Data Analysis of the Impact of Income and Geographical Location on the Chinese People's Happiness Index

Yinhao Liu[1], Meili Liu[2], Jeng-Eng Lin[3], and Chun-Te Lee[4](✉)

[1] Department of Computer Science, Wenzhou Kean University, Wenzhou, Zhejiang, China
liuyinh@kean.edu
[2] Institute of Artificial Intelligence, Midea Group, Shenzhen, China
[3] Department of Mathematical Sciences, George Mason University, Washington DC, USA
jelin@gmu.edu
[4] Department of Mathematics, Wenzhou Kean University, Wenzhou, Zhejiang, China
chulee@kean.edu

Abstract. With the advancement of China's reform and opening up, the life goal of the Chinese people has gradually shifted from solving the problem of food and clothing to pursuing happiness. However, everyone has their own definition of happiness, and there are many factors that affect happiness. Based on the "Top 30 Happiest Cities in China in 2019", as well as the per capita disposable income of these 30 cities and the distribution of the three major industries, this article explains the impact of income and geographic location on the People's Happiness Index.

Keywords: Happiness · Correlation Coefficient · Normalization · Linear Regression · K-Means

1 Introduction

From the late 1960s to the mid-1980s, the measurement of human subjective well-being became a hot research field in psychology. Psychologists discuss subjective well-being more from the three disciplines of life quality, mental health and social gerontology. As sociologists and economists join the ranks of happiness research, the rich connotations and manifestations of happiness are more revealed [4]. It should be said that happiness, as a part of the social psychological system, is affected by many complex factors, such as economic factors such as employment status and income level, as well as social factors such as education level and marriage quality.

Regarding the definition of happiness, the Chinese government put forward the so-called happiness, which means the continuous improvement of people's lives through the continuous development of production and reform and opening up, so that everyone can lead a more decent life. With the process of China's reform and opening up, people's quality of life has been greatly improved in recent years. For example, people's income is increasing, and cultural life is rich and colorful, because the Chinese government firmly believes that income growth and living standards can improve people's happiness.

N. Radojević et al. (Eds.): ICAID 2022, AHIS 7, pp. 820–830, 2023.
https://doi.org/10.2991/978-94-6463-010-7_82

2 Sorting and Analyzing Data

2.1 Data Collating

The data is depicted from the "China's Happiest Cities Ranking" published by the China and Foreign Urban Competitiveness Research Association in 2019 (https://xw.qq.com/partner/sxs/20210322A03T5Z/20210322A03T5Z00?ADTAG=sxs&pgv_ref=sxs&ivk_sa=1024320u). The score is comprehensively derived from the "GN Happiness City Evaluation Index System," composed of 5 first-level indicators, 21 s-level indicators, and 47 third-level indicators (including satisfaction index, life quality index, ecological environment index, social civilization index, economic welfare index). In addition, from the statistics bureaus of each city, in 2019, the city's per capita disposable income, urban per capita disposable income, rural per capita disposable income, per capita GDP, the proportion of tertiary industries, the city's Engel coefficient and other data. Then, upload the excel document containing all the data to Jupyter Notebook to get the data summary (Fig. 1).

	City	Score	Total_income	City_income	Rural_income	GDP	Industry_I	Industry_II	Industry_III	Engel coefficient
0	Qin dao	94.67	45452.00	54484.0	22573.0	124282.00	0.0350	0.3560	0.6090	0.280
1	Hang zhou	94.06	59261.00	66068.0	36255.0	152465.00	0.0210	0.3280	0.6510	0.324
2	Yan tai	93.73	37783.00	47977.0	21218.0	107500.00	0.0720	0.4160	0.5120	0.308
3	Harbin	93.46	32104.00	40007.0	18238.0	53925.93	0.1080	0.2150	0.6770	0.355
4	Ji nan	92.43	43056.00	51913.0	19454.0	108540.22	0.0360	0.3460	0.6180	0.253
5	Zhu hai	91.59	52495.00	55219.0	29069.0	175500.00	0.0170	0.4450	0.5380	0.206
6	Xin yang	91.07	20928.00	30425.0	14010.0	29303.21	0.1800	0.3660	0.4540	0.332
7	Hui zhou	90.88	37159.60	42999.4	23027.4	66808.51	0.0490	0.5190	0.4320	0.351
8	Wei hai	90.35	39593.00	49044.0	22171.0	105070.92	0.0970	0.4040	0.4990	0.269
9	Zhao qin	89.65	26122.00	33260.0	19217.0	53936.00	0.1717	0.4115	0.4168	0.272
10	Cheng du	89.03	36142.00	45878.0	24357.0	103386.00	0.0360	0.3083	0.6557	0.331
11	Chong qing	88.98	28920.00	37939.0	15133.0	75828.00	0.1086	0.3838	0.5076	0.321
12	Su zhou	88.30	60109.00	68629.0	35152.0	179400.00	0.0430	0.4440	0.5130	0.259
13	Jin hua	87.69	48155.00	59348.0	28511.0	81200.00	0.0320	0.4080	0.5600	0.269
14	Kun ming	87.16	22082.43	46289.0	16356.0	93853.00	0.0420	0.3210	0.6370	0.370
15	Xu zhou	86.78	29736.00	36215.0	19873.0	81138.00	0.0955	0.4036	0.5009	0.294
16	Ning bo	86.15	56982.00	64886.0	36632.0	143200.00	0.0270	0.4820	0.4910	0.354
17	He fei	85.69	38806.00	45404.0	22462.0	115623.00	0.0020	0.2570	0.7400	0.466
18	Meizhou	85.02	22904.00	29235.0	16447.0	27000.00	0.1850	0.3120	0.5030	0.291
19	Bin zhou	84.63	28517.00	37378.0	17480.0	62643.00	0.0937	0.4237	0.4827	0.280
20	Si pin	84.21	25295.00	28290.0	14803.0	24979.00	0.3470	0.1950	0.4580	0.300
21	Mu danjiang	83.96	28569.00	34422.0	20045.0	42626.66	0.2170	0.2140	0.5690	0.319
22	Tai zhou	83.71	47988.00	60351.0	30221.0	63203.95	0.0550	0.4650	0.4800	0.303
23	Yan chen	83.26	32101.00	38813.0	22267.0	79200.00	0.1090	0.4160	0.4750	0.288
24	Bao ji	82.75	23151.00	34446.0	13094.0	59000.00	0.0800	0.5730	0.3470	0.458
25	Yue yang	82.46	27051.00	35116.0	16878.0	65200.00	0.1000	0.4040	0.4960	0.312
26	Yu lin	81.74	24213.00	33904.0	13226.0	120900.00	0.0606	0.6504	0.2890	0.321
27	Yan an	81.45	24450.00	34888.0	11876.0	73703.00	0.0897	0.6009	0.3093	0.308
28	Xin yu	81.06	26262.00	37557.0	17990.0	81642.00	0.0830	0.4420	0.4750	0.324
29	Tong chuan	80.64	24666.00	32504.0	10229.0	44100.00	0.0760	0.3680	0.5560	0.351

Fig. 1. Snapshot of the data

|r|'s value and degree of correlation

Range	Relevance
0.00-0.19	Extremely low
0.20-0.39	Low
0.40-0.69	Medium
0.70-0.89	High
0.90-1.00	Extremely high

Fig. 2. r's value and degree of correlation

	Score	Total_income	City_income	Rural_income	GDP	Industry_I	Industry_II	Industry_III	Engel coefficient
Score	1	0.477341	0.454591	0.395428	0.375775	-0.232131	-0.296116	0.468476	-0.262956
Total_income	0.477341	1	0.950715	0.935551	0.774768	-0.559977	0.021814	0.367779	-0.29776
City_income	0.454591	0.950715	1	0.899792	0.784258	-0.665615	0.073293	0.388214	-0.219388
Rural_income	0.395428	0.935551	0.899792	1	0.68572	-0.465089	-0.01854	0.343434	-0.246874
GDP	0.375775	0.774768	0.784258	0.68572	1	-0.708363	0.242425	0.242765	-0.215608
Industry_I	-0.23213	-0.559977	-0.665615	-0.465089	-0.70836	1	-0.385238	-0.298146	-0.110682
Industry_II	-0.29612	0.021814	0.073293	-0.01854	0.242425	-0.385238	1	-0.76599	-0.035337
Industry_III	0.468476	0.367779	0.388214	0.343434	0.242765	-0.298146	-0.76599	1	0.112788
Engel coefficien	-0.26296	-0.29776	-0.219388	-0.246874	-0.21561	-0.110682	-0.035337	0.112788	1

Fig. 3. Correlation coefficient form

2.2 Data Prepossessing and Mapping

We first find the direct correlation coefficient of each variable and draw a heat map of the correlation coefficient to make the data clearer. The correlation coefficient is a statistical indicator used to reflect the closeness of the correlation between variables. The statistical index reflecting the linear correlation of two variables is called the correlation coefficient (Pearson 1985) (Figs. 2, 3 and 4).

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}} \tag{1}$$

From the heat map of correlation matrix, the happiness index is positively correlated with the disposable income of residents (urban residents, urban residents, and rural residents), per capita GDP, and the proportion of the tertiary industry in each city. The happiness index has a low degree of negative correlation with the Engel coefficient of each city. The lower the Engel coefficient, the lower the proportion of household income used to buy food. The lower the Engel coefficient, the lower the proportion of household income used to buy food [6]. From the summary of the data, it can be seen that among the 30 happiest cities in China, the Engel coefficient of most cities is between 20% and 40%, which is at the level of wealth; a few cities have Engel coefficient between 40% and 50%. This is well-off grade. When the Engel coefficient of a city is lower than 50%, as the Engel coefficient decreases, the happiness index rises, but the relationship between the two is minimal. This shows that when food is no longer the main expenditure, food expenditure will hardly affect people's happiness index. In other

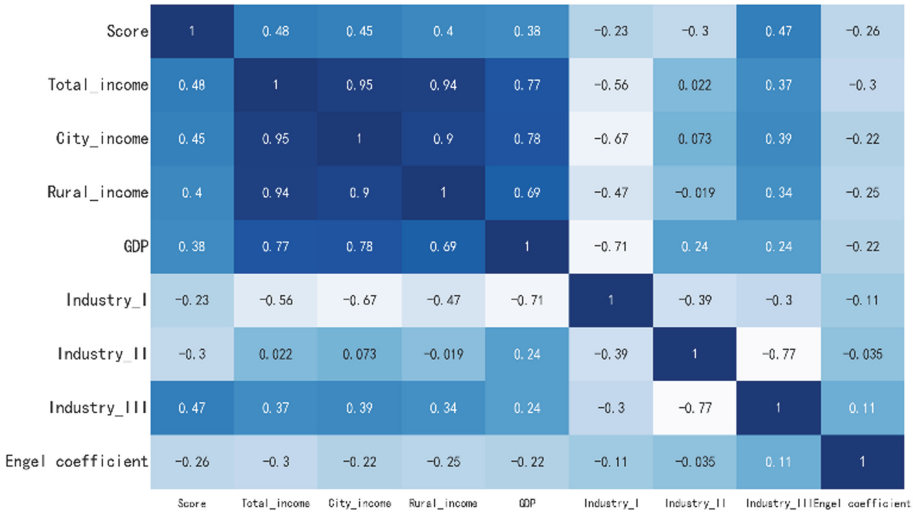

Fig. 4. Correlation coefficient thermal diagram

words, with the development of society and economy, the total amount of food has increased substantially compared with a few decades ago, and the types are also diverse. People no longer worry about not having enough food, but think about how to eat well.

2.3 Data Preprocessing and Normalization

It can be seen from the above correlation coefficient that as the people's disposable income and per capita GDP increase, the happiness index rises. In order to make the data more intuitive, we need to visualize the data. However, due to the significant differences in the dimensions of various indicators, when the total score is the same percentage as the change in income or per capita GDP, the amount of change in absolute value is too different. Therefore, we have to perform standard processing on the original data for the convenience of analysis and statistics. The purpose of normalization is to process data of different scales and dimensions so that it can be scaled to the same data interval and range to reduce the impact of scale, characteristics, and distribution differences on the model. Since we will perform cluster analysis on 30 cities later, Z-Score standardization performs better when distance is needed to measure similarity [3].

$$y_i = \frac{x_i - mean(x)}{\sigma} \tag{2}$$

Substitute each city's per capita disposable income and per capita GDP into the Z-Score formula and get the processed data set after calculation (Fig. 5).

	City	Score	Total_income	City_income	Rural_income	GDP	Industry_I	Industry_II	Industry_III	Engel coefficient
0	Qin dao	1.770031	0.900863	0.934153	0.234838	0.937134	0.0350	0.3560	0.6090	0.280
1	Hang zhou	1.625128	2.091269	1.943497	2.205002	1.636275	0.0210	0.3280	0.6510	0.324
2	Yan tai	1.546738	0.239756	0.367181	0.039722	0.520820	0.0720	0.4160	0.5120	0.308
3	Harbin	1.482601	-0.249802	-0.327265	-0.389388	-0.808201	0.1080	0.2150	0.6770	0.355
4	Ji nan	1.237929	0.694316	0.710135	-0.214288	0.546625	0.0360	0.3460	0.6180	0.253
5	Zhu hai	1.038391	1.508005	0.998196	1.170241	2.207708	0.0170	0.4450	0.5380	0.206
6	Xin yang	0.914867	-1.213230	-1.162170	-0.998207	-1.419020	0.1800	0.3660	0.4540	0.332
7	Hui zhou	0.869733	0.186016	-0.066530	0.300270	-0.488620	0.0490	0.5190	0.4320	0.351
8	Wei hai	0.743834	0.395788	0.460152	0.176951	0.460562	0.0970	0.4040	0.4990	0.269
9	Zhao qin	0.577552	-0.765481	-0.915149	-0.248416	-0.807951	0.1717	0.4115	0.4168	0.272
10	Cheng du	0.430274	0.098294	0.184290	0.491728	0.418764	0.0360	0.3083	0.6557	0.331
11	Chong qing	0.418397	-0.524279	-0.507456	-0.836498	-0.264872	0.1086	0.3838	0.5076	0.321
12	Su zhou	0.256866	2.164371	2.166643	2.046173	2.304456	0.0430	0.4440	0.5130	0.259
13	Jin hua	0.111963	1.133875	1.357966	1.089891	-0.131608	0.0320	0.4080	0.5600	0.269
14	Kun ming	-0.013936	-1.113712	0.220102	-0.660390	0.182277	0.0420	0.3210	0.6370	0.370
15	Xu zhou	-0.104203	-0.453936	-0.657672	-0.153954	-0.133146	0.0955	0.4036	0.5009	0.294
16	Ning bo	-0.253857	1.894808	1.840506	2.259289	1.406436	0.0270	0.4820	0.4910	0.354
17	He fei	-0.363128	0.327944	0.142989	0.218854	0.722329	0.0020	0.2570	0.7400	0.466
18	Meizhou	-0.522284	-1.042889	-1.265858	-0.647287	-1.476157	0.1850	0.3120	0.5030	0.291
19	Bin zhou	-0.614926	-0.559020	-0.556337	-0.498538	-0.591955	0.0937	0.4237	0.4827	0.280
20	Si pin	-0.714695	-0.836773	-1.348198	-0.884017	-1.526292	0.3470	0.1950	0.4580	0.300
21	Mu danjiang	-0.774082	-0.554537	-0.813901	-0.129186	-1.088503	0.2170	0.2140	0.5690	0.319
22	Tai zhou	-0.833468	1.119479	1.445360	1.336125	-0.578039	0.0550	0.4650	0.4800	0.303
23	Yan chen	-0.940364	-0.250061	-0.431302	0.190775	-0.181222	0.1090	0.4160	0.4750	0.288
24	Bao ji	-1.061512	-1.021596	-0.811810	-1.130108	-0.682327	0.0800	0.5730	0.3470	0.458
25	Yue yang	-1.130400	-0.685397	-0.753431	-0.585224	-0.528523	0.1000	0.4040	0.4960	0.312
26	Yu lin	-1.301433	-0.930046	-0.859035	-1.111100	0.853236	0.0606	0.6504	0.2890	0.321
27	Yan an	-1.370321	-0.909616	-0.773297	-1.305496	-0.317587	0.0897	0.6009	0.3093	0.308
28	Xin yu	-1.462964	-0.753412	-0.540740	-0.425100	-0.120643	0.0830	0.4420	0.4750	0.324
29	Tong chuan	-1.562733	-0.890996	-0.981021	-1.542659	-1.051954	0.0760	0.3680	0.5560	0.351

Fig. 5. Snapshot of the data by Z-Score

3 Data Visualization and Analysis

3.1 Plot the Relationship Between Happiness Index and Disposable Income

We resort to employ the linear regression equation of income and happiness index by the least square method, then use the disposable income as the abscissa and the happiness index as the ordinate to draw the relationship between the disposable income of the people in each city and the happiness index (Figs. 6, 7 and 8).

$$y = \bar{y} + \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}(x - \bar{x}) \tag{3}$$

From the above three charts, we can find that the happiness index and income show a trend that the higher the income, the higher the happiness index. In the following, we will choose the relationship between the city's residents' per capita disposable income and happiness index for further analysis.

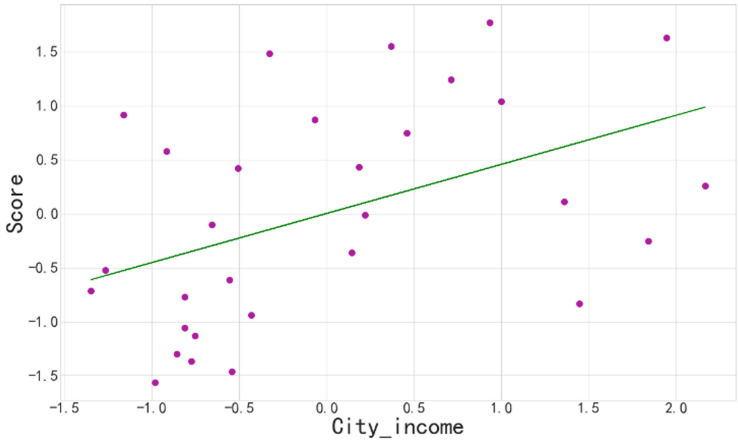

Fig. 6. Per capita disposable income of urban residents and happiness index

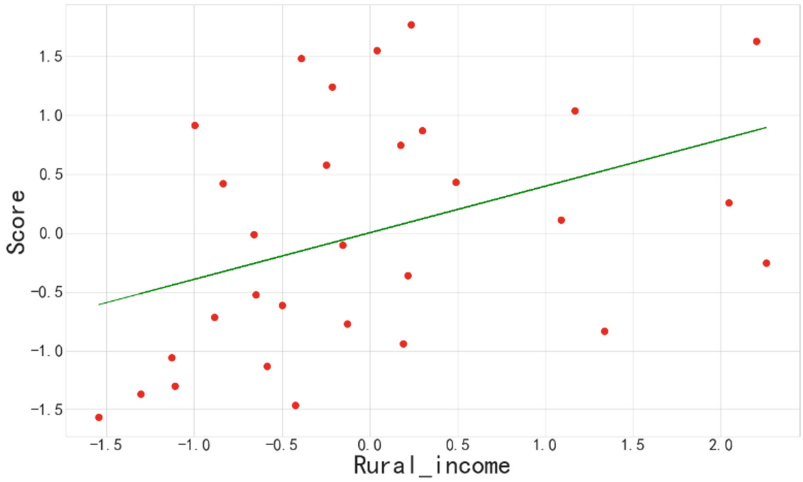

Fig. 7. Per capita disposable income of rural residents and happiness index

3.2 Cluster Analysis and Visualization of 30 Cities

We need to perform a cluster analysis of 30 cities to better judge the relationship between income and happiness index. At the same time, we can find out which cities have a higher happiness index under the premise of the same income. We choose to use the Kmeans algorithm to perform cluster analysis on 30 cities. In order to determine the optimal number of clusters (k value) in the Kmeans cluster, we need to use the Elbow method. The core index of the elbow method is SSE (sum of the squared errors). SSE is the clustering error of all samples and represents the quality of the clustering effect.

As the number of clusters k increases, the sample division will be more refined. The degree of aggregation of each cluster will gradually increase, then the error squared sum

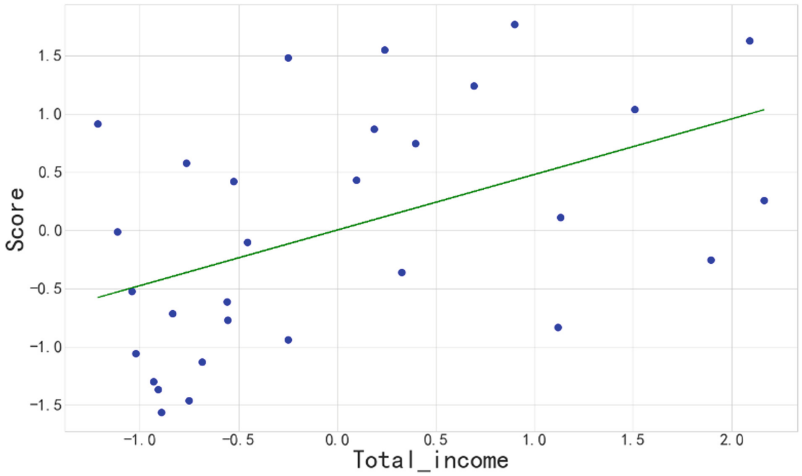

Fig. 8. Per capita disposable income of all residents and the happiness index

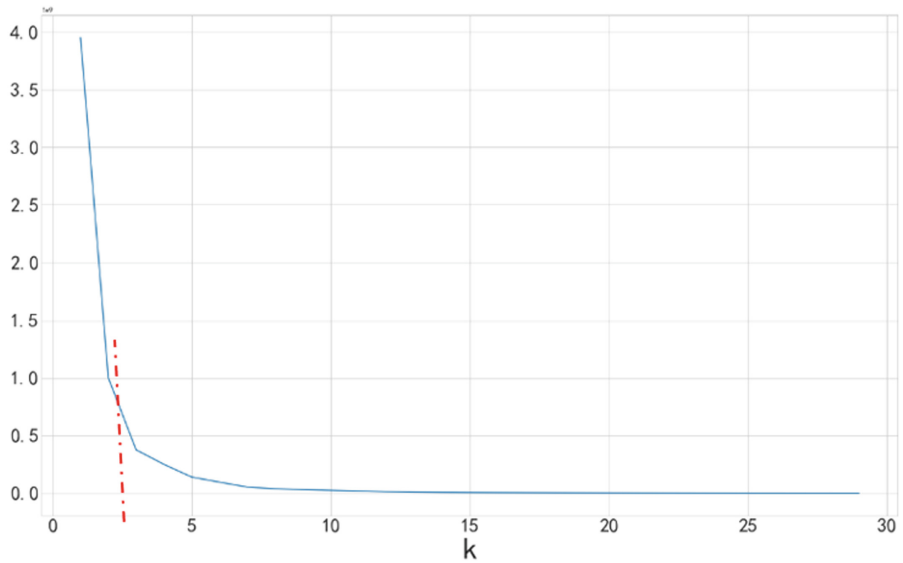

Fig. 9. Elbow method for Kmeans

SSE will gradually become smaller. When k is less than the true number of clusters, since the increase of k will significantly increase the degree of aggregation of each cluster, the SSE will decrease greatly. When k reaches the number of true clusters, the degree of aggregation obtained by k is increased. The return will decrease rapidly, so the decline of SSE will decrease sharply. Then it will be flat as the value of k continues to increase, which means the relationship between SSE and k is the shape of an elbow, and the k value corresponding to this elbow is the true number of clusters of the data [5] (Fig. 9).

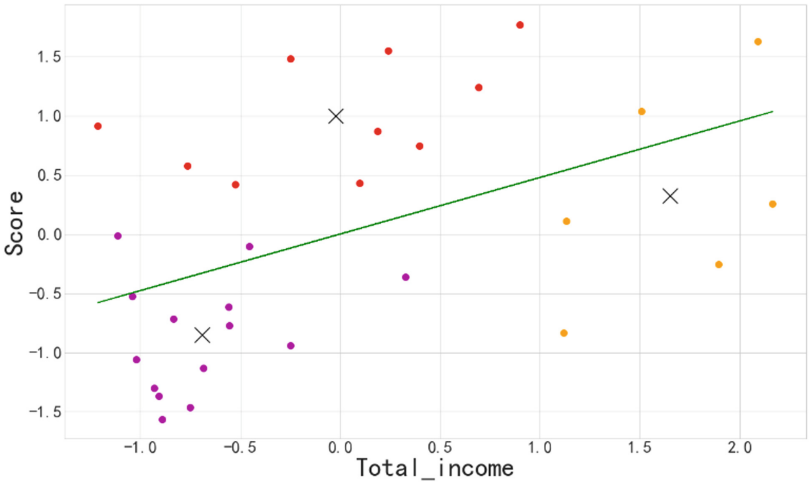

Fig. 10. Per capita disposable income of all residents and the happiness index after clusters

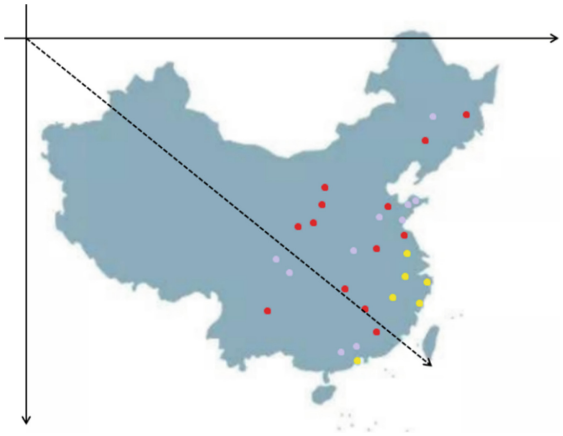

Fig. 11. Geographical location of each city

It can be seen from the figure above that the K value corresponding to the elbow is 3. Therefore, we divide the 30 cities into 3 clusters (Fig. 10).

According to the results of cluster analysis, the red dots in the figure represent: 'Kunming', 'Xuzhou', 'Hefei', 'Meizhou', 'Binzhou', 'Siping', 'Mudanjiang', 'Yancheng', 'Bao ji', 'Yueyang', 'Yu lin', 'Yan'an', 'Xin yu', 'Tong chuan'; The purple dots represent: 'Qing dao', 'Yantai', 'Harbin', 'Ji nan', 'Xin yang', 'Hui zhou', 'Wei hai', 'Zhao qing', 'Chengdu', 'Chongqing'; the yellow dots represent: 'Hangzhou', 'Zhuhai', 'Suzhou', 'Jinhua', 'Ningbo', 'Taizhou'.

Mark each city on the map to show its geographic location (Fig. 11).

According to the map and cluster diagram, we found that the closer to the coastal area, the higher the income required to obtain a higher happiness index. In the central

City	Tertiary Industry
Yello	0.538833
Purple	0.53811
Red	0.488421

Fig. 12. Snapshot of the data

	City	Total_income	Tertiary Industry
0	Hang zhou	40016	0.661696
1	Ning bo	33944	0.490603
2	Wen zhou	34107	0.551377
3	Jia xin	30547	0.438871
4	Hu zhou	29657	0.446201
5	Shao xin	31109	0.484637
6	Jin hua	30911	0.566033
7	Qu zhou	20635	0.531217
8	Zhou shan	32347	0.546566
9	Tai zhou	31768	0.489296
10	Li shui	25718	0.545323

Fig. 13. Snapshot of the data

region of China, although the per capita disposable income is not high, it is still able to obtain a high happiness index. On the other hand, in China's western and northern regions, the per capita disposable income is not high, and the happiness index is low.

Calculate the average proportion of the tertiary industry in these three types of cities (Fig. 12).

It can be seen from the table that the proportion of tertiary industry in coastal cities is more significant than that in inland cities. We can speculate that due to the formation of the influence of the tertiary industry on prices and its response to rising prices, prices in coastal areas are slightly higher than in inland cities. With the same income, the number of items that can be purchased in coastal areas will be smaller than in inland areas. In other words, to obtain the same happiness index as inland areas, people in coastal areas need higher incomes as support [1].

3.3 Predict the Eleven Cities in Zhejiang Province

Based on the above analysis of 30 cities, forecasts are made for 11 cities in Zhejiang Province. We firstly obtain the per capita disposable income of each city in Zhejiang Province and the proportion of the tertiary industry in 2019 (Fig. 13).

By substituting the per capita disposable income of each city in Zhejiang Province into the regression line equation obtained above, one can predict the happiness index of each city, and perform a cluster analysis of each city (Fig. 14).

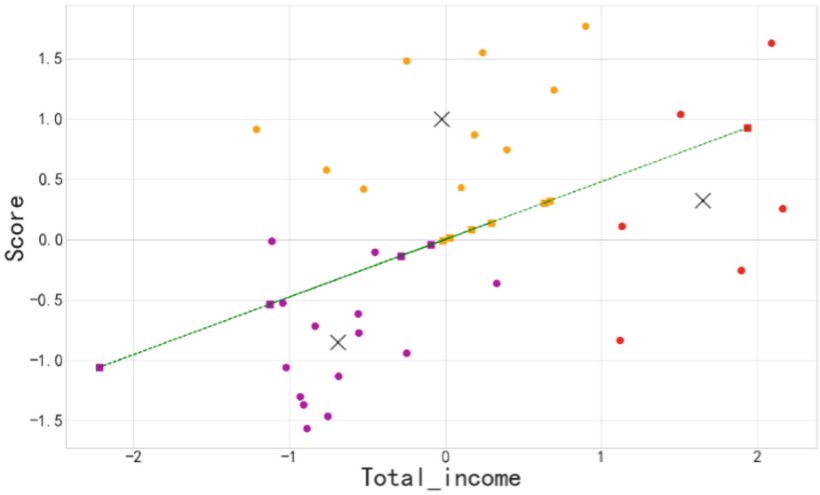

Fig. 14. Prediction for cities in Zhejiang Province

City	Tertiary Industry
Yello	0.661696
Purple	0.521419
Red	0.490403

Fig. 15. Snapshot of the data

According to the results of cluster analysis, the red dots in the figure represent: 'Jia xin', 'Hu zhou', 'Qu zhou', 'Li shui'; The purple dots represent: 'Ning bo', 'Wen zhou', 'Shao xin', 'Jin hua', 'Zhou shan', 'Tai zhou'; the yellow dots represent: 'Hang zhou'.

Calculate the average proportion of the tertiary industry in these three types of cities (Fig. 15).

From the above two charts, we find that Zhejiang Province has a similar pattern as the whole country. When the proportion of tertiary industry in a region reaches a certain level, residents need a relatively high income to ensure that they can obtain a high happiness index. Conversely, when the proportion of the tertiary industry is below a certain level, residents cannot obtain a relatively high income and, at the same time, cannot obtain a high happiness index [1].

4 Conclusion

It is found that income and happiness index are positively correlated. But the correlation coefficient is not significant, because happiness is a highly complex concept, and everyone has a different definition of happiness. But there is no doubt that income will affect everyone's happiness index, and the impact of income on the happiness of urban residents is greater than that of rural residents. This is related to the price gap between

urban and rural areas and the distribution of the tertiary industry. The tertiary industry indirectly affects people's happiness by affecting local prices and residents' income. The closer you are to the coastal areas, the higher your happiness, but the higher the disposable income you need. This is because the tertiary industry in coastal areas is densely distributed, making it easy for people to obtain happiness by consuming tertiary industry products. At the same time, the dense tertiary industry makes prices higher, which forces people to use higher incomes to maintain happiness.

References

1. Aygul H (2014) On the influence of the tertiary industry on price control. Modern Econ Inf (22):427
2. Pearson K (1895) Notes on regression and inheritance in the case of two parents. In Proceedings of the royal society of London 58:240–242
3. Kreyszig (1979) Advanced engineering mathematics, 4th edn., p 880, Eq 5. Wiley. ISBN 0–471–02140–7
4. Argyle M (2013) The psychology of happiness. Routledge
5. Bholowalia P, Kumar A (2014) EBK-means: a clustering technique based on elbow method and k-means in WSN. Int J Comput Appl 105(9)
6. Anker R (2011) Engel's law around the world 150 years later. Political Economy Research Institute, p 247