Business Analysis using SAS Studio

Clothing Retail & Distribution Company

Diane Haiden

February 2025

Table of Contents

1.0 Summary

In the dynamic world of clothing retail and distribution, data analysis plays a crucial role in driving business decisions and maintaining a competitive edge. This study examines a clothing retail and distribution company's dataset to improve marketing promotions and customer targeting strategies. By leveraging descriptive and predictive analytics, this study uncovers key trends in store sales and orders to support data-driven decision-making.

2.0 Business Problem

The company aims to identify customers who respond well to its marketing efforts and predict future business growth. The objective is to optimize marketing strategies, improve customer targeting, and maximize return on investment (ROI).

<u>Strategic Goals</u>

- Maximize the return on investment (ROI) from marketing efforts.
- Boost overall sales by increasing customer visit frequency.
- Enhance pricing strategies and inventory management for top-selling products.
- Optimize discount strategies to increase profitability.

By understanding these relationships, the company can make better decisions to improve its marketing, sales, pricing, and discount strategies while maximizing sales and profits.

3.0 Business Questions & Hypotheses

1. What is the relationship between marketing promotions (PROMOS) and average amount spent per visit (AVRG) for credit card users?
    - *H_0:* No significant relationship exists between PROMOS and AVRG.
    - *H_1:* A significant relationship exists between PROMOS and AVRG.

2. How does the frequency of purchase visits (FRE) affect total net sales (MON)?
    - *H_0:* No significant relationship exists between FRE and MON.
    - *H_1:* A significant relationship exists between FRE and MON.

3. What are the average unit prices and quantities for top-selling products?
    - *H_0:* No significant differences exist between top-selling and other products.
    - *H_1:* Significant differences exist between top-selling and other products.

4. How do discount rates impact gross and net sales?
    - *H_0:* Discount rates do not significantly impact gross and net sales.
    - *H_1:* Discount rates significantly impact gross and net sales.

4.0 Descriptive Statistics

Central tendency and dispersion measures were calculated for key variables, including PROMOS, AVRG, CC_CARD, FRE, and MON. Key findings:

• Average purchase frequency correlates with increased total net sales.
• Variance in spending behavior was observed among credit card users based on promotional activities.

Central Tendency – Sales

Variable	Mean	Median	N	Mode
PROMOS	11.5391159	12.0000000	28799	4.0000000
AVRG	113.5883176	92.0000000	28799	98.0000000
CC_CARD	0.3830341	0	28799	0
FRE	5.0390291	3.0000000	28799	1.0000000
MON	473.2124633	261.0000000	28799	98.0000000

Measures of spread or dispersion – Sales

Variable	Std Dev	Minimum	Maximum	Variance	Range	Lower Quartile	Upper Quartile
PROMOS	7.1393560	0	38.0000000	50.9704037	38.0000000	5.0000000	17.0000000
AVRG	86.9808026	0.4900000	1919.88	7565.66	1919.39	60.9800000	139.5000000
CC_CARD	0.4861350	0	1.0000000	0.2363272	1.0000000	0	1.0000000
FRE	6.3491216	1.0000000	115.0000000	40.3113456	114.0000000	1.0000000	6.0000000
MON	659.3274137	0.9900000	24140.33	434712.64	24139.34	135.0600000	567.5800000

Central Tendency - Orders

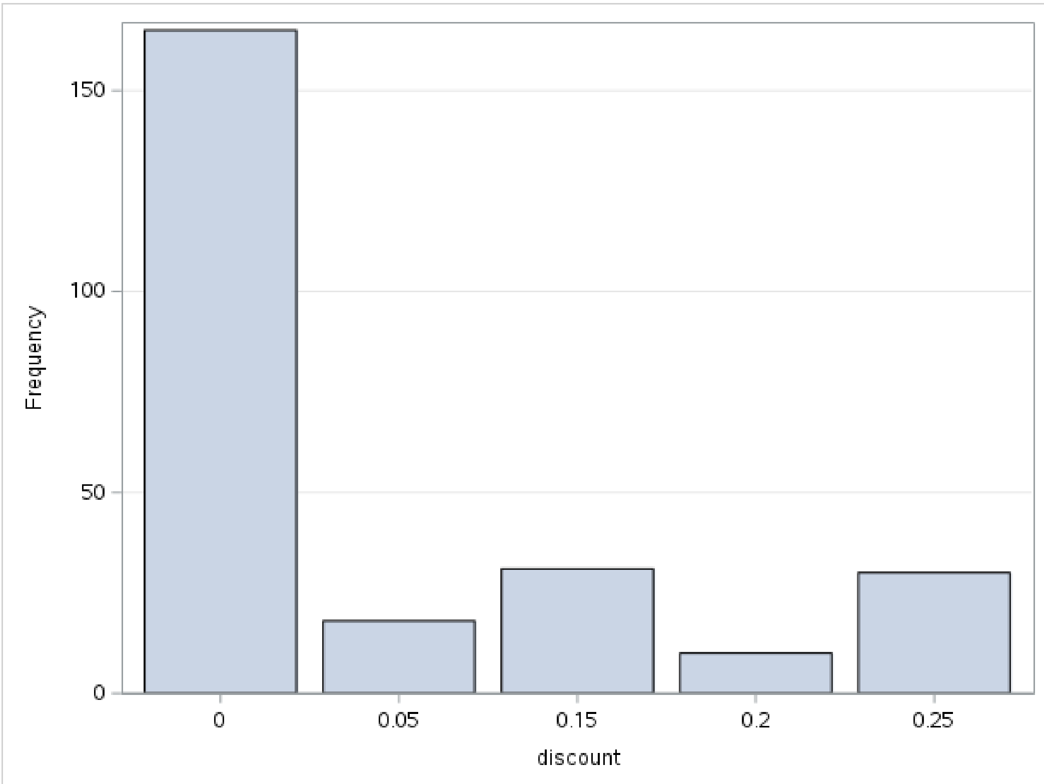
Variable	Mean	Median	N	Mode
unit_price	20.0133858	15.2000000	254	15.2000000
quantity	24.3307087	20.0000000	254	15.0000000
product_id	39.6220472	39.0000000	254	2.0000000
discount	0.0592520	0	254	0
gross_sale	492.9751969	288.0000000	254	168.0000000

Measures of spread or dispersion – Orders

Variable	Std Dev	Minimum	Maximum	N	Variance	Range	Lower Quartile	Upper Quartile
unit_price	15.4456608	2.0000000	99.0000000	254	238.5684367	97.0000000	10.4000000	26.2000000
quantity	15.7663549	1.0000000	70.0000000	254	248.5779465	69.0000000	12.0000000	35.0000000
product_id	23.0138055	2.0000000	77.0000000	254	529.6352432	75.0000000	20.0000000	59.0000000
discount	0.0916635	0	0.2500000	254	0.0084022	0.2500000	0	0.1500000
gross_sale	543.3813442	20.8000000	3080.00	254	295263.29	3059.20	167.4000000	640.0000000

**SAS code used to obtain reports can be found in Appendix A.*

Discount Groups



**SAS code used to obtain this chart can be found in Appendix A.*

5.0 Linear Regression – PROMOS and AVRG

A simple linear regression was applied to sales data to discover the relationship between the number of marketing promotions (PROMOS) and the average amount spent per visit (AVRG) for credit card users (CC_CARD = 1). The independent variable was the number of marketing promotions (PROMOS), and the average amount spent per visit (AVRG) was the dependent variable. This analysis was filtered with a where clause where CC_CARD=1.

Model: MODEL1
Dependent Variable: AVRG

Number of Observations Read	11031
Number of Observations Used	11031

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1918374	1918374	233.81	<.0001
Error	11029	90490077	8204.73999		
Corrected Total	11030	92408451			

Root MSE	90.58002	R-Square	0.0208
Dependent Mean	120.37145	Adj R-Sq	0.0207
Coeff Var	75.25042		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	149.39441	2.08480	71.66	<.0001
PROMOS	1	-1.90555	0.12462	-15.29	<.0001

**SAS code used to obtain reports can be found in Appendix A.*

Below is a summary of the linear regression analysis on PROMOS and AVRG:

Statistical Significance:

- The small F-value (233.81) and the associated p-value (less than 0.0001) indicate that the overall model is statistically significant. In other words, there's a significant relationship between the number of marketing promotions (PROMOS) and the average amount spent per visit (AVRG).

Model Explanation:

- Despite the statistical significance, the small F-value suggests that the model does not explain a large portion of the variance in AVRG. This means that while PROMOS has an impact on AVRG, there are likely other variables not included in the model that also influence AVRG.

Predictive Accuracy:

- The RMSE (Root Mean Square Error) of 90.58002 means that on average, the predicted AVRG deviates from the actual AVRG by approximately 90.58002 units. When compared to the AVRG range of 1919.39, the RMSE accounts for about 4.7% of the total range, suggesting the model performs well in terms of predictive accuracy.

Model Fit:

- The R-Squared value of 0.0208 indicates that the model explains only 2.08% of the variance in AVRG based on PROMOS. This implies that PROMOS alone is not a strong predictor of changes in AVRG.

Coefficient of PROMOS:

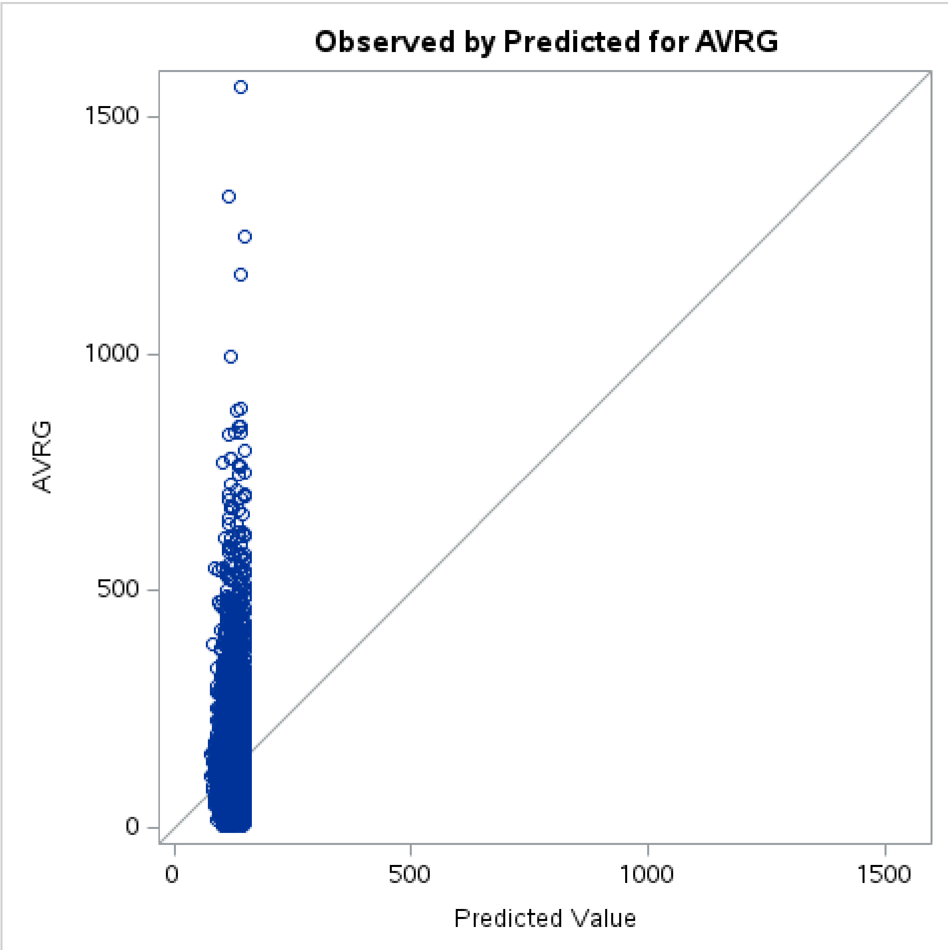
- The coefficient for PROMOS is -1.90555, meaning that for every one-unit increase in PROMOS, AVRG is expected to decrease by 1.90555 units. The fact that this coefficient is statistically significant (p-value much smaller than 0.05) supports the idea that PROMOS has an impact on AVRG.

Model Limitations:

- The scatterplot suggests a lack of a strong linear relationship between the predicted and observed values of AVRG, indicating that the linear regression model may not fully capture the relationship between PROMOS and AVRG for credit card users.

In summary, while the model shows that marketing promotions significantly impact the average amount spent per visit, it also highlights that there are other factors at play and that a simple linear model might not be the best fit for this relationship.

Linear Regression scatterplot for AVRG and PROMOS



The scatter plot data points are clustered near zero, with a few extreme outliers on the observed axis. This pattern suggests that the linear model does not accurately predict the higher AVRG values. Additionally, the absence of a visible trend indicates that the relationship between PROMOS and AVRG may not be linear. To address this, we tested alternative models for optimizing sales data, including nonlinear regression, polynomial regression, and the generalized additive model (GAM).

6.0 Nonlinear Regression

The Marquardt method within the PROC NLIN procedure was used to create a nonlinear regression on the sales data.

NLIN Procedure (AVRG and PROMOS)

The NLIN Procedure
Dependent Variable AVRG
Method: Marquardt

Iterative Phase			
Iter	a	b	Sum of Squares
0	150.0	-0.0200	91209044
1	154.1	-0.0165	90358310
2	153.8	-0.0166	90357253
3	153.8	-0.0166	90357253

NOTE: Convergence criterion met.

Estimation Summary	
Method	Marquardt
Iterations	3
R	1.064E-6
PPC(b)	6.744E-6
RPC(b)	0.000115
Object	3.53E-10
Objective	90357253
Observations Read	11031
Observations Used	11031
Observations Missing	0

Note: An intercept was not specified for this model.

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	2	1.6188E8	80941253	9879.68	<.0001
Error	11029	90357253	8192.7		
Uncorrected Total	11031	2.5224E8			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
a	153.8	2.3585	149.2	158.4
b	-0.0166	0.000999	-0.0185	-0.0146

Approximate Correlation Matrix		
	a	b
a	1.0000000	-0.8858759
b	-0.8858759	1.0000000

**SAS code used to obtain reports can be found in Appendix A.*

7.0 Polynomial Regression

Polynomial Regression (AVRG and PROMOS)

Data Set	WORK.SALES_CC1
Dependent Variable	AVRG
Selection Method	None

Number of Observations Read	11031
Number of Observations Used	11031

Dimensions	
Number of Effects	4
Number of Parameters	4

Least Squares Summary			
Step	Effect Entered	Number Effects In	SBC
0	Intercept	1	99655.2461
1	PROMOS	2	99433.1437
2	PROMOS*PROMOS	3	99373.4402*
3	PROMOS*PROMOS*PROMOS	4	99375.1896
* Optimal Value of Criterion			

Least Squares Model (No Selection)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2544330	848110	104.07	<.0001
Error	11027	89864121	8149.46233		
Corrected Total	11030	92408451			

Root MSE	90.27437
Dependent Mean	120.37145
R-Square	0.0275
Adj R-Sq	0.0273
AIC	110379
AICC	110379
SBC	99375

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	175.434315	4.042080	43.40	<.0001
PROMOS	1	-8.075621	1.052618	-7.67	<.0001
PROMOS*PROMOS	1	0.334443	0.077395	4.32	<.0001
PROMOS*PROMOS*PROMOS	1	-0.004572	0.001663	-2.75	0.0060

**SAS code used to obtain reports can be found in Appendix A.*

8.0 Generalized Additive Model (GAM)

The generalized additive model (GAM) illustrated below reveal significant findings. The chi-square test statistic is 90.9246, indicating a notable difference between the observed and expected values. The p-value is less than .0001, providing strong evidence against the null hypothesis and suggesting a statistically significant relationship between PROMOS (independent variable) and AVRG (dependent variable). The chi-square test evaluates whether PROMOS impacts AVRG across three categories or groups, supported by the degrees of freedom (DF) value of 3.00. Overall, the high chi-square statistic and the very low p-value strongly indicate that changes or differences in PROMOS are likely associated with changes or differences in AVRG, rather than occurring by random chance.

Generalized Additive Model (AVRG and PROMOS)

Dependent Variable: AVRG
Smoothing Model Component(s): spline(PROMOS)

Summary of Input Data Set	
Number of Observations	11031
Number of Missing Observations	0
Distribution	Gaussian
Link Function	Identity

Iteration Summary and Fit Statistics	
Final Number of Backfitting Iterations	2
Final Backfitting Criterion	5.574957E-31
The Deviance of the Final Estimate	89749964.612

The backfitting algorithm converged.

Regression Model Analysis
Parameter Estimates

Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	149.39441	2.07654	71.94	<.0001
Linear(PROMOS)	-1.90555	0.12413	-15.35	<.0001

Smoothing Model Analysis
Fit Summary for Smoothing Components

Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Spline(PROMOS)	0.999895	3.000000	5632.645146	39

Smoothing Model Analysis
Analysis of Deviance

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(PROMOS)	3.00000	740113	90.9246	<.0001

**SAS code used to obtain reports can be found in Appendix A.*

The nonlinear regression (NLIN), polynomial regression, and the generalized additive model (GAM) each had their strengths and weaknesses. Research shows that the polynomial regression would be the worst choice due to low R^2 (0.0275) and would not explain the variance in AVRG. If increasing PROMOS has a progressively smaller effect on AVRG, then the NLIN model would fit well. Otherwise, the generalized additive model (GAM) is the most flexible and can best capture nonlinearity.

9.0 Linear Regression – frequency of purchase visits and total net sales

A simple linear regression was applied to the sales data to examine the frequency of purchase visits (FRE) and how it affects the total net sales (MON). Using SAS, we assigned the frequency of purchase visits (FRE) as the independent variable and the total net sales (MON) as the dependent variable.

Linear Regression results for frequency of purchase visits (FRE) and total net sales (MON)

Model: MODEL1
Dependent Variable: MON

Number of Observations Read	28799
Number of Observations Used	28799

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6340894287	6340894287	29556.5	<.0001
Error	28797	6177960274	214535		
Corrected Total	28798	12518854561			

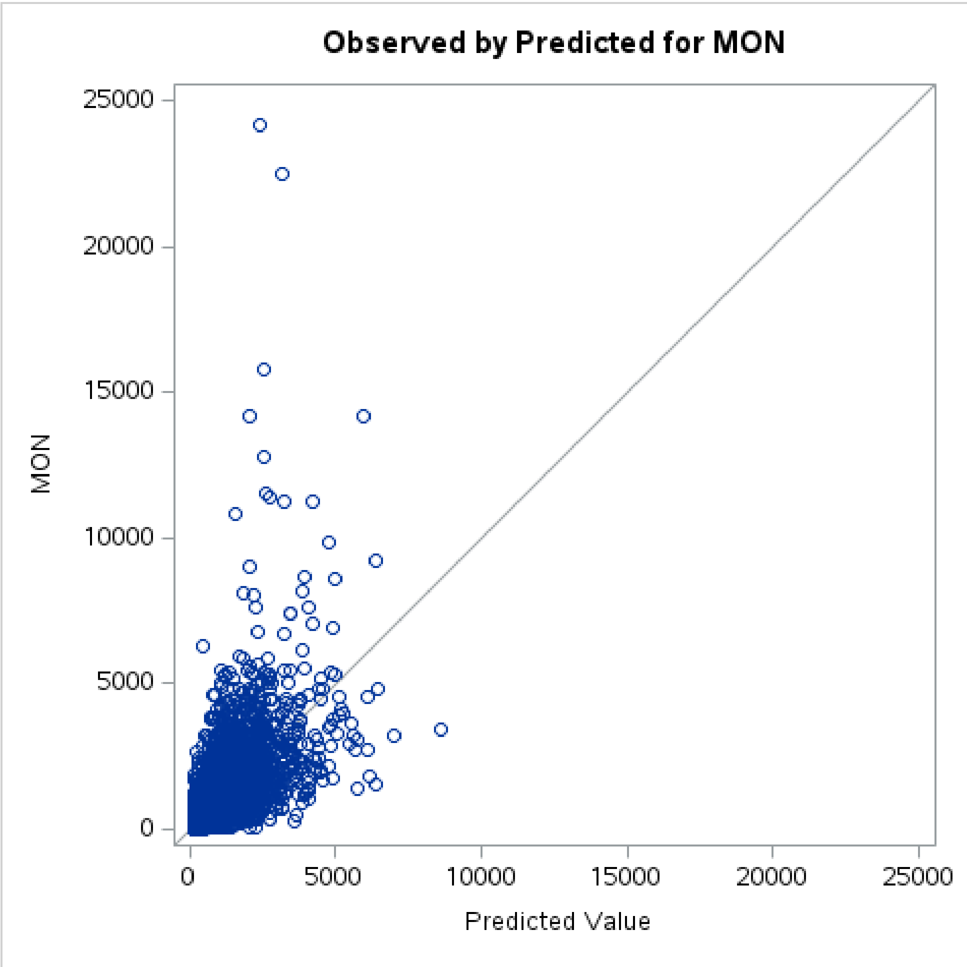
Root MSE	463.17908	R-Square	0.5065
Dependent Mean	473.21246	Adj R-Sq	0.5065
Coeff Var	97.87973		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	100.79734	3.48452	28.93	<.0001
FRE	1	73.90613	0.42989	171.92	<.0001

The ANOVA analysis above shows a statistically significant model with a large F-value of 29556.5 and a p-value of less than 0.0001. The RMSE of 463.17908 indicates that the predicted total net sales (MON) deviate from actual sales by about 463.18 units on average, accounting for approximately 1.92% of the total sales range. An R-Square value of 0.5065 suggests that the model explains 50.65% of the variance in sales based on the frequency of purchase visits (FRE). The coefficient for FRE is 73.90613, meaning that for each additional visit, total sales are expected to increase by 73.90613 units. The p-value indicates strong evidence supporting the significant impact of FRE on MON.

The scatterplot below confirms a positive correlation between FRE and MON, suggesting that higher visit frequency leads to higher sales.

Linear Regression scatterplot for FRE and MON



10.0 T-test: comparing average unit price and quantity

A t-test was applied to the orders data to analyze the average unit prices and quantities for the top-selling products. Using SAS, we first calculated the total sales per product. Then we identified the top-selling products and applied a flag to those records. The results from the T-test for unit price and quantities are shown below.

T-test results – unit price

Variable: unit_price

is_top_seller	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		14	15.2000	0	0	15.2000	15.2000
1		240	20.2942	15.8464	1.0229	2.0000	99.0000
Diff (1-2)	Pooled		-5.0942	15.4322	4.2430		
Diff (1-2)	Satterthwaite		-5.0942		1.0229		

is_top_seller	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		15.2000	15.2000	15.2000	0	.	.
1		20.2942	18.2792	22.3092	15.8464	14.5442	17.4066
Diff (1-2)	Pooled	-5.0942	-13.4505	3.2621	15.4322	14.1945	16.9082
Diff (1-2)	Satterthwaite	-5.0942	-7.1092	-3.0792			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	252	-1.20	0.2310
Satterthwaite	Unequal	239	-4.98	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	239	13	Infty	<.0001

T-test results – quantity

Variable: quantity

is_top_seller	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		14	29.6429	11.1742	2.9864	20.0000	50.0000
1		240	24.0208	15.9561	1.0300	1.0000	70.0000
Diff (1-2)	Pooled		5.6220	15.7450	4.3290		
Diff (1-2)	Satterthwaite		5.6220		3.1590		

is_top_seller	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		29.6429	23.1911	36.0946	11.1742	8.1008	18.0021
1		24.0208	21.9919	26.0498	15.9561	14.6449	17.5272
Diff (1-2)	Pooled	5.6220	-2.9037	14.1477	15.7450	14.4822	17.2510
Diff (1-2)	Satterthwaite	5.6220	-1.0660	12.3101			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	252	1.30	0.1952
Satterthwaite	Unequal	16.264	1.78	0.0938

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	239	13	2.04	0.1423

The T-test results reveal that the average unit prices between top-selling products and other products significantly differ. This conclusion stems from the observed t-value of -4.98 and a p-value of less than 0.0001, indicating that this difference is not due to random chance, leading us to confidently reject the null hypothesis for unit prices. Conversely, when examining the quantities sold, the T-test results show a t-value of 1.3 and a p-value of 0.1423. This implies that the difference in quantities sold is not statistically significant, suggesting that any observed difference could be due to random variation. Consequently, we fail to reject the null hypothesis for quantities. These findings highlight the importance of pricing strategies in the success of top-selling products, as their higher unit prices appear to contribute to their overall success, whereas the quantities sold do not exhibit a significant difference.

11.0 MANOVA – examine gross sales, net sales and discount

ANOVA – gross sales and discount

The Multivariate Analysis of Variance (MANOVA) was applied to the orders data to examine whether discount rates impact gross sales and net sales. Using SAS, we wrote a MANOVA code (see Appendix A) to include both gross sales and net sales as the dependent variables and assigned discount as the categorical variable. The results of our MANOVA test is shown below. First we list the ANOVA results of gross sales and discount.

Univariate ANOVA results of gross sales and discount.

Class Level Information		
Class	Levels	Values
discount	5	0 0.2 0.05 0.15 0.25

Number of Observations Read	254
Number of Observations Used	254

Dependent Variable: gross_sale

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2080615.89	520153.97	1.78	0.1327
Error	249	72620995.27	291650.58		
Corrected Total	253	74701611.15			

R-Square	Coeff Var	Root MSE	gross_sale Mean
0.027852	109.5485	540.0468	492.9752

Source	DF	Anova SS	Mean Square	F Value	Pr > F
discount	4	2080615.888	520153.972	1.78	0.1327

The ANOVA (Analysis of Variance) test results shown above provide an F-statistic of 1.78 and a p-value of 0.1327. The F-statistic represents the ratio of variance between groups to the variance within groups, aiding in determining the significance of differences among group means. The p-value indicates the probability of observing the test results under the null hypothesis, with a value less than 0.05 generally considered statistically significant. In this case, the p-value of 0.1327 exceeds the threshold of 0.05, indicating insufficient evidence to reject the null hypothesis. Consequently, there is not enough evidence to support the claim that discount rates have a significant impact on gross sales. Therefore, based on these results, discount rates do not appear to significantly affect gross sales in the given data.

ANOVA – net sales and discount

The ANOVA test results below show an F-statistic of 0.57 and a p-value of 0.6847. The F-statistic, which represents the ratio of variance between groups to the variance within groups, helps determine whether the differences among group means are significant. The p-value indicates the probability of observing the test results under the null hypothesis, with a value less than 0.05 generally considered statistically significant. In this case, the p-value of 0.6847 is much higher than the 0.05 threshold, suggesting there is insufficient evidence to reject the null hypothesis. As a result, there is not enough evidence to support the claim that discount rates have a significant impact on net sales. Thus, based on these findings, discount rates do not appear to have a meaningful influence on net sales in the given data.

Univariate ANOVA results of net sales and discount.

Dependent Variable: net_sale

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	564406.98	141101.74	0.57	0.6847
Error	249	61646720.31	247577.19		
Corrected Total	253	62211127.29			

R-Square	Coeff Var	Root MSE	net_sale Mean
0.009072	108.9901	497.5713	456.5290

Source	DF	Anova SS	Mean Square	F Value	Pr > F
discount	4	564406.9784	141101.7446	0.57	0.6847

MANOVA – gross sales, net sales and discount

Although the Univariate test tells us that there is no significant effect of discount on gross sales and net sales individually, the Multivariate Analysis of Variance (MANOVA) test analyzes the combined effect of discount on both dependent variables (gross sales & net sales).

Multivariate Analysis of Variance (MANOVA) results of gross sales, net sales, and discount.

Multivariate Analysis of Variance

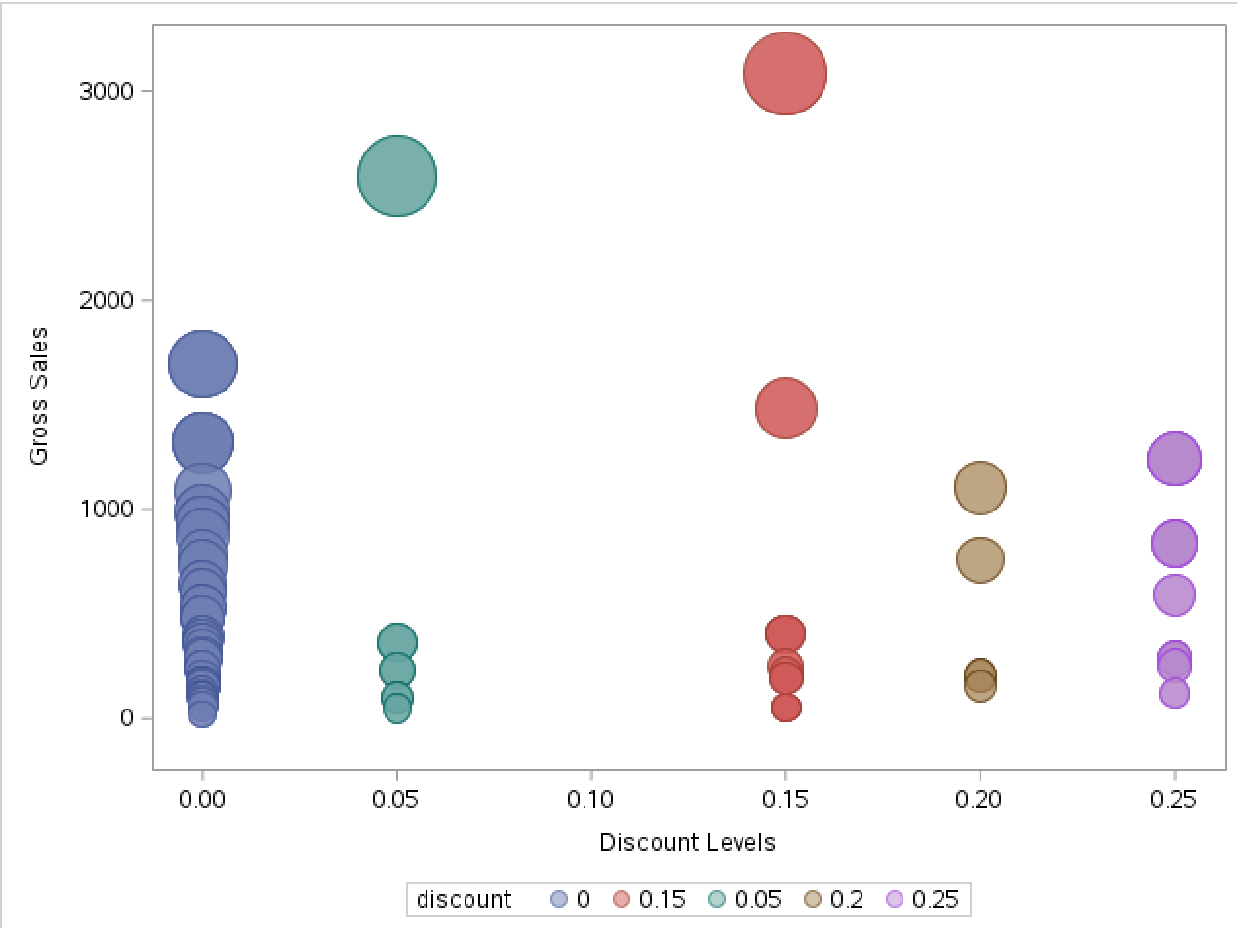
Characteristic Roots and Vectors of: E Inverse * H, where H = Anova SSCP Matrix for discount E = Error SSCP Matrix			
		Characteristic Vector V'EV=1	
Characteristic Root	Percent	gross_sale	net_sale
1.40781381	99.48	-0.00134356	0.00144806
0.00730215	0.52	-0.00004891	0.00018016

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall discount Effect H = Anova SSCP Matrix for discount E = Error SSCP Matrix S=2 M=0.5 N=123					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.41230380	34.56	8	496	<.0001
Pillai's Trace	0.59193471	26.17	8	498	<.0001
Hotelling-Lawley Trace	1.41511596	43.76	8	351.97	<.0001
Roy's Greatest Root	1.40781381	87.64	4	249	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

The MANOVA (Multivariate Analysis of Variance) results above show that Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Greatest Root all have p-values of less than 0.0001. This highly significant result suggests rejecting the null hypothesis, indicating that discount rates do have a significant impact on the combined measures of gross sales and net sales.

To further illustrate this relationship, a multidimensional bubble chart below represents the interactions between gross sales, net sales, and discounts, with the size of the bubbles corresponding to the net sales. Through this visual representation, we can gain a more intuitive understanding of how these variables relate to one another, confirming that discount rates significantly influence both gross sales and net sales.

Multidimensional Bubble Chart



12.0 Conclusion

The statistical analyses on the clothing retail store and distribution company's data offer valuable insights into how marketing promotions, purchase frequency, pricing strategies, and discounts affect sales performance:

- **Marketing Promotions:** Regression analysis showed that the number of promotions significantly impacts the average amount spent per visit, but only explains 2.08% of the variance. The generalized additive model (GAM) better captures this relationship, suggesting other factors also play a role.
- **Purchase Frequency:** A strong relationship exists between purchase frequency and total net sales, with an R-squared value of 0.5065. Increasing customer visit frequency can significantly boost total net sales.
- **Pricing Strategies:** The t-test revealed a significant difference in unit prices between top-selling and other products, but no significant difference in quantities sold. This implies that pricing strategies are critical for the success of top-selling products.
- **Discount Rates:** Individual ANOVA tests did not show a significant impact of discount rates on sales. However, the MANOVA test indicated a significant combined effect on multiple business metrics, suggesting that discount rates impact should be considered more broadly.

Recommendations

- Refine promotional strategies to focus on profitability rather than just increasing sales.
- Encourage repeat customer visits through loyalty programs or personalized marketing.
- Optimize pricing strategies for top-selling products.
- Explore how discounts influence customer behavior to improve pricing and promotions.

These insights can help the company enhance its direct marketing efforts, refine pricing strategies, and drive overall business growth.

13.0 Appendix A

```
/* (4.0) Find the central tendency of the sales data using PROMOS, AVRG, CC_CARD, FRE, and MON */
proc means data=WORK.SALES chartype mean median n mode vardef=df qmethod=os;
	var PROMOS AVRG CC_CARD FRE MON;
run;

/* ---------------------------------------------------------------------------------------------------- */
/* ---------------------------------------------------------------------------------------------------- */
/* (4.0) Find the measures of spread or dispersion of the sales data with a focus on PROMOS, AVRG,
CC_CARD, FRE, and MON */
proc means data=WORK.SALES chartype std min max var range vardef=df q1 q3 qmethod=os;
	var PROMOS AVRG CC_CARD FRE MON;
run;

/* ---------------------------------------------------------------------------------------------------- */
/* ---------------------------------------------------------------------------------------------------- */
/* (4.0) Find the central tendency of the orders data with a focus on unit_price, quantity, product_id,
discount, and gross_sale */
proc means data=WORK.ORDERS chartype mean median n mode vardef=df qmethod=os;
	var unit_price quantity product_id discount gross_sale;
run;

/* ---------------------------------------------------------------------------------------------------- */
/* ---------------------------------------------------------------------------------------------------- */
/* (4.0) Find the measures of spread or dispersion of the sales data with a focus on unit_price, quantity,
product_id, discount, and gross_sale */
proc means data=WORK.ORDERS chartype std min max n var range vardef=df q1 q3 qmethod=os;
	var unit_price quantity product_id discount gross_sale;
run;

/* ---------------------------------------------------------------------------------------------------- */
/* ---------------------------------------------------------------------------------------------------- */
/* Create a bar chart on the orders data to compare each discount group and to show which group is most
common */
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=WORK.ORDERS;
	vbar discount /;
	yaxis grid;
run;

ods graphics / reset;

/* ---------------------------------------------------------------------------------------------------- */
/* ---------------------------------------------------------------------------------------------------- */
/* (5.0) Performs simple linear regression to the sales data,
the dependent variable = AVRG, the independent variable = PROMOS,
the data was limited where CC_CARD = 1
*/
proc reg data=WORK.SALES(where=(CC_CARD=1)) alpha=0.05
	plots(only maxpoints=none)=(diagnostics residuals fitplot observedbypredicted);
	model AVRG=PROMOS /;
run;
quit;
```

```
/* ------------------------------------------------------------------------------------------------- */
/* ------------------------------------------------------------------------------------------------- */
/* (6.0) NONLINEAR REGRESSION */
/* First filter the data and create a new table where CC_CARD = 1 */
proc sql noprint;
        create table work.filter as select * from WORK.SALES where(CC_CARD EQ 1);
quit;

/* Next perform nonlinear regression to address business question 1*/
PROC NLIN data=WORK.FILTER METHOD=MARQUARDT;
        PARAMETERS a=150 b=-0.02;
        MODEL AVRG=a * EXP(b * PROMOS);
        OUTPUT OUT=PredictedData P=Predicted_AVRG;
RUN;

/* ------------------------------------------------------------------------------------------------- */
/* ------------------------------------------------------------------------------------------------- */
/* (7.0) This code performs polynomial regression*/
proc glmselect data=WORK.FILTER outdesign(addinputvars)=Work.reg_design;
        model AVRG=PROMOS PROMOS*PROMOS PROMOS*PROMOS*PROMOS / showpvalues
        selection=none;
run;

proc reg data=Work.reg_design alpha=0.05 plots(only maxpoints=none)=(diagnostics residuals
observedbypredicted);
        ods select DiagnosticsPanel ResidualPlot ObservedByPredicted;
        model AVRG=&_GLSMOD /;
run;
quit;

/* ------------------------------------------------------------------------------------------------- */
/* ------------------------------------------------------------------------------------------------- */
/* (8.0) This code performs generalized additive model */
PROC GAM data=WORK.FILTER;
        MODEL AVRG=SPLINE(PROMOS);
        OUTPUT OUT=PredictedData P=Predicted_AVRG;
RUN;

/* ------------------------------------------------------------------------------------------------- */
/* ------------------------------------------------------------------------------------------------- */
/* (9.0) This code performs linear regression on the sales data */
proc reg data=WORK.SALES alpha=0.05 plots(only maxpoints=none)=(diagnostics
        residuals fitplot observedbypredicted);
        model MON=FRE /;
run;
quit;

/* ------------------------------------------------------------------------------------------------- */
/* ------------------------------------------------------------------------------------------------- */
/* (10.0) T-test: comparing average unit price and quantity */
/* First we prep the data for a T-test on the orders data by flagging top-selling products */
/* Calcualte total sales per product */
proc sql;
create table product_sales as select product_id, sum(net_sale) as total_sales
        from WORK.ORDERS group by product_id;
quit;

/* Identify top 10% products */
proc univariate data=product_sales noprint;
        var total_sales;
        output out=quantile pctlpts=90 pctlpre=Q;
run;

data top_sellers_flag;
        merge product_sales quantile;
        if total_sales >=Q90 then
        is_top_seller=1;
        else is_top_seller=0;
run;
```

```
/* Merge the flagged data back to the orginal dataset */
proc sql;
	create table flagged_data as
	select a.*, b.is_top_seller
	from WORK.ORDERS a
	left join top_sellers_flag b
	on a.product_id - b.product_id;
quit;

/* Compute average unit price and quantity for top-selling products vs out products */
proc means data=flagged_data mean;
	class is_top_seller;
	var unit_price quantity;
run;


/* ---------------------------------------------------------------------------------------------------- */
/* ---------------------------------------------------------------------------------------------------- */
/* (11.0) MANOVA - Examine gross sales, net sales and discount */
/* Perform MANOVA */
proc anova data=WORK.ORDERS;
	class discount;
	model gross_sale net_sale=discount;
	manova h=discount / printe;
run;
quit;

/* Create multidimensional bubble chart */
proc sgplot data=WORK.ORDERS;
	bubble x=discount y=gross_sale size=net_sale / group=discount transparency=0.5;
	xaxis label="Discount Levels";
	yaxis label="Gross Sales:;
run;
```