Canonical Correlation Analysis Report for ”Life Expectancy” Dataset

S/18/827 - J Kumudumali Sandaleka

May 30, 2024

1 Introduction

Canonical Correlation Analysis (CCA) is a multivariate statistical technique that identifies and measures the associations between two sets of variables. Unlike simple correlation, which assesses the relationship between two variables, CCA simultaneously examines the relationships between two multivariate sets of variables. This analysis is particularly useful in understanding complex phenomena where multiple factors from two distinct domains interact.

In this project, my primary question driving is: ”How are various health factors associated with immunization and economic factors in determining life expectancy?” The purpose of this study is to identify the underlying correlations that might inform public health policy and resource allocation. This analysis is crucial as it provides insights into how interconnected health and economic factors impact life expectancy, guiding health interventions and policy decisions.

1.1 Dataset Description

The dataset for this analysis, sourced from the WHO’s Global Health Observatory and supplemented with economic data from the United Nations, encompasses life expectancy and critical health factors for 193 countries from 2000 to 2015. This period reflects significant health sector advancements, especially in developing nations.

The followings are the variables,

- Country : Country
- Year : Year
- Life expectancy : Life Expectancy in age
- Adult Mortality : Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- infant deaths : Number of Infant Deaths per 1000 population
- Alcohol : Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
- percentage expenditure : Expenditure on health as a percentage of Gross Domestic Product per capita(%)
- Hepatitis B : Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
- Measles : Measles - number of reported cases per 1000 population

- BMI : Average Body Mass Index of entire population
- Polio : Polio (Pol3) immunization coverage among 1-year-olds (%)
- Total expenditure : General government expenditure on health as a percentage of total government expenditure (%)
- Diphtheria : Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
- HIV/AIDS : Deaths per 1 000 live births HIV/AIDS (0-4 years)
- Schooling : Number of years of Schooling(years)

2 Methodology

The methodology section encompasses rigorous data preprocessing, canonical correlation analysis, testing for significance, ensuring a comprehensive understanding of latent factors and their relationships within the ”Life Expectancy” dataset.

2.1 Data Preprocessing

Before conducting canonical correlation analysis, the dataset was thoroughly preprocessed to ensure its integrity and suitability. This required addressing duplicate entries and missing values in addition to standardizing the data to reduce inconsistencies caused by disparate units of measurement. In this stage, it is observed that there are no outliers in this dataset.

2.2 Statistical Methods: Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a statistical method used to explore the relationships between two sets of variables. Unlike simpler correlation techniques that measure the relationship between two individual variables, CCA is designed to handle the complexity of relationships among multiple variables in two datasets. This method identifies pairs of canonical variates (linear combinations of the original variables) that maximize the correlation between the two sets.

Steps in Canonical Correlation Analysis

2.2.1 Formulating the Data Matrices:

The variables are categorized into health-related factors, economic factors, and immunization rates.

Set U (Health-related factors):

1. Life expectancy
2. Adult Mortality
3. Infant deaths
4. BMI

Set V (Immunization and Economic factors):

1. Alcohol
2. Hepatitis B
3. Measles
4. Polio
5. Diphtheria
6. Total expenditure
7. Percentage expenditure
8. Schooling

2.2.2 Canonical correlation coefficients

The canonical correlation coefficients are the values that indicate how strongly the two sets of variables are related. They range from 0 to 1, where 0 means no relationship and 1 means a perfect relationship. The first canonical correlation coefficient is the highest possible value, and the subsequent ones are lower and orthogonal to the previous ones. The number of canonical correlation coefficients is equal to the smaller number of variables in either set. Also, we can use the canonical correlation coefficients to test the null hypothesis that there is no relationship between the two sets of variables, and to compare the relative importance of each pair of canonical variables.

2.2.3 Canonical plots

The canonical plots are graphical representations of the canonical correlation analysis, where you can see the relationship between the two sets of variables and the cases. The most common types of canonical plots are biplots and scatter plots. A scatter plot shows the cases on the plot, where each axis represents a canonical variable. The shape and location of the points indicate the scores and clusters of the cases on the canonical dimensions.

2.2.4 Raw canonical coefficients / Canonical weights

The canonical weights are the coefficients that are used to create the canonical variables from the original variables. They are similar to the regression coefficients in multiple regression, and they indicate how much each variable contributes to the formation of the canonical variable. The canonical weights can help us determine which variables are most important or influential in creating the canonical variables, and how they affect the canonical correlation.

2.2.5 Canonical loadings

Canonical loadings are the correlations between the original variables and the canonical variables, which are linear combinations that maximize the canonical correlation coefficients. They help interpret the meaning and direction of the canonical variables and their relationships. A high positive loading indicates that a variable contributes positively to the canonical correlation and is positively associated with other variables with high positive loadings on the same canonical variable.

2.2.6 Canonical cross-loadings

Canonical cross-loadings measure the correlations between the original variables and the canonical variables of the other set. They help assess the overlap or redundancy between the two sets of variables and their mutual influence. A high positive cross-loading indicates a strong correlation with the canonical variable of the other set, sharing common variance with variables that have high loadings on that canonical variable. Canonical cross-loadings also help compute canonical communality, which quantifies the shared variance between the two sets of variables.

3 Results and Discussion

3.1 Testing for Significance

To test for independence between the health-related factors and the immunization and economic factors,

```
Wilks' Lambda, using F-approximation (Rao's F):
stat    approx df1      df2      p.value
1 to 4:  0.2669098 81.275272  32 6038.553 0.000000e+00
2 to 4:  0.6810853 32.056865  21 4704.002 0.000000e+00
3 to 4:  0.9546040  6.419760  12 3278.000 2.120193e-11
4 to 4:  0.9920124  2.641023   5 1640.000 2.188721e-02


 Hotelling-Lawley Trace, using F-approximation:
stat     approx df1  df2      p.value
1 to 4:  2.00057543 102.248160  32 6542 0.000000e+00
2 to 4:  0.44883164  34.998182  21 6550 0.000000e+00
3 to 4:  0.04723929   6.454068  12 6558 1.500378e-11
4 to 4:  0.00805190   2.643439   5 6566 2.151494e-02


Pillai-Bartlett Trace, using F-approximation:
stat    approx df1  df2      p.value
1 to 4:  0.940334145 63.003122  32 6560 0.000000e+00
2 to 4:  0.332223019 28.329614  21 6568 0.000000e+00
3 to 4:  0.045697231  6.332869  12 6576 2.809553e-11
4 to 4:  0.007987585  2.634774   5 6584 2.189202e-02


 Roy's Largest Root, using F-approximation:
stat   approx df1  df2 p.value
1 to 1:  0.6081111 318.1075   8 1640       0

F statistic for Roy's Greatest Root is an upper bound.
```

$$H_0 : \rho_i = 0 \; vs \; H_1 : at\ least\ one\ \rho_i \neq 0 \; for \; i = 1, 2, 3, 4 \tag{1}$$

All multivariate tests, including Wilks' Lambda, Hotelling-Lawley Trace, Pillai-Bartlett Trace, and Roy's

Largest Root, indicate highly significant canonical correlations with p-values well below 0.0001. These consistently low p-values confirm the statistical significance of the canonical correlations across all canonical functions.

Given that canonical correlations are ordered from largest to smallest, the results suggest that the first canonical correlation (ρ_1) is particularly noteworthy and significantly different from zero. This indicates a strong relationship between the first pair of canonical variates from the two sets of variables (U and V).

$$H_0 : \rho_i = 0 \; vs \; H_1 : at \; least \; one \; \rho_i \neq 0 \; for \; i = 2, 3, 4 \tag{2}$$

The second canonical correlation, though smaller than the first, remains statistically significant across all multivariate tests ($p < 0.0001$). This suggests that even after accounting for the first canonical variate, there is still a meaningful relationship between the remaining variance in the U and V sets. The second canonical function explains a notable, though lesser, portion of the shared variance.

$$H_0 : \rho_i = 0 \; vs \; H_1 : at \; least \; one \; \rho_i \neq 0 \; for \; i = 3, 4 \tag{3}$$

The third canonical correlation also shows statistical significance ($p < 0.0001$), indicating that the third pair of canonical variates still captures a significant, albeit smaller, relationship between the two sets of variables. This suggests that there is additional, but reducing, shared variance explained by the third canonical function.

$$H_0 : \rho_4 = 0 \; vs \; H_1 : \rho_4 \neq 0 \tag{4}$$

The fourth canonical correlation is statistically significant ($p < 0.0001$), though it is the smallest among the four. This implies that while the fourth pair of canonical variates accounts for the least amount of shared variance, it still represents a significant, but minimal, relationship between the sets of variables. The reducing size of the canonical correlations indicates that each successive function explains less of the remaining variance.

In conclusion, all multivariate tests (Wilks' Lambda, Hotelling-Lawley Trace, Pillai-Bartlett Trace, and Roy's Largest Root) indicate significant canonical correlations. The first canonical function consistently shows the strongest relationship across all tests. These results suggest that the canonical variates derived from the two sets of variables (U and V) have significant and meaningful relationships.

3.2 Canonical correlation

The squared values of the canonical variate pairs can be interpreted similarly to how r^2 values are interpreted in regression analysis.

```
Canonical.Correlation      Squared.Canonical.Correlation

0.77981480                  0.608111126
0.53528104                  0.286525788
0.19418972                  0.037709646
0.08937329                  0.007987585
```

From the results, we observe that 60.81% of the variation in U_1 is explained by the variation in V_1, and 28.65% of the variation in U_2 is explained by V_2. However, only 3.77% of the variation in U_3 is explained by V_3, and the fourth squared canonical correlation is extremely low at 0.80%. This indicates that the first canonical correlation is very high and suggests that it is the most significant. The subsequent canonical correlations drop off sharply, highlighting that only the first canonical correlation is truly important for understanding the relationship between the health-related factors and immunization and economic factors.

3.3 Canonical correlation plot

The first canonical variate for health-related factors is plotted against the first canonical variate for immunization and economic factors in the scatter plot for the first canonical variate pair:

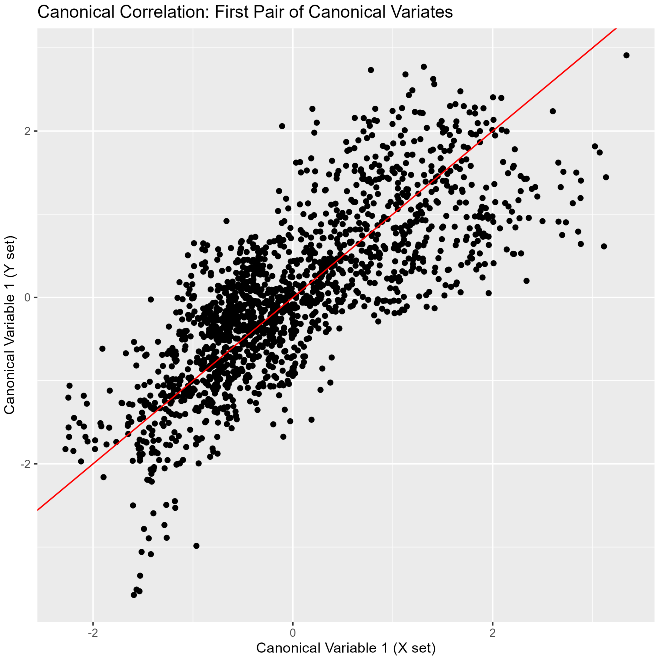


Figure 1: Canonical Correlation: First Pair of Canonical Variates

The regression line shows how well the data fits. The plot is a bit more scattered, but is still a reasonably good fit.

3.4 Raw canonical coefficients

The raw canonical coefficients (canonical weights) are the coefficients that are used to create the canonical variables from the original variables.

```
Set U:
                  [,1]
Lifeexpectancy -0.9266450
AdultMortality -0.1923545
infantdeaths    0.1726828
BMI            -0.2315565
```

$$U_1 = -0.93X_{LifeExpectancy} - 0.19X_{AdultMortality} + 0.17X_{InfantDeaths} - 0.23X_{BMI} \tag{5}$$

A one-unit increase in life expectancy results in a decrease of 0.93 units in the value of the first canonical variate for the health-related set of variables, when other variables are held constant. Similarly, a one-unit increase in adult mortality and BMI leads to decreases of approximately 0.19 units and 0.23 units, respectively, in the first canonical variate. In contrast, a one-unit increase in infant deaths causes an increase of about 0.17 units in the value of the first canonical variate for the health-related set of variables, holding other variables constant.

```
Set V:
                             [,1]
Alcohol                0.09912126
HepatitisB            -0.03722022
Measles                0.12392999
Polio                 -0.04655309
Diphtheria            -0.07482133
Totalexpenditure      -0.01522375
percentageexpenditure -0.15428906
Schooling             -0.90043023
```

$$V_1 = 0.1Y_{Alcohol} - 0.04Y_{Hep.B} + 0.12Y_{Measles} - 0.05Y_{Polio} - 0.07Y_{Diph.} - 0.02Y_{Tot.exp} - 0.15Y_{\%exp.} - 0.9Y_{Sch} \tag{6}$$

A one-unit increase in alcohol consumption results in a 0.10 unit increase in the value of the first canonical variate for the immunization and economic factors set, when other variables are held constant. Conversely, increases in Hepatitis B immunization, polio immunization, diphtheria immunization, total expenditure, and percentage expenditure lead to decreases of approximately 0.04 units, 0.05 units, 0.07 units, 0.02 units, and 0.15 units, respectively, in the first canonical variate. A significant impact is observed with schooling, where a one-unit increase results in a substantial decrease of approximately 0.90 units in the first canonical variate for the immunization and economic factors set, holding other variables constant.

3.5 Canonical - loadings

The canonical loadings are the correlations between the original variables and the canonical variables, which are the linear combinations of variables that maximize the canonical correlation coefficients.

```
Set U:
                       [,1]
Lifeexpectancy   -0.9462209
AdultMortality    0.5473673
infantdeaths      0.3754714
BMI              -0.7066970

Set V:
[,1]
Alcohol                -0.5637043
HepatitisB             -0.3055000
Measles                 0.2462906
Polio                  -0.4296835
```

```
Diphtheria            -0.4465742
Totalexpenditure      -0.2753939
percentageexpenditure -0.5203263
Schooling             -0.9729704
```

The first canonical variate for the U set is strongly negatively correlated with life expectancy (-0.95) and BMI (-0.71), and positively correlated with adult and infant mortality rates. For the V set, it is negatively correlated with alcohol consumption, vaccination rates (Hepatitis B, Polio, Diphtheria), health expenditures, and years of schooling. The first canonical variate for the V set shows a strong negative correlation with years of schooling (-0.97). This suggests a significant relationship between the health metrics and the variables related to immunization and economic in these canonical variates.

3.6 Canonical cross-loadings

The canonical cross-loadings indicate how each original variable correlates with the canonical variate of the other set.

```
Set U:
                     [,1]
Lifeexpectancy -0.7378771
AdultMortality  0.4268451
infantdeaths    0.2927981
BMI            -0.5510928

Set V:
            [,1]
Alcohol               -0.4395850
HepatitisB            -0.2382334
Measles                0.1920611
Polio                 -0.3350736
Diphtheria            -0.3482452
Totalexpenditure      -0.2147562
percentageexpenditure -0.4057581
Schooling             -0.7587367
```

For the first set (U), Life expectancy (-0.738) and BMI (-0.551) show strong negative correlations, while Adult Mortality (0.427) and infant deaths (0.293) exhibit moderate and weak positive correlations with the canonical variate of the second set (V). For the second set (V), Schooling (-0.759) and Alcohol (-0.440) have strong and moderate negative correlations, respectively, with the first set's (U) canonical variate. Other variables such as Polio (-0.335), Diphtheria (-0.348), and percentage expenditure (-0.406) also show weak negative correlations. These cross-loadings highlight the significant inverse relationships between health metrics from the first set and immunization and economic variables from the second set.

4 Conclusion

This study applied Canonical Correlation Analysis (CCA) to explore the relationships between health-related factors and immunization and economic factors in determining life expectancy. The analysis revealed significant canonical correlations between the two sets of variables, indicating substantial relationships.

Key findings include:

1. Significance of Canonical Correlations: All multivariate tests (Wilks' Lambda, Hotelling-Lawley Trace, Pillai-Bartlett Trace, and Roy's Largest Root) indicated highly significant canonical correlations, with p-values well below 0.0001. This underscores the robust relationships between the health-related factors and immunization and economic factors.

2. Canonical Correlation Coefficients: The first canonical correlation was notably high (0.78), explaining approximately 60.81% of the variance between the sets, indicating a strong initial relationship. Subsequent canonical correlations, while still significant, explained progressively less variance, emphasizing the dominance of the first canonical variate pair.

3. Canonical Coefficients: The canonical weights and loadings provided insights into the contributions of individual variables. For example, life expectancy, BMI, and schooling had strong negative correlations with their respective canonical variates, highlighting their significant roles in the observed relationships.

4. Cross-Loadings and Shared Variance: The canonical cross-loadings illustrated how variables from one set correlated with the canonical variates of the other set, further validating the interconnections between health metrics and economic and immunization factors.

These results suggest that public health policy and resource allocation should prioritize interconnected health and economic factors to effectively enhance life expectancy. The insights gained from CCA can guide targeted interventions and strategies, ensuring a more holistic approach to improving global health outcomes.

In conclusion, CCA has proven to be a powerful tool in unravelling the complex interplay between various health-related factors and economic factors, providing a comprehensive understanding that can inform and optimize public health policies.

5 References

1. Abdolrasol, M. G. M., Hussain, S. M. S., Ustun, T. S., Sarker, M. R., Hannan, M. A., Mohamed, R., Ali, J. A., Mekhilef, S., & Milad, A. (2021). Artificial Neural Networks Based Optimization Techniques: A review. Electronics, 10(21), 2689. https://doi.org/10.3390/electronics10212689

2. Canonical Correlation Analysis — R Data Analysis Examples. (n.d.). https://stats.oarc.ucla.edu/r/dae/canonical-correlation-analysis/

3. Canonical Correlation Analysis — SPSS Data Analysis Examples. (n.d.). https://stats.oarc.ucla.edu/spss/dae/canonical-correlation-analysis/

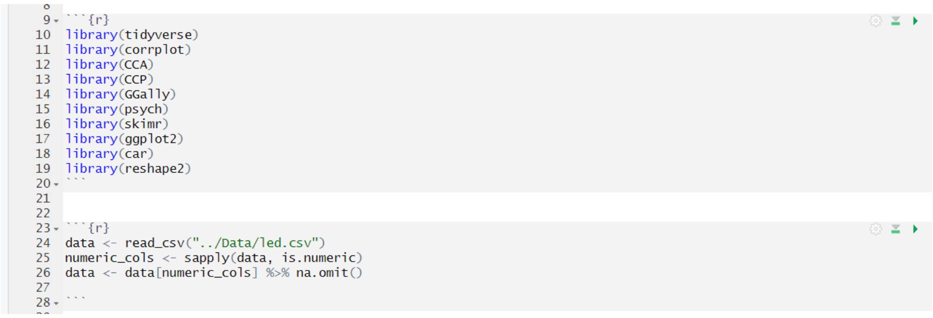
4. Canonical Correlation Analysis — STATA Data Analysis Examples. (n.d.). https://stats.oarc.ucla.edu/stata/dae/canonical-correlation-analysis/

5. Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., Black, W. C., & Prentice Hall, Inc. (1998). Canonical Correlation Analysis. In Canonical Correlation Analysis. https://mvstats.com/wp-content/uploads/2022/02/Canonical_Correlation_6e.pdf

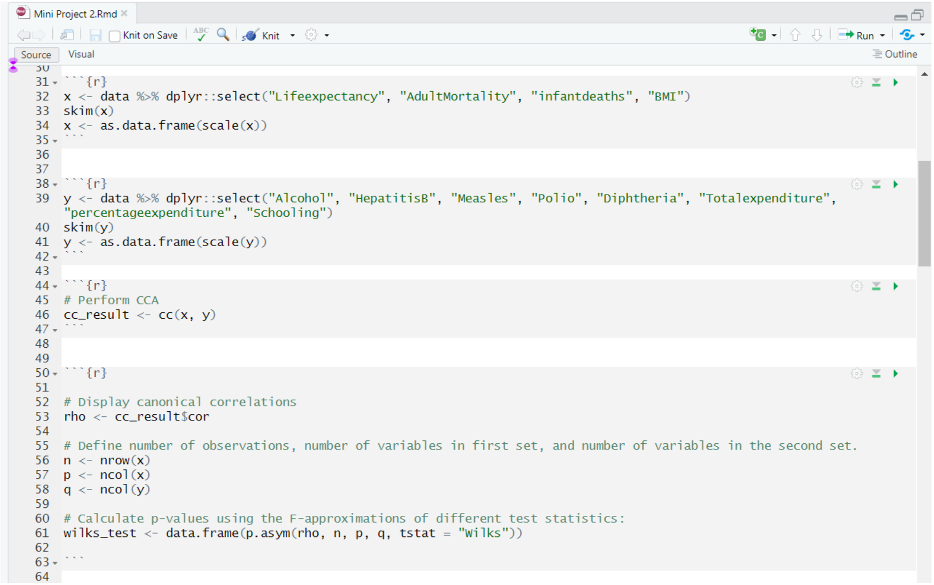
6. Haykir, M., & Çalışkan, O. (2021). Is There a Relationship between Empowering Chefs and the Culinary Creativity Process? Journal of Culinary Science & Technology, 21(3), 404–429. https://doi.org/10.1080/15428052.2021.1955793

7. Lesson 13: Canonical Correlation Analysis — STAT 505. (n.d.). PennState: Statistics Online Courses. https://online.stat.psu.edu/stat505/lesson/13

6 Appendices



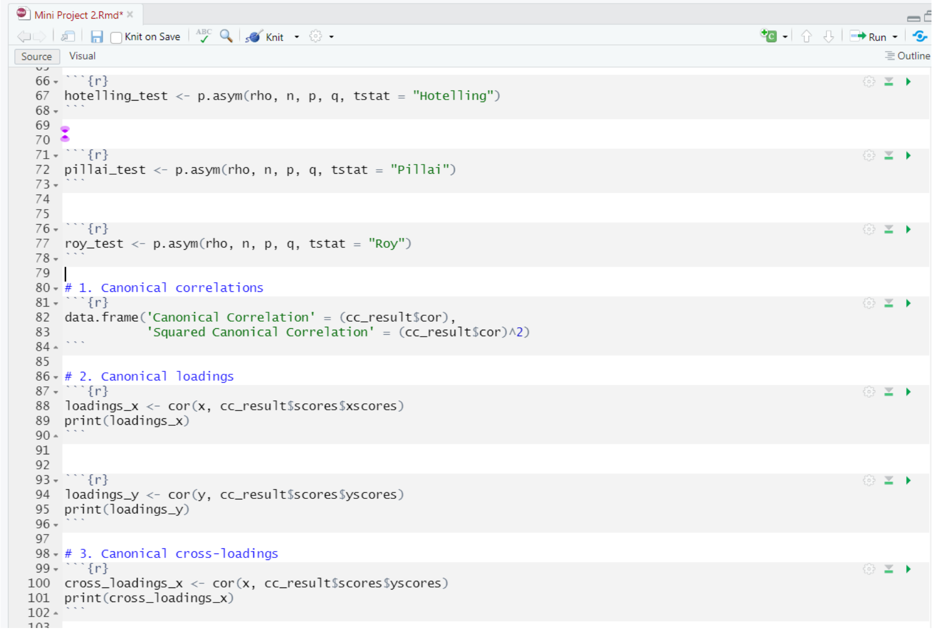
~~~
```{r}
library(tidyverse)
library(corrplot)
library(CCA)
library(CCP)
library(GGally)
library(psych)
library(skimr)
library(ggplot2)
library(car)
library(reshape2)
~~~

```{r}
data <- read_csv("../Data/led.csv")
numeric_cols <- sapply(data, is.numeric)
data <- data[numeric_cols] %>% na.omit()

```
```



Mini Project 2.Rmd
Knit on Save
Knit
Run
Source
Visual
Outline

```
```{r}
x <- data %>% dplyr::select("Lifeexpectancy", "AdultMortality", "infantdeaths", "BMI")
skim(x)
x <- as.data.frame(scale(x))
```

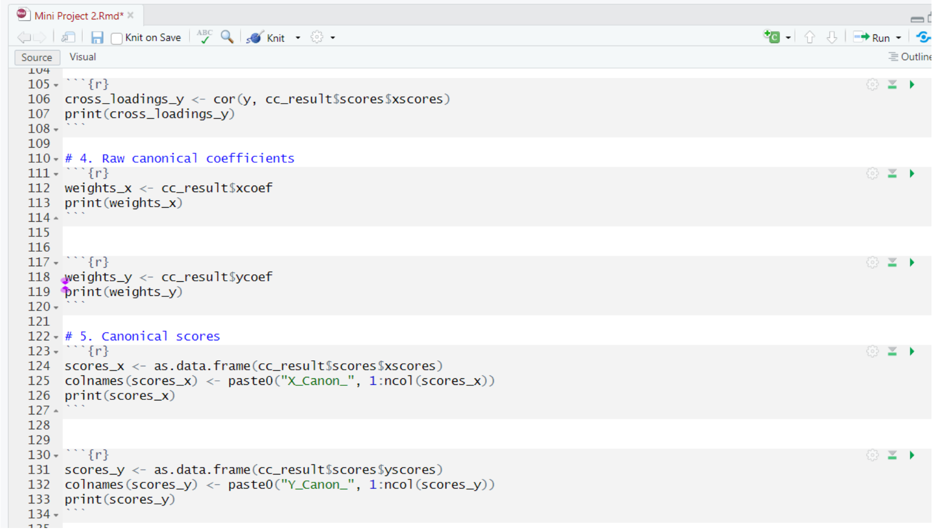
```{r}
y <- data %>% dplyr::select("Alcohol", "HepatitisB", "Measles", "Polio", "Diphtheria", "Totalexpenditure",
"percentageexpenditure", "Schooling")
skim(y)
y <- as.data.frame(scale(y))
```

```{r}
# Perform CCA
cc_result <- cc(x, y)
```

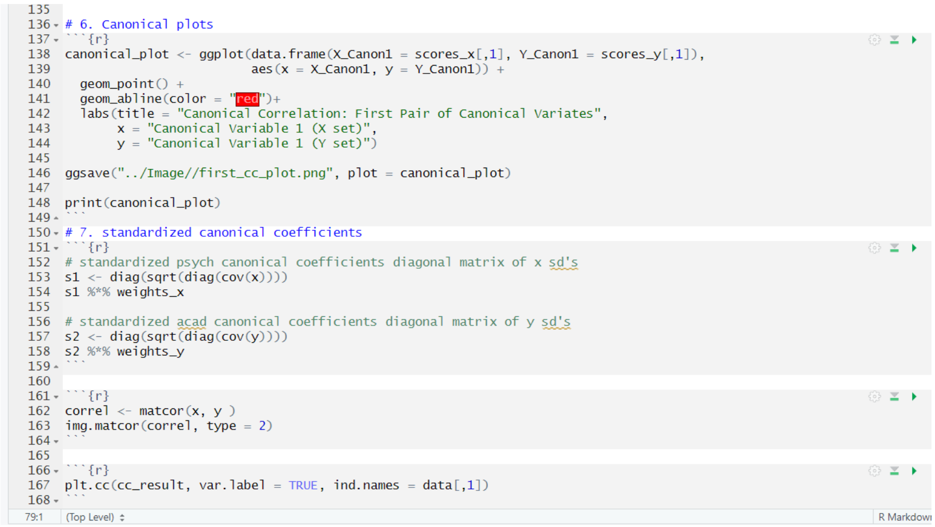
```{r}

# Display canonical correlations
rho <- cc_result$cor

# Define number of observations, number of variables in first set, and number of variables in the second set.
n <- nrow(x)
p <- ncol(x)
q <- ncol(y)

# Calculate p-values using the F-approximations of different test statistics:
wilks_test <- data.frame(p.asym(rho, n, p, q, tstat = "Wilks"))

```
```
```



~~~
```{r}
hotelling_test <- p.asym(rho, n, p, q, tstat = "Hotelling")
```



```{r}
pillai_test <- p.asym(rho, n, p, q, tstat = "Pillai")
```



```{r}
roy_test <- p.asym(rho, n, p, q, tstat = "Roy")
```

# 1. Canonical correlations
```{r}
data.frame('Canonical Correlation' = (cc_result$cor),
           'Squared Canonical Correlation' = (cc_result$cor)^2)
```

# 2. Canonical loadings
```{r}
loadings_x <- cor(x, cc_result$scores$xscores)
print(loadings_x)
```



```{r}
loadings_y <- cor(y, cc_result$scores$yscores)
print(loadings_y)
```

# 3. Canonical cross-loadings
```{r}
cross_loadings_x <- cor(x, cc_result$scores$yscores)
print(cross_loadings_x)
```
~~~



~~~
```{r}
cross_loadings_y <- cor(y, cc_result$scores$xscores)
print(cross_loadings_y)
```

# 4. Raw canonical coefficients
```{r}
weights_x <- cc_result$xcoef
print(weights_x)
```



```{r}
weights_y <- cc_result$ycoef
print(weights_y)
```

# 5. Canonical scores
```{r}
scores_x <- as.data.frame(cc_result$scores$xscores)
colnames(scores_x) <- paste0("X_Canon_", 1:ncol(scores_x))
print(scores_x)
```



```{r}
scores_y <- as.data.frame(cc_result$scores$yscores)
colnames(scores_y) <- paste0("Y_Canon_", 1:ncol(scores_y))
print(scores_y)
```
~~~



~~~

6. Canonical plots
```{r}
canonical_plot <- ggplot(data.frame(X_Canon1 = scores_x[,1], Y_Canon1 = scores_y[,1]),
                         aes(x = X_Canon1, y = Y_Canon1)) +
  geom_point() +
  geom_abline(color = "red")+
  labs(title = "Canonical Correlation: First Pair of Canonical Variates",
       x = "Canonical Variable 1 (X set)",
       y = "Canonical Variable 1 (Y set)")

ggsave("../Image//first_cc_plot.png", plot = canonical_plot)

print(canonical_plot)
```
7. standardized canonical coefficients
```{r}
# standardized psych canonical coefficients diagonal matrix of x sd's
s1 <- diag(sqrt(diag(cov(x))))
s1 %*% weights_x

# standardized acad canonical coefficients diagonal matrix of y sd's
s2 <- diag(sqrt(diag(cov(y))))
s2 %*% weights_y
```

```{r}
correl <- matcor(x, y )
img.matcor(correl, type = 2)
```

```{r}
plt.cc(cc_result, var.label = TRUE, ind.names = data[,1])
```
~~~