Missind values imputation using sklean

different stategy fro differnt variables (Numerical & categotical)

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.impute import SimpleImputer
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
```

In [2]:

```
train=pd.read_csv("train.csv")
test=pd.read_csv("test.csv")

print('train dataset shape :-',train.shape)
print('test dataset shape :-',test.shape)
```

```
train dataset shape :- (1460, 81)
test dataset shape :- (1459, 80)
```

In [3]:
```
train.head()
```

Out[3]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	Mis
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	

5 rows × 81 columns

In [4]:
```
x_train=train.drop(columns='SalePrice')
y_train=train['SalePrice']
x_test=test.copy()

print('train dataset shape :-',x_train.shape)
print('test dataset shape :-',y_train.shape)
print('X_test dataset shape :-',x_test.shape)
```

```
train dataset shape :- (1460, 80)
test dataset shape :- (1460,)
X_test dataset shape :- (1459, 80)
```

Missing values inputation

In [5]:
```
isnull_sum=x_train.isnull().sum()
isnull_sum
```

Out[5]:
```
Id                 0
MSSubClass         0
MSZoning           0
LotFrontage      259
LotArea            0
                ... 
MiscVal            0
MoSold             0
YrSold             0
SaleType           0
SaleCondition      0
Length: 80, dtype: int64
```

In [6]:
```
num_vars=x_train.select_dtypes(include=["int64","float64"]).columns
num_vars_miss=[var for var in num_vars if isnull_sum[var]>0]
```

In [7]:
```
num_vars_miss
```

Out[7]:
```
['LotFrontage', 'MasVnrArea', 'GarageYrBlt']
```

In [8]:

```
cat_vars=x_train.select_dtypes(include=["O"]).columns
cat_vars_miss=[var for var in cat_vars if isnull_sum[var]>0]
cat_vars_miss
```

Out[8]:

```
['Alley',
 'MasVnrType',
 'BsmtQual',
 'BsmtCond',
 'BsmtExposure',
 'BsmtFinType1',
 'BsmtFinType2',
 'Electrical',
 'FireplaceQu',
 'GarageType',
 'GarageFinish',
 'GarageQual',
 'GarageCond',
 'PoolQC',
 'Fence',
 'MiscFeature']
```

In [9]:

```
num_vars_mean=["LotFrontage"]

num_vars_median=['MasVnrArea', 'GarageYrBlt']

cat_vars_mode=['Alley',
 'MasVnrType',
 'BsmtQual',
 'BsmtCond',
 'BsmtExposure',
 'BsmtFinType1',
 'BsmtFinType2',
 'Electrical',
 'FireplaceQu']

cat_vars_missing=['GarageType',
 'GarageFinish',
 'GarageQual',
 'GarageCond',
 'PoolQC',
 'Fence',
 'MiscFeature']
```

In [10]:

```
num_vars_mean_imputer=Pipeline(steps=[("imputer", SimpleImputer(strategy="mean"))])
num_vars_median_imputer=Pipeline(steps=[("imputer", SimpleImputer(strategy="median"))])
cat_vars_mode_imputer=Pipeline(steps=[("imputer", SimpleImputer(strategy="most_frequent"))])
cat_vars_missing_imputer=Pipeline(steps=[("imputer", SimpleImputer(strategy="constant",
                                                                   fill_value="missing"))])
```

In [27]:
```
preprocessor=ColumnTransformer(transformers=
                               [("mean_imputer",num_vars_mean_imputer,num_vars_mean),

                                ("median_imputer",num_vars_median_imputer,num_vars_median),

                                ("mode_imputer",cat_vars_mode_imputer,cat_vars_mode) ,

                                ("missing_imputer",cat_vars_missing_imputer,cat_vars_missing)

                                ])
```

In [12]:
```
preprocessor.fit(x_train)
```

Out[12]:
```
ColumnTransformer(transformers=[('mean_imputer',
                                 Pipeline(steps=[('imputer', SimpleImputer())]),
                                 ['LotFrontage']),
                                ('median_imputer',
                                 Pipeline(steps=[('imputer',
                                                  SimpleImputer(strategy='median'))]),
                                 ['MasVnrArea', 'GarageYrBlt']),
                                ('mode_imputer',
                                 Pipeline(steps=[('imputer',
                                                  SimpleImputer(strategy='most_frequent'))]),
                                 ['Alley', 'MasVnrType', 'BsmtQual', 'BsmtCond',
                                  'BsmtExposure', 'BsmtFinType1',
                                  'BsmtFinType2', 'Electrical',
                                  'FireplaceQu']),
                                ('missing_imputer',
                                 Pipeline(steps=[('imputer',
                                                  SimpleImputer(fill_value='missing',
                                                                strategy='constant'))]),
                                 ['GarageType', 'GarageFinish', 'GarageQual',
                                  'GarageCond', 'PoolQC', 'Fence',
                                  'MiscFeature'])])
```

In [13]:
```
preprocessor.transform
```

Out[13]:
```
<bound method ColumnTransformer.transform of ColumnTransformer(transformers=[('mean_imputer',
                                 Pipeline(steps=[('imputer', SimpleImputer())]),
                                 ['LotFrontage']),
                                ('median_imputer',
                                 Pipeline(steps=[('imputer',
                                                  SimpleImputer(strategy='median'))]),
                                 ['MasVnrArea', 'GarageYrBlt']),
                                ('mode_imputer',
                                 Pipeline(steps=[('imputer',
                                                  SimpleImputer(strategy='most_frequent'))]),
                                 ['Alley', 'MasVnrType', 'BsmtQual', 'BsmtCond',
                                  'BsmtExposure', 'BsmtFinType1',
                                  'BsmtFinType2', 'Electrical',
                                  'FireplaceQu']),
                                ('missing_imputer',
                                 Pipeline(steps=[('imputer',
                                                  SimpleImputer(fill_value='missing',
                                                                strategy='constant'))]),
                                 ['GarageType', 'GarageFinish', 'GarageQual',
                                  'GarageCond', 'PoolQC', 'Fence',
                                  'MiscFeature'])])>
```

In [14]:
```
preprocessor.named_transformers_["mean_imputer"].named_steps["imputer"].statistics_
```

Out[14]:
```
array([70.04995837])
```

In [15]:
```
train["LotFrontage"].mean()
```

Out[15]:
```
70.04995836802665
```

In [ ]:
```

```

```
In [16]: preprocessor.named_transformers_["mode_imputer"].named_steps["imputer"].statistics_
```

```
Out[16]: array(['Grvl', 'None', 'TA', 'TA', 'No', 'Unf', 'Unf', 'SBrkr', 'Gd'],
               dtype=object)
```

```
In [17]: x_train_clean=preprocessor.transform(x_train)
         x_test_clean=preprocessor.transform(x_test)
```

```
In [18]: x_train_clean
```

```
Out[18]: array([[65.0, 196.0, 2003.0, ..., 'missing', 'missing', 'missing'],
                [80.0, 0.0, 1976.0, ..., 'missing', 'missing', 'missing'],
                [68.0, 162.0, 2001.0, ..., 'missing', 'missing', 'missing'],
                ...,
                [66.0, 0.0, 1941.0, ..., 'missing', 'GdPrv', 'Shed'],
                [68.0, 0.0, 1950.0, ..., 'missing', 'missing', 'missing'],
                [75.0, 0.0, 1965.0, ..., 'missing', 'missing', 'missing']],
               dtype=object)
```

In [19]:
```
preprocessor.transformers_
```

Out[19]:
```
[('mean_imputer',
  Pipeline(steps=[('imputer', SimpleImputer())]),
  ['LotFrontage']),
 ('median_imputer',
  Pipeline(steps=[('imputer', SimpleImputer(strategy='median'))]),
  ['MasVnrArea', 'GarageYrBlt']),
 ('mode_imputer',
  Pipeline(steps=[('imputer', SimpleImputer(strategy='most_frequent'))]),
  ['Alley',
   'MasVnrType',
   'BsmtQual',
   'BsmtCond',
   'BsmtExposure',
   'BsmtFinType1',
   'BsmtFinType2',
   'Electrical',
   'FireplaceQu']),
 ('missing_imputer',
  Pipeline(steps=[('imputer',
                   SimpleImputer(fill_value='missing', strategy='constant'))]),
  ['GarageType',
   'GarageFinish',
   'GarageQual',
   'GarageCond',
   'PoolQC',
   'Fence',
   'MiscFeature']),
 ('remainder',
  'drop',
  [0,
   1,
   2,
   4,
   5,
   7,
   8,
 9,
 10,
 11,
 12,
 13,
 14,
 15,
 16,
 17,
 18,
 19,
 20,
 21,
 22,
 23,
 24,
 27,
 28,
 29,
 34,
 36,
 37,
 38,
 39,
 40,
 41,
 43,
 44,
 45,
 46,
 47,
 48,
 49,
 50,
 51,
 52,
 53,
 54,
 55,
  56,
  61,
  62,
  65,
  66,
  67,
  68,
  69,
  70,
  71,
  75,
  76,
  77,
  78,
  79]))]
```

```
('remainder',      #  please keep in mind it's remainder................  den rakhana
 'drop',           #IMP he  hame columns transformer me remainder ko--- ['passthrough'] vlues den
          #    padthi he oo by deflout drop leta he {'drop', 'passthrough'}
 [0,
  1,
  2,
  4,
  5,
  7,
  8,
  9,
  10,
  11,
  12,
  13,
  14,
  15,
  16,
  17,
  18,
  19,
  20,
 21,
 22,
 23,
 24,
 27,
 28,
 29,
 34,
 36,
 37,
 38,
 39,
 40,
 41,
 43,
 44,
 45,
 46,
 47,
 48,
 49,
 50,
 51,
 52,
 53,
 54,
 55,
 56,
 61,
 62,
 65,
 66,
 67,
 68,
 69,
 70,
 71,
 75,
 76,
    77,
    78,
    79])]
```

In [30]:

```
x_train_clean_miss_var=pd.DataFrame(x_train_clean,columns=num_vars_mean+
                                                num_vars_median+cat_vars_mode+cat_vars_missing)
```

In [31]:

```
x_train_clean_miss_var
```

Out[31]:

	LotFrontage	MasVnrArea	GarageYrBlt	Alley	MasVnrType	BsmtQual	BsmtCond	BsmtExposure	BsmtFinType1	BsmtFinType2
0	65	196	2003	Grvl	BrkFace	Gd	TA	No	GLQ	Unf
1	80	0	1976	Grvl	None	Gd	TA	Gd	ALQ	Unf
2	68	162	2001	Grvl	BrkFace	Gd	TA	Mn	GLQ	Unf
3	60	0	1998	Grvl	None	TA	Gd	No	ALQ	Unf
4	84	350	2000	Grvl	BrkFace	Gd	TA	Av	GLQ	Unf
...	...	...	...	...	...	...	...	...	...	...
1455	62	0	1999	Grvl	None	Gd	TA	No	Unf	Unf
1456	85	119	1978	Grvl	Stone	Gd	TA	No	ALQ	Rec
1457	66	0	1941	Grvl	None	TA	Gd	No	GLQ	Unf
1458	68	0	1950	Grvl	None	TA	TA	Mn	GLQ	Rec
1459	75	0	1965	Grvl	None	TA	TA	No	BLQ	LwQ

1460 rows × 19 columns

In [22]:

```
x_train_clean_miss_var.isnull().sum().sum()
```

Out[22]:

```
0
```

In [24]:
```
train["Alley"].value_counts()
```

Out[24]:
```
Grvl    50
Pave    41
Name: Alley, dtype: int64
```

In [25]:
```
x_train_clean_miss_var["Alley"].value_counts()
```

Out[25]:
```
Grvl    1419
Pave      41
Name: Alley, dtype: int64
```

In [26]:
```
x_train_clean_miss_var["MiscFeature"].value_counts()
```

Out[26]:
```
missing    1406
Shed         49
Gar2          2
Othr          2
TenC          1
Name: MiscFeature, dtype: int64
```

In [ ]:
```

```