Data Preprocessing

Missing value imputation by mean and median

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

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

Out[2]:

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

5 rows × 81 columns

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

Out[3]:
```
(1460, 81)
```

In [4]:
```
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: 924.0+ KB
```

In [5]:
```
data.isnull().sum()
```

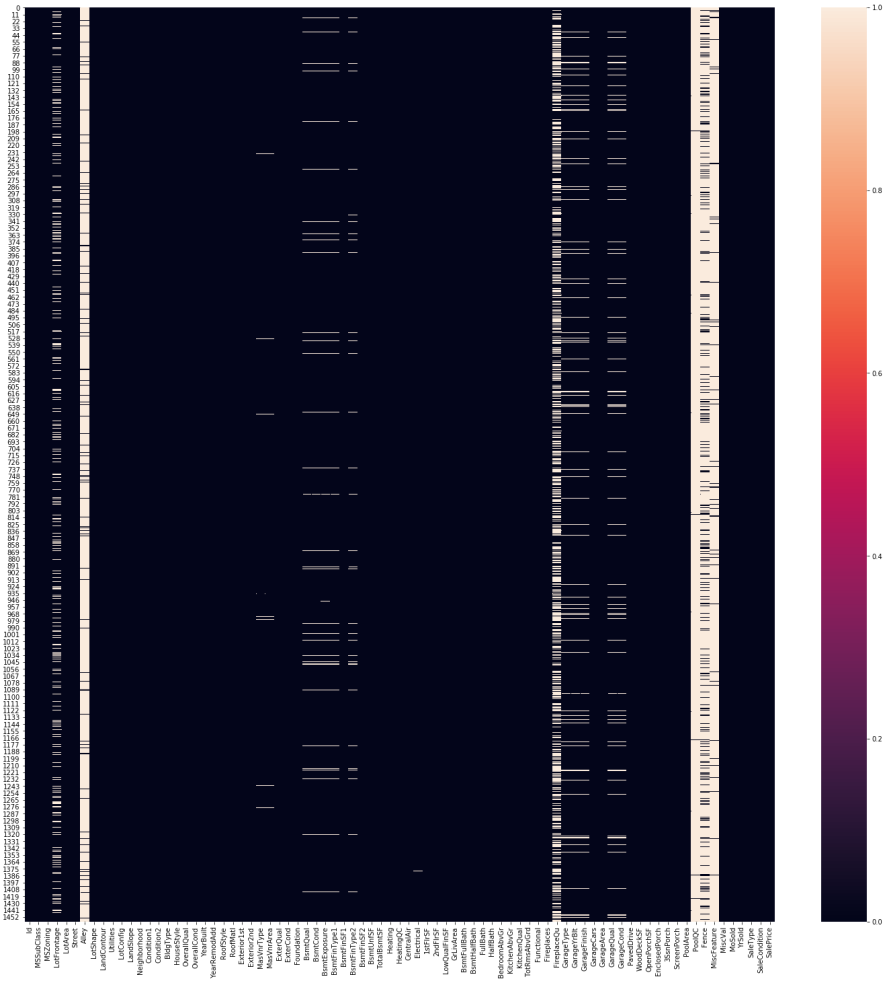
Out[5]:
```
Id                0
MSSubClass        0
MSZoning          0
LotFrontage     259
LotArea           0
                 ...
MoSold            0
YrSold            0
SaleType          0
SaleCondition     0
SalePrice         0
Length: 81, dtype: int64
```

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

Out[6]:
```
6965
```

In [7]:
```
plt.figure(figsize=(25,25))
sns.heatmap(data.isnull())
```

Out[7]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1fc40e73508>
```



In [12]:
```
missing_value = data.isnull().sum()/data.shape[0]*100
missing_value
```

Out[12]:
```
Id               0.000000
MSSubClass       0.000000
MSZoning         0.000000
LotFrontage     17.739726
LotArea          0.000000
                   ...
MoSold           0.000000
YrSold           0.000000
SaleType         0.000000
SaleCondition    0.000000
SalePrice        0.000000
```

```
Length: 81, dtype: float64
```

In [15]:
```
missing_value_gre = missing_value[missing_value > 17].keys()
missing_value_gre
```

Out[15]:
```
Index(['LotFrontage', 'Alley', 'FireplaceQu', 'PoolQC', 'Fenc
e',
       'MiscFeature'],
      dtype='object')
```

In [18]:
```
data_drop = data.drop(columns=missing_value_gre)
data_drop.shape
```

Out[18]:
```
(1460, 75)
```

In [20]:
```
num_data = data_drop.select_dtypes(include=['int64','float64'])
num_data.head()
```

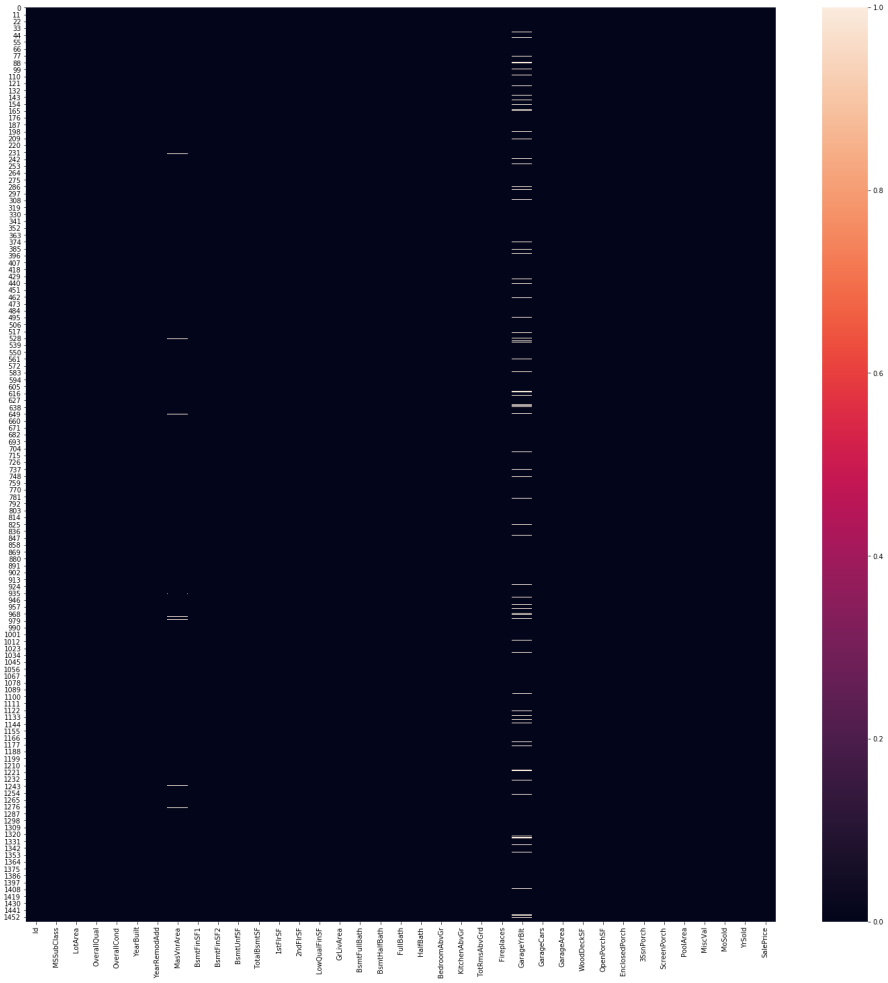
Out[20]:

	Id	MSSubClass	LotArea	OverallQual	OverallCond	YearBuilt	YearRemod
0	1	60	8450	7	5	2003	
1	2	20	9600	6	8	1976	
2	3	60	11250	7	5	2001	
3	4	70	9550	7	5	1915	
4	5	60	14260	8	5	2000	

5 rows × 37 columns

In [21]:
```
plt.figure(figsize=(25,25))
sns.heatmap(num_data.isnull())
```

Out[21]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1fc41c6ae88>
```



In [25]:
```
num_data[num_data.isnull().any(axis=1)]
```

Out[25]:

	Id	MSSubClass	LotArea	OverallQual	OverallCond	YearBuilt	YearF
39	40	90	6040	4	5	1955	
48	49	190	4456	4	5	1920	
78	79	90	10778	4	5	1968	
88	89	50	8470	3	2	1915	
89	90	20	8070	4	5	1994	
...	...	...	...	...	...	...	...
1349	1350	70	5250	8	5	1872	
1407	1408	20	8780	5	5	1985	

	Id	MSSubClass	LotArea	OverallQual	OverallCond	YearBuilt	YearR
1449	1450	180	1533	5	7	1970	
1450	1451	90	9000	5	5	1974	
1453	1454	20	17217	5	5	2006	

89 rows × 37 columns

In [26]:
```
num_data.isnull().sum()
```

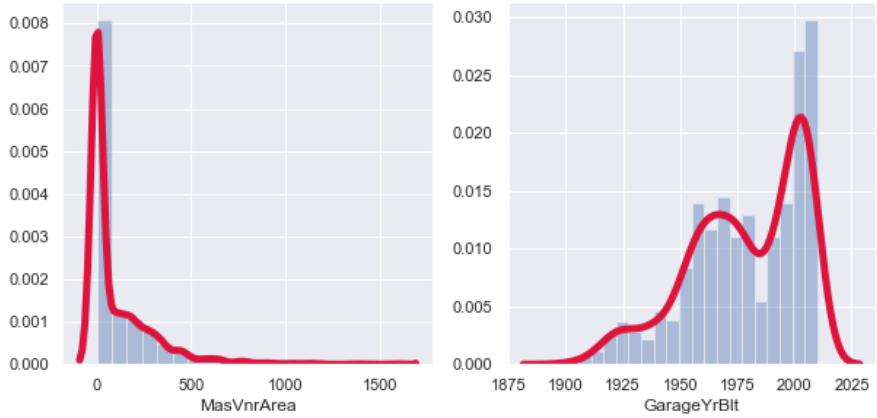
Out[26]:
```
Id              0
MSSubClass      0
LotArea         0
OverallQual     0
OverallCond     0
YearBuilt       0
YearRemodAdd    0
MasVnrArea      8
BsmtFinSF1      0
BsmtFinSF2      0
BsmtUnfSF       0
TotalBsmtSF     0
1stFlrSF        0
2ndFlrSF        0
LowQualFinSF    0
GrLivArea       0
BsmtFullBath    0
BsmtHalfBath    0
FullBath        0
HalfBath        0
BedroomAbvGr    0
KitchenAbvGr    0
TotRmsAbvGrd    0
Fireplaces      0
GarageYrBlt    81
GarageCars      0
GarageArea      0
WoodDeckSF      0
OpenPorchSF     0
EnclosedPorch   0
3SsnPorch       0
ScreenPorch     0
PoolArea        0
MiscVal         0
MoSold          0
YrSold          0
```

```
SalePrice         0
dtype: int64
```

In [30]:
```
missing_num_var = [var for var in num_data.columns if num_data[var]
.isnull().sum() > 0]
missing_num_var
```

Out[30]:
```
['MasVnrArea', 'GarageYrBlt']
```

In [33]:
```
plt.figure(figsize=(10,10))
sns.set()
for i,var in enumerate(missing_num_var):
  plt.subplot(2,2,i+1)
  sns.distplot(num_data[var],bins=20,kde_kws={'linewidth':5,'color'
:'#DC143C'})
```

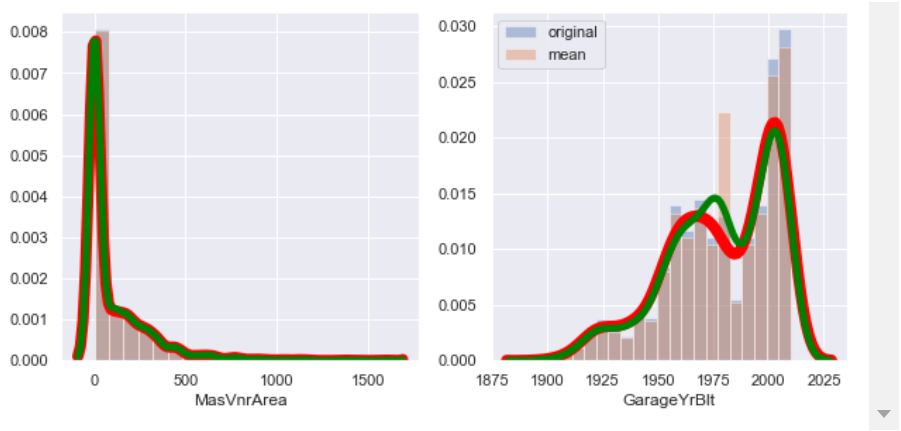


In [36]:
```
data_num_mean = num_data.fillna(num_data.mean())
data_num_mean.isnull().sum().sum()
```

Out[36]:
```
0
```

In [48]:
```
plt.figure(figsize=(10,10))
sns.set()
for i,var in enumerate(missing_num_var):
    plt.subplot(2,2,i+1)
    sns.distplot(num_data[var] , bins=20 , kde_kws={'linewidth': 8
,'color':'red'} , label="original")
    sns.distplot(data_num_mean[var] , bins=20 , kde_kws={'linewidt
h': 5,'color':'green'} , label="mean")
plt.legend()
```

Out[48]:
```
<matplotlib.legend.Legend at 0x1fc46bf1f48>
```

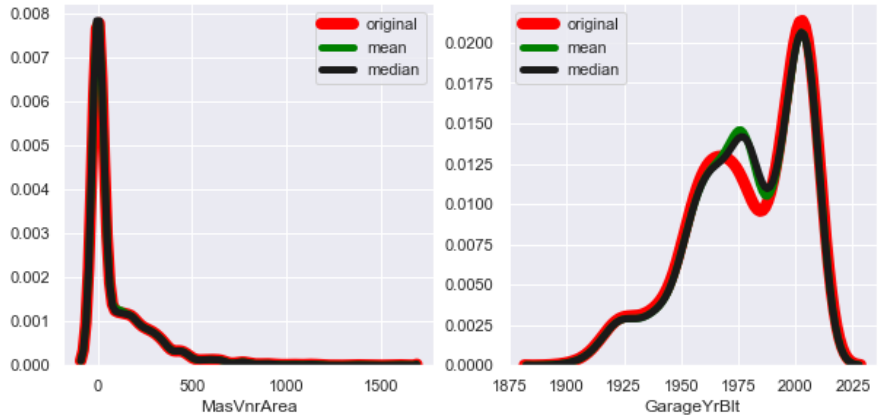


In [51]:
```
data_num_median = num_data.fillna(num_data.median())
data_num_median.isnull().sum().sum()
```

Out[51]: 0

In [58]:
```
plt.figure(figsize=(10,10))
sns.set()
for i,var in enumerate(missing_num_var):
    plt.subplot(2,2,i+1)
    sns.distplot(num_data[var] , bins=20 ,hist=False, kde_kws={'linewidth': 8 ,'color':'red'} , label="original")
    sns.distplot(data_num_mean[var] , bins=20 ,  hist=False , kde_kws={'linewidth': 5,'color':'green'} , label="mean")
    sns.distplot(data_num_median[var] , bins=20 , hist=False , kde_kws={'linewidth': 5,'color':'k'} , label="median")
plt.legend()
```

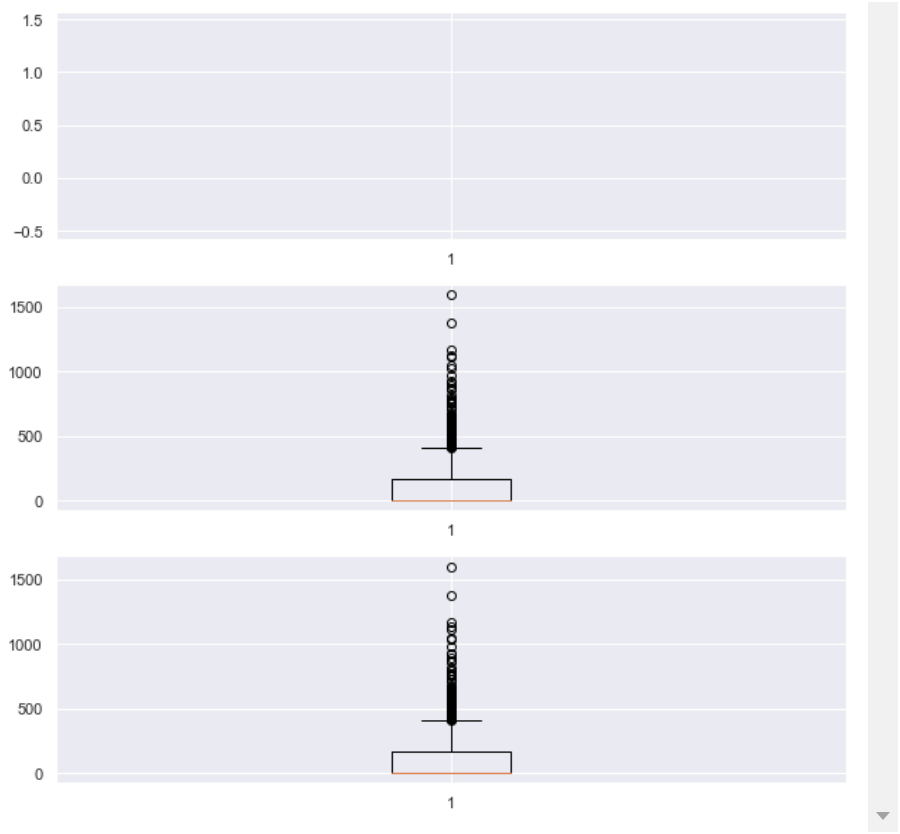
Out[58]: <matplotlib.legend.Legend at 0x1fc459fbb08>

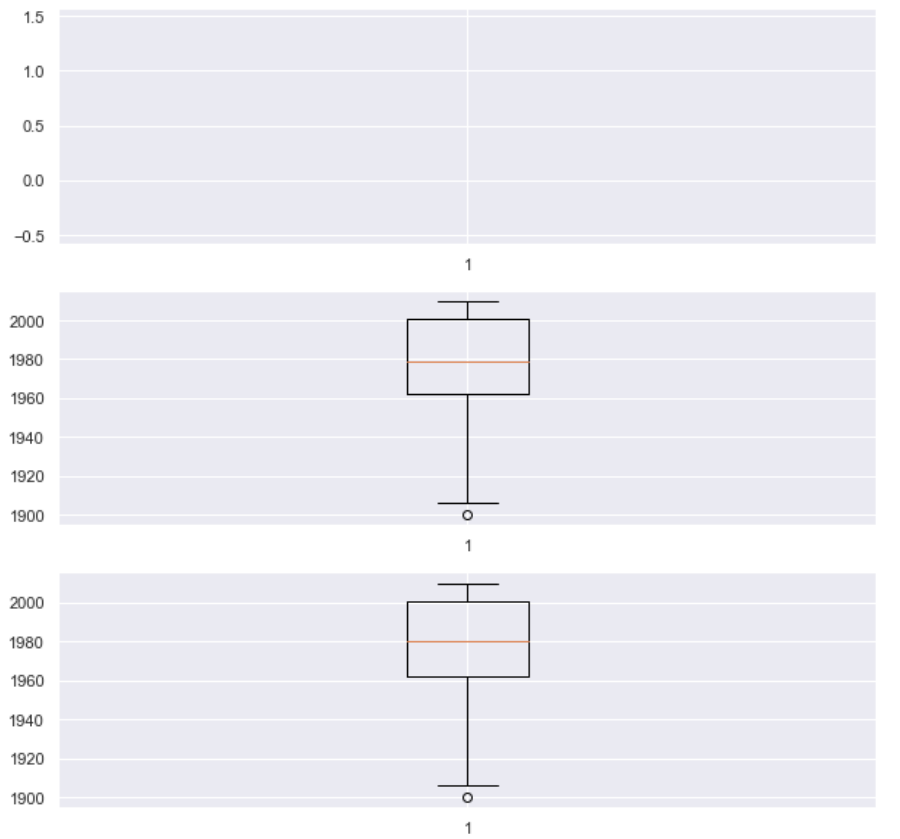


In [59]:
```
for i,var in enumerate(missing_num_var):
    plt.figure(figsize=(10,10))
```

```
    plt.subplot(3,1,1)
    plt.boxplot(num_data[var])
    plt.subplot(3,1,2)
    plt.boxplot(data_num_mean[var])
    plt.subplot(3,1,3)
    plt.boxplot(data_num_median[var])
```

```
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__init__.py:1316: RuntimeWarning: invalid value encountered in less_equal
  wiskhi = x[x <= hival]
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__init__.py:1323: RuntimeWarning: invalid value encountered in greater_equal
  wisklo = x[x >= loval]
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__init__.py:1331: RuntimeWarning: invalid value encountered in less
  x[x < stats['whislo']],
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__init__.py:1332: RuntimeWarning: invalid value encountered in greater
  x[x > stats['whishi']],
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__init__.py:1316: RuntimeWarning: invalid value encountered in less_equal
  wiskhi = x[x <= hival]
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__init__.py:1323: RuntimeWarning: invalid value encountered in greater_equal
  wisklo = x[x >= loval]
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__init__.py:1331: RuntimeWarning: invalid value encountered in less
  x[x < stats['whislo']],
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__init__.py:1332: RuntimeWarning: invalid value encountered in greater
  x[x > stats['whishi']],
```


1.5
1.0
0.5
0.0
−0.5
1
1500
1000
500
0
1
1500
1000
500
0
1


1.5
1.0
0.5
0.0
−0.5
1
2000
1980
1960
1940
1920
1900
1
2000
1980
1960
1940
1920
1900
1