zctntgcrt

June 30, 2023

[ ]:
```
This source code created by IndianAIProduction.com team
https:\\www.IndianAIProduction.com\handling-missing-Values-data-cleaning

Video on
Methods to Handling Missing Values/Data Part-1: https://youtu.be/cN3i8ktEg54
Handling Missing Values/Data Part-2: https://youtu.be/NqL8XOM9eww
Missing Value Imputation(numeric data) Part-3: https://youtu.be/nhnLdZeKlZk
Missing Value Imputation(numeric data) by class Part-4:https://youtu.be/
 ↪Mf2Tl2bPfz0
Missing Value Imputation - categorical value part-5: https://youtu.be/
 ↪rEJrFmXdkig

for video tutorial visit our youtube channel
www.youtube.com\IndianAIProduction
```

1 Data Cleaning

1.1 Categorical Missing value imputation Part-5

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
df = pd.read_csv(r"G:\DataSet\House Price Prediction\train.csv")
```

[3]:
```
cat_vars = df.select_dtypes(include='object')
cat_vars.head()
```

[3]:

	MSZoning	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	\
0	RL	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	
1	RL	Pave	NaN	Reg	Lvl	AllPub	FR2	Gtl	
2	RL	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	
3	RL	Pave	NaN	IR1	Lvl	AllPub	Corner	Gtl	
4	RL	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl	

	Neighborhood	Condition1	…	GarageType	GarageFinish	GarageQual	GarageCond	\

```
0      CollgCr       Norm  …      Attchd          RFn         TA         TA
1      Veenker      Feedr  …      Attchd          RFn         TA         TA
2      CollgCr       Norm  …      Attchd          RFn         TA         TA
3      Crawfor       Norm  …      Detchd          Unf         TA         TA
4      NoRidge       Norm  …      Attchd          RFn         TA         TA

  PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition
0          Y    NaN   NaN         NaN       WD        Normal
1          Y    NaN   NaN         NaN       WD        Normal
2          Y    NaN   NaN         NaN       WD        Normal
3          Y    NaN   NaN         NaN       WD       Abnorml
4          Y    NaN   NaN         NaN       WD        Normal

[5 rows x 43 columns]
```

[4]:
```
cat_vars.isnull().sum()
```

[4]:
```
MSZoning           0
Street             0
Alley           1369
LotShape           0
LandContour        0
Utilities          0
LotConfig          0
LandSlope          0
Neighborhood       0
Condition1         0
Condition2         0
BldgType           0
HouseStyle         0
RoofStyle          0
RoofMatl           0
Exterior1st        0
Exterior2nd        0
MasVnrType         8
ExterQual          0
ExterCond          0
Foundation         0
BsmtQual          37
BsmtCond          37
BsmtExposure      38
BsmtFinType1      37
BsmtFinType2      38
Heating            0
HeatingQC          0
CentralAir         0
Electrical         1
```

```
KitchenQual         0
Functional          0
FireplaceQu       690
GarageType         81
GarageFinish       81
GarageQual         81
GarageCond         81
PavedDrive          0
PoolQC           1453
Fence            1179
MiscFeature      1406
SaleType            0
SaleCondition       0
dtype: int64
```

[5]:
```
miss_val_per = cat_vars.isnull().mean()*100
miss_val_per
```

[5]:
```
MSZoning          0.000000
Street            0.000000
Alley            93.767123
LotShape          0.000000
LandContour       0.000000
Utilities         0.000000
LotConfig         0.000000
LandSlope         0.000000
Neighborhood      0.000000
Condition1        0.000000
Condition2        0.000000
BldgType          0.000000
HouseStyle        0.000000
RoofStyle         0.000000
RoofMatl          0.000000
Exterior1st       0.000000
Exterior2nd       0.000000
MasVnrType        0.547945
ExterQual         0.000000
ExterCond         0.000000
Foundation        0.000000
BsmtQual          2.534247
BsmtCond          2.534247
BsmtExposure      2.602740
BsmtFinType1      2.534247
BsmtFinType2      2.602740
Heating           0.000000
HeatingQC         0.000000
CentralAir        0.000000
```

```
Electrical        0.068493
KitchenQual       0.000000
Functional        0.000000
FireplaceQu      47.260274
GarageType        5.547945
GarageFinish      5.547945
GarageQual        5.547945
GarageCond        5.547945
PavedDrive        0.000000
PoolQC           99.520548
Fence            80.753425
MiscFeature      96.301370
SaleType          0.000000
SaleCondition     0.000000
dtype: float64
```

[6]:
```
drop_vars=['Alley','FireplaceQu','PoolQC','Fence','MiscFeature']
cat_vars.drop(columns=drop_vars, axis=1, inplace=True)
cat_vars.shape
```

```
C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\frame.py:3997:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  errors=errors,
```

[6]:
```
(1460, 38)
```

[7]:
```
isnull_per=cat_vars.isnull().mean()*100
miss_vars = isnull_per[isnull_per >0].keys()
miss_vars
```

[7]:
```
Index(['MasVnrType', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1',
       'BsmtFinType2', 'Electrical', 'GarageType', 'GarageFinish',
       'GarageQual', 'GarageCond'],
      dtype='object')
```

[8]:
```
cat_vars['MasVnrType'].fillna('Missing')
```

[8]:
```
0       BrkFace
1          None
2       BrkFace
3          None
4       BrkFace
         …
```

```
1455       None
1456      Stone
1457       None
1458       None
1459       None
Name: MasVnrType, Length: 1460, dtype: object
```

[9]:
```
cat_vars['MasVnrType'].mode()
```

[9]:
```
0    None
dtype: object
```

[10]:
```
cat_vars['MasVnrType'].value_counts()
```

[10]:
```
None       864
BrkFace    445
Stone      128
BrkCmn      15
Name: MasVnrType, dtype: int64
```

[11]:
```
cat_vars['MasVnrType'].fillna(cat_vars['MasVnrType'].mode()[0])
```

[11]:
```
0       BrkFace
1          None
2       BrkFace
3          None
4       BrkFace
          ...
1455       None
1456      Stone
1457       None
1458       None
1459       None
Name: MasVnrType, Length: 1460, dtype: object
```

[12]:
```
cat_vars['MasVnrType'].fillna(cat_vars['MasVnrType'].mode()[0]).value_counts()
```

[12]:
```
None       872
BrkFace    445
Stone      128
BrkCmn      15
Name: MasVnrType, dtype: int64
```

[13]:
```
cat_vars_copy= cat_vars.copy()

for var in miss_vars:
    cat_vars_copy[var].fillna(cat_vars[var].mode()[0],inplace=True)
```

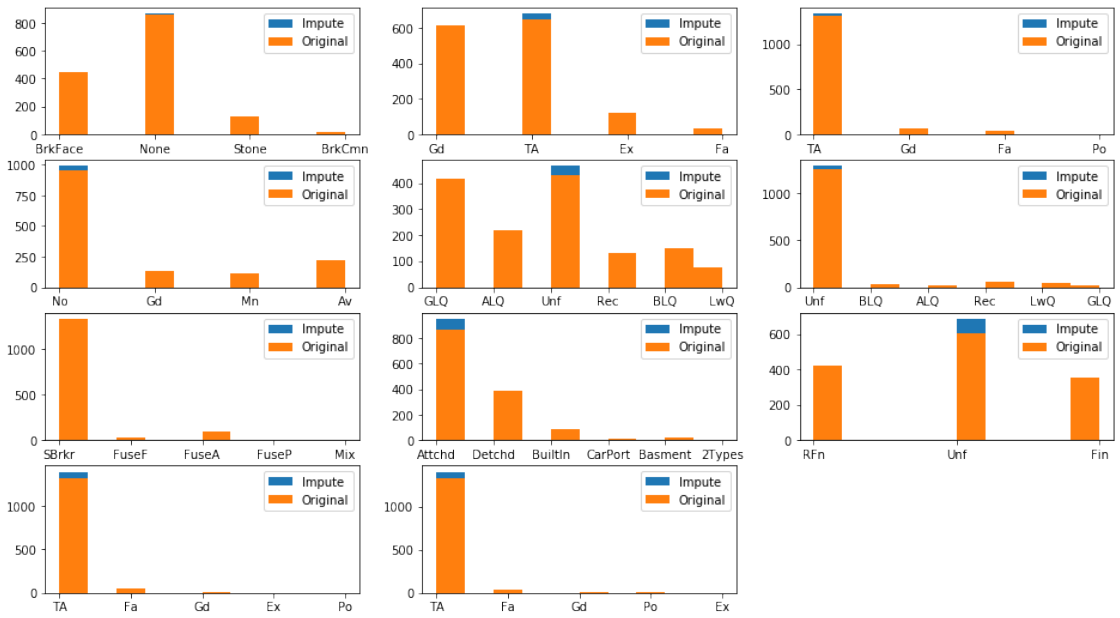
```
    print(var,"=",cat_vars[var].mode()[0])
```

```
MasVnrType = None
BsmtQual = TA
BsmtCond = TA
BsmtExposure = No
BsmtFinType1 = Unf
BsmtFinType2 = Unf
Electrical = SBrkr
GarageType = Attchd
GarageFinish = Unf
GarageQual = TA
GarageCond = TA
```

[14]:
```
cat_vars_copy.isnull().sum().sum()
```

[14]: 0

[15]:
```
plt.figure(figsize=(16,9))
for i,var in enumerate(miss_vars):
    plt.subplot(4,3,i+1)
    plt.hist(cat_vars_copy[var],label="Impute")
    plt.hist(cat_vars[var].dropna(),label="Original")
    plt.legend()
```



[16]:
```
df.update(cat_vars_copy)
df.drop(columns=drop_vars,inplace=True)
```

[17]:
```
df.select_dtypes(include='object').isnull().sum()
```

[17]:
```
MSZoning         0
Street           0
LotShape         0
LandContour      0
Utilities        0
LotConfig        0
LandSlope        0
Neighborhood     0
Condition1       0
Condition2       0
BldgType         0
HouseStyle       0
RoofStyle        0
RoofMatl         0
Exterior1st      0
Exterior2nd      0
MasVnrType       0
ExterQual        0
ExterCond        0
Foundation       0
BsmtQual         0
BsmtCond         0
BsmtExposure     0
BsmtFinType1     0
BsmtFinType2     0
Heating          0
HeatingQC        0
CentralAir       0
Electrical       0
KitchenQual      0
Functional       0
GarageType       0
GarageFinish     0
GarageQual       0
GarageCond       0
PavedDrive       0
SaleType         0
SaleCondition    0
dtype: int64
```

[1]:
```
print("Ab milenge next tutorial me,\nTab tak ke liye SIKHATE SIKHATE kuch␣
  ↪IMPLEMENT karte raho,\nThank You.....-:)")
```

```
Ab milenge next tutorial me,
Tab tak ke liye SIKHATE SIKHATE kuch IMPLEMENT karte raho,
```

Thank You…-:)