Data refinement

Choix du devoir et de l'environnement de travail

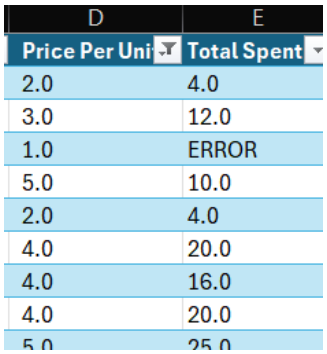
"Dirty data cafe"

A partir d'une extraction brute en format CSV, je choisis de travailler sur Excel.

Je maîtrise mieux l'outils Excel et le langage DAX.

Excel permet aussi de fournir des visuels, sur la donnée traitée, ergonomiques similaires à Power BI.

Exploration et Nettoyage de données



D	E
Price Per Uni	Total Spent
2.0	4.0
3.0	12.0
1.0	ERROR
5.0	10.0
2.0	4.0
4.0	20.0
4.0	16.0
4.0	20.0
5.0	25.0

Les colonnes "Price Per Unit" et "Total Spent" ont été remplies avec un *point* (".") au lieu d'un séparateur de calcul automatique. En général, une *virgule* (",").

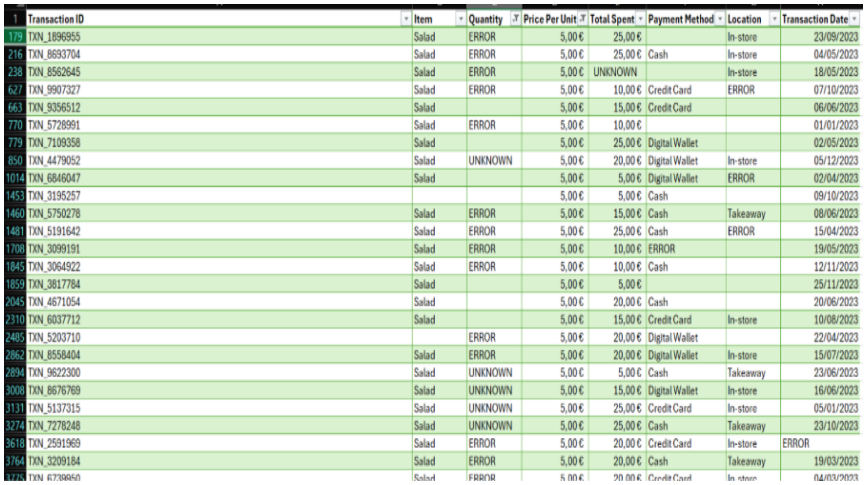
Afin de nettoyer cela, pour chaque colonne, j'ai transformé les *points* en virgules (ctrl+h ; *rechercher*, *remplacer*).

Enfin, j'ai appliqué le bon format aux colonnes (TEXT, DATE, MONETAIRE...)

A	B
Item	Price
Salad	5,00 €
Smoothie	4,00 €
Sandwich	4,00 €
Cake	3,00 €
Juice	3,00 €
Coffee	2,00 €
Tea	1,50 €
Cookie	1,00 €

J'ai créé un catalogue avec les items et leur prix. Ce qui sera utile pour retrouver des items dans la table si le prix choisit est unique (Exemple : Tea).

Certaines cellules contiennent "ERROR", "UNKNOWN" ou sont vides.



	Transaction ID	Item	Quantity	Price Per Unit	Total Spent	Payment Method	Location	Transaction Date
179	TXN_1896955	Salad	ERROR	5,00 €	25,00 €		In-store	23/09/2023
216	TXN_8693704	Salad	ERROR	5,00 €	25,00 €	Cash	In-store	04/05/2023
238	TXN_8562645	Salad	ERROR	5,00 €	UNKNOWN		In-store	18/05/2023
627	TXN_9907327	Salad	ERROR	5,00 €	10,00 €	Credit Card	ERROR	07/10/2023
663	TXN_9356512	Salad		5,00 €	15,00 €	Credit Card		06/06/2023
770	TXN_5728991	Salad	ERROR	5,00 €	10,00 €			01/01/2023
779	TXN_7109358	Salad		5,00 €	25,00 €	Digital Wallet		02/05/2023
850	TXN_4479052	Salad	UNKNOWN	5,00 €	20,00 €	Digital Wallet	In-store	05/12/2023
1014	TXN_6846047	Salad		5,00 €	5,00 €	Digital Wallet	ERROR	02/04/2023
1457	TXN_3195257			5,00 €	5,00 €	Cash		09/10/2023
1460	TXN_5750278	Salad	ERROR	5,00 €	15,00 €	Cash	Takeaway	08/06/2023
1481	TXN_5191642	Salad	ERROR	5,00 €	25,00 €	Cash	ERROR	15/04/2023
1708	TXN_3099191	Salad	ERROR	5,00 €	10,00 €	ERROR		19/05/2023
1845	TXN_3064922	Salad	ERROR	5,00 €	10,00 €	Cash		12/11/2023
1859	TXN_3817784	Salad		5,00 €	5,00 €			25/11/2023
2045	TXN_4671054	Salad		5,00 €	20,00 €	Cash		20/06/2023
2310	TXN_6037712	Salad		5,00 €	15,00 €	Credit Card	In-store	10/08/2023
2485	TXN_5203710		ERROR	5,00 €	20,00 €	Digital Wallet		22/04/2023
2862	TXN_8558404	Salad	ERROR	5,00 €	20,00 €	Digital Wallet	In-store	15/07/2023
2894	TXN_9622300	Salad	UNKNOWN	5,00 €	5,00 €	Cash	Takeaway	23/06/2023
3008	TXN_8676769	Salad	UNKNOWN	5,00 €	15,00 €	Digital Wallet	In-store	16/06/2023
3131	TXN_5137315	Salad	UNKNOWN	5,00 €	25,00 €	Credit Card	In-store	05/01/2023
3274	TXN_7278248	Salad	UNKNOWN	5,00 €	25,00 €	Cash	Takeaway	23/10/2023
3618	TXN_2591969	Salad	ERROR	5,00 €	20,00 €	Credit Card	In-store	ERROR
3764	TXN_3209184	Salad	ERROR	5,00 €	20,00 €	Cash	Takeaway	19/03/2023
3778	TXN_6739860	Salad	ERROR	5,00 €	20,00 €	Credit Card	In-store	04/03/2023

Je les ai normalisées en appelant ces cas “No information” quand la colonne est en format TEXT, “0” quand la colonne est en format NOMBRE, MONETAIRE, DATE. En format DATE, Excel transformera automatiquement la valeur en “00/01/1900”.

Ensuite, j’ai appliqué les calculs possibles sur les colonnes de prix, total ou quantité. De même, lorsque le prix est unique pour un item, j’attribue le nom de l’item.

Visuel et ergonomie

Sur ce travail, il était intéressant d’avoir d’abord une vision globale des résultats (dans ce cas-ci, une seule année : 2023) puis de détaillé la performance, l’activité jour par jour avec des moyennes.

Dans la feuille “Activity average per day”, les deux tableaux sont colorés avec des graduations de couleurs. Bleu : valeur maximale, orange valeur minimale.

La dernière colonne avec une mise en forme par des icônes flèches, permet de comprendre l’activité en progrès ou en déclin de nos produits ou de nos services.

Cette vision est un exemple pour les métiers actionnaires, investisseurs afin de prendre une nouvelle décision sur le prix proposés à leur produit.

Enfin, nous avons relevés beaucoup d’anomalies au fur et à mesure du nettoyage (des paiements non-aboutis, des informations qui n’ont pas été reportées). Le pourcentage de ces anomalies est mince, mais il est important de les recenser afin de pouvoir cibler le problème et d’amener des actions correctives.