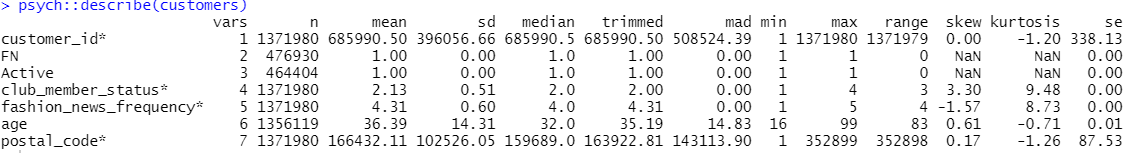
Two dataset has been taken for the analysis

1. Article.csv
2. Customers.csv

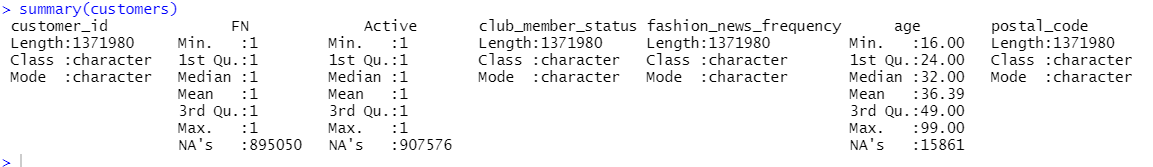
Description for the customer dataset is shown below



Few points which we get from the description are:

* Average age of person who purchases cloth from H&M is 36 year.
* Fashion news frequency

Summary for all the variable for customer dataset



* Club member status is divided into three parts

1. Active
2. Pre-create
3. Left club



* We have calculated the proportion of each cub member status.

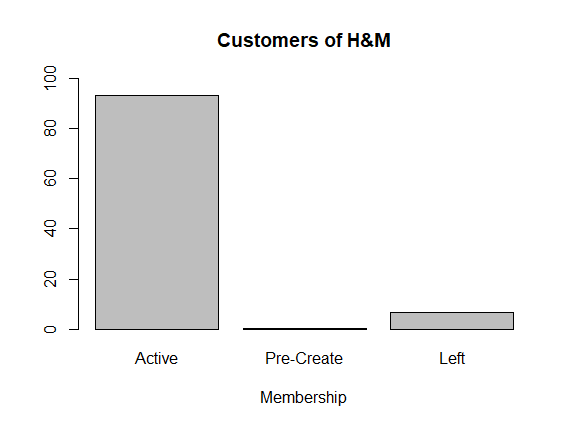
To perform this, filter function is used to get the row this have club member status as not empty.

Prop.table(table()) to get the proportion for each status

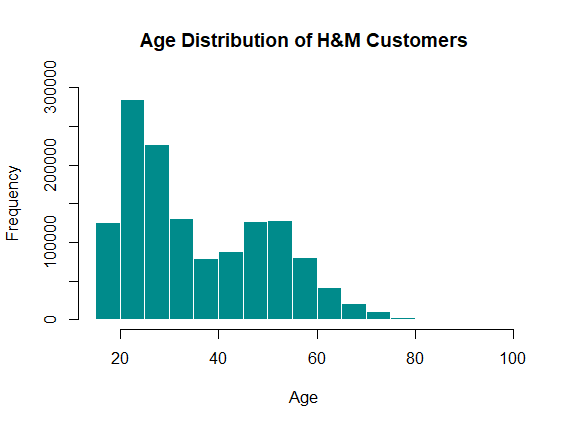


We can see Active have highest proportion (93.16%) , then left club (6.805%) and last one is Pre-create(0.34%).

We get better illustration, bar plot that been drawn to get better understand for the proportion.



* Next plot is the histogram which shows the distribution of people age purchasing product from the H&M.



We can see that people having age between 20-30 are the highest purchaser from H&M. No doubt people of age 60-80 also have a good frequency count in purchase product from H&M

* We have also calculated percentage of each generation having club Member status depending on their age group.

Age >0 & Age<=15 = Teen

Age>15 & Age <-40 = Youngters.

Age>40 & Age<=80 = Elders.





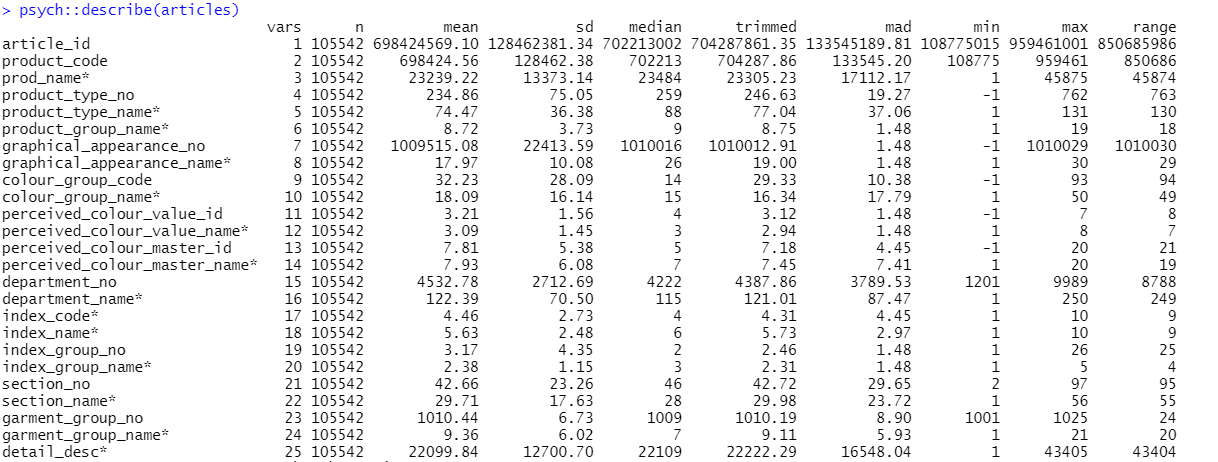
We can see Youngsters have highest proportion of Member status which is around 63%.

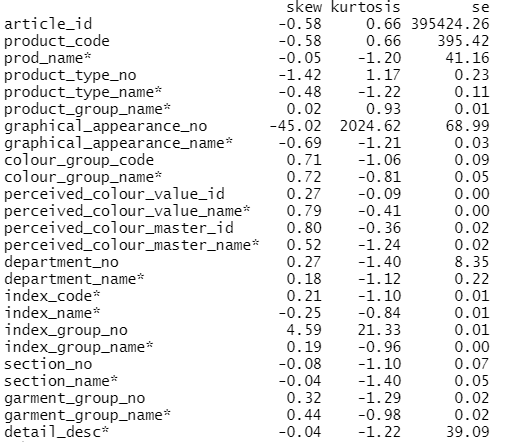
* Summary for age for club member status



Min age is 16 years and most frequent age from the data is 32 years.

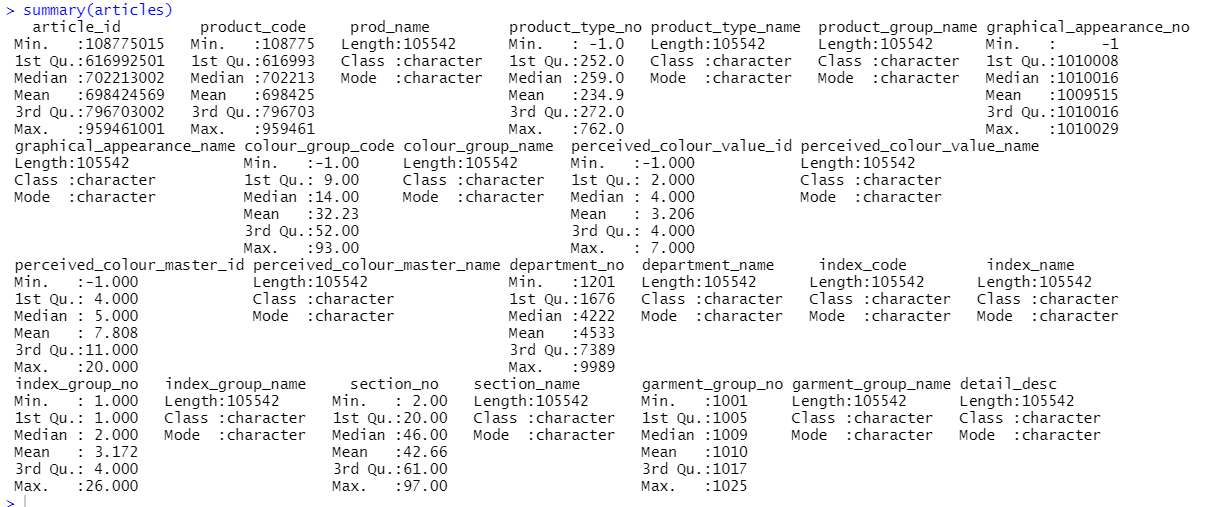
Description for the Article dataset is shown below



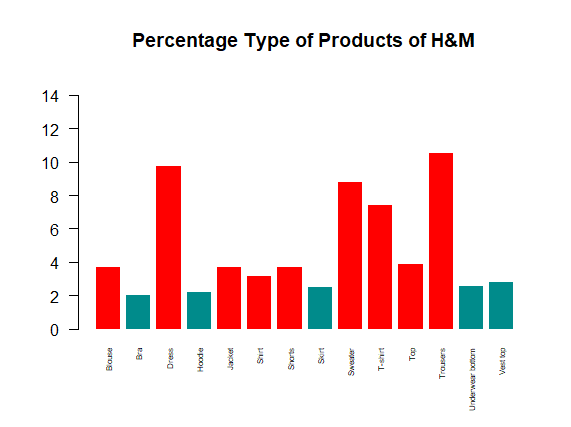


* Skewness for all variables is nearly equal to 0, which show the data is normally distributed.

Summary article dataset

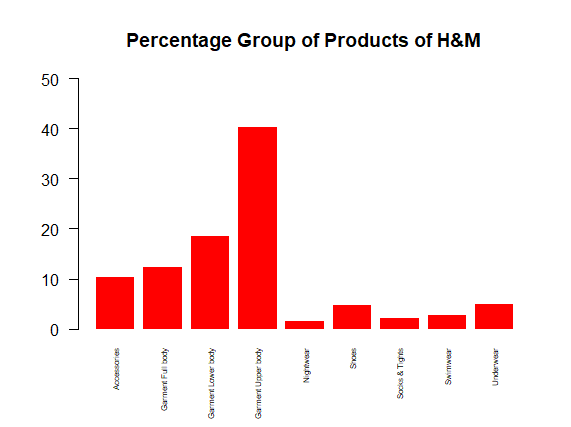


* Barplot for the Type of product of H&M has been plotted

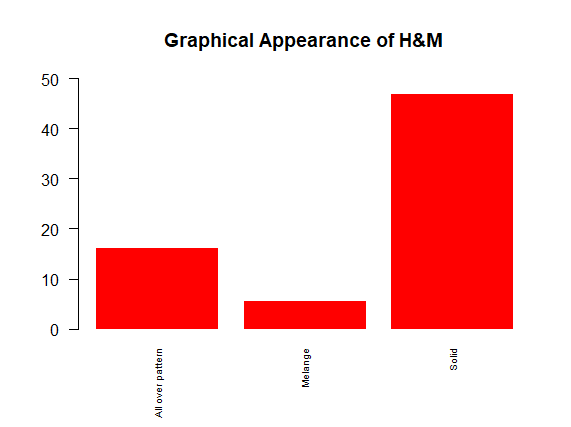


Highest percentage of product sold in Trouser which takes around 10.5% and dress is around 10% from the total proportion and least product sold is Bra which is around 2%. This is due to the more youngster purchasing product from H&M.

* Bar plot for the Group of product of H&M has been plotted

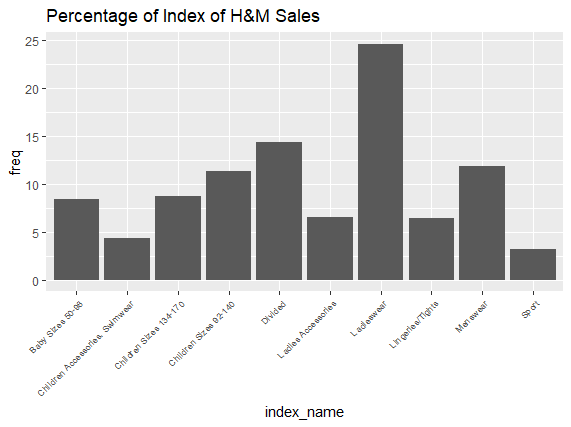


* Highest percentage is 40% which is for Garment upper body and least is nightwear which is 1%.
* Next bar plot is the percentage graphical pattern on product sold by H&M



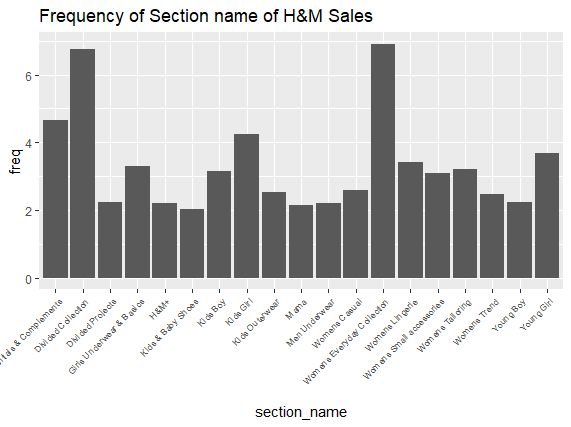
We can see most product sold by H&M have graphical pattern as solid (47%) and next is All over pattern and the last is Melange.

* GGplot has been plotted to represent the percentage of index of H&M.



The highest percentage sold according to index name is ladies wear which comprise of 25%. Sport wear and Children Accessories Swimwear is sold the least.

* GGplot has been plotted to represent the Frequency of Sales Name of H&M.



The highest frequency is for Divided Collection which is 7 and Women Everyday Collection which is 7 , rest all of them have range between 2-4.

Reference

[1] Kaggle: https://www.kaggle.com/competitions/h-and-m-personalized-fashionrecommendations/data

Appendix

setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

articles = read.csv("./articles.csv")

customers = read.csv("./customers.csv")

psych::describe(customers)

summary(customers)

unique(customers$club\_member\_status)

library(dplyr)

clubMemberStatus = customers %>% filter(club\_member\_status != "")

clubMemberStatus$club\_member\_status = as.numeric(as.factor(clubMemberStatus$club\_member\_status))

clubMemberStatus$fashion\_news\_frequency = as.numeric(as.factor(clubMemberStatus$fashion\_news\_frequency))

print(prop.table(table(clubMemberStatus$club\_member\_status))\*100)

barplot(prop.table(table(clubMemberStatus$club\_member\_status))\*100,

names.arg = c("Active","Pre-Create","Left"),

xlab="Membership",

main="Customers of H&M",

ylim=c(0,100))

hist(clubMemberStatus$age,

xlab="Age",

main="Age Distribution of H&M Customers",col='cyan4',border='white',

ylim=c(0,300000))

brackets <- clubMemberStatus %>% mutate(agegroup = case\_when(age > 0 & age <= 15 ~ 'Teen',

age > 15 & age <= 40 ~ 'Youngsters',

age > 40 & age <= 80 ~ 'Elders')) # end function

prop.table(table(brackets$agegroup)) \* 100

summary(clubMemberStatus$age)

psych::describe(articles)

summary(articles)

# PRODUCT SALES RELATIVE PLOT

salesAsType = articles %>% count(product\_type\_name) %>% mutate(freq = n / sum(n)\*100) %>% filter(freq > 2)

barplot(salesAsType$freq,names.arg=salesAsType$product\_type\_name,ylim=c(0,15),main='Percentage Type of Products of H&M',cex.names = 0.5,las=2,col=ifelse(salesAsType$freq > 3,'Red','cyan4'),border='white')

salesAsTypeGroup = articles %>% count(product\_group\_name) %>% mutate(freq = n / sum(n)\*100) %>% filter(freq > 1)

barplot(salesAsTypeGroup$freq,names.arg=salesAsTypeGroup$product\_group\_name,ylim=c(0,50),main='Percentage Group of Products of H&M',cex.names = 0.5,las=2,col='Red',border='white')

salesAsGraphics = articles %>% count(graphical\_appearance\_name) %>% mutate(freq = n / sum(n)\*100) %>% filter(freq > 5)

barplot(salesAsGraphics$freq,names.arg=salesAsGraphics$graphical\_appearance\_name,ylim=c(0,50),main='Graphical Appearance of H&M',cex.names = 0.6,las=2,col='Red',border='white')

library('ggplot2')

salesAsIndex = articles %>% count(index\_name) %>% mutate(freq = n / sum(n)\*100)

ggplot(salesAsIndex,aes(y=freq,x=index\_name))+geom\_bar(stat='identity')+theme(axis.text.x = element\_text(angle = 45,hjust=1,size=7))+labs(title='Percentage of Index of H&M Sales')

salesAsSection = articles %>% count(section\_name) %>% mutate(freq = n / sum(n)\*100) %>% filter(freq>2)

ggplot(salesAsSection,aes(y=freq,x=section\_name))+geom\_bar(stat='identity')+theme(axis.text.x = element\_text(angle = 45,hjust=1,size=7))+labs(title='Frequency of Section name of H&M Sales')