Lab01 : Préparer des données dans Power BI Desktop

Dans ce labo, vous commencez le développement d'une solution Power BI Desktop pour la société Adventure Works. Elle implique la connexion aux données sources, l'affichage de l'aperçu des données et l'utilisation de techniques d'aperçu des données pour comprendre les caractéristiques et la qualité des données sources.

Dans ce labo, vous découvrez comment :

- Ouvrir Power BI Desktop
- Définir les options de Power BI Desktop
- Se connecter à des données sources
- Obtenir un aperçu des données sources
- Utilisez des techniques d'aperçu des données pour mieux comprendre les données

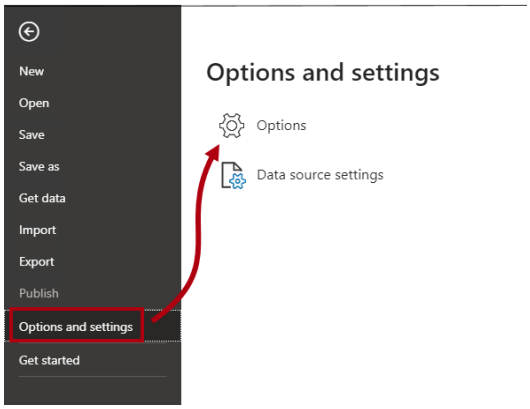
Préparer les données

Dans cet exercice, vous allez créer huit requêtes (tables) dans Power BI Desktop à partir de fichiers TXT et CSV. On va commencer par ouvrir Power BI desktop et faire un tour sur l'interface graphique pour bien comprendre l'organisation de l'outil et puis enregistrer le fichier Power BI Desktop à la fin de notre travail.

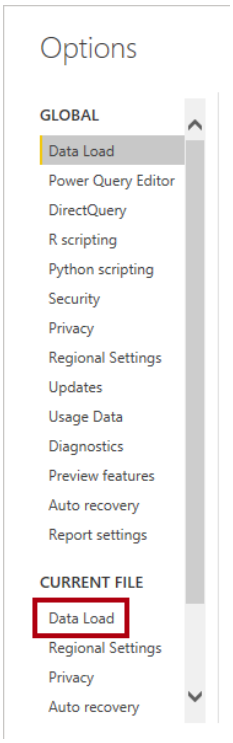
Remarque : Ne pas oublier d'enregistrer au fur et à mesure l'état d'avancement de votre travail en utilisant ctr+S.

On va commencer par définir les options de Power BI Desktop.

1. Dans Power BI Desktop, sélectionnez l'onglet de ruban Fichier.
2. Sur la gauche, sélectionnez Options et paramètres, puis Options.

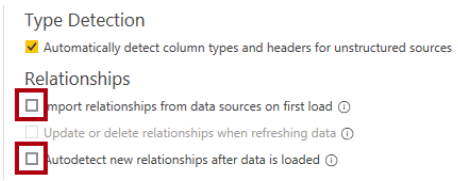


3. Dans la fenêtre Options, à gauche, dans le groupe Fichier actif, sélectionnez Chargement des données.



Les paramètres de Chargement des données pour le fichier actif permettent de définir des options qui déterminent les comportements par défaut lors de la modélisation.

4. Dans le groupe Relations, dé-sélectionnez les deux options sélectionnées comme indiqué sur la figure suivante.



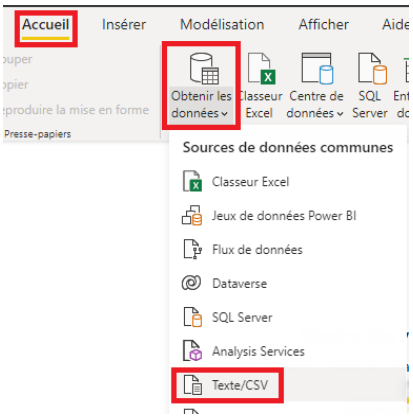
Même si ces deux options peuvent être utiles lors du développement d'un modèle de données, elles ont été désactivées pour prendre en charge l'expérience du projet qui va nous permettre de définir par nous même les relations entre les tables. Au moment de créer des relations dans les autres exercices, vous découvrez pourquoi vous ajoutez chacune d'elles.

5. Sélectionnez OK.
6. Enregistrez le fichier Power BI Desktop.

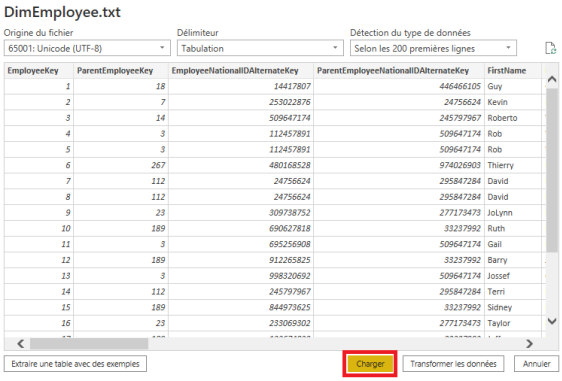
Obtenir des données depuis des fichiers TXT

Dans cette partie, vous allez créer des requêtes (tables) basées sur des fichiers TXT qui vont contenir les tables de dimensions et la table de fait et qu'on va utiliser dans notre analyse pour le projet.

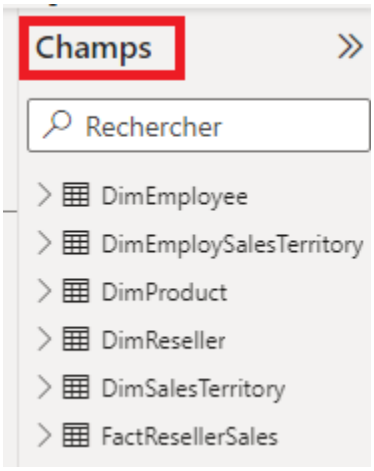
7. Sous l'onglet de ruban Accueil, dans le groupe Données, sélectionnez Obtenir les données > Text/CSV.



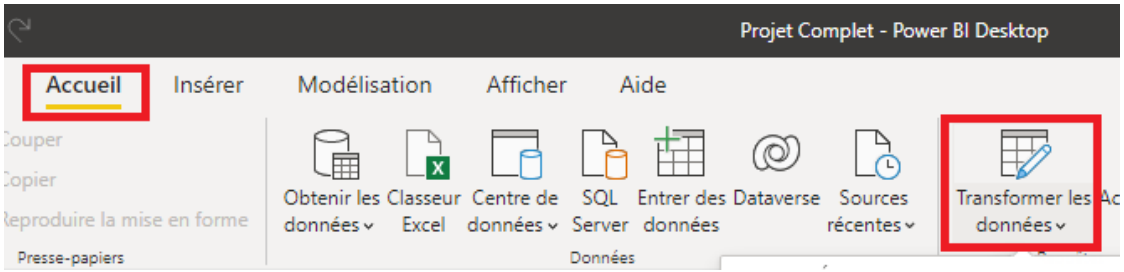
8. Dans la fenêtre qui s'ouvre allez dans le répertoire où vous avez stocké les fichiers de données à utiliser.
9. Ajouter les tables suivantes dans Power BI :
   - DimEmployee
   - DimEmployeeSalesTerritory
   - DimProduct
   - DimReseller
   - DimSalesTerritory
   - FactResellerSales



10. Vérifier qu'au niveau de l'onglet Champs vous avez l'affichage suivant sur les tables.

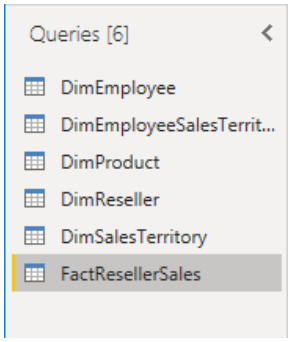


11. Dans le menu Acceuil, dans la section Requêtes sélectionner Transformer les données ce qui permet d'ouvrir la fenêtre de l'Éditeur Power Query.

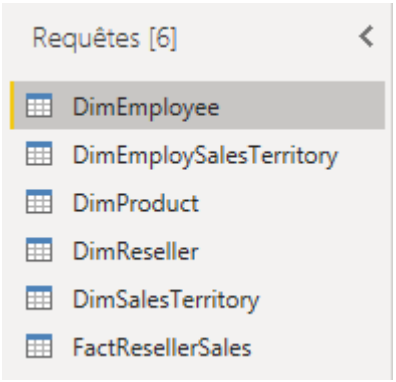


Dans cette première partie du projet, on ne vas pas transformer les données. L'objectif est d'explorer et de profiler les données dans la fenêtre Éditeur Power Query. On va afficher un aperçu des données importées depuis les fichiers txt. On va aussi découvrir des informations pertinentes sur les données qui nous permettent d'identifier la *qualité des colonnes*, la *distribution des colonnes* et les *outils de profilage de colonne* pour comprendre les données et évaluer leur pertinence.

12. A gauche, notez la présence du volet Requêtes qui contient une requête pour chaque table que nous avons importée.

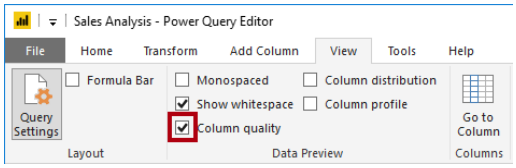


13. Sélectionnez la requête DimEmployee.



La table DimEmployee stocke une ligne pour chaque employé. Elle est composée par 27 colonnes qui contiennent des informations pertinentes pour les employés qui représntent les commerciaux.

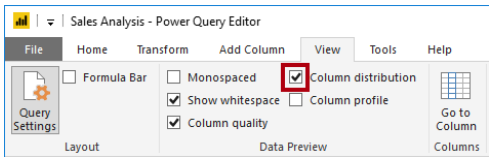
14. En bas à gauche, dans la barre d'état, notez les statistiques de la table : elle contient 27 colonnes et 296 lignes.
15. Pour évaluer la qualité des colonnes, sous l'onglet de ruban Affichage, dans le groupe Aperçu des données, sélectionnez Qualité de la colonne.



La qualité de la colonne vous permet de déterminer facilement le pourcentage de valeurs valides, en erreur ou vides. Ceci peut nous renseigner sur la qualité des données disponibles au niveau de la table.

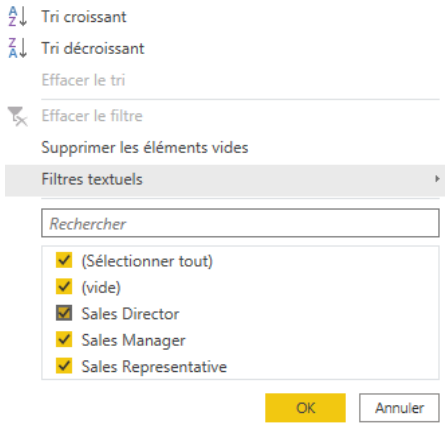
16. Pour la colonne Position (dernière colonne), remarquez que 94 % des lignes sont vides (null) ce qui ne correspond pas à une bonne qualité des données sur cette colonne.

17. Pour évaluer la distribution des colonnes, sous l'onglet de ruban Affichage, dans le groupe Aperçu des données, sélectionnez Distribution des colonnes.

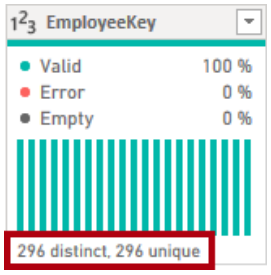


18. Examinez à nouveau la colonne Position et déterminez le nombre de valeurs distinctes et le nombre de valeurs uniques pour cette colonne.

19. En utilisant l'option de Filtre sur les données, déterminer les valeurs distinctes que vous avez au niveau de cette colonne.

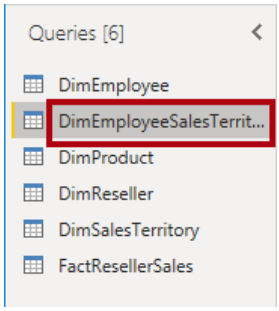


20. Quelle est l'information qui correspond à une valeur unique au niveau de la colonne Position.
21. Passez en revue la distribution des données pour la première colonne EmployeeKey : il y a 296 valeurs distinctes et 296 valeurs uniques.



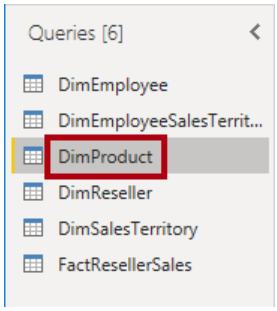
Quand les nombres de valeurs distinctes et uniques sont identiques, cela signifie que la colonne contient des valeurs uniques. Lors de la modélisation, il est important que certaines tables contiennent des colonnes uniques. Vous pouvez utiliser ces colonnes uniques pour créer des relations un-à-plusieurs.

22. La table contient deux colonnes, FirstName et LastName, on va voir dans la partie réservée aux transformations qu'il est possible de les associer pour créer une nouvelle colonne qui contient à la fois les deux informations.
23. Dans le volet Requêtes, sélectionnez la requête DimEmployeeSalesTerritory.

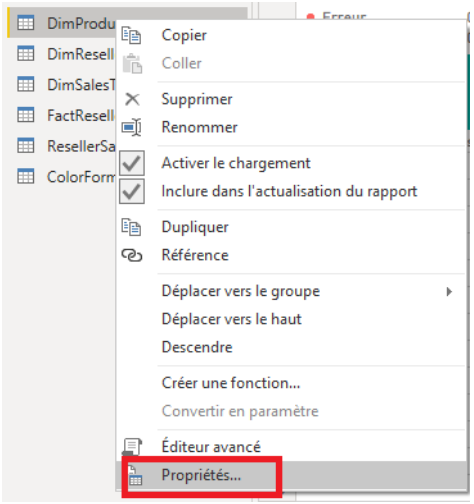


La table DimEmployeeSalesTerritory stocke une ligne pour chaque employé identifié par son EmployeeKey et les régions du secteur de vente qu'il gère. La table peut faire correspondre plusieurs régions pour un même employé. Certains employés gèrent une, deux ou éventuellement plusieurs régions. Lorsque vous modélisez ces données, vous devez définir une relation plusieurs-à-plusieurs.

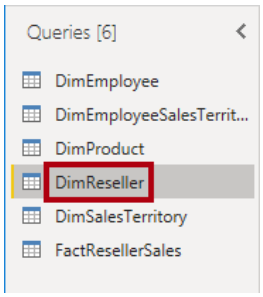
24. Dans la table DimEmployeeSalesTerritory et en utilisant la colonne EmployeeKey utiliser les filtres sur la colonne et le tri pour identifier la clé des employés qui ne sont opérationnels que dans une seule région.
25. Donner le numéro des régions en utilisant la colonne SalesTerritoryKey où un seul commercial est actif.
26. Dans le volet Requêtes, sélectionnez la requête DimProduct.



27. Pour ajouter une description de la table DimProduct, sélectionner dans le menu contextuel de la table DimProduct l'option Propriétés puis ajouter un descriptif de votre table dans la partie Description.

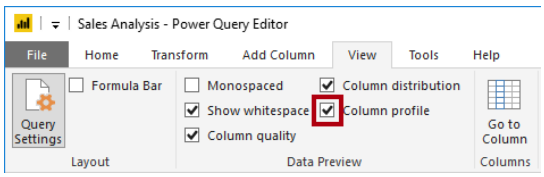


28. Dans le volet Requêtes, sélectionnez la requête DimReseller.

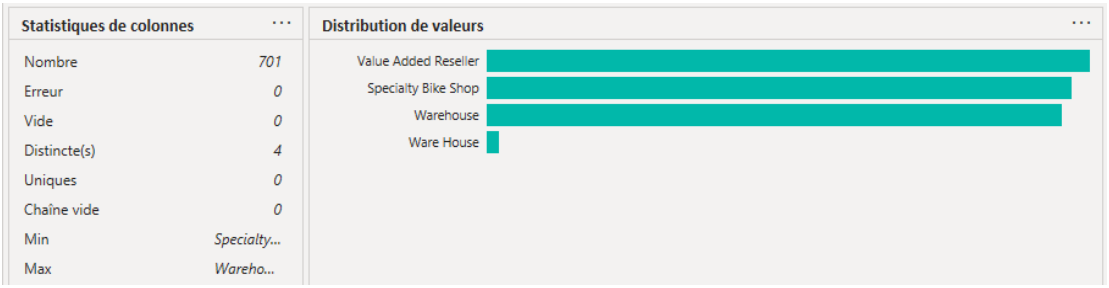


La table DimReseller contient une ligne par revendeur. Les revendeurs vendent, distribuent ou apportent de la valeur ajoutée aux produits Adventure Works.

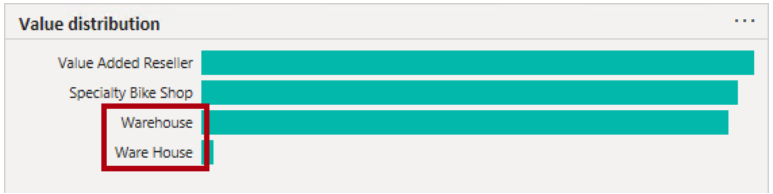
29. Pour visualiser les valeurs des colonnes, sous l'onglet de ruban Affichage, dans le groupe Aperçu des données, sélectionnez Profil de colonne.



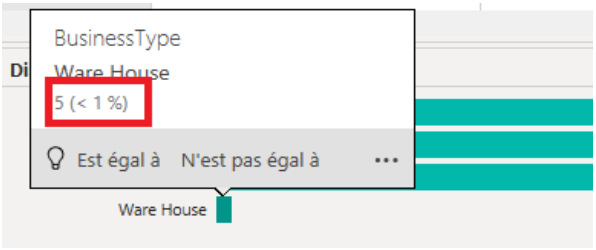
30. Sélectionnez l'en-tête de colonne BusinessType.
31. Notez qu'un nouveau volet s'ouvre sous le volet Aperçu des données.



32. Examinez les statistiques dans le volet Statistiques de colonnes et la distribution des valeurs de la colonne BusinessType dans le volet Distribution des valeurs.
33. Déterminer le nombre de lignes qu'on peut avoir pour la table.
34. Donner les valeurs distinctes présentes dans cette colonne.
35. Notez le problème de qualité des données : il existe deux étiquettes pour l'entrepôt (Warehouse et Ware House, cette dernière étant mal orthographiée).

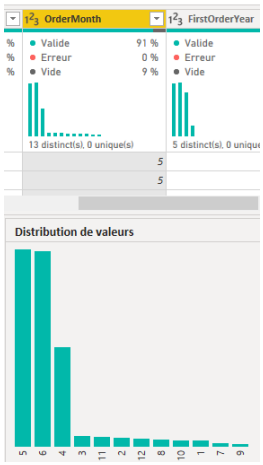


36. Placez le curseur sur la barre Ware House ; notez qu'il y a cinq lignes avec cette valeur.

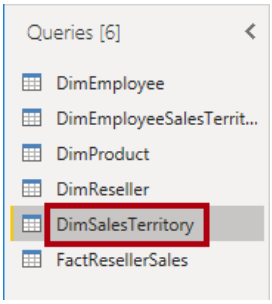


On va voir dans la partie réservée aux transformations qu'on peut faire sur les tables et les colonnes comment ré-étiqueter ces cinq lignes pour les corriger et préserver la qualité des données.

37. Dans la table DimReseller, déterminer pour la colonne OrderMonth les trois mois où on a le plus de ventes en utilisant la distribution des données sur la colonne.



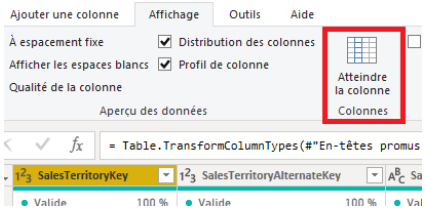
38. En se basant sur la colonne ResellerName dans la table DimReseller, déterminer le nombre de revendeurs uniques dans la table.
39. Déterminer les deux resellers qui ont été dupliqués dans cette table.
40. Identifier les modalités qu'on peut avoir au niveau de la colonne ProductLine.
41. Identifier pour chaque modalité le nombre d'instance disponibles en utilisant la distribution des données.
42. Sur la colonne AnnualSales, déterminer la valeur maximale et minimale des ventes effectuées en utilisant les Statistiques de colonne.
43. Dans le volet Requêtes, sélectionnez la requête DimSalesTerritory.



La table DimSalesTerritory contient une ligne par région commerciale, y compris Corporate HQ (siège social de l'entreprise). Les régions sont affectées à un pays, et les

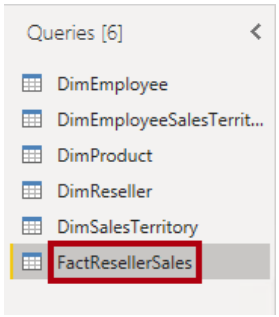
pays sont affectés à des groupes. On va voir dans la partie du projet relative aux transformations sur les données comment créer une hiérarchie pour prendre en charge l'analyse au niveau de la région, du pays ou du groupe.

44. Pour atteindre la colonne SalesterritoryRegion sélectionner la première colonne puis au niveau du volet Affichage et dans la partie Colonnes choisir Atteindre la colonne et sélectionner le nom de la colonne SalesterritoryRegion.



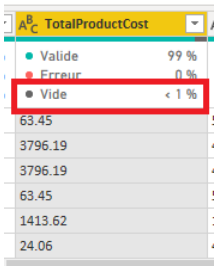
Cette méthode nous permet d'atteindre une colonne spécifique dans une table surtout lorsque le nombre des colonnes est important.

45. Dans le volet Requêtes, sélectionnez la requête FactResellerSales.



La table FactResellerSales contient une ligne par ligne de commande client où une commande client contient une ou plusieurs lignes et elle va représenter la table des faits de notre modèle.

46. Examinez la qualité de données pour la colonne TotalProductCost et notez que <1 % des lignes sont vides.



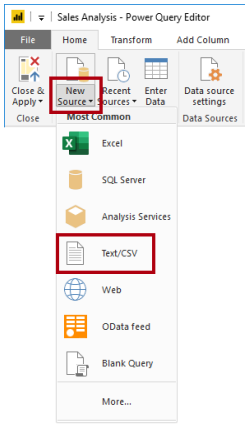
Les valeurs manquantes de la colonne TotalProductCost sont un problème de qualité des données. Dans la partie du projet réservée aux transformations sur les données, on va appliquer des transformations pour renseigner les valeurs manquantes en utilisant le coût StandardCost du produit, qui est stocké dans la table DimProduct.

47. Quelle est la différence entre les deux colonnes TotalProductCost et SalesAmount ?
48. Comment on peut calculer les valeurs des ventes disponibles au niveau de la colonne SalesAmount ?

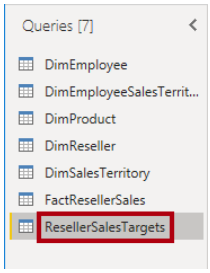
Obtenir des données d'un fichier CSV

Dans cette partie, vous allez créer une requête basée sur un fichier CSV.

49. Pour ajouter une nouvelle requête, dans la fenêtre Éditeur Power Query, sous l'onglet de ruban Accueil, dans le groupe Nouvelle requête, sélectionnez la flèche vers le bas Nouvelle source, puis sélectionnez Texte/CSV.



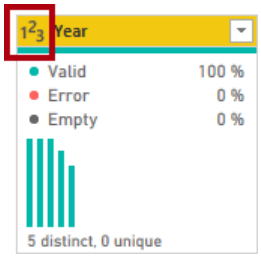
50. Chercher le fichier ResellerSalesTargets.csv depuis votre machine locale puis choisir Ouvrir.
51. Dans la fenêtre ResellerSalesTargets.csv, examinez l'aperçu des données.
52. Sélectionnez OK.
53. Dans le volet Requêtes, notez l'ajout de la requête ResellerSalesTargets.



Le fichier CSV ResellerSalesTargets contient une ligne par vendeur, par année. Chaque ligne enregistre 12 objectifs de ventes mensuels (exprimés en milliers). L'année fiscale de la société AdventureWorks commence le 1er juillet. Quand il n'y a pas d'objectif de ventes mensuel, un caractère de trait d'union est stocké à la place.

54. Est-ce qu'il y a des colonnes vides pour cette table ? Comment le vérifier ?
55. Passez en revue les icônes dans chaque en-tête de colonne, à gauche du nom de la colonne. On ne vous demande pas à ce niveau de modifier les types de données.
56. Citer les différents types qui ont été détectés par Power BI.

57. Pourquoi les données au niveau de la colonne M01 ont été déclarées comme des données de type texte alors qu'elles contiennent des nombres ?

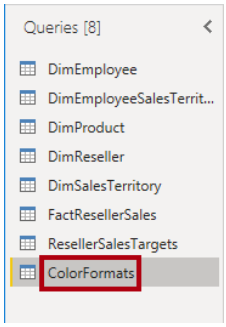


Les icônes représentent le type de données de la colonne. 123 représente un nombre entier et ABC représente du texte.

Obtenir des données supplémentaires d'un fichier CSV

Dans cette tâche, vous allez créer une requête supplémentaire basée sur un autre fichier CSV.

58. Effectuez les étapes de la tâche précédente pour créer une requête (table) basée sur le fichier ColorFormats.csv.



Le fichier CSV ColorFormats contient une ligne par couleur de produit. Chaque ligne enregistre les codes hexadécimaux pour mettre en forme les couleurs d'arrière-plan et de police.

59. Quelles sont les colonnes qu'il est possible de définir comme en-têtes pour cette table ?
60. Pourquoi l'information relative à la couleur d'un produit n'a pas été ajoutée directement au niveau de la table DimProduct au lieu de créer une nouvelle table ?

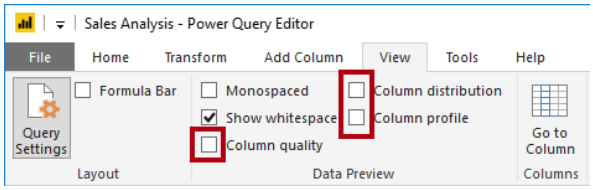
Terminer

Dans cette tâche, vous allez terminer le labo.

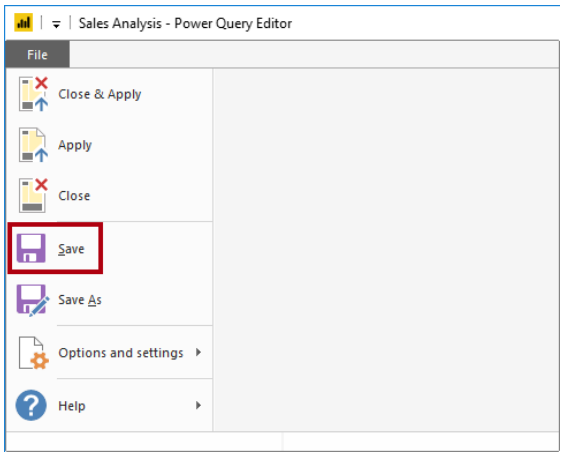
61. Sous l'onglet de ruban Afficher, dans le groupe Aperçu des données, désélectionnez les trois options d'aperçu des données :

    - Qualité de la colonne

- Distribution des colonnes
- Profil de colonne



62. Pour enregistrer le fichier Power BI Desktop, dans le mode Backstage Fichier, sélectionnez Enregistrer.



63. De retour sur le Power BI Desktop, vérifier que dans le volet Champs vous avez 08 tables.