---DATA CLEANING IN SQL

---------------------------------------------------------------

--Standardize Date Format

Alter Table Nashville

add SalesDate varchar(255)

Update Nashville

set SalesDate = convert(date, SaleDate)

---Now Delete the Old Date Format Column

Alter Table Nashville

Drop column SaleDate

--------------------------------------------------------------

---Populating the Property Address

--To do this, we self join the table first, excluding duplicated field.

select a.ParcelID, a.PropertyAddress, b.ParcelID, b.PropertyAddress, isnull(a.PropertyAddress,b.PropertyAddress)

from Nashville as a

join Nashville as b

on a.ParcelID = b.ParcelID

and a.[UniqueID ]<>b.[UniqueID ]

where a.PropertyAddress is null

update a

set PropertyAddress = isnull(a.PropertyAddress,b.PropertyAddress)

from Nashville as a

join Nashville as b

on a.ParcelID = b.ParcelID

and a.[UniqueID ]<>b.[UniqueID ]

where a.PropertyAddress is null

----Breaking out the property address into separate columns

select SUBSTRING(PropertyAddress,1,CHARINDEX(',',PropertyAddress)) as 'Address'

from Nashville

--to remove the comma

select SUBSTRING(PropertyAddress,1,CHARINDEX(',',PropertyAddress)-1) as 'Address'

from Nashville

select SUBSTRING(PropertyAddress,1,18) as 'address'

from Nashville

select SUBSTRING(PropertyAddress,1,CHARINDEX(',',PropertyAddress)-1) as 'Address',

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

from Nashville

---ALTERNATIVELY we can split the PropertyAddress column using the PARSENAME command

select

parsename(replace(PropertyAddress, ',','.'),2),

parsename(replace(PropertyAddress, ',','.'),1)

from Nashville

---We now create two new columns for the separated addresses

Alter Table Nashville

add AddressLine1 nvarchar(255)

Update Nashville

set AddressLine1 = SUBSTRING(PropertyAddress,1,CHARINDEX(',',PropertyAddress)-1)

Alter Table Nashville

add AddressLine2 nvarchar(255)

Update Nashville

set AddressLine2 = SUBSTRING(PropertyAddress,CHARINDEX(',',PropertyAddress)+1, LEN(PropertyAddress))

----Breaking out the owner address into separate columns using PARSENAME command

select

parsename(replace(OwnerAddress, ',','.'),3),

parsename(replace(OwnerAddress, ',','.'),2),

parsename(replace(OwnerAddress, ',','.'),1)

from Nashville

---We now create two new columns for the separated Owner address

Alter Table Nashville

add OwnerAddressLine3 nvarchar(255)

Update Nashville

set OwnerAddressLine3 = parsename(replace(OwnerAddress, ',','.'),3)

Alter Table Nashville

add OwnerAddressLine2 nvarchar(255)

Update Nashville

set OwnerAddressLine2 = parsename(replace(OwnerAddress, ',','.'),2)

Alter Table Nashville

add OwnerAddressLine1 nvarchar(255)

Update Nashville

set OwnerAddressLine1 = parsename(replace(OwnerAddress, ',','.'),1)

----Making the "Sold AsVacant" column unform by replacing "N" with "N" and "Y" with "Yes"

---First we run a check

Select distinct SoldASVacant, count(SoldASVacant)

from Nashville

group by SoldASVacant

order by SoldASVacant

---Now we effect the change

select SoldAsVacant,

case when SoldAsVacant = 'N' then 'No'

when SoldAsVacant = 'Y' then 'Yes'

else SoldAsVacant

end

from Nashville

----Now we update the table

update Nashville

set SoldAsVacant = case when SoldAsVacant = 'N' then 'No'

when SoldAsVacant = 'Y' then 'Yes'

else SoldAsVacant

end

----Removing Duplicate Values

With RowNumCTE AS(

select \*,

row\_number()over(

Partition by ParcelID,

PropertyAddress,

SalePrice,

SalesDate,

LegalReference

order by

UniqueID

) row\_num

from Nashville

---Order by ParcelID

)

select \*

from RowNumCTE

where row\_num > 1

order by PropertyAddress

---Now we delete the duplicate rows

With RowNumCTE AS(

select \*,

row\_number()over(

Partition by ParcelID,

PropertyAddress,

SalePrice,

SalesDate,

LegalReference

order by

UniqueID

) row\_num

from Nashville

---Order by ParcelID

)

Delete

from RowNumCTE

where row\_num > 1

----To delete unused columns

Alter table Nashville

Drop column TaxDistrict,OwnerAddress,PropertyAddress

select \*

from Nashville