Toronto

Final Project

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.

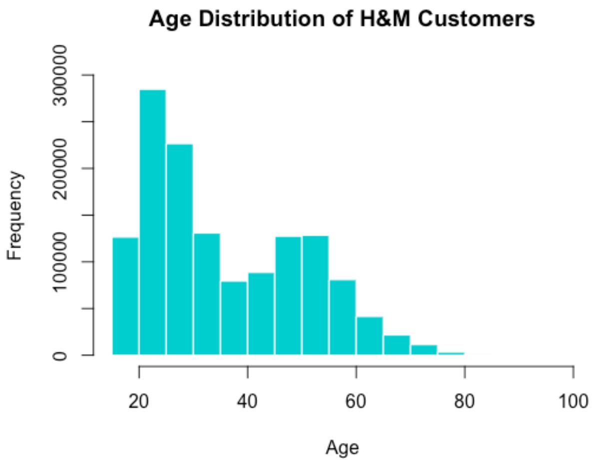
```
'data.frame':   105542 obs. of  25 variables:
 $ article_id                  : int  108775015 108775044 108775051 110065001 110065002 110065011 111565001 111565003 111586001 1115
93001 ...
 $ product_code                : int  108775 108775 108775 110065 110065 110065 111565 111565 111586 111593 ...
 $ prod_name                   : chr  "Strap top" "Strap top" "Strap top (1)" "OP T-shirt (Idro)" ...
 $ product_type_no             : int  253 253 253 306 306 306 304 302 273 304 ...
 $ product_type_name           : chr  "Vest top" "Vest top" "Vest top" "Bra" ...
 $ product_group_name          : chr  "Garment Upper body" "Garment Upper body" "Garment Upper body" "Underwear" ...
 $ graphical_appearance_no     : int  1010016 1010016 1010017 1010016 1010016 1010016 1010016 1010016 1010016 1010016 ...
 $ graphical_appearance_name   : chr  "Solid" "Solid" "Stripe" "Solid" ...
 $ colour_group_code           : int  9 10 11 9 10 12 9 13 9 9 ...
 $ colour_group_name           : chr  "Black" "White" "Off White" "Black" ...
 $ perceived_colour_value_id   : int  4 3 1 4 3 1 4 2 4 4 ...
 $ perceived_colour_value_name : chr  "Dark" "Light" "Dusty Light" "Dark" ...
 $ perceived_colour_master_id  : int  5 9 9 5 9 11 5 11 5 5 ...
 $ perceived_colour_master_name: chr  "Black" "White" "White" "Black" ...
 $ department_no               : int  1676 1676 1676 1339 1339 1339 3608 3608 3608 3608 ...
 $ department_name             : chr  "Jersey Basic" "Jersey Basic" "Jersey Basic" "Clean Lingerie" ...
 $ index_code                  : chr  "A" "A" "A" "B" ...
 $ index_name                  : chr  "Ladieswear" "Ladieswear" "Ladieswear" "Lingeries/Tights" ...
 $ index_group_no              : int  1 1 1 1 1 1 1 1 1 1 ...
 $ index_group_name            : chr  "Ladieswear" "Ladieswear" "Ladieswear" "Ladieswear" ...
 $ section_no                  : int  16 16 16 61 61 61 62 62 62 62 ...
 $ section_name                : chr  "Womens Everyday Basics" "Womens Everyday Basics" "Womens Everyday Basics" "Womens Lingerie"
 ...
 $ garment_group_no            : int  1002 1002 1002 1017 1017 1017 1021 1021 1021 1021 ...
 $ garment_group_name          : chr  "Jersey Basic" "Jersey Basic" "Jersey Basic" "Under-, Nightwear" ...
 $ detail_desc                 : chr  "Jersey top with narrow shoulder straps." "Jersey top with narrow shoulder straps." "Jersey to
p with narrow shoulder straps." "Microfibre T-shirt bra with underwired, moulded, lightly padded cups that shape the bust and provid
```

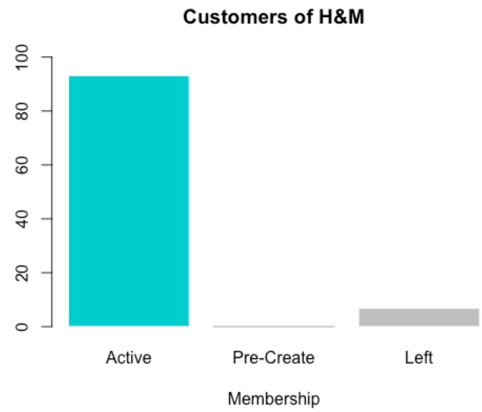
```
   Elders Youngsters
 37.24158   62.75842
```

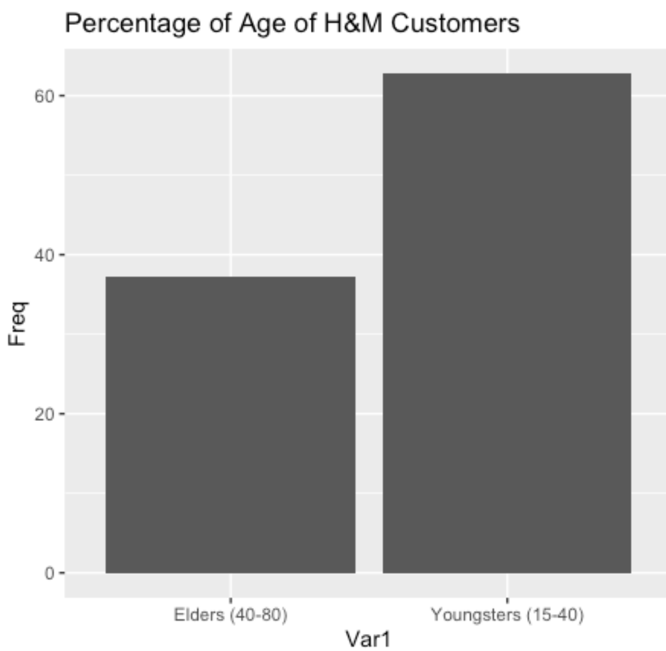
```
summary(clubMemberStatus$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  16.00   24.00   32.00   36.38   49.00   99.00   13575
```

- Mean age is 37 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.

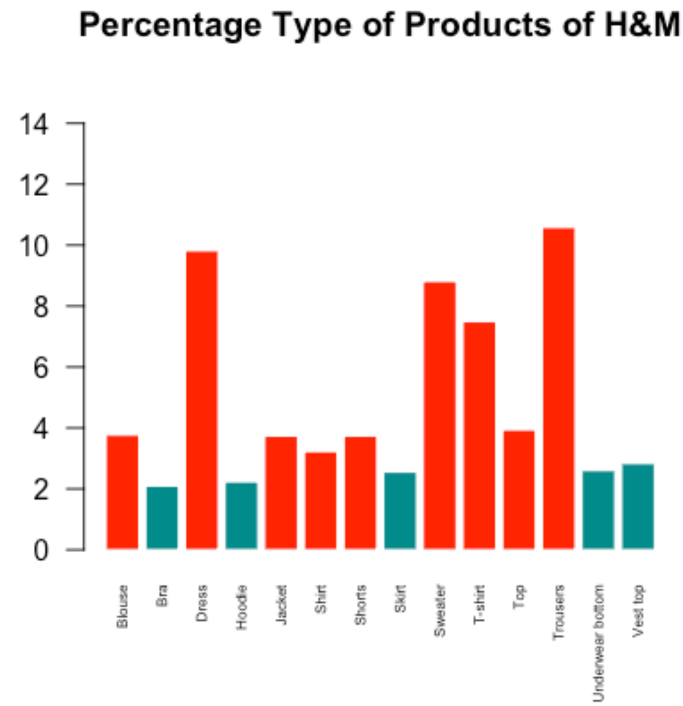
- Cleaning such as removing hasn't take place while making this analysis as it can affect.



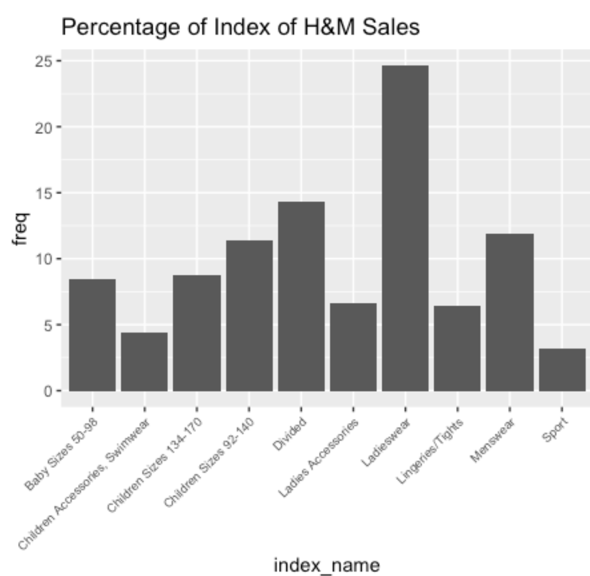




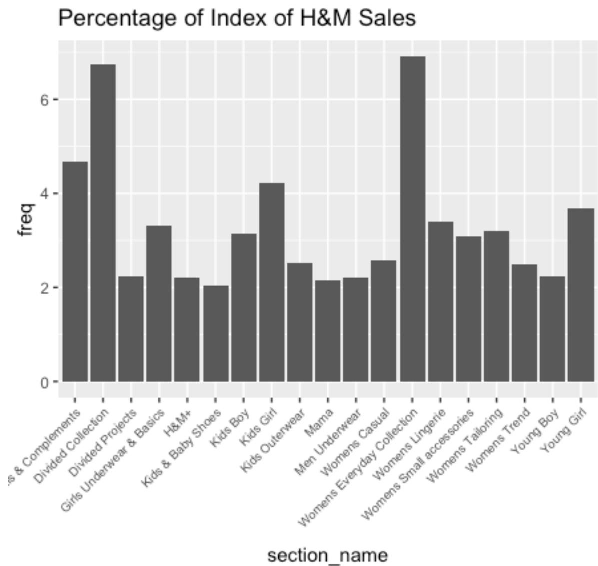
- 90% are active who follows news for whatever reason, some have pre-created membership and few left like 7% has left membership.
- As we can see, Trousers, Dresses, Sweaters, Tees and Top are top 5 whose cumulative relative frequency is almost 45%. Among that Trousers and Dresses tops with 11 and 10%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

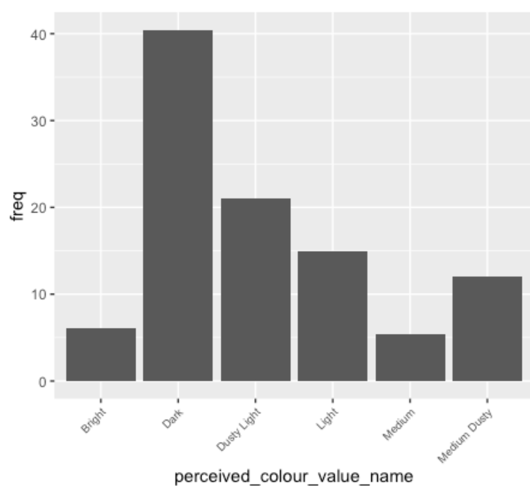


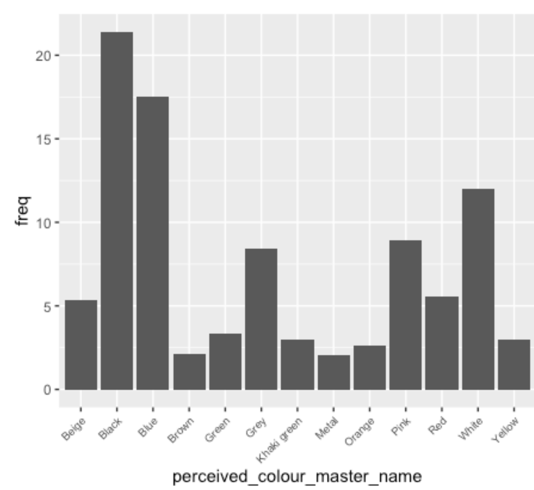
- This is uneven distribution of indexes at H&M. Following upper diagram, this follows same as that but in descriptive way. For ladieswear this has different ladieswear section like Everyday collection, Casual, etc. We can answer questions by subsetting or labelling this.



- 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.

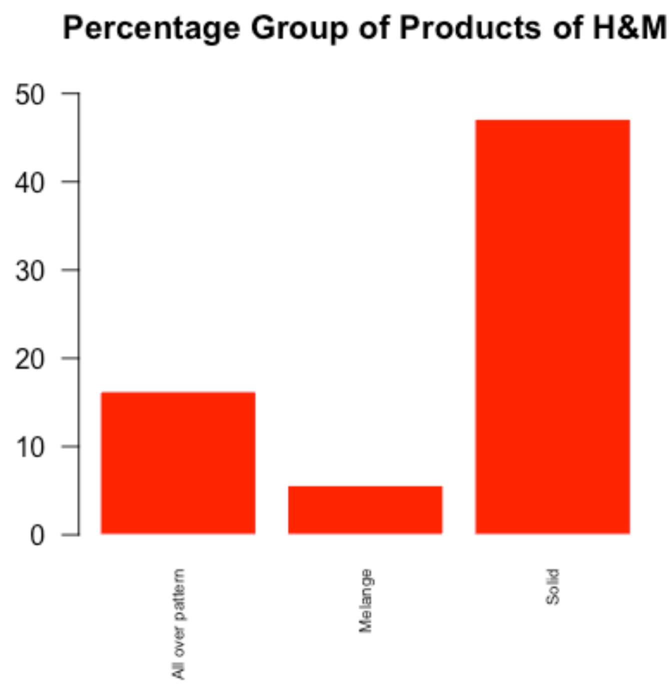




- To be precise, in Dark, Black and Blue are highly preferred by 37%. White and Pink are second highest preference of customers.

- 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.



Hypothesis Testing

Hypothesis testing is important as we need to be sure and gain some confidence about our variables. There many methods to test and we tried ANOVA and chi-square method due to all of the variables are categorical. Also used ANOVA there is a continuous variable in customers dataset which is age. To be specific, chi-square is where all the variables which are going to test are categorical.

Customer Hypothesis

Statement: We will test and gain statistical results if membership of H&M has some relationships with Age and Fashion news frequency.

Null Hypothesis (H_0): There is **no impact of interaction** between **fashion news and age**.

Alternative Hypothesis (H_1): There is **impact** of interaction between fashion news frequency and age.

Null Hypothesis (H_0): There is **no effect of fashion news** on the **club membership**.

Alternative Hypothesis (H_1): There is **a little effect of fashion** news on the club membership.

Null Hypothesis (H_0): There is **no effect of age** on the **club membership**.

Alternative Hypothesis (H_1): There is **some effect** of age on the club membership.

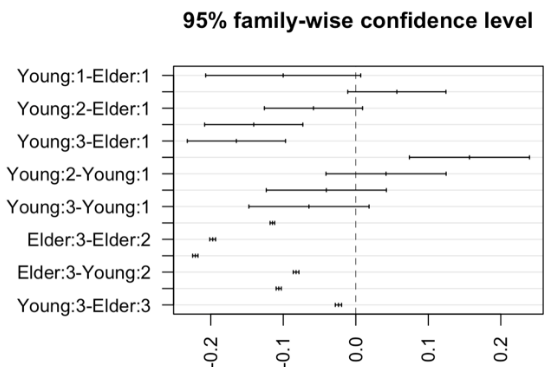
Summary: For this test the level of significance is 95% which means value of alpha for the test is 0.05 to test.

We will use 2-factor ANOVA. For that, all of the variables are independent of each other at initial.

From this test we can see that p-value of individual variables are almost 0 and for interaction of age and fashion news frequency is also 0. In nutshell, All of the variables are less than value of alpha which is 0.05 as mention earlier.

From Tukey interpretation, we can say is Youngs whom are less than 30 having no news frequency and Youngs with monthly news frequency's interaction is greater than significance. As X axis is showing difference in mean.

```
                                 Df Sum Sq Mean Sq F value Pr(>F)    
age                               1   2014    2014    8754 <2e-16 ***
fashion_news_frequency            1   7466    7466   32446 <2e-16 ***
age:fashion_news_frequency        1    639     639    2777 <2e-16 ***
Residuals                   1342320 308863       0                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Impact of Graphic patterns and Product type on Sales of H&M

Statement: In this, testing and gaining statistical confidence if graphics patterns on product and product type do have impact on sales of such product.

Null Hypothesis (H_0): Graphic patterns and Type of Articles do not have any impact on purchases.

Alternative Hypothesis (H_1): Graphic patterns and Type of Articles have impact on purchases.

Summary: For this test the level of significance is 95% which means value of alpha for the test is 0.05 to test.

We will use Chi-squared. Due to, all of the variables are categorical and having same number of samples. Note that the value of p is almost 0 which is less than the value of significance 0.05 and **χ2** statistics is **109893** which is far greater than **3913.956** which is critical value of test.

From this test we can see that p-value is 0 which is less than 0.05 so that we have enough evidence to reject the null hypothesis and conclude that there is some effect of graphic patterns and type of product on sale.

```
        Pearson's Chi-squared test

data:  table(articles$product_type_name, articles$graphical_appearance_name)
X-squared = 109893, df = 3770, p-value < 2.2e-16

> qchisq(p=0.05,df=3770,lower.tail = F)
[1] 3913.956
```

Impact of Colours and Product type on Sales of H&M

Statement: We will test for, if colours of product and product type do have effects on numbers of sales.

Null Hypothesis (H_0): Colours of product such as dark, light and Type of Articles (product) do not have any impact on sales made by clients of H&M.

Alternative Hypothesis (H_1): They have impact on purchases.

Summary: We will use again Chi-squared again for this test. Note that the value of p is almost 0 which is less than the value of significance 0.05 and **χ2** statistics is **80176** which is far greater than **2586.735** which is critical value of test.

```
        Pearson's Chi-squared test

data:  table(articles$product_type_name, articles$perceived_colour_master_name)
X-squared = 80176, df = 2470, p-value < 2.2e-16

> qchisq(p=0.05,df=2470,lower.tail = F)
[1] 2586.735
```

Model and Prediction

From the data, all of the variables are almost categorical such as T-shirts, its section, its colour, its pattern. So most of the variables are categorical not discrete and non-continues. So we think of a algorithm taught in this course is Logistic Regression. We will use Binomial and reason will be mention further.

Logistic Regression (GLM): In this test, there are multiple samples of different graphic patterns of articles in which we focused on Solid pattern to predict. If a random article whether or not having a solid pattern instead other one.

So made a dummy variable named solid and those samples whom have solid pattern it will denoted as 1 and others as 0. So we will perform binomial Logistic regression for checking between 1 & 0. So in simple, model will test if a article has solid pattern or not. Input variables will be graphical pattern and its colour.

First we split dataset into train and testing sets which had threshold of 0.7.

For the training, we put solid (0,1) variable as independent and other as dependent. From this we can say P value is 0 for all the variables. Accuracy is 94% and false positives are 4617. But this is for training set so we need to test this model with testing set. To add to this, 34644 are true positives and 34516 are true negatives, surprisingly false negatives are 0.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 34516     0
         1  4643 34644

               Accuracy : 0.9371
                 95% CI : (0.9353, 0.9388)
    No Information Rate : 0.5306
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8747

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8814
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.8818
             Prevalence : 0.5306
         Detection Rate : 0.4677
   Detection Prevalence : 0.4677
      Balanced Accuracy : 0.9407

       'Positive' Class : 0
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.10792  -0.00853  -0.00028   0.55762   0.60862

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -17.577421   0.242790  -72.40  < 2e-16 ***
graphical_name  0.747056   0.009487   78.75  < 2e-16 ***
colour         -0.012690   0.002151   -5.90 3.64e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 102465  on 74073  degrees of freedom
Residual deviance:  37020  on 74071  degrees of freedom
AIC: 37026
```

This is for Testing Variables now and accuracy is almost same as training. And looking at testing results we have 2030 are false positives.

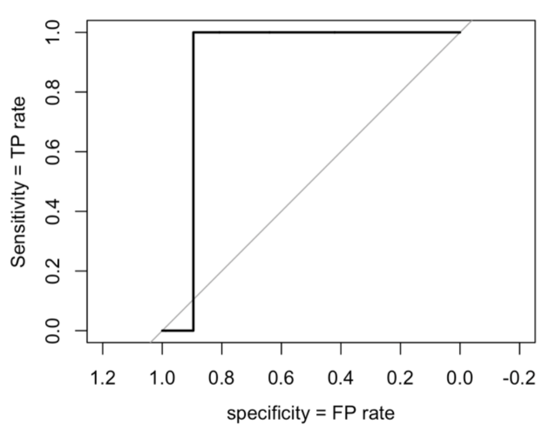
Sensitivity is 88% which tells us that how capable our model is to tell for the true positives.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 14606     0
         1  2030 15103

               Accuracy : 0.936
                 95% CI : (0.9333, 0.9387)
    No Information Rate : 0.5242
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8726

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8780
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.8815
             Prevalence : 0.5242
         Detection Rate : 0.4602
   Detection Prevalence : 0.4602
      Balanced Accuracy : 0.9390

       'Positive' Class : 0
```

ROC & AUC

ROC (Receiver Operating Characteristics) and AUC is Area under curve. Basically, These are the measures to gain the performance of a classification model. And For our model the AUC is 0.8955 which is almost 90%.

AUC tells how much the model is capable of determining between classes. The higher this value, the better model Is to tell if yes and no.




```
> AUC
Area under the curve: 0.8955
```

Conclusion

The model we developed and later used to predict, helped us to figure out whether the article (product) picked up by any customer is solid or not we can classify and also we have performed logistic regression for that which has accuracy of 94% on training set and to add to that same model was able to achieve 93.67% in our testing set.

The ROC & AUC helped us to check how much specific our model is capable of a given samples(customer picking up a specific category of t-shirt). We were able to get AUC value equal to 0.8955 which is 89%. The closer the value to 1, better the model performance and truthiness

Above mentioned tasks helps us to understand the analysis and interpretation of the dataset as hypothesis testing gives us gain confidence with proper proof.

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))
```

```
#1
v1 = select(salesAsIndex,-c(freq))
v2 =  select(salesAsColor,-c(freq))
colnames(v1) = c('Name','Sales')
colnames(v2) = c('Name','Sales')
mValues = rbind(v1,v2)
mValues$Name<- as.numeric(as.factor(mValues$Name))
s.anova<- aov(Name~Sales, data=mValues)
smmry<-summary(s.anova)
smmry

tCust <- tCust %>% mutate(ageGroup = case_when(age >= 50  & age <= 100 ~ 'Aged',
                                             age >= 30  & age <= 50 ~ 'Elder',
                                              age <= 30 ~ 'Young'))
tukey    =    TukeyHSD(aov(club_member_status    ~    ageGroup    *    fashion_news_frequency    ,
data=tCust),conf.level = 0.95)
par(mar=c(6,8,3,2))
plot(tukey,las=2)
tukey

#2
result <- chisq.test(table(articles$product_type_name,articles$perceived_colour_master_name))
result

#3
result <- chisq.test(table(articles$product_type_name, articles$graphical_appearance_name))
result

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
```