week 13 part 2

Christopher Kimani

2022-08-11

# Loading Libraries necessary  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.7 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(magrittr)

##   
## Attaching package: 'magrittr'  
##   
## The following object is masked from 'package:purrr':  
##   
## set\_names  
##   
## The following object is masked from 'package:tidyr':  
##   
## extract

library(warn = -1)  
library(RColorBrewer)  
library(ggplot2)  
library(lattice)  
library(corrplot)

## corrplot 0.92 loaded

library(psych)

##   
## Attaching package: 'psych'  
##   
## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(caret)

##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

# 1. Problem Definition

Kira Plastinina is a Russian brand that is sold through a defunct chain of retail stores in Russia, Ukraine, Kazakhstan, Belarus, China, Philippines, and Armenia. The brand’s Sales and Marketing team would like to understand their customer’s behavior from data that they have collected over the past year. More specifically, they would like to learn the characteristics of customer groups.

-Perform clustering stating insights drawn from your analysis and visualizations. -Upon implementation, provide comparisons between the approaches learned this week i.e. K-Means clustering vs Hierarchical clustering highlighting the strengths and limitations of each approach in the context of your analysis.

# 2. Data Sourcing

* The data is sourced from this link (<http://bit.ly/EcommerceCustomersDataset>)

## 2.1 Loading the dataset

#Loading the data  
data <-read.csv("http://bit.ly/EcommerceCustomersDataset", header = TRUE)  
#The first 6 rows   
head(data)

## 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

# 3. Check the Data

#Checking the structure of the data   
str(data)

## 'data.frame': 12330 obs. of 18 variables:  
## $ Administrative : int 0 0 0 0 0 0 0 1 0 0 ...  
## $ Administrative\_Duration: num 0 0 -1 0 0 0 -1 -1 0 0 ...  
## $ Informational : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Informational\_Duration : num 0 0 -1 0 0 0 -1 -1 0 0 ...  
## $ ProductRelated : int 1 2 1 2 10 19 1 1 2 3 ...  
## $ ProductRelated\_Duration: num 0 64 -1 2.67 627.5 ...  
## $ BounceRates : num 0.2 0 0.2 0.05 0.02 ...  
## $ ExitRates : num 0.2 0.1 0.2 0.14 0.05 ...  
## $ PageValues : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ SpecialDay : num 0 0 0 0 0 0 0.4 0 0.8 0.4 ...  
## $ Month : chr "Feb" "Feb" "Feb" "Feb" ...  
## $ OperatingSystems : int 1 2 4 3 3 2 2 1 2 2 ...  
## $ Browser : int 1 2 1 2 3 2 4 2 2 4 ...  
## $ Region : int 1 1 9 2 1 1 3 1 2 1 ...  
## $ TrafficType : int 1 2 3 4 4 3 3 5 3 2 ...  
## $ VisitorType : chr "Returning\_Visitor" "Returning\_Visitor" "Returning\_Visitor" "Returning\_Visitor" ...  
## $ Weekend : logi FALSE FALSE FALSE FALSE TRUE FALSE ...  
## $ Revenue : logi FALSE FALSE FALSE FALSE FALSE FALSE ...

# Data is in integers, numericals and characters.

# A statistical summary of the data   
summary(data)

## 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   
## Length:12330 Min. :1.000 Min. : 1.000 Min. :1.000   
## Class :character 1st Qu.:2.000 1st Qu.: 2.000 1st Qu.:1.000   
## Mode :character Median :2.000 Median : 2.000 Median :3.000   
## Mean :2.124 Mean : 2.357 Mean :3.147   
## 3rd Qu.:3.000 3rd Qu.: 2.000 3rd Qu.:4.000   
## Max. :8.000 Max. :13.000 Max. :9.000   
##   
## TrafficType VisitorType Weekend Revenue   
## Min. : 1.00 Length:12330 Mode :logical Mode :logical   
## 1st Qu.: 2.00 Class :character FALSE:9462 FALSE:10422   
## Median : 2.00 Mode :character TRUE :2868 TRUE :1908   
## Mean : 4.07   
## 3rd Qu.: 4.00   
## Max. :20.00   
##

# dimensions of the data  
dim(data)

## [1] 12330 18

# There are 18 columns and 12330 rows

# 4. Data Cleaning

# conversion from just a raw dataset to a dataframe  
data = as.data.frame(data)  
# Cleaning column names, by making them uniform  
colnames(data) = tolower(colnames(data))

## 4.1.1 Duplicated values

#Checking for duplicated rows  
duplicated\_rows <- data[duplicated(data),]  
dim(duplicated\_rows)

## [1] 119 18

# There are 10 duplicated rows

#Dropping duplicated rows  
data <- unique(data)  
dim(data)

## [1] 12211 18

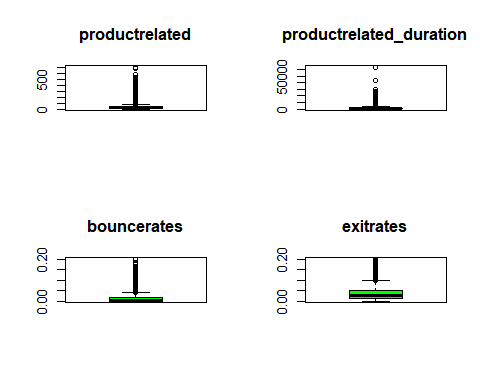
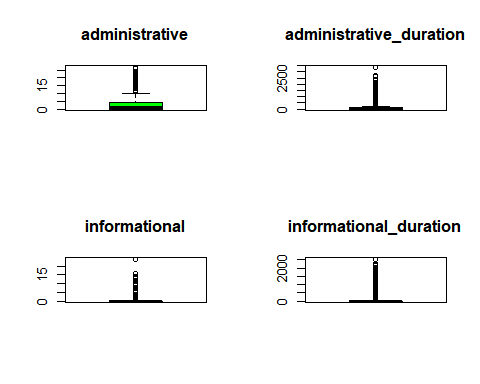
## 4.1.2 Missing Values

# Filling the missing values with the mean   
data = data %>%  
 mutate(administrative =replace(administrative,is.na(administrative),mean(administrative,na.rm=TRUE)))%>%  
 mutate(administrative\_duration =replace(administrative\_duration,is.na(administrative\_duration),mean(administrative\_duration,na.rm=TRUE)))%>%  
 mutate(informational = replace(informational, is.na(informational), mean(informational, na.rm = TRUE)))%>%  
 mutate(informational\_duration =replace(informational\_duration,is.na(informational\_duration),mean(informational\_duration,na.rm=TRUE)))%>%  
 mutate(productrelated =replace(productrelated,is.na(productrelated),mean(productrelated,na.rm=TRUE)))%>%  
 mutate(productrelated\_duration = replace(productrelated\_duration, is.na(productrelated\_duration), mean(productrelated\_duration, na.rm = TRUE)))%>%  
 mutate(bouncerates =replace(bouncerates, is.na(bouncerates),mean(bouncerates,na.rm=TRUE)))%>%  
 mutate(exitrates = replace(exitrates, is.na(exitrates), mean(exitrates, na.rm = TRUE)))

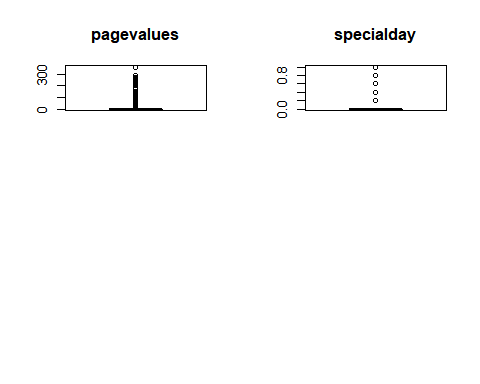
# Choosing the categorical columns and changing them to factors and checking the structure of the data   
cat\_col = c('month', 'operatingsystems', 'browser', 'region', 'traffictype', 'visitortype')  
data[,cat\_col] %<>% lapply(function(x) as.factor(as.character(x)))  
str(df)

## function (x, df1, df2, ncp, log = FALSE)

#Checking for outliers   
# Creating separate boxplots for each attribute  
par(mfrow=c(2,2))  
for(i in 1:10) {  
 boxplot(data[,i], main=names(data)[i], col = "green")}



# I will not be dropping outliers as they are important in the data.

 # 5. Exploratory Data Analysis

## 5.1 Univariate Analysis

### 5.1.1 Measures of central tendancy(Mean, Median, Mode), dispersion (Min, Max, Range, Quartiles, Variance, Standard deviation), others(Skewness, Kurtosis)

#Selecting only the numerical variables.  
nums <- subset(data, select = -c(specialday, month, operatingsystems,browser, region, traffictype, visitortype,weekend,revenue))  
head(nums)

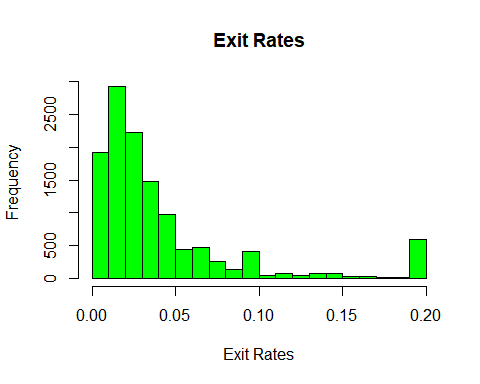
## administrative administrative\_duration informational informational\_duration  
## 1 0 0 0 0  
## 2 0 0 0 0  
## 3 0 -1 0 -1  
## 4 0 0 0 0  
## 5 0 0 0 0  
## 6 0 0 0 0  
## productrelated productrelated\_duration bouncerates exitrates pagevalues  
## 1 1 0.000000 0.20000000 0.2000000 0  
## 2 2 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

# further description of the data   
describe(nums)

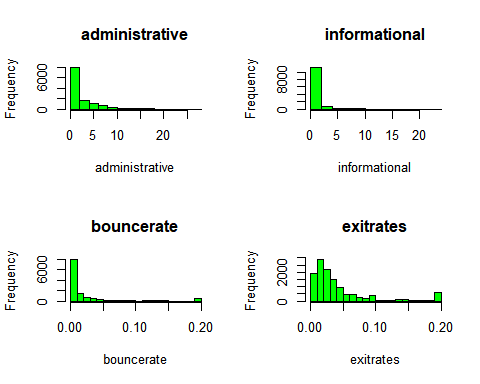
## vars n mean sd median trimmed mad min  
## administrative 1 12211 2.34 3.33 1.00 1.66 1.48 0  
## administrative\_duration 2 12211 81.68 177.44 9.00 42.89 13.34 -1  
## informational 3 12211 0.51 1.28 0.00 0.18 0.00 0  
## informational\_duration 4 12211 34.84 141.39 0.00 3.75 0.00 -1  
## productrelated 5 12211 32.06 44.58 18.00 23.06 19.27 0  
## productrelated\_duration 6 12211 1207.51 1918.98 611.00 832.56 747.53 -1  
## bouncerates 7 12211 0.02 0.05 0.00 0.01 0.00 0  
## exitrates 8 12211 0.04 0.05 0.03 0.03 0.02 0  
## pagevalues 9 12211 5.95 18.65 0.00 1.33 0.00 0  
## max range skew kurtosis se  
## administrative 27.00 27.00 1.95 4.64 0.03  
## administrative\_duration 3398.75 3399.75 5.59 50.14 1.61  
## informational 24.00 24.00 4.01 26.67 0.01  
## informational\_duration 2549.38 2550.38 7.54 75.53 1.28  
## productrelated 705.00 705.00 4.33 31.08 0.40  
## productrelated\_duration 63973.52 63974.52 7.25 136.71 17.37  
## bouncerates 0.20 0.20 3.15 9.27 0.00  
## exitrates 0.20 0.20 2.23 4.63 0.00  
## pagevalues 361.76 361.76 6.35 64.99 0.17

### 5.2 Visualizations

# Histogram of Exit Rates  
hist(data$exitrates,  
 main = "Exit Rates",  
 xlab = "Exit Rates",  
 col = "Green")

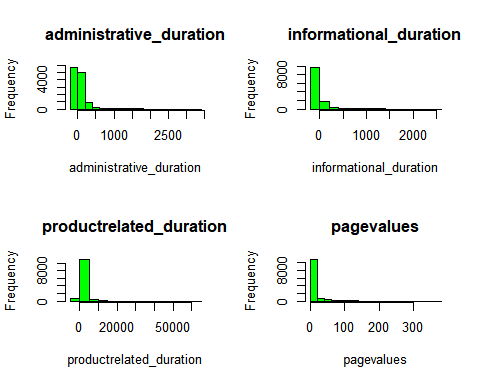
 Another distribution that is right skewed

#Plotting histograms to show distribution of variables   
par(mfrow = c(2, 2))  
hist(nums$administrative,col = "green", main = 'administrative', xlab = 'administrative')  
hist(nums$informational, col = "green", main = 'informational', xlab = 'informational')  
hist(nums$bouncerates,col = "green", main = 'bouncerate', xlab = 'bouncerate')  
hist(nums$exitrates,col = "green", main = 'exitrates', xlab = 'exitrates')



# Skewness is to the right.

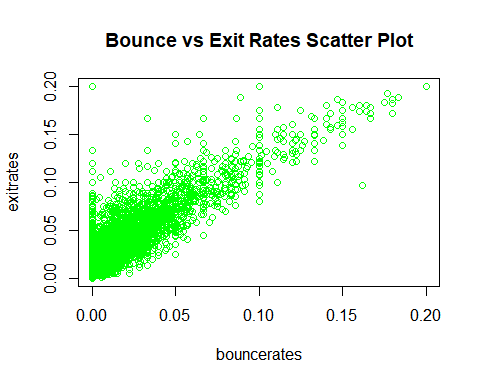
par(mfrow = c(2, 2))  
hist(nums$administrative\_duration, ,col = "green", main = 'administrative\_duration', xlab = 'administrative\_duration')  
hist(nums$informational\_duration,col = "green",main = 'informational\_duration', xlab = 'informational\_duration')  
hist(nums$productrelated\_duration,col = "green",main = 'productrelated\_duration', xlab = 'productrelated\_duration')  
hist(nums$pagevalues,col = "green",main = 'pagevalues', xlab = 'pagevalues')



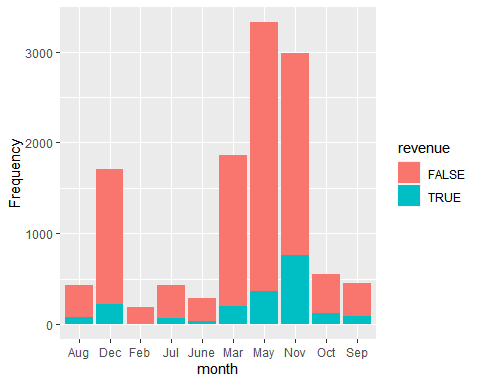
#People also spend a lot of time checking on the administration page  
 #People spend the least of time checking out the information related page

## 5.2 Bivariate Analysis

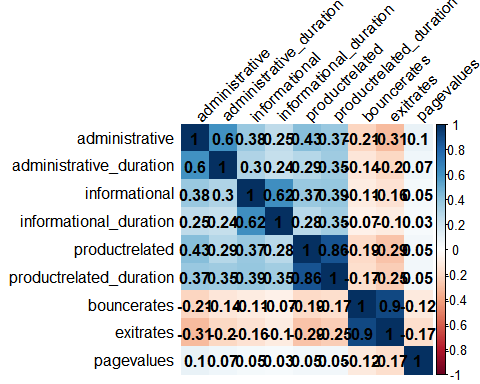
# bounce rate vs exit rate  
plot(exitrates ~ bouncerates, dat = data,   
 col = "green",  
 main = "Bounce vs Exit Rates Scatter Plot")

 -There exists strong positive correlation between Exit rates and Bounce rates.

#Revenue generation per month  
data %>%   
 ggplot() +  
 aes(x = month, revenue = ..count../nrow(data), fill = revenue) +  
 geom\_bar() +  
 ylab("Frequency")



#Get the correlation matrix  
res = cor(nums)  
#Plotting a correlation plot  
corrplot(res, method="color",addCoef.col = "black",   
 tl.col="black", tl.srt=45)

 -there exists a strong positive relationship between a page and its respective duration for example Product Related page and Product Related Duration

# 6. Implement the Solution

## 6.1 Feature Engineering

# # One hot encoding.  
dmy = dummyVars(" ~ .", data = data)  
data2 = data.frame(predict(dmy, newdata = data))

# Confirming data types of each attribute  
sapply(data2, class)

## administrative administrative\_duration   
## "numeric" "numeric"   
## informational informational\_duration   
## "numeric" "numeric"   
## productrelated productrelated\_duration   
## "numeric" "numeric"   
## bouncerates exitrates   
## "numeric" "numeric"   
## pagevalues specialday   
## "numeric" "numeric"   
## month.Aug month.Dec   
## "numeric" "numeric"   
## month.Feb month.Jul   
## "numeric" "numeric"   
## month.June month.Mar   
## "numeric" "numeric"   
## month.May month.Nov   
## "numeric" "numeric"   
## month.Oct month.Sep   
## "numeric" "numeric"   
## operatingsystems.1 operatingsystems.2   
## "numeric" "numeric"   
## operatingsystems.3 operatingsystems.4   
## "numeric" "numeric"   
## operatingsystems.5 operatingsystems.6   
## "numeric" "numeric"   
## operatingsystems.7 operatingsystems.8   
## "numeric" "numeric"   
## browser.1 browser.10   
## "numeric" "numeric"   
## browser.11 browser.12   
## "numeric" "numeric"   
## browser.13 browser.2   
## "numeric" "numeric"   
## browser.3 browser.4   
## "numeric" "numeric"   
## browser.5 browser.6   
## "numeric" "numeric"   
## browser.7 browser.8   
## "numeric" "numeric"   
## browser.9 region.1   
## "numeric" "numeric"   
## region.2 region.3   
## "numeric" "numeric"   
## region.4 region.5   
## "numeric" "numeric"   
## region.6 region.7   
## "numeric" "numeric"   
## region.8 region.9   
## "numeric" "numeric"   
## traffictype.1 traffictype.10   
## "numeric" "numeric"   
## traffictype.11 traffictype.12   
## "numeric" "numeric"   
## traffictype.13 traffictype.14   
## "numeric" "numeric"   
## traffictype.15 traffictype.16   
## "numeric" "numeric"   
## traffictype.17 traffictype.18   
## "numeric" "numeric"   
## traffictype.19 traffictype.2   
## "numeric" "numeric"   
## traffictype.20 traffictype.3   
## "numeric" "numeric"   
## traffictype.4 traffictype.5   
## "numeric" "numeric"   
## traffictype.6 traffictype.7   
## "numeric" "numeric"   
## traffictype.8 traffictype.9   
## "numeric" "numeric"   
## visitortype.New\_Visitor visitortype.Other   
## "numeric" "numeric"   
## visitortype.Returning\_Visitor weekendFALSE   
## "numeric" "numeric"   
## weekendTRUE revenueFALSE   
## "numeric" "numeric"   
## revenueTRUE   
## "numeric"

# Dropping revenue   
data3 <- data2[, -c(30:31)]  
data.rev<- data[, "revenue"]  
data4 <- data2[, -c(30,31)]

#scaling the data   
data4\_scaled <- scale(data4)

# Normalizing the the original data  
data\_normal <- as.data.frame(apply(data4, 2, function(x) (x - min(x))/(max(x)-min(x))))  
summary(data\_normal)

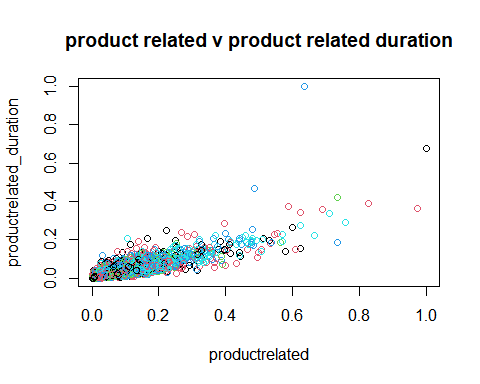
## administrative administrative\_duration informational   
## Min. :0.00000 Min. :0.0000000 Min. :0.0000   
## 1st Qu.:0.00000 1st Qu.:0.0002941 1st Qu.:0.0000   
## Median :0.03704 Median :0.0029414 Median :0.0000   
## Mean :0.08667 Mean :0.0243201 Mean :0.0212   
## 3rd Qu.:0.14815 3rd Qu.:0.0281197 3rd Qu.:0.0000   
## Max. :1.00000 Max. :1.0000000 Max. :1.0000   
## informational\_duration productrelated productrelated\_duration  
## Min. :0.0000000 Min. :0.00000 Min. :0.000000   
## 1st Qu.:0.0003921 1st Qu.:0.01135 1st Qu.:0.003048   
## Median :0.0003921 Median :0.02553 Median :0.009566   
## Mean :0.0140518 Mean :0.04547 Mean :0.018891   
## 3rd Qu.:0.0003921 3rd Qu.:0.05390 3rd Qu.:0.023094   
## Max. :1.0000000 Max. :1.00000 Max. :1.000000   
## bouncerates exitrates pagevalues specialday   
## Min. :0.00000 Min. :0.00000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.00000 1st Qu.:0.07128 1st Qu.:0.00000 1st Qu.:0.00000   
## Median :0.01471 Median :0.12500 Median :0.00000 Median :0.00000   
## Mean :0.10223 Mean :0.20748 Mean :0.01644 Mean :0.06191   
## 3rd Qu.:0.08333 3rd Qu.:0.24233 3rd Qu.:0.00000 3rd Qu.:0.00000   
## Max. :1.00000 Max. :1.00000 Max. :1.00000 Max. :1.00000   
## month.Aug month.Dec month.Feb month.Jul   
## Min. :0.00000 Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.00000   
## Median :0.00000 Median :0.0000 Median :0.0000 Median :0.00000   
## Mean :0.03546 Mean :0.1397 Mean :0.0149 Mean :0.03538   
## 3rd Qu.:0.00000 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.:0.00000   
## Max. :1.00000 Max. :1.0000 Max. :1.0000 Max. :1.00000   
## month.June month.Mar month.May month.Nov   
## Min. :0.00000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.00000 Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.02334 Mean :0.1526 Mean :0.2726 Mean :0.2443   
## 3rd Qu.:0.00000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :1.00000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## month.Oct month.Sep operatingsystems.1 operatingsystems.2  
## Min. :0.00000 Min. :0.00000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.00000 Median :0.00000 Median :0.0000 Median :1.0000   
## Mean :0.04496 Mean :0.03669 Mean :0.2089 Mean :0.5357   
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.0000 3rd Qu.:1.0000   
## Max. :1.00000 Max. :1.00000 Max. :1.0000 Max. :1.0000   
## operatingsystems.3 operatingsystems.4 operatingsystems.5 operatingsystems.6  
## Min. :0.0000 Min. :0.00000 Min. :0.0000000 Min. :0.000000   
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.0000000 1st Qu.:0.000000   
## Median :0.0000 Median :0.00000 Median :0.0000000 Median :0.000000   
## Mean :0.2074 Mean :0.03914 Mean :0.0004914 Mean :0.001556   
## 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.0000000 3rd Qu.:0.000000   
## Max. :1.0000 Max. :1.00000 Max. :1.0000000 Max. :1.000000   
## operatingsystems.7 operatingsystems.8 browser.1 browser.12   
## Min. :0.0000000 Min. :0.000000 Min. :0.0000 Min. :0.0000000   
## 1st Qu.:0.0000000 1st Qu.:0.000000 1st Qu.:0.0000 1st Qu.:0.0000000   
## Median :0.0000000 Median :0.000000 Median :0.0000 Median :0.0000000   
## Mean :0.0005733 Mean :0.006142 Mean :0.1989 Mean :0.0008189   
## 3rd Qu.:0.0000000 3rd Qu.:0.000000 3rd Qu.:0.0000 3rd Qu.:0.0000000   
## Max. :1.0000000 Max. :1.000000 Max. :1.0000 Max. :1.0000000   
## browser.13 browser.2 browser.3 browser.4   
## Min. :0.000000 Min. :0.0000 Min. :0.000000 Min. :0.00000   
## 1st Qu.:0.000000 1st Qu.:0.0000 1st Qu.:0.000000 1st Qu.:0.00000   
## Median :0.000000 Median :1.0000 Median :0.000000 Median :0.00000   
## Mean :0.004586 Mean :0.6458 Mean :0.008599 Mean :0.05986   
## 3rd Qu.:0.000000 3rd Qu.:1.0000 3rd Qu.:0.000000 3rd Qu.:0.00000   
## Max. :1.000000 Max. :1.0000 Max. :1.000000 Max. :1.00000   
## browser.5 browser.6 browser.7 browser.8   
## Min. :0.00000 Min. :0.00000 Min. :0.000000 Min. :0.00000   
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.000000 1st Qu.:0.00000   
## Median :0.00000 Median :0.00000 Median :0.000000 Median :0.00000   
## Mean :0.03816 Mean :0.01425 Mean :0.004013 Mean :0.01106   
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.000000 3rd Qu.:0.00000   
## Max. :1.00000 Max. :1.00000 Max. :1.000000 Max. :1.00000   
## browser.9 region.1 region.2 region.3   
## Min. :0.00e+00 Min. :0.0000 Min. :0.00000 Min. :0.0000   
## 1st Qu.:0.00e+00 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.0000   
## Median :0.00e+00 Median :0.0000 Median :0.00000 Median :0.0000   
## Mean :8.19e-05 Mean :0.3861 Mean :0.09246 Mean :0.1952   
## 3rd Qu.:0.00e+00 3rd Qu.:1.0000 3rd Qu.:0.00000 3rd Qu.:0.0000   
## Max. :1.00e+00 Max. :1.0000 Max. :1.00000 Max. :1.0000   
## region.4 region.5 region.6 region.7   
## Min. :0.0000 Min. :0.00000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.00000   
## Median :0.0000 Median :0.00000 Median :0.0000 Median :0.00000   
## Mean :0.0959 Mean :0.02604 Mean :0.0656 Mean :0.06208   
## 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.0000 3rd Qu.:0.00000   
## Max. :1.0000 Max. :1.00000 Max. :1.0000 Max. :1.00000   
## region.8 region.9 traffictype.1 traffictype.10   
## Min. :0.0000 Min. :0.00000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.00000   
## Median :0.0000 Median :0.00000 Median :0.0000 Median :0.00000   
## Mean :0.0353 Mean :0.04136 Mean :0.1957 Mean :0.03685   
## 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.0000 3rd Qu.:0.00000   
## Max. :1.0000 Max. :1.00000 Max. :1.0000 Max. :1.00000   
## traffictype.11 traffictype.12 traffictype.13 traffictype.14   
## Min. :0.00000 Min. :0.00e+00 Min. :0.0000 Min. :0.000000   
## 1st Qu.:0.00000 1st Qu.:0.00e+00 1st Qu.:0.0000 1st Qu.:0.000000   
## Median :0.00000 Median :0.00e+00 Median :0.0000 Median :0.000000   
## Mean :0.02023 Mean :8.19e-05 Mean :0.0597 Mean :0.001065   
## 3rd Qu.:0.00000 3rd Qu.:0.00e+00 3rd Qu.:0.0000 3rd Qu.:0.000000   
## Max. :1.00000 Max. :1.00e+00 Max. :1.0000 Max. :1.000000   
## traffictype.15 traffictype.16 traffictype.17 traffictype.18   
## Min. :0.00000 Min. :0.0000000 Min. :0.00e+00 Min. :0.0000000   
## 1st Qu.:0.00000 1st Qu.:0.0000000 1st Qu.:0.00e+00 1st Qu.:0.0000000   
## Median :0.00000 Median :0.0000000 Median :0.00e+00 Median :0.0000000   
## Mean :0.00303 Mean :0.0002457 Mean :8.19e-05 Mean :0.0008189   
## 3rd Qu.:0.00000 3rd Qu.:0.0000000 3rd Qu.:0.00e+00 3rd Qu.:0.0000000   
## Max. :1.00000 Max. :1.0000000 Max. :1.00e+00 Max. :1.0000000   
## traffictype.19 traffictype.2 traffictype.20 traffictype.3   
## Min. :0.000000 Min. :0.0000 Min. :0.00000 Min. :0.0000   
## 1st Qu.:0.000000 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.0000   
## Median :0.000000 Median :0.0000 Median :0.00000 Median :0.0000   
## Mean :0.001392 Mean :0.3201 Mean :0.01581 Mean :0.1653   
## 3rd Qu.:0.000000 3rd Qu.:1.0000 3rd Qu.:0.00000 3rd Qu.:0.0000   
## Max. :1.000000 Max. :1.0000 Max. :1.00000 Max. :1.0000   
## traffictype.4 traffictype.5 traffictype.6 traffictype.7   
## Min. :0.0000 Min. :0.00000 Min. :0.00000 Min. :0.000000   
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.000000   
## Median :0.0000 Median :0.00000 Median :0.00000 Median :0.000000   
## Mean :0.0873 Mean :0.02129 Mean :0.03628 Mean :0.003276   
## 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.000000   
## Max. :1.0000 Max. :1.00000 Max. :1.00000 Max. :1.000000   
## traffictype.8 traffictype.9 visitortype.New\_Visitor  
## Min. :0.00000 Min. :0.000000 Min. :0.0000   
## 1st Qu.:0.00000 1st Qu.:0.000000 1st Qu.:0.0000   
## Median :0.00000 Median :0.000000 Median :0.0000   
## Mean :0.02809 Mean :0.003358 Mean :0.1386   
## 3rd Qu.:0.00000 3rd Qu.:0.000000 3rd Qu.:0.0000   
## Max. :1.00000 Max. :1.000000 Max. :1.0000   
## visitortype.Other visitortype.Returning\_Visitor weekendFALSE   
## Min. :0.000000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.000000 1st Qu.:1.0000 1st Qu.:1.0000   
## Median :0.000000 Median :1.0000 Median :1.0000   
## Mean :0.006633 Mean :0.8547 Mean :0.7659   
## 3rd Qu.:0.000000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.000000 Max. :1.0000 Max. :1.0000   
## weekendTRUE revenueFALSE revenueTRUE   
## Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:0.0000   
## Median :0.0000 Median :1.0000 Median :0.0000   
## Mean :0.2341 Mean :0.8437 Mean :0.1563   
## 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000

## 6.2 K-Means Clustering

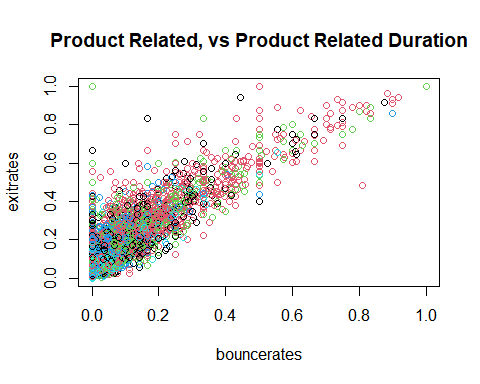
# Applying K-Means Clustering algorithm   
data.clusters <- kmeans(data\_normal, 5)  
data.clusters$size

## [1] 1770 4331 2516 2042 1552

# Product Related, vs Product Related duration  
plot(data\_normal[, 5:6], col = data.clusters$cluster, main = 'product related v product related duration')

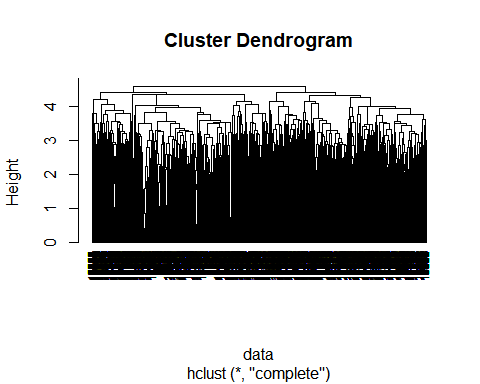


# Product Related, vs Product Related Duration  
plot(data\_normal[, 7:8], col = data.clusters$cluster, main = 'Product Related, vs Product Related Duration')



## 6.3 Hierachical Clustering

data <- dist(data\_normal, method = "euclidean")  
# We then apply hierarchical clustering using the Ward's method  
res.hc <- hclust(data, method = "complete")  
# Lastly we plot the obtained dendrogram  
#--  
plot(res.hc, cex = 0.6, hang = -1)



### 6.3 Principal Component Analysis (PCA)

# Applying PCA  
data\_pca <- prcomp(data\_normal,center = TRUE, scale = FALSE)   
summary(data\_pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 0.7372 0.62920 0.60672 0.54811 0.53561 0.51388 0.49702  
## Proportion of Variance 0.1178 0.08581 0.07979 0.06512 0.06218 0.05724 0.05354  
## Cumulative Proportion 0.1178 0.20359 0.28338 0.34849 0.41067 0.46791 0.52145  
## PC8 PC9 PC10 PC11 PC12 PC13 PC14  
## Standard deviation 0.4552 0.43190 0.39599 0.38460 0.37530 0.33907 0.31626  
## Proportion of Variance 0.0449 0.04043 0.03399 0.03206 0.03053 0.02492 0.02168  
## Cumulative Proportion 0.5664 0.60678 0.64077 0.67283 0.70336 0.72828 0.74996  
## PC15 PC16 PC17 PC18 PC19 PC20 PC21  
## Standard deviation 0.30735 0.29719 0.27539 0.26679 0.25308 0.2468 0.23407  
## Proportion of Variance 0.02047 0.01914 0.01644 0.01543 0.01388 0.0132 0.01188  
## Cumulative Proportion 0.77043 0.78958 0.80601 0.82144 0.83532 0.8485 0.86040  
## PC22 PC23 PC24 PC25 PC26 PC27 PC28  
## Standard deviation 0.21953 0.21405 0.20867 0.2026 0.19414 0.1897 0.18782  
## Proportion of Variance 0.01045 0.00993 0.00944 0.0089 0.00817 0.0078 0.00765  
## Cumulative Proportion 0.87085 0.88078 0.89022 0.8991 0.90728 0.9151 0.92273  
## PC29 PC30 PC31 PC32 PC33 PC34 PC35  
## Standard deviation 0.18497 0.17401 0.17207 0.16558 0.16386 0.15672 0.1519  
## Proportion of Variance 0.00742 0.00656 0.00642 0.00594 0.00582 0.00532 0.0050  
## Cumulative Proportion 0.93015 0.93671 0.94313 0.94907 0.95489 0.96021 0.9652  
## PC36 PC37 PC38 PC39 PC40 PC41 PC42  
## Standard deviation 0.14283 0.14101 0.13171 0.12726 0.12077 0.10243 0.09566  
## Proportion of Variance 0.00442 0.00431 0.00376 0.00351 0.00316 0.00227 0.00198  
## Cumulative Proportion 0.96964 0.97395 0.97771 0.98122 0.98438 0.98665 0.98864  
## PC43 PC44 PC45 PC46 PC47 PC48 PC49  
## Standard deviation 0.08709 0.06836 0.06714 0.06492 0.06325 0.05708 0.05696  
## Proportion of Variance 0.00164 0.00101 0.00098 0.00091 0.00087 0.00071 0.00070  
## Cumulative Proportion 0.99028 0.99129 0.99227 0.99319 0.99405 0.99476 0.99546  
## PC50 PC51 PC52 PC53 PC54 PC55 PC56  
## Standard deviation 0.05528 0.05114 0.04390 0.04361 0.04153 0.04006 0.03855  
## Proportion of Variance 0.00066 0.00057 0.00042 0.00041 0.00037 0.00035 0.00032  
## Cumulative Proportion 0.99612 0.99669 0.99711 0.99752 0.99789 0.99824 0.99856  
## PC57 PC58 PC59 PC60 PC61 PC62 PC63  
## Standard deviation 0.03368 0.03269 0.03168 0.02921 0.02721 0.02570 0.02185  
## Proportion of Variance 0.00025 0.00023 0.00022 0.00018 0.00016 0.00014 0.00010  
## Cumulative Proportion 0.99881 0.99904 0.99926 0.99944 0.99960 0.99975 0.99985  
## PC64 PC65 PC66 PC67 PC68 PC69  
## Standard deviation 0.01609 0.01324 0.009503 0.009037 9e-03 3.064e-15  
## Proportion of Variance 0.00006 0.00004 0.000020 0.000020 2e-05 0.000e+00  
## Cumulative Proportion 0.99991 0.99995 0.999960 0.999980 1e+00 1.000e+00  
## PC70 PC71 PC72 PC73 PC74  
## Standard deviation 2.3e-15 1.351e-15 1.012e-15 8.141e-16 7.667e-16  
## Proportion of Variance 0.0e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00  
## Cumulative Proportion 1.0e+00 1.000e+00 1.000e+00 1.000e+00 1.000e+00  
## PC75  
## Standard deviation 3.274e-16  
## Proportion of Variance 0.000e+00  
## Cumulative Proportion 1.000e+00

# We get 75 principle components. We can compare this with t-SNE results

## Conclusion

* Kira Plastinina should use the K Means clustering, the clusters are clearer.