

*Toronto*

**Initial Analysis Report**

Soni Manan ***NUID****: 002982645*

Savaliya Parth ***NUID****: 002982302*

Dave Dhairya ***NUID****: 002110382*

*College of Professional Studies****,*** *Master of Professional Studies in Analytics.*

***Subject****: ALY6015*

Under the guidance of

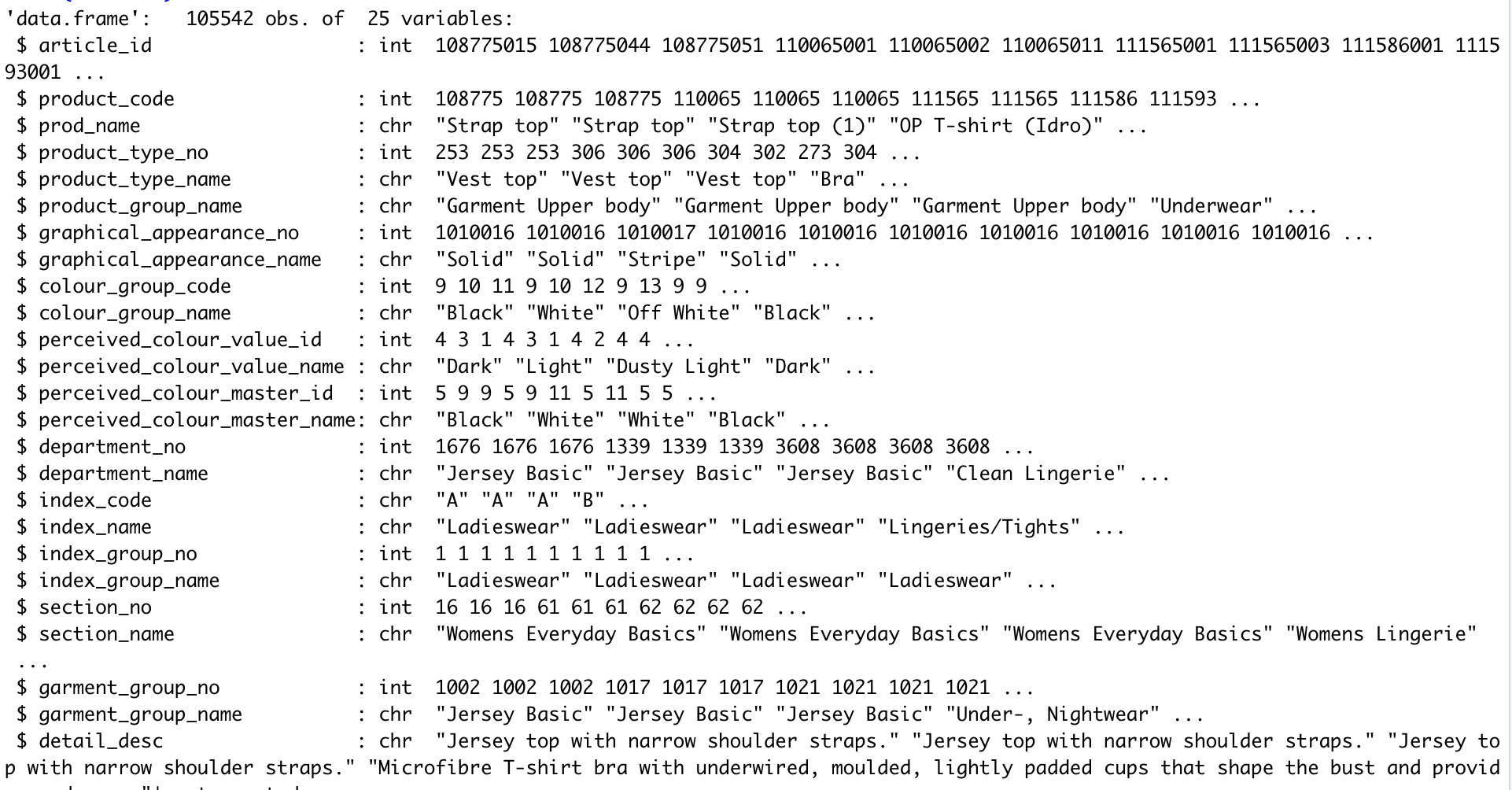
**Dr. Prof. Alex Maizlish**

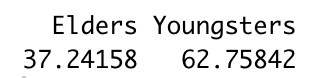
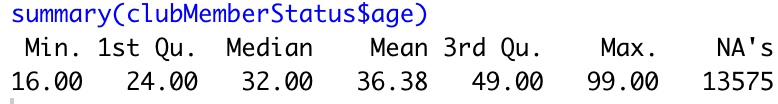
**Introduction**

Data has been released by H&M for their benefits and put as competition on Kaggle. All of the dataset contains data regarding their products and customers like clothing’s, accessories and customers membership status and their activeness. These data can be used for recommendation of their product based on sales. Initially it looked like we need to feature variables and subset them to get proper predictions.

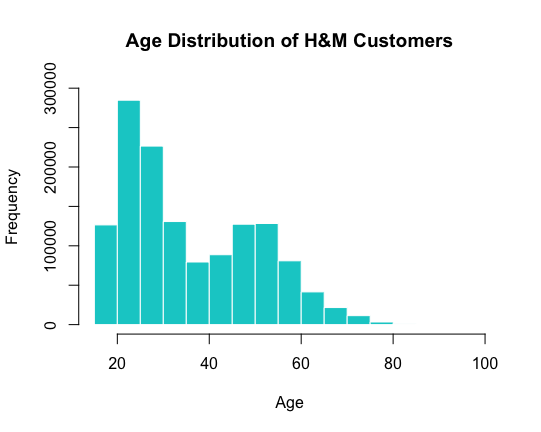
**Descriptive Analysis**

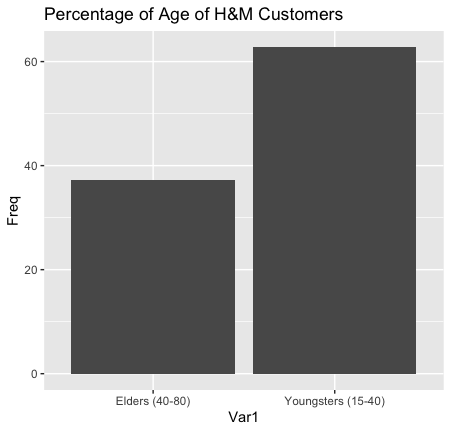
* Articles dataset has 105502 samples and 25 variables.
* Customers dataset has 1371980 samples and 7 variables in which postal and id are hashed so useless as well as many rows has NA which cant be replaced with mean and median so replaced with mode.
* Most of the values are character and others are assigned id.

****

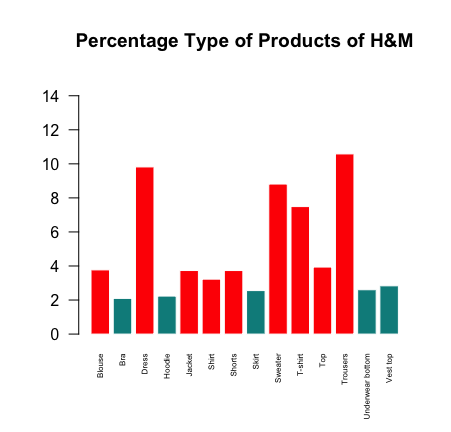
** **

* Mean age is 36.38 and we consider age 15 to 40 are youngsters who are into good fashion where 37% elders of H&M customers.
* Age distribution of H&M is not normal as we can judge like H&M is brand for upcoming fashion and provide clothing’s according to west. So most of are in the range of 0-40 and little increment between 50-70 age group.
* 90% are active who follows news for whatever reason, some have pre-created membership and few left like 7% has left membership.
* Cleaning such as removing hasn’t take place while making this analysis as it can affect.

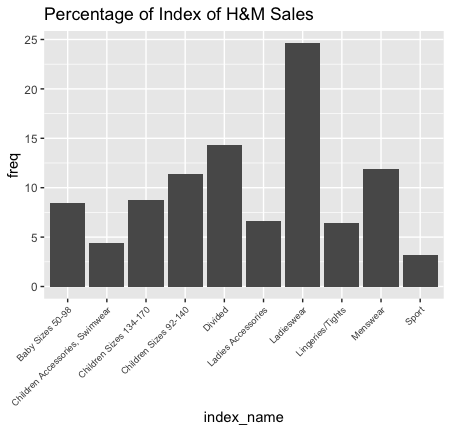
****

****

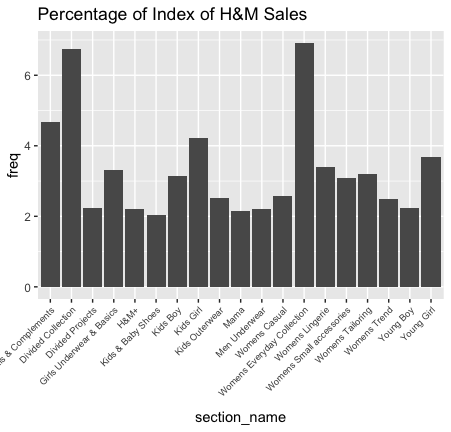
* 90% are active who follows news for whatever reason, some have pre-created membership and few left like 7% has left membership.
* As we can see, Trousers, Dresses, Sweaters, Tees and Top are top 5 whose cumulative relative frequency is almost 45%. Among that Trousers and Dresses tops with 11 and 10% respectively.

****

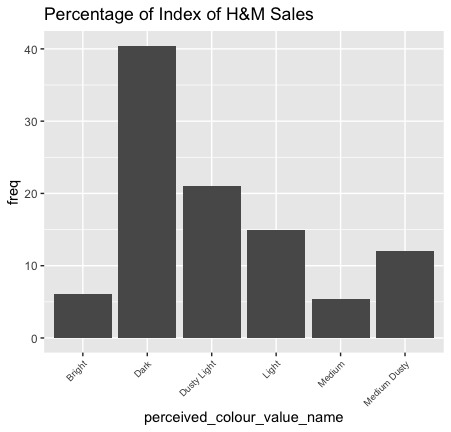
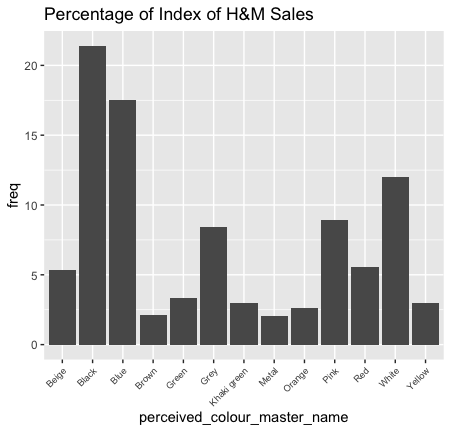
* Index name are just like category of articles.
* Ladieswear, tops with 25% in this where Divided, and Menswear are almost 15%. Following that, Childrenswear are at 10%.

****

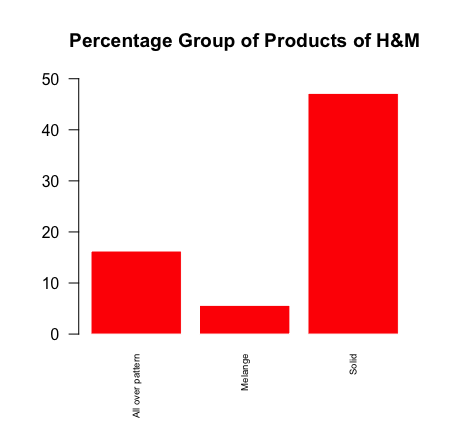
* This is uneven distribution of indexes at H&M. Following upper diagram, this follows same as that but in descriptive way. For ladieswear this has different ladieswear section like Everyday collection, Casual, etc. We can answer questions by subsetting or labelling this.

****

* This is uneven distribution of indexes at H&M. Following upper diagram, this follows same as that but in descriptive way. For ladieswear this has different ladieswear section like Everyday collection, Casual, etc. We can answer questions by subsetting or labelling this.
* This is overall colour preferences of customers. We can see Dark has preferred by 40% people. Dusty light and light has preferred by 20 and 15% customers respectively.

** **

* To be precise, in Dark, Black and Blue are highly preferred by 37%. White and Pink are second highest preference of customers.
* To be precise, in Dark, Black and Blue are highly preferred by 37%. White and Pink are second highest preference of customers.
* So people around 50% people prefers Solid over patterns and etc. So if a research question ask like what are the chances of a customer will buy a t-shirt with solid or with some kind of pattern.



**Models and Predictions**

* A Logistic regression is a type of regression model we use when the answer of variable is like yes or no.
* As we have variables which are categories weather it is menswear or ladies or is it accessory or not.
* So Multiple logistic regression, we will use to answer our research question from which H&M can get benefits.
* There can be multiple because if we need work to improve the model accuracy or there can be multiple independent variables.

**Conclusion**

From this we can learn how to make a logistic regression model for binary solutions. How to calculate miscalculation (False Negatives and Positives) by model.

Finding the accuracy and Performance by finding confusion matrix, ROC and AUC.

For the dataset, From the Applications, Accept, Fulltime Grads, Outstates and PhD are the variables who leading the model to 98%. As I mentioned, correlation matrix plays important role just like EDA and feature modelling.

Accuracy and AUC is just higher than 90-95% are consider as overfitting model however for qualified data it is possible.

Most of target audience should be youngsters and less elders as number of sales accounted for western styles like trousers, tops, t-shirts and etc.

**Next Milestone**

1) Next step should be getting confidence using hypothesis testing before going for the model.

Getting confidence is important. Most of data is categorical so we can go with Chi-Square test.

Statically confident is something we need to get as to please the authorities how strongly we are suggesting just instead telling.

2) Splitting the data between train and test and then use GLM to train model and predict using that.

Detect and remove outliers from process. Also identify the outliers using AIC, Cooks, QQ plots.

Getting ROC, AUC, Recall, Precision and Accuracy for the model.

**References**

H&M Personalized Fashion Recommendations, Provide product recommendations based on previous purchases, H&M Group *Sources*: https://www.kaggle.com/competitions/h-and-m-personalized-fashion- recommendations/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)

Mode <- function(x) {

ux <- unique(x)

ux[which.max(tabulate(match(x, ux)))]

}

customers[customers$club\_member\_status == "",] = Mode(customers$club\_member\_status)

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",

col=ifelse(prop.table(table(clubMemberStatus$club\_member\_status))\*100 > 80,'cyan3','grey'),

border=ifelse(prop.table(table(clubMemberStatus$club\_member\_status))\*100 <5,'gray','white'),

main="Customers of H&M",

ylim=c(0,100))

hist(clubMemberStatus$age,

xlab="Age",

main="Age Distribution of H&M Customers",col='cyan3',border='white',

ylim=c(0,300000))

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

age > 15 & age <= 40 ~ 'Youngsters (15-40)',

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

age\_brackets = as.data.frame(prop.table(table(brackets$agegroup)) \* 100)

ggplot(age\_brackets,aes(x=Var1,y=Freq))+geom\_bar(stat='identity') +labs(title='Percentage of Age of H&M Customers')

summary(clubMemberStatus$age)

psych::describe(articles)

str(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(sales$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='Percentage Group of Products 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='Percentage of Index of H&M Sales')

salesAsColor = articles %>% count(perceived\_colour\_value\_name) %>% mutate(freq = n / sum(n)\*100) %>% filter(freq>2)

ggplot(salesAsColor,aes(y=freq,x=perceived\_colour\_value\_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')

salesAsColorName = articles %>% count(perceived\_colour\_master\_name) %>% mutate(freq = n / sum(n)\*100) %>% filter(freq>2)

ggplot(salesAsColorName,aes(y=freq,x=perceived\_colour\_master\_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')

featureGraphics = as.data.frame(articles$graphical\_appearance\_name)

featureGraphics$Solid = ifelse(featureGraphics$`articles$graphical\_appearance\_name` =='Solid',1,2)

colnames(featureGraphics) = c("graphical\_name",'solid')

featureGraphics$graphical\_name = as.numeric(as.factor(featureGraphics$graphical\_name))

library(misclassGLM)

mIndex = sample(c(1,2), nrow(featureGraphics),

replace = T,

prob = c(0.7,0.3))

train\_x = featureGraphics[mIndex == 1,]

test\_x = featureGraphics[mIndex == 2,]

head(train\_x)

LR\_model <- glm(solid ~ graphical\_name,

data = train\_x,

family = binomial(link = "logit"))

summary(LR\_model)

prob.train\_x = predict(LR\_model,

newdata = train\_x,

type = "response")

cm\_data = as.factor(ifelse

(prob.train\_x >= 0.5,

1, 2))

library(caret)

confusionMatrix(cm\_data,

as.factor(ifelse(train\_x$solid == 1, 1,2)))

prob.test\_x = predict(LR\_model,

newdata = test\_x,

type = "response")

prob.test\_x

cm\_data = as.factor(ifelse

(prob.test\_x >= 0.5,

1, 2))

cm\_data

head(cm\_data)

confusionMatrix(cm\_data,

as.factor(ifelse(test\_x$solid == 1, 1,2)),

)

library(pROC)

ROC = roc (test\_x$solid, prob.test\_x)

X = plot(ROC,

col = "black",

ylab = "Sensitivity = TP rate",

xlab = 'specificity = FP rate')

#08 Calculate and interpret the AUC.

AUC = auc(ROC)

AUC