Nashville Housing Data Cleaning

Imported Data to Microsoft SQL

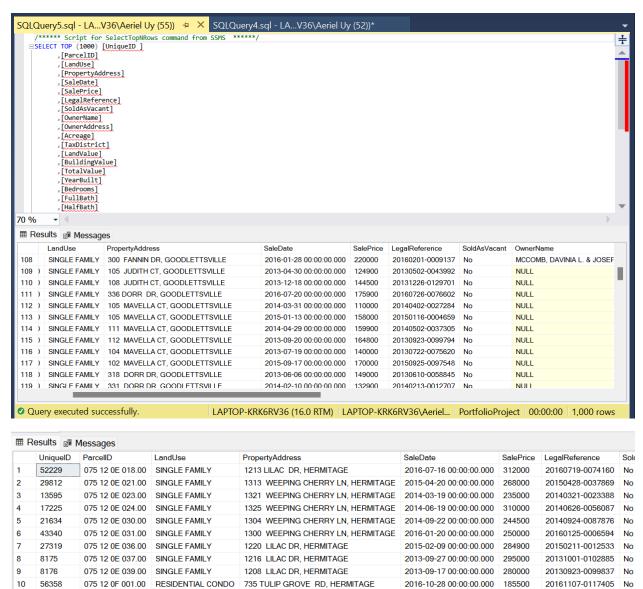
Objective: Data Cleaning

11

12

35201

18702



2015-07-24 00:00:00.000 159900

2014-07-30 00:00:00 000 131000

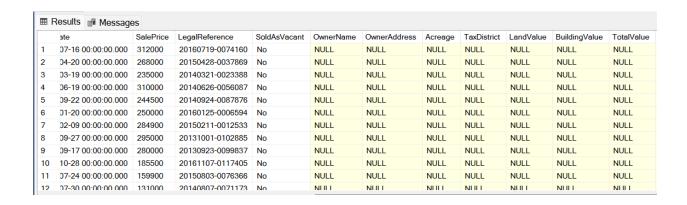
20150803-0076366

20140807-0071173 No

No

075 12 0F 002 00 RESIDENTIAL CONDO 735 TULIP GROVE RD. HERMITAGE

075 12 0F 005 00 RESIDENTIAL CONDO 735 TULIP GROVE RD HERMITAGE

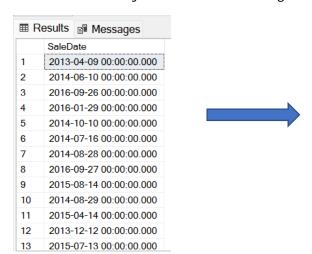


Tasks

First Task

Standardized Date Format on SaleDate Column showing the date and time

Select SaleDate, CONVERT(Date, SaleDate)
From PortfolioProject..NashvilleHousing



⊞ F	Results Messages	
	SaleDate	(No column name)
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
12	2013-12-12 00:00:00.000	2013-12-12
13	2015-07-13 00:00:00 000	2015-07-13

• Update the data with New Column named 'SaleDateConverted'

ALTER TABLE was used as the previous **UPDATE** did not work and changed the column name to SaleDateConverted to be the new column. SaleDate column to be deleted.

```
Update NashvilleHousing
SET SaleDate = CONVERT(Date, SaleDate)

ALTER TABLE NashvilleHousing
Add SaleDateConverted Date;

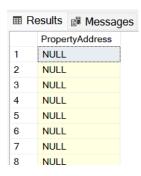
Update NashvilleHousing
SET SaleDateConverted = CONVERT(Date, SaleDate)
```

```
SaleDateConverted (No column name)
     2016-01-29
                      2016-01-29
5
     2014-10-10
                      2014-10-10
6
                      2014-07-16
     2014-07-16
7
                      2014-08-28
     2014-08-28
8
                      2016-09-27
     2016-09-27
9
     2015-08-14
                      2015-08-14
10
                      2014-08-29
     2014-08-29
11
     2015-04-14
                      2015-04-14
12
    2013-12-12
                      2013-12-12
                      2015-07-13
13
    2015-07-13
    2013-06-07
                      2013-06-07
14
15
     2014-06-30
                      2014-06-30
     2014-06-30
                      2014-06-30
```

Second Task

Populate Property Address Data

Select PropertyAddress
From PortfolioProject.dbo.NashvilleHousing
Where PropertyAddress is null

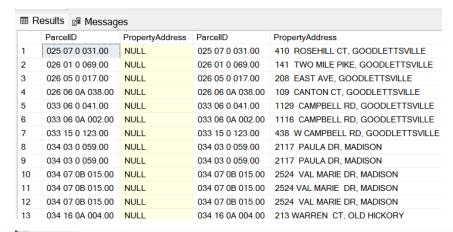


To populate the Null values, the ParcellD was used as it shows that if it is the same ParcellD then the PropertyAddress is the same.



```
Select a.ParcelID, a. PropertyAddress, b.ParcelID, b.PropertyAddress
From PortfolioProject..NashvilleHousing a
JOIN PortfolioProject..NashvilleHousing b
            ON a.ParcelID = b.ParcelID
            AND a.[UniqueID] <> b. [UniqueID]
Where PropertyAddress is null
```

We use JOIN to populate the property address. Alias were used such as 'a' and 'b' for the table name. Query shows to populate the address with the same parcel ID but not the same unique ID.





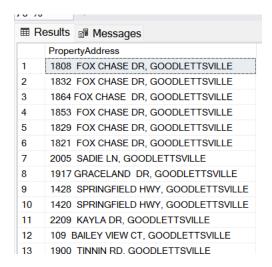
70 %	5 → ∢						
⊞F	■ Results						
	ParcellD	PropertyAddress	ParcellD	PropertyAddress	(No column name)		
1	025 07 0 031.00	NULL	025 07 0 031.00	410 ROSEHILL CT, GOODLETTSVILLE	410 ROSEHILL CT, GOODLETTSVILLE		
2	026 01 0 069.00	NULL	026 01 0 069.00	141 TWO MILE PIKE, GOODLETTSVILLE	141 TWO MILE PIKE, 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	034 03 0 059.00	NULL	034 03 0 059.00	2117 PAULA DR, MADISON	2117 PAULA DR, MADISON		
9	034 03 0 059.00	NULL	034 03 0 059.00	2117 PAULA DR, MADISON	2117 PAULA DR, MADISON		
10	034 07 0B 015.00	NULL	034 07 0B 015.00	2524 VAL MARIE DR, MADISON	2524 VAL MARIE DR, MADISON		
11	034 07 0B 015.00	NULL	034 07 0B 015.00	2524 VAL MARIE DR, MADISON	2524 VAL MARIE DR, MADISON		
12	034 07 0B 015.00	NULL	034 07 0B 015.00	2524 VAL MARIE DR, MADISON	2524 VAL MARIE DR, MADISON		
13	034 16 0A 004.00	NULL	034 16 0A 004.00	213 WARREN CT, OLD HICKORY	213 WARREN CT, OLD HICKORY		

Update a

Third Task

Separate the addresses to individual columns (Address, City, State)

Select PropertyAddress
From PortfolioProject.dbo.NashvilleHousing
--Where PropertyAddress is null
--order by ParcelID

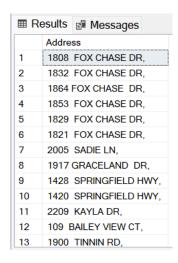


SELECT

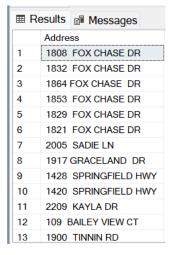
SUBSTRING(PropertyAddress, 1, CHARINDEX(',', PropertyAddress)) as Address From PortfolioProject..NashvilleHousing

SELECT

SUBSTRING(PropertyAddress, 1, CHARINDEX(',', PropertyAddress)-1) as Address From PortfolioProject..NashvilleHousing







SELECT

SUBSTRING(PropertyAddress, 1, CHARINDEX(',', PropertyAddress) -1) as Address
, SUBSTRING(PropertyAddress, CHARINDEX(',', PropertyAddress) + 1 , LEN(PropertyAddress))
as Address

From PortfolioProject.dbo.NashvilleHousing

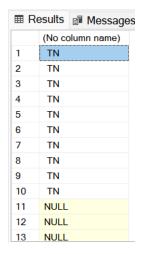
+1 was removed to remove the comma from the result and 'LEN' to separate the address even with different lengths



Select

```
PARSENAME(REPLACE(OwnerAddress, ',', '.'), 1)
From PortfolioProject.dbo.NashvilleHousing
```

--Another easier way of splitting the addresses. Commas were replaced with periods as PARSENAME is useful only for periods.

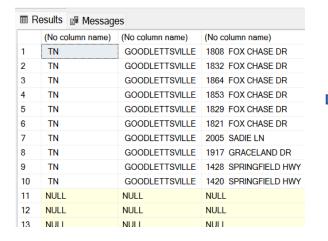


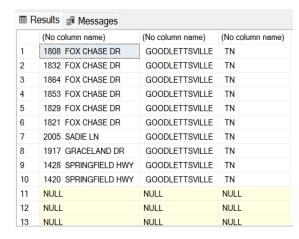
Select

```
PARSENAME(REPLACE(OwnerAddress, ',', '.'), 1)
,PARSENAME(REPLACE(OwnerAddress, ',', '.'), 2)
,PARSENAME(REPLACE(OwnerAddress, ',', '.'), 3)
From PortfolioProject.dbo.NashvilleHousing
```

Select

```
PARSENAME(REPLACE(OwnerAddress, ',', '.'), 1)
,PARSENAME(REPLACE(OwnerAddress, ',', '.'), 2)
,PARSENAME(REPLACE(OwnerAddress, ',', '.'), 3)
From PortfolioProject.dbo.NashvilleHousing
```





```
ALTER TABLE NashvilleHousing

Mod OwnerSplitAddress Nvarchar(255);

Update NashvilleHousing

SET OwnerSplitAddress = PARSENAME(REPLACE(OwnerAddress, ',', '.'), 3)

ALTER TABLE NashvilleHousing

Add OwnerSplitCity Nvarchar(255);

Update NashvilleHousing

SET OwnerSplitCity = PARSENAME(REPLACE(OwnerAddress, ',', '.'), 2)

ALTER TABLE NashvilleHousing

Add OwnerSplitState Nvarchar(255);

Update NashvilleHousing

SET OwnerSplitState = PARSENAME(REPLACE(OwnerAddress, ',', '.'), 1)
```

OwnerSplitAddress	OwnerSplitCity	OwnerSplitState
1808 FOX CHASE DR	GOODLETTSVILLE	TN
1832 FOX CHASE DR	GOODLETTSVILLE	TN
1864 FOX CHASE DR	GOODLETTSVILLE	TN
1853 FOX CHASE DR	GOODLETTSVILLE	TN
1829 FOX CHASE DR	GOODLETTSVILLE	TN
1821 FOX CHASE DR	GOODLETTSVILLE	TN
2005 SADIE LN	GOODLETTSVILLE	TN
1917 GRACELAND DR	GOODLETTSVILLE	TN
1428 SPRINGFIELD HWY	GOODLETTSVILLE	TN
1420 SPRINGFIELD HWY	GOODLETTSVILLE	TN
NULL	NULL	NULL
NHH	NHI	NHI

Fourth Task

 Replacing the Y and N to Yes and No in "Sold As Vacant" field

```
Select Distinct (SoldasVacant), COUNT(SoldasVacant)
From PortfolioProject..NashvilleHousing
Group by SoldasVacant
Order by 2
```



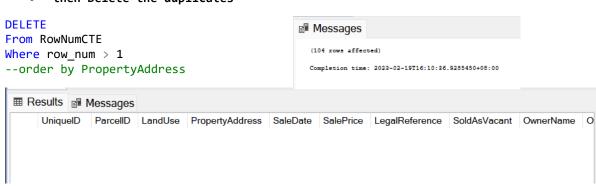
Fifth Task

--Remove Duplicates by Identifying the Duplicates

```
WITH ROWNUMCTE AS(
Select *,
       ROW_NUMBER() OVER (
       PARTITION BY ParcelID,
                             PropertyAddress,
                             SalePrice,
                             SaleDate,
                             LegalReference
                             Order by
                                   UniqueID
                                   ) as row_num
From PortfolioProject.dbo.NashvilleHousing
--order by ParcelID
Select *
From RowNumCTE
Where row_num > 1
order by PropertyAddress
```

⊞ R	Results 🗐	Messages			
	UniquelD	ParcellD	LandUse	PropertyAddress	SaleDate
46	27380	149 15 0A 418.00	SINGLE FAMILY	3044 BROMLEY WAY, ANTIOCH	2015-02-12 00:00:00.000
47	12692	163 11 0B 081.00	RESIDENTIAL CONDO	305 ASH FORGE DR, ANTIOCH	2014-02-03 00:00:00.000
48	27143	114 14 0A 022.00	SINGLE FAMILY	309 DOGWOOD TRL, NASHVILLE	2015-02-18 00:00:00.00(
49	27392	164 07 0A 139.00	SINGLE FAMILY	312 GROVEDALE TRCE, ANTIOCH	2015-02-20 00:00:00.000
50	27355	108 08 0 114.00	SINGLE FAMILY	3121 E LAKE DR, NASHVILLE	2015-02-06 00:00:00.000
51	27356	108 12 0B 031.00	SINGLE FAMILY	3124 HARBORWOOD DR, NASHVILLE	2015-02-13 00:00:00.000
52	27387	150 07 0 181.00	SINGLE FAMILY	3140 COUNTRY MEADOW RD, ANTI	2015-02-18 00:00:00.000
53	27388	150 09 0A 028.00	SINGLE FAMILY	3313 CALAIS CIR, ANTIOCH	2015-02-13 00:00:00.000
54	27382	150 03 0 442.00	SINGLE FAMILY	3412 WESTCLIFFE CT, ANTIOCH	2015-02-02 00:00:00.00(
55	27353	098 13 0C 200.00	SINGLE FAMILY	3420 EARHART RD, MOUNT JULIET	2015-02-06 00:00:00.000
56	27386	150 05 0 353.00	SINGLE FAMILY	365 CEDARCLIFF RD, ANTIOCH	2015-02-10 00:00:00.000
57	27402	176 01 0 017 00	SINGLE FAMILY	3848 ROLIND ROCK DR ANTIOCH	2015-02-13 00:00:00 000

• then Delete the duplicates



Sixth Task

• Delete Unused Columns

ALTER TABLE PortfolioProject.dbo.NashvilleHousing DROP COLUMN OwnerAddress, TaxDistrict, PropertyAddress

	UniquelD	ParcelID	LandUse	SaleDate	SalePrice	LegalReference	SoldAsVaca
1	2045	007 00 0 125.00	SINGLE FAMILY	2013-04-09 00:00:00.000	240000	20130412-0036474	No
2	16918	007 00 0 130.00	SINGLE FAMILY	2014-06-10 00:00:00.000	366000	20140619-0053768	No
3	54582	007 00 0 138.00	SINGLE FAMILY	2016-09-26 00:00:00.000	435000	20160927-0101718	No
4	43070	007 00 0 143.00	SINGLE FAMILY	2016-01-29 00:00:00.000	255000	20160129-0008913	No
5	22714	007 00 0 149.00	SINGLE FAMILY	2014-10-10 00:00:00.000	278000	20141015-0095255	No
6	18367	007 00 0 151.00	SINGLE FAMILY	2014-07-16 00:00:00.000	267000	20140718-0063802	No
7	19804	007 14 0 002.00	SINGLE FAMILY	2014-08-28 00:00:00.000	171000	20140903-0080214	No
8	54583	007 14 0 024.00	SINGLE FAMILY	2016-09-27 00:00:00.000	262000	20161005-0105441	No
9	36500	007 14 0 026.00	SINGLE FAMILY	2015-08-14 00:00:00.000	285000	20150819-0083440	No
10	19805	007 14 0 034.00	SINGLE FAMILY	2014-08-29 00:00:00.000	340000	20140909-0082348	No
11	29467	007 14 0A 024.00	SINGLE FAMILY	2015-04-14 00:00:00.000	425000	20150415-0033442	No
12	10754	007 14 0A 027 00	SINGLE FAMILY	2013-12-12 00:00:00 000	585000	20131227-0130352	No

Dataset: Alex the Analyst Guided Project