Data Science Final Project

Introduction:

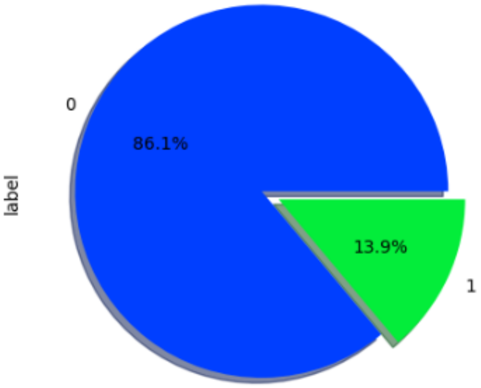
This report presents a comprehensive overview of a data science project conducted on a retail dataset. The project involved exploratory data analysis (EDA) to uncover relationships and patterns within the data. Additionally, feature engineering techniques were applied to convert RGB values into color names, providing more meaningful insights. Furthermore, a predictive model was developed on the label column.

Exploratory Data Analysis (EDA):

The EDA phase involved a thorough examination of the dataset, considering various columns such as country, sales, regular_price, current_price, retailweek, promo1, promo2, productgroup, category, cost, style, sizes, gender, and label.

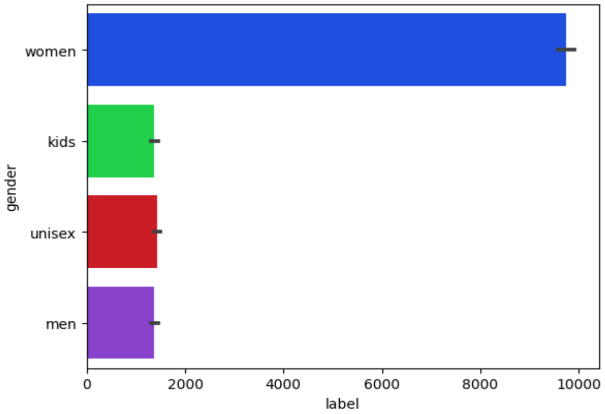
1-Label column:

- Out of **10000**, only around **1390** customers purchased after viewing the campaign, which means the data is unbalanced.
- Only **13.9%** of the total number of views.
- We need to dig down more to get better insights from the data and see which categories of the customer did purchase and who didn't.



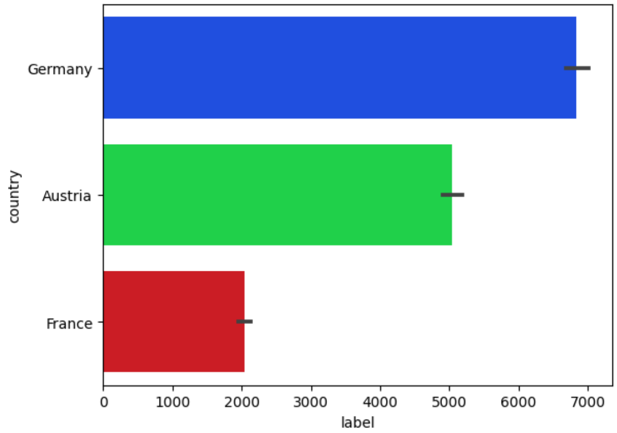
2- Gender column:

- This looks interesting. The number of offers that targeted woman is far greater than men, unisex and kids products.
- The ratio of the user who purchased and who didn't purchase is almost the same **0.16.**
- There is no direct relation between the gender and ratio of sales.



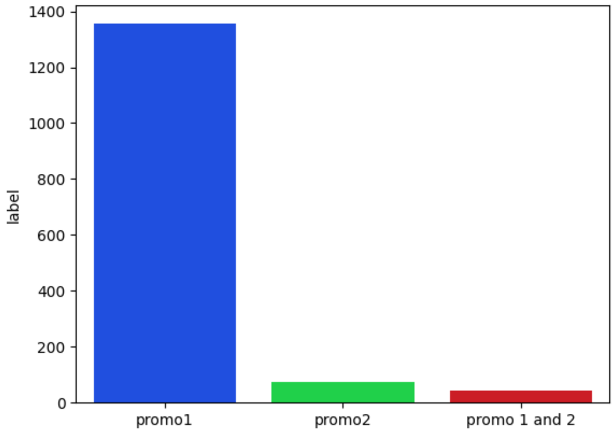
3- Country column:

- The number of offers in both Austria and Germany is greater than France.
- The ratio of the user who purchased and who didn't purchase in every country is almost the same at **0.16.**
- There is no direct relationship between the country name and ratio of sales.



4- Promo1 and Promo2 Columns:

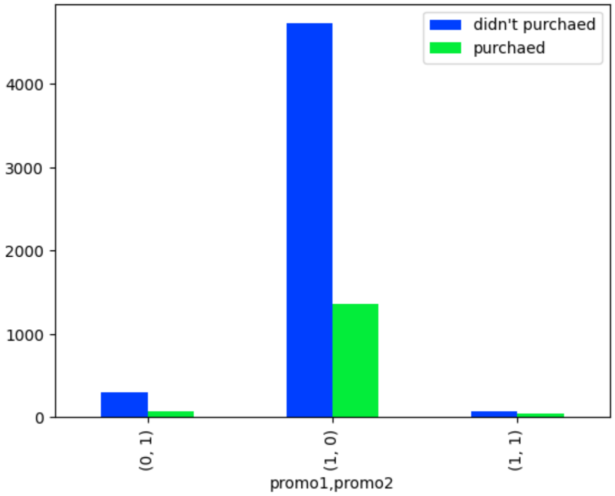
- Promo1 indicator for media advertisement. promo2: indicator for store events
- Promo1 is the most applied promo in the dataset and promo 2 is the least applied promo.



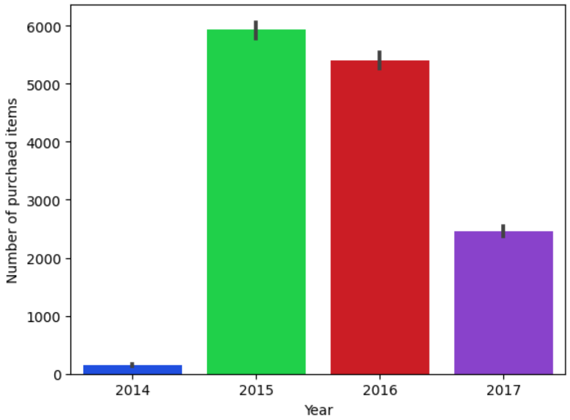
- If no **promos** were active the ratio of customers who bought to those who didn't buy is **0.15**.

  If promo1 was active the ratio would be **0.28**
  If promo2 was active the ratio would be **0.24**,
  If both promos 1 and 2 were active the ratio would be **0.617**.

  Which means that promos have a huge impact on Customers to decide whether to buy or not.

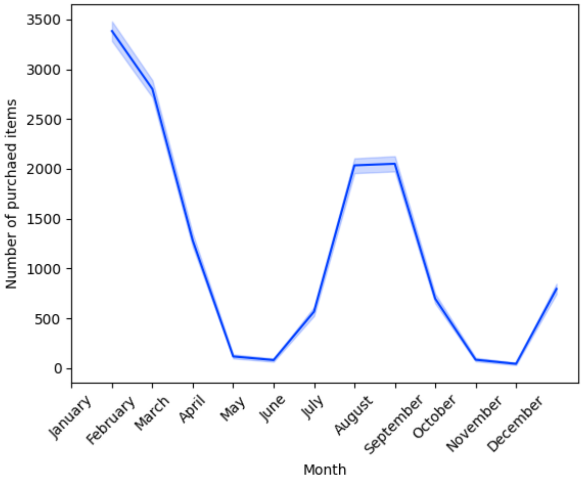


4- RetailWeek Column:

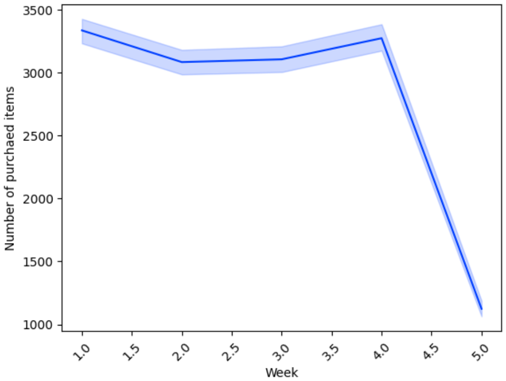


By looking at each retail year we find that year 2014 and 2017 are the lowest years, by returning to the data set we discovered that it started from at the end of 2014 to the middle of 2017 this explains why the distribution of years is looking like this.

And by showing the distribution over months we find that there is a huge increase in January, February and march and then there is a huge decrease in May, June, October and November.



The month in which the customer sees the article has a huge impact on whether he will purchase or not.



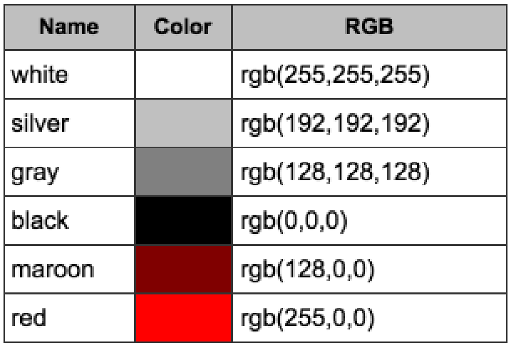
And by showing the distribution over weak numbers in month we find out that there is a huge decrease in at the last weeks of each month.

Week numbers in month also a huge impact on whether he will purchase or not.

In the data set there are columns to represent each channel of the RGB values.

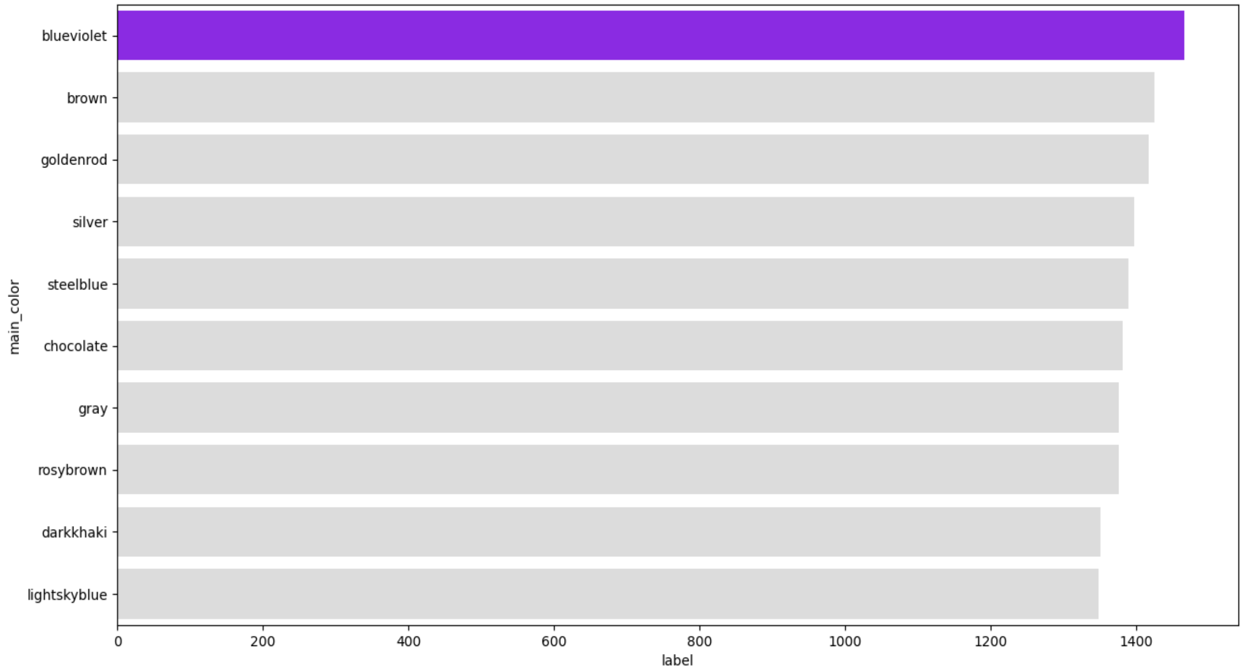
We have 6 columns 3 of them represent the main colors, the other 3 columns represent the secondary color, It’s a bit difficult to work with colors as RGB values it will be very hard to categorize and present in EDA.

Instead, we will convert them to names using **webnames** library in python.

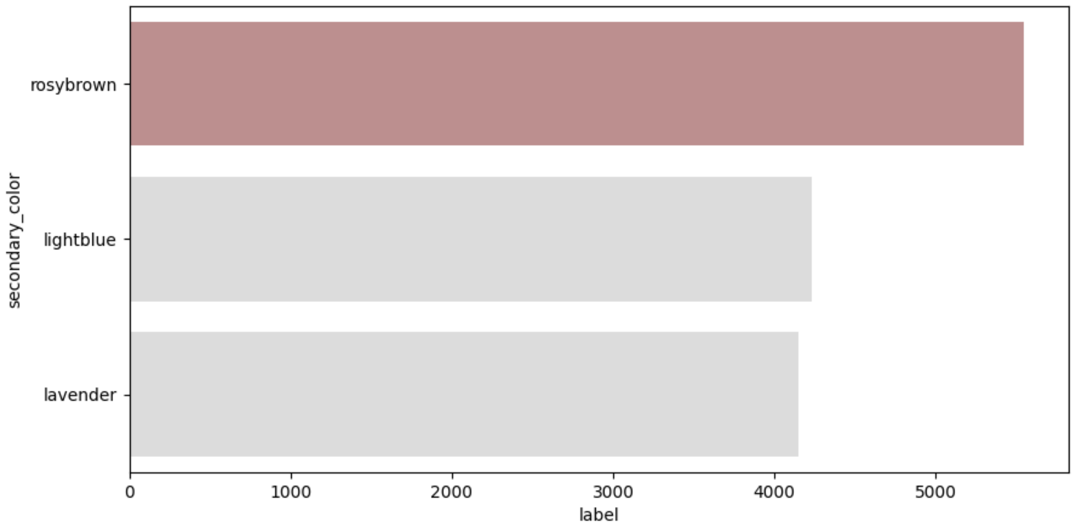


Name	Color	RGB
white	⬜	rgb(255,255,255)
silver	░	rgb(192,192,192)
gray	▒	rgb(128,128,128)
black	⬛	rgb(0,0,0)
maroon	🟫	rgb(128,0,0)
red	🟥	rgb(255,0,0)

After converting the RGB columns to name in main color we will be left with 10 color names.



After converting the RGB columns to names in secondary color we will be left with 3 color names.

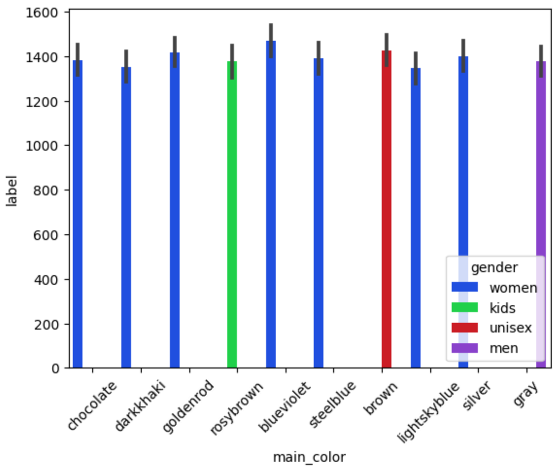


From the previous two plots we find that the main color that have higher sales Is the blue violet and for the secondary color it's the rosy brown

Main color column and gender:

We will see that men, kids, unisex product have only one single color, but for the women product it has 7 different colors.

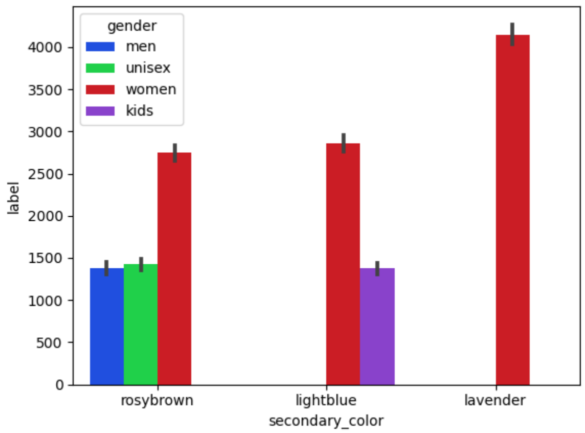
Blue violet is the color with the most purchases for female products.



Sec color column and gender:

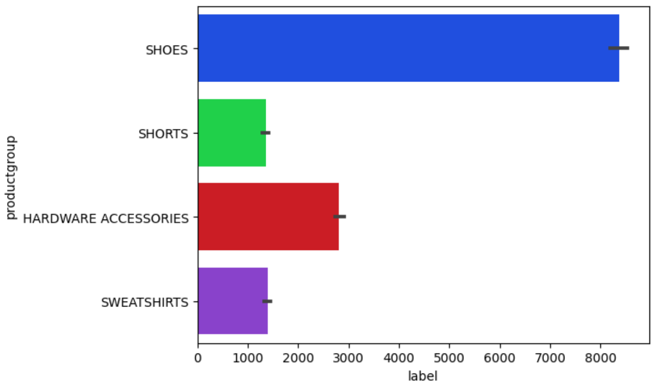
Same as the main color's men, kids, unisex products have only one color.

Lavender is the best color to be used in the women products.



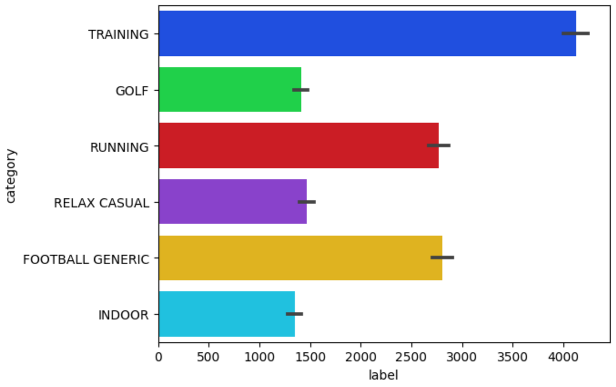
5- Product group column:

The Ratio is almost the same in each category.

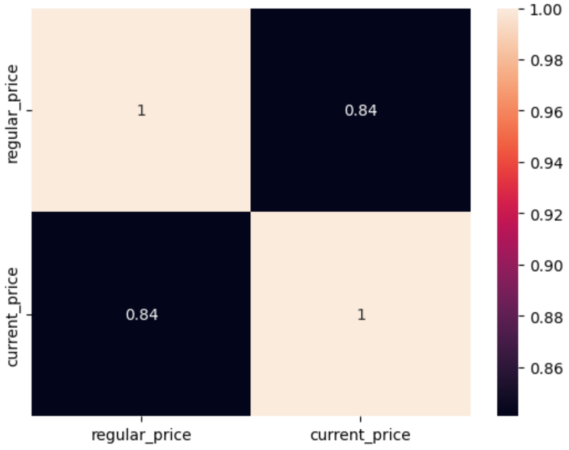


6- Category column

* The Ratio is almost the same in each category except for Relax Casual is slightly higher.



7- Regular price and current Price:



There is huge correlation between the two column so we can replace them with only one column.

Saving_amount = Regular price - current Price

Model Development and Evaluation:

Multiple machine learning models were explored to predict customer behavior, class imbalance between the target classes was addressed through oversampling techniques.

-Model Selection and Evaluation: Several machine learning models, such as Logistic Regression, Random Forest, and others, were trained and evaluated on the dataset. Evaluation metrics such as accuracy, precision, recall, and F1 score were used to assess the models' performance.

- Addressing Class Imbalance: As the dataset exhibited class imbalance in the target variable, oversampling techniques, such as Random Oversampling or SMOTE (Synthetic Minority Over-sampling Technique), were applied to balance the classes. This step aimed to improve the model's ability to accurately predict both classes.

- Best Performing Model: After evaluating various models, the Random Forest Classifier emerged as the best-performing model, achieving an accuracy of 0.953 . This model showcased the ability to effectively capture patterns and make accurate predictions in the retail dataset.

Test set accuracy using various models and techniques:

	No enhancements	Random over sampling	SMOTE over sampling	PCA + Random over sampling
Logistic Regression	0.86	0.78	0.89	0.79
KNN classifier	0.84	0.84	0.81	**0.84**
Random forest	0.85	0.95	0.91	**0.953**

Conclusion:

In conclusion, this data science project focused on analyzing a retail dataset, emphasizing EDA, feature engineering, and model development. Through EDA, we gained insights into the relationships between sales and pricing/promotion variables. Feature engineering techniques, including the conversion of RGB values to color names, added interpretability to the dataset. The developed predictive model, incorporating relevant features and applying feature selection methods, to predict customer behavior.