Project: Nashville Housing Market Data Cleaning & Valuation

Executive Summary

This project involved the end-to-end processing of a raw real estate dataset containing 56,000+ housing records. The objective was to transform "dirty" scraped data into a structured SQL database to analyse market trends, specifically focusing on short-term flipping profitability, vacancy valuation, and neighbourhood pricing benchmarks.

Phase 1: Data Cleaning (The Engineering)

Objective: Standardize a raw, error-prone dataset to prepare it for high-level analysis.

The Execution:

- **Standardization:** Converted `SaleDate` from text strings to standard Date format to enable time-series calculations.
- **Null Handling:** Attempted a "Self-Join Population" method to fill missing Property Addresses using Reference IDs (ParcelID). Upon discovering unique constraints that made imputation impossible, executed a targeted purge of null-address rows to maintain analytical integrity.
- **Normalization:** Standardized the `SoldAsVacant` field from four inconsistent values ('Y', 'Yes', 'N', 'No') into a binary 'Yes/No' format for accurate aggregation.
- **Deduplication:** Utilized a `ROW_NUMBER()` Window Function to identify and remove duplicate transaction records, ensuring that the final analysis was not skewed by redundant data entries.

Phase 2: Market Analysis (The Insights)

Analysis 1: The "Flipper" Audit (Profitability)

- **Objective:** Identify properties sold multiple times within the dataset to calculate realized gains from short-term holding.
- **Key Finding:** The market contains extreme outliers in profitability. The top-performing asset (320 11th Ave S) generated a **$54 Million profit**, indicating that the dataset captures not just residential flips but massive commercial development deals. This requires segmenting the data further to separate "Home Flippers" from "Commercial Developers."

Analysis 2: The "Vacancy" Valuation (Asset Pricing)

- **Objective:** Determine if "Sold as Vacant" properties offer a significant discount compared to occupied homes.
- **Key Finding:** The "Vacant Discount" was smaller than anticipated. Occupied homes sold for an average of **$330k**, while Vacant properties sold for **$299k**.
- **Strategic Implication:** The ~10% price difference suggests that vacancy alone is not a strong indicator of a "bargain." Investors should not prioritize vacant land solely for price reduction purposes.

Analysis 3: Geographic Pricing (Price Per SqFt)

- **Objective:** Normalize pricing by size (`SalePrice / FinishedArea`) to identify the true "most expensive" neighborhoods.
- **Key Finding: Nashville ($148/sqft)** narrowly edged out the affluent suburb of **Brentwood ($147/sqft)**.
- **Conclusion:** This contradicts the assumption that suburban luxury estates command the highest value. The data suggests a higher demand density for urban properties, driving the price-per-square-foot metric higher than the surrounding suburbs.