



## **Data Collection and Preprocessing Phase**

Date	18 June 2024
Team ID	739642
Project Title	Customer shopping segmentation using machine learning
Maximum Marks	6 Marks

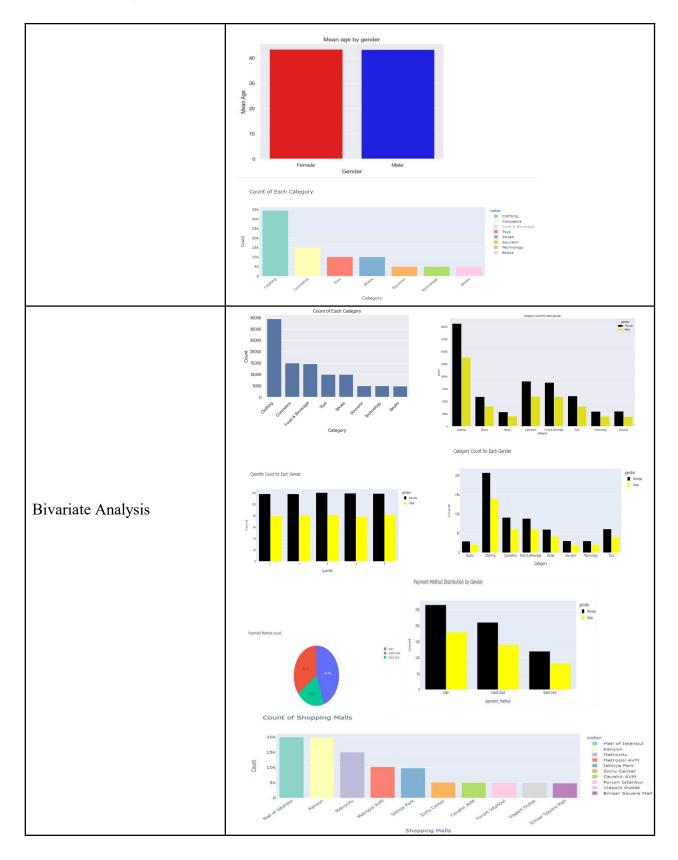
## **Data Exploration and Preprocessing Report**

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description										
	Dimension: 99457 rows × 10columns Descriptive statistics:										
		invoice_no	customer_id	gender	age	category	quantity	price	payment_method	invoice_date	shopping_mal
		1138884	C241288	Female	28	Clothing	5	1500.40	Credit Card	5/8/2022	Kanyor
		1 1317333	C111565	Male	21	Shoes	3	1800.51	Debit Card	12/12/2021	Forum Istanbu
Data Overview		2 1127801	C266599	Male	20	Clothing	1	300.08	Cash	9/11/2021	Metrocity
		3 1173702	C988172	Female	66	Shoes	5	3000.85	Credit Card	16/05/2021	Metropol AVN
		4 1337046	C189076	Female	53	Books	4	60.60	Cash	24/10/2021	Kanyor
	Gender distribution Gender distribution										
Univariate Analysis	60000						60000				
	50000						50000				
	40000						40000				
	30000					Count	30000				
	20000						20000				
	10000						10000				
	(	Fem	ale Gende		lale		0	Fe	male Gender	Male	

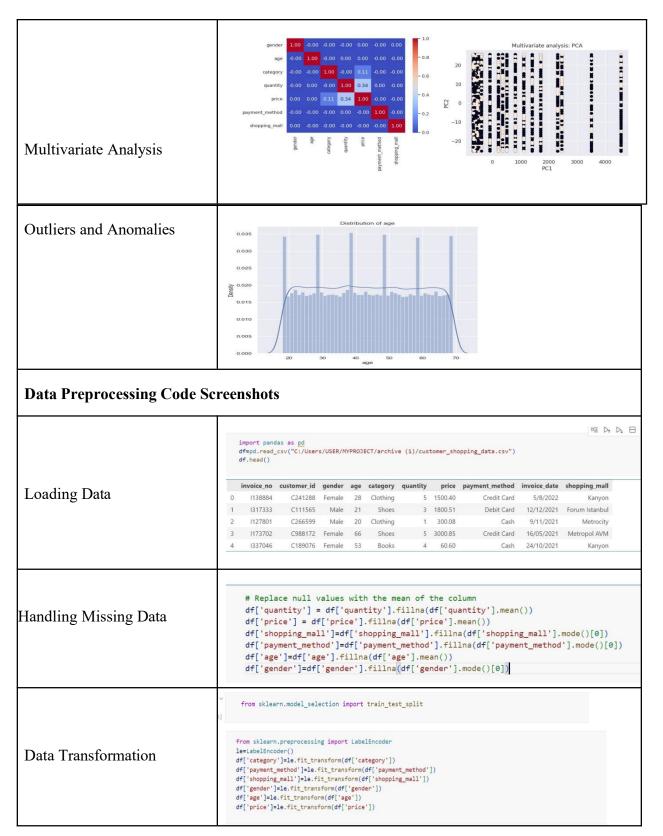
















	<pre># Encoding categorical variables label_encoders = {} for column in ['gender', 'category', 'payment_method', 'shopping_mall']:     le = LabelEncoder()     df[column] = le.fit_transform(df[column])     label_encoders[column] = le</pre>
	<pre># Feature scaling from sklearn.preprocessing import StandardScaler, LabelEncoder scaler = StandardScaler() df[['age', 'quantity', 'price']] = scaler.fit_transform(df[['age', 'quantity', 'price']])</pre>
	<pre>#Splitting the dataset into training and testing sets X = df.drop(columns=['invoice_no', 'invoice_date', 'customer_id']) y = df['category'] # Assuming you want to predict the category X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)</pre>
Feature Engineering	Attached the codes in final submission.
Save Processed Data	-