Calculating Composite Demographic Indexes

July 27, 2022

Contents

Introduction **2**

Load Libraries **2**

Set Graphics Theme **2**

Load Data **3**

Folder References 3

Load Data 3

Utility Functions **4**

Functions for Calculating Indexes **4**

More Calculations **5**

Distributions **5**

Pairs Plot 5

Means, SD, Medians and IGR 7

PCA Analysis of Sub-indexes **8**

Raw Values, Unscaled 10

Scaled PCA 11

Percentiles 13

Examining Results **14**

Correlations 14

Graphics 16

Pairs Plot of All Indexes 18

Output Results **21**

Introduction

CBEP, like other National Estuary Programs will receive additional funding to support our programs via the "Bipartisan Infrastructure Law" signed into law last December.

EPA has recently released guidance for applying for those funds. A core component of the guidance is that overall, the NEP program should comply with the White House's "Justice 40" initiative, which requires that "at least 40% of the benefits and investments from BIL funding flow to disadvantaged communities."

EPA suggested that we use the National-scale EJSCREEN tools to help identify "disadvantaged communities" in our region. The EPA guidance goes on to suggest we focus on five demographic indicators:

- Percent low-income;
- Percent linguistically isolated;
- Percent less than high school education;
- Percent unemployed; and
- Low life expectancy.

This notebook examines the distributions of EPA's suggested demographic indicators and calculates relevant composite indexes a couple of different ways, and calculates how Casco Bay Census tracts compare at national, Statewide, and Local scales.

Load Libraries

```
library(tidyverse)
#> -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
#> v ggplot2 3.3.6     v purrr   0.3.4
#> v tibble  3.1.7     v dplyr   1.0.9
#> v tidyr   1.2.0     v stringr 1.4.0
#> v readr   2.1.2     v forcats 0.5.1
#> -- Conflicts ------------------------------------------ tidyverse_conflicts() --
#> x dplyr::filter() masks stats::filter()
#> x dplyr::lag()    masks stats::lag()
library(GGally)
#> Registered S3 method overwritten by 'GGally':
#>   method from
#>   +.gg   ggplot2
library(readr)
```

Set Graphics Theme

This sets `ggplot()`graphics for no background, no grid lines, etc. in a clean format suitable for (some) publications.

```
theme_set(theme_classic())
```

Load Data

Folder References

I use folder references to allow limited indirection, thus making code from GitHub repositories more likely to run "out of the box".

```
data_folder <- "Original_Data"
dir.create(file.path(getwd(), 'figures'), showWarnings = FALSE)
```

I use the "Original_Data" folder to retain data in the form originally downloaded. That minimizes the chances of inadvertently modifying the source data. All data was accessed via EJScreen. The 2021 EJSCREEN Data was accessed on July 26, 2022, at https://gaftp.epa.gov/EJSCREEN/2021/. I downloaded geodatabases, and open the geospatial data they contained in ArcGIS and exported the tabular attribute data to CSV files. That tabular CSV data is provided in the "Original Data" folder here.

The "figures" folder isolates "final" versions of any graphics I produce. That just makes it a bit easier to find final products in what can sometimes be fairly large GitHub Repositories (although not here).

Load Data

The tabular (National) source data is quite extensive (over 50 MB), so I have not included it in the GitHub repository (GitHub does not appreciate files over 100 MB). The raw CSV file contains 74001 records, and 166 columns. Most, but not all are numeric. The Health data is significantly smaller. smaller. The large files also poses potential data access challenges in R.

I read in just the required data columns for now.

```
the_file <- 'EJSCREEN_Full_tracts.txt'
the_path <- file.path(data_folder, the_file)
the_data <- read_csv(the_path,
                     n_max = 74001,
                     col_types = cols_only(
                       ID = col_character(),
                       LOWINCPCT = col_double(),
                       LINGISOPCT = col_double(),
                       LESSHSPCT  = col_double(),
                       UNEMPPCT = col_double(),
                       P_LWINCPCT = col_double(),
                       P_LNGISPCT = col_double(),
                       P_LESHSPCT  = col_double(),
                       P_UNEMPPCT = col_double()))
```

```
the_file <- 'Tract2010_LifeExpectancy.txt'
the_path <- file.path(data_folder, the_file)
life_data <- read_csv(the_path,
                      n_max = 73057,
                      col_types = cols_only(LIFEEXP = col_double(),
                                            GEOID10 = col_character(),
                                            State = col_character(),
                                            County = col_character(),
                                            Life_Expectancy_Standard_Error = col_double(),
                                            Shape_Length = col_double(),
                                            Shape_Area = col_double()))
```

```r
miles_per_meter <- 0.000621371
sq_miles_per_sq_meters <- miles_per_meter^2
```

```r
the_data <- inner_join(life_data, the_data, by = c('GEOID10' = 'ID')) %>%
  relocate(LIFEEXP , .after = County) %>%
  rename(LIFEEXP_SE = Life_Expectancy_Standard_Error,
         Perimeter = Shape_Length,
         Area = Shape_Area) %>%
  mutate(Perimeter_km = Perimeter / 1000,
         Perimeter = Perimeter * miles_per_meter,
         Area_ha = Area / 10000,
         Area = Area * sq_miles_per_sq_meters,
         Shape_Index = Area / Perimeter^2) %>%
  mutate(NEG_LIFEEXP = 150 - LIFEEXP,              # Higher life expectancy is good
         LOWINCPCT = 100* LOWINCPCT,
         LESSHSPCT = 100* LESSHSPCT,
         LINGISOPCT =  100* LINGISOPCT,
         UNEMPPCT   = 100* UNEMPPCT) %>%
  relocate(Perimeter, Area, .after = Area_ha) %>%
  relocate(NEG_LIFEEXP, .after = LIFEEXP_SE)

rm(life_data)
```

Utility Functions

```r
quick_sum <- function(.dat)
  return(list(Mean = mean(.dat, na.rm = TRUE),
              SD = sd(.dat, na.rm = TRUE),
              Median = median(.dat, na.rm = TRUE),
              IQR = IQR(.dat, na.rm = TRUE)
              ))
```

```r
quick_percentile <- function(.dat) {
  L <- sum(! is.na(.dat))
  val <- rank(.dat) / L
  val[is.na(.dat)] <- NA
  return(val)
  }
```

Functions for Calculating Indexes

```r
calc_index_1 <- function(.data) {
  index_1 <- with(.data,
    (NEG_LIFEEXP + LOWINCPCT + LESSHSPCT +LINGISOPCT + UNEMPPCT) / 5)
  return(index_1)
}
```

The primary alternative is to calculate percentiles within each sub-index, and sum those. That makes the composite index approximately scale-free in each sub-index. Again, because of correlations among sub-indexes, that won't be quite correct, but it will be close.

```
calc_index_2 <- function(.data) {
  index_2 <- with(.data,
                  (P_NEG_LIFEEXP + P_LWINCPCT + P_LESHSPCT +
                     P_LNGISPCT +
                     P_UNEMPPCT) / 5)
  return(index_2)
}
```

More Calculations

```
the_data <- the_data %>%
  mutate(P_NEG_LIFEEXP = quick_percentile(NEG_LIFEEXP) * 100) %>%
  relocate(P_NEG_LIFEEXP, .before = P_LWINCPCT)
```

The following depends on having columns with the correct names, and there is no error checking....

- Index_1 is based on averaging the VALUES
- Index_2 is based on averaging the PERCENTILES
- P_Index_1 records the percentiles of Index 1, and is thus scale-free.
- P_Index_2 shows percentiles of Index 2.

```
the_data$Index_1 <- calc_index_1(the_data)
the_data$Index_2 <- calc_index_2(the_data)
the_data$P_Index_1 <- quick_percentile(the_data$Index_1)
the_data$P_Index_2 <- quick_percentile(the_data$Index_2)
```

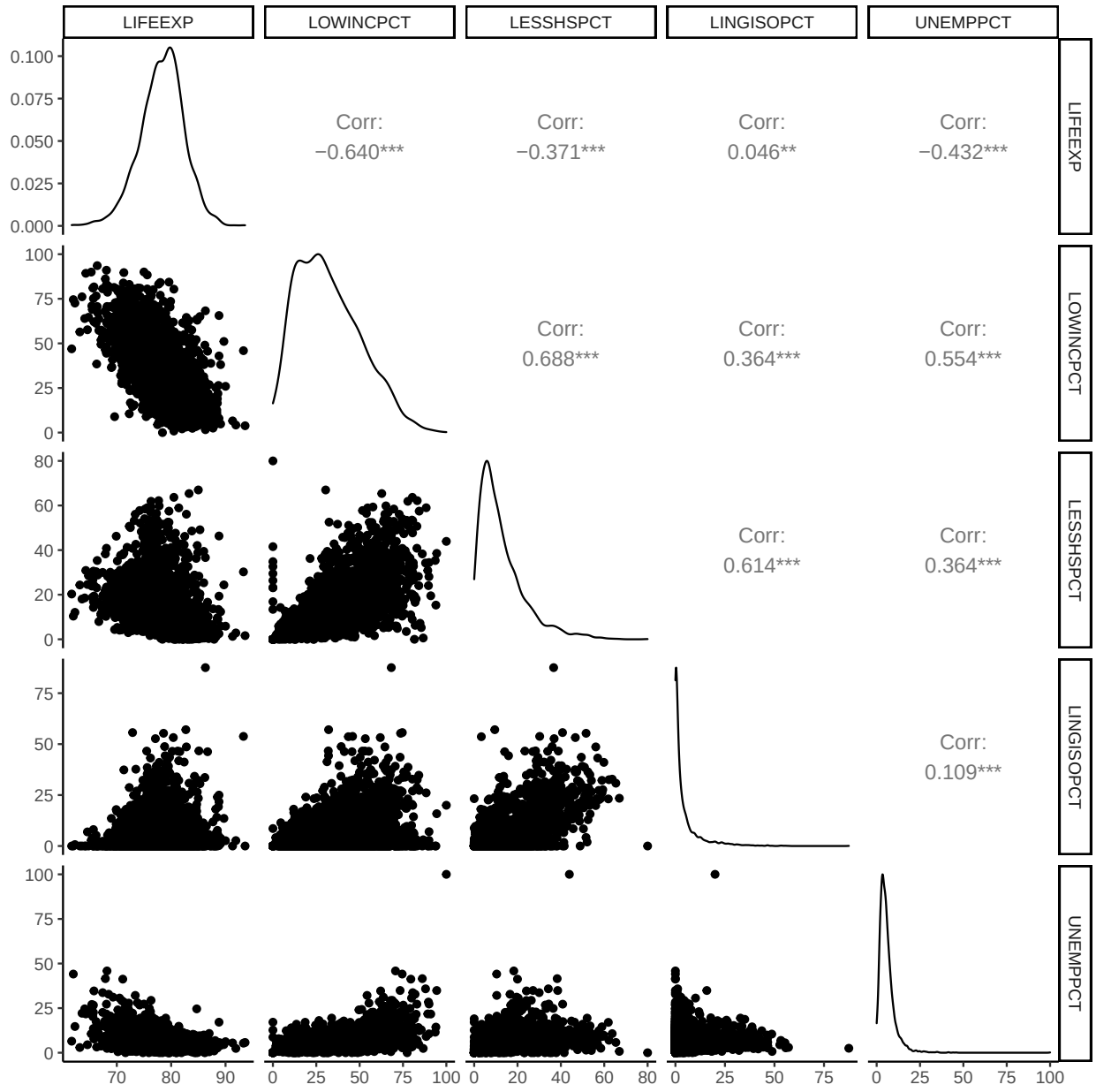
Distributions

Pairs Plot

`GGPairs` runs slowly because of the amount of data involved. In addition, this graphic ends up taking a huge amount of space in the final PDF. WE reduce plot complexity by plotting only a a 5% sample of the data.

```
the_data %>%
  select( "LIFEEXP", "LOWINCPCT", "LESSHSPCT", "LINGISOPCT","UNEMPPCT" ) %>%
  slice_sample(prop = 0.05, replace = FALSE) %>%
  ggpairs(progress = FALSE)
#> Warning: Removed 279 rows containing non-finite values (stat_density).
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 279 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 279 rows containing missing values
```

```
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 279 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 279 rows containing missing values
#> Warning: Removed 279 rows containing missing values (geom_point).
#> Removed 279 rows containing missing values (geom_point).
#> Removed 279 rows containing missing values (geom_point).
#> Removed 279 rows containing missing values (geom_point).
```



Data (except life expectancy) is not normally distributed, especially for those sub-indexes that have mostly low values. That is not unexpected for percents. which are bounded below by zero, and are a transformation of count data.

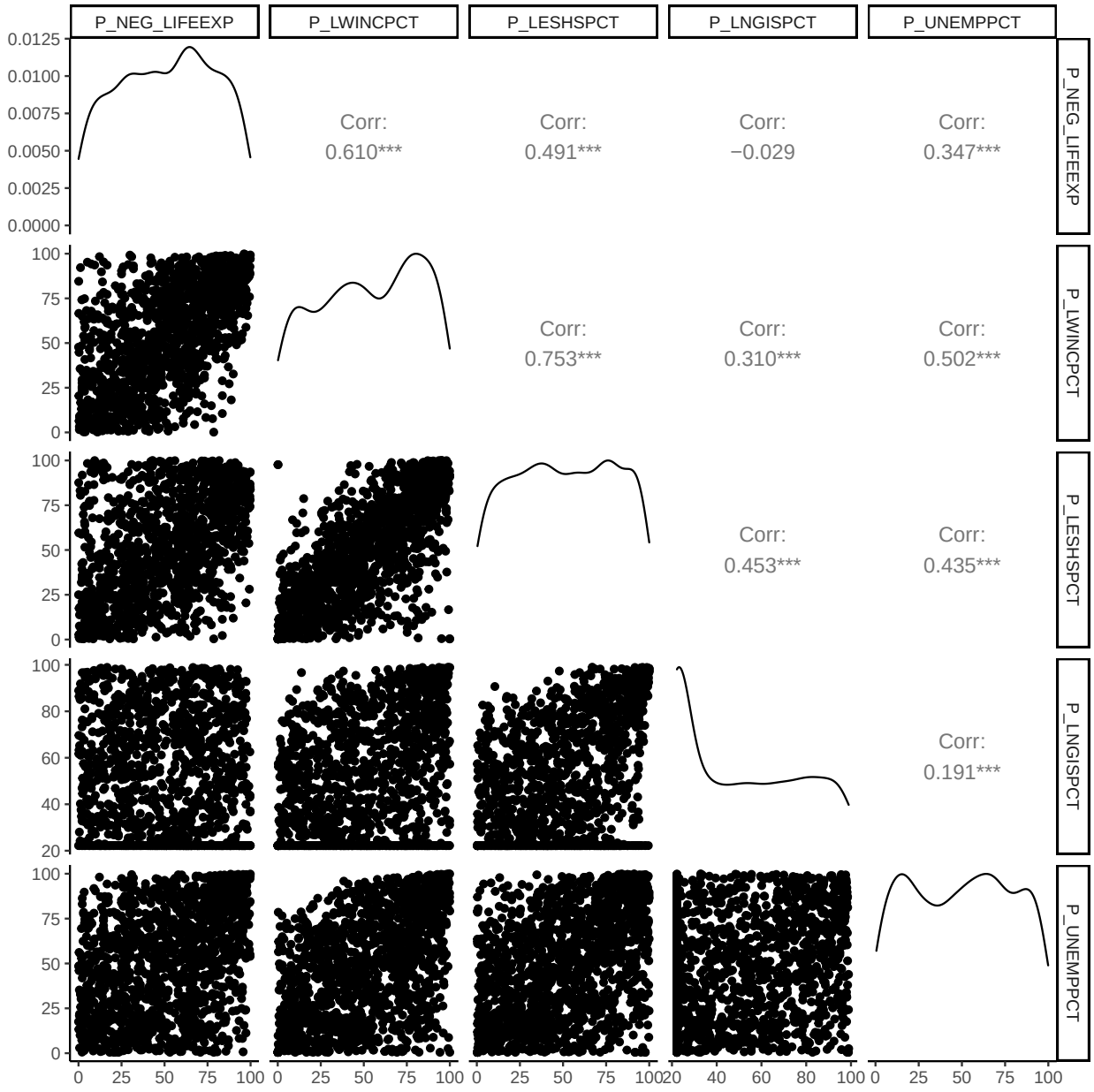
Adding or averaging raw values will lead to indexes dominated by the sub-indexes with the largest variance. For some analyses, I would consider data transformations, but that will not be needed here, since I will work with percentiles instead.

Means, SD, Medians and IGR

```
the_data %>%
  select( "LIFEEXP", "LOWINCPCT", "LESSHSPCT", "LINGISOPCT","UNEMPPCT" ) %>%
  map(quick_sum) %>%
  unlist() %>%
  array(dim = c(4,5),
        dimnames = list(c('Mean', 'SD', 'Median', 'IQR'),
                        c("NEG_LIFEEXP", "LOWINCPCT", "LESSHSPCT",
                          "LINGISOPCT","UNEMPPCT" )))
#>        NEG_LIFEEXP LOWINCPCT LESSHSPCT LINGISOPCT UNEMPPCT
#> Mean     78.309429  32.32590 12.536614   4.518998 5.748141
#> SD        3.990349  18.43269 10.366218   7.567037 4.434925
#> Median   78.500000  29.85428  9.667805   1.565558 4.714600
#> IQR       5.200000  26.69920 11.893758   5.258956 4.316772
```

Simply adding these indexes together will, roughly speaking, end up with an index that will emphasize poverty about twice as much as the lack of high school education and about four times as much as the other indicators. Moderate to high correlations among predictors will affect that somewhat, but the general idea is sound.

```
the_data %>%
  select( "P_NEG_LIFEEXP", "P_LWINCPCT", "P_LESHSPCT", "P_LNGISPCT","P_UNEMPPCT" ) %>%
  slice_sample(prop = 0.02, replace = FALSE) %>%
  ggpairs(progress = FALSE)
#> Warning: Removed 130 rows containing non-finite values (stat_density).
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 130 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 130 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 130 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 130 rows containing missing values
#> Warning: Removed 130 rows containing missing values (geom_point).
#> Removed 130 rows containing missing values (geom_point).
#> Removed 130 rows containing missing values (geom_point).
#> Removed 130 rows containing missing values (geom_point).
```



PCA Analysis of Sub-indexes

We can get more formal about the relationships between the different sub-indexes by calculating a Principal Components Analysis. The first PCA axis shows the "best fit" line through the multi-dimensional cloud of points defined by the set of sub-indexes. The second PCA axis defines the "best fit" line through the remaining variation, and so on.

The first PCA axis can be thought of as a linear combination of the sub-indexes. The average of the sub-indexes (as suggested in the funding memo) is another linear combination of the sub-indexes. In a specific sense, the first PCA axis is the optimal linear combination of the sub-indexes for summarizing the multidimensional data with just a single value.

Function for Plotting PCAs

I encapsulate the logic of plotting the PCA results just to simplify later code

```
plot_pca<- function(.pca, .scale = 2.5, .ann_space = 0.15,
                    .levels = c('NEG_LIFEEXP', 'LOWINCPCT', 'LESSHSPCT',
                                'LINGISOPCT', 'UNEMPPCT'),
                    .labels = c('Short Life', 'Income', 'School',
                                'Language', 'Unempl'),
                    .title = 'Principal Components Analysis')  {
  # .scale:  how much to expand the arrows to make them fit well against the plot
  # .ann_space: How far past the end of the arrow to place annotations

  # Gather the first two PCA axes as unit length vectors
  arrows <- as_tibble(.pca$rotation[,1:2]) %>%
     rename(PC1_Raw = PC1,
            PC2_Raw = PC2)

  # Scale length of each vector according to the standard deviations of
  # the relevant principal components (actually the square root of the
  # eigenvalues of the covariance matrix).

  scaled_arrows <-  pca$rotation %*% diag(pca$sdev) * .scale

  # Build the tibble containing  data to plot
  scaled_arrows <- as_tibble(scaled_arrows,
                             rownames = 'Variable',
                             .name_repair = ~paste0('PC', 1:5)) %>%
    select(Variable, PC1, PC2) %>%
    bind_cols(arrows) %>%
    mutate(ann1 = PC1 + .scale * .ann_space * PC1_Raw,
           ann2 = PC2 + .scale * .ann_space * PC2_Raw) %>%
    select(-PC1_Raw, -PC2_Raw) %>%
    mutate(Variable = factor(Variable,
                             levels = .levels,
                             labels = .labels ))

  plt <- ggplot(as_tibble(pca$x), aes(PC1, PC2)) +
  geom_point(alpha = 0.1, color = 'grey15') +
  #geom_density_2d(color = 'grey65') +
  geom_segment(data = scaled_arrows,
               mapping = aes(x = 0, y = 0, xend = PC1, yend = PC2),
               arrow = arrow(length = unit(0.25, 'cm'), type = 'open'),
               color = 'grey85') +
  geom_text(data = scaled_arrows,
            mapping = aes(x = ann1, y = ann2, label = Variable),
            color = 'grey85', size = 2.5) +
    ggtitle(.title) +
  theme_dark()

  return(plt)
}
```

Function to Calculate First PCA Axis Scores

I calculate scores anew to avoid alignment problems caused by missing data.

```
calc_scores <- function(.dat, .pca) {
  names <- rownames(.pca$rotation)
  vals <- map(names, ~.dat[,.])
  vals <- do.call(cbind, vals)  # converts list of vectors to data frame
  vals <- as.matrix(vals)

  mults <- matrix(.pca$rotation[,1], nrow = length(.pca$rotation[,1]))
  print(mults)
  res <- vals %*% mults
  return(as.vector(res))
}
```

Raw Values, Unscaled

First, I run a PCA on unscaled values. This highlights the fact that the "optimal" linear combination will depend on the standard errors of the sub-indexes. That means the PCA (like the simple average proposed in the memo) will depend on the units used to express each of the sub-indexes. In this context, that is probably not ideal. IF we are trying to identify disadvantaged communities, it should not matter whether life expectancy is measured in years of months, or whether the unemployment rate is expressed as a percent or a proportion.

```
pca <- the_data %>%
  select( "NEG_LIFEEXP", "LOWINCPCT", "LESSHSPCT", "LINGISOPCT","UNEMPPCT" ) %>%
  filter(complete.cases(.)) %>%
prcomp(scale. = FALSE)
```

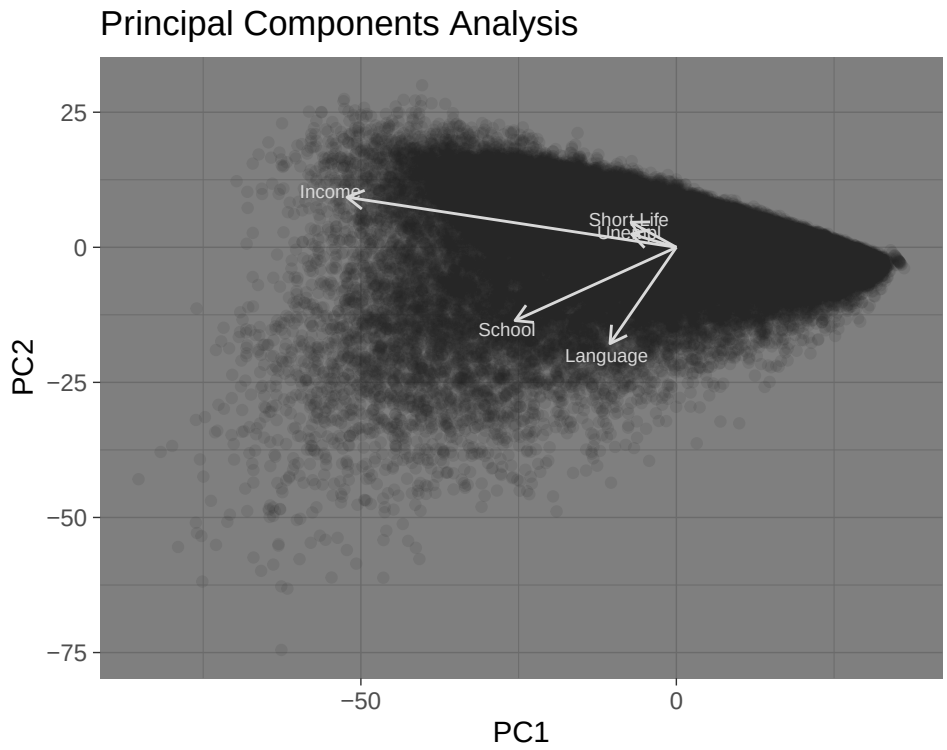
```
summary(pca)
#> Importance of components:
#>                           PC1    PC2     PC3    PC4     PC5
#> Standard deviation     19.9724 8.2667 4.71309 3.3827 2.77545
#> Proportion of Variance  0.7843 0.1344 0.04368 0.0225 0.01515
#> Cumulative Proportion   0.7843 0.9187 0.96235 0.9849 1.00000
```

The first two axes account for 92% of the variation in the sub-indexes.

```
pca$rotation
#>                    PC1         PC2         PC3         PC4         PC5
#> NEG_LIFEEXP -0.1195320  0.18648932 -0.18688676  0.06298960 -0.95500759
#> LOWINCPCT   -0.8711141  0.37363838  0.25792840 -0.14178350  0.12216788
#> LESSHSPCT   -0.4266060 -0.54701100 -0.71387546 -0.04748759  0.08314488
#> LINGISOPCT  -0.1755003 -0.71876600  0.62101006  0.11357685 -0.23242610
#> UNEMPPCT    -0.1186565  0.09884582 -0.05722422  0.98019130  0.11000251
```

The first axis is closely associated with percent low income people in each census Tract. That is because the standard deviation of the income indicator is about double the standard error of the second most variable indicator.

```
plot_pca(pca, .scale = 3, .ann_space = 1)
```



What we see is that the principal structure in the sub-indexes (when unscaled) is correlated with income and somewhat correlated with education. Linguistic isolation and education provide a fair amount of independent information via the second PCA Axis.

Scaled PCA

A scaled PCA first standardizes all variables to unit variance before conducting the PCA, thus making sure that changing units won't change results.

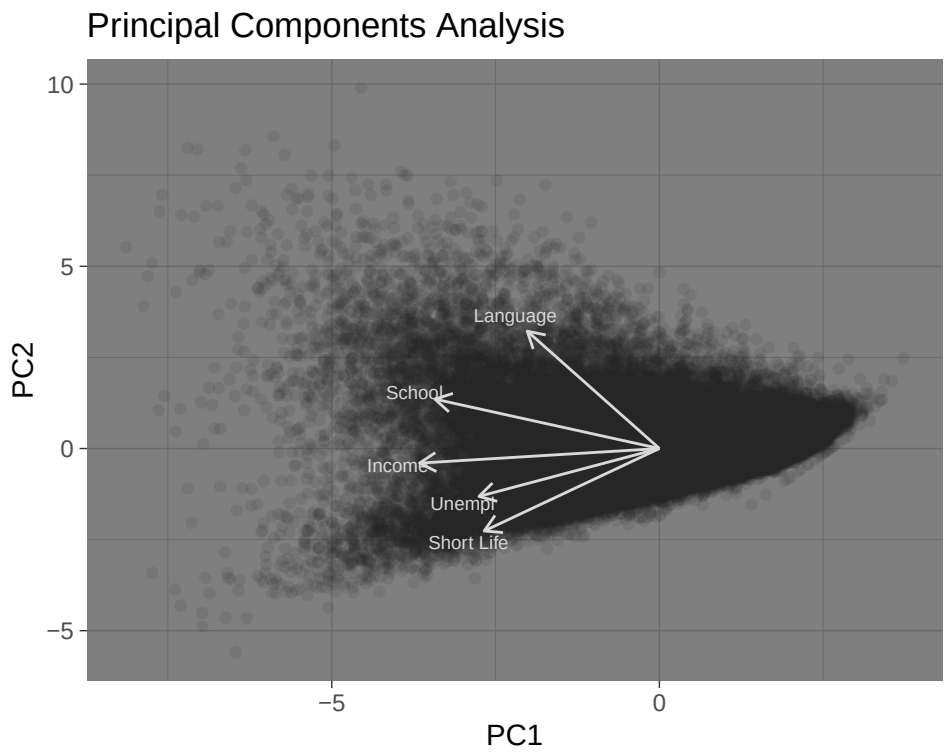
```
pca <- the_data %>%
  select( "NEG_LIFEEXP", "LOWINCPCT", "LESSHSPCT", "LINGISOPCT","UNEMPPCT" ) %>%
  filter(complete.cases(.)) %>%
prcomp(scale. = TRUE)
```

```
summary(pca)
#> Importance of components:
#>                           PC1    PC2    PC3    PC4     PC5
#> Standard deviation     1.6569 1.0958 0.7652 0.5201 0.44476
#> Proportion of Variance 0.5491 0.2402 0.1171 0.0541 0.03956
#> Cumulative Proportion  0.5491 0.7892 0.9063 0.9604 1.00000
```

This explains less of the overall variation in the first two PCA axes, confirming that some of the apparent pattern in the data was driven by the different standard deviations of the predictors.

```
pca$rotation
#>                    PC1         PC2         PC3        PC4        PC5
#> NEG_LIFEEXP -0.4033722 -0.51628930  0.50813918 -0.5273230 -0.1856375
#> LOWINCPCT   -0.5523113 -0.09137916  0.12713827  0.3789454  0.7258363
#> LESSHSPCT   -0.5168331  0.30909700  0.16033977  0.4678472 -0.6267000
#> LINGISOPCT  -0.3042858  0.73460142 -0.03571013 -0.5808921  0.1704700
#> UNEMPPCT    -0.4153755 -0.29985832 -0.83585069 -0.1483719 -0.1299521
```

```
plot_pca(pca, .scale = 4)
```



When variables are standardized to unit variance, the dominant axis is moderately correlated with all the sub-indexes, especially income and education. That suggests a common structure of community vulnerability. Language and to a lesser extent life expectancy are most heavily loaded on the second PCA axis.

This suggests that an index that is NOT based on scaled values of percentiles will largely function as a surrogate for income, while if the index is based on scaled values or percentiles, the index will reflect the effects of several different sources of disadvantage.

```
the_data$PCA_Index_V1 <- calc_scores(the_data, pca)
#>             [,1]
#> [1,] -0.4033722
#> [2,] -0.5523113
#> [3,] -0.5168331
#> [4,] -0.3042858
#> [5,] -0.4153755
```

Percentiles

```
pca <- the_data %>%
  select( "P_NEG_LIFEEXP", "P_LWINCPCT", "P_LESHSPCT",
          "P_LNGISPCT","P_UNEMPPCT" ) %>%
  filter(complete.cases(.)) %>%
prcomp(scale. = TRUE)
```

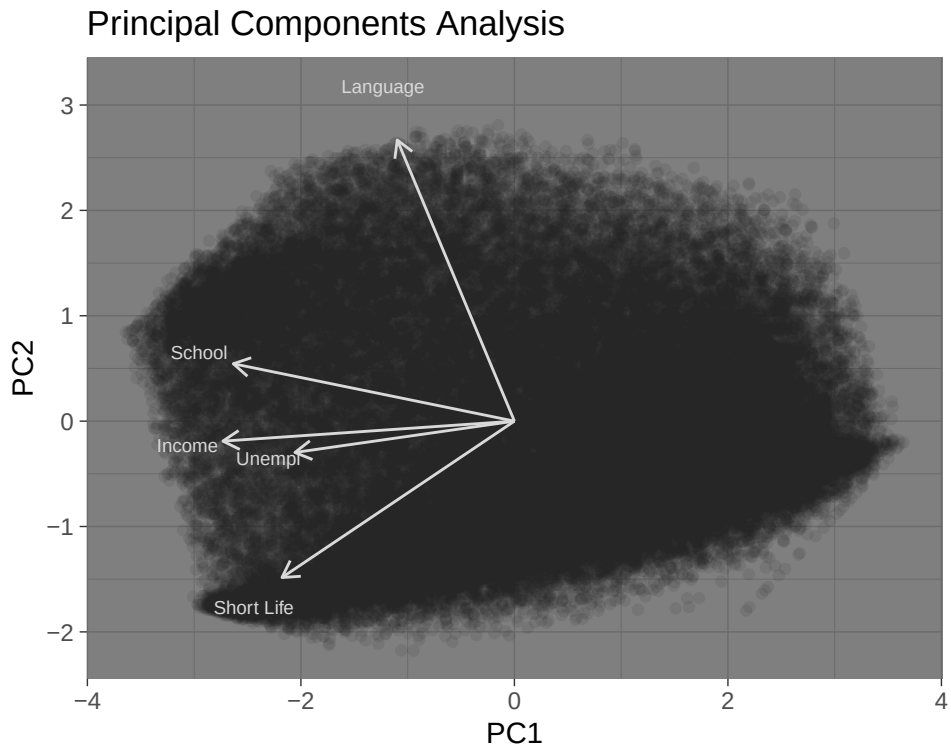
```
summary(pca)
#> Importance of components:
#>                           PC1    PC2    PC3     PC4     PC5
#> Standard deviation     1.6522 1.0405 0.8007 0.58648 0.45021
#> Proportion of Variance 0.5459 0.2165 0.1282 0.06879 0.04054
#> Cumulative Proportion  0.5459 0.7624 0.8907 0.95946 1.00000
```

The first two axes account for only 76% of the pattern in the sub-indexes.

```
pca$rotation
#>                       PC1         PC2         PC3         PC4         PC5
#> P_NEG_LIFEEXP -0.4393369 -0.47557195  0.32673090  0.67180972  0.15077470
#> P_LWINCPCT    -0.5510404 -0.06080545  0.16915414 -0.31720873 -0.75061487
#> P_LESHSPCT    -0.5312938  0.17452197  0.23661569 -0.48092027  0.63245366
#> P_LNGISPCT    -0.2214138  0.85474676  0.04536588  0.45846099 -0.09021844
#> P_UNEMPPCT    -0.4147774 -0.09531043 -0.89810286  0.08111434  0.07554676
```

Here, linguistics plays little role in Axis 1, but it dominates Axis 2. It is interesting that pecentile of (negative) life expectancy decreases the axis 2 score, while linguistic isolation sharply increases it.

```
plot_pca(pca, .scale = 3, .ann_space = .2,
         .levels = c(  "P_NEG_LIFEEXP", "P_LWINCPCT", "P_LESHSPCT",
          "P_LNGISPCT","P_UNEMPPCT" ))
```



So, an index based on the (national) percentiles of the scores produces a PCA with less structure, as expected. Here axis 1 is a composite of all the sub-indexes, with the strongest association with Schooling, Income and Unemployment. Axis 2 is principally linguistic isolation, but also has moderate loading for unemployment.

```
the_data$PCA_Index_V2 <- calc_scores(the_data, pca)
#>             [,1]
#> [1,] -0.4393369
#> [2,] -0.5510404
#> [3,] -0.5312938
#> [4,] -0.2214138
#> [5,] -0.4147774
```

Examining Results

Correlations

The Raw indexes are highly correlated with each of the sub-indexes, but especially with income and education.

```
the_data %>%
select(NEG_LIFEEXP, LOWINCPCT, LESSHSPCT, LINGISOPCT, UNEMPPCT,
       Index_1, Index_2) %>%
  cor(use = 'pairwise')  %>%
  round(3)
#>              NEG_LIFEEXP LOWINCPCT LESSHSPCT LINGISOPCT UNEMPPCT Index_1 Index_2
#> NEG_LIFEEXP        1.000     0.625     0.385     -0.053    0.423   0.586   0.677
#> LOWINCPCT          0.625     1.000     0.689      0.340    0.556   0.935   0.874
```

```
#> LESSHSPCT         0.385     0.689     1.000      0.594    0.374   0.883   0.788
#> LINGISOPCT       -0.053     0.340     0.594      1.000    0.113   0.578   0.442
#> UNEMPPCT          0.423     0.556     0.374      0.113    1.000   0.600   0.648
#> Index_1           0.586     0.935     0.883      0.578    0.600   1.000   0.928
#> Index_2           0.677     0.874     0.788      0.442    0.648   0.928   1.000
```

Rank correlations are roughly scale-free, so provide a more robust alternative where some metrics (as here) are not normally distributed.

```
the_data %>%
select(NEG_LIFEEXP, LOWINCPCT, LESSHSPCT, LINGISOPCT, UNEMPPCT,
       Index_1, Index_2) %>%
  cor(method = 'spearman', use = 'pairwise') %>%
  round(3)
#>              NEG_LIFEEXP LOWINCPCT LESSHSPCT LINGISOPCT UNEMPPCT Index_1 Index_2
#> NEG_LIFEEXP        1.000     0.632     0.502     -0.077    0.382   0.642   0.695
#> LOWINCPCT          0.632     1.000     0.747      0.233    0.527   0.953   0.885
#> LESSHSPCT          0.502     0.747     1.000      0.369    0.443   0.887   0.868
#> LINGISOPCT        -0.077     0.233     0.369      1.000    0.144   0.364   0.412
#> UNEMPPCT           0.382     0.527     0.443      0.144    1.000   0.581   0.697
#> Index_1            0.642     0.953     0.887      0.364    0.581   1.000   0.966
#> Index_2            0.695     0.885     0.868      0.412    0.697   0.966   1.000
```

The composite indexes are highly correlated, as expected.

```
the_data %>%
select(c(Index_1:PCA_Index_V2)) %>%
  cor(use = 'pairwise', method = 'pearson')  %>%
  round(3)
#>              Index_1 Index_2 P_Index_1 P_Index_2 PCA_Index_V1 PCA_Index_V2
#> Index_1        1.000   0.928     0.955     0.918       -0.997       -0.928
#> Index_2        0.928   1.000     0.963     0.997       -0.929       -0.992
#> P_Index_1      0.955   0.963     1.000     0.966       -0.960       -0.971
#> P_Index_2      0.918   0.997     0.966     1.000       -0.920       -0.991
#> PCA_Index_V1  -0.997  -0.929    -0.960    -0.920        1.000        0.935
#> PCA_Index_V2  -0.928  -0.992    -0.971    -0.991        0.935        1.000
```

Note that the PCA Index scores are **negatively** correlated with the other metrics. PCA, like most ordinations, is defined only to reflections. We swap the sign here.

```
the_data$PCA_Index_V1 <- -the_data$PCA_Index_V1
the_data$PCA_Index_V2 <- -the_data$PCA_Index_V2
```

Rank correlations are even higher.

```
the_data %>%
select(c(Index_1, Index_2, P_Index_1, P_Index_2,  PCA_Index_V1, PCA_Index_V2)) %>%
  cor(use = 'pairwise', method = 'spearman')  %>%
  round(3)
#>              Index_1 Index_2 P_Index_1 P_Index_2 PCA_Index_V1 PCA_Index_V2
#> Index_1        1.000   0.966     1.000     0.966        0.998        0.973
#> Index_2        0.966   1.000     0.966     1.000        0.958        0.993
```

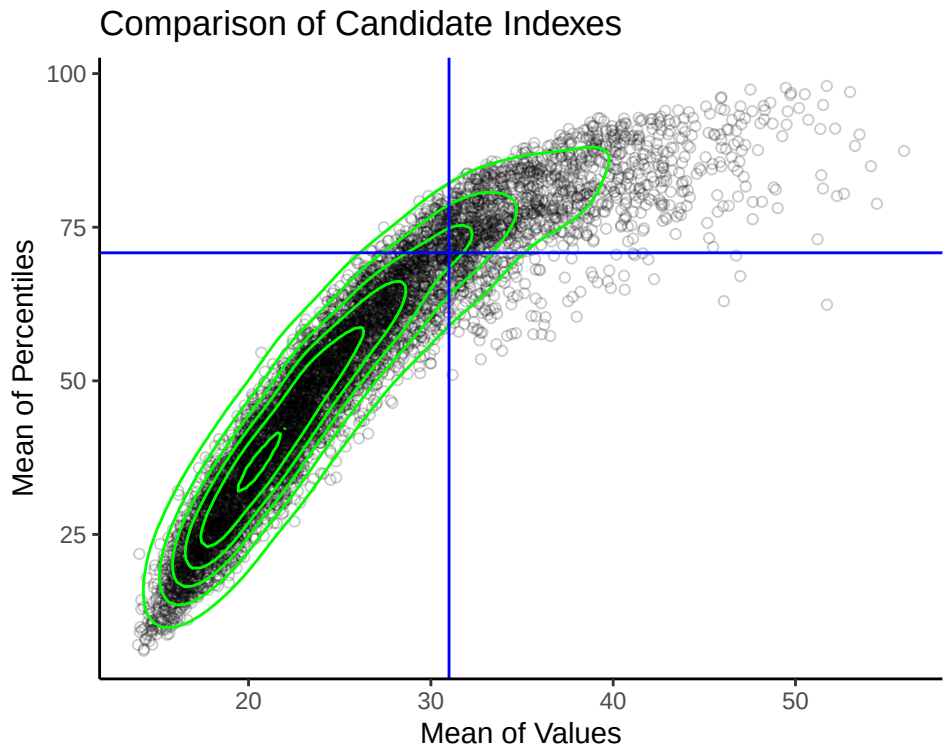
```
#> P_Index_1      1.000   0.966     1.000     0.966        0.998        0.973
#> P_Index_2      0.966   1.000     0.966     1.000        0.958        0.993
#> PCA_Index_V1   0.998   0.958     0.998     0.958        1.000        0.971
#> PCA_Index_V2   0.973   0.993     0.973     0.993        0.971        1.000
```

So, the raw index and the PCA indexes based on the same measurements are highly correlated, as one might expect. one is a simple average, the other a weighted average of the same five basic metrics.

Graphics

I plot only 5% of the data, to reduce the size of the PDF file....

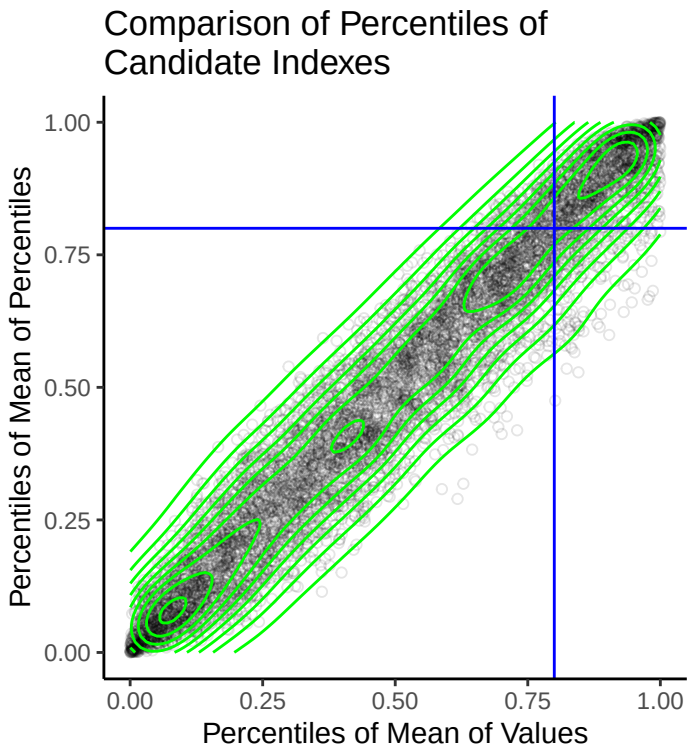
```
p80_sum <- quantile(the_data$Index_1, 0.8, na.rm = TRUE)
p80_sum_of_p  <- quantile(the_data$Index_2, 0.8, na.rm = TRUE)

plt <- the_data %>%
  slice_sample(prop = 0.1) %>%
  ggplot(aes(Index_1, Index_2)) +
  geom_point(alpha = 0.2, shape = 21) +
  geom_density_2d(color = 'green') +
  geom_vline(xintercept = p80_sum, color = 'blue') +
  geom_hline(yintercept = p80_sum_of_p, color = 'blue') +

  xlab( 'Mean of Values') +
  ylab( 'Mean of Percentiles') +
  ggtitle('Comparison of Candidate Indexes')
plt
#> Warning: Removed 620 rows containing non-finite values (stat_density2d).
#> Warning: Removed 620 rows containing missing values (geom_point).
```



The relationship between the indexes is not linear, but correlations are likely to be high over any finite range. Unfortunately, the correlations appear less robust at higher index values, exactly where we may want the most precision. The Blue lines represent the 80th percentiles in each axis. Note that the 80th percentile of the mean of five percentiles lies well below 80.

The relationship between the percentiles are close to linear, but each index is much closer – although not identical.

```
p80_sum <- quantile(the_data$P_Index_1, 0.8, na.rm = TRUE)
p80_sum_of_p  <- quantile(the_data$P_Index_2, 0.8, na.rm = TRUE)

plt <- the_data %>%
  slice_sample(prop = 0.1) %>%
  ggplot(aes(P_Index_1, P_Index_2)) +
  geom_point(alpha = 0.1, shape = 21) +
  geom_density_2d(color = 'green') +
  geom_vline(xintercept = p80_sum, color = 'blue') +
  geom_hline(yintercept = p80_sum_of_p, color = 'blue') +
  xlab( 'Percentiles of Mean of Values') +
  ylab( 'Percentiles of Mean of Percentiles') +
  ggtitle('Comparison of Percentiles of\nCandidate Indexes') +
  coord_fixed()
plt
#> Warning: Removed 564 rows containing non-finite values (stat_density2d).
#> Warning: Removed 564 rows containing missing values (geom_point).
```



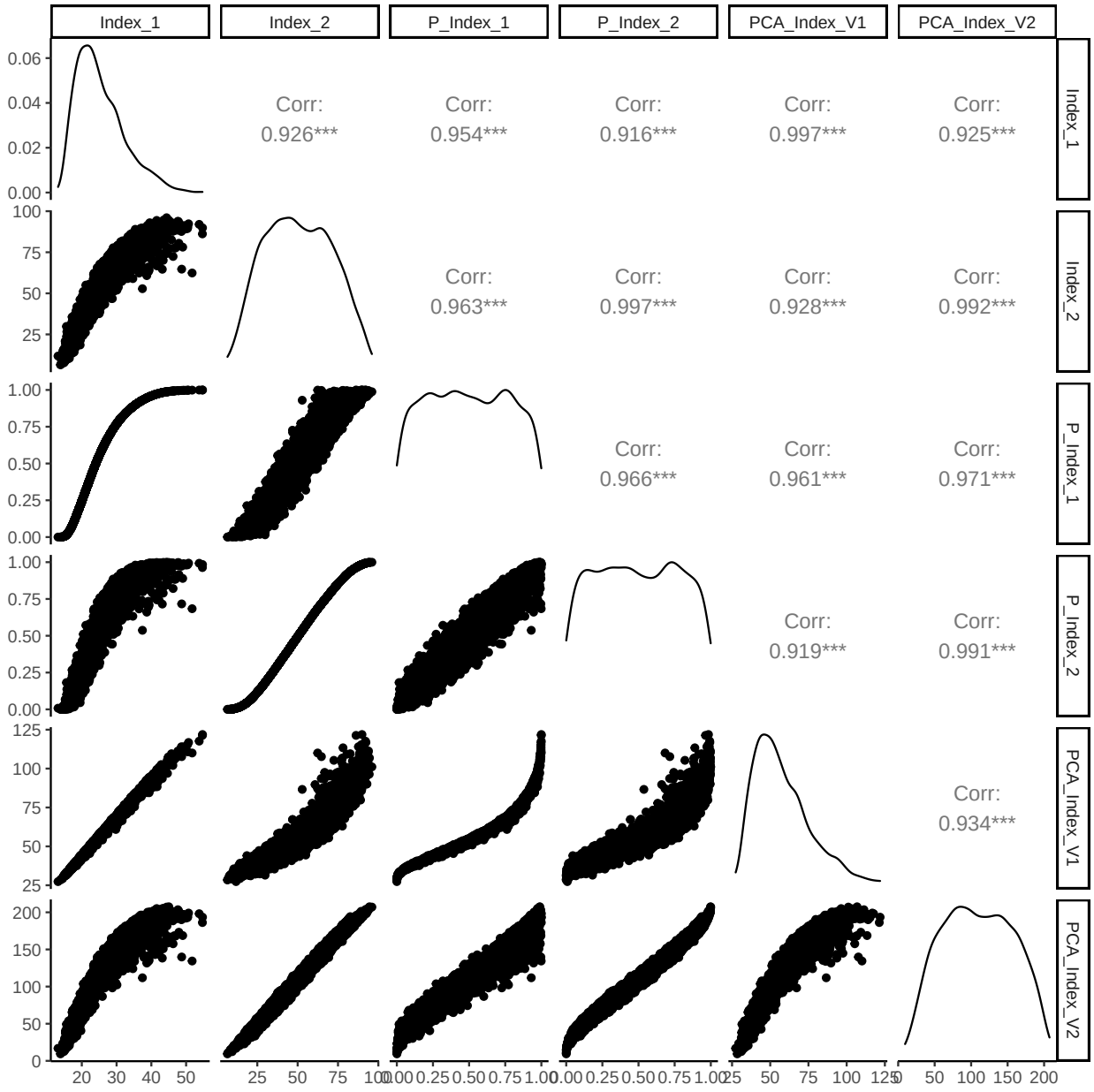
Here the percentiles DO fall

Pairs Plot of All Indexes

Again, I reduce plot complexity by plotting only 5% of the data.

```
the_data %>%
select(c(Index_1:PCA_Index_V2)) %>%
  slice_sample(prop = 0.05, replace = FALSE) %>%
  ggpairs(progress = FALSE)
#> Warning: Removed 318 rows containing non-finite values (stat_density).
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 318 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 318 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 318 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 318 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 318 rows containing missing values
#> Warning: Removed 318 rows containing missing values (geom_point).
#> Warning: Removed 318 rows containing non-finite values (stat_density).
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
```

```
#> Removed 318 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 318 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 318 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 318 rows containing missing values
#> Warning: Removed 318 rows containing missing values (geom_point).
#> Removed 318 rows containing missing values (geom_point).
#> Warning: Removed 318 rows containing non-finite values (stat_density).
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 318 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 318 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 318 rows containing missing values
#> Warning: Removed 318 rows containing missing values (geom_point).
#> Removed 318 rows containing missing values (geom_point).
#> Removed 318 rows containing missing values (geom_point).
#> Warning: Removed 318 rows containing non-finite values (stat_density).
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 318 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 318 rows containing missing values
#> Warning: Removed 318 rows containing missing values (geom_point).
#> Removed 318 rows containing missing values (geom_point).
#> Removed 318 rows containing missing values (geom_point).
#> Removed 318 rows containing missing values (geom_point).
#> Warning: Removed 318 rows containing non-finite values (stat_density).
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 318 rows containing missing values
#> Warning: Removed 318 rows containing missing values (geom_point).
#> Removed 318 rows containing missing values (geom_point).
#> Removed 318 rows containing missing values (geom_point).
#> Removed 318 rows containing missing values (geom_point).
#> Removed 318 rows containing missing values (geom_point).
#> Warning: Removed 318 rows containing non-finite values (stat_density).
```



A few observations:

- Taking percentiles has the effect of spreading out the tails of the distributions.
- The two PCA indexes are highly correlated with their
- Generally, correlations are highest within "Group 1" (raw data) or "Group 2" (percentiles).
- Correlations are high enough between the PCA-based nd simple average based index so as to make little practical difference. There is probably no reason to continue to consider the PCA based metrics (and if there were, I'd want to explore data transformations).

Output Results

```
write_csv(the_data, 'National_Draft_Indexes.csv')
```