



ANALYSIS OF THE EFFECTS OF DISCOUNTS AND PROMOTIONS ON IMPULSE BUYING

Auzan Azinuddin Wazidiatama¹, Ananda Deva Dwi Jaya², Aldha Nurhaliza Putri³, Hera Aprilia Noradinda⁴, Tarra Afifah Zahra Gaida⁵, Yusuf Java Abimadia⁶

¹²³UPN “Veteran” Jawa Timur, ⁴Universitas Muhammadiyah Sidoarjo, ⁵Universitas Negeri Surabaya, ⁶Universitas Islam Negeri Sunan Ampel Surabaya
25083010058@student.upnjatim.ac.id

ABSTRACT

This study aims to analyze the effects of discount and promotional strategies on impulse buying tendencies. The development of consumer behavior, particularly on digital platforms such as Shopee, has made impulse buying an increasingly relevant phenomenon to investigate, especially in relation to price incentives offered by sellers. To achieve this objective, this study employs a quantitative approach by analyzing data from 150 respondents. The analytical methods include Principal Component Analysis (PCA) to reduce ten independent variables into a simpler set of components and Principal Component Regression (PCR) to examine the effects of these components on impulse buying tendencies.

The results indicate that PCA successfully reduced the dimensionality of the data. The first principal component (PC1) explained 33.49% of the total variance, while the first two components (PC1 and PC2) jointly explained 49.01%. By including the sixth component (PC6), the cumulative explained variance reached 83.48%, indicating that most of the information in the data could be represented by six principal components. In the model diagnostics, 10 out of 150 respondents had Cook's Distance values exceeding the established threshold, indicating the presence of several influential observations that should be considered when interpreting the stability of the regression model. The PCR model was statistically significant overall; however, it explained only about 24-27% of the variation in impulse buying tendency ($R^2 = 0.2729$ and adjusted $R^2 = 0.2424$), and the mean R^2 from 5-fold cross-validation was only 0.1361. Thus, discount and promotional variables showed a statistically significant but modest relationship with impulse buying tendencies, and the limited predictive performance as well as the presence of influential observations should be taken into account when interpreting the model results.

Keywords: *Principal Component Analysis, Principal Component Regression, Discount, Promotion, E-Commerce*

I. INTRODUCTION

1.1 Background

The main issue addressed in this study concerns Q15, which represents impulse buying tendency and serves as the dependent variable. The independent indicators Q1–Q10 represent two conceptual factors: discounts (X_1) and promotions (X_2). Specifically, Q1–Q5 measures consumers' responses to discount-related factors, including their interest in discounted products, preferred discount levels, discount-seeking behavior, encountered discount percentages, and purchasing frequency influenced by discounts. Meanwhile, Q6–Q10 measures responses to promotional activities, including the influence of promotions on purchasing decisions, reactions to limited-time offers, and the use of free shipping, vouchers, and other promotional offers. These ten indicators are expected to be associated with Q15, which measures the frequency of unplanned purchases during the previous month. However, the use of multiple indicators may increase model complexity and raise concerns about multicollinearity.

To address this issue, this study uses Principal Component Analysis (PCA) for dimensionality reduction. The principal

* Corresponding author.

E-mail address:

<http://doi.org/10.33005/jasid.v2i2.77>

components obtained from PCA are then used in Principal Component Regression (PCR) to examine their relationship with Q15. This approach reduces the number of predictors and evaluates how well the selected principal components explain the variation in impulse buying tendency.

1.2 Research Questions

Based on the background described above, this study addresses the following research questions:

1. What is the relationship between variables Q1–Q10 (promotions and discounts) and Q15 (impulse buying tendency)?
2. How can multicollinearity arising from the large number of independent variables in the regression model be addressed?
3. How well can the PCR model explain the variation in Q15 using the principal components obtained from PCA?

1.3 Research Objectives

The objectives of this study are to:

1. Identify the relationships and effects of variables Q1–Q10 (promotions and discounts) on Q15 (impulse buying tendency).
2. Reduce the dimensionality of the data using PCA to address multicollinearity.
3. Evaluate the ability of the PCR model to explain the variation in Q15 using the selected principal components.

II. MATERIALS AND METHODS

2.1 Research Data and Variables

This study uses primary data collected through an online questionnaire distributed via Google Forms. The questionnaire was designed to obtain information about university students' shopping behavior related to discounts, promotions, and impulse buying. Initially, 166 responses were collected. After data cleaning to remove duplicate records, incomplete responses, and inconsistent entries, 150 observations were retained for analysis.

The study examines two conceptual independent variables: discounts (X_1) and promotions (X_2), with impulse buying tendency (Y) as the dependent variable. The questionnaire consisted of 15 items coded Q1–Q15. Discount (X_1) was represented by Q1–Q5, which covered interest in discounted products, preferred discount levels, discount-seeking behavior, encountered discount percentages, and purchase frequency influenced by discounts. Promotion (X_2) was represented by Q6–Q10, which covered the influence of promotions on purchasing decisions, responses to time-limited offers, and the use of free shipping, vouchers, and other promotional offers. The dependent variable (Y) was represented by Q15, which measured the frequency of unplanned purchases during the previous month.

Although the questionnaire included Q11–Q14, these items were excluded from the PCA and PCR models. Instead, Q1–Q10 were used as predictor variables, while Q15 served as the response variable. Before PCA, the ten predictor variables were standardized using the Z-score method to ensure that differences in measurement scales did not disproportionately influence the formation of principal components. Table 2.1 presents the research variables and their corresponding questionnaire items.

Table 2.1 Research Variables and Question Codes

Variable	Question Code	Description
Discount (X_1)	Q1	How interested are you in purchasing products when discounts are available?
Discount (X_1)	Q2	What discount amount or level do you find most attractive?
Discount (X_1)	Q3	How many times per month do you search for discounted products?

Discount (X_1)	Q4	What is the average discount percentage you encounter when shopping?
Discount (X_1)	Q5	How many times have you purchased products because of discounts in the past month?
Promotion (X_2)	Q6	To what extent do promotions influence your purchasing decisions?
Promotion (X_2)	Q7	When you see a time-limited promotion, how do you tend to make a purchasing decision?
Promotion (X_2)	Q8	How many times per month do you use free shipping offers?
Promotion (X_2)	Q9	How many times per month do you use vouchers?
Promotion (X_2)	Q10	How many times have you used promotions when shopping in the past month?
Impulse Buying Tendency (Y)	Q15	How many times have you made unplanned purchases in the past month?

2.2 Descriptive Analysis

Descriptive analysis provides an overview of the characteristics of the data used in a study. The data are summarized and presented systematically so that their patterns, distributions, and characteristics can be understood without hypothesis testing or generalization beyond the observed data. The resulting information may include means, medians, modes, frequencies, and proportions, presented in tables or graphs. In this study, descriptive analysis was used to summarize the data before modeling with Principal Component Regression (PCR). The cleaned dataset contained 150 observations suitable for analysis. Ten independent variables (X) derived from questionnaire items related to discounts and promotions were used. The dependent variable (Y) was Q15, which measures the frequency of impulse purchases during the previous month. Q15 was selected in line with the research objective of identifying the effects of discount and promotional factors on impulse buying tendency. The effects of the ten independent variables on the dependent variable were then analyzed using PCR. PCR was selected because forming principal components can address potential multicollinearity among independent variables before regression is performed.

2.3 Pearson Correlation

This study uses Pearson correlation to measure linear relationships between numerical variables Q1–Q15. Pearson's Product-Moment Correlation is a parametric statistical method used to analyze the relationship between two variables measured on an interval or ratio scale. The Pearson correlation coefficient (r) ranges from -1 to $+1$ [10]. A positive value indicates a relationship in the same direction, whereas a negative value indicates an inverse relationship. A value of zero indicates no linear correlation in the data. To obtain reliable correlation results, the relevant statistical assumptions should be considered. These assumptions are as follows:

- The two variables analyzed should be measured on an interval or ratio scale.
- The data should be normally distributed.
- The relationship should be linear rather than nonlinear.
- The data should be free of extreme outliers that could distort the calculation.
- The data should have homogeneous variance.

If the assumptions of Pearson correlation are not satisfied, Spearman's rank correlation may be a more suitable alternative, particularly when the data do not meet the normality assumption [3]. The Pearson correlation coefficient is

calculated by comparing the covariance of two variables with the product of their standard deviations, using the following formula:

$$r_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

Where:

r_{xy} = validity between variables X and Y

N = sample size or number of observations

X = value of the independent variable being analyzed

Y = value of the dependent variable being analyzed

2.4 Multicollinearity Test

The multicollinearity test is used to identify high correlations among independent variables in a regression model [6]. If the independent variables are highly correlated, the model is considered to exhibit multicollinearity. Ideally, a regression model should not contain problematic correlations among its independent variables.

In this study, multicollinearity was assessed using Tolerance (TOL) and the Variance Inflation Factor (VIF). A VIF value of 10 or less is used as a criterion for identifying the absence of serious multicollinearity in the regression model [6]. If multicollinearity is detected, several approaches may be used to address it:

1. Increase the sample size.
2. Remove one or more independent variables.
3. Combine time-series and cross-sectional data.
4. Transform the data.
5. Apply principal component regression.

In this study, Principal Component Regression (PCR) was selected because it can reduce dimensionality and address multicollinearity without removing independent variables.

2.5 Data Standardization

Data standardization is a transformation process used to place variables on comparable scales, allowing them to be compared directly. It reduces the influence of differences in units or value ranges so that variables can contribute on a comparable basis. In this study, standardization was performed before applying PCR to ensure that differences in the scales or measurement units of the independent variables would not dominate the formation of principal components. This helps each variable contribute on a balanced basis to the analysis. Standardization was performed using the Z-score method, which converts each observation to a standardized value with a mean of 0 and a standard deviation of 1. The standardized value is calculated using the following equation:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

X = original data value

μ = mean of all observations.

σ = standard deviation of all observations

2.6 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction method that transforms potentially correlated variables into a smaller set of principal components while retaining as much of the original data variance as possible [8]. In this study, PCA was used for dimensionality reduction and to address potential multicollinearity among the independent variables. It also aims to represent the information contained in ten independent variables through a smaller number of principal components that explain most of the data variance. The components are formed as linear combinations of the standardized independent variables. The first component (PC1) explains the greatest amount of variance, followed by the second component (PC2), and so on. The components are mutually orthogonal and therefore uncorrelated. The number of components retained was determined using cumulative explained variance. In this study, components explaining at least 80% of the total variance were retained. These components were then used as predictors in PCR.

2.7 Interpretation of PCA Loadings

PCA loadings are coefficients that indicate the contribution of, or relationship between, a variable and a principal component. They are used to identify variables that play a dominant role in forming each component. A larger absolute

loading indicates a greater contribution of the variable to the corresponding component. The positive or negative sign of a loading indicates the direction of the relationship. In this study, variables with loading values above 0.5 were treated as dominant indicators.

2.8 Principal Component Regression

Principal Component Regression (PCR) combines Principal Component Analysis (PCA) with regression analysis and can be used to address multicollinearity in an OLS model [5]. The procedure involves several stages, including data standardization, the formation of principal components based on eigenvalues (λ_i) and eigenvectors (a_j), and the selection of principal components (K_j) [1]. The selected components are then used as predictors in the regression model to explain the dependent variable, resulting in the following PCR model:

$$y = w_0 + w_1K_1 + w_2K_2 + \dots + w_mK_m$$

Where:

w_0 = intercept

$w_1, w_2, \dots, w_m$ = regression coefficients

$K_1, K_2, \dots, K_m$ = principal components

2.9 Model Testing

The F-test determines whether all independent variables included in the regression model are jointly significant predictors of the dependent variable. If the p-value is below 0.05 or the calculated F-statistic exceeds the critical F-value, the null hypothesis is rejected, indicating a statistically significant effect. If the p-value is above 0.05 or the calculated F-statistic is below the critical F-value, the null hypothesis is not rejected, indicating that a statistically significant effect has not been established (Aryani, 2020). The F-statistic is calculated using the following formula:

$$F = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}$$

Where:

R^2 = coefficient of determination

n = number of observations

k = number of independent variables

F = F-test statistic

The t-test (partial test) is used to determine whether an individual independent variable has a statistically significant relationship with the dependent variable. If the p-value is below 0.05 or the absolute calculated t-statistic exceeds the critical t-value, the null hypothesis is rejected, indicating a statistically significant effect. Conversely, if the p-value is above 0.05 or the absolute calculated t-statistic is below the critical t-value, the null hypothesis is not rejected (Aryani & Gustian, 2020). The t-statistic is calculated using the following formula:

$$t = \frac{\beta_i}{SE(\beta_i)}$$

Where:

β_i = regression coefficient for the i-th variable.

$SE(\beta_i)$ = standard error.

2.10 Classical Assumption Tests and Diagnostics

Classical assumption tests and model diagnostics are conducted to assess whether the regression model meets the relevant statistical assumptions, allowing its parameter estimates to be interpreted appropriately. In this study, the tests and diagnostics include residual normality, heteroskedasticity, and Cook's Distance analysis.

1) Residual Normality Test

The residual normality test determines whether the regression residuals or errors follow a normal distribution. The normality assumption is relevant to the validity of statistical inference in regression. In this study, residual normality was assessed using the Jarque–Bera (JB) test. The test compares the Jarque–Bera statistic with the critical value from the chi-square (χ^2) distribution. Residuals are considered consistent with normality when the Jarque–Bera statistic is less than or equal to the critical χ^2 value. The Jarque–Bera statistic is calculated from the skewness and kurtosis of the model residuals [6].

2) Heteroskedasticity Test

The heteroskedasticity test determines whether the residual variance differs across observations. A well-specified regression model is generally expected to have constant residual variance, or homoskedasticity. In this study, heteroskedasticity was assessed using the Breusch–Pagan–Godfrey (BPG) test. The test detects heteroskedasticity by

comparing the calculated χ^2 statistic with the critical χ^2 value. The model is considered to exhibit heteroskedasticity when the calculated χ^2 statistic exceeds the critical value. Conversely, if the calculated χ^2 statistic is less than or equal to the critical value, the model is considered not to show evidence of heteroskedasticity under this criterion [6].

3) Cook's Distance Analysis

Cook's Distance is a model diagnostic used to identify observations that have a substantial influence on the fitted regression model. A high Cook's Distance value indicates that an observation may strongly affect the estimated model parameters and may change the regression results. Therefore, Cook's Distance is used to identify potentially influential observations that may affect the stability and reliability of the fitted model.

2.11 Model Validation

Regression evaluation metrics are used to assess how well a regression model predicts values from the available data. They compare predicted values with actual values to quantify prediction errors and the model's ability to explain the data. Common metrics for evaluating regression performance include the following:

1) Coefficient of Determination (R-squared)

The coefficient of determination, or R-squared (R^2), measures the proportion of variation in the dependent variable that can be explained by the independent variables in a regression model. R^2 indicates how well the model explains the observed data. A higher R^2 generally indicates greater explanatory power. An R^2 of 1 represents a perfect fit, whereas an R^2 of 0 indicates that the model does not explain the variation in the data relative to the mean-based benchmark. R^2 may also be negative when predictions are worse than those obtained by using the mean of the observed data. In general, the coefficient of determination can be interpreted as the proportion of variance in the dependent variable explained by the independent variables [9].

2) Prediction Error Evaluation

- a) Mean Absolute Error (MAE) measures the average absolute difference between predicted and actual values. It quantifies the size of prediction errors without considering their direction. A lower MAE indicates that predictions are, on average, closer to the actual values. MAE is also less sensitive to outliers than squared-error metrics, providing a general measure of model performance [9].
- b) Mean Squared Error (MSE) measures the average squared difference between predicted and actual values. Squaring the errors gives larger errors more weight than smaller ones. Therefore, MSE is sensitive to outliers and can reflect predictions with large errors. A lower MSE indicates better regression performance [9].
- c) Root Mean Squared Error (RMSE) is the square root of MSE and measures prediction error in the same units as the original data. RMSE is easier to interpret than MSE because it is expressed on the same scale as the predicted variable. A lower RMSE indicates smaller prediction errors and predictions that are closer to actual values [9].

3) K-Fold Cross-Validation

K-fold cross-validation is a model validation technique used to evaluate model performance and reduce the risk of overfitting. It also helps assess how well a model generalizes to new data. In K-fold cross-validation, the dataset is divided into K subsets, or folds, of approximately equal size. One fold is used for validation while the remaining folds are used for training. This process is repeated K times so that each fold serves once as the validation set. The final evaluation metrics are calculated by averaging the results across all folds, providing a more stable and representative estimate of model performance.

2.12 Flowchart

The stages of this study, from data collection to the conclusion, are summarized in the flowchart in Figure 2.1.

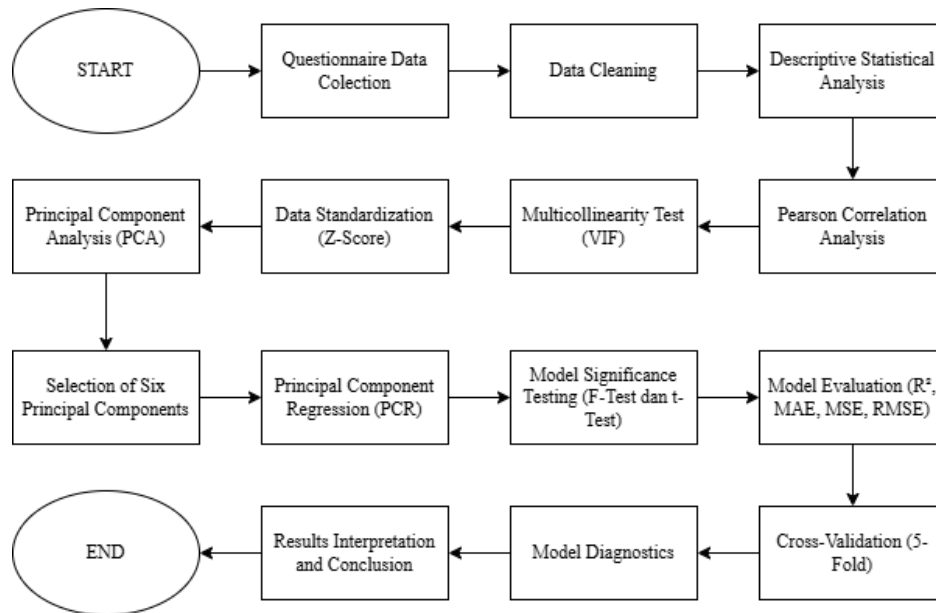


Figure 2.1 Research Flowchart

III. RESULT AND DISCUSSION

3.1 Descriptive Statistics

This study analyzed 150 respondents. The independent variables consisted of ten variables (Q1–Q10), while Q15 served as the dependent variable, in accordance with the study objectives and research questions. The analysis examined the relationships between the independent variables and the dependent variable using Principal Component Regression (PCR). Descriptive statistics are presented in Table 3.1 to summarize the minimum, maximum, mean, median, and standard deviation of each variable.

Table 3.1 Descriptive Statistics

Variable	Mean	Standard Deviation	Minimum	Median	Maximum
Q1	3.866667	1.127250	1	4.0	5
Q2	55.006667	21.388299	15	50.0	100
Q3	7.066667	5.430051	0	5.0	20
Q4	22.906667	9.518154	5	20.0	45
Q5	4.713333	3.711648	0	3.5	13
Q6	3.893333	0.949367	2	4.0	5
Q7	3.200000	0.843451	1	3.0	5

Q8	10.640000	7.964080	0	9.5	30
Q9	8.086667	5.463653	0	7.0	19
Q10	5.380000	3.560918	0	5.0	13
Q15	3.920000	3.532248	0	3.0	11

As shown in Table 3.1, the mean responses for Q1–Q10 generally fall within the middle to higher ranges, with considerable variation, indicating differences in respondents' perceptions of discounts and promotions. Q15 has a mean of 3.92 and a standard deviation of 3.53, suggesting a moderate level of impulse buying tendency but substantial differences among respondents. This variability provides a basis for further analysis using PCA and PCR.

3.2 Pearson Correlation

The initial stage of PCR involves examining relationships among the independent variables using a Pearson correlation matrix. This analysis identifies patterns of association among variables that may indicate potential multicollinearity and inform the use of PCA. The Pearson correlation results are presented in Table 3.2.

Table 3.2 Pearson Correlation Matrix

Variable	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Q1	1.000	0.003	0.303	-0.049	0.236	0.432	0.296	0.129	0.176	0.172
Q2	0.003	1.000	0.049	0.284	0.023	0.129	0.160	0.114	0.129	0.203
Q3	0.303	0.049	1.000	-0.049	0.557	0.300	0.274	0.407	0.398	0.316
Q4	-0.049	0.284	-0.049	1.000	0.112	0.153	0.136	0.083	0.055	0.114
Q5	0.236	0.023	0.557	0.112	1.000	0.146	0.278	0.493	0.467	0.439
Q6	0.432	0.129	0.300	0.153	0.146	1.000	0.496	0.046	0.113	0.189
Q7	0.296	0.160	0.274	0.136	0.278	0.496	1.000	0.113	0.174	0.269
Q8	0.129	0.114	0.407	0.083	0.493	0.046	0.113	1.000	0.636	0.485
Q9	0.176	0.129	0.398	0.055	0.467	0.113	0.174	0.636	1.000	0.564
Q10	0.172	0.203	0.316	0.114	0.439	0.189	0.269	0.485	0.564	1.000

Table 3.2 shows that most variables have positive relationships, with varying correlation strengths. The highest correlation is between Q8 and Q9 (0.636), indicating a strong positive relationship. Other moderate to strong correlations include Q9 and Q10 (0.564) and Q3 and Q5 (0.557). After examining the correlations among independent variables Q1–Q10, the next step was to analyze the relationship between each independent variable and the dependent variable (Y). This analysis assessed the strength of each variable's relationship with impulse buying tendency and identified variables significantly associated with the dependent variable. The Pearson correlation results between the independent variables and the dependent variable are presented in Table 3.3.

Table 3.3 Pearson Correlation of X with Y

Variable X	Pearson r	p-value	Significance
Q1	0.159	0.0518	No
Q2	-0.007	0.9296	No
Q3	0.344	0.0000	Yes
Q4	0.110	0.1804	No
Q5	0.388	0.0000	Yes
Q6	0.190	0.0202	Yes
Q7	0.244	0.0026	Yes
Q8	0.403	0.0000	Yes
Q9	0.420	0.0000	Yes
Q10	0.338	0.0000	Yes

Table 3.3 shows that seven variables (Q3, Q5, Q6, Q7, Q8, Q9, and Q10) are significantly associated with impulse buying tendency (Q15). Q9 has the strongest correlation with Q15 ($r = 0.420$), followed by Q8 ($r = 0.403$) and Q5 ($r = 0.388$). Specifically, Q5 measures the frequency of purchases made because of discounts during the previous month. Its positive correlation suggests that respondents who more frequently purchase items because of discounts tend to report higher impulse buying frequency. However, such purchases are not necessarily impulsive, as some may have been planned in advance. Therefore, the correlation indicates an association rather than a causal effect.

3.3 Multicollinearity Test Results

The multicollinearity test examines whether independent variables are highly correlated in the regression model. In this study, multicollinearity was assessed using the Variance Inflation Factor (VIF). The results are presented in Table 3.4.

Table 3.4 Multicollinearity Test Results (VIF)

Variable	VIF	Conclusion
Q1	1.329	No multicollinearity detected
Q2	1.156	No multicollinearity detected
Q3	1.718	No multicollinearity detected
Q4	1.177	No multicollinearity detected
Q5	1.871	No multicollinearity detected
Q6	1.622	No multicollinearity detected
Q7	1.456	No multicollinearity detected
Q8	1.929	No multicollinearity detected
Q9	2.028	No multicollinearity detected
Q10	1.684	No multicollinearity detected

All independent variables in Table 3.4 have VIF values below 10. Based on this criterion, the results do not indicate multicollinearity in the regression model. Therefore, all independent variables can be considered for further analysis without evidence of problematic high correlation among them.

3.4 Principal Component Analysis Results

PCA was performed to reduce the dimensionality of the data by transforming the independent variables into principal components. The variables were standardized beforehand because PCA is sensitive to differences in scale. Table 3.5 presents the percentage of variance explained by each component and the cumulative explained variance.

Table 3.5 Principal Component Analysis Results

PC	Variance	Cumulative
PC1	33.49%	33.49%
PC2	15.52%	49.01%
PC3	13.31%	62.32%
PC4	7.74%	70.05%
PC5	6.77%	76.82%
PC6	6.66%	83.48%
PC7	5.08%	88.55%
PC8	4.64%	93.19%
PC9	3.49%	96.68%
PC10	3.32%	100.00%

According to Table 3.5, the first principal component (PC1) explains 33.49% of the data variance, while the first two components (PC1 and PC2) jointly explain 49.01%. By PC6, the cumulative explained variance reaches 83.48%, indicating that most of the information in the data can be represented by six principal components. Thus, PCA reduces the ten independent variables to a smaller set of components while retaining most of the data variability.

3.5 Interpretation of Principal Components

After selecting six principal components that explain 83.48% of the total variance, the components were interpreted based on their loading values. Larger absolute loadings indicate stronger relationships between the original variables and the corresponding principal components. Table 3.6 presents the loadings for PC1–PC6.

Table 3.6 Loadings for PC1–PC6

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Q1	0.2495	0.3700	-0.3371	-0.2027	-0.1586	0.6568
Q2	0.1313	0.1329	0.6098	-0.5383	0.5087	0.0924
Q3	0.3806	-0.0120	-0.2673	0.1579	0.6123	0.0019
Q4	0.0966	0.1560	0.6333	0.6144	-0.1627	0.3155
Q5	0.4035	-0.1588	-0.1064	0.4146	0.2520	-0.0433
Q6	0.2511	0.5743	-0.0537	-0.0170	-0.1201	0.0058
Q7	0.2840	0.4542	0.0241	0.0529	-0.0933	-0.6278
Q8	0.3822	-0.3709	0.0499	-0.0242	-0.0873	0.1387
Q9	0.4028	-0.3098	0.0397	-0.1817	-0.2557	0.0386
Q10	0.3892	-0.1615	0.1519	-0.2409	-0.3937	-0.2092

According to Table 3.6, Q5 has the largest loading on PC1 (0.4035), followed closely by Q9 (0.4028), Q10 (0.3892), Q8 (0.3822), and Q3 (0.3806). However, all these loadings are below the 0.5 threshold used in this study to identify dominant indicators. Therefore, PC1 should not be interpreted as representing a single variable. Instead, it reflects a broader pattern of consumer engagement with discount-related and promotional shopping activities. In practical terms, this component summarizes shared variation across several shopping behaviors that may be relevant to impulse buying. However, the loading values alone do not establish a causal relationship with impulse buying; the relationship between the principal components and Q15 is examined through PCR.

Table 3.7 Variables with the Largest Loading in Each Principal Component

Principal Component	Variable	Loading Value	Absolute Loading Value
PC1	Q5	0.4035	0.4035
PC2	Q6	0.5743	0.5743
PC3	Q4	0.6333	0.6333
PC4	Q4	0.6144	0.6144
PC5	Q3	0.6123	0.6123
PC6	Q1	0.6568	0.6568

For the other components, Q6 has the largest absolute loading on PC2 (0.5743), Q4 on PC3 (0.6333) and PC4 (0.6144), Q3 on PC5 (0.6123), and Q1 on PC6 (0.6568). These variables have absolute loading values above 0.5 and can therefore be identified as the dominant indicators of their respective components according to the criterion adopted in this study. Table 3.7 summarizes the variables with the largest absolute loadings for each selected principal component.

3.6 Principal Component Regression Results

3.6.1 OLS Model Summary

The OLS regression summary for the PCR analysis provides an overview of the fitted model. Table 3.8 presents the estimation method, dependent variable, number of observations, number of principal components, and key model statistics.

Table 3.8 OLS Model Summary

Statistic	Value		
Estimation Method	Ordinary	Least	Squares
	(OLS)		
Dependent Variable	Q15		
Number of Observations	150		
Number of Principal Components	6		
R-squared	0.2729		
Adjusted R-squared	0.2424		
OLS F-statistic	8.9450		

OLS Prob (F-statistic) 2.71×10^{-8}

As shown in Table 3.8, the PCR model uses six principal components obtained from PCA, with Q15 as the dependent variable and 150 observations. The R-squared value of 0.2729 and adjusted R-squared value of 0.2424 indicate that the model explains approximately 24–27% of the variation in the dependent variable. The F-statistic of 8.9450 and its probability of 2.71×10^{-8} indicate that the regression model is statistically significant overall. These results show that PCR can be used to analyze the relationship between the principal components and impulse buying tendency.

3.6.2 OLS Regression Coefficients

PCR was performed using the six principal components obtained from PCA as independent variables. The parameters were estimated using Ordinary Least Squares (OLS). Table 3.9 presents the regression coefficients for each principal component.

Table 3.9 OLS Regression Coefficients

Variable	OLS Coefficient
Constant	3.9200
PC1	0.9545
PC2	-0.2380
PC3	-0.0999
PC4	0.4397
PC5	-0.3107
PC6	-0.1498

According to Table 3.9, PC1 has the largest positive coefficient (0.9545) in the model for Q15. PC2, PC3, PC5, and PC6 have negative coefficients, indicating a negative relationship with impulse buying tendency in the estimated model. The PCR equation is formed as a combination of the six principal components with an intercept of 3.9200. Overall, these results indicate that PCR can address multicollinearity and produce a relatively interpretable model.

3.6.3 PCR Regression Equation

The PCR analysis produced a regression equation describing the relationship between the principal components (PC1 through PC6) and the response variable Q15. The equation, based on the estimated intercept and regression coefficients, is presented below:

$$Q15 = 3.9200 + 0.9545PC1 - 0.2380PC2 - 0.0999PC3 + 0.4397PC4 - 0.3107PC5 - 0.1498PC6$$

The regression equation shows the relationship between the six principal components and impulse buying (Q15). The intercept of 3.9200 represents the predicted Q15 value when all principal component scores are zero. PC1 and PC4 have positive coefficients of 0.9545 and 0.4397, respectively, indicating a positive relationship with the predicted Q15 value. Meanwhile, PC2, PC3, PC5, and PC6 have negative coefficients of -0.2380, -0.0999, -0.3107, and -0.1498, respectively. However, based on the partial t-test results in Table 3.11, only PC1 is statistically significant at the 5% significance level, while the other components are not significant. These results indicate that PC1 is the only component with a statistically significant relationship with impulse buying in the fitted PCR model.

3.7 Model Significance

3.7.1 F-test (Simultaneous)

The F-test assesses the overall significance of the regression model by determining whether the independent variables jointly have a statistically significant relationship with the dependent variable. The results are presented in Table 3.10.

Test Component	Result
Robust HC3 F-statistic	7.3451
p-value	7.36×10^{-7}
Significance Level (α)	0.05
Decision	Reject H_0
Conclusion	The model is significant

As shown in Table 3.10, the robust HC3 F-statistic is 7.3451, with a p-value of 7.36×10^{-7} , which is below the 0.05 significance level. Therefore, H_0 is rejected and the regression model is statistically significant overall. This indicates that the principal components obtained from PCA are jointly associated with Q15.

3.7.2 t-test (Partial)

The t-test assesses the statistical significance of each principal component's individual relationship with the dependent variable. The results are presented in Table 3.11.

Variable	Coefficient	t-statistic	p-value	Decision
PC1	0.9545	6.391	< 0.001	Significant
PC2	-0.2380	-1.060	0.291	Not significant
PC3	-0.0999	-0.376	0.708	Not significant
PC4	0.4397	1.458	0.147	Not significant
PC5	-0.3107	-0.985	0.326	Not significant
PC6	-0.1498	-0.590	0.556	Not significant

According to Table 3.11, only PC1 shows a statistically significant positive relationship with impulse buying tendency ($\beta = 0.9545$, $t = 6.391$, $p < 0.001$), based on HC3 robust standard errors. In contrast, PC2–PC6 are not statistically significant at the 5% level. These results indicate that PC1 is the only principal component significantly associated with Q15 in the fitted PCR model.

The positive and significant coefficient of PC1 is broadly consistent with Anggarwati et al. [11], who reported positive and significant relationships of price discounts and promotions with impulse buying among Shopee consumers. This agreement suggests that shopping patterns involving discounts and promotional activities are relevant to understanding impulse buying tendency. However, the findings are not directly equivalent: Anggarwati et al. examined discounts and promotions through path analysis, whereas the present study used PCA to combine ten indicators into principal components before regression. Therefore, the significant PC1 coefficient reflects the relationship between a combined pattern of indicators and impulse buying, rather than the isolated effect of discounts or promotions.

3.8 Model Evaluation

3.8.1 Coefficient of Determination

The coefficient of determination is used to assess the model's ability to explain variation in the dependent variable. The results are presented in Table 3.12.

Table 3.12 Coefficient of Determination

Metric	Value
R-squared	0.2729
Adjusted R-squared	0.2424

Table 3.12 shows an R-squared value of 0.2729 and an adjusted R-squared value of 0.2424, indicating that the PCR model explains approximately 24–27% of the variation in Q15. Although the model is statistically significant, other factors outside the principal components included in the model may account for the remaining variation in impulse buying tendency.

3.8.2 Prediction Error Evaluation

Prediction error evaluation measures the accuracy of the model's predictions for the dependent variable. The results are presented in Table 3.13.

Table 3.13 Prediction Error Evaluation Results

Metric	Value
MAE	2.3846
MSE	9.0113
RMSE	3.0019

Table 3.13 reports an MAE of 2.3846, an MSE of 9.0113, and an RMSE of 3.0019. These values describe the prediction errors of the model. The RMSE being relatively close to the MAE suggests that the two error metrics are of a similar magnitude; however, the acceptability of the errors should be judged in the context of the scale of Q15 and in comparison with alternative models.

3.8.3 Cross-Validation Evaluation

Cross-validation was used to evaluate model consistency and generalization to different data subsets. The results of 5-fold cross-validation are presented in Table 3.14.

Table 3.14 Cross-Validation Results

Metric	Mean 5-Fold Cross-Validation
R ²	0.1361
MAE	2.5363
RMSE	3.1786

Table 3.14 shows a mean R² of 0.1361 across the five validation folds, indicating that the model's explanatory performance is lower on validation data. The MAE of 2.5363 and RMSE of 3.1786 are higher than the corresponding values reported for the fitted model, indicating limitations in generalization. Thus, although the model is statistically significant, its predictive performance may still need improvement.

3.9 Model Diagnostics

3.9.1 Residual Normality Test

The residual normality test assesses whether the regression errors follow a normal distribution. The results are presented in Table 3.15.

Table 3.15 Residual Normality Test Results

Testing Method	p-value	Decision $\alpha = 0.05$
Jarque-Bera	0.00534	Reject H_0

Table 3.15 reports a Jarque–Bera p-value of 0.00534, which is below the 0.05 significance level. Therefore, H_0 is rejected, indicating that the residuals are not normally distributed. OLS regression may still be used, but the violation of normality should be considered when interpreting statistical inference.

3.9.2 Heteroskedasticity Test

The heteroskedasticity test determines whether the residual variance differs across observations. This study uses the Breusch–Pagan method with a significance level of 0.05. The hypotheses are:

H_0 : There is no heteroskedasticity; the residual variance is constant (homoskedasticity).

H_1 : Heteroskedasticity is present; the residual variance is not constant.

The test results are presented in Table 3.16.

Table 3.16 Heteroskedasticity Test Results (Breusch–Pagan)

Test Statistic	Value
LM Statistic	11.362228
LM p-value	0.077806
F-statistic	1.9532290
F p-value	0.076231
Significance Level	0.05
Decision	Fail to Reject H_0

Table 3.16 reports an LM statistic of 11.362228 with an LM p-value of 0.077806. The F-statistic is 1.9532290, with an F p-value of 0.076231. Both p-values exceed the 0.05 significance level, so the null hypothesis is not rejected. Therefore, there is insufficient statistical evidence to conclude that heteroskedasticity is present in the regression model.

3.9.3 Cook's Distance Diagnostics

Cook's Distance was used to identify observations that may have a substantial influence on the stability of the regression model. Cook's Distance was calculated for each observation, using a threshold of 0.0267. The results are shown in Table 3.17.

Table 3.17 Cook's Distance Diagnostic Results

No.	Observation Index	Cook's Distance
1	0	0.050381
2	54	0.036439
3	55	0.035832
4	72	0.042283
5	77	0.029544
6	81	0.105273
7	83	0.052067
8	127	0.053352
9	140	0.062211
10	142	0.026867

Table 3.17 shows several observations with Cook's Distance values above the threshold of 0.0267. This indicates that some observations may have a substantial influence on the regression model and should be considered when interpreting the results.

3.9.4 Cook's Distance Summary

For ease of interpretation, a summary of the Cook's Distance analysis is presented in Table 3.18.

Table 3.18 Summary of Cook's Distance Results

Indicator	Value
Number of Observations	150
Cook's Distance Threshold (4/n)	0.0267
Number of Observations Above Threshold	10
Maximum Cook's Distance	0.105273
Observation Index with Maximum Cook's Distance	81

According to Table 3.18, 10 of the 150 observations have Cook's Distance values above the threshold of 0.0267. The maximum Cook's Distance value is 0.105273, found at observation index 81. This indicates that although most observations are not excessively influential, several observations may affect the stability of the model. Therefore, the regression results should be interpreted with these influential observations in mind.

IV. CONCLUSION

Pearson correlation analysis showed that seven variables (Q3, Q5, Q6, Q7, Q8, Q9, and Q10) were significantly associated with Q15 at the 5% significance level. Q9 had the strongest correlation with Q15 (0.420), followed by Q8 (0.403) and Q5 (0.388). Meanwhile, Q1, Q2, and Q4 were not significantly associated with Q15. These findings indicate that several variables related to discounts and promotions were associated with impulse buying tendency, although the strength of these relationships varied across variables.

Regarding multicollinearity, all independent variables had VIF values below 10, with the highest value being 2.028 for Q9. Therefore, no serious multicollinearity was indicated based on the criterion used in this study, and no specific corrective action was required. Nevertheless, PCA was applied to reduce the dimensionality of the ten independent variables. Six principal components were retained, cumulatively explaining 83.48% of the total variance, and were subsequently used as predictors in the PCR model.

The PCR model achieved an R-squared value of 0.2729 and an adjusted R-squared value of 0.2424, indicating that the six principal components explained roughly 24-27% of the variation in impulse buying tendency, so most of the variation remained unexplained. The robust HC3 F-test showed that the model was statistically significant overall, with a p-value of 7.36×10^{-7} , but statistical significance should not be equated with strong explanatory power. Only PC1 showed a statistically significant positive relationship with Q15 in the partial t-tests, while PC2-PC6 were not significant at the 5% level. Furthermore, 5-fold cross-validation produced a mean R^2 of only 0.1361, about half of the in-sample R-squared, together with higher prediction errors (MAE 2.5363 and RMSE 3.1786) than those of the fitted model (MAE 2.3846 and RMSE 3.0019), indicating limited generalization to new data. Therefore, the results indicate a statistically significant but modest relationship between the discount- and promotion-related variables and impulse buying tendency, and the model should not be regarded as a strong or reliable predictor. Other factors outside the variable studied may account for much of the remaining variation. Overall, PCA successfully reduced the dimensionality of the data, while PCR produced a statistically significant model with limited explanatory and predictive performance.

REFERENCES

- [1] R. Fadhilah, D. E. Putri, and H. S. Amanda, "Faktor-faktor yang mempengaruhi stunting di Indonesia: Analisis principal component regression," *Jurnal Gaussian*, vol. 14, no. 1, pp. 128–138, 2025, doi: 10.14710/j.gauss.14.1.128-138.
- [2] D. R. P. Sari, "Metode principal component analysis (PCA) sebagai penanganan asumsi multikolinearitas (studi kasus: Data produksi tapioka)," *Parameter: Jurnal Matematika, Statistika dan Terapannya*, vol. 2, no. 2, pp. 115–124, 2023, doi: 10.30598/parameter.2i02pp115-124.
- [3] P. Patimah, Zulpan, D. U. Alfansuri, E. Munawaroh, and M. Ilyas, "Memahami dan menerapkan uji korelasi dalam analisis data penelitian pendidikan," *Journal Education Innovation (JEI)*, vol. 3, no. 4, pp. 740–752, 2025, doi: 10.65474/f7vd8y11.
- [4] Tim Halo Ryan, "Mengenal standarisasi dalam statistik," *Halo Ryan*, Oct. 5, 2024. [Online]. Available: <https://haloryan.com/blog/mengenal-standarisasi-dalam-statistik>. [Accessed: Oct. 20, 2026].
- [5] A. Raihanah, C. P. Putri, P. Damayanti, M. Z. Meizilano, E. Sunandi, and Firdaus, "Perbandingan metode ridge regression dan principal component regression pada data PT Bumi Resources Tbk," *Jurnal Sains Data, Statistika, dan Artificial Intelligence*, vol. 2, no. 1, pp. 30–45, 2026. [Online]. Available: <https://journal.uici.ac.id/index.php/stardia/article/view/37>.
- [6] S. M. Sholihah, N. Y. Aditiya, E. S. Evani, and S. Maghfiroh, "Konsep uji asumsi klasik pada regresi linier berganda," *Jurnal Riset Akuntansi Soedirman*, vol. 2, no. 2, pp. 102–110, Dec. 2023, doi: 10.32424/1.jras.2023.2.2.10792.
- [7] Y. Aryani, "Sistem informasi penjualan barang dengan metode regresi linear berganda dalam prediksi pendapatan perusahaan," *Jurnal Riset Sistem Informasi dan Teknologi Informasi (JURSISTEKNI)*, vol. 2, no. 2, pp. 39–51, 2020, doi: 10.52005/jursistekni.v2i2.47.
- [8] IBM, "What is principal component analysis (PCA)?," Dec. 8, 2023. [Online]. Available: <https://www.ibm.com/think/topics/principal-component-analysis> [Accessed: Oct. 20, 2026].
- [9] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, Art. no. e623, 2021, doi: 10.7717/peerj-cs.623.
- [10] Fu, T., Tang, X., Cai, Z., Zuo, Y., Tang, Y., & Zhao, X. (2020). Correlation research of phase angle variation and coating performance by means of Pearson's correlation coefficient. *Progress in Organic Coatings*, 139, 105459.
- [11] D. Anggarwati, A. Hermawan, and W. P. Rahayu, "The effect of price discount and promotion on impulse buying through hedonic shopping motivation as an intervening variable on Shopee consumers," *LITERACY: International Scientific Journals of Social, Education, Humanities*, vol. 2, no. 1, pp. 34–49, 2023, doi: 10.56910/literacy.v2i1.411.