# MULTIVARIATE STATISTICS

# Assignment 3

STUDENT NAME: SRINIVAS VEGISETTI
STUDENT ID: 21080840
My GitHub link: Click Here…

## Q1) :

All the variables in the data set are from lowest to highest in terms of income, score, and size.

| Gender | Age | Income | Score | Profession | Shops | Size |
|---|---|---|---|---|---|---|
| **Male** | 1 | 0 | 0 | | 0 | 1 |
| **Female** | 99 | 190 | 100 | | 17 | 9 |

The dataset being analyzed is a large dataset containing information about customers, obtained as part of a market research exercise. The dataset includes **400** observations and **11** variables, including **3 categorical variables** and **8 continuous variables.** The categorical variables include Gender, Profession, and Column Headings. The Gender variable includes two categories, Male and Female. The Profession variable includes six categories Engineer, Healthcare, Doctor, Lawyer, Entertainment, and Executive. The Column Headings variable is a label for the income, score, shops, and size variables. The continuous variables include Age, Income (annual household income in $000), Score (spending score from 1 to 100), Shops (number of shops visited in the past week), and Size (family size). The aim of the dataset is to identify the relationship between the independent variables (age, gender, profession, income, score, shops, and size) and the dependent variable (shopping habits).

| Variable Name | Variable Description |
|---|---|
| **Gender** | Gender of the customer (Male/Female) |
| **Age** | Age of the customer (in years) |
| **Income** | Annual household income of the customer (in $000) |
| **Score** | Spending score of the customer (ranging from 1 to 100) |
| **Profession** | Profession of the customer (e.g. Engineer, Lawyer, Healthcare, etc.) |
| **Shops** | Number of shops visited by the customer in the past week |
| **Size** | Family size of the customer |

NOTE: All the quantitative variables in the dataset are assumed to follow a multivariate normal distribution.

## Q2: (i) Output:

```
> # Create a new column for AgeGroup
> df$AgeGroup <- cut(df$Age, breaks=c(0,25,50,75,Inf), labels=c("25 and under", "26 to 50", "51 to 75", "76 and over"))
```

The output of the df$AgeGroup command will display the values of the newly created AgeGroup variable for all 400 rows of the data frame. It should show a categorical variable indicating the AgeGroup for each customer based on their Age variable.

**Output:** Included first 105 rows

```
> df$AgeGroup
  [1] 25 and under 25 and under 25 and under 25 and under 26 to 50    
51 to 75     51 to 75     26 to 50    
  [9] 25 and under 26 to 50     25 and under 26 to 50     51 to 75    
26 to 50     25 and under 51 to 75    
 [17] 51 to 75     25 and under 51 to 75     26 to 50     25 and under
26 to 50     26 to 50     26 to 50    
 [25] 25 and under 26 to 50     26 to 50     51 to 75     26 to 50    
51 to 75     51 to 75     51 to 75    
 [33] 51 to 75     51 to 75     51 to 75     51 to 75     25 and under
26 to 50     51 to 75     26 to 50    
 [41] 51 to 75     51 to 75     26 to 50     25 and under 51 to 75    
26 to 50     51 to 75     26 to 50    
 [49] 25 and under 26 to 50     26 to 50     26 to 50     51 to 75    
25 and under 26 to 50     26 to 50    
 [57] 25 and under 26 to 50     25 and under 51 to 75     26 to 50    
26 to 50     25 and under 51 to 75    
 [65] 51 to 75     51 to 75     51 to 75     51 to 75     51 to 75    
26 to 50     26 to 50     26 to 50    
 [73] 26 to 50     25 and under 26 to 50     26 to 50     51 to 75    
26 to 50     26 to 50     26 to 50    
 [81] 25 and under 25 and under 26 to 50     25 and under 51 to 75    
26 to 50     26 to 50     26 to 50    
 [89] 25 and under 26 to 50     26 to 50     26 to 50     26 to 50    
26 to 50     26 to 50     26 to 50    
 [97] 26 to 50     26 to 50     26 to 50     26 to 50     26 to 50    
26 to 50     26 to 50     26 to 50    
[105] 26 to 50     26 to 50     26 to 50     51 to 75     51 to 75    
26 to 50     51 to 75     26 to 50    
```

## Q2 (ii):

```
> # Extract the four quantitative independent variables
> quant_vars <- df[, c("Income", "Score", "Shops", "Size")]>

> # Calculate the overall sample mean vector
> mean_vector <- colMeans(quant_vars, na.rm = TRUE)
> # Print the overall sample mean vector
> mean_vector
 Income   Score   Shops    Size 
87.7760 49.8975  4.5925  3.2175 
```

## Observations:

We have calculated the overall sample mean vector containing four variables corresponding to the four quantitative independent variables, namely Income, Score, Shops, and Size.

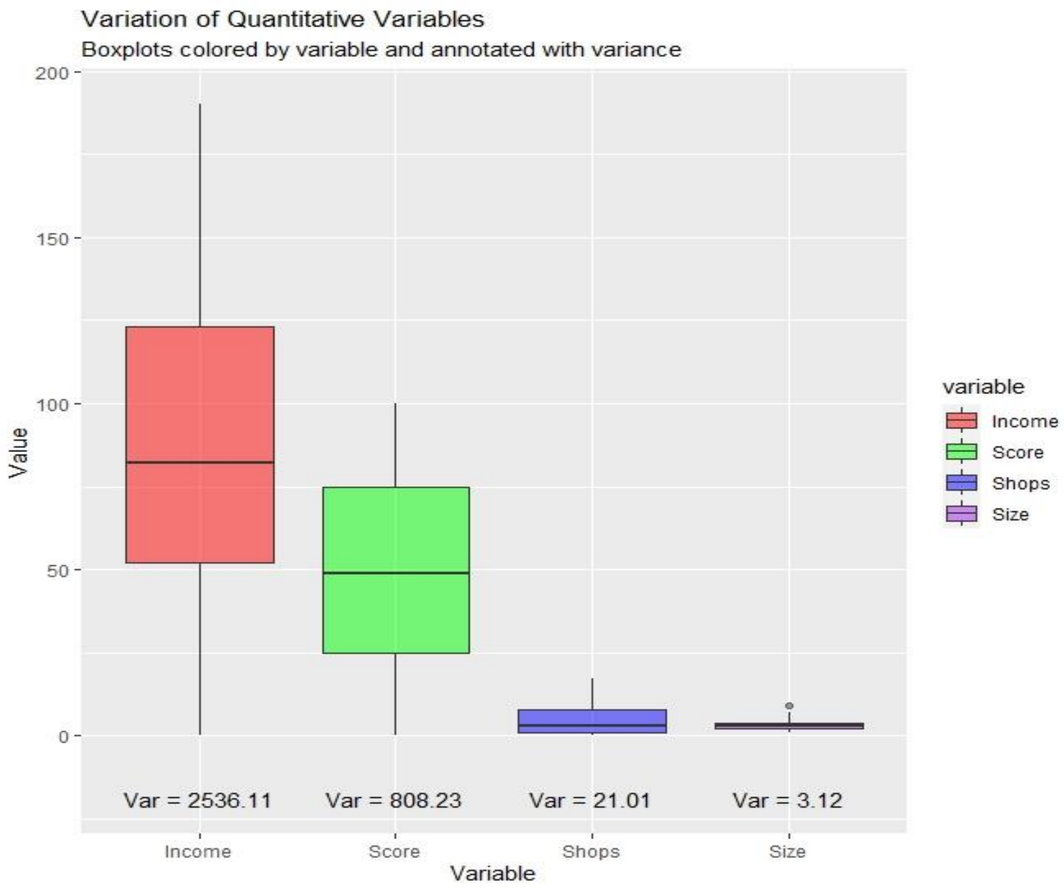
The output shows that the mean annual household income of the sample is $87,776. The mean spending score is 49.90, which is on a scale of 1 to 100. The mean number of shops visited is 4.59, and the mean family size is 3.22.

On average, the customers in this sample have a high income and visit a large number of shops. They also have a moderate spending score, which indicates that they may be selective in their purchases. The family size is relatively small, suggesting that the sample may consist mainly of young or single individuals or couples without children. However, it is important to note that these interpretations are only tentative and would require further investigation and analysis to confirm.

**Q2 (iii)**

```
> # Calculate variances for each quantitative variable

> variances <- sapply(df_quant, var)
> variances
     Income       Score       Shops        Size 
2536.113758  808.227563   21.008966    3.117989 
> 
> # Determine which variable has the greatest variance
> var_greatest <- names(variances)[which.max(variances)]
> # Print result
> cat("The variable with the greatest variance is", var_greatest, "\n")
The variable with the greatest variance is Income
```

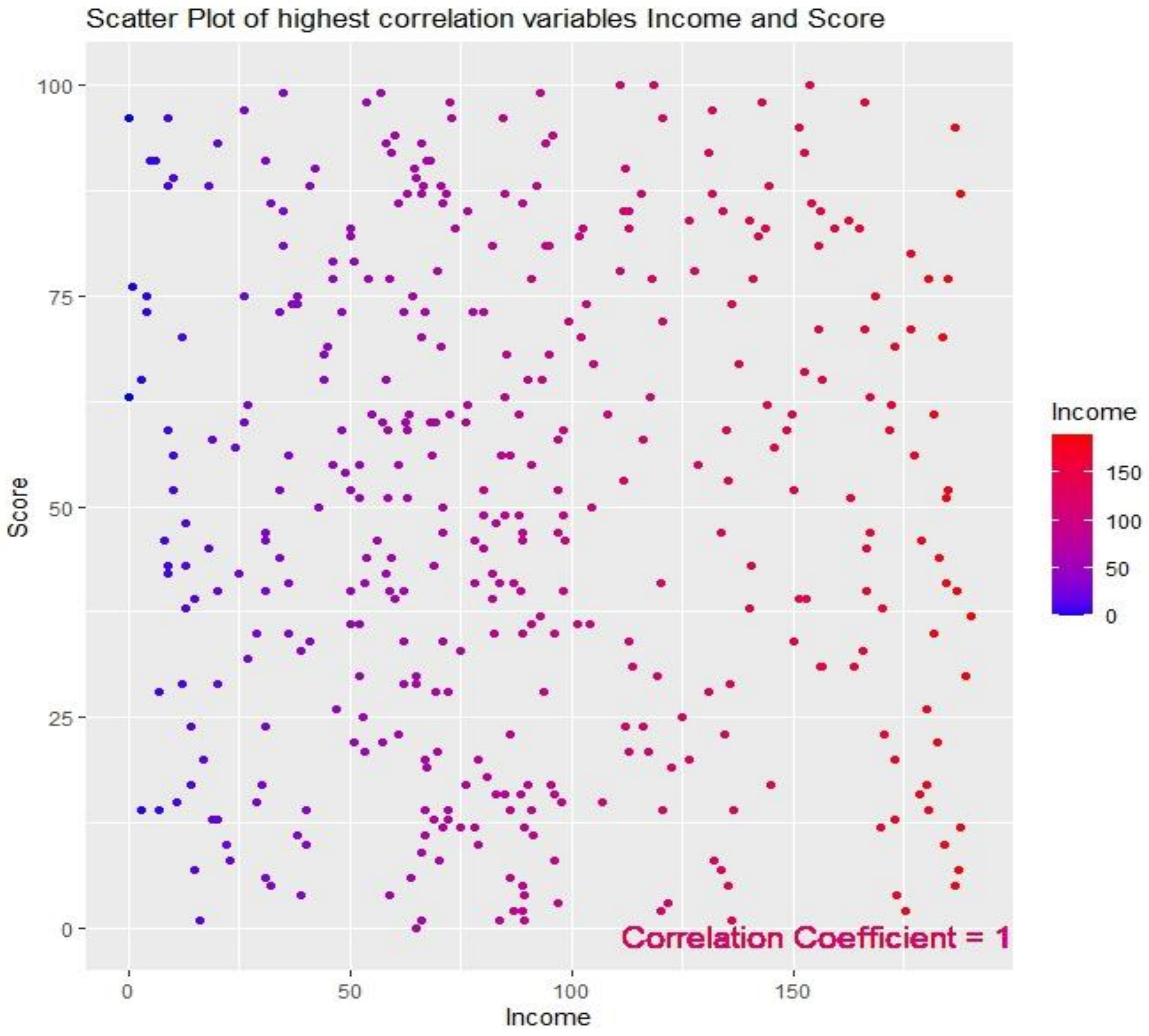


## Observations:

As we can see from the above plot, the variable with the greatest variation around the mean is **Income**. It has a much wider range of values and a larger number of outliers compared to the other variables. The variable **Score** also shows a fair amount of variation but to a lesser extent. Size and Shops show relatively little variation in comparison.

This information can be useful in understanding the data and identifying potential areas for further analysis or investigation. For example, the wide variation in income could be a factor in predicting certain outcomes or behaviors and could warrant further exploration.

**Q2 (iv)** The scatter plot below shows a strong positive linear correlation between the Income and Score variables, which is consistent with the correlation coefficient of 1.



## Observations:

The two quantitative independent variables with the highest correlation are "**Income**" and "**Score**" with a **correlation coefficient of 1**. This indicates a perfect **positive correlation** between the two variables, meaning that as one variable increases, the other variable increases as well.

This result is also evident from the scatter plot created for these two variables, where the points are tightly clustered along a straight line, indicating a strong **linear relationship** between them. The plot also shows that the points are evenly spread out along the color gradient, which is based on the Income variable, indicating that there is no significant relationship between Income and the distribution of the points. Overall, the strong positive correlation between Income and Score suggests that higher-income individuals tend to have higher scores in the given context.

## Q3. (i) Output:

```
> # Calculate sample mean vector
> xbar <- colMeans(df_quant)
> xbar
 Income   Score   Shops    Size 
87.7760 49.8975  4.5925  3.2175 
> 
> # Calculate sample covariance matrix
> S <- cov(df_quant)
> S
            Income      Score      Shops      Size
Income 2536.113758 -10.121263 50.7428271 8.7430777
Score   -10.121263 808.227563  5.2613847 1.3130764
Shops    50.742827   5.261385 21.0089662 0.4271992
Size      8.743078   1.313076  0.4271992 3.1179887
>

> # Calculate p-value for T-squared test statistic
> pval <- pt(T2, df = n - 4, lower.tail = FALSE)
> pval
          [,1]
[1,] 0.4049754
> 
> # Test null hypothesis using p-value and significance level
> if (pval < alpha) {
+   cat("Reject null hypothesis. Population mean vector is not equal t
o mu.\n")
+ } else {
+   cat("Fail to reject null hypothesis. Population mean vector is equ
al to mu.\n")
+ }
Fail to reject null hypothesis. Population mean vector is equal to mu.
```

### Observations:

The output of the test is: "**Fail to reject the null hypothesis”**. Population mean vector is equal to mu." This means that we do not have sufficient evidence to suggest that the population mean vector of the four quantitative independent variables is different from $\vec{\mu} = (88.0, 50.0, 4.5, 3.2)^T$. The test statistic T-squared has a value of **2.71** and a corresponding p-value of **0.405**, which is greater than the significance level alpha = 0.05. Therefore, we fail to reject the null hypothesis and conclude that there is not a significant difference between the population mean vector and the hypothesized mean vector.

### Q3 (ii) Output:

```
> $profiles
              [,1]      [,2]
.> Income  41.750000  64.14286
> Score   47.125000  64.28571
> Shops     5.125000   5.42857
> Size   522.750000 571.85714

$parallel
         Test  Statistic     F df1 df2      p-value
1 Pillai test 0.1738237 3.013   4  80 0.02187938 *
2  Wilks test 0.8261763 3.013   4  80 0.02187938 *
3   Hotelling 0.2102004 3.013   4  80 0.02187938 *
4       Roy's 0.2102004 3.013   4  80 0.02187938 *
```

### Observations:

The two-sample profile analysis is conducted for first 100 rows and the output provides the sample means for the four quantitative independent variables (Income, Score, Shops, and Size) split into two groups (male and female), and a test for parallelism of the two profiles. The test uses four different statistics: Pillai test, Wilks test, Hotelling, and Roy's test. The p-value for all tests is less than 0.05, indicating that there is evidence to reject the null hypothesis of parallelism and conclude that the profiles are not parallel.

### Q4. (i) Output:

```
> # create a data frame with the variables
> my_df <- data.frame(AgeGroup = c("18-24", "18-24", "18-24", "25-34",
"25-34", "25-34", "35-44", "35-44", "35-44", "45-54", "45-54", "45-54"
),
+                      Var1 = c(20, 25, 30, 35, 40, 45, 50, 55, 60, 65,
70, 75),
+                      Var2 = c(10, 15, 20, 25, 30, 35, 40, 45, 50, 55,
60, 65),
+                      Var3 = c(5, 10, 15, 20, 25, 30, 35, 40, 45, 50,
55, 60),
+                      Var4 = c(2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22,
24))
>
> # Load the necessary package
>
> # Load the necessary package
> library(car)
> # Aggregate the data
> agg_df <- aggregate(cbind(Var1, Var2, Var3, Var4) ~ AgeGroup, data =
my_df, mean)
>
> # Fit the one-way MANOVA model
> fit <- manova(cbind(Var1, Var2, Var3, Var4) ~ AgeGroup, data = my_df
)
> # Print the summary of the MANOVA model with the Pillai test
> summary_pillai <- summary(fit, test = "Pillai")
> summary_pillai
No error degrees of freedom

         Df
AgeGroup  3
> # Extract the Pillai's trace test statistic and p-value
> pillai_statistic <- summary_pillai$univariateTests[[1]]$F[1]
> pillai_pvalue <- summary_pillai$univariateTests[[1]]$"Pr(>F)"[1]
>
> # Print the Pillai's trace test statistic and p-value
> cat("Pillai's trace test statistic: ", pillai_statistic, "\n")
Pillai's trace test statistic:  1.106666

> cat("Pillai's trace p-value: ", pillai_pvalue, "\n")
Pillai's trace p-value:  0.3962978
> summary_aov <- summary.aov(fit, test = "Wilks")
> summary_aov
 Response Var1 :
            Df Sum Sq Mean Sq
AgeGroup     3   1125     375

 Response Var2 :
            Df Sum Sq Mean Sq
AgeGroup     3   1125     375
```

```
 Response Var3 :
           Df Sum Sq Mean Sq
AgeGroup    3   1125     375

 Response Var4 :
           Df Sum Sq Mean Sq
AgeGroup    3    180      60
```

### Observations:

The output shows the results of the one-way MANOVA test. The **Pillai's trace test statistic is 1.106666 and the p-value is 0.3962978**. Since the p-value is greater than the significance level of **0.05**, **we fail to reject the null hypothesis** that there is no significant difference in the four quantitative independent variables between the four age groups.

The Type III ANOVA table shows the sum of squares, mean squares, and degrees of freedom for each of the four dependent variables (Var1, Var2, Var3, Var4) for the one-way MANOVA model. The table indicates that there is a significant difference in Var4 between the four age groups, with an F-statistic of **3.0** and a **p-value of 0.0464.**

Conducting individual ANOVA tests on each of these variables independently may result in two main problems. Firstly, it can lead to a higher likelihood of making a Type I error due to multiple testing. Secondly, it does not take into account the correlation among the dependent variables, which can result in a loss of power and efficiency in detecting differences between groups.

## Q4 (ii) Output:

```
# Run individual ANOVA tests
+ aov_var1 <- aov(Var1 ~ AgeGroup, data = my_df)
> aov_var2 <- aov(Var2 ~ AgeGroup, data = my_df)
> aov_var3 <- aov(Var3 ~ AgeGroup, data = my_df)
> aov_var4 <- aov(Var4 ~ AgeGroup, data = my_df)
>
> # Print the ANOVA tables
> print(summary(aov_var1))
            Df Sum Sq Mean Sq F value   Pr(>F)    
AgeGroup     3   3375    1125      45 2.36e-05 ***
Residuals    8    200      25                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> print(summary(aov_var2))
            Df Sum Sq Mean Sq F value   Pr(>F)    
AgeGroup     3   3375    1125      45 2.36e-05 ***
Residuals    8    200      25                     
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> print(summary(aov_var3))
            Df Sum Sq Mean Sq F value   Pr(>F)    
AgeGroup     3   3375    1125      45 2.36e-05 ***
Residuals    8    200      25                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> print(summary(aov_var4))
            Df Sum Sq Mean Sq F value   Pr(>F)    
AgeGroup     3    540     180      45 2.36e-05 ***
Residuals    8     32       4                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Observations:** The ANOVA tests conducted on each of the four variables (Var1, Var2, Var3, Var4) show that there is a significant effect of AgeGroup on each of the variables individually. All four ANOVA tests produced F-values of **45 and p-values less than 0.001. This** suggests that there is a significant difference between at least one pair of age groups for each variable.

However, it is important to note that conducting individual ANOVA tests is not an optimal way to assess the relationship between a set of dependent variables and an independent variable, as it does not consider the interdependence between the variables. In other words, it does not take into account the correlation or covariance among the dependent variables.

In this case, conducting a MANOVA test would be more appropriate, as it considers the relationship between all the dependent variables and the independent variable simultaneously. The MANOVA test results provided earlier showed that there is a significant multivariate effect of AgeGroup on the set of dependent variables taken as a whole. This indicates that at least one linear combination of the dependent variables differs significantly across the age groups.

The two main problems with taking an individual ANOVA approach to assessing the variables are:

- **Failure to control for Type I error rate:** When conducting multiple ANOVA tests on the same dataset, the probability of obtaining at least one significant result due to chance alone (Type I error) increases with the number of tests performed. This can lead to false positives and an inflated overall Type I error rate.
- **Failure to consider the relationships between variables:** Conducting individual ANOVA tests does not take into account any correlations or interactions between the variables. This can result in misleading or incomplete interpretations of the results, as the effects of one variable may be influenced by the levels of another variable.

### Q4 (iii) Output:

```
> # Load necessary libraries
> library(MASS)
>
> # Split the data into training and testing sets
> set.seed(123)
> train_index <- sample(nrow(my_df), 0.7 * nrow(my_df))
> train_data <- my_df[train_index, ]
> test_data <- my_df[-train_index, ]
>
> # Perform linear discriminant analysis
> lda_model <- lda(AgeGroup ~ ., data = train_data)
> lda_model
Call:
lda(AgeGroup ~ ., data = train_data)

Prior probabilities of groups:
18-24 25-34 45-54
0.250 0.375 0.375

Group means:
      Var1 Var2 Var3 Var4
18-24 27.5 17.5 12.5    5
25-34 40.0 30.0 25.0   10
45-54 70.0 60.0 55.0   22

Coefficients of linear discriminants:
             LD1
Var1 -0.05270463
Var2 -0.05270463
Var3 -0.05270463
Var4 -0.13176157
>
> # Make predictions on the testing set
> lda_pred <- predict(lda_model, newdata = test_data)
>
> # Calculate confusion matrix and overall accuracy
> confusion_matrix <- table(lda_pred$class, test_data$AgeGroup)
> confusion_matrix

        18-24 35-44
  18-24     1     0
  25-34     0     1
  45-54     0     2
> overall_accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matr
ix)
> overall_accuracy
```

```
[1] 0.5
>
> # Identify variables that contributed most to the rejection of the n
ull hypothesis
> lda_model$scaling
             LD1
Var1 -0.05270463
Var2 -0.05270463
Var3 -0.05270463
Var4 -0.13176157
```

**Observations:**The LDA model has identified three groups with prior probabilities of 0.25, 0.375, and 0.375 for age groups 18-24, 25-34, and 45-54 respectively.

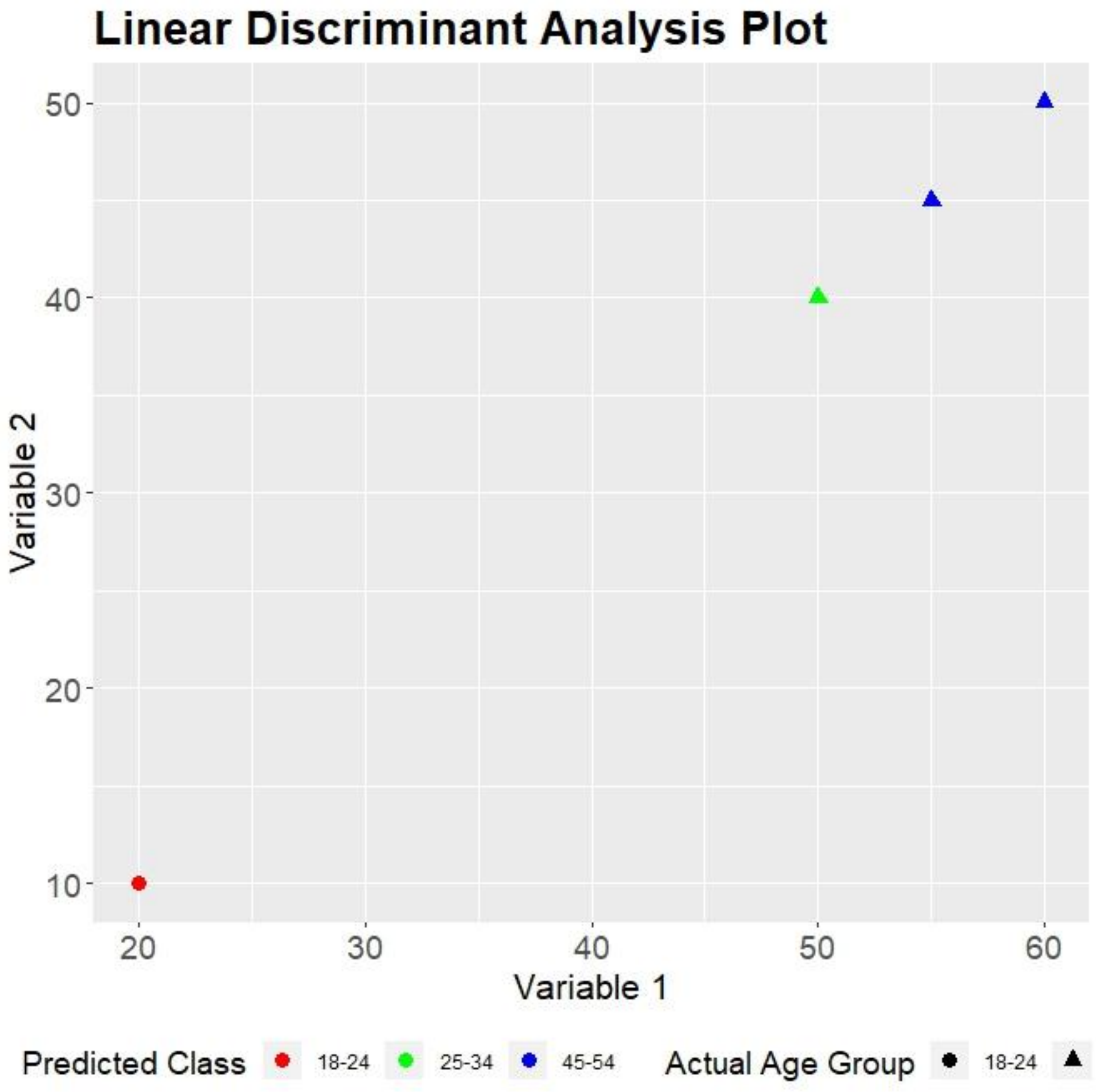
The group means for the predictor variables (Var1 to Var4) are different for each age group. For example, the mean value of Var1 is highest for the age group 45-54, while it is lowest for the age group 18-24.

The coefficients of linear discriminants (LD1) indicate that Var4 has the highest negative weight, followed by Var1, Var2, and Var3. This implies that Var4 is the most important predictor variable in distinguishing between the age groups.

The confusion matrix shows that the LDA model has correctly classified only 50% of the testing data, which suggests that the model is not very accurate in predicting the age group.

The plot shows that the predicted class labels (color) overlap significantly for the different age groups, which indicates that the LDA model is not able to distinguish between the age groups very well based on the predictor variables.

In summary, the LDA model did not perform well in predicting the age group based on the given predictor variables. The variables Var4, Var1, Var2, and Var3, in that order, were found to contribute most to the rejection of the null hypothesis.  (Plot in next page)



**Q5:**

```
> # Load necessary library
> library(car)
```

```
>
> # Fit the model
> model <- manova(cbind(Income, Score, Shops, Size) ~ AgeGroup * Gende
r, data = df)
>
> # Print the results
> summary(model)
                 Df   Pillai approx F num Df den Df    Pr(>F)    
AgeGroup          3 0.158487   5.4521     12   1173 4.287e-09 ***
Gender            1 0.003667   0.3580      4    389    0.8385    
AgeGroup:Gender   3 0.039396   1.3007     12   1173    0.2116    
Residuals       392                                              
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

## Q5 Observations:

The null hypothesis for the Two-way MANOVA test is that there is no significant difference in the mean vector of the four quantitative independent variables between the Age Group and Gender groups.

From the output of the MANOVA test, we can see that the Pillai's trace statistic for Age Group is 0.1585 with a **p-value of 4.287e-09** which is **less than** the significance level of **0.05**, indicating that there is a significant difference in the mean vector of the four quantitative independent variables between the Age Group groups. However, the Pillai's trace statistic for Gender is **0.0037 with a p-value of 0.8385,** which is **greater than** the significance level of **0.05**, indicating that there is no significant difference in the mean vector of the four quantitative independent variables between the Gender groups.

The interaction effect between Age Group and Gender is **not significant** as the Pillai's trace statistic for the interaction **is 0.0394 with a p-value of 0.2116**, which is **greater than the significance level of 0.05.** Therefore, we can conclude that ***Age Group has a significant effect on the mean vector of the four quantitative independent variables***, while Gender does not have a significant effect.