International Journal of Economics, Commerce and Management
United Kingdom ISSN 2348 0386 Vol. 14, Issue 5, May 2026

https://ijecm.co.uk/

PREDICTIVE MODELING OF DISCOUNTS, PRODUCT CATEGORY, SEASONAL PATTERNS, AND RETAIL STORE SALES: AN EMPIRICAL APPROACH

Domnic Ondiwa
Masters in Decision Analytics- Accounting Analytics, Student,
Virginia Commonwealth University, Richmond, USA
ondiwadomnic@gmail.com; https://orcid.org/0009-0004-6774-0385

Simon Ondiwa
Department of Accounting and Finance, Maseno University, Kenya
sondiwa81@gmail.com; https://orcid.org/0000-0002-0521-5321

Abstract

This study presents a comprehensive statistical and predictive analysis of a retail transaction dataset containing 5,000 customer purchases. The analysis was structured around focus areas such as data preparation, exploratory visualization, confidence intervals and hypothesis testing, analysis of variance, non-parametric validation, linear regression, logistic regression, diagnostic checks, and predictive performance assessment. Results indicate that the average purchase amount was approximately $285.09, with a 95% confidence interval from $269.80 to $300.38. Contrary to the expectation that higher discounts would lead to greater customer spending, the Welch two-sample t-test found no statistically significant difference in spending between the high- and low-discount groups. Product category, however, emerged as the dominant driver of purchase amount. Analysis of variance revealed extremely large and statistically significant differences in purchase amounts across categories, with electronics far above all other product groups. Seasonal differences, by contrast, were negligible and statistically nonsignificant. Multiple linear regression explained roughly 81.8% of the variation in purchase amount, with the strongest positive effect associated with the electronics category, whereas cash on delivery was associated with lower purchase amounts. Logistic regression classified high-value purchases with near-

perfect accuracy and an AUC above 0.999, but the model also showed numerical signs of quasi-separation, indicating that the classification boundary is driven strongly by a few dominant predictors rather than a balanced contribution from all covariates. The study will be important to retail outlet managers and employees in making promotional decisions, and to scholars for purposes of continuous research.

Keywords: Retail transactions, Discounting, Customer spending, retail sales performance, Purchasing amount, High Amount

INTRODUCTION

Retail businesses frequently rely on price promotions, category planning, and seasonal campaigns to influence customer spending (Zhang, Meng, & Wang, 2025. In practice, however, management decisions are often made based on intuition or conventional wisdom rather than statistically tested evidence (Power, Cyphert, & Roth, 2019). The idea that motivated this study was designed around a clear managerial question: what drives store sales, and are the common assumptions about promotions and seasonality supported by data? The research focus areas supplied for the study emphasized not only explanation, but also statistical testing, model validation, and practical interpretation. In that framework, the present analysis was designed to answer eight interrelated interpretive goals: whether the mean purchase amount differs from a benchmark, whether larger discounts are associated with greater spending, whether purchase amount differs across category and season, whether non-parametric tests support the parametric conclusions, which predictors matter most in the linear model, how the logistic model explains the probability of a high-value purchase, how predictive performance differs across models, and whether the models are sufficiently strong to support managerial decision-making.

The study uses a transaction-level retail dataset with 5,000 observations and eleven original variables. These variables include customer identifiers, age, gender, product category, item purchased, amount, season, payment method, item rating, percentage discount applied, and previous purchases. The main response variable for most of the analysis is Amount, which captures the monetary value of each purchase. For classification, the response was transformed into High Amount, where purchases above the sample mean were coded as high-value transactions, and those at or below the mean were coded as low-value transactions. The study remains tightly aligned with the substantive issue: whether discounting is an effective lever for raising spending, or whether the real drivers of revenue lie elsewhere.

To answer the research questions in a way that is both statistically defensible and managerially useful, the study is organized around the same progression used in the course: data

preparation, visualization, statistical inference, ANOVA, non-parametric analysis, linear regression, logistic regression, diagnostics, and predictive performance. Every major figure generated in the analysis is interpreted in context, and each major R output is translated into substantive business meaning.

LITERATURE REVIEW

Li et al. (2020) proposed a model for grey relational analysis of seasonal time series to identify and eliminate the effects of seasonal fluctuations on retail sales of goods in China. The study employed quarterly group time series data. Results of the study revealed that data grouping effectively reflects correlations between time series across different quarters and eliminates the influence of seasonal fluctuations. However, the current study focused on empirical analysis of discounts, product categories, seasonal patterns, and predictive modeling.

Ramos (2023) forecasted sales using shrinkage dimensionality reduction. The study underscored that the inclusion of many drivers of demand may not be accurately captured by commonly used ARIMA and ETS models, which are sometimes not straightforward to apply. The study used Principal Components Analysis and Shrinkage Estimators. Results reveal that Principal Components Analysis and Shrinkage are useful and 10% more accurate than the benchmarks, while offering insights into the impact of promotions. The current study, however, focused on secondary sales data to analyze discounts, product categories, seasonal patterns, and predictive modeling.

Makkalageri *et al* (2025) studied sales discount and consumer purchase behavior. The study underscored that discounts are typically used in marketing to boost short-term sales, yet their long-term effects on brand loyalty, perceived product value, and consumer trust remain underexplored. The employed mixed approach used structured surveys. Results reveal a strong correlation ($r = 0.85$, $p = 0.05$) between seasonal influence and purchase behavior, implying that consumers, especially young consumers, salaried employees, and lower-income consumers, are likely to postpone purchases and wait for major sales events. Additionally, trust in discounted product quality and perception about the brand also increase with high seasonal influence scores. The reviewed study employed a mixed-methods approach, whereas the current study used only secondary data to analyze discounts, product categories, seasonal patterns, and to perform predictive modeling.

Kirubadevi *et al* (2020) studied short-term discounting frameworks using multiple experiments. The study underscored the changing dynamics brought about by the proliferation of online shopping sites in India, which put physical shops under pressure and caused them to lose their client base. The study revealed that discounts are a key component of promotion and

should be included by retailers. Furthermore, the study revealed that all levels and types of discounts could attract consumers to purchase more; however, the most important thing for a retailer is to understand who their consumers are, their purchase history, their purchase behavior, and their responses to previous discounts. The reviewed study focused solely on discounts; the current study, however, examined discounts, product categories, seasonal patterns, and predictive modeling.

Yuan *et al* (2021) studied the number of off-discounts and consumer responses. The study employed a meta-analysis of 19 studies and 86 samples to examine consumer responses to discount frames. Results revealed that the number of discounts and consumer responses are related through a positive change in attitude. Furthermore, the results revealed that discounts have varied effects depending on product price level, product type, and mode of price promotion. The reviewed study's approach or methodology differs from that of the current study; hence, the results may not generalize. The current study used secondary data to investigate discounts, product categories, seasonal patterns, and predictive modeling.

METHODOLOGY

Research Design

Research design is a plan for a study. It involves philosophical, assumptions, strategies of inquiry, and specific research methods (Creswell, 2009). Research design involves a decision about what, where, when, how much, by what means concerning an inquiry (Kothari, 2004). Research design makes an inquiry efficient and effective in terms of information yields and costs involved (Kothari, 2004). The current study employed a quantitative research paradigm. Specifically, the study adopted a descriptive and inferential modeling approach to explore study variables and to provide inferences for stakeholder's consumption.

Data Preparation and Descriptive Structure

The first research focus area concerned data quality. Before any inference or modeling could be trusted, the dataset had to be screened for missing values, duplicates, and variable-format issues. The R output showed that the file contained 5,000 observations and 11 variables in its raw form. Importantly, the missing-value audit returned 0 missing values across all variables, and the duplicate check returned 0 duplicate records. This is an unusually clean transactional dataset that immediately strengthens the reliability of subsequent analyses by removing the need for imputation or record deletion. The discount field was renamed from "Discount Applied (%)" to "Discount" for easier coding, and the categorical variables Gender, Category, Item Purchased, Season, and Payment Method were recoded as factors. The seasons were explicitly ordered as

Spring, Summer, Autumn, and Winter, allowing the visualizations and comparisons to follow a natural temporal sequence.

Two variables derived were central to the study design. First, a high amount was defined as an amount exceeding the overall mean purchase amount. Transactions above the mean were classified as high-value, while all others were classified as low-value. This transformation made the logistic regression possible and aligned with the question about the probability of achieving high sales. Second, Discount Group was created by comparing each discount to the median. Observations above the median were labeled High Discount, and those at or below the median were labeled Low Discount. This derived grouping enabled a direct hypothesis test of whether higher discount levels lead to greater spending.

The descriptive structure of the cleaned data also provides early insight into the retail environment represented by the dataset. The average age of customers was approximately 45.22 years, and the gender balance was nearly even, with 2,504 females and 2,496 males. Payment method was heavily concentrated on card transactions, which accounted for 4,009 purchases, while cash on delivery accounted for 991 purchases. The category distribution was more uneven: footwear and sports were strongly represented, but the most analytically important feature was the presence of a large electronics segment with very high purchase amounts. This imbalance turns out to matter enormously for both the regression and classification stages, because electronics transactions create a distinct high-value cluster that drives much of the model's behavior.

Summary of Key Statistical Results

Table 1. Consolidates the main numerical findings discussed in depth later.

Analysis component	Key statistic	Value	Interpretive summary
One-sample mean test	Mean Amount	$285.09	The average purchase amount was well above the $100 benchmark.
One-sample mean test	95% CI for mean Amount	$269.80 to $300.38	The interval indicates a stable meaning well above the benchmark.
Discount comparison	Welch t-test p-value	0.5532	No evidence that the high-discount group spent more.

Category comparison	ANOVA F-statistic	2782	Category differences were extremely large and highly significant.
Category robustness	Kruskal-Wallis chi-square	2241.5	Non-parametric test confirmed the category result.
Seasonal comparison	Season ANOVA p-value	0.974	No meaningful seasonal differences in purchase amount.
Linear regression	R-squared	0.8182	The model explained about 81.8% of the variation in the amount.
Linear regression	Electronics coefficient	+1640.76	Electronics purchases were dramatically larger than the baseline category.
Linear regression	Cash on delivery coefficient	-61.74	Cash-on-delivery purchases were lower than card purchases.
Linear prediction	MAE / RMSE / MAPE	99.9 / 235 / 53.1%	Prediction was moderately accurate but affected by large-value transactions.
Logistic classification	Test accuracy	0.996	The model classified high-value purchases extremely well.
Logistic classification	Test AUC	0.9993	Discrimination between low and high purchases was nearly perfect.

Exploratory Visualization and Distributional Evidence

The exploratory stage of the analysis served two purposes. First, it allowed the analyst to understand how the response variable behaved before selecting formal tests. Second, it helped reveal whether the substantive research questions were plausible before attempting inference and modeling (Ekbote, Dhanshetti, & Sakhrekar ,2023. The script produced several plots, each

corresponding to a different part of the research focus. Rather than treating these figures as decorative outputs, the discussion below interprets each one as evidence about how the sales process operates.

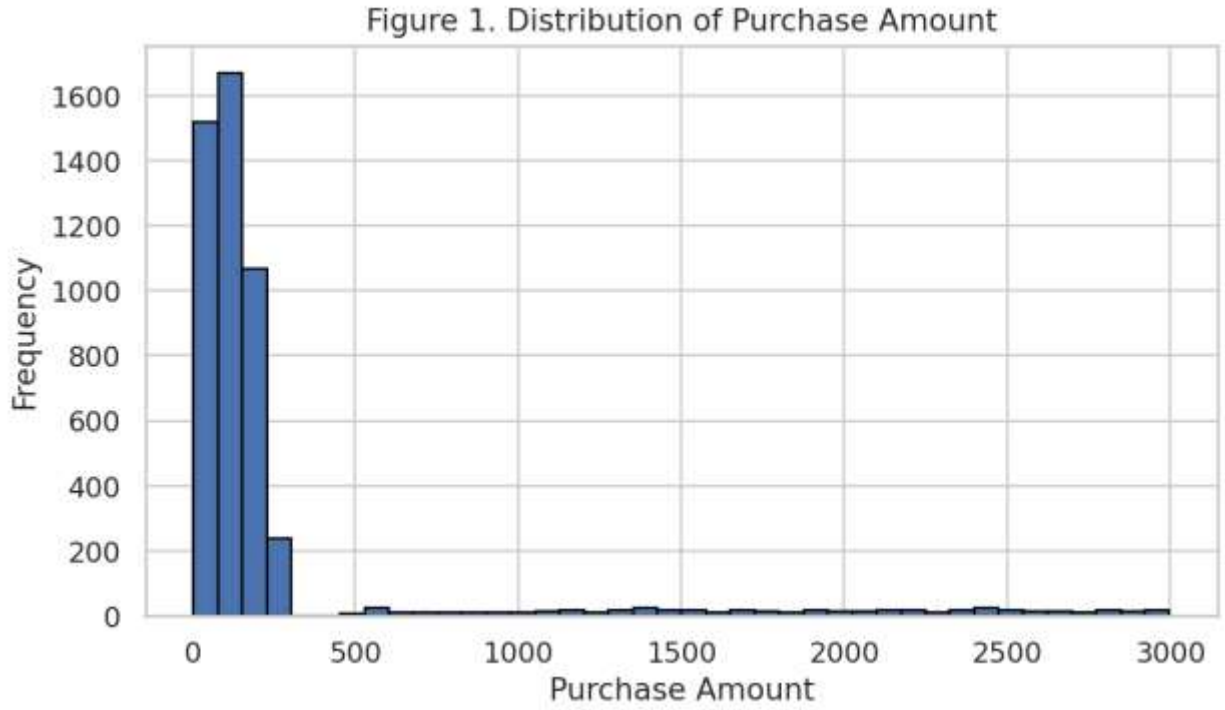

Figure 1. Distribution of purchase amount

Figure 1 shows the distribution of the amount across all 5,000 transactions. The distribution is heavily right-skewed. Most purchases are concentrated in the lower part of the monetary range, while a smaller number of transactions extend far into the upper tail. This visual pattern is fully consistent with the summary statistics: the median purchase amount (about $122.48) is far below the mean ($285.09), indicating that high-value purchases pull the mean upward. The figure, therefore, explains why normality tests later reject the assumption of normality. It also foreshadows why both robust and non-parametric checks are valuable in this research. From a business perspective, the histogram suggests that the store does not operate in a single homogeneous market segment. Instead, it combines many ordinary purchases with a much smaller set of very large transactions. Any managerial strategy designed to raise average revenue must therefore distinguish between ordinary and high-value baskets rather than assume all customers respond similarly.

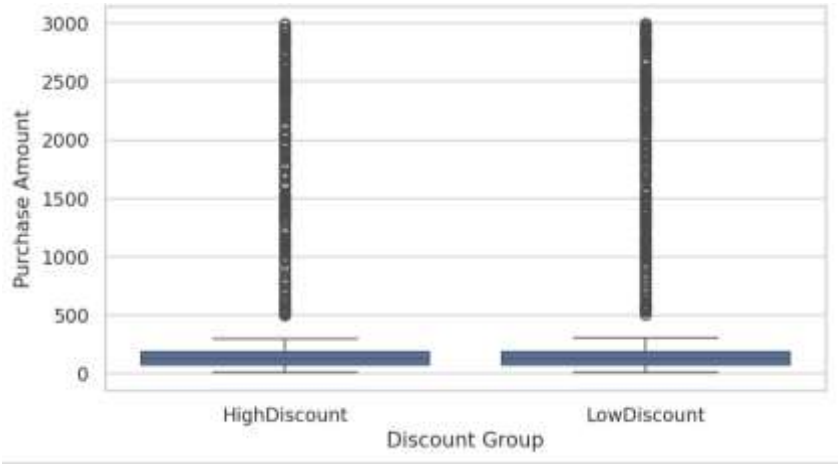

Figure 2. Purchase amount by discount group

Figure 2 compares purchase amounts for High Discount and Low Discount transactions. The two box plots overlap substantially, and neither group displays a systematically higher center than the other. The median and inter-quartile ranges are very similar, while both groups contain outliers. This visual evidence is important because it anticipates the later t-test result: if discounts were strongly increasing spending, one would expect the high-discount group to sit visibly higher. The figure instead suggests that the distribution of spending remains broadly similar regardless of whether the discount level is above or below the median. In practical terms, the plot weakens the intuitive argument that “more discount means more spending” and directs attention toward other drivers of value.

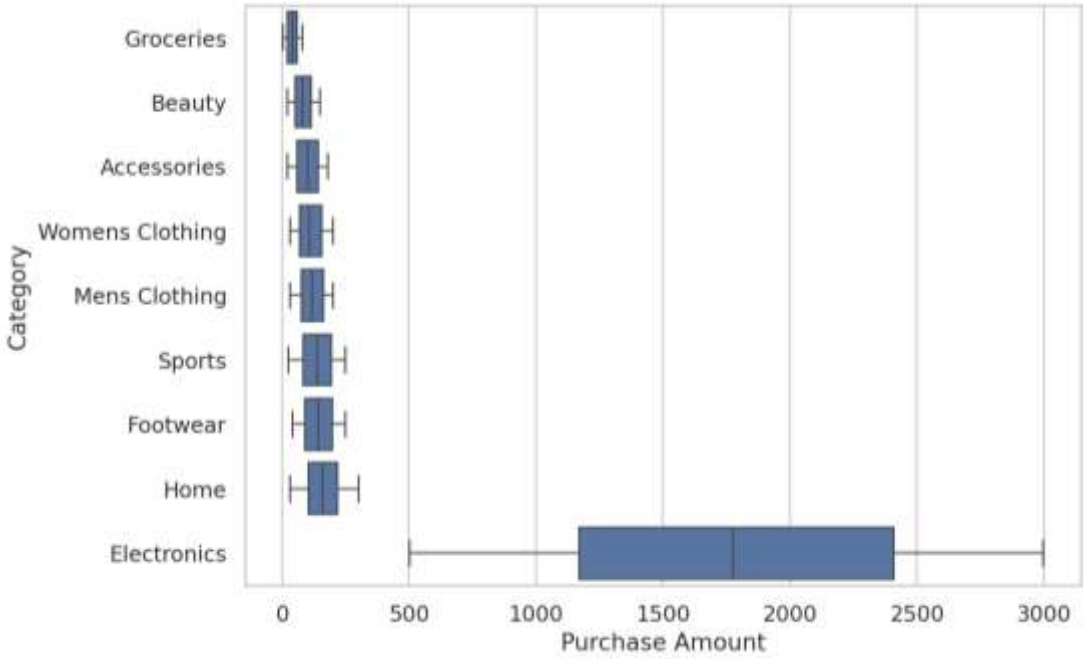

Figure 3. Purchase amount by product category

Figure 3 is one of the most informative plots in the entire. The category shows that most product groups occupy a relatively modest range of purchase values, but electronics stands apart dramatically. Its median is vastly higher than the medians of all other categories, and its spread covers a very different part of the monetary scale. This single figure explains why the category later dominates the ANOVA, the linear regression, and the logistic classification. It also reveals a practical retail truth: the store appears to sell both routine, everyday items and premium, big-ticket items, and the distinction between these markets is strongest at the category level. The category figure, therefore, directly addresses the research focus on whether store characteristics help explain sales. In this dataset, the answer is decisively yes.

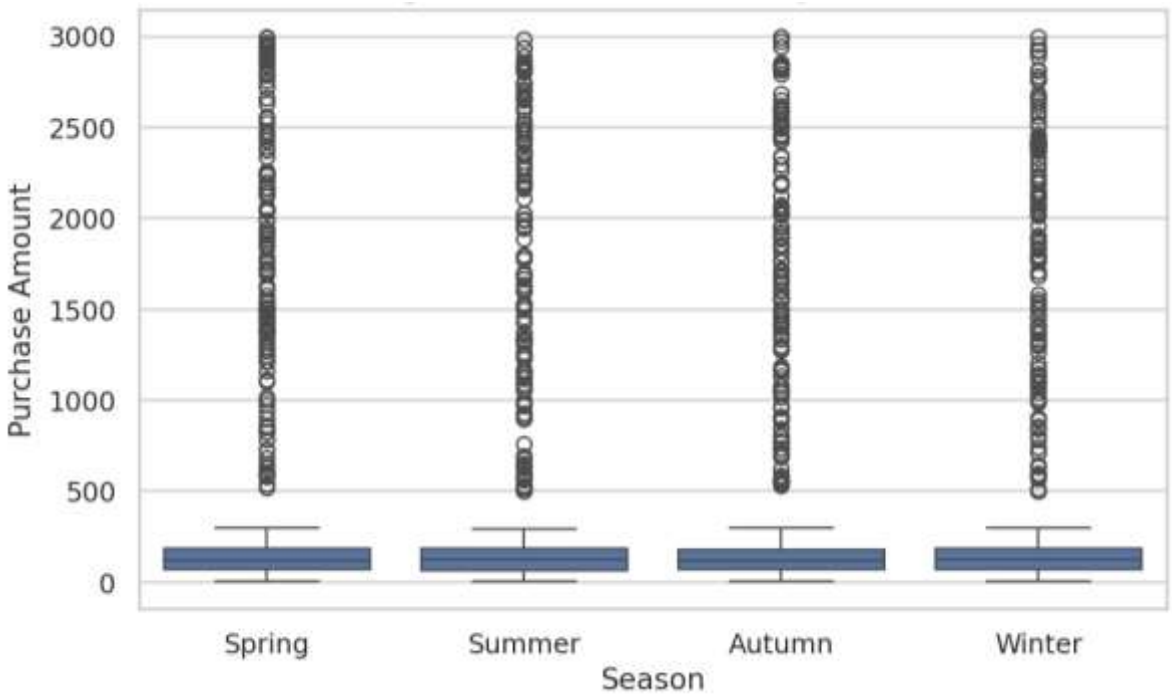

Figure 4. Purchase amount by season

Figure 4 compares the purchase amount across Spring, Summer, Autumn, and Winter. The central boxes and whiskers are highly similar, and no season visibly separates itself from the others. The mean markers introduced later in the seasonal summary also remain closely clustered. This figure is therefore useful precisely because it is unexciting: it visually suggests the absence of strong seasonal effects. That visual impression is borne out in the ANOVA results. Substantively, the season plot indicates that transaction values are more stable across the year than might have been expected. If managers were expecting a large seasonal swing in basket size, the figure offers little support for that assumption.

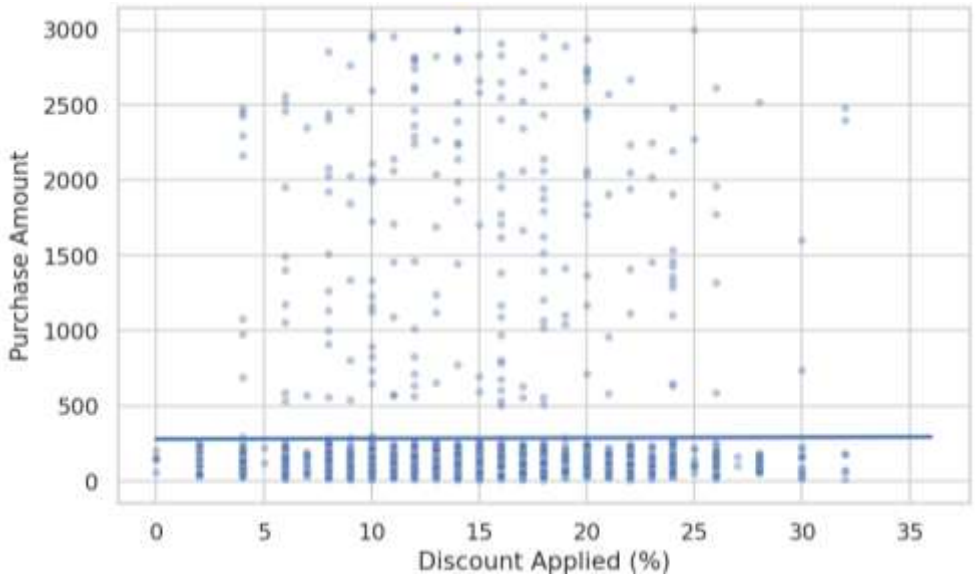

Figure 5. Purchase amount versus discount

Figure 5 plots purchase amount against discount percentage and includes a fitted trend line. The scatter cloud is diffuse, and the fitted line is shallow. There is no strong upward relationship between discount and spending. Large purchases occur at several discount levels, and low purchases also occur throughout the discount range. This figure is crucial for interpreting the later non-significant discount coefficient in the linear regression. The relationship between discount and revenue is not merely weak after controlling for other variables; it is already weak in the raw data. The plot therefore reinforces the conclusion that discounting is not the main mechanism determining purchase amount in this dataset.

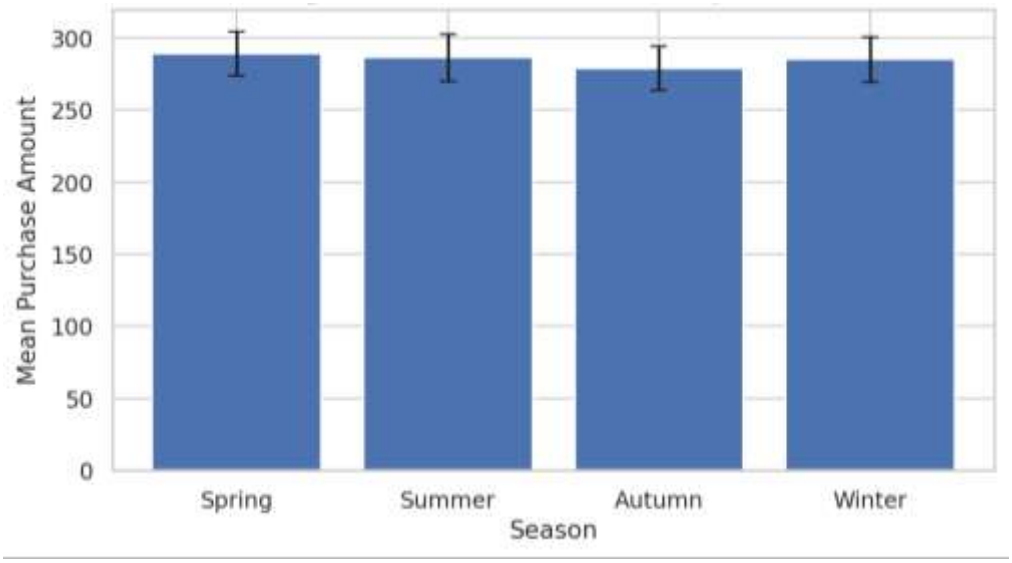

Figure 6. Mean purchase amount by season

Figure 6 summarizes seasonal meanings with standard error bars. The seasonal means are approximately 289 for Spring, 286 for Summer, 279 for Autumn, and 285 for Winter, and the error bars overlap almost completely. The practical meaning is that even if small numerical differences exist, they are tiny relative to the overall variability in purchase amount. The chart helps bridge the descriptive and inferential parts of the analysis by showing why the seasonal ANOVA is not significant: the observed seasonal means are simply too close together to support a substantive seasonal interpretation.

RESULTS

Confidence Interval and Benchmark Test for Mean Purchase Amount

The first formal inferential result examined whether the average purchase amount differed significantly from the benchmark of $100. The one-sample t-test returned a sample mean of 285.0905, a t-statistic of 23.733 with 4,999 degrees of freedom, and a p-value smaller than 2.2 × 10^-16. The 95% confidence interval for the true mean purchase amount ranged from 269.8016 to 300.3795. Statistically, this is an unambiguous rejection of the null hypothesis that the mean purchase amount equals $100. Substantively, the result indicates that the typical transaction value in this dataset is well above a low benchmark and that the estimated average is not only large but also highly precise due to the sample size.

The QQ plot for Amount, however, also showed a clear departure from normality, and the Shapiro-Wilk test strongly rejected the null hypothesis of normality. This does not invalidate the mean comparison because the sample size is large enough for the T-Test to remain informative, but it does matter for interpretation. The combination of a strong mean result and a non-normal raw distribution indicates that the average is influenced by a subset of very large purchases (Pandis & Tsagris, 2021). In an academic study, this distinction is important: the mean is a statistically reliable estimate, but the distributional shape reminds readers that “average spending” should not be confused with “typical spending” in the everyday sense. For managerial purposes, the result implies that the store generates substantial overall value, but much of that value is concentrated in the upper tail of the transaction distribution.

Discount Effect: Do Higher Discounts Increase Spending?

The next research focus area examined whether higher discounts are associated with higher spending. The data were divided into High Discount and Low Discount groups using the sample median as the cutoff. Before comparing means, the normality of the amount within each group was checked. Both groups strongly violated the assumption of normality (Pandis & Tsagris, 2021), with Shapiro-Wilk p-values well below conventional significance levels. Levene’s test,

however, returned a p-value of 0.9545, indicating no evidence of unequal variances between the two discount groups. On that basis, they used a Welch two-sample t-test with the directional alternative that the High Discount group should spend more than the Low Discount group.

The resulting statistics do not support that expectation. The estimated mean purchase amount in the High Discount group was 283.9570, and in the Low Discount group, 286.0491. The estimated difference was therefore negative, approximately -2.09. The test statistic was -0.13373 with a p-value of 0.5532, and the lower bound of the one-sided confidence interval was -27.82858. In plain language, the evidence is inconsistent with the idea that higher discounts increase purchase value. Not only is the difference statistically non-significant, but the direction of the observed difference is opposite to the managerial assumption embedded in the alternative hypothesis.

This result is highly important because it directly addresses the core promotional question in the research focus. A retailer often assumes that offering larger discounts will induce customers to spend more per transaction. In this dataset, there is no evidence of such an effect. On the contrary, Liu et al. (2021) asserted that permanent discounts negatively affect spending when set below 19%. The boxplot in Figure 2 and the scatterplot in Figure 5 already suggested that the spending distributions were similar across discount levels; the formal test confirms that impression. The managerial interpretation is not that discounts are useless in every sense, but that larger discounts are not a reliable lever to increase basket value here. That conclusion should prompt caution: the firm may be sacrificing margin without a compensating gain in purchase amount.

Category and Season Comparisons through ANOVA

The next question asked whether the purchase amount differs between the relevant groups. The two groupings of greatest interest were product category and season. The category is substantively tied to the kind of item being sold, while the season serves as the available temporal structure in the dataset. Empirical studies reviewed revealed that seasonal retail sales differ from quarter to quarter; the main factors determining seasonal fluctuations are the money supply, taxes, and residents' per capita disposable income (Li et al., 2020).

For the product category, the analysis began with a boxplot and a faceted QQ-plot of Amount across categories. The QQ-plot showed that normality was not a reasonable assumption across many categories, and Levene’s test provided very strong evidence of unequal variances across categories, with an F statistic of 1374.7 and a p-value below 2.2 × 10^-16. Despite that heterogeneity, the one-way ANOVA produced an F statistic of 2782 and a p-value below 2.2 × 10^-16, indicating overwhelming evidence that not all category means are equal. The magnitude

of this result is extraordinary, and Figure 3 makes the reason visually obvious: electronics transactions are far larger than transactions in the other categories.

The Tukey post-hoc comparisons clarify the structure of those differences. Electronics exceed every other category by a very large margin, with pairwise differences that are statistically significant across the board. Several smaller contrasts also appear, such as home exceeding Groceries and Footwear exceeding Accessories. However, the substantive story is not merely that “some categories differ.” It is that one category, Electronics, that defines a distinct upper tier of spending. This matters because it shifts the discussion away from generalized store performance and toward category strategy. The business does not appear to face a universal sales problem; rather, it operates with a bifurcated product mix in which some categories inherently generate much larger transaction values.

The season analysis produced a very different pattern. Mean purchase amounts were 289 in Spring, 286 in Summer, 279 in Autumn, and 285 in Winter. The season ANOVA returned an F statistic of only 0.074 and a p-value of 0.974. The Tukey comparisons similarly showed no meaningful seasonal separation, with all adjusted p-values effectively non-significant. Figure 4 and Figure 6 both make this easy to see: the seasonal boxplots and means are tightly clustered. In terms of the research focus, this means that the available time-structured variable, Season, does not explain purchase amount. The implication is that sales value in this dataset is not driven by a seasonal cycle; rather, it is driven by customer purchases.

Non-Parametric Validation

Because the category distributions were visibly non-normal and variances were unequal, they appropriately added a non-parametric robustness check (Orcan, 2020). The Kruskal-Wallis test for Amount by Category produced a chi-squared statistic of 2241.5 with 8 degrees of freedom and a p-value below 2.2 × 10^-16. This confirms the ANOVA conclusion from a distribution-free perspective: category differences in purchase amounts are not artifacts of parametric assumptions.

The pairwise Wilcoxon comparisons extend that conclusion further. Many pairwise contrasts remained significant after Bonferroni adjustment, especially those involving electronics relative to all other categories. The fact that both the parametric and non-parametric approaches point in the same direction is important academically because it strengthens the internal validity of the inference (Aleem & Salman, 2024). It would be easy to dismiss a large ANOVA result as an artifact of outliers or variance heterogeneity. Kruskal-Walli’s result blocks that criticism by showing that the conclusion survives even when the analysis relies on ranked rather than raw values. Therefore, the answer to the fourth interpretation prompt is clear: yes, the non-parametric

test supports the ANOVA conclusion. The category effect is not only statistically significant but also robust to violations of assumptions.

Linear Regression: Which Predictors Drive Purchase Amount?

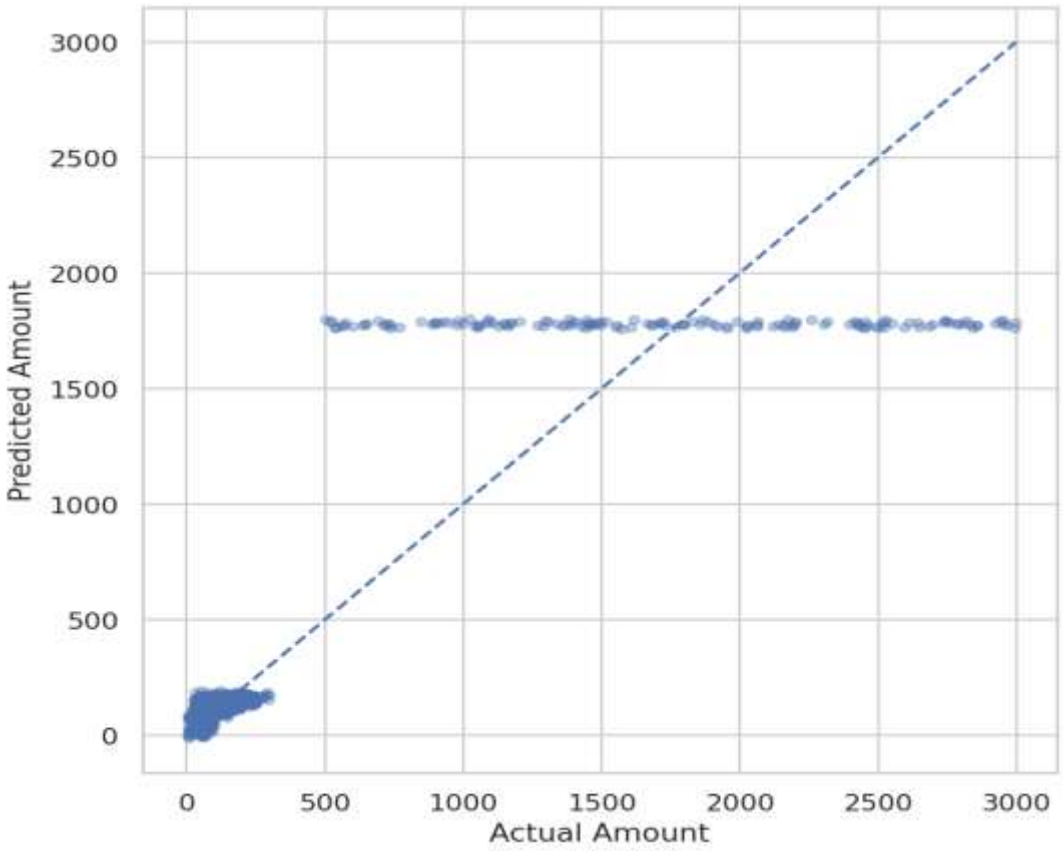

Figure 7. Predicted versus actual purchase amount on the test set

Figure 7 displays the relationship between predicted and actual purchase amounts in the out-of-sample data. The plot shows that the model can distinguish between low-to-moderate purchases and very high purchases, but the points are not uniformly distributed along the 45-degree line. This visual pattern is consistent with the ed MAE, RMSE, and MAPE values: the model is informative and useful, yet prediction error remains non-trivial because the outcome spans a very wide range and includes a distinct cluster of very expensive electronics purchases.

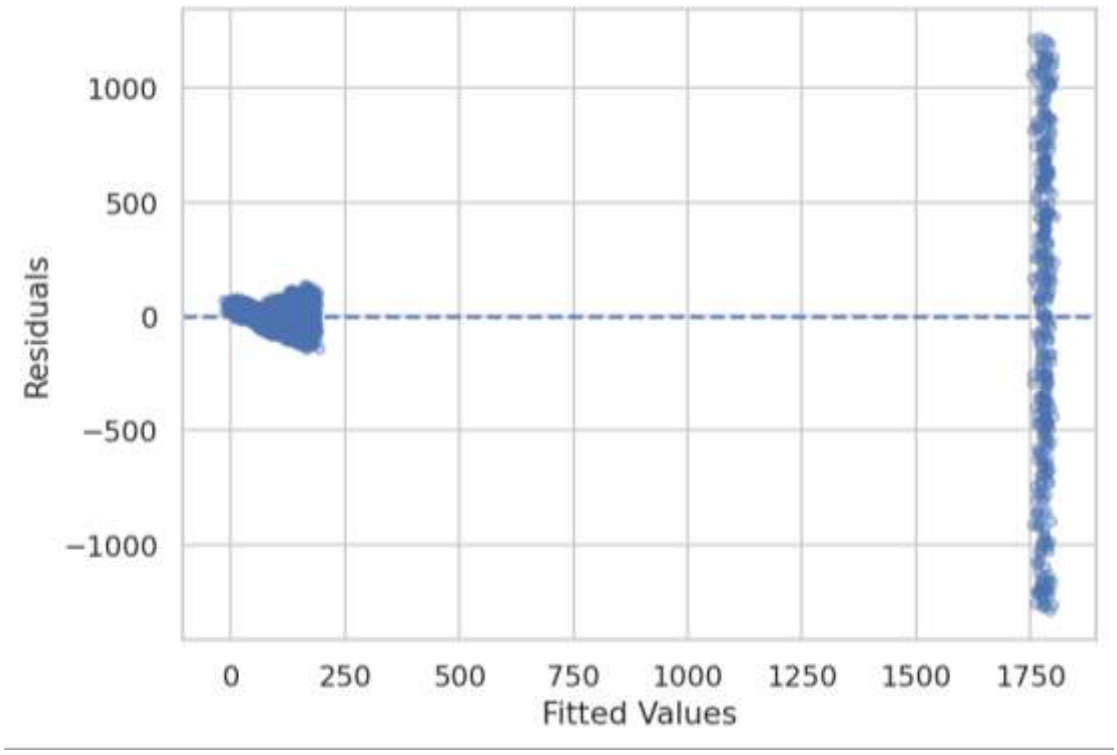

Figure 8. Residuals versus fitted values for the linear model

Figure 8 shows residuals plotted against fitted values. The residual meaning is essentially zero, which is desirable, but the spread is not perfectly constant across the fitted range. The model, therefore, captures the main structural drivers of Amount without fully eliminating the heterogeneity in the extreme upper tail. This is a common feature of retail data and should be interpreted as evidence of outcome complexity rather than model failure.

The multiple linear regression model was fitted to Amount using Age, Gender, Category, Season, Payment Method, Item Rating, Discount, and Previous Purchases on the training set. The full model produced an R-squared of 0.8182 and an adjusted R-squared of 0.8173, with an overall F statistic of 922 and a p-value below 2.2 × 10^-16. These are strong fit statistics. They indicate that roughly 81.8% of the variation in purchase amount is explained by the included predictors. In substantive terms, the model does not merely detect a few significant coefficients; it captures most of the systematic structure in transaction value.

The coefficient table reveals which variables are responsible for that explanatory power. The most important by far is Category Electronics, with an estimated coefficient of 1640.7598 and an extremely large t value. Holding all other predictors constant, an electronics transaction is predicted to be about $1,640 higher than the baseline category. No other coefficient comes close in magnitude. This is the clearest statistical answer to the question of what drives revenue in this dataset: product category, particularly electronics, overwhelms the

contribution of the other variables. Smaller but statistically meaningful category effects also appear for Home (positive), Groceries (negative), and Footwear (marginally positive). Payment Method Cash on Delivery has a coefficient of -61.7425 and is highly significant, indicating that cash-on-delivery transactions tend to be about $62 lower than card transactions, all else equal.

By contrast, the variables that many managers might expect to matter did not prove influential. Age, gender, item rating, discount, and previous purchases were all statistically non-significant in the full model. This pattern is crucial for interpretation. It suggests that basket value is not determined by demographic composition, customer history, or the level of discount once the purchase category is taken into account. The discount coefficient, at 0.6365 with a p-value of 0.5155, is especially noteworthy because it independently confirms the earlier t-test finding: even when analyzed as a continuous predictor inside a multivariable framework, discount still fails to emerge as a meaningful driver of purchase amount.

The study also examined the model's confidence intervals and a backward stepwise simplification. The reduced stepwise model retained only Category and Payment Method while maintaining essentially the same explanatory power, with an R-squared of 0.8179. That is a very revealing result. It means that once category and payment method are known, the other predictors add comparatively little explanatory value. Academically, this is a strong parsimony argument. Managerially, it states that the store's transaction value can be understood primarily by the type of product purchased and whether the transaction is paid for by card or in cash on delivery. The other variables may still matter in operational or customer-experience terms, but they are not the central determinants of monetary value per transaction in this dataset.

Linear Model Diagnostics and Outlier Structure

Model adequacy was assessed using variance inflation factors, influence diagnostics, QQ analysis of residuals, and the residuals-versus-fitted plot. The VIF values were all low. The generalized VIF values, converted to comparable scales, were close to 1 across all predictors, indicating that multicollinearity is not a serious problem. This is an important diagnostic result because it means the strong effect of the category is not an artifact of redundancy with the other predictors.

The influence analysis identified 204 observations with Cook's distance greater than 4/n. Given the sample size and the shape of the outcome distribution, this is not surprising. The dataset contains a mixture of ordinary purchases and much larger electronic transactions, which naturally creates influential cases. However, the presence of influential observations

does not by itself imply that they are erroneous or should be removed. In this context, those observations are likely to be genuine high-value retail transactions rather than data-entry mistakes. Removing them would discard exactly the portion of the market that makes the dataset interesting. The academically appropriate conclusion is therefore that influential points exist because the business genuinely contains a large-value segment, not because the model is malformed.

Residual normality was not supported by the Shapiro-Wilk test, which returned a p-value below 2.2 × 10^-16. Yet the mean of the residuals was essentially zero, satisfying the central-location assumption of OLS. The residual plot also indicates that the model captures the main structure but not every fine-grained aspect of variance across the fitted range. In a perfect Gaussian setting, this would be concerning, but in a retail transaction setting, it is expected. The diagnostics, therefore, support a nuanced conclusion: the linear model is strong and informative, but it should be interpreted as an effective explanatory and predictive summary rather than as a perfectly assumption-conforming Gaussian model.

Predictive Accuracy of the Linear Model

Out-of-sample performance for the linear model was evaluated on the test set using MAE, RMSE, and MAPE. The results were MAE = 99.9, RMSE = 235, and MAPE = 53.1%. MAE indicates that the average absolute prediction error is about $100. RMSE, which penalizes larger errors more heavily, rises to about $235, indicating that some predictions are substantially off. MAPE indicates an average absolute percentage error of approximately 53.1%, which is fairly large.

These metrics should not be interpreted superficially. At first glance, an MAPE above 50% may appear to indicate a weak model. However, percentage error is especially sensitive when the outcome variable includes many small purchases and a few very large purchases. In this dataset, the outcome spans from approximately $5 to nearly $3,000, so a moderate absolute error on a small transaction can produce a large percentage error. The prediction graph in Figure 7, therefore, provides essential context: the model captures the broad transaction structure, but the outcome is so heterogeneous that exact point prediction is inherently difficult. The correct managerial reading is that the linear model is good for identifying the main drivers of value and for making directional predictions, but it is not a precision forecasting engine for every individual transaction.

Logistic Regression: Probability of High-Value Purchases

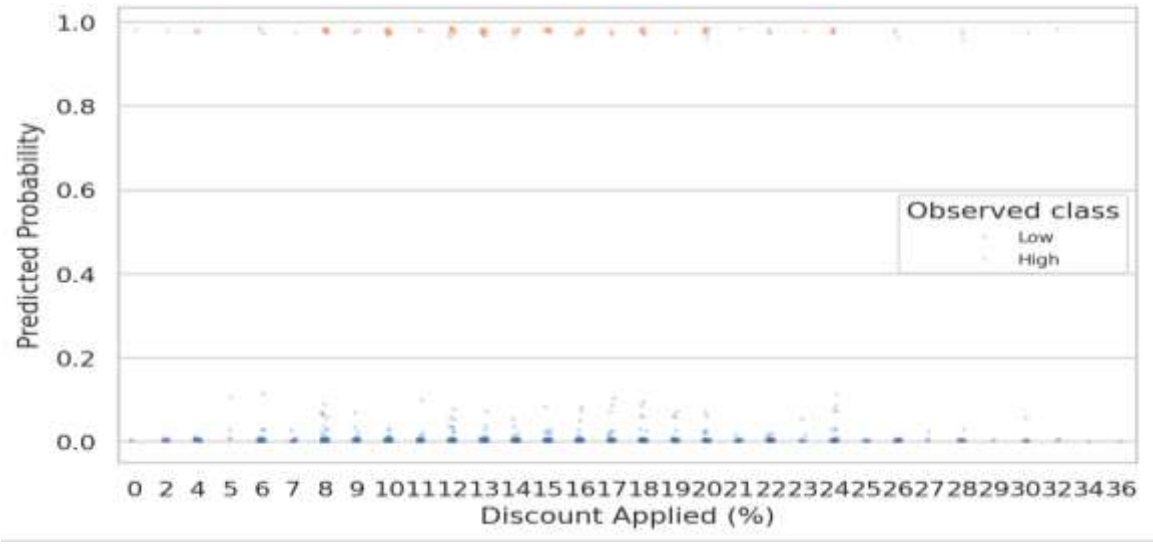

Figure 9. Predicted probability of a high-value purchase across discount levels

Figure 9 shows the fitted probability of a high-value transaction as a function of discount. The cloud of points reveals that discount alone does not cleanly and monotonically structure the probability. High predicted probabilities tend to be concentrated in transactions rather than being distributed smoothly across larger discounts. This visual result is consistent with the regression output, in which discount is not a significant driver of the high-value classification once the other variables are included.

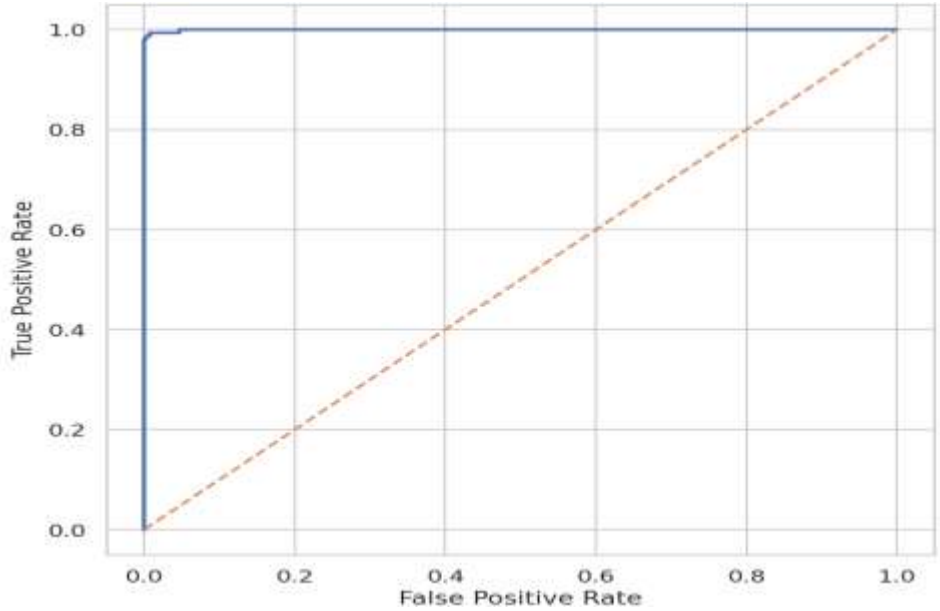

Figure 10. ROC curve for the logistic model

Figure 10 displays the ROC curve for the logistic model. The curve lies very close to the upper-left corner, indicating extremely strong discrimination between high-value and low-value

purchases. This visual impression aligns with the ed AUC values above 0.999 and confirms that the logistic model is exceptionally effective at ranking transactions by high-value risk, even though some coefficients are unstable due to quasi-separation.

The logistic regression addressed a different but related question: what factors affect the probability that a transaction will be a high-value purchase? The dependent variable, High Amount, classified observations above the overall mean purchase amount as “High” and all others as “Low.” The training-set logistic model included the same predictors as the linear model. A warning was issued in R stating that “fitted probabilities numerically 0 or 1 occurred.” This warning is analytically important because it indicates quasi-separation: some combinations of predictors, especially categorical ones, almost perfectly determine whether a transaction is high- or low-value.

The coefficient table should therefore be interpreted carefully. Several coefficients have extremely large standard errors and p-values near 1, not necessarily because the variables are unimportant in substantive terms, but because the model is operating in a near-separation environment. The most obvious example is Category Electronics, which has a very large positive estimate, implying massively increased odds of a high-value purchase. Exponentiating that coefficient produces an enormous odds ratio, which is mathematically correct but substantively not very useful as a literal managerial multiplier. The more honest interpretation is qualitative: electronics transactions are overwhelmingly more likely to be high-value purchases than the baseline category. Item Rating shows a positive but non-significant association, while age, discount, and previous purchases remain negligible. Gender (Male) has an odds ratio above 1, but it is not statistically significant. In short, the logistic model tells the same story as the linear model: transaction characteristics, especially category, matter far more than discount or customer demographics.

The omnibus fit of the logistic model is excellent. The chi-squared improvement from null deviance to residual deviance is effectively significant at $p \approx 0$, indicating that the full model is far better than an intercept-only specification. Yet the real strength of logistic regression lies in classification performance rather than coefficient-level inference. On the training set, the model achieved an accuracy of 0.997, a sensitivity of 0.975, a specificity of 1.000, a precision of 1.000, and an AUC of 0.9997. Using the selected cutoff of 0.25, the test-set performance remained almost identical: accuracy of 0.996, sensitivity of 0.968, specificity of 0.999, precision of 0.993, and an AUC of 0.9993. These are remarkably high values.

Academically, such near-perfect performance should be interpreted with caution. It does not necessarily mean the underlying behavioral process is simple. Instead, it suggests that high-value purchases are strongly separable from ordinary purchases due to one or more dominant

predictors. In this dataset, the dominant predictor is clearly category, especially electronics. Therefore, the logistic model is extremely useful for classification, but it should not be overinterpreted as evidence that every estimated odds ratio is stable or transportable to a different retail environment. The model is best read as a powerful discriminator shaped by a very strong structural distinction in the data.

Comparison of Linear and Logistic Performance

The study explicitly asked for a comparison between MAE/RMSE/MAPE for the linear model and AUC/accuracy for the logistic model. These metrics answer different questions. The linear model predicts a continuous amount, so its performance is judged by how close the predicted value is to the actual amount. The logistic model predicts a class, so its performance is judged by how well it separates high-value from low-value transactions.

The linear model is strong in explanatory terms and respectable in predictive terms, but it is not perfect. Its R-squared above 0.81 indicates that it captures most of the structure in Amount, yet the MAPE reminds us that exact transaction-level forecasts remain challenging. The logistic model, by contrast, performs almost perfectly in terms of classification. This does not mean that the logistic model is “better” in every sense; rather, it means that classifying transactions into high versus low value is much easier than forecasting their exact monetary amount. That is a common outcome in applied analytics. A business can often identify which purchases are likely to be large more accurately than it can predict the exact value of every purchase down to the dollar.

The linear model is better suited for explaining the drivers of purchase amount and quantifying the expected change associated with product categories and payment methods. The logistic model is better suited for operational screening and targeting, such as identifying which transactions or customers are likely to belong to a high-value segment. Together, they provide a fuller decision-support toolkit than either model alone.

DISCUSSIONS

Managerial Interpretation Relative to the Research Focus Areas

The research focus areas emphasized managerial interpretation rather than statistical output alone, and the results point to several clear conclusions. First, the mean purchase amount is not merely above the benchmark; it is far above it. This indicates that the store has meaningful revenue-generating capacity. Second, higher discounts are not associated with higher spending (Yuan *et at*., 2021). However, the observation contradicts Kirubadevi et al. (2020) and Yuan et al. (2021), who asserted that off-discounts influence consumer attitudes. Furthermore, Kirubadevi *et al. (2020) observed that, in addition to discounts, a retailer should*

understand consumers' purchase history, purchase behavior, and level of understanding of the purchase discount being offered. This is one of the most consequential findings in the study because it calls into question the efficiency of a common retail tactic. If larger discounts do not increase transaction value, their effect on profitability may be negative once margin erosion is accounted for. Third, the purchase amount varies dramatically by category, especially because electronics constitute a high-value segment that is structurally distinct from the rest of the product mix; this result is consistent with Steinhauser *et al* (2019), whose research confirms that product category influences purchase behavior. Fourth, season does not materially influence spending, so seasonal narrative explanations are weak in this dataset. This finding aligns with a study by Li *et al. (2020), which asserted that when data is grouped quarterly, the seasonal effect is eliminated, and contradicts a study by Makkalageri et al.* (2025), whose results revealed that seasonal influence affects clients' consumption behavior. Fifth, the non-parametric test confirms the category result, which means the conclusion does not depend on fragile assumptions. Sixth, in the linear regression, the strongest significant predictors are category, especially electronics, and payment methods. Seventh, in the logistic model, the main substantive odds interpretation is that category membership, again especially electronics, is overwhelmingly associated with the probability of a high-value purchase. Finally, the models are strong enough to support managerial decision-making, but only if they are used for the right purpose: explanation and segmentation rather than exact transaction-level forecasting in all cases.

These conclusions matter because they redirect managerial attention. A retailer looking only at discount policy might mistakenly assume that more aggressive promotions are needed. The evidence does not support that. Instead, the revenue structure appears to be category-driven. This suggests that managerial leverage lies in assortment, merchandising, premium-product strategy, and payment-channel optimization rather than in stronger discounting. The payment-method effect also deserves practical attention. Cash on delivery is associated with lower transaction values, which may reflect customer type, order risk, or purchasing caution. Encouraging card transactions could therefore indirectly improve basket size, or at least better align the business with higher-value purchases.

LIMITATIONS AND ACADEMIC CAUTIONS

An academic study should also state its limitations. First, the dataset includes Season but not a continuous transaction date, so the temporal analysis is necessarily seasonal rather than a full time-series model. Claims about trend or cyclical forecasting beyond the four seasons would therefore be inappropriate. Second, the unusually clean nature of the dataset and the strong

structural role of electronics suggest that the data may be synthetic or at least highly regularized. That does not invalidate the analysis, but it does mean that the near-perfect logistic performance should not be generalized uncritically to a more chaotic real-world retail environment. Third, the logistic regression coefficients are affected by quasi-separation, so coefficient-level interpretation should be more restrained than classification-level interpretation. Fourth, the transaction-level approach captures what happened in each purchase, but not necessarily why. For example, the non-significance of discounts does not prove that discounts never influence behavior; it only shows that within this dataset, discount level is not a reliable determinant of purchase amount once the existing transaction structure is considered.

Even with these limitations, the analysis remains academically defensible because each major conclusion is supported by multiple forms of evidence: descriptive plots, parametric tests, non-parametric validation, regression coefficients, diagnostics, and out-of-sample performance. In that sense, the study does more than merely fit models; it triangulates on a coherent substantive story.

CONCLUSION

The aim is to evaluate the drivers of retail purchases using the store sales dataset and the study's research focus areas. The answer that emerges from the evidence is clear. The purchase amount in this dataset is not primarily driven by discounts and does not vary meaningfully by season. Instead, it is driven most strongly by product category and, to a lesser but still important extent, payment method. Electronic transactions constitute a distinct high-value segment that shapes the overall data structure and accounts for much of the strong performance of both the linear and logistic models. The linear regression explains roughly 81.8% of the variance in purchase amount, while the logistic model almost perfectly distinguishes high-value from low-value transactions, though quasi-separation can affect coefficient stability.

From a managerial standpoint, the most important implication is strategic reallocation of attention. If the organization wants to raise revenue per transaction, the evidence points more strongly toward category and channel strategy than toward deeper discounts. An emphasis on premium products, especially in high-value categories, appears more promising than relying on discount escalation. Likewise, payment-channel effects should not be ignored, because lower-value transactions are associated with cash on delivery. From an academic standpoint, this demonstrates the value of combining visualization, inference, ANOVA, non-parametric tests, regression, diagnostics, and predictive validation in a single coherent analysis. Each method contributes a different piece of evidence, but all point to the same substantive conclusion: in this retail setting, what the customer buys matters far more than how much of a discount they receive.

REFERENCES

Aleem, D., & Salman, H. M. (2024). An easy way to understand the Research Tools, Scales and Parametric & Non-Parametric Tests. *Available at SSRN, http://dx.doi.org/10.2139/ssrn.4986837*

Creswell, J. (2009). *Research Design: Quantitative, qualitative and mixed methods approaches.* California: Sage Publications, Inc.

Ekbote, N., Dhanshetti, P., & Sakhrekar , S. (2023). Techniques of Exploratory Data Analysis. *A biannually Journal of M. P. Institute of Social Science Research, Ujjain 28(2)*, 10-14.

Ganesha, H. R., Aithal, P. S., & Kirubadevi, P. (2020). Short-Term Discounting Frameworks: Insights from Multiple Experiments. *International Journal of Case Studies in Business, IT, and Education 4(1) http://doi.org/10.5281/zenodo.3762872.*, 8-22.

Kothari, C. (2004). *Research methodology methods and techniques.* New Delhi: New Age international (p) Limited publishers.

Li, Z., Zhou, C., Wu, J., & Wang, Z. (2020). Identifying the factors of China's seasonal retail sales of consumer goods using a data grouping approach–based GRA method. *Grey Systems: Emirald publishers, 10(2) https://doi.org/10.1108/GS-11-2019-0055*, 125-143.

Liu, H., Labschat, L., Verhoef, P., & Zhao, H. (2021). The effect of permanent product discounts and order coupons on purchase incidence, purchase quantity, and spending. *Journal of Retailing 97(3) https://doi.org/10.1016/j.jretai.2020.11.007*, 377-393.

Makkalageri, S., Shilpa, G., Gokula, K. S., Kaviyarasan, K., & Vidyashree, K. (2025). Examining the Influence of Mega Sales Discounts on Consumer Purchase Behavior in the Consumer Durables Market. *International Scientific Journal of Engineering and Management (ISJEM), 4(5) DOI: 10.55041/ISJEM03714*, 1-12.

Orcan, F. (2020). Parametric or Non-parametric: Skewness to Test Normality for Mean. *International Journal of Assessment Tools in Education International 7(2) https://doi.org/10.21449/ijate.656077*, 255-265.

Pandis, N., & Tsagris, M. (2021). Normality test: Is it really necessary? *American Journal of Orthodontics and Dentofacial Orthopedics 159(4)*, 548-549.

Power, D., Cyphert, D., & Roth, R. (2019). Analytics, bias, and evidence: the quest for rational decision making. *Journal of Decision Systems, 28 (2) https://doi.org/10.1080/12460125.2019.1646509*, 63-66.

Ramos P, Oliveira, J. M., Kourentzes, N., & Fildes, R. (2022). Forecasting Seasonal Sales with Many Drivers: Shrinkage or Dimensionality Reduction? *Applied System Innovation, 6(1), https://doi.org/10.3390/asi6010003*. *https://doi.org/10.3390/asi6010003*.

Steinhauser, J., Janssen , M., & Hamm, U. (2019). with nutrition and health claims: What role do product category and gaze duration on claims play? *ScienceDirect 141 https://doi.org/10.1016/j.appet.2019.104337*.

Yuan, Q., Li, J., Jiang, Y., & Liu, C. (2021). When do amount-off discounts result in more positive consumer responses? Meta-analytic evidence. *Psychology and Marketing, https://doi.org/10.1002/mar.21572*.

Zhang, D., Meng, S., & Wang, Y. (2025). Impact Analysis of Price Promotion Strategies on Consumer Purchase Patterns in Fast-Moving Consumer Goods Retail. *Academia Nexus Journal 4(1) https://academianexusjournal.com/index.php/anj/article/view/36/37*, 1-34.