EXPLORATORY DATA ANALYSIS –ADIDAS US SALES DATASET

Objective-

Primary-

To analyze and evaluate the sales performance of Adidas in the US market, identify key trends, challenges, and growth opportunities and Provide actionable insights that can guide strategic decisions to enhance market share, customer engagement, and profitability."

Secondary-

Predicting Sales and Profit Margin

Dataset Overview

Obs	Retailer	Retailer_ID	Invoice_Date	Region	State	City	Product	Price_per_Unit	Units_Sold	Total_Sales	Operating_Profit	Operating_Margin	Sales_Method
1	Foot Locker	1185732	01/01/2020	Northeast	New York	New York	Men's Street Footwear	$50.00	1,200	$600,000	$300,000	50%	In-store
2	Foot Locker	1185732	01/02/2020	Northeast	New York	New York	Men's Athletic Footwear	$50.00	1,000	$500,000	$150,000	30%	In-store
3	Foot Locker	1185732	01/03/2020	Northeast	New York	New York	Women's Street Footwear	$40.00	1,000	$400,000	$140,000	35%	In-store
4	Foot Locker	1185732	01/04/2020	Northeast	New York	New York	Women's Athletic Footwear	$45.00	850	$382,500	$133,875	35%	In-store
5	Foot Locker	1185732	01/05/2020	Northeast	New York	New York	Men's Apparel	$60.00	900	$540,000	$162,000	30%	In-store

Obs	Retailer	Retailer_ID	Invoice_Date	Region	State	City	Product	Price_per_Unit	Units_Sold	Total_Sales	Operating_Profit	Operating_Margin	Sales_Metho
9644	Foot Locker	1185732	01/24/2021	Northeast	New Hampshire	Manchester	Men's Apparel	$50.00	64	$3,200	$896	28%	Outlet
9645	Foot Locker	1185732	01/24/2021	Northeast	New Hampshire	Manchester	Women's Apparel	$41.00	105	$4,305	$1,378	32%	Outlet
9646	Foot Locker	1185732	02/22/2021	Northeast	New Hampshire	Manchester	Men's Street Footwear	$41.00	184	$7,544	$2,791	37%	Outlet
9647	Foot Locker	1185732	02/22/2021	Northeast	New Hampshire	Manchester	Men's Athletic Footwear	$42.00	70	$2,940	$1,235	42%	Outlet
9648	Foot Locker	1185732	02/22/2021	Northeast	New Hampshire	Manchester	Women's Street Footwear	$29.00	83	$2,407	$650	27%	Outlet

The Adidas US Sales Dataset demonstrates the distribution of 9648 observation which are categorized into 13 variables out of which 7 are categorical and 6 are numerical which are as follows

Categorical Variables	Numerical Variables
Retailer	Operating Profit
Retailer ID	Operating Margin
Region	Price Per Unit
State	Units Sold
City	Total Sales
Product	Invoice Date
Sales Method	

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
6	City	Char	14	$14.	$14.	City
3	Invoice_Date	Num	8	MMDDYY10.		Invoice Date
12	Operating_Margin	Num	8	PERCENT12.		Operating Margin
11	Operating_Profit	Num	8	NLMNY15.		Operating Profit
8	Price_per_Unit	Num	8	NLMNY15.2		Price per Unit
7	Product	Char	25	$25.	$25.	Product
4	Region	Char	9	$9.	$9.	Region
1	Retailer	Char	13	$13.	$13.	Retailer
2	Retailer_ID	Num	8	BEST.		Retailer ID
13	Sales_Method	Char	8	$8.	$8.	Sales Method
5	State	Char	14	$14.	$14.	State
10	Total_Sales	Num	8	NLMNY15.		Total Sales
9	Units_Sold	Num	8	COMMA15.		Units Sold

Observations	9648
Variables	13
Indexes	0
Observation Length	144
Deleted Observations	0
Compressed	NO
Sorted	NO

The MEANS Procedure

Variable	Label	N	N Miss
Retailer_ID	Retailer ID	9648	0
Invoice_Date	Invoice Date	9648	0
Price_per_Unit	Price per Unit	9648	0
Units_Sold	Units Sold	9648	0
Total_Sales	Total Sales	9648	0
Operating_Profit	Operating Profit	9648	0
Operating_Margin	Operating Margin	9648	0

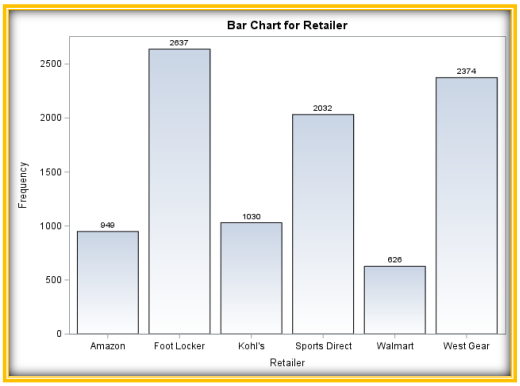
The SAS System

city_n	product_n	Region_n	Retailer_n	Retailer_ID_n	salesm_n	State_n
9648	9648	9648	9648	9648	9648	9648

UNIVARIATE ANALYSIS- Categorical Variables

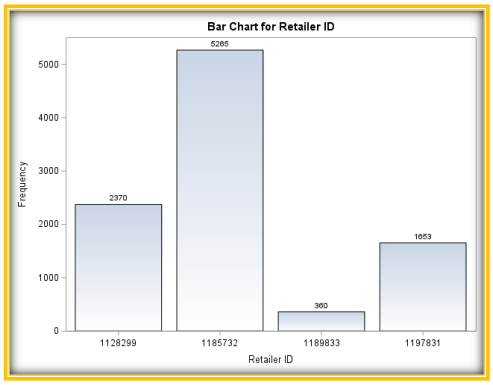
1.Retailer

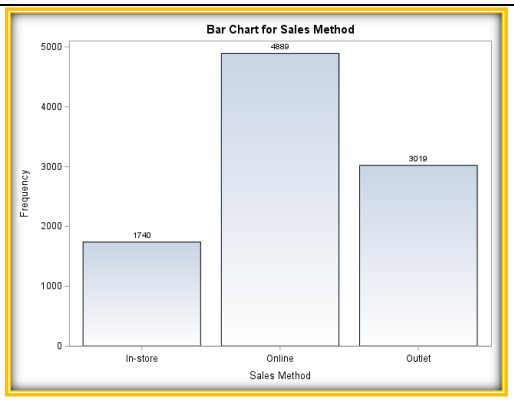
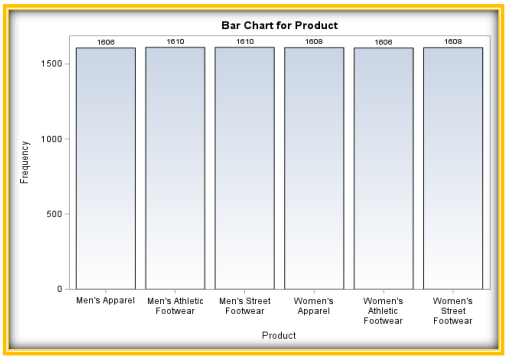
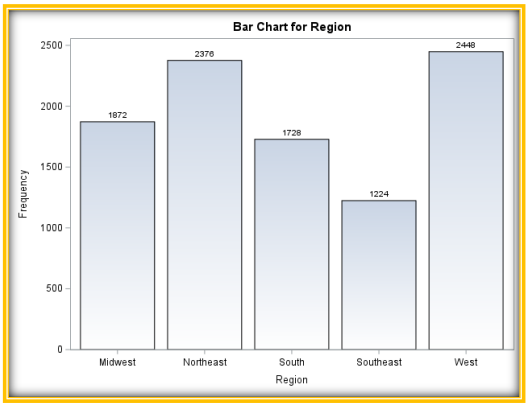
- It is the name of the Retailer that made the sale
- Summarisation- Proc Freq
- Visualisation- Bar Chart
- The Data Reveals out of Total 9648 observation
- **Amazon-949-9.84%**
- **Foot Locker-2637-27.33%**
- **Kohl's-1030-10.68%**
- **Sports Direct-2032-21.06%**
- **Walmart-626-6.49%**
- **West Gear -2374-24.61%**

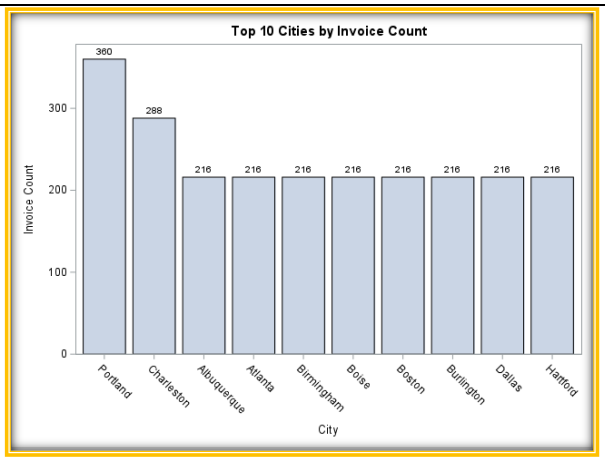
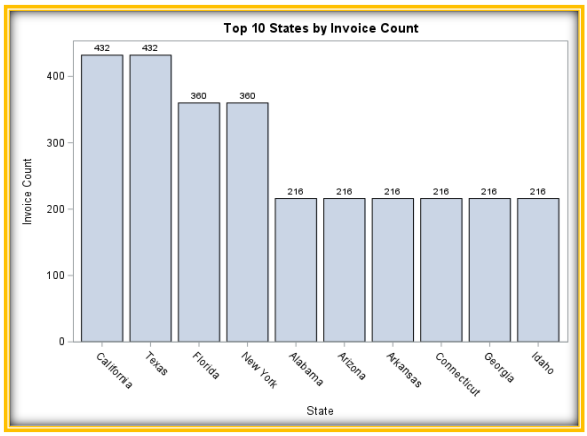


2 . Retailer ID

- The Analysis focuses on distribution of the variable ''Retailer ID' within the dataset that means unique identifier for the Retailer.
- Summarisation- Proc Freq
- Visualisation- Bar Chart
- The data reveals out of 9648 observations
- ID-1128299-2370-24.56%
- ID-1185732-5265-54.57%
- ID-1189833-360-3.73%
- ID-1197831-1653-17.13%



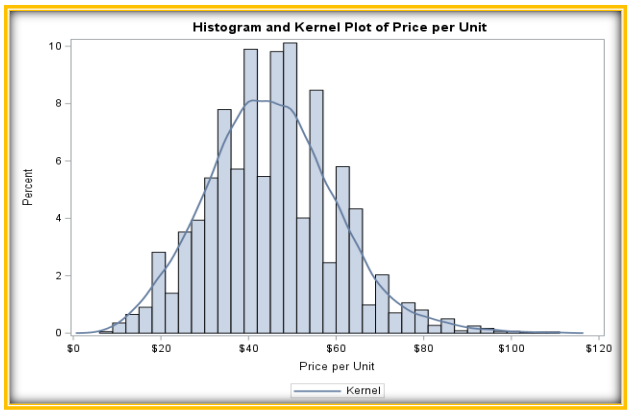
3. Sales Method • The Analysis focuses on distribution of the variable 'Sales Method' within the dataset that means the method by which the product was sold. • Summarisation- Proc Freq • Visualisation- Bar Chart • The data reveals out of 9648 observations • In-Store-1740-18.03 • Online-4889-50.67% • Outlet-3019-31.29%	
4. Product- • The Analysis focuses on distribution of the variable 'Product' within the dataset that means the type of product sold. • Summarisation- Proc Freq • Visualisation- Bar Chart • The data reveals out of total 9648 observations • Mens Apparel- 1606-16.65% • Men's Athletic Footwear -1610-16.69% • Men's Street Footwear- 1610-16.69% • Women's Apparel- 16.08-16.67% • Women's Athletic Footwear-1606-16.65% • Women's Street Footwear -1608-16.67%	
5. Region- • The Analysis focuses on distribution of the variable 'Region within the dataset that means the geographic region in which the sales took place • Summarisation- Proc Freq • Visualisation- Bar Chart • The data reveals out of total 9648 observations • Midwest-1872-19.4 • Northeast- 2376-24.63 • South- 1728-17.91% • Southeast-1224-12.69% • West-2448-25.37%	
6. City- • The Analysis focuses on distribution of the variable 'City' within the dataset that means the specific city where the sales took place	

• Summarisation- Proc Freq • Visualisation- Bar Chart • The data reveals distribution of total 9648 observations among 52 distinct cities in US • Top ones are • Portland-360-3.73% • Charleston-288-2.99%	
7. State- • The Analysis focuses on distribution of the variable 'State' within the dataset that means the specific State where the sales took place • Summarisation- Proc Freq • Visualisation- Bar Chart • The data reveals distribution of total 9648 observations among 50 distinct States in US • Top ones are • California-432-4.48% • Texas-432-4.48%	

UNIVARIATE ANALYSIS- Numerical Variables

Variable	Label	N	N Miss	Variance	Std Dev	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean	Mean	Sum	Minimum	Maximum
Retailer_ID	Retailer ID	9648	0	694869491.89	26360.38	2.25	1173323.66	1174375.78	1173849.72	11325302133	1128299.00	1197831.00
Invoice_Date	Invoice Date	9648	0	27599.40	166.13	0.74	22407.32	22413.95	22410.64	216217849.00	21915.00	22645.00
Price_per_Unit	Price per Unit	9648	0	216.25	14.71	32.52	44.92	45.51	45.22	436250.00	7.00	110.00
Units_Sold	Units Sold	9648	0	45903.93	214.25	83.39	252.65	261.21	256.93	2478861.00	0.00	1275.00
Total_Sales	Total Sales	9648	0	20140155804	141916.02	152.15	90441.29	96105.58	93273.44	899902125.00	0.00	825000.00
Operating_Profit	Operating Profit	9648	0	2936893573.9	54193.11	157.42	33343.74	35506.75	34425.24	332134761.45	0.00	390000.00
Operating_Margin	Operating Margin	9648	0	0.01	0.10	22.98	0.42	0.42	0.42	4081.02	0.10	0.80

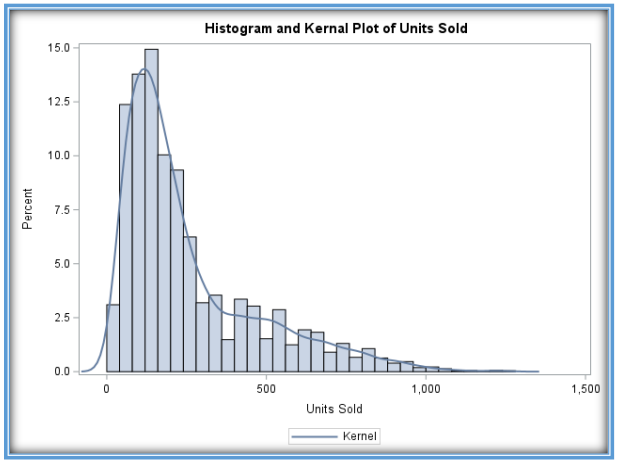
1 .Price Per Unit



Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.050967	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.590678	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	16.29257	Pr > A-Sq	<0.0050

- The Analysis Focuses on distribution of the continuous variable ' Price Per Unit'
- Summarisation- Proc Means,Proc Univariate
- Visualisation- Histogram and Density Plot
- Summary estimate is as shown in the image-'The Means Procedure'
- Kolmogorov-Smirnov Test of Normality shows variable is not normally distributed
- Histogram and Kernal plot displays that distribution is not uniform.

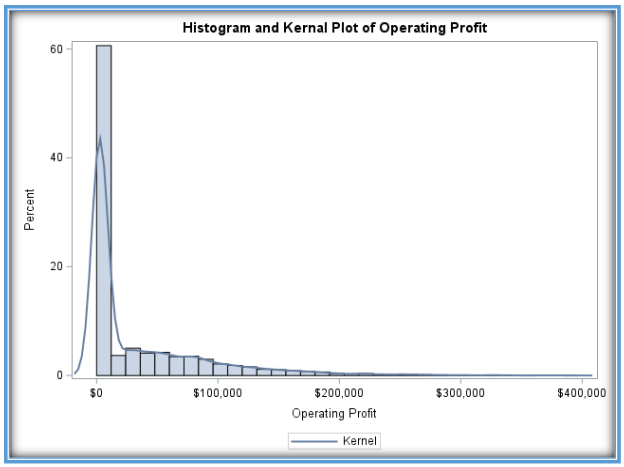
2 .Units Sold



Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.180815	Pr > D	<0.0100
Cramer-von Mises	W-Sq	95.50453	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	528.2038	Pr > A-Sq	<0.0050

- The Analysis Focuses on distribution of the continuous variable ' Units Sold
- Summarisation- Proc Means,Proc Univariate
- Visualisation- Histogram and Density Plot
- Summary estimate is as shown in the image-'The Means Procedure'
- Kolmogorov-Smirnov Test of Normality shows variable is not normally distributed
- Histogram and Kernal plot displays that distribution is not uniform.

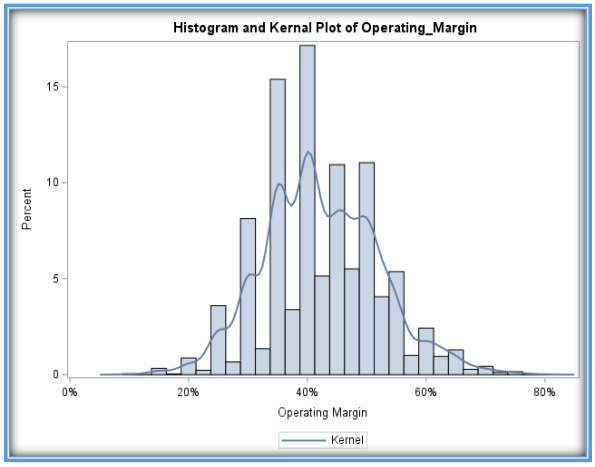
3. Operating Profit



Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.297937	Pr > D	<0.0100
Cramer-von Mises	W-Sq	217.9787	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1142.555	Pr > A-Sq	<0.0050

- The Analysis Focuses on distribution of the continuous variable 'Operating Profit'
- Summarisation- Proc Means,Proc Univariate
- Visualisation- Histogram and Density Plot
- Summary estimate is as shown in the image-'The Means Procedure'
- Kolmogorov-Smirnov Test of Normality shows variable is not normally distributed
- Histogram and Kernal plot displays that distribution is not uniform.

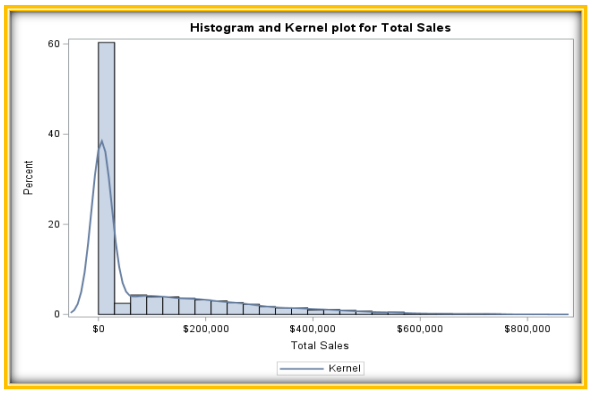
4. Operating Margin



Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.083958	Pr > D	<0.0100
Cramer-von Mises	W-Sq	7.251366	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	40.72366	Pr > A-Sq	<0.0050

- The Analysis Focuses on distribution of the continuous variable 'Operating Margin
- Summarisation- Proc Means,Proc Univariate
- Visualisation- Histogram and Density Plot
- Summary estimate is as shown in the image-'The Means Procedure'
- Kolmogorov-Smirnov Test of Normality shows variable is not normally distributed
- Histogram and Kernal plot displays that distribution is not uniform.

5. Total Sales



Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.297937	Pr > D	<0.0100
Cramer-von Mises	W-Sq	217.9787	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1142.555	Pr > A-Sq	<0.0050

- The Analysis Focuses on distribution of the continuous variable 'Total Sales
- Summarisation- Proc Means,Proc Univariate
- Visualisation- Histogram and Density Plot
- Summary estimate is as shown in the image-'The Means Procedure'
- Kolmogorov-Smirnov Test of Normality shows variable is not normally distributed
- Histogram and Kernal plot displays that distribution is not uniform.

6. Invoice Date

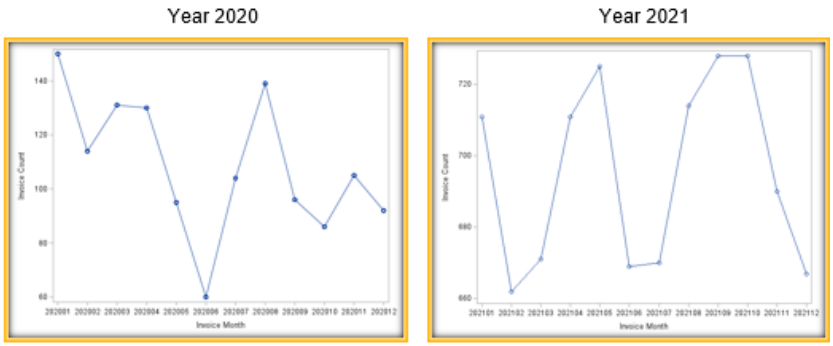
Highest

Invoice Date	invoice_count
01/17/2021	77
08/17/2021	72
03/16/2021	72
04/17/2021	72
11/17/2021	71
07/16/2021	70
12/16/2021	69
01/10/2021	67
01/24/2021	67
10/17/2021	65
02/17/2021	65
06/17/2021	64
01/23/2021	59
05/18/2021	54
09/18/2021	54
05/19/2021	54
09/19/2021	54
04/11/2021	54

Lowest

09/26/2020	2
05/14/2020	2
09/25/2020	2
09/24/2020	2
05/13/2020	2
09/23/2020	2
09/22/2020	2
05/12/2020	2
09/21/2020	2
09/20/2020	2
05/11/2020	2
09/19/2020	2
09/18/2020	2
05/10/2020	2
09/17/2020	2
05/09/2020	2
05/08/2020	2

Invoice Trend



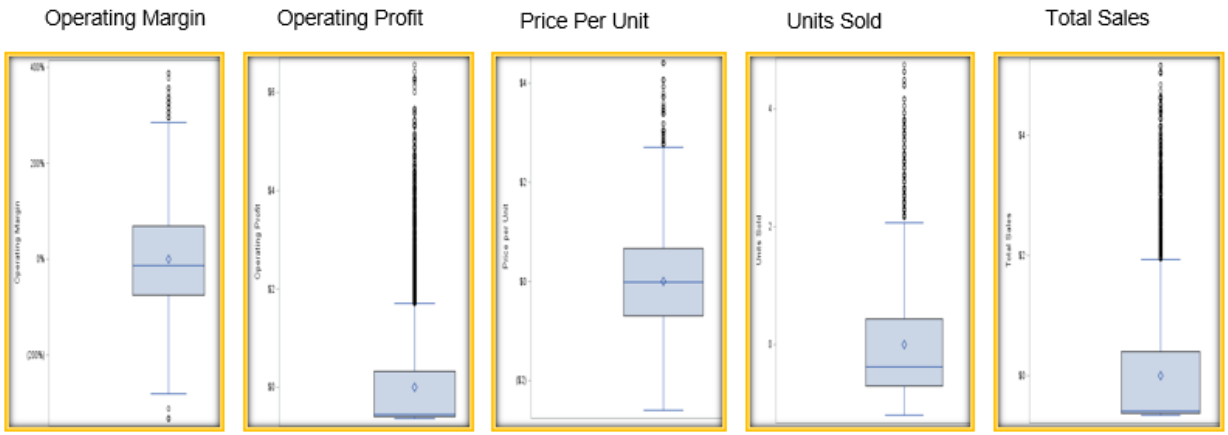
- The Analysis Focuses on distribution of the continuous variable 'Invoice Date'
- Visualisation- Series (to Visualise Monthwise Trend) for year 2020 and 2021 separately
- Summary estimate is as shown in the image-'The Means Procedure'
- Highest number of Invoicing happened on 17/01/2021
- Series plot indicate there is not particular season for having higher or lower invoicing.
- However In both years, june month was low performer. There is a rise in invoicing of August month in both years
- Winter and Summer Sales of 2021 is much better than Year 2020

Outliers Detection- Numerical Variables

Outlier detection done using Visualization method- Boxplot for all Five Numerical Variables

1. **Operating Margin:**
    - The distribution appears relatively normal, with some outliers at both ends.
    - Most values lie within a moderate range, but a few extremely high or low margins exist.
    - Potential data quality issue or exceptional business cases.
2. **Operating Profit:**
    - Highly skewed distribution with many outliers on the upper end.
    - The majority of data falls in the lower range, while a few high-profit transactions skew the data.
    - Indicates some exceptionally profitable transactions compared to the norm.
3. **Price Per Unit:**
    - Shows a right-skewed distribution with numerous outliers at the upper end.
    - The median price is stable, but some items are priced significantly higher than the majority.
    - Possible presence of premium products or data entry errors.
4. **Units Sold:**
    - The distribution is right-skewed, meaning most transactions involve lower unit sales, but a few have extremely high values.
    - The high number of outliers suggests bulk purchases or special transactions.
5. **Total Sales:**
    - Extreme right-skewness with numerous high-value outliers.
    - Most sales are in a lower range, but a few transactions generate exceptionally high revenue.
    - These could represent corporate deals, seasonal spikes, or outlier transactions.

Post Standardisation- Outliers Analysis



After standardizing the numerical variables (Operating Margin, Operating Profit, Price Per Unit, Units Sold, and Total Sales), the boxplots remain visually similar to the pre-standardization versions. However, their values are now on a standardized scale with a **mean of 0** and a **standard deviation of 1**.

Since the variables are standardized, the central tendency (mean) of each variable is now zero.
The spread of data (variance) has been adjusted, making the standard deviation one.

Operating Profit, Total Sales, and Units Sold exhibit a large number of high-value outliers (above the upper whisker).
Operating Margin and Price Per Unit have fewer outliers but still show some extreme values.
The outliers remain after standardization because standardization does not remove outliers—it only rescales the data.

The overall distribution and skewness remain unchanged (e.g., highly skewed distributions in Operating Profit and Total Sales).
The box widths are now comparable, allowing direct comparison of variable variability.

BIVARIATE ANALYSIS - Categorical Vs Categorical

A) <u>Association Between Region and Product</u>

Ho- Null Hypothesis-There is no association between Region and Product
H1- Alternate Hypothesis-There is association between Region and Product

Approach- since both are categorical Variables we will use:-

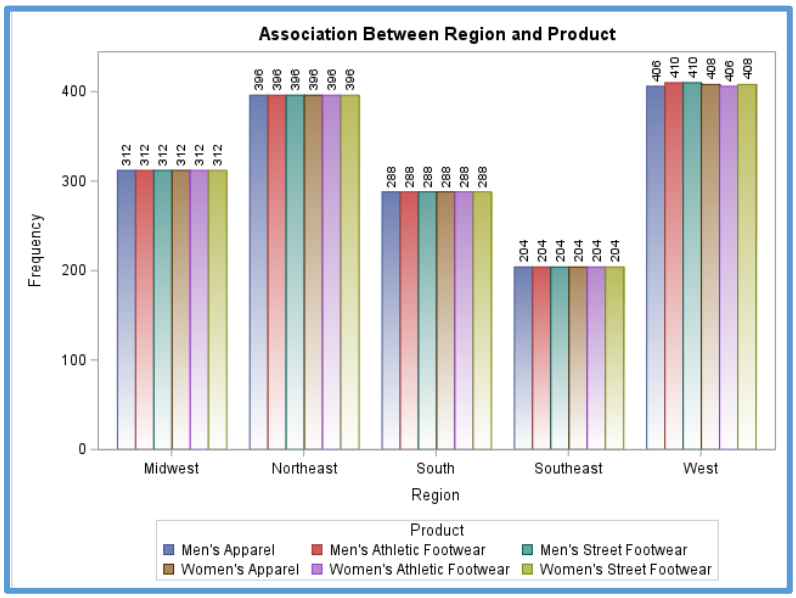
- For Summarisation- Frequency Table
- For Visualisation- Grouped Bar Chart
- For Independency- Chi sq test

.

Table of Region by Product

Region(Region)	Product(Product)						
	Men's Apparel	Men's Athletic Footwear	Men's Street Footwear	Women's Apparel	Women's Athletic Footwear	Women's Street Footwear	Total
Midwest	312 3.23	312 3.23	312 3.23	312 3.23	312 3.23	312 3.23	1872 19.40
Northeast	396 4.10	396 4.10	396 4.10	396 4.10	396 4.10	396 4.10	2376 24.63
South	288 2.99	288 2.99	288 2.99	288 2.99	288 2.99	288 2.99	1728 17.91
Southeast	204 2.11	204 2.11	204 2.11	204 2.11	204 2.11	204 2.11	1224 12.69
West	406 4.21	410 4.25	410 4.25	408 4.23	406 4.21	408 4.23	2448 25.37
Total	1606 16.65	1610 16.69	1610 16.69	1608 16.67	1606 16.65	1608 16.67	9648 100.00

- Above chart shows distribution of Adidas Products among 6 different regions. Midwest,Northeast,South,Southeast,West w.r.t number of invoices



Statistics for Table of Region by Product

Statistic	DF	Value	Prob
Chi-Square	20	0.0293	1.0000
Likelihood Ratio Chi-Square	20	0.0293	1.0000
Mantel-Haenszel Chi-Square	1	0.0003	0.9871
Phi Coefficient		0.0017	
Contingency Coefficient		0.0017	
Cramer's V		0.0009	

Sample Size = 9648

- Grouped Bar chart shows distribution of Adidas Products among 5 different regions. Midwest,Northeast,South,Southeast,West w.r.t number of invoices
- We can see all products are equally distributed among each region
- Statistics table displays the results of Chi-Square test between Region and Product
- P value is 1>0.05 and Cramer's V is very low

Conclusion

- **P value of >0.05- We fail to reject null hypothesis of Independency**
- **Region and Products are not Statistically significantly associated with each other.**
- **Region and Product are Independent of each other**

B) Association Between Retailer and Sales Method

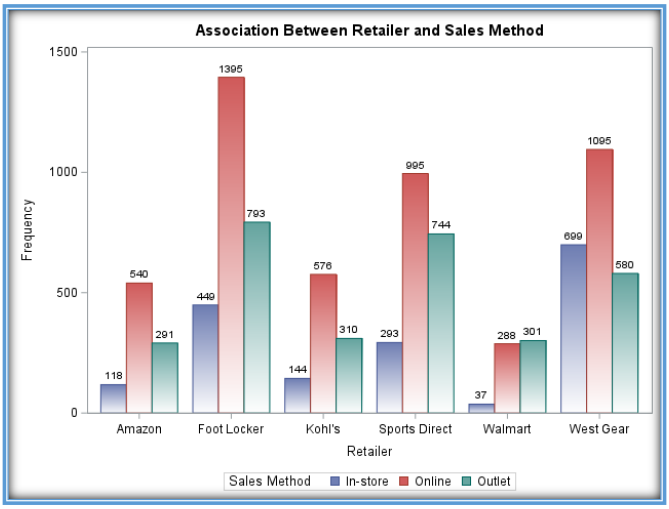
Ho- Null Hypothesis-There is no association between Retailer and Sales Method
H1- Alternate Hypothesis-There is association between Retailer and Sales Method

Approach- since both are categorical Variables we will use:-

- For Summarisation- Frequency Table
- For Visualisation- Grouped Bar Chart
- For Independency- Chi sq test

Table of Retailer by Sales_Method

Retailer(Retailer)	Sales_Method(Sales Method) In-store	Online	Outlet	Total
Amazon	118 1.22	540 5.60	291 3.02	949 9.84
Foot Locker	449 4.65	1395 14.46	793 8.22	2637 27.33
Kohl's	144 1.49	576 5.97	310 3.21	1030 10.68
Sports Direct	293 3.04	995 10.31	744 7.71	2032 21.06
Walmart	37 0.38	288 2.99	301 3.12	626 6.49
West Gear	699 7.25	1095 11.35	580 6.01	2374 24.61
Total	1740 18.03	4889 50.67	3019 31.29	9648 100.00



- Above Contingency Table and Group Bar Chart shows distribution of 6 Adidas Retailers among 3 types of Sales Method i.e. In store, Online and Outlet w.r.t number of invoices
- We can see Foot Locker has Maximum customer base in Online and Outlet Sales Method whereas West Gear is better with In store method
- Overall Online sales method is effective for all retailers except Walmart and In store is not much effective as compared to other 2 sales methods
- Walmart is lowest customer base with In store Sales Method

Statistics for Table of Retailer by Sales_Method

Statistic	DF	Value	Prob
Chi-Square	10	405.8288	<.0001
Likelihood Ratio Chi-Square	10	394.1055	<.0001
Mantel-Haenszel Chi-Square	1	52.3423	<.0001
Phi Coefficient		0.2051	
Contingency Coefficient		0.2009	
Cramer's V		0.1450	

Sample Size = 9648

- Statistics table displays the results of Chi-Square test between Retailer and Sales Method
- P value is 0.001<0.05 and Cramer's V is 0.145

Conclusion

- **P value of <0.05- We reject null hypothesis of Independency**
- **Retailer and Sales Method are statistically significantly associated with each other.**

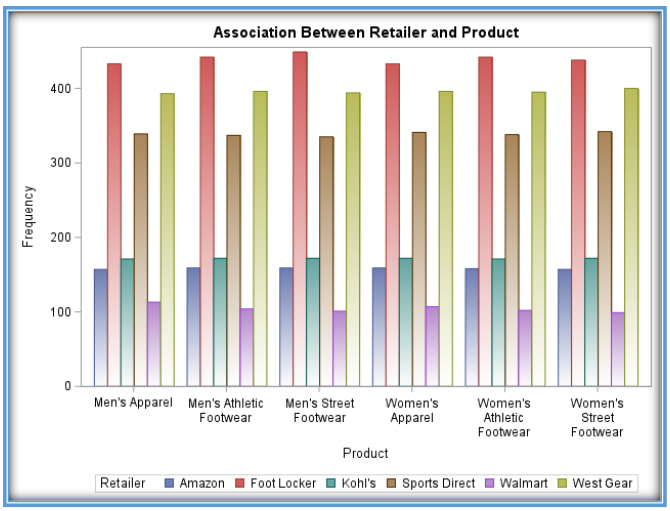
C) Association Between Retailer and Product

Ho- Null Hypothesis-There is no association between Retailer and Product
H1- Alternate Hypothesis-There is association between Retailer and Product

Approach- since both are categorical Variables we will use:-

- For Summarisation- Frequency Table
- For Visualisation- Grouped Bar Chart
- For Independency- Chi sq test

Frequency Percent	Table of Retailer by Product						
	Product(Product)						
Retailer(Retailer)	Men's Apparel	Men's Athletic Footwear	Men's Street Footwear	Women's Apparel	Women's Athletic Footwear	Women's Street Footwear	Total
Amazon	157 1.63	159 1.65	159 1.65	159 1.65	158 1.64	157 1.63	949 9.84
Foot Locker	433 4.49	442 4.58	449 4.65	433 4.49	442 4.58	438 4.54	2637 27.33
Kohl's	171 1.77	172 1.78	172 1.78	172 1.78	171 1.77	172 1.78	1030 10.68
Sports Direct	339 3.51	337 3.49	335 3.47	341 3.53	338 3.50	342 3.54	2032 21.06
Walmart	113 1.17	104 1.08	101 1.05	107 1.11	102 1.06	99 1.03	626 6.49
West Gear	393 4.07	396 4.10	394 4.08	396 4.10	395 4.09	400 4.15	2374 24.61
Total	1606 16.65	1610 16.69	1610 16.69	1608 16.67	1606 16.65	1608 16.67	9648 100.00



Statistic	DF	Value	Prob
Chi-Square	25	1.8532	1.0000
Likelihood Ratio Chi-Square	25	1.8385	1.0000
Mantel-Haenszel Chi-Square	1	0.0006	0.9799
Phi Coefficient		0.0139	
Contingency Coefficient		0.0139	
Cramer's V		0.0062	

- Above Contingency Table and Group Bar Chart shows distribution of 6 Adidas Retailers among 6 different Products i.e. Mens Apparel, Mens Athletic Footwear,Mens Street Footwear,Woman's Apparel,Woman's Athletic Footwear ,Woman's Street Footwear
- We can see Foot Locker has Maximum customer base for all products whereas Walmart has low customer base across all products
- Statistics table displays the results of Chi-Square test between Retailer and Products
- P value is 1>0.05 and Cramer's V is 0.0062

Conclusion

- **P value of >0.05- We fail to reject null hypothesis of Independency**
- **Retailer and Sales Method are not statistically significantly associated with each other.**
- **Retailer and Sales Method are independent of each other**

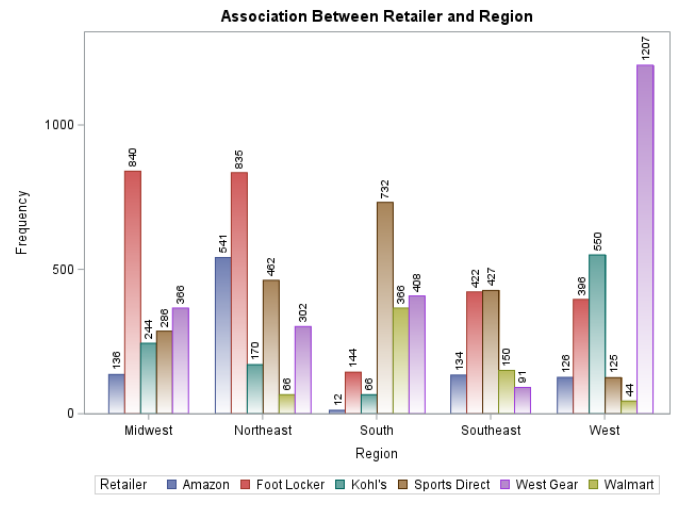
D) Association Between Retailer and Region

Ho- Null Hypothesis-There is no association between Retailer and Region
H1- Alternate Hypothesis-There is association between Retailer and Region

Approach- since both are categorical Variables we will use:-

- For Summarisation- Frequency Table
- For Visualisation- Grouped Bar Chart
- For Independency- Chi sq test

Frequency Percent	Table of Retailer by Region					
	Region(Region)					
Retailer(Retailer)	Midwest	Northeast	South	Southeast	West	Total
Amazon	136 1.41	541 5.61	12 0.12	134 1.39	126 1.31	949 9.84
Foot Locker	840 8.71	835 8.65	144 1.49	422 4.37	396 4.10	2637 27.33
Kohl's	244 2.53	170 1.76	66 0.68	0 0.00	550 5.70	1030 10.68
Sports Direct	286 2.96	462 4.79	732 7.59	427 4.43	125 1.30	2032 21.06
Walmart	0 0.00	66 0.68	366 3.79	150 1.55	44 0.46	626 6.49
West Gear	366 3.79	302 3.13	408 4.23	91 0.94	1207 12.51	2374 24.61
Total	1872 19.40	2376 24.63	1728 17.91	1224 12.69	2448 25.37	9648 100.00



Statistic	DF	Value	Prob
Chi-Square	20	4423.6640	<.0001
Likelihood Ratio Chi-Square	20	4496.0605	<.0001
Mantel-Haenszel Chi-Square	1	690.2118	<.0001
Phi Coefficient		0.6771	
Contingency Coefficient		0.5607	
Cramer's V		0.3386	

- Above Contingency Table and Group Bar Chart shows distribution of 6 Adidas Retailers among 5 Regions i.e. Midwest,Northeast,South,Southeast,West w.r.t number of invoices
- We can see Foot Locker has Maximum customer base Midwest and North East, West Gear has highest customer base in west as well as overall.
- Amazon has lowest share in South Region whereas Sports direct performs best in South
- Statistics table displays the results of Chi-Square test between Retailer and Region
- P value is 0.001<0.05 and Cramer's V is 0.3386

Conclusion

- **P value of <0.05- We reject null hypothesis of Independency**
- **Retailer and Region are statistically significantly associated with each other.**

E) Association Between Sales Method and Product

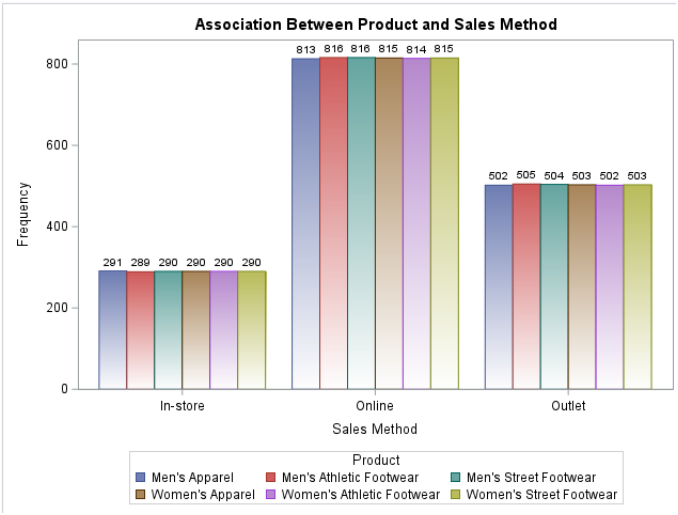
Ho- Null Hypothesis-There is no association between Sales Method and Product
H1- Alternate Hypothesis-There is association between Sales Method and Product

Approach- since both are categorical Variables we will use:-

- For Summarisation- Frequency Table
- For Visualisation- Grouped Bar Chart
- For Independency- Chi sq test

Frequency Percent	Table of Product by Sales_Method			
	Sales_Method(Sales Method)			
Product(Product)	In-store	Online	Outlet	Total
Men's Apparel	291 3.02	813 8.43	502 5.20	1606 16.65
Men's Athletic Footwear	289 3.00	816 8.46	505 5.23	1610 16.69
Men's Street Footwear	290 3.01	816 8.46	504 5.22	1610 16.69
Women's Apparel	290 3.01	815 8.45	503 5.21	1608 16.67
Women's Athletic Footwear	290 3.01	814 8.44	502 5.20	1606 16.65
Women's Street Footwear	290 3.01	815 8.45	503 5.21	1608 16.67
Total	1740 18.03	4889 50.67	3019 31.29	9648 100.00

Statistic	DF	Value	Prob
Chi-Square	10	0.0189	1.0000
Likelihood Ratio Chi-Square	10	0.0189	1.0000
Mantel-Haenszel Chi-Square	1	0.0001	0.9915
Phi Coefficient		0.0014	
Contingency Coefficient		0.0014	
Cramer's V		0.0010	



- Above Contingency Table and Group Bar Chart shows distribution of 6 Adidas Products among 3 types of Sales Method
- All Products are almost equally distributed among each Sales Method
- Statistics table displays the results of Chi-Square test between Sales Method and Products
- P value is 1>0.05 and Cramer's V is 0.0010

Conclusion

- **P value of >0.05- We fail to reject null hypothesis of Independency**
- **Products and Sales Method are not statistically significantly associated with each other.**
- **Products and Sales Method are independent of each other**

F) <u>Association Between Sales Method and Region</u>

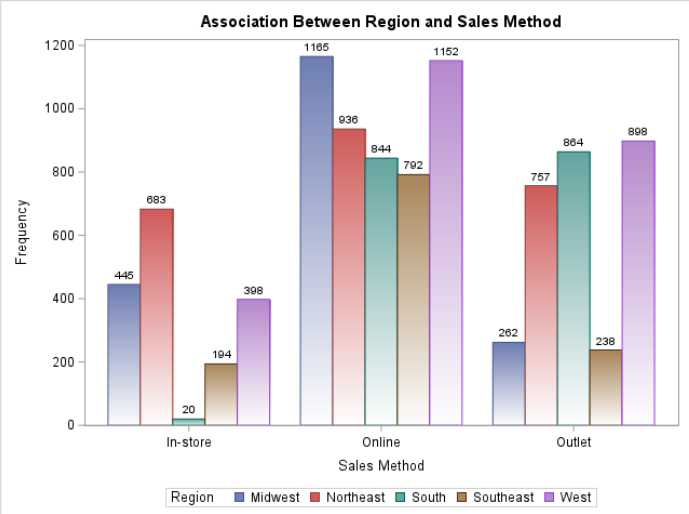
Ho- Null Hypothesis-There is no association between Sales Method and Region
H1- Alternate Hypothesis-There is association between Sales Method and Region

Approach- since both are categorical Variables we will use:-

- For Summarisation- Frequency Table
- For Visualisation- Grouped Bar Chart
- For Independency- Chi sq test
-

Frequency Percent	Table of Region by Sales_Method			
	Sales_Method(Sales Method)			
Region(Region)	In-store	Online	Outlet	Total
Midwest	445 4.61	1165 12.08	262 2.72	1872 19.40
Northeast	683 7.08	936 9.70	757 7.85	2376 24.63
South	20 0.21	844 8.75	864 8.96	1728 17.91
Southeast	194 2.01	792 8.21	238 2.47	1224 12.69
West	398 4.13	1152 11.94	898 9.31	2448 25.37
Total	1740 18.03	4889 50.67	3019 31.29	9648 100.00

Statistic	DF	Value	Prob
Chi-Square	8	1079.9202	<.0001
Likelihood Ratio Chi-Square	8	1276.6329	<.0001
Mantel-Haenszel Chi-Square	1	179.6860	<.0001
Phi Coefficient		0.3346	
Contingency Coefficient		0.3173	
Cramer's V		0.2366	



- Above Contingency Table and Group Bar Chart shows distribution of 5 Regions among 3 types of Sales Method w.r.t number of invoices
- Online Sales Method is very much effective in Midwest and West Region.
- In store method performed poorly in South Region where as it performs best in Northeast region.
- Statistics table displays the results of Chi-Square test between Sales Method and Region
- P value is 0.001<0.05 and Cramer's V is 0.2366

Conclusion

- **P value of <0.05- We reject null hypothesis of Independency**
- **Sales Method and Region are statistically significantly associated with each other.**

BIVARIATE ANALYSIS - Numerical Vs Numerical

A) <u>Correlation Between Price Per Unit and Units Sold</u>

H0- There is no correlation Between Price Per Unit and Units Sold

H1- There is significant Correlation between Price Per Unit and Units Sold

The CORR Procedure

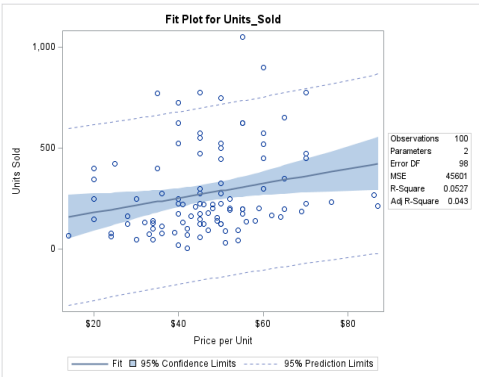
1 With Variables:	Price_per_Unit
1 Variables:	Units_Sold

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Price_per_Unit	100	46.89000	13.90363	4689	14.00000	87.00000	Price per Unit
Units_Sold	100	278.82000	218.28909	27882	6.00000	1050	Units Sold

Pearson Correlation Coefficients, N = 100
Prob > |r| under H0: Rho=0

	Units_Sold
Price_per_Unit Price per Unit	0.22951 0.0216



The REG Procedure
Model: MODEL1
Dependent Variable: Units_Sold Units Sold

Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	248487	248487	5.45	0.0216
Error	98	4468876	45601		
Corrected Total	99	4717363			

Root MSE	213.54338	R-Square	0.0527
Dependent Mean	278.82000	Adj R-Sq	0.0430
Coeff Var	76.58826		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	109.85925	75.46468	1.46	0.1487
Price_per_Unit	Price per Unit	1	3.60334	1.54362	2.33	0.0216

- P value of Price Per Unit is 0.0216<0.05. Therefore, we reject Null Hypothesis and conclude there is a correlation between Price per unit and Units Sold .
- R square is very low 0.0527and RMSE is high 213.5433
- R Square- 0.0527 indicates only 5% of variation in Units Sold can be explained by Price Per Unit.
- Model is not a best fit for the analysis between Units Sold and Price Per Unit

B) <u>Correlation Between Operating Profit and Total Sales</u>

H0- There is no correlation Between Operating Profit and Total Sales
H1- There is significant Correlation between Operating Profit and Total Sales

- P value of Price Per Unit is 0.0001<0.05. Therefore, we reject Null Hypothesis and conclude there is a correlation between Operating Profit and Total Sales
- R square is very low 0.9426and RMSE is high 31.5%
- R Square- 0.9426 indicates 94% of variation in Total Sales can be explained by Operating Profit

The CORR Procedure

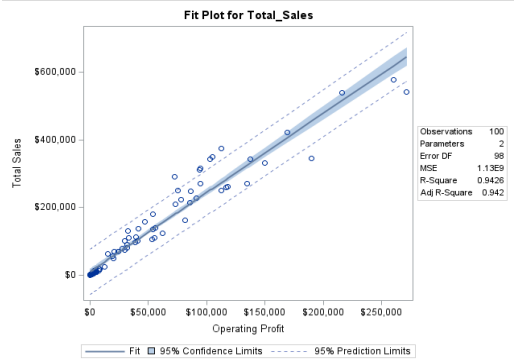
1 With Variables:	Operating_Profit
1 Variables:	Total_Sales

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Operating_Profit	100	41448	57726	4144825	120.96000	271250	Operating Profit
Total_Sales	100	106637	139544	10663672	252.00000	577500	Total Sales

Pearson Correlation Coefficients, N = 100
Prob > |r| under H0: Rho=0

	Total_Sales
Operating_Profit Operating Profit	0.97088 <.0001



The REG Procedure
Model: MODEL1
Dependent Variable: Total_Sales Total Sales

Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.817123E12	1.817123E12	1609.36	<.0001
Error	98	1.106516E11	1129097868		
Corrected Total	99	1.927775E12			

Root MSE	33602	R-Square	0.9426
Dependent Mean	106637	Adj R-Sq	0.9420
Coeff Var	31.51077		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	9360.11046	4143.76628	2.26	0.0261
Operating_Profit	Operating Profit	1	2.34694	0.05850	40.12	<.0001

C) Correlation Between All Numerical Variables

Pearson Correlation Coefficients, N = 9648
Prob > |r| under H0: Rho=0

	Total_Sales	Price_per_Unit	Units_Sold	Operating_Profit	Operating_Margin
Total_Sales Total Sales	1.00000	0.43581 <.0001	0.91343 <.0001	0.95631 <.0001	-0.36459 <.0001
Price_per_Unit Price per Unit	0.43581 <.0001	1.00000	0.26587 <.0001	0.39455 <.0001	-0.13749 <.0001
Units_Sold Units Sold	0.91343 <.0001	0.26587 <.0001	1.00000	0.89238 <.0001	-0.30548 <.0001
Operating_Profit Operating Profit	0.95631 <.0001	0.39455 <.0001	0.89238 <.0001	1.00000	-0.21192 <.0001
Operating_Margin Operating Margin	-0.36459 <.0001	-0.13749 <.0001	-0.30548 <.0001	-0.21192 <.0001	1.00000

- High Positive Correlation (91.343%) found between Units Sold and Total Sales
- High Positive Correlation (95.631%) found between Operating and Total Sales
- High Positive Correlation (89.238%) found between Units Sold and Operating Profit
- Moderate Negative Correlation found between Operating Margin and Total Sales

BIVARIATE ANALYSIS - Categorical Vs Numerical

A) Distribution of Operating Profit in Sales Method

H0- Means of Operating Profit is equal in all Sales Methods i.e Operating Profit is equally distributed in Sales Methods

H1 -Means of Operating Profit is not equal in all Sales Methods i.e Operating Profit is not equally distributed in Sales Methods

Approach- since Operating Profit is Numerical variable and Sales Method is a categorical Variable with more than 2 levels we will use:-

A) For Summarisation- Proc Univariate
B) For Normality-Proc Univariate
C) For Visualisation- Grouped Box Plot
D) For Independency- Proc Anova

The UNIVARIATE Procedure
Variable: Operating_Profit (Operating Profit)
Sales_Method = In-store

Moments			
N	1740	Sum Weights	1740
Mean	73328.3261	Sum Observations	127591288
Std Deviation	56537.0185	Variance	3196434463
Skewness	1.66358447	Kurtosis	3.28715061
Uncorrected SS	1.49147E13	Corrected SS	5.5586E12
Coeff Variation	77.1011988	Std Error Mean	1355.37128

The UNIVARIATE Procedure
Variable: Operating_Profit (Operating Profit)
Sales_Method = Online

Moments			
N	4889	Sum Weights	4889
Mean	19749.4736	Sum Observations	96555176.5
Std Deviation	43353.5166	Variance	1879527398
Skewness	3.37554984	Kurtosis	13.459984
Uncorrected SS	1.1094E13	Corrected SS	9.18713E12
Coeff Variation	219.517327	Std Error Mean	620.032297

The UNIVARIATE Procedure
Variable: Operating_Profit (Operating Profit)
Sales_Method = Outlet

Moments			
N	3019	Sum Weights	3019
Mean	35769.5586	Sum Observations	107988297
Std Deviation	57258.555	Variance	3278542121
Skewness	2.1044602	Kurtosis	5.11384683
Uncorrected SS	1.37573E13	Corrected SS	9.89464E12
Coeff Variation	160.076213	Std Error Mean	1042.09864

- From the results of Univariate Procedure, we see that Mean and standard deviation of Operating Profit is all Sales Methods is not same.
- To statistically prove this we will use normality test using Proc Univariate, Homoscedasticity test using Proc GLM and finally Proc Anova to prove means are not equal in all groups

- Test of Normality-

H0- Operating Profit is normally distributed
H1- Operating Profit is not normally distributed

Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.85262	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.427867	Pr > D	<0.0100
Cramer-von Mises	W-Sq	11.20894	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	67.41568	Pr > A-Sq	<0.0050

Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.385156	Pr > D	<0.0100
Cramer-von Mises	W-Sq	207.5168	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	995.5268	Pr > A-Sq	<0.0050

Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.307708	Pr > D	<0.0100
Cramer-von Mises	W-Sq	75.81803	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	388.1119	Pr > A-Sq	<0.0050

P value in Kolmogorov-Smirnov of Normality is less than 0.05 significance level for all Sales Methods. So we reject Null Hypothesis of normality and conclude that Operating Profit is not normally distributed, However as per CLT, since sample size is more than>30, we can skip the normality assumption

- Test of Homoscedasticity for equality of variance-

  H0- Variance of Operating Profit is equal in all Groups
  H1- Variance of Operating Profit is not equal in all Groups

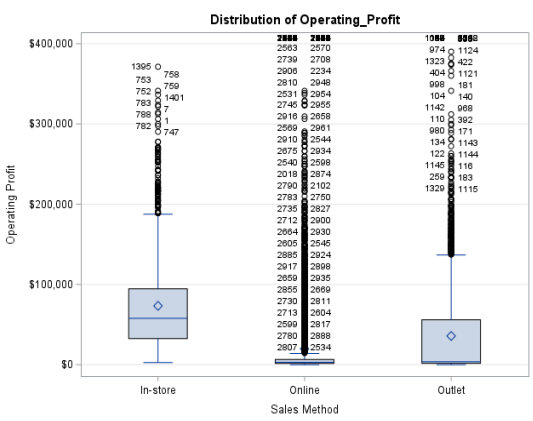
The GLM Procedure

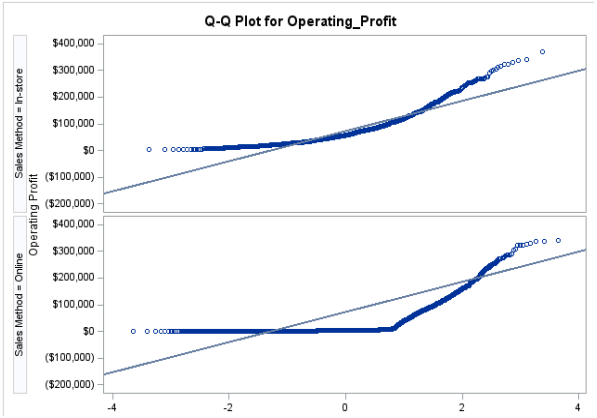
Levene's Test for Homogeneity of Operating_Profit Variance
ANOVA of Absolute Deviations from Group Means

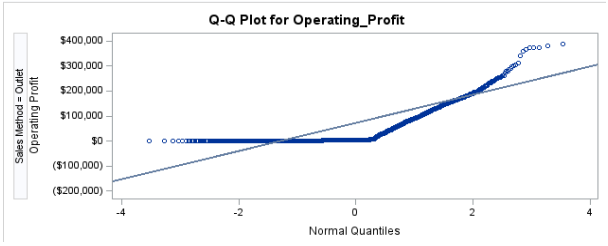
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Sales_Method	2	6E11	3E11	234.36	<.0001
Error	9645	1.235E13	1.2801E9		

Welch's ANOVA for Operating_Profit

Source	DF	F Value	Pr > F
Sales_Method	2.0000	662.23	<.0001
Error	4063.7		







P value in Levene's Test for Equality of Variance 0.0001 is less than 0.05 significance level.

Therefore, we reject Null Hypothesis and Conclude the variance of Operating Profit is not equal in all Sales Methods

- Test of difference-

  H0- Mean of Operating Profit is Equal in all Sales Methods
  H1- Mean of Operating Profit is not Equal in all Sales Methods

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
2106.0121	2	<.0001

Wilcoxon Scores (Rank Sums) for Variable Operating_Profit
Classified by Variable Sales_Method

Sales_Method	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
In-store	1740	13081884.5	8394630.0	105186.106	7518.32443
Outlet	3019	14095439.5	14565165.5	126854.532	4668.91007
Online	4889	19369452.0	23586980.5	136778.644	3961.84332

Average scores were used for ties.

The image shows the output of a Kruskal-Wallis Test, which is a non-parametric method used to compare three or more independent groups to determine if there are statistically significant differences between them.

- Chi-Square = 2106.0121-This is the test statistic, which measures the differences between the groups based on their rank sums.
- DF (Degrees of Freedom) = 2-Since the Kruskal-Wallis test is used to compare three or more groups, the degrees of freedom (DF) is calculated as k - 1, where k is the number of groups. In this case, there are three groups (3 - 1 = 2).
- Pr > ChiSq < 0.0001-This is the p-value. A p-value less than 0.0001 suggests that there is a statistically significant difference between the groups.
- Since this value is very small (typically lower than a significance level of 0.05), we reject the null hypothesis.

The ANOVA Procedure

Dependent Variable: Operating_Profit Operating Profit

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3.6918427E12	1.8459214E12	722.55	<.0001
Error	9645	2.464037E13	2554729867.7		
Corrected Total	9647	2.8332212E13			

R-Square	Coeff Var	Root MSE	Operating_Profit Mean
0.130305	146.8235	50544.34	34425.24

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Sales_Method	2	3.6918427E12	1.8459214E12	722.55	<.0001

Comparisons significant at the 0.05 level are indicated by *.**

Sales_Method Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
In-store - Outlet	37559	33834	41283	***
In-store - Online	53579	50125	57033	***
Outlet - In-store	-37559	-41283	-33834	***
Outlet - Online	16020	13156	18884	***
Online - In-store	-53579	-57033	-50125	***
Online - Outlet	-16020	-18884	-13156	***

- F value: the overall f statistic is calculated by using mean square model/mean square error, =722.55
  F value: It is the ratio of mean square model/mean square error, it is used to determine the variance explained by model is significantly greater than the unexplained variance
- The p value: (<0.05)-this means we reject null hypothesis and the mean is statistically not equal between the groups.

R squared: It is the proportion of variance in Operating Profit explained by the model. 0.13 percent shows that Sales Method can explain only 13% of variability of Operating Profit, it is a low, so our model is not sufficient to explain the variability of Operating Profit using Sales Method

- Diagram – Grouped Box Plot- The Total Length of Boxplot or Interquartile range (Distance between Q1 and Q3) is not similar for all 3 Sales Methods. We can see they do not line up and Diamond that represents the mean shows that the mean of all Sales Methods is not at similar level.

Conclusion-

- Since P values 0.0001<0.05 ,we reject the Null Hypothesis at 5% Significance Level
- Mean of Operating Profit is not significantly similar in all Sales Methods
- Sales Method explains 13% of variability of Operating Profit. Therefore, we can say this is not a good model.

B) <u>Distribution of Total Sales in Retailer</u>

H0- Means of Total Sales is equal in all Retailers i.e Total Sales is equally distributed in all Retailers

H1 -Means of Total Sales is not equal in all Retailers i.e Total Sales is not equally distributed in all Retailers

Approach- since Total Sales is Numerical variable and Retailer is a categorical Variable with more than 2 levels we will use:-

E) For Summarisation- Proc Univariate
F) For Normality-Proc Univariate
G) For Visualisation- Grouped Box Plot
H) For Independency- Proc Anova

The UNIVARIATE Procedure
Variable: Total_Sales (Total Sales)
Retailer = Amazon

Moments

N	949	Sum Weights	949
Mean	81874.5121	Sum Observations	77698912
Std Deviation	113390.215	Variance	1.28573E10
Skewness	1.6031961	Kurtosis	2.35189627
Uncorrected SS	1.85503E13	Corrected SS	1.21888E13
Coeff Variation	138.49269	Std Error Mean	3680.80212

The UNIVARIATE Procedure
Variable: Total_Sales (Total Sales)
Retailer = Foot Locker

Moments

N	2637	Sum Weights	2637
Mean	83464.0576	Sum Observations	220094720
Std Deviation	134053.6	Variance	1.79704E10
Skewness	2.20666938	Kurtosis	5.06041026
Uncorrected SS	6.57399E13	Corrected SS	4.73699E13
Coeff Variation	160.612369	Std Error Mean	2610.49832

The UNIVARIATE Procedure
Variable: Total_Sales (Total Sales)
Retailer = Kohl's

Moments

N	1030	Sum Weights	1030
Mean	99140.5369	Sum Observations	102114753
Std Deviation	130908.219	Variance	1.7137E10
Skewness	1.13152979	Kurtosis	0.01877409
Uncorrected SS	2.77576E13	Corrected SS	1.76339E13
Coeff Variation	132.04308	Std Error Mean	4078.94924

The UNIVARIATE Procedure
Variable: Total_Sales (Total Sales)
Retailer = Sports Direct

Moments

N	2032	Sum Weights	2032
Mean	89798.719	Sum Observations	182470997
Std Deviation	133219.357	Variance	1.77474E10
Skewness	1.62152321	Kurtosis	2.01431089
Uncorrected SS	5.24306E13	Corrected SS	3.6045E13
Coeff Variation	148.353294	Std Error Mean	2955.3266

The UNIVARIATE Procedure
Variable: Total_Sales (Total Sales)
Retailer = Walmart

Moments

N	626	Sum Weights	626
Mean	119102.891	Sum Observations	74558410
Std Deviation	185747.169	Variance	3.4502E10
Skewness	1.64419916	Kurtosis	1.85037324
Uncorrected SS	3.04439E13	Corrected SS	2.15638E13
Coeff Variation	155.955214	Std Error Mean	7423.94998

The UNIVARIATE Procedure
Variable: Total_Sales (Total Sales)
Retailer = West Gear

Moments

N	2374	Sum Weights	2374
Mean	102343.864	Sum Observations	242964333
Std Deviation	156931.144	Variance	2.46274E10
Skewness	1.76837241	Kurtosis	2.51529809
Uncorrected SS	8.33067E13	Corrected SS	5.84408E13
Coeff Variation	153.33713	Std Error Mean	3220.83727

- From the results of Univariate Procedure, we see that Mean and standard deviation of Total Sales is all Retailers is not same.
- To statistically prove this we will use normality test using Proc Univariate, Homoscedasticity test using Proc GLM and finally Proc Anova to prove means are not equal in all groups

- Test of Normality-

  H0- Age is normally distributed
  H1- Age is not normally distributed

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.741218	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.298564	Pr > D	<0.0100
Cramer-von Mises	W-Sq	17.22966	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	93.72886	Pr > A-Sq	<0.0050

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.271146	Pr > D	<0.0100
Cramer-von Mises	W-Sq	60.45497	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	318.6931	Pr > A-Sq	<0.0050

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.748712	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.326125	Pr > D	<0.0100
Cramer-von Mises	W-Sq	21.14662	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	113.0676	Pr > A-Sq	<0.0050

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.328439	Pr > D	<0.0100
Cramer-von Mises	W-Sq	47.00504	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	246.151	Pr > A-Sq	<0.0050

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.678251	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.334361	Pr > D	<0.0100
Cramer-von Mises	W-Sq	17.18567	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	87.48597	Pr > A-Sq	<0.0050

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.30144	Pr > D	<0.0100
Cramer-von Mises	W-Sq	56.17112	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	293.8409	Pr > A-Sq	<0.0050

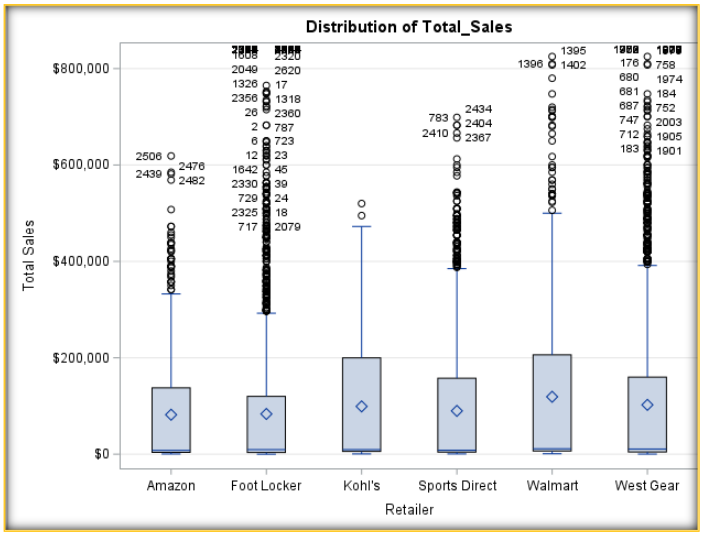
P value in Kolmogorov-Smirnov of Normality is less than 0.05 significance level for all Retailers. So we reject Null Hypothesis of normality and conclude that Total Sales is not normally distributed, However as per CLT, since sample size is more than>30, we can skip the normality assumption

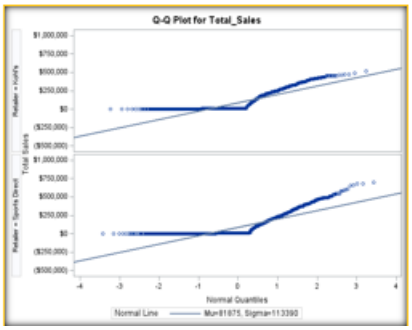
- Test of Homoscedasticity for equality of variance-

  H0- Variance of Total Sales is equal in all Groups
  H1- Variance of Total Sales is not equal in all Groups

Levene's Test for Homogeneity of Total_Sales Variance ANOVA of Absolute Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Retailer	5	2.066E12	4.132E11	52.78	<.0001
Error	9642	7.549E13	7.8289E9		

Welch's ANOVA for Total_Sales			
Source	DF	F Value	Pr > F
Retailer	5.0000	9.05	<.0001
Error	3085.5		





P value in Levene's Test for Equality of Variance 0.0001 is less than 0.05 significance level.

Therefore, we reject Null Hypothesis and Conclude the variance of Total Sales is not equal in all Retailers

- Test of difference-
  H0- Mean of Total Sales is Equal in all Retailers
  H1- Mean of Total Sales is not Equal in all Retailers

Wilcoxon Scores (Rank Sums) for Variable Total_Sales Classified by Variable Retailer

Retailer	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Foot Locker	2637	11931101.0	12722206.5	121925.678	4524.49791
Walmart	626	3380748.0	3020137.0	67388.953	5400.55591
Sports Direct	2032	9701767.5	9803384.0	111551.509	4774.49188
West Gear	2374	11807134.5	11453363.0	117835.768	4973.51917
Kohl's	1030	5343768.0	4969235.0	84483.517	5188.12427
Amazon	949	4382257.0	4578450.5	81473.787	4617.76291

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
87.6044	5	<.0001

The image shows the output of a Kruskal-Wallis Test, which is a non-parametric method used to compare three or more independent groups to determine if there are statistically significant differences between them.

- Chi-Square = 87.60-This is the test statistic, which measures the differences between the groups based on their rank sums.
- DF (Degrees of Freedom) = 5-Since the Kruskal-Wallis test is used to compare three or more groups, the degrees of freedom (DF) is calculated as k - 1, where k is the number of groups. In this case, there are three groups (6 - 1 = 5).
- Pr > ChiSq < 0.0001-This is the p-value. A p-value less than 0.0001 suggests that there is a statistically significant difference between the groups.
- Since this value is very small (typically lower than a significance level of 0.05), we reject the null hypothesis.

The ANOVA Procedure

Dependent Variable: Total_Sales Total Sales

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1.0499984E12	209999672960	10.48	<.0001
Error	9642	1.9324208E14	20041701377		
Corrected Total	9647	1.9429208E14			

R-Square	Coeff Var	Root MSE	Total_Sales Mean
0.005404	151.7782	141568.7	93273.44

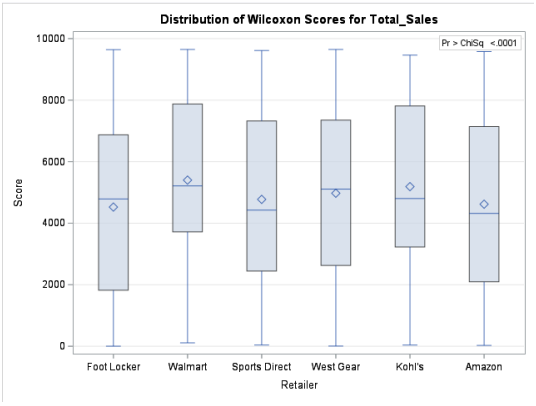
Source	DF	Anova SS	Mean Square	F Value	Pr > F
Retailer	5	1.0499984E12	209999672960	10.48	<.0001

Comparisons significant at the 0.05 level are indicated by ***.

Retailer Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Walmart - West Gear	16759	-4409	37927	
Walmart - Kohl's	19962	-3914	43839	
Walmart - Sports Direct	29304	7768	50840	***
Walmart - Foot Locker	35639	14692	56585	***
Walmart - Amazon	37228	12970	61487	***
West Gear - Walmart	-16759	-37927	4409	
West Gear - Kohl's	3203	-14375	20782	
West Gear - Sports Direct	12545	-1693	26784	
West Gear - Foot Locker	18880	5550	32209	***
West Gear - Amazon	20469	2375	38563	***
Kohl's - Walmart	-19962	-43839	3914	
Kohl's - West Gear	-3203	-20782	14375	
Kohl's - Sports Direct	9342	-8679	27362	
Kohl's - Foot Locker	15676	-1635	32988	
Kohl's - Amazon	17266	-3933	38465	
Sports Direct - Walmart	-29304	-50840	-7768	***
Sports Direct - West Gear	-12545	-26784	1693	
Sports Direct - Kohl's	-9342	-27362	8679	
Sports Direct - Foot Locker	6335	-7572	20242	
Sports Direct - Amazon	7924	-10599	26448	
Foot Locker - Walmart	-35639	-56585	-14692	***
Foot Locker - West Gear	-18880	-32209	-5550	***
Foot Locker - Kohl's	-15676	-32988	1635	
Foot Locker - Sports Direct	-6335	-20242	7572	
Foot Locker - Amazon	1590	-16245	19424	
Amazon - Walmart	-37228	-61487	-12970	***
Amazon - West Gear	-20469	-38563	-2375	***
Amazon - Kohl's	-17266	-38465	3933	
Amazon - Sports Direct	-7924	-26448	10599	
Amazon - Foot Locker	-1590	-19424	16245	

- F value: the overall f statistic is calculated by using mean square model/mean square error, =10.48
  F value: It is the ratio of mean square model/mean square error, it is used to determine the variance explained by model is significantly greater than the unexplained variance
- The p value: (<0.05)-this means we reject null hypothesis and the mean is statistically not equal between the groups.

R squared: It is the proportion of variance in Total Sales explained by the model. 0.005 percent shows that Retailer can explain only 0.5% of variability of Total Sales, it is a too low, so our model is not sufficient to explain the variability of Total Sales using Retailer



- Diagram – Grouped Box Plot- The Total Length of Boxplot or Interquartile range (Distance between Q1 and Q3) is not similar for all 6 Retailers. We can see they do not line up and Diamond that represents the mean shows that the mean of all Retailers is not at similar level.

Conclusion-

- Since P values 0.0001<0.05 ,we reject the Null Hypothesis at 5% Significance Level
- Mean of Total Sales is not significantly similar in all Retailers
- Retailer explains 13% of variability of Total Sales. Therefore, we can say this is not a good model.
-

Anova Test Results among All categorical Vs. Numerical Variables

Categorical Feature	Numerical Feature
Retailer	Total Sales
	Units Sold
	Operating P
	Operating Margin
	Price Per Units
Sales Method	Operating P
	Total Sales
	Units Sold
	Operating Margin
	Price Per Units
Region	Operating P
	Total Sales
	Units Sold
	Operating Margin
	Price Per Units
Product	Operating P
	Total Sales
	Units Sold
	Operating Margin
	Price Per Units

For all combinations :

- None of the Numerical Variable is normally distributed .
- None of the Numerical Variable has equal variance in any group of categorical variables
- P Values is < 0.05. We reject Null Hypothesis
- Means of Numerical Features are not equal in All respective categories of Categorical Features
- Numerical Features are not equal in all Categories of Categorical Features of this dataset

CHECKING MULTICOLLINEARITY

```
proc reg data=Nandinis.Adidas;
model Total_Sales= operating_profit operating_margin price_per_unit /vif collinoint ;
output out=outstat
p= predicted
r= Residual
stdr=se_resid
rstudent=Rstudent
h=Leverage
cookd=CooksD;
run;
quit;
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	34380	2056.17575	16.72	<.0001	0
Operating_Profit	Operating Profit	1	1.74392	0.01342	129.96	<.0001	5.82818
Operating_Margin	Operating Margin	1	-194986	3312.24414	-58.87	<.0001	1.14229
Price_per_Unit	Price per Unit	1	838.00993	22.96108	36.50	<.0001	1.25649
Units_Sold	Units Sold	1	169.08846	3.32074	50.92	<.0001	5.57882

Collinearity Diagnostics (intercept adjusted)

Number	Eigenvalue	Condition Index	Proportion of Variation			
			Operating_Profit	Operating_Margin	Price_per_Unit	Units_Sold
1	2.21741	1.00000	0.03019	0.03675	0.04949	0.03063
2	0.88927	1.57909	0.00719	0.80414	0.13013	0.00009830
3	0.80228	1.66249	0.01491	0.09049	0.69698	0.03235
4	0.09104	4.93515	0.94771	0.06862	0.12340	0.93693

The REG Procedure
Model: MODEL1
Dependent Variable: Total_Sales Total Sales

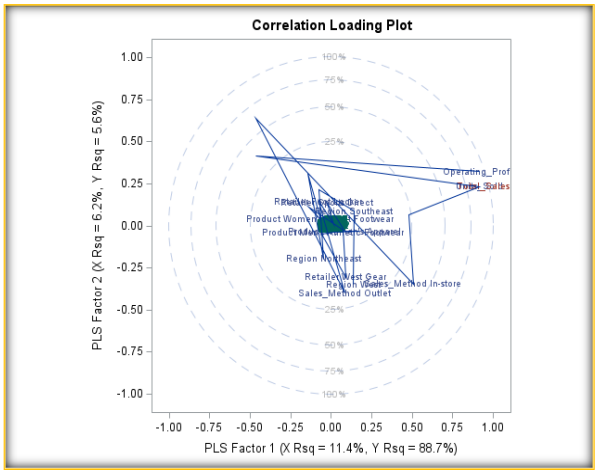
Number of Observations Read	9648
Number of Observations Used	9648

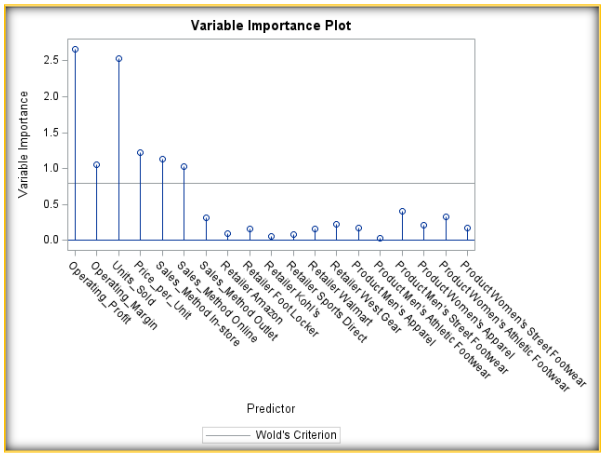
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1.858513E14	4.646283E13	53080.6	<.0001
Error	9643	8.440767E12	875325846		
Corrected Total	9647	1.942921E14			

Root MSE	29586	R-Square	0.9566
Dependent Mean	93273	Adj R-Sq	0.9565
Coeff Var	31.71954		

- Features are considered to be collinear or multi collinear if Variance inflation factor is more than 5 or 10.
- In above result VIF of Operating Profit and Units Sold is more than 5. It mean these 2 features have multicollinearity
- R Square is 95.66. Which Means 95.66% of variation in Total Sales is explained by independent Variables (numerical)





- In Above Correlation Leading Plot Operating Profit and Units Sold are closer to each other and in the range of PLS more than 0.5, which indicates high collinearity between them

- Variable Importance Plot indicates about Most important variables to explain the variation in Total Sales which are Operating Profit, Units Sold,Price Per Unit,Operating Margin, Sales Method-Instore and Online.

Lets Remove Units Sold to check the Multi Collinearity again

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	89385	1970.66475	45.36	<.0001	0
Operating_Profit	Operating Profit	1	2.35141	0.00692	339.89	<.0001	1.22101
Operating_Margin	Operating Margin	1	-242716	3578.33741	-67.83	<.0001	1.05080
Price_per_Unit	Price per Unit	1	566.32354	25.15500	22.51	<.0001	1.18864

Collinearity Diagnostics (intercept adjusted)

Number	Eigenvalue	Condition Index	Proportion of Variation		
			Operating_Profit	Operating_Margin	Price_per_Unit
1	1.51401	1.00000	0.23152	0.11904	0.21261
2	0.88902	1.30500	0.03324	0.84068	0.16898
3	0.59697	1.59253	0.73524	0.04028	0.61841

The REG Procedure
Model: MODEL1
Dependent Variable: Total_Sales Total Sales

Number of Observations Read	9648
Number of Observations Used	9648

Analysis of Variance

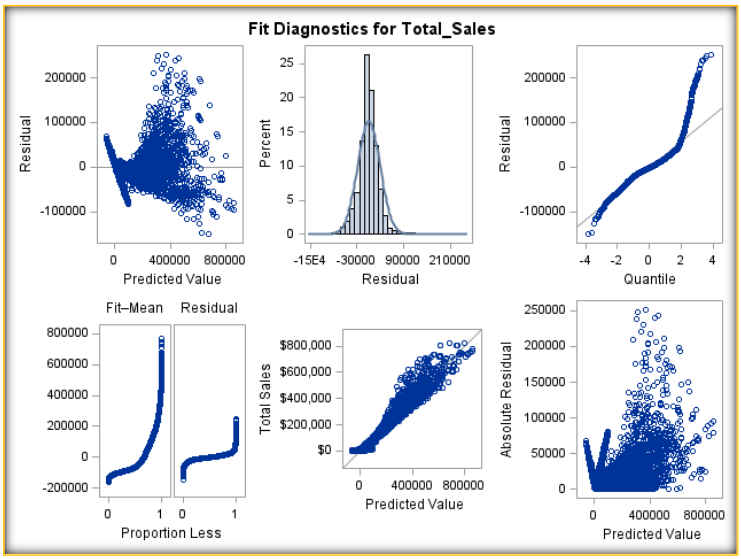
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.835818E14	6.119394E13	55101.8	<.0001
Error	9644	1.071025E13	1110561392		
Corrected Total	9647	1.942921E14			

Root MSE	33325	R-Square	0.9449
Dependent Mean	93273	Adj R-Sq	0.9449
Coeff Var	35.72838		

- In above result VIF of all features are less than 5. So there is no multicollinearity
- R Square is 94.49. Which Means 94.49% of variation in Total Sales is explained by independent Variables (numerical)
- There is not much difference in r squared valued after removing Units Sold. So lets us not skip it as Units Sold is importance from Calculation Total Sales.

REGRESSION ASSUMPTIONS

```
ods graphics on;
proc pls Data=Nandinis.Adidas plots=all;
class Sales_method retailer product region;
model Total_Sales= Operating_profit operating_margin units_sold price_per_unit sales_method Retailer Product region/solution;
run;
quit;
```



The image displays Fit Diagnostics for Total_Sales, which includes multiple residual plots used to assess the quality of a predictive model. Here's the interpretation of each plot:

Top Left: Residuals vs. Predicted Value

- This plot checks for homoscedasticity (constant variance).
- There is a clear pattern where residuals spread out more at higher predicted values, indicating heteroscedasticity (variance is not constant).
- This suggests that the model may not be appropriately capturing variability across different sales levels.

Top Middle: Histogram of Residuals

- This plot assesses whether residuals follow a normal distribution.
- The histogram is roughly symmetric, but there are deviations from normality.
- The presence of large residuals suggests possible outliers or skewness.

Top Right: Q-Q Plot of Residuals

- This plot assesses whether residuals follow a normal distribution.
- If residuals are normally distributed, points should fall along the diagonal line.
- The tails deviate significantly from the line, indicating that the residuals are not perfectly normal.

Bottom Left: Fit-Mean and Residual Proportion Plots

- These plots evaluate how residuals are distributed relative to their mean.
- The strong non-linearity indicates that the model may have systematic biases.

Bottom Middle: Total Sales vs. Predicted Value

- This plot compares actual sales with predicted values.
- The data follows an upward trend, suggesting that the model is capturing the general trend, but deviations exist.

Bottom Right: Absolute Residuals vs. Predicted Value

- This plot checks for heteroscedasticity.
- The residuals increase with predicted values, confirming heteroscedasticity.
- A transformation of the dependent variable (e.g., log transformation) or a different modeling approach (e.g., weighted regression) may help.

Overall Conclusion:

- The model shows heteroscedasticity (variance increases with predicted values).
- Residuals deviate from normality (as seen in the Q-Q plot and histogram).
- Possible outliers or influential points exist.
- Model refinement (e.g., transformation, adding interaction terms, or using non-linear regression) may be needed to improve predictive performance.

LINEAR REGRESSION

```
proc glm data=Nandinis.Adi_trans;
class Retailer Product Sales_Method REGION;
model Total_Sales = Retailer Product Sales_Method region Price_per_unit Units_sold
Operating_Profit operating_margin  /solution;
lsmeans sales_method Retailer Region Product/ pdiff stderr cl;
output out=outstat2
p=Predicted
r=Residual
stdr=se_Resid
student=Rstudent
h=Leverage
cookd=CooksD;
run;
quit;
```

The GLM Procedure

Dependent Variable: Total_Sales Total Sales

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	1.8617777E14	9.3088885E12	11044.3	<.0001
Error	9627	8.1143131E12	842870379		
Corrected Total	9647	1.9429208E14			

R-Square	Coeff Var	Root MSE	Total_Sales Mean
0.958237	31.12594	29032.23	93273.44

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Retailer	5	1.0499984E12	209999672960	249.15	<.0001
Product	5	4.5648856E12	912977112984	1083.18	<.0001
Sales_Method	2	3.0862513E13	1.5431256E13	18308.0	<.0001
Region	4	6.9229438E12	1.7307359E12	2053.38	<.0001
Price_per_Unit	1	3.3607559E13	3.3607559E13	39872.7	<.0001
Units_Sold	1	9.7027394E13	9.7027394E13	115115	<.0001
Operating_Profit	1	1.0171019E13	1.0171019E13	12067.1	<.0001
Operating_Margin	1	1.9714568E12	1.9714568E12	2338.98	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Retailer	5	28879921694	5775984338.8	6.85	<.0001
Product	5	142155800876	28431160175	33.73	<.0001
Sales_Method	2	73814903247	36907451624	43.79	<.0001
Region	4	49767666618	12441916654	14.76	<.0001
Price_per_Unit	1	918393406768	918393406768	1089.60	<.0001
Units_Sold	1	2.2909135E12	2.2909135E12	2717.99	<.0001
Operating_Profit	1	1.1822054E13	1.1822054E13	14025.9	<.0001
Operating_Margin	1	1.9714568E12	1.9714568E12	2338.98	<.0001

This output represents the General Linear Model (GLM) Procedure results for predicting Total Sales based on several independent variables.

1. Model Summary (Top Left Table)

- Model Sum of Squares = 1.8617777E12
- Error Sum of Squares = 8.1143131E12
- Corrected Total Sum of Squares = 1.9429208E12
- F Value = 11044.3, p < 0.0001-This suggests that the model is highly significant.
- R-Square = 0.9582-The model explains 95.82% of the variability in Total Sales, indicating an excellent fit.
- Root Mean Square Error (RMSE) = 29032.23-This measures the typical error in predictions. The lower, the better.
- Coefficient of Variation (Coeff Var) = 31.12594-A moderate CV suggests some variability in the model fit.

2. Factor Importance (Bottom Left Table)

- All predictor variables are highly significant ($p < 0.0001$).
- The highest F-values indicate the most impactful variables:
    - Sales_Method (F = 18308.0)
    - Price_per_Unit (F = 39872.7)
    - Units_Sold (F = 111515.0)
    - Operating_Profit (F = 12067.1)
    - Operating_Margin (F = 2338.98)
    - These variables have the largest impact on Total Sales.

3. Type III Sum of Squares (Right Table)

- Type III SS assesses the unique contribution of each predictor.
- Units_Sold has the highest impact (F = 2717.97)
- Price_per_Unit (F = 1089.60) and Operating_Profit (F = 14025.9) also play crucial roles.
- Retailer, Product, and Region also have significant effects but are less dominant than other factors.

Conclusion:

- The model fits the data very well ($R^2 = 95.82\%$).
- Units_Sold, Price_per_Unit, and Operating_Profit are the strongest predictors of Total Sales
- All factors significantly contribute to Total Sales, but their impact varies.

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	24819.4008	B	2402.533856	10.33	<.0001
Retailer Amazon	-976.1232	B	1216.520328	-0.80	0.4223
Retailer Foot Locker	-1070.1557	B	904.454724	-1.18	0.2368
Retailer Kohl's	-3409.4309	B	1107.152920	-3.08	0.0021
Retailer Sports Direct	-4854.6580	B	971.388289	-5.00	<.0001
Retailer Walmart	-3896.0679	B	1426.441859	-2.73	0.0063
Retailer West Gear	0.0000	B	.	.	.
Product Men's Apparel	-109.7870	B	1065.503572	-0.10	0.9179
Product Men's Athletic Footwear	-637.3037	B	1028.309175	-0.62	0.5354
Product Men's Street Footwear	-9817.0449	B	1069.781954	-9.18	<.0001
Product Women's Apparel	-1707.2314	B	1065.681217	-1.60	0.1092
Product Women's Athletic Footwear	3942.3277	B	1031.389281	3.82	0.0001
Product Women's Street Footwear	0.0000	B	.	.	.
Sales_Method In-store	7584.5429	B	978.560788	7.75	<.0001
Sales_Method Online	5898.2495	B	763.084196	7.73	<.0001
Sales_Method Outlet	0.0000	B	.	.	.

Region Midwest	4417.0206	B	1005.930577	4.39	<.0001
Region Northeast	332.0714	B	956.706084	0.35	0.7285
Region South	565.3635	B	1100.189163	0.51	0.6073
Region Southeast	-4245.0194	B	1142.902436	-3.71	0.0002
Region West	0.0000	B	.	.	.
Price_per_Unit	853.0026		25.841418	33.01	<.0001
Units_Sold	191.9430		3.681700	52.13	<.0001
Operating_Profit	1.6800		0.014185	118.43	<.0001
Operating_Margin	-186069.3509		3847.349225	-48.36	<.0001

- In above Table P value of following features is more than 0.05, which indicates these features are insignificant
    - Retailer Amazon
    - Retailer Footlocker
    - Product Men’s Apparel
    - Product Men’s Athletic Footwear

- Product Women's Apparel
- Region Northeast
- Region South

- Feature with P value less than 0.05 are Significant
  - Retailer Kohl's
  - Retailer Sports Direct
  - Retailer Walmart
  - Product Men's Street Footwear
  - Product Woman's Athletic Footwea
  - Sales Method In store
  - Sales Method Online
  - Region Midwest
  - Region Southeast
  - Price Per Unit
  - Units Sold
  - Operating Profit
  - Operating Margin

Least Square Means- Sales Method

This output analyzes the impact of different **Sales Methods (In-store, Online, Outlet)** on **Total Sales** using Least Squares Means (LSMEAN).

The GLM Procedure
Least Squares Means

Sales_Method	Total_Sales LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
In-store	95876.5866	853.6306	<.0001	1
Online	94190.2932	479.7057	<.0001	2
Outlet	88292.0437	596.2392	<.0001	3

Least Squares Means for effect Sales_Method
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Total_Sales

i/j	1	2	3
1		0.0902	<.0001
2	0.0902		<.0001
3	<.0001	<.0001	

Sales_Method	Total_Sales LSMEAN	95% Confidence Limits	
In-store	95877	94203	97550
Online	94190	93250	95131
Outlet	88292	87123	89461

Least Squares Means for Effect Sales_Method

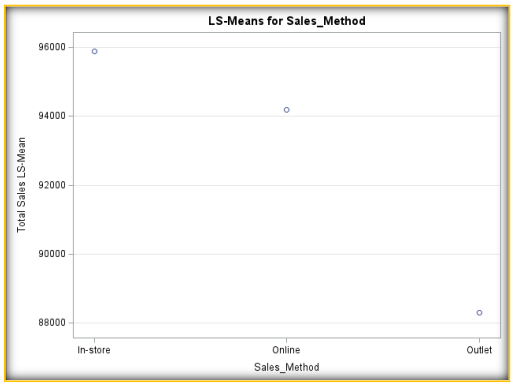
i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	1686.293376	-264.119483	3636.706234
1	3	7584.542874	5666.357808	9502.727940
2	3	5898.249498	4402.443896	7394.055100

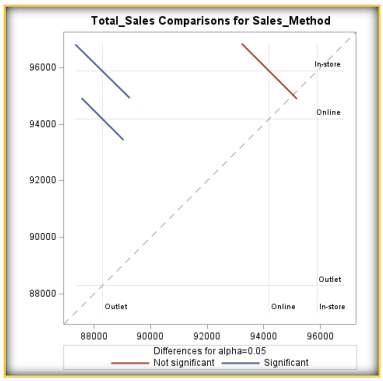
- Total_Sales LSMEAN: Represents the estimated mean total sales for each sales method after accounting for other variables.
- Standard Error: Measures the precision of the estimate.
- Pr > |t| (<.0001): All p-values are highly significant, indicating that the mean sales for each method are statistically different from zero.
- In-store has the highest mean sales (95,876.57), followed closely by Online (94,190.29), while Outlet has the lowest (88,292.04).
- The standard error is lowest for Online sales, indicating more stable estimates.

- Middle table tests whether sales methods differ significantly from each other using pairwise comparisons.
- P-values (Pr > |t|) tell us if differences between groups are significant**.**
- In-store vs. Online (p = 0.0902): No significant difference.
- In-store vs. Outlet (p < 0.0001): Significant difference.
- Online vs. Outlet (p < 0.0001): Significant difference.

- Conclusion: In-store and Online sales methods perform similarly, while Outlet sales are significantly lower.

- Bottom Table Provides confidence limits for each sales method's LSMEAN.
- In-store and Online confidence intervals overlap, reinforcing that they are not significantly different.
- Outlet sales are distinctly lower, as its confidence interval does not overlap with others.

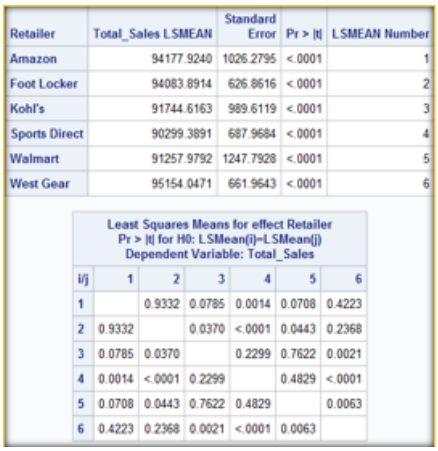




- Above Left Graph shows Instore Sales Method Outperform whereas Outlet Method has lower sales
- Colored Lines Indicate Statistical Significance:
    - Red Line (Not Significant): The difference between In-store and Online is not statistically significant at α = 0.05.
    - Blue Lines (Significant): The differences between Outlet vs. In-store and Outlet vs. Online are statistically significant.
- Dashed Diagonal Line:
    - Represents equality (no difference). The further apart the segments are from this line, the stronger the evidence of differences.

<u>Least Square Means- Retailers</u>

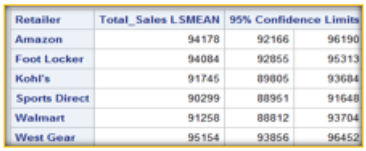
This output analyzes the impact of different **Retailers (Amazon, Foot Locker, Kohl's Sports Direct Walmart,West Gear)** on **Total Sales** using Least Squares Means (LSMEAN).



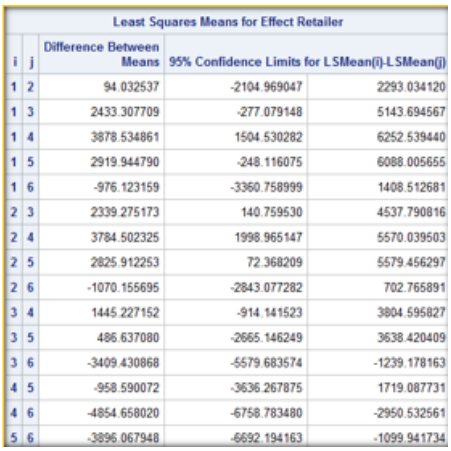
Retailer	Total_Sales LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
Amazon	94177.9240	1026.2795	<.0001	1
Foot Locker	94083.8914	626.8616	<.0001	2
Kohl's	91744.6163	989.6119	<.0001	3
Sports Direct	90299.3891	687.9684	<.0001	4
Walmart	91257.9792	1247.7928	<.0001	5
West Gear	95154.0471	661.9643	<.0001	6

Least Squares Means for effect Retailer
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Total_Sales

i/j	1	2	3	4	5	6
1		0.9332	0.0785	0.0014	0.0708	0.4223
2	0.9332		0.0370	<.0001	0.0443	0.2368
3	0.0785	0.0370		0.2299	0.7622	0.0021
4	0.0014	<.0001	0.2299		0.4829	<.0001
5	0.0708	0.0443	0.7622	0.4829		0.0063
6	0.4223	0.2368	0.0021	<.0001	0.0063	



Retailer	Total_Sales LSMEAN	95% Confidence Limits	
Amazon	94178	92166	96190
Foot Locker	94084	92855	95313
Kohl's	91745	89805	93684
Sports Direct	90299	88951	91648
Walmart	91258	88812	93704
West Gear	95154	93856	96452



Least Squares Means for Effect Retailer

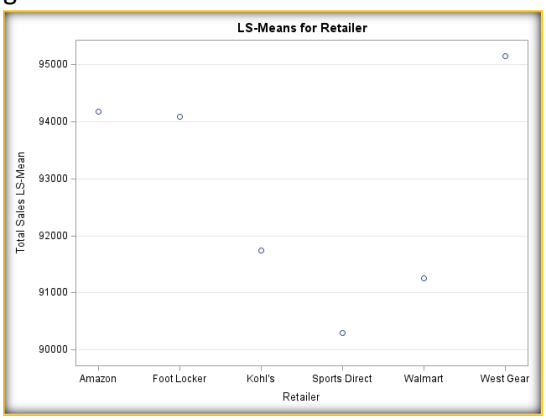
i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	94.032537	-2104.969047	2293.034120
1	3	2433.307709	-277.079148	5143.694567
1	4	3878.534861	1504.530282	6252.539440
1	5	2919.944790	-248.116075	6088.005655
1	6	-976.123159	-3360.758999	1408.512681
2	3	2339.275173	140.759530	4537.790816
2	4	3784.502325	1998.965147	5570.039503
2	5	2825.912253	72.368209	5579.456297
2	6	-1070.155695	-2843.077282	702.765891
3	4	1445.227152	-914.141523	3804.595827
3	5	486.637080	-2665.146249	3638.420409
3	6	-3409.430868	-5579.683574	-1239.178163
4	5	-958.590072	-3636.267875	1719.087731
4	6	-4854.658020	-6758.783480	-2950.532561
5	6	-3896.067948	-6692.194163	-1099.941734

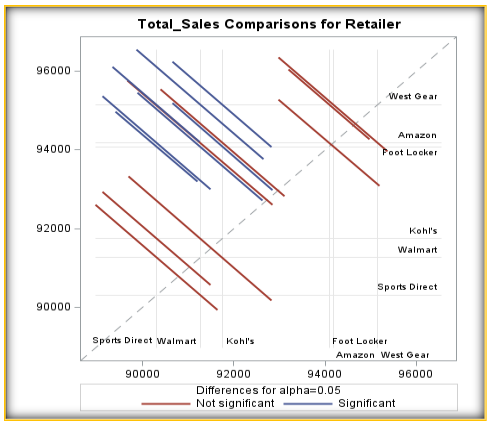
- Total_Sales LSMEAN: Represents the estimated mean total sales for each Retailer after accounting for other variables.
- Standard Error: Measures the precision of the estimate.
- Pr > |t| (<.0001): All p-values are highly significant, indicating that the mean sales for each Retailer are statistically different from zero.
- West Gear has the highest mean total sales (95,154.04), while Sports Direct has the lowest (90,299.39).
- Amazon and Foot Locker perform similarly, with total sales around 94,000.
- Kohl's, Walmart, and Sports Direct have lower estimated sales, but differences may not be statistically significant.

- Middle Table compares each retailer's mean total sales against every other retailer.
- Values close to 1 indicate no significant difference.
- Values close to 0 indicate significant differences.
- P-values (Pr > |t|) tell us if differences between groups are significant.
- Amazon vs. Foot Locker (0.9332): No significant difference.
- Amazon vs. West Gear (0.4223): No significant difference.
- Kohl's vs. Sports Direct (0.2299): Possibly significant difference.
- Sports Direct vs. Walmart (0.4829): No significant difference.

- Right Table provides the actual difference between each pair of retailers along with confidence intervals.
- If the confidence interval includes zero, the difference is not statistically significant
- Amazon vs. Sports Direct: Amazon has significantly higher total sales (Difference = 3878.53, CI = 1504.53 to 6252.53).
- Foot Locker vs. Sports Direct: Foot Locker also outperforms Sports Direct (Difference = 3784.50, CI = 1998.96 to 5570.03).
- West Gear vs. Foot Locker: West Gear significantly outperforms Foot Locker (Difference = -4854.68, CI = -7236.93 to -2472.44).
- Conclusion:
- West Gear, Amazon, and Foot Locker perform similarly, with no significant differences among them.

- Sports Direct has significantly lower sales compared to Amazon and Foot Locker.
- Kohl's and Walmart have mixed performance, with some differences being statistically significant.





- Above Left Graph shows West Gear Outperform whereas Sports Direct has lower sales
- Colored Lines Indicate Statistical Significance:
  - Red Line (Not Significant):
    Amazon and Foot Locker do not show significant differences in total sales.
    Amazon and West Gear are also not significantly different.
    Kohl's and Walmart have similar total sales, suggesting they are in the same performance group.
  - Blue Line (Significant):
    Sports Direct & Walmart have significantly lower sales than West Gear, Amazon, and Foot Locker.
    Kohl's has significantly lower sales than West Gear and Amazon.
    West Gear has significantly higher sales than several other retailers.
- Dashed Diagonal Line:
  - Represents equality (no difference). The further apart the segments are from this line, the stronger the evidence of differences.

Least Square Means- Region

This output analyzes the impact of different **Region (Midwest, Northeast, South, SouthEast, West)** on **Total Sales** using Least Squares Means (LSMEAN).

Region	Total_Sales LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
Midwest	96989.4412	766.0564	<.0001	1
Northeast	92904.4920	652.4550	<.0001	2
South	93137.7841	849.3724	<.0001	3
Southeast	88327.4013	922.4045	<.0001	4
West	92572.4207	737.3478	<.0001	5

Least Squares Means for effect Region
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Total_Sales

i/j	1	2	3	4	5
1		<.0001	0.0006	<.0001	<.0001
2	<.0001		0.8326	<.0001	0.7285
3	0.0006	0.8326		<.0001	0.6073
4	<.0001	<.0001	<.0001		0.0002
5	<.0001	0.7285	0.6073	0.0002	

Region	Total_Sales LSMEAN	95% Confidence Limits	
Midwest	96989	95488	98491
Northeast	92904	91626	94183
South	93138	91473	94803
Southeast	88327	86519	90136
West	92572	91127	94018

Least Squares Means for Effect Region

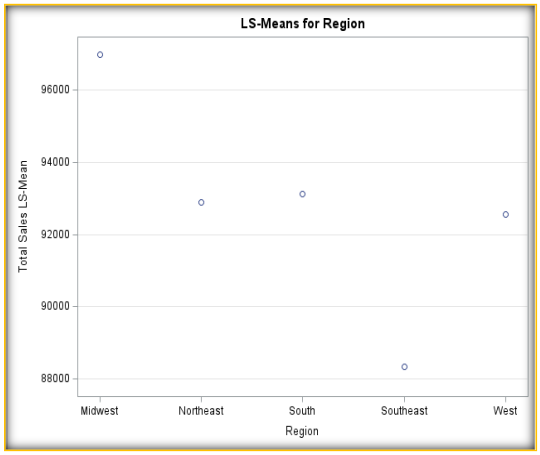
i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	4084.949194	2231.949545	5937.948844
1	3	3851.657097	1651.187723	6052.126471
1	4	8662.039963	6415.525958	10909
1	5	4417.020564	2445.184951	6388.856177
2	3	-233.292097	-2396.672777	1930.088582
2	4	4577.090768	2431.731634	6722.449903
2	5	332.071370	-1543.273878	2207.416618
3	4	4810.382866	2498.902729	7121.863002
3	5	565.363467	-1591.238810	2721.965744
4	5	-4245.019398	-6485.348679	-2004.690118

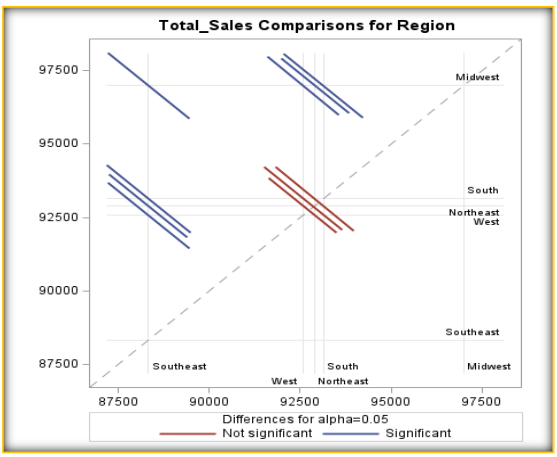
- Total_Sales LSMEAN: Represents the estimated mean total sales for each Region after accounting for other variables.
- Standard Error: Measures the precision of the estimate.
- Pr > |t| (<.0001): All p-values are highly significant, indicating that the mean sales for each Region are statistically different from zero.
- Midwest has the highest mean total sales (96989.44), while Southest has the lowest (88327.4).
- Northeast and South perform similarly, with total sales around 93,000.

- Middle Table compares each Region's mean total sales against every other Region.
- Values close to 1 indicate no significant difference.
- Values close to 0 indicate significant differences.
- P-values (Pr > |t|) tell us if differences between groups are significant.
- Northeast Vs. South (0.8326): No significant difference.
- Northeast Vs West (0.7285): No significant difference.
- South Vs West (0.6073): No significant difference.

- Right Table provides the actual difference between each pair of Region along with confidence intervals.
- If the confidence interval includes zero, the difference is not statistically significant
- Midwest Vs Southeast: Midwest has significantly higher total sales (Difference = 8662.04, CI = 6415.52 to 10909).
- South Vs South East : South also outperforms Southeast (Difference = 4810.38, CI = 2498.90 to 7121.86).
- Midwest Vs. West : Midwest significantly outperforms West (Difference = 4417.02, CI = 2445.18 to 6388.85).
- Midwest Vs. Northeast : Midwest significantly outperforms Northeast (Difference = 4084.95, CI = 2231.95 to 5937.95).
- Northeast Vs. Southeast : Northeast significantly outperforms Southeast (Difference = 4577.09, CI = 2431.73 to 6722.45).
- Southeast Vs. West : Southeast significantly outperforms West (Difference = -4245.02, CI = -3485.35 to -2004.69).

- Conclusion:
- Northeast ,South and West perform similarly, with no significant differences among them.
- Southeast has significantly lower sales compared to Midwest





- Above Left Graph shows Midwest Outperform whereas Southeast has lower sales
- Colored Lines Indicate Statistical Significance:
  - Red Line (Not Significant):
    Comparisons involving the South, Northeast, and West regions are mostly not significant, indicating that their total sales may be similar.
  - Blue Line (Significant):
    Comparisons involving the Midwest and Southeast regions seem to be statistically significant suggesting that their total sales differ from other regions
- Dashed Diagonal Line: Represents equality (no difference). The further apart the segments are from this line, the stronger the evidence of differences

Least Square Means- Product

This output analyzes the impact of different **Product (Men's Apparel ,Men's Athletic Footwear,Men's Street Footwear, Women's Apparel, Women's Athletic Footwear, Women's Street Footwear)** on **Total Sales** using Least Squares Means (LSMEAN).

Product	Total_Sales LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
Men's Apparel	94064.6940	789.9123	<.0001	1
Men's Athletic Footwear	93537.1774	783.9860	<.0001	2
Men's Street Footwear	84357.4362	828.3616	<.0001	3
Women's Apparel	92467.2497	782.5679	<.0001	4
Women's Athletic Footwear	98116.8088	777.7258	<.0001	5
Women's Street Footwear	94174.4811	785.6180	<.0001	6

Least Squares Means for effect Product
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Total_Sales

i/j	1	2	3	4	5	6
1		0.6184	<.0001	0.1262	0.0001	0.9179
2	0.6184		<.0001	0.3094	<.0001	0.5354
3	<.0001	<.0001		<.0001	<.0001	<.0001
4	0.1262	0.3094	<.0001		<.0001	0.1092
5	0.0001	<.0001	<.0001	<.0001		0.0001
6	0.9179	0.5354	<.0001	0.1092	0.0001	

Product	Total_Sales LSMEAN	95% Confidence Limits	
Men's Apparel	94065	92516	95613
Men's Athletic Footwear	93537	92000	95074
Men's Street Footwear	84357	82734	85981
Women's Apparel	92467	90933	94001
Women's Athletic Footwear	98117	96592	99641
Women's Street Footwear	94174	92635	95714

Least Squares Means for Effect Product

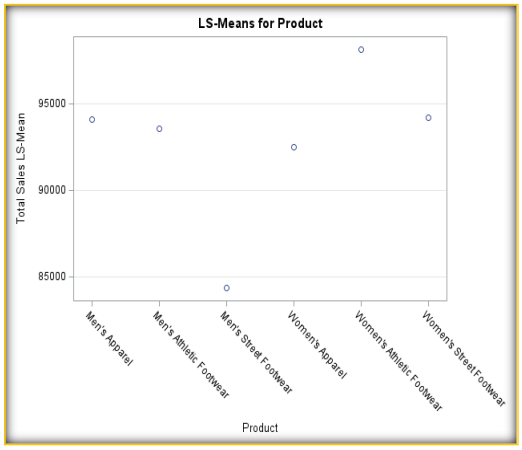
i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	527.516639	-1548.475780	2603.509058
1	3	9707.257881	7496.232565	11918
1	4	1597.444368	-450.109982	3644.998718
1	5	-4052.114761	-6109.575950	-1994.653572
1	6	-109.787019	-2198.398237	1978.824199
2	3	9179.741242	7103.081429	11256
2	4	1069.927729	-993.059335	3132.914794
2	5	-4579.631400	-6618.447236	-2540.815564
2	6	-637.303658	-2653.006032	1378.398716
3	4	-8109.813513	-10209	-6010.527137
3	5	-13759	-15903	-11616
3	6	-9817.044900	-11914	-7720.047152
4	5	-5649.559129	-7729.126140	-3569.992118
4	6	-1707.231387	-3796.190827	381.728053
5	6	3942.327742	1920.587712	5964.067772

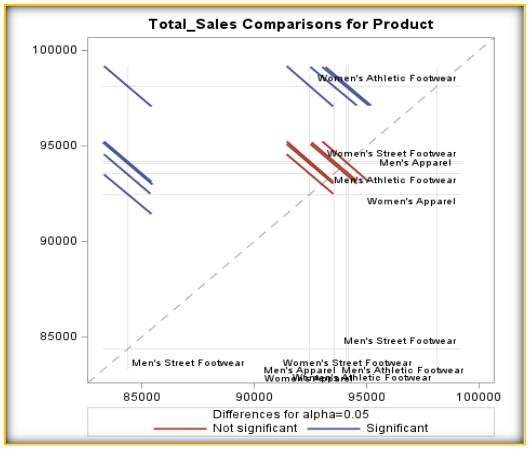
- Total_Sales LSMEAN: Represents the estimated mean total sales for each Product after accounting for other variables.
- Standard Error: Measures the precision of the estimate.
- Pr > |t| (<.0001): All p-values are highly significant, indicating that the mean sales for each Product are statistically different from zero.
- Women's athletic footwear has the highest mean total sales (98117), while Men's Street Footwear has the lowest (84357).
- Men's Apparel and Women's Street Footwear perform similarly, with total sales around 94,000.

- Middle Table compares each Product's mean total sales against every other Product.
- Values close to 1 indicate no significant difference.
- Values close to 0 indicate significant differences.
- P-values (Pr > |t|) tell us if differences between groups are significant.
- Men's Apparel Vs. Men's Athletic footwear (0.6184): No significant difference.
- Men's Apparel Vs. Women's Street footwear (0.9179): No significant difference.
- Men's Athletic footwear Vs Women's Street footwear (0.5354): No significant difference.
- Men's Athletic footwear vs. Women's Apparel (0.3094): Possibly significant difference.

- Right Table provides the actual difference between each pair of Product along with confidence intervals.
- If the confidence interval includes zero, the difference is not statistically significant
- Men's Apparel Vs Men's Street Footwear: Men's Apparel has significantly higher total sales (Difference = 9707.26, CI = 7496.23 to 11918).
- Men's Athletic Footwear Vs Men's Street Footwear: Men's Athletic Footwear has significantly higher total sales (Difference = 9179.74 CI = 7103.08 to 11256).
- Men's Street Footwear Vs. Women's Street Footwear : Men's Street Footwear significantly outperforms Women's Street Footwear (Difference = -9817.04, CI = -11914 to -7220.05).
- Men's Street Footwear Vs. Women's Apparel: Men's Street Footwear significantly outperforms Women's Apparel (Difference = -8109.81, CI = -10209 to -6010.52).

- Conclusion:
- Men's Apparel ,Women's Street Footwear, Men's Athletic Footwear perform similarly, with no significant differences among them.
- Men's Street Footwear has significantly lower sales compared to Women's Athletic Footwear.





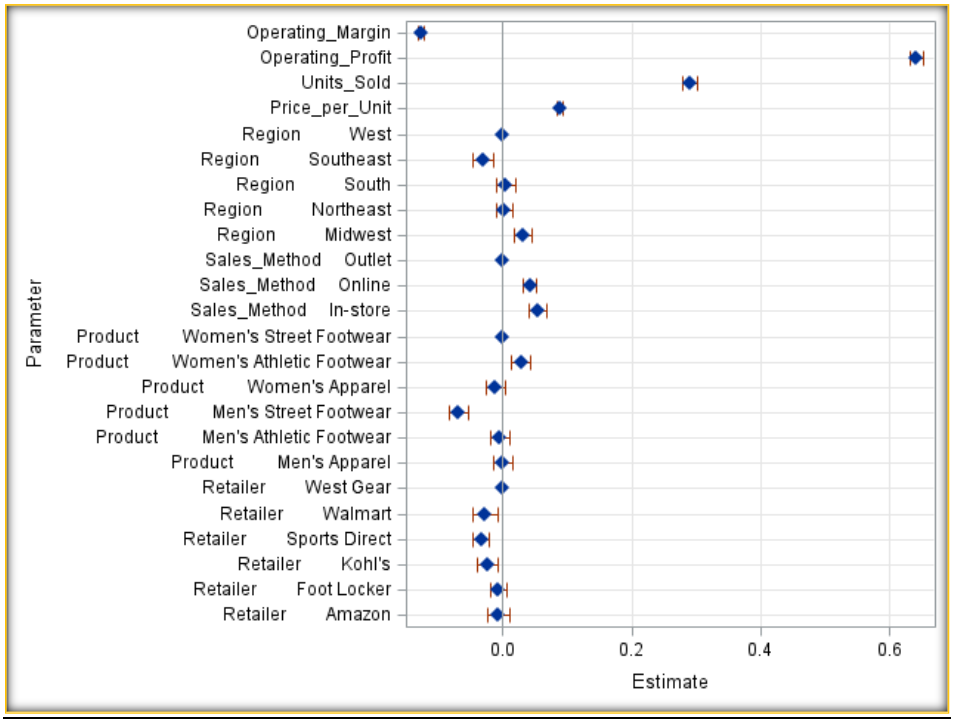
- Above Left Graph shows Women's Athletic Footwear.Outperform whereas Men's Street Footwear has lower sales
- Colored Lines Indicate Statistical Significance:
  - Red Line (Not Significant):
    Comparisons involving the Women's Street Footwear,Men's Apparel,Men's Atheletic Footwear,Women's Apparel are mostly not significant, indicating that their total sales may be similar.
  - Blue Line (Significant):
    Comparisons involving the Men's Street Footwear and Women's Athletic Footwear seem to be statistically significant suggesting that their total sales differ from other regions
- Dashed Diagonal Line: Represents equality (no difference). The further apart the segments are from this line, the stronger the evidence of differences

Visualization of Coefficients

```
proc standard data=Nandinis.Adi_Trans out=Nandinis.Adidas_stand mean=0 std=1;
var Units_Sold operating_Profit operating_margin total_sales price_per_unit;
run;
```

```
proc glm data=Nandinis.Adidas_stand;
class Retailer Product Sales_Method REGION;
model Total_Sales = Retailer Product Sales_Method region Price_per_unit
Units_sold Operating_Profit operating_margin  /solution ss3 clparm;
output out=outstat3
p=Predicted
r=Residual
stdr=se_Resid
student=Rstudent
h=Leverage
cookd=CooksD;
ods output ParameterEstimates =Nandinis.Paramest;
run;
quit;
```

```
proc sgplot Data=Nandinis.Paramest;
where parameter ne "Intercept";
scatter y=parameter x=estimate /xerrorlower=LowerCl xerrorupper=UpperCl markerattrs=(symbol=diamondfilled);
refline 0/axis=x;
xaxis grid;
yaxis grid;
run;
```



Top Predictors

- Operating Profit
- Units Sold
- Price per Unit
- Operating Margin

Not Significant

- Region South
- Region Northeast
- Product Woman's Apparel
- Product Men's Apparel
- Retailer Footlocker
- Retailer Amazon

This is a forest plot displaying parameter estimates with confidence intervals from a statistical model.

1. X-axis (Estimate): This represents the estimated effect size of each parameter. The red vertical line at 0 likely represents the null hypothesis (no effect).
2. Y-axis (Parameter): Lists various parameters analyzed in the model, including financial metrics *(e.g.,* Operating_Margin*,* Operating_Profit*),* sales factors *(e.g.,* Units_Sold*,* Price_per_Unit*),* regions, sales methods, products, and retailers.
3. Blue diamonds: Represent point estimates of each parameter.
4. Red error bars: Indicate the confidence intervals (likely 95%). Wider bars mean greater uncertainty, while narrow bars indicate more precise estimates.

Key Observations:

- Operating_Margin and Operating_Profit have the highest positive estimates, meaning they have a strong impact on the dependent variable.
- Price_per_Unit also has a notable positive estimate.
- Units_Sold appears to have a significant effect.
- Most of the other estimates are close to zero, suggesting they have little or no effect.
- Some regions and sales methods have slightly positive or negative estimates, but their confidence intervals cross zero, meaning their effects may not be statistically significant.

Splitting Data Into Train and Test (70:30)

```
title;
proc surveyselect Data=Nandinis.Adi_stand rate=0.70 outall out=Nandinis.Adi_result seed=1234;
run;

Data Nandinis.Train Nandinis.Test;
set Nandinis.Adi_result;
if selected =1 then output Nandinis.Train;
else output Nandinis.Test;
run;
```

LINEAR REGRESSION- PREDICTIVE MODELLING

```
proc glmselect Data=Nandinis.Train testdata=Nandinis.Test plots=all;
class Retailer Sales_method Product Region;
model Total_Sales=Operating_profit Operating_margin Units_sold Price_per_unit Retailer Sales_method Region
Product/selection =lasso(stop=none);
score data=Nandinis.test out=Nandinis.Testpred;
output out =Outputedata p=prob_predicted r=residual;
run;
```

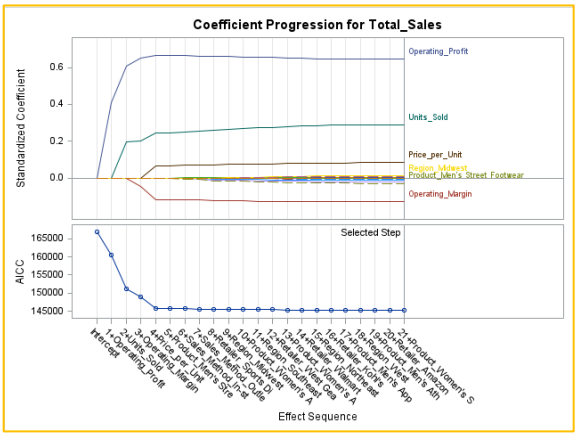
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	20	1.293918E14	6.46959E12	7912.40
Error	6733	5.505254E12	817652482	
Corrected Total	6753	1.348971E14		

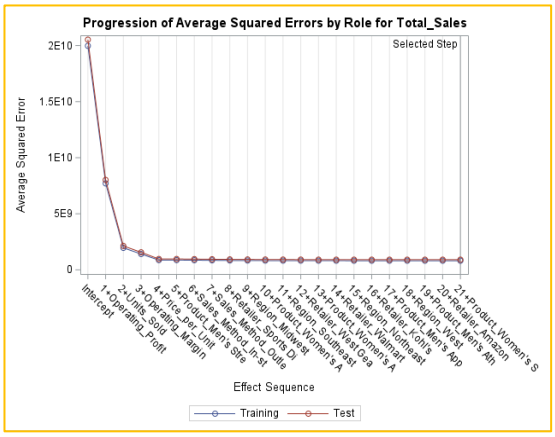
Root MSE	28595
Dependent Mean	92840
R-Square	0.9592
Adj R-Sq	0.9591
AIC	145382
AICC	145382
SBC	138769
ASE (Train)	815110181
ASE (Test)	905935666

Parameter	DF	Estimate
Intercept	1	31748
Operating_Profit	1	1.691713
Operating_Margin	1	-185758
Units_Sold	1	191.019394
Price_per_Unit	1	803.655850
Retailer_Amazon	1	2.624555
Retailer_Kohl's	1	-443.702567
Retailer_Sports Direct	1	-4093.914224
Retailer_Walmart	1	-1746.090033
Retailer_West Gear	1	1129.286334

Sales_Method_In-store	1	229.570145
Sales_Method_Outlet	1	-6026.116080
Region_Midwest	1	4238.647288
Region_Northeast	1	786.571435
Region_Southeast	1	-3438.072148
Region_West	1	-314.637370
Product_Men's Apparel	1	213.730994
Product_Men's Athletic Footwear	1	-21.774194
Product_Men's Street Footwear	1	-10502
Product_Women's Apparel	1	-2356.999050
Product_Women's Athletic Footwear	1	3301.132350
Product_Women's Street Footwear	0	0

- Root MSE: 28595
- Dependent Mean: 92840
- R-Square: 0.9592
- Adj R-Sq: 0.9591
- Model Fit: The high R-Square value (0.9592) indicates that the model explains about 95.92% of the variability in the data, which is a strong fit.
- F Value: The F Value (7912.40) is high, suggesting that the model is statistically significant.
- Root MSE: The Root Mean Square Error (28595) measures the standard deviation of the residuals, which indicates the average error of the model's predictions.
- Intercept: The base level of the dependent variable when all predictors are at zero.
- Operating_Profit: Positive estimate (1.691713) indicates a positive relationship with the dependent variable.
- Operating_Margin: Negative estimate (-185758) suggests a negative relationship with the dependent variable.
- Units_Sold & Price_per_Unit: Positive estimates (191.019394 and 803.655850, respectively) indicate these factors have a positive influence on the dependent variable.
- Retailers & Sales Methods: The estimates for retailers and sales methods show the influence of different retailers and sales methods on the dependent variable.
- Regions & Products: The estimates for regions and products show their influence on the dependent variable, with varying positive and negative impacts.





Left image represents the coefficient progression for Total Sales using stepwise regression or another model selection approach.

Top Plot: Coefficient Progression

- Y-axis (Standardized Coefficient): Represents the relative effect of each variable on Total Sales.
- X-axis (Effect Sequence): Displays the order in which variables were added to the model.
- Lines: Each colored line represents a different predictor variable. Their movement over the sequence shows how their effect changes as more variables are included in the model.
    - Operating_Profit (blue) has the strongest effect, stabilizing at a high value.
    - Units_Sold (green) also has a significant positive impact.
    - Price_per_Unit (orange) has a moderate but positive effect.

    - Region_Midwest (yellow) and Product_Women's Street Footwear (brown) have relatively small effects.
    - Operating_Margin (red) starts negative but becomes insignificant.
- Bottom Plot: AICC Progression
- Y-axis (AICC - Corrected Akaike Information Criterion): Measures model quality (lower is better).
- X-axis (Effect Sequence): Represents the sequence in which variables were added.
- The AICC drops steeply at the beginning and then stabilizes, indicating that the first few variables (Operating_Profit, Units_Sold, Price_per_Unit) contribute the most to model performance.

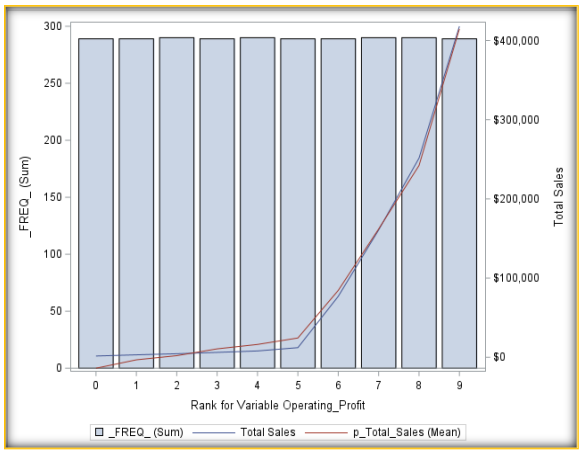
- Conclusion
- Operating_Profit and Units_Sold are the most important predictors of Total Sales.
- Price_per_Unit contributes positively but is less significant.
- Region and product categories have minimal effects.
- Operating_Margin initially appears negatively correlated but loses significance as more variables are added.
- The model reaches an optimal point quickly, after which adding more predictors does not improve AICC.

Right plot illustrates the progression of average squared errors for the Total Sales model as variables are added. It evaluates both training and test performance across different model specifications.

- Y-axis (Average Squared Error): Measures the error magnitude in predicting total sales.
- X-axis (Effect Sequence): Shows the order in which variables are included in the model.
- Blue Line (Training Data) & Red Line (Test Data): Represent the error progression for training and test datasets.
- Most of the predictive power comes from the first few variables (e.g., Operating_Profit, Units_Sold).
- Adding more variables beyond a certain point does not significantly improve the model.
- The model generalizes well, as training and test errors remain consistent.

Visualize Prediction By Operating Profit

```
proc means Data=Nandinis.Testpred n min p10 p30 p40 p50 p60 p70 p80 p90 max maxdec=2;
var Operating_profit;
run;

proc rank Data=Nandinis.Testpred out=Nandinis.Testpred_percent groups=10;
var Operating_profit;
ranks rank;
run;

proc summary Data=Nandinis.Testpred_percent;
class rank;
var total_Sales p_total_sales;
output out=Nandinis.pred_op mean=;
quit;

proc sgplot Data=Nandinis.pred_op;
vbar rank /response =_freq_;
vline rank / response =Total_Sales y2axis stat=mean;
vline rank /response=p_Total_Sales y2axis stat=mean;
run;

Proc means Data=Nandinis.pred_op;
class rank;
var total_sales p_total_sales;
run;
```



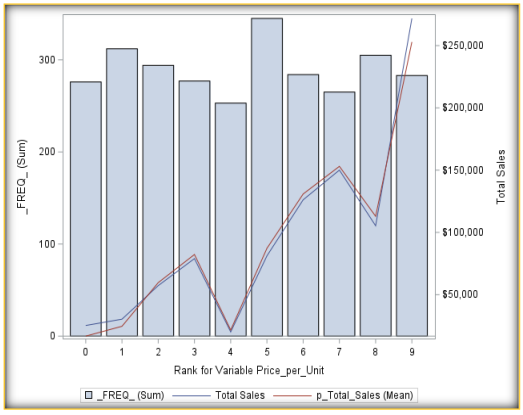
Rank for Variable Operating_Profit	N Obs	Variable	Label	N	Mean	Std Dev	Minimum	Maximum
0	1	Total_Sales	Total Sales	1	1594.94	.	1594.94	1594.94
		p_Total_Sales		1	-13836.62	.	-13836.62	-13836.62
1	1	Total_Sales	Total Sales	1	3092.88	.	3092.88	3092.88
		p_Total_Sales		1	-3236.02	.	-3236.02	-3236.02
2	1	Total_Sales	Total Sales	1	4471.85	.	4471.85	4471.85
		p_Total_Sales		1	2038.37	.	2038.37	2038.37
3	1	Total_Sales	Total Sales	1	6153.36	.	6153.36	6153.36
		p_Total_Sales		1	10609.71	.	10609.71	10609.71
4	1	Total_Sales	Total Sales	1	7982.23	.	7982.23	7982.23
		p_Total_Sales		1	16214.27	.	16214.27	16214.27
5	1	Total_Sales	Total Sales	1	12126.90	.	12126.90	12126.90
		p_Total_Sales		1	24341.09	.	24341.09	24341.09
6	1	Total_Sales	Total Sales	1	76965.55	.	76965.55	76965.55
		p_Total_Sales		1	84484.31	.	84484.31	84484.31
7	1	Total_Sales	Total Sales	1	160737.07	.	160737.07	160737.07
		p_Total_Sales		1	162015.18	.	162015.18	162015.18
8	1	Total_Sales	Total Sales	1	251547.41	.	251547.41	251547.41
		p_Total_Sales		1	242165.40	.	242165.40	242165.40
9	1	Total_Sales	Total Sales	1	418006.92	.	418006.92	418006.92
		p_Total_Sales		1	414645.20	.	414645.20	414645.20

- The above Graph reveal how "Operating_Profit" influences "Total Sales", and it provides insights into the mean, standard deviation, and other statistical measures across various ranks.
- The graph shows the relationship between the rank for the variable "Operating_Profit" and two metrics: *"*FREQ *(*Sum*)"* and "Total Sales".
- The graph includes two lines: a blue line representing "FREQ (Sum)" and a red line representing "Total Sales (Mean)".
- As the rank for "Operating_Profit" increases, both "FREQ (Sum)" and "Total Sales (Mean)" increase, with a significant rise in "Total Sales (Mean)" at higher ranks.
- **Table**: The table provides detailed statistics for each rank of the variable "Operating_Profit", including the number of observations (N), mean, standard deviation (Std Dev), minimum, and maximum values for each variable.
- Higher Operating Profit correlates with significantly higher Total Sales

Visualize Prediction By Units Sold

```
proc means Data=Nandinis.Testpred n min p10 p30 p40 p50 p60 p70 p80 p90 max maxdec=2;
var Units_sold;
run;

proc rank Data=Nandinis.Testpred out=Nandinis.Testpred_percent groups=10;
var Units_sold;
ranks rank;
run;

proc summary Data=Nandinis.Testpred_percent;
class rank;
var total_Sales p_total_sales;
output out=Nandinis.pred_us mean=;
quit;

proc sgplot Data=Nandinis.pred_us;
vbar rank /response =_freq_;
vline rank / response =Total_Sales y2axis stat=mean;
vline rank /response=p_Total_Sales y2axis stat=mean;
run;

Proc means Data=Nandinis.pred_us;
class rank;
var total_sales p_total_sales;
run;
```

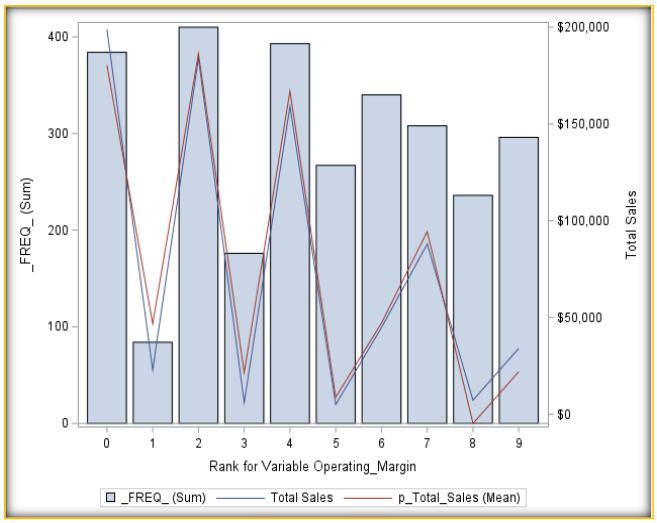
Rank for Variable Price_per_Unit	N Obs	Variable	Label	N	Mean	Std Dev	Minimum	Maximum
0	1	Total_Sales	Total Sales	1	25257.87	.	25257.87	25257.87
		p_Total_Sales		1	16718.16	.	16718.16	16718.16
1	1	Total_Sales	Total Sales	1	30334.06	.	30334.06	30334.06
		p_Total_Sales		1	24625.61	.	24625.61	24625.61
2	1	Total_Sales	Total Sales	1	57259.88	.	57259.88	57259.88
		p_Total_Sales		1	59653.10	.	59653.10	59653.10
3	1	Total_Sales	Total Sales	1	78976.61	.	78976.61	78976.61
		p_Total_Sales		1	82262.02	.	82262.02	82262.02
4	1	Total_Sales	Total Sales	1	20085.55	.	20085.55	20085.55
		p_Total_Sales		1	21627.51	.	21627.51	21627.51
5	1	Total_Sales	Total Sales	1	81140.45	.	81140.45	81140.45
		p_Total_Sales		1	87220.42	.	87220.42	87220.42
6	1	Total_Sales	Total Sales	1	126174.08	.	126174.08	126174.08
		p_Total_Sales		1	130848.15	.	130848.15	130848.15
7	1	Total_Sales	Total Sales	1	150028.80	.	150028.80	150028.80
		p_Total_Sales		1	153037.68	.	153037.68	153037.68
8	1	Total_Sales	Total Sales	1	105320.28	.	105320.28	105320.28
		p_Total_Sales		1	112937.09	.	112937.09	112937.09
9	1	Total_Sales	Total Sales	1	271815.64	.	271815.64	271815.64
		p_Total_Sales		1	252912.99	.	252912.99	252912.99



- Bar Chart:
- X-axis: Represents the rank for the variable "Units_Sold," ranging from 0 to 9.
- Y-axis (left): Represents *FREQ* (Sum).
- Y-axis (right): Represents Total Sales.
- Bars: Represent the *FREQ* (Sum).
- Blue Line: Represents Total Sales.
- Red Line: Represents p_Total_Sales (Mean).
- As the rank for "Units_Sold" increases, both Total Sales and p_Total_Sales (Mean) increase significantly.
- Table: The table provides detailed statistics for each rank of the variable "Units_Sold"
- As the rank for "Units_Sold" increases, both Total Sales and p_Total_Sales (Mean) also increase. This indicates a positive correlation between the rank for "Units_Sold" and the sales figures.

Visualize Prediction By Operating Margin

```
proc means Data=Nandinis.Testpred n min p10 p30 p40 p50 p60 p70 p80 p90 max maxdec=2;
var Operating_margin;
run;

proc rank Data=Nandinis.Testpred out=Nandinis.Testpred_percent groups=10;
var Operating_margin;
ranks rank;
run;

proc summary Data=Nandinis.Testpred_percent;
class rank;
var total_Sales p_total_sales;
output out=Nandinis.pred_om mean=;
quit;

proc sgplot Data=Nandinis.pred_om;
vbar rank /response =_freq_;
vline rank / response =Total_Sales y2axis stat=mean;
vline rank /response=p_Total_Sales y2axis stat=mean;
run;

Proc means Data=Nandinis.pred_om;
class rank;
var total_sales p_total_sales;
run;
```

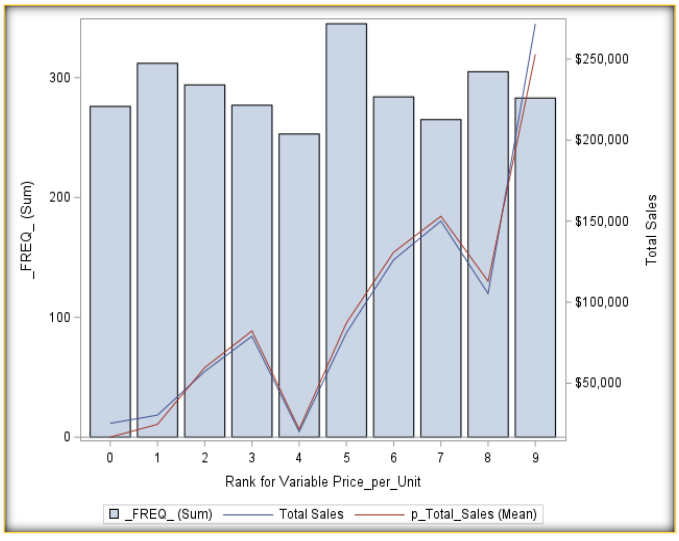


Rank for Variable Operating_Margin	N Obs	Variable	Label	N	Mean	Std Dev	Minimum	Maximum
0	1	Total_Sales	Total Sales	1	198935.38	.	198935.38	198935.38
		p_Total_Sales		1	180435.87	.	180435.87	180435.87
1	1	Total_Sales	Total Sales	1	22815.05	.	22815.05	22815.05
		p_Total_Sales		1	46741.46	.	46741.46	46741.46
2	1	Total_Sales	Total Sales	1	184012.30	.	184012.30	184012.30
		p_Total_Sales		1	186381.60	.	186381.60	186381.60
3	1	Total_Sales	Total Sales	1	6096.86	.	6096.86	6096.86
		p_Total_Sales		1	21592.26	.	21592.26	21592.26
4	1	Total_Sales	Total Sales	1	159172.87	.	159172.87	159172.87
		p_Total_Sales		1	167004.30	.	167004.30	167004.30
5	1	Total_Sales	Total Sales	1	5052.69	.	5052.69	5052.69
		p_Total_Sales		1	8844.89	.	8844.89	8844.89
6	1	Total_Sales	Total Sales	1	44808.18	.	44808.18	44808.18
		p_Total_Sales		1	47145.29	.	47145.29	47145.29
7	1	Total_Sales	Total Sales	1	87980.42	.	87980.42	87980.42
		p_Total_Sales		1	94392.63	.	94392.63	94392.63
8	1	Total_Sales	Total Sales	1	7264.79	.	7264.79	7264.79
		p_Total_Sales		1	-4729.23	.	-4729.23	-4729.23
9	1	Total_Sales	Total Sales	1	34060.87	.	34060.87	34060.87
		p_Total_Sales		1	21993.21	.	21993.21	21993.21

- Bar Chart Interpretation
- X-axis: Represents the rank for the variable "Operating_Margin," ranging from 0 to 9.
- Y-axis (left): Represents FREQ (Sum) with values ranging from 0 to 400.
- Y-axis (right): Represents Total Sales with values ranging from $0 to $200,000.
- Bars: Represent the FREQ (Sum).
- Blue Line: Represents Total Sales.
- Red Line: Represents p_Total_Sales (Mean).
- As the rank for "Operating_Margin" increases, both Total Sales and p_Total_Sales (Mean) show fluctuations.
- There are significant increases in Total Sales at certain ranks, suggesting a relationship between the rank for "Operating_Margin" and the sales figures.
- The table provides detailed statistics for each rank of the variable "Operating_Margin":

Visualize Prediction By Price Per Unit

```
proc means Data=Nandinis.Testpred n min p10 p30 p40 p50 p60 p70 p80 p90 max maxdec=2;
var price_per_unit;
run;

proc rank Data=Nandinis.Testpred out=Nandinis.Testpred_percent groups=10;
var price_per_unit;
ranks rank;
run;

proc summary Data=Nandinis.Testpred_percent;
class rank;
var total_Sales p_total_sales;
output out=Nandinis.pred_ppu mean=;
quit;

proc sgplot Data=Nandinis.pred_ppu;
vbar rank /response =_freq_;
vline rank / response =Total_Sales y2axis stat=mean;
vline rank /response=p_Total_Sales y2axis stat=mean;
run;

Proc means Data=Nandinis.pred_ppu;
class rank;
var total_sales p_total_sales;
run;
```



Rank for Variable Price_per_Unit	N Obs	Variable	Label	N	Mean	Std Dev	Minimum	Maximum
0	1	Total_Sales	Total Sales	1	25257.87	.	25257.87	25257.87
		p_Total_Sales		1	16718.16	.	16718.16	16718.16
1	1	Total_Sales	Total Sales	1	30334.06	.	30334.06	30334.06
		p_Total_Sales		1	24625.61	.	24625.61	24625.61
2	1	Total_Sales	Total Sales	1	57259.88	.	57259.88	57259.88
		p_Total_Sales		1	59653.10	.	59653.10	59653.10
3	1	Total_Sales	Total Sales	1	78976.61	.	78976.61	78976.61
		p_Total_Sales		1	82262.02	.	82262.02	82262.02
4	1	Total_Sales	Total Sales	1	20085.55	.	20085.55	20085.55
		p_Total_Sales		1	21627.51	.	21627.51	21627.51
5	1	Total_Sales	Total Sales	1	81140.45	.	81140.45	81140.45
		p_Total_Sales		1	87220.42	.	87220.42	87220.42
6	1	Total_Sales	Total Sales	1	126174.08	.	126174.08	126174.08
		p_Total_Sales		1	130848.15	.	130848.15	130848.15
7	1	Total_Sales	Total Sales	1	150028.80	.	150028.80	150028.80
		p_Total_Sales		1	153037.68	.	153037.68	153037.68
8	1	Total_Sales	Total Sales	1	105320.28	.	105320.28	105320.28
		p_Total_Sales		1	112937.09	.	112937.09	112937.09
9	1	Total_Sales	Total Sales	1	271815.64	.	271815.64	271815.64
		p_Total_Sales		1	252912.99	.	252912.99	252912.99

- Bar Chart:
- FREQ (Sum): This blue bar represents the sum of frequencies for each rank of the variable "Operating_Margin."
- Total Sales: This blue line represents the total sales corresponding to each rank.
- p_Total_Sales (Mean): This red line represents the mean of total sales for each rank.
- As the rank for "Operating_Margin" increases, the sum of frequencies fluctuates, indicating variability in the distribution.
- Total sales also show variation across different ranks, but generally, higher ranks are associated with higher total sales.
- The mean total sales (p_Total_Sales) follow a similar trend to total sales, showing an increase with higher ranks.
- Higher Ranks: Generally associated with higher total sales and higher mean total sales.
- Variability: There's significant variability in the distribution of frequencies and total sales across different ranks.
- Standard Deviation: Higher ranks tend to have a higher standard deviation, indicating greater variability in total sales.

SUMMARY

- **Using Linear Regression Model 95.82% of variance in Total sales can be explained using independent variables**
- **Operating Profit, Units Sold, Price per unit and Operating margin are Top 4 important features for prediction**
- **No Clear Correlation between Operating Margin and Total Sales., suggesting that higher margins do not necessarily lead to higher revenue**
- **Strong positive correlation between Units sold and Total Sales clearly indicates that total revenue is directly influenced by sales volume**
- **High Priced Products tend to generate more revenue but mid ranged products showed inconstancy**
- **Apart from top 4 features following are important contributors in total sales**
    - **Sales Method- Instore**
    - **Retailer- West Gear,Amazon,Foot locker**
    - **Region- Midwest**
    - **Product- Woman’s Athletic Footwear**