Clustering Categorical Data

Import the relevant libraries

In [4]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
from sklearn.cluster import KMeans
```

Load the data

Load data from the csv file: '*Categorical.csv*'.

In [5]:
```
# Load the data
raw_data = pd.read_csv('Categorical.csv')
# Check the data
raw_data
```

Out[5]:

	name	Longitude	Latitude	continent
0	Aruba	-69.982677	12.520880	North America
1	Afghanistan	66.004734	33.835231	Asia
2	Angola	17.537368	-12.293361	Africa
3	Anguilla	-63.064989	18.223959	North America
4	Albania	20.049834	41.142450	Europe
...	...	...	...	...
236	Samoa	-172.164851	-13.753243	Oceania
237	Yemen	47.586762	15.909280	Asia
238	South Africa	25.083901	-29.000341	Africa
239	Zambia	27.774759	-13.458242	Africa
240	Zimbabwe	29.851441	-19.004204	Africa

241 rows × 4 columns

Map the data

Use the '*continent*' category for this analysis.

In [6]:
```
data_mapped = data.copy()
data_mapped['continent'] = data_mapped['continent'].map({'North America':0,'Europe':1,'Asia':2,'Africa':3,'Sout
data_mapped
```

Out[6]:

	name	Longitude	Latitude	continent
0	Aruba	-69.982677	12.520880	0
1	Afghanistan	66.004734	33.835231	2
2	Angola	17.537368	-12.293361	3
3	Anguilla	-63.064989	18.223959	0
4	Albania	20.049834	41.142450	1
...	...	...	...	...
236	Samoa	-172.164851	-13.753243	5
237	Yemen	47.586762	15.909280	2
238	South Africa	25.083901	-29.000341	3
239	Zambia	27.774759	-13.458242	3
240	Zimbabwe	29.851441	-19.004204	3

241 rows × 4 columns

Select the features

In [7]:
```
x = data_mapped.iloc[:,3:4]
```

Clustering

In [8]:
```
kmeans = KMeans(4)
kmeans.fit(x)
```

Out[8]:
```
KMeans(n_clusters=4)
```

Use 4 clusters initially.

Clustering results

In [9]:
```
identified_clusters = kmeans.fit_predict(x)
identified_clusters
```

Out[9]:
```
array([1, 3, 0, 1, 1, 1, 1, 3, 0, 3, 2, 2, 2, 2, 1, 2, 1, 3, 0, 1, 0, 0,
       3, 1, 3, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 3, 3, 0, 0, 1, 0, 3, 0, 0,
       0, 0, 2, 0, 0, 0, 1, 1, 1, 1, 3, 3, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0,
       1, 1, 0, 1, 2, 0, 1, 1, 2, 0, 1, 3, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1,
       2, 0, 3, 2, 1, 1, 1, 1, 3, 1, 3, 3, 2, 1, 3, 3, 1, 3, 1, 1, 1, 3,
       3, 3, 3, 0, 3, 3, 2, 1, 3, 1, 3, 3, 3, 0, 0, 1, 1, 3, 0, 1, 1, 1,
       3, 1, 0, 1, 1, 0, 2, 1, 2, 1, 0, 1, 3, 1, 3, 2, 0, 0, 1, 2, 0, 3,
       0, 2, 0, 2, 0, 1, 2, 1, 1, 3, 2, 2, 3, 3, 1, 2, 0, 3, 2, 2, 1, 1,
       3, 1, 0, 3, 2, 3, 1, 1, 0, 0, 3, 0, 0, 0, 3, 2, 2, 2, 0, 1, 1, 0,
       0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 2, 3, 1, 0, 0, 3, 3, 3, 3, 2, 1, 0,
       3, 3, 0, 0, 1, 0, 1, 3, 1, 1, 0, 1, 1, 3, 2, 2, 2, 3, 0, 0, 0])
```

In [10]:
```
data_with_clusters = data_mapped.copy()
data_with_clusters['Cluster'] = identified_clusters
data_with_clusters
```

Out[10]:

	name	Longitude	Latitude	continent	Cluster
0	Aruba	-69.982677	12.520880	0	1
1	Afghanistan	66.004734	33.835231	2	3
2	Angola	17.537368	-12.293361	3	0
3	Anguilla	-63.064989	18.223959	0	1
4	Albania	20.049834	41.142450	1	1
...	...	...	...	...	...
236	Samoa	-172.164851	-13.753243	5	2
237	Yemen	47.586762	15.909280	2	3
238	South Africa	25.083901	-29.000341	3	0
239	Zambia	27.774759	-13.458242	3	0
240	Zimbabwe	29.851441	-19.004204	3	0

241 rows × 5 columns

Plot the data

In [11]:
```
plt.scatter(data['Longitude'], data['Latitude'], c=data_with_clusters['Cluster'], cmap = 'rainbow')
plt.xlim(-180,180)
plt.ylim(-90, 90)
plt.show()
```