MACS 40800 Unsupervised Machine Learning

Becky Lau, Rui He, Oliver Tang

Project Preliminary Results

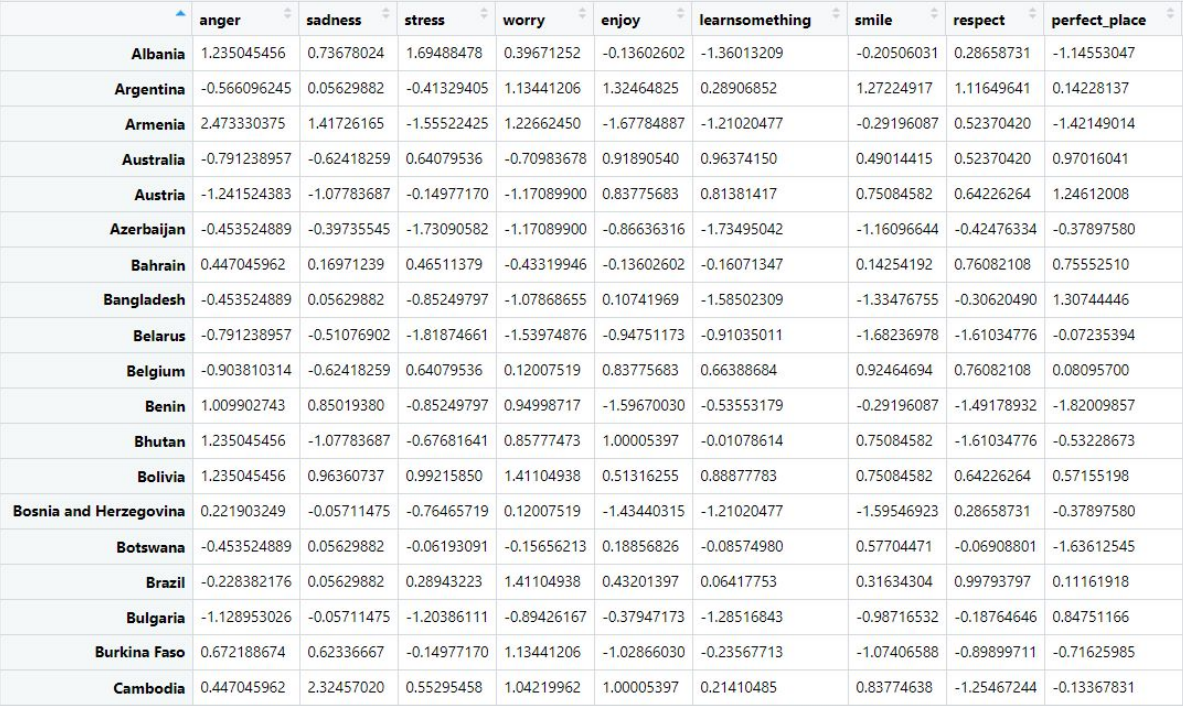
30th October, 2019

Introduction

We have a high dimensional dataset (Gallup World Poll) with many variables that are related to the broader concept of well-being (https://media.gallup.com/dataviz/www/WP_Questions_WHITE.pdf, pg.5). Each country has an overall score in well-being constructs such as the different kinds of emotions they have experienced. The goal of this project is to minimize the complexity of this feature space to make it more approachable and understandable. We are particularly interested in whether certain countries cluster together in terms of wellbeing to help discover underlying patterns.

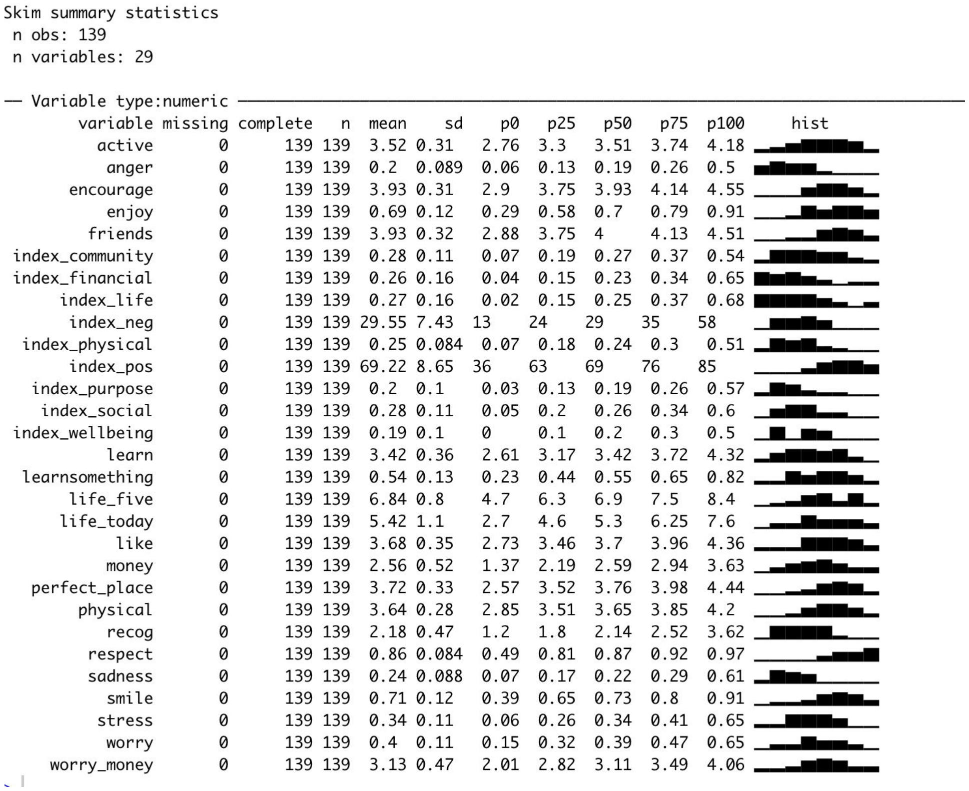
Data Munging

The variables come in many different forms in the surveys. Some variables were likert scale responses (1-7 scale, 7 being most extreme) where we took the weighted mean response. Some variables were categorical responses with three categories (Thriving, Struggling, Suffering). For these variables we took the percentage that reported "Thriving". Finally some variables were binary responses of Yes and No. For these variables we took the percentage that reported "yes". We took out all the countries that had at least one missing datapoint. A sample of data is included below:



	anger	sadness	stress	worry	enjoy	learnsomething	smile	respect	perfect_place
Albania	1.235045456	0.73678024	1.69488478	0.39671252	-0.13602602	-1.36013209	-0.20506031	0.28658731	-1.14553047
Argentina	-0.566096245	0.05629882	-0.41329405	1.13441206	1.32464825	0.28906852	1.27224917	1.11649641	0.14228137
Armenia	2.473330375	1.41726165	-1.55522425	1.22662450	-1.67784887	-1.21020477	-0.29196087	0.52370420	-1.42149014
Australia	-0.791238957	-0.62418259	0.64079536	-0.70983678	0.91890540	0.96374150	0.49014415	0.52370420	0.97016041
Austria	-1.241524383	-1.07783687	-0.14977170	-1.17089900	0.83775683	0.81381417	0.75084582	0.64226264	1.24612008
Azerbaijan	-0.453524889	-0.39735545	-1.73090582	-1.17089900	-0.86636316	-1.73495042	-1.16096644	-0.42476334	-0.37897580
Bahrain	0.447045962	0.16971239	0.46511379	-0.43319946	-0.13602602	-0.16071347	0.14254192	0.76082108	0.75552510
Bangladesh	-0.453524889	0.05629882	-0.85249797	-1.07868655	0.10741969	-1.58502309	-1.33476755	-0.30620490	1.30744446
Belarus	-0.791238957	-0.51076902	-1.81874661	-1.53974876	-0.94751173	-0.91035011	-1.68236978	-1.61034776	-0.07235394
Belgium	-0.903810314	-0.62418259	0.64079536	0.12007519	0.83775683	0.66388684	0.92464694	0.76082108	0.08095700
Benin	1.009902743	0.85019380	-0.85249797	0.94998717	-1.59670030	-0.53553179	-0.29196087	-1.49178932	-1.82009857
Bhutan	1.235045456	-1.07783687	-0.67681641	0.85777473	1.00005397	-0.01078614	0.75084582	-1.61034776	-0.53228673
Bolivia	1.235045456	0.96360737	0.99215850	1.41104938	0.51316255	0.88877783	0.75084582	0.64226264	0.57155198
Bosnia and Herzegovina	0.221903249	-0.05711475	-0.76465719	0.12007519	-1.43440315	-1.21020477	-1.59546923	0.28658731	-0.37897580
Botswana	-0.453524889	0.05629882	-0.06193091	-0.15656213	0.18856826	-0.08574980	0.57704471	-0.06908801	-1.63612545
Brazil	-0.228382176	0.05629882	0.28943223	1.41104938	0.43201397	0.06417753	0.31634304	0.99793797	0.11161918
Bulgaria	-1.128953026	-0.05711475	-1.20386111	-0.89426167	-0.37947173	-1.28516843	-0.98716532	-0.18764646	0.84751166
Burkina Faso	0.672188674	0.62336667	-0.14977170	1.13441206	-1.02866030	-0.23567713	-1.07406588	-0.89899711	-0.71625985
Cambodia	0.447045962	2.32457020	0.55295458	1.04219962	1.00005397	0.21410485	0.83774638	-1.25467244	-0.13367831

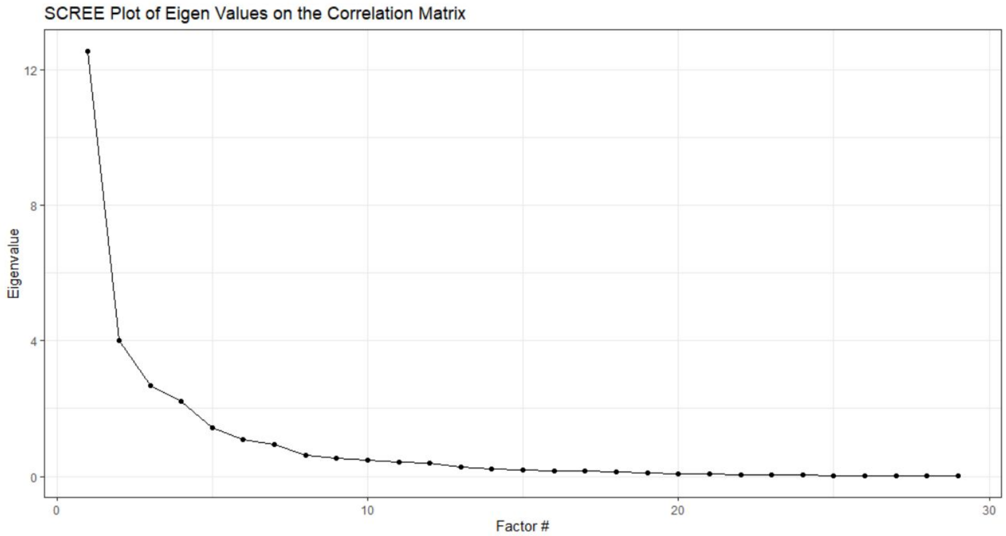
Then, we generated a descriptive profile of the dataset, which includes the number of observations, mean, variance and distribution of data. The summary shows that most of the variables we selected follow a normal distribution with a range of variance.



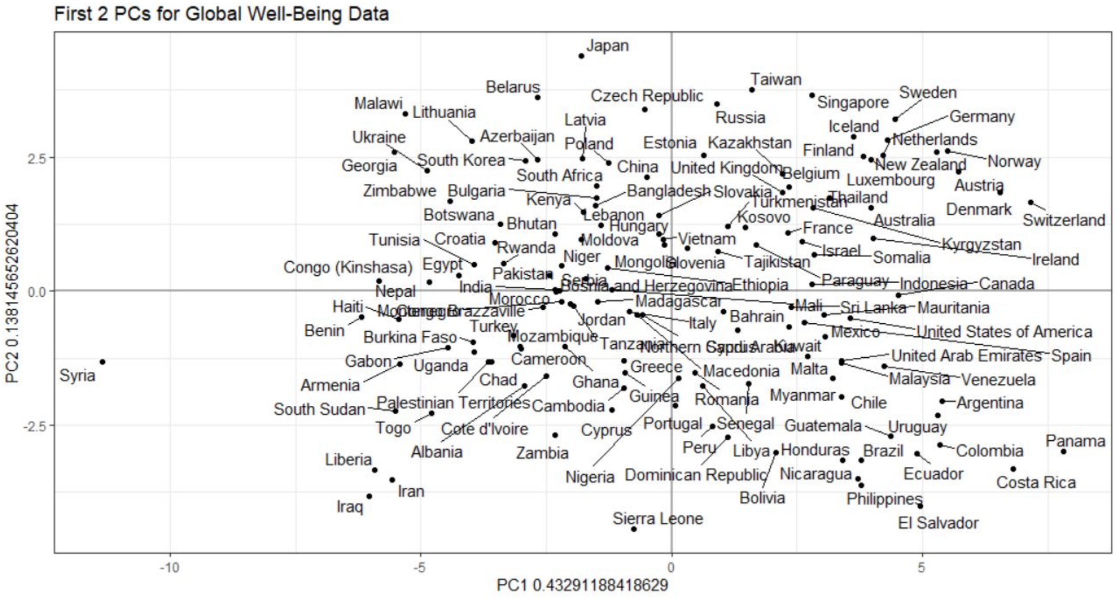
```
Skim summary statistics
 n obs: 139
 n variables: 29

── Variable type:numeric ──────────────────────────────────────────────────────────────────────────────
        variable missing complete   n  mean    sd    p0   p25   p50   p75  p100     hist
          active       0      139 139  3.52 0.31   2.76  3.3   3.51  3.74  4.18 ▂▃▅▇▇▇▆▂
           anger       0      139 139  0.2  0.089  0.06  0.13  0.19  0.26  0.5  ▅▇▆▆▂▁▁▁
       encourage       0      139 139  3.93 0.31   2.9   3.75  3.93  4.14  4.55 ▁▁▁▅▇▇▅▂
           enjoy       0      139 139  0.69 0.12   0.29  0.58  0.7   0.79  0.91 ▁▁▂▇▅▇▇▅
         friends       0      139 139  3.93 0.32   2.88  3.75  4     4.13  4.51 ▁▁▂▂▆▇▆▃
 index_community       0      139 139  0.28 0.11   0.07  0.19  0.27  0.37  0.54 ▂▇▇▇▆▆▃▂
 index_financial       0      139 139  0.26 0.16   0.04  0.15  0.23  0.34  0.65 ▇▆▇▅▃▁▂▂
      index_life       0      139 139  0.27 0.16   0.02  0.15  0.25  0.37  0.68 ▇▇▇▇▅▃▁▃
       index_neg       0      139 139 29.55 7.43  13    24    29    35    58    ▁▆▆▇▅▁▁▁
  index_physical       0      139 139  0.25 0.084  0.07  0.18  0.24  0.3   0.51 ▂▇▆▇▃▂▁▁
       index_pos       0      139 139 69.22 8.65  36    63    69    76    85    ▁▁▁▃▆▇▇▅
   index_purpose       0      139 139  0.2  0.1    0.03  0.13  0.19  0.26  0.57 ▂▇▆▃▂▂▁▁
    index_social       0      139 139  0.28 0.11   0.05  0.2   0.26  0.34  0.6  ▁▅▇▇▃▃▁▁
 index_wellbeing       0      139 139  0.19 0.1    0     0.1   0.2   0.3   0.5  ▁▇▁▆▅▁▁▁
           learn       0      139 139  3.42 0.36   2.61  3.17  3.42  3.72  4.32 ▂▅▇▇▆▇▃▁
  learnsomething       0      139 139  0.54 0.13   0.23  0.44  0.55  0.65  0.82 ▂▂▇▅▇▇▅▂
       life_five       0      139 139  6.84 0.8    4.7   6.3   6.9   7.5   8.4  ▁▂▃▆▇▃▇▂
      life_today       0      139 139  5.42 1.1    2.7   4.6   5.3   6.25  7.6  ▁▃▃▇▅▅▅▃
            like       0      139 139  3.68 0.35   2.73  3.46  3.7   3.96  4.36 ▂▂▂▇▇▇▆▃
           money       0      139 139  2.56 0.52   1.37  2.19  2.59  2.94  3.63 ▁▃▅▆▇▅▃▃
   perfect_place       0      139 139  3.72 0.33   2.57  3.52  3.76  3.98  4.44 ▁▁▂▃▆▇▆▂
        physical       0      139 139  3.64 0.28   2.85  3.51  3.65  3.85  4.2  ▁▁▂▅▇▇▅▂
           recog       0      139 139  2.18 0.47   1.2   1.8   2.14  2.52  3.62 ▁▇▇▇▇▂▁▁
         respect       0      139 139  0.86 0.084  0.49  0.81  0.87  0.92  0.97 ▁▁▁▁▃▅▅▇
         sadness       0      139 139  0.24 0.088  0.07  0.17  0.22  0.29  0.61 ▂▇▆▅▁▁▁▁
           smile       0      139 139  0.71 0.12   0.39  0.65  0.73  0.8   0.91 ▁▂▂▃▆▇▆▃
          stress       0      139 139  0.34 0.11   0.06  0.26  0.34  0.41  0.65 ▂▂▇▇▇▅▁▁
           worry       0      139 139  0.4  0.11   0.15  0.32  0.39  0.47  0.65 ▁▃▃▇▅▅▃▁
     worry_money       0      139 139  3.13 0.47   2.01  2.82  3.11  3.49  4.06 ▂▂▃▆▇▆▃▃
```

Principal Component Analysis

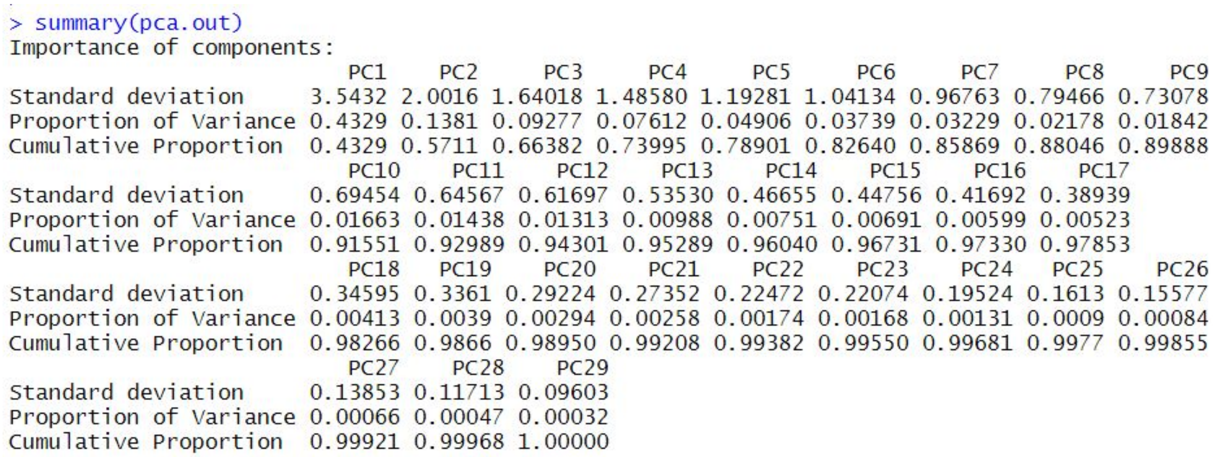
As the dataset is significantly high in dimensions, namely 29 features in total, PCA can reduce the dimensionality and make the information embedded in the data more approachable and understandable. At the same time, since all the features here are regarding people’s well-being, it makes sense that they may share correlation to some extent. We did PCA via eigenvalue decomposition (EVD) as well as singular value decomposition (SVD).



It can be found that the first eigenvalue on the correlation matrix is significantly larger than the rest, which indicates that the first two principal components have the most part of the information in the dataset. Then we plotted all geographies (a good term to avoid political disputes) as data points in terms of the first two principal components.

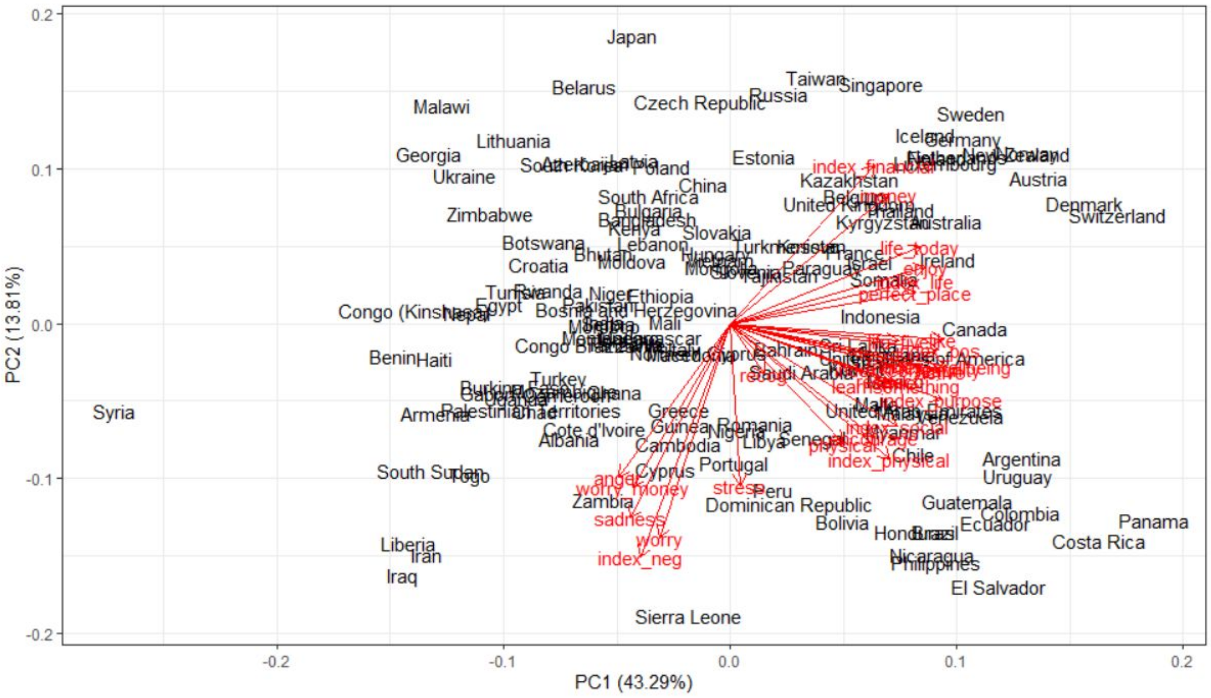


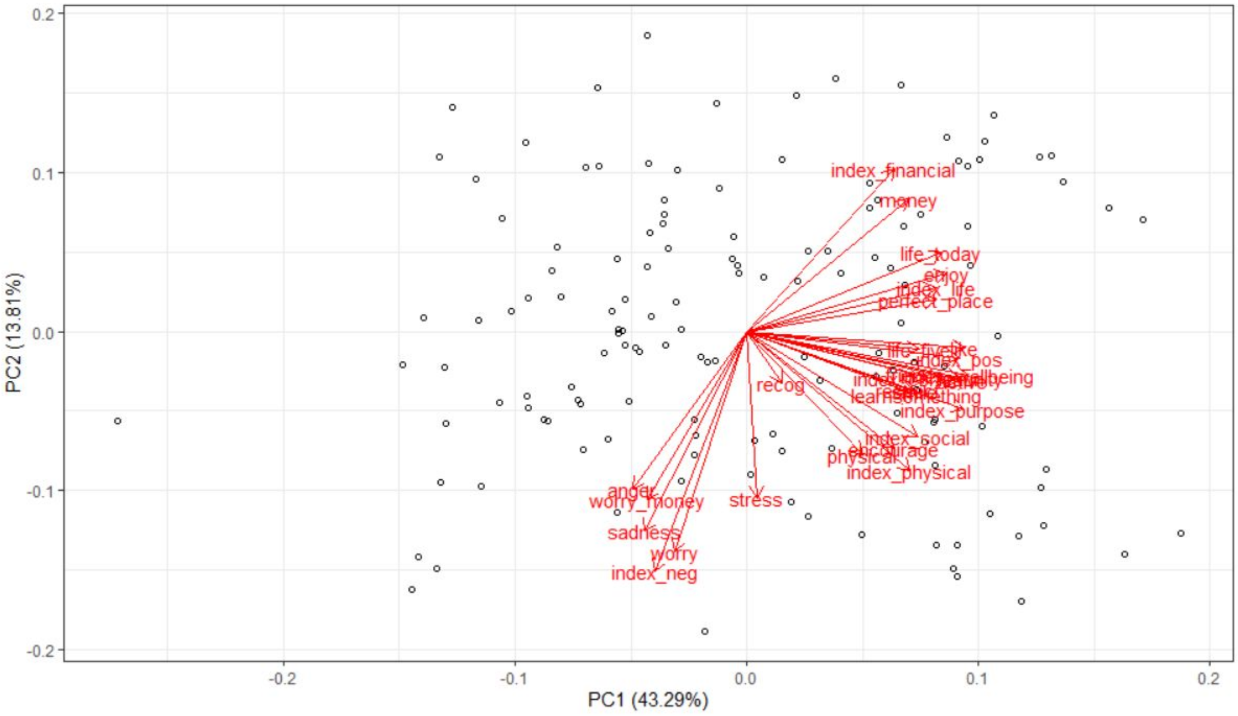
Same procedure is conducted using prcomp function which utilize SVD.



```
> summary(pca.out)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     3.5432 2.0016 1.64018 1.48580 1.19281 1.04134 0.96763 0.79466 0.73078
Proportion of Variance 0.4329 0.1381 0.09277 0.07612 0.04906 0.03739 0.03229 0.02178 0.01842
Cumulative Proportion  0.4329 0.5711 0.66382 0.73995 0.78901 0.82640 0.85869 0.88046 0.89888
                          PC10    PC11    PC12    PC13    PC14    PC15    PC16    PC17
Standard deviation     0.69454 0.64567 0.61697 0.53530 0.46655 0.44756 0.41692 0.38939
Proportion of Variance 0.01663 0.01438 0.01313 0.00988 0.00751 0.00691 0.00599 0.00523
Cumulative Proportion  0.91551 0.92989 0.94301 0.95289 0.96040 0.96731 0.97330 0.97853
                          PC18   PC19    PC20    PC21    PC22    PC23    PC24   PC25    PC26
Standard deviation     0.34595 0.3361 0.29224 0.27352 0.22472 0.22074 0.19524 0.1613 0.15577
Proportion of Variance 0.00413 0.0039 0.00294 0.00258 0.00174 0.00168 0.00131 0.0009 0.00084
Cumulative Proportion  0.98266 0.9866 0.98950 0.99208 0.99382 0.99550 0.99681 0.9977 0.99855
                          PC27    PC28    PC29
Standard deviation     0.13853 0.11713 0.09603
Proportion of Variance 0.00066 0.00047 0.00032
Cumulative Proportion  0.99921 0.99968 1.00000
```

Also plot using the first two principal components, with original coordinates is shown. At this stage we can tell that, with this dataset, PCA via EVD and SVD give almost the same results.



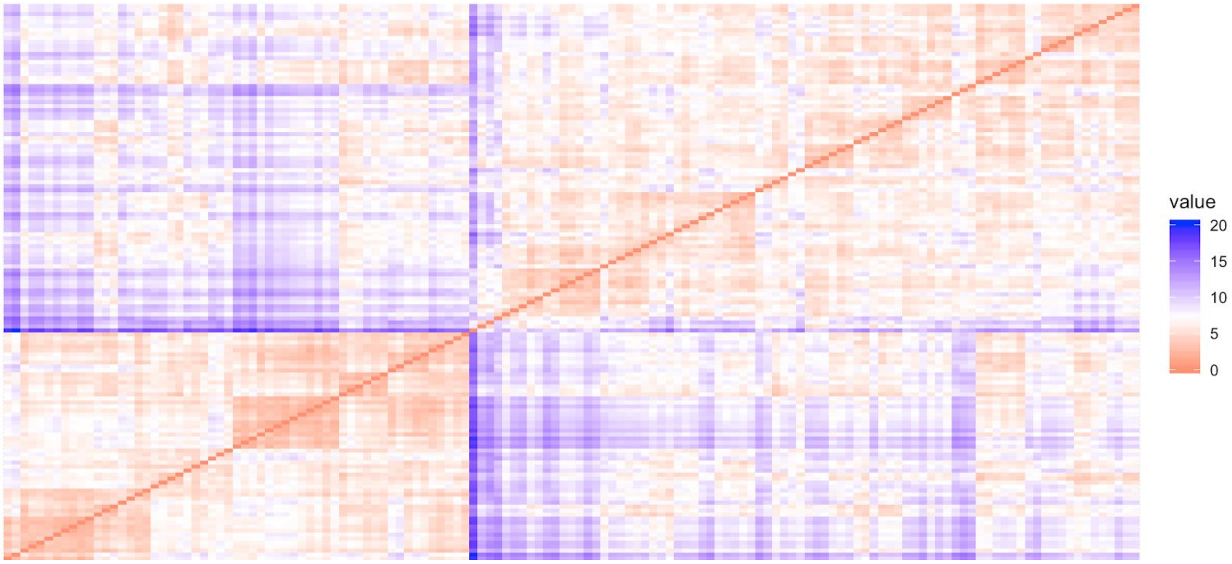


Regarding the original feature coordinates, we can easily find that those positive ones (pointing to upper right) are aligned opposite to those negative ones (pointing to lower left), such as money (people think they have enough money) and anger (people experience anger yesterday). Those features which are more complicated, cannot be addressed binarily, are essentially perpendicular to positive ones and negative ones, pointing to lower right. Theoretically, it supports that the two principal components decomposed the variance, with one correlated more with binary features, one correlated more with non-binary features.

Cluster Analysis with K-mean algorithm

The results from the PCA above suggest that there may exist some clusterability in the data. Theoretically, it is also reasonable to speculate that the “well-being” of countries may display some geospatial patterns as the mental and physical states of a country can be affected by the shared geopolitical, cultural environments of countries nearby. Therefore, we decided to perform a clustering analysis with hard partition.

To diagnose the clusterability, we used ODI for visualization and calculated the Hopkin Statics. As we can see in the ODI figure, the upper left and lower right square suggest that the data contains patterns and probably clusterable. The H stat is 0.29 (between 0-1), which is also suggesting a possibility for clusters to exist in the data set.

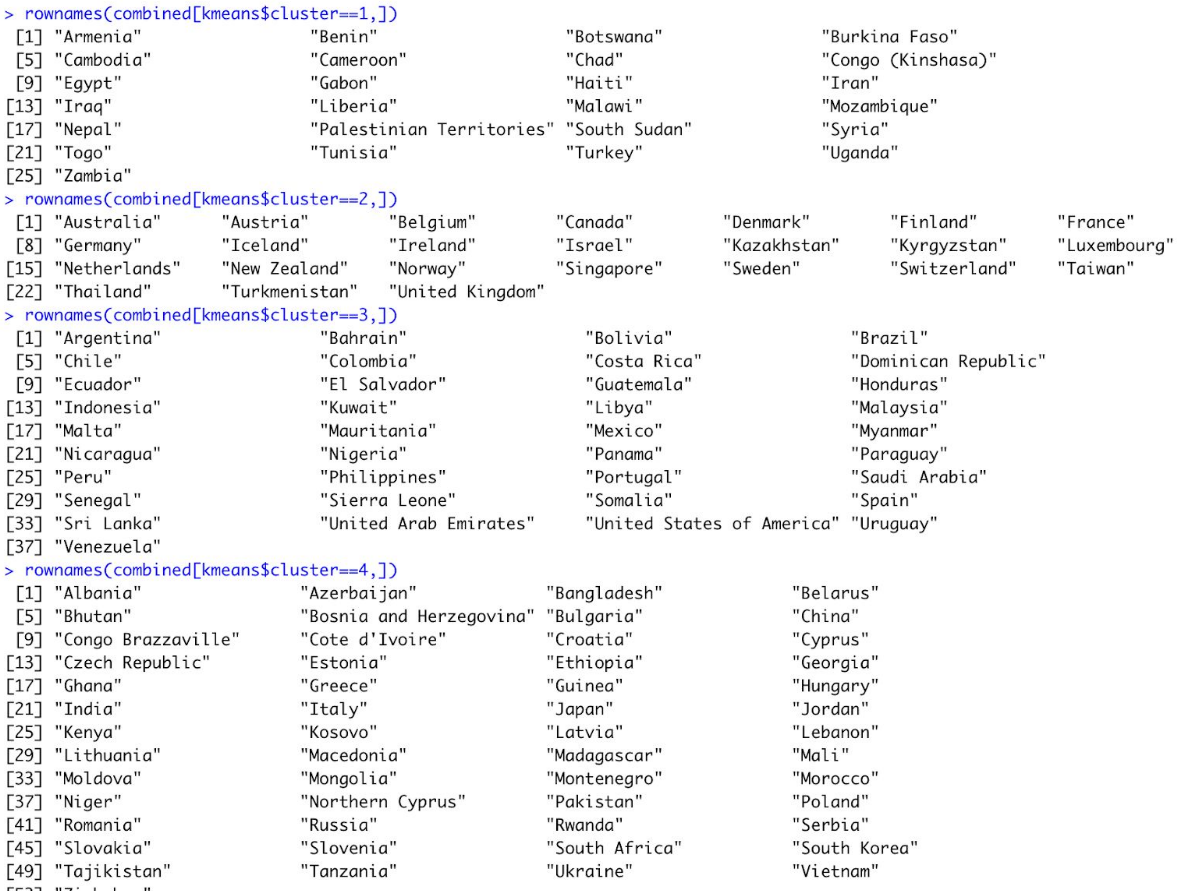


We then proceeded to run a k-mean algorithm with k set to 4. The reason we chose k=4 is that from the PCA result, the first 2 principal components explain the majority variance in the dataset (4 combinations of the 2 PCs). In addition, the validation test also showed that the Dunn's score is highest when k is equal to 4. The connectivity, however, suggest that k is optimal when set to 2. Here in the preliminary report, we demonstrate the result for 4 clusters, but we can compare the output with k being set to different values in the final report.

The sizes of the four clusters are 25, 24, 37 and 53. Some geospatial patterns did emerge from the clustering analysis; the first cluster includes more middle east countries, whose negative feelings such as anger, worry and stress levels are generally higher whereas the positive feelings such as financial stability, physical and enjoyment of life are low. The second cluster are mostly developed countries in Europe. These countries tend to have low values for variables indicating negative well-being (will be referred to as "negative variable") and high value for variables indicating positive well-being (which will be referred to as "positive variable"). The third cluster are mostly South American countries, with high negative variables but also high positive variables. United States is also grouped in this cluster. The fourth cluster are mostly asian and east european countries, which have low negative variables and also low positive variables.

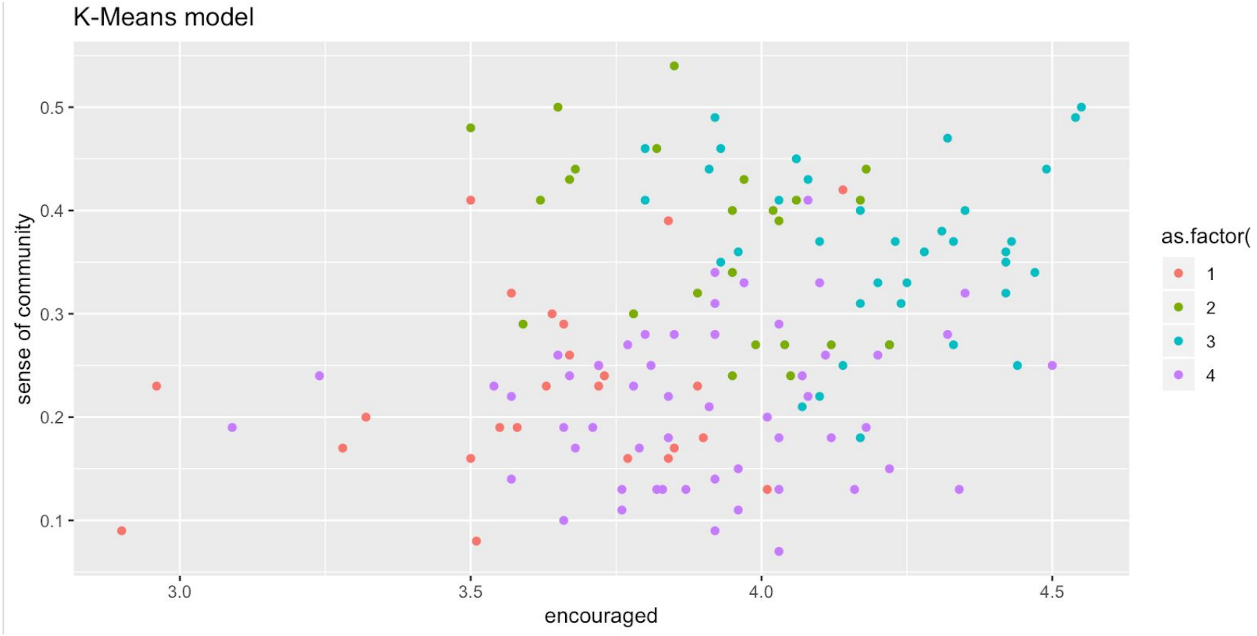
It is intuitive to assume that negative variables should be negatively correlated to the positive variables. For example, populations who feel more stress, anger and worry may also indicate having low life stability, smile less and worse physical health, and vice versa. Cluster 1 seems to be the countries that we commonly known as suffering from war and poverty. Cluster 2, on the other hand, are commonly known as the wealthier and more developed countries. And their data is consistent with the negative correlation between the positive and negative feelings.

Interestingly, cluster 3 and 4 show a counterintuitive trend of positive and negative variables. Cluster 3, despite the averages of the negative variables are high, shows high averages for positive variables (as if they both love and hate the world strongly). Whereas cluster 4 shows low score for both positive and negative variables (as if they don’t care about much). This may be related to the expressiveness of different countries, as South Americans are thought to be more expressive about their feelings whereas Asians are more conservative in terms of feeling expression. It can also be due to other sociopolitical factors and need further investigation.



```
> rownames(combined[kmeans$cluster==1,])
 [1] "Armenia"                "Benin"                  "Botswana"               "Burkina Faso"
 [5] "Cambodia"               "Cameroon"               "Chad"                   "Congo (Kinshasa)"
 [9] "Egypt"                  "Gabon"                  "Haiti"                  "Iran"
[13] "Iraq"                   "Liberia"                "Malawi"                 "Mozambique"
[17] "Nepal"                  "Palestinian Territories" "South Sudan"           "Syria"
[21] "Togo"                   "Tunisia"                "Turkey"                 "Uganda"
[25] "Zambia"
> rownames(combined[kmeans$cluster==2,])
 [1] "Australia"      "Austria"        "Belgium"        "Canada"         "Denmark"        "Finland"        "France"
 [8] "Germany"        "Iceland"        "Ireland"        "Israel"         "Kazakhstan"     "Kyrgyzstan"     "Luxembourg"
[15] "Netherlands"    "New Zealand"    "Norway"         "Singapore"      "Sweden"         "Switzerland"    "Taiwan"
[22] "Thailand"       "Turkmenistan"   "United Kingdom"
> rownames(combined[kmeans$cluster==3,])
 [1] "Argentina"                "Bahrain"                  "Bolivia"                  "Brazil"
 [5] "Chile"                    "Colombia"                 "Costa Rica"               "Dominican Republic"
 [9] "Ecuador"                  "El Salvador"              "Guatemala"                "Honduras"
[13] "Indonesia"                "Kuwait"                   "Libya"                    "Malaysia"
[17] "Malta"                    "Mauritania"               "Mexico"                   "Myanmar"
[21] "Nicaragua"                "Nigeria"                  "Panama"                   "Paraguay"
[25] "Peru"                     "Philippines"              "Portugal"                 "Saudi Arabia"
[29] "Senegal"                  "Sierra Leone"             "Somalia"                  "Spain"
[33] "Sri Lanka"                "United Arab Emirates"     "United States of America" "Uruguay"
[37] "Venezuela"
> rownames(combined[kmeans$cluster==4,])
 [1] "Albania"                "Azerbaijan"             "Bangladesh"             "Belarus"
 [5] "Bhutan"                 "Bosnia and Herzegovina" "Bulgaria"               "China"
 [9] "Congo Brazzaville"      "Cote d'Ivoire"          "Croatia"                "Cyprus"
[13] "Czech Republic"         "Estonia"                "Ethiopia"               "Georgia"
[17] "Ghana"                  "Greece"                 "Guinea"                 "Hungary"
[21] "India"                  "Italy"                  "Japan"                  "Jordan"
[25] "Kenya"                  "Kosovo"                 "Latvia"                 "Lebanon"
[29] "Lithuania"              "Macedonia"              "Madagascar"             "Mali"
[33] "Moldova"                "Mongolia"               "Montenegro"             "Morocco"
[37] "Niger"                  "Northern Cyprus"        "Pakistan"               "Poland"
[41] "Romania"                "Russia"                 "Rwanda"                 "Serbia"
[45] "Slovakia"               "Slovenia"               "South Africa"           "South Korea"
[49] "Tajikistan"             "Tanzania"               "Ukraine"                "Vietnam"
```

For data visualization example, we used a scatter plot for two selected variables “feeling encouraged”, and “feeling the sense of community” with cluster as factor. We can see an overall positive correlation between the two variables, as we may expect that the sense of community can promote people’s feeling of being encouraged by others, or being encouraged creates a sense of belonging. The figure also shows that the clusters in fact roughly separated from each other. The result can also be visualized with a choropleth map with cluster number as a factor to demonstrate the geospatial patterns we found in the cluster analysis.



Summary

So far we have performed a PCA and k-mean cluster analysis. Our preliminary results showed reduced dimensionality of variables and clustering based on geospatial distribution. Such distribution may be contributed by cross-cultural and sociopolitical factors. Our plan for the final report includes (1) a detailed list of variables and their corresponding questions in the surveys for reference, (2) exploration of other clustering algorithm, (3) more data visualization figures (choropleth for examples), and (4) some research on the potential explanation of the patterns we observed (such as attitude expressiveness, polarity of social environment, stability of political environments).