In [15]:
```
import numpy as np
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import statsmodels.api as sm
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from scipy.cluster import hierarchy
```

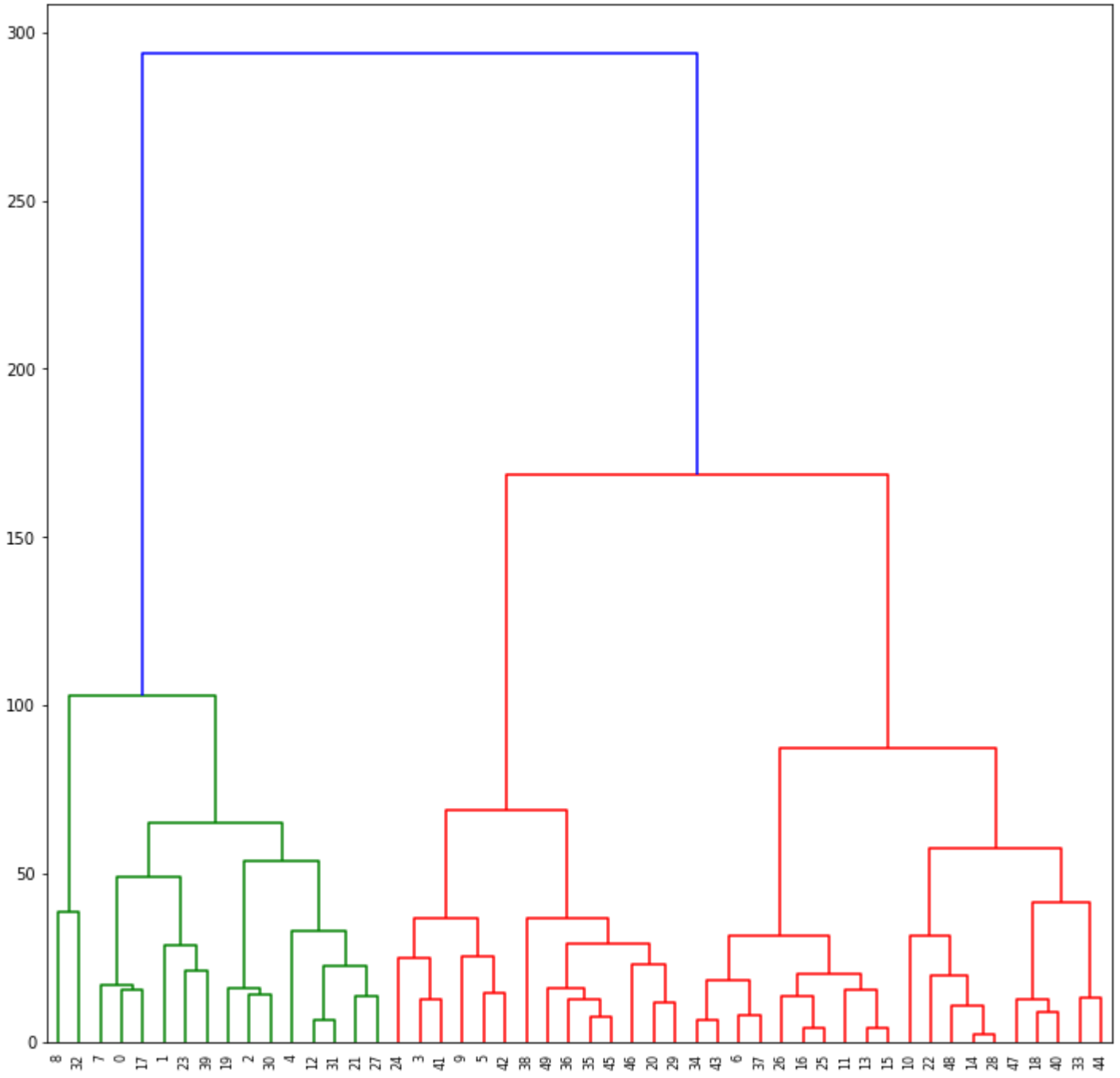
In [32]:
```
df = pd.read_csv('USArrests.csv')
df_copy = df.copy()
df_copy.rename(columns={'Unnamed: 0':'State'}, inplace=True)
df_copy.head()
```

Out[32]:

	State	Murder	Assault	UrbanPop	Rape
0	Alabama	13.2	236	58	21.2
1	Alaska	10.0	263	48	44.5
2	Arizona	8.1	294	80	31.0
3	Arkansas	8.8	190	50	19.5
4	California	9.0	276	91	40.6

Problem a

In [20]:
```
X = np.asarray(df_copy.drop('State',1))
link_tree = hierarchy.linkage(X, method='complete', metric='euclidean')
plt.figure(figsize=(12,12))
dn = hierarchy.dendrogram(link_tree)
```


300
250
200
150
100
50
0
8 32 7 0 17 1 23 39 19 2 30 4 12 31 21 27 24 3 41 9 5 42 38 49 36 35 45 46 20 29 34 43 6 37 26 16 25 11 13 15 10 22 48 14 28 47 18 40 33 44

Problem b

In [38]:
```
# Draw line at y=150
```

Problem c

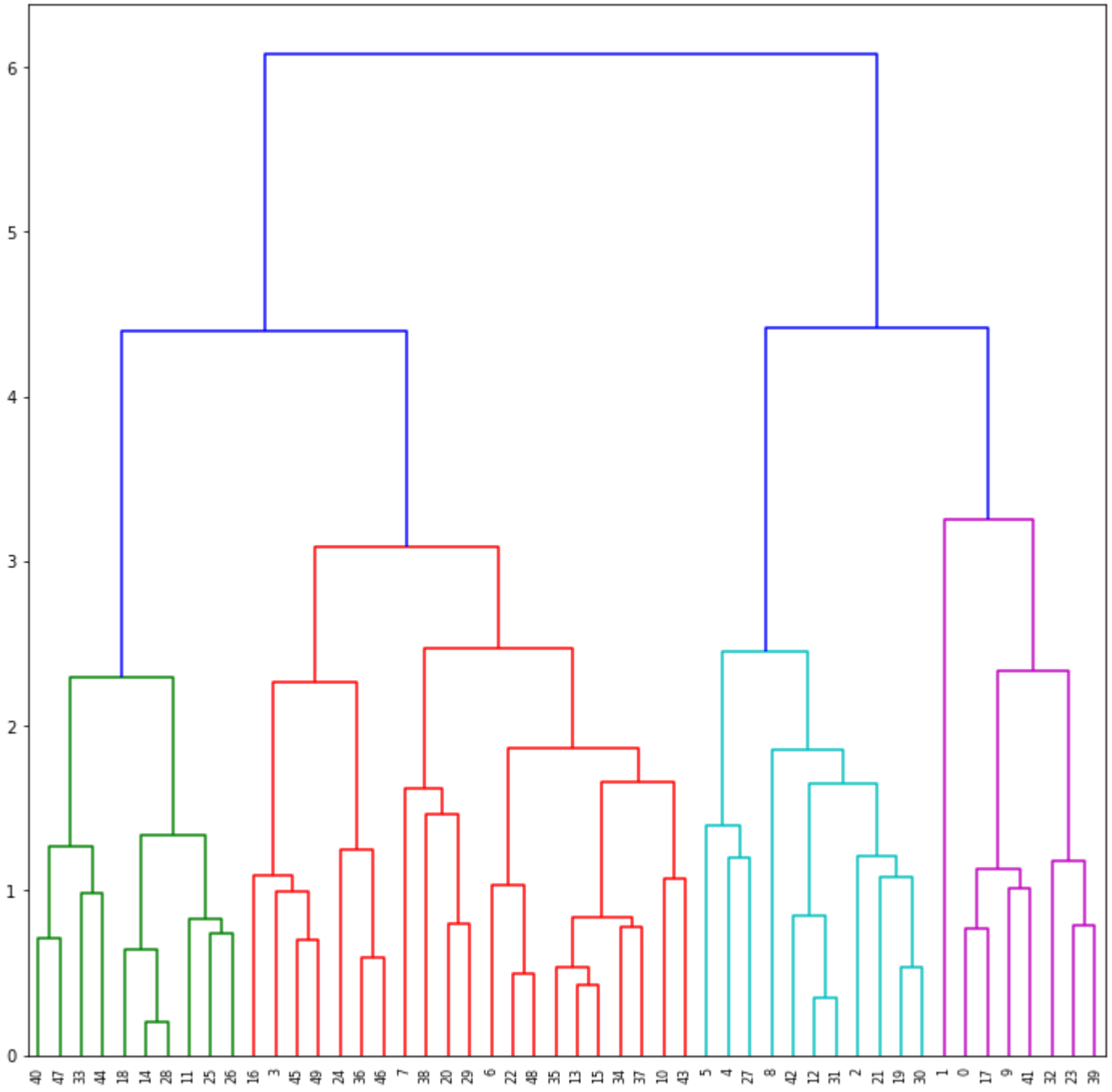
In [34]:
```
# Scale variables to have standard deviation of 1
murder_std = df_copy['Murder'].std()
df_copy['Murder'] = df_copy['Murder']/murder_std

assault_std = df_copy['Assault'].std()
df_copy['Assault'] = df_copy['Assault']/assault_std

urbanpop_std = df_copy['UrbanPop'].std()
df_copy['UrbanPop'] = df_copy['UrbanPop']/urbanpop_std

rape_std = df_copy['Rape'].std()
df_copy['Rape'] = df_copy['Rape']/rape_std
```

In [37]:
```
X = np.asarray(df_copy.drop('State',1))
link_tree = hierarchy.linkage(X, method='complete', metric='euclidean')
plt.figure(figsize=(12,12))
dn = hierarchy.dendrogram(link_tree)
```


6
5
4
3
2
1
0
40 47 33 44 18 14 28 11 25 26 16 3 45 49 24 36 46 7 38 20 29 6 22 48 35 13 15 34 37 10 43 5 4 27 8 42 12 31 2 21 19 30 1 0 17 9 41 32 23 39

In [ ]: