Classification and analysis of countries

Shulin Zheng

May 15, 2020

1. Introduce

1.1 Background

In March this year, the epidemic broke out all over the world, which has a huge impact on people's lives and work. The epidemic spread rapidly, affected a wide range, high mortality, affecting all aspects, such as economy, education and so on. At present, research shows that the spread of the epidemic is related to the level of medical treatment and the elderly in the country. We hope to cluster countries according to relevant data and medical indicators, so that countries in the same cluster can learn from each other's governance experience.

1.2 Questions

Whether countries can be clustered through data and indicators, so that similar countries can learn from each other's experience.

2. Data acquisition and cleaning

2.1 Data sources

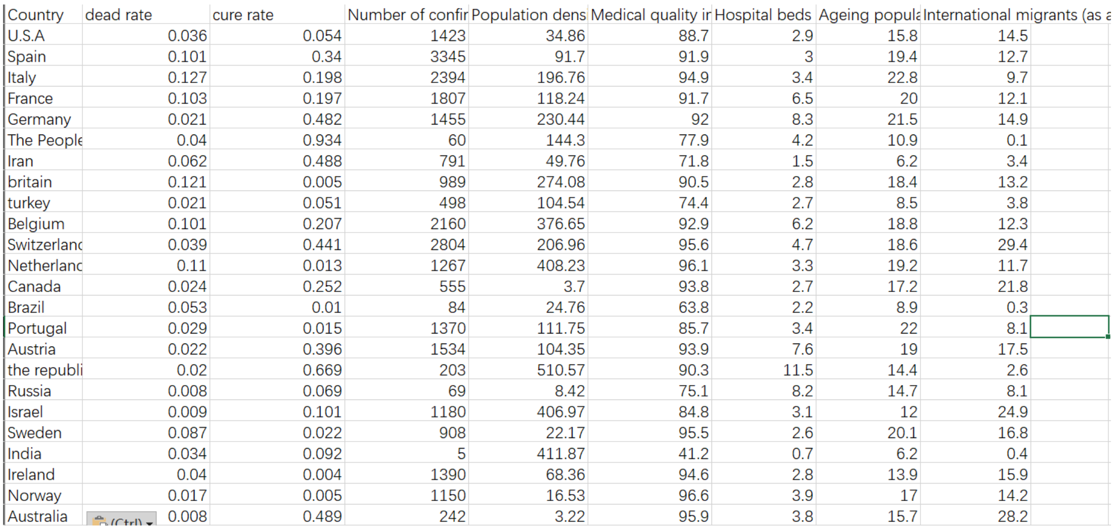
The number of confirmed cases, cured cases, deaths and suspected cases in each country related to the epidemic was obtained from Johns Hopkins University, while the medical related data in each country was obtained from the world bank.

2.2 Data cleaning

Data downloaded or scraped from multiple sources is combined into a single table. Due to a lack of records, many early seasons lacked values. I decided to use data from 2018. Get the number of people in each country, deal with the number of people cured, and get the cure rate and mortality rate.

2.3 Feature selection

After data cleansing, there are 134 countries. Checking the meaning of each feature, remove some redundancy.



Country	dead rate	cure rate	Number of confir	Population dens	Medical quality ir	Hospital beds	Ageing popula	International migrants (as a
U.S.A	0.036	0.054	1423	34.86	88.7	2.9	15.8	14.5
Spain	0.101	0.34	3345	91.7	91.9	3	19.4	12.7
Italy	0.127	0.198	2394	196.76	94.9	3.4	22.8	9.7
France	0.103	0.197	1807	118.24	91.7	6.5	20	12.1
Germany	0.021	0.482	1455	230.44	92	8.3	21.5	14.9
The People	0.04	0.934	60	144.3	77.9	4.2	10.9	0.1
Iran	0.062	0.488	791	49.76	71.8	1.5	6.2	3.4
britain	0.121	0.005	989	274.08	90.5	2.8	18.4	13.2
turkey	0.021	0.051	498	104.54	74.4	2.7	8.5	3.8
Belgium	0.101	0.207	2160	376.65	92.9	6.2	18.8	12.3
Switzerlanc	0.039	0.441	2804	206.96	95.6	4.7	18.6	29.4
Netherlanc	0.11	0.013	1267	408.23	96.1	3.3	19.2	11.7
Canada	0.024	0.252	555	3.7	93.8	2.7	17.2	21.8
Brazil	0.053	0.01	84	24.76	63.8	2.2	8.9	0.3
Portugal	0.029	0.015	1370	111.75	85.7	3.4	22	8.1
Austria	0.022	0.396	1534	104.35	93.9	7.6	19	17.5
the republi	0.02	0.669	203	510.57	90.3	11.5	14.4	2.6
Russia	0.008	0.069	69	8.42	75.1	8.2	14.7	8.1
Israel	0.009	0.101	1180	406.97	84.8	3.1	12	24.9
Sweden	0.087	0.022	908	22.17	95.5	2.6	20.1	16.8
India	0.034	0.092	5	411.87	41.2	0.7	6.2	0.4
Ireland	0.04	0.004	1390	68.36	94.6	2.8	13.9	15.9
Norway	0.017	0.005	1150	16.53	96.6	3.9	17	14.2
Australia	0.008	0.489	242	3.22	95.9	3.8	15.7	28.2

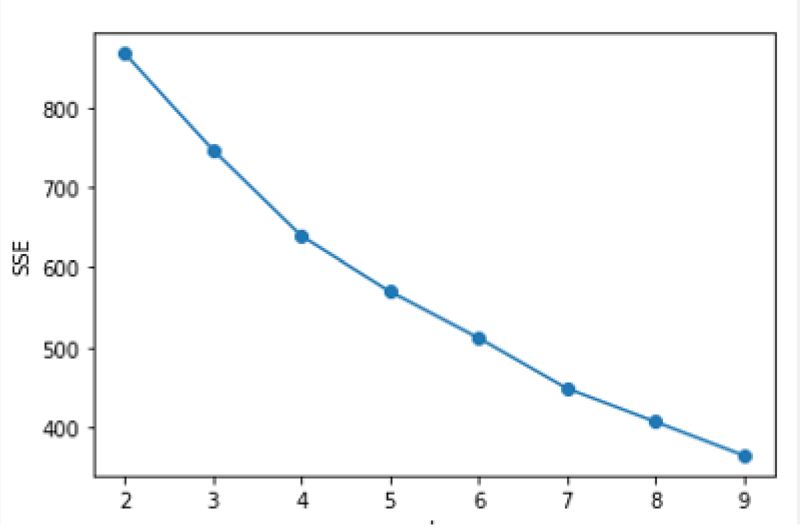
3. Exploratory data analysis

3.1 Selection of k-means

3.1.1 Elbow method

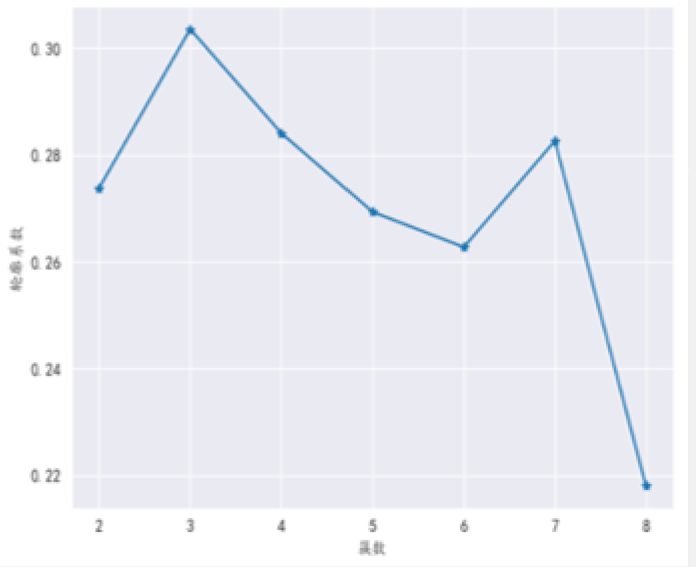
When k is less than the real cluster number, the aggregation degree of each cluster will be greatly increased with the increase of K, so the decrease range of SSE will be large. When k reaches the real cluster number, the aggregation degree return will be rapidly reduced with the increase of K, so the decrease range of SSE will be sharply reduced, and then it will become gentle with the increase of K value, that is to say, the relationship between SSE and K is an elbow Shape, and the K value corresponding to this elbow is the real clustering number of the data.

The K value of elbow is 4, so the best cluster number should be 4.



3.1.2 Contour coefficient method

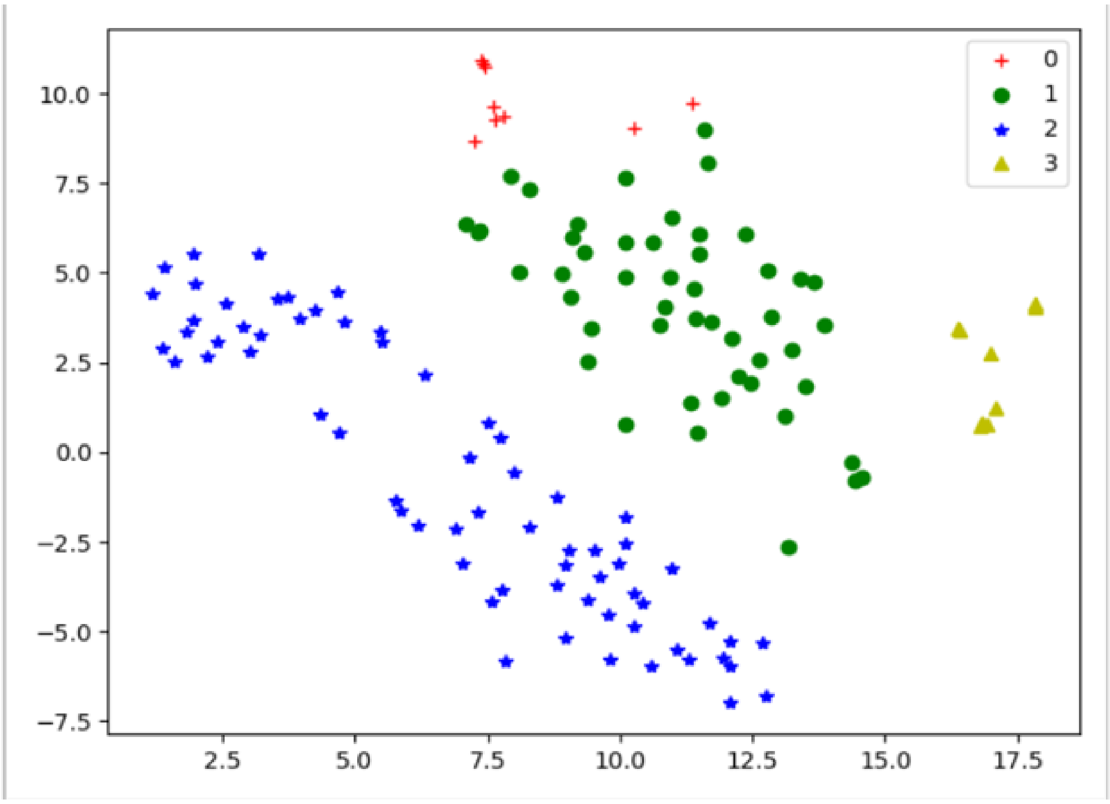
The maximum K value of contour coefficient is 3, which means that our optimal clustering number is 3. However, it is worth noting that from the elbow diagram of K and SSE, we can see that when k is taken as 3, SSE is still relatively large, so we go back to find the second, considering the K value 4 of the second largest contour coefficient, at this time SSE value is smaller than that when K is taken as 3, so the best clustering coefficient should be taken as 4 instead of 3.



3.2 Analysis and discussion

Using the integrated k-means algorithm in Python's machine learning library sklearn, 134 countries are grouped into four categories.

At the same time, tsne is used to visualize the clustering results.



The novel coronavirus pneumonia is not very serious in the first country: but the national medical situation is relatively low.

The second category of countries: China, the United States, Japan, South Korea, northern Europe, Australia and other countries, the epidemic situation is relatively serious, the medical level is medium to high, and the aging population also exists.

The novel coronavirus pneumonia is the third country: most of the countries are backward in developing countries such as Africa, the Middle East and Southeast Asia. The number of confirmed cases of new crown pneumonia epidemic is less than the time limit. However, due to poor medical quality, the shortage of medical resources, the low cure rate and high mortality rate.

The fourth kind of countries: countries with more serious epidemic situation in Europe, with serious aging and high mortality, but with high medical level, sufficient medical resources and relatively high cure rate.