CUSTOMERS SEGMENTATION USING K-MEANS

Deni Ramdani
https://www.linkedin.com/in/den-rmdani/

TABLE OF CONTENTS

INTRODUCTION

01

WHAT IS THE SEGMENTATION ANALYSIS?

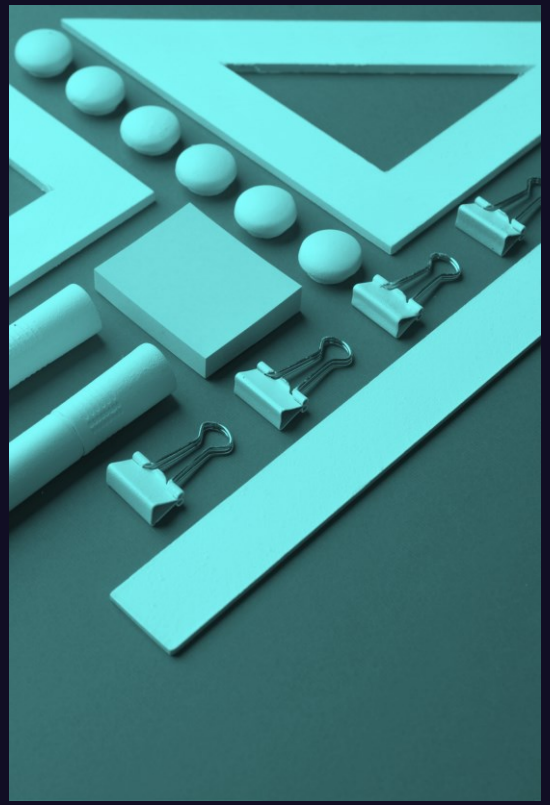
Segmentation analysis is a **technique** used **to group** data points into **meaningful categories**. It is applicable in various fields, including marketing (market segmentation analysis) and customer relationship management (customer segmentation analysis).

WHAT IS THE BENEFIT OF SEMENTATION ANALYSIS?

In marketing,

- **Retail Customer Segmentation:** Helps in identifying consumer expectations and improving business strategies.
- **Customer Segment Pricing:** Differentiates consumer segments based on price sensitivity and value perception.
- **Micro-markets:** Focuses on targeting specific segments within the main market to deliver tailored marketing messages.

Overall, **segmentation analysis** is portrayed as crucial for small businesses **to optimize marketing budgets** and **generate more sales** by targeting the most relevant consumer segments.



TYPES OF SEGMENTATION

DEMOGRAPHIC

Ex: age, gender, income, and education

GEOGRAPHIC

Ex: ZIP code, city, country, radius around a certain location

PSYCHOGRAPHIC

Ex: personality traits, interest, lifestyles, motivation, and priorities

BEHAVIORAL

Ex: puchasing habits, spending habits, user status, and brand interactions

DATA EXPLORATION 02

DATA UNDERSTANDING

This project utilized e-commerce data obtained from Kaggle.

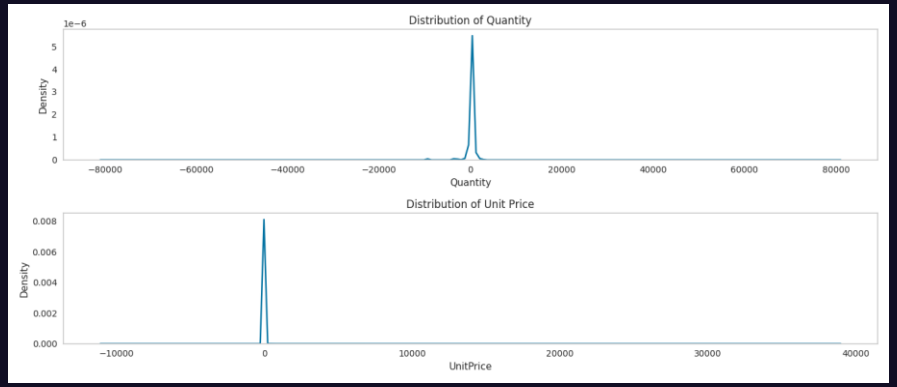
This dataset encompasses international transactions from a UK online shop between December 1st, 2010 and December 9th, 2011. The company is a leading provider of one-of-a-kind gifts for various occasions. Notably, a significant portion of their customer base consists of bulk buyers.

This data set comprises eight columns. The data set includes the following fields: Invoice Number, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, and Country.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom

Number of rows: 541909, number of columns: 8

	Name	dtypes	Missing	Missing_%	Uniques
0	InvoiceNo	object	0	0.000000	25900
1	StockCode	object	0	0.000000	4070
2	Description	object	1454	0.268311	4223
3	Quantity	int64	0	0.000000	722
4	InvoiceDate	object	0	0.000000	23260
5	UnitPrice	float64	0	0.000000	1630
6	CustomerID	object	135080	24.926694	4372
7	Country	object	0	0.000000	38



This report analyzes customer data from December 1st, 2010, to **December 9th, 2011**. Key findings include:

- There are 541909 **data entries**
- There are 4372 unique customers recorded.
- 4223 unique products were sold.
- Orders were made from 38 different countries, with the UK having the highest share.

- 135080 invoices lack customer IDs.
- 1454 invoices have no product descriptions.
- The distribution of the quantity column spans from extremely negative values to extremely positive values.
- The distribution of the unit price column also spans from extremely negative values to extremely positive values.
- There are 5268 duplicate records.

DATA CLEANING

03

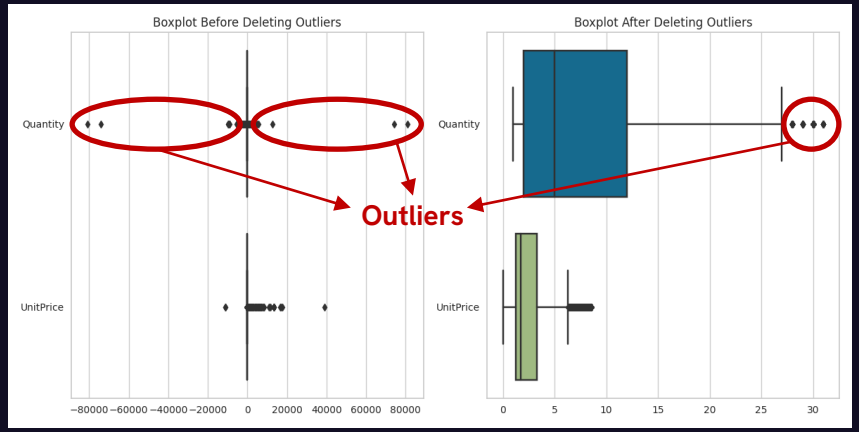
To understand how customers typically buy (purchase behavior), I'll focus on specific information from the data. Here's what I'll use:

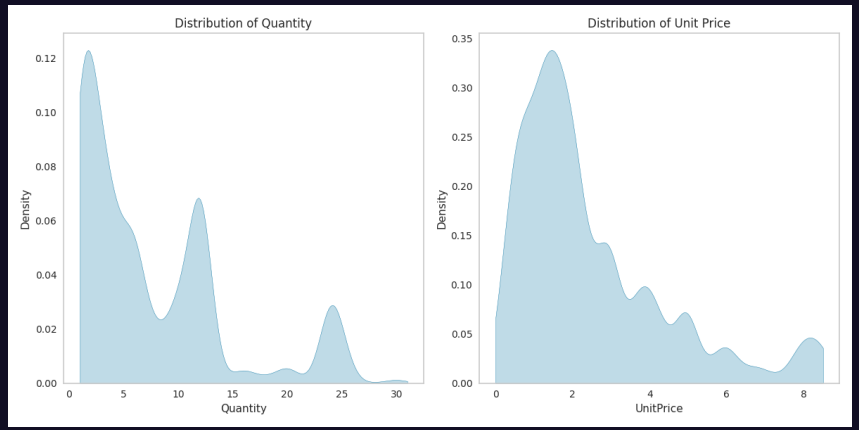
- Invoice details (number and date)
- Customer identification (ID)
- How much each item costs (unit price)
- How many of each item were bought (quantity)

Before I can analyze this data, I need to clean it up. Here's what I'll do and why:

1. **Removing duplicates**, they were most likely recorded due to a system error.
2. **Subsetting, columns** of interest for easier preprocessing.
3. **Removing invoices with null** customer IDs since the analysis requires known IDs.
4. **Removing invoices with negative** quantity and unit price values, as well as zero values if applicable. *Negative values refer to cancelled invoices.*
5. **Removing outliers** in both quantity and unit price columns. Having outliers usually distort the results of most analyses. I'll use z-score to remove them.

Boxplot & Distribution Before-After Data Cleaning





FEATURES ENGINEERING

04

The features in this dataset that tells us about customer buying behavior include **Quantity**, **Invoice Date**, and **Unit Price**. We are going to derive a **customer's RFM** (Recency, Frequency, Monetary) value using these variables.

R

Recency. ***How recently*** *a customer has made a purchase*

F

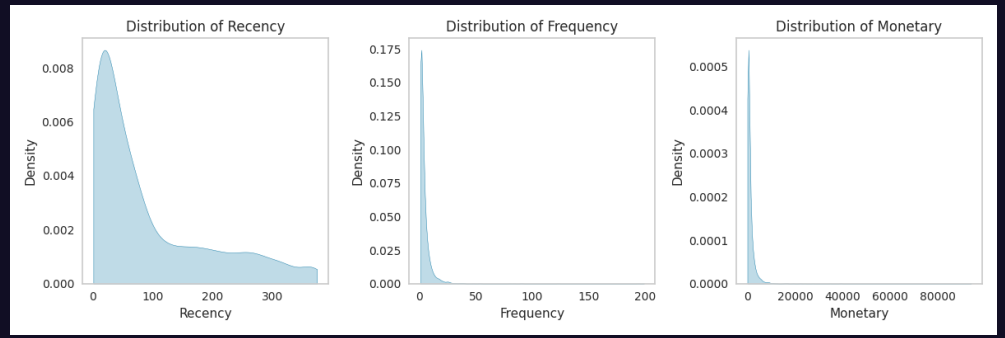
Frequency. ***How often*** *a customer makes a purchase*

M

Monetary. ***How much*** *money a customer spends on purchases*

CustomerID	Recency	Frequency	Monetary
12347	2	7	3314.73
12348	249	3	90.20
12349	19	1	984.15
12350	310	1	294.40
12352	36	7	1130.94

	Recency	Frequency	Monetary
count	4190.00	4190.00	4190.00
mean	92.52	4.01	1022.84
std	99.92	7.02	2190.41
min	1.00	1.00	1.90
25%	18.00	1.00	206.01
50%	51.00	2.00	465.52
75%	144.00	4.00	1126.52
max	374.00	196.00	84635.89



This is a **data description** and **distribution** of the engineering features

PREPROCESSING DATA

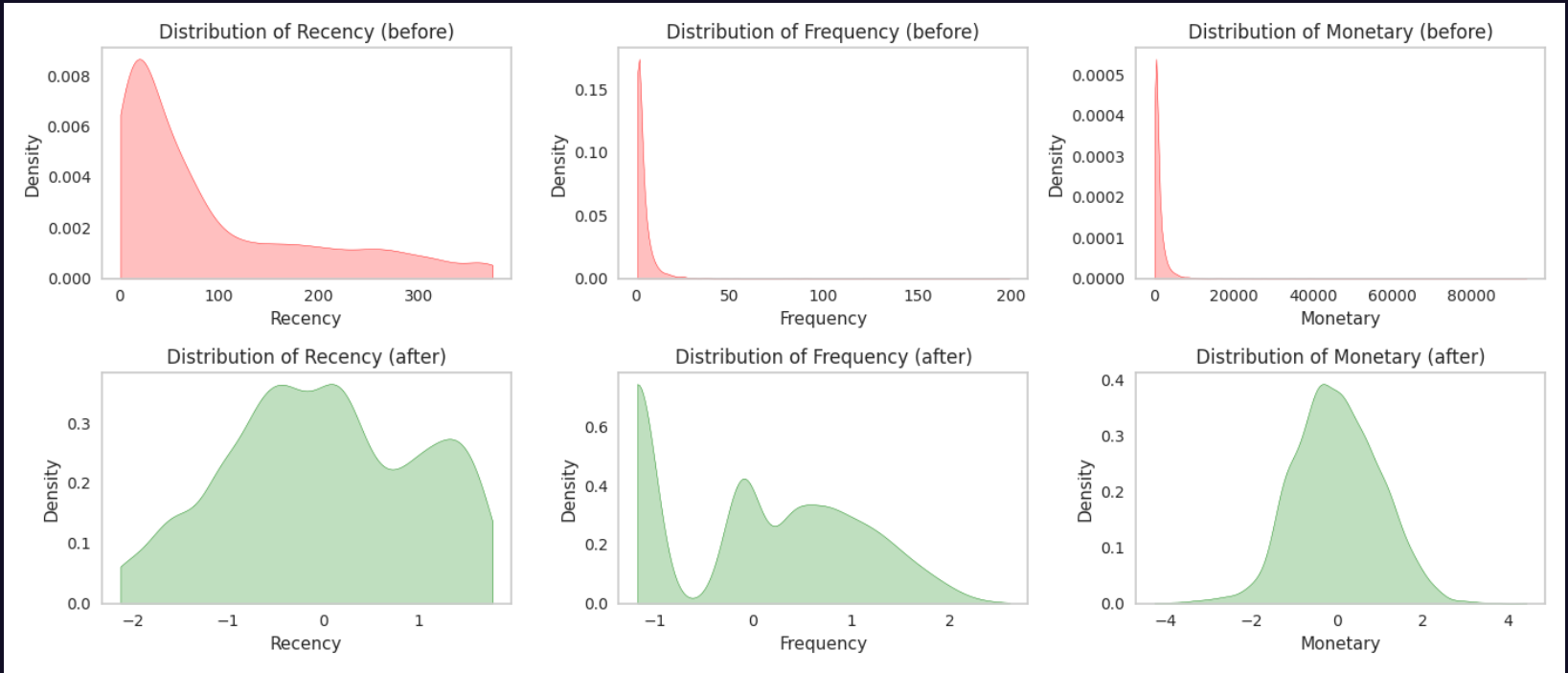
05

While **K-Means** is a powerful tool for customer segmentation, it has some expectations about the data it works with. Here's what we need to consider:

- **Data Distribution**: KMeans works best with data that follows **a bell-shaped curve (normal distribution)**. Our data appears skewed, so we'll need to transform it to be more like a normal distribution.
- **Feature Scaling**: KMeans relies on distances between data points. If some features have much larger values than others, it can distort the results. We'll check if our data needs scaling to ensure all features contribute equally.
- **Outliers**: Both purchase frequency and monetary value might have outliers, which can throw off some scaling methods.

To address these concerns, we'll use a technique called **PowerTransformer** from scikit-learn. This tool can transform the data to be closer to a normal distribution using methods like Yeo-Johnson or Box-Cox.

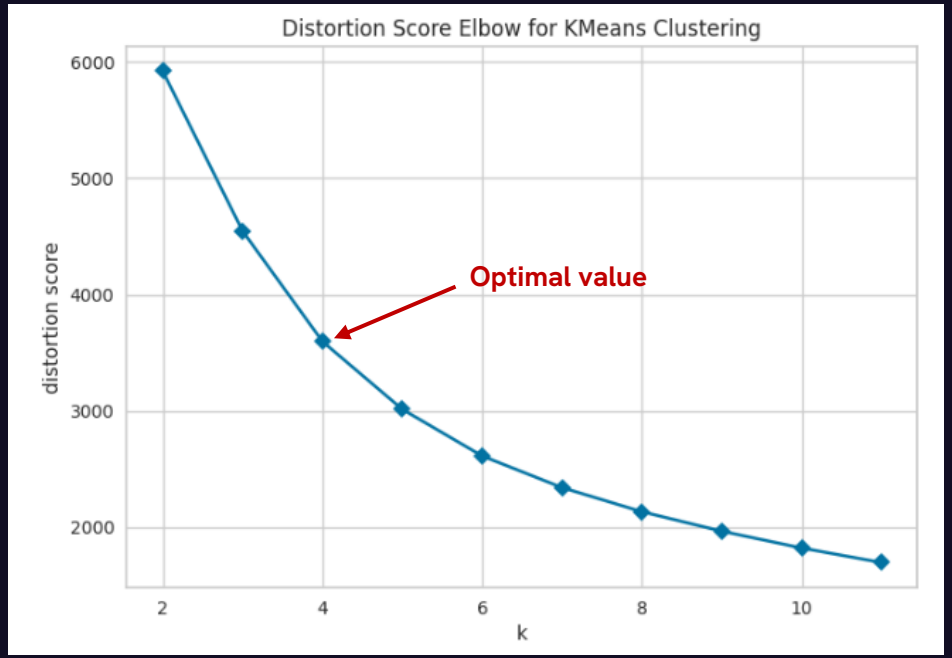
- Distribution features before & after transform data features with power transform



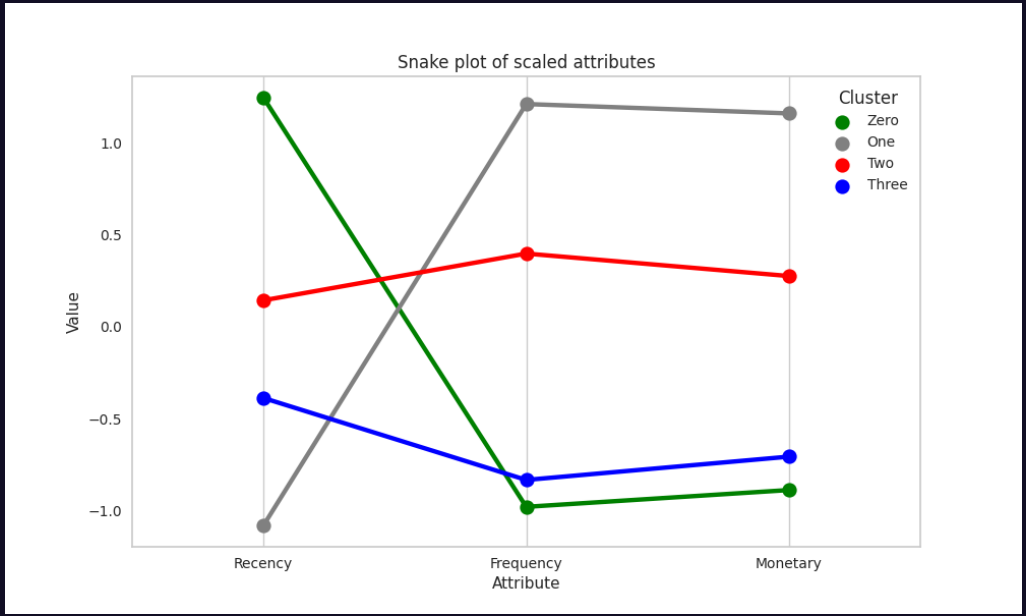
K-MEANS CLUSTERING

06

- The **elbow method** will now be employed in order to identify the optimal number of clusters for the data.

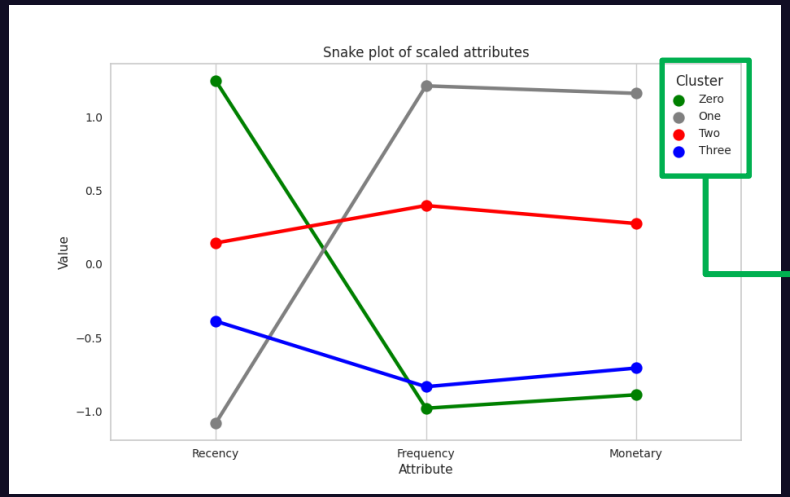


The optimal number of clusters is $k = 4$, as evidenced by the addition of the k value, which has a minimal impact on the distortion score value.

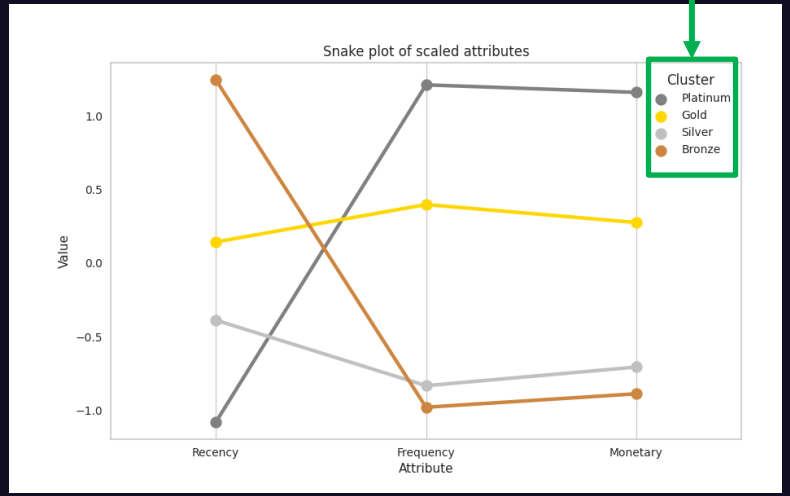


The **value of each cluster attribute** is obtained from **pandas.melt**, as illustrated in the accompanying figure.

- What do these four label names mean?



Silver: These are your customers who purchased a decent number of times and spent good amounts, but haven't purchased recently.
Bronze: These are customers who used to visit and purchase in your platform, but haven't been visiting recently.



Platinum: These customers are the most loyal and have recently, most often, and at a high level, spent money on the product.
Gold: The individuals in question are recent customers who have exhibited a high level of frequency and have spent a considerable amount of money.

RECOMMENDATION & CONCLUSION

07

Platinum Customers (Most Loyal, Recent, High Spenders):

Focus: Maintain their satisfaction.

Recommendations:

- **Exclusive benefits**: Offer exclusive discounts, early access to new products, or VIP memberships.
- **Personalized experiences**: Recommend products based on their purchase history and preferences.
- **High-touch communication**: Send personalized birthday greetings, anniversary discounts, or thank-you notes.
- **Early access to sales**: Inform them about upcoming sales or promotions before the general public.
- **Utilize customer reviews**: Encourage them to leave reviews and testimonials to build trust for new customers.

Gold Customers (Recent, Average Frequency & Spend):

Focus: Increase engagement and encourage repeat purchases.

Recommendations:

- **Loyalty programs**: Enroll them in a tiered loyalty program with rewards for frequent purchases.
- **Targeted promotions**: Send personalized offers and discounts based on their past purchases.
- **Win-back campaigns**: If they haven't purchased recently, send re-engagement emails reminding them of your brand and highlighting new products.
- **Flash sales**: Offer limited-time discounts or exclusive deals to create a sense of urgency.
- **Request feedback**: Ask for their input on their shopping experience and product preferences.

Silver Customers (Decent Purchase History, Not Recent):

Focus: Re-activate them and encourage future purchases.

Recommendations:

- **Win-back campaigns**: Design targeted campaigns with special offers or discounts tailored to their past purchases.
- **Reactivation emails**: Send personalized emails reminding them of abandoned carts, highlighting new products, or offering exclusive deals.
- **Seasonal promotions**: Target them with relevant promotions during holidays or special occasions.
- **Content marketing**: Engage them with informative blog posts, newsletters, or social media content related to their interests.
- **Analyze churn reasons**: Understand why they stopped purchasing and address any underlying issues.

Bronze Customers (Past Visitors/Buyers, Not Recent):

Focus: Re-engage them and entice them to return.

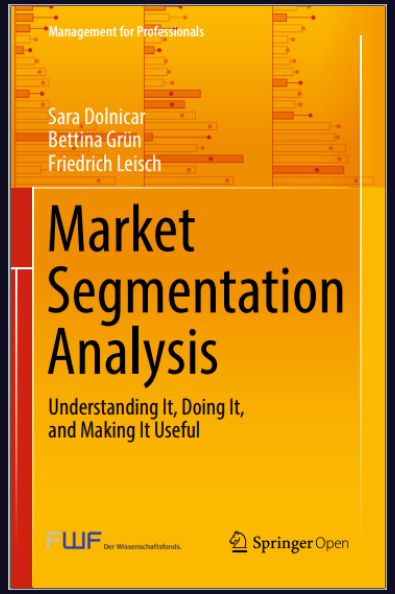
Recommendations:

- **Win-back campaigns**: Utilize targeted ads or email campaigns with attractive offers to rekindle their interest.
- **Personalized recommendations**: Recommend products based on their past browsing behavior (if available).
- **Social media engagement**: Re-engage them on social media platforms by running contests, polls, or influencer marketing campaigns.
- **Website retargeting**: Use retargeting ads to remind them of products they viewed or abandoned carts.
- **Exit-intent popups**: Offer special discounts or incentives to capture their attention before they leave your website.

CONCLUSIONS

1. Although **K-Means clustering can be a valuable tool**, it is not without limitations. One significant challenge is the lack of inherent interpretability. It is challenging to identify the specific attributes that distinguish one cluster from another within the K-Means framework.
2. In contrast, **RFM analysis** offers a more **interpretable approach**. By segmenting customers based on their recency (how recently they purchased), frequency (how often they purchase), and monetary value (how much they spend), we gain a **clear understanding of their purchasing behavior**. This provides a robust basis for subsequent clustering, enabling the grouping of customers with analogous purchasing patterns.

REFERENCES



5 min read

Data Analysis for Market Segmentation

Data scientists have a lot of useful instruments for helping companies to make the right dec...

mari-galdina.medium.com

Mari Galdina Medium

6 min read

Using Segmentation Analysis to Improve Marketing

Segmentation analysis is an important key when creating your business plan or launching your ...

medium.com/@anand.mishra

Anand Mishra Medium

THANK YOU

den.rmdani@gmail.com

https://www.linkedin.com/in/den-rmdani/