Basics of cluster analysis

In this notebook we explore the very basics of cluster analysis with k-means

Import the relevant libraries

In [12]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Set the styles to Seaborn
sns.set()

# k-means clustering with sklearn
from sklearn.cluster import KMeans
```

Load the data

In [13]:
```
# Load the country clusters data
data = pd.read_csv('Country_clusters.csv')
```

In [14]:
```
# Check out the data manually
data
```

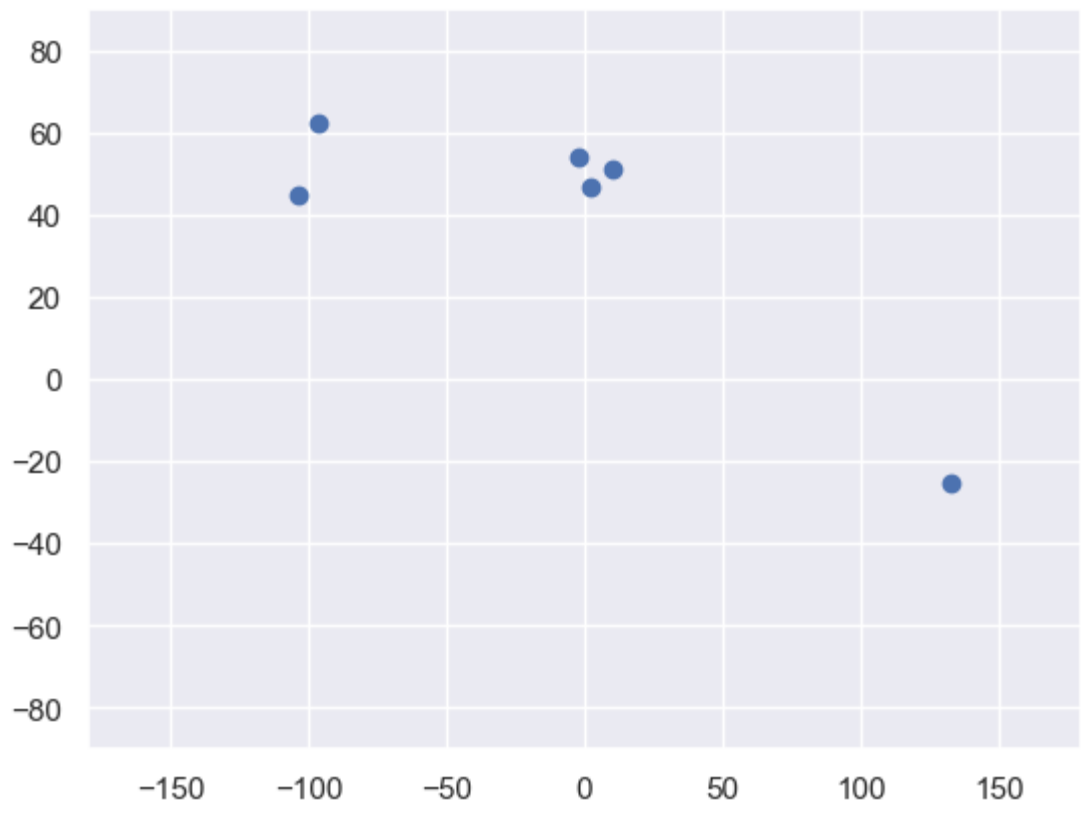
Out[14]:

	Country	Latitude	Longitude	Language
0	USA	44.97	-103.77	English
1	Canada	62.40	-96.80	English
2	France	46.75	2.40	French
3	UK	54.01	-2.53	English
4	Germany	51.15	10.40	German
5	Australia	-25.45	133.11	English

Plot the data

In [15]:
```
plt.scatter(data['Longitude'],data['Latitude'])
# Set limits of the axes, again to resemble the world map
plt.xlim(-180,180)
plt.ylim(-90,90)
plt.show
```

Out[15]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



Select the features

In [16]:
```
x = data.iloc[:,1:3] # choosing columns 1 and 2
```

In [17]:
```
x
```

Out[17]:

	Latitude	Longitude
0	44.97	-103.77
1	62.40	-96.80
2	46.75	2.40
3	54.01	-2.53
4	51.15	10.40
5	-25.45	133.11

Clustering

This is the part of the sheet which deals with the actual clustering

In [18]:
```
kmeans = KMeans(3)
```

In [19]:
```
# Fit the input data, i.e. cluster the data in X in K clusters
kmeans.fit(x)
```

Out[19]:
```
KMeans(n_clusters=3)
```

Clustering results

There are many ways to do this part, we found this to be the most illustrative one

In [20]:
```
# Create a variable which will contain the predicted clusters for each observation
identified_clusters = kmeans.fit_predict(x)
# Check the result
identified_clusters
```

Out[20]:
```
array([2, 2, 0, 0, 0, 1])
```

In [21]:
```
# Create a copy of the data
data_with_clusters = data.copy()
# Create a new Series, containing the identified cluster for each observation
data_with_clusters['Cluster'] = identified_clusters
# Check the result
data_with_clusters
```

Out[21]:

	Country	Latitude	Longitude	Language	Cluster
0	USA	44.97	-103.77	English	2
1	Canada	62.40	-96.80	English	2
2	France	46.75	2.40	French	0
3	UK	54.01	-2.53	English	0
4	Germany	51.15	10.40	German	0
5	Australia	-25.45	133.11	English	1

In [22]:
```
plt.scatter(data_with_clusters['Longitude'],data_with_clusters['Latitude'],c=data_with_clusters['Cluster'],cmap
plt.xlim(-180,180)
plt.ylim(-90,90)
plt.show()
```