# Sports Wear Group Report

By **Ahmed Tarek Ahmed Mohamed**

## Dataset

For this project, I'm using full_gen_data which belongs to the Sports Wear Group.

## 1. Business Understanding

### Problem Definition:

The provided data represents information from a marketing campaign referred to Sports Wear Group. They provided information about the product in the campaign that was sent to a specific customer and the result of this campaign. Based on census features, the machine learning task is to predict if the customer will buy or not as an advertisement result after sending the offer.

### Scope:

- The scope of this sample is to create a binary classification machine learning model which addresses the above prediction problem.
- I follow the stages of the **CRISP** lifecycle and organize documentation and code according to the stages of the lifecycle.

### Metrics:

The performance of the machine learning model will be evaluated on the test set provided. Accuracy is measured and reported using an accuracy score. accuracy of > 0.8 will be considered acceptable and suitable for deployment.

## 2. Data Understanding

- Detailed information about the data is provided in the file attached.
- There are a total of 100000 records in the dataset.
- Probability of the label '0' which means that the customer didn't buy: 86.07%
- Probability of the label '1' which means that the customer buys: 13.93%
- Target: Label column with 2 classes '0' and '1'.
- Features: country, article, sales, cost, sizes, gender, etc.

# 3. Data Preparation

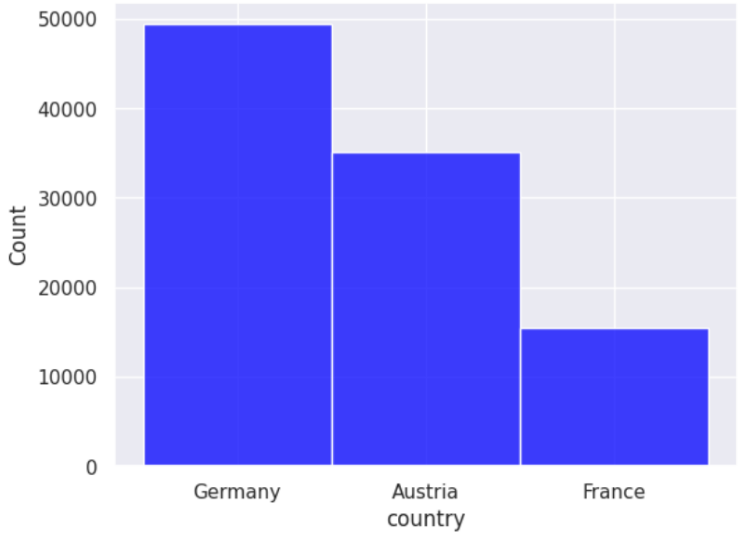
## Data Cleaning:

Data has no Null values, no duplicated rows, and has appropriate data types.

## Exploratory Data Analysis(EDA):

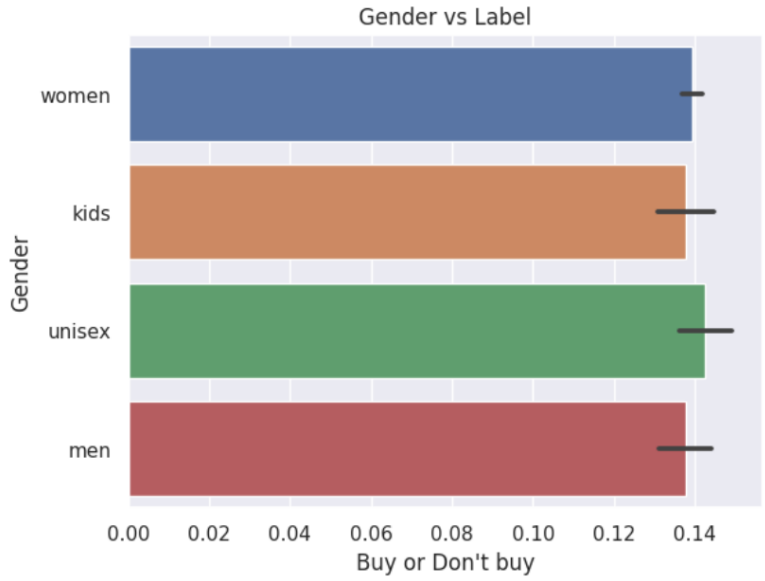
After using One-hot encoding for categorical features using the get dummies method, and applying feature engineering by removing useless columns for training the model. I have done some visualizations of data for better analysis and useful insights.

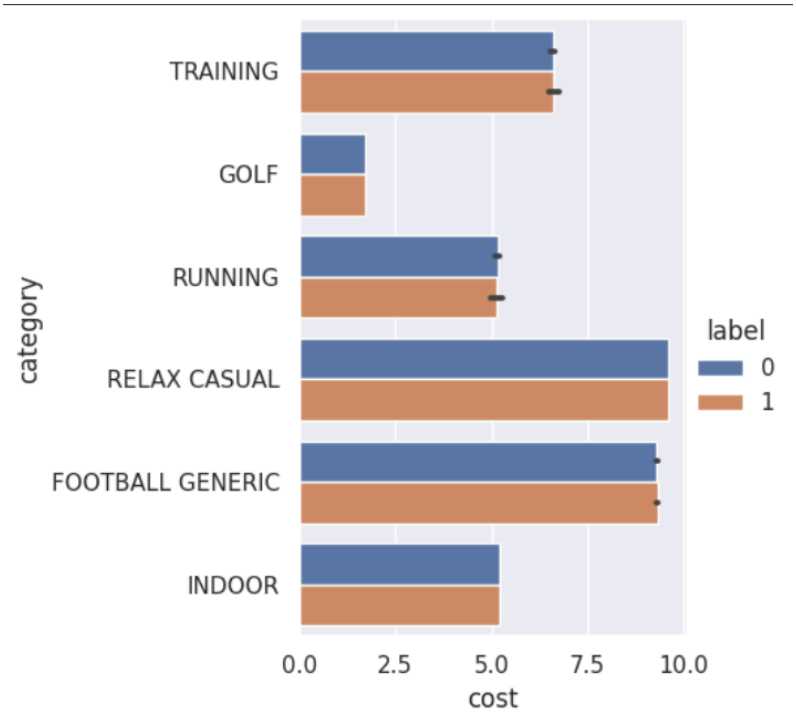
Here are some Insights:

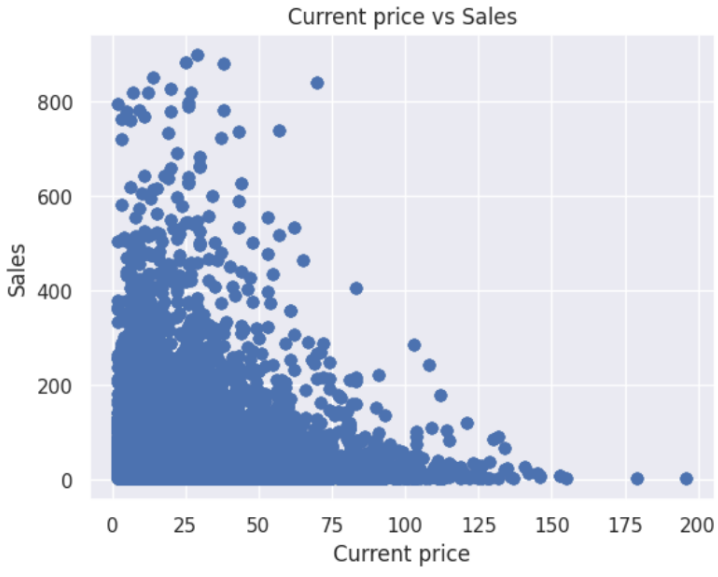



So, Germany is the highest country; while France is the lowest.


It's a relation between the country and whether the customer will buy the products or not.


Gender vs Label
women
kids
unisex
men
Gender
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Buy or Don't buy


TRAINING
GOLF
RUNNING
RELAX CASUAL
FOOTBALL GENERIC
INDOOR
category
label
0
1
0.0
2.5
5.0
7.5
10.0
cost


Current price vs Sales
800
600
400
200
0
Sales
0
25
50
75
100
125
150
175
200
Current price

## 4. Modeling

### Data Splitting:

I used an 80-20% train-test split of data.

### Machine Learning Technique:

I used the Random Forest algorithm from the sci-kit learn library and then trained the model.

## 5. Evaluation

The accuracy of the model was measured using an accuracy score on the test data set. The accuracy score of the Random Forest model was 0.85015 which is > 0.8 which is acceptable.

### Confusion Matrix: