File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Ecommerce CLustering Project.R | customerData_Hier | customerData_km | customerData_sc | top_customers | custo

Source on Save | Run | Source

```
#Selecting working Directory in my pc
setwd(choose.dir())
#Confirming the working directory path where Ecommrce Data Lies
getwd()

#install.packages("factoextra")--->example of installing necessary Libraries)
#Loading necessary Libraries
library(dplyr)
library(ggplot2)
library(cluster)
library(factoextra)
library(gridExtra)
library(purrr)

#Importing dataset from current working directory
data_ecom = read.csv("Ecommerce.csv")

#Viewing Dataset
View(data_ecom)
#Viewing first 6 rows of dataset in the console
head(data_ecom)

#Structure of the dataset
str(data_ecom)
#There is unnecessary column X with NA Values, Lets drop it
data_ecom <- subset(data_ecom,select=-X)

#We also saw  date data in the character format,lets change it to Date format
data_ecom$InvoiceDate <- as.Date(data_ecom$InvoiceDate, "%d-%B-%y")

#Lets recheck the columns and their structure, and view it again
str(data_ecom)
View(data_ecom)
```

132:1 (Top Level) | R Script

Console

Environment | History | Connections | Tutorial

Import | 715 MiB | List

R | Global Environment

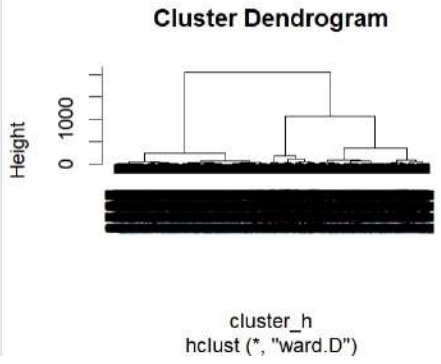
Data

customerDa…	3730 obs. of 2 variables
customerDa…	num [1:3730, 1:3] -1.091 1.…
customerDa…	num [1:3730, 1:3] -1.091 1.…
customerDa…	num [1:3730, 1:2] -1.091 1.…
data_ecom	541909 obs. of 8 variables
fit	List of 7
km	List of 9
top_custom…	20 obs. of 4 variables

Values

Files | Plots | Packages | Help | Viewer | Presentation

Zoom | Export



File Edit Code View Plots Session Build Debug Profile Tools Help

Ecommerce CLustering Project.R | customerData_Hier | customerData_km | customerData_sc | top_customers

```
34 # We can still see the NA values in the data, lets see its weight
35 na_count <-colSums(is.na(data_ecom))
36 na_count
37
38 # We have 135,080 missing Customer ID whcih is unique identifier.
39 # At this point, it's better to omit the data as we can't tag this transaction to right specific c
40 na.omit(data_ecom)
41
42 #Summary statistics of the dataset
43 summary(data_ecom)
44
45 #TotalSales & TotalUnitSales purchase by each customer
46 customerData <- data_ecom %>%
47   select(CustomerID, Country, Quantity, UnitPrice) %>%
48   group_by(CustomerID, Country) %>%
49   summarise(TotalRevenue = sum(Quantity * UnitPrice),
50             TotalItemsSold = sum(Quantity),
51             .groups = 'drop')
52 View(customerData)
53
54 ##TOPHIghly Valued Customers based on Revenue Generated
55 top_customers <- customerData %>%
56   arrange(desc(TotalRevenue)) %>%
57   slice(1:20)
58 View(top_customers)
59
60
61 #We remove CustomerID and Country Columns as these features are of Less Importance/ can result mul
62 customerData <- customerData[-c(1:2)]
63
64 #Finding and Sorting Outliers
65 boxplot(customerData)
66
67
```

132:1 (Top Level) R Script

Console

Environment | History | Connections | Tutorial

Import 715 MiB List

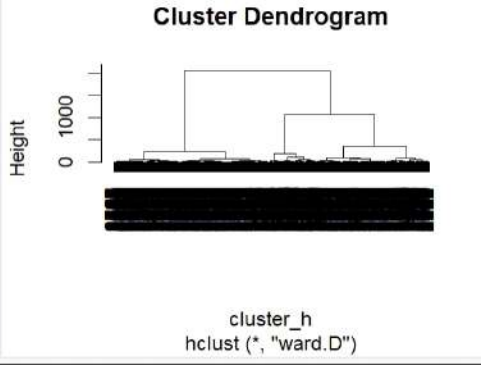
R Global Environment

Data	
customerDa…	3730 obs. of 2 variables
customerDa…	num [1:3730, 1:3] -1.091 1.…
customerDa…	num [1:3730, 1:3] -1.091 1.…
customerDa…	num [1:3730, 1:2] -1.091 1.…
data_ecom	541909 obs. of 8 variables
fit	List of 7
km	List of 9
top_custom…	20 obs. of 4 variables

Values

Files | Plots | Packages | Help | Viewer | Presentation

Zoom Export



File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Ecommerce CLustering Project.R | customerData_Hier | customerData_km | customerData_sc | top_customers | custo

Source on Save | Run | Source

```
66
67  #Eliminating the outliers
68  library(dplyr)
69
70  customerData <- customerData %>%
71    filter(between(TotalRevenue, quantile(TotalRevenue, 0.25) - 1.5 * IQR(TotalRevenue),
72                   quantile(TotalRevenue, 0.75) + 1.5 * IQR(TotalRevenue))) %>%
73    filter(between(TotalItemsSold, quantile(TotalItemsSold, 0.25) - 1.5 * IQR(TotalItemsSold),
74                   quantile(TotalItemsSold, 0.75) + 1.5 * IQR(TotalItemsSold)))
75
76  #Lets Normalize the data as Sales Amount and Sales Units value has huge diff to compare without sc
77  customerData_sc <-scale(customerData)
78  View(customerData_sc)
79
80  ##KMeans Clustering
81  set.seed(42)
82
83  #Let's group the data into 5 cluster with k-means
84  km <-kmeans(customerData_sc,centers=5)
85  km
86  km$size
87
88  #Determining optimal clusters with Elbow Mwthod
89  set.seed(42)
90
91  # Compute total within-cluster sum of squares for k=1 to 10
92  wss_values <- map_dbl(1:10, ~kmeans(customerData_sc, ., nstart = 10)$tot.withinss)
93
94  # Plot WSS values vs. number of clusters
95  plot(1:10, wss_values, type = "b", pch = 19, frame = FALSE,
96       xlab = "Number of clusters (k)", ylab = "Total within-cluster sum of squares (WSS)")
97
98  ###Looking at the plot we can say as per our K means graph 3 is the optimal number of clusters.
99
```

132:1 (Top Level) | R Script

Console

Environment | History | Connections | Tutorial

Import | 715 MiB | List

R | Global Environment

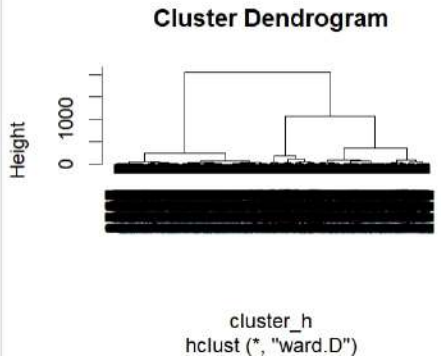
Data

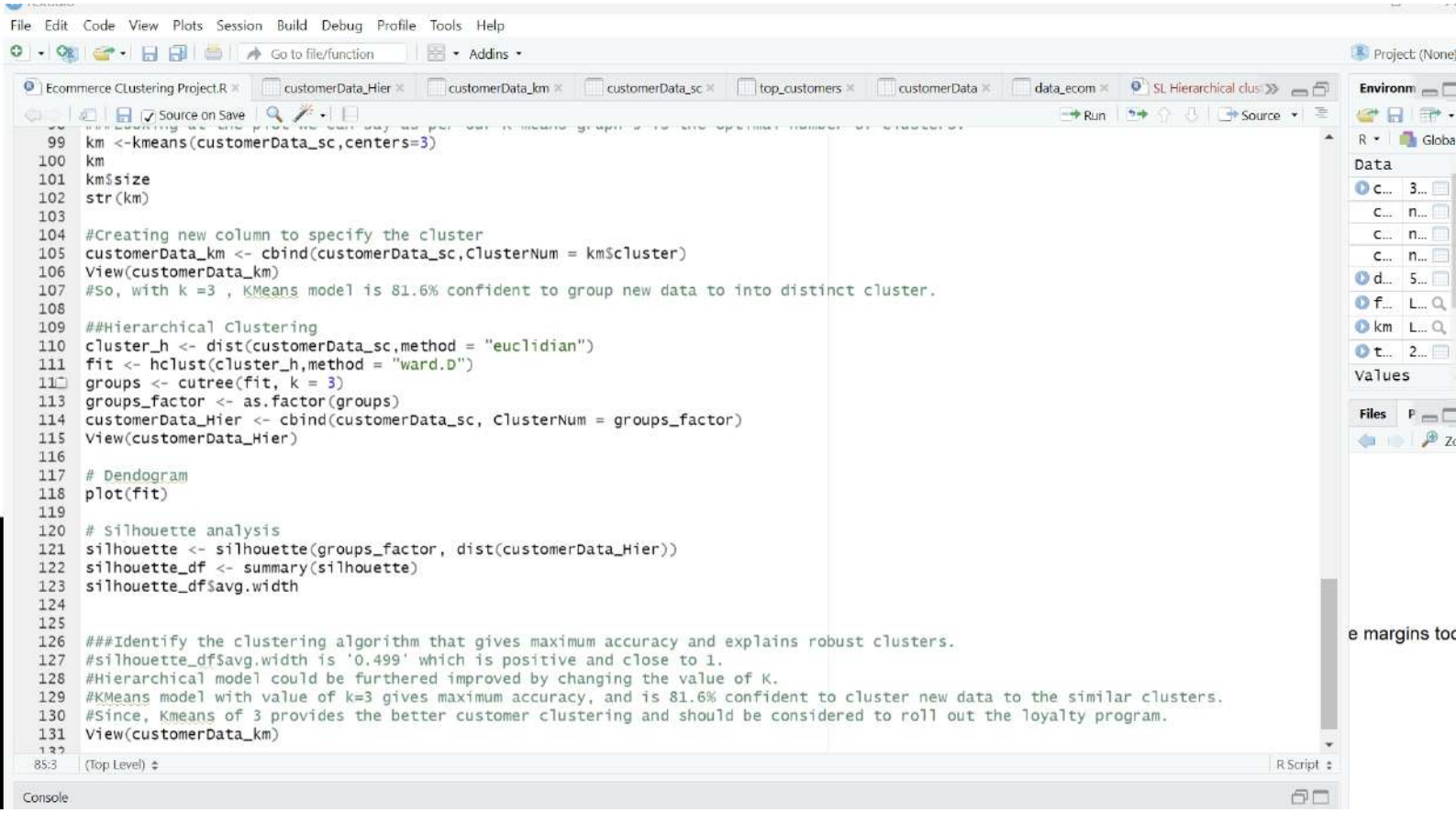
customerDa…	3730 obs. of 2 variables
customerDa…	num [1:3730, 1:3] -1.091 1.…
customerDa…	num [1:3730, 1:3] -1.091 1.…
customerDa…	num [1:3730, 1:2] -1.091 1.…
data_ecom	541909 obs. of 8 variables
fit	List of 7
km	List of 9
top_custom…	20 obs. of 4 variables

Values

Files | Plots | Packages | Help | Viewer | Presentation

Zoom | Export




File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Ecommerce CLustering Project.R
customerData_Hier
customerData_km
customerData_sc
top_customers
customerData
data_ecom
SL Hierarchical clust
Source on Save
Run
Source
km <-kmeans(customerData_sc,centers=3)
km
km$size
str(km)
#Creating new column to specify the cluster
customerData_km <- cbind(customerData_sc,ClusterNum = km$cluster)
View(customerData_km)
#So, with k =3 , KMeans model is 81.6% confident to group new data to into distinct cluster.
##Hierarchical Clustering
cluster_h <- dist(customerData_sc,method = "euclidian")
fit <- hclust(cluster_h,method = "ward.D")
groups <- cutree(fit, k = 3)
groups_factor <- as.factor(groups)
customerData_Hier <- cbind(customerData_sc, ClusterNum = groups_factor)
View(customerData_Hier)
# Dendogram
plot(fit)
# Silhouette analysis
silhouette <- silhouette(groups_factor, dist(customerData_Hier))
silhouette_df <- summary(silhouette)
silhouette_df$avg.width
###Identify the clustering algorithm that gives maximum accuracy and explains robust clusters.
#silhouette_df$avg.width is '0.499' which is positive and close to 1.
#Hierarchical model could be furthered improved by changing the value of K.
#KMeans model with value of k=3 gives maximum accuracy, and is 81.6% confident to cluster new data to the similar clusters.
#Since, Kmeans of 3 provides the better customer clustering and should be considered to roll out the loyalty program.
View(customerData_km)
85:3
(Top Level)
R Script
Console
Environm
Data
Values
Files
e margins too