Cours 7:
Traitement des données

Prof: Gauthier Gidel

23 Septembre 2024

(some of the material inspired from slides from Zico Kolter, Golnoosh Faradi, Kris Sankaran and Jhelum Chakravorty)

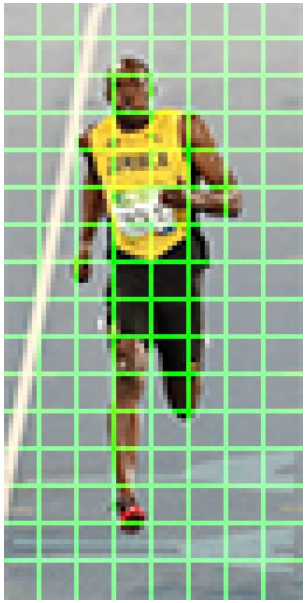
La préparation des données accapare une bonne fraction du temps d'un scientifique des données.

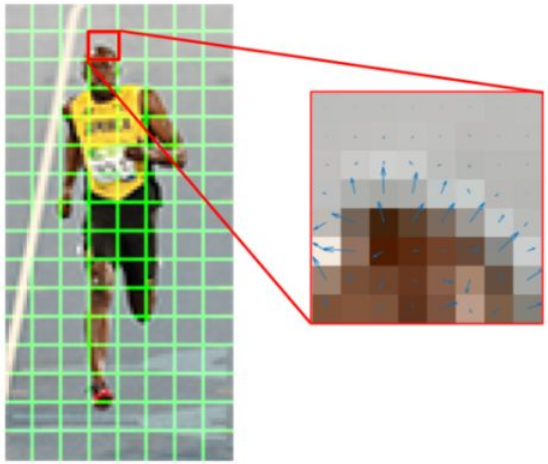
	tx_price	beds	baths	sqft	year_built	lot_size	property_type	exterior_walls	roof	basement	restaurants	groceries	nightlife
0	295850	1	1	584	2013	0	Apartment / Condo / Townhouse	Wood Siding	NaN	NaN	107	9	30
1	216500	1	1	612	1965	0	Apartment / Condo / Townhouse	Brick	Composition Shingle	1.0	105	15	6
2	279900	1	1	615	1963	0	Apartment / Condo / Townhouse	Wood Siding	NaN	NaN	183	13	31
3	379900	1	1	618	2000	33541	Apartment / Condo / Townhouse	Wood Siding	NaN	NaN	198	9	38
4	340000	1	1	634	1992	0	Apartment / Condo / Townhouse	Brick	NaN	NaN	149	7	22
5	265000	1	1	641	1947	0	Apartment / Condo / Townhouse	Brick	NaN	NaN	146	10	23
6	240000	1	1	642	1944	0	Single-Family	Brick	NaN	NaN	159	13	36
7	388100	1	1	650	2000	33541	Apartment / Condo / Townhouse	Wood Siding	NaN	NaN	198	9	38
8	240000	1	1	660	1983	0	Apartment / Condo / Townhouse	Brick	NaN	NaN	51	8	6
9	250000	1	1	664	1965	0	Apartment / Condo / Townhouse	Brick	NaN	NaN	119	10	26

Qu'est-ce qu'une caractéristique ?

Qu'est-ce qu'une caractéristique ?

Source de l'image : https://learnopencv.com/histogram-of-oriented-gradients/





Quelques calculs

2	3	4	4	3	4	2	2
5	11	17	13	7	9	3	4
11	21	23	27	22	17	4	6
23	99	165	135	85	32	26	2
91	155	133	136	144	152	57	28
98	196	76	38	26	60	170	51
165	60	60	27	77	85	43	136
71	13	34	23	108	27	48	110

Gradient Magnitude

80	36	5	10	0	64	90	73
37	9	9	179	78	27	169	166
87	136	173	39	102	163	152	176
76	13	1	168	159	22	125	143
120	70	14	150	145	144	145	143
58	86	119	98	100	101	133	113
30	65	157	75	78	165	145	124
11	170	91	4	110	17	133	110

Gradient Direction

-Algorithme ML,
-Visualisation
-...

Entrée : x

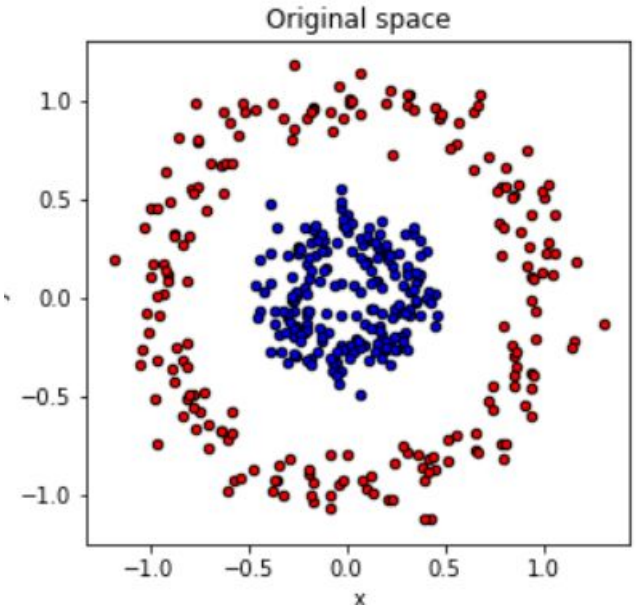
Caractéristiques extraites : $\varphi(x)$

Qu'est-ce que l'ingénierie des caractéristique ?

- Étape de traitement de l'entrée (et parfois aussi de la sortie)
- Incorporer la connaissance du domaine et des tâches
- Avant 2012 :
  - Caractéristique calculées à la main (par exemple, histogrammes de gradient orienté)
  - Trouver de nouvelles caractéristiques était un sujet de recherche pour les experts.
- Révolution de l'apprentissage en profondeur
  - Caractéristique apprises
  - Cependant, il y a encore du traitement de l'entrée ! (normalisation, encodage de sortie spécifique, détection des valeurs aberrantes)

Motivations pour l'ingénierie des caractéristiques

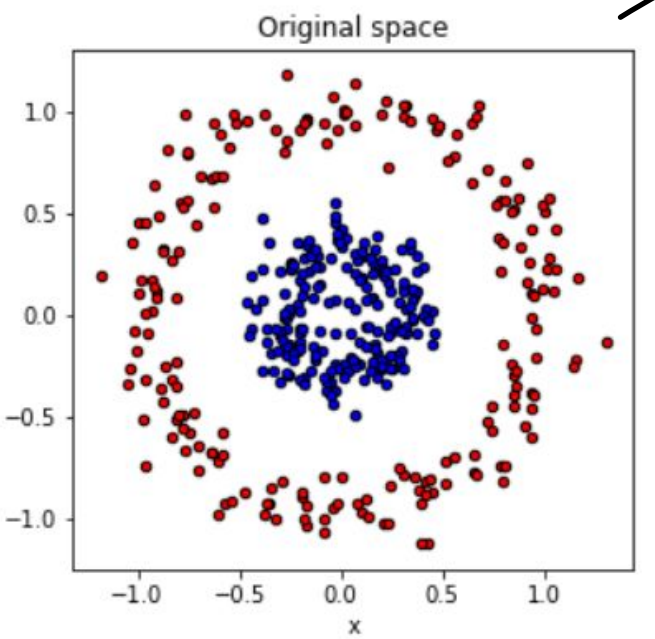
Un exemple simple

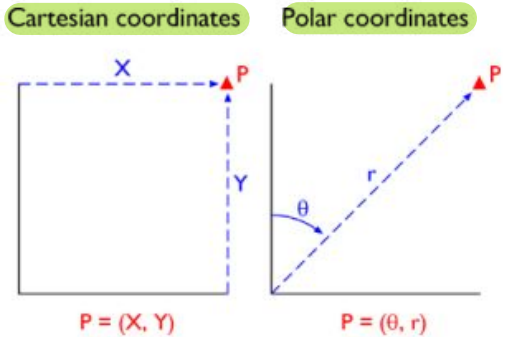


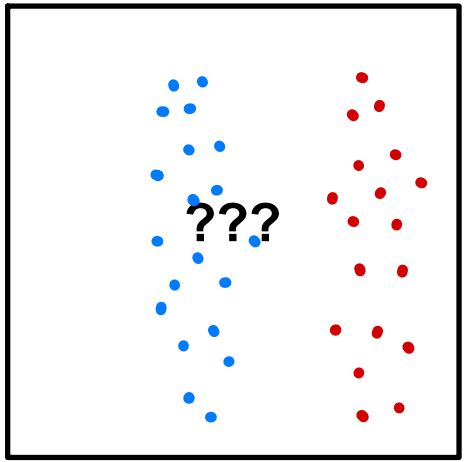
- Les données sont séparables
- Impossible de séparer à l'aide d'un classificateur linéaire

Source : https://www.slideshare.net/0xdata/feature-engineering-83511751

Un exemple simple







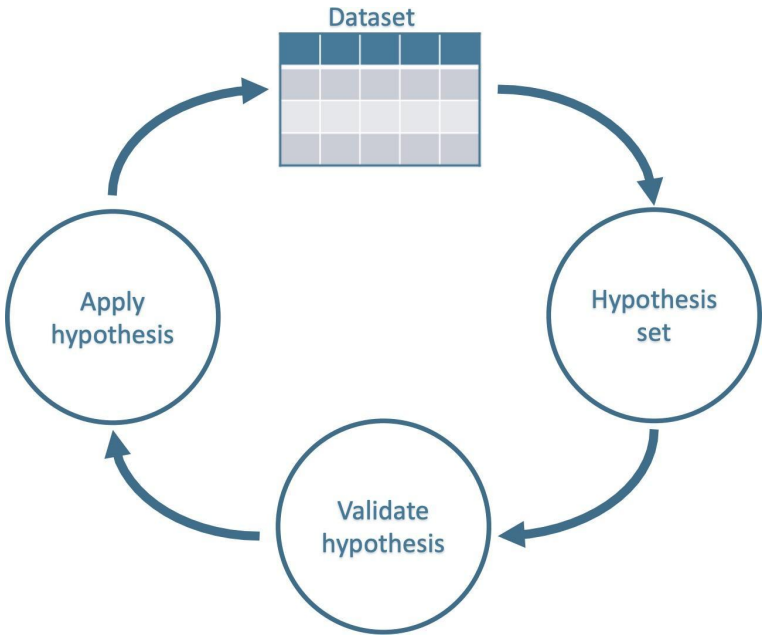
Source : https://www.slideshare.net/0xdata/feature-engineering-83511751

Quelques citations (traduises)

"Créer des caractéristiques est difficile, prend du temps et nécessite des connaissances d'expert. "L'apprentissage automatique appliqué" est essentiellement de l'ingénierie des fonctionnalités." André Ng

"Une bonne préparation des données et une ingénierie des caractéristiques font partie intégrante d'une meilleure prédiction" Marios Michailidis (KazAnova), Kaggle GrandMaster, Kaggle #3, ancien #1

Cycle d'ingénierie des caractéristiques

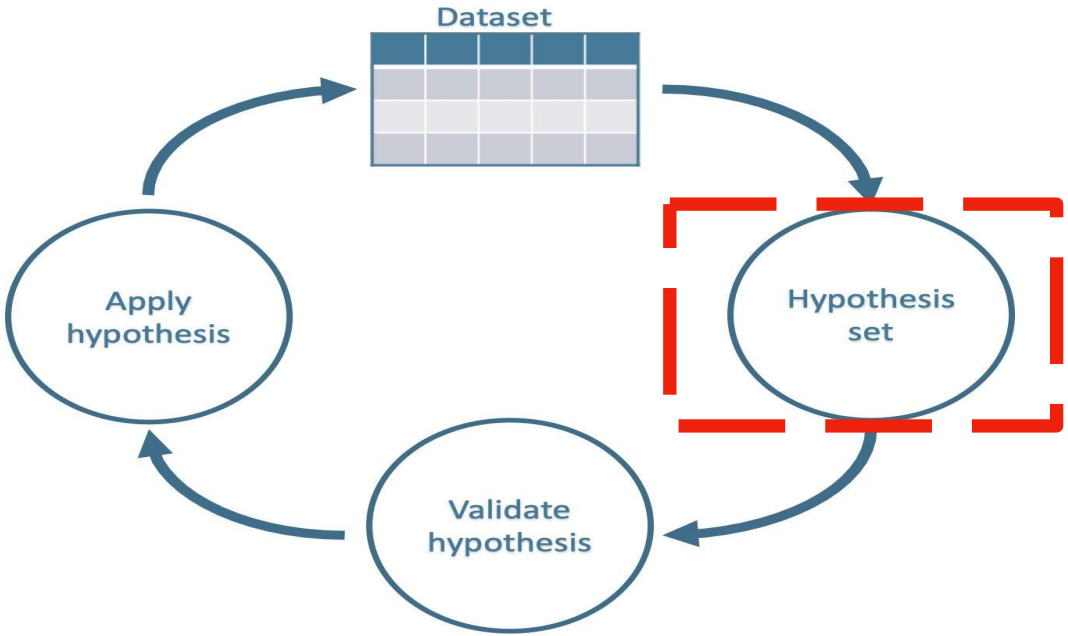


Ce qui n'est **pas** de l'ingénierie de caractéristiques

1. Collecte de données
2. Création de la variable cible (et métrique de performance)
3. Suppression des doublons
4. Correction des classes mal étiquetées.

Source : https://www.slideshare.net/0xdata/feature-engineering-83511751

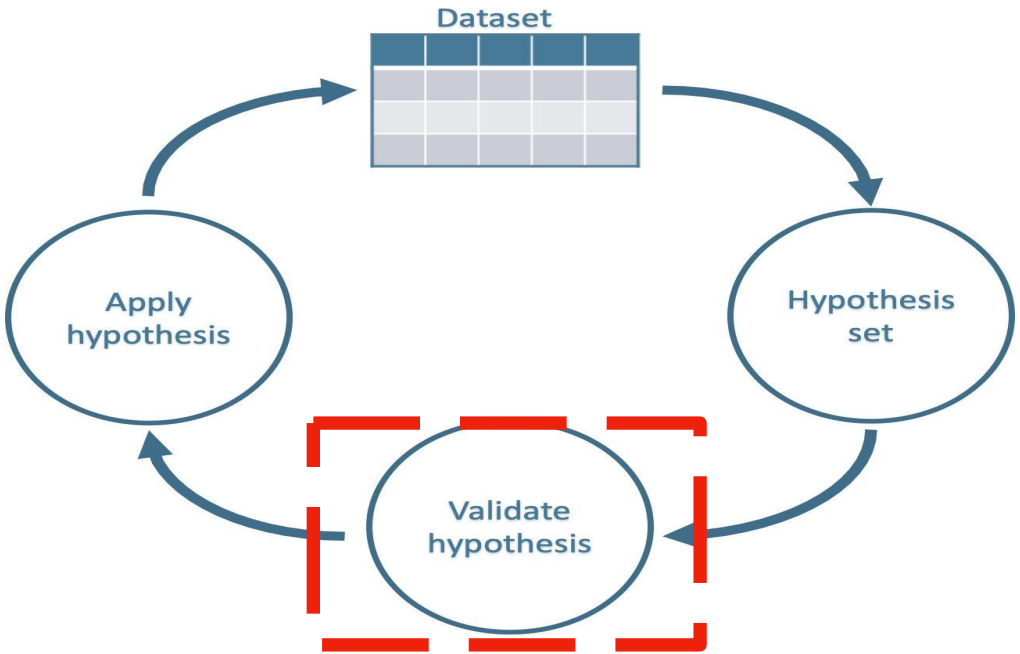
Cycle d'ingénierie des caractéristiques



Comment?

- Connaissance du domaine
- Expérience
- Visualization/exploration des données.
- Retour du modèle

Cycle d'ingénierie des caractéristiques



Comment?

- Validation croisée
- Mesure des métriques de performance
- Processus rigoureux pour éviter des fuites d'information.

L'ingénierie des caractéristiques est difficile

- De puissantes transformations de caractéristique (comme l'encodage de la sortie) peuvent introduire des fuites lorsqu'elles sont mal appliquées.
- Nécessite généralement une connaissance du domaine sur la façon dont les caractéristiques interagissent les unes avec les autres
- Prend du temps, il faut exécuter des milliers d'expériences

- Pourquoi l'ingénierie des caractéristique est importante
  - Extrayez nouvelles fonctionnalités plus efficace, supprimez les fonctionnalités non pertinentes ou bruyantes

    Des modèles plus simples avec de meilleurs résultats

Le nettoyage des données

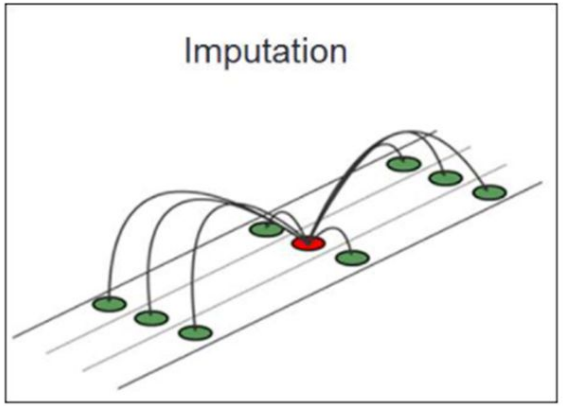
Le traitement des données manquantes

- Problème courant: bases de données avec des données manquantes.

- L'imputation des données
  - Mettre des valeurs génériques (0, Nan)
  - Utilise la distribution des variables présentes pour prédire les variables manquantes.
  - **Ne concerne pas la réponse y.**

Données connues

Donnée manquante à

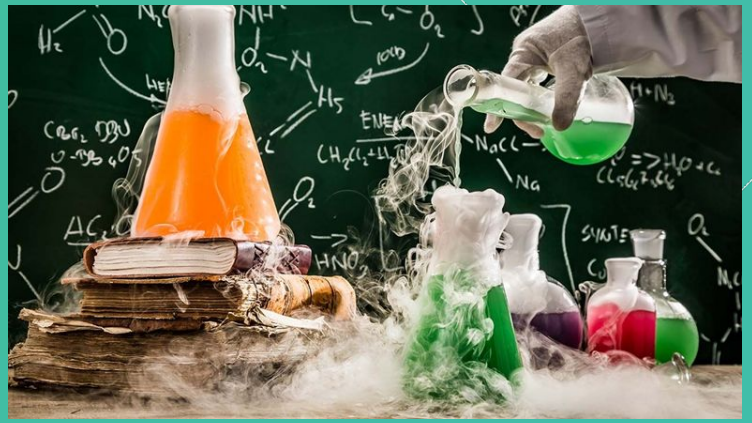


Exemple d'une base de données avec des données manquantes

	Caractéristiques (X)					Réponse (Y)
	Surface	Nombre chambres	Quartier résidentiel	…	Nombre écoles	Prix
Maison 1	x_1^1	x_2^1	x_3^1	…	x_{10}^1	250 000
Maison 2	x_1^2	x_2^2	?	…	?	300 000
Maison 3	?	x_2^3	x_3^3	…	x_{10}^3	325 000
…	…	…	…	…	…	…
Maison 100	x_1^{100}	x_2^{100}	x_3^{100}	…	x_{10}^{100}	475 000

- Si on laisse tomber exemples avec des données manquantes, on perd 30 % des données!

La transformation des données

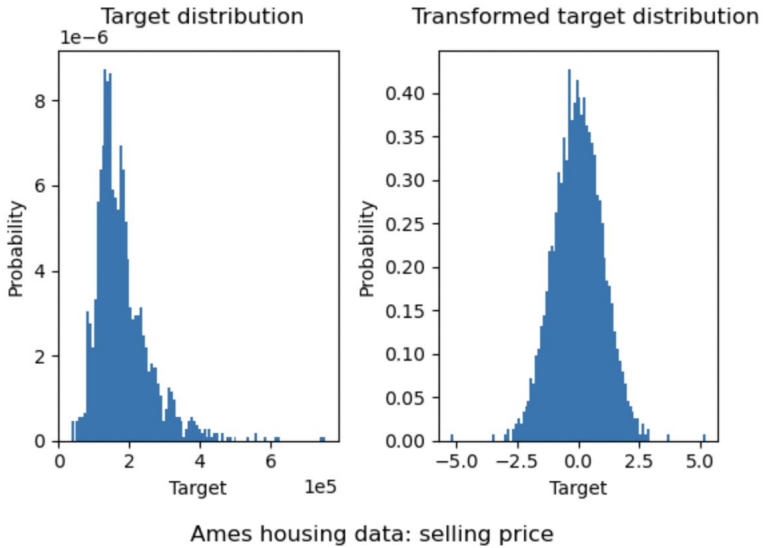


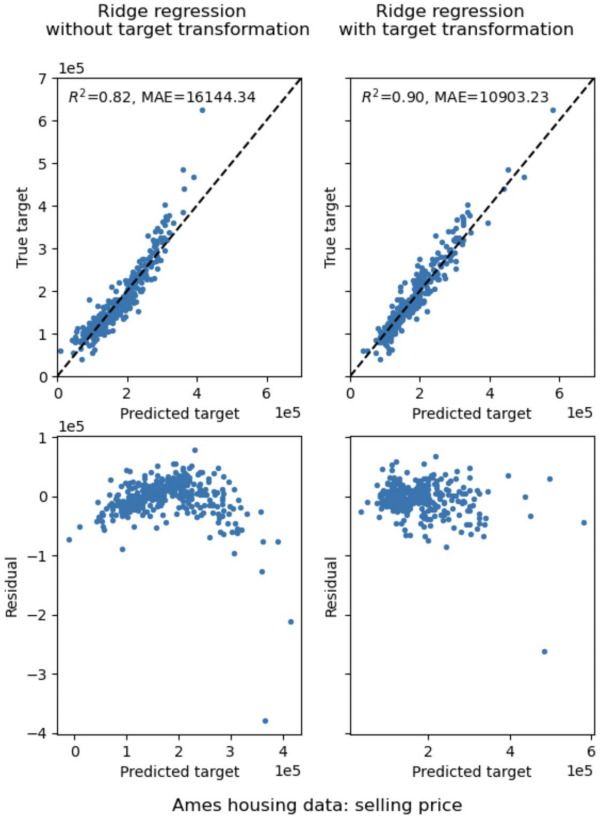
Transformation de la cible

- Transformation de la variable cible
- Utilisez-la lorsque la variable montre une distribution asymétrique pour rendre les résidus plus proches de la "distribution normale" (courbe en cloche).
- Échelle logarithmique utile lorsque la cible prendre des valeurs **très grandes et très petites.**
- Principalement utile pour les tâches de régression : peut améliorer l'ajustement du modèle
  par exemple, log(x), log(x+1), sqrt(x), sqrt(x+1), etc.

Exemple: Données Ames housing

source: https://scikit-learn.org/stable/auto_examples/compose/plot_transformed_target.html





Transformations des caractéristiques

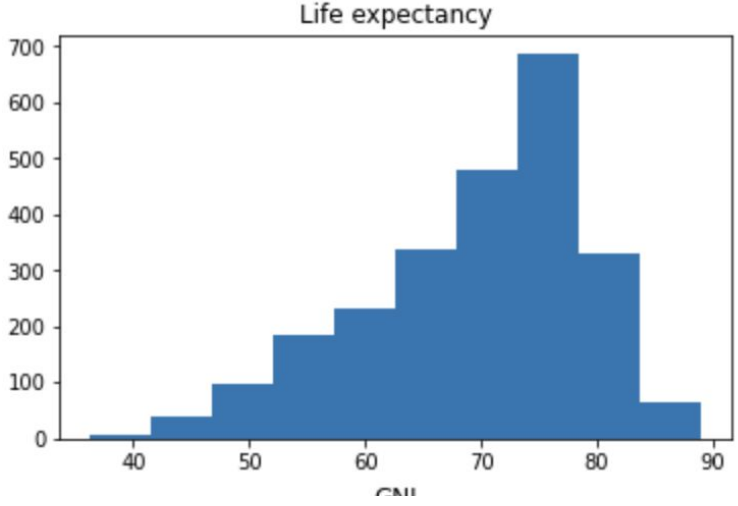
- Même idée.
- Transformation logarithmique (ou transformation log) :
- Cela aide à gérer les données asymétriques qui, après la transformation, deviennent plus proche d'une normale.
- Dans la plupart des cas, l'ordre de grandeur des données change dans la plage des données.
- Cela diminue également l'effet des valeurs aberrantes, en raison de la normalisation des différences de magnitude.

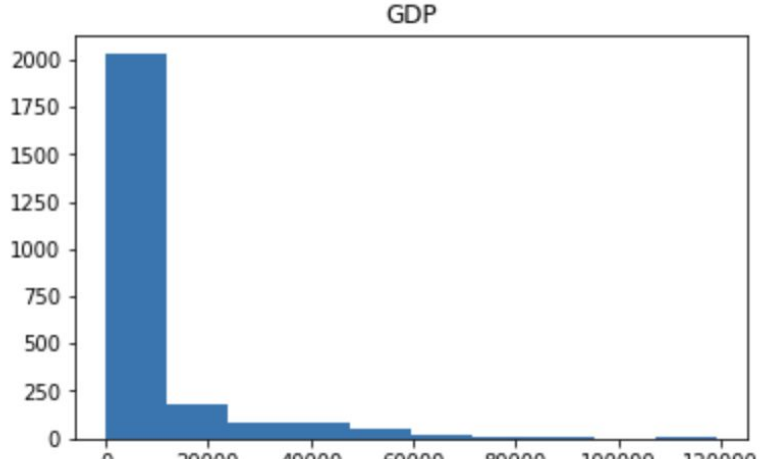
Example

jeu de données Kaggle:

- GDP (produit intérieur brut) et Espérance de vie.

Doit-on transformer une des deux caractéristique?





Transformations

- Les données doivent avoir des **valeurs positives**, sinon vous recevez une erreur.
  De plus, si les petites valeurs ne doivent pas être prise en compte vous pouvez ajouter 1 à vos données avant de les transformer.
  Ainsi, vous vous assurez que la sortie de la transformation est positive.

```
#Log Transform Example
data = pd.DataFrame({'value':[2,45, -23, 85, 28, 2, 35, -12]})


data['log+1'] = (data['value']+1).transform(np.log)
                                  ↳ s'assure que petites
                                    val. ø prisent en
#Negative Values Handling           compte ⇝ sortie positive
#Note that the values are different
data['log'] = (data['value']-data['value'].min()+1)
.transform(np.log)
```

	value	log(x+1)	log(x-min(x)+1)
0	2	1.09861	3.25810
1	45	3.82864	4.23411
2	-23	nan	0.00000
3	85	4.45435	4.69135
4	28	3.36730	3.95124
5	2	1.09861	3.25810
6	35	3.58352	4.07754
7	-12	nan	2.48491

Encodage des caractéristiques

- Transformez les caractéristiques catégorielles en caractéristiques numériques pour fournir des informations plus précises
- Aider à capturer explicitement les relations non linéaires et les interactions entre les valeurs des fonctionnalités
- La plupart des outils d'apprentissage automatique n'acceptent que des nombres comme entrée, par exemple xgboost, gbm, glmnet, libsvm, liblinear, etc.

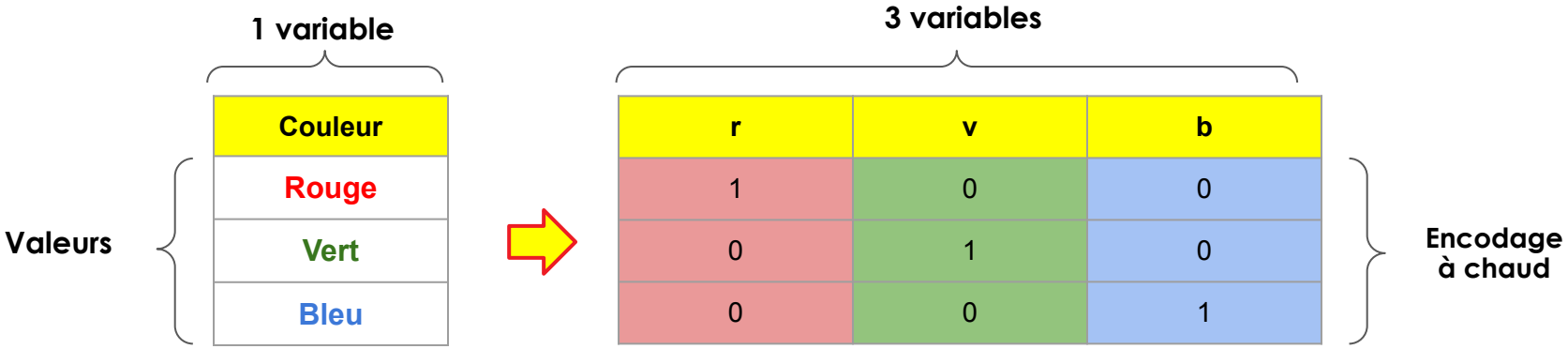
Conversion des valeurs catégoriques en valeurs numériques

- Les variables catégoriques contiennent des **catégories qui ne sont pas comparables.**
  - P. ex: Couleur = [Rouge, Vert, Bleu]
- Ces variables sont converties en variables factices binaires. (Encodage à chaud: *One-Hot Encoding*)

  - Conserve toutes les informations des variables catégoriques.
  - Convient aux modèles linéaires en AA.

  - Augmente la dimensionnalité des variables à traiter.
  - Peut introduire des informations redondantes.

Chaque catégories peut être transformée en **k** variables binaires factices (*dummy variables*)



Exemple: La couleur bleu est codée comme [0, 0, 1]

Attention: Les trois variables r, g et v sont **corrélées!**

- Représentation utilisée dans
  - Méthodes basées sur des arbres de décision
  - Méthodes de régression linéaire utilisant la sélection de variables

Conversion des valeurs ordinales en valeurs numériques

Les variables ordinales contiennent des **catégories dont l'ordre est important:**

- P. ex: note_finale = [A, B, C, D, E] avec A > B > C > D > E

On les convertit en **valeurs entières** (encodage ordinal)

- note_finale = [0, 1, 2, 3, 4] avec A = 4, B = 3, …

Représentation utilisée dans

- Méthodes basées sur des arbres de décision
- Méthodes de régression linéaire si
  - Beaucoup de catégories
  - Les valeurs numériques implicites des catégories sont distribuées linéairement

Encodage de fréquence

- Encodage des niveaux catégoriels de caractéristiques en valeurs comprises entre 0 et 1 en fonction de leur fréquence relative

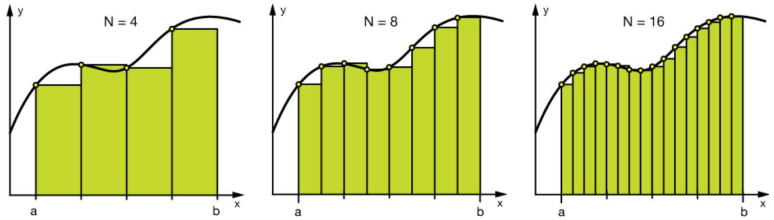
A	0.44 (4 out of 9)
B	0.33 (3 out of 9)
C	0.22 (2 out of 9)

Feature	Encoded Feature
A	0.44
A	0.44
A	0.44
A	0.44
B	0.33
B	0.33
B	0.33
C	0.22
C	0.22

Utile si: Beaucoup de valeurs pour la caractéristique. Ces valeurs ont une fréquence différente.
Example: pays, catégorie de produit au supermarché.

Regroupement (Binning)

- Le regroupement peut être appliqué à la fois aux données catégoriques et numériques :



Binning illustration of numerical data

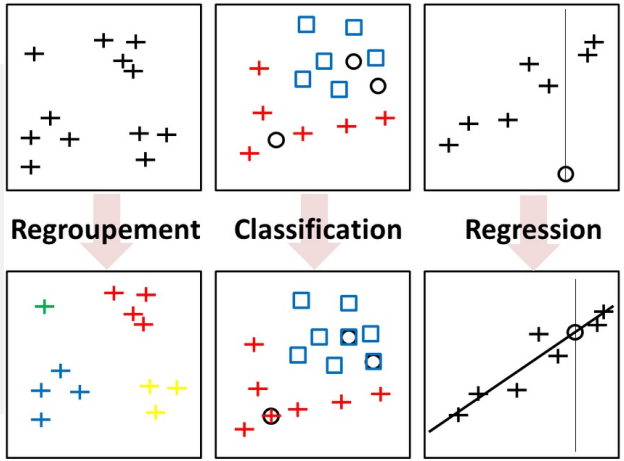
- Exemple

```
#Numerical Binning Example

Value      Bin
0-30   ->  Low
31-70  ->  Mid
71-100 ->  High

#Categorical Binning Example

Value      Bin
Spain  ->  Europe
Italy  ->  Europe
Chile  ->  South America
Brazil ->  South America
```



Regroupement (binning)

- La principale motivation du regroupement (binning) est de rendre le modèle plus robuste et d'éviter le surentraînement, cependant, cela a un coût sur les performances.
- Le compromis entre performance et surajustement est le point clé du processus de regroupement.
- Regroupement numérique : le regroupement peut être redondant en raison de son effet sur les performances du modèle.
- Regroupement catégorique : les étiquettes à basses fréquences affectent probablement négativement la robustesse des modèles statistiques. Ainsi, l'attribution d'une catégorie générale à ces valeurs moins fréquentes permet de conserver la robustesse du modèle.

Encodage moyen d'étiquette

- Au lieu d'encoder des variables catégoriques et d'**augmenter le nombre de caractéristiques**, nous pouvons encoder chaque niveau comme la moyenne de la réponse.

A	0.75 (3 out of 4)
B	0.66 (2 out of 3)
C	1.00 (2 out of 2)

Feature	Outcome	MeanEncode
A	1	0.75
A	0	0.75
A	1	0.75
A	1	0.75
B	1	0.66
B	1	0.66
B	0	0.66
C	1	1.00
C	1	1.00

Idée : Donner directement les informations pertinentes au modèle !

Problème : **peut conduire à du surapprentissage et à des bias!!!**

Encodage moyen d'étiquette

- Il est préférable de calculer la moyenne pondérée de la moyenne globale de l'ensemble d'apprentissage et de la moyenne du niveau :

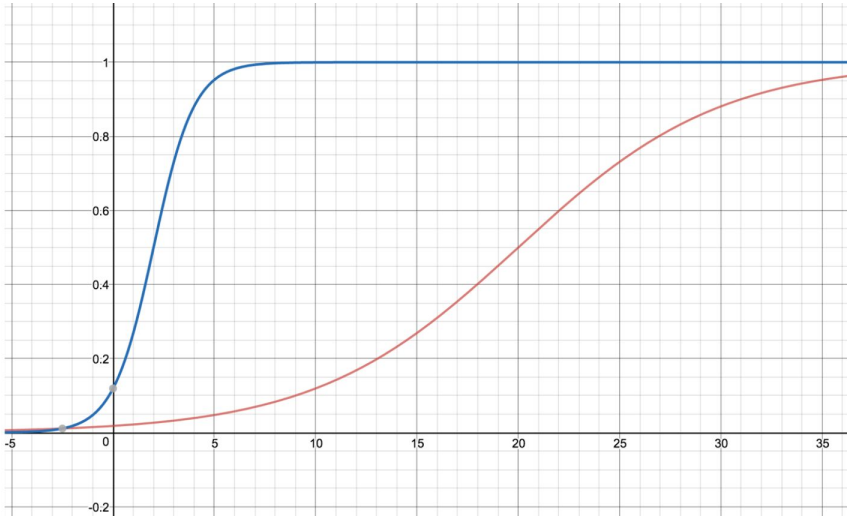
$$\lambda(n) * mean(level) + \big(1 - \lambda(n)\big) * mean(dataset)$$

- Les pondérations sont basées sur la fréquence des niveaux, c'est-à-dire que si une catégorie n'apparaît que quelques fois dans l'ensemble de données, sa valeur codée sera proche de la moyenne globale au lieu de la moyenne de ce niveau.
  n : nombre d'exemple dans la catégorie

Encodage moyen d'étiquette
Fonction de lissage λ

$$\frac{1}{1 + exp(\frac{-(x-k)}{f})}$$

x = frequency
k = inflection point
f = steepness



Encodage moyen d'étiquette: Lissage

$$\lambda = \frac{1}{1 + \exp(-\frac{(x-2)}{0.25})}$$

	x	level	dataset	λ	
A	4	0.75	0.77	0.99	0.99*0.75 + 0.01*0.77 = 0.7502
B	3	0.66	0.77	0.98	0.98*0.66 + 0.02*0.77 = 0.6622
C	2	1.00	0.77	0.5	0.5*1.0 + 0.5*0.77 = 0.885

$$\lambda = \frac{1}{1 + \exp(-\frac{(x-3)}{0.25})}$$

	x	level	dataset	λ	
A	4	0.75	0.77	0.98	0.98*0.75 + 0.01*0.77 = 0.7427
B	3	0.66	0.77	0.5	0.5*0.66 + 0.5*0.77 = 0.715
C	2	1.00	0.77	0.017	0.017*1.0 + 0.983*0.77 = 0.773

Feature	Outcome
A	1
A	0
A	1
A	1
B	1
B	1
B	0
C	1
C	1

Encodage moyen d'étiquette: schéma leave-one-out

- Pour éviter le surapprentissage, nous pourrions utiliser le schéma leave-one-out

Feature	Outcome
A	1
A	**0**
A	**1**
A	**1**
B	1
B	1
B	0
C	1
C	1

LOOencode
0.66

Encodage moyen d'étiquette: schéma leave-one-out

- Pour éviter le surajustement, nous pourrions utiliser le schéma leave-one-out

Feature	Outcome
A	**1**
A	0
A	**1**
A	**1**
B	1
B	1
B	0
C	1
C	1

LOOencode
0.66
1.00

Encodage moyen d'étiquette: schéma leave-one-out

- Pour éviter le surajustement, nous pourrions utiliser le schéma leave-one-out

Feature	Outcome
A	**1**
A	**0**
A	1
A	**1**
B	1
B	1
B	0
C	1
C	1

LOOencode
0.66
1.00
0.66

Encodage moyen d'étiquette: schéma leave-one-out

- Pour éviter le surajustement, nous pourrions utiliser le schéma leave-one-out

Feature	Outcome
A	**1**
A	**0**
A	**1**
A	1
B	1
B	1
B	0
C	1
C	1

LOOencode
0.66
1.00
0.66
0.66

Encodage moyen d'étiquette: schéma leave-one-out

- Pour éviter le surajustement, nous pourrions utiliser le schéma leave-one-out

Feature	Outcome
A	1
A	0
A	1
A	1
B	1
B	1
B	0
C	**1**
C	1

LOOencode
0.66
1.00
0.66
0.66
0.50
0.50
1.00
1.00
1.00

Encodage moyen d'étiquette

Quand l'utiliser?

- **Caractéristiques à cardinalité élevée :**
  a. caractéristique avec un **grand nombre de catégorie**s peut être difficile à encoder : one-hot génère trop de fonctionnalités
  b. Un encodage cible dérive des nombres pour les catégories à l'aide de la propriété la plus importante de l'entité : **sa relation avec la cible.**
- **Caractéristiques motivées par le domaine :** par expérience, vous pourriez soupçonner qu'une caractéristique catégorique devrait être importante même si elle a obtenu de mauvais résultats avec une métrique. Un encodage cible peut aider à révéler le véritable caractère informatif d'une caractéristique.

	tx_price	beds	baths	sqft	year_built	lot_size	property_type	exterior_walls	roof	basement	restaurants	groceries	nightlife
0	295850	1	1	584	2013	0	Apartment / Condo / Townhouse	Wood Siding	NaN	NaN	107	9	30
1	216500	1	1	612	1965	0	Apartment / Condo / Townhouse	Brick	Composition Shingle	1.0	105	15	6
2	279900	1	1	615	1963	0	Apartment / Condo / Townhouse	Wood Siding	NaN	NaN	183	13	31
3	379900	1	1	618	2000	33541	Apartment / Condo / Townhouse	Wood Siding	NaN	NaN	198	9	38
4	340000	1	1	634	1992	0	Apartment / Condo / Townhouse	Brick	NaN	NaN	149	7	22
5	265000	1	1	641	1947	0	Apartment / Condo / Townhouse	Brick	NaN	NaN	146	10	23
6	240000	1	1	642	1944	0	Single-Family	Brick	NaN	NaN	159	13	36
7	388100	1	1	650	2000	33541	Apartment / Condo / Townhouse	Wood Siding	NaN	NaN	198	9	38
8	240000	1	1	660	1983	0	Apartment / Condo / Townhouse	Brick	NaN	NaN	51	8	6
9	250000	1	1	664	1965	0	Apartment / Condo / Townhouse	Brick	NaN	NaN	119	10	26