CRM 205 Cartographie

STÉPHANIE LEFEBVRE

PLAN

I. Partie A : Logiciel
II. Partie B : Données

Partie A : Logiciel

- Comparatif logiciels statistiques
- Comparatif logiciels SIG
- Conclusion : Pourquoi utiliser R ?
- Présentation détaillée de R

Partie B : Initiation à R

- Définitions
- Problèmes
- Géocodage
- Pourquoi utiliser des données géographiques ?

Partie A : Logiciel

Comparatif logiciels statistiques

SPSS

- Développé par **IBM**
- Langage dit de « syntaxe » propre à SPSS ou **interface graphique** avec menus déroulants => **utilisation simple**
- Disponible sous Windows, macOS et Unix, **Payant**
- **Modules supplémentaires** peuvent s'ajouter ex : data preparing algorithme de discrétisation des variables continues ;
- **Version clone gratuite : PSPP** http://pspp.awardspace.com/

SAS / STAT

- Développé par **SAS Institute**
- Langages SAS de 3 types : procédural habituel, macro et IML
- Disponible sous Windows, macOS et Unix, **Version gratuite**
- Développer par une entreprise, les **mises à jour logiciel** ont lieux **tous les 2 ans** environ (idem pour SPSS)

En sept 2017 : **1200 commandes** (soit 150 fois moins que R)

R

- Développé par la **communauté scientifique** (universitaires) et par des **volontaires** => OpenSource
- **Langage R** mais des packages rendent exploitable les fichiers développés dans **d'autres langages**
- Disponible sous Windows, macOS et Unix, **Gratuit**
- **Fonctions complémentaires** nombreuses et en développement perpétuel : **Packages**

En sept 2017 : plus de **11300 packages** soit environ **180 000 fonctions**

Comparatif logiciels Système d'Information Géographique (SIG)

ESRI France : ArcGIS

- Outil très **puissant**, peut traiter un nombre très important de données (**données massives**);
- Destiné aux **entreprises** ou collectivités ;
- **Payant** / Version d'essai gratuite (Voir avec Jean-Luc BESSON)
- Sur le site : **Conférences**, **Web séminaire**, **Dossiers thématiques** en accès libre sur l'utilisation de données géographique

Livre de cartes :

https://fr.calameo.com/read/000196594b6b56742c771

QGIS

- Equivalent de ArcGIS mais en **OpenSource**
- Open Source donc gratuit et disponible pour tous
- https://www.qgis.org/fr/site/forusers/download.html

R

- R n'est pas un logiciel SIG mais l'est devenu grâce au **développement de packages**
- Les fond de cartes peuvent être **des images**, ou des **fichiers issus de Google, OpenMapStreet et Geofla**
- **Package** le plus utilisé dans notre cours « **Cartography** » mais il en existe d'autres tels que « **tmap** » ou « **ggmap »**

Conclusion : Pourquoi utiliser R ?

- **Un seul logiciel** pour le **traitement statistique** des données et la **réalisation de cartes** ;
- **Gratuit** (existe des versions payantes), **très performant** (le plus riche en nombre de fonctions), **constamment mis à jour** (3 mises à jour importantes en 2 ans), **très puissant** (peu gérer un grand nombre de données : base SQL);
- **Open Source** : **transparent,** tous les codes sont accessibles (on sait comment sont fait les calculs, ce que comportent les fonctions, …);
- **Compatible avec les autres langages** et disponible sur tous les systèmes d'exploitation (Windows, macOS, Linux)
- *Inconvénient : Apprentissage d'un langage ce qui peut être rédhibitoire pour les novices*

Installation R

- Téléchargement de R : sur le site CRAN : https://cran.r-project.org/
- Téléchargement Rstudio : https://www.rstudio.com/products/rstudio/download/

RGui (64-bit)

Fichier Edition Packages Fenêtres Aide

R Console

```
> # Set a custom color palette
> cols <- carto.pal(pal1 = "blue.pal", n1 = 2, pal2 = "red.pal", n2 = 4)
>
> # Plot symbols with choropleth coloration
> propSymbolsChoroLayer(spdf = nuts2.spdf,
+                       df = nuts2.df,
+                       var = "pop2008", #  field in df to plot the symbols siz$
+                       inches = 0.1, # set the symbols sizes
+                       "cagr", #  field in df to plot the colors
+                       ols, # symbols colors
+                       = c(-2.43,-1,0,0.5,1,2,3.1), # breaks
+                       = "grey50",  # border colors of the symbols
+                       .75, # symbols width
+                       legend.var.pos = "topright", # size legend position
+                       legend.var.values.rnd = -3, # size legend value roundinf
+                       legend.var.title.txt = "Total Population", # size legen$
+                       legend.var.style
+                       legend.var2.pos
+                       legend.var2.titl
>
> # layout
> layoutLayer(title = "Demographic trend
+             sources = "Eurostat, 2011"
+             author = "cartography", fr
>
```

2 - Console

Sans titre - Editeur R

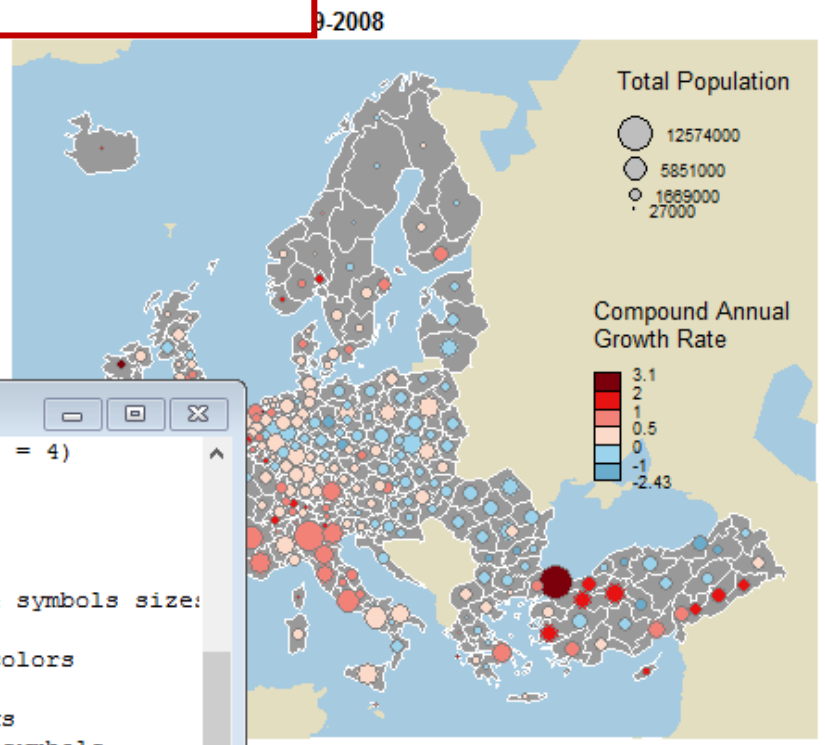
```
cols <- carto.pal(pal1 = "blue.pal", n1 = 2, pal2 = "red.pal", n2 = 4)

# Plot symbols with choropleth coloration
propSymbolsChoroLayer(spdf = nuts2.spdf,
                      df = nuts2.df,
                      var = "pop2008", #  field in df to plot the symbols size
                      inches = 0.1, # set the symbols sizes
                      var2 = "cagr
                      col = cols,
                      breaks = c(-
                      border = "gr
                      lwd = 0.75,
                      legend.var.pos = "topright", # size legend position
                      legend.var.values.rnd = -3, # size legend value roundinf
                      legend.var.title.txt = "Total Population", # size legend
                      legend.var.style = "e", # size legend type
                      legend.var2.pos = "right", # color legend position
                      legend.var2.title.txt = "Compound Annual\nGrowth Rate")
```

1 – Script

R Graphics: Device 2 (ACTIVE)

3 - Graphique



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

GraphiqueFrequence.R | Test_Pearson_Student.R | emeutes.R | Untitled2* | cartography.R

```
166                 legend.pos = "right", legend.frame = TRUE,
167                 legend.title.txt = "Number of Aggreements\n(regional level)",
168                 add = TRUE)
169
170 # Plot the layout
171 layoutLayer(title = "International Twinning Agreements Between Cities",
172             author = "ca                ources: Adam Ploszaj & Wikipedia,
173             scale = NULL                TRUE, col = NA,
174             coltitle = "
175
176 ## ----propchoroLayer, fig.height=5, fig.width=7, margin=TRUE--------------
177 # Compute the compound annual growth rate
178 nuts2.df$cagr <- (((nuts2.df$pop2008 / nuts2.df$pop1999)^(1/9)) - 1) * 100
179
180 # Plot a layer with the extent of the EU28 countries with only a background color
181 plot(nuts0.spdf, border = NA, col = NA, bg = "#A6CAE0")
182 # Plot non european space
183 plot(world.spdf, col  = "#E3DEBF", border = NA, add = TRUE)
184 # Plot Nuts2 regions
185
```

1 – Script

211:1 propchoroLayer, fig.height=5, fig.width=7, margin=TRUE R Script

Console | Terminal

~/R/win-library/3.4/cartography/doc/

```
+                     lwd = 0.75, # symbols width
+                     legend.var.pos = "topright", # size legend position
+                     legend.var.values.rnd = -3, # size legend value roundinf
+                     legend.var.title.txt = "Total Population", # size legend title
+                     legen                e legend type
+                     legen                color legend position
+                     legen                ound Annual\nGrowth Rate") # legen
d title
>
> # layout
> layoutLayer(title = "Demographic trends, 1999-2008", coltitle = "black",
+             sources = "Eurostat, 2011", scale = NULL,
+             author = "cartography", frame ="", col = NA)
>
```

2 - Console

Environment | History | Connections

Import Dataset List

Global Environment

Data	
nuts2.df	310 obs. of 11 variables
twincities.spdf	687
Values	
cols	chr F28278" "…

3 - environnement

Files | Plots | Packages | Help | Viewer

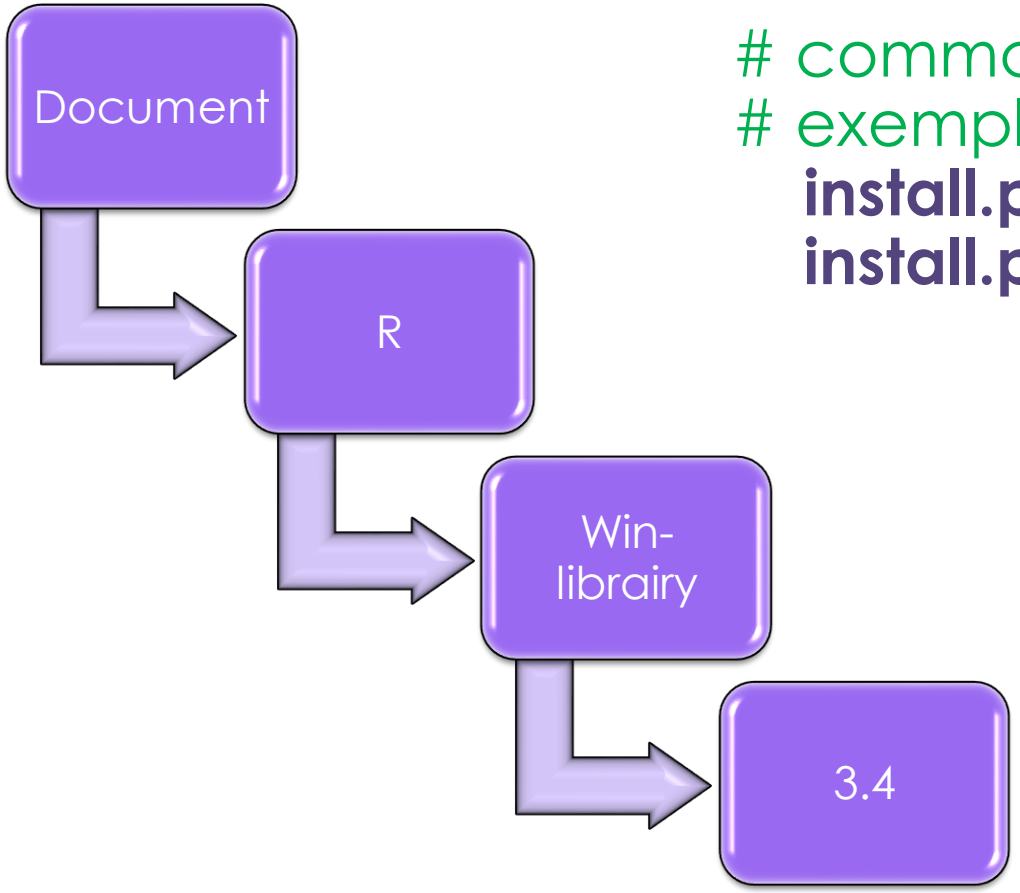
Zoom Export Publish

4 - Graphique

Interface Rstudio

- 1 **– Script** : contient le code source qui sera exécuté, les commentaires sont précédés d'un dièse et s'affiche en vert
- 2 – **Console** : fenêtre dans laquelle s'exécute le code, les messages d'erreurs y apparaissent en rouge tout comme les avertissements
- 3 – **Environnement** : contient toutes les données ainsi que les commandes qui ont été exécuté
- 4 – **Graphique** :
  - Plot : affiche les graphiques simples
  - Viewer : affiche les cartes dynamiques (package « leaflet »)
  - Packages : liste des packages qui ont été installé
  - Help : permet d'obtenir de l'aide en tapant un mot clef dans la barre de recherche

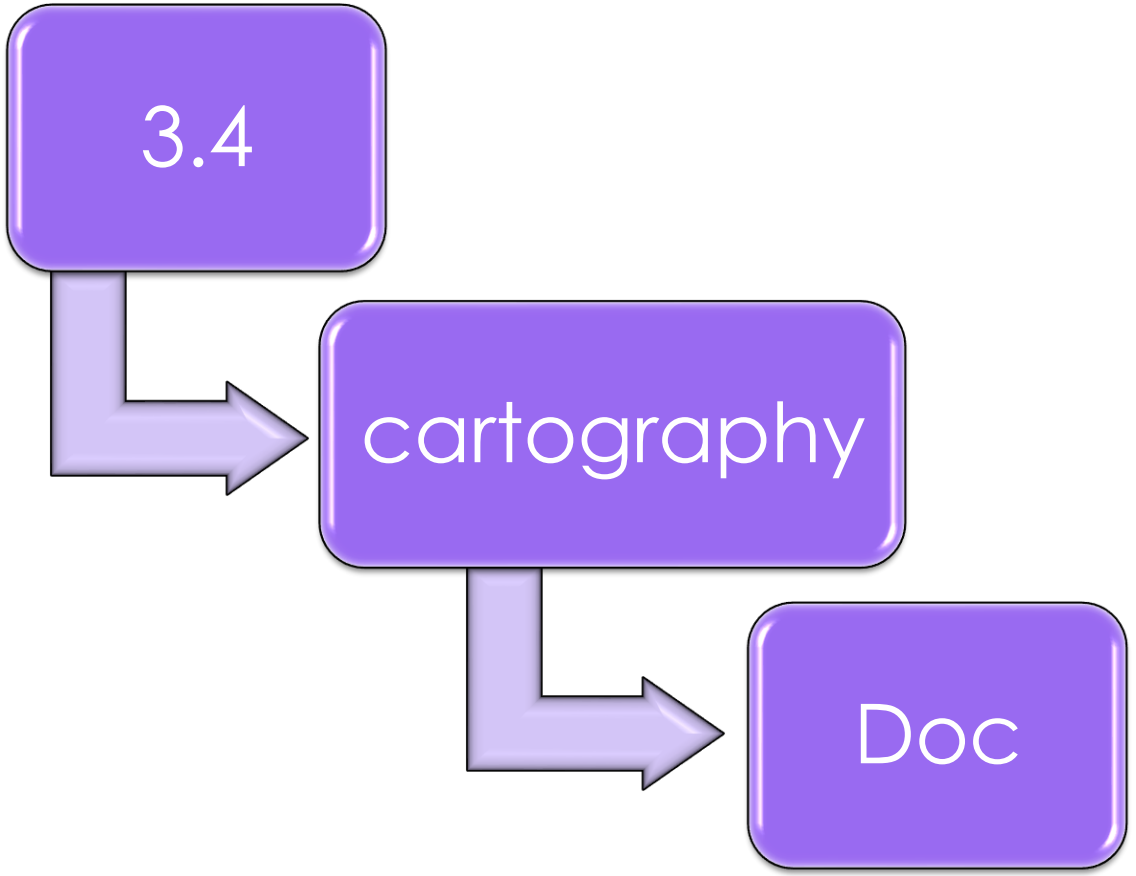
Installation packages




```
# commande installation
# exemple package « cartography »
   install.packages( ''cartography'' )
   install.packages( ''OSRM'' )
```

- Cartography
- GGplot2
- Skyni
- Etc.

Contenu du package



Script exemple permettant de découvrir les fonctions du package « cartography »

- Cartography.R

Partie B : Initiation à R

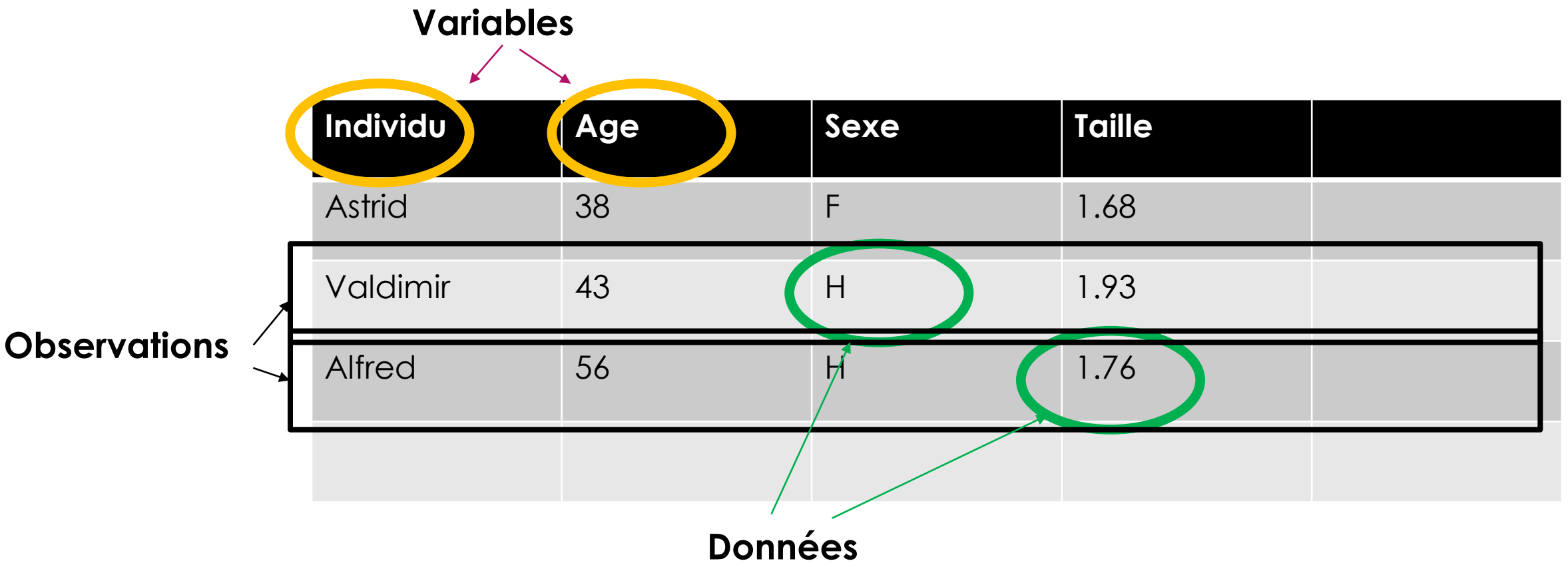
Manipulation sous R

- Qu'est-ce qu'un objet ?
- Création d'un vecteur
- Création d'un data frame (tableau de données organisées)

Les différents types d'objets dans R

Types d'objets	Données
Null (Objet vide)	NULL
Logical (booléen)	True ou T et False ou F
Numéric (valeur numérique)	1, 2.345 , pi
Character (chaîne de caractères)	'bonjour' , ''oui''

Définitions



Définitions

- Données
  - **Ce qui est connu et admis, et qui sert de base, à un raisonnement, à un examen ou à une recherche**. *Toute question de politique intérieure doit être vidée d'après les données de la statistique départementale* (Proudhon, *Propriété,* 1840, p. 340)
  - **Ensemble des indications enregistrées** en machine pour **permettre l'analyse et/ou la recherche automatique des informations** (Cros-Gardin 1964)
- Fichier de données
  - Un fichier est un traitement de données qui **s'organise dans un ensemble stable et structuré de données**.

Important : sous R nous allons travailler avec **des fichiers aux formats .CSV**

Forme des fichiers .CSV

Règles à respecter impérativement :

- Première ligne du fichier : **NOM DES VARIABLES**
- Le nom des fichiers et des variables **:**
  - **Aucun caractères spéciaux**
  - **Doit commencer par une lettre**
  - **Ne doit pas contenir d'espace** : remplacer les espace par _

Base de la programmation sous R

> Installer un package [install.package]
> Changement du répertoire courant [setwd]
> Chargement du package [library]
> Lecture du fichier CSV [read.csv]
> Lecture du fichier Excel (1ère feuille, nom de variable sur 1ère ligne) [read.xlsx]
> Statistiques descriptives sur l'ensemble des variables [summary]
> ls() liste le contenu de la mémoire
> Opérateur d'affectation **<-**

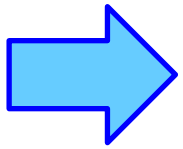
Comprendre la structure d'un data frame (ensemble de données) [2/2]

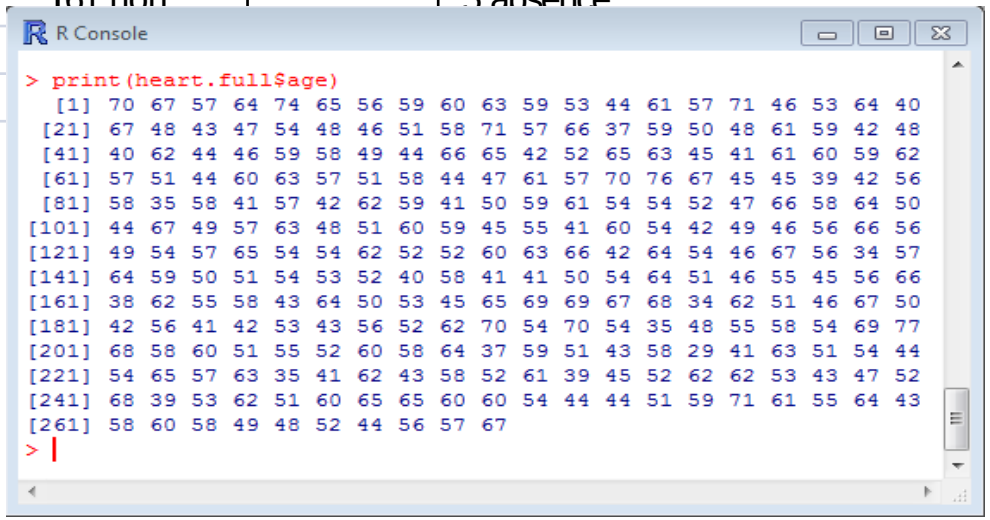
Data.frame = collection de variables
Accès aux variables (colonnes) avec $

heart.full

$age	$sexe	$typedouleu	$sucre	$tauxmax	$angine	$depression	$coeur
70	masculin	D	A	109	non	24	presence
67	feminin	C	A	160	non	16	absence
57	masculin	B	A	141	non	3	presence
64	masculin	D	A	105	oui	2	absence
74	feminin	B	n A	C B	121 oui	142 oui	2 absence
65	masculin	D	n A	D A	140 non	142 oui	4 absence
56	masculi		n	D A		170 non	
59	masculi			D A		154 non	
60	masculi		n	D A		161 non	
63	feminin						
59	masculi						

12 presence
40 presence
5 absence





Problèmes principaux

Problèmes

- Fichier incomplet : Valeur manquante
- Nom des variables différent d'un fichier à l'autre
- Deux variables dans une même case
- Fusion des fichiers

Solutions

- Représenter par « NA »
- Renommer [mutate]
- Séparer [separate]
- Joindre : [join] ou [merge]

A noter : il existe 4 variations pour la fonction join [inner-join] / [left-join] / [right-join] / [full-join] /

Pourquoi utiliser des données géographiques ?

Source : « Les données géographiques souveraines » - rapport du gouvernement – Juillet 2018

- Historiquement, la **donnée géographique** entretient un **lien très étroit avec l'exercice de la souveraineté**, dans sa **dimension militaire.** Puis de l' **homogénéisation du territoire**
- De nos jours, la puissance publique **acquiert**, **produit et mobilise quotidiennement des données géographiques** et plus largement **des données géolocalisées à l'appui de ses décisions et de son action**, dans des domaines aussi variés que :

  - la **défense nationale**, http://www.leparisien.fr/high-tech/une-application-de-fitness-devoile-des-bases-militaires-americaines-secretes-29-01-2018-7529169.php (applications STRAVA, POLAR Fitness);
  - la **sécurité**, (prévention situationnelle)
  - la **prévention des risques naturels** et **technologiques**, https://www.cnil.fr/sites/default/files/atoms/files/cnil_cahiers_ip5.pdf
  - la fiscalité,
  - l'**urbanisme**, (théorie des activités routinières - Cohen et Felson)
  - les **transports**, http://www.driea.ile-de-france.developpement-durable.gouv.fr/IMG/pdf/gares_-Theorie_et_pp_de_prevention_situationnelle.pdf
  - la santé, …

Géocodage

- Définition : Le **géocodage** consiste à affecter des coordonnées géographiques (longitude/latitude) à une adresse postale. Les coordonnées géographiques permettent de situer chaque adresse sur une carte numérique via un Système d'Informations Géographiques (SIG)
- Si les **données sont en France** : https://adresse.data.gouv.fr/csv
- Si les **données sont hors de France** : nécessite de passer par R via des api Google ou OpenStreetMap(Package « Photon ») https://data.hypotheses.org/317

Emplacement des commissariats en

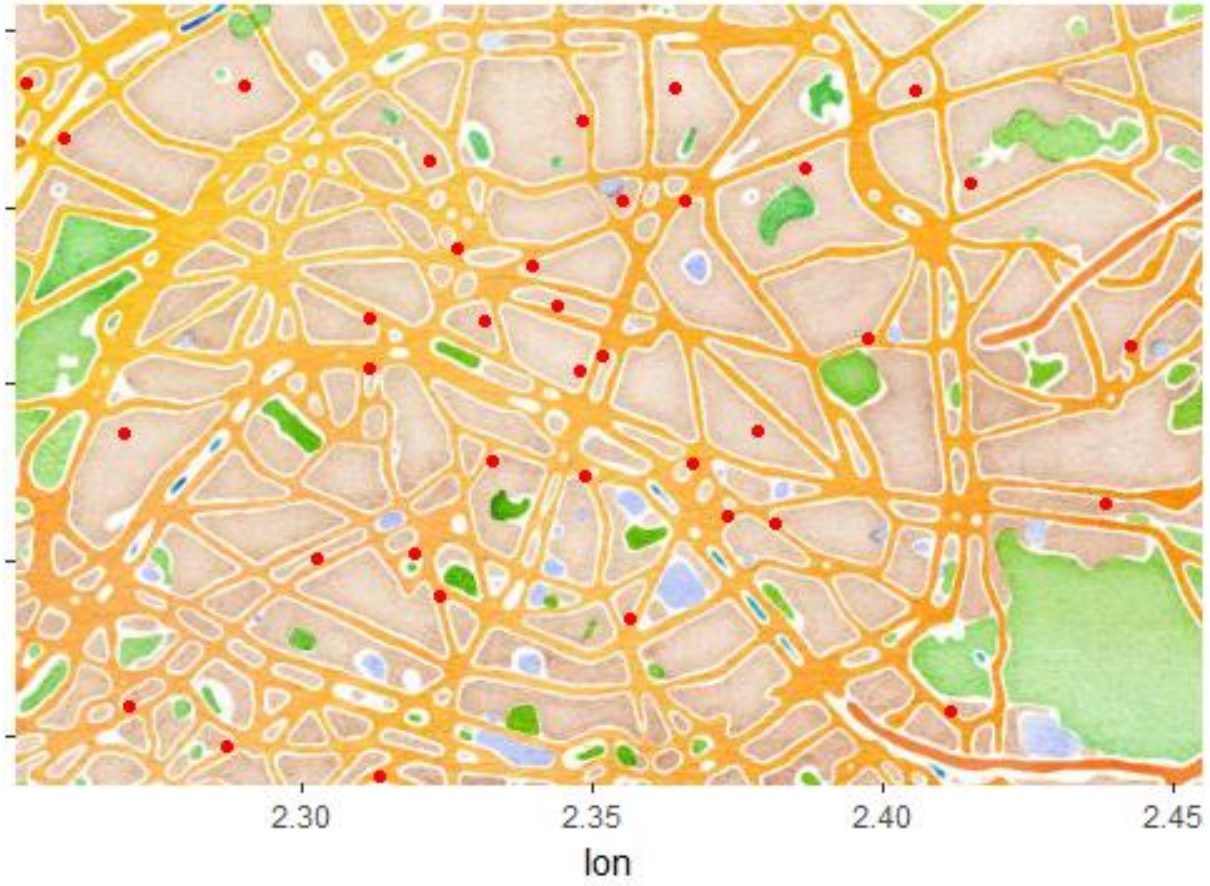
- Sur data.gouv.fr, téléchargement du fichier .CSV recensant les commissariats d'Ile-de-France

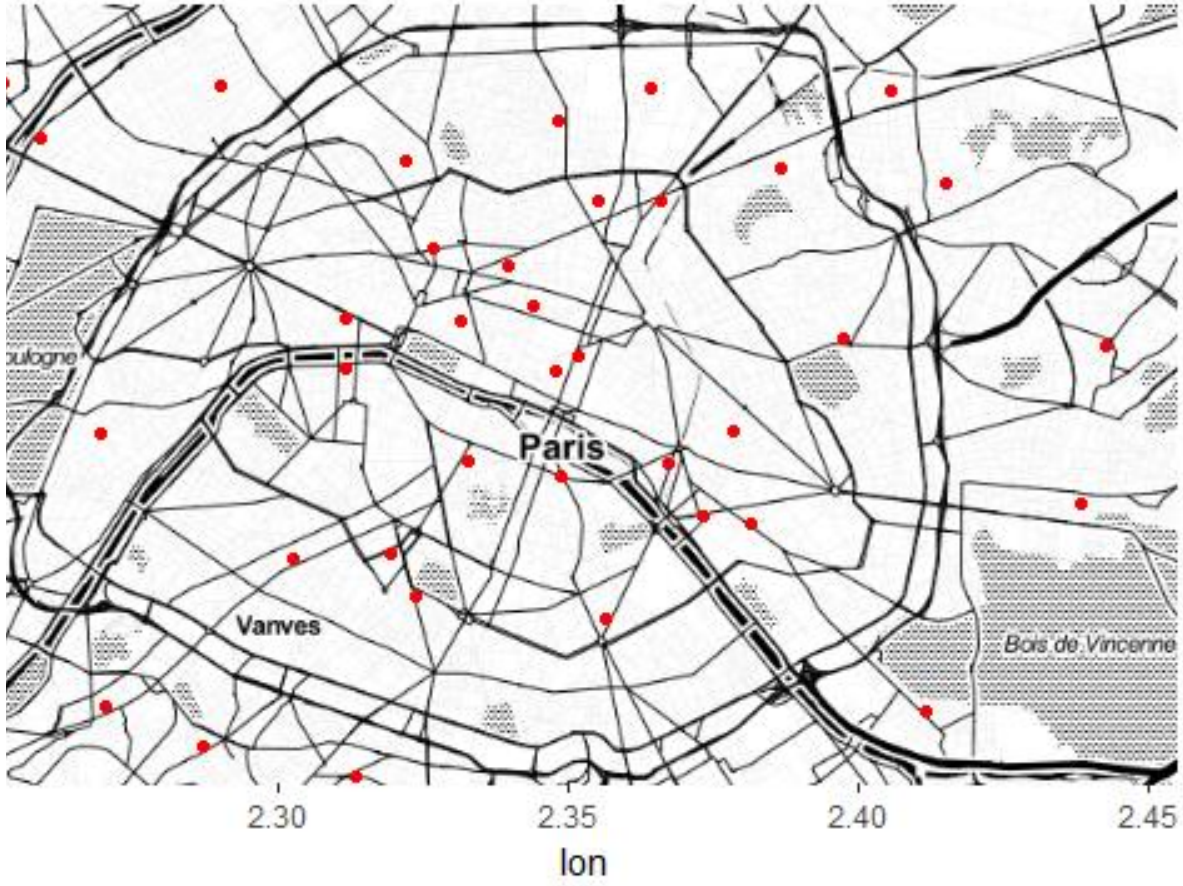
	A	B	C	D	E
1	name,C,255	descriptio,C,255			
2	LA DEFENSE	Commissariat central			
3	LA GARENNE COLOMBES	Commissariat central			
4	SURESNES	Commissariat central			
5	VANVES / Malakoff	Commissariat central			
6	AULNAY-SOUS-BOIS	Commissariat central			
7	BOBIGNY / Noisy-le-Sec	Commissariat central			
8	GAGNY	Commissariat central			

- **Géocodage** du fichier .CSV

	A	B	C	D
1	name	description	latitude	longitude
2	LA DEFENSE	Commissariat central	48.892096	2.2445963
3	LA GARENNE COLOMBES	Commissariat central	48.906106	2.2406057
4	SURESNES	Commissariat central	48.8684003	2.2250844
5	VANVES / Malakoff	Commissariat central	48.818878	2.2871037
6	AULNAY-SOUS-BOIS	Commissariat central	48.9146017	2.3880711
7	BOBIGNY / Noisy-le-Sec	Commissariat central	48.9098037	2.4511192
8	GAGNY	Commissariat central	48.8833012	2.5329109

Affichage commissariats IDF - GGMAP





Partie C : Gestion de projet

Gestion d'une étude de data mining

- Définition « data mining » :

  l'analyse de données depuis différentes perspectives et le fait de **transformer ces données en informations utiles, en établissant des relations entre les données** ou **en repérant des patterns**

- **Association** – chercher des patterns au sein desquelles un événement est lié à un autre événement.
- **Analyse de séquence** – chercher des patterns au sein desquelles un événement mène à un autre événement plus tardif.
- **Classification** – chercher de nouvelles patterns, quitte à changer la façon dont les données sont organisées.
- **Clustering** – trouver et documenter visuellement des groupes de faits précédemment inconnus.
- **Prédiction** – découvrir des patterns de données pouvant mener à des prédictions raisonnables sur le futur. Ce type de data mining est aussi connu sous le nom d'analyse prédictive.

Gestion d'une étude de data mining

1. Définition des objectifs
2. Inventaires des données existantes
3. Collecte des données
4. Exploitation et préparation des données
5. Segmentation de la population
6. Élaboration et validation des modèles prédictifs
7. Déploiement des modèles
8. Formation des utilisateurs des modèles
9. Suivi des modèles
10. Enrichissement des modèles

Gestion d'une étude de data mining

1. Définition des objectifs
   - Choix du sujet => poser une problématique
   - Évaluation des tâches à réaliser :
   - Répartition du travail

   => réalisation d'un diagramme de Gantt ou ….

Gestion d'une étude de data mining

2. Inventaire des données existantes
   - Quelles données existent ?
   - Où les trouver ?
   - Sous quel format sont-elles ?
   - D'où proviennent-elles ?

Gestion d'une étude de data mining

3. Collecter les données
   - Télécharger des fichiers, bases de données « prêtes » à être utilisé;
   - Réaliser ses propres fichiers de données;

Attention : Le plus important est de connaître son fichier et les données qu'il contient.

Gestion d'une étude de data mining

4. Préparation et exploitation des données
   - Nettoyage : enlever les données en double, remplacer ou enlever les données manquantes;
   - Modifier le format ;
   - Modifier les entrées (inverser lignes / colonnes);
   - Modifier / uniformiser le nom des variables…