Rapport sur le Traitement des Données

FADEL Farah TAHIRI Abdelilah

21 mars 2024

Analyse des Données

Nous commençons par une analyse des données pour obtenir des informations telles que le nombre total d'enregistrements, le nombre de lignes et de colonnes, ainsi qu'un résumé statistique des données.

Traitement des Données

Dans cette étape, nous effectuons plusieurs transformations sur nos données. Tout d'abord, nous convertissons le type de la colonne "Date_vente" de "objets" à "date" pour une utilisation ultérieure. En outre, pour le reste des colonnes, elles sont déjà dans le type de données adéquat pour une utilisation ultérieure. Ensuite, nous normalisons les données afin de les mettre à la même échelle, ce qui facilite la comparaison entre les différentes variables. Ensuite, nous traitons les valeurs manquantes en remplaçant les valeurs manquantes dans la colonne "Quantite_vendue" par la moyenne des valeurs existantes, en supprimant les lignes avec des valeurs manquantes dans la colonne "Prix_unitaire", et en remplaçant les valeurs manquantes dans la colonne "Nom_produit" par des valeurs uniques "eid1" à "eid10". Nous affichons également le nombre de valeurs manquantes par colonne pour surveiller l'évolution du nettoyage des données.

Gestion des Valeurs Aberrantes

Dans cette étape, nous détectons et gérons les valeurs aberrantes dans la colonne "Quantite_vendue". Nous identifions d'abord le nombre de valeurs aberrantes où la quantité vendue dépasse un seuil défini (par exemple, 50, représentant la quantité maximale généralement vendue dans le magasin). Ensuite, nous remplaçons ces valeurs aberrantes par la médiane de la colonne "Quantite_vendue".

Détection et Suppression des Doublons

la variable (data avant suppression) stocke le nombre d'enregistrements dans le DataFrame avant la suppression des doublons en utilisant la fonction len(df). Ensuite, la méthode (drop duplicates()) est utilisée pour supprimer les doublons du DataFrame, et cela est fait directement sur le DataFrame en utilisant l'argument inplace=True. Après la suppression des doublons, la variable (data apres suppression) stocke le nombre d'enregistrements restants dans le DataFrame. Enfin, les deux nombres sont imprimés pour comparer le nombre d'enregistrements avant et après la suppression des doublons.

Transformation Supplémentaire

Enfin, nous effectuons une transformation supplémentaire en ajoutant une nouvelle colonne "Montant_total" en agrégeant les colonnes "Prix_unitaire" et "Quantite_vendue".