Answers

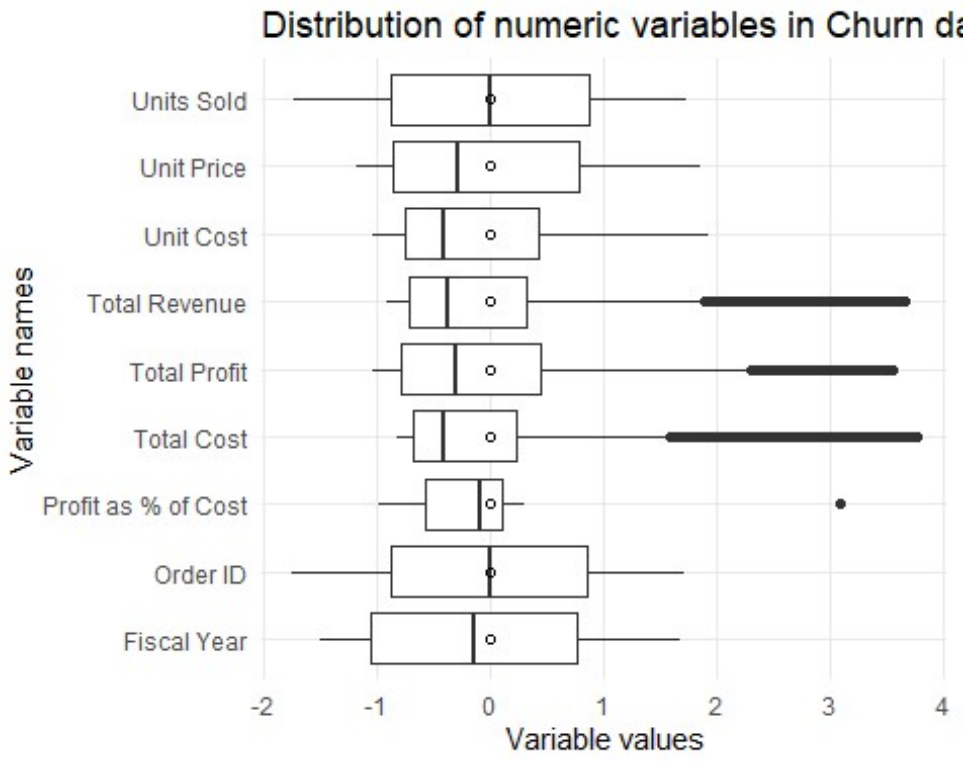
```
library(readr)
#library(tidyverse)
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(fastDummies)
library(ggplot2)
library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa

library(cowplot)
library(e1071)
library(knitr)
library(ggcorrplot)
library(corrplot)

## corrplot 0.88 loaded

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyr)

# Reading the file
Data_Science_Evaluation <- read_csv("C:/Users/KUNAL/Desktop/DS_INTER/Data
Science Evaluation.csv")

##
## -- Column specification --------------------------------------------------
------
## cols(
##   Region = col_character(),
##   Country = col_character(),
```

```
##   `Item Type` = col_character(),
##   `Fiscal Year` = col_double(),
##   `Sales Channel` = col_character(),
##   `Order Priority` = col_character(),
##   `Order Date` = col_character(),
##   `Order ID` = col_double(),
##   `Ship Date` = col_character(),
##   `Units Sold` = col_double(),
##   `Unit Price` = col_double(),
##   `Unit Cost` = col_double(),
##   `Total Revenue` = col_double(),
##   `Total Cost` = col_double(),
##   `Total Profit` = col_double(),
##   `Profit as % of Cost` = col_double()
## )

summary(Data_Science_Evaluation)

##     Region             Country           Item Type          Fiscal Year
##  Length:65535       Length:65535       Length:65535       Min.   :2010
##  Class :character   Class :character   Class :character   1st Qu.:2011
##  Mode  :character   Mode  :character   Mode  :character   Median :2013
##                                                           Mean   :2013
##                                                           3rd Qu.:2015
##                                                           Max.   :2017
##  Sales Channel      Order Priority      Order Date           Order ID
##  Length:65535       Length:65535       Length:65535       Min.
:100014913
##  Class :character   Class :character   Class :character   1st
Qu.:326706421
##  Mode  :character   Mode  :character   Mode  :character   Median
:552128064
##                                                           Mean
:552992170
##                                                           3rd
Qu.:778687822
##                                                           Max.
:999993884
##   Ship Date           Units Sold      Unit Price       Unit Cost
##  Length:65535       Min.   :    1   Min.   :  9.33   Min.   :  6.92
##  Class :character   1st Qu.: 2470   1st Qu.: 81.73   1st Qu.: 56.67
##  Mode  :character   Median : 4983   Median :205.70   Median :117.11
##                     Mean   : 4987   Mean   :266.19   Mean   :187.61
##                     3rd Qu.: 7500   3rd Qu.:437.20   3rd Qu.:263.33
##                     Max.   :10000   Max.   :668.27   Max.   :524.96
##  Total Revenue       Total Cost       Total Profit       Profit as % of
Cost
##  Min.   :     37   Min.   :     28   Min.   :      9.6   Min.   :0.1568
##  1st Qu.: 276414   1st Qu.: 162659   1st Qu.:  95668.3   1st Qu.:0.3483
##  Median : 787313   Median : 465925   Median : 280908.0   Median :0.5659
```

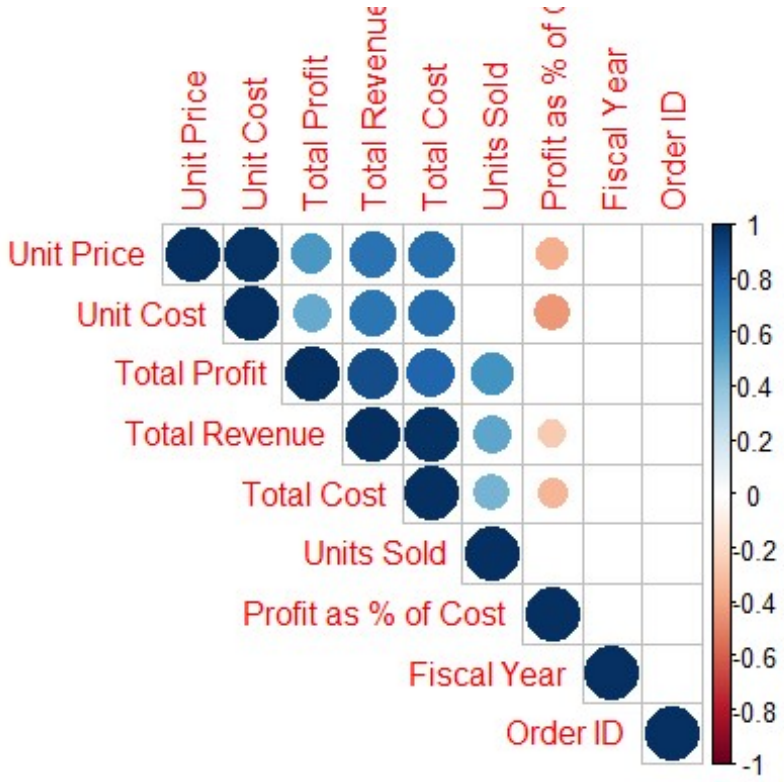
```
##  Mean   :1326827   Mean   : 935066   Mean   : 391761.0   Mean   :0.6107
##  3rd Qu.:1810119   3rd Qu.:1197962   3rd Qu.: 563289.5   3rd Qu.:0.6603
##  Max.   :6682032   Max.   :5249600   Max.   :1738700.0   Max.   :2.0491

#Checking for any NA values
any(colSums(is.na(Data_Science_Evaluation)) != 0)

## [1] FALSE

Data_Science_Evaluation <- na.omit(Data_Science_Evaluation)

str(Data_Science_Evaluation)

## tibble [65,535 x 16] (S3: tbl_df/tbl/data.frame)
##  $ Region             : chr [1:65535] "Sub-Saharan Africa" "Sub-Saharan
Africa" "Central America and the Caribbean" "Europe" ...
##  $ Country            : chr [1:65535] "Cote d'Ivoire" "Ethiopia" "Panama"
"Sweden" ...
##  $ Item Type          : chr [1:65535] "Snacks" "Snacks" "Clothes" "Office
Supplies" ...
##  $ Fiscal Year        : num [1:65535] 2010 2010 2011 2012 2016 ...
##  $ Sales Channel      : chr [1:65535] "Online" "Online" "Offline" "Online"
...
##  $ Order Priority     : chr [1:65535] "C" "H" "H" "L" ...
##  $ Order Date         : chr [1:65535] "4/23/2010" "6/6/2010" "1/2/2011"
"10/2/2012" ...
##  $ Order ID           : num [1:65535] 2.42e+08 5.30e+08 2.93e+08 3.61e+08
7.07e+08 ...
##  $ Ship Date          : chr [1:65535] "4/26/2010" "7/1/2010" "1/22/2011"
"10/20/2012" ...
##  $ Units Sold         : num [1:65535] 10000 10000 10000 10000 10000 ...
##  $ Unit Price         : num [1:65535] 153 153 109 651 437 ...
##  $ Unit Cost          : num [1:65535] 97.4 97.4 35.8 525 263.3 ...
##  $ Total Revenue      : num [1:65535] 1525800 1525800 1092800 6512100
4372000 ...
##  $ Total Cost         : num [1:65535] 974400 974400 358400 5249600 2633300
...
##  $ Total Profit       : num [1:65535] 551400 551400 734400 1262500 1738700
...
##  $ Profit as % of Cost: num [1:65535] 0.566 0.566 2.049 0.24 0.66 ...

#Checking for outliers
Data_Science_Evaluation %>% select_if(is.numeric) %>% mutate_all(scale) %>%
gather("features","values") %>% na.omit() %>%
  ggplot(aes(x = features, y = values)) +
  geom_boxplot(show.legend = FALSE) +
  stat_summary(fun = mean, geom = "point", pch = 1) + # Add average to the
boxplot
  scale_y_continuous(name = "Variable values", minor_breaks = NULL) +
  scale_fill_brewer(palette = "Set1") +
  coord_flip() +
```

```
  theme_minimal() +
  labs(x = "Variable names") +
  ggtitle(label = "Distribution of numeric variables in Churn dataset")
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

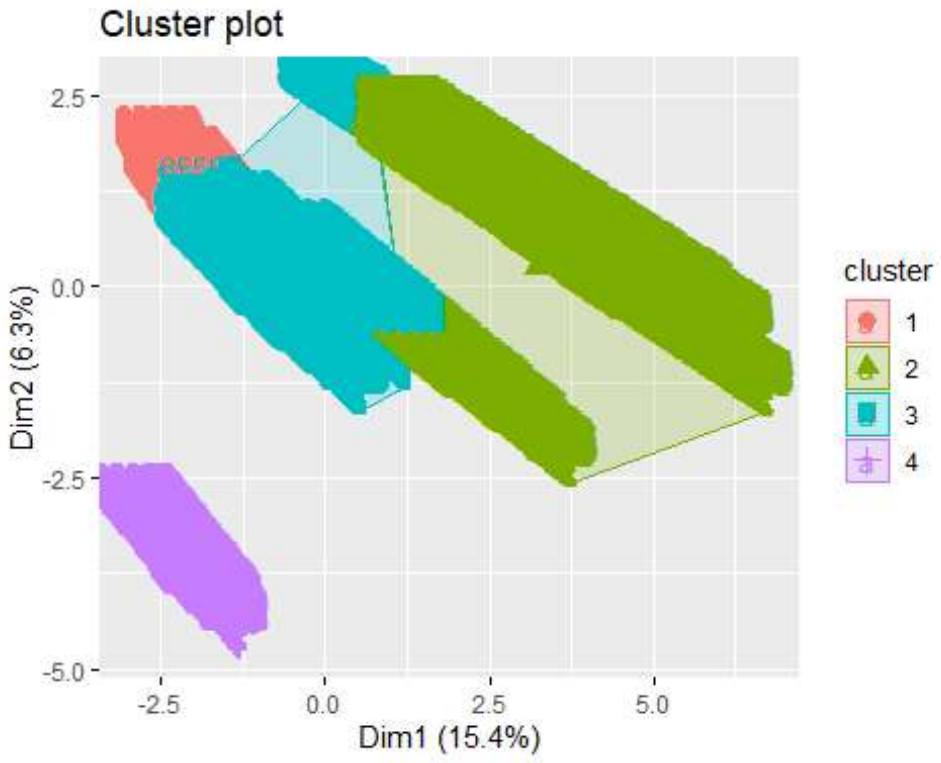



```
#We don't see many outliers so we can proceed with the data and scale it
#Just checking the correlation between the values
corrplot(cor(Data_Science_Evaluation %>% select_if(is.numeric)),
type="upper", order="hclust",method="circle")
```




```
#Creating factor,dummies and scaling the data
Data_Science_Evaluation_factors <- Data_Science_Evaluation %>%
  select(Region,`Item Type`, `Sales Channel`, `Order Priority`) %>%
  mutate_all(.funs = function(x){as.factor((x))})

Data_Science_Evaluation_factors_dummy <-
dummy_cols(Data_Science_Evaluation_factors) %>% select(-c(Region,`Item Type`,
`Sales Channel`, `Order Priority`))

Data_Science_Evaluation_new <- Data_Science_Evaluation %>%
  select(-c(Country,Region,`Order Date`,`Ship Date`,`Item Type`, `Sales
Channel`, `Order Priority`)) %>%
  cbind(Data_Science_Evaluation_factors_dummy)
set.seed(123)
str(Data_Science_Evaluation_new)

## 'data.frame':    65535 obs. of  34 variables:
##  $ Fiscal Year                          : num  2010 2010 2011 2012 2016
...
##  $ Order ID                             : num  2.42e+08 5.30e+08
2.93e+08 3.61e+08 7.07e+08 ...
##  $ Units Sold                           : num  10000 10000 10000 10000
10000 ...
##  $ Unit Price                           : num  153 153 109 651 437 ...
##  $ Unit Cost                            : num  97.4 97.4 35.8 525 263.3
...
##  $ Total Revenue                        : num  1525800 1525800 1092800
```

```
6512100 4372000 ...
##  $ Total Cost                                : num  974400 974400 358400
5249600 2633300 ...
##  $ Total Profit                              : num  551400 551400 734400
1262500 1738700 ...
##  $ Profit as % of Cost                       : num  0.566 0.566 2.049 0.24
0.66 ...
##  $ Region_Asia                               : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ Region_Australia and Oceania              : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Region_Central America and the Caribbean: int  0 0 1 0 0 0 0 0 0 0 ...
##  $ Region_Europe                             : int  0 0 0 1 1 0 0 0 0 0 ...
##  $ Region_Middle East and North Africa       : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Region_North America                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Region_Sub-Saharan Africa                 : int  1 1 0 0 0 0 1 0 0 1 ...
##  $ Item Type_Baby Food                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Item Type_Beverages                       : int  0 0 0 0 0 1 1 0 0 0 ...
##  $ Item Type_Cereal                          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Item Type_Clothes                         : int  0 0 1 0 0 0 0 1 1 0 ...
##  $ Item Type_Cosmetics                       : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ Item Type_Fruits                          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Item Type_Household                       : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ Item Type_Meat                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Item Type_Office Supplies                 : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ Item Type_Personal Care                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Item Type_Snacks                          : int  1 1 0 0 0 0 0 0 0 0 ...
##  $ Item Type_Vegetables                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sales Channel_Offline                     : int  0 0 1 0 0 1 0 0 0 0 ...
##  $ Sales Channel_Online                      : int  1 1 0 1 1 0 1 1 1 1 ...
##  $ Order Priority_C                          : int  1 0 0 0 0 1 0 0 1 0 ...
##  $ Order Priority_H                          : int  0 1 1 0 0 0 0 1 0 0 ...
##  $ Order Priority_L                          : int  0 0 0 1 0 0 1 0 0 1 ...
##  $ Order Priority_M                          : int  0 0 0 0 1 0 0 0 0 0 ...

Data_Science_Evaluation_new_scale <- scale(Data_Science_Evaluation_new)

data_sci_kmeans <- kmeans(Data_Science_Evaluation_new_scale, centers = 4,
nstart=25)
fviz_cluster(data_sci_kmeans, data= Data_Science_Evaluation_new_scale)
```




```
set.seed(123)
#Just checking the clusters and we can see the cluster 4 is the outliers

#Adding the clusters value to the  data
Data_Science_Evaluation_copy<-Data_Science_Evaluation
Data_Science_Evaluation_copy<-
cbind(Data_Science_Evaluation_copy,Cluster=data_sci_kmeans$cluster)

Data_Science_Evaluation_new_scale_cluster4<-Data_Science_Evaluation_copy %>%
  filter(Data_Science_Evaluation_copy["Cluster"]==4)
Data_Science_Evaluation_new_scale_cluster3<-Data_Science_Evaluation_copy %>%
  filter(Data_Science_Evaluation_copy["Cluster"]==3)
Data_Science_Evaluation_new_scale_cluster2<-Data_Science_Evaluation_copy %>%
  filter(Data_Science_Evaluation_copy["Cluster"]==2)
Data_Science_Evaluation_new_scale_cluster1<-Data_Science_Evaluation_copy %>%
  filter(Data_Science_Evaluation_copy["Cluster"]==1)

summary(Data_Science_Evaluation_new_scale_cluster4)

##     Region             Country           Item Type          Fiscal Year
##  Length:5414        Length:5414        Length:5414        Min.   :2010
##  Class :character   Class :character   Class :character   1st Qu.:2011
##  Mode  :character   Mode  :character   Mode  :character   Median :2013
##                                                           Mean   :2013
##                                                           3rd Qu.:2015
##                                                           Max.   :2017
##  Sales Channel      Order Priority      Order Date           Order ID
```

```
##  Length:5414        Length:5414        Length:5414        Min.
:100358235
##  Class :character   Class :character   Class :character   1st
Qu.:322617018
##  Mode  :character   Mode  :character   Mode  :character   Median
:545506012
##                                                          Mean
:549057229
##                                                          3rd
Qu.:774274730
##                                                          Max.
:999891316
##   Ship Date          Units Sold      Unit Price      Unit Cost
##  Length:5414        Min.   :    4   Min.   :109.3   Min.   :35.84
##  Class :character   1st Qu.: 2463   1st Qu.:109.3   1st Qu.:35.84
##  Mode  :character   Median : 5166   Median :109.3   Median :35.84
##                     Mean   : 5055   Mean   :109.3   Mean   :35.84
##                     3rd Qu.: 7566   3rd Qu.:109.3   3rd Qu.:35.84
##                     Max.   :10000   Max.   :109.3   Max.   :35.84
##  Total Revenue         Total Cost        Total Profit      Profit as % of
Cost
##  Min.   :    437.1   Min.   :   143.4   Min.   :   293.8   Min.   :2.049
##  1st Qu.: 269184.0   1st Qu.: 88282.9   1st Qu.:180901.1   1st Qu.:2.049
##  Median : 564595.1   Median :185167.4   Median :379427.8   Median :2.049
##  Mean   : 552385.7   Mean   :181163.1   Mean   :371222.6   Mean   :2.049
##  3rd Qu.: 826785.2   3rd Qu.:271156.5   3rd Qu.:555628.7   3rd Qu.:2.049
##  Max.   :1092800.0   Max.   :358400.0   Max.   :734400.0   Max.   :2.049
##     Cluster
##  Min.   :4
##  1st Qu.:4
##  Median :4
##  Mean   :4
##  3rd Qu.:4
##  Max.   :4

summary(Data_Science_Evaluation_new_scale_cluster3)

##     Region             Country           Item Type          Fiscal Year
##  Length:25591       Length:25591       Length:25591       Min.   :2010
##  Class :character   Class :character   Class :character   1st Qu.:2011
##  Mode  :character   Mode  :character   Mode  :character   Median :2013
##                                                          Mean   :2013
##                                                          3rd Qu.:2015
##                                                          Max.   :2017
##  Sales Channel      Order Priority      Order Date           Order ID
##  Length:25591       Length:25591       Length:25591       Min.
:100014913
##  Class :character   Class :character   Class :character   1st
Qu.:328956472
##  Mode  :character   Mode  :character   Mode  :character   Median
```

```
:552506577
##                                                                Mean
:553202614
##                                                                3rd
Qu.:776940310
##                                                                Max.
:999906766
##   Ship Date           Units Sold      Unit Price      Unit Cost
##  Length:25591       Min.   :    1   Min.   :152.6   Min.   : 90.93
##  Class :character   1st Qu.: 1913   1st Qu.:154.1   1st Qu.: 97.44
##  Mode  :character   Median : 4107   Median :205.7   Median :117.11
##                     Mean   : 4509   Mean   :226.8   Mean   :143.27
##                     3rd Qu.: 7050   3rd Qu.:255.3   3rd Qu.:159.42
##                     Max.   :10000   Max.   :437.2   Max.   :364.69
##  Total Revenue         Total Cost         Total Profit      Profit as % of
Cost
##  Min.   :    152.6   Min.   :     90.9   Min.   :    55.1   Min.   :0.1568
##  1st Qu.: 431970.0   1st Qu.: 268212.7   1st Qu.:146746.3   1st Qu.:0.5659
##  Median : 874130.8   Median : 545034.4   Median :334210.2   Median :0.6603
##  Mean   : 926289.3   Mean   : 568377.9   Mean   :357911.4   Mean   :0.6274
##  3rd Qu.:1341518.9   3rd Qu.: 820498.1   3rd Qu.:526636.3   3rd Qu.:0.6943
##  Max.   :2552289.4   Max.   :1593881.2   Max.   :958408.3   Max.   :0.7565
##     Cluster
##  Min.   :3
##  1st Qu.:3
##  Median :3
##  Mean   :3
##  3rd Qu.:3
##  Max.   :3

summary(Data_Science_Evaluation_new_scale_cluster2)

##     Region             Country           Item Type          Fiscal Year
##  Length:18112       Length:18112       Length:18112       Min.   :2010
##  Class :character   Class :character   Class :character   1st Qu.:2011
##  Mode  :character   Mode  :character   Mode  :character   Median :2013
##                                                           Mean   :2013
##                                                           3rd Qu.:2015
##                                                           Max.   :2017
##  Sales Channel      Order Priority      Order Date           Order ID
##  Length:18112       Length:18112       Length:18112       Min.
:100135505
##  Class :character   Class :character   Class :character   1st
Qu.:325533866
##  Mode  :character   Mode  :character   Mode  :character   Median
:552266252
##                                                           Mean
:553444654
##                                                           3rd
Qu.:782202053
```

```
##                                                                Max.
:999961698
##   Ship Date           Units Sold      Unit Price      Unit Cost
##  Length:18112       Min.   :    2   Min.   :421.9   Min.   :263.3
##  Class :character   1st Qu.: 3610   1st Qu.:437.2   1st Qu.:364.7
##  Mode  :character   Median : 5808   Median :651.2   Median :502.5
##                     Mean   : 5641   Mean   :568.0   Mean   :436.8
##                     3rd Qu.: 7882   3rd Qu.:668.3   3rd Qu.:525.0
##                     Max.   :10000   Max.   :668.3   Max.   :525.0
##  Total Revenue       Total Cost       Total Profit       Profit as % of
Cost
##  Min.   :   1302   Min.   :   1050   Min.   :    252.5   Min.   :0.1568
##  1st Qu.:1950265   1st Qu.:1441026   1st Qu.: 346452.8   1st Qu.:0.2405
##  Median :3013182   Median :2223266   Median : 663931.6   Median :0.2405
##  Mean   :3117414   Mean   :2381834   Mean   : 735579.6   Mean   :0.3227
##  3rd Qu.:4109196   3rd Qu.:3249951   3rd Qu.:1099660.1   3rd Qu.:0.3298
##  Max.   :6682032   Max.   :5249600   Max.   :1738700.0   Max.   :0.6603
##     Cluster
##  Min.   :2
##  1st Qu.:2
##  Median :2
##  Mean   :2
##  3rd Qu.:2
##  Max.   :2

summary(Data_Science_Evaluation_new_scale_cluster1)

##     Region             Country           Item Type          Fiscal Year
##  Length:16418       Length:16418       Length:16418       Min.   :2010
##  Class :character   Class :character   Class :character   1st Qu.:2011
##  Mode  :character   Mode  :character   Mode  :character   Median :2013
##                                                           Mean   :2013
##                                                           3rd Qu.:2015
##                                                           Max.   :2017
##  Sales Channel      Order Priority      Order Date           Order ID
##  Length:16418       Length:16418       Length:16418       Min.
:100023925
##  Class :character   Class :character   Class :character   1st
Qu.:325774407
##  Mode  :character   Mode  :character   Mode  :character   Median
:553482472
##                                                           Mean
:553462562
##                                                           3rd
Qu.:779988479
##                                                           Max.
:999993884
##   Ship Date           Units Sold      Unit Price      Unit Cost
##  Length:16418       Min.   :    2   Min.   : 9.33   Min.   : 6.92
##  Class :character   1st Qu.: 2475   1st Qu.: 9.33   1st Qu.: 6.92
```

```
##  Mode  :character   Median : 4962   Median :47.45   Median :31.79
##                     Mean   : 4987   Mean   :46.30   Mean   :31.88
##                     3rd Qu.: 7514   3rd Qu.:81.73   3rd Qu.:56.67
##                     Max.   :10000   Max.   :81.73   Max.   :56.67
##  Total Revenue        Total Cost        Total Profit       Profit as % of
Cost
##  Min.   :    37.3   Min.   :    27.7   Min.   :     9.64   Min.   :0.3483
##  1st Qu.: 52968.7   1st Qu.: 38213.5   1st Qu.: 14426.86   1st Qu.:0.3483
##  Median :149688.5   Median :101664.4   Median : 47927.43   Median :0.4422
##  Mean   :231194.6   Mean   :159192.5   Mean   : 72002.14   Mean   :0.4280
##  3rd Qu.:375566.8   3rd Qu.:254519.0   3rd Qu.:120607.17   3rd Qu.:0.4926
##  Max.   :817136.5   Max.   :566586.7   Max.   :250549.88   Max.   :0.4926
##     Cluster
##  Min.   :1
##  1st Qu.:1
##  Median :1
##  Mean   :1
##  3rd Qu.:1
##  Max.   :1
```

```
#After analyzing all the cluster's we can see that cluster 1 has minimum % to
profit ratio. Hence, we could eliminate
#few items from every location that are captured in cluster 1

#let's analyze further cluster 1 to make a specific decision
Data_Science_Evaluation_new_scale_cluster1_group<-
Data_Science_Evaluation_new_scale_cluster1 %>%
  group_by(Data_Science_Evaluation_new_scale_cluster1$`Item Type`) %>%
  summarise(Total_Profit_Cluster1_Iteams= sum(`Total Profit`))
Data_Science_Evaluation_new_scale_cluster1_group
```

```
## # A tibble: 3 x 2
##   `Data_Science_Evaluation_new_scale_cluster1$\`Item~
Total_Profit_Cluster1_Ite~
##   <chr>
<dbl>
## 1 Beverages
429395415.
## 2 Fruits
65128659.
## 3 Personal Care
687607031.
```

```
#we can see the minimum profit is generated by the fruits in each region
#Hence we can reduce the selling of fruits from several region that have
minimum sales profit from fruits
Data_Science_Evaluation_new_scale_cluster1_fruits<-
Data_Science_Evaluation_new_scale_cluster1 %>%
  filter(`Total Profit`<mean(`Total Profit`))

Data_Science_Evaluation_new_scale_cluster1_fruits_COuntries<-
```

```
Data_Science_Evaluation_new_scale_cluster1_fruits %>%
  group_by(Region) %>% summarise(T.Profit=sum(`Total Profit`)) %>%
arrange(desc(T.Profit))


#1 Fruits should be reduced from NorTh america.
#Personal care should be sold more because it helps to generate high revenue.
#Central America and the Caribbean, Australia and Oceania, North America
should be focused more and should be given
#more given more preference so that Total profit can be increased from these
places

#Question2
Data_Science_Evaluation_copy["Total Days of shipment"]<-
as.Date(Data_Science_Evaluation_copy$`Ship Date`,format="%m/%d/%Y")-
as.Date(Data_Science_Evaluation_copy$`Order Date`,format="%m/%d/%Y")

Data_Science_Evaluation_copy %>%
  group_by(`Item Type`,Region) %>%
  summarise(td<-sum(`Total Days of shipment`)) %>%
  filter(`td <- sum(\`Total Days of shipment\`)`>mean(`td <- sum(\`Total Days
of shipment\`)`))

## `summarise()` has grouped output by 'Item Type'. You can override using
the `.groups` argument.

## # A tibble: 32 x 3
## # Groups:   Item Type [12]
##    `Item Type` Region             `td <- sum(\`Total Days of shipment\`)`
##    <chr>       <chr>              <drtn>
##  1 Baby Food   Asia               21094 days
##  2 Baby Food   Europe             36763 days
##  3 Baby Food   Sub-Saharan Africa 34317 days
##  4 Beverages   Europe             36064 days
##  5 Beverages   Sub-Saharan Africa 35729 days
##  6 Cereal      Asia               19320 days
##  7 Cereal      Europe             37059 days
##  8 Cereal      Sub-Saharan Africa 35093 days
##  9 Clothes     Asia               20275 days
## 10 Clothes     Europe             35805 days
## # ... with 22 more rows

# Baby products, beverage and cereals are easiest to sell
#No! each region has totally different relation with the given product
```