yyla8fkkx

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
Missing Value Imputation - using scikit-learnn part-6:https://youtu.be/
  ↪sRk3GoyJPtU

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

About Scikit-Learn:
-------------------
scikit-learn official site: https://scikit-learn.org/stable/
installation of scikit-learn: https://scikit-learn.org/stable/install.html
sklearn.impute.SimpleImputer: https://scikit-learn.org/stable/modules/generated/
  ↪sklearn.impute.SimpleImputer.html
```

1 Data Cleaning

1.1 Missing value imputation using Scikit-Learn

1.1.1 for Numeric and Categorical Variables/Data

[2]:

```
import numpy as np
import pandas as pd
from sklearn.impute import SimpleImputer
```

[3]:

```
train = pd.read_csv(r"G:\DataSet\House Price Prediction\train.csv")
test = pd.read_csv(r"G:\DataSet\House Price Prediction\test.csv")
print("shape of train df = ",train.shape)
```

```python
print("shape of test df = ",test.shape)
```

```
shape of train df =  (1460, 81)
shape of test df =  (1459, 80)
```

[4]:
```python
train.head()
```

[4]:
```
   Id  MSSubClass MSZoning  LotFrontage  LotArea Street Alley LotShape  \
0   1          60       RL         65.0     8450   Pave   NaN      Reg
1   2          20       RL         80.0     9600   Pave   NaN      Reg
2   3          60       RL         68.0    11250   Pave   NaN      IR1
3   4          70       RL         60.0     9550   Pave   NaN      IR1
4   5          60       RL         84.0    14260   Pave   NaN      IR1

  LandContour Utilities  … PoolArea PoolQC Fence MiscFeature MiscVal MoSold  \
0         Lvl    AllPub  …        0    NaN   NaN         NaN       0      2
1         Lvl    AllPub  …        0    NaN   NaN         NaN       0      5
2         Lvl    AllPub  …        0    NaN   NaN         NaN       0      9
3         Lvl    AllPub  …        0    NaN   NaN         NaN       0      2
4         Lvl    AllPub  …        0    NaN   NaN         NaN       0     12

  YrSold  SaleType  SaleCondition  SalePrice
0   2008        WD         Normal     208500
1   2007        WD         Normal     181500
2   2008        WD         Normal     223500
3   2006        WD        Abnorml     140000
4   2008        WD         Normal     250000

[5 rows x 81 columns]
```

[5]:
```python
X_train=train.drop(columns="SalePrice")
y_train=train["SalePrice"]
print("shape of X_train df = ",X_train.shape)
print("shape of y_train df = ",y_train.shape)
```

```
shape of X_train df =  (1460, 80)
shape of y_train df =  (1460,)
```

2 Numerical Missing Value Imputation

[6]:
```python
num_vars=X_train.select_dtypes(include=["int64","float64"]).columns
```

[7]:
```python
num_vars
```

[7]:
```
Index(['Id', 'MSSubClass', 'LotFrontage', 'LotArea', 'OverallQual',
       'OverallCond', 'YearBuilt', 'YearRemodAdd', 'MasVnrArea', 'BsmtFinSF1',
       'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF',
        'LowQualFinSF', 'GrLivArea', 'BsmtFullBath', 'BsmtHalfBath', 'FullBath',
        'HalfBath', 'BedroomAbvGr', 'KitchenAbvGr', 'TotRmsAbvGrd',
        'Fireplaces', 'GarageYrBlt', 'GarageCars', 'GarageArea', 'WoodDeckSF',
        'OpenPorchSF', 'EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea',
        'MiscVal', 'MoSold', 'YrSold'],
       dtype='object')
```

[9]:

```
X_train[num_vars].isnull().sum()
```

[9]:

```
Id                 0
MSSubClass         0
LotFrontage      259
LotArea            0
OverallQual        0
OverallCond        0
YearBuilt          0
YearRemodAdd       0
MasVnrArea         8
BsmtFinSF1         0
BsmtFinSF2         0
BsmtUnfSF          0
TotalBsmtSF        0
1stFlrSF           0
2ndFlrSF           0
LowQualFinSF       0
GrLivArea          0
BsmtFullBath       0
BsmtHalfBath       0
FullBath           0
HalfBath           0
BedroomAbvGr       0
KitchenAbvGr       0
TotRmsAbvGrd       0
Fireplaces         0
GarageYrBlt       81
GarageCars         0
GarageArea         0
WoodDeckSF         0
OpenPorchSF        0
EnclosedPorch      0
3SsnPorch          0
ScreenPorch        0
PoolArea           0
MiscVal            0
MoSold             0
YrSold             0
dtype: int64
```

[10]:

```
imputer_mean = SimpleImputer(strategy='mean')
#imputer_mean = SimpleImputer(strategy='constant', fill_value=99)
```

[11]:

```
imputer_mean.fit(X_train[num_vars])
```

[11]:

```
SimpleImputer(add_indicator=False, copy=True, fill_value=None,
              missing_values=nan, strategy='mean', verbose=0)
```

[12]:

```
imputer_mean.statistics_
```

[12]:

```
array([7.30500000e+02, 5.68972603e+01, 7.00499584e+01, 1.05168281e+04,
       6.09931507e+00, 5.57534247e+00, 1.97126781e+03, 1.98486575e+03,
       1.03685262e+02, 4.43639726e+02, 4.65493151e+01, 5.67240411e+02,
       1.05742945e+03, 1.16262671e+03, 3.46992466e+02, 5.84452055e+00,
       1.51546370e+03, 4.25342466e-01, 5.75342466e-02, 1.56506849e+00,
       3.82876712e-01, 2.86643836e+00, 1.04657534e+00, 6.51780822e+00,
       6.13013699e-01, 1.97850616e+03, 1.76712329e+00, 4.72980137e+02,
       9.42445205e+01, 4.66602740e+01, 2.19541096e+01, 3.40958904e+00,
       1.50609589e+01, 2.75890411e+00, 4.34890411e+01, 6.32191781e+00,
       2.00781575e+03])
```

[13]:

```
imputer_mean.transform(X_train[num_vars])
```

[13]:

```
array([[1.000e+00, 6.000e+01, 6.500e+01, …, 0.000e+00, 2.000e+00,
        2.008e+03],
       [2.000e+00, 2.000e+01, 8.000e+01, …, 0.000e+00, 5.000e+00,
        2.007e+03],
       [3.000e+00, 6.000e+01, 6.800e+01, …, 0.000e+00, 9.000e+00,
        2.008e+03],
       …,
       [1.458e+03, 7.000e+01, 6.600e+01, …, 2.500e+03, 5.000e+00,
        2.010e+03],
       [1.459e+03, 2.000e+01, 6.800e+01, …, 0.000e+00, 4.000e+00,
        2.010e+03],
       [1.460e+03, 2.000e+01, 7.500e+01, …, 0.000e+00, 6.000e+00,
        2.008e+03]])
```

[14]:

```
X_train[num_vars] = imputer_mean.transform(X_train[num_vars])
test[num_vars] = imputer_mean.transform(test[num_vars])
```

[16]:

```
X_train[num_vars].isnull().sum()
```

[16]:

```
Id                0
MSSubClass        0
LotFrontage       0
LotArea           0
OverallQual       0
```

```
OverallCond      0
YearBuilt        0
YearRemodAdd     0
MasVnrArea       0
BsmtFinSF1       0
BsmtFinSF2       0
BsmtUnfSF        0
TotalBsmtSF      0
1stFlrSF         0
2ndFlrSF         0
LowQualFinSF     0
GrLivArea        0
BsmtFullBath     0
BsmtHalfBath     0
FullBath         0
HalfBath         0
BedroomAbvGr     0
KitchenAbvGr     0
TotRmsAbvGrd     0
Fireplaces       0
GarageYrBlt      0
GarageCars       0
GarageArea       0
WoodDeckSF       0
OpenPorchSF      0
EnclosedPorch    0
3SsnPorch        0
ScreenPorch      0
PoolArea         0
MiscVal          0
MoSold           0
YrSold           0
dtype: int64
```

```
[17]: test[num_vars].isnull().sum()
```

```
[17]: Id               0
      MSSubClass       0
      LotFrontage      0
      LotArea          0
      OverallQual      0
      OverallCond      0
      YearBuilt        0
      YearRemodAdd     0
      MasVnrArea       0
      BsmtFinSF1       0
      BsmtFinSF2       0
```

```
BsmtUnfSF        0
TotalBsmtSF      0
1stFlrSF         0
2ndFlrSF         0
LowQualFinSF     0
GrLivArea        0
BsmtFullBath     0
BsmtHalfBath     0
FullBath         0
HalfBath         0
BedroomAbvGr     0
KitchenAbvGr     0
TotRmsAbvGrd     0
Fireplaces       0
GarageYrBlt      0
GarageCars       0
GarageArea       0
WoodDeckSF       0
OpenPorchSF      0
EnclosedPorch    0
3SsnPorch        0
ScreenPorch      0
PoolArea         0
MiscVal          0
MoSold           0
YrSold           0
dtype: int64
```

3 Categorical Missing Value Imputation

[18]:

```
cat_vars=X_train.select_dtypes(include=["O"]).columns
cat_vars
```

[18]:

```
Index(['MSZoning', 'Street', 'Alley', 'LotShape', 'LandContour', 'Utilities',
       'LotConfig', 'LandSlope', 'Neighborhood', 'Condition1', 'Condition2',
       'BldgType', 'HouseStyle', 'RoofStyle', 'RoofMatl', 'Exterior1st',
       'Exterior2nd', 'MasVnrType', 'ExterQual', 'ExterCond', 'Foundation',
       'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2',
       'Heating', 'HeatingQC', 'CentralAir', 'Electrical', 'KitchenQual',
       'Functional', 'FireplaceQu', 'GarageType', 'GarageFinish', 'GarageQual',
       'GarageCond', 'PavedDrive', 'PoolQC', 'Fence', 'MiscFeature',
       'SaleType', 'SaleCondition'],
      dtype='object')
```

[19]:

```
X_train[cat_vars].isnull().sum()
```

[19]:

```
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
```

[20]:

```
imputer_mode = SimpleImputer(strategy='most_frequent')
#imputer_mean = SimpleImputer(strategy='constant', fill_value=99)
```

```
       imputer_mode
```

```
[20]: SimpleImputer(add_indicator=False, copy=True, fill_value=None,
                    missing_values=nan, strategy='most_frequent', verbose=0)
```

```
[21]: imputer_mode.fit(X_train[cat_vars])
```

```
[21]: SimpleImputer(add_indicator=False, copy=True, fill_value=None,
                    missing_values=nan, strategy='most_frequent', verbose=0)
```

```
[22]: imputer_mode.statistics_
```

```
[22]: array(['RL', 'Pave', 'Grvl', 'Reg', 'Lvl', 'AllPub', 'Inside', 'Gtl',
             'NAmes', 'Norm', 'Norm', '1Fam', '1Story', 'Gable', 'CompShg',
             'VinylSd', 'VinylSd', 'None', 'TA', 'TA', 'PConc', 'TA', 'TA',
             'No', 'Unf', 'Unf', 'GasA', 'Ex', 'Y', 'SBrkr', 'TA', 'Typ', 'Gd',
             'Attchd', 'Unf', 'TA', 'TA', 'Y', 'Gd', 'MnPrv', 'Shed', 'WD',
             'Normal'], dtype=object)
```

```
[23]: X_train[cat_vars] = imputer_mode.transform(X_train[cat_vars])
      test[cat_vars] = imputer_mode.transform(test[cat_vars])
```

```
[24]: X_train[cat_vars].isnull().sum()
```

```
[24]: MSZoning         0
      Street           0
      Alley            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
FireplaceQu      0
GarageType       0
GarageFinish     0
GarageQual       0
GarageCond       0
PavedDrive       0
PoolQC           0
Fence            0
MiscFeature      0
SaleType         0
SaleCondition    0
dtype: int64
```

[25]:
```
test[cat_vars].isnull().sum()
```

[25]:
```
MSZoning         0
Street           0
Alley            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
FireplaceQu      0
GarageType       0
GarageFinish     0
GarageQual       0
GarageCond       0
PavedDrive       0
PoolQC           0
Fence            0
MiscFeature      0
SaleType         0
SaleCondition    0
dtype: int64
```

[26]:
```
X_train.isnull().sum().sum()
```

[26]:
```
0
```

[27]:
```
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…-:)
```