Table of Contents

1. Abstract

The increasing complexity of consumer behavior in e-commerce necessitates data-driven strategies for customer segmentation and sales forecasting. This study analyzes a dataset comprising 2000 customer transactions from an online retail store to identify purchasing patterns and optimize marketing efforts. Through demographic, geographic, transactional, and post-purchase analyses, we segment customers into distinct groups using machine learning techniques such as clustering and RFM scoring. Additionally, predictive models, including logistic regression and XGBoost, are utilized to estimate purchase probabilities and forecast future sales. Our findings provide actionable insights for marketing and sales teams to enhance customer engagement, improve retention strategies, and optimize inventory management. Ethical considerations regarding data privacy and fairness are also discussed to ensure responsible data usage.

2. Introduction

The rapid expansion of e-commerce has led to an overwhelming volume of customer data, offering businesses valuable insights into consumer behavior. However, effectively analyzing and interpreting this data remains as a challenge. Customer segmentation plays a crucial role in enabling businesses to personalize marketing strategies, optimize pricing models, and predict future sales trends.

The results of this study are intended to aid e-commerce businesses in refining their marketing strategies, improving conversion rates, and optimizing resource allocation. Furthermore, ethical considerations such as data privacy are addressed to ensure compliance with global data protection regulations. By leveraging data science techniques, this research provides a strategic framework for enhancing customer engagement and maximizing business growth.

3. General Overview

3.1 Objective of the Study

By this report, we try to provide a thorough analysis of a database, containing transactions of a commercial e-shop. We intend to identify customers' behaviors, analyze variables given by a list of transactions and find out if customer segmentation is possible and provide useful information for the company. Moreover, we will suggest marketing strategies leveraging the outcomes of dataset analysis and we will proceed to sales forecasting for 2024.

Outcomes of this analysis will be presented to Marketing and Sales Departments as the main stakeholders of this project, but it could also provide useful information for the Product Development and Merchandizing, Finance and Legal, Customer Support, Supply Chain and Warehouse and IT departments.

This analysis extends to 5 main dimensions that will help us extract useful information for the dataset:

a. Demographic analysis
b. Geographical analysis
c. Product-related analysis
d. Transactional analysis
e. Post-purchase analysis

The methodologies we are going to use to analyze dataset is descriptive and predictive. We firstly focus on understanding better the data and try to extract insights over data anomalies or correlations among values, analyze various events and behaviors and on the final stage, this analysis,will include Reference-Frequency-Monetary Value (RFM), clustering techniques, and machine learning models for purchase probability prediction and revenue forecasting, based on the dataset provided.

3.2 Dataset Overview

Dataset comprises a wide range of variables related to two thousand (2000) customer transactions of an e-commerce website, collected daily after the “Add to cart” step of the customer journey, for the period of 2019 until 2023. Dataset includes a unique identifier to ensure data accuracy and integrity and facilitate the analytical process and contains the following variables:

- SessionStart (Date)
- CustomerID (Integer / Unique Identifier)
- FullName (String)
- Gender (String)
- Age (Numerical)
- CreditScore (Numerical
- MonthlyIncome (Numerical)
- Country (Categorical)
- State (Categorical)
- City (Categorical)
- Category (Categorical)
- Product (String)
- Cost (Numerical)
- Price (Numerical)
- Quantity (Numerical)
- CampaignSchema (Categorical)
- CartAdditionTime (Date)
- OrderConfirmation (Boolean)

- OrderConfirmationTime (Date)
- PaymentMethod (Categorical)
- SessionEnd (Date)
- OrderReturn (Boolean)
- ReturnReason (Categorical)

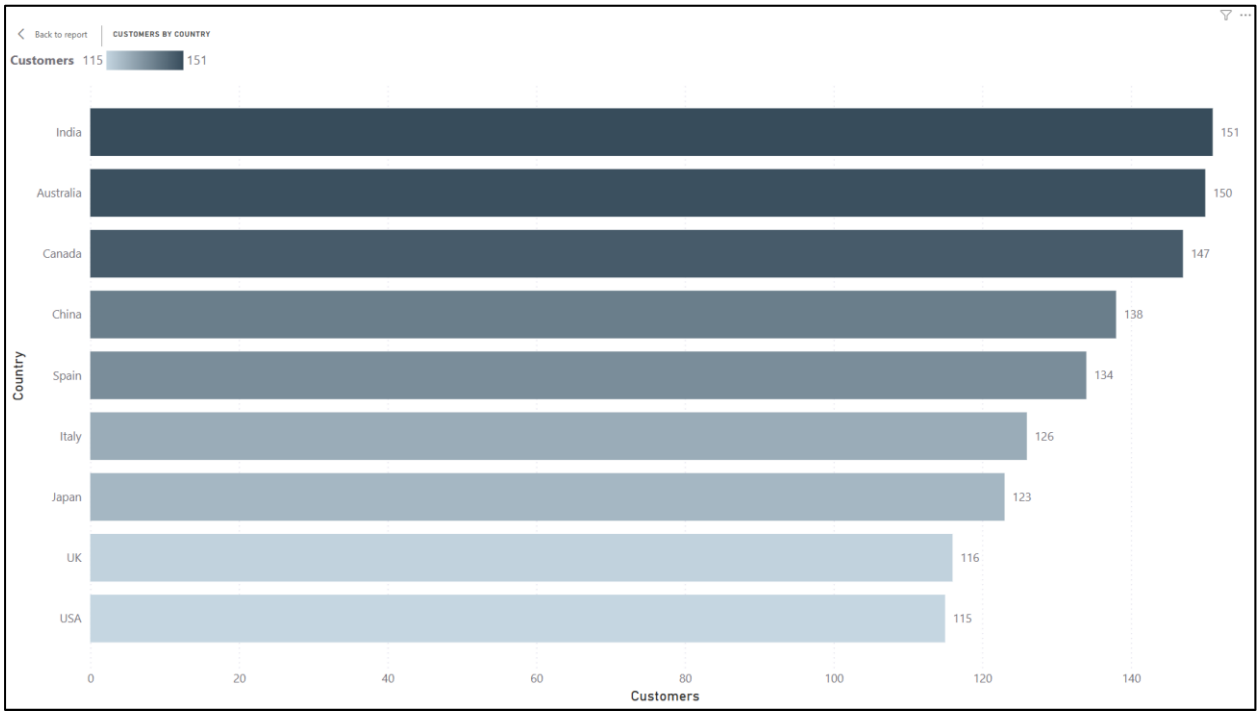
3.3 Ethical Considerations

It is important to ensure that this analysis abides by any legal framework concerning privacy and data security, and ethical issues are embedded in every phase of execution. Bellow the perspectives of ethical considerations:

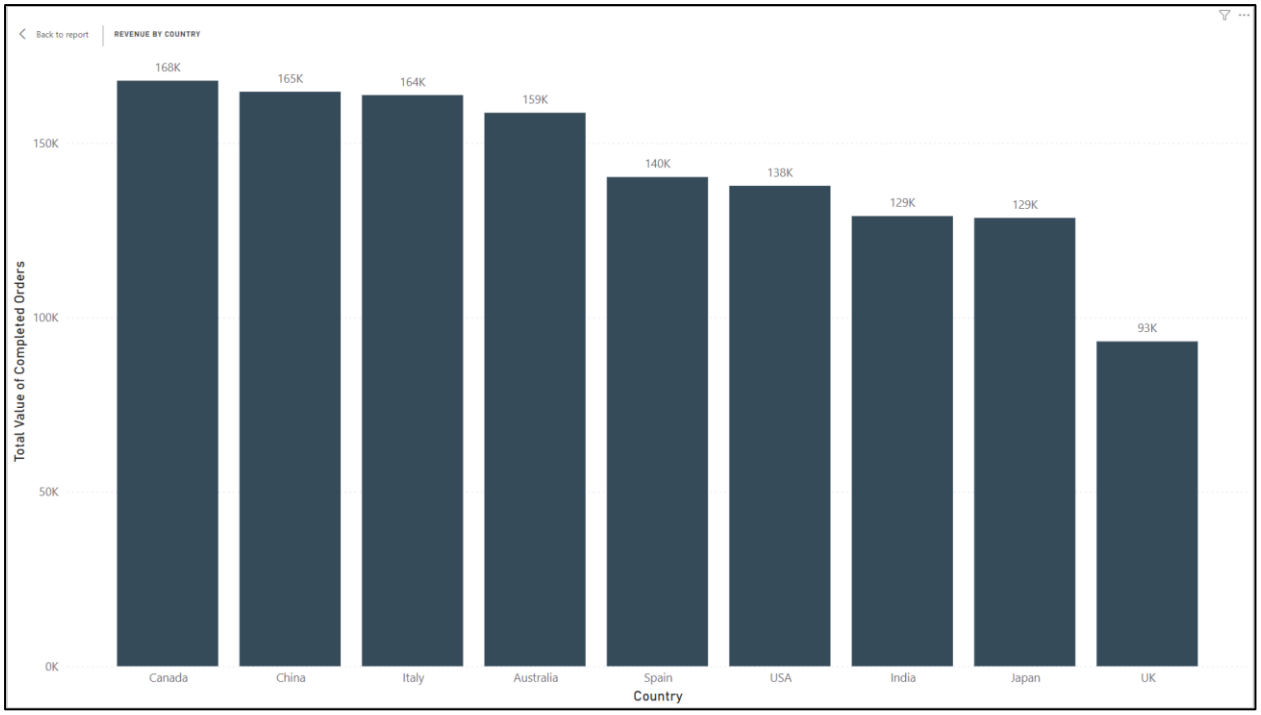
- Transparency: A clear statement must be made to individuals whose data is processed. Customers must be aware that their data is collected strictly for internal purposes, how data is processed, and we need to ensure that customers are given the option not to provide their personal data.
- Privacy, confidentiality and security: Dataset is processed exclusively by authorized personnel, abiding by all data security, encryption and storage legislation. The outcomes of analysis should be anonymized and delivered only to directly interested or affected departments of the company. For example, EU legislation for General Data Protection Regulation (GDPR) defines companies' obligations regarding personal data collection and manipulation, with detailed policies and harsh penalties.
- Accountability: Proper documentation of each step towards final exports and outcomes shall be stored for the reviewing and re-evaluating process by automated and human means. The department that is undergoing this analysis is accountable and responsible for every step of data management, distributing responsibilities, declaring limitations and establishing policies and procedures to tackle potential future issues.
- Discrimination, equity and fairness: Dataset is treated in a non-discriminatory manner, and analysis is conducted without being affected by political, ethnic and religious aspects.
- Social license: We need to consider that customers need to have the choice of opting out of their personal data collection and management and we need to state that outcomes of analysis will be used explicitly for entrepreneurial purposes only. In any opposite case, a public outlash could be triggered, affecting negatively the brand image and customer loyalty.

4. Geographic Analysis

4.1 Distribution of Customers by Country & Regional Sales Performance



Distribution of Customers by Country



Value of Completed Orders by Country

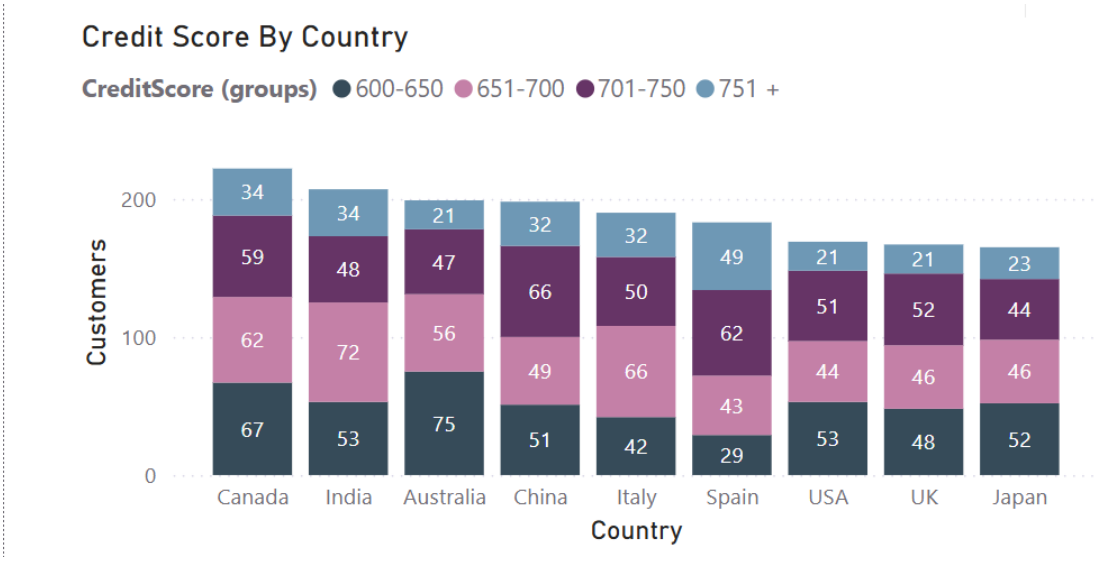
Strategic Importance:

- There are countries such as Canada and Australia that present high volume of customers and high revenues. Also, some markets present low volume of customers generating lower revenues, such as UK and Japan.
- Markets that present a high volume of customers, providing lower revenue for the company, are a concern. For example, India has a top 151 distinct customers that bring only 129K.

Market Segmentation Plan:

- Balanced markets: Markets presenting a logical analogy between customers reached and revenues are considered balanced. The company should emphasize retaining these customers, understand that is the position it obtains throughout these markets and examine if there are market opportunities for setting entrance barriers.
- Imbalanced Markets: Reasons that can cause this imbalance may differ. The company shall examine if the pricing policy affects negatively customer conversions and supply chain re-evaluation could result to lower retail prices, external factors from PESTLE analysis could also indicate points of concern

4.2 Distribution of customers' credit score groups by country



Customers' Credit score groups by Country

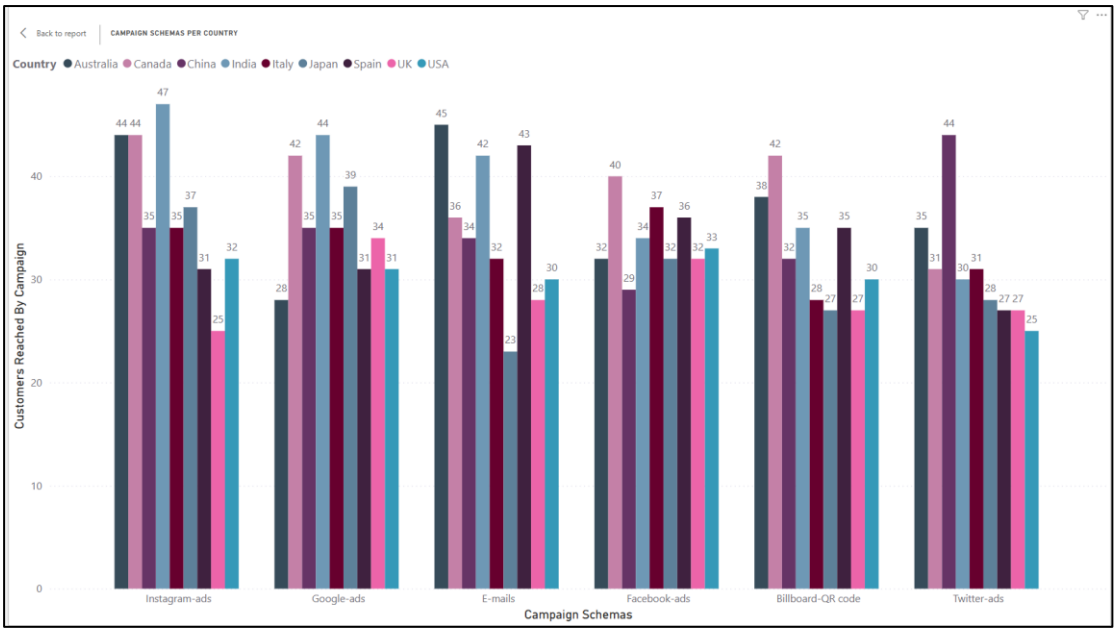
Strategic Importance:

- Credit Score groups among different regions are distributed in a similar way, with low, middle and high credit score customers being the majority and top-tier customers presenting lower volumes. The company can examine what the performance of each group is, regarding sales, product preferences etc.
- Customers from Europe present lower volume of low-tier customers
- Customers from Canada, Australia, USA and Japan present big volume of low credit score tier.

Customer Segmentation - Targeting Plan:

- Markets with dominant group: low credit score: Specific targeting on this group can be done to move customers from low to medium rank. Marketing campaigns, specific time-limited offers and coupons can boost sales and increase conversions.
- Move customers to top-tier: To convert customers to top-tier ranking, exclusive perks and offers must be advertised

4.3 Campaign schemas by country



Distribution of Campaign Schemas by Country

Strategic Importance:

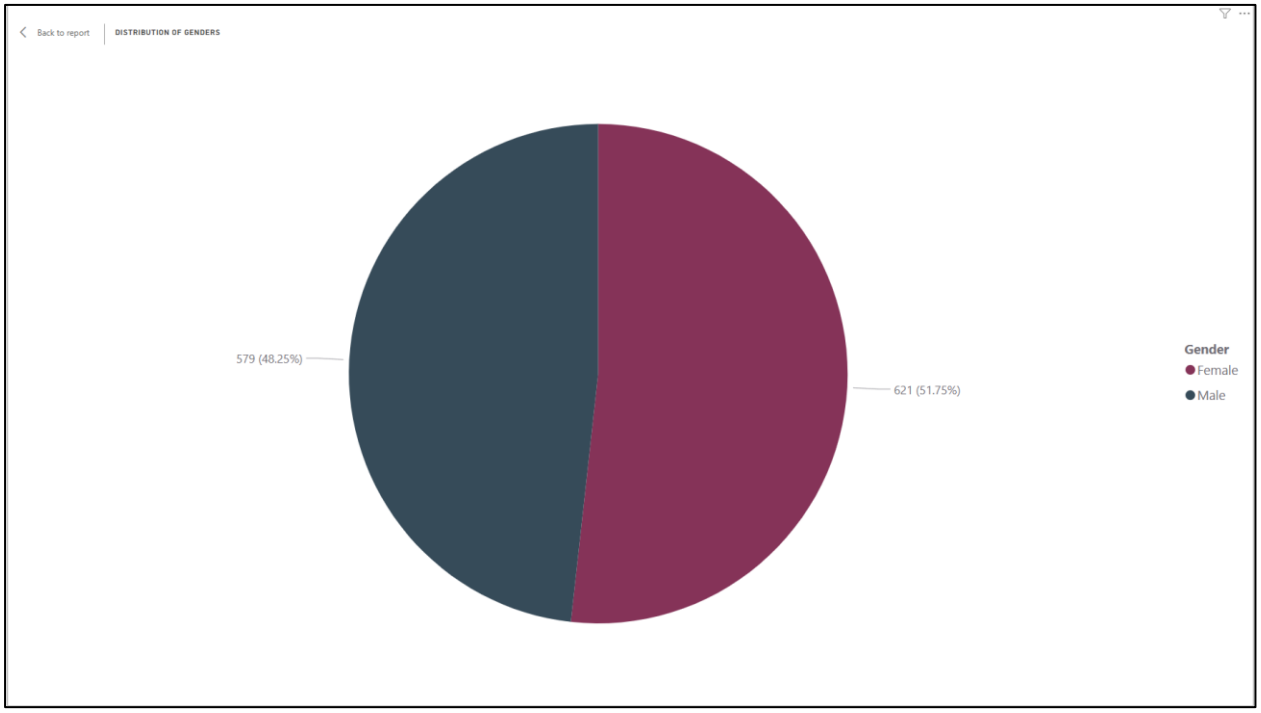
- Campaigns KPIs: Tracking of campaigns' performance per country can guide the company for more efficient marketing and advertising budget allocation and Return On Investment (ROI). By monitoring metrics such as Click-Through-Rate (CTR), Return on ad spend (ROAS) and Engagement Rate, we can understand top-performing and underperforming campaigns as well as clearly define the marketing funnel and the touchpoints of customer journey.
- Personalization: Customer experience can be optimized by personalization implementations to increase customer engagement and establish a closer relationship with the brand, thus providing increased revenue and brand reputation.

Consumer Behavior Patterns:

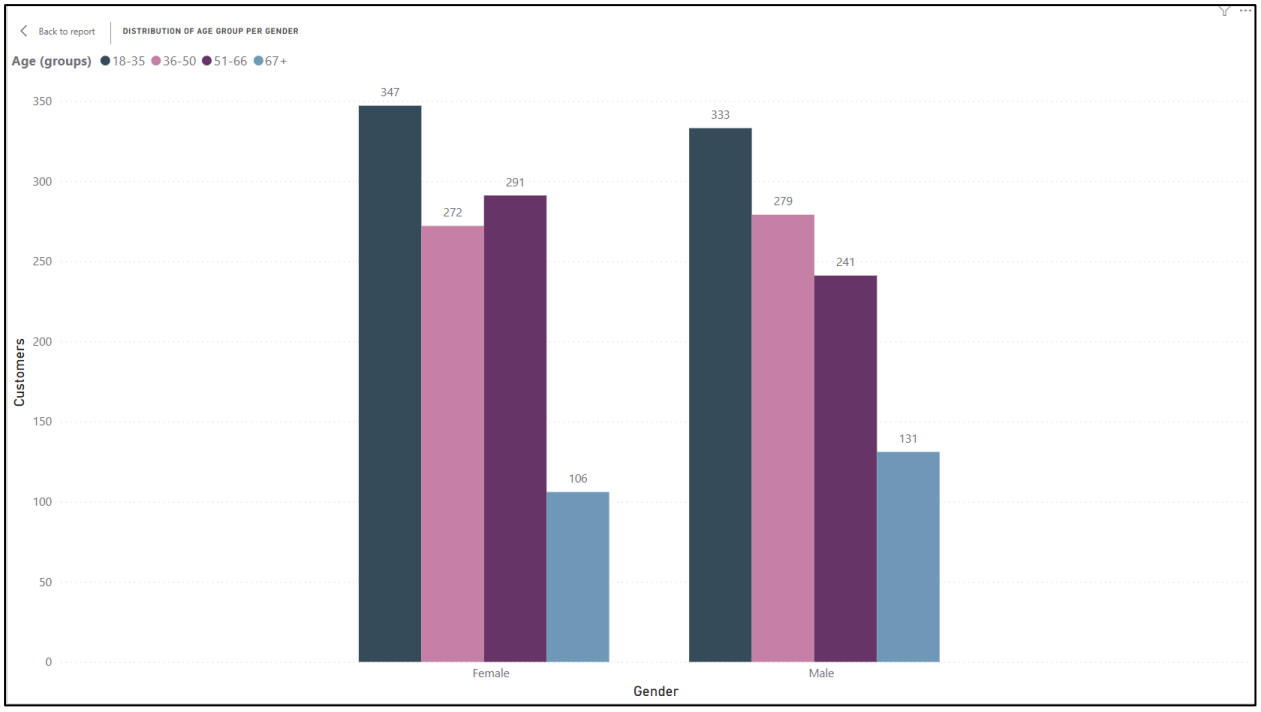
- Billboad and QR advertising seems to fall back compared to digital channels, so company should invest in optimizing digital channels even more.
- Email Marketing: Customers seem to react well to email marketing strategy and proceed to the website to trigger engaged sessions.
- Meta / Google Network cross platform Marketing: Leverage platforms to communicate personalized messages on multiple channels, in order to increase touchpoints with the customers.

5. Demographic Analysis

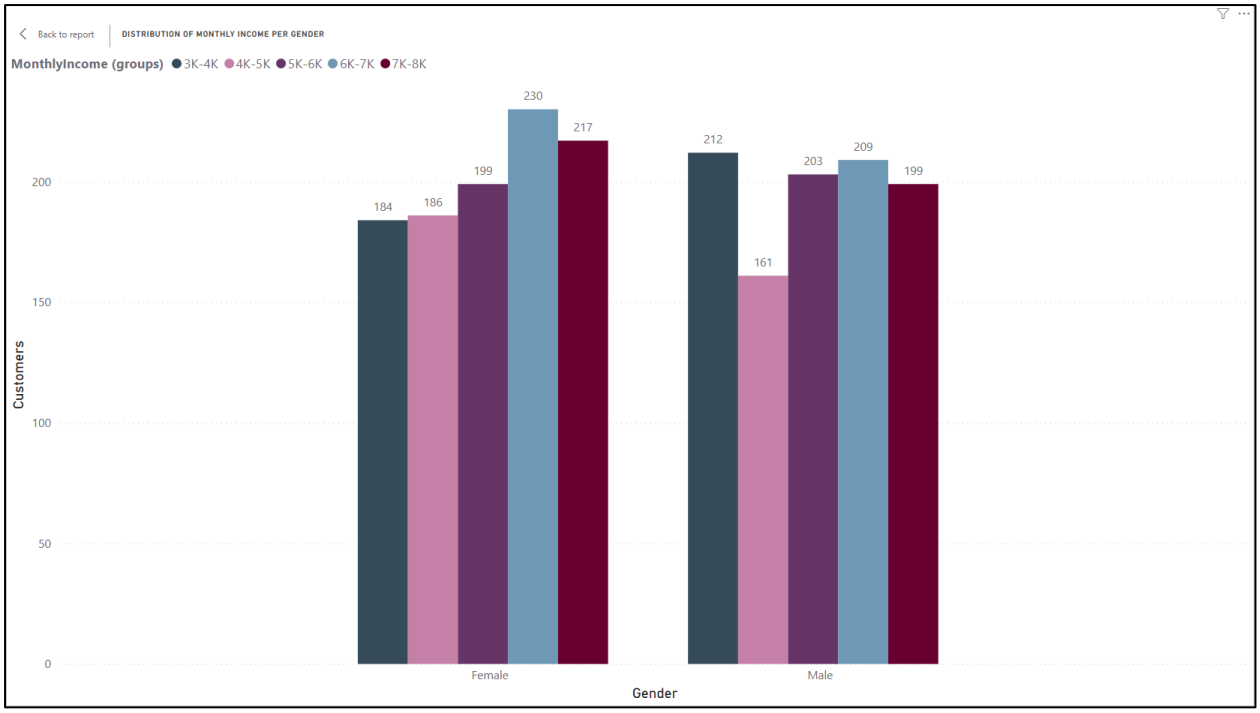
5.1 Gender distribution & Gender distribution by Age Group / Monthly Income



Gender Distribution on Database



Gender Distribution by Age Group



Gender Distribution by Monthly Income

Strategic Importance

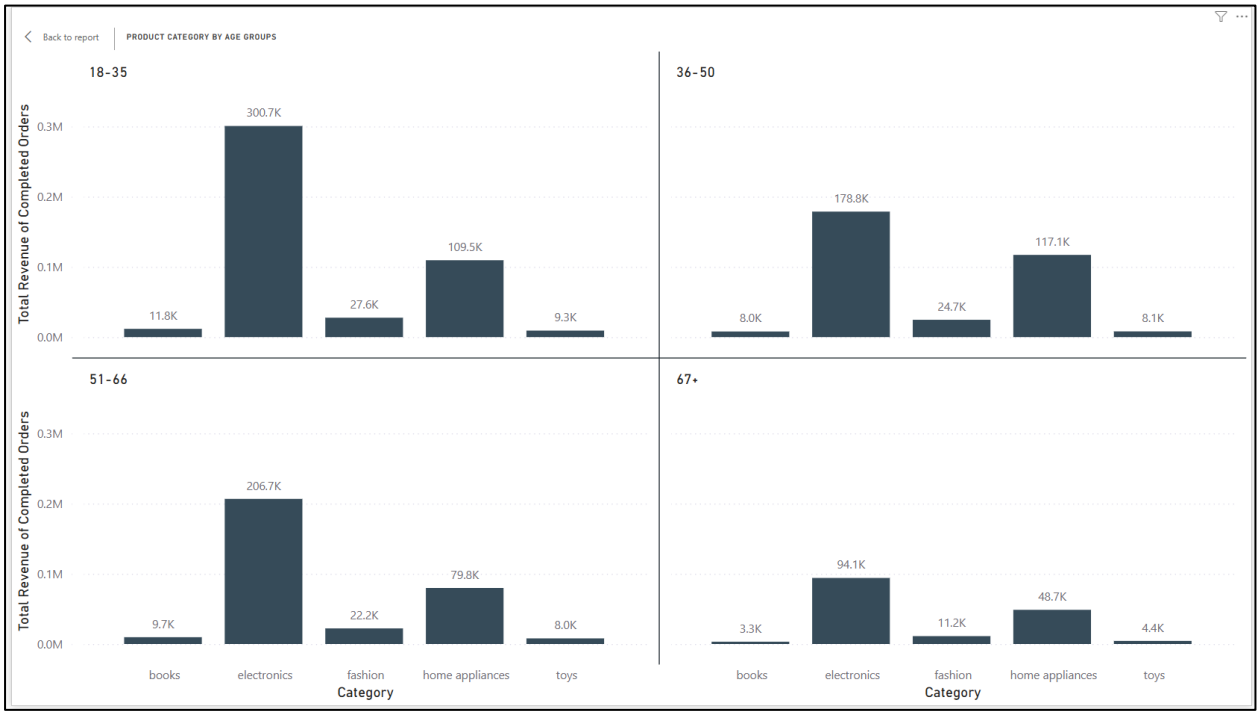
- **Branding Approach:** Based on gender distribution, the company can alter or adapt brand's look and feel and branding strategic approach, tailor marketing campaigns and achieve business growth by targeted paid campaigns. Analysis based on gender distribution can expand on product preferences, marketing channels conversion efficiency and customer journey touch and pain points identification per gender.
- **Personalization:** User experience can be optimized by personalized product suggestions based on gender and can be enhanced by promotional ads, email and messaging marketing custom-made for each user or cluster.
- **Product assortment:** Product portfolio needs to adapt according to age groups and incomes of targeted customers.
- **Pricing policy:** Product pricing and placement on the markets depend strongly on targeted customers' income and age group, so the company needs to examine core groups and plan its strategy accordingly.

Business Strategic Plan:

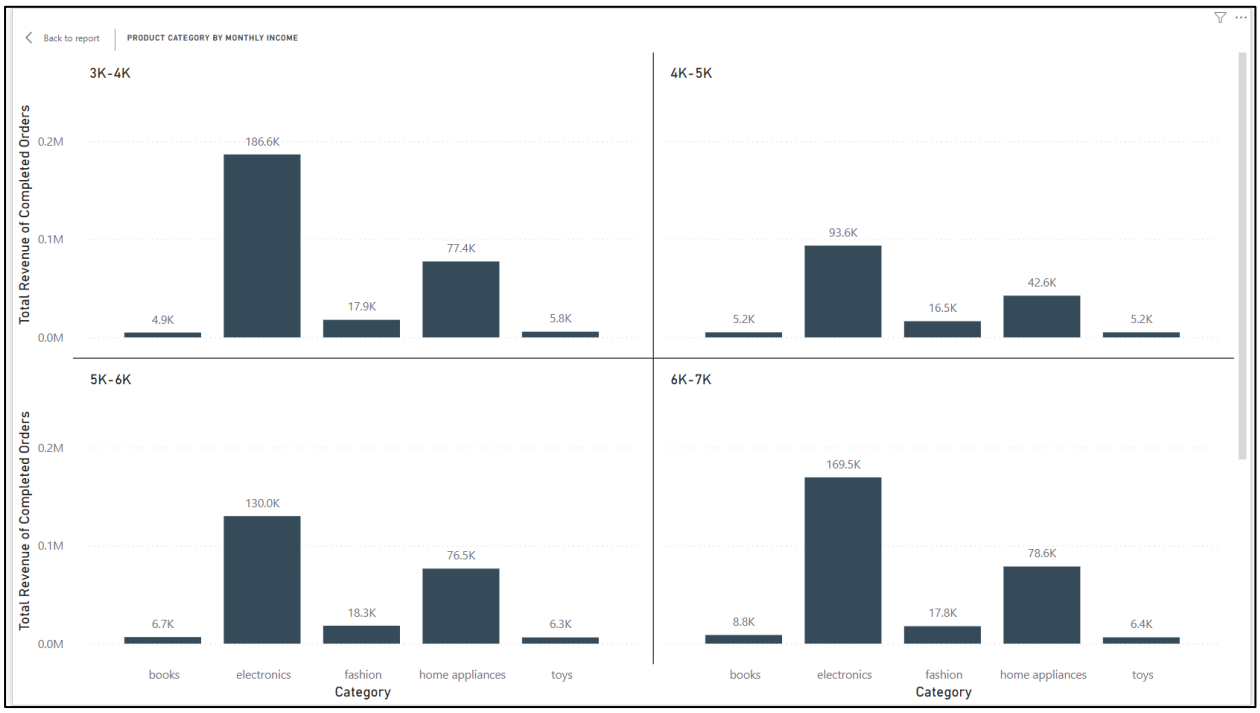
- Female customers are slightly more that male, this means that the branding and the branding, the tone of word and the look and feel of the e-shop must be neutral.
- Customers aged 18-66 are the majority on the database records, so it will be a true challenge to create a marketing mix that communicates efficiently with this age range, a multi-dimensional approach will be definitely needed.

- Slight emphasis can be given on product categories preferred by females with high income, since they are the majority on the dataset, and they have the potential to spend more money.

5.2 Revenues per product category by Age Groups & Monthly Incomes



Revenue by Completed orders by Age Groups per product category

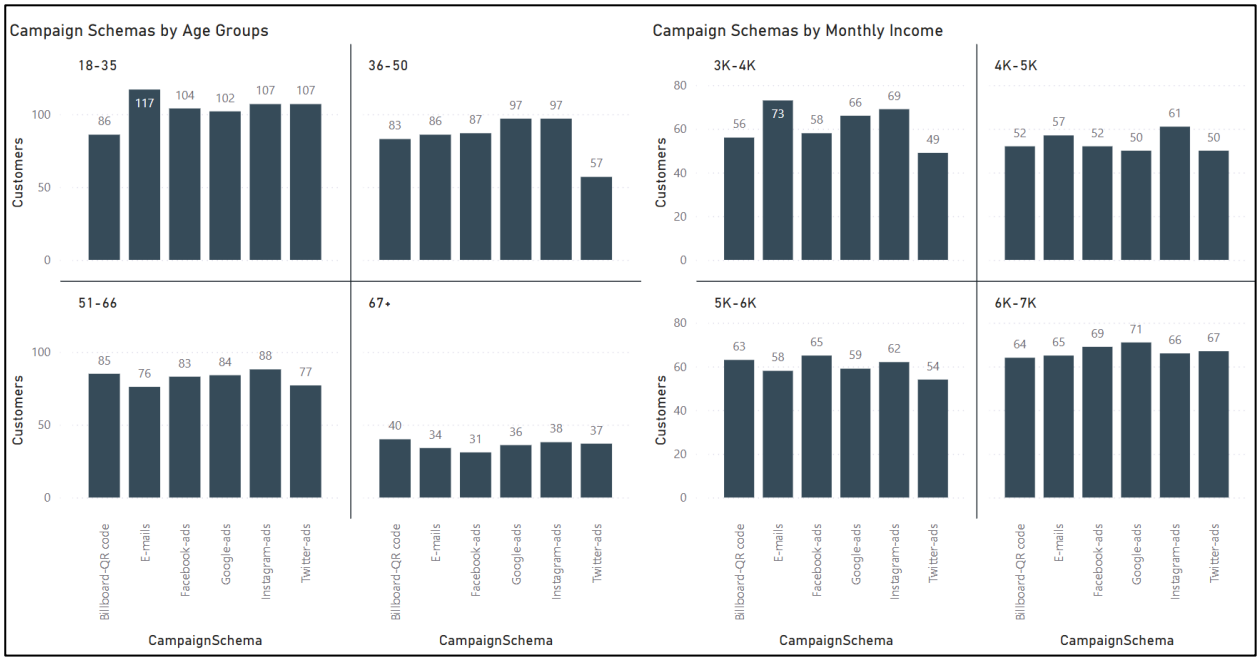


Revenue by Completed orders by Monthly Income per product category

Strategic Importance:

- Pricing Strategies: By identifying the top selling product categories by age group or monthly income, the company can establish an effective policy strategy among various customer clusters and implement various loyalty programs.
- Marketing Strategies: Marketing messages for communicating products can be adjusted if the company has identified variations in revenue produced by certain customer clusters.
- Market Positioning: Possible market penetration or abandonment decisions might be taken, and the company can understand better what the position obtains against the competition.

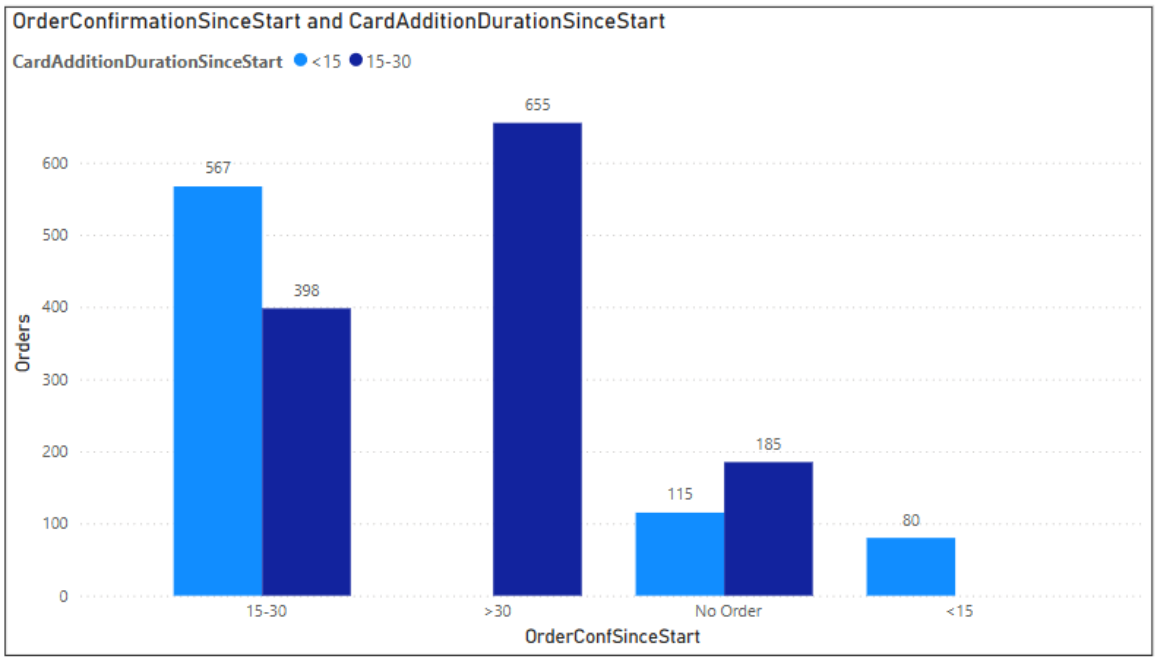
5.3 Age Groups and Monthly Income of Customers reached by Campaign Schema



- **Marketing Channels:** Different platform selection might be needed to communicate better across age or income groups, so adapting the tone of voice and look and feel of the company might be needed.
- **Branding Approach:** Branding and marketing strategies may not overlook the dominant age groups or income groups, as they could be the guideline for the branding strategy based on their product preferences.
- **Product assortment:** Introducing new products or allocating existing ones according to customers' age or monthly income to boost sales.
- **Loyalty planning:** Consumer behavior depends on monthly income; hence, different loyalty programs can be implemented so customers may respond better to promotional messages and perks.

6. Operational Analysis

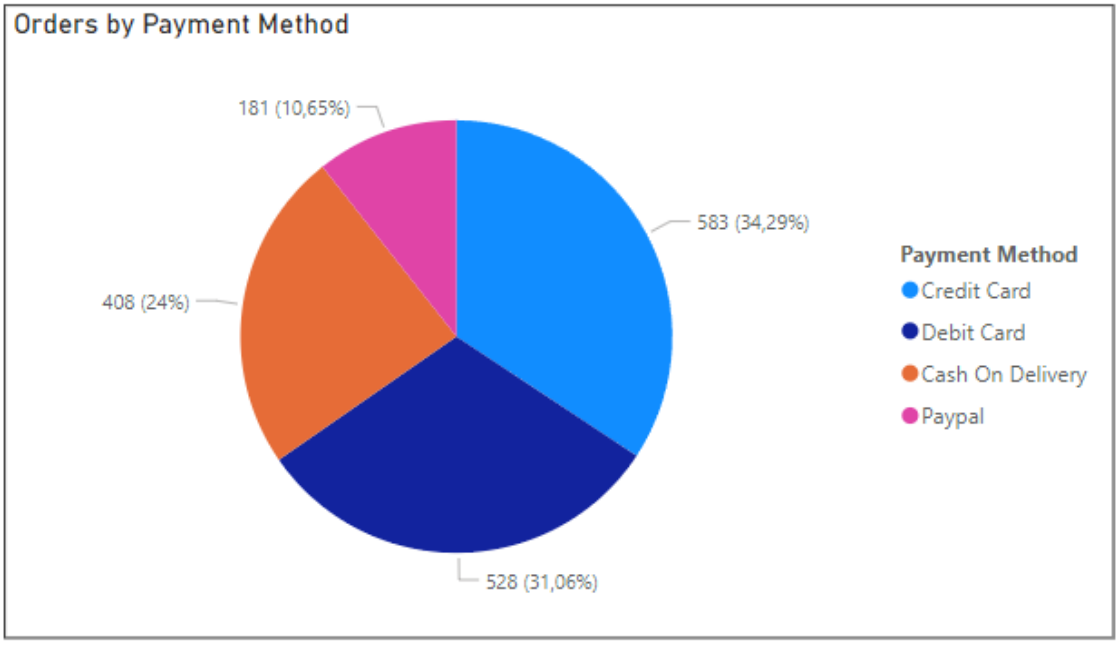
6.1 Order Confirmation and Card Addition Duration (Minutes)



Customer Segmentation Plan:

- Customers taking over 30 minutes to confirm orders (655) represent a majority, indicating potential hesitation or distractions. For these Hesitant Buyers: Providing pushes like reminders or limited time offers to encourage quicker confirmations as well as flash discounts, free shipping options and abandoned cart reminders could be helpful.
- Faster card additions (<15 minutes) seem linked to quicker order confirmations which shows the Quick Decision-Makers: One click checkout options and rewards for immediate actions may take place.

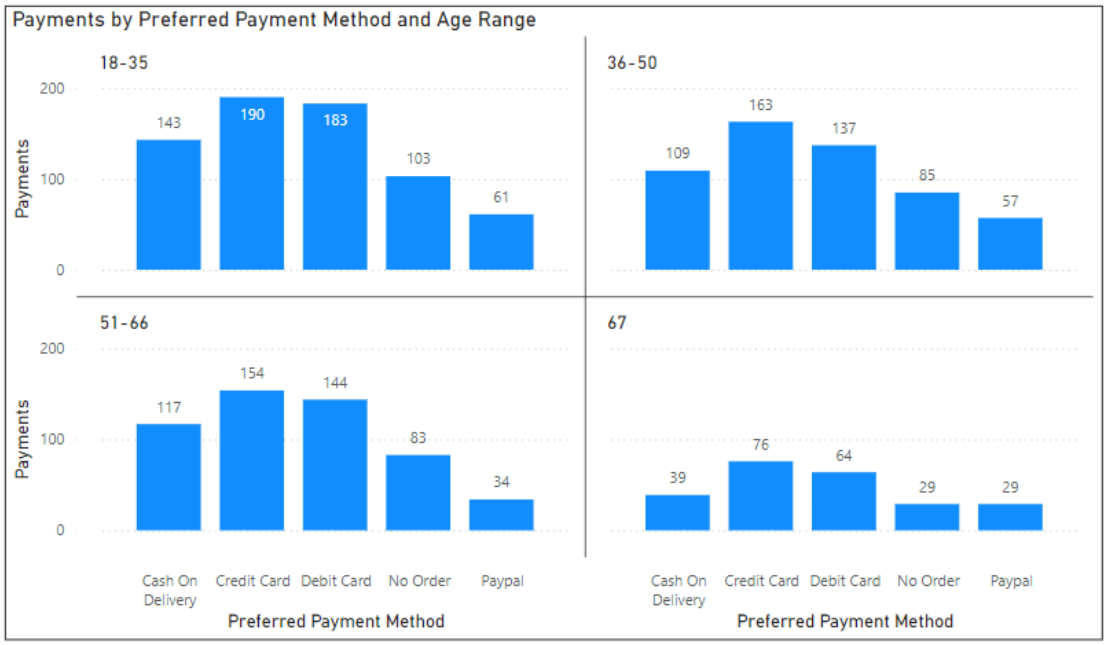
6.2 Orders by Payment Method



Customer Segmentation Plan:

- Credit cards are the most popular, followed by cash on delivery, showing different preferences. For these Credit/Debit Card Users: Offering cashback or loyalty points to boost usage as well as other per credit card company specific incentives is applicable.
- Cash Buyers: Ensuring secure and reliable cash on delivery options along with optimizing delivery timing.
- PayPal is underutilized, possibly due to limited awareness or availability. There could be special discounts given for payments with PayPal if the company prefers it over other payment methods.

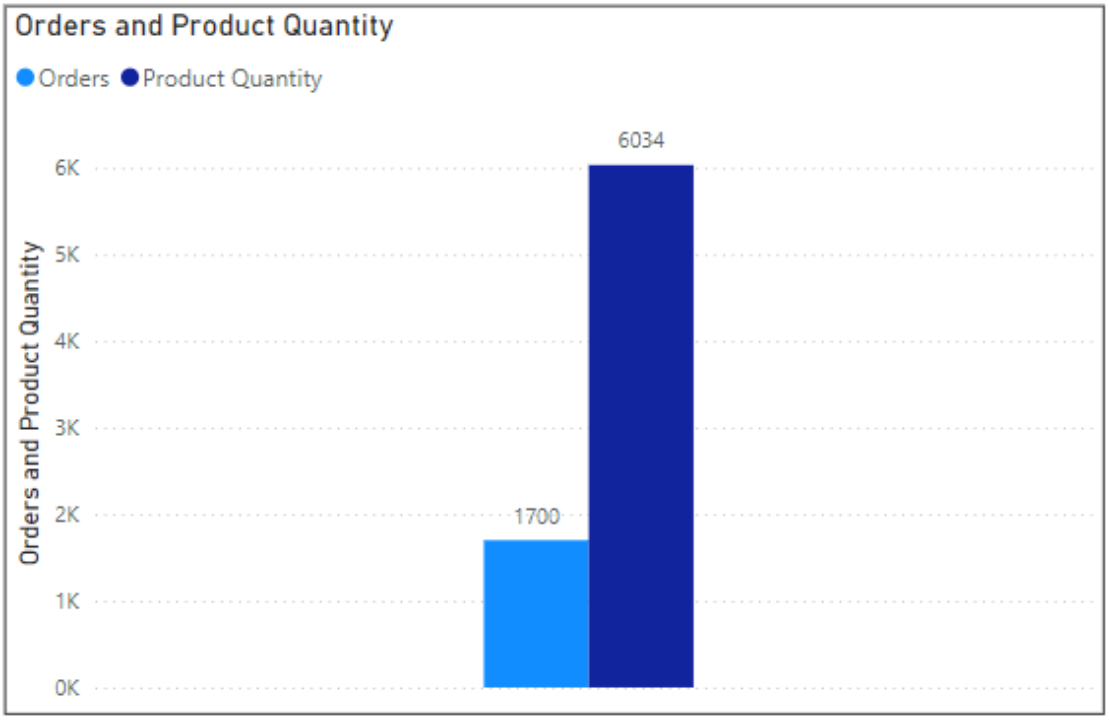
6.3 Segmentation by Age Group & Payment Methods



Customer Segmentation Plan:

- Younger customers (18-35) are more open towards credit cards and digital payments. This age group represents the Tech-Savvy Buyers: Promoting digital wallets & quick checkout options could be established.
- Middle-aged customers (36-50) are equally split between credit and debit cards, meaning financing options could be effective. These are more Financially Stable & Budget Conscious Buyers: Offering cashback rewards, store credits and more installments could boost sales for this age group.
- Older age groups (51+) prefer debit cards and Cash on Delivery, showing a need for trust-based transactions. These are more Trust-Oriented Buyers: Providing product & services reviews, fraud protection messaging, and secure transaction guarantees as well as simplified checkout will bring better results for this age group.

6.4 Orders and Product Quantity



Customer Segmentation Plan:

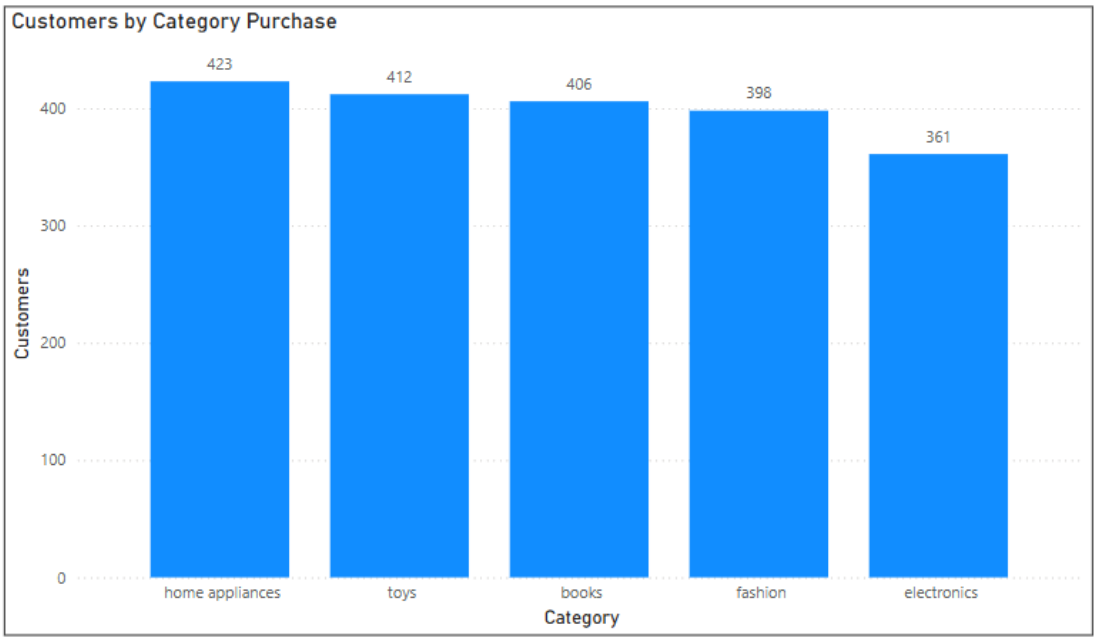
- The high product quantity shows that customers frequently buy multiple items in one order.
- Bulk Buyers: Introducing bundle offers and discounts for multi item purchases.
- Occasional Shoppers: Highlighting savings or coupons on additional products to increase cart size.

6.5 Orders by Product Quantity in Order

Customer Segmentation Plan:

- Orders with 3 items are the highest (350), indicating customer interest in medium-sized purchases.
- Small Cart Shoppers: Offering free shipping for orders above a certain amount to encourage larger purchases.
- Medium Cart Shoppers: Promoting combo deals and add on discounts.

6.6 Customers by Category Purchase

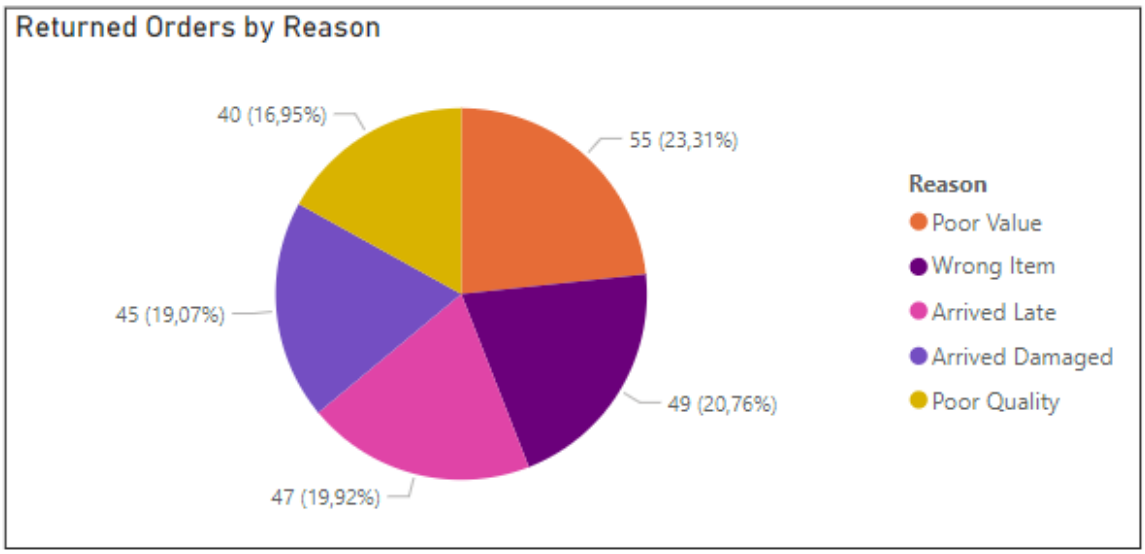


Customer Segmentation Plan:

- Home appliances and toys are the most popular categories, while electronics have fewer customers.
- Home Appliance Buyers: Providing extended warranties and free installations.
- Electronics Buyers: Launching tech guides and product showcases to drive interest. Making regular discounts or bundle products or supplying gift cards on electronics purchase to use in their next orders.

6.7 Order Returns

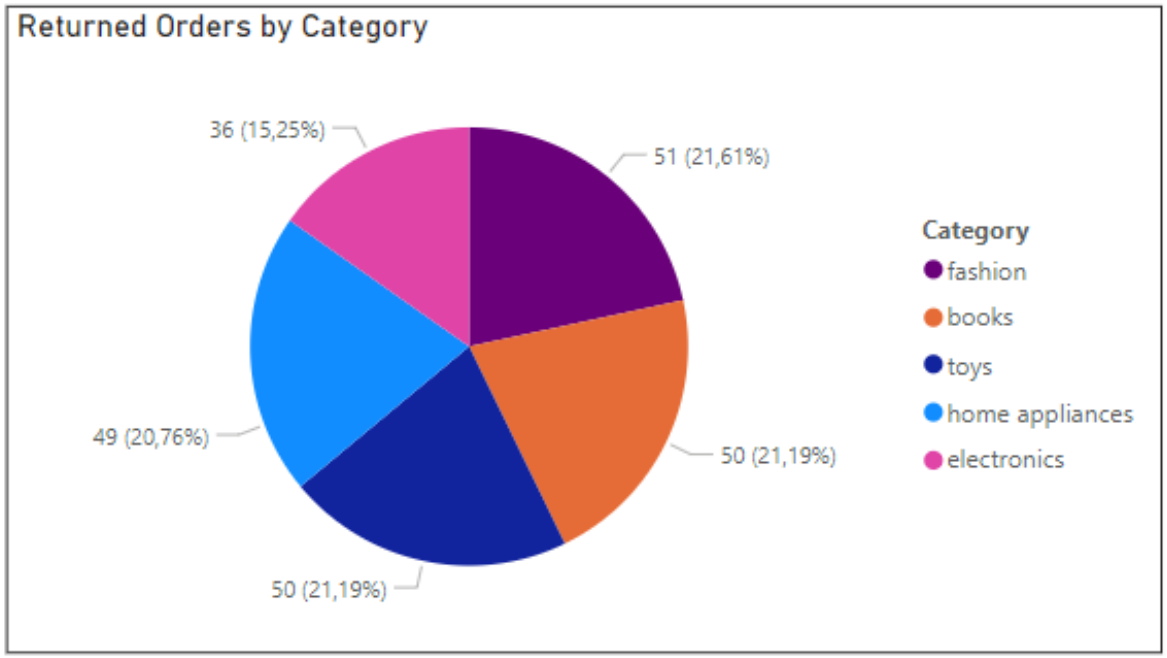
Returned Orders by Reason



Customer Segmentation Plan:

- Poor quality and damaged products are the top return reasons, emphasizing the need for quality control. Quality-Focused Customers: Implementing stricter quality assurance and product reviews.
- Late deliveries also affect customer satisfaction, showing logistics as a critical area. Time-Sensitive Buyers: Offering better shipping estimates and express delivery options.

Returned Orders by Category

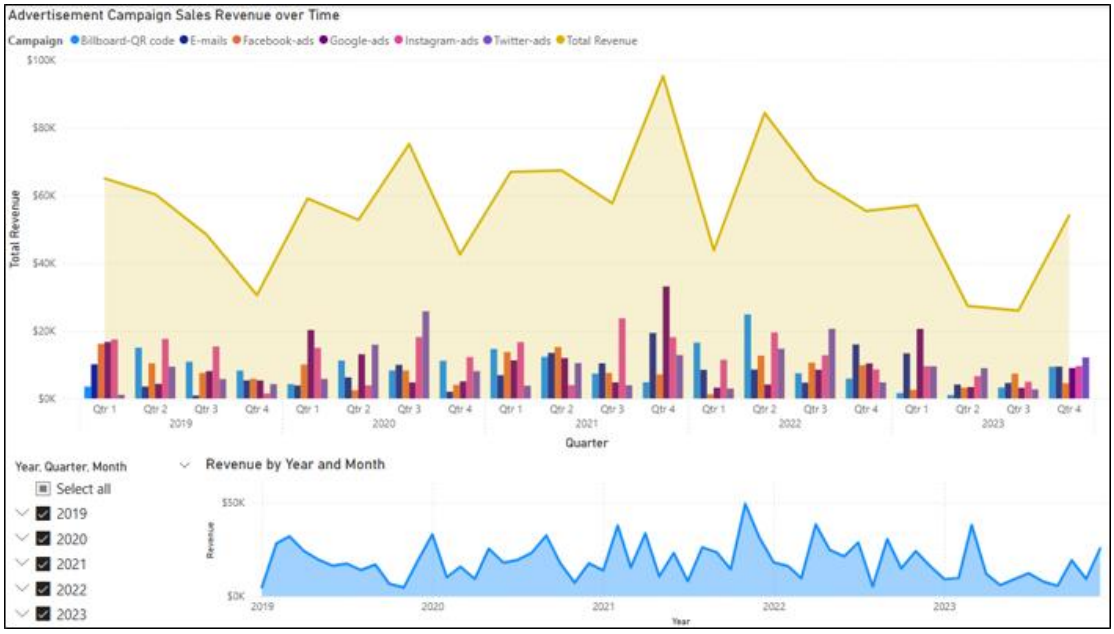


Customer Segmentation Plan:

- Books have the highest return rate, possibly due to unclear descriptions or misaligned expectations. Book Buyers: Enhancing product descriptions, adding previews, and verifying accuracy.
- Fashion has the lowest returns, indicating well matched customer size and preferences. Fashion Customers: Focusing on more accurate sizing and quality details.

7. Financial Analysis

7.1 Campaign Revenue



Graphs that show the total Revenue over the years, as well as the Revenue each advertisement campaign produced for every quarter. The graphs were made in Power BI and the year showcased can be filtered through slicers. The graphs have adjustable hierarchy of date values, ranging from distributing Revenue by Years to Days.

Strategic Importance:

• The overall revenue trend shows fluctuations, peaking in 2020 and 2022, suggesting successful campaigns or seasonal boosts during these periods.

• Social media platforms like Facebook, Instagram, and Google Ads appear to be consistently contributing to revenue, indicating effective online marketing strategies.

• Traditional methods like Billboard-QR codes have a minimal impact compared to digital platforms, suggesting a possible shift in focus towards digital marketing for better returns.

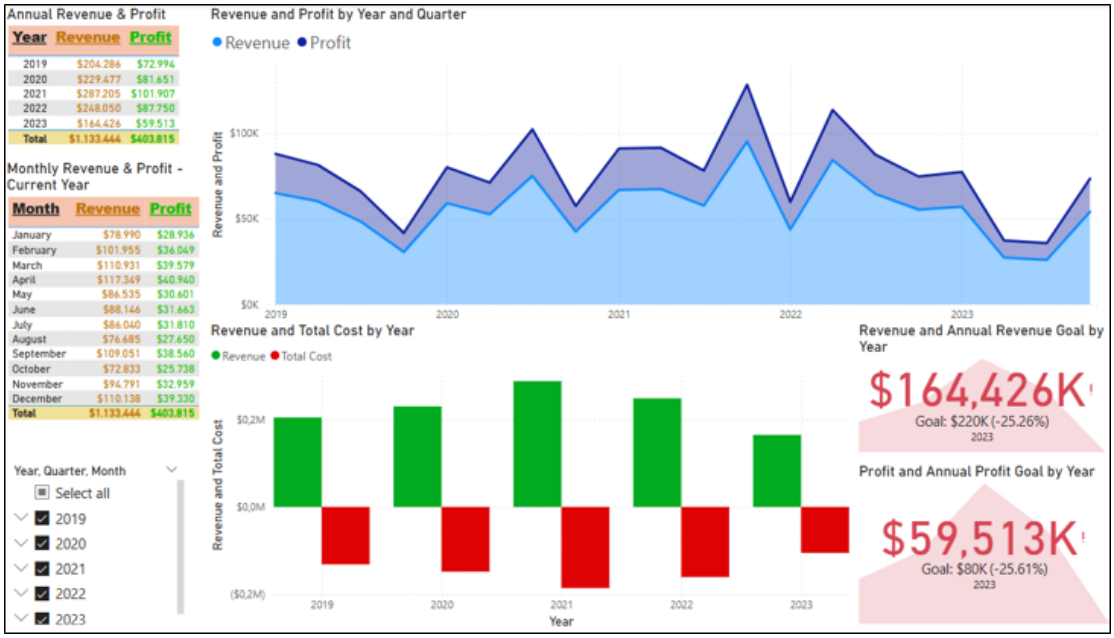
Business Strategy Plan:

• Leveraging High-Performing Channels: Increasing investment in the top-performing digital platforms, especially during peak seasons.

• Exploring Underperforming Channels: Finding the reasons behind the low impact of traditional advertising and reallocating budget.

• Seasonal Campaign Planning: Developing targeted campaigns for periods identified as high-revenue quarters to maximize profits.

7.2 Profit and Revenue Goals



In the graph above, the revenues and profits are compared for each year and the yearly revenues are compared to the costs. Yearly goals for 220,000$ in revenue and 80,000$ in profit are set. The interactive dashboard shows results for all years, but also each year separately.

Strategic Importance:

• Despite consistent revenue growth from 2019 to 2022, 2023 shows a significant downfall, indicating potential market challenges or operational inefficiencies.

• The company fell short of its revenue and profit goals for 2023, indicating the need to reconsider financial planning.

Business Strategy Plan:

• Cost Control: Identifying and reducing high-cost areas to improve profit margins.

• Revenue Diversification: Developing new revenue streams or enhance current offerings to combat declining sales.

• Goal Adjustment: Reevaluating revenue and profit targets for realistic and achiavable goals in 2024.

7.3 Category Revenue

The graphs indicate how the categories performed throughout the years. The waterfall chart especially distributes fall-offs or rises. The data is available for any selection of years.

Strategic Importance:

- Electronics Lead: Electronics drive the highest revenue ($700,670) but show downfall in 2019 and 2023.
- Home Appliances: Steady growth ($306,770) reflects consistent demand.
- Fashion Volatility: Fashion fluctuates, with peaks in 2021 but drops in other years.
- Books & Toys: Minimal impact on revenue, more like niche categories.
- Seasonal Trends: Revenue spikes during Q4 indicate holiday sales.

Business Strategy Plan:

- Electronics: Focusing on quality and customer service to reduce downfalls and to boost loyalty.
- Home Appliances: Expanding the product range and offering bundle deals to sustain growth.
- Fashion: Targeting marketing efforts during peak seasons to stabilize revenue.
- Books & Toys: Promoting as add-ons or gift items to maximize sales.
- Seasonal Campaigns: Aligning promotions with holiday trends for higher impact.

7.4 Product Revenue

These graphs showcase the performance of each product separately. Specifically, column charts rank the products by the top 5 in revenue and amount sold and filters all the rest to a 'Others' category, all through the usage of DAX in Power BI.

Strategic Importance:

• High-value products like televisions and laptops generate the most revenue but may have higher return rates, impacting net profits.

• Lower-priced items, such as microwaves and frying pans, have higher quantity sales but lower revenue impact, indicating a volume-based sales strategy.

• The return rate is low compared to total orders, suggesting customer satisfaction with product quality or efficient customer service.

Business Strategy Plan:

• Focusing on High-Value Products: Prolonged warranty or support services for high-value products to reduce returns and boost customer satisfaction.

• Volume Sales Strategy: Continuing to promote lower-priced items as impulse buys or bundled offers to increase overall sales.

• Customer Feedback Utilization: Gathering feedback on returned items to improve product quality and reducing future returns.

7.5 RFM Score

"The RFM score represents the value you give to each variable used in an RFM analysis: recency, frequency, and monetary value. The RFM score is a numerical score that helps you recognize all types of customers, from the best to the worst."[2]

For the purposes of this assignment, an RFM Score will be calculated using short Python scripts for each one of the 1103 customers recorded in our dataset. As stated above, the score is broken down by:

1. **Recency**: How recently a customer has made a purchase
   Recency was calculated by subtracting the number of days from the latest recorded purchase by the most recent order of purchase of each customer, as shown below.

```
# Calculating Customer Recency
# Groups the customers by their details and their most recent purchase date
cust_df = df.groupby(by=features, as_index = False)['OrderConfirmationTime'].max()
recent_date = cust_df['OrderConfirmationTime'].max()

cust_df['Recency'] = cust_df['OrderConfirmationTime'].apply( lambda x: (recent_date - x).days)
cust_df.head()
```

2. **Frequency**: How often a customer makes a purchase
   Frequency is the total of purchases a customer has ever made on record.

```
# Calculating Customer frequency
df_freq = df.groupby(by=['CustomerID', 'FullName'], as_index = False)['OrderConfirmationTime'].count()
cust_df['Frequency'] = df_freq['OrderConfirmationTime']
cust_df.head()
```

3. **Monetary value**: How much money a customer spends on purchases

```
# Revenue
df['Revenue'] = df['Price'] * df['Quantity']

# Calculating total spent per customer
df_tspent = df.groupby(by=['CustomerID', 'FullName'], as_index = False)['Revenue'].sum()
cust_df['Monetary Value'] = df_tspent['Revenue']
cust_df.head()
```

The values are later normalized in order to prevent issues with scalability.

```
#Compute numerical data ranks (1 through n) along axis. By default, equal values are assigned a rank that is the average of the ranks of those values.
cust_df['R_rank'] = cust_df['Recency'].rank(ascending=False) # Lower values get assigned higher ranks
cust_df['F_rank'] = cust_df['Frequency'].rank(ascending=True) # Higher values get assigned higher ranks
cust_df['M_rank'] = cust_df['Monetary Value'].rank(ascending=True) # Higher values get assigned higher ranks

# Normalizing ranks, diving by the highest rank and multiplying the result with 100
cust_df['R_rank_norm'] = (cust_df['R_rank']/cust_df['R_rank'].max())*100
cust_df['F_rank_norm'] = (cust_df['F_rank']/cust_df['F_rank'].max())*100
cust_df['M_rank_norm'] = (cust_df['F_rank']/cust_df['M_rank'].max())*100

cust_df.drop(columns=['R_rank', 'F_rank', 'M_rank'], inplace=True)
```

A formula is used in order to calculate the RFM Score, the weights of which are adjusted accordingly to the importance each of the three attributes have for our company. In our case, frequency is more valuable, followed by frequency. The score is then multiplied by 0.05 in order to scale it down to a measure of 1 to 5.

RFM' = (0.15 *x* **R**) + (0.25 *x* **F**) + (0.6 *x* **M**)

RFM = RFM' *x* 0.05

```
cust_df['RFM_Score'] = (0.15*cust_df['R_rank_norm']) + (0.25 * cust_df['F_rank_norm']) + (0.6*cust_df['M_rank_norm'])
cust_df['RFM_Score'] *= 0.05
cust_df = cust_df.round(2)
```

Each customer is now assigned an RFM score. The customers are now segmented according to their respective score:

- RFM score >= 4.5 : Top Customer
- 4.5 > RFM score >= 4 : High Value Customer
- 4 > RFM score >= 3 : Medium Value customer
- 3 > RFM score >= 1.5 : Low Value customer
- RFM score < 1.5 : Lost Customer

A label is assigned to each customer, which describes their current value for the e-commerce. Promotional strategies can now be applied to focus on high value customers and reengage lower value ones.

8. Machine Learning Techniques

In this assignment's use case, we enhance on the analytical nature of our problem with predictive modeling, by using Machine Learning techniques.

"Machine learning is a unified algorithmic framework designed to identify computational models that accurately describe empirical data and the phenomena underlying it, with little or no human involvement." [3]

By employing supervised and unsupervised learning algorithms on the preprocessed data in our possession through Python[1], we attempt to predict probabilities and forecast sales revenues, as well as realize the patterns hidden beneath our data. These patterns and trends distill analytical and descriptive character to our data that would normally not be accessible and discernible by the human factor alone.

8.1 Predicting Probability of Purchase

The first technique we apply to predict how probable it is for a customer to purchase a product from the company is a Classifier, more specifically, called Logistic Regression. Below, we endeavor to summarize the process we followed to employ this Machine Learning model, while performing the necessary tuning to our customer purchase record dataset:

- The most relevant features (characteristics) of a customer were kept for the training of the model, such as their Gender, Age, Income, Country of origin and the advertisement campaign that initially attracted them. The features were then encoded, making them readable to the machine.
  The target variable was the 'OrderConfirmation' column, which was labeled as True or False, depending on if the customer moved forward with their purchase or not.

	Gender	Age	CreditScore	MonthlyIncome	Country	CampaignSchema	OrderConfirmation
0	Male	57	780	7591	China	Instagram-ads	True
1	Female	69	746	3912	China	Google-ads	True
2	Female	21	772	7460	UK	Facebook-ads	True
3	Female	67	631	4765	UK	Twitter-ads	True
4	Male	57	630	3268	China	Billboard-QR code	True

Table 1: Features and Target used for Probability Prediction of purchase

[1] The Python scripts can be found in the deliverables of the assignment. Based on the nature of this course, they will not be analyzed in depth for this assignment report.

- We confirm the importance of our features as a changing factor for the target variable and then standardize them, a process which allows for scalability between higher and lower values, eliminating misinterpretation of ranges across the feature space.

- The appearance ratio of the two classes of the target variable (in our case, True or False) is:

  True: 1700 values
  False: 300 values

  This makes the training of the model redundant, as the classifier will overfit and become biased against predicting an essential probability about the 'False' label. We decided to implement an oversampling method in order to restore this balance to an adequate level.

- We train the model on the dataset and evaluate its results on unseen data; a testing set we extracted from our dataset. This evaluation is performed with the 'Recall Score', which is better suited for the Logistic Regression algorithm, as it is probabilities that are being predicted and the actual accuracy of the predicted labels is of no importance.

```
Recall score on testing set: 0.8436578171091446
```

- A data frame of iterated elements that consists of all possible value ranges and labels in the aforementioned features is created. A collection of 13500 generated customer characteristics will be tested.

	Gender	Age	CreditScore	MonthlyIncome	Country	CampaignSchema
0	0	20	600	3500	0	0
1	0	20	600	3500	0	1
2	0	20	600	3500	0	2
3	0	20	600	3500	0	3
4	0	20	600	3500	0	4
...	...	...	...	...	...	...
13495	1	60	780	7500	8	1
13496	1	60	780	7500	8	2
13497	1	60	780	7500	8	3
13498	1	60	780	7500	8	4
13499	1	60	780	7500	8	5

13500 rows × 6 columns

Table 2: Iterated combinations for each potential future customer

- The model is called to predict the percentages of probability for a customer, characterized by each combination as illustrated above. Once the probabilities are generated, they are appended to the iterated table. All labels are de-encoded and named after their original feature values as shown below.

	Gender	Age	CreditScore	MonthlyIncome	Country	CampaignSchema	No Order	Order
0	Female	20	600	3500	Australia	Billboard-QR code	0.563	0.437
1	Female	20	600	3500	Australia	E-mails	0.567	0.433
2	Female	20	600	3500	Australia	Facebook-ads	0.570	0.430
3	Female	20	600	3500	Australia	Google-ads	0.574	0.426
4	Female	20	600	3500	Australia	Instagram-ads	0.577	0.423

Table 3 First 5 rows of the generated dataset for potential future customers

- The table is imported into Power BI. With the combinations and probabilities it provides, we create an interactive page that allows the user to predict the possibility of all 13500 different customers, by selecting each of their characteristics by themselves. The dashboard can be found in the PBI report file attached to the deliverables of the assignment.

Table 4: A showcase of 'Probability of Confirmed order', the interactive page within Customer360Insights.pbix

8.2 Sales Forecast using XGBoost Regressor

Timeseries forecasting is commonly deployed to predict the margin of future sales revenue a company can produce. For this machine learning model, we used the XGBoost Regressor in order to predict the sales values over time.

- The only values kept are the date of the confirmed order and the revenue produced. We calculate the total of revenue for each day recorded in the dataset and list it as shown below.

	OrderConfirmationTime	Revenue
0	2019-01-01	200
1	2019-01-02	480
2	2019-01-03	0
3	2019-01-04	40
4	2019-01-05	640

Table 5: First 5 rows of table used for the XGB model

- For better feature engineering the Year, Month and Day are extracted from each date and the dates themselves are discarded.

- We split the dataset into a training set, that will be used for the model, and a testing set, which will later evaluate our results. The training set contains all dates and revenues for 2019-2022, while the testing set contains all values for the year 2023.

- After normalizing the model to eliminate any scalability issues, we train the model and evaluate its results as acceptable.

```
[0]     validation_0-rmse:1868.72861    validation_1-rmse:1708.46439
[100]   validation_0-rmse:1723.45427    validation_1-rmse:1611.02011
[171]   validation_0-rmse:1693.43081    validation_1-rmse:1612.72453
```

Table 6: Validation of results for every 100th prediction by Root-Mean-Square Erro

- The model is given each date for the next year, 2024, and it predicts new values that provide an adequate insight of the possible performance of the company in the next financial year.

- The results are appended in the table created on the first step and imported to Power BI. Two graphs are made in the report, which compare the performance of the model to the real values for 2023 and combine the current trendline of the Revenue for past dates with the predicted ones for 2024.

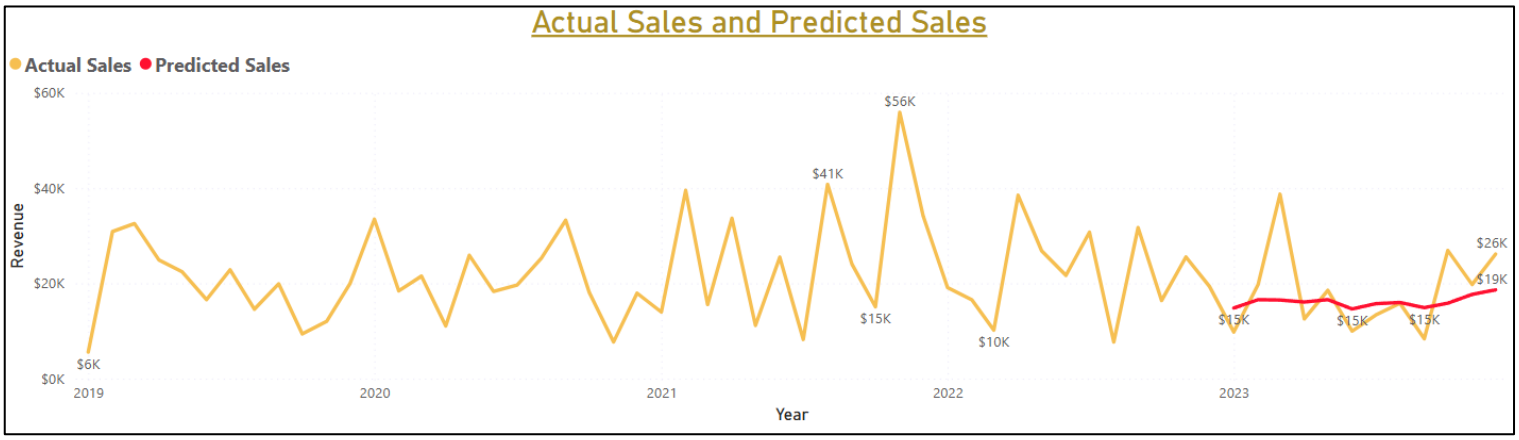


Fig. 1: Comparison of Actual Revenue and Predicted Revenue for 2023. The recorded values are colored yellow, while the predicted ones are colored red.

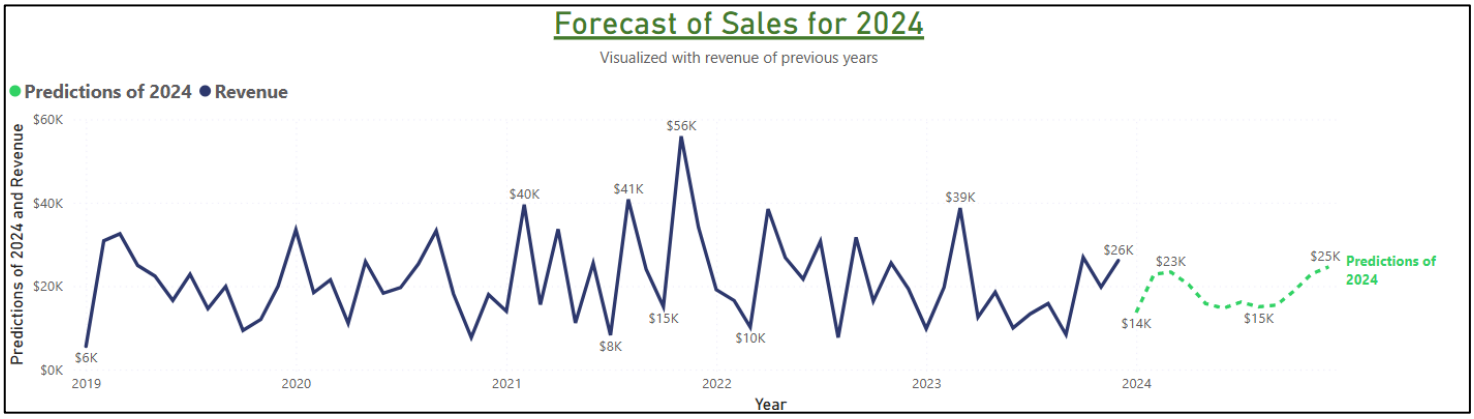


Fig. 2: Recorded Revenue and 2024 Forecast

8.3 Customer Clustering Analysis

Clustering techniques are popular for segmenting customers and analyzing their behavior by using Machine Learning models that group data based on patterns usually unrecognizable by humans alone. After choosing the '*Gender, Age, CreditScore, MonthlyIncome, Country, RFM_Score'* columns for our features used in the model, we proceed to tune the clustering model:

- The data is encoded and normalized, preparing it for training.

- A grid of various plots consisting of pairs of features is made. This showcases the distribution of values across the feature space. It is evident that the dimensionality of our dataset will pose a problem for the model.

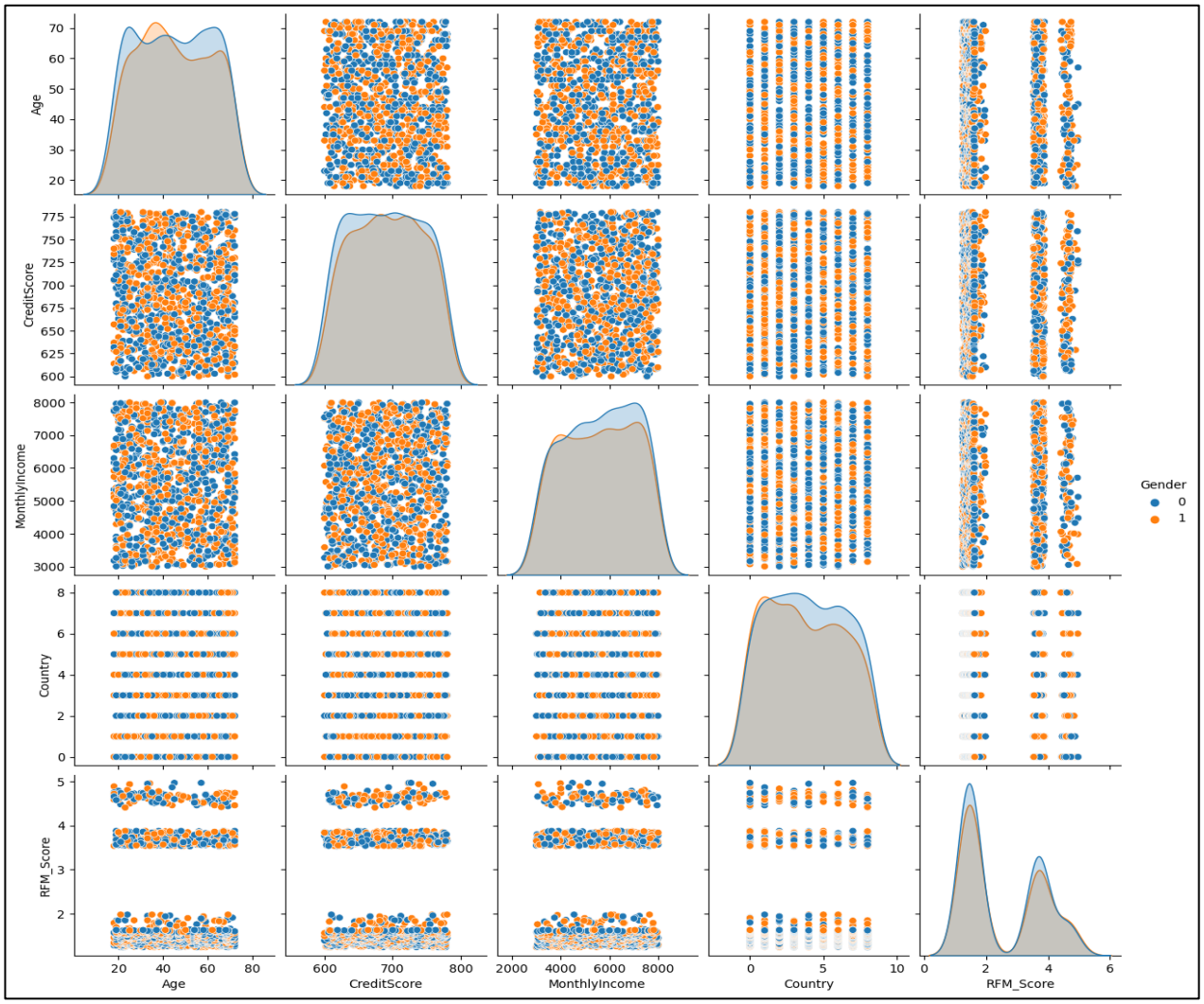


Fig. 3: Paired plots of features, showing their distribution

- We perform PCA (Principal Component Analysis) in order to decrease dimensionality to n=3. By plotting the features in 3d space, what was not clear before becomes much clearer now; The data can be clustered into groups.

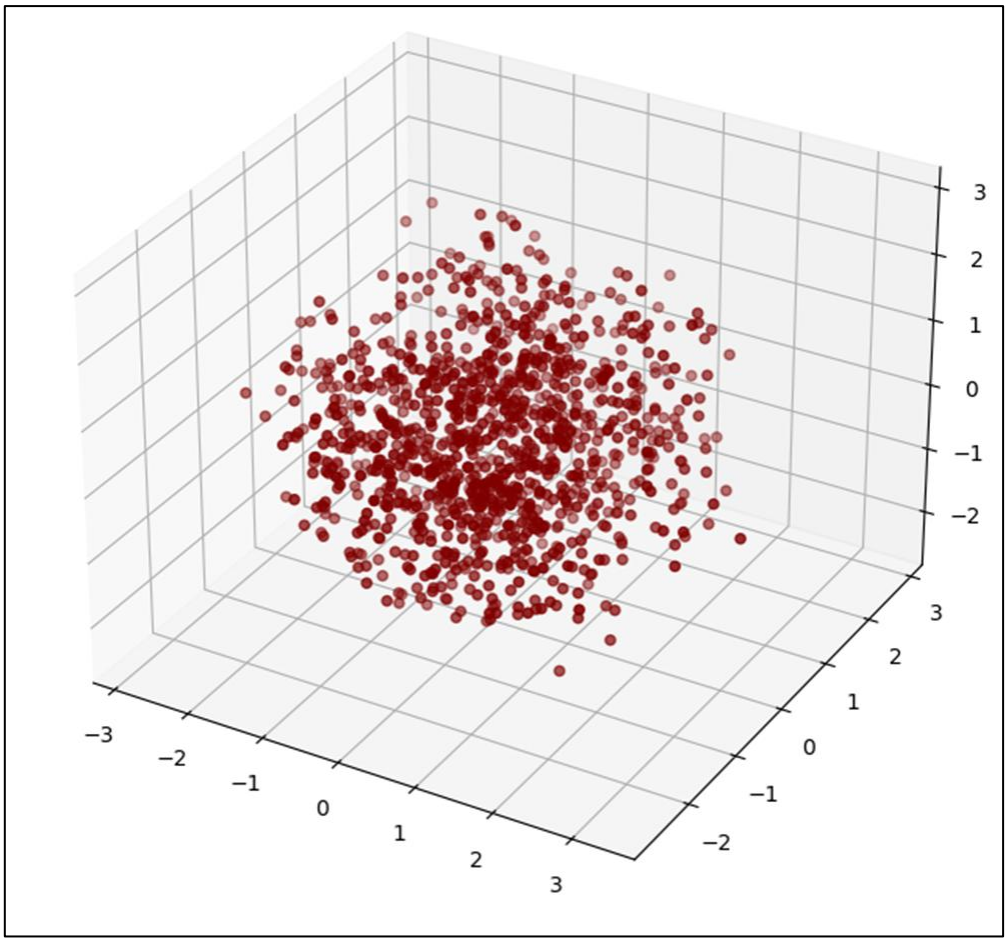


Fig. 4: A 3D Projection Of Data In The Reduced Dimension

- After training the model to separate the customers into 4 groups, graphs were created to better describe the characteristics of each cluster.

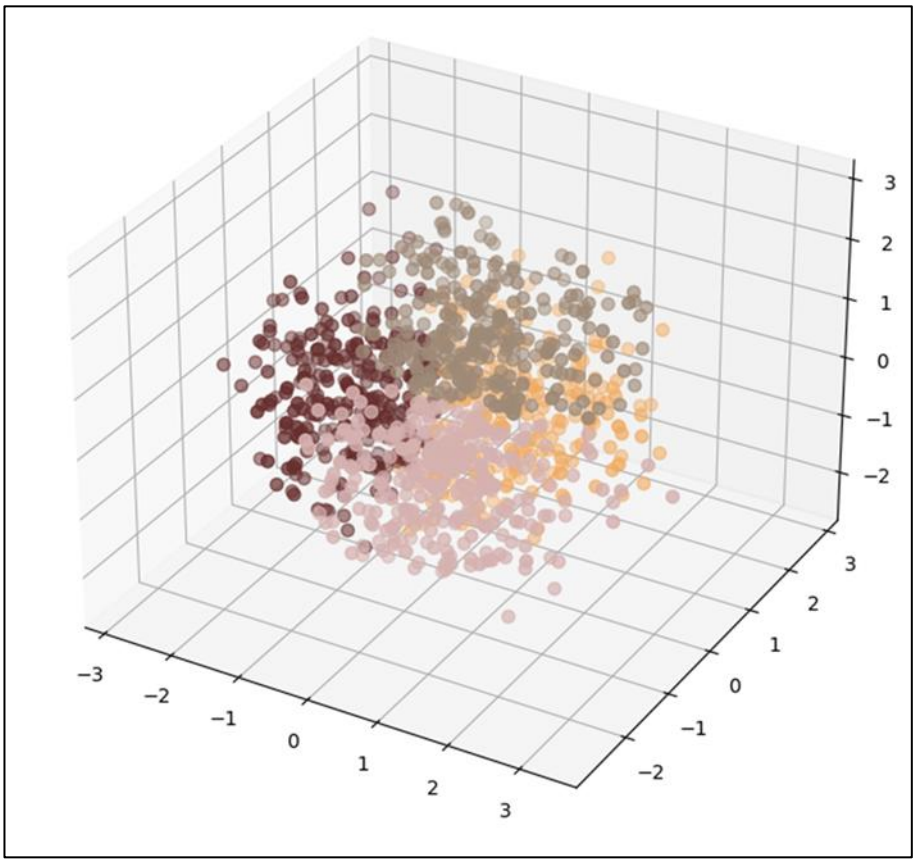


Fig. 5: The plot of the Clusters

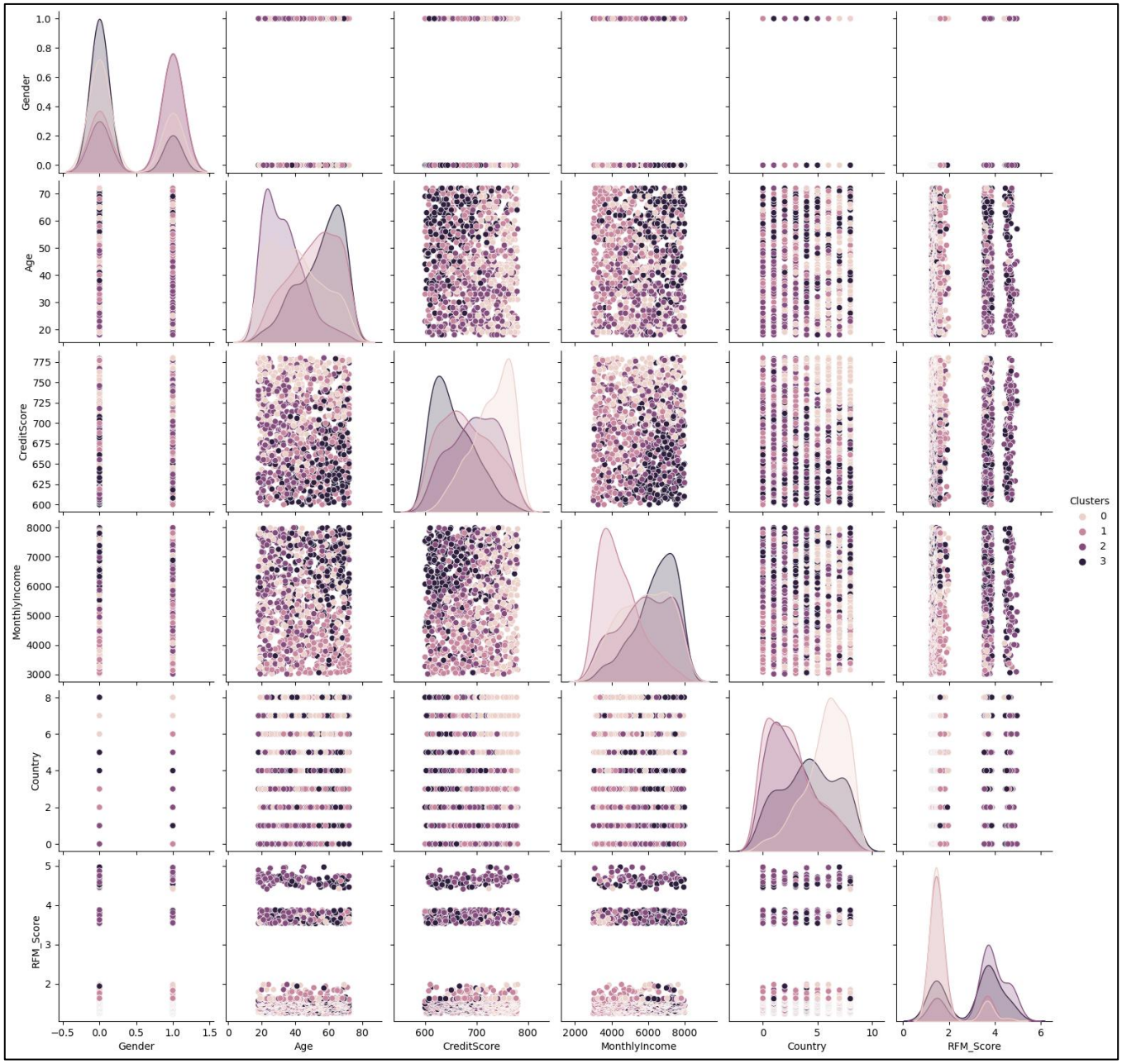


Fig. 6: Feature pair plots colored by clusters

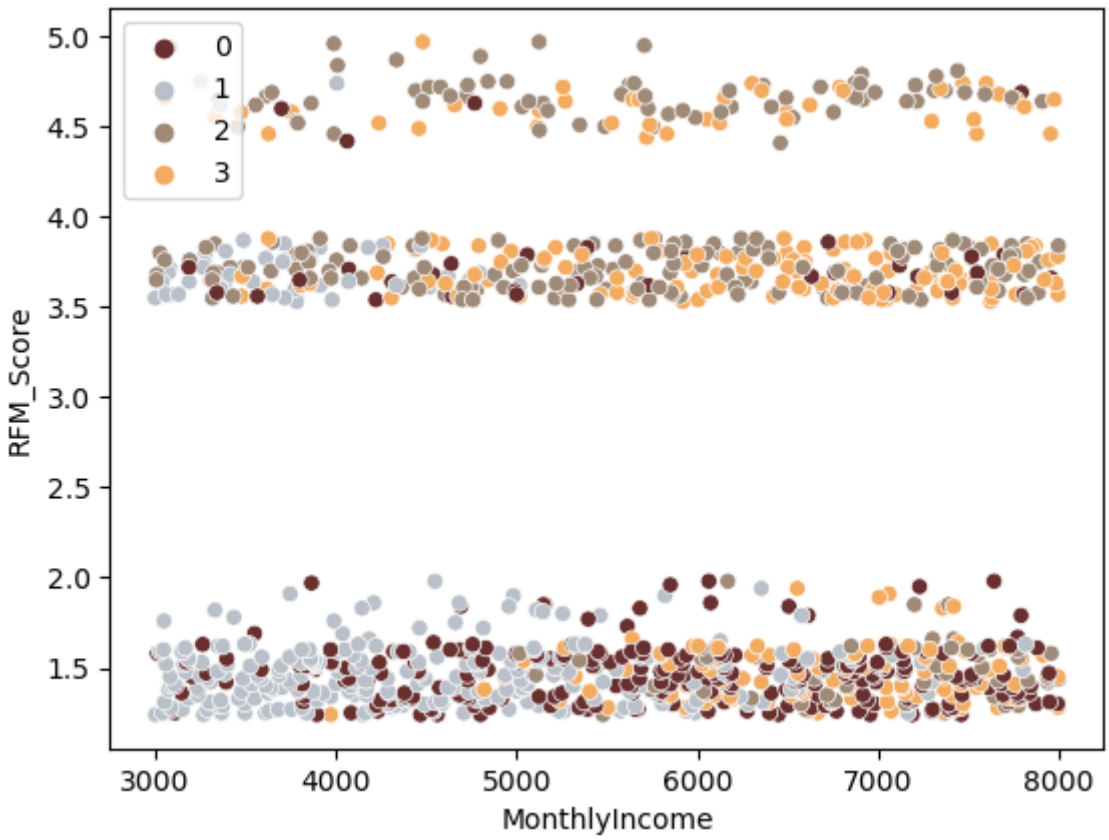


Fig. 7: Clustering on Income and RFM

- As observed in Fig. 7, each cluster/group has noteworthy attributes that make them distinct from the others. In this specific case, we remark the income and the spending score as the main features of each cluster.

    - **Cluster 0** : High Income, Low Spending Score
    - **Cluster 1** : Lower Income, Low Spending Score
    - **Cluster 2** : Average Income, High Spending Score
    - **Cluster 3** : High Income, High Spending Score

Each cluster will now be renamed to their respective main attributes.

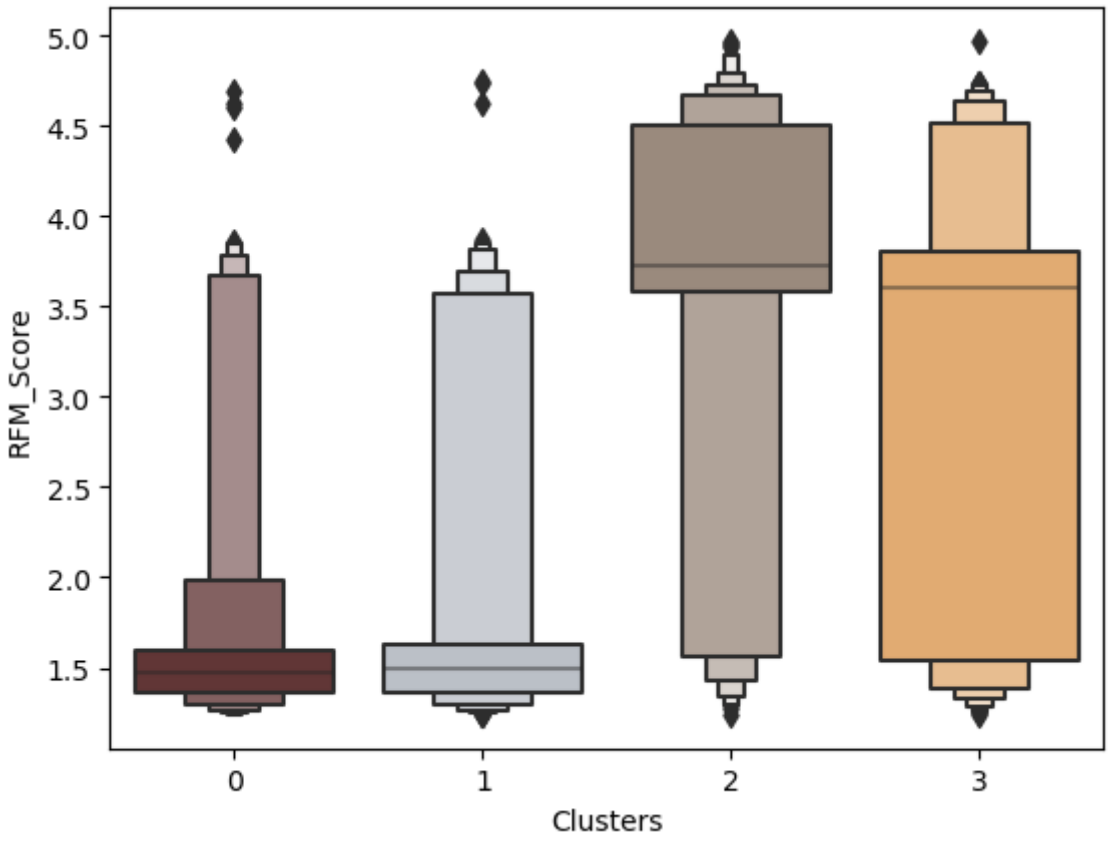


Fig. 8: Box plot of Cluster vs RFM_Score. Cluster 3 has the most clients, while Cluster 2 seconds it.

- Eventually, the appended clustered table of distinct customers is imported into Power BI.

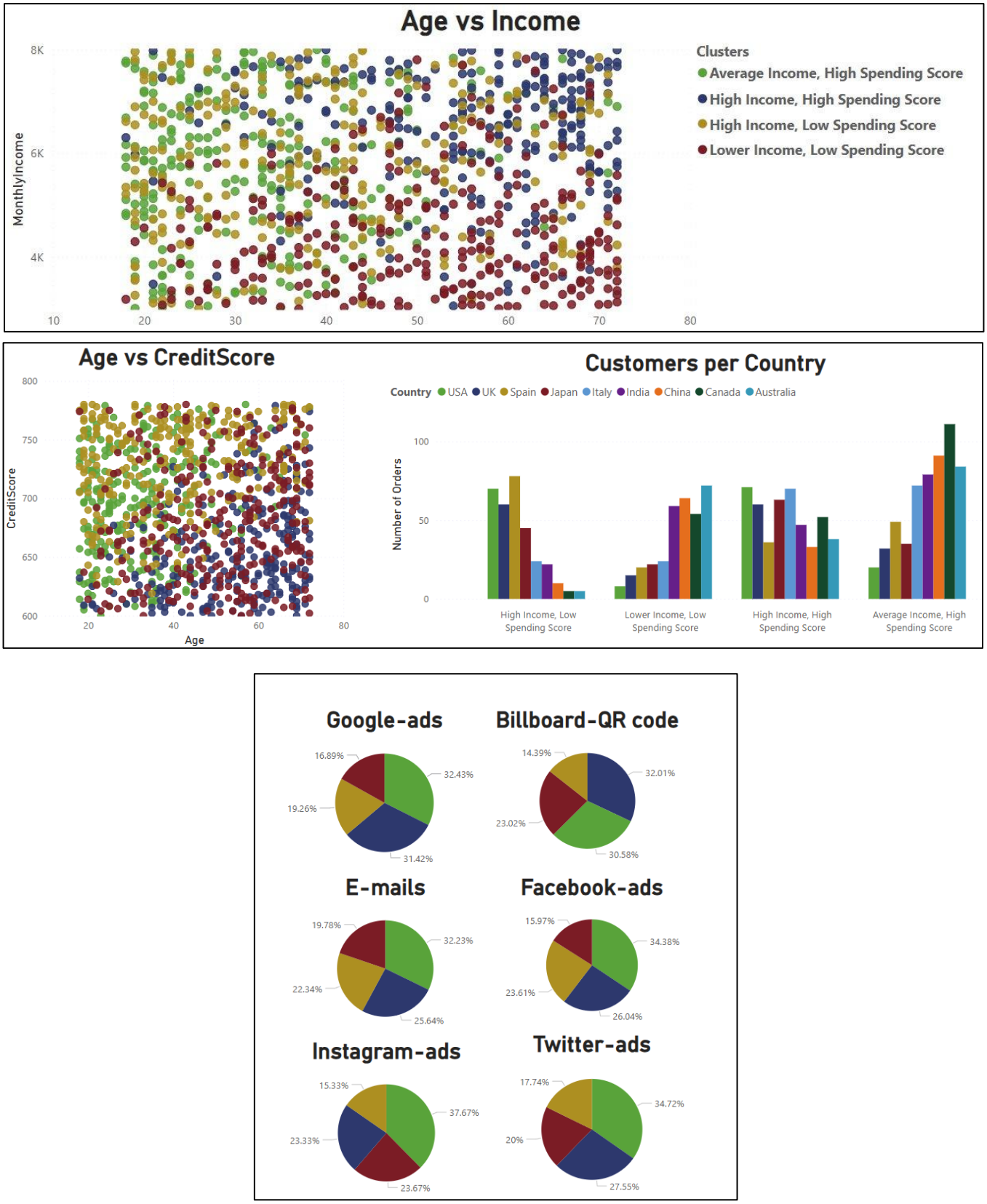


Fig. 9: Various graphs distributing the feature values, colored by each cluster.

First Group:
The first group consists of people of younger age and average credit score. They mostly originate from Canada, China, Australia and India. All advertisement campaigns appear to be effective on them.

Second Group:
The second group consists of people of older age and lower credit score. Their country of origins spans across all recorded countries in the database. Billboard and Google ads seem to be most effective on them.

Third Group:
The third group consists of people of middle age and high credit score. They mostly originate from USA, UK and Spain. E-mails and Facebook seem to draw them toward making a purchase.

Fourth Group:
The first group consists of people of older age and average credit score. They mostly originate from China, Australia and India, with Instagram, Billboard and Twitter ads noting the most effect.

9. Strategic Insights & Business Recommendations

With the insight collected so far, the next step would be to formulate a strategy in order to increase the reach of the company and grow its revenue for next year. Some of the decisions to be made are summarized below:

- **Targeted Marketing**
  Personalizing Campaigns by using the clusters generated and tailor advertisements to each distinct group of customers, according to Fig. 9.
  - Focus on social media for young customers
  - Prioritize Google, billboard and e-mail ads for older customers.
  - Promote through e-mails for high spenders personally.

- **Increase Conversion Rates**
  The probability of purchase insight can be used to target future and existing customers dynamically, by retargeting users who are likely to purchase but have not converted yet, or pushing offers and discounts to users who are less likely to order.

- **Sales Forecasting**
  Managing inventory and invest more resources in high-revenue months, to maximize the revenue produced.

- **Retaining customer segments**
  Prioritizing high-spending customers with offers and loyalty programs, but also exclusive customer support. Encouraging customers with low-spending score

into making purchases by targeting offers.

- **Investing into e-commerce and delivery**
    - Refining the website experience to prevent cart abandonment, as well as encourage hesitant buyers into making a purchase.
    - Expanding Geographically.

10. Customer Segmentation Strategy Decisions

Customer segmentation is essential for optimizing marketing strategies, improving customer retention, and maximizing revenue. This study utilizes a data-driven approach to identify distinct customer groups based on demographic, geographic, transactional, and behavioral patterns. By applying clustering techniques, RFM analysis, and predictive modeling, we define meaningful customer segments and propose targeted strategies tailored to each group's unique characteristics.

High-Value Frequent Shoppers

The first segment consists of customers who frequently make high-value purchases, contributing significantly to the company's overall revenue. These customers tend to prefer premium products, such as electronics and home appliances, and exhibit a strong brand affinity. Geographically, they are primarily concentrated in the USA, UK, and Spain, and they respond well to direct marketing efforts through email and social media campaigns, particularly on Facebook.

To enhance engagement with this segment, the company should implement exclusive loyalty programs that offer early access to promotions and personalized discounts based on past purchases. Additionally, priority customer support and extended warranties can be introduced as incentives to increase customer satisfaction and long-term retention.

Budget-Conscious Shoppers

This segment consists of price-sensitive customers who frequently purchase lower-cost products, such as household essentials and small electronics, but contribute to sustained cash flow due to their purchasing consistency. These customers are highly responsive to discounts and promotional offers and are more likely to engage with bulk purchasing opportunities.

To cater to this group, the company should introduce subscription-based discounts, loyalty reward programs, and bulk purchase promotions. Additionally, cost-effective marketing strategies should be optimized to ensure that digital advertising efforts remain profitable while appealing to this segment's price-conscious nature.

Seasonal Holiday Shoppers

A significant portion of revenue is generated during Q4 (October–December) due to holiday shopping trends. This segment consists of customers who primarily shop during major sales events such as Black Friday and Christmas, often making high-value purchases during these peak periods. They are particularly attracted to bundled product deals, such as television and soundbar packages or kitchen appliance sets.

To maximize engagement with this group, marketing efforts should be concentrated around seasonal promotions, with targeted advertising campaigns launched well in advance of major shopping holidays. Implementing limited-time offers and personalized gift recommendations can further drive conversions during peak shopping seasons.

Tech-Savvy Early Adopters

This segment is composed of customers who actively seek the latest technology and smart devices, including high-end electronics such as televisions, gaming consoles, and laptops. They represent the highest revenue-generating group, contributing over $700,000 in sales. Their purchasing decisions are heavily influenced by online reviews, influencer marketing, and product launches.

To effectively reach this group, the company should implement pre-order promotions and exclusive launch discounts for new products. Additionally, partnerships with influencers and targeted social media campaigns can enhance brand awareness and engagement. Offering product bundle discounts, such as a gaming console with a TV purchase, can further incentivize larger transactions.

11. Conclusion

This study analyzes customer behavior in e-commerce using data-driven techniques. A major challenge was the limited dataset (2,000 orders), which constrained machine learning model performance and deeper analysis. Such a small sample offers only a basic view of potential customer segments.

By leveraging demographic, transactional, and behavioral data, we identified key customer segments and purchasing trends. Using clustering, RFM analysis, and sales forecasting, we showed how businesses can optimize marketing, improve conversion rates, and boost customer retention.

Our findings emphasize the value of personalized marketing, data-driven decision-making, and operational efficiency in maximizing revenue and customer satisfaction. Key recommendations include targeted campaigns, improved checkout experience, and inventory optimization based on sales forecasts.

This research highlights the power of data science in e-commerce, offering a framework to enhance customer experience and drive sustainable growth.

12. References

[1] **Darshan, Dave.** *Customer 360 Insights.* Kaggle, 2023, https://www.kaggle.com/datasets/davedarshan/customer360insights.

[2] A. Panaitescu, “What is RFM Score: How to Calculate, RFM Formula, Example,” rfm-score. [Online]. Available: https://www.omniconvert.com/blog/rfm-score/

[3] J. Watt, A. K. Katsaggelos, and R. Borhani, *Machine learning refined: foundations, algorithms, and applications*. Cambridge: Cambridge university press, 2016.