161.762 Multivariate Statistics for Big Data Project

Exploring world energy consumption in different countries and regions and its influence on GDP and happiness indicator

ID: 24008144
Name: James/Peng Li
Date: 03/06/2024

Contents

ABSTRACT

This study aims to apply techniques learned in the course “Multivariate Statistics for Big Data” to explore global energy trends in 2021 and their influence on GDP and happiness. The energy data is derived from the “bp Statistical Review of World Energy 2022/71st edition.” GDP and population data for 2021 are sourced from the “World Bank,” and happiness indicators are drawn from the “World Happiness Report 2022” for the year 2021. The energy dataset includes information on six types of energy—oil, natural gas, coal, nuclear energy, hydroelectricity, and renewable energy—from 64 representative countries across seven regions in 2021.

Our analysis focuses on the relationships among these six types of energy and their correlations with GDP and happiness indicators. Various multivariate statistical techniques, including PCA, factor analysis, cluster analysis, canonical discriminant analysis, and partial least squares analysis, were employed. The results reveal the distinct contributions of different energy types to GDP and happiness. These findings offer a deeper understanding of the importance of energy and its influence on both current living conditions and future prospects.

Oil, coal, and natural gas emerged as the top three most consumed energy sources, forming the foundation for the development of each country. However, their contributions to GDP and happiness indicators vary. These insights prompt us to consider energy policy through the lens of short-term GDP growth and long-term well-being. While we cannot abandon traditional energy resources in the short term, there is a pressing need to develop sustainable energy sources. These are not only vital for our daily happiness but also for future GDP growth. Sound energy policies are crucial for the future of our planet and humanity, encompassing ecology, environmental sustainability, and public health.

Due to data and content limitations, this study focuses primarily on GDP and happiness indicators for preliminary analysis.

1. INTRODUCTION

1.1 Main target

The purpose of the project is to explore world energy with multivariate statistics techniques learned in the "Multivariate Statistics for Big Data" course.

We choose "world energy" as the topic. There are two main reasons:

- Firstly, based on the "bp Statistical Review of World Energy 2022/71st edition" we can get solid and detailed data about the world energy consumption in different countries in 2021. High quality data is the fundamental basis for analysis. It includes detailed information about different energy consumption (Oil, Natural gas, Coal, Nuclear energy, Hydroelectricity, Renewable) from 64 representative countries in 7 regions.
- Secondly, energy is also very key for this world and our daily life. Along with the quick development of world, we have even higher requirement for energy. According bp World Energy 2022 report: "the pronounced dip in carbon emissions in 2020 was only temporary. The challenges and uncertainties facing the global energy system are at their greatest for almost 50 years, since the time of the last great energy shocks of the 1970s"[1]. We hope that through this primary study to understand more of the different energy types themselves, and also plan to know their influence on GDP (the key economic indicator of each country) and our daily life (world happiness indicator).

So besides the data about the different energy consumption, we also collect data about GDP, population and happiness indicator. Based on these data:

- Firstly, we will have the exploratory data analysis for better understanding of the data in this data sets.
- Then based on the primary understanding, we will provide a list of the hypotheses (MULTIVARIATE in nature) that we would like to examine , and also propose related techniques that we will use to examine each hypothesis.
- Following this plan, in next part we will execute the study with planed techniques in detail. We will use SAS software and extract tables and graphs for this study.
- Finally we will summarize the key findings.

1.2 Data set description

Variables description

This data set contains data from 64 countries in 7 regions. It has 11 variables, in which 9 variables are in type of number and 2 variables are in type of character. Because the latest released edition is in 2021, we take this year data as basis for analysis. More details about the data are as blow:

1. Country (64 countries)
2. Regions: (North America, Cent. America, Europe, Asia Pacific, Middle East, CIS(Commonwealth of Independent States), 7 regions)

3. Oil (in exajoules)
4. Natural gas (in exajoules)
5. Coal (in exajoules)
6. Nuclear energy (in exajoules)
7. Hydroelectricity (in exajoules)
8. Renewable (in exajoules)
9. GDP (in US dollar)
10. Population
11. Happiness (indicator)

Data resource :

Item	Resource
World energy in 2021	bp Statistical Review of World Energy 2022 | 71st edition [1]
GDP in 2021	World Bank [2]
Population 2021	World Bank [3]
Happiness	World Happiness Report 2022 [4]

1.3 Data overall check

- To show the variables' type, length, format etc.

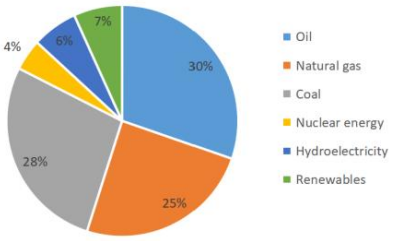
Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
5	Coal	Num	8	BEST6.		Coal
1	Country	Char	20	$20.	$20.	Country
9	GDP	Num	8	BEST20.		GDP
11	Happiness	Num	8	BEST18.		Happiness
7	Hydroelectricity	Num	8	BEST10.		Hydroelectricity
4	Natural gas	Num	8	BEST12.		Natural gas
6	Nuclear energy	Num	8	BEST9.		Nuclear energy
3	Oil	Num	8	BEST6.		Oil
10	Poplulation	Num	8	BEST18.		Poplulation
2	Region	Char	15	$15.	$15.	Region
8	Renewables	Num	8	BEST9.		Renewables

2 variables are in type of character. Others are in type of numeric.

- To show the variables' amount, mean, standard deviation, minimum, maximum, and frequency of variables in type of number.

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
Oil	Oil	64	2.6623438	5.8216244	0.0500000	35.3300000
Natural gas	Natural gas	64	2.1773438	4.5189967	0	29.7600000
Coal	Coal	64	2.4285937	11.0238520	0	86.1700000
Nuclear energy	Nuclear energy	64	0.3856250	1.1262790	0	7.4000000
Hydroelectricity	Hydroelectricity	64	0.5681250	1.6530212	0	12.2500000
Renewables	Renewables	64	0.6023438	1.7058803	0	11.3200000
GDP	GDP	63	1.443586E12	3.6719329E12	24496505941	2.33151E13
Population	Population	64	97596019.23	246862744	1525663.00	1412360000
Happiness	Happiness	61	6.0255410	0.9091245	3.7770000	7.8210000



In general, GDP and population are big scale number, we will consider to transform them later according to the analysis requirement. Other data are between 0 to 10.

Oil takes around 30% of the whole energy consumption, then coal with 28% and nature gas with 25%.

- To show the correlations to get overall impression about the relationship between different variables.

-> Oil is quite correlated with natural gas, nuclear and renewables. Wish to know more about the correlations between them. Especially combined with the country and region information.

-> GDP is very correlated with oil, as oil the most important energy currently.

-> Happiness does not have close relationship with energy consumption. But we will study more from the energy per capita side. And we see negative coefficient here, so it can be interesting to further study about it.

2. Hypotheses and techniques

Hypothesis 1: clusters of different energy types

For this study, we focus on the consumption of different types of energy, so firstly we would like to know more of the relationship between them. And we think that there will be some clusters of these various energy types. Based on this, we will use multivariate statistics techniques:

- Principal component analysis (PCA) is a linear dimensionality reduction technique with applications in exploratory data analysis, visualization and data preprocessing[5]. There are different types of energy items, plan to use PCA to understand them more.
- Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. Factor analysis searches for such joint variations in response to unobserved latent variables[6]. So it is suitable for examining this hypothesis.

Hypothesis 2: different countries and regions have different energy consumption speciality

In general different countries have different energy consumption structure, based on their location, resource, development status, economic and industrial status. So firstly we will use cluster analysis to find similarity of different countries. And then will use canonical discriminant analysis to analyze from the region side.

- Cluster analysis, also known as clustering, is a statistical technique used in machine learning and data mining that involves the grouping of objects or points in such a way that objects in the same group, also known as a cluster, are more similar to each other than to those in other groups[7].
- Canonical discriminant analysis, it is a dimension-reduction technique related to principal component analysis. Given a classification variable and several interval variables, canonical discriminant analysis derives canonical variables (linear combinations of the interval variables) that summarize between-class variation in much the same way that principal components summarize total variation[8].

Hypothesis 3: energy types and GDP&population have influence on each other

We take GDP and population as one group, as GDP is very closed with population. Normally big countries with high population also have high GDP. Of course we also see some small population countries also have high GDP, but this is not common phenomenal. We assume that there are some association between factors in the two groups. We will use multivariate statistics techniques:

- Canonical correlation analysis, originally defined by Hotelling in 1935 (Hotteling 1935; Hotteling 1936, see also Bartlett 1948), canonical correlation analysis (CCA) is a statistical method whose goal is to extract the information common to two data tables that measure quantitative variables on a same set of observations[9]. So this technique is very useful for analyzing the two groups data.

Hypothesis 4: energy consumption has high association with GDP

This part is deeper analysis based on hypothesis 3 results. We directly study the association of different types of energy with GDP. Because the contribution and importance of different types of energy for GDP should be different. And different countries have different energy resource and consumption structures, so the association should also be diversified from one country to another. We would like to know more about this part and the technique we will use is:

- Partial least squares analysis, partial least squares regression (PLS regression) is a statistical method that bears some relation to principal components regression; instead of finding hyperplanes of maximum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space[10].

Hypothesis 5: energy per capita has association with happiness

Can renewable energy lead to happiness? we can see some studies on websites from this angle. There are also many interesting finding behind it. Seems the two parts are distant from each other, but actually it is not so far from our daily life and also our daily happiness. Because we need to pay the bills for energy everyday. And the energy structure actually influence not only our daily energy cost but nearly all the cost around us, because nearly every goods needs energy to be produced. If energy price increases, nearly all the product prices increase, even the plants price, as they are also need energy to be transported. So energy is deeply influencing our daily life from nearly every sides. We also

see that there are many different kinds of energy types, so we want to know which energy types contribute more to our happiness.

- Partial least squares analysis will also be used for this part. But we will do one additional treatment for the data. We will add one new variable: energy per capita=energy/(population/1,000,000). Because the energy is demonstrated with large unit exajoules, so we also transformed the population into million, which make the energy per capita to have similar range with happiness ranking (0 to 10).

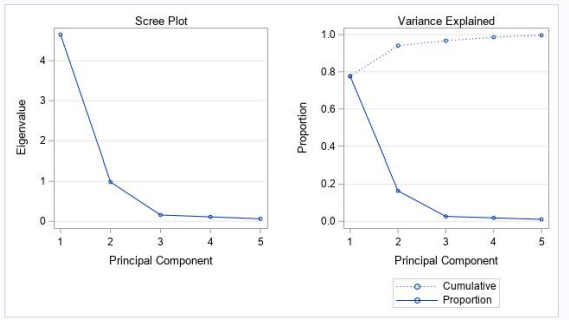
3. Hypothesis 1: clusters of different energy types

3.1 Principal components analysis

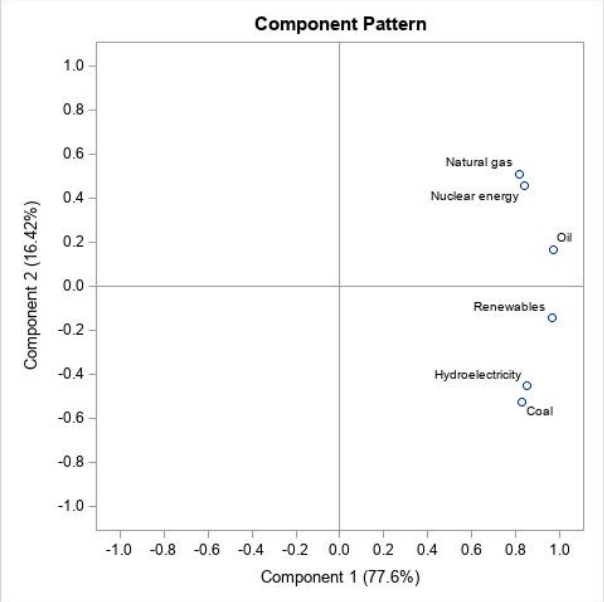
Principal components analysis was utilized to visualize how all the energy types interrelate. The technique was utilized because it efficiently reduces the number of dimensions required to visualize multiple variables together.

Eigenvalues of the Correlation Matrix

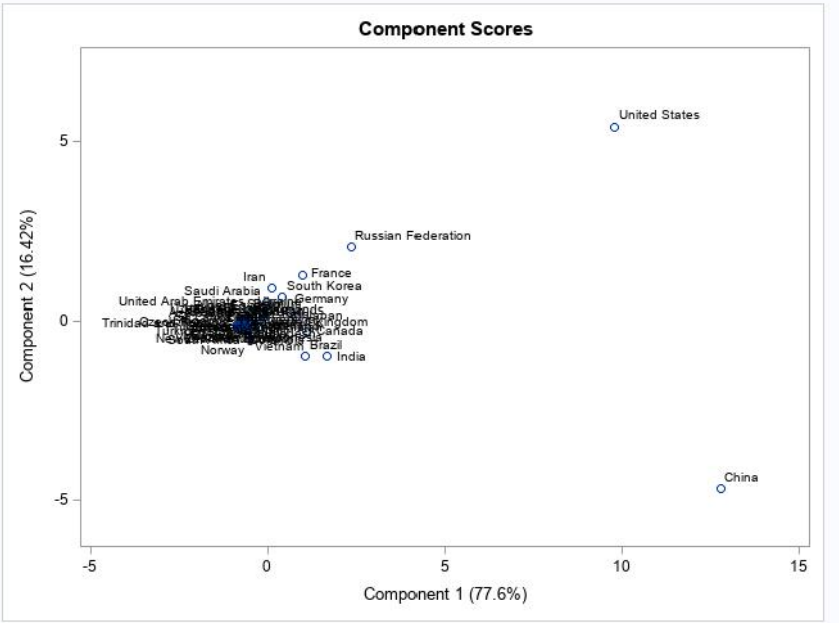
	Eigenvalue	Difference	Proportion	Cumulative
1	4.65621310	3.67123752	0.7760	0.7760
2	0.98497558	0.82534991	0.1642	0.9402
3	0.15962567	0.04704813	0.0266	0.9668
4	0.11257754	0.04836375	0.0188	0.9856
5	0.06421379		0.0107	0.9963



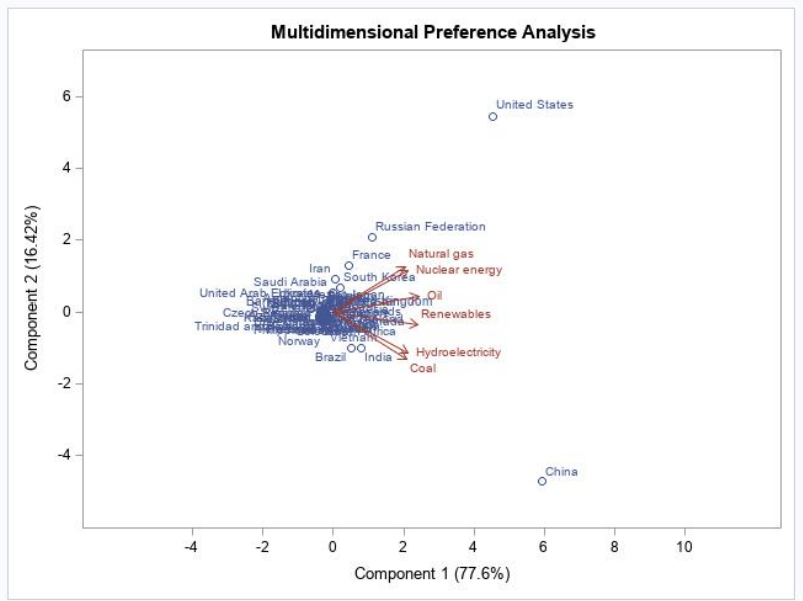
The above figure contains 5 eigenvalues by the principal components procedure. The first two eigenvalues explain 94% of the variation among the different indicators. Although two principal component axes are sufficient to visualize the variation in the variables, the third component was plotted to see what additional information it could provide.



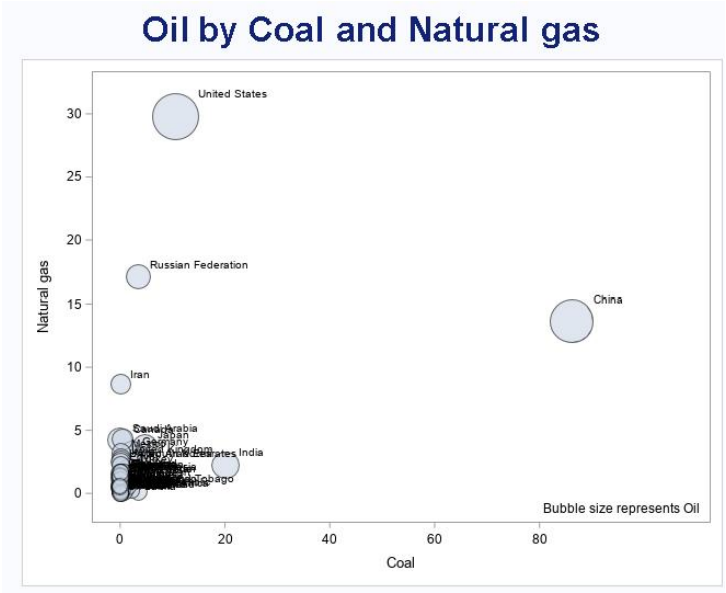
The component pattern plot for the first two principal components shows that natural gas and nuclear energy are most associated. Normally natural gas and nuclear energy need high requirement for application and production, so only "high technical" countries are the main consumers. Coal and hydroelectricity are also closely placed, normally "big energy consumption" countries rely on these two energy types.



Projecting country information onto the first two principal component axes, it is clear to see a clear pattern in in the way countries are placed. United States, China and Russian Federation are clearly separated from others. Both of United States and China are high energy consumption countries. But they locate in different area which indicates different energy consumption structures. Russian Federation is big energy consumption but also big energy supplier in the world, which makes it very unique.



The biplot above based on the first two components illustrate the country data points with energy types. You can find that United States and China are quite separated from others and have clearly different energy consumption structure. United States relies more on natural gas and nuclear energy, while China relies more on the coal and hydroelectricity energy. Russian Federation, France South Korea and Germany have similar scenario as United Sates. Indian and Brazil also have similar situation as China as developing countries .



The bubble plots above explains more from the oil side. We can find that United States, China, Russian Federation, Indian and Iran are big oil consumption and production countries, but have different rely on coal and natural gas. More clear about their energy consumption structure. Other counties because their relative smaller energy consumption, it is hard to show their speciality on this bubble plots graph.

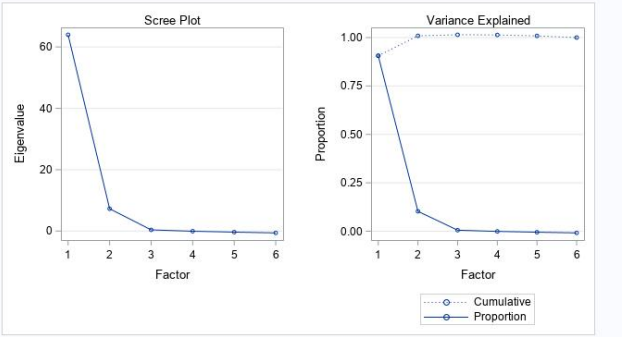
3.2 Factor Analysis

Factor analysis was also carried out to compare with the results of PCA analysis.

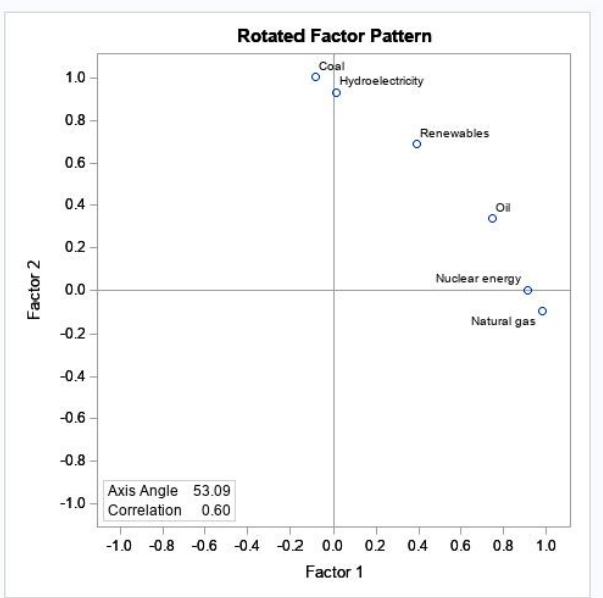
Preliminary Eigenvalues: Total = 70.6045891 Average = 11.7674315

	Eigenvalue	Difference	Proportion	Cumulative
1	63.9683959	56.7039971	0.9060	0.9060
2	7.2643988	6.8784414	0.1029	1.0089
3	0.3859574	0.4433254	0.0055	1.0144
4	-0.0573680	0.2927390	-0.0008	1.0136
5	-0.3501071	0.2565808	-0.0050	1.0086
6	-0.6066878		-0.0086	1.0000

factors will be retained by the PROPORTION criterion



The scree plot above suggests one factor. A second elbow also occurs at eigenvalue number 3, suggesting a possible 2-factor solution as well. We will analyze with two factors.



Similar conclusion as in the PCA analysis. Natural gas and nuclear energy are most associated and explained well by factor 1. Coal and hydroelectricity are also closely placed and well explained by factor 2. Renewable energy and Oil are relatively independent from others.

3.3 Conclusion

From this part with PCA and factor analysis we can conclude that the different types of energy has different speciality and can be grouped. Natural gas and nuclear energy are very associated, and coal and hydroelectricity are also closely correlated with each other. Renewables and Oil are different, and can not grouped with others.

Besides these, additionally we also primarily find the different energy consumption structures, especially about big energy consumption and production countries such like United States, China, Russian Federation, Indian, Brazil, Iran etc. But not so clear with other countries, so we will further investigate this in next part.

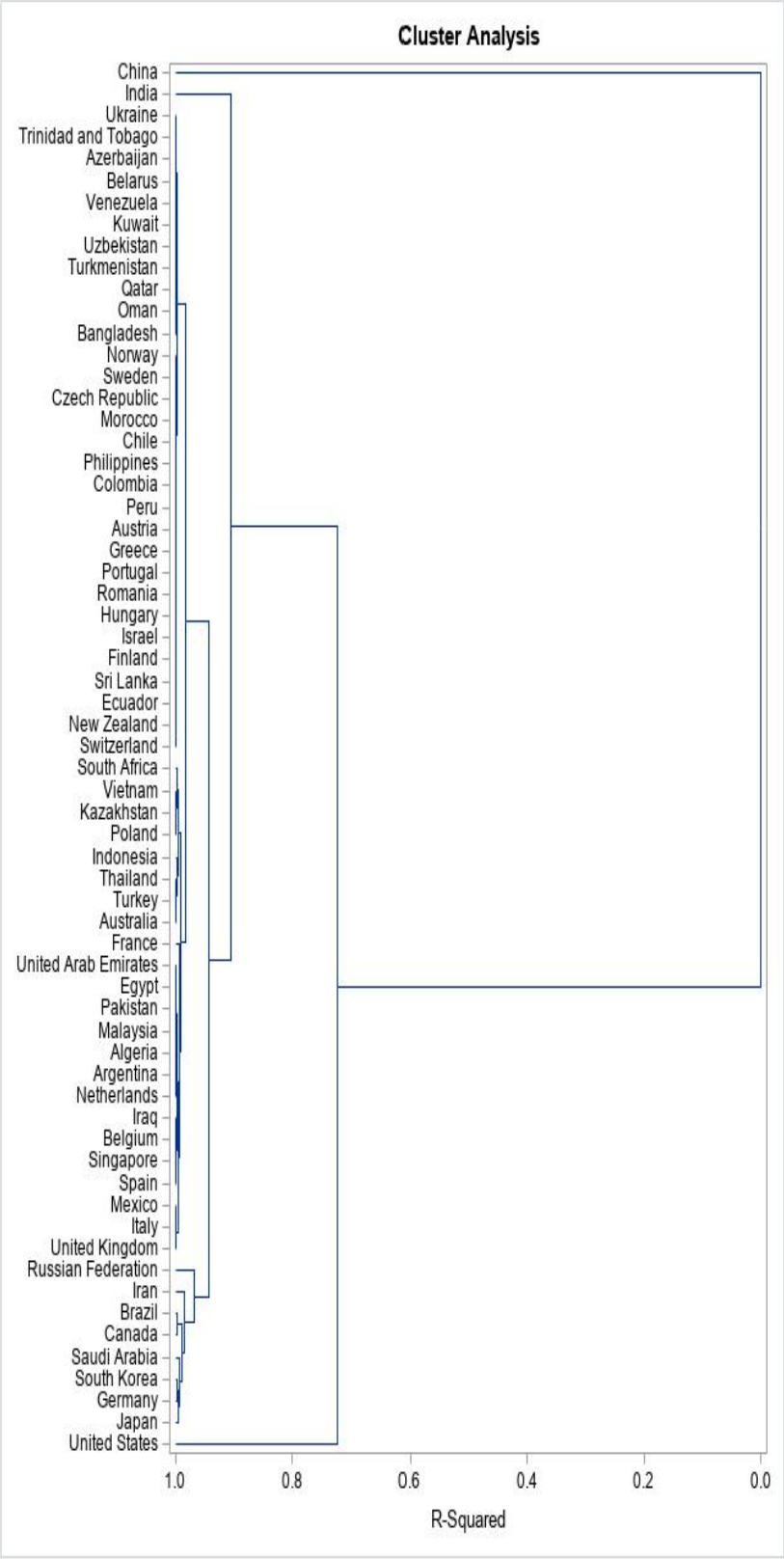
So hypotheses 1 can be generally accepted.

4. Hypothesis 2: different countries and regions have different energy consumption speciality

For this part, we will study the relationship between different energy types with countries and also regions. Not limited to big countries, but we wish to look into detail of other countries.

4.1 Cluster analysis

Cluster analysis was conducted to visualize which countries are most similar in terms of the development status.



- Same as it is shown in PCA analysis, China, United States and Indian are separately different from other countries.
- South Korean, Germany and Japan are grouped together. From economic level side this is also reasonable.
- Brazil and Canada are similar with rich resource of energy and big size.
- We also see some Europe countries such like Norway, Sweden, Czech Republic, Greece, Portugal, Romania , Hungary etc. are grouped together, as they are near each other and so similar with each other from size and location.
- New Zealand is similar as Israel, Finland, Sir Lanka, Ecuador and Switzerland. Geographically they are either near the sea or from resource side relatively independently from other countries.

4.2 Canonical Discriminant Analysis

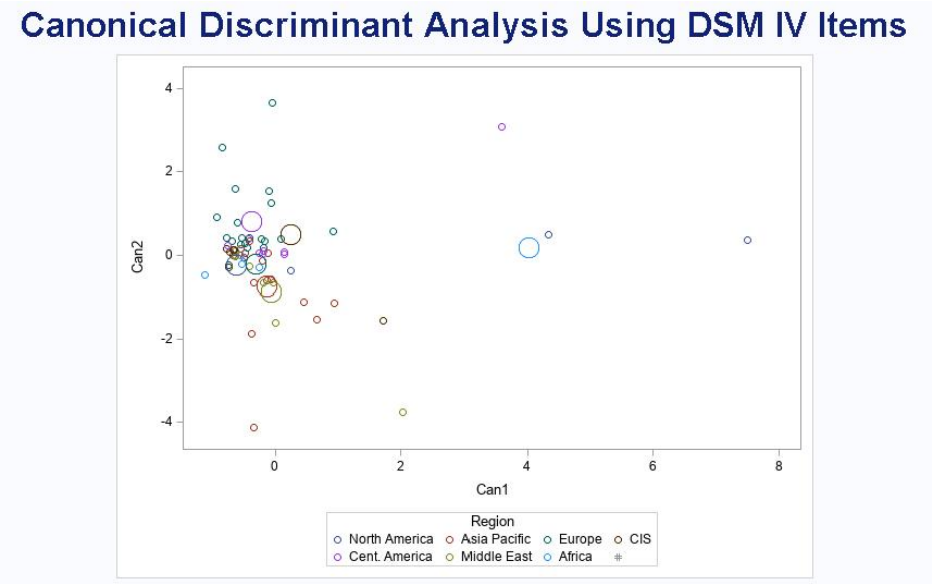
In the cluster analysis part, we studied the cluster from the country side. Here we would like to try to study more from the region side. To see if energy structure has specialty in different regions.

Using Proc Candisc

Pooled Within Canonical Structure							
Variable	**Label**	**Can1**	**Can2**	**Can3**	**Can4**	**Can5**	**Can6**
Oil	Oil	0.511060	-0.218657	-0.199606	0.607460	0.220925	0.483062
Natural gas	Natural gas	0.599141	-0.186330	0.302320	0.554347	0.100186	0.444495
Coal	Coal	0.035568	-0.260451	-0.323276	0.413951	0.642762	0.491827
Nuclear energy	Nuclear energy	0.526175	0.076993	-0.005110	0.818000	0.166486	0.142631
Hydroelectricity	Hydroelectricity	0.245306	-0.105637	-0.254067	0.206810	0.792757	0.439182
Renewables	Renewables	0.318548	-0.052132	-0.327041	0.507641	0.385317	0.618615

The variables that correlate highly with the first discriminant function are oil, nature gas, nuclear energy. It appears that these variables are responsible for an important portion of the discrimination among the region groups.

The variables that correlate highly with the second discriminant function is coal. It appears that this variable is responsible for an important portion of the discrimination among the groups.



It is easy to see the pattern of how the groups fall on the discriminant functions. As you saw before, the first function distinguishes from oil, nature gas, nuclear energy, while the second function discriminates coal energy. And we can see that the observations in region Europe nearly group in same area

Stepwise Discriminant Analysis

Stepwise Selection Summary

Step	Number In	Entered	Removed	Label	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	Natural gas		Natural gas	0.2952	3.98	0.0022	0.70478137	0.0022	0.04920311	0.0022
2	2	Oil		Oil	0.2162	2.57	0.0284	0.55240304	0.0005	0.08494188	0.0005
3	3	Coal		Coal	0.1901	2.15	0.0619	0.44740813	0.0002	0.11274217	0.0003
4	4	Hydroelectricity		Hydroelectricity	0.2638	3.22	0.0088	0.32938057	<.0001	0.14844064	<.0001
5	5	Renewables		Renewables	0.2241	2.55	0.0303	0.25555198	<.0001	0.18249973	<.0001

The final model after using the stepwise selection method includes all the variables except nuclear. This result can be used in next part.

Linear Discriminant Analysis with PROC DISCRIM

Based on the variables except nuclear, we had below analysis result.

Number of Observations and Percent Classified into Region

From Region	Africa	Asia Pacific	CIS	Cent. America	Europe	Middle East	North America	Total
Africa	0	1	0	0	3	0	0	4
	0.00	25.00	0.00	0.00	75.00	0.00	0.00	100.00
Asia Pacific	0	7	0	0	8	0	0	15
	0.00	46.67	0.00	0.00	53.33	0.00	0.00	100.00
CIS	0	0	1	0	5	0	0	6
	0.00	0.00	16.67	0.00	83.33	0.00	0.00	100.00
Cent. America	0	0	0	1	7	0	0	8
	0.00	0.00	0.00	12.50	87.50	0.00	0.00	100.00
Europe	0	0	0	1	19	0	0	20
	0.00	0.00	0.00	5.00	95.00	0.00	0.00	100.00
Middle East	0	2	1	0	5	0	0	8
	0.00	25.00	12.50	0.00	62.50	0.00	0.00	100.00
North America	0	0	0	0	1	0	2	3
	0.00	0.00	0.00	0.00	33.33	0.00	66.67	100.00
Total	0	10	2	2	48	0	2	64
	0.00	15.63	3.13	3.13	75.00	0.00	3.13	100.00
Priors	0.0625	0.23438	0.09375	0.125	0.3125	0.125	0.04688	

Error Count Estimates for Region

	Africa	Asia Pacific	CIS	Cent. America	Europe	Middle East	North America	Total
Rate	1.0000	0.5333	0.8333	0.8750	0.0500	1.0000	0.3333	0.5313
Priors	0.0625	0.2344	0.0938	0.1250	0.3125	0.1250	0.0469	

30 out of 64 observations were misclassified in this example, and 34 were correctly classified. Furthermore, 53.13% of the total regions would be misclassified using these functions. So overall this is not good. But if check the results of each region, we can find that Europe is 5% misclassified, so the countries in this region has very similar energy structure. After that is North American with 33.3% misclassified. Africa and Middle East have 100% misclassified, so that means the two regions have diversified energy consumption structure.

4.3 Conclusion

From the country and regions sides, we see some patterns and similarity of countries and countries in different regions. Some countries have similar energy structure even they are in different regions, this is based on the similarity of their existing natural resource basis and country development level. And some regions have similar energy structure for most the countries in its regions, so besides existing resource and country developing level, the geography also plays a role behind the energy structure, which is also related to energy transportation and relationship or cooperation between these countries.

5. Hypothesis 3: energy types and GDP&population have influence on each other

5.1 Canonical correlation analysis

The analysis demonstrates which energy type is most related to GDP and Population. Standardized variance output for analysis has been interpreted, because not all of the variables are of the same scale.

Test of H0: The canonical correlations in the current row and all that follow are zero				
Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
0.00551195	114.30	12	110	<.0001
0.40918644	16.17	5	56	<.0001

Multivariate Statistics and F Approximations					
	S=2 M=1.5 N=26.5				
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00551195	114.30	12	110	<.0001
Pillai's Trace	1.57734304	34.83	12	112	<.0001
Hotelling-Lawley Trace	74.68004822	338.03	12	82.525	<.0001
Roy's Greatest Root	73.23617447	683.54	6	56	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

According to the likelihood-ratio tests, the first two correlations are statistically different from 0. You reject the null hypothesis of all canonical correlations=0 from the multivariate statistics.

Canonical Structure

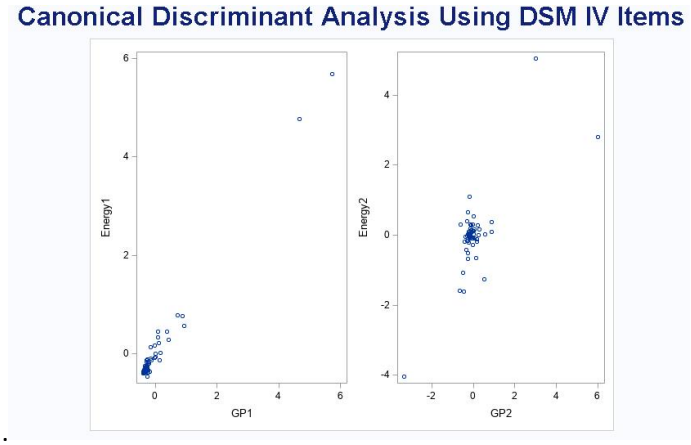
Correlations Between the Energy indicators and Their Canonical Variables			
		Energy1	Energy2
Oil	Oil	0.9863	0.0711
Natural gas	Natural gas	0.8356	-0.2324
Coal	Coal	0.7074	0.6513
Nuclear energy	Nuclear energy	0.8881	-0.2701
Hydroelectricity	Hydroelectricity	0.7077	0.4712
Renewables	Renewables	0.9472	0.2087

Correlations Between the GP indicators and Their Canonical Variables			
		GP1	GP2
GDP	GDP	0.9976	-0.0693
Population	Population	0.6204	0.7843

Correlations Between the Energy indicators and the Canonical Variables of the GP indicators			
		GP1	GP2
Oil	Oil	0.9796	0.0546
Natural gas	Natural gas	0.8299	-0.1787
Coal	Coal	0.7026	0.5006
Nuclear energy	Nuclear energy	0.8821	-0.2076
Hydroelectricity	Hydroelectricity	0.7029	0.3622
Renewables	Renewables	0.9408	0.1604

Correlations Between the GP indicators and the Canonical Variables of the Energy indicators			
		Energy1	Energy2
GDP	GDP	0.9909	-0.0532
Population	Population	0.6162	0.6028

Canonical structure output shows that out of the energy indicators, oil and renewables are most closely positively correlated with the canonical variables of GDP and population indicators. Oil represents basic energy and renewable is the future trend. GDP is strongly correlated with the canonical variables of energy. This is reasonable, because the economic results depend highly on energy consumption.



We can see the strong positive association between the first pair of canonical variates.The second pair's association is not strong, but the overall trend is still positive.

5.2 Conclusion

So we can see that the energy types oil and renewable energy are very associated with GDP. Compared with GDP, the energy association is not that strong associated with population, that means there is inequality of energy consumption between different countries. Some under developed countries relatively consume less compared to its population, and the highly developing and developed countries consumes relatively more and rely more on oil and seek more of the renewable energy. So this hypothesis can be generally accepted.

6. Hypothesis 4: GDP has high association with energy consumption

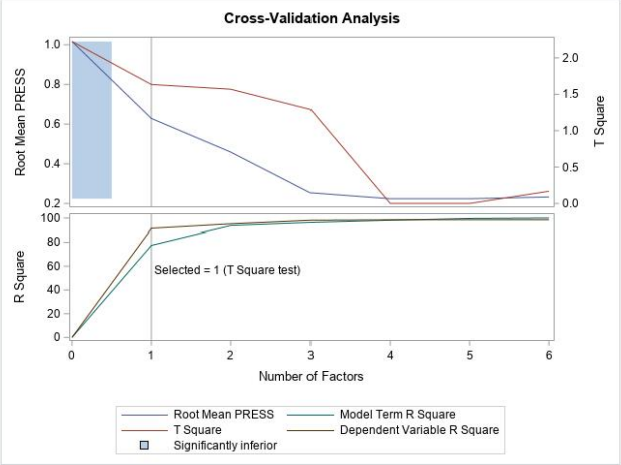
6.1 Partial least squares analysis

From above part we got the oil and renewable energy are most associated with GDP. For further investigate it, we will use partial least squares analysis for detailed analyzing the association between GDP and energy consumption.

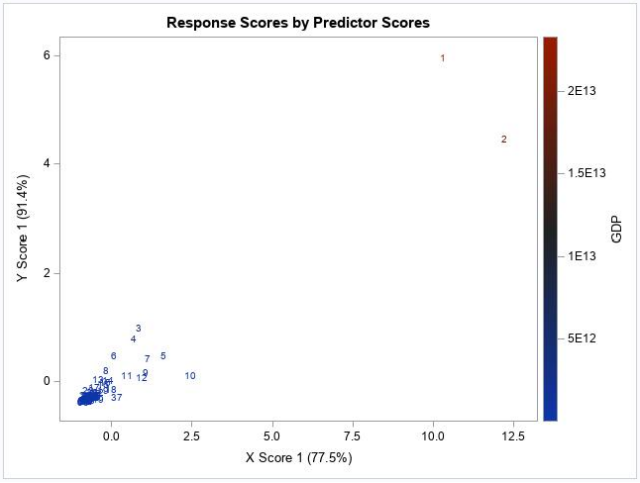
A large number of studies explored the relationship between energy consumption and GDP growth. Most of these studies focused on the causal relationship between these components. As an example, Narayan et al. (2010a) found a positive relationship between energy consumption and GDP growth in different developed and developing countries[11].

Percent Variation Accounted for by Partial Least Squares Factors				
	Model Effects		Dependent Variables	
Number of Extracted Factors	Current	Total	Current	Total
1	77.4638	77.4638	91.3805	91.3805

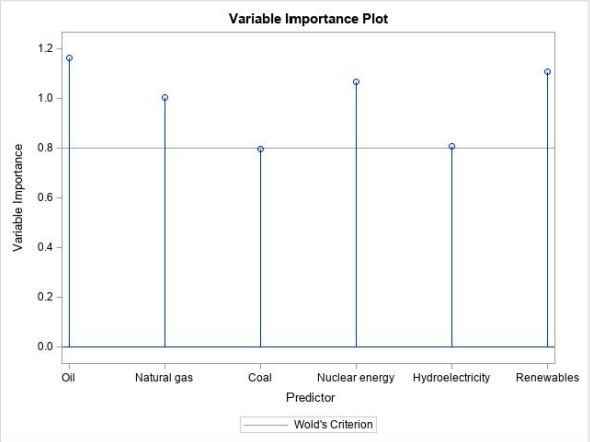
The percent variation table shows how well the predictors and the response are predicted by each factor. Only one factor is shown. The factor accounts for about 91.38% of the variance in GDP.



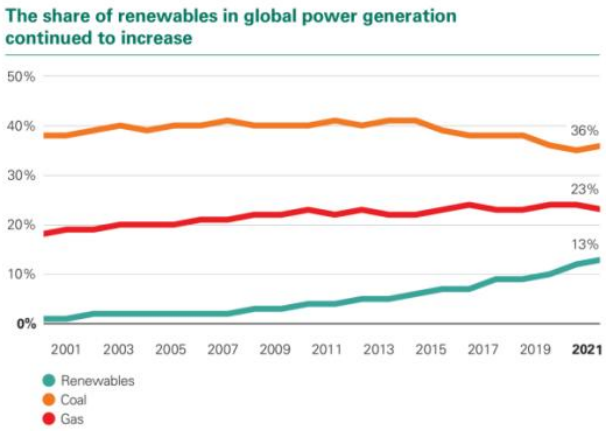
The Cross-Validation Analysis Plot enables you to see the variance explained. The root mean PRESS for up to 6 factors. The root mean PRESS has a clear minimum at 3 factors, after which overfitting is evident.



From above chart, we can find that observations 1 (United States), 2 (China) obviously stand out due to their high values of at least one of the factors as well as GDP. 3 and 4 represent Japan and Germany, they also show high values.



- From above variable importance chart, we can find that all energy types are above the threshold 0.8.
- Oil plays the most role for GDP. Oil can not just be used for energy but can also used for material for chemical industry. So it is fundamental important for production, which is directly related to GDP (Gross Domestic Product). So we see many conflicts in the world related to oil supply, and oil price also plays an very important role in global economic development.
- The second and third one are renewables and nuclear energy. Currently energy resource is limited because of even higher demand, and also because of limited and unsustainable supply of traditional energy resource such like oil and coal. So many countries which have high demand for energy and high GDP increase are developing renewable energy and even nuclear energy to fill the gap of energy demand and supply. According to "bp Statistical Review of World Energy 2022": renewable energy, led by wind and solar power, continued to grow strongly and in 2021 accounts for 13% of total power generation, more details please see bellow picture from "bp Statistical Review of World Energy 2022/71st edition". Renewable generation increased by almost 17% in 2021 and accounted for over half of the increase in global power generation over the past two years. Nuclear generation increased by 4.2% – the strongest increase since 2004 – led by China[1].



- Natural gas is coming after that. Natural gas is considered an important bridge in the transition of energy in the world. Some studies also found a positive relationship between CO2 emissions and natural gas consumption, GDP[12].

- Coal and hydroelectricity are relatively have lower contribution to GDP. Coal is a very traditional resource which has been used for many years. It can also be widely used in the world, no matter it is developed or underdeveloped countries. Hydroelectricity has very low percentage in the whole energy and its generation is limited to the geographic situation.

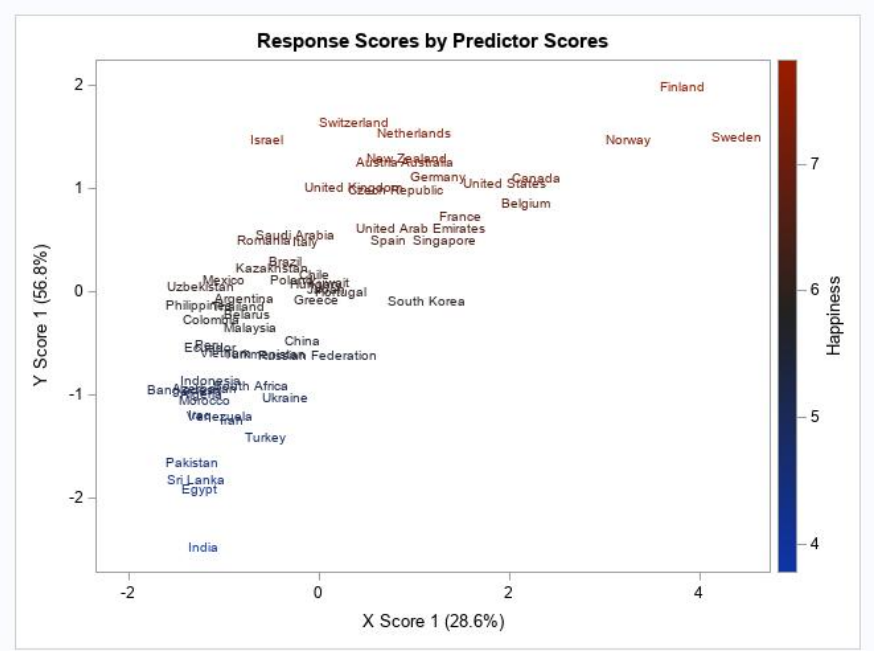
6.2 Conclusion

So we can see the clear association of different types of energy with GDP. Oil, renewable energy and nuclear energy are most related to GDP. Oil is the most important resource now, but in order to achieve higher GDP, more energy is needed, so renewable energy and nuclear energy are most common solution by these countries. Coal and hydroelectricity are not so correlated with GDP. Coal is very traditional energy resource, but it can cause big pollution, so now its usage is limited. And hydroelectricity has limitation of water resource, so it is also not commonly used widely. Nature gas is becoming more important recent years, because of resource limitation and new mining technology. So this hypothesis can be accepted.

7. Hypothesis 5: energy per capita has association with happiness

The world happiness report is based on a wide variety of data to extract the happiness ranking result[4]. And GDP is also included. So here we will not add GDP again into this study, but directly study the different types of energy's association with this happiness indicator result.

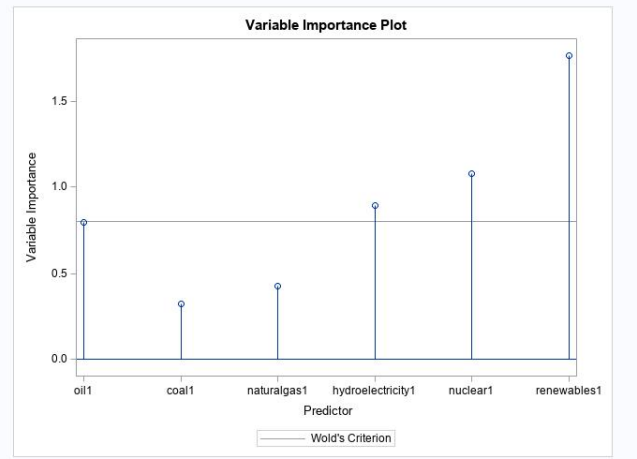
7.1 Partial least squares analysis



It appear that, generally lower factor 1 and factor 2 scores are associated with lower happiness. It is interesting to find some clusters in this graph:

- Finland, Sweden and Norway are as one cluster. And they are near each other with similar energy structure and also have high happiness ranking.
- United States and Canada are near each other, and have similar situation as Finland, Sweden and Norway. Same situation for Switzerland and Netherlands.
- It is also interesting to find that New Zealand is very close to Austrian in this graph. In impression there is big gap of the overall energy consumption between the two countries. But if look into the energy per capita, the two countries are close to each other again.
- Israel, Indian, South Korean, Turkey are relatively separated from other countries. This is because their special situation in the world, which are unique in each of their regions.

So based on the energy per capita and happiness, the countries are located in different areas with clusters. And through this clusters we can find close relationship between energy per capita and happiness.



Through above graph, we can find that renewable energy per capita plays the most important role for happiness.

- Renewable energy includes wind, solar energy. If the country has strong development in this area, this also indicate that the country also cares about the environment and diversified energy supply. And people in this country can also benefit much out of it, not just enough energy supply with relative cheaper price, but also environment protection, not in mind but also in practice and industry. So new and clean energy will contribute to the people happiness.
- Nuclear stands in the second position. It is controversial for this kind of energy, because it has high efficiency on the one hand and also has high risk on the other hand. In general, current years there are not many accidents caused by nuclear power plants. The disaster of Chernobyl nuclear power plant had happened many years ago and is not known well for current generation. So Nuclear energy at least can be an efficient way to solve the energy shortage, and can improve the happiness of people. Of course safety is another potential issue we need to always care about.
- Hydroelectricity is ranking in the third position. It is clean energy and also safe, although there could be some potential ecological risks behind it. But overall in general it has much smaller influence on the environment compared to other energy resources. And the overall on average cost on hydroelectricity is not that high. Although the primary first investment for is high, its cost will be continuously reduced according to the years of production.
- Oil, nature gas and coal are the last three types of energy, which are below the threshold. All of them are traditional energy and nonrenewable energies. That means the amount that can be used in the future will be reduced year after year and meanwhile their cost will also be increasing in long term. And they also cause much pollution during the production and usage. So if a country rely only on or much rely on these kinds of energies, it is hard to ensure that the people in this country will be satisfied about it, which is obvious related to happiness.

According to one report in Economics 2020: “against the background of climate change, many countries have started phasing out fossil-fuelled power generation and phasing in renewable energies. Renewable energy facilities, however, can be expected to have wellbeing effects of their own, such as visual or acoustic impairments. Similar to renewable energies, nuclear power avoids emission of air pollutants and greenhouse gases, but poses problems of nuclear risk.”[13]. We can find the general trend of energy structure globally. At a superficial level seems it is related to

climate change, environment or economy, but in the end this is all related to our "happiness". It is in general we can see the energy structure has close association with people happiness. Actually in our daily life we are facing energy cost every day, from fuel price for our cars to electricity and natural gas cost in our homes. We need to pay these bills everyday. We will not feel happy if the prices of these items increase hugely. And if the environment situation is not good such like air pollution, we will also not feel happy. So healthy and sustainable energy structure will significantly influence our daily life and also our happiness.

7.2 Conclusion

Seems that happiness is far from the energy consumption. But through this analysis, we found the close association between them. And we even found that the renewable energy is most related with happiness. Developing clean and sustainable energy can ensure the energy supply with stable price and also protect our environment, which has high benefits for our society now and also in the future. So this hypotheses can be accepted.

8. Summary

For this study we based on the energy consumption data for different kinds of energy types. Then we also added GDP, population and happiness indicator to analyze their association with different types of energy. We wrapped up 5 hypotheses step by step and use different suitable multivariate analysis techniques for deeper analysis. There are new and interesting findings after the analysis, which can be new for us before this study.

Hypothesis 1: clusters of different energy types. We used principal components analysis and factor analysis. We found that natural gas and nuclear energy are very associated, and coal and hydroelectricity are also closely correlated with each other. Additionally we also primarily find the different energy consumption structures for representative countries such like United States, China etc.

Hypothesis 2: different countries and regions have different energy consumption speciality. We found that some countries have similar energy structure even they are in different regions, and some regions have similar energy structure for most the countries in its regions, so besides existing resource and country developing level, the geography also plays a role behind the energy structure.

Hypothesis 3: energy types and GDP&population have influence on each other. We found that the energy types oil and renewable energy are very associated with GDP. Compared with GDP, the energy association is not that strong associated with population.

Hypothesis 4: energy consumption has high association with GDP. We can see the clear association of different types of energy with GDP. Oil, renewable energy and nuclear energy are most related to GDP. Renewable energy and nuclear energy are also very associated with GDP, as they are are the most popular solution for solving the energy limitation. Coal and hydroelectricity are not so correlated with GDP, because coal can cause big pollution, so now its usage is limited. And hydroelectricity has limitation of water resource. Nature gas is becoming more important recent years, same as its association with GDP.

Hypothesis 5: energy per capita has association with happiness. This is a new trial and we also got interesting conclusions. The renewable energy is most related with happiness. It is clean and sustainable, and it is very good solution to supply lower price and safe resource. It is very widely accepted globally. All these benefits leads to its high contribution to our society and our long term future. Nuclear although has some risk behind it. According to its technology development and strict control, recent years there is no big leakage out of it. So it stands the second role for association with happiness. Hydroelectricity comes after that. So the top three are all new and sustainable energy. So it is interesting finding about their high association with happiness.

We only focused on the year 2021. I think it will be more benefited if we take more years data for analysis to see the dynamic trends. And we may find more interesting findings. And we can further check the details of happiness indicator, to see if we can also add the energy structure factor or at least sustainable energy resource factor into the happiness ranking.

Reference:

[1] *bp Statistical Review of World Energy 2022 | 71st edition*. bp. https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2022-full-report.pdf

[2] *World Bank Open Data*. (n.d.). World Bank Open Data. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

[3] *World Bank Open Data*. (n.d.-b). World Bank Open Data. https://data.worldbank.org/indicator/SP.POP.TOTL?end=

[4] Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J., Aknin, L. B., & Wang, S. (2022, March 18). World Happiness Report 2022. *The World Happiness Re*port. https://worldhappiness.report/ed/2022/

[5] Wikipedia contributors. (2024b, May 16). *Principal component analysis.* Wikipedia. https://en.wikipedia.org/wiki/Principal_component_analysis

[6] Wikipedia contributors. (2024b, May 10). *Factor analysis.* Wikipedia. https://en.wikipedia.org/wiki/Factor_analysis

[7] Hassan, M. (2023, August 15). *Cluster Analysis* - types, methods and examples. Research Method. https://researchmethod.net/cluster-analysis/

[8] *Canonical Discriminant analysis.* (n.d.). https://www.sfu.ca/sasdoc/sashtml/insight/chap40/sect7.htm

[9] Abdi, H., Guillemot, V., Eslami, A., & Beaton, D. (2018). Canonical correlation analysis. In *Springer eBooks* (pp. 177–192). https://doi.org/10.1007/978-1-4939-7131-2_110191

[10] Wikipedia contributors. (2024b, April 19). *Partial least squares regression*. *Wikipedia.* https://en.wikipedia.org/wiki/Partial_least_squares_regression

[11] Al-Mulali, U., & Mohammed, A. H. (2015b). The relationship between energy consumption and GDP in emerging countries. International Journal of Energy Sector Management, 9(1), 77–93. https://doi.org/10.1108/ijesm-04-2013-0006

[12] Thalassinos, E., Kadłubek, M., Thong, L. M., Van Hiep, T., & Ugurlu, E. (2022). Managerial issues regarding the role of natural gas in the transition of energy and the impact of natural gas consumption on the GDP of selected countries. *Resources, 11(5)*, 42. https://doi.org/10.3390/resources11050042

[13] Welsch, H. (2020). Happiness and energy supply. In *Edward Elgar Publishing eBooks.*

https://www.elgaronline.com/edcollchap/edcoll/9781788119337/9781788119337.00025.xml

Complete code:

```
%let path=D:\1. BA\Multivariate Analysis for Big Data\PRACDATA;
libname PRAC"&path";
ods graphics on;
* Input the Excel file;
proc import datafile="&path\world_2021.xls"
    out=PRAC.world_2021
    dbms=xls
    replace;
    getnames=yes;
run;

* Overall data study;
proc contents data=PRAC.world_2021;
run;

proc means data=PRAC.world_2021;
  var oil--Happiness;
run;

ods graphics / imagemap;

proc princomp data=PRAC.world_2021
              n=5
              out=prin
              prefix=pca
              plots=(matrix score(ncomp=3) patternprofile pattern(ncomp=3));
    var Oil--Happiness;
    id country;
run;

* PCA;
ods graphics / imagemap;
proc princomp data=PRAC.world_2021
              n=5
              out=prin
              prefix=pca
              plots=(matrix score(ncomp=3) patternprofile pattern(ncomp=3));
 var oil--Renewables;
     id country;
run;
proc prinqual data=PRAC.world_2021 mdpref;
transform identity(oil--Renewables);
  id country;
run;

title 'Oil by Coal and Natural gas';
proc sgplot data=PRAC.world_2021;
    bubble x=coal y='Natural gas'n size=oil / transparency=0.4 datalabel=country;
    inset "Bubble size represents Oil" / position=bottomright;
run;

* Factor analysis;
ods graphics on;
proc factor data=PRAC.world_2021 plots=(scree loadings) method=ml priors=smc;
  title 'Factor Analysis: Extracting Factors';
 var oil--Renewables;
run;

proc factor data=PRAC.world_2021 plots=loadings method=principal
            priors=smc n=2 rotation=promax flag=.3 fuzz=.2;
```

```
  title 'Promax Rotation';
   var oil--Renewables;
run;

* Cluster analysis;
ods graphics on;
proc cluster data=prac.world_2021 method=ward ccc pseudo outtree=tree print=15
plots=den(height=rsq);
 var oil--Renewables;
id country;
run;

* Canonical Discriminant Analysis;

ods output canonicalmeans=b(rename=(can1=can1c can2=can2c));
proc candisc data=PRAC.world_2021 out=candout;
class region;
 var oil--Renewables;
title 'Canonical Discriminant Analysis Using DSM IV Items';
run;
data plot;
set candout b;
run;
proc sort data=plot;
by region fromregion;
run;
proc sgplot data=plot nocycleattrs;
scatter x=can1 y=can2 / group=region;
scatter x=can1c y=can2c / group=fromregion
markerattrs=(size=20);
run;

proc stepdisc data=PRAC.world_2021 method=stepwise;
class region;
 var oil--Renewables;
run;

proc discrim data=PRAC.world_2021;
  class Region;
  priors prop;
  var oil 'Natural gas'n coal Hydroelectricity Renewables;
run;

* Canonical correlation analysis;
ods output cancorr=a;
proc cancorr data=PRAC.world_2021 out=out_cancorr;
  var oil--Renewables;
  with GDP--Happiness;
run;

ods output cancorr=a;
proc cancorr data=PRAC.world_2021
vprefix=R wprefix=G
vname='R Questions' wname= 'G Questions'
outstat=out;
  var oil--Renewables;
  with GDP--Population;
run;
proc sgplot data=a;
series y=squcancorr x=number /markers;
xaxis integer;
run;

proc cancorr data=PRAC.world_2021 out=world red
vprefix=Energy wprefix=GP
vname='Energy indicators'
wname='GP indicators'
```

```
ncan=2;
  var oil--Renewables;
  with GDP--Population;
run;
proc contents data=world;
run;
proc sgscatter data=world;
plot energy1*gp1 energy2*gp2;
run;

proc sgplot data=a;
series y=squcancorr x=number /markers;
xaxis integer;
run;

* Partial Least Squares (PLS);
proc pls data = PRAC.world_2021 method = pls(algorithm=nipals)
cv=one cvtest(seed=608789001)
plot=(vip xyscores xscores parmprofiles dmod);
model GDP = oil--Renewables;
run;

proc pls data = prac.world_2021 method = pls(algorithm=nipals)
cv=one cvtest(seed=608789001)
plot=(vip xyscores xscores parmprofiles dmod);
model GDP = oil--Renewables happiness population;
run;

* Partial Least Squares (PLS);
data world;
set prac.world_2021;
oil1=oil*1000000/population;
coal1=Coal*1000000/population;
naturalgas1='Natural gas'n*1000000/population;
hydroelectricity1=hydroelectricity*1000000/population;
nuclear1='Nuclear energy'n*1000000/population;
renewables1=renewables*1000000/population;
run;

proc pls data = world method = pls(algorithm=nipals)
cv=one cvtest(seed=608789001)
plot=(vip xyscores xscores parmprofiles dmod);
model happiness = oil1--renewables1;
id Country;
run;
```