

## K-means clustering analysis and multiple linear regression model on household income in Malaysia

Gan Pei Yee<sup>1</sup>, Mohd Saifullah Rusiman<sup>1</sup>, Shuhaida Ismail<sup>1</sup>, Suparman<sup>2</sup>, Firdaus Mohamad Hamzah<sup>3</sup>,  
Muhammad Ammar Shafi<sup>1</sup>

<sup>1</sup>Faculty of Applied Sciences and Technology, Universiti Tun Hussien Onn Malaysia, UTHM Pagoh, Malaysia

<sup>2</sup>Faculty of Teacher Training and Education, Ahmad Dahlan University (UAD), Yogyakarta, Indonesia

<sup>3</sup>Centre For Defence Foundation Studies, Universiti Pertahanan Nasional Malaysia, Kuala Lumpur, Malaysia

### Article Info

#### Article history:

Received Sep 8, 2021

Revised Sep 27, 2022

Accepted Oct 27, 2022

#### Keywords:

Household income

K-means clustering

Mean square error

Multiple linear regression

Silhouette analysis

### ABSTRACT

Household income plays a significant role in determining a country's socioeconomic standing. This measure is often used by the government to formulate the federal budget and policies that are most appropriate for national development. In spite of this, Malaysia's current economic circumstances continue to be characterized by income disparity. Therefore, this shortcoming can be addressed by analyzing the household income survey (HIS) conducted by Department of Statistics Malaysia (DOSM). In this study, the hybrid model is proposed where K-means and multiple linear regression (MLR) for clustering and predicting household income in Malaysia. Based on the experimental results, the K-means clustering analysis in conjunction with the MLR model outperformed the MLR model without clustering with a smaller mean square error. As a result, clustering analysis results in a more accurate estimate of household income because it reduces the variation between households. It is important that household income information reflect the concern of policymakers about the impact of universal and targeted interventions on different socioeconomic groups.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



### Corresponding Author:

Mohd Saifullah Rusiman

Department of Mathematics and Statistics, Universiti Tun Hussien Onn Malaysia

Pagoh, 84600 Muar, Johor, Malaysia

Email: saifulah@uthm.edu.my

## 1. INTRODUCTION

The term household income refers to all earnings of the household or individual family members, regardless of whether they are in the form of monetary or in-kind goods and services, which are available for current consumption annually or more frequently [1]. There are several components of household income, including employment income, whether employed or self-employment, other earned income, property income as well as current transfers [2]–[4]. It is generally accepted that household income is the most significant indicator of economic well-being as it helps measure a household's resources for saving and consumption [5], [6].

In addition to measuring a citizen's socioeconomic status, household income also plays an important role in determining what policies should be implemented to promote national development [7]. The standard of living of an area can be evaluated based on this economic indicator. Moreover, household income was studied to determine whether policymakers were successful in addressing Malaysian economic inequality [8], [9]. The solution is particularly helpful since it ensures that poverty can be overcome effectively in a particular area. Policymakers will evaluate the impact of universal and targeted actions on different socioeconomic groups based on every piece of household income information. Policy issues involving

welfare, taxation, housing, education, labour market, health and other fiscal policies are influenced by data related to income distribution [6].

Furthermore, Malaysians also suffer from income inequality, influenced by racial backgrounds, geographical areas, and various other factors. Affluent Malaysians and the rest of the country often earn different amounts, with the gap widening each year. It is imperative that unequal income distribution be addressed as soon as possible since it undermines social cohesion and provides insufficient quality of life levels for Malaysians. Moreover, it is inconsistent with our national development initiative to promote growth with equity [1]. Therefore, modelling household income is duly necessary for policymakers to formulate an appropriate policy based on the socioeconomic status in Malaysia.

This study is about applying K-means in a multiple linear regression (MLR) model toward household income in Malaysia. This study aims to identify the factors that affect a household's income. Lastly, the MLR model with and without K-means technique analysis will be compared using the mean square error (MSE) to find a better model. Other previous studies were done to extend the MLR method with the fuzzy regression technique in worldwide according to various fields of studies [10]–[13].

## 2. PROPOSED METHOD

Multiple linear regression (MLR) is a popular method for analyzing multivariate factors [14]–[16]. Combining MLR with other methods can enhance the robustness and accuracy of the model. Previous study uses K-means clustering to divide the data into several disjoint groups that exhibit similar characteristics. This study uses K-means with a MLR model to cluster and predict household income in Malaysia. The purpose of this study is to cluster the factors that affect the income of a household. With K-mean clustering, it is possible to identify the significant factors that affect the income of a household. After identifying the number of clusters, the MLR analysis will be conducted based on the clusters. With the combination of MLR and K-means clustering, the MSE errors are expected to be minimized. Finally, the MSE of the MLR models with and without the K-means technique will be compared in order to find a better model.

## 3. RESEARCH METHOD

### 3.1. Data acquisition

The household income survey (HIS) conducted by the Department of Statistics Malaysia (DOSM) in 2012 provided detailed demographic and social information about each household. There are 22 variables included in the dataset, such as income category, education level, number of household members. A regression model for household income was formed using 12 variables based on their importance based on the MLR applied to the dataset as Table 1.

Table 1. Data description for household income

Variables	Name of variable	Variable type	Variables	Name of variable	Variable type
$Y$	gross total	Numeric	$X_7$	head of household certificate	Ordinal
$X_1$	Strata	Nominal	$X_8$	head of household activity	Nominal
$X_2$	Weightdp	Numeric	$X_9$	size of household	Numeric
$X_3$	head of household age	Numeric	$X_{10}$	Region	Nominal
$X_4$	head of household gender	Nominal	$X_{11}$	occupation of head household	Nominal
$X_5$	head of household marital status	Nominal	$X_{12}$	industry of head household	Nominal
$X_6$	head of household education	Ordinal			

### 3.2. Research methodology

#### 3.2.1. Multiple linear regression (MLR) model

The MLR model for dependent variable of  $Y$  with  $k$  predictor variables can be written as in (1),

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i \quad (1)$$

where  $Y_i$  is the dependent variable,  $\beta_0, \beta_1, \dots, \beta_k$  are coefficients of regression to be estimated with respect to observations,  $X_{i1}, \dots, X_{ik}$  are explanatory variables and  $\varepsilon_i$  is the error term [17].

It would be convenient to express the MLR model in matrix notation [18]. In matrix notation, the model given in (2). The least-squares estimator  $\hat{\beta}$  that minimizes the sum of squared residuals in MLR is expressed in (3).

$$Y = X\beta + \varepsilon \quad (2)$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (3)$$

where  $Y$  is a  $n \times 1$  vector of observation,  $X$  is an  $n \times p$  matrix of the levels of the regressor variables and  $\beta$  is  $p \times 1$  vector of the regression coefficients.

Several key assumptions must be fulfilled before running MLR model. The assumptions are multivariate normality, multicollinearity, independence, and homogeneity of variance [19], [20]. The specified method of checking the assumptions of linear regression in this study is tabulated in Table 2.

Table 2. Method on evaluating assumptions of MLR model

Assumptions	Assessing method
Multivariate normality	Normal Probability-Probability plot
Multicollinearity	Variance Inflation Factor (VIF)
Independence	Durbin-Watson statistic
Homogeneity of variance	Box-Cox plot

Serious multicollinearity is indicated by a VIF value greater than 10. Multicollinearity can be solved by discarding the variables with the highest VIF, resulting in a model with little to no multicollinearity. The VIF is expressed in the formula in (4) [18],

$$VIF_k = \frac{1}{1 - R_k^2} \quad (4)$$

where  $R_k^2$  is the  $R^2$  value obtained by regressing the  $k^{th}$  predictor on the explanatory variables. Durbin-Watson's statistic of the model should approach a value of two to fulfill the model's independence. For the Box-Cox plot to determine homoscedasticity, the estimated rounded lambda value should be 1, which indicates constancy in error variance [18].

Other than that, the coefficient of multiple determination,  $R^2$  is another important indicator. This is because it not merely measures how well the model fits a set of observations, but also elaborates the variation amount of the dependent variable, which is explained by the regression equation.  $R^2$  can be simplified as the variation proportion in the response variable accounted by independent variables. The higher the  $R^2$ , the more variation amount in the dependent variable can be explained by the predictor variables. Thus, most of the observations will fall on the fitted regression line [20]. The formula of  $R^2$  is given as in (5).

$$R^2 = \frac{SSR}{SSTO} = \frac{\sum_i^n (\hat{Y}_i - \bar{Y})^2}{\sum_i^n (Y_i - \bar{Y})^2} \quad (5)$$

where SSR = Sum square of regression

SSTO = Sum square of total

$\hat{Y}_i$  = Fitted regression line

$\bar{Y}$  = Mean of  $Y$

$Y_i$  = Data of dependent variable

### 3.2.2. K-means clustering method

Meanwhile, K-means clustering analysis is a useful method to reduce error rates since it will classify the data observations to the nearest cluster based on minimum distance computed. Euclidean's distance is commonly applied in clustering analysis to classify observations. The distance between two objects,  $O_i$  and  $O_j$  in  $p$ -dimensional space are calculated using euclidean distance formulated as in (6) [21],

$$\text{Euclidean}(O_i, O_j) = \sqrt{\sum_{d=1}^p (O_{id} - O_{jd})^2} \quad (6)$$

where the  $i, j$  is the  $i^{th}$  and  $j^{th}$  data object and  $p$  is number of features. The centroid of the  $i^{th}$  cluster is defined as in (7) [22],

$$c_i = \frac{1}{m_i} \sum_{x \in C_i} x \quad (7)$$

where  $c_i$  is cluster centroid;  $m_i$  refers to objects number in  $i^{th}$  cluster;  $x$  is a data object and  $C_i$  is  $i^{th}$  cluster.

The K-means clustering method determines the optimal number of clusters using average silhouette width. The average silhouette width near to the value of 1 signifies the observations are well clustered. In K-means clustering analysis, the calculation of centroids and distance from the centroids and grouping of data observations are iterative until a convergence point is reached [22], [23].

### 3.3. Accuracy comparison method

#### 3.3.1. Mean square error (MSE)

MSE is found by calculating the average of the squared error. It measures the distance between the observed and actual value of estimator [18]. The formula is written as in (8) [24], [25],

$$MSE = \frac{1}{n} \sum_{t=0}^n (y_t - \hat{y}_t)^2 \quad (8)$$

where,  $y_t$  is the real data at time  $t$ ,  $\hat{y}_t$  is the predicted value at time  $t$  and  $n$  is the number of data involved.

If there are only 2 clusters used, the MSE of clustering can be calculated using the in (9) [24],

$$MSE_{combined} = \frac{n_1 MSE_1 + n_2 MSE_2}{n_1 + n_2} \quad (9)$$

where  $n_1$  and  $n_2$  are sample sizes in cluster 1 and 2 respectively;  $MSE_1$  and  $MSE_2$  are the mean square error for cluster 1 and 2 respectively.

## 4. RESULTS AND DISCUSSION

The model contains one dependent variable and 12 independent variables. Prior to data analysis, categorical independent variables were recoded into dummy variables. This dataset was assessed for its suitability for MLR based on its assumptions of linear regression. After that, it was found that the constancy of variance and multivariate normality were not fulfilled in these data by referring to the normal P-P plot and Box-Cox plot as shown in Figure 1. Since Box-Cox plot showed that the optimal lambda is -0.06, data transformation using power of -0.06 was applied simultaneously to dependent and independent variables. After performing Box-Cox transformation, the assumptions of linear regression were successfully achieved as Figure 2.

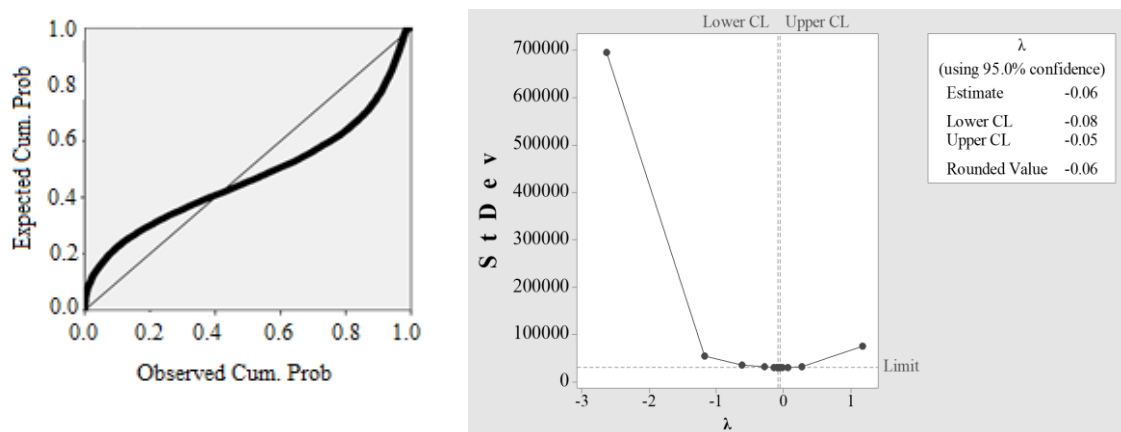


Figure 1. Normal P-P plot (left) and Box-Cox plot (right) of gross total standardized residuals

The model was first run using MLR model. The initial MLR model showed the problem of multicollinearity and non-significance of certain predictor variables. The checking procedure was first focused on multicollinearity and then evaluated predictor variables' significance. The multicollinearity problem is solved by discarding 2 variables ( $x_5$  and  $x_{10}$ ) where VIFF > 10. Besides, predictors that were not significant are also discarded since they do not contribute to the model. The MSE of the final model is the main concern of this study to see the effectiveness of the model in reducing error rates and variation of household income. The MLR final model consists of 10 significant explanatory variables and has  $MSE=3.08 \times 10^{-4}$ , as Table 3. The significant model is in (10).

$$\hat{Y} = 0.081 - 0.009 X_1 + 0.068 X_2 + 0.266 X_3 - 0.006 X_4 + 0.000069 X_6 + 0.000 X_7 + 0.004 X_8 + 0.194 X_9 + 0.002 X_{11} + 0.000 X_{12} \tag{10}$$

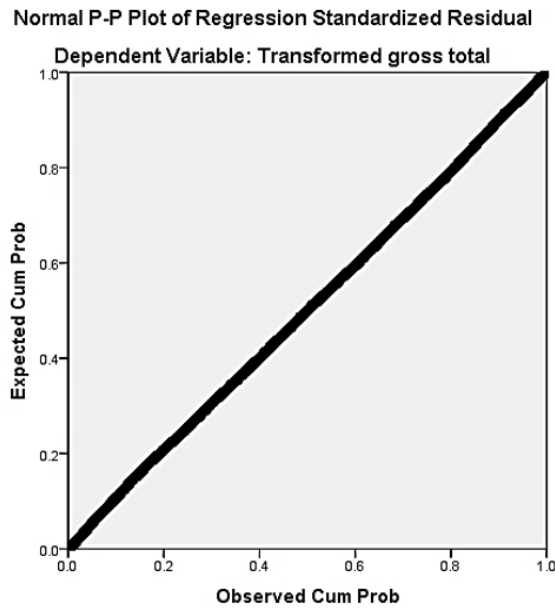


Figure 2. Normal P-P plot of the transformed gross total of standardized residuals

Table 3. Analysis of variance (ANOVA) table of MLR final model

Model	Sum of Squares	Df	Mean Square	F	Sig.
1 Regression	4.305	10	431	1398.008	0.0
Residual	4.072	13221	000		
Total	8.377	13231			

In the second step, K-means clustering analysis and the MLR model were combined. In this combination model, MSE value was used to assess whether this hybrid model provided an accurate prediction. The silhouette analysis was used to determine the optimal number of clusters. The output of silhouette analysis is tabulated as Table 4.

Table 4. Average silhouette width for various cluster number

Cluster number, k	Average Silhouette Width
2	0.65
3	0.44
4	0.55
5	0.50
6	0.53
7	0.54
8	0.54
9	0.48

Considering its average silhouette width is closest to one, the optimal number of clusters was identified as two clusters. When applying K-means clustering analysis, user can choose the variables of interest as clustering variables. In this study, clustering analysis was done based on (i) dependent variable and (ii) all variables. All the clustering variables were standardized first as the scale for variables was not the same. The MSE obtained from clustering based on the dependent variable is  $1.5811 \times 10^{-4}$ ; whereas MSE obtained from clustering based on all variables is  $2.9962 \times 10^{-4}$ . The comparison was then done on the MLR model with and without K-means clustering analysis as in Table 5.

Based on Table 6, for cluster 1 and 2, four independent variables were excluded (due to multicollinearity ( $x_5$  and  $x_{10}$ ) and nearly 0 of coefficient). Then the 8 independent variables are included as

in model in Table 6 for both clusters. The important contributors of household income for cluster 1 according to this model encompass 8 variables which are strata ( $X_1$ ), weight density population ( $X_2$ ), head of household age ( $X_3$ ), head of household gender ( $X_4$ ), head of household education ( $X_6$ ), head of household certificate ( $X_7$ ), head of household activity ( $X_8$ ) and household size ( $X_9$ ). Whereas for cluster 2, the 8 important contributors are strata ( $X_1$ ), weight density population ( $X_2$ ), head of household age ( $X_3$ ), head of household gender ( $X_4$ ), head of household certificate ( $X_7$ ), head of household activity ( $X_8$ ), household size ( $X_9$ ) and head of household occupation ( $X_{11}$ ).

Table 5. Comparison on performance of different models

Model	MLR only	MLR + K-Means Clustering (Dependent Variables)	MLR + K-Means Clustering (All Variables)
MSE (after transformation)	$3.0799 \times 10^{-4}$	$1.5811 \times 10^{-4}$	$2.9962 \times 10^{-4}$
MSE (transformed back to original)	$2.2605 \times 10^9$	$2.0053 \times 10^9$	$2.2498 \times 10^9$

Table 6. Summary model of K-means in MLR model

Cluster	Cluster 1	Cluster 2
MLR model	$\hat{Y} = 0.337 - 0.003 X_1 + 0.032 X_2 + 0.104 X_3 - 0.003 X_4 + 3.675E-5 X_6 + 0.001 X_7 + 0.002 X_8 + 0.114 X_9$	$\hat{Y} = 0.229 - 0.005 X_1 + 0.025 X_2 + 0.219 X_3 - 0.002 X_4 + 0.001 X_7 + 0.002 X_8 + 0.096 X_9 + 0.002 X_{11}$
MSE	$1.5740 \times 10^{-4}$	$1.5870 \times 10^{-4}$

The combination of K-means clustering analysis and MLR model is more effective in reducing error rates than the MLR model without K-means clustering analysis since it provides a smaller MSE. Since the model with the dependent variable as a clustering variable gives the smaller MSE, the significant predictors are extracted. The analysis also showed that the transformed gross total household income for cluster 1 is directly proportional to weight density population ( $X_2$ ), head of household age ( $X_3$ ), head of household education ( $X_6$ ), head of household certificate ( $X_7$ ), head of household activity ( $X_8$ ) and household size ( $X_9$ ). Meanwhile, the transformed gross total household income is inversely proportional to strata ( $X_1$ ) and head of household gender ( $X_4$ ). For cluster 2, the transformed gross total household income is directly proportional to weight density population ( $X_2$ ), head of household age ( $X_3$ ), head of household certificate ( $X_7$ ), head of household activity ( $X_8$ ), household size ( $X_9$ ) and occupation of head household ( $X_{11}$ ). Meanwhile, the transformed gross total household income is inversely proportional to strata ( $X_1$ ) and head of household gender ( $X_4$ ).

The contributors of household income, regardless of demographic factors or geographical restrictions can essentially affect the earnings of a household. Government can retrieve more information about this and formulate suitable policies so that all Malaysians can enjoy a high standard of living all the time. The information through statistical analysis is important to top management in decision making to optimize the economic situation.

## 5. CONCLUSION

Malaysian household income can be significantly influenced by geographic and demographic characteristics. Considering the income gap in Malaysia, it is crucial that the data be modeled using an appropriate technique to minimize the error rates arising from the income gap. MLR model can help to analyze the contributing factors effectively using statistical approach. However, to attain accurate and reliable results, clustering analysis such as the K-means approach can effectively reduce the variability in household income due to income gap by clustering the data before performing the MLR model. It is recommended to incorporate more potential contributors of household income into the model, for example expenditures, to make the model more reliable. Future researchers can also consider using other types of clustering techniques such as the fuzzy c-means approach to get the best model with the lowest error rates. Other than that, the researcher can rationally choose different variables as clustering variables where appropriate to obtain certain important discoveries.

## ACKNOWLEDGEMENTS




The research work is supported by Ministry of Higher Education, Malaysia with fundamental research grant scheme (FRGS) grant (Vot K297), reference number FRGS/1/2020/STG06/UTHM/02/4.

## REFERENCES




- [1] H. M. Z. RAGAYAH, "Income inequality in Malaysia," *Asian Economic Policy Review*, vol. 3, no. 1, pp. 114–132, Jun. 2008, doi: 10.1111/j.1748-3131.2008.00096.x.
- [2] R. Mahadevan, "Growth with equity: the Malaysian case," *Asia-Pacific Development Journal*, vol. 13, no. 1, pp. 27–52, Oct. 2006, doi: 10.18356/b3751b80-en.
- [3] S. Ishak and H. M. Z. Ragayah, "The patterns and trends of income distribution in Malaysia, 1970–1987," *Singapore Economic Review*, vol. 35, no. 1, pp. 102–123, 1990.
- [4] S. M. Hashim, *Income Inequality and Poverty in Malaysia*. Lanham, MD: Rowman & Littlefield Publishers, 1998.
- [5] K. Stratford and A. Cowling, *Chinese household income, consumption and savings*. RBA Bulletin, 2016.
- [6] *United Nations Economic Commission for Europe, Canberra Group Handbook on Household Income Statistics*. Geneva: United Nations, 2011.
- [7] "Australian Bureau of Statistics, Measures of Socioeconomic Status." 2011, [Online]. Available: [https://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/367D3800605DB064CA2578B60013445C/\\$File/1244055001\\_2011.pdf](https://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/367D3800605DB064CA2578B60013445C/$File/1244055001_2011.pdf).
- [8] M. Jakobsen, *Ethnic Chinese Entrepreneurship in Malaysia*. Routledge, 2014.
- [9] A. H. Roslan, "Income inequality, poverty and development policy in Malaysia," in *In International seminar on poverty and sustainable development, Université Montesquieu-Bordeaux IV and UNESCO*, 2001, pp. 22–23.
- [10] M. A. Shafi and M. S. Rusiman, "The use of fuzzy linear regression models for tumor size in colorectal cancer in hospital of Malaysia," *Applied Mathematical Sciences*, vol. 9, no. 53–56, pp. 2749–2759, 2015, doi: 10.12988/ams.2015.5175.
- [11] M. A. Shafi, M. S. Rusiman, S. Ismail, and M. G. Kamardan, "A hybrid of multiple linear regression clustering model with support vector machine for colorectal cancer tumor size prediction," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 4, pp. 323–328, 2019, doi: 10.14569/ijacsa.2019.0100439.
- [12] N. A. Samat and M. N. M. Salleh, "A study of data imputation using fuzzy c-means with particle swarm optimization," *Advances in Intelligent Systems and Computing*, vol. 549 AISC, pp. 91–100, 2017, doi: 10.1007/978-3-319-51281-5\_10.
- [13] K. Hussain and M. N. Mohd. Salleh, "Analysis of techniques for ANFIS rule-base minimization and accuracy maximization," *ARPN Journal of Engineering and Applied Sciences*, vol. 10, no. 20, pp. 9739–9746, 2015, doi: 10.13140/RG.2.1.1324.5525.
- [14] I. Mohamed, K. Khalid, and M. S. Yahya, "Combined estimating function for random coefficient models with correlated errors," *Communications in Statistics - Theory and Methods*, vol. 45, no. 4, pp. 967–975, 2016, doi: 10.1080/03610926.2013.853794.
- [15] J. Neter, W. Wasserman, and M. H. Kutner, *Applied Linear Regression Models*. USA: Richard D. Irwin, Inc., 1983.
- [16] A. Agresti, *An introduction to categorical data analysis*. Hoboken, NJ: Wiley-Interscience, 2007.
- [17] S. S. Amiri, M. Mottahedi, and S. Asadi, "Using multiple regression analysis to develop energy consumption indicators for commercial buildings in the U.S.," *Energy and Buildings*, vol. 109, pp. 209–216, Dec. 2015, doi: 10.1016/j.enbuild.2015.09.073.
- [18] W. C. Navidi, *Statistics for engineers and scientists*. New York: McGraw-Hill, 2015.
- [19] K. Rani Das and A. H. M. R. Imon, "A brief review of tests for normality," *American Journal of Theoretical and Applied Statistics*, vol. 5, no. 1, pp. 5–12, 2016, doi: 10.11648/j.ajtas.20160501.12.
- [20] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx, "Regression: Models, methods and applications," *Regression: Models, Methods and Applications*, vol. 9783642343, pp. 1–698, 2013, doi: 10.1007/978-3-642-34333-9.
- [21] R. S. King, *Cluster analysis and data mining: an introduction*. Boston, MA, 2015.
- [22] X. Jin and J. Han, *K-means clustering*. in: *sammur C., webb G.I. (eds) encyclopedia of machine learning*. Boston: Springer, 2011.
- [23] M. Z. Hossain, M. N. Akhtar, R. B. Ahmad, and M. Rahman, "A dynamic K-means clustering for data mining," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 2, p. 521, Feb. 2019, doi: 10.11591/ijeecs.v13.i2.pp521-526.
- [24] N. S. Wahab *et al.*, "A technique of fuzzy c-mean in multiple linear regression model toward paddy yield," *Journal of Physics: Conference Series*, vol. 995, p. 12010, Apr. 2018, doi: 10.1088/1742-6596/995/1/012010.
- [25] M. S. Rusiman, S. N. M. Nor, Suparman, and S. N. A. M. Razali, "Robust method in multiple linear regression model on diabetes patients," *Mathematics and Statistics*, vol. 8, no. 2, pp. 36–39, 2020, doi: 10.13189/ms.2020.081306.

## BIOGRAPHIES OF AUTHORS







**Gan Pei Yee**    holds a Bachelor of Science degree in Industrial Statistics from the University Tun Hussein Onn Malaysia (UTHM). In her last semester, she completed her senior project. It's Malaysian research that focuses on clustering analysis using a multiple linear regression model for family income. She is presently employed as an IT Support at KS Leow Business Solutions PLT, which is a job in the statistics sector. She can be contacted at email: ganyee25@yahoo.com.







**Assoc. Prof. Dr. Mohd Saifullah Rusiman**    is a lecturer with 17 years of teaching experience at the University Tun Hussein Onn Malaysia (UTHM) under the Department of Mathematics and Statistics. He got Ph. D in mathematics at the University Teknologi Malaysia (UTM) and Dokuz Eylul University (DEU), Turkey. His area of research interest was focused on applied statistics specifically statistics modeling, fuzzy statistics, and time series. He has published more than 90 articles where 52 of it has been indexed by scopus. He is also Chief Editor for the international journal of Enhanced Knowledge of Science and Technology (EKST) which was discovered in 2020. He can be contacted at email: saifulah@uthm.edu.my.



**Ts. Dr. Shuhaida Ismail**     is a lecturer at the Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia (UTHM). She obtained her PhD in Mathematics from Universiti Teknologi Malaysia. Her research interest is in Machine Learning, specifically in predictive modelling, classification and clustering. Her current research areas are in Hydrological Modelling, Analytics and Deep Learning. She can be contacted at email: shuhaida@uthm.edu.my.







**Assoc. Prof. Dr. Suparman**     is a lecturer at the Department of Mathematics Education, Ahmad Dahlan University (UAD), Indonesia. He obtained a Doctor of Philosophy in Applied Mathematics at Paul Sabatier University, France. His research interests are focused on applied statistics, especially reversible jump MCMC, bootstrap, simulated annealing, and its application to time series, signal processing, and finance. Scopus ID: 6507107541. He can be contacted at email: suparman@pmat.uad.ac.id.



**Assoc. Prof. Dr. Firdaus Mohamad Hamzah**     has been a lecturer at the Faculty of Engineering & Built Environment in the Universiti Kebangsaan Malaysia (UKM) since 2000. He obtained his BSc (Statistics) in 1998 and MSc (Quality Management and Productivity) in 1999 from UKM. He was conferred a PhD (Environmental Statistics) from the University of Glasgow, United Kingdom in 2012. He is currently a Senior Research Fellow at the Institute of Climate Change, UKM. He has published 50 journal articles, more than 90 proceeding papers, several books and technical papers. He can be contacted at email: firdaus.hamzah@upnm.edu.my.



**Dr. Muhammad Ammar Shafi**     is a lecturer with 1 year of teaching experience at the University Tun Hussein Onn Malaysia (UTHM). He is currently teaching under the Department of Management and Technology. He majored in statistics at the University Tun Hussein Onn Malaysia (UTHM), where he also earned his Doctor of Philosophy in Statistics. His area of research interest was focused on applied statistics specifically statistics modeling, fuzzy statistics, and applied statistics area. He can be contacted at email: ammar26121991@gmail.com.