РГУ НЕФТИ И ГАЗА (НИУ) ИМЕНИ И.М. ГУБКИНА
Факультет: Автоматики и вычислительной техники
Кафедра: Автоматизированных систем управления
Направление: **09.04.01** Информатика и вычислительная техника
Программа: Информационные технологии организационно-экономического управления в нефтегазовом комплексе

КУРСОВАЯ РАБОТА
на тему

Применение методов кластеризации для анализа мировой торговли нефтью и нефтепродуктами

Руководитель:

Д. э. н., профессор

кафедры АСУ

Алетдинова Анна Александровна

Выполнил:

Студент группы АСМ-22-05

Матвеев Роман Вячеславович

Актуальность работы

Экономисты и аналитики всех стран мира в условиях нестабильности и изменчивости ситуации проводят глубокие аналитические исследования для того, чтобы иметь возможность прогнозировать изменения на мировом рынке и предпринимать определенные действия с целью минимизации отрицательных последствий на экономику своих стран.

Кластерный анализ результатов торговли нефтью и нефтепродуктами за прошедший год позволит разделить страны по группам, каждая из которых будет иметь свои отличительные особенности, характеризующие общие тенденции и направления ведения внешней торговли энергоносителями.

Цель работы

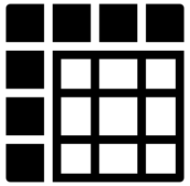
Проведение кластерного анализа международной торговли нефтью и нефтепродуктами за 2021 год.

Задачи

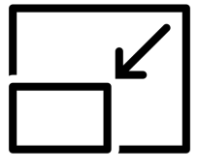
- Анализ литературных источников по выбранной предметной области;
- Ознакомление с общими сведениями о кластерном анализе, его трактовках и определениях в различных источниках;
- Рассмотрение наиболее популярных методов кластерного анализа и применяемых метрик;
- Анализ алгоритма k-means;
- Ознакомление с перечнем задач нефтегазовой отрасли, в которых может быть применим кластерный анализ;
- Поиск и выбор исходных данных, связанных с нефтегазовым комплексом, для осуществления кластерного анализа методом k-means;
- Реализация метода кластеризации k-means на выбранном наборе исходных данных;
- Визуализация, анализ и интерпретация полученных результатов.

Особенности кластеризации

- Задача кластеризации относится к группе методов обучения без учителя, а также группе методов «описательного моделирования».
- Позволяет группировать объекты не только по одному показателю, а по целому набору различных показателей.
- Отсутствие каких-либо ограничений на виды и типы исследуемых объектов.
- Позволяет сжимать обширные массивы различной информации и делать их более показательными и репрезентативными.
- Может быть использована для прогнозирования данных и детектирования аномалий.

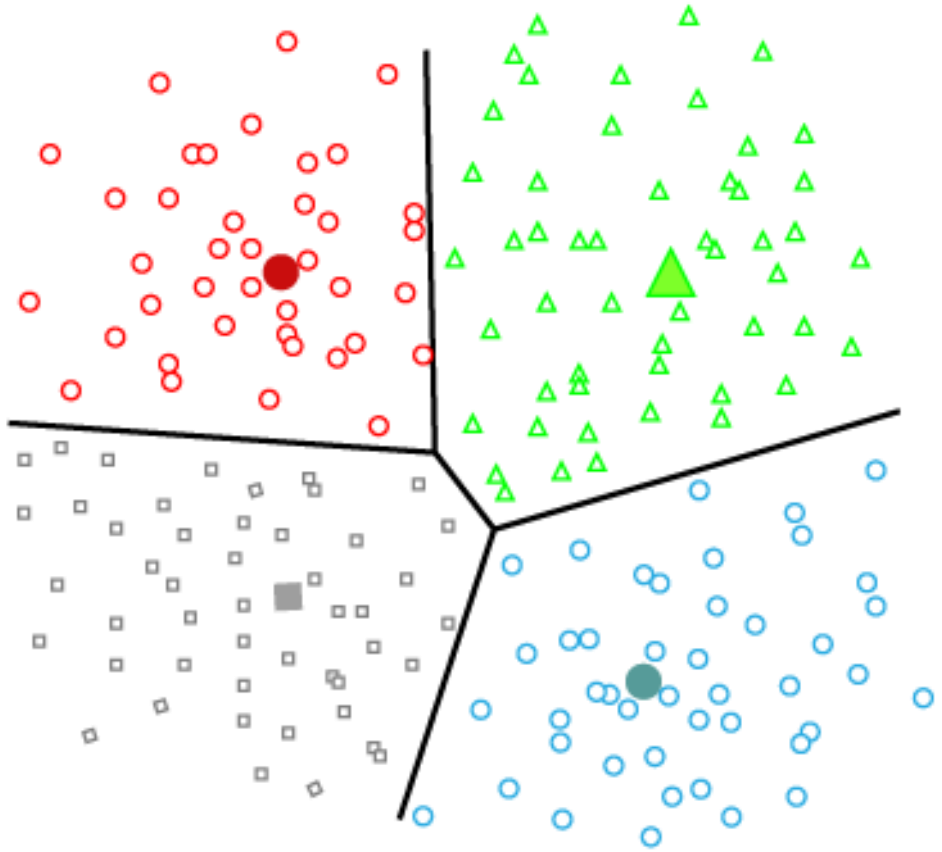


no limits



Группы методов

- Методы разделения
- Иерархические методы
- Методы на основе плотности
- Сеточные методы
- Методы на основе моделей



Исходные данные

- Результаты общемировой торговли нефтью и нефтепродуктами за 2021 год

- Данные импорта сырой нефти и нефтепродуктов

- Данные экспорта сырой нефти и нефтепродуктов

	Countries	Crude Imports	Product Imports	Crude Exports	Product Exports
0	Canada	23.909	30.636	197.439	33.542
1	Mexico	0.000	58.959	52.884	8.238
2	US	304.670	112.861	138.549	244.436
3	S. & Cent. America	21.826	105.767	124.147	23.639
4	Europe	467.741	197.500	36.379	110.523
5	Russia	0.024	1.875	263.565	140.670
6	Other CIS	15.941	6.889	87.121	17.713
7	Iraq	0.005	8.333	176.096	12.305
8	Kuwait	0.000	0.936	88.362	24.259
9	Saudi Arabia	0.013	16.142	323.215	57.653
10	United Arab Emirates	3.163	31.788	146.072	86.732
11	Other Middle East	18.665	19.664	96.967	62.436
12	North Africa	9.296	30.848	85.421	45.375
13	West Africa	0.471	46.032	187.365	8.555
14	East & S. Africa	12.368	41.093	4.827	2.747
15	Australasia	14.888	26.186	9.228	5.368
16	China	525.961	103.408	1.566	60.585
17	India	213.747	49.370	0.053	69.339
18	Japan	122.050	43.024	0.411	10.958
19	Singapore	47.017	91.829	1.009	68.932
20	Other Asia Pacific	257.100	202.057	38.180	131.191

Реализация алгоритма k-means

Подключение библиотек:

```
[154] import numpy as np
      from sklearn import datasets
      import pandas as pd
      import matplotlib.pyplot as plt
      from sklearn.cluster import KMeans
      from sklearn.metrics import silhouette_score
```

Загрузка csv-файла:

```
[4]  df = pd.read_csv('oil trades.csv', sep=';')
     df
```

Исходные данные:

	Countries	Crude Imports	Product Imports	Crude Exports	Product Exports
0	Canada	23.909	30.636	197.439	33.542
1	Mexico	0.000	58.959	52.884	8.238
2	US	304.670	112.861	138.549	244.436
3	S. & Cent. America	21.826	105.767	124.147	23.639
4	Europe	467.741	197.500	36.379	110.523
5	Russia	0.024	1.875	263.565	140.670
6	Other CIS	15.941	6.889	87.121	17.713
7	Iraq	0.005	8.333	176.096	12.305
8	Kuwait	0.000	0.936	88.362	24.259
9	Saudi Arabia	0.013	16.142	323.215	57.653
10	United Arab Emirates	3.163	31.788	146.072	86.732

Удаление поля «Название страны»:

```
[5]  data = df.drop("Countries", axis=1)
     counries = df["Countries"].values
     data = data.astype(np.float)
     data.head()
```

Названия стран временно удалены:

	Crude Imports	Product Imports	Crude Exports	Product Exports
0	23.909	30.636	197.439	33.542
1	0.000	58.959	52.884	8.238
2	304.670	112.861	138.549	244.436
3	21.826	105.767	124.147	23.639
4	467.741	197.500	36.379	110.523

NaN to Num:

```
X = data.values[:,:]
X = np.nan_to_num(X)
```

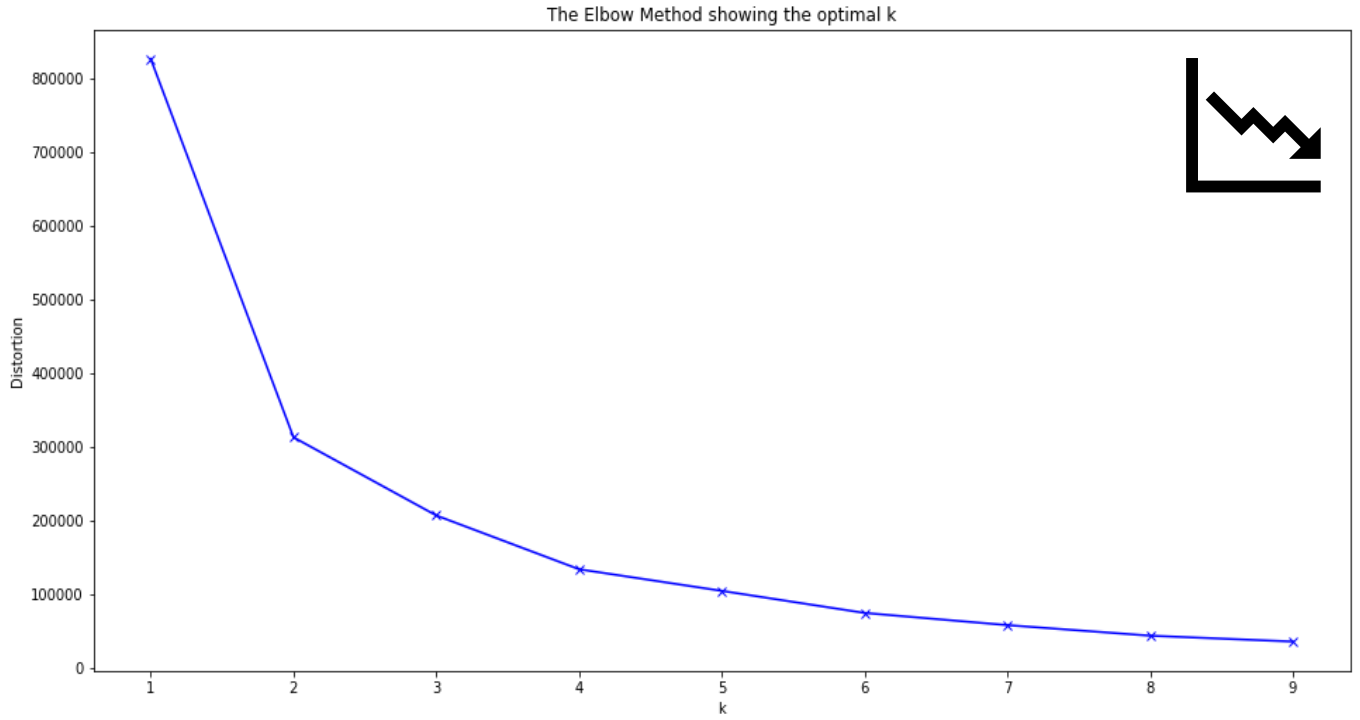
Определение количества кластеров

Метод «локтя» (The Elbow Method)

```
distortions = []
K = range(1,10)
for k in K:
    kmeanModel = KMeans(n_clusters=k)
    kmeanModel.fit(X)
    distortions.append(kmeanModel.inertia_)
```

<u>График зависимости суммы квадратов расстояний от количества кластеров:</u>

```
plt.figure(figsize=(16,8))
plt.plot(K, distortions, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```



Определение количества кластеров

Метод силуэта (The Silhouette Method)

K-means для диапазона значений k:

```
K = range(3,5)
for k in K:
  km = KMeans(init = "random", n_clusters=k, n_init = 12)
  km.fit_predict(X)
  score = silhouette_score(X, km.labels_, metric='euclidean')

  print('Silhouetter Score for k = %d: %.3f' % (k, score))
```

Метрика силуэта для k ∈ [3; 4]:

```
Silhouetter Score for k = 3: 0.381
Silhouetter Score for k = 4: 0.351
```

Реализация алгоритма k-means Sklearn

Выполнение кластеризации данных с k = 3:

```
clusterNum = 3
k_means = KMeans(init = "random", n_clusters=clusterNum, n_init = 12)
k_means.fit(X)
labels = k_means.labels_
centers = k_means.cluster_centers_
```

labels – результат кластеризации;
centers – координаты центров кластеров

Добавим в таблицу колонку с названиями стран и принадлежность к кластерам:

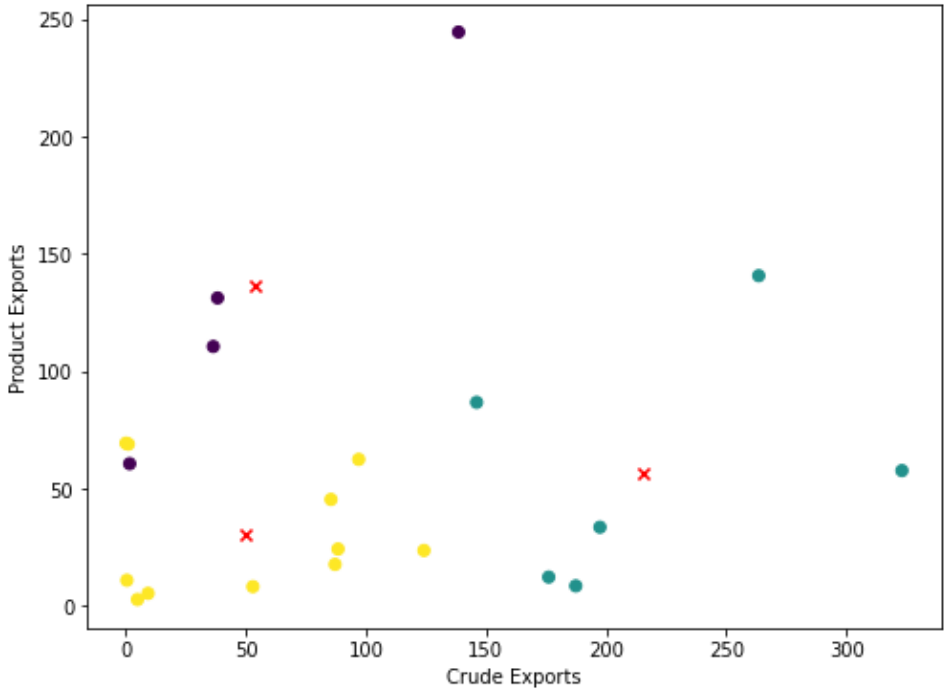
```
data["Countries"] = counries
data["Clus_km"] = labels
data
```

	Crude Imports	Product Imports	Crude Exports	Product Exports	Countries	Clus_km
0	23.909	30.636	197.439	33.542	Canada	1
1	0.000	58.959	52.884	8.238	Mexico	2
2	304.670	112.861	138.549	244.436	US	0
3	21.826	105.767	124.147	23.639	S. & Cent. America	2
4	467.741	197.500	36.379	110.523	Europe	0
5	0.024	1.875	263.565	140.670	Russia	1
6	15.941	6.889	87.121	17.713	Other CIS	2
7	0.005	8.333	176.096	12.305	Iraq	1
8	0.000	0.936	88.362	24.259	Kuwait	2
9	0.013	16.142	323.215	57.653	Saudi Arabia	1
10	3.163	31.788	146.072	86.732	United Arab Emirates	1
11	18.665	19.664	96.967	62.436	Other Middle East	2
12	9.296	30.848	85.421	45.375	North Africa	2
13	0.471	46.032	187.365	8.555	West Africa	1
14	12.368	41.093	4.827	2.747	East & S. Africa	2
15	14.888	26.186	9.228	5.368	Australasia	2
16	525.961	103.408	1.566	60.585	China	0
17	213.747	49.370	0.053	69.339	India	2
18	122.050	43.024	0.411	10.958	Japan	2
19	47.017	91.829	1.009	68.932	Singapore	2
20	257.100	202.057	38.180	131.191	Other Asia Pacific	0

Визуализация алгоритма k-means

Кластеры и их центры по экспорту нефти и нефтепродуктов:

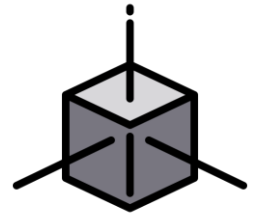
```
plt.figure(figsize=(8,6))
plt.scatter(X[:, 2], X[:, 3], c=labels.astype(np.float), alpha=1)
plt.xlabel("Crude Exports")
plt.ylabel("Product Exports")
plt.scatter(centers[:, 2], centers[:, 3], marker='x', c="r")
plt.show()
```

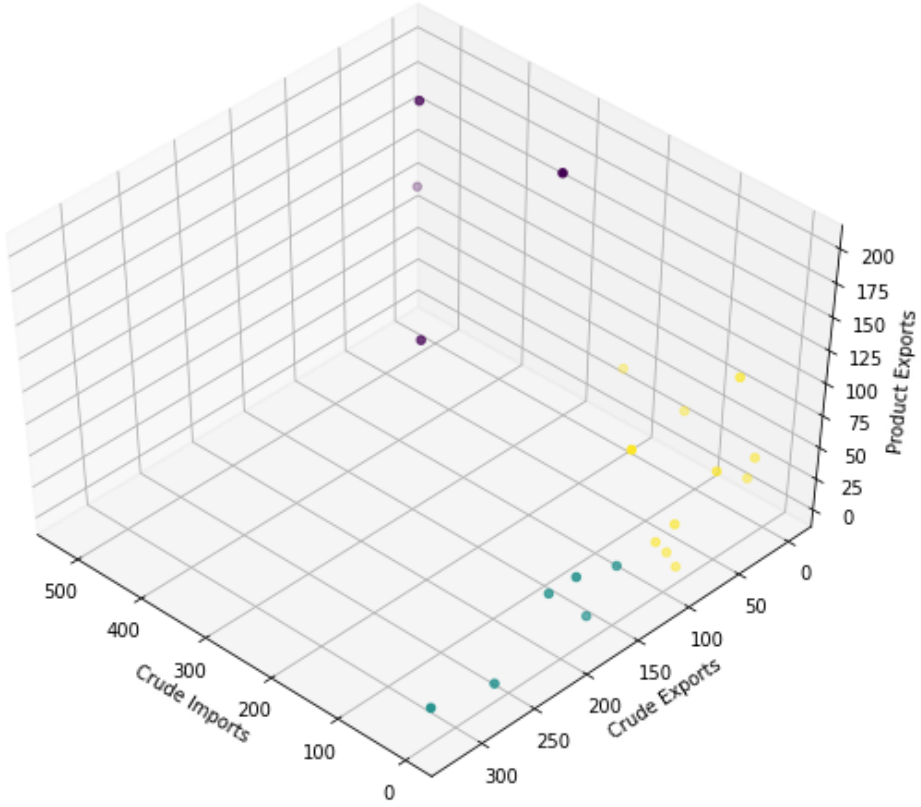



```
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure(1, figsize=(8, 6))
plt.clf
ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=134)

ax.set_xlabel("Crude Imports")
ax.set_ylabel("Crude Exports")
ax.set_zlabel("Product Exports")

ax.scatter(X[:, 0], X[:, 2], X[:, 1], c=labels.astype(np.float))
```





Средние значения по кластерам

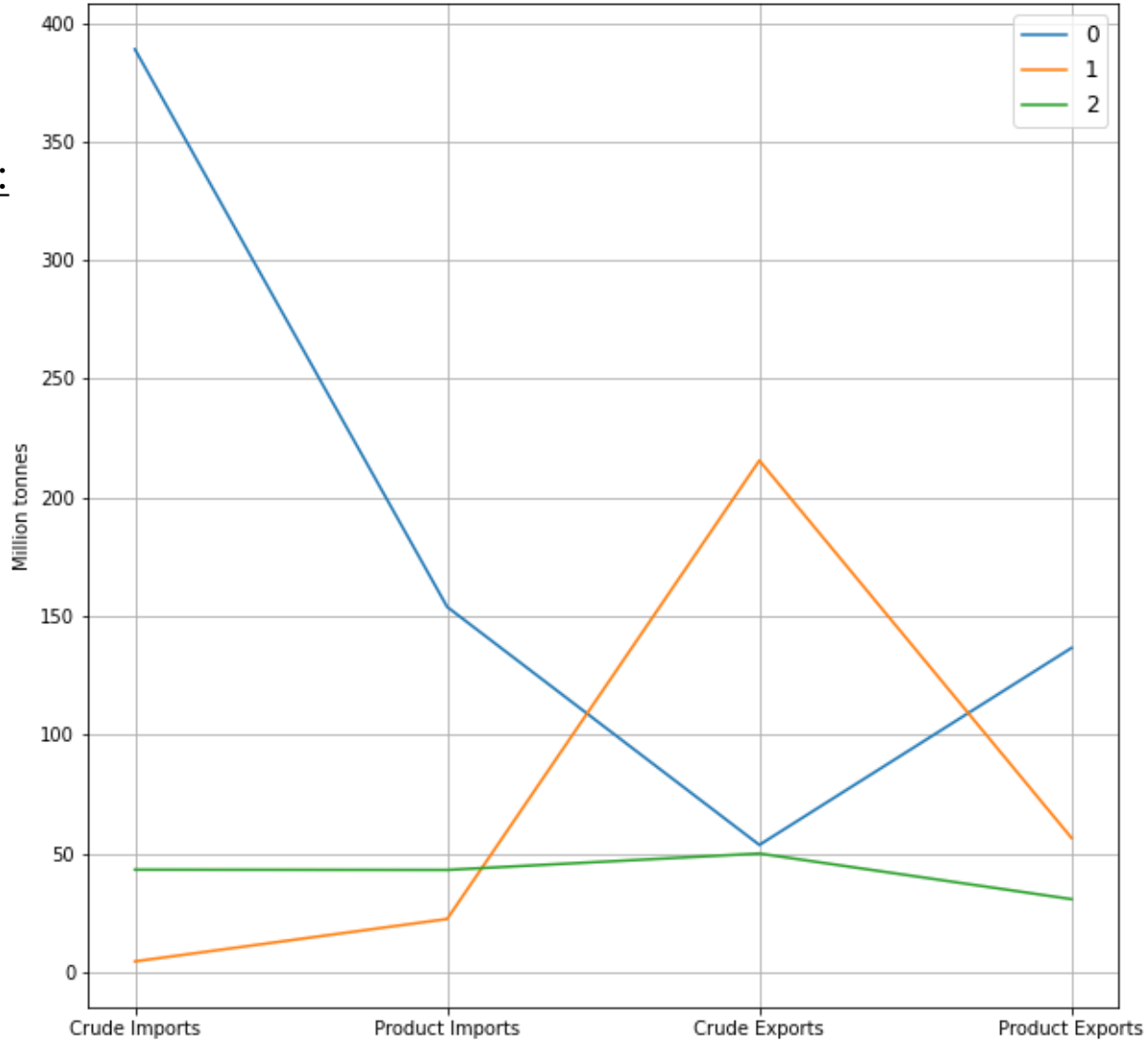
Вычисление средних значений показателей по кластерам:

```
data.groupby('Clus_km').mean()
```

Clus_km	Crude Imports	Product Imports	Crude Exports	Product Exports
0	388.868000	153.956500	53.668500	136.683750
1	4.597500	22.467667	215.625333	56.576167
2	43.254364	43.142273	50.039091	30.818545

Перенесем результаты на график:

```
plt.figure(figsize=(10, 10))
plt.ylabel("Million tonnes")
clusters_arr = np.arange(0, clusterNum, 1)
for i in clusters_arr:
  plt.plot(data.groupby('Clus_km').mean().iloc[i, :], label=i)
plt.grid(True)
plt.legend(loc='best', fontsize=12)
```



Реализация алгоритма k-means SciPy

Выполнение кластеризации данных с k = 3:

```
clusterNum = 3
centers_scipy, distortion = kmeans(X, k_or_guess=clusterNum)
labels_scipy,_ = vq(X,centers_scipy)
```

labels_scipy – результат кластеризации;
centers_scipy – координаты центров кластеров

Добавим в таблицу колонку с названиями стран и принадлежность к кластерам:

```
data["Countries"] = counries
data["Clus_km"] = labels_scipy
data
```

	Crude Imports	Product Imports	Crude Exports	Product Exports	Countries	Clus_km
0	23.909	30.636	197.439	33.542	Canada	1
1	0.000	58.959	52.884	8.238	Mexico	2
2	304.670	112.861	138.549	244.436	US	0
3	21.826	105.767	124.147	23.639	S. & Cent. America	2
4	467.741	197.500	36.379	110.523	Europe	0
5	0.024	1.875	263.565	140.670	Russia	1
6	15.941	6.889	87.121	17.713	Other CIS	2
7	0.005	8.333	176.096	12.305	Iraq	1
8	0.000	0.936	88.362	24.259	Kuwait	2
9	0.013	16.142	323.215	57.653	Saudi Arabia	1
10	3.163	31.788	146.072	86.732	United Arab Emirates	1
11	18.665	19.664	96.967	62.436	Other Middle East	2
12	9.296	30.848	85.421	45.375	North Africa	2
13	0.471	46.032	187.365	8.555	West Africa	1
14	12.368	41.093	4.827	2.747	East & S. Africa	2
15	14.888	26.186	9.228	5.368	Australasia	2
16	525.961	103.408	1.566	60.585	China	0
17	213.747	49.370	0.053	69.339	India	0
18	122.050	43.024	0.411	10.958	Japan	2
19	47.017	91.829	1.009	68.932	Singapore	2
20	257.100	202.057	38.180	131.191	Other Asia Pacific	0

Sklearn VS SciPy

KMeans из Sklearn:

```
cluster_0 = []
cluster_1 = []
cluster_2 = []

for i in data.values:
  n = i[len(data.values[0,:])-1]
  if n == 0:
    cluster_0.append(i[-2])
  elif n == 1:
    cluster_1.append(i[-2])
  elif n == 2:
    cluster_2.append(i[-2])
print("Cluster 0 ", cluster_0, "\n","Cluster 1 ",cluster_1,"\n","Cluster 2 ", cluster_2, "\n")
```

```
Cluster 0  ['US', 'Europe ', 'China', 'Other Asia Pacific']
 Cluster 1  ['Canada', 'Russia', 'Iraq', 'Saudi Arabia', 'United Arab Emirates', 'West Africa']
 Cluster 2  ['Mexico', 'S. & Cent. America', 'Other CIS', 'Kuwait', 'Other Middle East', 'North Africa', 'East & S. Africa', 'Australasia', 'India', 'Japan', 'Singapore']
```

kmeans из SciPy:

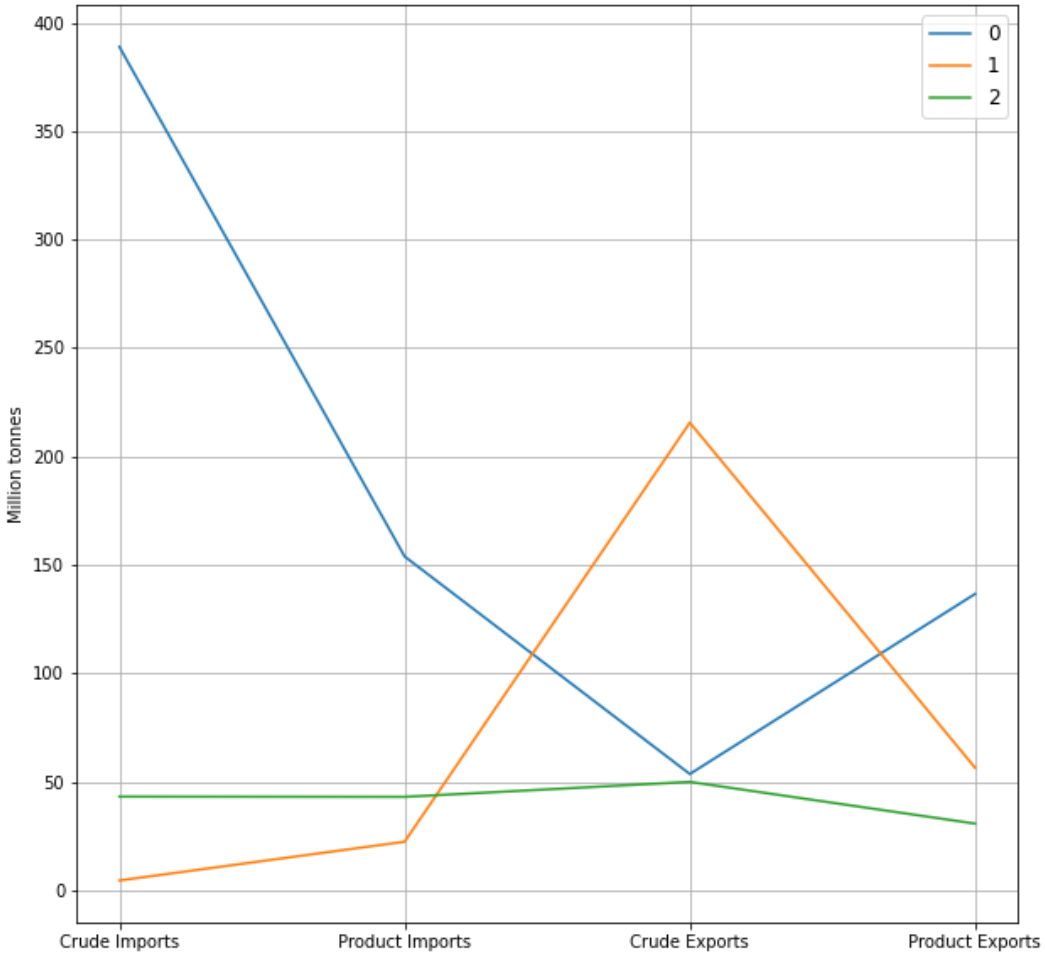
```
Cluster 0  ['US', 'Europe ', 'China', 'India', 'Other Asia Pacific']
Cluster 1  ['Canada', 'Russia', 'Iraq', 'Saudi Arabia', 'United Arab Emirates', 'West Africa']
Cluster 2  ['Mexico', 'S. & Cent. America', 'Other CIS', 'Kuwait', 'Other Middle East', 'North Africa', 'East & S. Africa', 'Australasia', 'Japan', 'Singapore']
```

Обработка полученных результатов

- Кластер №0 – US, Europe, China, Other Asia Pacific
- Кластер №1 – Canada, Russia, Iraq, Saudi Arabia, United Arab Emirates, West Africa
- Кластер №2 – Mexico, S. & Cent. America, Other CIS, Kuwait, Other Middle East, North Africa, East & S. Africa, Australasia, India, Japan, Singapore.

Обработка полученных результатов

Страны кластера №1, в числе которых и Россия, являются главными экспортерами сырой нефти, главными же потребителями – страны кластера №0, среди которых Китай, США, страны Европы и Азии. При этом страны кластера №0 можно назвать главными импортерами и экспортерами нефтепродуктов.



Заключение

В результате выполнения курсовой работы цель была достигнута, все задачи выполнены.

Кластерный анализ результатов торговли нефтью и нефтепродуктами за прошедший год позволил:

- разделить страны по группам, каждая из которых имеет свои отличительные особенности, характеризующие общие тенденции и направления ведения внешней торговли энергоносителями.
- Используя результаты данного анализа, были выдвинуты некоторые предположения о реакции мирового рынка нефти и нефтепродуктов на введение ограничений на цены российской нефти и возможные ответные меры со стороны России.

Благодарю за внимание!