self-project

MARKETING ANALYTICS

BY SEGMENTING & PREDICTING CUSTOMER BEHAVIOUR

May 15th, 2025

Johanes Purek

TABLE OF CONTENTS

01
PROBLEM RESEARCH

02
DATA INTERPRETATION

03
EXPLORATORY DATA ANALYSIS

04
CUSTOMER SEGMENTATION & BUSINESS INSIGHTS

05
CUSTOMER DASHBOARD

06
CUSTOMER PREDICTION

07
BUSINESS RECOMMENDATION

PROBLEM RESEARCH

- iFood, a top food ordering & delivery service in Brazil and Colombia, is experiencing **challenges** in **sustaining profit growth** despite strong revenues in the past 3 years.
- The company has a large customer base and offers **products from many categories through various channels**: physical stores, website, and catalogs.
- To address this, iFood is focusing on **strategic marketing initiatives** to improve overall effectiveness.
- A key approach is using data analytics to enhance **campaign targeting**, maximize **customer value**, and boost **conversion rates**.

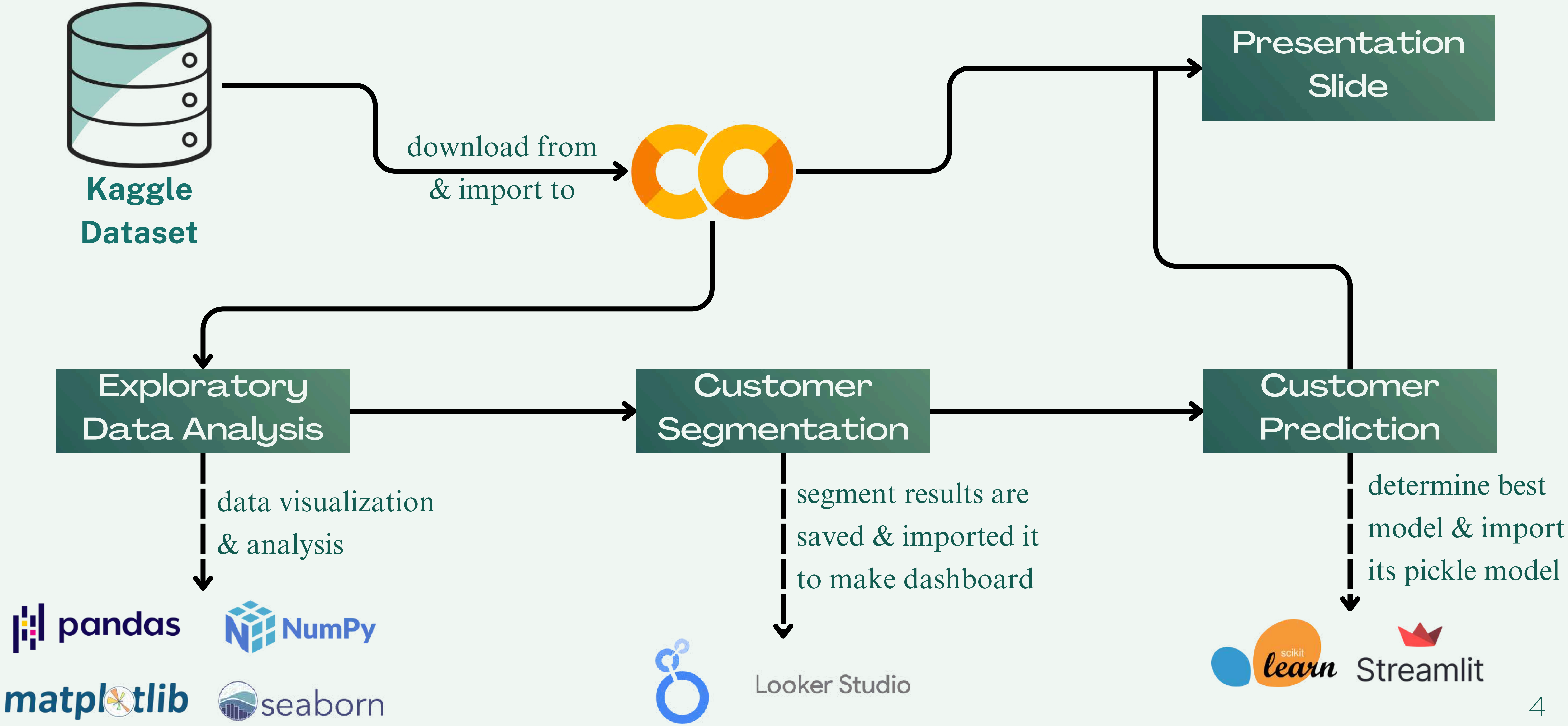
Goals & Objective

1. Segment customers into distinct groups based on demographics, behavior, and preferences to allow for targeted and personalized marketing strategies.
2. Build a classification model that predicts the likelihood of a customer responding positively to future marketing campaigns.
3. Use insights from segmentation and predictions to optimize the selection of target customers for upcoming campaigns.
4. Enable the marketing team to make strategic decisions based on analytical results to improve ROI on campaigns.

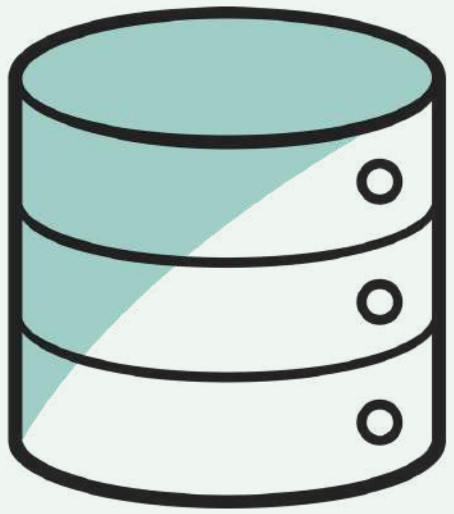
Actions & Metrics

1. Data Exploration & Preprocessing: Clean, transform, and visualize data to uncover customer behavior and campaign response patterns.
    - Metric: Completeness and clarity of preprocessed data.
2. Model Development & Evaluation: Perform customer segmentation and build predictive models (e.g., Logistic Regression, Random Forest) to forecast campaign response.
    - Metric: Silhouette Score & Visualizer for clustering; Accuracy, Precision, Recall, and F1-score for classification.
3. Business Recommendations: Deliver data-driven strategies for customer targeting and campaign optimization.
    - Metric: Relevance and potential business impact of insights.

PROJECT WORKFLOW



DATA INTERPRETATION



iFood Dataset
(ifood_df.csv)

39 *columns*

consist of all numerical values

column ...	column ...	column 'Response'
row ...	row ...	row ...
row ...	row ...	row ...

2.205 *rows*

RESPONSE

Unique values:
0 : Customer didn't accept the offer in the last campaign
1: Customer accepted the offer in the last campaign

Customer Profiles

- Age
- Marital
- Education
- Income
- Kidhome
- Teenhome
- Customer_Days
- Recency
- Complain

Product Preferences

- MntWines
- MntFruits
- MntMeatProducts
- MntFishProducts
- MntSweetProducts
- MntGoldProds
- MntTotal

Campaigns

- NumDealsPurchases
- AcceptedCmp1
- AcceptedCmp2
- AcceptedCmp3
- AcceptedCmp4
- AcceptedCmp5
- Response

Channel Performance

- NumWebPurchases
- NumCatalogPurchases
- NumStorePurchases
- NumWebVisitsMonth

EXPLORATORY DATA ANALYSIS

Import Library, Load Dataset

- Check Data Type and Data Shape

Data Preparation & Cleaning

- Dataset Summary Statistics
- Handle Unnecessary Features by drop some column
- Check missing value & drop duplicated data
- Detecting outliers & data distribution using boxplot
- Preparing data by doing some transformation on some features

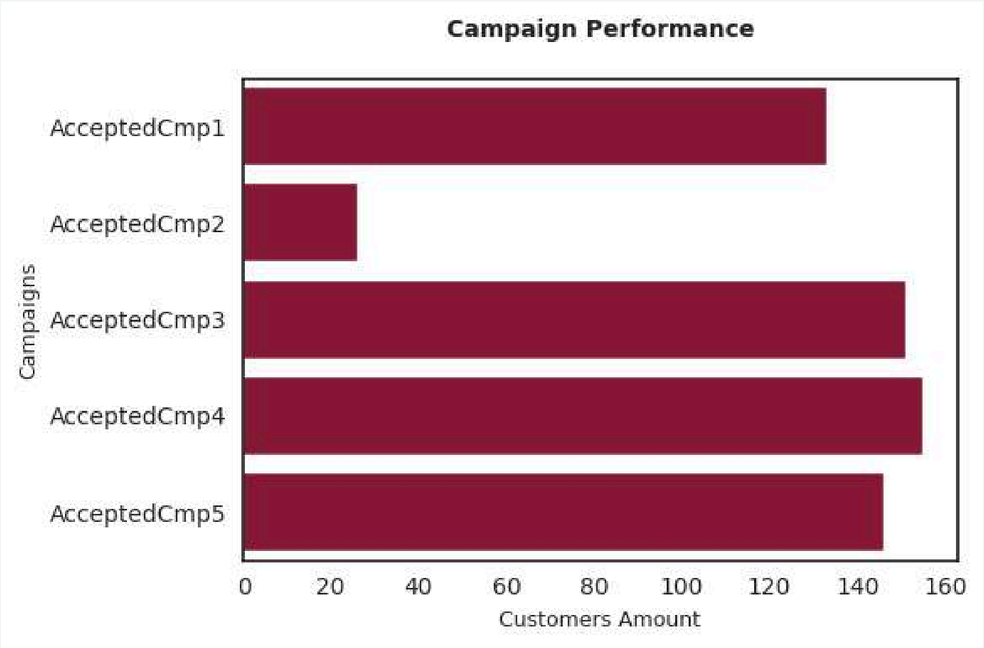
Exploratory Data Analysis

- Univariate Visualization
- Bivariate Visualization
- Check correlation matrix using heatmap to gain some insights

EXPLORATORY DATA ANALYSIS

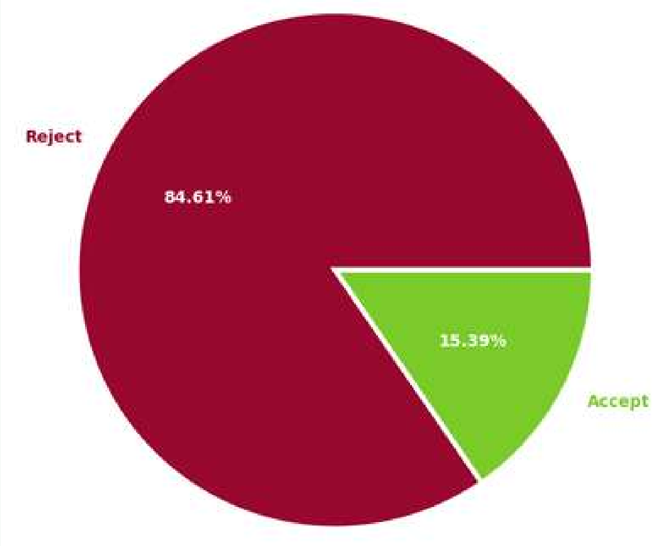
	Age	Customer_Days	Income	Dependents	Kidhome	Teenhome	Recency
Average	51.1	2511.6	51687.3	1.0	0.4	0.5	48.9

The average customer profile is 51 years old, has been a customer for almost 7 years, has an income of around $52.000 per year, has 1 dependent, and made a purchase from company in the last 49 days.

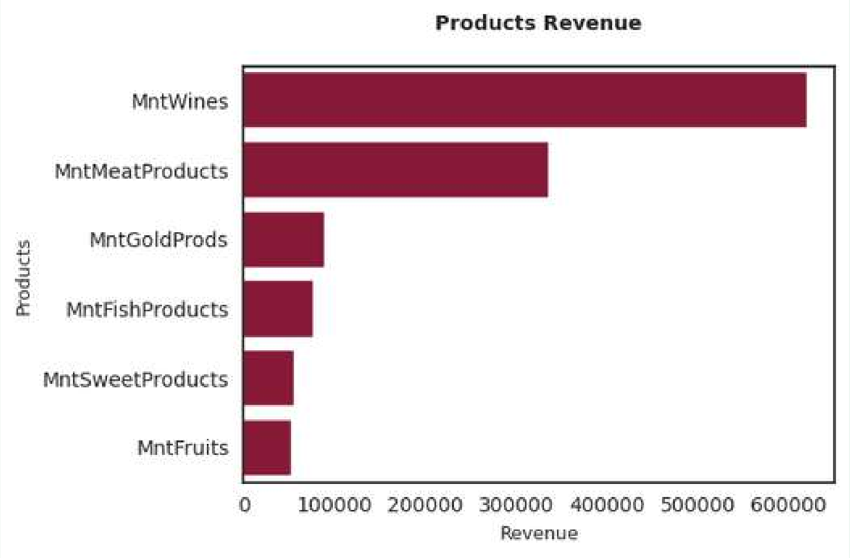


- **Campaign 4** stood out as the **most successful**, with 155 customers in the sample accepting the offer.
- In contrast, **Campaign 2** had the **lowest response**, with only 26 customers accepting the offer.

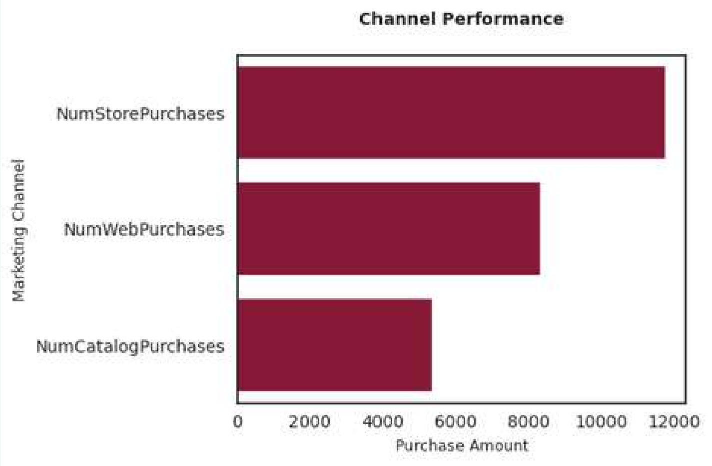
Overall Campaign Performance



Around **less than 16%** of the customer sample **accepted the offer** in all previous campaigns before the new campaign launching

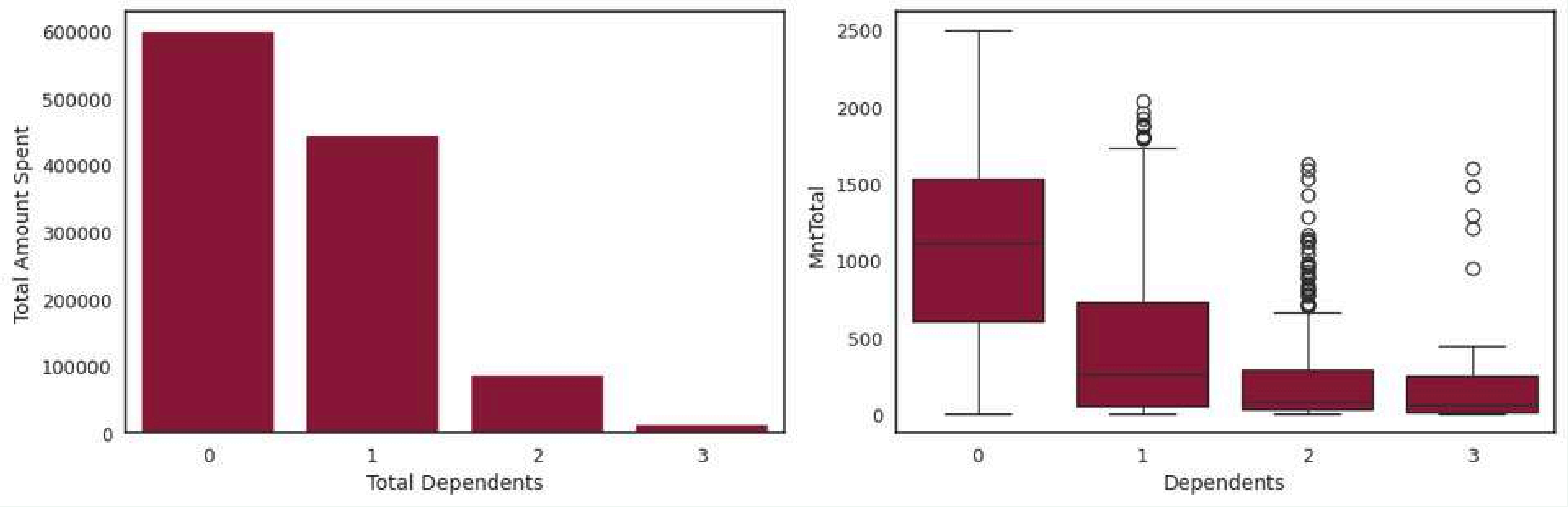


Wine is the **top revenue-generating** product, accounting for around **51%** of the company's total earnings.

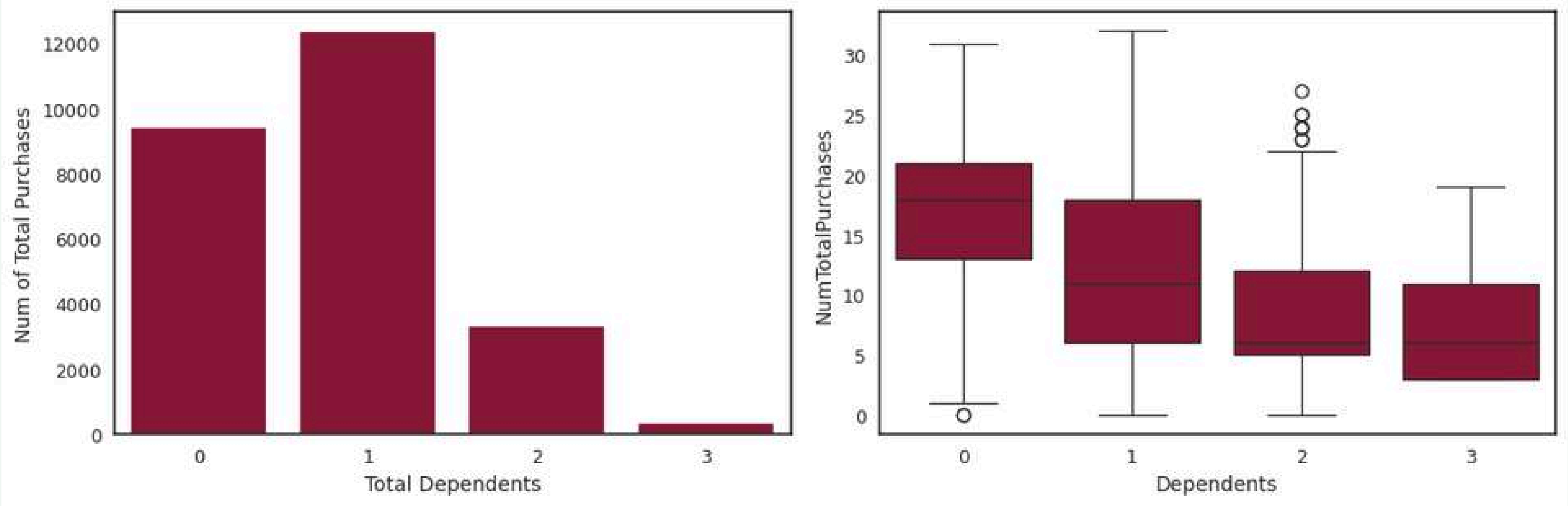


The **store channel** accounts for the **highest** number of purchases, contributing **46%** of the total orders.

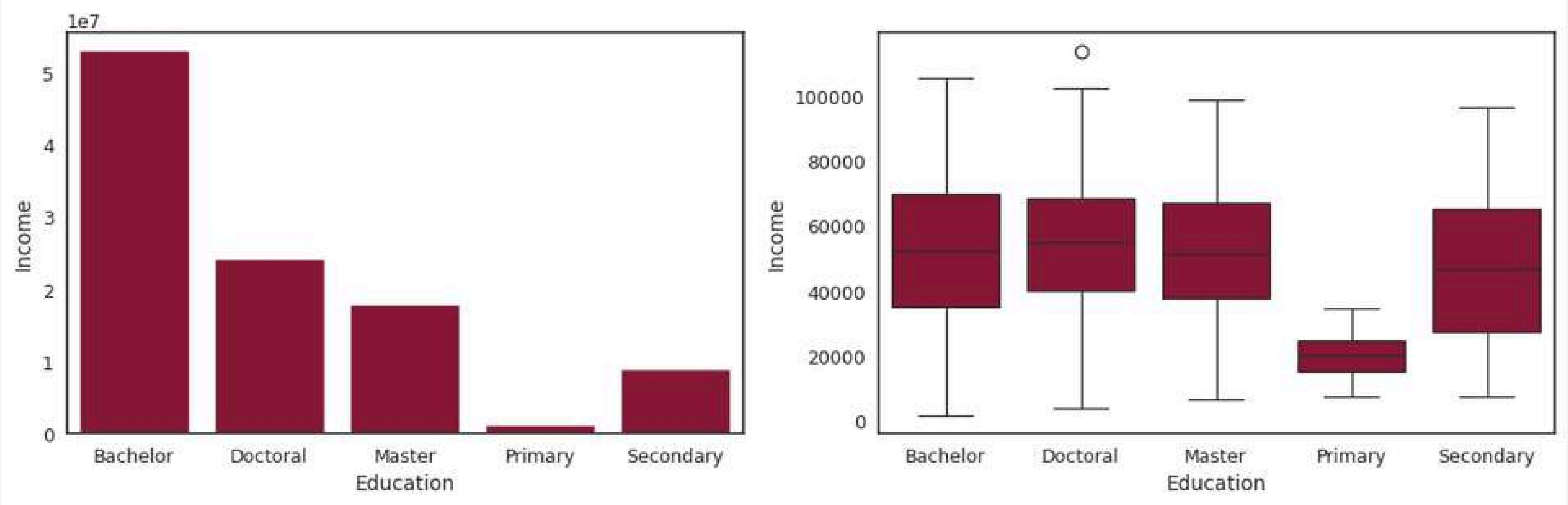
EXPLORATORY DATA ANALYSIS



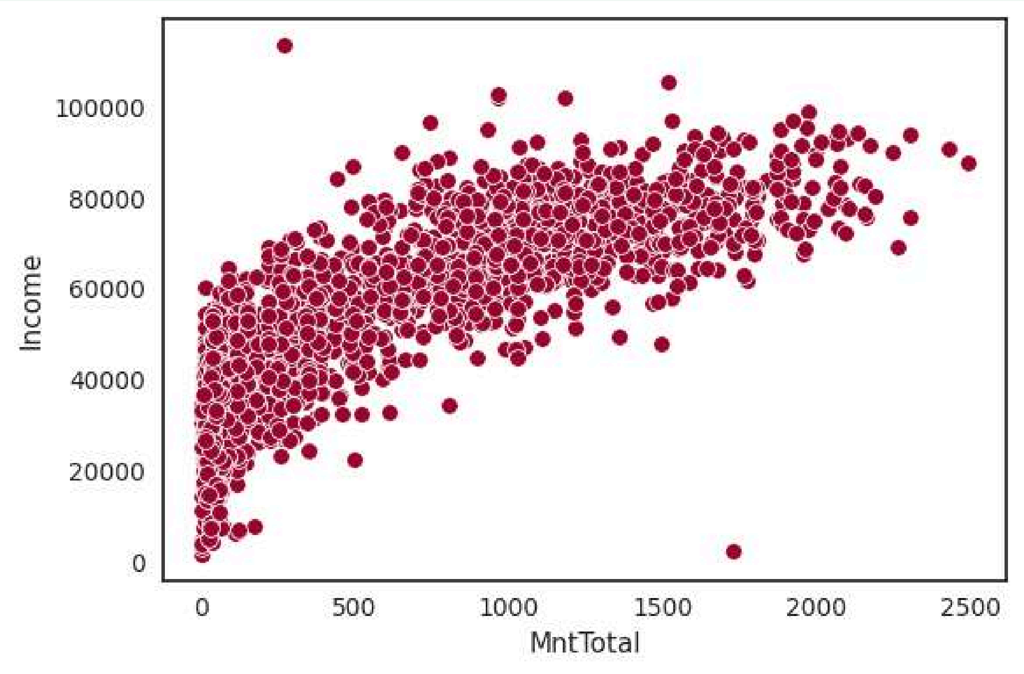
- Customers **without dependents** tend to **spend more**, and spending **gradually declines** as the number of dependents **increases**.
- Average amount of money spent by customers without dependents is about 1056 dollars, followed by 432 dollars with 1 dependent. There's a little difference between customers who have 2 and 3 dependents.



- Customers with **one dependent** contribute the **highest total purchases**, followed by those with no dependents, with spending gradually declining as the number of dependents rises.
- In terms of average purchases, customers without dependents lead with an average of 16 to 17 purchases.



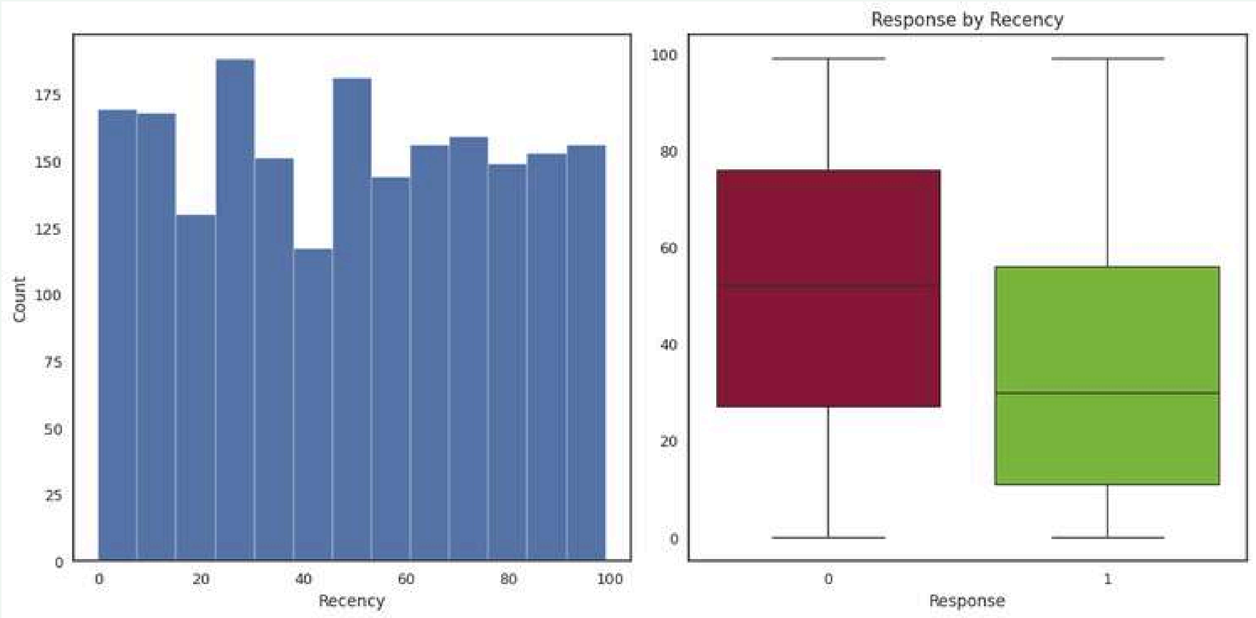
- Customers who hold a bachelor degree have the highest sum of income. But, have to check the average to avoid bias.
- The customers who **hold the doctoral degree** have the **highest average income** which indicate that the high someone's degree is, the highest income will be get.



There seems to be a **positive relationship** between a customer's **income and their spending**.

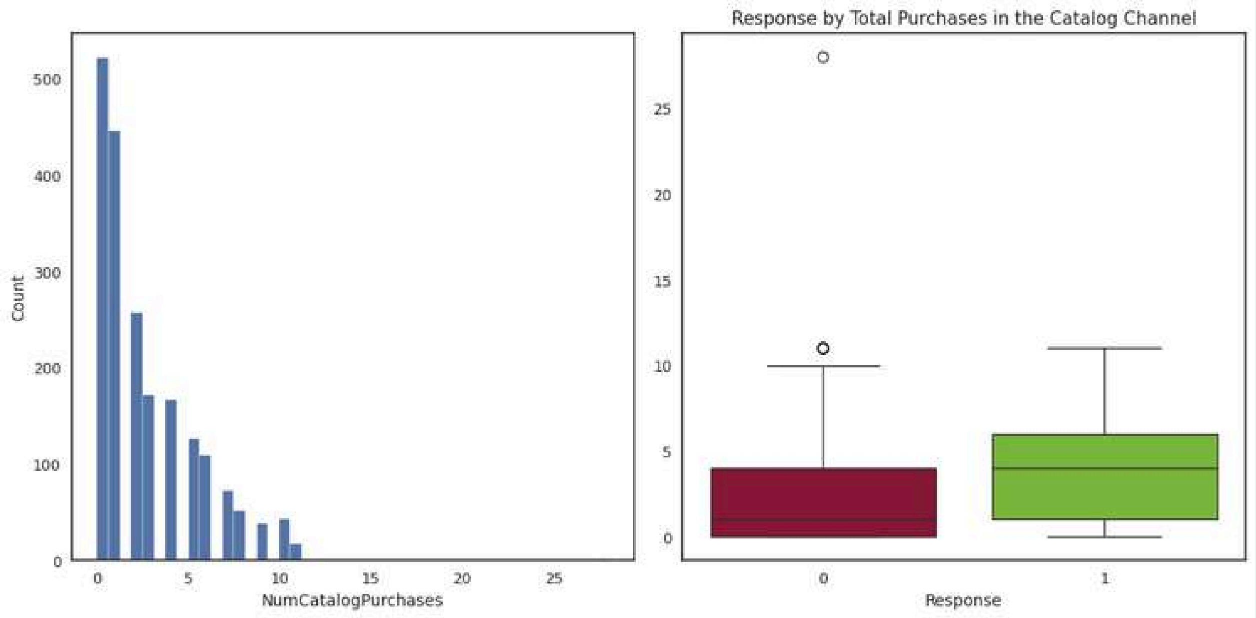
EXPLORATORY DATA ANALYSIS

Relationship Between **Response and Recency**



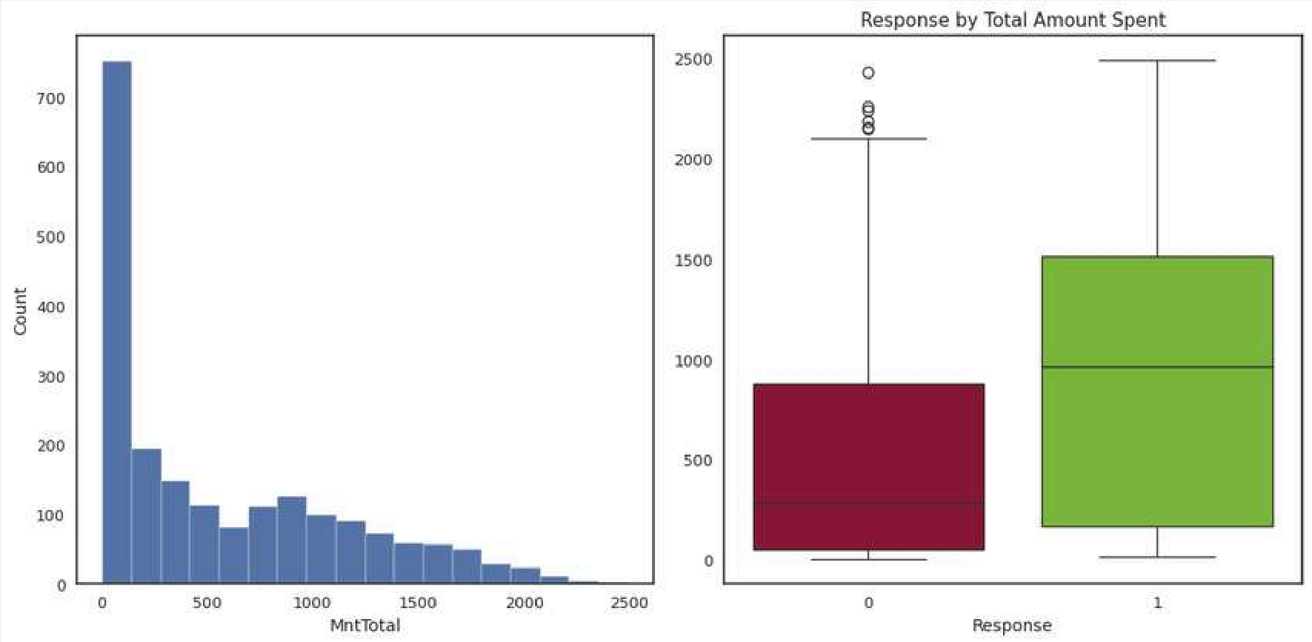
On average, customers who **accepted a campaign offer** made their **last purchase 35 days ago**. Customers with more **recent purchase activity** tend to be more **receptive** to campaign offers.

Relationship Between **Response and Total Purchases in Catalog Channel**



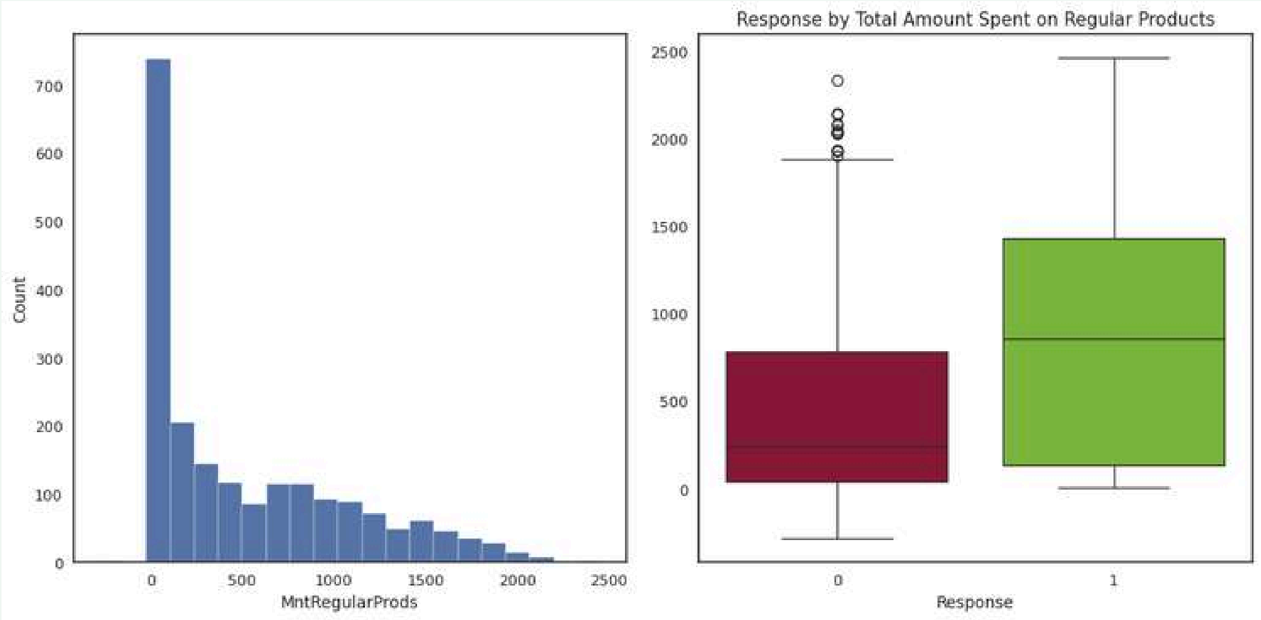
On average, customers who **accepted** a campaign offer made **4 purchases** through the **catalog channel**. Those who engage with the catalog channel are more **inclined to respond positively** to campaign offers.

Relationship Between **Response and Total Amount Spent**



Customers who **accepted the offer** spent an average total of 917 dollars. **Higher spending** customers tend to be **more responsive** to campaign offers.

Relationship Between **Response and Total Amount Spent on Regular Products**



On average, customers who **accepted the offer** spent 854 dollars on **regular products**. Customers with **higher spending** on regular products are more likely to **respond positively** to campaign offers.

CUSTOMER SEGMENTATION

Exploratory Data Analysis

- Univariate Visualization
- Bivariate Visualization
- Check correlation matrix using heatmap to gain some insights

Customer Segmentation

- Drop columns that have been used for another transformation features
- Encoding for categorical features(Education)
- Scaling process by using RobustScaler
- Recency, Frequency, and Monetary(RFM) analysis
- Customers Clustering using K-Means Algorithm

Import Segmentation Data

- Data Segments for Dashboard's visualization

CUSTOMER SEGMENTATION

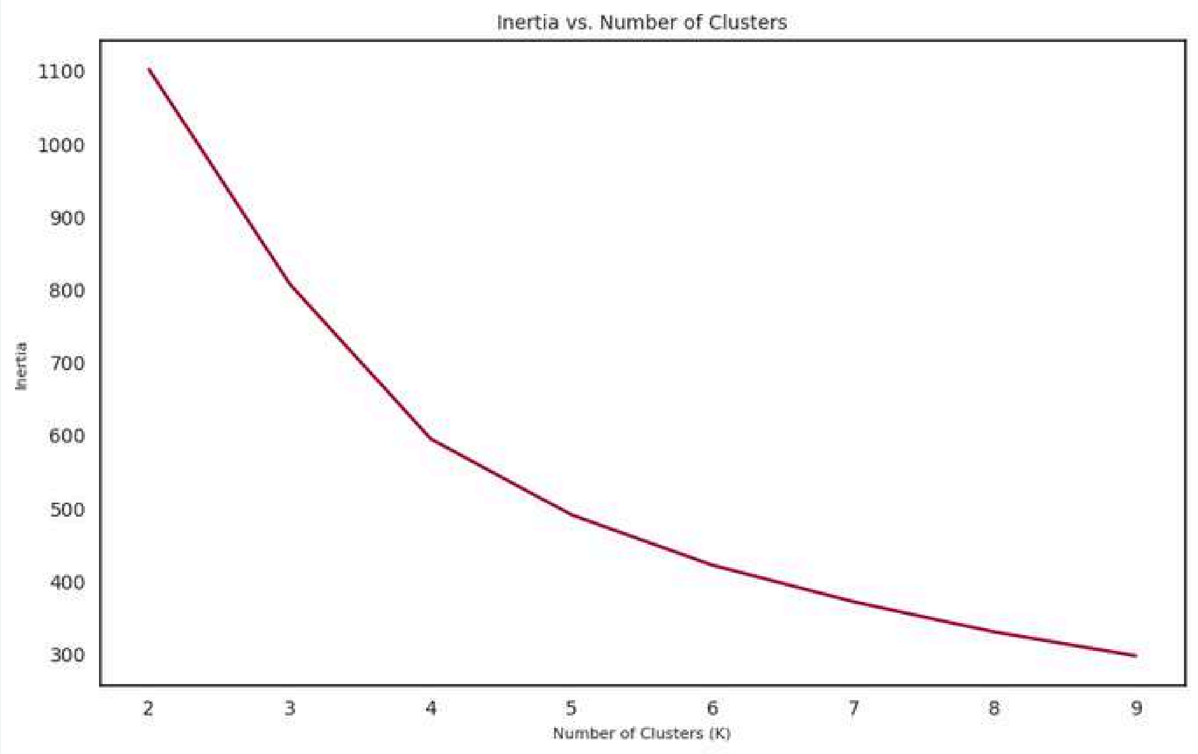
RFM analysis is a technique used to **segment customers** based on their **purchasing behavior**. It evaluates three key dimensions derived from past transactions:

R	Measures how recently a customer made a purchase	represented by →	**Recency**
F	Captures how often a customer makes purchases	represented by →	**NumTotalPurchases**
M	Reflects how much a customer spends	represented by →	**MntTotal**

CUSTOMER SEGMENTATION

Approach for Determine Number of Clusters(K-Means)

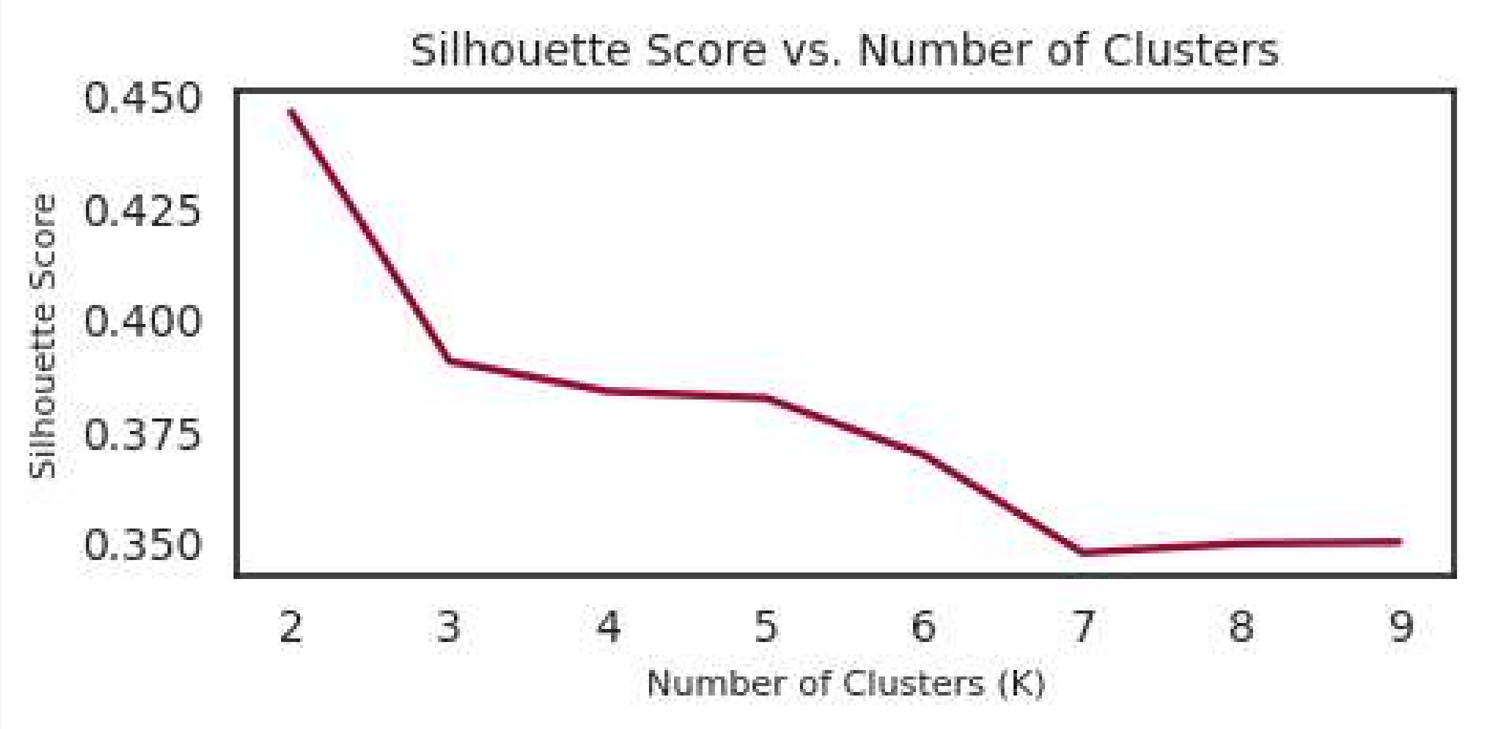
01



Elbow Method:

The elbow method does not show a **distinct inflection point**, making it **difficult** to determine the **optimal value** of K.

02



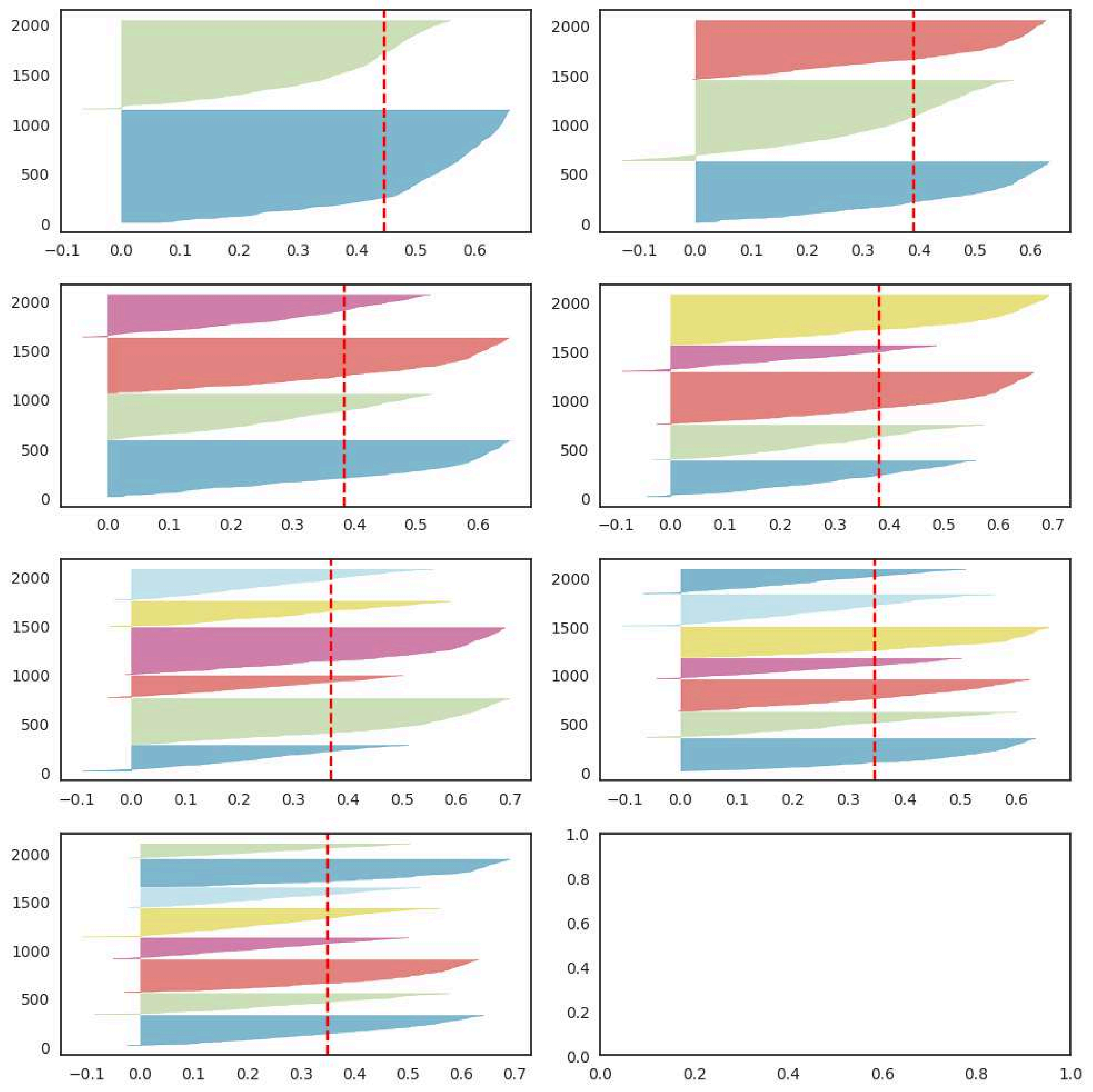
Silhouette Score Method:

Although the silhouette score reaches its peak value of 0.45 at K=2, this alone is not enough to confidently determine the optimal number of clusters.

CUSTOMER SEGMENTATION

Approach for Determine Number of Clusters(K-Means)

03



Silhouette Visualizer:

- Each cluster's silhouette score should **exceed the overall average** (indicated by the red dashed line on the x-axis). In this case, all clusters **meet** that requirement.
- The cluster sizes should be **relatively balanced**, without significant disparities. For instance, with K=2, the green cluster is nearly twice as wide as the blue one. Similarly, clusterings with K=[5, 6, 7, 8] are excluded due to uneven cluster sizes.
- While both K=3 and K=4 result in fairly uniform clusters, **K=4** achieves slightly **better balance**.
- Therefore, based on the silhouette plot criteria, **K=4 is selected as the optimal number of clusters.**

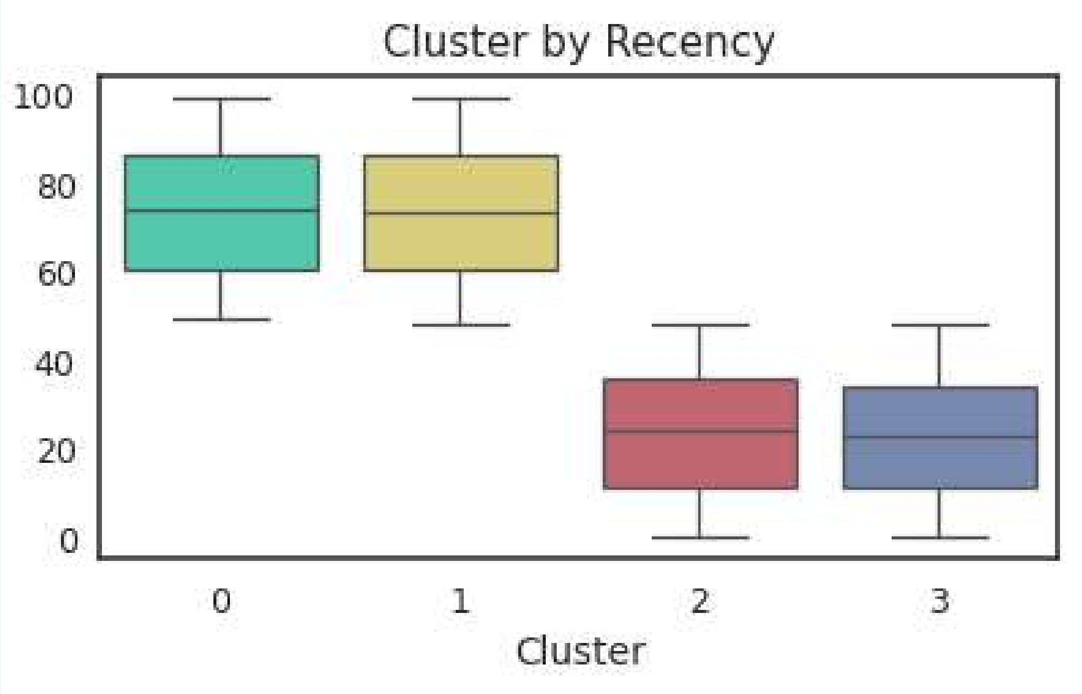
CUSTOMER SEGMENTATION

Customer’s Segments		Description for Customer	Count	Percentage(%)
Cluster 3	Core Champions	Customers who purchased recently, buy frequently, and spend the most. These are iFood's most valuable customers.	426	21.08
Cluster 2	Newcomer Shoppers	Customers who made recent purchases but in low quantities and value. They may be new users still exploring iFood's offerings.	560	27.71
Cluster 1	High-Value At Risk	Customers who used to purchase frequently and spend a lot but haven't bought anything recently. These are high-value customers at risk of churning and should be re-engaged.	462	22.86
Cluster 0	Dormant Buyers	Customers who bought infrequently and spent little—and haven't purchased in a long time. Some of them may have already switched to competitors.	573	28.35

CUSTOMER SEGMENTATION

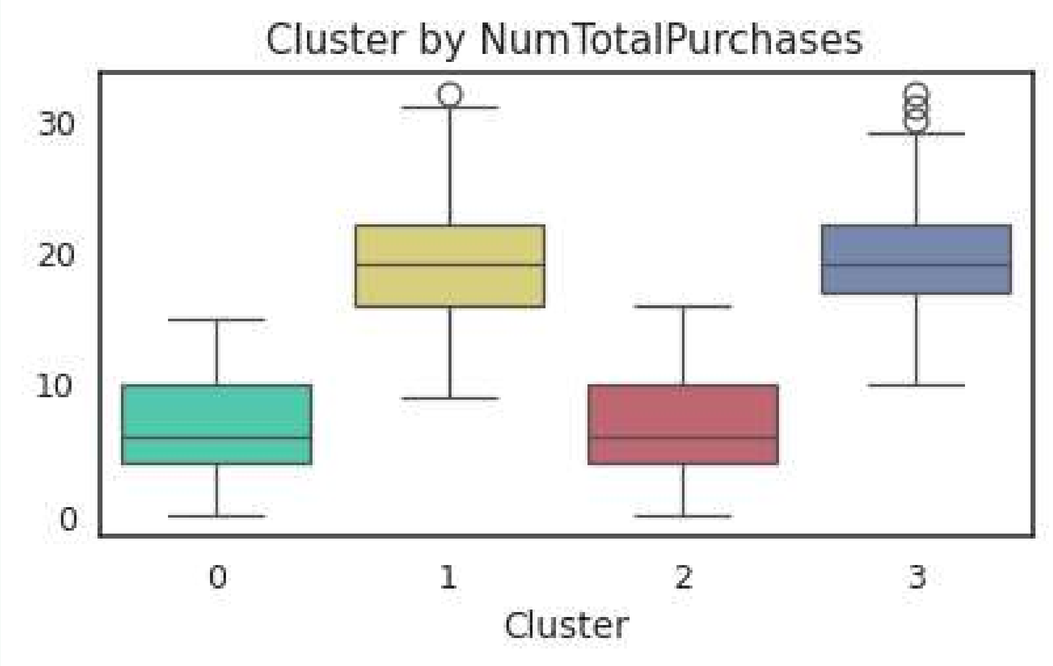
Clustering by RFM

R



- Cluster **Dormant Buyers** and **High-Value At Risk** consist of customers who **haven't made a purchase in a while**, with their last purchase occurring approximately 73 days ago.
- Meanwhile, Cluster **Newcomer Shoppers** and **Core Champions** include **more recent buyers**, having made their latest purchase around 22 days ago.

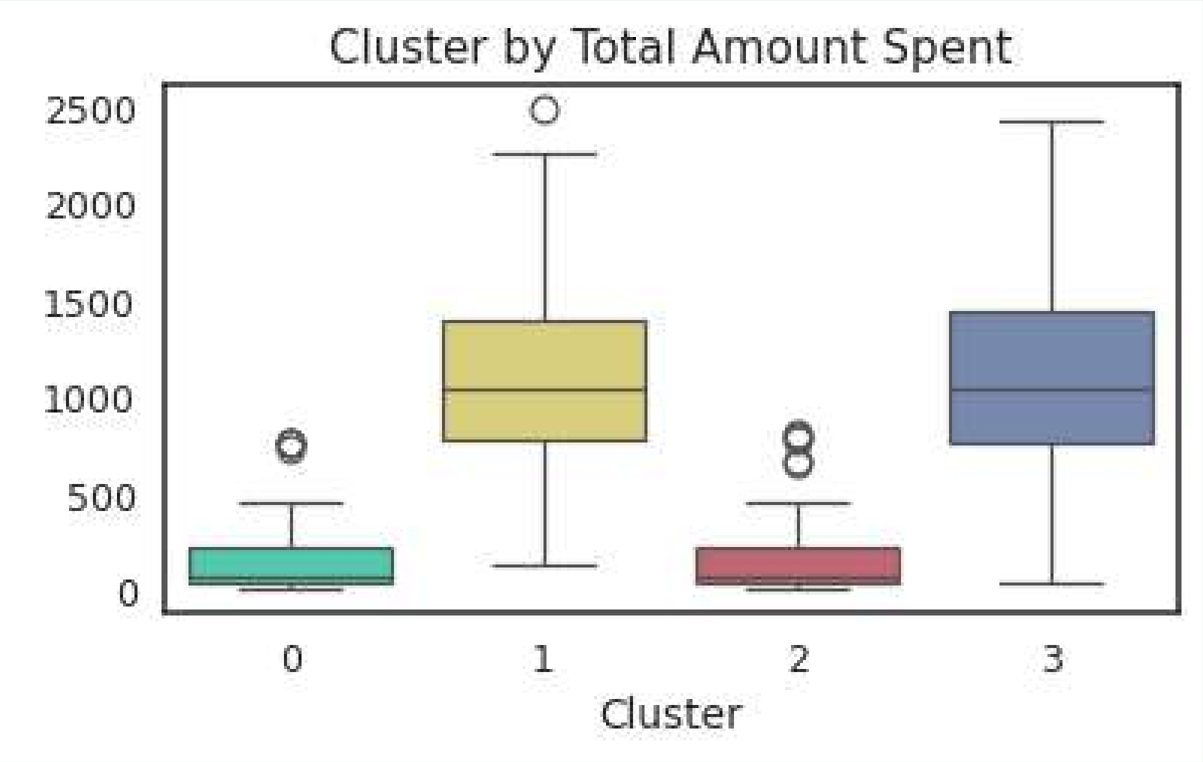
F



- Cluster **High-Value At Risk** and **Core Champions** represent **frequent buyers**, averaging 19 purchases.
- While, Cluster **Dormant Buyers** and **Newcomer Shoppers** show **lower purchasing frequency**, averaging 7 purchases.

CUSTOMER SEGMENTATION

M

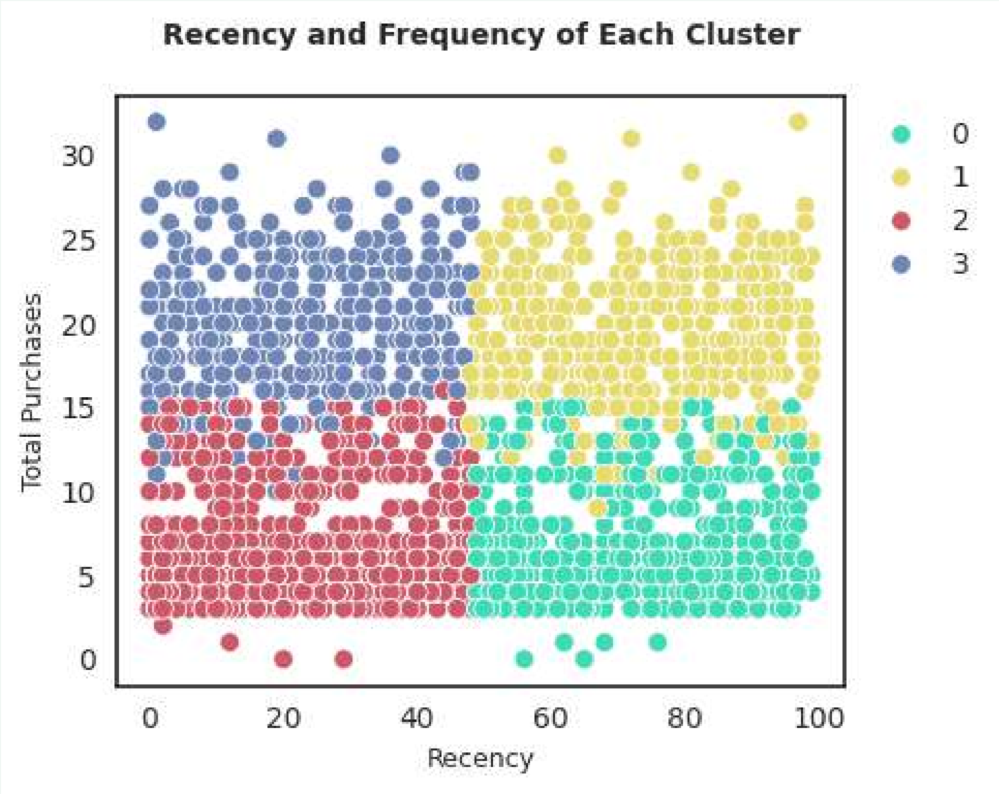


- Cluster **High-Value At Risk** and **Core Champions** consist of **high-spending customers** with an average total expenditure of approximately 1,110.
- While, Cluster **Dormant Buyers** and **Newcomer Shoppers** include **lower spenders** who spend around 130 on average.

CUSTOMER SEGMENTATION

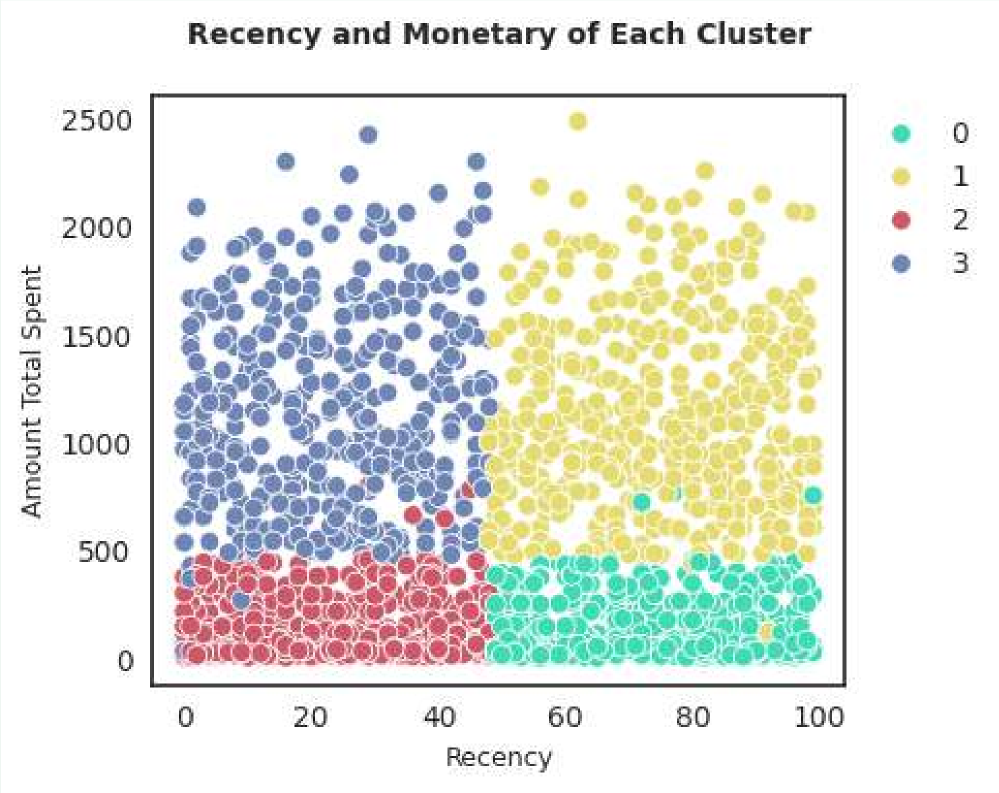
Relationship between RFM

R & F



- Cluster **Dormant Buyers** comprises customers who **haven't made a purchase for a long time**, with their last purchase between 49 and 99 days ago, and they have an **average total purchase of fewer than 15**.
- Cluster **High-Value At Risk** consists of customers who **haven't made a purchase in a while**, with their last purchase occurring between 48 and 99 days ago, and they have an **average of over 15 total purchases**.
- Cluster **Newcomer Shoppers** includes customers who **made a purchase within the last 0 to 48 days**, with an **average total purchase of fewer than 15**.
- Cluster **Core Champions** consists of customers who **made a purchase within the last 0 to 48 days**, and they have an **average of over 15 total purchases**.

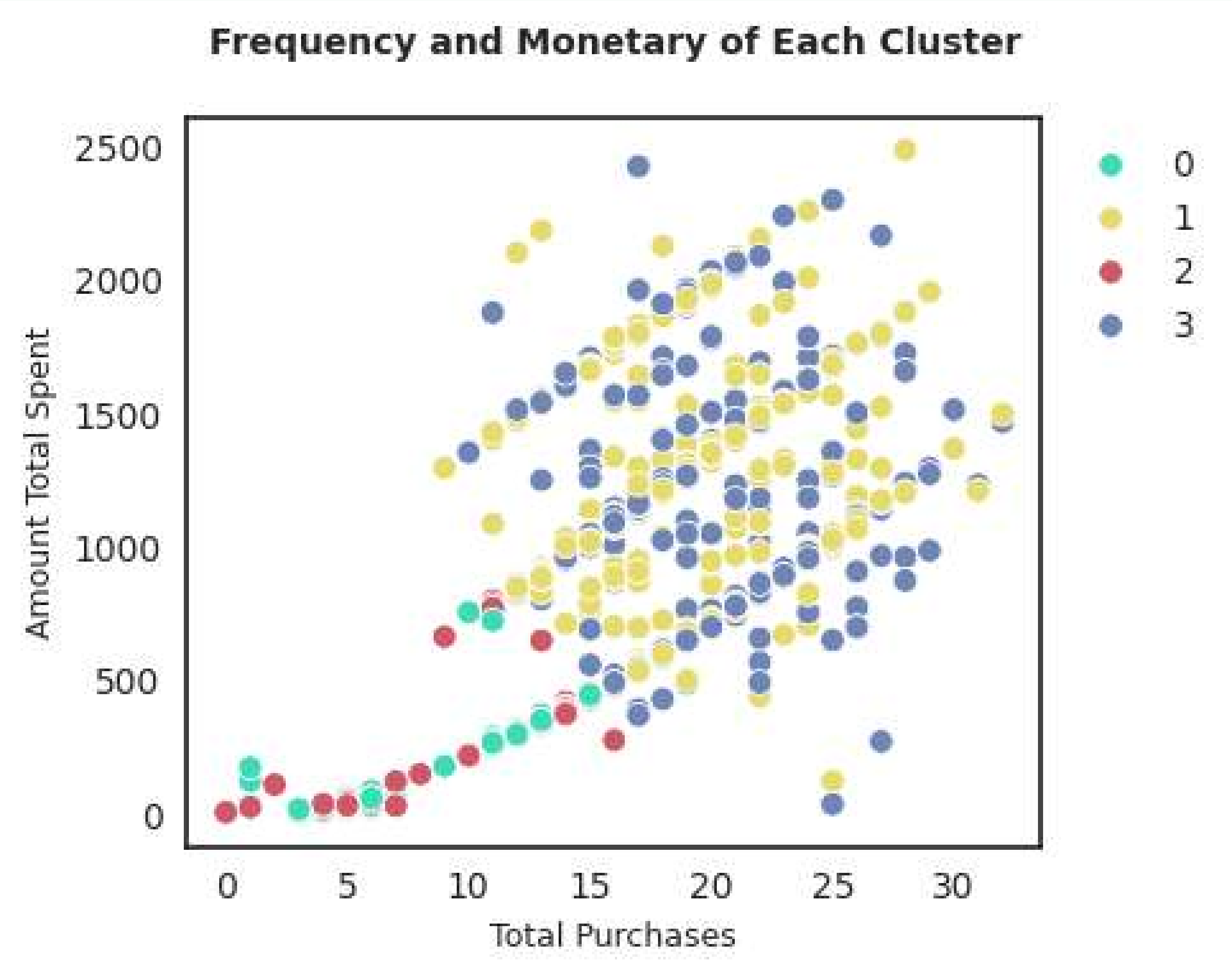
R & M



- Cluster **Dormant Buyers** comprises customers who haven't purchased in a long time (around 49-99 days), with an **average spending of less than 500**.
- Cluster **High-Value At Risk** consists of customers who haven't made a purchase in a long time – approximately 48 to 99 days ago – and have **spent more than 500 on average**.
- Cluster **Newcomer Shoppers** includes customers who recently made a purchase (within 0-48 days) and tend to **spend less than 500 on average**.
- Cluster **Core Champions** represents recent buyers (within 0-48 days) who **spend more than 500 on average**.

CUSTOMER SEGMENTATION

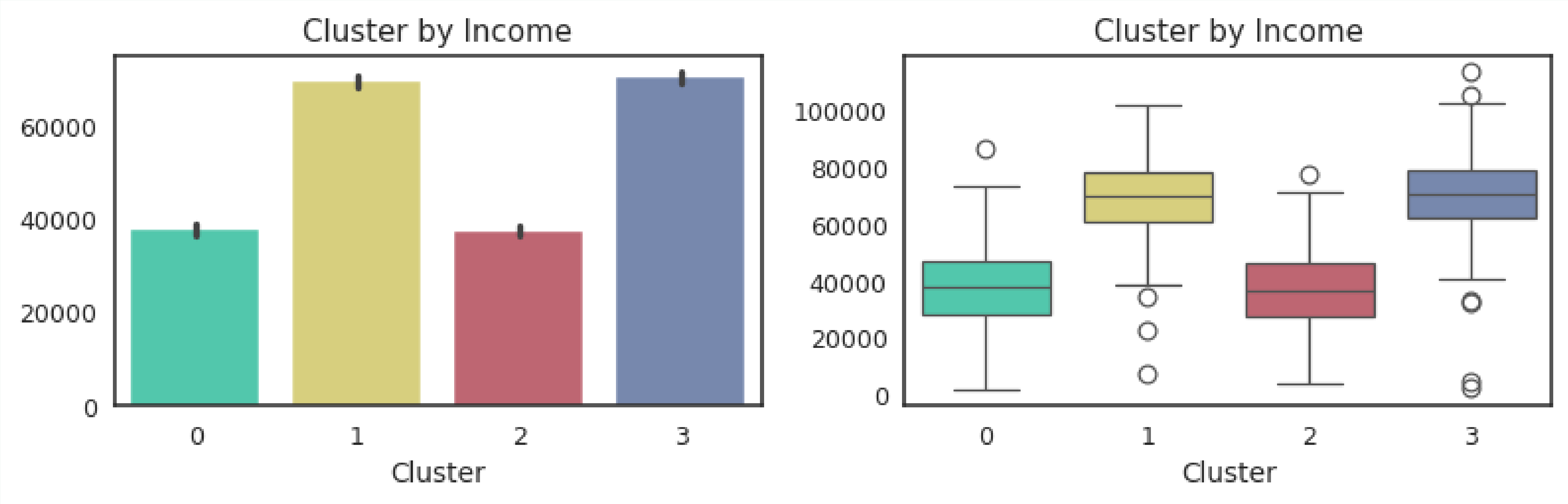
F & M



- Cluster **Dormant Buyers** and **Newcomer Shoppers** consist of customers with **fewer than 16 total purchases** and **total spending that does not exceed 800**.
- Cluster **High-Value At Risk** and **Core Champions** include customers with **more than 9 purchases** and **spending ranging from 30 to 3000**.

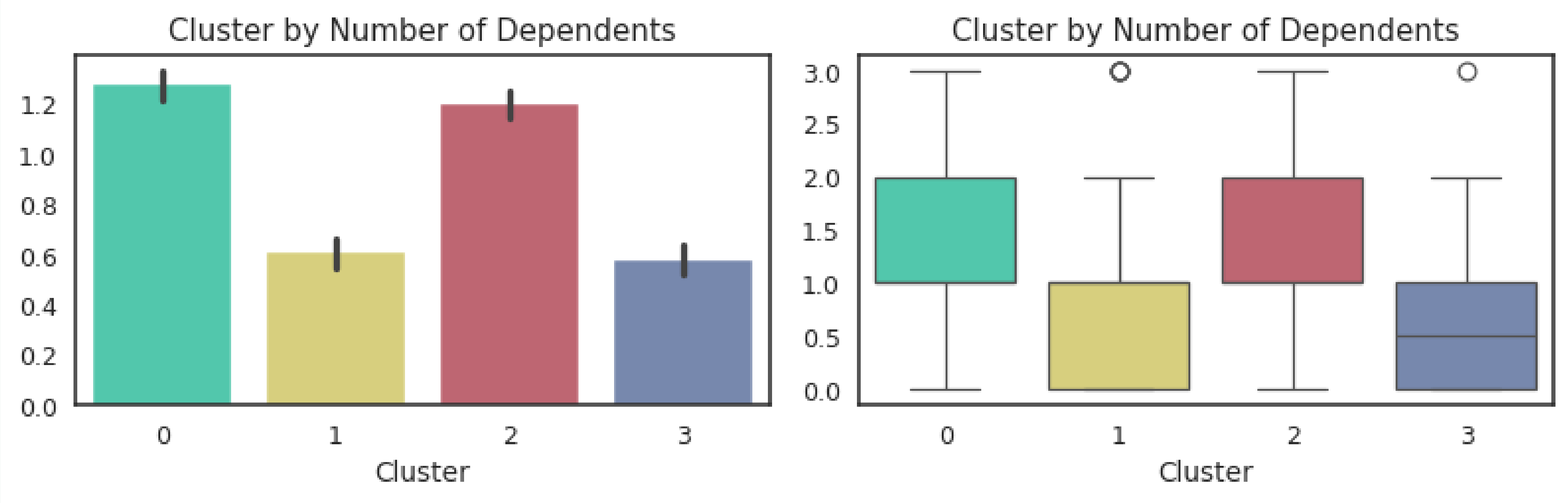
CUSTOMER SEGMENTATION

Clustering by Income



- Cluster **Dormant Buyers** and **Newcomer Shoppers** consist of **lower-income customers**, with an average income of approximately 37,000.
- Cluster **High-Value At Risk** and **Core Champions** represent customers with **higher incomes**, averaging around 70,000.

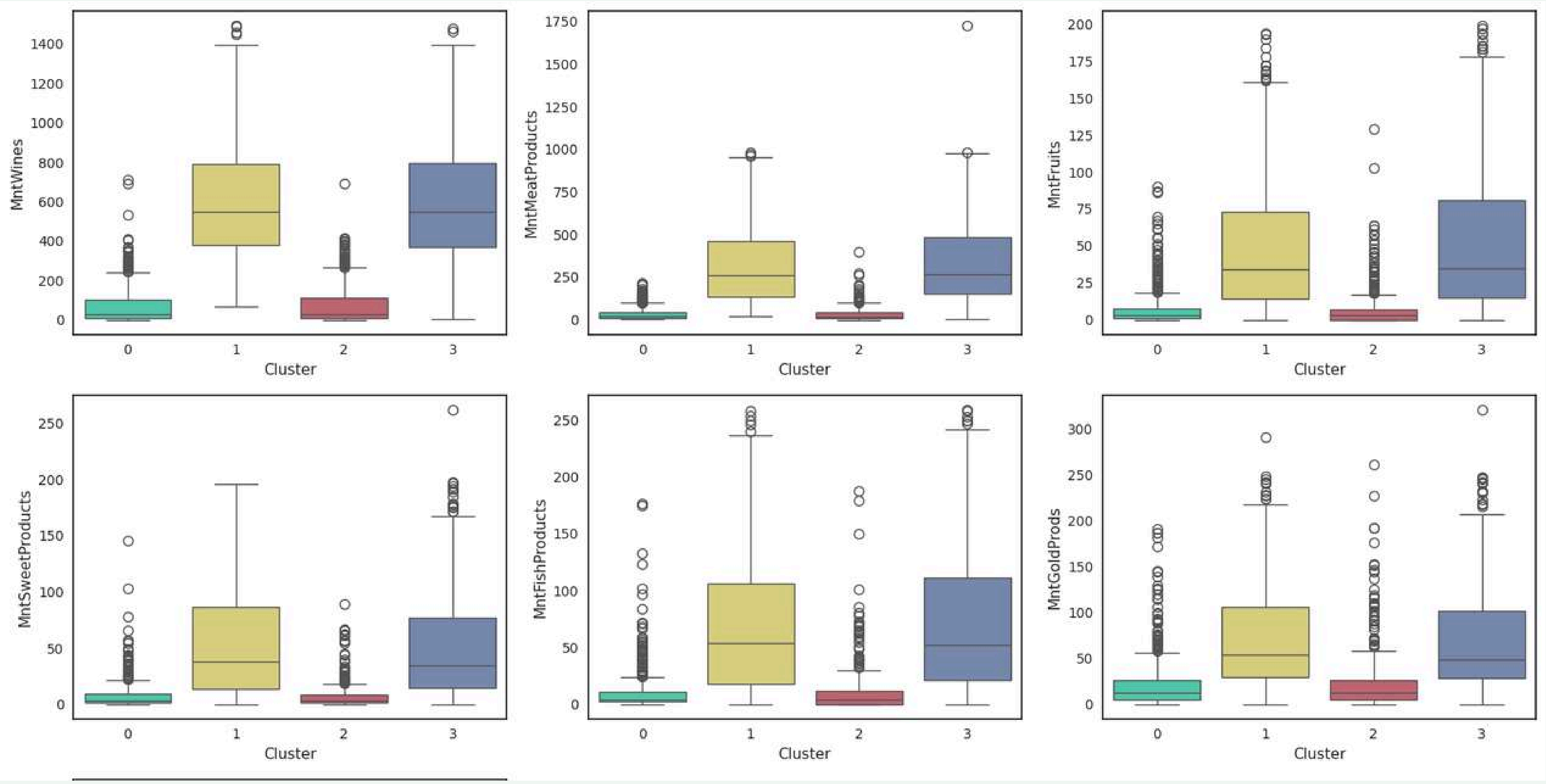
Clustering by Num of Dependents

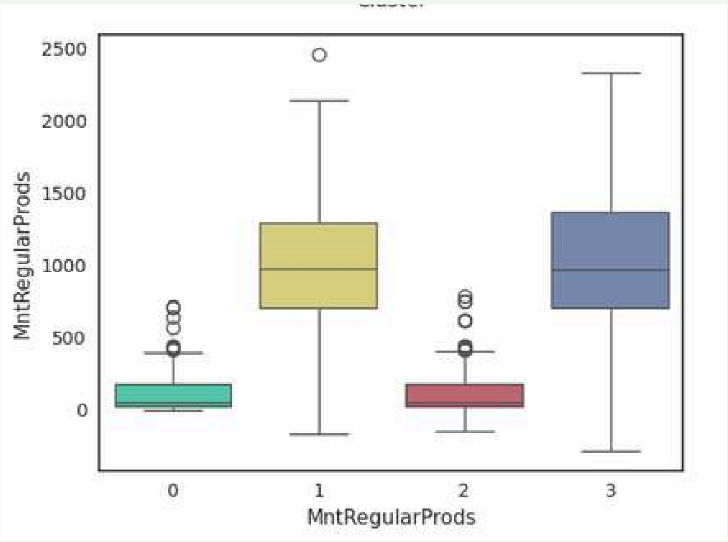


- Cluster **Dormant Buyers** and **Newcomer Shoppers** are customers who have an **average of 1 dependent**
- Cluster **High-Value At Risk** and **Core Champions** are customers who **do not have dependents**

CUSTOMER SEGMENTATION

Clustering by Product References

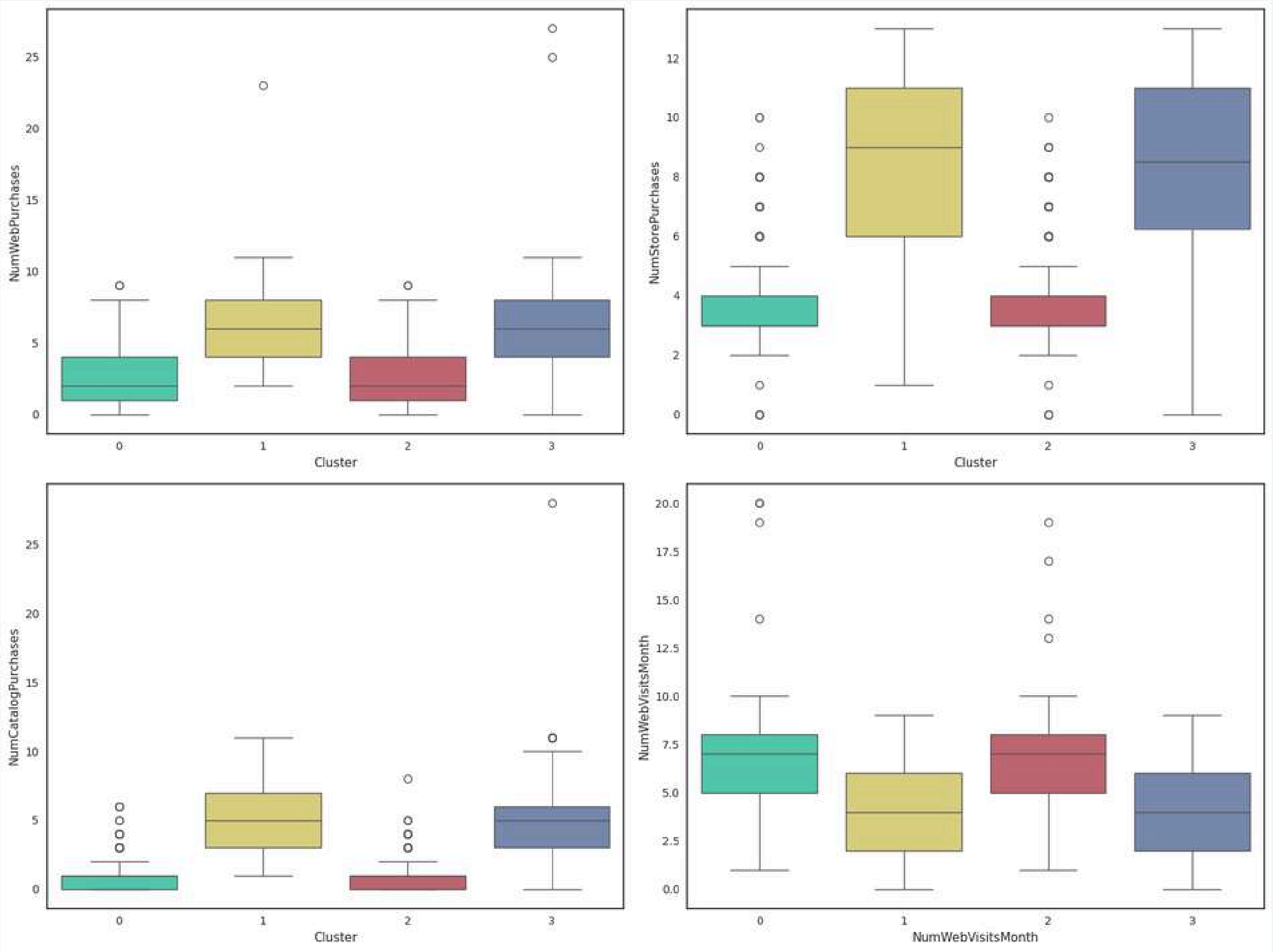




- Cluster **Dormant Buyers** and **Newcomer Shoppers** have **low purchasing frequencies** for each product category
- Cluster **High-Value At Risk** and **Core Champions** have a **high purchasing frequency** in all product categories

CUSTOMER SEGMENTATION

Clustering by Channel Performance



- Cluster **High-Value At Risk** and **Core Champions** show a **high frequency of purchases** across all channels, whereas Clusters **Dormant Buyers** and **Newcomer Shoppers** demonstrate **lower purchasing** activity on each channel.
- Cluster **Dormant Buyers** and **Newcomer Shoppers** make the **fewest purchases** overall but have the **highest number of web visits** compared to other clusters.

From the clustering analysis that has been done before, the following insights emerge:

- Customers with high purchase frequency and spending tend to visit the website less frequently. This could indicate that they are already familiar with the company's products and make quick purchasing decisions.
- Customers with lower purchase frequency and spending often visit the website more frequently. This behavior may suggest hesitation or indecision, possibly due to comparing product options, prices, or delivery fees between iFood and competitors.

BUSINESS INSIGHTS

01

The **average customer profiles** is 51 years old, almost 7 years of became a customer, having income around $52.000 per year, got 1 dependent, and made a purchase from company in the last 49 days. This kind of **average customer will by segmented in Cluster 2 as a Newcomer Shoppers**.

02

Engagement with the **latest campaign** is **relatively low**, with only around 16% of customers responding. However, this contrasts sharply with **Campaign 4**, which performed **significantly better**. This suggests that some campaigns are far more effective than others, possibly due to timing, targeting, or offer type.

03

Recency and spending behavior are **strong indicators** of **campaign response**. Customers who **recently made a purchase** and those who typically **spend more** are more likely to **accept new offers**. This pattern suggests campaigns should be closely aligned with recent activity by the customers.

BUSINESS INSIGHTS

04

There's a clear **connection between income, education, and spending**. Customers with **higher education** (especially doctorate holders) tend to **earn more** and also **spend more**. These segments present a strong opportunity for **premium positioning**.

05

Spending drops sharply for customers with dependents, particularly those with one. Interestingly, having **more than one dependent doesn't reduce spending much further**, suggesting that it's not household size but possibly lifestyle stage that influences purchase behavior.

06

Customer segmentation reveals **four distinct behavioral groups**. **Cluster 3** customers ("Core Champions") are the **most active and valuable**. **Cluster 1**, while **high-spending**, **hasn't purchased in a while**—indicating **a group at risk of churn**. **Clusters 0 and 2** are **less active and spend less**, but **exhibit high web traffic**, possibly showing **hesitation** or **price sensitivity** and took a **long time comparison** with the **competitors** product.

BUSINESS INSIGHTS

07

A significant portion of **high-income customers** (Clusters 1 and 3) also fall into the **category of high spenders and frequent buyers**, despite **not always** being the **most recent** purchasers. This shows that **income level** is a strong indicator of customer value over time, even when short-term engagement fluctuates.

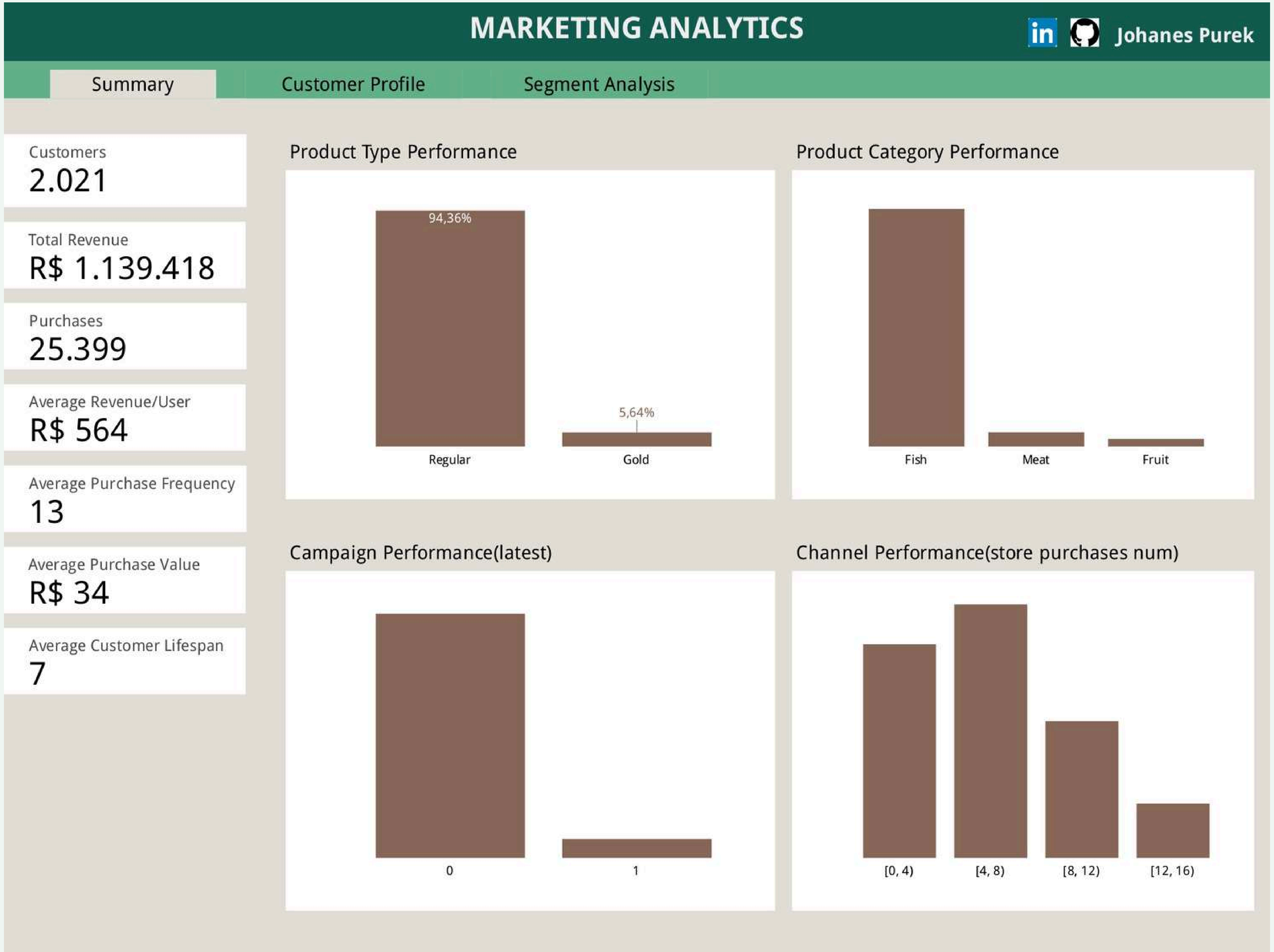
08

Customers with **higher spending on regular products** demonstrate a **stronger likelihood of responding positively** to promotional campaigns. These customers are primarily **concentrated in Cluster 1 and Cluster 3**, which are characterized by **recent purchasing** activity and **greater overall expenditure**. This pattern suggests that engaged, high-value customers are **more inclined to accept campaign offers** when they **align** with **their regular purchasing habits**.

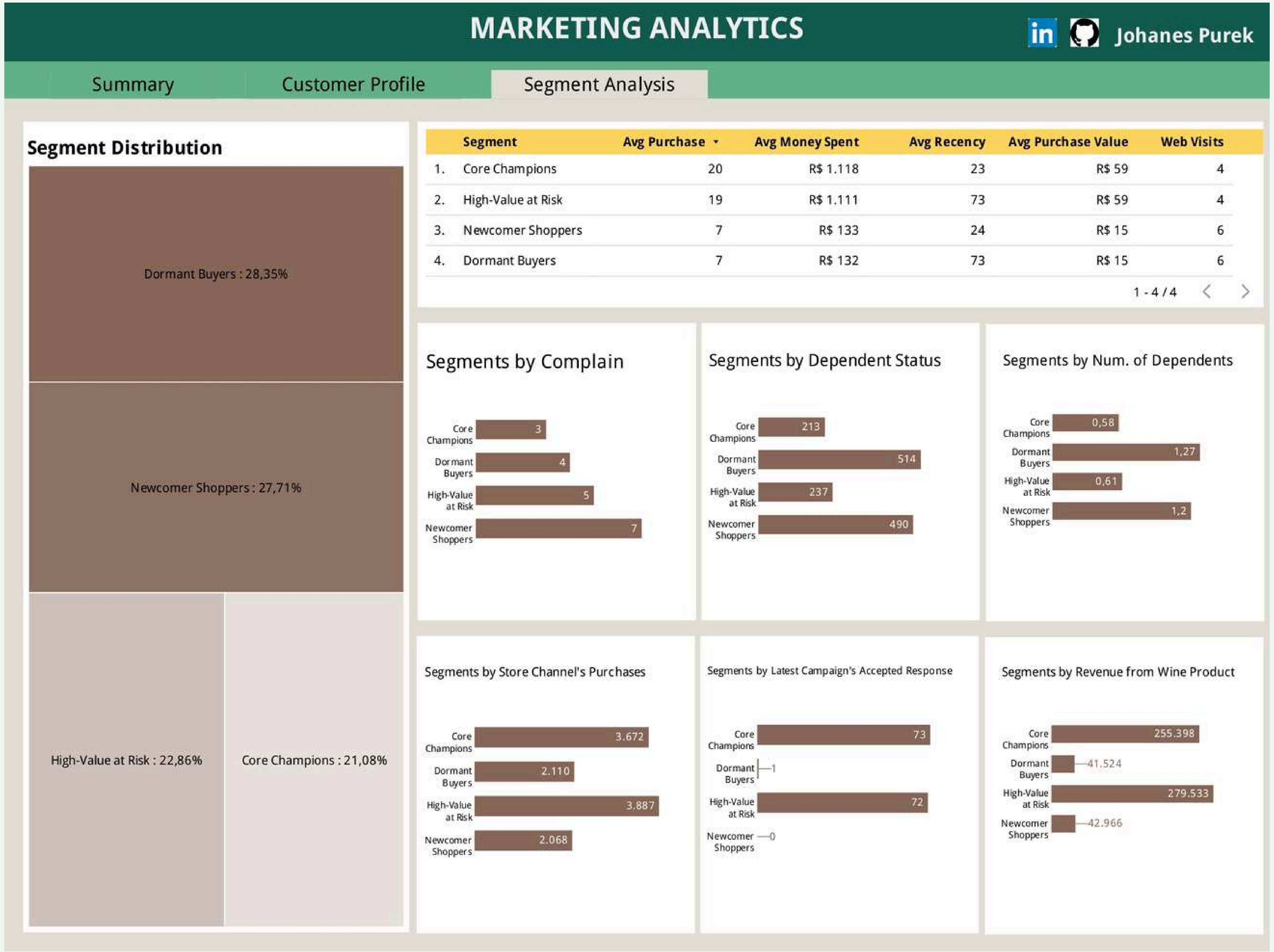
09

Website visits are high among lower-spending segments, which may point to **decision fatigue, unclear value propositions, or comparisons with competitors**. On the other hand, **high-spending customers visit less**, likely because they **already know** what they want.

CUSTOMER DASHBOARD



CUSTOMER DASHBOARD



For more: https://bit.ly/dashboard-marketing-jpbp

CUSTOMER PREDICTION

Import Segmentation Data

Data Segments for Dashboard's visualization

Customer Prediction

- **Handle Imbalanced Target by using minority class to be oversampled**
- **Feature Selection by using features that have a correlation value > 0.1**
- **Model Building(Machine Learning Model with its hyperparameter)**
- **Model Evaluation & Selection**
- **Import Model for Streamlit Deployment**

Business Recommendation

List of Recommendation for iFood

CUSTOMER PREDICTION

There are 27 column that will be used as features(X) based on the correlation value with Response as target(Y) which value is greater than 0.1

Customer Profiles

- LiveWith
- Education
- Income
- Kidhome
- Teenhome
- Dependents
- HasDependent
- Customer_Days
- Recency

Product Preferences

- MntWines
- MntFruits
- MntMeatProducts
- MntFishProducts
- MntSweetProducts
- MntRegularProds
- MntGoldProds
- MntTotal

Campaigns

- AcceptedCmp1
- AcceptedCmp3
- AcceptedCmp5
- AcceptedCmpOverall
- IsRetented

Channel Performance

- NumWebPurchases
- NumCatalogPurchases
- NumStorePurchases
- NumWebVisitsMonth
- NumTotalPurchases

CUSTOMER PREDICTION

5 Machine Learning's algorithm will be used in this case and will be train with its hyperparameter to get the best model's estimator result. **XGBoost best hyperparameters: {'learning_rate': 0.2, 'max_depth': 7, 'n_estimators': 100}**

ML Model	Accuracy Score	
	Training	Testing
XGBoost	**99.77%**	**96.02%**
Random Forest	**99.77%**	**95.44%**
K-Nearest Neighbor	**99.77%**	**91.7%**
Decision Tree	**88.5%**	**87.6%**
Logistic Regression	**87.91%**	**85.73%**

- All top 3 models show signs of potential overfitting which can be seen above. The XGBoost, Random Forest, and K-Nearest Neighbor all have identical very high training accuracy (99.77%), but show a notable drop in testing accuracy and especially KNN (91.70%), indicating overfitting and lower generalization.
- XGBoost is the best-performing and most balanced model. Despite the high training accuracy, XGBoost maintains the highest testing accuracy (96.02%), indicating strong generalization and making it the most reliable model for unseen data in this context.

CUSTOMER PREDICTION

Model's performance metrics overall

ML Model	Accuracy	Precision	Recall	F1-Score
XGBoost	**96.02%**	**96.33%**	**95.89%**	**96.11%**
Random Forest	**95.44%**	**96.29%**	**94.75%**	**95.51%**
K-Nearest Neighbor	**91.7%**	**87.07%**	**98.4%**	**92.39%**
Decision Tree	**87.6%**	**89.71%**	**85.62%**	**87.62%**
Logistic Regression	**85.73%**	**88.92%**	**82.42%**	**85.55%**

- XGBoost consistently outperforms all other models across 3 evaluation metrics. This model achieves the highest Accuracy (96.02%), Precision (96.33%), and F1-Score (96.11%), also pretty much high at Recall (95.89%) indicating it has both high correctness and balance between false positives and false negatives.
- K-Nearest Neighbors has high recall but at the cost of precision. Although KNN has the highest Recall (98.40%), its Precision is lowest (87.07%), indicating that it frequently misclassifies negatives as positives. This makes it less reliable in scenarios where false positives are costly.

CUSTOMER PREDICTION

For more: https://marketing-analytics-johanbernardd.streamlit.app/

BUSINESS RECOMMENDATION

01

Review and replicate the approach used in Campaign 4. Look at what made it successful like **messaging, timing, audience delivered** and use those insights to refine future campaigns. Underperforming campaigns like **Campaign 2** should either be **redesigned or dropped**.

02

Double down on targeting customers who've **purchased recently and spend more**. These customers show the highest responsiveness to campaigns and offer the best return on marketing efforts.

03

Investigate the **high bounce or hesitation behavior** among lower-spending customers. Improve the **user experience on the website, simplify purchase flows, and test different value messages** to reduce friction and encourage conversions.

BUSINESS RECOMMENDATION

04

Tailor offers based on **family responsibilities**. For example, customers with **no dependents** may **respond better to premium offers**, while those **with dependents** might prefer **practical bundles or savings-driven deals**.

05

Segment **campaign messaging** based on **income and education profiles**. Promote premium products like **wine to higher-income customers**, while highlighting **affordable value for more budget-conscious** segments.

06

Make **segmentation part** of the **regular campaign planning**. **Don't send** the **same message to everyone** and **change** it into more like **design** offers and **timing** based on behavior, not just demographics, to increase relevance and effectiveness.

BUSINESS RECOMMENDATION

Customer's Segments		Recommendation
Cluster 3	Core Champions	Introduce to them a loyalty program. These are the top customers and by offering them exclusive benefits, early access, or tailored rewards can help maintain their loyalty.
Cluster 2	Newcomer Shoppers	Nurture these newer users with educational content and onboarding campaigns. Help them discover more of the product range, build trust, and guide them toward their next purchase.
Cluster 1	High-Value At Risk	Re-engage with these high-potential but inactive customers. These are previously valuable customers who've gone quiet. A personalized win-back campaign or exclusive incentive may help bring them back.
Cluster 0	Dormant Buyers	Offer these most customers's segment with first-time discounts, free delivery, or loyalty points to encourage a return purchase. Provide them with low-cost product bundles to make re-entry more appealing for price-sensitive customers.

THANK YOU

Johanes Purek

johanesbernard26@gmail.com

github.com/johanbernardd