Data Preprocessing

Missing value imputation by mean and median of Class

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
```

Out[3]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	Lot
0	1	60	RL	65.0	8450	Pave	NaN	
1	2	20	RL	80.0	9600	Pave	NaN	
2	3	60	RL	68.0	11250	Pave	NaN	
3	4	70	RL	60.0	9550	Pave	NaN	
4	5	60	RL	84.0	14260	Pave	NaN	
...	...	...	...	...	...	...	...	...
1455	1456	60	RL	62.0	7917	Pave	NaN	
1456	1457	20	RL	85.0	13175	Pave	NaN	
1457	1458	70	RL	66.0	9042	Pave	NaN	
1458	1459	20	RL	68.0	9717	Pave	NaN	

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	Lo
1459	1460	20	RL	75.0	9937	Pave	NaN	

1460 rows × 81 columns

In [4]:
```
data.shape
```

Out[4]:
```
(1460, 81)
```

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: 924.0+ KB
```

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

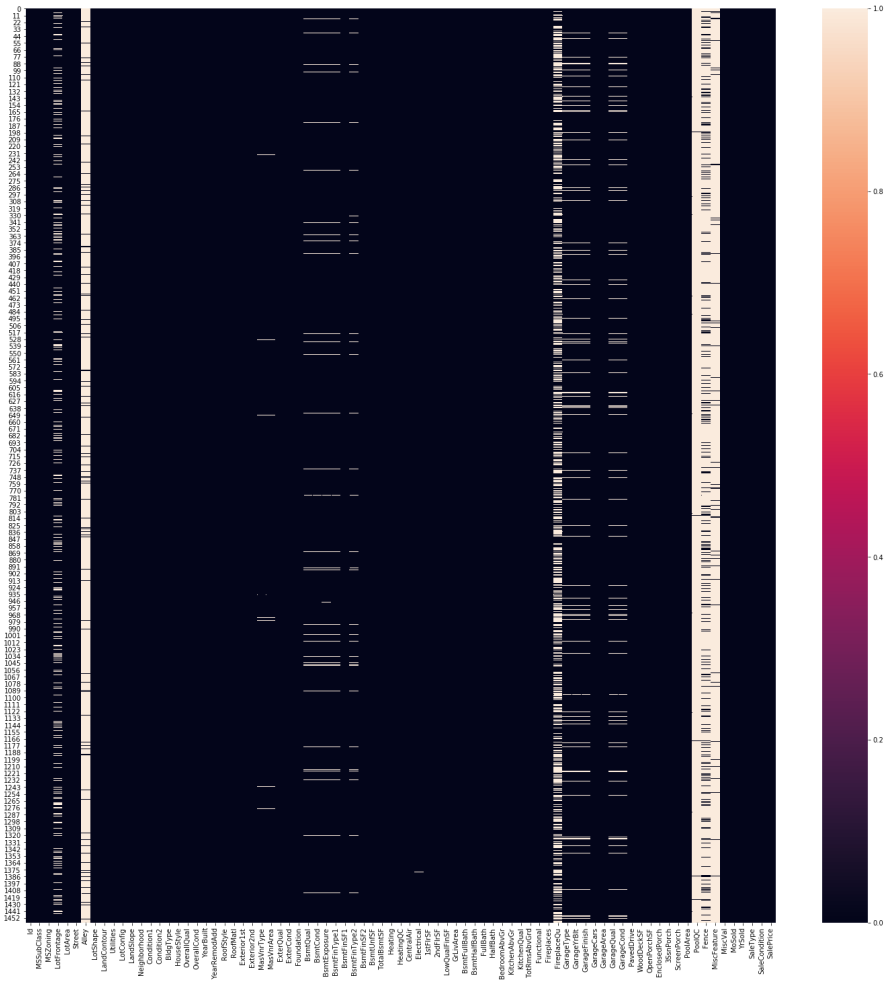
Out[6]:
```
Id                0
MSSubClass        0
MSZoning          0
LotFrontage     259
LotArea           0
                 ...
MoSold            0
YrSold            0
SaleType          0
SaleCondition     0
SalePrice         0
Length: 81, dtype: int64
```

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

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

In [8]:
```
plt.figure(figsize=(25,25))
sns.heatmap(data.isnull())
```

Out[8]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1f60004a888>
```



In [9]:
```
missing_value = data.isnull()/data.shape[0]*100
missing_value
```

Out[9]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	L
0	0.0	0.0	0.0	0.0	0.0	0.0	0.068493	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.068493	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.068493	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.068493	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.068493	
...	...	...	...	...	...	...	...	
1455	0.0	0.0	0.0	0.0	0.0	0.0	0.068493	

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	L
1456	0.0	0.0	0.0	0.0	0.0	0.0	0.068493	
1457	0.0	0.0	0.0	0.0	0.0	0.0	0.068493	
1458	0.0	0.0	0.0	0.0	0.0	0.0	0.068493	
1459	0.0	0.0	0.0	0.0	0.0	0.0	0.068493	

1460 rows × 81 columns

In [10]:
```
missing_value_per = data.isnull().sum()/data.shape[0]*100
missing_value_per
```

Out[10]:
```
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 [11]:
```
missing_value_per_gre = missing_value_per[missing_value_per > 20].keys()
missing_value_per_gre
```

Out[11]:
```
Index(['Alley', 'FireplaceQu', 'PoolQC', 'Fence', 'MiscFeature'], dtype='object')
```

In [12]:
```
data_drop = data.drop(columns=missing_value_per_gre)
data_drop.shape
```

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

In [13]:
```
data_num = data_drop.select_dtypes(include=['int64','float64'])
data_num.shape
```

Out[13]:
```
(1460, 38)
```

In [14]:
```
data_num.isnull().sum()
```

Out[14]:
```
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
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 [15]:
```
data_num[data_num.isnull().any(axis=1)]
```

Out[15]:

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	Ye
7	8	60	NaN	10382	7	6	
12	13	20	NaN	12968	5	6	
14	15	20	NaN	10920	6	5	

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	Ye
16	17	20	NaN	11241	6	7	
24	25	20	NaN	8246	5	8	
...	...	...	...	...	...	...	
1443	1444	30	NaN	8854	6	6	
1446	1447	20	NaN	26142	5	7	
1449	1450	180	21.0	1533	5	7	
1450	1451	90	60.0	9000	5	5	
1453	1454	20	90.0	17217	5	5	

339 rows × 38 columns

In [16]:
```
missing_num_var = [var for var in data_num.columns if data_num[var]
.isnull().sum() > 0]
missing_num_var
```

Out[16]:
```
['LotFrontage', 'MasVnrArea', 'GarageYrBlt']
```

In [17]:
```
missing_value_var=['LotFrontage', 'MasVnrArea', 'GarageYrBlt']
data_num[missing_value_var][data_num[missing_value_var].isnull().an
y(axis=1)]
```

Out[17]:

	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
...	...	...	...
1443	NaN	0.0	1916.0
1446	NaN	189.0	1962.0
1449	21.0	0.0	NaN
1450	60.0	0.0	NaN
1453	90.0	0.0	NaN

339 rows × 3 columns

In [18]:
```
data['LotConfig'].unique()
```

Out[18]:
```
array(['Inside', 'FR2', 'Corner', 'CulDSac', 'FR3'], dtype=obje
ct)
```

In [19]:
```
data[data.loc[ : , 'LotConfig'] == "Inside'"]["LotFrontage"].replac
e(np.nan , data[data.loc[ :  , 'LotConfig'] == "Inside'"]["LotFront
age"].mean())
```

Out[19]:
```
Series([], Name: LotFrontage, dtype: float64)
```

In [20]:
```
data_copy = data.copy()
for class_var in data['LotConfig'].unique():
    data_copy.update(data[data.loc[ : , 'LotConfig'] == class_var][
"LotFrontage"].replace(np.nan , data[data.loc[ :  , 'LotConfig'] ==
class_var]["LotFrontage"].mean()))
```

In [21]:
```
data_copy.isnull().sum()
```

Out[21]:
```
Id               0
MSSubClass       0
MSZoning         0
LotFrontage      0
LotArea          0
                ..
MoSold           0
YrSold           0
SaleType         0
SaleCondition    0
SalePrice        0
Length: 81, dtype: int64
```

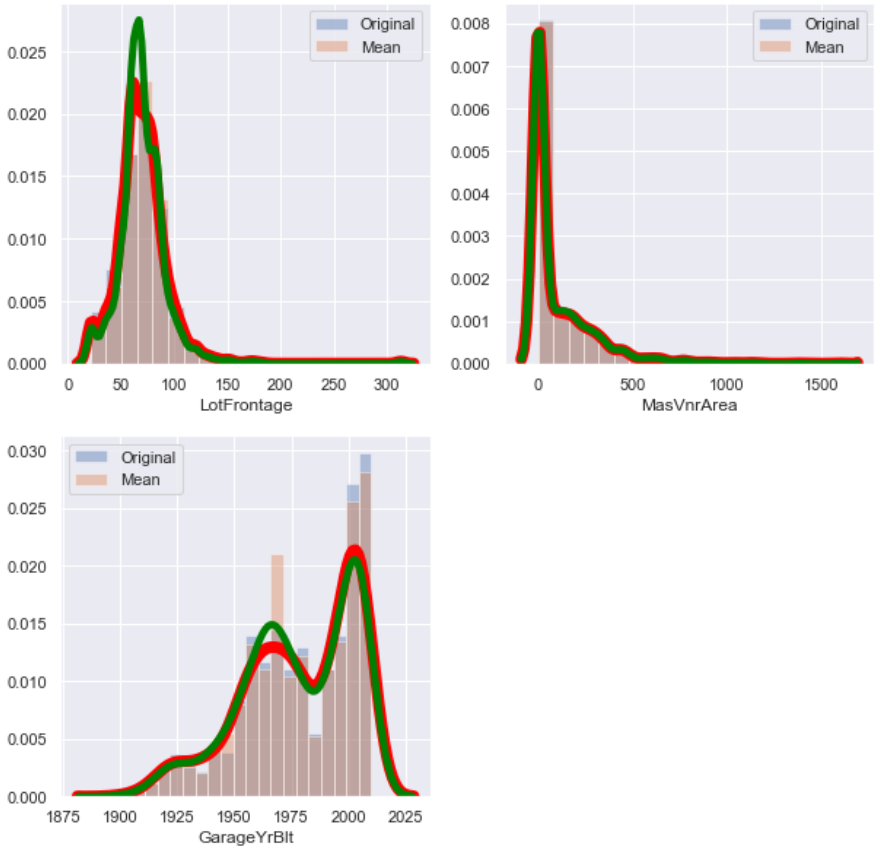
In [22]:
```
missing_num_vars = ['LotFrontage', 'MasVnrArea', 'GarageYrBlt']
cat_var = ['LotConfig','Exterior2nd','KitchenQual']
data_copy = data.copy()
for cat_var,missing_num_vars in zip(cat_var,missing_num_vars):
    for class_var in data[cat_var].unique():
        data_copy.update(data[data.loc[ : , cat_var] == class_var][
missing_num_vars].replace(np.nan , data[data.loc[ :  , cat_var] ==
class_var][missing_num_vars].mean()))
```

In [23]:
```
data_copy[missing_num_var].isnull().sum()
```

Out[23]:
```
LotFrontage    0
MasVnrArea     0
GarageYrBlt    0
dtype: int64
```

dtype: int64

In [24]:
```
plt.figure(figsize=(10,10))
sns.set()
for i,var in enumerate(missing_num_var):
  plt.subplot(2,2,i+1)
  sns.distplot(data[var] , bins=20, kde_kws={'linewidth':8 , 'color':'red'}, label = 'Original')
  sns.distplot(data_copy[var] , bins=20, kde_kws={'linewidth':5 , 'color':'green'}, label = 'Mean')
  plt.legend()
```



In [25]:
```
data_copy_median = data.copy()
for class_var in data['LotConfig'].unique():
    data_copy_median.update(data[data.loc[ : , 'LotConfig'] == class_var]["LotFrontage"].replace(np.nan , data[data.loc[ :  , 'LotConfig'] == class_var]["LotFrontage"].median()))
```

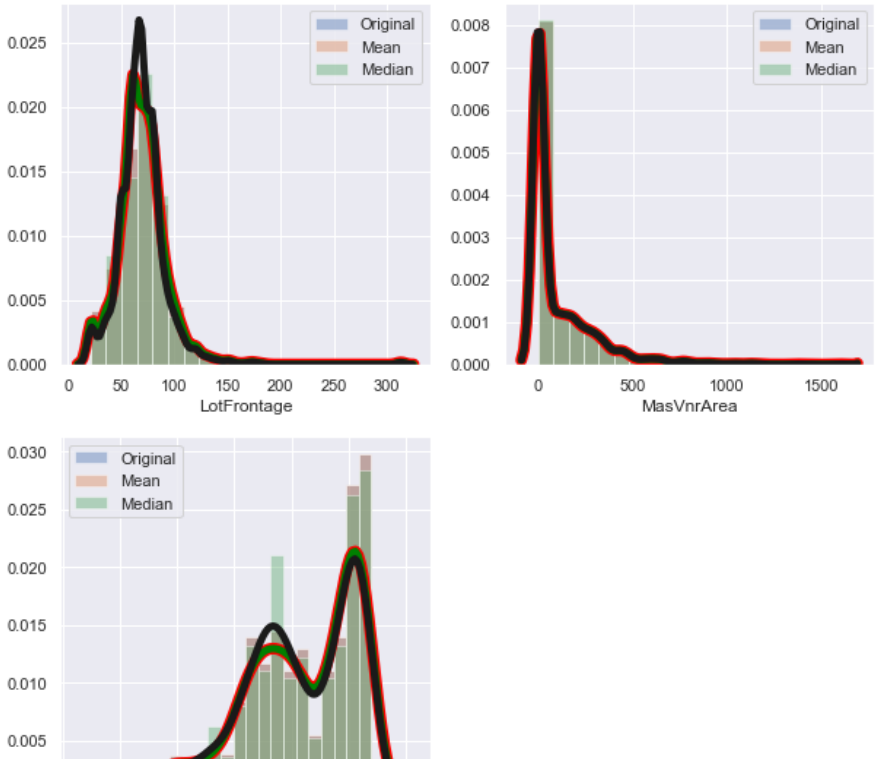
In [26]:
```
missing_num_vars = ['LotFrontage', 'MasVnrArea', 'GarageYrBlt']
cat_var = ['LotConfig','Exterior2nd','KitchenQual']
data_copy = data.copy()
```

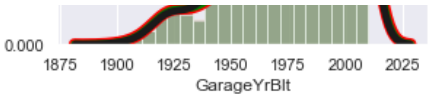
```
for cat_var,missing_num_vars in zip(cat_var,missing_num_vars):
    for class_var in data[cat_var].unique():
        data_copy_median.update(data[data.loc[ : , cat_var] == clas
s_var][missing_num_vars].replace(np.nan , data[data.loc[ :  , cat_v
ar] == class_var][missing_num_vars].median()))
```

In [27]:
```
data_copy_median[missing_num_var].isnull().sum()
```

Out[27]:
```
LotFrontage     0
MasVnrArea      0
GarageYrBlt     0
dtype: int64
```

In [28]:
```
plt.figure(figsize=(10,10))
sns.set()
for i,var in enumerate(missing_num_var):
  plt.subplot(2,2,i+1)
  sns.distplot(data[var] , bins=20, kde_kws={'linewidth':8 , 'colo
r':'red'}, label = 'Original')
  sns.distplot(data_copy[var] , bins=20, kde_kws={'linewidth':5 ,
'color':'green'}, label = 'Mean')
  sns.distplot(data_copy_median[var] , bins=20, kde_kws={'linewidt
h':5 , 'color':'k'}, label = 'Median')
  plt.legend()
```





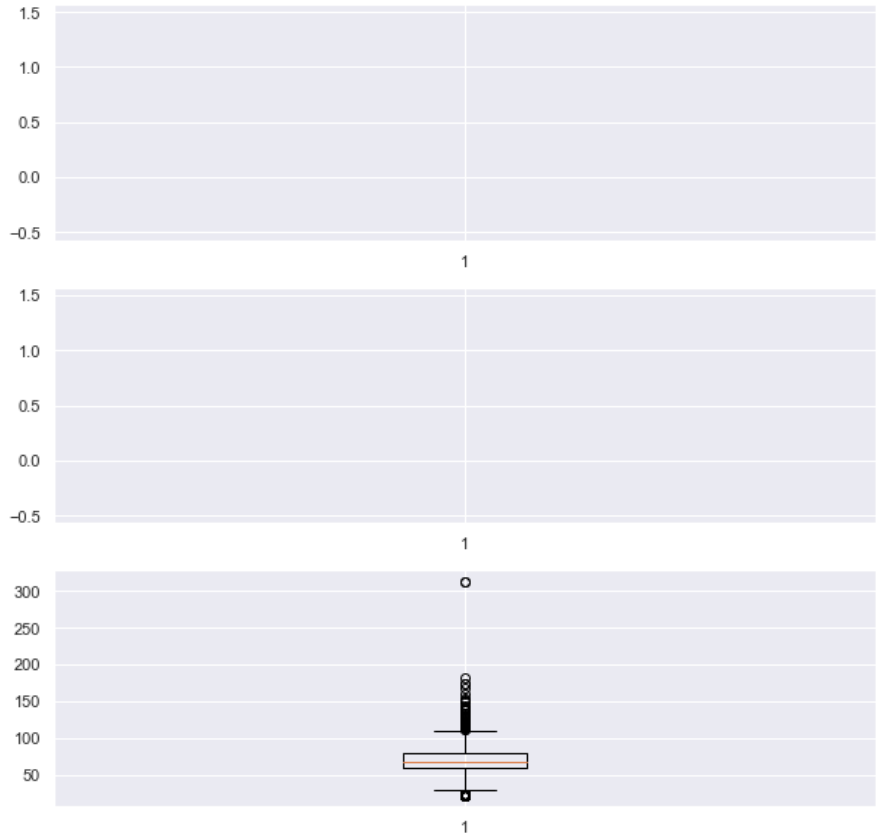
In [29]:

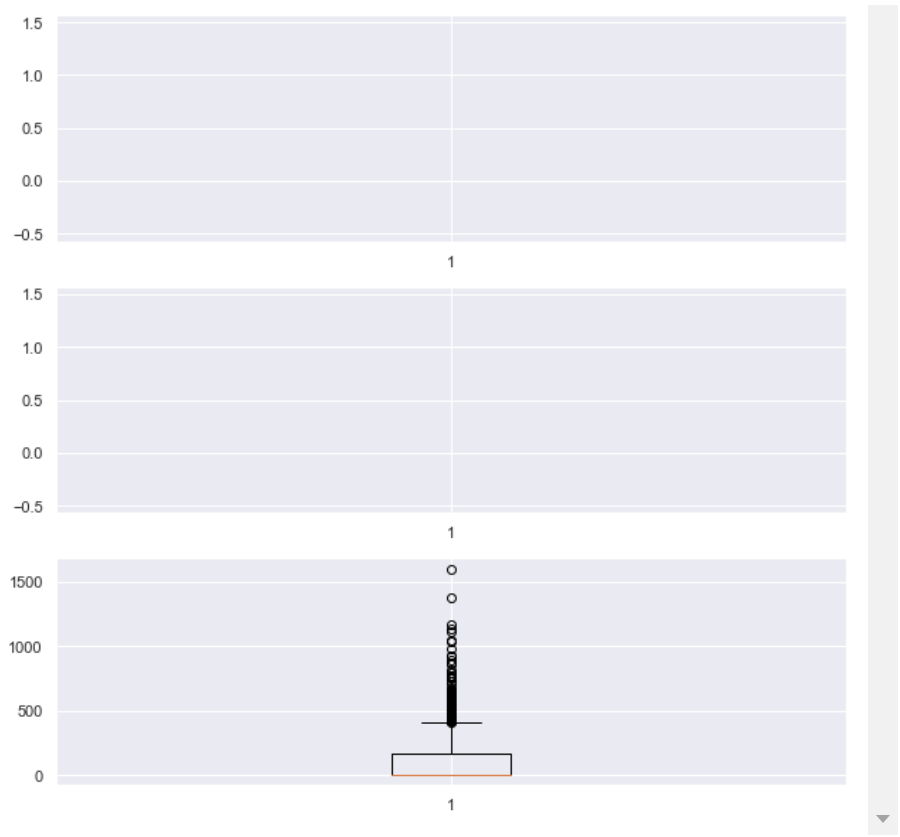
```
for i,var in enumerate(missing_num_var):
    plt.figure(figsize=(10,10))
    plt.subplot(3,1,1)
    plt.boxplot(data[var])
    plt.subplot(3,1,2)
    plt.boxplot(data_copy[var])
    plt.subplot(3,1,3)
    plt.boxplot(data_copy_median[var])
```

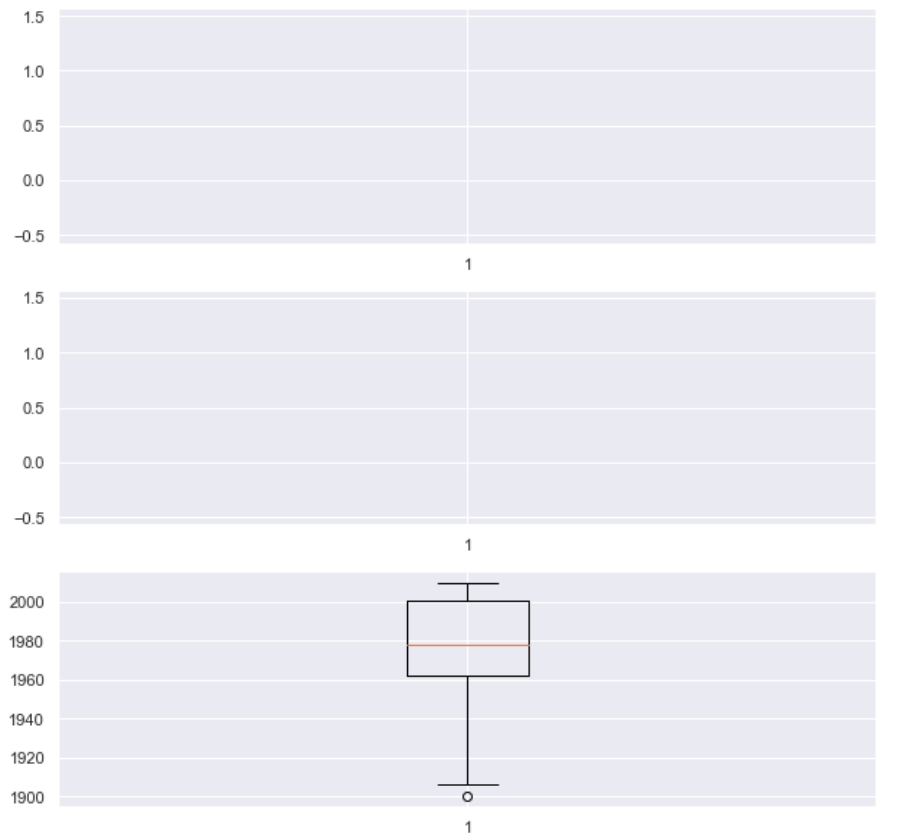
```
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1316: RuntimeWarning: invalid value encountered in les
s_equal
  wiskhi = x[x <= hival]
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1323: RuntimeWarning: invalid value encountered in gre
ater_equal
  wisklo = x[x >= loval]
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1331: RuntimeWarning: invalid value encountered in les
s
  x[x < stats['whislo']],
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1332: RuntimeWarning: invalid value encountered in gre
ater
  x[x > stats['whishi']],
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1316: RuntimeWarning: invalid value encountered in les
s_equal
  wiskhi = x[x <= hival]
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1323: RuntimeWarning: invalid value encountered in gre
ater_equal
  wisklo = x[x >= loval]
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1331: RuntimeWarning: invalid value encountered in les
s
  x[x < stats['whislo']],
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1332: RuntimeWarning: invalid value encountered in gre
ater
  x[x > stats['whishi']],
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1316: RuntimeWarning: invalid value encountered in les
s_equal
```

```
  wiskhi = x[x <= hival]
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1323: RuntimeWarning: invalid value encountered in gre
ater_equal
  wisklo = x[x >= loval]
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1331: RuntimeWarning: invalid value encountered in les
s
  x[x < stats['whislo']],
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1332: RuntimeWarning: invalid value encountered in gre
ater
  x[x > stats['whishi']],
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1316: RuntimeWarning: invalid value encountered in les
s_equal
  wiskhi = x[x <= hival]
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1323: RuntimeWarning: invalid value encountered in gre
ater_equal
  wisklo = x[x >= loval]
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1331: RuntimeWarning: invalid value encountered in les
s
  x[x < stats['whislo']],
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1332: RuntimeWarning: invalid value encountered in gre
ater
  x[x > stats['whishi']],
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1316: RuntimeWarning: invalid value encountered in les
s_equal
  wiskhi = x[x <= hival]
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1323: RuntimeWarning: invalid value encountered in gre
ater_equal
  wisklo = x[x >= loval]
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1331: RuntimeWarning: invalid value encountered in les
s
  x[x < stats['whislo']],
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1332: RuntimeWarning: invalid value encountered in gre
ater
  x[x > stats['whishi']],
C:\Users\Admin\anaconda3\lib\site-packages\matplotlib\cbook\__i
nit__.py:1316: RuntimeWarning: invalid value encountered in les
s_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
1.5
1.0
0.5
0.0
−0.5
1
1500
1000
500
0
1


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