Spring 2019

BUAN 6337: Predictive Analytics using SAS

A Report of

Group Assignment on
Sales Data Analysis

Submitted by Group 7

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

Under the Guidance of

Prof. Shervin Shahrokhi Tehrani

Section 1

Problem 1: Describe the type of each variable in terms of a qualitative variable or quantitative variable.

Quantitative Variables: Sales Unit | Price | Advertising Spending
Qualitative Variables: Design | Quality | Promotion | Sales Force Experience | Season | Location
ID is neither a Qualitative nor a qualitative variable since it does not include any order preference or natural order. It is just a way to represent each row of the data.

Problem 2: How many stores are there in the data? what are their frequencies in Europe and North America?

There are 400 stores in the data with 206 in Europe and 194 in North America.

Total Store	Europe	North America
400	206	194

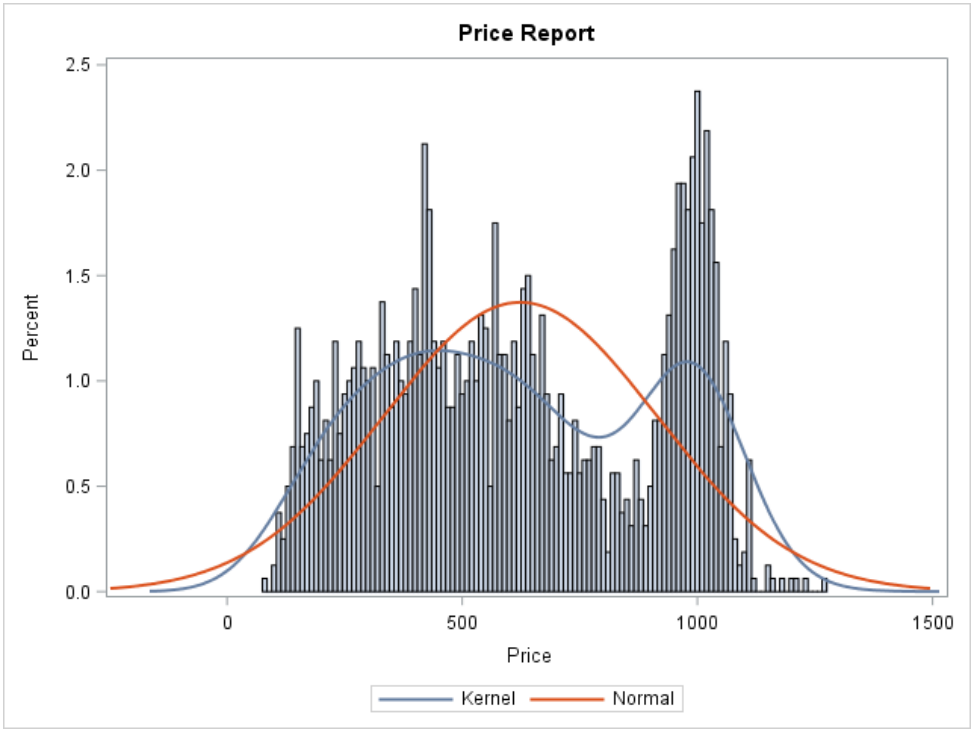
Problem 3: What is the average Sales Unit, Price, and Advertising Spending?

Below is the detail about the average of the quantitative variables:
Average Sales Unit: 35352.65
Average Price: $623.11
Average Advertising Spending: $10,091.67

Sales_Unit_Average	Price_Average	Adv_Spend_Average
35352.65	623.1062	10091.67

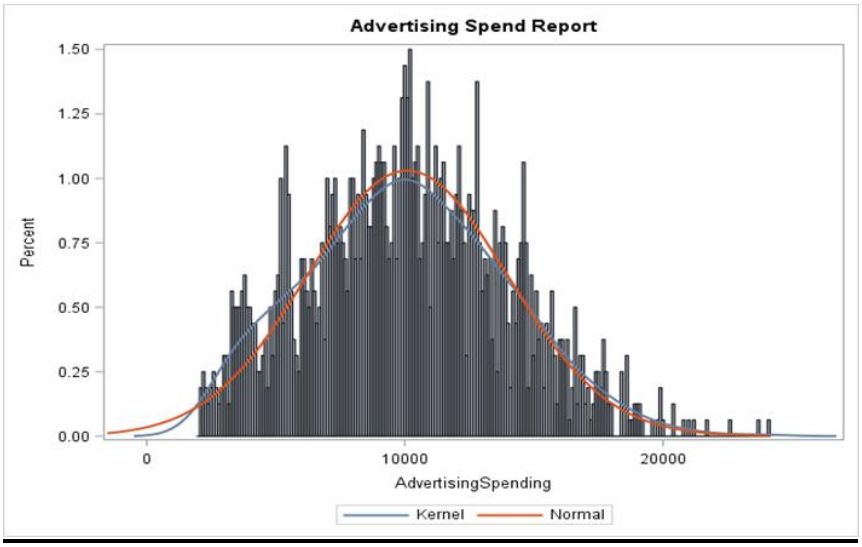
Problem 4: Plot histograms for Sales Unit, Price, and Advertising Spending. Your histograms should have the best normal and kernel fitted curve. For Sales Unit, Price, and Advertising Spending use binwidths 1000, 10, 100 respectively. Which of variables does look different from a normal distribution?

Below are the histogram for Price, Sales Unit and Advertiser Spending:



Histogram of Price (Binwidth 10)

Histogram of Sales (Binwidth 1000)



Histogram of Advertising Spending (Binwidth 100)

Findings:

1. Price Report: Price Histogram is not a normal curve and shows a bi-modal distribution
2. Sales Report: It is close to a normal distribution but shows a right-skewed distribution
3. Advertising Spend Report: It approaches a normal distribution curve by visual inspection

Problem 5: Find the correlation matrix of the Sales Unit, Price, and Advertising Spending? Report both Pearson and Spearman Correlation tests. How can you describe the effects of Price and Advertising Spending on Sales Unit?

Below is the correlation table for the variable:

Pearson and Spearman Correlation for Sales Unit Price and Advertising Spend

The CORR Procedure

3 Variables:	SalesUnit Price AdvertisingSpending

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
SalesUnit	1600	35353	9863	33169	20016	67375
Price	1600	623.10624	290.49632	594.58500	84.55000	1265
AdvertisingSpending	1600	10092	3874	10030	2089	24131

Pearson Correlation Coefficients, N = 1600
Prob > |r| under H0: Rho=0

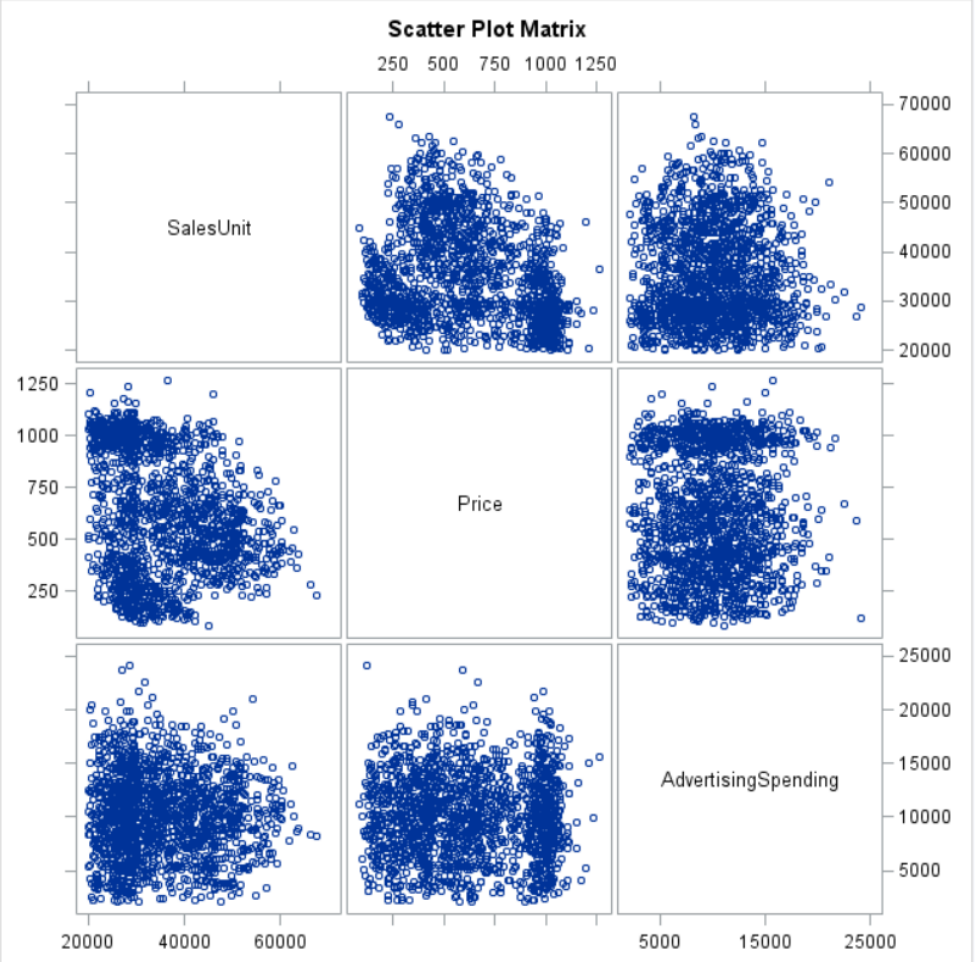
	SalesUnit	Price	AdvertisingSpending
SalesUnit	1.00000	-0.26151 <.0001	-0.01915 0.4439
Price	-0.26151 <.0001	1.00000	0.00639 0.7984
AdvertisingSpending	-0.01915 0.4439	0.00639 0.7984	1.00000

Spearman Correlation Coefficients, N = 1600
Prob > |r| under H0: Rho=0

	SalesUnit	Price	AdvertisingSpending
SalesUnit	1.00000	-0.25781 <.0001	-0.00056 0.9821
Price	-0.25781 <.0001	1.00000	0.00849 0.7343
AdvertisingSpending	-0.00056	0.00849	1.00000

Spearman-Pearson Correlation table

Here we see negative correlation between SalesUnit and Price, which tells us as Price increases SalesUnits decrease, however, this relationship is very weak(-.26).The correlationbetween Advertising Spending and SalesUnits is extremelyweak(-.02) implying that there is no linear relationship between these variables. It's possible that these variables are in fact related. It could be that a non-linear relationship between Advertising Spending and Sales Unitsexists or that an unaccounted-forfactor is masking the relationship between these variables

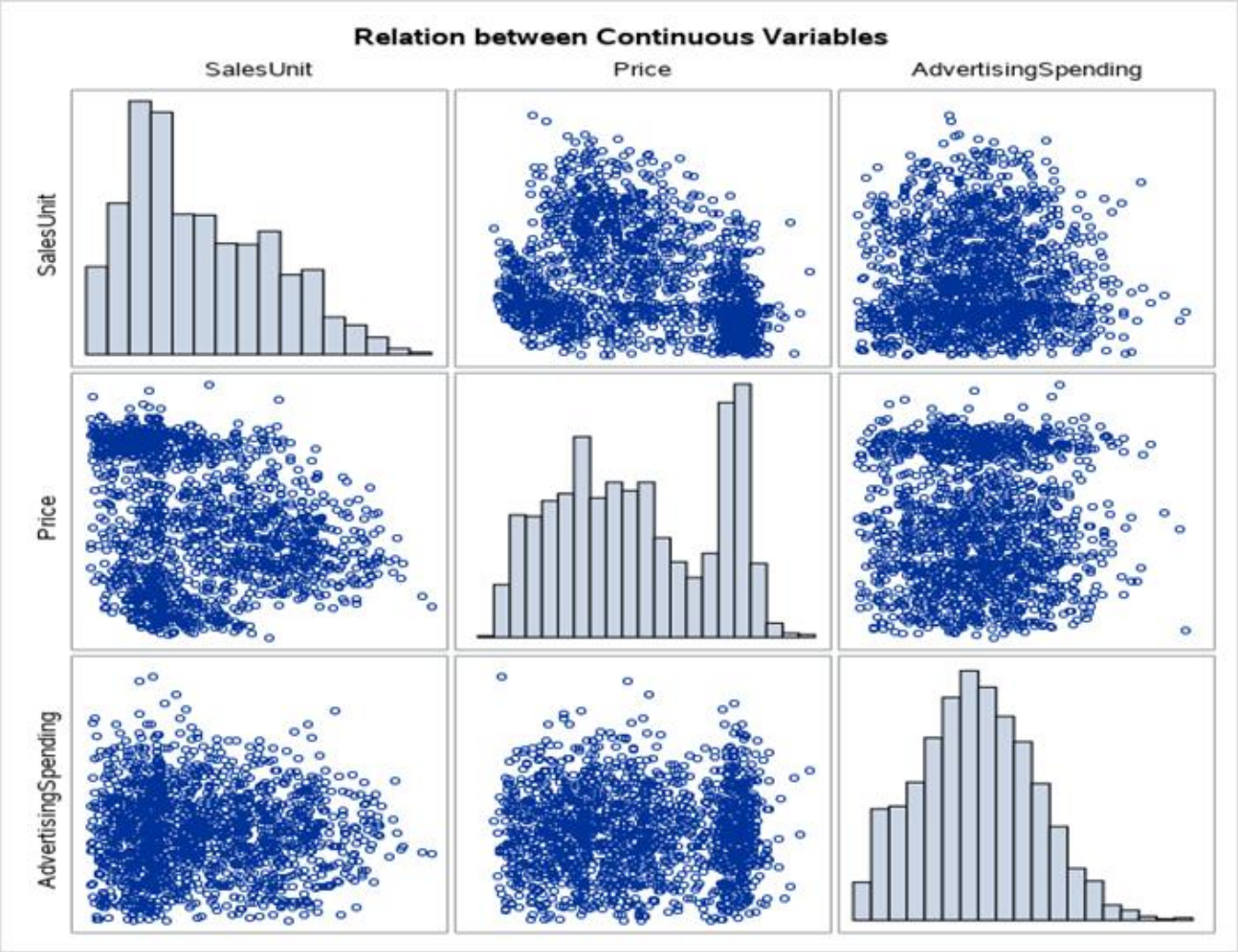


Findings:

- Negative correlation between Sales Unit and Price, which tells us as Price increases Sales Units decrease, however, this relationship is very weak (-.26).
- Correlation between Advertising Spending and Sales Units is extremely weak (-.02) implying that there is no linear relationship between these variables. The value is insignificant with a value of 0.9821.

Problem 6: Use Proc sgscatter to draw a matrix of Sales Unit, Price, Advertising Spending relations. Does the scatter plot of Sales Unit based on Advertising Spending confirm your result in Q5? If not, why?

Below is scatter plot of the continuous variables:

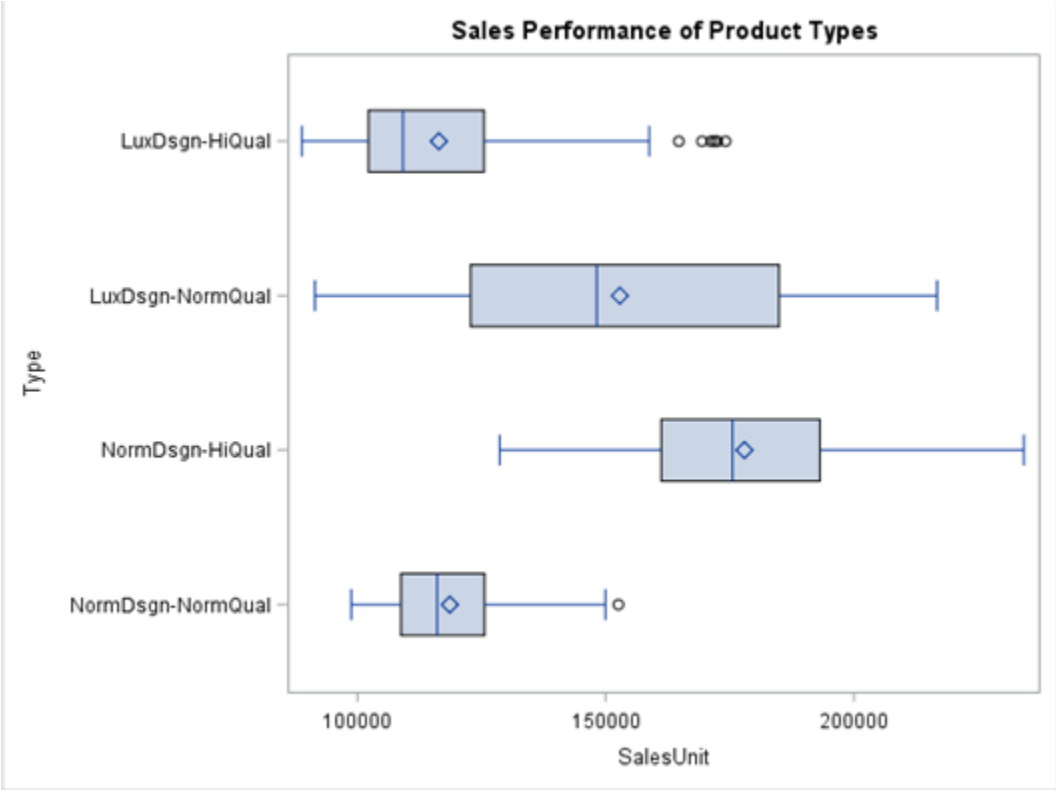


Scatter Plot of Continuous Variables

Findings:

- Scatter Plot of the matrix between Price and Sales Unit seems to support our previous observation of decrease in Price as Sales Unit increases. But it has a lot of spread visually and does not show a significant reduction.
- Not much can be inferred from the Advertising Spend and Price plot which proves our interpretation from the correlation matrix where it shows a decreasing trend though the value is not significant.

Problem 7: Use Box Plotting figures to determine the average people preference on Design and Quality level? Then, redo your analysis based on people located in Europe vs. North America. Do you find any evidence on a different taste of Design and Quality among Europeans vs. Americans?



Box Plot of Design type

Findings:

- The Luxury Design in High Quality has some outliers.
- The normal Design in normal Quality box plot is small as compared with other box plots which shows that the sales unit is very much concentrated to a range of some values.
- The Luxury Design in Normal Quality is tall as compared with other box plots which shows sales unit varies a lot.
- The Luxury Design in High Quality has some outliers. The mean is greater than median that means the data is right skewed (positive Skewed). The range is (Min-Max) 85306$.

Type=LuxDsgn-HiQual

Analysis Variable : SalesUnit							
N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
101	116381.49	20051.96	88829.00	102201.00	109171.00	125502.00	174135.00

There are no outliers. The Luxury Design in normal Quality mean is greater than median that means the data is right skewed (positive Skewed). The Range is 125283$.

Type=LuxDsgn-NormQual

Analysis Variable : SalesUnit							
N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
91	152849.88	35186.32	91448.00	122791.00	148198.00	184919.00	216731.00

There are no outliers. The Normal Design in high Quality mean is greater than median that means the data is right skewed (positive Skewed). The Range is 105565$.

Type=NormDsgn-HiQual

Analysis Variable : SalesUnit							
N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
105	177946.62	21712.29	128660.00	161195.00	175541.00	193140.00	234225.00

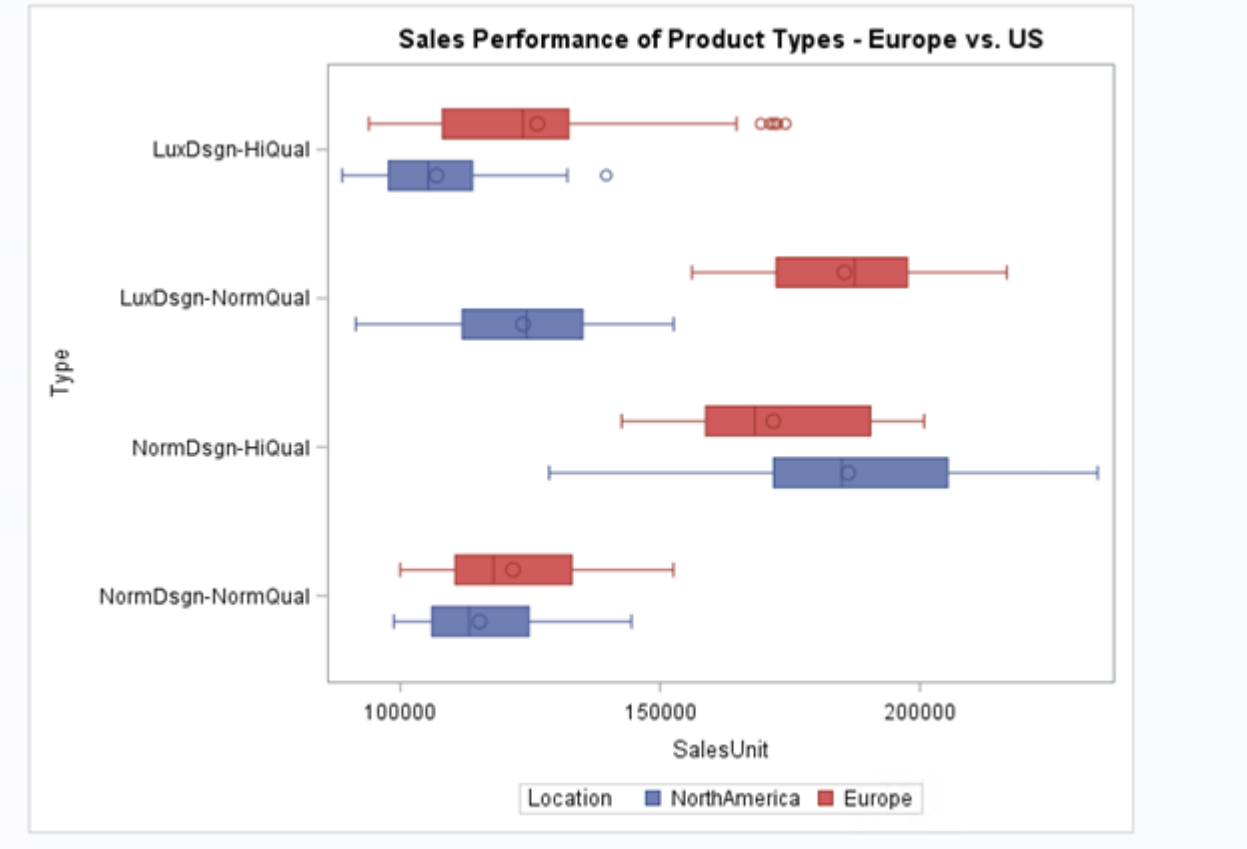
It has only one outlier. The Normal Design in normal Quality mean is greater than median that means the data is right skewed (positive Skewed). The Range is 53758$.

Type=NormDsgn-NormQual

Analysis Variable : SalesUnit							
N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
103	118601.68	13233.22	98779.00	108770.00	116062.00	125590.00	152533.00

Findings:

- The Luxury Design in High Quality has some outliers in both the locations.
- The Luxury Design in High Quality box plot is small as compared with other box plots which shows that the sales unit is very much concentrated to a range of some values for both the Locations.
- The Normal Design in High Quality is tall as compared with other box plots which shows sales unit varies a lot for both the locations.
- The Luxury Design in Normal Quality, the maximum sales unit North America is minimum sales unit in Europe.



Box Plot of Design Type for Europe vs North America

The Luxury Design in High Quality shows that in Europe has more outliers than in North America. Even the median and mean for Europe is more than the median and mean in North America. In Europe the mean is greater than median that means the data is right skewed (positive Skewed). In North America the mean is greater than median that means the data is right skewed (positive Skewed).

Type=LuxDsgn-HiQual

Analysis Variable : SalesUnit									
Location	N Obs	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
Europe	49	49	126367.16	22229.46	93932.00	108149.00	123579.00	132393.00	174135.00
NorthAmerica	52	52	106971.90	11714.68	88829.00	97794.00	105369.50	113862.50	139628.00

The Luxury Design in Normal Quality shows that the median and mean for Europe is more than the median and mean in North America. In Europe the mean is greater than median that means the data is right skewed (positive Skewed). In North America the median is greater than median that means the data is left skewed (negative Skewed).

Type=LuxDsgn-NormQual

Analysis Variable : SalesUnit									
Location	N Obs	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
Europe	43	43	185461.63	17456.09	156125.00	172396.00	187422.00	197609.00	216731.00
NorthAmerica	48	48	123635.19	15931.61	91448.00	111937.50	124272.50	135105.00	152632.00

The Normal Design in High Quality shows that the median and mean for North America is more than the median and mean in Europe. In Europe the mean is greater than median that means the data is right skewed (positive Skewed). In North America the mean is greater than median that means the data is right skewed (positive Skewed).

Type=NormDsgn-HiQual

Analysis Variable : SalesUnit									
Location	N Obs	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
Europe	60	60	171783.45	16911.56	142599.00	158767.00	168267.00	190507.00	200810.00
NorthAmerica	45	45	186164.18	24696.41	128660.00	171897.00	185000.00	205353.00	234225.00

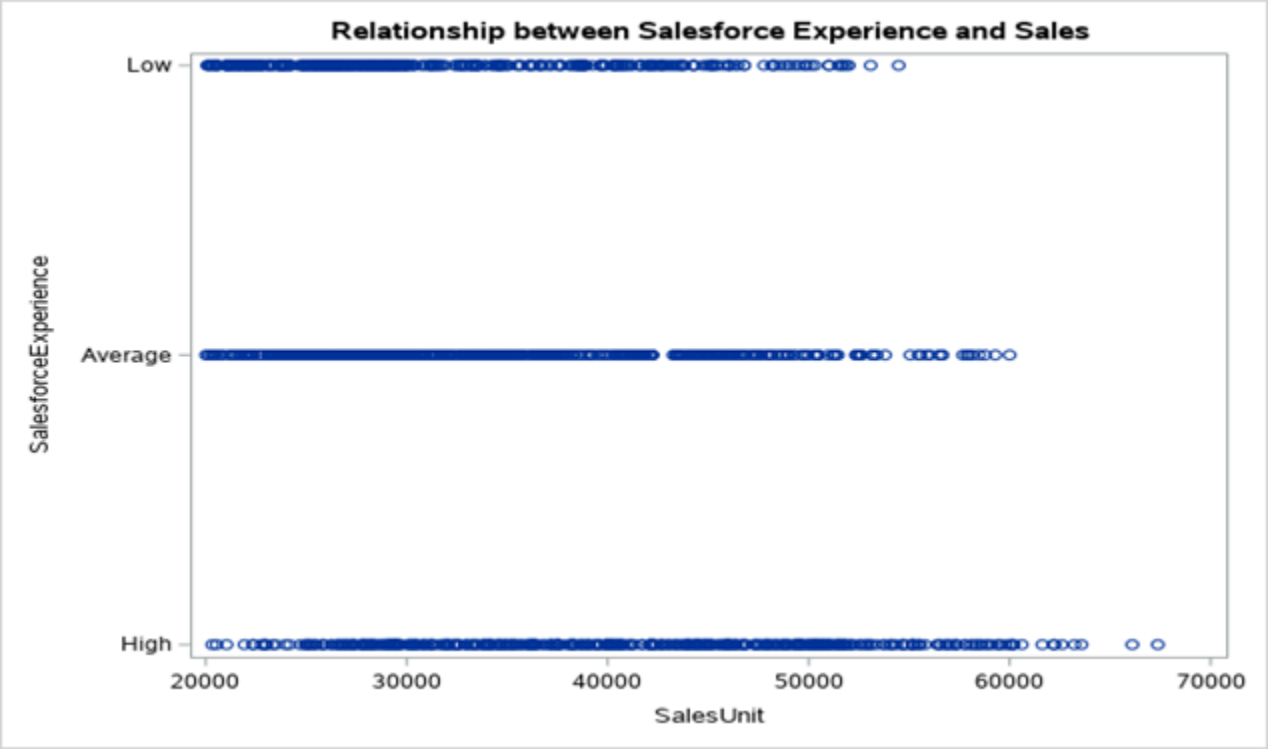
The Normal Design in Normal Quality shows that the median and mean for Europe is more than the median and mean in North America. In Europe the mean is greater than median that means the data is right skewed (positive Skewed). In North America the mean is greater than median that means the data is right skewed (positive Skewed).

Type=NormDsgn-NormQual

Analysis Variable : SalesUnit									
Location	N Obs	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
Europe	54	54	121675.07	14294.44	99959.00	110561.00	117944.50	133003.00	152533.00
NorthAmerica	49	49	115214.67	11143.29	98779.00	106137.00	113231.00	124750.00	144478.00

Problem 8: Use Scatter Plot figures to determine Sales Unit based on Salesforce experience level. Do find any clear evidence about Salesforce experience level? Then, redo your analysis based on the 4-types of products? Does higher Salesforce experience provide any special ability to sell more products? If yes, which type of products?

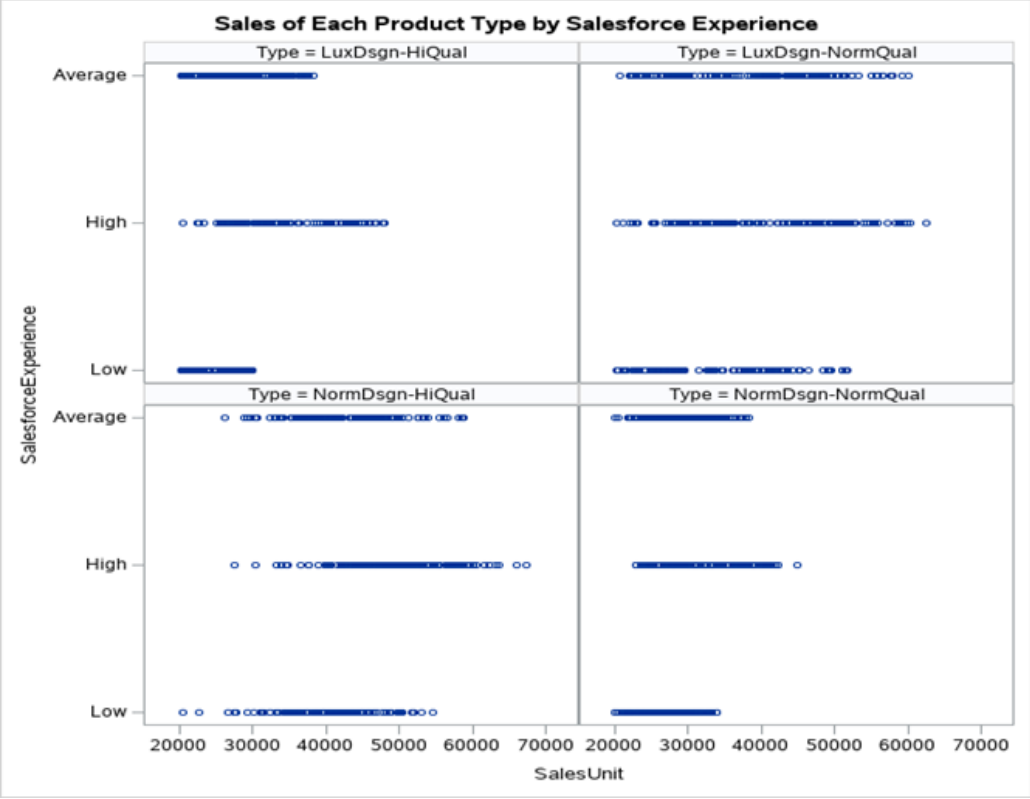
We find some evidence that higher salesforce experience leads to higher product sales, but it is difficult to see this relationship in a scatter plot. Comparing boxplots of sales by experience level shows that more experience is associated with higher sales.



Scatter Plot of Sales vs Salesforce Experience (Categories Average, High, Low)

Box Plot of Sales Unit vs Salesforce Experience (Categories Average, High, Low)

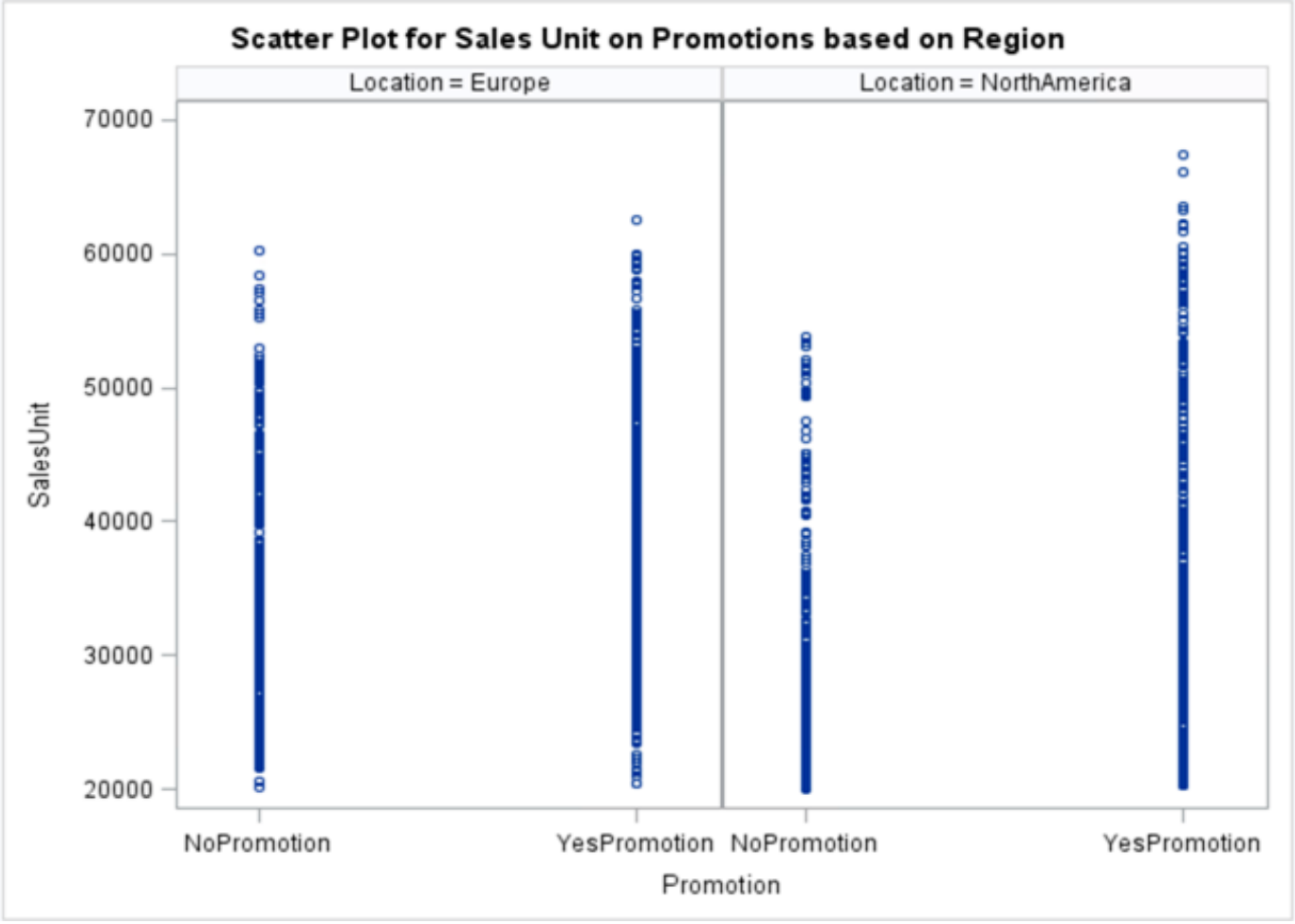
This relationship is most noticeable in the luxury design with high quality and normal design with high quality products.



Problem 9: Do question 8 again in terms of Sales Unit based on Promotion. Is there any difference between Europeans vs. Americans to response to promotion? Who is more price sensitive: Europeans or Americans?

Below are the scatter plot for Sales Unit based on promotion with region:

Scatter Plot of Sales Unit based on Promotion



Scatter Plot of Sales Unit on Promotions (Europe vs North America)

Finding:

- Yes, there is difference between Europeans vs American with response to Promotion. North America is more price sensitive based on Sales unit when compared to Europe.
- Since in the scatter plot for Sales Unit with region of North-America, there seems to be a increase in Sales Unit when a promotion is there as compared to No promotion.
- Europe region scatter plot with promotion doesn't vary much with/without promotion.

Problem 10: Provides a summary statistic of Sales unit, price, and advertising spending by seasons? Do the Sales Unit, Price, and Advertising Spending have different means across seasons?

Below table shows the summary statistics of the quantitative variables across season:

Summary Statistics by Season

Season=1Winter

	N	Mean	StdDev	Min	P25	Median	P75	Max
SalesUnit	400	34838.89	9728.40	20323.00	27636.00	32321.50	42224.50	62180.00
Price	400	629.10	299.52	102.44	370.92	588.33	948.85	1233.70
AdvertisingSpending	400	10223.36	3932.86	2494.65	7313.38	10049.75	13060.85	21030.50

Summary Statistics by Season

Season=2Spring

	N	Mean	StdDev	Min	P25	Median	P75	Max
SalesUnit	400	32586.23	8472.85	20016.00	26036.00	29653.00	38666.00	58600.00
Price	400	625.97	288.65	100.15	378.28	624.86	933.80	1209.46
AdvertisingSpending	400	10036.33	3787.33	2092.54	7520.13	10046.35	12704.25	20796.85

Summary Statistics by Season

Season=3Summer

	N	Mean	StdDev	Min	P25	Median	P75	Max
SalesUnit	400	37059.48	9949.48	20097.00	29033.50	35345.00	44378.50	62602.00
Price	400	618.24	291.06	112.15	388.37	588.38	928.90	1265.17
AdvertisingSpending	400	9816.82	3681.88	2145.53	7075.95	9836.44	12035.31	23722.94

Summary Statistics by Season

Season=4Fall

	N	Mean	StdDev	Min	P25	Median	P75	Max
SalesUnit	400	36926.00	10541.09	20139.00	28481.50	34731.50	45202.00	67375.00
Price	400	619.12	283.48	84.55	392.83	584.88	926.92	1153.46
AdvertisingSpending	400	10290.17	4077.80	2089.33	7394.67	10188.51	13184.69	24131.26

- Sales Unit: The mean of all the season is not perfectly same but are nearby to their other counter-parts. It is in the range of ~33,000.
- Price: The mean of all the season is not perfectly same but are nearby to their other counter-parts. It is in the range of ~600.

- Advertising Spending: The mean of all the season is not perfectly same but are nearby to their other counter-parts. It is in the range of ~10,000.

Section- 2

Assumed Condition: Confidence interval for following hypothesis testing is 95%

Problem 11: Provide a hypothesis testing to determine which of Sales Unit, Price, and Advertising Spending have significantly different means across seasons? (Hint: ANOVA)

1. Sales Unit vs seasons:

H0: Mean Sales Unit is equal for all season
Ha: Mean Sales Unit is different for all season.

Hypothesis Testing of Means : Sales Unit vs Season

The ANOVA Procedure

Dependent Variable: SalesUnit

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5322307397.6	1774102465.9	18.85	<.0001
Error	1596	150238522653	94134412.69		
Corrected Total	1599	155560830051			

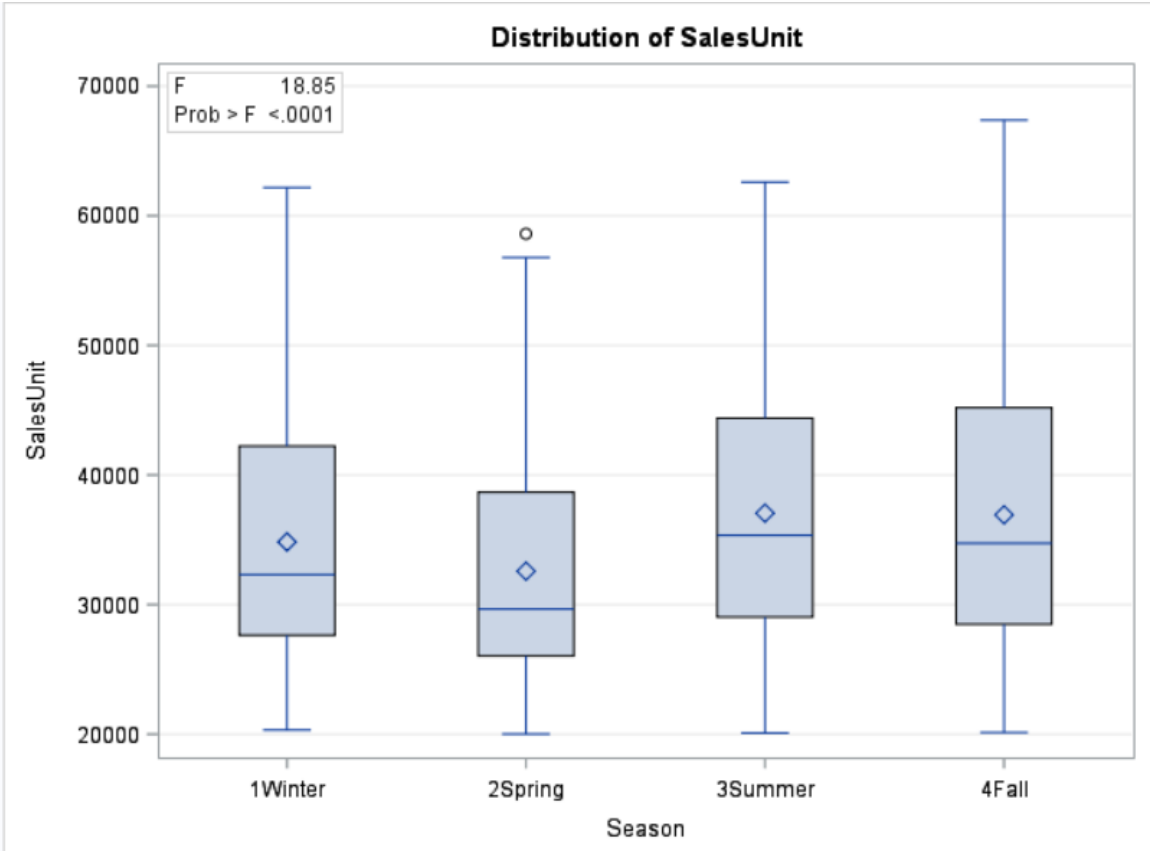
R-Square	Coeff Var	Root MSE	SalesUnit Mean
0.034214	27.44431	9702.289	35352.65

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Season	3	5322307398	1774102466	18.85	<.0001

ANOVA Statistics (Sales Unit vs Season)

The degrees of freedom (DF) column should be used to check the analysis results. The model degrees of freedom for a one-way analysis of variance are the number of levels minus 1; in this case, 4-1=3 The Corrected Total degrees of freedom are always the total number of observations minus one; in this case 1600-1=1599. The sum of Model and Error degrees of freedom equal the Corrected Total.

The overall F test is significant F-value = 18.85 and P<0.0001, indicating that the model as a whole accounts for a significant portion of the variability in the dependent variable.



Box Plot of Different Seasons on the basis of SalesUnit

Conclusion: We have sufficient evidence to accept that Sales Unit have significantly different means across seasons.

2. Price vs seasons:

H0: Price Unit is equal for all season
Ha: Price Unit is different for all season.

Hypothesis Testing of Means : Price vs Season

The ANOVA Procedure

Dependent Variable: Price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	33485.8	11161.9	0.13	0.9410
Error	1596	134903107.1	84525.8		
Corrected Total	1599	134936593.0			

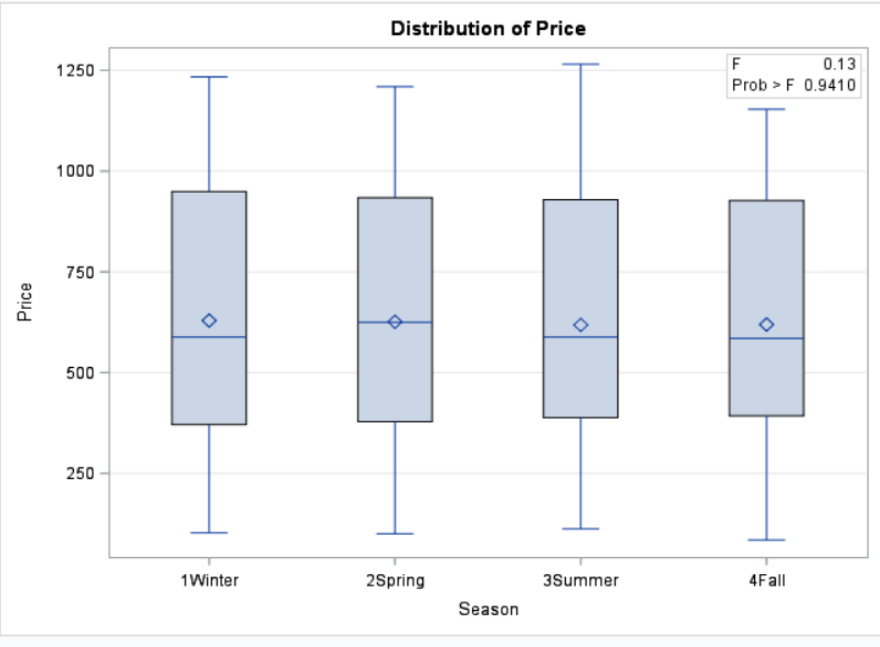
R-Square	Coeff Var	Root MSE	Price Mean
0.000248	46.65868	290.7331	623.1062

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Season	3	33485.83017	11161.94339	0.13	0.9410

ANOVA Statistics (Price vs Season)

The degrees of freedom (DF) column should be used to check the analysis results. The model degrees of freedom for a one-way analysis of variance are the number of levels minus 1; in this case, 4-1=3 The Corrected Total degrees of freedom are always the total number of observations minus one; in this case 1600-1=1599. The sum of Model and Error degrees of freedom equal the Corrected Total.

The overall F test is not significant F-value = 0.13 and P= 0.9410, indicating that the model as a whole accounts for a insignificant portion of the variability in the dependent variable.



Box Plot of Different Seasons on the basis of Price

Conclusion: We do not have sufficient evidence to accept that Price spend have significantly different means across seasons

3. Advertising Spending vs seasons:

H0: Mean Advertising spending is equal for all season
Ha: Mean Advertising spending is different for all season.

Hypothesis Testing of Means : Advertising Spend vs Season

The ANOVA Procedure

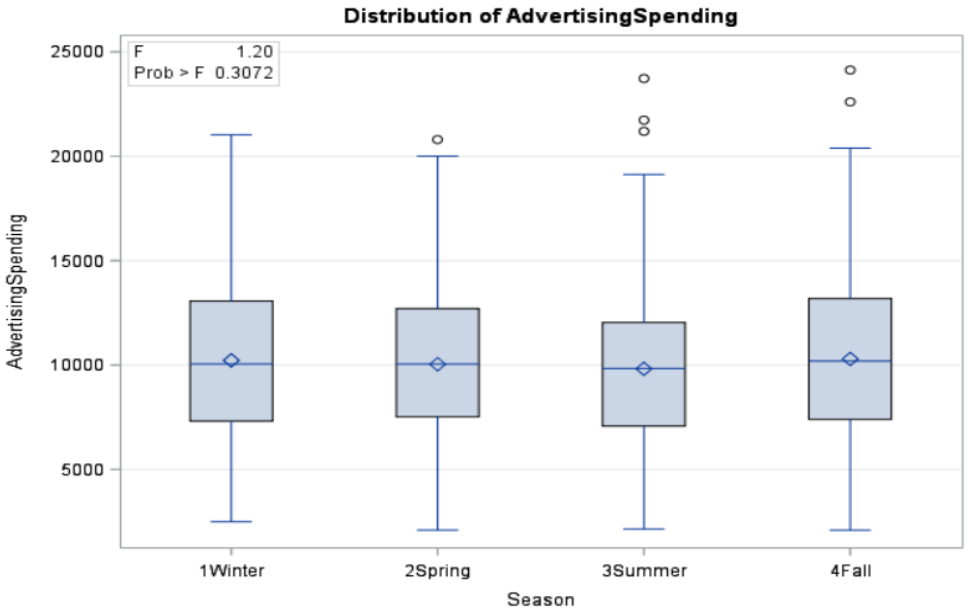
Dependent Variable: AdvertisingSpending

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	54140468	18046823	1.20	0.3072
Error	1596	23938387630	14998990		
Corrected Total	1599	23992528099			

R-Square	Coeff Var	Root MSE	AdvertisingSpending Mean
0.002257	38.37673	3872.853	10091.67

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Season	3	54140468.33	18046822.78	1.20	0.3072

ANOVA Statistics (Advertising Spending vs Season)



Box Plot of Different Seasons on the basis of Advertising Spending

The degrees of freedom (DF) column should be used to check the analysis results. The model degrees of freedom for a one-way analysis of variance are the number of levels minus 1; in this case, 4-1=3 The Corrected Total degrees of freedom are always the total number of observations minus one; in this case 1600-1=1599. The sum of Model and Error degrees of freedom equal the Corrected Total.

The overall F test is not significant F-value = 1.20 and P= 0.3072, indicating that the model as a whole accounts for a insignificant portion of the variability in the dependent variable.

Conclusion: We do not have sufficient evidence to accept that Advertising spend have significantly different means across seasons.

Problem 12: Provide a hypothesis testing to determine which of Sales Unit, Price, and Advertising Spending have significantly different means across location.

T-Test
H0: the mean of these three variables are the same in NA&EU;
Ha: the mean of these three variables are not the same in NA&EU

1. Sales Unit vs Location

Hypothesis Testing of Means : Sales Unit vs Location

The TTEST Procedure

Variable: SalesUnit

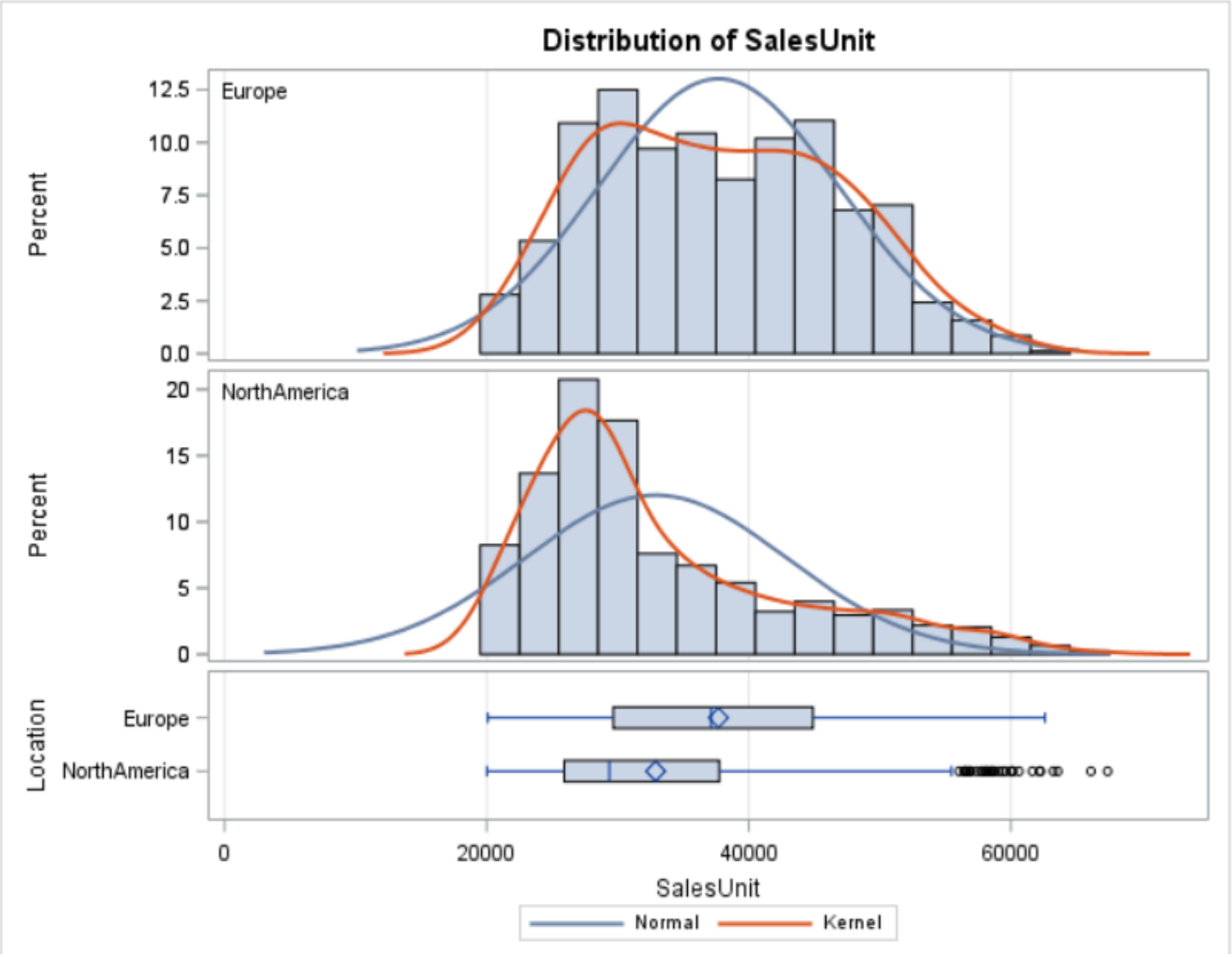
Location	N	Mean	Std Dev	Std Err	Minimum	Maximum
Europe	824	37675.1	9189.4	320.1	20082.0	62602.0
NorthAmerica	776	32886.5	9961.1	357.6	20016.0	67375.0
Diff (1-2)		4788.6	9571.4	478.8		

Location	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Europe		37675.1	37046.8	38303.5	9189.4	8766.1	9655.9
NorthAmerica		32886.5	32184.6	33588.5	9961.1	9489.0	10483.1
Diff (1-2)	Pooled	4788.6	3849.5	5727.7	9571.4	9250.8	9915.2
Diff (1-2)	Satterthwaite	4788.6	3847.2	5730.0			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1598	10.00	<.0001
Satterthwaite	Unequal	1567.2	9.98	<.0001

T- Test Result of Hypothesis (Sales Unit vs Location)

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	775	823	1.18	0.0227



Histogram of Sales Unit of Europe and North America

2. Price vs Location

Hypothesis Testing of Means : Price vs Location

The TTEST Procedure

Variable: Price

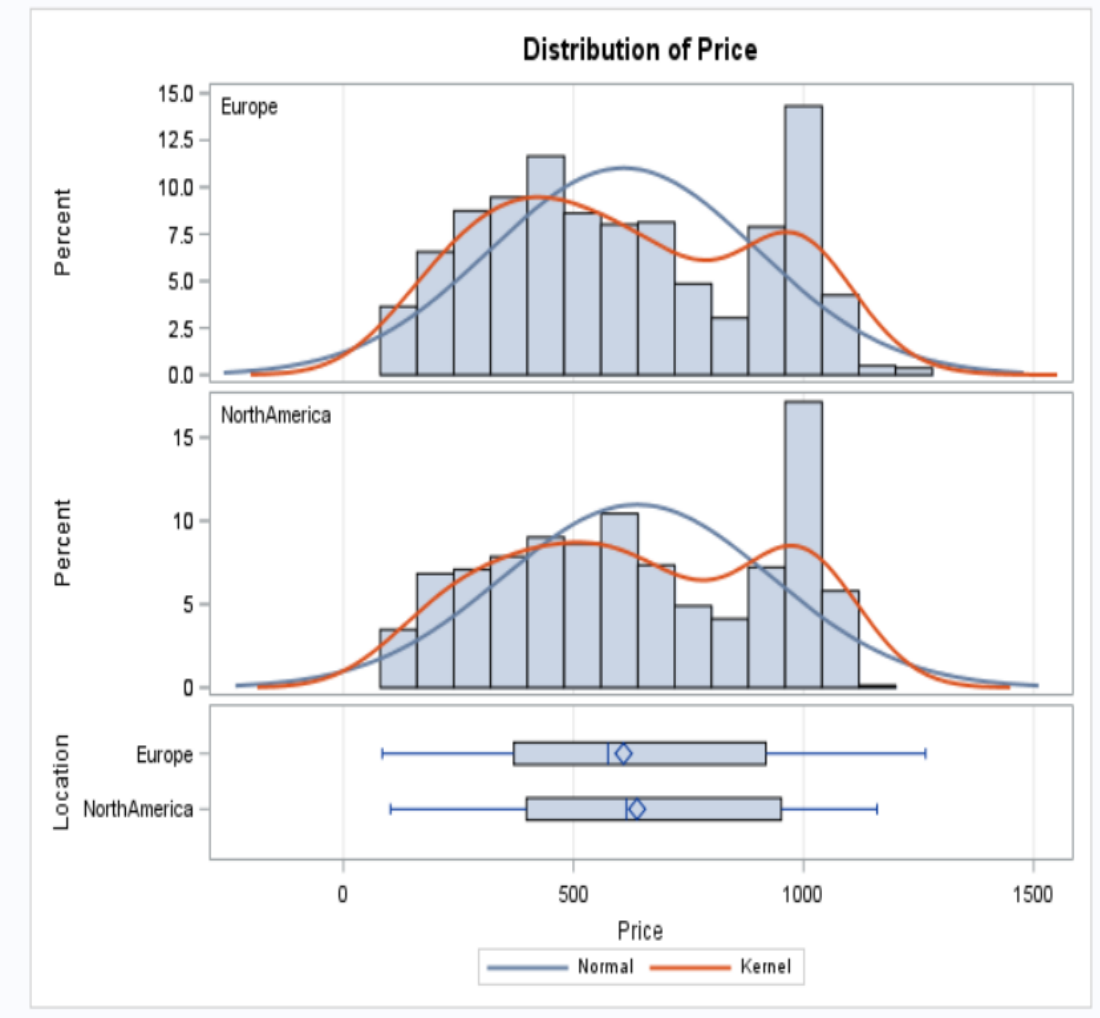
Location	N	Mean	Std Dev	Std Err	Minimum	Maximum
Europe	824	609.0	289.6	10.0895	84.5500	1265.2
NorthAmerica	776	638.1	290.9	10.4410	102.4	1159.8
Diff (1-2)		-29.1711	290.2	14.5176		

Location	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Europe		609.0	589.2	628.8	289.6	276.3	304.3
NorthAmerica		638.1	617.6	658.6	290.9	277.1	306.1
Diff (1-2)	Pooled	-29.1711	-57.6466	-0.6957	290.2	280.5	300.6
Diff (1-2)	Satterthwaite	-29.1711	-57.6503	-0.6919			

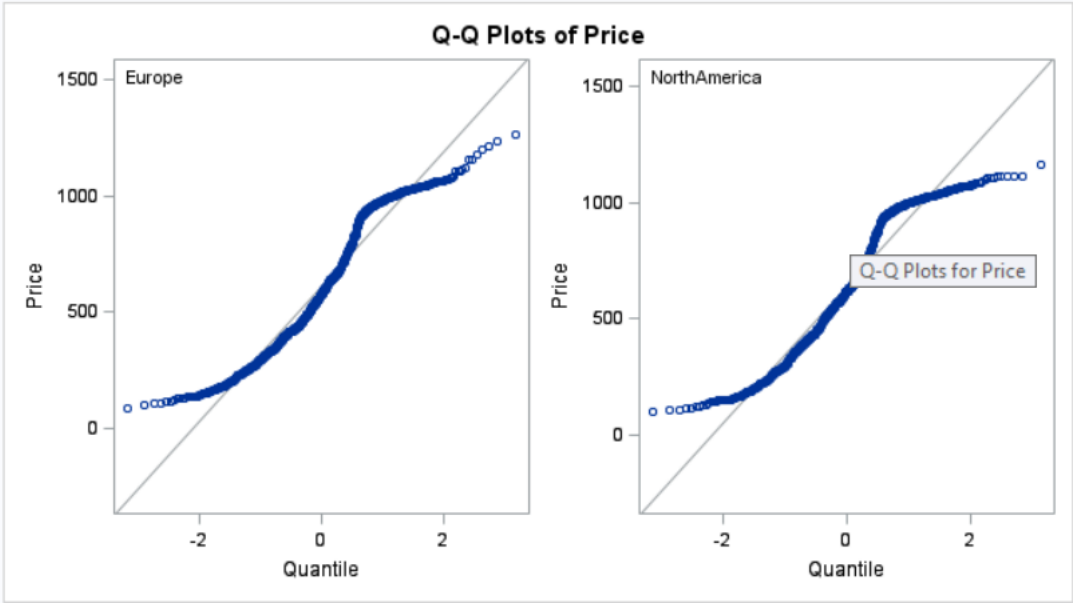
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1598	-2.01	0.0447
Satterthwaite	Unequal	1591.4	-2.01	0.0447

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	775	823	1.01	0.9041

T-Test results of Hypothesis (Price vs Location)



Histogram of Price of Europe and North America



Plots of Price on Location (North America and Europe)

3. Advertising Spending vs Location

Hypothesis Testing of Means : Advertising Spending vs Location

The TTEST Procedure

Variable: AdvertisingSpending

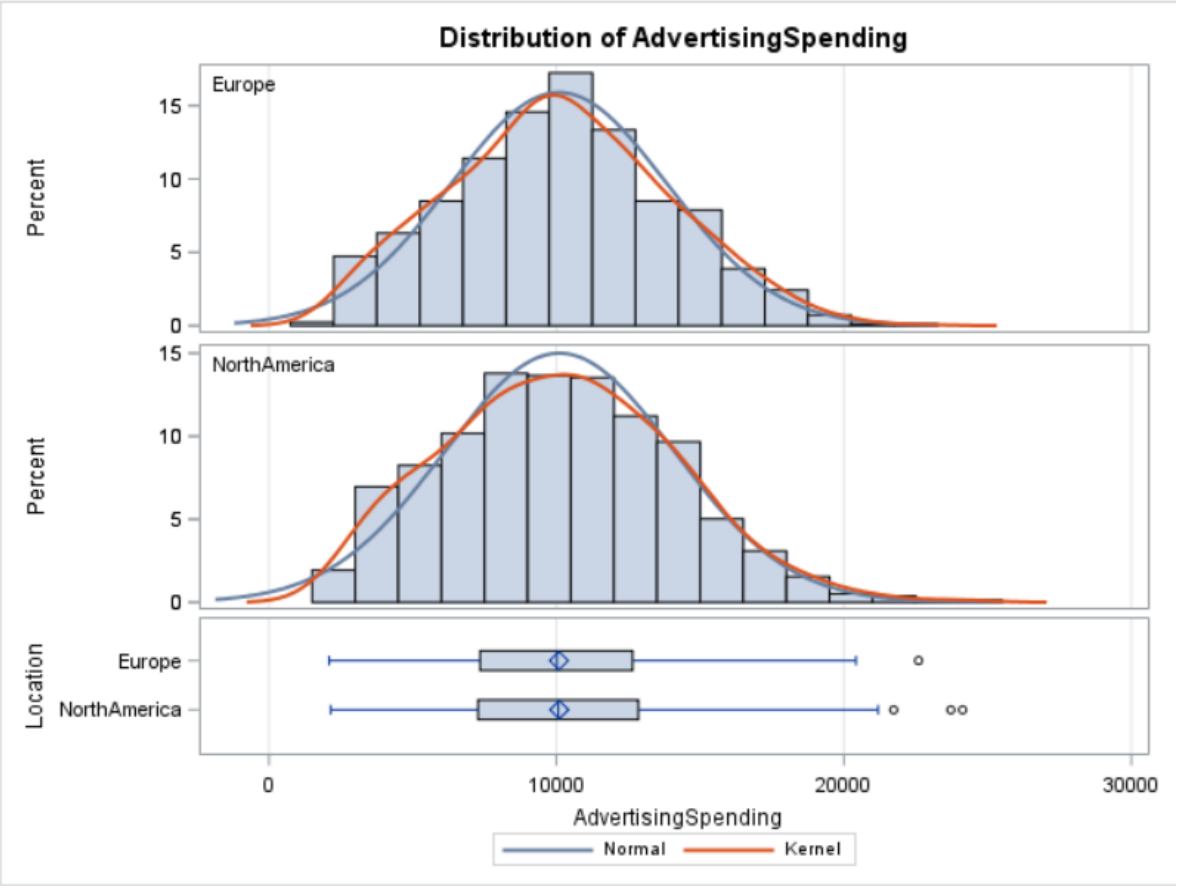
Location	N	Mean	Std Dev	Std Err	Minimum	Maximum
Europe	824	10085.3	3764.5	131.1	2089.3	22598.0
NorthAmerica	776	10098.5	3988.6	143.2	2145.5	24131.3
Diff (1-2)		-13.1961	3874.8	193.8		

Location	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Europe		10085.3	9827.9	10342.7	3764.5	3591.1	3955.6
NorthAmerica		10098.5	9817.4	10379.5	3988.6	3799.5	4197.6
Diff (1-2)	Pooled	-13.1961	-393.4	367.0	3874.8	3745.0	4014.0
Diff (1-2)	Satterthwaite	-13.1961	-394.0	367.6			

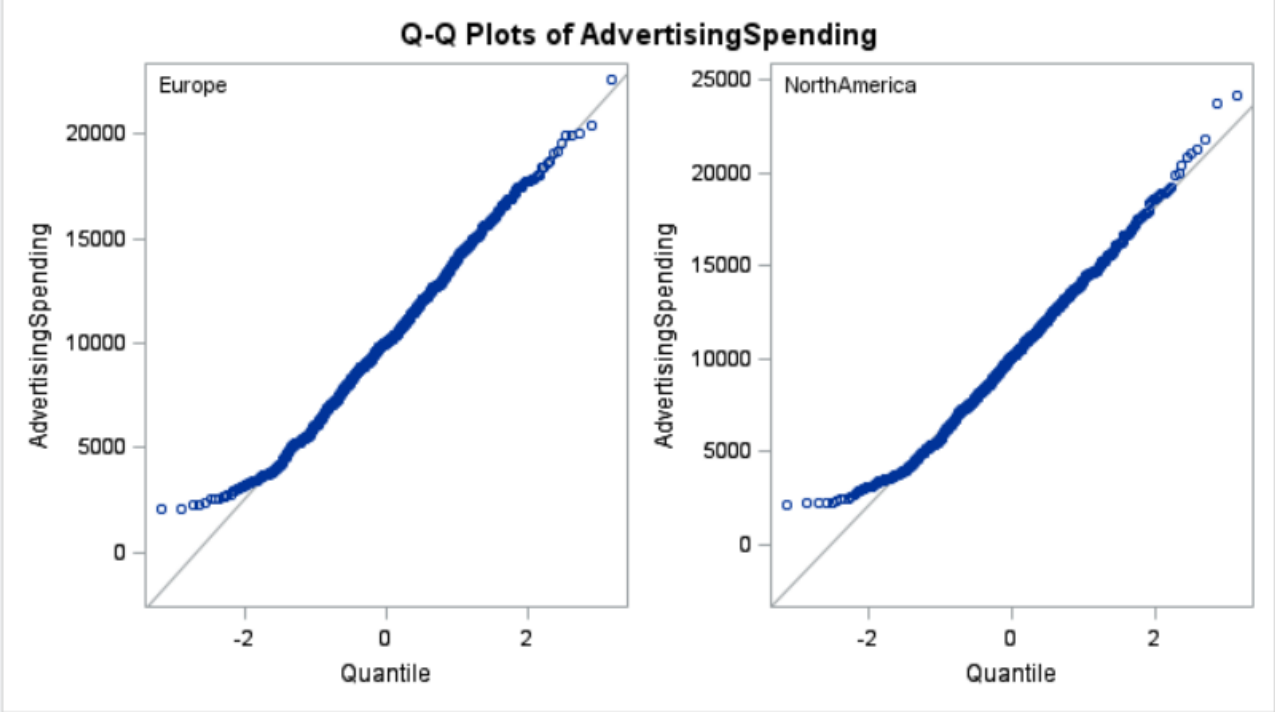
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1598	-0.07	0.9457
Satterthwaite	Unequal	1576.2	-0.07	0.9458

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	775	823	1.12	0.1022

T-Test Results of Hypothesis (Advertising Spending vs Location)



Histogram of Advertising Spending of Europe and North America



Plots of Advertising Spending on Location (Europe and North America)

Conclusion:

1. For sales unit: Enough evidence to accept that Mean of Europe Union is higher than North America
2. For Price: Not Enough evidence to accept that two means across location are different.
3. For Advertising Spending: Not enough evidences to accept that the mean of North America and Europe Union are significantly different.

Problem 13: Provide a hypothesis testing to determine any significant difference of Sales Unit, based on Salesforce experience level (Hint: Proc glm). Use the ods graphics to provide figures of your analyses.

H0: Mean sales unit is equal for all salesforce experience
Ha: Mean sales unit is different for all salesforce experience.

Hypothesis Testing : Sales Unit vs Salesforce Experience level

The GLM Procedure

Class Level Information		
Class	Levels	Values
SalesforceExperience	3	Average High Low

Number of Observations Read	1600
Number of Observations Used	1600

Hypothesis Testing : Sales Unit vs Salesforce Experience level

The GLM Procedure

Dependent Variable: SalesUnit

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	26340083401	13170041701	162.76	<.0001
Error	1597	129220746649	80914681.684		
Corrected Total	1599	155560830051			

R-Square	Coeff Var	Root MSE	SalesUnit Mean
0.169323	25.44437	8995.259	35352.65

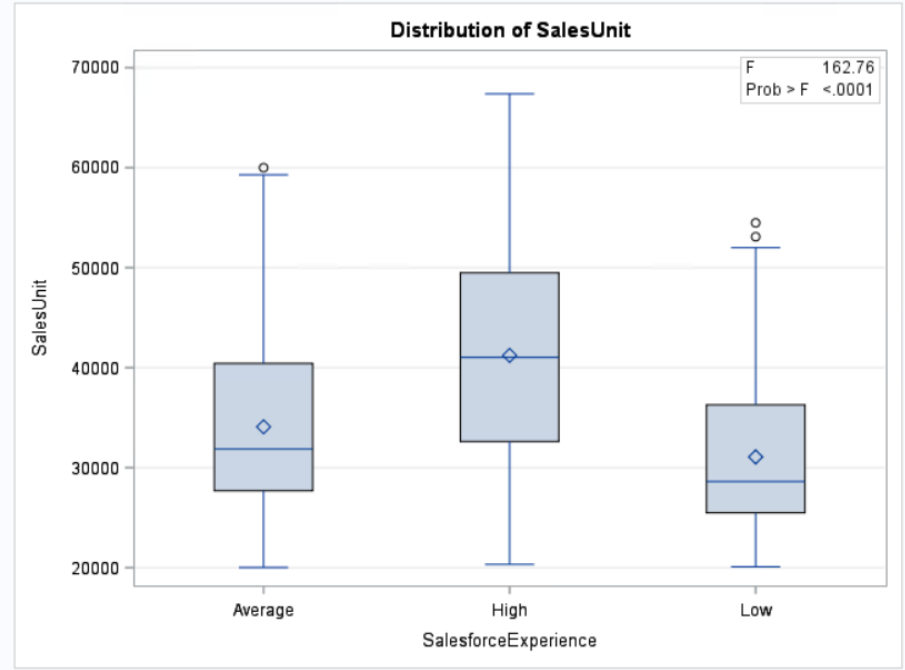
Source	DF	Type I SS	Mean Square	F Value	Pr > F
SalesforceExperience	2	26340083401	13170041701	162.76	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SalesforceExperience	2	26340083401	13170041701	162.76	<.0001

GLM Procedure of Hypothesis (SalesUnit vs Experienced Level)

The degrees of freedom (DF) column should be used to check the analysis results. The model degrees of freedom for a one-way analysis of variance are the number of levels minus 1; in this case, 3-1=2 The

Corrected Total degrees of freedom are always the total number of observations minus one; in this case 1600-1=1599. The sum of Model and Error degrees of freedom equal the Corrected Total.



Box Plot of Sales Unit (Categories Average, High, Low)

The overall F test is significant as F-value = 162.76 and P<0.0001 indicating that the model as a whole accounts for a significant portion of the variability in the dependent variable

Conclusion: We have sufficient evidence to accept that Mean sales unit is different for all salesforce experience.

Problem 14: Provide a hypothesis testing to determine any significant difference of Sales Unit, based on the 4-types of product. Which types of product are more popular (i.e., products have with higher average sales)? (Hint: Proc glm). Use the ods graphics to provide figures of your analyses.

H0: Mean Sales Unit is equal for all product types
Ha: Mean Sales Unit is different for all product type

There is strong evidence that mean Sales Unit differs among products. Normal design with high quality sells best, followed by luxury design with normal quality.

Hypothesis Testing : Sales Unit vs Design & Quality

The GLM Procedure

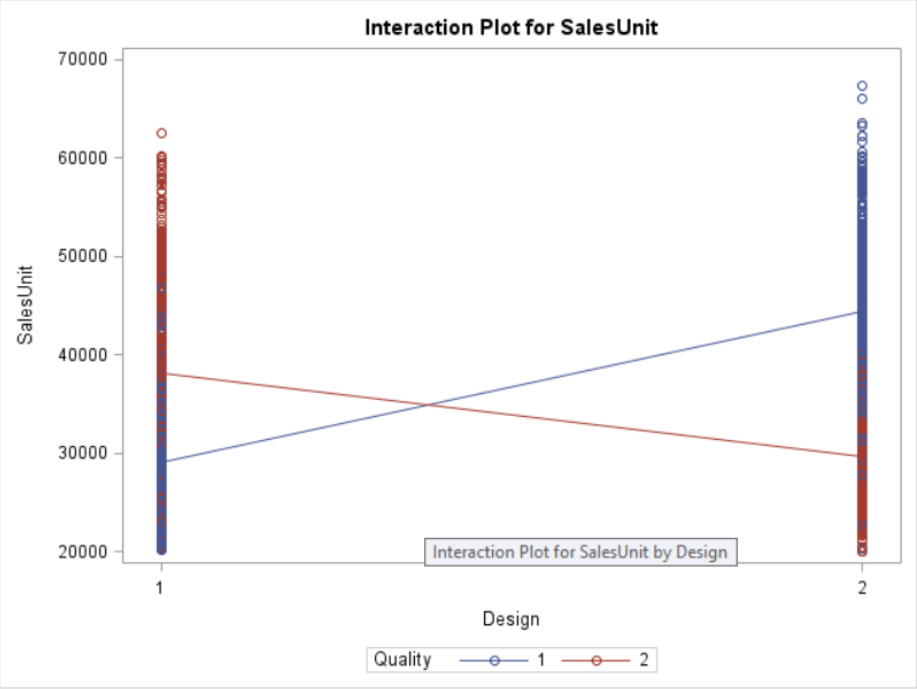
Dependent Variable: SalesUnit

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			

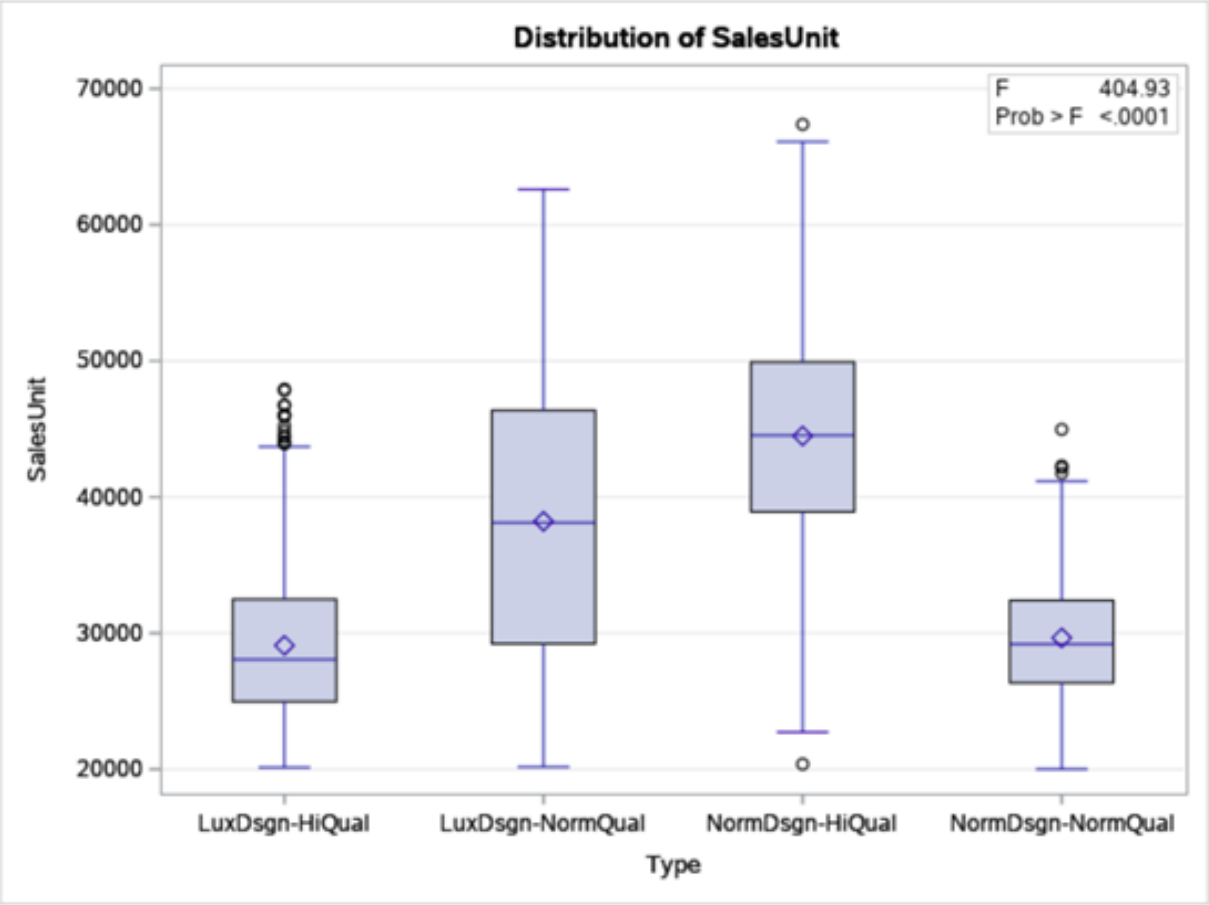
R-Square	Coeff Var	Root MSE	SalesUnit Mean
0.432191	21.04325	7439.347	35352.65

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Design	1	5536518924	5536518924	100.04	<.0001
Quality	1	4494513011	4494513011	81.21	<.0001
Design*Quality	1	57200970573	57200970573	1033.56	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Design	1	4649598124	4649598124	84.01	<.0001
Quality	1	3260861736	3260861736	58.92	<.0001
Design*Quality	1	57200970573	57200970573	1033.56	<.0001



The overall F test is significant as F-value = 404.93 and P<0.0001 indicating that the model as a whole accounts for a significant portion of the variability in the dependent variable



Box Plot of Sales Unit on the basis of Design Type

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

Normal design high quality product is most popular as per “products have with higher average sales”.

Conclusion: We have sufficient evidence to accept that Mean Sales Unit is different for all product type

Problem 15: Provide a hypothesis testing to determine any significant difference of Sales Unit, based on the 4-types of product & the consumer's location. Which types of product are more popular among Europeans (i.e., products have with higher average sales)? Which types of product are more popular among Americans? (Hint: Proc glm). Use the ods graphics to provide figures of your analyses.

Hypothesis Testing : Sales Unit vs Design & Quality & Location

The GLM Procedure

Class Level Information		
Class	Levels	Values
Design	2	1 2
Quality	2	1 2
Location	2	Europe NorthAmerica

Number of Observations Read	1600
Number of Observations Used	1600

Hypothesis Testing : Sales Unit vs Design & Quality & Location

The GLM Procedure

Dependent Variable: SalesUnit

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	92876913907	13268130558	336.97	<.0001
Error	1592	62683916144	39374319.186		
Corrected Total	1599	155560830051			

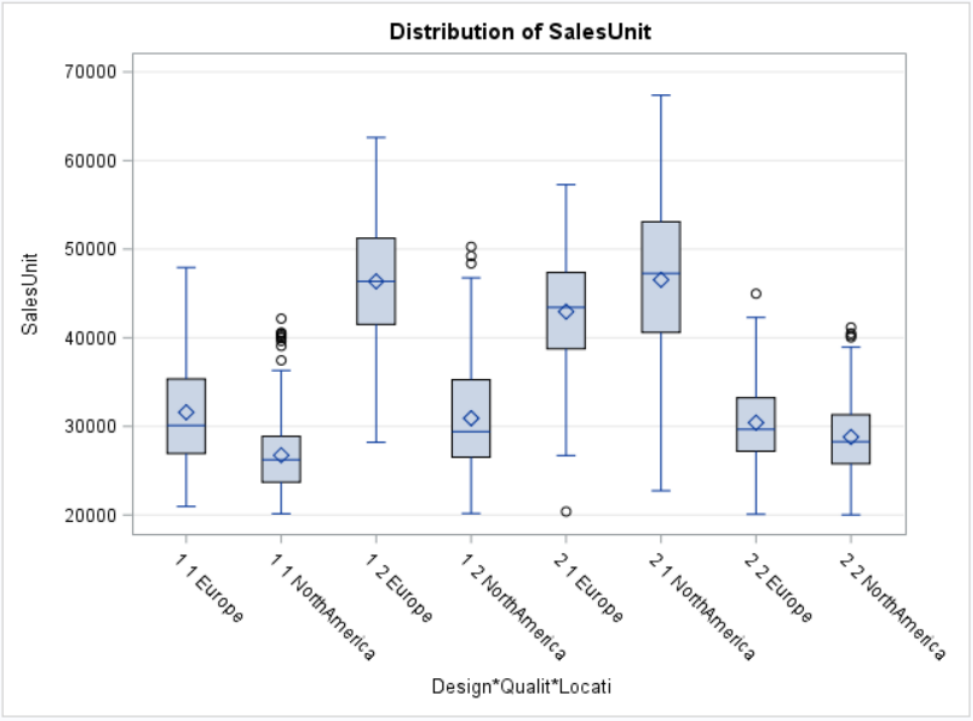
R-Square	Coeff Var	Root MSE	SalesUnit Mean
0.597046	17.74944	6274.896	35352.65

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Design	1	5536518924	5536518924	140.61	<.0001
Quality	1	4494513011	4494513011	114.15	<.0001
Design*Quality	1	57200970573	57200970573	1452.75	<.0001
Location	1	7170762905	7170762905	182.12	<.0001
Design*Location	1	11702942438	11702942438	297.22	<.0001
Quality*Location	1	6049836985	6049836985	153.65	<.0001
Design*Qualit*Locati	1	721369070	721369070	18.32	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Design	1	4249496096	4249496096	107.93	<.0001
Quality	1	3175770059	3175770059	80.66	<.0001
Design*Quality	1	59947251910	59947251910	1522.50	<.0001
Location	1	8315231236	8315231236	211.18	<.0001
Design*Location	1	12297451734	12297451734	312.32	<.0001
Quality*Location	1	6195516075	6195516075	157.35	<.0001
Design*Qualit*Locati	1	721369070	721369070	18.32	<.0001

Hypothesis Testing : Sales Unit vs Design & Quality & Location

The GLM Procedure



Box Plot of Sales Unit based on Design, Quality and Location

Level of Design	Level of Quality	Level of Location	N	SalesUnit	
				Mean	Std Dev
1	1	Europe	196	31591.7908	6510.15564
1	1	NorthAmerica	208	26742.9760	4371.77351
1	2	Europe	172	46365.4070	6933.93878
1	2	NorthAmerica	192	30908.7969	6598.68278
2	1	Europe	240	42945.8625	6381.05003
2	1	NorthAmerica	180	46541.0444	9243.87154
2	2	Europe	216	30418.7685	4820.86871
2	2	NorthAmerica	196	28803.6684	4472.81149

Conclusion: Type of product popular among:
Europe : Luxury Design Normal Quality
America : Normal Design High Quality

Problem 16: 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?

Key Findings:

- Europeans show a preference for luxury design with normal quality whereas Americans prefer normal design with luxury quality.
- European stores show higher average sales than American stores.
- Americans are more price sensitive than Europeans. It may be worth directing resources for promotions more heavily into the American market.
- Sales of this product appear to differ across seasons. Summer and fall show the strongest sales. Winter sales are slightly lower and spring sales are lowest of all.
- Stores with higher salesforce experience are seeing a higher sales volume.
- A histogram of stores' units sold shows that most stores' sales are normally distributed but there is a distinct pack of stores that are selling an exceptionally high volume. These top sellers should be investigated further.

Recommendation: Based on our findings, we recommend production of for luxury design with normal quality for sale in Europe and normal design with high quality for sale in North America. If only one product can be made, we recommend normal design with high quality since it has the highest sales overall.

Code

```
/*************************************************************
 *   PRODUCT:   Homework 1
 *   VERSION:   1.0
 *   CREATOR:   Group 7
 *   DATE:      05FEB19
 *************************************************************/



/*Importing File into SAS*/


PROC IMPORT OUT= LEC3.SALES
            DATAFILE= "H:\My SAS Files\HW1\Sales.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;


/*Ques 1 : Type of Each Variables

Quantative Variables : SalesUnit | Price | AdvertisingSpending

Qualitative Variables : Design | Quality | Promotion | SalesForceExperience | Season |
Location


*/


/* Ques 2 : How many stores are there in the data? what are their frequencies in Europe
and North America?  */


/* For finding the total number of stores */
 proc sql;

 select count(distinct ID) as Number_of_Stores
 from
 lec3.sales;
 title 'Number of Stores';

quit;

run;

/* For finding the number of stores Location wise */
proc sql;

select Location,count(distinct ID) as Number_of_Stores
from
lec3.sales
group by Location;

title 'Number of Stores Location Wise';
```

```
quit;

run;

/*Ques 3 : What is the average Sales Unit, Price, and Advertising Spending? */

proc sql;

select avg(SalesUnit) as Sales_Unit_Average,
            avg(Price) as Price_Average,
            avg(AdvertisingSpending) as Adv_Spend_Average
from
lec3.sales;

title 'Average of Quantative Variables';

quit;

run;

proc means data= lec3.Sales mean;
var SalesUnit Price AdvertisingSpending;
 title 'Average of Quantative Variables';
run;

/*Ques 4: ? Plot histograms for Sales Unit, Price, and Advertising Spending.
Your histograms should have the best normal and kernel fitted curve.
For Sales Unit, Price, and Advertising Spending use binwidths 1000, 10, 100 respectively.
Which of variables does look different from a normal distribution?  */

/* Plotting Histogram for Sales */

proc sgplot data= lec3.Sales;
 histogram SalesUnit / binwidth = 1000 ;
 density SalesUnit / type = kernel;
 density SalesUnit/type = normal;
 title 'Sales Report';
run;


/* Plotting Histogram for Price */

proc sgplot data= lec3.Sales;
 histogram Price / binwidth = 10 ;
 density Price / type = kernel;
 density Price/type = normal;
 title 'Price Report';
run;

/* Plotting Histogram for Advertising Spend */

proc sgplot data= lec3.Sales;
 histogram AdvertisingSpending / binwidth = 100 ;
 density AdvertisingSpending / type = kernel;
 density AdvertisingSpending/type = normal;
 title 'Advertising Spend Report';
run;
```

```
/* The Sales Unit and Price does little diffrent from the Normal Curve.
Sales Unit Histogram is a Right Skewed distribution and the Price Histogram looks like a
histogram with two peaks. */

/* Plotting Histogram for all the three metrics : using Normal curve

proc univariate data=lec3.Sales;
histogram SalesUnit Price AdvertisingSpending /normal;
run;
*/


/* Ques 5: Find the correlation matrix of the Sales Unit, Price, and Advertising
Spending?
Report both Pearson and Spearman Correlation tests. How can you describe the effects of
Price and Advertising Spending on Sales Unit? */

proc corr data=lec3.Sales Spearman;
var SalesUnit Price AdvertisingSpending;
title 'Pearson and Spearman Correlation for Sales Unit Price and Advertising Spend';
run;

/* Price is having negative correlation of -0.25 with the Sales Unit.As Price Increase
Sales unit decreases.
   Advertising Spending is also having a negative correlation effect on Sales Unit. But
it is not significant */


/* Ques 6 : Use Proc sgscatter to draw a matrix of Sales Unit, Price, Advertising
Spending relations.
Does the scatter plot of Sales Unit based on Advertising Spending confirm your result in
Q5? If not, why? */

proc sgscatter data= lec3.Sales;
 matrix SalesUnit Price AdvertisingSpending / diagonal= (histogram);
 title 'Relation between SalesUnit Price Advertising Spend';
run;

/* The Scatter Plot of the matrix between Price and Sales Unit seems to support our
previous observation of descrease in Price as Sales Unit increases
Not much can be infered from the Advertising Spend and Price plot */


/*Ques 7 : Use Box Plotting figures to determine the average people preference on Design
and Quality level?
Then, redo your analysis based on people located in Europe vs. North America.
Do you find any evidence on a different taste of Design and Quality among Europeans vs.
Americans? */

/* Doesn't understand the requirement */

proc means data=lec3.Sales sum;
  class ID;
  var SalesUnit;
  output out=Store_Sales_Totals sum=SalesUnit;
  run;
DATA Store_Sales_Totals ;
Set Store_Sales_Totals;
keep ID SalesUnit; *removes extra columns;
```

```
if ID = '' then delete; *removes totals row from table so we now have 1 obs for each
store and nothing else;
run;

/*Create second dataset that keeps only the first instance of each ID (we don't need any
vars that differ across a particular ID so this is fine) */
proc sort data=lec3 out=WORK.Sales_w_type nodupkey;
BY ID;
run;
/*Then from the second dataset 1. drop vars that don't apply to the ID-grouped
observation, 2. add variable to this dataset that concatenates Design and Quality Type*/
DATA Sales_W_Type;
  SET Sales_W_Type;

  Length Type $ 18;
              IF (Design=1) and (Quality=1) THEN Type = "LuxDsgn-HiQual";
              IF (Design=2) and (Quality=1) THEN Type = "NormDsgn-HiQual";
              IF (Design=1) and (Quality=2) THEN Type = "LuxDsgn-NormQual";
              IF (Design=2) and (Quality=2) THEN Type = "NormDsgn-NormQual";
  drop AdvertisingSpending Price Promotion SalesforceExperience SalesUnit Season;
RUN;

/*Now merge the 2 datasets. this results in a dataset with 1 observation for each store
and variable describing each product type.*/
proc sort data=Store_Sales_Totals;
 by ID;
proc sort data=Sales_W_Type;
 by ID;
data Full_Year_Data;
 merge Store_Sales_Totals (in= in_1) Sales_W_Type (in= in_2);
 by ID;
 if in_1 AND in_2;
run;

/* Taking the summary statistics of the data by Type*/
 proc means data= Full_Year_Data n mean stddev min p25 median p75 max maxdec= 2;
 var salesunit;
 title 'Summary Statistics';
 by Type;
run;

/* Printing the summary statistics of the data by Type*/
proc print data=Full_Year_Data;
 title 'Full Year Data';
run;

/* Box Plot by Type all locations */
/*Interpret as: stores carrying this type tended to sell fewer units than of other
types*/
proc sgplot data= Full_Year_Data;
 hbox SalesUnit  / category= Type;
 title 'Sales Performance of Product Types';
run;

/* Box Plot by Type for each location*/
proc sgplot data= Full_Year_Data;
 hbox SalesUnit  / category= Type group=Location;
 title 'Sales Performance of Product Types - Europe vs. US';
run;
```

```
/*Ques 8 : Use Scatter Plot figures to determine Sales Unit based on Salesforce
experience level.
Do find any clear evidence about Salesforce experience level? Then, redo your analysis
based on the 4-types of products?
Does higher Salesforce experience provide any special ability to sell more products? If
yes, which type of products?  */

proc sgplot data = lec3.Sales;
 scatter Y=SalesUnit X=SalesforceExperience;
 title 'Scatter Plot for Sales Unit based on Experience Level ';
run;

proc sgpanel data = lec3.Sales;
 panelby Design Quality;
 scatter Y=SalesUnit X=SalesforceExperience;
 title 'Scatter Plot for Sales Unit on Experience Level based on products ';
run;

/* The scatter plot provides clear evidence about increase in SaleUnit with increase in
Sales Force Experience.
No, higher Sales Force experience provide an special ability to sell products.
The products with High Design and Quality along with products with low Design and
Quality. */


/* Ques 9 : Do question 8 again in terms of Sales Unit based on Promotion.
Is there any difference between Europeans vs. Americans to response to promotion?
Who is more price sensitive: Europeans or Americans? */

proc sgplot data = lec3.Sales;
 scatter Y=SalesUnit X=Promotion;
 title 'Scatter Plot for Sales Unit based on Promotion ';
run;

proc sgpanel data = lec3.Sales;
 panelby Location;
 scatter Y=SalesUnit X=Promotion;
 title 'Scatter Plot for Sales Unit on Promotions based on Region ';
run;

/* Yes, there is diffrence between Europeans vs American in response to Promotion.
North America is price sensitive based on Sales unit */

/*Ques 10 : Provides a summary statistic of Sales unit, price, and advertising spending
by seasons?
Do the Sales Unit, Price, and Advertising Spending have different means across seasons?
*/

proc sort data = lec3.Sales;
by season;
run;

proc tabulate data= lec3.Sales ;
 var SalesUnit Price AdvertisingSpending;
 table SalesUnit Price AdvertisingSpending,N Mean StdDev Min p25 Median p75 Max;
 by Season;
```

```
 title 'Summary Statistics by Season';
run;

/* Yes, Sales Unit, Price and Advertising Spend is having diffrent means across all the
seasons.
But they are close to each other respectively. */


/* Ques 11: Provide a hypothesis testing to determine which of Sales Unit, Price,
and Advertising Spending have significantly different means across seasons? (Hint: ANOVA)
*/

/*Sales Unit */

ods graphics on;
proc anova data=lec3.Sales;
class Season;
model SalesUnit = Season;
means Season;
title 'Hypothesis Testing of Means : Sales Unit vs Season';
run;
ods graphics off;

/*Price */

ods graphics on;
proc anova data=lec3.Sales;
class Season;
model Price = Season;
means Season;
title 'Hypothesis Testing of Means : Price vs Season';
run;
ods graphics off;

/*Advertising Spend */

ods graphics on;
proc anova data=lec3.Sales;
class Season;
model AdvertisingSpending = Season;
means Season;
title 'Hypothesis Testing of Means : Advertising Spend vs Season';
run;
ods graphics off;


/* Ques 12: Provide a hypothesis testing to determine which of Sales Unit, Price, and
Advertising Spending have
significantly different means across location, i.e., North America vs. Europe? (Hint: T-
test) */

/*Doubts : Interpretation is pending


/*Sales Unit */

ods graphics on;
proc ttest data=lec3.Sales;
class Location;
```

```
 var SalesUnit;
 title 'Hypothesis Testing of Means : Sales Unit vs Location';
run;
ods graphics off;

/*Price */

ods graphics on;
proc ttest data=lec3.Sales;
class Location;
 var Price;
 title 'Hypothesis Testing of Means : Price vs Location';
run;
ods graphics off;

/*Advertising Spend */

ods graphics on;
proc ttest data=lec3.Sales;
class Location;
 var AdvertisingSpending;
 title 'Hypothesis Testing of Means : Advertising Spending vs Location';
run;
ods graphics off;


/*Ques 13 : Provide a hypothesis testing to determine any significant difference of Sales
Unit,
based on Salesforce experience level (Hint: Proc glm). Use the ods graphics to provide
figures of your analyses */

ods graphics on;
proc glm data=lec3.Sales;
class SalesForceExperience;
model SalesUnit=SalesForceExperience;
means Saleesforceexperience;
title 'Hypothesis Testing : Sales Unit vs Salesforce Experience level';
run;
ods graphics off;

/* Ques 14: Provide a hypothesis testing to determine any significant difference of Sales
Unit,
based on the 4-types of product. Which types of product are more popular
(i.e., products have with higher average sales)? (Hint: Proc glm).
Use the ods graphics to provide figures of your analyses */


ods graphics on;
proc glm data=lec3.Sales;
class Design Quality;
model SalesUnit=Design | Quality;
means Design*Quality;
title 'Hypothesis Testing : Sales Unit vs Design & Quality';
run;
ods graphics off;


/* Normal Design High Quality product are most popular based on Sales Unit */
```

```
/*Ques 15: Provide a hypothesis testing to determine any significant difference of Sales
Unit,
based on the 4-types of product & the consumer's location.
Which types of product are more popular among Europeans (i.e., products have with higher
average sales)?
Which types of product are more popular among Americans? (Hint: Proc glm).
Use the ods graphics to provide figures of your analyses */

ods graphics on;
proc glm data=lec3.Sales;
class Design Quality Location;
model SalesUnit=Design | Quality | Location;
means Design*Quality*Location;
title 'Hypothesis Testing : Sales Unit vs Design & Quality & Location';
run;
ods graphics off;

/* Europe : Luxury Design Normal Quality
      America : Normal Design High Quality */
```

Thank You for reading report.