40.016: The Analytics Edge

Week 2 Lecture 2

Principal Component Analysis: An index of Social Progress

Term 5, 2022

Opinion

‘We’re No. 28! And Dropping!’

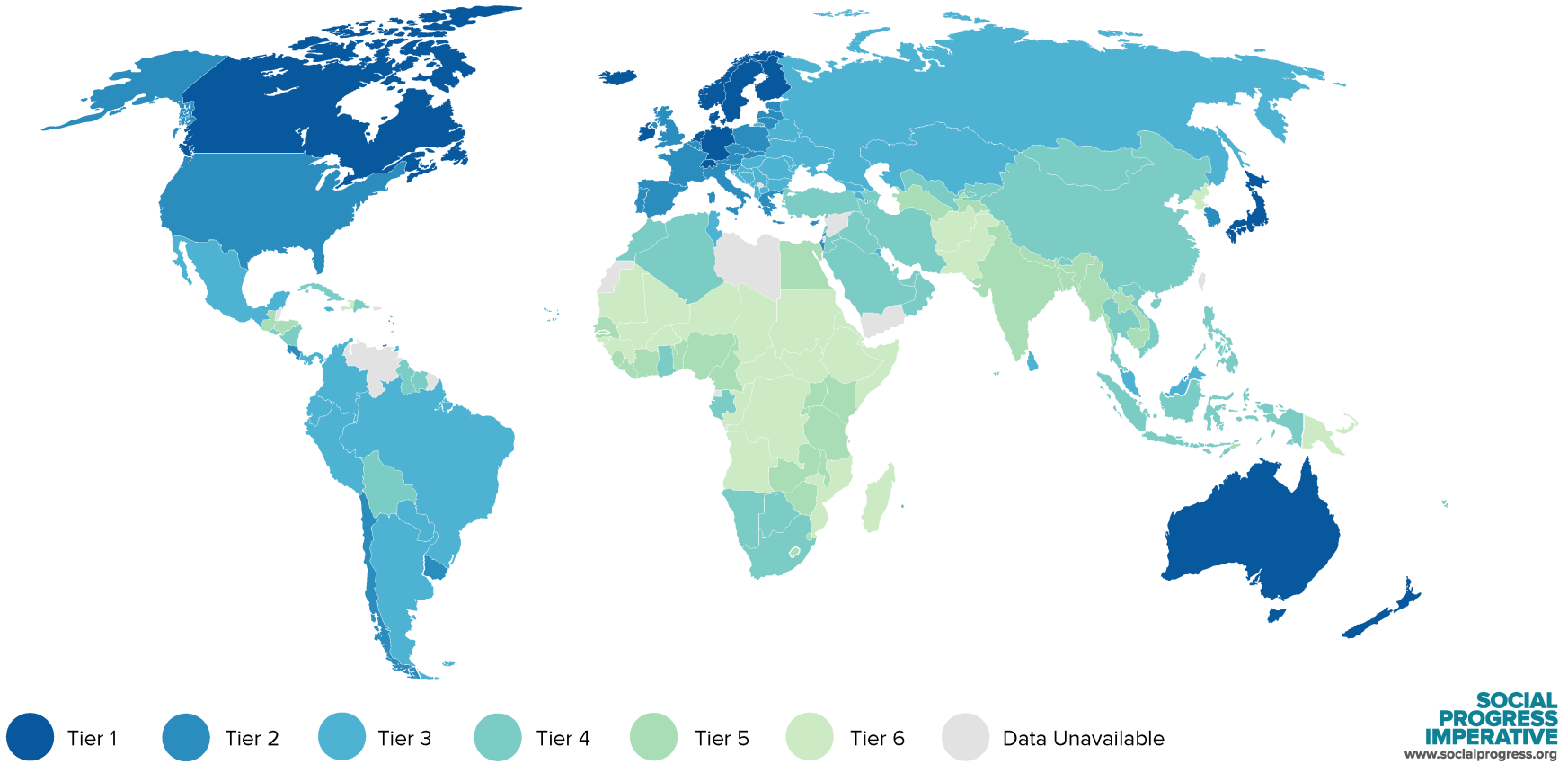
A measure of social progress finds that the quality of life has dropped in America over the last decade, even as it has risen almost everywhere else.

By Nicholas Kristof
Opinion Columnist

Sept. 9, 2020

1835

2020 Social Progress Index



https://youtu.be/UdMNuzIIois

Social Progress Index

The Social Progress Index is a well-established measure, published since 2013, that is meant to catalyze improvement and drive action by presenting social outcome data in a useful and reliable way. Composed of multiple dimensions, the Social Progress Index can be used to benchmark success and provide a holistic, transparent, outcome-based measure of a countrys well-being that is independent of economic indicators. Policymakers, businesses, and countries citizens alike can use it to compare their country against others on different facets of social progress, allowing the identification of specific areas of strength or weakness.

- taken from the Methodology summary

Figure 1 / Social Progress Index Component-Level Framework

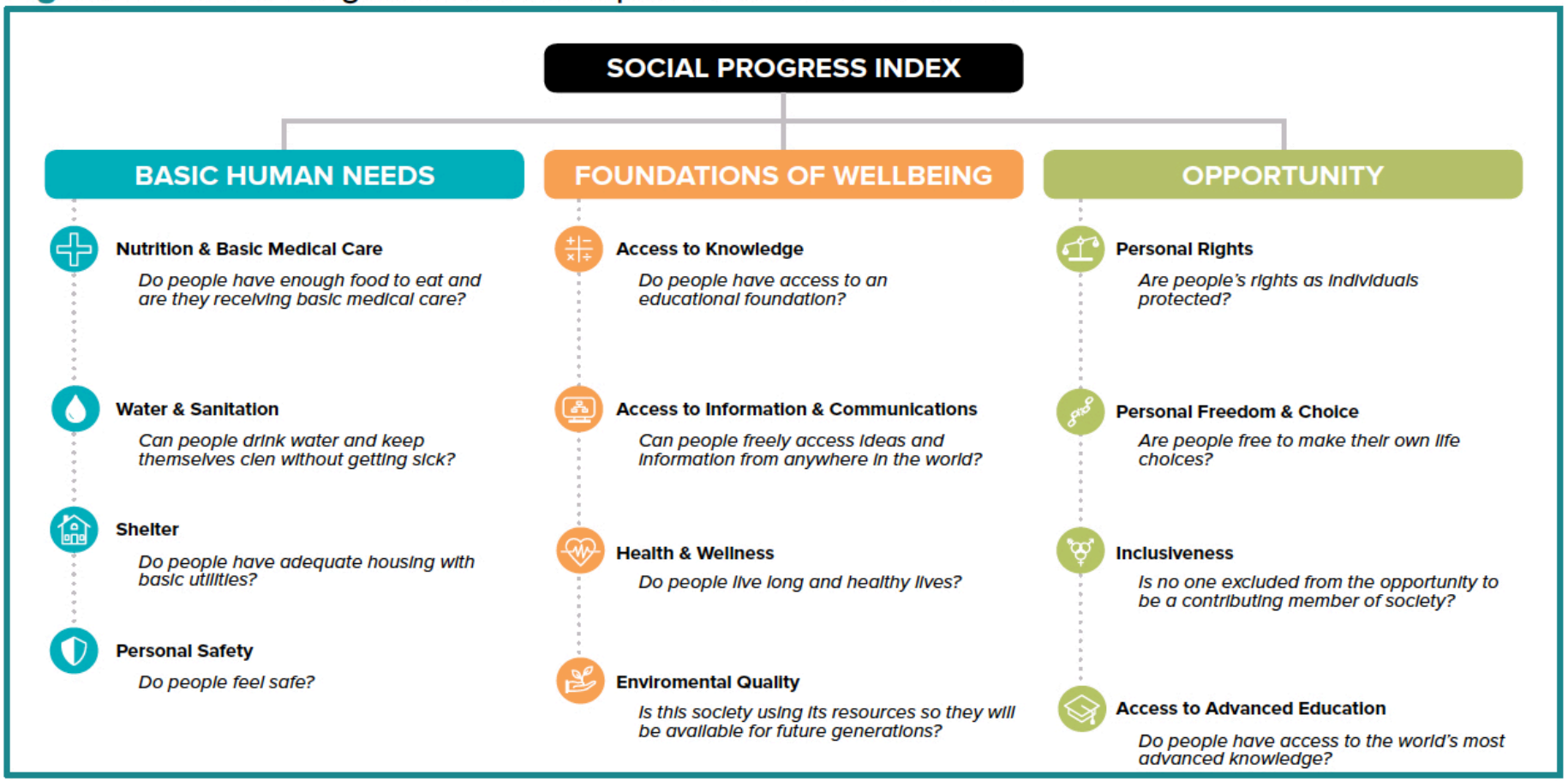


Figure 2 / Social Progress Index Indicator-Level Framework

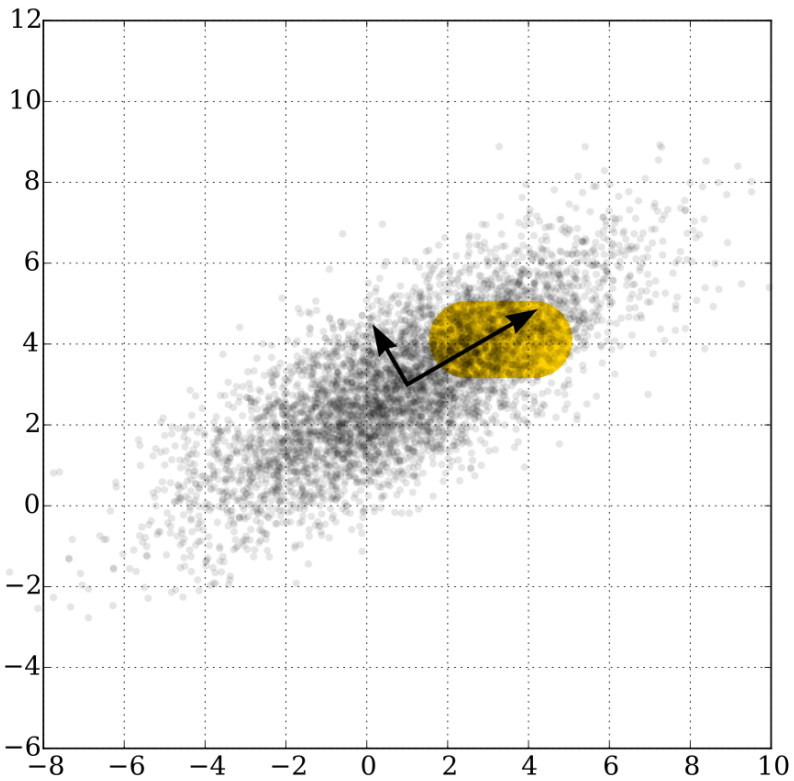
Key questions

1. What can we understand about social progress from the raw data?
2. Can we identify where the major differences in social progress occurs?
3. Can we propose some kind of social progress index from our analysis.

Supervised learning vs. Unsupervised learning

	Supervised learning	Unsupervised learning
Data	Data is labelled. Output: y, input: $x_1, \ldots, x_p$.	Unlabelled data. Only input: $x_1, \ldots, x_p$.
Goal	Predicting the response, classifying, etc.	Not direct. Understanding the structure. Clustering, dimension reduction
Assessment	Break into training and test sets and validate	Difficult
Methods	Regression (linear, logistic, ...), Random forests, decision trees, ...	PCA, SVD, k-means, hierarchichal clustering, ...

Principal Component Analysis



We observe the $n \times p$ data matrix:

$$\boldsymbol{x} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} =: \begin{bmatrix} \underline{x}_1^\top \\ \vdots \\ \underline{x}_n^\top \end{bmatrix}$$

1. n = number of observations
2. p = number of features
3. $X_1, \dots, X_p$ = the feature variables

Image courtesy: `https://commons.wikimedia.org/wiki/File:GaussianScatterPCA.svg`

PCA

Can we find $Z_1, \ldots, Z_q$ with $q << p$ such that

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \ldots + \phi_{p1} X_p = \underline{X}^{\mathsf{T}} \underline{\phi}_1,$$

$$\vdots$$

$$Z_q = \phi_{1q} X_1 + \phi_{2q} X_2 + \ldots + \phi_{pq} X_p = \underline{X}^{\mathsf{T}} \underline{\phi}_q,$$

PCA: linear algebra

1. Any real symmetric $p \times p$ matrix A has p orthonormal eigenvectors $\underline{v}_1, \ldots, \underline{v}_p$ and associated eigenvalues $\lambda_1, \ldots, \lambda_p$ respectively such that for $1 \leq i, j \leq p$,
   1. $A\underline{v}_i = \lambda_i \underline{v}_i$,
   2. $\underline{v}_i^T \underline{v}_i = 1$,
   3. $\underline{v}_i^T \underline{v}_j = 0$ for $i \neq j$.

   Moreover, $\underline{v}_i \in \mathbb{R}^p$ and $\lambda_i \in \mathbb{R}$.

PCA: approach 1

PCA: approach 2

PCA: approach 2

PCA: how many?

PCA: limitations?

Albeit being quite useful, we need to keep in mind certain limitations.

- PCA depends on the scaling of the variables and hence one needs to scale the variables, essentially bringing them all down to the same scale.
- We assume that a linear relationships between variables can explain the total variability.
- The eventual principal components loses some interpretability.

Resources

- https://www.datacamp.com/community/tutorials/pca-analysis-r
- https://towardsdatascience.com/principal-component-analysis-pca-101-using-r-361f4c53a9ff
- http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/
- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. The Elements of Statistical Learning. Vol. 1. 10. Springer series in statistics, New York.

	Score	Rank AE	Rank SPI
Switzerland	6.86	1	6
Norway	6.81	2	1
Denmark	6.76	3	2
Sweden	6.65	4	5
Finland	6.51	5	3
Iceland	6.42	6	9
Luxembourg	6.4	7	14
Japan	6.3	8	13
Canada	6.23	9	7
Germany	6.22	10	11
Netherlands	6.16	11	10
New Zealand	6.09	12	4
Australia	6.09	13	8
Ireland	6.02	14	12
Belgium	5.9	15	16
Italy	5.88	16	23
Cyprus	5.78	17	26
Austria	5.75	18	15
Korea, Republic of	5.73	19	17
France	5.69	20	18

	Score	Rank AE	Rank SPI
United Kingdom	5.69	21	20
Spain	5.65	22	19
Estonia	5.63	23	24
Slovenia	5.51	24	22
Portugal	5.47	25	21
United States	5.34	26	28
Greece	5.24	27	27
Czechia	5.18	28	25
Malta	4.99	29	30
Singapore	4.91	30	29
Lithuania	4.76	31	32
Poland	4.53	32	31
Israel	4.47	33	33
Costa Rica	4.42	34	37
Uruguay	4.34	35	38
Slovakia	4.22	36	36
Latvia	4.15	37	35
Croatia	4.13	38	39
Chile	3.78	39	34
Barbados	3.67	40	42

	Score	Rank AE	Rank SP
Ethiopia	−5.43	144	145
Angola	−5.45	145	151
Mali	−5.46	146	150
Sierra Leone	−5.49	147	134
Mozambique	−5.67	148	142
Pakistan	−5.88	149	141
Madagascar	−5.93	150	148
Congo, Republic of	−6.1	151	149
Guinea−Bissau	−6.46	152	152
Papua New Guinea	−6.8	153	153
Niger	−6.93	154	157
Afghanistan	−7.05	155	155
ongo, Democratic Republic of	−7.26	156	156
Guinea	−7.65	157	154
Burundi	−8	158	158
Eritrea	−9.03	159	160
Somalia	−9.47	160	159
Chad	−10.13	161	162
South Sudan	−10.18	162	163
Central African Republic	−11.07	163	161