Clustering with Countries_exercise.csv

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

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

Out[5]:

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

241 rows × 3 columns

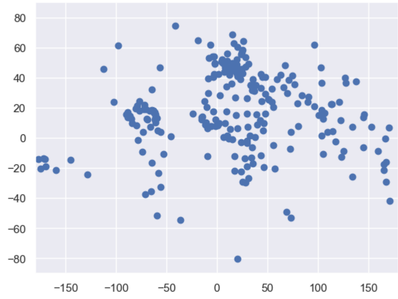
Remove the duplicate index column from the dataset.

In [6]:
```
data = raw_data.copy()
```

Plot the data

Plot the *'Longtitude'* and *'Latitude'* columns.

In [7]:
```
plt.scatter(data['Longitude'], data['Latitude'])
plt.xlim(-180,180)
plt.ylim(-90, 90)
plt.show()
```



Select the features

Create a copy of that data and remove all parameters apart from *Longitude* and *Latitude*.

In [8]:
```
x = data.iloc[:,1:3]
x
```

Out[8]:

	Longitude	Latitude
0	-69.982677	12.520880
1	66.004734	33.835231
2	17.537368	-12.293361
3	-63.064989	18.223959
4	20.049834	41.142450
...	...	...
236	-172.164851	-13.753243
237	47.586762	15.909280
238	25.083901	-29.000341
239	27.774759	-13.458242
240	29.851441	-19.004204

241 rows × 2 columns

Clustering

I'm changing index in kmeans() then running the remaining kernels.

In [9]:
```
kmeans = KMeans(7)
```

In [10]:
```
kmeans.fit(x)
```

Out[10]:
```
KMeans(n_clusters=7)
```

Clustering Resutls

In [11]:
```
identified_clusters = kmeans.fit_predict(x)
identified_clusters
```

Out[11]:
```
array([1, 0, 3, 1, 4, 4, 4, 0, 1, 0, 6, 3, 2, 3, 1, 2, 4, 0, 3, 4, 5, 5,
       0, 4, 0, 1, 1, 4, 1, 4, 1, 1, 1, 1, 1, 2, 0, 3, 5, 4, 1, 0, 5, 5,
       3, 5, 6, 1, 3, 5, 1, 1, 1, 1, 4, 4, 4, 4, 3, 1, 4, 1, 5, 1, 4, 3,
       4, 4, 3, 4, 2, 1, 4, 4, 2, 5, 4, 0, 4, 5, 5, 5, 5, 5, 4, 1, 4, 1,
       2, 1, 2, 3, 1, 4, 1, 4, 2, 4, 0, 2, 3, 4, 0, 0, 4, 4, 4, 1, 4, 4,
       2, 0, 0, 3, 0, 2, 1, 1, 2, 4, 0, 2, 4, 5, 4, 1, 4, 0, 3, 4, 4, 4,
       2, 1, 5, 4, 4, 3, 0, 1, 2, 4, 5, 4, 0, 4, 0, 2, 3, 5, 1, 3, 3, 2,
       3, 2, 5, 2, 5, 1, 6, 4, 4, 0, 2, 2, 0, 0, 1, 6, 1, 2, 2, 2, 4, 1,
       2, 4, 1, 4, 6, 0, 4, 0, 3, 5, 0, 5, 3, 5, 2, 1, 5, 2, 5, 1, 4, 3,
       3, 1, 4, 5, 1, 4, 4, 4, 3, 1, 3, 4, 1, 5, 5, 2, 0, 0, 2, 6, 1, 4,
       4, 2, 3, 3, 4, 1, 1, 0, 4, 1, 1, 1, 1, 2, 2, 6, 6, 0, 3, 3, 3])
```

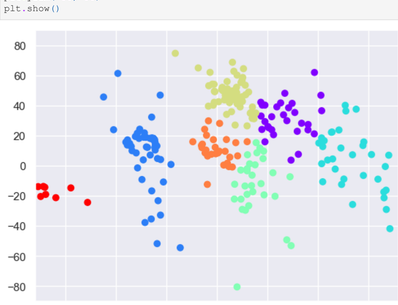
In [12]:
```
data_with_clusters = data.copy()
data_with_clusters['Cluster'] = identified_clusters
data_with_clusters
```

Out[12]:

	name	Longitude	Latitude	Cluster
0	Aruba	-69.982677	12.520880	1
1	Afghanistan	66.004734	33.835231	0
2	Angola	17.537368	-12.293361	3
3	Anguilla	-63.064989	18.223959	1
4	Albania	20.049834	41.142450	4
...	...	...	...	...
236	Samoa	-172.164851	-13.753243	6
237	Yemen	47.586762	15.909280	0
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

In [13]:
```
plt.scatter(data['Longitude'], data['Latitude'],c=data_with_clusters['Cluster'], cmap = 'rainbow')
plt.xlim(-180,180)
plt.ylim(-90, 90)
plt.show()
```



In [ ]:
```
```