دانشگاه خوارزمی

دانشکده علوم ریاضی و کامپیوتر

عنوان:

خوشه‌بندی کشورهای دنیا بر اساس پارامترهای مختلف رضایت زندگی با استفاده از الگوریتم k-میانگین

استاد راهنما:

دکتر کیوان برنا و دکتر محمد سلطانیان

دانشجو:

فاطمه خدادادی

تابستان 1402


بسم الله الرحمن الرحيم

فهرست مطالب

فهرست جداول

فهرست اشکال

خوشه‌بندی کشورهای دنیا بر اساس پارامترهای مختلف رضایت زندگی با استفاده از الگوریتم k-میانگین

1- مقدمه

خوشه‌بندی به فرایند تقسیم نمونه‌های داده بدون برچسب به خوشه‌های مختلف گفته می‌شود. این خوشه‌ها باید به‌گونه‌ای ساخته شوند که نمونه‌های هر یک بیشترین شباهت را با همدیگر و بیشترین اختلاف را با نمونه‌های خوشه‌های دیگر داشته باشند. به زبان ساده‌تر، هدف از خوشه‌بندی تفکیک نمونه‌هایی با ویژگی‌ها و صفات مشابه و تخصیص آن‌ها به خوشه‌ها است [1].

یکی از محبوبترین روش‌های خوشه‌بندی K-میانگین[1] است. اگر n را تعداد نمونه‌ها، d را تعداد متغیرهای مستقل مسئله و k را تعداد خوشه‌ها در نظر بگیریم، درجه پیچیدگی محاسباتی این الگوریتم یا به عبارتی تعداد تکرارهای الگوریتم برای رسیدن به جواب بهینه برابر با $O(n^{dk+1})$ خواهد بود. در خوشه‌بندی K-میانگین، مسئله با کمینه‌سازی یا بیشینه‌سازی تابع هدف حل می‌شود. درصورتی‌که معیار میزان فاصله بین نمونه‌ها باشد، تابع هدف کمینه‌سازی خواهد بود و درصورتی‌که معیار مشابهت باشد، تابع هدف بیشینه‌سازی است [2].

2- هدف پروژه

هدف از این پروژه خوشه‌بندی کشورهای دنیا بر اساس پارامترهای مختلف میزان رضایت زندگی با الگوریتم یادگیری بدون نظارت K-میانگین است.

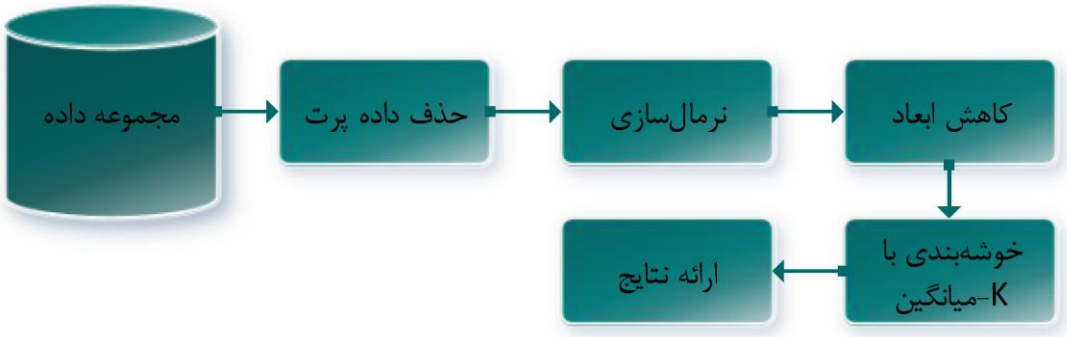
3- روش‌شناسی پروژه

روش‌شناسی پروژه در شکل 1 خلاصه‌شده است. مطابق با این شکل 1، پروژه حاضر شامل سه مرحله اصلی است:

[1] K-means

خوشه‌بندی کشورهای دنیا بر اساس پارامترهای مختلف رضایت زندگی با استفاده از الگوریتم k-میانگین

1. تهیه و آماده‌سازی داده: در این گام اطلاعات مربوط به رضایت زندگی تهیه و پس از پیش‌پردازش مجموعه داده تحقیق آماده شد.
2. خوشه‌بندی: ابتدا تعداد مناسب خوشه‌ها با محاسبه پارامتر مجذور فاصله نمونه‌ها تا نزدیکترین مرکز خوشه آن‌ها (inertia) تعیین شد. سپس خوشه‌بندی کشورهای مختلف در شش خوشه به کمک الگوریتم K-میانگین صورت گرفت.
3. ارائه نتایج: پس از خوشه‌بندی، بر اساس نظر کارشناسان و ویژگی‌های هر خوشه، کشورهای مختلف برچسب بسیار عالی، عالی، خوب، متوسط، بد و بسیار بد گرفتند.



شکل 1: روششناسی پروژه

4- مجموعه داده

مجموعه داده استفاده‌شده در این پروژه گزارش شادی جهان مربوط به سال 2017 بود که توسط https://worldhappiness.report/ تهیه‌شده است. این مجموعه داده شامل اطلاعات 166 کشور مختلف است و متغیرهای نردبان زندگی[2]، سرانه ثبت تولید ناخالص داخلی[3]، حمایت اجتماعی[4]، امید به زندگی سالم در بدو تولد[5]، آزادی

[2] Life ladder
[3] Log GDP per capita
[4] Social support
[5] Healthy life expectancy at birth

خوشه‌بندی کشورهای دنیا بر اساس پارامترهای مختلف رضایت زندگی با استفاده از الگوریتم k-میانگین

در انتخاب زندگی[6]، سخاوتمندی[7]، تصورات از فساد[8]، تأثیر مثبت[9]، تأثیر منفی[10]، اعتماد به دولت ملی[11]، کیفیت دموکراسی[12] و کیفیت تحویل[13] را در برمی‌گیرد.

4-1- پیش‌پردازش داده

پیش‌پردازش داده در این پروژه به‌منظور بهبود خوشه‌بندی و در سه گام انجام شد:

- حذف داده پرت[14]: وجود داده پرت می‌توانند در عملکرد الگوریتم‌های یادگیری ماشین اختلال ایجاد کند. بر همین اساس مطابق با ادبیات تحقیق [3]، رکوردهای داده‌ای که در بازه [3 × انحراف معیار + میانگین، 3 × انحراف معیار − میانگین] قرار نداشتند به‌عنوان داده پرت تشخیص داده شدند و از مجموعه داده حذف گردیدند.
- نرمال‌سازی: وجود متغیرهای مستقل در بازه‌های مختلف ممکن است سبب اخلال در عملکرد الگوریتم‌های یادگیری ماشین گردد و تأثیر یک متغیر به‌اشتباه بیش از سایر متغیرها لحاظ گردد [4]. بر همین اساس با رابطه نرمال‌سازی خطی، تمامی متغیرهای مستقل در بازه صفر تا یک نرمال‌سازی شدند:

رابطه 1
$$x_{Normalized} = \frac{(x-Min)}{(Max-Min)}$$

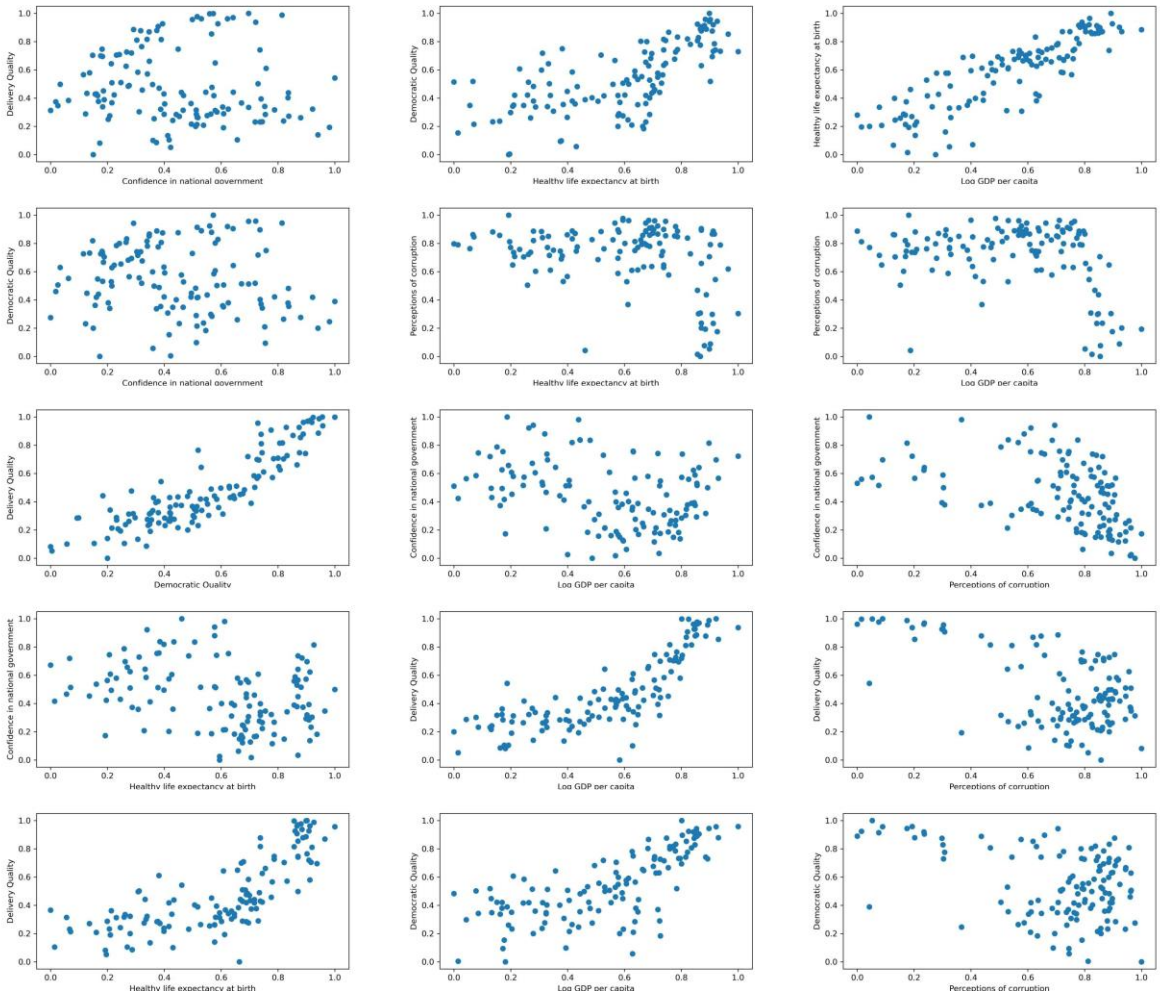
در رابطه 1 $x_{Normalized}$، x، Max و Min به ترتیب مقدار رکورد متغیر نرمال شده، مقدار رکورد متغیر پیش از نرمال‌سازی، ماکزیمم مقدار متغیر و مینیمم مقدار متغیر هستند.

---

6 Freedom to make life choices
7 Generosity
8 Perceptions of corruption
9 Positive affect
10 Negative affect
11 Confidence in national government
12 Democratic Quality
13 Delivery Quality
14 Outlier

- کاهش ابعاد: وجود تعداد زیادی از متغیرهای مستقل می‌تواند به عملکرد الگوریتم یادگیری ماشین آسیب برساند. بر همین اساس گه‌ها کاهش تعداد متغیرهای مستقل می‌تواند مفید باشد. یکی از ساده‌ترین روش‌های کاهش ابعاد، حذف متغیرهایی است که کمترین واریانس را دارند [4]؛ بنابراین از میان 12 متغیر مستقل مجموعه داده تحقیق، شش مورد که کمترین واریانس را داشتند حذف شدند.

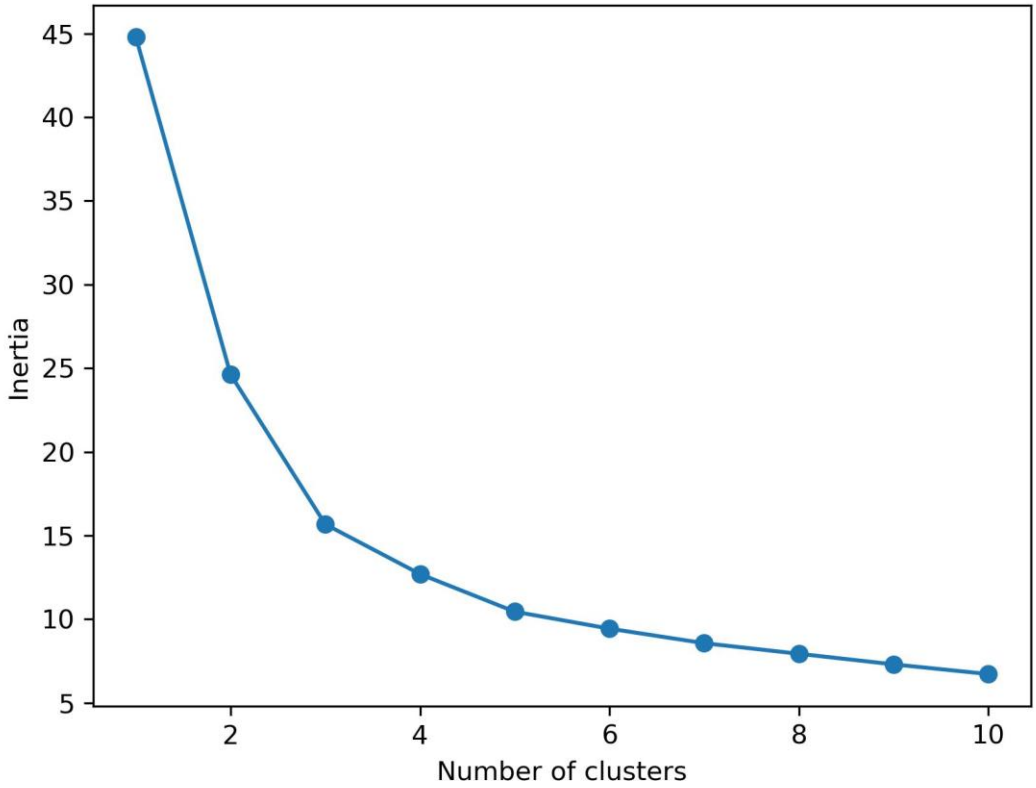
شکل 2 نمودار پراکندگی متغیرهای مختلف را پس از انجام پیش‌پردازش نشان می‌دهد.



شکل 2: نمودار پراکندگی متغیرهای تحقیق پس از انجام پیش‌پردازش داده

5- خوشه‌بندی با الگوریتم K-میانگین (K-means)

به‌منظور خوشه‌بندی کشورها با استفاده از مجموعه داده پردازش‌شده، از کتابخانه scikit-learn و پارامترهای پیش‌فرض آن استفاده شد. جهت تعیین تعداد مناسب خوشه‌ها، پارامتر inertia محاسبه گردید که نتیجه آن در شکل 3 قابل‌مشاهده است. با توجه به نظر کارشناسان، تعداد مناسب خوشه‌ها شش تعیین شد.



شکل 3: محاسبه پارامتر inertia جهت تعیین تعداد مناسب خوشه‌ها

با تعیین تعداد مناسب خوشه‌ها، خوشه‌بندی انجام شد که نتایج آن در جدول 1 قابل‌مشاهده است. مطابق با جدول 1، 23 کشور در خوشه 0، 28 کشور در خوشه 1، 30 کشور در خوشه 2، 15 کشور در خوشه 3، 15 کشور در خوشه 4 و 15 کشور در خوشه 5 قرار گرفتند.

خوشه‌بندی کشورهای دنیا بر اساس پارامترهای مختلف رضایت زندگی با استفاده از الگوریتم k-میانگین

جدول 1: نتایج خوشه‌بندی

کشور	خوشه	کشور	خوشه	کشور	خوشه	کشور	خوشه
Belgium	0	Mauritania	1	Ukraine	2	Nicaragua	5
Chile	0	Mozambique	1	Australia	3	Russia	5
Costa Rica	0	Nepal	1	Austria	3	Sri Lanka	5
Cyprus	0	Niger	1	Canada	3	Thailand	5
Czech Republic	0	Nigeria	1	Denmark	3	Turkey	5
Estonia	0	Pakistan	1	Finland	3	Venezuela	5
France	0	Senegal	1	Germany	3		
Hungary	0	Sierra Leone	1	Hong Kong S.A.R. of China	3		
Iceland	0	Togo	1	Ireland	3		
Israel	0	Uganda	1	Luxembourg	3		
Italy	0	Zimbabwe	1	Netherlands	3		
Japan	0	Albania	2	New Zealand	3		
Latvia	0	Argentina	2	Norway	3		
Lithuania	0	Armenia	2	Sweden	3		
Malta	0	Bosnia and Herzegovina	2	Switzerland	3		
Mauritius	0	Brazil	2	United Kingdom	3		
Portugal	0	Bulgaria	2	Bangladesh	4		
Slovakia	0	Colombia	2	Botswana	4		
Slovenia	0	Croatia	2	Cambodia	4		
South Korea	0	Dominican Republic	2	Gambia	4		
Spain	0	El Salvador	2	Ghana	4		
United States	0	Gabon	2	India	4		
Uruguay	0	Georgia	2	Laos	4		
Afghanistan	1	Greece	2	Myanmar	4		
Benin	1	Jamaica	2	Namibia	4		
Burkina Faso	1	Kosovo	2	Philippines	4		
Cameroon	1	Lebanon	2	Rwanda	4		
Chad	1	Macedonia	2	Tajikistan	4		
Congo (Brazzaville)	1	Mexico	2	Tanzania	4		
Congo (Kinshasa)	1	Moldova	2	Uzbekistan	4		
Ethiopia	1	Mongolia	2	Zambia	4		
Guinea	1	Montenegro	2	Azerbaijan	5		
Haiti	1	Panama	2	Belarus	5		
Ivory Coast	1	Paraguay	2	Bolivia	5		
Kenya	1	Peru	2	Ecuador	5		
Lesotho	1	Romania	2	Guatemala	5		
Liberia	1	Serbia	2	Honduras	5		
Madagascar	1	South Africa	2	Iraq	5		
Malawi	1	Trinidad and Tobago	2	Kazakhstan	5		
Mali	1	Tunisia	2	Kyrgyzstan	5		

پس از خوشه‌بندی، میانگین مقادیر متغیرهای مستقل در هر خوشه محاسبه گشت. بر اساس مقادیر میانگین و نظر کارشناسان،

خوشه‌بندی کشورهای دنیا بر اساس پارامترهای مختلف رضایت زندگی با استفاده از الگوریتم k-میانگین

به هر خوشه یک برچسب رضایت زندگی اختصاص داده شد. مقادیر میانگین محاسبه‌شده به همراه برچسب هر خوشه در جدول 2 آمده است.

جدول 2: نتایج برچسب‌گذاری خوشه‌ها بر اساس مقادیر میانگین متغیرهای مستقل

	Log GDP per capita	Healthy life expectancy at birth	Perceptions of corruption	Confidence in national government	Democratic Quality	Delivery Quality	Label
Cluster 0	0.868	0.891	0.211	0.554	0.887	0.941	Excellent
Cluster 1	0.568	0.600	0.767	0.531	0.422	0.354	Moderate
Cluster 2	0.342	0.455	0.536	0.884	0.343	0.313	Good
Cluster 3	0.588	0.654	0.875	0.192	0.518	0.404	Bad
Cluster 4	0.764	0.817	0.782	0.300	0.780	0.715	Very good
Cluster 5	0.215	0.256	0.786	0.539	0.347	0.248	Very bad

مـر اجع

[1] M. G. Omran, A. P. Engelbrecht, and A. Salman, "An overview of clustering methods," *Intelligent Data Analysis,* vol. 11, no. 6, pp. 583-605, 2007.

[2] G. Hamerly and C. Elkan, "Learning the k in k-means," *Advances in neural information processing systems,* vol. 16, 2003.

[3] F. Farhangi, A. Sadeghi-Niaraki, J. Safari Bazargani, S. V. Razavi-Termeh, D. Hussain, and S.-M. Choi, "Time-Series Hourly Sea Surface Temperature Prediction Using Deep Neural Network Models," *Journal of Marine Science and Engineering,* vol. 11, no. 6, p. 1136, 2023.

[4] F. Farhangi, "Investigating the role of data preprocessing, hyperparameters tuning, and type of machine learning algorithm in the improvement of drowsy EEG signal modeling," *Intelligent Systems with Applications,* vol. 15, p. 200100, 2022.

پیوست 1- کد پایتون پروژه

1. واردکردن کتابخانه‌های موردنیاز

```
# importing libraries

import pandas as pd
from sklearn.cluster import KMeans
from numpy import std
from statistics import pvariance
import matplotlib.pyplot as plt
```

2. خواندن مجموعه داده و اطلاعات موردنیاز

```
# reading dataset and initial information

dataset = pd.read_csv('Data.csv')
dataset = dataset[dataset['Year']==2017]
dataset = dataset.dropna()
parameters = dataset.columns.drop(['Country name', 'Year'])
```

3. شناسایی و حذف داده پرت

```
# detecting outliers

outlier_indexes = []

for item in parameters:

    parameter = dataset[item]
    mean = sum(parameter)/len(parameter)
    SD = std(parameter)

    try:
        index = parameter[(parameter<=mean-3*SD)].index[0]
        outlier_indexes.append(index)
    except:
        pass

    try:
        index = parameter[(parameter>=mean+3*SD)].index[0]
        outlier_indexes.append(index)
    except:
        pass
```

```
dataset = dataset.drop(outlier_indexes)
```

4. نرمال‌سازی

```
# normalization

for item in parameters:

    parameter = dataset[item]
    _max = max(parameter)
    _min = min(parameter)
    dataset[item] = (dataset[item]-_min)/(_max-_min)
```

5. کاهش ابعاد

```
# dimention reduction

variance_list = []

for item in parameters:

    parameter = dataset[item]
    variance_list.append(pvariance(parameter))

parameter_list = []

for item in parameters:

    parameter = dataset[item]
    variance_list.append(pvariance(parameter))

    if pvariance(parameter) < (sum(variance_list)/len(variance_list)):

        parameter_list.append(item)

dataset = dataset.drop(parameter_list, axis=1)
```

6. ذخیره مجموعه داده پردازش‌شده

```
# saving pre-processed dataset

dataset.to_csv('Pre-processed dataset.csv',index = False)
```

7. خواندن مجموعه داده پردازش‌شده

```
# reading pre-processed dataset

preprocessed_dataset = pd.read_csv('Pre-processed dataset.csv')
parameters = preprocessed_dataset.columns.drop(['Country name', 'Year'])
```

8. ترسیم نمودار پراکندگی

```
# creating scatter plots

list_plot = []

for col in parameters:
    for row in parameters:

        if [row, col] in list_plot or [col, row] in list_plot or col == row:

            pass

        else:

            list_plot.append([row, col])
            plt.figure(figsize=(7, 3.5))
            plt.xlabel(col)
            plt.ylabel(row)
            x = preprocessed_dataset[col]
            y = preprocessed_dataset[row]
            plt.scatter(x, y)
            plt.savefig('Scatter_plots/Scatter_'+col+'_'+row+'.jpeg', dpi=300)
```

9. تعیین تعداد مناسب خوشه‌ها

```
# calculating the intertia for different numbers of clusters

x = list(zip(preprocessed_dataset['Log GDP per capita'], preprocessed_dataset['Healthy life expectancy at birth'],\
        preprocessed_dataset['Perceptions of corruption'], preprocessed_dataset['Confidence in national government'],\
        preprocessed_dataset['Democratic Quality'], preprocessed_dataset['Delivery Quality']))

inertias = []

for i in range(1,11):
    kmeans = KMeans(n_clusters=i)
    kmeans.fit(x)
    inertias.append(kmeans.inertia_)
```

```
plt.figure(figsize=(7, 3.5))
plt.plot(range(1,11), inertias, marker='o')
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
plt.show()
plt.savefig('intertia.jpeg', dpi=300)
```

10.خوشه‌بندی

```
# clustering

model = KMeans(n_clusters=6)
model.fit(x)
clusters = pd.DataFrame(list(zip(preprocessed_dataset['Country name'], model.predict(x))))
clusters.columns = ['Country name', 'Cluster']
clusters = preprocessed_dataset.merge(clusters, left_on='Country name', right_on='Country name')
clusters.to_csv('Clusters.csv', index = False)
```

11.تعیین میانگین متغیرهای مستقل در هر خوشه

```
# calculating mean of variables in each cluster

Clusters = pd.read_csv ('Clusters.csv')

result = pd.DataFrame()

for item in clusters['Cluster'].drop_duplicates().sort_values(ascending=True):

    selected_cluster = clusters[clusters['Cluster'] == item]
    average = pd.DataFrame(selected_cluster[['Log GDP per capita', 'Healthy life expectancy at birth',\
                    'Perceptions of corruption', 'Confidence in national government',\
                    'Democratic Quality', 'Delivery Quality']].mean())
    average.columns = ['Cluster '+' '+str(item)]
    result = pd.concat([result, average.T])

result.to_csv('Average_results.csv')
```