E_Commerce Unsupervised Learning

Sharon Maswai

3/9/2020

Definition of Question

Context

An entrepreneur from Kenya has created an online cause which she would like to advertise on her blog. Based on the data she collected from running the ads on her blog, ehe would like a data scientist to help her determine which individuals click on the ads.

Objective

To do an analysis to establish individuals that click on the ads most.

Data Appropriateness

This dataset is suitable for the analysis.

```
#Importing libraries and loading them
install.packages("tidyverse")
```

```
## Installing package into '/home/sharon-maswai/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

```
install.packages("ggplot2")
```

```
## Installing package into '/home/sharon-maswai/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------ tidyverse 1.3.0 --

## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts --------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(ggplot2)
```

Load Dataset

```r
ecommerce <- read.csv("~/Downloads/ecommerce.csv")
head(ecommerce)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       0             0                      0
## 2              0                       0             0                      0
## 3              0                      -1             0                     -1
## 4              0                       0             0                      0
## 5              0                       0             0                      0
## 6              0                       0             0                      0
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1              1                0.000000  0.20000000 0.2000000          0
## 2              2               64.000000  0.00000000 0.1000000          0
## 3              1               -1.000000  0.20000000 0.2000000          0
## 4              2                2.666667  0.05000000 0.1400000          0
## 5             10              627.500000  0.02000000 0.0500000          0
## 6             19              154.216667  0.01578947 0.0245614          0
##   SpecialDay Month OperatingSystems Browser Region TrafficType
## 1          0   Feb                1       1      1           1
## 2          0   Feb                2       2      1           2
## 3          0   Feb                4       1      9           3
## 4          0   Feb                3       2      2           4
## 5          0   Feb                3       3      1           4
## 6          0   Feb                2       2      1           3
##         VisitorType Weekend Revenue
## 1 Returning_Visitor   FALSE   FALSE
## 2 Returning_Visitor   FALSE   FALSE
## 3 Returning_Visitor   FALSE   FALSE
## 4 Returning_Visitor   FALSE   FALSE
## 5 Returning_Visitor    TRUE   FALSE
## 6 Returning_Visitor   FALSE   FALSE
```

Data Understanding

Dataset summary

```r
summary(ecommerce)
```

```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :  -1.00         Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.000
##  Median : 1.000   Median :   8.00         Median : 0.000
##  Mean   : 2.318   Mean   :  80.91         Mean   : 0.504
##  3rd Qu.: 4.000   3rd Qu.:  93.50         3rd Qu.: 0.000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.000
```

```
##  NA's   :14       NA's   :14              NA's   :14
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :  -1.00        Min.   :  0.00   Min.   :   -1.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  185.0
##  Median :   0.00        Median : 18.00   Median :  599.8
##  Mean   :  34.51        Mean   : 31.76   Mean   : 1196.0
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1466.5
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
##  NA's   :14             NA's   :14       NA's   :14
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   1st Qu.:0.00000
##  Median :0.003119   Median :0.02512   Median :  0.000   Median :0.00000
##  Mean   :0.022152   Mean   :0.04300   Mean   :  5.889   Mean   :0.06143
##  3rd Qu.:0.016684   3rd Qu.:0.05000   3rd Qu.:  0.000   3rd Qu.:0.00000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000
##  NA's   :14         NA's   :14
##      Month      OperatingSystems    Browser           Region
##  May    :3364   Min.   :1.000    Min.   : 1.000   Min.   :1.000
##  Nov    :2998   1st Qu.:2.000    1st Qu.: 2.000   1st Qu.:1.000
##  Mar    :1907   Median :2.000    Median : 2.000   Median :3.000
##  Dec    :1727   Mean   :2.124    Mean   : 2.357   Mean   :3.147
##  Oct    : 549   3rd Qu.:3.000    3rd Qu.: 2.000   3rd Qu.:4.000
##  Sep    : 448   Max.   :8.000    Max.   :13.000   Max.   :9.000
##  (Other):1337
##   TrafficType               VisitorType     Weekend         Revenue
##  Min.   : 1.00   New_Visitor      : 1694   Mode :logical   Mode :logical
##  1st Qu.: 2.00   Other            :   85   FALSE:9462      FALSE:10422
##  Median : 2.00   Returning_Visitor:10551   TRUE :2868      TRUE :1908
##  Mean   : 4.07
##  3rd Qu.: 4.00
##  Max.   :20.00
##
```

Checking column names

```
names(ecommerce)
```

```
##  [1] "Administrative"          "Administrative_Duration"
##  [3] "Informational"           "Informational_Duration"
##  [5] "ProductRelated"          "ProductRelated_Duration"
##  [7] "BounceRates"             "ExitRates"
##  [9] "PageValues"              "SpecialDay"
## [11] "Month"                   "OperatingSystems"
## [13] "Browser"                 "Region"
## [15] "TrafficType"             "VisitorType"
## [17] "Weekend"                 "Revenue"
```

Checking datatypes

```
#obtaining the datatypes
sapply(data, class)
```

```
##       ...      list   package   lib.loc   verbose     envir overwrite
##    "name"    "call"    "NULL"    "NULL"    "call"    "name" "logical"       "{"
```

Checking the shape of the dataset

```
dim(ecommerce)
```

```
## [1] 12330    18
```

Checking for null values and duplicates

```
colSums(is.na(ecommerce))
```

```
##          Administrative Administrative_Duration           Informational
##                      14                      14                      14
##  Informational_Duration          ProductRelated ProductRelated_Duration
##                      14                      14                      14
##             BounceRates               ExitRates              PageValues
##                      14                      14                       0
##              SpecialDay                   Month        OperatingSystems
##                       0                       0                       0
##                 Browser                  Region             TrafficType
##                       0                       0                       0
##             VisitorType                 Weekend                 Revenue
##                       0                       0                       0
```

```
# Dropping the missing values
ecommerce= na.omit(ecommerce)
```

Checking for and dropping duplicates

```
# Checking for dupliccate values
dup_rows <- ecommerce[duplicated(ecommerce ),]
dim(dup_rows)
```

```
## [1] 117  18
```

```
ecommerce = ecommerce[!duplicated(ecommerce), ]
```

Conclusion

- There are no duplicates in the dataset.

Univariate Analysis

Checking for outliers

Modeling

```r
data= ecommerce[, c(1,2,3,4,5,6,7,8,9,10, 12, 13, 14,1)]
```

```r
data = cor(data)
data
```

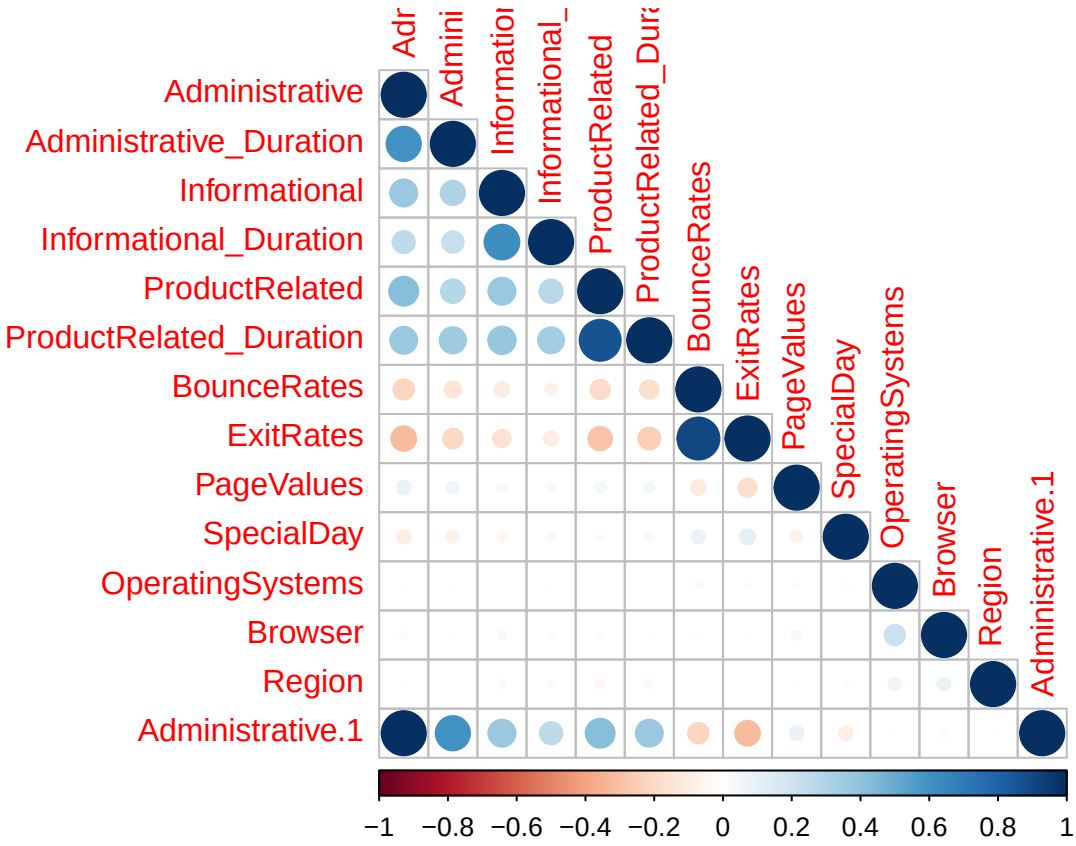
```
##                          Administrative Administrative_Duration Informational
## Administrative             1.000000000             0.600409653    0.37528761
## Administrative_Duration    0.600409653             1.000000000    0.30143630
## Informational              0.375287611             0.301436296    1.00000000
## Informational_Duration     0.254786021             0.237189860    0.61867795
## ProductRelated             0.428191515             0.286783914    0.37260472
## ProductRelated_Duration    0.371027224             0.353513793    0.38608372
## BounceRates               -0.213666635            -0.137333397   -0.10950530
## ExitRates                 -0.311274132            -0.202024452   -0.15956681
## PageValues                 0.096920968             0.066168365    0.04739015
## SpecialDay                -0.097072098            -0.074736885   -0.04937677
## OperatingSystems          -0.006697922            -0.007610715   -0.00962587
## Browser                   -0.025763658            -0.015833675   -0.03876681
## Region                    -0.007262053            -0.006723711   -0.03047732
## Administrative.1           1.000000000             0.600409653    0.37528761
##                          Informational_Duration ProductRelated
## Administrative                      0.254786021    0.428191515
## Administrative_Duration             0.237189860    0.286783914
## Informational                       0.618677947    0.372604721
## Informational_Duration              1.000000000    0.279061948
## ProductRelated                      0.279061948    1.000000000
## ProductRelated_Duration             0.346580691    0.860308186
## BounceRates                        -0.070159472   -0.193515772
## ExitRates                          -0.102932678   -0.286163211
## PageValues                          0.030064160    0.054115494
## SpecialDay                         -0.031293040   -0.025930622
## OperatingSystems                   -0.009749983    0.004090351
## Browser                            -0.019609349   -0.013706213
## Region                             -0.027920098   -0.040106501
## Administrative.1                    0.254786021    0.428191515
##                          ProductRelated_Duration  BounceRates    ExitRates
## Administrative                      0.371027224 -0.213666635 -0.311274132
## Administrative_Duration             0.353513793 -0.137333397 -0.202024452
## Informational                       0.386083717 -0.109505298 -0.159566815
## Informational_Duration              0.346580691 -0.070159472 -0.102932678
## ProductRelated                      0.860308186 -0.193515772 -0.286163211
## ProductRelated_Duration             1.000000000 -0.174375499 -0.245334012
## BounceRates                        -0.174375499  1.000000000  0.903358192
## ExitRates                          -0.245334012  0.903358192  1.000000000
## PageValues                          0.050840624 -0.115991977 -0.173571542
## SpecialDay                         -0.038210652  0.087839995  0.116783762
## OperatingSystems                    0.002775788  0.026839839  0.016482012
## Browser                            -0.007838332 -0.016018380 -0.003565541
## Region                             -0.034862498  0.001432015 -0.001837556
## Administrative.1                    0.371027224 -0.213666635 -0.311274132
##                           PageValues   SpecialDay OperatingSystems      Browser
## Administrative            0.09692097 -0.097072098     -0.006697922 -0.025763658
## Administrative_Duration   0.06616837 -0.074736885     -0.007610715 -0.015833675
```

```
## Informational             0.04739015 -0.049376774     -0.009625870 -0.038766808
## Informational_Duration    0.03006416 -0.031293040     -0.009749983 -0.019609349
## ProductRelated            0.05411549 -0.025930622      0.004090351 -0.013706213
## ProductRelated_Duration   0.05084062 -0.038210652      0.002775788 -0.007838332
## BounceRates              -0.11599198  0.087839995      0.026839839 -0.016018380
## ExitRates                -0.17357154  0.116783762      0.016482012 -0.003565541
## PageValues                1.00000000 -0.064532709      0.018583782  0.045845065
## SpecialDay               -0.06453271  1.000000000      0.012757766  0.003465984
## OperatingSystems          0.01858378  0.012757766      1.000000000  0.212244823
## Browser                   0.04584506  0.003465984      0.212244823  1.000000000
## Region                    0.01059087 -0.016452464      0.071953240  0.091889464
## Administrative.1          0.09692097 -0.097072098     -0.006697922 -0.025763658
##                                Region Administrative.1
## Administrative           -0.007262053      1.000000000
## Administrative_Duration  -0.006723711      0.600409653
## Informational            -0.030477323      0.375287611
## Informational_Duration   -0.027920098      0.254786021
## ProductRelated           -0.040106501      0.428191515
## ProductRelated_Duration  -0.034862498      0.371027224
## BounceRates               0.001432015     -0.213666635
## ExitRates                -0.001837556     -0.311274132
## PageValues                0.010590868      0.096920968
## SpecialDay               -0.016452464     -0.097072098
## OperatingSystems          0.071953240     -0.006697922
## Browser                   0.091889464     -0.025763658
## Region                    1.000000000     -0.007262053
## Administrative.1         -0.007262053      1.000000000
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot(data, type = 'lower')
```



K Means Algorythmn

```
feat = ecommerce[, c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17)]

# labels
y = ecommerce[, 18]
```

Normalize the data

```
normalize <- function(x){
  return ((x-min(x)) / (max(x)-min(x)))
}
columns = c('Administrative','Administrative_Duration','Informational','Informational_Duration','Produc
for (i in columns){
    feat[, i] = normalize(feat[, i])
}
```

#####Coverting categorical data to numerical

```
num_data = c('Month','VisitorType', 'Weekend')
for (i in num_data){
    feat[,i] = as.numeric(feat[,i])
}
```

Fitting the data

```
k_model <- kmeans(feat,2)
```

```
#Checking cluster number
k_model$cluster_obs = as.numeric(y)

k_model$size
```

```
## [1] 10116  2083
```

```
table(k_model$cluster,k_model$cluster_obs )
```

```
##
##        0    1
##   1 8541 1575
##   2 1750  333
```

Accuracy

```
mean(k_model$cluster_obs == k_model$cluster)
```

```
## [1] 0.1291089
```

Hierachical Clustering

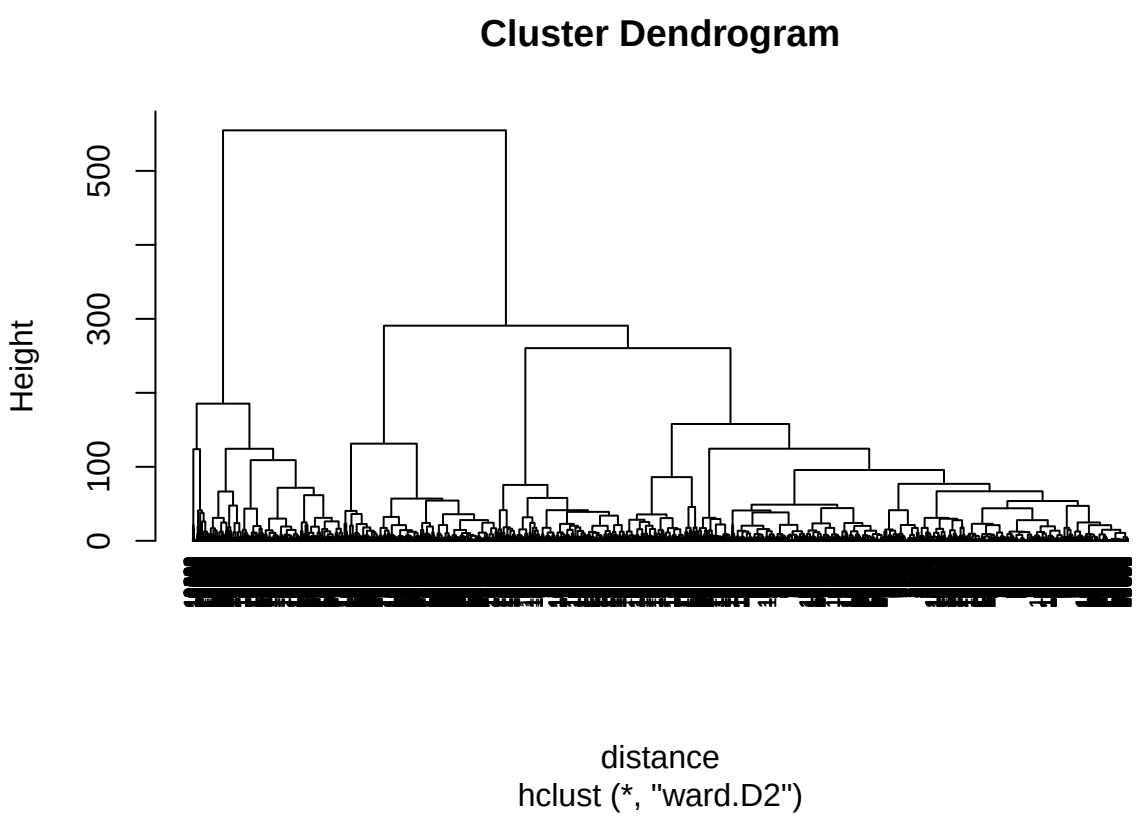
```
#obtaining the euclidean distance
distance <- dist(feat, method = "euclidean")
```

Fitting

```
# fitting the model
model <- hclust(distance, method = "ward.D2" )
```

Visualiztion

```
# plotting
plot(model, cex = 0.6, hang = -1)
```


Cluster Dendrogram
Height
0
100
300
500
distance
hclust (*, "ward.D2")

Accuracy