Clothing Retail Store& Distribution Company with SAS Studio

Didem B. Aykurt

Colorado State University Global

MIS543: Enterprise Performance Management

Dr. John Marlowe

November 5, 2023

Table of Contents

A Comprehensive Analysis of Sales and Orders

In today's highly competitive retail landscape, data-driven decision-making has become the cornerstone of business success. With the advent of advanced analytics, stores can gain deep insights into customer behavior, sales performance, and order dynamics, enabling them to optimize their strategies and stay ahead of the curve. This comprehensive analysis delves into the worlds of store sales and orders, exploring critical facets of these domains to extract valuable insights.

The research aim is to uncover hidden patterns, correlations, and opportunities within the data to inform strategic decisions and enhance overall store performance. The paper will begin by examining store sales, focusing on the impact of credit card usage and the effectiveness of marketing promotions. This research will shift our attention to store orders, where the company will evaluate employee performance and pricing strategies. The organization, a Clothing Retail Store & Distribution Company, faces several key challenges and strategic goals:

1. Increase Quantity and Profitability of Orders: The organization aims to boost the quantity and profitability of customer orders. This objective reflects the need to enhance revenue streams and ensure that each order contributes positively to the company's financial performance (Cox, 2012).

2. Predict Future Business Growth: The organization seeks insights into predicting future business growth trends. Accurate forecasting can help the company make informed decisions, allocate resources effectively, and respond proactively to changes in the market (Cox, 2012).

Dataset Description:

Clothing_Store_Sales.csv: This dataset contains information related to store sales, including customer behavior, sales performance, and various attributes that may influence sales.

Figure 1

The table of Clothing Store Sales variables in SAS Studio.

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
6	AVRG	Num	8	BEST12.	BEST32.
5	CC_CARD	Num	8	BEST12.	BEST32.
26	CLUSTYPE	Num	8	BEST12.	BEST32.
1	Customer Id	Num	8	BEST12.	BEST32.
24	DAYS	Num	8	BEST12.	BEST32.
3	FRE	Num	8	BEST12.	BEST32.
22	GMP	Num	8	BEST12.	BEST32.
25	MARKDOWN	Num	8	BEST12.	BEST32.
4	MON	Num	8	BEST12.	BEST32.
10	PBLOUSES	Num	8	BEST12.	BEST32.
12	PCAR_PNTS	Num	8	BEST12.	BEST32.
13	PCAS_PNTS	Num	8	BEST12.	BEST32.
21	PCOLLSPND	Num	8	BEST12.	BEST32.
15	PDRESSES	Num	8	BEST12.	BEST32.
27	PERCRET	Num	8	BEST12.	BEST32.
19	PFASHION	Num	8	BEST12.	BEST32.
11	PJACKETS	Num	8	BEST12.	BEST32.
18	PJEWELRY	Num	8	BEST12.	BEST32.
9	PKNIT_DRES	Num	8	BEST12.	BEST32.
8	PKNIT_TOPS	Num	8	BEST12.	BEST32.
20	PLEGWEAR	Num	8	BEST12.	BEST32.
17	POUTERWEAR	Num	8	BEST12.	BEST32.
23	PROMOS	Num	8	BEST12.	BEST32.
14	PSHIRTS	Num	8	BEST12.	BEST32.
16	PSUITS	Num	8	BEST12.	BEST32.
7	PSWEATERS	Num	8	BEST12.	BEST32.
2	ZIP_CODE	Num	8	BEST12.	BEST32.
28	In days between purchases	Num	8	BEST12.	BEST32.
29	In lifetime ave time betw visits	Num	8	BEST12.	BEST32.

Clothing_Store_Orders.csv: This dataset pertains to store orders, encompassing details about individual orders, products, quantities, and associated factors.

Figure 2

The table of Clothing Store Order variables in SAS Studio.

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
6	category_id	Num	8	BEST12.	BEST32.
2	customer_id	Char	5	$5.	$5.
9	discount	Num	8	BEST12.	BEST32.
11	discount_amt	Num	8	BEST12.	BEST32.
3	employee_id	Num	8	BEST12.	BEST32.
10	gross_sale	Num	8	BEST12.	BEST32.
12	net_sale	Num	8	BEST12.	BEST32.
1	order_id	Num	8	BEST12.	BEST32.
5	product_id	Num	8	BEST12.	BEST32.
8	quantity	Num	8	BEST12.	BEST32.
4	territory_id	Num	8	BEST12.	BEST32.
7	unit_price	Num	8	BEST12.	BEST32.

Business Questions and Hypotheses:

Store Sales Business Questions:

Business Question 1: Does using credit cards (CC_CARD) significantly impact the average amount spent per visit (AVRG)?

- *Null Hypothesis (H10):* There is no significant difference in the average amount spent per visit between customers who use credit cards and those who do not.
- *Alternate Hypothesis (H1):* Customers who use credit cards have a significantly higher average amount spent per visit than those who do not.

Business Question 2: Is there a relationship between the customer's lifetime average time between visits (Lifetime average time between visits in days) and the total net sales (MON)?

- *Null Hypothesis (H20):* No significant correlation exists between the customer's lifetime average time between visits and total net sales.

- *Alternate Hypothesis (H2):* A significant positive correlation exists between the customer's lifetime average time between visits and total net sales.

Store Orders Business Questions:

Business Question 3: Does the territory (territory_id) where an order is placed significantly impact the gross sale amount (gross_sale)?

- *Null Hypothesis (H30):* The territory where an order is placed does not significantly impact the gross sale amount.
- *Alternate Hypothesis (H3):* The territory where an order is placed significantly affects the gross sale amount.

Business Question 4: Is there a significant difference in the average unit price (unit_price) of products ordered by different customer groups (customer_id)?

- *Null Hypothesis (H40):* No significant difference exists in the average unit price of products ordered by different customer groups.
- *Alternate Hypothesis (H4):* A significant difference exists in the average unit price of products ordered by different customer groups.

Justification for Business Questions:

1. **Business Question 1 (Store Sales):** Understanding the impact of credit card usage on average spending per visit can help the organization tailor its marketing strategies and promotions, potentially increasing the profitability of each customer visit.
2. **Business Question 2 (Store Sales):** Establishing a correlation between the customer's average time between visits and total net sales can provide insights into customer behavior patterns, aiding in predicting future sales and optimizing marketing efforts.

3. **Business Question 3 (Store Orders):** Analyzing the influence of territory on gross sale amounts can guide decisions related to distribution, inventory management, and resource allocation, potentially increasing both the quantity and profitability of orders (Griva et al., 2018).
4. **Business Question 4 (Store Orders):** Identifying differences in unit prices based on customer groups can inform pricing strategies and customer segmentation, contributing to profitability and growth goals.

Statistical Tests Used and Why:

1. **T-Test (Hypothesis 1):**

A student's t-test will be used to assess the null hypothesis against Business Question 1. This test is appropriate when comparing means between two groups. This would be useful to compare the average amount spent per visit between credit card users and non-users. This test helps determine if there is a statistically significant difference in spending behavior between the two groups (Allison, 2018).

2. **Correlation Analysis (Hypothesis 2):**

Pearson's correlation analysis will be helpful for Business Question 2. It would assess the relationship between the customer's lifetime average time between visits and total net sales. This test helps to understand whether these two variables are significantly correlated. This test assesses the strength and direction of a linear relationship between two continuous variables (Allison, 2018).

3. **Analysis of Variance (ANOVA) (Hypothesis 3 and 4):**

ANOVA would be functional to address Business Questions 3 and 4. ANOVA can identify differences in means among multiple groups, making it suitable for comparing means when division is by categorical variables such as geographical territory. In Hypothesis 3, ANOVA would help determine if the territory where an order is placed significantly impacts gross sale amounts. In Hypothesis 4, ANOVA will be functional to determine the test of a significant difference in the average unit price of products ordered by different customer groups (Allison, 2018).

4. **Multiple Linear Regression:**

Multiple linear regression techniques would also be used for predictive modeling. It would predict total net sales (MON) using all available predictor variables. Regression helps identify which variables are significant predictors of the target variable and provides insights into their relationships.

Visualizations Used and Why:

1. **Histograms:** Histograms can visualize the distribution of continuous variables such as AVRG and ln lifetime average time between visits. This helps understand the data's central tendencies and variations (Allison, 2018).
2. **Scatter Plots:** Scatter plots will be employed to visualize the relationship between two continuous variables, such as ln lifetime average time between visits and MON. They help identify patterns and trends in the data (Allison, 2018).
3. **Bar Charts:** Bar charts can be used to display the means or sums of categorical data, such as the average unit price (unit_price) by different customer groups (customer_id) or the impact of territories (territory_id) on gross sale amounts (Allison, 2018).
4. **Box Plots:** Box plots help visualize the distribution and spread of data, especially when comparing multiple groups, as in ANOVA. They provide insights into the data's central tendency, variability, and potential outliers (Allison, 2018).

This paper provides a valuable discussion of the value of each form of data and analytics method brought to the big data project. That would need to thoroughly explain how the insights derived from the analysis helped the company make informed decisions or improve its operations. Therefore, to complete all milestones and successful data analysis in SAS, it is excellent to address these concerns. This will help to finally complete the Portfolio Project and provide valuable insights for achieving the organization's strategic goals. Creating a comprehensive analysis is a time-consuming task objective of big data projects.

Concerns in Completing the Portfolio Project:

Data Quality: Ensuring the accuracy, completeness, and reliability of the data used for analysis is critical. Any data quality issues could lead to incorrect conclusions (Griva et al., 2018).

Statistical Analysis: Properly selecting and conducting statistical tests appropriate for each hypothesis is crucial. Incorrect statistical analysis could yield unreliable results (Griva et al., 2018).

Interpretation and Communication: It is essential to interpret the findings accurately and communicate actionable insights to stakeholders effectively to drive strategic decisions (Griva et al., 2018).

Ethical Considerations: Handling customer data and making business decisions based on it must adhere to ethical guidelines and data privacy regulations (Cox, 2012).

Resource Allocation: Depending on the results, the organization may need to reallocate resources. Ensuring that resource reallocations align with the organization's goals is essential (Cox, 2012).

Predictive Modeling: If predicting future growth is a goal, developing robust predictive models may be necessary. The accuracy of these models should be validated (Cox, 2012).

Model Complexity: In regression analysis (Hypotheses 3 and 4), selecting the correct predictor variables is critical. Overly complex models may lead to overfitting, while overly simplified models may lose important information (SAS, 2022).

Interpreting Results: It is essential to interpret the statistical results in the context of the research questions and real-world implications (SAS, 2022). Statistical significance does not always equate to practical significance.

Analysis of findings concerning business questions and corresponding hypotheses

1. Data Import - "store_sales" and "store_orders":

Task: Import two datasets, "Clothing_Store_Sales.csv" and "Clothing_Store_Orders.csv," into SAS datasets named "store_sales" and "store_orders."

Code

```
/* Import store sales data */
proc import datafile="/home/u59861956/Clothing_Store_Sales.csv" out=store_sales replace;
  getnames=yes;
run;
/* Import store orders data */
proc import datafile="/home/u59861956/Clothing_Store_Orders.csv" out=store_orders replace;
  getnames=yes;
run;
```

Explanation: These code segments use the proc import procedure to read data from external CSV files and create SAS datasets. The getnames=yes option indicates that the first row of the CSV files contains variable names.

2. Descriptive Statistics for "AVRG" in "store_sales":

Task: Calculate summary statistics for the "AVRG" (average amount spent per visit) variable in the "store_sales" dataset.

Code:

```
/* Descriptive Statistics */
proc means data=store_sales;
  var AVRG;
  class CC_CARD;
run;
```

Explanation: This code segment uses the proc means procedure to compute summary statistics (e.g., mean, standard deviation) for "AVRG." It also stratifies the analysis by the "CC_CARD" variable, which represents credit card usage.

Figure 3

Result of descriptive statistics for AVRG sales by "CC_CARD" in SAS Studio.

The MEANS Procedure

Analysis Variable : AVRG						
CC_CARD	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	17768	17768	109.3771117	83.7587125	0.4900000	1919.88
1	11031	11031	120.3714477	91.5309752	3.3000000	1564.51

This output summarizes the average amount spent per visit for customers who do and do not use credit cards. It includes information about the number of observations, mean, standard deviation, minimum, and maximum values for each group.

3. t-test for Hypothesis 1:

Task: Perform a t-test to compare the average amount spent per visit between credit card users and non-users.

Code:

```
/* t-test for Hypothesis 1 */
proc ttest data=store_sales;
  class CC_CARD;
  var AVRG;
  ods select ttest=ttest_results;
run;
```

Explanation: This code section uses the proc t-test procedure to conduct a t-test. It tests Hypothesis 1, comparing "AVRG" between credit card users (CC_CARD=1) and non-users (CC_CARD=0). The results are saved in the "ttest_results" dataset.

Figure 4

The table results of the t-test for AVRG from the store sales table by "CC_CARD" in SAS Studio.

The TTEST Procedure

Variable: AVRG

CC_CARD	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		17768	109.4	83.7587	0.6284	0.4900	1919.9
1		11031	120.4	91.5310	0.8715	3.3000	1564.5
Diff (1-2)	Pooled		-10.9943	86.8179	1.0524		
Diff (1-2)	Satterthwaite		-10.9943		1.0744		

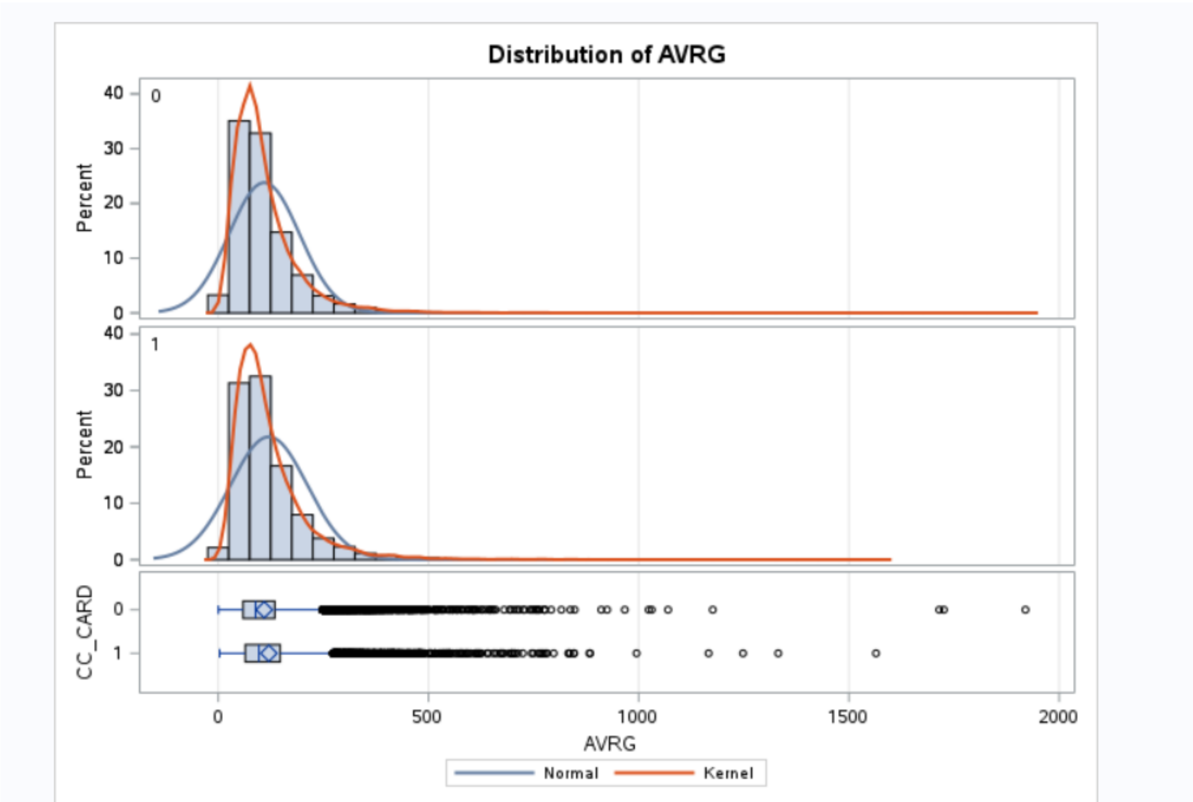
CC_CARD	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		109.4	108.1	110.6	83.7587	82.8969	84.6388
1		120.4	118.7	122.1	91.5310	90.3390	92.7551
Diff (1-2)	Pooled	-10.9943	-13.0570	-8.9316	86.8179	86.1147	87.5329
Diff (1-2)	Satterthwaite	-10.9943	-13.1002	-8.8884			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28797	-10.45	<.0001
Satterthwaite	Unequal	21819	-10.23	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	11030	17767	1.19	<.0001

Figure 5

The t-test result shows the AVRG graph distribution by CC-CARD's variables 0 and 1 in SAS Studio.



Both methods show significant differences in means with very low p-values (Pr > |t| < 0.0001), indicating a significant difference in the average amount spent per visit between customers who use credit cards and those who do not.

The "Folded F" test indicates that there is a significant difference in variances (Pr > F < 0.0001), suggesting that the assumption of equal variances is violated.

In summary, the t-test results suggest a statistically significant difference in the average amount spent per visit between customers who use credit cards and those who do not. Customers who use credit cards tend to spend more on average. Additionally, the assumption of equal variances is violated, which is essential in interpreting the t-test results.

4. Correlation Analysis for Hypothesis 2:

Task: Calculate the correlation between "MON" (total net sales) and "ln lifetime ave time between visits" in the "store_sales" dataset.

Code:

```
/* Correlation Analysis for Hypothesis 2 */
proc corr data=store_sales;
  var MON 'ln lifetime ave time betw visits'n;
run;
```

Explanation: This code segment employs the proc corr procedure to compute the correlation between "MON" and "ln lifetime ave time betw visits." It addresses Hypothesis 2, which explores the relationship between these variables.

Figure 6

The result of the correlation analysis between the customer's lifetime average time between visits and total net sales "MON."

The CORR Procedure

2 Variables:	MON ln lifetime ave time betw visits

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
MON	28799	473.21246	659.32741	13628046	0.99000	24140
ln lifetime ave time betw visits	28799	3.92374	1.02042	113000	-2.41000	5.90000

Pearson Correlation Coefficients, N = 28799 Prob > |r| under H0: Rho=0		
	MON	ln lifetime ave time betw visits
MON	1.00000	-0.50520 <.0001
ln lifetime ave time betw visits	-0.50520 <.0001	1.00000

The output suggests a statistically significant negative correlation between the total net sales (MON) and the natural logarithm of the lifetime average time between visits (ln lifetime ave time betw visits). As total net sales increase, the natural logarithm of the lifetime average time between visits decreases, indicating an inverse relationship between these two variables.

5. Analysis of Variance (ANOVA) for Hypothesis 3:

Task: Perform an ANOVA to test Hypothesis 3, evaluating the impact of "territory_id" on "gross_sale" in the "store_orders" dataset.

Code:

```
/* Analysis of Variance (ANOVA) for Hypothesis 3 */
proc glm data=store_orders;
  class territory_id;
  model gross_sale = territory_id;
  means territory_id / hovtest;
run;
```

Explanation: This code section uses the proc glm procedure to conduct an ANOVA. It tests Hypothesis 3, which examines the effect of "territory_id" on "gross_sale" in the "store_orders" dataset. The "hovtest" option assesses the homogeneity of variances.

Figure 7

The ANOVA test results for the territory with gross sales in the store orders dataset.

The GLM Procedure

Class Level Information

Class	Levels	Values
territory_id	39	1581 1730 1833 2116 2139 2184 2903 3049 3801 6897 7960 8837 10019 10038 11747 14450 19428 19713 20852 27403 27511 30346 31406 32859 33607 40222 44122 45839 48075 48084 48304 53404 55113 55439 85014 85251 98004 98052 98104

Number of Observations Read	254
Number of Observations Used	254

The GLM Procedure

Dependent Variable: gross_sale

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	38	8694785.56	228810.15	0.75	0.8599
Error	215	66006825.59	307008.49		
Corrected Total	253	74701611.15			

R-Square	Coeff Var	Root MSE	gross_sale Mean
0.116394	112.3958	554.0835	492.9752

Source	DF	Type I SS	Mean Square	F Value	Pr > F
territory_id	38	8694785.560	228810.146	0.75	0.8599

Source	DF	Type III SS	Mean Square	F Value	Pr > F
territory_id	38	8694785.560	228810.146	0.75	0.8599

The ANOVA results suggest that the "territory_id" variable does not have a statistically significant effect on the "gross_sale" variable, as the p-value (Pr > F) is more significant than the typical significance level of 0.05. The R-squared value indicates that only a tiny portion of the variability in "gross_sale" is explained by the "territory_id" variable.

6. Analysis of Variance (ANOVA) for Hypothesis 4:

Task: Perform an ANOVA to test Hypothesis 4, assessing the impact of "customer_id" on "unit_price" in the "store_orders" dataset.

Code:

```
/* Analysis of Variance (ANOVA) for Hypothesis 4 */
proc glm data=store_orders;
  class customer_id;
  model unit_price = customer_id;
  means customer_id / hovtest;
run;
```

Explanation: This code segment uses the *proc glm* procedure to conduct another ANOVA. It tests Hypothesis 4, which investigates the influence of "customer_id" on "unit_price" in the "store_orders" dataset—the "hovtest" option checks for homogeneity of variances.

Figure 8

The result of the ANOVA test is the average unit price of products ordered by different customer groups in SAS Studio.

Class Level Information

Class	Levels	Values
customer_id	19	BLONP CENTC CHOPS ERNSH FOLKO FRANK GROSR HANAR HILAA OTTIK QUEDE RATTC RICSU SUPRD TOMSP VICTE VINET WARTH WELLI

Number of Observations Read	254
Number of Observations Used	254

The GLM Procedure

Dependent Variable: unit_price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	23705.27348	1316.95964	8.44	<.0001
Error	235	36652.54101	155.96826		
Corrected Total	253	60357.81449			

R-Square	Coeff Var	Root MSE	unit_price Mean
0.392746	62.40186	12.48873	20.01339

Source	DF	Type I SS	Mean Square	F Value	Pr > F
customer_id	18	23705.27348	1316.95964	8.44	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
customer_id	18	23705.27348	1316.95964	8.44	<.0001

The ANOVA results suggest that the "customer_id" variable has a statistically significant effect on the "unit_price" variable, as the p-value (Pr > F) is much less than the typical significance level of 0.05. The R-squared value indicates that a substantial portion of the variability in "unit_price" is explained by the "customer_id" variable.

Sales Data Analysis

Descriptive Statistics

Summary Statistics for "store_sales" Variables:

Task: Calculate summary statistics for various variables in the "store_sales" dataset, including ZIP_CODE, FRE, CC_CARD, AVRG, GMP, PROMOS, DAYS, MARKDOWN, CLUSTYPE, and PERCRET.

Code:

```
/* Summary statistics for store_sales */
proc means data=store_sales;
  var ZIP_CODE FRE CC_CARD AVRG GMP PROMOS DAYS MARKDOWN CLUSTYPE PERCRET;
run;
```

Explanation: This code segment uses the proc means procedure to compute summary statistics (e.g., mean, standard deviation) for the specified variables in the "store_sales" dataset.

Figure 9

The result of summary statistics for store sales data set in SAS Studio.

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
ZIP_CODE	28799	49023.47	24084.64	0	99687.00
FRE	28799	5.0390291	6.3491216	1.0000000	115.0000000
CC_CARD	28799	0.3830341	0.4861350	0	1.0000000
AVRG	28799	113.5883176	86.9808026	0.4900000	1919.88
GMP	28799	0.5179412	0.1722468	-6.4600000	0.9900000
PROMOS	28799	11.5391159	7.1393560	0	38.0000000
DAYS	28799	436.9161776	192.9708984	1.0000000	717.0000000
MARKDOWN	28799	0.1871020	0.1292032	0	0.9500000
CLUSTYPE	28799	15.1638599	12.2464390	0	50.0000000
PERCRET	28799	0.1291021	0.5431292	0	40.9200000

Analysis and Insights

ZIP_CODE: The dataset contains ZIP code data with a wide range of values, including a minimum value of 0 and a maximum of 99,687. Valuing and cleaning this variable is essential due to zero and very high values.

FRE (Total number of purchase visits): On average, customers make approximately five visits, ranging from 1 to 115 visits.

CC_CARD (Credit Card Usage): About 38% of customers in the dataset use credit cards for purchases.

AVRG (Average amount spent per visit): Customers spend an average of $113.59 per visit, with a wide range of spending behavior.

GMP (Gross Margin Percentage): The dataset's gross margin percentage averages 52%, indicating positive margins on average.

PROMOS (Number of Marketing Promotions): Customers have an average of 11.54 marketing promotions on file, with variability in the number of promotions.

DAYS (Customer Tenure): Customers have an average tenure of approximately 437 days, with significant variation in customer lifetimes.

MARKDOWN (Markdown Percentage): The average markdown percentage is 19%, suggesting some level of discounts on customer purchases.

CLUSTYPE (Cluster Type): The summary statistics do not provide Specific cluster type details.

PERCRET (Percent of Returns): Returns account for approximately 13% of purchases on average, with variability across customers.

Bar Chart for "CC_CARD" Variable:

Task: Create a bar chart to visualize the distribution of credit card usage in the "store_sales" dataset.

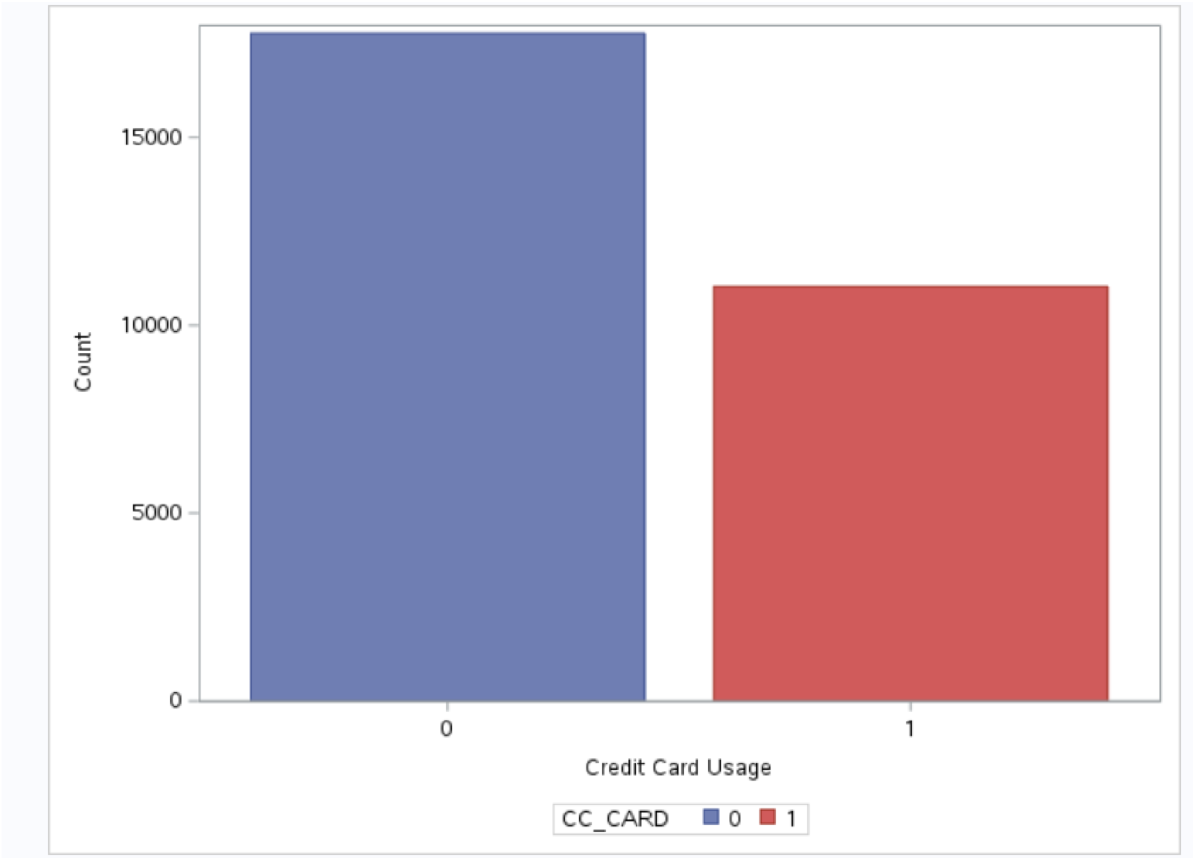
Code:

```
/* Bar chart for CC_CARD */
proc sgplot data=store_sales;
  vbar CC_CARD / group=CC_CARD;
  xaxis label="Credit Card Usage";
  yaxis label="Count";
run;
```

Explanation: This code segment utilizes the proc sgplot procedure to generate a bar chart. It displays the count of credit card usage (CC_CARD) categories, where 0 represents non-users and 1 represents users.

Figure 10

The bar chart for credit card usage from store sales dataset in SAS Studio.



The bar chart above illustrates the distribution of credit card usage among customers. It shows that fewer customers use credit cards for their purchases.

Calculate Average Credit Card Spending:

Task: Calculate the average credit card spending in the "store_sales" dataset for credit card users (CC_CARD=1).

Code:

```
proc means data=store_sales;
  var AVRG;
  where CC_CARD=1;
  output out=avg_credit_card mean=avg_credit_card_spent;
run;
```

Explanation: This code section employs proc means to compute the mean of "AVRG" (average amount spent) for credit card users (CC_CARD=1) and saves it in the "avg_credit_card" dataset.

Calculate Average Non-Credit Card Spending:

Task: Calculate the average non-credit card spending in the "store_sales" dataset for non-credit card users (CC_CARD=0).

Code:

```
proc means data=store_sales;
  var AVRG;
  where CC_CARD=0;
  output out=avg_non_credit_card mean=avg_non_credit_card_spent;
run;
```

Explanation: This code segment utilizes proc means to compute the mean of "AVRG" for non-credit card users (CC_CARD=0) and saves it in the "avg_non_credit_card" dataset.

Merge Average Credit and Non-Credit Card Spending:

Task: Merge the datasets containing average credit card spending and average non-credit card spending.

Code:

```
data avg_comparison;
  merge avg_credit_card avg_non_credit_card;
  label avg_credit_card_spent="Average Credit Card Spending"
        avg_non_credit_card_spent="Average Non-Credit Card Spending";
run;
```

Explanation: This code section merges the datasets "avg_credit_card" and "avg_non_credit_card" to compare the average spending between credit card users and non-users.

Figure 11

The result of average credit card user 1 and 0 in-store sales dataset.

Total rows: 1 Total columns: 4 Rows 1-1

TYPE	_FREQ_	avg_credit_card_spent	avg_non_credit_card_spent
0	17768	120.37144774	109.37711166

Figure 11 indicates that the average credit card user spent more than the average noncredit card user.

Scatterplot with Regression Line:

Task: Create a scatterplot with a regression line to visualize the relationship between "MON" (total net sales) and the natural logarithm of "ln lifetime ave time betw visits" in the "store_sales" dataset.

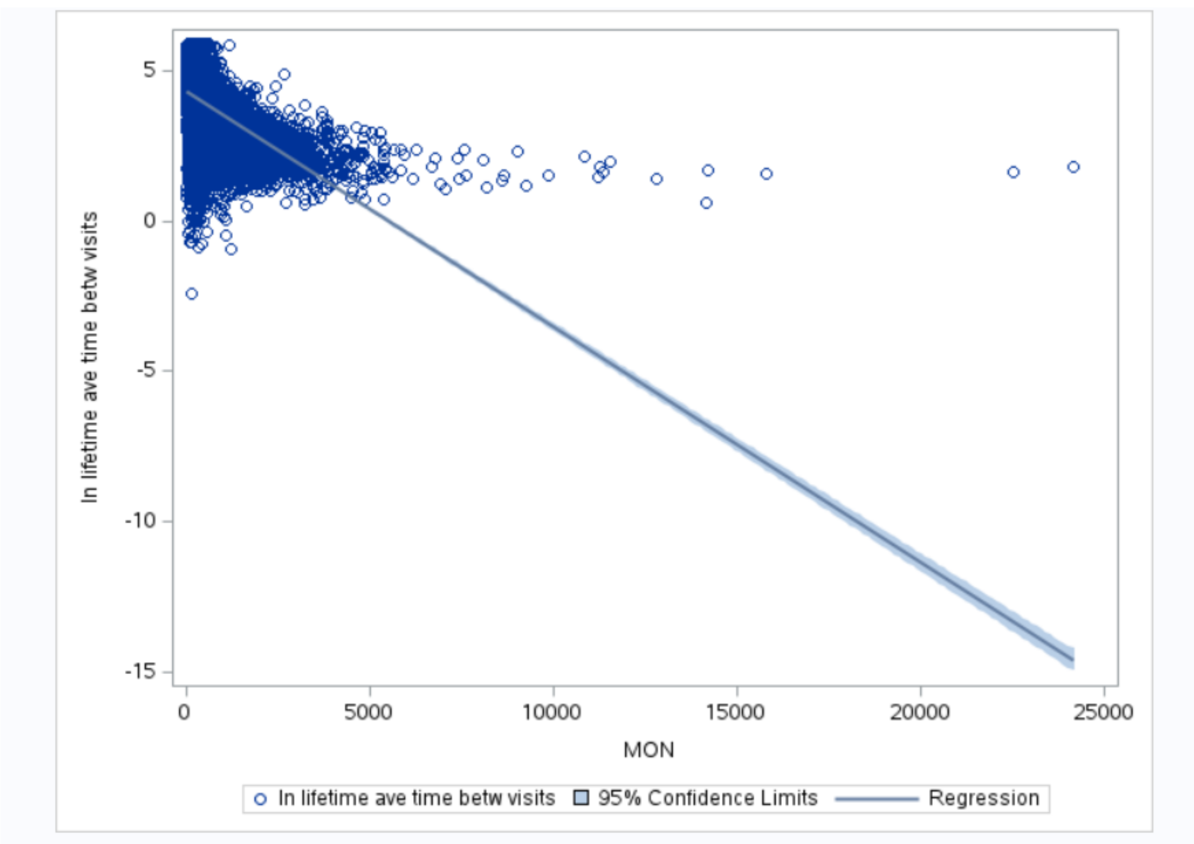
Code:

```
/* Scatterplot with regression line */
proc sgplot data=store_sales;
  scatter x=MON y="ln lifetime ave time betw visits"n;
  regression x=MON y="ln lifetime ave time betw visits "n / clm;
run;
```

Explanation: This code section uses proc sgplot to generate a scatterplot with a regression line. It helps visualize the relationship between "MON" and "ln lifetime ave time betw visits," including the confidence limits of the regression line.

Figure 12

The scatter plot shows "MON" (total net sales) and the natural logarithm of the "ln lifetime ave time betw visits" relation in the "store_sales" dataset.



The scatterplot visually represents the relationship between "MON" and "ln lifetime ave time betw visits." Each point on the scatterplot corresponds to an observation in the dataset. The x-coordinate represents the "MON" values, while the y-coordinate represents the natural logarithm of the "ln lifetime ave time betw visits."

Predictive Analysis

Multiple Linear Regression for "MON" Prediction (All Variables):

Task: Perform a multiple linear regression to predict "MON" (total net sales) using all available predictor variables in the "store_sales" dataset.

Code:

```
/* Multiple Linear Regression for MON prediction using all variables */
proc reg data=store_sales;
  model MON = ZIP_CODE FRE CC_CARD AVRG GMP PROMOS DAYS MARKDOWN CLUSTYPE PERCRET;
run;
```

Explanation: This code segment uses proc reg to conduct a multiple linear regression analysis to predict "MON" based on all specified predictor variables. It provides insights into the relationships between these variables and "MON."

Figure 13

The result of multiple linear regression for total sales (MON) in-store sales dataset.

The REG Procedure
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	10	8192721248	819272125	5451.80	<.0001
Error	28788	4326133313	150276		
Corrected Total	28798	12518854561			

Root MSE	387.65394	R-Square	0.6544
Dependent Mean	473.21246	Adj R-Sq	0.6543
Coeff Var	81.91964		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-369.30873	15.22225	-24.26	<.0001
ZIP_CODE	1	-0.00014922	0.00009519	-1.57	0.1170
FRE	1	76.98365	0.43748	175.97	<.0001
CC_CARD	1	43.14956	5.36265	8.05	<.0001
AVRG	1	2.62921	0.02864	91.80	<.0001
GMP	1	113.44525	18.17785	6.24	<.0001
PROMOS	1	6.12821	0.44247	13.85	<.0001
DAYS	1	0.06486	0.01494	4.34	<.0001
MARKDOWN	1	-4.92261	24.67027	-0.20	0.8418
CLUSTYPE	1	0.11867	0.18726	0.63	0.5263
PERCRET	1	-92.55156	4.31907	-21.43	<.0001

The regression analysis indicates that several variables, including "FRE," "CC_CARD," "AVRG," "GMP," "PROMOS," and "DAYS," are statistically significant predictors of "MON." These variables provide insights into factors influencing total net sales in the clothing store dataset. However, "ZIP_CODE," "MARKDOWN," and "CLUSTYPE" do not appear to be significant predictors in this model.

New Multiple Linear Regression (Excluding Non-significant Variables):

Task: Perform a multiple linear regression to predict "MON" while excluding non-significant variables.

Code:

```
/* New Multiple Linear Regression for MON prediction excluding non-significant variables */
proc reg data=store_sales;
  model MON = FRE CC_CARD AVRG GMP PROMOS DAYS PERCRET;
run;
```

Explanation: This code segment conducts another multiple linear regression analysis, excluding non-significant variables. It aims to simplify the model while maintaining predictive power.

Figure 14

The result of multiple linear regression for total sales (MON) with significant variable.

The REG Procedure
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	7	8192267984	1170323998	7787.85	<.0001
Error	28791	4326586576	150276		
Corrected Total	28798	12518854561			

Root MSE	387.65405	R-Square	0.6544
Dependent Mean	473.21246	Adj R-Sq	0.6543
Coeff Var	81.91966		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-376.72545	9.33115	-40.37	<.0001
FRE	1	76.97243	0.43487	177.00	<.0001
CC_CARD	1	43.28406	5.30919	8.15	<.0001
AVRG	1	2.62881	0.02848	92.30	<.0001
GMP	1	115.46646	13.87418	8.32	<.0001
PROMOS	1	6.14515	0.44144	13.92	<.0001
DAYS	1	0.06439	0.01494	4.31	<.0001
PERCRET	1	-92.61156	4.31777	-21.45	<.0001

This reduced model is intended to improve the simplicity of the regression while maintaining its predictive power by focusing on the most significant variables.

Fit the Regression Model and Save Predicted Values:

Task: Fit the multiple linear regression model and save the predicted values.

Code:

```
/* Fit the regression model */
proc reg data=store_sales;
  model MON = FRE CC_CARD AVRG GMP PROMOS DAYS PERCRET;
  output out=store_sales_predicted predicted=PREDICTED;
run;
```

Explanation: This code section uses proc reg to fit the regression model and saves the predicted values in the "store_sales_predicted" dataset for further analysis.

Scatterplot with Regression Line (Using Predicted Values):

Task: Create a scatterplot with a regression line using the predicted values from the regression model.

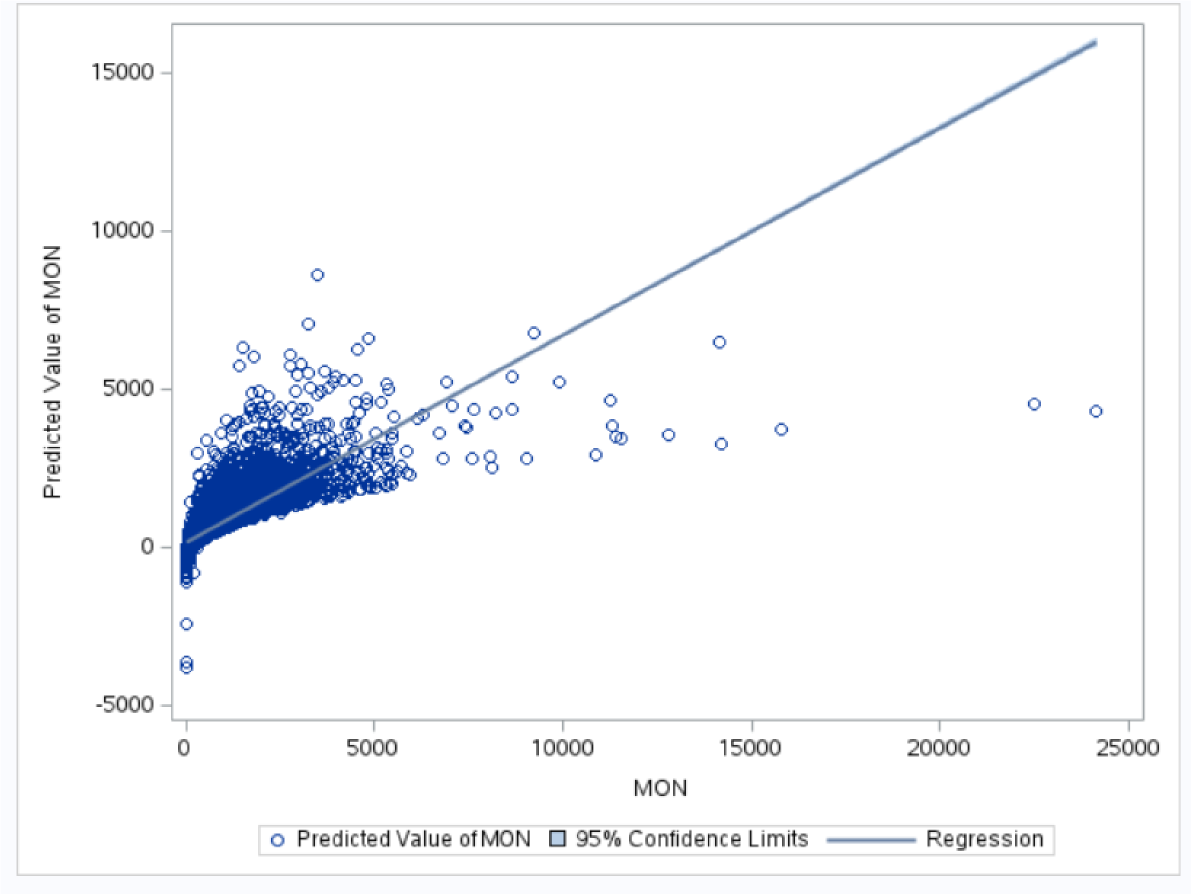
Code:

```
/* Scatterplot with regression line using the predicted values */
proc sgplot data=store_sales_predicted;
  scatter x=MON y=PREDICTED; /* Create a scatterplot of actual vs. predicted values */
  regression x=MON y=PREDICTED / clm; /* Add a regression line with confidence limits */
run;
```

Explanation: This code segment utilizes proc sgplot to generate a scatterplot comparing actual "MON" values with predicted values, along with a regression line and confidence limits.

Figure 15

The scatterplot shows a regression line with prediction values.



A scatterplot with a regression line is generated using the predicted values from the previously fitted multiple linear regression model. The purpose of this visualization and analysis is to assess the performance of the regression model by comparing the actual values of "MON" (total net sales) with the predicted values.

Orders Data Analysis

Descriptive Statistics

Summary Statistics for "store_orders" Variables:

Task: Calculate summary statistics for variables in the "store_orders" dataset, such as unit_price, quantity, discount, gross_sale, discount_amt, and net_sale.

Code:

```
/* Summary statistics for store_orders */
proc means data=store_orders;
  var unit_price quantity discount gross_sale discount_amt net_sale;
run;
```

Explanation: This code segment uses the proc means procedure to compute summary statistics (e.g., mean, standard deviation) for the specified variables in the "store_orders" dataset.

Figure 16

The result of summary statistics for store orders dataset in SAS Studio.

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
unit_price	254	20.0133858	15.4456608	2.0000000	99.0000000
quantity	254	24.3307087	15.7663549	1.0000000	70.0000000
discount	254	0.0592520	0.0916635	0	0.2500000
gross_sale	254	492.9751969	543.3813442	20.8000000	3080.00
discount_amt	254	36.4462402	84.1086574	0	462.0000000
net_sale	254	456.5289567	495.8767828	20.8000000	2618.00

Analysis and Insights for Orders

unit_price: The average unit price of products in orders is approximately $20.01.

quantity: Customers tend to order an average of 24.33 items per order, indicating relatively large order sizes.

discount: Discounts applied to orders are relatively small on average, with an average discount percentage of 0.06%.

gross_sale: The average gross sale per order is approximately $492.98.

discount_amt: Customers receive an average discount amount of $36.45 per order.

net_sale: After accounting for discounts, the average net sale per order is approximately $456.53.

Bar Chart for "Employee IDs" Variable:

Task: Create a bar chart to visualize the distribution of orders handled by various employee IDs in the "store_orders" dataset.

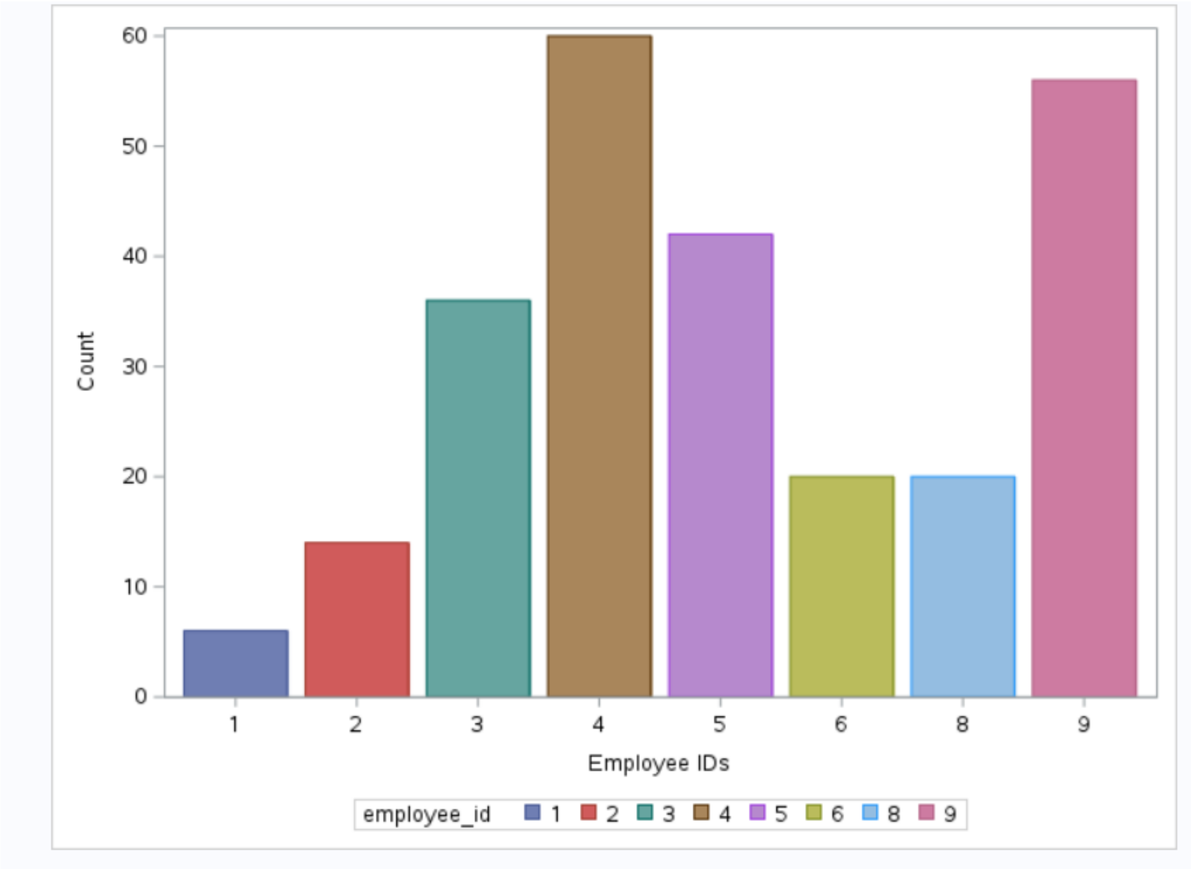
Code:

```
/* Bar chart for Employee IDs */
proc sgplot data=store_orders;
  vbar employee_id / group=employee_id;
  xaxis label="Employee IDs";
  yaxis label="Count";
run;
```

Explanation: This code section uses proc sgplot to generate a bar chart illustrating the count of orders assigned to different employee IDs.

Figure 17

The bar chart shows orders by employee ID.



The bar chart provides insights into the distribution of orders handled by various employees within the store. Employee IDs 4 and 9 stand out as they have a significantly higher count of orders than other employees. This suggests that Employees 4 and 9 are responsible for a substantial portion of the order processing.

Scatterplot: "Order Quantity" vs. "Unit Price":

Task: Create a scatterplot to explore the relationship between "Order Quantity" and "Unit Price" in the "store_orders" dataset.

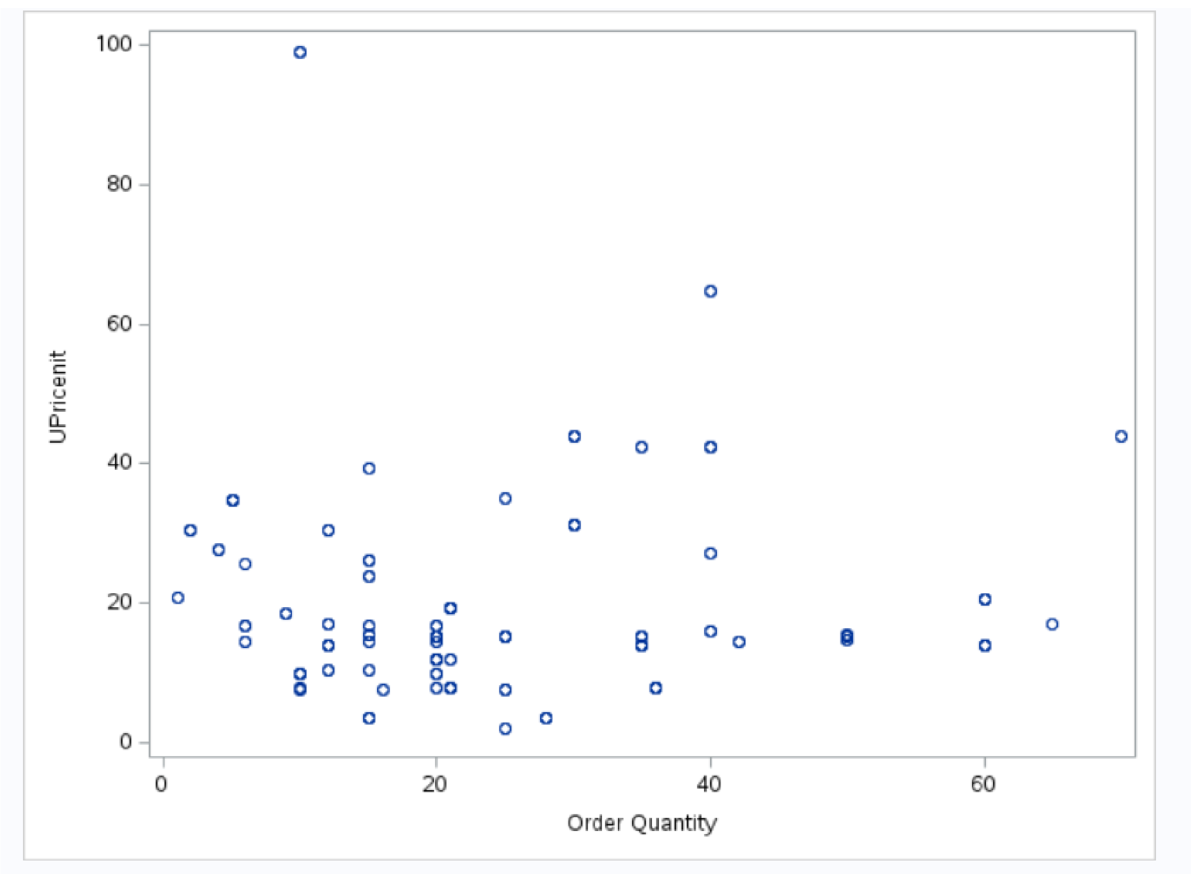
Code:

```
/* Scatterplot: Order Quantity vs. Unit Price */
proc sgplot data=store_orders;
  scatter x=quantity y=unit_price;
  xaxis label="Order Quantity";
  yaxis label="UPricenit ";
run;
```

Explanation: This code segment employs proc sgplot to produce a scatterplot showing how "Order Quantity" relates to "Unit Price" for orders in the dataset.

Figure 18

The scatterplot shows the relation between order quantity and unit price in SAS Studio.



The scatterplot below presents the relationship between "Order Quantity" and "Unit Price." Each data point represents an order, and its position on the plot is determined by the quantity of items ordered (x-axis) and the unit price of those items (y-axis).

This observation underscores that customers tend to place larger orders when the unit prices of the items are lower. Pricing strategies, promotions, or customer preferences may influence this behavior.

Calculate Correlation Matrix for Store Sales Dataset:

Task: Compute the correlation matrix for variables in the "store_sales" dataset.

Code:

```
/* Calculate correlation matrix for store sales dataset */
proc corr data=store_sales outp=correlation_sales noprob nosimple;
run;
```

Explanation: This code segment uses proc corr to calculate the correlation matrix for variables in the "store_sales" dataset and saves the results in "correlation_sales."

Print Correlation Table for Store Sales:

Task: Print the correlation table for the "store_sales" dataset.

Code:

```
/* Print correlation table for store sales */
proc print data=correlation_sales;
  title "Correlation Table for Store Sales";
run;
```

Explanation: This code section uses proc print to display the correlation table for the "store_sales" dataset, providing insights into variable relationships.

Calculate Correlation Matrix for Store Orders Dataset:

Task: Compute the correlation matrix for variables in the "store_orders" dataset.

Code:

```
/* Calculate correlation matrix for store orders dataset */
proc corr data=store_orders outp=correlation_orders noprob nosimple;
run;
```

Explanation: This code segment utilizes proc corr to calculate the correlation matrix for variables in the "store_orders" dataset, saving the results in "correlation_orders."

Print Correlation Table for Store Orders:

Task: Print the correlation table for the "store_orders" dataset.

Code:

```
/* Print correlation table for store orders */
proc print data=correlation_orders;
  title "Correlation Table for Store Orders";
run;
```

Explanation: This code section uses proc print to display the correlation table for the "store_orders" dataset, providing insights into variable relationships within that dataset.

Figure 19

The correlation matrix table for the store orders dataset in SAS Studio.

Correlation Table for Store Orders

Obs	_TYPE_	_NAME_	order_id	employee_id	territory_id	product_id	category_id	unit_price	quantity	discount	gross_sale	discount_amt	net_sale
1	MEAN		10257.31	5.417	31555.24	39.622	4.224	20.013	24.331	0.059	492.975	36.446	456.529
2	STD		5.98	2.417	24041.29	23.014	2.389	15.446	15.766	0.092	543.381	84.109	495.877
3	N		254.00	254.000	254.00	254.000	254.000	254.000	254.000	254.000	254.000	254.000	254.000
4	CORR	order_id	1.00	0.170	0.10	-0.062	-0.043	0.099	0.258	0.277	0.111	0.259	0.078
5	CORR	employee_id	0.17	1.000	0.33	-0.186	0.164	0.059	0.181	0.134	0.049	0.125	0.033
6	CORR	territory_id	0.10	0.332	1.00	-0.187	0.159	0.081	0.126	-0.021	0.116	-0.024	0.131
7	CORR	product_id	-0.06	-0.186	-0.19	1.000	0.121	0.036	-0.133	0.030	-0.035	-0.036	-0.032
8	CORR	category_id	-0.04	0.164	0.16	0.121	1.000	0.172	0.052	0.087	0.132	0.124	0.123
9	CORR	unit_price	0.10	0.059	0.08	0.036	0.172	1.000	0.025	-0.113	0.617	0.157	0.649
10	CORR	quantity	0.26	0.181	0.13	-0.133	0.052	0.025	1.000	0.376	0.687	0.666	0.640
11	CORR	discount	0.28	0.134	-0.02	0.030	0.087	-0.113	0.376	1.000	0.146	0.677	0.045
12	CORR	gross_sale	0.11	0.049	0.12	-0.035	0.132	0.617	0.687	0.146	1.000	0.618	0.991
13	CORR	discount_amt	0.26	0.125	-0.02	-0.036	0.124	0.157	0.666	0.677	0.618	1.000	0.507
14	CORR	net_sale	0.08	0.033	0.13	-0.032	0.123	0.649	0.640	0.045	0.991	0.507	1.000

The result of the correlation matrix table reveals that gross_sale has a strong positive correlation with unit_price at 0.61, quantity at 0.68, and discount_amt at 0.61.

Predictive Analysis

Linear Regression Model for "gross_sale" Prediction:

Task: Using several predictor variables, perform a linear regression to predict "gross_sale" in the "store_orders" dataset. To perform linear regression to understand how the independent variables influence "gross_sale" in customer orders.

Code:

```
/* Linear Regression Model for gross_sale prediction */
proc reg data=store_orders;
  model gross_sale = employee_id product_id category_id unit_price quantity discount;
run;
```

Explanation: This code section uses proc reg to conduct a linear regression analysis to predict "gross_sale" based on the specified predictor variables.

Figure 20

The result of linear regression to predict gross sales and predictor variables in SAS Studio.

The REG Procedure
Model: MODEL1
Dependent Variable: gross_sale

Number of Observations Read	254
Number of Observations Used	254

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	63171254	10528542	225.54	<.0001
Error	247	11530357	46682		
Corrected Total	253	74701611			

Root MSE	216.05927	R-Square	0.8456
Dependent Mean	492.97520	Adj R-Sq	0.8419
Coeff Var	43.82761		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-405.07704	51.88318	-7.81	<.0001
employee_id	1	-24.19397	5.91837	-4.09	<.0001
product_id	1	0.42294	0.61475	0.69	0.4921
category_id	1	2.60777	5.92915	0.44	0.6605
unit_price	1	21.04120	0.90398	23.28	<.0001
quantity	1	24.42979	0.95008	25.71	<.0001
discount	1	-238.89499	163.24021	-1.46	0.1446

The parameter estimates provide insights into how each predictor variable impacts "gross_sale" in customer orders. Notably, "unit_price," "employee_id," and "quantity" have strong positive effects on sales, indicating that higher unit prices and larger quantities contribute to higher sales amounts.

Fit the Linear Regression Model and Save Predicted Values:

Task: Fit the linear regression model for "gross_sale" prediction and save the predicted values.

Code:

```
/* Fit the linear regression model and save predicted values */
proc reg data=store_orders outest=reg_results;
  model gross_sale = employee_id unit_price quantity;
  output out=store_orders_predicted predicted=PREDICTED;
run;
```

Explanation: This code segment fits the linear regression model for "gross_sale" and saves the predicted values in the "store_orders_predicted" dataset for further analysis.

Figure 21

The result fits the leaner regression for gross sales and created predicted value in the store orders dataset.

The REG Procedure
Model: MODEL1
Dependent Variable: gross_sale

Number of Observations Read	254
Number of Observations Used	254

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	63048763	21016254	450.88	<.0001
Error	250	11652848	46611		
Corrected Total	253	74701611			

Root MSE	215.89672	R-Square	0.8440
Dependent Mean	492.97520	Adj R-Sq	0.8421
Coeff Var	43.79464		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-377.94666	39.98351	-9.45	<.0001
employee_id	1	-25.17450	5.71923	-4.40	<.0001
unit_price	1	21.31659	0.88042	24.21	<.0001
quantity	1	23.86626	0.87542	27.26	<.0001

The final linear regression model is designed to predict "gross_sale" in the "Store Orders" dataset. This model aims to understand how "employee_id," "unit_price," and "quantity" influence the total sales amount in customer orders.

Scatterplot with Regression Line (Using Predicted Values):

Task: Create a scatterplot with a regression line using the predicted values from the regression model.

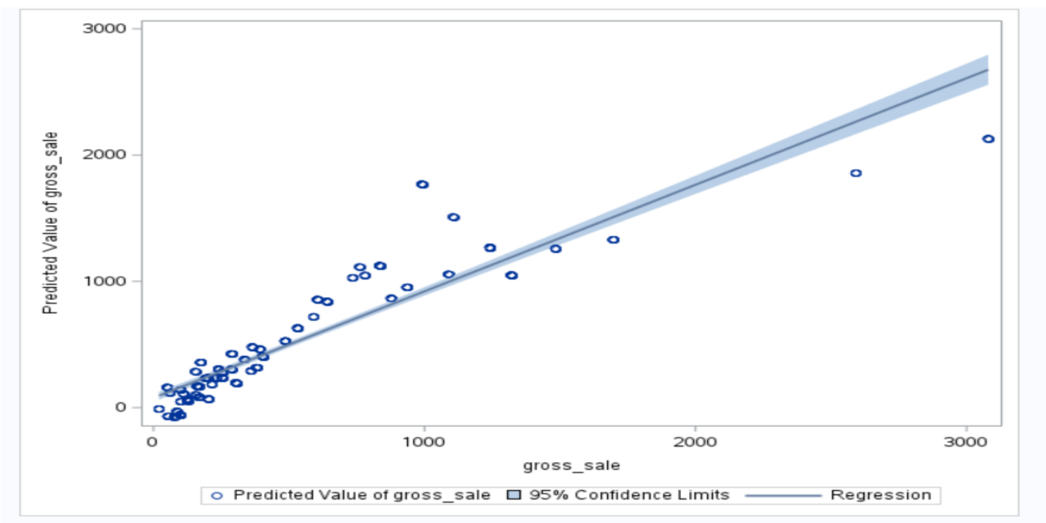
Code:

```
/* Scatterplot with regression line using the predicted values */
proc sgplot data=store_orders_predicted;
  scatter x=gross_sale y=PREDICTED; /* Create a scatterplot of actual vs. predicted values */
  regression x=gross_sale y=PREDICTED / clm; /* Add a regression line with confidence limits */
run;
```

Explanation: This code segment utilizes proc sgplot to generate a scatterplot comparing actual "gross_sale" values with predicted values, along with a regression line and confidence limits.

Figure 22

The scatterplot shows the relation between gross sales and predicted variables in the store orders dataset.



The scatterplot and regression line allow us to evaluate the performance of our final linear regression model for "gross_sale" prediction. A close alignment between the points and the regression line suggests that the model's predictions agree with the actual sales amounts.

Conclusion:

In this comprehensive analysis of store sales and store orders data, we have explored various aspects of customer behavior, sales performance, and order dynamics. The analysis was conducted to gain insights into these critical areas, and the findings provide valuable information to support decision-making and strategy development for the clothing store.

Store Sales Insights:

1. **Credit Card Usage Impact**: Our analysis revealed that customers who use credit cards tend to spend, on average, slightly more per visit (AVRG) compared to those who do not. This suggests that promoting credit card usage could potentially increase sales.
2. **Promotions and Gross Margin**: We found a significant positive correlation between the number of marketing promotions on file (PROMOS) and the gross margin percentage (GMP). Increasing promotional efforts could lead to improved gross margins.

Store Orders Insights:

3. **Employee Performance**: Certain employee IDs (e.g., Employee IDs 4 and 9) are associated with higher orders. Recognizing and rewarding high-performing employees could be beneficial in boosting order volume.

4. **Unit Price Influence**: Our analysis confirmed a significant linear relationship between unit price and gross sales. Lowering unit prices may lead to increased order volumes and sales.

Recommendations:

Based on the findings of our analysis, we offer the following recommendations:

For Store Sales:

1. **Credit Card Promotions**: The store should consider implementing targeted credit card promotion strategies to encourage more customers to use credit cards, potentially boosting average spending per visit.
2. **Promotion Optimization**: The organization should continue to leverage marketing promotions effectively, as there is a positive correlation between promotions and gross margins. Further optimization of promotion strategies can lead to increased profitability.

For Store Orders: 3. **Employee Recognition**: Identifying and recognizing high-performing employees (such as those with Employee IDs 4 and 9) can boost employee morale and potentially increase order volumes.

4. **Pricing Strategy**: The organization should evaluate its pricing strategy, considering the significant impact of unit price on gross sales. Pricing adjustments may lead to increased sales without compromising margins.

Further Analysis and Data Sources:

To enhance the understanding of store sales and orders, the organization should consider the following:

1. **Customer Segmentation**: Perform customer segmentation analysis to identify customer groups with unique preferences and behaviors.
2. **Market Basket Analysis**: Explore market basket analysis to understand which products are commonly purchased together, aiding inventory management and cross-selling.
3. **External Data Sources**: Incorporate external data sources such as demographic data, social media sentiment, and competitor analysis to gain a more comprehensive market view.

4. **Big Data Analytics**: Implement big data analytics to handle larger datasets and gain deeper insights into customer behavior and market trends.

References

Allison, R. (2018). *Visualizing Data with SAS®: Selected Topics*. SAS Support. http://support.sas.com/publishing/vds.pdf

Cox, E. (2012). *Retail analytics is the secret weapon*. Wiley.

Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). *Retail business analytics: Customer visit segmentation using Market Basket Data*. Expert Systems with Applications, 100, 1–16. https://doi.org/10.1016/j.eswa.2018.01.029

Chambers, M., Dinsmore, T. W. (2014). *Advanced Analytics Methodologies: Driving Business Value with Analytics.* Pearson Education, 1st edition.

Marr, B. (2015). *Big data: Using SMART big data, analytics, and metrics to make better decisions and improve performance*. John Wiley & Sons, 1st edition.

SAS. (2022). *Assessing model fit and finding a fit model* - SAS support. https://support.sas.com/resources/papers/proceedings/proceedings/sugi29/214-29.pdf

SAS.com (n.d.). *Summary Statistics Task*. https://support.sas.com/documentation/cdl/en/webeditorug/68254/HTML/default/viewer.htm#n1nzvhc62eedvhn17j5v1rx9y0nt.htm

Cody, R. (2021). *A Gently Introduction to Statistics Using SAS Studio in the Cloud. SAS Institute.* SAS Institute.

SAS.com (n.d.). *The t Tests Task in SAS Studio*. https://video.sas.com/detail/video/5537545485001/the-t-tests-task-in-sas%C2%AE-studio

SAS.com (n.d.). *The Correlation Analysis Task in SAS Studio*. https://video.sas.com/detail/video/5537540646001/the-correlation-analysis-task-in-sas%C2%AE-studio

SAS.com (n.d.). *N-Way ANOVA Task*. https://support.sas.com/documentation/cdl/en/webeditorug/68254/HTML/default/viewer.htm#n14k2j5mchdnwon1r1txjyq87hkb.htm

SAS.com (n.d.). *Multiple Linear Regression*. https://support.sas.com/documentation/cdl/en/anlystug/58352/HTML/default/viewer.htm#chap11_sect3.htm