(It's all about data Clean Up)



https://www.linkedin.com/in/nibedita-sarkar/



## (It's all about data Clean Up)

Q) What is missing Value?

The values or data that is not stored (or not present) for some variable/s in the given dataset.

- Q) What are the methods to handle missing values?
- 1) Ignore missing values row/ Delete row
- 2) Fill missing value manually
- 3) Global Constant
- 4) Measure of central tendency (Mean, Median, Mode)
- 5) Measure of central tendency for each class
- 6) Most probable value (ML Algorithms)
- 1) Ignore missing values row/ Delete row:

In this approach, you simply ignore/ delete the rows or observations that contain missing values. This method is suitable when the missing values are minimal and do not significantly impact the analysis. However, if a large proportion of the data is missing, this method may result in the loss of valuable information.

If the dataset contains missing values equal to or less than 5%, we can consider deleting the corresponding rows.

If the columns contain more than 20% or 30% missing values, we can consider deleting those columns.



## (It's all about data Clean Up)

## 2) Fill missing value manually:

For small datasets, you can manually examine the missing values and fill them based on your domain knowledge or by consulting subject matter experts. While this method ensures that the filled values are accurate, it can be time-consuming and subjective.

## 3) Global Constant:

In this approach, missing values are replaced with a global constant value. For example, you can replace missing numerical values with zero or missing categorical values with a specific category label. However, this method may introduce bias or distort the data if the missing values are not truly missing at random.

## 4) Measure of central tendency (Mean, Median, Mode):

Missing values in numerical variables can be replaced by a measure of central tendency such as the mean, median, or mode. The mean is commonly used if the variable follows a normal distribution, while the median is preferred if the variable has outliers. The mode is suitable for categorical variables.

## 5) Measure of central tendency for each class:

If the data has categorical variables and missing values are present within specific classes or groups, you can replace the missing values with the measure of central tendency (mean, median, or mode) calculated separately for each class. This approach helps preserve the within-class characteristics and reduces bias.



(It's all about data Clean Up)

## 6) Most probable value (ML Algorithms):

Machine learning algorithms can be used to predict the missing values based on the available data. You can train a model on the non-missing values and then use it to predict the missing values. This method utilizes the relationships and patterns in the data to estimate the missing values. However, it assumes that the missing values are predictable based on other variables.

# Note:

Among the six methods mentioned, I have written a process to handle missing values for the first five methods. The sixth method involves handling missing values using a machine learning algorithm, which I will share at a later time.

Please consider following @Nibedita Sarkar on LinkedIn for more updates and information. You can find the link to her LinkedIn profile here:

[https://www.linkedin.com/in/nibedita-sarkar/]



### tation-by-deleting-row-and-columns

June 11, 2023

### 1 Data Cleaning:

#### 1.1 Missing value imputation by Deleting Rows and Columns:

Importing necessary libraries

```
[1]: import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     import warnings
     warnings.filterwarnings('ignore')
    data (Original):
[3]: data = pd.read_csv("train.csv")
                                         # Load dataset
     data.head(5) # Checking first 5 rows from the DataFrame.
[5]:
            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
                                                             Pave
                                                                    NaN
                                                    11250
                                                                              IR1
     3
         4
                     70
                               RL
                                           60.0
                                                     9550
                                                             Pave
                                                                    NaN
                                                                              IR1
         5
     4
                     60
                               RL
                                           84.0
                                                    14260
                                                                              IR1
                                                             Pave
                                                                    NaN
       LandContour Utilities
                                ... PoolArea PoolQC Fence MiscFeature MiscVal MoSold
     0
                Lvl
                        AllPub
                                          0
                                               NaN
                                                      NaN
                                                                   NaN
                                                                              0
                                                                                      2
     1
                Lvl
                       AllPub
                                          0
                                               NaN
                                                                   NaN
                                                                              0
                                                                                      5
                                                      NaN
     2
                Lvl
                       AllPub
                                          0
                                               NaN
                                                      NaN
                                                                   NaN
                                                                              0
                                                                                      9
                                                                                      2
     3
                Lvl
                       AllPub
                                          0
                                               NaN
                                                      NaN
                                                                   NaN
                                                                              0
     4
                Lvl
                       AllPub
                                               NaN
                                                                              0
                                                                                     12
                                                      NaN
                                                                   NaN
       YrSold
                SaleType
                           SaleCondition
                                           SalePrice
         2008
     0
                      WD
                                  Normal
                                              208500
     1
         2007
                      WD
                                  Normal
                                              181500
     2
         2008
                      WD
                                  Normal
                                              223500
     3
         2006
                      WD
                                 Abnorml
                                              140000
         2008
                      WD
                                  Normal
                                              250000
```

[5 rows x 81 columns]

```
[4]: data.shape # Checking the shape of the dataset, dataset contains 1460 rows and $\to 81$ columns
```

#### [4]: (1460, 81)

When we check the shape of the data using the command "data.shape", we can observe that the DataFrame consists of 1460 rows and 81 columns. However, when we examine the first five rows of the DataFrame using the command "data.head(5)", it doesn't display all 81 columns. As a result, we utilize the following code, specifically "pd.set\_option()", to address this issue.

```
[6]: pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
```

pd.set\_option('display.max\_columns', None): This line sets the maximum number of columns to be displayed in the output to None, which means there is no limit. As a result, all columns in a DataFrame will be shown when you print or display it.

pd.set\_option('display.max\_rows', None): This line sets the maximum number of rows to be displayed in the output to None, removing any limit. As a result, all rows in a DataFrame will be shown when you print or display it.

```
[7]: data.head(2) # The code `data.head(2)` displays the first two rows of the dataset, showing all 81 columns.
```

```
[7]:
        Ιd
            MSSubClass MSZoning
                                  LotFrontage LotArea Street Alley LotShape
     0
         1
                     60
                               RL
                                           65.0
                                                    8450
                                                            Pave
                                                                   NaN
                                                                             Reg
     1
         2
                     20
                               R.T.
                                           80.0
                                                    9600
                                                            Pave
                                                                   NaN
                                                                             Reg
       LandContour Utilities LotConfig LandSlope Neighborhood Condition1
     0
                Lvl
                       AllPub
                                  Inside
                                                Gtl
                                                          CollgCr
                                                                         Norm
     1
                Lvl
                       AllPub
                                     FR2
                                                Gtl
                                                          Veenker
                                                                        Feedr
                                                                     YearBuilt \
       Condition2 BldgType HouseStyle
                                         OverallQual
                                                       OverallCond
     0
             Norm
                       1Fam
                                 2Story
                                                    7
                                                                  5
                                                                           2003
     1
             Norm
                       1Fam
                                 1Story
                                                    6
                                                                  8
                                                                           1976
        YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType
     0
                 2003
                           Gable
                                  CompShg
                                               VinylSd
                                                            VinylSd
                                                                        BrkFace
     1
                 1976
                           Gable
                                  CompShg
                                               MetalSd
                                                            MetalSd
                                                                           None
        MasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
     0
              196.0
                           Gd
                                      TΑ
                                               PConc
                                                            Gd
                                                                     TA
                                                                                    No
     1
                0.0
                           TΑ
                                      TA
                                              CBlock
                                                            Gd
                                                                     TA
                                                                                    Gd
```

		%LQ 706 %LQ 978		0 0	150 284	856 1262
	Heating He O GasA 1 GasA	eatingQC Centra Ex Ex	lAir Electrica Y SBrk Y SBrk	r 856	2ndFlrSF Low 854 0	QualFinSF \ 0 0
	GrLivArea 0 1710 1 1262	) 1		2	HalfBath Bedr 1 0	roomAbvGr \ 3 3
	KitchenAb O 1		d	d Functional 8 Typ 6 Typ		FireplaceQu \ NaN TA
	GarageType  O Attchd  1 Attchd		GarageFinish RFn RFn	GarageCars 2 2	GarageArea Gar 548 460	rageQual \ TA TA
	GarageCond  TA		oodDeckSF Ope 0 298	nPorchSF En 61 0	aclosedPorch 3 0 0	SSsnPorch \ 0 0
	ScreenPor 0 1	cch PoolArea P 0 0 0 0	oolQC Fence Mi NaN NaN NaN NaN	scFeature M NaN NaN	MiscVal MoSolo 0 2 0 5	2 2008
	SaleType S 0 WD 1 WD	SaleCondition Normal Normal	SalePrice 208500 181500			
[8]:	$data.tail(2)$ $\Rightarrow dataset.$	# The code `c		lisplays the	last two rows	of theu
	Id 1458 1459 1459 1460	MSSubClass MSZ 20 20	RL	tage LotAre 68.0 971 75.0 993	.7 Pave NaN	I Reg
	LandCor 1458 1459	tour Utilities Lvl AllPub Lvl AllPub	Inside	Gtl	NAmes	zion1 \ Norm Norm
	1458 N	on2 BldgType H Jorm 1Fam Jorm 1Fam	ouseStyle Ove 1Story 1Story	rallQual Ov 5 5	verallCond Yea 6 6	arBuilt \ 1950 1965
	YearRe	emodAdd RoofSty	le RoofMatl Ex	terior1st Ex	tterior2nd MasV	nrType \

```
1996 Hip CompShg MetalSd MetalSd
1965 Gable CompShg HdBoard HdBoard
   1458
                                                      None
   1459
                                                      None
        MasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond \
   1458
             0.0
                 TA TA
                                  CBlock TA
             0.0
   1459
                     Gd
                             ТΑ
                                   CBlock
                                            ТΑ
                                                   ТΑ
       BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2 \
                   GLQ
   1458
           Mn
                            49
                                           Rec
                                                    1029
   1459
             No
                        BLQ
                              830
                                           LwQ
                                                    290
        BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical \
                            GasA
                                     Gd Y
   1458
                     1078
   1459 136
                   1256
                            GasA
                                     Gd
                                             Y
                                                   SBrkr
        1stFlrSF 2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath \
   1458
           1078
                 0 0
                                    1078
                    0
                               0
   1459
           1256
                                    1256
       FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual \
                            2
            1
                    1
   1459
                              3
                                         1
                                                   TΑ
        TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt \
                       Typ 0 NaN Attchd 1950.0
Typ 0 NaN Attchd 1965.0
   1458
          5
               6
   1459
                       Typ
    GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive \
                    1
                                240
   1458
             Unf
                                         TA
                                                  TA
                        1
                                         TA
                                                          Υ
   1459
             Fin
                                276
                                                  TA
        WoodDeckSF OpenPorchSF EnclosedPorch 3SsnPorch ScreenPorch \
                            112 0
             366
   1458
                  0
                        68
             736
                                           0
   1459
                                    0
        PoolArea PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType \
                     NaN NaN 0 4
   1458
         0
                 NaN
                                                 2010 WD
            0 NaN
                                           6
   1459
                     NaN
                               {\tt NaN}
                                      0
                                                 2008
                                                          WD
       SaleCondition SalePrice
   1458
           Normal
                   142125
   1459
           Normal
                    147500
[9]: data.info()
```

The `data.info()` function provides a concise summary of the dataset, including  $\cup$  information about the number of rows, columns, data types, memory usage and missing values.

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 81 columns):

#	Column	Non-Null Count	Dtype
0	Id	1460 non-null	int64
1	MSSubClass	1460 non-null	int64
2	MSZoning	1460 non-null	object
3	LotFrontage	1201 non-null	float64
4	LotArea	1460 non-null	int64
5	Street	1460 non-null	object
6	Alley	91 non-null	object
7	LotShape	1460 non-null	object
8	LandContour	1460 non-null	object
9	Utilities	1460 non-null	object
10	LotConfig	1460 non-null	object
11	LandSlope	1460 non-null	object
12	Neighborhood	1460 non-null	object
13	Condition1	1460 non-null	object
14	Condition2	1460 non-null	object
15	BldgType	1460 non-null	object
16	HouseStyle	1460 non-null	object
17	OverallQual	1460 non-null	int64
18	OverallCond	1460 non-null	int64
19	YearBuilt	1460 non-null	int64
20	${\tt YearRemodAdd}$	1460 non-null	int64
21	RoofStyle	1460 non-null	object
22	RoofMatl	1460 non-null	object
23	Exterior1st	1460 non-null	object
24	Exterior2nd	1460 non-null	object
25	${ t MasVnrType}$	1452 non-null	object
26	MasVnrArea	1452 non-null	float64
27	ExterQual	1460 non-null	object
28	ExterCond	1460 non-null	object
29	Foundation	1460 non-null	object
30	BsmtQual	1423 non-null	object
31	${\tt BsmtCond}$	1423 non-null	object
32	${\tt BsmtExposure}$	1422 non-null	object
33	${\tt BsmtFinType1}$	1423 non-null	object
34	BsmtFinSF1	1460 non-null	int64
35	BsmtFinType2	1422 non-null	object

36	BsmtFinSF2	1460 non-null	int64
37	BsmtUnfSF	1460 non-null	int64
38	TotalBsmtSF	1460 non-null	int64
39	Heating	1460 non-null	object
40	${\tt HeatingQC}$	1460 non-null	object
41	CentralAir	1460 non-null	object
42	Electrical	1459 non-null	object
43	1stFlrSF	1460 non-null	int64
44	2ndFlrSF	1460 non-null	int64
45	${ t LowQualFinSF}$	1460 non-null	int64
46	GrLivArea	1460 non-null	int64
47	BsmtFullBath	1460 non-null	int64
48	BsmtHalfBath	1460 non-null	int64
49	FullBath	1460 non-null	int64
50	HalfBath	1460 non-null	int64
51	BedroomAbvGr	1460 non-null	int64
52	KitchenAbvGr	1460 non-null	int64
53	KitchenQual	1460 non-null	object
54	TotRmsAbvGrd	1460 non-null	int64
55	Functional	1460 non-null	object
56	Fireplaces	1460 non-null	int64
57	FireplaceQu	770 non-null	object
58	GarageType	1379 non-null	object
59	GarageYrBlt	1379 non-null	float64
60	GarageFinish	1379 non-null	
	•		object
61	GarageCars	1460 non-null	int64
62	GarageArea	1460 non-null	int64
63	GarageQual	1379 non-null	object
64	GarageCond	1379 non-null	object
65	PavedDrive	1460 non-null	object
66	WoodDeckSF	1460 non-null	int64
67	OpenPorchSF	1460 non-null	int64
68		1460 non-null	int64
69		1460 non-null	
	ScreenPorch	1460 non-null	int64
	PoolArea	1460 non-null	int64
72	PoolQC	7 non-null	object
73	Fence	281 non-null	object
	MiscFeature	54 non-null	object
75	MiscVal	1460 non-null	int64
76	MoSold	1460 non-null	int64
77	YrSold	1460 non-null	int64
78	SaleType	1460 non-null	object
79	SaleCondition	1460 non-null	object
80	SalePrice	1460 non-null	int64
dtyp	es: float64(3),	int64(35), obj	ect(43)
	ry usage: 924.0	-	
	. •		

[9]: '\nThe `data.info()` function provides a concise summary of the dataset, including information about the number of rows, \ncolumns, data types, memory usage and missing values.\n\n'

#### [10]: data.isnull().sum()

[10]: Id 0 MSSubClass 0 MSZoning 0 LotFrontage 259 LotArea 0 Street 0 Alley 1369 LotShape LandContour 0 Utilities 0 LotConfig 0 LandSlope 0 Neighborhood 0 Condition1 0 Condition2 0 BldgType 0 HouseStyle 0 OverallQual 0 OverallCond 0 YearBuilt 0 YearRemodAdd 0 RoofStyle 0 RoofMatl 0 Exterior1st 0 0 Exterior2nd MasVnrType 8 MasVnrArea 8 ExterQual 0 ExterCond 0 Foundation 0 BsmtQual 37 BsmtCond 37 BsmtExposure 38 BsmtFinType1 37 BsmtFinSF1 0 BsmtFinType2 38 BsmtFinSF2 0 BsmtUnfSF 0

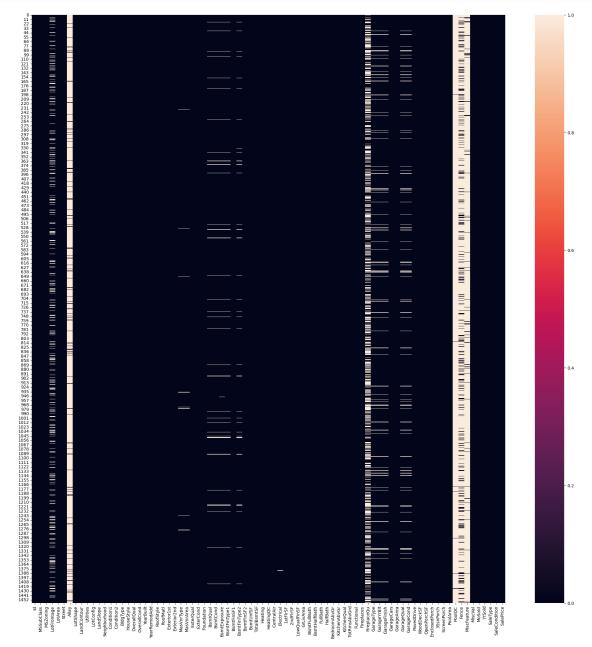
```
TotalBsmtSF
                     0
Heating
                     0
                     0
HeatingQC
                     0
CentralAir
Electrical
                     1
1stFlrSF
                     0
2ndFlrSF
                     0
{\tt LowQualFinSF}
                     0
GrLivArea
                     0
BsmtFullBath
                     0
BsmtHalfBath
                     0
FullBath
                     0
HalfBath
                     0
BedroomAbvGr
                     0
KitchenAbvGr
                     0
KitchenQual
                     0
TotRmsAbvGrd
                     0
                     0
Functional
                     0
Fireplaces
FireplaceQu
                   690
GarageType
                    81
GarageYrBlt
                    81
GarageFinish
                    81
GarageCars
                     0
GarageArea
                     0
GarageQual
                    81
GarageCond
                    81
PavedDrive
                     0
WoodDeckSF
                     0
                     0
OpenPorchSF
EnclosedPorch
                     0
3SsnPorch
                     0
ScreenPorch
                     0
PoolArea
                     0
PoolQC
                  1453
Fence
                  1179
MiscFeature
                  1406
MiscVal
                     0
MoSold
                     0
YrSold
                     0
                     0
SaleType
SaleCondition
                     0
SalePrice
                     0
dtype: int64
```

```
[11]: plt.figure(figsize=(25,25))
sns.heatmap(data.isnull())
```

```
plt.show()

The code `plt.figure(figsize=(25,25))` sets the figure size for the plot.

The next line `sns.heatmap(data.isnull())` creates a heatmap visualization of the missing values in the dataset using the seaborn library.
```



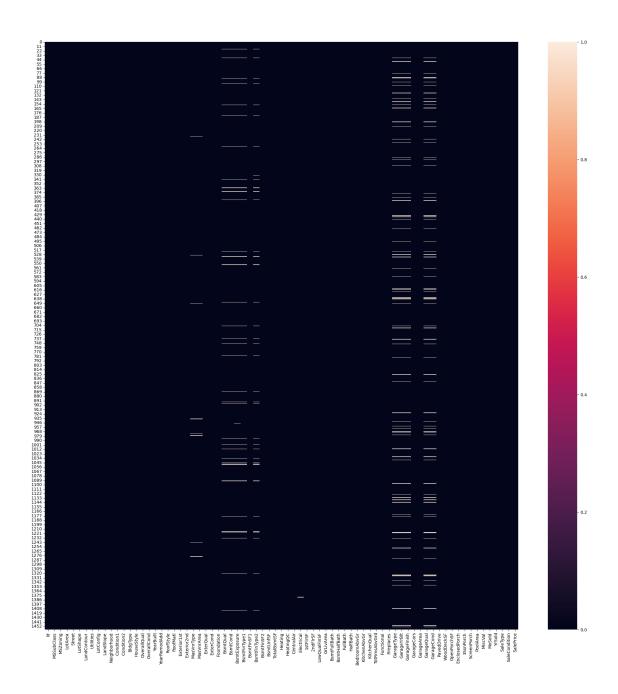
- [11]: '\nThe code `plt.figure(figsize=(25,25))` sets the figure size for the plot. \n\nThe next line `sns.heatmap(data.isnull())` creates a heatmap visualization of the missing values in the dataset \nusing the seaborn library.\n\n'
- [12]: missing\_value\_percent = data.isnull().sum() / data.shape[0] \* 100
  print(missing\_value\_percent)
  - # The code calculates the percentage of missing values in each column of the  $\Box$   $\Box$  dataset and then prints the resulting percentages.

Id	0.000000
MSSubClass	0.000000
MSZoning	0.000000
LotFrontage	17.739726
LotArea	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
OverallQual	0.000000
OverallCond	0.000000
YearBuilt	0.000000
YearRemodAdd	0.000000
RoofStyle	0.000000
RoofMatl	0.000000
Exterior1st	0.000000
Exterior2nd	0.000000
MasVnrType	0.547945
MasVnrArea	0.547945
ExterQual	0.000000
ExterCond	0.000000
Foundation	0.000000
BsmtQual	2.534247
BsmtCond	2.534247
BsmtExposure	2.602740
BsmtFinType1	2.534247
BsmtFinSF1	0.000000
BsmtFinType2	2.602740
<i>J</i> 1	

BsmtFinSF2	0.000000
BsmtUnfSF	0.000000
TotalBsmtSF	0.000000
Heating	0.000000
HeatingQC	0.000000
CentralAir	0.000000
Electrical	0.068493
1stFlrSF	0.000000
2ndFlrSF	0.00000
LowQualFinSF	0.000000
GrLivArea	0.000000
BsmtFullBath	0.00000
BsmtHalfBath	0.00000
FullBath	0.000000
HalfBath	0.000000
BedroomAbvGr	0.000000
KitchenAbvGr	0.000000
KitchenQual	0.000000
TotRmsAbvGrd	0.000000
Functional	0.000000
Fireplaces	0.00000
FireplaceQu	47.260274
GarageType	5.547945
GarageYrBlt	5.547945
GarageFinish	5.547945
GarageCars	0.000000
GarageArea	0.000000
GarageQual	5.547945
GarageCond	5.547945
PavedDrive	0.000000
WoodDeckSF	0.000000
OpenPorchSF	0.000000
EnclosedPorch	0.000000
3SsnPorch	0.000000
ScreenPorch	0.000000
PoolArea	0.000000
PoolQC	99.520548
Fence	80.753425
MiscFeature	96.301370
MiscVal	0.000000
MoSold	0.000000
YrSold	0.000000
SaleType	0.000000
SaleCondition	0.000000
SalePrice	0.000000
d+wno. float64	

dtype: float64

```
[40]: missing_value_column = missing_value_percent[missing_value_percent > 17].keys()
      print(missing_value_column)
      111
      The code is used to identify the columns in a dataset that have missing values \Box
       ⇔exceeding 17%.
      It retrieves the keys (column names) from the `missing_value_percent'
      ⇔dictionary where the corresponding values
      are greater than 17%.
      ,,,
     Index(['LotFrontage', 'Alley', 'FireplaceQu', 'PoolQC', 'Fence',
            'MiscFeature'],
           dtype='object')
[40]: '\nThe code is used to identify the columns in a dataset that have missing
      values exceeding 17%. \nIt retrieves the keys (column names) from the
      `missing_value_percent` dictionary where the corresponding values \nare greater
      than 17\%.\n\n'
     data1 after dropping missing value column > 17\%:
[20]: data1 = data.drop(columns = missing_value_column)
[21]: plt.figure(figsize=(25,25))
      sns.heatmap(data1.isnull())
      plt.show()
      111
      The code is used to get visualization of the missing values after removing \Box
       ⇔columns
      where the missing values exceed 17% of the total data in those columns.
      111
```



[21]: '\nThe code is used to get visualization of the missing values after removing columns \nwhere the missing values exceed 20% of the total data in those columns.\n\n'

Our objective is to eliminate all the white lines from the visualization in order to achieve a cleaner and more organized appearance.

```
[22]: data1.shape # Checking the shape of the DataFrame , DataFrame contains Rows:

→1460 and Columns: 76
```

```
[22]: (1460, 75)

    data2 after dropping missing value rows.

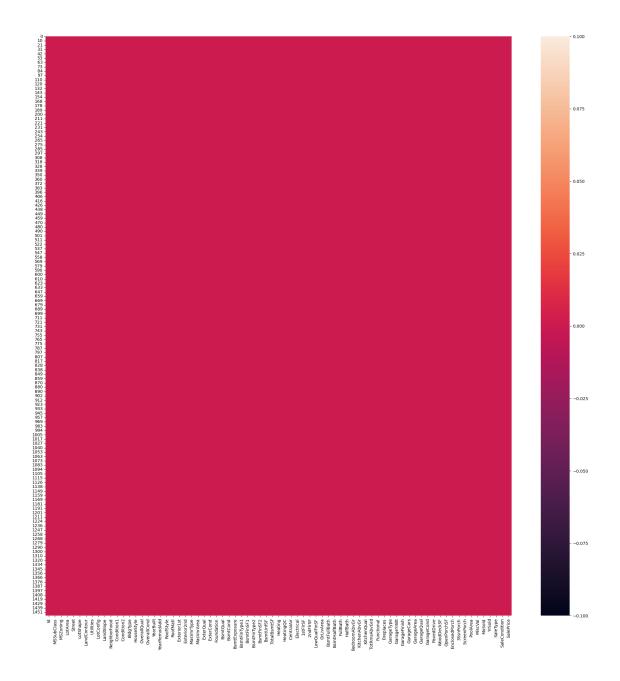
[23]: data2 = data1.dropna() # deleting missing value rows.

[41]: data2.shape

[41]: (1338, 75)

[18]: plt.figure(figsize=(25,25))
    sns.heatmap(df1.isnull())
    plt.show()

# This heatmap creates a visualization of missing values after dropping rows.
```



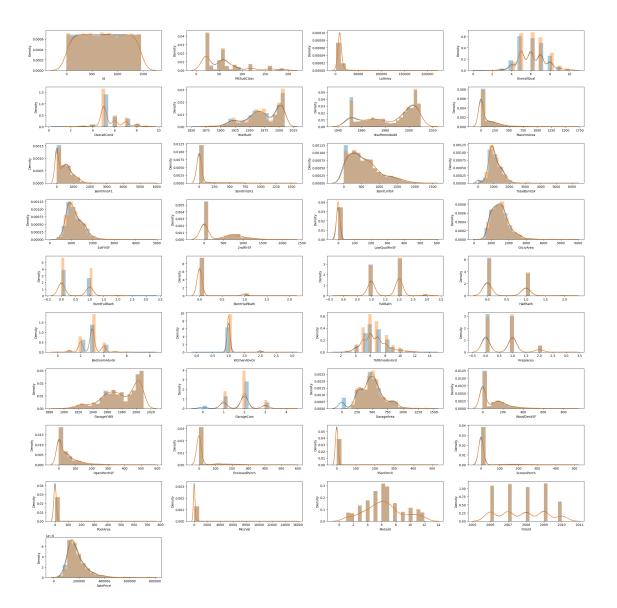
Here, we have achieved a cleaner and more organized appearance.

```
[26]: data2.isnull().sum().sum() # Here, we can see that our DatFrame doesn't contain_ant null values.
```

[26]: 0

Checking the data distribution before and after dealing with missing values for numerical column.

```
[27]: data2.select_dtypes(include=['int64', 'float64']).columns # Cheking the column_
       ⇔names of numerical DataFrame.
[27]: Index(['Id', 'MSSubClass', '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', 'SalePrice'],
            dtype='object')
[28]: num_var = ['Id', 'MSSubClass', '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', 'SalePrice']
[30]: plt.figure(figsize=(25,25))
      # Calculate the number of rows and columns for the subplots
      num_rows = (len(num_var) + 3) // 4 # Adjust the number of columns as needed
      num_cols = 4
      for i, var in enumerate(num_var):
          plt.subplot(num_rows, num_cols, i+1)
          sns.distplot(data[var], bins=20)
          sns.distplot(data2[var], bins=20)
      plt.tight_layout() # Optional: Adjusts the spacing between subplots
      plt.show()
```



We can see there is not much difference between original dataset and cleaned dataset.

# Checking the data distribution before and after dealing with missing values for categorical column.

```
[32]: data2.select_dtypes(include=['object']).columns

# It is used to retrieve the column names of a DataFrame df1 that have object_u data type.
```

```
[36]: pd.concat([data['MSZoning'].value_counts() / data.shape[0] * 100,
data2['MSZoning'].value_counts() / data2.shape[0] * 100], axis=1,
keys=['MSZoning_orginal', 'MSZoning_clean'])
```

```
[36]:
               MSZoning_orginal
                                   MSZoning_clean
                       78.835616
      RL
                                         79.671151
      RM
                       14.931507
                                         14.275037
      FV
                        4.452055
                                         4.633782
      RH
                        1.095890
                                         0.822123
      C (all)
                        0.684932
                                         0.597907
```

The code snippet you provided combines the value counts of the 'MSZoning' column from the original data (data) and the cleaned data (df1) using pd.concat(). It creates a DataFrame that shows the percentage distribution of the values in the 'MSZoning' column for both datasets.

```
[38]: cat_var_dist('BsmtQual')
```

```
[38]: BsmtQual_orginal BsmtQual_clean
TA 44.452055 44.394619
Gd 42.328767 44.245142
Ex 8.287671 8.968610
Fa 2.397260 2.391629
```

The cat\_var\_dist() function you provided calculates the percentage distribution of values in a categorical variable for both the original data (data) and the cleaned data (df1). It returns a DataFrame that combines these distributions side by side.

```
[39]: for var in cat_var:
          result = cat_var_dist(var)
          print(result)
          print('\n')
              MSZoning_orginal MSZoning_clean
     RL
                      78.835616
                                      79.671151
     RM
                      14.931507
                                      14.275037
     F۷
                       4.452055
                                       4.633782
     RH
                                       0.822123
                       1.095890
     C (all)
                       0.684932
                                       0.597907
           Street_orginal Street_clean
                99.589041
                               99.626308
     Pave
                 0.410959
     Grvl
                                0.373692
          LotShape_orginal LotShape_clean
                 63.356164
                                  61.958146
     Reg
     IR1
                 33.150685
                                  34.304933
     IR2
                  2.808219
                                   2.989537
     IR3
                  0.684932
                                   0.747384
          LandContour_orginal LandContour_clean
     Lvl
                    89.794521
                                        90.134529
     Bnk
                      4.315068
                                         3.886398
     HLS
                      3.424658
                                         3.587444
     Low
                      2.465753
                                         2.391629
             Utilities_orginal Utilities_clean
     AllPub
                      99.931507
                                       99.925262
     NoSeWa
                       0.068493
                                        0.074738
                                 LotConfig_clean
              LotConfig_orginal
     Inside
                       72.054795
                                        71.524664
     Corner
                       18.013699
                                        18.236173
     CulDSac
                       6.438356
                                         6.726457
     FR2
                       3.219178
                                         3.213752
                                         0.298954
     FR3
                       0.273973
          LandSlope_orginal LandSlope_clean
```

Mod	4.452055	4.559043
Sev	0.890411	0.896861

	Neighborhood_orginal	Neighborhood_clean
NAmes	15.410959	15.620329
CollgCr	10.273973	10.911809
${\tt OldTown}$	7.739726	7.473842
Edwards	6.849315	5.231689
Somerst	5.890411	6.203288
Gilbert	5.410959	5.754858
NridgHt	5.273973	5.605381
Sawyer	5.068493	5.156951
NWAmes	5.000000	5.455904
SawyerW	4.041096	3.961136
BrkSide	3.972603	3.512706
${\tt Crawfor}$	3.493151	3.736921
Mitchel	3.356164	3.139013
NoRidge	2.808219	3.064275
Timber	2.602740	2.765321
IDOTRR	2.534247	2.167414
${\tt ClearCr}$	1.917808	1.943199
${\tt StoneBr}$	1.712329	1.868460
SWISU	1.712329	1.494768
${\tt MeadowV}$	1.164384	0.896861
Blmngtn	1.164384	1.270553
${\tt BrDale}$	1.095890	1.121076
Veenker	0.753425	0.822123
NPkVill	0.616438	0.672646
Blueste	0.136986	0.149477

	Condition1_orginal	Condition1_clean
Norm	86.301370	86.846039
Feedr	5.547945	4.708520
Artery	3.287671	3.213752
RRAn	1.780822	1.943199
PosN	1.301370	1.420030
RRAe	0.753425	0.747384
PosA	0.547945	0.597907
RRNn	0.342466	0.373692
RRNe	0.136986	0.149477

	Condition2_orginal	Condition2_clean
Norm	98.972603	98.953662
Feedr	0.410959	0.373692
Artery	0.136986	0.149477

RRNn	0.136986	0.149477
PosN	0.136986	0.149477
PosA	0.068493	0.074738
RRAn	0.068493	0.074738
RRAe	0.068493	0.074738
	DidaTama amainal 1	DldaTama alaam
4 T	BldgType_orginal	
1Fam	83.561644	85.052317
TwnhsE	7.808219	8.370703
Duplex	3.561644	2.092676
Twnhs	2.945205	2.840060
2fmCon	2.123288	1.644245
	Hana Charle caminal	II
4.00	HouseStyle_orginal	•
1Story	49.726027	49.103139
2Story	30.479452	31.838565
1.5Fin	10.547945	10.014948
SLvl	4.452055	4.783259
SFoyer	2.534247	2.242152
1.5Unf	0.958904	0.822123
2.5Unf	0.753425	0.747384
2.5Fin	0.547945	0.448430
	RoofStyle_orginal	•
Gable	78.150685	77.503737
Hip	19.589041	20.328849
Flat	0.890411	0.822123
Gambrel	0.753425	0.747384
Mansard	0.479452	0.448430
Shed	0.136986	0.149477
	RoofMatl_orginal	RoofMatl_clean
CompShg	98.219178	98.206278
Tar&Grv	0.753425	0.672646
WdShngl	0.410959	0.448430
WdShake	0.342466	0.373692
Metal	0.068493	0.074738
Membran	0.068493	0.074738
Roll	0.068493	0.074738
ClyTile	0.068493	0.074738
·		
	Exterior1st_orgin	al Exterior1st_cle

Exterior1st\_orginal Exterior1st\_clean VinylSd 35.273973 36.322870

${\tt HdBoard}$	15.205479	15.769806
MetalSd	15.068493	15.022422
Wd Sdng	14.109589	13.677130
Plywood	7.397260	7.473842
${\tt CemntBd}$	4.178082	3.886398
${\tt BrkFace}$	3.424658	3.288490
WdShing	1.780822	1.494768
Stucco	1.712329	1.569507
AsbShng	1.369863	1.121076
${\tt BrkComm}$	0.136986	0.074738
Stone	0.136986	0.149477
AsphShn	0.068493	NaN
${\tt ImStucc}$	0.068493	0.074738
CBlock	0.068493	0.074738
	Exterior2nd_orginal	Exterior2nd_clean
VinvlSd	34 520548	35.500747

	Exterior2nd_orginal	${\tt Exterior2nd\_clean}$
VinylSd	34.520548	35.500747
MetalSd	14.657534	14.723468
${\tt HdBoard}$	14.178082	14.723468
Wd Sdng	13.493151	13.153961
Plywood	9.726027	9.491779
${\tt CmentBd}$	4.109589	3.811659
Wd Shng	2.602740	2.391629
Stucco	1.780822	1.718984
${\tt BrkFace}$	1.712329	1.644245
AsbShng	1.369863	1.195815
${\tt ImStucc}$	0.684932	0.747384
Brk Cmn	0.479452	0.448430
Stone	0.342466	0.149477
AsphShn	0.205479	0.149477
Other	0.068493	0.074738
CBlock	0.068493	0.074738

	MasVnrType_orginal	MasVnrType_clean
None	59.178082	57.025411
${\tt BrkFace}$	30.479452	32.286996
Stone	8.767123	9.566517
BrkCmn	1.027397	1.121076

	ExterQual_orginal	ExterQual_clean
TA	62.054795	60.014948
Gd	33.424658	35.650224
Ex	3.561644	3.811659
Fa	0.958904	0.523169

TA Gd Fa Ex Po	ExterCond_orginal 87.808219 10.000000 1.917808 0.205479 0.068493	ExterCond_clean 88.415546 10.239163 1.195815 0.149477 NaN
PCon CBlo Brk' Slad Ston Wood	nc 44.31 ock 43.42 Til 10.00 b 1.64 ne 0.41	4658 43.348281 0000 9.641256 3836 NaN 0959 0.448430
TA Gd Ex Fa	BsmtQual_orginal 44.452055 42.328767 8.287671 2.397260	BsmtQual_clean 44.394619 44.245142 8.968610 2.391629
TA Gd Fa Po	BsmtCond_orginal 89.794521 4.452055 3.082192 0.136986	BsmtCond_clean 92.451420 4.633782 2.840060 0.074738
No Av Gd Mn		9.491779
Unf GLQ ALQ BLQ Rec LwQ	BsmtFinType1_orgi: 29.452 28.630 15.068 10.136 9.109 5.068	055       29.297459         137       30.044843         493       15.620329         986       10.538117         589       9.342302

Unf Rec LwQ BLQ ALQ GLQ	BsmtFinType2_orginal 86.027397 3.698630 3.150685 2.260274 1.301370 0.958904	3.961136 3.437967 4 2.391629 1.420030
GasA GasW Grav Wall OthW Floor	Heating_orginal Heating_orgina	Heating_clean 98.505232 1.195815 0.224215 NaN 0.074738 NaN
Ex TA Gd Fa Po	HeatingQC_orginal He 50.753425 29.315068 16.506849 3.356164 0.068493	eatingQC_clean 52.615845 28.400598 16.218236 2.690583 0.074738
Y N	entralAir_orginal Ce 93.493151 6.506849	entralAir_clean 95.440957 4.559043
SBrk Fuse Fuse Fuse Mix	r 91.369863 A 6.438356 F 1.849315	1.270553 0.149477
TA Gd Ex Fa	KitchenQual_orginal 50.342466 40.136986 6.849315 2.671233	KitchenQual_clean 48.579970 42.451420 7.249626 1.718984
Тур	Functional_orginal 93.150685	Functional_clean 93.721973

Min2 Min1 Mod Maj1 Maj2	2.328767 2.123288 1.027397 0.958904 0.342466	2.242152 2.092676 0.822123 0.747384 0.298954
Sev	0.068493	0.074738
Gara, Attchd Detchd BuiltIn Basment CarPort 2Types	geType_orginal 59.589041 26.506849 6.027397 1.301370 0.616438 0.410959	GarageType_clean 63.677130 27.578475 6.352765 1.420030 0.523169 0.448430
GarageFi	nish_orginal (	GarageFinish_clean
Unf	41.438356	43.348281
RFn	28.904110	30.866966
Fin	24.109589	25.784753
•	1_orginal Gar: 89.794521 3.287671 0.958904 0.205479 0.205479	ageQual_clean 94.917788 3.587444 1.046338 0.224215 0.224215
GarageCon	d_orginal Gara	ageCond clean
_	90.821918	•
Fa	2.397260	2.466368
Gd	0.616438	0.672646
Po	0.479452	0.523169
Ex	0.136986	0.149477
Y 9	_orginal Paved 1.780822 6.164384 2.054795	dDrive_clean 93.946188 4.035874 2.017937
SaleTvr	pe_orginal Sa	leType clean
WD	86.780822	86.547085
New	8.356164	8.744395

COD	2.945205	3.139013
ConLD	0.616438	0.448430
ConLI	0.342466	0.298954
ConLw	0.342466	0.298954
CWD	0.273973	0.298954
Oth	0.205479	0.074738
Con	0.136986	0.149477

	SaleCondition_orginal	SaleCondition_clean
Normal	82.054795	82.511211
Partial	8.561644	8.968610
Abnorml	6.917808	6.427504
Family	1.369863	1.494768
Alloca	0.821918	0.523169
AdjLand	0.273973	0.074738

Above code iterate over each variable in the cat\_var list, call the cat\_var\_dist() function for each variable, and print the resulting DataFrame.

## e-imputation-by-mean-and-median-2

June 11, 2023

### 1 Data Cleaning:

#### 1.1 Missing value imputation by Mean and Median:

Importing necessary libraries:

```
[1]: import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     import warnings
     warnings.filterwarnings('ignore')
     from IPython.display import Image
    data (Original):
[3]: data = pd.read_csv('train.csv') # Loading dataset
[4]: data.shape # Checking the shape of the DataFrame, DataFrame contains Rows:
      →1460 and Columns: 81.
[4]: (1460, 81)
[5]: data.head(2) # Checking first 2 rows from the DataFrame.
[5]:
            MSSubClass MSZoning
                                LotFrontage
                                               LotArea Street Alley LotShape
                                         65.0
     0
                    60
                             RL
                                                  8450
                                                          Pave
                                                                 NaN
                                                                          Reg
     1
                    20
                              RL
                                         80.0
                                                  9600
                                                          Pave
                                                                 NaN
                                                                          Reg
       LandContour Utilities
                             ... PoolArea PoolQC Fence MiscFeature MiscVal MoSold
                                        0
                                             NaN
                                                                          0
                                                                                 2
     0
               Lvl
                      AllPub
                                                   NaN
                                                                NaN
                                                                                 5
     1
               Lvl
                      AllPub
                                        0
                                             NaN
                                                   NaN
                                                                NaN
                                                                          0
       YrSold
               SaleType
                         SaleCondition SalePrice
         2008
                     WD
                                 Normal
                                            208500
         2007
                     WD
                                Normal
                                            181500
     [2 rows x 81 columns]
```

When we check the shape of the data using the command "data.shape", we can observe that the DataFrame consists of 1460 rows and 81 columns. However, when we examine the first two rows of the DataFrame using the command "data.head(2)", it doesn't display all 81 columns. As a result, we utilize the following code, specifically "pd.set\_option()", to address this issue.

```
[6]: pd.set_option('display.max_column', None)
pd.set_option('display.max_rows', None)
```

pd.set\_option('display.max\_columns', None): This line sets the maximum number of columns to be displayed in the output to None, which means there is no limit. As a result, all columns in a DataFrame will be shown when you print or display it.

pd.set\_option('display.max\_rows', None): This line sets the maximum number of rows to be displayed in the output to None, removing any limit. As a result, all rows in a DataFrame will be shown when you print or display it.

```
[7]: data.head(2) # The code `data.head(2)` displays the first two rows of the dataset, showing all 81 columns.
```

```
[7]:
            MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape
         1
                               RL
     0
                     60
                                           65.0
                                                     8450
                                                            Pave
                                                                    NaN
                                                                              Reg
         2
     1
                     20
                               RL
                                           80.0
                                                     9600
                                                            Pave
                                                                    NaN
                                                                              Reg
       LandContour Utilities LotConfig LandSlope Neighborhood Condition1
                Lvl
                       AllPub
                                  Inside
                                                Gtl
                                                          CollgCr
     0
                                                                         Norm
                Lvl
                       AllPub
                                     FR2
                                                Gtl
                                                          Veenker
     1
                                                                        Feedr
       Condition2 BldgType HouseStyle
                                         OverallQual
                                                        OverallCond
                                                                      YearBuilt \
     0
             Norm
                                                     7
                                                                   5
                                                                            2003
                       1Fam
                                 2Story
                                                     6
                                                                   8
     1
                                                                            1976
             Norm
                       1Fam
                                 1Story
        YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType
     0
                                                            VinylSd
                                                                        BrkFace
                 2003
                           Gable CompShg
                                               VinylSd
     1
                 1976
                           Gable
                                  CompShg
                                               MetalSd
                                                            MetalSd
                                                                           None
        MasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
     0
              196.0
                            Gd
                                      TA
                                               PConc
                                                            Gd
                                                                      TA
                                                                                    No
                0.0
                            TA
                                       TA
     1
                                              CBlock
                                                            Gd
                                                                      TA
                                                                                    Gd
       BsmtFinType1
                      BsmtFinSF1 BsmtFinType2
                                                 BsmtFinSF2
                                                              BsmtUnfSF
                                                                          TotalBsmtSF
     0
                 GLQ
                              706
                                            Unf
                                                           0
                                                                     150
                                                                                   856
     1
                 ALQ
                              978
                                            Unf
                                                           0
                                                                     284
                                                                                  1262
       Heating HeatingQC CentralAir Electrical
                                                              2ndFlrSF
                                                                         LowQualFinSF
                                                   1stFlrSF
     0
          GasA
                       Ex
                                    Y
                                            SBrkr
                                                         856
                                                                    854
                                                                                     0
     1
                                    Y
          GasA
                       Ex
                                            SBrkr
                                                        1262
                                                                      0
                                                                                     0
```

GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr \

	0	1710 1262		1		0		2 2	1 0		3	
	•	1202		Ü		_		2	Ü		Ü	
		itchenAbv							Fireplace	es Firepla		
	0		1	Gd			8	Тур		0	NaN	
	1		1	TA			6	Тур		1	TA	
	Ga	rageType	Garage	YrBlt G	arageFi	nish	GarageCai	rs Ga:	rageArea	GarageQua	1 \	
	0	Attchd	_	003.0	O	RFn	O	2	548	T		
	1	Attchd	1	976.0		RFn		2	460	Т	A	
		_	PavedDr			_	nPorchSF	Encl				
	0	TA		Y		0	61			)	0	
	1	TA		Y	298	8	0		(	)	0	
	S	creenPord	h Pool	Area Po	olQC Fe	nce Mi	scFeature	e Mis	cVal MoS	Sold YrSo	ld \	
	0		0	0		NaN	Nal		0	2 20		
	1		0	0		NaN	Nal		0	5 20		
		leType Sa										
	0	WD		rmal	20850							
	1	WD	No	rmal	18150	0						
[8]:	data	+511(2)	# The	code `c	lata tai	7(2)	di cml auc	tho 1	act two	rows of th	2	
[O].		.taii(2) staset.	# 1116	coue o	iaca. car	6(2)	arsprays	<i>1116 1</i>	ust two	rows of th	E	
[8]:			ISSubCla		_		_			lley LotSh	_	
	1458	1459		20	RL		68.0	9717	Pave		Reg	
	1459	1460		20	RL		75.0	9937	Pave	NaN	Reg	
		LandCont	our Uti	lities	LotConf	ig Lan	dSlope Ne	eighbo:	rhood Cor	ndition1	\	
	1458			AllPub	Insi	_	Gtl	_	NAmes	Norm		
	1459		Lvl	AllPub	Insi	de	Gtl	Ed	wards	Norm		
		01:+:	O D1 1	т п	C+1	- 0	1 1 0 1	0	- 110 1	V D 1 +	,	
	1458		onz blag orm	туре но 1Fam	usestyr 1Stor		rallQual 5	uver	alicond 6	YearBuilt 1950		
	1459		orm	1Fam	1Stor	•	5		6	1965		
	1400	1// C	) I III	11 alli	15,001	у	3		O	1900		
		YearRen	nodAdd R	.oofStyl	e RoofMa	atl Ex	terior1s	t Exte	rior2nd N	MasVnrType	\	
	1458		1996	Hi	p Comp	Shg	MetalSo	<b>d</b> 1	MetalSd	None		
	1459		1965	Gabl	e Comp	Shg	HdBoard	<b>d</b> 1	HdBoard	None		
		M 17 4	77 .	0- 3	E C			D C	-1 D . C	\		
	1/150	Masvnr					ndation I					
	1458		0.0	TA		ΓA	CBlock		TA Ta	TA		
	1459		0.0	Gd		ΓΑ	CBlock		TA	TA		
		BsmtExpo	sure Bs	mtFinTy	rpe1 Bsi	ntFinS	F1 BsmtF:	inType	2 BsmtF:	inSF2 \		

1458 1459		Mn No	GLQ BLQ		49 830		Rec LwQ		029 290		
			•								
	BsmtUnfSF			ating He	_						
1458	0		1078	GasA	Go		Y		useA		
1459	136		1256	GasA	Go	d	Y	' SI	Brkr		
	1stFlrSF	2ndFlrSF	LowQu	alFinSF	GrLiv	Area	BsmtFu	llBath	BsmtHal	fBath	\
1458	1078	0		0		1078		1		0	
1459	1256	0		0		1256		1		0	
	FullBath	HalfBath	Rodro	omAbvGr	Ki+ch/	an A ha	Cr Kita	hanOual	\		
1458	1	0	Dearo	2	WI CCII	SIIADV	1	Gd	`		
1459	1	1		3			1	TA			
	TotRmsAbv	Grd Funct		Fireplac		_			_		\
1458		5	Тур		0		NaN	Attchd		950.0	
1459		6	Тур		0		NaN	Attchd	1	965.0	
	GarageFini	sh Garag	eCars	GarageAr	rea Gara	ageQu	ıal Gara	geCond I	PavedDri	ve \	
1458	•	nf	1	_	240	-	TA	TA		Y	
1459		in	1	2	276		TA	TA		Y	
	מו מו זו	T 0 D	1.00		ID 1	0.0	ъ .	a 5			
1.450	WoodDeckS	-		Enclosed		35sn		ScreenPo			
1458 1459	36 73		0 68		112 0		0 0		0		
1409	13	0	00		U		U		U		
	PoolArea	PoolQC Fe	nce Mis	cFeature	e Misc	Val	MoSold	YrSold	SaleTyp	e \	
1458	0	NaN 1	NaN	NaN	I	0	4	2010	W	D	
1459	0	NaN 1	NaN	NaN	I	0	6	2008	W	D	
	SaleCondit	ion Sale	Price								
1458	Nor		42125								
1459	Nor		47500								
data.	info()										
111											
The	data.info(	() · functi	on prov	ides a d	concise	sumn	nary of	the dat	aset, ir	cludin	$\imath g_\sqcup$
⇔in.	formation of	about the	number	of rows	,						
colun	mns, data t	ypes, mem	ory usa	ge and n	nissing	valv	ies.				
,,,											
Z-1-		e	D-+- 77								

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 81 columns):

[9]:

#	Column	Non-Null Count	Dtype
0	Id	1460 non-null	int64
1	MSSubClass	1460 non-null	int64
2	MSZoning	1460 non-null	object
3	LotFrontage	1201 non-null	float64
4	Lotfrontage	1460 non-null	int64
5	Street	1460 non-null	object
6		91 non-null	ū
7	Alley	1460 non-null	object
	LotShape		object
8	LandContour	1460 non-null	object
9	Utilities	1460 non-null	object
10	LotConfig	1460 non-null	object
11	LandSlope	1460 non-null	object
12	Neighborhood	1460 non-null	object
13	Condition1	1460 non-null	object
14	Condition2	1460 non-null	object
15	BldgType	1460 non-null	object
16	HouseStyle	1460 non-null	object
17	OverallQual	1460 non-null	int64
18	OverallCond	1460 non-null	int64
19	YearBuilt	1460 non-null	int64
20	YearRemodAdd	1460 non-null	int64
21	RoofStyle	1460 non-null	object
22	RoofMatl	1460 non-null	object
23	Exterior1st	1460 non-null	object
24	Exterior2nd	1460 non-null	object
25	${ t MasVnrType}$	1452 non-null	object
26	MasVnrArea	1452 non-null	float64
27	ExterQual	1460 non-null	object
28	ExterCond	1460 non-null	object
29	Foundation	1460 non-null	object
30	${\tt BsmtQual}$	1423 non-null	object
31	BsmtCond	1423 non-null	object
32	${ t BsmtExposure}$	1422 non-null	object
33	${\tt BsmtFinType1}$	1423 non-null	object
34	BsmtFinSF1	1460 non-null	int64
35	${\tt BsmtFinType2}$	1422 non-null	object
36	BsmtFinSF2	1460 non-null	int64
37	BsmtUnfSF	1460 non-null	int64
38	TotalBsmtSF	1460 non-null	int64
39	Heating	1460 non-null	object
40	HeatingQC	1460 non-null	object
41	CentralAir	1460 non-null	object
42	Electrical	1459 non-null	object
43	1stFlrSF	1460 non-null	int64
44	2ndFlrSF	1460 non-null	int64
45	LowQualFinSF	1460 non-null	int64
	•		

```
46
    GrLivArea
                    1460 non-null
                                    int64
                                    int64
 47
    BsmtFullBath
                    1460 non-null
 48
    BsmtHalfBath
                    1460 non-null
                                    int64
 49
    FullBath
                    1460 non-null
                                    int64
                    1460 non-null
 50
    HalfBath
                                    int64
 51
    BedroomAbvGr
                    1460 non-null
                                    int64
    KitchenAbvGr
                    1460 non-null
                                    int64
 53 KitchenQual
                    1460 non-null
                                    object
 54 TotRmsAbvGrd
                    1460 non-null
                                    int64
 55 Functional
                    1460 non-null
                                    object
 56 Fireplaces
                    1460 non-null
                                    int64
 57
    FireplaceQu
                    770 non-null
                                    object
 58
    GarageType
                    1379 non-null
                                    object
 59
    GarageYrBlt
                    1379 non-null
                                    float64
 60
    GarageFinish
                    1379 non-null
                                    object
    GarageCars
                    1460 non-null
                                    int64
 61
 62
    GarageArea
                    1460 non-null
                                    int64
 63
    GarageQual
                    1379 non-null
                                    object
 64
    GarageCond
                    1379 non-null
                                    object
 65
    PavedDrive
                    1460 non-null
                                    object
                    1460 non-null
 66
    WoodDeckSF
                                    int64
     OpenPorchSF
                    1460 non-null
                                    int64
 67
    EnclosedPorch 1460 non-null
                                    int64
    3SsnPorch
                    1460 non-null
                                    int64
 70
    ScreenPorch
                    1460 non-null
                                    int64
 71 PoolArea
                    1460 non-null
                                    int64
 72 PoolQC
                    7 non-null
                                    object
 73
    Fence
                    281 non-null
                                    object
 74 MiscFeature
                    54 non-null
                                    object
    MiscVal
                    1460 non-null
                                    int64
    MoSold
                    1460 non-null
                                    int64
 77
    YrSold
                    1460 non-null
                                    int64
 78
    SaleType
                    1460 non-null
                                    object
 79
    SaleCondition
                    1460 non-null
                                    object
    SalePrice
                    1460 non-null
                                    int64
dtypes: float64(3), int64(35), object(43)
memory usage: 924.0+ KB
```

[9]: '\nThe `data.info()` function provides a concise summary of the dataset, including information about the number of rows, \ncolumns, data types, memory usage and missing values.\n\n'

```
[10]: data.isnull().sum()
```

# The code `data.isnull().sum()` calculates the sum of missing values in each  $\hookrightarrow$  column of the dataset.

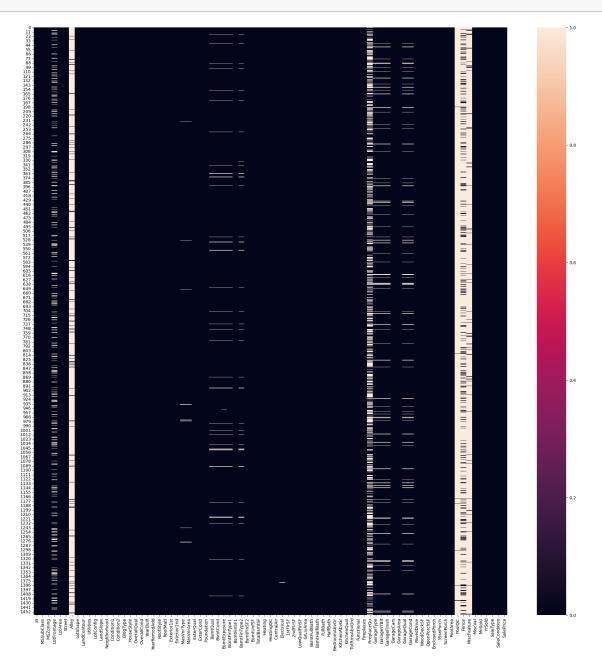
[10]:		0
	MSSubClass	0
	MSZoning	0
	LotFrontage	259
	LotArea	0
	Street	0
	Alley	1369
	LotShape	0
	LandContour	0
	Utilities	0
	LotConfig	0
	LandSlope	0
	Neighborhood	0
	Condition1 Condition2	
		0
	BldgType	
	HouseStyle	0
	OverallQual OverallCond	0
	YearBuilt	0
	YearRemodAdd	0
	RoofStyle	0
	RoofMatl	0
	Exterior1st	0
	Exterior1st	0
	MasVnrType	8
	MasVnrArea	8
	ExterQual	0
	ExterCond	0
	Foundation	0
	BsmtQual	37
	BsmtCond	37
	BsmtExposure	38
	BsmtFinType1	37
	BsmtFinSF1	0
	BsmtFinType2	38
	BsmtFinSF2	0
	BsmtUnfSF	0
	TotalBsmtSF	0
	Heating	0
	HeatingQC	0
	CentralAir	0
	Electrical	1
	1stFlrSF	0
	2ndFlrSF	0
	LowQualFinSF	0
	GrLivArea	0

```
BsmtHalfBath
                          0
      FullBath
                           0
      HalfBath
                           0
      BedroomAbvGr
                           0
      KitchenAbvGr
                          0
     KitchenQual
                          0
      TotRmsAbvGrd
                          0
     Functional
                          0
     Fireplaces
                          0
     FireplaceQu
                        690
      GarageType
                         81
      GarageYrBlt
                          81
      GarageFinish
                         81
      GarageCars
                          0
      GarageArea
                          0
      GarageQual
                         81
      GarageCond
                          81
      PavedDrive
                          0
      WoodDeckSF
                          0
      OpenPorchSF
                          0
      EnclosedPorch
                          0
      3SsnPorch
                          0
      ScreenPorch
                          0
      PoolArea
                          0
     PoolQC
                       1453
     Fence
                       1179
     MiscFeature
                       1406
     MiscVal
                          0
     MoSold
                          0
      YrSold
                          0
                          0
      SaleType
      SaleCondition
                          0
      SalePrice
                           0
      dtype: int64
[11]: plt.figure(figsize=(25,25))
      sns.heatmap(data.isnull())
      plt.show()
      The code `plt.figure(figsize=(25,25))` sets the figure size for the plot.
      The next line `sns.heatmap(data.isnull())` creates a heatmap visualization of \Box
       ⇔the missing values in the dataset
      using the seaborn library.
```

0

BsmtFullBath

1 1 1

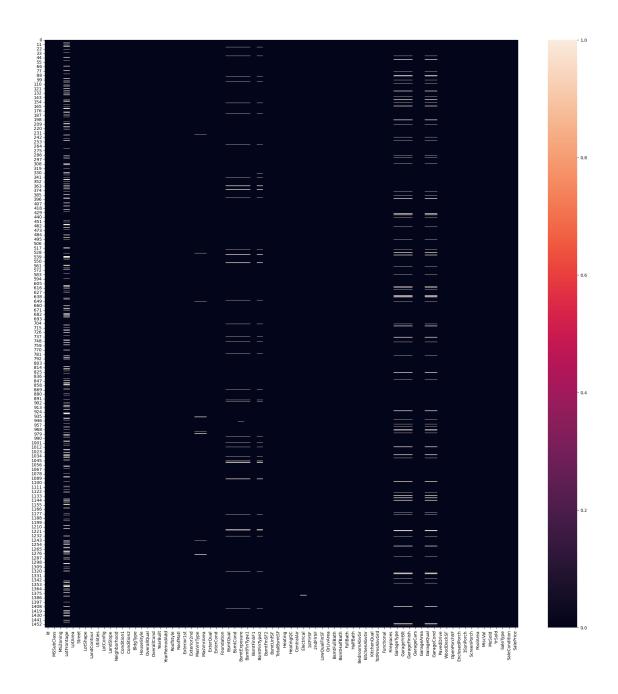


- [11]: '\nThe code `plt.figure(figsize=(25,25))` sets the figure size for the plot. \n\nThe next line `sns.heatmap(data.isnull())` creates a heatmap visualization of the missing values in the dataset \nusing the seaborn library.\n\n'
- [12]: missing\_value\_percent = data.isnull().sum() / data.shape[0] \* 100
  print(missing\_value\_percent)

# The code calculates the percentage of missing values in each column of the  $\Box$  dataset and then prints the resulting percentages.

Id	0.000000
MSSubClass	0.000000
MSZoning	0.000000
LotFrontage	17.739726
LotArea	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
OverallQual	0.000000
OverallCond	0.000000
YearBuilt	0.000000
YearRemodAdd	0.000000
RoofStyle	0.000000
RoofMatl	0.000000
Exterior1st	0.000000
Exterior2nd	0.000000
MasVnrType	0.547945
MasVnrArea	0.547945
ExterQual	0.000000
ExterCond	0.000000
Foundation	0.000000
BsmtQual	2.534247
BsmtCond	2.534247
BsmtExposure	2.602740
BsmtFinType1	2.534247
BsmtFinSF1	0.000000
BsmtFinType2	2.602740
BsmtFinSF2	0.000000
BsmtUnfSF	0.000000
TotalBsmtSF	0.000000
Heating	0.000000
HeatingQC	0.000000
CentralAir	0.000000
Electrical	0.068493
1stFlrSF	0.000000

```
2ndFlrSF
                  0.000000
LowQualFinSF
                  0.000000
GrLivArea
                  0.000000
BsmtFullBath
                  0.000000
BsmtHalfBath
                  0.000000
FullBath
                  0.000000
HalfBath
                  0.000000
BedroomAbvGr
                  0.000000
KitchenAbvGr
                  0.000000
KitchenQual
                  0.000000
TotRmsAbvGrd
                  0.000000
Functional
                  0.000000
Fireplaces
                  0.000000
FireplaceQu
                 47.260274
GarageType
                  5.547945
GarageYrBlt
                  5.547945
GarageFinish
                  5.547945
GarageCars
                  0.000000
GarageArea
                  0.000000
GarageQual
                  5.547945
GarageCond
                  5.547945
PavedDrive
                  0.000000
WoodDeckSF
                  0.000000
OpenPorchSF
                  0.000000
EnclosedPorch
                  0.000000
                  0.000000
3SsnPorch
ScreenPorch
                  0.000000
PoolArea
                  0.000000
PoolQC
                 99.520548
Fence
                 80.753425
MiscFeature
                 96.301370
MiscVal
                  0.000000
MoSold
                  0.000000
YrSold
                  0.000000
                  0.000000
SaleType
SaleCondition
                  0.000000
SalePrice
                  0.000000
dtype: float64
```



[15]: '\nThe code is used to get visualization of the missing values after removing columns \nwhere the missing values exceed 20% of the total data in those columns.\n\n'

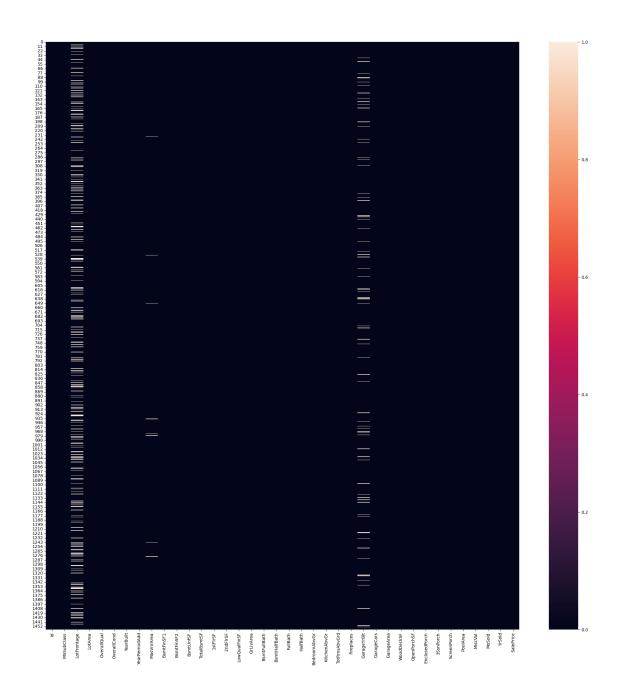
```
[16]: data1.shape # Checking the shape of the DataFrame , DataFrame contains Rows:

→1460 and Columns: 76
```

[16]: (1460, 76)

```
numerical DataFrame (selecting DataFrame that contains only numerical data types)
```

```
[17]: numerical_DataFrame = data1.select_dtypes(include=['int64', 'float64'])
      numerical_DataFrame.head(2)
[17]:
         Ιd
             MSSubClass LotFrontage LotArea OverallQual
                                                               OverallCond
                                                                           YearBuilt \
                      60
                                 65.0
                                           8450
                                                                                  2003
          1
      1
          2
                      20
                                 80.0
                                           9600
                                                            6
                                                                         8
                                                                                  1976
         YearRemodAdd MasVnrArea BsmtFinSF1
                                                 BsmtFinSF2
                                                             BsmtUnfSF
                                                                         TotalBsmtSF
      0
                 2003
                             196.0
                                            706
                                                           0
                                                                    150
                                                                                  856
                                                           0
                 1976
                               0.0
                                            978
                                                                    284
                                                                                 1262
      1
         1stFlrSF
                   2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath \
                         854
                                                  1710
                                                                    1
      0
              856
                                          0
      1
             1262
                           0
                                          0
                                                  1262
                                                                    0
                                                                                   1
         {\tt FullBath\ HalfBath\ BedroomAbvGr\ KitchenAbvGr\ TotRmsAbvGrd\ Fireplaces\ } \\
      0
                 2
                                          3
      1
                2
                           0
                                          3
                                                         1
                                                                       6
                                                                                    1
         GarageYrBlt GarageCars GarageArea WoodDeckSF
                                                             {\tt OpenPorchSF}
              2003.0
      0
                                2
                                           548
                                                                      61
                                                         0
      1
              1976.0
                                2
                                           460
                                                        298
                                                                       0
         EnclosedPorch
                         3SsnPorch ScreenPorch PoolArea MiscVal
                                                                      MoSold
                                                                              YrSold \
                                                                                 2008
      0
                                                         0
                                                                   0
                                                                           2
      1
                      0
                                 0
                                               0
                                                         0
                                                                   0
                                                                           5
                                                                                 2007
         SalePrice
      0
            208500
      1
            181500
[18]: numerical_DataFrame.shape
      # Here, the numerical_DataFrame contains 1460 rows and 38 columns.
[18]: (1460, 38)
[19]: plt.figure(figsize=(25,25))
      sns.heatmap(numerical_DataFrame.isnull())
      plt.show()
      # The code is used to get visualization of the missing values of numerical \Box
        \hookrightarrow DataFrame.
```



```
[63]: numerical_DataFrame_null_Rows = numerical_DataFrame[numerical_DataFrame.

→isnull().any(axis=1)]

numerical_DataFrame_null_Rows.head(5)

///

The code `numerical_DataFrame[numerical_DataFrame.isnull().any(axis=1)]` is_□

→used to filter rows in a DataFrame

called `numerical_DataFrame` that contain at least one null value.
```

7 12 14 16	8 13 15 17	60 20 20 20		NaN NaN NaN NaN	10382 12968 10920 11242	2 3 ) L	7 5 6	7 5 6	6 6 5 7	19 19 19 19	73 62 60 70	
24	25	20		Nan	8246	)	t	)	8	19	80	
	YearRemod	lAdd i	MasVnrA	rea Bs	smtFinSFi	l BsmtF:	inSF2	BsmtUnf	SF To	talBsmtS	F\	
7												
12	1	1962		0.0			0	1	75			
14							0	5	20			
16							0					
24							668					
	1stFlrSF	2ndF	lrSF L	owQualI	FinSF G	LivArea	Bsmt	FullBath	Bsmt	HalfBath	\	
7	1107		983		0	2090		1		0		
12	912		0		0	912		1		0		
14	1253		0		0	1253		1		0		
16	1004		0		0	1004		1		0		
24	1060		0		0	1060		1		0		
		Half:		edroom <i>l</i>		itchenAb		CotRmsAbv		-		
			0									
	1		1									
	1		0				1					
24	1		0		3		1		6		1	
	GarageVrF	R1+ C	arageCa	ra Cai	rageArea	WoodDe	~kSE	OpenPorc	hSE \			
7	•		arageoa		-	WOOdbe		-				
				-	=. •				• •			
	1500											
	Enclosed		3SsnPo	rch So	creenPor	ch Pool	Area	MiscVal	MoSol	d YrSol	d \	
7			3SsnPo	rch So	creenPord	ch Pool	Area O	MiscVal 350		d YrSol		
7 12		orch	3SsnPo		creenPord	0			1		9	
		orch 228	3SsnPo	0		0	0	350	1	1 200	9 8	
12		Porch 228 0	3SsnPo	0 0		0 76	0 0	350 0	1	1 200 9 200	9 8 8	
12 14		Porch 228 0 176	3SsnPo	0 0 0		0 76 0	0 0 0	350 0 0	1	1 200 9 200 5 200	9 8 8 0	
12 14 16		Porch 228 0 176	3SsnPo	0 0 0		0 76 0	0 0 0 0	350 0 0 700	1	1 200 9 200 5 200 3 201	9 8 8 0	
	12 14 16 24 7 12 14 16 24 7 12 14 16 24 7 12 14 16 24	7 8 12 13 14 15 16 17 24 25  YearRemod 7 11 14 11 16 11 24 22  1stFlrSF 7 1107 12 912 14 1253 16 1004 24 1060  FullBath 7 2 12 1 14 1 16 1 24 1  GarageYrF 7 1973 12 1962 14 1960 16 1970	7 8 60 12 13 20 14 15 20 16 17 20 24 25 20  YearRemodAdd 7 1973 12 1962 14 1960 16 1970 24 2001  1stFlrSF 2ndF 7 1107 12 912 14 1253 16 1004 24 1060  FullBath Half 7 2 12 1 14 1 16 1 24 1  GarageYrBlt G 7 1973.0 12 1962.0 14 1960.0 16 1970.0	7 8 60 12 13 20 14 15 20 16 17 20 24 25 20  YearRemodAdd MasVnrA 7 1973 24 12 1962 14 1960 21 16 1970 18 24 2001  1stFlrSF 2ndFlrSF L 7 1107 983 12 912 0 14 1253 0 16 1004 0 24 1060 0  FullBath HalfBath B 7 2 1 12 1 0 14 1 1 16 1 0 24 1 0  GarageYrBlt GarageCa 7 1973.0 12 1962.0 14 1960.0	7 8 60 NaN 12 13 20 NaN 14 15 20 NaN 16 17 20 NaN 24 25 20 NaN  YearRemodAdd MasVnrArea Bs 7 1973 240.0 12 1962 0.0 14 1960 212.0 16 1970 180.0 24 2001 0.0  1stFlrSF 2ndFlrSF LowQualf 7 1107 983 12 912 0 14 1253 0 16 1004 0 24 1060 0  FullBath HalfBath Bedroomf 7 2 1 12 1 0 14 1 1 16 1 0 24 1 0  GarageYrBlt GarageCars Gar 7 1973.0 2 12 1962.0 1 14 1960.0 1 16 1970.0 2	7 8 60 NaN 10382 12 13 20 NaN 12968 14 15 20 NaN 10920 16 17 20 NaN 11241 24 25 20 NaN 1242  7 1973 240.0 853 12 1962 0.0 733 14 1960 212.0 733 16 1970 180.0 578 24 2001 0.0 188  1stFlrSF 2ndFlrSF LowQualFinSF Gr 7 1107 983 0 1 12 912 0 0 0 1 14 1253 0 0 0 1 14 1253 0 0 0 1 14 1253 0 0 0 0 14 1253 0 0 0 0 15 16 1004 0 0 0 0 24 1060 0 0 0 0  FullBath HalfBath BedroomAbvGr Kr 7 2 1 3 1 12 1 0 2 1 14 1 1 2 1 16 1 0 2 2 14 1 0 3 3  GarageYrBlt GarageCars GarageArea 7 1973.0 2 484 12 1962.0 1 352 14 1960.0 1 352 14 1960.0 1 352 14 1960.0 1 352	7       8       60       NaN       10382         12       13       20       NaN       12968         14       15       20       NaN       10920         16       17       20       NaN       10920         16       17       20       NaN       11241         24       25       20       NaN       11241         24       20       0       737       144         1960       212.0       733       166       1970       180.0       578         24       2001       0       188       0       2090         12       912       0       0       912         14       1253       0       0       104         24       1060       0       0       1060         14       1060       0       0       1060         14       1       1	7 8 60 NaN 10382 7 12 13 20 NaN 12968 5 14 15 20 NaN 10920 6 16 17 20 NaN 11241 6 24 25 20 NaN 8246 5  YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 7 1973 240.0 859 32 12 1962 0.0 737 0 14 1960 212.0 733 0 16 1970 180.0 578 0 24 2001 0.0 188 668   IstFlrSF 2ndFlrSF LowQualFinSF GrLivArea Bsmt 7 1107 983 0 2090 12 912 0 0 912 14 1253 0 0 912 14 1253 0 0 1253 16 1004 0 0 1004 24 1060 0 0 0 1004 24 1060 0 0 0 1060  FullBath HalfBath BedroomAbvGr KitchenAbvGr 7 7 2 1 3 3 1 12 1 0 2 1 14 1 1 1 2 1 2 1 16 1 0 2 1 1 16 1 1 0 2 1 1 16 1 1 0 2 1 1 16 1 1 0 3 1 1 17  GarageYrBlt GarageCars GarageArea WoodDeckSF 7 1973.0 2 484 235 12 1962.0 1 352 140 14 1960.0 1 352 140 14 1960.0 1 352 0 0 16 1970.0 2 480 0 0	7 8 60 NaN 10382 7 12 13 20 NaN 12968 5 14 15 20 NaN 10920 6 16 17 20 NaN 11241 6 24 25 20 NaN 8246 5   YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnf 7 1973 240.0 859 32 2 12 1962 0.0 737 0 1 14 1960 212.0 733 0 5 16 1970 180.0 578 0 4 24 2001 0.0 188 668 2   IstFlrSF 2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath 7 1107 983 0 2090 1 12 912 0 0 912 1 14 1253 0 0 1253 1 16 1004 0 0 1253 1 16 1004 0 0 1004 1 24 1060 0 0 0 1004 1 24 1060 0 0 0 1060 1  FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbv 7 2 1 3 1 1 2 1 1 14 1 1 1 2 1 1 16 1 0 2 1 1 16 1 0 2 1 1 16 1 0 3 1 1  GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorce 7 1973.0 2 484 235 12 1962.0 1 352 140 14 1960.0 1 352 0 1 16 1970.0 2 480 0 0	7 8 60 NaN 10382 7 6 12 13 20 NaN 12968 5 6 14 15 20 NaN 10920 6 5 16 17 20 NaN 11241 6 7 24 25 20 NaN 8246 5 8   YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF To 7 1973 240.0 859 32 216 12 1962 0.0 737 0 175 14 1960 212.0 733 0 520 16 1970 180.0 578 0 426 24 2001 0.0 188 668 204   1stFlrSF 2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath Bsmt 7 1107 983 0 2090 1 12 912 0 0 912 1 14 1253 0 0 0 1253 1 16 1004 0 0 0 1004 1 24 1060 0 0 0 1004 1 24 1060 0 0 0 1000 1  FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd F 7 2 1 3 1 7 12 1 0 2 1 4 14 1 1 1 2 1 5 16 1 1 0 2 1 5 16 1 1 0 2 1 5 16 1 1 0 2 1 6  GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF A 7 1973.0 2 484 235 204  14 1960.0 1 352 140 0 0 14 1960.0 1 355 0 213 16 1970.0 2 484 235 204	7         8         60         NaN         10382         7         6         19           12         13         20         NaN         12968         5         6         19           14         15         20         NaN         10920         6         5         19           16         17         20         NaN         11241         6         7         19           24         25         20         NaN         11241         6         7         19           YearRemodAd         MasVnrArea         BsmtFinsF1         BsmtFinsF2         BsmtUnfSF         TotalBsmtS           7         1973         240.0         859         32         216         110           12         1960         212.0         733         0         520         125           16         1970         180.0         578         0         426         100           24         2001         0.0         188         668         204         106           15tFlrSF         2ndFlrSF         LowQualFinSF         GrLivArea         BsmtFullBath         BsmtHalfBath           12         912         0	Table

```
14 15700016 14900024 154000
```

[22]: num\_DF\_MissingValue = numerical\_DataFrame.isnull().sum() print(num\_DF\_MissingValue)

Ιd 0 MSSubClass 0 LotFrontage 259 LotArea 0 0 OverallQual OverallCond 0 YearBuilt 0 YearRemodAdd 0 MasVnrArea 8 BsmtFinSF1 0 BsmtFinSF2 0  ${\tt BsmtUnfSF}$ 0 TotalBsmtSF 0 1stFlrSF 0 2ndFlrSF 0 LowQualFinSF 0 GrLivArea 0 BsmtFullBath 0 BsmtHalfBath 0 FullBath 0 HalfBath 0  ${\tt BedroomAbvGr}$ 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 SalePrice 0 dtype: int64

```
[23]: num_DF_MissingValue_columns = num_DF_MissingValue[num_DF_MissingValue > 0].

keys()

print(num_DF_MissingValue_columns)

///

This code is used to retrieve the column names from a DataFrame called

num_DF_MissingValue where

the corresponding values are greater than zero (indicating missing values).
```

Index(['LotFrontage', 'MasVnrArea', 'GarageYrBlt'], dtype='object')

[23]: '\nThis code is used to retrieve the column names from a DataFrame called num\_DF\_MissingValue where \nthe corresponding values are greater than zero (indicating missing values).\n\n'

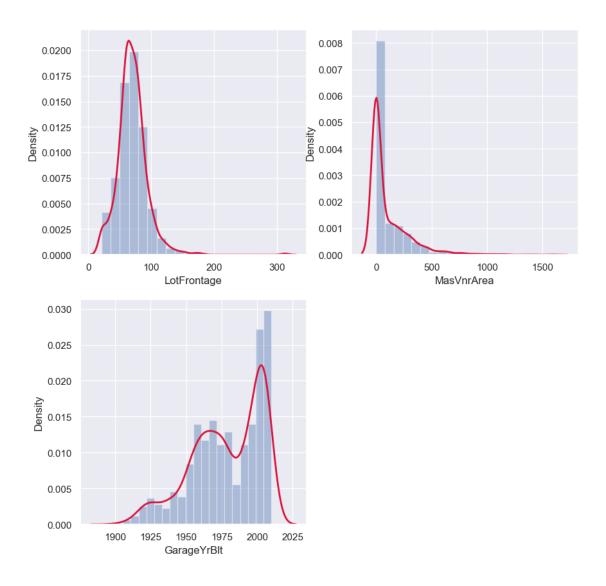
The code creates a figure with a specific size, sets the seaborn style, and then iterates over the num\_DF\_MissingValue\_columns. For each column, it creates a subplot and plots the distribution of the corresponding column from the numerical\_DataFrame using the seaborn distplot function. The bins parameter specifies the number of bins in the histogram, and the kde\_kws parameter sets the properties of the kernel density estimation line (linewidth and color).

```
[24]: # Visualization of Missing Numeric Data Distribution

plt.figure(figsize=(10,10))

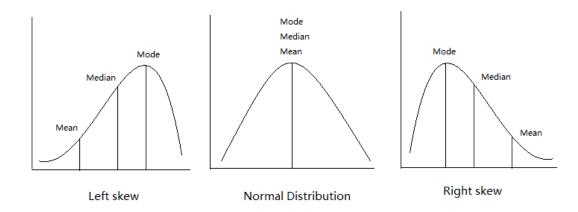
sns.set()

for i, var in enumerate(num_DF_MissingValue_columns):
    plt.subplot(2,2,i+1)
    sns.distplot(numerical_DataFrame[var], bins=20, kde_kws={'linewidth':2,u_d'color':'#DC143C'})
```



[25]: Image('C:/Users/Server/Desktop/NIBZZ/Intellipaat/SQL Project/img.png')

[25]:



When we compare the visualization with the image, we can observe that LotFrontage follows an almost normal distribution, MasVnrArea is skewed to the right, and GarageYrBlt is skewed to the left. Therefore, we use the mean as a measure of central tendency for LotFrontage, and the mode as a measure of central tendency for the other two variables.

# Filling Missing Values in "LotFrontage" Column with Mean as Measure of Central Tendency:

```
[65]: numerical_DataFrame_copy1 = numerical_DataFrame.copy() # Creating a copy of unumerical_DataFrame
numerical_DataFrame_copy1.shape
```

[65]: (1460, 38)

[66]: lotfrontage\_mean = numerical\_DataFrame\_copy1['LotFrontage'].mean() #\_\_

~Calculating mean of column LotFrontage and printing it.

print('LotFrontage Mean: ',lotfrontage\_mean)

LotFrontage Mean: 70.04995836802665

[69]: numerical\_DataFrame\_copy1['LotFrontage'].fillna(LotFrontage\_mean,inplace=True)
numerical\_DataFrame\_copy1[num\_DF\_MissingValue\_columns].isnull().sum()
# Filling the missing value LotFrontage column with it's mean and checking if\_
\_\_\_\_\_\_the missing values of that column is handled.

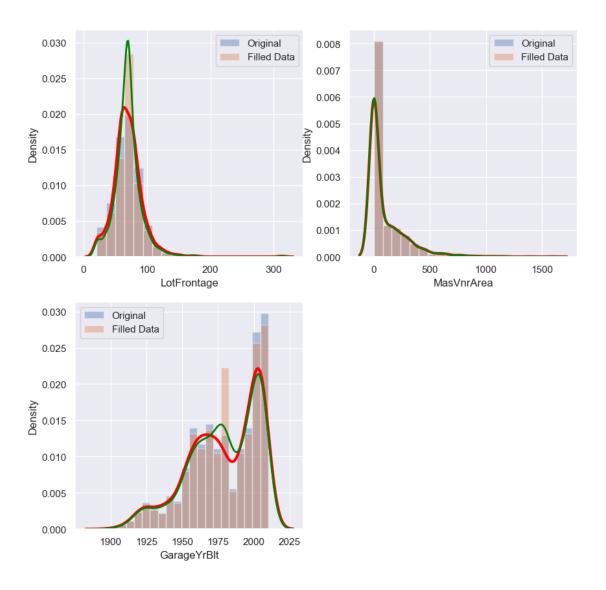
Filling Missing Values in "MasVnrArea" and "GarageYrBlt" Column with Median as

```
Measure of Central Tendency:
[70]: MasVnrArea_Median = numerical_DataFrame_copy1['MasVnrArea'].median()
      print("MasVnrArea_Median: ", MasVnrArea_Median)
      # Finding the median of column MasVnrArea and printing it.
     MasVnrArea_Median: 0.0
[71]: GarageYrBlt_Median = numerical_DataFrame_copy1['GarageYrBlt'].median()
      print("GarageYrBlt_Median: ",GarageYrBlt_Median)
      # Finding the median of column GarageYrBlt and printing it.
     GarageYrBlt_Median: 1980.0
[72]: numerical DataFrame copy1['MasVnrArea'].fillna(MasVnrArea Median, inplace=True)
      numerical_DataFrame_copy1['GarageYrBlt'].fillna(GarageYrBlt_Median, __
       →inplace=True)
      # Filling the missing value MasVnrArea, GarageYrBlt columns with it's median.
```

- [73]: numerical\_DataFrame\_copy1[num\_DF\_MissingValue\_columns].isnull().sum() # Checking if the missing values of that column is handled.
- [73]: LotFrontage 0 MasVnrArea 0 GarageYrBlt 0 dtype: int64

Visualizing Original and Filled Data Distributions for Numerical DataFrame:

```
[74]: plt.figure(figsize=(10,10))
    sns.set()
    for i, var in enumerate(num_DF_MissingValue_columns):
       plt.subplot(2,2,i+1)
        sns.distplot(numerical_DataFrame[var], bins=20, kde kws={'linewidth':3,__
     sns.distplot(numerical_DataFrame_copy1[var], bins=20, kde_kws={'linewidth':
     plt.legend()
```



## Concatenating Numerical DataFrames to Fill Missing Values :

[76]: df\_concat\_DF = df\_concat[df\_concat.isnull().any(axis=1)] df\_concat\_DF.head(5)

[76]:	${ t LotFrontage}$	MasVnrArea	${ t GarageYrBlt}$	${ t LotFrontage}$	MasVnrArea	${ t GarageYrBlt}$
7	NaN	240.0	1973.0	70.049958	240.0	1973.0
12	NaN	0.0	1962.0	70.049958	0.0	1962.0
14	NaN	212.0	1960.0	70.049958	212.0	1960.0
16	NaN	180.0	1970.0	70.049958	180.0	1970.0
24	NaN	0.0	1968.0	70.049958	0.0	1968.0

Based on the information provided above, it is evident that the missing value in the LotFrontage column has been replaced with the mean value of the LotFrontage column. Similarly, the missing values in the MasVnrArea and GarageYrBlt columns have been replaced with the median values of the MasVnrArea and GarageYrBlt columns respectively.

# central-tendency-for-each-class-1

June 11, 2023

# 1 Data Cleaning:

1.1 Missing value imputation of numerical column by Measure of central tendency for each class:

```
Importing necessary libraries
```

```
[1]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

#### data (Original):

```
[2]: data = pd.read_csv("train.csv") # Loading dataset
```

- [3]: (1460, 81)
- [4]: data.head() # Checking first 5 rows from the DataFrame.

[4]:	ld	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	\
C	) 1	60	RL	65.0	8450	Pave	NaN	Reg	
1	. 2	20	RL	80.0	9600	Pave	NaN	Reg	
2	2 3	60	RL	68.0	11250	Pave	NaN	IR1	
3	3 4	70	RL	60.0	9550	Pave	NaN	IR1	
4	<u> 5</u>	60	RL	84.0	14260	Pave	NaN	IR1	

	LandContour	Utilities	•••	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	\
(	Lvl	AllPub	•••	0	NaN	NaN	NaN	0	2	
-	Lvl	AllPub	•••	0	NaN	NaN	NaN	0	5	
2	2 Lvl	AllPub	•••	0	NaN	NaN	NaN	0	9	
3	B Lvl	AllPub	•••	0	NaN	NaN	NaN	0	2	
4	l Lvl	AllPub	•••	0	NaN	NaN	NaN	0	12	

	YrSold	SaleType	${\tt SaleCondition}$	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]

When we check the shape of the data using the command "data.shape", we can observe that the DataFrame consists of 1460 rows and 81 columns. However, when we examine the first five rows of the DataFrame using the command "data.head(5)", it doesn't display all 81 columns. As a result, we utilize the following code, specifically "pd.set\_option()", to address this issue.

```
[6]: pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
```

pd.set\_option('display.max\_columns', None): This line sets the maximum number of columns to be displayed in the output to None, which means there is no limit. As a result, all columns in a DataFrame will be shown when you print or display it.

pd.set\_option('display.max\_rows', None): This line sets the maximum number of rows to be displayed in the output to None, removing any limit. As a result, all rows in a DataFrame will be shown when you print or display it.

```
[7]: data.head(2) # The code `data.head(2)` displays the first two rows of the dataset, showing all 81 columns.
```

```
[7]:
             MSSubClass MSZoning
                                   LotFrontage
        Ιd
                                                  LotArea Street Alley LotShape
     0
         1
                               RL
                                           65.0
                     60
                                                     8450
                                                             Pave
                                                                     NaN
                                                                              Reg
     1
         2
                     20
                               RL
                                           80.0
                                                     9600
                                                             Pave
                                                                     NaN
                                                                              Reg
       LandContour Utilities LotConfig LandSlope Neighborhood Condition1
     0
                Lvl
                        AllPub
                                                 Gtl
                                   Inside
                                                           CollgCr
                                                                          Norm
     1
                Lvl
                        AllPub
                                      FR2
                                                 Gtl
                                                           Veenker
                                                                         Feedr
       Condition2 BldgType HouseStyle
                                          OverallQual
                                                        OverallCond
     0
              Norm
                        1Fam
                                  2Story
                                                     7
                                                                    5
                                                                            2003
                                  1Story
                                                     6
                                                                   8
     1
              Norm
                        1Fam
                                                                            1976
        YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType
     0
                 2003
                           Gable
                                  CompShg
                                                VinylSd
                                                             VinylSd
                                                                         BrkFace
     1
                 1976
                                   CompShg
                                                MetalSd
                                                             MetalSd
                                                                            None
                           Gable
        MasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
     0
              196.0
                            Gd
                                       TΑ
                                                PConc
                                                             Gd
                                                                       TA
                                                                                     No
     1
                0.0
                            TA
                                       TΑ
                                               CBlock
                                                             Gd
                                                                       TA
                                                                                     Gd
```

	F	Rsmt.FinTv	me1	BsmtFinSF	'1 Bsmt.F	inTvne2	Bsmt.Fin9	SF2 B	smt.UnfSl	₹ Total	BsmtSF	\
	0	30mor 1111 y	GLQ	70		Unf	Domor 111	0	150		856	`
	1		ALQ	97		Unf		0	284		1262	
			•									
	I	Heating H	<b>Teatin</b>	gQC Centr	alAir E	lectrica	l 1stFl	rSF 2:	ndFlrSF	LowQua	alFinSF	\
	0	${\tt GasA}$		Ex	Y	SBrk	r 8	356	854		0	
	1	${\tt GasA}$		Ex	Y	SBrk	r 12	262	0		0	
		GrLivAre		mtFullBat	h Bsmt				lfBath	Bedroom		\
	0	171			1	C		2	1		3	
	1	126	52		0	1		2	0		3	
		Vitaban A	hC-	Vitahan Ou	To+	Dwg AbreCw	d Euroti	- Tomo	Finanla	oog Fins		\
	0	KICHEHA	lovgr 1	KitchenQu	Gd	MIISADVGI	d Function		гттерта	0	pracequ NaN	
	1		1		TA		6	Тур Тур		1	TA	
	_		1		IA		O	тур		1	11	
	(	GarageTyp	oe Ga	rageYrBlt	Garage	Finish	GarageCai	rs Ga:	rageArea	a Garage	Qual \	
	0	Attch		2003.0	_	RFn	0	2	548	_	TA	
	1	Attch	nd	1976.0	)	RFn		2	460	)	TA	
	(	GarageCon	nd Pav	edDrive	WoodDec	kSF Ope	nPorchSF	Encl	osedPor	ch 3Ssr	Porch	\
	0		CA.	Y		0	61			0	0	
	1	T	TA.	Y		298	0			0	0	
		~ 5	_	<b>.</b>	D 300		<b>-</b> .					,
	^	ScreenPo		PoolArea 0	PoolQC	NaN	screature. Nal		cval Mo	oSold Y 2	rSold 2008	\
	0		0	0	NaN	NaN	Nai Nai		0	5	2008	
	_		U	U	IVAIV	Ivaiv	Ivai	.V	O	0	2001	
	5	SaleType	SaleC	ondition	SalePr	ice						
	0	WD		Normal		500						
	1	WD		Normal		500						
[8]:	dat	ta.tail(2	2) # T	he code	data.ta	il(2)` a	lisplays	the la	st two	rows of	$the_{f \sqcup}$	
	$\hookrightarrow$	dataset.										
507							_		_			
[8]:	4 4 5	Id	MSSu	bClass MS	•		•			•	-	\
	145			20	RL		68.0	9717	Pave	NaN	Reg	
	145	59 1460		20	RL		75.0	9937	Pave	NaN	Reg	
		LandCo	ntour	Utilitie	s LotCo	nfig Lan	dSlope Ne	oi ghho	rhood Co	ondition	n1 \	
	145		Lvl			side	Gtl	•	NAmes	Nor		
	145		Lvl			side	Gtl		wards	Nor		
	_,		- · <b>-</b>					_•		<b></b>		
		Condit	ion2	BldgType	HouseSt	yle Ove	rallQual	Over	allCond	YearBu	ilt \	
	145	58	Norm	1Fam	1St	ory	5		6	1	.950	
	145	59	Norm	1Fam	1St	ory	5		6	1	.965	

```
YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType \
   1458
              1996
                       Hip CompShg MetalSd MetalSd
   1459
              1965
                                   HdBoard
                     Gable CompShg
                                             HdBoard
                                                        None
        MasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond \
             0.0
                      TΑ
                             TA
                                   CBlock
   1458
                                             TΑ
                                                    TΑ
             0.0
                      Gd
                             TΑ
                                             TΑ
   1459
                                   CBlock
                                                     TΑ
       BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2 \
              Mn GLQ
                                 49 Rec
   1459
              No
                        BLQ
                                 830
                                            LwQ
                                                     290
        BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical \
                      1078
   1458
         0
                            GasA
                                     Gd
                                        Y
   1459 136
                      1256
                            GasA
                                      Gd
                                              Y
                                                    SBrkr
        1stFlrSF 2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath \
   1458
           1078
                    0
                               0
                                      1078
                     0
                                0
                                      1256
                                                              0
   1459
           1256
        FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual \
                         2
   1458
         1 0
                                          1
   1459
            1
                    1
                               3
                                          1
                                                    ТΑ
        TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt \
                       Тур О
                                          \mathtt{NaN}
   1458
                                                 Attchd
               6
                                 0
                                          {\tt NaN}
   1459
                                                 Attchd
                                                           1965.0
                       Тур
       GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive \
                   1
   1458
          Unf
                            240
                                          TA
                                                   ΤA
   1459
              Fin
                        1
                                 276
                                          TA
                                                   TΑ
                                                            Y
        WoodDeckSF OpenPorchSF EnclosedPorch 3SsnPorch ScreenPorch \
                            112 0
   1458
             366
                  0
                        68
                                             0
             736
                                   0
   1459
        PoolArea PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType \
   1458
            0 \quad \text{NaN}
                      NaN NaN
                                   0 4 2010
   1459
            0
                               {\tt NaN}
                                    0 6
                                                  2008
                                                           WD
                 NaN
                      NaN
       SaleCondition SalePrice
            Normal 142125
   1458
   1459
           Normal
                    147500
[9]: data.info()
    111
```

The `data.info()` function provides a concise summary of the dataset, including  $\cup$  information about the number of rows, columns, data types, memory usage and missing values.

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 81 columns):

#	Column	Non-Null Count	Dtype
0	Id	1460 non-null	int64
1	MSSubClass	1460 non-null	int64
2	MSZoning	1460 non-null	object
3	LotFrontage	1201 non-null	float64
4	LotArea	1460 non-null	int64
5	Street	1460 non-null	object
6	Alley	91 non-null	object
7	LotShape	1460 non-null	object
8	LandContour	1460 non-null	object
9	Utilities	1460 non-null	object
10	LotConfig	1460 non-null	object
11	LandSlope	1460 non-null	object
12	Neighborhood	1460 non-null	object
13	Condition1	1460 non-null	object
14	Condition2	1460 non-null	object
15	BldgType	1460 non-null	object
16	HouseStyle	1460 non-null	object
17	OverallQual	1460 non-null	int64
18	OverallCond	1460 non-null	int64
19	YearBuilt	1460 non-null	int64
20	${\tt YearRemodAdd}$	1460 non-null	int64
21	RoofStyle	1460 non-null	object
22	RoofMatl	1460 non-null	object
23	Exterior1st	1460 non-null	object
24	Exterior2nd	1460 non-null	object
25	${ t MasVnrType}$	1452 non-null	object
26	MasVnrArea	1452 non-null	float64
27	ExterQual	1460 non-null	object
28	ExterCond	1460 non-null	object
29	Foundation	1460 non-null	object
30	BsmtQual	1423 non-null	object
31	${\tt BsmtCond}$	1423 non-null	object
32	${\tt BsmtExposure}$	1422 non-null	object
33	${\tt BsmtFinType1}$	1423 non-null	object
34	BsmtFinSF1	1460 non-null	int64
35	BsmtFinType2	1422 non-null	object

36	BsmtFinSF2	1460 non-null	int64
37	BsmtUnfSF	1460 non-null	int64
38	TotalBsmtSF	1460 non-null	int64
39	Heating	1460 non-null	object
40	${\tt HeatingQC}$	1460 non-null	object
41	CentralAir	1460 non-null	object
42	Electrical	1459 non-null	object
43	1stFlrSF	1460 non-null	int64
44	2ndFlrSF	1460 non-null	int64
45	${ t LowQualFinSF}$	1460 non-null	int64
46	GrLivArea	1460 non-null	int64
47	BsmtFullBath	1460 non-null	int64
48	BsmtHalfBath	1460 non-null	int64
49	FullBath	1460 non-null	int64
50	HalfBath	1460 non-null	int64
51	BedroomAbvGr	1460 non-null	int64
52	KitchenAbvGr	1460 non-null	int64
53	KitchenQual	1460 non-null	object
54	TotRmsAbvGrd	1460 non-null	int64
55	Functional	1460 non-null	object
56	Fireplaces	1460 non-null	int64
57	FireplaceQu	770 non-null	object
58	GarageType	1379 non-null	object
59	GarageYrBlt	1379 non-null	float64
60	GarageFinish	1379 non-null	
	•		object
61	GarageCars	1460 non-null	int64
62	GarageArea	1460 non-null	int64
63	GarageQual	1379 non-null	object
64	GarageCond	1379 non-null	object
65	PavedDrive	1460 non-null	object
66	WoodDeckSF	1460 non-null	int64
67	OpenPorchSF	1460 non-null	int64
68		1460 non-null	int64
69		1460 non-null	
	ScreenPorch	1460 non-null	int64
	PoolArea	1460 non-null	int64
72	PoolQC	7 non-null	object
73	Fence	281 non-null	object
	MiscFeature	54 non-null	object
75	MiscVal	1460 non-null	int64
76	MoSold	1460 non-null	int64
77	YrSold	1460 non-null	int64
78	SaleType	1460 non-null	object
79	SaleCondition	1460 non-null	object
80	SalePrice	1460 non-null	int64
dtyp	es: float64(3),	int64(35), obj	ect(43)
	ry usage: 924.0	-	
	. •		

[9]: '\nThe `data.info()` function provides a concise summary of the dataset, including information about the number of rows, \ncolumns, data types, memory usage and missing values.\n\n'

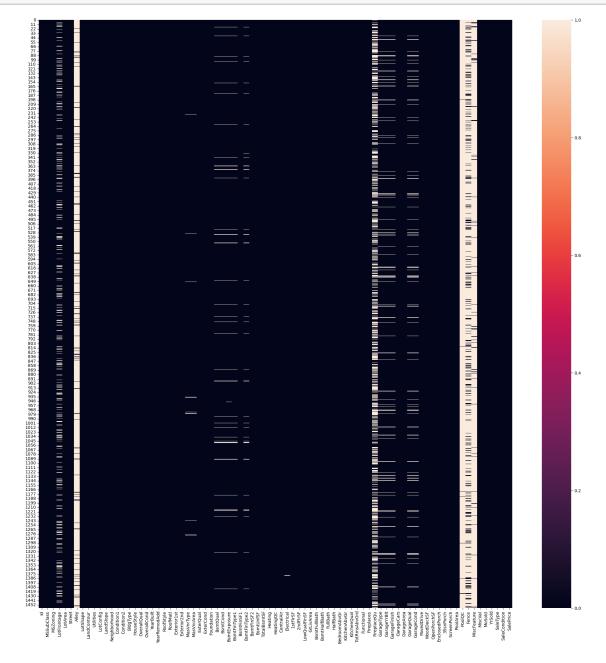
## [10]: data.isnull().sum()

[10]:	Id	0
[10].	MSSubClass	0
	MSZoning	0
	LotFrontage	259
	LotArea	0
	Street	0
	Alley	1369
	LotShape	0
	LandContour	0
	Utilities	0
		0
	LotConfig	
	LandSlope	0
	Neighborhood	0
	Condition1	0
	Condition2	0
	BldgType	0
	HouseStyle	0
	OverallQual	0
	OverallCond	0
	YearBuilt	0
	YearRemodAdd	0
	RoofStyle	0
	RoofMatl	0
	Exterior1st	0
	Exterior2nd	0
	${ t MasVnrType}$	8
	${ t MasVnrArea}$	8
	ExterQual	0
	ExterCond	0
	Foundation	0
	BsmtQual	37
	BsmtCond	37
	BsmtExposure	38
	BsmtFinType1	37
	BsmtFinSF1	0
	BsmtFinType2	38
	BsmtFinSF2	0
	BsmtUnfSF	0
	TotalBsmtSF	0
	Heating	0
	HeatingQC	0
		•

```
0
      CentralAir
      Electrical
                           1
                           0
      1stFlrSF
                           0
      2ndFlrSF
      LowQualFinSF
                           0
      GrLivArea
                           0
      BsmtFullBath
                           0
      BsmtHalfBath
                           0
      FullBath
                           0
      HalfBath
                           0
      BedroomAbvGr
                           0
      KitchenAbvGr
                           0
      KitchenQual
                           0
      TotRmsAbvGrd
                           0
      Functional
                           0
      Fireplaces
                           0
      FireplaceQu
                         690
      GarageType
                          81
      GarageYrBlt
                          81
      GarageFinish
                          81
      GarageCars
                           0
      GarageArea
                           0
      GarageQual
                          81
      GarageCond
                          81
      PavedDrive
                           0
      WoodDeckSF
                           0
      OpenPorchSF
                           0
      EnclosedPorch
                           0
      3SsnPorch
                           0
      ScreenPorch
                           0
      PoolArea
                           0
      PoolQC
                        1453
      Fence
                        1179
      MiscFeature
                        1406
      MiscVal
                           0
      MoSold
                           0
      YrSold
                           0
      SaleType
                           0
      SaleCondition
                           0
                           0
      SalePrice
      dtype: int64
[12]: plt.figure(figsize=(25,25))
      sns.heatmap(data.isnull())
      plt.show()
```

The code `plt.figure(figsize=(25,25))` sets the figure size for the plot.

The next line `sns.heatmap(data.isnull())` creates a heatmap visualization of the missing values in the dataset using the seaborn library.



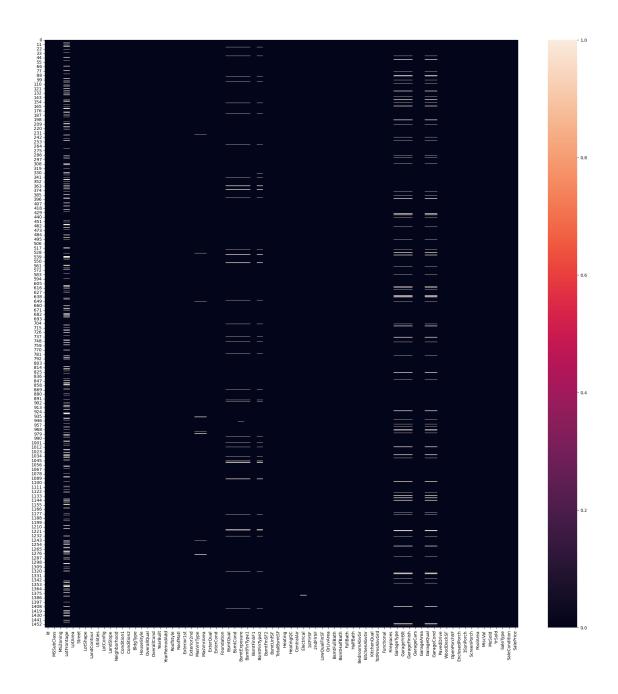
- [12]: '\nThe code `plt.figure(figsize=(25,25))` sets the figure size for the plot. \n\nThe next line `sns.heatmap(data.isnull())` creates a heatmap visualization of the missing values in the dataset \nusing the seaborn library.\n\n'
- [15]: missing\_value\_percent = data.isnull().mean() \* 100
  print(missing\_value\_percent)
  - # The code calculates the percentage of missing values in each column of the  $\Box$   $\Box$  dataset and then prints the resulting percentages.

Id	0.000000
MSSubClass	0.000000
MSZoning	0.000000
LotFrontage	17.739726
LotArea	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
OverallQual	0.000000
OverallCond	0.000000
YearBuilt	0.000000
YearRemodAdd	0.000000
RoofStyle	0.000000
RoofMatl	0.000000
Exterior1st	0.000000
Exterior2nd	0.000000
${\tt MasVnrType}$	0.547945
MasVnrArea	0.547945
ExterQual	0.000000
ExterCond	0.000000
Foundation	0.000000
BsmtQual	2.534247
BsmtCond	2.534247
${\tt BsmtExposure}$	2.602740
BsmtFinType1	2.534247
BsmtFinSF1	0.000000
${\tt BsmtFinType2}$	2.602740
BsmtFinSF2	0.000000

BsmtUnfSF	0.000000
TotalBsmtSF	0.000000
Heating	0.000000
HeatingQC	0.000000
CentralAir	0.000000
Electrical	0.068493
1stFlrSF	0.000000
2ndFlrSF	0.000000
LowQualFinSF	0.000000
GrLivArea	0.000000
BsmtFullBath	0.000000
BsmtHalfBath	0.000000
FullBath	0.000000
HalfBath	0.000000
BedroomAbvGr	0.000000
KitchenAbvGr	0.000000
KitchenQual	0.000000
TotRmsAbvGrd	0.000000
Functional	0.000000
Fireplaces	0.000000
FireplaceQu	47.260274
GarageType	5.547945
GarageYrBlt	5.547945
GarageFinish	5.547945
GarageCars	0.000000
GarageArea	0.000000
GarageQual	5.547945
GarageCond	5.547945
PavedDrive	0.000000
WoodDeckSF	0.000000
OpenPorchSF	0.000000
EnclosedPorch	0.000000
3SsnPorch	0.000000
ScreenPorch	0.000000
PoolArea	0.000000
PoolQC	99.520548
Fence	80.753425
MiscFeature	96.301370
MiscVal	0.000000
MoSold	0.000000
YrSold	0.000000
SaleType	0.000000
SaleCondition	0.000000
SalePrice	0.000000

dtype: float64

```
[16]: missing_value_column = missing_value_percent[missing_value_percent > 20].keys()
      print(missing_value_column)
      111
      The code is used to identify the columns in a dataset that have missing values \Box
       \rightarrow exceeding 20%.
      It retrieves the keys (column names) from the `missing_value_percent'
      ⇔dictionary where the corresponding values
      are greater than 20%.
      111
     Index(['Alley', 'FireplaceQu', 'PoolQC', 'Fence', 'MiscFeature'],
     dtype='object')
     data1 after dropping missing value column > 20\%:
[17]: data1 = data.drop(columns=missing_value_column)
[18]: data1.shape # Checking the shape of the DataFrame , DataFrame contains Rows:
       →1460 and Columns: 76
[18]: (1460, 76)
[20]: plt.figure(figsize=(25,25))
      sns.heatmap(data1.isnull())
      plt.show()
      The code is used to get visualization of the missing values after removing \Box
      ⇔columns
      where the missing values exceed 17% of the total data in those columns.
      111
```



Our objective is to eliminate all the white lines from the visualization in order to achieve a cleaner and more organized appearance.

#### numerical\_DataFrame (selecting DataFrame that contains only numerical data types)

```
[22]: numerical_dataFrame = data.select_dtypes(include=['int64','float64'])
numerical_dataFrame.head(2)
```

```
{\tt Id\ MSSubClass\ LotFrontage\ LotArea\ OverallQual\ OverallCond\ YearBuilt\ } \\
                     60
                                65.0
                                         8450
                                                                               2003
          1
                                                         7
         2
                     20
                                80.0
                                         9600
                                                          6
      1
                                                                       8
                                                                               1976
         YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF
                                                                       TotalBsmtSF \
      0
                 2003
                            196.0
                                          706
                                                        0
                                                                  150
                                                                               856
                 1976
                              0.0
                                          978
                                                        0
                                                                  284
                                                                              1262
      1
                   2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath \
         1stFlrSF
                        854
                                                1710
      0
              856
                                        0
                                        0
                                                1262
                                                                  0
      1
             1262
                          0
                                                                                1
         FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces \
                2
                                        3
      0
                          1
                                                      1
                2
                          0
                                        3
      1
                                                       1
                                                                     6
                                                                                 1
         GarageYrBlt GarageCars GarageArea WoodDeckSF
                                                          OpenPorchSF \
      0
              2003.0
                               2
                                         548
                                                       0
      1
              1976.0
                               2
                                         460
                                                     298
                                                                     0
         EnclosedPorch 3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold \
      0
                                0
                                             0
                                                       0
                                                                 0
                                                                         2
                                                                              2008
                     0
                                0
                                             0
                                                       0
                                                                 0
      1
                                                                         5
                                                                              2007
         SalePrice
      0
            208500
      1
            181500
[23]: numerical_dataFrame.shape # Here, the numerical DataFrame contains 1460 rows_
       ⇔and 38 columns.
[23]: (1460, 38)
[27]: Numerical_DF_missing_value = numerical_dataFrame.isnull().sum() # Checking the_
       ⇔null values of numerical DataFrame.
      Numerical_DF_missing_value
[27]: Id
                         0
      MSSubClass
                         0
      LotFrontage
                       259
      LotArea
                         0
      OverallQual
                         0
      OverallCond
      YearBuilt
      YearRemodAdd
      MasVnrArea
      BsmtFinSF1
```

```
BsmtFullBath
     BsmtHalfBath
                         0
     FullBath
                         0
     HalfBath
      BedroomAbvGr
     KitchenAbvGr
                         0
      TotRmsAbvGrd
                         0
     Fireplaces
                         0
      GarageYrBlt
                        81
      GarageCars
                         0
      GarageArea
      WoodDeckSF
      OpenPorchSF
      EnclosedPorch
      3SsnPorch
                         0
      ScreenPorch
                         0
     PoolArea
                         0
     MiscVal
     MoSold
                         0
      YrSold
                         0
      SalePrice
                         0
      dtype: int64
[30]: Num_DF_missing_value_columns =
       →Numerical_DF_missing_value[Numerical_DF_missing_value > 0].keys()
      Num_DF_missing_value_columns
      # Checking columns names of numerical_dataFrame having missing value > 0
[30]: Index(['LotFrontage', 'MasVnrArea', 'GarageYrBlt'], dtype='object')
[33]: Missing_col_DF =
       →numerical_dataFrame[Num_DF_missing_value_columns][numerical_dataFrame[Num_DF_missing_value_
       →isnull().any(axis=1)]
      Missing_col_DF.head()
      # Having look at columns having missing value
[33]:
          LotFrontage MasVnrArea GarageYrBlt
      7
                  NaN
                            240.0
                                        1973.0
```

BsmtFinSF2

BsmtUnfSF

GrLivArea

TotalBsmtSF 1stFlrSF 2ndFlrSF  ${\tt LowQualFinSF}$  0

0

0

0

12	NaN	0.0	1962.0
14	NaN	212.0	1960.0
16	NaN	180.0	1970.0
24	NaN	0.0	1968.0

In this process, we impute missing values in a numerical column by using a measure of central tendency for each class. To find the missing value, we need to consider a categorical column that is related to it. This requires having domain knowledge. Then, we proceed to impute the missing value by taking the mean or median, depending on the different categories within the categorical column.

#### Handling missing value by class:

[81]:	da	ata1.head()	3								
[81]:		Id MSSub(	Class MSZoni	ng L	otFronta	ge Lot.	Area	Street	LotShape	LandContour	\
	0	1	60	RL	65	.0	3450	Pave	Reg	Lvl	
	1	2	20	RL	80	.0	9600	Pave	Reg	Lvl	
	2	3	60	RL	68	.0 1	1250	Pave	IR1	Lvl	
	3	4		RL			9550	Pave	IR1	Lvl	
	4	5	60	RL	84	.0 1	4260	Pave	IR1	Lvl	
		Utilities I	LotConfig La	ndSlo	pe Neigh	.borhood	Conc	dition1	Condition	2 BldgType	\
	0	AllPub	Inside		-	CollgCr		Norm	Nor		
	1	AllPub	FR2	G	tl	Veenker		Feedr	Nor	m 1Fam	
	2	AllPub	Inside	G	tl	CollgCr		Norm	Nor	m 1Fam	
	3	AllPub	Corner	G	tl	Crawfor		Norm	Nor	m 1Fam	
	4	AllPub	FR2	G	tl	NoRidge		Norm	Nor	m 1Fam	
		HouseStyle	OverallQua	1 Ov	erallCon	.d Year	Built	: YearR	emodAdd R	oofStyle \	
	0	2Story		7		5	2003		2003	Gable	
	1	1Story		6		8	1976	3	1976	Gable	
	2	2Story		7		5	2001	L	2002	Gable	
	3	2Story		7		5	1915	5	1970	Gable	
	4	2Story		8		5	2000	)	2000	Gable	
		RoofMatl Ex	kterior1st E	xteri	or2nd Ma	.sVnrTyp	e Ma	asVnrAre	a ExterQu	al ExterCond	. \
	0	CompShg	VinylSd		nylSd	BrkFac		196.		Gd TA	
	1	CompShg	MetalSd		talSd	Non	е	0.	0	TA TA	
	2	CompShg	VinylSd	Vi	nylSd	BrkFac	е	162.	0	Gd TA	
	3	CompShg	Wd Sdng	Wd	Shng	Non	е	0.	0	TA TA	
	4	CompShg	VinylSd	Vi	nylSd	BrkFac	е	350.	0	Gd TA	
		Foundation	BsmtQual Bs	mtCon	.d BsmtEx	posure 1	BsmtI	FinType1	BsmtFin	SF1 \	
	0	PConc	Gd		'A	No		GLQ		706	
	1	CBlock	Gd	Т	'A	Gd		ALQ		978	
	2	PConc	Gd	Т	'A	Mn		GLQ		486	
	3	BrkTil	TA	G	d	No		ALQ		216	

4	PConc	Gd	TA	I	lν	GLQ	6	555		
	BsmtFinType2	BsmtFinSF	2 Bsmt.Uı	nfSF Tot	alBsmtSF	Heating	Heating	QC \		
0	Unf		)	150	856	GasA	_	Ex		
1	Unf		)	284	1262	GasA		Ex		
2	Unf		)	434	920	GasA		Ex		
3	Unf		)	540	756	GasA		Gd		
4	Unf	(	)	490	1145	GasA		Ex		
	CentralAir E	lectrical :	1stFlrSF	2ndFlrS	SF LowQua	alFinSF	GrLivAr	·ea \		
0	Y	SBrkr	856	85	54	0	17	'10		
1	Y	SBrkr	1262		0	0	12	262		
2	Y	SBrkr	920	86		0		'86		
3	Y	SBrkr	961	75		0		17		
4	Y	SBrkr	1145	105	53	0	21	.98		
	BsmtFullBat1	h BsmtHalfl	Bath Fu	llBath F	HalfBath	BedroomA	AbvGr K		vGr	\
0		1	0	2	1		3		1	
1	(	0	1	2	0		3		1	
2		1	0	2	1		3		1	
3		1	0	1	0		3		1	
4		1	0	2	1		4		1	
-		L	O	2	1		4		1	
	KitchenQual	TotRmsAbvG	rd Funct:	ional Fi	replaces	GarageTy	pe Gar	ageYrBlt	\	
0	Gd		8	Тур	0	Atto	chd	2003.0		
1	TA		6	Тур	1	Atto	chd	1976.0		
2	Gd		6	Тур	1	Atto		2001.0		
3	Gd		7	Тур	1	Deto		1998.0		
4	Gd		9		1	Atto		2000.0		
4	Ga		9	Тур	1	ACCC	ZIIG	2000.0		
	GarageFinish	GarageCar	s Garage	eArea Gar	ageQual (	GarageCor	nd Paved	Drive \		
0	RFn		2	548	TA	7	ΓΑ	Y		
1	RFn	4	2	460	TA	7	ΓΑ	Y		
2	RFn	2	2	608	TA	7	ГΑ	Y		
3	Unf		3	642	TA		ΓΑ	Y		
4	RFn		3	836	TA		ГΑ	Y		
-	141 11	·		000		-		-		
	WoodDeckSF	OpenPorchSl	F Enclos	sedPorch	3SsnPor	ch Scree	enPorch	PoolAre	a \	
0	0	6:	1	0		0	0		0	
1	298	(	)	0		0	0		0	
2	0	4:	2	0		0	0		0	
3	0	3!		272		0	0		0	
4	192	84		0		0	0		0	
<del>-1</del>	132	04	1	U		5	U		J	
	MiscVal Mos	Sold YrSolo	d SaleTvi	pe SaleCo	ndition	SalePric	ce			
0	0	2 2008		ND	Normal	20850				
1	0	5 200		₩D	Normal	18150				
1	J	5 200	١	עיי	MOTHIAT	1010(	,,,			

```
2
         0
                  9
                        2008
                                    WD
                                               Normal
                                                           223500
3
         0
                  2
                        2006
                                    WD
                                              Abnorml
                                                           140000
4
         0
                 12
                        2008
                                    WD
                                               Normal
                                                           250000
```

```
To analyze missing value of LotFrontage let's analyze LotConfig column.
[82]: data1['LotConfig'].unique() # Here, we check different categories under_
       \hookrightarrowLotConfig column.
[82]: array(['Inside', 'FR2', 'Corner', 'CulDSac', 'FR3'], dtype=object)
[83]: mean_lotfrontage = data1[data1.loc[:,'LotConfig'] == "Inside"]["LotFrontage"].
       oreplace(np.nan, data1[data1.loc[:,'LotConfig'] == "Inside"]["LotFrontage"].
       →mean())
      mean_lotfrontage.head(10)
[83]: 0
            65.000000
      2
            68.000000
      5
            85.000000
      6
            75.000000
      8
            51.000000
      10
            70.000000
      11
            85.000000
      12
            67.715686
      13
            91.000000
      17
            72.000000
      Name: LotFrontage, dtype: float64
[87]: data1_copy_1 = data1.copy()
      data1_copy_1.isnull().sum()
[87]: Id
                          0
```

MSSubClass 0 MSZoning 0 LotFrontage 259 LotArea 0 Street 0 LotShape 0 LandContour 0 Utilities 0 LotConfig 0 LandSlope 0 Neighborhood 0 Condition1 0 Condition2 0 BldgType 0 HouseStyle 0

OverallQual         0           OverallCond         0           YearBuilt         0           YearRemodAdd         0           RoofStyle         0           RoofMatl         0           Exterior1st         0           Exterior2nd         0           MasVnrType         8           MasVnrArea         8           ExterQual         0           Foundation         0           Foundation         0           Foundation         0           BsmtQual         37           BsmtCond         37           BsmtCond         37           BsmtExposure         38           BsmtFinType1         37           BsmtFinSF1         0           BsmtFinType2         38           BsmtFinType2         38           BsmtFinType2         38           BsmtFinType2         38           BsmtFinSF2         0           BsmtInfsF         0           CentralAir         0           CentralAir         0           CentralAir         0           GrLivArea         0           BsmtFullBath <td< th=""><th></th><th></th></td<>		
OverallCond         0           YearBuilt         0           YearRemodAdd         0           RoofStyle         0           RoofMatl         0           Exterior1st         0           Exterior2nd         0           MasVnrType         8           MasVnrArea         8           ExterQual         0           Foundation         0           Foundation         0           BsmtQual         37           BsmtCond         37           BsmtCond         37           BsmtCond         37           BsmtFungal         3           BsmtFinSF1         0           BsmtFinSF1         0           BsmtFinSF2         0           BsmtFinSF2         0           BsmtUnfSF         0           CentralAir         0           CentralAir         0           CentralAir         0           CuoQualFinSF         0           ChuvQualFinSF         0           GrLivArea         0           BsmtHalfBath         0           FullBath         0           HalfBath         0	OverallQual	0
YearBuilt         0           YearRemodAdd         0           RoofStyle         0           RoofMatl         0           Exterior1st         0           Exterior2nd         0           MasVnrType         8           MasVnrArea         8           ExterQual         0           ExterCond         0           Foundation         0           BsmtQual         37           BsmtCond         37           BsmtCond         37           BsmtFinType1         37           BsmtFinType1         37           BsmtFinType1         37           BsmtFinType1         37           BsmtFinType2         38           BsmtFinType2         0           CentralAir         0           CentralAir         0           CentralAir         0           CentralAir         0           Cublic		0
YearRemodAdd         0           RoofStyle         0           RoofMatl         0           Exterior1st         0           Exterior2nd         0           MasVnrType         8           MasVnrArea         8           ExterQual         0           ExterCond         0           Foundation         0           BsmtQual         37           BsmtCond         37           BsmtExposure         38           BsmtFinType1         37           BsmtFinSF1         0           BsmtFinType1         37           BsmtFinSF1         0           BsmtFinSF2         0           BsmtFinSF2         0           BsmtUnfSF         0           CentralAir         0           CentralAir         0           CentralAir         0           CentralAir         0           Charles         0           BsmtFuls         0           GrLivArea         0           BsmtFullBath         0           BedroomAbvGr         0           KitchenQual         0           TotRmsAbvGrd         0		
RoofStyle         0           RoofMatl         0           Exterior1st         0           Exterior2nd         0           MasVnrType         8           MasVnrArea         8           ExterQual         0           ExterCond         0           Foundation         0           BsmtQual         37           BsmtCond         37           BsmtExposure         38           BsmtFinSF1         0           BsmtFinSF2         0           BsmtFinSF2         0           BsmtFinSF2         0           BsmtUnfSF         0           TotalBsmtSF         0           Heating         0           HeatingQC         0           CentralAir         0           Electrical         1           1stFlrSF         0           2ndFlrSF         0           CrivArea         0           BsmtFullBath         0           FullBath         0           HalfBath         0           KitchenQual         0           KitchenAbvGr         0           KitchenAbvGrd         0 <t< td=""><td></td><td></td></t<>		
RoofMatl         0           Exterior1st         0           Exterior2nd         0           MasVnrType         8           MasVnrArea         8           ExterQual         0           ExterCond         0           Foundation         0           BsmtQual         37           BsmtCond         37           BsmtExposure         38           BsmtFinType1         37           BsmtFinSF1         0           BsmtFinType2         38           BsmtFinType2         38           BsmtFinSF2         0           BsmtFinSF2         0           BsmtUnfSF         0           CentralAir         0           Electrical         1           1stFlrSF         0           CondralAir         0           Electrical         1           1stFlrSF         0           CondPlrSF         0           CowQualFinSF         0           GrLivArea         0           BsmtHalfBath         0           FullBath         0           KitchenAbvGr         0           KitchenQual         0		
Exterior1st 0 Exterior2nd 0 MasVnrType 8 MasVnrArea 8 ExterQual 0 ExterCond 0 Foundation 0 BsmtQual 37 BsmtCond 37 BsmtExposure 38 BsmtFinType1 37 BsmtFinSF1 0 BsmtFinSF2 0 BsmtFinSF2 0 BsmtUnfSF 0 TotalBsmtSF 0 Heating 0 HeatingQC 0 CentralAir 0 Electrical 1 1stFlrSF 0 2ndFlrSF 0 CndFlrSF 0 EndFlrSF 0 EndFlrSF 0 ChivArea 0 BsmtFullBath 0 BsmtHalfBath 0 BsmtHalfBath 0 FullBath 0 HalfBath 0 HalfBath 0 FullBath 0 Functional 0 Fireplaces 0 GarageType 81 GarageType 81 GarageFinish 81 GarageCars 0 GarageArea 0 GarageArea 0 GarageQual 81	•	
Exterior2nd		
MasVnrType MasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinSF2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical 1stFlrSF 2ndFlrSF CudFlrSF		
MasVnrArea ExterQual O ExterCond Foundation O BsmtQual SmtCond SmtExposure SmtFinType1 SmtFinType1 SmtFinSF1 O BsmtFinSF2 O BsmtFinSF2 O BsmtUnfSF TotalBsmtSF O Heating HeatingQC CentralAir Electrical 1stFlrSF OndFlrSF OudFlrSF OudFlrSF OudFlrSF O ContralAir SmtFinSF2 O CentralAir O CentralAir O CentralAir O ContralAir O Contral		
ExterQual 0 ExterCond 0 Foundation 0 BsmtQual 37 BsmtCond 37 BsmtExposure 38 BsmtFinType1 37 BsmtFinSF1 0 BsmtFinSF2 0 BsmtFinSF2 0 BsmtUnfSF 0 TotalBsmtSF 0 Heating 0 HeatingQC 0 CentralAir 0 Electrical 1 1stFlrSF 0 2ndFlrSF 0 CndFlrSF 0 CndFlrSF 0 EndFlrSF 0 ChivArea 0 BsmtFullBath 0 BsmtHalfBath 0 BsmtHalfBath 0 FullBath 0 HalfBath 0 HalfBath 0 FullBath 0 Functional 0 Fireplaces 0 GarageType 81 GarageFinish 81 GarageCars 0 GarageArea 0 GarageArea 0 GarageQual 81	· -	
ExterCond 0 Foundation 0 BsmtQual 37 BsmtCond 37 BsmtExposure 38 BsmtFinType1 37 BsmtFinSF1 0 BsmtFinSF2 38 BsmtFinSF2 0 BsmtUnfSF 0 TotalBsmtSF 0 Heating 0 HeatingQC 0 CentralAir 0 Electrical 1 1stFlrSF 0 2ndFlrSF 0 LowQualFinSF 0 GrLivArea 0 BsmtFullBath 0 BsmtHalfBath 0 FullBath 0 HalfBath 0 HalfBath 0 FullBath 0 F		
Foundation         0           BsmtQual         37           BsmtCond         37           BsmtExposure         38           BsmtFinType1         37           BsmtFinSF1         0           BsmtFinSF2         0           BsmtFinSF2         0           BsmtUnfSF         0           TotalBsmtSF         0           Heating         0           HeatingQC         0           CentralAir         0           Electrical         1           1stFlrSF         0           2ndFlrSF         0           LowQualFinSF         0           GrLivArea         0           BsmtFullBath         0           BsmtHalfBath         0           FullBath         0           KitchenAbvGr         0           KitchenQual         0           TotRmsAbvGrd         0           Fireplaces         0           GarageType         81           GarageFinish         81           GarageArea         0           GarageQual         81	ExterQual	
BsmtQual         37           BsmtCond         37           BsmtExposure         38           BsmtFinType1         37           BsmtFinSF1         0           BsmtFinSF2         38           BsmtFinSF2         0           BsmtUnfSF         0           TotalBsmtSF         0           Heating         0           HeatingQC         0           CentralAir         0           Electrical         1           1stFlrSF         0           2ndFlrSF         0           LowQualFinSF         0           GrLivArea         0           BsmtFullBath         0           BsmtHalfBath         0           FullBath         0           HalfBath         0           BedroomAbvGr         0           KitchenQual         0           TotRmsAbvGrd         0           Functional         0           Fireplaces         0           GarageType         81           GarageYrBlt         81           GarageArea         0           GarageQual         81	ExterCond	0
BsmtCond         37           BsmtExposure         38           BsmtFinType1         37           BsmtFinSF1         0           BsmtFinSF2         38           BsmtFinSF2         0           BsmtUnfSF         0           TotalBsmtSF         0           Heating         0           HeatingQC         0           CentralAir         0           Electrical         1           1stFlrSF         0           2ndFlrSF         0           CrLivArea         0           BsmtFullBath         0           FullBath         0           HalfBath         0           FullBath         0           KitchenAbvGr         0           KitchenQual         0           TotRmsAbvGrd         0           Fireplaces         0           GarageType         81           GarageYrBlt         81           GarageArea         0           GarageQual         81	Foundation	0
BsmtExposure         38           BsmtFinType1         37           BsmtFinSF1         0           BsmtFinSF2         0           BsmtFinSF2         0           BsmtUnfSF         0           TotalBsmtSF         0           Heating         0           HeatingQC         0           CentralAir         0           Electrical         1           1stFlrSF         0           2ndFlrSF         0           CrLivArea         0           BsmtFullBath         0           FullBath         0           FullBath         0           KitchenAbvGr         0           KitchenQual         0           TotRmsAbvGrd         0           Fireplaces         0           GarageType         81           GarageFinish         81           GarageCars         0           GarageQual         81	BsmtQual	37
BsmtFinType1         37           BsmtFinSF1         0           BsmtFinType2         38           BsmtFinSF2         0           BsmtUnfSF         0           TotalBsmtSF         0           Heating         0           HeatingQC         0           CentralAir         0           Electrical         1           1stFlrSF         0           2ndFlrSF         0           LowQualFinSF         0           GrLivArea         0           BsmtFullBath         0           BsmtHalfBath         0           FullBath         0           HalfBath         0           KitchenAbvGr         0           KitchenQual         0           TotRmsAbvGrd         0           Functional         0           Fireplaces         0           GarageType         81           GarageYrBlt         81           GarageArea         0           GarageQual         81	BsmtCond	37
BsmtFinType1         37           BsmtFinSF1         0           BsmtFinType2         38           BsmtFinSF2         0           BsmtUnfSF         0           TotalBsmtSF         0           Heating         0           HeatingQC         0           CentralAir         0           Electrical         1           1stFlrSF         0           2ndFlrSF         0           LowQualFinSF         0           GrLivArea         0           BsmtFullBath         0           BsmtHalfBath         0           FullBath         0           HalfBath         0           KitchenAbvGr         0           KitchenQual         0           TotRmsAbvGrd         0           Functional         0           Fireplaces         0           GarageType         81           GarageYrBlt         81           GarageArea         0           GarageQual         81	BsmtExposure	38
BsmtFinSF1         0           BsmtFinType2         38           BsmtFinSF2         0           BsmtUnfSF         0           TotalBsmtSF         0           Heating         0           HeatingQC         0           CentralAir         0           Electrical         1           1stFlrSF         0           2ndFlrSF         0           CrLivArea         0           BsmtFullBath         0           BsmtHalfBath         0           FullBath         0           HalfBath         0           KitchenAbvGr         0           KitchenQual         0           TotRmsAbvGrd         0           Fireplaces         0           GarageType         81           GarageYrBlt         81           GarageArea         0           GarageQual         81	_	
BsmtFinType2         38           BsmtFinSF2         0           BsmtUnfSF         0           TotalBsmtSF         0           Heating         0           HeatingQC         0           CentralAir         0           Electrical         1           1stFlrSF         0           2ndFlrSF         0           LowQualFinSF         0           GrLivArea         0           BsmtFullBath         0           FullBath         0           HalfBath         0           KitchenAbvGr         0           KitchenQual         0           TotRmsAbvGrd         0           Functional         0           Fireplaces         0           GarageType         81           GarageFinish         81           GarageArea         0           GarageQual         81	· · ·	
BsmtFinSF2 0 BsmtUnfSF 0 TotalBsmtSF 0 Heating 0 HeatingQC 0 CentralAir 0 Electrical 1 1stFlrSF 0 2ndFlrSF 0 LowQualFinSF 0 GrLivArea 0 BsmtFullBath 0 BsmtHalfBath 0 FullBath 0 HalfBath 0 HalfBath 0 HalfBath 0 FullBath 0 GarageType 81 GarageType 81 GarageFinish 81 GarageCars 0 GarageQual 81		
BsmtUnfSF 0 TotalBsmtSF 0 Heating 0 HeatingQC 0 CentralAir 0 Electrical 1 1stFlrSF 0 2ndFlrSF 0 LowQualFinSF 0 GrLivArea 0 BsmtFullBath 0 BsmtHalfBath 0 FullBath 0 HalfBath 0 HalfBath 0 HalfBath 0 FullBath 0 FullBath 0 HalfBath 0 GrichenAbvGr 0 KitchenAbvGr 0 KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageCars 0 GarageArea 0 GarageQual 81	V -	
TotalBsmtSF 0 Heating 0 HeatingQC 0 CentralAir 0 Electrical 1 1stFlrSF 0 2ndFlrSF 0 LowQualFinSF 0 GrLivArea 0 BsmtFullBath 0 BsmtHalfBath 0 FullBath 0 HalfBath 0 HalfBath 0 KitchenAbvGr 0 KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageCars 0 GarageQual 81		
Heating 0 HeatingQC 0 CentralAir 0 Electrical 1 1stFlrSF 0 2ndFlrSF 0 LowQualFinSF 0 GrLivArea 0 BsmtFullBath 0 BsmtHalfBath 0 FullBath 0 HalfBath 0 HalfBath 0 BedroomAbvGr 0 KitchenAbvGr 0 KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageGars 0 GarageQual 81		
HeatingQC 0 CentralAir 0 Electrical 1 1stFlrSF 0 2ndFlrSF 0 LowQualFinSF 0 GrLivArea 0 BsmtFullBath 0 BsmtHalfBath 0 HalfBath 0 HalfBath 0 HalfBath 0 HalfBath 0 FullBath 0 FullBath 0 GrichenQual 0 TotRmsAbvGr 0 KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageCars 0 GarageArea 0 GarageQual 81		
CentralAir         0           Electrical         1           1stFlrSF         0           2ndFlrSF         0           LowQualFinSF         0           GrLivArea         0           BsmtFullBath         0           BsmtHalfBath         0           FullBath         0           HalfBath         0           KitchenAbvGr         0           KitchenQual         0           TotRmsAbvGrd         0           Functional         0           Fireplaces         0           GarageType         81           GarageYrBlt         81           GarageCars         0           GarageArea         0           GarageQual         81	•	
Electrical 1 1stFlrSF 0 2ndFlrSF 0 LowQualFinSF 0 GrLivArea 0 BsmtFullBath 0 BsmtHalfBath 0 HalfBath 0 HalfBath 0 BedroomAbvGr 0 KitchenAbvGr 0 KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageArea 0 GarageQual 81	•	
1stFlrSF 0 2ndFlrSF 0 LowQualFinSF 0 GrLivArea 0 BsmtFullBath 0 BsmtHalfBath 0 FullBath 0 HalfBath 0 BedroomAbvGr 0 KitchenAbvGr 0 KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageCars 0 GarageQual 81		
2ndFlrSF 0 LowQualFinSF 0 GrLivArea 0 BsmtFullBath 0 BsmtHalfBath 0 FullBath 0 HalfBath 0 BedroomAbvGr 0 KitchenAbvGr 0 KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageArea 0 GarageQual 81		
LowQualFinSF 0 GrLivArea 0 BsmtFullBath 0 BsmtHalfBath 0 FullBath 0 HalfBath 0 BedroomAbvGr 0 KitchenAbvGr 0 KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageArea 0 GarageQual 81		
GrLivArea 0 BsmtFullBath 0 BsmtHalfBath 0 FullBath 0 HalfBath 0 BedroomAbvGr 0 KitchenAbvGr 0 KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageCars 0 GarageArea 0 GarageQual 81		
BsmtFullBath BsmtHalfBath OFullBath OHalfBath OBedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces GarageType SarageYrBlt GarageFinish GarageCars GarageArea GarageQual OFunctional OFireplaces OFICTER SAFE SAFE SAFE SAFE SAFE SAFE SAFE SAFE		
BsmtHalfBath 0 FullBath 0 HalfBath 0 BedroomAbvGr 0 KitchenAbvGr 0 KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageArea 0 GarageQual 81	GrLivArea	
FullBath 0 HalfBath 0 BedroomAbvGr 0 KitchenAbvGr 0 KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageCars 0 GarageArea 0 GarageQual 81	BsmtFullBath	
HalfBath 0 BedroomAbvGr 0 KitchenAbvGr 0 KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageCars 0 GarageArea 0 GarageQual 81	BsmtHalfBath	0
BedroomAbvGr 0 KitchenAbvGr 0 KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageCars 0 GarageArea 0 GarageQual 81	FullBath	0
KitchenAbvGr 0 KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageCars 0 GarageArea 0 GarageQual 81	HalfBath	0
KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageCars 0 GarageArea 0 GarageQual 81	BedroomAbvGr	0
KitchenQual 0 TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageCars 0 GarageArea 0 GarageQual 81	KitchenAbvGr	0
TotRmsAbvGrd 0 Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageCars 0 GarageArea 0 GarageQual 81	KitchenQual	
Functional 0 Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageCars 0 GarageArea 0 GarageQual 81	· ·	
Fireplaces 0 GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageCars 0 GarageArea 0 GarageQual 81		
GarageType 81 GarageYrBlt 81 GarageFinish 81 GarageCars 0 GarageArea 0 GarageQual 81		
GarageYrBlt 81 GarageFinish 81 GarageCars 0 GarageArea 0 GarageQual 81	_	
GarageFinish 81 GarageCars 0 GarageArea 0 GarageQual 81		
GarageCars 0 GarageArea 0 GarageQual 81	-	
GarageArea 0 GarageQual 81		
GarageQual 81	_	
_	~	
GarageCond 81	-	
	GarageCond	81

```
PavedDrive
                   0
WoodDeckSF
                   0
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch
                   0
PoolArea
                   0
MiscVal
                   0
MoSold
                   0
YrSold
                   0
SaleType
SaleCondition
                   0
SalePrice
                   0
dtype: int64
```

[91]: data1\_copy\_1[Num\_DF\_missing\_value\_columns].isnull().sum()

# We can see that missing value of LotFrontage column is handled

Similarly, we will handle missing value of other numerical columns columns.

```
[92]: data1_copy2_mean = data1.copy()
data1_copy2_mean.head(10)
```

[92]:		Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	LotShape	LandContour	\
	0	1	60	RL	65.0	8450	Pave	Reg	Lvl	
	1	2	20	RL	80.0	9600	Pave	Reg	Lvl	
	2	3	60	RL	68.0	11250	Pave	IR1	Lvl	
	3	4	70	RL	60.0	9550	Pave	IR1	Lvl	
	4	5	60	RL	84.0	14260	Pave	IR1	Lvl	
	5	6	50	RL	85.0	14115	Pave	IR1	Lvl	
	6	7	20	RL	75.0	10084	Pave	Reg	Lvl	
	7	8	60	RL	NaN	10382	Pave	IR1	Lvl	
	8	9	50	RM	51.0	6120	Pave	Reg	Lvl	

9	10	190	RL	50.0	7420 I	Pave	Reg	Lv	1
	Utilities I	otConfig	LandSlope	Neighborhood	Conditi	ion1 Co	ndition2	BldgType	\
0	AllPub	Inside	Gtl	CollgCr		Norm	Norm	1Fam	
1	AllPub	FR2	Gtl	Veenker		eedr	Norm	1Fam	
2	AllPub	Inside	Gtl	CollgCr		Norm	Norm	1Fam	
3	AllPub	Corner	Gtl	Crawfor		Norm	Norm	1Fam	
4	AllPub	FR2	Gtl	NoRidge		Norm	Norm	1Fam	
5	AllPub	Inside	Gt1	Mitchel		Norm	Norm	1Fam	
6	AllPub	Inside	Gt1	Somerst		Norm	Norm	1Fam	
7	AllPub	Corner	Gt1	NWAmes		PosN	Norm	1Fam	
8	AllPub	Inside	Gtl	OldTown		tery	Norm	1Fam	
9	AllPub	Corner	Gtl	BrkSide	AF	tery	Artery	2fmCon	
	HouseStyle	OverallQ	ual Over	allCond Year		YearRem	odAdd Roo	fStyle	\
0	2Story		7	5	2003		2003	Gable	
1	1Story		6	8	1976		1976	Gable	
2	2Story		7	5	2001		2002	Gable	
3	2Story		7	5	1915		1970	Gable	
4	2Story		8	5	2000		2000	Gable	
5	1.5Fin		5	5	1993		1995	Gable	
6	1Story		8	5	2004		2005	Gable	
7	2Story		7	6	1973		1973	Gable	
8	1.5Fin		7	5	1931		1950	Gable	
9	1.5Unf		5	6	1939		1950	Gable	
	RoofMatl Ex	terior1st	Exterior	2nd MasVnrTyp	e MasVr	nrArea 1	ExterQual	ExterCo	nd \
0	CompShg	VinylSd	•		e	196.0	Gd		TA
1	CompShg	MetalSd				0.0	TA		TA
2	CompShg	VinylSd	•		e	162.0	Gd		TA
3	CompShg	Wd Sdng		-		0.0	TA		TA
4	CompShg	VinylSd	•			350.0	Gd		TA
5	CompShg	VinylSd	v			0.0	TA		TA
6	CompShg	VinylSd	•			186.0	Gd		TA
7	CompShg	HdBoard				240.0	TA		TA
8	CompShg	BrkFace		•		0.0	TA		TA
9	CompShg	MetalSd	Meta.	lSd Non	ıe	0.0	TA		TA
	Foundation	BsmtQual	BsmtCond 1	BsmtExposure	BsmtFin7	Гуре1	BsmtFinSF		
0	PConc	Gd	TA	No		GLQ	70	6	
1	CBlock	Gd	TA	Gd		ALQ	97		
2	PConc	Gd	TA	Mn		GLQ	48		
3	BrkTil	TA	Gd	No		ALQ	21		
4	PConc	Gd	TA	Av		GLQ	65	5	
5	Wood	Gd	TA	No		GLQ	73	2	
6	PConc	Ex	TA	Av		GLQ	136		
7	CBlock	Gd	TA	Mn		ALQ	85	9	

8	BrkTil	TA	TA		No		Unf		0			
9	BrkTil	TA	TA		No		GLQ		851			
	BsmtFinType2	BsmtFinSF2	BsmtUn	fSF 7	Total	BsmtSF	Heating	Heati	ngQC	\		
0	Unf	0		150		856	${\tt GasA}$		Ex			
1	Unf	0		284		1262	GasA		Ex			
2	Unf	0		434		920	GasA		Ex			
3	Unf	0		540		756	GasA		Gd			
4	Unf	0		490		1145	GasA		Ex			
5	Unf	0		64		796	GasA		Ex			
6	Unf	0		317		1686	GasA		Ex			
7	BLQ	32		216		1107	GasA		Ex			
8	Unf	0		952		952	GasA		Gd			
9	Unf	0		140		991	GasA		Ex			
	CentralAir El	ectrical 1st	tFlrSF	2ndF	lrSF	LowQua	alFinSF	GrLiv	Area	\		
0	Y	SBrkr	856		854		0		1710	•		
1	Y	SBrkr	1262		0		0		1262			
2	Y	SBrkr	920		866		0		1786			
3	Y	SBrkr	961		756		0		1717			
4	Y	SBrkr	1145		1053		0		2198			
5	Y	SBrkr	796	•	566		0		1362			
6	Y	SBrkr	1694		0		0		1694			
7	Y	SBrkr	1107		983		0		2090			
8	Y	FuseF	1022		752		0		1774			
9	Y	SBrkr	1077		0		0		1077			
J	1	DDIKI	1011		O		O		1011			
	BsmtFullBath	BsmtHalfBa	th Ful	1Bath	Hal	fRath	Redroom	hvGr	Kitc	henAhv	Gr	١
0	1		0	2	mar	1	Dodioom	3	11100	110111101	1	`
1	0		1	2		0		3			1	
2	1		0	2		1		3			1	
3	1		0	1		0		3			1	
4	1		0	2		1		4			1	
5	1		0	1		1		1			1	
6	1		0	2		0		3			1	
7	1		0	2		1		3			1	
8	0		0	2		0		2			2	
9	1		0	1		0		2			2	
9	1		O	1		O		2			_	
	KitchenQual	TotRmsAbvGrd	Functi	onal	Fire	nlaces	GarageTy	me G	arado	YrBlt	\	
0	Gd	B 18 VGA A IIII 10 01	runcti		LILE	praces 0	Atto	-	_	003.0	`	
	TA	6		Тур		1	Atto			976.0		
1 2	Gd	6		Тур		1	Atto			001.0		
3	Gd Gd	7		Тур		1	Deto			998.0		
				Тур								
4	Gd TA	9		Тур		1	Atto			0.000		
5	TA	5		Тур		0	Atto			993.0		
6	Gd	7		Тур		1	Atto	ли	2	004.0		

```
7
             TA
                              7
                                         Тур
                                                          2
                                                                 Attchd
                                                                                1973.0
8
             ΤA
                              8
                                        Min1
                                                          2
                                                                                1931.0
                                                                 Detchd
9
             TA
                              5
                                                          2
                                         Тур
                                                                 Attchd
                                                                                1939.0
  GarageFinish
                  GarageCars
                                 GarageArea GarageQual GarageCond PavedDrive
                                                       TA
0
             RFn
                                         548
                                                                    TA
                                                                                   Y
                             2
                                                                                   Y
1
             R.Fn
                                         460
                                                       ТΑ
                                                                    TΑ
2
             RFn
                             2
                                                       TA
                                                                    TA
                                                                                   Y
                                         608
3
                             3
                                                                                   Y
             Unf
                                         642
                                                       TA
                                                                    TA
4
             RFn
                             3
                                         836
                                                       TA
                                                                    TA
                                                                                   Y
                             2
5
             Unf
                                                       TA
                                                                    TA
                                                                                   Y
                                         480
6
             RFn
                             2
                                         636
                                                       TA
                                                                    TA
                                                                                   Y
7
                             2
             RFn
                                         484
                                                       TA
                                                                    TA
                                                                                   Y
                                                                                   Y
8
             Unf
                             2
                                         468
                                                       Fa
                                                                    TA
9
             RFn
                             1
                                         205
                                                       Gd
                                                                    TA
                                                                                   Y
   WoodDeckSF
                 OpenPorchSF
                                                   3SsnPorch
                                                               ScreenPorch
                                                                              PoolArea
                                 EnclosedPorch
0
              0
                            61
                                               0
                                                            0
                                                                            0
1
            298
                             0
                                               0
                                                            0
                                                                            0
                                                                                       0
                            42
                                               0
                                                            0
2
              0
                                                                            0
                                                                                       0
3
              0
                            35
                                             272
                                                            0
                                                                            0
                                                                                       0
            192
4
                            84
                                               0
                                                            0
                                                                            0
                                                                                       0
5
             40
                            30
                                               0
                                                          320
                                                                            0
                                                                                       0
            255
                                               0
6
                            57
                                                            0
                                                                            0
                                                                                       0
7
            235
                           204
                                             228
                                                            0
                                                                            0
                                                                                       0
                                                            0
8
             90
                             0
                                             205
                                                                            0
                                                                                       0
              0
                             4
9
                                               0
                                                            0
                                                                                       0
   MiscVal
              MoSold
                       YrSold SaleType SaleCondition
                                                            SalePrice
0
          0
                    2
                          2008
                                       WD
                                                   Normal
                                                                208500
1
          0
                    5
                          2007
                                       WD
                                                   Normal
                                                                181500
2
          0
                    9
                          2008
                                       WD
                                                   Normal
                                                                223500
3
          0
                    2
                          2006
                                       WD
                                                 Abnorml
                                                                140000
4
          0
                   12
                          2008
                                                   Normal
                                       WD
                                                                250000
5
        700
                   10
                          2009
                                       WD
                                                   Normal
                                                                143000
6
          0
                    8
                          2007
                                       WD
                                                  Normal
                                                               307000
7
        350
                          2009
                                       WD
                                                  Normal
                                                               200000
                   11
8
          0
                    4
                          2008
                                       WD
                                                 Abnorml
                                                                129900
9
          0
                    1
                          2008
                                       WD
                                                   Normal
                                                                118000
```

### 1.2 Mean:

```
[95]: num_vars_miss = ['LotFrontage', 'MasVnrArea', 'GarageYrBlt']
    cat_vars = ['LotConfig', 'Exterior2nd', 'KitchenQual']

for cat_var, num_var_miss in zip(cat_vars,num_vars_miss):
    for var_class in data[cat_var].unique():
```

```
data1_copy2_mean.update(data[data.loc[:,cat_var] ==_u

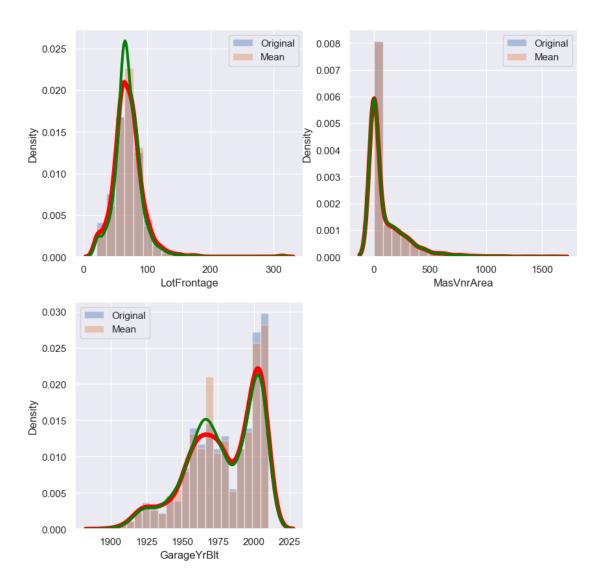
-var_class][num_var_miss].replace(np.nan,data[data.loc[:,cat_var] ==_u

-var_class][num_var_miss].mean()))
```

```
[97]: data1_copy2_mean[num_vars_miss].isnull().sum()
```

Here, we have addressed the issue of missing values in the numerical column by calculating the mean for each category in the categorical column.

#### 1.2.1 Data Distribution:



## 1.3 Median:

[100]: data1\_copy2\_median = data1.copy()
 data1\_copy2\_median.head()

[100]:		Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	LotShape	LandContour	\
	0	1	60	RL	65.0	8450	Pave	Reg	Lvl	
	1	2	20	RL	80.0	9600	Pave	Reg	Lvl	
	2	3	60	RL	68.0	11250	Pave	IR1	Lvl	
	3	4	70	RL	60.0	9550	Pave	IR1	Lvl	
	4	5	60	R.I.	84.0	14260	Pave	TR1	I.v.l	

Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType \
O AllPub Inside Gtl CollgCr Norm Norm 1Fam

1	AllPub	FR2	Gtl	Vee	enker	]	Feedr	No	rm	1Fa	m	
2	AllPub	Inside	Gtl	Col	llgCr		Norm	No	rm	1Fa	m	
3	AllPub	Corner	Gtl	Cra	awfor		Norm	No	rm	1Fa	m	
4	AllPub	FR2	Gtl	NoF	Ridge		Norm	No	rm	1Fa	.m	
	HouseStyle	OverallQua	l Overal	lCond	YearB	uilt	YearRem	nodAdd	Roof	Style	\	
0	2Story		7	5	:	2003		2003	(	Gable		
1	1Story		6	8		1976		1976	(	Gable		
2	2Story		7	5	:	2001		2002	(	Gable		
3	2Story		7	5		1915		1970	(	Gable		
4	2Story		8	5	:	2000		2000	(	Gable		
	RoofMatl Ext						VnrArea	ExterQ		ExterC		\
0	CompShg	VinylSd	VinylS		kFace		196.0		Gd		TA	
1	CompShg	MetalSd	MetalS		None		0.0		TA		TA	
2	CompShg	VinylSd	VinylS	d Br	kFace		162.0		Gd		TΑ	
3	CompShg	Wd Sdng	Wd Shn	g	None		0.0		TA		TA	
4	CompShg	VinylSd	VinylS	d Br	kFace		350.0		Gd		TA	
				. =	_	. = .			a=.	,		
_	Foundation B			mtExpos		smtFi		BsmtFi		\		
0	PConc	Gd	TA		No		GLQ		706			
1	CBlock	Gd	TA		Gd		ALQ		978			
2	PConc	Gd	TA		Mn		GLQ		486			
3	${\tt BrkTil}$	TA	Gd		No		ALQ		216			
4	PConc	Gd	TA		Av		GLQ		655			
	D . H. H. C	. D . E . G	FO D	can n		. 00			0.0	,		
^	BsmtFinType2				otalB		Heating		_			
0	Unf		0	150		856	GasA		Ex			
1	Unf		0	284		1262	GasA		Ex			
2	Unf		0	434		920	Gas <i>A</i>		Ex			
3	Unf		0	540		756	GasA		Gd			
4	Unf	•	0	490		1145	Gas <i>A</i>	1	Ex			
	CentralAir E	'loc+ricol	1stFlrSF	OndEl	~CE i	T 011011	alFinSF	GrLiv	1200	\		
^	Y		15tr115r 856		854	LOWQU	0			`		
0		SBrkr							1710			
1	Y	SBrkr	1262		0		0		1262			
2	Y	SBrkr	920		866		0		1786			
3	Y	SBrkr	961		756		0		1717			
4	Y	SBrkr	1145	1	1053		0		2198			
	BsmtFullBat	h BsmtHal	fRath Fu	11Bath	Half	Rath	Bedroom	. Δ byGr	Ki+	chenAb	wGr	\
0		1	0	2		1		3			1	`
1		0	1	2		0		3			1	
2		1	0	2		1		3			1	
3		1	0	1		0		3			1	
3 4			0	2		1		3 4			1	
4		1	U	2		T		4			Т	

```
0
                   Gd
                                   8
                                                            0
                                                                   Attchd
                                                                                 2003.0
                                             Тур
                                   6
                                                            1
                                                                                 1976.0
       1
                   TA
                                             Тур
                                                                   Attchd
       2
                   Gd
                                   6
                                                                   Attchd
                                                                                 2001.0
                                             Typ
       3
                   Gd
                                   7
                                                            1
                                                                   Detchd
                                                                                 1998.0
                                             Тур
                   Gd
                                                                                 2000.0
                                   9
                                                            1
                                                                   Attchd
                                             Тур
                                     GarageArea GarageQual GarageCond PavedDrive
         GarageFinish
                        GarageCars
       0
                   R.Fn
                                             548
                                                          TA
                                                                      TΑ
       1
                   RFn
                                  2
                                             460
                                                          ТΑ
                                                                      TA
                                                                                   Y
       2
                   RFn
                                  2
                                                          TΑ
                                                                      TΑ
                                                                                   Y
                                             608
       3
                   Unf
                                  3
                                             642
                                                          ТΑ
                                                                      ТΑ
                                                                                   Y
                                                                                   γ
                   RFn
                                  3
                                             836
                                                          TΑ
                                                                      TA
          WoodDeckSF
                       OpenPorchSF
                                     EnclosedPorch
                                                     3SsnPorch
                                                                 ScreenPorch
                                                                               PoolArea
       0
                                 61
                  298
                                                              0
                                                                            0
                                  0
                                                  0
                                                                                       0
       1
       2
                    0
                                 42
                                                  0
                                                              0
                                                                            0
                                                                                       0
       3
                    0
                                 35
                                                272
                                                              0
                                                                            0
                                                                                       0
                  192
       4
                                 84
                                                  0
          MiscVal MoSold
                            YrSold SaleType SaleCondition SalePrice
       0
                 0
                         2
                               2008
                                           WD
                                                      Normal
                                                                  208500
                 0
                         5
                               2007
                                                      Normal
       1
                                           WD
                                                                  181500
       2
                               2008
                                                      Normal
                 0
                         9
                                           WD
                                                                  223500
                 0
                         2
                                                     Abnorml
       3
                               2006
                                           WD
                                                                  140000
                               2008
                                                      Normal
                 0
                        12
                                           WD
                                                                  250000
[101]: data1_copy2_median[num_vars_miss].isnull().sum()
[101]: LotFrontage
                       259
       MasVnrArea
                         8
       GarageYrBlt
                        81
       dtype: int64
[102]: num_vars_miss = ['LotFrontage', 'MasVnrArea', 'GarageYrBlt']
       cat_vars = ['LotConfig','Exterior2nd','KitchenQual']
       for cat var, num var miss in zip(cat vars, num vars miss):
           for var_class in data[cat_var].unique():
                data1_copy2_median.update(data[data.loc[:,cat_var] ==_
        Govar_class] [num_var_miss].replace(np.nan,data[data.loc[:,cat_var] ==__
         →var_class][num_var_miss].median()))
[103]: data1_copy2_median[num_vars_miss].isnull().sum()
```

TotRmsAbvGrd Functional Fireplaces GarageType

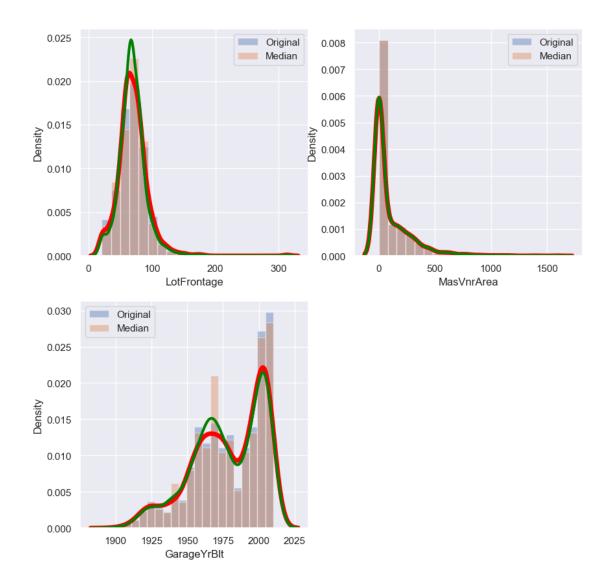
GarageYrBlt \

KitchenQual

```
[103]: LotFrontage 0
MasVnrArea 0
GarageYrBlt 0
dtype: int64
```

Here, we have addressed the issue of missing values in the numerical column by calculating the mean for each category in the categorical column.

### 1.3.1 Data Distribution:



# tegorical-missing-value-imputation

June 11, 2023

## 1 Data Cleaning:

### 1.1 Categorical Missing value imputation:

Importing necessary libraries:

```
[1]: import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     import warnings
     warnings.filterwarnings('ignore')
     from IPython.display import Image
    data (Original):
[2]: data = pd.read_csv('train.csv')
[3]: data.shape # Checking the shape of the DataFrame, DataFrame contains Rows:
      →1460 and Columns: 81.
[3]: (1460, 81)
[4]: data.head(2)
                   # Checking first 2 rows from the DataFrame.
[4]:
            MSSubClass MSZoning
                                 LotFrontage
                                                LotArea Street Alley LotShape
                                          65.0
     0
                     60
                              RL
                                                   8450
                                                          Pave
                                                                  NaN
                                                                           Reg
     1
                     20
                              RL
                                          80.0
                                                   9600
                                                          Pave
                                                                  NaN
                                                                           Reg
       LandContour Utilities
                              ... PoolArea PoolQC Fence MiscFeature MiscVal MoSold
                                         0
                                              NaN
                                                                           0
                                                                                   2
     0
               Lvl
                       AllPub
                                                    NaN
                                                                 {\tt NaN}
                                                                                   5
     1
               Lvl
                       AllPub
                                         0
                                              NaN
                                                    NaN
                                                                 {\tt NaN}
                                                                           0
                          SaleCondition SalePrice
       YrSold
               SaleType
         2008
                      WD
                                 Normal
                                             208500
     1
         2007
                     WD
                                 Normal
                                             181500
     [2 rows x 81 columns]
```

When we check the shape of the data using the command "data.shape", we can observe that the DataFrame consists of 1460 rows and 81 columns. However, when we examine the first two rows of the DataFrame using the command "data.head(2)", it doesn't display all 81 columns. As a result, we utilize the following code, specifically "pd.set\_option()", to address this issue.

```
[5]: pd.set_option('display.max_column', None)
pd.set_option('display.max_rows', None)
```

pd.set\_option('display.max\_columns', None): This line sets the maximum number of columns to be displayed in the output to None, which means there is no limit. As a result, all columns in a DataFrame will be shown when you print or display it.

pd.set\_option('display.max\_rows', None): This line sets the maximum number of rows to be displayed in the output to None, removing any limit. As a result, all rows in a DataFrame will be shown when you print or display it.

```
[8]: categorical_DataFrame.head() # Checking first 5 rows of categorical DataFrame.
```

[8]:	MSZoning	Street	Alley	LotShape	${\tt LandContour}$	Utilities	LotConfig	LandSlope	١
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	
	Neighborh	nood Cor	ndition	1 Condit	ion2 BldgTyne	e HouseStv	le RoofStvl	e RoofMatl	

	Merginoriiood	Conditioni	Conditions	pragrybe	Housestyle	ROOISTYLE	ROOTMati	
0	CollgCr	Norm	Norm	1Fam	2Story	Gable	CompShg	
1	Veenker	Feedr	Norm	1Fam	1Story	Gable	CompShg	
2	CollgCr	Norm	Norm	1Fam	2Story	Gable	CompShg	
3	Crawfor	Norm	Norm	1Fam	2Story	Gable	CompShg	
4	NoRidge	Norm	Norm	1Fam	2Story	Gable	CompShg	

\

	Exterior1st	Exterior2nd	MasVnrType	ExterQual	ExterCond	Foundation	BsmtQual
0	VinylSd	VinylSd	${\tt BrkFace}$	Gd	TA	PConc	Gd
1	MetalSd	MetalSd	None	TA	TA	CBlock	Gd
2	VinylSd	VinylSd	${\tt BrkFace}$	Gd	TA	PConc	Gd
3	Wd Sdng	Wd Shng	None	TA	TA	${\tt BrkTil}$	TA
4	VinylSd	VinylSd	BrkFace	Gd	TA	PConc	Gd

	BsmtCond	${\tt BsmtExposure}$	BsmtFinType1	BsmtFinType2	Heating	${\tt HeatingQC}$	\
0	TA	No	GLQ	Unf	${\tt GasA}$	Ex	
1	TA	Gd	ALQ	Unf	${\tt GasA}$	Ex	
2	TA	Mn	GLQ	Unf	${\tt GasA}$	Ex	
3	Gd	No	ALQ	Unf	${\tt GasA}$	Gd	
4	TA	Av	GLQ	Unf	GasA	Ex	

```
CentralAir Electrical KitchenQual Functional FireplaceQu GarageType \
0
           Y
                   SBrkr
                                   Gd
                                              Тур
                                                           NaN
                                                                    Attchd
           Y
                   SBrkr
                                   TA
                                                            TΑ
                                                                    Attchd
1
                                              Тур
2
           Υ
                   SBrkr
                                   Gd
                                              Тур
                                                            TΑ
                                                                    Attchd
3
           Y
                   SBrkr
                                   Gd
                                              Тур
                                                            Gd
                                                                    Detchd
           Υ
4
                   SBrkr
                                   Gd
                                              Тур
                                                            TA
                                                                    Attchd
 GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature \
0
           RFn
                        TA
                                    TA
                                                 Y
                                                       NaN
                                                             NaN
                                                                          NaN
           RFn
                        TA
                                    TA
                                                 Y
                                                       NaN
                                                             NaN
1
                                                                          NaN
2
           RFn
                        TA
                                    TA
                                                 Y
                                                       NaN
                                                             NaN
                                                                          NaN
           Unf
                                                 Y
                                                             NaN
                                                                          NaN
3
                        TA
                                    TA
                                                       NaN
4
           RFn
                                    TA
                                                 Y
                        TΑ
                                                       NaN
                                                             {\tt NaN}
                                                                          NaN
  SaleType SaleCondition
0
                   Normal
        WD
1
        WD
                   Normal
2
        WD
                   Normal
```

## [9]: categorical\_DataFrame.isnull().sum()

WD

WD

Abnorml

Normal

# It used to calculate the number of missing values in each column of  $a_{\square}$   $\rightarrow$  DataFrame called categorical\_DataFrame.

[9]: MSZoning 0 Street 0 1369 Alley LotShape 0 LandContour 0 Utilities 0 0 LotConfig LandSlope 0 Neighborhood 0 Condition1 0 Condition2 0 BldgType 0 0 HouseStyle RoofStyle 0 0 RoofMatl Exterior1st 0 0 Exterior2nd MasVnrType 8 ExterQual 0 ExterCond 0

3

4

```
Foundation
                     0
BsmtQual
                    37
BsmtCond
                    37
BsmtExposure
                    38
BsmtFinType1
                    37
BsmtFinType2
                    38
                     0
Heating
HeatingQC
                     0
CentralAir
                     0
Electrical
                     1
KitchenQual
                     0
Functional
                     0
FireplaceQu
                   690
GarageType
                    81
GarageFinish
                    81
GarageQual
                    81
GarageCond
                    81
PavedDrive
                     0
PoolQC
                  1453
                  1179
Fence
MiscFeature
                  1406
                     0
SaleType
SaleCondition
                     0
dtype: int64
```

[11]: missing\_value\_percent = categorical\_DataFrame.isnull().mean()\*100
missing\_value\_percent

#calculates the percentage of missing values in each column of the  $\$   $\hookrightarrow$  categorical\_DataFrame.

[11]: MSZoning 0.000000 Street 0.000000 Alley 93.767123 LotShape 0.000000 LandContour 0.000000 Utilities 0.000000 0.000000 LotConfig 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

```
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
[15]: drop_column = missing_value_percent[missing_value_percent > 20].keys()
      print(drop_column)
      111
      The code `drop_column = missing_value_percent[missing_value_percent > 20].
       ⇔keys()` selects the columns from
      `missing_value_percent` that have missing value percentages greater than 20\%
       ⇔and assigns them to the `drop_column` variable.
      111
     Index(['Alley', 'FireplaceQu', 'PoolQC', 'Fence', 'MiscFeature'],
     dtype='object')
[21]: new_data = categorical_DataFrame.drop(columns = drop_column)
      new_data.head()
```

Exterior2nd

0.000000

```
The code new data = categorical DataFrame.drop(columns=drop_column) creates a_1
       →new DataFrame called new_data by dropping the
      columns specified in the drop_column variable from the categorical_DataFrame
       111
        MSZoning Street LotShape LandContour Utilities LotConfig LandSlope
[21]:
               RL
                    Pave
                                            Lvl
                                                    AllPub
                                                               Inside
                               Reg
      1
               R.T.
                    Pave
                               Reg
                                            T.v.T
                                                    All Pub
                                                                  FR2
                                                                             Gt.1
      2
               R.T.
                                                    AllPub
                                                               Inside
                    Pave
                               IR1
                                            Lvl
                                                                             Gt.1
      3
               RL
                    Pave
                               IR1
                                            Lvl
                                                    AllPub
                                                               Corner
                                                                             Gtl
                                                                  FR2
      4
               RL
                    Pave
                               IR1
                                            Lvl
                                                    AllPub
                                                                             Gtl
        Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl
      0
              CollgCr
                             Norm
                                         Norm
                                                   1Fam
                                                             2Story
                                                                         Gable
                                                                                CompShg
      1
              Veenker
                            Feedr
                                         Norm
                                                   1Fam
                                                             1Story
                                                                         Gable
                                                                                CompShg
      2
              CollgCr
                             Norm
                                         Norm
                                                   1Fam
                                                             2Story
                                                                         Gable
                                                                                CompShg
      3
              Crawfor
                             Norm
                                         Norm
                                                   1Fam
                                                             2Story
                                                                         Gable
                                                                                CompShg
      4
              NoRidge
                             Norm
                                                             2Story
                                                                         Gable CompShg
                                         Norm
                                                   1Fam
        Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual
      0
             VinylSd
                          VinylSd
                                      BrkFace
                                                      Gd
                                                                 TA
                                                                          PConc
      1
             MetalSd
                          MetalSd
                                         None
                                                      TA
                                                                 TA
                                                                         CBlock
                                                                                       Gd
      2
            VinylSd
                          VinylSd
                                      BrkFace
                                                      Gd
                                                                 TΑ
                                                                          PConc
                                                                                       Gd
      3
             Wd Sdng
                          Wd Shng
                                         None
                                                      TA
                                                                 TΑ
                                                                         BrkTil
                                                                                       ТΑ
      4
             VinylSd
                          VinylSd
                                      BrkFace
                                                      Gd
                                                                 TA
                                                                          PConc
                                                                                       Gd
        BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC
      0
               TA
                             No
                                          GLQ
                                                        Unf
                                                                GasA
                                                                             Ex
               TA
                                                        Unf
      1
                             Gd
                                          ALQ
                                                                GasA
                                                                             Ex
      2
               TA
                             Mn
                                          GLQ
                                                        Unf
                                                                GasA
                                                                             Ex
      3
               Gd
                             No
                                          ALQ
                                                        Unf
                                                                GasA
                                                                             Gd
               TA
                                          GLQ
                                                        Unf
                                                                GasA
                                                                             Ex
                             Αv
        CentralAir Electrical KitchenQual Functional GarageType GarageFinish
      0
                  Y
                          SBrkr
                                          Gd
                                                              Attchd
                                                                               RFn
                                                     Тур
      1
                  Y
                          SBrkr
                                          TA
                                                              Attchd
                                                                               RFn
                                                     Тур
      2
                  Y
                          SBrkr
                                          Gd
                                                              Attchd
                                                                               RFn
                                                     Тур
      3
                  Y
                          SBrkr
                                                              Detchd
                                                                               Unf
                                          Gd
                                                     Тур
      4
                  Υ
                          SBrkr
                                          Gd
                                                     Тур
                                                              Attchd
                                                                               RFn
        GarageQual GarageCond PavedDrive SaleType SaleCondition
      0
                 TA
                             TA
                                          Y
                                                   WD
                                                              Normal
      1
                 TA
                             TA
                                          Y
                                                   WD
                                                              Normal
      2
                 TA
                             TA
                                          Y
                                                   WD
                                                              Normal
```

111

```
[24]: Missing_val_columns = new_data.isnull().sum()
print(Missing_val_columns)

The code Missing_val_columns = new_data.isnull().sum() calculates the number of_

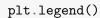
missing values in each column of the
new_data DataFrame and assigns the result to the Missing_val_columns variable.
```

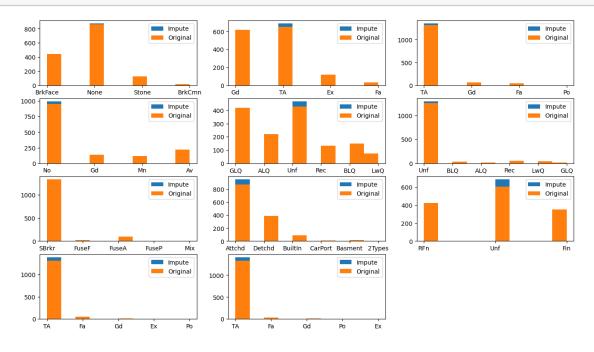
MSZoning 0 Street 0 LotShape 0 LandContour 0 0 Utilities 0 LotConfig 0 LandSlope Neighborhood 0 0 Condition1 Condition2 0 BldgType 0 0 HouseStyle RoofStyle 0 0 RoofMatl Exterior1st 0 Exterior2nd 0 8 MasVnrType ExterQual 0 ExterCond 0 Foundation 0 BsmtQual 37 **BsmtCond** 37 BsmtExposure 38 37 BsmtFinType1 BsmtFinType2 38 0 Heating 0 HeatingQC 0 CentralAir Electrical 1 0 KitchenQual Functional 0 GarageType 81 GarageFinish 81 81 GarageQual GarageCond 81

```
PavedDrive
                        0
     SaleType
                        0
     SaleCondition
                        0
     dtype: int64
[26]: Null_column = Missing_val_columns[Missing_val_columns > 0].keys()
      print(Null_column)
      111
      The code `Null_column = Missing_val_columns[Missing_val_columns > 0].keys()`_
       ⇔selects the columns from the
      `Missing_val_columns` variable that have a count of missing values greater than \sqcup
       ⇔0 and assigns their column names to the
      `Null_column` variable.
      ,,,
     Index(['MasVnrType', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1',
             'BsmtFinType2', 'Electrical', 'GarageType', 'GarageFinish',
             'GarageQual', 'GarageCond'],
           dtype='object')
[27]: new_data_copy = new_data.copy()
      new_data_copy.shape
[27]: (1460, 38)
[28]: for var in Null_column:
          new_data_copy[var].fillna(new_data_copy[var].mode()[0], inplace = True)
          print(var, '=',new_data[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
```

This loop iterates over each column name stored in Null\_column, fills the missing values in that column with the mode (most frequent value), and then prints the column name along with the mode value. The fillna function is used to replace the missing values with the specified mode value in place.

```
[30]: new_data_copy.isnull().sum() # We can see null value is handled
                        0
[30]: MSZoning
      Street
                        0
                        0
      LotShape
      LandContour
                        0
      Utilities
                        0
      LotConfig
                        0
      LandSlope
                        0
      Neighborhood
                        0
      Condition1
                        0
      Condition2
                        0
                        0
      BldgType
      HouseStyle
                        0
      RoofStyle
                        0
      RoofMatl
                        0
      Exterior1st
                        0
      Exterior2nd
                        0
      MasVnrType
                        0
      ExterQual
                        0
      ExterCond
                        0
      Foundation
                        0
      BsmtQual
      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
      dtype: int64
[31]: plt.figure(figsize=(16,9))
      for i,var in enumerate(Null_column):
          plt.subplot(4,3,i+1)
          plt.hist(new_data_copy[var],label="Impute")
          plt.hist(new_data[var].dropna(),label="Original")
```





The code you provided is creating a figure and subplots using plt.subplot() to visualize the distribution of the imputed values (new\_data\_copy[var]) compared to the original values (new\_data[var]) for each column in Null\_column.