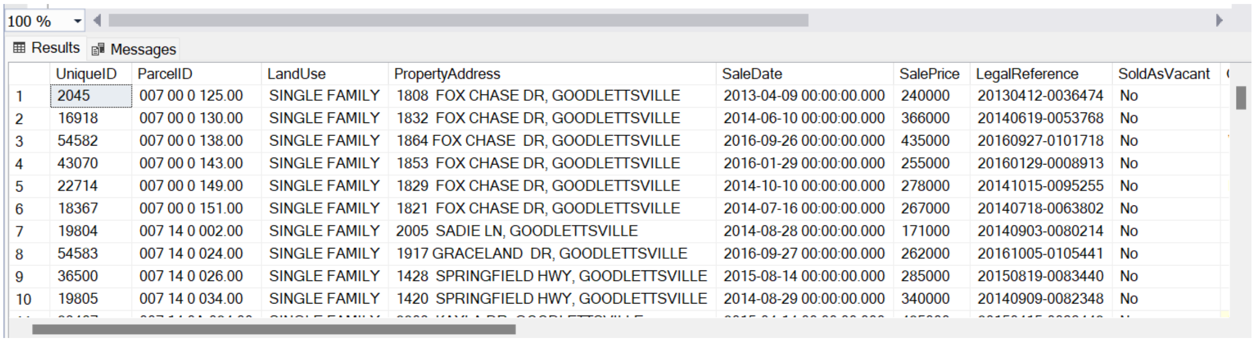
Cleaning Data in SQL Queries

In this project, I cleaned dirty excel file in SQL to

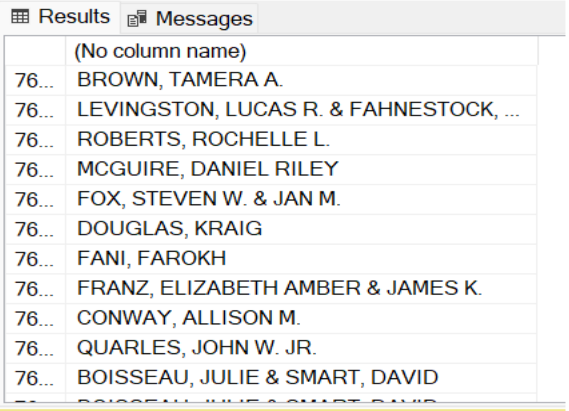
1. **STDARDIZE DATE FORMAT--** (To set the date in a more pleasant format).
2. **POPULATE PROPERTY ADDRESS DATA-- (**To put same address with same ParcelID under empty PropertyAddresses).
3. **BREAKING OUT ADDRESS INTO INDIVIDUAL COLUMNS--**(Address, City, State).
4. **CHANGE Y AND N TO YES AND NO IN.**
5. Delete unused Columns.
6. I decided not to remove NULL or duplicate values in SQL because the deleted columns might be useful in future, I will rather do that in Excel or Power BI.

```
--To check the entire file before starting the cleaning process.

SELECT *
FROM NatshvilleHousing
```



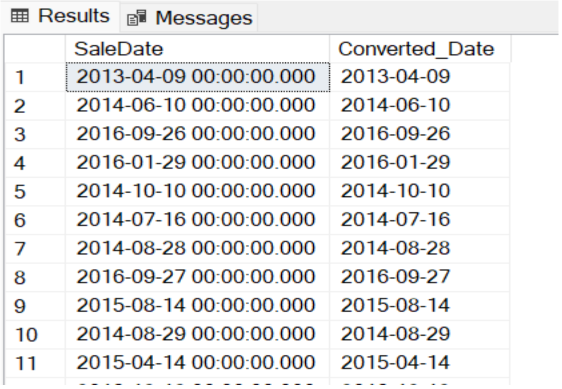
	UniqueID	ParcelID	LandUse	PropertyAddress	SaleDate	SalePrice	LegalReference	SoldAsVacant
1	2045	007 00 0 125.00	SINGLE FAMILY	1808 FOX CHASE DR, GOODLETTSVILLE	2013-04-09 00:00:00.000	240000	20130412-0036474	No
2	16918	007 00 0 130.00	SINGLE FAMILY	1832 FOX CHASE DR, GOODLETTSVILLE	2014-06-10 00:00:00.000	366000	20140619-0053768	No
3	54582	007 00 0 138.00	SINGLE FAMILY	1864 FOX CHASE DR, GOODLETTSVILLE	2016-09-26 00:00:00.000	435000	20160927-0101718	No
4	43070	007 00 0 143.00	SINGLE FAMILY	1853 FOX CHASE DR, GOODLETTSVILLE	2016-01-29 00:00:00.000	255000	20160129-0008913	No
5	22714	007 00 0 149.00	SINGLE FAMILY	1829 FOX CHASE DR, GOODLETTSVILLE	2014-10-10 00:00:00.000	278000	20141015-0095255	No
6	18367	007 00 0 151.00	SINGLE FAMILY	1821 FOX CHASE DR, GOODLETTSVILLE	2014-07-16 00:00:00.000	267000	20140718-0063802	No
7	19804	007 14 0 002.00	SINGLE FAMILY	2005 SADIE LN, GOODLETTSVILLE	2014-08-28 00:00:00.000	171000	20140903-0080214	No
8	54583	007 14 0 024.00	SINGLE FAMILY	1917 GRACELAND DR, GOODLETTSVILLE	2016-09-27 00:00:00.000	262000	20161005-0105441	No
9	36500	007 14 0 026.00	SINGLE FAMILY	1428 SPRINGFIELD HWY, GOODLETTSVILLE	2015-08-14 00:00:00.000	285000	20150819-0083440	No
10	19805	007 14 0 034.00	SINGLE FAMILY	1420 SPRINGFIELD HWY, GOODLETTSVILLE	2014-08-29 00:00:00.000	340000	20140909-0082348	No

```
1--Remove unwanted space before or in between the names

SELECT TRIM([OwnerName])
FROM NatshvilleHousing
UPDATE NatshvilleHousing
SET [OwnerName] = TRIM([OwnerName])
```



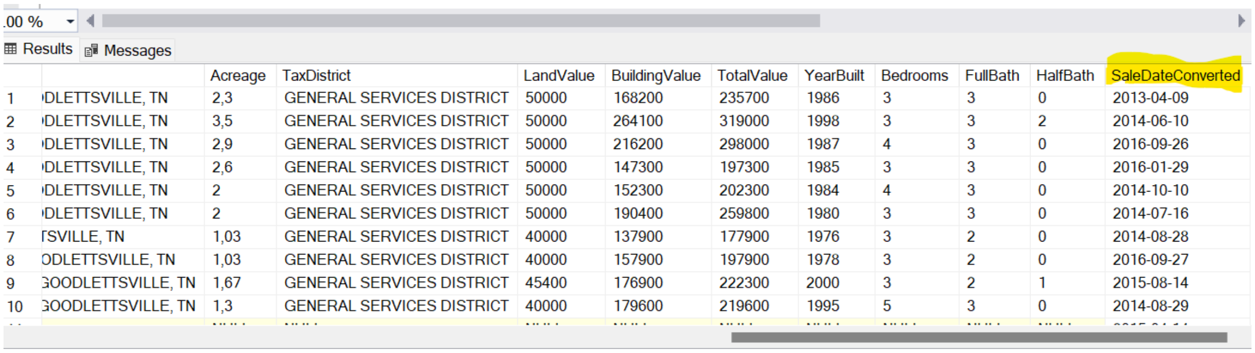
	(No column name)
76...	BROWN, TAMERA A.
76...	LEVINGSTON, LUCAS R. & FAHNESTOCK, ...
76...	ROBERTS, ROCHELLE L.
76...	MCGUIRE, DANIEL RILEY
76...	FOX, STEVEN W. & JAN M.
76...	DOUGLAS, KRAIG
76...	FANI, FAROKH
76...	FRANZ, ELIZABETH AMBER & JAMES K.
76...	CONWAY, ALLISON M.
76...	QUARLES, JOHN W. JR.
76...	BOISSEAU, JULIE & SMART, DAVID

```
2--STANDARDIZE DATE FORMAT-- (To set the date in a more pleasant format)

SELECT SaleDate, CONVERT(DATE, SaleDate) AS Converted_Date
FROM NatshvilleHousing
```



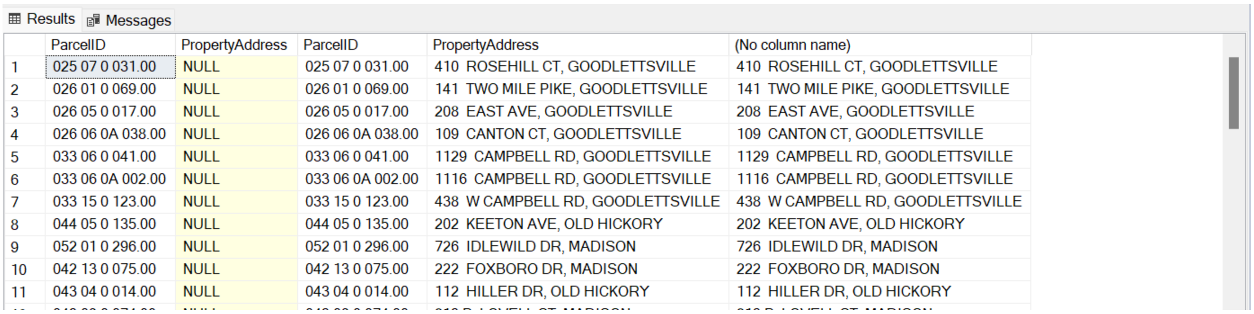
	SaleDate	Converted_Date
1	2013-04-09 00:00:00.000	2013-04-09
2	2014-06-10 00:00:00.000	2014-06-10
3	2016-09-26 00:00:00.000	2016-09-26
4	2016-01-29 00:00:00.000	2016-01-29
5	2014-10-10 00:00:00.000	2014-10-10
6	2014-07-16 00:00:00.000	2014-07-16
7	2014-08-28 00:00:00.000	2014-08-28
8	2016-09-27 00:00:00.000	2016-09-27
9	2015-08-14 00:00:00.000	2015-08-14
10	2014-08-29 00:00:00.000	2014-08-29
11	2015-04-14 00:00:00.000	2015-04-14

```
UPDATE NatshvilleHousing
SET SaleDate = CONVERT (DATE,SaleDate)
ALTER TABLE NatshvilleHousing
ADD SaleDateConverted DATE; --This will add another column known as SaleDateConverted

UPDATE NatshvilleHousing
SET SaleDateConverted = CONVERT (DATE,SaleDate) --This will set the converted dates into
SaleDateConverted
```



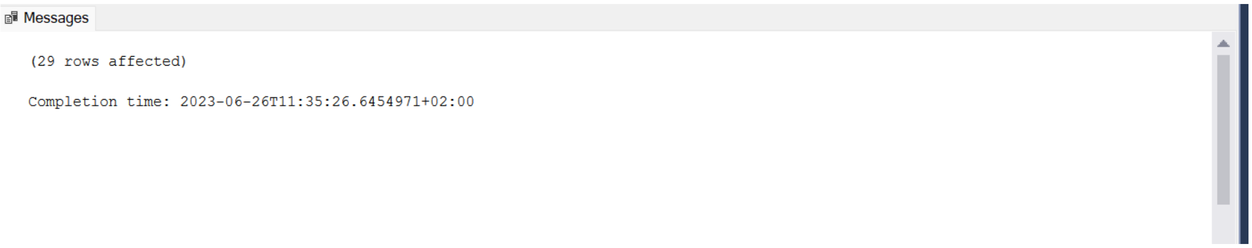
		Acreage	TaxDistrict	LandValue	BuildingValue	TotalValue	YearBuilt	Bedrooms	FullBath	HalfBath	SaleDateConverted
1	DLETTSVILLE, TN	2,3	GENERAL SERVICES DISTRICT	50000	168200	235700	1986	3	3	0	2013-04-09
2	DLETTSVILLE, TN	3,5	GENERAL SERVICES DISTRICT	50000	264100	319000	1998	3	3	2	2014-06-10
3	DLETTSVILLE, TN	2,9	GENERAL SERVICES DISTRICT	50000	216200	298000	1987	4	3	0	2016-09-26
4	DLETTSVILLE, TN	2,6	GENERAL SERVICES DISTRICT	50000	147300	197300	1985	3	3	0	2016-01-29
5	DLETTSVILLE, TN	2	GENERAL SERVICES DISTRICT	50000	152300	202300	1984	4	3	0	2014-10-10
6	DLETTSVILLE, TN	2	GENERAL SERVICES DISTRICT	50000	190400	259800	1980	3	3	0	2014-07-16
7	TSVILLE, TN	1,03	GENERAL SERVICES DISTRICT	40000	137900	177900	1976	3	2	0	2014-08-28
8	ODLETTSVILLE, TN	1,03	GENERAL SERVICES DISTRICT	40000	157900	197900	1978	3	2	0	2016-09-27
9	GOODLETTSVILLE, TN	1,67	GENERAL SERVICES DISTRICT	45400	176900	222300	2000	3	2	1	2015-08-14
10	GOODLETTSVILLE, TN	1,3	GENERAL SERVICES DISTRICT	40000	179600	219600	1995	5	3	0	2014-08-29

```
3--POPULATE PROPERTY ADDRESS DATA-- (To put same address with same ParcelID under empty
PropertyAddresses)

SELECT A.ParcelID, A.PropertyAddress, B.ParcelID, B.PropertyAddress,
ISNULL(A.PropertyAddress,B.PropertyAddress) --Adds Address into the Empty PropertyAddress
FROM NatshvilleHousing A
JOIN NatshvilleHousing B
       on A.ParcelID = B.ParcelID
       AND A.[UniqueID ] <> B.[UniqueID ]
where A.PropertyAddress IS NULL
```



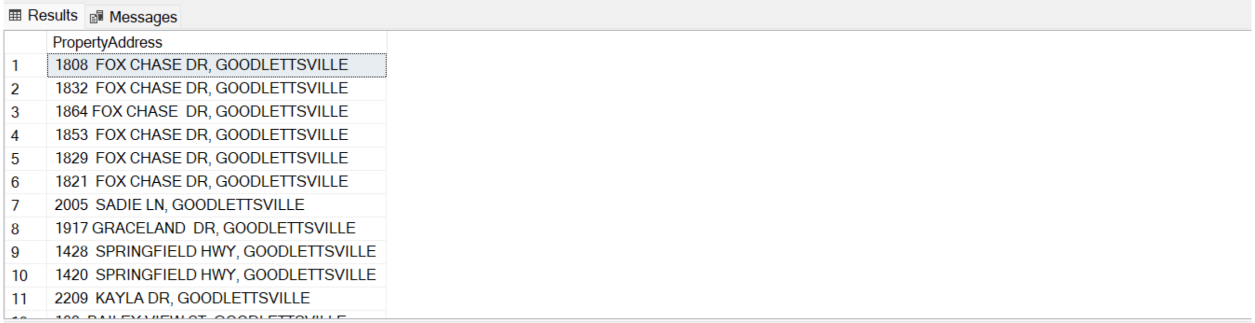
	ParcelID	PropertyAddress	ParcelID	PropertyAddress	(No column name)
1	025 07 0 031.00	NULL	025 07 0 031.00	410 ROSEHILL CT, GOODLETTSVILLE	410 ROSEHILL CT, GOODLETTSVILLE
2	026 01 0 069.00	NULL	026 01 0 069.00	141 TWO MILE PIKE, GOODLETTSVILLE	141 TWO MILE PIKE, GOODLETTSVILLE
3	026 05 0 017.00	NULL	026 05 0 017.00	208 EAST AVE, GOODLETTSVILLE	208 EAST AVE, GOODLETTSVILLE
4	026 06 0A 038.00	NULL	026 06 0A 038.00	109 CANTON CT, GOODLETTSVILLE	109 CANTON CT, GOODLETTSVILLE
5	033 06 0 041.00	NULL	033 06 0 041.00	1129 CAMPBELL RD, GOODLETTSVILLE	1129 CAMPBELL RD, GOODLETTSVILLE
6	033 06 0A 002.00	NULL	033 06 0A 002.00	1116 CAMPBELL RD, GOODLETTSVILLE	1116 CAMPBELL RD, GOODLETTSVILLE
7	033 15 0 123.00	NULL	033 15 0 123.00	438 W CAMPBELL RD, GOODLETTSVILLE	438 W CAMPBELL RD, GOODLETTSVILLE
8	044 05 0 135.00	NULL	044 05 0 135.00	202 KEETON AVE, OLD HICKORY	202 KEETON AVE, OLD HICKORY
9	052 01 0 296.00	NULL	052 01 0 296.00	726 IDLEWILD DR, MADISON	726 IDLEWILD DR, MADISON
10	042 13 0 075.00	NULL	042 13 0 075.00	222 FOXBORO DR, MADISON	222 FOXBORO DR, MADISON
11	043 04 0 014.00	NULL	043 04 0 014.00	112 HILLER DR, OLD HICKORY	112 HILLER DR, OLD HICKORY

```
UPDATE A
SET PropertyAddress = ISNULL(A.PropertyAddress,B.PropertyAddress)
FROM NatshvilleHousing A
JOIN NatshvilleHousing B
       on A.ParcelID = B.ParcelID
       AND A.[UniqueID ] <> B.[UniqueID ]
where A.PropertyAddress IS NULL
```



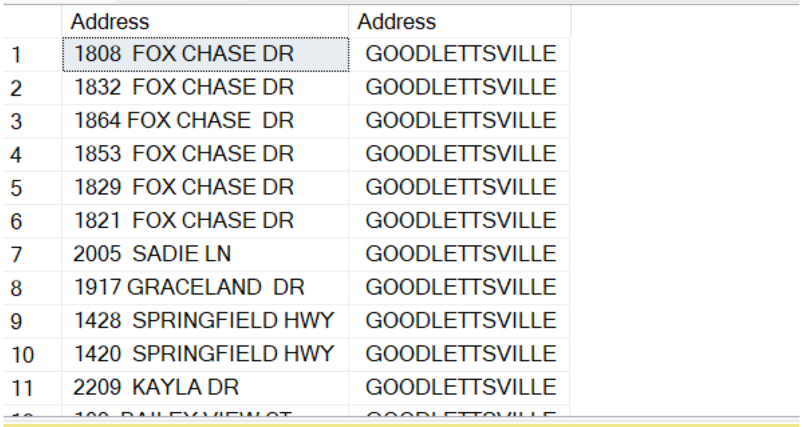
```
(29 rows affected)

Completion time: 2023-06-26T11:35:26.6454971+02:00
```

```sql
4--BREAKING OUT ADDRESS INTO INDIVIDUAL COLUMNS (ADDRESS, CITY, STATE)

--FOR PROPERTYADDRESS
SELECT PropertyAddress
FROM NatshvilleHousing
```



	PropertyAddress
1	1808 FOX CHASE DR, GOODLETTSVILLE
2	1832 FOX CHASE DR, GOODLETTSVILLE
3	1864 FOX CHASE DR, GOODLETTSVILLE
4	1853 FOX CHASE DR, GOODLETTSVILLE
5	1829 FOX CHASE DR, GOODLETTSVILLE
6	1821 FOX CHASE DR, GOODLETTSVILLE
7	2005 SADIE LN, GOODLETTSVILLE
8	1917 GRACELAND DR, GOODLETTSVILLE
9	1428 SPRINGFIELD HWY, GOODLETTSVILLE
10	1420 SPRINGFIELD HWY, GOODLETTSVILLE
11	2209 KAYLA DR, GOODLETTSVILLE

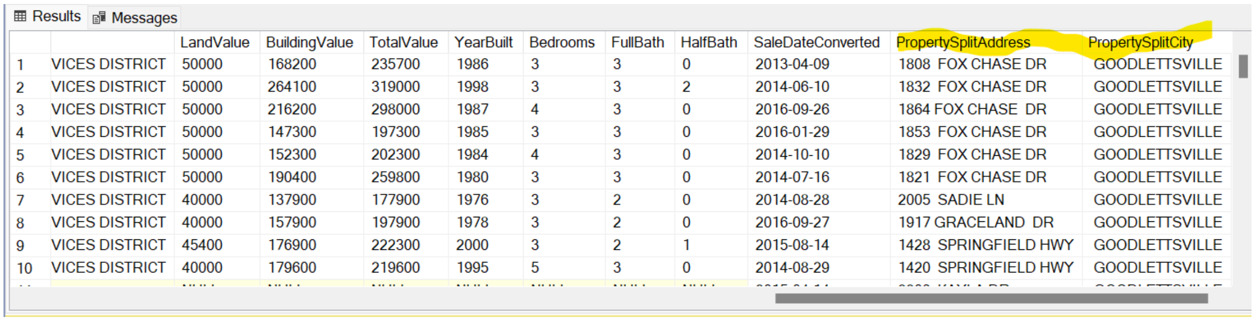
```sql
SELECT
PARSENAME(REPLACE(PropertyAddress, ',','.'),2) Address,
PARSENAME(REPLACE(PropertyAddress, ',','.'),1) Address
FROM NatshvilleHousing
```



	Address	Address
1	1808 FOX CHASE DR	GOODLETTSVILLE
2	1832 FOX CHASE DR	GOODLETTSVILLE
3	1864 FOX CHASE DR	GOODLETTSVILLE
4	1853 FOX CHASE DR	GOODLETTSVILLE
5	1829 FOX CHASE DR	GOODLETTSVILLE
6	1821 FOX CHASE DR	GOODLETTSVILLE
7	2005 SADIE LN	GOODLETTSVILLE
8	1917 GRACELAND DR	GOODLETTSVILLE
9	1428 SPRINGFIELD HWY	GOODLETTSVILLE
10	1420 SPRINGFIELD HWY	GOODLETTSVILLE
11	2209 KAYLA DR	GOODLETTSVILLE

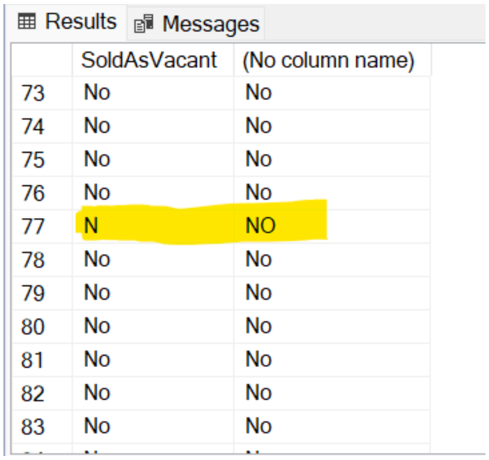
```sql
ALTER TABLE NatshvilleHousing
ADD PropertySplitAddress nvarchar(255);

UPDATE NatshvilleHousing
SET PropertySplitAddress = PARSENAME(REPLACE(PropertyAddress,',', '.'),2)

ALTER TABLE NatshvilleHousing
ADD PropertySplitCity nvarchar(255);
```

```sql
UPDATE NatshvilleHousing
SET PropertySplitCity = PARSENAME(REPLACE(PropertyAddress,',', '.'),1)
SELECT *
FROM NatshvilleHousing
```



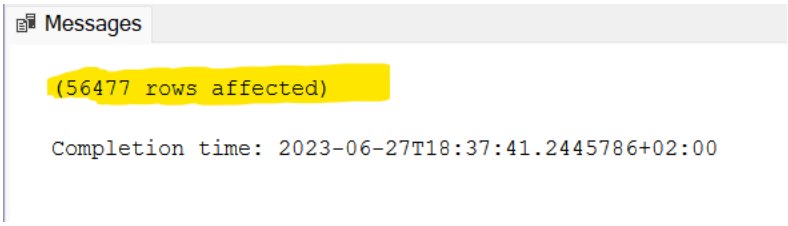
		LandValue	BuildingValue	TotalValue	YearBuilt	Bedrooms	FullBath	HalfBath	SaleDateConverted	PropertySplitAddress	PropertySplitCity
1	VICES DISTRICT	50000	168200	235700	1986	3	3	0	2013-04-09	1808 FOX CHASE DR	GOODLETTSVILLE
2	VICES DISTRICT	50000	264100	319000	1998	3	3	2	2014-06-10	1832 FOX CHASE DR	GOODLETTSVILLE
3	VICES DISTRICT	50000	216200	298000	1987	4	3	0	2016-09-26	1864 FOX CHASE DR	GOODLETTSVILLE
4	VICES DISTRICT	50000	147300	197300	1985	3	3	0	2016-01-29	1853 FOX CHASE DR	GOODLETTSVILLE
5	VICES DISTRICT	50000	152300	202300	1984	4	3	0	2014-10-10	1829 FOX CHASE DR	GOODLETTSVILLE
6	VICES DISTRICT	50000	190400	259800	1980	3	3	0	2014-07-16	1821 FOX CHASE DR	GOODLETTSVILLE
7	VICES DISTRICT	40000	137900	177900	1976	3	2	0	2014-08-28	2005 SADIE LN	GOODLETTSVILLE
8	VICES DISTRICT	40000	157900	197900	1978	3	2	0	2016-09-27	1917 GRACELAND DR	GOODLETTSVILLE
9	VICES DISTRICT	45400	176900	222300	2000	3	2	1	2015-08-14	1428 SPRINGFIELD HWY	GOODLETTSVILLE
10	VICES DISTRICT	40000	179600	219600	1995	5	3	0	2014-08-29	1420 SPRINGFIELD HWY	GOODLETTSVILLE

```sql
5--CHANGE Y AND N TO YES AND NO IN [SOLD AS VACANT] FIELD

SELECT SoldAsVacant,
CASE
        WHEN SoldAsVacant = 'Y' then 'YES'
        WHEN SoldAsVacant = 'N' then 'NO'
        ELSE SoldAsVacant
        END
FROM NatshvilleHousing
```



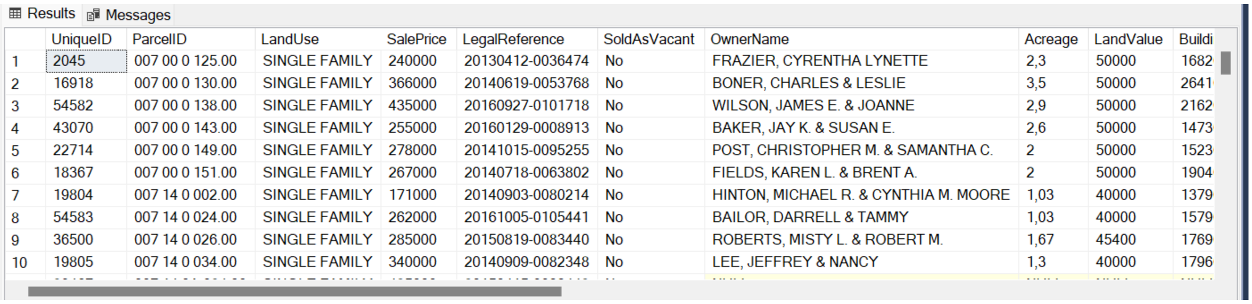
	SoldAsVacant	(No column name)
73	No	No
74	No	No
75	No	No
76	No	No
77	N	NO
78	No	No
79	No	No
80	No	No
81	No	No
82	No	No
83	No	No

```sql
UPDATE NatshvilleHousing
SET SoldAsVacant = case
       when SoldAsVacant = 'Y' then 'YES'
       when SoldAsVacant = 'N' then 'NO'
       when SoldAsVacant = 'Yes' then 'YES'
       when SoldAsVacant = 'No' then 'NO'
       ELSE SoldAsVacant
       END
```



```
(56477 rows affected)

Completion time: 2023-06-27T18:37:41.2445786+02:00
```

```sql
6--Delete unused Columns

select*
from NatshvilleHousing
ALTER TABLE NatshvilleHousing
DROP COLUMN OwnerAddress, TaxDistrict, PropertyAddress

ALTER TABLE NatshvilleHousing
DROP COLUMN SaleDate
```



	UniqueID	ParcelID	LandUse	SalePrice	LegalReference	SoldAsVacant	OwnerName	Acreage	LandValue	Buildi
1	2045	007 00 0 125.00	SINGLE FAMILY	240000	20130412-0036474	No	FRAZIER, CYRENTHA LYNETTE	2,3	50000	1682
2	16918	007 00 0 130.00	SINGLE FAMILY	366000	20140619-0053768	No	BONER, CHARLES & LESLIE	3,5	50000	2641
3	54582	007 00 0 138.00	SINGLE FAMILY	435000	20160927-0101718	No	WILSON, JAMES E. & JOANNE	2,9	50000	2162
4	43070	007 00 0 143.00	SINGLE FAMILY	255000	20160129-0008913	No	BAKER, JAY K. & SUSAN E.	2,6	50000	1473
5	22714	007 00 0 149.00	SINGLE FAMILY	278000	20141015-0095255	No	POST, CHRISTOPHER M. & SAMANTHA C.	2	50000	1523
6	18367	007 00 0 151.00	SINGLE FAMILY	267000	20140718-0063802	No	FIELDS, KAREN L. & BRENT A.	2	50000	1904
7	19804	007 14 0 002.00	SINGLE FAMILY	171000	20140903-0080214	No	HINTON, MICHAEL R. & CYNTHIA M. MOORE	1,03	40000	1379
8	54583	007 14 0 024.00	SINGLE FAMILY	262000	20161005-0105441	No	BAILOR, DARRELL & TAMMY	1,03	40000	1579
9	36500	007 14 0 026.00	SINGLE FAMILY	285000	20150819-0083440	No	ROBERTS, MISTY L. & ROBERT M.	1,67	45400	1769
10	19805	007 14 0 034.00	SINGLE FAMILY	340000	20140909-0082348	No	LEE, JEFFREY & NANCY	1,3	40000	1796