Project Report

1. INTRODUCTION

1.1 Project Overview

Wholesale customer segmentation analysis divides wholesale customers into groups based on their shared characteristics, such as product spending habits, demographics, and purchase history. This analysis can identify customer segments, understand their needs and preferences, and develop targeted marketing and sales strategies. This project aims to analyse the spending behaviour of wholesale customers and identify growth opportunities.

1.2 Purpose

The objective of this project is to perform a wholesale customer segmentation analysis using data on customer spending habits for different product categories. The goal is to identify different customer segments based on their spending patterns and develop insights into their needs and preferences.

The purpose of the wholesale customer segmentation analysis project is to help the wholesale distributor better understand its customers and develop more effective marketing and sales strategies. By segmenting customers based on their spending patterns, the distributor can identify their needs and preferences, and develop targeted marketing and sales strategies that are more likely to be successful.

Specifically, the project will help the distributor to:

- Identify customer segments with high growth potential. This will allow the distributor to focus its marketing and sales efforts on these segments and generate more revenue.
- Understand the needs of different customer segments. This will allow the distributor to develop products and services that meet the specific needs of each segment.
- Improve customer retention. By understanding the needs and preferences of its customers, the distributor can develop targeted retention programs that are more likely to keep customers coming back.
- Increase customer satisfaction. By offering personalized products, services, and experiences, the distributor can improve customer satisfaction and loyalty.

2. LITERATURE SURVEY

2.1 Existing problem

The literature review unfolds a dynamic panorama within Wholesale Banking (WB) and the proactive integration of Early Warning Systems (EWS) for financial distress detection. ING's Advanced Risk Integrated Application (ARIA) is noteworthy for its dual role in identifying ongoing negative changes and exploring up-selling opportunities.

The incorporation of Customer Segmentation (CS) models, employing unsupervised methods like K-Means and DBSCAN, signifies a progressive understanding of customer behavior. Analyzing aspects such as tightness, separation, and risk-reward within clusters contributes to a nuanced comprehension of WB clients' financial health.

Shifting focus to the Brazilian supermarket segment, the review highlights escalating investments in Customer Relationship Management (CRM) strategies. The study on Nordest~ao, utilizing the recency-frequency-monetary model and Gaussian mixture models, aligns with the global trend of augmenting customer understanding. Although the use of KNN algorithm in ML for customer segmentation is not explicitly mentioned in the provided text, it could be explored as a potential approach to enhance the study's analytical depth.

The exploration extends to the wholesale sector, underscoring the objective of creating a machine learning model for anticipatory product demand. While the text discusses the use of SMOreg and Kstar algorithms, potential inclusion of KNN in similar studies could amplify the discussion. This section offers insights into the evolving landscape of wholesale practices and sets the stage for a more tailored exploration of the KNN algorithm's role in Wholesale Customer Segmentation.

In conclusion, the literature review provides a comprehensive overview of the evolving trends in Wholesale Banking, supermarket chains, and the wholesale sector. To align with the project title, emphasis on the application of the KNN algorithm in ML for Wholesale Customer Segmentation is recommended.

This tailored perspective not only addresses the specific focus of the project but also enriches the discourse on the dynamic interplay between machine learning algorithms and wholesale customer analysis.

2.2 References

https://ieeexplore.ieee.org/abstract/document/8275249
https://ieeexplore.ieee.org/abstract/document/9885315
https://ieeexplore.ieee.org/abstract/document/9666609

2.3 Problem Statement Definition

Wholesale distributors often have a large and diverse customer base. This can make it difficult to understand the needs and preferences of each customer segment. As a result, distributors may not be able to develop effective marketing and sales strategies. If distributors do not understand their customer segments, they may not be able to develop effective marketing and sales campaigns. This can lead to wasted marketing spend and lost sales.

The wholesale customer segmentation analysis project aims to address this problem by developing a model that can be used to identify and understand different customer segments. The insights gained from this analysis will be used to develop more targeted marketing and sales strategies that can help distributors to improve their business performance.
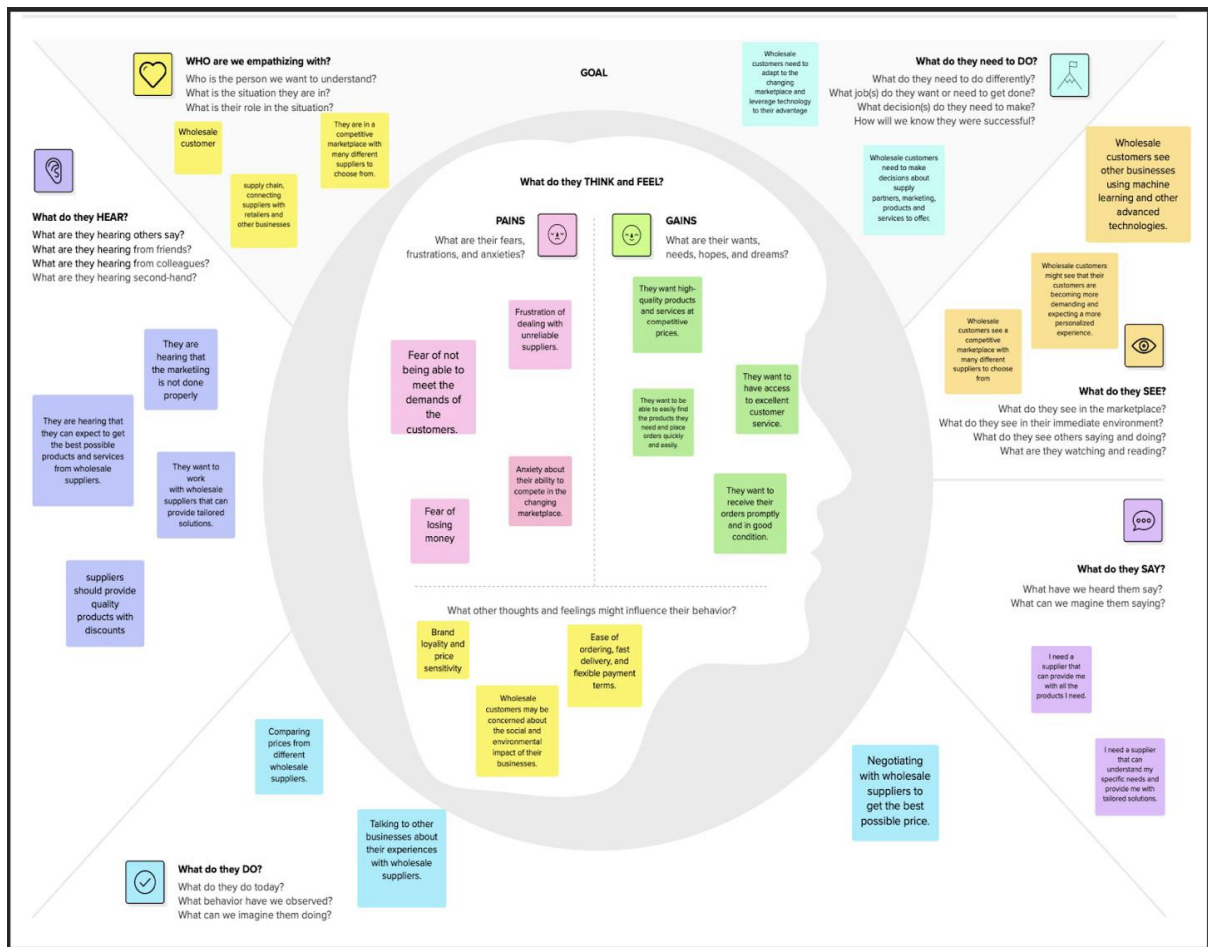
Key aspects of the problem statement:

- Ineffective marketing and sales campaigns: If distributors do not understand their customer segments, they may not be able to develop effective marketing and sales campaigns. This can lead to wasted marketing spend and lost sales.
- Low customer satisfaction and loyalty: If distributors do not understand the needs and preferences of their customers, they may not be able to provide them with the products and services that they want. This can lead to low customer satisfaction and loyalty.
- Poor decision-making: If distributors do not understand how different customer segments behave, they may not be able to make good decisions about product mix, pricing, and other aspects of the business. This can lead to suboptimal profitability.
- A distributor may send out a marketing campaign for a new product to all of its customers, but the campaign may be irrelevant to many of the customers. This can lead to wasted marketing spend and low response rates.
- A distributor may offer a discount to all of its customers, but the discount may not be attractive to all customer segments. This can lead to low redemption rates and lost sales.
- A distributor may develop a new product that is not relevant to the needs of any of its customer segments. This can lead to the product being unsuccessful and a waste of resources.

The wholesale customer segmentation analysis project is a valuable initiative that can help distributors to overcome these challenges and achieve their business goals.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

Empathy Map

WHO are we empathizing with?
Who is the person we want to understand?
What is the situation they are in?
What is their role in the situation?

- Wholesale customer
- They are in a competitive marketplace with many different suppliers to choose from.
- supply chain, connecting suppliers with retailers and other businesses

GOAL
- Wholesale customers need to adapt to the changing marketplace and leverage technology to their advantage

What do they need to DO?
What do they need to do differently?
What job(s) do they want or need to get done?
What decision(s) do they need to make?
How will we know they were successful?

- Wholesale customers need to make decisions about supply partners, marketing, products and services to offer.
- Wholesale customers see other businesses using machine learning and other advanced technologies.
- Wholesale customers might see that their customers are becoming more demanding and expecting a more personalized experience.
- Wholesale customers see a competitive marketplace with many different suppliers to choose from

What do they HEAR?
What are they hearing others say?
What are they hearing from friends?
What are they hearing from colleagues?
What are they hearing second-hand?

- They are hearing that the marketing is not done properly
- They are hearing that they can expect to get the best possible products and services from wholesale suppliers.
- They want to work with wholesale suppliers that can provide tailored solutions.
- suppliers should provide quality products with discounts

What do they THINK and FEEL?

PAINS
What are their fears, frustrations, and anxieties?

- Frustration of dealing with unreliable suppliers.
- Fear of not being able to meet the demands of the customers.
- Anxiety about their ability to compete in the changing marketplace.
- Fear of losing money

GAINS
What are their wants, needs, hopes, and dreams?

- They want high-quality products and services at competitive prices.
- They want to be able to easily find the products they need and place orders quickly and easily.
- They want to have access to excellent customer service.
- They want to receive their orders promptly and in good condition.

What other thoughts and feelings might influence their behavior?

- Brand loyalty and price sensitivity
- Ease of ordering, fast delivery, and flexible payment terms.
- Wholesale customers may be concerned about the social and environmental impact of their businesses.
- Negotiating with wholesale suppliers to get the best possible price.

What do they SEE?
What do they see in the marketplace?
What do they see in their immediate environment?
What do they see others saying and doing?
What are they watching and reading?

What do they SAY?
What have we heard them say?
What can we imagine them saying?

- I need a supplier that can provide me with all the products I need.
- I need a supplier that can understand my specific needs and provide me with tailored solutions.

What do they DO?
What do they do today?
What behavior have we observed?
What can we imagine them doing?

- Comparing prices from different wholesale suppliers.
- Talking to other businesses about their experiences with wholesale suppliers.

3.2 Ideation & Brainstorming

Brainstorm
& idea prioritization

Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.

- 🕐 **10 minutes** to prepare
- ⏳ **1 hour** to collaborate
- 👤 **2-8 people** recommended

Need some inspiration?

See a finished version of this template to kickstart your work.

Open example →

Before you collaborate

A little bit of preparation goes a long way with this session. Here's what you need to do to get going.

🕐 10 minutes

A **Team gathering**
Akhil Ajithkumar
Shreeya Vemula
Kapilavai Krishna Priiya Abhigna
Kabala Devi Rishitha

B **Set the goal**
Wholesale customer segmentation analysis

C **Learn how to use the facilitation tools**
Use the Facilitation Superpowers to run a happy and productive session.

Open article →

---

1 Define your problem statement

What problem are you trying to solve? Frame your problem as a How Might We statement. This will be the focus of your brainstorm.

🕐 5 minutes

Who:
Wholesale customers

What:
Loss in business due to ineffective marketing strategies and product offerings.

When:
They go to market to buy products

Where:
When they go to market to buy products

Why:
We need to fix this to identify customer spending behaviours to maximize profit for the business.

PROBLEM

How might we enable wholesale businesses to understand their customers and their product needs in a highly competitive marketplace in order to develop products

Brainstorm

Write down any ideas that come to mind that address your problem statement.

🕐 10 minutes

Person 1	Person 2	Person 3	Person 4
Churn Prediction: Develop models to predict customer churn based on their spending behavior. By identifying potential churners, you can implement retention strategies.	Implement regular customer surveys and feedback mechanisms to gather insights into customer preferences, pain points, and product needs. Use this data to inform product development.	A wholesale distributor of food products could use customer data to identify the most popular products in each geographic region.	Create interactive dashboards that provide real-time insights into spending behavior, customer segments, and the effectiveness of marketing efforts.
Time Series Analysis: Analyze time-series data to identify seasonal spending patterns and trends. This can help businesses prepare for peak seasons and optimize inventory.	Segment your customer base into different categories based on purchasing behavior, demographics, and preferences. Tailor your products to meet the specific needs of each segment.	A wholesale distributor of food products partnered with a local restaurant to develop a new line of pre-made sauces.	Implement advanced inventory management systems that track product performance and can automatically reorder or adjust stock based on customer demand.
Anomaly Detection: Detect unusual spending patterns or outliers in the data. Anomalies might indicate fraudulent activity or unusual customer behavior that requires attention.	Wholesale businesses can conduct market research to understand the competitive landscape, identify new market opportunities, and learn more about customer needs.	Attend a trade show for small businesses who are struggling to find affordable and high quality office supplies	The distributor learned that many customers were looking for parts that were easy to install. The distributor then developed a new line of parts that were specifically designed for easy installation.

Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. Once all sticky notes have been grouped, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you and break it up into smaller sub-groups.

🕐 20 minutes

Implement regular customer surveys and feedback mechanisms to gather insights into customer preferences, pain points, and product needs. Use this data to inform product development.

Churn Prediction: Develop models to predict customer churn based on their spending behavior. By identifying potential churners, you can implement retention strategies.

A wholesale distributor of food products partnered with a local restaurant to develop a new line of pre-made sauces.

Time Series Analysis: Analyze time-series data to identify seasonal spending patterns and trends. This can help businesses prepare for peak seasons and optimize inventory.

Segment your customer base into different categories based on purchasing behavior, demographics, and preferences. Tailor your products to meet the specific needs of each segment.

4

Prioritize

Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.

🕐 **20 minutes**

Importance

If each of these tasks could get done without any difficulty or cost, which would have the most positive impact?

+

Time Series Analysis: Analyze time-series data to identify seasonal spending patterns and trends. This can help businesses prepare for peak seasons and optimize inventory.

Churn Prediction: Develop models to predict customer churn based on their spending behavior. By identifying potential churners, you can implement retention strategies.

Segment your customer base into different categories based on purchasing behavior, demographics, and preferences. Tailor your products to meet the specific needs of each segment.

Implement regular customer surveys and feedback mechanisms to gather insights into customer preferences, pain points, and product needs. Use this data to inform product development.

A wholesale distributor of food products partnered with a local restaurant to develop a new line of pre-made sauces.

-

Feasibility

-   **+**

Regardless of their importance, which tasks are more feasible than others? (Cost, time, effort, complexity, etc.)

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

The wholesale customer segmentation analysis project aims to develop a model that can be used to identify and understand different customer segments. The insights gained from this analysis will be used to develop more targeted marketing and sales strategies that can help distributors to improve their business performance.

Requirements:

The following requirements must be met for the wholesale customer segmentation analysis project to be successful:

- Data: The project must have access to high-quality data on customer demographics, purchase history, and spending habits. This data will be used to develop the segmentation model and identify the needs and preferences of different customer segments.
- Segmentation algorithm: The project must use a robust and effective segmentation algorithm to group customers into different segments. The algorithm should be able to take into account a variety of factors, such as customer demographics, purchase history, and spending habits.
- Customer segmentation analysis: The project must develop a methodology for analyzing the customer segments and identifying their needs and preferences. This analysis should be based on the data collected and the segmentation algorithm used.
- Actionable insights: The project must develop actionable insights that can be used to develop targeted marketing and sales strategies for each customer segment. These insights should be specific and actionable, and they should be based on the customer segmentation analysis.

Acceptance Criteria:

The following acceptance criteria must be met for the wholesale customer segmentation analysis project to be considered successful:

- The segmentation model must be able to accurately group customers into different segments. This can be measured by calculating the accuracy of the model on a held-out test set.
- The customer segmentation analysis must identify the needs and preferences of each customer segment. This can be measured by comparing the customer segments to each other and identifying the key differences between them.
- The actionable insights developed from the customer segmentation analysis must be relevant and useful for developing targeted marketing and sales strategies. This can be measured by surveying customers in each segment to see if the insights are helpful and relevant to their needs.

Conclusion:

By meeting the requirements and acceptance criteria outlined above, the wholesale customer segmentation analysis project can be successful in developing a model that can be used to identify and understand different customer segments. The insights gained from this analysis can then be used to develop more targeted marketing and sales strategies that can help distributors to improve their business performance.

4.2 Non-Functional requirements

Non-functional requirements for a Wholesale Customer Segmentation Analysis typically refer to the criteria that describe how the system or analysis should perform, rather than the specific functions it should carry out. These requirements are often related to performance, security, usability, scalability, and other qualities. Here are some non-functional requirements that might be relevant to Wholesale Customer Segmentation Analysis:

Performance:

Response Time: Define the maximum acceptable time for the system to respond to segmentation queries or data processing.

Throughput: Specify the number of customer segments or data points that the system should be able to handle within a given time frame.

Scalability:

Data Scalability: Define how the system should handle an increase in the volume of customer data without compromising performance.

User Scalability: Specify how many users or concurrent sessions the system should support.

Security:

Data Privacy and Protection: Define measures to ensure the privacy and security of customer data throughout the segmentation process.

Access Control: Determine who has access to specific segments or features within the analysis tool.

Reliability:

System Availability: Specify the percentage of time the system should be available for use.

Fault Tolerance: Define how the system should handle and recover from errors or failures during the segmentation process.

Usability:

User Interface: Define ease of use, clarity, and intuitiveness of the user interface for performing segmentation analysis.

Training and Documentation: Specify the need for training resources and comprehensive documentation for users.

Interoperability:

Integration with Other Systems: Define the system's ability to work with other tools or systems that may provide relevant data or analytics.

Compliance:

Regulatory Compliance: Ensure that the system complies with relevant regulations, such as GDPR, HIPAA, or industry-specific standards for customer data handling.

Maintainability:

Upgradability: Define how easily the system can be updated with new features or changes in the segmentation process.

System Documentation: Ensure the availability of comprehensive documentation for system maintenance and updates.

Reporting and Analysis:

Reporting Accuracy: Specify the level of accuracy required in the segmentation results and subsequent reporting.

Visualization: Define the types of visualizations and reporting formats necessary for effective interpretation and communication of results.

These non-functional requirements are crucial to ensure that the Wholesale Customer Segmentation Analysis system not only performs its intended functions but does so in a way that meets broader criteria for performance, security, usability, and scalability. Adjust and tailor these requirements based on the specific needs and context of your wholesale customer segmentation analysis project.

5. PROJECT DESIGN

5.1 Data Flow Diagrams & User Stories

Data Collection: This phase entails gathering a diverse dataset of products that are bought by customers from sources such as customer surveys and interviews, social media monitoring, or other means. The collected data are then stored in a raw data repository.

Data Cleaning: Clean the data by removing duplicates, missing values, and outliers. This step ensures that the data is accurate and reliable.

Data Transformation: Transform the data into a suitable format for clustering analysis. This step involves converting the data into a numerical format that can be used for clustering.

Feature Selection: Select the most relevant features that can help in clustering. This step involves identifying the features that are most important for clustering and discarding the rest.

Clustering Algorithm Selection: Choose an appropriate clustering algorithm based on the type of data and business requirements. There are several clustering algorithms available, such as K-means, Hierarchical Clustering, and DBSCAN.

Model Training: Train the clustering model on the selected features. This step involves running the selected clustering algorithm on the data to create clusters.

Model Evaluation: Evaluate the model's performance using metrics such as silhouette score, elbow method, and gap statistic. This step helps in determining the optimal number of clusters and assessing the quality of the clusters.

Cluster Interpretation: Interpret the clusters by analyzing the characteristics of each cluster. This step involves identifying the common characteristics of each cluster and understanding what makes them unique.

Segmentation Strategy Development: Develop a segmentation strategy based on the insights gained from cluster interpretation. This step involves using the insights gained from cluster interpretation to develop a segmentation strategy that can be used to target specific customer groups.
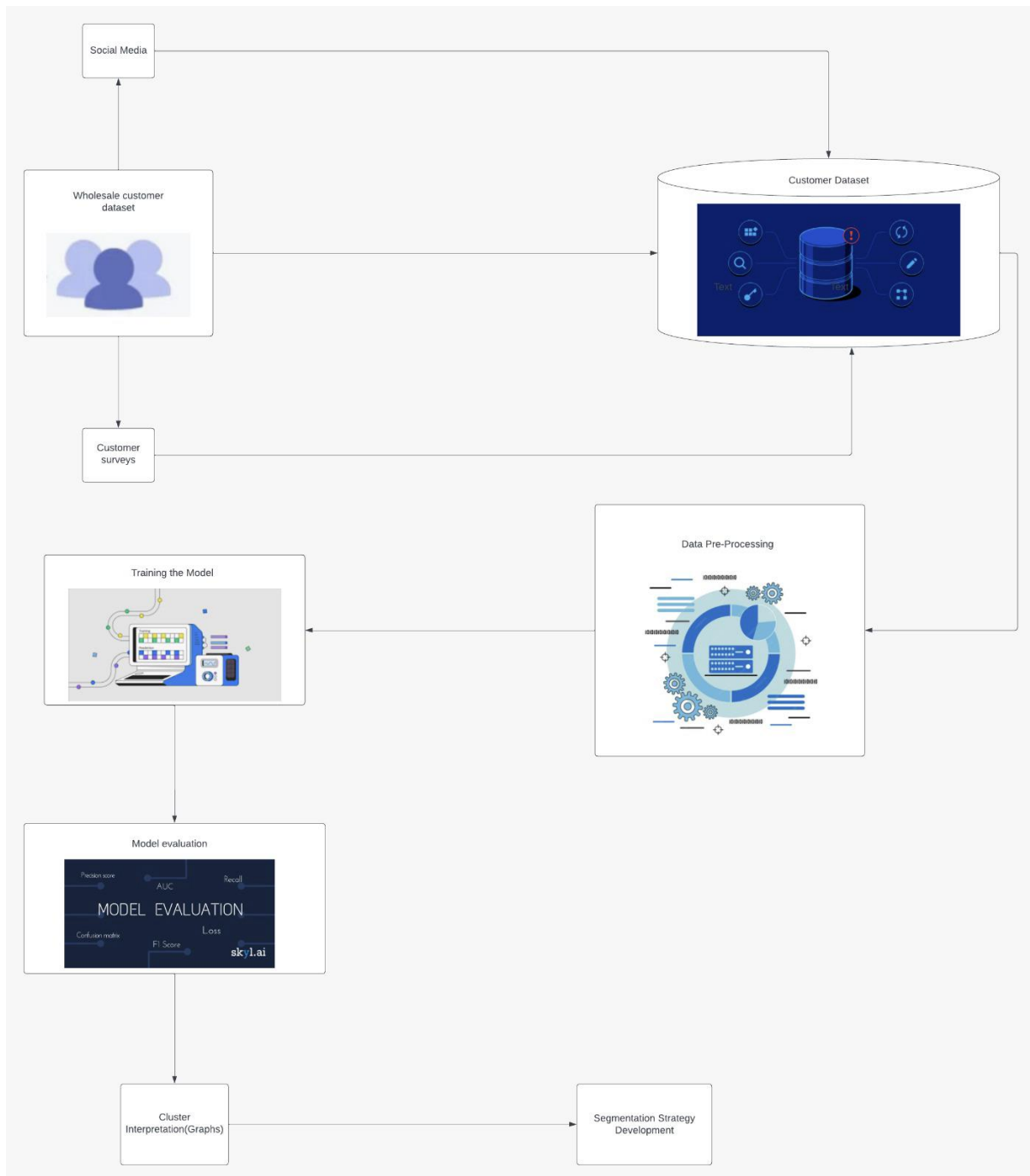
5.2 Solution Architecture

Our wholesale customer segmentation analysis solution employs a sophisticated architecture, drawing upon machine learning primarily using k neighbors classification, SVM, and Naïve Bayes. This application is designed to significantly enhance wholesale businesses. That will help them to understand their customers and their product needs in a highly competitive marketplace in order to develop products.

Key Components:

1. Machine Learning Models: The core of the architecture is the machine learning techniques, which are trained on extensive datasets comprising customer segmentation, product recommendation, and demand forecasting. K-means clustering, hierarchical clustering, and decision trees can be used to segment customers into different groups based on their purchase history, demographics, and other factors.

2. Supervised Learning: One of the distinctive features of our architecture is that supervised learning is a type of machine learning where the model is trained on a set of labeled data.

The labeled data consists of input data and the corresponding output data. The model learns to predict customer churn based on customer data and accurately produces the customer's needs and we can provide insights to the producer. And mainly to identify hidden patterns and trends in customer data

3. Unsupervised learning: Clustering algorithms are unsupervised learning algorithms, which means that they do not require labeled data to train. This is important for customer segmentation analysis, as it can be difficult and expensive to label customer data.

4. This information can then be used to segment customers into different groups based on their shared characteristics and behaviors. Improve customer understanding. By segmenting customers into different groups, businesses can better understand the needs and wants of each customer group.
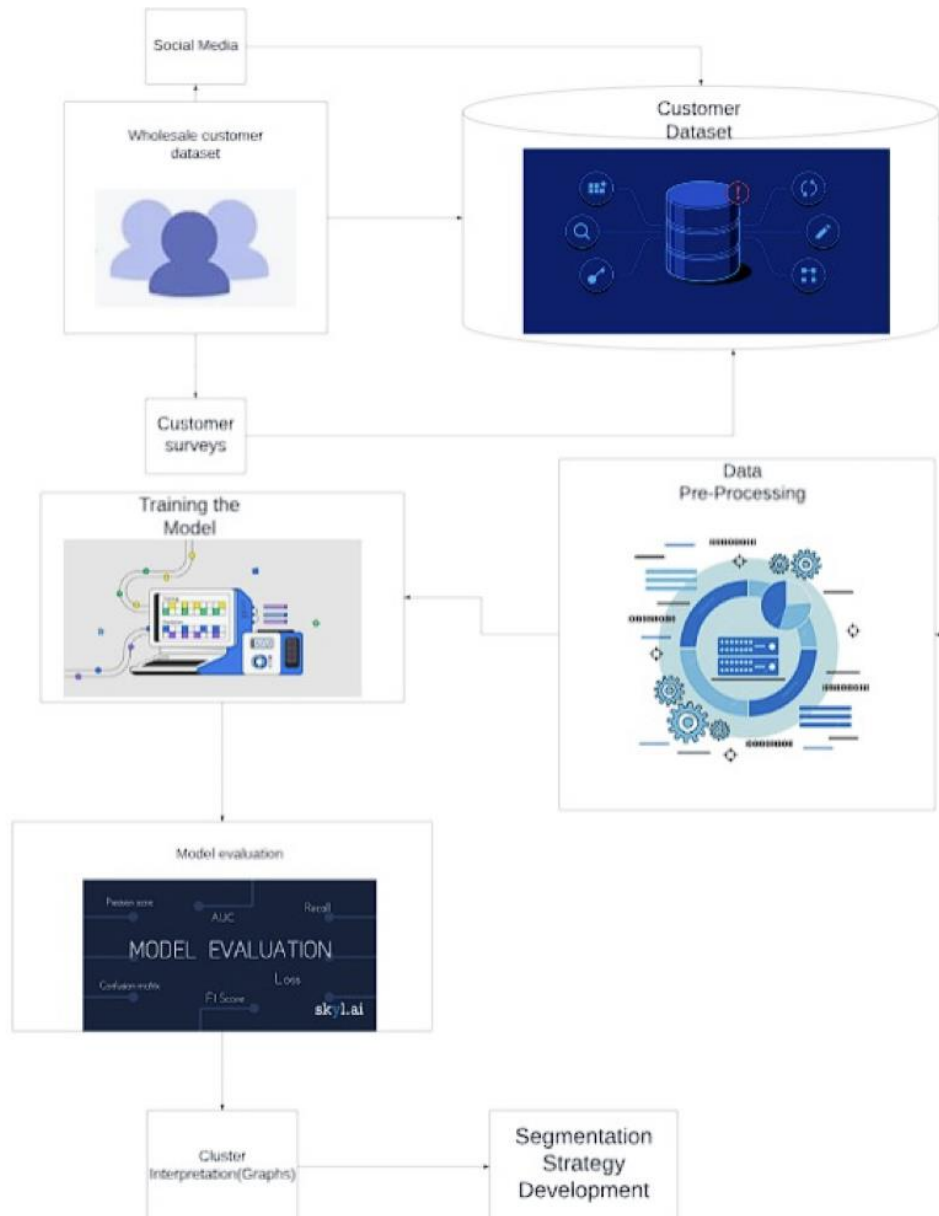
6. PROJECT PLANNING & SCHEDULING
6.1 Technical Architecture

Prediction Train Data Data CSV Evaluation UI Data Preprocessing Model Algorithm Test Data Inputs User
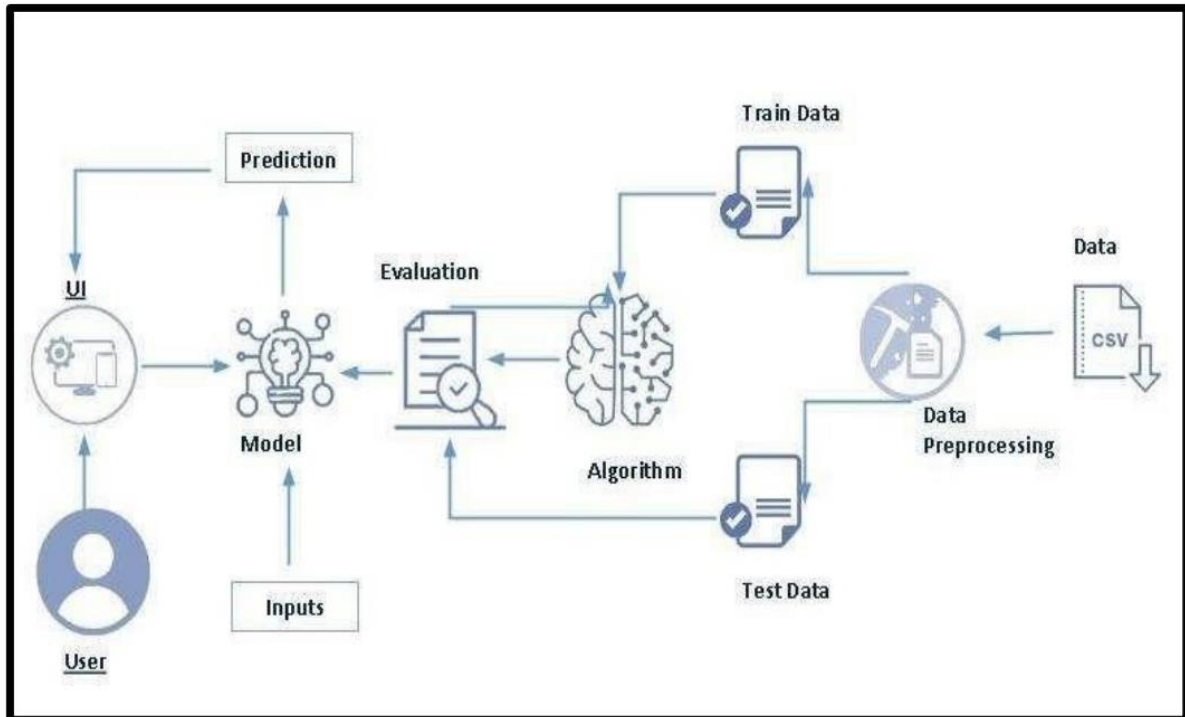
Table-1 : Components & Technologies:

S.No	Component	Description	Technology
1.	Application Logic-1	Logic for a process in the application	Python
2.	File Storage	File storage requirements for Storing the dataset	Local Filesystem
3.	Machine Learning Model	Purpose of Model	SVM, Naive-Bayes, K-means, Clustering, etc.

Table-2: Application Characteristics:

S.No	Characteristics	Description	Technology
1.	Open-Source Frameworks	List the open-source frameworks used	Google Colab
2.	Scalable Architecture	Justify the scalability of architecture (3 – tier, Micro-services)	Model can be scaled according to dataset.

6.2 Sprint Planning & Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Project setup & Infrastructure	USN-1	Set up the development environment with the required tools and frameworks to start the Wholesale Customer Segmentation Analysis	2	High	Shreeya
Sprint-1	Data collection	USN-2	Gather a diverse dataset based on the Wholesale Customer spending behaviours.	2	Medium	Shreeya
Sprint-2	Data Preprocessing	USN-3	Preprocess the collected dataset by finding the missing values, normalizing the dataset, and data visualization and split it into training and validation sets.	5	High	Devi
Sprint-3	Model development	USN-4	Explore and evaluate different machine learning models(e.g., clustering) to segment the wholesale customers.	5	High	Krishna Priya
Sprint-4	Training	USN-5	Train the selected machine learning model using the preprocessed dataset.	3	High	Akhil
Sprint-4	Model Evaluation	USN-6	Evaluate the machine learning models performance on the validation dataset.	2	Medium	Akhil

6.3 Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	4	3 Days	26 Oct 2023	28 Oct 2023	4	29 Oct 2023
Sprint-2	5	3 Days	29 Oct 2023	31 Oct 2023	5	01 Nov 2023
Sprint-3	5	3 Days	01 Nov 2023	03 Nov 2023	5	04 Nov 2023
Sprint-4	5	2 Days	04 Nov 2023	05 Nov 2023	5	06 Nov 2023

7. CODING & SOLUTIONING Data collection :

ML depends heavily on data, It is the most crucial aspect that makes algorithm training possible. So this section allows you to download the required dataset.

Activity 1:

Download the dataset There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc. In this project, we have used Wholesale Customer Segmentation data. This data is downloaded from kaggle.com. Please refer to the link given below to download the dataset.

Exploratory data analysis:

Univariate analysis :

In simple words, univariate analysis is understanding the data with single feature. Here we have displayed two different graphs such as distplot and countplot.

● Seaborn package provides a wonderful function distplot. With the help of distplot, we can find the distribution of the feature. To make multiple graphs in a single plot, we use subplot

```
Channel_df=df['Channel'].value_counts().reset_index()
Channel_df.rename(columns={'index': 'Channel_type'}, inplace=True)
Channel_df.rename(columns={'Channel': 'Count'}, inplace=True)
Channel_df.head()
```

	Channel_type	Count
0	1	298
1	2	142

```
plt.figure(figsize=(15,5))
plt.title('Distribution of Channel_type')
sns.barplot(x='Channel_type',y='Count',data=Channel_df)
```

```
<Axes: title={'center': 'Distribution of Channel_type'}, xlabel='Channel_type', ylabel='Count'>
```
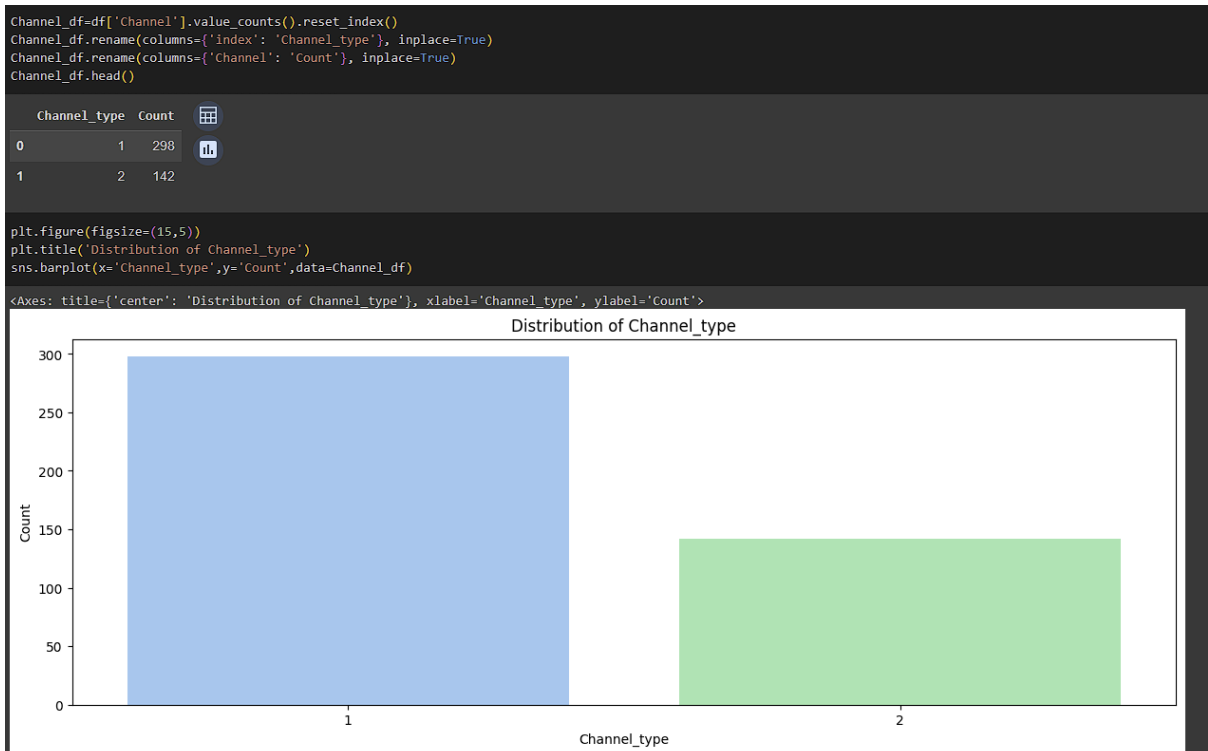
```
#Boxplot of each column
plt.figure(figsize=(15,10))
sns.boxplot(x='variable', y='value', data=df.melt())

plt.show()
```
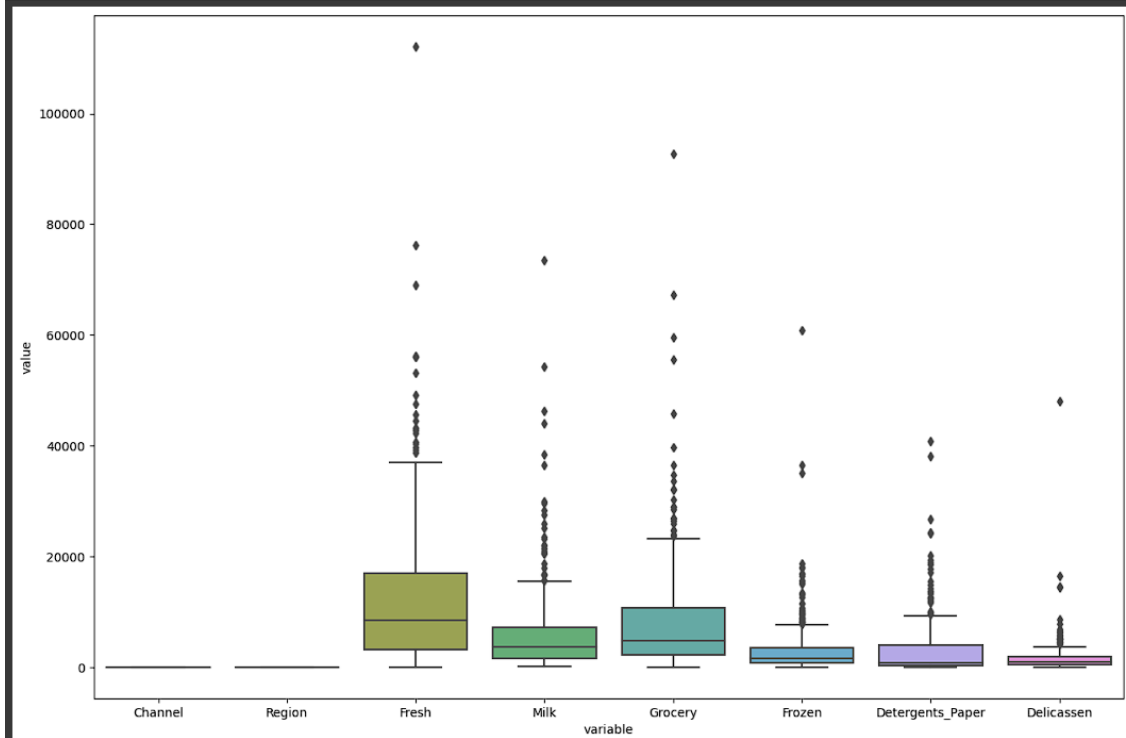
Bivariate analysis

To find the relation between two features we use bivariate analysis.

```python
def categorical_data(i):
    df[i].value_counts().plot(kind='bar')

j_1 = ['Channel','Region']

for k in j_1:
    categorical_data(i=k)
    plt.show()
```
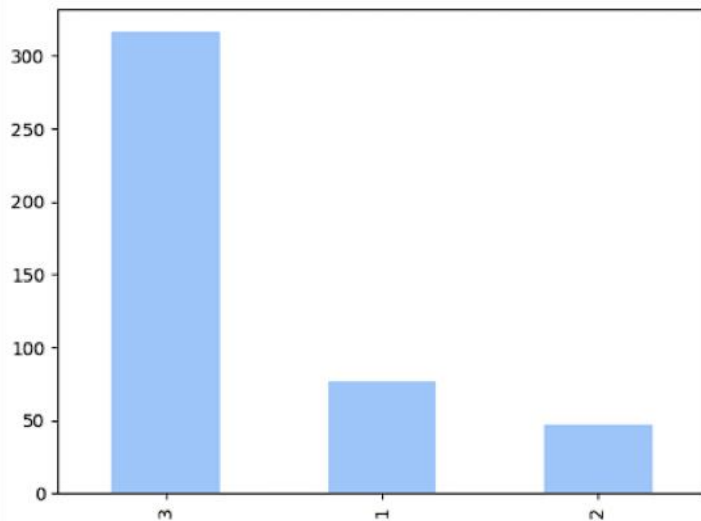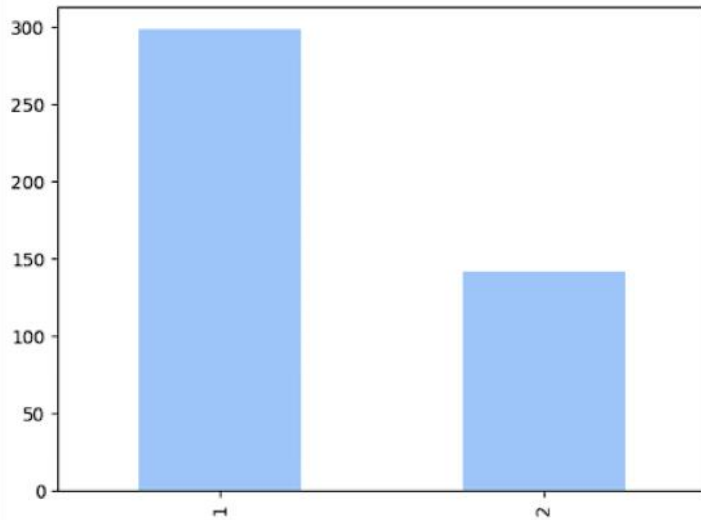


```python
df.corr()
```

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Channel	1.000000	0.062028	-0.169172	0.460720	0.608792	-0.202046	0.636026	0.056011
Region	0.062028	1.000000	0.055287	0.032288	0.007696	-0.021044	-0.001483	0.045212
Fresh	-0.169172	0.055287	1.000000	0.100510	-0.011854	0.345881	-0.101953	0.244690
Milk	0.460720	0.032288	0.100510	1.000000	0.728335	0.123994	0.661816	0.406368
Grocery	0.608792	0.007696	-0.011854	0.728335	1.000000	-0.040193	0.924641	0.205497
Frozen	-0.202046	-0.021044	0.345881	0.123994	-0.040193	1.000000	-0.131525	0.390947
Detergents_Paper	0.636026	-0.001483	-0.101953	0.661816	0.924641	-0.131525	1.000000	0.069291
Delicassen	0.056011	0.045212	0.244690	0.406368	0.205497	0.390947	0.069291	1.000000

```
print('Correlation Heat map of the data')
plt.figure(figsize=(10,6))
sns.heatmap(df.corr(),annot=True,fmt='.2f',vmin=-1,vmax=1,cmap='Spectral')
plt.show()
```

Correlation Heat map of the data
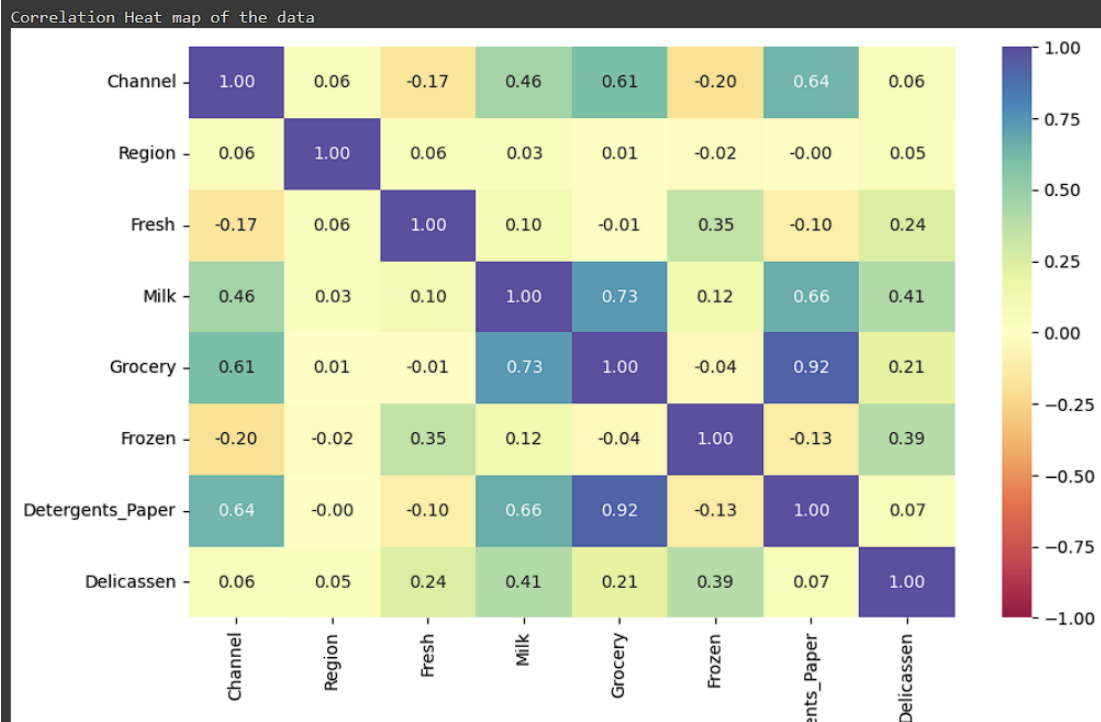
Multivariate analysis

In simple words, multivariate analysis is to find the relation between multiple features.
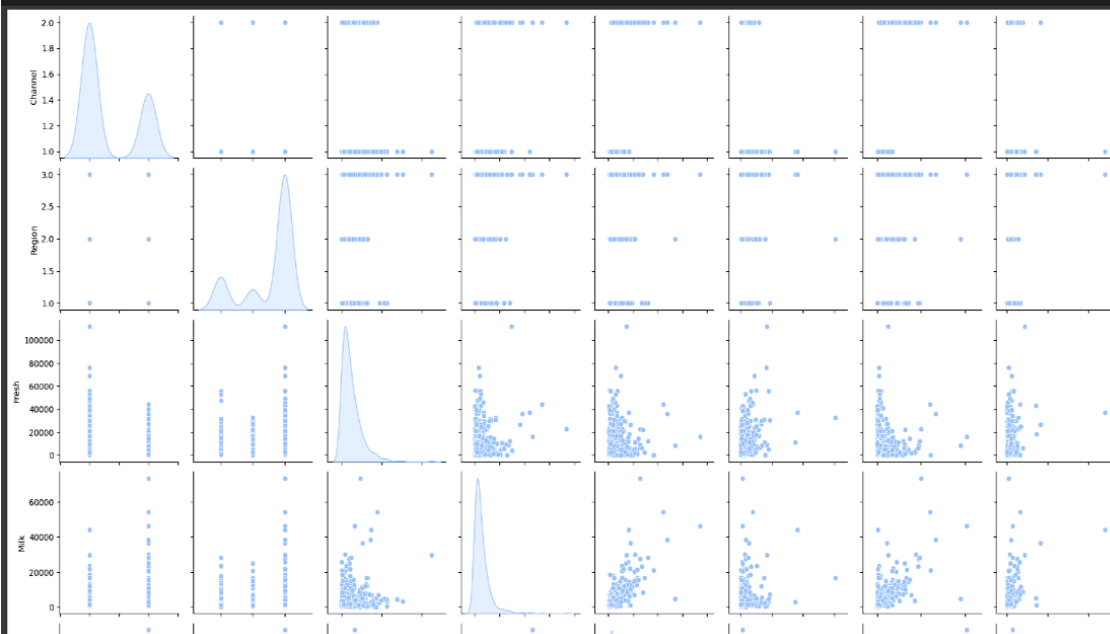
Descriptive analysis

Descriptive analysis is to study the basic features of data with the statistical process. Here pandas has a worthy function called describe. With this describe function we can understand the unique, top and frequent values of categorical features. And we can find mean, std, min, max and percentile values of continuous features.

Descriptive Analysis

```
[173] df.describe()
```

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	1.322727	2.543182	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	0.468052	0.774272	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	1.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	1.000000	2.000000	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	1.000000	3.000000	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	2.000000	3.000000	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	2.000000	3.000000	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

Data Pre-processing

As we have understood how the data is lets pre-process the collected data.

The download data set is not suitable for training the machine learning model as it might have so much randomness so we need to clean the dataset properly in order to fetch good results. This activity includes the following steps.

- Handling missing values

Handling Missing Values

```
#Checking for missing values
df.isnull().sum()
```

```
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicassen          0
dtype: int64
```

As we can see, there are no missing values.

- Handling categorical data

```
[ ] df.head()
```

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	2	3	12669	9656	7561	214	2674	1338
1	2	3	7057	9810	9568	1762	3293	1776
2	2	3	6353	8808	7684	2405	3516	7844
3	1	3	13265	1196	4221	6404	507	1788
4	2	3	22615	5410	7198	3915	1777	5185

- Handling outliers
- Scaling Techniques
- Splitting dataset into training and test set

```
[ ] X = df.drop('Channel', axis=1)
    Y = df['Channel']

    #using 5 folds split the data.
    skf = KFold(n_splits=5)
    for train_index,test_index in skf.split(X, Y):
      X_train, X_test = X.loc[train_index],X.loc[test_index]
      Y_train, Y_test = Y.loc[train_index],Y.loc[test_index]


    print('{:<15} {:<15} {:<15}'.format('DataSet','Features','Label'))
    print('{:<15} {:<15} {:<15}'.format('Train',X_train.shape[0],Y_train.shape[0]))
    print('{:<15} {:<15} {:<15}'.format('Test',X_test.shape[0],Y_test.shape[0]))

    DataSet         Features        Label
    Train           352             352
    Test            88              88
```

Milestone 4: Model Building

Now our data is cleaned and it's time to build the model. We can train our data on different algorithms. For this project we are applying Two classification algorithms KNN, Naive-Bayes algorithm. The best model is saved based on its performance.

1. K-Neighbors Classifier

```python
#Choosing K value
error_rate = []

for i in range(1,30):
    KNN_class = KNeighborsClassifier(n_neighbors = i)
    KNN_class.fit(X_train, Y_train)
    Y_predict = KNN_class.predict(X_test)
    #appending error rate that is not equal to test dataset
    error_rate.append(np.mean(Y_predict != Y_test))

#Plotting the plot to check the K value from graph
plt.figure(figsize = (12,6))
plt.plot(range(1,30), error_rate, color = 'blue', linestyle = 'dashed', marker = 'o', markerfacecolor = 'yellow', markersize = 10)
plt.title('Error Value vs K-Value')
plt.xlabel('K-Value')
plt.ylabel('Error Value')
```

```
Text(0, 0.5, 'Error Value')
```


2. **Naive-Bayes algorithms**

```python
GNB = GaussianNB()

GNB.fit(X_train, Y_train)

Y_predict = GNB.predict(X_train)

print("Train Data Model Accuracy: {0:.3f}".format(metrics.accuracy_score(Y_train, Y_predict)*100))

Y_predict = GNB.predict(X_test)

Train Data Model Accuracy: 84.659
```

Milestone 5: Application Building
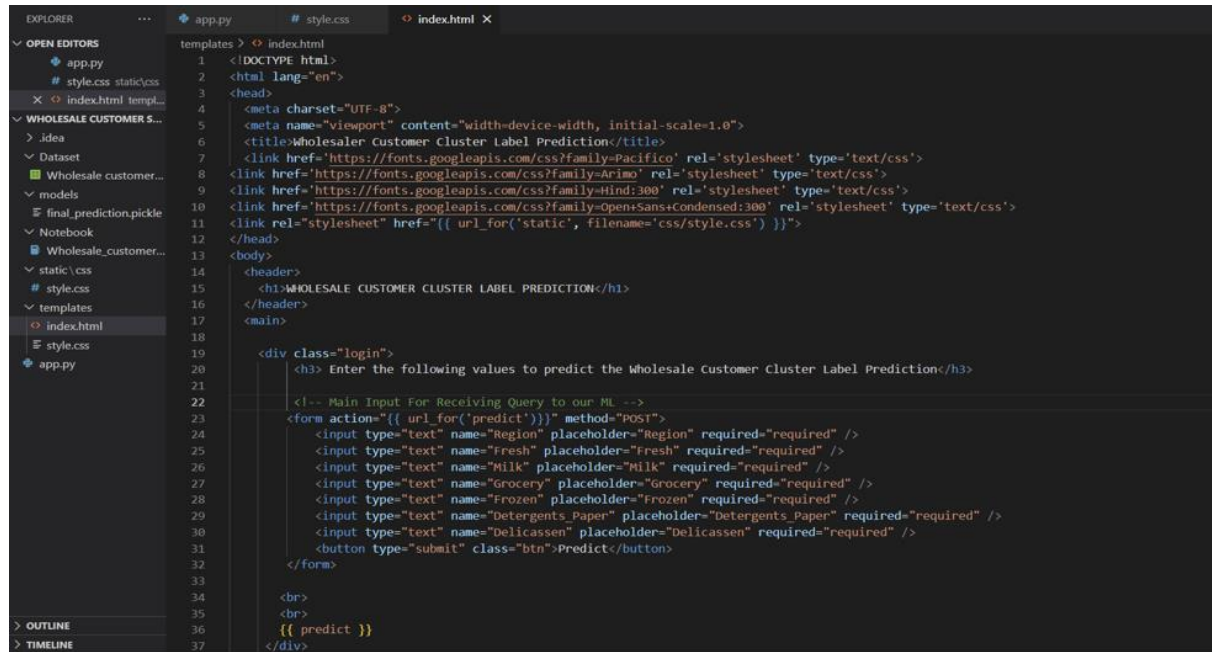
After the model is built, we will be integrating it to a web application so that normal users can also use it. The UI is based on taking input for various parameters and a predict button which

will use the ML model to predict which cluster label the given input parameters correspond to.
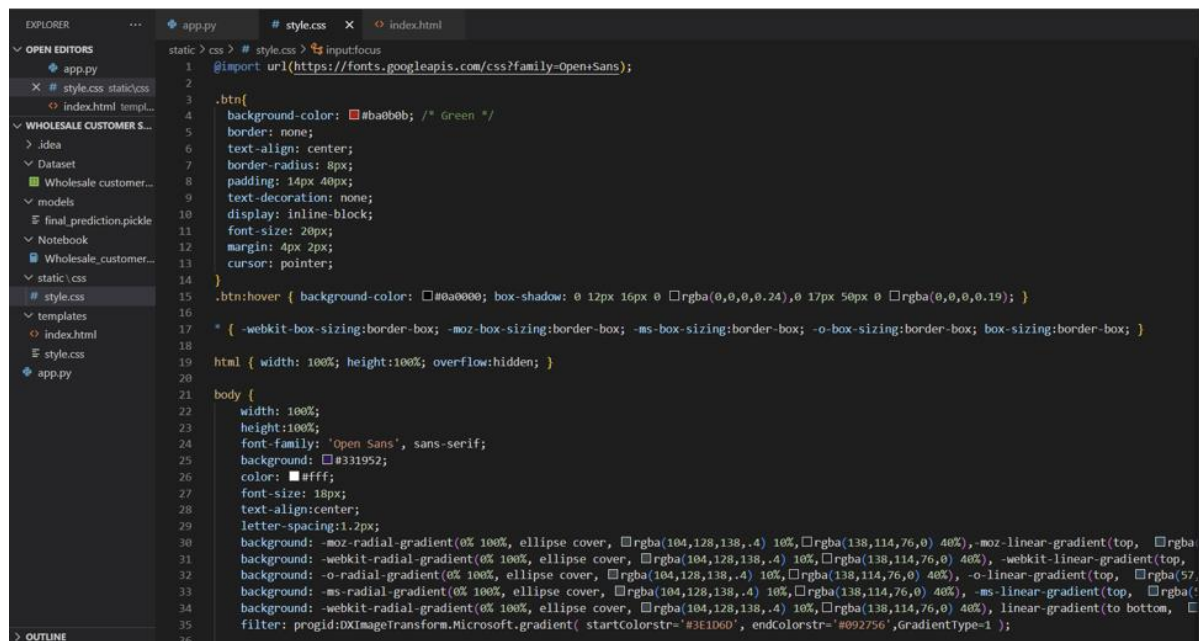
Activity1: Building Html Pages:

For this project create an HTML file and a CSS file for styling namely

- Index.html
- Style.css

Activity 2: Build Python code:

Step 1: Import the libraries

```
1    from flask import Flask, render_template, url_for, request
2    import pickle as p
3    import pickle
4    from flask import Flask, request,jsonify,render_template
5    import numpy as np
6    import pandas as pd
7    from sklearn.preprocessing import StandardScaler
8
```

Step 2: Initialise the flask app and load the model

```
modelfile = 'models/final_prediction.pickle'
model = p.load(open(modelfile, 'rb'))
app = Flask(__name__)
```

Step 3: Render the HTML pages.

```
@app.route('/')
def welcome():
    return render_template('index.html')
```

Step 4: Build the predict function which will take inputs from the UI and then use it to predict the cluster label and give the prediction to the submit in HTML page.

```
@app.route('/predict', methods =['GET','POST'])
def predict():
    Region = float(request.form['Region'])
    Fresh = float(request.form['Fresh'])
    Milk = float(request.form['Milk'])
    Grocery = float(request.form['Grocery'])
    Frozen = float(request.form['Frozen'])
    Detergents_Paper = float(request.form['Detergents_Paper'])
    Delicassen = float(request.form['Delicassen'])

    total = [[Region,Fresh,Milk,Grocery,Frozen,Detergents_Paper,Delicassen]]
    prediction = model.predict(total)
    prediction = int(prediction[0])


    if prediction==1:
        return render_template('index.html',predict="Customer belongs to Cluster Label 1")
    else:
        return render_template('index.html',predict="Customer belongs to Cluster Label 2")
```

Main Function:

```
if __name__ == "__main__":
    app.run(debug=True)
```

Activity 3: Run the application

Run the flask application using the run method. By default, the flask runs on port 5000. If the port is to be changed, an argument can be passed and the port can be modified.

- Run the application using the anaconda command prompt or the powershell terminal in Virtual Studio Code if required libraries are installed.
- We need to move to the folder where the project is stored in the terminal and then type the command "python app.py" to run the flask application.

```
PS C:\Users\Asus Tuf A15\Wholesale Customer segmentation> python app.py
 * Serving Flask app 'app'
 * Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on http://127.0.0.1:5000
Press CTRL+C to quit
 * Restarting with watchdog (windowsapi)
 * Debugger is active!
 * Debugger PIN: 577-256-571
```

7.3 Performance Metrics

We need to evaluate both the models to find which model we should use while predicting for the application.

	Model	Train Accuracy	Test Accuracy	F1-Score	Recall	Precision	AUC	Best_Cross_Val_Score	False Negatives	False Positives	True Negatives	True Positives
0	KNN	84.94	90.91	0.79	0.88	0.71	10.11	0.84 (+/- 0.05)	2	6	65	15
1	Naïve Bayes	84.66	86.36	0.68	0.76	0.62	17.40	0.84 (+/- 0.03)	4	8	63	13

8. PERFORMANCE TESTING

8.1 Performance Metrics

	Model	Train Accuracy	Test Accuracy	F1-Score	Recall	Precision	AUC	Best_Cross_Val_Score	False Negatives	False Positives	True Negatives	True Positives
0	KNN	84.94	90.91	0.79	0.88	0.71	10.11	0.84 (+/- 0.05)	2	6	65	15
1	Naive Bayes	84.66	86.36	0.68	0.76	0.62	17.40	0.84 (+/- 0.03)	4	8	63	13

9. RESULTS

9.1 Output Screenshots

10. ADVANTAGES & DISADVANTAGES

When considering wholesale customer segmentation analysis, there are several advantages and some potential disadvantages to be aware of:

Advantages:

1. Targeted Marketing: Segmentation allows for more targeted and personalized marketing strategies, which can increase customer engagement and sales by addressing specific customer needs and preferences.

2. Improved Customer Satisfaction: By tailoring products, services, and marketing strategies to different customer segments, there's a higher likelihood of meeting their specific needs, ultimately leading to increased satisfaction.

3. Enhanced Product Development: Understanding customer segments can guide product development by identifying trends and demands in different market niches, enabling the creation of products that cater to specific customer requirements.

4. Cost Efficiency: Targeted marketing and better resource allocation due to segmentation can lead to cost savings by focusing efforts on high-potential customer segments rather than employing a one-size-fits-all approach.

5. Competitive Advantage: A better understanding of the market through segmentation can provide a competitive edge by identifying market gaps and allowing for the development of strategies to exploit those gaps.

Disadvantages:

1. Complexity and Cost: Developing and maintaining a sophisticated segmentation strategy can be expensive and require ongoing investment in data collection, analytics, and tools.

2. Overgeneralization or Oversimplification: There's a risk of oversimplifying customer segments, leading to broad assumptions, or overcomplicating segments, making it challenging to implement effective strategies.

3. Data Privacy Concerns: Analyzing customer data to create segments may raise concerns about privacy and data protection, especially in regions with strict regulations.

4. Segmentation Inaccuracy: If the data used for segmentation is flawed or outdated, the resulting segments may not accurately represent the customer base, leading to misguided strategies.

5. Segment Overlap and Conflict: Sometimes, customers may fit into multiple segments, leading to challenges in targeting and creating consistent marketing strategies.

Careful consideration of these advantages and disadvantages is essential in determining whether wholesale customer segmentation analysis is the right approach for a particular business and how it should be implemented to maximize benefits and mitigate potential drawbacks.

11. CONCLUSION

Wholesale customer segmentation analysis offers numerous benefits that can significantly impact a business's marketing, sales, and overall strategy. It enables personalized approaches, improves customer satisfaction, and guides resource allocation effectively, leading to cost savings and enhanced competitiveness. Additionally, data-driven insights empower businesses to make informed decisions and predict future trends.

In conclusion, while the advantages of wholesale customer segmentation analysis are compelling and potentially transformative for businesses, it's crucial to approach its implementation carefully. Businesses must strike a balance between leveraging the benefits of segmentation and managing the associated challenges, ensuring that the strategy aligns with both customer needs and the company's resources. With careful planning, continual refinement, and a commitment to data accuracy and privacy, businesses can harness the power of customer segmentation to drive growth and success.

12. FUTURE SCOPE

The future scope of wholesale customer segmentation analysis is likely to continue evolving, driven by advancements in technology, data analytics, and shifting consumer behaviors. Here are potential areas of development and prospects for wholesale customer segmentation analysis:

Advanced Analytics and AI:

Machine Learning and AI: Advancements in AI and machine learning will refine segmentation models, allowing for more accurate and dynamic customer segmentation based on real-time data.

Predictive Analytics: The use of predictive models will become more sophisticated, enabling businesses to forecast customer behavior and trends with higher precision.

Data Integration and Personalization:

Omni-Channel Segmentation: Integration of data across various channels, including online, offline, social media, and IoT, will provide a more comprehensive view of customer behavior, enabling better segmentation.

Hyper-Personalization: Segmentation will move beyond broad categories to hyper-personalization, tailoring products, services, and marketing messages at an individual level based on real-time data.

Ethical and Regulatory Considerations:

Ethical Data Use: Emphasis on ethical data practices, transparency, and user consent will become a significant focus, particularly with the growing emphasis on data privacy and regulatory compliance.

Compliance Adaptation: Businesses will need to continuously adapt their segmentation strategies to comply with evolving data privacy regulations and consumer protection laws.

Customer Experience and Engagement:

Customer Journey Mapping: Segmentation will extend to map the entire customer journey, facilitating a more holistic understanding of the customer experience.

Emotional Analytics: Integration of emotional analytics and sentiment analysis into segmentation strategies to gauge customer sentiment and tailor experiences accordingly.

Technological Integration:

Blockchain Integration: Secure data sharing through blockchain technology might enhance the integrity and security of customer data used in segmentation.

Augmented Reality (AR) and Virtual Reality (VR): Utilization of AR/VR technology in understanding customer behavior for segmentation, particularly in industries where immersive experiences impact buying decisions.

Sustainability and Social Responsibility:

Segmentation for Sustainability: Segmentation strategies may align with sustainability efforts, allowing businesses to target environmentally conscious consumer segments.

Social Impact Segmentation: Identifying customer segments interested in social responsibility and using this data for cause-related marketing efforts.

Global Expansion and Localization:

Global Segmentation Strategies: Businesses expanding globally will focus on creating segmentation strategies tailored to diverse cultural, economic, and geographic regions.

Localization Algorithms: Algorithms that adapt segmentation models based on local nuances and preferences, improving accuracy and relevance in different markets.

Continuous Adaptation and Improvement:

Real-Time Segmentation: Moving towards real-time segmentation, allowing businesses to adjust strategies rapidly based on immediate changes in customer behavior.

As the business landscape evolves, the future of wholesale customer segmentation analysis will undoubtedly involve leveraging cutting-edge technology, prioritizing ethical considerations, and focusing on a more personalized and customer-centric approach.

Businesses that can adapt and integrate these advancements into their segmentation strategies will likely gain a competitive edge in understanding, engaging, and satisfying their customers.

13. APPENDIX

Source Code

https://colab.research.google.com/drive/1H51je8kBxpBab8mjqG5z0Ny6TXzxiFvB#scrollTo=5fRHVXmNHADA

GitHub & Project Demo Link

GitHub: smartinternz02/SI-GuidedProject-594318-1697471997 (github.com)

Project Demo: https://drive.google.com/file/d/1H-fBBVtcTmrcwxSnHEtU8QaIS_StWWus/view?usp=drive_link