Data Cleaning

Missing value imputation by Mean, Median

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

```
                                uc?export=download&
                                    |
                                    |
                                    |
                                   *


data_df=r"https://drive.google.com/open?id=1BiGZSedP4BIIuTbVTBodOhVgFImaz08"
```

In [2]:
```
data=pd.read_csv("train.csv")
data
```

Out[2]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fenc
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	Nal
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	Nal
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	Nal
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	Nal
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	Nal
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1455	1456	60	RL	62.0	7917	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	Nal
1456	1457	20	RL	85.0	13175	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	MnPr
1457	1458	70	RL	66.0	9042	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	GdPr
1458	1459	20	RL	68.0	9717	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	Nal
1459	1460	20	RL	75.0	9937	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	Nal

1460 rows × 81 columns

In [3]:
```
data.shape
```

Out[3]: (1460, 81)

In [4]:
```
pd.set_option("display.max_columns",None)
pd.set_option("display.max_rows",None)
```

In [5]:
```
data.info()
```

```
<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  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  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  BsmtCond       1423 non-null   object
32  BsmtExposure   1422 non-null   object
33  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  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
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  EnclosedPorch  1460 non-null   int64
69  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
 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
dtypes: float64(3), int64(35), object(43)
memory usage: 678.7+ KB
```

In [6]:
```
data.isnull().sum()
```

Out[6]:
```
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
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
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
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
SalePrice        0
dtype: int64
```

In [7]:
```
data.isnull().sum().sum()
```

Out[7]:
```
6965
```

In [8]:
```
data.head()
```

Out[8]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhc
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 [9]:
```
null_var=data.isnull().sum()/data.shape[0]*100
null_var
```

Out[9]:
```
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
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
```

In [10]:
```
drop_cloumn = null_var[null_var >20].keys()
drop_cloumn
```

Out[10]:
```
Index(['Alley', 'FireplaceQu', 'PoolQC', 'Fence', 'MiscFeature'], dtype='object')
```

In [11]:
```
data2=data.drop(columns=drop_cloumn)
```

In [12]:
```
data2.shape
```

Out[12]:
```
(1460, 76)
```

In [13]:
```
data3_num=data2.select_dtypes(include=["int64","float64"])
```

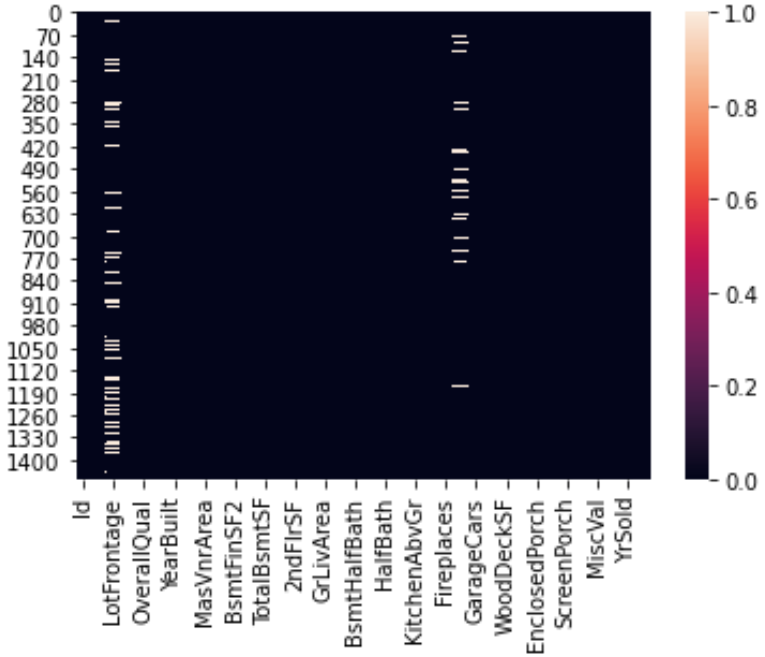
In [14]:
```
data3_num.head()
```

Out[14]:

	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 [15]:
```
sns.heatmap(data3_num.isnull())
```

Out[15]: <AxesSubplot:>



In [17]:
```
data3_num.isnull().sum()/data.shape[0]*100
```

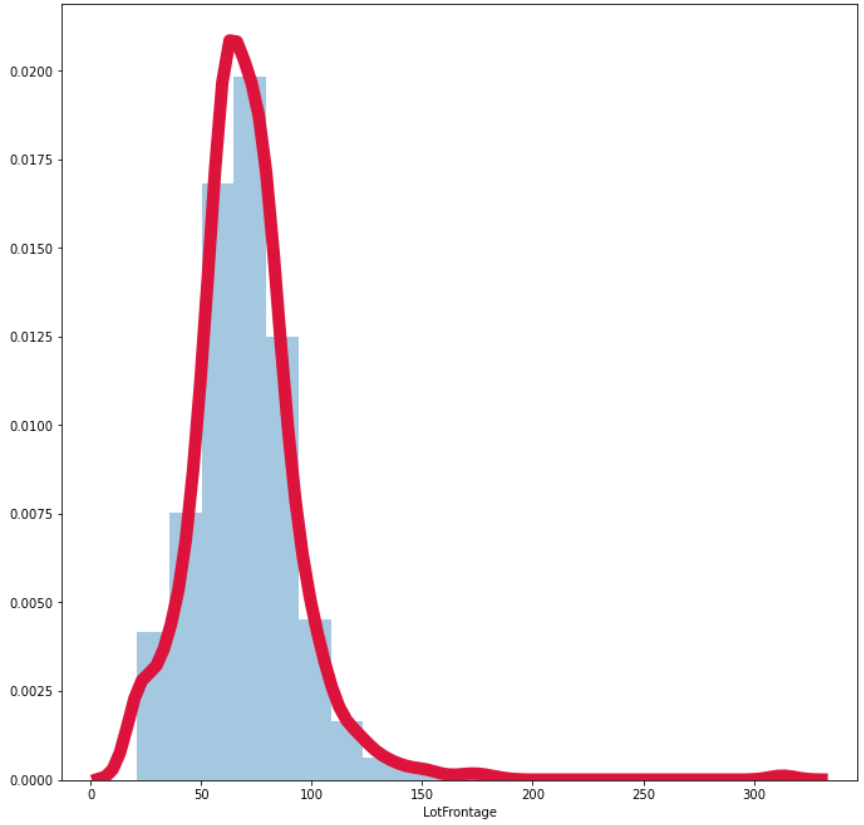
Out[17]:
```
Id               0.000000
MSSubClass       0.000000
LotFrontage     17.739726
LotArea          0.000000
OverallQual      0.000000
OverallCond      0.000000
YearBuilt        0.000000
YearRemodAdd     0.000000
MasVnrArea       0.547945
BsmtFinSF1       0.000000
BsmtFinSF2       0.000000
BsmtUnfSF        0.000000
TotalBsmtSF      0.000000
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
TotRmsAbvGrd     0.000000
Fireplaces       0.000000
GarageYrBlt      5.547945
GarageCars       0.000000
GarageArea       0.000000
WoodDeckSF       0.000000
OpenPorchSF      0.000000
EnclosedPorch    0.000000
3SsnPorch        0.000000
ScreenPorch      0.000000
PoolArea         0.000000
MiscVal          0.000000
```

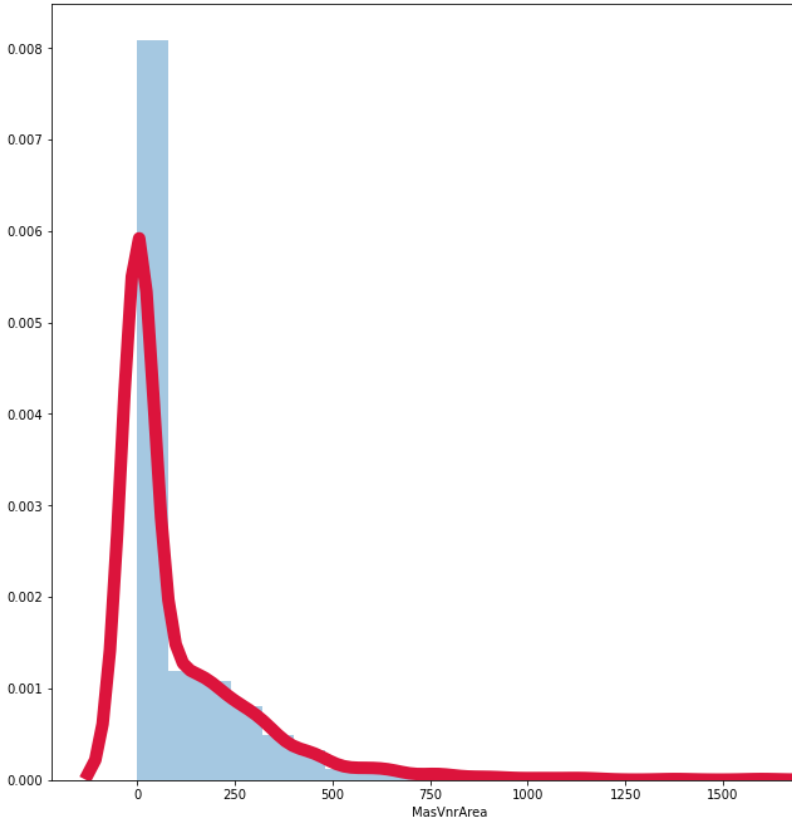
```
MoSold           0.000000
YrSold           0.000000
SalePrice        0.000000
dtype: float64
```

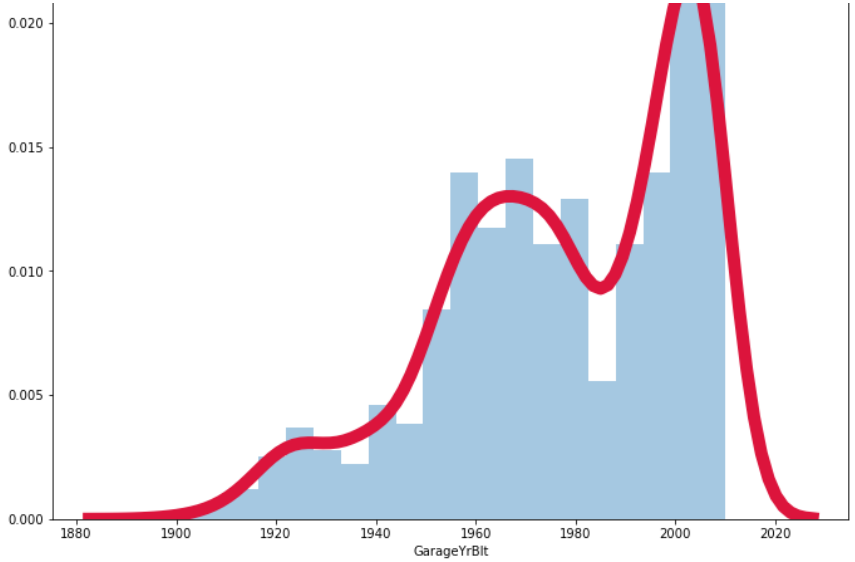
In [19]:
```
missin_num_var=[var for var in data3_num.columns if data3_num[var].isnull().sum()>0]
missin_num_var
```

Out[19]:
```
['LotFrontage', 'MasVnrArea', 'GarageYrBlt']
```

In [22]:
```
plt.figure(figsize=(25,25))
for i, var in enumerate(missin_num_var):
    plt.subplot(2,2,i+1)
    sns.distplot(data3_num[var],bins=20,kde_kws={"linewidth":10,"color":"#DC143C"})
#     sns.distplot(data3[var],bins=20)
```





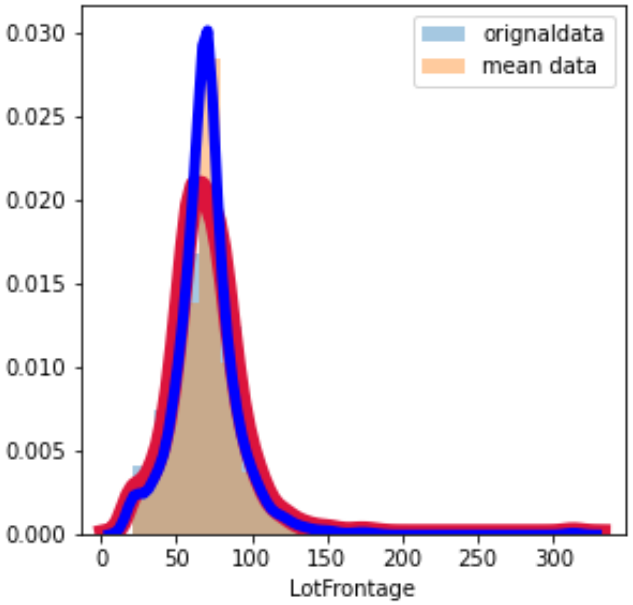


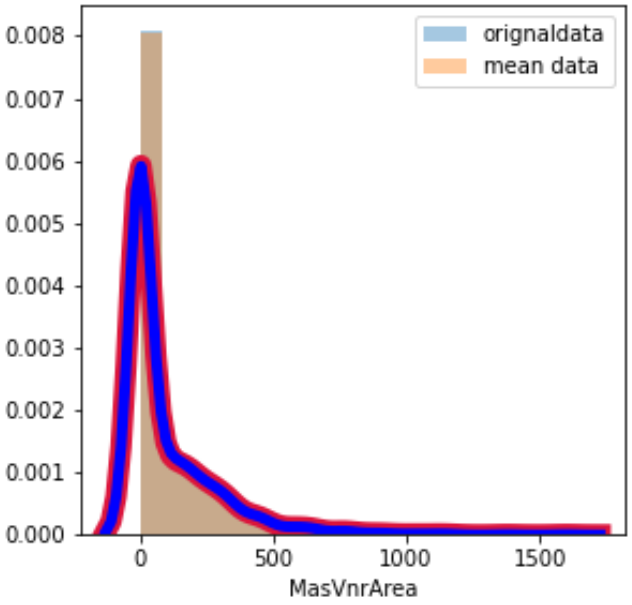
In [23]:
```
data4_num_mean=data3_num.fillna(data3_num.mean())
```

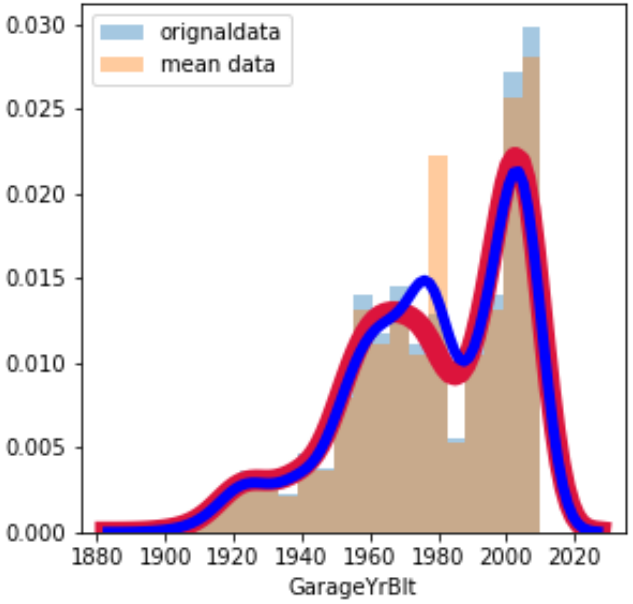
In [25]:
```
data4_num_mean.isnull().sum().sum()
```

Out[25]:
```
0
```

In [33]:
```
plt.figure(figsize=(10,10))
for i, var in enumerate(missin_num_var):
    plt.subplot(2,2,i+1)
    sns.distplot(data3_num[var],bins=20,kde_kws={"linewidth":10,"color":"#DC143C"},label="orignaldata")
    sns.distplot(data4_num_mean[var],bins=20,kde_kws={"linewidth":5,"color":"blue"},label="mean data")
    plt.legend()
```

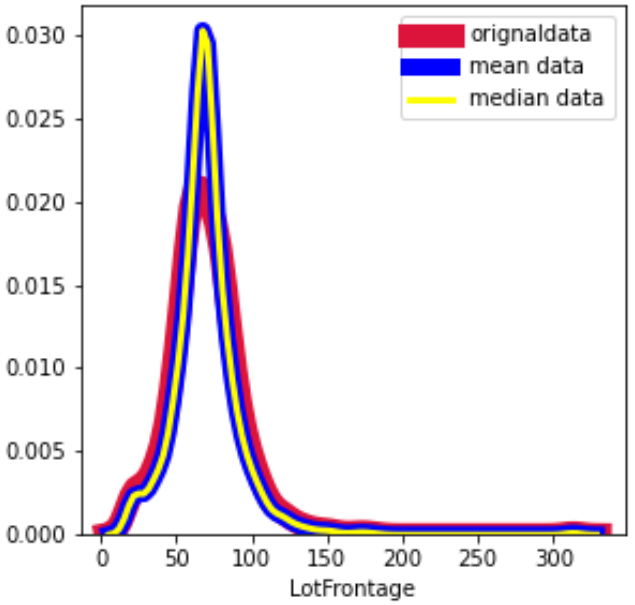


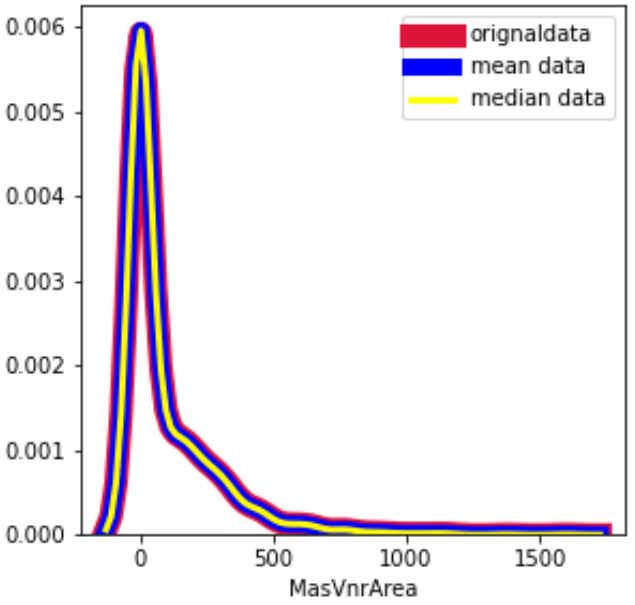


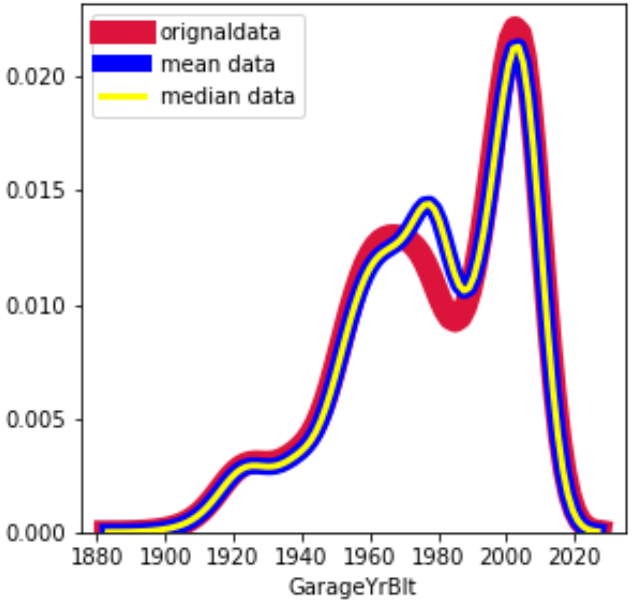


In [40]:
```
data5_num_median=data3_num.fillna(data3_num.median())
```

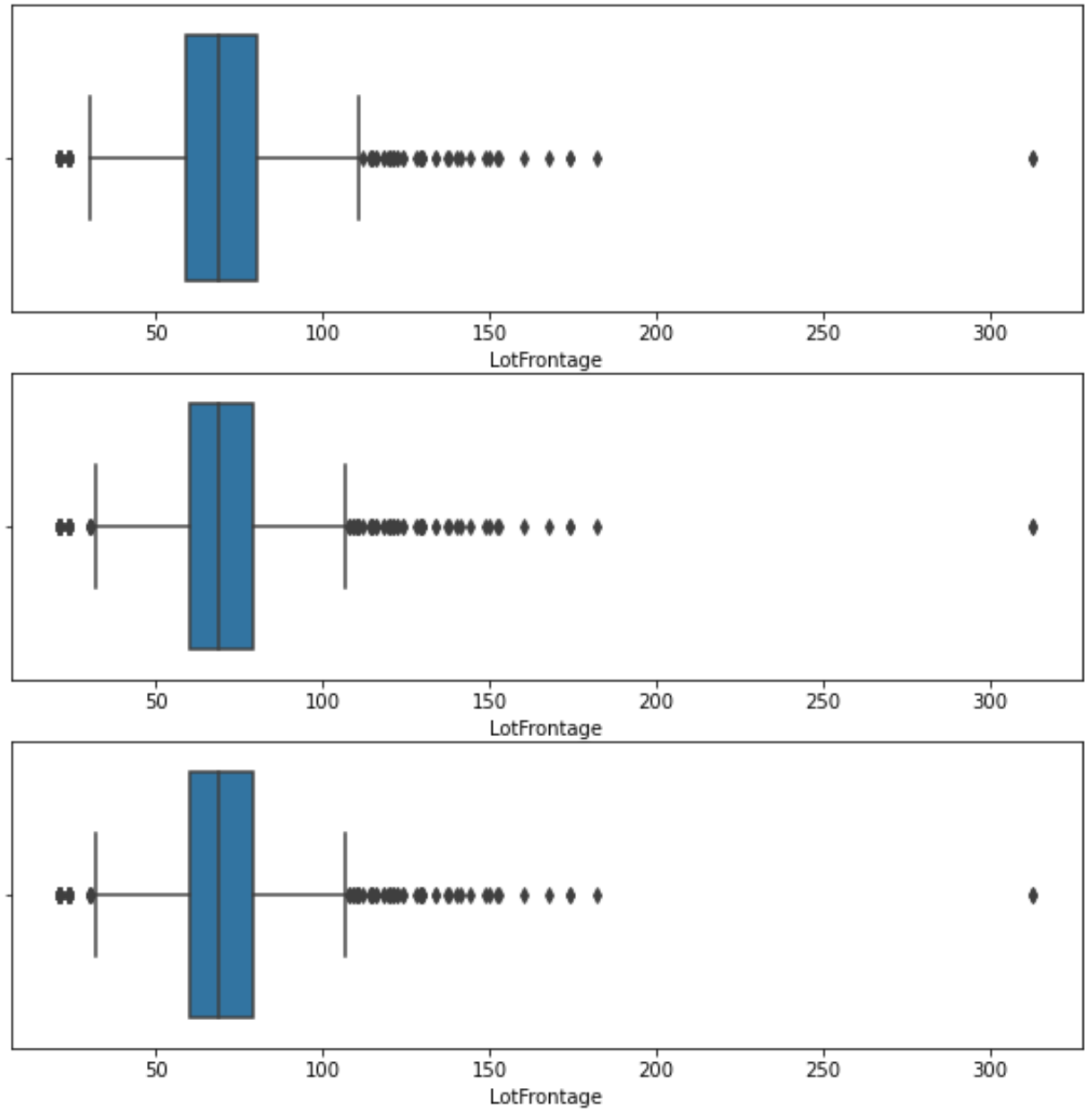
In [47]:
```
plt.figure(figsize=(10,10))
for i, var in enumerate(missin_num_var):
    plt.subplot(2,2,i+1)
    sns.distplot(data3_num[var],bins=20,hist=False,kde_kws={"linewidth":11,"color":"#DC143C"},label="orgina
    sns.distplot(data4_num_mean[var],bins=20,hist=False,kde_kws={"linewidth":8,"color":"blue"},label="mean
    sns.distplot(data5_num_median[var],bins=20,hist=False,kde_kws={"linewidth":3,"color":"yellow"},label="m
    plt.legend()
```

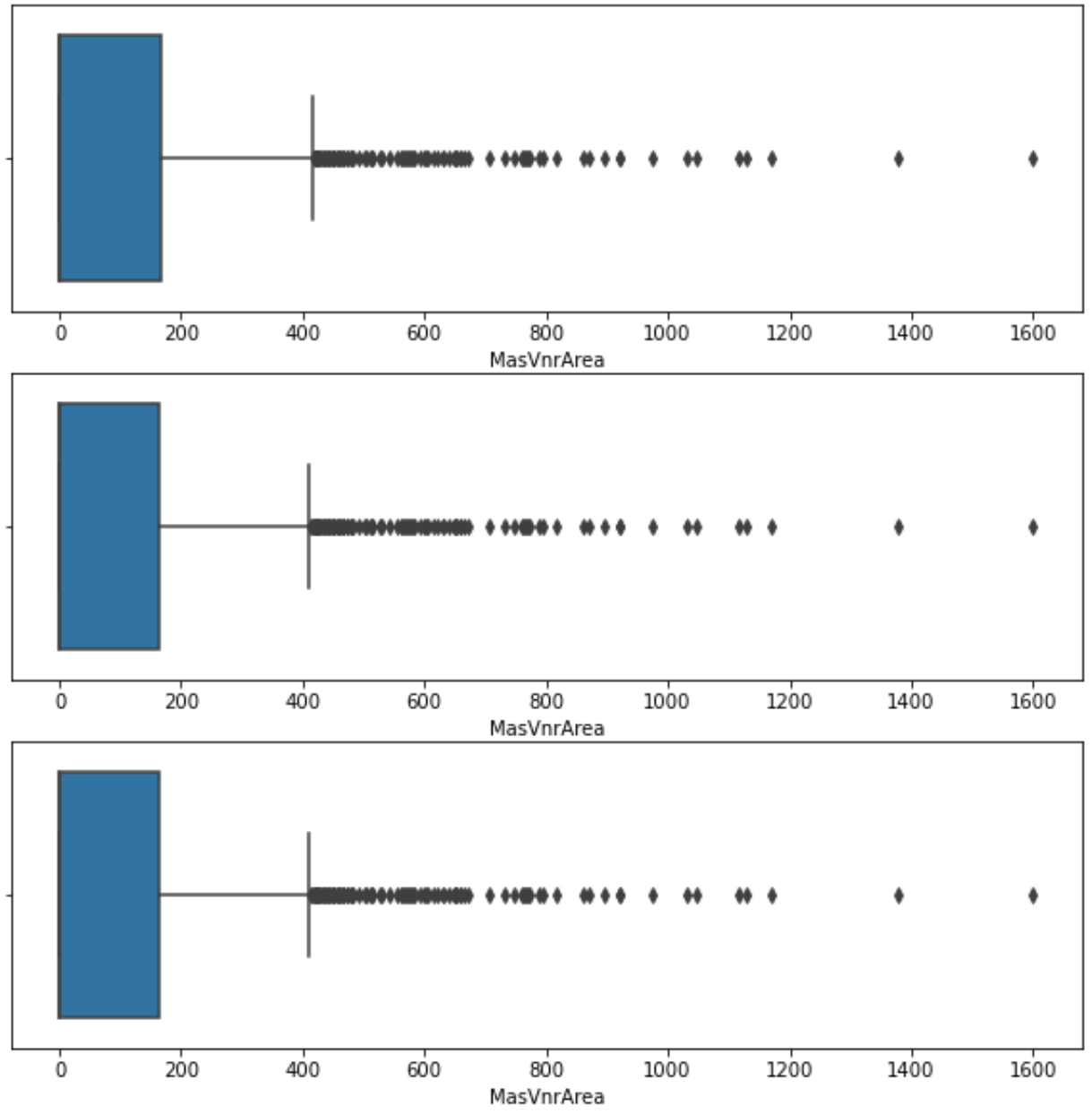

orignaldata
mean data
median data
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
LotFrontage

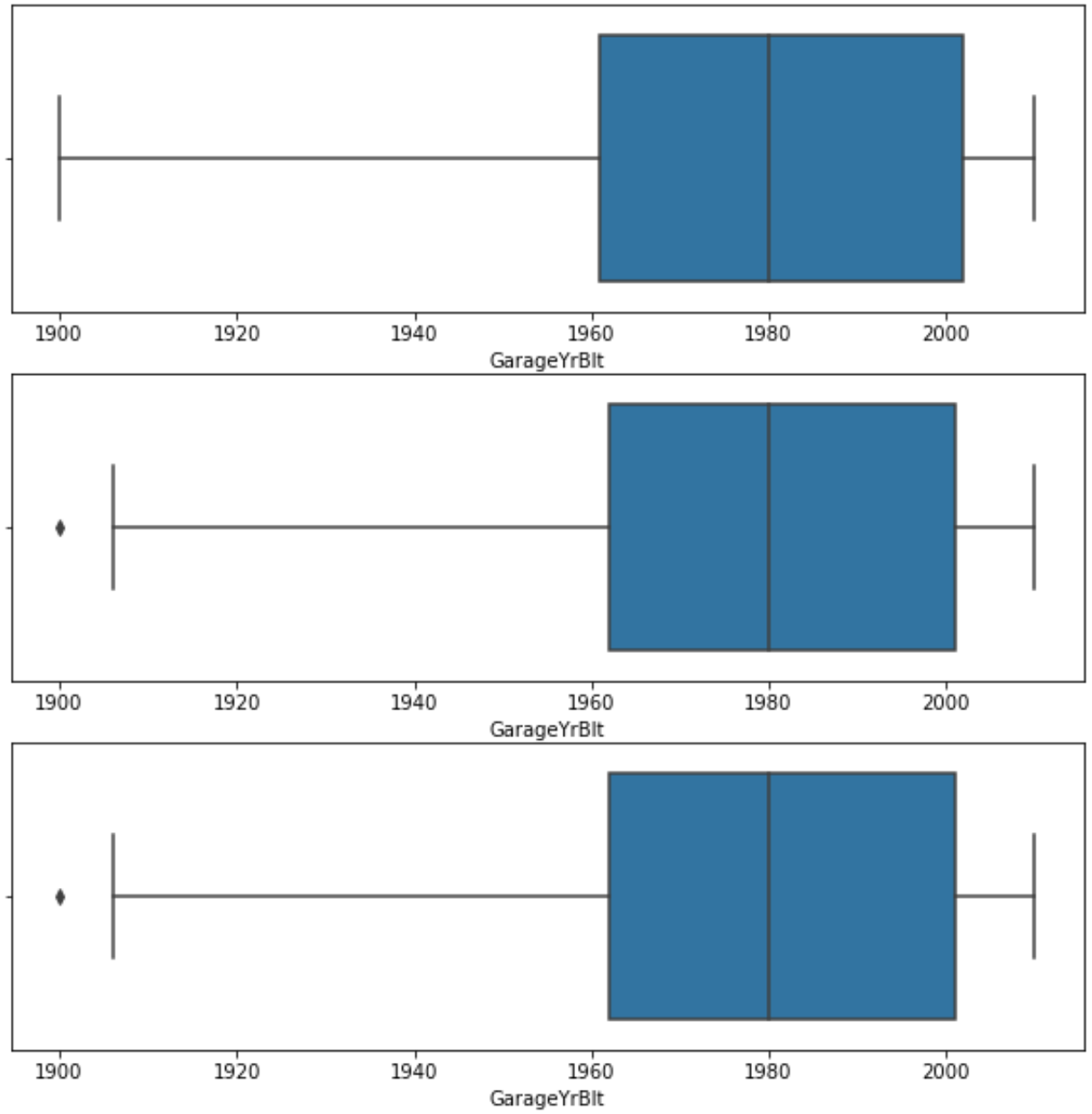

orignaldata
mean data
median data
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
500
1000
1500
MasVnrArea


orignaldata
mean data
median data
0.020
0.015
0.010
0.005
0.000
1880
1900
1920
1940
1960
1980
2000
2020
GarageYrBlt

In [52]:
```
for i, var in enumerate(missin_num_var):
    plt.figure(figsize=(10,10))
    plt.subplot(3,1,1)
    sns.boxplot(data[var])
    plt.subplot(3,1,2)
    sns.boxplot(data4_num_mean[var])
    plt.subplot(3,1,3)
    sns.boxplot(data5_num_median[var])
```


50
100
150
200
250
300
LotFrontage
50
100
150
200
250
300
LotFrontage
50
100
150
200
250
300
LotFrontage


0
200
400
600
800
1000
1200
1400
1600
MasVnrArea
0
200
400
600
800
1000
1200
1400
1600
MasVnrArea
0
200
400
600
800
1000
1200
1400
1600
MasVnrArea



In [55]:
```
data_concat=pd.concat([data[missin_num_var],data4_num_mean[missin_num_var],data5_num_median[missin_num_var]
```

In [59]:
```
con_data=data_concat[data_concat.isnull().any(axis=1)]
```

In [61]:
```
con_data.head(25)
```

Out[61]:

	LotFrontage	MasVnrArea	GarageYrBlt	LotFrontage	MasVnrArea	GarageYrBlt	LotFrontage	MasVnrArea	GarageYrBlt
7	NaN	240.0	1973.0	69.0	240.0	1973.0	69.0	240.0	1973.0
12	NaN	0.0	1962.0	69.0	0.0	1962.0	69.0	0.0	1962.0
14	NaN	212.0	1960.0	69.0	212.0	1960.0	69.0	212.0	1960.0
16	NaN	180.0	1970.0	69.0	180.0	1970.0	69.0	180.0	1970.0
24	NaN	0.0	1968.0	69.0	0.0	1968.0	69.0	0.0	1968.0
31	NaN	0.0	1966.0	69.0	0.0	1966.0	69.0	0.0	1966.0
39	65.0	0.0	NaN	65.0	0.0	1980.0	65.0	0.0	1980.0
42	NaN	0.0	1983.0	69.0	0.0	1983.0	69.0	0.0	1983.0
43	NaN	0.0	1977.0	69.0	0.0	1977.0	69.0	0.0	1977.0
48	33.0	0.0	NaN	33.0	0.0	1980.0	33.0	0.0	1980.0
50	NaN	0.0	1997.0	69.0	0.0	1997.0	69.0	0.0	1997.0
64	NaN	573.0	1998.0	69.0	573.0	1998.0	69.0	573.0	1998.0
66	NaN	287.0	1970.0	69.0	287.0	1970.0	69.0	287.0	1970.0
76	NaN	0.0	1956.0	69.0	0.0	1956.0	69.0	0.0	1956.0
78	72.0	0.0	NaN	72.0	0.0	1980.0	72.0	0.0	1980.0
84	NaN	22.0	1995.0	69.0	22.0	1995.0	69.0	22.0	1995.0
88	105.0	0.0	NaN	105.0	0.0	1980.0	105.0	0.0	1980.0
89	60.0	0.0	NaN	60.0	0.0	1980.0	60.0	0.0	1980.0
95	NaN	68.0	1993.0	69.0	68.0	1993.0	69.0	68.0	1993.0
99	77.0	0.0	NaN	77.0	0.0	1980.0	77.0	0.0	1980.0
100	NaN	28.0	1977.0	69.0	28.0	1977.0	69.0	28.0	1977.0

	LotFrontage	MasVnrArea	GarageYrBlt	LotFrontage	MasVnrArea	GarageYrBlt	LotFrontage	MasVnrArea	GarageYrBlt
104	NaN	600.0	1951.0	69.0	600.0	1951.0	69.0	600.0	1951.0
108	85.0	0.0	NaN	85.0	0.0	1980.0	85.0	0.0	1980.0
111	NaN	0.0	2000.0	69.0	0.0	2000.0	69.0	0.0	2000.0
113	NaN	184.0	1953.0	69.0	184.0	1953.0	69.0	184.0	1953.0