SAS Retail Analysis

–SAS Project

Problem Statement

Forecast the sales based on the independent variables such as Profit, Quantity, Marketing cost, and Expenses using the regression model.

The dataset is maintained for the Retail Analysis, and it has records of both independent and dependent variables.

Analysis :

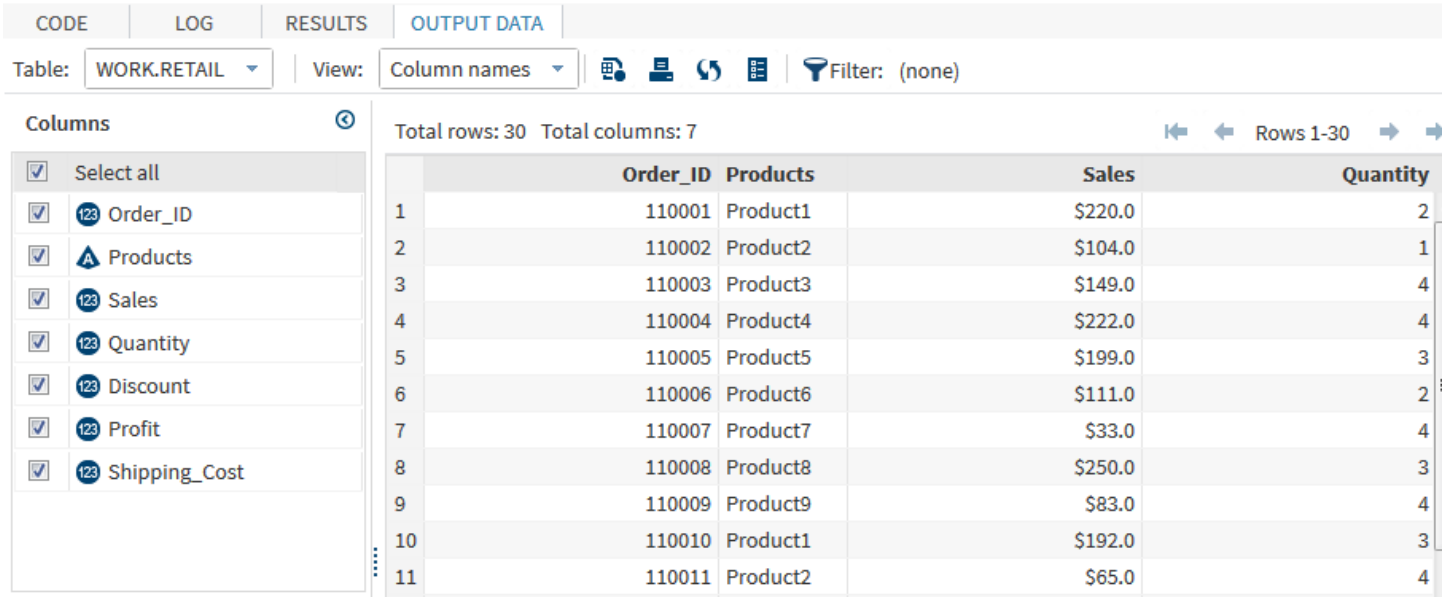
- Import the required dataset

```
 FILENAME REFFILE '/home/sr0/sasuser.v94/Project 04_Retail Analysis_Dataset.xlsx';

PROC IMPORT DATAFILE=REFFILE
        DBMS=XLSX REPLACE
        OUT=WORK.Retail(RENAME= 'Shipping Cost'n=Shipping_Cost);
        GETNAMES=YES;
RUN;
```

Verify the Data Import :

```
PROC CONTENTS DATA=WORK.Retail; RUN;
```



CODE | LOG | RESULTS | OUTPUT DATA

Table: WORK.RETAIL | View: Column names | Filter: (none)

Columns

- Select all
- Order_ID
- Products
- Sales
- Quantity
- Discount
- Profit
- Shipping_Cost

Total rows: 30 Total columns: 7 Rows 1-30

	Order_ID	Products	Sales	Quantity
1	110001	Product1	$220.0	2
2	110002	Product2	$104.0	1
3	110003	Product3	$149.0	4
4	110004	Product4	$222.0	4
5	110005	Product5	$199.0	3
6	110006	Product6	$111.0	2
7	110007	Product7	$33.0	4
8	110008	Product8	$250.0	3
9	110009	Product9	$83.0	4
10	110010	Product1	$192.0	3
11	110011	Product2	$65.0	4

.

Descriptive Statistics on the Dataset

.

To Perform the Descriptive Statistics, the Dataset is first grouped Based on the Product ID .

*Retail.sas ×

CODE | LOG | RESULTS | OUTPUT DATA

Table: WORK.RETAIL_GROUPED | View: Column names | Filter: (none)

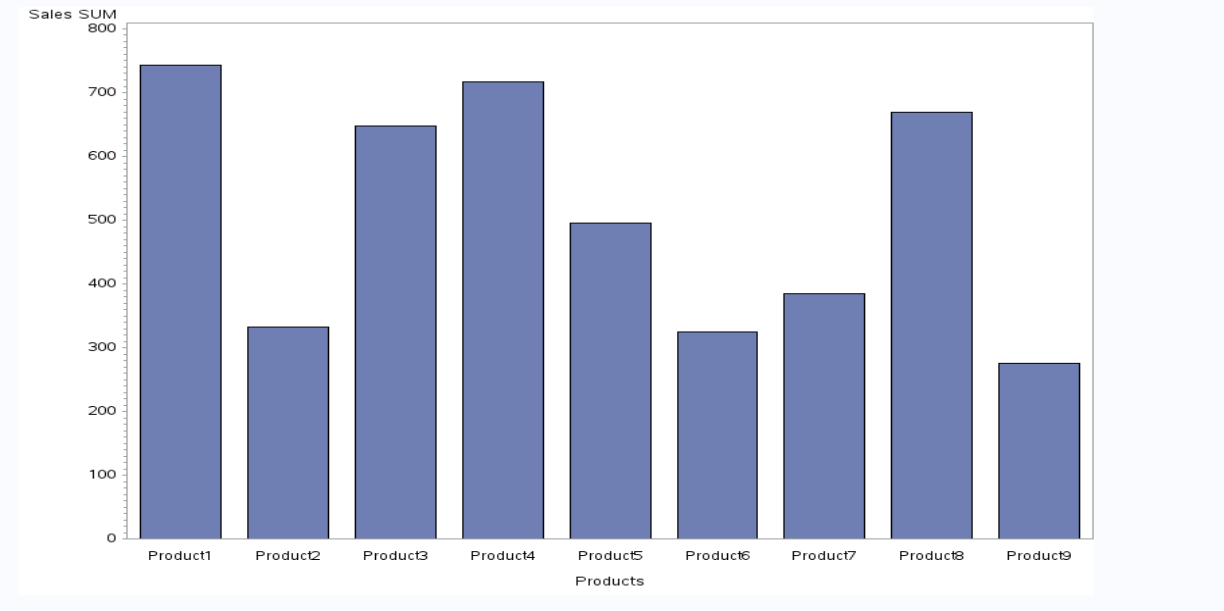
Columns

- Select all
- Products
- Sales
- Quantity
- Profit
- Avg_Discount
- Expenses

Total rows: 9 Total columns: 6 Rows 1-9

Products	Sales	Quantity	Profit	Avg_Discount
Product1	743	13	366.97	0.025
Product2	332	9	116.23	0.025
Product3	648	16	302.98	0.01
Product4	717	12	409.68	0.0325
Product5	495	10	207.07	0.03
Product6	325	8	100.48	0.0333333333
Product7	385	10	169.71	0.02
Product8	669	11	365	0.0266666667
Product9	275	6	125.07	0.035

As a first Step to describe the Dataset, we create a Chart of the Product Vs.Sales is Obtained to identify the Product that gives the Max.Sales

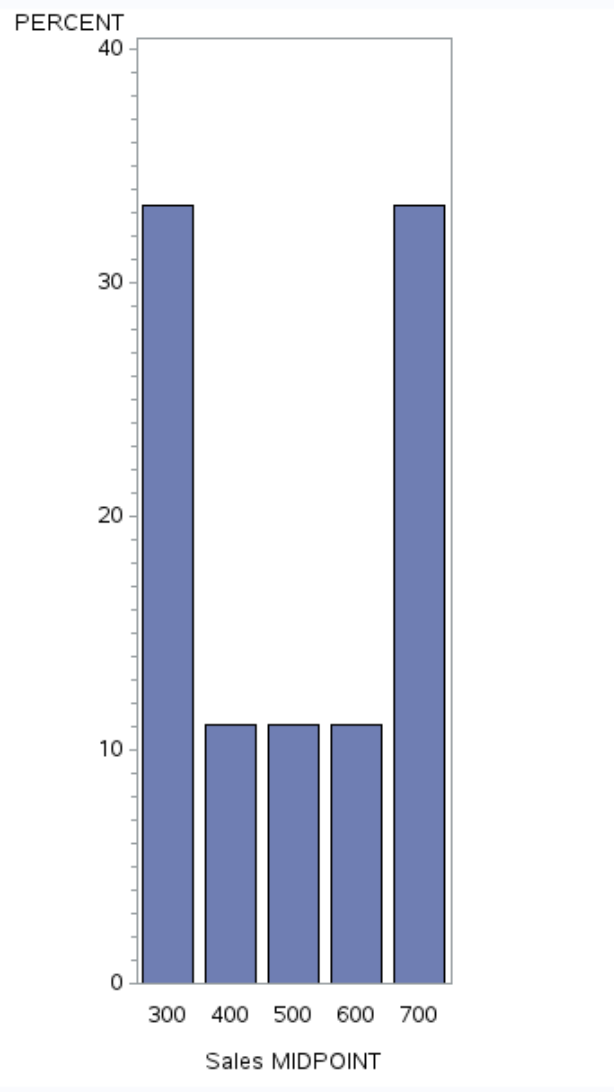


INFERENCE :

It can be seen that product 1 has the Maximum Sales

- Next, to find the Average value of Sales , a Chart is plotted to find out the % Distribution of the Sales Value

OUTPUT:



INFERENCE:

Sales Values are distributed more at the Extremes .Most of the Sales are either around 700$ or the minimum 300$.

- Next, to obtain the Summary Statistics for Each product Category , the Mean.Median and Average Values are obtained for the Sales,Profit,Quantity ,Discount Values .

OUTPUT:

Table of Contents

Products	N Obs	Variable	Label	Mean	Median	Mode
Product1	4	Sales	Sales	185.7500000	206.0000000	220.0000000
		Quantity	Quantity	3.2500000	3.0000000	3.0000000
		Discount	Discount	0.0250000	0.0200000	0.0100000
		Profit	Profit	91.7425000	114.0600000	.
		Shipping_Cost	Shipping Cost	9.1742500	11.4060000	.
Product2	4	Sales	Sales	83.0000000	84.5000000	.
		Quantity	Quantity	2.2500000	2.0000000	1.0000000
		Discount	Discount	0.0250000	0.0300000	0.0300000
		Profit	Profit	29.0575000	26.9400000	.
		Shipping_Cost	Shipping Cost	2.9057500	2.6940000	.
Product3	4	Sales	Sales	162.0000000	149.0000000	149.0000000
		Quantity	Quantity	4.0000000	4.0000000	4.0000000
		Discount	Discount	0.0100000	0.0100000	0.0100000
		Profit	Profit	75.7450000	63.0400000	63.0400000
		Shipping_Cost	Shipping Cost	7.5745000	6.3040000	6.3040000
Product4	4	Sales	Sales	179.2500000	206.0000000	.
		Quantity	Quantity	3.0000000	3.0000000	.
		Discount	Discount	0.0325000	0.0300000	0.0300000
		Profit	Profit	102.4200000	105.6600000	.
		Shipping_Cost	Shipping Cost	10.2420000	10.5660000	.
Product5	3	Sales	Sales	165.0000000	192.0000000	.
		Quantity	Quantity	3.3333333	3.0000000	.
		Discount	Discount	0.0300000	0.0200000	0.0200000
		Profit	Profit	69.0233333	89.1500000	.
		Shipping_Cost	Shipping Cost	6.9023333	8.9150000	.
Product6	3	Sales	Sales	108.3333333	111.0000000	.
		Quantity	Quantity	2.6666667	2.0000000	2.0000000
		Discount	Discount	0.0333333	0.0400000	.
		Profit	Profit	33.4933333	32.5000000	.
		Shipping_Cost	Shipping Cost	3.3493333	3.2500000	.

- Next, to identify the Product that sells the most a Descending Sort is performed on the Quantity.

OUTPUT:

Table: WORK.RETAIL_GROUPED | View: Column names | Filter: (none)

Columns

- Select all
- Products
- Sales
- Quantity
- Profit
- Avg_Discount
- Expenses

Total rows: 9 Total columns: 6 | Rows 1-9

Total rows: 9 Total columns: 6

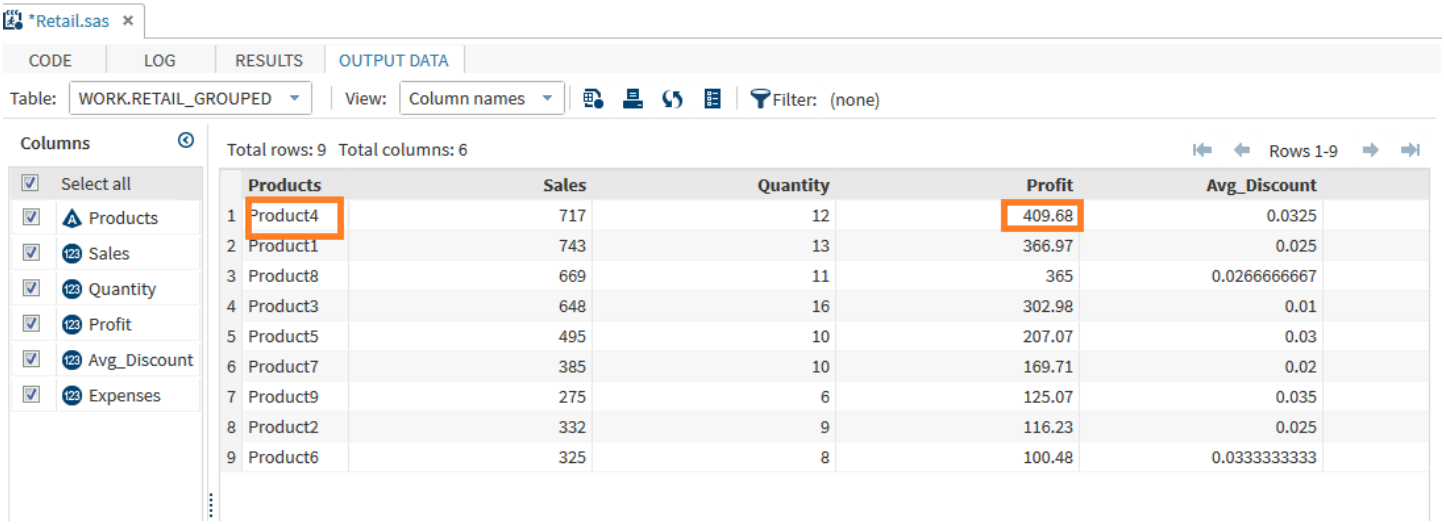
	Products	Sales	Quantity	Profit	Avg_Discount
1	Product3	648	16	302.98	0.01
2	Product1	743	13	366.97	0.025
3	Product4	717	12	409.68	0.0325
4	Product8	669	11	365	0.0266666667
5	Product5	495	10	207.07	0.03
6	Product7	385	10	169.71	0.02
7	Product2	332	9	116.23	0.025
8	Product6	325	8	100.48	0.0333333333
9	Product9	275	6	125.07	0.035

INFERENCE:

Product 1 Sells the most

- Next, to identify the Product that yields maximum Profit , a Descending Sort is performed on the Profit.

OUTPUT:


*Retail.sas

CODE | LOG | RESULTS | OUTPUT DATA

Table: WORK.RETAIL_GROUPED | View: Column names | Filter: (none)

Columns: Select all, Products, Sales, Quantity, Profit, Avg_Discount, Expenses

Total rows: 9 Total columns: 6 Rows 1-9

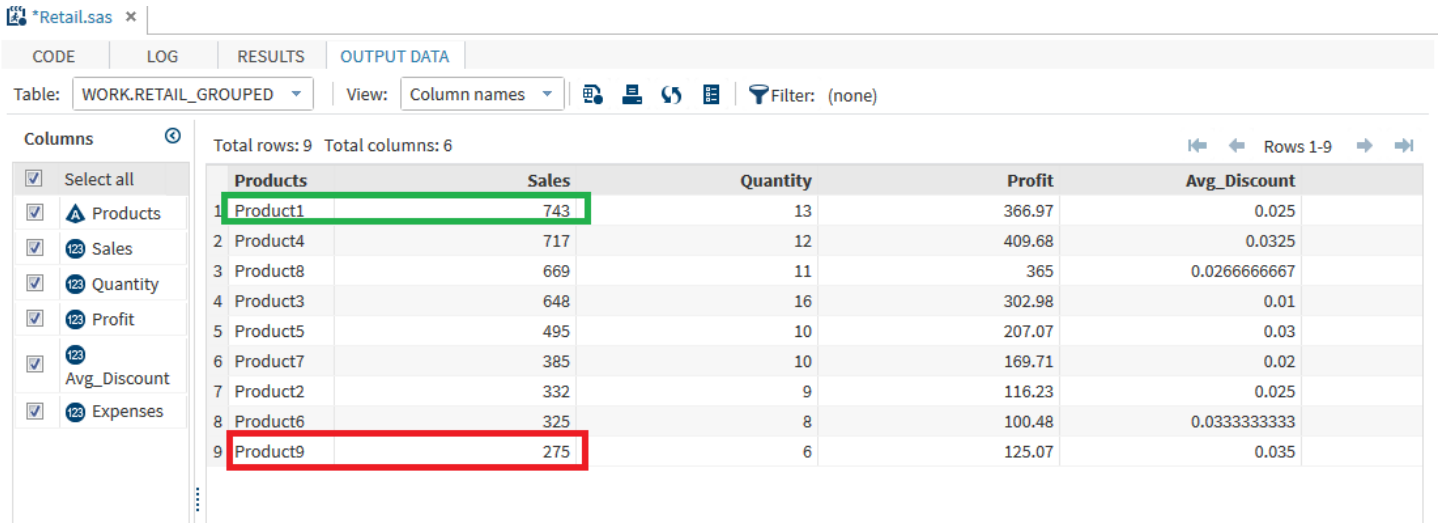
	Products	Sales	Quantity	Profit	Avg_Discount
1	Product4	717	12	409.68	0.0325
2	Product1	743	13	366.97	0.025
3	Product8	669	11	365	0.0266666667
4	Product3	648	16	302.98	0.01
5	Product5	495	10	207.07	0.03
6	Product7	385	10	169.71	0.02
7	Product9	275	6	125.07	0.035
8	Product2	332	9	116.23	0.025
9	Product6	325	8	100.48	0.0333333333

INFERENCE:

Product 4 yields the highest Profit.

- Next, to identify the Product that had Maximimand Minimum Sales Profit , a Descending Sort is performed on the Sales.

OUTPUT:


*Retail.sas

CODE | LOG | RESULTS | OUTPUT DATA

Table: WORK.RETAIL_GROUPED | View: Column names | Filter: (none)

Columns: Select all, Products, Sales, Quantity, Profit, Avg_Discount, Expenses

Total rows: 9 Total columns: 6 Rows 1-9

	Products	Sales	Quantity	Profit	Avg_Discount
1	Product1	743	13	366.97	0.025
2	Product4	717	12	409.68	0.0325
3	Product8	669	11	365	0.0266666667
4	Product3	648	16	302.98	0.01
5	Product5	495	10	207.07	0.03
6	Product7	385	10	169.71	0.02
7	Product2	332	9	116.23	0.025
8	Product6	325	8	100.48	0.0333333333
9	Product9	275	6	125.07	0.035

INFERENCE:

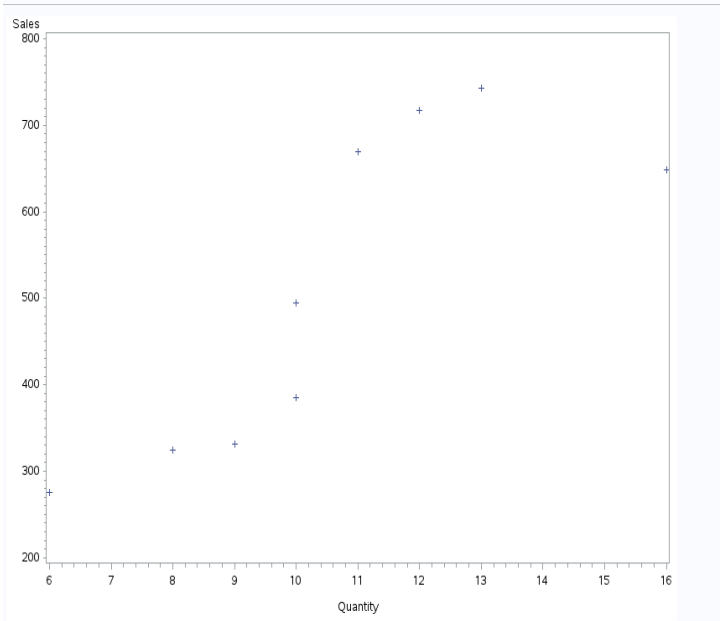
Product 1 has the Maximum Sales of 743$

Product 9 has the least Sales of 275$

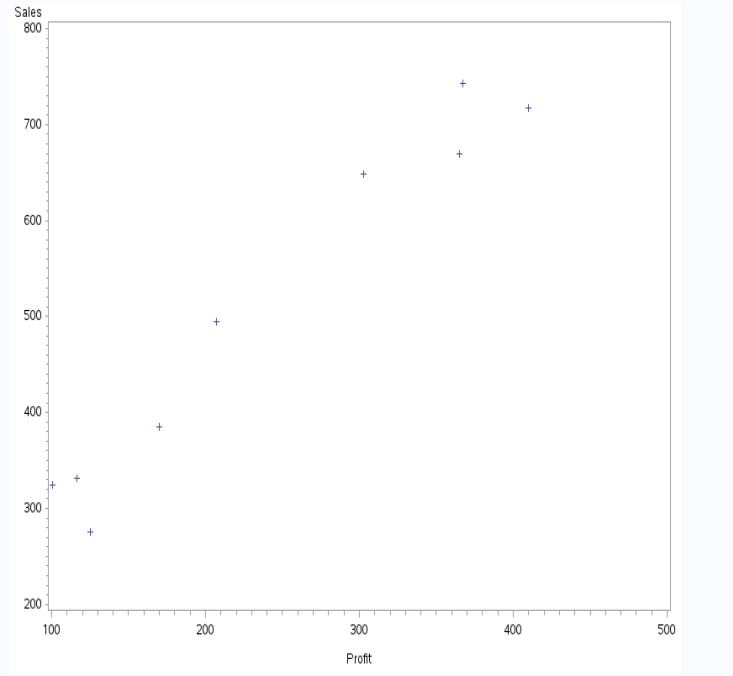
Analyze the Significance of Independent Variables

In Order to visualize the significance of each of the Independent Variable , first a plot of the Sales Vs.each of the Independent Variables is created .

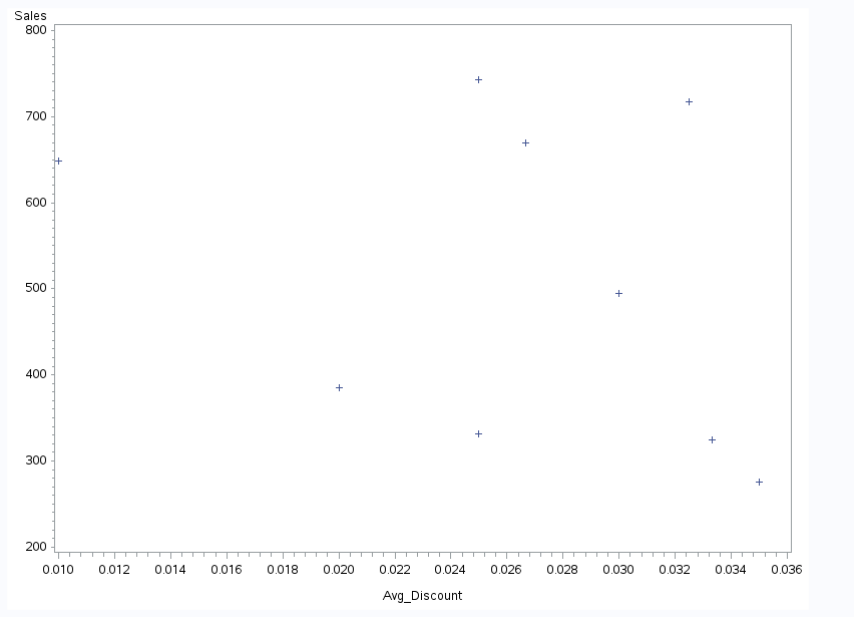
Plot of Sales Vs.Quantity



Plot of Sales Vs.Profit



Plot of Sales Vs.Average Discount



Analysis of Significance using Regression :

A Regression Model is created for the Dependent variable Sales based on Independent Variables Quantity,Profit and Average Discount

MODEL OUTPUT:

The REG Procedure
Model: MODEL1
Dependent Variable: Sales

Number of Observations Read	9
Number of Observations Used	9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	274416	91472	113.45	<.0001
Error	5	4031.23416	806.24683		
Corrected Total	8	278447			

Root MSE	28.39449	R-Square	0.9855
Dependent Mean	509.88889	Adj R-Sq	0.9768
Coeff Var	5.56876		

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.

Expenses =	0.1 * Profit

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-181.55791	159.65872	-1.14	0.3070
Quantity	1	30.78977	11.54626	2.67	0.0445
Profit	B	1.01754	0.18676	5.45	0.0028
Expenses	0	0	.	.	.
Avg_Discount	1	4618.33813	3010.08729	1.53	0.1855

INFERENCE :

It can be seen that Quantity and Profit have p-values less than 0.05 and are significant.

However the Discount appears to have no significant impact on Sales .

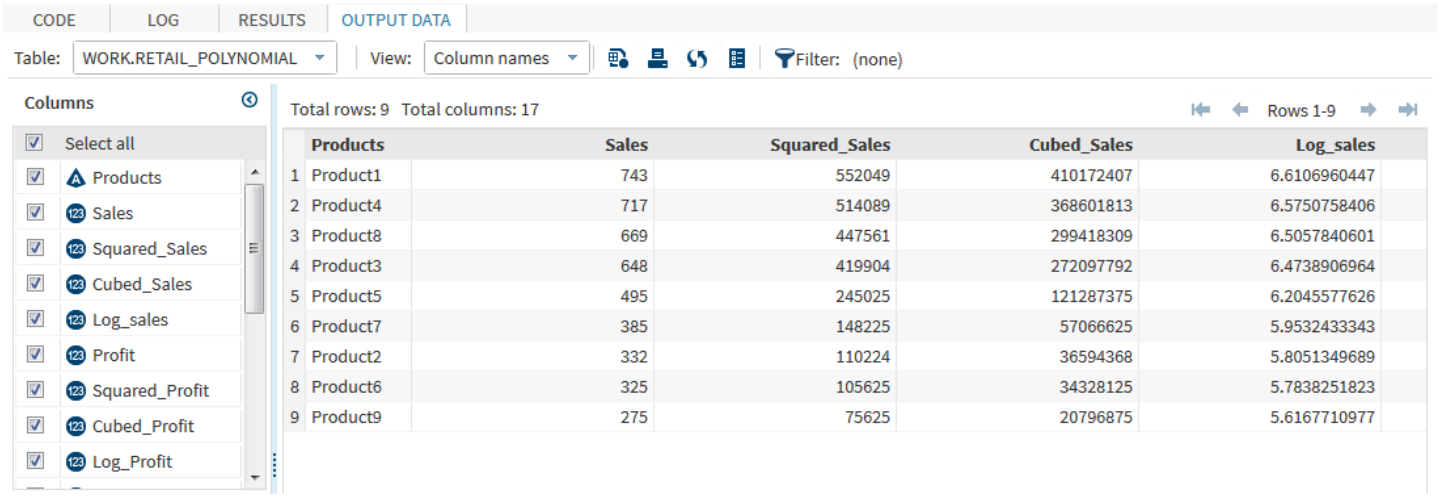
The R-Squared Value is 0.9855 which indicates that the Model is described very well by the parameters used .

Create Dataset with Polynomial Features

A New Table with Polynomial Values for all the Parameters is created using the code below :

```
PROC SQL;
CREATE TABLE RETAIL_POLYNOMIAL AS
SELECT PRODUCTS,SALES, (SALES**2) AS SQUARED_SALES,(SALES**3) AS CUBED_SALES ,LOG(SALES) AS LOG_SALES,
    PROFIT,   (PROFIT**2) AS SQUARED_PROFIT,(PROFIT**3) AS CUBED_PROFIT ,LOG(PROFIT) AS LOG_PROFIT,
    QUANTITY, (QUANTITY**2) AS SQUARED_QUANTITY,(QUANTITY**3) AS CUBED_QUANTITY ,LOG(QUANTITY) AS
LOG_QUANTITY,
    EXPENSES, (EXPENSES**2) AS SQUARED_EXPENSES,(EXPENSES**3) AS CUBED_EXPENSES ,LOG(EXPENSES) AS
LOG_EXPENSES
FROM RETAIL_GROUPED
QUIT;
```

OUTPUT:

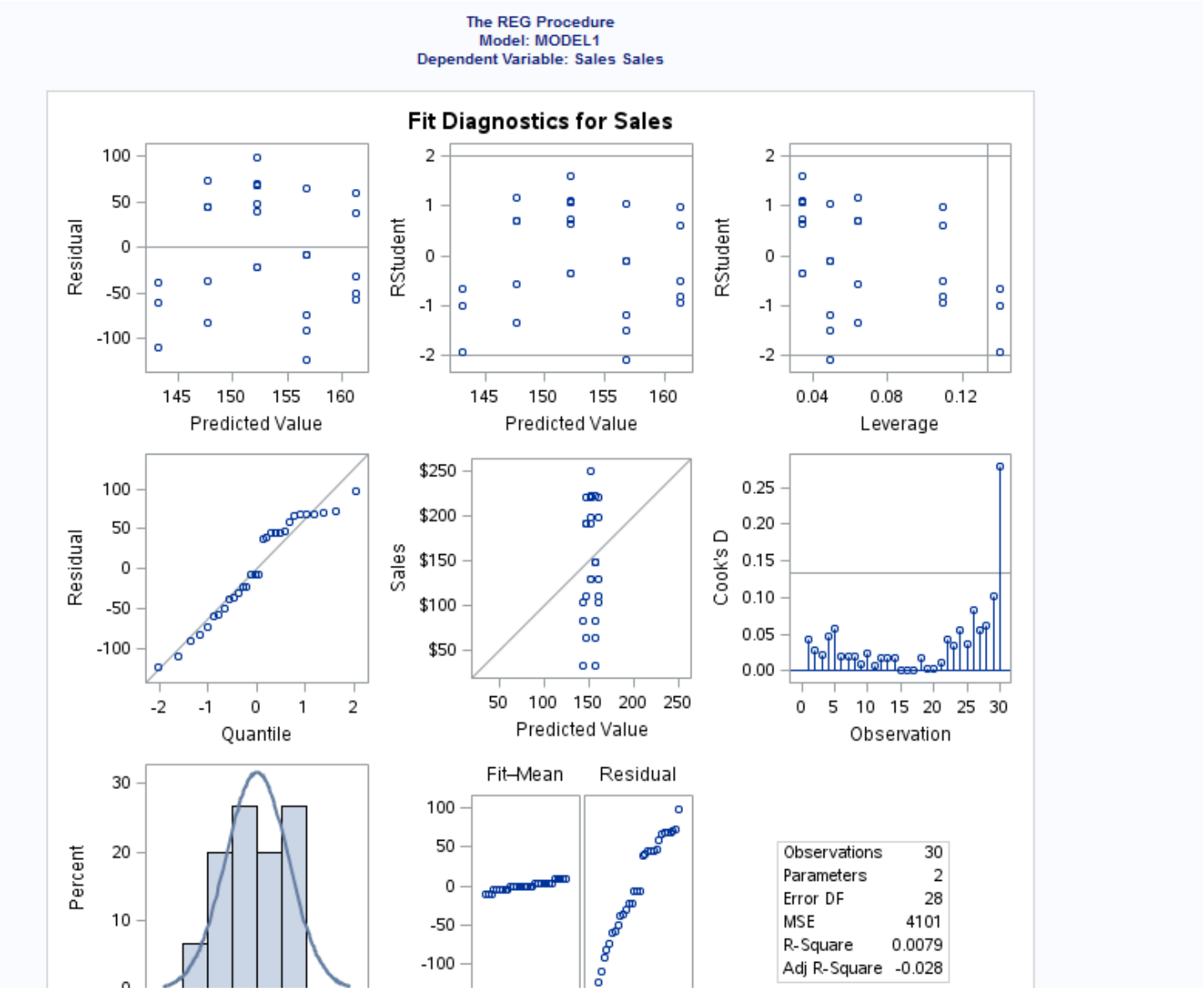


	Products	Sales	Squared_Sales	Cubed_Sales	Log_sales
1	Product1	743	552049	410172407	6.6106960447
2	Product4	717	514089	368601813	6.5750758406
3	Product8	669	447561	299418309	6.5057840601
4	Product3	648	419904	272097792	6.4738906964
5	Product5	495	245025	121287375	6.2045577626
6	Product7	385	148225	57066625	5.9532433343
7	Product2	332	110224	36594368	5.8051349689
8	Product6	325	105625	34328125	5.7838251823
9	Product9	275	75625	20796875	5.6167710977

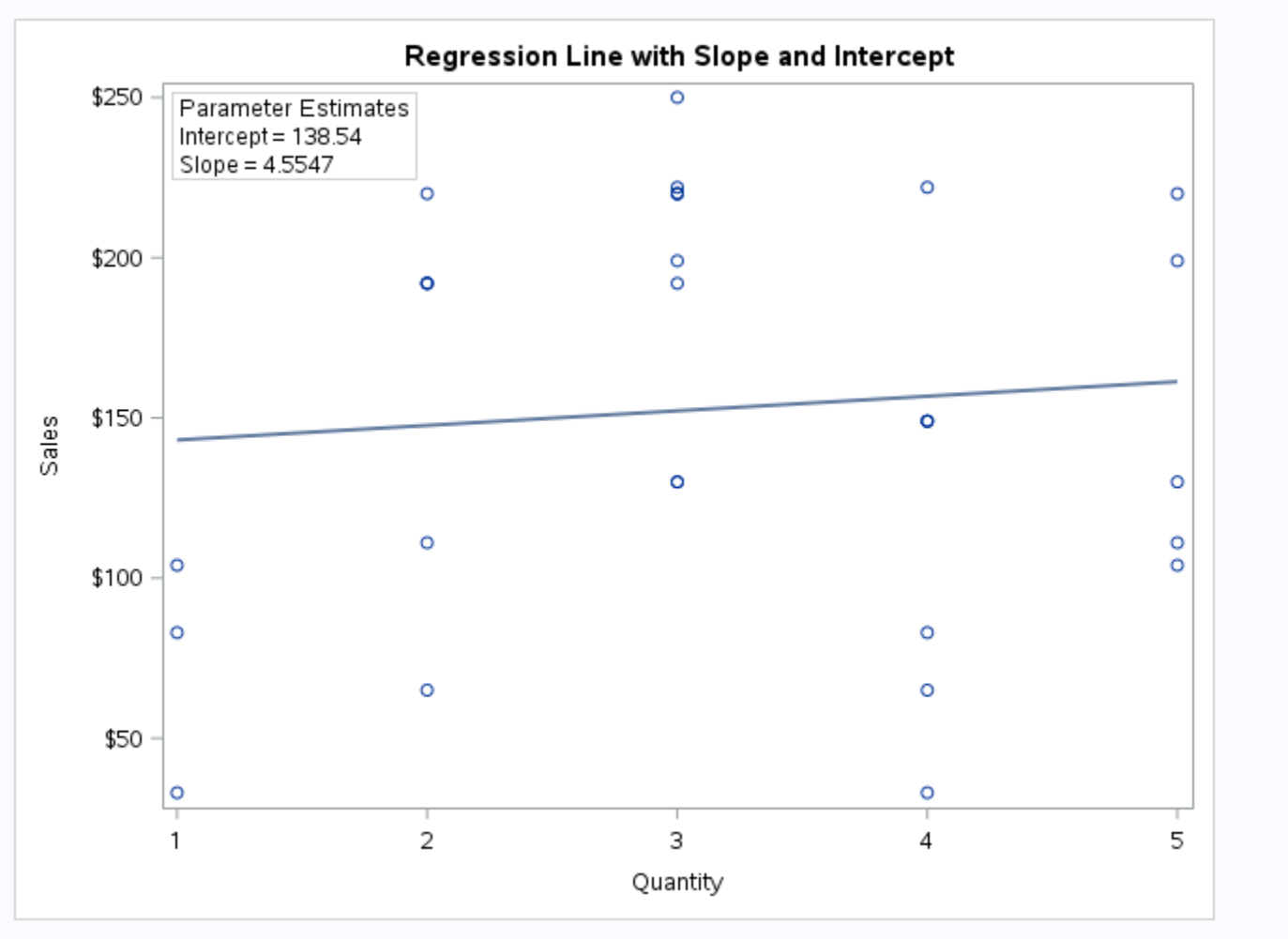
Perform Regression Test

A Regression Test is performed to identify the Impact of each variable on the Sales .

Model of Sales Vs. Quantity :



Below is the Regression Line for Sales Vs.Quantity Model:



INFERENCE :

The Quantity parameter alone is insufficient to describe the Sales .The R-Squared Value for this Model is very low .

However it is true that an increase in Quantity obviously has an increase in Sales .

Model of Sales Vs. Profit :

The model parameters for the Sales Vs.Profit model is given below :

OUTPUT:

The REG Procedure
Model: MODEL1
Dependent Variable: Sales

Number of Observations Read	9
Number of Observations Used	9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	265543	265543	144.05	<.0001
Error	7	12904	1843.38955		
Corrected Total	8	278447			

Root MSE	42.93471	R-Square	0.9537
Dependent Mean	509.88889	Adj R-Sq	0.9470
Coeff Var	8.42041		

Parameter Estimates

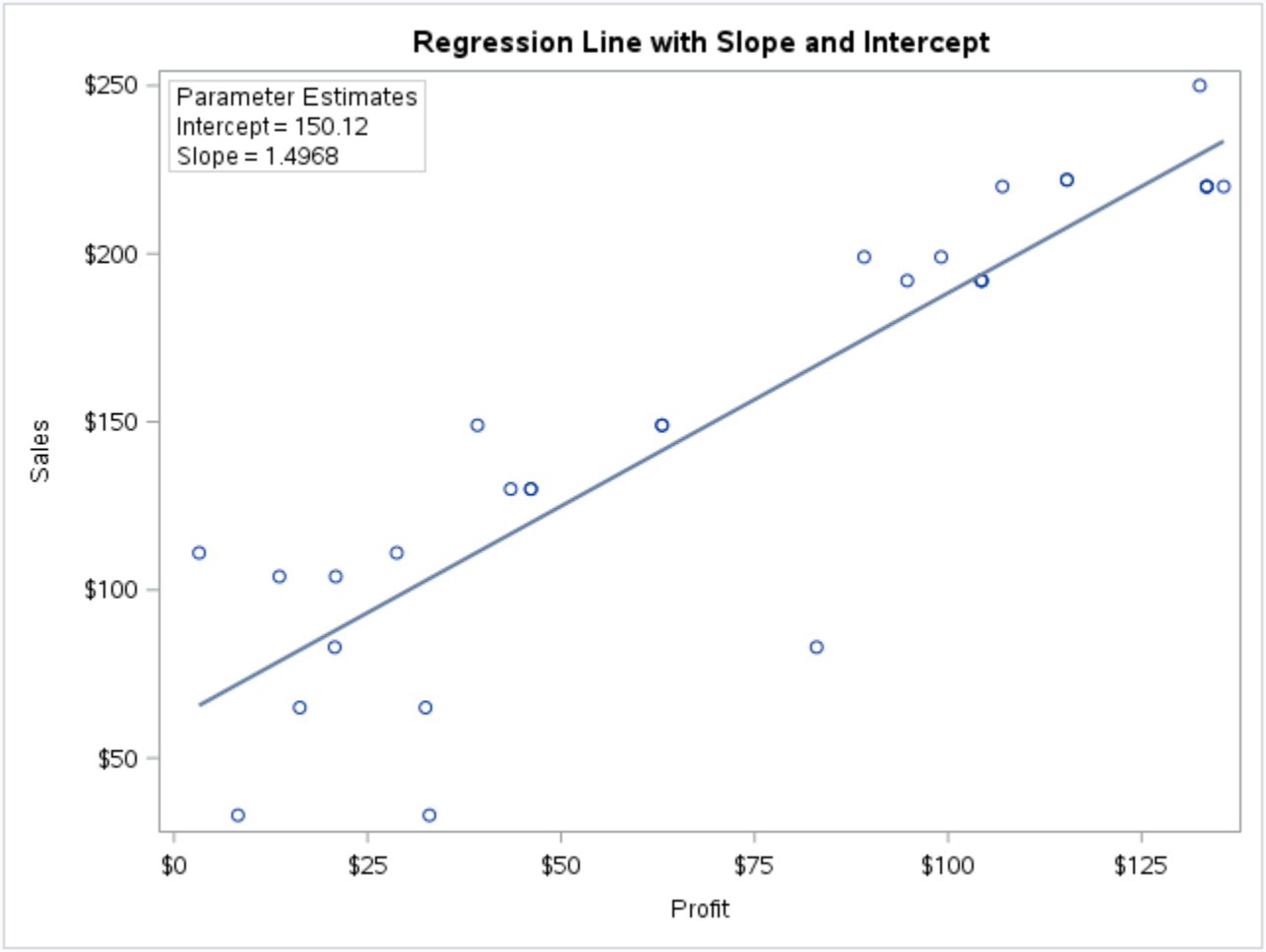
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	150.12257	33.21643	4.52	0.0027
Profit	1	1.49682	0.12471	12.00	<.0001

INFERENCE :

It can be seen that the p-value is very low which implies the Profit has a high impact on Sales .

The R-Squared Value is close to 1 which indicates that this parameter describes the Model to a great extent .

Below is the Regression Line which exactly describes the linearity in the relationship between Sales and Profit .


Regression Line with Slope and Intercept
Parameter Estimates
Intercept = 150.12
Slope = 1.4968
$250
$200
$150
$100
$50
Sales
$0
$25
$50
$75
$100
$125
Profit

Model of Sales Vs. Discount :

Below are the model Parameters for this model :

OUTPUT :

The REG Procedure
Model: MODEL1
Dependent Variable: Sales

Number of Observations Read	9
Number of Observations Used	9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	26554	26554	0.74	0.4188
Error	7	251893	35985		
Corrected Total	8	278447			

Root MSE	189.69644	R-Square	0.0954
Dependent Mean	509.88889	Adj R-Sq	-0.0339
Coeff Var	37.20349		

Parameter Estimates

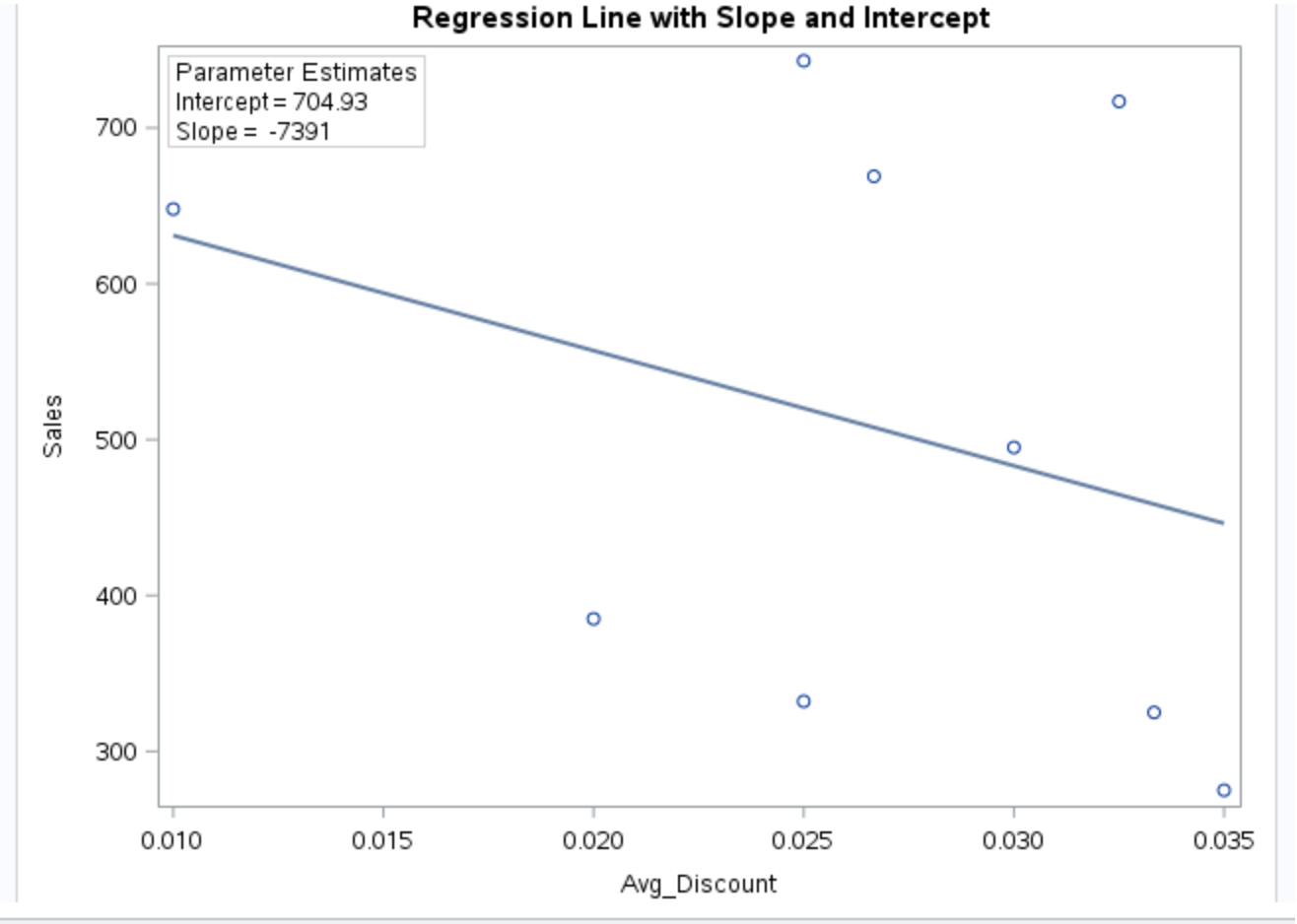
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	704.92540	235.68596	2.99	0.0202
Avg_Discount	1	-7390.85714	8603.82190	-0.86	0.4188

INFERENCE :

Although the significance is very less , The Discount has a negative Impact on Sales value described the negative Intercept and also can be seen in the Regression Line Plot below .

The insignificance of Discount can be proved by the high p-value >0.05.

Hence this model is irrelevant in the estimation of Sales .


Regression Line with Slope and Intercept
Parameter Estimates
Intercept = 704.93
Slope = -7391
700
600
500
400
300
Sales
0.010
0.015
0.020
0.025
0.030
0.035
Avg_Discount

Print the Output Dataset

The Output parameters of the Regression Models are printed below using the following Code :

```
PROC PRINT DATA=Work.PE ; RUN;
PROC PRINT DATA=Work.PE1 ; RUN;
PROC PRINT DATA=Work.PE2 ; RUN;
```

It can be seen that product 1 has the Maximum Sales

Obs	Model	Dependent	Variable	DF	Estimate	StdErr	tValue	Probt	Label
1	MODEL1	Sales	Intercept	1	138.54340	32.67743	4.24	0.0002	Intercept
2	MODEL1	Sales	Quantity	1	4.55472	9.63605	0.47	0.6401	Quantity

Obs	Model	Dependent	Variable	DF	Estimate	StdErr	tValue	Probt
1	MODEL1	Sales	Intercept	1	150.12257	33.21643	4.52	0.0027
2	MODEL1	Sales	Profit	1	1.49682	0.12471	12.00	<.0001

Obs	Model	Dependent	Variable	DF	Estimate	StdErr	tValue	Probt
1	MODEL1	Sales	Intercept	1	704.92540	235.68596	2.99	0.0202
2	MODEL1	Sales	Avg_Discount	1	-7390.85714	8603.82190	-0.86	0.4188