Exam-01

Pablo Vivas

2023-10-30

# Question 1

## Pre-work

First, we read the data and take a look at how does it look.

world = read\_csv("data/world-3.csv")  
  
world |>   
 sample\_n(10) |>   
 arrange(country) |>   
 gt() |>   
 fmt\_number(  
 columns = c(6, 10, 11, 12),  
 decimals = 2  
 )

| country | region | lifeexpf | lifeexpm | literacy | popincr | babymort | birthr | deathr | gdp | aidsr | bdratio | fertilty | literacym | literacyf |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Australia | OECD | 80 | 74 | 100 | 1.38 | 7.3 | 15 | 8 | 4.23 | 1.93 | 1.88 | 1.90 | 100 | 100 |
| Brazil | Latin America | 67 | 57 | 81 | 1.28 | 66.0 | 21 | 9 | 3.37 | 1.99 | 2.33 | 2.70 | 82 | 80 |
| China | Pacific/Asia | 69 | 67 | 78 | 1.10 | 52.0 | 21 | 7 | 2.58 | 0.32 | 3.00 | 1.84 | 87 | 68 |
| India | Pacific/Asia | 59 | 58 | 52 | 1.90 | 79.0 | 29 | 10 | 2.44 | 0.60 | 2.90 | 4.48 | 64 | 39 |
| Lithuania | East Europe | 77 | 68 | 99 | 0.30 | 17.0 | 15 | 10 | 3.83 | 0.67 | 1.50 | 2.00 | 99 | 98 |
| Nigeria | Africa | 57 | 54 | 51 | 3.10 | 75.0 | 44 | 12 | 2.45 | 0.99 | 3.67 | 6.40 | 62 | 40 |
| Pakistan | Pacific/Asia | 58 | 57 | 35 | 2.80 | 101.0 | 42 | 10 | 2.61 | 0.50 | 4.20 | 6.43 | 47 | 21 |
| South Africa | Africa | 68 | 62 | 76 | 2.60 | 47.1 | 34 | 8 | 3.50 | 1.51 | 4.25 | 4.37 | NA | NA |
| Spain | OECD | 81 | 74 | 95 | 0.25 | 6.9 | 11 | 9 | 4.12 | 2.28 | 1.22 | 1.40 | 97 | 93 |
| Venezuela | Latin America | 76 | 70 | 88 | 2.16 | 28.0 | 26 | 5 | 3.45 | 1.75 | 5.20 | 3.05 | 90 | 87 |

**?(caption)**

Then, we make a description of all the numeric values.

world |>   
 select(where(is.numeric)) |>   
 skim() |>   
 select(-1) |>   
 gt() |>   
 fmt\_number(  
 columns = 3:10,  
 decimals = 2  
 ) |>   
 cols\_label(  
 skim\_variable = "variable",  
 n\_missing = "missing",  
 numeric.mean = "mean",  
 numeric.sd = "std",  
 numeric.p0 = "p0",  
 numeric.p25 = "p25",  
 numeric.p50 = "p50",  
 numeric.p75 = "p75",  
 numeric.p100 = "p100",  
 numeric.hist = "histogram"  
 ) |>   
 cols\_align(  
 columns = -1,  
 align = "center"  
 )

| variable | missing | complete\_rate | mean | std | p0 | p25 | p50 | p75 | p100 | histogram |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| lifeexpf | 0 | 1.00 | 70.16 | 10.57 | 43.00 | 67.00 | 74.00 | 78.00 | 82.00 | ▂▂▁▅▇ |
| lifeexpm | 0 | 1.00 | 64.92 | 9.27 | 41.00 | 61.00 | 67.00 | 72.00 | 76.00 | ▂▂▂▇▇ |
| literacy | 2 | 0.98 | 78.34 | 22.88 | 18.00 | 63.00 | 88.00 | 98.00 | 100.00 | ▁▂▂▂▇ |
| popincr | 0 | 1.00 | 1.68 | 1.20 | -0.30 | 0.52 | 1.80 | 2.68 | 5.24 | ▇▅▇▃▁ |
| babymort | 0 | 1.00 | 42.31 | 38.08 | 4.00 | 9.30 | 27.70 | 63.00 | 168.00 | ▇▃▁▂▁ |
| birthr | 0 | 1.00 | 25.92 | 12.36 | 10.00 | 14.00 | 25.00 | 35.00 | 53.00 | ▇▃▃▃▂ |
| deathr | 1 | 0.99 | 9.56 | 4.25 | 2.00 | 6.85 | 9.00 | 11.00 | 24.00 | ▅▇▅▁▁ |
| gdp | 0 | 1.00 | 3.42 | 0.62 | 2.09 | 3.00 | 3.48 | 3.87 | 4.37 | ▃▅▇▇▇ |
| aidsr | 3 | 0.97 | 1.38 | 0.71 | 0.00 | 0.77 | 1.38 | 1.86 | 3.18 | ▃▇▇▃▂ |
| bdratio | 1 | 0.99 | 3.20 | 2.12 | 0.92 | 1.54 | 2.67 | 4.18 | 14.00 | ▇▃▁▁▁ |
| fertilty | 2 | 0.98 | 3.56 | 1.90 | 1.30 | 1.88 | 3.05 | 5.00 | 8.19 | ▇▅▂▃▁ |
| literacym | 24 | 0.78 | 78.73 | 20.45 | 28.00 | 63.00 | 87.00 | 96.00 | 100.00 | ▁▁▃▂▇ |
| literacyf | 24 | 0.78 | 67.26 | 28.61 | 9.00 | 45.00 | 71.00 | 93.00 | 100.00 | ▂▂▂▃▇ |

**?(caption)**

We see from **?@tbl-desc-data** that some of the variables have missing data, specifically, the literacym variable has complete rate of 78%. Also, the standard deviation from each variable is in a different scale, ranging from 0.62 units in gdp to 38.08 units in babymort. This makes sense because some variables are in a log-scale.

|  |
| --- |
| Note |
| Given this scenario, we would prefer to perform pca with the correlation matrix. |

## Real work

### a)

We perform two principal component analysis using: a) the covariance matrix and b) the correlation matrix.

# get complete cases  
world\_complete = world |>   
 drop\_na()  
  
# for the covariance matrix  
pca\_cov = world\_complete |>   
 select(-c(1,2)) |>   
 prcomp()  
  
# for the correlation matrix  
pca\_cor = world\_complete |>   
 select(-c(1,2)) |>   
 prcomp(scale = TRUE)  
  
# matrix with loadings  
loadings = bind\_rows(  
 pca\_cov |>   
 tidy(matrix = "rotation") |>   
 mutate(matrix = factor("covariance")),  
 pca\_cor |>   
 tidy(matrix = "rotation") |>   
 mutate(matrix = factor("correlation"))  
)

### b)

### c)

### d)

world\_complete = pca\_cor |>  
 augment(world\_complete)

# Question 2

# Question 3