Nashville Housing Portfolio Project- Data Cleaning and Data mining using SQL

Let us look at the data and take a brief overview of the columns we have:

These are the columns that we will be working with

+--------------------+---------------+------+-----+---------+-------+

| Field | Type | Null | Key | Default | Extra |

+--------------------+---------------+------+-----+---------+-------+

| UniqueID | int | NO | PRI | NULL | |

| ParcelID | varchar(50) | YES | | NULL | |

| LandUse | varchar(100) | YES | | NULL | |

| SaleDate | date | YES | | NULL | |

| SalePrice | int | YES | | NULL | |

| LegalReference | varchar(255) | YES | | NULL | |

| SoldAsVacant | varchar(10) | YES | | NULL | |

| OwnerName | varchar(255) | YES | | NULL | |

| Acreage | decimal(10,2) | YES | | NULL | |

| LandValue | int | YES | | NULL | |

| BuildingValue | int | YES | | NULL | |

| TotalValue | int | YES | | NULL | |

| YearBuilt | year | YES | | NULL | |

| Bedrooms | int | YES | | NULL | |

| FullBath | int | YES | | NULL | |

| HalfBath | int | YES | | NULL | |

| StreetAddress | varchar(255) | YES | | NULL | |

| CityAddress | varchar(255) | YES | | NULL | |

| OwnerStreetAddress | varchar(255) | YES | | NULL | |

| OwnerCityAddress | varchar(255) | YES | | NULL | |

| OwnerStateAddress | varchar(255) | YES | | NULL | |

+--------------------+---------------+------+-----+---------+-------+

Loading the housing data in MySQL

MySQL> select \* from PropertySales;

A screenshot of a table

AI-generated content may be incorrect.

Cleaning and loading the data into the correct columns

I knew that working with dates is a triumph as the date is in multiple formats like dd/mm/yyy, yyyy/mm/dd, or even Yesterday.

I have spent an entire day in analyzing the date information and trying to clean the data and bring a generic and standard date format in the entire dataset. However, it was not that easy.

Since I was working with MySQL on my terminal (local server) as I use a Mac, whenever I tried to update the date format in an excel and converted back to CSV, the formatting was redundant.

The original dataset had a Property Address column which contained data in the form of “1234 Main St, NYC, NY”. Therefore, while loading the data, the date column was not able to identify the date and thereby throwing the date as ‘0000-00-00’. I then started working around to identify the issue and then I found it in my **LOAD INFILE command**

“LOAD DATA LOCAL INFILE '/path/to/your/file.csv'

INTO TABLE your\_existing\_table

FIELDS TERMINATED BY ','

ENCLOSED BY '"' --This was the main difference while loading the data in its proper columns which I completely missed at first

LINES TERMINATED BY '\n'

IGNORE 1 ROWS;”

Populated the PropertyAddress where the PropertyAddress is null

Using JOIN--

MySQL> update PropertySales AS a JOIN PropertySales AS b on a.ParcelID = b.ParcelID AND a.UniqueID <> b.UniqueID SET a.PropertyAddress = b.PropertyAddress where a.PropertyAddress IS NULL AND b.PropertyAddress IS NOT NULL;

Breaking the Address into Street, City, State

Using SUBSTRING\_INDEX function--

street- MySQL> select substring\_index (OwnerAddress, ',', 1) AS OwnerStreet from PropertySales;

city- MySQL> select substring\_index(substring\_index(OwnerAddress, ',', -2),',', 1) AS OwnerCity from PropertySales

state- MySQL> select substring\_index(substring\_index(OwnerAddress, ',', -1),',', 1) AS OwnerState from PropertySales

Converting Y to ‘Yes’ and N to ‘No’

Using CASE statement--

MySQL> select SoldAsVacant, CASE when SoldAsVacant = 'Y' then ‘Yes’ when SoldAsVacant = ’N’ then ’No’ else SoldAsVacant end from PropertySales;

Drop unused columns

MySQL> alter table PropertySales drop column OwnerAddress, drop column PropertyAddress, drop column TaxDistrict;