Data cleaing

Numerical Missing values Imputation By class

In [3]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

In [89]:
```
df=pd.read_csv("train.csv")
df.head()
```

Out[89]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborh
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	Coll
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	FR2	Gtl	Veer
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	Coll
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	Corner	Gtl	Crav
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl	NoRi

In [8]:
```
df.shape
```

Out[8]:
```
(1460, 81)
```

In [9]:
```
pd.set_option("display.max_columns",None)
pd.set_option("display.max_rows",None)
```

In [11]:
```
df.head()
```

Out[11]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborh
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	Coll
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	FR2	Gtl	Veer
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	Coll
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	Corner	Gtl	Crav
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl	NoRi

In [13]:
```
df.info()
```

...

In [15]:
```
df.isnull().sum()
```

Out[15]:
```
Id                      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              0
Condition2              0
BldgType                0
HouseStyle              0
OverallQual             0
OverallCond             0
YearBuilt               0
```

In [17]:
```
df.isnull().sum().sum()
```

Out[17]:
```
6965
```

In [19]:
```
null_var=df.isnull().sum()/df.shape[0]*100
null_var
```

Out[19]:
```
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
```

In [21]:
```
drop_cloumn = null_var[null_var >20].keys()
drop_cloumn
```

Out[21]:
```
Index(['Alley', 'FireplaceQu', 'PoolQC', 'Fence', 'MiscFeature'], dtype='object')
```

In [27]:
```
df2=df.drop(columns=drop_cloumn)
```

In [28]:
```
df2.shape
```

Out[28]:
```
(1460, 76)
```

In [33]:
```
df3_num=df2.select_dtypes(include=["int64","float64"])
```

In [35]:
```
df3_num.head()
```

Out[35]:

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	BsmtFinSF2
0	1	60	65.0	8450	7	5	2003	2003	196.0	706	0
1	2	20	80.0	9600	6	8	1976	1976	0.0	978	0
2	3	60	68.0	11250	7	5	2001	2002	162.0	486	0
3	4	70	60.0	9550	7	5	1915	1970	0.0	216	0
4	5	60	84.0	14260	8	5	2000	2000	350.0	655	0

In [38]:
```
df3_num.isnull().sum()
```

Out[38]:
```
Id                 0
MSSubClass         0
LotFrontage      259
LotArea            0
OverallQual        0
OverallCond        0
YearBuilt          0
YearRemodAdd       0
MasVnrArea         8
BsmtFinSF1         0
BsmtFinSF2         0
BsmtUnfSF          0
TotalBsmtSF        0
1stFlrSF           0
2ndFlrSF           0
LowQualFinSF       0
GrLivArea          0
BsmtFullBath       0
BsmtHalfBath       0
FullBath           0
```

In [45]:

```
num_var_miss = ['LotFrontage','MasVnrArea','GarageYrBlt']
df3_num[num_var_miss][df3_num[num_var_miss].isnull().any(axis=1)]
```

Out[45]:

	LotFrontage	MasVnrArea	GarageYrBlt
7	NaN	240.0	1973.0
12	NaN	0.0	1962.0
14	NaN	212.0	1960.0
16	NaN	180.0	1970.0
24	NaN	0.0	1968.0
31	NaN	0.0	1966.0
39	65.0	0.0	NaN
42	NaN	0.0	1983.0
43	NaN	0.0	1977.0
48	33.0	0.0	NaN

In [94]:

```
df["LotConfig"].unique()
```

Out[94]:

```
array(['Inside', 'FR2', 'Corner', 'CulDSac', 'FR3'], dtype=object)
```

In [95]:
```
df[df.loc[:,"LotConfig"] == "Inside"]["LotFrontage"].replace(np.nan,df[df.loc[:,"LotConfig"] == "Inside"]["
```

Out[95]:
```
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
18        66.000000
19        70.000000
21        57.000000
22        75.000000
23        44.000000
24        67.715686
27        98.000000
29        60.000000
30        50.000000
```

In [87]:
```
df_copy = df.copy()
for var_class in df['LotConfig'].unique():
    df_copy.update(df[df.loc[:,'LotConfig'] == var_class]["LotFrontage"].replace(np.nan,df[df.loc[:,'LotCon
```

In [88]:
```
df_copy.isnull().sum()
```

Out[88]:
```
Id                 0
MSSubClass         0
MSZoning           0
LotFrontage        0
LotArea            0
Street             0
Alley           1369
LotShape           0
LandContour        0
Utilities          0
LotConfig          0
LandSlope          0
Neighborhood       0
Condition1         0
Condition2         0
BldgType           0
HouseStyle         0
OverallQual        0
OverallCond        0
YearBuilt          0
```

In [121]:
```
df_copy = df.copy()
num_vars_miss = ['LotFrontage','MasVnrArea','GarageYrBlt']
cat_vars=['LotConfig','MasVnrType','GarageType']
for cat_vars ,null_var_miss in zip(cat_vars,num_vars_miss):

    for var_class in df[cat_vars].unique():
        df_copy.update(df[df.loc[:,cat_vars] == var_class][num_var_miss].replace(np.nan,df[df.loc[:,cat_var
```

In [124]:
```
df_copy[num_vars_miss].isnull().sum()
```

Out[124]:
```
LotFrontage    0
MasVnrArea     0
GarageYrBlt    0
dtype: int64
```

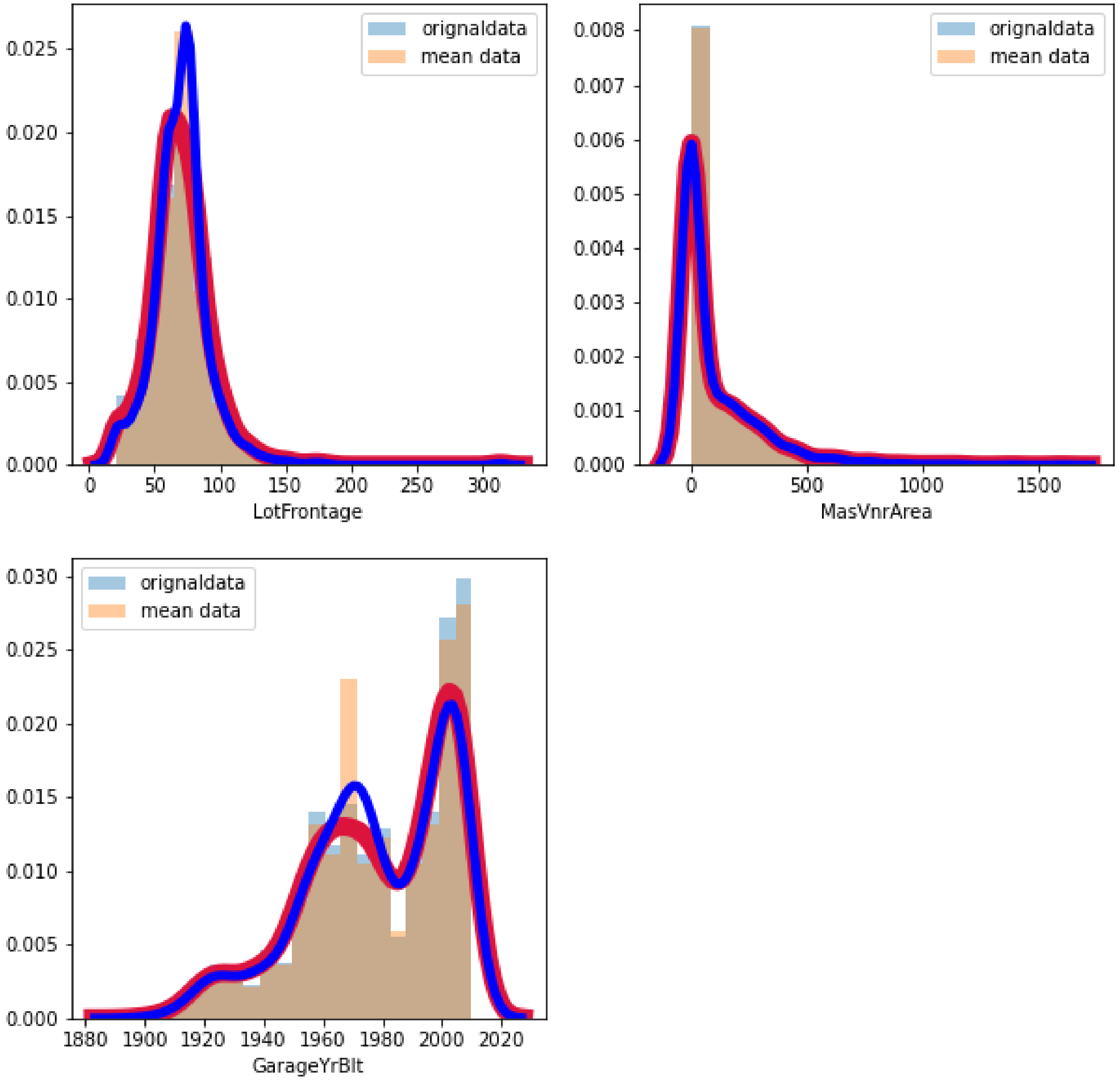
In [125]:
```
df_copy[df_copy[['MasVnrType']].isnull().any(axis=1)]
```

Out[125]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neig
234	235	60	RL	79.076923	7851	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	
529	530	20	RL	74.923631	32668	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	
650	651	60	FV	65.000000	8125	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	
936	937	20	RL	67.000000	10083	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	
973	974	20	FV	95.000000	11639	Pave	NaN	Reg	Lvl	AllPub	Corner	Gtl	
977	978	120	FV	35.000000	4274	Pave	Pave	IR1	Lvl	AllPub	Inside	Gtl	
1243	1244	20	RL	107.000000	13891	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	
1278	1279	60	RL	75.000000	9473	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	

In [126]:

```
plt.figure(figsize=(10,10))
for i, var in enumerate(num_vars_miss):
    plt.subplot(2,2,i+1)
    sns.distplot(df[var],bins=20,kde_kws={"linewidth":10,"color":"#DC143C"},label="orignaldata")
    sns.distplot(df_copy[var],bins=20,kde_kws={"linewidth":5,"color":"blue"},label="mean data")
    plt.legend()
```


orignaldata
mean data
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
LotFrontage
orignaldata
mean data
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
500
1000
1500
MasVnrArea
orignaldata
mean data
0.030
0.025
0.020
0.015
0.010
0.005
0.000
1880
1900
1920
1940
1960
1980
2000
2020
GarageYrBlt

```
In [ ]:
```

In [132]:

```
df_copy_median = df.copy()
num_vars_miss = ['LotFrontage','MasVnrArea','GarageYrBlt']
cat_vars=['LotConfig','MasVnrType','GarageType']
for cat_vars ,null_var_miss in zip(cat_vars,num_vars_miss):

    for var_class in df[cat_vars].unique():
        df_copy_median.update(df[df.loc[:,cat_vars] == var_class][num_var_miss].replace(np.nan,df[df.loc[:,
```

In [133]:

```
df_copy_median[num_vars_miss].isnull().sum()
```

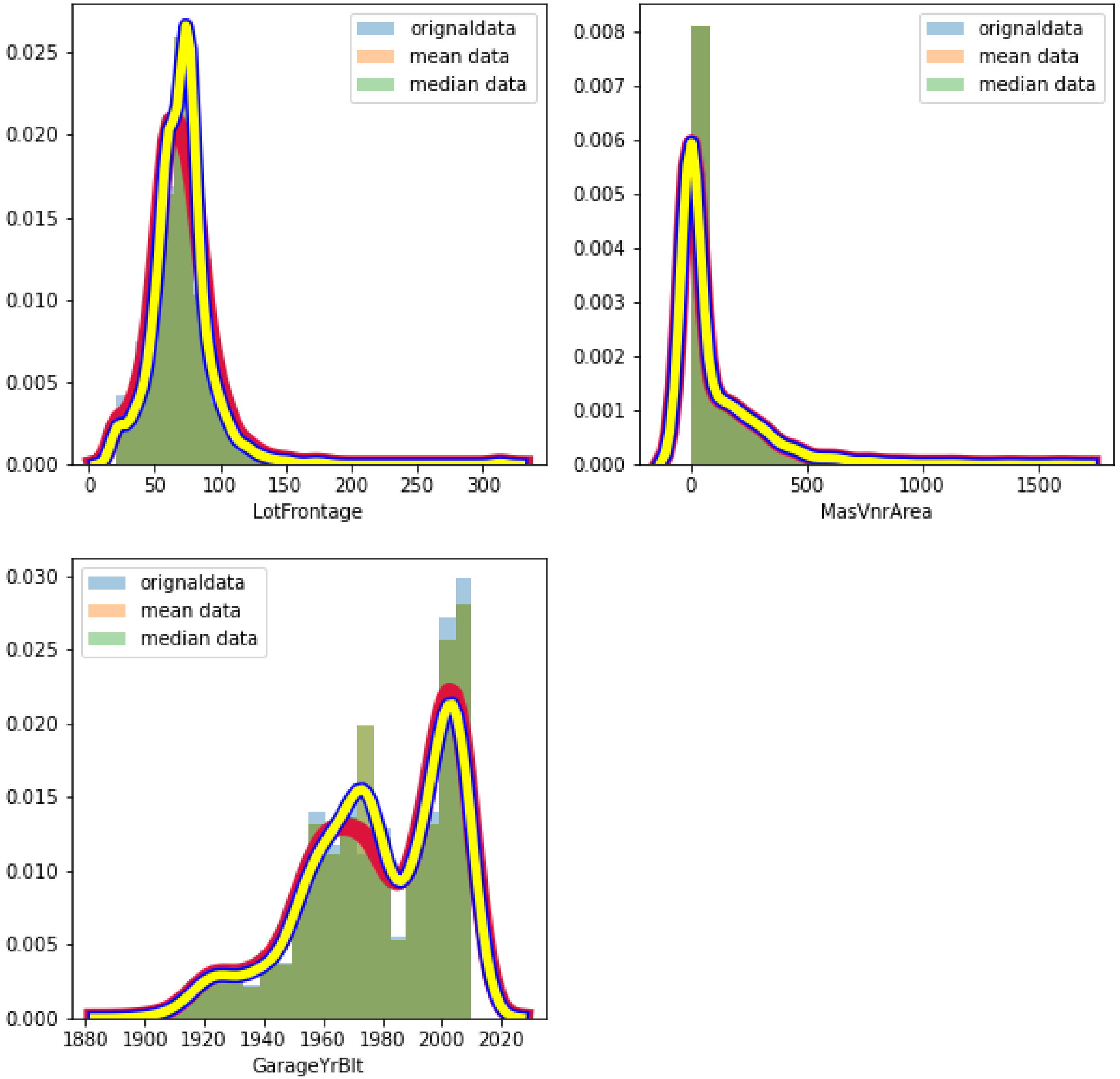
Out[133]:

```
LotFrontage    0
MasVnrArea     0
GarageYrBlt    0
dtype: int64
```

In [135]:

```
plt.figure(figsize=(10,10))
for i, var in enumerate(num_vars_miss):
    plt.subplot(2,2,i+1)
    sns.distplot(df[var],bins=20,kde_kws={"linewidth":10,"color":"#DC143C"},label="orignaldata")
    sns.distplot(df_copy[var],bins=20,kde_kws={"linewidth":8,"color":"blue"},label="mean data")

    sns.distplot(df_copy_median[var],bins=20,kde_kws={"linewidth":5,"color":"yellow"},label="median data")
    plt.legend()
```



0.025
0.020
0.015
0.010
0.005
0.000
orignaldata
mean data
median data
0
50
100
150
200
250
300
LotFrontage
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
orignaldata
mean data
median data
0
500
1000
1500
MasVnrArea
0.030
0.025
0.020
0.015
0.010
0.005
0.000
orignaldata
mean data
median data
1880
1900
1920
1940
1960
1980
2000
2020
GarageYrBlt