Sec_1_Homework_9

March 8, 2024

1 0.) Import and Clean data

[62]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
```

[ ]:
```

```

[63]:
```
#drive.mount('/content/gdrive/', force_remount = True)
df = pd.read_csv("Country-data.csv", sep = ",")
```

[64]:
```
df.head()
```

[64]:
```
               country  child_mort  exports  health  imports  income  \
0          Afghanistan        90.2     10.0    7.58     44.9    1610
1              Albania        16.6     28.0    6.55     48.6    9930
2              Algeria        27.3     38.4    4.17     31.4   12900
3               Angola       119.0     62.3    2.85     42.9    5900
4  Antigua and Barbuda        10.3     45.5    6.03     58.9   19100

   inflation  life_expec  total_fer   gdpp
0       9.44        56.2       5.82    553
1       4.49        76.3       1.65   4090
2      16.10        76.5       2.89   4460
3      22.40        60.1       6.16   3530
4       1.44        76.8       2.13  12200
```

[65]:
```
naame = df[['country']].copy()
X = df.drop('country', axis = 1)
```

[66]:
```
scaler = StandardScaler().fit(X)
X_scaled = scaler.transform(X)
```

2 1.) Fit a kmeans Model with any Number of Clusters

[67]:
```
KMeans?
```

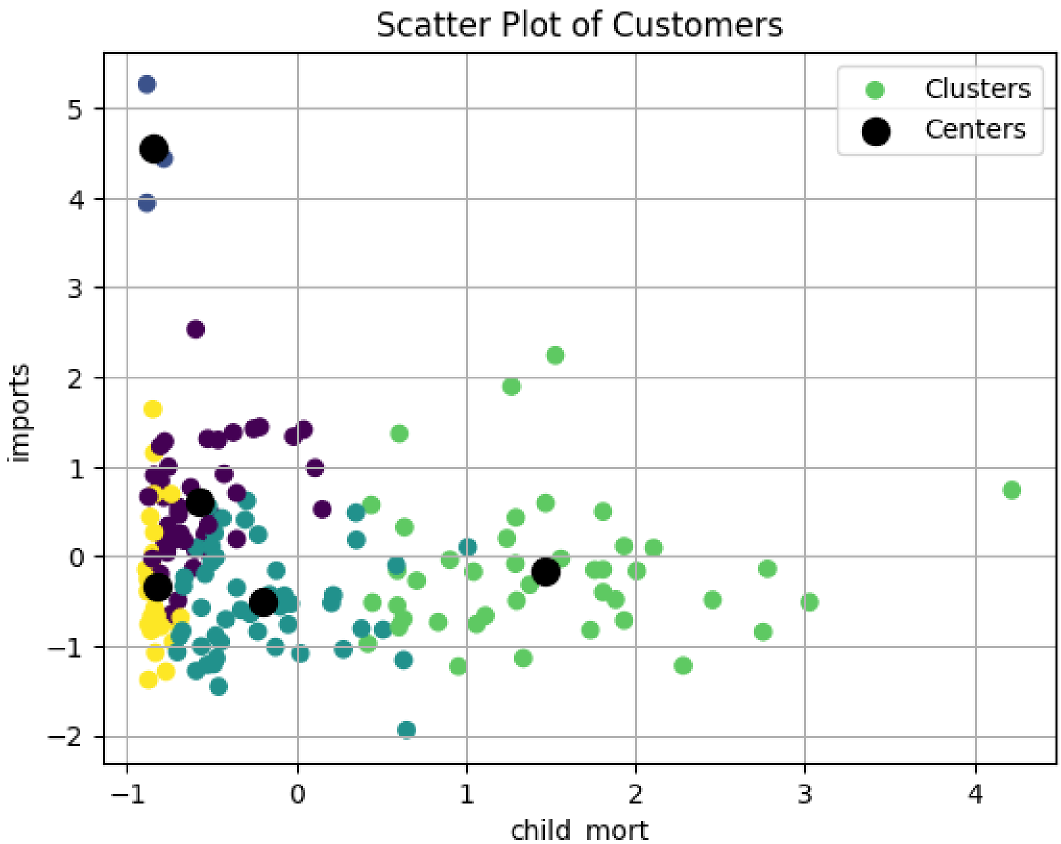
[68]:
```
kmeans = KMeans(n_clusters = 5).fit(X_scaled)
```

3 2.) Pick two features to visualize across

[69]:
```
X.columns
```

[69]:
```
Index(['child_mort', 'exports', 'health', 'imports', 'income', 'inflation',
       'life_expec', 'total_fer', 'gdpp'],
      dtype='object')
```

[70]:
```
import matplotlib.pyplot as plt

x1_index = 0
x2_index = 3


scatter = plt.scatter(X_scaled[:, x1_index], X_scaled[:, x2_index], c=kmeans.
 ↪labels_, cmap='viridis', label='Clusters')


centers = plt.scatter(kmeans.cluster_centers_[:, x1_index], kmeans.
 ↪cluster_centers_[:, x2_index], marker='o', color='black', s=100,␣
 ↪label='Centers')

plt.xlabel(X.columns[x1_index])
plt.ylabel(X.columns[x2_index])
plt.title('Scatter Plot of Customers')

# Generate legend
plt.legend()

plt.grid()
plt.show()
```



[ ]:

[ ]:

4 3.) Check a range of k-clusters and visualize to find the elbow. Test 30 different random starting places for the centroid means

[71]:
```
WCSSs = []
Ks = range(1, 15)
for k in Ks:
    kmeans = KMeans(n_clusters = k, n_init = 30).fit(X_scaled)
    WCSSs.append(kmeans.inertia_)
```

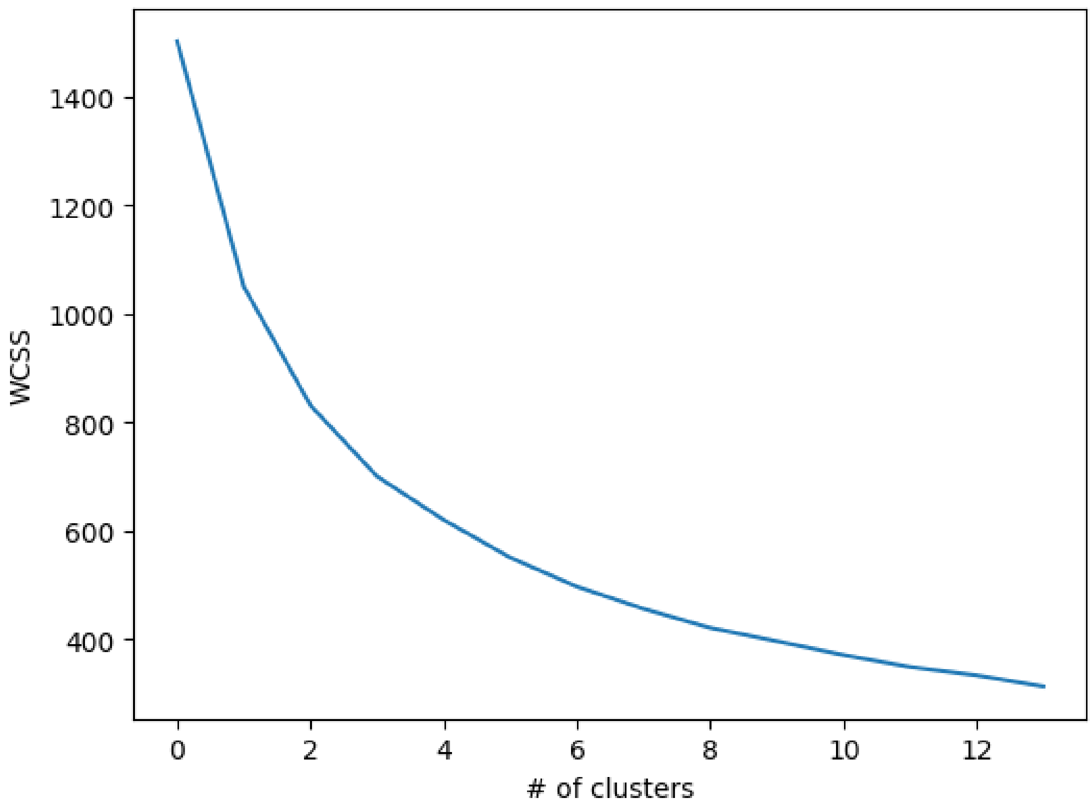
[ ]:

[ ]:

5 4.) Use the above work and economic critical thinking to choose a number of clusters. Explain why you chose the number of clusters and fit a model accordingly.

- The reduction in WCSS begins to plateau around 2 clusters, making it a suitable choice for capturing the most significant variation between different economic segments.
- Choosing 2 clusters is economically sensible for distinguishing broadly between more developed and less developed countries, aligning with traditional global economic divisions.
- Opting for 2 clusters would simplify the model and potentially enhance interpretability, focusing on the primary divide in global economic development status.

[72]:
```
plt.plot(WCSSs)
plt.ylabel('WCSS')
plt.xlabel('# of clusters')
plt.show()
```



[ ]:

[ ]:

[ ]:

[ ]:

```

```

6 6.) Do the same for a silhoutte plot

[73]:

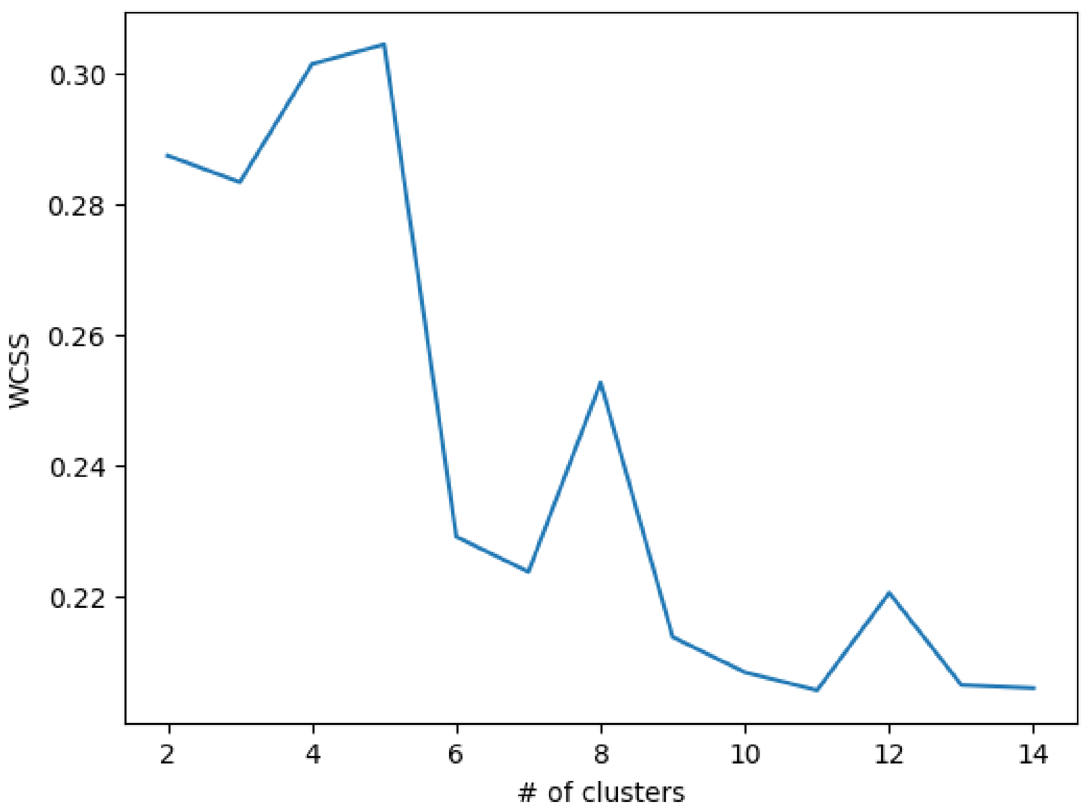
```
from sklearn.metrics import silhouette_score
```

[74]:

```
SSs = []
Ks = range(2, 15)
for k in Ks:
    kmeans = KMeans(n_clusters = k, n_init = 30).fit(X_scaled)
    sil = silhouette_score(X_scaled, kmeans.labels_)
    SSs.append(sil)
```

[75]:

```
plt.plot(Ks, SSs)
plt.ylabel('WCSS')
plt.xlabel('# of clusters')
plt.show()
```



7 7.) Create a list of the countries that are in each cluster. Write interesting things you notice.

- Cluster 1 is composed mainly of developing countries from Africa, Asia, and some from the Pacific and South America, corresponding with lower economic indicators.
- Cluster 2 contains a diverse set of developed countries from all continents with higher economic and health indicators.
- The clusters likely reflect the global economic divide between the more developed “Global North” and the less developed “Global South.”

[76]:
```
kmeans = KMeans(n_clusters = 2, n_init = 30).fit(X_scaled)
```

[77]:
```
preds = pd.DataFrame(kmeans.labels_)
```

[78]:
```
output = pd.concat([preds, df], axis = 1)
```

[79]:
```
print('Cluster 1: ')
list(output.loc[output[0] == 0, 'country'])
```

```
Cluster 1:
```

[79]:
```
['Afghanistan',
 'Angola',
 'Bangladesh',
 'Benin',
 'Bolivia',
 'Botswana',
 'Burkina Faso',
 'Burundi',
 'Cambodia',
 'Cameroon',
 'Central African Republic',
 'Chad',
 'Comoros',
 'Congo, Dem. Rep.',
 'Congo, Rep.',
 "Cote d'Ivoire",
 'Egypt',
 'Equatorial Guinea',
 'Eritrea',
 'Gabon',
 'Gambia',
 'Ghana',
 'Guatemala',
 'Guinea',
 'Guinea-Bissau',
 'Guyana',
 'Haiti',
```

```
 'India',
 'Indonesia',
 'Iraq',
 'Kenya',
 'Kiribati',
 'Kyrgyz Republic',
 'Lao',
 'Lesotho',
 'Liberia',
 'Madagascar',
 'Malawi',
 'Mali',
 'Mauritania',
 'Micronesia, Fed. Sts.',
 'Mongolia',
 'Mozambique',
 'Myanmar',
 'Namibia',
 'Nepal',
 'Niger',
 'Nigeria',
 'Pakistan',
 'Philippines',
 'Rwanda',
 'Samoa',
 'Senegal',
 'Sierra Leone',
 'Solomon Islands',
 'South Africa',
 'Sudan',
 'Tajikistan',
 'Tanzania',
 'Timor-Leste',
 'Togo',
 'Tonga',
 'Turkmenistan',
 'Uganda',
 'Uzbekistan',
 'Vanuatu',
 'Yemen',
 'Zambia']
```

[80]:
```
print('Cluster 2: ')
list(output.loc[output[0] == 1, 'country'])
```

```
Cluster 2:
```

[80]:
```
['Albania',
 'Algeria',
 'Antigua and Barbuda',
 'Argentina',
 'Armenia',
 'Australia',
 'Austria',
 'Azerbaijan',
 'Bahamas',
 'Bahrain',
 'Barbados',
 'Belarus',
 'Belgium',
 'Belize',
 'Bhutan',
 'Bosnia and Herzegovina',
 'Brazil',
 'Brunei',
 'Bulgaria',
 'Canada',
 'Cape Verde',
 'Chile',
 'China',
 'Colombia',
 'Costa Rica',
 'Croatia',
 'Cyprus',
 'Czech Republic',
 'Denmark',
 'Dominican Republic',
 'Ecuador',
 'El Salvador',
 'Estonia',
 'Fiji',
 'Finland',
 'France',
 'Georgia',
 'Germany',
 'Greece',
 'Grenada',
 'Hungary',
 'Iceland',
 'Iran',
 'Ireland',
 'Israel',
 'Italy',
 'Jamaica',
```

```
'Japan',
'Jordan',
'Kazakhstan',
'Kuwait',
'Latvia',
'Lebanon',
'Libya',
'Lithuania',
'Luxembourg',
'Macedonia, FYR',
'Malaysia',
'Maldives',
'Malta',
'Mauritius',
'Moldova',
'Montenegro',
'Morocco',
'Netherlands',
'New Zealand',
'Norway',
'Oman',
'Panama',
'Paraguay',
'Peru',
'Poland',
'Portugal',
'Qatar',
'Romania',
'Russia',
'Saudi Arabia',
'Serbia',
'Seychelles',
'Singapore',
'Slovak Republic',
'Slovenia',
'South Korea',
'Spain',
'Sri Lanka',
'St. Vincent and the Grenadines',
'Suriname',
'Sweden',
'Switzerland',
'Thailand',
'Tunisia',
'Turkey',
'Ukraine',
'United Arab Emirates',
```

```
'United Kingdom',
'United States',
'Uruguay',
'Venezuela',
'Vietnam']
```

[ ]:

8 8.) Create a table of Descriptive Statistics. Rows being the Cluster number and columns being all the features. Values being the mean of the centroid. Use the nonscaled X values for interpolation

[82]:
```python
Q8DF = pd.concat([preds, X], axis = 1)
Q8DF.groupby(0).mean()
```

[82]:
```
   child_mort    exports    health    imports        income  inflation  \
0
0   76.280882  30.198515  6.090147  43.642146   4227.397059  11.098750
1   12.161616  48.603030  7.314040  49.121212  26017.171717   5.503545

   life_expec  total_fer          gdpp
0
0   61.910294   4.413824   1981.235294
1   76.493939   1.941111  20507.979798
```

[83]:
```python
Q8DF.groupby(0).std()
```

[83]:
```
   child_mort    exports    health    imports        income  inflation  \
0
0   38.076068  18.201742  2.645319  19.323451   4890.581414  13.682630
1    8.523122  30.116032  2.716652  26.928785  20441.749847   6.957187

   life_expec  total_fer          gdpp
0
0    6.897418   1.285590   2528.509189
1    3.735757   0.486744  20578.727127
```

[3]:

9 9.) Write an observation about the descriptive statistics.

- Group 1 exhibits higher average values in income, life expectancy, and GDP per capita compared to group 0, indicating it may represent a more economically developed segment.
- There is less variability in child mortality and total fertility rates in group 1 as shown by the lower standard deviation, whereas GDP per capita has high variability in both groups.

- The average health expenditure as a percentage of GDP is similar for both groups, suggesting comparable health investment despite differences in economic development.

```
[ ]:
```