Analysis Report: Sports Group Marketing Campaign Performance

Introduction

The aim of this analysis report is to provide insights into the process of performance of marketing campaign in the context of sports Group's business model. As a company specializing in managing sports items for customers, accurately determining the performance of the markering campaign is crucial for ensuring fair transactions and maintaining competitive advantage in the market.

Estimating the performance of marketing campaign to enhance customers to purchase more involves considering various factors, such as the number of available articles they selling, the countries where the company invests their products, the number of unit solds in retail week, The price of each unit, the promos applied during these intervals, the product groups they belongs to, their cost, style, and sizes, we can develop a robust a marketing strategy that aligns with customer expectations and market trends.

Through this analysis, we aim to gain insights into the key factors influencing performance of marketing campaign, identify patterns or trends in the market, and develop a data-driven approach that aligns with customer expectations and the competitive landscape. By leveraging the power of data science and statistical techniques, we can enhance our decision-making process and optimize our pricing strategy for imporving the marketing imapct to the customers.

The following sections of this report will go into the details of the analysis, including data preprocessing, descriptive statistics, correlation analysis, feature importance, provide recommendations based on the findings for the company.

Data Overview

The analysis is based on given dataset that contains information relevant to performance of marketing campaign.

The dataset includes the following columns:

Column	Description
country	Country name, three unique countries.
article	6 digit article number, as unique identifier of an article
sales	total number of units sold in respective retail week
regular_price	recommended retail price of the article.
current_price	current selling price (weighted average over the week)
ratio	price ratio as current_price/regular_price, such that price discount is 1-ratio
retailweek	start date of the retailweek.
promo1	indicator for media advertisement, taking 1 in weeks of activation and 0 otherwise

Column	Description
promo2	indicator for store events, taking 1 in weeks with events and 0 otherwise
customer_id	customer unique identifier, one id per customer
article	6 digit article number, as unique identifier of an article, 10 unique types or articles.
productgroup	product group the article belongs to
category	product category the article belongs to
cost	total costs of the article (assumed to be fixed over time)
style	description of article design.
sizes	size range in which article is available.
gender	gender of target consumer of the article.
rgb_*_main_color	intensity of the red (r), green (g), and blue (b) primaries of the article‘s main color, taking values [0,250]
rgb_*_sec_color	intensity of the red (r), green (g), and blue (b) primaries of the article‘s secondary color, taking values [0,250]
label	advertisement result after offering/sending/presenting the offer to the customer. 0 means the customer did not buy and 1 means the costomer did buy

Understanding the description of the features of the dataset, and how label column was defined, It was assumed that, this dataset contains the sales made by customers, and label column identify the advertisement effect; So zero means the customer made the purchase but without being affected by a promo, and One means the customer made the purchase due to being affected by a promo.

Data Preprocessing

Before conducting the analysis, several preprocessing steps were performed on the dataset.

These steps included:

Handling duplicated records in the data:

- There are no complete duplication records in the dataset.
- We have 3 columns of unique ID ['article_id_1', 'article_id_2', 'customer_id'], but according to our case in this dataset, they could be possible to be duplicated, as the duplication here indicates that, the same customer could have multiple records as he\she buy multiple times different items from different articles, so accordingly the articles ids could have duplications aswell.

Hanlding negative values:

- To make sure there are no logical data like cost or price to be in negative.

Handling missing values:

- In both numerical and categorical features, but there were not much to handle.

Filtering outliers:

- Based on relevant features to have more robust and skewness free data.

Handling mislabeled target data:

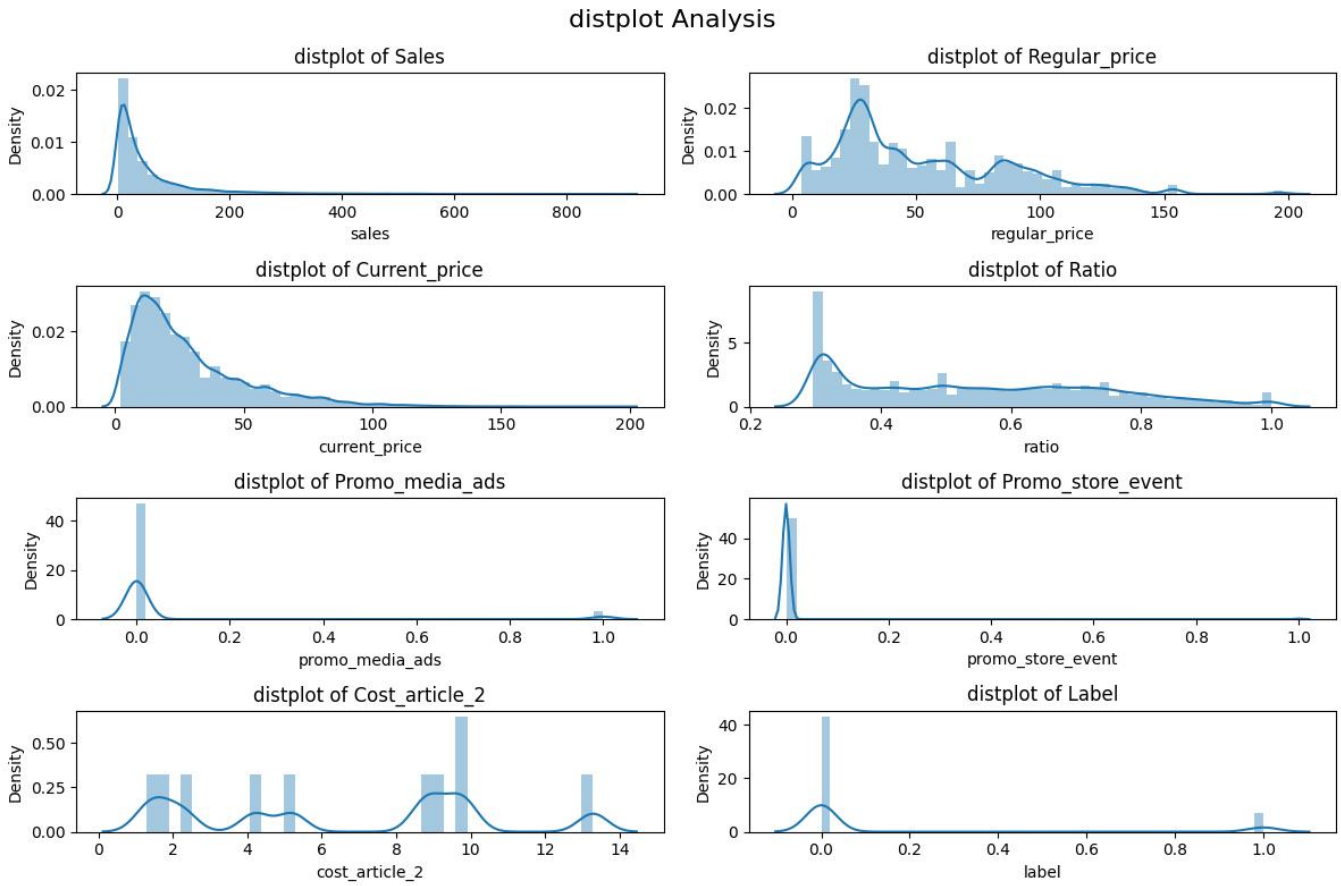
- This based on our understanding of the given dataset, for the target column 'label', where the zeros mean that the customer purchased the unit but without the impact of advertising campaign.
- So to have an purchase due to marketing impact on the customers, there should be atleast one of the given promos: [promo by media ads or promo by store event] is applied.
- Noting that the percentage of customers who said to have purchased due to promo impact, but in real-time there were no promos applied was around 12.5% of the total customers.

Analysis Insights

Starting with the eda report generated for the corresponding data, some important insights to show:

- There are no physical actuall missing values in the dataset, So we should check for a logical missing values in features, which means that some dummy values to represent a feature which are not logical for that feature, So to solve this, we will consider it missing, and try to handle the missing values.
- There are skewness in following features [sales, regular_price, ratio].
- Timeseries data starting from 2014 till 2017.
- Feature: Sizes, Two unique values, to indicate either the customer choise store with multiple sizes varaity or not. Unfortunatly, in this dataset we can't determine which sizes are actually common.
- Women is the largest customer type in the dataset, But have they the highest selling?.
- A first impression from the data, It's shown that most of the sales are made without using the marketing campain, Also for the promos features, most of the data says that there were no promo at retail week.
    - We need to analysis the small percentage of customers who bought with promo, what makes them bought it.
    - Also, as the data shows that, there are a huge percentage of customers who already bought multiple time without any promos, So we need to invistige how to make them buy more with special promos related to them.

Analysis: Individual Features Analayis of raw data

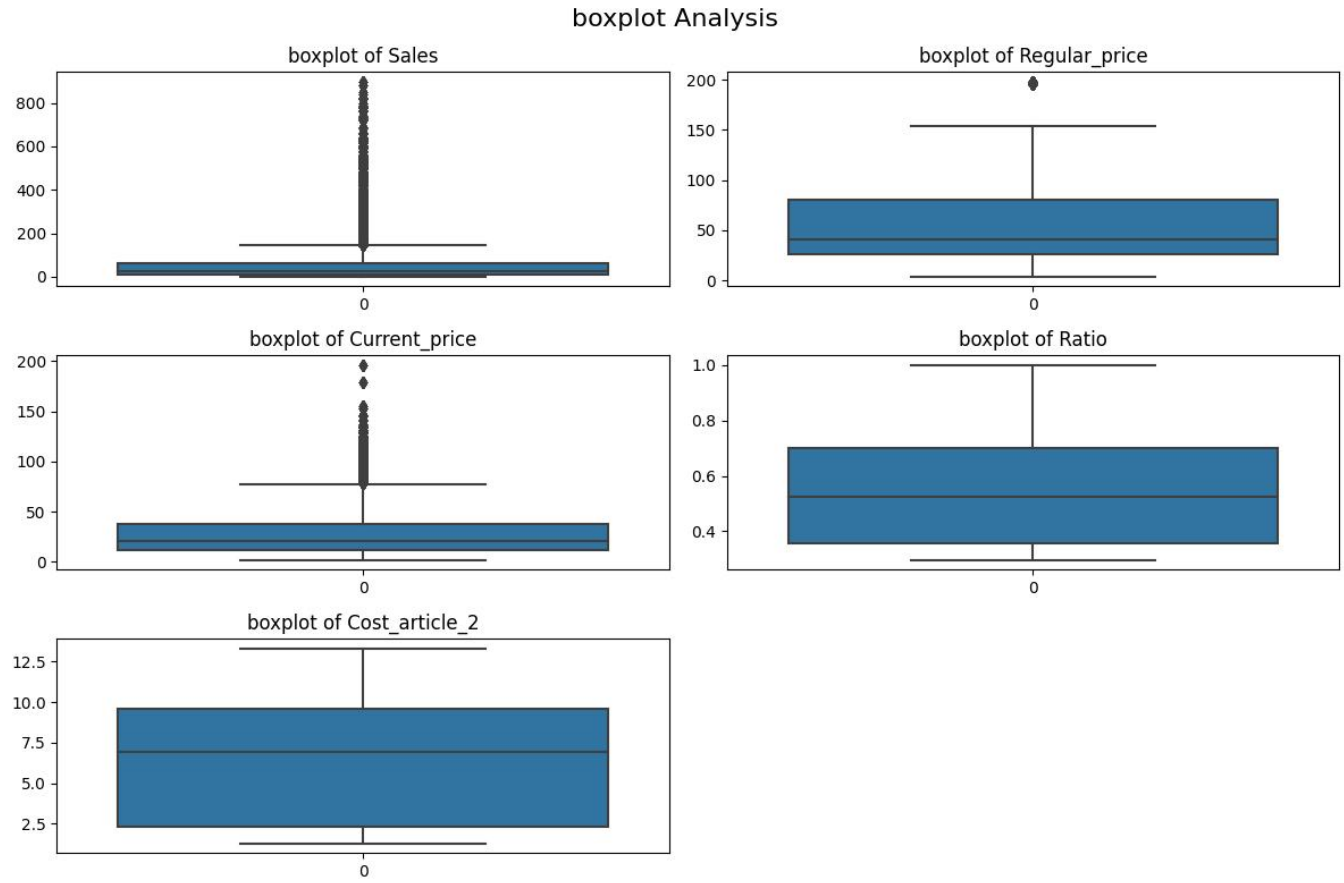


By analysing the distribution of each of these features [sales, regular price, curren_price, ratio, promo by media ads, promo by store event, cost of article id 2, labels], It was identified the following

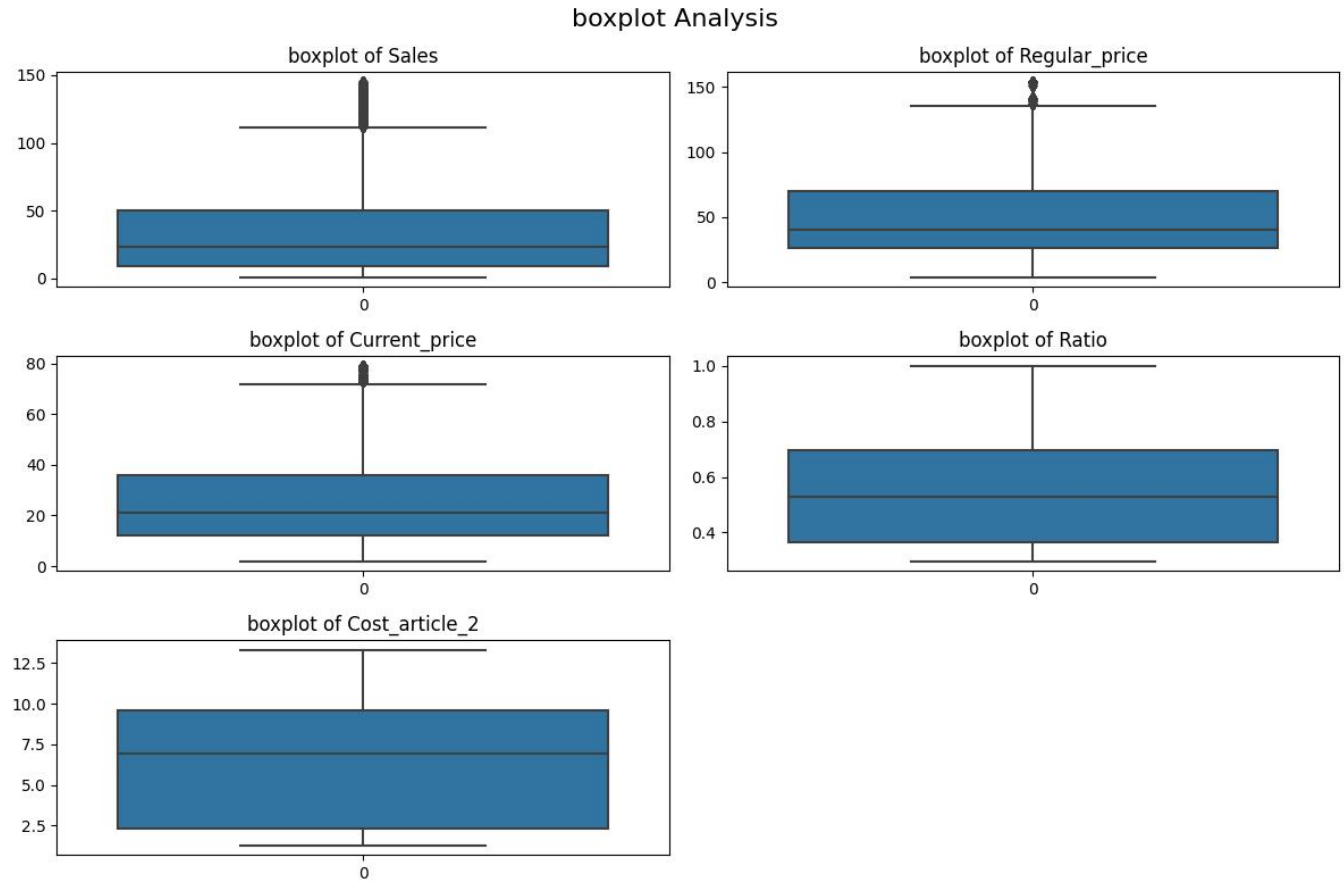
- There are skewness in {sales, regular_price, ratio} features.
- Huge outliers in both [ratio, cost of article id 2].

Analysis: Boxplot for each numeric features

Before Cleaning

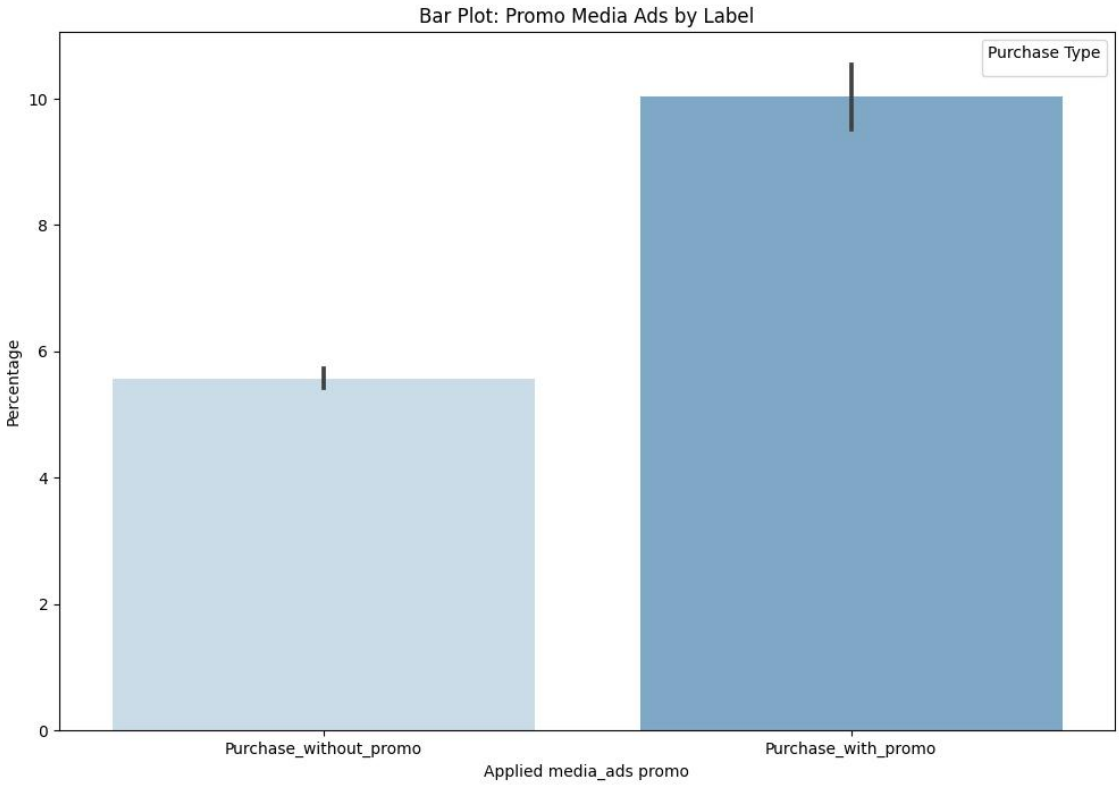


After Cleaning



Analysis: Promo Usage

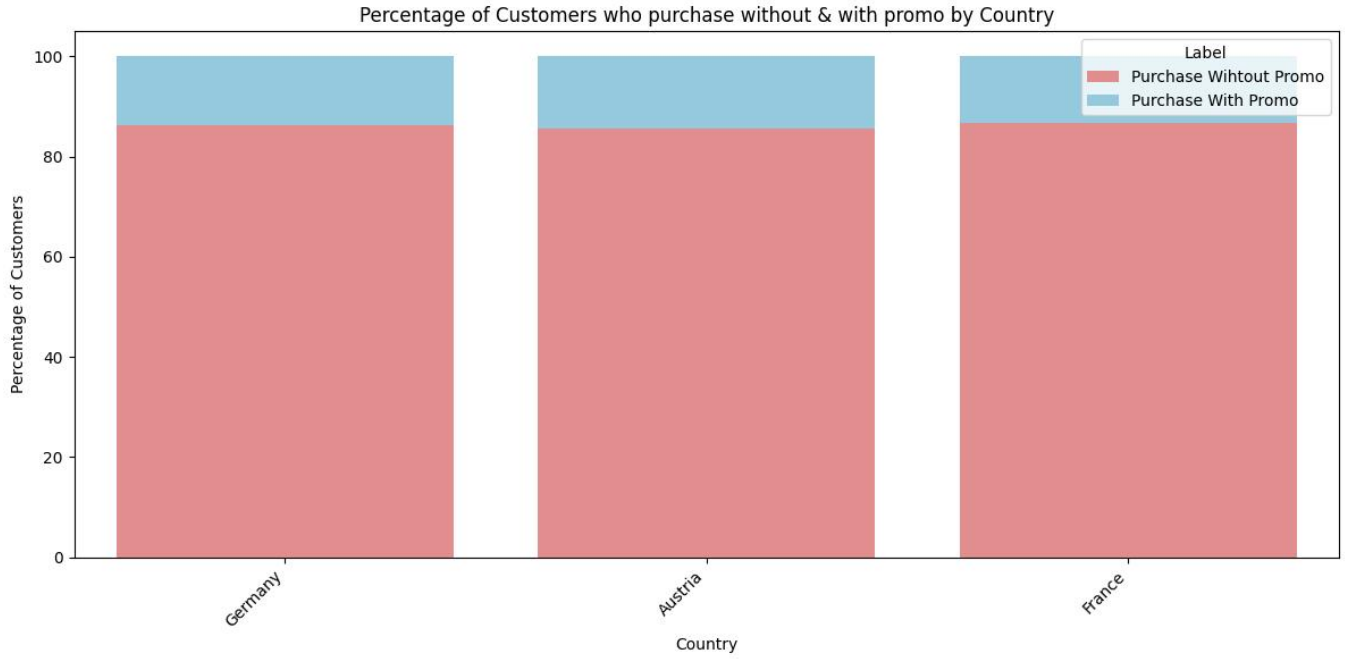
What is the percentage of applied promo_1 & promo_2



Insights regarding this points:

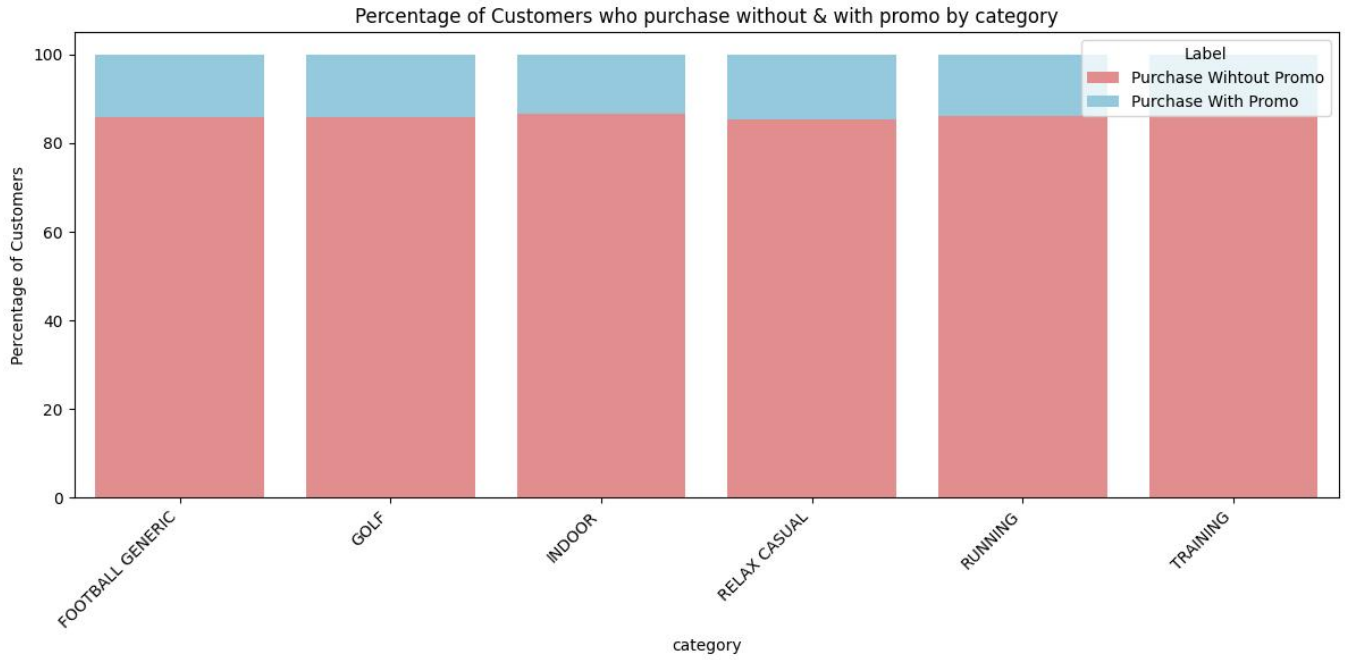
- 'Promo_media_ads': This promo was applied arount 6% during the purchases without promo, and around 10% during of the purchases with promo.
- 'Promo_store_event': This promo was applied arount 0.45% during the purchases without promo, and around 0.85% during of the purchases with promo. This indicates that, this promo was not applied much like the promo_media_ads.
- For both promos, the applied perecentage are high, so it is good indicator that the customers are willing to use them whenever they are applied.
- The verticle lines in the bars are called Error Lines, which means the hieght of the bar could vary depend on this verticle line, as it requires more data to accurately plot.

What is the percentage of promo usage according to each country?



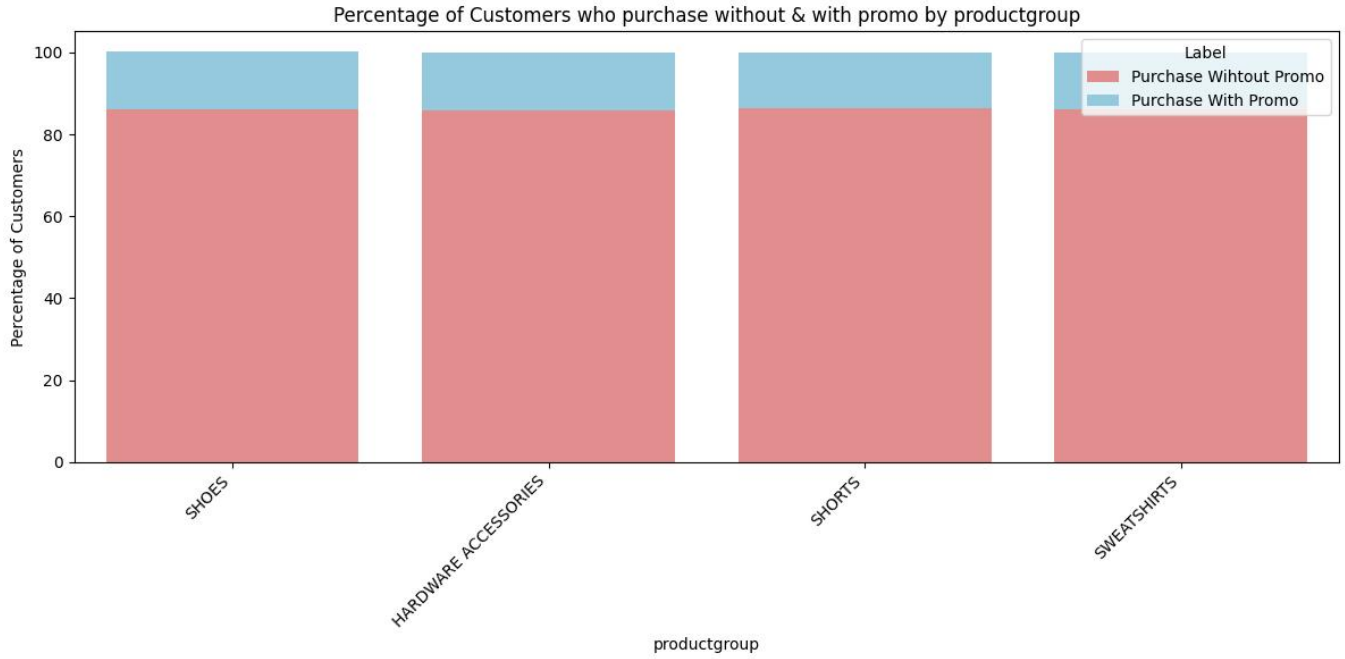
- The distribution of promo usage in purchases across each country is almost similar, as around 14% of the purchase with promo, and the rest are without promo.
- This indicates all the country have similar behaviour of of promo, So if a new promo is applied, then all the country will be behave the same.
- So the marketing team should be more focused on hunting these couinties and deal with them similarly.

What is most used category with promos?



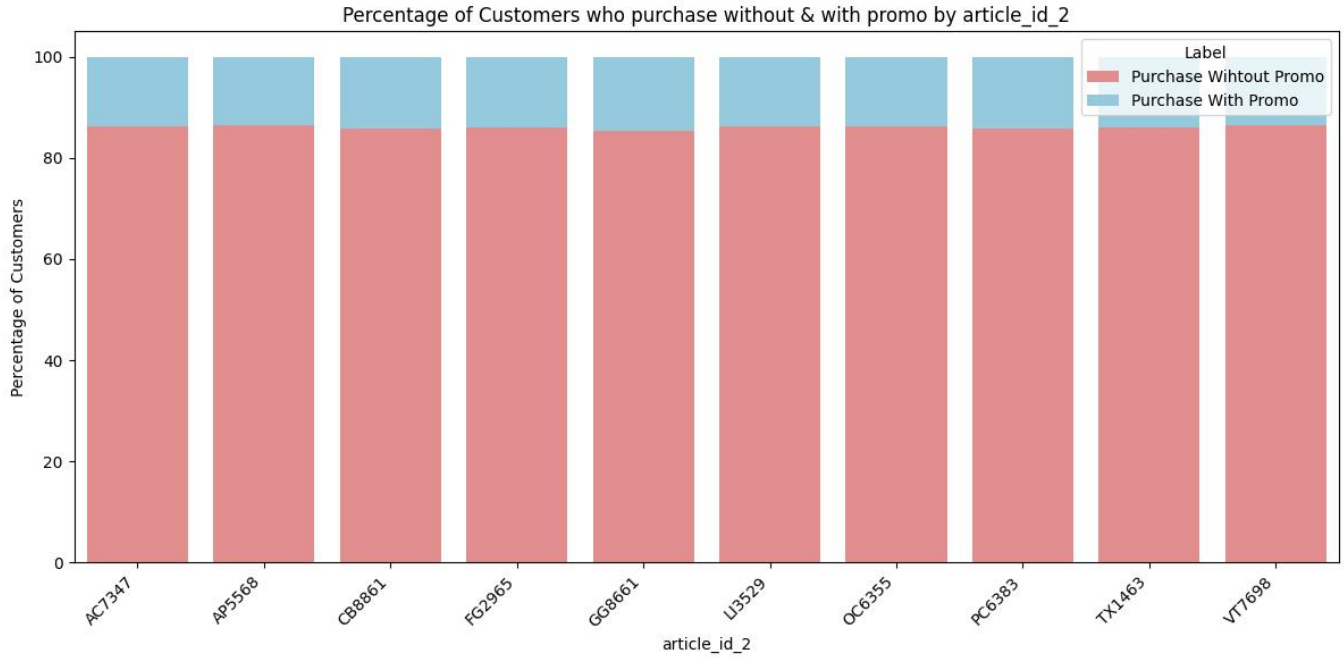
- The same distibution for each category we have, the percentage is divided by [15% : 85%].
- This indicates that the marketing team should also behave the same with each category.

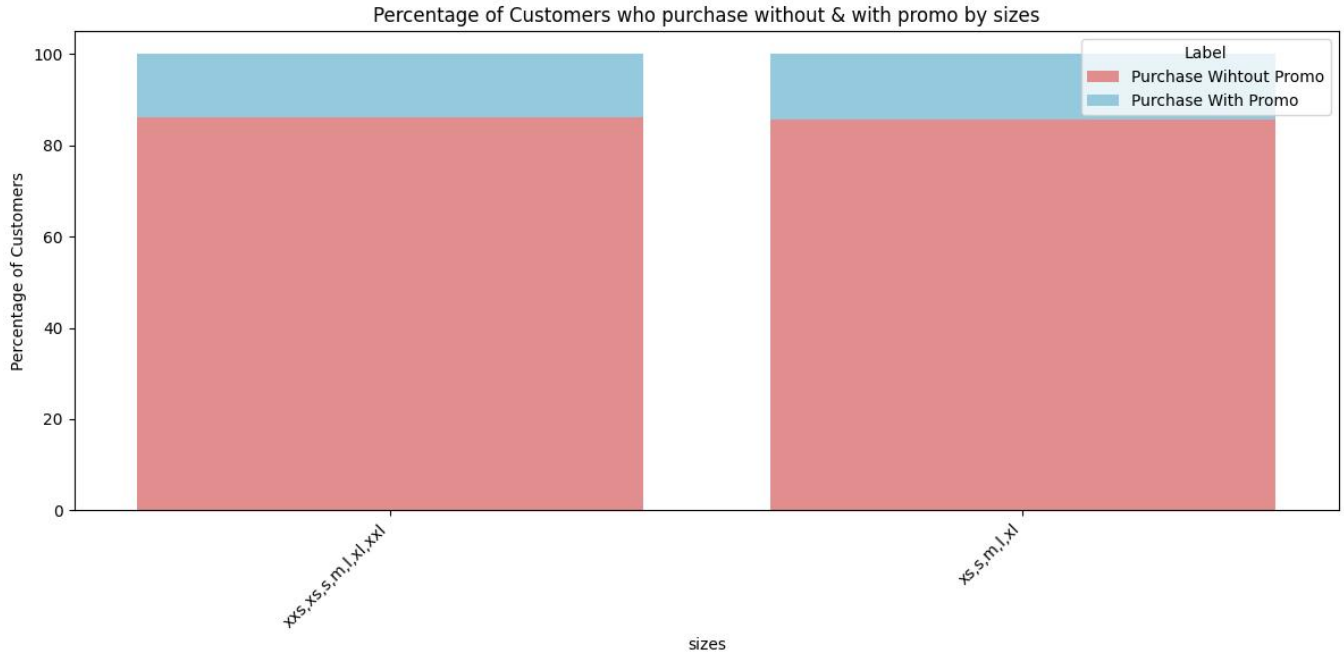
What is most used product group with promos?



- The same distibution for each product gourp we have, the percentage is divided by [15% : 85%].
- This indicates that the marketing team should also behave the same with each product group.

According to arcticle_2 & sizes which are the most used promo?





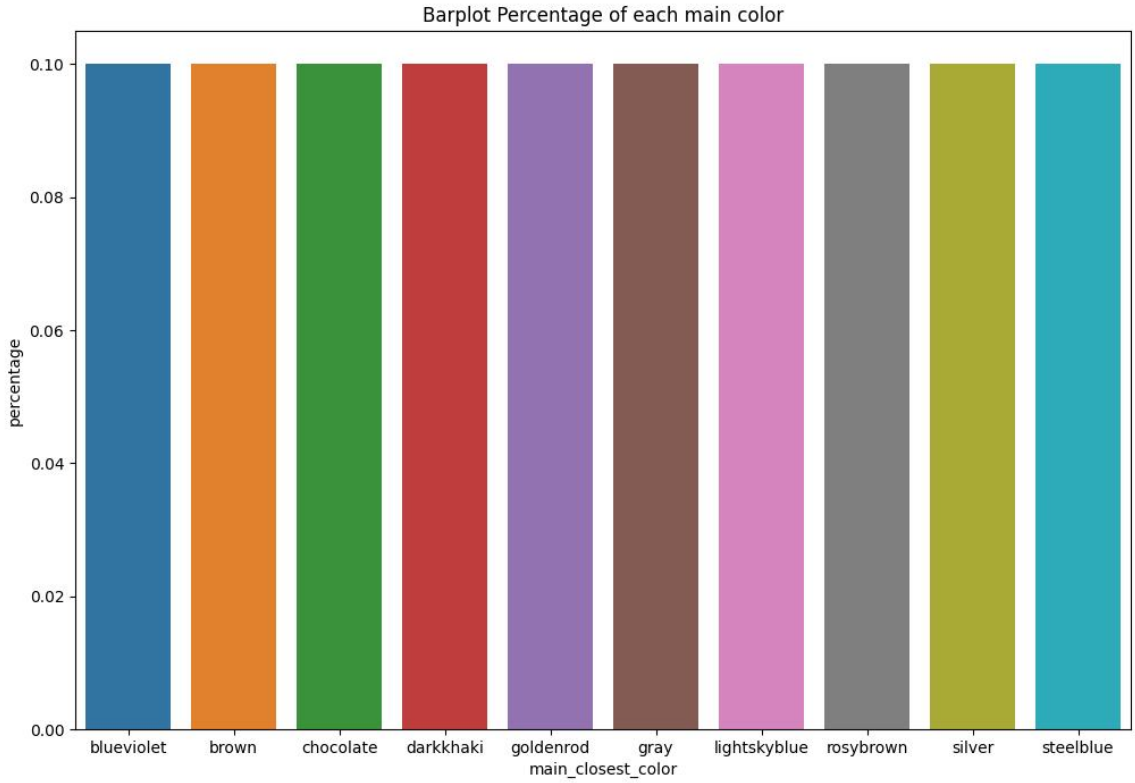
- Even over the articles id 2 types & size types, they have the same percentage distribution across purchase without & with promos.

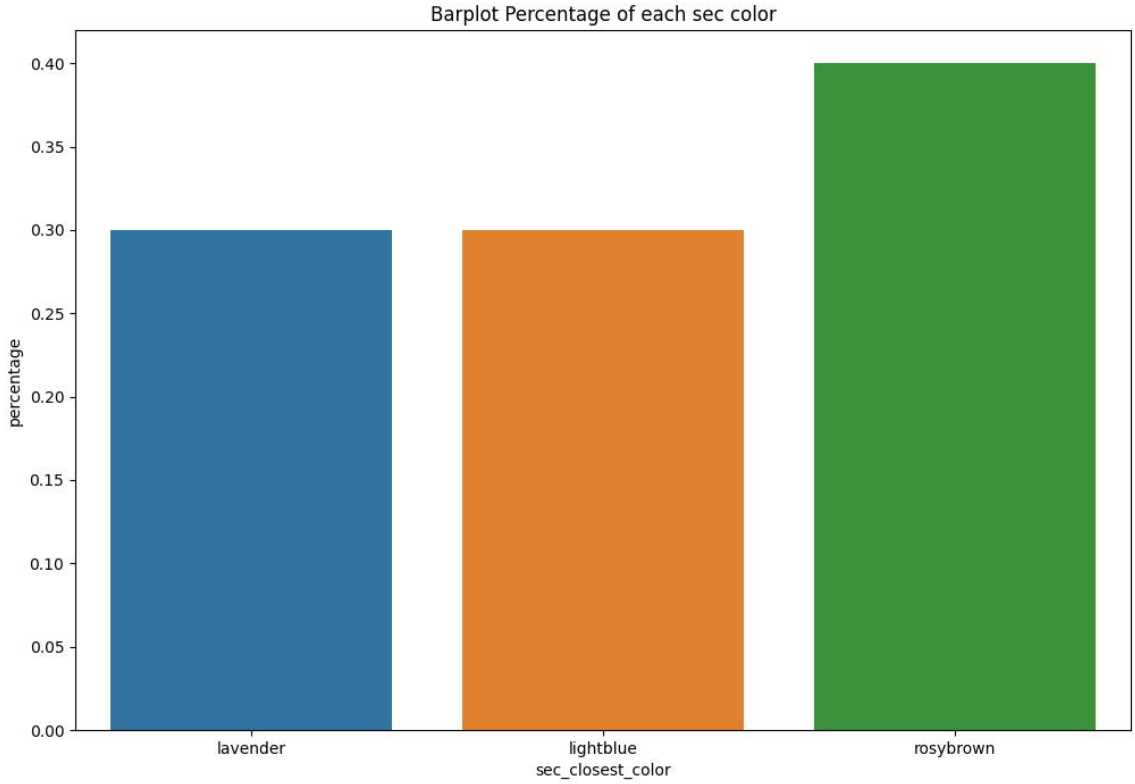
From All these insights, they can guide our promo impact strategies, We can determine which country, article type, size, and category response to promo impact.

Analaysis: Color Features

The RGB values for the primary and secondary articles in the dataset represent the intensity of the red, green, and blue primaries of the respective colors.

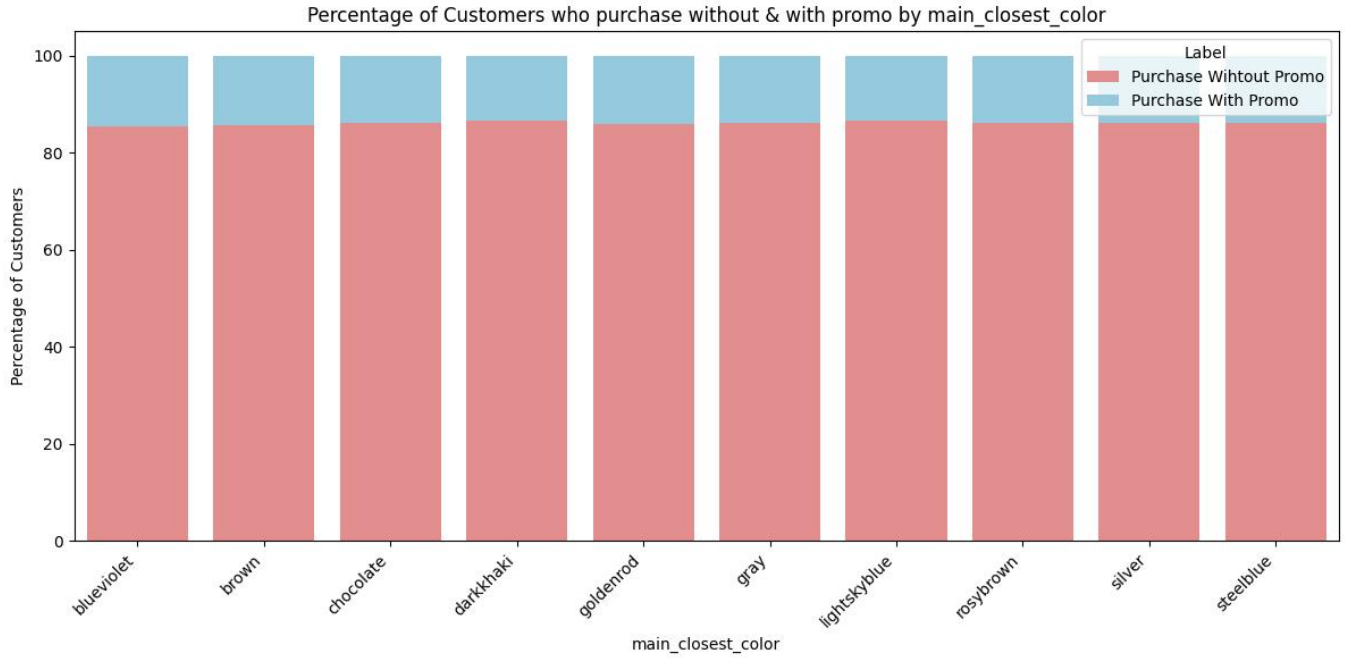
What is the most popular main and secondary colors?

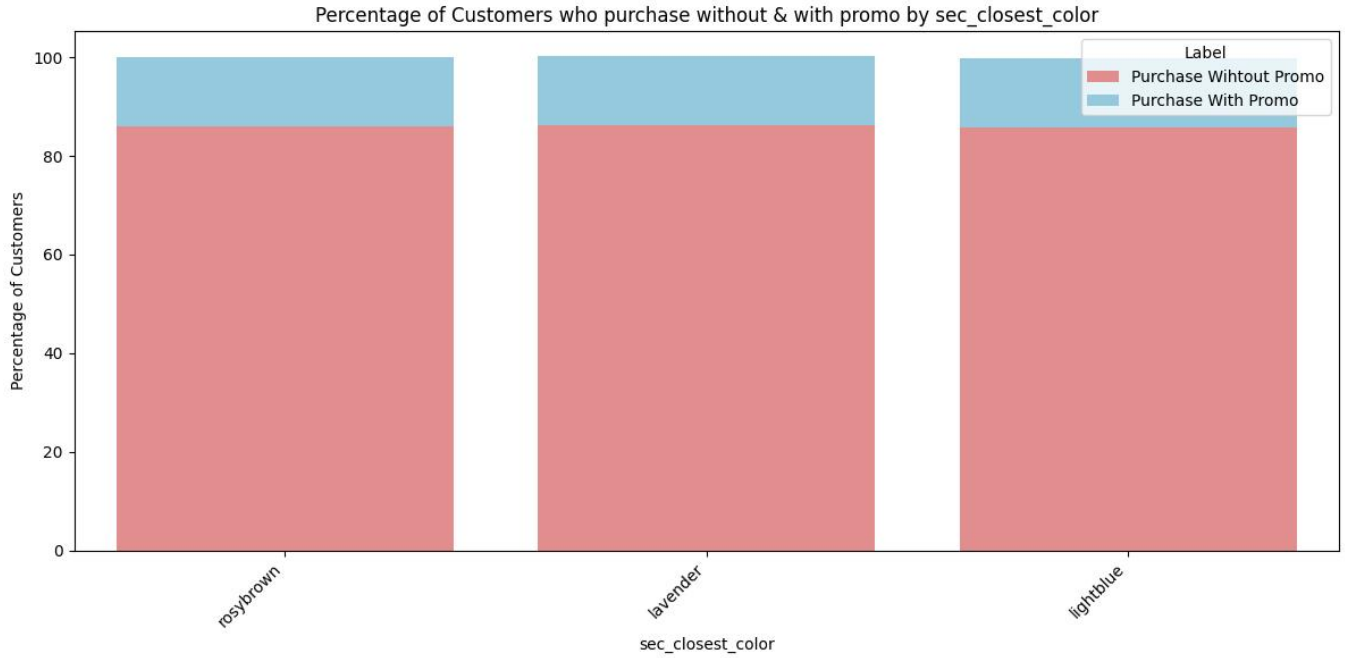




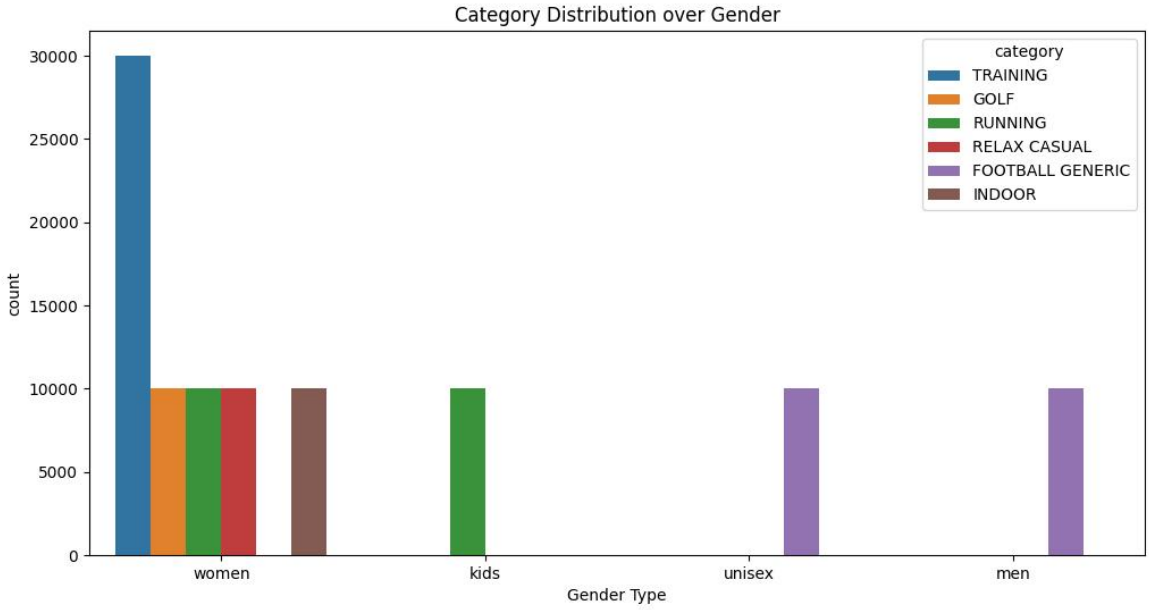
- The main colors are distributed equivally, unlike the secondary color, where rosybrown is the most popular in all of them.

Which main/secondary color has more promos applied during purchases?





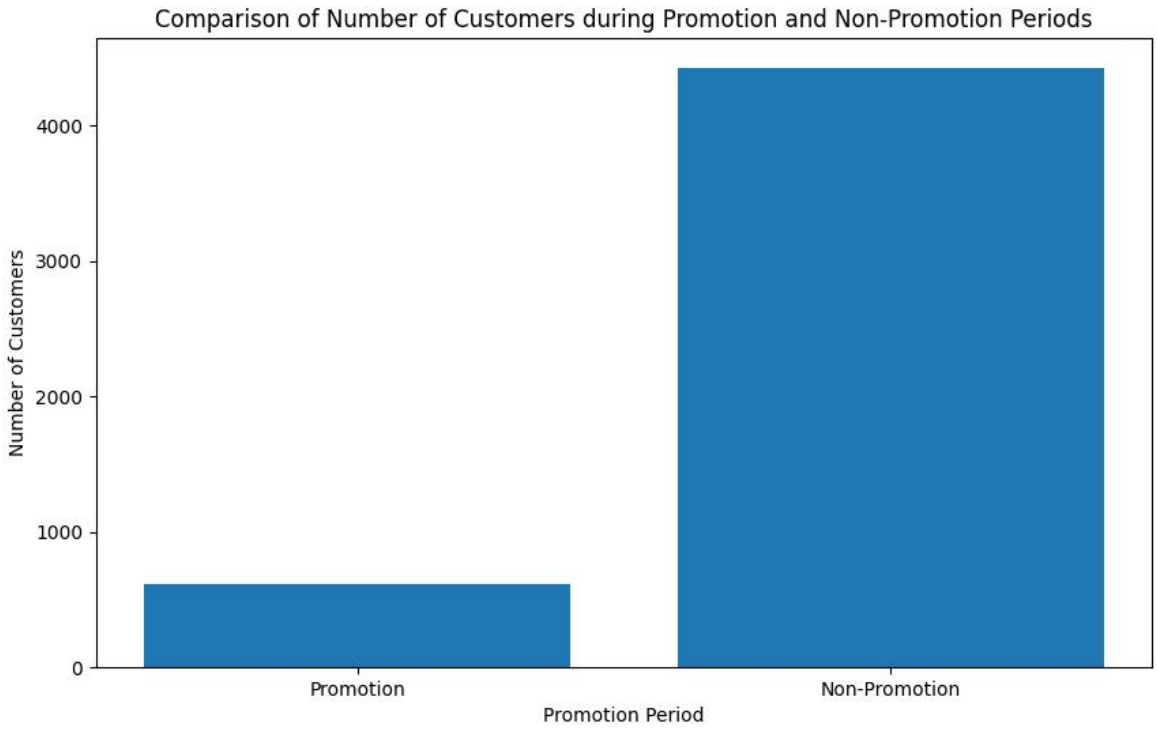
Anlaysis: What is the most popular category by gender?



- Women are the highest gender type who bought almost all the categories, except 'FootBall Generic' category.

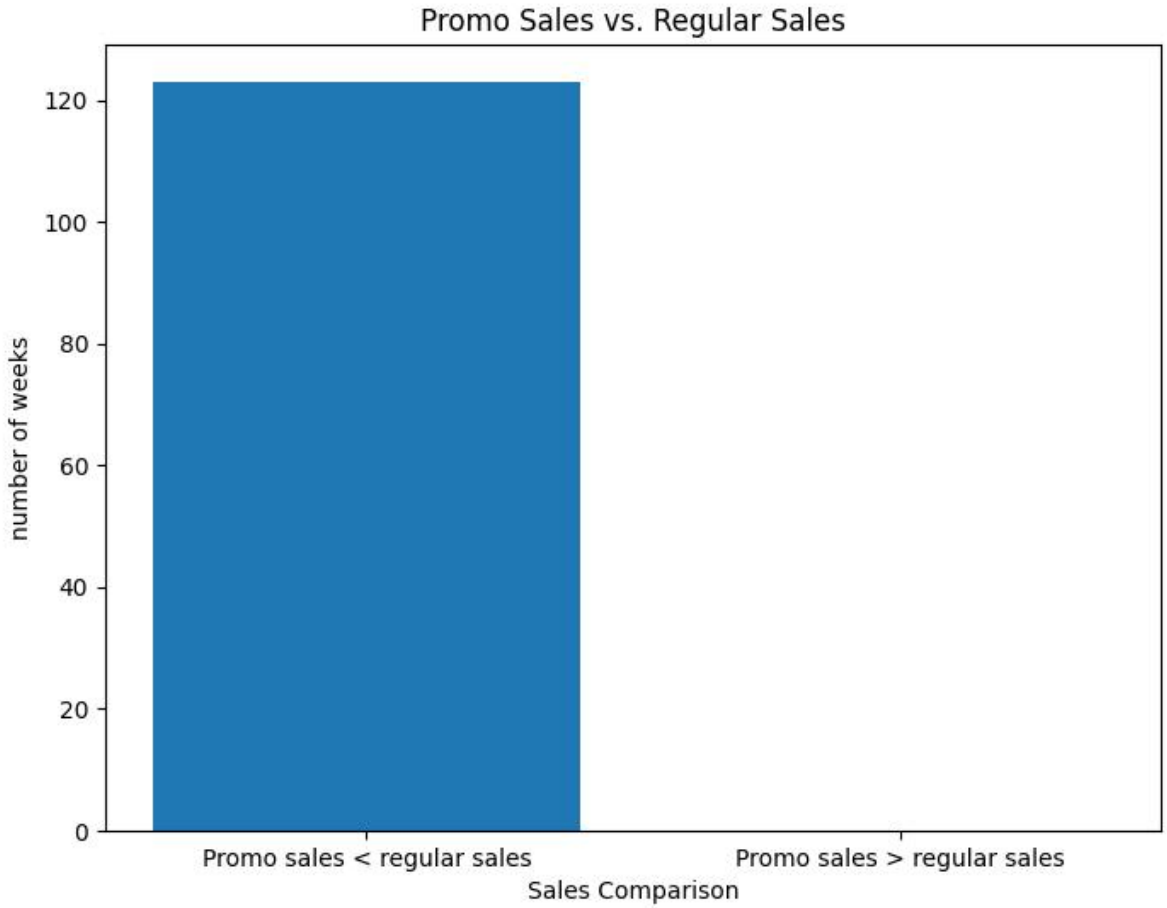
Analysis: Promo Effectiveness

How many week were the promo applied during?



The promo duration were very little, that the impact with not enough to have in customers purchases.

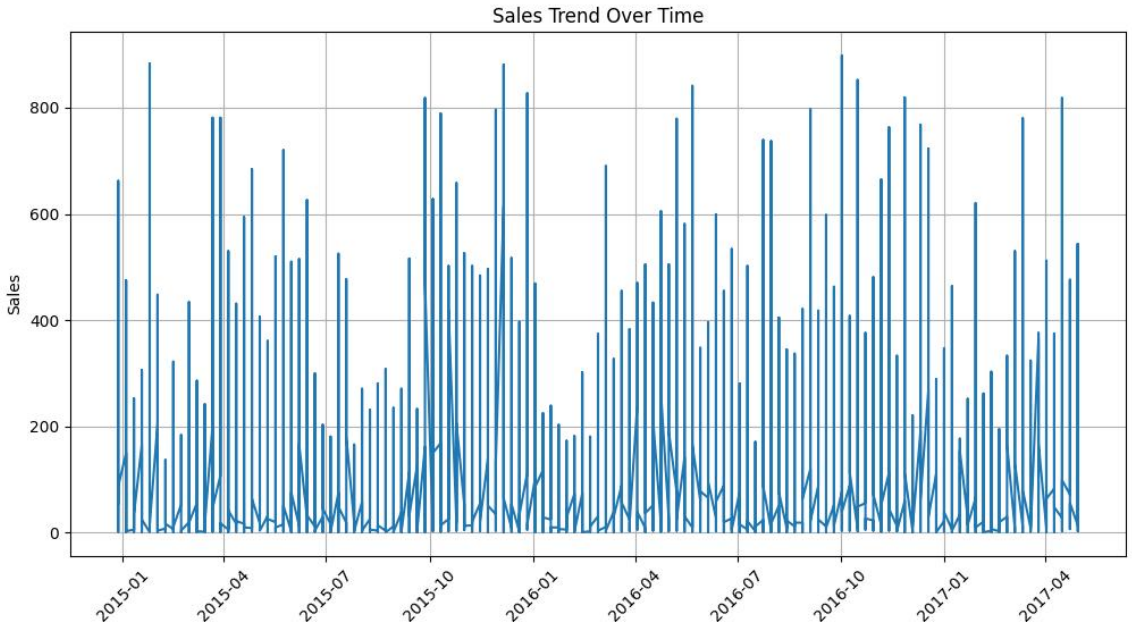
During promo is being applied, which sales were higher due to promo effect or normal regular sales?



This shows that, there are no a signle week where the sales were high because of promo applied.

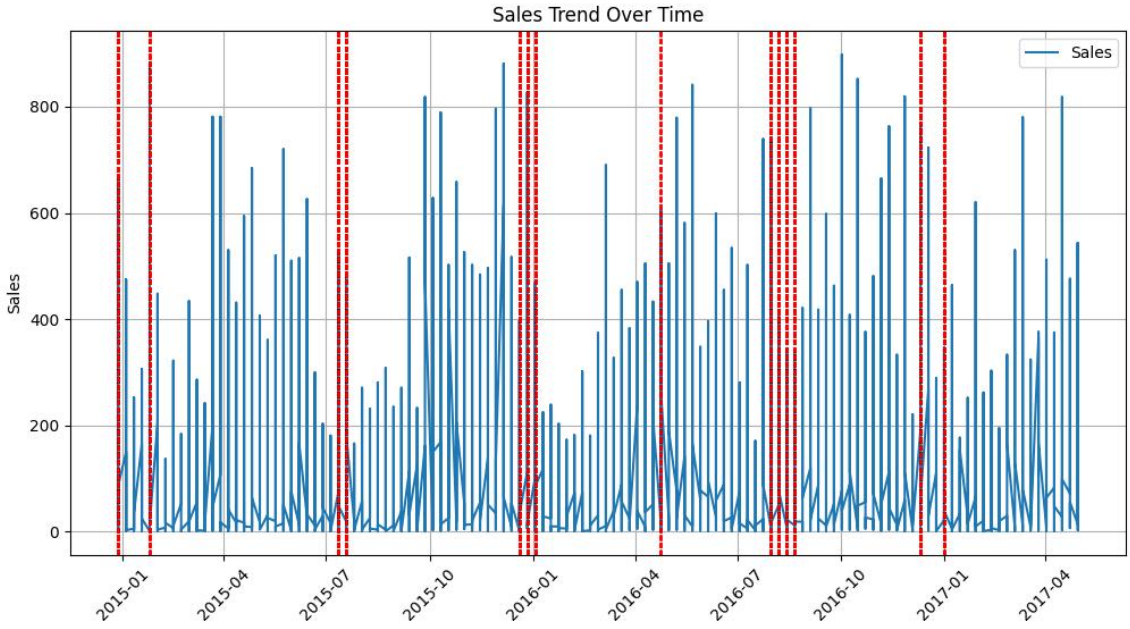
Analysis: Time Series

Sales time series distribution

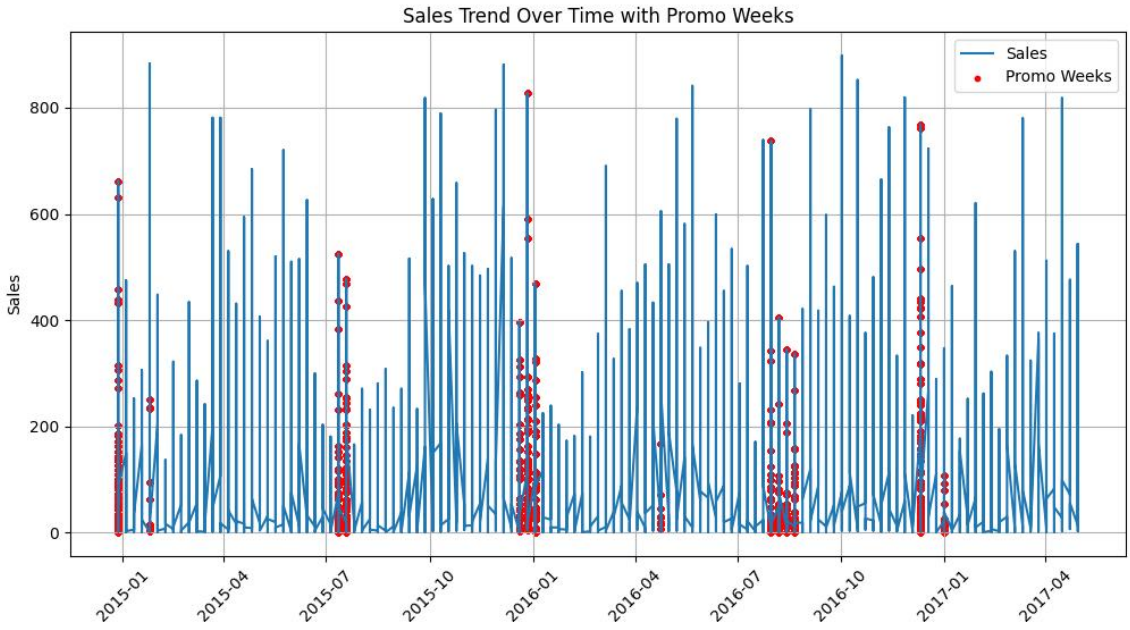


- There are fluctuation over the weeks of year in sales.
- There are some pattern where the sales increase in the summer time, early winter time, and decrease the rest of the year.

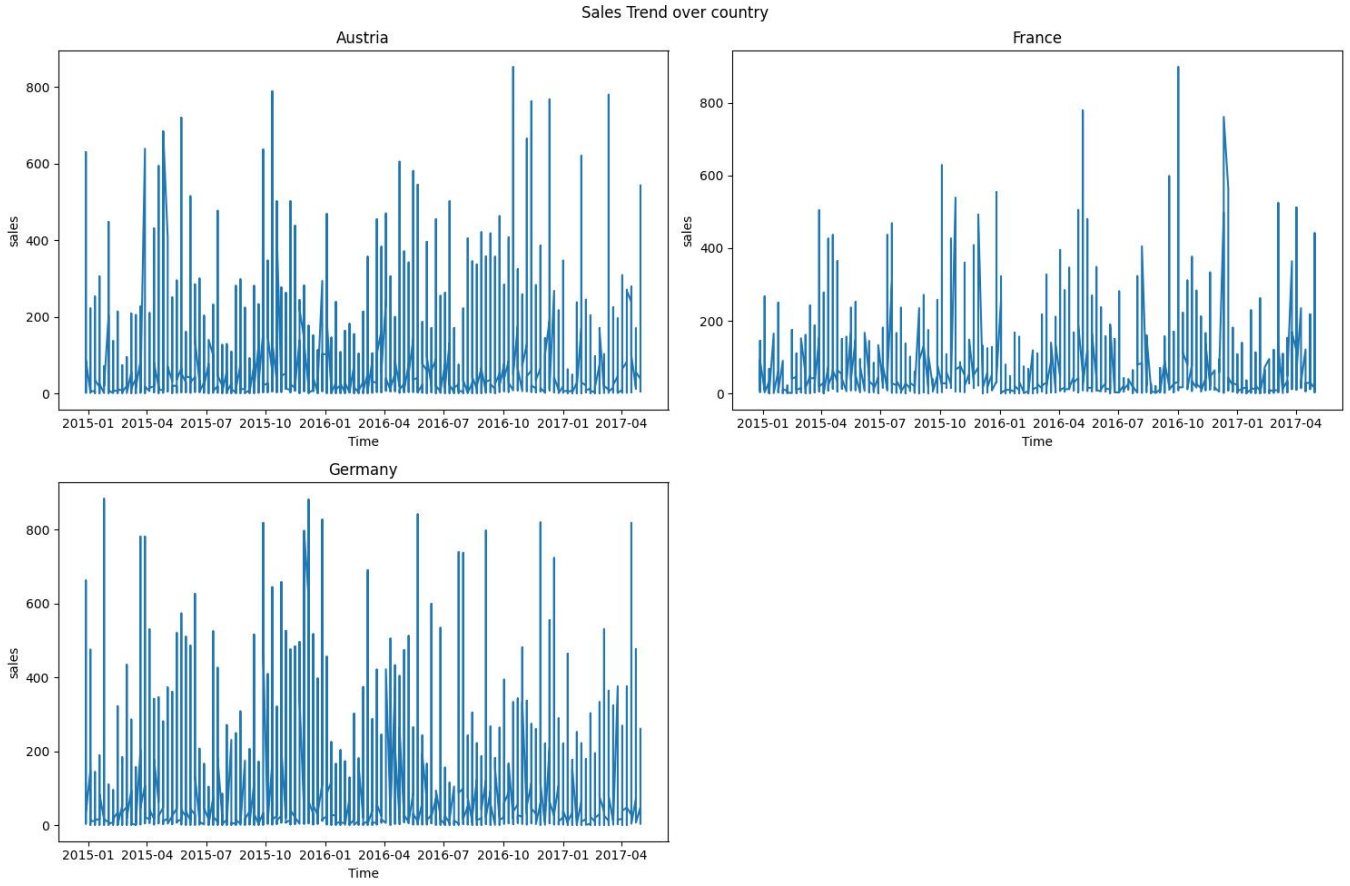
Sales during promo applied weeks



To get better visualization how much the sales were during the weeks applied to:



Sales across each country



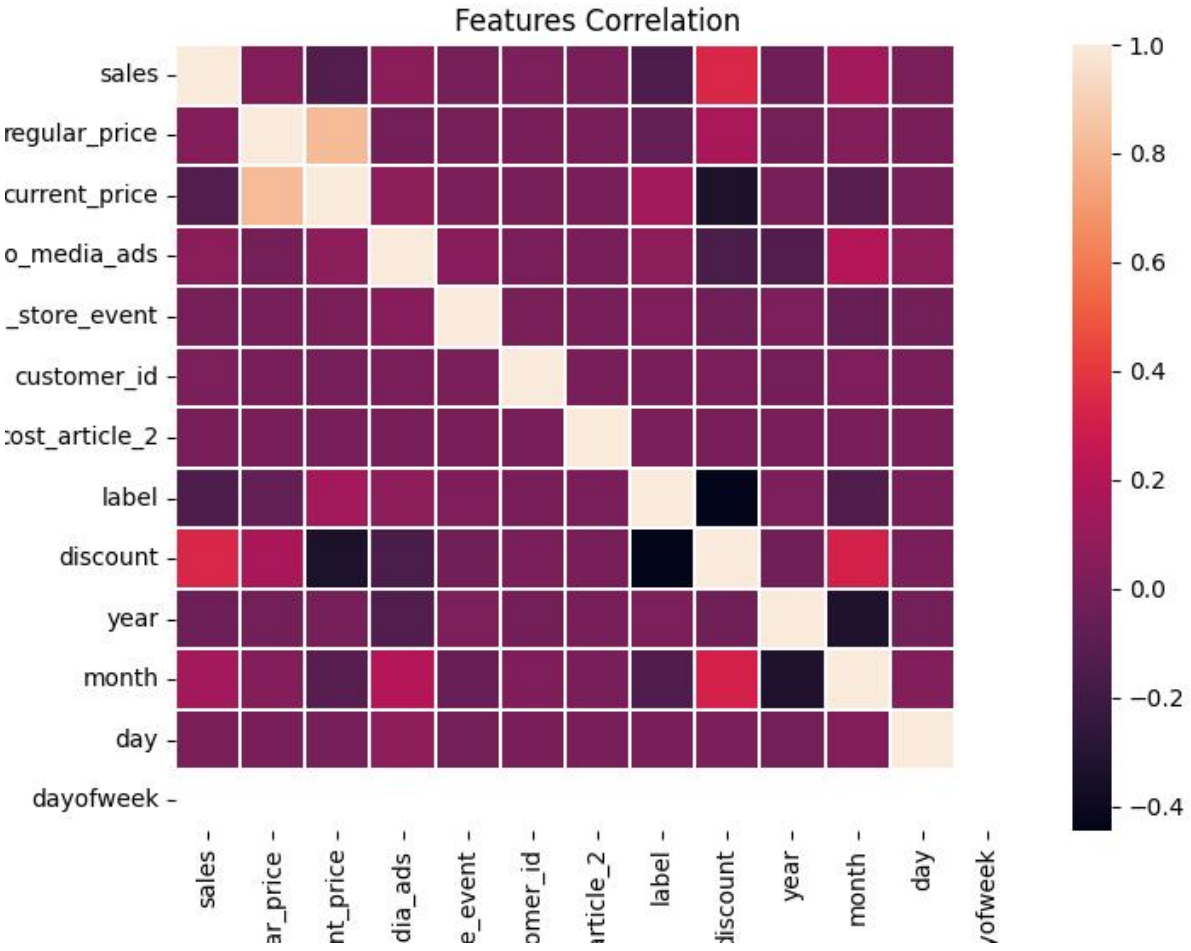
- This shows that, Germany has the mist flucations during each week.

Feature Engineering

Creating additional informative feature will have huge impact in model training process. So the recommened features to engineered are:

Feature	Description
Discount	The discount value for current price of the unit comapred the regular price.
Main Closest Color	The main unit closest color based in the RGB values
Secondary Closest Color	The secondary unit closest color based in the RGB values
Time based Features	Time related features to the timestamp featuers 'retailweek' extracted like [year, month, day, day of week]

Features Correlation



- There is a negative correlation between target column 'label' and discount feature.
- Also a high postive correlation between current price and regular price features.

Model Insights

Prediction the performance of marketing campaign is a classification problem, So one of the most powerfull model is XGBOOST Classification, which was used here in this project.

The overall accuracy were [Training set: 88% , Validation set: 86%]

The confustion metric over the training set is:

	Positive Prediction	**Negative Prediction**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

	Positive Prediction	**Negative Prediction**
Actual Positive	47382	288
Actual Negative	6544	1222

So this model has high precision, but low Recall in case of the detective the postive values, which means when this model detect that the customer will purchase due to the market campaign, it can correctly detect

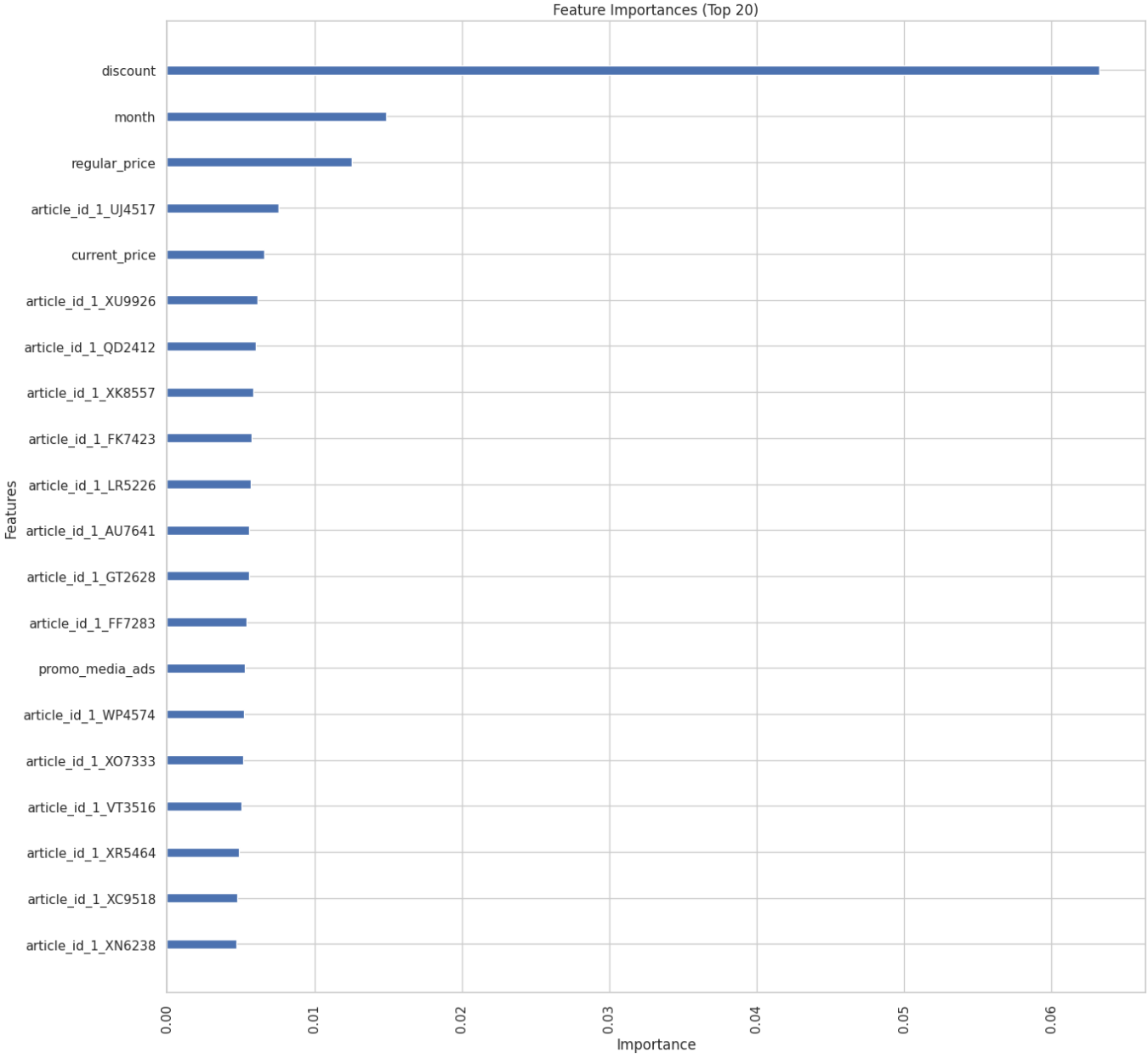
this is a 1 , but can't detect the all customers who actually purchase due to market campain, so it will loss actaully customers.

For model tuning and imporvement, A baseline model will be used to be a baseline for each new developed model, and compare the performance.

Feature Importance

According to the used features after all preprocessing steps inlcuding [feature engineering to create new features, imputing for missing values, scalling for numerical values, one-hot encoding for categorical values], More important features were created.

To get more details regarding the first top 20 features detected by current model.



Recommendations

Based on the analysis conducted, the following recommendations are suggested:

- Regularly monitor and update promos weeks strategies based on the market trends and customers' behaviour.
- Consider the impact of specific features on the marketing campaign performance.
- Continuously collect and analysis data on market and customers preferences, and also each country contribution in each market to maintain a competitive edge.

Conclution

In conclusion, the analysis of the company dataset has provided valuable insights into unit types, countries market share, categories the company have and their impact of each promos they offer. These insights can assist the company in making informed decisions regarding promos values, their time to apply, which to apply to, to whom, and understanding customer preferences. It is crucial for the company to leverage these findings to refine its strategies, maximize profitability, and establish a strong foothold in the competitive market.