Best Predictive Feature for a Country's Happiness

Luke Villanueva, University of California - Los Angeles

Contents

Abstract

This data analysis focuses on finding which social factor has the most effect on the happiness level of specific countries. The predictive variables considered are Social Progress Index (SPI), gross domestic product (GDP), World Giving Index (WGI). Using K-Means clustering to find a representative country for each group, a cubic spline interpolation was used to fill in for reasonably missing data. A multiple linear regression model was fitted to each representative country. A program outputs the data's statistics and graphs.

Introduction

Every year since around the year 2008, the World Happiness Report has been surveying and collecting data on samples of the population of each country. Based on a number of criteria and categories, the total number of each country's "happiness level" is reported annually. These categories can span from general kindness each citizen portrays (i.e. the World Giving Index), their GDP and economic success, to their progress towards a more humane government (i.e. the Social Progress index).

For the purpose of a limited scope for this project, specific predictor variables have been chosen are Social Progress Index (SPI), gross domestic product (GDP), and World Giving Index (WGI).

Data Preparation

Data Collection

The data was collected from the World Happiness Report. There were reports available from 2013-2023. Some issues with this data was that 2014 was not available. In addition, 2023 had many countries excluded from the report. And so, to provide the best time scope for the project, the years from 2015-2022 was selected.

The reports consisted of many variables that were considered to create the happiness score such as economy score, family score, health score, and more. For this project, only the country name, each year, and each relevant score was extracted.

Country	Region	Happiness	Happiness	Standard I	Economy (	Family	Health (Lif	Freedom	Trust (Gov	Generosit	Dystopia
Switzerlan	Western E	1	7.587	0.03411	1.39651	1.34951	0.94143	0.66557	0.41978	0.29678	2.5173
Iceland	Western E	2	7.561	0.04884	1.30232	1.40223	0.94784	0.62877	0.14145	0.4363	2.7020
Denmark	Western E	3	7.527	0.03328	1.32548	1.36058	0.87464	0.64938	0.48357	0.34139	2.4920
Norway	Western E	4	7.522	0.0388	1.459	1.33095	0.88521	0.66973	0.36503	0.34699	2.4653
Canada	North Am	5	7.427	0.03553	1.32629	1.32261	0.90563	0.63297	0.32957	0.45811	2.4517
Finland	Western E	6	7.406	0.0314	1.29025	1.31826	0.88911	0.64169	0.41372	0.23351	2.6195
Netherlan	Western E	7	7.378	0.02799	1.32944	1.28017	0.89284	0.61576	0.31814	0.4761	2.465
Sweden	Western E	8	7.364	0.03157	1.33171	1.28907	0.91087	0.6598	0.43844	0.36262	2.3711
New Zeala	Australia a	9	7.286	0.03371	1.25018	1.31967	0.90837	0.63938	0.42922	0.47501	2.2642
Australia	Australia a	10	7.284	0.04083	1.33358	1.30923	0.93156	0.65124	0.35637	0.43562	2.2664
Israel	Middle Ea	11	7.278	0.0347	1.22857	1.22393	0.91387	0.41319	0.07785	0.33172	3.0885
Costa Rica	Latin Ame	12	7.226	0.04454	0.95578	1.23788	0.86027	0.63376	0.10583	0.25497	3.1772
Austria	Western E	13	7.2	0.03751	1.33723	1.29704	0.89042	0.62433	0.18676	0.33088	2.533
Mexico	Latin Ame	14	7.187	0.04176	1.02054	0.91451	0.81444	0.48181	0.21312	0.14074	3.6021
United Sta	North Am	15	7.119	0.03839	1.39451	1.24711	0.86179	0.54604	0.1589	0.40105	2.5101
Brazil	Latin Ame	16	6.983	0.04076	0.98124	1.23287	0.69702	0.49049	0.17521	0.14574	3.2600
Luxembou	Western E	17	6.946	0.03499	1.56391	1.21963	0.91894	0.61583	0.37798	0.28034	1.9696

Figure 1: A small snippet of the World Happiness Report for 2015. This organization of information is similar for the following years.

The same applied to the GDP, SPI, and WGI reports. All reports provided many statistics that were considered to calculate each country's respective score for each category.

Country N	Country C	Indicator I	Indicator C	1960	1961	1962	1963	1964	1965	1966	1967
Aruba	ABW	Life expec	SP.DYN.LE	64.152	64.537	64.752	65.132	65.294	65.502	66.063	66.439
Africa East	AFE	Life expec	SP.DYN.LE	44.08555	44.3867	44.75218	44.91316	45.47904	45.49834	45.2491	45.92491
Afghanista	AFG	Life expec	SP.DYN.LE	32.535	33.068	33.547	34.016	34.494	34.953	35.453	35.924
Africa We	AFW	Life expec	SP.DYN.LE	37.84515	38.16495	38.7351	39.06372	39.33536	39.61804	39.83783	39.4715
Angola	AGO	Life expec	SP.DYN.LE	38.211	37.267	37.539	37.824	38.131	38.495	38.757	39.092
Albania	ALB	Life expec	SP.DYN.LE	54.439	55.634	56.671	57.844	58.983	60.019	60.998	61.972
Andorra	AND	Life expec	SP.DYN.LE00.IN								
Arab Worl	ARB	Life expec	SP.DYN.LE	44.9729	45.6764	46.12258	46.97247	47.89576	48.23211	48.45707	48.91291
United Ara	ARE	Life expec	SP.DYN.LE	48.811	49.695	50.686	51.584	52.848	53.985	55.185	56.339
Argentina	ARG	Life expec	SP.DYN.LE	63.978	64.36	64.244	64.449	64.363	64.593	64.891	64.992
Armenia	ARM	Life expec	SP.DYN.LE	61.431	61.803	62.125	62.223	62.418	62.754	62.945	63.213
American	ASM	Life expec	SP.DYN.LE00.IN								
Antigua ar	ATG	Life expec	SP.DYN.LE	61.55	62.363	63.192	64.101	65.058	66.026	66.956	67.859
Australia	AUS	Life expec	SP.DYN.LE	70.81707	70.97317	70.94244	70.91171	70.88098	70.85024	70.81951	70.86927
Austria	AUT	Life expec	SP.DYN.LE	68.58561	69.57732	69.30951	69.44366	69.92195	69.7222	70.04585	69.9178
Azerbaijar	AZE	Life expec	SP.DYN.LE	55.837	56.192	56.516	56.898	57.196	57.619	57.672	57.973
Burundi	BDI	Life expec	SP.DYN.LE	43.024	43.252	43.441	43.718	43.974	42.005	43.531	43.341

Figure 2: A small snapshot of the GDP data. The scope of the data continues, stopping before 2023.

rank_scor	country	spicountry	spiyear	status	score_spi	score_bhr	score_fow	score_opp	score_nbr	score_ws
	World	WWW	2022		65.24	75.8	63.62	56.28	82.09	80.74
	World	WWW	2021		64.87	75.35	63.18	56.07	82.01	79.84
	World	WWW	2020		64.55	74.91	62.91	55.83	81.79	78.69
	World	WWW	2019		64.31	74.48	62.46	55.98	81.53	77.99
	World	WWW	2018		63.62	73.98	61.03	55.84	81.23	77.2
	World	WWW	2017		63.29	73.43	60.52	55.92	80.9	76.59
	World	WWW	2016		62.65	72.88	59.26	55.8	80.53	76.17
	World	WWW	2015		62.32	72.35	58.7	55.92	80.09	75.89
	World	WWW	2014		61.35	71.82	56.08	56.17	79.66	75.33
	World	WWW	2013		60.98	71.25	55.44	56.26	79.18	74.71
	World	WWW	2012		60.47	70.92	54.34	56.14	78.59	74.16
	World	WWW	2011		59.84	70.5	53.24	55.79	77.98	73.63
164	Afghanista	AFG	2011	Ranked	32.62	40.73	25.56	31.57	50.35	49.61
164	Afghanista	AFG	2012	Ranked	33.16	41.85	25.9	31.72	52.22	51.34
164	Afghanista	AFG	2013	Ranked	34.35	43.8	27.73	31.51	53.76	54.21
163	Afghanista	AFG	2014	Ranked	35.68	47.19	28.04	31.81	54.76	56.41
163	Afghanista	AFG	2015	Ranked	36.58	48.54	28.47	32.74	55.01	56.95

Figure 3: A small snapshot of the SPI data. There were many more variables used to calculate the SPI for each country for each year until 2023.

Country;Total Ranking;Total Score;Helping a stranger Ranking;Helping a
New Zealand;1;57%;10;63%;11;68%;7;41%
Australia;1;57%;8;64%;8;70%;13;38%
Ireland;3;56%;21;60%;5;72%;17;35%
Canada;3;56%;4;68%;14;64%;17;35%
United States of America;5;55%;7;65%;18;60%;9;39%
Switzerland;5;55%;21;60%;7;71%;20;34%
Netherlands;7;54%;73;46%;2;77%;9;39%
Sri Lanka;8;53%;55;50%;19;58%;2;52%
United Kingdom;8;53%;26;58%;3;73%;29;29%
Austria;10;52%;26;58%;10;69%;27;30%
Sierra Leone;11;50%;2;75%;62;29%;4;45%
Laos;11;50%;41;53%;14;64%;23;32%
Malta;13;48%;95;40%;1;83%;58;21%
Turkmenistan;14;47%;12;62%;107;17%;1;61%
Iceland;14;47%;70;47%;12;67%;43;26%

Figure 4: A small snapshot of the WGI data. The data was delimited via semicolons, as opposed to the usual commas.

Value Changes

For the WGI data, the WGI was given as a percentage. This implied that the value was stored as a string. This data was vectorized to be mutated into a simple integer for better data manipulation and analysis.

Data Cleaning

For the data cleaning, the objective was to create one dataframe with columns of Country, Year, WHI, SPI, GDP, and WGI, alongside with the rows being each individual report of the country on that specific year.

Issues

The following are issues that came up during the creation of the one dataframe containing all of the necessary data for the data analysis. After these issues were resolved, a dataframe was created. There still was missing data, but that will be solved via interpolation/extrapolation.

Missing Countries

As cleaning progressed, an issue that came up was that there were missing countries with each year. And so, the most recent year (i.e. 2022) was chosen to represent the individual unique countries because it was predicted to have at least most of the countries in each of the previous years.

Nomenclature Errors

Another issue among all the data was that each dataframe not only inconsistently names countries across each category of WHI, SPI, GDP, and WGI, but they also inconsistently named them through the years as well. Such examples would be Czech Republic being sometimes called Czechia, Taiwan being called Taiwan Province of China, and more.

This issue was manually fixed by finding the countries that had only a nomenclature inconsistency. However, this did not take into consideration the countries that had multiple parts.

Split Countries

Another issue would be that there were multiple countries that shared the same name but differed in politically. Such examples would be Congo (i.e. Republic of Congo vs the Democratic Republic of Congo), Macedonia (i.e. Macedonia vs North Macedonia), and more.

Some of these countries were considered still but a specific naming addition was added to distinguish between the different countries (i.e. Congo (Brazzaville) and Congo (Kinshana)), and some countries were discarded just because there weren't enough data for one or both of the country's split parts.

Missing Data

Too Much Missing Data

As expected, there were many places where data was missing for all the categories. The criteria used to distinguish which countries were to be kept in would be that they did not have too many missing values. Setting threshold of the max missing values to 6 allowed enough countries to still be used while filtering countries that had too many missing data points.

Too Much Data Missing Sequentially

Another filtering process done in regards to missing data was that there needed to be enough points for the degrees of freedom to be viable for the cubic spline interpolation algorithm to work. And so, certain countries were also omitted because the algorithm was not able to interpolate the data.

Final Country List

The following is the final list of countries used in the data analysis:

- 'Afghanistan', 'Albania', 'Argentina', 'Armenia', 'Australia', 'Austria', 'Azerbaijan', 'Bangladesh', 'Belarus', 'Belgium', 'Benin', 'Bolivia', 'Bosnia and Herzegovina', 'Botswana', 'Brazil', 'Bulgaria', 'Burkina Faso', 'Cambodia', 'Cameroon', 'Canada', 'Chad', 'Chile', 'China', 'Colombia', 'Costa Rica', 'Croatia', 'Czechia', 'Denmark', 'Dominican Republic', 'Ecuador', 'Egypt', 'El Salvador', 'Estonia', 'Ethiopia', 'Finland', 'France', 'Gabon', 'Georgia', 'Germany', 'Ghana', 'Greece', 'Guatemala', 'Guinea', 'Haiti', 'Honduras', 'Hungary', 'Iceland', 'India', 'Indonesia', 'Iran', 'Iraq', 'Ireland', 'Israel', 'Italy', 'Jamaica', 'Japan', 'Jordan', 'Kazakhstan', 'Kenya', 'Kuwait', 'Latvia', 'Lebanon', 'Liberia', 'Lithuania', 'Luxembourg', 'Madagascar', 'Malawi', 'Mali', 'Malta', 'Mauritania', 'Mauritius', 'Mexico', 'Moldova', 'Mongolia', 'Montenegro', 'Morocco', 'Myanmar', 'Namibia', 'Nepal', 'Netherlands', 'New Zealand', 'Nicaragua', 'Niger', 'Nigeria', 'Norway', 'Pakistan', 'Panama', 'Paraguay', 'Peru', 'Philippines', 'Poland', 'Portugal', 'Romania', 'Russia', 'Rwanda', 'Saudi Arabia', 'Senegal', 'Serbia', 'Sierra Leone', 'Singapore', 'Slovenia', 'South Africa', 'Spain', 'Sri Lanka', 'Sweden', 'Switzerland', 'Tajikistan', 'Tanzania', 'Thailand', 'Togo', 'Tunisia', 'Turkmenistan', 'Uganda', 'Ukraine', 'United Arab Emirates', 'United Kingdom', 'United States', 'Uruguay', 'Uzbekistan', 'Venezuela', 'Vietnam', 'Yemen', 'Zambia', 'Zimbabwe'

Data Interpolation/Extrapolation via Cubic Spline Interpolation

To fill in the missing data for the remaining countries after filtering, cubic spline interpolation was used for both data interpolation and extrapolation. This algorithm was placed on the dataframe's variable columns (i.e. WHI, SPI, GDP, and WGI).

Data Analysis

Model Selection

For the purpose of the most easy interpretation, a multiple linear regression model was chosen to model the behavior of the WHI with respect to SPI, GDP, and WGI.

Representative Country Selection

In order to find the countries to model and analyze deeper, a selective process must be made to categorize each country into groups.

This selective process was utilizing the medians of each variable for each country (i.e. creating a dataframe of each country's variables' medians). This would remove the element of time and let only each country have one instance.

This would let the iterative clustering process group the countries more efficiently.

The clustering that provided that most defined groups will be used as the grouping.

K-Means Clustering

A K-Means clustering machine learning algorithm was applied to the datasets for WHI with respect to each predictor variable.

By this plot, WHI with respect to SPI seems to have the most defined clusters. And so, this grouping will be used to label the observations.

The labels were attached to each of their respective observation (i.e. country), and this allows the countries to be grouped.

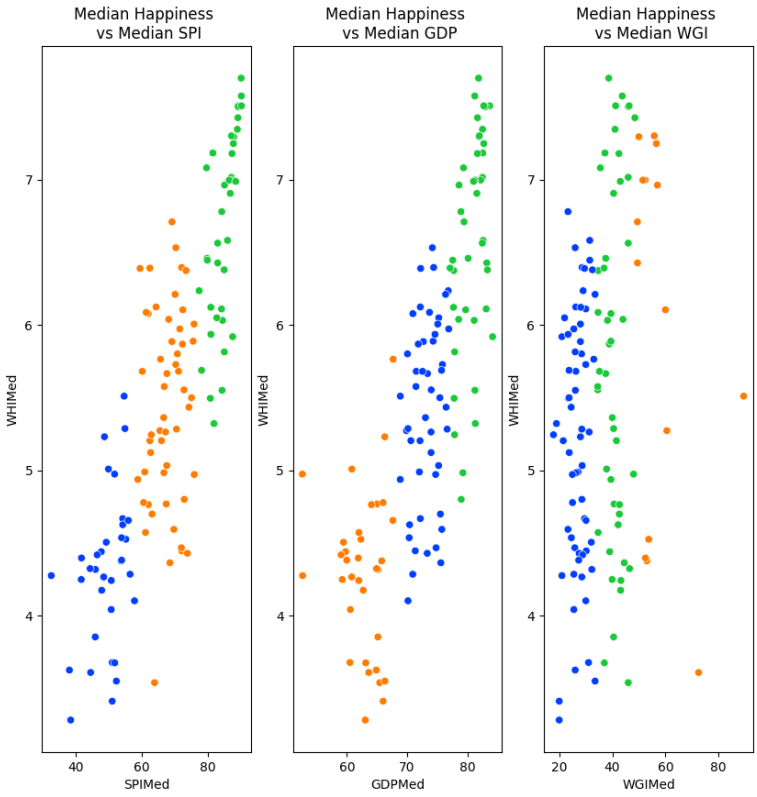


Figure 5: K-Means plot result

Country	WHIMed	SPIMed	GDPMed	WGIMed	Cluster
Switzerlan	7.5103	89.23	83.67805	41.24352	1
Iceland	7.50425	89.325	82.9622	46	1
Denmark	7.5775	90.295	81.19837	43.73316	1
Norway	7.51	90.33	82.68415	46.34974	1
Canada	7.297	87.95	81.88439	50	1
Finland	7.7005	90.22	81.83293	38.66321	1
Netherlan	7.42795	89.245	81.66098	48.5	1
Sweden	7.34825	89.065	82.48415	41	1
New Zeala	7.3033	87.28	81.95732	55.79275	1
Australia	7.25	87.87	82.69658	56.60622	1
Israel	7.18505	81.615	82.52596	37.20725	1
Costa Rica	7.083	79.7	79.327	35.5	1
Austria	7.1815	87.445	81.64268	42.5	1
Mexico	6.533	70.355	74.17	26	0
United Sta	6.96415	85.175	78.58902	57	1
Brazil	6.3973	72.105	74.387	28.61399	0
Luxembou	7.018	87.255	82.46707	46	1

Figure 6: A snapshot of the final dataframe holding the median values of each country and their respective cluster group label from the K-Means algorithm

K-Means Analysis

After grouping, the country with a median WHI value closest to the value of the K-Means algorithm's centroids will be used as the representative countries for each group.

The median dataframe was vectorized by a distance formula to see which observation had the lowest distance to the centroid.

France, Montenegro, and Ethiopia were the closest countries to their respective centroids.

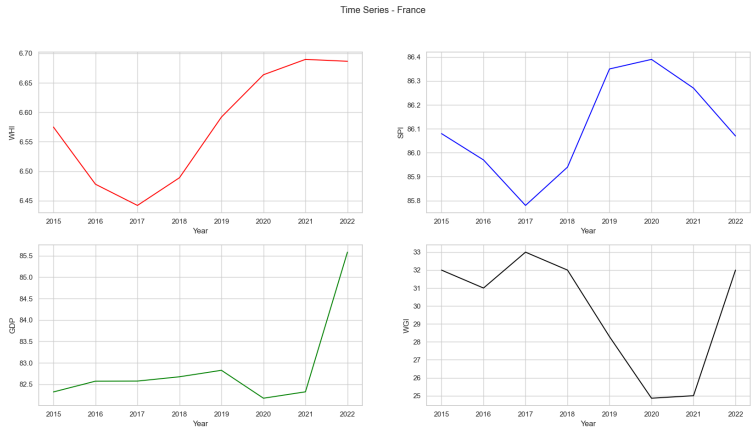


Figure 7: France's plots on WHI with respect to each predictor variable

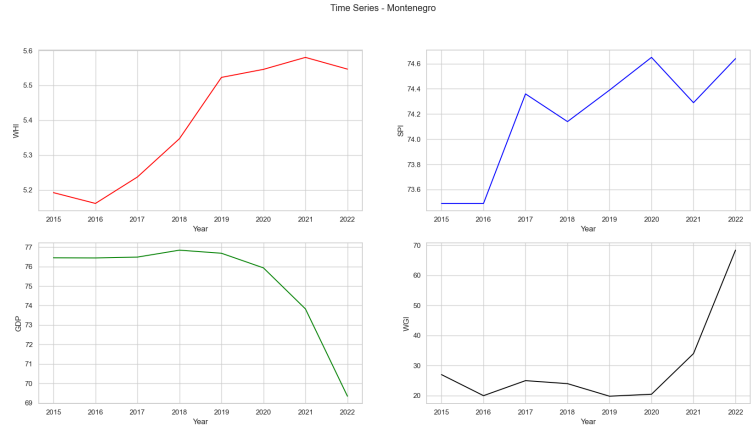


Figure 8: Montenegro's plots on WHI with respect to each predictor variable

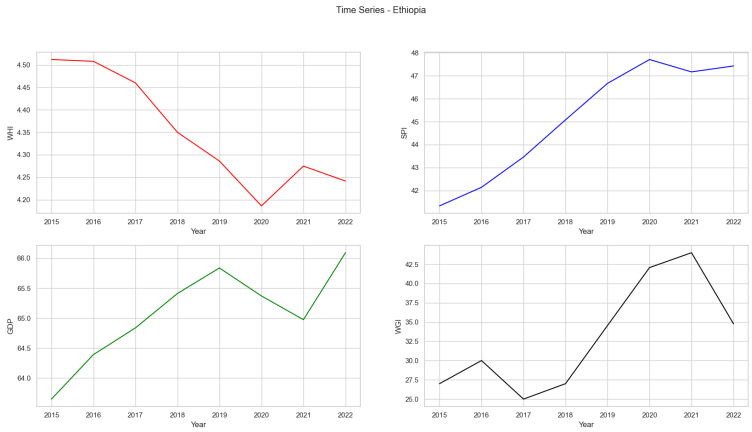


Figure 9: Ethiopia's plots on WHI with respect to each predictor variables

This makes them the representative countries for their group. These countries will now be statistically analyzed.

Statistical Analysis and Research Conclusion

After finding the corresponding representatives of each group of countries, statistically analyzing them will provide a very general idea of how we can expect each group to behave.

As a word of precaution, the following models have gone through power transformation algorithms to fix the diagnosis plots' results, but there was little change to any of the models.

France

Linear Model Diagnosis

The residual vs fitted plot is generally scattered, which implies a good sense of linearity. However, the trend is slightly downward, which may imply inconsistent variance.

The normal plot has most of the points following the normal line, so normality can be generally assumed.

The scale-location plot showcases the plots to be generally randomly scattered, but the trend is seeming to be somewhat positive. This implies slight inconsistent variance.

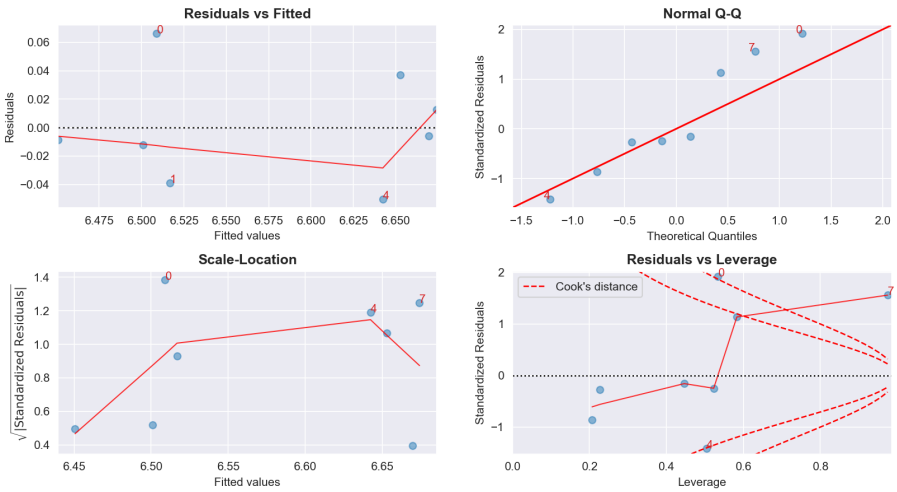


Figure 10: France's Diagnostic plots

There are 4 points that have high leverage and influence via Cook's distance, which negatively heavily affects the linear model.

The VIF table shows SPI and WGI to be highly collinear.

In conclusion, the model's faults include: highly collinear variables, a handful of bad leverage and influence points, and possible heteroscedasticity.

Statistical Interpretation

The following analysis is assuming that the model is valid.

Under a 10% significance level, the intercept and SPI are not statistically significant. The GDP and WGI are statistically significant.

This can be interpreted that as GDP increases in France, the higher WHI gets. On the other hand, as WGI increases, WHI decreases.

Montenegro

Linear Model Diagnosis

The residual vs fitted values plot shows a general scatter, which implies that linearity can be assumed.

The QQ plot shows the plots generally following the line, which implies that normality can be assumed.

OLS Regression Results

Dep. Variable:	WHI	R-squared:	0.851
Model:	OLS	Adj. R-squared:	0.739
Method:	Least Squares	F-statistic:	7.592
Date:	Sun, 18 Jun 2023	Prob (F-statistic):	0.0397
Time:	00:47:07	Log-Likelihood:	15.305
No. Observations:	8	AIC:	-22.61
Df Residuals:	4	BIC:	-22.29
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-13.1993	16.277	-0.811	0.463	-58.392	31.994
SPI	0.1857	0.192	0.969	0.387	-0.346	0.718
GDP	0.0513	0.020	2.575	0.062	-0.004	0.107
WGI	-0.0155	0.013	-1.172	0.306	-0.052	0.021

Omnibus:	0.512	Durbin-Watson:	1.749
Prob(Omnibus):	0.774	Jarque-Bera (JB):	0.403
Skew:	0.421	Prob(JB):	0.817
Kurtosis:	2.292	Cond. No.	1.12e+05

Figure 11: France's linear model's summary stats

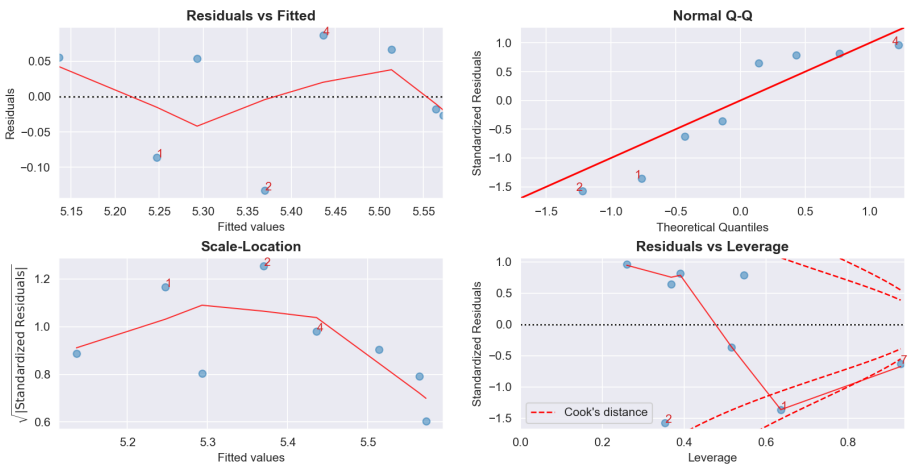


Figure 12: Montenegro's Diagnostic plots

The scale-location plot shows the plots generally scattered, but there is a slight downward trend, which implies there might be inconsistent variance.

There are 2 points that have high bad leverage and high influence via Cook's distance. These points are negatively affecting the model.

The VIF table shows that WGI and GDP are highly collinear.

In conclusion, the model's faults include: highly collinear variables, a handful of bad leverage and influence points, and possible heteroscedasticity.

Statistical Interpretation

OLS Regression Results

Dep. Variable:	WHI	R-squared:	0.799
Model:	OLS	Adj. R-squared:	0.649
Method:	Least Squares	F-statistic:	5.308
Date:	Sun, 18 Jun 2023	Prob (F-statistic):	0.0703
Time:	00:48:54	Log-Likelihood:	9.4485
No. Observations:	8	AIC:	-10.90
Df Residuals:	4	BIC:	-10.58
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.2053	10.966	-0.292	0.785	-33.651	27.241
SPI	0.2372	0.101	2.341	0.079	-0.044	0.519
GDP	-0.1134	0.070	-1.609	0.183	-0.309	0.082
WGI	-0.0157	0.011	-1.464	0.217	-0.046	0.014

Omnibus:	1.085	Durbin-Watson:	1.971
Prob(Omnibus):	0.581	Jarque-Bera (JB):	0.774
Skew:	-0.529	Prob(JB):	0.679
Kurtosis:	1.904	Cond. No.	3.24e+04

Figure 13: Montenegro's linear model's summary stats

Under a 10% significance level, the intercept and SPI are not statistically significant. The GDP and WGI are statistically significant.

GDP and WGI have negative coefficients.

This implies as GDP increases, WHI decreases. Also, this implies as WGI increases, WHI decreases.

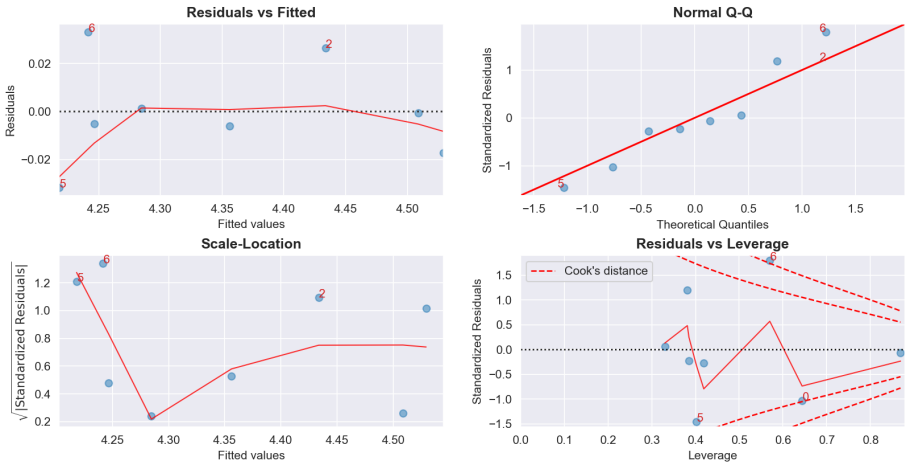


Figure 14: Ethiopia's diagnostic plots

Ethiopia

Linear Model Diagnosis

The residuals vs fitted values plot shows a general scatter, which implies that linearity can be assumed.

The QQ plot shows that plots following the line well.

The scale-location plot is generally scattered, which implies that there is homoscedasticity.

There is only 1 point that has high bad leverage and high influence, hurting the model.

The VIF table shows that GDP and SPI are highly collinear.

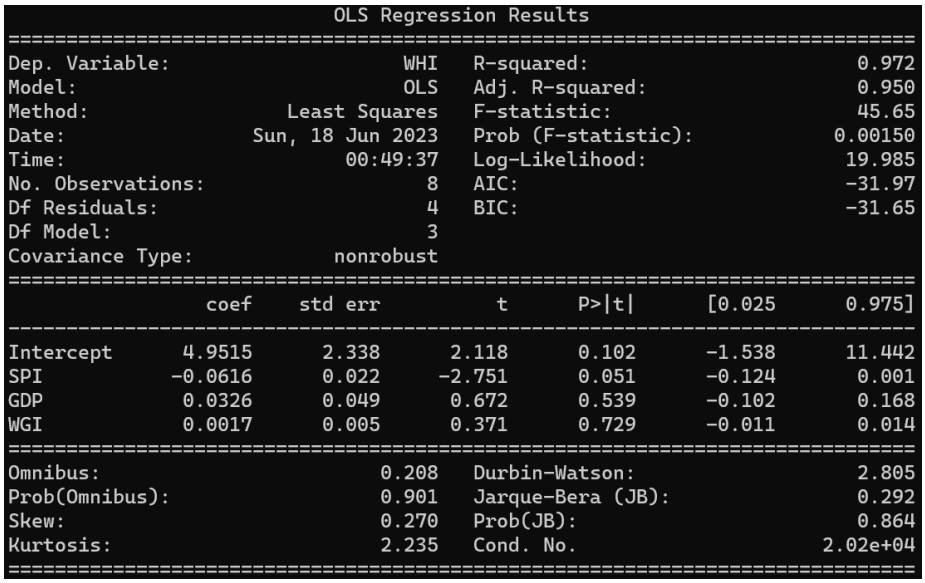
In conclusion, the model's faults is highly collinear variables.

Statistical Interpretation

Under a 10% significance level, the intercept is barely significant, but close enough to consider is somewhat significant. GDP and WGI are not significant. SPI is significant.

SPI has a negative coefficient.

This implies that as SPI increases, the WHI decreases.



OLS Regression Results

Dep. Variable:	WHI	R-squared:	0.972
Model:	OLS	Adj. R-squared:	0.950
Method:	Least Squares	F-statistic:	45.65
Date:	Sun, 18 Jun 2023	Prob (F-statistic):	0.00150
Time:	00:49:37	Log-Likelihood:	19.985
No. Observations:	8	AIC:	-31.97
Df Residuals:	4	BIC:	-31.65
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.9515	2.338	2.118	0.102	-1.538	11.442
SPI	-0.0616	0.022	-2.751	0.051	-0.124	0.001
GDP	0.0326	0.049	0.672	0.539	-0.102	0.168
WGI	0.0017	0.005	0.371	0.729	-0.011	0.014

Omnibus:	0.208	Durbin-Watson:	2.805
Prob(Omnibus):	0.901	Jarque-Bera (JB):	0.292
Skew:	0.270	Prob(JB):	0.864
Kurtosis:	2.235	Cond. No.	2.02e+04

Figure 15: Ethiopia's linear model's summary stats

Analysis Conclusion

In layman's terms and generally applying this analysis to the rest of the clusters, for countries in group 1, there is evidence that as the economic success of the country increases, happiness of the country also increases. However, the more kind and giving a group 1 country is, the less happy the country is.

In regards to group 2, the higher the economic success and the more kind a group 2 country is, the less happy the country is.

In regards to group 3, the more socially progressed the group 3 country is, the less happy the country is.

Limitations and Possible Further Research

As this project progressed, multiple tangents came up that can be resolved or further investigated in the future for this research.

Limitations

K-Means Grouping Selection Method Has Little Mathematical Support

Using the medians of each country is not necessarily the best way to grab the best representative value for the country's statistics. Another method would have provided a more applicable grading scale for the clustering method to use.

In addition, just because the clustering method grouped the country's together, there is no mathematical process that can support general assumption that a "representative" country can vaguely describe the behaviors of other countries in the same clustered group.

Interpolation/Extrapolation Strategies

Cubic spline interpolation is not the best interpolation method for a highly volatile index such as WHI. Another interpolation method that could have been used would be a seasonal adjusted interpolating algorithm, that takes into account the error in the previous history and applies it to the interpolating equation. In addition, a more robust extrapolation algorithm could be used besides extending an interpolating algorithm past its scope.

Too Few Years as Scope

For a more in-depth analysis on the statistics of the countries, there should have been more years provided for the WHI. However, the earliest year provided for WHI was 2013, and even then, 2013's data was not available to the public to access and manipulate. And so, too few years were available to provide a large enough sample to diagnose the model or be strongly confident in the statistical analysis.

Too Few Countries Post-Filter

The cross-referencing work done between the different datasets showed that not every organization survey the same countries nor do they make sure that the data is correctly inputted. This let too few countries make it past the filtering process.

Further Research

Analyzing Each Country

A much larger scale project could be to statistically analyze each country rather than clustering them and obtaining a representative of the cluster. This would allow a general census on each country rather than arbitrarily applying a conclusion on one country based on another's analysis.

Further External Research

Analyzing Why Certain Countries Had Missing Data

There could be a project that utilizes the fact there were many missing data across the multiple datasets used for this project. There must be a correlating factor between the capabilities for an organization to obtain information from that country and the actual status of that country.

Program Demonstration

When running `main.py`, the program greets the user with a welcome message. It then requests an input to see which group the user wants to see the statistical results of.

```
Hello! Welcome to the Statistics Summary applet on the behavior of World Happiness Index via Social Progress, Gross Dome
stic Product, and World Giving Index.
Please select which Group you would like to analyze: 0, 1, or 2:1

            Group 1 is observed to be
            the group with an average median happiness
            score with an average SPI. The groups can be seen
            via the following K-Means plot:

Press Enter to continue...

            The representative of this group is Montenegro.



            The following are the basic graphs of
            the country's variables:

Press Enter to continue...

            The following is the summary of the stats of
            the multiple linear regression model (where World
            Happiness Index is the response variable, and
            Social Progress Index, Gross Domestic Product,
            and World Giving Index are the predictor
            variables):
```

```
Press Enter to continue...

              The following are the diagnostic plots and
              VIF table of the model:

    Features  VIF Factor
1        SPI        1.37
3        WGI       19.39
2        GDP       20.89
0  Intercept    87191.88
Program ending...
```

With every line that it asks the user to press enter to continue, the previously mentioned plots are printed out (e.g. K-means plot, basic time series plots, linear model summary stats, and diagnostic plots)

References

World Happiness Report, https://worldhappiness.report

GDP Report Per Country Per Year, https://datahub.io/core/gdp

Social Progress Index, https://www.socialprogress.org

World Giving Index, https://www.kaggle.com/datasets/vislupus/caf-world-giving-index