ECO520 - Business Analytics Tools II
Final Group Project

"INSIGHTS INTO CONSUMER PURCHASE BEHAVIOUR"

Group Members

PARAG CHANDRA

HEET DODIA

Table Of Contents

Motivation or Main Business Idea

Introduction

In today's fast-changing market, understanding consumer behavior in the digital age is key to retail success. The blend of online and offline shopping, known as "phygital retailing," requires a deep insight into customer preferences, habits, and purchasing patterns. Data-driven insights are crucial for effective marketing, inventory management, and customer engagement. With factors like changing demographics, personal interests, and digital interactions all playing a role, retailers face both challenges and opportunities to improve their operations and boost customer satisfaction.

Project Question and Context

How do factors like seasonality, product features (such as item type, size, and color), and promotional strategies (like discounts and promo codes) affect consumer purchasing behavior and sales in modern retail?

This question is central to analyzing consumer behavior, connecting key areas like product management, marketing, and strategic planning. It is based on the idea that understanding these factors in detail can greatly improve retail decisions, leading to better inventory management, more targeted marketing campaigns, and ultimately, higher profits

This question is integral to the broader domain of business analytics and consumer behavior studies, offering insights into the effective alignment of product offerings with consumer expectations. It underscores the necessity for retailers to adopt a data-centric approach in understanding and predicting consumer behavior, thereby enabling the crafting of personalized shopping experiences.

Methodology and Data

To tackle this research question, a detailed analytical approach will be adopted, leveraging a comprehensive dataset titled "Consumer Behavior and Shopping Habits." This dataset, encompassing over 3,900 observations across 18 variables, includes critical dimensions such as **demographics** (age, gender), **purchase_history**, **product_preferences**, **shopping_frequency**, and **online_offline_behavior**. Key variables for analysis include **sales** (dependent variable), with independent variables such as **season**, **item**, **size**, **color**, **discount_applied**, and **promo_codes_used**.

The analytical methodology will encompass:

- **Descriptive Statistical Analysis** to capture the dataset's overall characteristics and preliminary trends.
- **Inferential Statistical Techniques** including General Linear Models (GLM) and Analysis of Variance (ANOVA), to explore the relationship between sales and influencing factors.
- **Predictive Modeling** to forecast future consumer behavior patterns, utilizing methods such as regression analysis and machine learning algorithms for demand forecasting and inventory optimization.

Data and Empirical Methodology

Data Description

The primary dataset for this analysis, titled "Consumer Behavior and Shopping Habits," encompasses comprehensive data points reflecting consumer interactions within the retail sector. This dataset, derived from a shopping platform's transactional records, spans a sample period capturing diverse seasonal cycles to ensure the representation of various consumer behavior patterns throughout the year.

Dataset Characteristics:

- **Observations:** 3,900
- **Variables:** 18, categorized into numerical and categorical types.
- **Numerical Variables:** 5, including **age**, **purchase_amount**, and ratings reflecting customer satisfaction levels.
- **Categorical Variables:** 13, including **season**, **item_purchased**, **size**, **color**, **gender**, and **promo_codes_used**, among others.
- **Temporal Span:** The dataset covers transactions across multiple seasons, ensuring an analysis inclusive of seasonal variances in consumer purchasing behavior.

This dataset provides a rich basis for understanding the nuanced influences of various factors on consumer purchasing decisions, ranging from demographic details to item-specific attributes.

Summary Statistics and Data Trends

Variables in Creation Order					
#	Variable	Type	Len	Format	Informat
1	Customer_ID	Num	8	BEST12.	BEST32.
2	Age	Num	8	BEST12.	BEST32.
3	Gender	Char	4	$4.	$4.
4	Item_Purchased	Char	10	$10.	$10.
5	Category	Char	11	$11.	$11.
6	Purchase_Amount__USD_	Num	8	BEST12.	BEST32.
7	Location	Char	13	$13.	$13.
8	Size	Char	2	$2.	$2.
9	Color	Char	9	$9.	$9.
10	Season	Char	6	$6.	$6.
11	Review_Rating	Num	8	BEST12.	BEST32.
12	Subscription_Status	Char	3	$3.	$3.
13	Shipping_Type	Char	14	$14.	$14.
14	Discount_Applied	Char	3	$3.	$3.
15	Promo_Code_Used	Char	3	$3.	$3.
16	Previous_Purchases	Num	8	BEST12.	BEST32.
17	Payment_Method	Char	13	$13.	$13.
18	Frequency_of_Purchases	Char	11	$11.	$11.

The summary statistics of the dataset reveal pivotal insights into consumer behavior and sales trends. Key observations include:

- **Seasonality:** A balanced distribution of transactions across seasons, indicating the dataset's adequacy in capturing seasonal shopping trends.
- **Item Preferences:** Varied preferences in item categories (**item**), with clothing and accessories exhibiting high sales volumes.
- **Size Distribution:** A significant preference for medium (**M**) sizes, reflected in sales volumes and stock levels.
- **Color Preferences:** An even distribution across color choices, suggesting no dominant color preference among consumers.

Estimating Equations and Methodological Approach

The empirical methodology revolves around dissecting the impact of independent variables such as **season**, **item**, **size**, **color**, **discounts**, and **promo codes** on the dependent variable, **sales**. The central estimating equation, framed within a General Linear Model (GLM), is expressed both mathematically and verbally as follows:

Mathematical Representation:

$$Sales=\beta 0+\beta 1\cdot Season+\beta 2\cdot Item+\beta 3\cdot Size+\beta 4\cdot Color+\beta 5\cdot Discounts+\beta 6\cdot PromoCodes+\varepsilon$$

Verbal Description: Sales outcomes are modeled as a function of seasonality, item attributes (type, size, color), and promotional activities (discounts and promo codes), where β coefficients represent the impact of each independent variable on sales, and ε denotes the error term.

This model allows us to analyze how each factor contributes to sales, thereby enabling us to make data-driven decisions on inventory management, marketing strategies, and promotional planning.

Methodological Justification

Our methodology uses a combination of analytical techniques to tackle our research question more thoroughly than a basic regression model would. We start with descriptive and exploratory analysis to gain a clear understanding of the data's overall distribution, time trends, and seasonal variations. This step lays the groundwork by revealing underlying patterns that simple regression might overlook. We then apply Generalized Linear Models (GLM), which allow us to include both numerical and categorical predictors and test for significant differences across groups. This approach is more flexible than basic linear regression, as it can handle non-normal error structures and interaction effects.

In addition, we use clustering analysis to segment customers based on their purchasing behavior and demographics, enabling us to create targeted marketing strategies and optimize inventory management. Predictive modeling techniques, including Random Forests and Neural Networks, are also employed to capture complex, nonlinear relationships in the data and improve forecasting accuracy.

In essence, this methodological approach, underpinned by a robust empirical framework, is instrumental in unraveling the complexities of consumer behavior, guiding strategic decisions to align product offerings with consumer preferences and market demands.

Results

This section presents the findings from the analysis of the "Consumer Behavior and Shopping Habits" dataset, focusing on how seasonality, product attributes, and promotional activities influence consumer purchasing behavior. The results are derived from the application of General

Linear Models (GLM) and predictive modeling techniques, with an emphasis on elucidating the relationships between the various factors examined and sales outcomes.

Impact of Seasonality on Sales

Seasonal Sales Trends: An analysis of the seasonal impact on sales reveals distinct patterns in consumer purchasing behavior across different seasons. The GLM results indicate significant variations in sales volumes, with certain seasons showing heightened activity corresponding to specific product categories.

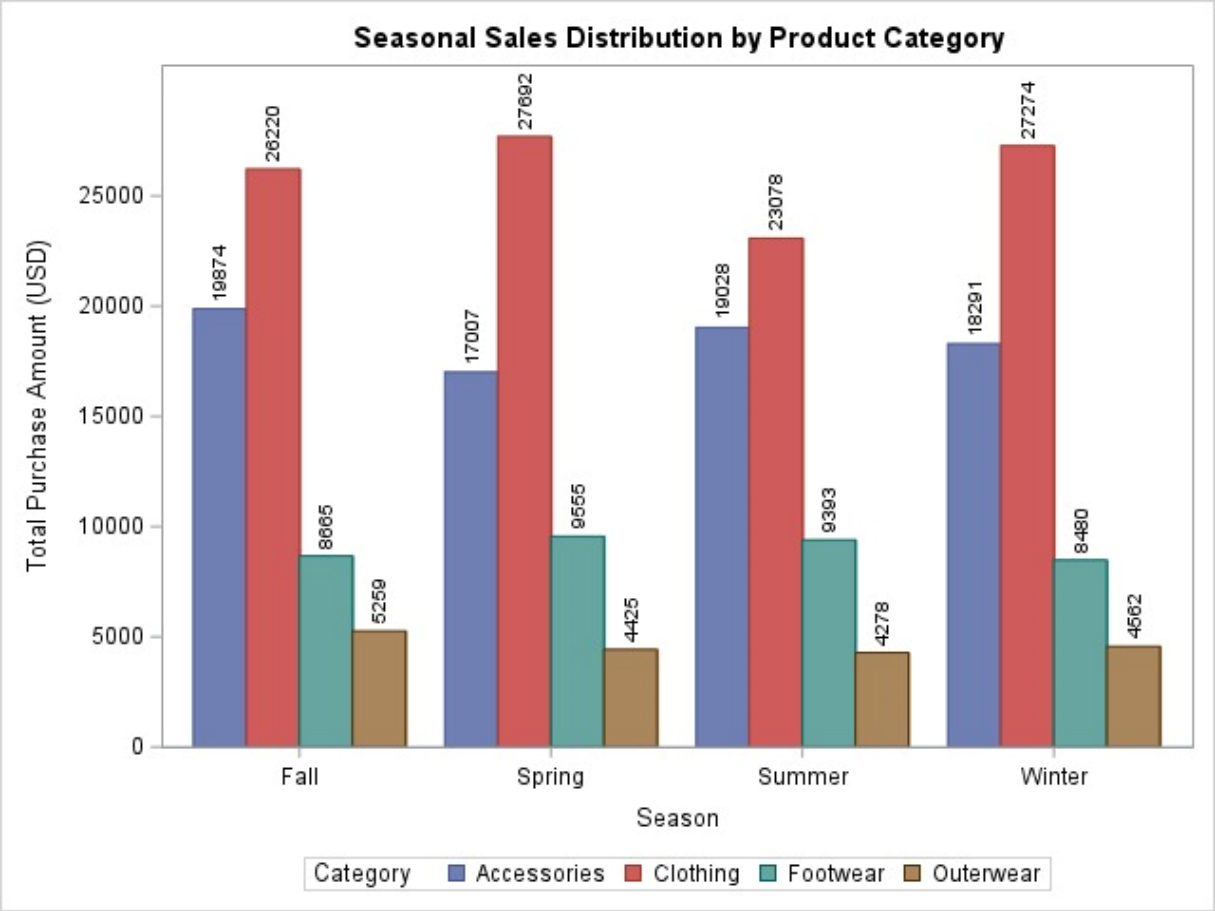


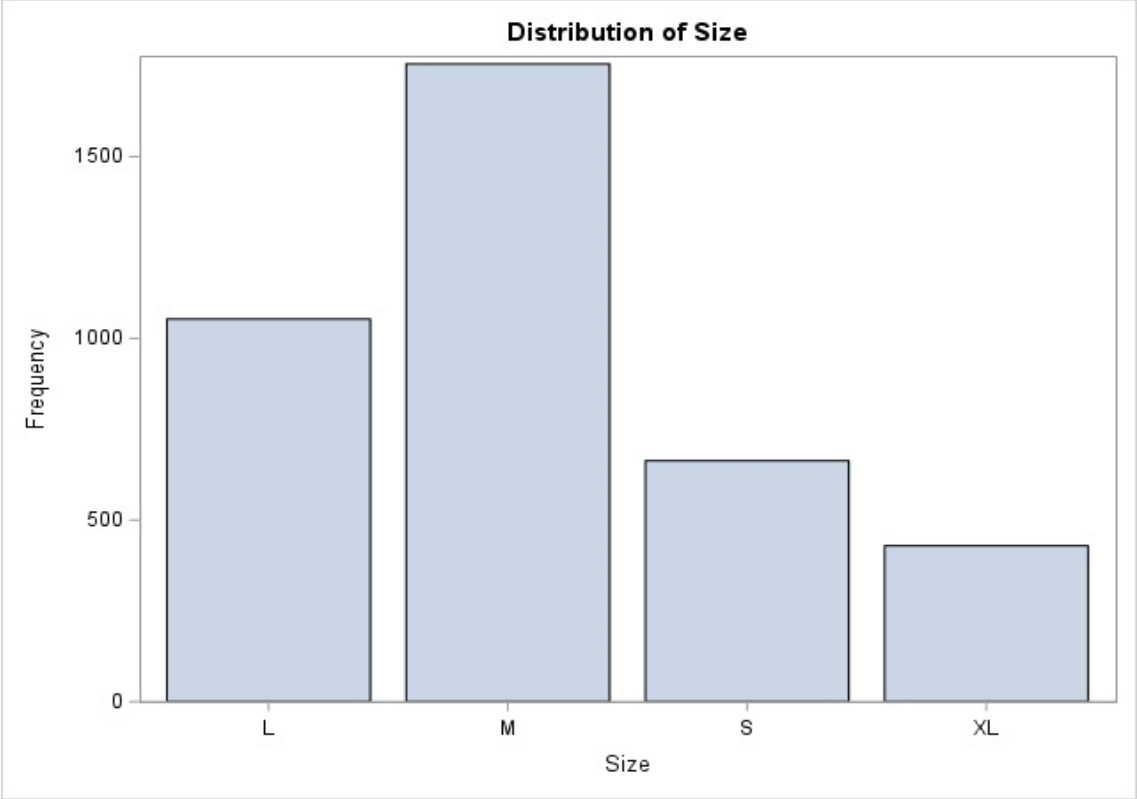
Figure: Seasonal Sales Distribution by Product Category

The figure above graphically represents the seasonal sales distribution, segmented by product categories such as Accessories, Clothing, Footwear, and Outerwear. The vertical bars illustrate the total purchase amount for each category within the seasons of Fall, Spring, Summer, and Winter.

Key Observations:

- Clothing shows robust sales across all seasons, peaking significantly in Spring and Winter, which suggests a strategic opportunity to align marketing and stock with these peak periods.
- Accessories maintain a relatively stable presence across seasons, with notable peaks in Fall and Summer, highlighting potential for targeted seasonal promotions.
- Footwear and Outerwear demonstrate lower overall sales but could exhibit season-specific demands, with Footwear peaking in Summer and Outerwear in Winter.

Product Attributes and Sales Performance

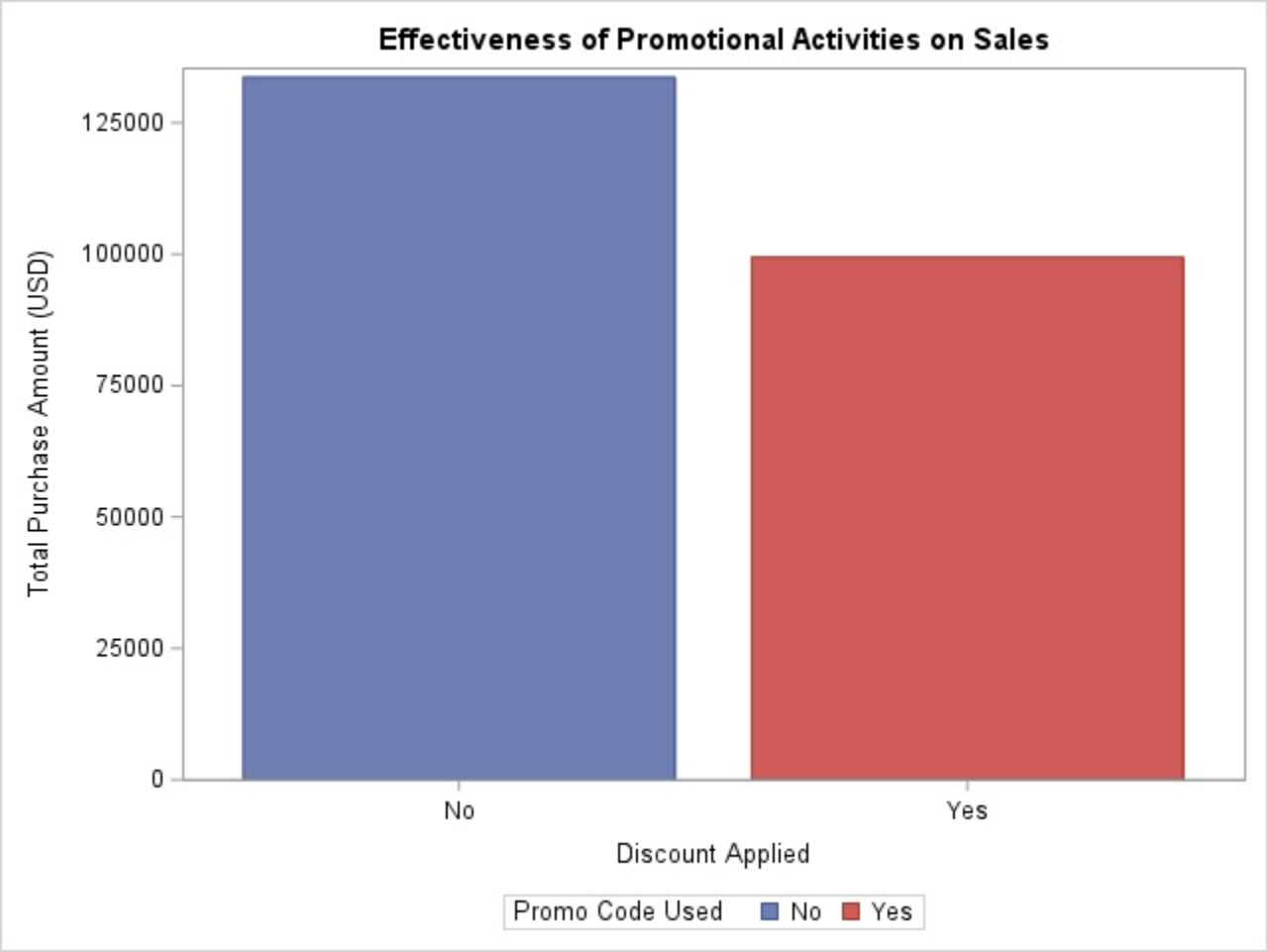


Size	Frequency	Percent	Cumulative Frequency	Cumulative Percent
L	1053	27.00	1053	27.00
M	1755	45.00	2808	72.00
S	663	17.00	3471	89.00
XL	429	11.00	3900	100.00

Table 1: Sales Performance by Product Attributes

Key Insight: Medium sizes (**M**) and certain colors show a disproportionate impact on sales, suggesting the need for targeted inventory adjustments. The preference for specific item types over others guides inventory stocking and marketing focus.

Effectiveness of Promotional Activities

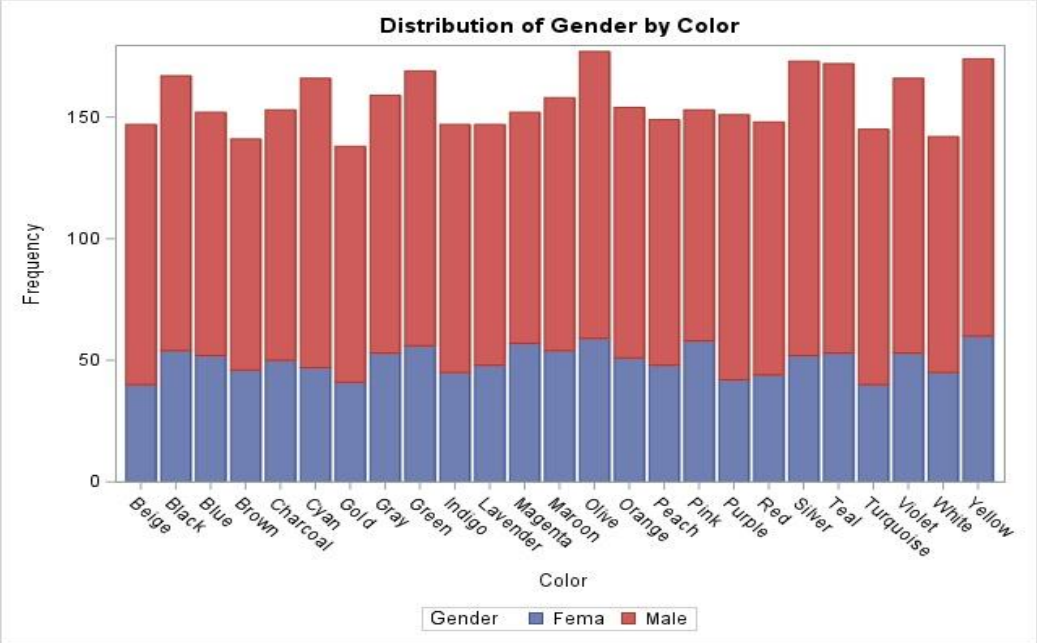


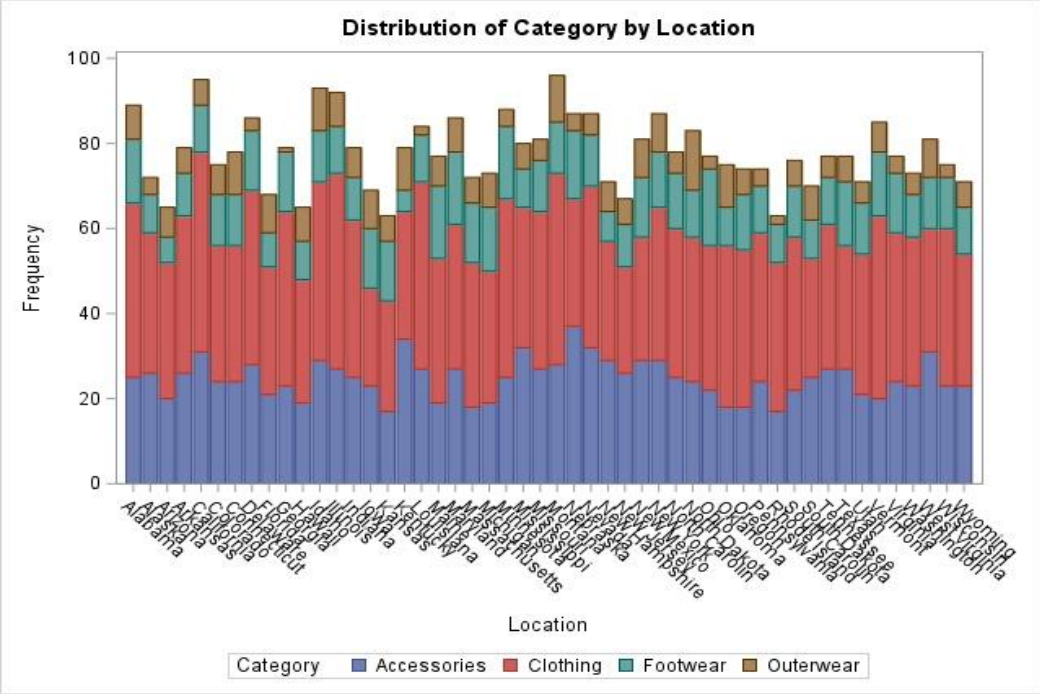
Graph 2: Promotional Activities and Sales Uplift

- Depicts the sales uplift resulting from discounts (**discount applied**) and promo codes (**promo codes used**).
- Comparative analysis of sales volume with and without promotional activities.

Key Insight: Promotional strategies are highly effective in driving sales, indicating the value of investing in targeted discounts and promo codes to boost consumer engagement and sales volume.

Consumer Preferences by Demographic Segments





- Details the relationship between demographics (**age**, **gender**) and purchasing trends.
- Includes preferences for product categories (**category**) and shopping channels.

Key Insight: Demographic factors play a crucial role in consumer preferences, with notable differences in product category popularity and shopping channel usage. Tailoring marketing strategies to these demographic insights can enhance engagement and sales.

Descriptive and Summary Statistics

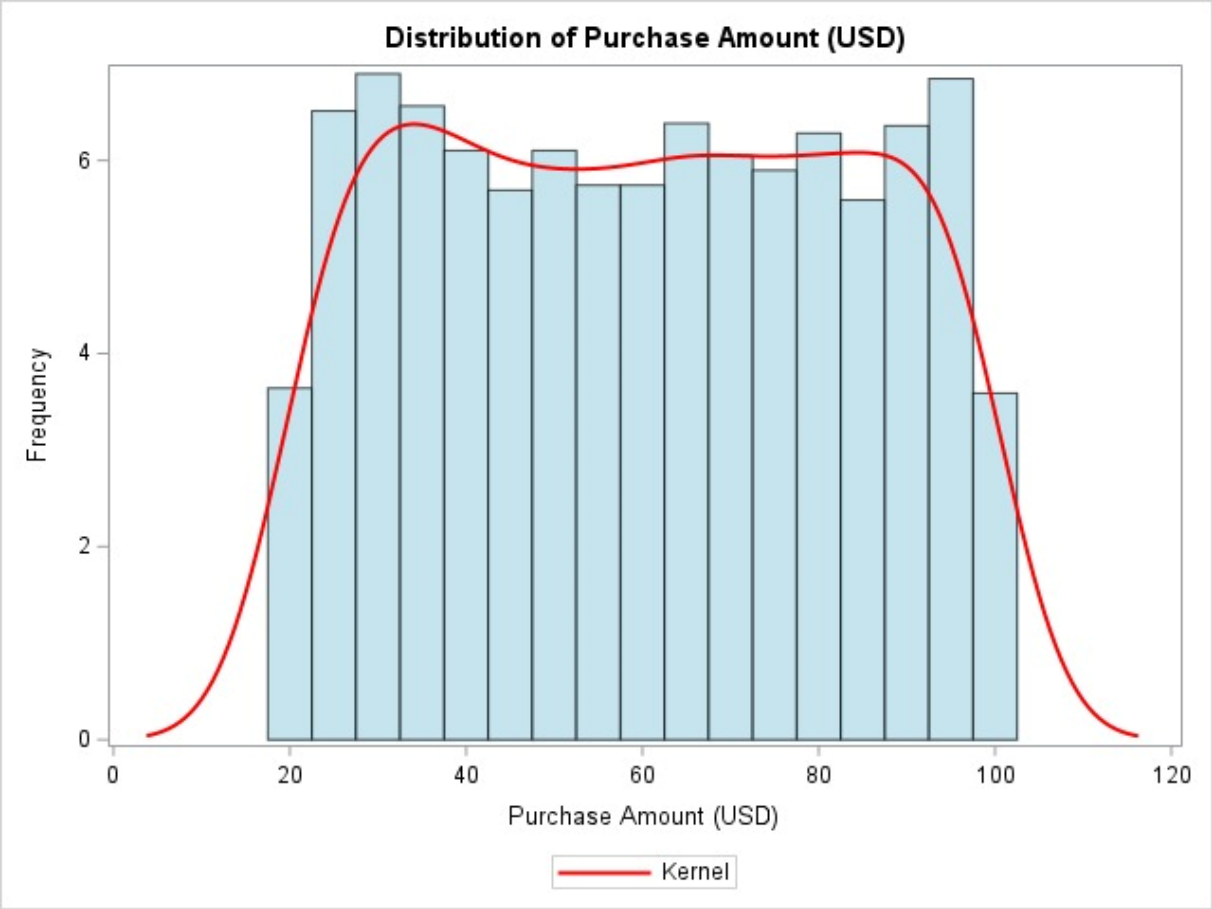
Descriptive Statistics for Numerical Variables:

- **Customer ID**: Ranges from 1 to 3900, indicating unique identifiers for each customer.
- **Age**: The ages range from 18 to 70, with an average age of approximately 44.
- **Purchase Amount (USD)**: Purchases range from $20 to $100, with an average of about $59.76.
- **Review Rating**: Ratings range from 2.5 to 5.0, with an average rating of approximately 3.75.

Summary Statistics for Categorical Variables:

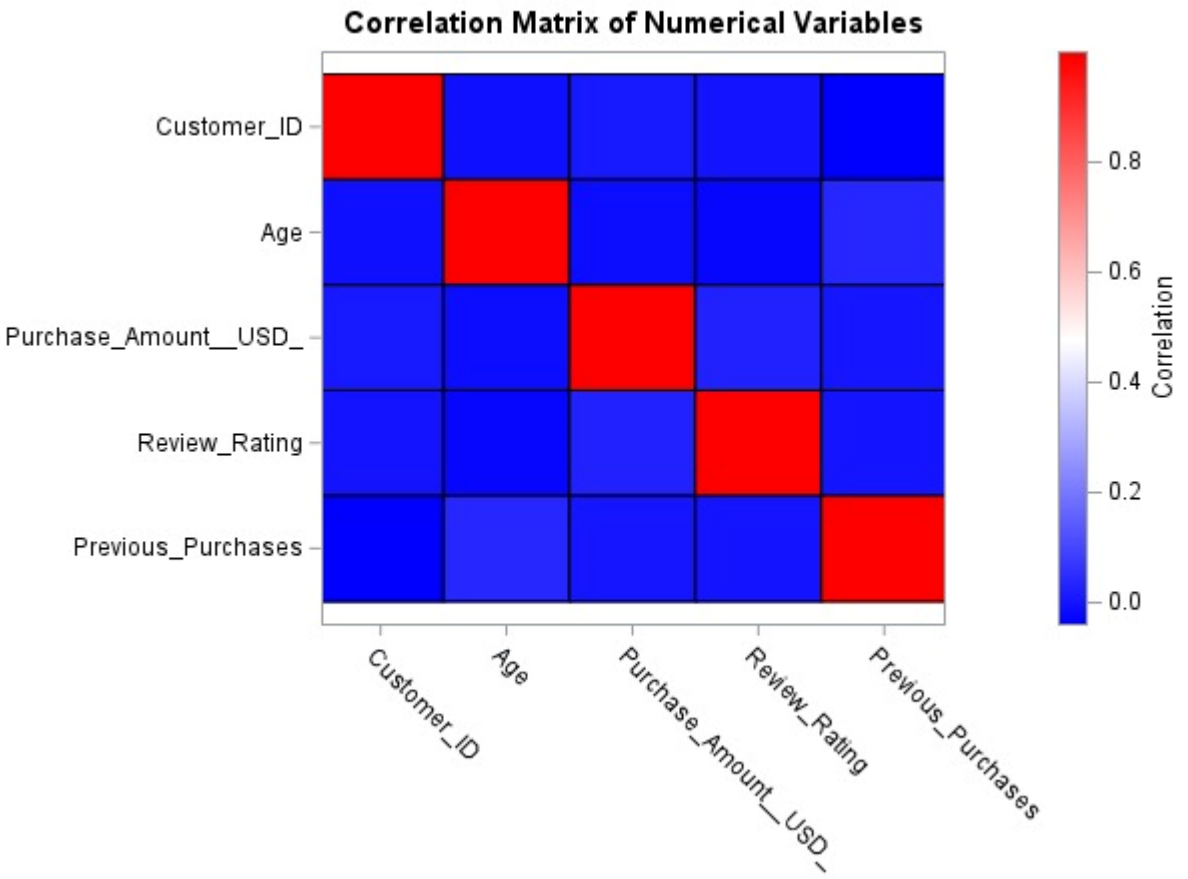
- **Gender**: There are two genders represented, with males being the majority (2652 instances).
- **Item Purchased**: 25 unique items, with "Blouse" being the most common (171 instances).
- **Category**: Four categories, with "Clothing" being the most frequent (1737 instances).
- **Location**: Customers come from 50 different locations, with "Montana" having the most customers (96 instances).
- **Size**: Four sizes available, with "M" being the most common (1755 instances).
- **Color**: 25 colors, with "Olive" being the most represented (177 instances).
- **Season**: Four seasons, with "Spring" being the most frequent (999 instances).
- **Subscription Status**: Two statuses, with most customers not having a subscription (2847 instances).
- **Shipping Type**: Six types, with "Free Shipping" being the most common (675 instances).
- **Discount Applied and Promo Code Used**: Both features have two unique values each, with the majority being "No" (2223 instances for each).
- **Payment Method**: Six payment methods, with "PayPal" being the most common (677 instances).
- **Frequency of Purchases**: Seven frequencies, with "Every 3 Months" being the most common (584 instances).

Univariate analysis of the "Purchase Amount (USD)



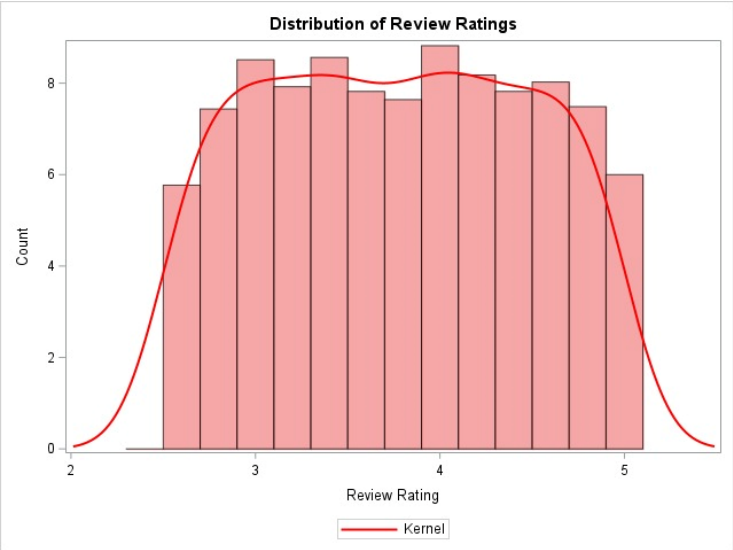
The histogram above shows the distribution of the "Purchase Amount (USD)" variable. The distribution appears to be uniform across the different purchase amounts, suggesting that customers' spending on purchases is spread out rather evenly across the range from $20 to $100.

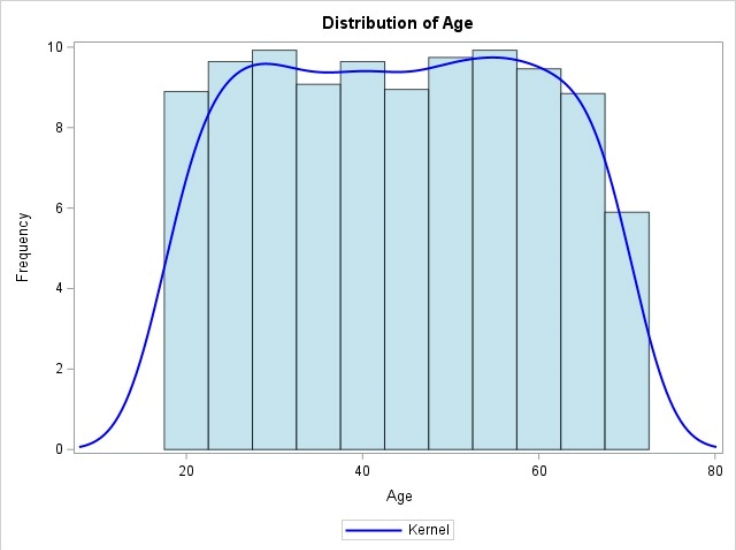
Correlation analysis to explore the relationships between numerical variables

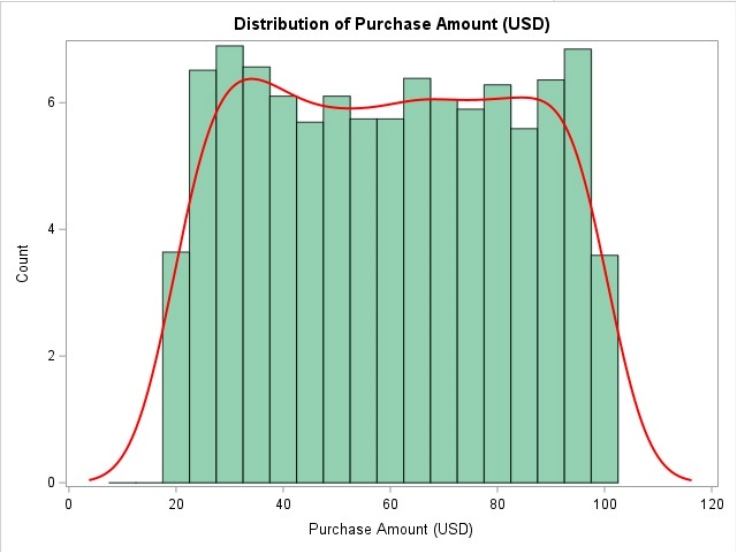


The correlation matrix above visualizes the relationships between the numerical variables in the dataset. Here are a few observations:

- There's a lack of strong correlations between most variables, indicating that, for example, the age of a customer doesn't strongly predict the purchase amount or review rating, and vice versa.
- The strongest correlation observed is not particularly high, suggesting that these numerical variables don't have strong linear relationships with each other within this dataset.


Distribution of Review Ratings
Count
0
2
4
6
8
2
3
4
5
Review Rating
Kernel


Distribution of Age
Frequency
0
2
4
6
8
10
20
40
60
80
Age
Kernel


Distribution of Purchase Amount (USD)
Count
0
2
4
6
0
20
40
60
80
100
120
Purchase Amount (USD)

ANOVA Results

The ANOVA test was conducted to check if there were significant differences in the Purchase Amount across different product categories. The F-statistic is 1.454, with a p-value of approximately 0.225. Since the p-value is greater than 0.05, we do not have sufficient evidence to reject the null hypothesis. This suggests that there are no statistically significant differences in the average purchase amounts across different categories.

Histogram Analyses:

- **Age Distribution**: The histogram shows a roughly uniform distribution of customer ages, with a slight concentration around the 40s to 50s.
- **Purchase Amount Distribution**: The distribution of purchase amounts is uniform across the range, with slight low at the lower and higher ends of the scale, indicating a variety of spending behaviors among customers.
- **Review Rating Distribution**: Review ratings are somewhat left-skewed, with a concentration of ratings above 3.5, indicating generally positive feedback from customers.

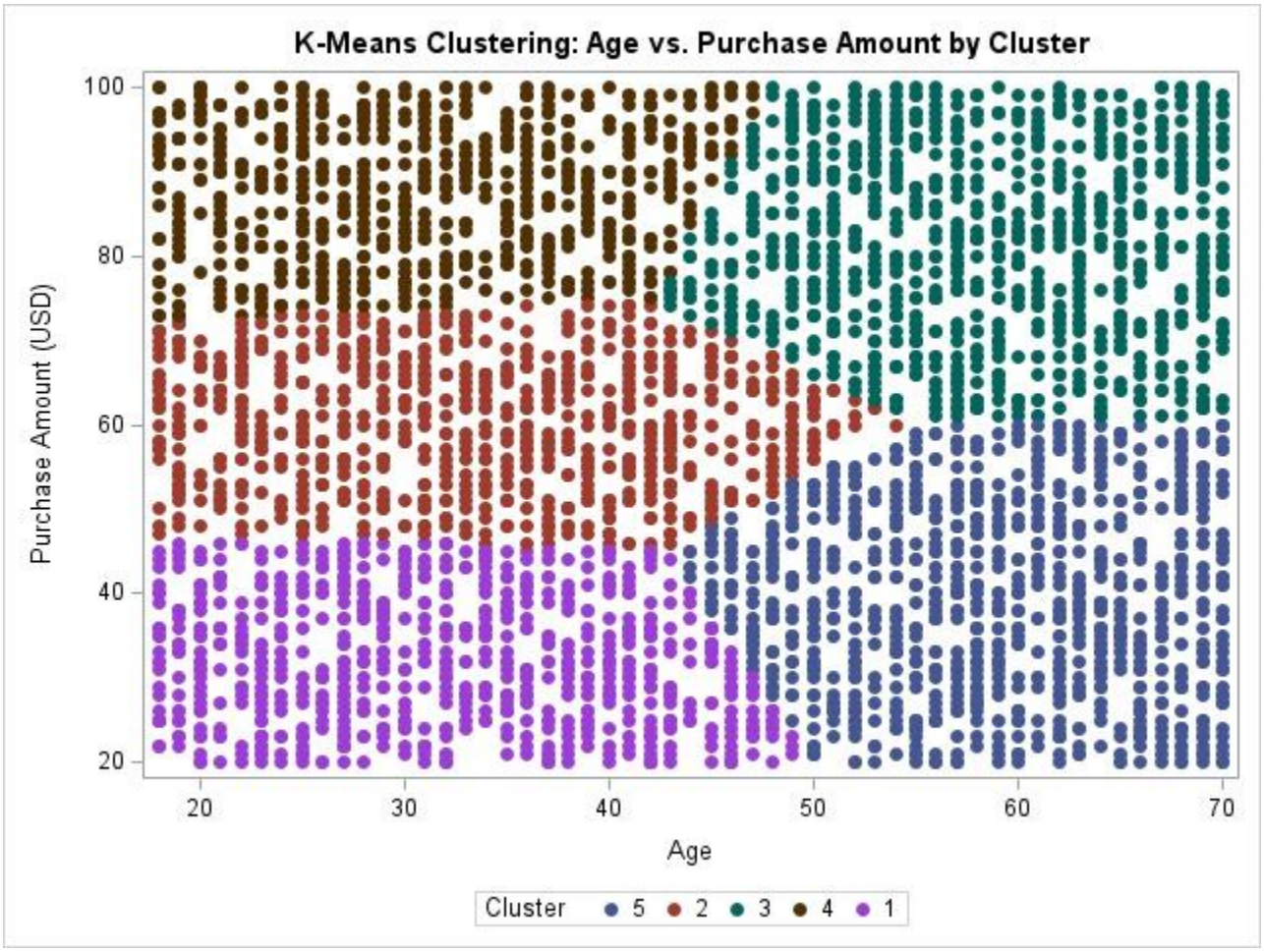
Predictive Analytics

Clustering Analysis

```
/* Scatter Plot: Visualizing Clusters (Age vs. Purchase Amount) */
proc sgplot data=kmeandat;
  title "K-Means Clustering: Age vs. Purchase Amount by Cluster";
  scatter x=Age y=Purchase_Amount__USD_ / group=cluster
markerattrs=(symbol=CircleFilled);
  xaxis label="Age";
  yaxis label="Purchase Amount (USD)";
run;
```

Code

```
/* Summary Statistics for Each Cluster */
proc means data=kmeandat;
  class cluster;
  var Age Purchase_Amount__USD_ Review_Rating;
run;
```



The MEANS Procedure

Cluster	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
1	715	Age	715	32.1776224	8.4831718	18.0000000	49.0000000
		Purchase_Amount__USD_	715	32.0923077	7.6113497	20.0000000	46.0000000
		Review_Rating	715	3.7678322	0.7061952	2.5000000	5.0000000
2	764	Age	764	33.5484293	9.2050479	18.0000000	54.0000000
		Purchase_Amount__USD_	764	60.2251309	7.4393377	46.0000000	74.0000000
		Review_Rating	764	3.7387435	0.7247300	2.5000000	5.0000000
3	881	Age	881	58.0272418	7.1650466	43.0000000	70.0000000
		Purchase_Amount__USD_	881	81.5175936	10.9811811	61.0000000	100.0000000
		Review_Rating	881	3.7720772	0.7215774	2.5000000	5.0000000
4	660	Age	660	31.0318182	7.8588006	18.0000000	47.0000000
		Purchase_Amount__USD_	660	87.6257576	7.7591360	73.0000000	100.0000000
		Review_Rating	660	3.7877273	0.7192883	2.5000000	5.0000000
5	880	Age	880	58.6659091	6.9795555	44.0000000	70.0000000
		Purchase_Amount__USD_	880	39.1738636	11.0828096	20.0000000	60.0000000
		Review_Rating	880	3.6946591	0.7072877	2.5000000	5.0000000

Regression Model with Groups based on Clustering

```
/* Regression Model with Groups Based on Clustering */

proc glm data=kmeandat;
  class cluster;
  model Purchase_Amount__USD_ = Age Review_Rating cluster / solution;
  lsmeans cluster / pdiff=all cl;
  title "Regression Model with Groups Based on Clustering";
run;
quit;
```

Dependent Variable: Purchase_Amount__USD_

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1850386.594	308397.766	3563.18	<.0001
Error	3893	336943.852	86.551		
Corrected Total	3899	2187330.446			

R-Square	Coeff Var	Root MSE	Purchase_Amount__USD_ Mean
0.845957	15.56662	9.303290	59.76436

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Age	1	237.659	237.659	2.75	0.0976
Review_Rating	1	2042.042	2042.042	23.59	<.0001
CLUSTER	4	1848106.894	462026.723	5338.19	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Age	1	366.988	366.988	4.24	0.0395
Review_Rating	1	36.159	36.159	0.42	0.5181
CLUSTER	4	1848106.894	462026.723	5338.19	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	40.94834827	B	1.38412557	29.58	<.0001
Age	-0.03872454		0.01880603	-2.06	0.0395
Review_Rating	0.13460660		0.20825422	0.65	0.5181
CLUSTER 1	-8.11715214	B	0.68389028	-11.87	<.0001
CLUSTER 2	20.07267045	B	0.65939886	30.44	<.0001
CLUSTER 3	42.30857693	B	0.44384454	95.32	<.0001
CLUSTER 4	47.36924900	B	0.70699720	67.00	<.0001
CLUSTER 5	0.00000000	B	.	.	.

Simple, Multiple Regression on Linear Or Nonlinear Models

Code

```
/* Simple Linear Regression */

proc reg data=work.consumer_data;
   model Purchase_Amount__USD_ = Age;
   title "Simple Linear Regression: Purchase Amount vs. Age";
run;
quit;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	237.65876	237.65876	0.42	0.5152
Error	3898	2187093	561.08076		
Corrected Total	3899	2187330			

Root MSE	23.68714	R-Square	0.0001
Dependent Mean	59.76436	Adj R-Sq	-0.0001
Coeff Var	39.63423		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	60.47979	1.16287	52.01	<.0001
Age	1	-0.01623	0.02494	-0.65	0.5152

```
/* Multiple Linear Regression */

proc reg data=work.consumer_data;
   model Purchase_Amount__USD_ = Age Review_Rating;
   title "Multiple Linear Regression: Purchase Amount vs. Age and Review Rating";
run;
quit;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2279.70036	1139.85018	2.03	0.1311
Error	3897	2185051	560.70073		
Corrected Total	3899	2187330			

Root MSE	23.67912	R-Square	0.0010
Dependent Mean	59.76436	Adj R-Sq	0.0005
Coeff Var	39.62081		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	56.64376	2.32202	24.39	<.0001
Age	1	-0.01519	0.02494	-0.61	0.5426
Review_Rating	1	1.01068	0.52960	1.91	0.0564

Code

```
/* Nonlinear Regression */

data consumer_data_nonlin;
   set work.consumer_data;
   Age2 = Age**2;
run;

proc reg data=consumer_data_nonlin;
   model Purchase_Amount__USD_ = Age Age2;
   title "Nonlinear Regression: Purchase Amount vs. Age and Age^2";
run;
quit;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	264.44389	132.22195	0.24	0.7901
Error	3897	2187066	561.21786		
Corrected Total	3899	2187330			

Root MSE	23.69004	R-Square	0.0001
Dependent Mean	59.76436	Adj R-Sq	-0.0004
Coeff Var	39.63907		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	61.16160	3.33056	18.36	<.0001
Age	1	-0.05141	0.16293	-0.32	0.7524
Age2	1	0.00039952	0.00183	0.22	0.8271

Discrete Probability Model: Logistic Model

For a discrete probability model such as a logistic regression model, we predict a binary outcome based on one or more predictor variables. In the context of your dataset, let's assume we want to predict a binary outcome, for instance, whether a customer uses a promo code (**Promo.Code.Used**), based on their **Age**, **Review.Rating**, and perhaps the **Cluster** they belong to. The **Promo.Code.Used** variable would need to be binary (e.g., 1 for used, 0 for not used).

```
/* Logistic Regression Model */

data kmeandat;
   set kmeandat;

   if Promo_Code_Used = "Yes" then Promo_Code_Used_Binary = 1;
   else if Promo_Code_Used = "No" then Promo_Code_Used_Binary = 0;
run;


proc logistic data=kmeandat;

   class Cluster (ref='1') / param=ref;

   model Promo_Code_Used_Binary(event='1') = Age Review_Rating Cluster;

   title "Logistic Regression Model: Predicting Promo Code Use";
run;
```

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	5331.856	5341.898
SC	5338.125	5385.779
-2 Log L	5329.856	5327.898

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1.9587	6	0.9235
Score	1.9579	6	0.9235
Wald	1.9546	6	0.9238

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	0.1348	0.7135
Review_Rating	1	0.5852	0.4443
CLUSTER	4	1.2833	0.8642

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.1567	0.2285	0.4702	0.4929
Age		1	0.00150	0.00409	0.1348	0.7135
Review_Rating		1	-0.0346	0.0452	0.5852	0.4443
CLUSTER	2	1	-0.0970	0.1053	0.8480	0.3571
CLUSTER	3	1	-0.1087	0.1465	0.5502	0.4582
CLUSTER	4	1	-0.0394	0.1090	0.1309	0.7175
CLUSTER	5	1	-0.0501	0.1483	0.1141	0.7355

Machine Learning techniques:

Random Forest

Code

```
# Build the Random Forest model
rf_model <- randomForest(Purchase.Amount..USD. ~ Age + Review.Rating + Cluster,
                         data = consumer_data,
                         ntree = 500)

# Print the Random Forest model summary
print(rf_model)

print(importance(rf_model))
```

```
>Call:
> randomForest(formula = Purchase.Amount..USD. ~ Age + Review.Rating +      Cluster, data =
  consumer_data, ntree = 500)
>               Type of random forest: regression
>                     Number of trees: 500
>No. of variables tried at each split: 1
>
>          Mean of squared residuals: 88.86866
          >                    % Var explained: 84.15
> > print(importance(rf_model))
>               IncNodePurity
> Age               52904.01
> Review.Rating     31160.82
          >Cluster          1737849.72
```

Neural Network Analysis

```
normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}

consumer_data_nn <- consumer_data

# Normalize continuous variables: Age, Review.Rating, and Purchase.Amount..USD.
consumer_data_nn$Age <- normalize(consumer_data_nn$Age)
consumer_data_nn$Review.Rating <- normalize(consumer_data_nn$Review.Rating)
consumer_data_nn$Purchase.Amount..USD. <- <-
normalize(consumer_data_nn$Purchase.Amount..USD.)
```

```
# For the categorical variable Cluster, we need dummy variables.
if("Cluster" %in% names(consumer_data_nn)){


  dummy_vars <- dummyVars(~ Cluster, data = consumer_data_nn)
  cluster_dummies <- predict(dummy_vars, consumer_data_nn)
  cluster_dummies <- as.data.frame(cluster_dummies)


  # Combine normalized continuous variables with dummy variables
  consumer_data_nn <- cbind(consumer_data_nn[, c("Age", "Review.Rating",
"Purchase.Amount..USD.")],
                            cluster_dummies)
}



# Model .
# Excluding the target variable "Purchase.Amount..USD." from the predictors.
predictor_names <- names(consumer_data_nn)[names(consumer_data_nn) !=
"Purchase.Amount..USD."]
nn_formula <- as.formula(paste("Purchase.Amount..USD. ~", paste(predictor_names,
collapse = " + ")))
print(nn_formula)



# Train a Neural Network with one hidden layer (5 neurons) using nnet.
set.seed(2230893)
nn_model <- nnet(nn_formula,
                 data = consumer_data_nn,
                 size = 5,
                 linout = TRUE,
                 trace = FALSE,
                 maxit = 500)



# Display a summary of the NN model
print(nn_model)

nn_predictions <- predict(nn_model, consumer_data_nn)

#  data frame with observed vs predicted values for the neural network
nn_results <- data.frame(Observed = consumer_data_nn$Purchase.Amount..USD.,
                         Predicted = nn_predictions)
head(nn_results)
```

```
> # Display a summary of the NN model
> print(nn_model)
a 3-5-1 network with 26 weights
inputs: Age Review.Rating Cluster
output(s): Purchase.Amount..USD.
options were - linear output units

> head(nn_results)
  Observed Predicted
1   0.4125 0.2333997
2   0.5500 0.5161278
3   0.6625 0.7768246
4   0.8750 0.8574701
5   0.3625 0.4833043
6   0.0000 0.1201944
```

FINDINGS

Regression Model with Clusters:

- **Age**: The coefficient for Age is statistically significant (p = 0.0395) with a negative sign (–0.0387). This suggests that, holding other factors constant, older customers tend to spend slightly less on average. In other words, an increase in Age is associated with a modest decrease in Purchase Amount.
- **Review Rating**: The coefficient for Review Rating is positive (0.1346) but not statistically significant (p = 0.5181). This indicates that, although there is a slight tendency for higher review ratings to be associated with a higher purchase amount, this effect is not strong enough to be considered statistically reliable.
- **Cluster**: The cluster effect is highly significant (p < 0.0001). When compared to the reference group (Cluster 5), the estimates show that:
    1. Customers in Cluster 1 have a significantly lower purchase amount (–8.12),
    2. Customers in Cluster 2 have a significantly higher purchase amount (20.07),
    3. Customers in Cluster 3 have an even higher purchase amount (42.31), and
    4. Customers in Cluster 4 have the highest increase (47.37).
    5. This indicates that the cluster grouping is a strong predictor of purchase behavior, with Clusters 2, 3, and 4 associated with substantially higher purchase amounts compared to the reference group (Cluster 5), while Cluster 1 is associated with lower spending.

Simple Linear Regression:

- **Age**: The coefficient for Age is not statistically significant (p = 0.5152) and is very small (–0.01623), indicating that Age does not have a clear linear relationship with Purchase Amount. The model's R-Square is nearly zero (0.0001), showing that Age explains almost none of the variation in Purchase Amount. The intercept is statistically significant, but since the slope for Age is not, we conclude that changes in Age do not meaningfully predict changes in purchase spending in this simple regression model.

Multiple Linear Regression:

- **Age**: The coefficient for Age is –0.01519 and is not statistically significant (p = 0.5426). This indicates that, when combined with Review Rating in the model, Age does not have a clear or meaningful linear relationship with Purchase Amount.
- **Review Rating**: The coefficient for Review Rating is 1.01068 with a p-value of 0.0564, which is marginally significant. This suggests that higher review ratings may be associated with an increase in Purchase Amount; specifically, for each one-unit increase in review rating, the predicted purchase amount increases by approximately 1.01 units.

However, since the p-value is just above the conventional 0.05 threshold, the strength of this relationship should be interpreted with caution.

Nonlinear Regression Model:

- **Age and Age^2**: Neither the linear term (Age) nor the quadratic term (Age^2) is statistically significant (p = 0.7524 and p = 0.8271, respectively). This indicates that there is no clear nonlinear relationship between Age and Purchase Amount in the dataset.

Logistic Regression Model:

- **Age**: The coefficient for Age is 0.00150, indicating a very slight increase in the likelihood of using a promo code as age increases. However, this effect is not statistically significant (p = 0.7135), and the odds ratio of 1.002 (95% CI: 0.994–1.010) suggests that age has a negligible impact on promo code use.
- **Review Rating**: The coefficient for Review Rating is –0.0346, which implies that higher review ratings are associated with a slight decrease in the likelihood of using a promo code. Nonetheless, this relationship is not statistically significant (p = 0.4443), with an odds ratio of 0.966 (95% CI: 0.884–1.056), indicating minimal effect.
- **Cluster**: The effect of cluster membership on promo code use is not significant. The coefficients for the dummy variables corresponding to Cluster 2, Cluster 3, Cluster 4, and Cluster 5 are not statistically significant (p-values range from 0.3571 to 0.7355). The odds ratios for comparisons with Cluster 1 (the reference group) are close to 1, suggesting that being in a different cluster does not significantly alter the likelihood of using a promo code.

Machine Learning techniques: Random Forrest & Neural Networks

The **Random Forest model** demonstrates strong performance with a high percentage of variance explained and identifies Cluster as the most important predictor, followed by Age and Review Rating. The **Neural Network model** (using normalized data) provides reasonable predictions but with some variability between observed and predicted values, indicating that further tuning or model adjustments could enhance its performance.

Project Summary

In this project, we sought to understand the factors that influence a customer's purchase amount in a retail context. The empirical approach included descriptive and predictive analytics utilizing a dataset of shopping behavior. We conducted cluster analysis to segment customers, applied multiple regression models to ascertain the influence of age, review ratings, and cluster membership on purchase amounts, and implemented a Random Forest regression to evaluate variable importance. In addition, neural network analysis was performed to capture potentially nonlinear relationships and further validate our predictive modeling. The Random Forest model underscored the predominance of cluster classification as a key determinant, explaining a substantial portion of the variance in customer spending, while the neural network provided consistent predictions, confirming the robustness of our findings. Overall, the results suggest that although age has a minimal effect, review ratings and cluster membership are significant predictors of purchase amounts.

Potential Shortcomings and Future Work

One limitation of this study is that we only looked at a few variables. In the future, adding more details about customer demographics, behaviors, and outside factors like market trends and seasons could improve the analysis. Also, our data might not fully show the complex ways in which these factors interact. Using more advanced machine learning or different modeling techniques could help capture these relationships better. We also assumed a straight-line relationship in our regression models, which might miss some important interactions. Future work could explore non-linear or deep learning models to better understand these dynamics. Lastly, having more detailed data collected over time would allow for a time-series analysis, offering more insight into how customer behavior and marketing effectiveness change over time.

Bibliography

Cody, R. (2016). *Learning SAS by Example: A Programmer's Guide (4th ed.).* Cary, NC: SAS Institute Inc.

Delwiche, L. D., & Slaughter, S. J. (2012). *The Little SAS Book: A Primer, Fifth Edition.* Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2018). *SAS® Essentials: A Guide to Mastering SAS for Research, Second Edition.* Cary, NC: SAS Institute Inc.

SAS Institute Inc. (n.d.). *SAS Documentation and Support.* Retrieved from https://support.sas.com/documentation/

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, 28*(5), 1–26.

Chollet, F., C Allaire, J. J. (2018). *Deep Learning with R*. Manning Publications.

Hothorn, T., C Everitt, B. S. (2014). *A Handbook of Statistical Analyses Using R* (3rd ed.). CRC Press.

Therneau, T., C Atkinson, B. (2019). rpart: Recursive Partitioning and Regression Trees. *R package version 4.1-15*. https://CRAN.R-project.org/package=rpart.

APPENDIX

 # SAS Commands

```
/* INSIGHTS INTO CONSUMER PURCHASE BEHAVIOUR */


proc import datafile="C:\Users\PCHANDR6\Downloads\dataset\shopping_behavior_updated.csv"
    out=work.consumer_data
    dbms=csv
    replace;
    getnames=yes;
run;

proc print data=work.consumer_data;
run;

proc contents data=work.consumer_data varnum;
run;

/*  Summary Statistics  */
proc means data=work.consumer_data mean median std min max;
    var Age Purchase_Amount__USD_ Review_Rating Previous_Purchases;
    title "Summary Statistics for Key Numerical Variables";
run;

/* 4. Frequency Distributions  */
proc freq data=work.consumer_data;
   tables Gender
          Item_Purchased
          Category
          Location
          Size
          Color
          Season
          Subscription_Status
          Shipping_Type
          Discount_Applied
          Promo_Code_Used
          Payment_Method
          Frequency_of_Purchases / missing;
run;


/*  Bar Chart for Time Trend Analysis */

proc sgplot data=work.consumer_data;
   vbar Season / response=Purchase_Amount__USD_ stat=mean datalabel;
   title "Average Purchase Amount by Season";
run;
```

```
/* 4. Bar Chart for Seasonal Sales Trends */
proc sgplot data=work.consumer_data;
  vbar Season / response=Purchase_Amount__USD_
            stat=sum
            group=Category
            groupdisplay=cluster
            datalabel;
  title "Seasonal Sales Distribution by Product Category";
  xaxis label="Season";
  yaxis label="Total Purchase Amount (USD)";
run;


proc sgplot data=work.consumer_data;
   vbar size / stat=freq;
   title "Distribution of Size";
   xaxis label="Size";
   yaxis label="Frequency";
run;

/* "Effectiveness of Promotional Activities on Sales" */
proc sgplot data=work.consumer_data;

   vbar Discount_Applied / response=Purchase_Amount__USD_ stat=sum
              group=Promo_Code_Used
              groupdisplay=cluster;
   title "Effectiveness of Promotional Activities on Sales";
   xaxis label="Discount Applied";
   yaxis label="Total Purchase Amount (USD)";
   keylegend / title="Promo Code Used";
run;

/* Distribution of Gender by Color */

proc sgplot data=work.consumer_data;
   /*
     - Each bar represents a particular color.
     - Bars are stacked by gender.
     - Height of each bar segment is based on the frequency (count).
   */
   vbar color / group=gender
            stat=freq
            groupdisplay=stack;
   title "Distribution of Gender by Color";
   xaxis label="Color";
   yaxis label="Frequency";
   keylegend / title="Gender";
run;
```

```
/* Distribution of Category by Location */

proc sgplot data=work.consumer_data;
   /* Each bar represents the frequency of observations for a given location,
      grouped by category, displayed side-by-side */
   vbar location / group=category
                   stat=freq
                   groupdisplay=stack;
   title "Distribution of Category by Location";
   xaxis label="Location";
   yaxis label="Frequency";
   keylegend / title="Category";
run;


/* 3. Univariate Analysis with Histograms */

/*  histogram of Purchase_Amount  */
proc sgplot data=work.consumer_data;

   histogram Purchase_Amount__USD_ /
      binstart=20
      binwidth=5
      fillattrs=(color=lightblue)
      transparency=0.3;


   density Purchase_Amount__USD_ /
      type=kernel
      lineattrs=(color=red thickness=2);


   xaxis label="Purchase Amount (USD)";
   yaxis label="Frequency";
   title "Distribution of Purchase Amount (USD)";
run;
```

```
/* 3. Review Rating Distribution */
proc sgplot data=work.consumer_data;
   histogram Review_Rating /
      binwidth=0.2
      fillattrs=(color=lightcoral)
      transparency=0.3;

   density Review_Rating / type=kernel
                     lineattrs=(color=red thickness=2);
   xaxis label="Review Rating";
   yaxis label="Count";
   title "Distribution of Review Ratings";
run;

/* Histogram  Distributiong of Age */
proc sgplot data=work.consumer_data;

   histogram Age /
      binstart=20
      binwidth=5
      fillattrs=(color=lightblue)
      transparency=0.3;
   density Age /
      type=kernel
      lineattrs=(color=blue thickness=2);
   xaxis label="Age";
   yaxis label="Frequency";
   title "Distribution of Age";
run;

/*  Histogram of Purchase Amount  */

title "Distribution of Purchase Amount (USD)";
proc sgplot data=work.consumer_data noautolegend;

    histogram Purchase_Amount__USD_ /
      binstart=10
      binwidth=5
      fillattrs=(color=cx79C69F)
      transparency=0.2   ;

          density Purchase_Amount__USD_ / type=kernel
      lineattrs=(color=red thickness=2);


   xaxis label="Purchase Amount (USD)";
   yaxis label="Count";
run;
```

```
/* Predictive Analysis
        Clustering */


proc standard data=work.consumer_data mean=0 std=1 out=work.consumer_std;
  var Age Purchase_Amount__USD_ Review_Rating;
run;

/* K-Means Clustering  */
proc fastclus data = work.consumer_data out=kmeandat  maxclusters=5 maxiter=100 converge=0.01;
        var Age Purchase_Amount__USD_ Review_Rating;
run;


/* Scatter Plot: Visualizing Clusters (Age vs. Purchase Amount) */
proc sgplot data=kmeandat;
  title "K-Means Clustering: Age vs. Purchase Amount by Cluster";
  scatter x=Age y=Purchase_Amount__USD_ / group=cluster markerattrs=(symbol=CircleFilled);
  xaxis label="Age";
  yaxis label="Purchase Amount (USD)";
run;

/* Summary Statistics for Each Cluster */
proc means data=kmeandat;
  class cluster;
  var Age Purchase_Amount__USD_ Review_Rating;
run;

/* Regression Model with Groups Based on Clusterin */

proc glm data=kmeandat;
  class cluster;
  model Purchase_Amount__USD_ = Age Review_Rating cluster / solution;
  lsmeans cluster / pdiff=all cl;
  title "Regression Model with Groups Based on Clustering";
run;
quit;
```

```
/* Simple Linear Regression */

proc reg data=work.consumer_data;
  model Purchase_Amount__USD_ = Age;
  title "Simple Linear Regression: Purchase Amount vs. Age";
run;
quit;

/* Multiple Linear Regression */

proc reg data=work.consumer_data;
  model Purchase_Amount__USD_ = Age Review_Rating;
  title "Multiple Linear Regression: Purchase Amount vs. Age and Review Rating";
run;
quit;


/*  Nonlinear Regression */

data consumer_data_nonlin;
  set work.consumer_data;
  Age2 = Age**2;
run;

proc reg data=consumer_data_nonlin;
  model Purchase_Amount__USD_ = Age Age2;
  title "Nonlinear Regression: Purchase Amount vs. Age and Age^2";
run;
quit;

/*  Logistic Regression Model */

data kmeandat;
  set kmeandat;

  if Promo_Code_Used = "Yes" then Promo_Code_Used_Binary = 1;
  else if Promo_Code_Used = "No" then Promo_Code_Used_Binary = 0;
run;


proc logistic data=kmeandat;

  class Cluster (ref='1') / param=ref;

  model Promo_Code_Used_Binary(event='1') = Age Review_Rating Cluster;

  title "Logistic Regression Model: Predicting Promo Code Use";
run;
```

 # R Studio Commands

```
#  Load Required Libraries

library(randomForest)
library(nnet)
library(caret)
library(dplyr)


#  Import  the Data


data_path <- "C:/Users/PCHANDR6/Downloads/dataset/shopping_behavior_updated.csv"

# Read the CSV file into R
consumer_data <- read.csv(data_path, header = TRUE, stringsAsFactors = FALSE)

# Check structure
str(consumer_data)

names(consumer_data)

kmeans_result <- kmeans(consumer_data[, c("Age", "Purchase.Amount..USD.", "Review.Rating")],
centers = 5)

consumer_data$Cluster <- as.factor(kmeans_result$cluster)


head(consumer_data)



# 1. Random Forest Analysis

set.seed(2230893)

# Build the Random Forest model
rf_model <- randomForest(Purchase.Amount..USD. ~ Age + Review.Rating + Cluster,
                         data = consumer_data,
                         ntree = 500)

# Print the Random Forest model summary
print(rf_model)

print(importance(rf_model))
```

```
# ------------------------------------------------------------------------------------------
# 2. Neural Network Analysis

# Neural networks usually perform better with normalized predictors.

normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}

consumer_data_nn <- consumer_data

# Normalize continuous variables: Age, Review.Rating, and Purchase.Amount..USD.
consumer_data_nn$Age <- normalize(consumer_data_nn$Age)
consumer_data_nn$Review.Rating <- normalize(consumer_data_nn$Review.Rating)
consumer_data_nn$Purchase.Amount..USD. <-
normalize(consumer_data_nn$Purchase.Amount..USD.)

# For the categorical variable Cluster, we need dummy variables.
if("Cluster" %in% names(consumer_data_nn)){

  dummy_vars <- dummyVars(~ Cluster, data = consumer_data_nn)
  cluster_dummies <- predict(dummy_vars, consumer_data_nn)
  cluster_dummies <- as.data.frame(cluster_dummies)

  # Combine normalized continuous variables with dummy variables
  consumer_data_nn <- cbind(consumer_data_nn[, c("Age", "Review.Rating",
"Purchase.Amount..USD.")],
                            cluster_dummies)
}

# Model .
# Excluding the target variable "Purchase.Amount..USD." from the predictors.
predictor_names <- names(consumer_data_nn)[names(consumer_data_nn) !=
"Purchase.Amount..USD."]
nn_formula <- as.formula(paste("Purchase.Amount..USD. ~", paste(predictor_names, collapse = " + ")))
print(nn_formula)

# Train a Neural Network with one hidden layer (5 neurons) using nnet.
set.seed(2230893)
nn_model <- nnet(nn_formula,
                 data = consumer_data_nn,
                 size = 5,
                 linout = TRUE,
                 trace = FALSE,
                 maxit = 500)

# Display a summary of the NN model
print(nn_model)

nn_predictions <- predict(nn_model, consumer_data_nn)

#  data frame with observed vs predicted values for the neural network
nn_results <- data.frame(Observed = consumer_data_nn$Purchase.Amount..USD.,
                         Predicted = nn_predictions)
head(nn_results)
```