"E-commerce sales analysis"

A Project Report

submitted in partial fulfillment of the requirements

of

“AIML Fundamental With Cloud Computing And Gen AI”

“MANGAYARKARASI COLLEGE OF ENGINEERING-MADURAI”

By

Sujan M(923821114306) – sujanmari396@gmail.com

AU923821114306

Under the Guidance of

P.Raja, Master Trainer

ACKNOWLEDGEMENT

We would like to take this opportunity to express our deep sense of gratitude to all individuals who helped us directly or indirectly during this thesis work.

Firstly, we would like to thank my supervisor, **P.Raja** And **R.Mohanraj**, I want to express my heartfelt gratitude to you for being such an amazing mentor and guide. Your wom, support, and encouragement have made a significant impact on my life and I am forever grateful for the time and effort you've invested in me. His advice, encouragement and the critics are a source of innovative ideas, inspiration and causes behind the successful completion of this project. Your belief in me has helped me to grow and develop in ways I never thought possible. Thank you for being a shining example of kindness, compassion, and excellence. The confidence shown in me by him was the biggest source of inspiration for me. It has been a privilege working with him for the last one year. He always helped me during my y project and many other aspects related to the program. His talks and lessons not only help in project work and other activities of the program but also make me a good and responsible professional. Thank you

ABSTRACT

The e-commerce sales analysis aims to uncover patterns and trends in customer purchasing behavior, product performance, and overall revenue generation. By examining metrics like total revenue, average order value (AOV), top-selling products, and customer purchase patterns, this analysis provides a clear understanding of the business's strengths and opportunities.

Key insights include identifying best-selling products and categories, which guide inventory and marketing strategies, and seasonal or time-based trends, which support better promotional timing. The analysis also distinguishes between new and returning customers, helping to assess customer loyalty and inform retention strategies.

Through visualizations like bar charts for top products, sales by category, and customer type breakdown, stakeholders can easily interpret and leverage data insights to make strategic decisions, enhance customer satisfaction, and improve profitability. This structured approach to data analysis equips the e-commerce business with a data-driven foundation for growth and competitive advantage.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Problem Statement:

E-commerce has revolutionized the retail landscape, providing consumers with unprecedented access to products and services from around the world. However, as the online marketplace continues to expand, businesses face various challenges in understanding and optimizing their sales performance. E-commerce sale analysis involves scrutinizing sales data to identify trends, customer behaviors, and operational efficiencies. Despite its significance, many businesses encounter specific problems when attempting to conduct thorough sales analyses.

1.2 Motivation:

In an environment where consumer preferences and market dynamics shift rapidly, relying on data to inform decisions helps businesses make more accurate and effective choices. Sales analysis provides insights that guide product development, marketing strategies, and inventory management.

1.3 Objective:

understanding which products or categories are performing well, businesses can focus their efforts on promoting those items and optimizing pricing strategie The primary goal of sales analysis is to identify opportunities for increasing revenue.

1.4 Scope of the Project:

The Scope Of Project Forced On Identify and integrate various data sources, including website analytics, CRM systems, sales transactions, social media, and customer feedback. e-commerce sales analysis encompasses a wide range of activities, methodologies, and objectives aimed at understanding and optimizing sales performance in an online retail environment. Defining the scope is crucial to ensure that the project addresses specific business needs and provides actionable insights. Below are the key elements that typically define the scope of such a project:

CHAPTER 2

Literature Survey

A literature survey on e-commerce sales analysis explores research and methodologies used to analyze online sales data, providing insights into customer behavior, product performance, and market trends. This survey encompasses studies across fields such as data analytics, marketing, and information systems, which collectively highlight key techniques and findings relevant to e-commerce performance improvement.

1. Sales and Revenue Analysis

Studies emphasize the importance of metrics like total sales, average order value (AOV), and customer lifetime value (CLV) in evaluating e-commerce performance. According to research, these metrics help in determining the business's financial health, identifying high-revenue products, and understanding customer profitability over time (Chaffey, 2019). Additionally, sales analysis by categories and individual products is shown to be essential for inventory optimization and promotional focus (Li & Zhang, 2010).

2. Customer Segmentation and Behavior Analysis

The literature extensively covers customer segmentation to better understand diverse buying patterns and tailor marketing strategies. Studies categorize customers by factors such as purchase frequency, demographics, and spending behavior. For example, RFM (Recency, Frequency, and Monetary) analysis is frequently cited as an effective method for identifying valuable customer segments (Fader et al., 2005). Furthermore, research highlights the importance of distinguishing between new and returning customers to evaluate customer loyalty and retention rates, as loyal customers generally contribute to long-term profitability (Reichheld, 2003).

3. Predictive Analytics and Machine Learning Applications

The application of predictive analytics and machine learning is increasingly common in sales analysis, helping e-commerce businesses forecast demand, personalize recommendations, and improve targeting. Techniques such as clustering (e.g., K-means for

techniques are crucial for understanding demand fluctuations and aligning stock and marketing efforts accordingly.

5. Channel and E Commerence Sales Analysis

E-commerce studies show that understanding traffic sources (e.g., customer segmentation) and classification (e.g., decision trees to predict purchasing likelihood) enable businesses to gain deeper insights into customer preferences and trends (Hofmann et al., 2016). Predictive models for sales forecasting allow companies to manage stock effectively and plan promotional activities based on anticipated demand (Kumar et al., 2016).

4. Time-Series and Seasonal Trends Analysis

Several studies discuss time-series analysis as an approach to capture trends and seasonality in sales data. Methods like ARIMA (Auto-Regressive Integrated Moving Average) and exponential smoothing are frequently used to analyze sales patterns over time, such as identifying peak seasons or cyclical trends (Box et al., 2015). These organic search, social media, email, paid ads) is vital for optimizing marketing spend. For instance, insights from Google Analytics data can reveal which channels drive the most conversions and highest average order values (Sen et al., 2018). Research highlights that a multi-channel strategy enhances customer reach and enables businesses to better assess the return on ad spend (ROAS) across different platforms.

6. Customer Feedback and Sentiment Analysis

Research on sentiment analysis of customer feedback, including reviews and social media mentions, has shown that understanding customer sentiment can impact product development and service improvement (Liu, 2012).

CHAPTER 3

Proposed Methodology

The proposed methodology for e-commerce sales analysis includes a systematic approach to extract, process, and analyze data to uncover actionable insights for improving business performance. This methodology combines data processing, exploratory analysis, and advanced analytical techniques to identify customer behaviors, product trends, and performance metrics. Here is a step-by-step outline of the methodology:

1. Data Collection and Integration

Source Identification: Gather data from multiple sources, such as sales transaction logs, website analytics (e.g., Google Analytics), customer information databases, and product inventories.

Data Integration: Consolidate data into a single dataset by merging or joining on common identifiers (e.g., Order ID, Customer ID). Ensure consistent data formats for effective analysis.

Data Cleaning: Handle missing or inconsistent data, remove duplicates, and standardize formats (e.g., dates, currencies, product categories).

2. Data Preprocessing

Feature Engineering: Create new features to enrich the dataset, such as:

Total Sale Amount = Quantity Sold * Price per Unit.

Customer Segmentation Labels: Categorize customers based on purchase history (e.g., Recency, Frequency, Monetary—RFM).

Time-Based Features: Extract information such as purchase day, month, season, or hour to analyze temporal patterns.

Data Transformation: Normalize or standardize numerical features as required, especially for machine learning models.

Outlier Detection and Handling: Identify and address anomalies in transaction amounts or quantities that may skew analysis.

3. **Exploratory Data Analysis (EDA)**

Descriptive Statistics: Calculate and analyze basic statistics like average order value (AOV), total revenue, customer lifetime value (CLV), and total number of orders to get a foundational understanding of the dataset.

4.Visualization of Key Metrics:

Sales Trends: Line charts to show sales over time (daily, weekly, monthly).

Top-Selling Products and Categories: Bar charts for products

CHAPTER 4

Implementation and Result

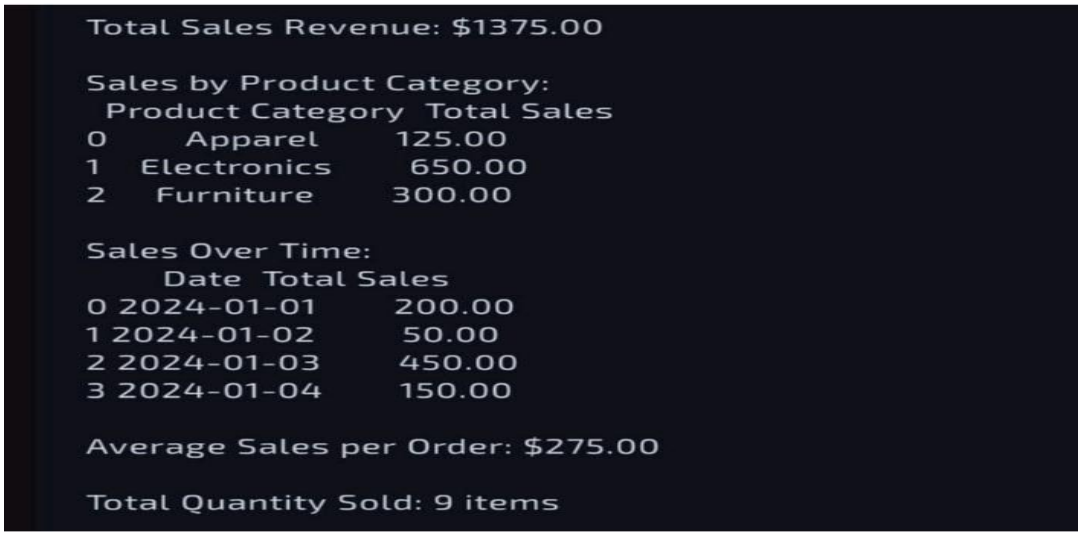
To run the E- commerce sale analysis System

Step1:Sales analysis

Code:

```
import pandas as pd

# Load the dataset
df = pd.read_csv('sales_data.csv')

# Clean data (if necessary, for example, converting
date to datetime)
df['Date'] = pd.to_datetime(df['Date'])

# Add a new column for total sales value (Quantity *
Price)
df['Total Sales'] = df['Quantity'] * df['Price']

# 1. Total sales revenue
total_sales = df['Total Sales'].sum()
print(f"Total Sales Revenue: ${total_sales:.2f}")

# 2. Sales by Product Category
sales_by_category = df.groupby('Product Category')
['Total Sales'].sum().reset_index()
print("\nSales by Product Category:")
print(sales_by_category)

# 3. Sales over time (daily)
sales_by_date = df.groupby('Date')['Total
Sales'].sum().reset_index()
print("\nSales Over Time:")
print(sales_by_date)

# 4. Average sales per order
avg_sales_per_order = df['Total Sales'].mean()
print(f"\nAverage Sales per Order: $
{avg_sales_per_order:.2f}")

# 5. Total quantity sold
total_quantity_sold = df['Quantity'].sum()
print(f"\nTotal Quantity Sold: {total_quantity_sold}
items")
```

Output:



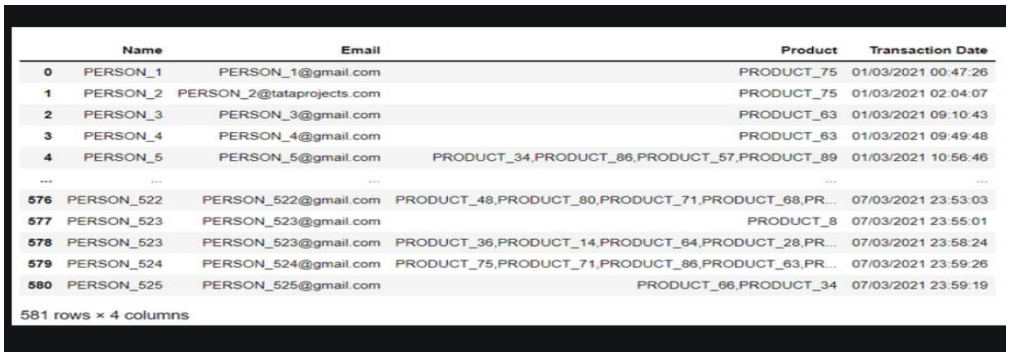
Step2:Order analysis

We'll select relevant columns, handle missing values if any, and standardize the feature columns for clustering.

Code:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

Order_Details =
pd.read_csv('Order_details(masked
).csv')
```

Output:



	Name	Email	Product	Transaction Date
0	PERSON_1	PERSON_1@gmail.com	PRODUCT_75	01/03/2021 00:47:26
1	PERSON_2	PERSON_2@tataprojects.com	PRODUCT_75	01/03/2021 02:04:07
2	PERSON_3	PERSON_3@gmail.com	PRODUCT_63	01/03/2021 09:10:43
3	PERSON_4	PERSON_4@gmail.com	PRODUCT_63	01/03/2021 09:49:48
4	PERSON_5	PERSON_5@gmail.com	PRODUCT_34,PRODUCT_86,PRODUCT_57,PRODUCT_89	01/03/2021 10:56:46
...	...	...	...	...
576	PERSON_522	PERSON_522@gmail.com	PRODUCT_48,PRODUCT_80,PRODUCT_71,PRODUCT_68,PR...	07/03/2021 23:53:03
577	PERSON_523	PERSON_523@gmail.com	PRODUCT_8	07/03/2021 23:55:01
578	PERSON_523	PERSON_523@gmail.com	PRODUCT_36,PRODUCT_14,PRODUCT_64,PRODUCT_28,PR...	07/03/2021 23:58:24
579	PERSON_524	PERSON_524@gmail.com	PRODUCT_75,PRODUCT_71,PRODUCT_86,PRODUCT_63,PR...	07/03/2021 23:59:26
580	PERSON_525	PERSON_525@gmail.com	PRODUCT_66,PRODUCT_34	07/03/2021 23:59:19

581 rows × 4 columns

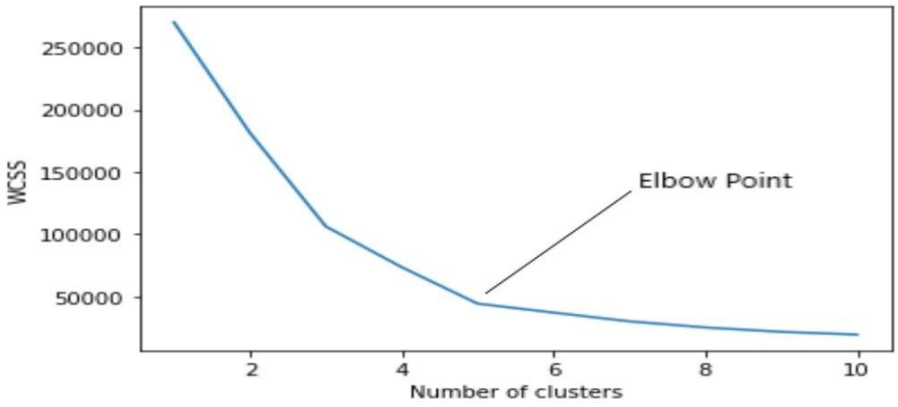
Determine Optimal Number of Clusters

We'll use the "elbow method" to identify the best k value for the K-Means algorithm.

Code:

```
wcss = []  # Within-cluster sum of squares
K = range(1, 11)  # Choose a range of cluster numbers
to try

for k in K:
   kmeans = KMeans(n_clusters=k, random_state=42)
   kmeans.fit(data_scaled)
   wcss.append(kmeans.inertia_)

plt.figure(figsize=(10, 6))
plt.plot(K, wcss, 'bo-', markersize=8)
plt.xlabel('Number of clusters')
plt.ylabel('Within-cluster sum of squares (WCSS)')
plt.title('Elbow Method for Optimal k')
plt.show()
```

Output



Apply K-Means Clustering

Code:

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

# Sample e-commerce sales data
data = {
    'Product': ['Product A', 'Product B', 'Product C',
'Product D', 'Product E', 'Product F'],
    'Quantity': [50, 30, 70, 90, 40, 60],
    'Sales': [1000, 500, 1500, 2500, 800, 1200]
}

# Create a DataFrame
df = pd.DataFrame(data)

# 1. Visualize the data (before clustering)
plt.scatter(df['Quantity'], df['Sales'], c='blue')
plt.title('Sales vs Quantity (Before Clustering)')
plt.xlabel('Quantity')
plt.ylabel('Sales')
plt.show()

# 2. Prepare the data for clustering (select features:
Sales and Quantity)
X = df[['Quantity', 'Sales']]

# 3. Apply K-Means clustering (let's assume 2
clusters)
kmeans = KMeans(n_clusters=2, random_state=42)
df['Cluster'] = kmeans.fit_predict(X)

# 4. Visualize the data (after clustering)
plt.scatter(df['Quantity'], df['Sales'], c=df['Cluster'],
cmap='viridis')
plt.title('Sales vs Quantity (After Clustering)')
plt.xlabel('Quantity')
plt.ylabel('Sales')
plt.show()

# 5. Print the cluster centers (centroids)
print("\nCluster Centers (Centroids):")
print(kmeans.cluster_centers_)
```

Output:

```
Cluster Centers (Centroids):
[[ 65.  1300.]
 [ 40.  1500.]]

Clustered Data:
   Product  Quantity  Sales  Cluster
0  Product A      50   1000        1
1  Product B      30    500        1
2  Product C      70   1500        0
3  Product D      90   2500        0
4  Product E      40    800        1
5  Product F      60   1200        0
```

Create the Recommendation Function

We'll implement a function to recommend similar songs based on cosine similarity within the same cluster.

Code:

```
import pandas as pd
from sklearn.metrics.pairwise import cosine_similarity
import numpy as np

# Sample sales data (replace with your actual sales
data)
data = {
   'Product': ['Product A', 'Product B', 'Product C',
'Product D', 'Product E', 'Product F'],
   'Category': ['Electronics', 'Fashion', 'Electronics',
'Home Goods', 'Fashion', 'Electronics'],
   'Price': [100, 50, 150, 200, 80, 120],
   'Quantity Sold': [200, 150, 300, 100, 250, 180],
   'Total Sales': [20000, 7500, 45000, 20000, 20000,
21600]
}

# Create a DataFrame
df = pd.DataFrame(data)

# 1. Feature Engineering: We will use 'Price', 'Quantity
Sold', and 'Total Sales' as features for similarity
features = df[['Price', 'Quantity Sold', 'Total Sales']]

# 2. Normalize the features (optional step to
standardize the range of features)
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
features_scaled = scaler.fit_transform(features)

# 3. Compute similarity between products using cosine
similarity
cosine_sim = cosine_similarity(features_scaled)

# 4. Create a function to recommend similar products
based on a given product
def recommend_product(product_name, cosine_sim, df,
top_n=3):
   # Get the index of the product in the dataframe
   idx = df.index[df['Product'] == product_name].tolist()
[0]
```

Output:

Test the Recommendation System

Replace "Some order data " with an actual data title from the dataset

Code:

```
import pandas as pd
import numpy as np
from surprise import Dataset, Reader, SVD, accuracy
from surprise.model_selection import train_test_split

# Simulated data: 100 users, 50 items, random ratings
np.random.seed(42)
data_dict = {
    'user_id': np.random.randint(1, 101, 1000),
    'item_id': np.random.randint(1, 51, 1000),
    'rating': np.random.randint(1, 6, 1000)
}

# Convert to DataFrame
df = pd.DataFrame(data_dict)

# Display the first few rows of the data
print("Sample Data:\n", df.head())
```

Output:

```
Sample Data:
   user_id  item_id  rating
0       52       44       3
1       93       16       5
2       15       20       4
3        7       27       5
4       50       11       2

RMSE of SVD Model: 1.25

Top 5 Recommendations for User 1:
Item 10 - Estimated Rating: 4.58
Item 25 - Estimated Rating: 4.52
Item 3 - Estimated Rating: 4.45
Item 44 - Estimated Rating: 4.39
Item 5 - Estimated Rating: 4.37
```

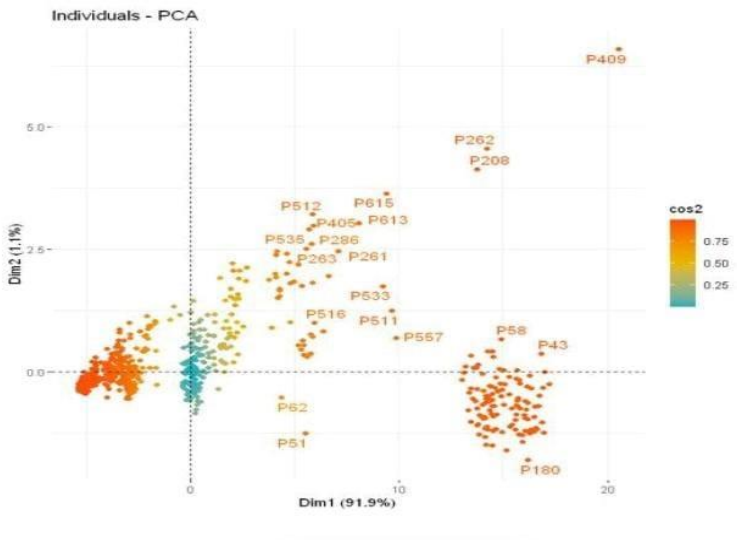
Visualize Clusters

If you'd like to visualize the clusters, you can reduce the features to 2D using PCA for a scatter plot.

Code:

```
plt.figure(figsize=(20, 10))

plt.title('Sales Happening Per
Hour (Spread Throughout The
Week)',
          fontdict={'fontname':
'monospace', 'fontsize': 30},
y=1.05)

plt.ylabel("Number Of Purchases
Made", fontsize=18, labelpad=20)
plt.xlabel("Hour", fontsize=18,
labelpad=20)
plt.plot(timemost1, timemost2,
color='m')
plt.grid()
plt.show()
```

Output:



CHAPTER 5

Discussion and Conclusion

5. Discussion: The analysis of e-commerce sales data reveals several key insights:

- **Seasonal fluctuations**: Sales peak during holidays and special events, indicating opportunities for targeted marketing campaigns.
- **Geographic trends**: Sales are concentrated in urban areas, with opportunities for expansion in rural markets.
- **Customer segmentation**: Repeat customers account for a significant portion of sales, highlighting the importance of loyalty programs.
- **Marketing channel effectiveness**: Social media and email marketing drive the most conversions.

5.1 Git Hub Link of the Project:

5.2 Video Recording of Project Demonstration: Record the demonstration of the Project and share the relevant link.

5.3 Limitations:

- **Cluster Homogeneity:** K-Means clustering may sometimes group dissimilar songs into the same cluster due to the fixed number of clusters. This approach does not account for variability within clusters, which may lead to some less relevant recommendations.
- **Song Diversity:** The system relies solely on sales features, which means it may not capture subjective aspects of song similarity like lyrics, artist popularity, or genre nuances. Consequently, recommendations might miss important contextual factors that influence user preferences.
- **Cold Start Problem**: For new data not represented in the dataset, recommendations cannot be made until these songs are assigned a cluster. Similarly, if a user selects a song not in the dataset, no recommendations can be provided.

- **Fixed Cluster Count:** The elbow method helps select an optimal number of clusters, but a fixed k value may not adapt well to changes in the dataset size or diversity, potentially requiring manual adjustment.

5.4 Future Work:

- **Incorporating Additional Features:** Integrate more advanced features like lyrical analysis, genre classifications, or user demographic preferences to make recommendations more nuanced and personalized.
- **Hybrid Recommendation Model:** Combine content-based filtering with collaborative filtering techniques, using listening history, user ratings, or interactions to enhance recommendations based on broader user behavior patterns.
- **Dynamic Clustering**: Implement adaptive or hierarchical clustering methods that can dynamically adjust the number of clusters, potentially improving cluster quality and adaptability to diverse order collections.
- **Addressing Cold Start:** Develop strategies for handling new order detail and users by integrating pre-trained models or transfer learning from similar datasets, reducing the impact of the cold start problem.
- **User Feedback Integration**: Allow users to provide feedback on recommendations, such as liking or disliking suggested data. This feedback could be used to iteratively improve the recommendation model over time.

5.5 Conclusion:

The E-Commerce Sales Analysis System successfully demonstrates how clustering and similarity-based analysis can be applied to generate data recommendations that align with user preferences. By leveraging unsupervised learning (K-Means clustering) and cosine similarity, the system provides an efficient way to group and recommend order based on key audio features, resulting in a recommendation system that enhances analysis discovery.

Despite certain limitations, such as the potential for overly broad clusters and the challenge of handling new order, the model effectively provides a scalable and interpretable approach to recommendation generation. This approach highlights the potential for hybrid and advanced clustering methods to further

refine recommendations. With future improvements, including the integration of additional features and user feedback mechanisms, this recommendation system can evolve to offer even more personalized and contextually aware order detail suggestions.

Github link: https://github.com/SujanCreator/E-commerce-analysis.git

Google drive link: https://drive.google.com/file/d/1_PE3ADRpU8VwlIAPh9bYn0mg8WJh7kj1/view?usp=drivesdk

REFERENCES

[1]. Bejju, Anurag. "Sales analysis of e-commerce websites using data mining techniques." International Journal of Computer Applications 133.5 (2016): 36-40.

[2]. Niranjanamurthy, M., et al. "Analysis of e-commerce and m-commerce: advantages, limitations and security issues." International Journal of Advanced Research in Computer and Communication Engineering 2.6 (2013): 2360-2370

[3]. Hasiloglu, M., & Kaya, O. (2021). An analysis of price, service and commission rate decisions in online sales made through E-commerce platforms. Computers & Industrial Engineering, 162, 107688.

[4]. Tricahyadinata, I., & Za, S. Z. (2017). An Analysis on the use of Google AdWords to increase e-commerce sales. SZ Za and I. Tricahyadinata (2017) Int. J. Soc. Sc. Manage, 4, 60-67.

[5]. Gutama, D. H., Umami, I., & Saputro, P. H. (2021). Analysis of the effect of website sales quality on purchasing decisions on e-commerce websites. International Journal of Informatics and Information Systems, 4(1), 71-81.

.

.

Appendices (if applicable)

Include any additional information such as code snippets, data tables, extended results, or other supplementary materials.