**DATA CLEANING USING SQL**

In this project, I worked on Nashville Housing Data, a dataset about properties located in Nashville, capital city of Tennessee state in US. I used MS SQL Management Server Studio to do the data exploration and cleansing.

**DATASET**:

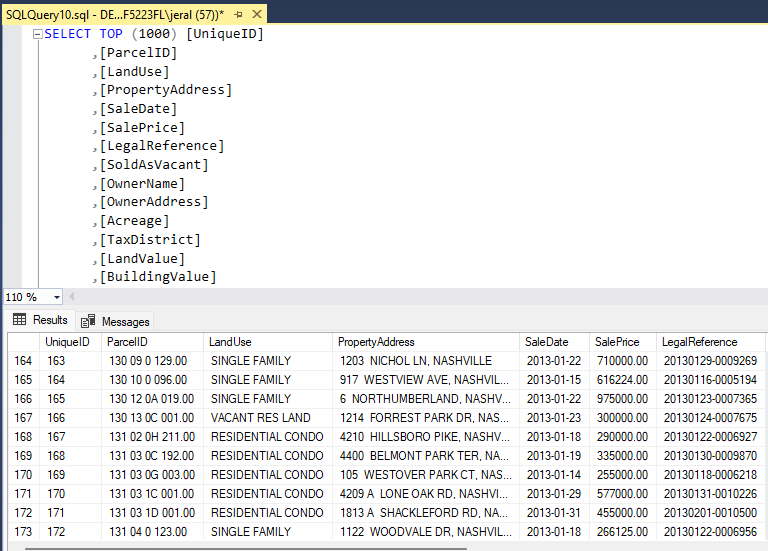
The dataset is composed of 20 columns and 56,477 rows.

Source: <https://www.kaggle.com/datasets/tmthyjames/nashville-housing-data>

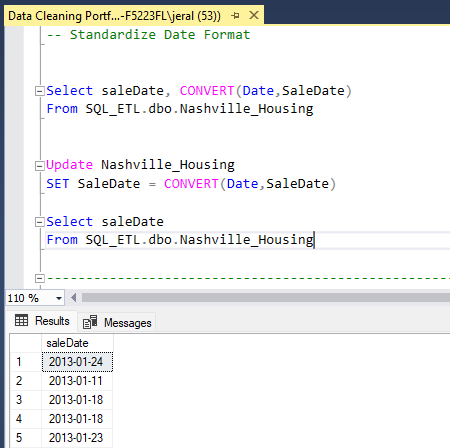
**DATA CLEANING**

On exploring, found that duplicate data has to be removed and unwanted columns. Also the following columns has to be addressed:

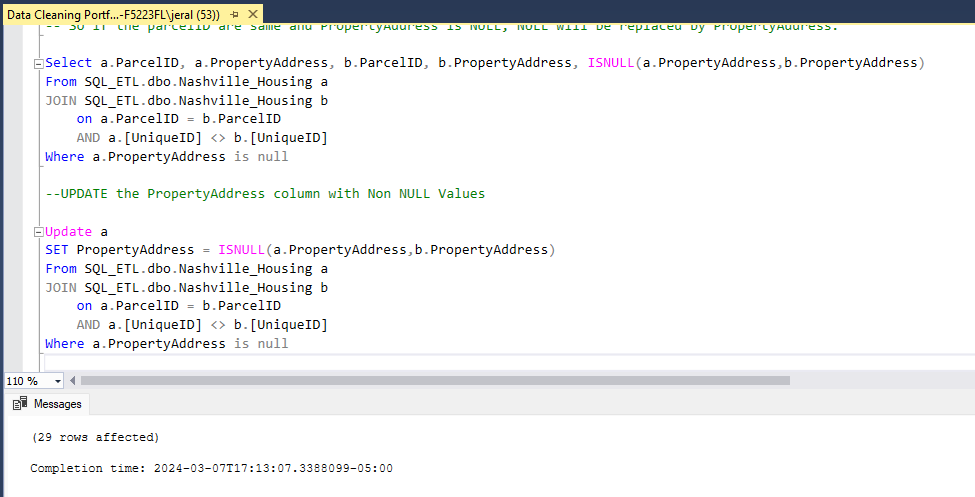
* Sale\_Date - Date is not in the standard format.
* PropertyAddress - NULL values and has both house address and city in the same column.
* OwnerAddress - Column has house address, city and states in the same column.
* SoldAsVacant - Have inconsistent entries such as Y and N instead of Yes and No.



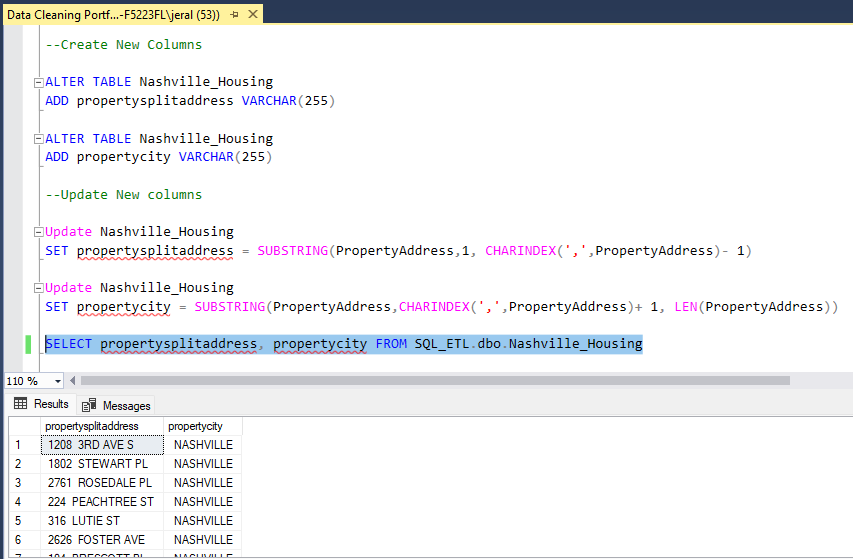
1. SALE\_DATE: Standardize the Date column



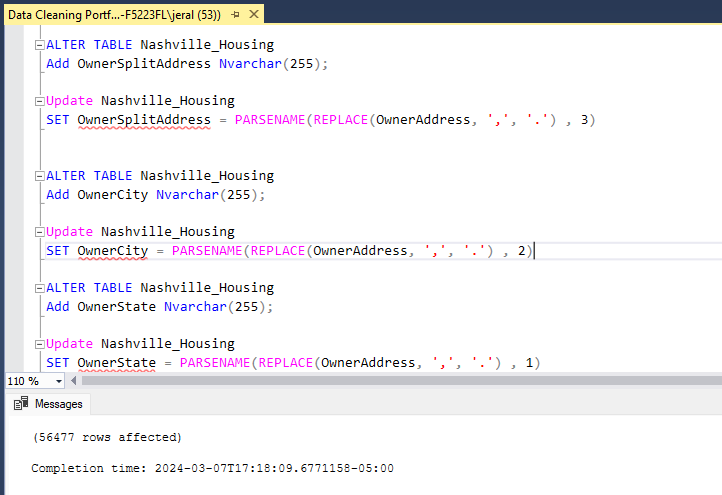
1. PropertyAddress - Fill Null Values with same ParcelID



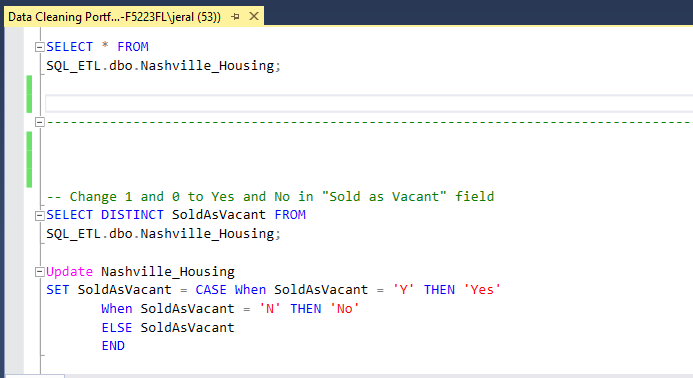
1. PropertyAddress - Segregate column into house address and city column.



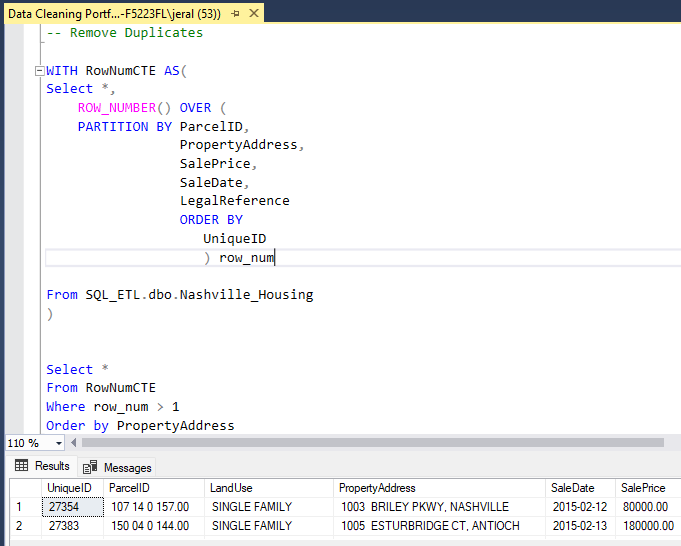
1. OwnerAddress - Segregate column into house address, city and state columns.



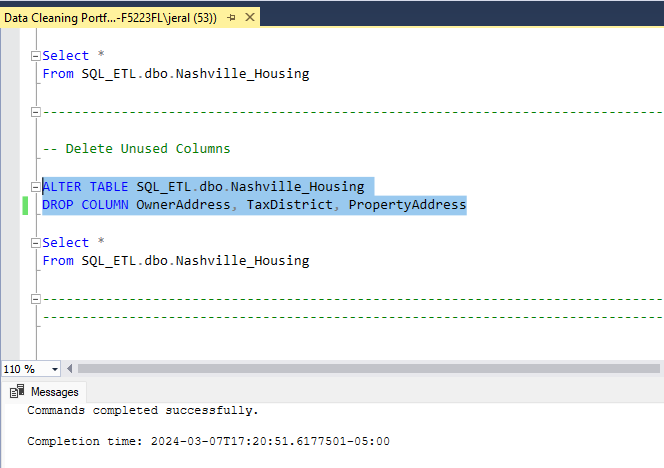
1. SoldAsVacant - Change Y and N to Yes and No



1. Remove Duplicates - Found 104 duplicate rows



1. Remove Unwanted Columns



CONCLUSION:

After carrying out the data cleaning, I was left with 56,373 rows and 21 columns.