Northeastern University

H&M Fashion Recommendation

Guided By: Dr. Alex Maizlish
Project By: Dhairya Dave
Manan Soni
Parth Savaliya

Table Content

- Dataset and Source
- Objectives
- Data cleaning
- Analyzing data and its snippets.
- Hypothesis testing
- Model and prediction
- Conclusion
- Reference

Dataset and Source

- This dataset is about product recommendation of H&M and Source of this data is Kaggle. This was officially released by H&M for competition.
- It have 105542 rows and 25 variables. i.e, product, sections, description, color, etc.
- Another dataset related to this is customers with 1.3M samples and 4 variables.
- what we can do to H&M using this data is motive.
- Most of the data is category expect the age in customers.

Objectives & Aim

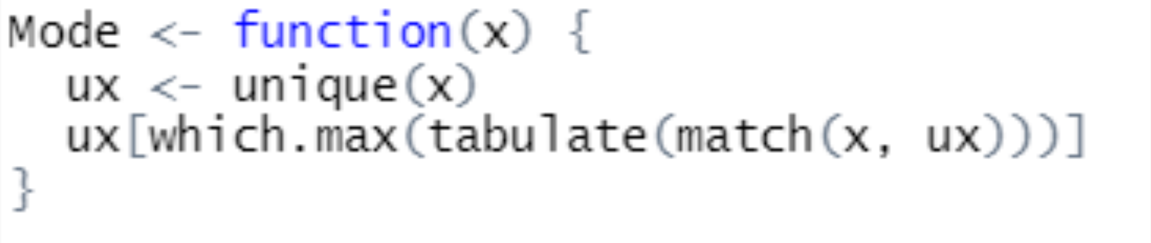
1. Distribution of Customers and membership
2. Knowledge of sales/articles of H&M
3. Which are the most purchased product or from sections.
4. Graphical appearance on products and their percentages to the sales.

AIMs

1. Which model we can used on our featured variables and why?
2. Determination of customers & membership with respected age.
3. Determination of customer purchases with respected types of product and graphical appearance.
4. Determination of customer purchases of product with respected color and type.
5. We are focusing that what people were tried to purchased and based on that tried to predict and gain confidence.

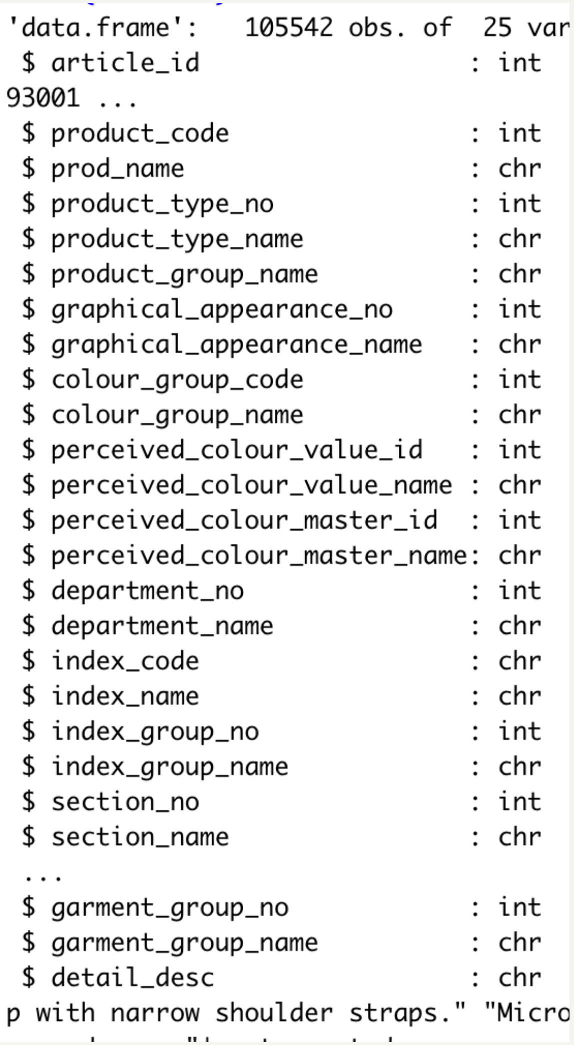
Data Cleaning

- Dataset of articles was cleaned but not the customer.
- For Customer, we used **MODE**.



```
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
```

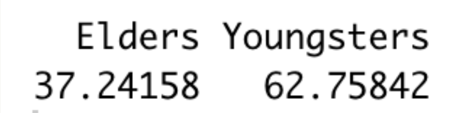
- To process we used to separate and labelling at most.



```
'data.frame':   105542 obs. of  25 var
 $ article_id                  : int
93001 ...
 $ product_code                : int
 $ prod_name                   : chr
 $ product_type_no             : int
 $ product_type_name           : chr
 $ product_group_name          : chr
 $ graphical_appearance_no     : int
 $ graphical_appearance_name   : chr
 $ colour_group_code           : int
 $ colour_group_name           : chr
 $ perceived_colour_value_id   : int
 $ perceived_colour_value_name : chr
 $ perceived_colour_master_id  : int
 $ perceived_colour_master_name: chr
 $ department_no               : int
 $ department_name             : chr
 $ index_code                  : chr
 $ index_name                  : chr
 $ index_group_no              : int
 $ index_group_name            : chr
 $ section_no                  : int
 $ section_name                : chr
 ...
 $ garment_group_no            : int
 $ garment_group_name          : chr
 $ detail_desc                 : chr
p with narrow shoulder straps." "Micro
```

- As we can see that all of the variable is categorical with their assigned codes.

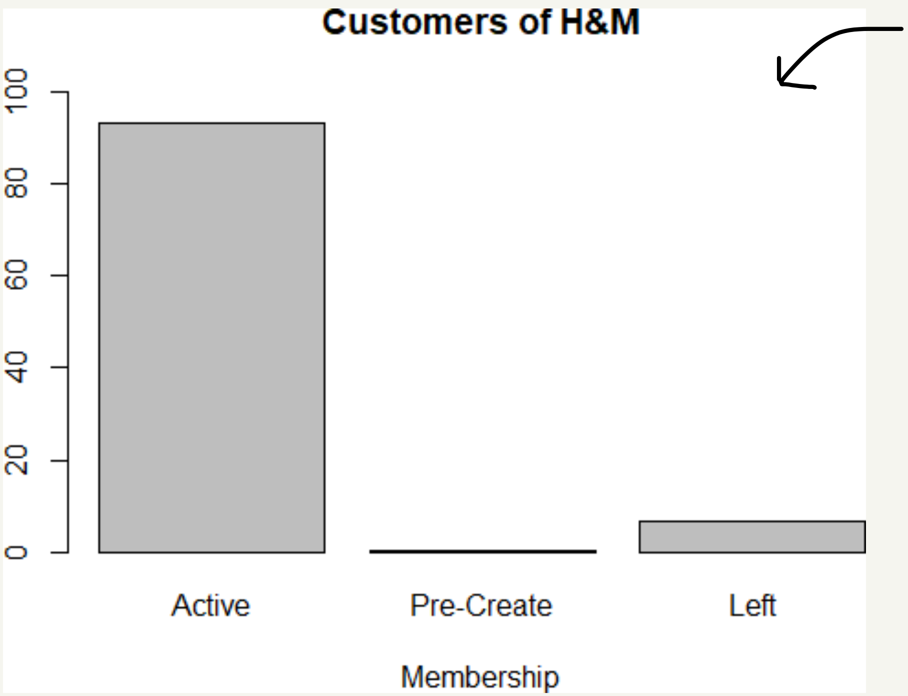
- There are 62.75% and they are in age bracket 15-40.
- 37.24% are categorized to those who are greater than 40.



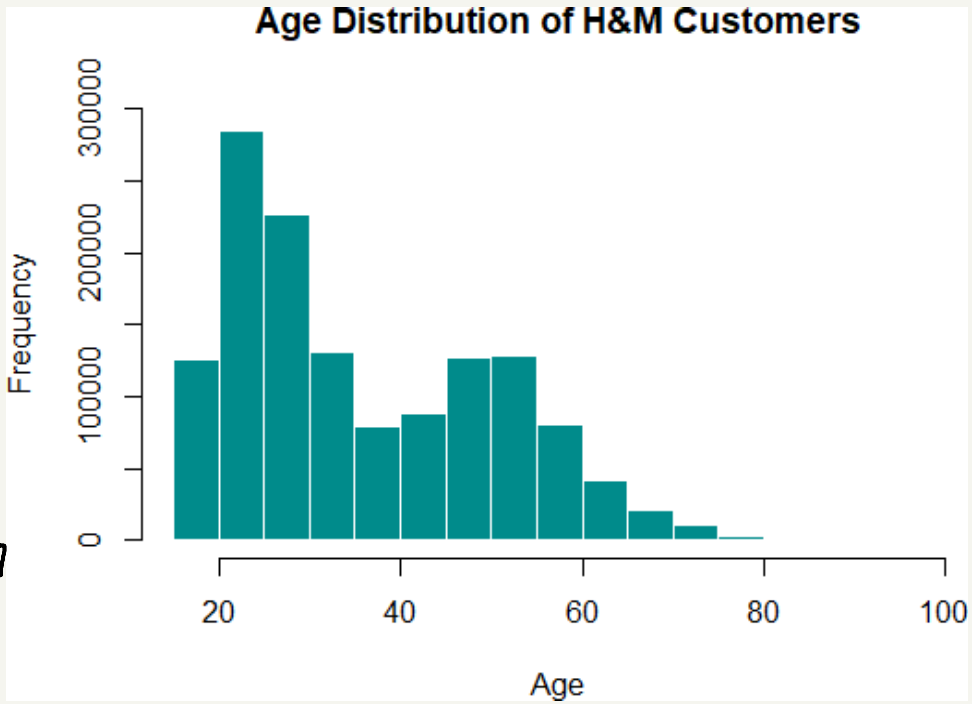
```
   Elders Youngsters
 37.24158   62.75842
```

- To note there are almost no samples in which customers of age 15 and less than that are exist.

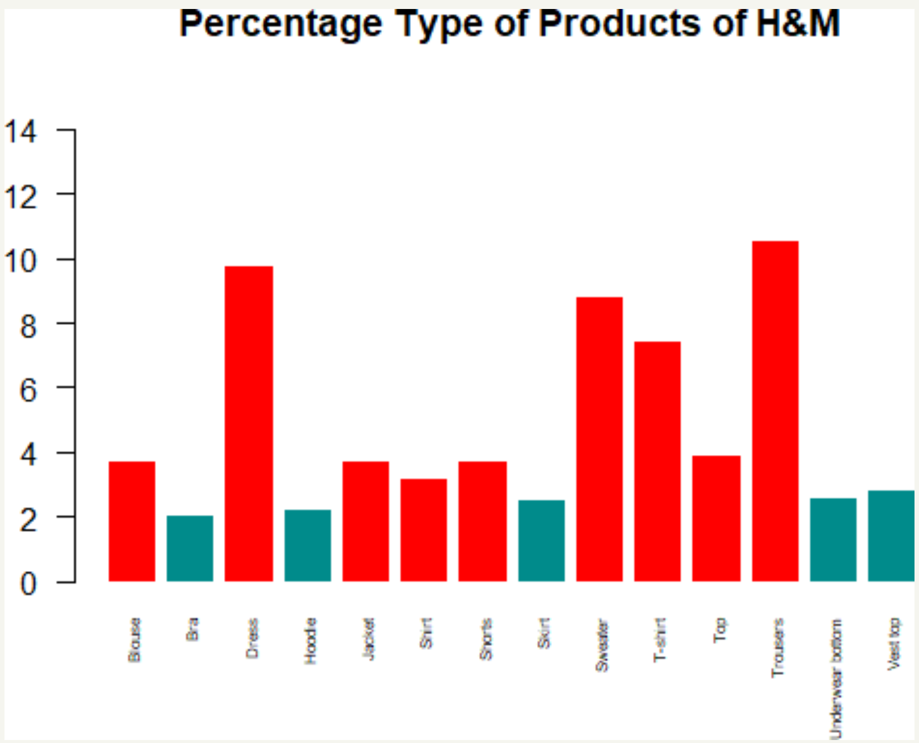
Analyzing data & Exploration



- Distribution of Customer and their status.
- 90% customers are active.

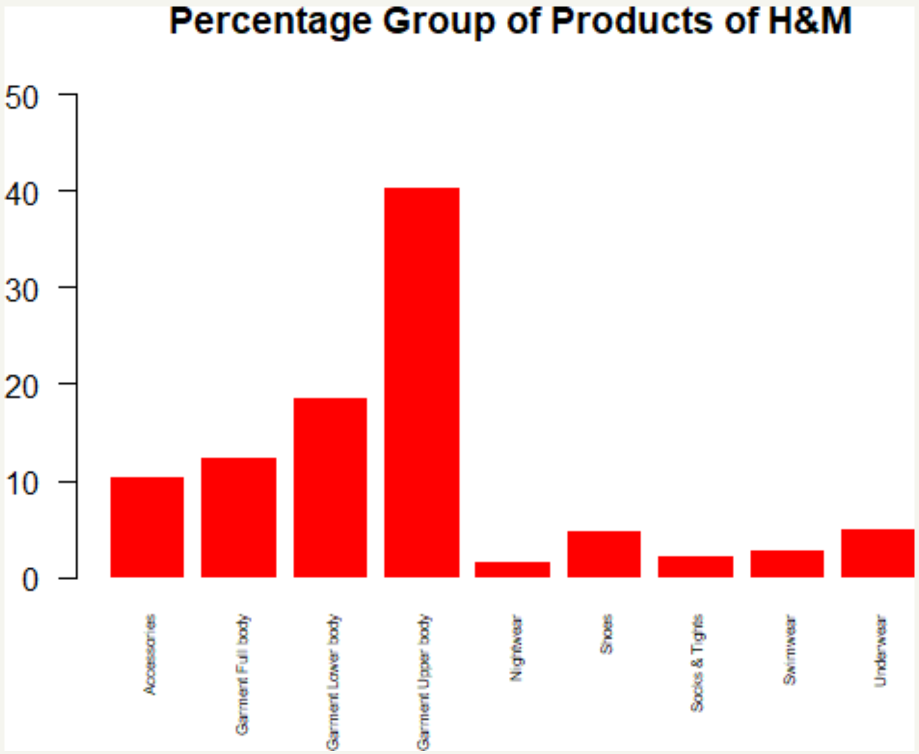


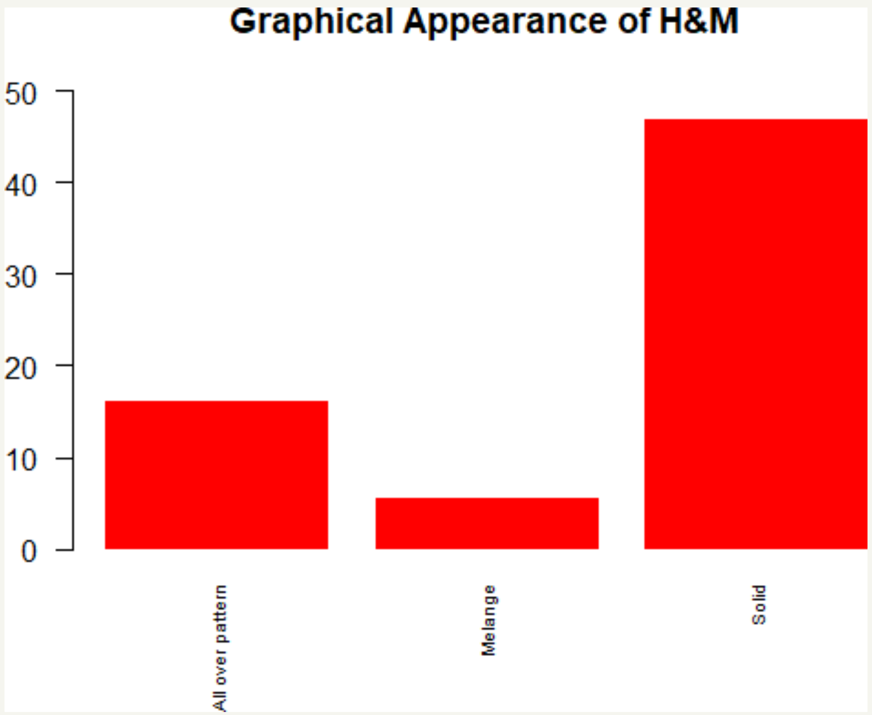
- Distribution of **AGE**.
- Mean age is 37 for 1.3M samples.



- Preferences of customers according to type of products.

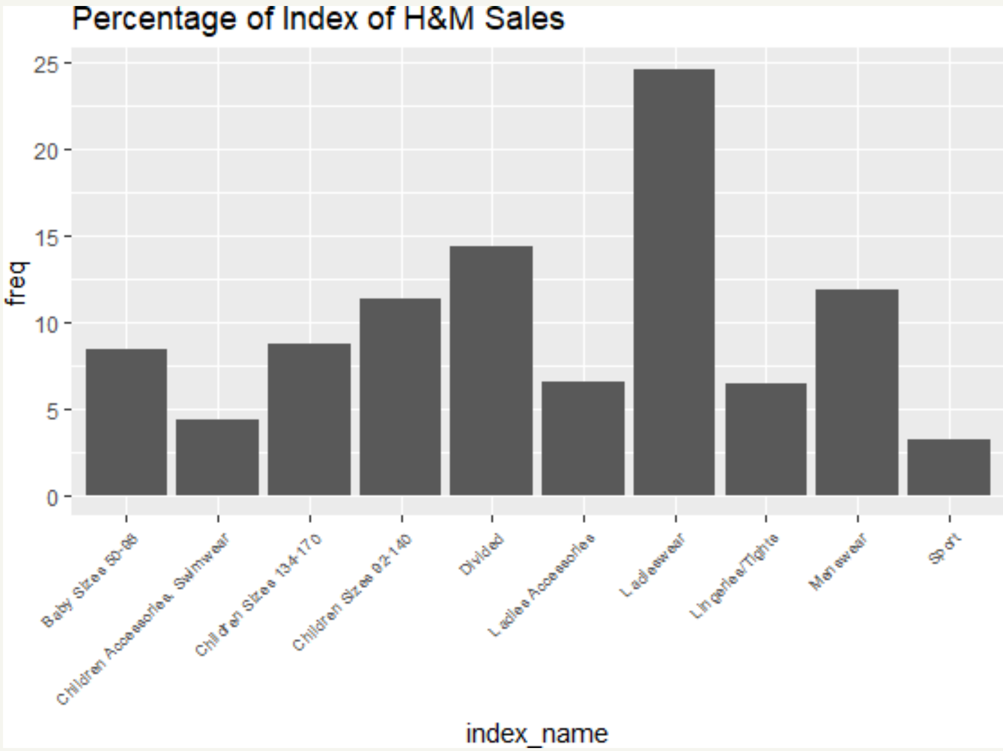
- Frequency of Group of product according to sales

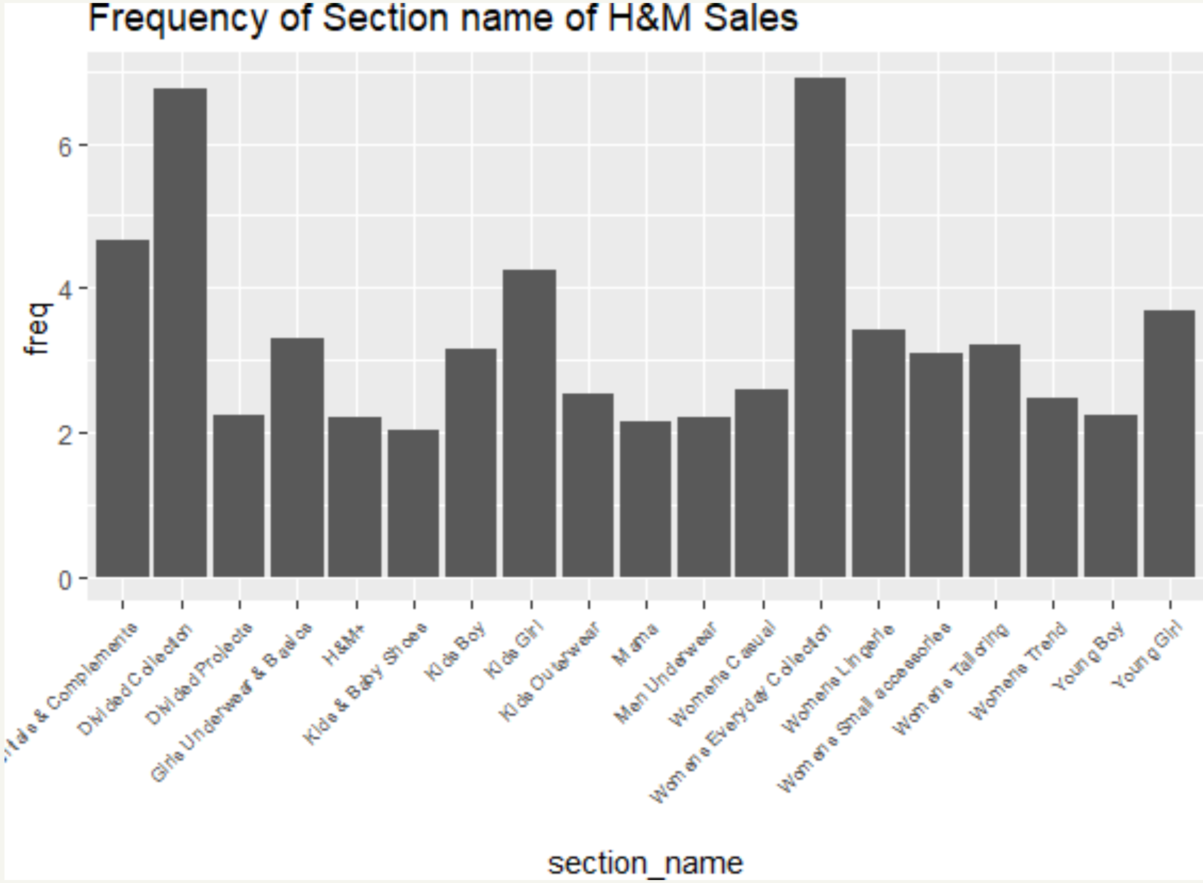




- Distribution of products according to graphical patterns over product.

- Relative frequency of Categorical distribution of products.





- Frequency according to Sections.

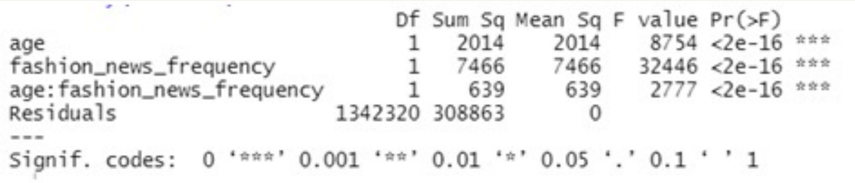
Hypothesis Testing and Confidence

- **Statement 1:** How has fashion news and age influenced the purchase in H&M store.
- **Statement 2:** How has graphics and product type has impacted the sales.
- **Statement 3:** How has colours and product type affected the sale in H&M

Analysing Statement 1

How has fashion news and age influenced the club membership in H&M store.

- **H0**: There is **no impact of interaction** between **fashion news and age**.
- **H1**: There is **impact** of interaction between fashion news frequency and age.
- **H0**: There is **no effect of fashion news** on the **club membership**.
- **H1**: There is **a little effect of fashion** news on the club membership.
- **H0**: There is **no effect of age** on the **club membership**.
- **H1**: There is **some effect** of age on the club membership.

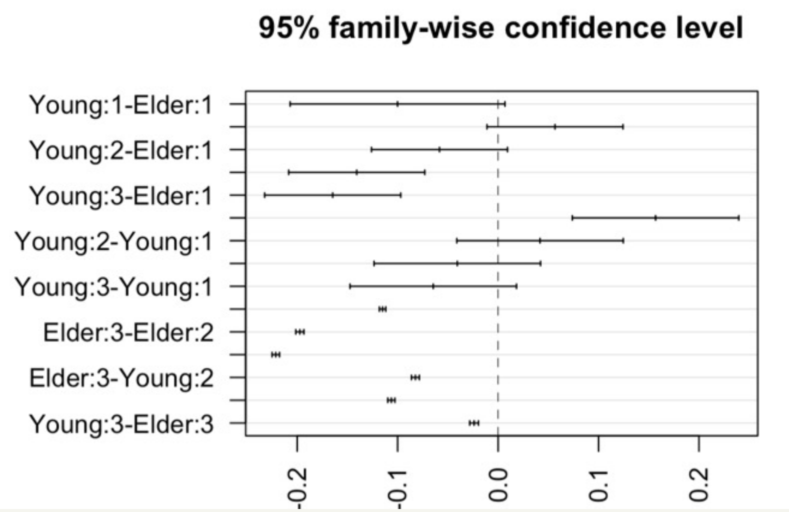


```
                                 Df Sum Sq Mean Sq F value Pr(>F)    
age                               1   2014    2014    8754 <2e-16 ***
fashion_news_frequency            1   7466    7466   32446 <2e-16 ***
age:fashion_news_frequency        1    639     639    2777 <2e-16 ***
Residuals       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

- Interpreting the outcome we can say that as **p-value** is **less than significance** value hence, we reject the null hypothesis.

Analysing Statement 1: Tukey

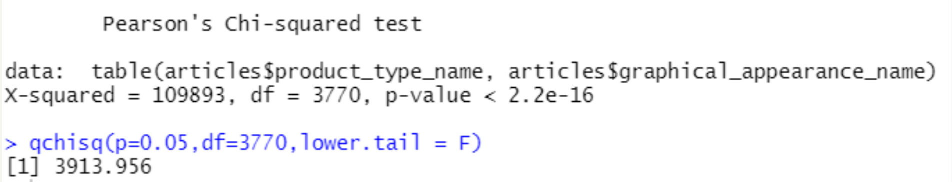
- From this test we can see that p-value of individual variables are almost 0 and for interaction of age and fashion news frequency is also 0. In nutshell, All of the variables are less than value of alpha which is 0.05 as mention earlier.
- From Tukey interpretation, we can say is Youngs whom are less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



Analysing Statement 2

How has graphics and product type has impacted the sales.

- Null Hypothesis: Graphic patterns and Type of Articles do not have any impact on purchases.
- Alternative Hypothesis: Graphic patterns and Type of Articles have impact on purchases.



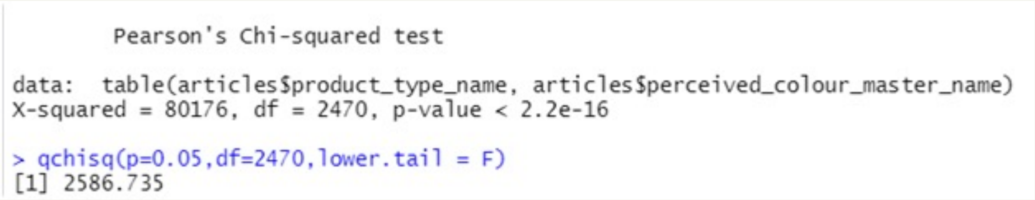
```
        Pearson's Chi-squared test

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
```

- Interpreting the outcome we can say as p- value is very much less than significance value we reject the null hypothesis

Analysing Statement 3

How has colours and product type affected the sale in H&M.

- **Null Hypothesis**: 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**: They have impact on sales made by clients.



```
        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
```

- Interpreting the p-value which is very less than alpha and that is why we reject the null hypothesis.

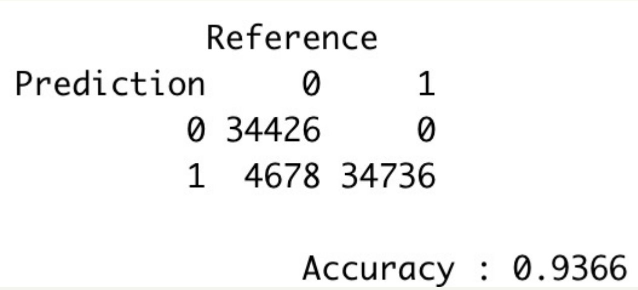
Models and Prediction.

- 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.

Train-test and Confusion Matrix

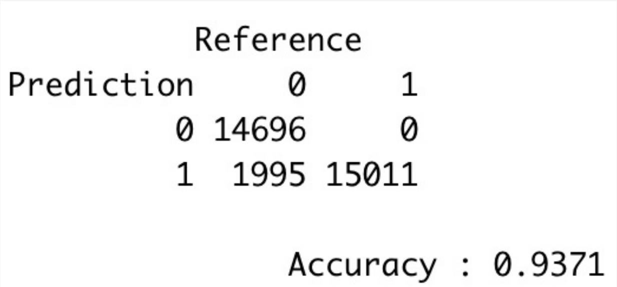
First we split dataset into train and testing sets which had threshold of 0.7.

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.



```
          Reference
Prediction     0     1
         0 34426     0
         1  4678 34736

               Accuracy : 0.9366
```

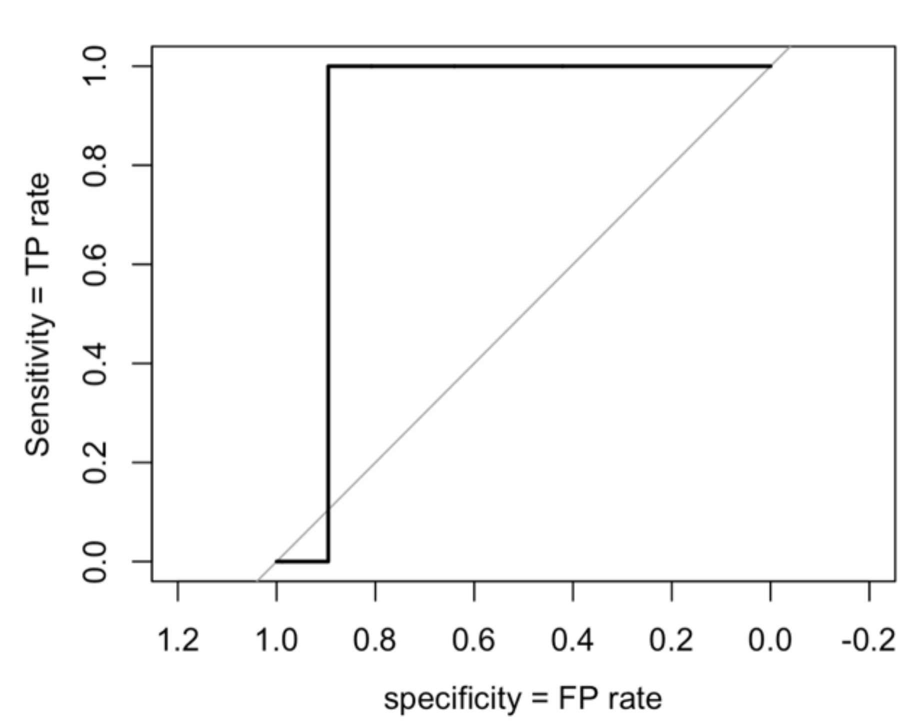
Train set



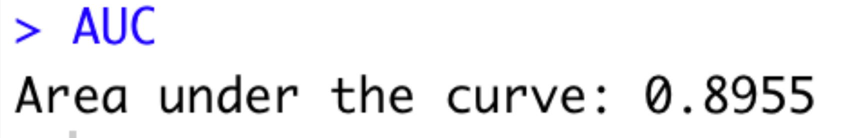
```
          Reference
Prediction     0     1
         0 14696     0
         1  1995 15011

               Accuracy : 0.9371
```

Test set

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

- By performing above tests we can conclude that the tests taken helps us to understand the analysis and interpretation of the dataset as hypothesis testing gives us gain confidence with proper proof.
- To figure out whether the article picked up by the customer is solid or not we have performed logistic regression which has accuracy of 94% on training set and 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 to predict a given random event(customer picking up a specific category of t-shirt). We were able to get AUC value equal to 0.8955. Closer the value to 1, better 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