OPTIMIZATION OF FASHION RETAIL CUSTOMER DATA MANAGEMENT THROUGH EXPLORATORY DATA ANALYSIS AND RECENCY, FREQUENCY, MONETARY

Yusuf[1], Lody Saladin Basri[1], Widi Hastomo[2]* dan Ire Puspa Wardhani[1]

(1)Sekolah Tinggi Manajemen Informatika dan Komputer Jakarta STI&K
Jalan BRI No. 17, Radio Dalam, Kebayoran Baru, Jakarta Selatan 12140
(2)Institut Teknologi dan Bisnis Ahmad Dahlan Jakarta
Jl. Ir H. Juanda No.77, Cireundeu, Ciputat Tim., Tangerang Selatan 15419
akhiyus@gmail.com, lody.saladin@gmail.com, widie.has@gmail.com, irepuspa@gmail.com,
*Correscponding Author

ABSTRAK

This study was conducted to analyze revenue patterns, product segmentation, and customer retention in the H&M retail business using Kaggle competition data "H&M Personalized Fashion Recommendations." The urgency of this study lies in the need to understand revenue fluctuations and customer behavior in order to optimize business strategies. The data used includes transactions, articles, and customer profiles from 2018 to 2020. The analysis methods applied include exploratory data analysis (EDA) analysis and customer segmentation using the RFM (recency, frequency, and monetary) model to identify customer groups based on purchasing behavior. The results of the study show that the highest revenue occurs in the middle of the year, with a sharp decline in growth in mid-2018. Low-recency customers contribute more to revenue, while product segmentation shows the need for stock adjustments, especially for baby/children and divided products. This study successfully identified key factors that influence revenue and customer retention and provided strategic recommendations for inventory improvement and market segmentation. These results are important for H&M to improve operational efficiency and improve marketing strategies in the future.

Kata Kunci: *EDA, RFM, Fashion retail.*

INTRODUCTION

In the fast-paced and dynamic digital era, the fashion retail industry is faced with a major challenge to deeply understand customer preferences and behavior [1]. Technological advances have driven an increase in the volume of transaction data [2], so that retail companies now have access to a variety of information about customer shopping patterns, purchase frequency, and transaction value. However, without proper management, this data often becomes an asset that is not utilized optimally. This is where the role of exploratory data analysis (EDA) and the recency, frequency, and monetary (RFM) method becomes very important to explore strategic insights that can lead to increased customer loyalty and retention [3].

Exploratory Data Analysis (EDA) enables companies to understand hidden patterns, anomalies, and relationships within data through visualization and statistical approaches [4]. With EDA, fashion retail companies can identify important trends such as when customers are likely to make purchases, which products are most in demand, and what factors influence purchasing decisions. Meanwhile, the RFM method, which focuses on three key elements: recency, frequency, and transaction value (monetary), provides an effective way to segment customers based on their shopping behavior [5]. With this segmentation, companies can direct more targeted and relevant marketing campaigns, improve customer experience, and ultimately optimize revenue.

Combining EDA and RFM in customer data management provides a powerful analytical approach to improve fashion retail business performance [6]. With deeper insights into customer preferences, companies can tailor targeted product, promotion, and service strategies [7]. Furthermore, optimizing this data management helps companies make more accurate and efficient data-driven decisions, creating sustainable added value in a highly competitive market.

Research conducted by [8] using the RFM model and K-means clustering, with the results strengthening the understanding of customer segmentation, increasing the company's ability to drive loyalty, and increasing profitability through personalized strategies. Research by [9] with the results that the use of machine learning in sales forecasting can increase accuracy and efficiency and provide a competitive advantage for businesses.

Research by [10] used the K-Means and Hierarchical Clustering methods and EDA data analysis with results emphasizing the importance of customer segmentation in analyzing the market, which allows companies to develop more effective marketing strategies and increase competitiveness.

Based on a number of literatures obtained, research on RFM and EDA has been widely conducted by comparing several methods, but not many use the competition dataset from Kaggle.com [11]. Therefore, this research was developed using the RFM method with EDA analysis using the competition dataset from Kaggle.com.

RESEARCH METHODS

This study uses a dataset from the Kaggle.com competition [11]. The RFM (Recency Frequency Monetary Value) method is widely used to segment customers by dividing the value into three parts:

- Low Value: classified as less active customers and visitors who do not make frequent purchases, providing very low income for entrepreneurs.
- Mid Value: classified as customers who frequently use the company's platform, quite often and generate moderate income for entrepreneurs.
- High Value: customers who are classified as being able to provide high income for the company, low frequency, and inactivity

The RFM method will be implemented by programming with Python language and machine learning to identify each group (group/cluster). The pseudocode is represented in figure 1.

```
1. Import libraries
2. Input dataset
3. Set reference date
4. Calculate RFM values
5. Create RFM DataFrame
6. Rank RFM and calculate RFMscore
7. Segment customers based on RFM
   Score
8. Output result
```

Figure 1. *Pseudocode RFM*

The operating environment used to support this research is Intel Core i5 RAM: 8 GB with Storage: SSD 256 GB. The operating system uses Windows 10 Python language with Jupyter notebook IDE. Python libraries: Pandas, Numpy, Scikit Learn, Matplotlib/Seaborn, and Datetime.

Because the dataset is quite large, external processing with the help of GPU available on the Google NVIDIA Tesla T4, architecture: turing, memory: 16 GB GDDR6, CUDA Cores: 2560, tensor cores: 320 (for enhanced deep learning performance), performance: can provide up to 8.1 teraflops (single-precision) and 65 teraflops (tensor performance), use cases: great for AI inferencing, deep learning training, and machine learning tasks, especially for more energy-efficient processing.

The research flow is shown in Figure 2. It starts with analyzing the dataset with EDA, followed by the dataset exploration process consisting of the revenue cycle and customer retention, followed by customer segmentation with the RFM method, then the visualization process using a treemap diagram.

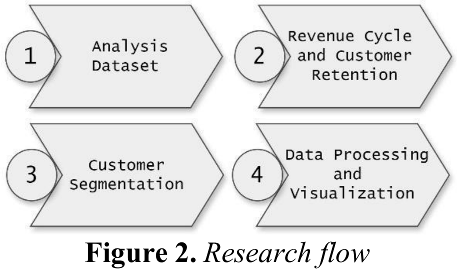

1
Analysis Dataset
2
Revenue Cycle and Customer Retention
3
Customer Segmentation
4
Data Processing and Visualization

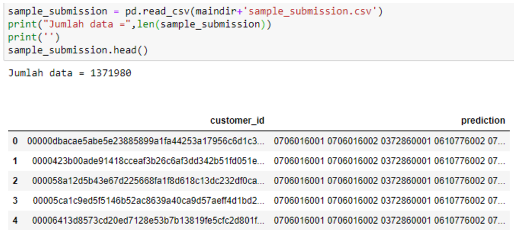
Figure 2. *Research flow*

RESULTS AND DISCUSSION

The proposed methodology is implemented on a synthetic dataset of H&M store customer transactions with 53 online marketplaces and 4850 stores. The stages of the research process start from:

1. Dataset Analysis

The dataset obtained consists of 4 files with csv extensions ("articles.csv", "customer.csv," "sample_submission.c sv," and "transactions_train.csv") and 1 "images" folder containing product image files with jpg extensions.The "sample_submission.csv" file contains the file format for submission in this competition. This file contains predictions of 1,371,980 customer data that are expected to buy 12 products in the next time period (figure 3).



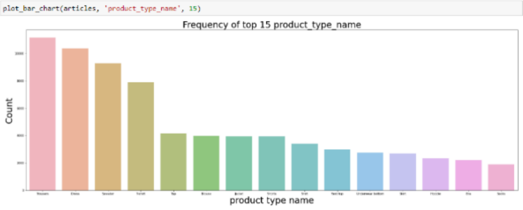
```
sample_submission = pd.read_csv(maindir+'sample_submission.csv')
print("Jumlah data =",len(sample_submission))
print('')
sample_submission.head()
```

```
Jumlah data = 1371980
```

	customer_id	prediction
0	00000dbacae5abe5e23885899a1fa44253a17956c6d1c3...	0706016001 0706016002 0372860001 0610776002 07...
1	0000423b00ade91418cceaf3b26c6af3dd342b51fd051e...	0706016001 0706016002 0372860001 0610776002 07...
2	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	0706016001 0706016002 0372860001 0610776002 07...
3	00005ca1c9ed5f5146b52ac8639a40ca9d57aeff4d1bd2...	0706016001 0706016002 0372860001 0610776002 07...
4	00006413d8573cd20ed7128e53b7b13819fe5cfc2d801f...	0706016001 0706016002 0372860001 0610776002 07...

Figure 3. *File "sample_submission.csv"*

From the frequency histogram in 15 product types, "Trouser" is the article with the highest number of product type names (figure 4).



```
plot_bar_chart(articles, 'product_type_name', 15)
```

Figure 4. *Frequency of "product_type_names"*

The product type with the name "Trouser," the department with the name "Jersey," and the graphic name "Solid" have the largest number of article_ids (figure 5).

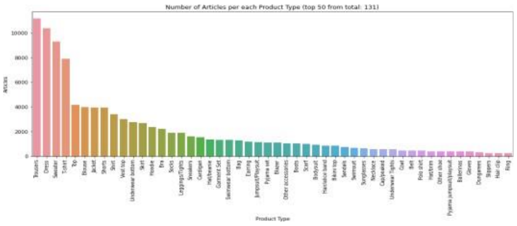
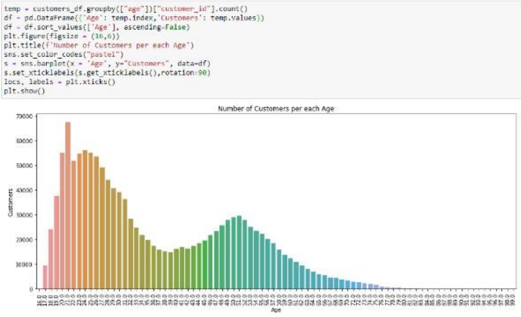


Figure 5. *Number of articles for each product type name*

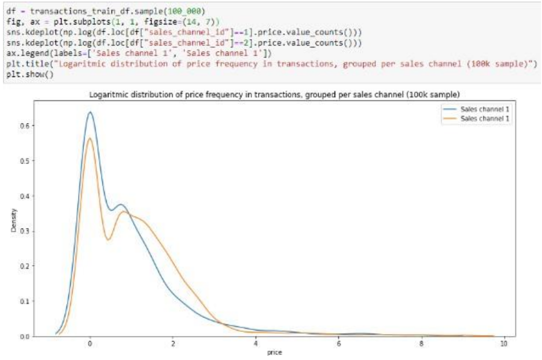
Customers who shop and receive services provided by H&M are mostly relatively young, between 20 and 30 years old (figure 6).



```
temp = customers_df.groupby(["age"])["customer_id"].count()
df = pd.DataFrame({'Age': temp.index,'Customers': temp.values})
df = df.sort_values(['Age'], ascending=False)
plt.figure(figsize = (16,6))
plt.title(f'Number of Customers per each Age')
sns.set_color_codes("pastel")
s = sns.barplot(x = 'Age', y="Customers", data=df)
s.set_xticklabels(s.get_xticklabels(),rotation=90)
locs, labels = plt.xticks()
plt.show()
```

Figure 6. *Number of customers per age*

Figure 7-9: Sales channel group 2 (online) managed to make a sales transaction volume that exceeded sales channel group 1 (offline). The highest transaction volume value occurred in 2019 at the end of September (September) and the beginning of October (October), the end of November (November), and the beginning of December (December). However, the transaction price of each product item sold by sales channel 1 exceeded sales channel 2. So sales channel group 2 has a higher sales volume but with a lower price for each product item compared to sales channel group 1. For offline transactions, there is empty data (no transactions occurred) around April 2020.



```
df = transactions_train_df.sample(100_000)
fig, ax = plt.subplots(1, 1, figsize=(14, 7))
sns.kdeplot(np.log(df.loc[df["sales_channel_id"]==1].price.value_counts()))
sns.kdeplot(np.log(df.loc[df["sales_channel_id"]==2].price.value_counts()))
ax.legend(labels=['Sales channel 1', 'Sales channel 1'])
plt.title("Logaritmic distribution of price frequency in transactions, grouped per sales channel (100k sample)")
plt.show()
```

Logaritmic distribution of price frequency in transactions, grouped per sales channel (100k sample)
Sales channel 1
Sales channel 1
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
10
price

Figure 7. *Logarithmic distribution of transaction price frequency for each sales channel.*

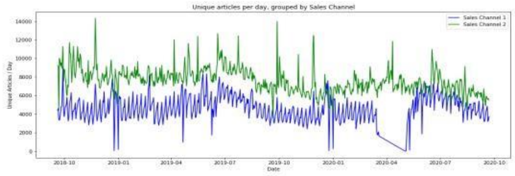


Figure 8. *unique articles per day, grouped by sales channel*

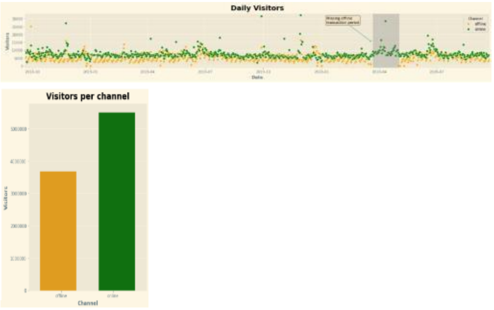


Figure 9. *Plot of daily offline and online visitors*

2. Revenue Cycle and Customer Retention

This study will display some transaction data graphic information that can be used to help measure the business revenue cycle. Information about the company's highest daily revenue occurred around 29-09-2019 and around 6-12-2019 (figure 10).

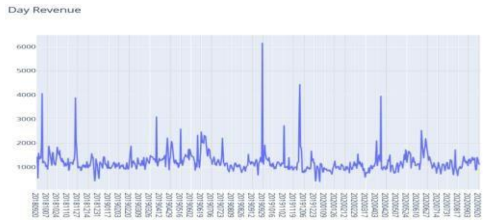


Figure 10. *Daily return*

Active customers each month can bevisualized in the form of a histogram (figure 11).

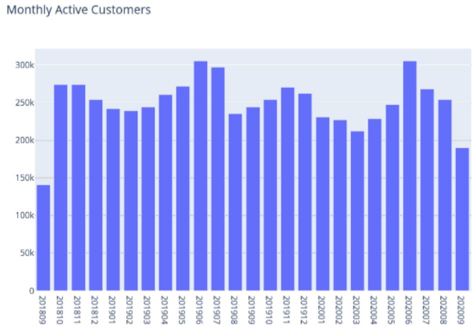


Figure 11. *Monthly active costumers*

The new customer ratio occurred in December 2018 at almost 3.5% and continued to decline to 0.1% in December 2020 (figure 12).

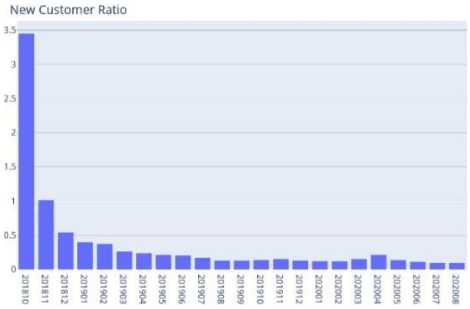


Figure 12. *New costumers ratio*

To find out the average number of customers who return to buy products and use services (retention) from the company, see the histogram in Figure 13.The highest average monthly retention occurred in months 7-8 2019 and 7 2020.

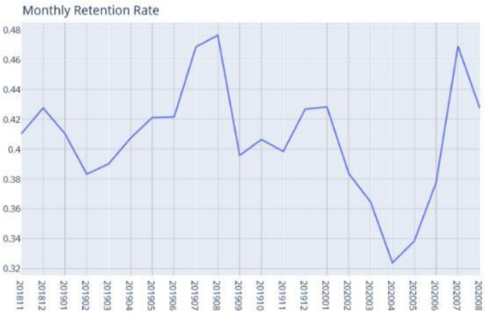
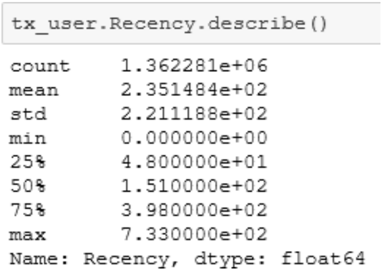


Figure 13. *Monthly retention rate*

3. Customer Segmentation

One of the challenges in this research is the customer segmentation process using large and complex data, and it is impossible to treat each customer the same with the same content, channels, and needs. Each customer has different needs and profiles. Companies must be able to provide services according to the characteristics of the customers they face. If the company wants to increase the average retention, it is necessary to segment based on the possibility of churn and the actions taken.

The recency calculation begins with finding the last purchase date and looking at the number of days of inactivity for each customer. To find out the customer recency overview, you can use the Pandas tool using the describe() method, providing mean, min, max, count, and percentile information (Figure 14).

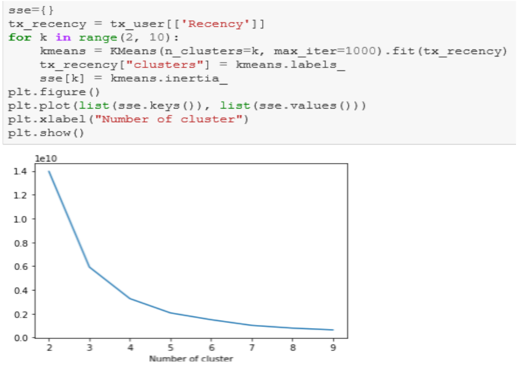


```
tx_user.Recency.describe()
```

```
count     1.362281e+06
mean      2.351484e+02
std       2.211188e+02
min       0.000000e+00
25%       4.800000e+01
50%       1.510000e+02
75%       3.980000e+02
max       7.330000e+02
Name: Recency, dtype: float64
```

Figure 14. *Recency description*

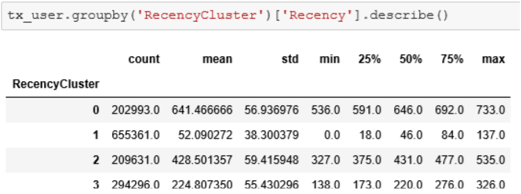
To estimate the optimum number of clusters needed by looking at the visual inertia graph (figure 15).



```
sse={}
tx_recency = tx_user[['Recency']]
for k in range(2, 10):
    kmeans = KMeans(n_clusters=k, max_iter=1000).fit(tx_recency)
    tx_recency["clusters"] = kmeans.labels_
    sse[k] = kmeans.inertia_
plt.figure()
plt.plot(list(sse.keys()), list(sse.values()))
plt.xlabel("Number of cluster")
plt.show()
```

Figure 15. *Inertia recency graph*

By selecting the number of clusters as 4 (according to business needs), a dataframe description for 4 clusters is produced (Figure 16).



```
tx_user.groupby('RecencyCluster')['Recency'].describe()
```

	count	mean	std	min	25%	50%	75%	max
RecencyCluster								
0	202993.0	641.466666	56.936976	536.0	591.0	646.0	692.0	733.0
1	655361.0	52.090272	38.300379	0.0	18.0	46.0	84.0	137.0
2	209631.0	428.501357	59.415948	327.0	375.0	431.0	477.0	535.0
3	294296.0	224.807350	55.430296	138.0	173.0	220.0	276.0	326.0

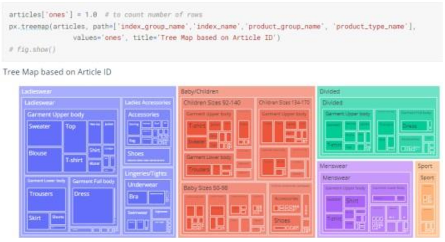
Figure 16. *Dataframe description for 4 recency clusters*

Of the 4 recency clusters, they show different characteristics, where cluster 1 is very recent with a total of 655361 compared to clusters 0, 2, and 3.

4. Data Processing and Visualization

In the data processing and visualization stage, a series of steps are carried out to clean, prepare, and describe the data comprehensively. The dataset used consists of transaction files, articles, and customers that are analyzed using descriptive techniques to understand the distribution and important patterns in the data. Empty or irrelevant data is removed to maintain the accuracy of the analysis. In addition, data visualization is carried out to provide a clearer picture of sales trends, product segmentation, and customer behavior. This visualization helps in understanding the relationship between variables and makes it easier to interpret the results of the analysis.

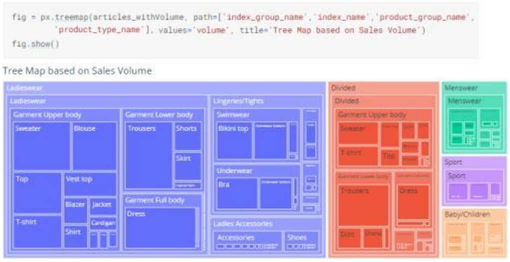
A tree map diagram can be used to further visualize the nuances of the product collection of various H&M product categories. With the tree map, product segmentation can be seen based on articleI_id, index name sales volume, and section_name sales volume. A tree map based on Article ID (figure 17) shows the 5 most articles with ladieswear (30%), babychildren (30%), divided (20%), and meanswear and sport (20%).



```
articles['ones'] = 1.0  # to count number of rows
px.treemap(articles, path=['index_group_name','index_name','product_group_name', 'product_type_name'],
           values='ones', title='Tree Map based on Article ID')
# fig.show()
```

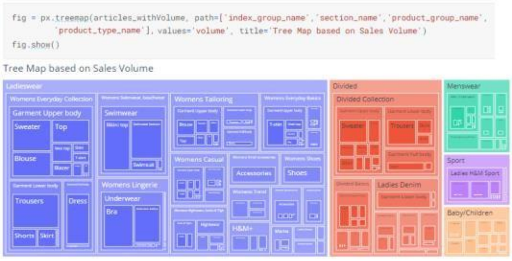
Figure 17. *Treemap based on article_id*

Product segmentation based on sales volume index group name: ladieswear occupies the largest percentage (60%), divided (20%), while menswear, sport, and baby children are around 20% of the total sales volume. Product segmentation based on sales volume is divided into two derivatives, namely index_name (figure 18) and section name (figure 19).



```
fig = px.treemap(articles_withVolume, path=['index_group_name','index_name','product_group_name',
        'product_type_name'], values='volume', title='Tree Map based on Sales Volume')
fig.show()
```

Figure 18. *Treemap based on sales volume (index name)*



```
fig = px.treemap(articles_withVolume, path=['index_group_name','section_name','product_group_name',
        'product_type_name'], values='volume', title='Tree Map based on Sales Volume')
fig.show()
```

Figure 19. *Treemap based on sales volume (section name)*

Tree map Article ID: the Baby/Children group name fills about 30% of all group names, an excessively large number with a relatively small sales volume of around 5%. While the group name with the divided index is only available for 20% with a sales volume of around 20%. To maintain sufficient product availability according to customer needs, there needs to be a reduction in the availability of baby/child products and a slight increase in divided products.

CONCLUSION

The study revealed that H&M's highest revenue consistently occurred in the middle of the year, namely in June to July, from 2018 to 2020. However, there was a significant decline in average growth in mid-2018, indicating a large fluctuation in the revenue cycle. On the other hand, customers with low recency (meaning customers who shop more frequently) contributed more to the company's revenue, while customers with high recency and low shopping frequency contributed less. The data also showed that product segmentation needs to be reviewed, with stock reduction for baby/child products that have low sales performance and stock increase for divided products that are more in demand by customers.

The limitations of this study lie in the absence of an in-depth analysis of external factors that may affect revenue trends and customer retention, such as lifestyle changes, fashion trends, or H&M's marketing policies. Furthermore, this study did not explore specific strategies to address declining growth over a specific period. For future studies, it is important to examine the impact of changes in marketing strategies on customer retention as well as explore the use of more sophisticated machine learning techniques to predict more dynamic purchasing trends and product personalization more accurately.

REFERENCE

[1] J. Z. Zhang, C.-W. Chang, and S. A. Neslin, "How Physical Stores Enhance Customer Value: The Importance of Product Inspection Depth," J. Mark., vol. 86, no. 2, pp. 166–185, Apr. 2021, doi: 10.1177/00222429211012106.

[2] M. S. Hosen et al., "Data-Driven Decision Making : Advanced Database Systems for Business Intelligence," vol. 3, pp. 687–704, 2024.

[3] C. Rungruang, P. Riyapan, A. Intarasit, K. Chuarkham, and J. Muangprathub, "RFM model customer segmentation based on hierarchical approach using FCA," Expert Syst. Appl., vol. 237, p. 121449, 2024, doi: https://doi.org/10.1016/j.eswa.2023.12 1 449.

[4] Harinakshi, A. A. Lydia, M. Poongundran, S. Masarath, P. V Karthick, and I. M. Zayats, "EDA and its Impact in Dataset Discover Patterns in the Service Sector," in 2022 4th International Conference on Inventive Research in Computing Applications

(ICIRCA), 2022, pp. 940–945. doi: 10.1109/ICIRCA54612.2022.9985599.

[5] M. Husnah and R. Novita, “Clustering of Customer Lifetime Value With Length Recency Frequency and Monetary Model Using Fuzzy C-Means Algorithm,” in 2022 International Conference on Informatics Electrical and Electronics (ICIEE), 2022, pp. 1–4. doi: 10.1109/ICIEE55596.2022.10010209.

[6] N. Sharma, “Analyzing Customer Behvior Patterns & Predicting Online Product Return Intentions: A Data Mining Approach,” 2024. [Online]. Available: https://opus4.kobv.de/opus4- rhein-waal/frontdoor/index/index/docId/1924

[7] E. S. Hastomo, W., Karno, A. S. B., Sudjiran, S., Arif, D., & Moreta, “Exloratory Data Analysis Untuk Analisa Data Belanja Pelanggan dan Pendapatan Bisnis Retail Fashion Dengan Bahasa Python,” Infotekmesin, vol. 13, no. 2, 2022, [Online]. Available: https://www.ejournal.pnc.ac.id/index.ph p/infotekmesin/article/view/1319

[8] P. Anitha and M. M. Patil, “RFM model for customer purchase behavior using K-Means algorithm,” J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 5, pp. 1785–1792, 2022, doi:https://doi.org/10.1016/j.jksuci.2019.12. 011.

[9] V. Kadam and S. Vhatkar, “Design and Develop Data Analysis and Forecasting of the Sales Using Machine Learning BT - Intelligent Computing and Networking,” 2022, pp. 157–171.

[10] N. Godcares, A. Sirsath, A. Bongale, P. Kadam, R. Jayawal, and S. Patil, “Exploring Customer Segmentation in the Context of Market Analysis,” in 2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC), 2023, pp. 444–449. doi: 10.1109/R10-HTC57504.2023.10461815.

[11] xlsrln Carlos García Ling, ElizabethHMGroup, FridaRim, inversion, Jaime Ferrando, Maggie, neuraloverflow, “H&M Personalized Fashion Recommendations,” kaggle.com, 2022. https://www.kaggle.com/competitions/h -and-m-personalized-fashion-recommendations