Spring 2019

BUAN 6337: Predictive Analytics using SAS

A Report of

Group Assignment 2 on Sales Data Analysis

Submitted by Group 7

Vinay Singh	**2021441554**
Vaibhav Shrivastava	**2021434681**
Megan Malisani	**2021440151**
Pragati Mishra	**2021434655**
Ishan Jain	**2021426222**
Erhao Liang	**2021435949**

Under the Guidance of

Prof. Shervin Shahrokhi Tehrani

Overview

The data contains the sale units of a European brand in a year at its stores across Europe and North America. They ran an experimental study to find the best product design in each market. They offered 4-types of product base on **Design** and **Quality** Attributes. The design has two versions: **Luxury Design** vs. **Normal Design**. Also, the quality has two conditions: **High Quality** vs. **Normal Quality**. The seasonal sale units of each store have been provided in Sales.csv, where you can observe the type of the product, its price, the store's advertising efforts, and the store's sales force experience in that season.

Following are the description variables present in the data file:

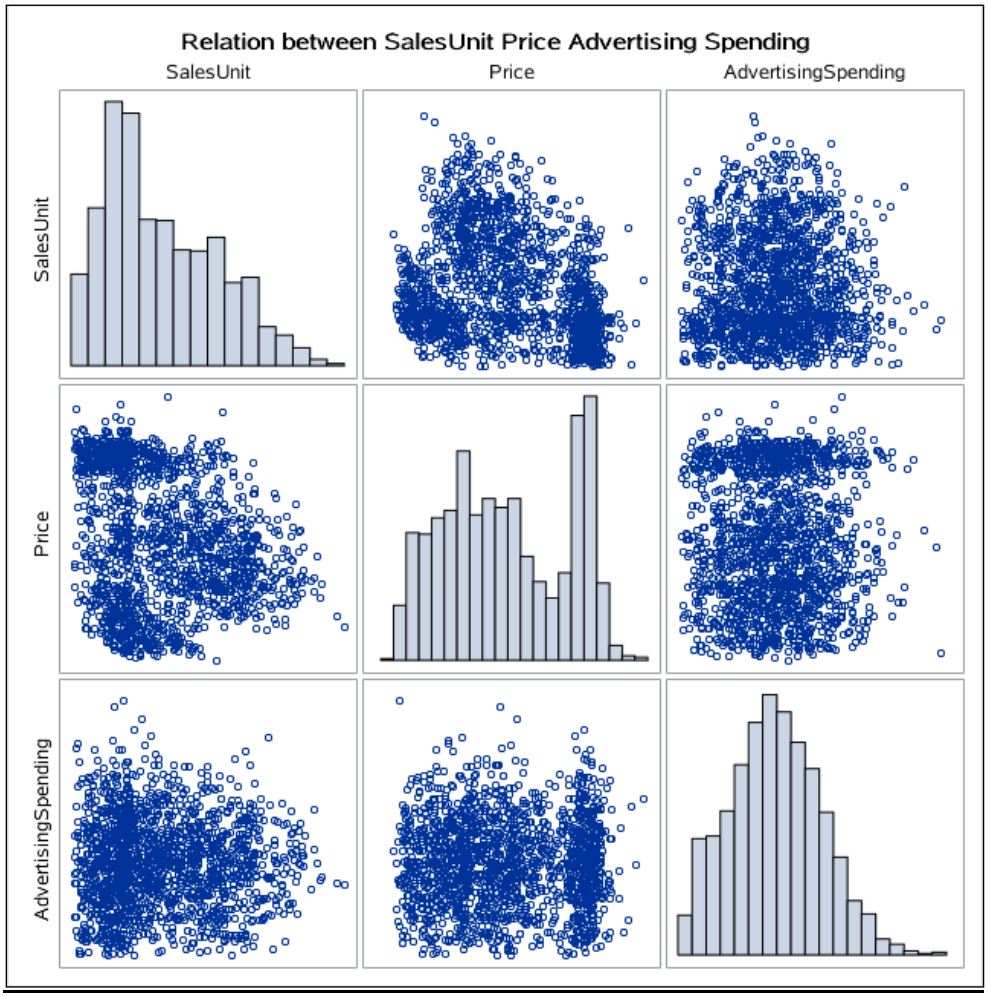
Variables	Description
ID	The store ID
Sales Units	How many unites of the product had been sold at this store
Design	1 means Luxury design, 2 means Normal design
Quality	1 means High quality, 2 means Normal quality
Price	Price per unit
Promotion	Promotion means the price had been labeled as the promotional price (It was discounted price)
Advertising Spending	The amount of the money which had been spent to advertise in the store location
Salesforce Experience	Showing the level of salesforce experience in the store; Low, Average and High
Location	Where is the Store NA: North America EU: Europe
Season	Showing the season; 1Winter, 2Spring, 3Summer, 4Fall

Condition: Confidence interval for following hypothesis testing is 95%

1. SCATTER plot

Use Proc sgscatter to plot a matrix of Sales unit, price, and advertising spending. What type of relationship do you see between Sales unit and price? (Linear or Non-linear)

What type of relationship do you see between Sales unit and advertising spending? (Linear or Non-linear)?



- SalesUnit vs Price: It is showing a negative correlation with the data. But it is not clear from the graph about the linear or non-linear relationship.

- SalesUnit vs AdvertisingSpending: It is showing a positive correlation with the data. But it is not clear from the graph about the linear or non-linear relationship.

2. Linear regression

Now, provide a linear regression of Sales unit over price? (Consider only price) Is the price coefficient significant? Does price's coefficient make sense? What does it mean?
Now, provide a non-linear regression of Sales unit over price and its power 2? Are the price and its power 2's coefficients significant? Explain your result in terms of consumers' reaction to increase of price in this market. Compare the R^2 of the above models (linear vs. non-linear). Which model fits the data better and why?

- The Linear model coefficient Price is highly significant since P value is < 0.0001. The coefficient of the Price is -8.11 which means that if 1 unit of Price is increased it will lead to decrease of 8.11 Sales Unit. This does not make sense since with increase in Price, Sales Unit should increase. */

- Price and Price power 2's coefficient are significant in this model with P value < 0.0001 for both of them. This model tells a different story when compared to the linear regression:
  The intercept for Price is 51.9 and Price Square is -0.04 which tells us that the Consumer Reaction with increase in Price lets the SalesUnit increase initially but after a certain threshold
  it starts to decrease.

- The R squared value for linear model is 0.0684(Adj R sqrd = 0.0678) and for non linear model it is 0.1926(Adj sqrd = 0.1915).

We can see from the R squared values the nonlinear model fits better and so we can say that Sales Unit and Price are non-linearly related

Linear Regression of Sales Unit over Price

The REG Procedure
Model: MODEL1
Dependent Variable: SalesUnit

Number of Observations Read	1600
Number of Observations Used	1600

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	10638182325	10638182325	117.30	<.0001
Error	1598	1.449226E11	90690017		
Corrected Total	1599	1.555608E11			

Root MSE	9523.13065	R-Square	0.0684
Dependent Mean	35353	Adj R-Sq	0.0678
Coeff Var	26.93753		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	40885	563.58606	72.54	<.0001
Price	1	-8.87910	0.81981	-10.83	<.0001

Regression Analysis of Sales Unit over Price Price_2

The REG Procedure
Model: MODEL1
Dependent Variable: SalesUnit

Number of Observations Read	1600
Number of Observations Used	1600

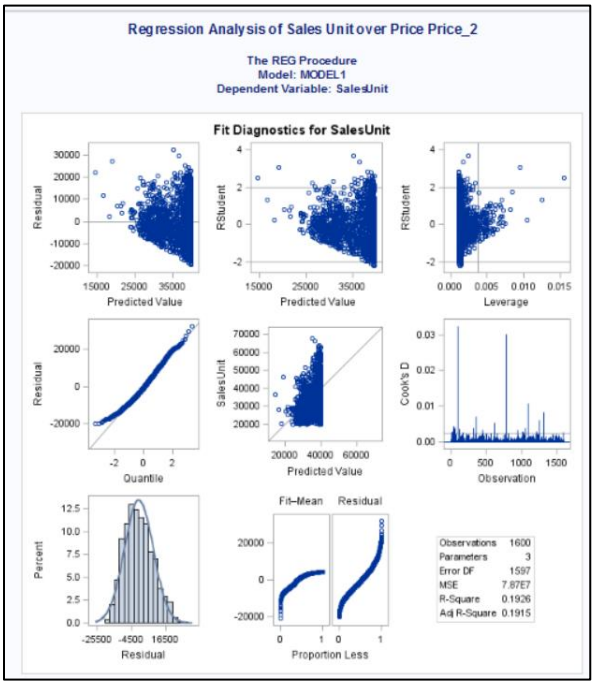
Analysis of Variance

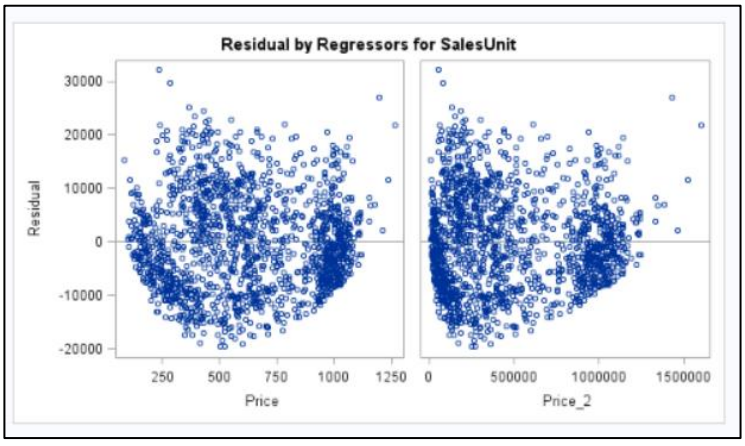
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	29954101807	14977050904	190.42	<.0001
Error	1597	1.256067E11	78651677		
Corrected Total	1599	1.555608E11			

Root MSE	8868.57807	R-Square	0.1926
Dependent Mean	35353	Adj R-Sq	0.1915
Coeff Var	25.08604		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	25676	1103.36746	23.27	<.0001
Price	1	51.95567	3.95630	13.13	<.0001
Price_2	1	-0.04803	0.00306	-15.67	<.0001





3. Linear regression

Now, provide a *linear* regression of Sales unit over advertising spending? (Consider only advertising spending) Is the advertising spending coefficient significant? Does advertising's coefficient make sense? If No, why? Now, provide a *non-linear* regression of Sales unit over advertising spending and its power 2? Are the coefficients significant? Explain your result in terms of Sales unit changes by increasing of advertising in this market. Compare the R^2 of the above models (linear vs. non-linear). Which model fits the data better and why?

- The Linear model coefficient Advertising Spending is not significant since P value is 0.4439 and greater than 0.001. The coefficient of the Price is -0.04877 which means that if 1 unit of Price is increased it will lead to decrease of -0.04877 Sales Unit. This does not make sense since with increase in Advertising Spending, Sales Unit should increase.

- Advertising Spending and Advertising Spending power 2's coefficient are significant in this model with P value < 0.0001 for both. This model tells a different story when compared to the linear regression: The intercept for Advertising Spending is 1.15 and Price Square is -0.00005715 which tells us that the with increase in Advertising Spending lets the Sales Unit increase initially but after a certain threshold it starts to decrease.

- The R squared value for linear model is 0.0004(Adj R sqrd = -0.0003) and for non linear model it is 0.0131(Adj sqrd = 0.0118).

- We can see from the R sqrd values the nonlinear model fits better and so we can say that SalesUnit and Advertising Spending are non-linearly related

Linear Regression of Sales Unit over Advertising Spending

The REG Procedure
Model: MODEL1
Dependent Variable: SalesUnit

Number of Observations Read	1600
Number of Observations Used	1600

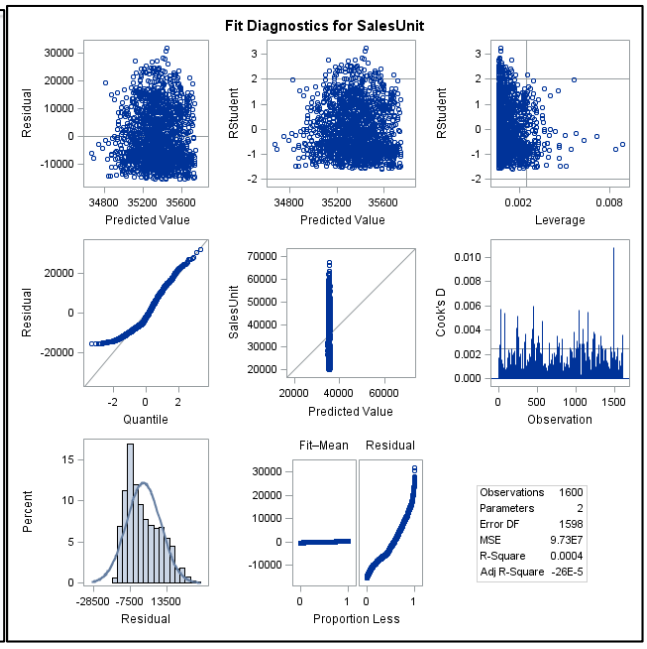
Analysis of Variance

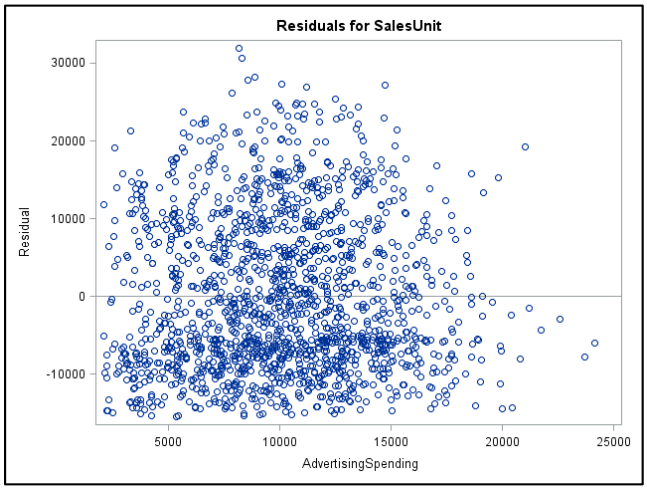
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	57069967	57069967	0.59	0.4439
Error	1598	1.555038E11	97311489		
Corrected Total	1599	1.555608E11			

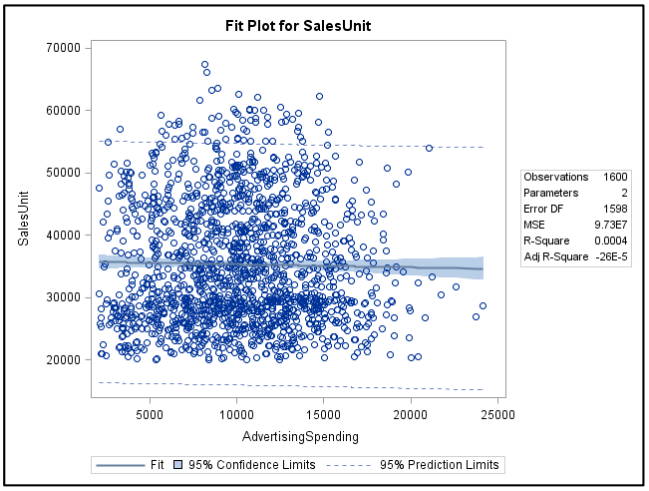
Root MSE	9864.65861	R-Square	0.0004
Dependent Mean	35353	Adj R-Sq	-0.0003
Coeff Var	27.90359		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	35845	688.38987	52.07	<.0001
AdvertisingSpending	1	-0.04877	0.06369	-0.77	0.4439







Regression Analysis of Sales Unit over Advertising Spending Advertising Spending_2

The REG Procedure
Model: MODEL1
Dependent Variable: SalesUnit

Number of Observations Read	1600
Number of Observations Used	1600

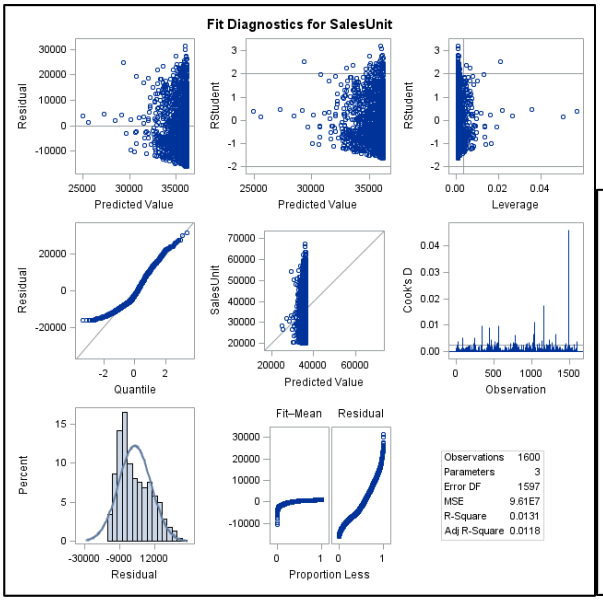
Analysis of Variance

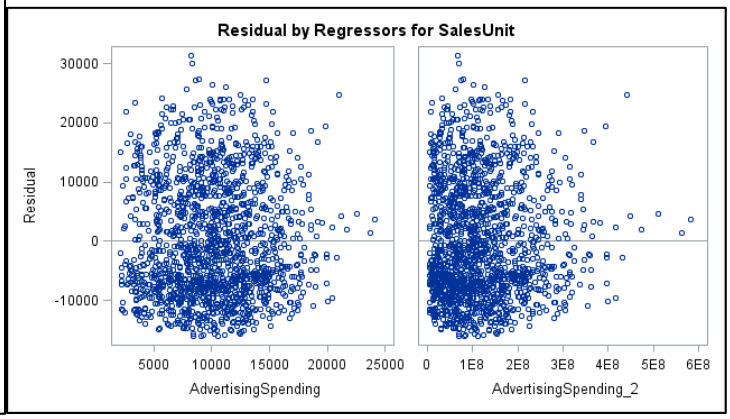
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2033598882	1016799441	10.58	<.0001
Error	1597	1.535272E11	96134772		
Corrected Total	1599	1.555608E11			

Root MSE	9804.83412	R-Square	0.0131
Dependent Mean	35353	Adj R-Sq	0.0118
Coeff Var	27.73437		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	30409	1380.33860	22.03	<.0001
AdvertisingSpending	1	1.15157	0.27219	4.23	<.0001
AdvertisingSpending_2	1	-0.00005715	0.00001260	-4.53	<.0001





4. Sensitivity Analysis

Imagine you **only** observe **Sales unit, price, and location variables**. You are interested in knowing who is more price sensitive? North Americans or European. Propose a regression model to find the most price sensitive group of consumers. (Hint: Use the Interaction effect). You should write your model precisely, estimate it, and explain the estimation results to identify the most price sensitive group of consumers.

North American are more Price Sensitive when compared to European. The sales unit decrease by 8.199 unit if the price is increased by 1 unit for North American when compared to Europe.

Price Sensitivity Analyis: North America vs European

The GLM Procedure

Class Level Information		
Class	Levels	Values
Location	2	NorthAmerica Europe

Number of Observations Read	1600
Number of Observations Used	1600

Price Sensitivity Analyis: North America vs European

The GLM Procedure

Dependent Variable: SalesUnit

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	40410946573	10102736643	139.94	<.0001
Error	1595	115149883478	72194284.312		
Corrected Total	1599	155560830051			

R-Square	Coeff Var	Root MSE	SalesUnit Mean
0.259776	24.03419	8496.722	35352.65

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Price	1	10638182325	10638182325	147.35	<.0001
Price_2	1	19315919482	19315919482	267.55	<.0001
Location	1	8196764339	8196764339	113.54	<.0001
Price*Location	1	2260080427	2260080427	31.31	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Price	1	13465311503	13465311503	186.51	<.0001
Price_2	1	19018455182	19018455182	263.43	<.0001
Location	1	23997891	23997891	0.33	0.5643
Price*Location	1	2260080427	2260080427	31.31	<.0001

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

5. Sensitivity Analysis

Imagine you **only** observe **Sales unit, location, and promotion variables**. You are interested in knowing (a) do consumers respond to promotional prices? and (b) who is more responsive to promotional offers? North Americans or European. Propose a regression model to answer the above questions. (Hint: Use the Interaction effect). You should write your model precisely, estimate it, and explain the estimation results to answer the above questions. Are your results compatible with your answer in Q4?

Price Sensitivity Analyis based on Promotions

The GLM Procedure

Class Level Information		
Class	Levels	Values
Promotion	2	YesPromotion NoPromotion

Number of Observations Read	1600
Number of Observations Used	1600

a) Yes, consumer response to the promotions has a positive correlation with the Sales Unit as seen by the glm model of Sales Unit vs Promotion.

Price Sensitivity Analyis based on Promotions

The GLM Procedure

Dependent Variable: SalesUnit

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11170273980	11170273980	123.62	<.0001
Error	1598	144390556071	90357043.849		
Corrected Total	1599	155560830051			

R-Square	Coeff Var	Root MSE	SalesUnit Mean
0.071806	26.88803	9505.632	35352.65

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Promotion	1	11170273980	11170273980	123.62	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Promotion	1	11170273980	11170273980	123.62	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	32475.40710	B	351.3382764	92.43	<.0001
Promotion YesPromotion	5303.67009	B	477.0079258	11.12	<.0001
Promotion NoPromotion	0.00000	B	.	.	.

Price Sensitivity Analyis in Location based on Promotions

The GLM Procedure

Class Level Information		
Class	Levels	Values
Location	2	NorthAmerica Europe
Promotion	2	YesPromotion NoPromotion

Number of Observations Read	1600
Number of Observations Used	1600

Price Sensitivity Analyis in Location based on Promotions

The GLM Procedure

Dependent Variable: SalesUnit

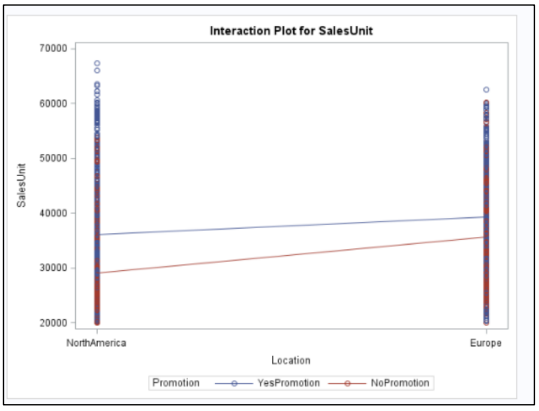
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	21431417620	7143805873.2	85.00	<.0001
Error	1596	134129412431	84040985.233		
Corrected Total	1599	155560830051			

R-Square	Coeff Var	Root MSE	SalesUnit Mean
0.137769	25.93126	9167.387	35352.65

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Promotion	1	11170273980	11170273980	132.91	<.0001
Location	1	9114401954	9114401954	108.45	<.0001
Location*Promotion	1	1146741685	1146741685	13.65	0.0002

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Promotion	1	11323068255	11323068255	134.73	<.0001
Location	1	9603708715	9603708715	114.27	<.0001
Location*Promotion	1	1146741685	1146741685	13.65	0.0002

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	35694.95479	B	472.7718885	75.50	<.0001
Promotion YesPromotion	3642.09655	B	641.1745032	5.68	<.0001
Promotion NoPromotion	0.00000	B	.	.	.
Location NorthAmerica	-6619.96883	B	677.9258380	-9.77	<.0001
Location Europe	0.00000	B	.	.	.
Location*Promotion NorthAmerica YesPromotion	3400.15559	B	920.4740684	3.69	0.0002
Location*Promotion NorthAmerica NoPromotion	0.00000	B	.	.	.
Location*Promotion Europe YesPromotion	0.00000	B	.	.	.
Location*Promotion Europe NoPromotion	0.00000	B	.	.	.



Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

b) As seen in the glm model, North American seems more responsive towards promotional offers. If a promotion is there the Sales Unit increases by 3400 unit for North American when compared to Europe. Yes, the result is found compatible with the Question 4.

6. Hypothesis Testing

Imagine you **only** observe **Sales unit, price, location, Design, and Quality variables**. We know the firms offer 4-types of product. Define a new categorical variable called **Product-Type** to code the 4-types of firm's products. Define the **Product-Type** as follows:

Design =1	Quality =1	**Product-Type = 1**
Design =1	Quality =2	**Product-Type = 2**
Design =2	Quality =1	**Product-Type = 3**
Design =2	Quality =2	**Product-Type = 4**

First, provide a hypothesis testing to find: is there any significant difference among the average prices of these 4-types of product? (Hint ANOVA) If yes, RANK the types of product based on their average price. Which attribute of product does increase the price of a type of product more significantly, Design or Quality?
You are interested in knowing (a) which type of product is the least popular among consumers? (b) Also, you want to know do North Americans vs. Europeans have different preference for Design and Quality? In other words, is the most important attributes of product (i.e., Design vs Quality) different among North Americans vs. Europeans? You should write precisely a regression model, estimate it, and explain the estimation results to identify the above questions. Are your results consistent with your results in Q4 and Q5?

Hypothesis Testing

Ho- Price are same for all product types
H1 – Prices are different for all product types

Hypothesis Testing for Product Type vs Average Price

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
Product_Type	4	1 2 3 4

Number of Observations Read	1600
Number of Observations Used	1600

Hypothesis Testing for Product Type vs Average Price

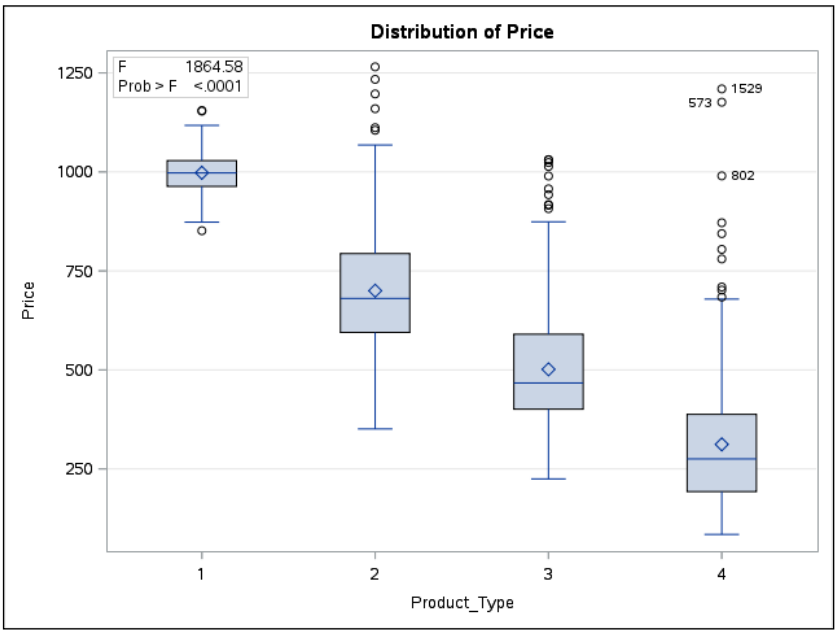
The ANOVA Procedure

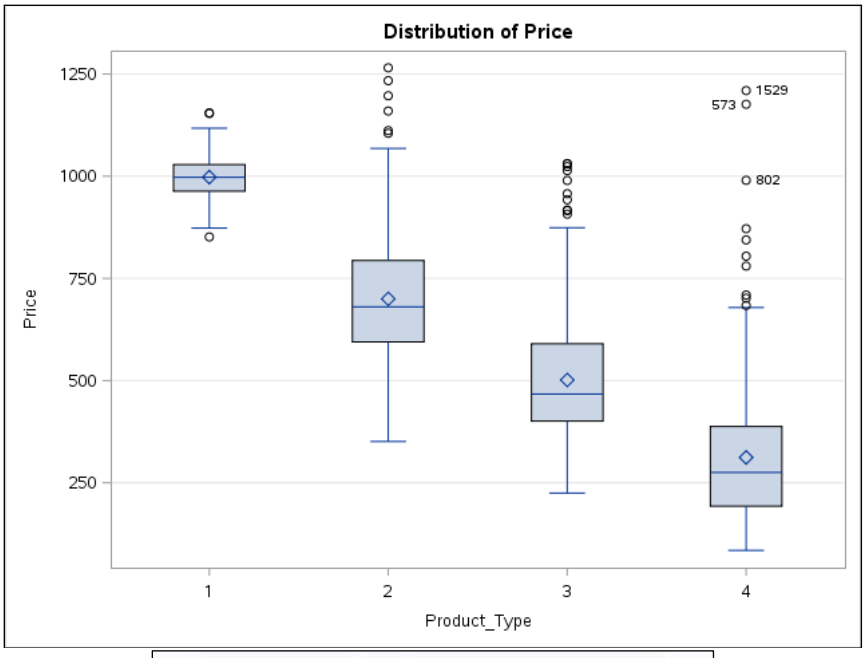
Dependent Variable: Price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	104982976.6	34994325.5	1864.58	<.0001
Error	1596	29953616.4	18767.9		
Corrected Total	1599	134936593.0			

R-Square	Coeff Var	Root MSE	Price Mean
0.778017	21.98599	136.9961	623.1062

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Product_Type	3	104982976.6	34994325.5	1864.58	<.0001





Level of Product_Type	N	Price	
		Mean	Std Dev
1	404	997.777129	48.831099
2	364	699.905357	152.738234
3	420	501.446357	154.787548
4	412	311.881092	159.724204

If the Design = 1 or Luxury Design, the average price of those product is higher when compared to product with Quality.

(a) which type of product is the least popular among consumers?

Hypothesis Testing for Product Type vs Average Sales Unit

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
Product_Type	4	1 2 3 4

Number of Observations Read	1600
Number of Observations Used	1600

Hypothesis Testing for Product Type vs Average Sales Unit

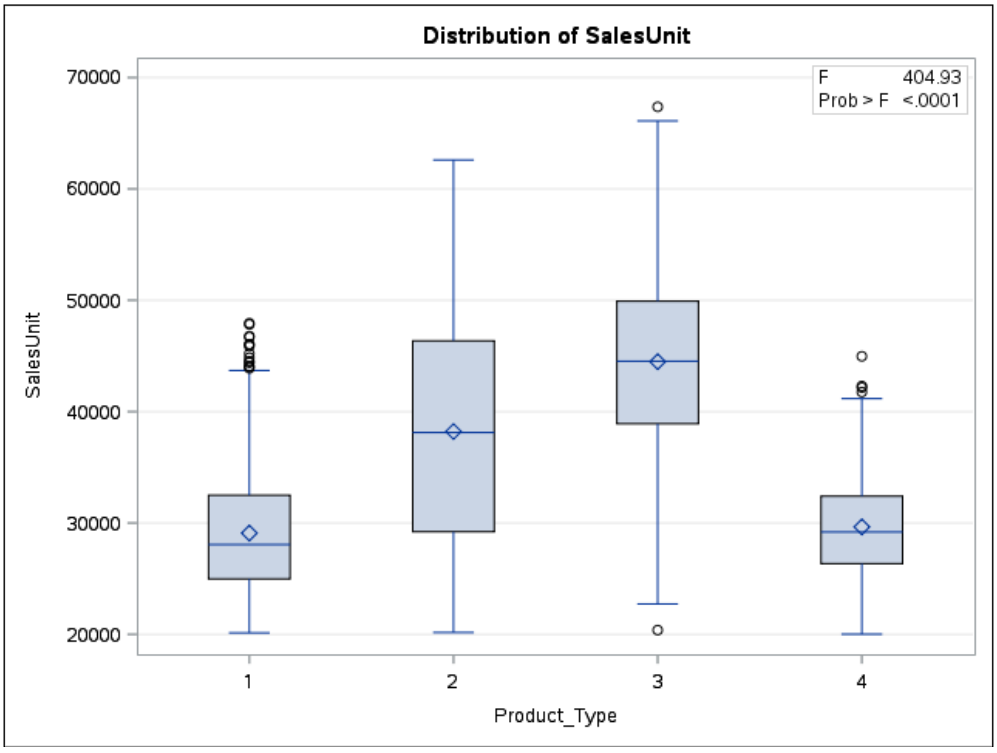
The ANOVA Procedure

Dependent Variable: SalesUnit

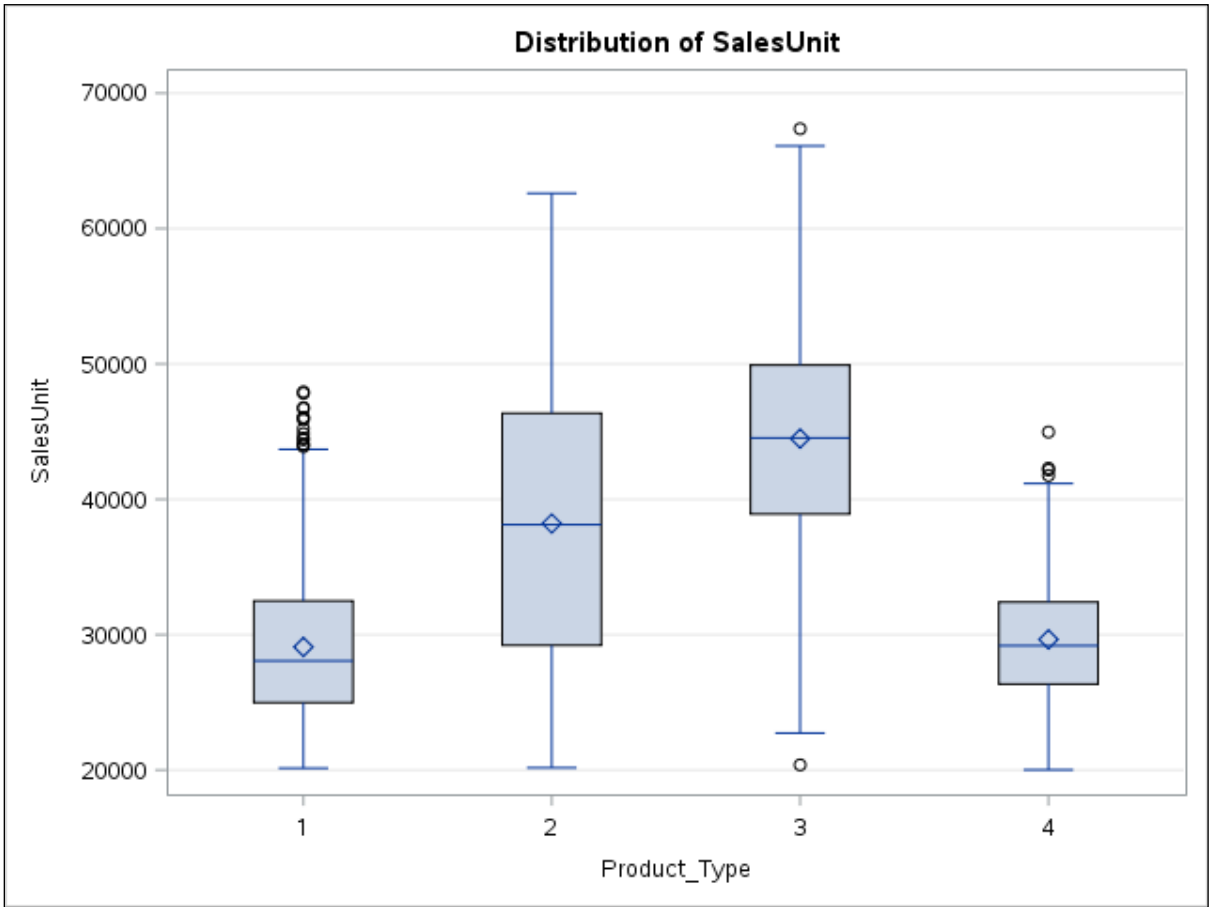
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	67232002509	22410667503	404.93	<.0001
Error	1596	88328827542	55343876.906		
Corrected Total	1599	155560830051			

R-Square	Coeff Var	Root MSE	SalesUnit Mean
0.432191	21.04325	7439.347	35352.65

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Product_Type	3	67232002509	22410667503	404.93	<.0001



Hypothesis Testing for Product Type vs Average Sales Unit
The ANOVA Procedure



Level of Product_Type	N	SalesUnit	
		Mean	Std Dev
1	404	29095.3713	6017.6124
2	364	38212.4698	10260.1317
3	420	44486.6548	7931.1501
4	412	29650.4199	4722.4659

- The popularity of the product can be determined using the Sales Unit. As seen, mean sales unit for the Product Type =1 (Luxury Design and High Quality) is the lowest and hence it is the least popular product among consumer.

(b) Also, you want to know do North Americans vs. Europeans have different preference for Design and Quality? In other words, is the most important attributes of product (i.e., Design vs Quality) different among North Americans vs. Europeans?

Preference for type of product

The GLM Procedure

Class Level Information		
Class	Levels	Values
Location	2	NorthAmerica Europe
Product_Type	4	2 3 4 1

Number of Observations Read	1600
Number of Observations Used	1600

Preference for type of product

The GLM Procedure

Dependent Variable: SalesUnit

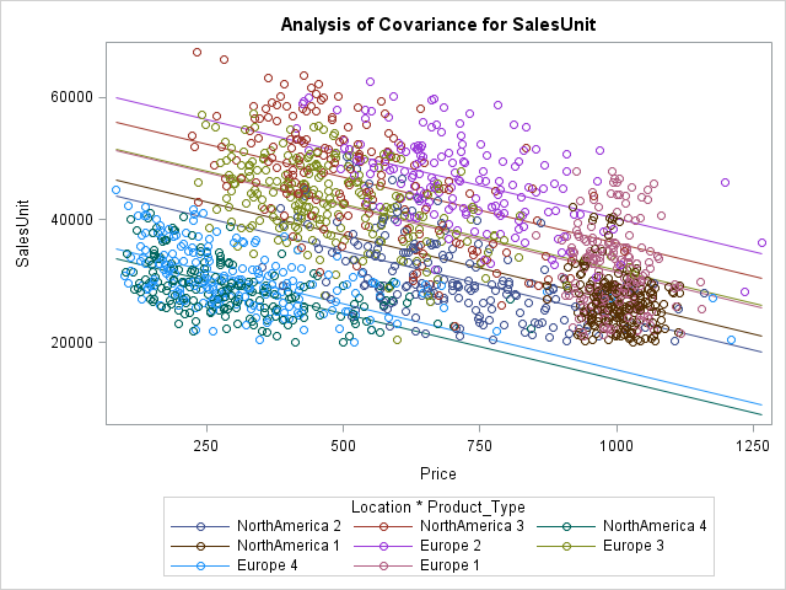
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	106763298087	13345412261	435.12	<.0001
Error	1591	48797531964	30670981.75		
Corrected Total	1599	155560830051			

R-Square	Coeff Var	Root MSE	SalesUnit Mean
0.686312	15.66541	5538.139	35352.65

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Price	1	10638182325	10638182325	346.85	<.0001
Location	1	8220227606	8220227606	268.01	<.0001
Location*Product_Typ	6	87904888155	14650814693	477.68	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Price	1	13886384180	13886384180	452.75	<.0001
Location	1	8014364159	8014364159	261.30	<.0001
Location*Product_Typ	6	87904888155	14650814693	477.68	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	53077.92588	B	1084.501704	48.94	<.0001
Price	-21.60952		1.015580	-21.28	<.0001
Location NorthAmerica	-4702.46907	B	551.351565	-8.53	<.0001
Location Europe	0.00000	B	.	.	.
Location*Product_Typ NorthAmerica 2	-2601.73572	B	639.030375	-4.07	<.0001
Location*Product_Typ NorthAmerica 3	9475.17305	B	743.784788	12.74	<.0001
Location*Product_Typ NorthAmerica 4	-12880.79303	B	892.765973	-14.43	<.0001
Location*Product_Typ NorthAmerica 1	0.00000	B	.	.	.
Location*Product_Typ Europe 2	8701.98649	B	645.157057	13.49	<.0001
Location*Product_Typ Europe 3	348.77096	B	742.826884	0.47	0.6388
Location*Product_Typ Europe 4	-15875.45389	B	880.861922	-18.02	<.0001
Location*Product_Typ Europe 1	0.00000	B	.	.	.



Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

7. Sales Force Experience

Imagine you **only** observe **Sales unit, price, Salesforce experience, Design, and Quality variables**. Create a categorical variable **Product-Type** to code the 4-types product as in Q6.
You are interested in knowing (a) How far does the Salesforce experience increase the sales? (b) Also, you want to know do high experience Salesforce have any expertise to sell specific type of products? In other words, do their experience help them to sell more high-ended products? You should write precisely a regression model, estimate it, and explain the estimation results to identify the above questions.

a) There is a high correlation between the Salesforce experience and Sales Unit. As seen in the glm model, coefficient of High Salesforce experience is 10430.034, which suggest that with High Experience the Sales Unit increases by 10430-unit coefficient of Average Salesforce experience is 10430.034, which suggest that with Average Experience the Sales Unit increases by 10430 unit when compared to Low experience coefficient of High Salesforce experience is 3720, which suggest that with High Experience the Sales Unit increases by 3720 unit when compared to Low experience.

b) As seen in the glm model, High Sales force experience have an ability to sell specific type of products. They are good in selling Product type=3 (Low Design and High Quality), with coefficient of 348 which tells us that High Sales force experience sells 348 Unit of Low-Design High-Quality products more when compared to High Design and High Quality.

It shows that High Sales Force Experience can sell higher-more ended product, specifically high-quality product when compared low quality product.

Preference for type of product based on SalesForce Experience

The GLM Procedure

Dependent Variable: SalesUnit

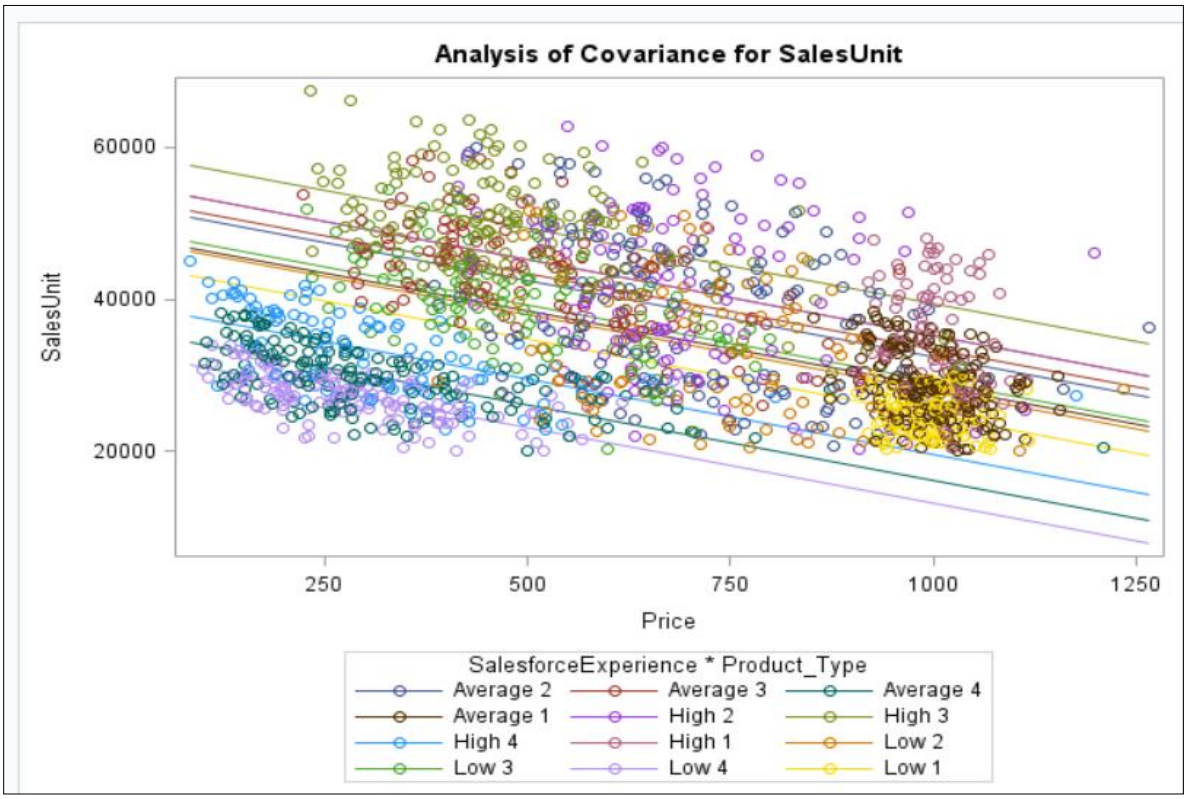
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	97298900387	8108241699	220.86	<.0001
Error	1587	58261929663	36711990.966		
Corrected Total	1599	155560830051			

R-Square	Coeff Var	Root MSE	SalesUnit Mean
0.625472	17.13886	6059.042	35352.65

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Price	1	10638182325	10638182325	289.77	<.0001
SalesforceExperience	2	25951492161	12975746081	353.45	<.0001
Salesforc*Product_Ty	9	60709225901	6745469545	183.74	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Price	1	11802621630	11802621630	321.49	<.0001
SalesforceExperience	2	16877335944	8438667972	229.86	<.0001
Salesforc*Product_Ty	9	60709225901	6745469545	183.74	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	44762.14580	B	1242.979590	36.01	<.0001
Price	-19.92829		1.111437	-17.93	<.0001
SalesforceExperience Average	3720.95493	B	715.120626	5.20	<.0001
SalesforceExperience High	10430.03959	B	852.593757	12.23	<.0001
SalesforceExperience Low	0.00000	B	.	.	.
Salesforc*Product_Ty Average 2	3854.12753	B	733.320136	5.26	<.0001
Salesforc*Product_Ty Average 3	4813.27082	B	854.097145	5.64	<.0001
Salesforc*Product_Ty Average 4	-12413.31701	B	991.233288	-12.52	<.0001
Salesforc*Product_Ty Average 1	0.00000	B	.	.	.
Salesforc*Product_Ty High 2	-56.25814	B	904.947211	-0.06	0.9504
Salesforc*Product_Ty High 3	4122.53992	B	971.542708	4.24	<.0001
Salesforc*Product_Ty High 4	-15715.77774	B	1150.053895	-13.67	<.0001
Salesforc*Product_Ty High 1	0.00000	B	.	.	.
Salesforc*Product_Ty Low 2	3103.23707	B	908.228513	3.42	0.0006
Salesforc*Product_Ty Low 3	4450.51766	B	974.481719	4.57	<.0001
Salesforc*Product_Ty Low 4	-11654.39823	B	1083.935242	-10.75	<.0001
Salesforc*Product_Ty Low 1	0.00000	B	.	.	.



8. Seasonality Trend

Imagine you **only** observe **Sales unit, price, Season, Design, and Quality variables**. Create a categorical variable **Product-Type** to code the 4-types product as in Q6.
You are interested in knowing (a) How far does the Seasonality increase the sales? (b) Also, you want to know do people have any preference to buy specific type of products in different seasons? In other words, do people buy more high quality or luxury products based on the current season? You should write precisely a regression model, estimate it, and explain the estimation results to identify the above questions. Can you explain a logical, verbal story about the seasonality effects which you find in your analysis? In other words, please try to justify your result based on a logical argument.

a) Seasonality effect can be seen in the glm model.
Summer season has the highest sales unit of 4231 unit more than the Spring Season.
Similarly, Winter and Fall season sells 1886 unit and 3466 unit more than the Spring season. */

b) Below are the season preferences with regards to the Product Type:

Winter: Product Type 3 (Low Design and High Quality)
Summer: Product Type 2 (Luxury Design and Low Quality)
Fall: Product Type 3 (Low Design and High Quality)
Spring: Product Type 3 (Low Design and High Quality)

We can tell in the Summer season is more inclined towards the Luxury Design products because in Summer season sells the maximum of Product Type=2 with
6626 unit more than product type 1.
As seen in the GLM model, the trend of selling a product is similar for Winter, Fall and Spring season since the Sales Unit is highest for Product Type =3 and lowest for
Product Type=4 in all these seasons. But the seasonality effect is visible in the summer season since the trend shifts from Product Type=3 to Product Type =2.
In summer customer preferred Luxury Design product when compared to High Quality product in all the other season.

Considering this as clothing brand, there can be lot of general stories which can proof our argument above:

- People during summer have more money to spend on clothing since they do not go for jackets and sweater which are costly. Since less clothing is required

during summer, they go for Luxury product

- Since the weather is sunny during summer, people spend a lot of time outside their home and hence prefer luxury design to showcase their extravagant life.

Preference for type of product based on Seasonality

The GLM Procedure

Dependent Variable: SalesUnit

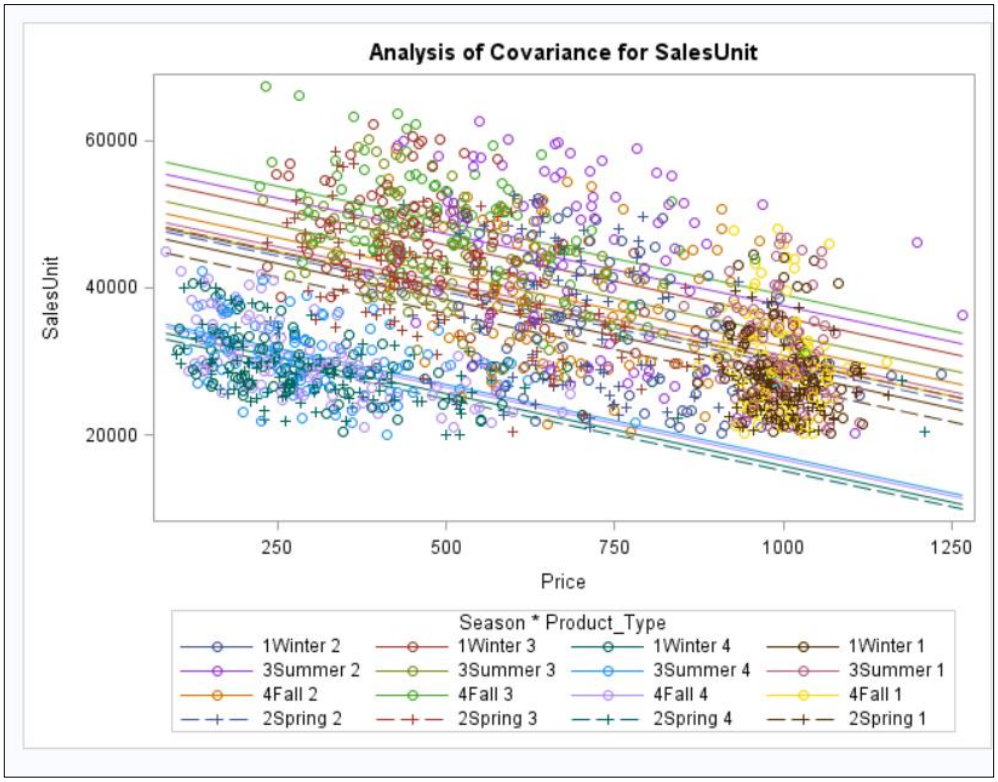
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	88400811892	5525050743.2	130.23	<.0001
Error	1583	67160018159	42425785.318		
Corrected Total	1599	155560830051			

R-Square	Coeff Var	Root MSE	SalesUnit Mean
0.568272	18.42438	6513.508	35352.65

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Price	1	10638182325	10638182325	250.75	<.0001
Season	3	5143869983	1714623328	40.41	<.0001
Season*Product_Type	12	72618759584	6051563299	142.64	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Price	1	11466349005	11466349005	270.27	<.0001
Season	3	5000419979	1666806660	39.29	<.0001
Season*Product_Type	12	72618759584	6051563299	142.64	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	46320.62202	B	1361.586553	34.02	<.0001
Price	-19.66503		1.196181	-16.44	<.0001
Season 1Winter	1886.13026	B	916.583080	2.06	0.0398
Season 3Summer	4231.77824	B	916.585947	4.62	<.0001
Season 4Fall	3466.36401	B	916.618023	3.78	0.0002
Season 2Spring	0.00000	B	.	.	.
Season*Product_Type 1Winter 2	1558.00167	B	990.348300	1.57	0.1159
Season*Product_Type 1Winter 3	7340.44679	B	1088.433559	6.74	<.0001
Season*Product_Type 1Winter 4	-12730.03197	B	1234.472901	-10.31	<.0001
Season*Product_Type 1Winter 1	0.00000	B	.	.	.
Season*Product_Type 3Summer 2	6626.45599	B	1007.424589	6.58	<.0001
Season*Product_Type 3Summer 3	2816.02612	B	1091.441041	2.58	0.0100
Season*Product_Type 3Summer 4	-13910.63601	B	1232.294649	-11.29	<.0001
Season*Product_Type 3Summer 1	0.00000	B	.	.	.
Season*Product_Type 4Fall 2	2005.91248	B	1017.743010	1.97	0.0489
Season*Product_Type 4Fall 3	8906.64371	B	1075.700118	8.28	<.0001
Season*Product_Type 4Fall 4	-13515.92407	B	1220.449327	-11.07	<.0001
Season*Product_Type 4Fall 1	0.00000	B	.	.	.
Season*Product_Type 2Spring 2	2847.39251	B	1012.425544	2.81	0.0050
Season*Product_Type 2Spring 3	3460.57586	B	1083.416353	3.19	0.0014
Season*Product_Type 2Spring 4	-11575.88267	B	1220.181125	-9.49	<.0001


Analysis of Covariance for SalesUnit
SalesUnit
60000
40000
20000
250
500
750
1000
1250
Price
Season * Product_Type
1Winter 2
1Winter 3
1Winter 4
1Winter 1
3Summer 2
3Summer 3
3Summer 4
3Summer 1
4Fall 2
4Fall 3
4Fall 4
4Fall 1
2Spring 2
2Spring 3
2Spring 4
2Spring 1

9. Regression Analysis

Imagine you **only** observe **Sales unit, price, and advertising spending variables**. Based on your answers in Q2 and Q3, provide a regression model's result to explain the Sales unit as a function of price and advertising spending. Based on Cook's D statistic, determine all influential points. You need to print the influential points in your report. Now, repeat your above regression analysis ***without*** including the influential points. Does it improve your model goodness-of-fit? If yes, how far? (Hint: follow the rule of thumb which is stated in class).

If the regression analysis is performed without influential points, the goodness-of-fit is increased. The Adjusted R-Sqr with influential point is 0.20, if we run regression analysis without the influential point, the R-Square value is increases to 0.2319 which shows us increase in goodness of fit.

Regression model with Influential Points

The REG Procedure
Model: MODEL1
Dependent Variable: SalesUnit

Number of Observations Read	1600
Number of Observations Used	1600

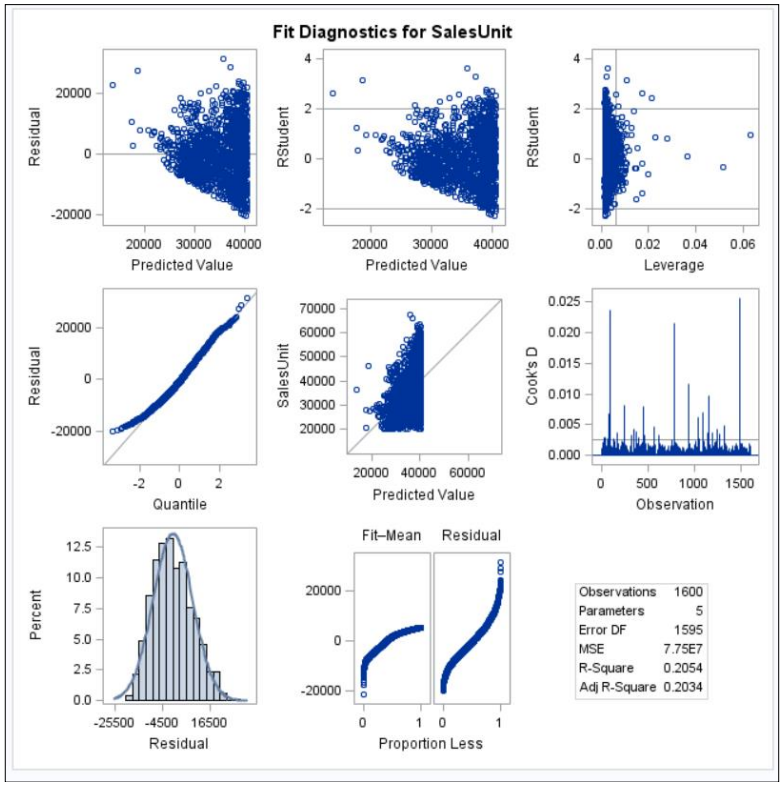
Analysis of Variance

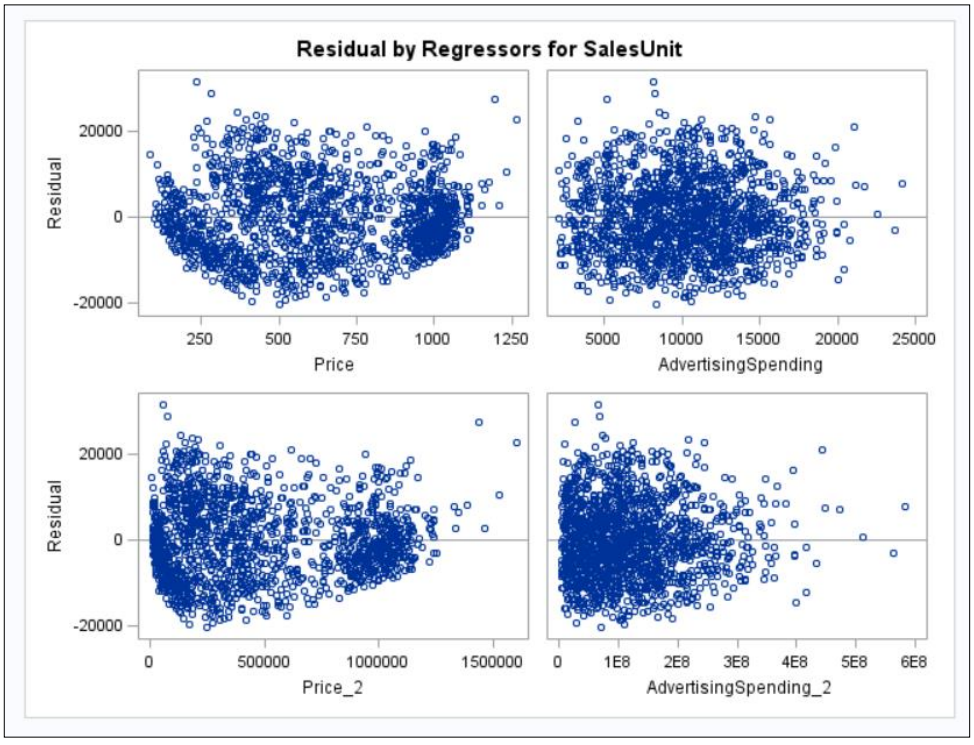
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	31957702409	7989425602	103.10	<.0001
Error	1595	1.236031E11	77494124		
Corrected Total	1599	1.555608E11			

Root MSE	8803.07469	R-Square	0.2054
Dependent Mean	35353	Adj R-Sq	0.2034
Coeff Var	24.90075		

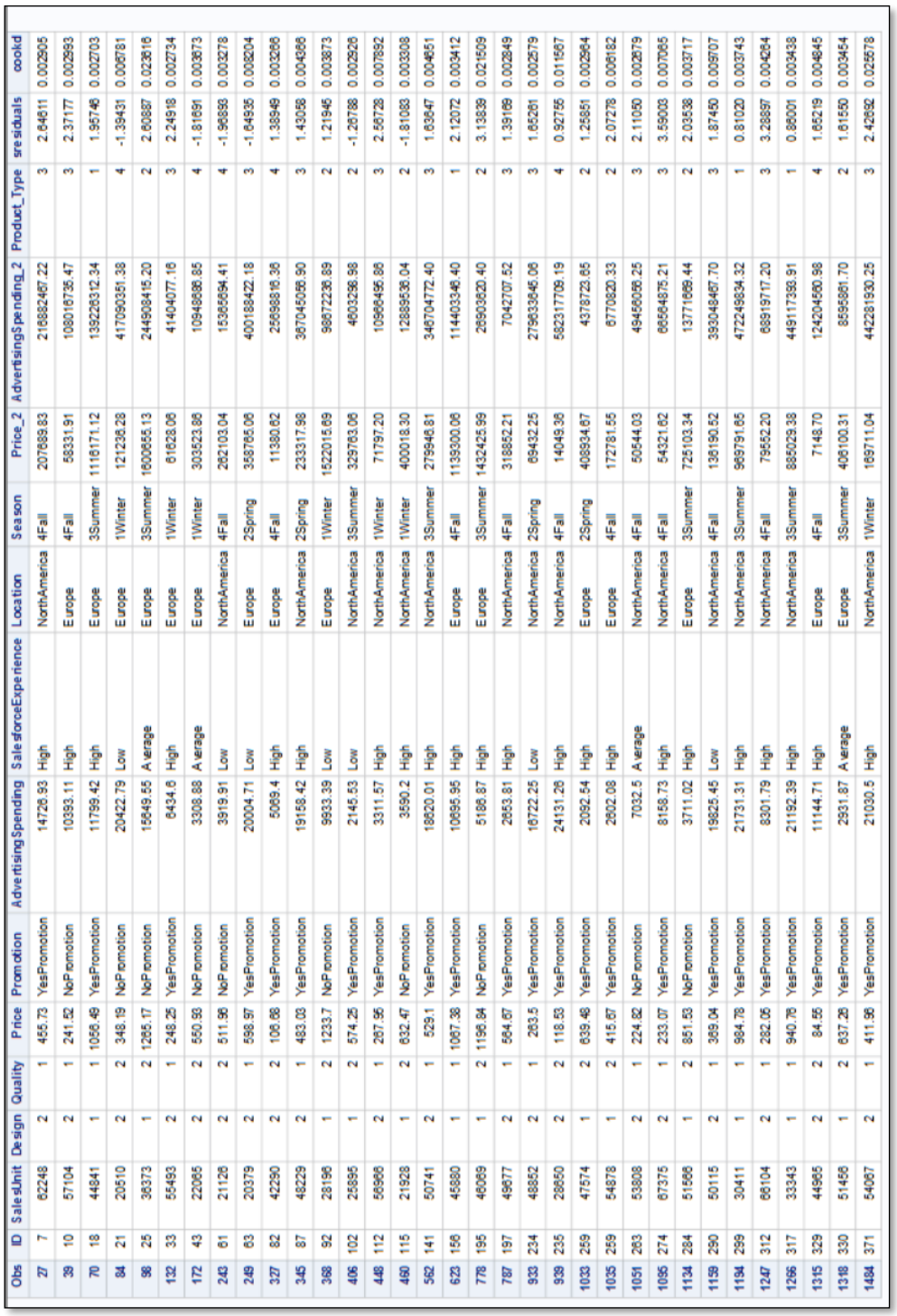
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	20808	1636.59342	12.71	<.0001
Price	1	52.01687	3.92730	13.24	<.0001
AdvertisingSpending	1	1.13490	0.24438	4.64	<.0001
Price_2	1	-0.04806	0.00304	-15.80	<.0001
AdvertisingSpending_2	1	-0.00005656	0.00001132	-5.00	<.0001


Fit Diagnostics for SalesUnit
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
SalesUnit
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 1600
Parameters 5
Error DF 1595
MSE 7.75E7
R-Square 0.2054
Adj R-Square 0.2034


Residual by Regressors for SalesUnit
Residual
Price
AdvertisingSpending
Residual
Price_2
AdvertisingSpending_2

Prints the observations that are influential (cook's d > 4/ n)



Obs	ID	SalesUnit	Design	Quality	Price	Promotion	Advertising	Spending	SalesforceExperience	Location	Season	Price_2	Advertising	Spending_2	Product_Type	sresiduals	cookd
27	7	62248	2	1	455.73	YesPromotion		14728.93	High	NorthAmerica	4Fall	207689.83		216882467.22	3	2.64611	0.002505
39	10	57104	2	1	241.52	NoPromotion		10393.11	High	Europe	4Fall	58331.91		108016735.47	3	2.37177	0.002593
70	18	44841	1	1	1056.49	YesPromotion		11799.42	High	Europe	3Summer	1116171.12		1392263312.34	1	1.95746	0.002703
84	21	20510	2	2	348.19	NoPromotion		20422.79	Low	Europe	1Winter	121236.28		417090351.38	4	-1.39431	0.006781
98	25	38373	1	2	1265.17	NoPromotion		15849.55	Average	Europe	3Summer	1600855.13		2449508415.20	2	2.60887	0.023616
132	33	55493	2	1	248.25	YesPromotion		6434.6	High	Europe	1Winter	61628.06		41404077.16	3	2.24918	0.002734
172	43	22065	2	2	550.39	NoPromotion		3308.88	Average	Europe	1Winter	303523.86		1094886.85	4	-1.81691	0.003673
243	61	21126	2	2	511.96	NoPromotion		3919.91	Low	NorthAmerica	4Fall	262103.04		1536594.41	4	-1.96893	0.003278
249	63	20379	2	1	598.97	YesPromotion		20004.71	Low	Europe	2Spring	358765.06		4001884422.18	3	-1.64935	0.008204
327	82	42290	2	2	106.88	YesPromotion		5089.4	High	Europe	4Fall	11380.82		25698816.36	4	1.38949	0.003266
345	87	48229	2	1	483.03	YesPromotion		19158.42	High	NorthAmerica	2Spring	233317.98		367045056.90	3	1.43068	0.004366
368	92	28196	1	2	1233.7	NoPromotion		9933.39	Low	Europe	1Winter	1522015.69		98872236.89	2	1.21945	0.003873
406	102	25895	1	2	574.25	YesPromotion		2145.53	Low	NorthAmerica	3Summer	329763.06		4603296.98	2	-1.26788	0.002926
448	112	56966	2	1	267.96	YesPromotion		3311.57	High	NorthAmerica	1Winter	71797.20		10966495.88	3	2.56728	0.007892
460	115	21928	1	2	832.47	NoPromotion		3590.2	High	NorthAmerica	1Winter	400018.30		12889536.04	2	-1.81083	0.003308
562	141	50741	2	1	529.1	YesPromotion		18820.01	High	NorthAmerica	3Summer	279946.81		3487047724.40	3	1.63647	0.004651
623	156	45880	1	1	1067.38	YesPromotion		10895.95	High	Europe	4Fall	1139300.06		1144033346.40	1	2.12072	0.003412
778	195	46089	1	2	1196.84	NoPromotion		5188.87	High	Europe	3Summer	1432425.99		28903820.40	2	3.13839	0.021509
787	197	49877	2	1	584.67	YesPromotion		2853.81	High	NorthAmerica	4Fall	318852.21		7042707.52	3	1.39169	0.002849
933	234	48852	2	1	283.5	YesPromotion		16722.25	Low	NorthAmerica	2Spring	69432.25		2796333846.06	3	1.65261	0.002579
939	235	28850	2	2	118.53	YesPromotion		24131.26	High	NorthAmerica	4Fall	14049.36		5823117709.19	4	0.92755	0.011567
1033	259	47574	1	2	639.46	YesPromotion		2092.54	High	Europe	2Spring	408934.67		4378723.85	2	1.25851	0.002564
1035	259	54878	1	2	415.67	YesPromotion		2802.08	High	Europe	4Fall	172781.55		6770820.33	2	2.07278	0.006182
1051	263	53808	2	1	224.82	NoPromotion		7032.5	Average	NorthAmerica	4Fall	50544.03		4946086.25	3	2.11050	0.002579
1095	274	67375	2	1	233.07	YesPromotion		8158.73	High	NorthAmerica	4Fall	54321.62		6658487.5.21	3	3.59003	0.007065
1134	284	51686	1	2	851.53	NoPromotion		3711.02	High	Europe	3Summer	725103.34		13771689.44	2	2.03538	0.003717
1159	290	50115	2	1	369.04	YesPromotion		19825.45	Low	NorthAmerica	4Fall	136190.52		3930484867.70	3	1.87450	0.009707
1194	299	30411	1	1	984.78	YesPromotion		21731.31	High	NorthAmerica	3Summer	969791.65		47224983.4.32	1	0.81020	0.003743
1247	312	66104	2	1	282.05	YesPromotion		8301.79	High	NorthAmerica	4Fall	79552.20		6891971.7.20	3	3.28897	0.004264
1266	317	33343	1	1	940.76	YesPromotion		21192.39	High	NorthAmerica	3Summer	885029.38		4491173.93.91	1	0.86001	0.003438
1315	329	44965	2	2	84.55	YesPromotion		11144.71	High	Europe	4Fall	7148.70		1242046.60.98	4	1.65219	0.004845
1318	330	51456	1	2	637.26	YesPromotion		2931.87	Average	Europe	3Summer	406100.31		8595861.70	2	1.61550	0.003454
1484	371	54067	2	1	411.96	YesPromotion		21030.5	High	NorthAmerica	1Winter	169711.04		4422819330.25	3	2.42692	0.025678

Regression model with Influential Points

```
Estimate the model without the influential points
```

Regression model without Influential Points

The REG Procedure
Model: MODEL1
Dependent Variable: SalesUnit

Number of Observations Read	1567
Number of Observations Used	1567

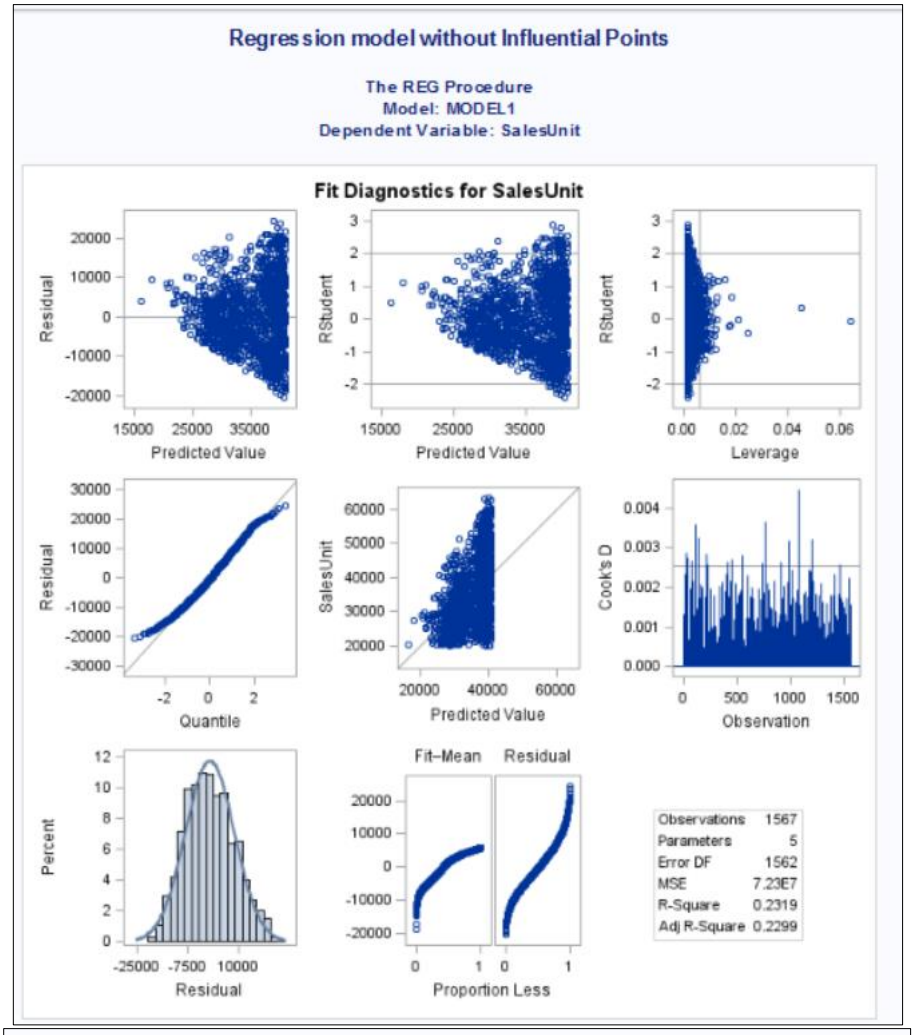
Analysis of Variance

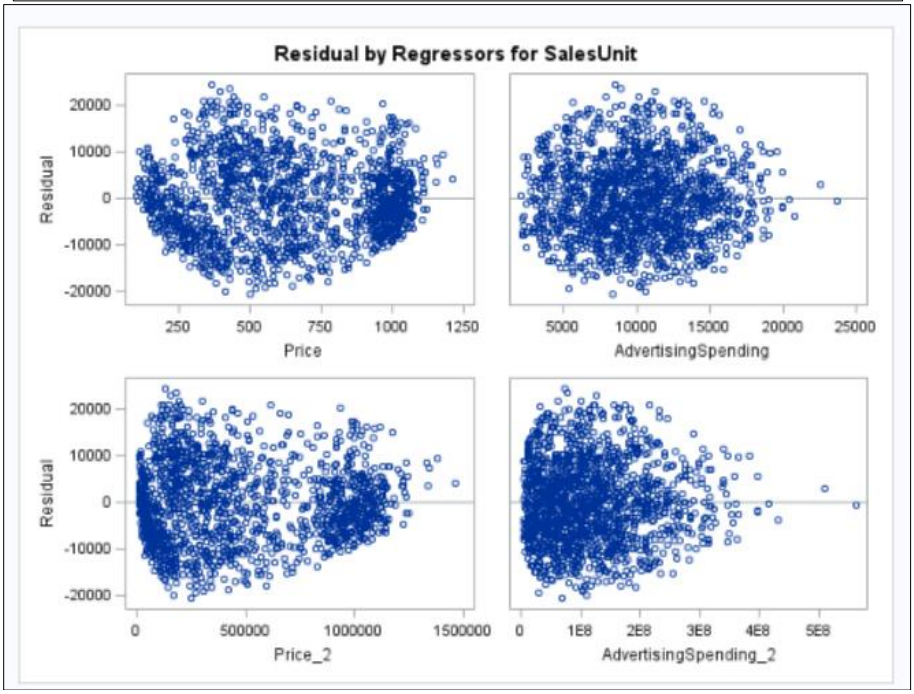
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	34100037684	8525009421	117.90	<.0001
Error	1562	1.129465E11	72308895		
Corrected Total	1566	1.470465E11			

Root MSE	8503.46369	R-Square	0.2319
Dependent Mean	35179	Adj R-Sq	0.2299
Coeff Var	24.17218		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	17886	1659.40776	10.78	<.0001
Price	1	57.79683	3.89788	14.83	<.0001
AdvertisingSpending	1	1.41080	0.25614	5.51	<.0001
Price_2	1	-0.05244	0.00302	-17.34	<.0001
AdvertisingSpending_2	1	-0.00007065	0.00001205	-5.86	<.0001


Regression model without Influential Points
The REG Procedure
Model: MODEL1
Dependent Variable: SalesUnit
Fit Diagnostics for SalesUnit
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
SalesUnit
Predicted Value
Cook's D
Observation
Percent
Residual
Fit-Mean
Residual
Proportion Less
Observations 1567
Parameters 5
Error DF 1562
MSE 7.23E7
R-Square 0.2319
Adj R-Square 0.2299


Residual by Regressors for SalesUnit
Residual
Price
AdvertisingSpending
Price_2
AdvertisingSpending_2

10. ANOVA

Imagine you **only** observe **Sales unit, price, and advertising spending variables**. Based on your answers in Q2 and Q3, provide a regression model's result to explain the Sales unit as a function of price and advertising spending. Now, check the collinearity of independent variables by reporting the VIF. If the VIF >10, drop some variables to reduce collinearity in an appropriate way. Does your new model fit the data better? If yes why? If no, why?

Guideline: For question 11-15, consider the whole dataset, i.e., all columns. Moreover, please add the power two of price and advertising spending as new predictors into dataset. Precisely, your dependent variable is **Sales unit** and the independent variables are **Price, Price^2, advertising spending, advertising spending^2, Product-Type, Promotion, Salesforce experience, Location**, **and Season**. Also, do the same procedure on **Sales_test.csv**

- VIF >10 if all the variable are added to the regression model i.e., Price, Price_2, Advertising Spending and AdvertisingSpending_2 and hence multicollinearity is present in the model.
- VIF <10 if we just include Price and Advertising Spend, and hence the multicollinearity is removed from the model.

No, the model does not fit better since removing variable will decrease the R square value, a greater number of variables will lead to high R Square value. But, it will lead to true model and standard error will not be large.

Since, the price and advertising spend shows a non linear trend in the data, hence this model shows decreased R Square value.

Regression model without Influential Points

The REG Procedure
Model: MODEL1
Dependent Variable: SalesUnit

Number of Observations Read	1600
Number of Observations Used	1600

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	31957702409	7989425602	103.10	<.0001
Error	1595	1.236031E11	77494124		
Corrected Total	1599	1.555608E11			

Root MSE	8803.07469	R-Square	0.2054
Dependent Mean	35353	Adj R-Sq	0.2034
Coeff Var	24.90075		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	20808	1636.59342	12.71	<.0001	0
Price	1	52.01687	3.92730	13.24	<.0001	26.85647
Price_2	1	-0.04806	0.00304	-15.80	<.0001	26.85587
AdvertisingSpending	1	1.13490	0.24438	4.64	<.0001	18.48992
AdvertisingSpending_2	1	-0.00005656	0.00001132	-5.00	<.0001	18.49008

Collinearity Diagnostics (intercept adjusted)

Number	Eigenvalue	Condition Index	Proportion of Variation			
			Price	Price_2	AdvertisingSpending	AdvertisingSpending_2
1	1.98942	1.00000	0.00630	0.00628	0.00445	0.00446
2	1.96437	1.00636	0.00310	0.00312	0.00926	0.00924
3	0.02742	8.51822	0.00000222	4.826152E-7	0.98628	0.98629
4	0.01879	10.28846	0.99060	0.99060	0.00000776	1.54124E-7

Regression model without Influential Points

The REG Procedure
Model: MODEL1
Dependent Variable: SalesUnit

Number of Observations Read	1600
Number of Observations Used	1600

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	10685729695	5342864848	58.90	<.0001
Error	1597	1.448751E11	90717032		
Corrected Total	1599	1.555608E11			

Root MSE	9524.54892	R-Square	0.0687
Dependent Mean	35353	Adj R-Sq	0.0675
Coeff Var	26.94154		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	41332	835.92292	49.44	<.0001	0
Price	1	-8.87530	0.81995	-10.82	<.0001	1.00004
AdvertisingSpending	1	-0.04452	0.06149	-0.72	0.4692	1.00004

Collinearity Diagnostics (intercept adjusted)

Number	Eigenvalue	Condition Index	Proportion of Variation	
			Price	AdvertisingSpending
1	1.00639	1.00000	0.49680	0.49680
2	0.99361	1.00641	0.50320	0.50320

Regression model without Influential Points

The REG Procedure
Model: MODEL1
Dependent Variable: SalesUnit

Number of Observations Read	1600
Number of Observations Used	1600

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	16438669131	8219334566	94.35	<.0001
Error	1597	1.391222E11	87114691		
Corrected Total	1599	1.555608E11			

Root MSE	9333.52509	R-Square	0.1057
Dependent Mean	35353	Adj R-Sq	0.1046
Coeff Var	26.40120		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	39841	713.55987	55.83	<.0001	0
Price_2	1	-0.00853	0.00062240	-13.71	<.0001	1.00002
AdvertisingSpending	1	-0.04511	0.06026	-0.75	0.4542	1.00002

Collinearity Diagnostics (intercept adjusted)

Number	Eigenvalue	Condition Index	Proportion of Variation	
			Price_2	AdvertisingSpending
1	1.00443	1.00000	0.49778	0.49778
2	0.99557	1.00444	0.50222	0.50222

Model: MODEL1
Dependent Variable: SalesUnit

Number of Observations Read	1600
Number of Observations Used	1600

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	30021691666	10007230555	127.22	<.0001
Error	1596	1.255391E11	78658608		
Corrected Total	1599	1.555608E11			

Root MSE	8868.96882	R-Square	0.1930
Dependent Mean	35353	Adj R-Sq	0.1915
Coeff Var	25.08714		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	26200	1239.91703	21.13	<.0001	0
Price	1	51.99452	3.95669	13.14	<.0001	26.85644
Price_2	1	-0.04805	0.00306	-15.68	<.0001	26.85587
AdvertisingSpending	1	-0.05308	0.05726	-0.93	0.3541	1.00013

Collinearity Diagnostics (intercept adjusted)

Number	Eigenvalue	Condition Index	Proportion of Variation		
			Price	Price_2	AdvertisingSpending
1	1.98126	1.00000	0.00940	0.00940	0.00003068
2	0.99994	1.40761	7.589408E-7	0.00000158	0.99986
3	0.01879	10.26734	0.99060	0.99060	0.00010620

Model: MODEL1
Dependent Variable: SalesUnit

Number of Observations Read	1600
Number of Observations Used	1600

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12618978697	4206326232	46.97	<.0001
Error	1596	1.429419E11	89562564		
Corrected Total	1599	1.555608E11			

Root MSE	9463.74997	R-Square	0.0811
Dependent Mean	35353	Adj R-Sq	0.0794
Coeff Var	26.76956		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	35945	1426.31779	25.20	<.0001	0
Price	1	-8.85732	0.81473	-10.87	<.0001	1.00006
AdvertisingSpending	1	1.14261	0.26272	4.35	<.0001	18.48984
AdvertisingSpending_2	1	-0.00005652	0.00001217	-4.65	<.0001	18.49008

Collinearity Diagnostics (intercept adjusted)

Number	Eigenvalue	Condition Index	Proportion of Variation		
			Price	AdvertisingSpending	AdvertisingSpending_2
1	1.97268	1.00000	0.00005036	0.01371	0.01371
2	0.99990	1.40459	0.99993	0.00000306	0.00000233
3	0.02742	8.48230	0.00001667	0.98629	0.98629

11. Analysis of Variance and Regression Model

By using Proc Glmmod, you should create two new datasets out of **Sales.csv** and **Sales_test.csv** which include all the above quantitative variables and all dummy variables of all the above qualitative variables. By using Proc Reg, find the best regression model on the above generated data out of **Sales.csv**. you must use the Mallows' C_p criterion for selection procedure. Now, by considering the above best regression model, record its R^2 and calculate the MSE_{test} by using the above generated data out of **Sales_test.csv** (Hint: You can use Proc score)

Now, by using Proc Glmselect, find the best models according to Forward, Backward, Stepwise algorithm. In all cases, (1) - you must use the Mallows' C_p criterion for selection procedure, (2) adding all interaction effects at the second order, and (3) in a hierarchal order, when the main effects should be included first. Meanwhile, find and record their R^2 and the MSE_{test} of the best selected models by the above procedures (i.e., Forward, Backward, Stepwise algorithm) by using **Sales_test.csv** as the test dataset.

Now, please rank all the above recorded best models based on either their R^2 or the MSE_{test} of them. If your goal is to do inference about the Sales unit based on independent variables which one of the above best models will be your choice? If your goal is to do predictive analysis about the Sales unit based on independent variables which one of the above best models will be your choice?

Algorithm	R Square	Adjusted R	MSE(test)
Forward	0.9495	0.9473	81605603
Backward	0.9501	0.9478	81930426
Stepwise	0.9491	0.947	81489109

The above table gives a summary statistic while using all the different algorithm.
For Inferences about Sales Unit we will be using our Adjusted R square value as the judging parameter. The more the Adj R square value, the better the model will be. So, for Inference - Backward Algorithm gives us the best model.

For Prediction about Sales Unit we will be using our MSE test value as the judging parameter. The less the MSE test value, the better the model will be.
So, for Prediction - Stepwise Algorithm gives us the best model.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	71	1.478051E11	2081762061	410.14	<.0001
Error	1528	7755723701	5075735		
Corrected Total	1599	1.555608E11			

Root MSE	2252.93928
Dependent Mean	35353
R-Square	0.9501
Adj R-Sq	0.9478
AIC	26376
AICC	26383
BIC	24784
C(p)	65.76523
SBC	25161
ASE (Train)	4847327
ASE (Test)	81930426

Best Regression model using Stepwise Algorithm

The GLMSELECT Procedure

Data Set	WORK.SALES
Test Data Set	WORK.SALES_TEST
Dependent Variable	SalesUnit
Selection Method	Stepwise
Select Criterion	C(p)
Stop Criterion	C(p)
Effect Hierarchy Enforced	Single

Observation Profile for Analysis Data	
Number of Observations Read	1600
Number of Observations Used	1600
Number of Observations Used for Training	1600

Observation Profile for Test Data	
Number of Observations Read	400
Number of Observations Used	400

Class Level Information

Class	Levels		Values
Design	2	*	1 2
Quality	2	*	1 2
Promotion	2	*	NoPromotion YesPromotion
SalesforceExperience	3	*	Average High Low
Location	2	*	Europe NorthAmerica
Season	4	*	1Winter 2Spring 3Summer 4Fall

* Associated Parameters Split

Dimensions	
Number of Effects	56
Number of Effects after Splits	178
Number of Parameters	178

12. Best Model selection

Now, by using Proc Glmselect, find the best models according to Forward, Backward, Stepwise algorithm. In all cases, (1) - you must use the Cross-Validation criterion for selection procedure based on creating 10-folds of **Sales.csv** dataset (please set the seed=2), (2) adding all interaction effects at the second order, and (3) in a hierarchal order, when the main effects should be included first. Meanwhile, find and record their R^2 and the MSE_{test} of the best selected models by the above procedures (i.e., Forward, Backward, Stepwise algorithm) by using **Sales_test.csv** as the test dataset.

Now, please rank all the above recorded best models based on either their R^2 or the MSE_{test} of them in Q11 and Q12. If your goal is to do inference about the Sales unit based on independent variables which one of the above best models will be your choice? If your goal is to do predictive analysis about the Sales unit based on independent variables which one of the above best models will be your choice?

Algorithm	R Square	Adjusted R	MSE(test)
Forward	0.9341	0.9317	85222988
Backward	0.9563	0.9535	83083861
Stepwise	0.9486	0.9467	81497787

The above table gives a summary statistic while using all the different algorithm.
For Inferences about Sales Unit we will be using our Adjusted R square value as the judging parameter. The more the Adj R square value, the better the model will be.
So, for Inference - Backward Algorithm gives us the best model.

For Prediction about Sales Unit we will be using our MSE test value as the judging parameter. The less the MSE test value, the better the model will be.
So, for Prediction - Stepwise Algorithm gives us the best model.

13. Best Model selection

Now, by using Proc Glmselect, find the best models according to LASSO and Elastic Net. In all cases, (1) - you must use the Cross-Validation criterion for selection procedure based on creating 10-folds of **Sales.csv** dataset (please set the seed=2), (2) adding all interaction effects at the second order, and (3) in a hierarchal order, when the main effects should be included first. Meanwhile, find and record their R^2 and the MSE_{test} of the best selected models by the above procedures (i.e., LASSO and Elastic Net) by using **Sales_test.csv** as the test dataset.

If your goal is to do inference about the Sales unit based on independent variables which one of the above best models will be your choice? If your goal is to do predictive analysis about the Sales unit based on independent variables which one of the above best models will be your choice?

Algorithm	R Square	Adjusted R	MSE(test)
LASSO	0.9481	0.9456	81217832
Elastic Net	0.9469	0.9444	81083439

The above table gives a summary statistic while using all the different algorithm (Elastic Net & LASSO).
For Inferences about Sales Unit we will be using our Adjusted R square value as the judging parameter. The more the Adj R square value, the better the model will be. So, for Inference - LASSO Algorithm gives us the best model.

For Prediction about Sales Unit we will be using our MSE test value as the judging parameter. The less the MSE test value, the better the model will be. So, for Prediction – Elastic Net Algorithm gives us the best model.

Best Regression model using Lasso

The GLMSELECT Procedure

Data Set	WORK.SALES
Test Data Set	WORK.SALES_TEST
Dependent Variable	SalesUnit
Selection Method	LASSO
Stop Criterion	None
Choose Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	10
Effect Hierarchy Enforced	None
Random Number Seed	2

Observation Profile for Analysis Data	
Number of Observations Read	1600
Number of Observations Used	1600
Number of Observations Used for Training	1600

Observation Profile for Test Data	
Number of Observations Read	400
Number of Observations Used	400

Class Level Information		
Class	Levels	Values
Design	2	1 2
Quality	2	1 2
Promotion	2	NoPromotion YesPromotion
SalesforceExperience	3	Average High Low
Location	2	Europe NorthAmerica
Season	4	1Winter 2Spring 3Summer 4Fall

Dimensions	
Number of Effects	56
Number of Effects after Splits	178
Number of Parameters	178

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	71	1.473026E11	2074684577	383.87	<.0001
Error	1528	8258225059	5404598		
Corrected Total	1599	1.555608E11			

Root MSE	2324.77903
Dependent Mean	35353
R-Square	0.9469
Adj R-Sq	0.9444
AIC	26477
AICC	26484
SBC	25262
ASE (Train)	5161391
ASE (Test)	81083439
CV PRESS	8934368140

14. Bagging Method

Now, by using Proc Glmselect, do the regression analysis based on Bagging Method. In all cases, (1) - you must use the Cross-Validation criterion for selection procedure based on creating 10-folds of Sales.csv dataset (please set the seed=2), (2) adding all interaction effects at the second order, (3) in a hierarchal order, when the main effects should be included first, and use 100 samples (based on Bootstrapping). Based on the Average Model of 100 models, find and record the MSE_{train} and the MSE_{test} by using Sales.csv and Sales_test.csv as the training and test datasets respectively.

If your goal is to do inference about the Sales unit based on independent variables does Bagging Method provide you any advantage? If your goal is to do predictive analysis about the Sales unit based on independent variables does Bagging Method provide you any advantage?

Bagging method is helpful for doing inference since it decreases the bias and variance of the data set.
But Bagging method provides advantage in Predictive Analysis since we are taking an average of 100 samples model to come up with the final recommendation.

Regression using Bagging Method

The GLMSELECT Procedure

Data Set	WORK.SALES
Test Data Set	WORK.SALES_TEST
Dependent Variable	SalesUnit
Selection Method	Stepwise
Select Criterion	C(p)
Stop Criterion	C(p)
Effect Hierarchy Enforced	Single

Model Averaging Information	
Sampling Method	Unrestricted (with replacement)
Sample Percentage	100
Number of Samples	100

Observation Profile for Analysis Data	
Number of Observations Read	1600
Number of Observations Used	1600
Number of Observations Used for Training	1600

Observation Profile for Test Data	
Number of Observations Read	400
Number of Observations Used	400

Class Level Information			
Class	Levels		Values
Design	2	*	1 2
Quality	2	*	1 2
Promotion	2	*	NoPromotion YesPromotion
SalesforceExperience	3	*	Average High Low
Location	2	*	Europe NorthAmerica
Season	4	*	1Winter 2Spring 3Summer 4Fall
* Associated Parameters Split			

Dimensions	
Number of Effects	56
Number of Effects after Splits	178
Number of Parameters	178

Regression using Bagging Method

The GLMSELECT Procedure
Selected Model

Information for Score Statement 1	
Input Data Set	WORK.SALES_TEST
Output Data Set	WORK.TEST_PERFORMANCE
Number of Observations Read	400
Number of Observations Scored	400
Number of Residuals Computed	400
Residual Sum of Squares	32690623604

Information for Score Statement 2	
Input Data Set	WORK.SALES
Output Data Set	WORK.INSAMPLE_PERFORMANCE
Number of Observations Read	1600
Number of Observations Scored	1600
Number of Residuals Computed	1600

15. Contrafactual Analysis

This firm wants to increase its sales. They have two options. The first way is to train their sales-force people to become more expert. The second way is to increase their advertising budget. The firm told you that they can increase their advertising budget by 5%. Also, they know, by having a one-day workshop for "Low" and "Average"-types of sales-force people, they will become "Average" and "High"-types sales-force respectively.

Below is the bagging model for doing the contrafactual analysis:

The GLMSELECT Procedure

Data Set	WORK.SALES
Dependent Variable	SalesUnit
Selection Method	Stepwise
Select Criterion	Cross Validation
Stop Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	10
Effect Hierarchy Enforced	Single
Random Number Seed	2

Model Averaging Information	
Sampling Method	Unrestricted (with replacement)
Sample Percentage	100
Number of Samples	100

Number of Observations Read	1600
Number of Observations Used	1600

Class Level Information			
Class	Levels		Values
Product_Type	4	*	1 2 3 4
Promotion	2	*	NoPromotion YesPromotion
SalesforceExperience	3	*	Average High Low
Location	2	*	Europe NorthAmerica
Season	4	*	1Winter 2Spring 3Summer 4Fall
* Associated Parameters Split			

Dimensions	
Number of Effects	46
Number of Effects after Splits	174
Number of Parameters	174

Which strategy will increase the sales more?

Increasing the advertising spend strategy works better for increasing the sales unit instead of changing based on Product Type. As we can see below, the average difference for Sales Force experience is 3341 and for advertising spend mean is -111.995. Since, Advertising Spend is close to zero, it is better strategy.

Means for Performance by Sales Force Experience

The MEANS Procedure

Analysis Variable : diff				
N	Mean	Std Dev	Minimum	Maximum
1600	3341.81	3548.34	-7503.31	12883.76

Means for Performance by Advertising Spending

The MEANS Procedure

Analysis Variable : diff				
N	Mean	Std Dev	Minimum	Maximum
1600	-111.9858842	2330.75	-8772.43	7159.45

What is the average sales unit changes by Product-Type in each strategy? Which strategy is more profitable from your point of view based on your information about this firm?

Seeing the mean below by Product Type, with Sales force experience, product Type=2 is close to zero.
Seeing the mean below by Product Type, with Advertising Spend, product Type=2 is close to zero.

Means for Performance by Sales Force Experience

The MEANS Procedure

Product_Type=1

Analysis Variable : diff				
N	Mean	Std Dev	Minimum	Maximum
404	4747.30	4346.29	-4796.10	12883.76

Product_Type=2

Analysis Variable : diff				
N	Mean	Std Dev	Minimum	Maximum
364	2770.42	3198.29	-6465.35	8943.78

Product_Type=3

Analysis Variable : diff				
N	Mean	Std Dev	Minimum	Maximum
420	3102.69	3280.93	-4505.35	10487.10

Product_Type=4

Analysis Variable : diff				
N	Mean	Std Dev	Minimum	Maximum
412	2712.19	2786.87	-7503.31	9707.60

Means for Performance by Advertising Spending

The MEANS Procedure

Product_Type=1

Analysis Variable : diff				
N	Mean	Std Dev	Minimum	Maximum
404	-108.6904015	2678.47	-8691.02	6240.93

Product_Type=2

Analysis Variable : diff				
N	Mean	Std Dev	Minimum	Maximum
364	-87.8067838	2336.17	-8772.43	6789.95

Product_Type=3

Analysis Variable : diff				
N	Mean	Std Dev	Minimum	Maximum
420	-121.5490278	2222.25	-7211.83	7159.45

Product_Type=4

Analysis Variable : diff				
N	Mean	Std Dev	Minimum	Maximum
412	-126.8306590	2060.26	-7667.36	5121.93

16. Summary

Provide a paragraph which summarizes your analyses in this homework (or any other important trend you find in this data). For example: Which variables have positive effects on the Sales Unit? Which variables have negative effects on the Sales Unit? Does exist a systematic (i.e., significant) difference between Europeans and Americans? If Yes, what are they? Who is more price sensitive? How should be targeting strategy of this firm, i.e., which product should be targeted and to whom? What should be the strategy on Salesforce training to increase the sales? And …

Guideline: For question 16, you should provide a professional, managerial paragraph to summarize your analyses. It must be short, informative, and inclusive. ***The last sentence must provide a clear strategy for this firm***. Which product should be targeted in each market & what should be the price level in each market? What can be done to increase the sales? Question 16 will be marked very competitively. The best answer will receive the full mark. We will decrease marks according to your rank among groups.

Summary Report

- variables Positively Related to Sales Unit: Advertising spending, consumer response to promotion, popularity of product, salesforce experience
- Variables Negatively Related to Sales Unit: Price
- Price Sensitivity: North America is more price sensitive compared to Europe. North America also seems more responsive towards promotional offers compared to Europe.
- Price Difference Across Products: High quality and luxury design both increase average product price. The product with normal quality and luxury design has a higher average price than the product with high quality normal design.
- Experience: With higher salesforce experience, we see higher sales. This trend is even more significant for products with high quality.
- Product Preference: Throughout most of the year, normal design and high quality is preferred, but in summer people prefer the luxury design and tend spend more.
- Conclusion: Based on our analysis we recommend that the firm produces the product with normal design and high quality. Promotional efforts could be focused on the American market, as they will be more influential in American stores. Lastly, we recommend investing in training for salespeople as this is likely to increase sales.

Codes

```
LIBNAME HW2 'H:\My SAS Files\HW2';

/* This imports the csv dataset into SAS. */
/* You can do it by using the "Import Data" option in File on the main menu */

PROC IMPORT OUT= HW2.sales
            DATAFILE= "H:\My SAS Files\HW2\Sales.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

PROC IMPORT OUT= HW2.sales_test
            DATAFILE= "H:\My SAS Files\HW2\Sales_test.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;


/* generating the working dataset in Work library */
data sales;
 set HW2.sales;
run;


data sales_test;
 set HW2.sales_test;
run;



/* 1. Use Proc sgscatter to plot a matrix of Sales unit, price, and advertising spending.
What type of relationship do you see between Sales unit and price? (Linear or Non-linear)
What type of relationship do you see between Sales unit and advertising spending? (Linear
or Non-linear)?    */

proc sgscatter data= sales;
 matrix SalesUnit Price AdvertisingSpending/ diagonal= (histogram);
 title 'Relation between SalesUnit Price Advertising Spending';
run;


/* SalesUnit vs Price : It is showing a negative correlation with the data.
But it is not clear from the graph about the linear or non-linear relationship.

SalesUnit vs AdvertisingSpending : It is showing a positive correlation with the data.
But it is not clear from the graph about the linear or non-linear relationship. */


/* Ques 2: Now, provide a linear regression of Sales unit over price? (Consider only
price) Is the price coefficient significant?
```

```
Does price's coefficient make sense? What does it mean?
Now, provide a non-linear regression of Sales unit over price and its power 2?
Are the price and its power 2's coefficients significant?
Explain your result in terms of consumers' reaction to increase of price in this market.
Compare the R^2 of the above models (linear vs. non-linear). Which model fits the data
better and why? */

proc reg data = sales;
 model SalesUnit =  Price;
 title 'Linear Regression of Sales Unit over Price';
run;

data sales;
 set sales;
 Price_2 = Price * Price;
run;
ods graphics on;
/*Spending Regression Model --> Regression of Spending over income and its power 2 */
proc reg data = sales;
 model SalesUnit =  Price Price_2;
 title 'Regression Analysis of Sales Unit over Price Price_2';
run;
ods graphics off;

/* The Linear model coefficient Price is highly significant since Pr value is < 0.0001.
The coefficient of the Price is -8.11 which means that if 1 unit of Price is increased it
will lead to decrease of 8.11 Sales Unit.
This does not make sense since with increase in Price, Sales Unit should increase. */

/*  Price and Price power 2's coefficient are significant in this model with Pr value <
0.0001 for both of them.
This model tells a diffrent story when compared to the linear regression :
The intercept for Price is 51.9 and Price Square is -0.04 which tells us that the
Consumer Reaction with increase in Price lets the SalesUnit increase intitally but after
a certain threshold
it starts to decrese.
The R squared value for linear model is 0.0684(Adj R sqrd = 0.0678) and for non linear
model it is 0.1926(Adj sqrd = 0.1915).
We can see from the R sqrd values the non linear model fits better and so we can say that
SalesUnit and Price are non linearly related */



/* Ques 3: Now, provide a linear regression of Sales unit over advertising spending?
(Consider only advertising spending) Is the advertising spending coefficient significant?
Does advertising's coefficient make sense? If No, why?
Now, provide a non-linear regression of Sales unit over advertising spending and its
power 2?
Are the coefficients significant? Explain your result in terms of Sales unit changes by
increasing of advertising in this market.
Compare the R^2 of the above models (linear vs. non-linear). Which model fits the data
better and why? */


proc reg data = HW2.sales;
 model SalesUnit =  AdvertisingSpending;
 title 'Linear Regression of Sales Unit over AdvertisingSpending';
run;
```

```
data sales;
 set sales;
 AdvertisingSpending_2 = AdvertisingSpending * AdvertisingSpending;
run;
ods graphics on;
/*Spending Regression Model --> Regression of Spending over income and its power 2 */
proc reg data = sales;
 model SalesUnit =  AdvertisingSpending AdvertisingSpending_2;
 title 'Regression Analysis of Sales Unit over AdvertisingSpending
AdvertisingSpending_2';
run;
ods graphics off;


/* The Linear model coefficient Advertising Spending is not significant since Pr value is
0.4439 and greater than 0.001.
The coefficient of the Price is -0.04877 which means that if 1 unit of Advertsing Spend
is increased it will lead to decrease of -0.04877 Sales Unit.
This does not make sense since with increase in Advertising Spending, Sales Unit should
increase. */

/*  Advertising Spending and Advertising Spending power 2's coefficient are significant
in this model with Pr value < 0.0001 for both of them.
This model tells a diffrent story when compared to the linear regression :
The intercept for Advertising Spending is 1.15 and Price Square is -0.00005715 which
tells us that the with increase in Advertising Spending lets the SalesUnit increase
intitally but after a certain threshold
it starts to decrese.
The R squared value for linear model is 0.0004(Adj R sqrd = -0.0003) and for non linear
model it is 0.0131(Adj sqrd = 0.0118).
We can see from the R sqrd values the non linear model fits better and so we can say that
SalesUnit and Advertising Spending are non linearly related */


/* Ques 4: Imagine you only observe Sales unit, price, and location variables.
You are interested in knowing who is more price sensitive? North Americans or European.
Propose a regression model to find the most price sensitive group of consumers. (Hint:
Use the Interaction effect).
You should write your model precisely, estimate it, and explain the estimation results to
identify the most price sensitive group of consumers. */



proc glm data=Sales;
 class Location(ref='Europe');
 model SalesUnit = Price Price_2 Location Price*Location/ solution;
 title 'Price Sensitivity Analyis: North America vs European' ;
run;

/* North American are more Price Sensitive when compared to European. The sales unit
decrease by 8.199 unit if the price is increased by 1 unit for North American
when compared to Europe. */


/* Ques 5:  Imagine you only observe Sales unit, location, and promotion variables.
You are interested in knowing (a) do consumers respond to promotional prices? and
(b) who is more responsive to promotional offers? North Americans or European.
Propose a regression model to answer the above questions. (Hint: Use the Interaction
effect).
```

```
You should write your model precisely, estimate it, and explain the estimation results to
answer the above questions.
Are your results compatible with your answer in Q4?  */

proc glm data = Sales;
class Promotion(ref = 'NoPromotion');
 model SalesUnit =  Promotion / solution;
 title 'Price Sensitivity Analyis based on Promotions' ;
run;

/* a) Yes, consumer reponse to the promotions has a positve correlation with the Sales
Unit as seen by the glm model of Sales Unit vs Promotion. */


proc glm data = Sales;
class Location(ref='Europe') Promotion(ref = 'NoPromotion');
 model SalesUnit =  Promotion Location Promotion*Location/ solution;
 title 'Price Sensitivity Analyis in Location based on Promotions' ;
run;

/* b) As seen in the glm model, North American seems more responsive towards promotional
offers. If a promotion is there the Sales Unit increases
by 3400 unit for North American when compared to Europe. Yes, the result are found
compatible with the Question 4 */



/*Ques 6.   Imagine you only observe Sales unit, price, location, Design, and Quality
variables.
We know the firms offer 4-types of product. Define a new categorical variable called
Product-Type to code the 4-types of firm's products.
Define the Product-Type as follows: */


data Sales;
 set Sales;
 if Design = 1 & Quality = 1 then Product_Type = 1;
 if Design = 1 & Quality = 2 then Product_Type = 2;
 if Design = 2 & Quality = 1 then Product_Type = 3;
 if Design = 2 & Quality = 2 then Product_Type = 4;
run;

/*First, provide a hypothesis testing to find: is there any significant difference among
the average prices of these 4-types of product? (Hint ANOVA)
If yes, RANK the types of product based on their average price.
Which attribute of product does increase the price of a type of product more
significantly, Design or Quality? */

proc anova data= Sales;
      class Product_Type;
      model Price  = Product_Type;
      means Product_Type;
      title 'Hypothesis Testing for Product Type vs Average Price' ;
run;

/* If the Design = 1 or Luxury Design, the average price of those product is higher when
compared to product with Quality. */

/* (a) which type of product is the least popular among consumers?  */
```

```
proc anova data= Sales;
      class Product_Type;
      model SalesUnit  = Product_Type;
      means Product_Type;
      title 'Hypothesis Testing for Product Type vs Average Sales Unit' ;
run;


/*The popularity of the product can be determined using the Sales Unit. As seen, mean
sales unit for the Product Type =1 (Luxury Design and High Quality)
is the lowest and hence it is the least popular product among consumer. */

/* (b) Also, you want to know do North Americans vs. Europeans have different preference
for Design and Quality?
In other words, is the most important attributes of product (i.e., Design vs Quality)
different among North Americans vs. Europeans?
You should write precisely a regression model, estimate it, and explain the estimation
results to identify the above questions.
Are your results consistent with your results in Q4 and Q5? */

proc glm data = Sales;
class Location(ref='Europe') Product_Type(ref = '1');
 model SalesUnit =  Price Location Location*Product_Type/ solution;
 title 'Preference for type of product' ;
run;


/* Yes, as seen in glm model, North American preferred product type=3(Low Design and High
Quality) and European preffered product type=2(Luxury Design and Low Quality).
/* Yes, the result are consistent with the Q.2 and Q.5, which gave us a picture that
North American are more price sensitive and salesunit increases with promotion for
American when compared to European */

/* 7. Imagine you only observe Sales unit, price, Salesforce experience, Design, and
Quality variables.
Create a categorical variable Product-Type to code the 4-types product as in Q6.

a) How far does the Salesforce experience increase the sales?
(b) Also, you want to know do high experience Salesforce have any expertise to sell
specific type of products?
In other words, do their experience help them to sell more high-ended products?
You should write precisely a regression model, estimate it, and explain the estimation
results to identify the above */

proc glm data = Sales;
class SalesforceExperience(ref='Low') Product_Type(ref = '1');
 model SalesUnit =  Price SalesforceExperience Product_Type*SalesforceExperience/
solution;
 title 'Preference for type of product based on SalesForce Experience' ;
run;

/* a) There is a high correlation between the Salesforce experiecnce and Sales Unit. As
seen in the glm model,
coefficient of High Salesforce experience is 10430.034, which suggest that with High
Experience the Sales Unit increases by 10430 unit
coefficient of Average Salesforce experience is 10430.034, which suggest that with
Average Experience the Sales Unit increases by 10430 unit when compared to Low experience
```

```
coefficient of High Salesforce experience is 3720, which suggest that with High
Experience the Sales Unit increases by 3720 unit when compared to Low experience */

/* b) As seen in the glm model, High Sales force experience have an ability to sell
specific type of products.
They are good in selling Product type=3( Low Design and High Quality), with coefficient
of 348 which tells us that High Sales force experience sells 348 Unit of Low-Design High-
Quality product
more when compared to High Design and High Quality.

It shows that High Sales Force Experience can sell more high-ended product, specifically
high quality product when compared low quality product*/



/* Ques 8: Imagine you only observe Sales unit, price, Season, Design, and Quality
variables.
Create a categorical variable Product-Type to code the 4-types product as in Q6.
You are interested in knowing
(a) How far does the Seasonality increase the sales?
(b) Also, you want to know do people have any preference to buy specific type of products
in different seasons?
In other words, do people buy more high quality or luxury products based on the current
season?
You should write precisely a regression model, estimate it, and explain the estimation
results to identify the above questions.
Can you explain a logical, verbal story about the seasonality effects which you find in
your analysis?
In other words, please try to justify your result based on a logical argument.  */


proc glm data = Sales;
class  Season(ref = '2Spring') Product_Type(ref = '1');
 model SalesUnit =  Price Season Product_Type*Season/ solution;
 title 'Preference for type of product based on Seasonality' ;
run;
ods graphics on;

/* a) Seasonality effect can be seen in the glm model.
Summer season has the highest sales unit of 4231 unit more than the Spring Season.
Similarly, Winter and Fall season sells 1886 unit and 3466 unit more than the Spring
season.  */

/*b) Below are the season prefernce with regards to the Product Type:

       Winter : Product Type 3 (Low Design and High Quality)
       Summer : Product Type 2 (Luxury Design and Low Quality)
     Fall   : Product Type 3 (Low Design and High Quality)
       Spring : Product Type 3 (Low Design and High Quality)

We can tell in the Summer season is more inclined towards the Luxurry Design products
because in Summer season sells the maximum of Product Type=2 with
6626 unit more than product type 1.

As seen in the GLM model, the trend of selling a product is similar for Winter, Fall and
Spring season since the Sales Unit is highest for Product Type =3 and lowest for
Product Type=4 in all these season. But the seasonlity effect is visible in the summer
season since the trend shifts from Product Type=3 to Product Type =2.
```

```
In summer customer prefered Luxury Design product when compared to High Quality product
in all the other season.

Considering this as clothing brand, there can be lot of general stories which can proof
our argument above:
- People during summer have more money to spend on clothing since they do not go for
jackets and sweater which are costly. Since less clothing os required
during summer , they go for Luxury product
- Since the weather is sunny during summer, people spend a lot of time outside their home
and hence prefer luxury design to showcase their extravagant life

*/



/* Ques 9:  Imagine you only observe Sales unit, price, and advertising spending
variables.
Based on your answers in Q2 and Q3, provide a regression model's result to explain the
Sales unit as a function of price and advertising spending.
Based on Cook'D statistic, determine all influential points. You need to print the
influential points in your report.
Now, repeat your above regression analysis without including the influential points.
Does it improve your model goodness-of-fit?
If yes, how far? (Hint: follow the rule of thumb which is stated in class) */

proc reg data = Sales;
 model SalesUnit =  Price AdvertisingSpending Price_2 AdvertisingSpending_2;
output out = data cookd = cookd student=sresiduals;
title 'Regression  model with Influential Points' ;
run;
ods graphics off;
quit;

/* prints the observations that are influential (cook's d > 4/ n) */
proc print data=data ;
 var _ALL_;
 where Cookd > 4 / 1600;
run;
ods graphics on;

/* Estimate the model without the influential points */

proc reg data=data;
model SalesUnit =  Price AdvertisingSpending Price_2 AdvertisingSpending_2;
where Cookd < 4 / 1600;
title 'Regression  model without Influential Points' ;
run;
ods graphics off;
quit;


/* If the regression analysis is performed without influential points, the goodness-of-
fit is increased. The Adjusted R-Squ with influential point is 0.20,
if we run regression analysis without the influential point, the R-Square value is
increases to 0.2319 which shows us increase in goodness of fit. */

/* Ques 10: Imagine you only observe Sales unit, price, and advertising spending
variables.
```

```
Based on your answers in Q2 and Q3, provide a regression model's result to explain the
Sales unit as a function of price and advertising spending.
Now, check the collinearity of independent variables by reporting the VIF. If the VIF
>10,
drop some variables to reduce collinearity in an appropriate way.
Does your new model fit the data better? If yes why? If no, why? */

proc reg data = Sales;
 model SalesUnit =  Price Price_2 AdvertisingSpending AdvertisingSpending_2 / collinoint
vif;
run;

proc reg data = Sales;
 model SalesUnit =  Price  AdvertisingSpending  / collinoint vif;
run;

proc reg data = Sales;
 model SalesUnit =  Price_2  AdvertisingSpending_2 / collinoint vif;
run;

proc reg data = Sales;
 model SalesUnit =  Price_2  AdvertisingSpending / collinoint vif;
run;

proc reg data = Sales;
 model SalesUnit =  Price  Price_2 AdvertisingSpending  / collinoint vif;
run;

proc reg data = Sales;
 model SalesUnit =  Price  AdvertisingSpending AdvertisingSpending_2  / collinoint vif;
run;

/* VIF >10 if all the variable are added to the regression model i.e Price, Price_2,
AdvertisingSpending and AdvertisingSpending_2 and hence multicollinearity is present in
the model.
   VIF <10 if we just include Price and AdvertisingSpend, and hence the multicollinearity
is removed from the model.

 No, the model does not fit better since removing variable will decrease the R square
value, more number of variable will lead to high R Square value.
But, it will lead to true model and standard error will not be large

*/


/* Ques 11 : By using Proc Glmmod, you should create two new datasets out of Sales.csv
and Sales_test.csv
which include all the above quantitative variables and all dummy variables of all the
above qualitative variables.
By using Proc Reg, find the best regression model on the above generated data out of
Sales.csv.
you must use the Mallows' C_p criterion for selection procedure. Now, by considering the
above best regression model, record its R^2 and
calculate the MSE_test by using the above generated data out of Sales_test.csv (Hint: You
can use Proc score)

Now, by using Proc Glmselect, find the best models according to Forward, Backward,
Stepwise algorithm. In all cases,
(1) - you must use the Mallows' C_p criterion for selection procedure,
```

```
(2) adding all interaction effects at the second order, and (3) in a hierarchal order,
when the main effects should be included first.
Meanwhile, find and record their R^2 and the MSE_test of the best selected models by the
above procedures
(i.e., Forward, Backward, Stepwise algorithm) by using Sales_test.csv as the test
dataset.
Now, please rank all the above recorded best models based on either their R^2 or the
MSE_test of them.
If your goal is to do inference about the Sales unit based on independent variables which
one of the above best models will be your choice?
If your goal is to do predictive analysis about the Sales unit based on independent
variables which one of the above best models will be
your choice? */

proc glmmod data=Sales outdesign=Sales_with_indicators noprint;
 class Design Quality Promotion SalesforceExperience Location Season;
 model SalesUnit = Price Price_2 AdvertisingSpending AdvertisingSpending_2 Design Quality
Promotion SalesforceExperience Location Season/ noint;
 title 'Regression using Sales data';
run;


data Sales;
 set Sales;
 Price_2 = Price**2;
 AdvertisingSpending_2 = AdvertisingSpending**2;
run;


data Sales_test;
 set Sales_test;
 if Design = 1 & Quality = 1 then Product_Type = 1;
 if Design = 1 & Quality = 2 then Product_Type = 2;
 if Design = 2 & Quality = 1 then Product_Type = 3;
 if Design = 2 & Quality = 2 then Product_Type = 4;
run;

data Sales_test;
 set Sales_test;
 Price_2 = Price**2;
 AdvertisingSpending_2 = AdvertisingSpending**2;
run;

proc glmmod data=Sales_test outdesign=Sales_test_with_indicators noprint;
 class Design Quality Promotion SalesforceExperience Location Season;
 model SalesUnit = Price Price_2 AdvertisingSpending AdvertisingSpending_2 Design Quality
Promotion SalesforceExperience Location Season/ noint;
  title 'Regression using Sales test data';
run;
proc contents data=Sales_with_indicators;
run;
proc contents data=Sales_test_with_indicators;
run;
proc reg data=Sales_with_indicators outest = result  plots=all;
 model SalesUnit = col1-col19 /selection=cp adjrsq aic bic best=10;
run;
quit;
```

```
/* Not working Score */

proc score data=Sales_test_with_indicators score=result Type=parms  predict
out=predicted_data;
var col1-col19;
run;


/*  ------*/
data predicted_data;
set  predicted_data;
residula_2 = (SalesUnit-model)**2;
run;
proc means data = predicted_data mean ;
var residula_2;
run;


proc glmselect data=Sales testdata=Sales_test  plots=all;
 class Design(split)  Quality(split) Promotion(split) SalesforceExperience(split)
Location(split) Season(split);
model SalesUnit =
Price|Price_2|AdvertisingSpending|AdvertisingSpending_2|Design|Quality|Promotion|Salesfor
ceExperience|Location|Season @2
  /selection=forward(select=cp) hierarchy=single showpvalues;
 performance buildsscp=incremental;
 title ' Best Regression model using Forward Algorithm' ;
run;
proc glmselect data=Sales testdata=Sales_test  plots=all;
 class Design(split)  Quality(split) Promotion(split) SalesforceExperience(split)
Location(split) Season(split);
model SalesUnit =
Price|Price_2|AdvertisingSpending|AdvertisingSpending_2|Design|Quality|Promotion|Salesfor
ceExperience|Location|Season @2
  /selection=backward(select=cp) hierarchy=single showpvalues;
  title ' Best Regression model using Backward Algorithm' ;
 performance buildsscp=incremental;
run;
proc glmselect data=Sales testdata=Sales_test  plots=all;
 class Design(split)  Quality(split) Promotion(split) SalesforceExperience(split)
Location(split) Season(split);
model SalesUnit =
Price|Price_2|AdvertisingSpending|AdvertisingSpending_2|Design|Quality|Promotion|Salesfor
ceExperience|Location|Season @2
  /selection=stepwise(select=cp) hierarchy=single showpvalues;
 performance buildsscp=incremental;
 title ' Best Regression model using Stepwise Algorithm' ;
run;


/*
Algorithm   R Square    Adjusted R  MSE(test)
Forward          0.9495       0.9473            81605603
Backward    0.9501            0.9478            81930426
Stepwise    0.9491            0.947       81489109

The above table gives a summary statistic while using all the diffrent algorithm.
For Infrences about Sales Unit we will be using our Adjusted R square value as the
judging parameter. The more the Adj R square value, the better the model will be.
```

So for Infernce - Backward Algorithm gives us the best model.

For Prediction about Sales Unit we will be using our MSE test value as the judging parameter. The less the MSE test value, the better the model will be.
So for Prediction - Stepwise Algorithm gives us the best model.

*/

/* Ques 12 : Now, by using Proc Glmselect, find the best models according to Forward, Backward, Stepwise algorithm.
In all cases, (1) - you must use the Cross-Validation criterion for selection procedure based on creating 10-folds of Sales.csv dataset
(please set the seed=2), (2) adding all interaction effects at the second order, and (3) in a hierarchal order,
when the main effects should be included first. Meanwhile, find and record their R^2 and the MSE_test of the best selected models by the
above procedures (i.e., Forward, Backward, Stepwise algorithm) by using Sales_test.csv as the test dataset.

Now, please rank all the above recorded best models based on either their R^2 or the MSE_test of them in Q11 and Q12.
If your goal is to do inference about the Sales unit based on independent variables which one of the above best models will be your choice?
If your goal is to do predictive analysis about the Sales unit based on independent variables which one of the above best models will be
your choice?

*/

```
proc glmselect data=Sales testdata=Sales_test  seed = 2 plots=all;
 class Product_Type(split) Promotion(split) SalesforceExperience(split) Location(split)
Season(split);
model SalesUnit =
Price|Price_2|AdvertisingSpending|AdvertisingSpending_2|Product_Type|Promotion|Salesforce
Experience|Location|Season @2
  /selection=forward(select=cv) hierarchy=single cvmethod=random(10)showpvalues ;
 performance buildsscp=incremental;
 title ' Best Regression model using Forward Algorithm' ;
run;
proc glmselect data=Sales testdata=sales_test  seed = 2 plots=all;
 class Product_Type(split) Promotion(split) SalesforceExperience(split) Location(split)
Season(split);
model SalesUnit =
Price|Price_2|AdvertisingSpending|AdvertisingSpending_2|Product_Type|Promotion|Salesforce
Experience|Location|Season @2
  /selection=backward(select=cv) hierarchy=single cvmethod=random(10)showpvalues ;
 performance buildsscp=incremental;
 title ' Best Regression model using Backward Algorithm' ;
run;
proc glmselect data=sales testdata=sales_test  seed = 2 plots=all;
 class Design(split)  Quality(split) Promotion(split) SalesforceExperience(split)
Location(split) Season(split);
```

```
model SalesUnit =
Price|Price_2|AdvertisingSpending|AdvertisingSpending_2|Design|Quality|Promotion|SalesforceExperience|Location|Season @2
  /selection=stepwise(select=cv) hierarchy=single cvmethod=random(10)showpvalues ;
 performance buildsscp=incremental;
  title ' Best Regression model using Stepwise Algorithm' ;
run;

/* Note : It shows the same solution for Question 11 and Question 12 */

/*
Algorithm   R Square    Adjusted R  MSE(test)
Forward          0.9490       0.9469            81832696
Backward    0.9501            0.9478            81930426
Stepwise    0.9491            0.947       81489109

The above table gives a summary statistic while using all the diffrent algorithm.
For Infrences about Sales Unit we will be using our Adjusted R square value as the
judging parameter. The more the Adj R square value, the better the model will be.
So for Infernce - Backward Algorithm gives us the best model.

For Prediction about Sales Unit we will be using our MSE test value as the judging
parameter. The less the MSE test value, the better the model will be.
So for Prediction - Stepwise Algorithm gives us the best model.


*/

/* Ques 13: Now, by using Proc Glmselect, find the best models according to LASSO and
Elastic Net.
In all cases, (1) - you must use the Cross-Validation criterion for selection procedure
based on creating 10-folds of Sales.csv dataset
(please set the seed=2), (2) adding all interaction effects at the second order, and (3)
in a hierarchal order,
when the main effects should be included first. Meanwhile, find and record their R^2 and
the MSE_test of the best selected models by the
above procedures (i.e., LASSO and Elastic Net) by using Sales_test.csv as the test
dataset.

If your goal is to do inference about the Sales unit based on independent variables which
one of the above best models will be your choice?
If your goal is to do predictive analysis about the Sales unit based on independent
variables which one of the above best models will be
your choice?

*/

proc glmselect data=sales testdata=sales_test  seed = 2 plots=all;
 class Design(split)  Quality(split) Promotion(split) SalesforceExperience(split)
Location(split) Season(split);
model SalesUnit =
Price|Price_2|AdvertisingSpending|AdvertisingSpending_2|Design|Quality|Promotion|SalesforceExperience|Location|Season @2
  /selection=lasso(choose=cv stop=none) hierarchy=single cvmethod=random(10)showpvalues ;
 performance buildsscp=incremental;
  title ' Best Regression model using Lasso' ;
run;
proc glmselect data=sales testdata=sales_test  seed = 2 plots=all;
```

```
 class product_type(split) Promotion(split) SalesforceExperience(split) Location(split)
Season(split);
model SalesUnit =
Price|Price_2|AdvertisingSpending|AdvertisingSpending_2|product_type|Promotion|Salesforce
Experience|Location|Season @2
  /selection=elasticnet(choose=cv stop=none) hierarchy=single
cvmethod=random(10)showpvalues ;
 performance buildsscp=incremental;
 title ' Best Regression model using Elastic Net' ;
run;

/*
Algorithm    R Square    Adjusted R  MSE(test)
LASSO      0.9481       0.9456            81217832
Ealstic Net  0.9469           0.9444            81083439

The above table gives a summary statistic while using all the diffrent algorithm( Elastic
Net & LASSO)
For Infrences about Sales Unit we will be using our Adjusted R square value as the
judging parameter. The more the Adj R square value, the better the model will be.
So for Infernce - LASSO Algorithm gives us the best model.

For Prediction about Sales Unit we will be using our MSE test value as the judging
parameter. The less the MSE test value, the better the model will be.
So for Prediction - ElasticNet Algorithm gives us the best model.

*/



/* Ques 14: Now, by using Proc Glmselect, do the regression analysis based on Bagging
Method. In all cases,
(1) - you must use the Cross-Validation criterion for selection procedure based on
creating 10-folds of Sales.csv dataset
(please set the seed=2), (2) adding all interaction effects at the second order, (3) in a
hierarchal order,
when the main effects should be included first, and use 100 samples (based on
Bootstrapping). Based on the Average Model of 100 models,
find and record the MSE_train and the MSE_test by using Sales.csv and Sales_test.csv as
the training and test datasets respectively.

If your goal is to do inference about the Sales unit based on independent variables does
Bagging Method provide you any advantage?
If your goal is to do predictive analysis about the Sales unit based on independent
variables does Bagging Method provide you any advantage? */

proc glmselect data=sales testdata=sales_test  seed = 2 plots=all;
 class Design(split)  Quality(split) Promotion(split) SalesforceExperience(split)
Location(split) Season(split);
model SalesUnit =
Price|Price_2|AdvertisingSpending|AdvertisingSpending_2|Design|Quality|Promotion|Salesfor
ceExperience|Location|Season @2
  /selection=stepwise(select=cp) hierarchy=single cvmethod=random(10)showpvalues ;
 modelaverage nsamples=100 tables=(ParmEst(all));
 performance buildsscp=incremental;
 score data=sales_test out=test_performance residual=res;
 score data=sales out=insample_performance residual=res;
 title ' Regression using Bagging Method' ;
run;
```

```
/* Ques 15: This firm wants to increase its sales. They have two options. The first way
is to train their sales-force people to become more
expert. The second way is to increase their advertising budget. The firm told you that
they can increase their advertising budget by 5%.
Also, they know, by having a one-day workshop for "Low" and "Average"-types of sales-
force people, they will become "Average" and
"High"-types sales-force respectively.

You want to do a contrafactual analysis for the firm. You will do the analysis here based
on the Average Model of 100 models in Q14.
Replicate the above scenarios by either increasing the advertising spending variables by
5% in Sales.csv, or increasing the sales-force
experience level form "Low" TO "Average" and from "Average" To "High" in Sales.csv (due
to a one-day workshop).
Based on these changes in Sales.csv, you will have two new datasets called
Sales_increasing_ad and Sales_increasing_expertise.

Now, predict the sales unit in the above two datasets based on the Average Model of 100
models in Q14. You know the difference of sales
before and after of the above firms' strategies (i.e., either increasing advertising
budget by 5% or increasing sales-force expertise by
holding workshop).

Which strategy will increase the sales more? What is the average sales unit changes by
Product-Type in each strategy?
Which strategy is more profitable from your point of view based on your information about
this firm?

*/

data new_data;
set sales;
if SalesforceExperience = 'Average' then SalesforceExperience='High';
if SalesforceExperience = 'Low' then SalesforceExperience='Average';
run;

data new_data2;
set sales;
advertisingspending = (advertisingspending + 0.05*advertisingspending);
advertisingspending_2 = advertisingspending**2;
run;


/*-------------Bagging Model----------------------*/
proc glmselect data=sales seed=2 plots=all;
 class Product_Type(split) promotion(split) SalesforceExperience(split) location(split)
season(split);
model
salesunit=Price|Price_2|AdvertisingSpending|AdvertisingSpending_2|Product_Type|Promotion|
SalesForceExperience|Location|Season@2
  /selection=stepwise(select=cv) hierarchy=single cvmethod=random(10) showpvalues;
 modelaverage nsamples=100 tables=(ParmEst(all));
 score data=new_data out=performance_train residual=res predicted=psalesunit;
 score data=new_data2 out=performance_ad residual=res predicted=psalesunit;
 performance buildsscp=incremental;
run;
```

```
data performance_train;
set performance_train;
diff = psalesunit - salesunit;
run;

proc means data=performance_train;
var diff;
title 'Means for Performance by Sales Force Experience ';
run;

data performance_ad;
set performance_ad;
diff = psalesunit - salesunit;

proc means data=performance_ad;
var diff;
title 'Means for Performance by Advertising Spending ';
run;

proc sort data=performance_train;
by Product_Type;
run;

proc sort data=performance_ad;
by Product_Type;
run;

proc means data=performance_train;
var diff;
by Product_Type;
title 'Means for Performance by Sales Force Experience ';
run;

proc means data=performance_ad;
var diff;
by Product_Type;
title 'Means for Performance by Advertising Spending ';
run;
```

End Report.