

# Analyse de Contenu et de Données Prémices

**Guillaume Metzler**  
guillaume.metzler@live.fr

Univ. Lyon, UJM-Saint-Etienne, CNRS, Institut d'Optique Graduate School, Laboratoire  
Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France

Printemps 2020



# A propos de moi

## Recherche :

- Docteur en Informatique (Computer Science), Apprentissage Machine
- Travaille autour de la problématique de la détection de fraudes
- Utilisation de données collectées par les entreprises (banques, enseignes, ...)
- Construction de modèles pour effectuer différentes tâches

**Objectif de la thèse :** construire des modèles pour détecter les personnes utilisant des faux chèques pour régler leurs achats.



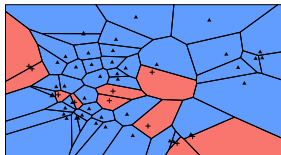
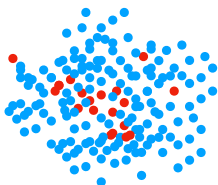
## Actuellement :

- Post-Doctorant au Laboratoire Hubert Curien
- Enseignements en Licence et Master (et aussi avec vous !)
- Poursuite des travaux de recherches sur *Imbalanced Learning*

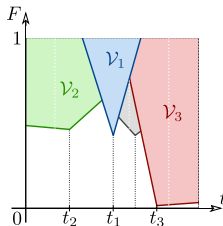
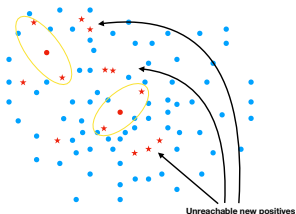
Je souhaite devenir **Maître de Conférence**

# Quelques exemples

Imbalanced dataset



Problème de minimisation :  $\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^m \ell(f(\beta, \mathbf{x}_i), y_i) + \lambda \|\beta\|$



Je travaille essentiellement avec des Données Numériques

Et vous dans l'histoire ? Pourquoi étudiez des données ?

## La notion de données est omniprésente dans notre quotidien !

- Emergence de l'informatique et son amélioration perpétuelle
- Facilité de récolter des données et de les faire circuler (enquête, sondage, inscription, réseaux sociaux, ...)
- Des capacités de stockage qui ne cessent d'augmenter
- Une capacité de traitement par des algorithmes de plus en plus en rapides

# A la question pourquoi

**Bref, les données sont importantes, notamment d'un point de vue financier !**

- L'achat et la revente de données prennent une part importante.
- Très utilisées d'un point de vue Marketing (publicité ciblée)
- Peuvent également être utilisées d'un point de vue RH.

**Pour autant il est difficile de fixer une réelle valeur à ses données, mais surtout ... on ne peut pas en faire n'importe quoi !**

Petite pensée pour Facebook ou autres sites revendants vos données personnelles.

# Une omniprésence des données





## Réutiliser :

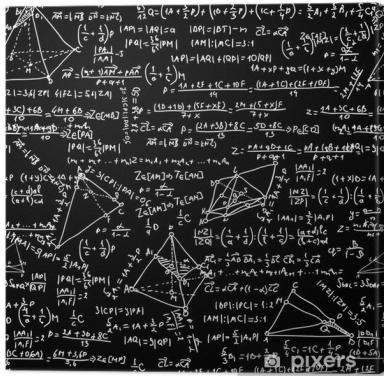
- Vos connaissances acquises en cours de **Marketing**
- Mobiliser vos connaissances en **Statistiques**

## Et vous donner les outils pour :

- Effectuer votre **Enquête Terrain**, séances pendant lesquelles vous allez acquérir des données et bien sûr les **étudier** pour répondre à une demande réelle de votre employeur, client, etc ...

# A propos du cours

Pas de craintes ! On ne va pas faire des maths dans ce cours, ni même utiliser des théories bizarres.



Quoi que ... mais non !

## Ce que vous allez faire :

- Apprendre à lire un tableau de données (il faut bien commencer par la base)
- Apprendre à analyser ses données (en extraire la substantifique moëlle) tris à plats, tris croisés, valeurs centrales, dispersion ou autres tests statistiques
- Rédiger une analyse des données (plus compliquée que ce qu'il n'y paraît)

Pour cela on utilisera le logiciel Sphinx Campus.

# Mais c'est quoi Sphinx Campus ?

- Un logiciel d'analyse de données (vous allez me dire ...on s'en doutait)
- Permet de concevoir ses propres enquêtes ou sondages
- Permet de visualiser des données ...
- ... mais aussi de les étudier !
- ou encore d'étudier les liens entre des variables (corrélation)



## Vous aurez des travaux à effectuer entre chaque séance

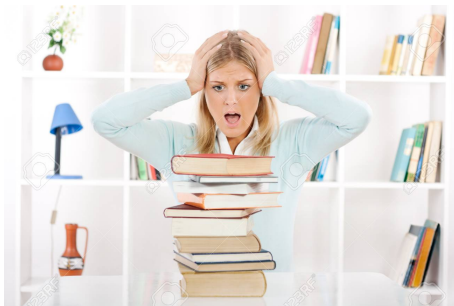
- Des lectures de documents (articles, cours portant sur une notion) et vous aurez à répondre à quelques questions à chaque fois.
- Des exercices pour vous faire retravailler vos notions en maths.
- Un ou deux travail (aux) en groupe pour faire une analyse de données.
- Ou encore des quizz sur Brightspace.

# Et en terme d'évaluation alors ?

## Rien de fixer pour le moment mais :

- Les travaux en cours seront évalués.
- Les exercices à faire à la maison aussi.
- Vous aurez un petit examen
- Un examen terminal qui comptera pour 50% de votre note finale

**Pour valider votre matière ... il suffit d'avoir 10 de moyenne.**



- Si vous avez des questions, remarques et suggestions, je reste joignable par mail à n'importe quel moment en dehors des cours (même à la fin de ce cours):

[guillaume.metzler@live.fr](mailto:guillaume.metzler@live.fr)

- N'hésitez pas à m'interrompre pendant le cours, on est peu nombreux donc c'est faisable.

**Des questions sur le déroulement du cours ?**

Passons aux choses sérieuses maintenant



Lorsque vous concevez une enquête, vous faites principalement les choses suivantes

- vous réfléchissez (éventuellement) aux personnes à qui est destinée cette enquête/sondage : on désignera ce groupe de personnes par le terme de **population**. Chaque personne sera alors désigné par le terme de **répondant** ou encore **individu** ou **sondé**. Vous trouverez aussi le terme **observation**
- vous rédigez des **questions**. A chaque questions est associée une **réponse** que l'on appelle aussi **variable**
- et enfin vous faites compléter votre questionnaire par un échantillon d'individus

# Que faire des réponses par la suite ?

Je dirai qu'il n'y a pas de recette **universelle**, que l'on pourrait appliquer à chaque enquête. Votre questionnaire est susceptible d'évoluer au cours des entretiens et le traitement des données brutes va dépendre de ce que **vous souhaitez montrer** mais aussi de la **qualité des réponses**.

Par contre cette étape est très importante pour votre analyse. Vous allez :

- Vérifier la qualité des réponses
- Détecter les réponses manquantes
- Eventuellement éliminer les données ou valeurs aberrantes

Quelques étapes après la conception de votre questionnaire et avant l'analyse des réponses.

- (Déterminer un plan préliminaire d'analyse des données)
- Vérification des questionnaires
- Edition
- Codage
- Transcription (fausse étape actuellement)
- Nettoyage des données
- Ajustements

**Tout est en place pour l'analyse**

Il s'agit de vérifier, au cours du sondage (moment où les personnes complètent le questionnaire) ou à la fin, la qualité des questionnaires:

- Complétude
- Etude des réponses et des enchaînements
- Etude des questionnaires
- Profil des répondants

**Des étapes importantes qui, si elles ne sont pas vérifiées, peuvent mettre à mal votre étude/analyse**

**Il s'agit de vérifier le bon déroulement de la procédure**

## Améliorer la précision et identifier les problèmes

**Problèmes** : liés aux questions ouvertes - mauvaises complétions - réponses incohérentes - abréviations ou termes ambiguës dans les questions ouvertes, ...

### Traitement des problèmes :

- Retour sur le terrain
- Imputer des valeurs spécifiques
- Elimination des données

# Deux étapes qui se font presque automatiquement actuellement

Règle importante pour le codage : **Pour une variable donnée, les codes doivent être mutuellement exclusifs et collectivement exhaustifs**, i.e. à chaque réponse correspond au moins un code et ce dernier est unique !

Il s'agit essentiellement de regarder que les personnes ont bien répondu à la question posée, coché une réponse unique quand cela est demandé.

Cette étape est **fastidieuse** et elle n'est pas effectuée sur la totalité des retours d'une enquête pour des raisons à la fois **humaines** et **économiques**. Seul un sous-échantillon est vérifié.

Elle permet cependant d'éliminer des réponses absurdes qui pourraient nuire à l'analyse des données.

# Traiter les réponses manquantes

Elles peuvent avoir deux origines :

- l'individu n'a pas répondu à la question
- il s'agit d'une réponse manquante dite *systeme*

Pour le traitement, plusieurs possibilités s'offrent à nous :

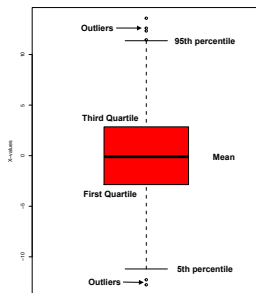
- **ignorer** simplement cette absence de réponse
- **supprimer** l'individu de l'enquête (à utiliser avec parcimonie, surtout que vous avez un faible nombre de répondants)
- remplacer la valeur manquante par une valeur neutre, en général la **moyenne**
- **extrapoler** la valeur manquante par rapport aux autres variables



# Ecarter les valeurs extrêmes voire absudes

C'est un cas que l'on rencontrera peu (voire pas) dans le cadre de ce cours.

Sachez cependant qu'il existe des méthodes statistiques pour détecter les valeurs extrêmes ou aberrantes, que l'on appellent **points leviers** dans les modèles statistiques linéaires. Une façon simple de détecter ces valeurs est également l'utilisation de **boxplot - boîte à moustache** ou **boîte de Tukey (1977)**.



Maintenant que nos données sont récoltées et "nettoyées" on peut alors les regarder d'encore plus près ! On remarque que l'on a des variables avec des **chiffres**, d'autres avec du **texte** ou encore des **notes**, des **noms de villes** ou encore un **commentaire** de quelques mots.

**Qu'est-ce qui différencie toutes ces variables ?**

En fait, on distingue deux types de variables :

- les variables dites **quantitatives**
- les variables dites **qualitatives**

La distinction entre ces deux variables est **primordiale**, elles ne s'étudient pas de la même façon (notion de valeur centrale, dispersion ou test de corrélation).

En outre, elles n'ont pas le même objectif (cf. lecture):

- **qualitative**  $\implies$  **une phase d'exploration**
- **quantitative**  $\implies$  une "analyse pure", qui servira à **montrer ou confirmer** des faits

## Variables Quantitatives

Il s'agit de variables qui contiennent des quantités ou valeurs mesurables. Les variables quantitatives sont donc numériques (attention une variable numérique n'est pas forcément quantitative !):

- Discrète : Âge, durée d'un séjour, ...
- Continue : Salaire, distance, température, ...

C'est avec ce type de variables que l'on va pouvoir effectuer un grand nombre d'analyses statistiques (calcul de moyenne, variance, modèles linéaires, etc ...).

Généralement peu de manipulations sont faites sur les variables quantitatives. Cependant, pour le besoin de certaines études il peut être intéressant de regrouper les différentes valeurs au sein d'une **classe** afin de **réduire** le nombre de modalités pour l'étude..

Cela se rencontre par exemple lorsque l'on cherche à étudier le salaire annuel d'une population:

$$\dots < 20k, \quad 20k \leq \dots < 30k, \quad 30k \leq \dots < 50k, \dots \geq 50k.$$

On peut aussi rencontrer cela lorsque l'on étudie l'**âge** d'une population.

## Variables Qualitatives

Il s'agit de variables qui expriment une qualité ou une catégorie comme le sexe, le nom, votre fonction. On distingue différentes catégories:

- Catégorique ou Nominale : nom de ville, métier ou tout ce qui porte un nom en général.
- Ordinale : il peut s'agir d'un score (numérique, 0-20) ou d'un jugement comprenant la notion de **rang** (pas d'accord, plutôt pas d'accord, d'accord, ...) ou d'**ordre** d'un point de vue matheux.

J'attire votre attention sur le fait qu'une variable **numérique** n'est pas obligatoirement une variable **quantitative** !

Si ces variables sont plus pauvres d'un point de vue "mathématiques", elles restent très riches en informations !

Eviter de faire la confusion entre **effectifs** et **fréquences**.

## Définitions

- **Effectif** : il s'agit du nombre d'individus  $x$  appartenant à une classe donnée ou à une catégorie, généralement  $x \in \mathbb{N}$ .
- **Fréquence** : Il s'agit d'une valeur  $x \in [0, 1]$  qui représente le **pourcentage** d'une population appartenant à une classe donnée.

## Vacances !

Enquête portant sur les destinations les plus visitées par les français:

Lieu de vacances	Effectifs
Seychelles	30 000
Paris	1 000 000
Londres	700 000
New York	800 000

- Quelle la variable étudiée et sa nature ?
- Quel est son nombre de modalités ?
- Quelle est la population étudiée ?

Passons maintenant à la pratique et regardons quelques exemples sur Sphinx



Regardons tout cela sur Sphinx

Je vous invite à vous connecter sur votre compte sphinx à l'aide du navigateur :

- Firefox
- ou Google Chrome

## Attention

**Ne pas utiliser Safari.** Certains menus de Sphinx ne fonctionnent pas avec Safari (Petite pensée pour les utilisateurs de Mac).

On va alors se concentrer sur cette barre de menu :



# Menu principal



- **Accueil** : permet de retourner au menu principal (choix d'une enquête / jeu de données)
- **Conception** : permet de créer votre propre enquête
- **Diffusion & Collecte** : envoyer votre questionnaire pour le faire compléter et récolter de nouvelles données
- **Données** : permet la visualisation de vos données et éventuellement de les modifier
- **Analyses** : pour des visualisations graphiques de vos données, calcul de grandeurs statistiques et tests statistiques

# Tableau de données

Sélectionnez l'enquête **Tourisme** et dirigez vous ensuite sur le menu : **Données**, en cliquant dessus vous verrez la page suivante:

Tableur

Consultation

Qualité de l'échantillon

Supprimer toutes les réponses

Variables

Choisir une strate





?

Tourisme - 552 réponses - Echantillon total

Déverrouiller

Supprimer

Exporter

	N°	1. Holiday L...	2. Activités	3. Accomod...	4. First vl...	5. Duration ...	6. Sources of information	7. Total spen...	8. Pour l'héb...	9. Pour l'all...	10. Au resta...	11. Pour les ...
<input type="checkbox"/>	 1	Ville-Bains	Baignade ; Caves / fromages ; Loisirs	Camping		8	Presse	2000	300	1000		500
<input type="checkbox"/>	 2	Le Villard	Baignade ; Vie locale ; Visites	Famille / amis	Non	8	Bouche à oreille	2500		1500		1000
<input type="checkbox"/>	 3	Le Villard	Famille ; Vie locale ; Promenade à pied	Camping	Oui	8	Presse	2500	500	1000	200	300
<input type="checkbox"/>	 4		Baignade ; Famille ; Bicyclette	Location / gîte	Oui	15	Presse	4000	1500	1500		500

Vous trouvez dans l'onglet **tableur** (en **violet** en haut à gauche) du menu **Données** qui vous permet de visualiser vos données.

## Quelques remarques

- nombre de lignes = nombre de répondants / individus / observations
- nombre de colonnes = nombre de variables / questions

Comment lire ce tableau ? La réponse est simple :




Une ligne représente donc les réponses fournies par **un seul individu**. Les réponses fournies par cette personnes sont décrites dans **chaque colonne** du tableau.

Mais on peut aussi s'amuser à faire des manipulations sur cette table !

# Quelques manipulations

## Modifier les données

Il vous est possible de **modifier** les données existantes en cliquant sur le bouton **Déverrouiller** (au dessus de votre tableau à gauche).

Tableur   Consultation   Qualité de l'échantillon				
 Déverrouiller				
		N°	1. Holiday I...	2. Activities
<input type="checkbox"/>		1	Ville-Bains	Baignade ; Caves / fromages ; Loisirs
<input type="checkbox"/>		2	Le Villard	Baignade ; Vie locale ; Visites

# Quelques manipulations

## Modifier les données

Voilà alors ce qui devrait s'afficher au même endroit.

Tableur



Consultation

Qualité de l'échantillon

Verrouiller

☒ Enregistrer les modifications

☐ Annuler les modifications

		N°	1. Holi...	2. Activities
<input type="checkbox"/>		1	Ville-Bains	Baignade ; Caves / fromages ; Loisirs
<input type="checkbox"/>		2	Le Villard	Baignade ; Vie locale ; Visites

Vous êtes maintenant libre de *supprimer* ou *modifier les données existantes*. Et il ne vous reste plus qu'à **enregistrer les modifications**.

# Quelques manipulations

## Autres possibilités

D'autres manipulations sont aussi possibles comme:

- **trier** les données
- **filtrer** les données

Pour cela, regardons les menus se trouvant sur la partie supérieure droit de votre tableau.

Supprimer toutes les réponses

Variables

Choisir une strate ▼

?



**Tourisme** - 552 réponses - Echantillon total



Supprimer



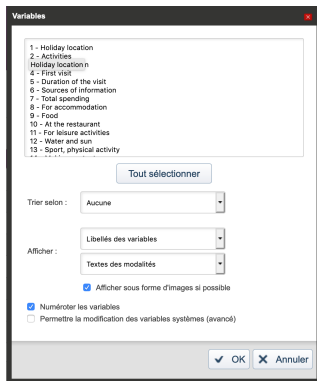
Exporter ▼



# Quelques manipulations

## Ordonner les données

Il vous est possible de **trier** les données existantes en fonction des valeurs ou modalités d'une variable) en cliquant sur le bouton **Variables** précédent



On s'intéresse aux deux premiers menus uniquement ans l'ordre (de haut en bas):

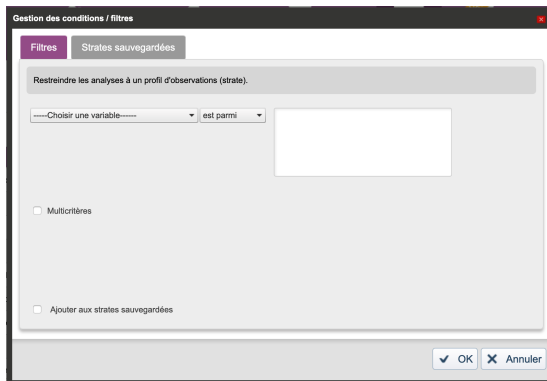
- Sélectionner les variables à afficher (il suffit de cliquer dessus)
- Sélectionner la variable selon laquelle le tri est effectué

Le tri peut se faire directement sur le tableau, en cliquant sur le nom de la variable.

# Quelques manipulations

## Filtrer les données

En cliquant sur le bouton **Choisir une strate** puis **Définir une strate**, le menu ci-dessous s'affichera.



The screenshot shows a software window titled "Gestion des conditions / filtres". It has two tabs: "Filtres" (active) and "Strates sauvegardées". The main area contains the instruction "Restreindre les analyses à un profil d'observations (strate).". Below this is a form with a dropdown menu labeled "Choisir une variable", a dropdown menu labeled "est parmi", and a large empty text box. At the bottom left, there are two checkboxes: "Multicritères" and "Ajouter aux strates sauvegardées". At the bottom right, there are two buttons: "OK" and "Annuler".

Il nous servira à filtrer les données selon une ou plusieurs caractéristiques (modalités).

# Quelques manipulations

## Filtrer les données

Il ne reste qu'à sélectionner les variables et les modalités selon lesquelles le tri est effectué. Il est même possible de filtrer selon plusieurs modalités de plusieurs variables à l'aide des opérateurs logiques **ET** et **OU**.

Gestion des conditions / filtres

Filtres Strates sauvegardées

Restreindre les analyses à un profil d'observations (strate).

12-L'eau et le soleil n'est pas pan

Non réponse  
Pas d'accord du tout  
Plutôt pas d'accord  
Cela dépend  
Plutôt d'accord  
Tout à fait d'accord

☒ Multicritères Ajouter Supprimer

Holiday location parmi "Margency,Ville-Bains,Le Villard"  
L'eau et le soleil n'est pas parmi "Plutôt pas d'accord;Cela dépend;Plutôt d'accord"

☒ Toutes les conditions (ET) ☐ Au moins une condition (OU)

☐ Ajouter aux strates sauvegardées

✓ OK ✗ Annuler

# Quelques manipulations

## Filtrer les données

Après validation, voilà la nouvelle table filtrée (on peut le voir car on a moins de lignes dans la tableau qu'avant 552 → 103)

Tourisme - 103 réponses - Filtre courant : Holiday location parmi "Margency;Ville-B...

Déverrouiller

Supprimer Exporter ▼

	N°	1. Holiday l...	2. Activités	3. Accomod...	4. First vi...	5. Duration ...	6. Sources of information	7. Total spen...	8. Pour l'héb...	9. Pour l'ail...	10. Au resta...	11	
<input checked="" type="checkbox"/>		1	Ville-Bains	Baignade ; Caves / fromages ; Loisirs	Camping		8	Presse	2000	300	1000		50
<input checked="" type="checkbox"/>		2	Le Villard	Baignade ; Vie locale ; Visites	Famille / amis	Non	8	Bouche à oreille	2500		1500		10
<input checked="" type="checkbox"/>		3	Le Villard	Famiente ; Vie locale ; Promenade à pied	Camping	Oui	8	Presse	2500	500	1000	200	30
<input checked="" type="checkbox"/>		5	Le Villard	Baignade ; Loisirs ; Planche à voile	Location / gîte	Non	8	Office de tourisme	3000	1000	1500		50
<input checked="" type="checkbox"/>		24	Ville-Bains	Baignade ; Loisirs	Location / gîte	Oui	21	Presse	4500	1500	2000		10
<input checked="" type="checkbox"/>		28	Ville-Bains	Baignade ; Famiente ; Loisirs	Camping	Oui	10	Bouche à oreille	2500	600	1000	300	
<input checked="" type="checkbox"/>		29	Ville-Bains	Promenade à pied ; Famiente ; Voiture	Hôtel	Non	24	Bouche à oreille					
<input checked="" type="checkbox"/>		30	Ville-Bains	Baignade ; Bicyclette ; Canot-kayak	Famille / amis	Non	15	Bouche à oreille ; Presse					
<input checked="" type="checkbox"/>		32	Ville-Bains	Randonnée ; Bicyclette ; Promenade à pied	Famille / amis	Oui	15	Bouche à oreille					

1

2

3

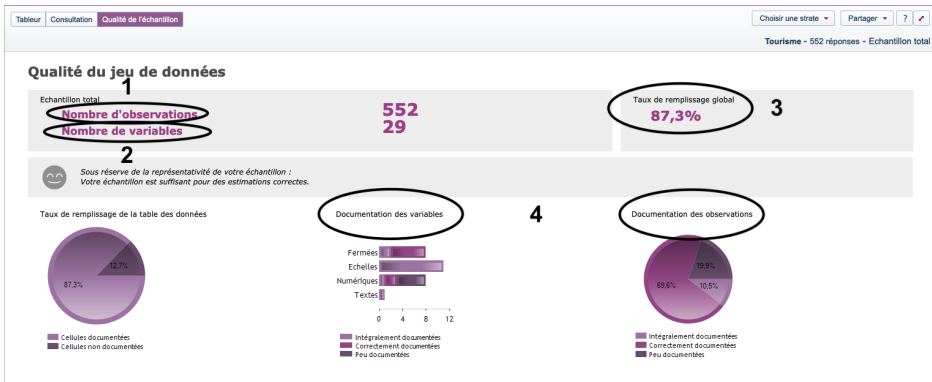
Page 1 de 3

50 ▼

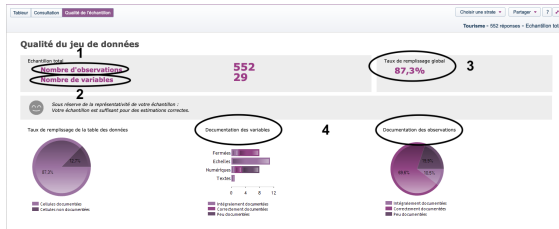
réponses par page

Afficher les réponses 1 - 50 sur 103

# Qualité de l'échantillon



# Qualité de l'échantillon



- 1) **Nombres d'observations** : correspond au nombre de personnes ayant complété l'enquête
- 2) **Nombre de variables** : le nombre de questions posées aux individus
- 3) **Taux de remplissage** : pourcentage de réponses à la totalité des questions (pour l'ensemble des observations)
- 4) **Documentation des ...** : vous renseigne sur la **qualité des réponses** pour l'ensemble des **variables** et **observations**

Regardons maintenant la deuxième partie de cette page :

Détails du taux de remplissage des variables

Variables	Fermées	Echelles	Numériques	Textes	Total
Intégralement documentées	2	11	1	1	15
Correctement documentées	6	0	2	0	8
Mal documentées	0	0	5	0	5

Détails du taux de remplissage des observations

Observations	Pour l'ensemble des variables	Pour les fermées	Pour les échelles	Pour les numériques	Pour les textes
Intégralement documentées	58	340	552	82	552
Correctement documentées	384	179	0	82	0
Mal documentées	110	33	0	388	0
Singulières	0	0	0	0	-
Avec trop peu de variance	-	-	0	-	-

Cette deuxième partie se décompose en deux tables :

- La première vous donne un *indice de remplissage* des différentes questions / variables et ce en fonction de la **nature de la question**.
- La deuxième table vous indique le nombre d'individus ayant répondu totalement ou partiellement aux différentes variables, toujours en fonction de la **nature de la variable**.

# Pour la semaine prochaine

## Travail à faire

- Relire votre cours (cela va sans dire mais c'est toujours mieux en le disant)
- Revoir les manipulations sur Sphinx
- Lire le document intitulé : *Préparer des Données*