Stuart School of Business

ILLINOIS INSTITUTE OF TECHNOLOGY

Westfield_ project

SONALI OJHA

REPORT - Westfield Shopping Center

INTRODUCTION

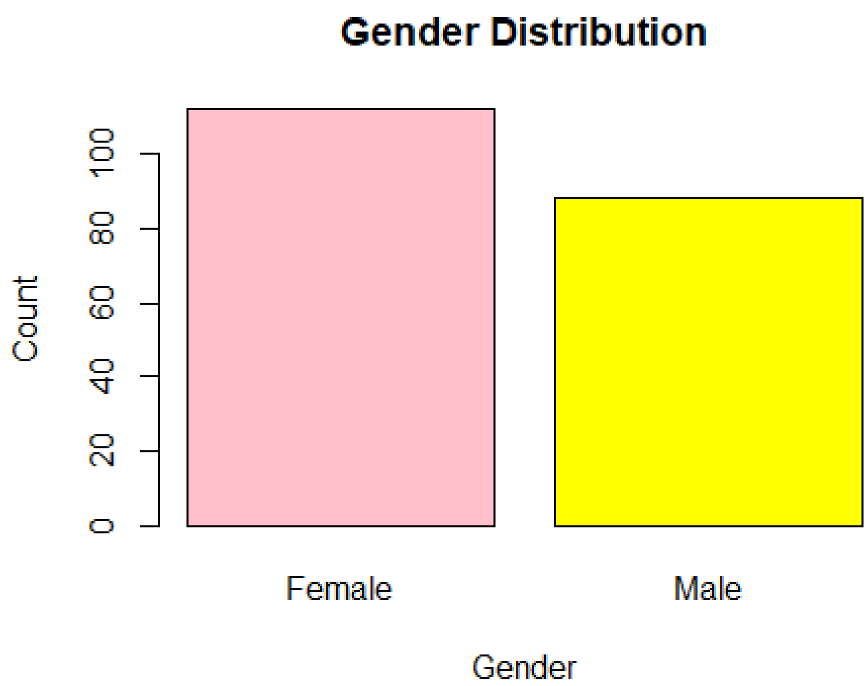
The Wilmette, Illinois mall Westfield Mall draws a diverse clientele from all walks of life. The management staff at the mall has a solid grasp of the preferences and traits of its patrons because it offers a variety of retailers to cater to their diverse needs. Segmenting customers

is one method by which management does this. This report's objective is to create consumer categories for Westfield Mall by analyzing customer statistics.

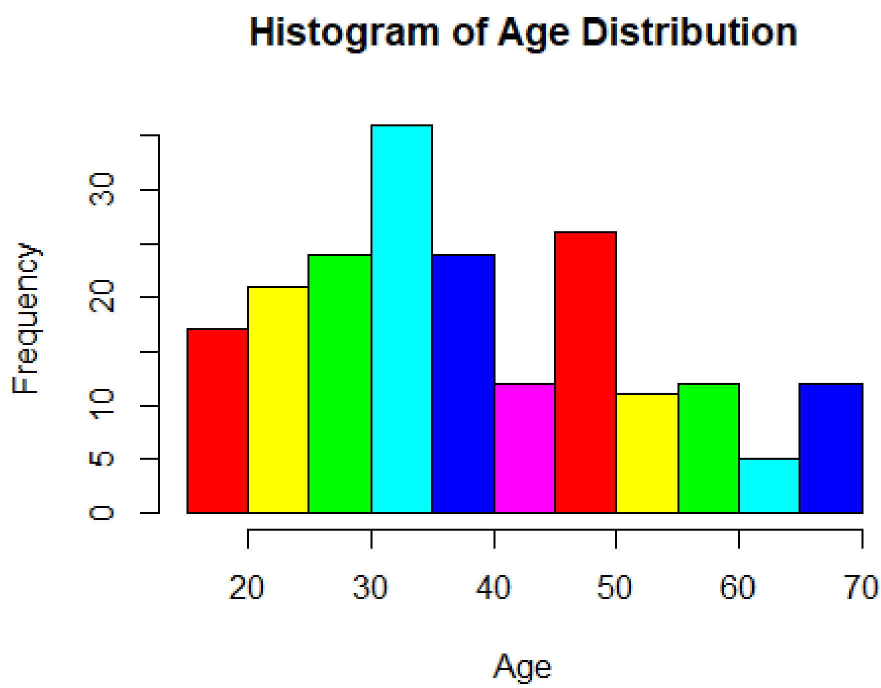
FINDINGS THROUGH RStudios:

```
#Step 1: Read the dataset
getwd()

## [1] "C:/Users/davey/Downloads"

westfld.df <- read.csv("westfield.csv")

#Step 2: Find the dimensions of the data set
dim(westfld.df)

## [1] 200   5

#Step 3: Summarize the data and showcase the structure of the dataset
str(westfld.df)

## 'data.frame':    200 obs. of  5 variables:
##  $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender                : chr  "Male" "Male" "Female" "Female" ...
##  $ Age                   : int  19 21 20 23 31 22 35 23 64 30 ...
##  $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
##  $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...

summary(westfld.df)

##    CustomerID        Gender               Age        Annual.Income..k..
##  Min.   :  1.00   Length:200         Min.   :18.00   Min.   : 15.00
##  1st Qu.: 50.75   Class :character   1st Qu.:28.75   1st Qu.: 41.50
##  Median :100.50   Mode  :character   Median :36.00   Median : 61.50
```

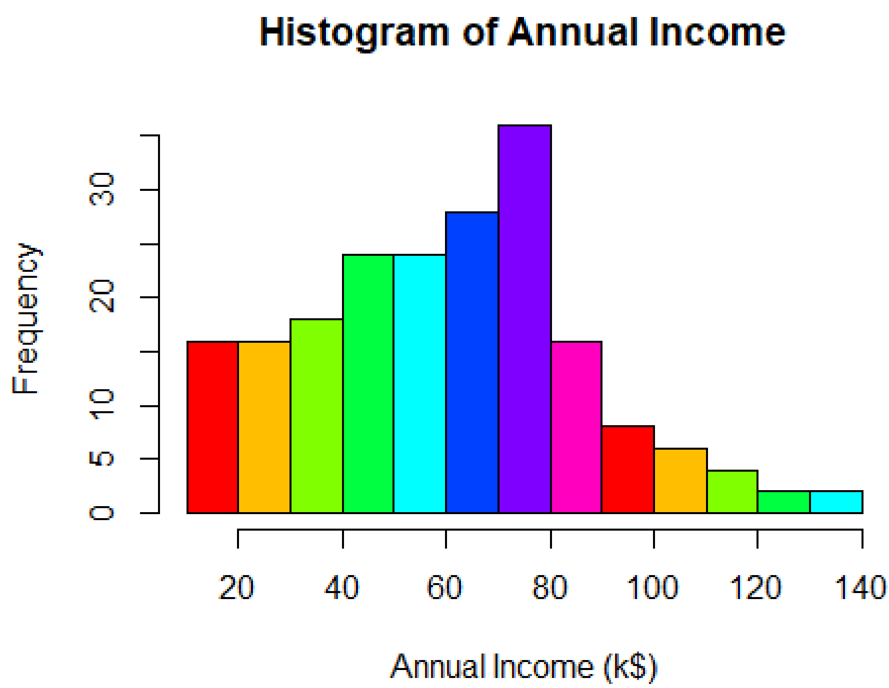
```
##  Mean   :100.50                     Mean   :38.85   Mean   : 60.56
##  3rd Qu.:150.25                     3rd Qu.:49.00   3rd Qu.: 78.00
##  Max.   :200.00                     Max.   :70.00   Max.   :137.00
##  Spending.Score..1.100.
##  Min.   : 1.00
##  1st Qu.:34.75
##  Median :50.00
##  Mean   :50.20
##  3rd Qu.:73.00
##  Max.   :99.00

# Convert Gender to factor
westfld.df$Gender<-factor(westfld.df$Gender)
str(westfld.df)

## 'data.frame':    200 obs. of  5 variables:
##  $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender                : Factor w/ 2 levels "Female","Male": 2 2 1 1 1 1
1 1 2 1 ...
##  $ Age                   : int  19 21 20 23 31 22 35 23 64 30 ...
##  $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
##  $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...

#####  Data Visualization
#Step 4: # Creating a Simple bar chart to show the gender distribution
count.gender<-table(westfld.df$Gender)
barplot(table(westfld.df$Gender), main = "Gender Distribution", xlab="Gender"
,ylab="Count",
        col=c("pink","yellow"))
```




```r
#Step 5: Visualize the Age using histogram
# Create a histogram to show the distribution of Age
hist(westfld.df$Age, main = "Histogram of Age Distribution",xlab = "Age", yla
b = "Frequency",
     col = rainbow(n = 6), breaks = 12)  # Rainbow palette with 6 colors
```




```r
#Step 6: Visualize the Annual Income using histogram
hist(westfld.df$Annual.Income..k.., main = "Histogram of Annual Income",
     xlab = "Annual Income (k$)", ylab = "Frequency",
     col = rainbow(n = 8))  # Rainbow palette with 8 colors
```

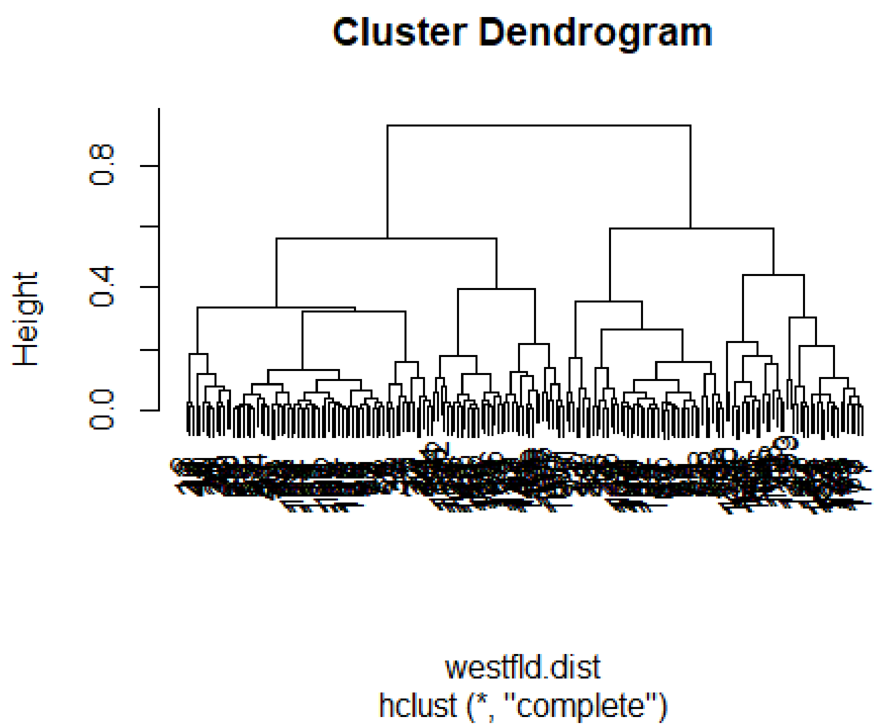



```
#Step 7: Remove the first column (CustomerID)
westfld.df1<-westfld.df[,-1]
str(westfld.df1)

## 'data.frame':    200 obs. of  4 variables:
##  $ Gender                : Factor w/ 2 levels "Female","Male": 2 2 1 1 1 1
1 1 2 1 ...
##  $ Age                   : int  19 21 20 23 31 22 35 23 64 30 ...
##  $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
##  $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...

### Developing a customer segmentation
# Remove the Age variable (assuming it's less relevant for your business goal
)
westfld.df2<-westfld.df1[,-2]
str(westfld.df2)

## 'data.frame':    200 obs. of  3 variables:
##  $ Gender                : Factor w/ 2 levels "Female","Male": 2 2 1 1 1 1
1 1 2 1 ...
##  $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
##  $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...
```

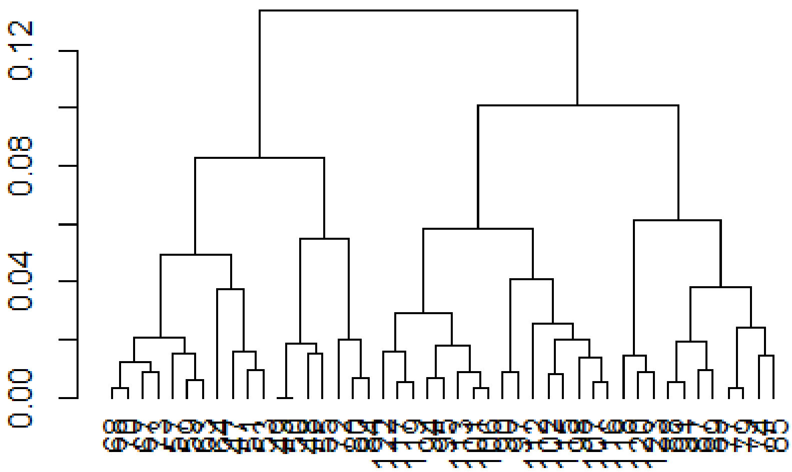
```
# Creating a function to look at mean values by group (if there is a grouped variable)
westfld.sum<-function(data,groups){
  aggregate(data,list(groups),function(x)mean(as.numeric(x)))
}
#summary
westfld.sum(westfld.df2,westfld.df1$Age)
```

```
##    Group.1   Gender Annual.Income..k.. Spending.Score..1.100.
## 1       18 1.750000           51.25000               60.00000
## 2       19 1.750000           57.00000               39.75000
## 3       20 1.600000           41.60000               40.20000
## 4       21 1.200000           38.80000               66.80000
## 5       22 1.333333           31.33333               70.00000
## 6       23 1.000000           41.50000               63.33333
## 7       24 1.500000           39.25000               71.50000
## 8       25 1.666667           57.66667               39.66667
## 9       26 2.000000           58.00000               54.50000
## 10      27 1.333333           63.16667               60.33333
## 11      28 1.750000           85.25000               70.00000
## 12      29 1.200000           63.60000               76.60000
## 13      30 1.285714           76.14286               80.28571
## 14      31 1.125000           48.37500               63.87500
## 15      32 1.454545           87.18182               66.00000
## 16      33 1.666667           80.33333               54.33333
## 17      34 1.400000           79.00000               39.20000
## 18      35 1.333333           46.66667               63.88889
## 19      36 1.333333           81.00000               52.50000
## 20      37 1.666667           65.00000               15.33333
## 21      38 1.333333           74.50000               63.16667
## 22      39 2.000000           72.66667               84.66667
## 23      40 1.500000           61.66667               47.50000
## 24      41 1.000000          101.00000               28.00000
## 25      42 1.500000           60.00000               18.50000
## 26      43 1.666667           65.66667               34.00000
## 27      44 1.000000           75.50000               13.50000
## 28      45 1.000000           69.33333               37.66667
## 29      46 1.333333           59.00000               21.33333
## 30      47 1.333333           70.16667               28.50000
## 31      48 2.000000           58.20000               41.80000
## 32      49 1.142857           51.00000               42.71429
## 33      50 1.200000           58.60000               45.80000
## 34      51 1.000000           55.50000               46.50000
## 35      52 1.500000           55.50000               21.00000
## 36      53 2.000000           39.50000               25.00000
## 37      54 1.250000           59.75000               35.75000
## 38      55 1.000000           57.00000               58.00000
```

```
## 39      56 1.000000          79.00000                35.00000
## 40      57 1.500000          64.50000                28.00000
## 41      58 1.500000          54.00000                15.00000
## 42      59 2.000000          65.25000                33.00000
## 43      60 1.333333          43.33333                36.33333
## 44      63 1.500000          56.50000                47.00000
## 45      64 2.000000          19.00000                 3.00000
## 46      65 1.500000          50.50000                43.50000
## 47      66 1.500000          63.00000                49.00000
## 48      67 1.750000          45.50000                41.50000
## 49      68 1.333333          56.66667                48.66667
## 50      69 2.000000          44.00000                46.00000
## 51      70 2.000000          47.50000                55.50000

# Using daisy for distance calculation
library(cluster)

## Warning: package 'cluster' was built under R version 4.3.3

# Calculate distances between data points (works with mixed data types)
westfld.dist<-daisy(westfld.df2)
as.matrix(westfld.dist)[1:2,1:2]

##           1         2
## 1 0.0000000 0.1428571
## 2 0.1428571 0.0000000

## this output provides a measure of the dissimilarity between the first two
#customer observations based on the variables used in the analysis

## Segmentation Technique ##
## hclust() clustring
westfld.hc<-hclust(westfld.dist, method = "complete")
# Visualize the hierarchical clustering dendrogram
plot(westfld.hc)
```




```
##This plot represents the results of the hierarchical
#clustering analysis performed on the customer data.

# Cut the dendrogram at a specific height
plot(cut(as.dendrogram(westfld.hc), h = 0.3)$lower[[2]])
```




```
# Examine observations linked at lower and higher heights
westfld.df2[c(12, 15), ]  # Similar (linked at low height)

##    Gender Annual.Income..k.. Spending.Score..1.100.
## 12 Female                 19                     99
## 15   Male                 20                     13

print("Similar observations (linked at low height):")

## [1] "Similar observations (linked at low height):"

#These two observations are considered similar and are linked at a low height
.
#Despite differences in gender, their annual incomes are close, and both have
#relatively extreme spending scores, one very high and one very low.
westfld.df2[c(87, 198), ]  # Less similar (joined at higher level)

##     Gender Annual.Income..k.. Spending.Score..1.100.
## 87  Female                 57                     58
## 198   Male                126                     74

print("Less similar observations (joined at higher level):")

## [1] "Less similar observations (joined at higher level):"
```

```
#These two observations are considered less similar and are joined at a highe
r level.
#They have more differences in both annual income and spending score compared
to the first pair.

#Cophenetic correlation coefficient (measures distance preservation)
cor(cophenetic(westfld.hc),westfld.dist)

## [1] 0.8671576

#This value indicates a strong positive correlation
cophenetic(westfld.hc)

##               1           2           3           4           5
6
## 2   0.354633657
## 3   0.931247909 0.931247909
## 4   0.931247909 0.931247909 0.336121334
## 5   0.931247909 0.931247909 0.158748745 0.336121334
## 6   0.931247909 0.931247909 0.336121334 0.006133601 0.336121334
## 7   0.931247909 0.931247909 0.005464481 0.336121334 0.158748745 0.33612133
4
## 8   0.931247909 0.931247909 0.336121334 0.183896509 0.336121334 0.18389650
9
## 9   0.168227947 0.354633657 0.931247909 0.931247909 0.931247909 0.93124790
9
## 10  0.931247909 0.931247909 0.336121334 0.025203524 0.336121334 0.02520352
4
## 11  0.168227947 0.354633657 0.931247909 0.931247909 0.931247909 0.93124790
9
## 12  0.931247909 0.931247909 0.336121334 0.183896509 0.336121334 0.18389650
9
## 13  0.931247909 0.931247909 0.086595294 0.336121334 0.158748745 0.33612133
4
## 14  0.931247909 0.931247909 0.336121334 0.011598082 0.336121334 0.01159808
2
## 15  0.168227947 0.354633657 0.931247909 0.931247909 0.931247909 0.93124790
9
## 16  0.354633657 0.020463923 0.931247909 0.931247909 0.931247909 0.93124790
9
## 17  0.931247909 0.931247909 0.158748745 0.336121334 0.027935764 0.33612133
4
## 18  0.354633657 0.073547452 0.931247909 0.931247909 0.931247909 0.93124790
9
## 19  0.055871529 0.354633657 0.931247909 0.931247909 0.931247909 0.93124790
9
## 20  0.931247909 0.931247909 0.336121334 0.183896509 0.336121334 0.18389650
```

```
9
## 21  0.055871529 0.354633657 0.931247909 0.931247909 0.931247909 0.93124790
9
## 22  0.354633657 0.073547452 0.931247909 0.931247909 0.931247909 0.93124790
9
## 23  0.931247909 0.931247909 0.027991524 0.336121334 0.158748745 0.33612133
4
## 24  0.354633657 0.073547452 0.931247909 0.931247909 0.931247909 0.93124790
9
## 25  0.931247909 0.931247909 0.086595294 0.336121334 0.158748745 0.33612133
4
## 26  0.354633657 0.038920486 0.931247909 0.931247909 0.931247909 0.93124790
9
## 27  0.931247909 0.931247909 0.158748745 0.336121334 0.102263856 0.33612133
4
## 28  0.354633657 0.260399242 0.931247909 0.931247909 0.931247909 0.93124790
9
## 29  0.931247909 0.931247909 0.158748745 0.336121334 0.102263856 0.33612133
4
## 30  0.931247909 0.931247909 0.336121334 0.078342813 0.336121334 0.07834281
3
## 31  0.168227947 0.354633657 0.931247909 0.931247909 0.931247909 0.93124790
9
## 32  0.931247909 0.931247909 0.336121334 0.078342813 0.336121334 0.07834281
3
## 33  0.168227947 0.354633657 0.931247909 0.931247909 0.931247909 0.93124790
9
## 34  0.354633657 0.134883462 0.931247909 0.931247909 0.931247909 0.93124790
9
## 35  0.931247909 0.931247909 0.086595294 0.336121334 0.158748745 0.33612133
4
## 36  0.931247909 0.931247909 0.336121334 0.078342813 0.336121334 0.07834281
3
## 37  0.931247909 0.931247909 0.086595294 0.336121334 0.158748745 0.33612133
4
## 38  0.931247909 0.931247909 0.336121334 0.078342813 0.336121334 0.07834281
3
## 39  0.931247909 0.931247909 0.158748745 0.336121334 0.102263856 0.33612133
4
## 40  0.931247909 0.931247909 0.336121334 0.078342813 0.336121334 0.07834281
3
## 41  0.931247909 0.931247909 0.158748745 0.336121334 0.102263856

..
## 150
## 151
## 152
```

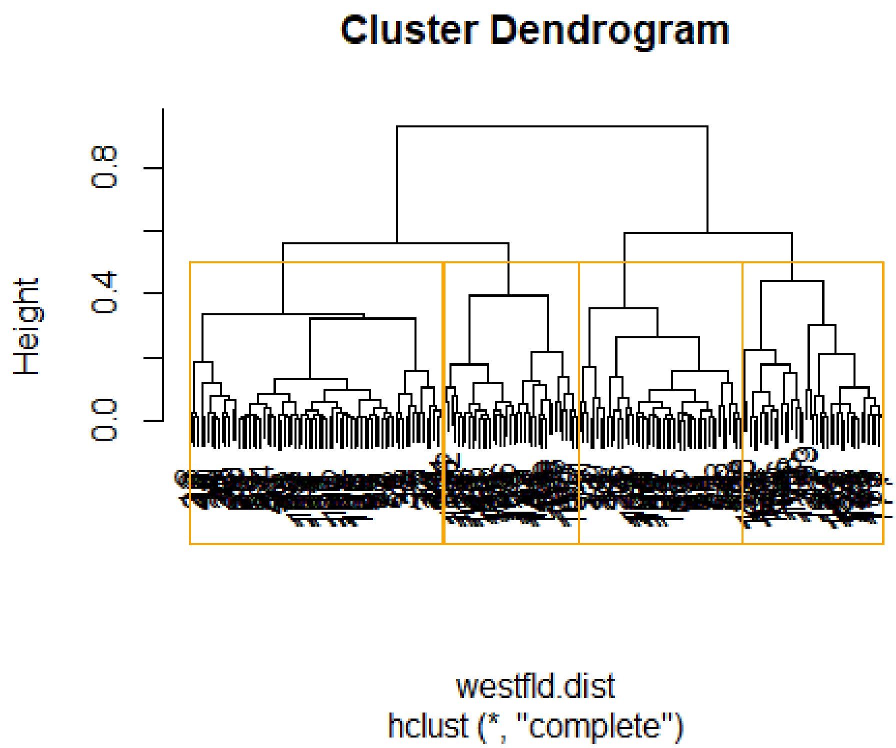
```
## 153
## 154
## 155
## 156
## 157
## 158
## 159
## 160
## 161
## 162
## 163
## 164
## 165
## 166
## 167
## 168
## 169
## 170
## 171
## 172
## 173
## 174
## 175
## 176
## 177
## 178
## 179
## 180
## 181
## 182 0.396342143
## 183 0.931247909 0.931247909
## 184 0.396342143 0.009534962 0.931247909
## 185 0.029274005 0.396342143 0.931247909 0.396342143
## 186 0.931247909 0.931247909 0.442232631 0.931247909 0.931247909
## 187 0.085758894 0.396342143 0.931247909 0.396342143 0.085758894 0.93124790
9
## 188 0.931247909 0.931247909 0.442232631 0.931247909 0.931247909 0.14843314
4
## 189 0.085758894 0.396342143 0.931247909 0.396342143 0.085758894 0.93124790
9
## 190 0.396342143 0.023865284 0.931247909 0.023865284 0.396342143 0.93124790
9
## 191 0.085758894 0.396342143 0.931247909 0.396342143 0.085758894 0.93124790
9
## 192 0.396342143 0.078287053 0.931247909 0.078287053 0.396342143 0.93124790
9
## 193 0.931247909 0.931247909 0.305007249 0.931247909 0.931247909 0.44223263
```

```
1
## 194 0.396342143 0.102152336 0.931247909 0.102152336 0.396342143 0.93124790
9
## 195 0.135608342 0.396342143 0.931247909 0.396342143 0.135608342 0.93124790
9
## 196 0.396342143 0.102152336 0.931247909 0.102152336 0.396342143 0.93124790
9
## 197 0.135608342 0.396342143 0.931247909 0.396342143 0.135608342 0.93124790
9
## 198 0.931247909 0.931247909 0.442232631 0.931247909 0.931247909 0.22170179
5
## 199 0.931247909 0.931247909 0.305007249 0.931247909 0.931247909 0.44223263
1
## 200 0.931247909 0.931247909 0.442232631 0.931247909 0.931247909 0.22170179
5
##             187         188         189         190         191         19
2
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
## 29
## 30
## 31
```

```
## 32
## 33
## 34
## 35
.
## 184
## 185
## 186
## 187
## 188 0.931247909
## 189 0.029274005 0.931247909
## 190 0.396342143 0.931247909 0.396342143
## 191 0.008865841 0.931247909 0.029274005 0.396342143
## 192 0.396342143 0.931247909 0.396342143 0.078287053 0.396342143
## 193 0.931247909 0.442232631 0.931247909 0.931247909 0.931247909 0.93124790
9
## 194 0.396342143 0.931247909 0.396342143 0.102152336 0.396342143 0.10215233
6
## 195 0.135608342 0.931247909 0.135608342 0.396342143 0.135608342 0.39634214
3
## 196 0.396342143 0.931247909 0.396342143 0.102152336 0.396342143 0.10215233
6
## 197 0.135608342 0.931247909 0.135608342 0.396342143 0.135608342 0.39634214
3
## 198 0.931247909 0.221701795 0.931247909 0.931247909 0.931247909 0.93124790
9
## 199 0.931247909 0.442232631 0.931247909 0.931247909 0.931247909 0.93124790
9
## 200 0.931247909 0.221701795 0.931247909 0.931247909 0.931247909 0.93124790
9
##             193         194         195         196         197         19
8
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
```

```
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
##
## 186
## 187
## 188
## 189
## 190
## 191
## 192
## 193
## 194 0.931247909
## 195 0.931247909 0.396342143
## 196 0.931247909 0.059942010 0.396342143
## 197 0.931247909 0.396342143 0.057209769 0.396342143
## 198 0.442232631 0.931247909 0.931247909 0.931247909 0.931247909
## 199 0.099587376 0.931247909 0.931247909 0.931247909 0.931247909 0.44223263
1
## 200 0.442232631 0.931247909 0.931247909 0.931247909 0.931247909 0.06066689
0
##             199
## 2
## 3
## 4
## 5
## 6
## 7
..
## 186
## 187
## 188
## 189
## 190
## 191
## 192
## 193
## 194
## 195
## 196
## 197
## 198
```

```
## 199
## 200 0.442232631

#This indicates a good fit for using hierarchical clustering in this case.

# Cut the dendrogram into 4 clusters
plot(westfld.hc)

# Cut dendrogram at k=4 and highlighted with orange border
rect.hclust(westfld.hc,k=4,border = "orange")
```

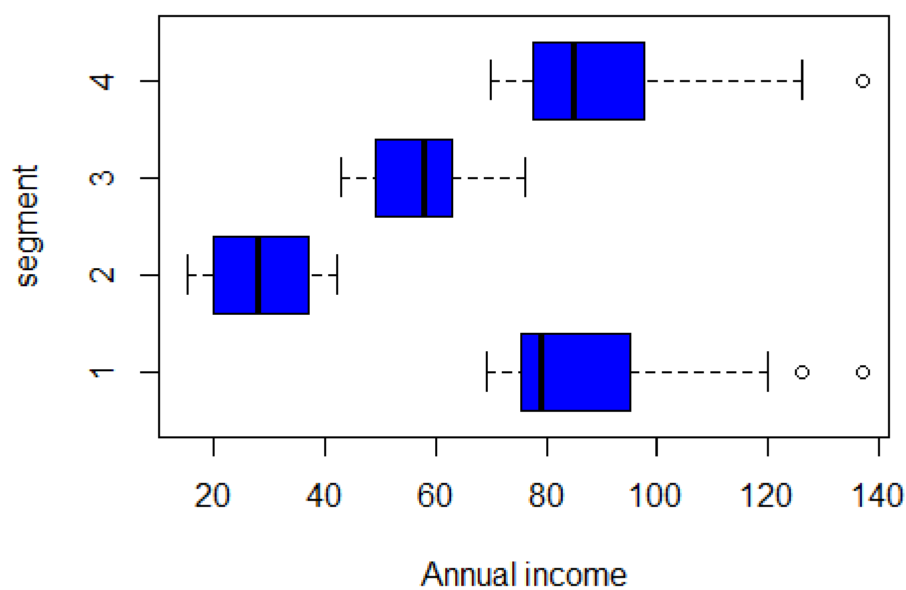



```
#This visually highlights four clusters within the hierarchical
#clustering for easier interpretation of the data groupings.
#The orange rectangle emphasizes the specific level
#in the dendrogram where these four clusters are formed.

#cluster labels (1-4) to data points
westfld.hc.segment<-cutree(westfld.hc,k=4)

#create a frequency table summarizing the distribution of data points across
the identified clusters.
table(westfld.hc.segment)
```

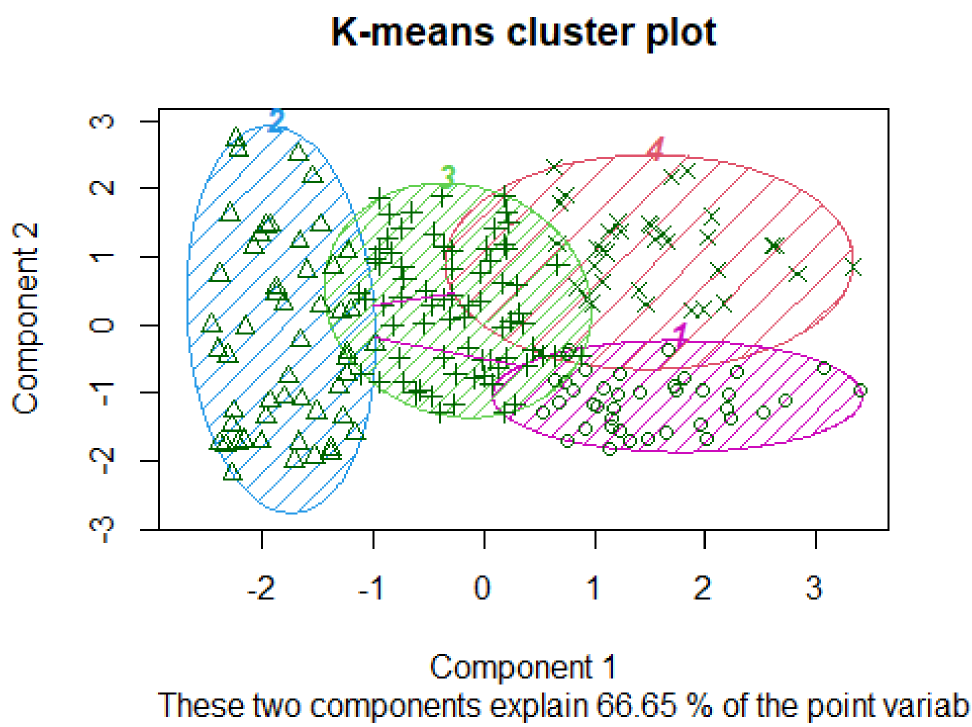
```
## westfld.hc.segment
##  1  2  3  4
## 47 73 41 39

#Hierarchical clustering assigned data points to 4 groups (clusters 1-4).
#These labels are arbitrary (don't have inherent meaning).
#Cluster 2 has the most members (73), followed by 1 (47), 3 (41), and 4 (39)


#calculation of summary grouped by the cluster membership stored
westfld.sum(westfld.df,westfld.hc.segment)

##   Group.1 CustomerID Gender      Age Annual.Income..k.. Spending.Score..1.
100.
## 1       1   60.02128      2 42.14894           43.21277               50.5
5319
## 2       2   62.06849      1 38.98630           44.19178               49.7
1233
## 3       3  154.92683      2 37.12195           84.02439               46.1
7073
## 4       4  164.00000      1 36.43590           87.43590               54.9
2308

#This table summarizes customer data by their assigned cluster (1-4). It lets
you compare characteristics
#like average spending and income across different customer segments.
#inspecting variables in westfld.df2 with reference to 8 clusters
westfld.sum(westfld.df2,westfld.hc.segment)

##   Group.1 Gender Annual.Income..k.. Spending.Score..1.100.
## 1       1      2           43.21277               50.55319
## 2       2      1           44.19178               49.71233
## 3       3      2           84.02439               46.17073
## 4       4      1           87.43590               54.92308

###K-MEANS
#k-means requires numeric input and specified no. of clusters
westfld.df.num<-westfld.df2
westfld.df.num$Gender<-ifelse(westfld.df2$Gender=="Male",0,1)
#summary and structure
summary(westfld.df.num)

##      Gender     Annual.Income..k.. Spending.Score..1.100.
##  Min.   :0.00   Min.   : 15.00     Min.   : 1.00
##  1st Qu.:0.00   1st Qu.: 41.50     1st Qu.:34.75
##  Median :1.00   Median : 61.50     Median :50.00
##  Mean   :0.56   Mean   : 60.56     Mean   :50.20
```

```
##  3rd Qu.:1.00   3rd Qu.: 78.00     3rd Qu.:73.00
##  Max.   :1.00   Max.   :137.00     Max.   :99.00

str(westfld.df.num)

## 'data.frame':    200 obs. of  3 variables:
##  $ Gender                : num  0 0 1 1 1 1 1 1 0 1 ...
##  $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
##  $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...

##k=4 groups
set.seed(96734) #because starting assignments are random
westfld.k<-kmeans(westfld.df.num, centers = 4)
westfld.sum(westfld.df2,westfld.k$cluster)

##   Group.1   Gender Annual.Income..k.. Spending.Score..1.100.
## 1       1 1.461538           86.53846               82.12821
## 2       2 1.365385           27.96154               49.73077
## 3       3 1.432432           56.70270               49.35135
## 4       4 1.542857           88.20000               17.11429

#K-Means clustering created customer segments. Cluster 1 has the highest inco
me and spending,
#while Cluster 4 has the lowest spending.
#"Gender" values don't directly represent male/female proportions.

#Group: Cluster number assigned by k-means.
#Gender: The average proportion of females in each cluster.
#a value of 1.46 means that, on average, 1.46 out of every 2 customers in tha
t cluster are female.
#Annual Income (k$): The average annual income of customers in each cluster.
#Spending Score (1-100): The average spending score of customers in each clus
ter.

#Interpretation:
##Cluster 1: This cluster has a relatively higher proportion of females (1.46
out of 2)
#compared to males. On average, customers in this cluster have a high annual
income (86.54k)
#and a high spending score (82.13).
##Cluster 2: This cluster has a slightly lower proportion of females (1.37 ou
t of 2).
#Customers in this cluster have a lower annual income (27.96k) compared to Cl
uster 1,
#but they still have a moderate spending score (49.73).
##Cluster 3: Similar to Cluster 2, this cluster also has a slightly lower pro
portion of females (1.43 out of 2).
#Customers in this cluster have a moderate annual income (56.70k) and a moder
```

```
ate spending score (49.35).
##Cluster 4: This cluster has a higher proportion of females (1.54 out of 2)
#compared to Clusters 2 and 3. Customers in this cluster have a high annual i
ncome (88.20k),
#but they have a relatively low spending score (17.11).


##Visualizing the annual income and segment using $cluster
boxplot(westfld.df.num$Annual.Income..k.. ~ westfld.k$cluster,col="blue",
        xlab="Annual income", ylab = "segment", horizontal = TRUE)
```



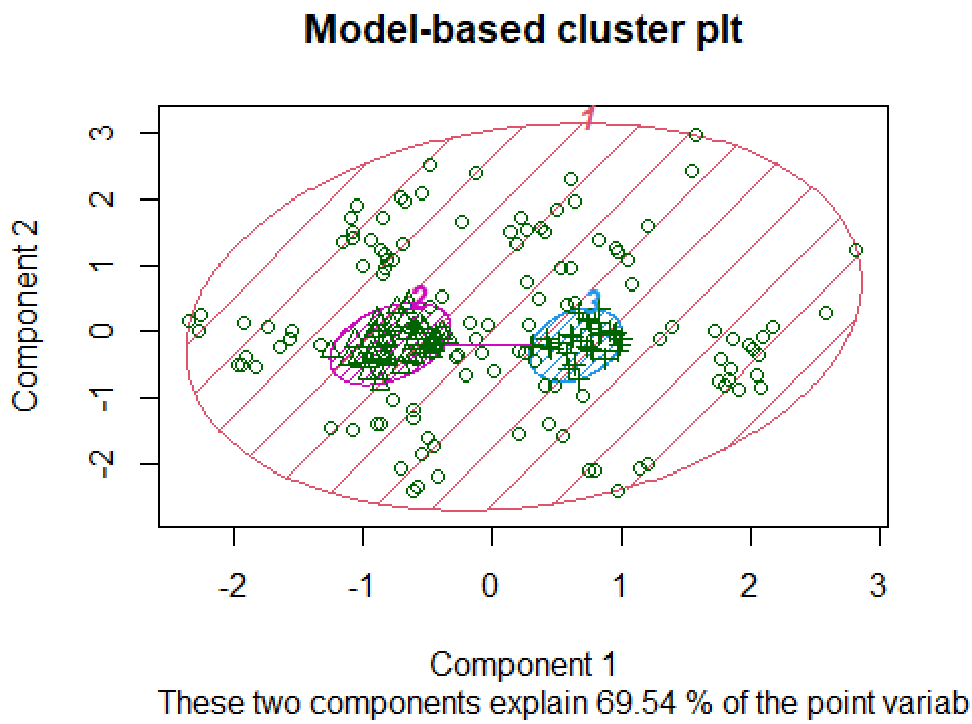

```
#The boxplot result shows how annual income is
#spread out across different customer groups

###visualizing the overall clusters
##clusplot ()
library(cluster)
clusplot(westfld.df,westfld.k$cluster,color = TRUE,shade = TRUE, labels = 4,
         main = "K-means cluster plot")
```




```
#This K-means cluster plot suggests that the customers can be grouped
#into four segments with some potential overlap between clusters.
#The tightness of the ellipses indicates a variation in
#the level of distinctiveness between the customer segments.
##The blue and green ellipses are relatively tight, indicating a more concent
rated
##distribution of customers within those clusters.
##The red and purple ellipses are more scattered, suggesting a wider range of
##characteristics among customers assigned to those segments.


#Model-based clustring: Mclust()
## data must be numeric
library(mclust)

## Warning: package 'mclust' was built under R version 4.3.3

## Package 'mclust' version 6.1
## Type 'citation("mclust")' for citing this R package in publications.

westfld.mc<-Mclust(westfld.df.num) #use all defaults
summary(westfld.mc)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEI (diagonal, equal shape) model with 3 components:
##
##  log-likelihood   n df       BIC      ICL
##        -1867.94 200 16 -3820.652 -3829.28
##
## Clustering table:
##   1   2   3
## 122  46  32

#This implies there might be three underlying groups (customer segments)
#with different characteristics within the data.lower ICL values indicate bet
ter models.
#Here, the ICL value is -3829.28.


#Mclust for 4 groups
westfld.mc4<-Mclust(westfld.df.num,G=4)   #4 clusters
summary(westfld.mc4)

## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EII (spherical, equal volume) model with 4 components:
##
##  log-likelihood   n df       BIC       ICL
##       -2531.863 200 16 -5148.499 -5158.072
##
## Clustering table:
##   1   2   3   4
##  20  34 107  39

BIC(westfld.mc,westfld.mc4) #A lower BIC value generally indicates a better

##             df      BIC
## westfld.mc  16 3820.652
## westfld.mc4 16 5148.499

#model that balances goodness of fit and model complexity.
westfld.sum(westfld.df2,westfld.mc$class)

##   Group.1   Gender Annual.Income..k.. Spending.Score..1.100.
## 1       1 1.459016           64.36066               50.31148
```

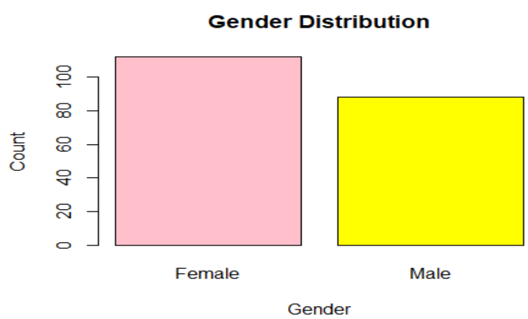
```
## 2       2 1.000000          54.54348              49.52174
## 3       3 2.000000          54.71875              50.75000

#In this case, westfld.mc (BIC=-3820.652) has a lower BIC compared to westfld
.mc4 (BIC=-5148.499),
#suggesting the 3-cluster model might be a better fit based on this criterion
.

##plot for Mclust model
clusplot(westfld.df2,westfld.mc$class,color = TRUE,shade = TRUE,labels = 4,
         main = "Model-based cluster plt")
```

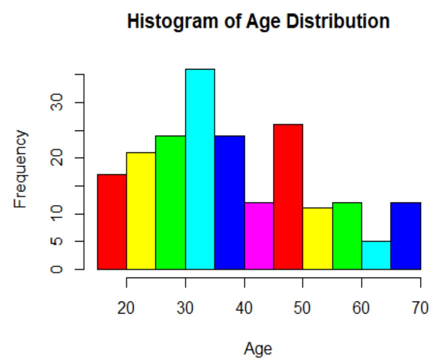


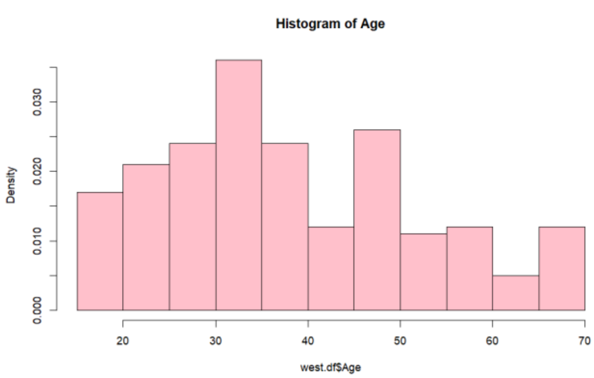

```
#Customers in cluster 2 share some characteristics with
#cluster 1 but also have additional variations that differentiate them.
#Cluster 1 might represent a broader customer segment,
#while cluster 2 is a more specific sub-segment within it.
#these two principal components capture almost 70% of the
#essential differences between the customer data points.
#The plot visually represents this nesting by showing the smaller ellipses of
clusters 2 and 3
#entirely positioned within the larger blue ellipse of cluster 1. This sugges
ts that the data points in
#clusters 2 and 3 have characteristics that fall entirely within the broader
range of data points captured by cluster 1.
```

FINDINGS

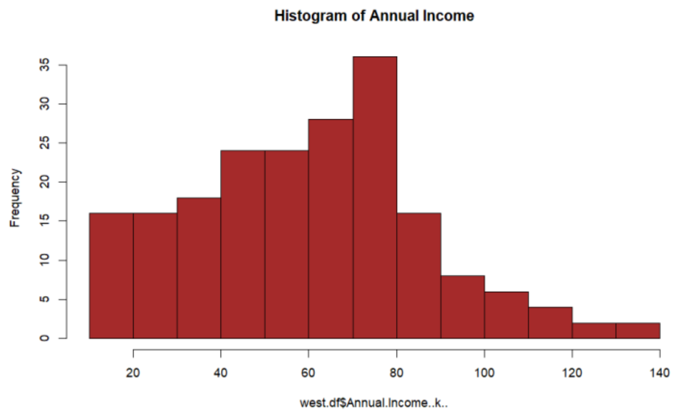
- Visualize the Age using histogram



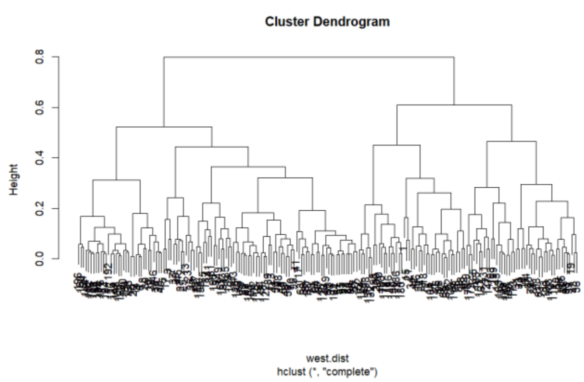


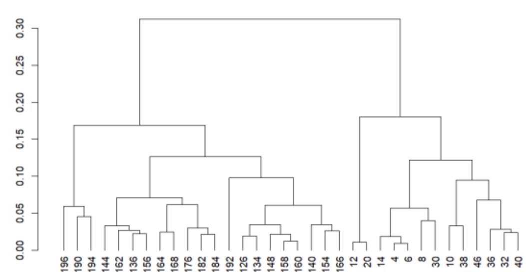


- Visualize the Annual Income using histogram

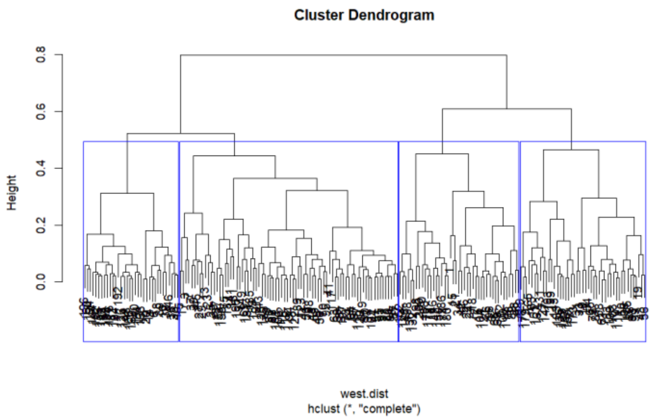


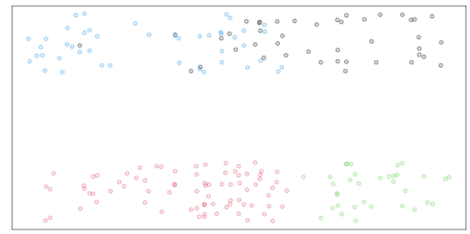
- hclust()



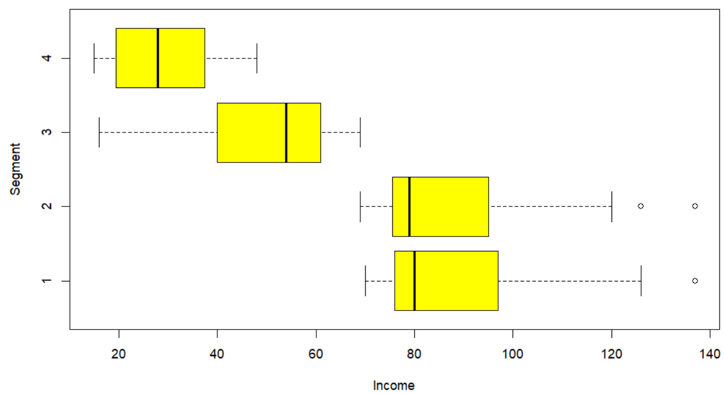


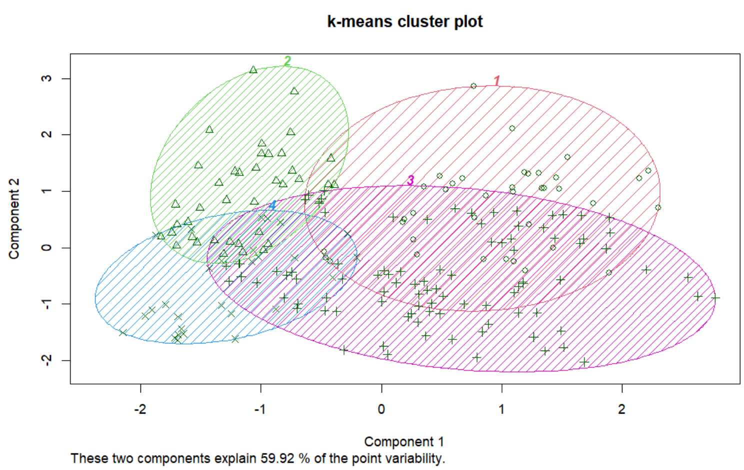
Westfield_ project



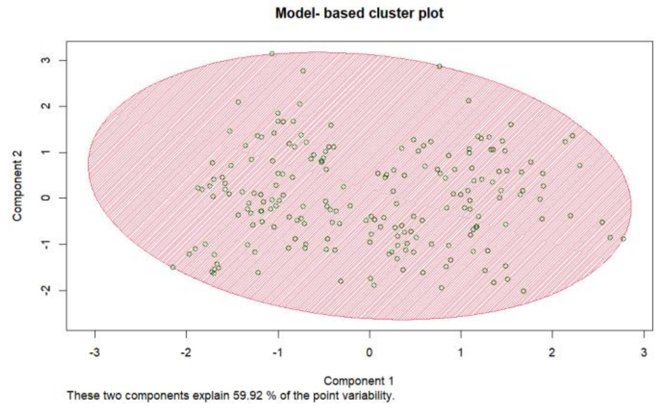


- k means


Segment
Income
1
2
3
4
20
40
60
80
100
120
140


k-means cluster plot
Component 1
Component 2
These two components explain 59.92 % of the point variability.

- Mclust



<u>First segment</u>:
high-spenders (26.5 percent of clients)
Customers in this sector have both a high spending score and a high annual income. As these

Customers that are prepared to shell out a significant sum of money for shopping are an important source of business for Westfield Mall.

<u>Segment 2</u>:
Customers with Average Spending (34.5%)
Customers in this sector have an average expenditure score and an average annual income. These consumers make up the biggest group and are probably price conscious when they purchase.

<u>Third Segment</u>:
Customers with Low Spending (18.5%)
Customers in this category have low spending scores and low annual incomes. As these

When shopping, clients are probably going to be price conscious and not likely to spend a lot of money at the mall.

<u>Segment 4:</u>
Customers with High Incomes but Low Spending (20.5%)
Customers in this sector have low spending scores but high annual incomes. Though they might not be interested in spending a lot of money at the mall, these customers might have a great earning potential.

Why K MEAN method?

Performing clustering techniques like hierarchical clustering, k-means, and model-based clustering is part of the analysis of consumer datasets. These methods are employed to divide up the clientele according to attributes like spending percentage, age, gender, and yearly income.

Performing clustering techniques like hierarchical clustering, k-means, and model-based clustering is part of the analysis of consumer datasets. These methods are employed to divide up the clientele according to attributes like spending percentage, age, gender, and yearly income.

Four separate client segments were also identified by the model-based clustering results.

BUSINESS GOALS

- Increase Revenue:
  To enhance revenue, improve client spending across all segments to drive sales and create new revenue sources.

- Boost Customer Engagement:
  Raise customer satisfaction and loyalty to encourage referrals and repeat business.

- Increase Market Reach:
  To increase the number of patrons at the mall, draw in new markets while holding onto your current clientele.

- Create Brand Differentiation:
  Present Westfield Mall as a top shopping destination renowned for its wide range of products, first-rate customer support, and customized experiences.

Marketing strategy

- **Targeted Segmentation and Personalization:**
  Make use of consumer information to divide the market into manageable segments and tailor offers and marketing materials to each one. Use CRM systems to monitor the preferences and actions of your customers so that you may send them customized messages and offers.

- **Quality Products and Services:**
  Place a strong emphasis on the caliber of goods and services offered to all clientele, making sure that affluent clients receive opulent options and middle-class and lower-class clients receive value-for-money goods that satisfy their requirements and tastes.

- **Discounts and Promotions:**
  Apply specific discounts and promotions based on the purchasing patterns of each category. To promote more spending, give high-spending clients access to special offers and incentives while offering middle-class and lower-income consumers value-driven promotions.

- **Enhancing the Customer Experience:**
  Make an investment to make the entire shopping experience better with facilities, practicality, and customized services. Establish tiered loyalty programs to encourage higher spending and return visits from all client categories.

- **Omnichannel Presence:**
  Create an omnichannel marketing strategy to connect with consumers through a variety of touchpoints, such as social media, mobile apps, websites, and physical stores. To offer a unified brand experience, make sure everything is smooth and consistent across all platforms.

- **Community Engagement and Events:**
  To promote a feeling of community and engagement, plan events, workshops, and recreational pursuits. Work together with local organizations and personalities to generate excitement and draw more shoppers to the mall.

- **Analysis and optimization ongoing:**
  Keep an eye on key performance indicators (KPIs) such consumer spending trends, satisfaction scores, and loyalty program participation on a regular basis. Utilize data analytics to spot patterns, chances, and areas that need work so that marketing tactics can be continuously improved.

Conclusion

To sum up, we have determined four distinct client segments by taking into account their spending score, age, gender, and annual income. To target each segment, we have suggested a variety of marketing techniques. Effectively focusing on each category will help the mall's sales and income.