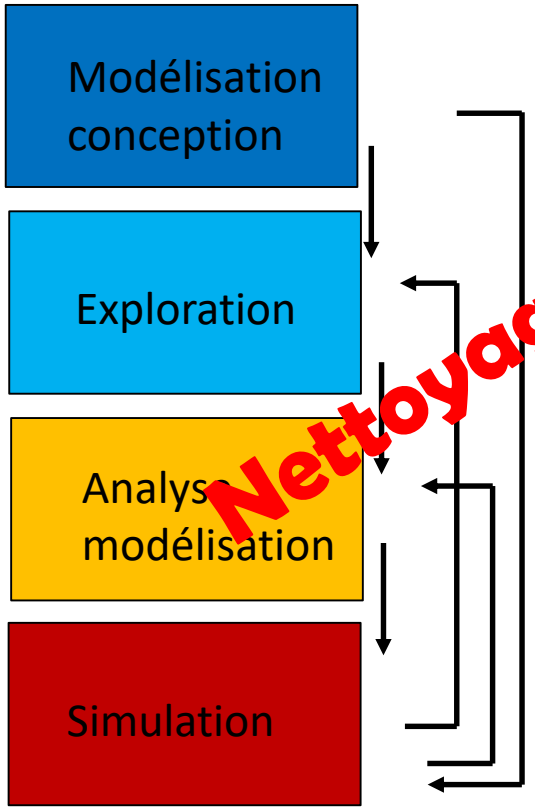
École thématique 2020
EXPLO-SHS
12-16 oct. – La Rochelle

MATE-SHS

Comment utiliser le nettoyage des données pour explorer, rendre compte d'un potentiel scientifique : le langage R

Hélène Mathian, Hughes Pécout

Cycle du traitement de la donnée



Nettoyage à tous les étages

Cycle/étapes d'analyse de la donnée (Mathian,Sanders)

Cycle science de la donnée (Trousse & al.)

Nettoyage ⇔ Qualité => cohérence

- « Le nettoyage de données est l'opération de détection et de correction d'erreurs présentes sur des données stockées dans des bases de données ou dans des fichiers. »
- Nettoyage est
  - une étape **d'amélioration de la qualité**
  - Mais c'est une étape de **compréhension de la donnée,**
- Passe par une boucle autour de

exploration <-> nettoyage<-> vérification

- Nettoyer, corriger … nécessite souvent de définir **un point de vue**!
  => notion d'incohérence

« La qualité des données au lieu de l'incohérence » lien

- Différencier la
  - **cohérence « externe/forme » -> considérations techniques liées au process**
  - et la **cohérence « interne/fond » -> considérations sémantiques/expertises….**

Explorer / Nettoyer

- Diagnostic => **explorer**

  - Visualiser – décrire (stat) -analyser

⇒ **Nettoyer**

➢ Adaptation

- recoder

➢ Correction

- Corriger
- Compléter -> estimer….

Exemples de méthodes sur la base de cohérence « technique »

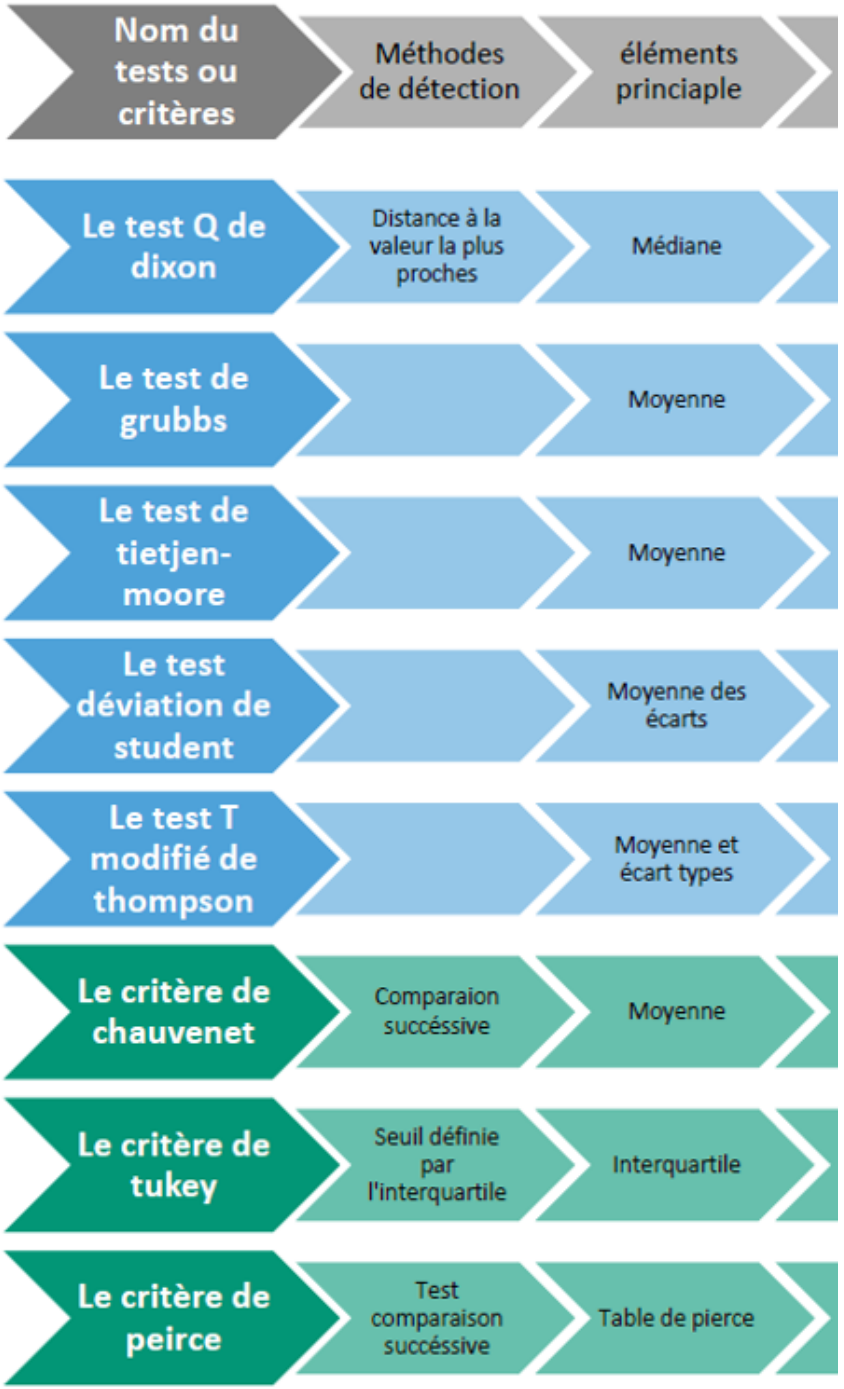
Analyses de colonnes (une à une)	Indicateurs	TDQ	DC	DP
Statistiques simples	Nombre total des valeurs	X	X	X
	Nombre de valeurs nulles	X	X	X
	Nombre de valeurs distinctes	X	X	X
	Nombre de valeurs doubles	X	X	X
	Table de fréquence	X		
Statistiques sur les chaînes	Longueur maximale des chaînes	X	X	X
	Longueur minimale des chaînes	X	X	X
	Longueur moyenne des chaînes	X	X	X
Statistiques sur les numériques	Valeur maximale des numériques	X	X	
	Valeur minimale des numériques	X	X	
	Moyenne des numériques	X	X	
	Écart type des numériques	X	X	
Statistiques sur les dates	Fréquence des années	X	X	
	Motif de fréquence de Date	X		
Format des chaînes	Motif de fréquence	X		
Nature des chaînes				
Langue des chaînes				
Nature de la langue des chaînes (Latin, Arabe)			X	
Nombre de chaînes valides syntaxique				
Nombre de chaînes valides sémantique				

TDQ= Talend Data Quality 8
DC= DataCleaner 9
DP= DatirisProfiler 10

Source: Qualité contextuelle des données : **détection et nettoyage guidés par la sémantique des données.**

able 2.11 – Tableau comparatif des différents outils de profilage (analyse de colonnes)

Détection de valeurs aberrantes « outliers »



Axel Fourneyron, 2020

La question des données manquantes

Ce que vous n'aurez pas

- Systématique (corrélée à un facteur ?) ou aléatoire ?

- Quelques méthodes automatiques de reconstruction

Méthode	Commentaire	Mécanique/ Sémantique
Moyenne / médiane	Diminue la variabilité	M
Par tirage conditionnel	1- plus proches voisins (moyenne des k plus proches voisins sur les p variables renseignées) 2- Classification sur p mêmes variables renseignées et moyenne conditionnelles par classe 3- modèle de régression sur p variables (valeurs prédites)	1-M 2-S
Par moyenne partielle	Avoir une variable groupe ayant un sens fort => utiliser les mêmes méthodes mais conditionnées par le groupe	

Caractéristiques et TP

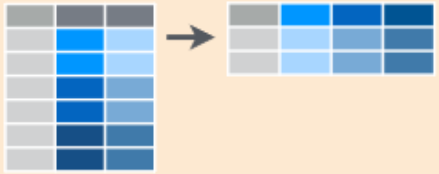
Dimensions	Aspect	Méthode	Exemple
Statistique	Distribution	outliers	MetropLyon
	Sémantique		DansMaRue
	Données manquantes		X
Cohérence temporelle		Explorer les cycles	DansMaRue
Cohérence spatiale		Explorer les niveaux	DansMaRue
Cohérence sémantique		Analyse des sentiments	TrumpTweet

Manipulation des données: déployer et ranger dplyr & tidyr

Data Wrangling
with dplyr and tidyr
Cheat Sheet

Reshaping Data - Change the layout of a data set

tidyr::**gather(cases, "year", "n", 2:4)**
Gather columns into rows.



tidyr::**spread(pollution, size, amount)**
Spread rows into columns.

tidyr::**separate(storms, date, c("y", "m", "d"))**
Separate one column into several.

tidyr::**unite(data, col, ..., sep)**
Unite several columns into one.

dplyr::**data_frame(a = 1:3, b = 4:6)**
Combine vectors into data frame (optimized).

dplyr::**arrange(mtcars, mpg)**
Order rows by values of a column (low to high).

dplyr::**arrange(mtcars, desc(mpg))**
Order rows by values of a column (high to low).

dplyr::**rename(tb, y = year)**
Rename the columns of a data frame.

Manipulation des données: déployer et ranger dplyr & tidyr

Data Wrangling
with dplyr and tidyr
Cheat Sheet

Subset Observations (Rows)

dplyr::**filter(iris, Sepal.Length > 7)**
Extract rows that meet logical criteria.

dplyr::**distinct(iris)**
Remove duplicate rows.

dplyr::**sample_frac(iris, 0.5, replace = TRUE)**
Randomly select fraction of rows.

dplyr::**sample_n(iris, 10, replace = TRUE)**
Randomly select n rows.

dplyr::**slice(iris, 10:15)**
Select rows by position.

dplyr::**top_n(storms, 2, date)**
Select and order top n entries (by group if grouped data).

Subset Variables (Columns)

dplyr::**select(iris, Sepal.Width, Petal.Length, Species)**
Select columns by name or helper function.

Manipulation des données: déployer et ranger dplyr & tidyr

Data Wrangling
with dplyr and tidyr
Cheat Sheet

Summarise Data

dplyr::**summarise(iris, avg = mean(Sepal.Length))**
Summarise data into single row of values.

dplyr::**summarise_each(iris, funs(mean))**
Apply summary function to each column.

dplyr::**count(iris, Species, wt = Sepal.Length)**
Count number of rows with each unique value of variable (with or without weights).

Summarise uses **summary functions**, functions that take a vector of values and return a single value, such as:

dplyr::**first**
First value of a vector.

dplyr::**last**
Last value of a vector.

dplyr::**nth**
Nth value of a vector.

dplyr::**n**
# of values in a vector.

dplyr::**n_distinct**
# of distinct values in a vector.

IQR
IQR of a vector.

min
Minimum value in a vector.

max
Maximum value in a vector.

mean
Mean value of a vector.

median
Median value of a vector.

var
Variance of a vector.

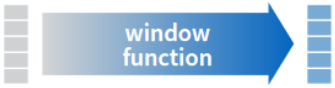
sd
Standard deviation of a vector.

Make New Variables

dplyr::**mutate(iris, sepal = Sepal.Length + Sepal. Width)**
Compute and append one or more new columns.

dplyr::**mutate_each(iris, funs(min_rank))**
Apply window function to each column.

dplyr::**transmute(iris, sepal = Sepal.Length + Sepal. Width)**
Compute one or more new columns. Drop original columns.



Mutate uses **window functions**, functions that take a vector of values and return another vector of values, such as:

dplyr::**lead**
Copy with values shifted by 1.

dplyr::**lag**
Copy with values lagged by 1.

dplyr::**dense_rank**
Ranks with no gaps.

dplyr::**min_rank**
Ranks. Ties get min rank.

dplyr::**percent_rank**
Ranks rescaled to [0, 1].

dplyr::**row_number**
Ranks. Ties got to first value.

dplyr::**ntile**
Bin vector into n buckets.

dplyr::**between**
Are values between a and b?

dplyr::**cume_dist**

dplyr::**cumall**
Cumulative `all`

dplyr::**cumany**
Cumulative `any`

dplyr::**cummean**
Cumulative `mean`

cumsum
Cumulative `sum`

cummax
Cumulative `max`

cummin
Cumulative `min`

cumprod
Cumulative `prod`

pmax
Element-wise `max`

pmin

Manipulation des données: dplyr & tidyr

Group Data

dplyr::**group_by(iris, Species)**

Group data into rows with the same value of Species.

dplyr::**ungroup(iris)**

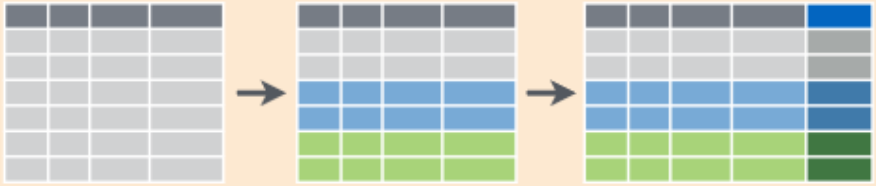
Remove grouping information from data frame.

iris %>% group_by(Species) %>% summarise(…)

Compute separate summary row for each group.

iris %>% group_by(Species) %>% mutate(…)

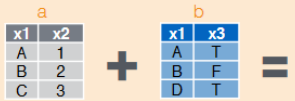
Compute new variables by group.



rstudio.com devtools::install_github("rstudio/EDAWR") for data sets Learn mo

Data Wrangling with dplyr and tidyr

Combine Data Sets



Mutating Joins

dplyr::**left_join(a, b, by = "x1")**
Join matching rows from b to a.

dplyr::**right_join(a, b, by = "x1")**
Join matching rows from a to b.

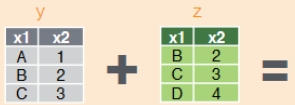
dplyr::**inner_join(a, b, by = "x1")**
Join data. Retain only rows in both sets.

dplyr::**full_join(a, b, by = "x1")**
Join data. Retain all values, all rows.

Filtering Joins

dplyr::**semi_join(a, b, by = "x1")**
All rows in a that have a match in b.

dplyr::**anti_join(a, b, by = "x1")**
All rows in a that do not have a match in b.



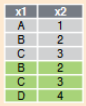
Set Operations

dplyr::**intersect(y, z)**
Rows that appear in both y and z.

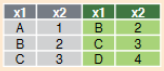
dplyr::**union(y, z)**
Rows that appear in either or both y and z.

dplyr::**setdiff(y, z)**
Rows that appear in y but not z.

Binding



dplyr::**bind_rows(y, z)**
Append z to y as new rows.



dplyr::**bind_cols(y, z)**
Append z to y as new columns.
Caution: matches rows by position.

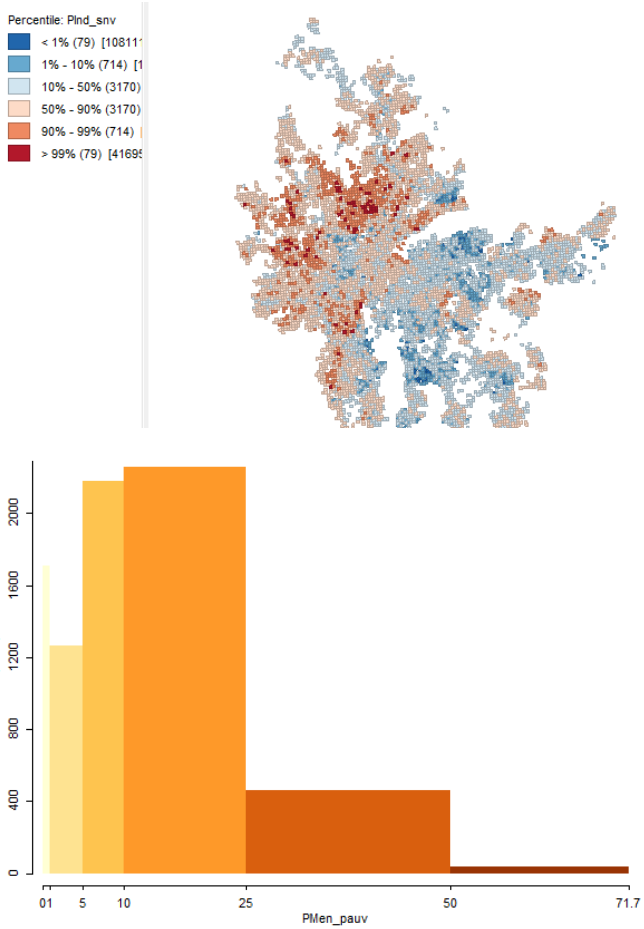
TP1

Métropole de Lyon – ségrégation socio-spatiale

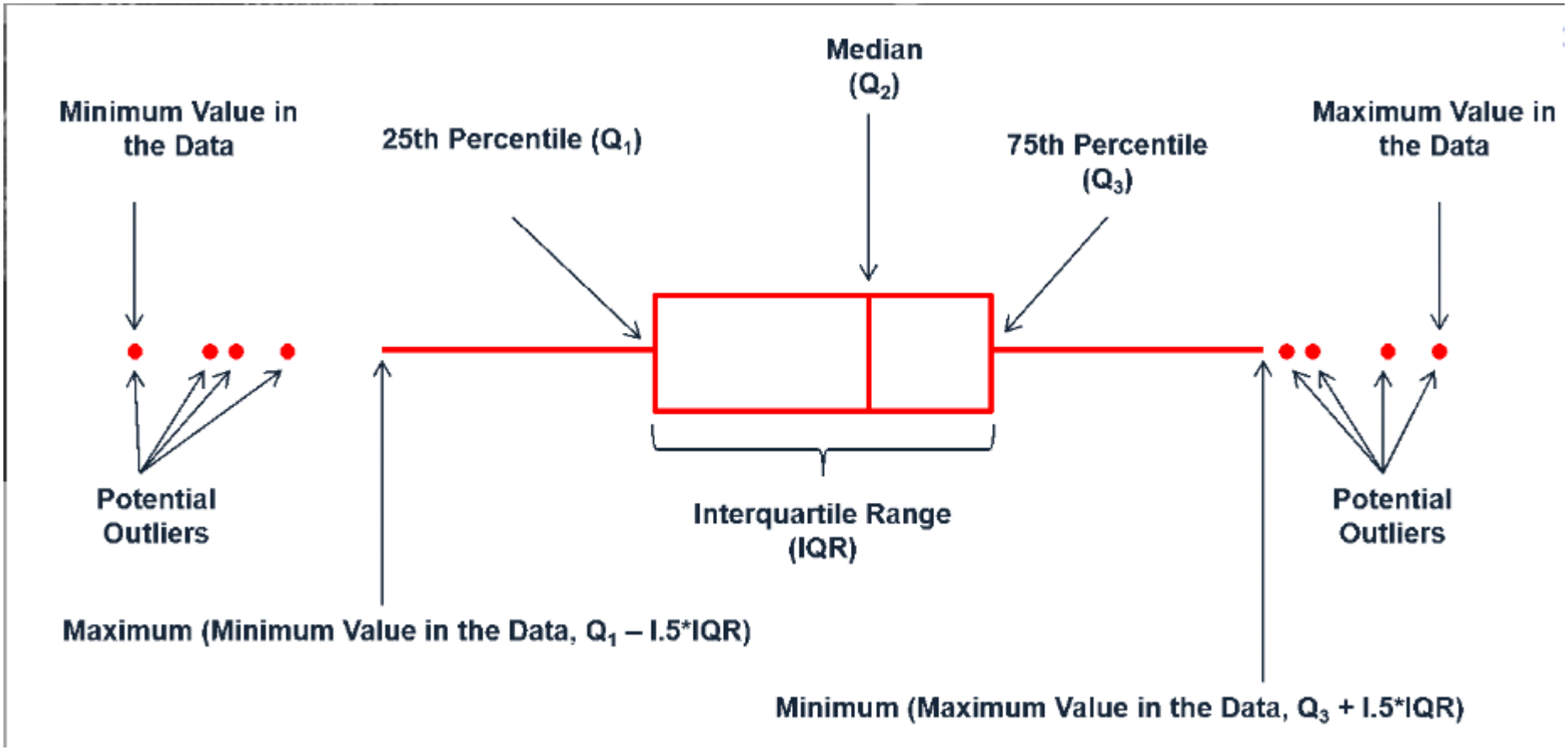
- À l'échelon des mailles de 200m (Carro200m2015GLyonW.shp)

- Détecter les outliers

Depcom	Code communal
Ind	Nombre d'individus du carreau
Men	Nombre de ménages du carreau
PMen_pauv	% de ménages pauvres (au seuil 60% de pauvreté)
PMen_1ind	% de ménages d'un seul individu
PMen_5ind	% de ménages de 5 individus ou plus
PMen_prop	% de ménages propriétaires
PMen_fmp	% de ménages monoparentaux
PMen_mais	% de ménages en maison
SurfLogMoy	Superficie moyenne des logements
PLog_av45	% de logements construits avant 1945
PLog_ap90	% de logements construits après 90
PLog_soc	% de logements sociaux
Ind_snvmoy	moyenne par individu des niveaux de vie winsorisés des individus
PInd_inf10	% d'individus demoins de 10 ans
PInd_11_17	% d'individus de 11 à 17 ans
PInd_18_24	% d'individus de 18 à 24 ans
PInd_80p	% d'individus de plus de 80 ans



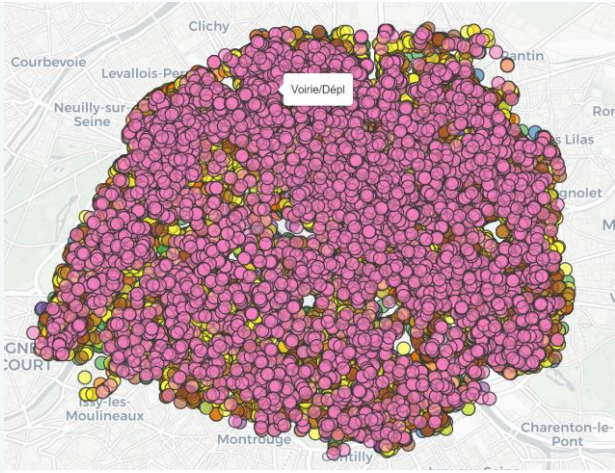
Détection des outliers
Les graphiques (1) « boxplot » ou « boite à moustache »



TP2

Données « Dans ma rue »

https://teleservices.paris.fr/dansmarue/

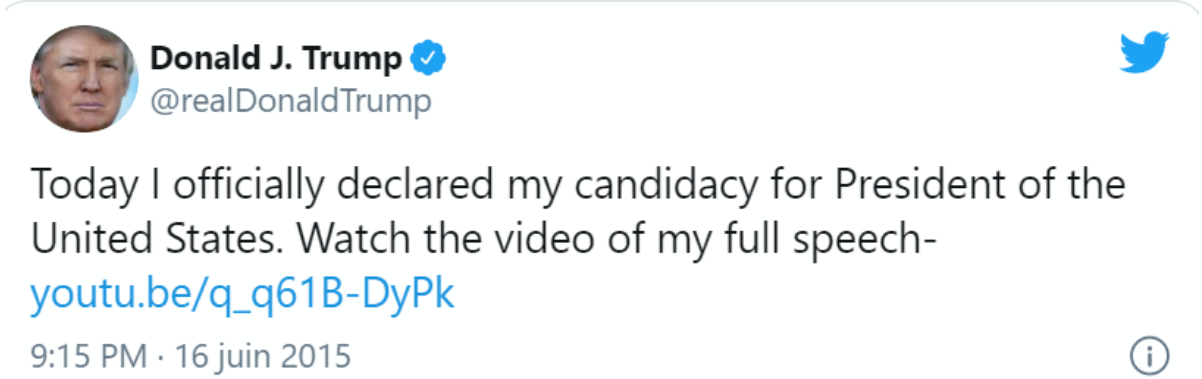


– données déclaratives- cadrées (saisie avec avec contraintes)
Sémantique, temps, espace

- explorer:
  - Définir le cadre- de quoi ça parle
  - Qu'est ce qui pourrait m'intéresser… quel est l''intérêt ;o) ?

TP3

Qui se cache derrière les tweets….



- #1398 tweets entre le "2015-12-14 20:09:15 UTC" et le 2016-08-08 15:20:44 UTC"
- #élections on eu lieu le 8 novembre 2016 et il s'était déclaré candidat en juin 2015

- *TP d'après http://varianceexplained.org/r/trump-tweets/*
  thanks to David Robinson