**Rapport de Projet : Nettoyage des Données Immobilier – Compétences SQL pour les Analystes de Données**

**Contexte du Projet**

Ce projet a pour objectif de consolider les compétences en **nettoyage de données** à travers l'utilisation du langage SQL. Il est basé sur un jeu de données immobilier provenant de ventes de propriétés à Nashville. Le processus de nettoyage permet d'améliorer la qualité des données, étape indispensable pour garantir des analyses fiables et précises. Les tâches incluent la normalisation des formats de date et d'adresse, la gestion des valeurs manquantes, l'élimination des doublons, ainsi que la suppression des colonnes inutiles.

**Objectifs du Projet**

* **Standardiser les formats de date** pour homogénéiser les données temporelles.
* **Compléter les valeurs manquantes**, notamment les adresses des propriétés, à partir d’informations partiellement disponibles.
* **Uniformiser les données** dans les colonnes qui présentent des variations dans les réponses (par exemple, ‘Y’ et ‘Yes’).
* **Supprimer les doublons** afin de ne conserver que des enregistrements uniques et pertinents.
* **Élaguer les colonnes inutilisées** afin d'optimiser la base de données pour une analyse future.

**Étapes Clés et Démarche**

**1. Analyse Initiale des Données**

L'analyse commence par une vérification des premières lignes de la base de données. Cela permet de comprendre la structure des données et d'identifier les premières anomalies, comme des valeurs manquantes ou des incohérences dans certains champs (dates, adresses, etc.).

**2. Standardisation des Dates**

Une colonne dédiée a été ajoutée pour convertir et standardiser les dates au format souhaité. Cette étape est essentielle pour permettre des analyses temporelles cohérentes, facilitant par exemple l’étude des tendances dans les ventes de propriétés à travers le temps.

**3. Gestion des Adresses Manquantes**

Certaines lignes de la base de données n’avaient pas d’adresse pour certaines propriétés. En identifiant les parcelles qui partagent un même identifiant unique (ParcelID), les adresses manquantes ont pu être complétées à partir d’autres enregistrements correspondant à ces mêmes parcelles. Cela a permis d’enrichir les données et d'éviter les biais liés aux informations incomplètes.

**4. Séparation des Composants d'Adresse**

Afin de mieux structurer les données, les adresses des propriétés ont été décomposées en plusieurs colonnes distinctes : **adresse**, **ville**, et **état**. Cette séparation facilite par la suite l'analyse et le filtrage des données, par exemple pour examiner les tendances de vente selon les villes ou états. Une méthode similaire a été appliquée aux adresses des propriétaires.

**5. Uniformisation des Données Booléennes**

Dans certaines colonnes, comme celle indiquant si une propriété a été vendue comme vacante, les réponses prenaient des formes variées (Y, Yes, N, No). Une harmonisation de ces réponses a été effectuée pour ne conserver que des valeurs cohérentes et lisibles dans les analyses.

**6. Suppression des Doublons**

Les enregistrements dupliqués dans les données ont été identifiés à l'aide de critères tels que l'ID de la parcelle, l'adresse de la propriété et le prix de vente. Ces doublons ont ensuite été supprimés pour garantir que chaque transaction immobilière soit unique. Cela évite les distorsions dans les résultats analytiques comme les statistiques de vente ou les moyennes de prix.

**7. Suppression des Colonnes Inutiles**

Pour alléger la base de données et améliorer la lisibilité, plusieurs colonnes inutilisées ou redondantes ont été supprimées. Cette étape optimise la performance de la base et simplifie la visualisation et l’interprétation des données dans des outils analytiques ou lors de futures requêtes SQL.

**Résultats et Impacts**

**1. Normalisation des Dates**

La conversion des dates au format standard a facilité la visualisation des tendances de ventes dans le temps, permettant d'analyser les fluctuations mensuelles ou annuelles des transactions immobilières.

**2. Complétion des Adresses**

En remplissant les adresses manquantes à partir d’informations similaires disponibles dans les données, l’intégrité des enregistrements a été renforcée, permettant une analyse plus complète et sans trous d’information.

**3. Structuration des Adresses**

La décomposition des adresses en colonnes distinctes (adresse, ville, état) a permis d'améliorer la précision des filtrages et des regroupements de données par région, rendant les analyses géographiques plus simples et plus efficaces.

**4. Uniformisation des Valeurs Booléennes**

En normalisant les valeurs booléennes, les résultats sont devenus plus cohérents et plus faciles à interpréter. Cela permet d'éviter les confusions dans les rapports finaux et de garantir une analyse uniforme.

**5. Suppression des Doublons**

L’élimination des doublons a assuré que chaque enregistrement soit unique, ce qui a éliminé les biais dans les calculs, tels que les totaux ou les moyennes de prix de vente.

**6. Nettoyage et Optimisation de la Base de Données**

La suppression des colonnes inutilisées a non seulement simplifié la structure des données, mais elle a également réduit le bruit dans les analyses futures. Cela rend la base de données plus accessible et prête pour des analyses ou des visualisations ultérieures.

**Compétences SQL Appliquées**

Ce projet a mis en lumière plusieurs compétences essentielles pour un analyste de données :

* **Gestion et transformation des chaînes de caractères** : Utilisation de fonctions SQL pour extraire et structurer les adresses, démontrant la capacité à manipuler efficacement des données textuelles complexes.
* **Nettoyage des données et normalisation** : Application des bonnes pratiques de nettoyage de données via des opérations de mise à jour, normalisation des formats de date et correction des valeurs manquantes.
* **Identification et suppression des doublons** : Utilisation de sous-requêtes avancées pour détecter et supprimer les doublons, garantissant l'intégrité des analyses.
* **Optimisation de la base de données** : Simplification des données via la suppression des colonnes inutiles, montrant la capacité à préparer une base de données propre et prête pour l'analyse.

**Conclusion**

Ce projet de nettoyage des données a permis de consolider des compétences clés en **SQL**, nécessaires pour tout analyste de données travaillant avec des données immobilières ou des jeux de données complexes. La qualité des données a été nettement améliorée, rendant la base de données prête à être utilisée dans des analyses approfondies ou des visualisations avancées. Grâce à ces techniques de nettoyage, les analyses futures seront plus fiables et plus pertinentes.