Exploring the Relationship between Health Services and Socioeconomic Variables in Indonesia Using Canonical Correlation Analysis: A Study for Public Health Improvement

Angelica
2540133915
Statistics Department
School of Computer Science
Jakarta, 11530
angelica008@binus.ac.id

Edrick Setiawan
2540124021
Statistics Department
School of Computer Science
Jakarta, 11530
edrick.setiawan@binus.ac.id

Zaphenath Paneah Joseph Irawan
2501961520
Statistics Department
School of Computer Science
Jakarta, 11530
zaphenath.irawan@binus.ac.id

***Abstract*— Health care is one aspect of the needs that must be met by every individual. Good health services can be created by taking into account the economic conditions of each person. The better the economic condition of an individual, the better the health services that can be obtained. This study aims to determine whether there is a relationship between health services and socio-economic characteristics in Indonesia. In this study, the method used is the canonical correlation analysis method with a quantitative approach. The data used is secondary data taken from the Central Bureau of Statistics. Health service variable indicators include immunization, water, and health. Socioeconomic characteristics variable indicators include unemployment, housing, and expenditure. The results obtained in this study indicate that health services have a close relationship with socioeconomic characteristics. The variable indicator of service quality that has the greatest influence is health, while the indicator of socioeconomic characteristics, namely expenditure, is the variable that has the greatest influence.**

***Keywords*— canonical correlation analysis, socio-economic, health service**

I. INTRODUCTION

Health is one of the main factors that are important in people's daily lives, so it needs to be properly maintained physically and mentally [1]. Moreover, health is also important in improving the Human Development Index (HDI) [2]. According to CEO World in 2023, Indonesia is ranked 39th out of 110 countries in the health service index. This health service index is measured by the infrastructure and competence of health workers as well as environmental factors and clean water factors [3].

In Indonesia, most people have suffered from health issues. Over time, the standard of living of the community has increased, so people's demands for health values have also increased. Addressing these issues requires improving health services. Quality health services can provide a good health impact for the Indonesian people. One of the aspects that can improve health is access to clean water. Specifically, clean water is one of the important things because it is beneficial for human survival [4]. In addition, measles immunization in infants is also one of the health services that should be improved. This immunization is important to provide immunity against diseases caused by the measles-rubella virus because measles is easily transmitted [5]. Furthermore, quality health services also play an important role in reducing maternal and infant mortality. Thus, childbirth services at health workers are also one of the important aspects [6].

The existence of necessities such as the occurrence of pain or inconvenience either directly perceived or medically detected will drive a person to use health services. However, utilizing health services incurs costs that must be paid. These increasing costs are closely related to socioeconomic aspects, someone with good finances can afford these costs, while someone less capable may not have health insurance or cannot even afford these expenses [7]. One of the aspects that affect socioeconomics is unemployment. Unemployment has a negative relationship with economic growth [8]. Consequently, if there is a change in economic growth, it will result in a negative change in unemployment. Therefore,

unemployment can be used as one of the measurement tools in the socio-economy. Economic growth is affected by government expenditure. In an economy, if per capita income increases, government expenditure will also increase relatively [9]. Similarly, if per capita expenditure increases, government expenditure will also increase. Therefore, per capita expenditure is one of the socio-economic aspects. Moreover, another socioeconomic aspect is homeownership status. Homeownership is considered to have a positive impact on various socio-economic aspects of human beings [10].

In this research, we will use Canonical Correlation Analysis (CCA) to explore the relationship between health variables and socioeconomic variables. Canonical Correlation Analysis (CCA) emerges as a powerful multivariate statistical technique to address these challenges. Originally proposed by Hotelling (1936), CCA identifies linear combinations of variables from two sets (or more, in multiset CCA) that maximize the correlation between these combinations [11]. This method allows us to uncover complex interdependencies between diverse sets of variables, which is crucial for our study.

Several studies have applied CCA to different domains. For instance, Akour et al. (2023) utilized CCA to explore the relationships between students' performance in face-to-face and online education. Their findings indicated significant canonical correlations between the two modes of education, suggesting that CCA can effectively identify common factors influencing student performance across different learning environments [12]. Similarly, in medical settings, Zhang et al. (2022) developed a sparse multi-view CCA method to identify associations in multimodal brain imaging, which has shown promising results in understanding brain disorders such as Alzheimer's disease [13]. These examples highlight the versatility and efficacy of CCA in various research contexts, including this study which used CCA to look at the relationship between healthcare and socio-economic variables.

This research supports the third Sustainable Development Goals (SDGs), namely good health and well-being. This SDGs aims is to ensure healthy lives and support well-being for all people at all age groups. By understanding how socioeconomic factors and access to health services interact and affect individual health, this research aims to identify policies and interventions that can improve access and quality of health services. Consequently, this research is likely to provide valuable insights to reduce health disparities and ensure that all levels of society can access the right to good health.

Socioeconomic aspects can significantly affect a person's health through various factors, including employment status, lifestyle, access to health services, education level, environmental conditions, and stress [2]. The higher a person's socioeconomic status, the better their health condition. This is mainly because of better access to health resources, better education on health practices, a more secure and healthy living environment, and a better ability to manage stress and maintain a healthy lifestyle. Therefore, this research aims to measure the relationship between healthcare indicators and socioeconomic characteristics to improve public health. Canonical correlation analysis was used to examine the relationship between one group of variables (health service indicators) and another group of variables (socioeconomic characteristics). This approach can not only reveal the strength of the relationship between two groups of variables but also describe the structure of the relationship within each group of variables.

II. METHODOLOGY

A. *Dataset*

The data used in this study are data on health services and data on socioeconomic aspects. This data covers 34 provinces in Indonesia so that the amount of data is 34 rows. The data that used in this study is secondary data taken from the Central Bureau of Statistics (BPS) in 2023. The data used in the first group of variables, namely health service data as many as 3 variables, namely immunization, clean water, and health [14] [15] [16]. In the second group of variables, the socioeconomic characteristics data also has 3 variables, namely unemployment, housing, and expenditure [17] [18] [19].

In the group of x variables, namely socioeconomic characteristics, the unemployment variable is the percentage value of the open unemployment rate, the housing variable is the percentage value of the ownership status of own homes and the expenditure variable refers to per capita monthly food expenditure in rupiah. Variable group y (health service aspects) also has 3 variables where the immunization variable refers to the percentage of toddlers who have been immunized against measles. The water variable relates to the percentage of households with a safe drinking water source. The health variable refers to the percentage of the population who experienced health complaints in the past month.

B. *Canonical Correlation Analysis*

Canonical Correlation is a development form of multiple linear regression analysis. The purpose of canonical correlation analysis is to correlate or correlate several dependent variables with several independent variables. The

difference between canonical correlation and multiple linear regression is that multiple linear regression only uses one dependent variable with several independent variables [20].

Canonical correlation analysis is a statistical method used to identify relationships between two or more sets of variables with different dimensions. CCA finds linear combinations of variables from two data sets that are maximally correlated. This method helps in reducing the dimensionality of the data while retaining the most relevant information [21].

In canonical correlation analysis, the variables with different dimensions are transformed into canonical variables through the formation of a correlation matrix. The correlation matrix is constructed by combining the correlation matrices of the variables. The general form of the conventional CCA is [11]:

$$CCA: \underset{u_1,u_2}{max\rho} = corr(Y_1u_1, Y_2u_2) = \frac{u_1^T\Sigma_{12}u_2}{\sqrt{u_1^T\Sigma_{11}u_1}\sqrt{u_2^T\Sigma_{22}u_2}} \quad (1)$$

Canonical correlation analysis is a multivariate statistical technique used to see the strength of the relationship between two groups of variables, such as groups of variables X and Y. The group of variables X consists of $X_1, X_2, \ldots, X_p$ is denoted by a vector of random variables, and the group of variables Y consisting of $Y_1, Y_2, \ldots, Y_q$ is also denoted by a vector of random variables [22].

Canonical correlation analysis focuses on the correlation between two variables, such as U and V. Variable U is a linear combination of variables in group X, and variable V is a linear combination of variables in group Y [22]. The condition is that the group members of variables x and Y are p ≤ q.

C. *Assumptions of Canonical Correlation Analysis*

The validity of CCA results depends on several key assumptions [11]. There are three assumptions in Canonical Correlation Analysis. All three assumptions are critical to ensuring the robustness and interpretability of the analysis, including linearity, multivariate normality, and the absence of multicollinearity [23].

The first assumption that must be met is that the data follows a multivariate normal distribution. This assumption ensures that the sampling distribution of the canonical correlations is close to normal so that the canonical correlations are reliable [11]. The normality test was conducted using the Shapiro-Wilk Test due to the small population size [24]. The hypothesis for the multivariate normal distribution is as follows.

H_0: variables in group-i are multivariate normally distributed
H_1: variables in group-i are not multivariate normally distributed

The test statistics for the Shapiro-Wilk Test are as follows.

$$W = \frac{\left(\sum_{t=1}^{n} a_i x_i\right)^2}{\sum_{t=1}^{n}(x_i - \bar{x})^2} \quad (2)$$

The rejection region is if the p-value < α or W < critical value, so if the p-value > α , then the variable in the dataset is multivariate normally distributed [24].

Canonical Correlation Analysis assumes that the relationships between the variables in each set are linear. This means that the correlation structure between the variables can be adequately captured by a linear combination of the original variables [11]. The linearity assumption can be done using Pearson correlation. Pearson correlation is a correlation used to measure the strength of the linear relationship between two variables [25]. The hypothesis for this assumption is as follows.

H_0: $\rho = 0$ (there is no significant linear relationship between the two variables)
H_1: $\rho \neq 0$ (there is a significant linear relationship between the two variables)

The test statistics for this assumption are as follows.

$$r_{xy} = \frac{N \cdot \Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N \cdot \Sigma X^2 - (\Sigma X)^2} \cdot N \cdot \Sigma Y^2 - (\Sigma Y)^2} \quad (3)$$

Where r_{xy} is the correlation value, N is the total sample, ΣX is the total of variable x, and ΣY is the total of variable Y. The rejection region is if the p-value < α, and the conclusion is there is a significant linear relationship between the two variables.

The last assumption is multicollinearity. Multicollinearity occurs when there are high correlations between the variables within each set, which can inflate the variance of the estimated coefficients and make the results difficult to interpret [11]. The multicollinearity test can be measured by looking at the Variance Inflation Tolerance (VIF) value. If the VIF value is less than ten (VIF < 10), it means that the variable does not contain multicollinearity [26].

D. *Canonical Correlation Coefficient*

In determining the number of canonical variables to be used, is based on the magnitude of the canonical correlation and the size of the percentage of diversity that can be explained by the canonical variables [27]. These measures are usually seen from the canonical correlation or squared canonical correlation values. Canonical correlation is a measure of the strength of the relationship between two sets of variables while squared canonical correlation is the square

of the canonical correlation value. This correlation is used to measure how much variation in one set of variables can be explained by variation in another set of variables through canonical combinations.

E. Canonical Correlation Significance Test

The significance of the relationship between canonical variables is crucial in understanding the underlying structure and making predictions. There are two tests of parameter significance, namely the simultaneous canonical correlation test and the partial canonical correlation test. Partial significance testing involves testing the significance of the relationship between all variables in the data set. Simultaneous significance testing involves testing the significance of the relationships between all variables in the dataset [28]. The hypothesis for simultaneous significance testing is as follows [23].

H_0: $\rho_1 = \rho_2 = \ldots = \rho_k = 0$ (all canonical correlations are not significant)

H_1: $at\ least\ \rho_i \neq 0, i = 1,2,\ldots,k$ (at least one canonical correlation is significant)

The test statistics for simultaneous significance testing are as follows.

$$B = -\left[(n-1) - \frac{1}{2}(p+q+1)\right] \ln \Lambda \quad (4)$$

$$\Lambda = \prod_{i=1}^{k}\left(1-\rho_i{}^2\right) \quad (5)$$

where Λ is Wilk's Lambda variable, n is the number of observations, p is the number of independent variables, and q is the number of independent variables. The rejection region is if the p-value < 0.05 or $B > \chi^2{}_\alpha$. If the simultaneous hypothesis test concludes that at least one canonical correlation is significant, the partial hypothesis test is conducted to find out which canonical correlation is significant.

Partial significance testing involves testing the significance of the relationships between a subset of variables while controlling for the effects of other variables. This is particularly important in high-dimensional settings where the number of variables is large and the relationships between variables are complex [28]. The hypothesis for partial significance testing is as follows [23].

H_0: $\rho_1 = 0, \rho_2 = 0, \ldots \rho_k = 0$ (the canonical correlation is not significant)

H_1: $\rho_i \neq 0, i = 1,2,\ldots,k$ (the canonical correlation is significant)

The test statistics for the partial significance testing are as follows.

$$B_r = -\left[(n-1) - \frac{1}{2}(p+q+1)\right] \ln \Lambda_r \quad (6)$$

$$\Lambda_r = \prod_{i=r}^{k}\left(1-\rho^2{}_i\right) \quad (7)$$

where Λ_r is Wilk's lambda variable, n is the number of observations, p is the number of independent variables, and q is the number of independent variables. The rejection region is if the p-value < 0.05 or $B_r > \chi^2{}_\alpha$.

F. Canonical Correlation Weight

Canonical correlation weights are essential components in Canonical Correlation Analysis (CCA), which is a method used to understand the relationships between two sets of variables. The canonical weights are the coefficients that maximize the correlation between the linear combinations of the variables in each set. These weights are derived through a process that involves solving an eigenvalue problem, where the eigenvectors corresponding to the largest eigenvalues are the canonical weights.

In the context of Canonical Correlation Analysis (CCA), weights are often referred to as canonical vectors. Latent variables are also known as canonical variates, and the correlations between them are referred to as canonical correlations [29]. Canonical weights or standardized coefficients are interpreted based on their magnitude. A larger value indicates a larger contribution from the variable. Variables with opposite signs indicate an inverse relationship, while variables with the same sign indicate a direct relationship [30].

Obtaining a linear combination for the canonical correlation function requires calculating the weights. The coefficient vector a can be obtained by finding the eigenvalues of the matrix $\Sigma_{YY}{}^{-1}\Sigma_{YX}\,\Sigma_{XX}{}^{-1}\,\Sigma_{XY}$ and the coefficient vector b can be obtained by finding the eigenvalues of the matrix $\Sigma_{XX}{}^{-1}\Sigma_{XY}\,\Sigma_{YY}{}^{-1}\,\Sigma_{YX}$. After obtaining the eigenvalues and eigenvectors, we can convert the eigenvectors into canonical weights by multiplying them with the corresponding inverse root of the covariance matrix [23].

$$a_i = e_i\Sigma_{XX}{}^{-1/2} \quad (8)$$

$$b_i = f_i\Sigma_{YY}{}^{-1/2} \quad (9)$$

where e_i are the eigenvectors corresponding to the eigenvalues of $\Sigma_{YY}{}^{-1}\Sigma_{YX}\,\Sigma_{XX}{}^{-1}\,\Sigma_{XY}$ and f_i are the eigenvectors corresponding to the eigenvalues of $\Sigma_{XX}{}^{-1}\Sigma_{XY}\,\Sigma_{YY}{}^{-1}\,\Sigma_{YX}$.

G. Canonical Correlation Function

In the formation of the canonical correlation function, canonical coefficients are necessary. The group of x variables has a random variable vector X and the group of y variables has a random variable vector Y. The following are the characteristics of the vector [23].

$$E(X) = \mu_x \quad (10)$$

$$E(Y) = \mu_y \quad (11)$$

$$Var(X) = \Sigma_{XX} \quad (12)$$

$$Var(Y) = \Sigma_{YY} \quad (13)$$

$$Cov(X,Y) = \Sigma_{XY} \quad (14)$$

$$Cov(Y,X) = \Sigma_{YX} \quad (15)$$

where Σ_{XX} is the variance-covariance matrix between X and X of size p × p, Σ_{YY} is the variance-covariance matrix between Y and Y of size q × q, Σ_{XY} is the variance-covariance matrix between X and Y of size p × q, and Σ_{YX} is the variance-covariance matrix between Y and X of size q × p.

The linear combination of the two groups of variables is as follows [32].

$$U_i = a_i{}^t X \text{ dan } V_i = b_i{}^t Y \quad (16)$$

That combination can be explained as follows.

$$U = a_1 X_1 + a_2 X_2 + \ldots + a_p X_p \quad (17)$$

$$V = b_1 X_1 + b_2 X_2 + \ldots + b_q Y_q \quad (18)$$

where a is the canonical weight vector for the variables in set X and b is the canonical weight vector for the variables in the set of Y.

H. Canonical Correlation Loadings

Canonical loadings also known as canonical structure coefficients, are the correlations between the original variables and the canonical variates. These loadings provide insight into the relationships between the original variables and the canonical variates, helping to interpret the canonical functions. Essentially, canonical loadings measure the contribution of each original variable to the canonical variates, indicating how much of the variance in the original variables is explained by the canonical variates.

The canonical loadings are crucial for understanding the underlying structure of the data in canonical correlation analysis (CCA). They help in identifying which variables are most strongly associated with the canonical variates, thus providing a clearer interpretation of the canonical functions [31].

The greater the canonical loading value the closer the relationship of the canonical variable concerned with the original variable. Canonical loading is divided into 2 types, including the following [22].

1. The correlation between the canonical variables and the original variables in the same region (intraset correlations) has the following formula.

$$corr(X,U_i) = corr\,(X, a_i^t X) = \Sigma_x\, a_i \quad (19)$$

$$corr(X,V_i) = corr\,(X, b_i^t X) = \Sigma_x\, b_i \quad (20)$$

2. Correlations between canonical variables in one region and variables from different regions (interset correlations) have the following formula.

$$corr(X,U_i) = corr\,(X, a_i^t X) = r_{U_i V_i} \Sigma_y\, b_i \quad (21)$$

$$corr(X,V_i) = corr\,(X, b_i^t X) = r_{U_{i,} V_i} \Sigma_x a_i \quad (22)$$

III. Results and Discussions

A. Assumptions of Canonical Correlation Analysis

In performing canonical correlation analysis, assumption testing is crucial. Therefore, assumption testing is performed at the start before creating the canonical correlation function. Three assumptions must be met, the first is that the data must be normally distributed. This test is done using the Shapiro-Wilk Test with the following hypothesis:

H_0: variables in group-i are multivariate normally distributed

H_1: variables in group-i are not multivariate normally distributed

The test result produced a p-value of 0.004998 for group one (dependent variable) and 0.0004906 for group two (independent variable). Since the p-value is smaller than 0.10, it is concluded that the data is not normally distributed for both groups of variables. Therefore, the transformation and scaling process was carried out. The transformation process was carried out to transform the data into a consistent scale so that the normality test could be met while the scaling process was carried out to equalize units across variables and control rounding errors.

After the transformation process using power transform and scaling, the test results have a p-value of 0.8984 for the dependent variable and 0.876 for the independent variable. Then the normality assumption was met because the p-value is bigger than 0.10 and continued with the linearity assumption using the Pearson correlation test with the following hypothesis.

H_0: $\rho = 0$ (there is no significant linear relationship between the two variables)

H_0: $\rho \neq 0$ (there is a significant linear relationship between the two variables)

The results of assumption testing produce a p-value that can be used to determine whether two variables have a linear relationship and a correlation value that can be seen to measure the strength of the relationship between the two

variables. This test is carried out for all combinations between groups of x variables and groups of y variables.

TABLE 1. PEARSON CORRELATION TEST FOR Y VARIABLES

Variable	P-Value	Correlation Value
immunization	0.005844	0.462854
water	0.0675	0.3172669
health	0.00194	0.5125446

Based on the results in Table 1, since the p-value is smaller than 0.10, the null hypothesis is rejected. Therefore it can be concluded that all combinations of variables have a linear relationship. The immunization variable and the water variable have a linear relationship of 0.462854 and have a positive correlation. The immunization variable and the health variable also have a linear relationship of 0.3172669 and have a positive correlation. The variable water and variable health also have a linear relationship of 0.5125446 and have a positive correlation. All three have a positive correlation, which means that when one variable increases, the other variable will also increase.

TABLE 2. PEARSON CORRELATION TEST FOR X VARIABLES

Variable	P-Value	Correlation Value
unemployment	0.001508	-0.5229302
housing	0.01007	-0.4353577
expenditure	2.643e-05	0.6548736

Based on the results in Table 2, since the p-value is smaller than 0.10, the null hypothesis is rejected. Therefore, it can be concluded that all combinations of variables have a linear relationship. The unemployment variable and the housing variable have a linear relationship of 0.5229302 and have a negative correlation, which means that when the unemployment variable increases, the housing variable will decrease. The unemployment variable and the expenditure variable also have a linear relationship of 0.4353577 and have a negative correlation, so when the unemployment variable increases, the expenditure variable will decrease. The housing variable and the expenditure variable also have a linear relationship of 0.6548736 and a positive correlation, which means that when the housing variable increases, the expenditure variable will also increase.

TABLE 3. MULTICOLLINEARITY TEST FOR X AND Y VARIABLE

Variable		VIF
Variable Y	immunization	1.144348
	water	1.336174
	health	1.281771
Variable X	unemployment	1.355604
	housing	2.468411
	expenditure	2.206032

The presence of multicollinearity is assessed through the VIF value and is considered non-existent if the VIF value for the variable is below 10. From the results of Table 3, it can be seen that in the test there are no variables that have multicollinearity because the VIF value < 10.

B. Canonical Correlation Coefficient

Canonical correlation analysis (CCA) is a statistical method used to understand the relationship between two sets of variables. The canonical correlation (R-value) represents the strength of the relationship between these linear combinations, ranging from -1 to 1, where values closer to 1 or -1 indicate a stronger relationship. The R-squared value (R^2) is the square of the canonical correlation. It indicates the proportion of variance in one set of variables explained by the linear combination of the other set, ranging from 0 to 1, where higher values indicate a stronger explanatory power.

TABLE 4. CANONICAL CORRELATION COEFFICIENT

Canonical Correlation Function	Canonical Correlation	Squared Canonical Correlation
(U_1, V_1)	0.5632226	0.31721972
(U_2, V_2)	0.5145585	0.26477042
(U_3, V_3)	0.1406602	0.01978528

Based on the results in Table 4, the canonical correlation analysis reveals the relationships between variable group x and variable group y with the respective canonical correlations and their squared values. The first pair of canonical variables has a canonical correlation of 0.5632226, indicating a moderately strong relationship with the other set of variables. The squared canonical correlation for the first canonical pair is 0.31721972, suggesting that approximately

31.72% of the variance in unemployment can be explained by the linear combination of the other variables. The second pair of canonical variables shows a slightly lower canonical correlation of 0.5145585, with a squared value of 0.26477042, meaning that about 26.48% of the variance in housing is explained by the combined effect of the other variables. The third pair of canonical variables has the weakest canonical correlation at 0.1406602, with a squared canonical correlation of 0.01978528, indicating that only 1.98% of the variance in expenditure is accounted for by the linear combination of the other variables.

C. Canonical Correlation Significance Test

Significance testing is necessary to determine whether the observed canonical correlations are statistically significant or if they arose randomly. There are two types of tests to be conducted, partial and simultaneous tests.

Simultaneous tests evaluate the significance of all canonical correlations together. The hypothesis for simultaneous significance testing is as follows.

H_0: $\rho_1 = \rho_2 = \ldots = \rho_k = 0$ (all canonical correlations are not significant)

H_1: $at\ least\ \rho_i \neq 0, i = 1,2,\ldots,k$ (at least one canonical correlation is significant)

Overall significance testing using Wilk's statistical test. This test is performed by comparing the p-value with alpha. The rejection criterion is if $\Lambda_k < \Lambda_{table}$. The alpha used is 10% or 0.10. The results show that the statistical test value is 0.492068, which means that the $\Lambda_k < \Lambda_{table}$, is 0.492068 < 0.563, so reject H0. So, it can be concluded that there is at least 1 significant canonical function. After that, proceed with testing the significance of the parameters partially.

Partial tests evaluate the significance of each canonical correlation individually. The hypothesis for partial significance testing is as follows.

H_0: $\rho_1 = 0, \rho_2 = 0, \ldots \rho_k = 0$ (the canonical correlation is not significant)

H_1: $\rho_i \neq 0, i = 1,2,\ldots,k$ (the canonical correlation is significant)

TABLE 5. PARTIAL SIGNIFICANCE TEST

	k = 1	**k = 2**	**k = 3**
Λ_k	0.4920680	0.7206829	0.9802147
Approximate F	2.5669208	2.5803162	0.6055392
df_1	9	4	1
df_2	68.29525	58	30
F-Table	1.7252706	2.0443901	2.8806945

Reject H_0 if approximate F > F table. Based on the results of Table 5, it can be concluded that:

1. At k = 1, because the approximate F value> F table is 2.5669208 > 1.7252706, then reject *H*0. That is, there is at least 1 pair of canonical variates that are correlated or for the first canonical function and there is a relationship between the x variable group and the y variable group.

2. At k = 2, because the approximate F value > F table, namely 2.5803162 > 2.0443901, then reject *H*0. For the second canonical function, there is at least 1 pair of correlated canonical variates and a relationship between the x variable group and the y variable group.

3. At k = 3, because the approximate F value < F table, namely 0.6055392 < 2.8806945, it fails to reject *H*0. For the third canonical function, there is no relationship between the group of x variables and the group of y variables.

Therefore, the canonical functions that will be used for analysis are the first canonical function and the second canonical function because both functions have a significant effect.

D. Canonical Correlation Weight

The canonical weight gives a value of how much a variable has contributed to the canonical variable. The greater the coefficient value of the variable, the greater its influence on the canonical variable. On the other hand, the smaller the value, the less influential the variable is.

TABLE 6. CANONICAL WEIGHTS OF DEPENDENT VARIABLES

Variable	Canonical Weight	
	First Canonical Function	Second Canonical Function
immunization	-0.01566398	1.09977
water	-0.3018651	-0.7269497
health	1.123927	-0.02822717

In the first canonical function, the order of contribution of dependent variables to the canonical variable is health, water, and immunization. This means that from the group of health service variables, the health variable is the most contributing variable in the relationship between health services and socioeconomic characteristics of 1.123927.

In the second canonical function, the order of contribution of dependent variables to the canonical variable is immunization, water, and health. This means that from the group of health service variables, the immunization variable is the most contributing variable in the relationship between health services and socioeconomic characteristics of 1.09977.

Based on the canonical weight value of the dependent variable in Table 6, it produces the following canonical correlation function.

$$U_1 = -0.01566398Y_1 - 0.3018651Y_2 + 1.123927Y_3 \quad (23)$$
$$U_2 = 1.09977Y_1 - 0.7269497Y_2 - 0.02822717Y_3 \quad (24)$$

TABLE 7. CANONICAL WEIGHTS OF INDEPENDENT VARIABLES

Variable	Canonical Weight	
	First Canonical Function	Second Canonical Function
unemployment	0.4786772	0.6586182
housing	0.523573	-0.8255645
expenditure	0.7469636	0.8039897

In the first canonical function, the order of contribution of independent variables to the dependent variable is expenditure, housing, and unemployment. This means that from the group of socioeconomic characteristics variables, the expenditure variable is the variable that contributes the most to the relationship between health services and socioeconomic characteristics of 0.7469636.

In the second canonical function, the order of contribution of independent variables to the dependent variable is housing, expenditure, and unemployment. This means that from the group of health service variables, the housing variable is the variable that contributes the most in the relationship between health services and socio-economic characteristics of -8255645.

Based on the canonical weight value of the independent variable in Table 7, it produces the following canonical correlation function.

$$V_1 = 0.4786772X_1 + 0.523573X_2 + 0.7469636X_3 \quad (25)$$
$$V_2 = 0.6586182X_1 - 0.8255645X_2 + 0.8039897X_3 \quad (26)$$

E. Canonical Correlation Loadings

Canonical loading is the correlation between the original variables and the canonical variates from canonical correlation analysis (CCA). Canonical loadings help interpret the canonical variates by indicating the extent to which each original variable contributes to these linear combinations.

TABLE 8. CANONICAL LOADINGS OF DEPENDENT VARIABLES

Variable	Canonical Loading	
	First Canonical Function	Second Canonical Function
immunization	0.2012014	0.75434254
water	0.2669476	-0.23238460
health	0.9642383	-0.05190086

In the first canonical function, the variable health has the highest canonical loading at 0.9642383, indicating that health strongly contributes to this canonical function. This indicates that the first canonical function primarily reflects the impact of health on the captured variance. The water variable has a moderate loading of 0.2669476, indicating a small impact on Canonical Function 1. Immunization, with a loading of 0.2012014, has the smallest influence on this function compared to the other two variables.

In the second canonical function, immunization carries a substantial weight of 0.75434254, indicating its significant impact on this canonical function. This shows that the second canonical function is mainly affected by immunization. On the other hand, water shows a negative loading of -0.23238460, indicating a reverse connection with Canonical Function 2, albeit with a relatively weak impact. Health has a small negative impact on Canonical Function 2, with a loading of -0.05190086.

TABLE 9. CANONICAL LOADINGS OF INDEPENDENT VARIABLES

Variable	Canonical Loading	
	First Canonical Function	Second Canonical Function
unemployment	-0.1203113	0.74030777
housing	0.7624250	-0.64346426

expenditure	0.8814420	-0.02338518

In the first canonical function, the variable expenditure has the highest canonical loading at 0.8814420, indicating a strong contribution to this canonical function. This indicates that spending has the highest impact on the variation accounted for by the initial canonical function. Housing has a strong positive loading of 0.7624250, showing a notable impact too. On the other hand, there is a small reverse connection between unemployment and the first canonical function with a negative loading of -0.1203113.

In the second canonical function, the unemployment factor has the highest loading of 0.7403077, indicating its significant impact on this function. With a negative loading of -0.64346426, housing exhibits an opposite yet considerable correlation with this function. Expenditure has a small and slightly negative weight of -0.02338518, suggesting a minor impact on the second canonical function.

IV. CONCLUSION

This study demonstrates the application of Canonical Correlation Analysis to understand the relationship between health services and socioeconomic characteristics in Indonesia. The results show that there is a close relationship between the two indicators. Expenditure is the independent variable that has the largest contribution. In contrast, health in the dependent variable has the highest contribution so health and expenditure are variable indicators that have the closest relationship compared to other variable indicators.

REFERENCES

[1] Anathasia, S.E., & Mulyanti, D. (2023). Faktor-Faktor yang Mempengaruhi Peningkatan Kualitas Pelayanan Kesehatan di Rumah Sakit: Tinjauan Teoritis. Jurnal Ilmiah Kedokteran dan Kesehatan, 2(2), 145-151.

[2] Wardhana, A. (2022). Pengaruh Aspek Sosial Ekonomi terhadap Kesehatan. Media Sains Indonesia, 101-108.

[3] Datanesia. (2023). Indonesia Darurat Dokter.

[4] Junaedi, M. (2022). Sanitasi, Pengelolaan dan Akses Air Bersih Untuk Peningkatan Kesehatan di Indonesia. Jurnal Tampiasih, 1 (1), 6-10.

[5] Sari, U.M. (2022). Hubungan antara Pengetahuan Ibu, Jarak Tempuh dan Peran Petugas Kesehatan dengan Pemberian Imunisasi Campak di Puskesmas Gelumbang Tahun 2022. Jurnal IMJ: Indonesia Midwifery Journal, 6(1), 33-40.

[6] Ambarsari, R.D., Sary, Y.N.E., & Azizah, F.M. (2023). Hubungan Kualitas Pelayanan Persalinan Dengan Kepuasan Ibu Bersalin Di Puskesmas Padang Tahun 2022. Jurnal Ilmiah Obsgin, 15(2), 104-111.

[7] Syahri, Y.A., Ritonga, I.R., Silalahi, S.A., & Gurning, F.R. (2024). Pembiayaan Inovatif dalam Peningkatan Akses Pelayanan Kesehatan Masyarakat: Studi Literatur. Jurnal Kesehatan, 2(1), 75-83.

[8] Ardian, R., Syahputra, M., & Dermawan, D. (2022). Pengaruh Pertumbuhan Ekonomi terhadap Tingkat Pengangguran Terbuka di Indonesia. EBISMEN: Jurnal Ekonomi, Bisnis dan Manajemen, 1(3), 190-198.

[9] Rezki, I. (2021). Analisis Pengaruh Pengeluaran Pemerintah Inflasi dan Tenaga Kerja terhadap Pertumbuhan Ekonomi Kota Palangka Raya. Widina Bhakti Persada Bandung.

[10] Idzaji, E.A. (2022). Hubungan Kepemilikan Rumah Terhadap Kepuasan Hidup Subjektif di Indonesia.

[11] Zhuang, X., Yang, Z., & Cordes, D. (2020). A Technical Review of Canonical Correlation Analysis for Neuroscience Applications. Human Brain Mapping, 41(!3) 3807-3833. DOI:10.1002/hbm.25090

[12] Akour, I., Rahamneh, A.A., Kurdi, B. A., Alhamad, A., Al-Makhariz, Alshurideh, M., & Al-Hawary, S. (2023). Using the Canonical Correlation Analysis Method to Study Students' Levels in Face-to-Face and Online Education in Jordan. Information Sciences Letter, 12(2), 901-910. DOI: 10.18576/isl/120229

[13] Zhang, X., Pan, J., Shen, J., Din, Z.U., Li, J., Lu, D., Wu, M., & Hu, B. (2022). Fusing of Electroencephalogram and Eye Movement With Group Sparse Canonical Correlation Analysis for Anxiety Detection. IEEE Transactions on Affective Computing, 13(2), 958-971. DOI: 10.1109/TAFFC.2020.2981440.

[14] Badan Pusat Statistik. (2023). Persentase Balita yang Pernah Mendapat Imunisasi Campak.

[15] Badan Pusat Statistik. (2023). Persentase Rumah Tangga dengan Air Minum Layak.

[16] Badan Pusat Statistik. (2023). Persentase Wanita 15–49 Tahun Pernah Kawin dan Melahirkan Hidup dalam Dua Tahun Terakhir Dengan Penolong Persalinan Tenaga Kesehatan Lain.

[17] Badan Pusat Statistik. (2023). Persentase Tingkat Pengangguran terbuka. l

[18] Badan Pusat Statistik. (2023). Proporsi Rumah Tangga dengan Status Kepemilikan Rumah Milik dan Sewa/Kontrak Menurut Provinsi.

[19] Badan Pusat Statistik. (2023). Rata-rata Pengeluaran per Kapita Sebulan Makanan dan Bukan Makanan di Daerah Perkotaan dan Perdesaan Menurut Provinsi (rupiah).

[20] Rahayu, D. (2022). Aplikasi Analisis Korelasi Kanonik untuk Menguji Pengaruh Konsentrasi, dan terhadap Parameter Meteorologi di Kota Semarang.

[21] Yang, X., Liu, W., Liu, W., & Tao, D. (2021). A Survey on Canonical Correlation Analysis. IEEE Transactions on Knowledge and Data Engineering, 33(6), 2349-2368. DOI:: 10.1109/TKDE.2019.2958342.

[22] Putra, R.D.R., Putra, A.A., & Sriningsih, R. (2021). Analisis Korelasi Kanonik Hubungan Lingkungan Pendidikan terhadap Prestasi Belajar. UNPjoMath, 4(2), 46-50

[23] Lestari, S., Oktaviani, F., Wijaya, A., & Anwar, S. (2020). Hubungan Faktor Multidimensi terhadap Derajat Kemiskinan di Indonesia dengan Analisis Korelasi Kanonik. Journal of Data Analysis, 3(1), 26-35.

[24] Ahadi, G.D., & Zain, N.N.L.E. (2023). The Simulation Study of Normality Test Using Kolmogorov-Smirnov, Anderson-Darling, and Shapiro-Wilk. Eigen Mathematics Journal, 6(1), 11-19.

[25] Sari, F.M., Hadiati, R.N., & Sihotang, W.P. (2023). Analisis Korelasi Pearson Jumlah Penduduk dengan Jumlah Kendaraan Bermotor di Provinsi Jambi. Multi Proximity: Jurnal Statistika Universitas Jambi, 2(1), 39-44.

[26] Octaviana, R. (2022). COVID-19 dan Pasar Modal Syariah: Pendekatan Regresi Korelasi Kanonikal. Jurnal Ekonomi dan Bisnis Islam (al-Tijary), 8(1), 57-67. DOI: 10.21093/at.v8i1.5779.

[27] Suryana, N. A. D., Sulvianti, I. D., & Aidi, M. N. (2021). Analisis Korelasi Kanonik pada Kualitas Air Sungai Ciliwung. Xplore: Journal of Statistics, 10(2), 182–196. DOI: 10.29244/xplore.v10i2.245.

[28] McKeague, I. W., & Zhang, X. (2022). Significance Testing for Canonical Correlation Analysis in High Dimensions. Biometrika, 109(4), 1067–1083.

[29] Mihalik, A., Chapman, J., Adams, R. A., Winter, N. R., Ferreira, F. S., Shawe-Taylor, J., & Mourão-Miranda, J. (2022). Canonical Correlation Analysis and Partial Least Squares for Identifying Brain–Behavior Associations: A Tutorial and A Comparative Study. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 7(11), 1055-1067. DOI: 10.1016/j.bpsc.2022.07.012

[30] Diel, M. I., Lúcio, A. D., Lambrecht, D. M., Pinheiro, M. V. M., Sari, B. G., Olivoto, T., Valera, O. V. S., de Melo, P. J., Tartaglia, F. L., Tischler, A. L., & Schmidt, D. (2020). Canonical Correlations in Agricultural Research: Method of Interpretation used Leads to Greater Reliability of Results. International Journal for Innovation Education and Research, 8(7), 171-181.

[31] Prastio, F., Martha, S., & Perdana, H. (2019). Pengujian Korelasi Kanonik Penumpang Pesawat di Bandar Udara Internasional Supadio. Bimaster : Buletin Ilmiah Matematika, Statistika dan Terapannya, 8(2), 185-192. DOI: 10.26418/bbimst.v8i2.31533

AOL Multivariate Statistics

2024-06-18

Import Library

```
library(readxl)
library(mvShapiroTest)
library(car)
library(MASS)
library(GGally)
library(CCA)
library(expm)
library(dplyr)
library(bestNormalize)
library(mvoutlier)
library(CCP)
```

Group 1

```
kesehatan <- read_excel("C:/Users/angel/OneDrive - Bina
Nusantara/Semester 6/Multivariate Statistics/AOL/kesehatan.xlsx")
kesehatan

kelompoksatu <- kesehatan[, c("imunisasi", "air", "dokter")]
kelompoksatu
```

Normality Assumption

H0: variabel pada kelompok 1 berdistribusi multivariate normal H1: variabel pada kelompok 1 tidak berdistribusi multivariate normal

```
mvShapiro.Test(as.matrix(kelompoksatu))
```

Transformation the Data

```
transform_pt <- summary(powerTransform(kelompoksatu))

imunisasi <- kelompoksatu[,1]^transform_pt$result[1]
air <- kelompoksatu[,2]^transform_pt$result[2]
dokter <- kelompoksatu[,3]^transform_pt$result[3]

transformed <- tibble(imunisasi, air, dokter)
transformed <- as.data.frame(scale(transformed))

mvShapiro.Test(as.matrix(transformed))
```

Linearity Assumption

```
ggpairs(transformed)
```

Cor test

```
cor.test(transformed$imunisasi, transformed$air)
cor.test(transformed$imunisasi, transformed$dokter)
cor.test(transformed$air, transformed$dokter)
```

Multicollinearity Test

```
{r warning=FALSE} kelompoksatu$DummyY <- seq(1) model1 <- lm(DummyY
~., data = kelompoksatu) vif(model1)
```

Group 2

```
sosioekonomi <- read_excel("C:/Users/angel/OneDrive - Bina
Nusantara/Semester 6/Multivariate Statistics/AOL/sosioekonomi.xlsx")
sosioekonomi

kelompokdua <- sosioekonomi[, c("pengangguran", "rumah",
"pengeluaran")]
kelompokdua
```

Normality Assumption

H0: variabel pada kelompok 2 berdistribusi multivariate normal H1: variabel pada kelompok 2 tidak berdistribusi multivariate normal

```
mvShapiro.Test(as.matrix(kelompokdua))
```

Transformation the Data

```
transform_pt2 <- summary(powerTransform(kelompokdua))

pengangguran <- kelompokdua[,1]^transform_pt2$result[1]
rumah <- kelompokdua[,2]^transform_pt2$result[2]
pengeluaran <- kelompokdua[,3]^transform_pt2$result[3]
transformed2 <- tibble(pengangguran,rumah,pengeluaran)
transformed2 <- as.data.frame(scale(transformed2))

mvShapiro.Test(as.matrix(transformed2))
```

Linearity Assumption

```
ggpairs(transformed2)

cor.test(transformed2$pengangguran, transformed2$rumah)
cor.test(transformed2$pengangguran, transformed2$pengeluaran)
cor.test(transformed2$rumah, transformed2$pengeluaran)
```

Multicollinearity Test

```
kelompokdua$DummyY <- seq(1)
model2 <- lm(DummyY ~., data = kelompokdua)
vif(model2)
```

Merge the Dataset

```
# Menggabungkan data kelompok satu dan kelompok dua menjadi satu frame
data_cca <- cbind(transformed, transformed2)

# Menampilkan hasil penggabungan
print(data_cca)
```

Canonical Correlation Analysis

Matrics Correlation

```
rho = cor(data_cca)
rho

(p11 = rho[1:3, 1:3])
(p12 = rho[1:3, 4:6])
(p21 = rho[4:6, 1:3])
(p22 = rho[4:6, 4:6])

(invsqrt_p11 = solve(sqrtm(p11)))
(inv_p22 = solve(p22))
```

Membuat matriks A

```
A = invsqrt_p11 %*% p12 %*% inv_p22 %*% p21 %*% invsqrt_p11
A

# Menghitung nilai korelasi kanonik dari A
(r2 = eigen(A)$values)
(r = sqrt(r2))
```

Uji Serentak

A_3,3,30 = 0.563 Karena lambda hitung < lambda tabel = 0.492 < 0.563 maka setidaknya ada 1 fungsi kanonik yang signifikan.

```
lambda = det(rho)/(det(p11)*det(p22))
lambda
```

Uji Parsial (MANUAL)

```
lambda1 = (1-r[1]^2)*(1-r[2]^2)*(1-r[3]^2)
lambda1

lambda2 = (1-r[2]^2)*(1-r[3]^2)
lambda2
```

```
lambda3 = (1-r[3]^2)
lambda3

lambda_k <- c(lambda1, lambda2, lambda3)
```

Hitung Nilai W

```
p = 3
q = 3
n = 34

# w
w = n - (1/2)*(p + q + 3)
w
```

Hitung Nilai t untuk semua K

```
# t untuk k = 1
t_1 = sqrt(((p-1+1)^2 * (q-1+1)^2 - 4)/((p-1+1)^2 + (q-1+1)^2 - 5))

# t untuk k = 2
t_2 = sqrt(((p-2+1)^2 * (q-2+1)^2 - 4)/((p-2+1)^2 + (q-2+1)^2 - 5))

# t untuk k = 3
t_3 = sqrt(((p-3+1)^2 * (q-3+1)^2 - 4)/((p-3+1)^2 + (q-3+1)^2 - 5))

t_k <- c(t_1, t_2, t_3)
t_k
```

Hitung Nilai df1 untuk semua K

```
# df1 untuk k = 1
df1_1 = (p-1+1)*(q-1+1)

# df1 untuk k = 2
df1_2 = (p-2+1)*(q-2+1)

# df1 untuk k = 3
df1_3 = (p-3+1)*(q-3+1)

df1 <- c(df1_1, df1_2, df1_3)
df1
```

Hitung nilai df2 untuk semua K

```
# df2 untuk k = 1
df2_1 = w * t_1 - (1/2) * ((p-1+1) * (q-1+1)) + 1

# df2 untuk k = 2
df2_2 = w * t_2 - (1/2) * ((p-2+1) * (q-2+1)) + 1

# df2 untuk k = 3
```

```
df2_3 = w * t_3 - (1/2) * ((p-3+1) * (q-3+1)) + 1

df2 <- c(df2_1, df2_2, df2_3)
df2
```

Hitung nilai f hitung

```
# F hitung untuk k = 1
f_1 = ((1-lambda1^(1/t_1)) / lambda1^(1/t_1)) * (df2_1 / df1_1)

# F hitung untuk k = 2
f_2 = ((1-lambda2^(1/t_2)) / lambda2^(1/t_2)) * (df2_2 / df1_2)

# F hitung untuk k = 3
f_3 = ((1-lambda3^(1/t_3)) / lambda3^(1/t_3)) * (df2_3 / df1_3)

f_k <- c(f_1, f_2, f_3)
f_k

ftable <- c(1.725270585, 2.044390135, 2.880694517)

# Result
result <- data.frame(lambda_k, t_k, df1, df2, f_k, ftable)
result
```

Tolak H0 jika F hitung > F table Berdasarkan hasil maka dapat ditarik kesimpulan bahwa: - Dikarenakan nilai F hitung > F table yaitu 2.5669208 > 1.725270585 maka tolak H0. Artinya, minimal ada 1 pasangan canonical variate yang saling berkorelasi atau untuk fungsi kanonik pertama ada hubungan antara kelompok variabel x dan kelompok variabel y. - Dikarenakan nilai F hitung > F table yaitu 2.580316 > 2.044390135 maka tolak H0. Artinya, minimal ada 1 pasangan canonical variate yang saling berkorelasi atau untuk fungsi kanonik kedua ada hubungan antara kelompok variabel x dan kelompok variabel y. - Dikarenakan nilai F hitung < F table yaitu 0.045 < 2.880694517 maka gagal tolak H0. Artinya, untuk fungsi kanonik ketiga tidak terdapat hubungan antara kelompok variabel x dan kelompok variabel y.

Uji Parsial (PACKAGE)

```
# Perform Wilks' Lambda test using CCP package
n <- 34
p <- 3
q <- 3
wilks_result <- p.asym(r, n, p, q, tstat = "Wilks")

# Print the results
wilks_result

## F-test
F_1 = wilks_result$approx[1]
F_2 = wilks_result$approx[2]
F_3 = wilks_result$approx[3]
```

```
## F-Table
f_table_1 = qf(p=0.10, wilks_result$df1[1], wilks_result$df2[1],
lower.tail = FALSE)
f_table_2 = qf(p=0.10, wilks_result$df1[2], wilks_result$df2[2],
lower.tail = FALSE)
f_table_3 = qf(p=0.10, wilks_result$df1[3], wilks_result$df2[3],
lower.tail = FALSE)

F_1
F_2
F_3


f_table_1
f_table_2
f_table_3

(e = eigen(A)$vector)

e1 = eigen(A)$vector[,1]
(u1 = e1 %*% invsqrt_p11)

e2 = eigen(A)$vector[,2]
(u2 = e2 %*% invsqrt_p11)

e3 = eigen(A)$vector[,3]
(u3 = e3 %*% invsqrt_p11)
```

Membuat matriks B

```
(invsqrt_p22 = solve(sqrtm(p22)))
(inv_p11 = solve(p11))

B = invsqrt_p22 %*% p21 %*% inv_p11 %*% p12 %*% invsqrt_p22
B

(f = eigen(B)$vector)

f1 = eigen(B)$vector[,1]
(v1 = f1 %*% invsqrt_p22)

f2 = eigen(B)$vector[,2]
(v2 = f2 %*% invsqrt_p22)

f3 = eigen(B)$vector[,3]
(v3 = f3 %*% invsqrt_p22)
```

Muatan Kanonik

```
# Muatan kanonik untuk variabel X
canonical_loadings_X1 = p11 %*% t(u1)
```

```
canonical_loadings_X2 = p11 %*% t(u2)

# Muatan kanonik untuk variabel Y
canonical_loadings_Y1 = p22 %*% t(v1)
canonical_loadings_Y2 = p22 %*% t(v2)

library(knitr)

# Membuat data frame untuk muatan kanonik
muatan_kanonik <- data.frame(
  Variabel = c("imunisasi", "air", "dokter", "pengangguran", "rumah",
"pengeluaran"),
  Fungsi_Kanonik_1 = c(canonical_loadings_X1, canonical_loadings_Y1),
  Fungsi_Kanonik_2 = c(canonical_loadings_X2, canonical_loadings_Y2)
)

# Menampilkan data frame
print(muatan_kanonik)
```

Dataset kesehatan.xlsx

	imunisasi	air	dokter
ACEH	49.72	89.74	40.17
SUMATERA UTARA	66.94	92.19	37.68
SUMATERA BARAT	55.64	85.59	47.61
RIAU	65.03	90.47	39.33
JAMBI	57.62	80.02	34.88
SUMATERA SELATAN	67.7	87.19	31.59
BENGKULU	68.88	73.08	41.2
LAMPUNG	75.63	82.78	26.56
KEP. BANGKA BELITUNG	72.27	81.64	45.4
KEP. RIAU	75.1	92.1	50.35
DKI JAKARTA	70.48	99.42	61.67
JAWA BARAT	70.58	93.86	29.48
JAWA TENGAH	72.7	93.76	46.19
DI YOGYAKARTA	81.32	96.69	56.98
JAWA TIMUR	75.26	96.01	43.68
BANTEN	65.29	92.95	37.99
BALI	79.97	98.31	70.43
NUSA TENGGARA BARAT	73.87	96.03	30.25
NUSA TENGGARA TIMUR	75.47	88.35	27.41
KALIMANTAN BARAT	64.34	82.08	25.24
KALIMANTAN TENGAH	65.45	77.72	23.11
KALIMANTAN SELATAN	66.93	76.29	36.54
KALIMANTAN TIMUR	68.99	87.9	42.45
KALIMANTAN UTARA	69.91	90.19	41.83
SULAWESI UTARA	72.05	94.37	55.91
SULAWESI TENGAH	69.08	86.85	32.89
SULAWESI SELATAN	70.03	92.12	42.22
SULAWESI TENGGARA	69.34	94.8	24.59
GORONTALO	67.1	96	53
SULAWESI BARAT	64.43	79.86	27.05
MALUKU	71.8	92.98	16.9
MALUKU UTARA	63.9	89.01	24.21
PAPUA BARAT	68.03	81.57	28.91
PAPUA	69.31	66.49	29.11

Dataset sosioekonomi.xlsx

	pengangguran	rumah	pengeluaran
ACEH	6.03	84.12	1225976
SUMATERA UTARA	5.89	71.46	1305339
SUMATERA BARAT	5.94	72.61	1411823
RIAU	4.23	77.56	1527549
JAMBI	4.53	87.28	1424125
SUMATERA SELATAN	4.11	84.71	1209986
BENGKULU	3.42	88.38	1332558
LAMPUNG	4.23	92.4	1203017
KEP. BANGKA BELITUNG	4.56	88.65	1727550
KEP. RIAU	6.8	72.97	1989703
DKI JAKARTA	6.53	56.57	2791716
JAWA BARAT	7.44	83.38	1567666
JAWA TENGAH	5.13	91.05	1209906
DI YOGYAKARTA	3.69	86.43	1731560
JAWA TIMUR	4.88	90.92	1323486
BANTEN	7.52	85.67	1743687
BALI	2.69	85.24	1741523
NUSA TENGGARA BARAT	2.8	91.35	1260820
NUSA TENGGARA TIMUR	3.14	90.74	961372
KALIMANTAN BARAT	5.05	91.43	1345552
KALIMANTAN TENGAH	4.1	81.92	1525785
KALIMANTAN SELATAN	4.31	83.09	1457344
KALIMANTAN TIMUR	5.31	75.14	1980275
KALIMANTAN UTARA	4.01	76.63	1693577
SULAWESI UTARA	6.1	79.47	1315176
SULAWESI TENGAH	2.95	88.44	1173679
SULAWESI SELATAN	4.33	87.68	1252551
SULAWESI TENGGARA	3.15	90.54	1172739
GORONTALO	3.06	85.14	1228893
SULAWESI BARAT	2.27	93.35	1036520
MALUKU	6.31	84.69	1238170
MALUKU UTARA	4.31	90.26	1317159
PAPUA BARAT	5.38	82.94	1598254
PAPUA	2.67	85.31	1509992