Apprentissage & Fouille de Données

Introduction

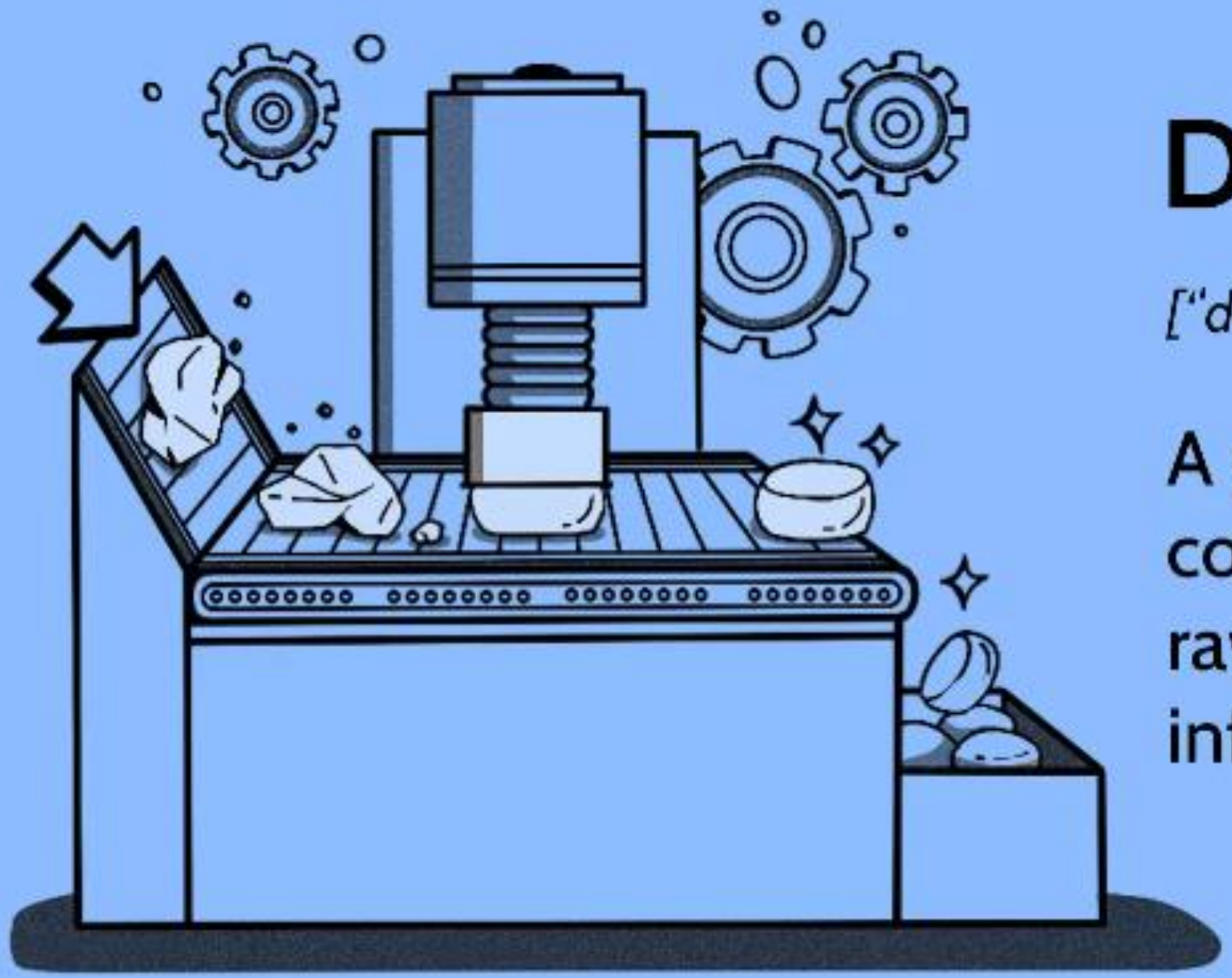
Mohamed Anis MEJRI

Année Universitaire: 2023-2024

Prérequis du cours

STATISTIQUES

PYTHON



Data Mining

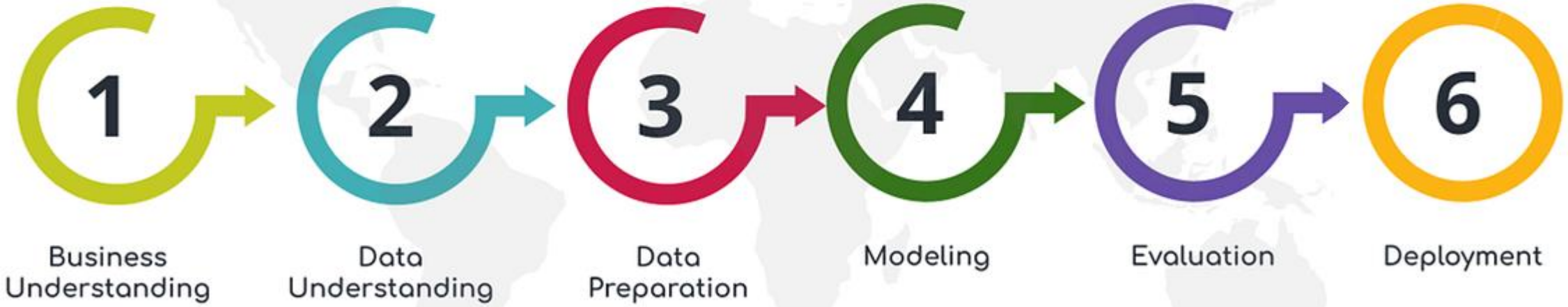
[ˈdā-tə ˈmī-niŋ]

A process used by companies to turn raw data into useful information.

Méthodologies de Data Mining

Mohamed Anis MEJRI

CRISP-DM: The Cross-Industry Standard Process for Data Mining



CRISP-DM 1: Business Understanding

- **Etablir les objectifs commerciaux :**

  Il est essentiel de « comprendre parfaitement, d'un point de vue commercial, ce que le client souhaite réellement accomplir »

- **Evaluer la situation :**

  Etudier la faisabilité, déterminer la disponibilité des ressources, les exigences du projet, évaluer les risques et les mesures d'urgence, et effectuer une analyse coûts-avantages.

- **Définir les objectifs de fouille de données :**

  En plus de définir les objectifs commerciaux, il est également nécessaire de préciser ce à quoi ressemble le succès d'un point de vue technique de la fouille de données.

- **Elaborer le plan de projet :**

  Sélectionner les technologies et les outils, et définir des plans détaillés pour chaque phase du projet.

CRISP-DM 2: Data Understanding

- **Collecte des données initiales :**

  Acquérir les données nécessaires et (si nécessaire) les charger dans votre outil d'analyse.

- **Description des données :**

  Examiner les données et documenter leurs propriétés telles que le format des données, le nombre d'enregistrements ou les identifiants des champs.

- **Exploration des données :**

  Approfondir l'analyse des données. Interroger, visualiser et identifier les relations entre les données.

- **Vérification de la qualité des données :**

  Documenter tout problème de qualité identifié.

CRISP-DM 3: Data Preparation

- **Sélection des données :**

  Déterminer quels ensembles de données seront utilisés et documenter les raisons de leur inclusion ou exclusion.

- **Nettoyage des données :**

  Souvent, il s'agit de la tâche la plus longue. Sans cela, vous risquez probablement de faire face au principe "garbage in, garbage out".

- **Construction des données :**

  Créer de nouveaux attributs qui seront utiles.

- **Intégration des données :**

  Créer de nouveaux ensembles de données en combinant des données de sources multiples.

- **Formatage des données :**

  Reformater les données au besoin. Par exemple, convertir des valeurs de chaînes de caractères qui stockent des nombres en valeurs numériques afin de pouvoir effectuer des opérations mathématiques.

CRISP-DM 4: Modeling

- **Sélection des techniques de modélisation :**

  Déterminer les algorithmes à utiliser.

- **Elaboration de la conception des tests :**

  En fonction de votre approche de modélisation, vous pourriez avoir besoin de diviser les données en ensembles d'entraînement, de test et de validation.

- **Construction des modèles :**

  Implémenter les algorithmes choisis pour créer des modèles.

- **Evaluation des modèles :**

  En général, plusieurs modèles sont en concurrence les uns avec les autres, et le data scientist doit interpréter les résultats du modèle en fonction de la connaissance du domaine, des critères de réussite prédéfinis et de la conception des tests.

CRISP-DM 5: Evaluation

- **Evaluation des résultats :**

  Les modèles répondent-ils aux critères de réussite commerciale ? Le(s)quel(s) devrions-nous approuver pour l'entreprise ?

- **Examen du processus :**

  Revoir le travail accompli. Toutes les étapes ont-elles été correctement exécutées ? Documenter les résultats et corriger si nécessaire.

- **Détermination des prochaines étapes :**

  En fonction des tâches précédentes, décider de passer au déploiement, de réitérer le projet, ou d'initier de nouveaux projets.

CRISP-DM 6: Deployment

- **Etablir le déploiement :**

  Elaborer et exécuter un plan pour déployer les modèles produits.

- **Planifier la surveillance et la maintenance :**

  Elaborer un plan complet de surveillance et de maintenance afin d'éviter des problèmes lors de la phase opérationnelle (ou post-projet) d'un modèle.

- **Produire le rapport final :**

  L'équipe du projet documente un résumé du projet, pouvant inclure une présentation finale des résultats de la fouille de données.

- **Révision du projet :**

  Effectuer une rétrospective du projet pour évaluer ce qui s'est bien déroulé, ce qui aurait pu être amélioré et comment l'améliorer à l'avenir.

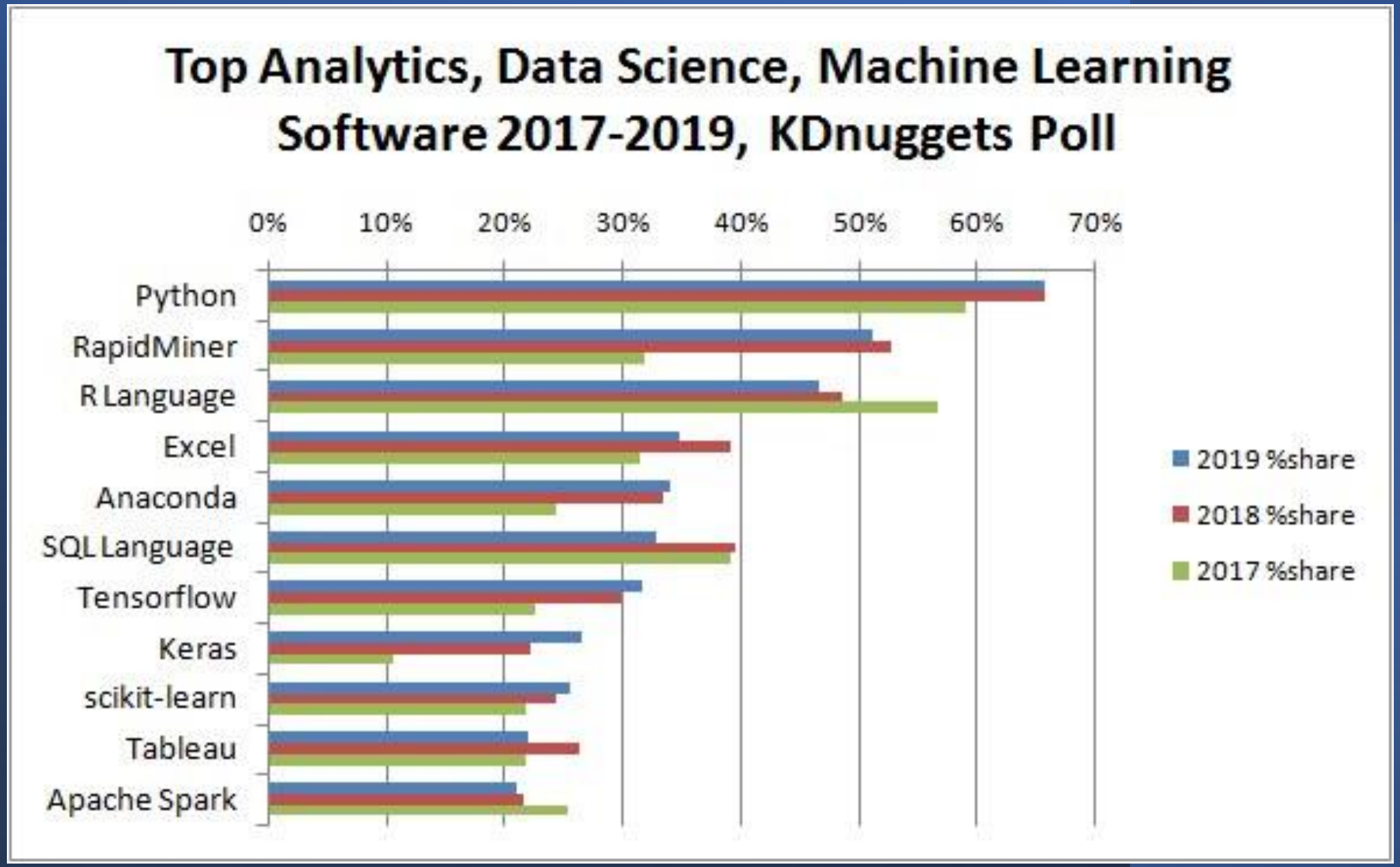
Langages et Outils de Data Mining

Mohamed Anis MEJRI

Principaux langages de fouille (mining) des données

Principaux outils de fouille (mining) des données

Principaux outils de fouille (mining) des données


Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll
0%
10%
20%
30%
40%
50%
60%
70%
Python
RapidMiner
R Language
Excel
Anaconda
SQL Language
Tensorflow
Keras
scikit-learn
Tableau
Apache Spark
2019 %share
2018 %share
2017 %share

CRISP-DM 2: Data Understanding

Exploration des données

Mohamed Anis MEJRI

Exploration des données

Les data scientists utilisent plusieurs techniques pour visualiser, analyser et déceler des motifs dans les données afin de mieux comprendre la nature des données. Cela peut aider à:

- Identifier des problèmes au niveau des données tels que des valeurs manquantes ou incorrectes, des fautes de frappe et des anomalies (valeurs aberrantes).
- Etudier la distribution des données.
- Analyser les éventuelles relations entre les variables.
- Identifier les variables qui pourraient ne pas influencer le résultat souhaité.

Principales étapes d'exploration des données

- Identification des attributs descriptifs

  Définissez chaque variable dans votre ensemble de données et son rôle.
- Analyse univariée
  - Pour les variables continues, créez un boxplot ou un histogramme pour chaque variable séparément.
  - Pour les variables catégorielles, créez un diagramme en barres pour montrer leurs fréquences.
- Analyse bivariée

  Créez des outils de visualisation pour déterminer les interactions entre les variables.
  - Continu et continu : Nuage de points (Scatterplot)
  - Catégoriel et catégoriel : Diagramme en colonnes empilées (Stacked Column Chart)
  - Catégoriel et continu : boxplot combiné avec un graphique de groupe
- Identification des valeurs manquantes
- Détection des valeurs aberrantes

Exploration des données

Identification des attributs descriptifs

- Les attributs sont communément appelés aussi variables, dimensions ou caractéristiques.
- Les attributs peuvent avoir plusieurs types:
  - Numériques:
    - discrets
    - continus
  - Catégorielles:
    - nominaux (pas d'ordre spécifique)
    - ordinaux (avec un ordre spécifique)
    - binaires
  - Autres:
    - Textuels
    - Temporels
    - Géographiques

Identification des attributs descriptifs

*F*onctions utiles

- df.shape
- print(df.columns.to_list())
- print(df.info())
- print(df.dtypes)
- print(df.describe())
- print(df.describe(include ='all’))
- print(df.describe(include = 'object’))
- print(df['categorical_column'].value_counts())

Exploration des données
Analyse univariée

- Statistiques basiques:
  - Nombre d'objets
  - Nombre d'attributs

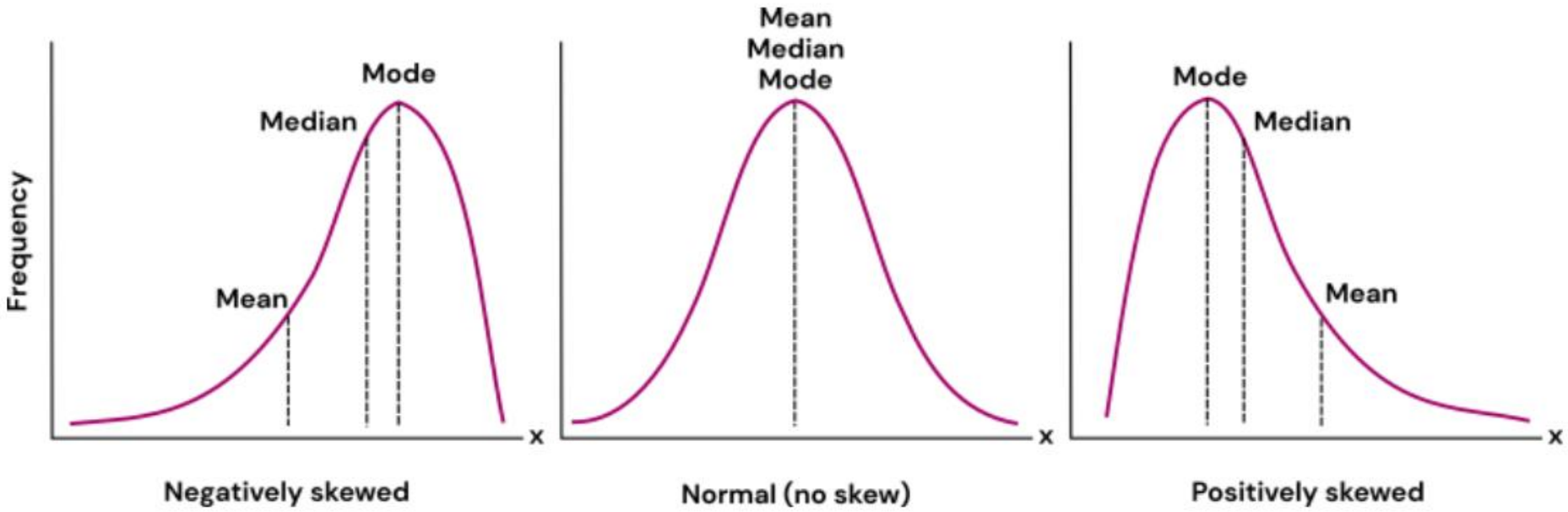
- Distributions des valeurs des attributs:
  - Numériques: Tendance centrale et dispersion.
  - Catégorielles: Pourcentage de chaque valeur.

Analyse univariée
Tendance centrale des attributs numériques

- **Moyenne (Mean) :**
  - **Définition :** La moyenne, ou moyenne arithmétique, est calculée en ajoutant toutes les valeurs d'un ensemble de données et en divisant le total par le nombre d'observations.
  - **Formule :** Somme des valeurs / Nombre d'observations
  - **Avantages :** Sensible à toutes les valeurs dans l'ensemble de données.
- **Médiane (Median) :**
  - **Définition :** La médiane est la valeur centrale d'un ensemble de données trié par ordre croissant ou décroissant. Si le nombre d'observations est pair, la médiane est la moyenne des deux valeurs du milieu.
  - **Formule :** Aucune formule spécifique, elle dépend de la position des valeurs dans l'ensemble de données.
  - **Avantages :** Moins sensible aux valeurs extrêmes que la moyenne.

Analyse univariée
Tendance centrale des attributs numériques

- **Mode :**
  - **Définition :** Le mode est la valeur qui apparaît le plus fréquemment dans un ensemble de données.
  - **Formule :** Aucune formule spécifique, c'est simplement la valeur la plus fréquemment observée.
  - **Avantages :** Utile pour identifier les valeurs les plus fréquentes.
- **Midrange :**
  - **Définition :** Le midrange est la moyenne arithmétique de la valeur maximale et de la valeur minimale d'un ensemble de données.
  - **Formule :** (Valeur maximale+Valeur minimale) / 2
  - **Avantages :** Facile à calculer, mais sensible aux valeurs extrêmes.


Frequency
Mode
Median
Mean
x
Negatively skewed
Mean
Median
Mode
x
Normal (no skew)
Mode
Median
Mean
x
Positively skewed

Analyse univariée
Dispersion des attributs numériques

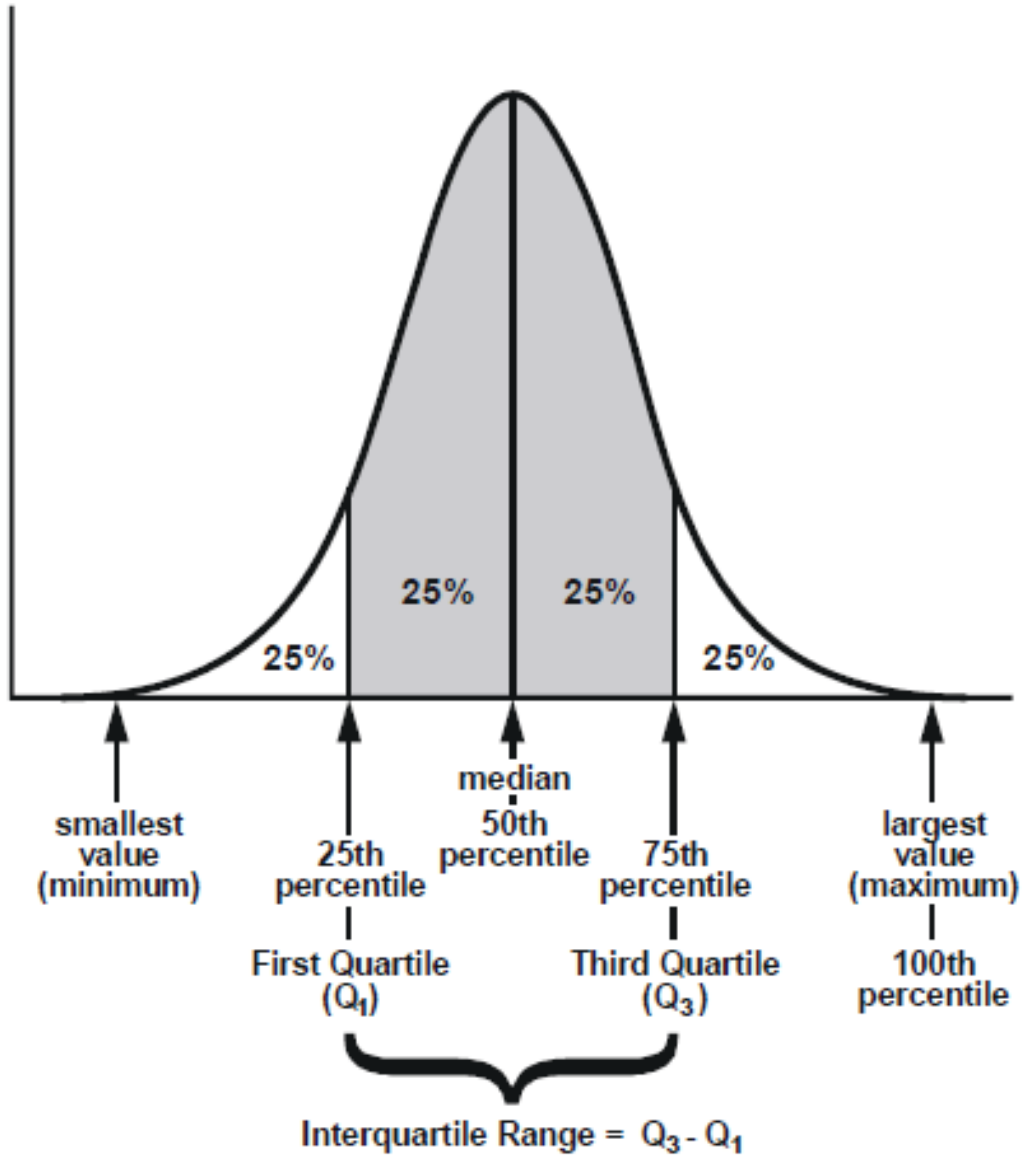
- **Etendue (Range) :**
  - **Définition :** L'étendue est la différence entre la valeur maximale et la valeur minimale dans un ensemble de données.
  - **Utilité :** Elle donne une mesure simple de la dispersion des valeurs dans l'ensemble de données.
- **Quartiles :**
  - **Définition :** Les quartiles divisent un ensemble de données en quatre parties égales. Les trois quartiles sont Q1 (25eme percentile), Q2 (50eme percentile ou médiane), et Q3 (75eme percentile).
  - **Utilité :** Les quartiles fournissent des informations sur la tendance centrale et la dispersion de l'ensemble de données.

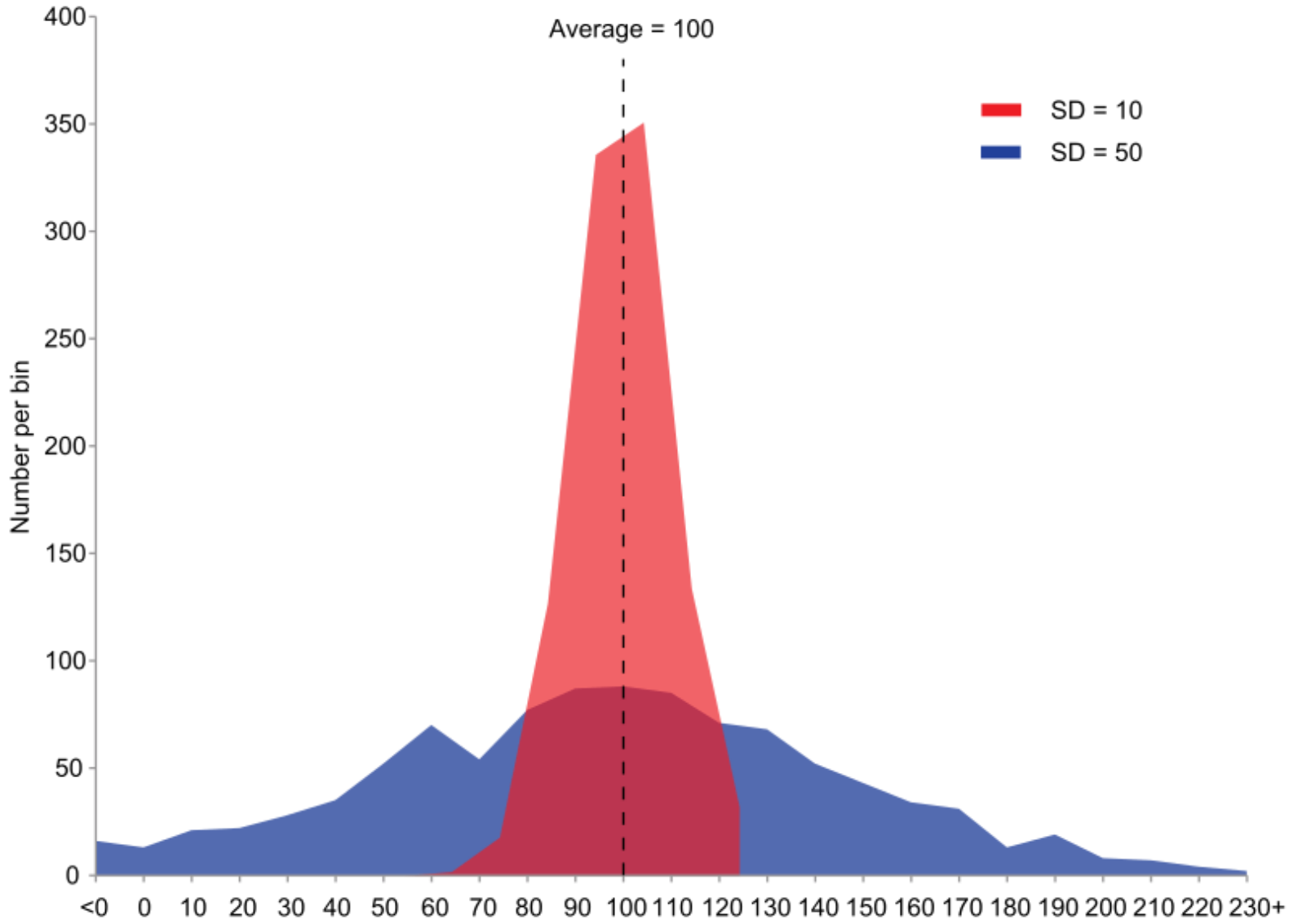
Analyse univariée
Dispersion des attributs numériques

- **Ecart interquartile (IQR) :**
  - **Définition :** L'IQR est la plage de valeurs entre le premier quartile (Q1) et le troisième quartile (Q3).
  - **Formule :** $Q3 - Q1$
  - **Utilité :** C'est une mesure robuste de la dispersion des 50% des données centrales. Il est moins sensible aux valeurs extrêmes que l'étendue.
- **Variance :**
  - **Définition :** La variance mesure à quelle distance chaque point de données est de la moyenne. C'est la moyenne des carrés des différences par rapport à la moyenne.
  - **Utilité :** La variance fournit une mesure de la variabilité globale de l'ensemble de données.

Analyse univariée
Dispersion des attributs numériques

- **Ecart-type (Standard Deviation) :**
  - **Définition :** L'écart-type est la racine carrée de la variance. Il représente la distance moyenne de chaque point de données à la moyenne.
  - **Utilité :** L'écart-type est une mesure largement utilisée de la dispersion des données. Il est exprimé dans la même unité que les données originales, ce qui facilite son interprétation.


25%
25%
25%
25%
smallest value (minimum)
25th percentile
median 50th percentile
75th percentile
largest value (maximum)
First Quartile (Q_1)
Third Quartile (Q_3)
100th percentile
Interquartile Range = $Q_3 - Q_1$


Average = 100
SD = 10
SD = 50
Number per bin
400
350
300
250
200
150
100
50
0
<0 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230+

Analyse univariée
Outils de visualisation

- Box plot

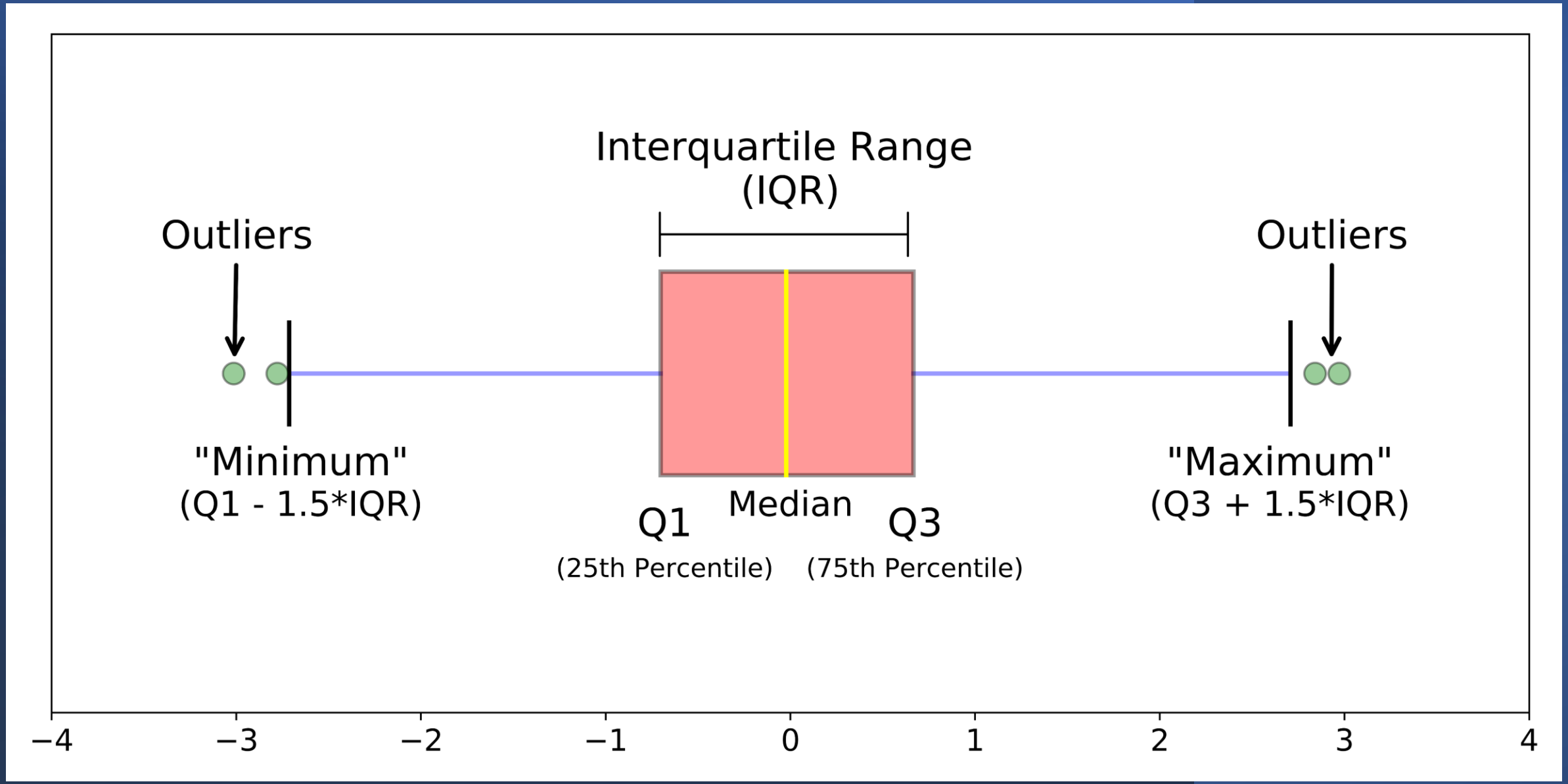
  Un "box plot" est un type de graphique utilisé en statistiques pour représenter graphiquement la distribution d'un ensemble de données et montrer la médiane, les quartiles et les valeurs extrêmes.

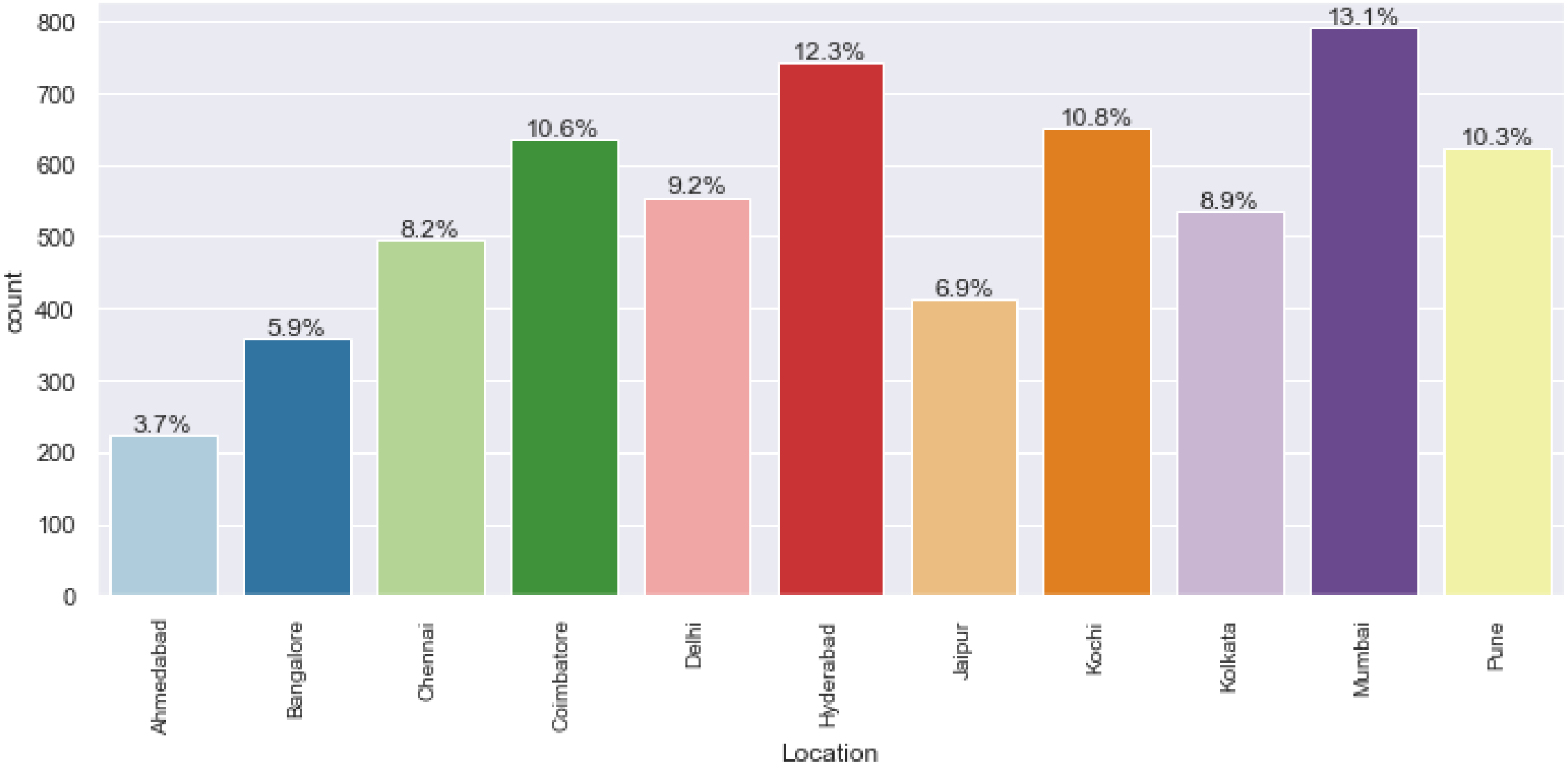
- Bar plot

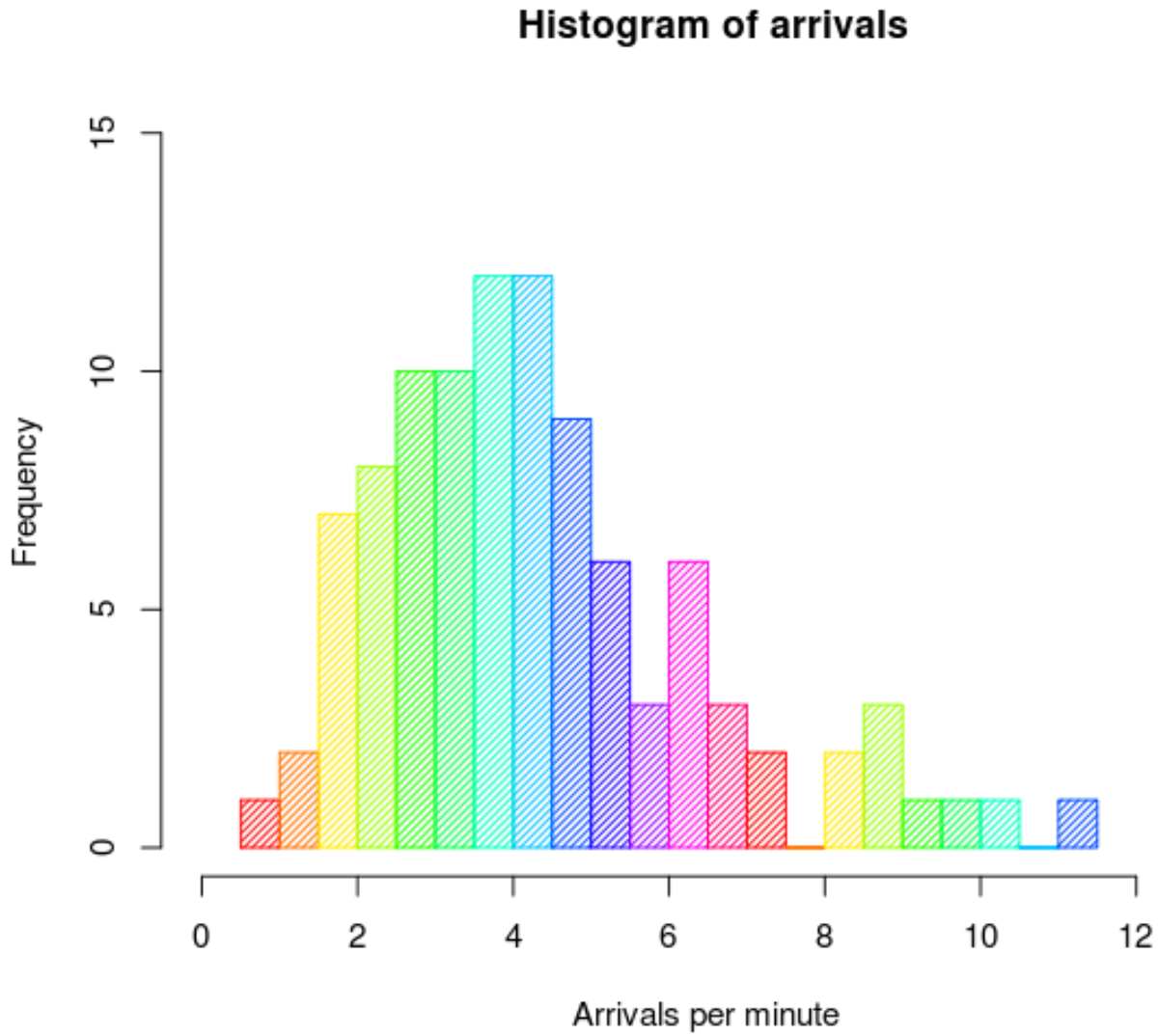
  Un bar plot est un type de graphique utilisé pour représenter visuellement la distribution de données catégorielles. Il affiche les valeurs de différentes catégories sous forme de barres rectangulaires, dont la hauteur est proportionnelle à la fréquence de chaque catégorie.

- Histogramme

  Un histogramme est un type de graphique utilisé en statistiques pour représenter la distribution des données en regroupant les valeurs en intervalles et en affichant la fréquence (ou la densité) de chaque intervalle par des barres verticales.

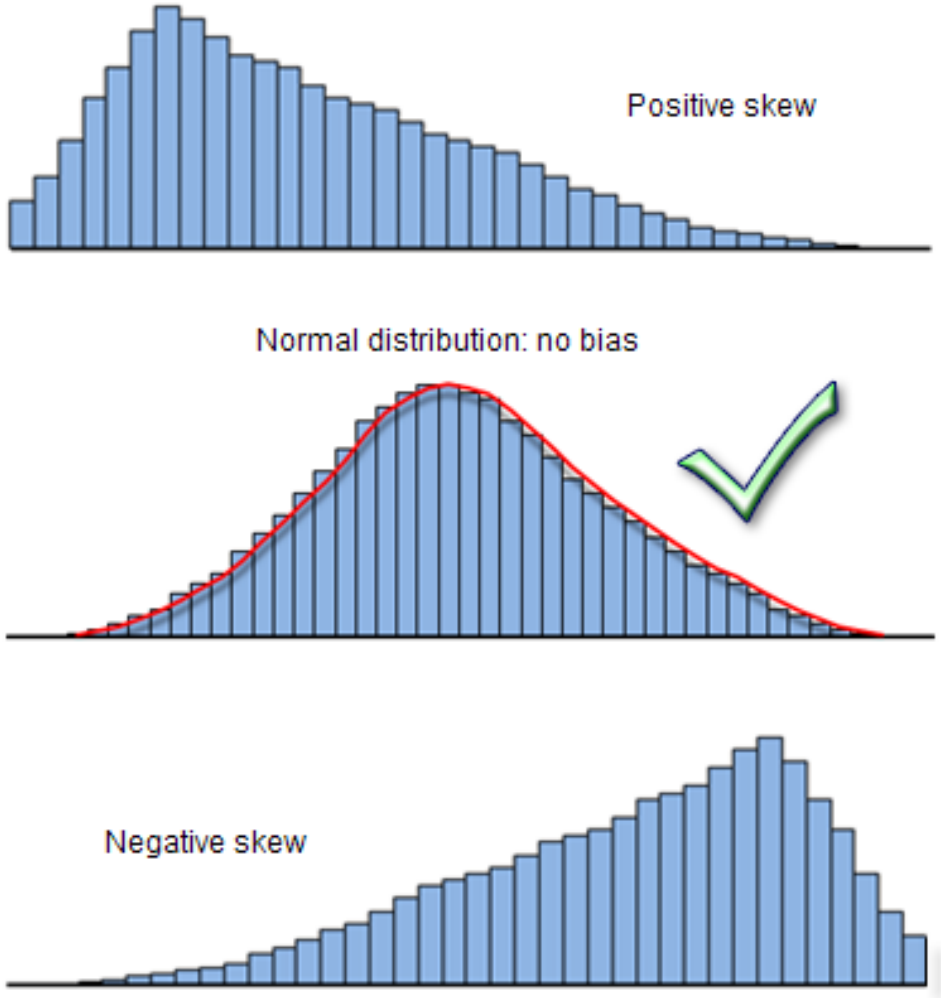

Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
"Maximum"
(Q3 + 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
−4
−3
−2
−1
0
1
2
3
4


800
700
600
500
400
300
200
100
0
count
3.7%
5.9%
8.2%
10.6%
9.2%
12.3%
6.9%
10.8%
8.9%
13.1%
10.3%
Ahmedabad
Bangalore
Chennai
Coimbatore
Delhi
Hyderabad
Jaipur
Kochi
Kolkata
Mumbai
Pune
Location


Histogram of arrivals
Frequency
0
5
10
15
0
2
4
6
8
10
12
Arrivals per minute

Analyse univariée
Identification des biais

- Si vos données proviennent d'une distribution symétrique, telle que la distribution normale, les données seront réparties uniformément autour du centre des données.
- Si les données ne sont pas réparties de manière approximativement égale autour du centre de l'histogramme, on parle généralement de "biais" (ou "skewness" en anglais).


Positive skew
Normal distribution: no bias
Negative skew

Analyse univariée
Fonctions utiles

- sns.countplot(x='categorical_column', data=df)
- sns.boxplot(data = df[column or list of columns], orient = 'h’)
- sns.histplot(df['numerical_column'], kde=True)
- sns.boxplot(x='categorical_column', y='numerical_column', data=df)
- plt.bar(df['Category'], df['Count'], color='skyblue’)

df['Count'] représente le nombre correspondant pour chaque catégorie. Par exemple : plt.bar(df[‘genre'], df[‘age'], color='skyblue') pour compter le "genre" par « age".

Exploration des données
Analyse bivariée

- L'analyse bivariée est une méthode d'analyse statistique qui examine la relation entre deux variables.
- Contrairement à l'analyse univariée, qui se concentre sur une seule variable à la fois, l'analyse bivariée explore les interactions, les corrélations ou les différences entre deux variables simultanément.

Analyse bivariée
Outils de visualisation

- **Matrice de Corrélation**

  Affiche les corrélations entre plusieurs paires de variables.

  - Corrélation positive - Des valeurs proches de 1 indiquent une corrélation positive forte, signifiant que lorsque une variable augmente, l'autre variable a tendance à augmenter également.
  - Corrélation négative - Des valeurs proches de -1 indiquent une corrélation négative forte, signifiant que lorsque une variable augmente, l'autre variable a tendance à diminuer.
  - Pas de corrélation (0) - Un coefficient de corrélation de 0 suggère qu'il n'y a pas de relation linéaire entre les variables.

- **Scatter plot** (diagramme de dispersion)

  Un "nuage de points" ou "diagramme de dispersion" est un type de graphique utilisé pour représenter graphiquement la relation entre deux variables continues.

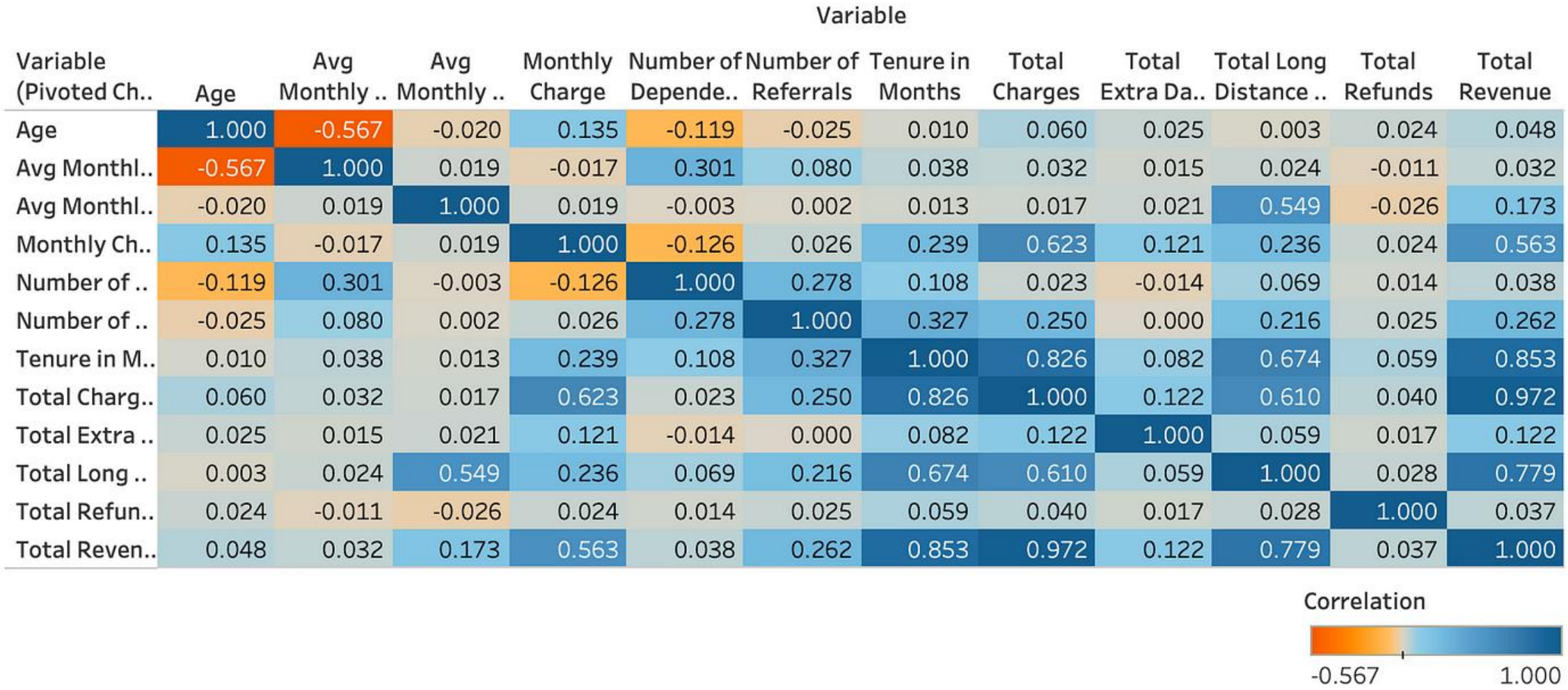
Analyse bivariée
Outils de visualisation

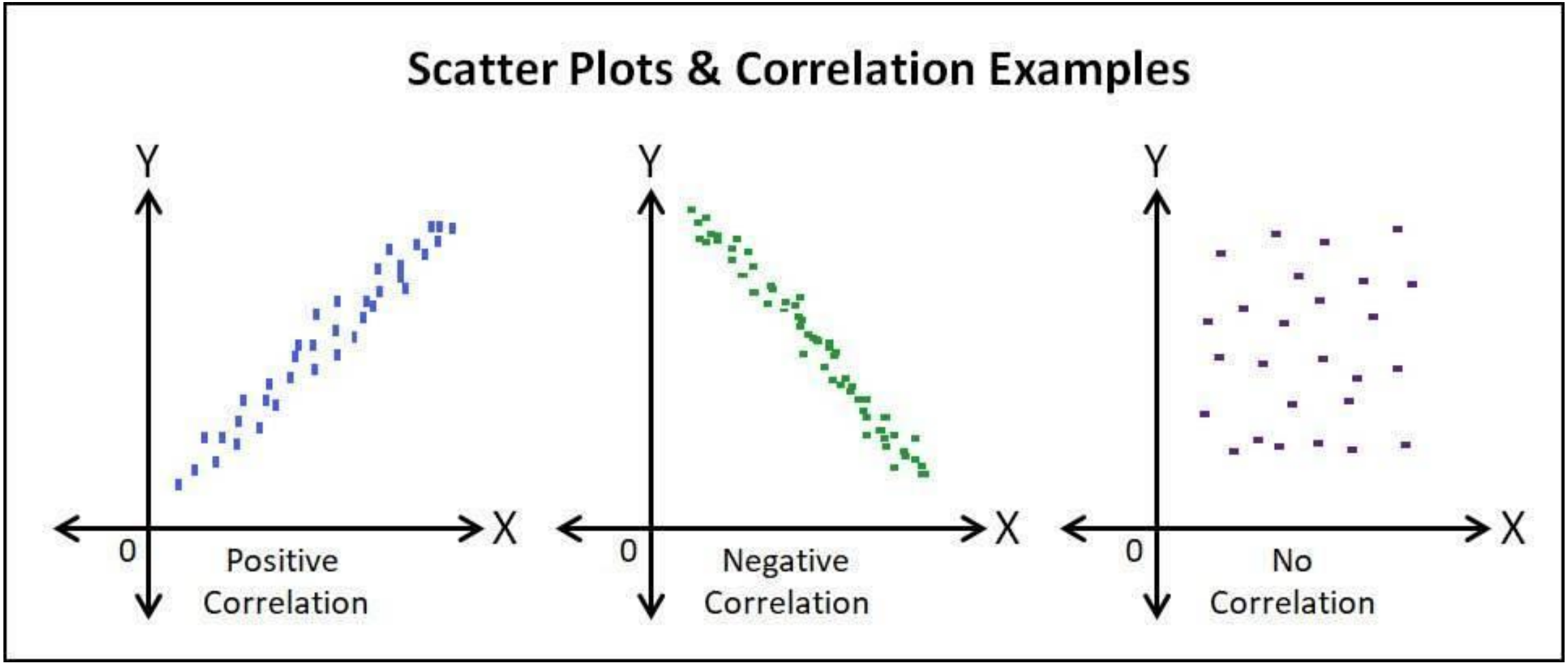
- **Pair plot** (graphique de paires)

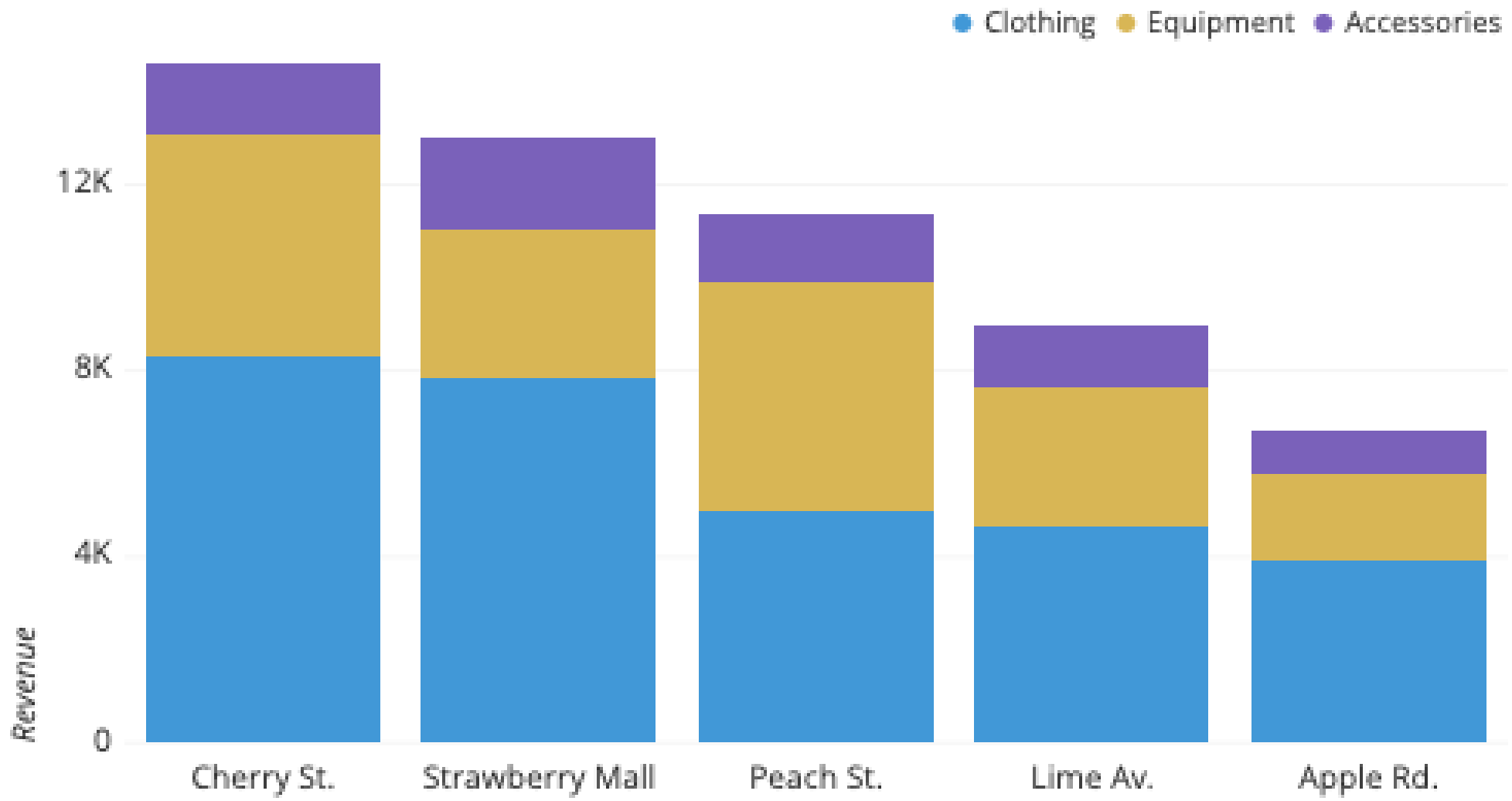
  Un pair plot est une visualisation statistique qui permet d'examiner les relations entre les différentes paires de variables dans un ensemble de données multivariées. Il affiche des graphiques bidimensionnels pour chaque paire de variables, ce qui facilite l'observation des tendances, des corrélations, et des distributions croisées.

- **Un stacked column chart (diagramme en colonnes empilée)**

  Un graphique qui représente plusieurs séries de données sur une même colonne, où chaque colonne est divisée en segments empilés qui représentent les différentes composantes de chaque série.

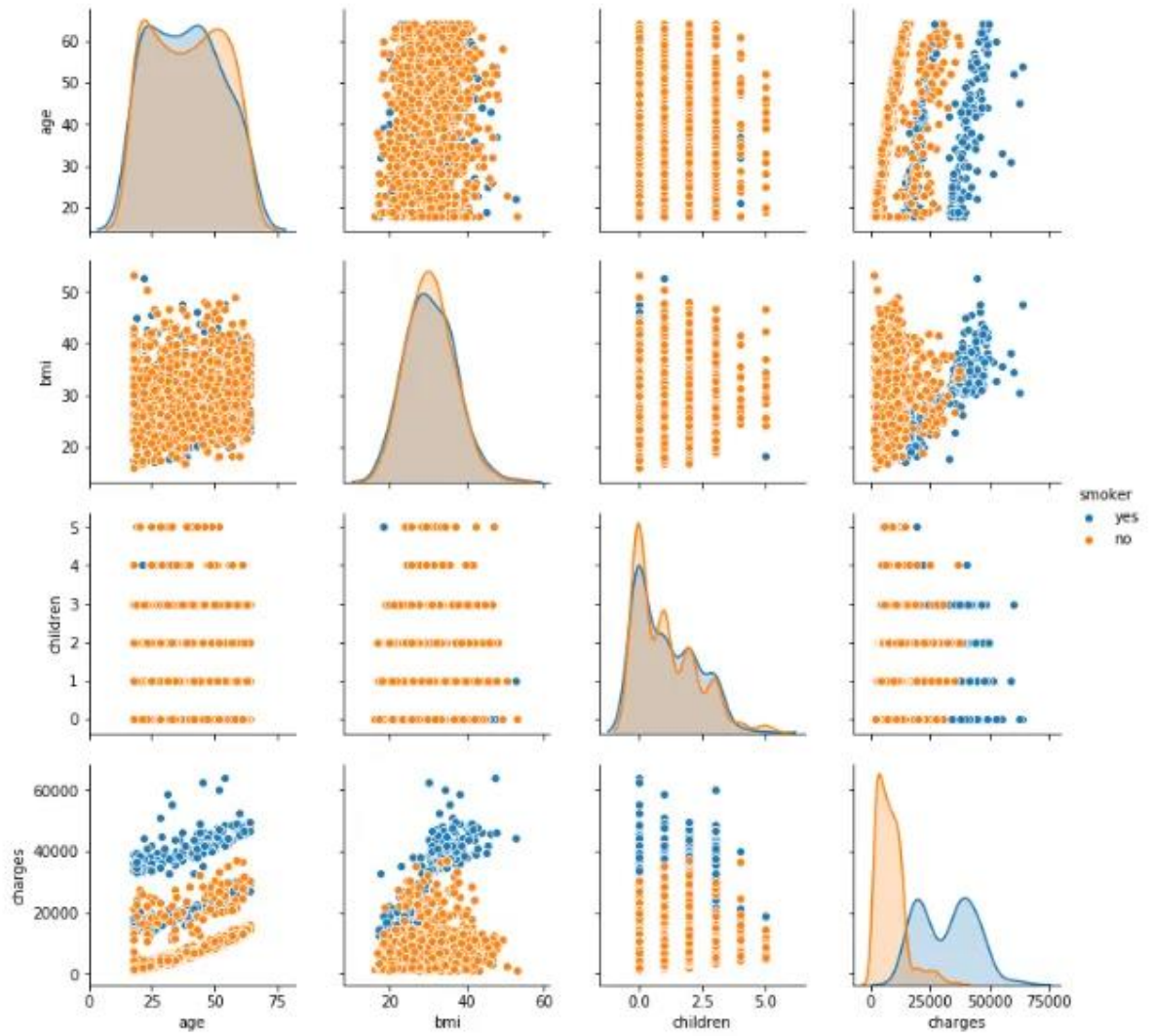

Variable
Variable (Pivoted Ch..
Age
Avg Monthly ..
Avg Monthly ..
Monthly Charge
Number of Depende..
Number of Referrals
Tenure in Months
Total Charges
Total Extra Da..
Total Long Distance ..
Total Refunds
Total Revenue
Age 1.000 -0.567 -0.020 0.135 -0.119 -0.025 0.010 0.060 0.025 0.003 0.024 0.048
Avg Monthl.. -0.567 1.000 0.019 -0.017 0.301 0.080 0.038 0.032 0.015 0.024 -0.011 0.032
Avg Monthl.. -0.020 0.019 1.000 0.019 -0.003 0.002 0.013 0.017 0.021 0.549 -0.026 0.173
Monthly Ch.. 0.135 -0.017 0.019 1.000 -0.126 0.026 0.239 0.623 0.121 0.236 0.024 0.563
Number of .. -0.119 0.301 -0.003 -0.126 1.000 0.278 0.108 0.023 -0.014 0.069 0.014 0.038
Number of .. -0.025 0.080 0.002 0.026 0.278 1.000 0.327 0.250 0.000 0.216 0.025 0.262
Tenure in M.. 0.010 0.038 0.013 0.239 0.108 0.327 1.000 0.826 0.082 0.674 0.059 0.853
Total Charg.. 0.060 0.032 0.017 0.623 0.023 0.250 0.826 1.000 0.122 0.610 0.040 0.972
Total Extra .. 0.025 0.015 0.021 0.121 -0.014 0.000 0.082 0.122 1.000 0.059 0.017 0.122
Total Long .. 0.003 0.024 0.549 0.236 0.069 0.216 0.674 0.610 0.059 1.000 0.028 0.779
Total Refun.. 0.024 -0.011 -0.026 0.024 0.014 0.025 0.059 0.040 0.017 0.028 1.000 0.037
Total Reven.. 0.048 0.032 0.173 0.563 0.038 0.262 0.853 0.972 0.122 0.779 0.037 1.000
Correlation
-0.567
1.000


Scatter Plots & Correlation Examples
Y
X
0
Positive
Correlation
Y
X
0
Negative
Correlation
Y
X
0
No
Correlation


Clothing
Equipment
Accessories
12K
8K
4K
0
Revenue
Cherry St.
Strawberry Mall
Peach St.
Lime Av.
Apple Rd.

```
sns.pairplot(df, hue="smoker")
```

```
<seaborn.axisgrid.PairGrid at 0x1afb4b02128>
```



Analyse bivariée
Fonctions utiles

- correlation_matrix = df.corr()
- correlation_coefficient = df['variable1'].corr(df['variable2'])
- sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm’)
- sns.scatterplot(x='numerical_column1', y='numerical_column2', data=df)
- sns.pairplot(df[['numerical_column1', 'numerical_column2', 'numerical_column3’]]) # ou bien toutes les données.

Exploration des données
Identification des valeurs manquantes (fonctions utiles)

- df.isnull().sum().sum()
- df.isnull().sum()
- sns.heatmap(df.isnull(), cbar=False, cmap='viridis’)

Exploration des données
Détection des valeurs aberrantes

Plusieurs méthodes permettent de détecter les valeurs aberrantes:

- **IQR (Interquartile Range)**
- **Tukey's Fences**
- **Local Outlier Factor (LOF)**
- **Isolation Forest**
- **DBSCAN**

CRISP-DM 3: Data Preparation

CRISP-DM 3: Data Preparation

Nettoyage des données

Problèmes courants de qualité des données

- **Données manquantes, incomplètes.**
- **Données bruitées, erronées, aberrantes.**
- **Données redondantes.**
- **Distribution biaisée des données.**
- **Données inconsistantes.**

Solutions de qualité des données
Données manquantes

- Supprimer les lignes contenant des données manquantes.
- Supprimer les colonnes contenant des données manquantes.
- Remplir manuellement les données manquantes.
- Remplir automatiquement les attributs incomplets:
  - Moyenne globale, moyenne par classe.
  - Estimer la valeur manquante par régression, KNN…

Solutions de qualité des données
Données manquantes (fonctions utiles)

- df.dropna(inplace=True) # Supprimer les lignes avec des valeurs manquantes
- df.dropna(axis=1, inplace=True) # Remplacer les valeurs manquantes par la moyenne de la colonne.
- df['column_name'].fillna(df['column_name'].mode()[0], inplace=True) # Remplacer les valeurs manquantes par le mode (valeur la plus fréquente) de la colonne.
- df.fillna(method='ffill', inplace=True) # Remplacer les valeurs manquantes par la valeur non manquante précédente.
- df.fillna(method='bfill', inplace=True) # Remplacer les valeurs manquantes par la valeur non manquante suivante.
- df['column_name'].interpolate(inplace=True) # Interpoler les valeurs manquantes en fonction des valeurs des points de données voisins.
- df['column_name'].fillna(value, inplace=True) # Remplacer les valeurs manquantes par une valeur spécifique de votre choix.

Solutions de qualité des données
Données redondantes

- Identifier les données redondantes à travers une études des corrélations entre les attributs descriptifs.
  - Données numériques:
    - Coefficient de corrélation
    - Scatter plot
  - Données catégorielles:
    - chi-deux (chi-square test)

Solutions de qualité des données
Données aberrantes

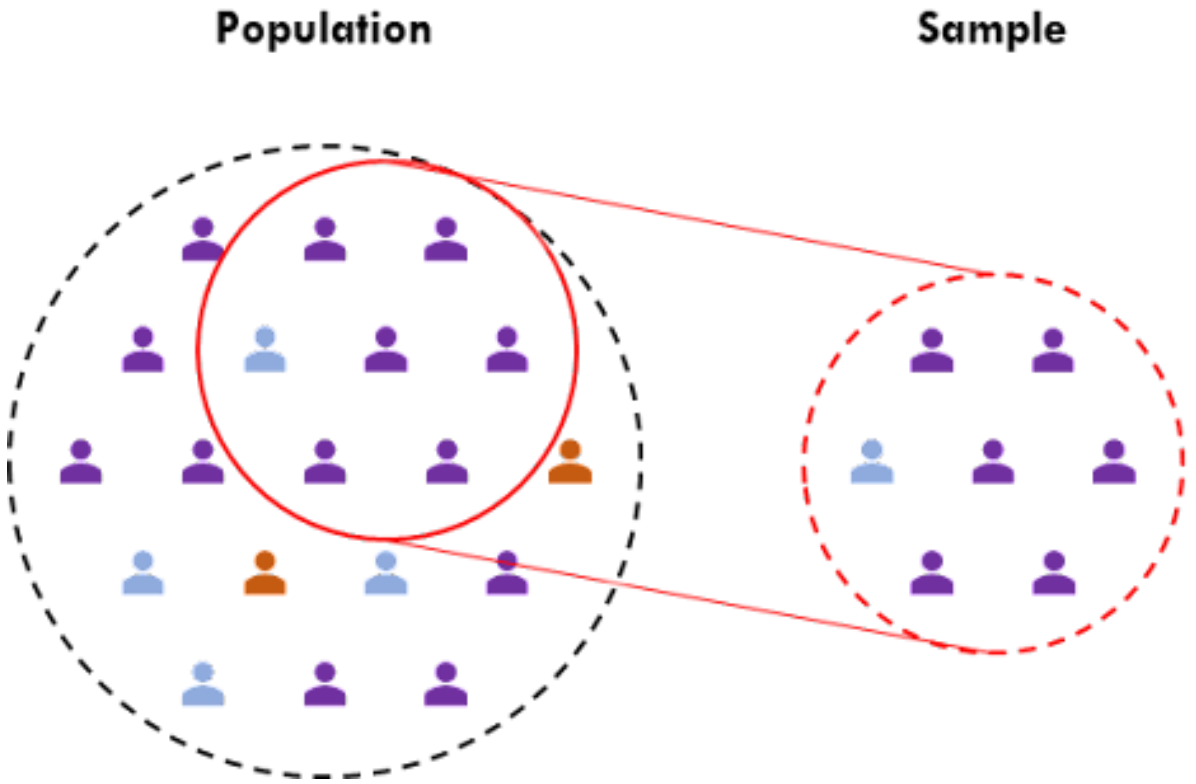
- Détecter les aberrations:
  - Régression
  - Clustering
- Supprimer les lignes contenant des données aberrantes.
- Corriger les erreurs détectées.

Solutions de qualité des données
Données aberrantes (fonctions utiles)

- df_no_outliers = df[~outliers] # Supprimer les lignes contenant des valeurs aberrantes.
- df['column_name'] = np.log1p(df['column_name']) # Appliquer des transformations mathématiques pour réduire l'impact des valeurs extrêmes.
- df['column_name'] = df['column_name'].clip(lower=lower_threshold, upper=upper_threshold) # Définir un seuil pour les valeurs extrêmes.
- df.loc[outliers, 'column_name'] = df['column_name'].median() # Remplacer les valeurs aberrantes par des valeurs imputées (par exemple, moyenne, médiane).
- df['column_name'] = pd.cut(df['column_name'], bins=bin_edges, labels=bin_labels) # Convertir la variable en catégories pour gérer les valeurs extrêmes.

Solutions de qualité des données

Traiter les distributions biaisées



Solutions de qualité des données
Traiter les distributions biaisées

- **Utilisation des méthodes de rééchantillonnage:**
  - **Sous-échantillonnage** - Retirer de manière aléatoire des instances de la classe majoritaire pour équilibrer la distribution des classes. Veillez à ne pas perdre d'informations précieuses.
  - **Sur-échantillonnage** - Répliquer de manière aléatoire des instances de la classe minoritaire pour équilibrer la distribution des classes. Veillez à ne pas surajuster le modèle.
  - **SMOTE (Synthetic Minority Over-sampling Technique)** - Génère des échantillons synthétiques pour la classe minoritaire en interpolant entre les instances existantes.

Solutions de qualité des données
Traiter les distributions biaisées

- **Utilisation d'algorithmes spécialisés**

  Certains algorithmes sont conçus pour traiter de manière plus efficace des ensembles de données déséquilibrés. Par exemple, des algorithmes tels que XGBoost et LightGBM disposent de paramètres qui peuvent être ajustés pour traiter le déséquilibre de classes.

- **Utilisation de méthodes d'ensemble**

  Utilisez des méthodes d'ensemble telles que le bagging et le boosting. Des algorithmes tels que Random Forest et AdaBoost peuvent être efficaces pour traiter des données déséquilibrées.

Solutions de qualité des données

Traiter les distributions biaisées

- **Poids de classe personnalisés**

  Ajuster les poids de classe pendant l'entraînement du modèle pour pénaliser plus lourdement les erreurs de classification de la classe minoritaire.

- **Métriques d'évaluation**

  Plutôt que d'utiliser l'accuracy, considérez des métriques telles que la précision, le rappel, le score F1, ou l'aire sous la courbe ROC (AUC-ROC) qui fournissent une meilleure compréhension de la performance du modèle sur des ensembles de données déséquilibrés.

- **Collecte de plus de données**

  Si possible, collectez plus de données pour la classe minoritaire afin de fournir au modèle plus d'informations pour l'entraînement.

CRISP-DM 3: Data Preparation

Réduction de dimensionnalité des données

Pourquoi réduire la dimensionnalité

- **Simplicité et Interprétabilité**

  Facilite la compréhension des modèles par des représentations plus simples.

- **Eviter la Malédiction de la Dimensionnalité**

  Réduit les problèmes liés à l'augmentation du nombre de variables.

- **Amélioration de la Performance**

  Diminue le risque de surajustement (overfitting) en éliminant le bruit et en mettant en évidence les tendances générales.

- **Gain de Temps de Calcul**

  Accélère le processus d'entraînement des modèles en traitant moins de caractéristiques.

Pourquoi réduire la dimensionnalité

- **Simplicité et Interprétabilité**

  Facilite la compréhension des modèles par des représentations plus simples.

- **Eviter la Malédiction de la Dimensionnalité**

  Réduit les problèmes liés à l'augmentation du nombre de variables.

- **Amélioration de la Performance**

  Diminue le risque de surajustement (overfitting) en éliminant le bruit et en mettant en évidence les tendances générales.

- **Gain de Temps de Calcul**

  Accélère le processus d'entraînement des modèles en traitant moins de caractéristiques.

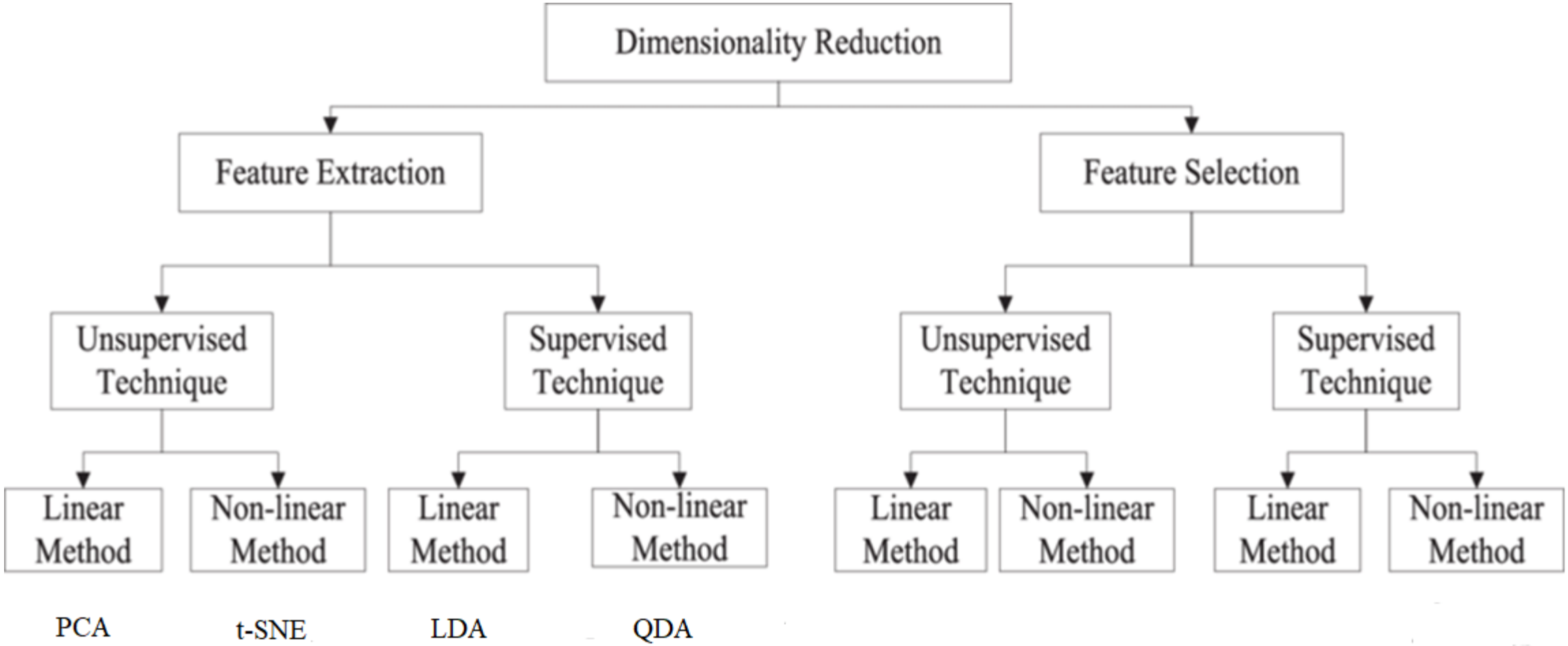
Comment réduire la dimensionnalité

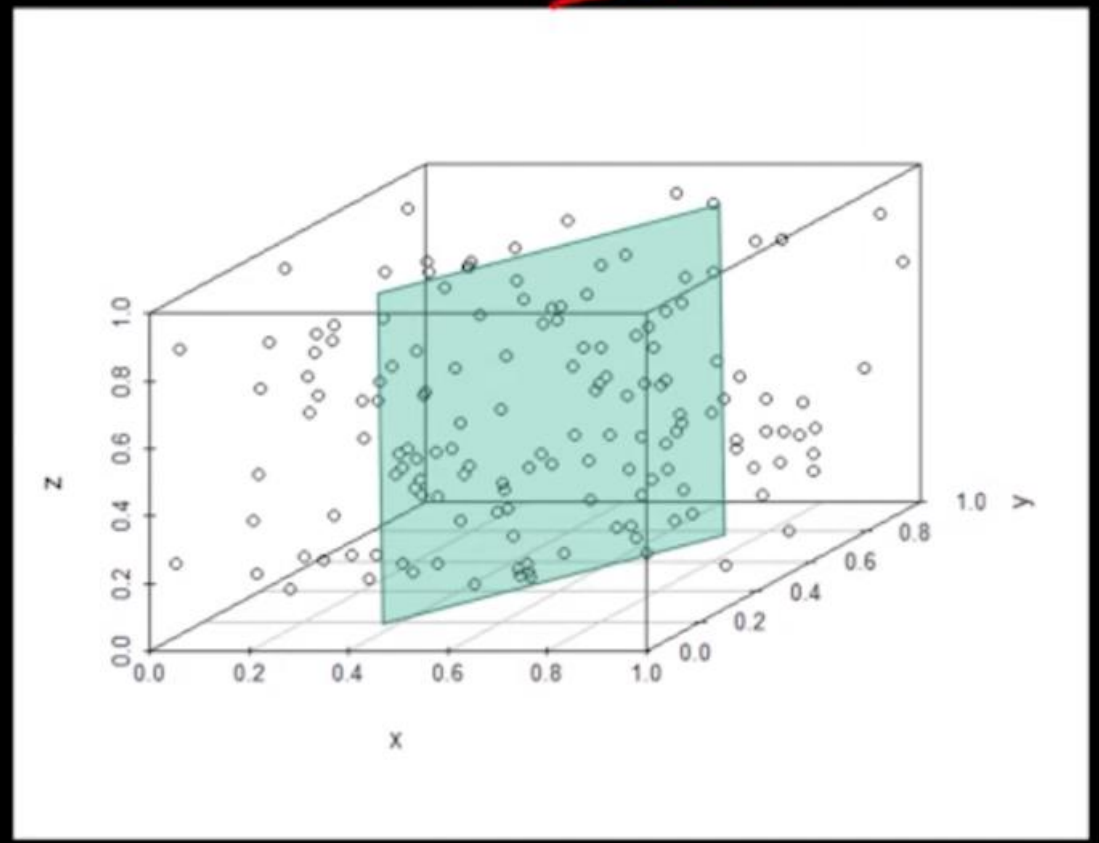
- **Feature Selection**

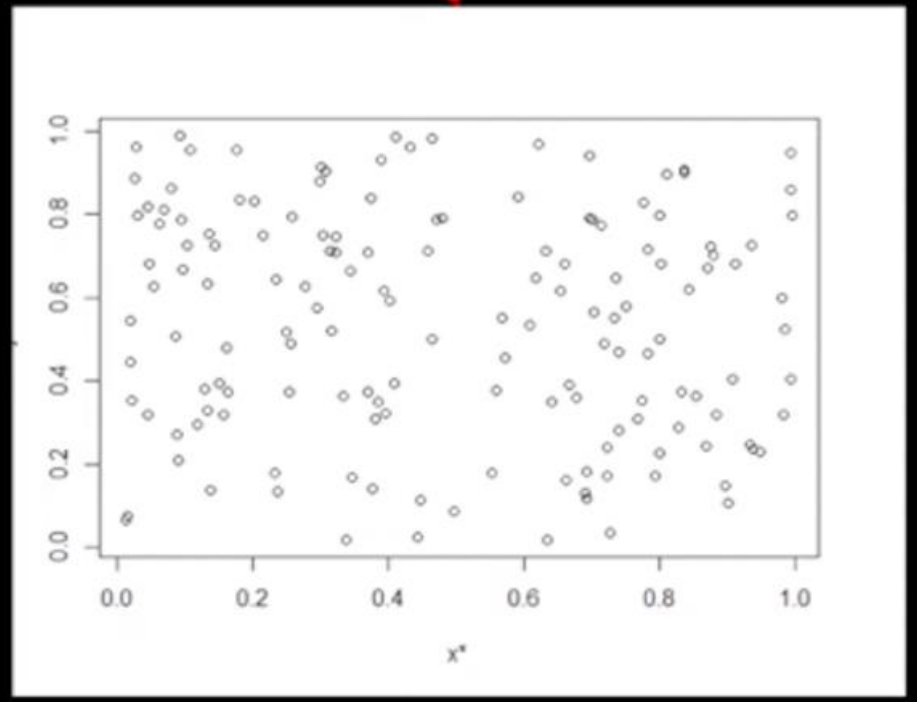
  Choix des variables les plus informatives.

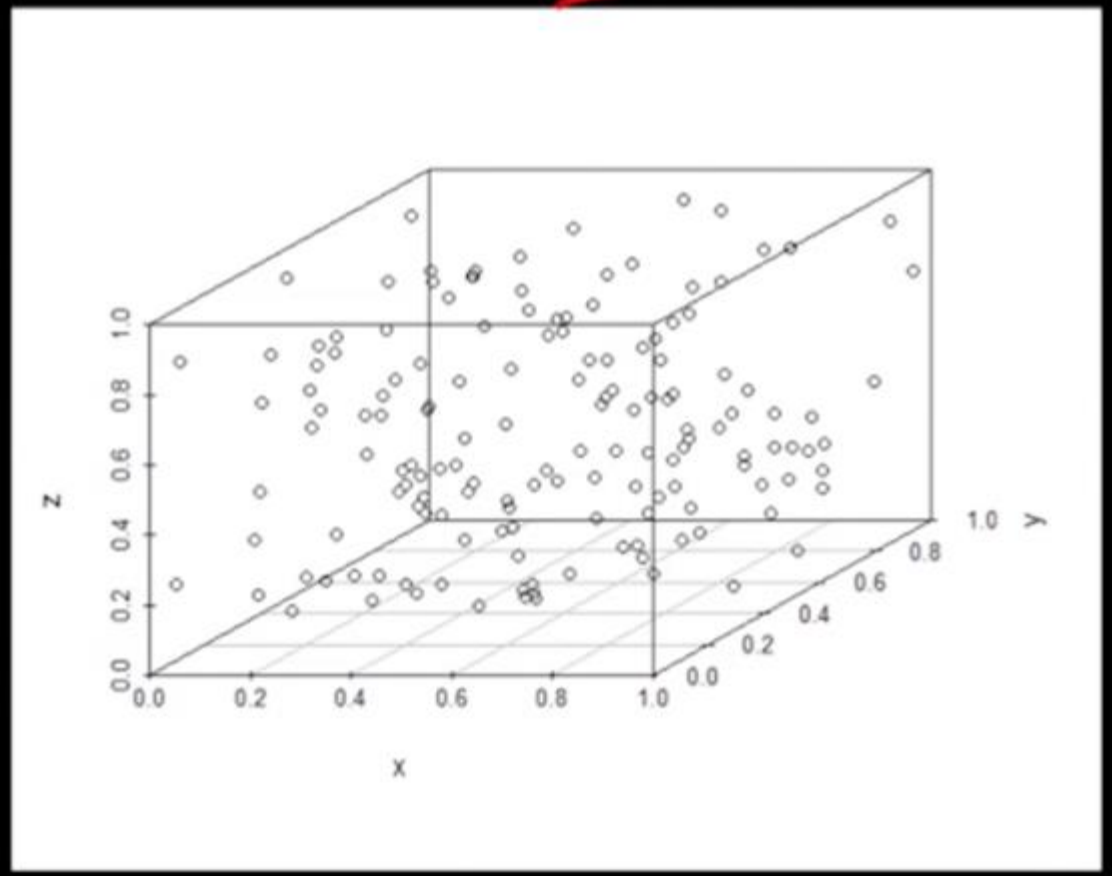
- **Feature Engineering (ou Feature Extraction)**

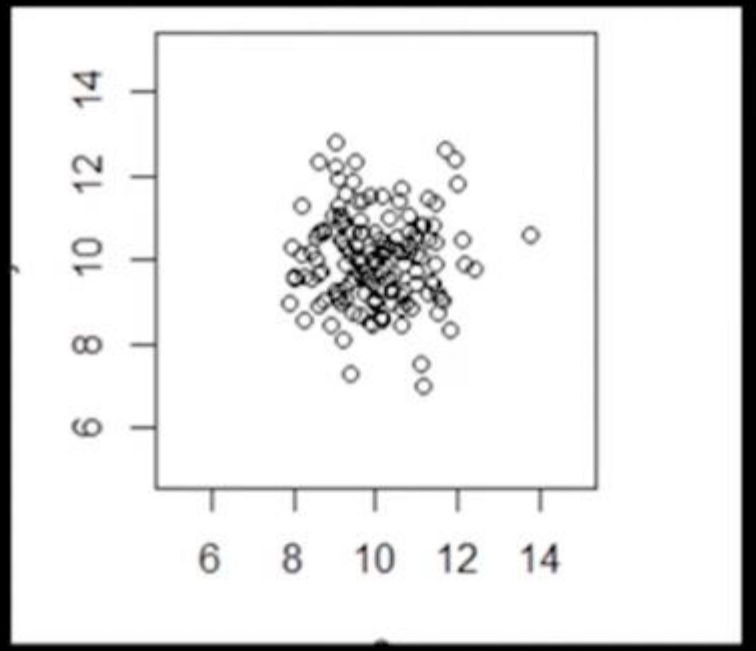
  Création de nouvelles caractéristiques pertinentes à partir des variables existantes.


Dimensionality Reduction
Feature Extraction
Feature Selection
Unsupervised Technique
Supervised Technique
Unsupervised Technique
Supervised Technique
Linear Method
Non-linear Method
Linear Method
Non-linear Method
Linear Method
Non-linear Method
Linear Method
Non-linear Method
PCA
t-SNE
LDA
QDA

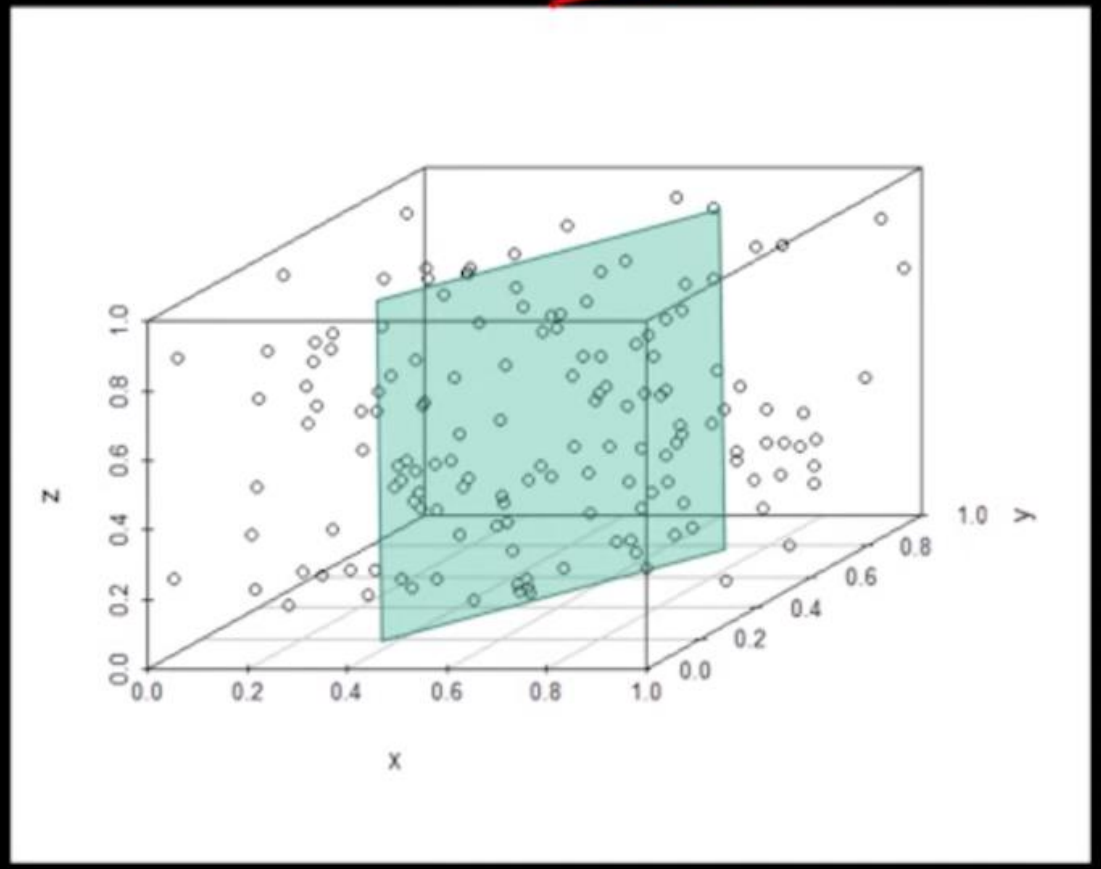

z
0.0
0.2
0.4
0.6
0.8
1.0
x
0.0
0.2
0.4
0.6
0.8
1.0
y
0.0
0.2
0.4
0.6
0.8
1.0

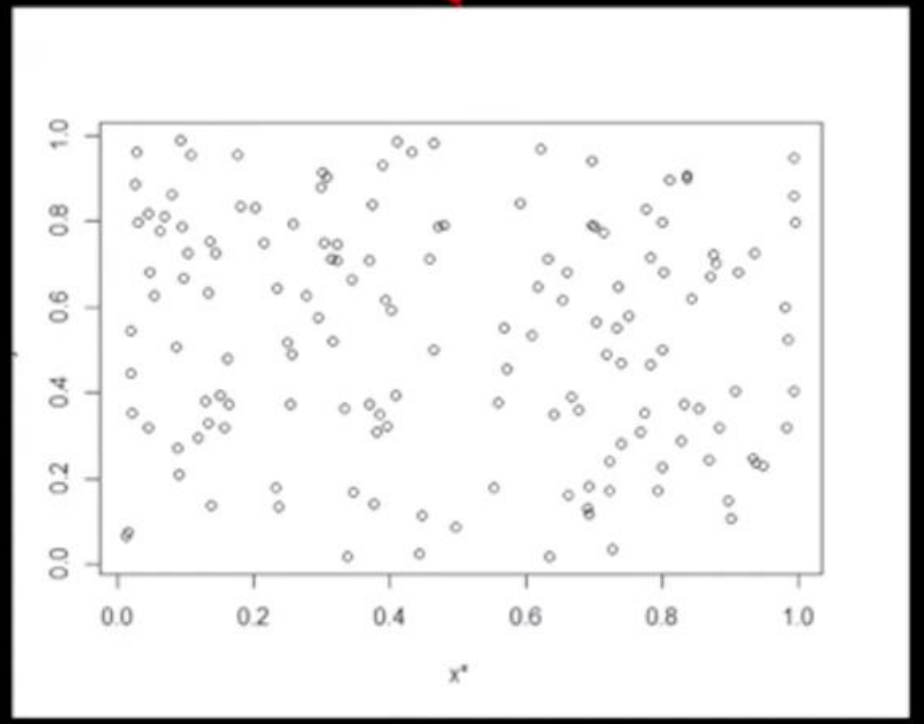

0.0
0.2
0.4
0.6
0.8
1.0
x*
0.0
0.2
0.4
0.6
0.8
1.0


z
0.0
0.2
0.4
0.6
0.8
1.0
x
0.0
0.2
0.4
0.6
0.8
1.0
y
0.0
0.2
0.4
0.6
0.8
1.0

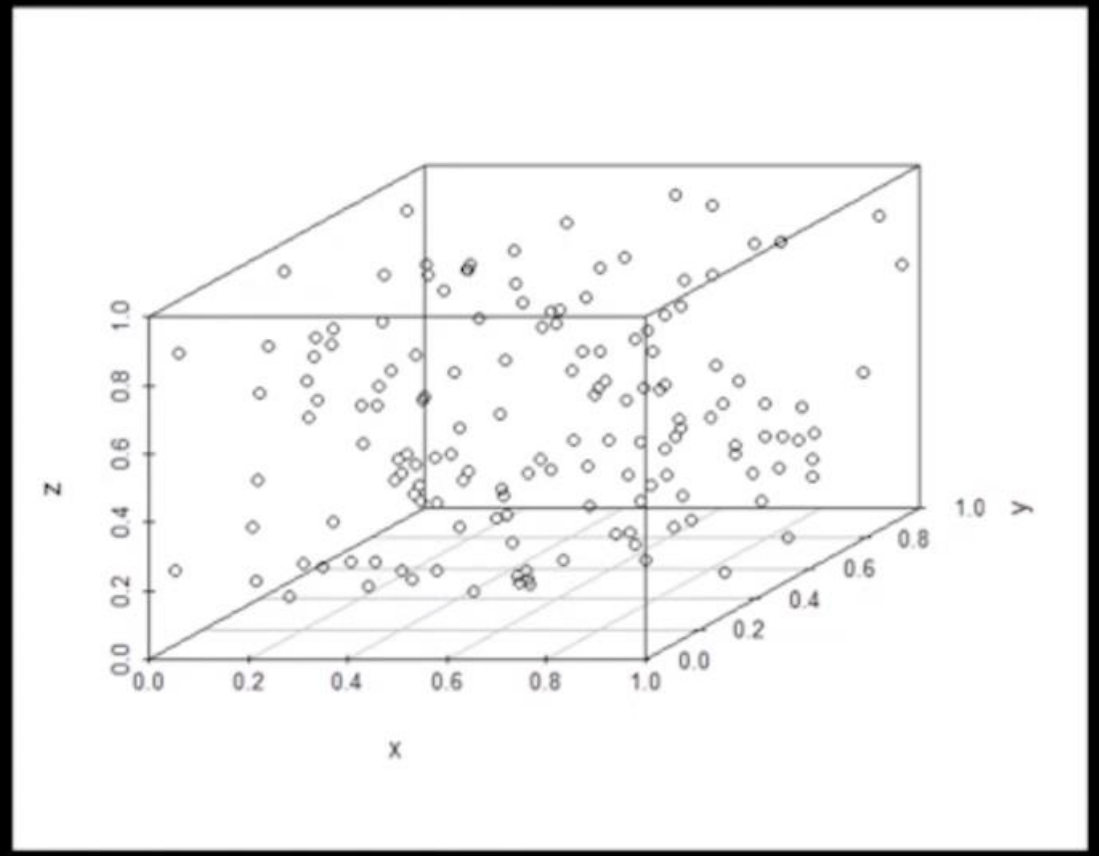

6
8
10
12
14
6
8
10
12
14

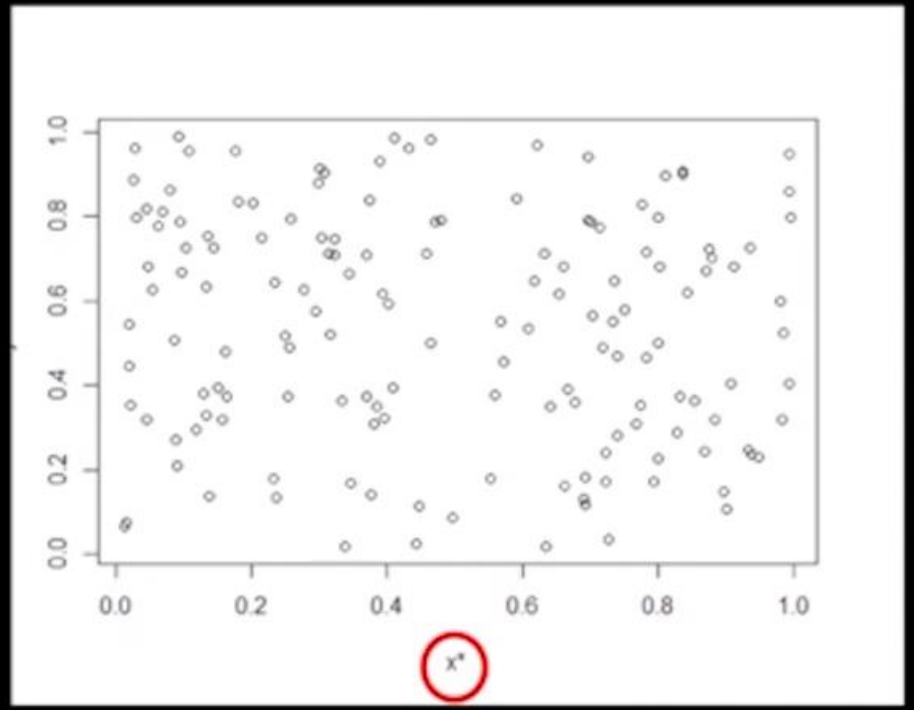
Echec





Succès





X* (et non X) : construit par combinaison linéaire des variables

Critères de choix de la méthode de réduction de dimensionalité

- Type et structure des données
- Dimensionnalité et rareté des données
- Variance et corrélation
- Visualisation et interprétation
- Evaluation et validation

Choix de la méthode de réduction de dimensionalité
Type et structure des données

- **Pour des données numériques et continues:**

  Privilégiez l'algèbre linéaire : PCA ou SVD.

- **En cas de données catégorielles ou discrètes:**

  Optez pour des méthodes basées sur la distance ou la similarité : MDS ou t-SNE.

- **Pour les données textuelles ou d'images:**

  Utilisez des méthodes axées sur l'extraction ou l'apprentissage de caractéristiques : LSA ou autoencodeurs.

Choix de la méthode de réduction de dimensionalité
Dimensionnalité et parcimonie

- **Si vos données ont de nombreuses caractéristiques mais peu d'observations:**

  Privilégiez des méthodes qui gèrent la malédiction de la dimensionalité, comme la projection aléatoire ou l'analyse factorielle.

- **Si vos données comportent de nombreux zéros ou des valeurs manquantes:**

  Optez pour des méthodes adaptées à la parcimonie des données, telles que l'ACP clairsemée (sparse PCA) ou la factorisation matricielle non négative (NMF).

Choix de la méthode de réduction de dimensionalité
Variance et corrélation

- **Si vos données présentent une variance élevée et une faible corrélation:**

  Optez pour des méthodes capturant la variance maximale, telles que PCA ou SVD.

- **Si vos données ont une faible variance et une forte corrélation:**

  Choisissez des méthodes capturant la corrélation minimale, comme l'analyse en composantes indépendantes (ICA) ou le Kernel PCA.

Choix de la méthode de réduction de dimensionalité
Visualisation et interprétation

- **Si vous souhaitez visualiser vos données dans un espace bidimensionnel ou tridimensionnel:**

  Utilisez des méthodes qui réduisent la dimensionalité à deux ou trois, telles que PCA, MDS ou t-SNE.

- **Pour interpréter vos données en fonction des caractéristiques ou variables d'origine:**

  Privilégiez des méthodes préservant l'interprétabilité, comme l'analyse factorielle, l'ICA ou la LSA.

Choix de la méthode de réduction de dimensionalité
Evaluation et validation

- **Si vous souhaitez évaluer vos données en termes de la quantité de variance expliquée par les dimensions réduites:**

  Utilisez des méthodes fournissant les mesures de variance, telles que PCA, SVD ou l'analyse factorielle.

- **Pour valider vos données en termes de la capacité des dimensions réduites à reconstruire les données originales:**

  Utilisez des méthodes fournissant l'erreur de reconstruction, telles que PCA, SVD ou les autoencodeurs.

CRISP-DM 3: Data Preparation

Transformation des données

Mohamed Anis MEJRI

Transformation des données

- **Encodage des Caractéristiques ("Feature Encoding")**
- **Mise à l'Échelle des Caractéristiques ("Feature Scaling")**
- **Transformation des variables.**

Transformation des données

Encodage des Caractéristiques ("Feature Encoding") :

- **Définition :** L'encodage des caractéristiques consiste à transformer des variables catégorielles en une forme numérique, de sorte qu'elles puissent être utilisées par des modèles d'apprentissage automatique.
- **Utilité :** Les algorithmes d'apprentissage automatique travaillent souvent mieux avec des données numériques, donc l'encodage des caractéristiques est nécessaire pour représenter des informations catégorielles de manière adaptée.

Transformation des données

Mise à l’échelle des Caractéristiques ("Feature Scaling") :

- **Définition :** La mise à l'échelle des caractéristiques vise à normaliser la plage des valeurs des caractéristiques. Cela garantit que toutes les caractéristiques contribuent de manière égale aux calculs des modèles d'apprentissage automatique.
- **Utilité :** Certains algorithmes d'apprentissage automatique sont sensibles à l'échelle des caractéristiques, et la mise à l'échelle aide à éviter que des caractéristiques avec des plages de valeurs différentes ne dominent l'apprentissage.

Transformation des données

Transformation des variables:

La transformation des variables consiste à modifier directement les variables descriptives de sorte à:

- Réarranger la distribution des données.
- Faciliter la convergence des algorithmes.

Transformation des données
Feature Encoding

- **Encodage à un chiffre (Label Encoding)**
- **Encodage à chaud (One-Hot Encoding)**
- **Encodage ordinal**
- **Encodage binaire (Binary Encoding)**
- **Encodage des fréquences (Frequency Encoding)**
- **Encodage en fonction de la cible (Target Encoding)**

Transformation des données
Feature Encoding

Encodage à un chiffre (Label Encoding)

- **Description :** Assigner à chaque catégorie un nombre entier unique.
- **Exemple :** Si les catégories sont {Rouge, Vert, Bleu}, le label encoding pourrait les transformer en {0, 1, 2}.

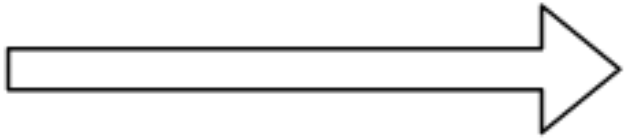
Transformation des données
Feature Encoding

Encodage à chaud (One-Hot Encoding)

- **Description :** Créer des variables binaires distinctes pour chaque catégorie.
- **Exemple :** Si les catégories sont {Rouge, Vert, Bleu}, le one-hot encoding pourrait les transformer en trois variables binaires distinctes (par exemple, [1, 0, 0], [0, 1, 0], [0, 0, 1]).

Color
Red
Green
Blue

One-hot encoding



d1	d2	d3
1	0	0
0	1	0
0	0	1

Transformation des données
Feature Encoding

Encodage ordinal

- **Description :** Assigner des entiers à des catégories ordinales selon leur ordre.
- **Exemple :** Si les catégories sont {Petit, Moyen, Grand}, l'encodage ordinal pourrait les transformer en {0, 1, 2}.

Transformation des données
Feature Encoding

Encodage binaire (Binary Encoding)

- **Description :** Assigner à chaque catégorie un chiffre distinct convertit en binaire.
- **Exemple :** Si les catégories sont {A, B, C}, le binary encoding pourrait les transformer en {00, 01, 10}.

Level	"Decimal encoding"	Binary encoding	One hot encoding
No	0	000	000001
Primary	1	001	000010
Secondary	2	010	000100
BSc/BA	3	011	001000
MSc/MA	4	100	010000
PhD	5	101	100000

Transformation des données
Feature Encoding

Encodage des fréquences (Frequency Encoding)

- **Description :** Remplacer chaque catégorie par sa fréquence d'occurrence dans l'ensemble de données.
- **Exemple :** Si la catégorie A apparaît 5 fois, elle serait remplacée par 5.

Transformation des données
Feature Encoding

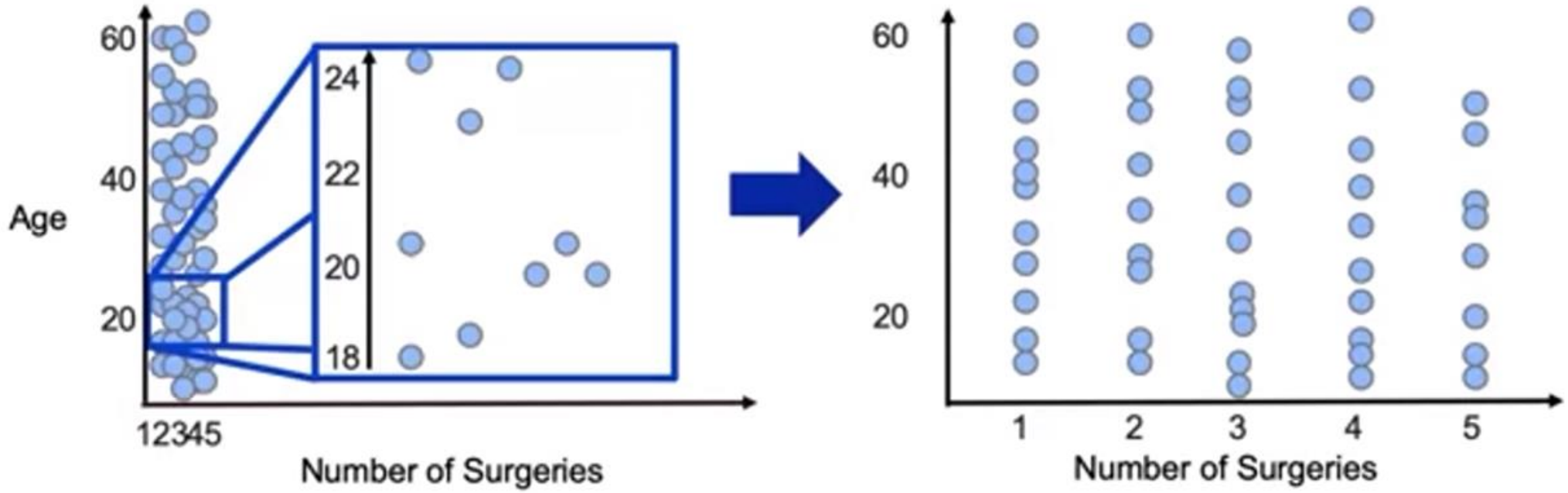
Encodage en fonction de la cible (Target Encoding)

- **Description :** Utiliser les statistiques de la variable cible pour encoder les catégories.
- **Exemple :** Remplacer chaque catégorie par la moyenne de la variable cible pour cette catégorie.

Transformation des données
Feature Scaling

- Les techniques de mise à l'échelle de caractéristiques les plus courantes sont la normalisation (Min-Max Scaling) et la standardisation (Z-score normalization).
- Ces méthodes sont largement utilisées et servent de techniques fondamentales pour le prétraitement des caractéristiques numériques dans les applications d'apprentissage automatique.

Feature Scaling: Example



Transformation des données
Feature Scaling

Normalisation (Min-Max Scaling):

- **Formule :** (X-min(X)) / (max(X)-min(X))
- **Description :**
  - Met à l'échelle les caractéristiques dans une plage spécifique, généralement entre 0 et 1.
  - Cette méthode est sensible aux valeurs aberrantes. Elle est donc à utiliser lorsque les données sont uniformément distribuées sans valeurs aberrantes extrêmes.
  - Elle est utile lorsque les caractéristiques ou les algorithmes dépendent de plages spécifiques.

Transformation des données
Feature Scaling

Standardisation (Z-score normalization):

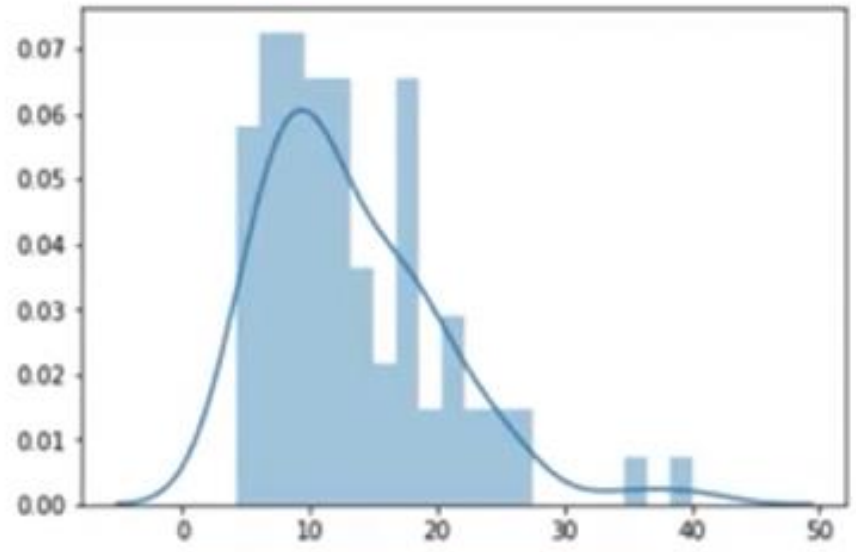
- **Formule: (X – moyenne(X)) / écart-type(X))**
- **Description :**
  - Transforme les caractéristiques pour avoir une moyenne de 0 et un écart-type de 1.
  - Elle est moins sensible aux valeurs aberrantes par rapport à la normalisation.
  - Produit de bons résultats lorsque les caractéristiques ou les algorithmes supposent une distribution normale centrée à zéro (PCA, KNN, SVM).

Transformation des données
Transformation des variables

- Transformation logarithmique
- Transformation polynomiale

Log Transformation Example

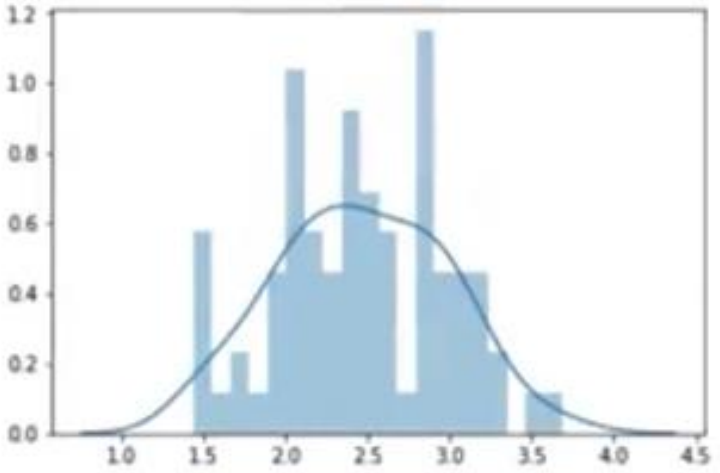
```
# plot a histogram and density plot
sns.distplot(data, bins=20);
```



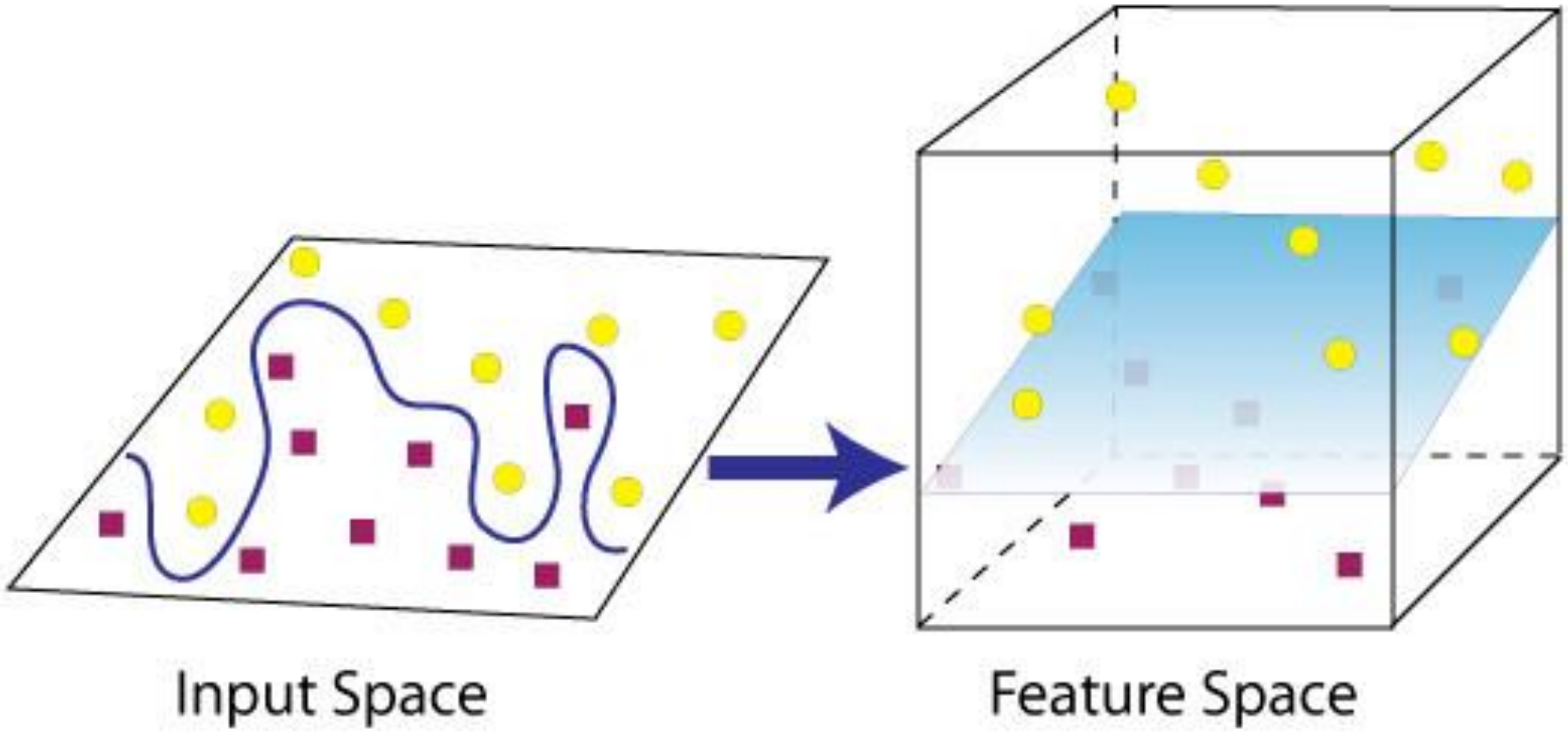
Right (positive) skewed

```
import math
log_data = [math.log(d) for d in data['Unemployment']]

# plot transformed plots
sns.distplot(log_data, bins=20);
```



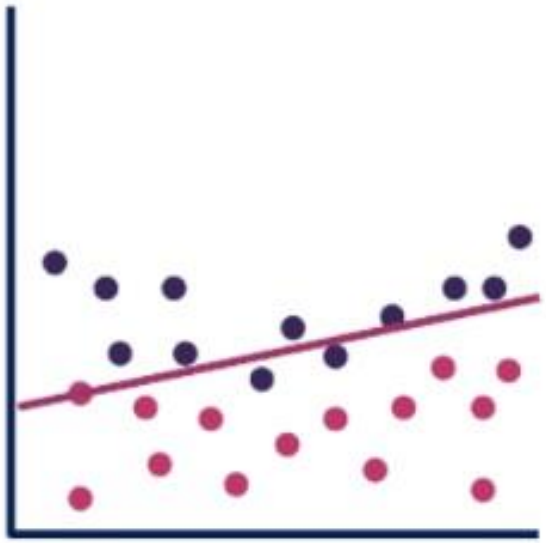
Normal



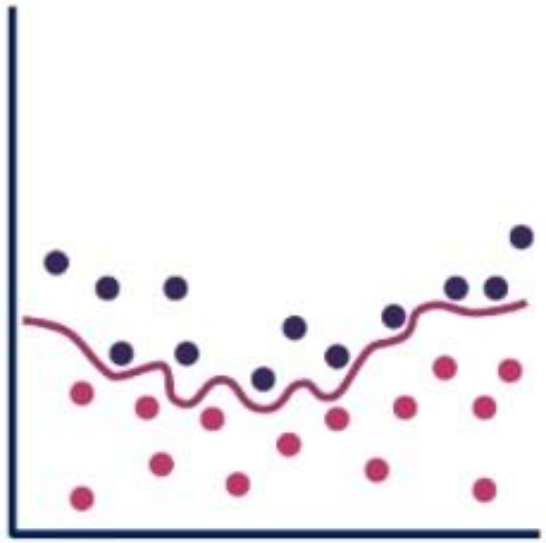
Feature Transformation

CRISP-DM 3: Data Preparation

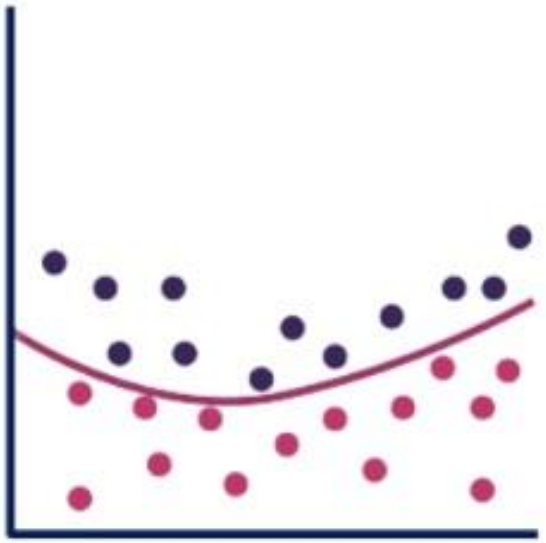
Separation des données d’entrainement, validation et de test



Underfitting

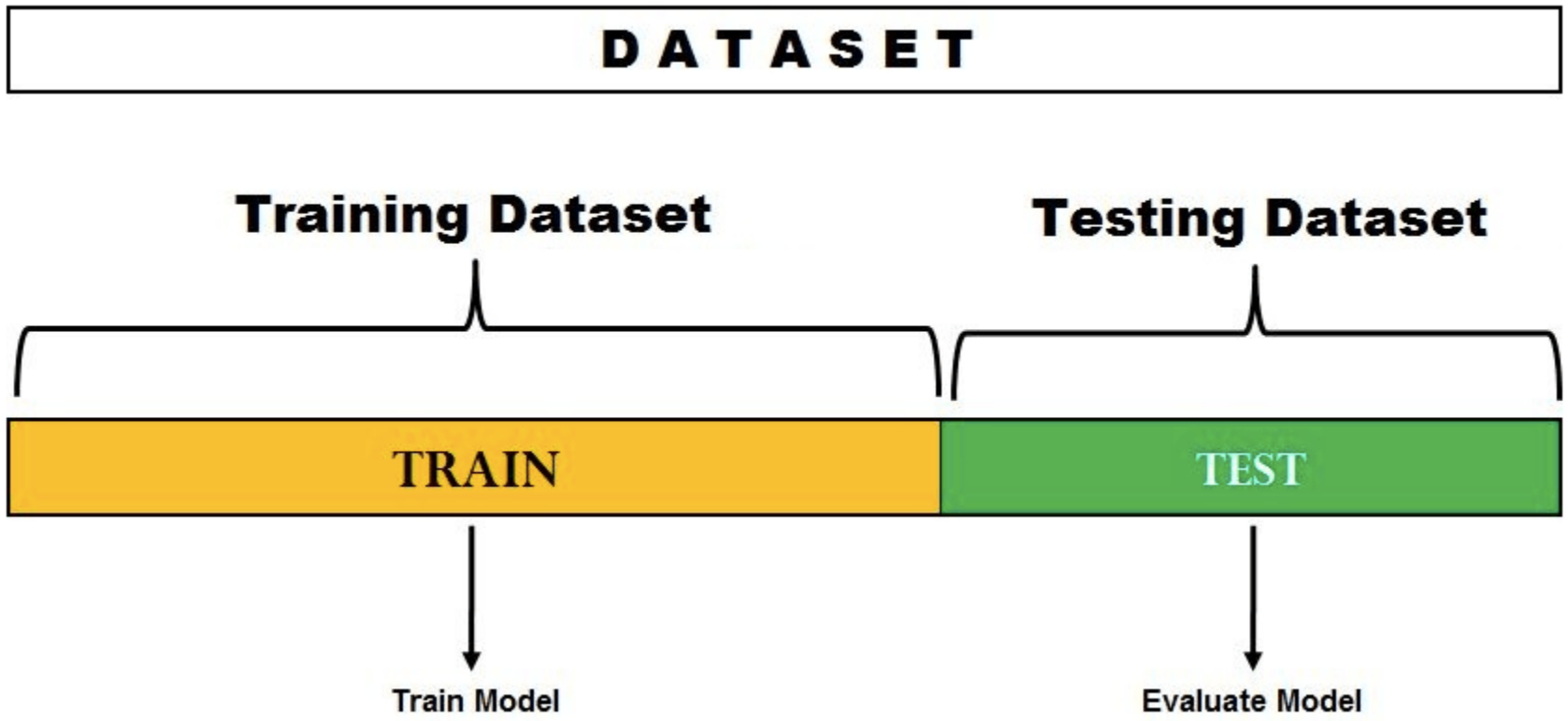


Overfitting

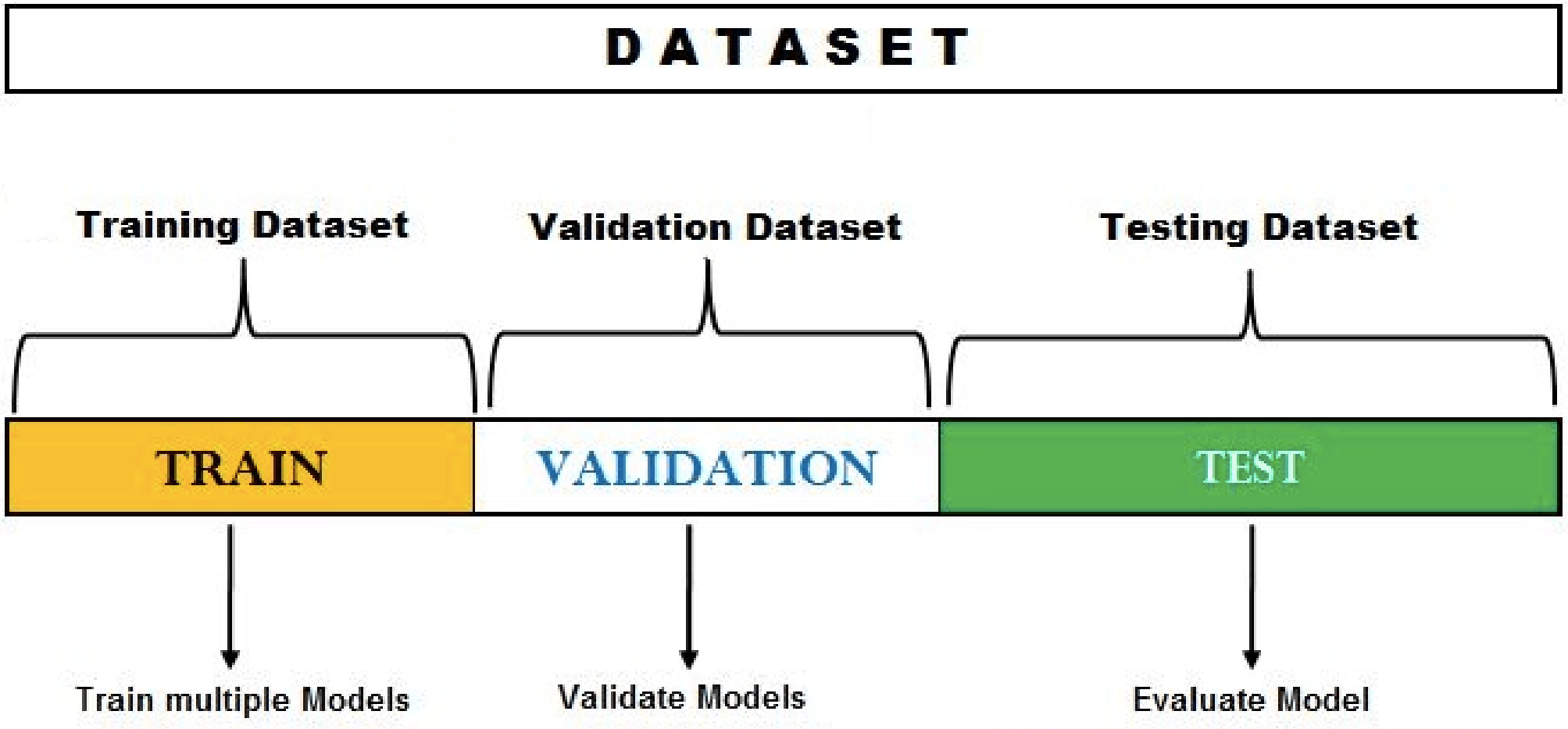


Balanced

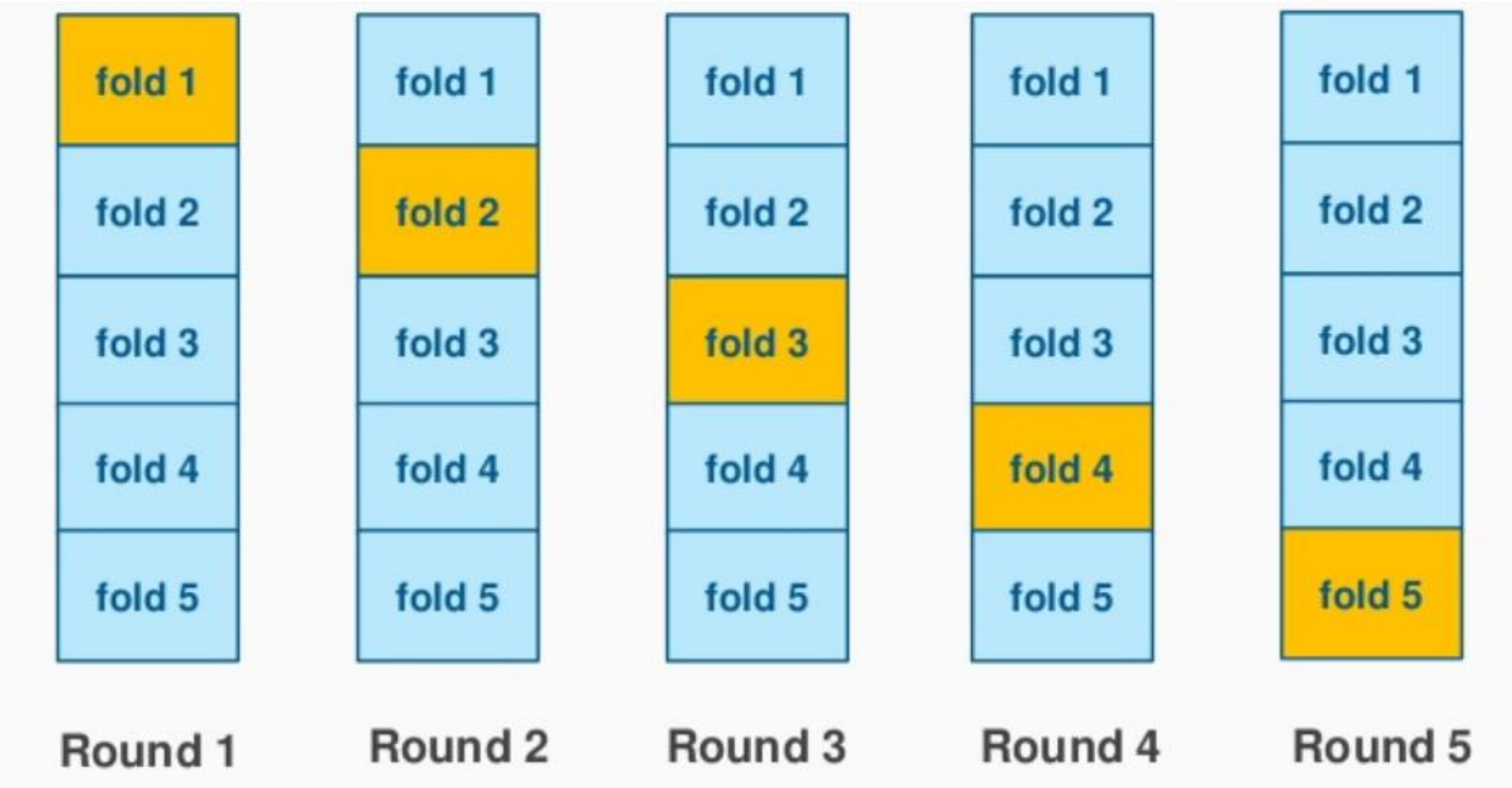
Hold-out validation (Train / Test)



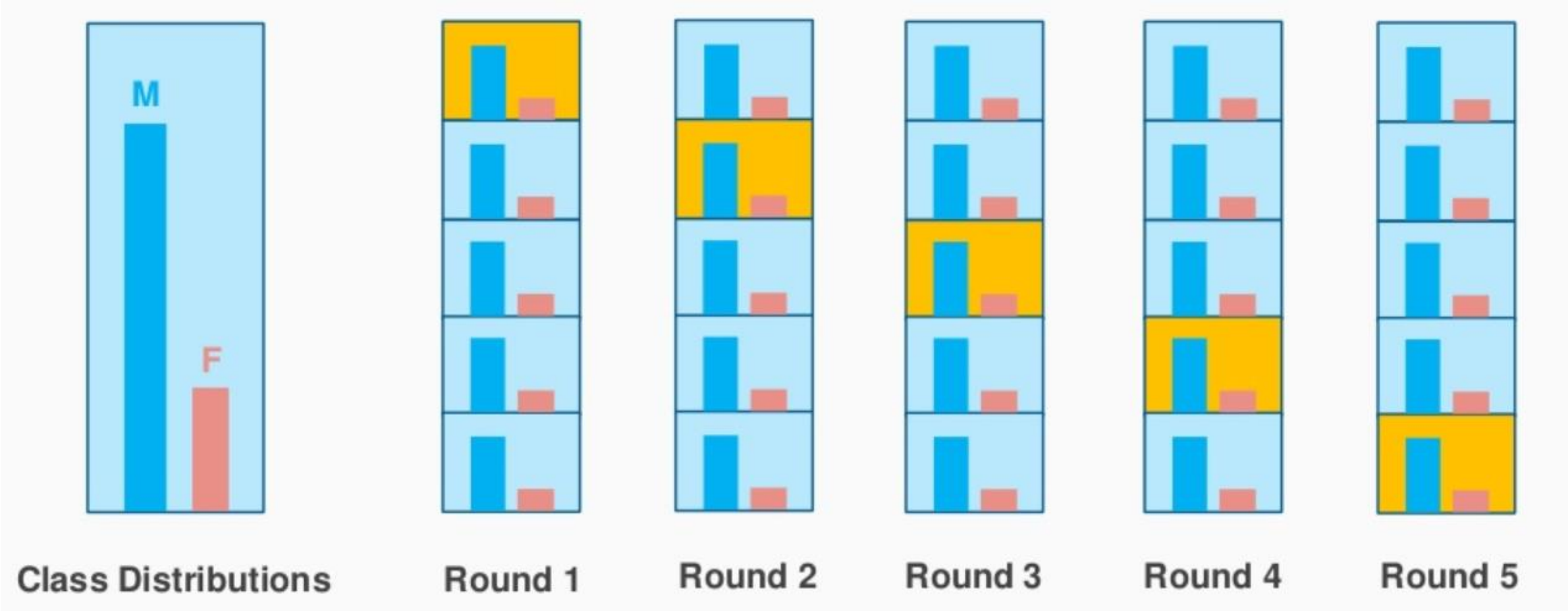
Hold-out validation (Train / Validation / Test)



K-Fold Cross-Validation



Stratified K-Fold cross-Validation

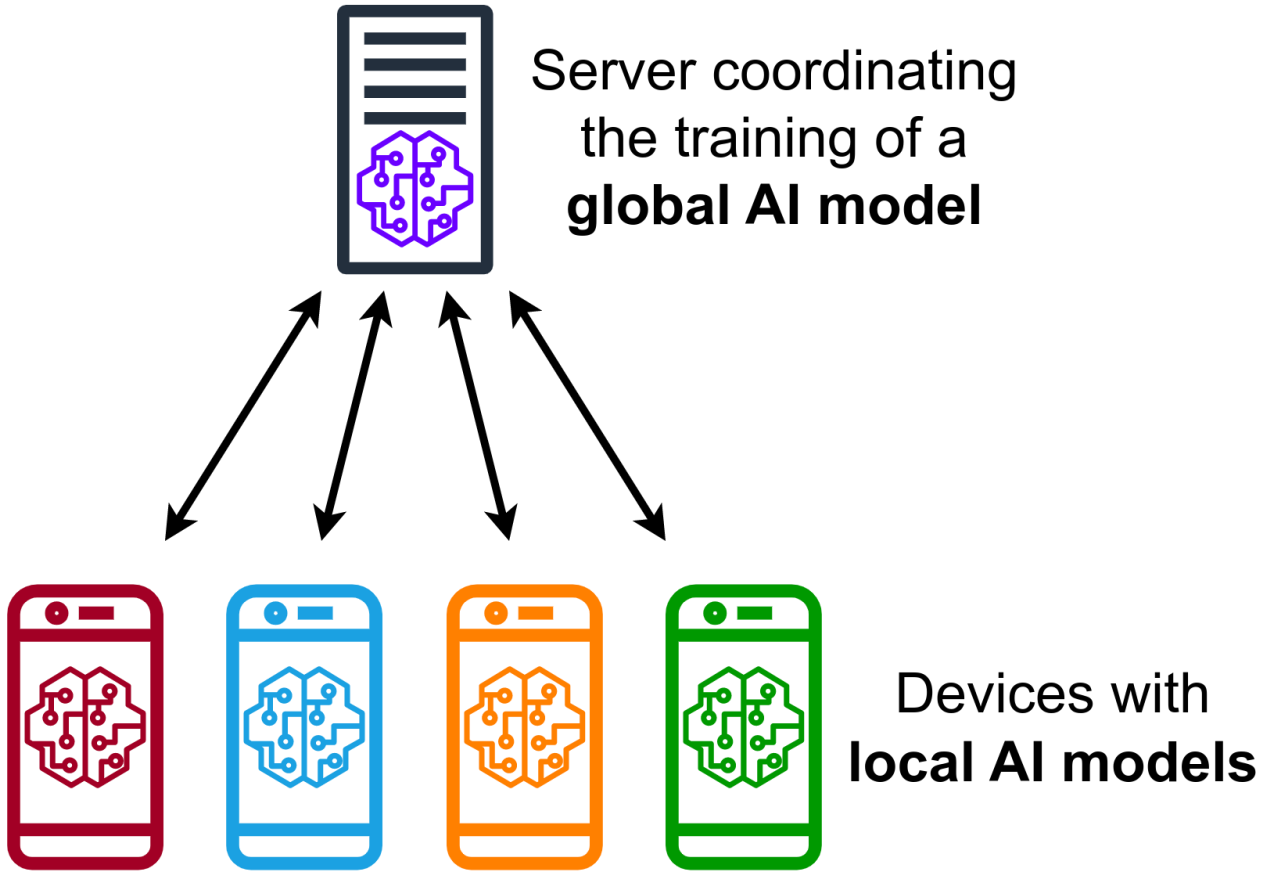


CRISP-DM 4: Modeling

CRISP-DM 4: Modeling

Confidentialité et droits d’auteur

Federated Learning



Differential privacy