Customer Behaviour Analysis

Junha Baek

28th Nov 2024

Table of Contents

Executive Summary

This report is mainly focused on developing personal data analysis skills. Thus, the business backgrounds and conditions are assumed.

The Customer Behaviour Analysis project is to identify the customer groups who are likely to purchase services or products of an organisation. The analysis was conducted using the *Customer Purchases Behaviour Dataset* (Goyal, 2021). The income of the customers significantly affects their purchasing amount.

Introduction

Understanding customer purchasing behavior is important for any business seeking to improve product offerings, boost customer satisfaction, and increase revenue. In this project, I aim to predict customers who are most likely to subscribe to new products based on their past purchasing history and related information. This insight will be used to optimize marketing efforts and target customers more effectively, thereby maximising the return on investment.

The analysis is based on the dataset which provides customers' demographic data such as income, location, age, and gender, and purchase history like product categories and purchase amount.

The analysis will involve multiple stages, including data cleaning, exploratory data analysis (EDA), model building, and evaluation. The findings from this analysis will be used to derive actionable insights and recommendations, which can ultimately assist in developing data-driven marketing strategies and improving customer engagement.

Data Summary

The dataset was obtained from the Kaggle repository named 'Customer Purchases Behaviour Dataset'. It is generated by generative AI to mimic real-world commercial data.

The dataset is saved as a CSV file. This data provides customers with detailed information which is useful for analysing customer purchasing patterns and predicting customer behaviour to focus on specific targets to maximise revenue.

The dataset contains 10,000 records and 12 attributes.

- id: A unique identifier for each customer.
- age: Age of the customer.
- gender: Gender of the customer (0 for Male, 1 for Female).
- income: Annual income of the customer.
- education: Education level of the customer.
- region: Region where the customer resides.
- loyalty_status: Loyalty status of the customer.
- purchase_frequency: Frequency of purchases made by the customer.
- purchase_amount: Amount spent by the customer in each purchase.
- product_category: Category of the purchased product.
- promotion_usage: Indicates whether the customer used promotional offers (0 for No, 1 for Yes).
- satisfaction_score: Satisfaction score of the customer.

As it is an AI generated dataset, missing values are not detected. Likewise, outliers are little but the outliers in `purchase_amount` are replaced by maximum value as the data is concentrated on lower side.

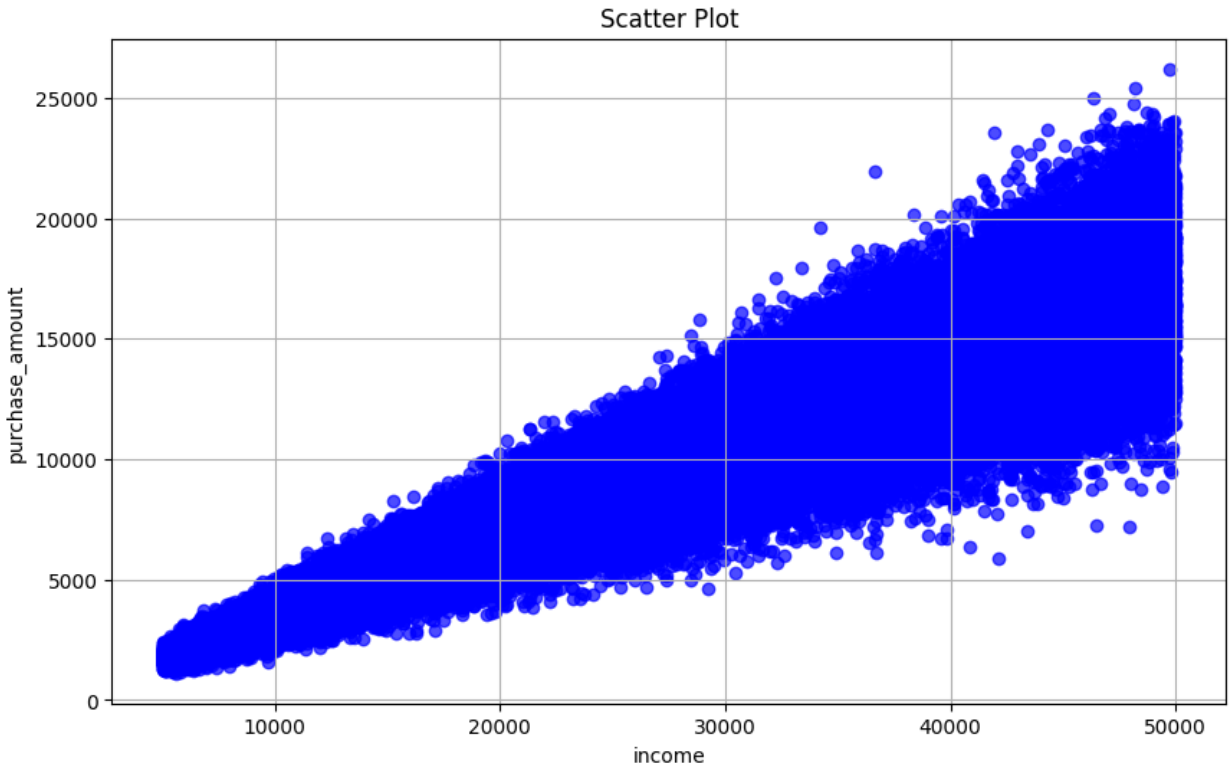
The dataset does not tend to show trends or relationships among features to identify any features that affect purchase behaviour, most values are evenly distributed.

Exploratory Data Analysis

A comprehensive Exploratory Data Analysis (EDA) was conducted to better understand the structure, quality, and relationships present in the dataset. The primary objective of EDA was to gain insights into customer behavior and identify potential key factors influencing purchase amount and satisfaction.

For a detailed exploration of the data, please refer to the **EDA notebook** provided as part of this project.

The customer's income level affects the amount of purchase.



However, there was no other significant factor influencing the income or purchase amount.

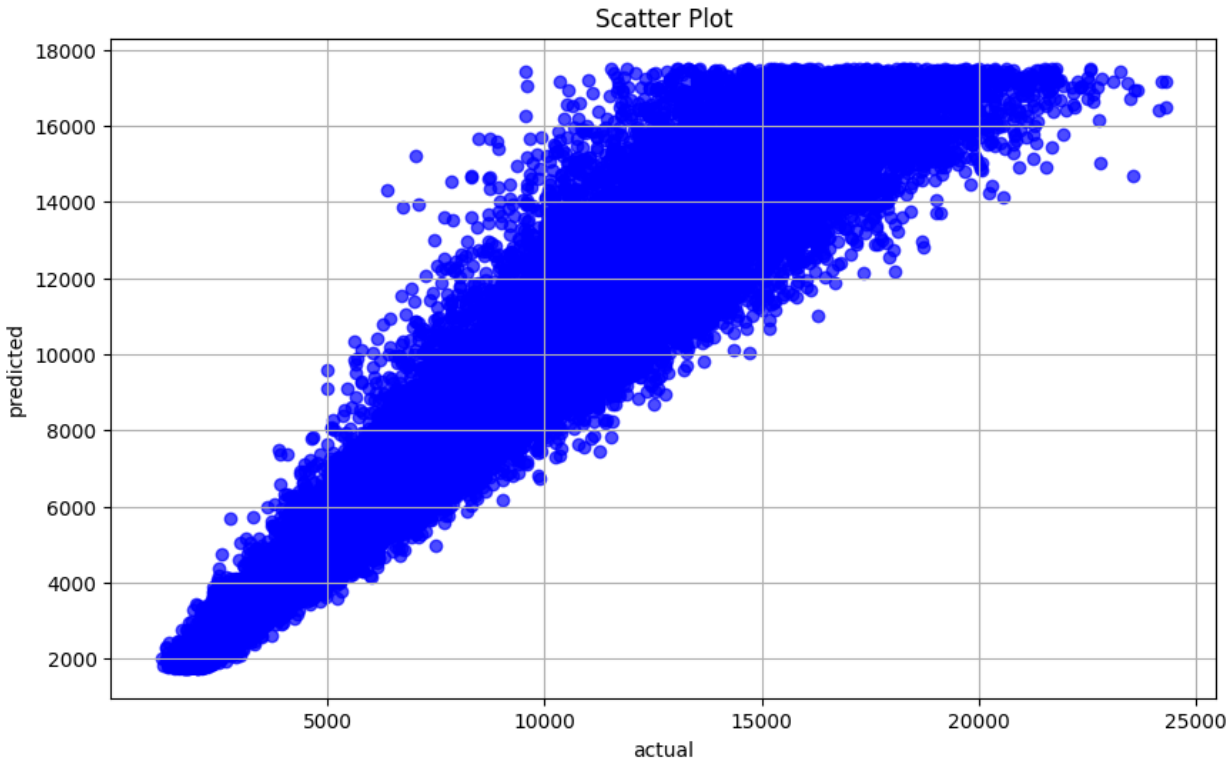
Model Development

The prediction model is developed to anticipate the purchasing amount of each customer based on their income level. 4 different types of models are tested and evaluated based on Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2. MSE emphasises larger errors, useful when it is required to heavily penalise incorrect predictions. MAE takes the absolute value of the differences between predicted and actual values; useful it is desired to treat all errors equally. R^2 measures the overall fit of the model, useful for comparing models and understanding explained variance. The models tested are linear regression, polynomial regression, decision-tree regression, and random forest regressor.

Income is chosen to predict purchase amount as purchase amount and income showed significant correlation, but others did not. The outliers of purchase amount have been replaced by maximum value of the purchase amount. Encoding is skipped as the categorical variables are not selected. The data is split into 80% of training and 20% of testing. The linear regression model is selected as baseline to compare the performances of more complex models.

Models	MSE	MAE	R^2
Linear Regression	2337297.247456583	1104.7002729827261	0.898043074037824
Polynomial Regression	2337272.595640726	1104.7043834923188	0.8980441493924316
Decision-Tree	2361897.4204908926	1114.160537113369	0.896969972179066
Random Forest	3237782.4104110613	1295.7734355432033	0.8587623624427136

Overall performance is the best with the polynomial regression model. Below is the scatter plot comparing predicted value to actual value by using the polynomial regression model.



Scikit-Learn module is used to develop the models.

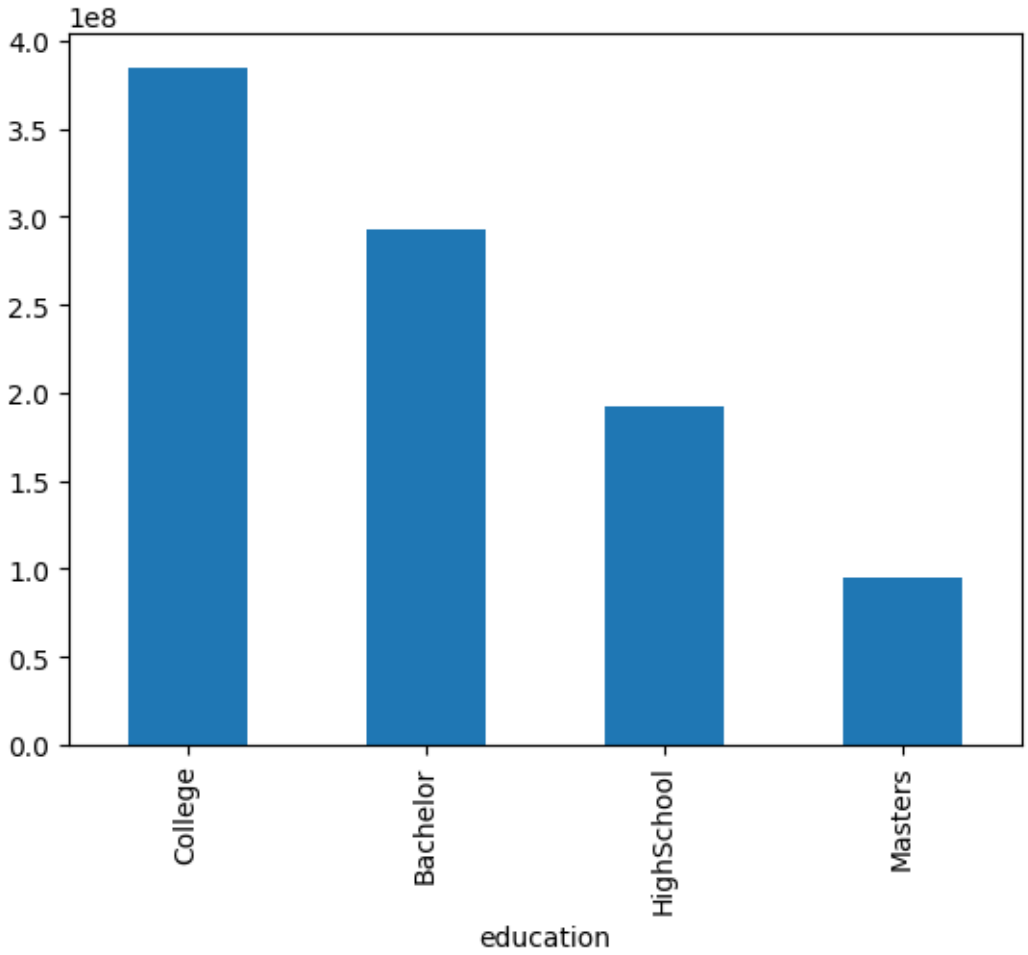
Detailed information can be found on **model_experiments.ipynb** which is attached along the project.

Insights Analysis

Customer Segmentation

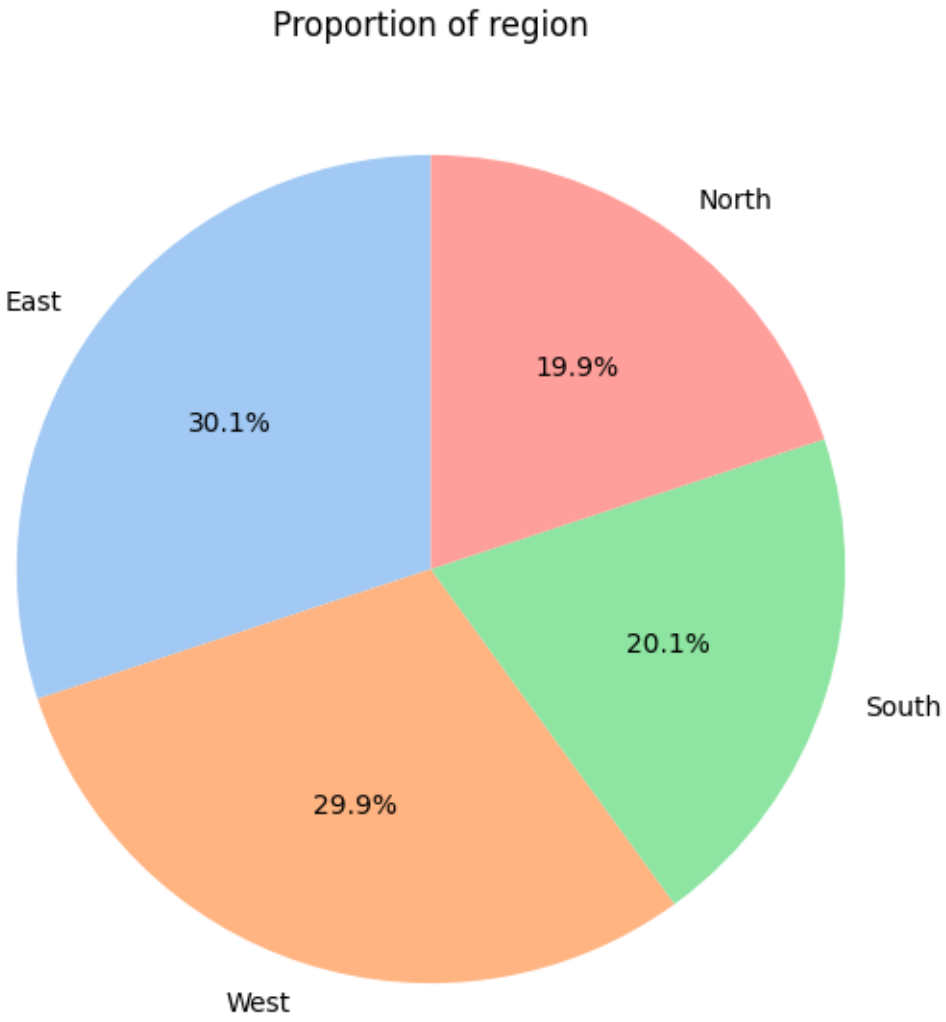
While purchase amount for every group is evenly distributed, there are specific customer groups that purchase the product more than other groups. This could be simply because of the population distribution in terms of macro perspective. Even distribution also means that a bigger customer group spends more money on purchasing. Thus, it is important to know which groups are the main customer groups.

1. Education level



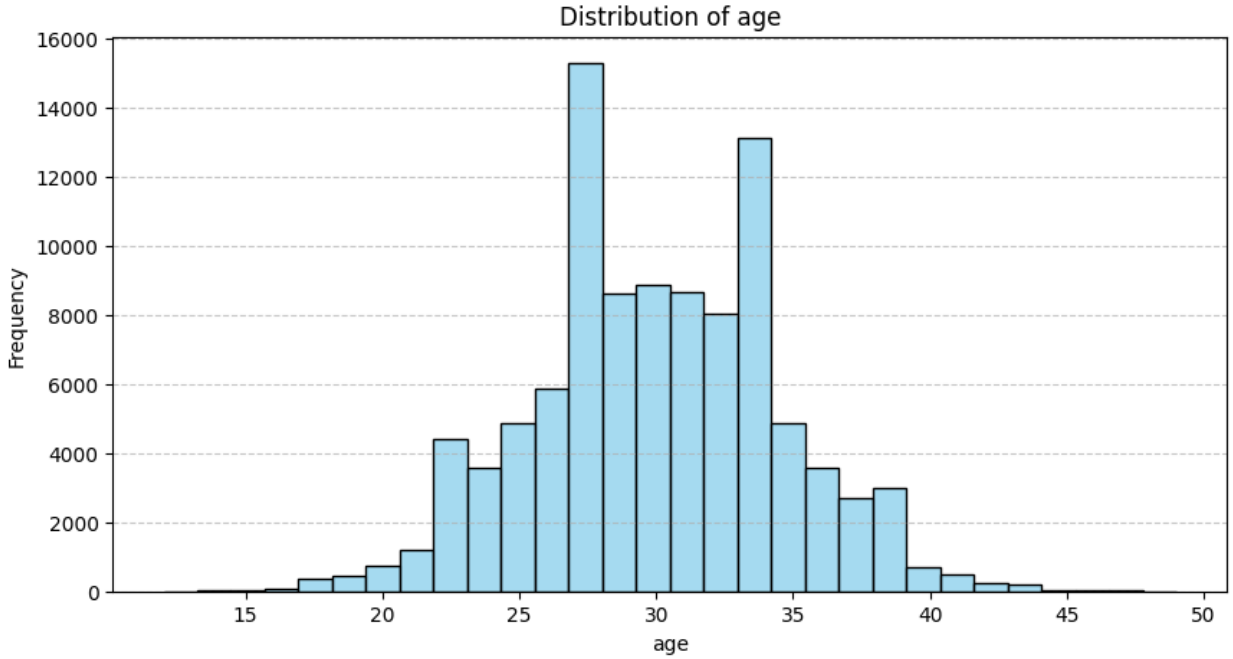
The biggest education group of customers is College graduates. It is important to identify common interests of college graduates' group to maximise the revenue.

2. Region



The main customer groups are from the East and West regions. This could be because the location of the store is focused on the east and west regions. In this case I will assume that it is not only because of the location. If the reason is not about location, the organisation can improve marketing for the North and South regions to attract more customers from there.

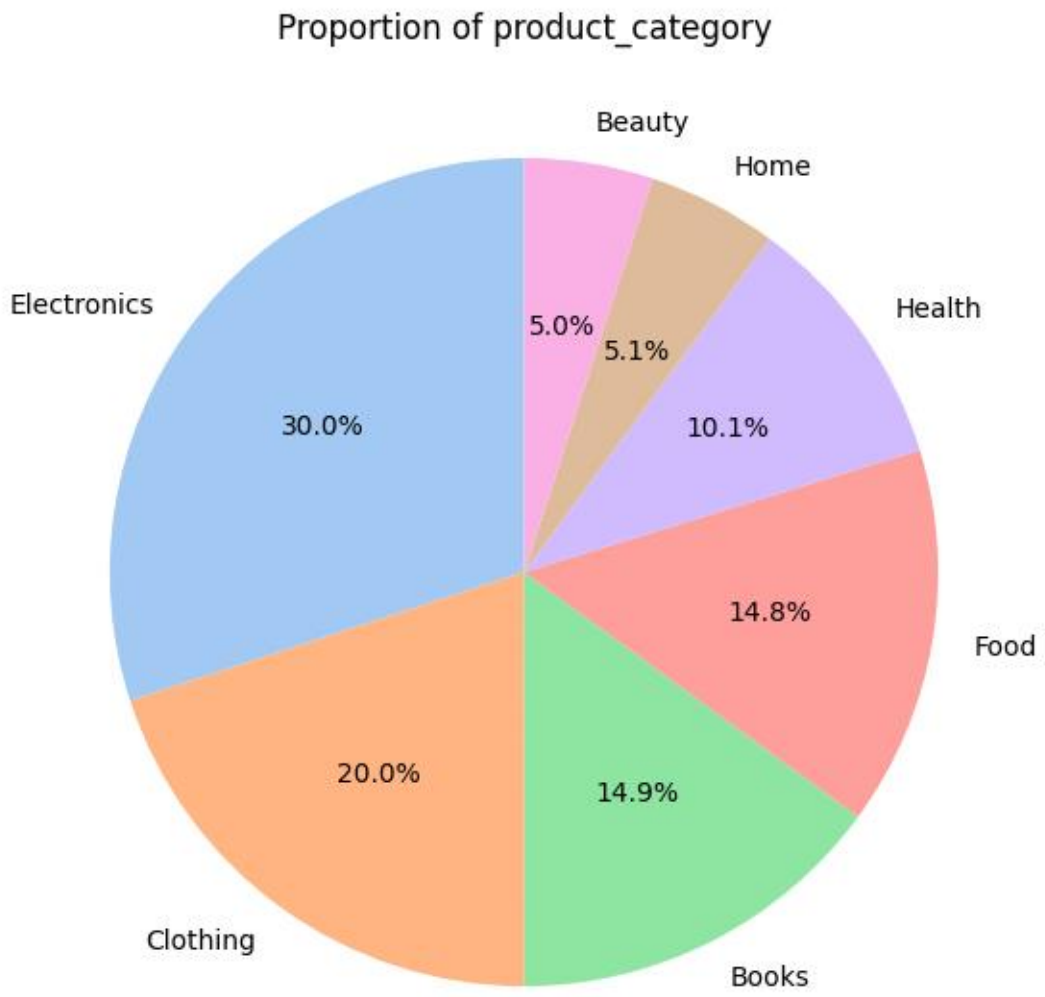
3. Age



The customers are mainly in their mid 20s to mid 30s. Specifically 27 to 28 and 33 to 34 are the biggest age groups of the customers. It could be beneficial to analyse their interests to increase revenue or focus on marketing to other age groups to attract other age groups.

Product Analysis

The most purchased product category is electronics and is followed by clothing. These two categories dominate the sales by 50 per cent of total sales.

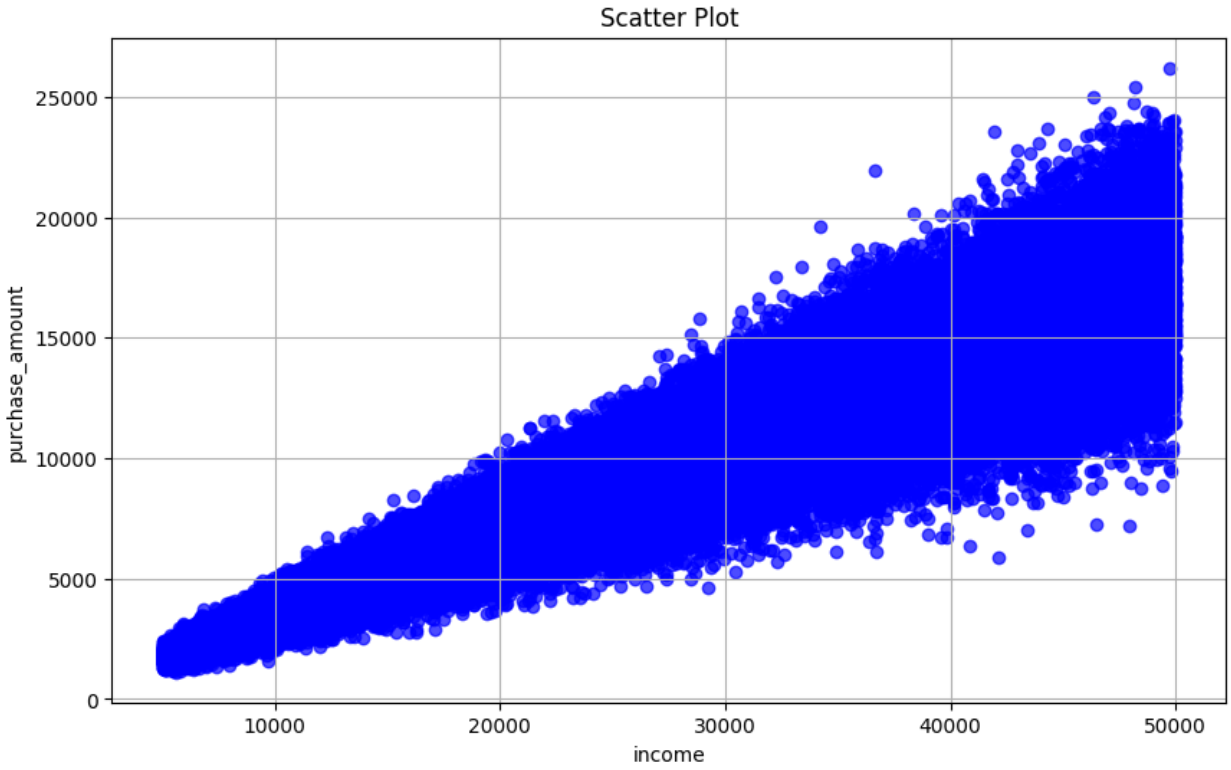


This shows it will be beneficial to focus on expanding the business toward electronics and clothing. While books and food take big portions by approximately 30 per cent of the sales, beauty and home products are weakness of the business.

This analysis could be collaborated with other analysis, however, there is no correlation between product categories and customer groups.

Purchase Behaviour Analysis

There is a significant correlation between income and purchase amount. This shows a positive linear relation which means higher income customers may purchase more than customers who have lower income.



This analysis is validated by p-value test to ensure whether this is statistically meaningful.

The outcome is used to develop the polynomial regression model to predict customer purchasing amount. Thus, the business can determine the most efficient group to maximise its revenue by investing in marketing or products.

Unfortunately, there is no other meaningful correlation between income and other features of the current dataset.

Limitation

As the dataset is AI generated and evenly distributed, the data does not show many meaningful insights, which can be used for real-world circumstances and interesting relationships between the features which would be possibly related in actual data.

Conclusion

This Customer Behaviour Analysis project identified key factors influencing purchasing behaviour to improve marketing strategies. Using the 'Customer Purchases Behaviour Dataset,' I analysed data to predict customers' likelihood of purchasing products.

Income level was found to be the strongest predictor of purchase amount, and polynomial regression was the most effective model. Customer segmentation by education, region, and age provided additional insights into targeting opportunities.

While other features showed limited correlation, the findings provide a foundation for targeted marketing, particularly in electronics and clothing. Despite the limitations of the synthetic dataset, this analysis helped me improve my data analysis skills. Future work should use real-world data for more actionable insights.

Reference

Goyal, S. (2021). *Customer Purchases Behaviour Dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/sanyamgoyal401/customer-purchases-behaviour-dataset