
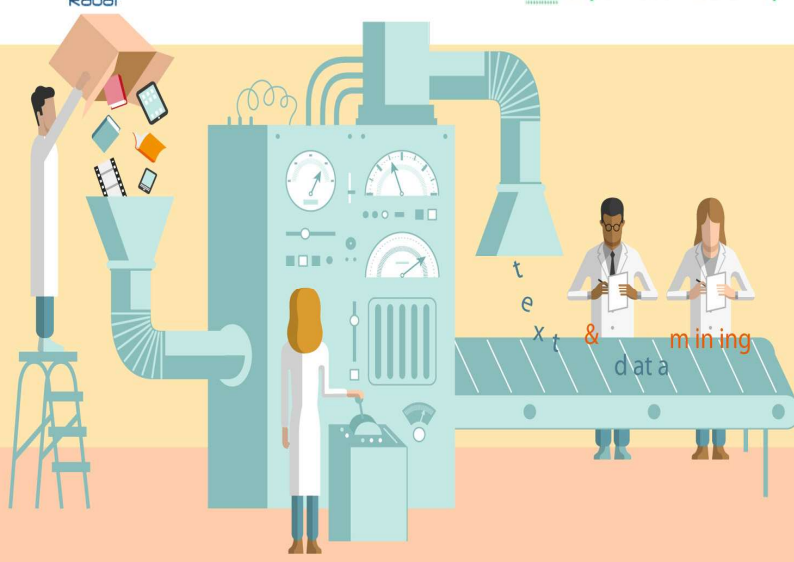


Université Mohammed V  
Faculté des Sciences  
Rabat

**Royaume du Maroc**  
Université Mohammed V de Rabat  
Faculté des Sciences  
Département d'Informatique



**IPSS**  
**Intelligent Processing  
Systems & Security**



# Data Mining & Machine Learning

Master IPS  
Faculté des sciences – Rabat  
Université Mohamed V

45

**Chapitre 2**

# Data Mining : Compréhension des données

De la donnée à la connaissance

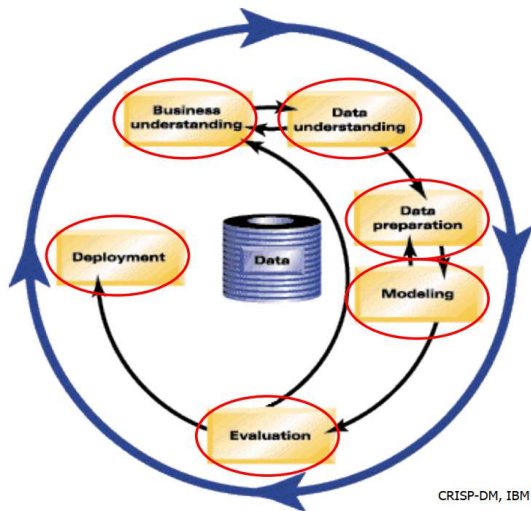




[ipes.boutyour@gmail.com](mailto:ipes.boutyour@gmail.com) 46

46

## Rappel : MÉTHODE CRISP-DM



CRISP-DM, IBM Official Site

1. Compréhension du métier
2. Compréhension des données
3. Constitution du Data Hub
4. Modélisation
5. Evaluation
6. Déploiement

ipes.boutyour@gmail.com

47

47

## Compréhension du métier

- Bien comprendre l'enjeu métier (fidélisation des clients, détection de fraudes, augmentation des ventes d'un produit, etc.)
- Bien évaluer la situation (ressources, prérequis, contraintes, risques, coûts, gains, etc.)
- Impliquer les experts du domaine en question
- Bien définir l'objectif du data mining

ipes.boutyour@gmail.com

48

48

## Compréhension des données

### Variables

- Une variable est une propriété ou caractéristique d'un individu
  - Exemple : Couleur des yeux d'une personne, température, état civil, ...
- Une collection de variables décrivant un individu. On dit individu ou enregistrement, point, cas, objet, entité, observation

Variables

age	Revenus	Etudiant	Taux crédit	Achat_PC
<=30	élevé	non	faible	non
<=30	élevé	non	excellent	non
31...40	élevé	non	faible	oui
>40	moyen	non	faible	oui
>40	faible	oui	faible	oui
>40	faible	oui	excellent	non
31...40	faible	oui	excellent	oui
<=30	moyen	non	faible	non
<=30	faible	oui	faible	oui
>40	moyen	oui	faible	oui
<=30	moyen	oui	excellent	oui
31...40	moyen	non	excellent	oui
31...40	élevé	oui	faible	oui
>40	moyen	non	excellent	non

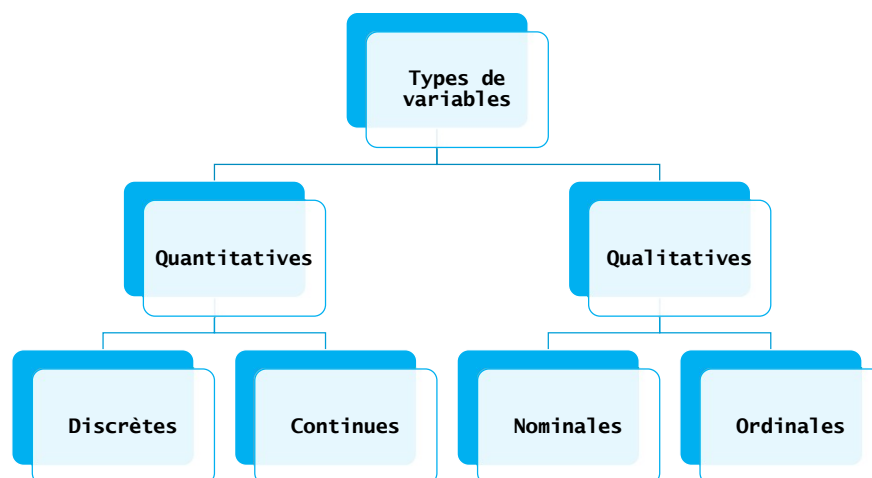
ipes.boutyour@gmail.com

49

49

## Compréhension des données

### Types de variables



ipes.boutyour@gmail.com

50

50

# Compréhension des données

## Types de variables

### ○ Variables quantitatives

- Les valeurs sont des nombres qui peuvent être ordonnés et additionnés
- Mesurables par une unité physique
- Exemples: salaire, poids, taille, proportion, quantité, durée de séjour, etc.

# Compréhension des données

## Types de variables

### ○ Variables quantitatives

#### ■ Variables **continues**

- Elles forment un sous-ensemble infini de  $\mathbb{R}$  (ex: salaire)
- Elles sont ordonnées (on peut les comparer par la relation d'ordre  $<$ )
- On peut effectuer des opérations arithmétiques

#### ■ Variables **discrètes**

- Elles forment un sous-ensemble fini ou infini dénombrable de  $\mathbb{N}$  (ex: nombre d'enfants)
- Elles sont ordonnées (on peut les comparer par la relation d'ordre  $<$ )
- On peut effectuer des opérations arithmétiques

Poids
41,5
33,4
37,5
33,5
39,7
30,8
37,4
38,2
43
38,5

Nombre de frères et sœurs
1
2
3

# Compréhension des données

## Types de variables

### ○ Variables qualitatives

- Les valeurs sont des qualités appelées modalités
- Les modalités peuvent être sous format numérique ou alphanumérique
- Non mesurables par une unité physique; caractéristique de l'individu
- Exemples: sexe, profession, couleur des yeux, couleur des cheveux, etc.

# Compréhension des données

## Types de variables

### ○ Variables qualitatives

- Variables **ordinales**
  - Les modalités peuvent être ordonnées (ex: «faible, moyen, fort»)
  - Il n'est pas possible de calculer la distance entre les modalités
  - Elles sont souvent traitées comme données discrètes
- Variables **nominales**
  - Les modalités ne peuvent pas être ordonnées (ex: profession)
- Variables de **type intervalle (avec transformation)**
  - Elles sont numériques (ex: durée de vie)
  - Il est possible de calculer la distance entre les modalités

Force
Faible
Normal
Fort
Très fort
Invincible

Situation familiale
Célibataire
Divorcé
Marié
Veuf

## Compréhension des données

### Types de variables (Application)

- Nationalité d'un individu
- Temps de réalisation d'un travail
- Nombre d'étudiant dans une classe
- Nom de la couleur
- Date de naissance
- Degré de satisfaction
- Distance parcourue
- Etat civil
- Durée de voyage
- Sexe d'un individu
- Marque de voiture
- Nombre de bonnes réponses à un examen
- Classement dans une compétition
- Taille d'un individu

## Compréhension des données

### Types de variables

#### ○ Variables **qualitatives**

- Le type d'une variable qualitative est déterminé par la façon avec laquelle on la mesure.
- Exemple: variable « **Education** »
  - **Nominale** : privée, publique
  - **Ordinale** : niveau d'études atteint
  - **Intervalle** : nombre d'années d'études après le BAC

## Compréhension des données

### Étapes :

1. Collecter les données
2. Décrire les données
3. Explorer les données
  - Valeurs manquante
  - Analyse univariée
  - Analyse bivariée
4. Vérifier la qualité des données

## Compréhension des données

### Première étape: collecter les données

- Bases de données classiques (Relationnelles, Transactionnelles)
- Bases de données avancées (Objet, Objet-relationnelles, Spatiales, Séries temporelles, Textes, Multimédia, Hétérogènes)
- Log de pages web
- Entrepôts de données (DW)

## Compréhension des données

### Deuxième étape: décrire les données

- Vérifier le volume des données et leurs propriétés générales
- Vérifier les différents attributs et découvrir leurs ordres de grandeurs.

## Compréhension des données

### Troisième étape: explorer les données

- Analyser en détail les variables en utilisant la **statistique descriptive**
  - **Statistique univariée** pour analyser en détail les propriétés d'une variable
  - **Statistique bivariée** pour analyser la relation entre deux variables



## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée:

- Décrire et résumer chaque variable
- Généraliser les informations à la population entière
- Détecter les anomalies (valeurs rares, manquantes, aberrantes, extrêmes)
  - **Valeur aberrante:** valeur erronée à cause d'une mauvaise mesure, erreur de calcul, fausse déclaration, ...
  - **Valeur extrême:** valeur très supérieure ou inférieure par rapport à l'ordre de grandeur des observations de la variable

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée:

- **Valeur extrême:** pas forcément aberrante
  - Peut correspondre à une catégorie particulière d'individus
  - Profil rare, intéressant à détecter (fraude, impayé, niche...)
  - Affecte certaines techniques qui se basent sur les calculs de la variance (régression logistique, analyse discriminante)
- **Valeur manquante**
  - Mauvais fonctionnement de l'équipement
  - Non saisie car non/mal comprise
  - Considérée peu importante au moment de la saisie

## Compréhension des données

## Troisième étape: explorer les données

- Analyser univariée: Variable qualitative

- **Effectif** : nombre d'individus de l'échantillon pour chaque modalité
- **Effectif total** : nombre de valeurs dans la série statistique.
- **Fréquence**: effectif ramené à la taille de l'échantillon (%) ( effectif de la modalité / effectif total )
- **Exemple** :
  - Prenons la série: bleu, noir, bleu, vert, noir, rouge, vert , bleu, noir, noir
  - L'effectif total =
  - L'effectif de la valeur bleu = sa fréquence =
  - L'effectif de la valeur vert = sa fréquence =

## Compréhension des données

## Troisième étape: explorer les données

- Analyser univariée: Variable qualitative

- **Effectif** : nombre d'individus de l'échantillon pour chaque modalité
- **Effectif total** : nombre de valeurs dans la série statistique.
- **Fréquence**: effectif ramené à la taille de l'échantillon (%) ( effectif de la modalité / effectif total )
- **Exemple** :
  - Prenons la série: bleu, noir, bleu, vert, noir, rouge, vert , bleu, noir, noir
  - L'effectif total = 10
  - L'effectif de la valeur bleu = 3 sa fréquence = 3/10
  - L'effectif de la valeur vert = 2 sa fréquence = 2/10

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

1. **Indicateurs de tendance centrale** (moyenne, médiane, mode, minimum, maximum, quartile)
2. **Indicateurs de Indicateurs de dispersion** (étendue, variance, écart-type, coefficient de variation)
3. **Indicateurs de forme de la distribution** (asymétrie, aplatissement)

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### 1. Indicateurs de tendance centrale

- Moyenne : somme des observations / taille de l'échantillon
- Médiane : valeur qui partage l'échantillon en deux parties égales
- Mode : valeur la plus fréquente (variable discrète) ou classe la plus dense (variable continue)
  - Calcul de la **médiane** (n taille échantillon;  $X_i$  i<sup>e</sup> observation) :
    - Classer les observations par ordre croissant
    - n impair  $\rightarrow$  médiane = valeur de l'observation centrale  $X_{(n+1)/2}$
    - n pair  $\rightarrow$  médiane = moyenne des deux valeurs centrales  $(X_{n/2} + X_{n/2 + 1})/2$

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### 1. Indicateurs de tendance centrale

- ❖ Moyenne : sensible aux valeurs extrêmes/aberrantes et à la forme de la distribution

Exemple : 5 personnes âgées de 34, 35, 37, 39, et 100 ans

→ Privilégier la médiane en cas de distribution asymétrique.

- ❖ Médiane : à utiliser dans le cas de variable discrète / variable qualitative ordinale avec un nombre important de modalités

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### 1. Indicateurs de tendance centrale

- ❖ Exemple: Série de 10 observations (âge)

35; 28; 29; 29; 30; 31; 35; 35; 27; 39

Déterminer la moyenne

Déterminer la médiane

Déterminer le mode

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### 1. Indicateurs de tendance centrale

❖ Exemple: Série de 10 observations (âge)

35; 28; 29; 29; 30; 31; 35; 35; 27; 39

Déterminer la moyenne =  $(35+28+29+29+30+31+35+35+27+39) / 10 = 31.8$

Déterminer la médiane :

ordre : 27; 28; 29; 29; 30; 31; 35; 35; 35; 39

$n=10$  (pair) → médiane =  $(30+31)/2 = 30.5$

Déterminer le mode : 35 se répète le plus souvent

Le mode est 35

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### 1. Indicateurs de tendance centrale

❖ Quartiles: En statistique descriptive, un quartile est chacune des 3 valeurs qui divisent les **données triées** en 4 parts égales, de sorte que chaque partie représente 1/4 de l'échantillon de population.

❖ Le quartile est calculé en tant que 4-quantile. Donc :

- le 1er quartile sépare les 25 % inférieurs des données ;
- le 2e quartile est la médiane de la série ;
- le 3e quartile sépare les 75 % inférieurs des données.

❖ La différence entre le 3e quartile et le 1er quartile s'appelle écart interquartile ; c'est un critère de dispersion de la série.

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### 1. Indicateurs de tendance centrale

###### ❖ Quartiles:

- Dans le cas continu on utilise la fonction représentative du **polygone des fréquences cumulées**.
- Dans le cas discret, on **range** les données par **ordre croissant** ensuite : Le quartile inférieur est la valeur du milieu du premier ensemble, dans lequel **25 %** des valeurs sont inférieures à **Q1** et 75 % lui sont supérieures. Le premier quartile prend la notation Q1. Le quartile supérieur est la valeur du milieu du deuxième ensemble, dans lequel **75 %** des valeurs sont inférieures à **Q3** et 25 % lui sont supérieurs. Le troisième quartile prend donc la notation Q3

ipes.boutyour@gmail.com

71

71

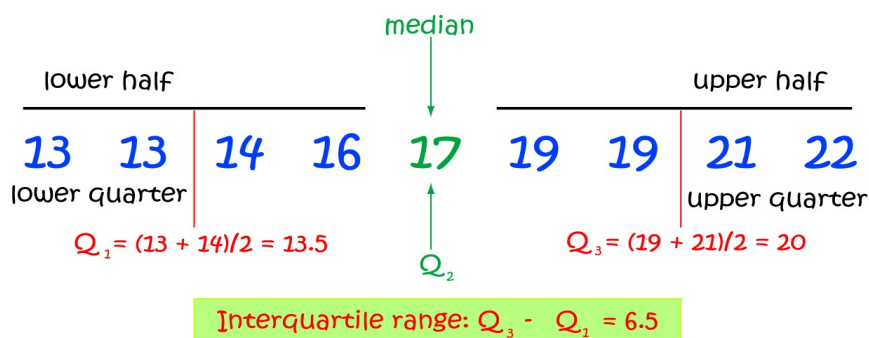
## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### 1. Indicateurs de tendance centrale

###### ❖ Quartiles:



ipes.boutyour@gmail.com

72

72

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### 1. Indicateurs de tendance centrale

###### ❖ Quartiles:

- Exemple : Soient les 12 observations suivantes :

57, 11, 15, 34, 24, 20, 28, 19, 37, 47, 50, 1

Les valeurs dans **l'ordre ascendant** : 1, 11, 15, 19, 20, 24, 28, 34, 37, 47, 50, 57

Q1 est entre 15 et 19 donc : **Q1 = 17**

Q2 est entre 24 et 28 donc : **Q2 = 26** (c'est la médiane)

Q3 est entre 37 et 47 donc : **Q3 = 42**

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### 1. Indicateurs de tendance centrale

###### ❖ Quartiles:

- **Exemple 2:** On a interrogé 20 élèves en leur demandant leur pointure. On a trié les résultats dans le tableau suivant :

Pointure	35	36	38	39	40
effectif	1	5	4	7	3
Effectif cumulé					

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### 1. Indicateurs de tendance centrale

###### ❖ Quartiles:

- [Exemple 2](#): On a interrogé 20 élèves en leur demandant leur pointure. On a trié les résultats dans le tableau suivant :

Pointure	35	36	38	39	40
effectif	1	5	4	7	3
Effectif cumulé	1	6	10	17	20

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### 1. Indicateurs de tendance centrale

###### ❖ Quartiles:

- [Exemple 2](#): On a interrogé 20 élèves en leur demandant leur pointure. On a trié les résultats dans le tableau suivant :

Pointure	35	36	38	39	40
effectif	1	5	4	7	3
Effectif cumulé	1	6	10	17	20

Médiane =

Q1 =

Q3 =



## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### 1. Indicateurs de tendance centrale

###### ❖ Quartiles:

- [Exemple 2](#): On a interrogé 20 élèves en leur demandant leur pointure. On a trié les résultats dans le tableau suivant :

Pointure	35	36	38	39	40
effectif	1	5	4	7	3
Effectif cumulé	1	6	10	17	20

Médiane =  
Q1 =  
Q3 =

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### ▪ Indicateurs de tendance centrale

###### ❖ Quartiles:

- [Exemple 2](#): On a interrogé 20 élèves en leur demandant leur pointure. On a trié les résultats dans le tableau suivant :

Pointure	35	36	38	39	40
effectif	1	5	4	7	3
Effectif cumulé	1	6	10	17	20

Médiane = 38.5  
Q1 = 36  
Q3 = 39

$N=20$  et  $N/4=5$  ; donc le premier quartile est la 5e valeur, soit **36**  
 $N=20$  et  $3N/4=15$  ; donc le troisième quartile est la 15e valeur, soit **39**

## Compréhension des données

### Troisième étape: explorer les données

- Analyser univariée: Variable quantitative

#### 2. Indicateurs de dispersion

- ❖ Évaluent la répartition des observations autour des valeurs centrales
- ❖ Étendue, variance, écart-type, écart interquartile, coefficient de variation

## Compréhension des données

### Troisième étape: explorer les données

- Analyser univariée: Variable quantitative

#### ■ Indicateurs de dispersion : étendue

- ❖ Différence entre la plus grande et la plus petite des valeurs observées (maximum - minimum)
- ❖ Basée uniquement sur les extrêmes → Très sensible aux extrêmes  
→ Souvent peu significative

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

#### 2. Indicateurs de dispersion : variance (dénoté $v$ ou $s^2$ )

- ❖ Mesure la dispersion autour de la moyenne
- ❖ La moyenne des carrés des écarts par rapport à la moyenne
- ❖ Plus les données sont concentrées autour de la moyenne, plus la variance est faible.

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

#### 2. Indicateurs de dispersion : Ecart-type (dénoté $s$ ou $\sigma$ )

- ❖ Mesure la dispersion autour de la moyenne
  - Dans le cas d'une population entière, l'écart type est obtenu en appliquant la formule suivante :

$$\sigma = \sqrt{\left(\frac{\sum (xi - \mu)^2}{n}\right)}$$

dans laquelle  $\mu$  désigne la moyenne arithmétique de la distribution et  $n$  le nombre de données dans cette population.

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

#### 2. Indicateurs de dispersion : Ecart-type (dénnoté $s$ ou $\sigma$ )

❖ Mesure la dispersion autour de la moyenne

- Dans le cas d'un échantillon de cette distribution, l'écart type est obtenu en appliquant la formule suivante :

$$s = \sqrt{\left(\frac{\sum (xi - \bar{x})^2}{n-1}\right)}$$

dans laquelle  $\bar{x}$  désigne la moyenne des données de l'échantillon et  $n$  désigne le nombre de données considérées.

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

#### 2. Indicateurs de dispersion : écart interquartile

❖ Mesure la taille de l'intervalle situé au centre de la série et incluant 50% des observations :

$$\text{Ecart interquartile} = Q3 - Q1$$

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### 2. Indicateurs de dispersion : coefficient de variation

- ❖ Rapport de l'écart-type à la moyenne de la distribution en %
- ❖ Utile pour comparer la dispersion des variables
- ❖ On dit qu'une variable X est dispersée si :

$$CV(X) > 25\%$$

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

##### 3. Indicateurs de forme de la distribution :

- ❖ Asymétrie
- ❖ Aplatissement

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

#### 3. Indicateurs de forme de la distribution : **Asymétrie**

- ❖ Mesure l'asymétrie d'une distribution
- ❖ Coefficient de symétrie (Skewness) :

$$\frac{1}{n} \sum_i \left( \frac{x_i - \bar{x}}{\sigma_x} \right)^3$$

- **Skewness = 0** → distribution normale
- **Skewness > 0** → distribution asymétrique à droite
- **Skewness < 0** → distribution asymétrique à gauche

ipes.boutyour@gmail.com

87

87

## Compréhension des données

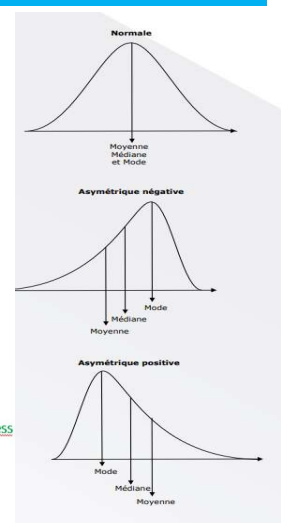
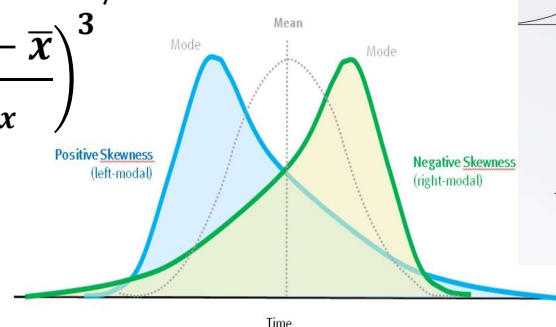
### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

#### 3. Indicateurs de forme de la distribution : **Asymétrie**

- ❖ Mesure l'asymétrie d'une distribution
- ❖ Coefficient de symétrie (Skewness) :

$$\frac{1}{n} \sum_i \left( \frac{x_i - \bar{x}}{\sigma_x} \right)^3$$



ipes.boutyour@gmail.com

88

88

## Compréhension des données

### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

#### 3. Indicateurs de forme de la distribution : Coefficient d'aplatissement

- ❖ Mesure le relief ou la platitude d'une courbe issue d'une distribution de fréquences
- ❖ Coefficient d'aplatissement (Kurtosis)

$$\frac{1}{n} \sum_i \left( \frac{x_i - \bar{x}}{\sigma_x} \right)^4$$

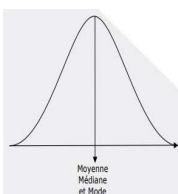
- **Kurtosis = 3** → aplatie comme une distribution normale (d.n)
- **Kurtosis > 3** → plus concentrée qu'une (d.n)
- **Kurtosis < 3** → plus aplatie qu'une (d.n)

## Compréhension des données

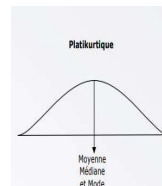
### Troisième étape: explorer les données

#### ○ Analyser univariée: Variable quantitative

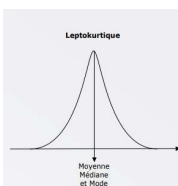
#### 3. Indicateurs de forme de la distribution : Coefficient d'aplatissement



- ⊙ Mésokurtique: courbe normale (cloche)



- ⊙ Platikurtique: courbe plate
  - > les cas s'éloignent de la moyenne
  - > forte variation : distribution relativement hétérogène



- ⊙ Leptokurtique: courbe élancée
  - > haute concentration de cas qui prennent les valeurs égales ou proches de la moyenne
  - > peu de variation : distribution relativement homogène

## Compréhension des données

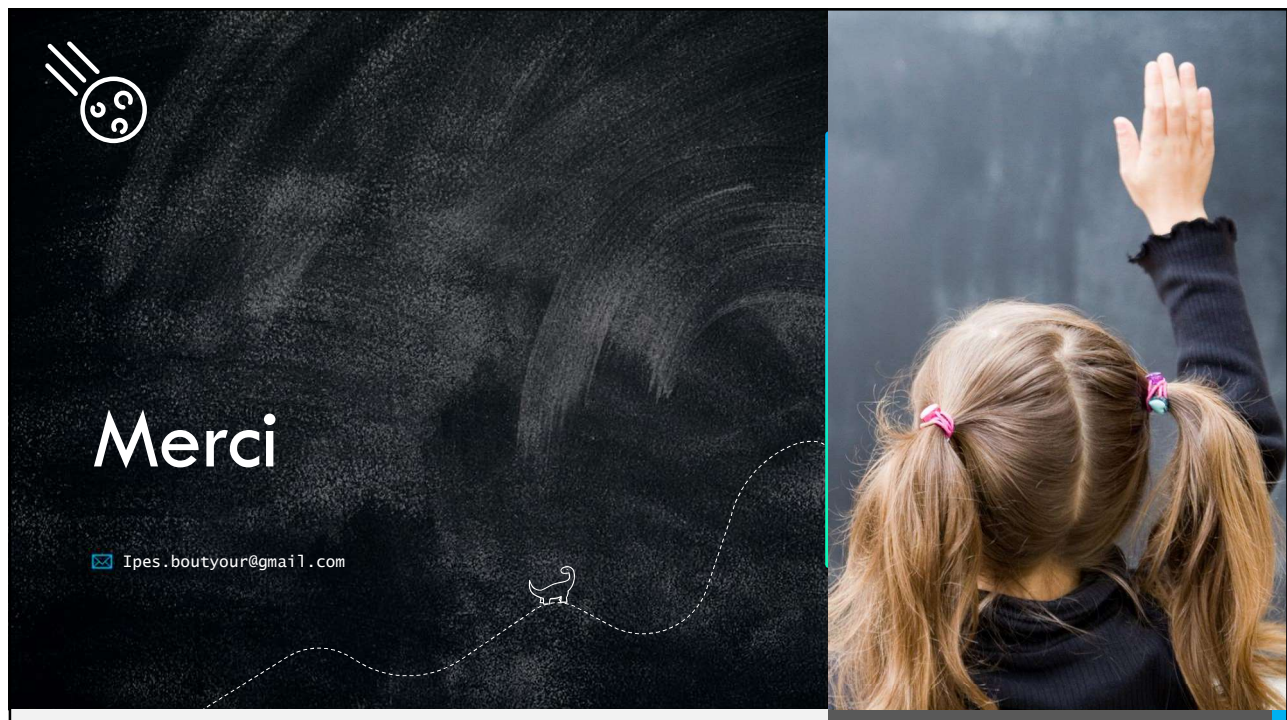
### Etapes :

1. Collecter les données
2. Décrire les données
3. Explorer les données
  - Valeurs manquante
  - Analyse univariée
  - **Analyse bivariable (pour la séance prochaine)**
4. Vérifier la qualité des données

ipes.boutyour@gmail.com

91

91



92




# Data Mining : Compréhension des données

Analyse bivariable

ipes.boutyour@gmail.com


93

93

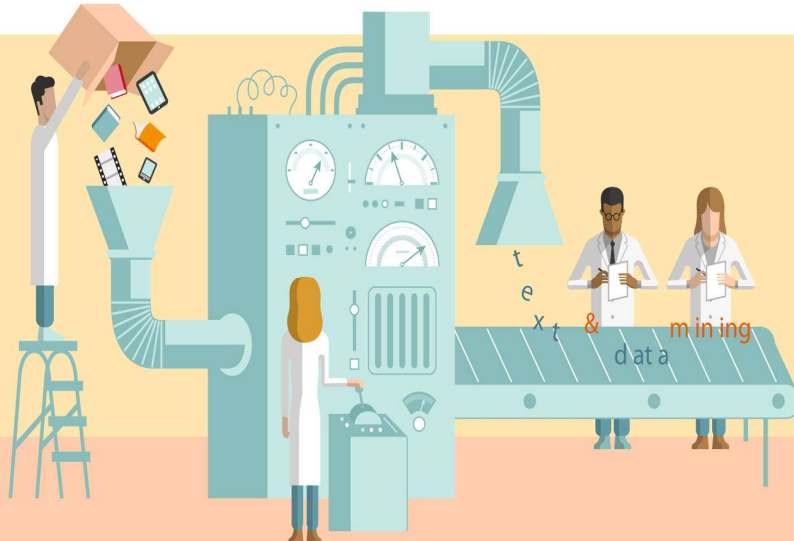


Université Mohammed V  
Faculté des Sciences  
Rabat

**Royaume du Maroc**  
Université Mohammed V de Rabat  
Faculté des Sciences  
Département d'Informatique



**IPSS**  
**Intelligent Processing  
Systems & Security**



## Data Mining & Machine Learning

Master IPS  
Faculté des sciences – Rabat  
Université Mohamed V

94