Panama papers

Thomas Gargot

15 octobre 2016

Panama papers

Ce fichier Rmd pour R markdown combine un logiciel de statistiques R (dans des blocs appelées chunks) avec un langage de balisage (de rédaction) markdown. Il permet à Rstudio de générer des rapports .pdf (texte non modifiable), .html (pages internets), .doc (texte modifiable, word).

Présentation de R

R est un logiciel libre de statistique extrêmement puissant qui permet une prise en main plus rapide et aisée (comparée à d'autres logiciels de programmation). Grâce à un grand nombre de librairies disponibles, R propose de nombreuses fonctionnalités (analyse statistique, textmining, analyse de réseaux, etc.) qui répondent aux besoins de différentes communautés de recherche (SHS et STM). Le partage des scripts sur des plateformes tels que Gitbub permet de rendre les analyses complètement reproductibles.

L'objectif de cet atelier est de vous faire découvrir R et la philosophie qui lui est associée. Nous vous montrerons comment débuter avec R: c'est à dire importer une base de donnée, tracer un graphe et faire des statistiques de bases.

Base de données

Nous allons analyser une partie de la base de données panama papers publiée par le sunday times. Elle a fait l'actualité en 2016 car il s'agit de la plus grosse fuite de données de l'histoire qui met en évidence le fonctionnement des paradis fiscaux et de l'évasion fiscale.

Dictionnaire de données

Ce paragraphe décrit la signification de chaque variable

- company_url : OpenCorporates URL for the company.
- company_name : Company name as stated in the Panama companies registry.
- officer_position_es : Officer's position in the company, as stated in Spanish in the Panama companies registry.
- officer_position_en : Officer's position in the company, translated into English by the Sunday Times Data Team.
- officer_name : Officer's name, as stated in the Panama companies registry. Note that this can be the name of a person or another company. Sometimes more than one person is listed in one record.
- inc_date : Incorporation date of the company.
- dissolved_date : Dissolution date of the company, or 0000-00-00 if the company is current.
- updated_date : Time and date given by the Panama companies registry when OpenCorporates retrieved the record.
- company_type : Company type as stated in the Panama companies registry in Spanish.

- mf_link : Indicator denoting officers we have linked to Mossack Fonseca. Please note that our search has not been exhaustive - there are likely many more associates of Mossack Fonseca in this data who we have not yet discovered. 0 = not linked, 1 = linked.

Récupérer des fichiers sur son ordinateur

```
#Afficher le répertoire courant
getwd()
```

```
## [1] "/Users/Ofix/Documents/Fac/internat/Recherche/projets/openScience/R"
```

```
# Importer des fichiers et les sauvegarder dans la base de donnée (dataframe appellée panama), nous pré

#panama <- read.csv("/Users/Ofix/Desktop/sunday_times_panama_data/sunday_times_panama_data.csv")
```

Télécharger des fichiers en ligne

```
#Crée un fichier temporaire
temp <- tempfile()
# Importer des fichiers et les sauvegarder dans la base de donnée (dataframe appellée panama)
## Télécharge le fichier
download.file("https://cdn.rawgit.com/times/data/master/sunday_times_panama_data.zip",temp)
# Lit le fichier
panama <- read.csv(unz(temp,"sunday_times_panama_data.csv"),stringsAsFactors = FALSE)
# Efface le fichier temporaire
unlink(temp)
rm(temp)
```

Cette stratégie est très pratique car elle permet de rendre le travail d'analyse le plus reproductible possible. Chacun peut récupérer les données brutes et faire la même analyse que l'auteur.

Description de la base de données

```
# dimensions
dim(panama)
```

```
## [1] 528998     10
```

```
# structure de chaque variables : exemple et type
str(panama)
```

```
## 'data.frame':    528998 obs. of  10 variables:
##  $ company_url        : chr  "https://opencorporates.com/companies/pa/100056" "https://opencorporates
##  $ company_name       : chr  "OVERSEAS FINANCIAL AND INSURANCE SERVICES" "OVERSEAS FINANCIAL AND INSU
##  $ officer_position_es: chr  "agent" "presidente" "tesorero" "secretario" ...
##  $ officer_position_en: chr  "Legal agent" "President/Chairman" "Treasurer" "Secretary" ...
##  $ officer_name       : chr  "JURGEN MOSSACK" "JURGEN MOSSACK" "JURGEN MOSSACK" "DIVA ARGELIS PATINO
##  $ inc_date           : chr  "1982-11-09" "1982-11-09" "1982-11-09" "1982-11-09" ...
##  $ dissolved_date     : chr  "1990-01-05" "1990-01-05" "1990-01-05" "1990-01-05" ...
##  $ updated_date       : chr  "2016-03-17 08:19:25" "2016-03-17 08:19:25" "2016-03-17 08:19:25" "2016
```

```
##  $ company_type       : chr  "SOCIEDAD ANONIMA\n" "SOCIEDAD ANONIMA\n" "SOCIEDAD ANONIMA\n" "SOCIEDAD
##  $ mf_link            : int  1 1 1 1 1 1 0 1 1 0 ...
```

Quelle est le nombre de companies dans cette base de données ?

```
head(panama$company_name)
```

```
## [1] "OVERSEAS FINANCIAL AND INSURANCE SERVICES"
## [2] "OVERSEAS FINANCIAL AND INSURANCE SERVICES"
## [3] "OVERSEAS FINANCIAL AND INSURANCE SERVICES"
## [4] "OVERSEAS FINANCIAL AND INSURANCE SERVICES"
## [5] "OVERSEAS FINANCIAL AND INSURANCE SERVICES"
## [6] "OVERSEAS FINANCIAL AND INSURANCE SERVICES"
```

```
head(unique(panama$company_name))
```

```
## [1] "OVERSEAS FINANCIAL AND INSURANCE SERVICES"
## [2] "PEPINO CORPORATION"
## [3] "SOCIETE THS DE COMMERCE, S.A."
## [4] "WISCOL OVERSEAS, S.A."
## [5] "MARINE WORLD CARRIERS S. A."
## [6] "WEBB HOLDINGS LTD., INC."
```

```
length(panama$company_name)
```

```
## [1] 528998
```

Il y a 528998 compagnies répertoriées dans cette base de données.

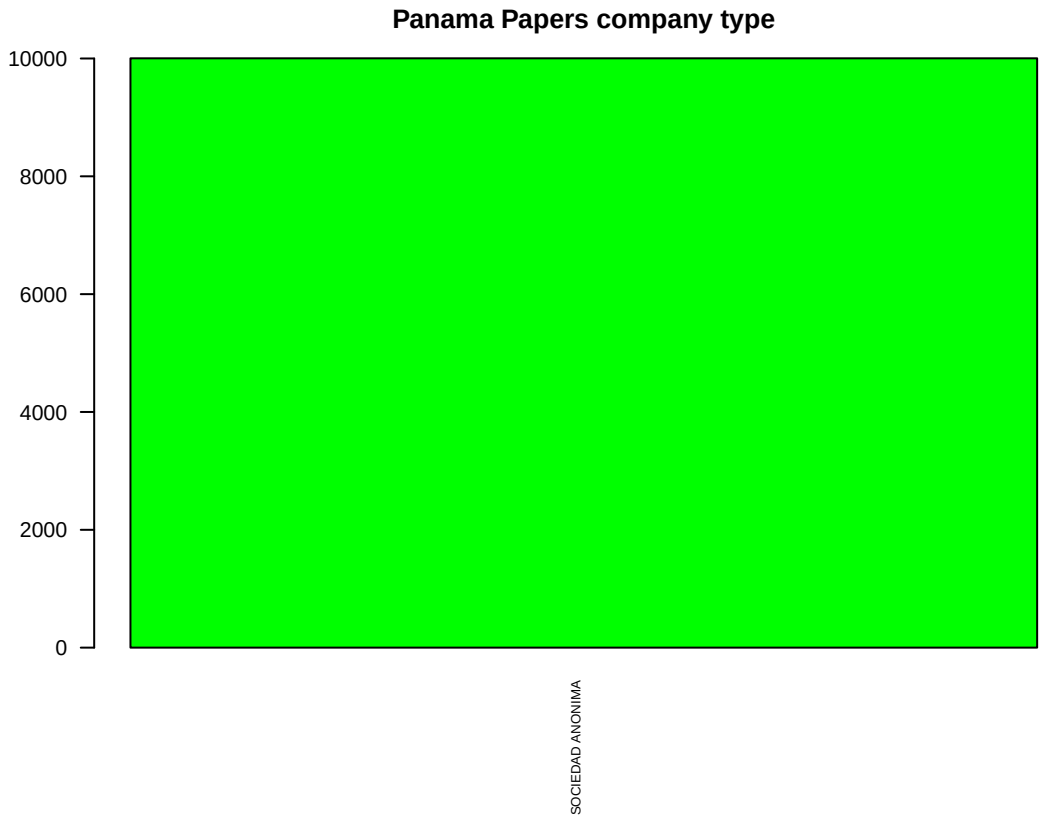
Sélection d'une partie exemple de la base de données

```
# je sélectionne les lignes de la première à la 10003ème
# View(panama[1:10003,])
pana <- panama[1:10003,]
```

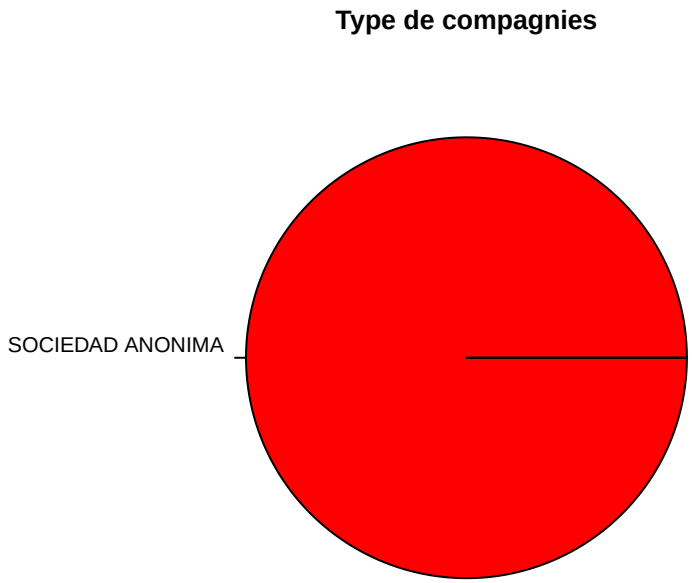
```
table(pana$company_type)
```

```
##
## SOCIEDAD ANONIMA\n
##              10003
```

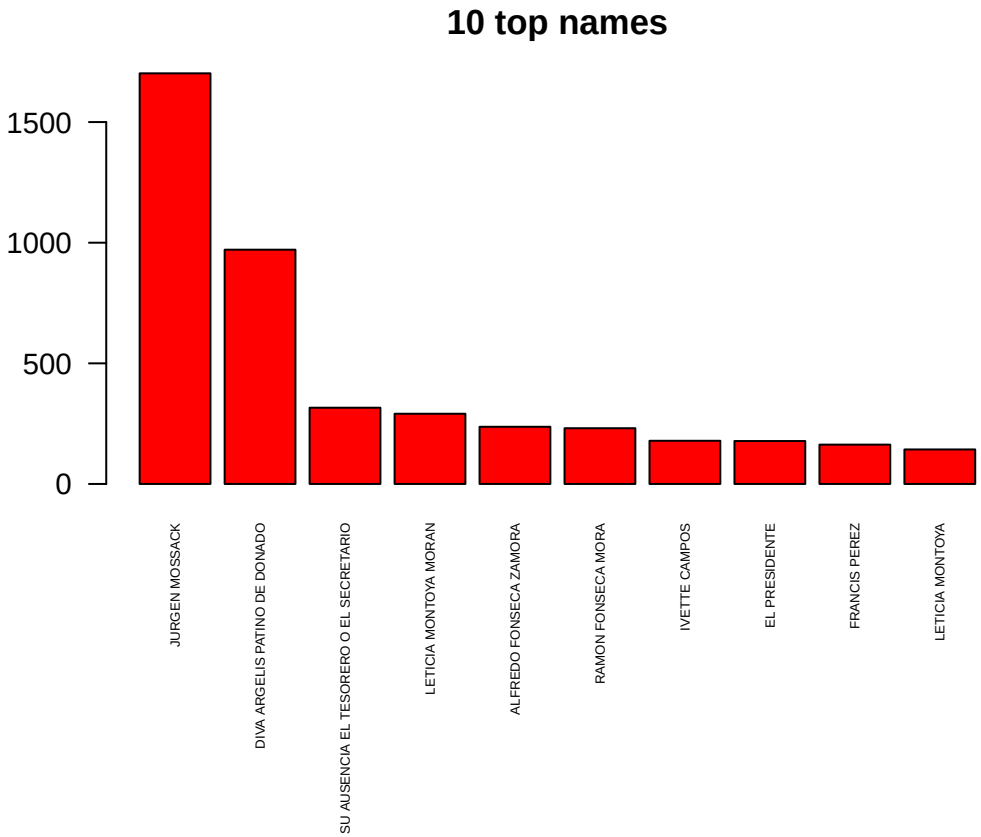
```
par(mar=c(7,4,3,6), cex=0.7)
barplot(sort(table(pana$company_type), decreasing =TRUE),
        cex.names = 0.6, las=2, col="green",
        main="Panama Papers company type")
```




```
pie(sort(table(pana$company_type)), col="red", main="Type de compagnies")
```




```
par(mar=c(10,4,3,6), cex=0.9)
topNames <-sort(table(pana$officer_name), decreasing = TRUE)[1:10]
barplot(topNames, las=2, cex.names=0.5, col="red", main="10 top names")
```



- Jürgen Mossack, né le 20 mars 1948 à Fürth (Bavière) est un avocat d'affaires panaméen, d'origine allemande1. Avec Ramón Fonseca Mora, il est le co-fondateur de Mossack Fonseca, un cabinet juridique basé au Panama et mis en cause dans l'affaire dite des « Panama Papers ».

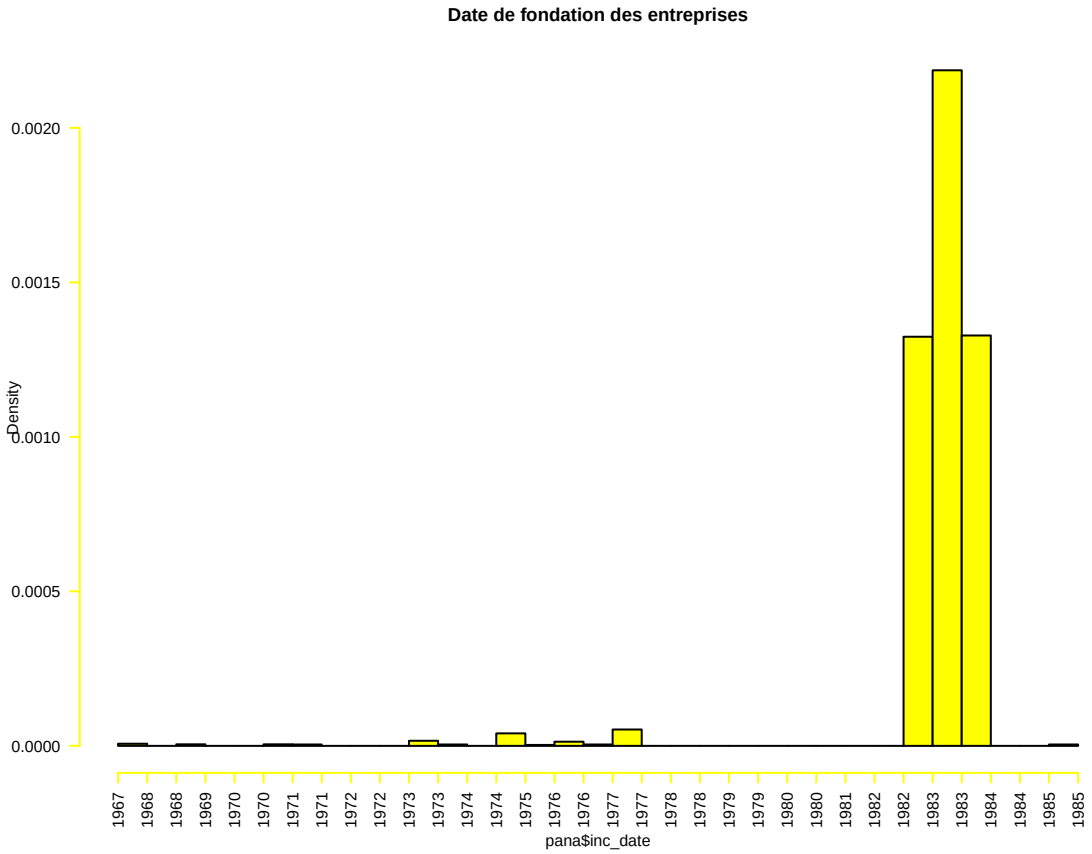
Date de création et durée des entreprises

Pour toutes

```
str(pana$inc_date)
```

```
##  chr [1:10003] "1982-11-09" "1982-11-09" "1982-11-09" ...
```

```
pana$inc_date <- as.Date(pana$inc_date)
par(mar=c(4,5,3,6), cex=0.5)
hist(pana$inc_date, breaks=25, las=2, col="yellow", main="Date de fondation des entreprises")
```

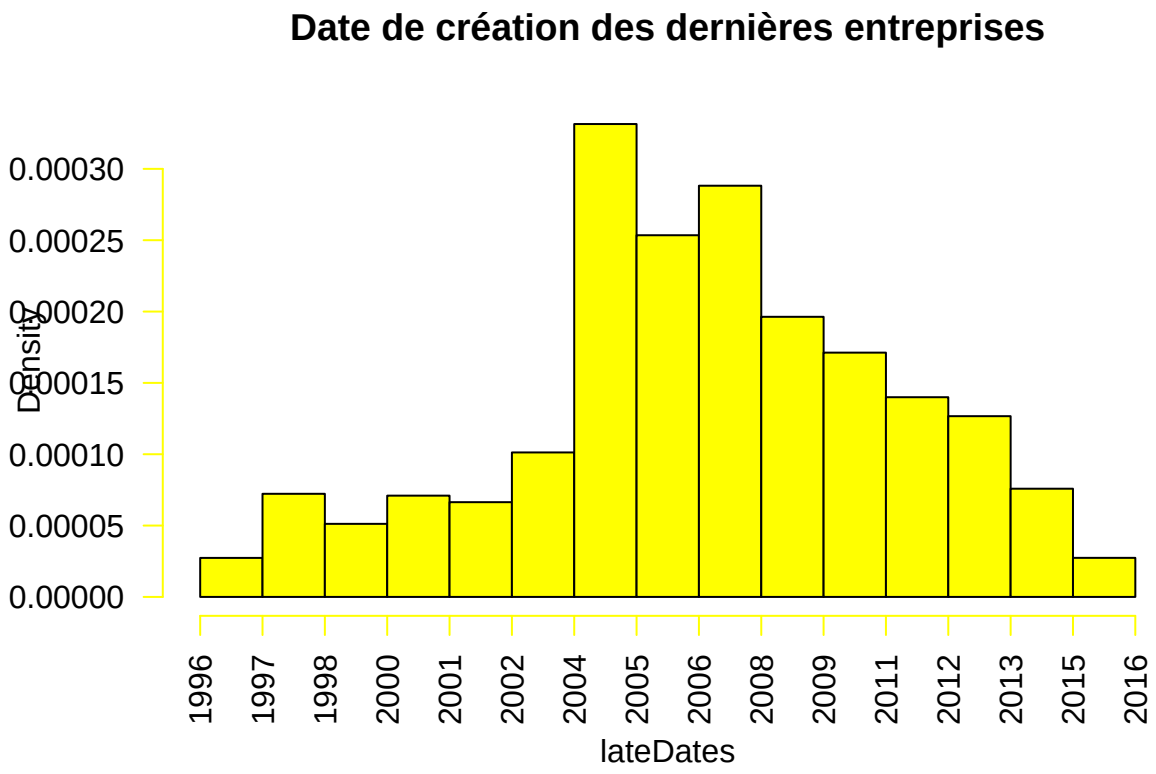


Pour les plus récentes

```
sortedDates <- sort(panama$inc_date)
lateDates <- sortedDates[150000:528998]
str(lateDates)
```

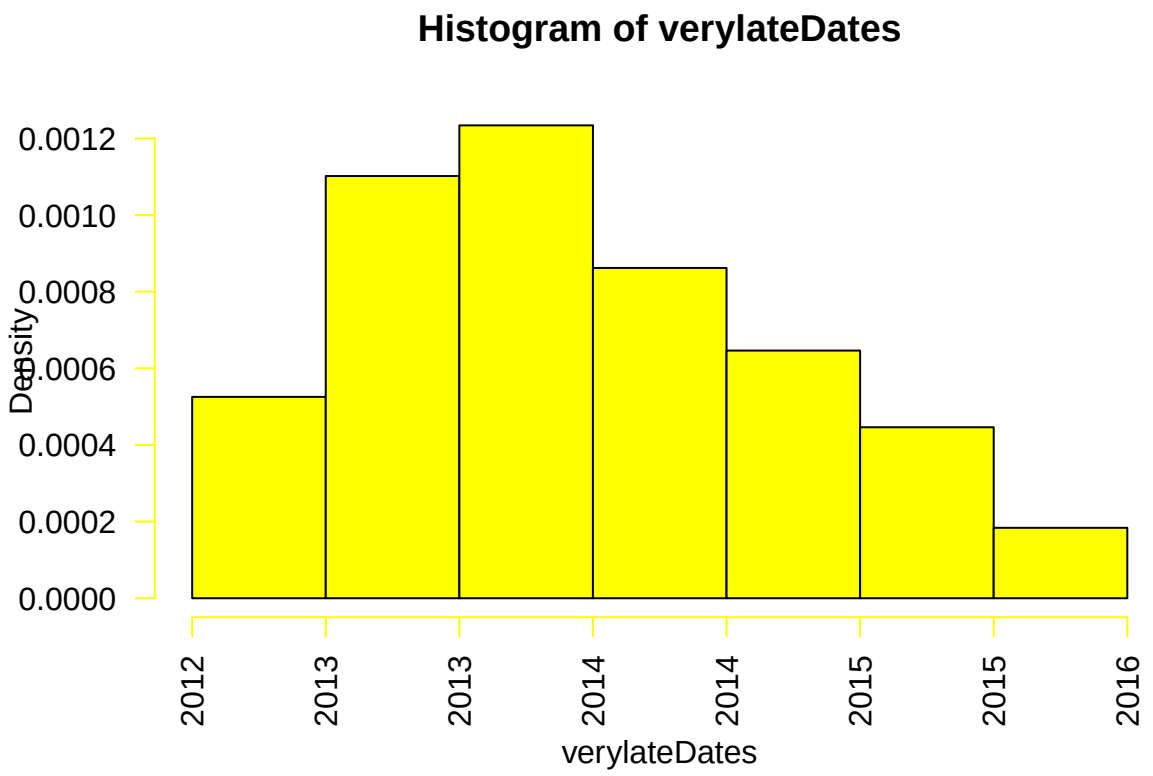
```
##  chr [1:378999] "1997-01-10" "1997-01-10" "1997-01-10" ...
```

```
lateDates <- as.Date(lateDates)
hist(lateDates, las=2, breaks=25, col="yellow", main="Date de création des dernières entreprises")
```



Pour les très récentes

```
verylateDates <- sortedDates[500000:528998]
verylateDates <- as.Date(verylateDates)
hist(verylateDates, breaks=6, las=2, col="yellow")
```



Durée d'existence des entreprises

```
# changement des variables en format date
pana$dissolved_date <- as.Date(pana$dissolved_date)
pana$inc_date <- as.Date(pana$inc_date )
#Duree d'existence des entreprises
pana$duration <- pana$dissolved_date - pana$inc_date

# Affiche les premières valeurs
head(pana$dissolved_date)
```

```
## [1] "1990-01-05" "1990-01-05" "1990-01-05" "1990-01-05" "1990-01-05"
## [6] "1990-01-05"
```

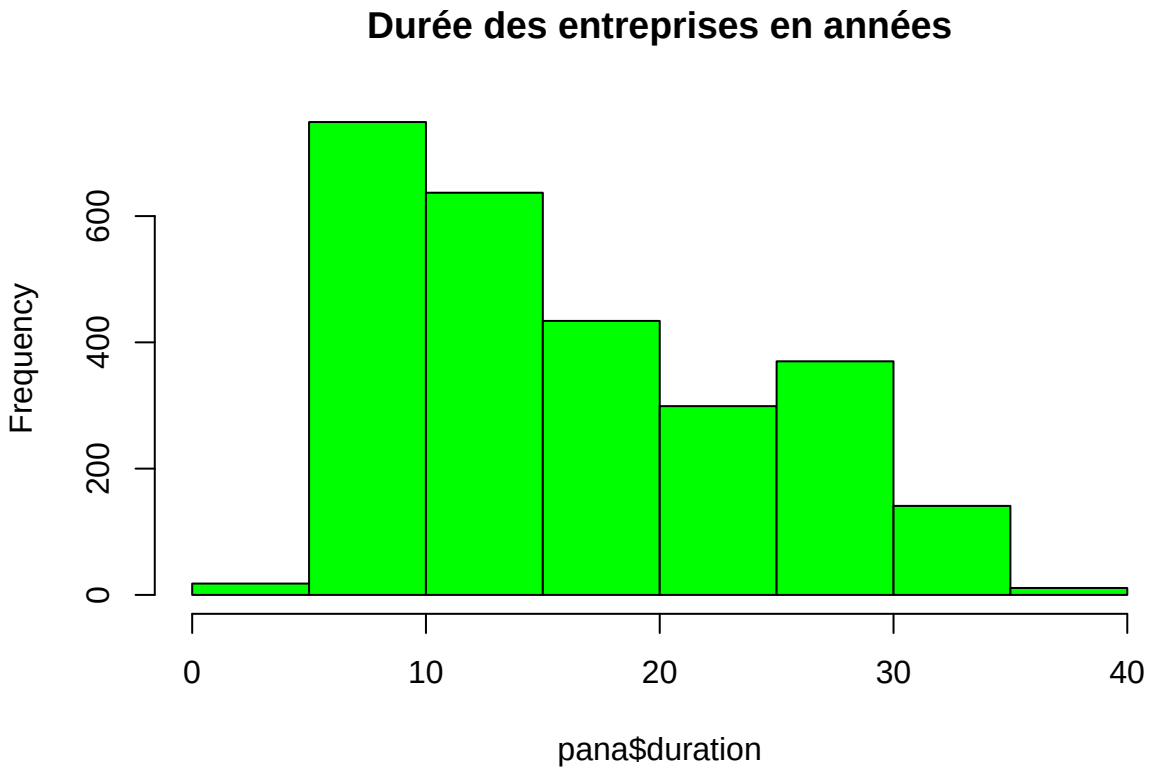
```
head(pana$inc_date)
```

```
## [1] "1982-11-09" "1982-11-09" "1982-11-09" "1982-11-09" "1982-11-09"
## [6] "1982-11-09"
```

```
# Durée d'existence des entreprises en jours
str(pana$duration)
```

```
## Class 'difftime'  atomic [1:10003] 2614 2614 2614 2614 2614 ...
##   ..- attr(*, "units")= chr "days"
```

```
pana$duration <- as.numeric(pana$duration)
pana$duration <- pana$duration/365.25
hist(pana$duration, col="green", main="Durée des entreprises en années")
```



Un peu plus d'information sur ces durées

```
mean(pana$duration, na.rm=TRUE)
```

```
## [1] 15.98575
```

```
median(pana$duration, na.rm=TRUE)
```

```
## [1] 14.5462
```

```
max(pana$duration, na.rm=TRUE)
```

```
## [1] 38.41752
```

```
sd(pana$duration, na.rm=TRUE)
```

```
## [1] 7.989974
```

```
summary(pana$duration, na.rm=TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.004   8.966  14.550  15.990  22.830  38.420    7344
```

Ensuite on peut publier notre travail sur gitHub ou rpubs

Un bug, ou vous êtes bloqués ? Vous pouvez chercher sur le site ou demander à la communauté sur stackoverflow