

Data Cleaning In SQL Project

LAWAL Peter

2024-01-04

Introduction

This is a data cleaning SQL based project using the Nashville Housing dataset. The purpose of this project is to showcase my data cleaning skills in SQL. Although this report and code chunks are viewed in an R Markdown format, this project had been previously executed in Microsoft SQL Server Management Studio.

Installing Packages and Libraries for the purpose of this project, certain packages and libraries will be loaded to make R compatible with MSSQL databases connections. For this project, the DBI, odbc and RODBC packages will be installed and their subsequent packages loaded.

```
library(DBI)
```

```
library(odbc)
```

```
library(RODBC)
```

Connecting to the database Server a database connection will be established by creating a DSN connection to SQL_Server locally hosted on my desktop. This connection will be named `con2`

```
con2 <- dbConnect(odbc :: odbc(), "SQL_Server_DSN")
```

viewing the table to be worked with using a basic SQL query, I will preview the data provided in the database

```
Select *  
From [NashvilleHousing2]
```

Table 1: Displaying records 1 - 10

[illegible]

UniqID	ParcelID	LandUse	Property	SaleDate	SalePrice	SalePrice	SalePrice	Owner	Address	Area	Age	Dist	Tract	Building	Area	Bed	Full	Bath
3894571	02	RESIDENTIAL	CONDOVILLA 10-014.00	2015-01-01	220000	20151029	029-NA	NA	NA NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
			VIEW 26 CT, BRENT-WOOD			0110208												
4619771	02	RESIDENTIAL	CONDOVILLA 04-018.00	2016-01-01	232500	20160506	006-NA	NA	NA NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
			VIEW 27 CT, BRENT-WOOD			0045160												

the next line of queries is to order the data by the land use type. The result is that the dataset is ordered by the land use type in alphabetical order.

```
Select *
From NashvilleHousing2
Order by LandUse
```

Table 2: Displaying records 1 - 10

UniqID	ParcelID	LandUse	Property	SaleDate	SalePrice	SalePrice	SalePrice	Owner	Address	Area	Age	Dist	Tract	Building	Area	Bed	Full	Bath
3047992	04	APARTMENT	LOW RISE MAN BUILT SINCE 1960)	2015-05-07	200000	20050508	008-E3	1601	1.21	URBAN	72002	047903	2420	09580	0	0	0	0
			HERMAN ST, NASHVILLE			0042350	CON-STRUC-TION	HERMAN ST, NASHVILLE, TN			SER-VICES							
3048092	04	APARTMENT	LOW RISE MAN BUILT SINCE 1960)	2015-05-07	200000	20050508	008-E3	1601	1.13	URBAN	79000	058990	6624	09580	0	0	0	0
			HERMAN ST, NASHVILLE			0042350	CON-STRUC-TION	HERMAN ST, NASHVILLE, TN			SER-VICES							
2542106	00	CHURCH	HAVEN-12-HILL DR, NASHVILLE	2014-12-19	4612300	41123	023-OLIVE	938	5.67	URBAN	10210	000200	1530	09653	1	1	1	1
						0117321	BRANCH-MIS-SION-ARY BAP-TIST CHURCH	HAVEN-HILL DR, NASHVILLE, TN			SER-VICES							

UniqueID	ParcelID	LandUse	PropertyAddress	SaleDate	SalePrice	RefNo	OwnerName	OwnerAddress	Acres	Tag	Dist	Tract	Building	TotVal	NetVal	Bed	Floor	Bath
7684092	CHURCH1300	B808	PARK AVE, NASHVILLE	2013400000	20131003	003	3808	3808	0.69	URBAN	N0400	005000	019700	72NA	0	0		
							PARK AV-ENUE PART-NERS, LLC	PARK AVE, NASHVILLE, TN		SER-VICES DIS-TRICT								
2519072	CHURCH1500	H530	RIVER-12-SIDE DR, NASHVILLE	2014750000	20141022	022	SNAPSHOOT	0530	0.39	URBAN	N6990	005300	025200	37NA	0	0		
							DE-VEL-OP-MENT, LLC	RIVER-SIDE DR, NASHVILLE, TN		SER-VICES DIS-TRICT								
2118559	CHURCH1500	H606	GRANNY WHITE PIKE, BRENT-WOOD	2014452500	20140025	025	YOUNG, ALAN	5606	1.37	GENE	R4100	009600	089600	654	2	1		
							GRANNY WHITE PIKE, BRENT-WOOD, TN			SER-VICES DIS-TRICT								
2013073	CHURCH1500	H2800	MC-GAV-OCK PIKE, NASHVILLE	2014423500	20140018	018	CATHOLIC	2800	34.64	URBAN	N38640	002971	830100	980NA	0	0		
							DIO-CESE OF NASHVILLE, TN	MC-GAV-OCK PIKE, NASHVILLE, TN		SER-VICES DIS-TRICT								
5433342	CHURCH1500	H215	BELLE-VUE RD, NASHVILLE	2016950000	20161004	004	CRJ EVENT CEN-TER, LLC	215	2.76	GENE	R4100	005820	004550	71NA	0	0		
							BELLE-VUE RD, NASHVILLE, TN			SER-VICES DIS-TRICT								
5168605	CHURCH1300	H2314	9TH AVE S, NASHVILLE	2016460000	20160018	018	JONES, PE-TER & CATHERINE	2314	0.38	URBAN	N40050	009001	194900	50NA	0	0		
							9TH AVE S, NASHVILLE, TN			SER-VICES DIS-TRICT								
730082	CHURCH1600	H816	SHELBY AVE, NASHVILLE	2013440000	20130028	028	BURT, STEVE C.	816	0.20	URBAN	N60000	001360	073600	9400	0	0		
							SHELBY AVE, NASHVILLE, TN			SER-VICES DIS-TRICT								

Changing the format of saledate in raw data from datetime format to date as time is not necessary in this data, also, the presence of time makes the data hard to read.

```
Select
SaleDate, CONVERT(date, SaleDate) as NewSaleDate
from NashvilleHousing2
```

Table 3: Displaying records 1 - 10

SaleDate	NewSaleDate
2016-09-20	2016-09-20
2015-01-16	2015-01-16
2014-11-01	2014-11-01
2013-09-12	2013-09-12
2013-11-22	2013-11-22
2014-01-22	2014-01-22
2014-08-28	2014-08-28
2016-09-09	2016-09-09
2015-10-26	2015-10-26
2016-04-27	2016-04-27

We see the result of the query as two columns containing the NewSaleDate column
the NashvilleHousing2 dataset will then be updated with theupdate' function

```
Update NashvilleHousing2
Set SaleDate = CONVERT(Date, SaleDate)
```

a new column named SaleDateConverted is also added to allow for readability of the data

```
Alter Table NashvilleHousing2
Add SaleDateConverted Date;
```

the table will then be updated with the code chunk below and we can view the update of the table with the subsequent code chunk.

```
Update NashvilleHousing2
Set SaleDateConverted = Convert(date, SaleDate)
```

```
Select *
From NashvilleHousing2
```

Table 4: Displaying records 1 - 10

UniqueID	ParcelID	LandUse	Property	SaleDate	SalePrice	SoldDate	Owner	Address	City	State	Zip	Lat	Long	Area	Bedrooms	Bathrooms	YearBuilt	YearRenovated	SaleDateConverted
5392271	00	SINGLE-FAM-ILY	1500 LAND 20 DR, BRENT-WOOD	2016-09-20	300000	2016-09-20	CHANCE MATTHEW RAY & LEANN KAY & ET AL	5900 CLOVER- LAND DR, BRENT-WOOD, TN	1.10	URBAN	37209	35.9400	-86.7780	1778	00	313	1	0	2016-09-20

Unique ID	Parcel ID	Land Use	Property	Sale Price	Sub Price	Sold Price	Acres	Vacant	Owner	Address	Size	Dist	Int	Val	Log	W	B	F	H	B	S	Date	Converted
25778000202.00	771	VACANT	CLOVER DR, BRENTWOOD LAND	2014	2000	2015	0.93	NA	0	0.93	URBAN	NA	NA	42900	42900	NA	NA	0	0	0	0	2015-01-16	
2339002013.00	071	SINGLEFAMILY	OAKES DR, BRENTWOOD	2014	25000	2014	1.36	NA	5672	1.36	URBAN	NA	NA	16300	16300	08070	12700	09843	2	0	0	2014-11-01	
75861020A002.00	171	RESIDENTIAL	VILLA VIEW CT, BRENTWOOD	2013	16900	2013	0.16	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2013-09-12	
94191020A003.00	171	CONDVILLA	VILLA VIEW CT, BRENTWOOD	2013	19500	2013	0.16	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2013-11-22	
11405020A008.00	171	CONDVILLA	VILLA VIEW CT, BRENTWOOD	2014	16500	2014	0.16	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2014-01-22	
19129020A011.00	171	RESIDENTIAL	VILLA VIEW CT, BRENTWOOD	2014	19000	2014	0.16	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2014-08-28	
53923020A011.00	171	RESIDENTIAL	VILLA VIEW CT, BRENTWOOD	2016	25500	2016	0.15	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2016-09-09	
38945020A014.00	171	RESIDENTIAL	VILLA VIEW CT, BRENTWOOD	2015	22000	2015	0.16	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2015-10-26	

UniqueID	ParcelID	LandUse	Property	Style	YearBuilt	SALEPRICE	OriginalPrice	Sold	Owner	Address	Area	Size	District	Building	Fire	Water	Pool	Room	Full	Null	SaleDate	Converted
4619771	02	RESIDENTIAL	CONDO	VILLA 04-VIEW 27 CT, BRENT-WOOD	2016	232500	20160506	NA	NA	NA NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2016-04-27	

The table has been confirmed updated with the necessary column.

Next, I will like to see what datapoints have nulls in them and how they can be fixed.

```
Select *
From NashvilleHousing2
Where PropertyAddress is null
```

Table 5: Displaying records 1 - 10

UniqueID	ParcelID	LandUse	Property	Style	YearBuilt	SALEPRICE	OriginalPrice	Sold	Owner	Address	Area	Size	District	Building	Fire	Water	Pool	Room	Full	Null	SaleDate	Converted
4307070	025	SINGLEFAM-ILY	01-15	0005776	2016	179900	20160120	NA	COSTNER, FRED & CAR-OLYN	ROSE-OF HILL CT, GOODLETTSVILLE, TN	0.96	CITY	3000070000	1000010000	09643	1	0				2016-01-15	
3943026	01	VACANRES-I-	10-23	0109602	2015	453000	20151028	NA	SHACKLEFORD, MICHAEL C., JR.	LEWIS, 0.17 MILE PIKE, GOODLETTSVILLE, TN	0.17	CITY	2110021600	012700	00153	2	0				2015-10-23	
4529005	026	SINGLEFAM-ILY	03-29	0029941	2016	455000	20160330	NA	TRIPP, MARVIN S. & DEB-O-RAH YOUNG	208 EAST AVE, GOODLETTSVILLE, TN	0.20	CITY	2110030200	051300	00083	2	0				2016-03-29	
5314076	026	RESIDENTIAL	CONDO	08-25	2016	644490	20160831	NA	NA	NA NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2016-08-25	
4308033	06	SINGLEFAM-ILY	01-04	0001526	2016	470000	20160707	NA	FRANK, ZACHARY & NIKI	1129 CAMP-BELL RD, GOODLETTSVILLE, TN	0.24	GENE	3300010500	004550	00003	2	0				2016-01-04	

UniqueID	ParcelID	Land Use	Property Style	Year Began	Referral	Owner	Address	Area	Tag	District	Building	Area	Value	Bedrooms	Bathrooms	Null	Build Date	Converted
4529033	06	SINGL FAM-0A	ILY	2016210003-03-29	20160331-0030709	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2016-03-29	
4873033	15	SINGL FAM-0	ILY	2016199905-05-05	20160506-0045368	COLEMAN, AARON	138 W CAMP- A. & BELL	1.39	GENE	1100	003001353009542	1	0				2016-05-05	
	123.00					CE- RD, CIL, COR- RIE J.	GOODLETTSVILLE, TN											
3653034	03	SINGL FAM-0	ILY	2015245008-08-13	20150819-0083759	DILICK, JOHN	2117 PAULA DR, & MADI- AN- SON, NETTE TN	1.01	GENE	13200	070002283009644	3	0				2015-08-13	
4691034	07	VACANT RES-0B	I-	2016400004-04-27	20160304-0020905	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2016-04-27	
	015.00	DEN- TIAL LAND																
4426034	16	VACANT RES-0A	I-	2016130002-02-04	20160205-0011327	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2016-02-04	
	004.00	DEN- TIAL LAND																

29 rows of data were found where the property address is null.

In order to figure out if the null data stemmed from simple omissions, I use a join statement to check for identical parcelIDs and join them.

```

Select a.ParcelID, a.PropertyAddress, b.ParcelID, b.PropertyAddress
From Portfolio_Project..NashvilleHousing2 a
Join Portfolio_Project..NashvilleHousing2 b
On a.ParcelID = b.ParcelID
And a.[UniqueID ] <> b.[UniqueID ]
Where a.PropertyAddress is null

```

Table 6: Displaying records 1 - 10

ParcelID	PropertyAddress	ParcelID	PropertyAddress
025 07 0 031.00	NA	025 07 0 031.00	410 ROSEHILL CT, GOODLETTSVILLE
026 01 0 069.00	NA	026 01 0 069.00	141 TWO MILE PIKE, GOODLETTSVILLE

ParcelID	PropertyAddress	ParcelID	PropertyAddress
026 05 0 017.00	NA	026 05 0 017.00	208 EAST AVE, GOODLETTSVILLE
026 06 0A 038.00	NA	026 06 0A 038.00	109 CANTON CT, GOODLETTSVILLE
033 06 0 041.00	NA	033 06 0 041.00	1129 CAMPBELL RD, GOODLETTSVILLE
033 06 0A 002.00	NA	033 06 0A 002.00	1116 CAMPBELL RD, GOODLETTSVILLE
033 15 0 123.00	NA	033 15 0 123.00	438 W CAMPBELL RD, GOODLETTSVILLE
034 03 0 059.00	NA	034 03 0 059.00	2117 PAULA DR, MADISON
034 03 0 059.00	NA	034 03 0 059.00	2117 PAULA DR, MADISON
034 07 0B 015.00	NA	034 07 0B 015.00	2524 VAL MARIE DR, MADISON

The result is a table where property addresses have been parsed based on the parcel IDs found for properties in the dataset.

```
Select a.ParcelID, a.PropertyAddress, b.ParcelID, b.PropertyAddress, ISNULL(a.PropertyAddress, b.PropertyAddress)
From Portfolio_Project..NashvilleHousing2 a
Join Portfolio_Project..NashvilleHousing2 b
On a.ParcelID = b.ParcelID
And a.[UniqueID ] <> b.[UniqueID ]
Where a.PropertyAddress is null
```

Table 7: Displaying records 1 - 10

ParcelID	PropertyAddress	ParcelID	PropertyAddress	PropertyAddress
025 07 0 031.00	NA	025 07 0 031.00	410 ROSEHILL CT, GOODLETTSVILLE	410 ROSEHILL CT, GOODLETTSVILLE
026 01 0 069.00	NA	026 01 0 069.00	141 TWO MILE PIKE, GOODLETTSVILLE	141 TWO MILE PIKE, GOODLETTSVILLE
026 05 0 017.00	NA	026 05 0 017.00	208 EAST AVE, GOODLETTSVILLE	208 EAST AVE, GOODLETTSVILLE
026 06 0A 038.00	NA	026 06 0A 038.00	109 CANTON CT, GOODLETTSVILLE	109 CANTON CT, GOODLETTSVILLE
033 06 0 041.00	NA	033 06 0 041.00	1129 CAMPBELL RD, GOODLETTSVILLE	1129 CAMPBELL RD, GOODLETTSVILLE
033 06 0A 002.00	NA	033 06 0A 002.00	1116 CAMPBELL RD, GOODLETTSVILLE	1116 CAMPBELL RD, GOODLETTSVILLE
033 15 0 123.00	NA	033 15 0 123.00	438 W CAMPBELL RD, GOODLETTSVILLE	438 W CAMPBELL RD, GOODLETTSVILLE
034 03 0 059.00	NA	034 03 0 059.00	2117 PAULA DR, MADISON	2117 PAULA DR, MADISON
034 03 0 059.00	NA	034 03 0 059.00	2117 PAULA DR, MADISON	2117 PAULA DR, MADISON
034 07 0B 015.00	NA	034 07 0B 015.00	2524 VAL MARIE DR, MADISON	2524 VAL MARIE DR, MADISON

with the below code chunk, the dataset will be updated using the datapoints gotten using the unique ID

```
Update a
Set PropertyAddress = ISNULL(a.PropertyAddress, b.PropertyAddress)
From Portfolio_Project..NashvilleHousing2 a
Join Portfolio_Project..NashvilleHousing2 b
On a.ParcelID = b.ParcelID
and a.[UniqueID ] <> b.[UniqueID ]
WHEre a.PropertyAddress is null
```

I will then check to confirm the dataset has been cleaned

```
Select *
From NashvilleHousing2
Where PropertyAddress is null
```

Table 8: 0 records

UniqueID	ParcelID	PropertyAddress	Address	City	State	Zip	Latitude	Longitude	Neighborhood	Area	Population	HouseAge	FullBath	HalfBath	Bedroom	YearBuilt	Value	YearRenovated	Converted
----------	----------	-----------------	---------	------	-------	-----	----------	-----------	--------------	------	------------	----------	----------	----------	---------	-----------	-------	---------------	-----------

all null datapoints have been corrected for the property address.

Up next, I will break the property id into separate columns for readability

```
Select PropertyAddress
From NashvilleHousing2
```

Table 9: Displaying records 1 - 10

PropertyAddress
5900 CLOVERLAND DR, BRENTWOOD
CLOVERLAND DR, BRENTWOOD
5672 OAKES DR, BRENTWOOD
103 VILLA VIEW CT, BRENTWOOD
105 VILLA VIEW CT, BRENTWOOD
115 VILLA VIEW CT, BRENTWOOD
104 VILLA VIEW CT, BRENTWOOD
104 VILLA VIEW CT, BRENTWOOD
110 VILLA VIEW CT, BRENTWOOD
118 VILLA VIEW CT, BRENTWOOD

I will separate the address into separate columns using the comma separator in the CHARINDEX function

```
Select
SUBSTRING (PropertyAddress, 1, CHARINDEX(',', PropertyAddress)) as Address
From NashvilleHousing2
```

Table 10: Displaying records 1 - 10

Address
5900 CLOVERLAND DR, CLOVERLAND DR, 5672 OAKES DR, 103 VILLA VIEW CT, 105 VILLA VIEW CT, 115 VILLA VIEW CT, 104 VILLA VIEW CT, 104 VILLA VIEW CT, 110 VILLA VIEW CT, 118 VILLA VIEW CT,

the above chunk splits just the first part of the address line

the next code chunk will split the address line by two places

```
Select
SUBSTRING (PropertyAddress, 1, CHARINDEX(',', PropertyAddress)-1) as Address,
Substring (PropertyAddress, CharIndex(',', PropertyAddress) + 1, Len(PropertyAddress)) as Address
From NashvilleHousing2
```

Table 11: Displaying records 1 - 10

Address	Address
5900 CLOVERLAND DR	BRENTWOOD
CLOVERLAND DR	BRENTWOOD
5672 OAKES DR	BRENTWOOD
103 VILLA VIEW CT	BRENTWOOD
105 VILLA VIEW CT	BRENTWOOD
115 VILLA VIEW CT	BRENTWOOD
104 VILLA VIEW CT	BRENTWOOD
104 VILLA VIEW CT	BRENTWOOD
110 VILLA VIEW CT	BRENTWOOD
118 VILLA VIEW CT	BRENTWOOD

I will then update the table with the next code chunks `nvarchar` specifies the formatting of the data in the column that is being created.

```
Alter Table NashvilleHousing2
Add PropertySplitAddress nvarchar(225);
```

```
Update NashvilleHousing2
Set PropertySplitAddress = SUBSTRING (PropertyAddress, 1, CHARINDEX(',', PropertyAddress)-1)
```

```
Alter Table NashvilleHousing2
Add PropertyCity nvarchar(225);
```

```

Update NashvilleHousing2
Set PropertyCity = Substring (PropertyAddress, CharIndex(',', PropertyAddress) + 1, Len(PropertyAddress))

Select *
From NashvilleHousing

```

Table 12: Displaying records 1 - 10

UniqueID	ParcelID	LandSale	Price	RefersTo	OwnerName	Area	Age	BuildType	YearBuilt	Bedrooms	Bathrooms	SaleDate	Price	County	Subdiv	City	State
2045007	SING-240020130412-00 FAMILY	125.00	0036474	12-	FRAZIER, CYREN-THA LYNETTE	2,305	0000	068203	2013	3	0	1808-04-09	1808	GOODE	DETTCS	ODNE	TTTSVILLE
1691807	SING-366020140619-00 FAMILY	130.00	0053768	19-	BONER, CHARLES & LESLIE	3,505	0000	064103	2014	3	2	1832-06-10	1832	GOODE	DETTCS	ODNE	TTTSVILLE
5458207	SING-435020160927-00 FAMILY	138.00	0101718	27-	WILSON, JAMES E. & JOANNE	2,905	0000	016209	2016	3	0	1864-09-26	1864	GOODE	DETTCS	ODNE	TTTSVILLE
4307007	SING-255020160129-00 FAMILY	143.00	0008913	29-	BAKER, JAY K. & SU-SAN E.	2,605	0000	047309	2016	3	0	1853-01-29	1853	GOODE	DETTCS	ODNE	TTTSVILLE
2271407	SING-278020141015-00 FAMILY	149.00	0095255	15-	POST, CHRISTOPHER M. & SAMANTHA C.	2,005	0000	052302	2014	3	0	1829-10-10	1829	GOODE	DETTCS	ODNE	TTTSVILLE
1836707	SING-267020140718-00 FAMILY	151.00	0063802	18-	FIELDS, KAREN L. & BRENT A.	2,005	0000	090402	2014	3	0	1821-07-16	1821	GOODE	DETTCS	ODNE	TTTSVILLE
1980407	SING-177020140903-14 FAMILY	002.00	0080214	03-	HINTON, MICHAEL R. & CYNTHIA M. MOORE	1,034	0000	037907	2014	2	0	2005-08-28	2005	GOODE	DETTCS	ODNE	TTTSVILLE
5458307	SING-262020161015-14 FAMILY	024.00	0105441	15-	BAILOR, DARRELL & TAMMY	1,034	0000	057909	2016	2	0	1917-09-27	1917	GOODE	DETTCS	ODNE	TTTSVILLE

UniqueID	ParcelID	State	Price	Ref	Owner	Name	Age	Build	Year	Bed	Bath	Year	Pool	Dr	Core	By	Sp	City	State
365007	SING-185000150819	00	00150819	00	ROBERTS	6745400769022300003	2	1	2015-	1428	GOOD	DET	CS	MO	NE			ETTSVILLE	
14	FAM-	0083440			MISTY				08-	SPRING-	SPRING-								
0	ILY				L. &				14	FIELD	FIELD								
026.00					ROBERT					HWY	HWY								
					M.														
198007	SING-140000140909	00	00140909	00	LEE,	1.30400007960219600955	3	0	2014-	1420	GOOD	DET	CS	MO	NE			ETTSVILLE	
14	FAM-	0082348			JEF-				08-	SPRING-	SPRING-								
0	ILY				FREY				29	FIELD	FIELD								
034.00					&					HWY	HWY								
					NANCY														

The same operation that was performed with property address above will also be performed on the owner address, but with the PARSENAME function.

```
Select OwnerAddress
From NashvilleHousing2
```

Table 13: Displaying records 1 - 10

OwnerAddress
5900 CLOVERLAND DR, BRENTWOOD, TN
0 CLOVERLAND DR, BRENTWOOD, TN
5672 OAKES DR, BRENTWOOD, TN
NA
NA
NA
NA
NA
NA
NA

```
Select
PARSENAME(Replace(OwnerAddress, ',', '.'), 3),
PARSENAME(Replace(OwnerAddress, ',', '.'), 2) ,
PARSENAME(Replace(OwnerAddress, ',', '.'), 1)
From NashvilleHousing2
```

Table 14: Displaying records 1 - 10

5900 CLOVERLAND DR	BRENTWOOD	TN
0 CLOVERLAND DR	BRENTWOOD	TN
5672 OAKES DR	BRENTWOOD	TN
NA	NA	NA
NA	NA	NA
NA	NA	NA
NA	NA	NA
NA	NA	NA
NA	NA	NA
NA	NA	NA

NA NA NA

Now that the effect of the above functions is known, the table will then be updated

```
Alter Table NashvilleHousing2
Add OwnerAddressSplit nvarchar(225);
```

```
Update NashvilleHousing2
Set OwnerAddressSplit = PARSENAME(Replace(OwnerAddress, ',', '.'), 3)
```

```
Alter Table NashvilleHousing2
Add OwnerAddressCity nvarchar(225);
```

```
Update NashvilleHousing2
Set OwnerAddressCity = PARSENAME(Replace(OwnerAddress, ',', '.'), 2)
```

```
Alter Table NashvilleHousing2
Add OwnerAddressState nvarchar(225);
```

```
Update NashvilleHousing2
Set OwnerAddressState= PARSENAME(Replace(OwnerAddress, ',', '.'), 1)
```

```
Select *
From NashvilleHousing
```

Table 15: Displaying records 1 - 10

Unit	Parcel ID	Land Sale Price	Referral	Owner Name	Area	Age	Build	Log	Year	Bed	Floor	Bath	Sale Date	County	Sp	Qtr	Ch	Qtr	Ch	City	State
2045007	SING 240020130412	240020130412	FRAZIER, 2,305	00000682035700863	3	0	2013-1808	GOOD	DETC	GOOD	DETC	GOOD	DETC	GOOD	DETC	GOOD	DETC	GOOD	DETC	DETTSVILLE	
	00 FAM- 0036474	0036474	CYREN-				04- FOX						09- CHASE								
	0 ILY		THA										DR								
	125.00		LYNETTE										DR								
1691807	SING 366020140519	366020140519	BONER, 3,505	0000026410319000983	3	2	2014-1832	GOOD	DETC	GOOD	DETC	GOOD	DETC	GOOD	DETC	GOOD	DETC	GOOD	DETC	DETTSVILLE	
	00 FAM- 0053768	0053768	CHARLES				06- FOX						10- CHASE								
	0 ILY		&										DR								
	130.00		LESLIE										DR								
5458207	SING 435020160927	435020160927	WILSON, 2,905	0000021620298000874	3	0	2016-1864	GOOD	DETC	GOOD	DETC	GOOD	DETC	GOOD	DETC	GOOD	DETC	GOOD	DETC	DETTSVILLE	
	00 FAM- 0101718	0101718	JAMES				09- FOX						26- CHASE								
	0 ILY		E. &										DR								
	138.00		JOANNE										DR								
4307007	SING 255020160129	255020160129	BAKER, 2,605	000004730097300853	3	0	2016-1853	GOOD	DETC	GOOD	DETC	GOOD	DETC	GOOD	DETC	GOOD	DETC	GOOD	DETC	DETTSVILLE	
	00 FAM- 0008913	0008913	JAY				01- FOX						29- CHASE								
	0 ILY		K. &										DR								
	143.00		SU-SAN										DR								
			E.																		

UniqueID	ParcelID	State	Price	SoldAsVacant	OwnerName	Acres	Days	BuildingArea	YearBuilt	Bedrooms	Bathrooms	SaleDate	DaysToSell	County	City	State
2271007	SING-000	278000	141115	N	POST, CHRISTOPHER M. & SAMANTHA C.	2.005	0000	52300	2008	4	3	0	2014-10-10	1829	GOODLETTSVILLE	TN
1836707	SING-000	267000	140718	N	FIELDS, KAREN L. & BRENT A.	2.005	0000	90400	2008	3	0	0	2014-07-16	1821	GOODLETTSVILLE	TN
1980007	SING-140	177000	20140903	N	HINTON, MICHAEL R. & CYNTHIA M. MOORE	1.034	0000	37900	2007	2	0	0	2014-08-28	2005	GOODLETTSVILLE	TN
5458007	SING-140	262000	20161105	N	BAILOR, DARRELL & TAMMY	1.034	0000	57900	2009	2	0	0	2016-09-27	1917	GOODLETTSVILLE	TN
3650007	SING-140	285000	20150819	N	ROBERTS, MISTY L. & ROBERT M.	6.745	4000	76900	2002	2	1	0	2015-08-14	1428	GOODLETTSVILLE	TN
1980007	SING-140	340000	20140909	N	LEE, JEFFREY & NANCY	1.304	0000	79600	2001	3	0	0	2014-08-29	1420	GOODLETTSVILLE	TN

Investigating the Data further, I found in the SoldAsVacant column some entries are in the forms yes, no, Y and N, to not skew the data, or have errors, I changed them to a consistent form Yes and No

```
Select Distinct (SoldAsVacant), Count(SoldAsVacant)
From NashvilleHousing2
Group by SoldAsVacant
Order by 2
```

Table 16: 4 records

SoldAsVacant	
Y	52
N	399
Yes	4623
No	51403

```

Select
SoldAsVacant
, Case WHEN SoldAsVacant = 'Y' Then 'Yes'
      when SoldAsVacant = 'N' then 'No'
      Else SoldAsVacant
      END
from NashvilleHousing2

```

Table 17: Displaying records 1 - 10

SoldAsVacant	
No	No
Yes	Yes
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No

```

Update
NashvilleHousing2
Set SoldAsVacant = Case WHEN SoldAsVacant = 'Y' Then 'Yes'
      when SoldAsVacant = 'N' then 'No'
      Else SoldAsVacant
      END

```

The table has been updated with the corrections that were needed.

Next up I will be removing duplicates within the dataset and creating a Common Table Expression CTE

```

With RowNumCTE As(
Select *,
      ROW_NUMBER () Over (
      Partition By ParcelID,
                  PropertyAddress,
                  SalePrice,
                  SaleDate,
                  LegalReference
      Order by
                  UniqueID) row_num
From
NashvilleHousing2
--Order by ParcelID
)
--
Select *
from RowNumCTE
Where row_num > 1
Order by PropertyAddress

```


Table 18: Displaying records 1 - 10

[illegible]

UnitID	ParcelID	PropType	SaleDate	SalePrice	Address	LegalReference	OwnerName	OwnerAddress	OwnerAddressCity	OwnerAddressState	ParcelID	PropType	SaleDate	SalePrice	Address	LegalReference	OwnerName	OwnerAddress	OwnerAddressCity	OwnerAddressState
273598	008	SING	20151002	20050225	NA NA	NANA	NANA	NANA	NANA	NANA	2015108	HERMITAGE	NA	2						
09	FAMWIND	02-	0016199								02-	WIND-								
0B	ILY CHASE										18	CHASE								
009.00	RUN,											RUN								
	HER-																			
	MITAGE																			
273675	RESIDENT	00118200	5050224	NA NA	NANA	NANA	NANA	NANA	NANA	NANA	20151129	NASHVILLE	NA	2						
10	CONDO	02-	0015807								02-	RAN-								
0E	SOM 13										13	SOM								
014.00	WAY,											WAY								
	NASHVILLE																			
273598	RESIDENT	00118200	5050227	NA NA	NANA	NANA	NANA	NANA	NANA	NANA	20151207	HERMITAGE	NA	2						
06	CONDO	02-	0017053								02-	CHICK-								
0B	ADEE	26									26	ADEE								
137.00	CIR,											CIR								
	HER-																			
	MITAGE																			

```

With RowNumCTE As(
Select *,
    ROW_NUMBER () Over (
        Partition By ParcelID,
                    PropertyAddress,
                    SalePrice,
                    SaleDate,
                    LegalReference
        Order by
            UniqueID) row_num
From
NashvilleHousing2
--Order by ParcelID
)

DELETE
from RowNumCTE
Where row_num > 1

```

Finally as part of the data cleaning, unwanted columns will be removed from the table such as **OwnerName** to anonymize the data and protect privacy of owners, and the **OwnerAddress** and **OwnerAddressCity**

```

Select *
From NashvilleHousing2

```

Table 19: Displaying records 1 - 10

Unit	Parcel ID	Prop	Style	File	Reg	File	Owner	Acres	Tag	Dist	Val	High	Val	Hold	Half	Back	Date	City	Qtr	Qtr	Assess	Assess	City	Stat
539271	SING	1900	2016	0000	2016	N006	CHAN	5900	1.10	URB	4220540	1778033	B	1	0	2016	5900	BRENTWOOD	09-	CLOVER	CLOVER-	20	LAND	LAND
00	FAMCLOVER-	09-	0106043	MATTHEW	SER-												09-	CLOVER	CLOVER-					
0	ILY LAND	20		RAY	LAND	VICES											20	LAND	LAND					
152.00	DR,			&	DR,	DIS-												DR	DR					
	BRENT-			LEAN	BRENT-	TRICT																		
	WOOD			KAY	WOOD,																			
				&	TN																			
				ET																				
				AL																				
257781	VACANT	NOV	2015	0000	2015	N015	NA	0	0.93	URB	42900	42900	ANA	0	0	2015	CLOVER	BRENTWOOD	01-	DR	CLOVER-	16	LAND	DR
00	RES-DR,	01-	0005484	CLOVER	SER-																			
0	I- BRENT-	16-		LAND	VICES																			
202.00	WOOD			DR,	DIS-																			
	TIAL			BRENT-	TRICT																			
	LAND			WOOD,																				
				TN																				
2339071	SING	1672	2014	2500	2014	N016	CARIN	5672	1.36	URB	46806072	10984	B	2	0	2014	5672	BRENTWOOD	11-	OAKES	OAKES	01	DR	DR
02	FAMOAKES	11-	0103531	KEVIN	OAKES	SER-																		
0	ILY DR,	01		F.	DR,	VICES																		
013.00	BRENT-			BRENT-	DIS-																			
	WOOD			WOOD,	TRICT																			
				TN																				
758671	RESID	NOV	2013	6900	2013	N016	NA	NA	NANA	NANA	NANA	NANA	ANA	2013	103	BRENTWOOD	09-	VILLA						
02	CONDO	09-	0097186																					
0A	VIEW	12																						
002.00	CT,																							
	BRENT-																							
	WOOD																							
941971	CONDO	NOV	2013	9500	2013	N016	NA	NA	NANA	NANA	NANA	NANA	ANA	2013	105	BRENTWOOD	11-	VILLA						
02	VILLA	11-	0126781																					
0A	VIEW	22																						
003.00	CT,																							
	BRENT-																							
	WOOD																							
1140571	CONDO	NOV	2014	6500	2014	N017	NA	NA	NANA	NANA	NANA	NANA	ANA	2014	115	BRENTWOOD	01-	VILLA						
02	VILLA	11-	0007078																					
0A	VIEW	22																						
008.00	CT,																							
	BRENT-																							
	WOOD																							
1912971	RESID	NOV	2014	9000	2014	N017	NA	NA	NANA	NANA	NANA	NANA	ANA	2014	104	BRENTWOOD	08-	VILLA						
02	CONDO	08-	0080767																					
0A	VIEW	28																						
011.00	CT,																							
	BRENT-																							
	WOOD																							

Unique ID	Parcel ID	Land Use	Property Type	Sale Date	Base Price	Original Price	Bedrooms	Baths	Average Sq Ft	Days On Market	Distance To School	Latitude	Longitude	Year Built	Pool	Fire Place	Hard Wood	Deck	Garage	Spa	Hot Tub	Address	State
7586171	02	0A	002.00	RESIDENTIAL CONDOMINIUM	2013-09-12	169000	20130916	06-	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2013-09-12	103	BRENTWOOD VILLA VIEW CT	CA	
9419171	02	0A	003.00	CONDOMINIUM	2013-11-22	195000	20131126	06-	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2013-11-22	105	BRENTWOOD VILLA VIEW CT	CA	
1140571	02	0A	008.00	CONDOMINIUM	2014-01-22	165000	20140127	07-	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2014-01-22	115	BRENTWOOD VILLA VIEW CT	CA	
1912971	02	0A	011.00	RESIDENTIAL CONDOMINIUM	2014-08-28	190000	20140904	04-	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2014-08-28	104	BRENTWOOD VILLA VIEW CT	CA	
5392371	02	0A	011.00	RESIDENTIAL CONDOMINIUM	2016-09-09	255000	20160915	05-	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2016-09-09	104	BRENTWOOD VILLA VIEW CT	CA	
3894571	02	0A	014.00	RESIDENTIAL CONDOMINIUM	2015-10-26	220000	20151029	09-	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2015-10-26	110	BRENTWOOD VILLA VIEW CT	CA	
4619771	02	0A	018.00	RESIDENTIAL CONDOMINIUM	2016-04-27	232500	20160506	06-	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2016-04-27	118	BRENTWOOD VILLA VIEW CT	CA	