Potential Clothes Analysis

Presentation By:

Anisha Mittal

Agenda

01
Problem Statement

02
Data

03
Analysis

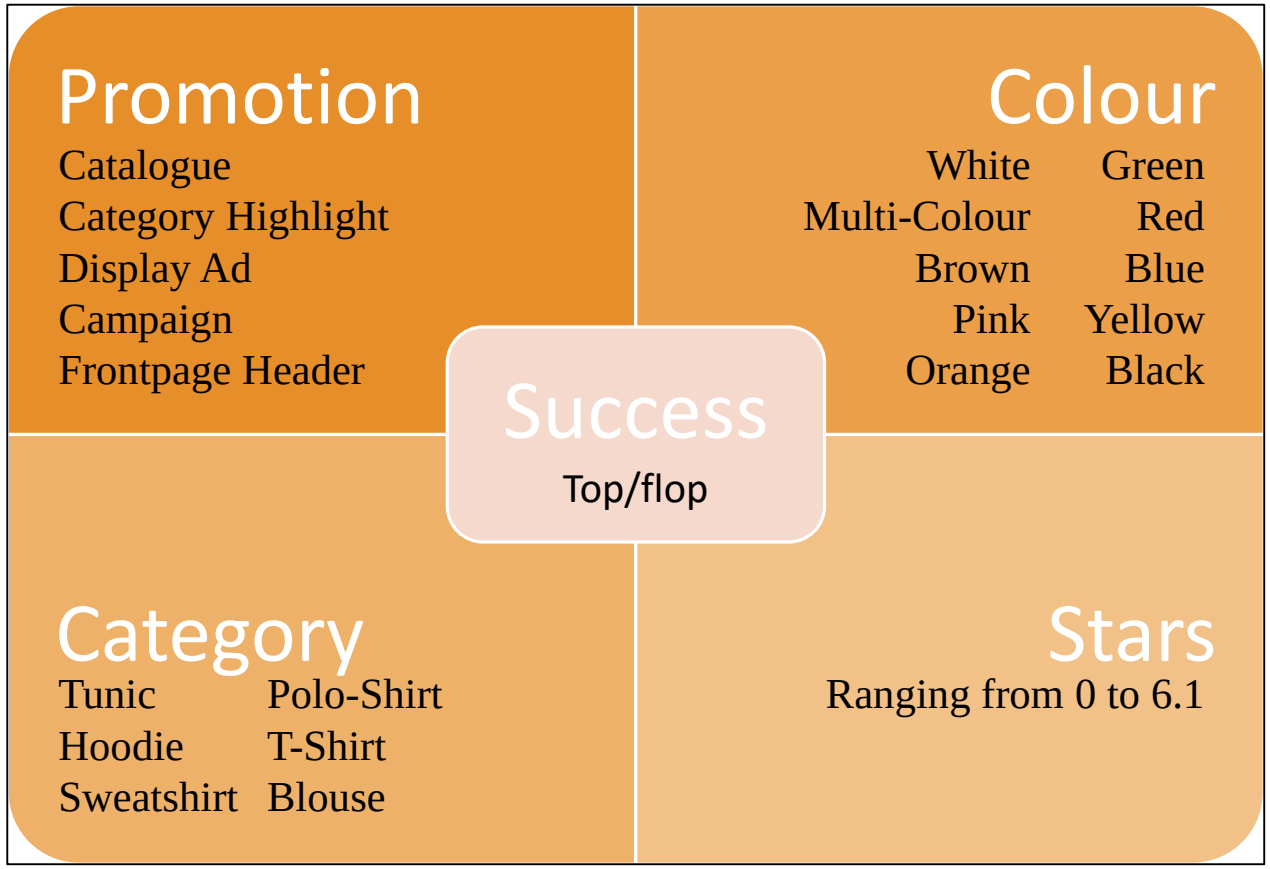
04
Prediction Results

05
Stakeholders Inference

Problem Statement

WHICH PRODUCT WILL BE SUCCESSFUL?

Data



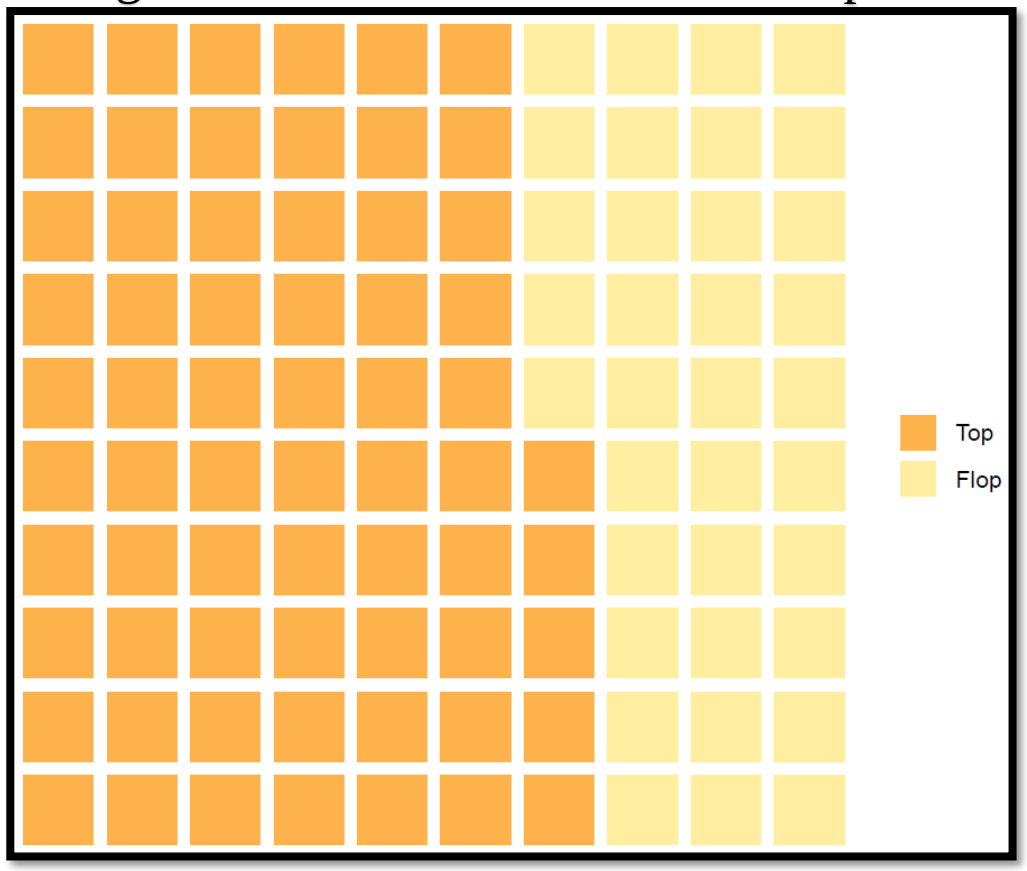
About

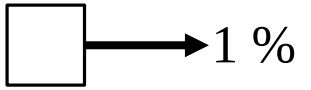
- Data for a fashion e-commerce company's collection.
- Two datasets:
  - Past years successful or not products
  - Future Product details
- 8000 and 2000 entries respectively.
- No missing values.

Analysis

Exploratory Data Analysis (top/flop)

Percentage of successful and unsuccessful products in past years





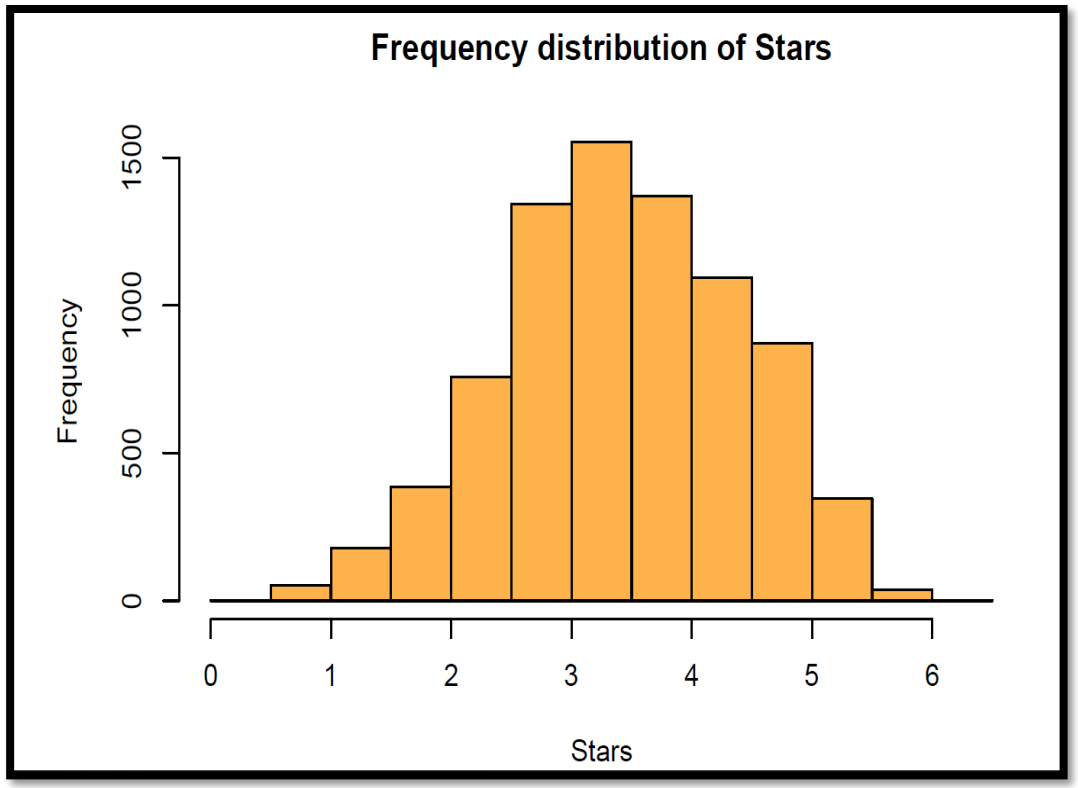
Top: 65% products
Flop: 35% products

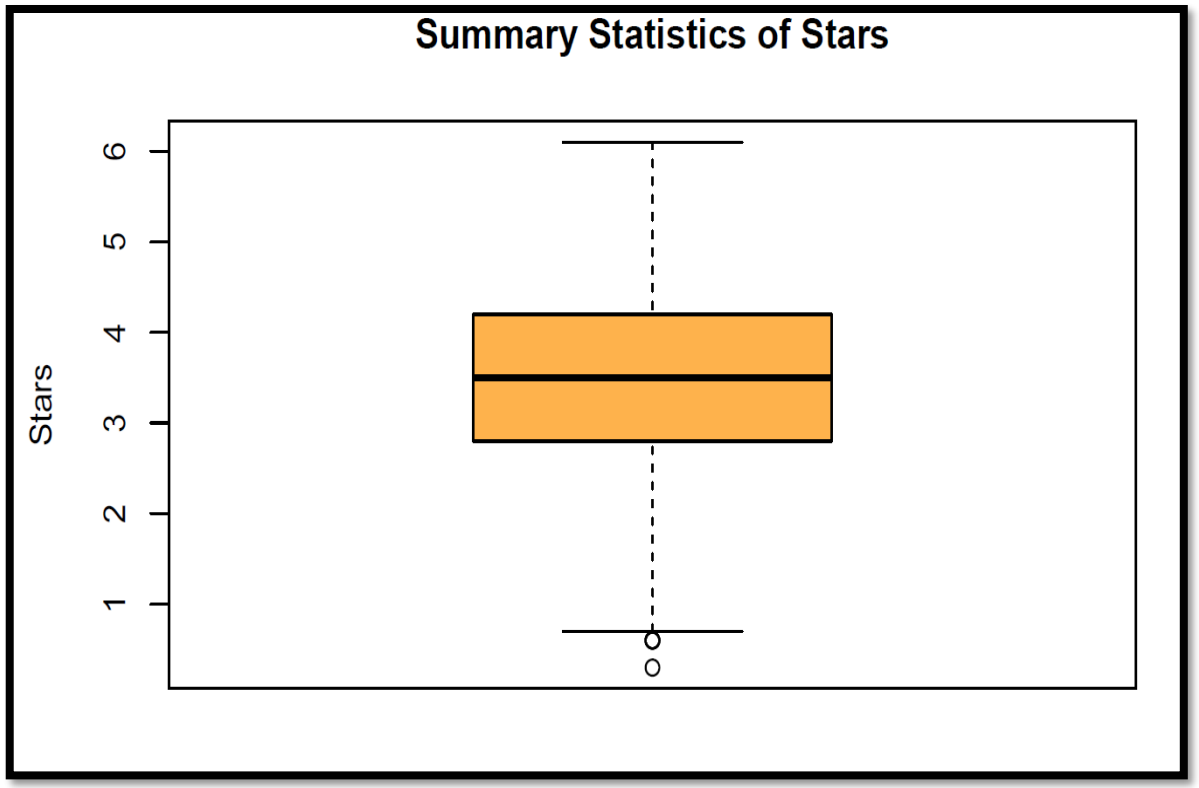
Waffle plot

Exploratory Data Analysis (stars)

Frequency distribution and Summary statistics of stars

- Normally distributed
- Two outliers
- Mean: 3.473
- Median: 3.5

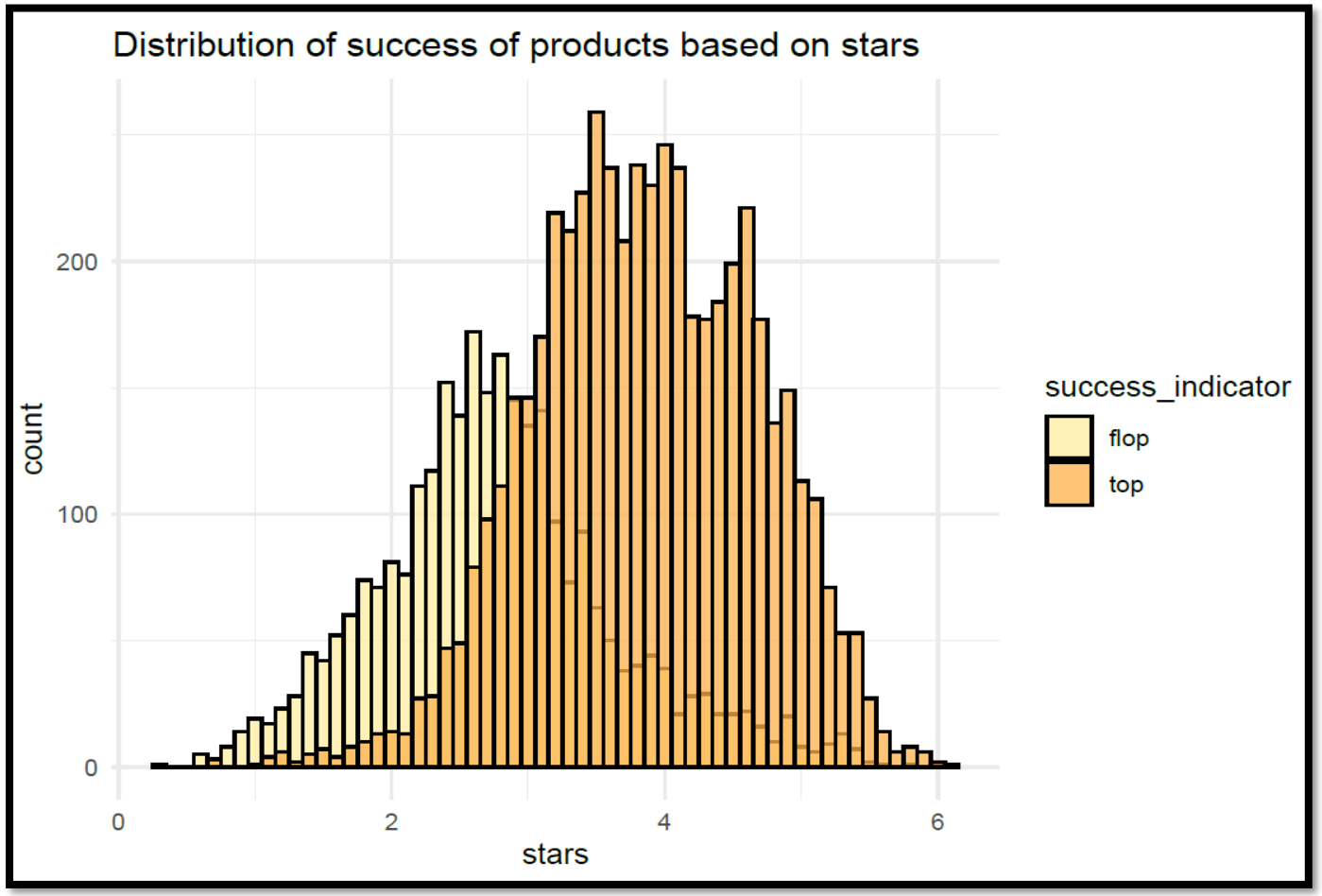




Histogram

Boxplot

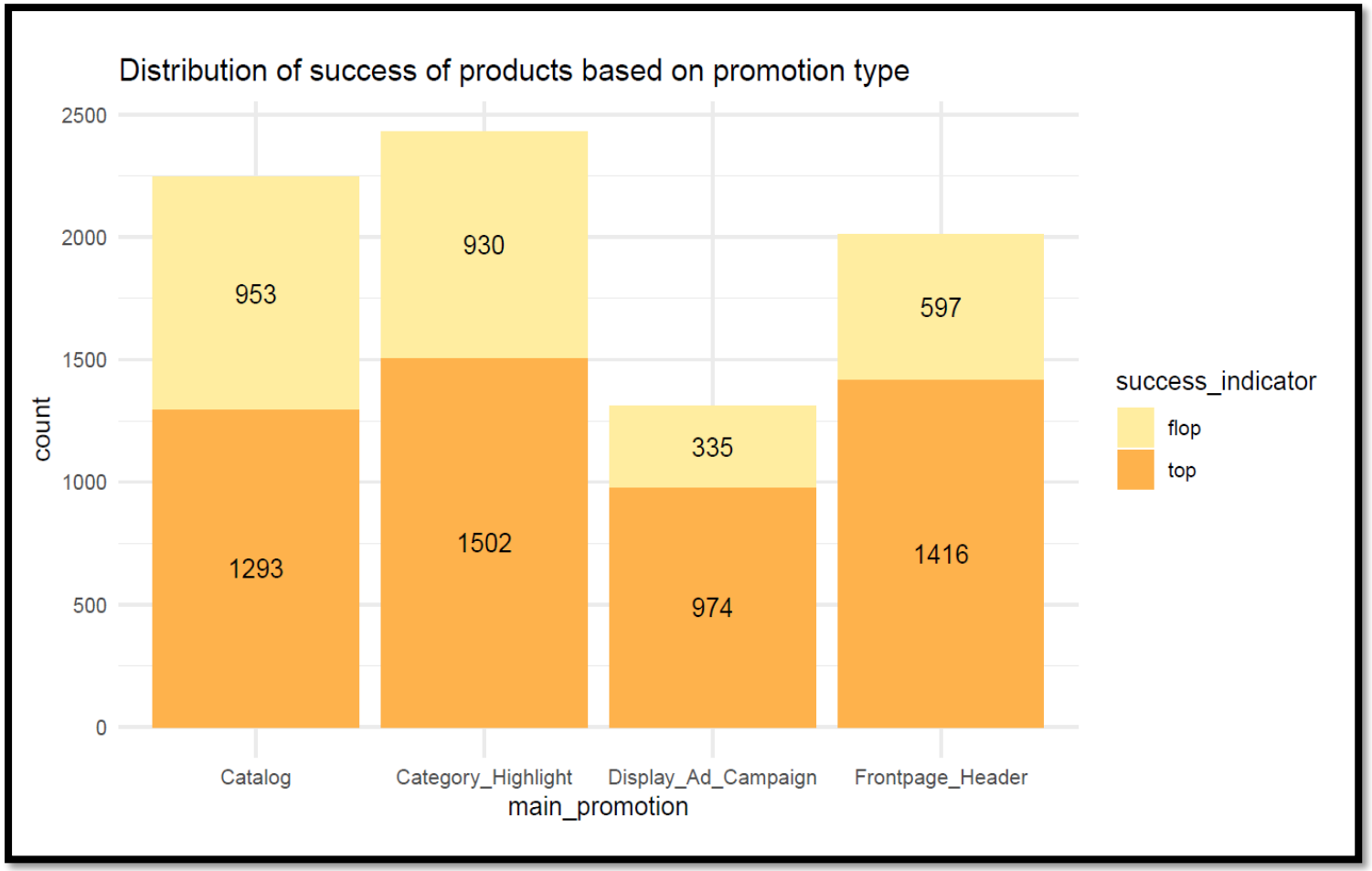
Exploratory Data Analysis (Stars)



- The distribution of top and flop overlap in terms of stars of reviews.
- Even products with stars greater than 4 have been unsuccessful.
- Evenly distributed.

Histogram

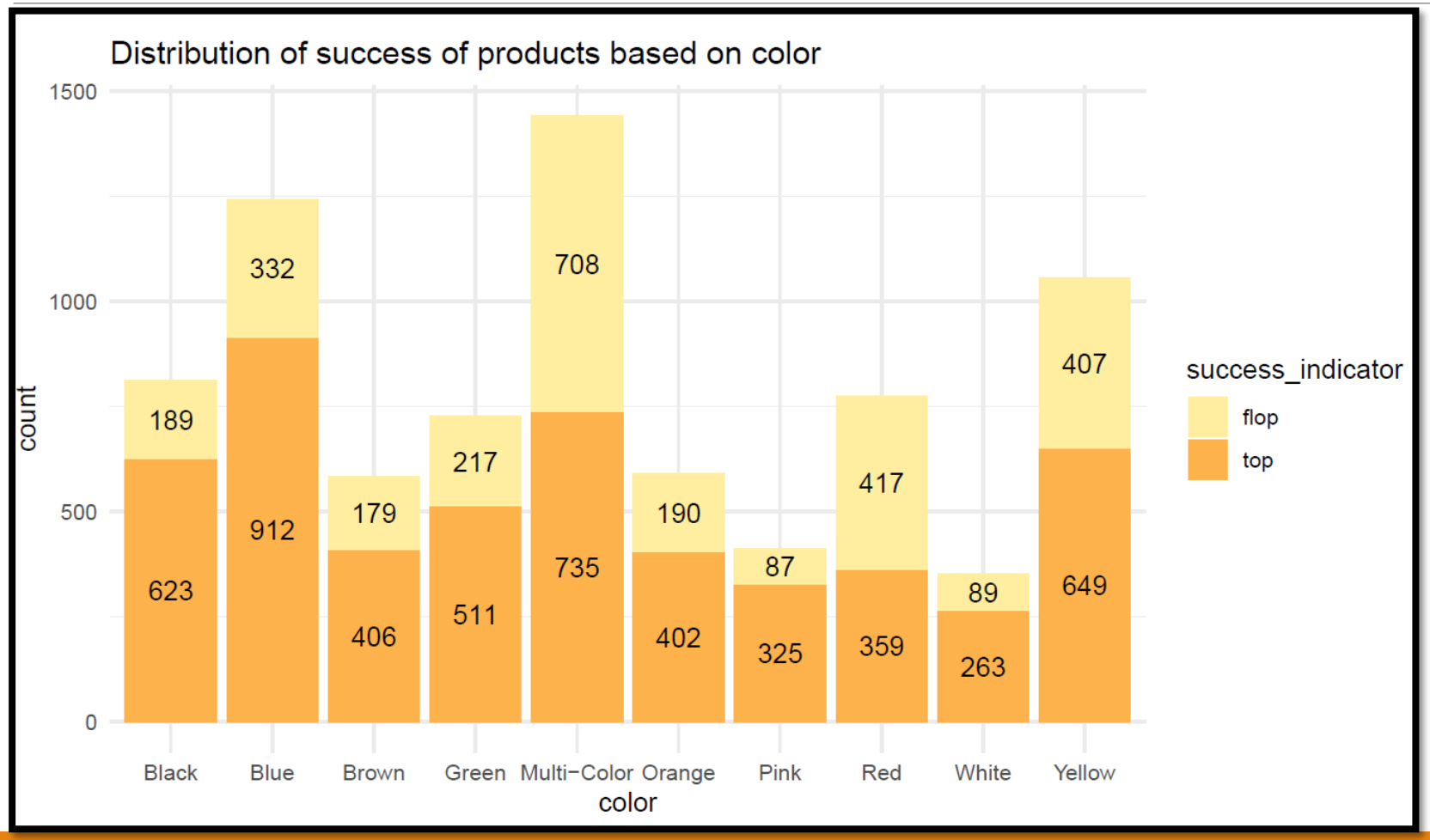
Exploratory Data Analysis (Promotion type)



- 28.97% of total successful products were promoted through Category highlight.
- Category Highlight was the most frequent used promotion method, where 61.75% of its products were successful.
- Display Ad Campaign was the least frequent promotion method, still 75% of its products were successful.

Stacked Bar graph

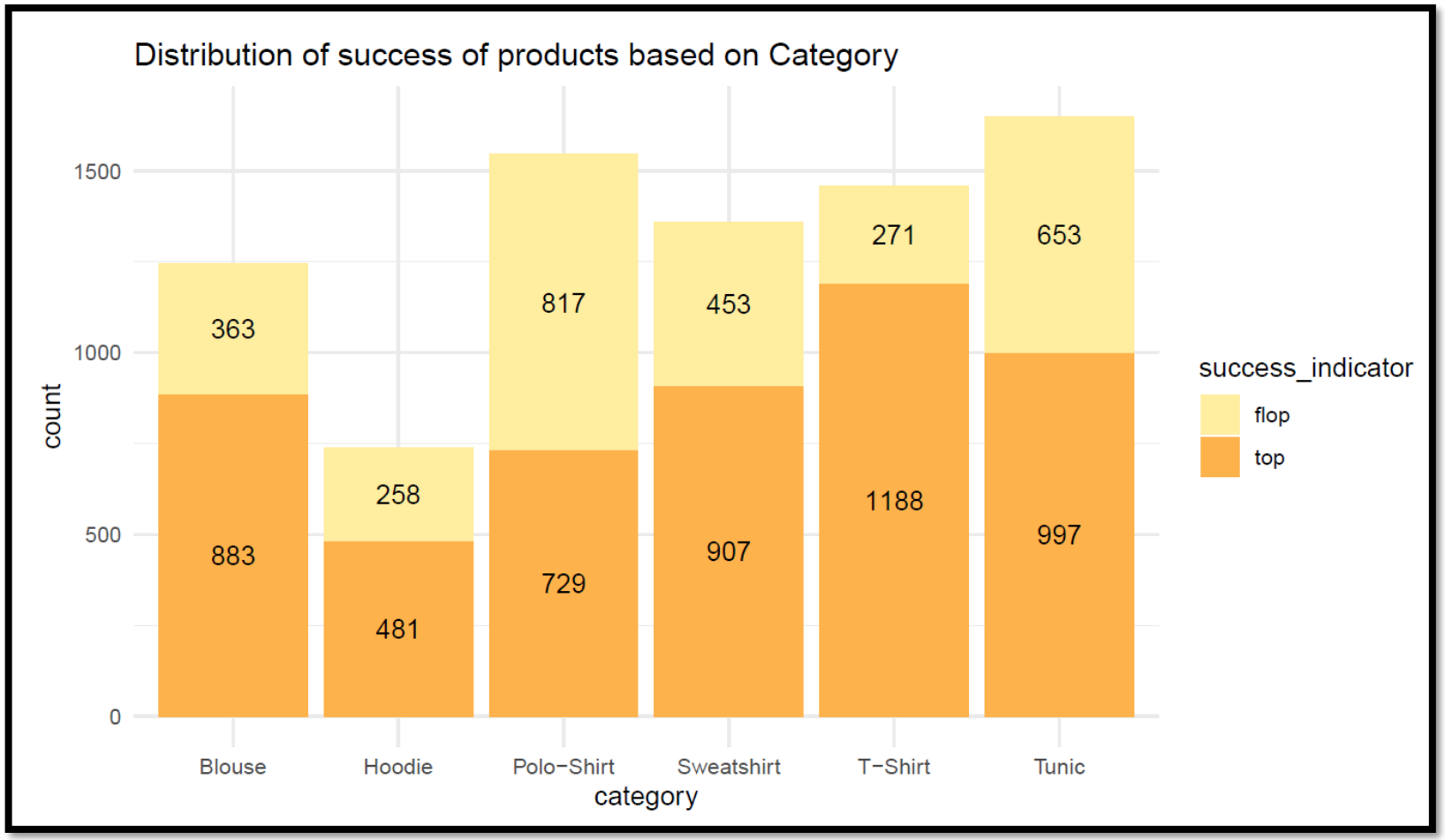
Exploratory Data Analysis (Colour)



- Blue and Multi-coloured products were tested the most.
- Only 50.9% of Multi-coloured products were successful.
- 78.8% of pink products were successful.
- Red products were the least successful with only 46.2%.

Stacked Bar graph

Exploratory Data Analysis (Category)



- Tunic and Polo-Shirt kind of products were tested the most.
- 81.4% of t-shirts were successful and also were the most success with 22.9%, Tunics are a close second.
- Even though polo-shirts were the most frequent product, only 47.1% of its products were successful.
- Hoodies has the least success.

Stacked Bar graph

Association

Cramer's v

It is a measure of association between two nominal variables.

A value between 0 and 1.

Value near 0 means low association and a value near 1 high association.

Predictors	Cramer's V with success
Stars	0.58
Category	0.23
Colour	0.22
Promotion type	0.13

Statistical Modelling

Models Tried

Logistic Regression

Random forest

Method

Divided historic data in training (75%) and testing (25%) data.

Used the most accurate model to predict the success or not of the many potential products.

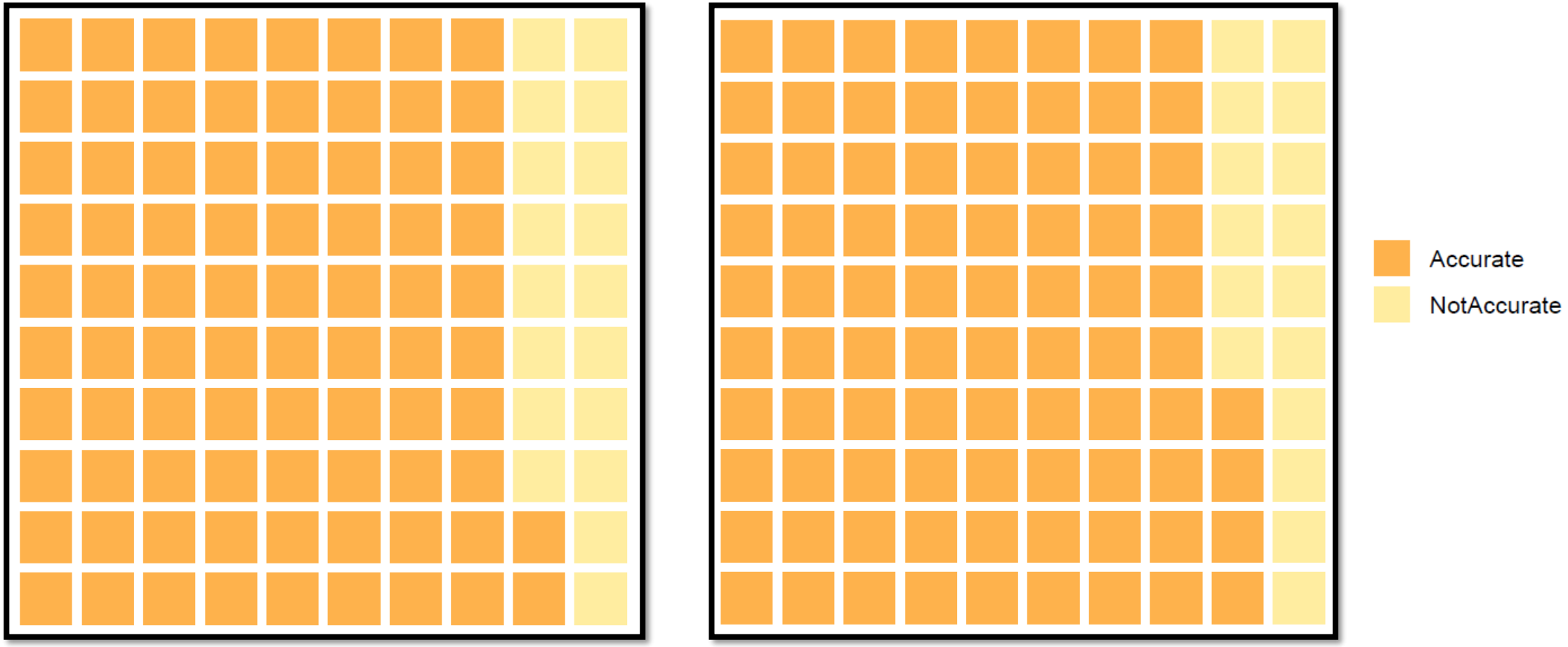
Reason

Response variable is nominal.

Could have used unsupervised learning but it would require a lot of data manipulation.

Since we already know the classes of the response variable, Supervised learning is a better choice.

Statistical Modelling Accuracy

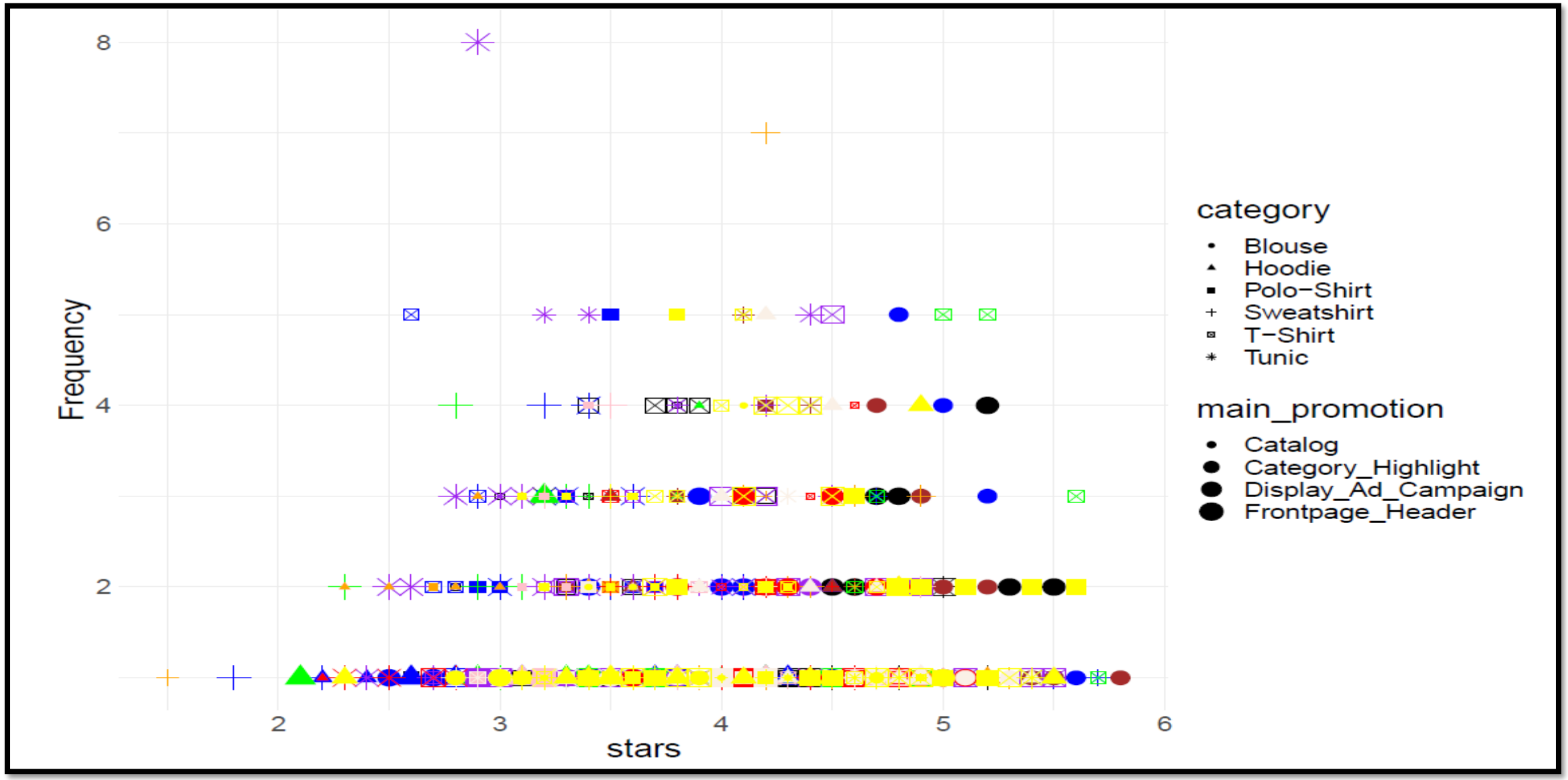


Logistic Regression

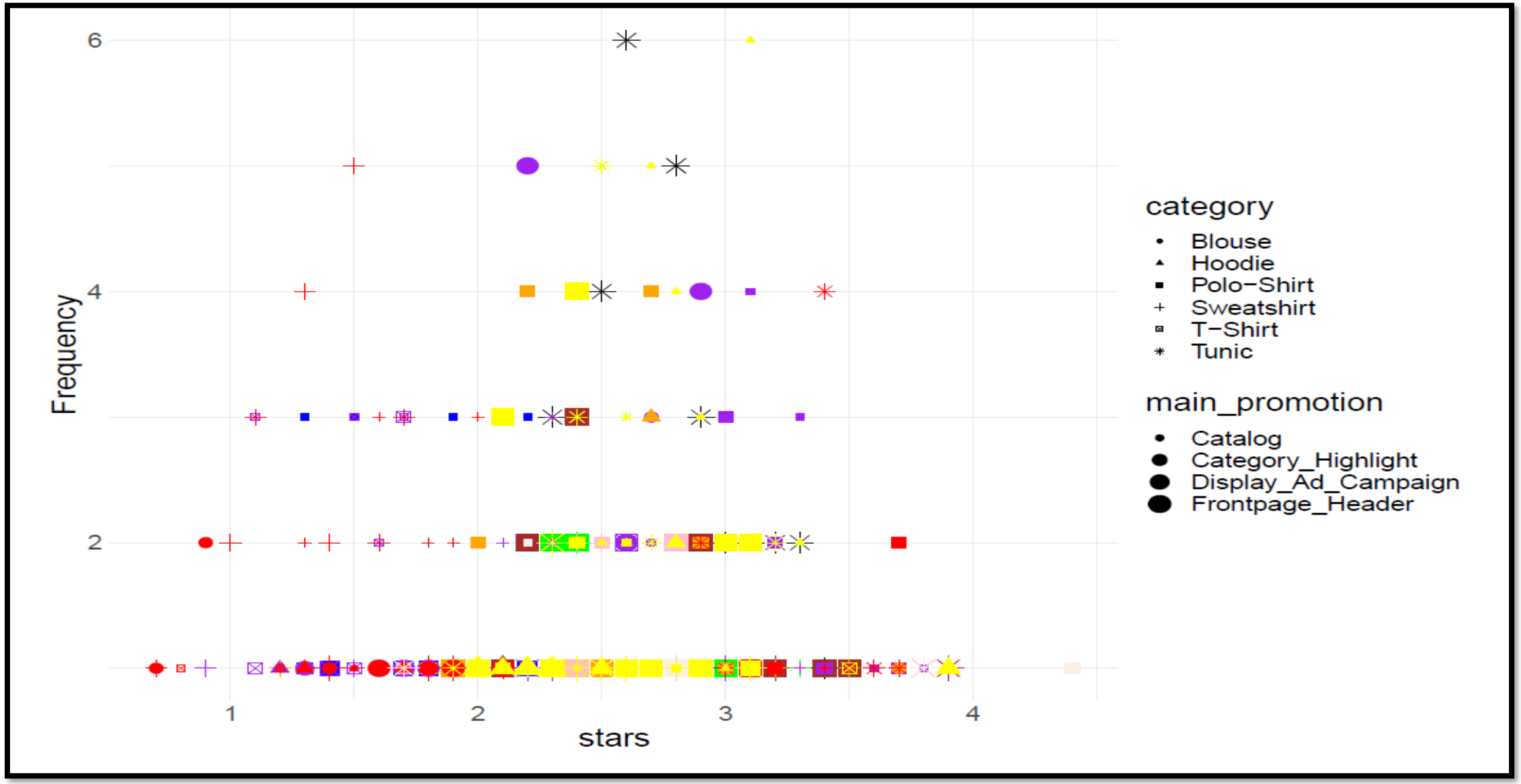
Random forest

Prediction Result

Frequency of SUCCESSFUL potential products based on category, colour, stars, and promotion type



Frequency of UNSUCCESSFUL potential products based on category, colour, stars, and promotion type



Stakeholders Inference

01

Engage customers with display ad campaign form of marketing the most.

02

T-Shirts and Tunics can be promoted the most to increase traffic on website.

03

Organize and present a collection that customers can clearly see which have advantages over the competitors based on color and category.

04

Produce/stock more multi-color and yellow products.

THANK YOU

QUESTIONS?