Explore "Data Profiling"

Types de données : Confirmé.
Valeurs Manquantes (Nulls) : Très important ! Note les colonnes avec des pourcentages élevés de nulls.
Valeurs Uniques : Utile pour les colonnes catégorielles (ex: StoreType doit avoir 4 valeurs uniques).
Valeurs Max/Min, Moyenne, Médiane : Pour les colonnes numériques.
Barres de distribution : Pour visualiser rapidement la répartition des valeurs.

Sales

Value Statistics

Grouped Values

0-41,551

Max	41,551
Min	0
Lower quartile	3,727.776
Upper quartile	7,855.856
Average	5,773.8189
Standard deviation	3,849.9261
Sum	5,873,180,623
Median	5,743.895
Variance	14,821,931.5547

Max	41,551
Min	0
Lower quartile	3,727.776
Upper quartile	7,855.856
Average	5,773.8189
Standard deviation	3,849.9261
Sum	5,873,180,623
Median	5,743.895
Variance	14,821,931.5547

Top Values

0	172,871
5,674	215
5,558	197
5,483	196
6,049	195

995 more >

1. Observations Clés Immédiates :

A. il y a beaucoup de 0 dans les ventes. Cela représente **172 871 occurrences** ! C'est un point crucial.
B. **Max = 41,551 :** C'est la valeur de vente la plus élevée enregistrée pour un jour et un magasin donnés.
C. **Min = 0 :** La valeur de vente la plus basse.
D. **Lower quartile (Q1) = 3,727.776 :** 25% des ventes sont inférieures ou égales à 3 727,776 .
E. **Upper quartile (Q3) = 7,855.856 :** 75% des ventes sont inférieures ou égales à 7 855,856. Inversement, 25% des ventes sont supérieures à cette valeur.
F. **Average (Moyenne) = 5,773.8189 :** C'est la somme de toutes les ventes divisée par le nombre total de jours-magasins. La moyenne est fortement influencée par les valeurs extrêmes (très hautes) et par le grand nombre de zéros.
G. **Standard deviation (Écart-type) = 3,849.9261 :** C'est une mesure de la dispersion des données autour de la moyenne. Un écart-type élevé par rapport à la moyenne indique que les données sont très dispersées. Ici, il est assez élevé (environ 66% de la moyenne), ce qui confirme que les ventes varient beaucoup d'un jour à l'autre et d'un magasin à l'autre.
H. **Sum = 5,873,180,623 :** La somme totale de toutes les ventes enregistrées dans le dataset.
I. **Median = 5,743.895 :** C'est la valeur centrale lorsque toutes les ventes sont triées. 50% des ventes sont en dessous et 50% sont au-dessus de cette valeur.

Pourquoi la médiane est importante ici : Puisque la distribution est asymétrique à droite, la moyenne (5 773) est légèrement supérieure à la médiane (5 743). Si la distribution était parfaitement symétrique, la moyenne et la médiane seraient très proches. Si la moyenne était beaucoup plus grande que la médiane, cela indiquerait la présence de très grandes valeurs (outliers positifs) qui "tire" la moyenne vers le haut. Ici, la différence n'est pas énorme, mais elle confirme l'asymétrie.

Variance = 14,821,931.5547 : C'est le carré de l'écart-type. Elle mesure aussi la dispersion des données par rapport à la moyenne. Une grande variance indique une grande variabilité. En général, on utilise l'écart-type car il est dans la même unité que les données (ici, des Euros/Dollars de ventes).

2.Raisonnement de Consultant

Le Problème des Ventes à Zéro :

- **Hypothèse majeure :** Les jours avec Sales = 0 correspondent très probablement aux jours où les magasins sont **fermés (Open = 0)**.
- **Implication pour la prédiction :** Si nous voulons prédire les ventes des jours où le magasin est ouvert, il faudra **filtrer ces jours Sales = 0 (et Open = 0)** pour l'entraînement de notre modèle. Sinon, le modèle sera biaisé et tentera de prédire des zéros là où il n'y en a pas.

Variabilité des Ventes :L'écart-type élevé et la distribution asymétrique indiquent que les ventes sont très variables. Cela signifie que notre modèle devra être robuste pour capturer cette hétérogénéité et les pics de ventes.

Facteurs potentiels de variabilité : Promotions (Promo, Promo2), jours fériés (StateHoliday, SchoolHoliday), jours de la semaine (DayOfWeek), distance à la concurrence (CompetitionDistance), type de magasin (StoreType), etc. Notre modèle devra absolument intégrer ces variables.

Complexité de la Prédiction : Prédire des ventes qui varient de 0 à plus de 41 000 sera un défi. Cela souligne l'importance du Feature Engineering (créer de nouvelles variables) et de modèles avancés.

Outil Summarize (Résumé) :

Record	StoreType	Count
1	d	312,912
2	a	551,627
3	b	15,830
4	c	136,840

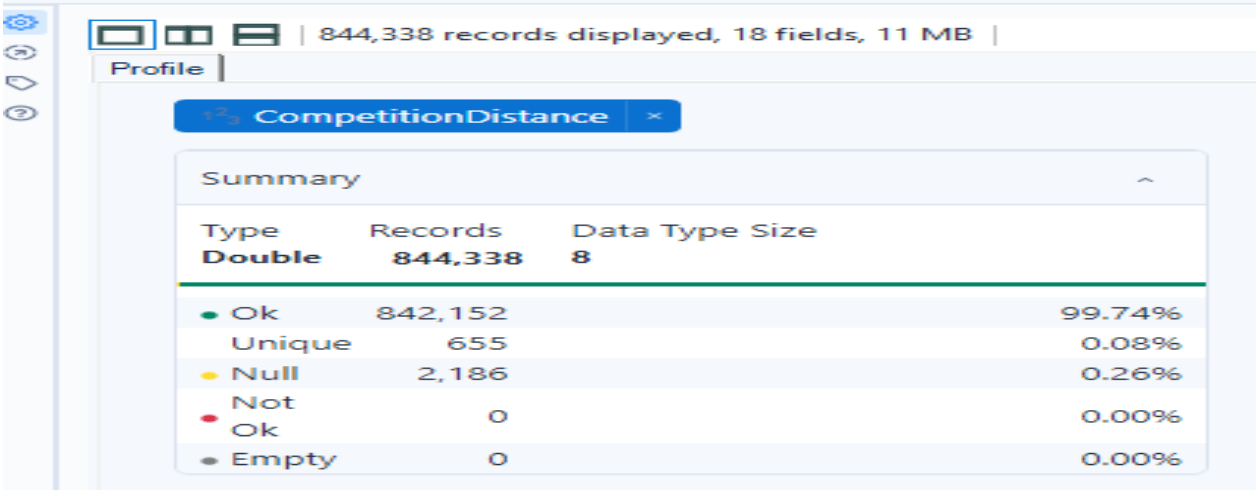
Record	StateHoliday	Count
1	0	986,159
2	a	20,260
3	b	6,690
4	c	4,100

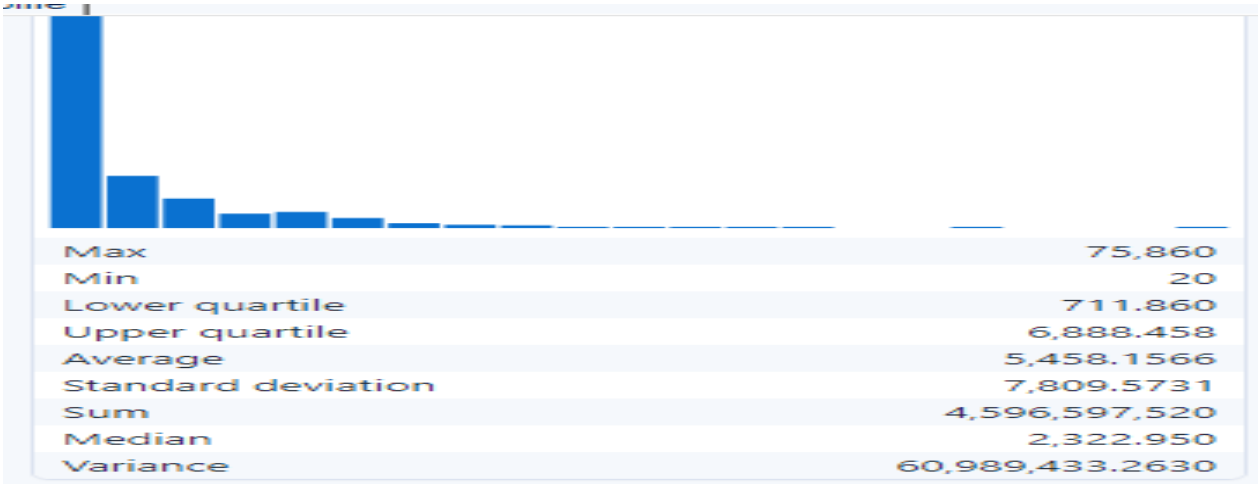
Outil Filter (Filtre) :

Open == 0 (magasins fermés)-->Présence des valeurs nulles(172817/172871)
Open == 1 (magasins fermés)-->Présence des valeurs nulles????????(54)

Informations Manquantes Spécifiques :

1-CompetitionDistance

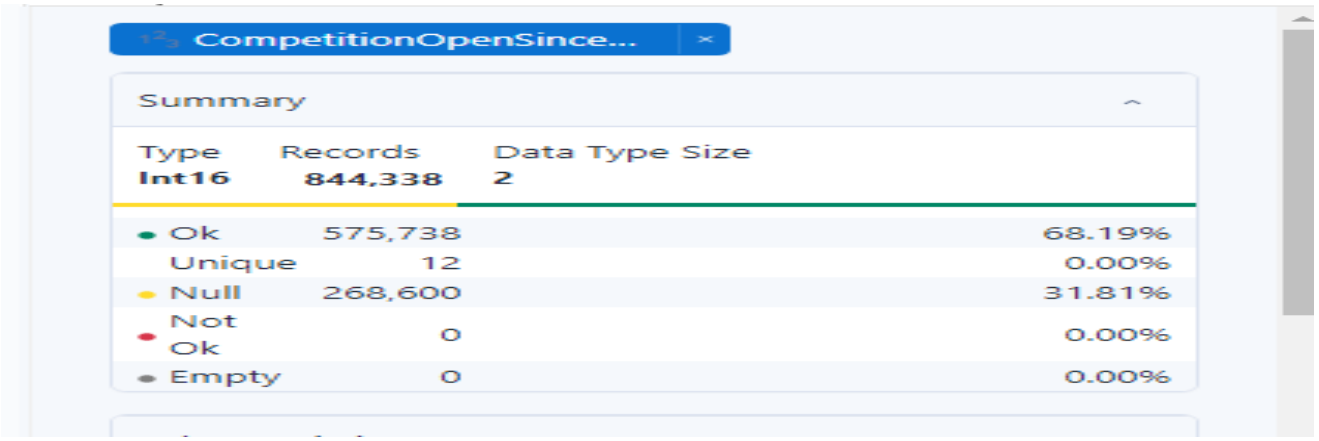


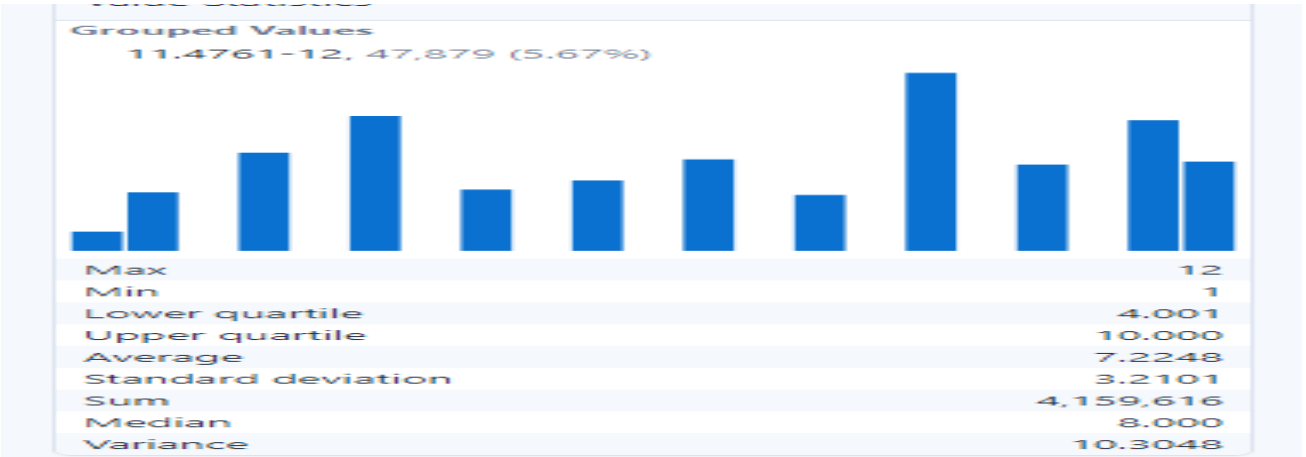


Max	75,860
Min	20
Lower quartile	711.860
Upper quartile	6,888.458
Average	5,458.1566
Standard deviation	7,809.5731
Sum	4,596,597,520
Median	2,322.950
Variance	60,989,433.2630

- **Colonne :**CompetitionDistance
- **Observations :** Faible pourcentage de Nulls (0.26%). Distribution asymétrique avec des distances variées.
- **Interprétation des Nulls :** Signifie "pas de concurrent proche/mesurable".
- **Décision de Nettoyage :** Imputation des Nulls par la valeur constante de **100 000**.
- **Justification :** Permet au modèle de traiter l'absence de concurrence comme une information distincte et pertinente, évitant de fausser la moyenne/médiane.

CompetitionOpenSinceMonth





Relation avec CompetitionDistance :

- **Question cruciale :** Est-ce que les Nulls dans CompetitionOpenSinceMonth (et CompetitionOpenSinceYear) correspondent aux mêmes lignes où CompetitionDistance était Null ? Ou bien sont-ils indépendants ?
- **Si liés :** Si CompetitionDistance est Null (pas de concurrent), alors il est logique que CompetitionOpenSinceMonth/Year soit Null aussi. Dans ce cas, les 0.26% de Nulls de CompetitionDistance seraient inclus dans les 31.81% de Nulls de CompetitionOpenSinceMonth/Year.
- **Si indépendants :** Cela signifierait qu'il y a un concurrent (CompetitionDistance est présent), mais que sa date d'ouverture est inconnue. C'est une information différente.

Quand CompetitionDistance est Null, alors CompetitionOpenSinceMonth et CompetitionOpenSinceYear sont également Null à 100%.
Remplacer Null pour CompetitionDistance par 100000 Et Competition Month/Year par 0

Promo2

Quand Promo2 == 0 (False), alors Promo2SinceWeek, Promo2SinceYear, et PromoInterval sont systématiquement Null.

Synthèse de Toutes les Imputations (Nettoyage des Nulls terminé !)

CompetitionDistance : Imputer les Nulls par **100000**

CompetitionOpenSinceMonth : Imputer les Nulls par **0**

CompetitionOpenSinceYear : Imputer les Nulls par **0**

Promo2SinceWeek : Imputer les Nulls par **0**

Promo2SinceYear : Imputer les Nulls par **0**

PromoInterval : Imputer les Nulls par **""** (chaîne vide)