Project : Multivariate Analysis of Spotify most streamed Songs

a report authored by
Yuyutsu Saini
2020MT60571

supervised by
Prof. Rahul Singh

06/11/2024

MTL766 : Multivariate Statistic

Contents

1 Introduction

The dataset analyzed in this project encompasses user behavior and demographic information, providing a basis to study purchasing patterns across a diverse population. By examining attributes like age, annual income, purchase amount, and loyalty score, this analysis aims to reveal insights into factors driving customer loyalty and purchase frequency. Additionally, understanding the regional distribution of users allows for examining any geographical trends in purchasing behavior. This multivariate analysis will leverage statistical techniques to uncover relationships, detect outliers, and reduce dimensionality, offering a concise summary of user profiles and purchasing dynamics.

- **Dataset**: Composed of user demographic and purchasing behavior data, with 238 observations.
- **Purpose**: To analyze factors contributing to customer loyalty and purchasing frequency.
- **Relevance**: Insights may inform targeted marketing strategies and customer retention efforts.
- **Key Attributes**:
  - **Demographics**: age, region, annual_income.
  - **Behavior**: purchase_amount, loyalty_score, purchase_frequency.
- **Objectives**: To apply multivariate techniques, including EDA, covariance testing, and PCA, to explore and summarize customer profiles effectively.

2 Dataset Description

- **Total Observations**: The dataset contains 238 entries, divided into two main regions for focused analysis:
  - **North East Region**: 84 observations (includes North and East regions).
  - **South West Region**: 154 observations (includes South and West regions).
- **Features**: The dataset includes 7 features:
  - **Numerical Features** (5): *age*, *annual_income*, *purchase_amount*, *loyalty_score*, and *purchase_frequency*.
  - **Categorical Features** (2): *user_id* and *region*.
- **Region Distribution**:
  - **North**: 78 users
  - **South**: 77 users
  - **West**: 77 users
  - **East**: 6 users
- **Segregation Rationale**: The dataset is divided into North East (North + East) and South West (South + West) regions to facilitate intra- and inter-regional comparisons and to identify specific purchase and loyalty patterns.

3 Exploratory Data Analysis

Table 1 presents the descriptive statistics for each attribute in the customer dataset, including measures of central tendency, dispersion, and range.

Table 1: Summary Statistics of Customer Data

Statistic	user_id	age	annual_income	purchase_amount	loyalty_score	purchase_frequency
Count	238	238	238	238	238	238
Mean	119.50	38.68	57407.56	425.63	6.79	19.80
Std	68.85	9.35	11403.88	140.05	1.90	4.56
Min	1.00	22.00	30000.00	150.00	3.00	10.00
25%	60.25	31.00	50000.00	320.00	5.50	17.00
50%	119.50	39.00	59000.00	440.00	7.00	20.00
75%	178.75	46.75	66750.00	527.50	8.28	23.00
Max	238.00	55.00	75000.00	640.00	9.50	28.00

3.1 Relationship Between All Attributes

The plot below shows a pairwise plot illustrating the relationships between all attributes in the dataset.

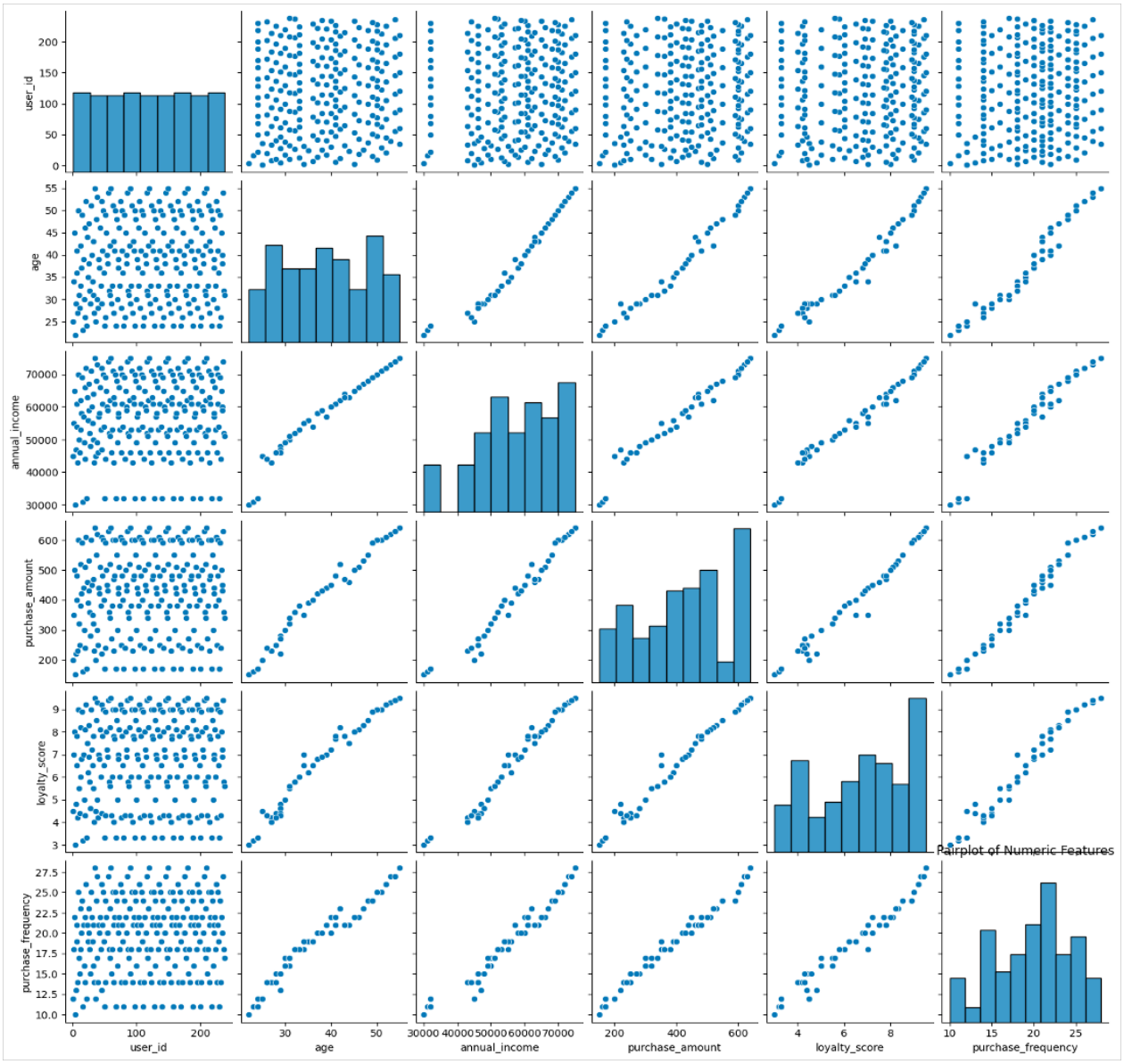
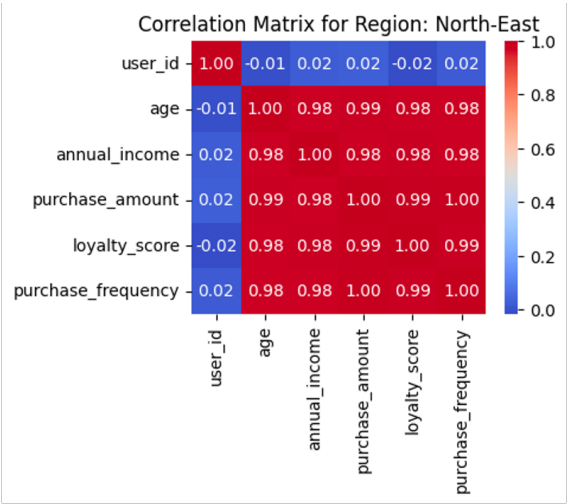


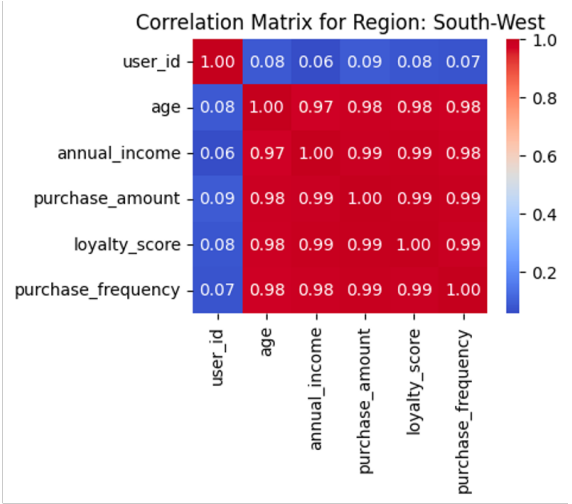
Figure 1: The Relationship Between All Attributes in the Data

3.2 Correlation of Both datasets

The following figure displays two plots representing the correlation matrices for each dataset, corresponding to the North-East and South-West regions.



(a) Correlation Plot 1



(b) Correlation Plot 2

Figure 2: Comparison of Correlation Plots

3.2.1 Observation

Based on the relationship between all attributes in the data table and the correlation matrix, we can draw the following observations:

- All attributes in the dataset appear to be strongly correlated with each other, indicating possible relationships or dependencies between these attributes.
- The ‘user_id‘ attribute stands out as a unique identifier for each entry in the dataset. It is not correlated with any other attribute, highlighting its role solely as an identifier.
- The strong correlations suggest that certain attributes may share underlying patterns or influence each other, which could be relevant for predictive modeling or clustering tasks.
- Future analysis could focus on understanding the causal relationships or exploring feature reduction techniques to address potential multicollinearity.

4 Comparison of Covariance Matrices

4.1 Box M’s Test

Box’s M test is a statistical test used to determine if multiple groups have equal covariance matrices. This test is commonly used to check the assumption of homogeneity of variances, which is important in multivariate analyses like discriminant analysis and MANOVA.

4.1.1 Purpose of Box’s M Test

The primary purpose of Box’s M test is to assess whether the variance-covariance matrices of different groups are significantly different from each other. A small p-value (typically less than 0.05) suggests that there are significant differences in the covariance matrices, meaning that the assumption of homogeneity of variances may not hold.

4.1.2 How to Perform Box's M Test

To perform Box's M test:

1. Compute the pooled covariance matrix by averaging the covariance matrices of each group, weighted by the sample sizes.
2. Calculate the determinant of each group's covariance matrix as well as the determinant of the pooled covariance matrix.
3. Use these determinants to compute the Box's M statistic using the formula below.
4. Apply a correction factor to adjust for sample size differences and calculate the degrees of freedom for interpreting the Box's M statistic.

4.1.3 Equation for Box's M Test

The Box's M statistic is calculated as follows:

$$M = (N - g) \cdot \ln(\det \mathbf{S}_{\text{pooled}}) - \sum_{i=1}^{g} (n_i - 1) \cdot \ln(\det \mathbf{S}_i)$$

where:

- N is the total sample size,
- g is the number of groups,
- n_i is the sample size of group i,
- $\mathbf{S}_{\text{pooled}}$ is the pooled covariance matrix,
- $\mathbf{S}_i$ is the covariance matrix of group i,
- det represents the determinant of a matrix.

4.1.4 Correction Factor

The correction factor is used to adjust the Box's M statistic for sample size differences between groups and is given by:

$$\text{correction\_factor} = 1 - \frac{(2p^2 + 3p - 1)}{6(p + 1)(g - 1)} \left(\sum_{i=1}^{g} \frac{1}{n_i - 1} - \frac{1}{N - g} \right)$$

where:

- p is the number of variables,
- k is the number of groups,
- n_1 and n_2 are the sample sizes of the two groups.

4.1.5 Degrees of Freedom

The degrees of freedom (dof) for Box's M test is calculated as:

$$\text{dof} = \frac{p(p+1)}{2} \cdot (k-1)$$

where:

- p is the number of variables,
- k is the number of groups.

4.1.6 Calculating the p-value

To determine the significance of the Box's M statistic, use the chi-squared distribution with the calculated degrees of freedom:

$$\text{df} = \frac{p(p+1)}{2} \cdot (k-1)$$

The resulting p-value indicates whether the covariance matrices differ significantly. A small p-value (e.g., $p < 0.05$) suggests that the assumption of homogeneity of variances is violated.

4.1.7 Box's M Test Results

Statistic	**Value**
M Statistic	100.776012
Degrees of Freedom (for chi-square conversion)	15
P-value	9.297145×10^{-15}

Table 2: Box's M Test Results

4.1.8 Observation

From the results of Box's M test, we observe that the p-value is extremely small (9.297145×10^{-15}). This indicates strong evidence against the null hypothesis, which states that the covariance matrices of the groups are equal. Since the p-value is much smaller than the common significance level of 0.05, we reject the null hypothesis.

The rejection of the null hypothesis suggests that the covariance matrices of the different groups are significantly different from each other. Therefore, the assumption of homogeneity of variances does not hold, implying that the groups have unequal covariance structures.

4.2 Likelihood Ratio Test for Covariance Matrices

4.2.1 Purpose of Likelihood Ratio Test

The Likelihood Ratio Test (LRT) is a statistical test used to compare the fit of two models: one under the null hypothesis and one under the alternative hypothesis. In the context of covariance matrices, the test is used to determine whether the covariance matrices of two datasets are equal. If the null hypothesis is rejected, it implies that the covariance matrices are significantly different between the two datasets.

4.2.2 How to Perform the Test

To perform the Likelihood Ratio Test for testing the equality of covariance matrices, we need the sample covariance matrices for the two datasets. The test is based on the ratio of the determinants of the covariance matrices, comparing the likelihood of the two models:

1. Compute the sample covariance matrices S_1 and S_2 for the two datasets. 2. Compute the pooled covariance matrix S_{pooled}. 3. Calculate the Likelihood Ratio Statistic Λ.

4.2.3 Equation

The Likelihood Ratio statistic Λ is given by:

$$\Lambda = \frac{|S_1|^{n_1} \, |S_2|^{n_2}}{|S_{\text{pooled}}|^{n_1+n_2}}$$

Where: - $|S_1|$ and $|S_2|$ are the determinants of the covariance matrices of the two groups. - $|S_{\text{pooled}}|$ is the determinant of the pooled covariance matrix. - n_1 and n_2 are the sample sizes for the two datasets.

The test statistic Λ follows a chi-square distribution under the null hypothesis. We calculate the transformed statistic $-2\ln(\Lambda)$, which follows a chi-square distribution with degrees of freedom:

$$-2\ln(\Lambda) \sim \chi^2_{df}$$

4.2.4 Degree of Freedom

The degrees of freedom *df* for the Likelihood Ratio Test are calculated as:

$$df = \frac{p(p+1)}{2}$$

Where p is the number of variables (attributes) in the dataset.

4.2.5 Calculate P-value

To calculate the p-value, we compare the computed statistic $-2\ln(\Lambda)$ to the chi-square distribution with the corresponding degrees of freedom. The p-value represents the probability of obtaining a test statistic at least as extreme as the one calculated, assuming the null hypothesis is true.

4.2.6 Results

For this test, the computed results are as follows:

$$\hat{\Lambda} = 3.499576610159108 \times 10^{-23}$$

$$\text{Degrees of Freedom} = 13$$

$$\text{P-value} = 3.330669 \times 10^{-16}$$

Statistic	Value
Wilks' Lambda ($\hat{\Lambda}$)	$3.499576610159108 \times 10^{-23}$
Degrees of Freedom	13
P-value	3.330669×10^{-16}

Table 3: Likelihood Ratio Test Results

4.2.7 Observation

From the results of the Likelihood Ratio Test, we observe that the p-value is extremely small (3.330669×10^{-16}). This provides strong evidence against the null hypothesis, which asserts that the covariance matrices of the two datasets are equal. Since the p-value is much smaller than the typical significance level of 0.05, we reject the null hypothesis.

Thus, the rejection of the null hypothesis suggests that the covariance matrices of the two datasets are significantly different. Therefore, we conclude that the assumption of equal covariance matrices does not hold.

5 Comparison of Mean

5.1 Welch's T-test for Equality of Means

5.1.1 Purpose of Welch's T-test

The Welch's T-test is a statistical test used to determine whether the means of two groups are significantly different from each other. Unlike the Student's T-test, Welch's T-test does not assume equal variances between the two groups, making it more suitable when the assumption of equal variances is violated. It is commonly used when comparing the means of two datasets, particularly when the sample sizes and variances are unequal.

5.1.2 How to Perform the Test

To perform Welch's T-test for each attribute, the following steps are involved:

1. Calculate the sample means $\bar{X}_1$ and $\bar{X}_2$, and the sample variances S_1^2 and S_2^2 for the two groups. 2. Compute the test statistic for each attribute, which is given by the formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Where: - $\bar{X}_1$ and $\bar{X}_2$ are the sample means of the two groups. - S_1^2 and S_2^2 are the sample variances of the two groups. - n_1 and n_2 are the sample sizes of the two groups.

3. The degrees of freedom for Welch's T-test are calculated as:

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

4. Compare the calculated t-statistic to the critical value from the t-distribution with the calculated degrees of freedom, and obtain the p-value.

5.1.3 Equation

The formula for the t-statistic is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

The formula for degrees of freedom is:

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

5.1.4 Calculate P-value

The p-value is calculated by comparing the calculated t-statistic with the t-distribution with the corresponding degrees of freedom. The p-value indicates the probability of obtaining a t-statistic at least as extreme as the one calculated, assuming the null hypothesis (that the means are equal) is true.

5.1.5 Results

For each attribute, the t-statistic and p-value are as follows:

Attribute	**T-statistic**	**P-value**
Age	-0.8109	0.4643
Annual Income	-0.0414	0.9690
Purchase Amount	0.0898	0.9329
Loyalty Score	0.1208	0.9108
Purchase Frequency	0.2500	0.8194

Table 4: Results of Welch's T-test for each Attribute

5.1.6 Observation

From the results of the Welch's T-test, we observe that the p-values for all attributes are significantly greater than the typical significance level of 0.05. Specifically, the p-values range from 0.4643 (for Age) to 0.9690 (for Annual Income). Since all the p-values are greater than 0.05, we fail to reject the null hypothesis for all attributes.

Therefore, we conclude that there is no significant difference in the means of the two datasets for any of the attributes tested. The assumption that the means of the attributes are equal cannot be rejected based on the data.

5.2 James's Test for Comparison of Means

5.2.1 Purpose of James's Test

James's Test is used to compare the means of two groups when the covariance matrices are assumed to be unequal. It is particularly useful when the sample sizes are small and the assumption of equal

covariance matrices is questionable. The test statistic is a quadratic form involving the difference between the sample means, weighted by the inverse of the covariance matrices, and follows an approximate F-distribution under the null hypothesis.

5.2.2 How to Perform James's Test

To perform James's test, follow these steps:

1. Compute the difference between the means of the two groups, denoted by $\bar{x}_1$ and $\bar{x}_2$, as $\Delta\bar{x} = \bar{x}_1 - \bar{x}_2$.

2. Compute the pooled covariance matrix or the individual covariance matrices S_1 and S_2 for each group.

3. Calculate the James's Test Statistic using the formula:

$$J = (\Delta\bar{x})^T \left(S_1 \frac{1}{n_1} + S_2 \frac{1}{n_2} \right)^{-1} (\Delta\bar{x})$$

Where: - $\Delta\bar{x} = \bar{x}_1 - \bar{x}_2$ is the difference between the sample means. - S_1 and S_2 are the covariance matrices for the two groups. - n_1 and n_2 are the sample sizes of the two groups.

4. Calculate the critical F-value using the F-distribution with degrees of freedom based on the number of variables and the sample sizes.

5. Compute the p-value by comparing the test statistic to the critical F-value. A p-value smaller than the chosen significance level (typically 0.05) indicates rejection of the null hypothesis.

5.2.3 Equation

The formula for James's test statistic is:

$$J = (\bar{x}_1 - \bar{x}_2)^T \left(S_1 \frac{1}{n_1} + S_2 \frac{1}{n_2} \right)^{-1} (\bar{x}_1 - \bar{x}_2)$$

Where: - $\bar{x}_1$ and $\bar{x}_2$ are the sample means of the two groups. - S_1 and S_2 are the covariance matrices for each group. - n_1 and n_2 are the sample sizes of each group.

5.2.4 Calculate P-value

The p-value for James's test is obtained by comparing the test statistic to the critical F-value. If the calculated test statistic exceeds the critical value, the p-value will be smaller than the significance level, leading to the rejection of the null hypothesis.

5.2.5 Results

For the given data, the results of James's test are as follows:

Test Statistic	**Value**
James's Test Statistic	78.5758
Critical F-value	8.7901
P-value	0.0020

Table 5: James's Test Results

5.2.6 Observation

The calculated James's test statistic is 78.5758, which is significantly greater than the critical F-value of 8.7901. Additionally, the p-value of 0.0020 is less than the typical significance level of 0.05. Therefore, we reject the null hypothesis and conclude that the means of the two datasets are significantly different.

6 Outlier Detection using Mahalanobis Distance

6.1 Mahalanobis Distance

Mahalanobis Distance is a multivariate measure that quantifies the distance of a data point from the mean of a distribution, considering the correlation of the dataset. Unlike Euclidean distance, which only takes into account the direct difference between points, Mahalanobis Distance accounts for the variance and covariance of the dataset.

The formula for Mahalanobis distance is given by:

$$D_M = \sqrt{(x - \mu)^T S^{-1} (x - \mu)}$$

Where: - x is the data point. - μ is the mean vector of the distribution. - S is the covariance matrix of the dataset. - S^{-1} is the inverse of the covariance matrix.

This distance is particularly useful for identifying outliers in multivariate datasets. Points with a high Mahalanobis distance (often above a threshold, e.g., 3 or 4) are considered potential outliers, as they are far from the mean of the dataset when accounting for the covariance structure.

6.2 Procedure for Outlier Detection

To detect outliers using Mahalanobis Distance:

1. **Compute the mean** and **covariance matrix** of the dataset.
2. **Calculate the Mahalanobis distance** for each data point.
3. **Identify outliers** by comparing each Mahalanobis distance to a threshold value, such as 3. Data points with a distance greater than the threshold are considered outliers.

In our analysis, the number of outliers detected in the two datasets is as follows:

- Number of Outliers in Dataset 1: 8
- Number of Outliers in Dataset 2: 24

6.3 Visualizing the Mahalanobis Distance

Below are the plots of the Mahalanobis distances for Dataset 1 and Dataset 2. The plots help visualize the distribution of distances, with outliers typically having higher values.

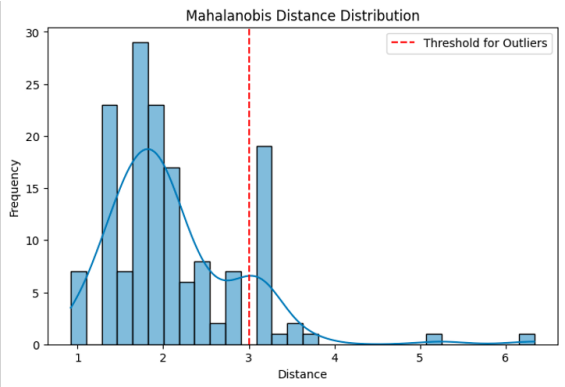


Figure 3: Mahalanobis Distance Plot for Dataset 1

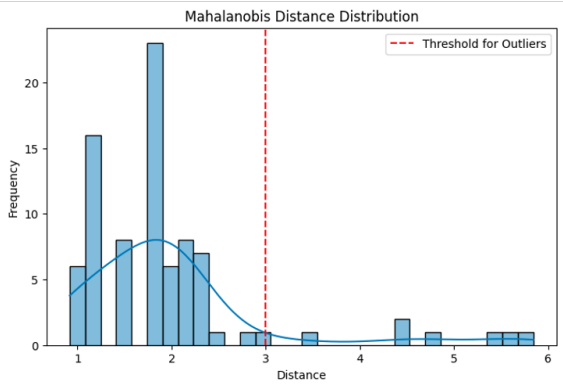


Figure 4: Mahalanobis Distance Plot for Dataset 2

7 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform a dataset into a set of orthogonal components. These components capture the maximum variance in the data. The primary goal of PCA is to reduce the number of dimensions (features) while retaining as much information as possible. This is done by computing the eigenvectors (principal components) of the covariance matrix of the data.

The eigenvalues associated with each principal component represent the amount of variance explained by that component. Larger eigenvalues indicate components that explain a higher proportion of the data's variance, while smaller eigenvalues indicate components with lesser variance.

7.1 PCA Plot and Eigenvalues

The following plot represents the first few principal components derived from the data. The eigenvalues for the principal components are as follows:

$$\text{Eigenvalues: } 4.9398, \quad 0.0256, \quad 0.0035, \quad 0.0182, \quad 0.0129$$

These eigenvalues suggest that the first principal component captures the majority of the variance in the dataset, indicating that the data is highly correlated along this axis.

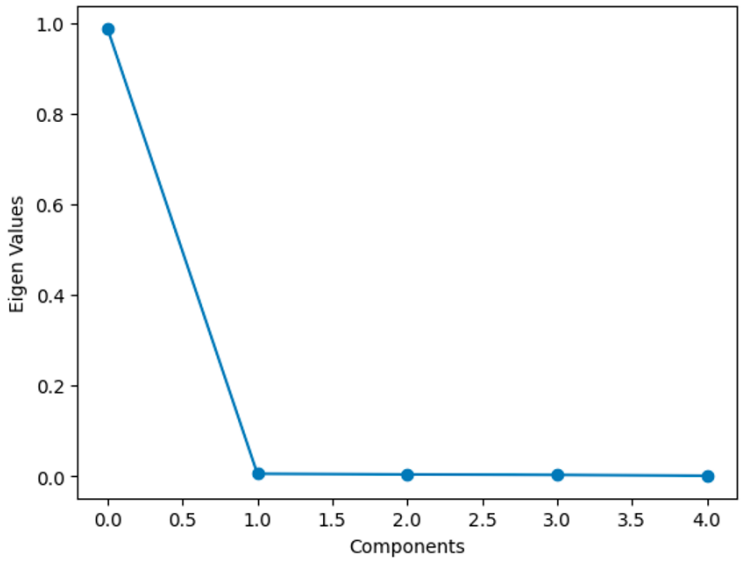


Figure 5: Principal Component Analysis (PCA) Plot

7.2 Observation

The large eigenvalue of 4.9398 for the first principal component indicates that a significant portion of the variance in the dataset is captured by this component. This is consistent with the earlier finding that the attributes in the dataset are strongly related, as most of the variability can be explained by a single dimension. The relatively small eigenvalues for the subsequent components (0.0256, 0.0035, 0.0182, and 0.0129) show that the remaining components capture very little additional variance, reinforcing the idea that the data is highly correlated along the first principal component.

7.3 Dimension Reduction Plots

Presented below are plots illustrating dimension reduction through PCA into 1D, 2D, and 3D spaces.

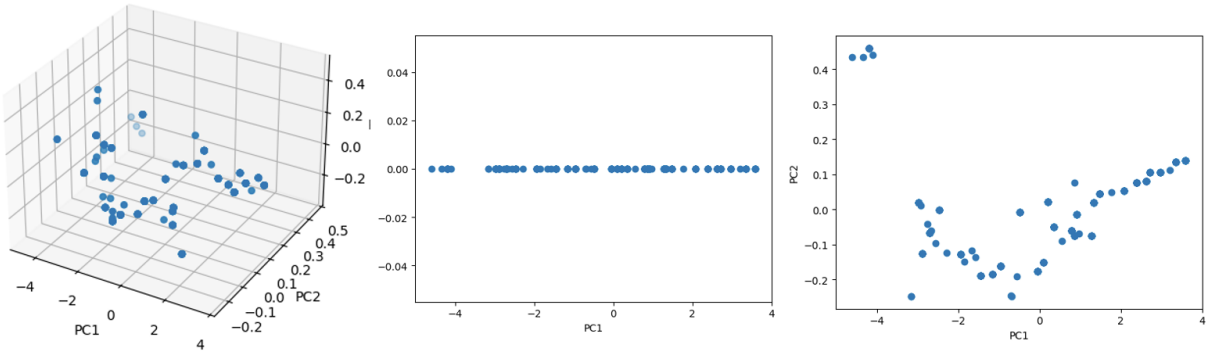


Figure 6: Dimension Reduction using PCA into 1D, 2D, and 3D

8 Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to fit a line that best represents the data, minimizing the difference between predicted and actual values. In the case of a single independent variable, the relationship can be expressed as:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where: - y is the dependent variable. - x is the independent variable. - β_0 is the intercept. - β_1 is the slope of the line. - ϵ is the error term.

The quality of the fit is often measured by R^2, which indicates the proportion of the variance in the dependent variable that is predictable from the independent variable.

8.1 Gradient Descent for Linear Regression

To find the optimal values of β_0 and β_1, we can use gradient descent, an iterative optimization algorithm that minimizes the cost function (mean squared error in this case). The cost function for linear regression is:

$$J(\beta_0, \beta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

The gradients for β_0 and β_1 are calculated as follows:

$$\frac{\partial J}{\partial \beta_0} = -\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)$$

$$\frac{\partial J}{\partial \beta_1} = -\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i) x_i$$

The parameters are then updated iteratively using a learning rate α:

$$\beta_0 = \beta_0 - \alpha \frac{\partial J}{\partial \beta_0}, \quad \beta_1 = \beta_1 - \alpha \frac{\partial J}{\partial \beta_1}$$

8.2 R^2 Values for Age-based Regressions

The following table shows the R^2 values for linear regressions performed on Age against other variables.

Regression	R^2 **Value**
Age vs Purchase Amount	0.9725
Age vs Annual Income	0.9503
Age vs Loyalty Score	0.9640

Table 6: R^2 values for Age-based Linear Regressions

8.3 Regression Plots

Below are the plots of the linear regressions performed on Age versus Purchase Amount, Annual Income, and Loyalty Score.

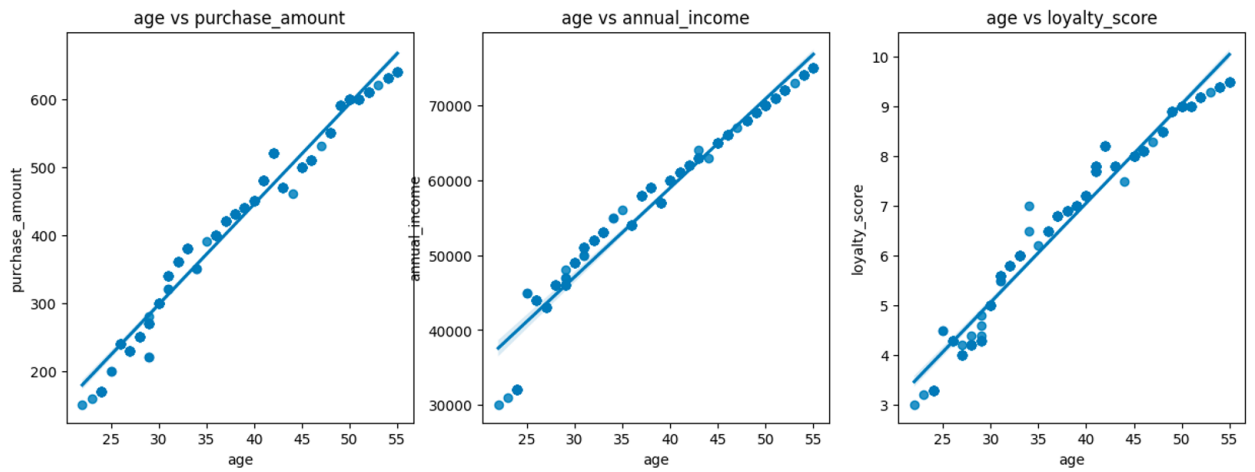


Figure 7: Linear Regression Plot

9 Softmax Regression for Region Classification

9.1 Model Training and Evaluation

To classify the regions into categories—North, South, East, and West—a softmax regression model was trained and tested on the same dataset. The resulting accuracy is as follows:

Accuracy: 0.4328

9.2 Analysis

The relatively low accuracy score of 43.28% suggests that the model struggles to find a distinct linear boundary between the classes in this dataset. This outcome is expected, as the data likely lacks clear separability across the four regional categories.

9.3 Visualization of Results

The figures below illustrate the model's performance:

- **Actual Region Labels:** Displaying the true region labels in the dataset.
- **Predicted Region Labels:** Showing the model's predicted labels based on the softmax regression results.

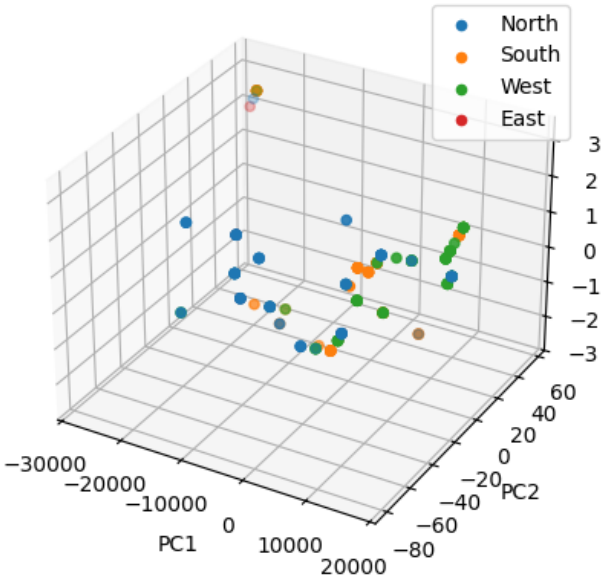


Figure 8: Actual Region Labels

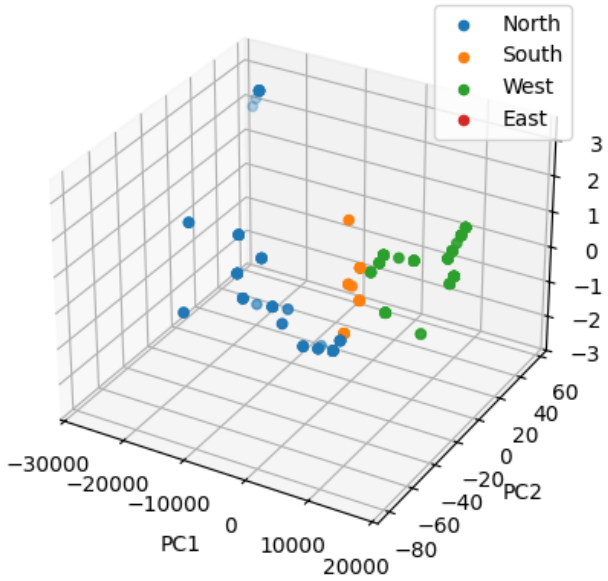


Figure 9: Predicted Region Labels

9.4 Observation

From these plots, it is evident that there is no clear linear boundary separating the regions, which accounts for the model's limited accuracy. Softmax regression, which relies on linear decision boundaries, may not be suitable for effectively classifying this dataset, as it does not capture complex relationships within the data.

10 Conclusion

Based on the analysis of the datasets, the following key observations can be made:

- The relationship between all attributes shows a strong correlation, indicating that the attributes are interdependent.
- The Box M test results suggest that the null hypothesis, which assumes equal covariance matrices, is rejected, confirming significant differences in the covariance structures of the two datasets.
- The Likelihood Ratio Test further supports this conclusion by indicating a significant difference between the datasets with a very low p-value.
- Welch's T-Test indicates no significant difference in the means for the different attributes between the two datasets, as evidenced by the high p-values.
- James's Test also rejects the null hypothesis, supporting the conclusion that the datasets have different means.
- Outlier detection using Mahalanobis distance revealed 8 outliers in Dataset 1 and 24 in Dataset 2, indicating the presence of data points that are significantly different from the rest of the observations.

- Principal Component Analysis (PCA) revealed that a small number of components explain most of the variance, confirming the high correlation between the attributes.
- The softmax regression model performed with an accuracy of 43.28%, suggesting that the data does not have clear linear separability for classification.
- Linear regression models for predicting attributes such as Age, Purchase Amount, and Loyalty Score achieved high R^2 values, indicating strong relationships between the predictors and the outcomes.
- Dimension reduction using PCA helped visualize the structure of the data in 1D, 2D, and 3D, highlighting the complexities of the dataset.