The problem involves using Linear Regression analysis since the variables involved are continuous. Also since multiple variables are involved, it requires **Multivariate Linear Regression** to forecast sales.

```
FILENAME REFFILE '/home/gowthamharshabh0/Project 04_Retail Analysis_Dataset.xlsx';

/* The following code imports dataset into SAS*/
PROC IMPORT DATAFILE=REFFILE
        DBMS=XLSX
        OUT=SASPROJ.Retail_Analysis;
        GETNAMES=YES;
RUN;

PROC CONTENTS DATA=SASPROJ.Retail_Analysis; RUN;
```

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
5	Discount	Num	8	BEST.		Discount
1	Order_ID	Num	8	BEST.		Order_ID
2	Products	Char	8	$8.	$8.	Products
6	Profit	Num	8	NLMNY15.1		Profit
4	Quantity	Num	8	BEST.		Quantity
3	Sales	Num	8	NLMNY15.1		Sales
7	Shipping_Cost	Num	8	NLMNY15.1		Shipping_Cost

```
/* Since it is been observed that the dataset has individual price of product
  but no record measuring the total sales of each product, a new variable
  Total_Sales = Sales*Quantity is being created*/

proc sql;
create table SASProj.Retail_Analysis_Modified as
        select *, Sales*Quantity as Total_Sales from Sasproj.retail_analysis;
        quit;


/* Following code gets descriptive statics on the modified dataset */
Proc Means data=sasproj.retail_analysis_modified;
run;
```

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
Order_ID	Order_ID	30	110015.50	8.8034084	110001.00	110030.00
Sales	Sales	30	152.9666667	63.1759903	33.0000000	250.0000000
Quantity	Quantity	30	3.1666667	1.2340942	1.0000000	5.0000000
Discount	Discount	30	0.0256667	0.0154659	0.0100000	0.0500000
Profit	Profit	30	72.1063333	44.6008984	3.2500000	135.6000000
Shipping_Cost	Shipping_Cost	30	7.2106333	4.4600898	0.3250000	13.5600000
Total_Sales		30	491.1000000	265.3040351	33.0000000	1100.00

```
/* Checking whether individual variable is significant or linearly related to Total_Sales*/
proc reg data=sasproj.retail_analysis_modified;
        model Total_Sales = Quantity; /*Checking the suitability of variable quatity*/
        var Total_Sales;
        Run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Total_Sales

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	870480	870480	20.82	<.0001
Error	28	1170721	41811		
Corrected Total	29	2041201			

Root MSE	204.47849	R-Square	0.4265
Dependent Mean	491.10000	Adj R-Sq	0.4060
Coeff Var	41.63683		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	46.53585	104.33962	0.45	0.6590
Quantity	Quantity	1	140.38868	30.76807	4.56	<.0001

It can be seen that hypothesis testing shows positive correlation with 42% predictability. Implies Quantity is a significant variable.

```
proc reg data=sasproj.retail_analysis_modified;
        model Total_Sales = Discount; /*Checking the suitability of variable Discount*/
```

```
var Total_Sales;
Run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Total_Sales

Number of Observations Read	30
Number of Observations Used	30

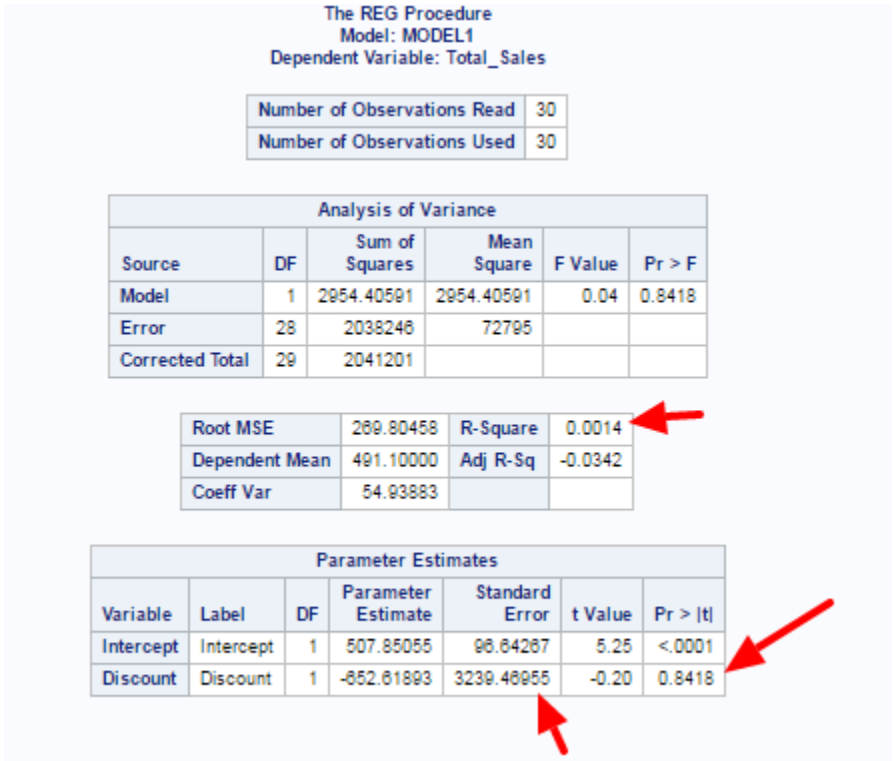
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2954.40591	2954.40591	0.04	0.8418
Error	28	2038246	72795		
Corrected Total	29	2041201			

Root MSE	269.80458	R-Square	0.0014
Dependent Mean	491.10000	Adj R-Sq	-0.0342
Coeff Var	54.93883		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	507.85055	96.64267	5.25	<.0001
Discount	Discount	1	-652.61893	3239.46955	-0.20	0.8418



The t-test data shows that there is no significant correlation with this variable to the value of Dependant variable. The p-value of 0.84 is way above 0.05. R-Square data shows no predictability too with a poor value of .0014%. Hence Discount is **not** a suitable variable for regression analysis.

```
proc reg data=sasproj.retail_analysis_modified;
        model Total_Sales = Profit; /*Checking the suitability of variable Profit*/
        var Total_Sales;
        Run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Total_Sales

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	618834	618834	12.18	0.0016
Error	28	1422366	50799		
Corrected Total	29	2041201			

Root MSE	225.38590	R-Square	0.3032
Dependent Mean	491.10000	Adj R-Sq	0.2783
Coeff Var	45.89409		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	254.93367	79.19411	3.22	0.0032
Profit	Profit	1	3.27525	0.93839	3.49	0.0016

It can be seen that hypothesis testing shows positive correlation with 30% predictability. Even though Pr value is slightly higher, the alpha value condition (<0.05) is still satisfied to negate the null hypothesis. Implies Profit is a significant variable with linear relation with Total_Sales.
Also needs to be noticed is the Coefficient of Variance which is higher at 45%.

```
/*Checking the suitability of variable Shipping_Cost.
  Marketing cost is assumed as Shipping _Cost*/
proc reg data=sasproj.retail_analysis_modified;
        model Total_Sales = Shipping_Cost;
        var Total_Sales;
        Run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Total_Sales

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	618834	618834	12.18	0.0016
Error	28	1422366	50799		
Corrected Total	29	2041201			

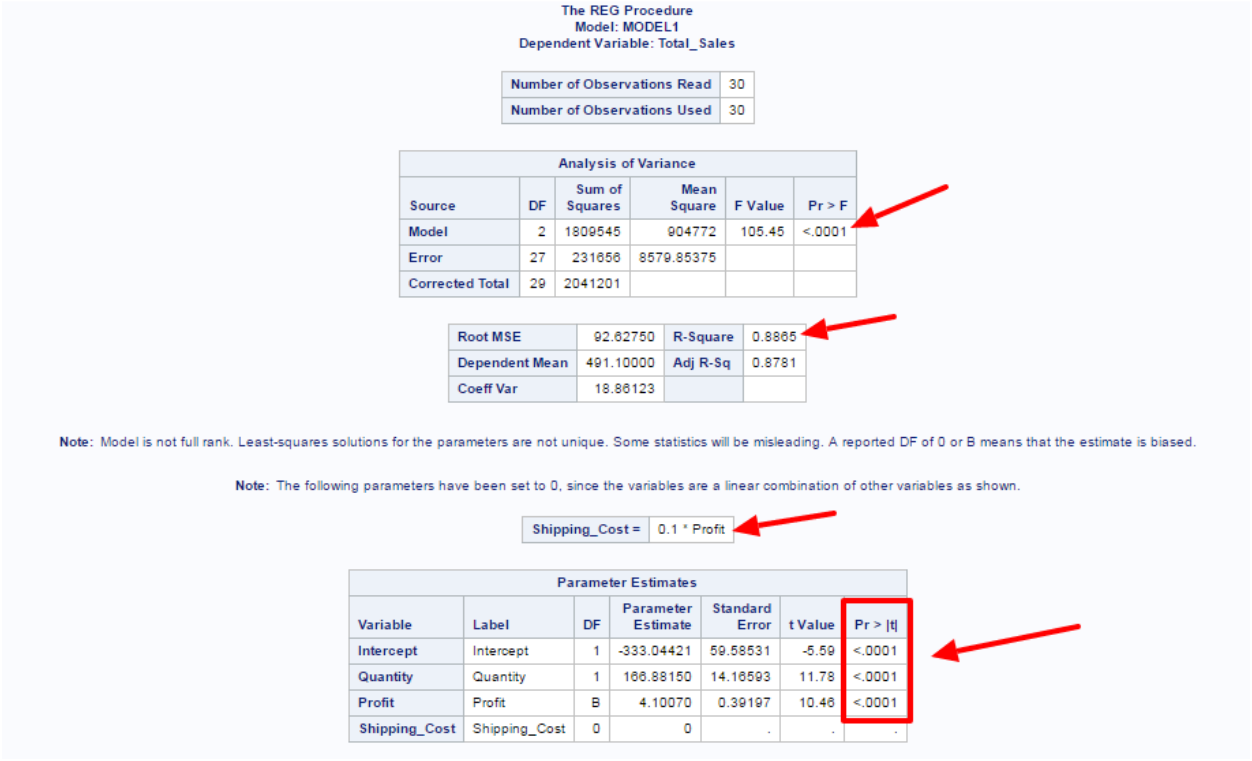
Root MSE	225.38590	R-Square	0.3032
Dependent Mean	491.10000	Adj R-Sq	0.2783
Coeff Var	45.89409		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	254.93367	79.19411	3.22	0.0032
Shipping_Cost	Shipping_Cost	1	32.75251	9.38392	3.49	0.0016

This variable shows statistical similarity to the variable Profit. Individually this variable needs to be tested along with Profit through multivariate analysis.

```
/*Now performing Multivariate Regression Analysis*/
proc reg data=sasproj.retail_analysis_modified;
        model Total_Sales = Quantity Profit Shipping_Cost;
        var Total_Sales;
        Run;
```



The REG Procedure
Model: MODEL1
Dependent Variable: Total_Sales

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1809545	904772	105.45	<.0001
Error	27	231656	8579.85375		
Corrected Total	29	2041201			

Root MSE	92.62750	R-Square	0.8865
Dependent Mean	491.10000	Adj R-Sq	0.8781
Coeff Var	18.86123		

Note: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

Shipping_Cost =	0.1 * Profit

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-333.04421	59.58531	-5.59	<.0001
Quantity	Quantity	1	166.88150	14.16593	11.78	<.0001
Profit	Profit	B	4.10070	0.39197	10.46	<.0001
Shipping_Cost	Shipping_Cost	0	0	.	.	.

The assumption for performing regression analysis is violated here as the variables tested against are not independent of each other. Shipping cost and Profit have a direct discernible relation here hence any one would suffice in predicting the variability with Total_Sales. Rest of the variables are showing good positive correlation with R^2 value boosted to over 88%.

Now performing the regression by removing the insignificant variable of Shipping_Cost.

```
proc reg data=sasproj.retail_analysis_modified;
        model Total_Sales = Quantity Profit;
        var Total_Sales;
        Run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Total_Sales

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1809545	904772	105.45	<.0001
Error	27	231656	8579.85375		
Corrected Total	29	2041201			

Root MSE	92.62750	R-Square	0.8865
Dependent Mean	491.10000	Adj R-Sq	0.8781
Coeff Var	18.86123		

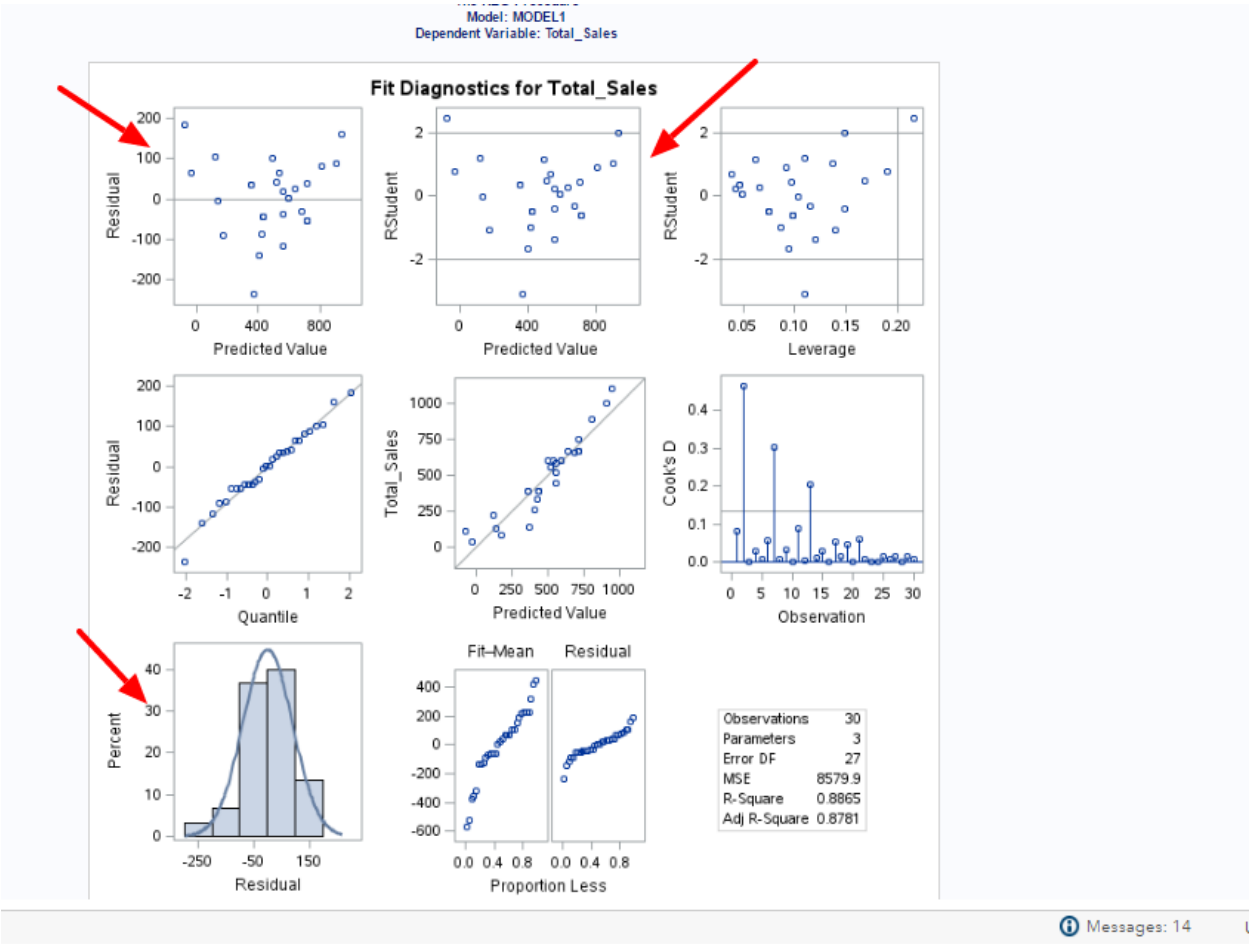
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-333.04421	59.58531	-5.59	<.0001
Quantity	Quantity	1	166.88150	14.16593	11.78	<.0001
Profit	Profit	1	4.10070	0.39197	10.46	<.0001

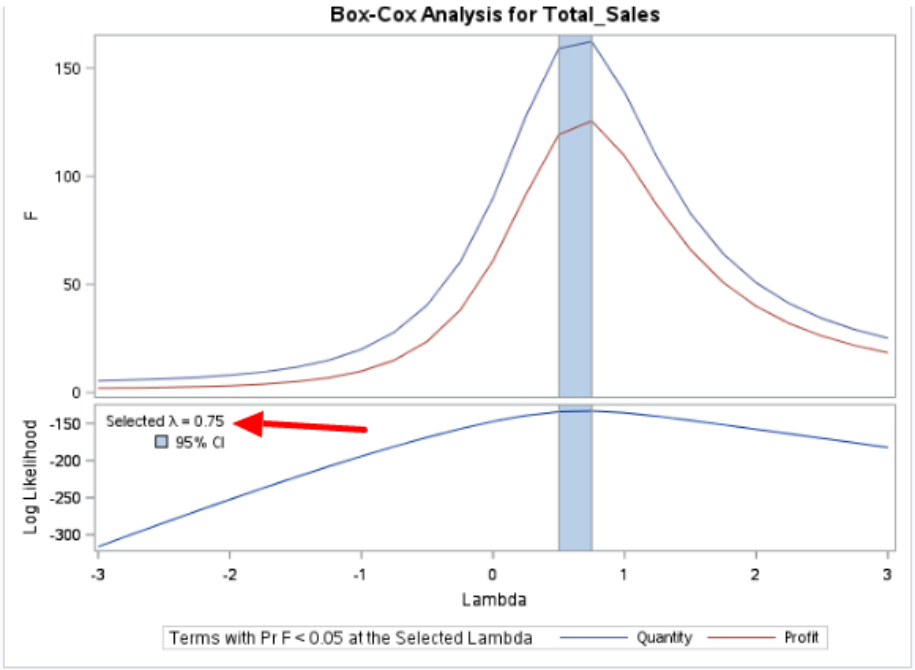
- The R^2 value still holds good at 88% while we got rid of one dependent variable. This means approximately 88% of the variation of Total_Sales is explained by the independent variables.
- Based on t-test , the Pr-values for Quantity and Profit are less than 0.05 indicating sufficient evidence for predicting the Total_Sales. They predict positive high correlation and corroboration of our hypothesis.
- The resultant equation derived from the model thus would be as follows:

Total_Sales = 166.88*Quantity + 4.1*Profit -333.04

This shows that increase in 1 quantity of product would raise the total sales by around $166 and increase in profit by $1 would have meant sales would go up by $4.1.



The histogram and quantile plots show a healthy model. However looking at the fitness check of the model, the residual plot data seems to be showing a conical trend raising doubts over violation of assumptions. Performing BoxCox test to validate our model ➔



Dependent Variable BoxCox(Total_Sales)

Number of Observations Read	30
Number of Observations Used	30

The TRANSREG Procedure Hypothesis Tests for BoxCox(Total_Sales)

Univariate ANOVA Table Based on the Usual Degrees of Freedom

Source	DF	Sum of Squares	Mean Square	F Value	Liberal p
Model	2	89620.20	44810.10	122.35	>= <.0001
Error	27	9888.45	366.24		
Corrected Total	29	99508.64			

The above statistics are not adjusted for the fact that the dependent variable was transformed and so are generally liberal.

Messages: 16

The test shows that the values are following a normal distribution not requiring any transformation (as lambda = 0.75) for Dependent variable here. Also corroborate the validity of the model.

Recommended Transformation	Equation	Lambda
Square	Y^2	1.5 to 2.5
None	Y	0.75 to 1.5
Square-root	$Y^{1/2}$	0.25 to 0.75
Natural log	Ln(Y)	-0.25 to 0.25
Inverse square-root	$1/Y^{1/2}$	-0.75 to -0.25
Reciprocal	1/Y	-1.5 to -0.75
Inverse square	$1/Y^2$	-2.5 to -1.5

("Box-Cox Method")