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;

UniqueID	ParcelID	LandUse	SaleDate	SalePrice	LegalReference	SoldAsVacant	OwnerName
0	105 03 0D 008.00	RESIDENTIAL CONDO	2013-01-24	132000	20130128-0008725	No	
1	105 11 0 080.00	SINGLE FAMILY	2013-01-11	191500	20130118-0006337	No	STINSON, LAURA M.
2	118 03 0 130.00	SINGLE FAMILY	2013-01-18	202000	20130124-0008033	No	NUNES, JARED R.
3	119 01 0 479.00	SINGLE FAMILY	2013-01-18	32000	20130128-0008863	No	WHITFORD, KAREN
4	119 05 0 186.00	SINGLE FAMILY	2013-01-23	102000	20130131-0009929	No	HENDERSON, JAMES P. & LYNN P.
5	119 05 0 387.00	SINGLE FAMILY	2013-01-04	93736	20130118-0006110	No	MILLER, JORDAN
6	119 10 0A 104.00	RESIDENTIAL CONDO	2013-01-07	64900	20130109-0002881	No	
7	119 13 0 183.00	SINGLE FAMILY	2013-01-15	44000	20130115-0004888	No	MICKLER, PATRICK L. & LOIS J. & ARNETT, RYAN D.
8	119 13 0 183.00	SINGLE FAMILY	2013-01-25	49900	20130128-0008950	No	MICKLER, PATRICK L. & LOIS J. & ARNETT, RYAN D.
9	119 15 0 158.00	SINGLE FAMILY	2013-01-09	25000	20130111-0003850	No	SONA LAND CO, LLC

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 analyzing the date information and trying to clean the data and bring a generic and standard date format to 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 Excel and converted back to CSV, the formatting was redundant.

The original dataset had a Property Address column that 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 threw 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;