
Predict Customer
Purchase
Behavior

ABOUT ME

Nguyễn Quốc Bảo

(DA & BI Analytics)

INDEX

DATA OVERVIEW & CLEANING

- Introduction to Database
- Overview of row and column data
- Check for duplicates, empty…

BUILDING MODEL

- Building and evaluating machine learning models

EXPLORATORY DATA ANALYSIS

- Perform distribution analysis of each feature (univariate analysis)

CONCLUSION

- Based on the most effective model, understand the factors that influence purchases
- Make future recommendations

TARGET

- Understand customer target groups, ages, best-selling products, and other factors such as online websites, advertising programs, etc.

- Predict the likelihood of a customer making a purchase based on personal characteristics and purchasing behavior.

- Propose appropriate campaigns for each type of company's future customers.

DATA OVERVIEW & CLEANING

OVERVIEW & CLEANING

DATA DICTIONARY

Age	Age of Customer
Gender	Gender of Customer(0: Male, 1: Female)
Annual Income	Customer annual income($)
Number of Purchases	Order Quantity
Product Category	List of products purchased (0: Electronics, 1: Clothings, 2: Household goods, 3: Beauty Products, 4: Sportswear)
Time Spent on Website	Time Customer spent on Website(Minutes)
Loyalty Program	Does Customer is Loyal or Not(0: NO, 1: YES)
Discount Availed	Coupon/Promotes Customer used (Range: 0-5)
Purchase Status(Target Variable)	Customer purchasing ability(0: NO, 1: YES)

OVERVIEW & CLEANING

```
df = pd.read_csv('/content/customer_purchase_data.csv')

print("Dataset shape of df is", df.shape) #Total Row & Column
df
```

```
Dataset shape of df is (1500, 9)
```

	Age	Gender	AnnualIncome	NumberOfPurchases	ProductCategory	TimeSpentOnWebsite	LoyaltyProgram	DiscountsAvailed	PurchaseStatus
0	40	1	66120.267939	8	0	30.568601	0	5	1
1	20	1	23579.773583	4	2	38.240097	0	5	0
2	27	1	127821.306432	11	2	31.633212	1	0	1
3	24	1	137798.623120	19	3	46.167059	0	4	1
4	31	1	99300.964220	19	1	19.823592	0	0	1
...	...	...	...	...	...	...	...	...	...
1495	39	1	65048.141834	13	0	34.590743	0	5	1
1496	67	1	28775.331069	18	2	17.625707	0	1	1
1497	40	1	57363.247541	7	4	12.206033	0	0	0
1498	63	0	134021.775532	16	2	37.311634	1	0	1
1499	50	0	52625.665974	13	0	25.348017	1	4	1

1500 rows × 9 columns

 # Read CSV File to check dataset

OVERVIEW & CLEANING

```
# Statistical summary
df.describe()
```

	Age	Gender	AnnualIncome	NumberOfPurchases	ProductCategory	TimeSpentOnWebsite	LoyaltyProgram	DiscountsAvailed	PurchaseStatus
count	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.00000
mean	44.298667	0.504667	84249.164338	10.420000	2.012667	30.469040	0.326667	2.555333	0.43200
std	15.537259	0.500145	37629.493078	5.887391	1.428005	16.984392	0.469151	1.705152	0.49552
min	18.000000	0.000000	20001.512518	0.000000	0.000000	1.037023	0.000000	0.000000	0.00000
25%	31.000000	0.000000	53028.979155	5.000000	1.000000	16.156700	0.000000	1.000000	0.00000
50%	45.000000	1.000000	83699.581476	11.000000	2.000000	30.939516	0.000000	3.000000	0.00000
75%	57.000000	1.000000	117167.772858	15.000000	3.000000	44.369863	1.000000	4.000000	1.00000
max	70.000000	1.000000	149785.176481	20.000000	4.000000	59.991105	1.000000	5.000000	1.00000

#Assess data quality, identify necessary preprocessing steps & gain a preliminary understanding of the distribution of variables

```
# Dataset information
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Age                 1500 non-null   int64  
 1   Gender              1500 non-null   int64  
 2   AnnualIncome        1500 non-null   float64
 3   NumberOfPurchases   1500 non-null   int64  
 4   ProductCategory     1500 non-null   int64  
 5   TimeSpentOnWebsite  1500 non-null   float64
 6   LoyaltyProgram      1500 non-null   int64  
 7   DiscountsAvailed    1500 non-null   int64  
 8   PurchaseStatus      1500 non-null   int64  
dtypes: float64(2), int64(7)
memory usage: 105.6 KB
```

```
df.drop_duplicates(inplace = True)
```

```
df.isna().sum()
```

	0
Age	0
Gender	0
AnnualIncome	0
NumberOfPurchases	0
ProductCategory	0
TimeSpentOnWebsite	0
LoyaltyProgram	0
DiscountsAvailed	0
PurchaseStatus	0

dtype: int64

OVERVIEW & CLEANING

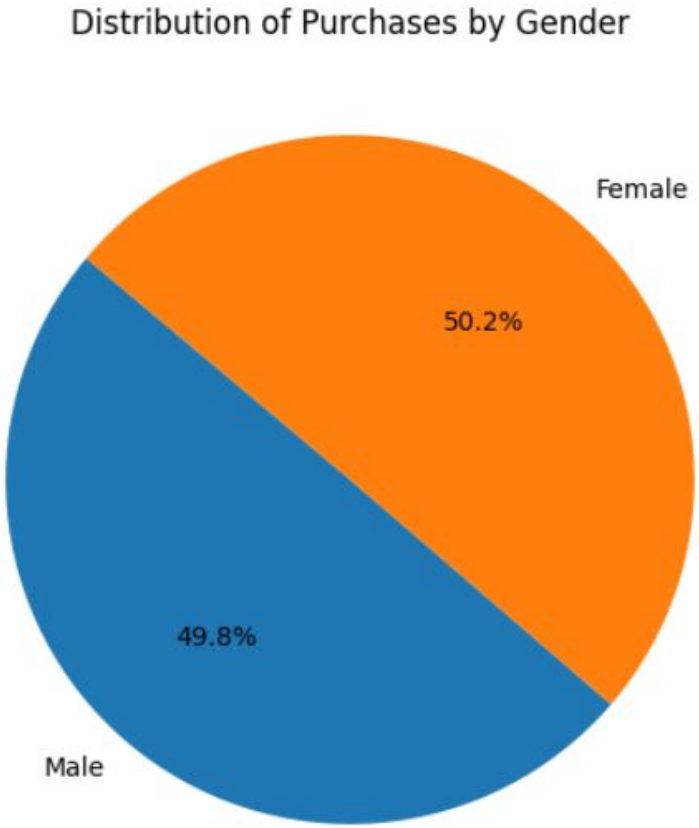
```
import numpy as np
import pandas as pd
from copy import deepcopy
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix
from sklearn.preprocessing import LabelEncoder, Normalizer
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report, precision_score, recall_score, f1_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import roc_curve, auc
from sklearn.neighbors import KNeighborsClassifier
```

 # Prepare the working environment with the necessary tools for analysis and predictive modeling.

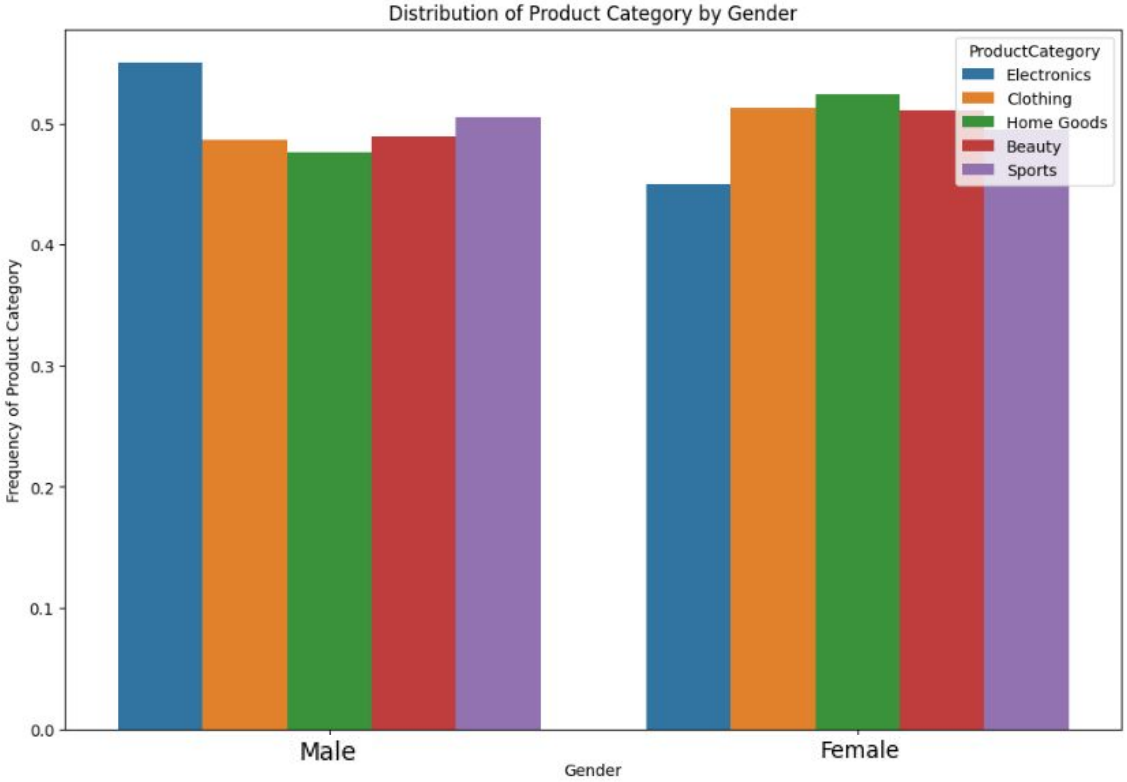

Exploratory
Data Analysis

EXPLORATORY DATA ANALYSIS



- The ratio between the two groups (0: Male and 1: Female) is quite balanced, approximately 50/50.
- This state of affairs shows that there is no such thing as a gender-specifics that needs to be focused on

EXPLORATORY DATA ANALYSIS

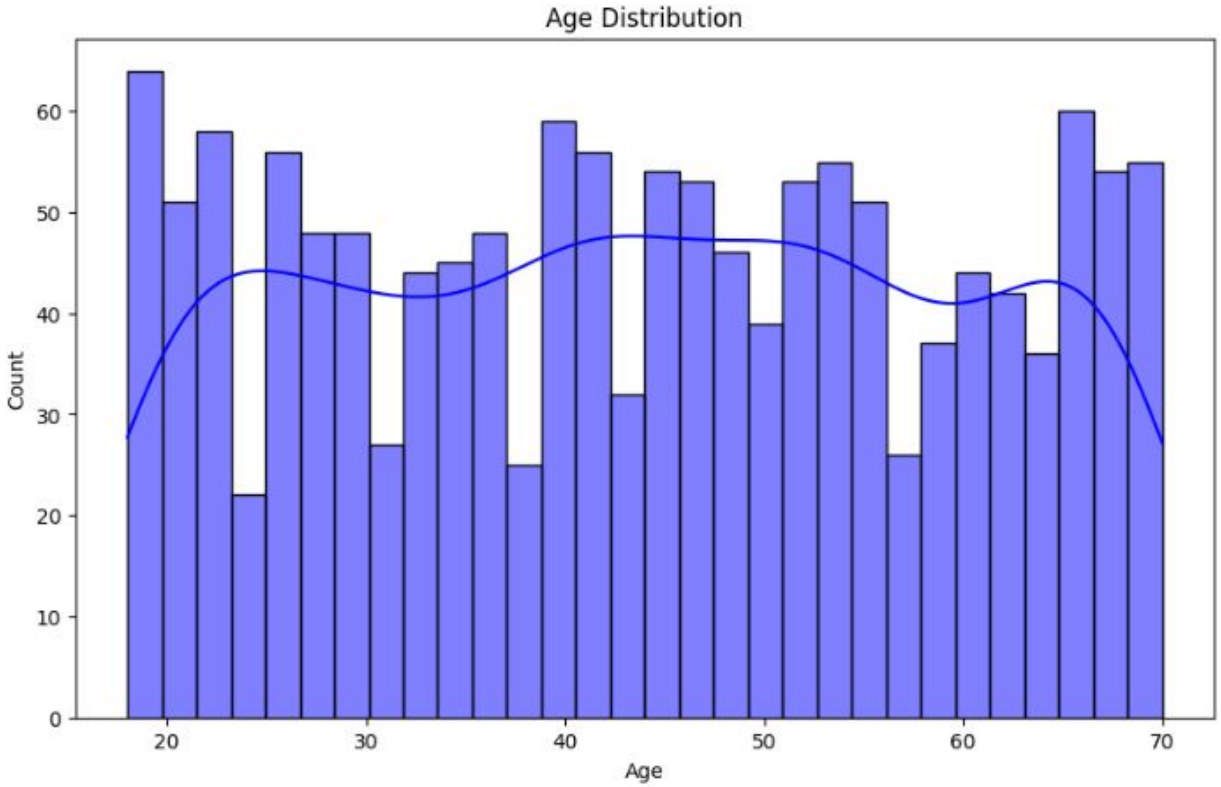


A special difference in shopping behavior between the two genders: Women focus on beautiful clothes and decorations, household items. While men prioritize Electronics and Actionable Toys.

=> Interests and lifestyles are clearly 2 genders

Marketing strategy: Based on this analysis, we can adjust the marketing strategy to suit each customer group by gender to focus on customer needs to promote business development.

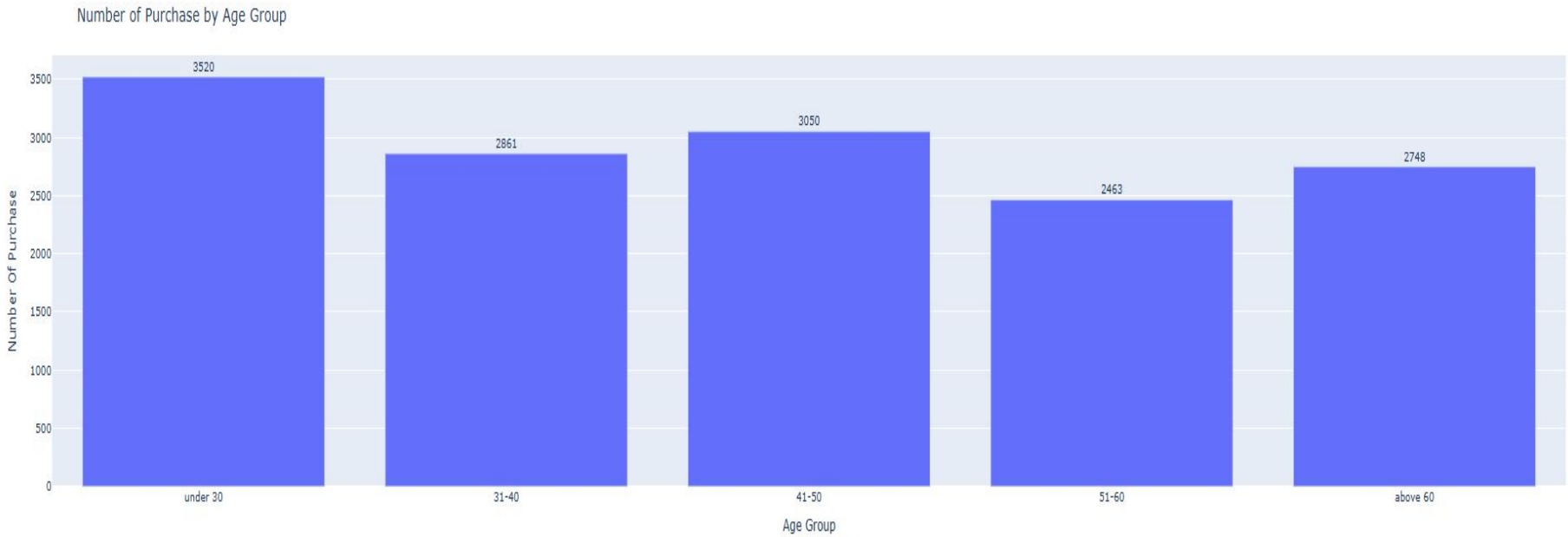
EXPLORATORY DATA ANALYSIS



- The age of the customers ranges from about 18 to 70.

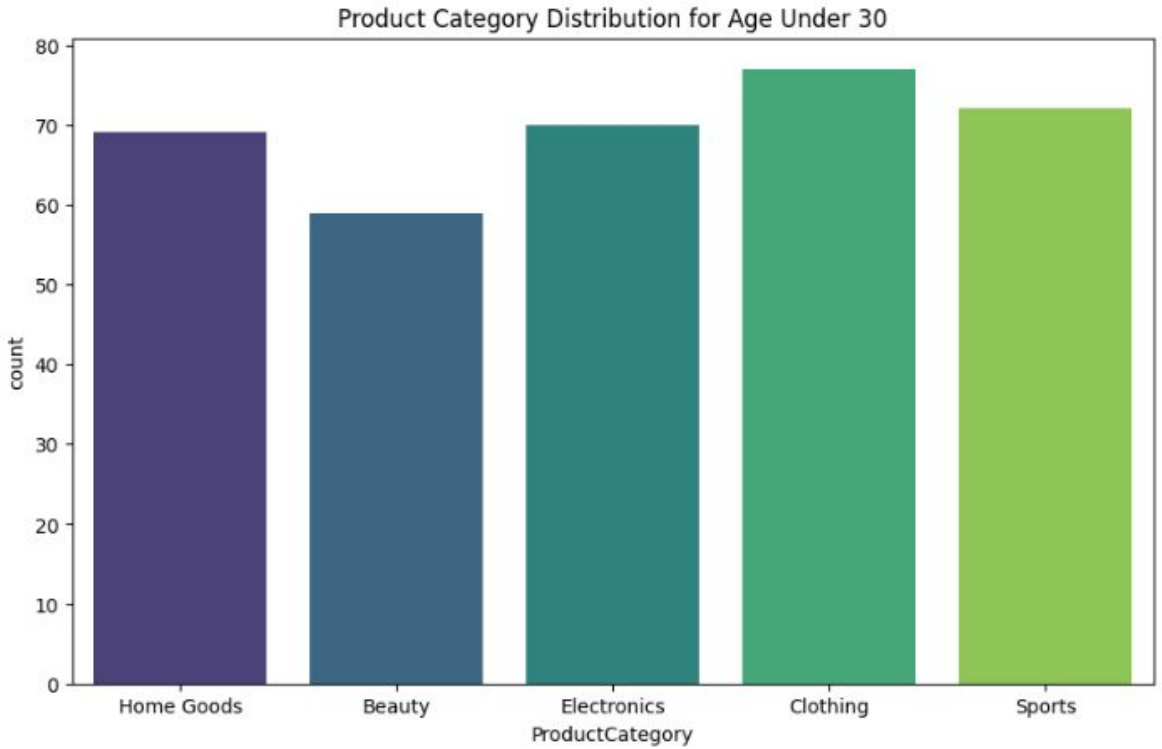
- It can be seen that some age groups have larger numbers of people, for example around 20 - 40 years old and 60 - 70 years old.

EXPLORATORY DATA ANALYSIS



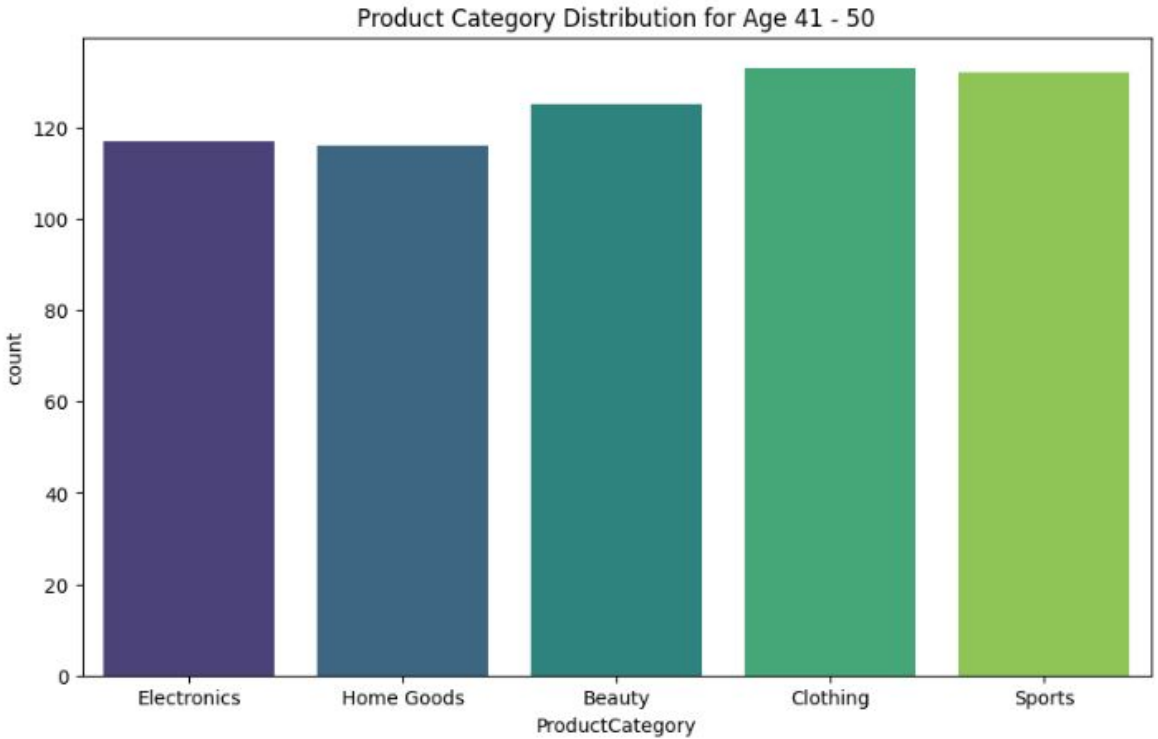
- The **Under 30** age group tends to buy the most, while the **51-60** age group buys the least.
- **The middle-aged and older age** groups have an average and fairly even number of purchases.

EXPLORATORY DATA ANALYSIS



- Customers **under 30** prefer products in the **Clothing and Sports** categories.

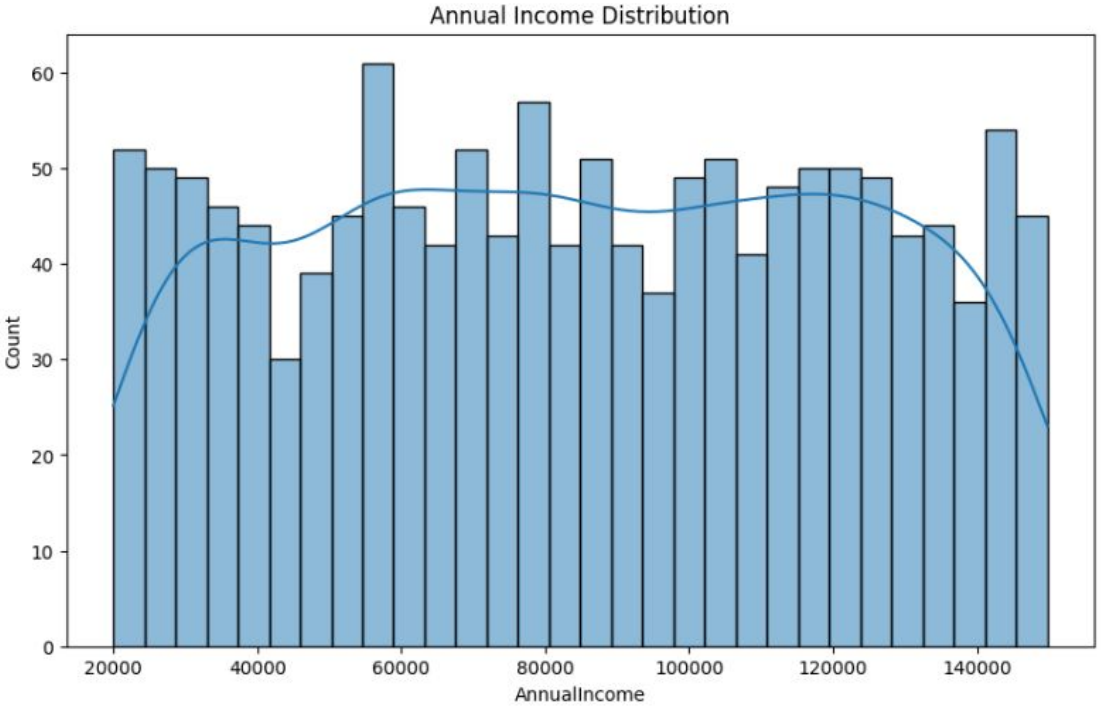
EXPLORATORY DATA ANALYSIS



- Similar to the **Under 30** Group, the **41 - 50** customer group is also interested in products from the **Clothing & Sport Category**

- It is noteworthy that the overview of other product lines is also very high, with almost no difference.

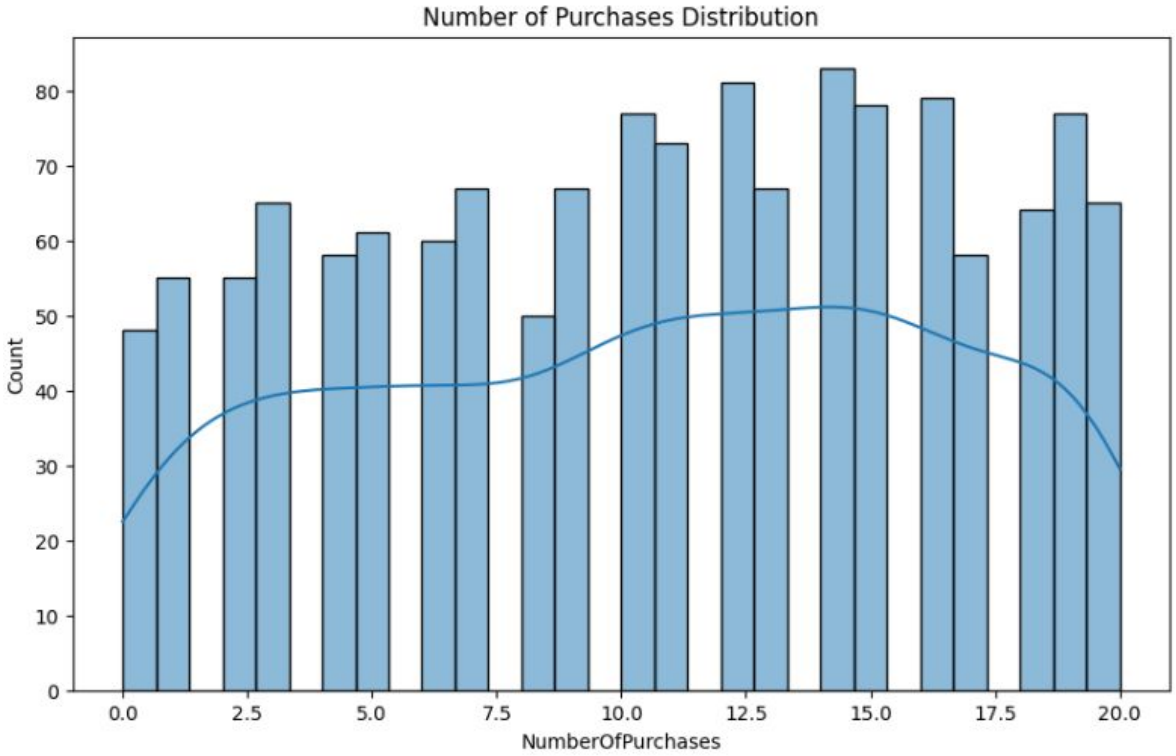
EXPLORATORY DATA ANALYSIS



- Income data ranges from **20,000 - 140,000**

- Customer income is mainly in the low and middle income ranges, specifically around the ranges of **20,000 - 40,000, 50,000 - 60,000, and 100,000 - 120,000.**

EXPLORATORY DATA ANALYSIS

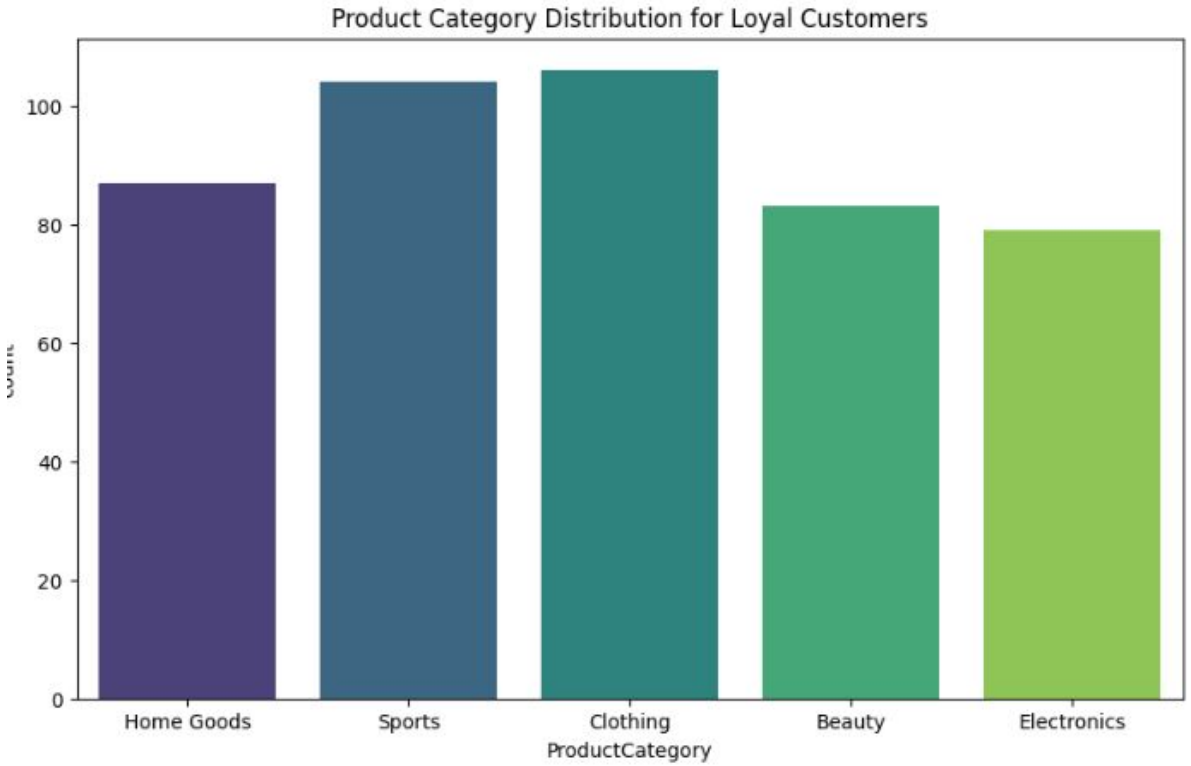


Some groups of purchase frequency are concentrated in the **0 - 2 purchase range** (new customers or low purchase), and a large number of loyal customers are in the **10 - 15 purchase range**.

"The analysis shows that the majority of customers purchase between **10-15 times**. If combined with age analysis, we see that the group of customers **under 30 years old** accounts for the majority of this group.

=> This shows that marketing strategies are and should be targeting this age group

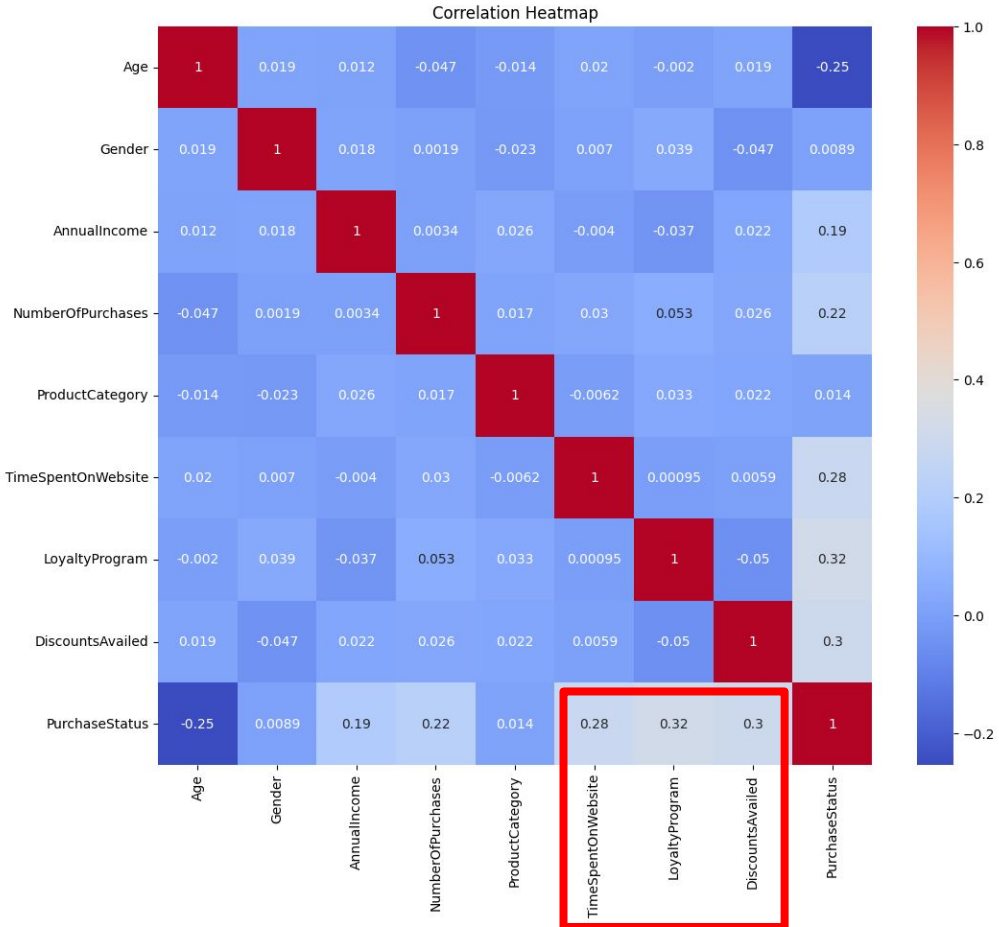
EXPLORATORY DATA ANALYSIS



- Further analysis of the **Loyal Customers** group, we can see that the two product categories **Clothing & Sports** are quite high.

- This also leads to the conclusion that the **Loyal Customer** group mainly belongs to the **two segments Under 30 and Age 41 - 50**

EXPLORATORY DATA ANALYSIS



- **Purchase Status** is positively correlated with: **Loyalty Program (0.32) ; Discounts Availed (0.3) ; TimeSpentOnWebsite (0.28)**

- **Purchase Status** is negatively correlated with **Age (-0.25)**

BUILDING MODEL

BUILDING MODEL

```
# Standardizing the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

print("Scaled data:")
print(X_scaled)
```

```
Scaled data:
[[-0.25445635  0.9971223  -0.49506956 ... -0.01054426 -0.7074889
   1.40669061]
 [-1.54628306  0.9971223  -1.62864811 ...  0.44149835 -0.7074889
   1.40669061]
 [-1.09414371  0.9971223   1.14908091 ...  0.05218787  1.41344974
  -1.53557366]
 ...
 [-0.25445635  0.9971223  -0.72841829 ... -1.09255797 -0.7074889
  -1.53557366]
 [ 1.23114435 -1.00288601  1.3143051  ...  0.38678871  1.41344974
  -1.53557366]
 [ 0.391457   -1.00288601 -0.85466085 ... -0.31816703  1.41344974
   0.81823776]]
```

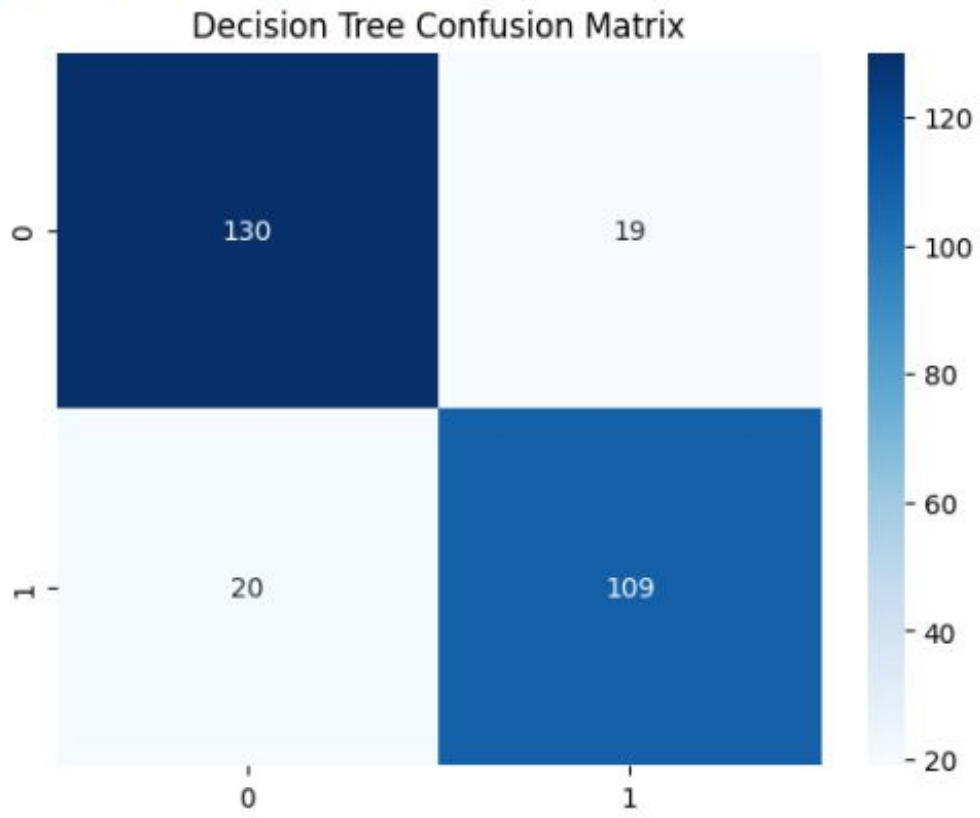
```
# Split the data into training and testing sets
x_train, x_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)
```

 # Training model

```
# Decision Tree Classifier
from sklearn.tree import DecisionTreeClassifier
dt_model = DecisionTreeClassifier(random_state=42)
dt_model.fit(x_train, y_train)
y_pred_dt = dt_model.predict(x_test)
```

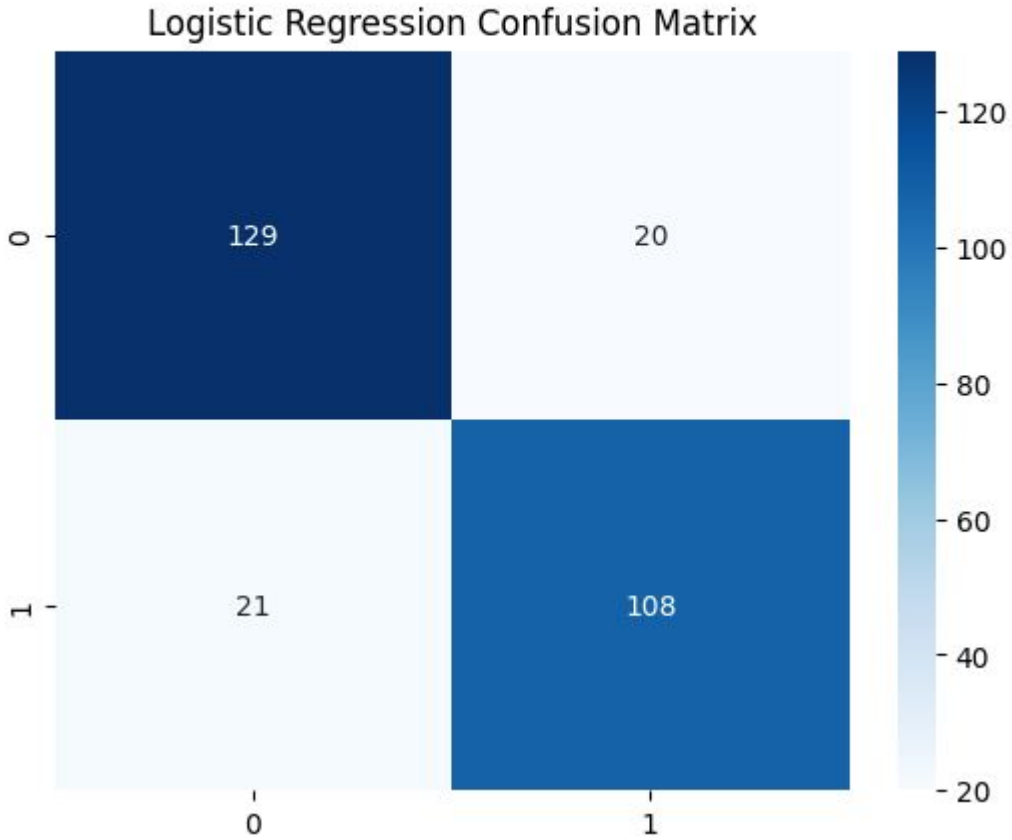
```
# Decision Tree Evaluation
accuracy_dt = accuracy_score(y_test, y_pred_dt)
precision_dt = precision_score(y_test, y_pred_dt)
recall_dt = recall_score(y_test, y_pred_dt)
f1_dt = f1_score(y_test, y_pred_dt)
cm_dt = confusion_matrix(y_test, y_pred_dt)
```

BUILDING MODEL



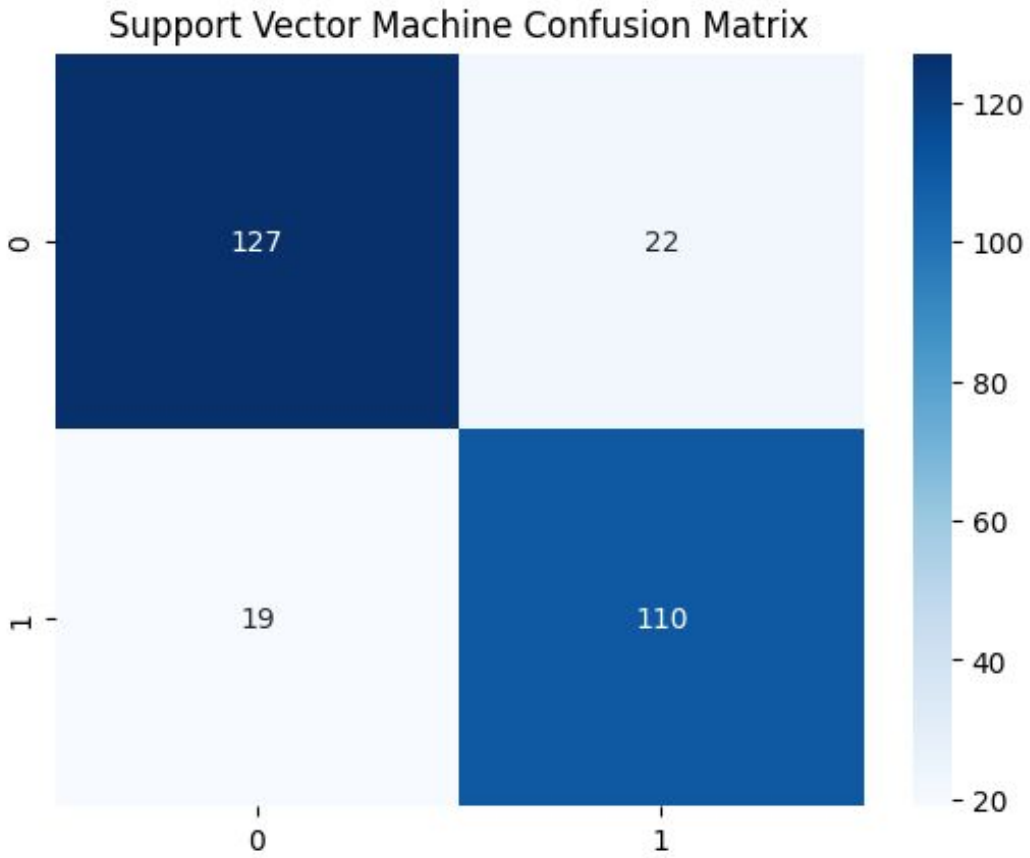
Decision Tree Classifier:
Accuracy: 0.8597122302158273
Precision: 0.8515625
Recall: 0.8449612403100775
F1 Score: 0.8482490272373541

BUILDING MODEL



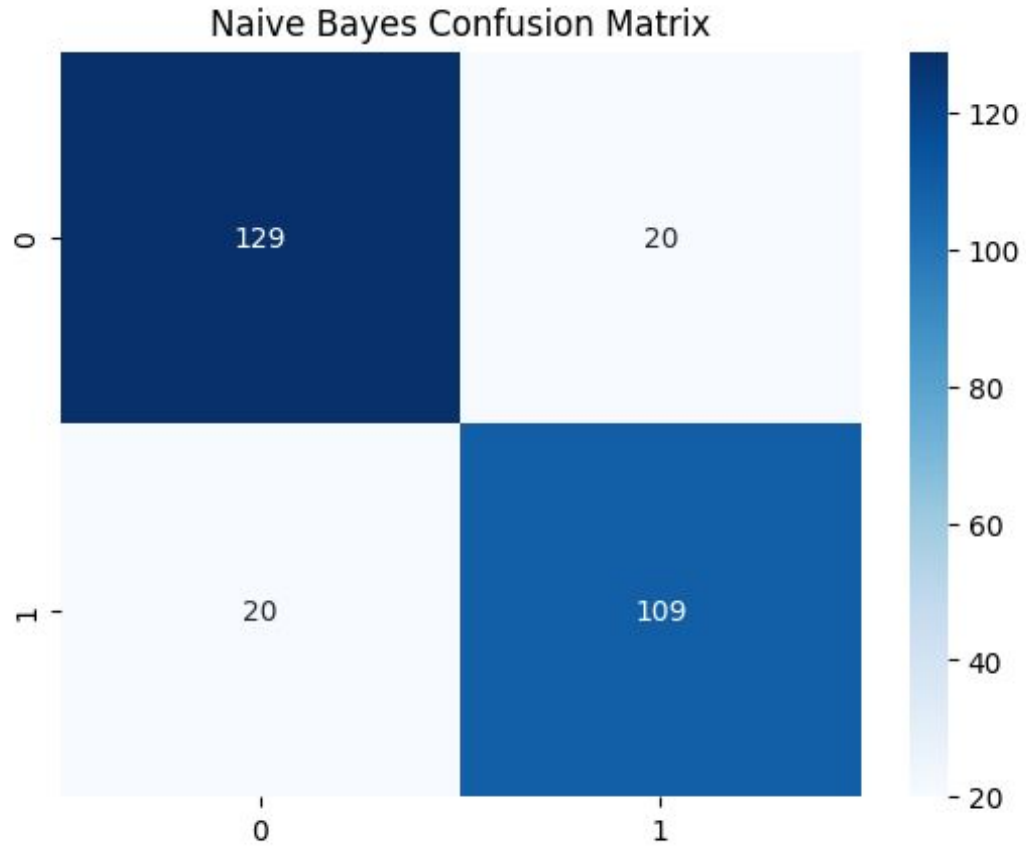
Logistic Regression:
Accuracy: 0.8525179856115108
Precision: 0.84375
Recall: 0.8372093023255814
F1 Score: 0.8404669260700389

BUILDING MODEL



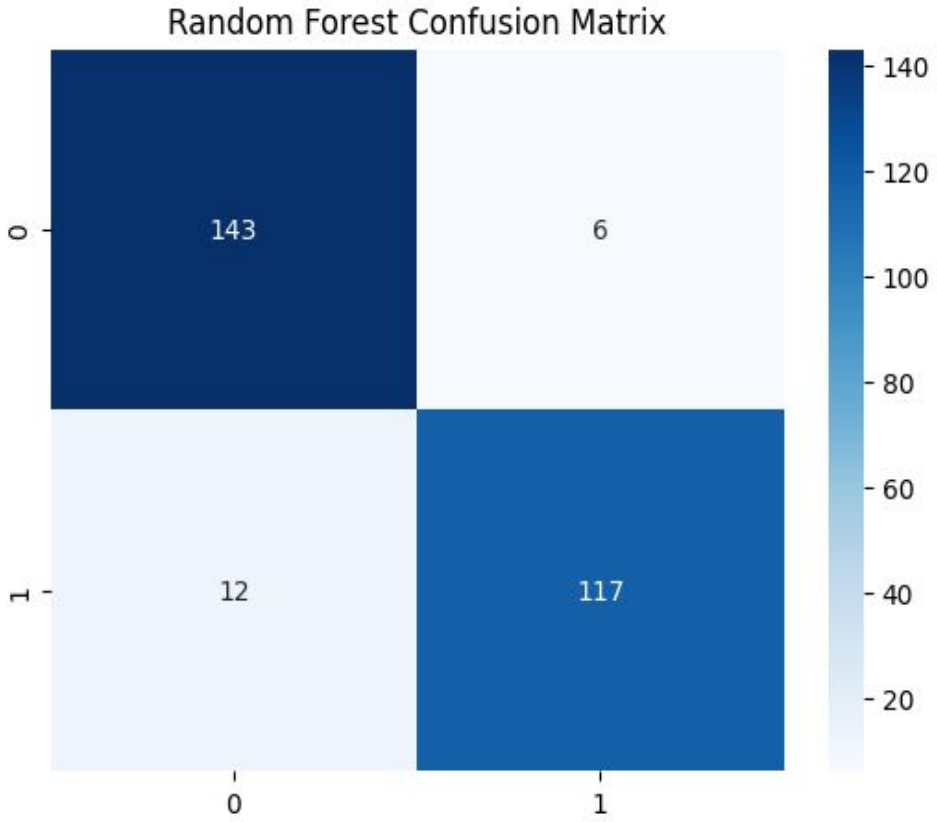

```
Support Vector Machine:
 Accuracy: 0.85251798561151108
 Precision: 0.83333333333333334
 Recall: 0.8527131782945736
 F1 Score: 0.8429118773946360
```

BUILDING MODEL



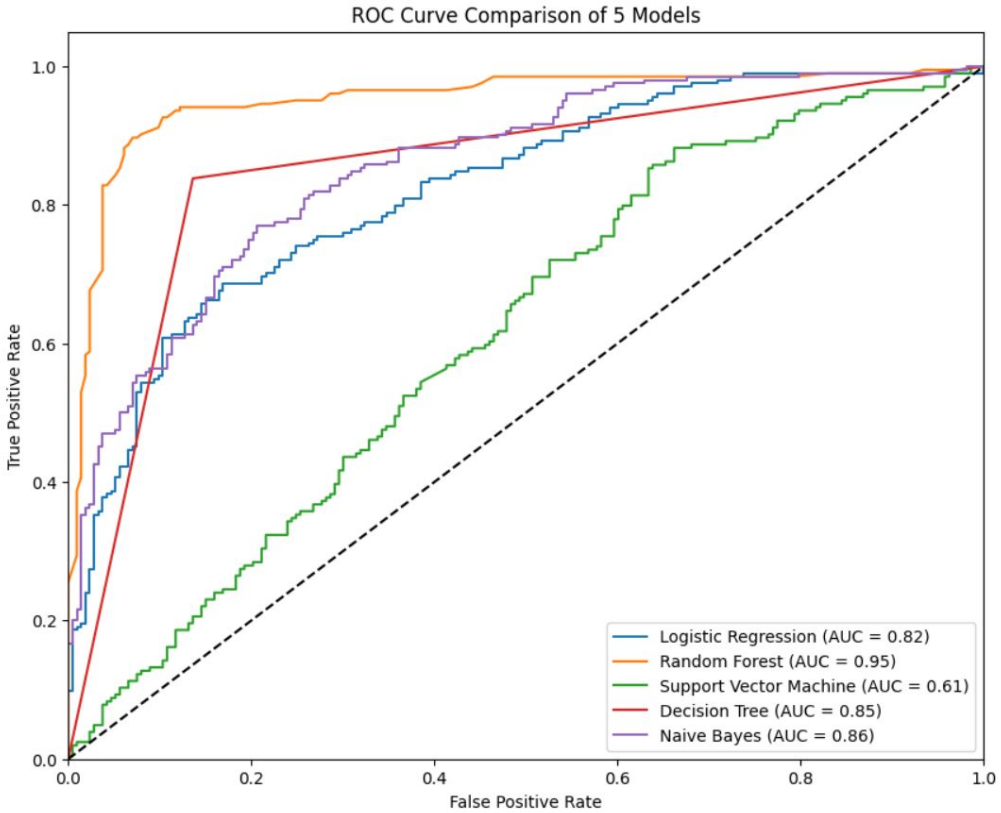

```
Naive Bayes:
 Accuracy: 0.8561151079136691
 Precision: 0.8449612403100775
 Recall: 0.8449612403100775
 F1 Score: 0.8449612403100775
```

BUILDING MODEL




```
Random Forest:
 Accuracy: 0.935251798561151
 Precision: 0.9512195121951219
 Recall: 0.9069767441860465
 F1 Score: 0.9285714285714286
```

BUILDING MODEL



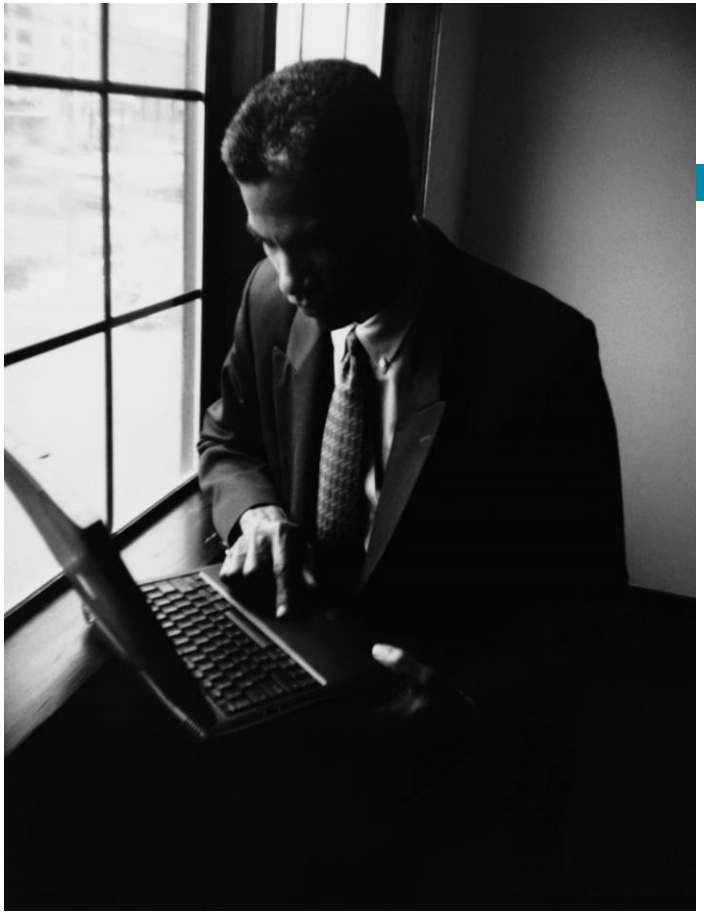
Random Forest is the best performing model with **AUC = 0.95**, close to perfect.

Support Vector Machine is the worst performing model with **AUC = 0.61**, only slightly better than the random models.

The remaining models have average performance.

SUMMARY

- Notable customer groups are **Under 30 and 41-50** years old.
- Favorite products are **Sport and Clothing**. Products with potential for **future development are Beauty**
- **Random Forest** is the best performing model that helps us understand the factors that influence purchases:



CONCLUSION

WEBSITE & STORE ONLINE

Improve website experience to increase customer stay time.

COUPON

Offering discounts positively influences purchase decisions and increases online promotion.

CUSTOMER LOYALTY

Increase customer loyalty program participation for higher purchase likelihood.

PRODUCT

Focus on hot selling product categories that appeal to customers who are interested in them. Especially younger customers.

DATA SOURCE

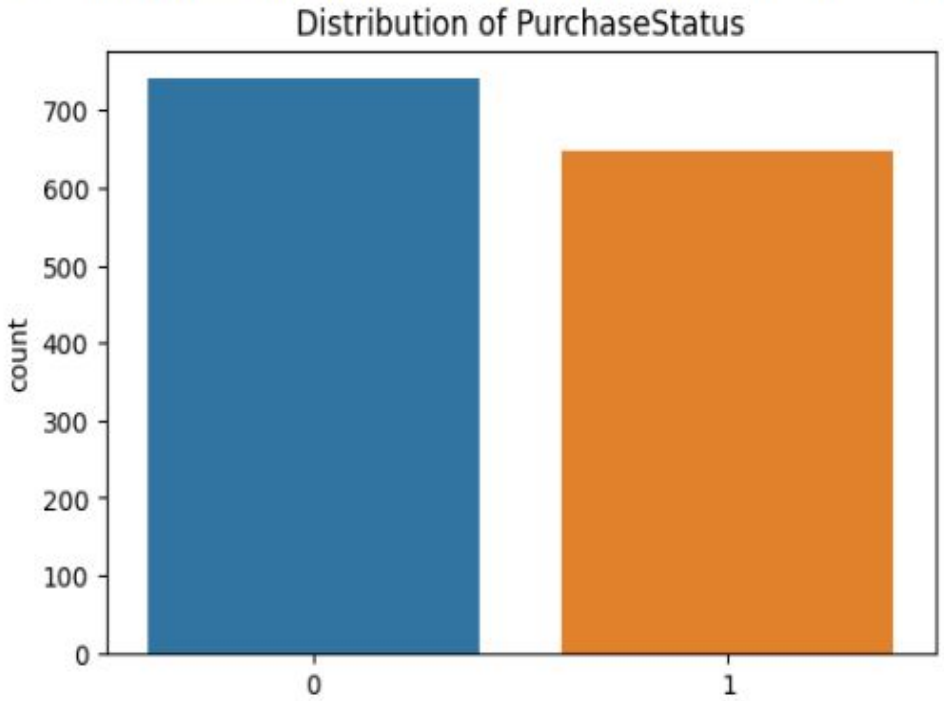
Link Python	Click here: **Link**
Link Source Data	Click here: **Link**
Link SlideShow	Click here: **Link**
Link Source Information	Customer Purchase behavior Modeling & EDA. For more info, click **Here**
Contact me	Gmail to Bao, click **Here**
Linkedin	M-Company. For more info, click **Here**

THANK YOU

Nguyễn Quốc Bảo
(DA & BA Analyst)

Contact: baoquocnguyen1408@gmail.com

EXPLORATORY DATA ANALYSIS



KHẢ NĂNG MUA HÀNG CỦA KHÁCH HÀNG

- **Nhóm 0** *(Không mua hàng)* chiếm ưu thế hơn **Nhóm 1** *(Mua hàng)*, với tỷ lệ xấp xỉ **51.85%** so với **48.15%**.

=> Cho thấy cần tập trung phân tích sâu hơn vào **Nhóm 0** để hiểu rõ nguyên nhân và tìm giải pháp cải thiện.

CONCLUSION

- Dựa trên mô hình hiệu quả nhất, hiểu rõ yếu tố ảnh hưởng mua hàng

  LoyaltyProgram (0.32): Khách hàng tham gia chương trình khách hàng thân thiết có khả năng mua hàng cao hơn.

  DiscountsAvailed (0.3): Việc cung cấp giảm giá có ảnh hưởng tích cực đến quyết định mua hàng.

  TimeSpentOnWebsite (0.28): Khách hàng dành nhiều thời gian trên website có xu hướng mua hàng nhiều hơn.

- Dự đoán hành vi trong tương lai

- **Tập trung vào các yếu tố có tương quan cao**:
  - Xây dựng các chương trình khách hàng thân thiết.
  - Cải thiện trải nghiệm trên website để tăng thời gian khách hàng ở lại.
  - Tối ưu hóa chiến lược giảm giá.
- **Phân khúc khách hàng dựa trên tuổi tác**:
  - Các chiến lược tiếp thị nên ưu tiên nhắm vào nhóm khách hàng trẻ.