# Do Countries Have "Synthetic" Traits?

Bill Alive (@bill_alive)

# Abstract

Using personal survey data with hundreds of personality features, psychologists have discovered the “Big Five” synthetic personality traits (conscientiousness, extroversion, agreeableness, neuroticism, and openness to experience) that can give a useful view of a human personality in only five dimensions.

Can world indicator data yield an analogous set of “synthetic” country traits?

In this project, Principal Component Analysis (PCA) was used on the World Bank World Development Indicators dataset for 2018 to extract five synthetic country traits. Each of these traits reveals an intriguing set of indicators with which it is most and least correlated. The highest and lowest scoring countries for each trait are also presented.

# Motivation

As we try to make sense of the world, the data can be overwhelming. We want to know, and quantify, which indicators matter most, but there are so many to consider.

Custom indicators, like the Economic Freedom Ranking or the Human Development Index (HDI), can seem like the solution. Here, at last, is a single number that can tell us something important about the differences between countries.

The problem? By definition, they focus only on *some* of the data. They may be useful, but only if you agree with the presuppositions of the researchers. How they define "economic freedom" or "human development" determines the value of their findings.

But what if we want to set presuppositions aside? What if we want to know which traits *most* distinguish countries from each other, *whatever* these traits might be?

# Dataset

The World Development Indicators dataset is the “World Bank’s premier compilation of cross-country comparable data on development.”

Note: instead of the older version of the dataset used for the Week 6 Mini-Project, I acquired the most recent data available directly from the World Bank site.

The full dataset includes data from 1960 to 2021. Each row consists of a country (or country grouping, like “High income”), an indicator, and then, for each year, a separate column holding that indicator’s value for that country in that year.

For this project, I focused on a single year of data. The most recent years do not have quite as much data, so I chose to use 2018, for which there are 171,102 non-null data points distributed across 1,286 separate indicators.

# Data Preparation and Cleaning

- Removed the data for all years except 2018.
- Removed rows for country groups. This project focuses only on individual countries. Indicators like “Gross National Product” for groups like “High income” would have introduced (even more) extreme variance into the data.
- Scaled data to a normal distribution using sklearn StandardScaler().
- Replaced missing values with zero. This may not be the optimal solution, but zero was the mean value after scaling. In future analysis, a more nuanced approach might involve removing missing values from the calculations entirely.

# Research Questions

Can we use Principal Component Analysis to extract “synthetic” country traits from world indicator data?

Will these traits correspond to recognizable “features” of countries, the way the Big Five personality traits like extroversion make intuitive sense?

If not, will they provide any other interesting or useful insights? Can they show us any surprising correlations between indicators that we might not easily see otherwise?

# Methods

I used Principal Component Analysis to extract the top five “synthetic traits” from the world indicator data. I carefully followed the same steps outlined in video 9.4 from the UCSD DSE220x Machine Learning Fundamentals course, “Case Study: Personality Assessment”. I also relied on the Week 9 notebook for that course, which performs PCA on the MNIST dataset of handwritten digits.

On personality survey data, PCA can yield not only the synthetic traits themselves, but also:

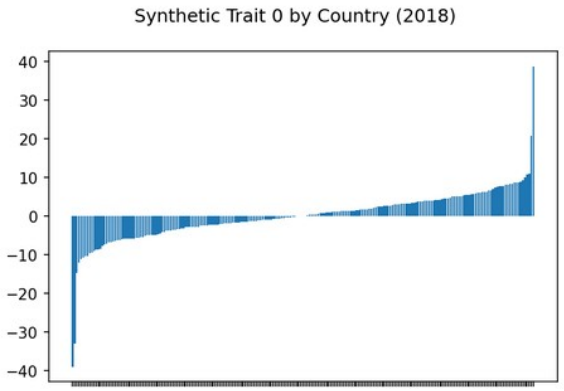
- the personality features most and least correlated with these traits
- each person’s individual score for each trait.

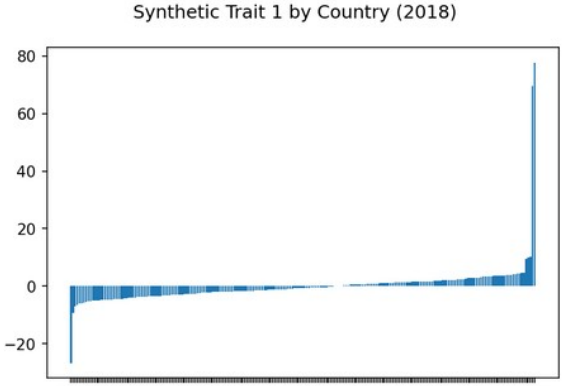
This project shows how using PCA on world indicator data can yield similar results for countries.

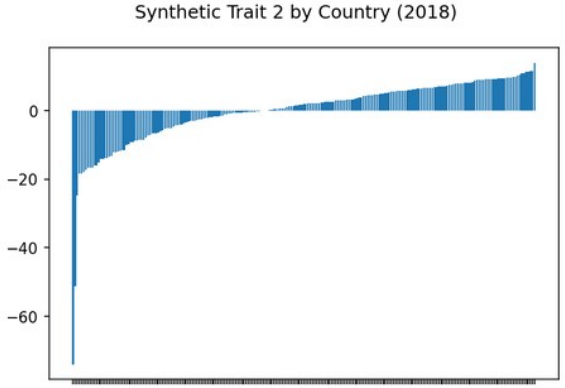
# Findings: Synthetic Traits by Country

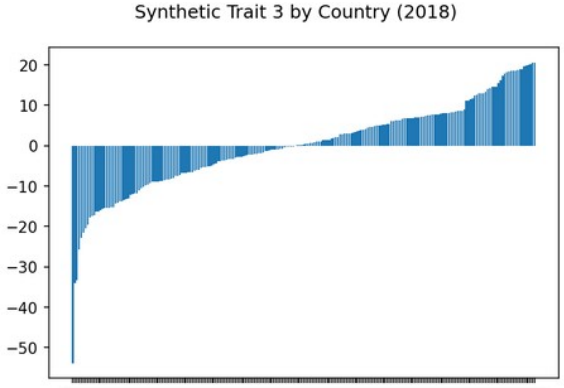
First, let's see how each of our five synthetic traits is distributed across the 217 countries. Trait 0 is based on the first principal component, trait 1 on the second, etc.

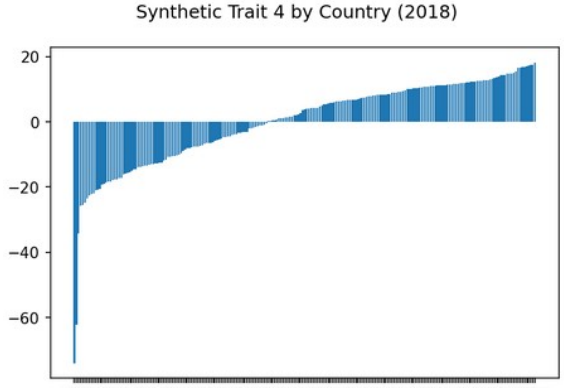
As these charts show, the positive and negative outliers seem extreme; this may or may not be a problem.



# Findings: Synthetic Trait 0

## *+Debt, +Net secondary income, +Transport -ODA provided, -Net primary income*

**Top 10 Indicators: Most Correlated With Trait 0**
(751) Net secondary income (Net current transfers from abroad) (current US$)
(742) Net secondary income (BoP, current US$)
(710) Multilateral debt service (TDS, current US$)
(700) External debt stocks, total (DOD, current US$)
(686) Secondary education, pupils (% female)
(668) Debt service on external debt, total (TDS, current US$)
(665) External debt stocks, private nonguaranteed (PNG) (DOD, current US$)
(650) Public private partnerships investment in transport (current US$)
(650) Investment in transport with private participation (current US$)
(650) External debt stocks, long-term (DOD, current US$)

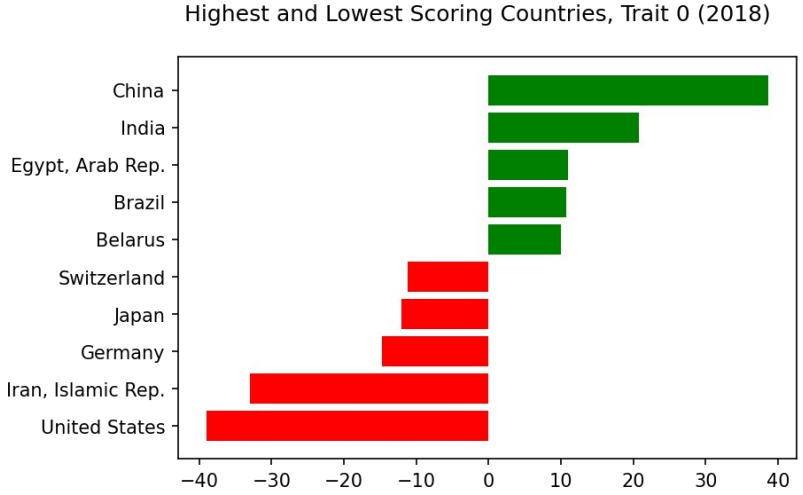
**Bottom 10 Indicators: Least Correlated With Trait 0**
(-635) Imports of goods and services (constant 2015 US$)
(-642) Secure Internet servers
(-650) General government final consumption expenditure (constant 2015 US$)
(-665) Net primary income (Net income from abroad) (current US$)
(-676) Net ODA provided, total (current US$)
(-680) Net ODA provided, total (constant 2020 US$)
(-681) Net primary income (BoP, current US$)
(-711) Charges for the use of intellectual property, receipts (BoP, current US$)
(-718) Net ODA provided, to the least developed countries (current US$)
(-733) Net errors and omissions (BoP, current US$)

*What's the most striking trait that sets countries apart? Apparently, it's **debt.***

*Well, not quite. The indicators most correlated with this trait are **net secondary income.** This seems to balance with two of the least correlated indicators being **net primary income.***

*But 5 of the 10 topmost indicators here deal with debt. Plus, the countries scoring most highly on this trait are also least likely to provide **ODA** (Official Development Assistance) elsewhere.*

*The high investment in **transport** is intriguing, but the low correlation with **imports of goods and services** may be misleading; other related indicators dealing with imports are more highly correlated (see the notebook for the full data).*



# Findings: Synthetic Trait 1

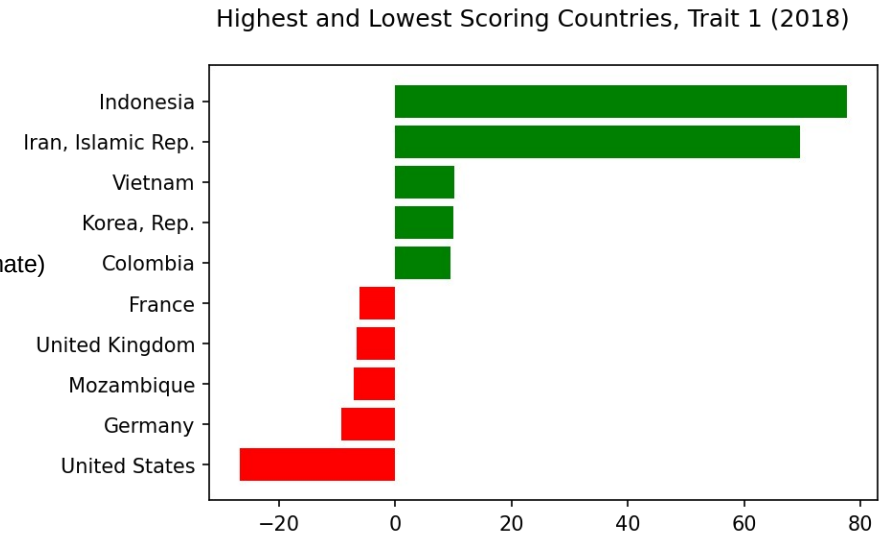
## *+GFCF, +Manufacturing, +Agriculture, -ODA provided, -Net primary income*

**Top 10 Indicators: Most Correlated With Trait 1**
(1552) Gross fixed capital formation (current LCU)
(1549) Gross fixed capital formation (constant LCU)
(1548) Manufacturing, value added (constant LCU)
(1546) Industry (including construction), value added (constant LCU)
(1544) Manufacturing, value added (current LCU)
(1542) Imports of goods and services (constant LCU)
(1541) Gross capital formation (constant LCU)
(1541) Taxes less subsidies on products (current LCU)
(1540) Agriculture, forestry, and fishing, value added (constant LCU)
(1535) Households and NPISHs Final consumption expenditure (current LCU)

**Bottom 10 Indicators: Least Correlated With Trait 1**
(-435) Age dependency ratio (% of working-age population)
(-438) Net ODA provided, total (current US$)
(-441) Net ODA provided, total (constant 2020 US$)
(-449) Ratio of female to male labor force participation rate (%) (modeled ILO estimate)
(-456) Net ODA provided, to the least developed countries (current US$)
(-460) Charges for the use of intellectual property, receipts (BoP, current US$)
(-461) Primary income receipts (BoP, current US$)
(-791) Net primary income (Net income from abroad) (current LCU)
(-1095) Net lending (+) / net borrowing (-) (current LCU)
(-1326) Terms of trade adjustment (constant LCU)

*For our next highest trait, the leading indicators are **gross fixed capital formation** (also called "investment", see here) as well as **manufacturing, industry,** and **agriculture.** Once again, we see **net ODA provided** and **net primary income** in the least correlated indicators.*

*So far, this seems to make sense. But then, what about the other leading indicators: imports, taxes, and household consumption? Should these correlate highly with manufacturing and agriculture? Also, why do Indonesia and Iran have such high scores? Iran had an extremely **low** score on trait 0, so there may just be a problem with this country's data. But trait 1 may need further analysis.*



# Findings: Synthetic Trait 2

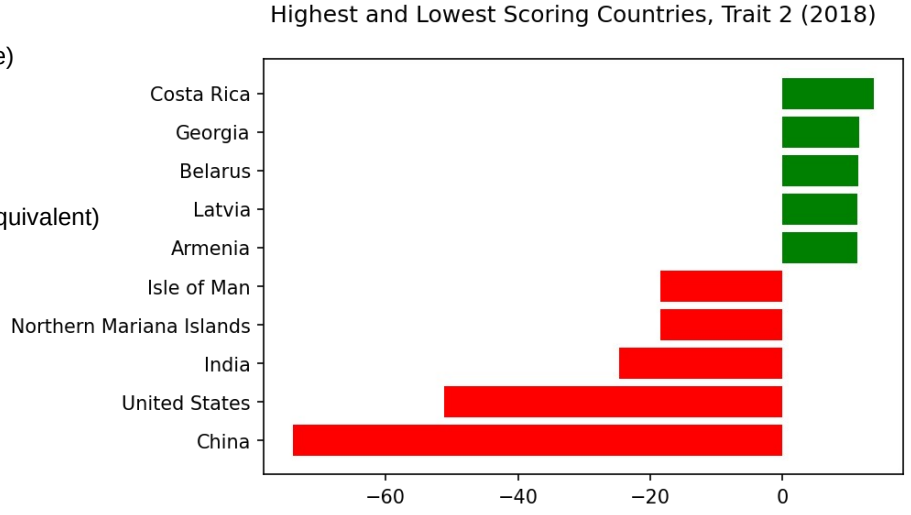
## *+SPI, +Older population, +Labor % -Emissions, -Capital Formation*

**Top 10 Indicators: Most Correlated With Trait 2**
(1158) Statistical performance indicators (SPI): Pillar 3 data products score (scale 0-100)
(1151) Statistical performance indicators (SPI): Overall score (scale 0-100)
(1131) Statistical performance indicators (SPI): Pillar 2 data services score (scale 0-100)
(1085) Survival to age 65, female (% of cohort)
(1080) Labor force participation rate, total (% of total population ages 15+) (national est.)
(1075) Population ages 15-64, male (% of male population)
(1069) Population ages 15-64 (% of total population)
(1067) Population, female (% of total population)
(1062) Ratio of female to male labor force participation rate (%) (national estimate)
(1062) Women Business and the Law Index Score (scale 1-100)

**Bottom 10 Indicators: Least Correlated With Trait 2**
(-1304) Industry (including construction), value added (constant 2015 US$)
(-1304) Gross domestic savings (current US$)
(-1313) Nitrous oxide emissions in energy sector (thousand metric tons of CO2 equivalent)
(-1314) Gross savings (current US$)
(-1323) Total greenhouse gas emissions (kt of CO2 equivalent)
(-1324) Gross capital formation (current US$)
(-1324) Gross fixed capital formation (current US$)
(-1330) CO2 emissions (kt)
(-1331) Adjusted savings: carbon dioxide damage (current US$)
(-1334) Gross capital formation (constant 2015 US$)

*For Trait 2, the strongest indicators are, perhaps ironically, synthetic **performance indicators** **created** by the World Bank. But we also see intriguing correlations with **adult** and **female population percentages** and **labor force participation** rates.*

*Meanwhile, **emissions** dominate the least correlated indicators, along with **gross capital formation** (the opposite of trait 1). Given the emissions issue, it's no surprise to see the US and China with the lowest scores on this trait… but why are the Isle of Man and the Northern Mariana Islands down here as well? Missing data, perhaps? And do China and India also have low correlations with gross domestic **savings**… and **industry**? This trait requires further investigation.*



# Findings: Synthetic Trait 3

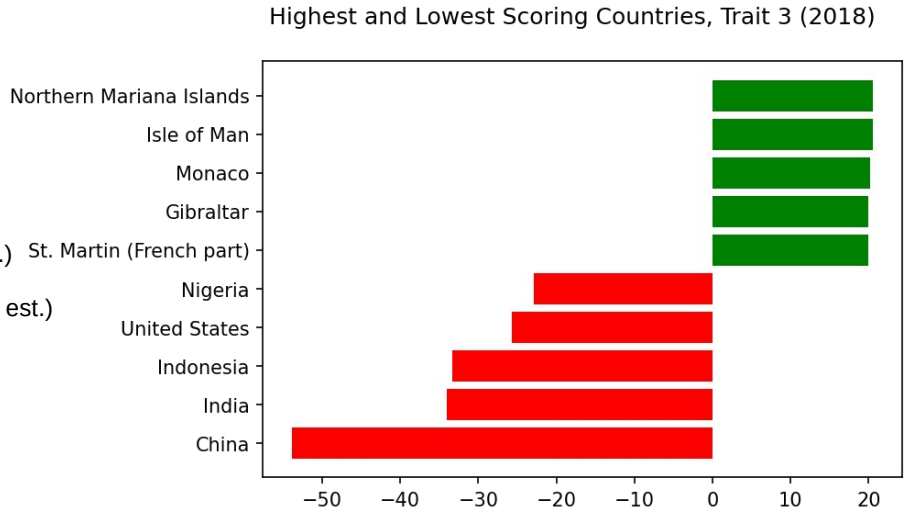
## *+GDP, +Electricity/Comms/Water, +Urban -Children, -Vulnerable/Self-Employed*

**Top 10 Indicators: Most Correlated With Trait 3**
(1188) GDP per capita (current US$)
(1097) GDP per capita (constant 2015 US$)
(1071) Access to electricity (% of population)
(1015) Access to electricity, rural (% of rural population)
(988) Urban population (% of total population)
(941) Fixed broadband subscriptions (per 100 people)
(899) Fixed telephone subscriptions (per 100 people)
(878) People using safely managed drinking water services (% of population)
(869) GNI per capita, Atlas method (current US$)
(817) GDP per capita, PPP (current international $)

**Bottom 10 Indicators: Least Correlated With Trait 3**
(-1753) Population ages 10-14, male (% of male population)
(-1755) Population ages 05-09, male (% of male population)
(-1762) Population ages 0-14 (% of total population)
(-1764) Vulnerable employment, male (% of male employment) (modeled ILO est.)
(-1765) Population ages 0-14, male (% of male population)
(-1775) Vulnerable employment, female (% of female employment) (modeled ILO est.)
(-1778) Self-employed, female (% of female employment) (modeled ILO est.)
(-1782) Vulnerable employment, total (% of total employment) (modeled ILO est.)
(-1782) Self-employed, male (% of male employment) (modeled ILO est.)
(-1798) Self-employed, total (% of total employment) (modeled ILO est.)

*At first, Trait 3 seems straightforward: **wealth**. The highest correlating indicator couldn't be simpler: **GDP per capita.** Correlated with wealth, we see access to **electricity**, subscriptions to **broadband** and **telephone,** even **safely managed drinking water,** plus a higher percentage **urban population**. Meanwhile, these rich countries have a lower **percentage of children** and lower rates of **self-employment.** All very neat… until you check the country scores.*

*They don't seem to correlate with GDP per capita. Yes, Monaco and Isle of Man are both wealthy, and they score highly on this trait. But how can the United States possibly have one of the **lowest** scores, with a 2018 GDP per capita of $62,805? And Gibraltar, St. Martin, and the Northern Mariana Islands, it turns out, are **missing** GDP per capita data for 2018. So while this trait is promising, something is amiss. Filling missing values with 0's may have skewed these scores.*



# Findings: Synthetic Trait 4

## *+Child Mortality, +CPIA ratings, -Imports/Exports, -Health Expenditure*

**Top 10 Indicators: Most Correlated With Trait 4**

(1650) Mortality rate, infant, male (per 1,000 live births)
(1642) Mortality rate, infant (per 1,000 live births)
(1630) Mortality rate, under-5, male (per 1,000 live births)
(1630) Mortality rate, infant, female (per 1,000 live births)
(1628) Mortality rate, neonatal (per 1,000 live births)
(1618) Mortality rate, under-5 (per 1,000 live births)
(1604) Mortality rate, under-5, female (per 1,000 live births)
(1556) CPIA efficiency of revenue mobilization rating (1=low to 6=high)
(1554) CPIA trade rating (1=low to 6=high)
(1550) CPIA equity of public resource use rating (1=low to 6=high)

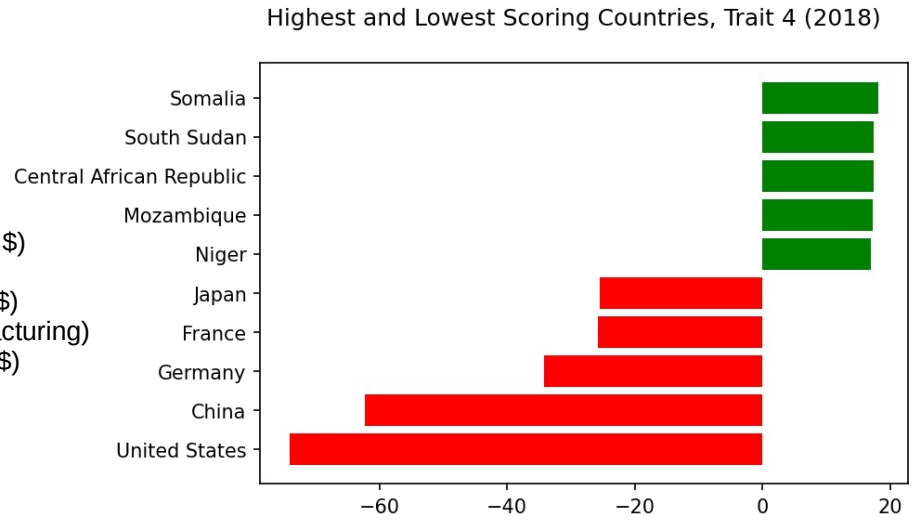
**Bottom 10 Indicators: Least Correlated With Trait 4**

(-2155) Imports of goods and services (BoP, current US$)
(-2160) Current health expenditure per capita, PPP (current international $)
(-2163) Imports of goods and services (current US$)
(-2164) Imports of goods, services and primary income (BoP, current US$)
(-2164) Machinery and transport equipment (% of value added in manufacturing)
(-2175) Exports of goods, services and primary income (BoP, current US$)
(-2176) Commercial service imports (current US$)
(-2176) Service imports (BoP, current US$)
(-2181) Exports of goods and services (BoP, current US$)
(-2191) Exports of goods and services (current US$)

*Trait 4 is heartbreaking. I never expected **infant and child mortality** to dominate one of these synthetic traits.*

*On the flip side, we see low correlations with **health expenditure,** as well as both **imports** and **exports**. This trait seems grim; you do not want a high score.*

*And yet, why does it also correlate with high scores on **CPIA ratings**? The World Bank states that its "Country Policy and Institutional Assessment is done annually for all its borrowing countries." Presumably, high scores are desirable, so why would they correlate with high child mortality? Non-borrowing countries do not receive **any** CPIA ratings, so perhaps that explains the correlation, but it bears further study.*



# Limitations

This project was inspired by PCA performed on personality surveys that offered simple, consistent data. By contrast, the features in this data vary wildly by many orders of magnitude, from gross national income to percentages to CPIA ratings from 1 to 6. Despite a simple attempt to scale the data to a normal distribution, it may well be that far more nuanced work is required to make this data usable for PCA.

Similarly, the simple expedient of replacing missing values with zeroes may have been naive. Not only are different countries missing *different* values, but some countries are simply far more represented in the data. This higher rate of data collection may in itself be an indicator of wealth and infrastructure, which could mean that assigning zeros to countries without this benefit, when zero suggests a “mean” for a given indicator, is highly misleading. In trait 4, for instance, having *any* value for the CPIA ratings, which apply only to borrower countries, seems to be more significant than the rating itself.

Also, even this vast dataset is inherently limited. Ideally, it should be easy to scale and add new indicators from other datasets in the future, refining these synthetic traits in interesting ways.

Finally, unlike most human personalities, country indicators can change significantly over time. To be truly valuable, these synthetic traits need to be easy to understand and view over time.

# Conclusions

*Can we use Principal Component Analysis to extract "synthetic" country traits from world indicator data?* **Yes.** Despite all the limitations of this project, the traits we extracted do seem to signal some kind of meaning; they do not appear totally random.

*Will these traits correspond to recognizable "features" of countries, the way the Big Five personality traits like extraversion make intuitive sense?* **Uncertain.** Though some traits seem straightforward, this is sometimes belied by country scores that don't seem right, as if the first PCA personality study had rated Richard Nixon as *low* on "neuroticism". Other traits, while suggestive, do not at first glance seem to have a main "idea", although a fuller examination of the correlating indicators might lead to one.

*If not, will these traits provide any other interesting or useful insights? Can they show us any surprising correlations between indicators that we might not easily see otherwise?* **Yes.** Although no obvious new "features" present themselves, perhaps that actually makes them more enticing. These synthetic traits with their strange correlations seem like **paths to explore**: high ridges among the mountains of data that may yield valuable, unexpected new vistas for possible positive change.

# Acknowledgements

Sadly, I did not have any opportunities for feedback.

My data was sourced from the World Bank, and as stated above, this PCA approach was based entirely on the discussion and notebook in Week 9 of the Machine Learning Fundamentals Course, DSE220x. The idea to apply this analysis to world indicators was my own, as was the preliminary analysis of these synthetic traits that is presented here.

# References

Again, the work presented here is my own. But these sources were helpful as I prepared this project:

UCSD DSE220x Machine Learning Fundamentals, 9.4 "Case Study: Personality Assessment"

https://en.wikipedia.org/wiki/Big_Five_personality_traits

ODA: https://data.oecd.org/oda/net-oda.htm

GFCF: https://data.oecd.org/gdp/investment-gfcf.htm

SPI: https://datacatalog.worldbank.org/search/dataset/0037996

CPIA: https://databank.worldbank.org/reports.aspx?source=country-policy-and-institutional-assessment