📊 Analyse Exploratoire des Données - Fashion Store Sales

1. Structure du fichier

Colonnes et types :

Nombre total de colonnes :19

Nombre total de lignes : 2253

Colonne	Type estimé	Description
sale_date	String	Date de vente
item_id	Integer	Identifiant unique de l'article vendu
sale_id	Integer	Identifiant de la vente
product_id	Integer	Identifiant du produit
quantity	Integer	Quantité vendue
original_price	Float	Prix original
unit_price	Float	Prix unitaire après remise
discount_applied	Float	Montant de la remise
discount_percent	String	Pourcentage de remise (format XX.XX%)
discounted	Integer	Indicateur binaire (0/1)
item_total	Float	Total pour l'article
channel	String	Canal de vente
channel_campaigns	String	Campagne marketing
total_amount	Float	Montant total de la transaction
product_name	String	Nom du produit
category	String	Catégorie produit
brand	String	Marque
color	String	Couleur
size	String	Taille
catalog_price	Float	Prix catalogue

cost_price	Float	Prix de revient
customer_id	Integer	ID client
gender	String	Genre
age_range	String	Tranche d'âge
signup_date	Date	Date d'inscription
first_name	String	Prénom
last_name	String	Nom
email	String	Email
country	String	Pays

2. Cardinalité des colonnes principales

Données catégorielles :

category : 8 valeurs uniques

- Dresses (21.393697 %)
- Shoes (20.949845 %)
- Pants (15.623613 %)
- T-Shirts (21.837550 %)
- Sleepwear (20.195295 %)

brand : 1 valeur unique

- Tiva (100%) → Marque exclusive

gender : 1 valeur unique

- Female (100%) → Boutique féminine

country : 6 pays

- Germany (23.834887 %)
- France (22.103862 %)
- Italy (18.419885 %)
- Portugal (8.921438 %)

- Netherlands (14.469596 %)
- Spain (12.250333 %)

channel : 2 valeurs
- App Mobile (51.930759 %)
- E-commerce (48.069241 %)

age_range : 5 tranches
- 16-25 (20.195295 %)
- 36-45 (21.171771%)
- 46-55 (19.218819 %)
- 56-65 (18.863737 %)
- 26-35 (20.550377 %)

3. Entités métier principales identifiées

ENTITÉ 1 : Produit

- **Clé :** *product_id*
- **Attributs :** product_name, category, brand, color, size, catalog_price, cost_price
- **Relations :** Apparaît dans plusieurs ventes

ENTITÉ 2 : Client

- **Clé :** *customer_id*
- **Attributs :** first_name, last_name, email, gender, age_range, country, signup_date
- **Observations :** Certains clients récurrents (même customer_id sur plusieurs lignes)

ENTITÉ 3 : Vente

- **Clé composite :** *sale_id*
- **Attributs :** sale_date, channel, channel_campaigns, total_amount
- **Relations :** Contient plusieurs articles (ventes multi-articles)

ENTITÉ 4 : Transaction

- **Clé :** *item_id* (un par ligne de vente)
- **Attributs :** quantity, original_price, unit_price, discount_applied, item_total
- **Relations :** Liée à une vente et un produit

4. Redondances identifiées

Redondance 1 : Prix tripliqués

original_price, unit_price, catalog_price

Analyse :

- original_price = Prix avant remise
- unit_price = Prix après remise (original_price - discount_applied)
- catalog_price = Prix catalogue (souvent égal à original_price)

Problème : catalog_price semble redondant avec original_price dans 95% des cas

Redondance 2 : Calculs dérivés

discount_percent peut être calculé à partir de discount_applied / original_price

discounted (0/1) peut être déduit de discount_applied > 0

item_total = unit_price × quantity (cohérent)

Redondance 3 : Relations prix

Pour les ventes sans remise :

- original_price = unit_price = catalog_price
- discount_applied = 0
- discount_percent = "0.00%"
- discounted = 0

5. Anomalies potentielles

Anomalie 1 : Valeurs manquantes

first_name : 5.14 % manquants

last_name : 2.84 % manquants

email : 9.94 % manquants

total_amount : 9.98 % manquants

Anomalie 2 : Incohérences temporelles

signup_date POSTERIEURE à sale_date dans 8.21% des cas

Ex: Vente 2025-04-06 mais inscription 2025-05-21

Anomalie 3 : Discounts incohérents

Cas où discount_percent ≠ (discount_applied / original_price) × 100

Ex: original_price=66.87, discount_applied=6.69, discount_percent="10.00%"

66.87 × 0.10 = 6.687 (arrondi à 6.69) → OK

Mais d'autres cas montrent des écarts dus aux arrondis

6. Relations métier-clés

7. Modélisation

Normalisation :

Dim_channel (

channel_id ,

channel_name

)

Dim_campaign (

campaign_id ,

#channel_id ,

campaign_name,

);

Dim_category (

category_id,

category_name

);

Dim_brand (

brand_id,

brand_name

);

Dim_color (

color_id,

color_name

);

Dim_size (
size_id,
size_label
);

Dim_country (
country_id,
country_name
);

Dim_customer (
customer_id ,
first_name,
last_name,
email,
gender,
age_range,
signup_date,
#country_id
);

Dim_product (
product_id,
product_name,

#category_id,
#brand_id,
#color_id,
#size_id ,
cost_price,
original_price
)

Fact_sale (
sale_id ,
sale_date,
total_amount,
#customer_id ,
#campaign_id
)

Fact_sale_item (
item_id,
#sale_id ,
#product_id ,
quantity,
discount_percent
)