



RAPPORT TECHNIQUE : COMMERCE EN LIGNE

**Modélisation, intégration et
visualisation d'un Système
d'Information Décisionnel**



Yamina BENFERLOU



Sommaire

Table des matières

Sommaire	1
Introduction	2
Présentation des données	3-6
Modélisation du SID	6-9
• Choix du processus métier à modéliser	6
• Définition de la granularité	6-7
• Choix des dimensions	7-8
• Identification des faits	8-9
Intégration des données (Processus ETL)	10-16
• Extraction des données	10
• Transformation des données	11-13
• Chargement des données	13-16
Restitution des données	17-19
Conclusion	20

Introduction

Contexte :

Le commerce en ligne, notamment au Brésil, est un secteur en pleine expansion, où la concurrence entre les plateformes est de plus en plus féroce. Dans ce contexte, comprendre les facteurs qui influencent les comportements des consommateurs devient essentiel pour les entreprises souhaitant maintenir une relation durable avec leurs clients.

C'est dans le cadre du projet "**One Day in Brazil**", visant à créer un tableau de bord illustrant l'activité journalière dans le secteur du commerce électronique brésilien, que nous avons choisi de travailler sur les données issues de [Olist Store](https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce). Cette plateforme, acteur majeur du e-commerce au Brésil, connecte des vendeurs et des acheteurs, permettant aux petites et moyennes entreprises de vendre leurs produits via divers canaux en ligne.

De fait, pour cette étude, nous avons sélectionné **cinq bases de données** parmi les huit disponibles sur Kaggle :

<https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>

Ces bases fournissent des informations détaillées sur les profils des acheteurs, les commandes passées, les articles commandés, et les paiements effectués. Ce jeu de données permet une analyse approfondie des comportements d'achat, des préférences de paiement et des interactions des clients avec les produits.

Problématique

Ainsi, la question à laquelle nous nous intéresserons est la suivante :

Quelles sont les habitudes de consommation des clients sur les plateformes de commerce électronique au Brésil, et comment varient-elles en fonction des produits et des modes de paiement ?

Dans ce rapport, nous nous concentrerons uniquement sur **la modélisation du système d'information décisionnel (SID)**, première étape clé du projet. Cette étape permettra de structurer et d'organiser les données sélectionnées pour les rendre exploitables dans le cadre du tableau de bord "One Day in Brazil ».

Le rapport final sera disponible sur github, une plateforme web collaborative, en se rendant sur le lien suivant :

<https://github.com/arthursim0n/projet>

Présentation des données

Les bases de données que nous avons sélectionnées font partie d'un ensemble de données publiques de commerce électronique brésilien sur les commandes effectuées sur Olist Store. L'ensemble de données contient des informations sur 100 000 commandes de 2016 à 2018 effectuées sur plusieurs marchés au Brésil. Ses fonctionnalités permettent de visualiser une commande sous plusieurs angles : du statut de la commande, du prix, du paiement et des performances de fret à la localisation du client, aux attributs du produit et enfin aux avis rédigés par les clients.

Ici, nous nous sommes focalisés sur les 5 bases de données qui semble apporter le plus d'informations à savoir : les clients, les commandes, les paiements, les avis et les produits.

- **Base de données des clients : olist_customers_dataset.csv**

Cette base de données contient des informations relatives aux clients de la plateforme. Elle permet d'identifier chaque client à l'aide d'un identifiant unique et fournit des données géographiques telles que le code postal, la ville et l'état où le client est situé. Ces informations sont utiles pour des analyses démographiques, géographiques ou pour mieux comprendre la répartition des clients.

Les différentes variables présentes dans ce jeu de données sont les suivantes :

customer_id : Identifiant unique de chaque commande client.

customer_unique_id : Identifiant unique du client, qui peut être commun à plusieurs commandes.

customer_zip_code_prefix : Préfixe du code postal du client.

customer_city : Ville du client.

customer_state : État du client.

price : Prix de l'article.

freight_value : Coût du fret.

- **Base de données des commandes : olist_order_dataset.csv**

Cette base de données contient des informations sur les commandes passées sur la plateforme entre 2016 et 2018. Ces informations permettent d'analyser le parcours complet d'une commande, depuis l'achat jusqu'à la livraison, en passant par les différentes étapes de traitement et de transport. Elles sont essentielles pour évaluer les performances logistiques, la satisfaction client et identifier les éventuels points d'amélioration dans le processus de commande.

Ainsi, les variables présentes dans ce jeu de données sont les suivantes :

order_id : Identifiant unique de la commande.

customer_id : Identifiant unique du client ayant passé la commande.

order_status : Statut actuel de la commande (par exemple, livrée, expédiée, en attente, annulée).

order_purchase_timestamp : Date et heure de l'achat de la commande.

order_approved_at : Date et heure de l'approbation du paiement.

order_delivered_carrier_date : Date à laquelle la commande a été remise au transporteur.

order_delivered_customer_date : Date de livraison effective au client.

order_estimated_delivery_date : Date de livraison estimée initialement prévue.

- **Base de données des paiements : olist_order_payments_dataset.csv**

Cette base détaille les paiements associés aux commandes. Elle contient des informations sur les méthodes de paiement utilisées, le montant payé et les éventuelles mensualités. Cela permet d'analyser les habitudes de paiement des clients.

Les différentes variables présentes dans ce jeu de données sont les suivantes :

order_id : Identifiant unique de la commande.

payment_sequential : Séquence des paiements pour une commande.

payment_type : Mode de paiement utilisé (par exemple : carte de crédit, virement bancaire, etc.).

payment_installments : Nombre de mensualités choisies pour le paiement.

payment_value : Montant payé.

- **Base de données des produits : olist_products_dataset.csv**

Cette base de données contient des informations détaillées sur les produits disponibles sur la plateforme. Elle inclut des caractéristiques comme le poids, les dimensions, et le nombre de photos disponibles. Ces données permettent d'analyser l'offre de produits et d'évaluer leur description ou présentation.

Les différentes variables présentes dans ce jeu de données sont les suivantes :

product_id : Identifiant unique du produit.

product_category_name : Nom de la catégorie à laquelle appartient le produit.

product_name_length : Longueur du nom du produit (nombre de caractères).

product_description_lenght : Longueur de la description du produit (nombre de caractères).

product_photos_qty : Nombre de photos associées au produit.

product_weight_g : Poids du produit (en grammes).

product_length_cm : Longueur du produit (en centimètres).

product_height_cm : Hauteur du produit (en centimètres).

product_width_cm : Largeur du produit (en centimètres).

- **Base de données des articles commandés : olist_order_items_dataset.csv**

Cette base de données regroupe des informations détaillées sur les articles inclus dans chaque commande passée sur la plateforme. Elle relie les commandes aux produits, aux vendeurs, ainsi qu'aux informations sur les frais de transport et les délais de livraison (=Transaction).

Les différentes variables présentes dans ce jeu de données sont les suivantes :

order_id : Identifiant unique de la commande.

order_item_id : Identifiant de l'article au sein de la commande.

product_id : Identifiant du produit commandé.

seller_id : Identifiant du vendeur ayant fourni le produit.

shipping_limit_date : Date limite de livraison prévue.

Modélisation du SID

- **Choix du processus métier à modéliser**

Nous avons retenu deux processus métiers clés pour notre étude : **l'analyse des commandes** et **l'analyse des produits**. Ces deux processus sont étroitement liés et offrent une vue complète sur l'activité quotidienne de la plateforme de commerce électronique **Olist**, un acteur majeur du e-commerce au Brésil.

Analyse des commandes :

Ce processus se concentre sur l'étude du cycle de vie des commandes passées sur Olist, depuis leur création jusqu'au paiement et à la livraison. Il vise à comprendre les comportements d'achat des clients, notamment en termes de montants dépensés et de fréquences d'achat. Par ailleurs, l'objectif de ce processus métier est aussi d'obtenir des informations sur les méthodes de paiements les plus utilisées, tout en analysant sa variation en fonction des montants et des profils clients.

Analyse des produits :

Ce processus explore, quant à lui, les caractéristiques des produits vendus sur la plateforme, et leur impact sur les comportements d'achat. Il nous permettra notamment de répondre à nos interrogations concernant les catégories de produits les plus populaires. De plus, avec ce processus métier, on souhaitera analyser l'impact des caractéristiques du produit (poids, dimensions, etc.) sur les décisions d'achat des clients.

Nous avons choisi ces deux processus en raison de leur centralité dans l'activité d'Olist. En effet, ensemble, ils offrent une vue détaillée et complète du commerce électronique au Brésil. En se focalisant sur ces deux processus, nous répondons donc aux questions les plus pertinentes pour les utilisateurs tout en garantissant une structure de données claire et exploitable.

- **Définition de la granularité**

Nous avons choisi une **granularité transactionnelle** pour les deux processus métier. Cela correspond au niveau de détail le plus fin dans l'analyse des données.

En effet, chaque transaction est enregistrée sous forme d'une ligne unique dans les bases de données.

Pour l'analyse des commandes, on a donc une ligne par commande, identifiée par un order_id. Celui-ci inclut des informations détaillées comme les montants payés, les modes de paiement et les dates clés.

En ce qui concerne l'analyse des produits, chaque produit commandé est décrit individuellement, avec un lien entre son product_id et l'identifiant de commande (order_id).

Ainsi, cette granularité permet de regrouper tous les détails nécessaires pour analyser les comportements d'achat, les préférences des clients et les caractéristiques des produits, telles que leur catégorie, leur prix ou leur quantité. De plus, elle garantit la possibilité d'agréger les données pour des analyses synthétiques, par exemple par jour, par semaine ou par catégorie de produit (=période).

En choisissant ce niveau de granularité, nous assurons une flexibilité dans l'exploration des données tout en maintenant une richesse d'information suffisante pour la réalisation de nos analyses.

Désormais, il s'agit pour nous de définir les différentes dimensions.

• Choix des dimensions

Pour structurer notre analyse et comprendre clairement les liens entre les dimensions et les processus métier, nous avons établi un tableau croisé. Ce tableau associe chaque dimension aux processus métier pertinents, en identifiant les tables sources nécessaires pour exploiter ces dimensions.

Processus métier	Table source	Date	Client	Transaction	Produit	Paiement
	Table associée	orders	customers	orders_items	products	order_payments
Analyse des commandes		X	X	X		X
Analyse des produits		X		X	X	

L'analyse des commandes se concentre sur l'étude des transactions effectuées par les clients, en détaillant les données associées aux commandes, aux clients, et aux paiements. Ainsi, les dimensions utilisées pour ce processus sont :

- ⇒ **Date** : Cette dimension est fondamentale. Elle permet de suivre l'évolution des commandes dans le temps, en analysant les périodes (jour, mois, année) où elles sont effectuées. Cette information est extraite de la table orders.
- ⇒ **Client** : Cette table de dimension relie les commandes aux informations spécifiques des clients, notamment leur localisation géographique. Cette dimension s'appuie sur la table customers.
- ⇒ **Transaction** : Cette dimension offre du détail sur les articles inclus dans chaque commande, telles que les quantités et les prix. Cette dimension provient de la table orders_items. Elle est fondamentale car elle contient des mesures très pertinentes pour l'analyse, notamment le prix.
- ⇒ **Paiement** : Enfin, cette dernière dimension de ce processus, permet d'étudier les modes de paiement utilisés par les clients et les montants payés. Les données associées à cette dimension sont issues de la table order_payments.

L'analyse des produits, quant à elle, se concentre sur les caractéristiques des articles vendus et leur performance dans le cadre des commandes. Les dimensions utilisées pour ce processus sont :

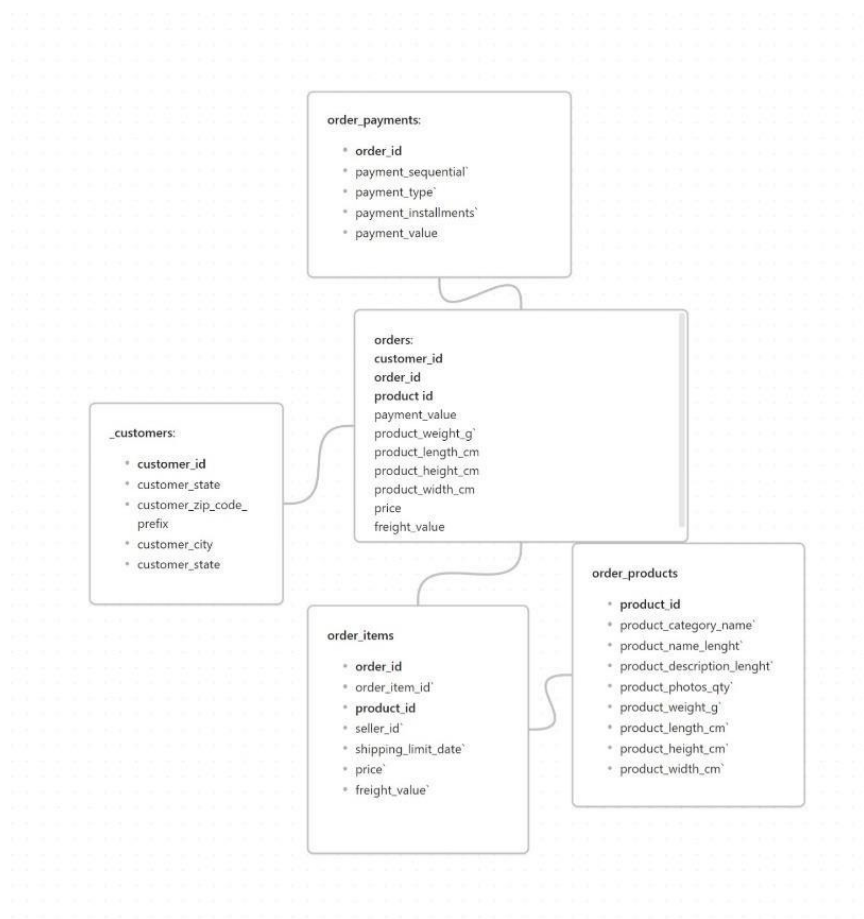
- ⇒ **Date** : Cette dimension, fondamentale car présente dans les deux processus métier, permet de suivre les ventes des produits au fil du temps. Ainsi, on pourra observer les périodes les plus propices pour certaines catégories de produit. Cette information provient de la table orders.
- ⇒ **Transaction** : Pareillement, cette table fournit des détails sur les volumes d'articles achetés et leur association avec les commandes. Cette dimension s'appuie sur la table orders_items.
- ⇒ **Produit** : Essentielle au processus métier, cette dimension nous permet d'obtenir les caractéristiques spécifiques des produits, comme leur catégorie, leur poids, ou leurs dimensions. Ces informations proviennent de la table products.

• Identification des faits

Fait (indicateur)	Description	Relation avec les dimensions	Rôle dans le SID
Montant total payé (payment_value)	Montant total payé pour une commande, incluant produits et frais de port.	Associé à la dimension Paiement pour analyser les types de paiement, Date pour observer les tendances, et Transaction pour comprendre les revenus générés par chaque commande.	Permet de mesurer les revenus générés par commande ou par période et de comprendre l'impact des produits et des modes de paiement sur les ventes.
Frais de port (freight_value)	Montant payé pour la livraison des produits.	Associé à la dimension Paiement pour analyser l'impact des frais de port sur le montant total payé, Produit pour observer l'impact du produit sur les frais d'expédition, Transaction pour relier les frais de livraison aux commandes spécifiques..	Utile pour analyser l'impact des coûts de livraison sur le prix total payé, optimiser les stratégies de livraison, et ajuster les prix en fonction des frais d'expédition.
Poids et dimensions du produit (product_weight, height, width, length)	Poids et dimensions physiques des produits commandés, utiles pour l'analyse logistique.	Associé à la dimension Produit pour observer l'impact du poids et des dimensions sur les coûts de transport et de stockage, et Commande pour	Permet d'analyser les coûts logistiques (transport, stockage) en fonction des caractéristiques physiques des produits, et d'optimiser les

		évaluer l'impact des poids et dimensions sur les frais de livraison.	stratégies de livraison et d'entreposage.
Prix du produit (price)	Prix de chaque produit dans la commande.	Associé à la dimension Produit pour analyser les prix des produits et les comparer entre eux, Commande pour observer la répartition du prix total sur chaque produit, et Transaction pour analyser les revenus générés par produit.	Permet d'analyser la rentabilité des produits, l'impact des prix sur les décisions d'achat des clients, et d'identifier les produits les plus rentables.

Modèle possible pour notre système d'information décisionnel



Ici, nous avons défini les dimensions pertinentes pour l'analyse des commandes et des produits en prenant en compte les processus métier clés de notre projet.

Chaque dimension a été choisie en fonction de son rôle dans l'analyse des comportements des consommateurs et des produits, et nous avons pu expliquer leur relation avec les différentes tables sources. Les faits identifiés, tels que le montant total ou le nombre de produits commandés, seront essentiels pour alimenter les indicateurs de performance et permettre des analyses approfondies sur l'évolution des commandes et des produits.

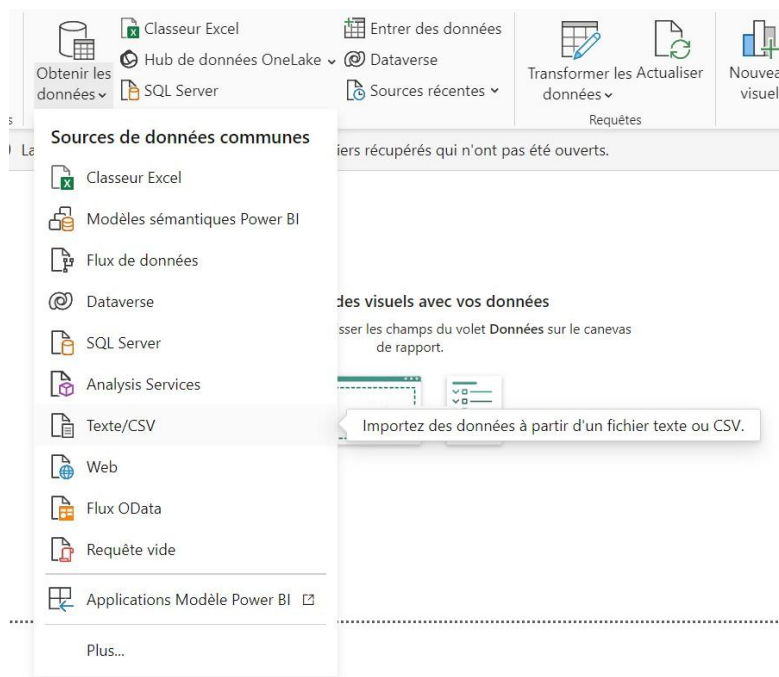
Intégration des données (Processus ETL)

Dans cette étape, nous avons mis en place notre processus ETL (Extraction, Transformation et Chargement) pour intégrer et préparer les données issues de fichiers CSV dans Power BI.

1. Extraction des données

Comme dit plus tôt, nous avons extrait les données à partir de Kaggle.

Une fois ces bases extraites et téléchargées, nous avons commencé l'import dans Power BI en cliquant sur « Obtenir les données ».



Cependant, avant cet import il est important de passer par l'étape de transformation des données

2. Transformation des données

Pour transformer les données extraites et les rendre exploitables, nous avons utilisé l'éditeur Power Query intégré à Power BI, qui permet de nettoyer, structurer et enrichir les données via une interface intuitive. Après avoir chargé les fichiers CSV dans Power BI, nous avons accédé à l'éditeur en sélectionnant l'option "Transformer les données". Cette étape préliminaire nous a permis de visualiser chaque table individuellement et d'identifier les ajustements nécessaires.

olist_orders_dataset.csv

Origine du fichier: 1252: Europe de l'Ouest (Windows) Délimiteur: Virgule Détection du type de données: Selon les 200 premières lignes

order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	ord
e481f51c6d54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	02/10/2017 10:56:33	02/10/2017 11:07:15	
53c9db2f8b7c7dce06741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	24/07/2018 20:41:37	26/07/2018 03:24:27	
47770eb9100c2d0c44946d9c07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	08/08/2018 08:38:49	08/08/2018 08:55:23	
949d5b44dbf5de018ef9c1697b7945f8a	f88197465ea7920adcbec7375364a82	delivered	18/11/2017 19:28:06	18/11/2017 19:45:59	
ad21c59c0840e6cb83a9ceb5573f8159	8ab97904e6daea8866dbdc4fb7aad2c	delivered	13/02/2018 21:18:39	13/02/2018 22:20:29	
a4591c265e18cb1dce52889e2d8acc3	503740e9ca751ccdda7ba28e5ab8f608	delivered	09/07/2017 21:57:05	09/07/2017 22:10:13	
136cce7aa42f4b2cefd53fde79a6098	ed0271e0b7da060a393796590e7b737a	invoiced	11/04/2017 12:22:08	13/04/2017 13:25:17	
6514b8d028c9f2cc2374ded245783f	9bdf08b4b3b52b5526ff42d37d47f222	delivered	16/05/2017 13:10:30	16/05/2017 13:22:11	
76c6e866289321a7c93b82b54852d33	f5a9f0e6b351c431402b8461ea51999	delivered	23/01/2017 18:29:09	25/01/2017 02:50:47	
ed69fb5eb88e0ed6a785585b27e16dbf	31ad1d1b63eb9962463f764d4e60c9d	delivered	29/07/2017 11:55:02	29/07/2017 12:05:32	
efce16cb79ec1d90b1da9085a6118aeb	494dded5b201313c64ed7f100595b95c	delivered	16/05/2017 19:41:10	16/05/2017 19:50:18	
34513e0c3fab462a5830c0989c7edcb	7711c624181d843aaf6e1855097bc37	delivered	13/07/2017 19:58:11	13/07/2017 20:10:08	
8256a6600982b15f86e904cd32918	d3e3b7c766bc0214e0c830b17ee2341	delivered	07/06/2018 10:06:19	09/06/2018 03:13:12	
5f9f6c15d0b717ac6ad1f3d77225a350	19402a48fe860416ad993348aba37740	delivered	25/07/2018 17:44:10	25/07/2018 17:55:14	
432aaf21d85167c2c86ec9448c4e42cc	3df704f53d3f1d4818840b34ec672a9f	delivered	01/03/2018 14:14:28	01/03/2018 15:10:47	
dc3b6b511fca050b97cd5c05de84dc3	3b6828a50ffe546942b7a473d70ac0fc	delivered	07/06/2018 19:03:12	12/06/2018 23:31:02	
403b97836b0c04622354c4f513062e5f	738b086814c6fc74b8c583f8516ee3	delivered	02/01/2018 19:00:43	02/01/2018 19:09:04	
116f0b09343b49556dbad5f35bee0cdf	3187789bec990987628d7a9beb4dd6ac	delivered	26/12/2017 23:41:31	26/12/2017 23:50:22	
85c8e9f6dc634de8d2f1e2904404d3	059f7fc5719c7d4dcbafe370971a8d70	delivered	21/11/2017 00:03:41	21/11/2017 00:14:22	
83018ec114ee8641c97e08f7b4e926f	7f8c8b9c2ae27bf3300f670c3d478be8	delivered	26/10/2017 15:54:26	26/10/2017 16:08:14	

Extraire une table avec des exemples Charger Transformer les données Annuler

Tout d'abord, dans toutes les tables, pour les variables numériques, nous avons remplacé les points par des virgules dans l'objectif de pouvoir manipuler les chiffres sur Power BI.

Dans la table des commandes (**Orders**), nous avons commencé par examiner les colonnes pour vérifier leur pertinence par rapport à l'analyse. Par exemple, les dates de commande et de livraison étant essentielles pour l'analyse des délais et des tendances temporelles, nous avons converti la colonne `order_purchase_timestamp` en type Date afin de faciliter les calculs et visualisations futures (tout comme toutes les autres dates de cette table). De plus, nous avons dupliqué cette colonne et gardé uniquement l'heure, pour créer la colonne `Heure_achat_commande`. Cela va nous permettre d'analyser l'heure d'achat indépendamment de la date d'achat.

En ce qui concerne les détails des articles commandés (**Order Items**), certaines colonnes, comme `seller_id`, ont été supprimées, car les informations sur les vendeurs n'étaient pas pertinentes pour les objectifs définis. Nous avons ensuite transformé la colonne `shipping_limit_date`, qui représentait la date limite de livraison, en type Date pour permettre des comparaisons temporelles précises. Par ailleurs, pour enrichir l'analyse, une nouvelle colonne calculée a été ajoutée pour déterminer le montant total par article. Cette colonne, nommée `Cout_total`, a été créée en additionnant le prix unitaire (`price`) aux frais de transport (`freight_value`), fournissant ainsi une mesure essentielle pour les analyses financières. Enfin, à l'aide d'une formule dans Power Query, nous avons créé une colonne personnalisée nommée `Catégorie_Prix_Article` qui permet d'identifier l'intensité de prix de l'article acheté plus facilement.

Si le prix est inférieur à 100 : prix faible ; entre 100 et 500 : Prix moyen ; entre 500 et 1 000 : prix élevé ; au-dessus de 1000 : Prix très élevé.

Dans la table des paiements (**Payments**), une attention particulière a été portée à la colonne `payment_value`, qui représentait le montant des paiements effectués. Celle-ci a été convertie en type Décimal afin d'assurer une précision dans les calculs. Les doublons potentiels sur `order_id` ont été éliminés, garantissant que chaque paiement soit unique et correctement associé à une commande spécifique. Pareillement, nous avons créé une

colonne personnalisée nommée *Catégorie_Prix_Commande* qui permet d'identifier l'intensité de prix de la commande plus facilement.

Si la valeur du paiement est inférieure à 500 : Faible ; entre 500 et 1000: Modérée ; au-dessus de 1000 : Elevé.

Pour les données relatives aux clients (**Customers**), une simplification a été effectuée en supprimant la colonne `customer_unique_id`, redondante avec `customer_id`, qui est déjà unique. Cette rationalisation a permis d'alléger les données tout en conservant les informations nécessaires à l'analyse des comportements des clients.





Enfin, dans la table des produits (**Products**), bien que moins transformée, les colonnes ont été vérifiées pour s'assurer que leurs types de données correspondaient aux besoins de l'analyse. .

Toutes ces transformations ont été guidées par les objectifs de l'analyse, à savoir : simplifier les données en éliminant les colonnes inutiles, garantir la cohérence des types de données pour des calculs fiables, et enrichir les tables avec des colonnes calculées pour fournir des indicateurs pertinents. En combinant ces ajustements, nous avons assuré que chaque table soit parfaitement adaptée pour l'étape suivante de modélisation et de visualisation.

Une fois ces transformations faites, nous avons utilisé des outils comme le Qualité de la colonne et Profil de colonne pour nous assurer de la validité des données, analyser la distribution et identifier des anomalies potentielles.

La plupart des données étaient valides à 100% (0% d'erreurs et de vides) mais quelques clarifications étaient nécessaires.

Par exemple, dans la table *order*, une partie des données avait des valeurs vides pour les dates de livraisons. Cela correspond aux commandes qui n'ont jamais été livrées (« invoiced » dans *order_status*).

 order_approved_at	 order_delivered_carrier_date	 order_delivered_customer_date	
<ul style="list-style-type: none"> ● Valide 100 % ● Erreur 0 % ● Vide 0 % 	<ul style="list-style-type: none"> ● Valide 99 % ● Erreur 0 % ● Vide 1 % 	<ul style="list-style-type: none"> ● Valide 98 % ● Erreur 0 % ● Vide 2 % 	
02/10/2017	04/10/2017	10/10/2017	
26/07/2018	26/07/2018	07/08/2018	
08/08/2018	08/08/2018	17/08/2018	
18/11/2017	22/11/2017	02/12/2017	
13/02/2018	14/02/2018	16/02/2018	
09/07/2017	11/07/2017	26/07/2017	
13/04/2017	null	null	

Pareillement, les variables qui concernent les descriptions de produits sont parfois vides car les descriptions sont incomplètes. Ces lignes ne doivent pas être supprimées car elles

représentent un potentiel axe d'analyse : les produits avec une description incomplète ont-ils moins de chance d'être vendus ?

AB_C product_category_name	123 product_name_lenght	123 product_description_lenght
● Valide 97 %	● Valide 97 %	● Valide 97 %
● Erreur 0 %	● Erreur 0 %	● Erreur 0 %
● Vide 3 %	● Vide 3 %	● Vide 3 %

AB_C product_id	AB_C product_category_name	123 product_name_lenght	123 product_description_lenght
● Valide 100 %	● Valide 97 %	● Valide 97 %	● Valide 97 %
<div> <div> Statistiques de colonnes </div> <div> Distribution de valeurs </div> </div>			
Nombre	1000	cama_mesa_banho	
Erreur	0	esporte_lazer	
Vide	0	beleza_saude	
Distincte(s)	59	moveis_decoracao	
Uniques	7	utilidades_domesticas	
Chaîne vide	29	automotivo	
Min		brinquedos	
Max	utilidade...	informatica_acessorios	
		telefonica	
		relogios_presentes	
		cool_stuff	

3. Chargement des données

Dans notre projet, le chargement des données s'effectue après avoir extrait et transformé les données brutes pour répondre aux besoins analytiques. Une fois les dimensions (clients, produits, commandes, paiements, etc.) et la table de faits correctement définies et enrichies, ces données sont transférées dans le modèle décisionnel final. Ce modèle est conçu pour être optimisé en vue de requêtes rapides et efficaces.

Lors du chargement, les données sont indexées, ce qui permet d'améliorer les performances des requêtes en facilitant l'accès aux informations recherchées. Par exemple, les clés primaires et secondaires, telles que *order_id*, *product_id* ou *customer_id*, sont utilisées pour relier les tables de dimensions à la table de faits. Cette indexation garantit que les jointures et les agrégations, nécessaires pour construire les visualisations et les indicateurs, s'exécutent de manière fluide, rendant les données prêtes pour l'analyse dans les tableaux de bord.

Création d'une table Calendrier :

Pour ce projet, nous avons créé une table de temps afin de structurer les données temporelles et de faciliter les analyses. Nous avons choisi de couvrir la période allant de 2016 à 2018, car cette plage contient l'ensemble des données de notre dataset, qui s'étendent du 4 septembre 2016 au 17 octobre 2018. Cette période a été définie avec précision pour garantir

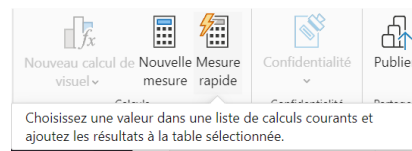
que toutes les dates présentes dans les commandes soient incluses dans la table de temps, sans oublier aucune période pertinente.

Pour cela, dans Power Query, nous avons entré la formule suivante :

```
= List.Dates(#date(2016,9,4), 365*3, #duration(1,0,0,0))
```

Une fois la liste créée, nous l'avons convertie en tableau pour pouvoir travailler avec ces données dans Power BI. Nous avons ensuite enrichi cette table en ajoutant des colonnes supplémentaires, comme l'année, le mois, le nom du jour et la semaine de l'année. Ces colonnes ont été créées en dupliquant la colonne des dates, puis en utilisant les transformations de type date dans l'onglet « Transformer » de Power Query.

Cette table, que nous avons appelée "Calendrier", est essentielle à nos analyses. Elle a été reliée à la table des commandes via la colonne *order_purchase_timestamp* pour permettre des calculs temporels précis, comme la croissance des ventes d'une année sur l'autre ou les analyses par mois.



Création de mesures DAX :

Pour enrichir nos analyses et garantir une meilleure interprétation des données, nous avons créé six mesures dynamiques à l'aide de DAX dans Power BI.

La première mesure que nous avons mise en place, intitulée **Contribution_FRET**, calcule la part des frais de livraison par rapport au montant total des ventes. Cet indicateur est essentiel pour comprendre l'impact des frais de livraison sur le chiffre d'affaires global. Il permet d'identifier dans quelle mesure ces frais influencent les marges et de comparer cet impact entre différentes périodes ou catégories de produits.

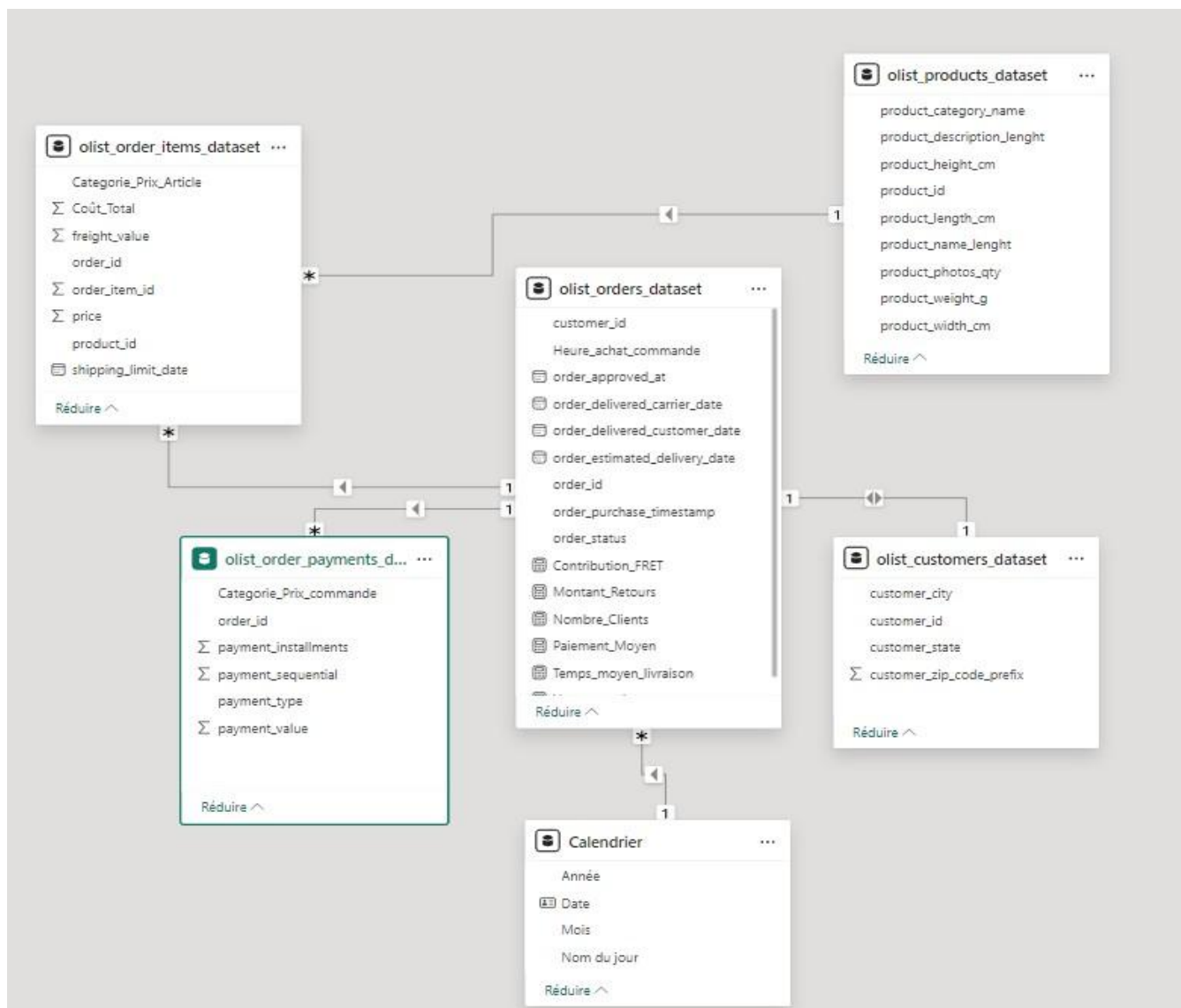
Nous avons également élaboré la mesure **Montant_Retours**, qui reflète les pertes financières liées aux commandes annulées ou indisponibles. En se basant sur les statuts des commandes, notamment "canceled" et "unavailable", cette mesure nous permet d'analyser les retours ou les commandes non honorées, offrant ainsi une vision claire des opportunités de ventes perdues.

La mesure **Paieement_MOYEN** calcule quant à elle le montant moyen payé par commande. Cet indicateur est particulièrement utile pour identifier les comportements d'achat des clients et évaluer la valeur moyenne des transactions. Grâce à cette mesure, nous pouvons détecter des variations entre les différentes périodes ou catégories et mieux comprendre les tendances de consommation.

Pour évaluer la performance logistique, nous avons conçu la mesure **Temps_moyen_livraison**, qui calcule la durée moyenne entre la date de commande et la date de livraison des produits. Cet indicateur met en évidence l'efficacité des processus de traitement et de livraison, en excluant les commandes qui n'ont pas été livrées pour garantir une analyse précise.

Enfin, nous avons créé la mesure **Ventes_totales**, qui calcule le chiffre d'affaires global en additionnant tous les montants payés par les clients. Cet indicateur fondamental fournit une vue d'ensemble de la performance financière sur l'ensemble de la période analysée et constitue une base solide pour l'interprétation des résultats globaux. Nous avons aussi créé une mesure intitulée **Nombre_Clients**, qui permet de calculer le nombre total de clients distincts ayant passé une commande sur la période analysée.

Ces six mesures sont dynamiques et adaptées aux besoins de notre projet. Elles permettent d'approfondir les analyses, qu'il s'agisse d'évaluer les performances globales, d'identifier des zones problématiques ou encore de comparer les résultats entre différentes dimensions. Intégrées aux visualisations, elles facilitent la prise de décision en offrant des informations claires et précises sur les performances de la plateforme e-commerce.



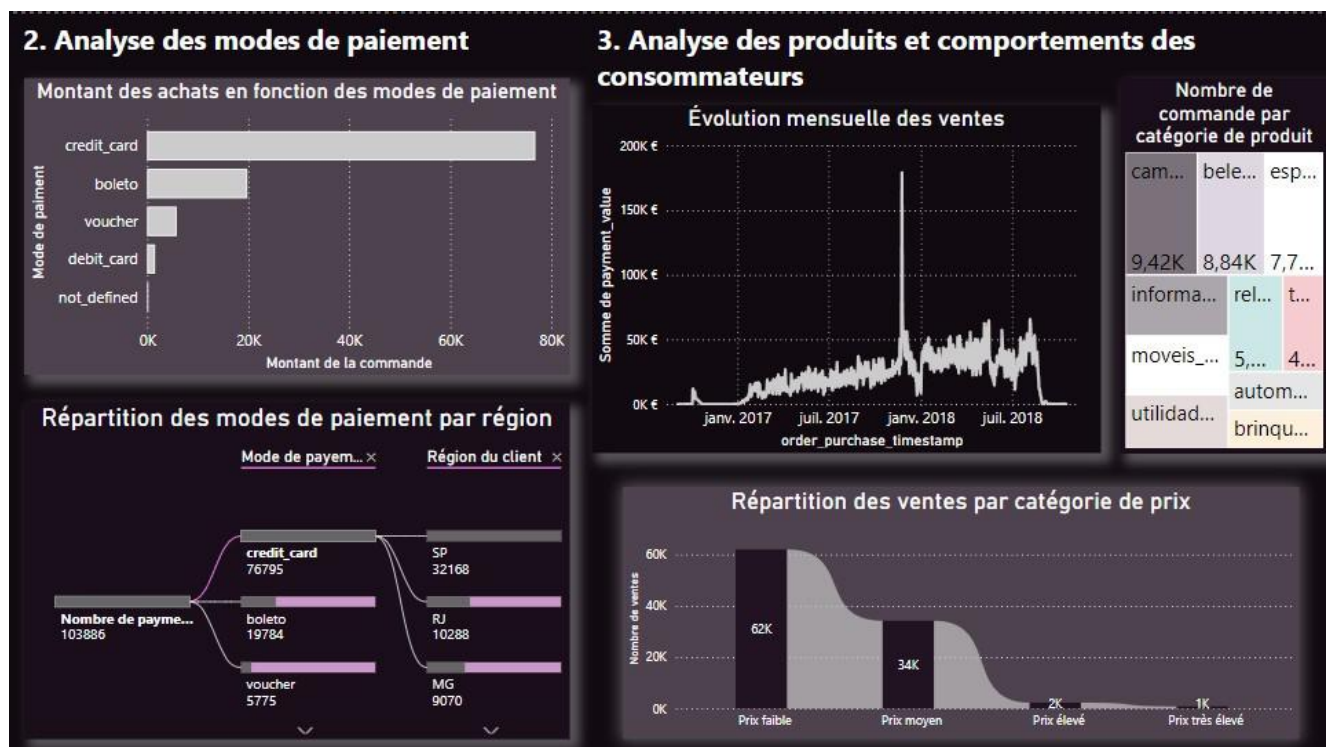
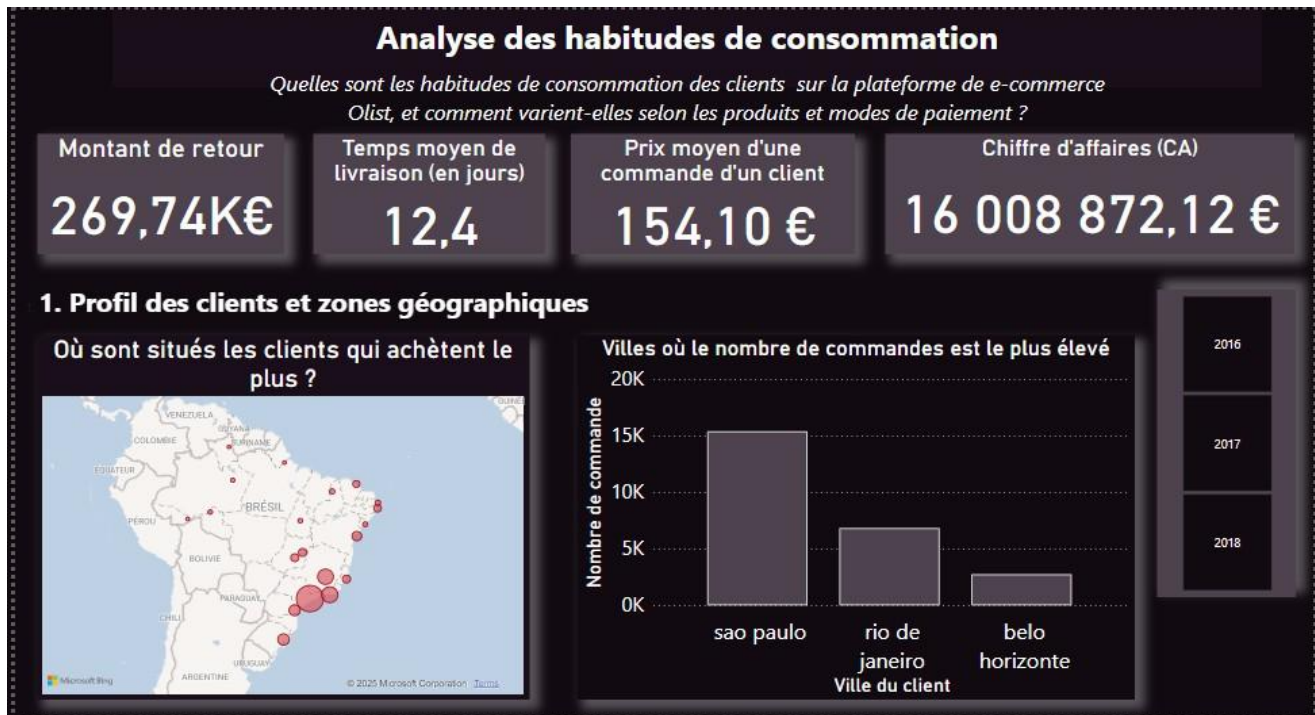
Dans notre système d'information décisionnel, la table de faits, *olist_order_items_dataset*, constitue le cœur du modèle de données en centralisant toutes les informations transactionnelles nécessaires à nos analyses. Elle regroupe des éléments clés tels que les identifiants des commandes et des produits, les quantités commandées, les prix et les frais de livraison. Cette table sert de base au calcul des mesures que nous avons créées, comme le chiffre d'affaires total, la contribution des frais de livraison, le montant moyen des paiements, ou encore le délai moyen de livraison.

Elle est reliée à plusieurs dimensions, comme les clients, les produits, les paiements et les commandes, via des clés primaires telles que *order_id*, *product_id* et *customer_id*. Ces connexions permettent de croiser les mesures quantitatives avec des informations qualitatives provenant des dimensions, comme les catégories de produits, les régions des clients ou les périodes temporelles.

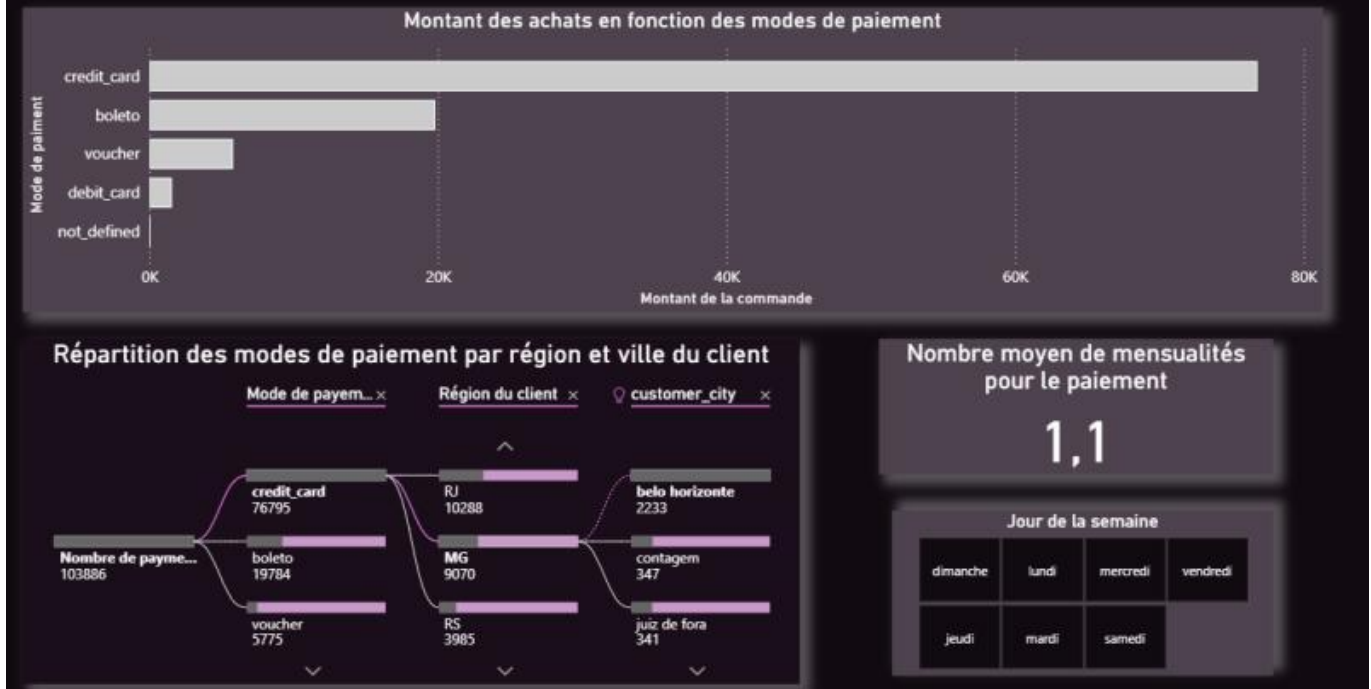
Bien que notre modèle soit proche d'un schéma en étoile, certaines dimensions enrichies, comme *olist_products_dataset*, apportent des détails supplémentaires tout en limitant la redondance, ce qui nous rapproche d'un schéma en flocon. Cette hybridation permet une exploration précise des tendances, comme les performances des catégories de produits ou l'impact des délais de livraison, tout en maintenant une organisation claire et une efficacité optimale dans l'analyse des performances de notre plateforme e-commerce.

Restitution des données

Après avoir structuré notre modèle de données et créé les mesures dynamiques nécessaires, nous passons désormais à la phase de réalisation des visualisations et du tableau de bord interactif à l'aide de Power BI. Cette étape nous permet de transformer les données brutes en insights visuels exploitables, facilitant ainsi l'analyse et la prise de décision.



3. Analyse des modes de paiement



Page 1 :

- Prix moyen d'une commande (154,10 €) : les clients dépensent modérément, ce qui montre une propension à acheter des produits abordables ou de gamme moyenne. Cela indique une clientèle sensible aux prix, ce qui explique la dominance des produits à prix bas ou moyens dans la répartition des ventes (Page 2).

Analyse géographique des clients :

Les commandes sont majoritairement concentrées en Amérique latine, et plus spécifiquement au Brésil, ce qui est logique étant donné que la plateforme est brésilienne. On observe également une présence plus modeste en Amérique du Nord.

En ce qui concerne d'autres régions (Europe, Asie, etc.) aucune activité notable, ce qui indique un potentiel inexploité pour l'expansion internationale.

Villes avec le plus de commandes :

- São Paulo (20K commandes) : Principal marché pour la plateforme.
- Rio de Janeiro et Belo Horizonte : Deuxième et troisième villes avec une forte activité, ce qui montre qu'il existe une demande importante dans d'autres grandes métropoles brésiliennes.

Page 2 :

1. Modes de paiement :

Montant des achats par mode de paiement :

- La carte de crédit domine largement les transactions, ce qui reflète la confiance des consommateurs envers ce mode de paiement et son accessibilité.
- Le boleto (19 784 transactions) est également très populaire au Brésil, car il est souvent utilisé par des clients sans carte bancaire ou qui préfèrent payer en espèces.
- Les autres modes (cartes de débit et vouchers) sont moins utilisés mais pourraient croître avec le temps.

- Répartition des modes de paiement par région :
- São Paulo (SP) : 76 975 paiements par carte de crédit — une base forte. Le boleto est utilisé à Rio de Janeiro (RJ) et Minas Gerais (MG).
- On constate une préférence locale

Analyse des catégories de produits :

- Les caméras (9,42K commandes) sont les produits les plus populaires, suivis par les articles de sport et les produits informatiques. Ces produits sont probablement plus abordables, ce qui correspond aux préférences des consommateurs pour des produits à prix bas ou moyens.

Évolution des ventes dans le temps : les ventes connaissent des pics saisonniers (janvier, juillet, décembre 2018). Ces pics coïncident probablement avec des événements comme le Black Friday, Noël ou des campagnes promotionnelles.

Cela montre que les consommateurs sont sensibles aux promotions et achètent davantage pendant ces périodes.

Répartition des ventes par catégorie de prix :

On constate que la majorité des ventes se concentre dans les catégories "Prix faible" et "Prix moyen", qui représentent respectivement environ 60 000 et 34 000 ventes. Ces deux segments dominent largement, ce qui souligne une forte préférence des consommateurs pour des produits accessibles en termes de coût.

En revanche, les catégories "Prix élevé" et "Prix très élevé" enregistrent des volumes de ventes nettement inférieurs. La chute rapide du nombre d'articles vendus dans ces segments illustre la relation inverse entre le prix et la demande, avec un marché beaucoup plus restreint pour les produits haut de gamme. Cette tendance, représentée par une courbe en forte décroissance, reflète clairement que le prix joue un rôle déterminant dans les décisions d'achat.

D'un point de vue stratégique, cette répartition indique que l'essentiel des revenus est généré par les produits à prix faible et moyen. Cela suggère une opportunité d'optimisation dans ces segments, notamment en renforçant l'offre ou en développant des promotions spécifiques pour attirer davantage de clients. Toutefois, les produits plus onéreux, bien que moins populaires en termes de volume, peuvent offrir des marges plus élevées et justifient une stratégie de niche, ciblant un public spécifique avec des efforts marketing adaptés.

Résumé des analyses :

En résumé, ces analyses montrent que les habitudes de consommation varient selon les régions, les modes de paiement, et les produits.

Les régions jouent un rôle important dans la détermination des préférences des consommateurs, avec des différences marquées dans les choix de produits et les habitudes d'achat en fonction des caractéristiques géographiques, économiques et culturelles.

Les modes de paiement ont également un impact sur le comportement des consommateurs. Les clients peuvent être plus enclins à utiliser des cartes de crédit, des paiements en plusieurs fois, ou encore des méthodes locales comme le boleto bancaire, selon leur familiarité et la confiance dans ces systèmes.

Enfin, les produits eux-mêmes influencent grandement les choix des consommateurs, certains segments du marché étant plus enclins à acheter des produits de luxe, tandis que d'autres privilégient les articles à bas prix ou les promotions. Les catégories de produits influencent également le mode de paiement choisi, les achats de haute valeur étant souvent associés à des paiements différés ou fractionnés, tandis que les produits à faible coût sont souvent réglés en une seule fois.

Conclusion

En conclusion, ce projet nous a permis de répondre à la problématique en identifiant clairement les habitudes de consommation des clients sur les plateformes de commerce électronique au Brésil. Les analyses ont révélé que la majorité des ventes se concentre sur des produits à prix faible et moyen, indiquant une forte sensibilité des consommateurs au coût des articles. Cependant, bien que les produits hauts de gamme soient moins demandés en volume, ils offrent des opportunités intéressantes en termes de marges pour des segments spécifiques.

Les modes de paiement jouent également un rôle déterminant dans ces habitudes. Les paiements échelonnés, souvent utilisés pour les articles plus coûteux, reflètent une tendance des consommateurs brésiliens à répartir leurs dépenses, tandis que les paiements directs dominent pour les produits moins chers. Enfin, nous avons constaté que la demande varie en fonction des caractéristiques des produits, mais aussi selon les régions, mettant en lumière des comportements d'achat spécifiques à certaines zones géographiques.

Au-delà des résultats, ce projet nous a apporté une compréhension approfondie des processus décisionnels et des outils analytiques nécessaires pour traiter des volumes importants de données. Nous avons appris à structurer un modèle décisionnel robuste, à exploiter des outils comme Power BI pour créer des visualisations interactives et à relier des données quantitatives à des dimensions qualitatives. Ce travail a également renforcé nos compétences en interprétation de données complexes, en nous montrant comment traduire des tendances chiffrées en recommandations stratégiques concrètes.

En répondant à la problématique, ce projet met en évidence l'importance de bien comprendre les comportements des consommateurs pour adapter les stratégies commerciales. Il nous ouvre aussi des perspectives pour approfondir l'analyse, comme la segmentation des clients ou la prédiction des ventes futures, afin de continuer à exploiter tout le potentiel des données e-commerce.