               "code_postal",
               "libelle_commune_pour_calcul_cp_insee",
               "region")], as.factor)
# str(df_festivals)
```

Dans cette étude, nous allons nous concentrer sur les festivals annuels. Nous choisissons donc de ne pas prendre en compte les autres modalités de la variable **periodicite**.

```
table(df_festivals$periodicite)
```

```
##
##                  Annuelle                 Biennale Biennale années impaires
##                       553                        1                         1
##   Biennale années paires
##                        2
```

```
df_festivals <- df_festivals |>
  filter(periodicite == "Annuelle") |> # filtrage pour n'avoir que les festivals annuels
  select(-periodicite) # suppression de la colonne periodicite
                       # car les informations sont redondantes à présent

# Vérification :
# table(df_festivals$periodicite)
```

Nous allons extraire les coordonnées GPS de la variable **coordonnes_insee** afin de la scinder en deux colonnes : latitude et longitude. Cette étape nous permettra de faire plus facilement nos graphiques par la suite.

```
# On extrait d'abord les coordonnées et on créé un dataframe contenant la latitude et la longitude
coord_df <- data.frame(matrix(unlist(sapply(df_festivals$coordonnees_insee, function(x) {
  unlist(strsplit(paste(x, collapse = ", "), ", "))
})), ncol = 2, byrow = TRUE))

coord_df <- coord_df |>
  rename("longitude" = X1, "latitude" = X2)

# Vérification de la bonne forme du dataframe
# coord_df

# On concaténe les deux dataframes
df_festivals <- bind_cols(df_festivals,coord_df)
# df_festivals

# On supprime la colonne coordonnees_insee car on ne la réutilisera pas
df_festivals <- df_festivals |>
  select(-coordonnees_insee)

# Vérification
head(df_festivals,3)
```

```
##   date_de_fin_ancien nom_departement departement mois_habituel_de_debut
## 1         2019-02-10       Finistère          29           02 (février)
## 2         2019-03-09         Hérault          34              03 (mars)
## 3         2019-03-09          Vendée          85              03 (mars)
##   code_postal libelle_commune_pour_calcul_cp_insee date_debut_ancien
## 1       29200                                BREST        2019-02-05
## 2       34430                    ST JEAN DE VEDAS        2019-03-08
## 3       85500                        LES HERBIERS        2019-03-08
##             region nom_de_la_manifestation                   site_web
## 1         Bretagne      ASTROPOLIS L'HIVER                       <NA>
## 2        Occitanie         THIS IS ENGLAND tafproduction.blogspot.fr/
## 3 Pays de la Loire        FREE SONS DIVERS     www.freesonsdivers.com
##       longitude       latitude
## 1 48.4004997828  -4.5027907853
## 2 43.5716282319  3.83221847952
## 3 46.8666125813 -1.02216086186
```

```
write.table(df_festivals,"donnees_festivals.csv",sep=";")
```

Quelques graphiques

Tout d'abord, nous allons pouvoir visualiser les régions avec le plus d'évènements.

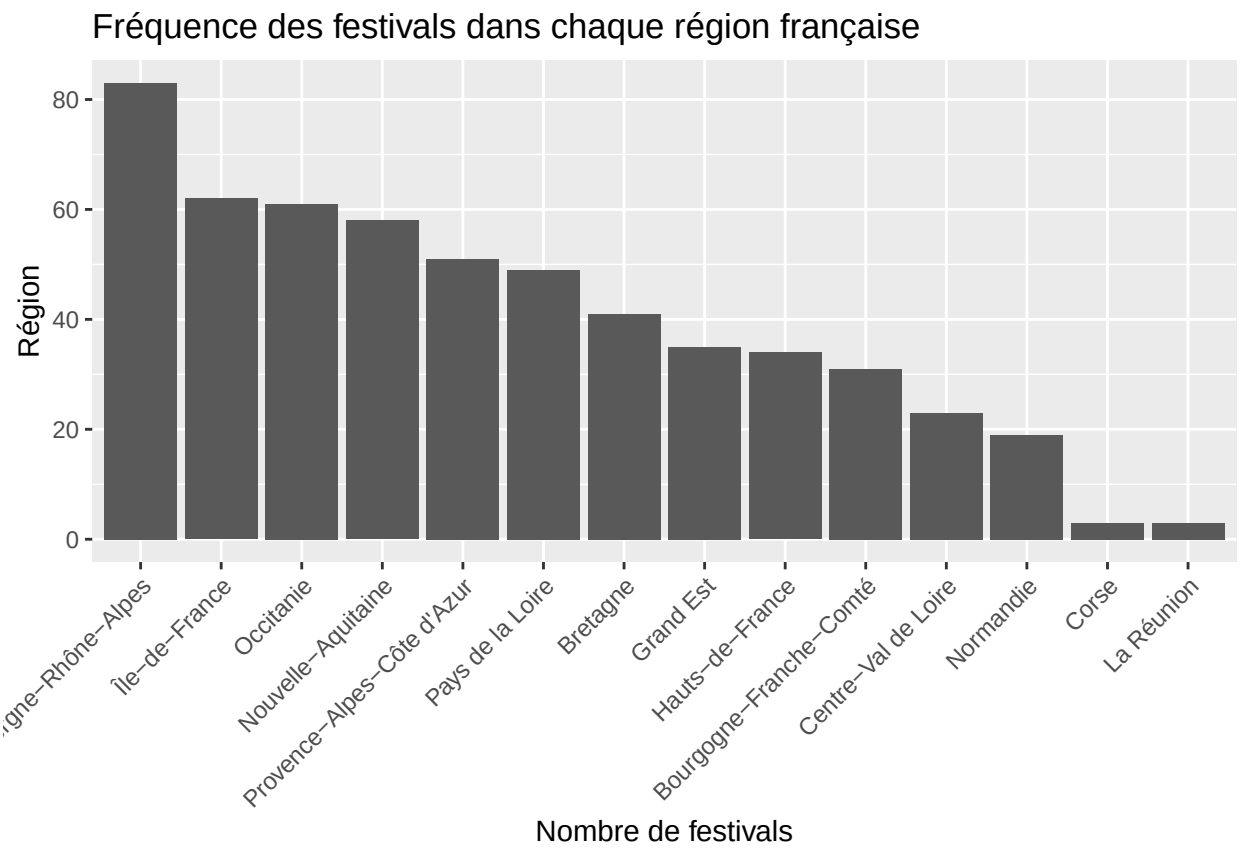
```
# Calculer les fréquences de chaque région
freq <- table(df_festivals$region)
```

```r
# Créer un dataframe avec les fréquences de chaque région
df_regions <- data.frame(table(df_festivals$region)) |>
  rename(Region = Var1, Frequence = Freq) |>
  mutate(Pourcentage = round(Frequence / sum(Frequence) * 100, 1)) |>
  arrange(-Frequence) # on trie selon le nombre de festivals dans la région
head(df_regions,3)
```

```
##                 Region Frequence Pourcentage
## 1 Auvergne-Rhône-Alpes        83        15.0
## 2        Île-de-France        62        11.2
## 3            Occitanie        61        11.0
```

```r
ggplot(df_regions, aes(x = reorder(Region,-Frequence), y = Frequence)) +
  geom_bar(stat = "identity") +
  labs(title = "Fréquence des festivals dans chaque région française", x = "Nombre de festivals", y = "
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  guides(fill=FALSE)
```

```
## Warning: The ‘<scale>‘ argument of ‘guides()‘ cannot be ‘FALSE‘. Use "none" instead as
## of ggplot2 3.3.4.
```



Données provenant d'un site web

Extraction des données via un site web

```
site_festivals <- "https://martinbeatz.com/festivals-electro-france/"
```

```
festival_html <- read_html(site_festivals)
```

Grâce à cette page web, nous allons pouvoir établir les festivals les plus en vogue.

```
listes_festivals <- festival_html |>
  html_nodes(xpath = "//h2") |>
  html_text()
# listes_festivals

cat("Les festivals les plus recommandés sont :", paste("\n", listes_festivals))
```

```
## Les festivals les plus recommandés sont :
##  Electrobeach Festival
##  Les Plages Electroniques
##  Elektric Park
##  Cocorico Electro
##  Tomorrowland Winter
##  Delta Festival
##  Hope Festival - Toulouse
##  Reperkusound
##  Cercle Festival
##  Stereoparc
##  Touquet Music Beach
##  Summer Festival
##  Panoramas Festival
##  Marvellous Festival
##  I Love Techno
##  Dream Nation
##  Nuits Sonores
##  Pharaonic Festival
```

Obtention du lieu d'un festival

```
liste_lieux <- festival_html |>
  html_nodes(xpath = "//*[@id='post-2934']/div/div/div/p[9]/text()[1]") |>
  html_text()
liste_lieux
```

```
## [1] "Lieu : Plage du Palais des Festivals - Cannes"
```

```
# Pareil nous intéresse toutes les 5 valeurs, comment tous les récupérer dans une seule liste ?
```

Obtention de la date du festival

```
liste_dates_prov <- festival_html |>
  html_nodes(xpath = "//*[@id='post-2934']/div/div/div/p[position()=4 or position()=9 or position()=14 
  html_text()
liste_dates <- gsub("\nDate : ", "", liste_dates_prov)
liste_dates
```

```
##  [1] "14, 15 et 16 Juillet 2023" "4, 5 et 6 Août 2023"
##  [3] "1 et 2 Septembre 2023"     "13, 14 et 15 Juillet 2023"
##  [5] "18 au 24 Mars 2023"        "23 au 27 Août 2023"
##  [7] "2021"                      "8, 8 et 9 Avril 2023"
##  [9] "11 & 12 Septembre 2021"    "21 et 22 Juillet 2023"
## [11] "25 et 26 Août 2023"        "10 et 11 Septembre 2021"
## [13] "22 au 24 Septembre 2023"   "27 et 28 Mai 2023"
## [15] "7, 8 et 9 Avril 2023"      "17, 18, 19 Septembre 2021"
## [17] "19 au 25 Juillet 2021"     "26 Mars 2022"
```

```
for(i in seq_along(listes_festivals)){
    cat("Le festival", listes_festivals[[i]], "aura lieu les", liste_dates[[i]], "à", liste_lieux, "\n"
}
```

```
## Le festival Electrobeach Festival aura lieu les 14, 15 et 16 Juillet 2023 à Lieu : Plage du Palais d
## Le festival Les Plages Electroniques aura lieu les 4, 5 et 6 Août 2023 à Lieu : Plage du Palais des 
## Le festival Elektric Park aura lieu les 1 et 2 Septembre 2023 à Lieu : Plage du Palais des Festivals
## Le festival Cocorico Electro aura lieu les 13, 14 et 15 Juillet 2023 à Lieu : Plage du Palais des Fes
## Le festival Tomorrowland Winter aura lieu les 18 au 24 Mars 2023 à Lieu : Plage du Palais des Festiva
## Le festival Delta Festival aura lieu les 23 au 27 Août 2023 à Lieu : Plage du Palais des Festivals -
## Le festival Hope Festival - Toulouse aura lieu les 2021 à Lieu : Plage du Palais des Festivals - Can
## Le festival Reperkusound aura lieu les 8, 8 et 9 Avril 2023 à Lieu : Plage du Palais des Festivals -
## Le festival Cercle Festival aura lieu les 11 & 12 Septembre 2021 à Lieu : Plage du Palais des Festiva
## Le festival Stereoparc aura lieu les 21 et 22 Juillet 2023 à Lieu : Plage du Palais des Festivals - C
## Le festival Touquet Music Beach aura lieu les 25 et 26 Août 2023 à Lieu : Plage du Palais des Festiva
## Le festival Summer Festival aura lieu les 10 et 11 Septembre 2021 à Lieu : Plage du Palais des Festi
## Le festival Panoramas Festival aura lieu les 22 au 24 Septembre 2023 à Lieu : Plage du Palais des Fes
## Le festival Marvellous Festival aura lieu les 27 et 28 Mai 2023 à Lieu : Plage du Palais des Festival
## Le festival I Love Techno aura lieu les 7, 8 et 9 Avril 2023 à Lieu : Plage du Palais des Festivals -
## Le festival Dream Nation aura lieu les 17, 18, 19 Septembre 2021 à Lieu : Plage du Palais des Festiva
## Le festival Nuits Sonores aura lieu les 19 au 25 Juillet 2021 à Lieu : Plage du Palais des Festivals
## Le festival Pharaonic Festival aura lieu les 26 Mars 2022 à Lieu : Plage du Palais des Festivals - Ca
```

Zoom sur le Electrobeach Music Festival

Il est le plus grand festival français de musiques électroniques fondé par Alain Ferrand. Il a lieu au Barcarès en face du Lydia, le plus vieux paquebot du monde ensablé depuis 1967.

```
site_electro_beach <- "https://fr.wikipedia.org/wiki/Electrobeach_Music_Festival"
```

```
electro_beach_html <- read_html(site_electro_beach)
```

Nous allons maintenant cherche à en savoir un peu plus sur ce festival, en commençant par son histoire.

Mise en relation entre le JSON et le site des festivals