Table Of Contents

A. Problem Statement

B. Project Objective

C. Data Description

D. Data Pre-processing Steps

E. Exploratory Data Analysis (EDA)

F. Insights

G. Customer Segmentation

H. Inferences

I. Model Selection and Building

J. Future Possibilities of the Project

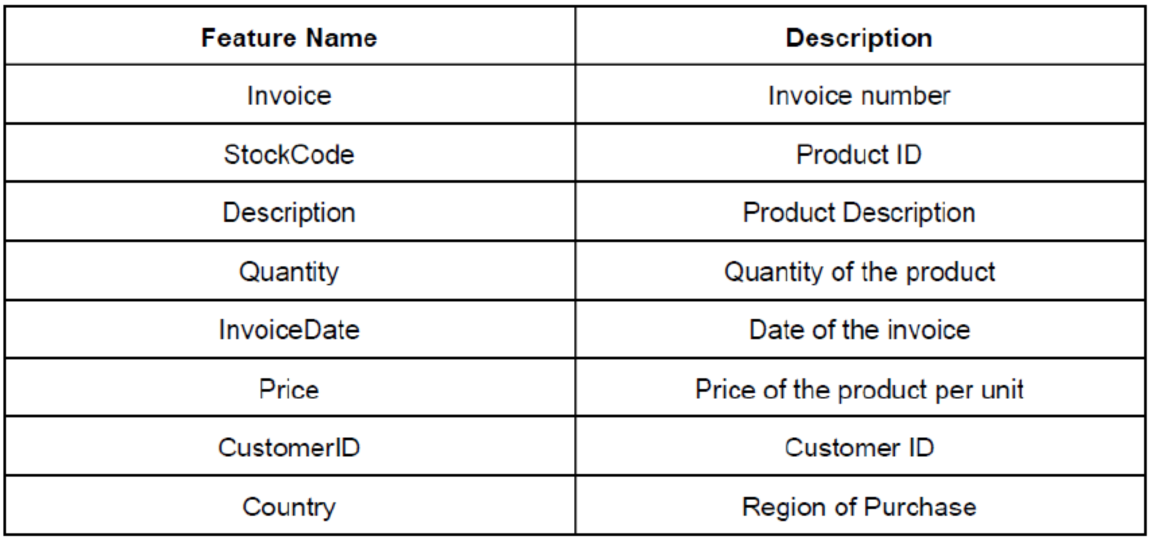
A. Problem Statement

An online retail store is trying to understand the various customer purchase patterns for their firm, we are required to give enough evidence based insights to provide the same. Also segmenting the customers based on their purchasing behavior.

B. Project Objective

1. Using the data, have to find useful insights about the customer purchasing history that can be an added advantage for the online retailer.
2. Segment the customers based on their purchasing behavior.

C. Data Description



Feature Name	Description
Invoice	Invoice number
StockCode	Product ID
Description	Product Description
Quantity	Quantity of the product
InvoiceDate	Date of the invoice
Price	Price of the product per unit
CustomerID	Customer ID
Country	Region of Purchase

D. Data Pre-processing Steps

1. Loading the Dataset & Libraries: The initial step involves loading the dataset into the analysis environment, typically using libraries which helps in data manipulation, visualisation in Python, ensuring accessibility for further examination.
2. Shape Inspection : To get the overview of data.
3. Checking for Data Types: It is imperative to inspect the data types of each column to ensure consistency and appropriateness for subsequent analyses and operations.
4. Converting Date Column: When dealing with temporal data, such as dates, converting the date column from an object type to a date type facilitates time-based analyses and visualizations.

5. Extracting strings from alphanumeric column to make it numeric.
6. Handling Missing Values : It's important to deal with missing/null values by 'dropping' or 'imputing' in order to perform EDA or model making.
7. Descriptive Statistics : Providing insights into its central tendency, dispersion, and distribution. This is crucial for understanding data patterns and identifying anomalies before further analysis.

   7.1 Dropping rows having negative numeric data

   7.2 Dropping rows having Unit Price = 0

8. Handling Outliers: Identification and treatment of outliers are crucial to maintain data integrity.

   8.1 PLotting boxplot to investigate outliers.

   8.2 Not Removing them as this is transaction data where outliers might provide meaningful insights

E. Exploratory Data Analysis (EDA)

1. Providing Unique Counts in all numeric features.
2. Providing Counts of all Unique things in all features

Exporting the final cleaned data into csv file named 'Final.csv' for future use to draw insights and customer segmentation.

F. Insights

1. Customer Based

   1.1 Top 10 Customer Based on Order Value

   Inference -

   1.Customer ID 14646,18102 are biggest customers

   2.Top 10 Customers contribute upto 17 % of total Order Value

2. Product Based

   2.1 Top favourite products of customers by order value

   Inference -

1.PAPER CRAFT , LITTLE BIRDIE makes a biggest order value of Customers.

2.Top 10 products contribute to 10% of total order value.

2.2 Top Ordered products of customers by quantity

Inference -

1.'PAPER CRAFT , LITTLE BIRDIE' & 'MEDIUM CERAMIC TOP STORAGE JAR' are the products ordered in large quantities.

3. Country Based

3.1 Countries giving maximum orders

Inference -

1.We have maximum order from UK - 349201.

2.Germany, France & Ireland are next biggest customers.

3.2 Countries giving maximum order value

Inference -

1.UK has maximum order value which is 82 % of total order value.

2.Other top 4 countries have 11% share in total order value.

3.3 Countries having maximum unique customers

Inference -

1.Uk has maximum number of customers - 90%

2.Top 5 Countries contribute to 96% of unique customers namely -

A.UK B.Germany C.France D.Spain E.Belgium

3.4 Country wise Trending Product based on frequency

3.5 Countrywise Highest Revenue generating products

3.6 Most ordered Products in UK

Inference -

1.Product ordered in Highest quantity in UK is - 'MEDIUM CERAMIC TOP STORAGE JAR' contributes to 5% of total quantity.

3.7 Highest Revenue Generating Products in UK

Inference -

1.Highest Revenue Generating Product of UK is MEDIUM CERAMIC TOP STORAGE JAR which contributes to 3.2 %

4. Time Based

Calculated Daily, Monthly, Week-Day wise revenue.

4.1 Highest grossing month

Inference -

1.Month of November marks the highest revenue generation

2.Month of April & February marks the lowest revenue generation

G. Customer Segmentation Based on Purchasing Behaviour

Segmenting customers based on their **purchasing behavior** involves analyzing their transaction history to identify distinct groups.

RFM Segmentation (Recency, Frequency, Monetary), which evaluates customer behavior based on three metrics:

1.**Recency**: How recently a customer made a purchase.

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

3.**Monetary** : How much a customer spends.

Apply RFM Scoring

R_Score - Assigning Highest R_score(5) to customers who purchased very recently (Low Recency).

F_Score- Assigning Highest F_score(5) to customers who purchased very frequently (High Frequency- more unique Invoice Count).

M_Score- Assigning Highest M_score(5) to customers who purchased of large order value.

Segmentation Categories - 'Loyal', 'At Risk', 'Regular', 'Irregular', 'Recent Visitors', 'Frequent Buyers', 'Big Spenders', 'Others'

H. Inferences

1.There are 4% of customers 'At Risk' who neither visits frequently nor gives significant Order Value.

2. 20% customers are 'Irregular' who does not buys frequently which are less than 'Regular' customers (16%).

3.Only 7% customers are 'Loyal' who are frequent as well as give big order values.

Analysing High Spending Customers

Big Spenders and Irregular have high order value than Big Spenders and Regular.

I. Model Building

As this is customer segmentation problem, we will proceed with classification algorithm.

We will build 'Random Forest' and 'SVC' models as both can capture the complex relationship between data and generalise well.

'Random Forest' and **'SVC'** both can deal with problem of overfitting efficiently.

Insights -

Both models are achieving over 90% accuracy for customer behavior prediction

J. Future Possibilities of the Project

'Customer Loyalty' - We can identify and 'AWARD' customers based on - 'Regular' and 'Loyal' customers who contribute consistently to revenue.

'Customer Retention' - Identifying customers based on their irregularity i.e high Recency & low R_score and focusing on them.

'Revenue Analysis' - Analysing the spending behaviours of different segments.

'Targeted Marketing' - Different campaigns for different customer segments i.e Rewards for Loyal and Win-back campaigns for Irregular customers.