PAPER

MMS2432 APPLIED MULTIVARIATE STATISTICS

CANONICAL CORRELATION ANALYSIS BETWEEN THE CONDITION OF EDUCATIONAL RESOURCES AND THE ACADEMIC ACHIEVEMENTS OF HIGH SCHOOL STUDENTS IN INDONESIA (2019)

Student Name:
BRYAN FLORENTINO LEO

***Nomor Induk Mahasiswa (NIM)*:**
21/473767/PA/20429

STUDY PROGRAM OF BACHELOR IN STATISTICS
DEPARTMENT OF MATHEMATICS
FACULTY OF MATHEMATICS AND NATURAL SCIENCES
UNIVERSITAS GADJAH MADA
2022

ABSTRACT

Education is an unseen yet crucial aspect in the life of human beings. In fact, education is one of three indicators of the Human Development Index (HDI) which has been measured annually since 1990 by the United Nations Development Programme (UNDP). The trend tells a decline in Indonesia's HDI value since 2019, with the most recent value of 0.701 and mean duration of schooling of only 8.6 years. In early 2022, the prime demand proposed by Commission X of The House of Representatives of the Republic of Indonesia to the Minister of Education, Culture, Research, and Technology was to evaluate holistically and profoundly the selection of teachers from the group of State Civil Apparatuses in the preceding year, including the affirmation track for employees recruiting to Foremost, Outermost, and Underdeveloped regions. This paper is written to know whether there exists a relationship between the condition of educational resources and academic achievements of high school students in Indonesia, which are indirectly linked to the quality of education of the country. The aim is achieved by undergoing canonical correlation analysis on multiple variables which represent both halves. On this paper, the case is limited to the year 2019 or academic year 2018/2019. At the end, after going through a canonical correlation analysis, it is obtained that there is a positive yet very strong correlation of 0.990 (out of 1.00) between the condition of educational resources and the academic achievements of high school students in Indonesia in 2019.

CONTENTS

CHAPTER I
INTRODUCTION

1.1 Background

Education is an unseen yet crucial aspect in the life of human beings. In one of his speeches, the former president of South Africa, Nelson R. Mandela, quoted, “Education is the most powerful weapon we can use to change the world.” In fact, education, specifically knowledge, is one of three indicators of the Human Development Index (HDI) which has been measured annually since 1990 by the United Nations Development Programme (UNDP). In 2021, Indonesia’s human development was ranked on the 114th place from 191 countries which take part in the United Nations (UN). Even though the value of 0.705 leaded Indonesia to be considered as a nation with a high human development, data has shown that our index was still lower than the world on average. Furthermore, the trend tells a decline in Indonesia’s HDI value since 2019, with the mean duration of schooling in 2021 of only 8.6 years.

Indeed, education has been a chronic issue in our land. From minister to minister, complaints about education seemed to be loyally assigned. One of the latest news reported by jpnn.com stated that in early 2022, Commission X of The House of Representatives of the Republic of Indonesia (*Dewan Perwakilan Rakyat Republik Indonesia* or DPR RI) proposed five great demands to the Minister of Education, Culture, Research, and Technology (*Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi* or *Mendikbudristek*), Nadiem Makarim.[1] The prime demand was to evaluate holistically and profoundly the selection of teachers from the group of State Civil Apparatuses (*Aparatur Sipil Negara* or ASN) in the preceding year, including the affirmation track for employees recruiting to Foremost, Outermost, and Underdeveloped regions (*daerah Terdepan, Terluar, dan Tertinggal*, also known as 3T regions).

[1] From *Komisi X Ajukan 5 Tuntutan kepada Mas Nadiem, Ada Soal PPPK, BOS dan Direktorat Baru*, by JPNN.com. Retrieved from jpnn.com: https://www.jpnn.com/news/komisi-x-ajukan-5-tuntutan-kepada-mas-nadiem-ada-soal-pppk-bos-dan-direktorat-baru.

Regarding this case, we aim to know whether there exists a relationship between the condition of educational resources and academic achievements of high school students in Indonesia, which are indirectly linked to the quality of education of the country. The aim is achieved by undergoing canonical correlation analysis on multiple variables which represent both halves. On this paper, the case is limited to the year 2019 or academic year 2018/2019.

1.2 Literature Review

1.2.1. Human Development Index

According to the United Nations Development Programme (UNDP), the Human Development Index (HDI) is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable, and have a decent standard of living. In one term each, the HDI compiled the latest situation of health, education, and wealth of people in a nation. The HDI was introduced in 1990 by a Pakistani economist, Mahbub ul Haq.

In the latest Human Development Report published by the UNDP, there are four indices measured with multiple, specific constraints from each of the 191 country members of the United Nations (UN). The four aspects are life expectancy for health, expected years of schooling and mean years of schooling for education, and Gross National Income (GNI) per capita for standard of living. The Human Development Index is the geometric mean of normalized indices for each of the three dimensions.

1.2.2. Multivariate Analysis

The complexities of most phenomena happening in daily life require collecting observations on many different variables. The body of methodology to understand simultaneous measurements on multiple variables is called *multivariate analysis*. (Rencher, 2002:1).

Initially, the utilization of multivariate analyses were in the behavioral and biological sciences. However, interest and demands in multivariate methods have spread to numerous other fields of study, such as education, engineering, linguistics, psychology, and many others. Multivariate analysis is concerned generally with two areas, which are descriptive and inferential statistics. A popular technique in the descriptive realm is to find linear combinations of variables, while in the inferential area, numerous methods are extensions of univariate procedures. More specifically, Johnson and Wichern (2007) quoted the objectives of scientific investigations to which multivariate methods most naturally lend themselves include:

1. data reduction or structural simplification,
2. sorting and grouping,
3. investigation of the dependence among variables,
4. prediction, and
5. hypothesis construction and testing.

The essential mathematical preliminary to learn more about multivariate analyses is matrix algebra. The main reason is due to measurements on several variables or characteristics frequently arranged in graphs and tabular arrangements. On the other hand, statistical prerequisites include descriptive statistics and graphical techniques.

1.2.3. Canonical Correlation Analysis

Canonical correlation analysis could be defined as a statistical method to calculate the amount of linear relationship between two groups of variables. (Rencher, 2002:361) In other words, it is a multivariate statistics technique to quantify the association between two sets of variables. (Härdle & Simar, 2003) Canonical correlation analysis was developed by Harold Hotelling in 1935 in his research to determine the relationship between arithmetic speed and arithmetic power to reading speed and reading power.

The idea behind canonical correlation analysis is to determine the possible linear combinations of each set of variables and find a pair of linear combinations which results in the largest correlation. The pairs of linear combinations are called *canonical variables*. The aim of a canonical correlation analysis is to maximize the correlation between two sets of data.

Let $\boldsymbol{Y}_{(p\times 1)} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix}$ and $\boldsymbol{X}_{(q\times 1)} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix}$ be two random vectors of variables measured from the same sampling units. It is assumed that $\boldsymbol{Y}$ is the smaller set, so $p \leq q$. We recall multiple correlation, which is the correlation between one Y and multiple X's. The sample covariances and correlations among $Y, X_1, X_2, \dots, X_q$ could be denoted in the matrices:

$$\boldsymbol{S} = \begin{bmatrix} s_Y{}^2 & \boldsymbol{s}_{Yx}{}^{\boldsymbol{T}} \\ \boldsymbol{s}_{Yx} & \boldsymbol{S}_{xx} \end{bmatrix} \text{ and}$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & \boldsymbol{r}_{Yx}{}^{\boldsymbol{T}} \\ \boldsymbol{r}_{Yx} & \boldsymbol{R}_{xx} \end{bmatrix},$$

where bold, lowercase elements are subvectors and bold, uppercase elements are submatrices. For details,

- $\boldsymbol{s}_{Yx}{}^{\boldsymbol{T}} = [s_{Y1} \quad s_{Y2} \quad \cdots \quad s_{Yq}]$ is the vector of sample covariances of Y with $X_1, X_2, \dots, X_q$;
- $\boldsymbol{S}_{xx} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ s_{q1} & s_{q2} & \cdots & s_{qq} \end{bmatrix}$ is the matrix of sample covariances of the X's;
- $\boldsymbol{r}_{Yx}{}^{\boldsymbol{T}} = [r_{Y1} \quad r_{Y2} \quad \cdots \quad r_{Yq}]$ is the vector of sample correlations of Y with $X_1, X_2, \dots, X_q$; and
- $\boldsymbol{R}_{xx} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ r_{q1} & r_{q2} & \cdots & r_{qq} \end{bmatrix}$ is the matrix of sample correlations of the X's.

The squared multiple correlation (denoted by R^2) between Y and the X's could be computed from the partitioned sample covariance matrix or correlation matrix as:

$$R^2 = \frac{\boldsymbol{s}_{Yx}{}^{\boldsymbol{T}}\boldsymbol{S}_{xx}{}^{-\boldsymbol{1}}\boldsymbol{s}_{Yx}}{s_Y{}^2} = \boldsymbol{r}_{Yx}{}^{\boldsymbol{T}}\boldsymbol{R}_{xx}{}^{-\boldsymbol{1}}\boldsymbol{r}_{Yx}.$$

R^2 could be interpreted as a single measure of linear relationship between Y and the X's. Alternatively, the multiple correlation R could be defined as the maximum correlation between Y and a linear combination of the X's, i.e. $R = \max_{\boldsymbol{b}} r_{Y,\boldsymbol{b}^T\boldsymbol{X}}$, where $\boldsymbol{b}$ is a vector of coefficients which will be discussed further.

Canonical correlation is the generalization of multiple correlation, where we deal with multiple Y's and multiple X's. To begin with, we define a joint vector between $\boldsymbol{Y}$ and $\boldsymbol{X}$ and call it $\boldsymbol{W}$.

$$\boldsymbol{W}_{(p+q)\times 1} = \begin{bmatrix}\boldsymbol{Y}\\ \boldsymbol{X}\end{bmatrix} = \begin{bmatrix}Y_1\\ Y_2\\ \vdots\\ Y_p\\ X_1\\ X_2\\ \vdots\\ X_q\end{bmatrix}$$

The random vector $\boldsymbol{W}$ has a corresponding mean vector

$$\overline{\boldsymbol{W}}_{(p+q)\times 1} = E(\boldsymbol{Z}) = \begin{bmatrix}E(\boldsymbol{Y})\\ E(\boldsymbol{X})\end{bmatrix} = \begin{bmatrix}\overline{\boldsymbol{Y}}\\ \overline{\boldsymbol{X}}\end{bmatrix}$$

and covariance matrix

$$\boldsymbol{S}_{(p+q)\times(p+q)} = E(\boldsymbol{W} - \overline{\boldsymbol{W}})(\boldsymbol{W} - \overline{\boldsymbol{W}})^T$$

$$= \begin{bmatrix}E(\boldsymbol{Y} - \overline{\boldsymbol{Y}})(\boldsymbol{Y} - \overline{\boldsymbol{Y}})^T & E(\boldsymbol{Y} - \overline{\boldsymbol{Y}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^T\\ E(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{Y} - \overline{\boldsymbol{Y}})^T & E(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^T\end{bmatrix}$$

$$= \begin{bmatrix}\boldsymbol{S}_{yy\,(p\times p)} & \boldsymbol{S}_{yx\,(p\times q)}\\ \boldsymbol{S}_{xy\,(q\times p)} & \boldsymbol{S}_{xx\,(q\times q)}\end{bmatrix}.$$

The squared multiple correlation (denoted by $R_M{}^2$) between $\boldsymbol{Y}$ and $\boldsymbol{X}$ could be computed from the partitioned sample covariance matrix by:

$$R_M{}^2 = \left|\boldsymbol{S}_{yy}{}^{-\boldsymbol{1}}\boldsymbol{S}_{yx}\boldsymbol{S}_{xx}{}^{-\boldsymbol{1}}\boldsymbol{S}_{xy}\right| = \prod_{i=1}^{s} r_i{}^2,$$

where $s = \min{(p, q)}$ and $r_i{}^2, i = 1, 2, \dots, s$ are the eigenvalues of $\boldsymbol{S}_{yy}{}^{-\boldsymbol{1}}\boldsymbol{S}_{yx}\boldsymbol{S}_{xx}{}^{-\boldsymbol{1}}\boldsymbol{S}_{xy}$. $R_M{}^2$ is a poor measure of association because

$0 \leq {r_i}^2 \leq 1$ for all i and their product is usually too small to reflect the amount of association. On the other hand, the eigenvalues themselves provide meaningful measures of association and their square roots, $r_i, i = 1, 2, \dots, s$ are called *canonical correlations*.

Let $\boldsymbol{a}_{(p\times 1)} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$ and $\boldsymbol{b}_{(q\times 1)} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_q \end{bmatrix}$ be defined as *coefficient vectors*. We define the linear combinations of $\boldsymbol{Y}$'s and $\boldsymbol{X}$'s as $U = \boldsymbol{a^T Y}$ and $V = \boldsymbol{b^T X}$. We obtain the formulas of variances and covariance of U and V as:

$$Var(U) = \boldsymbol{a^T} Cov(\boldsymbol{Y})\boldsymbol{a} = \boldsymbol{a^T S}_{\boldsymbol{yy}}\boldsymbol{a};$$

$$Var(V) = \boldsymbol{b^T} Cov(\boldsymbol{X})\boldsymbol{b} = \boldsymbol{b^T S}_{\boldsymbol{xx}}\boldsymbol{b};$$

$$Cov(U, V) = \boldsymbol{a^T} Cov(\boldsymbol{Y}, \boldsymbol{X})\boldsymbol{b} = \boldsymbol{a^T S}_{\boldsymbol{yx}}\boldsymbol{b}.$$

The first pair of *canonical variables* is the pair of linear combinations U_1, V_1 which yields the canonical correlations, i.e.

$$r_1 = \max_{\boldsymbol{a},\boldsymbol{b}} r_{\boldsymbol{U},\boldsymbol{V}},$$

where

$$r_{\boldsymbol{U},\boldsymbol{V}} = Corr(U, V) = \frac{\boldsymbol{a^T S}_{\boldsymbol{yx}}\boldsymbol{b}}{\sqrt{\boldsymbol{a^T S}_{\boldsymbol{yy}}\boldsymbol{a}\sqrt{\boldsymbol{b^T S}_{\boldsymbol{xx}}\boldsymbol{b}}}}.$$

There are additional pairs of canonical variables $U_i = \boldsymbol{a_i}^{\boldsymbol{T}}\boldsymbol{Y}$ and $V_i = \boldsymbol{b_i}^{\boldsymbol{T}}\boldsymbol{X}$ which consecutively maximize $r_2, r_3, \dots, r_s$. In general, there are $s = \min{(p, q)}$ values of squared canonical correlation ${r_i}^2$.

A theorem of matrix algebra states that nonzero eigenvalues of the multiplication of two matrices $\boldsymbol{AB}$ are equal to the eigenvalues of $\boldsymbol{BA}$ as long as $\boldsymbol{AB}$ and $\boldsymbol{BA}$ are square, but both matrices yield different eigenvectors. In this case, if we define $\boldsymbol{A} = \boldsymbol{S_{yy}}^{-\boldsymbol{1}}\boldsymbol{S_{yx}}$ and $\boldsymbol{B} = \boldsymbol{S_{xx}}^{-\boldsymbol{1}}\boldsymbol{S_{xy}}$, then we could obtain ${r_i}^2, i = 1, 2, \dots, s$ from $\boldsymbol{AB}$ or $\boldsymbol{BA}$. The eigenvalues could be obtained from either characteristic equation:

$$\left|\boldsymbol{S_{yy}}^{-\boldsymbol{1}}\boldsymbol{S_{yx}}\boldsymbol{S_{xx}}^{-\boldsymbol{1}}\boldsymbol{S_{xy}} - r^2\boldsymbol{I}\right| = 0,$$

$$\left|\boldsymbol{S}_{xx}{}^{-\mathbf{1}}\boldsymbol{S}_{xy}\boldsymbol{S}_{yy}{}^{-\mathbf{1}}\boldsymbol{S}_{yx} - r^2\boldsymbol{I}\right| = 0.$$

The coefficient vectors $\boldsymbol{a_i}$ and $\boldsymbol{b_i}$ are the eigenvectors of the equal two matrices:

$$\left(\boldsymbol{S}_{yy}{}^{-\mathbf{1}}\boldsymbol{S}_{yx}\boldsymbol{S}_{xx}{}^{-\mathbf{1}}\boldsymbol{S}_{xy} - r^2\boldsymbol{I}\right)\boldsymbol{a} = \boldsymbol{0},$$

$$\left(\boldsymbol{S}_{xx}{}^{-\mathbf{1}}\boldsymbol{S}_{xy}\boldsymbol{S}_{yy}{}^{-\mathbf{1}}\boldsymbol{S}_{yx} - r^2\boldsymbol{I}\right)\boldsymbol{b} = \boldsymbol{0}.$$

From another perspective, $r_1{}^2 \geq r_2{}^2 \geq \cdots \geq r_s{}^2$ are the s largest eigenvalues of $\boldsymbol{S}_{yy}{}^{-\frac{\mathbf{1}}{2}}\boldsymbol{S}_{yx}\boldsymbol{S}_{xx}{}^{-\mathbf{1}}\boldsymbol{S}_{xy}\boldsymbol{S}_{yy}{}^{-\frac{\mathbf{1}}{2}}$ and $\boldsymbol{e_1}, \boldsymbol{e_2}, \dots, \boldsymbol{e_s}$ are the associated eigenvectors. $r_1{}^2 \geq r_2{}^2 \geq \cdots \geq r_s{}^2$ are also the s largest eigenvalues of $\boldsymbol{S}_{xx}{}^{-\frac{\mathbf{1}}{2}}\boldsymbol{S}_{xy}\boldsymbol{S}_{yy}{}^{-\mathbf{1}}\boldsymbol{S}_{yx}\boldsymbol{S}_{xx}{}^{-\frac{\mathbf{1}}{2}}$ and $\boldsymbol{f_1}, \boldsymbol{f_2}, \dots, \boldsymbol{f_s}$ are the associated eigenvectors. The relationships between $\boldsymbol{e_i}, \boldsymbol{f_i}, \boldsymbol{a_i}$, and $\boldsymbol{b_i}$ are:

$$\boldsymbol{f_i} = \boldsymbol{S}_{xx}{}^{-\frac{\mathbf{1}}{\mathbf{2}}}\boldsymbol{S}_{xy}\boldsymbol{S}_{yy}{}^{-\frac{\mathbf{1}}{\mathbf{2}}}\boldsymbol{e_i};$$

$$\boldsymbol{a_i} = \boldsymbol{e_i}{}^{\boldsymbol{T}}\boldsymbol{S}_{yy}{}^{-\frac{\mathbf{1}}{2}}; \text{ and}$$

$$\boldsymbol{b_i} = \boldsymbol{f_i}{}^{\boldsymbol{T}}\boldsymbol{S}_{xx}{}^{-\frac{\mathbf{1}}{\mathbf{2}}},$$

for $i = 1, 2, \dots, s$.

We could also approach the values of canonical correlation from the partitioned correlation matrix. The random vector $\boldsymbol{W}$ has a corresponding correlation matrix of

$$\boldsymbol{R}_{(p+q)\times(p+q)} = \begin{bmatrix} \boldsymbol{R}_{yy}{}_{(p\times p)} & \boldsymbol{R}_{yx}{}_{(p\times q)} \\ \boldsymbol{R}_{xy}{}_{(q\times p)} & \boldsymbol{R}_{xx}{}_{(q\times q)} \end{bmatrix}.$$

The characteristic equations corresponding to the characteristic equations involving the multiplication of submatrices of $\boldsymbol{S}$ is:

$$\left|\boldsymbol{R}_{yy}{}^{-\mathbf{1}}\boldsymbol{R}_{yx}\boldsymbol{R}_{xx}{}^{-\mathbf{1}}\boldsymbol{R}_{xy} - r^2\boldsymbol{I}\right| = 0,$$

$$\left|\boldsymbol{R}_{xx}{}^{-\mathbf{1}}\boldsymbol{R}_{xy}\boldsymbol{R}_{yy}{}^{-\mathbf{1}}\boldsymbol{R}_{yx} - r^2\boldsymbol{I}\right| = 0,$$

which yield the same eigenvalues $r_i{}^2, i = 1, 2, \dots, s$.

If we utilize the partitioned correlation matrix instead of the partitioned covariance matrix, we obtain the same eigenvalues but different eigenvectors:

$$\left(\boldsymbol{R}_{yy}{}^{-\mathbf{1}}\boldsymbol{R}_{yx}\boldsymbol{R}_{xx}{}^{-\mathbf{1}}\boldsymbol{R}_{xy} - r^2\boldsymbol{I}\right)\boldsymbol{c} = \boldsymbol{0},$$

$$\left(\boldsymbol{R}_{yy}{}^{-\mathbf{1}}\boldsymbol{R}_{yx}\boldsymbol{R}_{xx}{}^{-\mathbf{1}}\boldsymbol{R}_{xy} - r^2\boldsymbol{I}\right)\boldsymbol{d} = \mathbf{0}.$$

The eigenvectors $\boldsymbol{c}$ and $\boldsymbol{d}$ are *standardized coefficient vectors* and the relationship with the eigenvectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is $\boldsymbol{c} = \boldsymbol{D}_y\boldsymbol{a}$ and $\boldsymbol{d} = \boldsymbol{D}_x\boldsymbol{b}$, where $\boldsymbol{D}_y = diag\left(s_{Y_1}, s_{Y_2}, \dots, s_{Y_p}\right)$ and $\boldsymbol{D}_x = diag\left(s_{X_1}, s_{X_2}, \dots, s_{X_q}\right)$.

A similar method of this approach is to standardize $\boldsymbol{Y}$ and $\boldsymbol{X}$ into standardized variables $\boldsymbol{Z}^{(\mathbf{1})}$ and $\boldsymbol{Z}^{(\mathbf{2})}$ by the formulas:

$$Z^{(1)}{}_i = \frac{Y_i - \mu_{Y_i}}{\sqrt{s_{ii}}}$$

and

$$Z^{(2)}{}_i = \frac{X_i - \mu_{X_i}}{\sqrt{s_{ii}}}.$$

Then, we define a covariance matrix of the standardized variables:

$$\boldsymbol{R}^*_{(p+q)\times(p+q)} = \begin{bmatrix} \boldsymbol{R}^*_{yy}{}_{(p\times p)} & \boldsymbol{R}^*_{yx}{}_{(p\times q)} \\ \boldsymbol{R}^*_{xy}{}_{(q\times p)} & \boldsymbol{R}^*_{xx}{}_{(q\times q)} \end{bmatrix}.$$

Thus, by this approach, the canonical variables U and V are defined as:

$$U_i = \boldsymbol{a_i}^{\boldsymbol{T}}\boldsymbol{Z}^{(\mathbf{1})} = \boldsymbol{e_i}^{\boldsymbol{T}}\boldsymbol{R}^*_{yy}{}^{-\frac{\mathbf{1}}{\mathbf{2}}}\boldsymbol{Z}^{(\mathbf{1})},$$

$$V_i = \boldsymbol{b_i}^{\boldsymbol{T}}\boldsymbol{Z}^{(\mathbf{2})} = \boldsymbol{f_i}^{\boldsymbol{T}}\boldsymbol{R}^*_{xx}{}^{-\frac{\mathbf{1}}{\mathbf{2}}}\boldsymbol{Z}^{(\mathbf{2})}.$$

$\boldsymbol{e_i}$ and $\boldsymbol{f_i}$, $i = 1, 2, \dots, s$ are respectively the eigenvectors of $\boldsymbol{R}^*_{yy}{}^{-\frac{\mathbf{1}}{\mathbf{2}}}\boldsymbol{R}^*_{yx}\boldsymbol{R}^*_{xx}{}^{-\mathbf{1}}\boldsymbol{R}^*_{xy}\boldsymbol{R}^*_{yy}{}^{-\frac{\mathbf{1}}{\mathbf{2}}}$ and $\boldsymbol{R}^*_{xx}{}^{-\frac{\mathbf{1}}{\mathbf{2}}}\boldsymbol{R}^*_{xy}\boldsymbol{R}^*_{yy}{}^{-\mathbf{1}}\boldsymbol{R}^*_{yx}\boldsymbol{R}^*_{xx}{}^{-\frac{\mathbf{1}}{\mathbf{2}}}$. $r_1{}^2 \geq r_2{}^2 \geq \cdots \geq r_s{}^2$ are the s largest eigenvalues of either matrices.

Canonical correlations have at least three interesting properties.

1. $Var(U_i) = Var(V_i) = 1$

   $Cov(U_i, U_j) = Corr(U_i, U_j) = 0, \qquad i \neq j$

   $Cov(V_i, V_j) = Corr(V_i, V_j) = 0, \qquad i \neq j$

   $Cov(U_i, V_j) = Corr(U_i, V_j) = 0, \qquad i \neq j$

   for $i, j = 1, 2, \dots, s$.
2. Simple, multiple, and canonical correlations are insensitive to changes of scale on either the $\boldsymbol{Y}$’s or $\boldsymbol{X}$’s.

3. Since the first canonical correlation r_1 is the maximum correlation between the linear combinations of $\boldsymbol{Y}$ and $\boldsymbol{X}$, r_1 exceeds the absolute value of the simple correlation or multiple correlation between any $\boldsymbol{Y}$'s and $\boldsymbol{X}$'s.

1.2.4. Tests of Significance of Canonical Correlations

An important test of significance in a canonical correlation analysis is the test of no relationship between the $\boldsymbol{Y}$'s and $\boldsymbol{X}$'s. To begin with, we set the hypotheses:

- H0: $\boldsymbol{S}_{yx} = \boldsymbol{0}$ (There is no linear relationship between the $\boldsymbol{Y}$'s and $\boldsymbol{X}$'s. All canonical correlations $r_1, r_2, \dots, r_s$ are not significant.)
- H1: $\boldsymbol{S}_{yx} \neq \boldsymbol{0}$ (There is a linear relationship between the $\boldsymbol{Y}$'s and $\boldsymbol{X}$'s. Some canonical correlations $r_1, r_2, \dots, r_s$ are significant.)

There are four multivariate test statistics could be used to infer in this test.

1.2.3.a. Wilks' Likelihood Ratio Statistic (Wilks' Λ)

Under the significance rate of α, the significance of the canonical correlations could be tested by

$$\Lambda_1 = \frac{|\boldsymbol{S}|}{|\boldsymbol{S}_{yy}||\boldsymbol{S}_{xx}|} = \frac{|\boldsymbol{R}|}{|\boldsymbol{R}_{yy}||\boldsymbol{R}_{xx}|},$$

which is distributed as $\Lambda_{p,q,n-1-q}$. H0 is rejected if $\Lambda_1 \leq \Lambda_\alpha$. Λ_1 is also expressible in terms of the squared canonical correlations as:

$$\Lambda_1 = \prod_{i=1}^{s} (1 - r_i{}^2).$$

If the parameters exceed the range of critical values for Wilks' Λ (or if the observation size n is large), we could use Bartlett's approximation:

$$\chi^2 = -\left[n - \frac{1}{2}(p + q + 3)\right] \ln(\Lambda_1),$$

which is distributed as χ^2_{pq}. With this test statistic, H0 is rejected if $\chi^2 \geq \chi^2_\alpha$. Alternatively, we could also use the *F*-approximation with test statistic defined as:

$$F = \frac{1 - \Lambda_1{}^{\frac{1}{t}}}{\Lambda_1{}^{\frac{1}{t}}} \frac{df_2}{df_1},$$

which is distributed as F_{df_1, df_2}, where

$$df_1 = pq$$

and

$$df_2 = \left\{n - \frac{1}{2}(p + q + 3)\right\} \sqrt{\frac{p^2q^2 - 4}{p^2 + q^2 - 5}} - \frac{1}{2}pq + 1.$$

With this test statistic, H0 is rejected if $F > F_\alpha$.

1.2.3.b. Pillai's Test Statistic

Pillai's test statistic for the significance of canonical correlations is

$$V^{(s)} = \sum_{i=1}^{s} r_i{}^2.$$

Upper percentage points of $V^{(s)}$ are indexed by $s = \min(p, q)$, $m = \frac{1}{2}(|q - p| - 1)$, and $N = \frac{1}{2}(n - q - p - 2)$. *F*-approximations could be applied to $V^{(s)}$.

1.2.3.c. Lawley−Hotelling's Test Statistic

Lawley−Hotelling's test statistic for the significance of canonical correlations is

$$U^{(s)} = \sum_{i=1}^{s} \frac{r_i{}^2}{1 - r_i{}^2}.$$

Upper percentage points of $\frac{\nu_E U^{(s)}}{\nu_H}$ are indexed by p, $\nu_H = q$, and $\nu_E = n - q - 1$. *F*-approximations could be applied to $U^{(s)}$.

1.2.3.d. Roy's Largest Root Statistic

Roy's largest root statistic for the significance of canonical correlations is given by

$$\theta = {r_1}^2.$$

Upper percentage points of θ are indexed by s, m, and N as defined for Pillai's test. F-approximations could be applied to θ.

1.3 Research Methodology

1.3.1. Data

Before earning the dataset to be analyzed, numerous datasets were gathered from multiple sources on the internet. For details, the gathered datasets and their respective sources are:

- the number of high schools, teachers, and high school students in Indonesia by 2019 from *https://www.bps.go.id/indikator/indikator/view_data_pub/0000/api_pub/a1lFcnlHNXNYMFlueG8xL0ZOZnU0Zz09/da_04/4*,
- the quality levels of high schools in Indonesia by 2019 from *https://npd.kemdikbud.go.id/?appid=ruangkelas&tahun=2019*,
- the proportion of teachers in high schools in Indonesia by 2019 who had degree(s) of at least applied bachelor (D4) or bachelor (S1) degree from *https://npd.kemdikbud.go.id/?appid=kualifikasi&tahun=2019*,
- the National Examination (*Ujian Nasional*) scores achieved by high school students from the majors of Natural Sciences and Social Sciences in Indonesia in 2019 from *https://npd.kemdikbud.go.id/?appid=hasilun&tahun=2019*,
- the proportion of Indonesian from age 15 to 24 who had been skillful in Information Technology and Computer (*Teknologi Informasi dan Komputer*) by 2019 from *https://www.bps.go.id/indicator/28/1451/1/proporsi-remaja-dan-*

dewasa-usia-15-24-tahun-dengan-keterampilan-teknologi-informasi-dan-komputer-tik-menurut-provinsi.html, and

- literacy rate (*Angka Melek Huruf*) of Indonesian from age 15 to 24 in 2019 from *https://www.bps.go.id/indicator/28/1462/1/angka-melek-huruf-penduduk-umur-15-24-tahun-menurut-provinsi.html*, and
- proportion of school participation (*Angka Partisipasi Sekolah*) of Indonesian from age 16 to 18 in 2019 from *https://www.bps.go.id/indicator/28/301/1/angka-partisipasi-sekolah-a-p-s-.html*.

In this case, the phrase "high school" is inclusive to *Sekolah Menengah Atas* only. Equivalent instances, such as vocational high school (*Sekolah Menengah Kejuruan*), *Madrasah Aliyah*, and so on are not referred to.

After selection and filtration processes, the final dataset obtained consists of 34 tuples and ten attributes. Each tuple represents a province in Indonesia. Meanwhile, the attributes are grouped into two different sets. The first set consists of variables which represent the condition of educational resources, which are:

- the number of high schools (**Banyak.Sekolah**)
- the proportion of high schools which were considered good (**SMA.Baik**),
- the number of high school teachers (**Banyak.Guru**), and
- the proportion of high school teachers which had degree(s) of at least applied bachelor (D4) or bachelor (S1) (**Guru.Baik**).

On the other hand, the second set consists of variables which represent the academic achievements of high school students, which are:

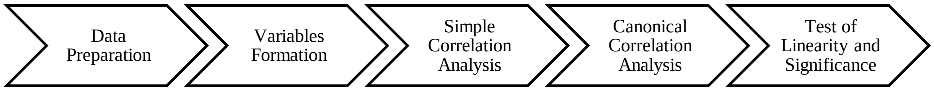
- the number of high school students (**Banyak.Murid**),
- the National Examination scores achieved by high school students from the majors of Natural Sciences (**UN.IPA**),
- the National Examination scores achieved by high school students from the majors of Social Sciences (**UN.IPS**),

- the proportion of high school students who were skillful in Information Technology and Computer (**TIK**),
- the literacy rate of high school students (**AMH**), and
- the proportion of school participation of high school students (**APS**).

All data values were recorded in 2019 or the academic year of 2018/2019.

1.3.2. Data Analysis Procedure

For there are multiple variables in two main groups, in order to know the nature of relationship between both groups, the analysis applied to the dataset is canonical correlation analysis. The procedure to be done in the following chapter is illustrated with the chart below.


Data Preparation
Variables Formation
Simple Correlation Analysis
Canonical Correlation Analysis
Test of Linearity and Significance

1. Data preparation

   Data was gathered and compiled into a file. Then, the dataset is imported into the computation application, RStudio.
2. Variables formation

   A canonical correlation analysis requires two sets of variables. After importing the dataset, the sets of variables are defined.
3. Simple correlation analysis

   The correlation between every pair of variables, despite the sets they are grouped into, is observed in a correlation matrix.
4. Canonical correlation analysis

   The canonical correlations are formed. The raw coefficient vectors are observed and combined with the original variables, while the standardized coefficient vectors are combined with the standardized variables. The linear combinations of variables which maximizes the canonical correlations are modeled into equations.

5. Tests of linearity between variables and significance of canonical correlations

   Under four types of significance test, the significant dimensions of canonical correlations are inferred. Only the significant canonical correlations and corresponding linear combinations are brought into conclusion.

All analysis are done by computation using RStudio.

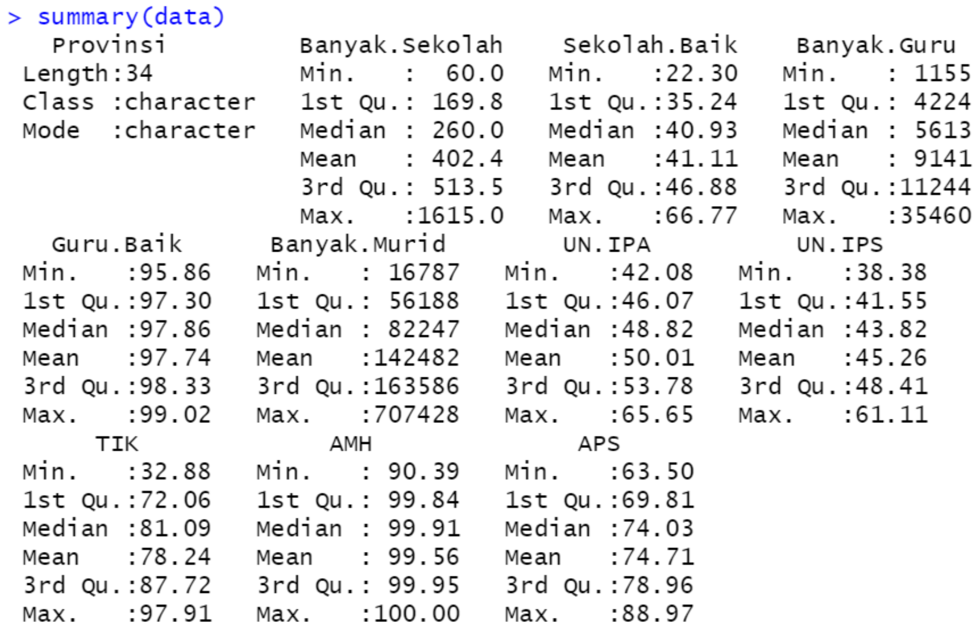
CHAPTER II

BODY

2.1 Analysis Results

Below are the syntaxes along with the outputs (if there is any).

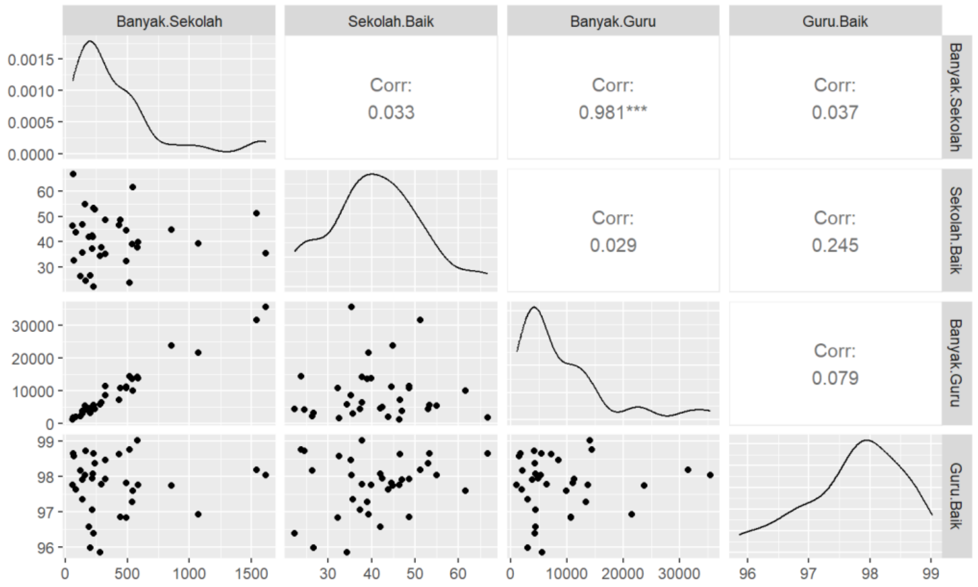
```
## Importing data ##
library(readxl)
data = read_excel("Dataset.xlsx")
View(data)

## Variables formation ##
Y = data[,2:5]
X = data[,6:11]

## Data descriptives ##
summary(data)
```



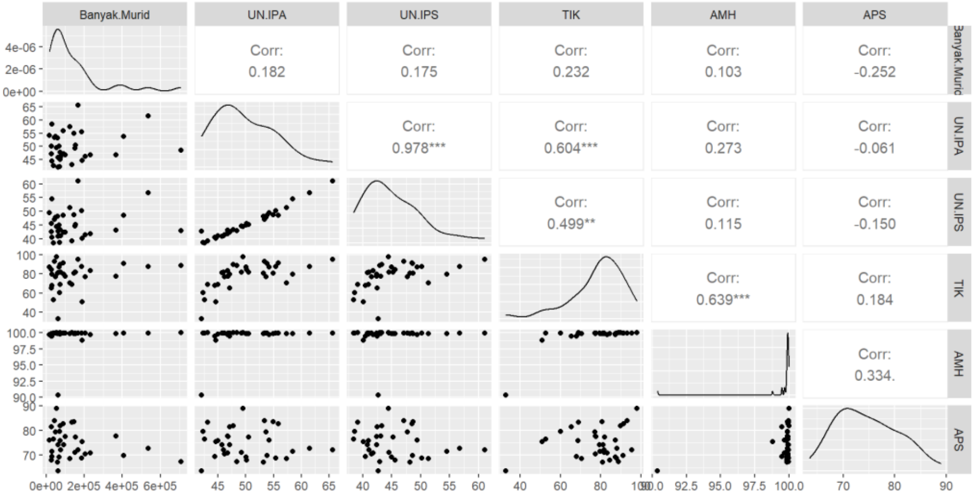
```
> summary(data)
   Provinsi          Banyak.Sekolah    Sekolah.Baik    Banyak.Guru   
 Length:34          Min.   :  60.0   Min.   :22.30   Min.   : 1155  
 Class :character   1st Qu.: 169.8   1st Qu.:35.24   1st Qu.: 4224  
 Mode  :character   Median : 260.0   Median :40.93   Median : 5613  
                    Mean   : 402.4   Mean   :41.11   Mean   : 9141  
                    3rd Qu.: 513.5   3rd Qu.:46.88   3rd Qu.:11244  
                    Max.   :1615.0   Max.   :66.77   Max.   :35460  
   Guru.Baik      Banyak.Murid        UN.IPA          UN.IPS     
 Min.   :95.86   Min.   : 16787   Min.   :42.08   Min.   :38.38  
 1st Qu.:97.30   1st Qu.: 56188   1st Qu.:46.07   1st Qu.:41.55  
 Median :97.86   Median : 82247   Median :48.82   Median :43.82  
 Mean   :97.74   Mean   :142482   Mean   :50.01   Mean   :45.26  
 3rd Qu.:98.33   3rd Qu.:163586   3rd Qu.:53.78   3rd Qu.:48.41  
 Max.   :99.02   Max.   :707428   Max.   :65.65   Max.   :61.11  
      TIK             AMH              APS       
 Min.   :32.88   Min.   : 90.39   Min.   :63.50  
 1st Qu.:72.06   1st Qu.: 99.84   1st Qu.:69.81  
 Median :81.09   Median : 99.91   Median :74.03  
 Mean   :78.24   Mean   : 99.56   Mean   :74.71  
 3rd Qu.:87.72   3rd Qu.: 99.95   3rd Qu.:78.96  
 Max.   :97.91   Max.   :100.00   Max.   :88.97  
```

```
## Canonical Correlation Analysis ##
library(ggplot2)
library(GGally)
library(CCA)
library(CCP)
```

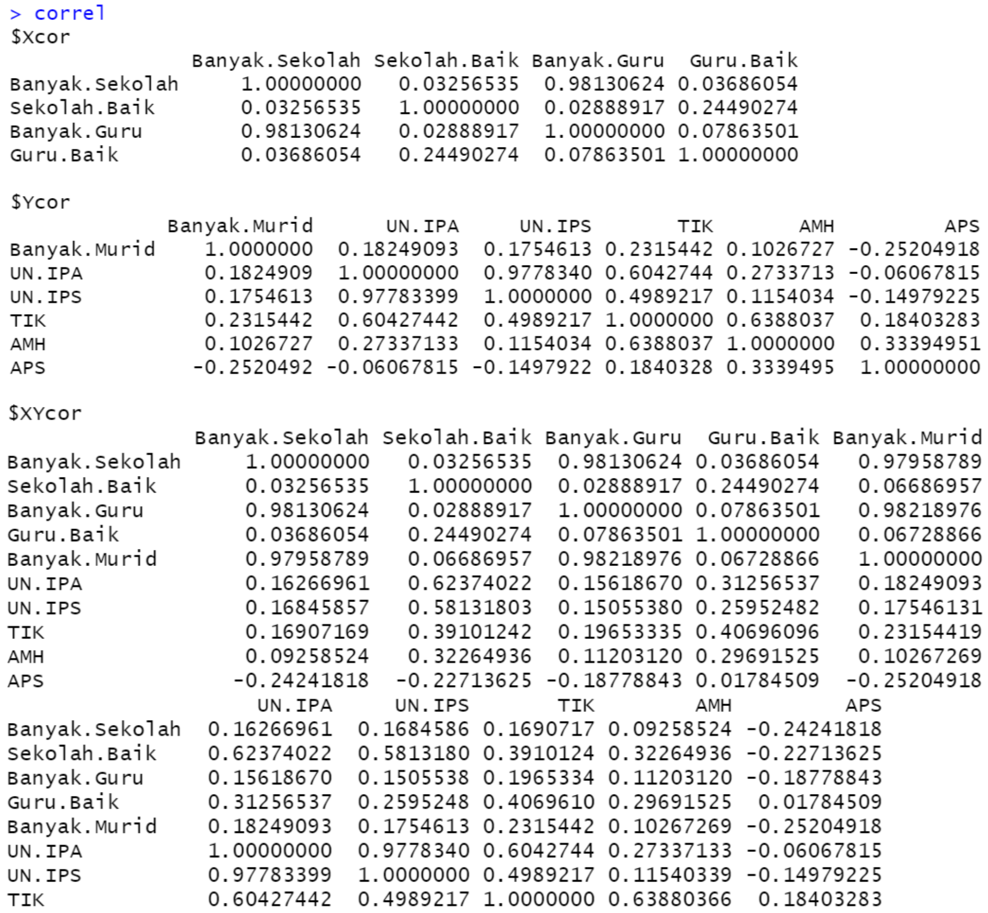
```
ggpairs(Y)
```



```
ggpairs(X)
```

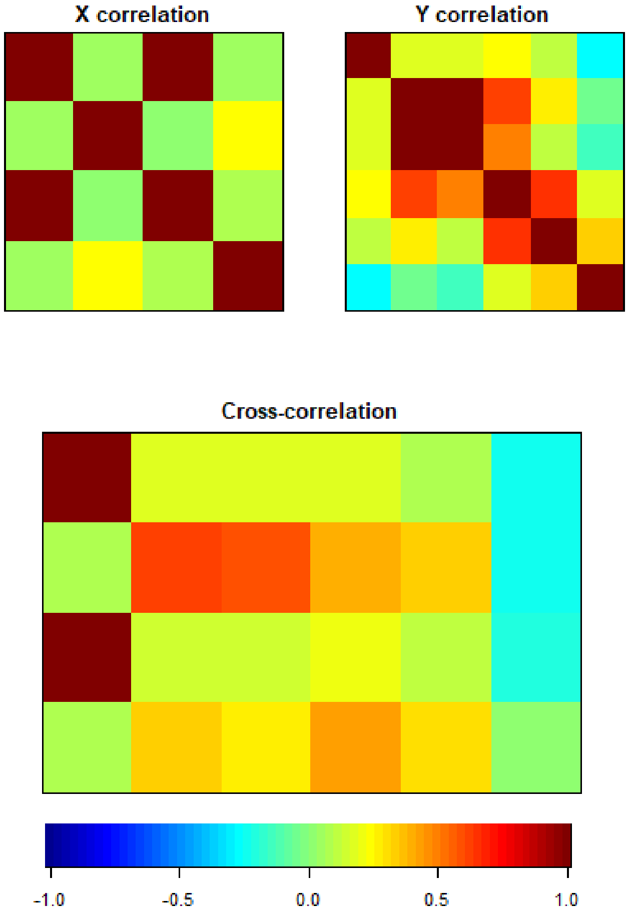


```
# Simple correlations #

correl = matcor(Y, X)

correl
```

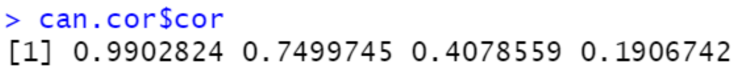


```
> correl
$Xcor
               Banyak.Sekolah Sekolah.Baik Banyak.Guru  Guru.Baik
Banyak.Sekolah     1.00000000   0.03256535  0.98130624 0.03686054
Sekolah.Baik       0.03256535   1.00000000  0.02888917 0.24490274
Banyak.Guru        0.98130624   0.02888917  1.00000000 0.07863501
Guru.Baik          0.03686054   0.24490274  0.07863501 1.00000000

$Ycor
             Banyak.Murid      UN.IPA     UN.IPS       TIK       AMH         APS
Banyak.Murid    1.0000000  0.18249093  0.1754613 0.2315442 0.1026727 -0.25204918
UN.IPA          0.1824909  1.00000000  0.9778340 0.6042744 0.2733713 -0.06067815
UN.IPS          0.1754613  0.97783399  1.0000000 0.4989217 0.1154034 -0.14979225
TIK             0.2315442  0.60427442  0.4989217 1.0000000 0.6388037  0.18403283
AMH             0.1026727  0.27337133  0.1154034 0.6388037 1.0000000  0.33394951
APS            -0.2520492 -0.06067815 -0.1497922 0.1840328 0.3339495  1.00000000

$XYcor
               Banyak.Sekolah Sekolah.Baik Banyak.Guru  Guru.Baik Banyak.Murid
Banyak.Sekolah     1.00000000   0.03256535  0.98130624 0.03686054   0.97958789
Sekolah.Baik       0.03256535   1.00000000  0.02888917 0.24490274   0.06686957
Banyak.Guru        0.98130624   0.02888917  1.00000000 0.07863501   0.98218976
Guru.Baik          0.03686054   0.24490274  0.07863501 1.00000000   0.06728866
Banyak.Murid       0.97958789   0.06686957  0.98218976 0.06728866   1.00000000
UN.IPA             0.16266961   0.62374022  0.15618670 0.31256537   0.18249093
UN.IPS             0.16845857   0.58131803  0.15055380 0.25952482   0.17546131
TIK                0.16907169   0.39101242  0.19653335 0.40696096   0.23154419
AMH                0.09258524   0.32264936  0.11203120 0.29691525   0.10267269
APS               -0.24241818  -0.22713625 -0.18778843 0.01784509  -0.25204918
                   UN.IPA    UN.IPS       TIK        AMH         APS
Banyak.Sekolah 0.16266961 0.1684586 0.1690717 0.09258524 -0.24241818
Sekolah.Baik   0.62374022 0.5813180 0.3910124 0.32264936 -0.22713625
Banyak.Guru    0.15618670 0.1505538 0.1965334 0.11203120 -0.18778843
Guru.Baik      0.31256537 0.2595248 0.4069610 0.29691525  0.01784509
Banyak.Murid   0.18249093 0.1754613 0.2315442 0.10267269 -0.25204918
UN.IPA         1.00000000 0.9778340 0.6042744 0.27337133 -0.06067815
UN.IPS         0.97783399 1.0000000 0.4989217 0.11540339 -0.14979225
TIK            0.60427442 0.4989217 1.0000000 0.63880366  0.18403283
```

```
win.graph()
img.matcor(correl, type = 2)
```

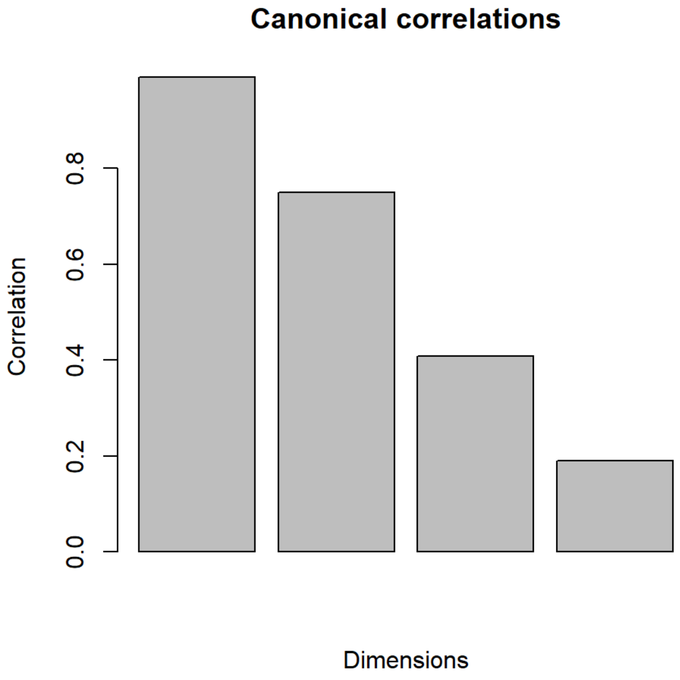


```
# Raw canonical correlations #
can.cor = cc(Y, X)
can.cor$cor
```

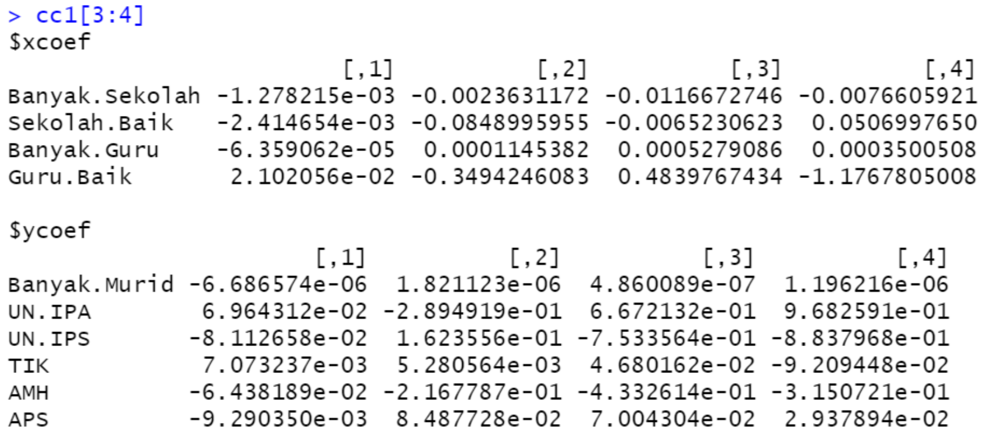


```
> can.cor$cor
[1] 0.9902824 0.7499745 0.4078559 0.1906742
```

```
barplot(cc1$cor, main = "Canonical correlations",
         xlab = "Dimensions", ylab = "Correlation")
```

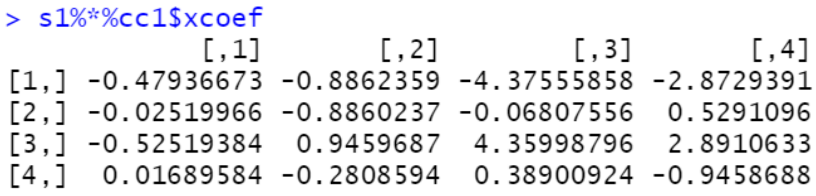


```
# Coefficients of raw canonical variables #
cc1[3:4]
```



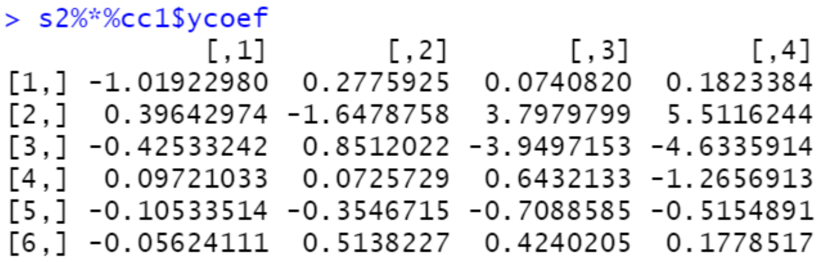
```
> cc1[3:4]
$xcoef
                       [,1]          [,2]          [,3]          [,4]
Banyak.Sekolah -1.278215e-03 -0.0023631172 -0.0116672746 -0.0076605921
Sekolah.Baik   -2.414654e-03 -0.0848995955 -0.0065230623  0.0506997650
Banyak.Guru    -6.359062e-05  0.0001145382  0.0005279086  0.0003500508
Guru.Baik       2.102056e-02 -0.3494246083  0.4839767434 -1.1767805008

$ycoef
                     [,1]          [,2]          [,3]          [,4]
Banyak.Murid -6.686574e-06  1.821123e-06  4.860089e-07  1.196216e-06
UN.IPA        6.964312e-02 -2.894919e-01  6.672132e-01  9.682591e-01
UN.IPS       -8.112658e-02  1.623556e-01 -7.533564e-01 -8.837968e-01
TIK           7.073237e-03  5.280564e-03  4.680162e-02 -9.209448e-02
AMH          -6.438189e-02 -2.167787e-01 -4.332614e-01 -3.150721e-01
APS          -9.290350e-03  8.487728e-02  7.004304e-02  2.937894e-02
```

```
# Standardized canonical coefficients #

s1 = diag(sqrt(diag(cov(Y))))

s1%*%cc1$xcoef
```



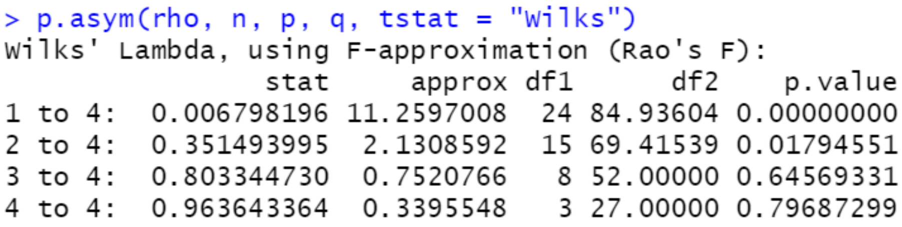
```
> s1%*%cc1$xcoef
            [,1]       [,2]        [,3]       [,4]
[1,] -0.47936673 -0.8862359 -4.37555858 -2.8729391
[2,] -0.02519966 -0.8860237 -0.06807556  0.5291096
[3,] -0.52519384  0.9459687  4.35998796  2.8910633
[4,]  0.01689584 -0.2808594  0.38900924 -0.9458688
```

```
s2 = diag(sqrt(diag(cov(X))))

s2%*%cc1$ycoef
```



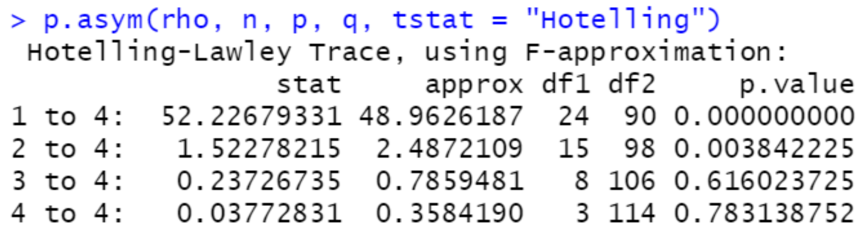
```
> s2%*%cc1$ycoef
            [,1]       [,2]       [,3]       [,4]
[1,] -1.01922980  0.2775925  0.0740820  0.1823384
[2,]  0.39642974 -1.6478758  3.7979799  5.5116244
[3,] -0.42533242  0.8512022 -3.9497153 -4.6335914
[4,]  0.09721033  0.0725729  0.6432133 -1.2656913
[5,] -0.10533514 -0.3546715 -0.7088585 -0.5154891
[6,] -0.05624111  0.5138227  0.4240205  0.1778517
```

```
## Significance test ##

rho = cc1$cor

n = dim(Y)[1]

p = length(Y)

q = length(X)


p.asym(rho, n, p, q, tstat = "Wilks")
```



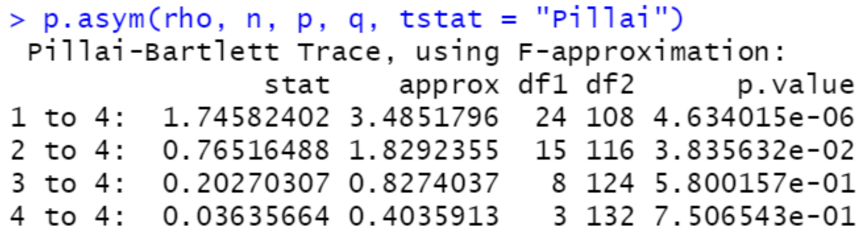
```
> p.asym(rho, n, p, q, tstat = "Wilks")
Wilks' Lambda, using F-approximation (Rao's F):
               stat     approx df1      df2    p.value
1 to 4:  0.006798196 11.2597008  24 84.93604 0.00000000
2 to 4:  0.351493995  2.1308592  15 69.41539 0.01794551
3 to 4:  0.803344730  0.7520766   8 52.00000 0.64569331
4 to 4:  0.963643364  0.3395548   3 27.00000 0.79687299
```

```
p.asym(rho, n, p, q, tstat = "Hotelling")
```



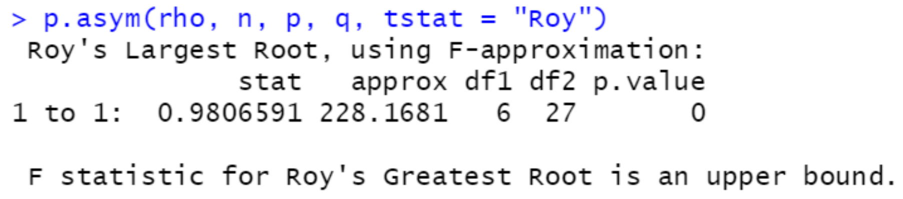
```
> p.asym(rho, n, p, q, tstat = "Hotelling")
 Hotelling-Lawley Trace, using F-approximation:
                stat    approx df1 df2     p.value
1 to 4:  52.22679331 48.9626187  24  90 0.000000000
2 to 4:   1.52278215  2.4872109  15  98 0.003842225
3 to 4:   0.23726735  0.7859481   8 106 0.616023725
4 to 4:   0.03772831  0.3584190   3 114 0.783138752
```

```
p.asym(rho, n, p, q, tstat = "Pillai")
```



```
> p.asym(rho, n, p, q, tstat = "Pillai")
 Pillai-Bartlett Trace, using F-approximation:
                stat    approx df1 df2      p.value
1 to 4:  1.74582402 3.4851796  24 108 4.634015e-06
2 to 4:  0.76516488 1.8292355  15 116 3.835632e-02
3 to 4:  0.20270307 0.8274037   8 124 5.800157e-01
4 to 4:  0.03635664 0.4035913   3 132 7.506543e-01
```

```
p.asym(rho, n, p, q, tstat = "Roy")
```



```
> p.asym(rho, n, p, q, tstat = "Roy")
 Roy's Largest Root, using F-approximation:
              stat   approx df1 df2 p.value
1 to 1:  0.9806591 228.1681   6  27       0

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

2.2 Discussion

The minimum value, maximum value, and mean of each variable are concluded in the table below.

Variable	Symbol	Minimum Value	Maximum Value	Mean
`Banyak.Sekolah`	Y_1	60.0	1615.0	402.4
`Sekolah.Baik`	Y_2	22.30	66.77	41.11
`Banyak.Guru`	Y_3	1155	35460	9141
`Guru.Baik`	Y_4	95.86	99.02	97.74
`Banyak.Murid`	X_1	16787	707428	142482
`UN.IPA`	X_2	42.08	65.65	50.01
`UN.IPS`	X_3	38.38	61.11	45.26

TIK	X_4	32.88	97.91	78.24
AMH	X_5	90.39	100.00	99.56
APS	X_6	63.50	88.97	74.71

The correlation of each pair of variables are plotted into four matrices, which are the submatrices of the partitioned correlation matrix. The matrices with their elements rounded to the nearest thousandth are displayed below.

$$\boldsymbol{S}_{yy} = \begin{bmatrix} 1 & 0.033 & 0.981 & 0.037 \\ 0.033 & 1 & 0.029 & 0.245 \\ 0.981 & 0.029 & 1 & 0.079 \\ 0.037 & 0.245 & 0.079 & 1 \end{bmatrix}$$

$$\boldsymbol{S}_{xx} = \begin{bmatrix} 1 & 0.182 & 0.175 & 0.232 & 0.103 & -0.252 \\ 0.182 & 1 & 0.978 & 0.604 & 0.273 & -0.061 \\ 0.175 & 0.978 & 1 & 0.499 & 0.115 & -0.150 \\ 0.232 & 0.604 & 0.499 & 1 & 0.639 & 0.184 \\ 0.103 & 0.273 & 0.115 & 0.639 & 1 & 0.334 \\ -0.252 & -0.061 & -0.150 & 0.184 & 0.334 & 1 \end{bmatrix}$$

$$\boldsymbol{S}_{yx} = \begin{bmatrix} 1 & 0.033 & 0.981 & 0.037 & 0.980 & 0.163 & 0.168 & 0.169 & 0.092 & -0.242 \\ 0.033 & 1 & 0.029 & 0.245 & 0.067 & 0.624 & 0.581 & 0.391 & 0.323 & -0.227 \\ 0.981 & 0.029 & 1 & 0.079 & 0.982 & 0.156 & 0.151 & 0.197 & 0.112 & -0.188 \\ 0.037 & 0.245 & 0.079 & 1 & 0.067 & 0.313 & 0.260 & 0.407 & 0.297 & 0.118 \\ 0.980 & 0.067 & 0.982 & 0.067 & 1 & 0.182 & 0.175 & 0.232 & 0.103 & -0.252 \\ 0.163 & 0.624 & 0.156 & 0.313 & 0.182 & 1 & 0.978 & 0.604 & 0.273 & -0.061 \\ 0.168 & 0.581 & 0.151 & 0.260 & 0.175 & 0.978 & 1 & 0.499 & 0.115 & -0.150 \\ 0.169 & 0.391 & 0.197 & 0.407 & 0.232 & 0.604 & 0.499 & 1 & 0.639 & 0.184 \\ 0.092 & 0.323 & 0.112 & 0.297 & 0.103 & 0.273 & 0.115 & 0.639 & 1 & 0.334 \\ -0.242 & -0.227 & -0.188 & 0.118 & -0.252 & -0.061 & -0.150 & 0.184 & 0.334 & 1 \end{bmatrix}$$

There are four canonical correlations formed, which are consecutively r_1 = 0.9902824, r_2 = 0.7499745, r_3 = 0.4078559, and r_4 = 0.1906742. The maximum correlation yielded by the linear combinations between the variables are 0.9902824, which indicates that there is a very strong linear relationship between the two sets of variables.

The linear combinations of the (raw) variables in each set corresponding to each of the canonical correlation above are:

1. $-1.278 \times 10^{-3}Y_1 - 2.415 \times 10^{-3}Y_2 - 6.359 \times 10^{-5}Y_3 + 0.021Y_4$ and $-6.687 \times 10^{-6}X_1 + 0.070X_2 - 0.081X_3 + 7.073 \times 10^{-3}X_4 - 0.064X_5 - 9.290 \times 10^{-3}X_6$;

2. $-0.002Y_1 - 0.085Y_2 + 0.0001Y_3 - 0.349Y_4$ and $1.821 \times 10^{-6}X_1 - 0.289X_2 + 0.162X_3 + 5.281 \times 10^{-3}X_4 - 0.217X_5 + 0.085X_6$;
3. $-0.012Y_1 - 0.007Y_2 + 0.0005Y_3 + 0.484Y_4$ and $4.860 \times 10^{-7}X_1 + 0.667X_2 - 0.753X_3 + 0.047X_4 - 0.433X_5 + 0.070X_6$; and
4. $-0.008Y_1 + 0.051Y_2 + 0.0004Y_3 - 1.177Y_4$ and $1.196 \times 10^{-6}X_1 + 0.968X_2 - 0.884X_3 - 0.092X_4 - 0.315X_5 + 0.029X_6$.

The linear combinations of the standardized variables in each set corresponding to each of the canonical correlation above are:

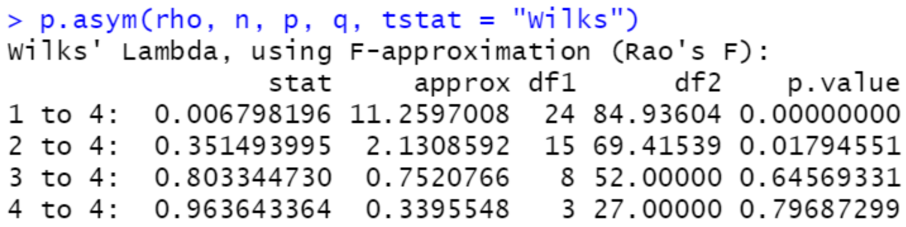
1. $-0.479Y_1^* - 0.025Y_2^* - 0.525Y_3^* + 0.017Y_4^*$ and $-1.019X_1^* + 0.396X_2^* - 0.425X_3^* + 0.097X_4^* - 0.105X_5^* - 0.056X_6^*$;
2. $-0.886Y_1^* - 0.886Y_2^* + 0.946Y_3^* - 0.281Y_4^*$ and $0.278X_1^* - 1.648X_2^* + 0.851X_3^* + 0.073X_4^* - 0.355X_5^* + 0.514X_6^*$;
3. $-4.376Y_1^* - 0.068Y_2^* + 4.360Y_3^* + 0.389Y_4^*$ and $0.074X_1^* + 3.800X_2^* - 3.950X_3^* + 0.643X_4^* - 0.709X_5^* + 0.424X_6^*$; and
4. $-2.873Y_1^* + 0.529Y_2^* + 2.891Y_3^* - 0.946Y_4^*$ and $0.182X_1^* + 5.512X_2^* - 4.634X_3^* - 1.266X_4^* - 0.515X_5^* + 0.178X_6^*$.

The canonical correlation analysis for this test is ended by the tests of linearity between the sets of variables and significance of the canonical correlations. The test statistics of Wilks' Λ is chosen among the four tests.

Wilks' Λ Test of Significance

- Hypotheses
  - H$_0$: There is no linear relationship between the ***Y***'s and ***X***'s. All canonical correlations $r_1, r_2, \ldots, r_s$ are not significant.
  - H$_1$: There is a linear relationship between the ***Y***'s and ***X***'s. Some canonical correlations $r_1, r_2, \ldots, r_s$ are significant.
- Significance rate

  α = 0.05
- Test statistic



```
> p.asym(rho, n, p, q, tstat = "Wilks")
Wilks' Lambda, using F-approximation (Rao's F):
              stat      approx df1      df2    p.value
1 to 4:  0.006798196 11.2597008  24 84.93604 0.00000000
2 to 4:  0.351493995  2.1308592  15 69.41539 0.01794551
3 to 4:  0.803344730  0.7520766   8 52.00000 0.64569331
4 to 4:  0.963643364  0.3395548   3 27.00000 0.79687299
```

- Critical region

  H_0 is rejected if `p.value` $< \alpha$.

- Conclusion

  The combination of dimensions 3 and 4 and the dimension 4 by itself have `p.value`'s of 0.646 and 0.797, consecutively. Therefore, for dimensions 3 and 4, H_0 is not rejected. It could be concluded that there is no significant linear relationship between the $\boldsymbol{Y}$'s and $\boldsymbol{X}$'s in dimensions 3 and 4. Also, all canonical correlations $r_1, r_2, \dots, r_s$ in dimensions 3 and 4 are not significant.

Hence, the significant pair of linear combinations of variables which maximizes the correlation between the condition of educational resources and the academic achievements of high school students in Indonesia in 2019 is:

$$-0.479Y_1^* - 0.025Y_2^* - 0.525Y_3^* + 0.017Y_4^* \text{ and } -1.019X_1^* + 0.396X_2^* - 0.425X_3^* + 0.097X_4^* - 0.105X_5^* - 0.056X_6^*,$$

where the variables had been standardized (subtracted by the respective means, then divided by the respective standard deviations). The maximum correlation between the two sets of variables is 0.990, which could be interpreted that there is a nearly perfect, positive linear relationship between the condition of educational resources and the academic achievements of high school students in Indonesia in 2019.

CHAPTER III
CONCLUSION AND SUGGESTIONS

3.1 Conclusion

The dataset obtained consists of 34 tuples and ten attributes. Each tuple represents a province in Indonesia. Meanwhile, the attributes are grouped into two different sets. The first set consists of variables which represent the condition of educational resources in Indonesia in 2019, which are:

- the number of high schools (**Banyak.Sekolah**)
- the proportion of high schools which were considered good (**SMA.Baik**),
- the number of high school teachers (**Banyak.Guru**), and
- the proportion of high school teachers which had degree(s) of at least applied bachelor (D4) or bachelor (S1) (**Guru.Baik**).

On the other hand, the second set consists of variables which represent the academic achievements of high school students, which are:

- the number of high school students (**Banyak.Murid**),
- the National Examination scores achieved by high school students from the majors of Natural Sciences (**UN.IPA**),
- the National Examination scores achieved by high school students from the majors of Social Sciences (**UN.IPS**),
- the proportion of high school students who were skillful in Information Technology and Computer (**TIK**),
- the literacy rate of high school students (**AMH**), and
- the proportion of school participation of high school students (**APS**).

All data values were recorded in Indonesia in 2019 or the academic year of 2018/2019.

A canonical correlation analysis was done to obtain the pair of linear combinations of each set of variables which maximizes the correlation between both sets. The significant pair of linear combinations of variables which maximizes the correlation between the condition of educational resources and the academic achievements of high school students in Indonesia in 2019 is:

$-0.479Y_1^* - 0.025Y_2^* - 0.525Y_3^* + 0.017Y_4^*$ and $-1.019X_1^* + 0.396X_2^* - 0.425X_3^* + 0.097X_4^* - 0.105X_5^* - 0.056X_6^*$,

where the variables had been standardized. The maximum correlation between the two sets of variables is 0.990, which could be interpreted that there is a nearly perfect, positive linear relationship between the condition of educational resources and the academic achievements of high school students in Indonesia in 2019.

3.2 Suggestions

In the upcoming papers or research about canonical correlation analysis, the respective writer or researcher could improve the computation to obtain more conclusions, such as outputs of canonical loading. Additional tests could also be run, including the test of significance of succeeding canonical correlations after the first. (Rencher, 2002:369)

REFERENCES

Charles University. (n.d.). *NMST539, LS 2015/16, Cvičenie 11 (týždeň 12): Canonical Corelations in R*. Retrieved from SCHOOL OF MATHEMATICS, Faculty of Mathematics and Physics, Charles University: https://www2.karlin.mff.cuni.cz/~maciak/NMST539/cvicenie11.html

Härdle, W., & Simar, L. (2003). *Applied Multivariate Statistical Analysis.* Springer.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). Upper Saddle River, NJ, USA: Pearson Prentice Hall.

Omayan, D. O. (2022, May 16). *Activity 6*. Retrieved from RPubs by RStudio: https://rpubs.com/Devy/902673

Rencher, A. C. (2002). *Methods of Multivariate Analysis* (2nd ed.). Canada: John Wiley & Sons, Inc.

The Pennsylvania State University. (n.d.). *Lesson 13: Canonical Correlation Analysis*. Retrieved from PennState Eberly College of Science: https://online.stat.psu.edu/stat505/book/export/html/682

UCLA: Statistical Consulting Group. (n.d.). *Canonical Correlation Analysis | R Data Analysis Examples*. Retrieved from UCLA Advanced Research Computing: Statistical Methods and Data Analytics: https://stats.oarc.ucla.edu/r/dae/canonical-correlation-analysis/

United Nations Development Programme (UNDP). (n.d.). *Human Development Index (HDI)*. Retrieved from UNDP Human Development Reports: https://hdr.undp.org/data-center/human-development-index#/indicies/HDI