



An abstract background featuring a dense, colorful pattern of overlapping circles in shades of orange, yellow, green, blue, and purple, creating a textured, liquid-like effect.

ANALYSE ET VISUALISATION DES DONNÉES

Alexandra-Maria DOBRESCU

| Cours 9 |

Collection et préparation des données

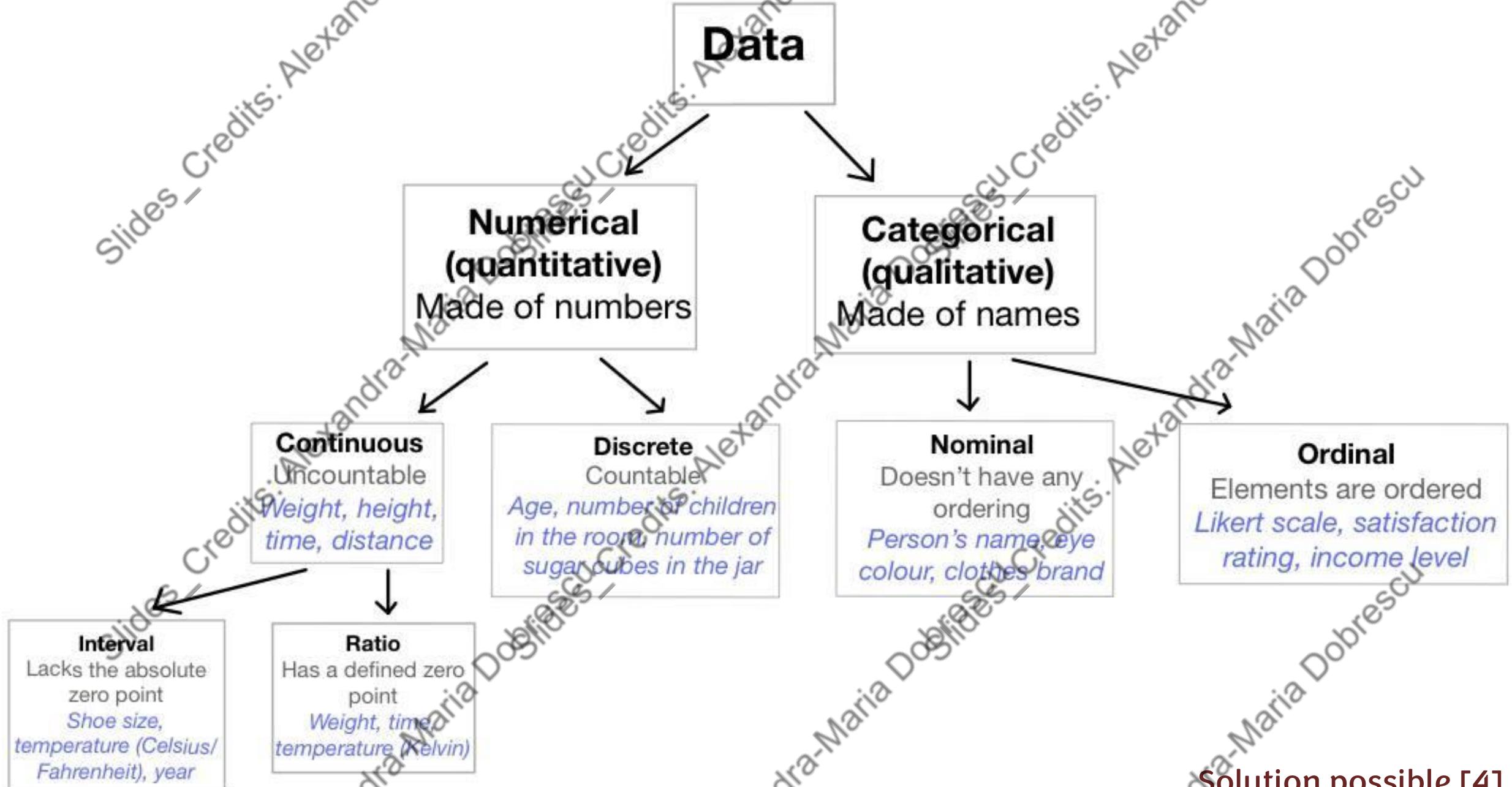
- SOURCES ET ACQUISITION DES DONNÉES
- NETTOYAGE ET PRÉTRAITEMENT DES DONNÉES
- TYPES DE DONNÉES
- LA MESURE

Types de données

CATÉGORIELLES ~ NUMÉRIQUES LES TYPES D'ÉCHELLES

- ✓ Il est indispensable de comprendre les différents types de données et la manière dont elles peuvent être classifiées.
- ✓ Nous allons nous concentrer sur deux catégories principales de types de données :
 - les données catégorielles et les données numériques,
 - ainsi que sur **les types d'échelles**:
 - Données nominales
 - Données ordinale
 - Données d'intervalle
 - Données de rapport

Solution possible [4]



Types de données

NOMINALES

Les valeurs appartenant à une échelle nominale sont caractérisées par des étiquettes.

Les valeurs ne sont pas ordonnées et sont également pondérées.

On ne peut pas calculer la moyenne ou la médiane à partir d'un ensemble de valeurs de ce type.

Les données nominales sont catégoriques mais peuvent être parfois être traitées comme des données numériques en attribuant des nombres aux étiquettes.

En revanche, on peut déterminer le mode, c'est-à-dire la valeur la plus fréquente.

Types de données

ORDINALES

Les valeurs de ce type sont ordonnées, mais la différence ou la distance entre deux valeurs ne peut être déterminée.

Les valeurs ne déterminent que le rang ou la position dans l'ensemble.

Pour ces valeurs, on peut calculer le mode ou la médiane (la valeur placée au milieu de l'ensemble ordonné) mais pas la moyenne.

Ces valeurs sont catégoriques par essence mais peuvent être traitées comme des valeurs numériques.

- L'ensemble des grades militaires;
- L'ordre des marathoniens aux Jeux olympiques (sans les temps);

Types de données

INTERVALLE

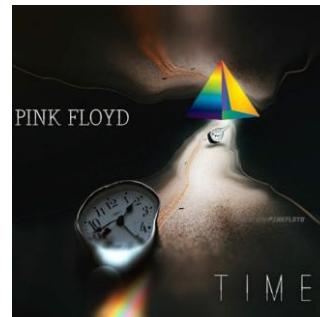
Il s'agit de valeurs numériques.

A échelle d'intervalle, la différence entre deux valeurs est significative.

On peut calculer la moyenne, l'écart type ou utiliser la régression pour prédire de nouvelles valeurs.

On considère le zéro comme une valeur arbitraire fixe.
PS: Zéro ne signifie pas « rien ».

- La température à l'aide de l'échelle Celsius est un attribut à échelle d'intervalle.
- C'est pourquoi les valeurs négatives sont également autorisées.
- Des exemples ?



Types de données

Il s'agit de valeurs numériques.

On considère le zéro comme une valeur arbitraire fixe.

INTERVALLE

A échelle d'intervalle, la différence entre deux valeurs est significative.

- La température à l'aide de l'échelle Celsius est un attribut à échelle d'intervalle.

People also ask :

What is 0 Anno Domini?

Well, actually there is no year 0; the calendar goes straight from 1 BC to 1 AD, complicating the process of calculating years. Most scholars believe that Jesus was born between 6 and 4 BC (Before Christ) and that he died between 30 and 36 AD (Anno Domini, latin for "in the year of the lord"). Dec 7, 2016



popularmechanics.com

<https://www.popularmechanics.com/archaeology/case...>

Types de données

RAPPORT

Ces attributs sont comme les attributs à échelle d'intervalle.

PS: Zéro signifie « rien ».

Toutes les opérations mathématiques peuvent être effectuées.

Les valeurs négatives ne sont pas autorisées.

Le rapport entre deux valeurs est significatif.

- Logarithmes
- Les moyennes géométriques et harmoniques
- Le coefficient de variation

- salaire : Le salaire d'un développeur senior est deux fois plus élevé que celui d'un développeur junior.
- IoT exemples?

Des conclusions intéressantes

CES CATÉGORIES N'ONT PAS
D'ORDRE SIGNIFICATIF.

LES DONNÉES
NOMINALES NE PEUVENT
PAS ÊTRE QUANTIFIÉES.

ETHNICITY
(HISPANIC, ASIAN)

LES VALEURS SONT
UNIQUEMENT ATTRIBUÉES À
DES CATÉGORIES DISTINCTES.

MARITAL STATUS
(MARRIED,
SINGLE,
WIDOWED)

CARACTÉRISTIQUES
PRINCIPALES DES
DONNÉES
NOMINALES

GENDER
(WOMEN, MEN)

HAIR COLOR
(BLONDE,
BROWN,
BRUNETTE)

ELLES NE PEUVENT PAS NON
PLUS ÊTRE CLASSÉES DANS
UN ORDRE QUELCONQUE.

HOUSING STYLE
(RANCH HOUSE,
MODERNIST, ART
DECO)

L'échelle nominale

L'échelle nominale permet de classer les données non numériques en catégories.

Les échelles nominales pourraient simplement être appelées « étiquettes » (labels).

Une échelle nominale ne comportant que deux catégories (féminin/masculin) est dite « dichotomique ».

Les échelles nominales s'excluent mutuellement (pas de chevauchement) et n'ont pas de valeur numérique

- Mettre des pays dans des continents. Exemple : La Bulgarie est un pays d'Europe.

Des conclusions intéressantes

{ DE NOMBREUX EXEMPLES DE SEGMENTATION DU MARCHÉ CONSTITUENT UNE BASE POUR LA CRÉATION D'ÉCHELLES ET DE MESURES NOMINALES. }

Segmentation géographique

- ✓ État (Californie, Colorado, Texas)
- ✓ Pays (Roumanie, Portugal, France)
- ✓ Collège
- ✓ Département
- ✓ Continent



- ✓ Communauté
- ✓ Urbain/rural
- ✓ Région nord/sud
- ✓ Régions chaudes/froides
- ✓ Zones de haute altitude/de basse altitude

Des conclusions intéressantes

{ DE NOMBREUX EXEMPLES DE SEGMENTATION DU MARCHÉ CONSTITUENT UNE BASE POUR LA CRÉATION D'ÉCHELLES ET DE MESURES NOMINALES. }

Segmentation démographique

- ✓ Âge
- ✓ Race
- ✓ Religion
- ✓ Sexe
- ✓ Taille de la famille (Couple seulement, 3 membres de la famille, 4 membres de la famille)
- ✓ Revenu
- ✓ Niveau d'éducation (lycée, université, formation professionnelle)



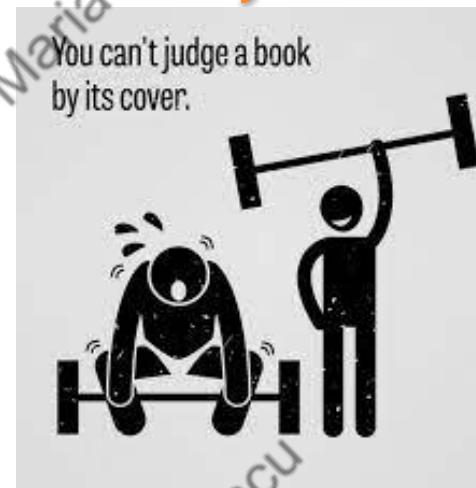
- ✓ Ethnie (Hispanique, Asiatique)
- ✓ Style d'habitation
- ✓ Situation de famille
- ✓ Profession
- ✓ Groupe socio-économique (cadres supérieurs, cadres intermédiaires, personnel administratif, professions libérales)

Des conclusions intéressantes

{ DE NOMBREUX EXEMPLES DE SEGMENTATION DU MARCHÉ CONSTITUENT UNE BASE POUR LA CRÉATION D'ÉCHELLES ET DE MESURES NOMINALES. }

Segmentation psychographique

- ✓ Classe (Famille de travailleurs, Classe moyenne, Famille de la classe supérieure)
- ✓ Personnalité ("ouvert", "créatif", "sérieux")
- ✓ Attitudes (Espoir, Optimisme, Réalisme)
- ✓ Modes de vie (mode de vie sain, mode de vie sans gluten)
- ✓ Hobbies (Lecture, Equitation, Ballons de football, Marche)
- ✓ Intérêts (Intérêt pour la nourriture, la technologie, la mode)
- ✓ Traits de caractère particuliers (Religieux, Aventureux)
- ✓ Style d'habitation
- ✓ Attentes (explicites, implicites, technologiques)
- ✓ Opinions (Clients qui évaluent le produit avec 3 étoiles, 4 étoiles, 5 étoiles)



Des conclusions intéressantes

{ DE NOMBREUX EXEMPLES DE SEGMENTATION DU MARCHÉ CONSTITUENT UNE BASE POUR LA CRÉATION D'ÉCHELLES ET DE MESURES NOMINALES. }

Segmentation comportementale

- ✓ Occasion (anniversaire, remise de diplôme);
- ✓ Étape du parcours Bayer (sensibilisation, considération, décision);
- ✓ Connaissance de la marque (Aucune, Un peu, Forte connaissance);
- ✓ Types de fidélité (pas de fidélité, fidélité par inertie, fidélité latente, fidélité premium);
- ✓ Sensibilité au prix (Sensible aux changements de prix, Pas sensible)
- ✓ Style d'achat (Éviter les achats, Aimer les achats)
- ✓ Taux d'utilisation (Fort, Faible)
- ✓ Statut de l'utilisateur (Jamais, Occasionnel, Régulier)



[3]

Des conclusions intéressantes

{ DE NOMBREUX EXEMPLES DE SEGMENTATION DU MARCHÉ CONSTITUENT UNE BASE POUR LA CRÉATION D'ÉCHELLES ET DE MESURES NOMINALES. }

Segmentation des médias

- ✓ Télévision (télévision par câble, télévision en réseau, télévision par satellite);
- ✓ Radio (utilisation de la radio par satellite, radio internet, radio locale, radio nationale)
- ✓ Médias sociaux (Facebook, Twitter, Instagram);
- ✓ Journaux (journal local, journal national, magazine de consommateurs)
- ✓ Utilisation des moteurs de recherche sur internet (Google, Bing, Yahoo)
- ✓ Assistants intelligents, chatbots



GPT - 4



Des conclusions intéressantes

{ DE NOMBREUX EXEMPLES DE SEGMENTATION DU MARCHÉ CONSTITUENT UNE BASE POUR LA CRÉATION D'ÉCHELLES ET DE MESURES NOMINALES. }

Segmentation temporelle

- ✓ Saisons (hiver, printemps, été, automne);
- ✓ Événements spéciaux (Black Friday);
- ✓ Fêtes (Thanksgiving, Noël).



Segmentation des prestations

- ✓ Confort (sans effort d'achat, avec certains efforts d'achat);
- ✓ Service à la clientèle (attendu, souhaité, service à la clientèle de base)
- ✓ Caractéristiques spéciales (grande vitesse, accès facile)
- ✓ Qualité (haute qualité, qualité moyenne, basse qualité)

Des conclusions intéressantes

1

Les données ordinaires sont placées dans un certain ordre.

2

Les nombres ordinaires n'indiquent qu'une séquence.

3

Nous pouvons attribuer des numéros aux données ordinaires.

4

Nous ne pouvons pas faire d'arithmétique avec des nombres ordinaires.

Sont-ils catégoriques ?

Les variables ordinaires se situent « entre » les variables catégorielles et les variables quantitatives.

Données binaires

Il arrive qu'un attribut n'ait que deux valeurs, comme le sexe dans l'exemple précédent. Dans ce cas, l'attribut est appelé **binaire**.

💡 Les attributs binaires peuvent être traités comme des intervalles ou des rapports, mais dans la plupart des cas, ils doivent être traités comme des **attributs nominaux** (binaires symétriques) ou **ordinaux** (binaires asymétriques).

📏 Il existe un ensemble de fonctions de similarité et de dissimilarité (distance) spécifiques aux attributs binaires.

Binaire symétrique: lorsque les deux valeurs ont le même poids et la même importance et ont la même importance (comme dans le cas du sexe).

Binaire asymétrique: l'une des valeurs est plus importante que l'autre.
Distribution inégale des 0 et des 1, créant un motif non uniforme

00110011 **vs** 01001011

Collection et préparation des données

- SOURCES ET ACQUISITION DES DONNÉES
- NETTOYAGE ET PRÉTRAITEMENT DES DONNÉES
- TYPES DE DONNÉES
- LA MESURE

Mesure des données

MESURE DE LA TENDANCE CENTRALE

{ **LA TENDANCE CENTRALE EST LA MESURE STATISTIQUE QUI PRÉSENTE LE CENTRE OU LE MILIEU D'UNE DISTRIBUTION DE VALEURS.** }

Remarque: Il existe plusieurs mesures de la tendance centrale, et le choix de celle à utiliser dépend de la nature des données et des caractéristiques spécifiques que vous souhaitez saisir.

M 1: La moyenne (Mean) est calculée en additionnant toutes les valeurs d'un ensemble de données, puis en les divisant par le nombre de valeurs.

💡 Elle est sensible aux valeurs extrêmes (aberrantes) des données.

M 2: La médiane (Median) est la valeur centrale d'un ensemble de données lorsqu'il est ordonné. S'il y a un nombre pair de valeurs, la médiane est la moyenne des deux valeurs médianes.

💡 La médiane est moins sensible aux valeurs aberrantes que la moyenne.

Mesure des données

MESURE DE LA TENDANCE CENTRALE

LA TENDANCE CENTRALE EST LA MESURE STATISTIQUE QUI PRÉSENTE LE CENTRE OU LE MILIEU D'UNE DISTRIBUTION DE VALEURS.

M 3: Le mode est la valeur qui apparaît le plus fréquemment dans un ensemble de données. Un ensemble de données peut avoir un seul mode (unimodal), plusieurs modes (multimodal) ou aucun mode.

Remarque 1: Ces mesures offrent différentes perspectives sur la tendance centrale d'un ensemble de données, et le choix de celle à utiliser dépend des caractéristiques des données et des objectifs de l'analyse.

Remarque 2: Outre ces mesures traditionnelles, il existe d'autres mesures de tendance centrale pour des situations spécifiques, telles que la moyenne géométrique pour les produits de valeurs et la moyenne harmonique pour les taux.

💡 Chaque mesure a ses propres forces et faiblesses, et le choix dépend du contexte des données et de l'analyse.

Mesure des données

MESURE DE LA DISPERSION

EN STATISTIQUE, LA DISPERSION FAIT RÉFÉRENCE À L'ÉTENDUE OU À LA VARIABILITÉ D'UN ENSEMBLE DE DONNÉES.

Remarque: Il existe plusieurs mesures de dispersion qui permettent de quantifier le degré de dispersion ou de regroupement des valeurs au sein d'un ensemble de données.

M 1: L'étendue (range) est la mesure la plus simple de la dispersion et se calcule comme la différence entre les valeurs maximales et minimales d'un ensemble de données.

 Bien que facile à calculer, l'étendue peut être sensible aux valeurs aberrantes.

M 2: La variance mesure la distance entre chaque point de données de l'ensemble et la moyenne. Elle consiste à calculer la moyenne des différences au carré entre chaque point de données et la moyenne.

Mesure des données

MESURE DE LA DISPERSION

EN STATISTIQUE, LA DISPERSION FAIT RÉFÉRENCE À L'ÉTENDUE OU À LA VARIABILITÉ D'UN ENSEMBLE DE DONNÉES.

M 3: L'écart-type est la racine carrée de la variance. Il fournit une mesure de la distance moyenne entre chaque point de données et la moyenne.

M 4: L'IQR est l'intervalle dans lequel se situent les 50% du milieu des données.

 Il est moins sensible aux valeurs aberrantes que l'intervalle.

IQR = Q3-Q1 (Q1 pour 25% est le premier quartile et Q3 pour 75% est le troisième quartile).

Remarque: Le choix de la mesure à utiliser dépend des caractéristiques des données et des objectifs spécifiques de l'analyse.

Outliers

"Minimum"
($Q1 - 1.5 \cdot IQR$)

Interquartile Range
(IQR)

Q1

Median

Q3

(25th Percentile) (75th Percentile)

Outliers

"Maximum"
($Q3 + 1.5 \cdot IQR$)

Source [4]

Exercice : Analyse des notes d'examen

MESURE DE LA TENDANCE CENTRALE

75, 89, 92, 64, 78, 92, 88, 91, 72, 85,
64, 78, 92, 75, 88, 94, 69, 72, 85, 90

Un enseignant a enregistré les notes d'examen (sur 100) d'une classe de 20 élèves.

E 1: Trouvez la moyenne des notes obtenues à l'examen.

E 2: Classez les notes de la plus basse à la plus haute.

Trouvez la médiane, c'est-à-dire la note du milieu de la liste ordonnée.

E 3: Déterminez s'il existe un mode (la note la plus fréquente). S'il y a plus d'un mode, dressez la liste de tous les modes.

E 4: Examinez les mesures de tendance centrale (moyenne, médiane, mode) et réfléchissez à ce que chacune d'entre elles indique sur la distribution des notes d'examen.

Exercice : Analyse des températures quotidiennes

MESURE DE LA DISPERSION

75, 89, 92, 64, 78, 92, 88, 91, 72, 85,
64, 78, 92, 75, 88, 94, 69, 72, 85, 90

Températures maximales quotidiennes (en degrés Fahrenheit) pour une ville au cours des deux dernières semaines.

E 1: Trouvez l'intervalle des températures quotidiennes.

E 2: Calculez la variance des températures (n est le nombre de jours, x_i chaque température quotidienne, \bar{x} la moyenne).

E 3: Calculez l'écart-type en utilisant la racine carrée de la variance.

E 4: Déterminez l'intervalle interquartile.

E 5: Examinez ce que chaque mesure de dispersion révèle sur la répartition des températures quotidiennes. Existe-t-il des tendances ou des motifs ? Comment la présence de valeurs aberrantes affecte-t-elle ces mesures ?

Bibliographie

- [1] Subramaniam, A. (2020). What Is Big Data Analytics| Big Data Analytics Tools and Trends| Edureka.
- [2] <https://ucarecdn.com/2bc4eb6c-4c71-4679-8c0b-308b293b8515/>
- [3] <https://salespanel.io/blog/marketing/b2b-behavioral-segmentation/>
- [4] <https://www.kdnuggets.com/2019/11/understanding-boxplots.html>