

The background of the slide is decorated with numerous paint splashes in various shades of yellow, orange, green, and blue, creating a vibrant, abstract pattern.

ANALYSE ET VISUALISATION DES DONNÉES

Alexandra-Maria DOBRESCU

| Cours 8 |

Collection et préparation des données



**SOURCES ET ACQUISITION DES
DONNÉES**



**NETTOYAGE ET
PRÉTRAITEMENT DES DONNÉES**



TYPES DE DONNÉES



LA MESURE



Discrétisation des données

CONTEXTE

LA DISCRÉTISATION DES DONNÉES EST UN PROCESSUS DE PRÉTRAITEMENT DES DONNÉES QUI CONSISTE À CONVERTIR DES DONNÉES CONTINUES EN INTERVALLES OU CATÉGORIES DISCRETS.

Remarque 1: Cette opération est souvent effectuée pour simplifier les données, réduire le bruit et les rendre plus faciles à analyser ou à modéliser.

Remarque 2: La discrétisation est généralement appliquée à des variables continues, les transformant en variables catégorielles.

Recommandation : Même pour les attributs discrets, il est préférable d'avoir un nombre réduit de valeurs conduisant à une représentation réduite des données. Ceci peut être réalisé par des hiérarchies de concepts.

Discrétisation

MÉTHODE

LA DISCRÉTISATION CONSISTE À RÉDUIRE LE NOMBRE DE VALEURS D'UN ATTRIBUT CONTINU DONNÉ EN DIVISANT SES VALEURS EN INTERVALLES.

Remarque 1: Chaque intervalle est étiqueté et chaque valeur d'attribut sera remplacée par l'étiquette de l'intervalle.

Parmi les méthodes les plus populaires pour effectuer une discrétisation sont les suivantes:

- **Binning:** des bins de largeur égale (Equal Width Binning) ou des bins de fréquence égale (Equal Frequency Binning) peuvent être utilisés.

Discrétisation des données

BINNING

Equal Width Binning

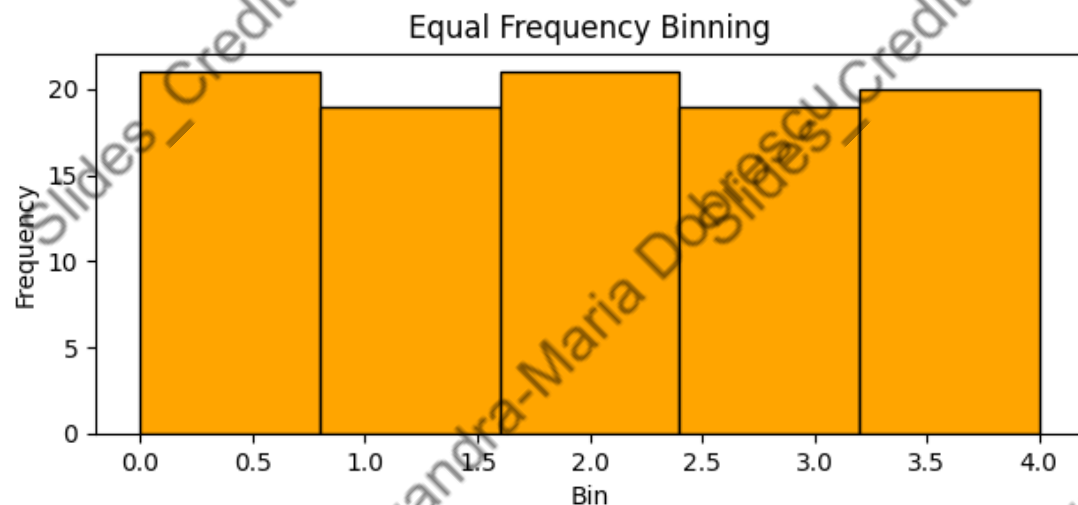
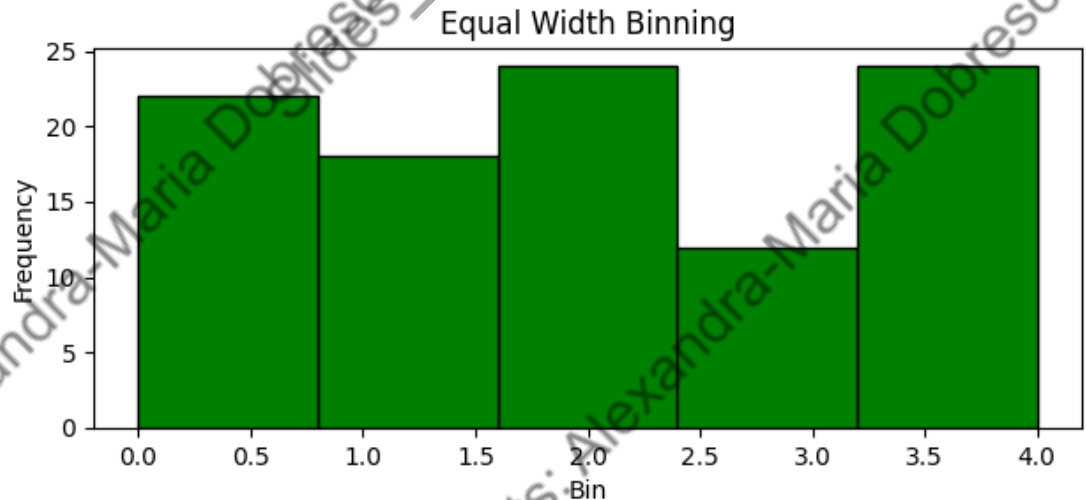
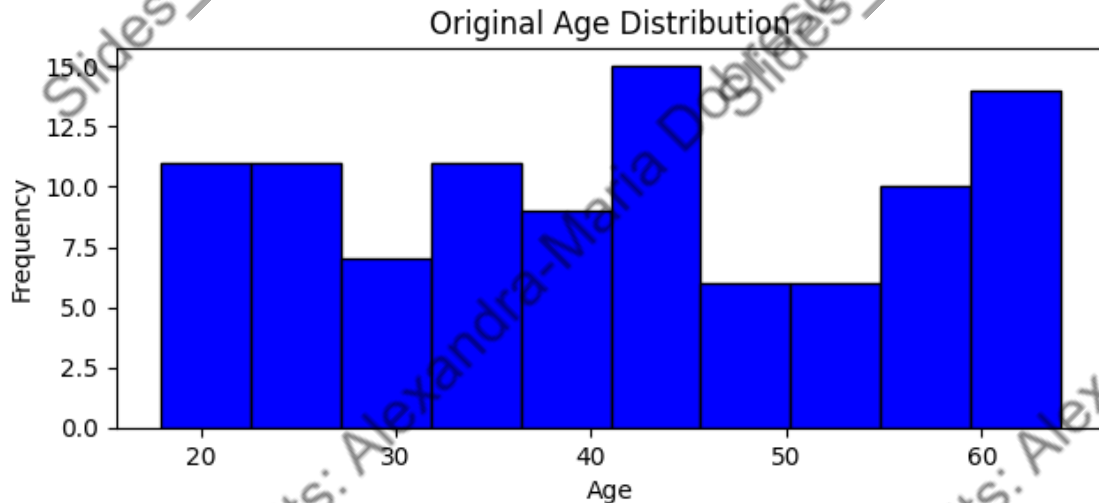
- Diviser l'intervalle de la variable continue en un nombre fixe de cellules de largeur égale.
- La largeur de chaque cellule est déterminée par l'étendue de la variable divisée par le nombre de cellules.
- Par exemple, si les valeurs vont de 0 à 100 et que nous voulons 10 bins, chaque bin aura une largeur de 10.

Equal Frequency Binning

- Diviser les données en intervalles de manière à ce que chaque intervalle contienne approximativement le même nombre de points de données.
- Cela peut aider à gérer les distributions asymétriques et les valeurs aberrantes.
- Par exemple, si nous disposons de 100 observations et que nous voulons 10 bins, chaque case contiendra 10 observations.

Discrétisation des données

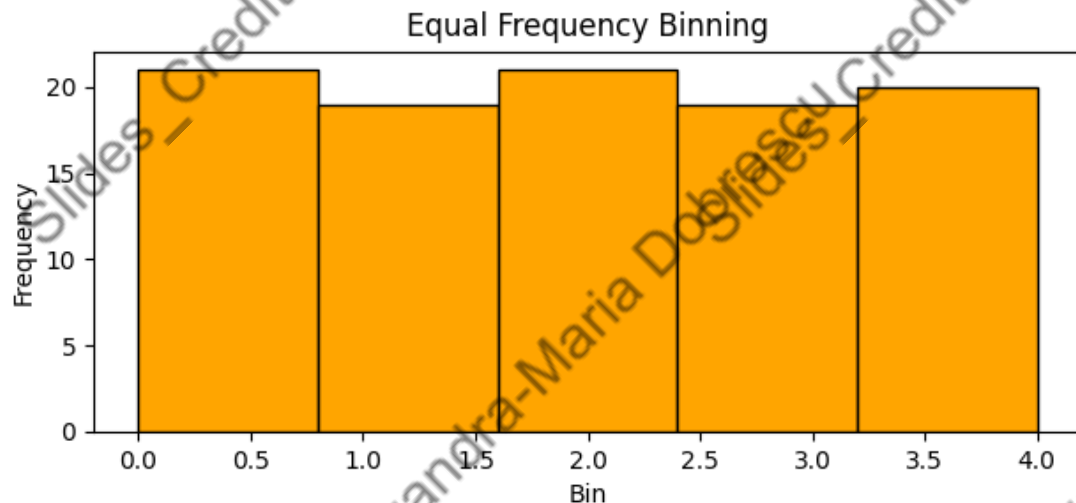
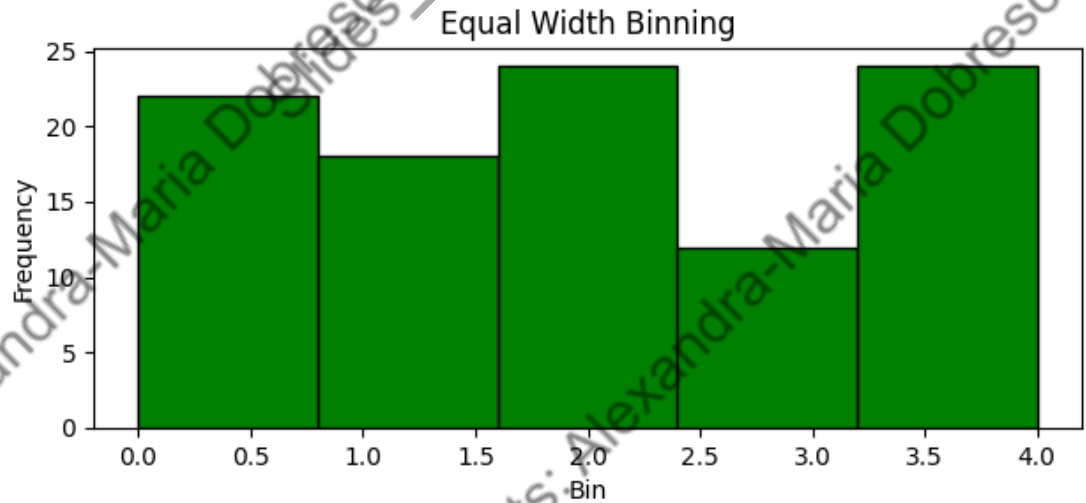
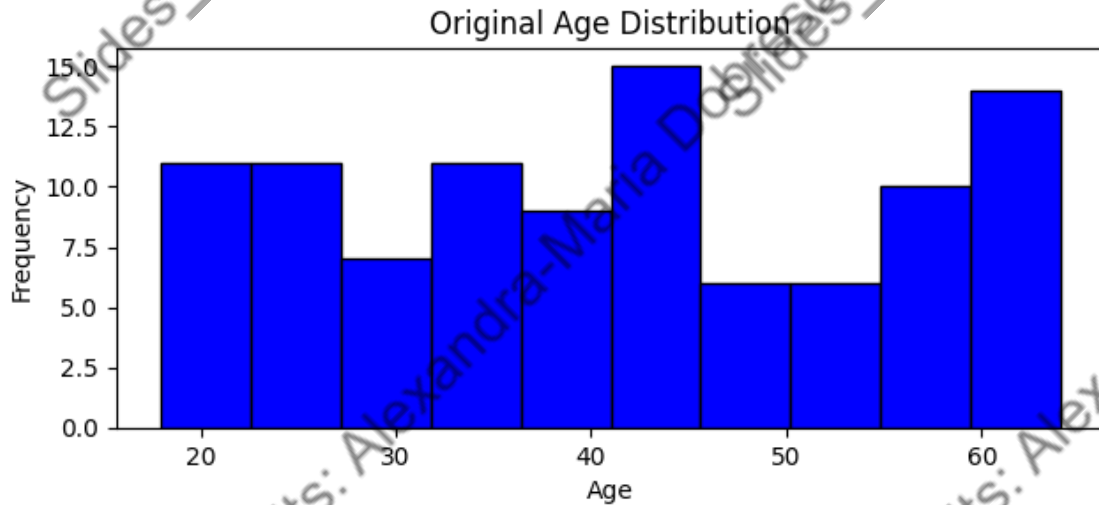
BINNING



Un ensemble de données d'âges sur lequel nous effectuons des binages de largeur égale et des fréquences égales.

Discrétisation des données

BINNING



Pouvez-vous les comparer?
(avantages us inconvénients)

Discrétisation des données

BINNING

Equal Width Binning

- L'avantage de cette méthode est qu'elle est facile à mettre en œuvre et à interpréter, et qu'elle préserve la distribution des données.
- L'inconvénient est qu'elle peut créer des cases vides ou peu nombreuses, en particulier si les données sont asymétriques ou présentent des valeurs aberrantes. Cela peut réduire le contenu en informations et la précision de l'analyse.

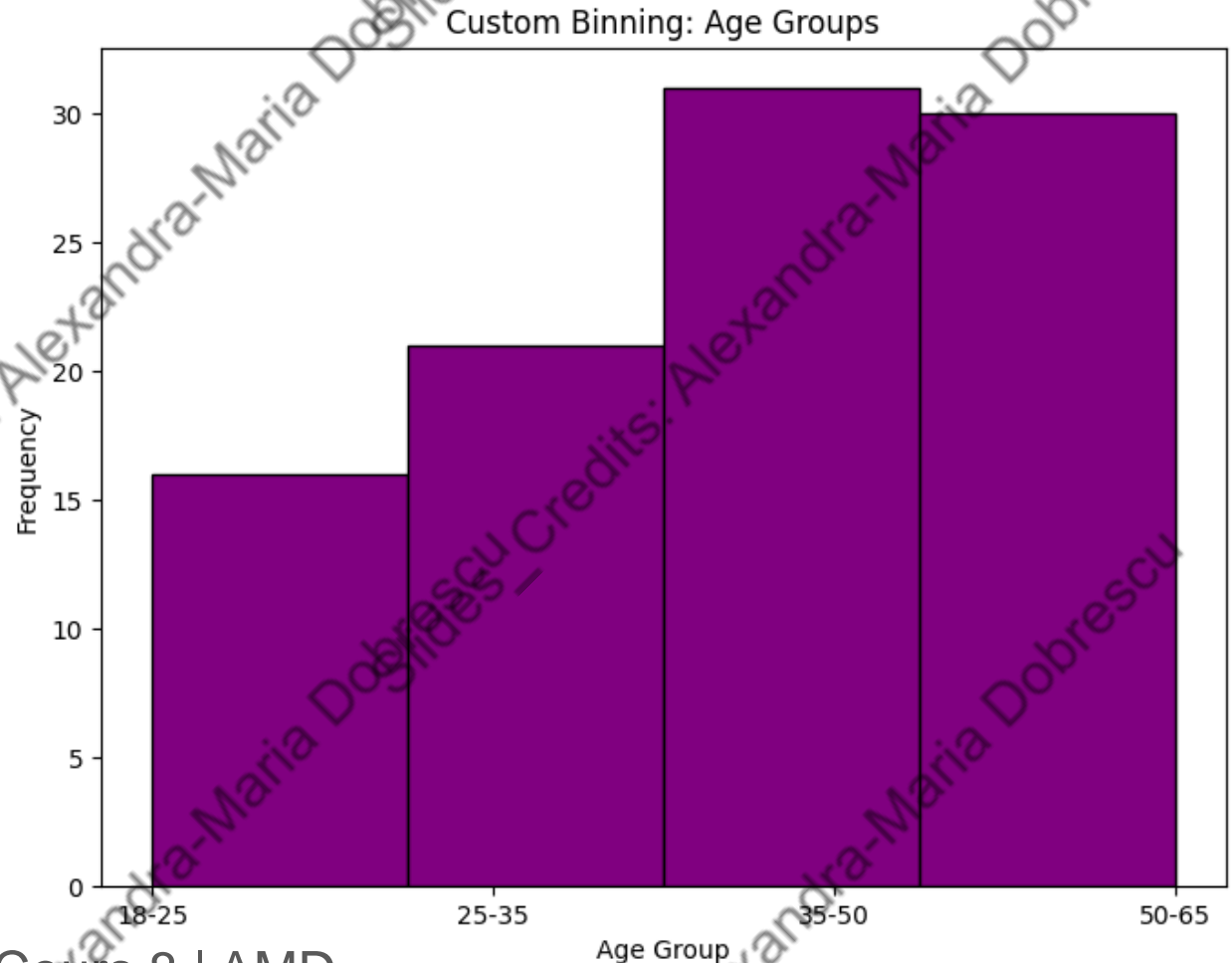
Equal Frequency Binning

- L'avantage de cette méthode est qu'elle crée des unités de mesure équilibrées qui permettent de mieux gérer les données asymétriques et les valeurs aberrantes.
- L'inconvénient est qu'elle peut fausser la distribution des données et créer des largeurs de cellules irrégulières. Cela peut rendre l'analyse plus complexe et moins intuitive.

Discrétisation des données

BINNING PERSONNALISÉ

- La définition de limites d'emplacements personnalisées est basée sur la connaissance du domaine ou sur des exigences spécifiques.
- Cette méthode permet de choisir avec souplesse les limites de l'emplacement en fonction des caractéristiques des données.



Discrétisation des données

COMMENT CHOISIR UNE MÉTHODE DE BINNING

- En ce qui concerne les méthodes de regroupement, il n'existe pas de solution unique. Elle dépend des caractéristiques et des objectifs des données et de l'analyse. Parmi les facteurs à prendre en compte figurent la forme et la répartition des données, l'objectif et le niveau de détail de l'analyse, ainsi que le nombre et la taille des binettes.
- Si les données sont symétriques et uniformément réparties, il peut être préférable d'utiliser des intervalles de largeur égale (equal-width binning) ;
- Si les données sont asymétriques ou présentent des valeurs aberrantes, il peut être préférable d'utiliser des intervalles de fréquence égale (frequency binning).
- Si l'objectif est d'explorer des modèles et des tendances générales, les cellules de largeur égale peuvent être plus informatives (equal-width binning) ;
- Si l'identification de groupes ou de segments spécifiques des données est nécessaire, les cellules de fréquence égale peuvent être plus utiles (frequency binning).

Discrétisation des données

COMMENT CHOISIR UNE MÉTHODE DE BINNING

- Le nombre et la taille optimaux des bins dépendent d'un compromis entre la perte et le gain d'informations; des bins trop peu nombreux ou trop grands peuvent simplifier à l'excès les données, tandis que des bins trop nombreux ou trop petits peuvent entraîner un surajustement.
- En général, il est recommandé d'utiliser la racine carrée du nombre d'observations comme nombre de bins, mais cela peut varier en fonction des données et de l'analyse.

Discrétisation des données

EXTRAS - TRAITEMENT DES DONNÉES D'IMAGE

Chaque pixel a sa propre valeur de luminosité.

2	3	2	2	3	2
2	3	5	5	3	2
2	3	6	6	2	1
2	3	6	6	2	1
3	8	8	6	4	2
3	6	5	5	5	5

10	14	10
10	24	6
20	24	16

4 pixels sont regroupés, leurs valeurs individuelles sont combinées

Le binning consiste généralement à diviser l'image en une grille de cellules de taille égale, puis à calculer la moyenne des valeurs des pixels dans chaque cellule afin d'obtenir une nouvelle image à plus faible résolution.

Discrétisation des données

EXTRAS - TRAITEMENT DES DONNÉES D'IMAGE

Chaque pixel a sa propre valeur de luminosité.

2	3	2	2	3	2
2	3	5	5	3	2
2	3	6	6	2	1
2	3	6	6	2	1
3	8	8	6	4	2
3	6	5	5	5	5

10	14	10
10	24	6
20	24	16

4 pixels sont regroupés, leurs valeurs individuelles sont combinées

La valeur originale est remplacée par les valeurs combinées. Toutefois, chaque pixel représente la valeur moyenne d'une zone plus large, ce qui permet de réduire le bruit et d'améliorer la qualité globale de l'image.

Discrétisation

MÉTHODE

LA DISCRÉTISATION CONSISTE À RÉDUIRE LE NOMBRE DE VALEURS D'UN ATTRIBUT CONTINU DONNÉ EN DIVISANT SES VALEURS EN INTERVALLES.

Parmi les méthodes les plus populaires pour effectuer une discrétisation sont les suivantes:

- **Histogrammes** : comme le binning, les histogrammes répartissent les valeurs d'un attribut dans des catégories. Chaque unité a une étiquette différente et les étiquettes remplacent les valeurs.
- **Intervalles basés sur l'entropie** : chaque valeur d'attribut est considérée comme un point de séparation potentiel (entre deux intervalles) et un gain d'information est calculé pour cette valeur (réduction de l'entropie par la séparation en ce point). La valeur ayant le plus grand gain d'information est ensuite choisie. De cette manière, les intervalles peuvent être construits de manière descendante.

Discrétisation

MÉTHODE

Parmi les méthodes les plus populaires pour effectuer une discrétisation sont les suivantes:

- **Analyse de grappes (cluster)** : après le regroupement, toutes les valeurs d'une même grappe sont remplacées par la même étiquette (l'identifiant du cluster, par exemple).

Rappel: Regroupement par K-Means : Appliquer la méthode des k-moyens pour regrouper les valeurs similaires, et chaque groupe devient un intervalle discret.

- **Méthodes basées sur les arbres de décision** : Les arbres de décision peuvent être utilisés non seulement pour trouver des points de coupure, mais aussi dans le cadre de méthodes d'ensemble telles que la forêt aléatoire pour effectuer une discrétisation.

Discrétisation

MÉTHODE

Parmi les méthodes les plus populaires pour effectuer une discrétisation sont les suivantes:

- **Discrétisation basée sur le chi-carré :** Utiliser des tests statistiques tels que le test du chi-carré pour identifier les points de rupture significatifs pour la répartition en binômes.
- **Discrétisation descendante (MDLP) :** Le principe de la longueur de description minimale (*Minimum Description Length Principle* - MDLP) est une approche descendante qui trouve de manière récursive les points de séparation optimaux sur la base de la longueur de code minimale requise pour représenter les données.

Discrétisation

HIÉRARCHIE DE CONCEPTS

- L'utilisation d'une hiérarchie de concepts pour effectuer discrétisation consiste à remplacer les concepts (ou valeurs) de bas niveau par des concepts de plus haut niveau.

Remarque 1: Pour les valeurs numériques, la discrétisation et les hiérarchies de concepts sont les mêmes.

Exemple : remplacez la valeur numérique de l'âge par jeune, d'âge moyen ou vieux.

Remarque 2: Pour les données catégorielles, l'objectif est de remplacer un grand ensemble de valeurs par un plus petit (les données catégorielles sont discrètes par définition).

Discrétisation

HIÉRARCHIE DE CONCEPTS

LES CONCEPTS HIÉRARCHIQUES NE SONT PAS INTRINSÈQUEMENT UNE MÉTHODE DE DISCRÉTISATION DES DONNÉES, MAIS PLUTÔT UN TERME PLUS LARGE QUI PEUT ENGLOBER DIVERSES STRUCTURES ET RELATIONS HIÉRARCHIQUES AU SEIN DES DONNÉES OU DES INFORMATIONS.

Cependant, lorsqu'il s'agit de discrétisation, les concepts hiérarchiques peuvent être considérés dans le sens de la création des bins imbriqués ou hiérarchiques.

Enfants (0-12 ans)

Nourrissons (0-2 ans)

Tout-petits (3-5 ans)

Enfants (6-12 ans)

Adolescents (13-18 ans)

Adultes (19-65 ans)

Jeunes adultes (19-30 ans)

Adultes d'âge moyen (31-50 ans)

Adultes plus âgés (51-65 ans)

Personnes âgées (65 ans et plus)

Discrétisation

HIÉRARCHIE DE CONCEPTS

Remarque 3: Dans ce cas, on a une organisation hiérarchique dans laquelle les catégories de niveau supérieur (par exemple, « Enfants », « Adultes », « Personnes âgées ») englobent des sous-catégories de niveau inférieur (par exemple, « Nourrissons », « Tout-petits », etc.).

Remarque 4: La création de bins hiérarchiques peut impliquer des bins personnalisés avec des structures imbriquées ou l'utilisation de méthodes spécialisées qui prennent en charge les modèles hiérarchiques dans les données.

Exemple : Si les modèles hiérarchiques dans les données peuvent être complexes, en particulier lorsqu'il s'agit de discrétisation, il existe quelques méthodes ou approches spécialisées qui peuvent être envisagées : Arbres de décision et forêts aléatoires, MDLP (principe de la longueur de description minimale), ChiMerge avec clustering hiérarchique.

Discrétisation

HIÉRARCHIE DE CONCEPTS

Arbres de décision et forêts aléatoires :

Les arbres de décision capturent naturellement les modèles hiérarchiques et les forêts aléatoires, en tant qu'ensembles d'arbres de décision, peuvent être efficaces pour traiter des relations hiérarchiques complexes.

Le MDLP est une approche descendante qui trouve de manière récursive les points de séparation optimaux sur la base de la longueur de code minimale requise pour représenter les données, ce qui permet de capturer les structures hiérarchiques.

La combinaison de **ChiMerge**, un algorithme de regroupement des données utilisant des tests statistiques pour fusionner les cellules adjacentes, avec des techniques de regroupement hiérarchique améliore sa capacité à capturer des modèles hiérarchiques dans les données.

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 ordinales
 - Données d'intervalle
 - Données de rapport

Types de données

CATÉGORIELLES

Idée 1: Les données catégorielles, également connues sous le nom de données qualitatives ou nominales, représentent des caractéristiques ou des attributs qui peuvent être divisés en catégories distinctes.

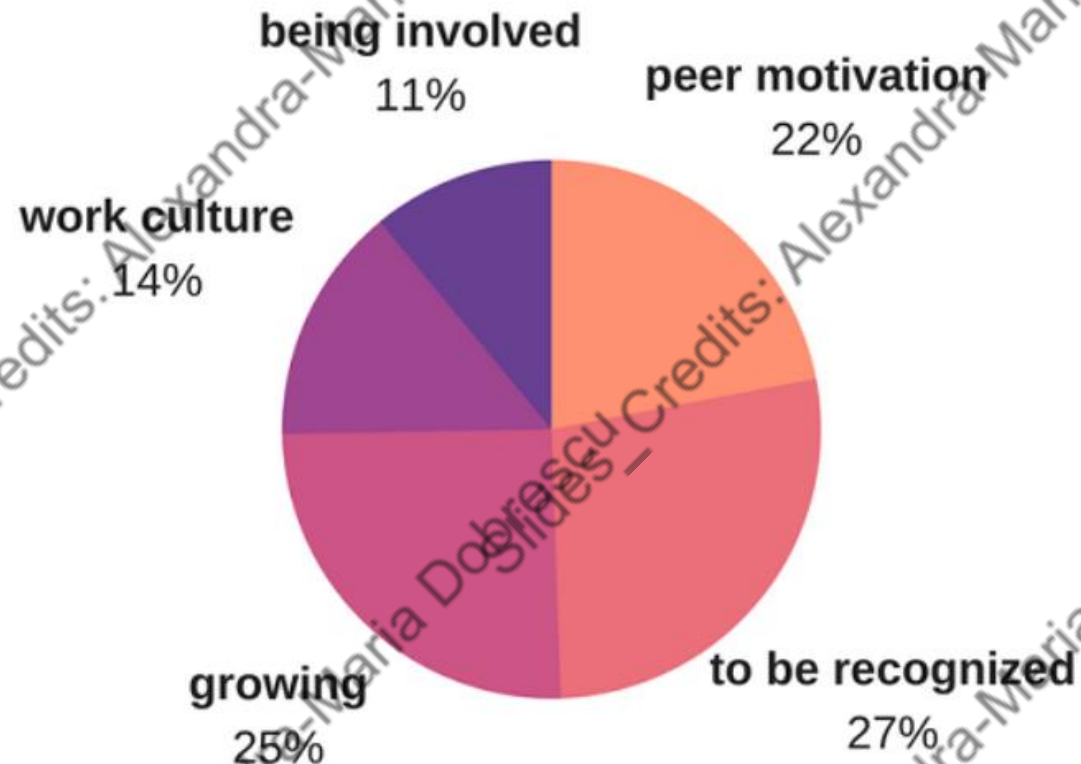
Idée 2: Les données catégorielles sont souvent représentées par des étiquettes ou des noms et ne peuvent être ordonnées ou mesurées numériquement.

Exemple : Parmi les exemples de données catégorielles, on peut citer le sexe (homme ou femme), les couleurs (rouge, bleu, vert) et les niveaux d'éducation (lycée, université, diplômé).

Types de données

CATÉGORIELLES

Exemples de données graphiques et catégorielles :
Enquête sur « Qu'est-ce qui motive les employés à mieux travailler ? »



Types de données

CATÉGORIELLES

Exemples de données graphiques et catégorielles :
Enquête sur « Qu'est-ce qui motive **mes étudiants** à mieux travailler ? »



?

Types de données

CATÉGORIELLES

Idée 3: L'analyse des données catégorielles comprend très souvent des tableaux de données. Les valeurs sont représentées sous la forme d'un tableau à double entrée ou d'un tableau de contingence en comptant le nombre d'éléments appartenant à chaque catégorie.

Eye Color					
Hair Color	Green	Blue	Brown	Black	Total
Blonde	4	7	2	1	14
Brown	2	4	18	2	26
Black	1	2	5	2	10
Total	7	13	25	5	50

Types de données

CATÉGORIELLES

Idée 4: Le tableau montre les résultats des groupes formés en comptant la couleur des cheveux et des yeux de chaque personne.

Les tableaux à double entrée et les tableaux de contingence sont d'excellents outils pour voir comment deux variables catégorielles sont liées.

Le tableau représente les nombres ou les pourcentages de personnes appartenant à un groupe pour deux variables quantitatives ou plus. Il facilite la recherche de différentes relations entre les données.

Discussion: Que retirez-vous de cette analyse du tableau précédent ?

Types de données

CATÉGORIELLES

Les principales caractéristiques des données catégorielles:

- Les données catégorielles sont divisées en groupes ou en catégories.
- Les catégories sont basées sur des caractéristiques qualitatives.
- Les valeurs et les variables catégorielles ne sont pas ordonnées.
- Les données catégorielles peuvent prendre des valeurs numériques, mais ces nombres n'ont pas de signification mathématique.
- Les données catégorielles sont représentées graphiquement par des diagrammes à barres et des diagrammes circulaires.

Types de données

NUMÉRIQUES

Idée 1: Les données numériques consistent en des nombres provenant d'un ensemble de valeurs continues ou discrètes.

Idée 2: Les valeurs sont ordonnées, de sorte qu'il est possible de tester cet ordre (<, >, !=, ...). Parfois, nous devons ou pouvons convertir des données catégorielles en données numériques en leur attribuant une valeur numérique (ou un code) à chaque étiquette.

Les principales caractéristiques des données numériques:

- Elles peuvent être quantifiées et vérifiées.
- Les données peuvent être comptées.
- Type de données: nombre et statistiques.
- Elles répondent à des questions telles que « combien ? », « à quelle fréquence ? » et « combien de fois ? ».

df_numerical_scale										
	Income	Recency	MntWines	MntFruits	MntFish	MntSweet	NumPurchases	Year_Birth	Kidhome	Teenhome
0	0.084832	0.585859	0.425318	0.442211	0.664093	0.334601	0.200000	-0.985125	-0.825033	-0.929687
1	0.067095	0.383838	0.007368	0.005025	0.007722	0.003802	0.133333	-1.235457	1.032328	0.906732
2	0.105097	0.262626	0.285332	0.246231	0.428571	0.079848	0.066667	-0.317572	-0.825033	-0.929687
3	0.037471	0.262626	0.007368	0.020101	0.038610	0.011407	0.133333	1.267866	1.032328	-0.929687
4	0.085065	0.949495	0.115874	0.216080	0.177606	0.102662	0.333333	1.017534	1.032328	-0.929687
...
2235	0.089472	0.464646	0.474883	0.216080	0.162162	0.448669	0.133333	-0.150684	-0.825033	0.906732
2236	0.093669	0.565657	0.271936	0.000000	0.000000	0.000000	0.466667	-1.903010	2.889690	0.906732
2237	0.083092	0.919192	0.608171	0.241206	0.123552	0.045627	0.066667	1.017534	-0.825033	-0.929687
2238	0.101536	0.080808	0.286671	0.150754	0.308880	0.114068	0.133333	-1.068569	-0.825033	0.906732
2239	0.076908	0.404040	0.056263	0.015075	0.007722	0.003802	0.200000	-1.235457	1.032328	0.906732

2240 rows x 10 columns

TYPES OF DATA

Qualitative Data

Quantitative Data

Names as John,
Patricia, Mary...



Smell e.g. aromatic,
buttery,....



Colors e.g. green,
white, blue....



Scores on exams
e.g. 85, 67, 90 ...





The weight of a
person or a subject



Your shoe size



Basis for Comparison	 Qualitative Data	 Quantitative Data
Definition	Qualitative data is information that can't be expressed as a number	Quantitative data is data that can be expressed as a number or can be quantified
Can data be counted?	NO	YES
Data type	Words, objects, pictures, observations, and symbols	Number and statistics

Questions that data answer	How and why this has happened?	“how many, “how much” and “how often”
Examples	<ul style="list-style-type: none"> Names as John, Maria, ... Ethnicity such as American Indian, Asian, etc. Colors e.g. green, white, blue 	<ul style="list-style-type: none"> Scores on tests and exams e.g. 85, 67, 90 and etc. The weight of a person or a subject Your shoe size
Purposes of data analysis	Understand, explain, and interpret social interactions and patterns	Test hypothesis, develop predictions for the future, check cause and effect

Types of data analysis	Patterns, characteristics, theme identification	Statistical relationship identification
Scope of the results	Less generalizable, particular findings. Do not drive conclusions and generalizations across a population	Generalizable findings. Draw conclusions and trends about a large population based on a sample taken from it
Popular methods of data analysis	<ul style="list-style-type: none"> • Content analysis • Thematic analysis • Discourse analysis • Grounded theory • Conversation analysis 	<ul style="list-style-type: none"> • Linear regression models • Logistic regression • Analysis of Variance (ANOVA) • Statistical significance • Correlation analysis • Central tendency • Dispersion • Distribution

Source [3]

Types de données

LES NIVEAUX DE MESURE DES DONNÉES / LES TYPES D'ÉCHELLE

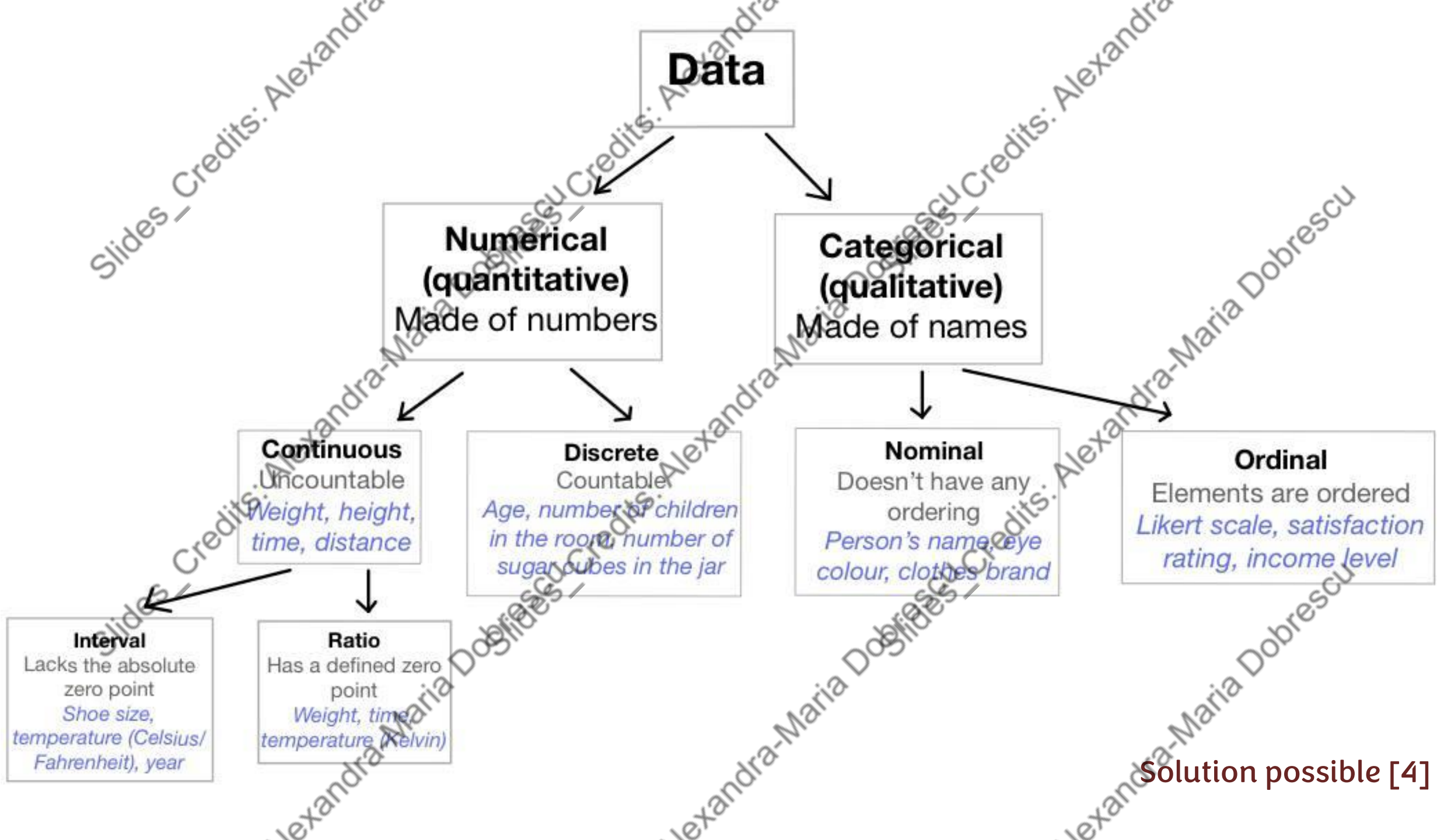
Idée 1: La compréhension du niveau de mesure de vos variables est une capacité vitale lorsque vous travaillez dans le domaine des données.

Idée 2: En d'autres termes, les manières d'étiqueter les données sont connues sous le nom « d'échelles ».

En fait, il existe quatre échelles de mesure: **nominale**, **ordinaire**, **d'intervalle** et de **rapport**.

Il s'agit simplement de méthodes permettant de catégoriser différents types de variables.

Exercice: Construisez une représentation graphique dans laquelle vous jouez avec les concepts hiérarchiques suivants: **Données**, **Données numériques**, **catégorielles**, **Continu**, **Discret**, **Intervalle**, **Nominal**, **Ordinal**, **Rapport**.
Donnez des exemples.



Solution possible [4]

Bibliographie

- [1] Subramaniam, A. (2020). What Is Big Data Analytics| Big Data Analytics Tools and Trends| Edureka.
- [2] <https://www.linkedin.com/advice/1/what-advantages-disadvantages-equal-width>
- [3] <https://www.intellspot.com/categorical-data-examples/>
- [4] <https://ucarecdn.com/2bc4eb6c-4c71-4679-8c0b-308b293b8515/>