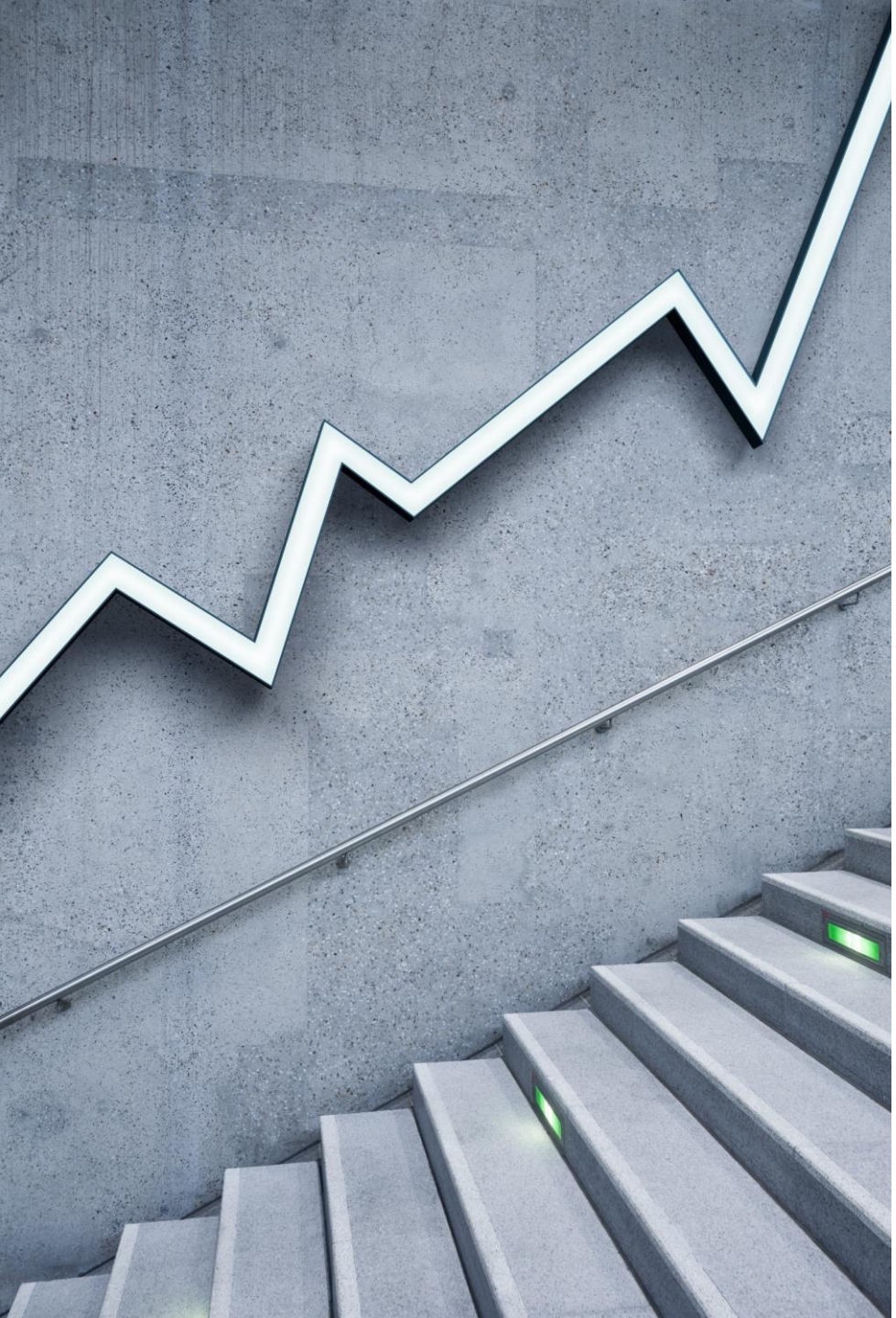


INSEEC M2

DATA ANALYTICS

LANGAGES DE PROGRAMMATION

13/02/2025

PROJET NOTÉ

DATA ANALYTICS

LANGAGES DE PROGRAMMATION

SYLLABUS

Introduction

PARTIE 1

Découverte de Python : bases du langage, librairies, IDE

PARTIE 2

Data management : Pandas & Numpy

PARTIE 3

Data visualisation : Matplotlib, Seaborn, Plotly

Optionnel : Streamlit

PARTIE 4

Démarche d'analyse complète

Lancement Projet en groupe

PARTIE 5

Oral du projet

PRÉSENTATION DU SUJET

Projet Noté
Présentation du sujet

Contexte :

La marque « Feminiz » est soumise à une forte concurrence sur le marché de la lingerie. L'une des premières actions a été de développer une gamme plus large de produits complémentaires avec des maillots de bain et des vêtements d'intérieur.

Parallèlement, un programme de fidélisation a été mis en place pour établir un lien avec les clients et les récompenser.

Pour aller plus loin dans la perspective de se différencier de ses concurrents, « Feminiz » souhaite établir une segmentation RFM afin de communiquer différemment avec ses clients et d'allouer de manière optimale ses investissements marketing.

La marque de lingerie « Feminiz » vous demande donc de réaliser une analyse RFM de ses clients sur l'année écoulée.

Données :

Vous aurez plusieurs datasets à disposition pour cette analyse : les transactions, les caractéristiques des clients ainsi que des tableaux de référence sur les magasins et les produits.

Projet Noté
Présentation du sujet

Livrable attendu :
Vous devez préparer une présentation de 15 minutes maximum (prévoir des slides) pour le comité de direction du département marketing.

Le plan de la présentation est imposé comme suit :

1. Contexte et objectifs de la réunion
2. Présentation des résultats de l'analyse RFM
3. Analyse d'un segment particulier de clients
4. Conclusion et vos recommandations

Vous serez noté sur le contenu et la forme. Soyez professionnels !

Votre notebook Python sera également à rendre et comptera dans la note.

Projet Noté
Présentation du sujet

<u>Organisation :</u>

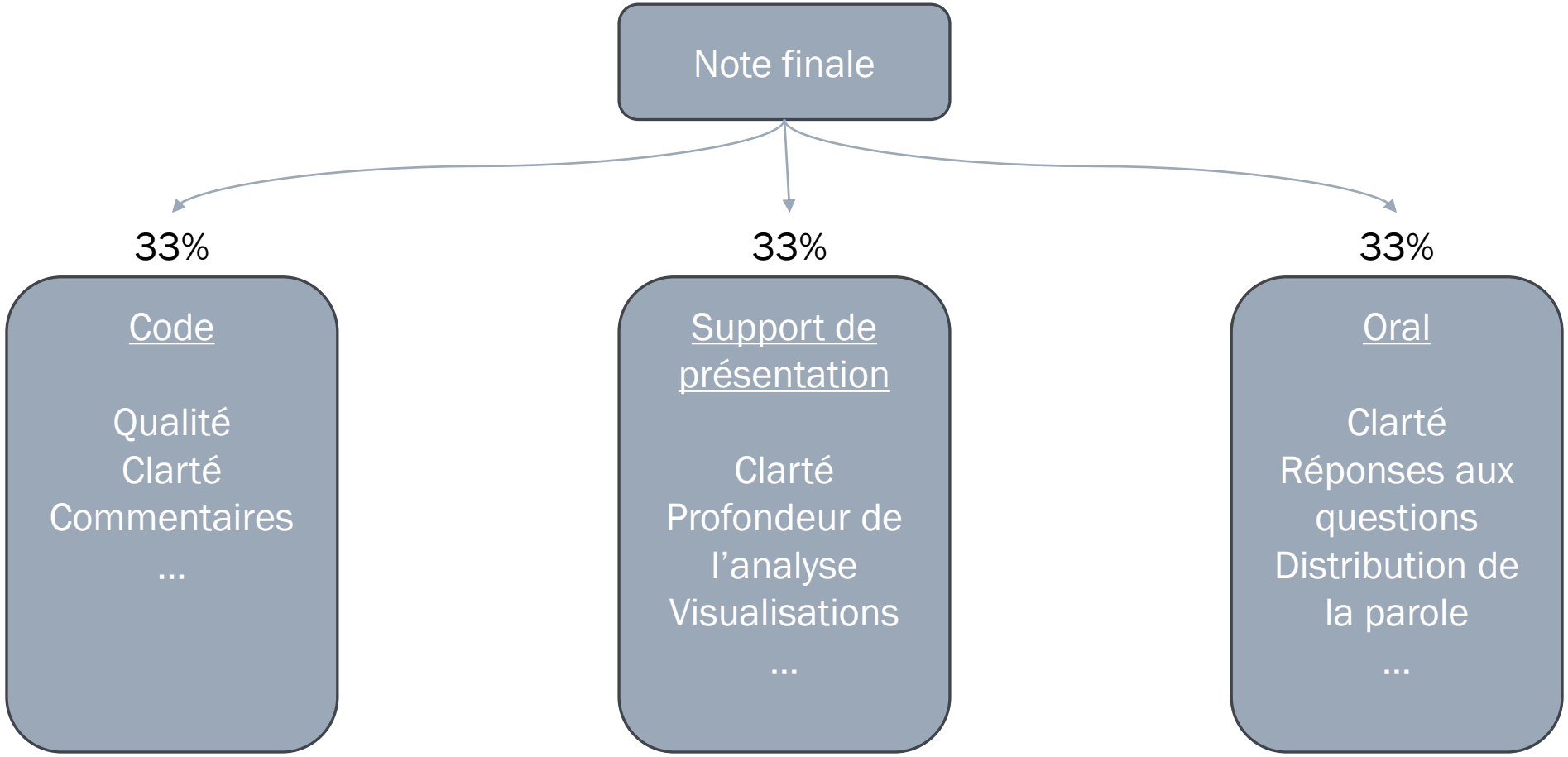
Pendant les séances suivantes (et ensuite à la maison), vous travaillerez en groupe de 3+ étudiants sur le projet que vous présenterez à l'oral.

Chaque groupe s'organise comme il l'entend pour le projet et peut paralléliser les différentes composantes du projet (code Python, analyse des données, présentation PPT, organisation générale) mais chaque membre doit être à l'aise avec toutes les étapes.

L'objectif global de ce projet est de construire une segmentation RFM pour une marque de lingerie :

- Tous les groupes doivent construire la segmentation RFM selon les lignes directrices suivantes et selon ce que nous discuterons pendant les sessions pratiques.
- Ensuite, chaque groupe analysera un segment en particulier, l'analysera à l'aide de l'outil statistique de son choix et formulera des recommandations marketing pour communiquer efficacement avec lui.

Projet Noté
Modalités de notation



PROPOSITION DE DÉMARCHE

Projet Noté
Proposition de démarche

Partie 1 : Compréhension des données

Importez les 6 tableaux csv et regardez les premières lignes de chaque tableau.

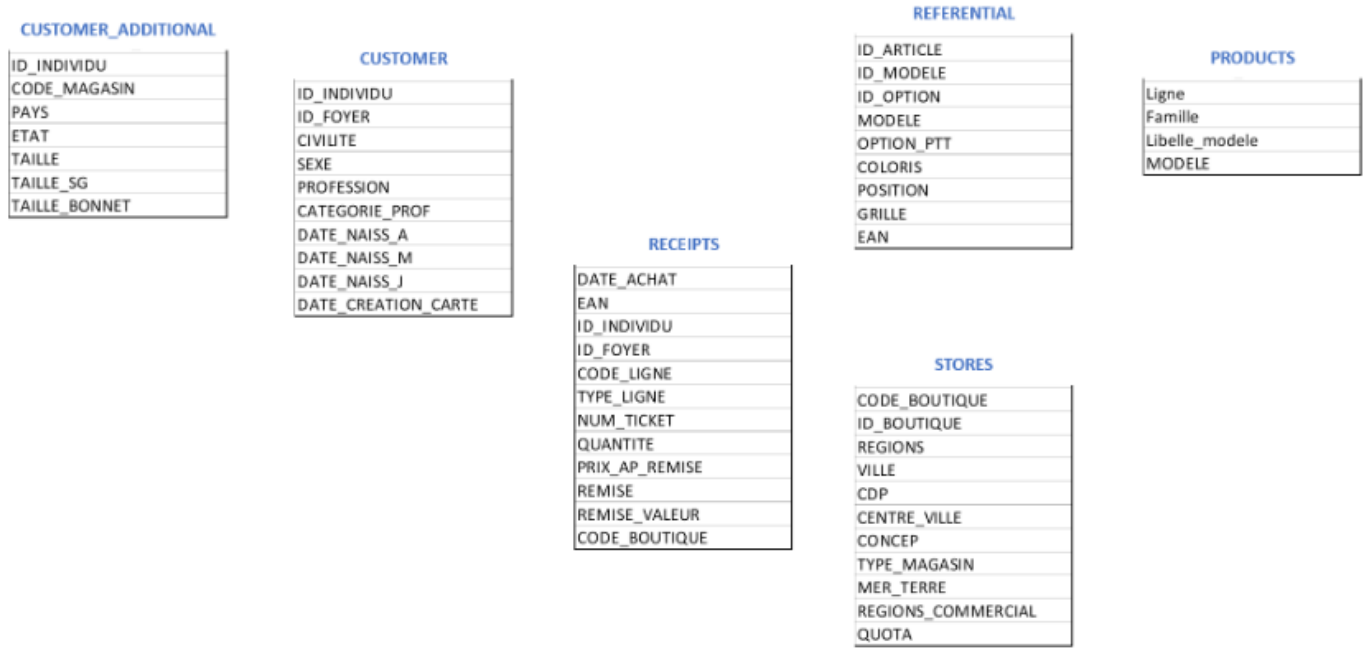
- CUSTOMER
- CUSTOMER_ADDITIONAL
- STORE
- REFERENTIAL
- RECEIPTS
- PRODUCTS

Projet Noté
Proposition de démarche

Partie 1 : Compréhension des données

Complétez le schéma relationnel de données suivant avec le lien entre les tables.

Par exemple, entre CUSTOMER et CUSTOMER_ADDITIONAL, existe-t-il des colonnes portant le même nom ? Est-il judicieux d'utiliser cette colonne (ou ces colonnes) ?



Projet Noté
Proposition de démarche

Partie 1 : Compréhension des données

Créez les tables CUSTOMERS_INFO et RECEIPTS_INFO contenant toutes les informations comme suit :
→ Pour créer CUSTOMERS_INFO, faites une jointure CUSTOMER et CUSTOMER_ADDITIONAL.

- Joindre la table CUSTOMERS.
- La fonction « merge » peut être utilisée en Python : sur quelle colonne la jointure doit-elle être effectuée ? Doit-on effectuer une jointure à gauche, une jointure interne, une jointure à droite, une jointure externe ?
- Assurez-vous que les trois tables ont le même nombre de lignes.
- Code_magasin sera renommé MANAGING_STORE pour ne pas le confondre avec Code_magasin de la table RECEIPTS.

Projet Noté
Proposition de démarche

Partie 1 : Compréhension des données

Faites une jointure des dimensions de RECEIPTS (STORE, REFERENTIAL, PRODUCTS et RECEIPTS)

- Commencez par joindre les rubriques RECEIPTS et REFERENTIAL.
- Nous décidons de convertir le type de ... en raison d'un problème.
- Créer RECEIPTS_INFO avec les tables RECEIPTS et REFERENTIAL
- Maintenant faites la jointure de PRODUCTS avec la table créée ci-dessus.
- Ne garder que Ligne et Famille de PRODUCTS, et la colonne utilisée pour la fusion. Fusionner ce DataFrame avec la table RECEIPTS_INFO récemment créée.
- Maintenant la jointure de STORE avec la table créée ci-dessus.
- Ne garder que REGIONS, CENTRE_VILLE, TYPE_MAGASIN et REGIONS_COMMERCIAL de STORE, et la variable pour la fusion. Fusionnez ce DataFrame avec RECEIPTS_INFO créé précédemment.

Projet Noté
Proposition de démarche

Partie 1 : Compréhension des données

Avant de commencer l'audit d'une base de données, il est important de comprendre comment les différents types de données sont liés. Il est obligatoire de définir un moyen d'identifier de manière unique les « actions des clients ».

Puisque nous nous intéressons aux achats effectués dans les magasins, nous devons prendre en compte les éléments suivants :

- 1 visite unique d'une personne donnée dans un magasin donné à une date donnée, définie comme un ticket de caisse.
- Si 1 personne a visité deux fois 1 magasin à une date donnée, nous devons identifier 2 reçus différents.
- Si 2 personnes ont visité 1 magasin donné à 1 date donnée, nous devons identifier 2 reçus différents.
- Si 1 personne a visité 2 magasins donnés à une date donnée, nous devons identifier 2 reçus différents.

Projet Noté
Proposition de démarche

Partie 1 : Compréhension des données

Sélectionnez toutes les informations relatives aux tickets de caisse du client 174591.

→ Déduire de l'analyse du client 174591 et des hypothèses ci-dessus les colonnes nécessaires pour identifier 1 visite en magasin, c'est-à-dire 1 ticket de caisse unique.

Il semble que nous devions utiliser l'information NUM_TICKET. Cependant, en regardant les lignes de ce client, est-ce suffisant ? Pouvons-nous utiliser d'autres variables pour nous assurer que nous ne prenons en compte que les visites uniques, c'est-à-dire les tickets d'achat uniques ?

Projet Noté
Proposition de démarche

Partie 1 : Compréhension des données

A partir des 2 bases de données précédemment constituées, étudier les variables suivantes :
dans RECEIPTS_INFO :

- Analyser ensemble `REGIONS` et `REGIONS_COMMERCIAL`. Y a-t-il une colonne qui ne fournit pas d'informations supplémentaires et qui peut être supprimée ?
- Analyser `CENTRE_VILLE`. Y a-t-il une modalité qui peut être modifiée ?
- Analysez ensemble `Ligne` et `Famille` juste pour avoir une idée de la hiérarchie du produit. Il ne devrait pas y avoir de modification à faire.
- Analyser `CODE_LIGNE` et `TYPE_LIGNE`. Ces colonnes sont-elles utiles ?
- Avec un peu d'analyse sur PRIX_AP_REMISE, REMISE_VALEUR, REMISE (à faire!), nous supposons que : PRIX_AP_REMISE est le prix payé après les remises potentielles, REMISE_VALEUR semble plus utile que REMISE, REMISE_VALEUR est la remise représentée en % (entre 0 et 100 avec quelques anomalies).
- Il faut alors supprimer la colonne REMISE, remplacer les valeurs de REMISE_VALEUR par 100 si les valeurs sont > 100.

Projet Noté
Proposition de démarche

Partie 1 : Compréhension des données

En analysant les données et en discutant avec l'équipe marketing, vous devez créer de nouvelles variables :

- une colonne nommée `PLV` (Publicité en Lieu de Vente) identifiant les achats où MODELE est égal à PLV.
- une colonne appelée `Gift` identifiant les achats dont le MODELE est FAVO ou FAVORI, et PRIX_AP_REMISE = 0.
- une colonne appelée `Entry-level` identifiant les produits d'entrée de gamme visant à acquérir de nouveaux clients : MODELE est égal à ACCESS.

Projet Noté
Proposition de démarche

Partie 2 : Préparation de la donnée

- Trouver les colonnes qui doivent être stockées sous forme de dates. Les convertir en type date Pandas (indice : pandas.to_datetime() devrait s'avérer utile)
- Calculer le prix final d'un achat avec `PRIX_AP_REMISE` x `QUANTITE`.
- Filtrez sur les 12 derniers mois disponibles, en prenant des mois complets.

TRÈS IMPORTANT : Tous les calculs futurs concernant les recettes seront effectués avec ce périmètre de date !

Projet Noté
Proposition de démarche

Partie 2 : Préparation de la donnée

Calculer la valeur monétaire par visite et le prix moyen

Pour chaque visite (voir précédemment pour identifier correctement 1 visite), calculez :

- le nombre de produits vendus (vous pouvez l'appeler NB_PRODUCTS)
- la valeur monétaire (somme des prix des produits ; vous pouvez l'appeler VISIT_VALUE).
- le prix moyen par visite devrait être facile à calculer (appelé AVG_PRICE)

Projet Noté
Proposition de démarche

Partie 2 : Préparation de la donnée

Calculer la FRÉQUENCE par individu (à partir du tableau précédent par visite)

Pour chaque individu, calculer :

- NB_VISITS : le nombre de visites,
- CUMUL_VALUE : la somme de la valeur monétaire par visite (en utilisant VISIT_VALUE calculée précédemment),
- AVG_VISIT_VALUE : moyenne de la valeur monétaire par visite (à l'aide de VISIT_VALUE calculée),
- AVG_NB_PRODUCTS_VISIT` : moyenne des quantités (en utilisant NB_PRODUCTS calculé auparavant)

Projet Noté
Proposition de démarche

Partie 2 : Préparation de la donnée

Calculer la RECENCE par individu (à partir de la table précédente par visite)
Calculer d'abord la dernière date d'achat pour chaque individu.

- Comptez le nombre distinct de magasins, de lignes et de familles pour chaque individu.
- Compter le nombre de cadeaux.

Part des visites dans le magasin gestionnaire
Joindre VISIT_VALUE (en ne gardant que l'ID INDIVIDU et le CODE BOUTIQUE) et les informations sur les clients (en ne gardant que l'ID INDIVIDU et le MAGASIN GESTIONNAIRE).
Créez une colonne pour voir si le magasin d'achat est le même que le magasin gestionnaire.
Calculez ensuite la part des visites effectuées dans le magasin gestionnaire.

Projet Noté
Proposition de démarche

Partie 2 : Préparation de la donnée

Table CUSTOMERS_INFO - Calcul par individu :

Âge (AGE) :
Tout d'abord, trouvez les colonnes de dates qui vous aideront à calculer l'âge (il n'est pas nécessaire de calculer l'âge avec une précision de mois ou de jour, l'année suffit, mais vous pouvez la calculer si cela vous intéresse !)

L'âge est-il calculé par rapport à l'année en cours ? Ou une autre année est-elle plus appropriée ? Pensez à la fourchette de dates que nous utilisons actuellement.

A ce stade, la distribution (que vous aurez visualisée) de l'âge des clients révèle des valeurs aberrantes : clients avec un âge négatif (ou inférieur à 18 ans) : créer des valeurs manquantes lorsque l'âge < 15 ou l'âge > 90.

Projet Noté
Proposition de démarche

Partie 2 : Préparation de la donnée

Table CUSTOMERS_INFO - Calcul par individu :

Ancienneté
Convertir DATE_CREATION_CARTE dans le bon type.
Calculer l'ancienneté du client (années après la création de la carte de fidélité).
Y a-t-il des valeurs aberrantes ?
Créez des valeurs manquantes lorsque l'ancienneté est supérieure à 10 ans, et réduisez les valeurs négatives à 0.

Rassembler toutes les caractéristiques dans une table MASTER (au niveau individuel)

Projet Noté
Proposition de démarche

Partie 2 : Préparation de la donnée

Beaucoup de clients ont des valeurs manquantes concernant les caractéristiques d'achat. Vérifiez qu'ils ne sont effectivement pas présents dans la table des reçus depuis le début.

En effet, ces clients n'ont pas de reçus et nous n'allons donc pas les prendre en compte dans la segmentation RFM. Il peut s'agir de prospects, de clients inactifs (dernier achat <1 an) ou de clients très récents (dernier achat > 30/11/2016).

Pour la segmentation RFM, nous excluons les clients inactifs sur l'année d'étude.
→ Remplir les valeurs manquantes avec 0 pour NB_GIFTS.

Projet Noté
Proposition de démarche

Partie 3 : Construction de la segmentation RFM

Analysez et visualisez la distribution des variables suivantes :

- **Montant :** *Nous avons des clients dont la valeur cumulée est de 0. Analysez 1 client pour comprendre le raisonnement et la façon de les traiter.*
- **Fréquence**
- **Récence :** Rappelez-vous que nous ne prenons en compte qu'un an d'historique d'achat pour les clients. Si l'ancienneté est inférieure à 1 an, de quel type de clients s'agit-il ? Nous voulons créer une variable pour les identifier.

Puis :

→ Déterminez des seuils afin de les répartir en groupes faibles/moyens/élevés.

→ Créer les segments RFM finaux

Projet Noté
Proposition de démarche

Partie 4 : Analyse des segments

Utiliser les données ci-dessous pour décrire le segment de votre choix :

- ✓ CATEGORY INDICATOR
- ✓ CLIENTS AGE
- ✓ CLIENTS GENDER
- ✓ CLIENTS SENIORITY
- ✓ PURCHASE CUMULATED PURCHASE AMOUNT
- ✓ PURCHASE AVERAGE BASKET
- ✓ PURCHASE NUMBER OF VISITS
- ✓ BEHAVIOR RECENCY
- ✓ BEHAVIOR LOCATION OF THE MANAGING STORE
- ✓ BEHAVIOR PURCHASE PART OF THE MANAGING STORE
- ✓ PRODUCTS TYPE NB STORES
- ✓ PRODUCTS TYPE NB DIFFERENT FAMILIES
- ✓ PRODUCTS TYPE NB DIFFERENT LINES
- ✓ PROGRAM NB GIFTS

Projet Noté
Proposition de démarche

Partie 5 : Recommandations marketing

Concentrez-vous sur un segment et **donnez des recommandations à la direction marketing** en utilisant des outils d'analyse et de statistiques. Cette partie est obligatoire pour la présentation.