

# Préparation des données

Année universitaire  
2020/2021



## Préparation des données

1. Nettoyage des données
2. Feature selection
3. Data transformation
4. Feature engineering
5. Dimensionality reduction

# Nettoyage des données

Il s'agit d'identifier et corriger les erreurs relatives à l'insertion des données

La qualité des données est très importante. Les données doivent être:

- valide (obéir aux différentes contraintes et règles du métiers)
- pertinence ( si les valeurs sont dans logiques, raisonnables)
- complète (il n'y a pas de valeurs manquantes)
- consistante ( il n'y a pas de contradiction dans les valeurs)
- uniforme (décrit avec les mêmes unités de mesures)



# Nettoyage des données

## Inspection des données

Il faut faire des visualisations, des statistiques et des interprétations

- La médiane (plus robuste aux outliers)
- La moyenne (utilisée lorsque les données ne sont pas biaisées (skewed))
- Données catégorique: le mode est la médiane
- Les outliers
- Si la distribution des données soit normale
- La variance des données
- La fréquence de certaines colonnes
- La corrélation qui existe entre les attributs

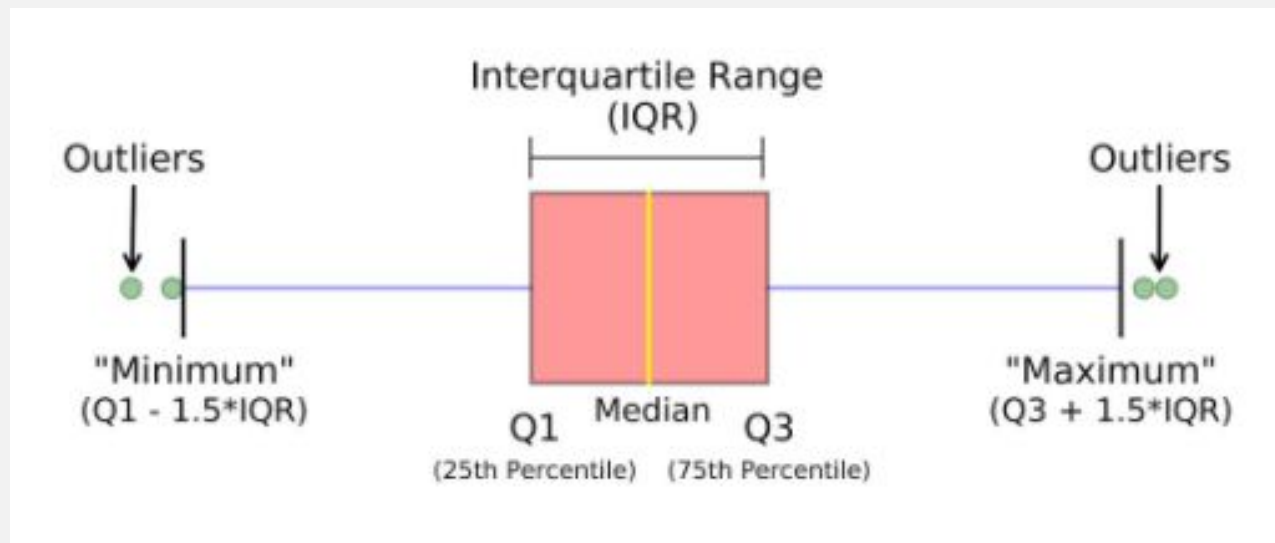
Utilisées pour voir la tendance centrale des données

Indique la distribution des données par rapport à la moyenne

# Nettoyage des données

## Inspection des données

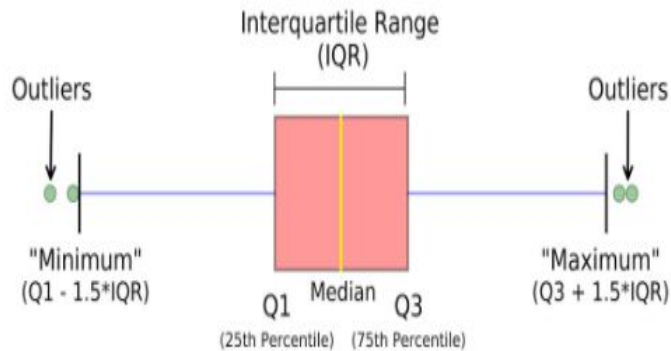
Boite à moustaches ou boxplot



# Nettoyage des données

## Inspection des données

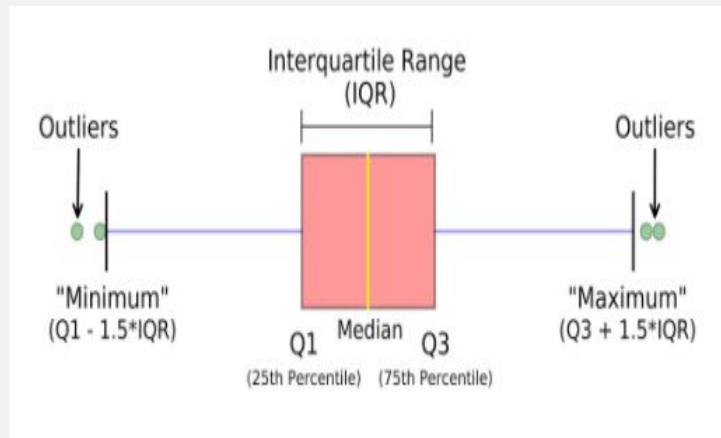
- lorsque la boîte à moustaches est courte, cela implique que la plupart de vos points de données sont similaires, car il existe de nombreuses valeurs dans une petite plage.



- Lorsque la boîte à moustaches est haute, cela implique que la plupart de vos points de données sont assez différents, car les valeurs sont réparties sur une large plage
- Si la valeur médiane est plus proche du bas, nous savons que la plupart des données ont des valeurs inférieures. Si la valeur médiane est plus proche du sommet, nous savons que la plupart des données ont des valeurs plus élevées. Fondamentalement, si la ligne médiane n'est pas au milieu de la boîte, c'est une indication de dat biaisé

# Nettoyage des données

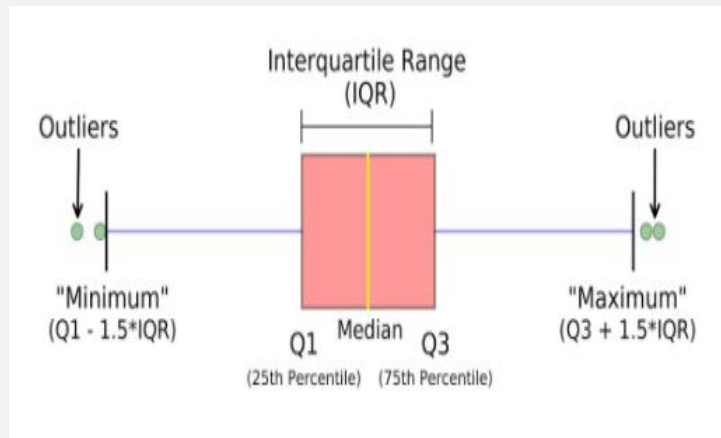
## Inspection des données



Les moustaches sont-elles très longues? Cela signifie que vos données ont un écart type et une variance élevés, c'est-à-dire que les valeurs sont réparties et varient fortement. Si vous avez de longues moustaches d'un côté de la boîte mais pas de l'autre, vos données peuvent varier fortement dans une seule direction.

# Nettoyage des données

## Inspection des données



Si les boxplots sont longues, cela signifie que vos données ont un écart type et une variance élevés, c'est-à-dire que les valeurs sont réparties et varient fortement. Si vous avez de longues moustaches d'un côté de la boîte mais pas de l'autre, vos données peuvent varier fortement dans une seule direction.



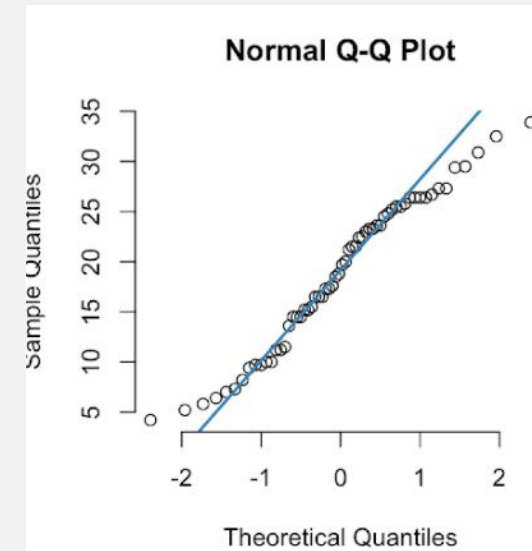
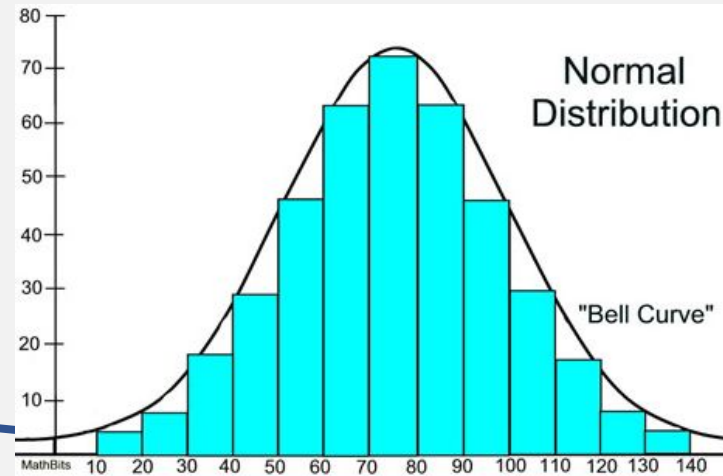
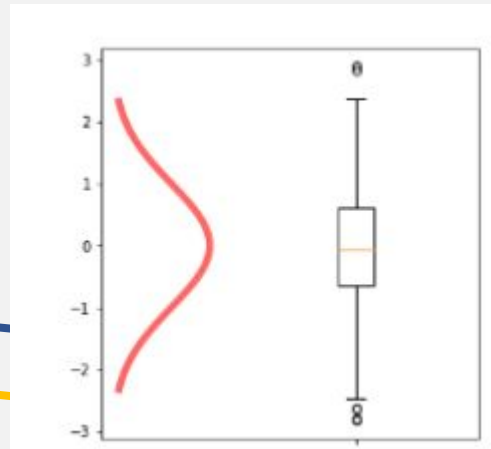
# Nettoyage des données

## Inspection des données

La distribution normale des données

Une seule variable: on utilise QQplot, histogramme  
Plusieurs variable: on utilise boxplot

Si les données sont biaisées d'une manière exagérée, on peut introduire le logarithme sur les données et réobserver



# Nettoyage des données

## Inspection des données

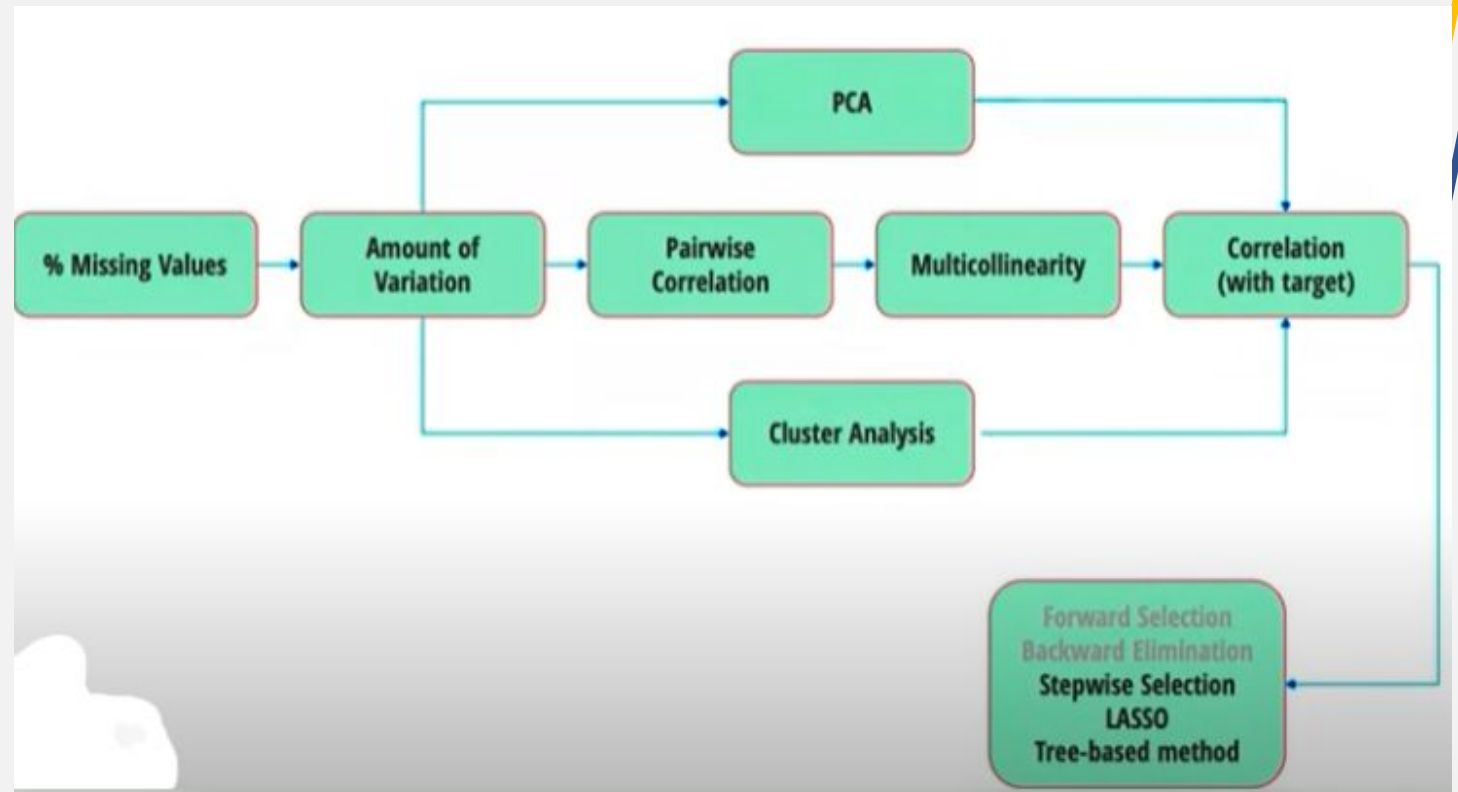
Des données manquantes

- Supprimer les données manquantes (lignes ou colonnes)
- Remplacer par la plus fréquente valeur ou par la moyenne
-

# Feature selection

Il s'agit d'identifier les variables les plus pertinents par rapport à l'objectifs:

- pair-wise correlation,
- variable target correlation,
- forward/backward and stepwise selection
- Multivariate correlation.
- Lasso
- Tree-based selection



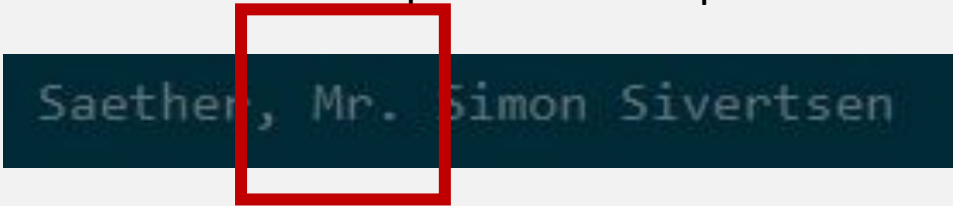
# Data transformation

- Unifier l'unité de mesure de variables.
- Transformer une variable qualitative en une variable numérique
- Transformer du texte en une matrice numérique
- Changer l'échelle ou la distribution des variables
- Etaler une variable temporelle sur plusieurs intervalles
- 
-

# Feature engineering

Déduire de nouvelles variables à partir des données (par exemple faire du web scrapping)

Décomposer des données textuelles non structurées en plusieurs champs en utilisant les expressions régulières



Saether, Mr. Simon Sivertsen

Si on souhaite rassembler les données par une caractéristique particulière pour en déduire la fréquence de certaines données et utiliser cette colonne après comme nouvelle variable.

(le cas d'un clustering, labellisation manuelle, et utilisation de cette nouvelle variable créée)

# Dimensionality reduction

Création d'un espace de projection compact des données **ACP**