```r
# Load necessary libraries
library(caret)
```

```
## Warning: package ’caret’ was built under R version 4.3.2
```

```
## Loading required package: ggplot2
```

```
## Warning: package ’ggplot2’ was built under R version 4.3.2
```

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

```r
library(nortest)
library(lmtest)
```

```
## Warning: package ’lmtest’ was built under R version 4.3.2
```

```
## Loading required package: zoo
```

```
##
## Attaching package: ’zoo’
```

```
## The following objects are masked from ’package:base’:
##
##     as.Date, as.Date.numeric
```

```r
library(car)
```

```
## Warning: package ’car’ was built under R version 4.3.2
```

```
## Loading required package: carData
```

```r
library(MLmetrics)
```

```
## Warning: package ’MLmetrics’ was built under R version 4.3.2
```

```
##
## Attaching package: ’MLmetrics’
```

```
## The following objects are masked from ’package:caret’:
##
##     MAE, RMSE
```

```
## The following object is masked from ’package:base’:
##
##     Recall
```

```r
library(ggplot2)
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```r
# Load the dataset
library(readr)
df <- read_csv("ecommerce_customers.csv")
```

```
## Rows: 500 Columns: 8

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): Email, Address, Avatar
## dbl (5): Avg_Session_Length, Time_on_App, Time_on_Website, Length_of_Members...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
# Basic Data Examination
dim(df)
```

```
## [1] 500   8
```

```r
summary(df)
```

```
##     Email             Address             Avatar          Avg_Session_Length
##  Length:500         Length:500         Length:500         Min.   :29.53
##  Class :character   Class :character   Class :character   1st Qu.:32.34
##  Mode  :character   Mode  :character   Mode  :character   Median :33.08
##                                                           Mean   :33.05
##                                                           3rd Qu.:33.71
##                                                           Max.   :36.14
##   Time_on_App     Time_on_Website Length_of_Membership Yearly_Amount_Spent
##  Min.   : 8.508   Min.   :33.91   Min.   :0.2699       Min.   :256.7
##  1st Qu.:11.388   1st Qu.:36.35   1st Qu.:2.9304       1st Qu.:445.0
##  Median :11.983   Median :37.07   Median :3.5340       Median :498.9
##  Mean   :12.052   Mean   :37.06   Mean   :3.5335       Mean   :499.3
##  3rd Qu.:12.754   3rd Qu.:37.72   3rd Qu.:4.1265       3rd Qu.:549.3
##  Max.   :15.127   Max.   :40.01   Max.   :6.9227       Max.   :765.5
```

```r
# Check missing value
sapply(df, function(x) sum(is.na(x)))
```

```
##                Email              Address               Avatar
##                    0                    0                    0
##   Avg_Session_Length          Time_on_App      Time_on_Website
##                    0                    0                    0
## Length_of_Membership  Yearly_Amount_Spent
##                    0                    0
```

```
# Check and remove outliers
# Outlier removal functions
outliers <- function(x) {
  Q1 <- quantile(x, probs=.25)
  Q3 <- quantile(x, probs=.75)
  iqr = Q3-Q1
  upper_limit = Q3 + (iqr*1.5)
  lower_limit = Q1 - (iqr*1.5)
  x > upper_limit | x < lower_limit
}

remove_outliers <- function(df, cols = names(df)) {
  for (col in cols) {
    df <- df[!outliers(df[[col]]),]
  }
  df
}

df_new = remove_outliers(df, c('Avg_Session_Length', 'Time_on_App', 'Time_on_Website', 'Length_of_Membe

# Perform k-means clustering for segmentation
set.seed(42) # For reproducibility
num_clusters <- 3 # Define the number of clusters
clusters <- kmeans(df_new[,c('Avg_Session_Length', 'Time_on_App', 'Length_of_Membership')], centers = nu
df_new$cluster <- as.factor(clusters$cluster)

# Cross-validation setup
control <- trainControl(method = "cv", number = 10) # 10-fold cross-validation

# Fitting Model for each cluster with cross-validation
df_new$cluster <- as.numeric(as.character(df_new$cluster))
models <- list()
for (i in 1:num_clusters) {
  cluster_data <- subset(df_new, cluster == i)
  model <- train(Yearly_Amount_Spent ~ Avg_Session_Length + Time_on_App + Length_of_Membership,
                 data = cluster_data,
                 method = "lm",
                 trControl = control)
  models[[i]] <- model
}

final_models <- lapply(models, function(x) x$finalModel)
# Summarize models
lapply(models, summary)
```

```
## [[1]]
##
```

```
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.7280  -5.6735   0.0487   5.1195  23.4250
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1056.0146    36.5253  -28.91   <2e-16 ***
## Avg_Session_Length      26.0106     1.0175   25.56   <2e-16 ***
## Time_on_App             39.8862     1.2521   31.86   <2e-16 ***
## Length_of_Membership    61.3277     0.8852   69.28   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 8.606 on 151 degrees of freedom
## Multiple R-squared:  0.9831, Adjusted R-squared:  0.9828
## F-statistic:  2933 on 3 and 151 DF,  p-value: < 2.2e-16
##
##
## [[2]]
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.2254  -7.1032  -0.1117   7.1512  24.5780
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1080.915     41.642  -25.96   <2e-16 ***
## Avg_Session_Length      26.744      1.370   19.52   <2e-16 ***
## Time_on_App             39.699      1.433   27.71   <2e-16 ***
## Length_of_Membership    61.550      0.990   62.17   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 10.53 on 148 degrees of freedom
## Multiple R-squared:  0.9783, Adjusted R-squared:  0.9779
## F-statistic:  2229 on 3 and 148 DF,  p-value: < 2.2e-16
##
##
## [[3]]
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.6437  -6.8282   0.3189   7.0227  30.0919
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          -1022.306     55.162  -18.53   <2e-16 ***
## Avg_Session_Length      25.242      1.594   15.83   <2e-16 ***
## Time_on_App             39.204      1.113   35.24   <2e-16 ***
## Length_of_Membership    60.761      1.191   51.02   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 10.95 on 163 degrees of freedom
## Multiple R-squared:  0.9727, Adjusted R-squared:  0.9722 
## F-statistic:  1933 on 3 and 163 DF,  p-value: < 2.2e-16
```

```r
# Presenting model results using stargazer
stargazer(final_models, type = "text", title = "Regression Models for E-Commerce Customer Segments",
          header = FALSE, digits = 2, out = "models_results.txt")
```

```
## 
## Regression Models for E-Commerce Customer Segments
## ==================================================================================================
##                                                    Dependent variable:                            
##                       ----------------------------------------------------------------------------
##                                                         .outcome                                  
##                                  (1)                       (2)                       (3)          
## --------------------------------------------------------------------------------------------------
## Avg_Session_Length           26.01***                  26.74***                  25.24***         
##                               (1.02)                    (1.37)                    (1.59)          
##                                                                                                   
## Time_on_App                  39.89***                  39.70***                  39.20***         
##                               (1.25)                    (1.43)                    (1.11)          
##                                                                                                   
## Length_of_Membership         61.33***                  61.55***                  60.76***         
##                               (0.89)                    (0.99)                    (1.19)          
##                                                                                                   
## Constant                   -1,056.01***              -1,080.91***              -1,022.31***       
##                               (36.53)                   (41.64)                   (55.16)         
##                                                                                                   
## --------------------------------------------------------------------------------------------------
## Observations                    155                       152                       167           
## R2                             0.98                      0.98                      0.97           
## Adjusted R2                    0.98                      0.98                      0.97           
## Residual Std. Error       8.61 (df = 151)          10.53 (df = 148)          10.95 (df = 163)     
## F Statistic          2,932.77*** (df = 3; 151) 2,228.54*** (df = 3; 148) 1,933.10*** (df = 3; 163)
## ==================================================================================================
## Note:                                                                  *p<0.1; **p<0.05; ***p<0.01
```

```r
# Model evaluation can be extracted from the models' summaries, as cross-validation scores are included
```