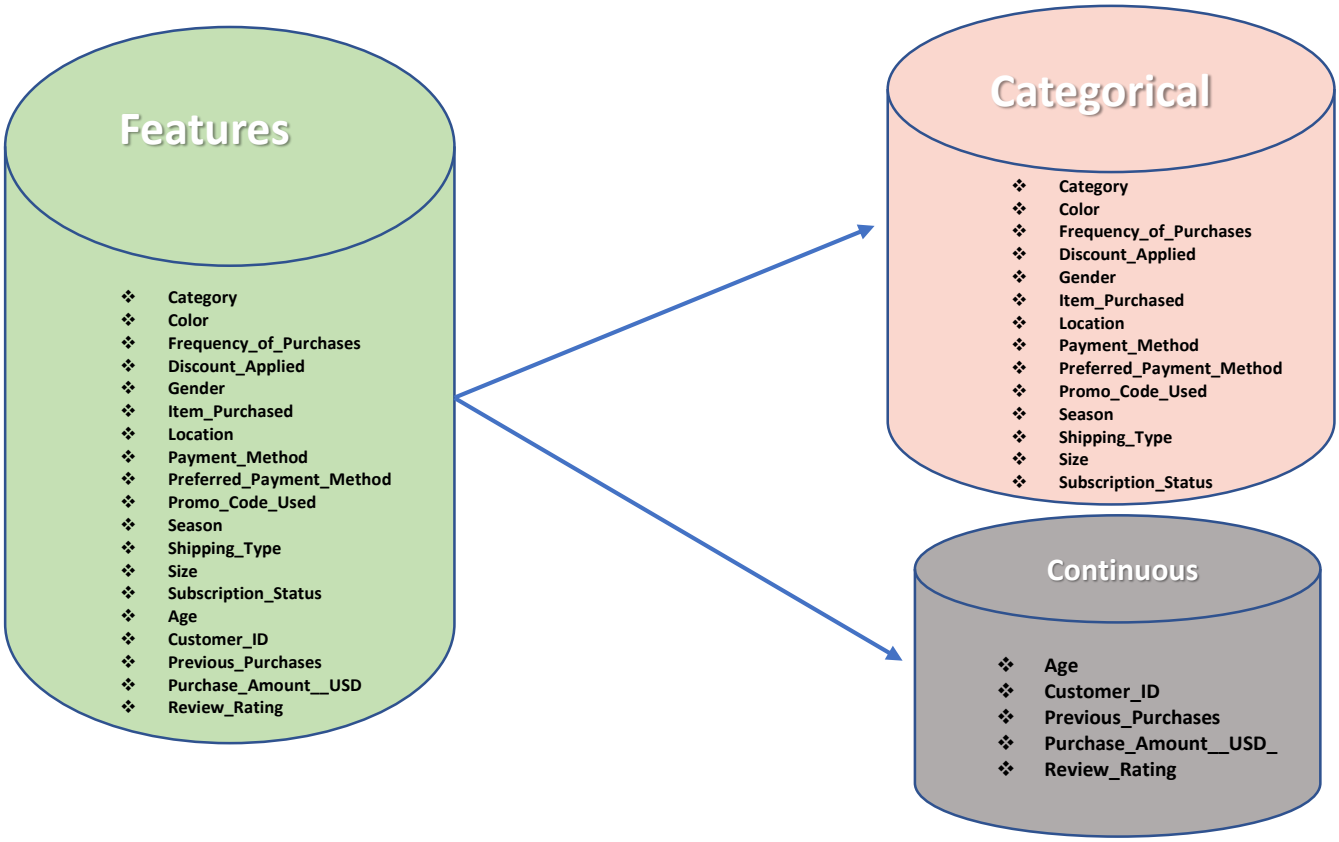
Introduction

The Customer Shopping Preferences Dataset offers valuable insights into consumer behavior and purchasing patterns. Understanding customer preferences and trends is critical for businesses to tailor their products, marketing strategies, and overall customer experience. This dataset captures a wide range of customer attributes including age, gender, purchase history, preferred payment methods, frequency of purchases, and more. Analyzing this data can help businesses make informed decisions, optimize product offerings, and enhance customer satisfaction. The dataset stands as a valuable resource for businesses aiming to align their strategies with customer needs and preferences. It's important to note that this dataset is a Synthetic Dataset Created for Beginners to learn more about Data Analysis and Machine Learning.

Content

This dataset encompasses various features related to customer shopping preferences, gathering essential information for businesses seeking to enhance their understanding of their customer base. The features include customer age, gender, purchase amount, preferred payment methods, frequency of purchases, and feedback ratings. Additionally, data on the type of items purchased, shopping frequency, preferred shopping seasons, and interactions with promotional offers is included. With a collection of 3900 records, this dataset serves as a foundation for businesses looking to apply data-driven insights for better decision-making and customer-centric strategies.



Dataset Glossary

Customer ID - Unique identifier for each customer
Age - Age of the customer
Gender - Gender of the customer (Male/Female)
Item Purchased - The item purchased by the customer
Category - Category of the item purchased
Purchase Amount (USD) - The amount of the purchase in USD
Location - Location where the purchase was made
Size - Size of the purchased item
Color - Color of the purchased item
Season - Season during which the purchase was made
Review Rating - Rating given by the customer for the purchased item
Subscription Status - Indicates if the customer has a subscription (Yes/No)
Shipping Type - Type of shipping chosen by the customer
Discount Applied - Indicates if a discount was applied to the purchase (Yes/No)
Promo Code Used - Indicates if a promo code was used for the purchase (Yes/No)
Previous Purchases - The total count of transactions concluded by the customer at the store, excluding the ongoing transaction
Payment Method - Customer's most preferred payment method
Frequency of Purchases - Frequency at which the customer makes purchases (e.g., Weekly, Fortnightly, Monthly)

Objectives

- Determine which product categories are most popular among customers.
- Use customer preferences to forecast demand and optimize inventory levels for different products.
- Understand how customer preferences change with seasons and adjust marketing and inventory strategies accordingly.
- Use historical data to build predictive models that anticipate future customer shopping trends.

Business Problem

How can we optimize inventory levels to meet the demands of our most loyal and frequent customers, while also developing predictive models that anticipate future trends in customer shopping behavior ?

Methodology

- **Data cleaning and processing**: It involve the identification, handling, and transformation of raw data to ensure its quality, consistency, and readiness for analysis or modelling.
- **Exploratory Data Analysis (EDA)**: Process of visually and statistically analyzing datasets to discover patterns, relationships, and insights, aiding in the understanding of data characteristics and informing subsequent analyses.

- **Feature Engineering**: Process of transforming raw data into a new set of feature that enhances the performance of the ML models, improving their predictive power for generalization.
- **Building Predictive Model**: It involves developing a mathematical algorithm that leverages historical data to make accurate predictions or classifications on new, unseen data points.
- **Model Evaluation**: Assessment of a machine learning model's performance, typically through metrics such as accuracy, precision, recall, and F1 score, to gauge its effectiveness in making predictions on unseen data.
- **Business recommendations:** With all the insights taken from the dataset and the model used, some recommendations are given.

Data Cleaning and Preprocessing

Illustration 1: Number of Missing Value for Numerical Variables

Number of missing values numerical variables

The MEANS Procedure

Variable	N Miss
Customer_ID	0
Age	48
Purchase_Amount__USD_	38
Review_Rating	0
Previous_Purchases	23

Illustration 2: Number of Missing Value for Categorical Variables

Category	Color	Gender	Location	Payment_Method	Season	Shipping_Type	Subscription_Status
4	21	18	18	2	1	17	2

Illustration 3: Duplicated Values

Obs	Customer_ID	Age	Gender	Item_Purchased	Category	Purchase_Amount__USD_	Location	Size	Color	Season	Review_Rating	Subscription_Status	Payment_Method	Shipping_Type	Discount_Applied	Promo_Code_Used	Previous_Purchases	Preferred_Payment_Method	Frequency_of_Purchases
1	2208	69	Male	Handbag	Accessories	38	New Mexico	M	Gray	Fall	2.9	No	Debit Card	Next Day A	No	No	11	Credit Card	Annually
2	3819	70	Fema	Sneakers	Footwear	41	Oregon	XL	Indigo	Winter	3.8	No	Credit Card	Free Shipp	No	No	42	Cash	Monthly
3	3875	70	Fema	Sweater	Clothing	54	Nevada	XL	Beige	Summer	3.9	No	Cash	2-Day Shir	No	No	33	Credit Card	Bi-Weekly

Illustration 4: Customer Shopping Preferences - Head

Customer_ID	Age	Gender	Item_Purchased	Category	Purchase_Amount__USD	Location	Size	Color	Season	Review_Rating	Subscription_Status	Payment_Method	Shipping_Type	Discount_Applied	Promo_Code_Used	Previous_Purchases	Preferred_Payment_Method	Frequency_of_Purchases
1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Credit Card	Express	Yes	Yes	14	Venmo	Fortnightly
2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Bank Transfer	Express	Yes	Yes	2	Cash	Fortnightly
3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Cash	Free Shipping	Yes	Yes	23	Credit Card	Weekly
4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	PayPal	Next Day Air	Yes	Yes	49	PayPal	Weekly
5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Cash	Free Shipping	Yes	Yes	31	PayPal	Annually
6	46	Male	Sneakers	Footwear	20	Wyoming	M	White	Summer	2.9	Yes	Venmo	Standard	Yes	Yes	14	Venmo	Weekly
7	63	Male	Shirt	Clothing	85	Montana	M	Gray	Fall	3.2	Yes	Debit Card	Free Shipping	Yes	Yes	49	Cash	Quarterly
8	27	Male	Shorts	Clothing	34	Louisiana	L	Charcoal	Winter	3.2	Yes	Debit Card	Free Shipping	Yes	Yes	19	Credit Card	Weekly
9	26	Male	Coat	Outerwear	97	West Virginia	L	Silver	Summer	2.6	Yes	Venmo	Express	Yes	Yes	8	Venmo	Annually
10	57	Male	Handbag	Accessories	31	Missouri	M	Pink	Spring	4.8	Yes	PayPal	2-Day Shipping	Yes	Yes	4	Cash	Quarterly

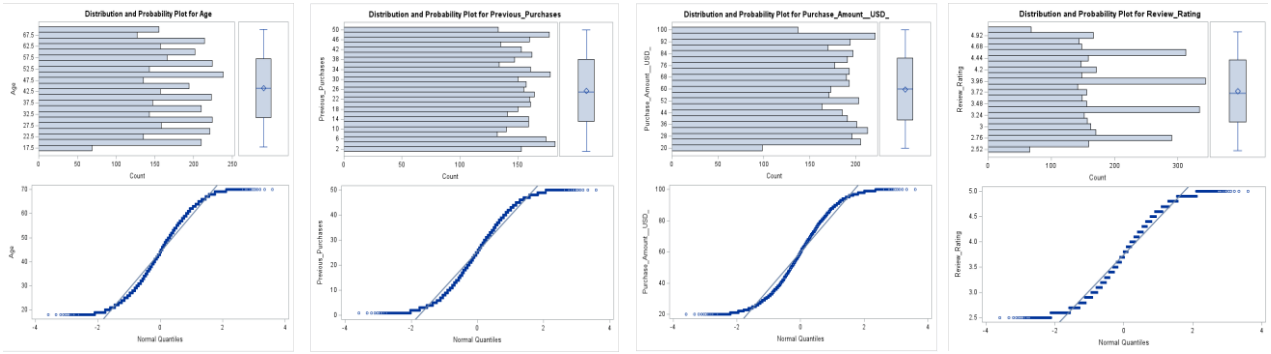
Illustration 5: Customer Shopping Preferences - Tail

Obs	Customer_ID	Age	Gender	Item_Purchased	Category	Purchase_Amount_USD	Location	Size	Color	Season	Review_Rating	Subscription_Status	Payment_Method	Shipping_Type	Discount_Applied	Promo_Code_Used	Previous_Purchases	Preferred_Payment_Method	Frequency_of_Purchases
3894	3894	21	Fema	Hat	Accessories	64	Massachusetts	L	White	Fall	3.3	No	Bank Transfe	Store Pickup	No	No	29	Bank Transfer	Bi-Weekly
3895	3895	66		Skirt	Clothing	78	Connecticut	L	White	Spring	3.9	No	Cash	2-Day Shipping	No	No	.	Credit Card	Every 3 Mon
3896	3896	40	Fema	Hoodie	Clothing	28	Virginia	L	Turquoise	Summer	4.2	No	Cash	2-Day Shipping	No	No	32	Venmo	Weekly
3897	3897	52	Fema	Backpack	Accessories	49	Iowa	L	White	Spring	4.5	No	PayPal	Store Pickup	No	No	41	Bank Transfer	Bi-Weekly
3898	3898	.	Fema	Belt	Accessories	33	New Jersey	L	Green	Spring	2.9	No	Credit Card	Standard	No	No	24	Venmo	Quarterly
3899	3899	44	Fema	Shoes	Footwear	77	Minnesota	S	Brown	Summer	3.8	No	PayPal	Express	No	No	24	Venmo	Weekly
3900	3900	52	Fema	Handbag	Accessories	81	California	M	Beige	Spring	3.1	No	Bank Transfe	Store Pickup	No	No	33	Venmo	Quarterly
3901	2208	69	Male	Handbag	Accessories	38	New Mexico	M	Gray	Fall	2.9	No	Debit Card	Next Day Air	No	No	11	Credit Card	Annually
3902	3875	70	Fema	Sweater	Clothing	54	Nevada	XL	Beige	Summer	3.9	No	Cash	2-Day Shipping	No	No	33	Credit Card	Bi-Weekly
3903	3819	70	Fema	Sneakers	Footwear	41	Oregon	XL	Indigo	Winter	3.8	No	Credit Card	Free Shipping	No	No	42	Cash	Monthly

The customer shopping preferences dataset contains 3903 observations and 19 features, with some little missing and duplicated value. The outliers of the data are not very extreme, so we decided to keep them. The target is feature is **Frequency_of_Purchases** with 7 levels.

Exploratory Data Analysis

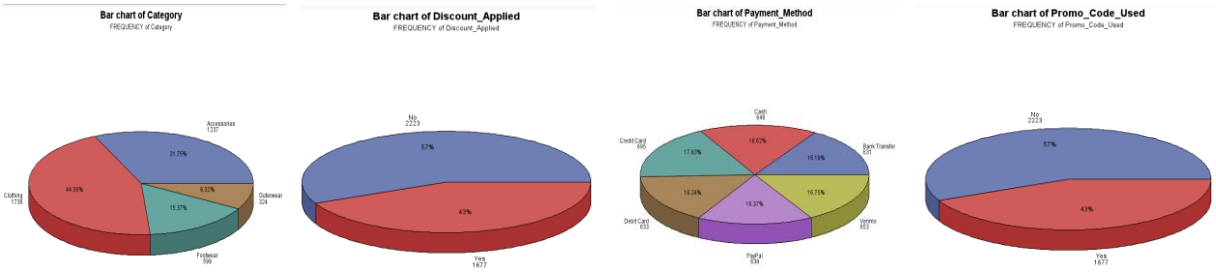
Illustration 6: Univariate Analysis (Distribution of Numerical Variables)



Key Insights:

- ❑ There are no major outliers in the data.
- ❑ Since all P-value < 0.05, the data are normally distributed.
- ❑ People ages around 50 are the most representative among the population.
- ❑ Consumers usually spent the amount of $2, $32, and $48.
- ❑ The highest peak for the amount spent is around $95.
- ❑ Most Review_rating have been made at 3.96.

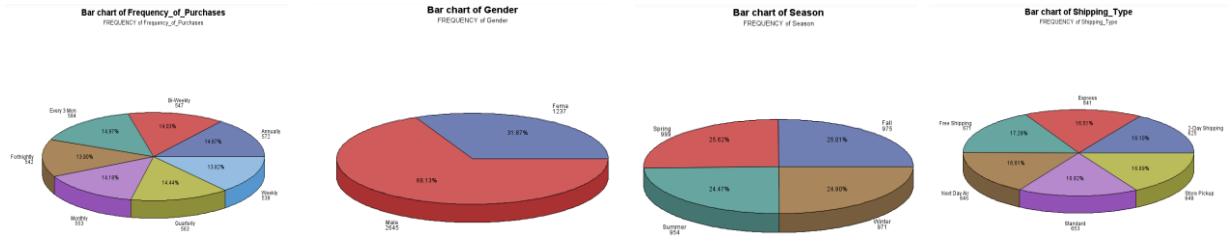
Illustration 7: Univariate Analysis (Distribution of Categorical Variables)



Key Insights:

- ❑ In terms of the Category, Clothing is the best-seller with 44.56% of total Sales
- ❑ 57% of purchases were not discounted
- ❑ The most widely used method of payment is by Credit card with 17.83%
- ❑ 57% of purchases do not use a promotional code

Illustration 8: Univariate Analysis (Distribution of Categorical Variables)



Key Insights:

- ❑ Sales are made at almost the same frequency. However, the frequency of purchases made every 3 months is slightly higher with 14.97%
- ❑ Male made more purchases with a rate of 68.13% of total sales
- ❑ Purchases are the same every season. However, sales in the spring season are slightly higher with 25.62% of total sales.
- ❑ Free shipping is the most used delivery channel with 17.28%.

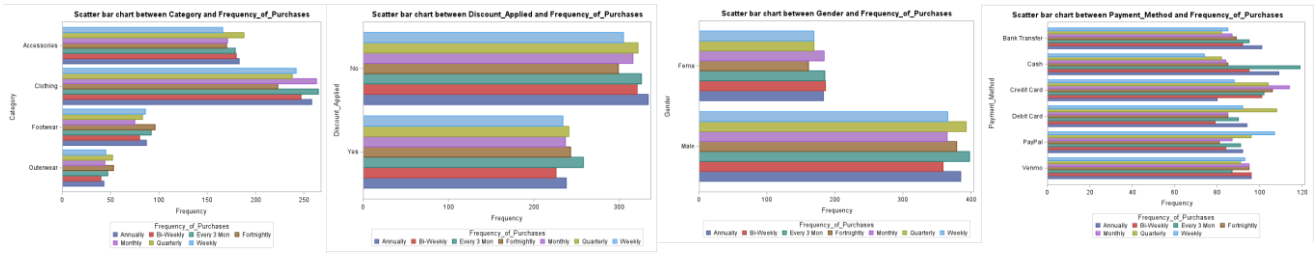
Illustration 9: Bivariate Analysis (Continuous VS Continuous)

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations				
	Age	Purchase_Amount__USD_	Review_Rating	Previous_Purchases
Age	1.00000 3852	-0.00964 0.5518 3815	-0.02478 0.1242 3852	0.03744 0.0205 3829
Purchase_Amount__USD_	-0.00964 0.5518 3815	1.00000 3862	0.02935 0.0682 3862	0.00696 0.6665 3840
Review_Rating	-0.02478 0.1242 3852	0.02935 0.0682 3862	1.00000 3900	0.00460 0.7744 3877
Previous_Purchases	0.03744 0.0205 3829	0.00696 0.6665 3840	0.00460 0.7744 3877	1.00000 3877

Key Insights:

- ❑ There are negative correlation between Age, Purchase_Amount__USD_ (-0.00964), Review_Rating (-0.02478), little positive correlation between age and Previous_Purchases(0.03744)
- ❑ There are little positive corr between Purchase_Amount__USD_, Review_Rating (0.02935) and Previous_Purchases (0.00696)
- ❑ There is a little positive correlation between Review_Rating and Previous_Purchases (0.00460)

Illustration 10: Bivariate Analysis (Categorical VS Categorical)

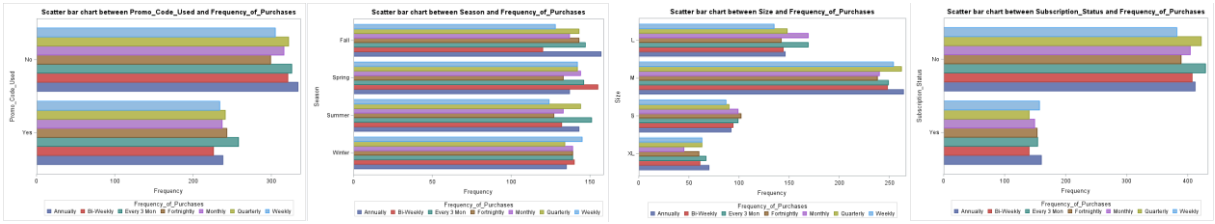


Key Insights:

- ❑ Frequency of purchases of Clothing are made mostly monthly, every three months, and annually.
- ❑ The highest frequency of purchases where discounts have not been applied is observed annually and every three months.
- ❑ Male Make most purchases every three months, quarterly and annually.

- ❑ Cash is the most widely used method of payment, with a high frequency of use every three months, followed by credit cards, which are used every month.

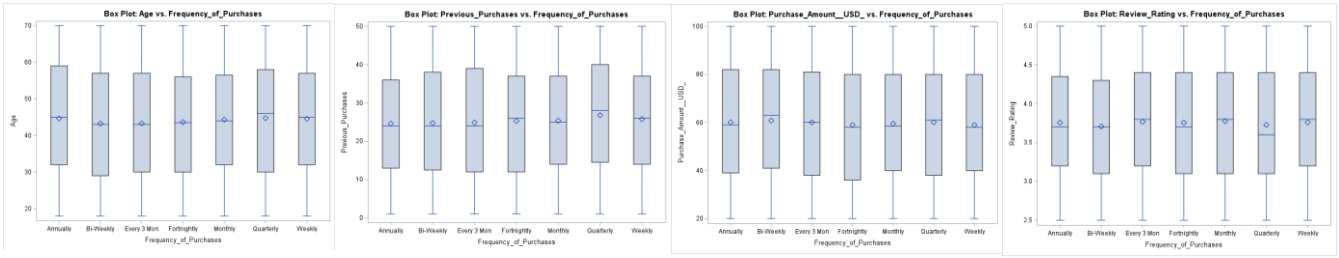
Illustration 11: Bivariate Analysis (Categorical VS Categorical)



Key Insights:

- ❑ The highest frequency of purchases where promo_code have not been used is observed annually and every three months.
- ❑ Frequency of purchases are highest in the Fall and Spring season on an annual and Bi-weekly basis.
- ❑ Frequency of purchases of products with size M are observed mostly annually, quarterly and weekly.
- ❑ The highest frequency of purchases where subscription status are not observed are made every three months, quarterly and annually.

Illustration 12: Bivariate Analysis (Categorical VS Continuous)



Key Insights:

- ❑ 50% of Customers aged around 48 make the most purchases quarterly and annually.
- ❑ 50% of previous purchases over $28 are made quarterly.
- ❑ 50% of purchases over $60 are made bi-weekly and quarterly.
- ❑ 50% of Review rating with 3.8 are mostly observed with the frequency of purchase every three months, monthly, and Weekly.

Model Presentation

- **The Target (Frequency_of_purchases)** can be identified as a multinomial classification whether the customer will make purchases Fortnightly, weekly, Bi-weekly, monthly, every three month, quarterly or annually.
- To enhance our model, we needed to categorize the multinomial classification into binomial classifications:

  - **Regular** :("Weekly", "Bi-Weekly", "Monthly")

  - **Irregular** : ("Quarterly", "Every 3 Mon", "Annually", "Fortnightly")
- We employed logistic regression **PROC LOGISTIC** to construct a prediction model as our target variable is categorical.

For this stage, we will apply the following steps:

- ✓ Feature Engineering by using one hot encoding for categorical variables (Gender, Promo_Code_Used, Subscription_Status), and categorization of feature with more than 6 levels (Location, color, shipping type …);
- ✓ Splitting test and training sets for modelling using proc surveyselect with a rate of 70%.
- ✓ Since we don't have multicollinearity except between payment method and Preferred_Payment_Method, we will exclude one of them and use all the others features.

Logistic Regression Output

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	4865.962	5061.265
SC	4872.144	6353.186
-2 Log L	4863.962	4643.265

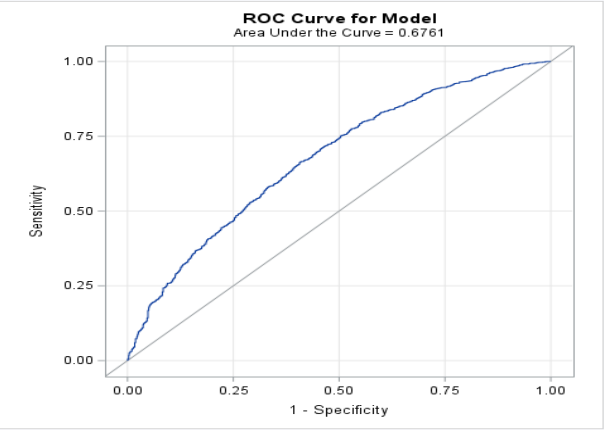
- ❑ the AIC for the model with both intercept and covariates (predictors) is 5061.265, while the AIC for the intercept-only model is 4865.962. The model with lower AIC is preferred, so adding covariates has increased the AIC, suggesting that the model with covariates might not be significantly better than the intercept-only model.

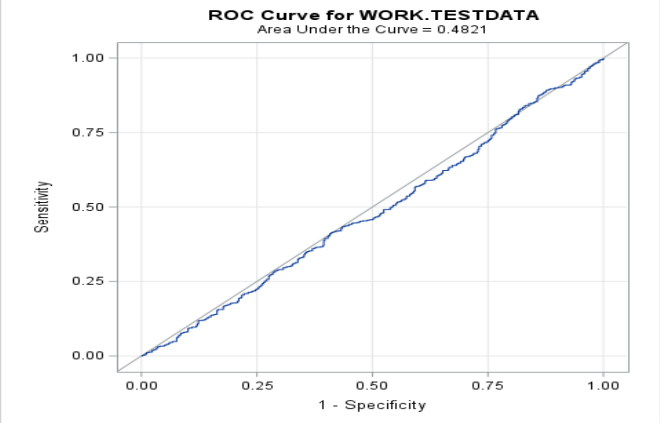
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	220.6973	208	0.2602
Score	214.1227	208	0.3707
Wald	201.9427	208	0.6053

- From this table, we can see that the Likelihood ratio chi-squared value of 220.6973 with a corresponding p-value of 0.2602. This P-value is greater than 5%, then we don't have enough evidence to says that the logistic regression model is not statistical significant.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	64.5	Somers' D	0.289
Percent Discordant	35.5	Gamma	0.289
Percent Tied	0.0	Tau-a	0.141
Pairs	3112713	c	0.645

- **C-Statistic:(0.645)** indicates moderate discriminative ability. The model is fairly effective in distinguishing between the two levels (regular and irregular) based on predicted probabilities.
- **Somers' D (0.289)**: its suggests a moderate positive association between predicted probabilities and observed responses





Partition for the Hosmer and Lemeshow Test					
		Purchase_Category = Irregular		Purchase_Category = Regular	
Group	Total	Observed	Expected	Observed	Expected
1	249	72	72.67	177	176.33
2	249	99	102.02	150	146.98
3	249	114	118.19	135	130.81
4	249	136	129.74	113	119.26
5	249	149	140.00	100	109.00
6	249	152	150.11	97	98.89
7	249	156	160.37	93	88.63
8	249	170	171.17	79	77.83
9	249	185	184.32	64	64.68
10	246	197	201.41	49	44.59

Sensitivity and Specificity				
Statistic	Estimate	Standard Error	95% Confidence Limits	
Sensitivity (Recall)	0.3816	0.0256	0.3314	0.4319
Specificity (True Negative Responders)	0.5755	0.0183	0.5396	0.6115
Positive Predictive Value	0.3072	0.0218	0.2644	0.3500
Negative Predictive Value	0.6537	0.0188	0.6168	0.6905

Table of F_Purchase_Category by I_Purchase_Category			
F_Purchase_Category (From: Purchase_Category)	I_Purchase_Category (Into: Purchase_Category)		
Frequency Percent Row Pct Col Pct	Regular	Irregular	Total
Regular	137 12.60 30.72 38.16	309 28.43 69.28 42.45	446 41.03
Irregular	222 20.42 34.63 61.84	419 38.55 65.37 57.55	641 58.97
Total	359 33.03	728 66.97	1087 100.00
Frequency Missing = 83			

- We can notice that, based on Hosmer and Lemeshow test, the observed and expected frequencies are consistently close across groups, and this suggests a good fit for the model.
- **Sensitivity (recall)** for testdata: proportion of true positive responders (response = Regular) that have true results is 38.16%(the model correctly identified 38.16% of the actual "Regular" instances among the responders).
- **Specificity (TNR)**: True Negative Responders (response = Irregular) that have negative test result is 57.55% (the model correctly identified 57.55% of the actual "Irregular" instances as negative).

Executive Summary

The document "Customer Shopping Preferences" provides a comprehensive analysis of a synthetic dataset designed for beginners in data analysis and machine learning. It includes 3903 observations across 19 features, focusing on various aspects of customer shopping behavior such as purchase history, payment methods, item preferences, and demographic details. Key objectives include understanding product popularity, forecasting demand, and building predictive models for future trends. The methodology involves data cleaning, exploratory analysis, feature engineering, and logistic regression for predictive modeling. The analysis reveals insights like the popularity of clothing items, payment preferences, and seasonal trends in shopping. The document concludes with business recommendations for inventory optimization, personalized product suggestions, and supply chain management to align with customer preferences and purchasing patterns.

Business Recommendations

- **Inventory Optimization**: Optimize inventory levels for best-selling clothing items to meet the demands of the most loyal and frequent customers.
- **Product Recommendations**:Implement personalized product recommendations based on age group and past purchase behavior. Leverage predictive models to anticipate future trends and tailor product offerings.

- **Supply Chain Management**: Collaborate with suppliers and optimize the supply chain to ensure the availability of popular products and meet customer demands during peak seasons