FA6

ABLIAN

2025-05-05

```r
data <- read.csv("D:/FEU/3RD YR 2ND SEM/EDA/customer_segmentation.csv")
head(data)
```

Data Exploration

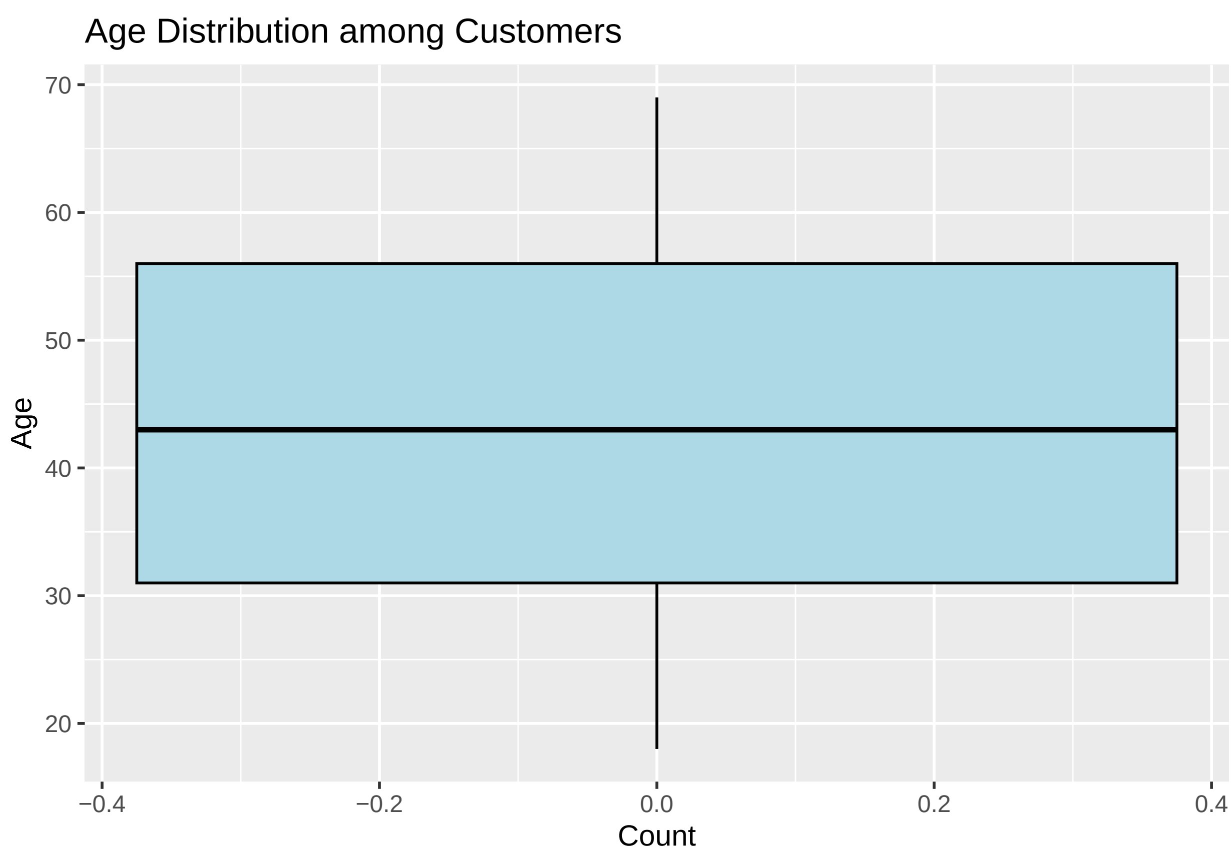
```
##   Customer.ID Age Annual.Income..K.. Gender Product.Category.Purchased
## 1           1  56                106 Female                    Fashion
## 2           2  69                 66 Female                       Home
## 3           3  46                110   Male                    Fashion
## 4           4  32                 50   Male                Electronics
## 5           5  60                 73 Female                     Others
## 6           6  25                 48   Male                       Home
##   Average.Spend.per.Visit.... Number.of.Visits.in.Last.6.Months
## 1                    163.4528                                16
## 2                    163.0205                                31
## 3                    104.5413                                29
## 4                    110.0646                                26
## 5                    142.2546                                38
## 6                    106.7621                                22
##   Customer.Segment
## 1  Premium Shopper
## 2   Budget Shopper
## 3   Budget Shopper
## 4  Regular Shopper
## 5  Regular Shopper
## 6   Budget Shopper
```

```r
summary(data)
```

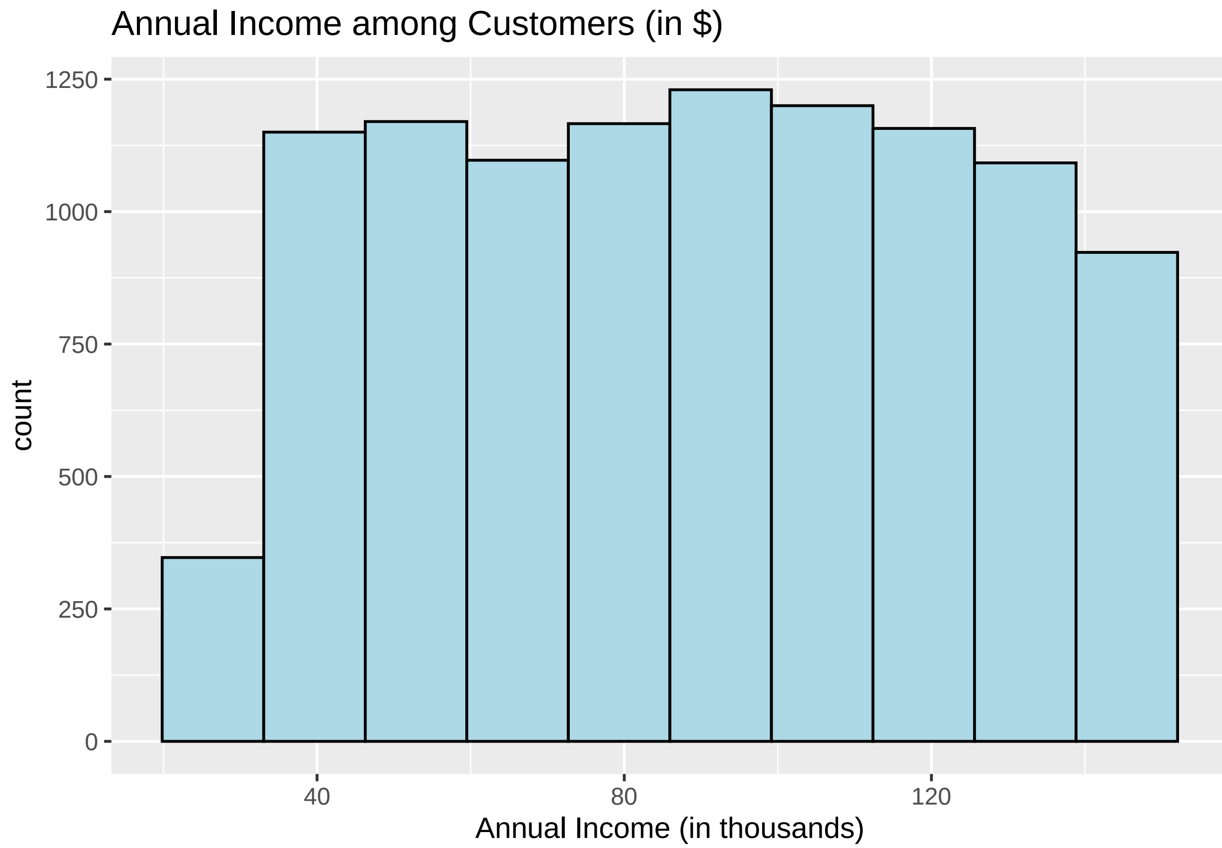
```
##   Customer.ID         Age        Annual.Income..K..    Gender
##  Min.   :    1   Min.   :18.00   Min.   : 30.00     Length:10532
##  1st Qu.: 2634   1st Qu.:31.00   1st Qu.: 59.00     Class :character
##  Median : 5266   Median :43.00   Median : 89.00     Mode  :character
##  Mean   : 5266   Mean   :43.59   Mean   : 89.18
##  3rd Qu.: 7899   3rd Qu.:56.00   3rd Qu.:118.00
##  Max.   :10532   Max.   :69.00   Max.   :149.00
##  Product.Category.Purchased Average.Spend.per.Visit....
##  Length:10532               Min.   : 10.00
##  Class :character           1st Qu.: 56.71
```

```
##   Mode  :character           Median :104.69
##                              Mean   :104.30
##                              3rd Qu.:150.89
##                              Max.   :199.96
##   Number.of.Visits.in.Last.6.Months Customer.Segment
##   Min.   : 5.00                     Length:10532
##   1st Qu.:13.00                     Class :character
##   Median :22.00                     Mode  :character
##   Mean   :21.92
##   3rd Qu.:31.00
##   Max.   :39.00
```

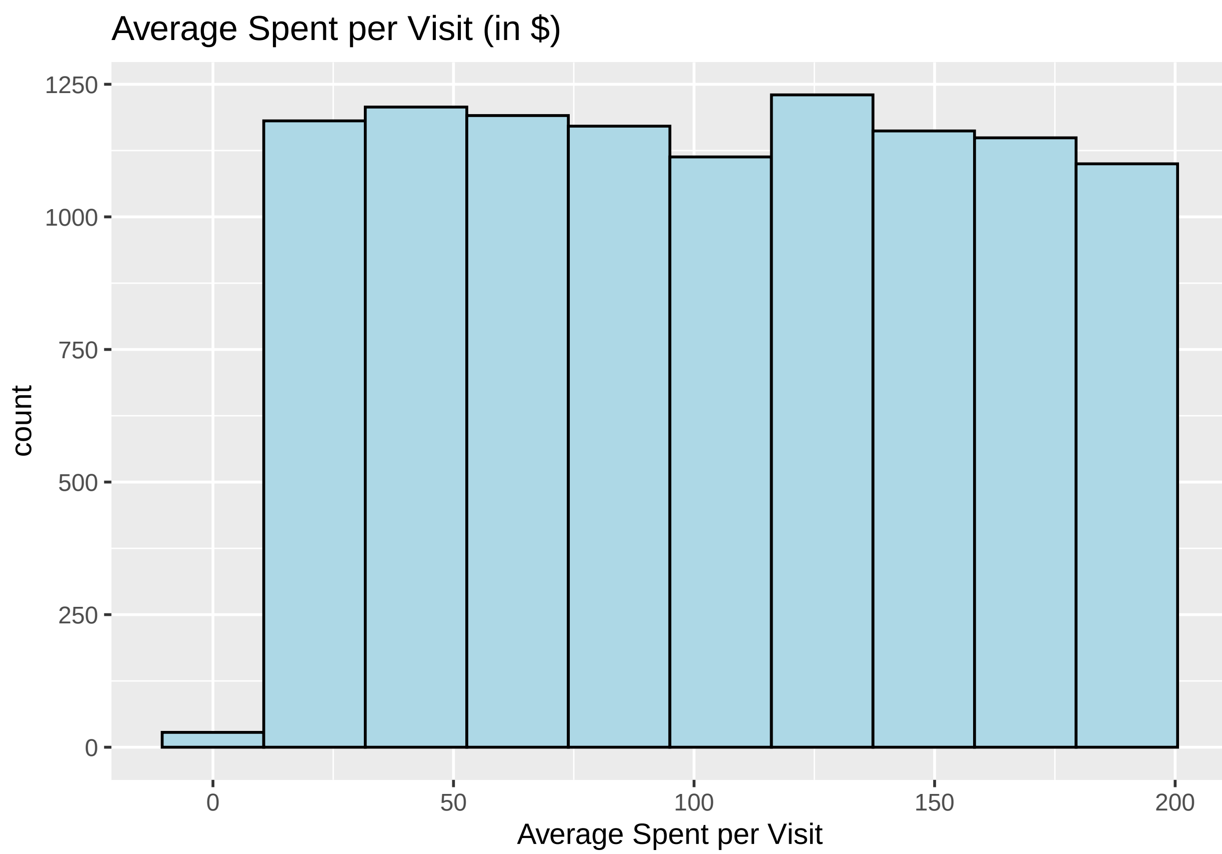
```r
library(ggplot2)
ggplot(data = data, mapping = aes(y = Age)) +
  geom_boxplot( fill = "lightblue", color = "black") +
  labs(title = "Age Distribution among Customers", x = "Count")
```




```r
library(ggplot2)
ggplot(data =  data, mapping = aes(x = Annual.Income..K..)) +
  geom_histogram( fill = "lightblue", color = "black", bins = 10)+
  labs(title = "Annual Income among Customers (in $)", x = "Annual Income (in thousands)")
```




```
ggplot(data =  data, mapping = aes(x =Average.Spend.per.Visit....)) +
  geom_histogram( fill = "lightblue", color = "black", bins = 10)+
  labs(title = "Average Spent per Visit (in $)", x = "Average Spent per Visit")
```

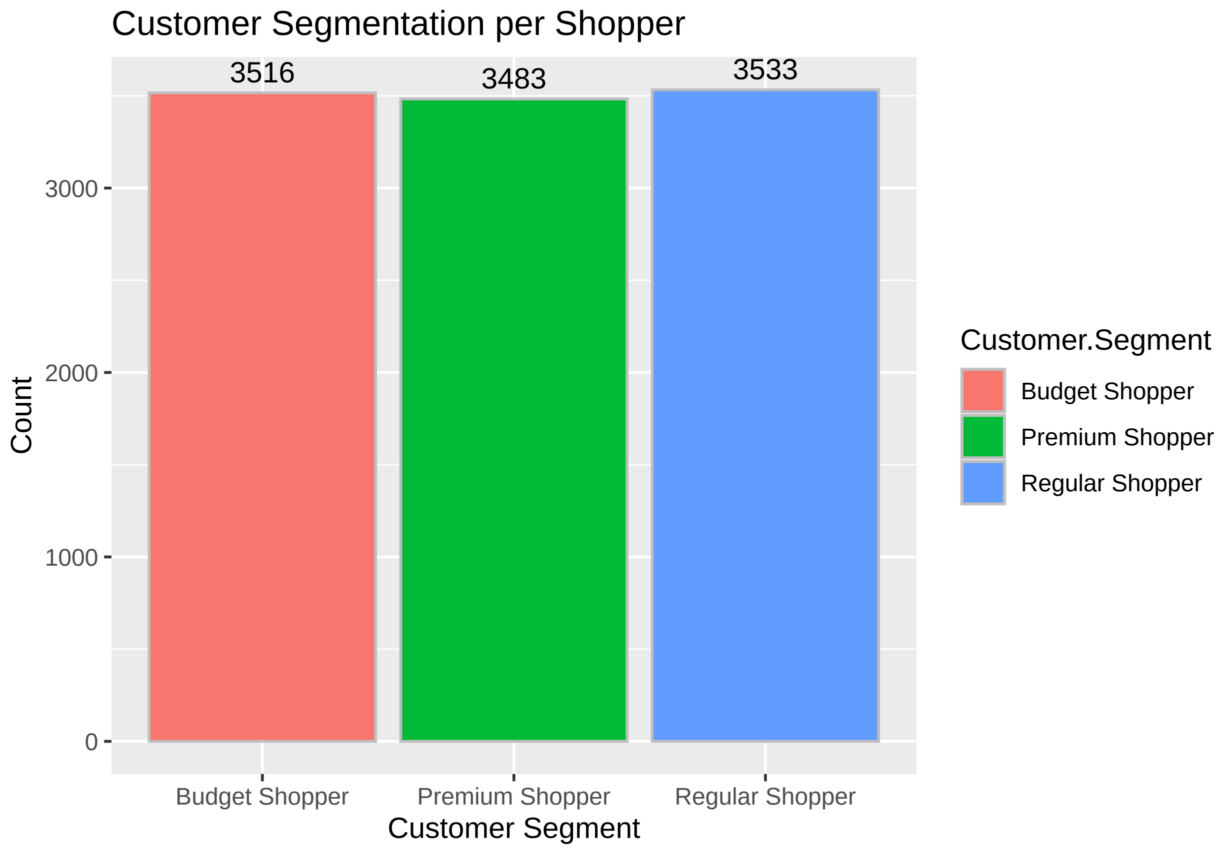



```r
colSums(is.na(data))
```

```
##                      Customer.ID                              Age
##                                0                                0
##                 Annual.Income..K..                           Gender
##                                0                                0
##         Product.Category.Purchased       Average.Spend.per.Visit....
##                                0                                0
## Number.of.Visits.in.Last.6.Months                 Customer.Segment
##                                0                                0
```

There are no missing values.

```r
ggplot(data = data, mapping = aes(x = Customer.Segment, fill = Customer.Segment )) +
  geom_bar(color = "gray") +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.5) +
  labs(title= "Customer Segmentation per Shopper", x = "Customer Segment", y = "Count")
```



The three types of customer are about equal but the highest one is the regular shopper, having 50 more than the lowest (premium shopper). The one in the middle is the budget shoppers, tallying at 3516. Nevertheless, all of them are high.

```r
library("caret")
```

Data Reprocessing

```
## Loading required package: lattice
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
dummy <- dummyVars("~Product.Category.Purchased", data = data)
prod_dummy <- data.frame(predict(dummy, newdata = data))
data_OH<-cbind(data, prod_dummy)
data_OH$Product.Category.Purchased<- NULL
head(data_OH)
```

```
##   Customer.ID Age Annual.Income..K.. Gender Average.Spend.per.Visit....
## 1           1  56                106 Female                    163.4528
## 2           2  69                 66 Female                    163.0205
## 3           3  46                110   Male                    104.5413
## 4           4  32                 50   Male                    110.0646
## 5           5  60                 73 Female                    142.2546
## 6           6  25                 48   Male                    106.7621
##   Number.of.Visits.in.Last.6.Months Customer.Segment
## 1                                16  Premium Shopper
## 2                                31   Budget Shopper
## 3                                29   Budget Shopper
## 4                                26  Regular Shopper
## 5                                38  Regular Shopper
## 6                                22   Budget Shopper
##   Product.Category.PurchasedBooks Product.Category.PurchasedElectronics
## 1                               0                                     0
## 2                               0                                     0
## 3                               0                                     0
## 4                               0                                     1
## 5                               0                                     0
## 6                               0                                     0
##   Product.Category.PurchasedFashion Product.Category.PurchasedHome
## 1                                 1                              0
## 2                                 0                              1
## 3                                 1                              0
## 4                                 0                              0
## 5                                 0                              0
## 6                                 0                              1
##   Product.Category.PurchasedOthers
## 1                                0
## 2                                0
## 3                                0
## 4                                0
## 5                                1
## 6                                0
```

```r
data_OH$Gender_Label <- ifelse(data_OH$Gender == "Male", 1, 0)
data_OH$Gender <- NULL
data_OH$Customer.ID <- NULL

df <- data_OH
df <- df%>%rename(Gender = Gender_Label)

head(df)
```

```
##   Age Annual.Income..K.. Average.Spend.per.Visit....
## 1  56                106                    163.4528
```

```
## 2  69                 66                    163.0205
## 3  46                110                    104.5413
## 4  32                 50                    110.0646
## 5  60                 73                    142.2546
## 6  25                 48                    106.7621
##   Number.of.Visits.in.Last.6.Months Customer.Segment
## 1                                16  Premium Shopper
## 2                                31   Budget Shopper
## 3                                29   Budget Shopper
## 4                                26  Regular Shopper
## 5                                38  Regular Shopper
## 6                                22   Budget Shopper
##   Product.Category.PurchasedBooks Product.Category.PurchasedElectronics
## 1                               0                                     0
## 2                               0                                     0
## 3                               0                                     0
## 4                               0                                     1
## 5                               0                                     0
## 6                               0                                     0
##   Product.Category.PurchasedFashion Product.Category.PurchasedHome
## 1                                 1                              0
## 2                                 0                              1
## 3                                 1                              0
## 4                                 0                              0
## 5                                 0                              0
## 6                                 0                              1
##   Product.Category.PurchasedOthers Gender
## 1                                0      0
## 2                                0      0
## 3                                0      1
## 4                                0      1
## 5                                1      0
## 6                                0      1
```

```r
continuous_vars <- c("Age","Annual.Income..K..", "Average.Spend.per.Visit....")
df[continuous_vars] <- scale(df[continuous_vars])

head(df)
```

```
##          Age Annual.Income..K.. Average.Spend.per.Visit....
## 1  0.8323972          0.4886923                 1.083217431
## 2  1.7046295         -0.6737348                 1.075302088
## 3  0.1614492          0.6049350                 0.004477697
## 4 -0.7778779         -1.1387056                 0.105615627
## 5  1.1007764         -0.4703100                 0.695052913
## 6 -1.2475415         -1.1968270                 0.045143784
##   Number.of.Visits.in.Last.6.Months Customer.Segment
## 1                                16  Premium Shopper
## 2                                31   Budget Shopper
## 3                                29   Budget Shopper
## 4                                26  Regular Shopper
## 5                                38  Regular Shopper
## 6                                22   Budget Shopper
##   Product.Category.PurchasedBooks Product.Category.PurchasedElectronics
```

```
## 1                                  0                                     0
## 2                                  0                                     0
## 3                                  0                                     0
## 4                                  0                                     1
## 5                                  0                                     0
## 6                                  0                                     0
##   Product.Category.PurchasedFashion Product.Category.PurchasedHome
## 1                                 1                              0
## 2                                 0                              1
## 3                                 1                              0
## 4                                 0                              0
## 5                                 0                              0
## 6                                 0                              1
##   Product.Category.PurchasedOthers Gender
## 1                                0      0
## 2                                0      0
## 3                                0      1
## 4                                0      1
## 5                                1      0
## 6                                0      1
```

```r
library(caret)
library(dplyr)
set.seed(600)
trIndex <- createDataPartition(df$Customer.Segment, p = 0.8, list = FALSE)

train_data <- df[trIndex, ]
test_data <- df[-trIndex, ]

train_data$Customer.Segment <- as.factor(train_data$Customer.Segment)
test_data$Customer.Segment <- as.factor(test_data$Customer.Segment)

cat("\nTraining data rows:", nrow(train_data),
    "\nTest data rows:", nrow(test_data),
    "\nClass distribution in training set:\n")
```

```
##
## Training data rows: 8427
## Test data rows: 2105
## Class distribution in training set:
```

```r
print(table(train_data$Customer.Segment))
```

```
##
##  Budget Shopper Premium Shopper Regular Shopper
##            2813            2787            2827
```

```r
library(nnet)
segment_mlr <- multinom(Customer.Segment ~ ., data = train_data)
```

```
## # weights:  36 (22 variable)
## initial  value 9258.005757
```

```
## iter  10 value 9248.948137
## iter  20 value 9246.986594
## final  value 9246.582703
## converged
```

```r
summary(segment_mlr)
```

```
## Call:
## multinom(formula = Customer.Segment ~ ., data = train_data)
##
## Coefficients:
##                  (Intercept)          Age Annual.Income..K..
## Premium Shopper  0.06239558 -0.002531114       -0.004037751
## Regular Shopper  0.09614614  0.029310277       -0.013686390
##                  Average.Spend.per.Visit.... Number.of.Visits.in.Last.6.Months
## Premium Shopper                 -0.03229190                      -0.003489528
## Regular Shopper                 -0.05849034                      -0.003156629
##                  Product.Category.PurchasedBooks
## Premium Shopper                      -0.10962215
## Regular Shopper                      -0.08548282
##                  Product.Category.PurchasedElectronics
## Premium Shopper                             0.02838115
## Regular Shopper                             0.02325285
##                  Product.Category.PurchasedFashion
## Premium Shopper                          0.1295403
## Regular Shopper                          0.1252064
##                  Product.Category.PurchasedHome Product.Category.PurchasedOthers
## Premium Shopper                     0.001907183                       0.01218914
## Regular Shopper                    -0.028865774                       0.06203545
##                        Gender
## Premium Shopper -0.008107097
## Regular Shopper -0.078946806
##
## Std. Errors:
##                  (Intercept)        Age Annual.Income..K..
## Premium Shopper  0.05827820 0.02680783         0.02677545
## Regular Shopper  0.05798534 0.02672753         0.02668951
##                  Average.Spend.per.Visit.... Number.of.Visits.in.Last.6.Months
## Premium Shopper                  0.02685104                       0.002651785
## Regular Shopper                  0.02676879                       0.002643341
##                  Product.Category.PurchasedBooks
## Premium Shopper                       0.05387690
## Regular Shopper                       0.05349586
##                  Product.Category.PurchasedElectronics
## Premium Shopper                             0.05588146
## Regular Shopper                             0.05584757
##                  Product.Category.PurchasedFashion
## Premium Shopper                         0.05649986
## Regular Shopper                         0.05646343
##                  Product.Category.PurchasedHome Product.Category.PurchasedOthers
## Premium Shopper                      0.05478646                       0.05314039
## Regular Shopper                      0.05504375                       0.05255361
##                      Gender
## Premium Shopper 0.05352182
```

```
## Regular Shopper 0.05335506
##
## Residual Deviance: 18493.17
## AIC: 18533.17
```

The only thing noticeable on the customer segmentation is that the age and annual income slightly has an influence on being a premium shopper, but in terms of shopping behavior, it is obvious that being a premium shopper would make you spend a higher average every visit, which is true on the analysis. We can also see that being a premium shopper also shows a tendency to lean on fashion products.

```r
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-8
```

```r
library(caret)
library(ggplot2)

set.seed(123)
train_data_balanced <- upSample(
  x = train_data[ , -which(names(train_data) == "Customer.Segment")],
  y = train_data$Customer.Segment,
  yname = "Customer.Segment"
)
X <- model.matrix(Customer.Segment ~ ., data = train_data_balanced)[, -1]
y <- train_data_balanced$Customer.Segment

cv_fit <- cv.glmnet(X, y, family = "multinomial", alpha = 0.5,
                    type.measure = "class", nfolds = 5, standardize = TRUE)

best_lambda <- cv_fit$lambda.min
cat("Best lambda:", best_lambda, "\n")
```

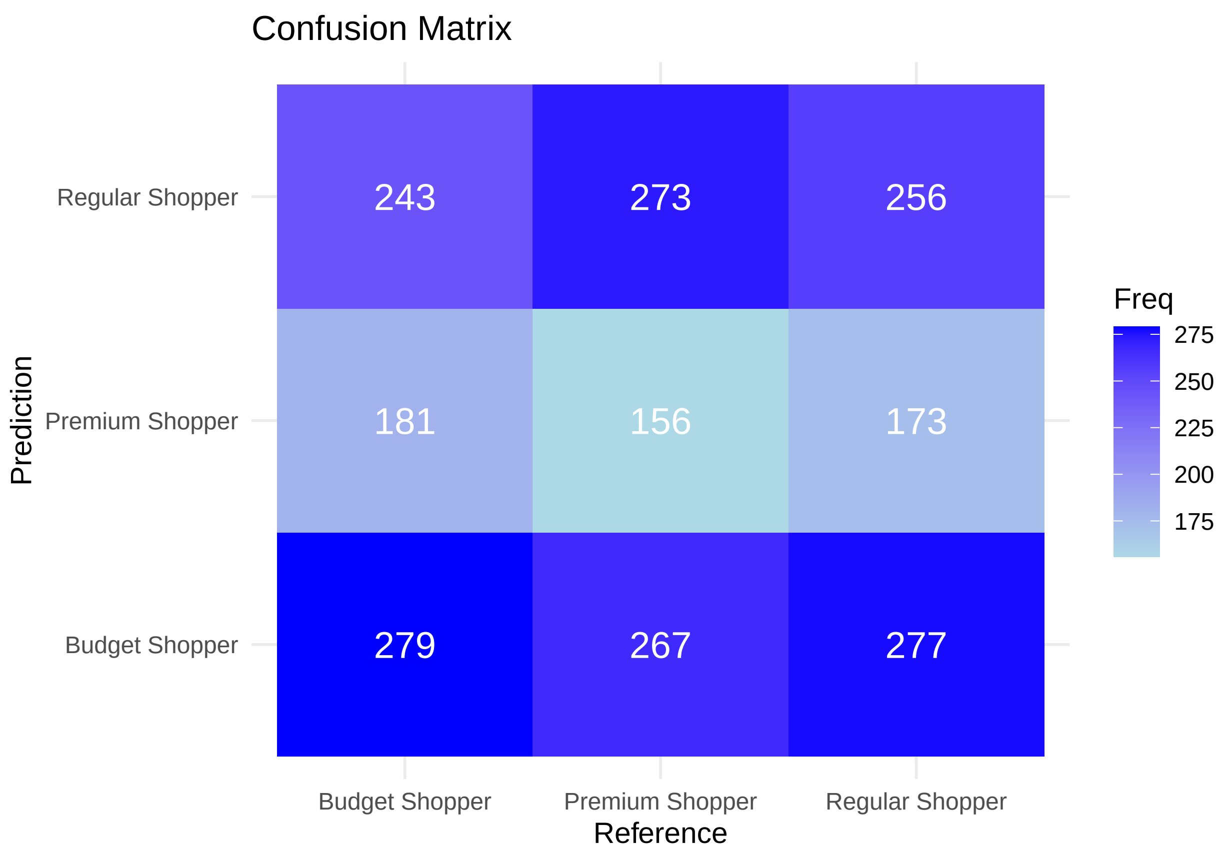
```
## Best lambda: 0.009726015
```

```r
final_model <- glmnet(X, y, family = "multinomial", alpha = 0.5, lambda = best_lambda, standardize = TRU

test_X <- model.matrix(Customer.Segment ~ ., data = test_data)[, -1]

pred_probs <- predict(final_model, newx = test_X, type = "response")[,,1]  # 3D array
pred_labels <- colnames(pred_probs)[max.col(pred_probs)]

class_levels <- levels(train_data$Customer.Segment)
pred_labels <- factor(pred_labels, levels = class_levels)
true_labels <- factor(test_data$Customer.Segment, levels = class_levels)

conf_matrix <- confusionMatrix(pred_labels, true_labels)
print(conf_matrix)
```

```
## Confusion Matrix and Statistics
```

```
##
##                    Reference
## Prediction          Budget Shopper Premium Shopper Regular Shopper
##   Budget Shopper               279             267             277
##   Premium Shopper              181             156             173
##   Regular Shopper              243             273             256
##
## Overall Statistics
##
##                  Accuracy : 0.3283
##                    95% CI : (0.3082, 0.3488)
##       No Information Rate : 0.3354
##       P-Value [Acc > NIR] : 0.7625
##
##                     Kappa : -0.0081
##
##  Mcnemar's Test P-Value : 6.067e-09
##
## Statistics by Class:
##
##                      Class: Budget Shopper Class: Premium Shopper
## Sensitivity                         0.3969                0.22414
## Specificity                         0.6120                0.74876
## Pos Pred Value                      0.3390                0.30588
## Neg Pred Value                      0.6693                0.66144
## Prevalence                          0.3340                0.33064
## Detection Rate                      0.1325                0.07411
## Detection Prevalence                0.3910                0.24228
## Balanced Accuracy                   0.5044                0.48645
##                      Class: Regular Shopper
## Sensitivity                          0.3626
## Specificity                          0.6312
## Pos Pred Value                       0.3316
## Neg Pred Value                       0.6624
## Prevalence                           0.3354
## Detection Rate                       0.1216
## Detection Prevalence                 0.3667
## Balanced Accuracy                    0.4969
```

```r
ggplot(as.data.frame(conf_matrix$table),
       aes(Reference, Prediction, fill = Freq)) +
  geom_tile() +
  geom_text(aes(label = Freq), color = "white", size = 5) +
  scale_fill_gradient(low = "lightblue", high = "blue") +
  theme_minimal() +
  labs(title = "Confusion Matrix")
```



Reporting The customer segmentation model was built using multinomial logistic regression with elastic net regularization. It classified shoppers into three (budget, regular, premium) groups based on their age, income, spending habits, and history. The model achieved 32.6 accuracy, which is worse than random guessing, even though there are already feature scaling and cross-validation. The confusion matrix showed frequent misclassifications particularly for premium shoppers, although we can see that the age and annual income slightly has an influence on being a premium shopper. This suggests that the model lack predictive power for clear segmentation. For future improvements, analysts should focus on testing advanced nonlinear models like XGBoost (Extreme Gradiant Boosting).