Séance 2

Pierre-Marc Juneau

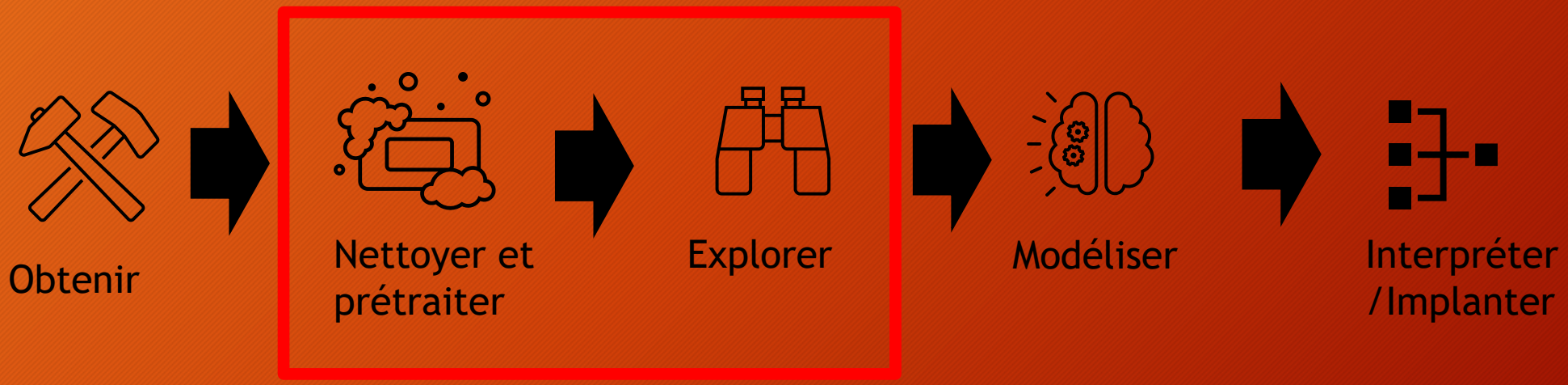
17 mars 2021

Plan

1. Principales étapes en science des données
2. Importance de la qualité des données
3. Critères de qualité des données
4. Nettoyage préliminaire des données

1- Principales étapes en science des données

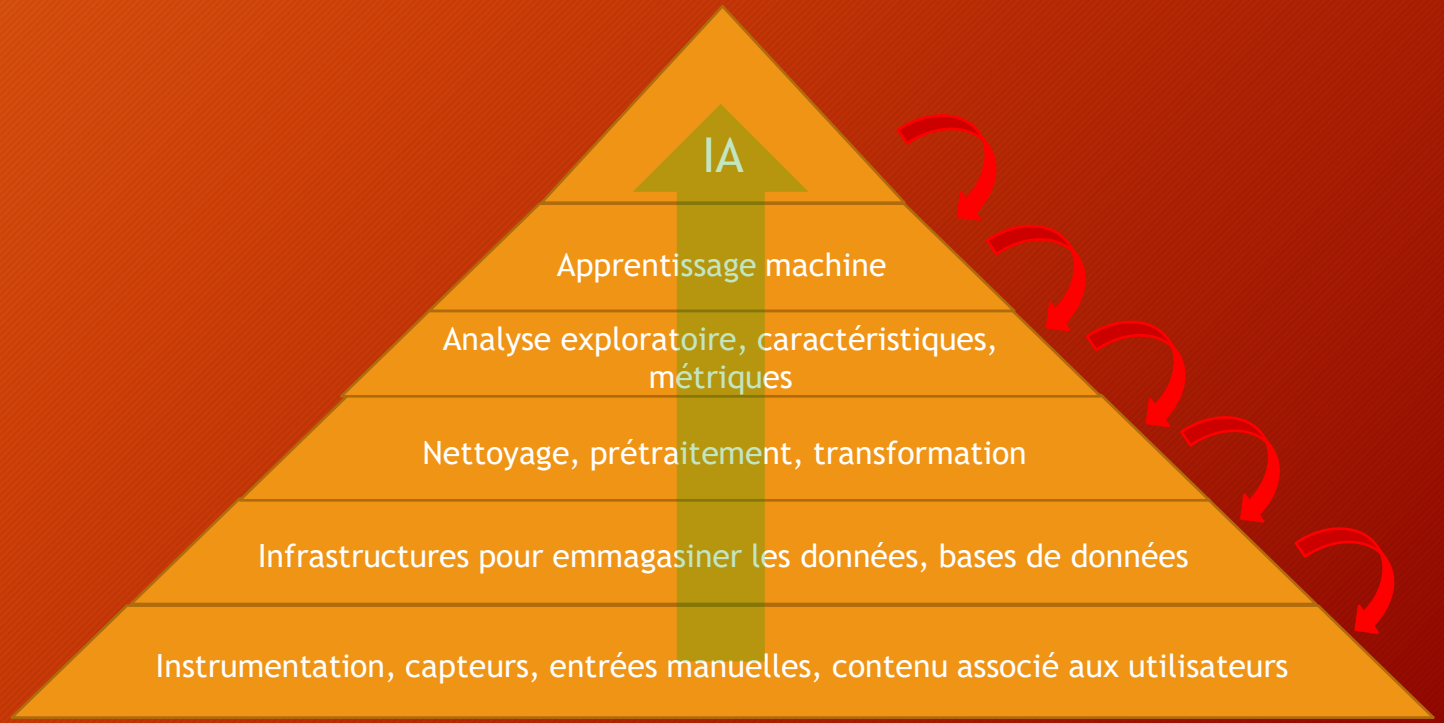
- Les étapes d'un projet en sciences des données (et où l'analyse exploratoire se situe):



https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492

1- Principales étapes de l'exploration de données

- Les données sont à la base d'un projet en science des données et en IA.
- Les critères de qualité des données permettent de détecter les problèmes et de considérer des améliorations en amont (à la base)
- Ex: lors du prétraitement des données l'analyste s'aperçoit qu'une variables associée à un capteur donne principalement des valeurs erronées



2- Importance de la qualité des données

- Les 5 caractéristiques (« V ») des données massives (« Big Data »)
  - Variété: les données sont de source variées (textes, photos, etc.), et de plus en plus non-structurées.
  - Véracité: s'assurer avec le grand volume de données générées qu'elles sont fiables, exactes, valides et cohérentes (de qualité).
  - Vélocité: les données doivent circuler et être exploitées avant de ne plus être d'actualité.
  - Volume: être en mesure de générer et d'exploiter les grands volumes de données générés (accessibilité).
  - Valeur: générer de la valeur pour la société (pouvoir être exploitée de manière pertinente).

2- Importance de la qualité des données

- Les impacts associés à une mauvaise qualité des données peuvent se faire sentir à plusieurs niveaux, et sont détectables plus ou moins facilement (d'évidents à subtiles)

3- Critères de qualité des données

- Les critères de qualité des données peuvent dépendre de beaucoup de facteurs: secteur d'activité, entreprise, utilisation prévue, etc.
- Quelques publications s'intéressent aux méthodes et métriques pouvant être développées pour assurer la qualité des données dans une organisation (ex: voir Heinrich et al. 2018)
- Dans le cadre de ce cours, quelques métriques générales sont proposées

3- Critères de qualité des données

- Accessibilité et disponibilité
  - Les données doivent être disponibles dans un format approprié
  - Il faut avoir la permission d'utiliser les données (ex: utilisation de données personnelles ou médicales: il peut y avoir un incitatif à anonymiser les données pour la protection des données personnelles)
  - Il est possible d'acheter des données dans différents domaines

3- Critères de qualité des données

- Intelligibilité
  - Dans le cas de données provenant de sondages et d'enquête, il faut vérifier que la documentation est complète, et que les questions posées sont claires.
  - Comme métrique, il serait possible de compter le nombre de requête des participants demandant des précisions.
  - Le nombre de valeurs qui ne sont pas dans le bon format pourrait également être une métrique intéressante.

SONDAGE

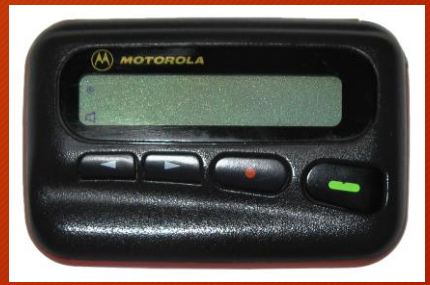
Question: Où habitez-vous?

Réponse: Ville? type d'habitation?

3- Critères de qualité des données

- Pertinence
  - Un des critères important pour la pertinence des données est son âge
  - En effet, la société et la technologie évolue, et des données qui étaient pertinentes il y a 10 ans ne sont peut-être plus pertinentes aujourd'hui.
  - Exemple: si jamais nous retrouvons des données d'utilisation des pagettes… difficile de les utiliser pour tirer des conclusions dans le contexte actuel

Selon les données que j'ai trouvées la pagette est très populaire chez les personnes dans la trentaine



3- Critères de qualité des données

- Pertinence
  - Une métrique a été proposée par Ballou et al. [1998] pour évaluer le degré d'actualité des données en fonction de l'âge des données

$$\text{Degré d'actualité} = \max\left[1 - \frac{\text{Âge des données}}{\text{Durée de vie typique}}, 0\right]$$

  - La durée de vie typique des données dépend des domaines d'application (ex: en marketing et dans les études sociales leur durée de vie est courte). Par exemple certaines sources suggèrent une limite de 3 ans sur les données provenant d'études marketing, économiques et liées au comportement des consommateurs.
  - Toutefois, pour un historien, ces données seront toujours valides pour la période visée.
  - Les données scientifiques, si elles décrivent des phénomènes physiques et chimiques tangibles, n'ont pas nécessairement de date d'expiration.

3- Critères de qualité des données

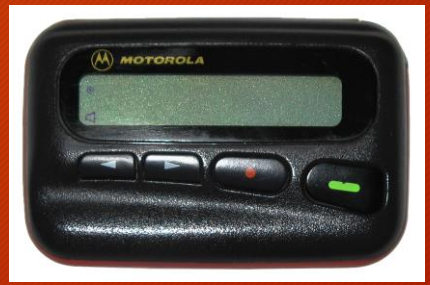
- Pertinence
  - Dans le cas des données de pagette
  - Degré d'actualité =

  max [1 – 25 ans / 3 ans, 0]

  = max [-7.33, 0] = 0

  - Donc ces données ne sont pas pertinentes, 0 d'actualité… sommes-nous surpris?

Selon les données que j'ai trouvées la pagette est très populaire chez les personnes dans la trentaine



3- Critères de qualité des données

- Complétude
  - Il peut y avoir des « trous » dans la matrice de données (ex: valeurs NaN)
    - Associés à des omissions lors de la saisie des données (ex: un consommateur qui a oublié d'entrer son adresse lorsqu'il s'est inscrit sur un site web)
    - Dans le domaine industrielle, peut être associé à une erreur de capteur (ex: une sonde de température qui ne peut acheminer son signal à un système de contrôle)

Date	Temps	Température du four (0C)
2020-04-01	13:00:00	55.4
2020-04-01	13:01:00	53.2
2020-04-01	13:02:00	55.2
2020-04-01	13:03:00	57.6
2020-04-01	13:04:00	65.3
2020-04-01	13:05:00	75.3
2020-04-01	13:06:00	87.4
2020-04-01	13:07:00	93.5
2020-04-01	13:08:00	97.4
2020-04-01	13:09:00	101.3
2020-04-01	13:10:00	104.3
2020-04-01	13:11:00	109.6
2020-04-01	13:12:00	NaN
2020-04-01	13:13:00	NaN
2020-04-01	13:14:00	125.3
2020-04-01	13:15:00	132.7
2020-04-01	13:16:00	143.6

3- Critères de qualité des données

- Complétude
  - Une métrique a été proposée par Blake et Mangiameli [2011] pour évaluer le degré de complétude des données, en fonction des données manquantes

$$\text{Degré Complétude} = \frac{N_R - N_{NAN}}{N_R}$$

N_R est le nombre d'occurrences (instances) total dans le jeu de données
N_{NAN} est le nombre d'occurrences (instances) avec au moins une valeur NaN

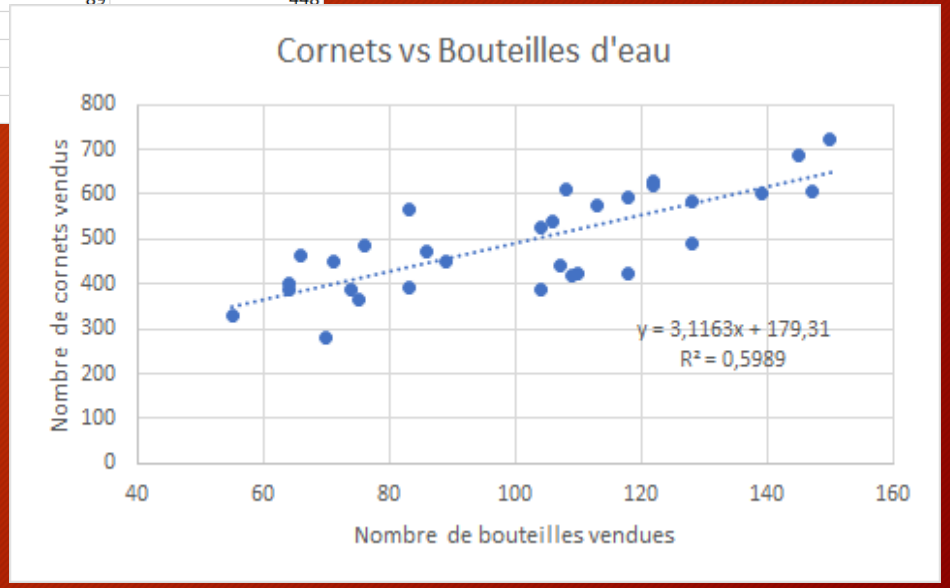
3- Critères de qualité des données

- Complétude
  - Un autre aspect de la complétude (plus difficile à évaluer) est de savoir si toutes les variables pertinentes sont incluses dans un jeu de données
  - Pour bien voir/comprendre les corrélations qui existent dans un jeu de données, il faut inclure les variables qui sont la source d'une tendance ou corrélation (cause à effet)
  - Exemple: ventes dans un kiosque de crème glacée

3- Critères de qualité des données

- Complétude
  - Seulement 2 variables disponibles: nombre de bouteilles d'eau vendues et nombres de cornets de crème glacée vendus
  - Conclusion: il y a une forte corrélation entre les ventes de bouteilles d'eau et celles de cornets de crème glacée
  - Pour vendre plus de cornets, alors il faut vendre plus de bouteille
  - Option: faire une promotion sur les bouteilles d'eau pour vendre plus de cornets???

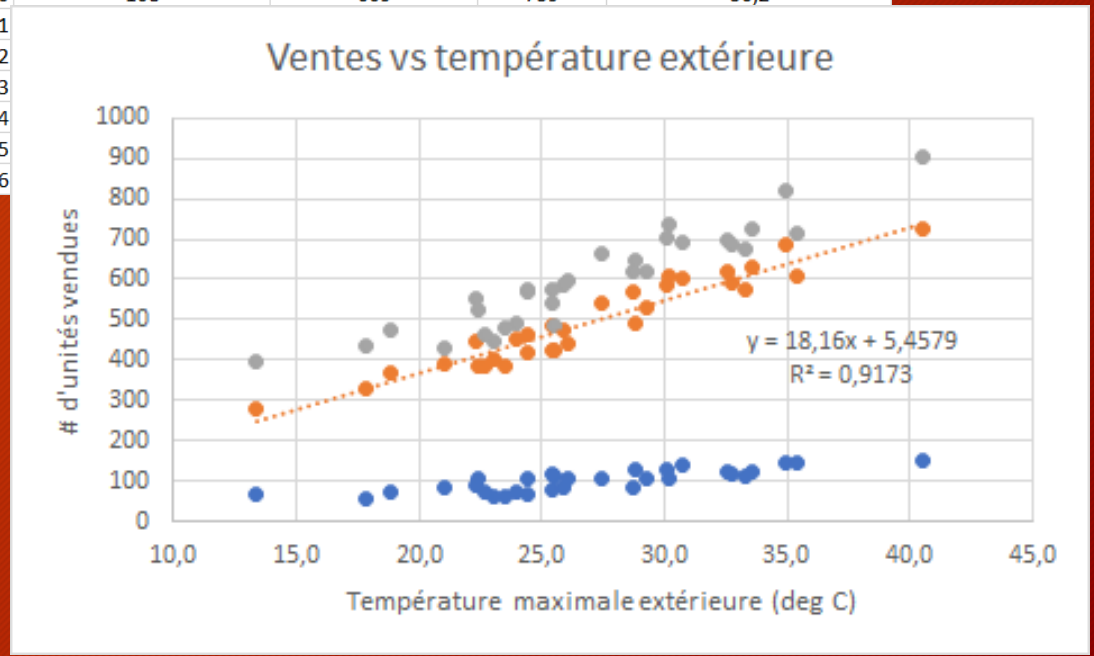
Date	# de bouteilles d'eau	# de cornets vendus
2018-08-01	66	464
2018-08-02	55	329
2018-08-03	76	485
2018-08-04	75	366
2018-08-05	122	630
2018-08-06	104	388
2018-08-07	139	603
2018-08-08	128	584
2018-08-09	145	687
2018-08-10	108	609
2018-08-11	70	281
2018-08-12	89	448
2018-08-13		
2018-08-14		
2018-08-15		
2018-08-16		



3- Critères de qualité des données

- Complétude
  - Cependant, en regardant les données de plus près et en y intégrant les données météo…
  - Conclusion: c'est plutôt la température maximale qui influence les ventes…
  - Pour capturer les relations de cause à effet, il faut tenter de capturer l'ensemble des variables importantes dans l'ensemble de données.
  - Cet exemple était évident, mais dans certains cas ce n'est pas toujours évident (ex: facteurs humains dans une usine: ce n'est pas toujours mesuré/quantifié)

Date	# de bouteilles d'eau	# de cornets vendus	# de ventes	Température max (deg C)
2018-08-01	66	464	572	24,4
2018-08-02	55	329	434	17,8
2018-08-03	76	485	573	25,5
2018-08-04	75	366	475	18,9
2018-08-05	122	630	726	33,6
2018-08-06	104	388	526	22,4
2018-08-07	139	603	693	30,7
2018-08-08	128	584	704	30,1
2018-08-09	145	687	819	35,0
2018-08-10	108	609	739	30,2
2018-08-11				
2018-08-12				
2018-08-13				
2018-08-14				
2018-08-15				
2018-08-16				



3- Critères de qualité des données

- Exactitude
  - Est-ce que les données sont représentatives des situations réelles?
  - Il faut s'assurer d'avoir un échantillon de données qui est représentatif (non-biaisé) de la population que l'on tente d'étudier.
  - Cet échantillon doit être assez important, tout dépendant de la marge d'erreur souhaitée.
  - Nous reviendrons sur ces concepts lorsque nous verrons les statistiques inférentielles.

	Confidence level = 95%			Confidence level = 99%		
	Margin of error			Margin of error		
Population size	5%	2,5%	1%	5%	2,5%	1%
100	80	94	99	87	96	99
500	217	377	475	285	421	485
1.000	278	606	906	399	727	943
10.000	370	1.332	4.899	622	2.098	6.239
100.000	383	1.513	8.762	659	2.585	14.227
500.000	384	1.532	9.423	663	2.640	16.055
1.000.000	384	1.534	9.512	663	2.647	16.317

https://www.checkmarket.com/blog/how-to-estimate-your-population-and-survey-sample-size/

3- Critères de qualité des données

- Exactitude
  - Une métrique possible pour évaluer l'exactitude de chaque point (proposée par Hinrichs [2002]) repose sur la distance entre la valeur dans la base de donnée et la valeur réelle:

$$\text{Degré Exactitude} = \frac{1}{\mathrm{d}(\omega,\ \omega_m) + 1}$$

ω est le point dans la base de donnée
ω_m est le point réel
$\mathrm{d}(\omega,\ \omega_m)$ est la distance entre le point dans la base de données et le point réel

  - Problématique: comment trouver/évaluer ω_m ? (ce n'est pas toujours évident)

3- Critères de qualité des données

- Exactitude
  - Peut se faire par exemple quand l'exactitude de certains points est vérifiée par validation
  - Exemple: un procédé de production de sacs de croustilles. Les données industrielles sont mesurées en lignes à l'aide de capteurs. Il est possible de prendre des échantillons pour vérifier en laboratoire quelques points, et se donner une idée de l'exactitude des points dans l'ensemble de la base de données

Temps	Poids estimé/calibré en production (g)	Poids mesuré en laboratoire (g)	Exactitude
13:00:00	200		
13:00:01	200		
13:00:02	201		
13:00:03	202		
13:00:04	200	200.45	0.68965517
13:00:05	201		
13:00:06	202		
13:00:07	205	205.4	0.71428571
13:00:08	199		

3- Critères de qualité des données

- Validité
  - Les données doivent être valides (ex: dans le bon format), en accord avec le monde réel qu'elles représentent.
  - Indices pouvant remettre en question la validité de certaines données
    - Des valeurs qui sortent des gammes usuelles ou acceptables (ex: une voiture qui aurait moins de 2 portes)
    - Des données qui se répètent dans une série temporelle (peut être lié à la saturation d'un capteur). Exemple: un capteur de température ne peut monter en haut de 200 ^{0}C (ces valeurs ne sont probablement pas valides et/ou exactes).

Date	Temps	Température du four (0C)
2020-04-01	17:00:00	183.4
2020-04-01	17:01:00	185.6
2020-04-01	17:02:00	192.9
2020-04-01	17:03:00	195.6
2020-04-01	17:04:00	198.3
2020-04-01	17:05:00	200.0
2020-04-01	17:06:00	200.0
2020-04-01	17:07:00	200.0
2020-04-01	17:08:00	200.0
2020-04-01	17:09:00	200.0
2020-04-01	17:10:00	200.0
2020-04-01	17:11:00	200.0
2020-04-01	17:12:00	200.0
2020-04-01	17:13:00	200.0
2020-04-01	17:14:00	200.0
2020-04-01	17:15:00	200.0
2020-04-01	17:16:00	200.0

3- Critères de qualité des données

- Validité
  - La validité des données associées à une variable particulière peut être affectée par le fait que les unités ne sont pas fournies (ni dans l'entête du fichier de données, ni dans le nom de la colonne, etc.).
  - Si les unités ne sont pas connues, utiliser les données (et assurer leur validité) peut-être difficile.

http://news.bbc.co.uk/2/hi/science/nature/462264.stm

3- Critères de qualité des données

- Validité
  - Une métrique possible (qu'il est possible de dériver de Yang et al. [2013]) est de déterminer k règles de validité sur les données, et de voir combien d'occurrences (instances) respectent chacune de ces règles (Qi), pour k règles

$$Q_i = \frac{N_R - N_{NV}}{N_R} \qquad \text{Degré Validité} = \sum_{i=1}^{k} Q_i$$

N_R est le nombre d'occurrences (instances) total dans le jeu de données
N_{NV} est le nombre d'occurrences (instances) qui enfreignent la règle i

3- Critères de qualité des données

- Cohérence
  - Est-ce que les données sont consistantes d'un jeu de données à l'autre, ou même à l'intérieur d'un même jeu de données (entre variables)?
  - En fait, il peut avoir une certaine forme de redondance dans l'information, avec des variables qui sont liées les unes aux autres par des relations mathématiques/contraintes
  - Utilisez ces liens pour valider les données
  - Exemple: données financières pour différentes entreprises (selon les exercices financiers)

3- Critères de qualité des données

- Cohérence
  - Dans le domaine financier: Actif = Passif + Capital
  - Certains points ne sont pas cohérents avec cette règle

Fait une opération matricielle pour faire la somme des colonnes Passif et Capital

Nom de l'entreprise	Nombre d'employés	Actif (100k$)	Passif (100k$)	Capital (100k$)
Entreprise A	17	52	27	6
Entreprise B	993	512	408	104
Entreprise C	412	209	205	43
Entreprise D	98	74	62	12
Entreprise E	2	15	13	2

Passif + Capital (100 k$)
33
512
248
74
15

Ces deux instances (occurrences) ne sont pas cohérentes… possiblement à enlever ou à réévaluer

3- Critères de qualité des données

- Cohérence
  - Une métrique possible (dérivée d'Alpar et Winkelsträter [2014]) serait de voir combien de tuples (instances ou occurrences) dans l'ensemble de données qui répondent à chaque règle de cohérence déterminée

$$\text{Degré Cohérence} = \frac{N_R - N_{NC}}{N_R}$$

N_R est le nombre d'occurrences (instances) total dans le jeu de données
N_{NC} est le nombre d'occurrences (instances) qui enfreignent au moins une règle de cohérence

3- Critères de qualité des données

- Cohérence
  - Comme pour le degré de validité, une métrique possible (qu'il est possible de dériver de Yang et al. [2013]) est de déterminer k règles de cohérence sur les données, et de voir combien d'occurrences (instances) respectent chacune de ces règles (Qi), pour k règles

$$Q_i = \frac{N_R - N_{NC}}{N_R} \qquad \text{Degré Cohérence} = \sum_{i=1}^{k} Q_i$$

N_R est le nombre d'occurrences (instances) total dans le jeu de données
N_{NC} est le nombre d'occurrences (instances) qui enfreignent la règle i

3- Critères de qualité des données

- Cohérence
  - Dans certains jeux de données, l'unicité des instances (ex: chaque client d'un magasin devrait avoir un seul profil dans la base de données) est importante, et il faut s'assurer qu'il n'y a pas de doublons.

3- Critères de qualité des données

- Exemple de calcul des métriques de qualité
  - Ensemble de données sur les véhicules électriques/hybrides

Il est possible de voir qu'il y a des données manquantes

	YEAR	Make	Model	Size	(kW)	Unnamed: 5	TYPE	CITY (kWh/10	HWY (kWh/1(	COMB (kWh/:	CITY (Le/100	HWY (Le/100	COMB (Le/10(	(g/km)	RATING	(km)	TIME (h)
2	2012	MITSUBISHI	i-MiEV	SUBCOMPACT	49	A1	B	16.9	21.4	18.7	1.9	2.4	2.1	0	n/a	100	7
3	2012	NISSAN	LEAF	MID-SIZE	80	A1	B	19.3	23	21.1	2.2	2.6	2.4	0	n/a	117	7
4	2013	FORD	FOCUS ELECT	COMPACT	107	A1		19	21.1		45	2.4	2.2	0	n/a	122	4
5	2013	MITSUBISHI	i-MiEV	SUBCOMPACT	49	A1	B	16.9	21.4	18.7	-3	2.4	2.1	0	n/a	100	7
6	2013	NISSAN	LEAF	MID-SIZE	80	A1	B	19.3	23	21.1	2.2		2.4	0	n/a	117	7
7	2013	SMART	FORTWO ELE(	TWO-SEATER	35	A1	B	17.2	22.5	19.6	1.9	2.5	2.2	0	n/a	109	8
8	2013	SMART	FORTWO ELE(	TWO-SEATER	35	A1	B	17.2	22.5	19.6	1.9	2.5	2.2	0	n/a	109	8
9	2013	TESLA	MODEL S (40	FULL-SIZE	270	A1	B		21.9	22.2	2.5	2.5	2.5	0	n/a	224	6
10	2013	TESLA	MODEL S (60	FULL-SIZE	270	A1	B	22.2	21.7	21.9	2.5	2.4	2.5	0	n/a	335	10
11	2013	TESLA	MODEL S (85	FULL-SIZE	270	A1	B	23.8	23.2	23.6	2.7	2.6	2.6	0	n/a	426	12
12	2033	TESLA	MODEL S PER	FULL-SIZE	310	A1	B	23.9	23.2	23.6	2.7	2.6	2.6	0	n/a	426	12
13	2014	CHEVROLET	SPARK EV	SUBCOMPACT	104	A1	B	16		17.8	1.8	2.2	2	0	n/a	131	7
14	2014	FORD	FOCUS ELECTRIC		107	A1	B	19	21.1	20	2.1	2.4	2.2	0	n/a	122	4
15	2014	MITSUBISHI	i-MiEV	SUBCOMPACT	49	A1	B	16.9	21.4	18.7	1.9	2.4	2.1	0	n/a	100	7
16	2014	NISSAN	LEAF	MID-SIZE	80	A1	B	16.5	20.8	18.4	1.9	2.3	2.1	0	n/a	135	5
17	2014	SMART	FORTWO ELE(	TWO-SEATER	35	A1	B	17.2	22.5	19.6	1.9	2.5	2.2	0	n/a	109	8
18	2014	SMART	FORTWO ELE(	TWO-SEATER	35	A1	B	17.2	22.5	19.6	1.9	2.5	2.2	0	n/a	109	8
19	2014	TESLA	MODEL S (60	FULL-SIZE	225	A1	B	22.2	21.7	21.9		2.4	2.5	0	n/a	335	10
20	2014	TESLA	MODEL S (85	FULL-SIZE	270	A1	B	23.8	23.2	23.6	2.7	2.6	2.6	0	n/a	426	12
21	2014	TESLA	MODEL S PER	FULL-SIZE	310	A1	B	23.9	23.2	23.6	2.7	2.6	2.6	0	n/a	426	12
22	2015	BMW	i3	SUBCOMPACT	125	A1	B	15.2	18.8	16.8	1.7	2.1	1.9	0	n/a	130	4
23	2015	CHEVROLET	SPARK EV	SUBCOMPACT	104	A1	B	16	19.6	17.8	1.8	2.2	2	0	n/a	131	7
24	2015	FORD	FOCUS ELECT	COMPACT	107	A1	B	19	21.1	20	2.1	2.4	2.2	0	n/a	122	4
25	2015	KIA		STATION WAC	81	A1	B	17.5	22.7	19.9	2	2.6	2.2	0	n/a	149	4
26	2015	MITSUBISHI	i-MiEV	SUBCOMPACT	49	A1	B	16.9	21.4	18.7	1.9	2.4	2.1	0	n/a	100	7
27	2015	NISSAN	LEAF	MID-SIZE	80	A1	B	16.5	20.8	18.4	1.9	2.3	2.1	0	n/a	135	5

DonnéesVoitures

3- Critères de qualité des données

- Exemple de calcul des métriques de qualité
  - Les données (dans un fichier CSV) sont chargées dans Spyder

```
import numpy as np
import pandas as pd
from numpy import nan

donnee = pd.read_csv('DonnéesVoitures.csv')
stats=donnee.describe()
dimensions=donnee.shape
```

3- Critères de qualité des données

- Exemple de calcul des métriques de qualité
  - Une première visualisation montre qu'il y a des données manquantes

e - DataFrame

YEAR	Make	Model	Size	(kW)	Innamed:	TYPE	(kWh/10C	(kWh/10(	3 (kWh/10	/ (Le/100	Y (Le/100	IB (Le/10C	(g/km)	RATING	(km)	TIME (h)
2012	MITSUBISHI	i-MiEV	SUBCOMPACT	49	A1	B	16.9	21.4	18.7	1.9	2.4	2.1	0	nan	100	7
2012	NISSAN	LEAF	MID-SIZE	80	A1	B	19.3	23	21.1	2.2	2.6	2.4	0	nan	117	7
2013	FORD	FOCUS ELECTRIC	COMPACT	107	A1	nan	9	21.1	nan	5	2.4	2.2	0	nan	122	4
2013	MITSUBISHI	i-MiEV	SUBCOMPACT	49	A1	B	16.9	21.4	18.7	-3	2.4	2.1	0	nan	100	7
2013	NISSAN	LEAF	MID-SIZE	80	A1	B	19.3	23	21.1	2.2	nan	.4	0	nan	117	7
2013	SMART	FORTWO ELECTRIC DRIVE CABRIOLET	TWO-SEATER	35	A1	B	17.2	22.5	19.6	1.9	2.5	2.2	0	nan	109	8

3- Critères de qualité des données

- Exemple de calcul des métriques de qualité
  - Validité des données: une première étape est de regarder les stats à haut niveau

Index	YEAR	(kW)	(kWh/10(	(kWh/10(	3 (kWh/10	/ (Le/100	Y (Le/100	1B (Le/100	(g/km)	RATING	(km)	TIME (h)
count	53	53	52	52	52	52	52	53	53	19	53	53
mean	2015.11	190.623	19.5942	21.6731	20.5519	2.93269	2.41923	2.30189	0	10	239.17	8.4717
std	2.77808	155.526	3.00519	1.2246	1.9973	6.00156	0.142854	0.212576	0	0	141.426	2.99104
min	2012	35	15.2	18.8	16.8	-3	2.1	1.9	0	10	100	4
25%	2014	80	16.975	20.95	18.7	1.9	2.3	2.1	0	10	117	7
50%	2015	107	19	21.7	20	2.1	2.4	2.2	0	10	135	8
75%	2016	283	22.375	22.55	22.125	2.6	2.5	2.5	0	10	402	12
max	2033	568	23.9	23.3	23.6	45	2.6	2.6	0	10	473	12

Règle de validité 1: l'année ne peut dépasser 2021

Règle de validité 2: la consommation en L au 100 km en ville ne peut être négatif et ne peut excéder 4 L au 100 km

3- Critères de qualité des données

- Exemple de calcul des métriques de qualité
  - Il est alors possible de calculer le degré de validité

```
NR=dimensions[0]
NNV=sum(i > 2021 for i in donnee["YEAR"])
Q1=(NR-NNV)/NR

NNV2=sum(i > 4 for i in donnee["CITY (Le/100 km)"]) + sum(i < 0 for i in donnee["CITY (Le/100 km)"])
Q2=(NR-NNV2)/NR

DegValidite=(Q1+Q2)/2
```

Q1 = 0.9811320754716981
Q2 = 0.9622641509433962
DegValidite = 0.9716981132075472

3- Critères de qualité des données

- Exemple de calcul des métriques de qualité
  - Il est alors possible de calculer le degré de complétude pour chaque variable

    Nnan=donnee.isnull().sum()

    DegCompletude=(NR-Nnan)/NR

DegCompletude - Series

Index	0
YEAR	1
Make	1
Model	0.981132
Size	0.981132
(kW)	1
Unnamed: 5	1
TYPE	0.981132
CITY (kWh/100 km)	0.981132
HWY (kWh/100 km)	0.981132
COMB (kWh/100 km)	0.981132
CITY (Le/100 km)	0.981132
HWY (Le/100 km)	0.981132
COMB (Le/100 km)	1
(g/km)	1
RATING	0.358491
(km)	1
TIME (h)	1

3- Critères de qualité des données

- Exercice L02 - #1
  - Un ensemble de données sur la productivité de vaches laitières en fonction de différents facteurs (ex: l'alimentation) vous a été fourni.
  - Évaluez la qualité de ce jeu de données: calculez le degré de complétude, de cohérence et/ou de validité.

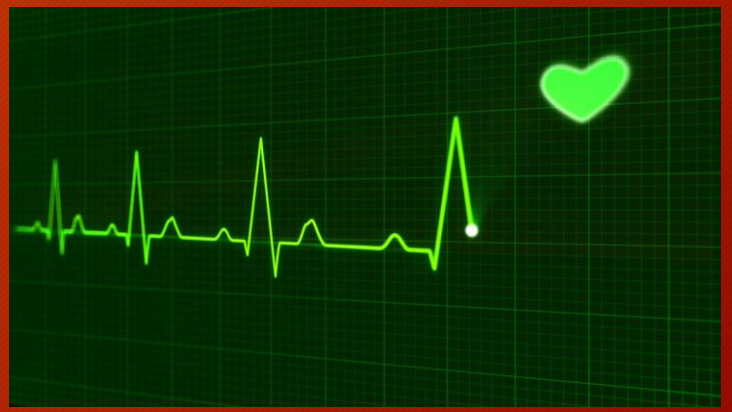
4- Nettoyage préliminaire des données

- Pour adresser les problèmes de qualité dans les données, plusieurs approches de nettoyage existent
  - Enlèvement des tuples (instances) problématiques
  - Enlever les variables avec un faible degré de complétude
  - Remplacement des valeurs NaN par un estimé (imputation)

4- Nettoyage préliminaire des données

- Exemple: Petit tableau de données médicales

Patient ID	Date	Rythme cardiaque (battements/min)	Taux oxygénation (%)	Glycémie (g/L)
546	21-10-2019	84	100	0.9
785	21-10-2019			0.9
385	21-10-2019	71		1.3
902	21-10-2019	82		1.3
198		72		1.1
861	21-10-2019	96		0.6
558	21-10-2019	74		1.2
513	21-10-2019	80	100	1.0
507	21-10-2019	89	99	0.4
125	21-10-2019		100	0.7
351	22-10-2019	77	80	100.0
429	22-10-2019	75	95	0.9



4- Nettoyage préliminaire des données

- Exemple: Petit tableau de données médicales
  - Patient ID est unique (clé primaire), il sera utilisé comme index
    - *import numpy as np*
    - *import pandas as pd*
    - *from numpy import nan*
    - *import sklearn.impute as imp*
    - *donnee = pd.read_csv('DonnéesMédicales.csv', index_col='Patient ID')*
    - *stats=donnee.describe()*

donnee - DataFrame

Patient ID	Date	Rythme cardiaque (battements/min)	Taux oxygénation (%)	Glycémie (g/L)
546	21-10-2019	84	100	0.9
785	21-10-2019	nan	nan	0.9
385	21-10-2019	71	nan	1.3
902	21-10-2019	82	nan	1.3
198	nan	72	nan	1.1
861	21-10-2019	96	nan	0.6
558	21-10-2019	74	nan	1.2
513	21-10-2019	80	100	1
507	21-10-2019	89	99	0.4
125	21-10-2019	nan	100	0.7
351	22-10-2019	77	80	100
429	22-10-2019	75	95	0.9

4- Nettoyage préliminaire des données

- Exemple: Petit tableau de données médicales
  - Possible d'enlever tout simplement les tuples qui comportent une valeur NaN
    - *X1 = donnee.dropna()*

X1 - DataFrame

Patient ID	Date	Rythme cardiaque (battements/min)	Taux oxygénation (%)	Glycémie (g/L)
546	21-10-2019	84	100	0.9
513	21-10-2019	80	100	1
507	21-10-2019	89	99	0.4
351	22-10-2019	77	80	100
429	22-10-2019	75	95	0.9

Dans ce cas, c'est un peu drastique

4- Nettoyage préliminaire des données

- Exemple: Petit tableau de données médicales
  - Autre approche: remplacer les Nan par des valeurs estimées… cependant laquelle? Dépend des contextes
  - Par des zéros
    - *X2 = donnee.replace(nan, 0)*

X2 - DataFrame

Patient ID	Date	Rythme cardiaque (battements/min)	Taux oxygénation (%)	Glycémie (g/L)
546	21-10-2019	84	100	0.9
785	21-10-2019	0	0	0.9
385	21-10-2019	71	0	1.3
902	21-10-2019	82	0	1.3
198	0	72	0	1.1
861	21-10-2019	96	0	0.6
558	21-10-2019	74	0	1.2
513	21-10-2019	80	100	1
507	21-10-2019	89	99	0.4
125	21-10-2019	0	100	0.7
351	22-10-2019	77	80	100
429	22-10-2019	75	95	0.9

4- Nettoyage préliminaire des données

- Exemple: Petit tableau de données médicales
  - Cependant, remplacer les NaN par des zéros fausse beaucoup les statistiques (ex: moyenne, écart-type)

Avant

stats - DataFrame

Index	Rythme cardiaque (battements/min)	Taux oxygénation (%)	Glycémie (g/L)
count	10	6	12
mean	80	95.6667	9.19167
std	7.97217	7.91623	28.5985
min	71	80	0.4
25%	74.25	96	0.85
50%	78.5	99.5	0.95
75%	83.5	100	1.225
max	96	100	100

Après

stats2 - DataFrame

Index	Rythme cardiaque (battements/min)	Taux oxygénation (%)	Glycémie (g/L)
count	12	12	12
mean	66.6667	47.8333	9.19167
std	31.964	50.2446	28.5985
min	0	0	0.4
25%	71.75	0	0.85
50%	76	40	0.95
75%	82.5	99.25	1.225
max	96	100	100

4- Nettoyage préliminaire des données

- Exemple: Petit tableau de données médicales
  - Autre approche: imputer une valeur plus représentative de chaque variable, telle que la moyenne et le mode (pour les variables autres que numériques)
    - *X3 = donnee.fillna(donnee.mean())*
    - *X3 = X3.fillna(donnee['Date'].mode()[0])*

X3 - DataFrame

Patient ID	Date	Rythme cardiaque (battements/min)	Taux oxygénation (%)	Glycémie (g/L)
546	21-10-2019	84	100	0.9
785	21-10-2019	80	95.6667	0.9
385	21-10-2019	71	95.6667	1.3
902	21-10-2019	82	95.6667	1.3
198	21-10-2019	72	95.6667	1.1
861	21-10-2019	96	95.6667	0.6
558	21-10-2019	74	95.6667	1.2
513	21-10-2019	80	100	1
507	21-10-2019	89	99	0.4
125	21-10-2019	80	100	0.7
351	22-10-2019	77	80	100
429	22-10-2019	75	95	0.9

4- Nettoyage préliminaire des données

- Exemple: Petit tableau de données médicales
  - Ceci permet de conserver les mêmes moyennes pour les variables (cependant l'écart-type diminue)

Avant

stats - DataFrame

Index	Rythme cardiaque (battements/min)	Taux oxygénation (%)	Glycémie (g/L)
count	10	6	12
mean	80	95.6667	9.19167
std	7.97217	7.91623	28.5985
min	71	80	0.4
25%	74.25	96	0.85
50%	78.5	99.5	0.95
75%	83.5	100	1.225
max	96	100	100

Après

stats3 - DataFrame

Index	Rythme cardiaque (battements/min)	Taux oxygénation (%)	Glycémie (g/L)
count	12	12	12
mean	80	95.6667	9.19167
std	7.2111	5.33712	28.5985
min	71	80	0.4
25%	74.75	95.6667	0.85
50%	80	95.6667	0.95
75%	82.5	99.25	1.225
max	96	100	100

4- Nettoyage préliminaire des données

- Exemple: Petit tableau de données médicales
  - Une autre méthode pour remplacer les valeurs NaN est le *SimpleImputer* dans la librairie *sklearn* (cependant sur seulement des données numériques)
    - *donnee_num=donnee.values*
    - *donnee_X4=donnee_num[:,1:4]*
    - *imputer = imp.SimpleImputer(missing_values=nan, strategy='mean')*
    - *X4_num=pd.DataFrame(imputer.fit_transform(donnee_X4))*
    - *X4_num.index=donnee.index*
    - *X4_num.columns=donnee.columns[1:4]*
    - *X4 = pd.concat([donnee['Date'],X4_num], axis=1)*

X4 - DataFrame

Patient ID	Date	Rythme cardiaque (battements/min)	Taux oxygénation (%)	Glycémie (g/L)
546	21-10-2019	84	100	0.9
785	21-10-2019	80	95.6667	0.9
385	21-10-2019	71	95.6667	1.3
902	21-10-2019	82	95.6667	1.3
198	nan	72	95.6667	1.1
861	21-10-2019	96	95.6667	0.6
558	21-10-2019	74	95.6667	1.2
513	21-10-2019	80	100	1
507	21-10-2019	89	99	0.4
125	21-10-2019	80	100	0.7
351	22-10-2019	77	80	100
429	22-10-2019	75	95	0.9

4- Nettoyage préliminaire des données

- Exemple: Petit tableau de données médicales
  - Il y a plusieurs méthodes d'imputation disponibles, pouvant se servir des autres données pour avoir un meilleur estimé
    - *donnee_num=donnee.values*
    - *donnee_X5=donnee_num[:,1:4]*
    - *imputer = imp.KNNImputer(n_neighbors=2)*
    - *X5_num=pd.DataFrame(imputer.fit_transform(donnee_X4))*
    - *X5_num.index=donnee.index*
    - *X5_num.columns=donnee.columns[1:4]*
    - *X5 = pd.concat([donnee['Date'],X5_num], axis=1)*

X5 - DataFrame

Patient ID	Date	Rythme cardiaque (battements/min)	Taux oxygénation (%)	Glycémie (g/L)
546	21-10-2019	84	100	0.9
785	21-10-2019	79.5	97.5	0.9
385	21-10-2019	71	97.5	1.3
902	21-10-2019	82	100	1.3
198	nan	72	97.5	1.1
861	21-10-2019	96	99.5	0.6
558	21-10-2019	74	97.5	1.2
513	21-10-2019	80	100	1
507	21-10-2019	89	99	0.4
125	21-10-2019	90	100	0.7
351	22-10-2019	77	80	100
429	22-10-2019	75	95	0.9

4- Nettoyage préliminaire des données

- Exemple: Petit tableau de données médicales
  - Il est également possible de retirer les tuples (ou instances) qui ont des valeurs aberrantes ou tout simplement erronées en sélection les données selon des critères

donnee - DataFrame

Patient ID	Date	Rythme cardiaque (battements/min)	Taux oxygénation (%)	Glycémie (g/L)
546	21-10-2019	84	100	0.9
785	21-10-2019	nan	nan	0.9
385	21-10-2019	71	nan	1.3
902	21-10-2019	82	nan	1.3
198	nan	72	nan	1.1
861	21-10-2019	96	nan	0.6
558	21-10-2019	74	nan	1.2
513	21-10-2019	80	100	1
507	21-10-2019	89	99	0.4
125	21-10-2019	nan	100	0.7
351	22-10-2019	77	80	100
429	22-10-2019	75	95	0.9

4- Nettoyage préliminaire des données

- Exemple: Petit tableau de données médicales
  - Il est également possible de retirer les tuples (ou instances) qui ont des valeurs aberrantes ou tout simplement erronées en sélection les données selon des critères
    - *X6 = donnee[(donnee["Glycémie (g/L)"] < 3) & (donnee["Glycémie (g/L)"] > 0)]*

X6 - DataFrame

Patient ID	Date	Rythme cardiaque (battements/min)	Taux oxygénation (%)	Glycémie (g/L)
546	21-10-2019	84	100	0.9
785	21-10-2019	nan	nan	0.9
385	21-10-2019	71	nan	1.3
902	21-10-2019	82	nan	1.3
198	nan	72	nan	1.1
861	21-10-2019	96	nan	0.6
558	21-10-2019	74	nan	1.2
513	21-10-2019	80	100	1
507	21-10-2019	89	99	0.4
125	21-10-2019	nan	100	0.7
429	22-10-2019	75	95	0.9

4- Nettoyage préliminaire des données

- Exemple: Petit tableau de données médicales
  - De plus, la variable « Taux oxygénation(%) » a un degré de complétude très bas (): il serait possible de retirer cette variable
    - *dimensions=donnee.shape*
    - *NR=dimensions[0]*
    - *Nnan=donnee.isnull().sum()*
    - *DegCompletude=(NR-Nnan)/NR*

DegCompletude - Series

Index	0
Date	0.916667
Rythme cardiaque (battements/min)	0.833333
Taux oxygénation (%)	0.5
Glycémie (g/L)	1

4- Nettoyage préliminaire des données

- Exemple: Petit tableau de données médicales
  - De plus, la variable « Taux oxygénation(%) » a un degré de complétude très bas (): il serait possible de retirer cette variable
    - *X7 = donnee.drop('Taux oxygénation (%)', 1)*

X7 - DataFrame

Patient ID	Date	Rythme cardiaque (battements/min)	Glycémie (g/L)
546	21-10-2019	84	0.9
785	21-10-2019	nan	0.9
385	21-10-2019	71	1.3
902	21-10-2019	82	1.3
198	nan	72	1.1
861	21-10-2019	96	0.6
558	21-10-2019	74	1.2
513	21-10-2019	80	1
507	21-10-2019	89	0.4
125	21-10-2019	nan	0.7
351	22-10-2019	77	100
429	22-10-2019	75	0.9

4- Nettoyage préliminaire des données

- Exercice L02 - #2
  - Pour l'ensemble de données vu précédemment, créez une première matrice (X1) qui en remplaçant les valeurs manquantes par la moyenne.
  - Ensuite, créez une deuxième matrice (X2) dans laquelle vous enlevez plutôt les tuples qui sont incohérents. Retirez de l'ensemble de données les variables qui ont un degré de complétude en bas de 90%. S'il reste des valeurs manquantes, remplacez-les par la moyenne.

Références

- Articles
  - Bernd Heinrich, Diana Hristova, Mathias Klier, Alexander Schiller, and Michael Szubartowicz. 2018. Requirements for Data Quality Metrics. J. Data and Information Quality 9, 2, Article 12 (January 2018), 32 pages. DOI:https://doi.org/10.1145/3148238
  - Cai, L and Zhu, Y 2015 The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14: 2, pp. 1-10, DOI: http://dx.doi.org/10.5334/dsj-2015-002
- Sites web
  - https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492
  - https://www150.statcan.gc.ca/n1/pub/12-539-x/2009001/quality-qualite-fra.htm
  - https://segmeasurement.com/content/when-study-considered-be-outdated
  - https://scikit-learn.org/stable/modules/impute.html#univariate-vs-multivariate-imputation