Data Mining project Phase 1

Faridreza Momtaz Zandi 9812762601 Alireza Noorbakhsh 9812762496

Main focus in this phase was to understand datasets and how to work with them. We start by finding out our data types in the datasets and finally work on the numerical attrebutes to have basic knowledge of our atterbutes such as min, max, mean, range etc. We proceed to check our datasets one by one and get a full grasp on them.

INOUT Dataset

Unfortunatly there is no numerical attribute in this dataset and all of them are either Nominal, Ordinal or Binary and even the attributes that seems to be numertic are representing date or some sort of ID and there is no point or goal in workong on them and finding out their min, max, median etc.

INOUTLINE Dataset

Unlike our last dataset in this one we have a few numerical attributes that we can work on and get a sense of how is it to work on datasets! we begin by choosing these attributes and finally code!

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```
df = pd.read_csv('E:\ce\data mining\DataSets\INOUTLINE.csv' , low_memory = False)
df = df[['ACCUMULATEDEPRECIATION' , 'BOOKVALUE' , 'PRIMALVALUE' ,'DEPRECATION_PERIO
ogdf = df
median_df = df.median()
mode_df = df.mode()
df = df.describe()
df = df.transpose()
df['range'] = df['max'] - df['min']
df['Min acceptable value'] = df['min'] - ((df['max'] - df['min']) * 1.5)
df['Max acceptable value'] = df['max'] + ((df['max'] - df['min']) * 1.5)
df['median'] = median_df
df['mode'] = mode_df.head(1).transpose()
df = df.drop(columns = ['std' ,'count'])
df = df[['range' , 'min' , 'max' , 'mean' ,'mode' , 'median' , 'Min acceptable valu
df
```

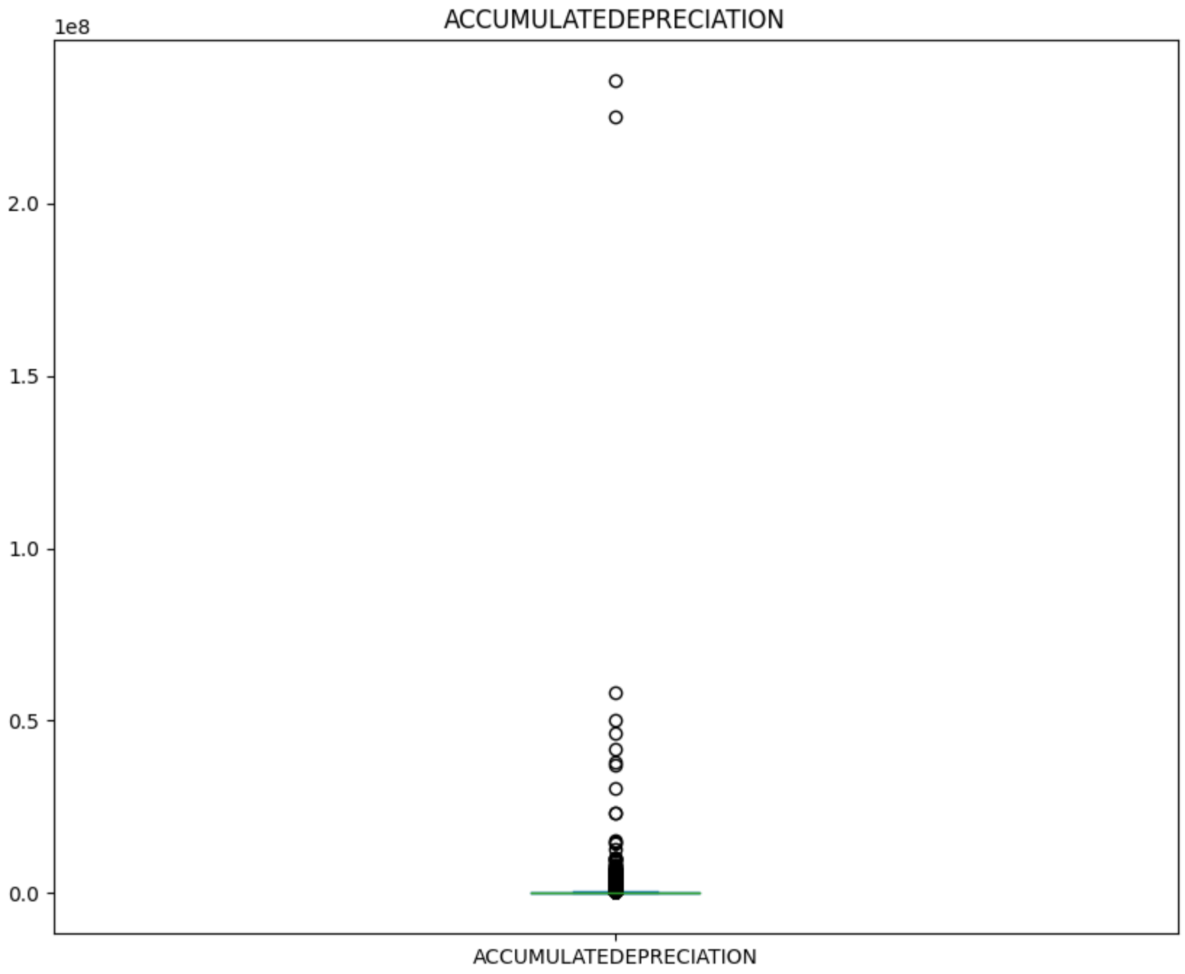
Out[2]:

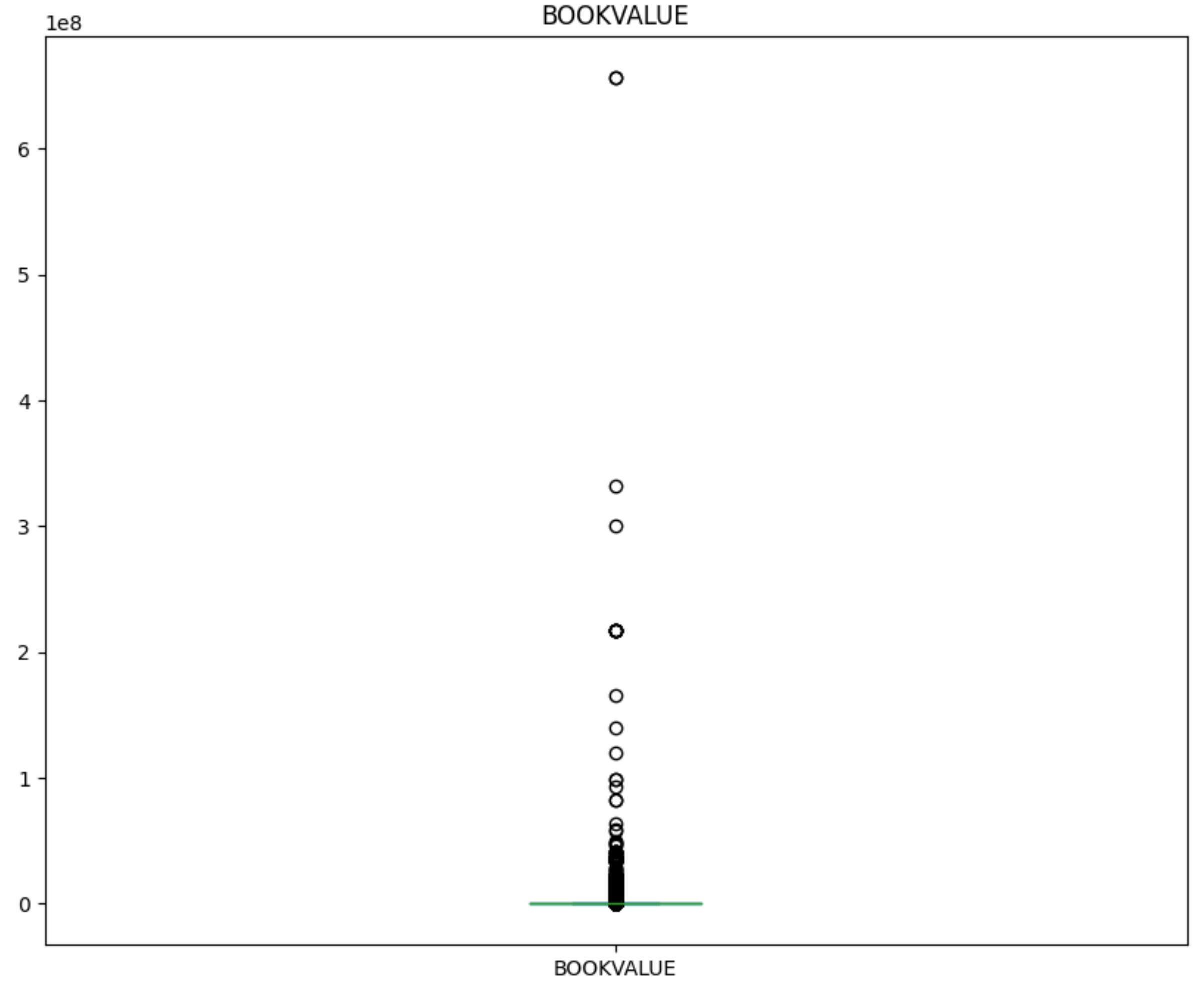
	range	min	max	mean	mode	median	
ACCUMULATEDEPRECIATION	2.356000e+08	0.0	2.356000e+08	4.098806e+05	0.0	0.0	-3
BOOKVALUE	6.564481e+08	0.0	6.564481e+08	6.981809e+05	1.0	1.0	-9
PRIMALVALUE	1.954043e+10	0.0	1.954043e+10	1.458704e+07	1.0	1880000.0	-2
DEPRECATION_PERIOD	1.000197e+06	2.0	1.000199e+06	2.500512e+05	2.0	2.0	-1

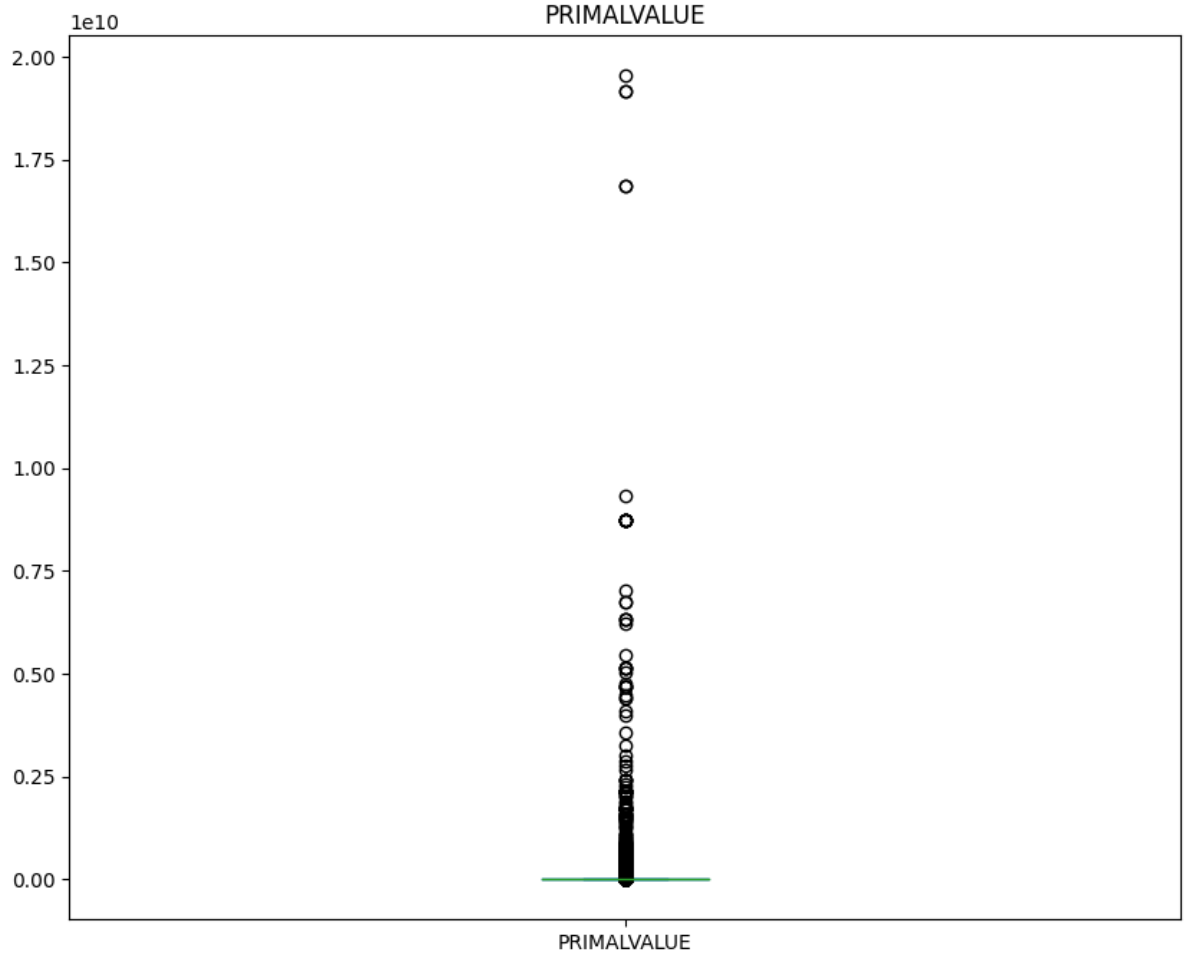
Now that we have the table we can draw our box plots:

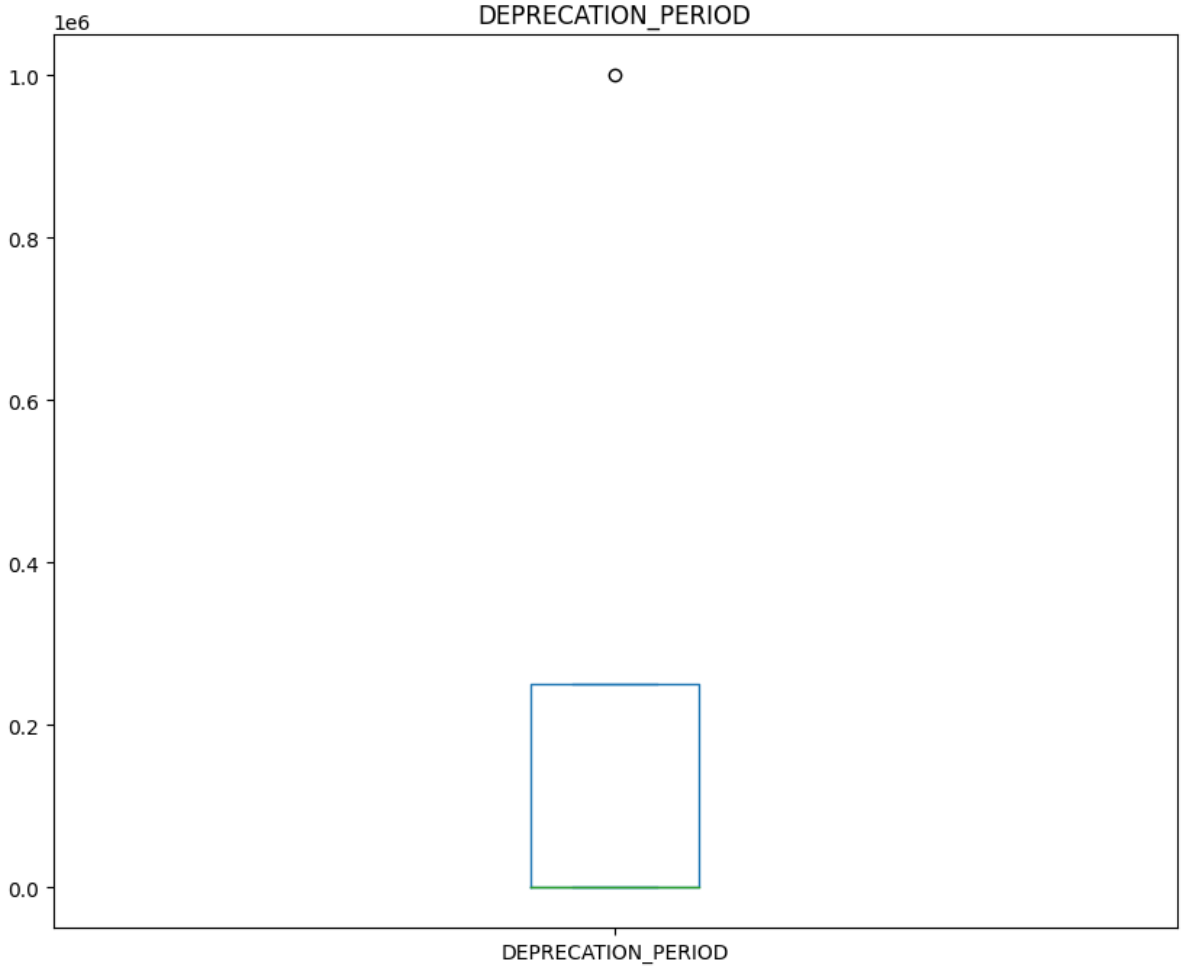
In [3]:

```
ogdf['ACCUMULATEDEPRECIATION'].plot(kind = 'box' , title = 'ACCUMULATEDEPRECIATION'
plt.show()
ogdf['BOOKVALUE'].plot(kind = 'box' , title = 'BOOKVALUE' , figsize = (10,8))
plt.show()
ogdf['PRIMALVALUE'].plot(kind = 'box' , title = 'PRIMALVALUE' , figsize = (10,8))
plt.show()
ogdf['DEPRECATION_PERIOD'].plot(kind = 'box' , title = 'DEPRECATION_PERIOD' , figsi
plt.show()
```




BOOKVALUE
1e8
6
5
4
3
2
1
0
BOOKVALUE


PRIMALVALUE
1e10
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
PRIMALVALUE



PRODUCTINSTANCE Dataset

Here is our biggest dataset but is also as messy and although we have lots and lots of attributes most of them doesn't have concistancy in their data types but we managed to find some that are consistent and also numerical! here they are:

In [4]:

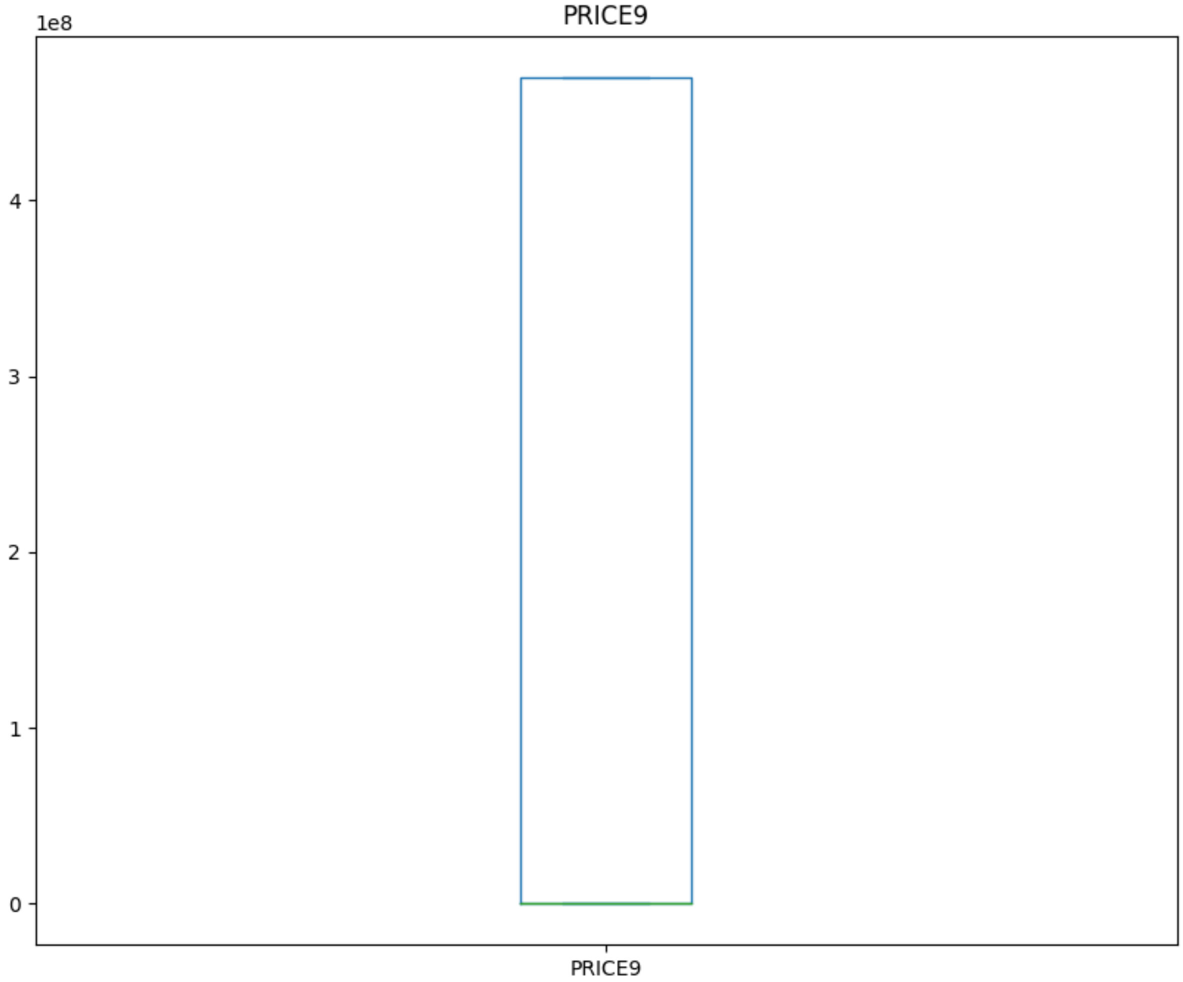
```
import pandas as pd
df = pd.read_csv('E:\ce\data mining\DataSets\PRODUCTINSTANCE.csv' , encoding='cp125
df = df[['PRICE9' , 'SALVAGEVALUE' , 'PI_VALUEAFTERCOEFFICIENTINC' , 'COEFFICIENTVA
       , 'PRESENTVALUE' , 'BOOKVALUE' , 'AREA_TOTAL']]
ogdf = df
median_df = df.median()
mode_df = df.mode()
df = df.describe()
df = df.transpose()
df['range'] = df['max'] - df['min']
df['Min acceptable value'] = df['min'] - ((df['max'] - df['min']) * 1.5)
df['Max acceptable value'] = df['max'] + ((df['max'] - df['min']) * 1.5)
df['median'] = median_df
df['mode'] = mode_df.head(1).transpose()
df = df.drop(columns = ['std' ,'count'])
df = df[['range' , 'min' , 'max' , 'mean' ,'mode' , 'median' , 'Min acceptable valu
df
```

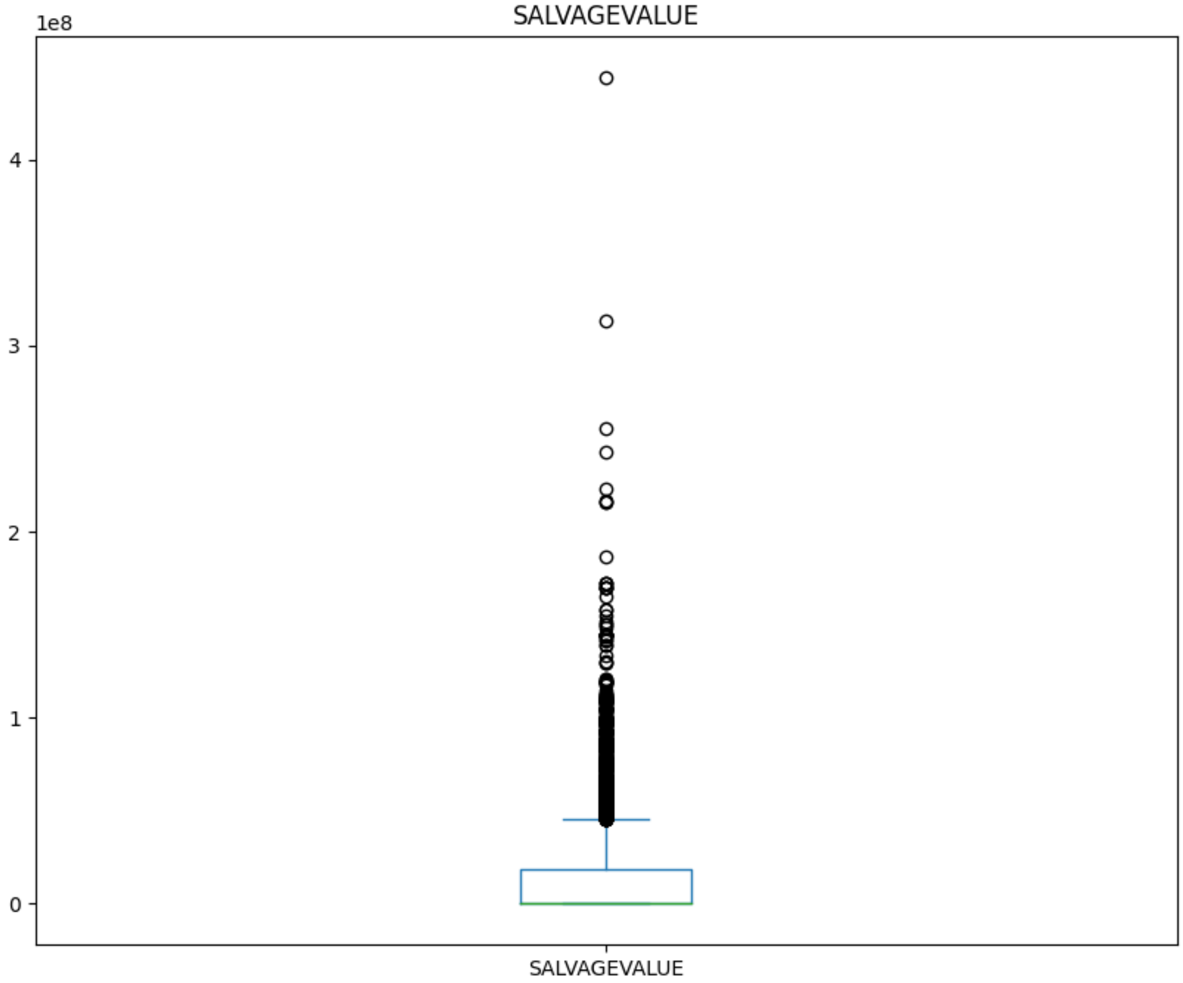
Out[4]:

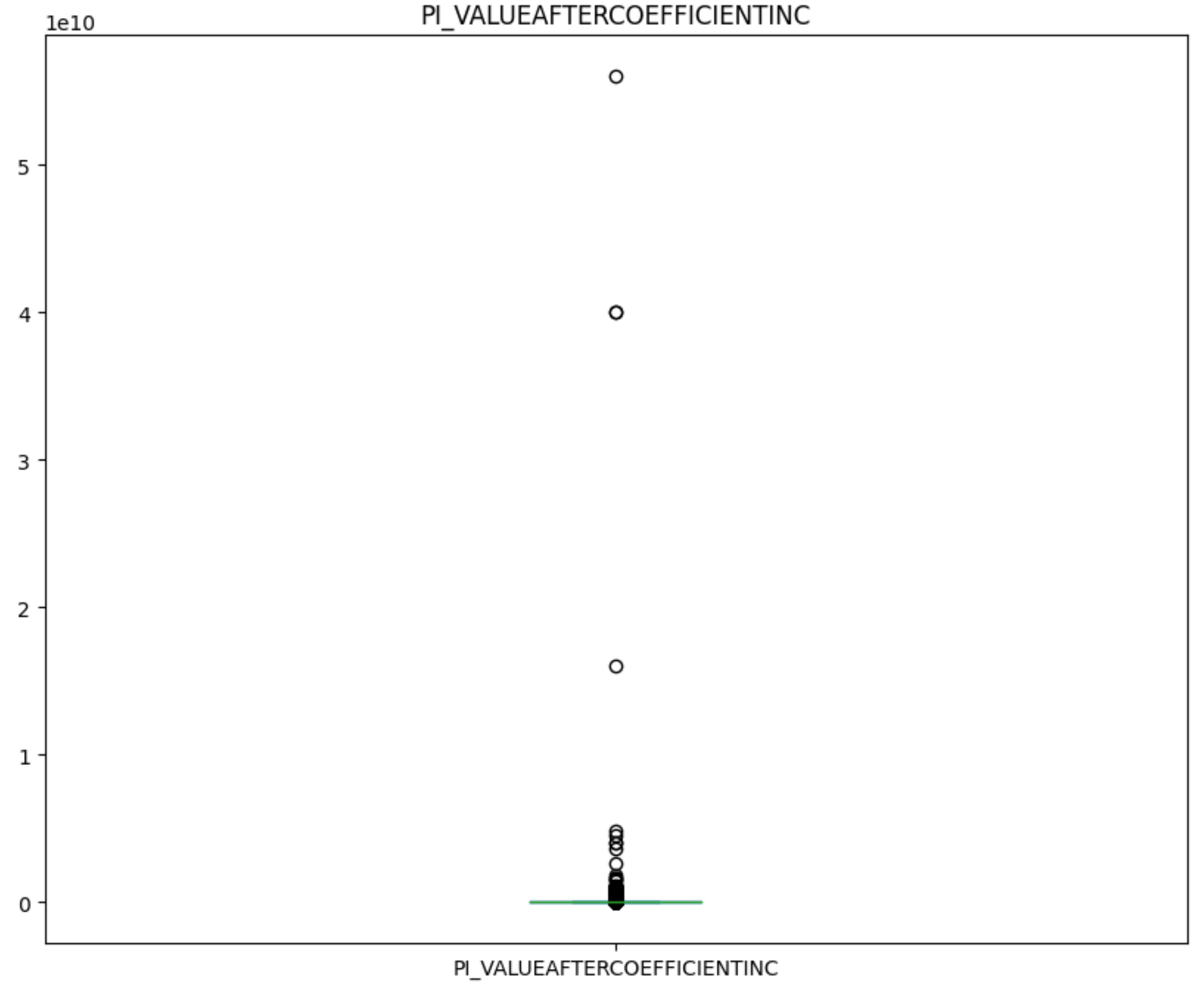
	range	min	max	mean	mode	
PRICE9	4.696381e+08	0.0	4.696381e+08	1.431538e+08	1304.0	
SALVAGEVALUE	4.438800e+08	0.0	4.438800e+08	1.377224e+07	0.0	
PI_VALUEAFTERCOEFFICIENTINC	5.600000e+10	0.0	5.600000e+10	9.521633e+07	0.0	12
COEFFICIENTVALUE	5.000000e+00	0.0	5.000000e+00	1.212077e+00	1.0	
ANTIQUITYCOEFFICIENT	5.000000e+00	0.0	5.000000e+00	1.242834e+00	1.0	
PRESENTVALUE	7.854600e+07	0.0	7.854600e+07	1.590321e+06	171000.0	4
BOOKVALUE	1.620218e+10	0.0	1.620218e+10	5.706507e+06	1.0	
AREA_TOTAL	2.123433e+09	-104592.0	2.123328e+09	7.150833e+04	250.0	

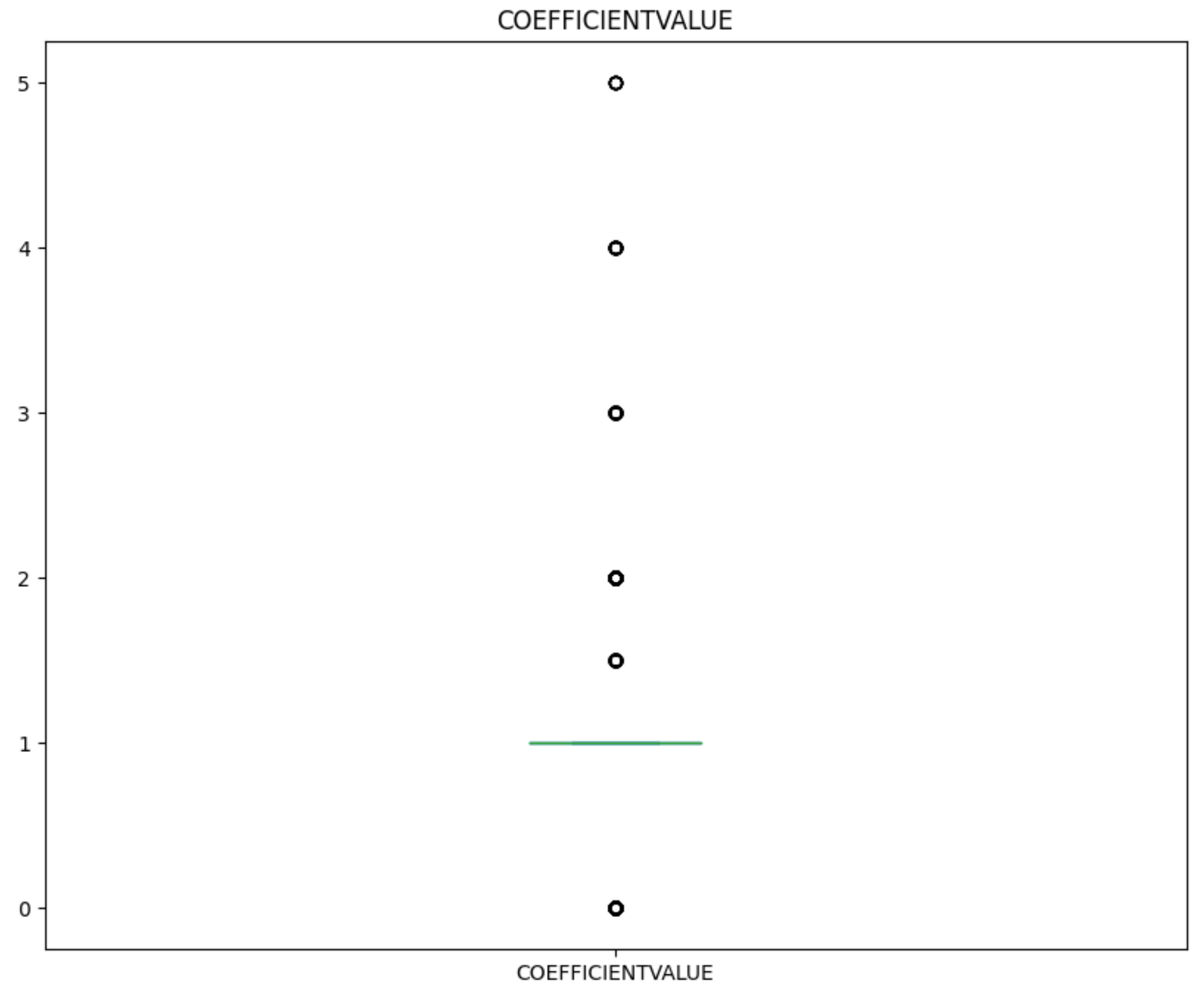
Now that we have the table we can draw our box plots:

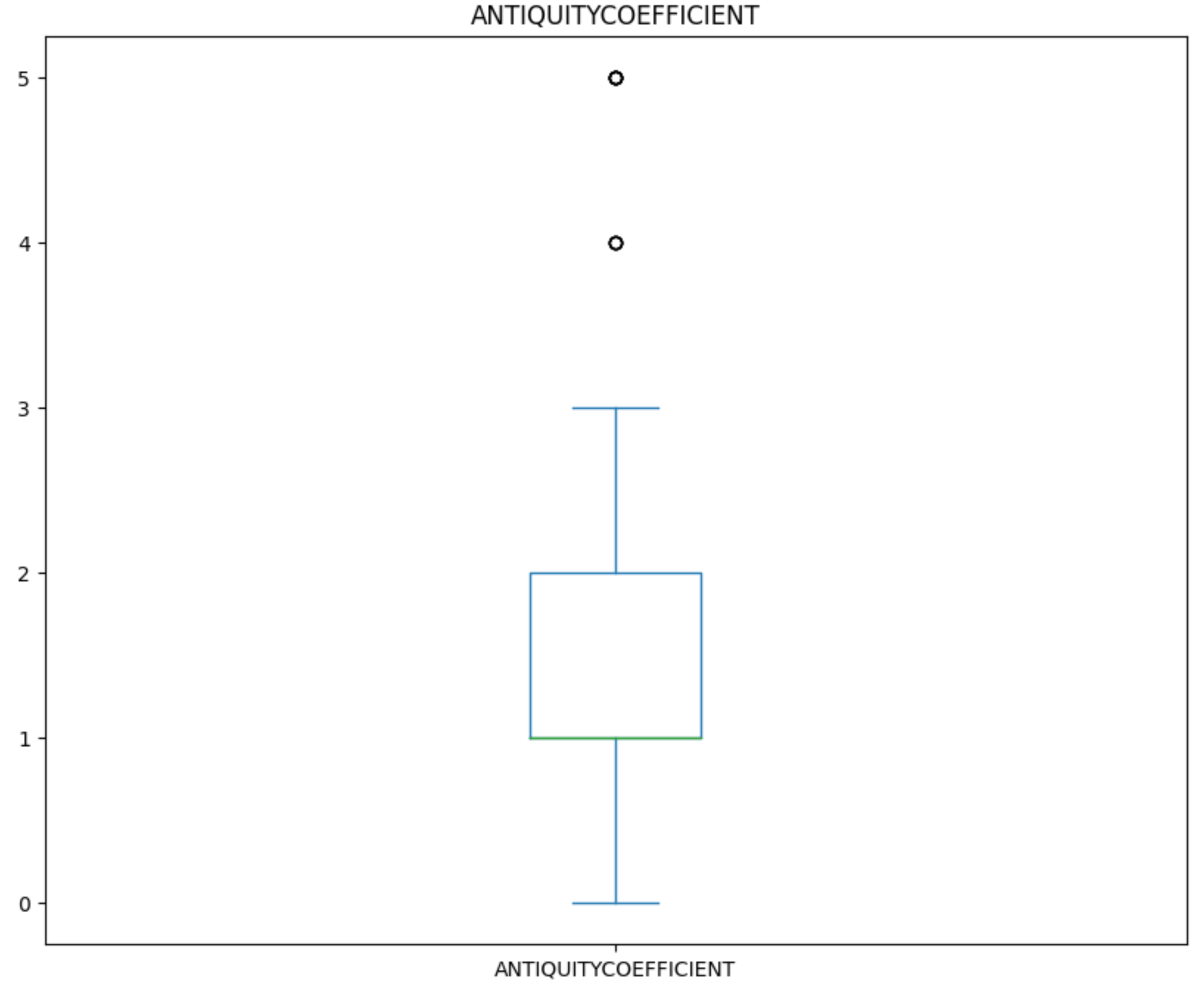
In [5]:
```
ogdf['PRICE9'].plot(kind = 'box' , title = 'PRICE9' , figsize = (10,8))
plt.show()
ogdf['SALVAGEVALUE'].plot(kind = 'box' , title = 'SALVAGEVALUE' , figsize = (10,8))
plt.show()
ogdf['PI_VALUEAFTERCOEFFICIENTINC'].plot(kind = 'box' , title = 'PI_VALUEAFTERCOEFF
plt.show()
ogdf['COEFFICIENTVALUE'].plot(kind = 'box' , title = 'COEFFICIENTVALUE' , figsize =
plt.show()
ogdf['ANTIQUITYCOEFFICIENT'].plot(kind = 'box' , title = 'ANTIQUITYCOEFFICIENT' , f
plt.show()
ogdf['PRESENTVALUE'].plot(kind = 'box' , title = 'PRESENTVALUE' , figsize = (10,8))
plt.show()
ogdf['BOOKVALUE'].plot(kind = 'box' , title = 'BOOKVALUE' , figsize = (10,8))
plt.show()
ogdf['AREA_TOTAL'].plot(kind = 'box' , title = 'AREA_TOTAL' , figsize = (10,8))
plt.show()
```

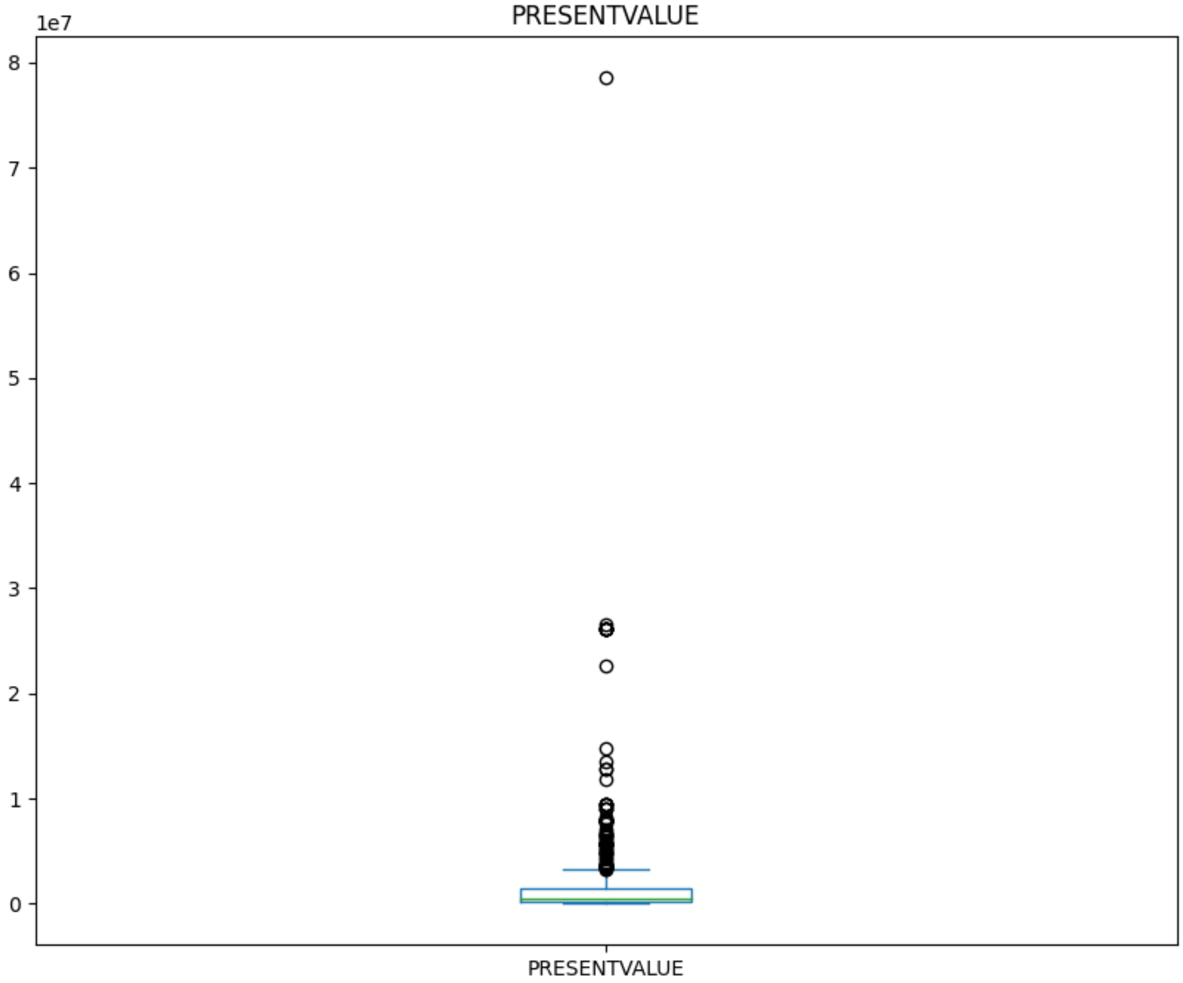

PRICE9
1e8
4
3
2
1
0
PRICE9

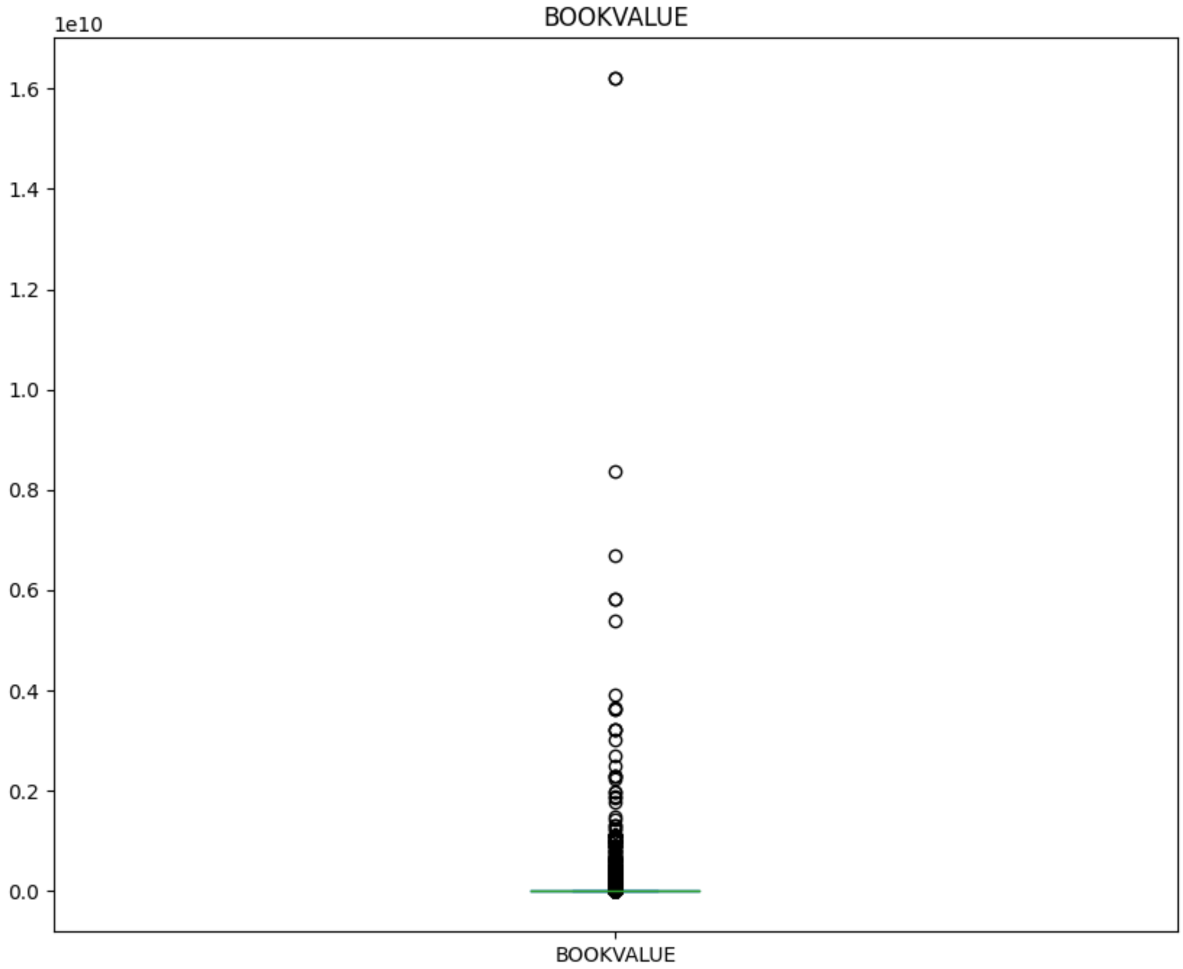

SALVAGEVALUE
1e8
4
3
2
1
0
SALVAGEVALUE

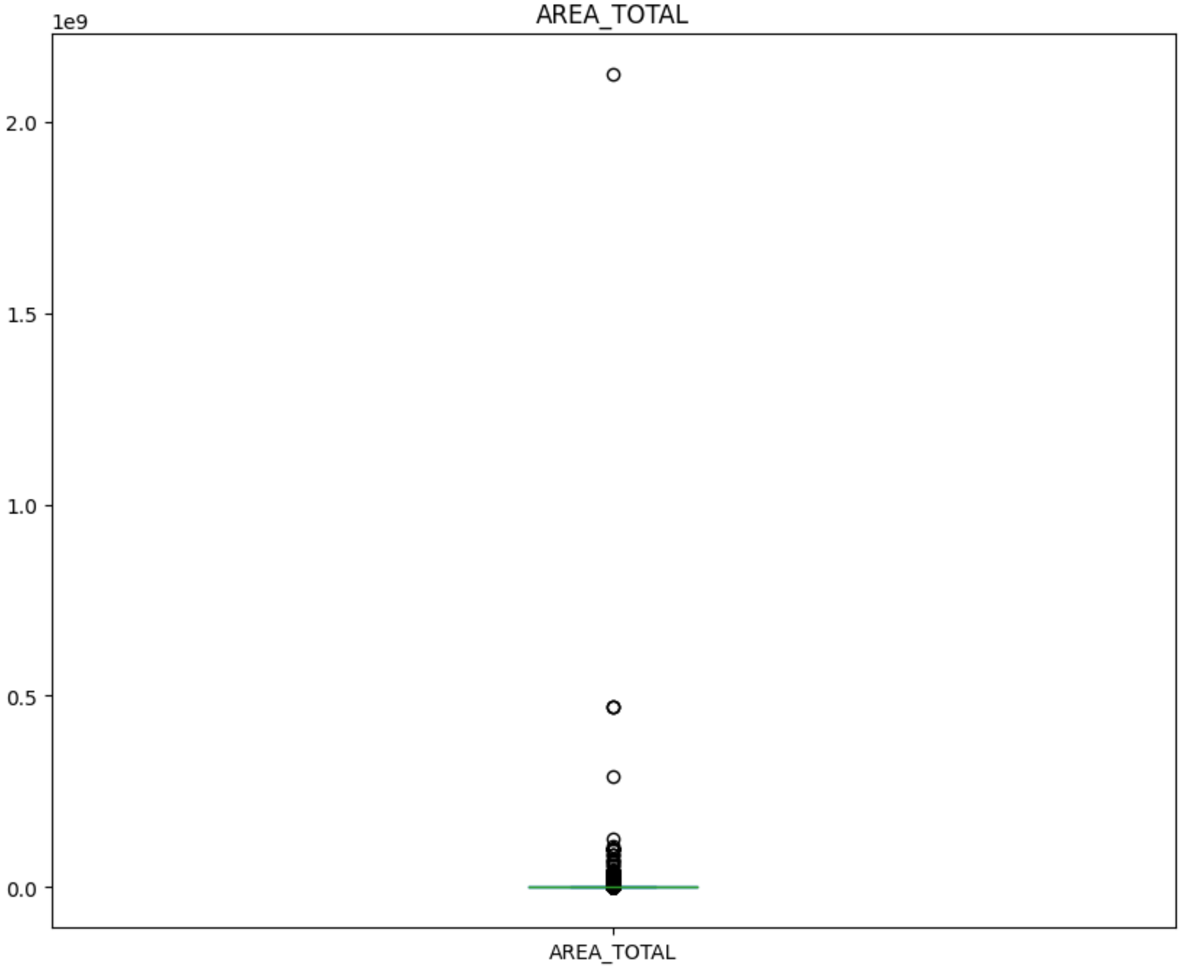

PI_VALUEAFTERCOEFFICIENTINC
1e10
5
4
3
2
1
0
PI_VALUEAFTERCOEFFICIENTINC


COEFFICIENTVALUE
5
4
3
2
1
0
COEFFICIENTVALUE


ANTIQUITYCOEFFICIENT
5
4
3
2
1
0
ANTIQUITYCOEFFICIENT


PRESENTVALUE
1e7
8
7
6
5
4
3
2
1
0
PRESENTVALUE


BOOKVALUE
1e10
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
BOOKVALUE



TRANSFER_ITEM and TRANSFER_ITEM_D Datasets

Both these Datasets are like the first one and doesn't have any numeric attributes that we can work on.