Report on
Retail Sales Analysis for Revenue Enhancement

Submitted by

Mohammad Monjur-E-Elahi
Course: Data Science Project Using SAS [DSA14]
Program: Data Science and Application - Advanced Diploma [6060]
Metro College of Technology
Date: 1st August, 2021

1. Introduction

This report covers Retail store customers' transactional data that were recorded from May,2007 to April,2008, to analyze various different products on Different dates, with 8 variables about the data in one Data Set and to utilize some of the variables from the second Data Set with Electronic products with the category EC90, having 11 variables.

More information about the variables in both Data Sets, including no. of observations and the variables can be found in the table within Appendix: **A1** & **A2.**

2. Objectives

The following objectives were used in order to build the analysis:

- Perform Univariate Analysis and Bivariate Analysis to answer the business questions
- Perform Predictive Modeling and check whether this holds good or not
- Recommend measures to increase sales based on Data and Analysis

3. Business Questions

The following business questions were used in order to build the analysis:

1. Find the **Province** to focus to increase sales
2. Find the bestselling **Category** to increase sales
3. Find the **Source** of order with potential to increase sales
4. Find the **Top Customers** for promotion for sales volumes
5. Find the **Day of Week** to focus to increase sales

4. Methodology

The methodology followed is depicted below, followed by a description of the steps taken and the results arrived at.

Collection of Data	Secondary. Provided by Metro College of Technology
Definition of Data	29298 Observations, 7 columns
Exclusions	Kept Location from EC90 and Description, removed Price from transaction dataset
Inclusions	Location information from ec90, Sales included by calculation
Software Used	SAS 9.4
Statistical Techniques	Univariate, Bivariate, ANOVA

- Data cleansing/Validation
- Data Preparation
- Statistical Procedures
- Descriptive Data Analysis
- Using SAS 9.4 - All the steps were carried out using the SAS 9.4 software.

Data cleansing / Validation

The first step taken was to Remove the Duplicate records that were found and which had duplicate data across all the variables, as this seemed to be erroneous and could have an impact on results. In order to do this, the 'Proc Sort' procedure was used with 'NODUPRECS BY _ALL_', in order to drop those records which had duplicate values across all variables, only.

After looking at the contents of the data and the running the 'Proc Univariate' procedure, the second step was to Exclude the records that have a Quantity = 0 as it will not be used in

the analysis.

The final step was to count missing values for all categorical variables and remove unnecessary categories. This was carried out in a regular Data Step of creating a Data set with no missing values.

Data Preparation

There were two steps, using the 'Proc Sql' commands, one step of which was to Create the Sales Variable, using the Price and Quantity variables and another step was used to find the Customer_ID, from the first data set, in the EC90 Data Set and then to Include the City, Prov & Postal Code for each customer and then left join it in the first Data Set.

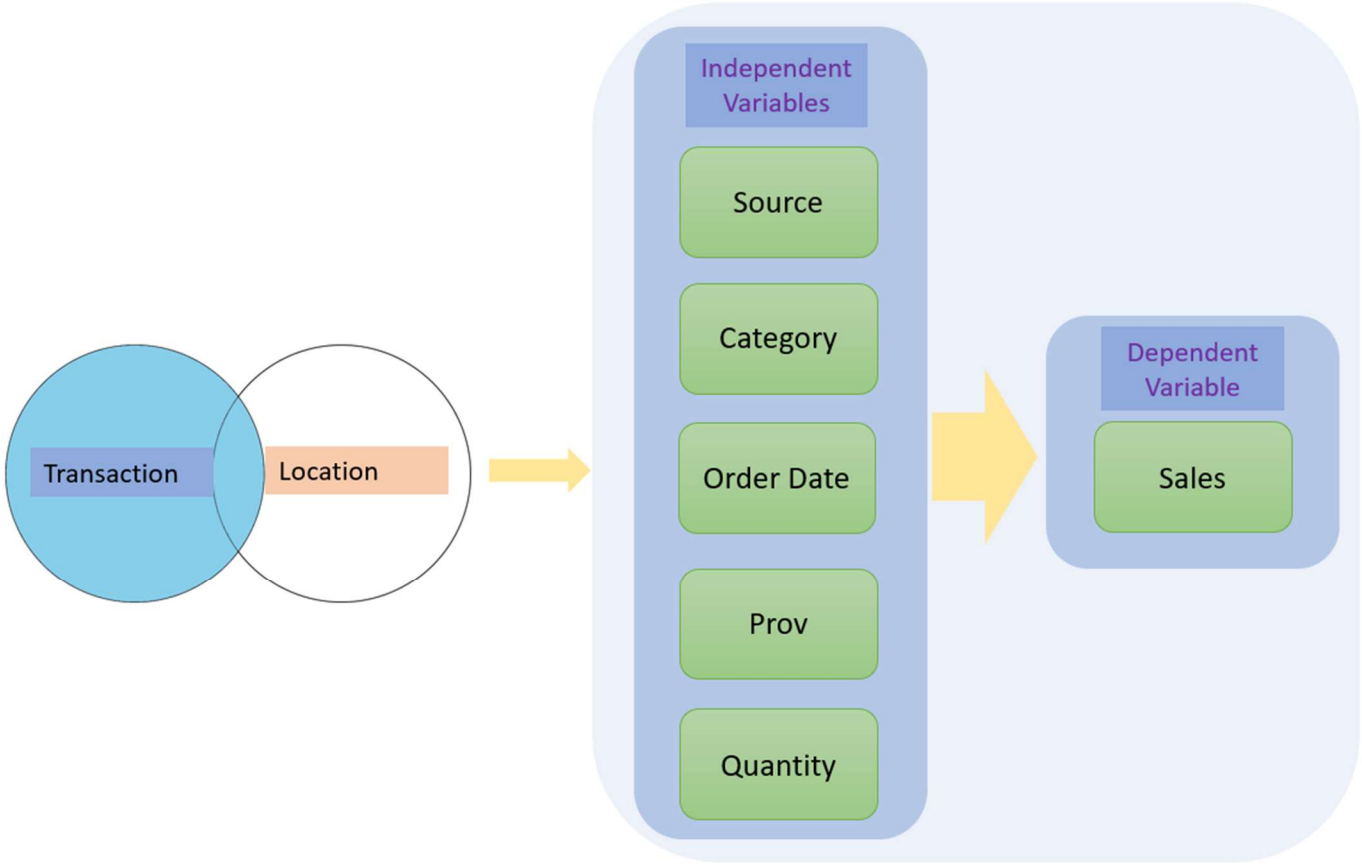
The Year, Quarter, Month, Day also had to be extracted from the Order_Date available in the first Data Set. These columns were created and then formatted into Letters, i.e., Jan, Feb, Mar, etc., using the IF – ELSE IF code logic.

Statistical Procedures

Use of the Statistical Procedures that best describe the variables that have the most effect on the target, i.e., Sales or Quantity sold, were as given below:

- Procedures – Import, Contents, Print, SQL, Sort, Univariate, Format, Freq, REG, Gchart, Sgplot, Standard, GLMSELECT.
- The count () and sum () functions were the most common functions used within theabove procedures to carry out the statistical analysis.
- Inferential Statistics – Test on the variables was carried out using ANOVA at asignificance of 0.05.
- Predictive Statistics – Linear Regression Model was built to try and predict the sales amount.

5. Conceptual Framework



The Conceptual Framework depicts how the overall analysis went on to reach the insight of the data.

The transactional dataset was left joined to the location dataset after cleaning the duplicate rows from both datasets. Also, the rows with missing values in any observation was removed since this constituted very low in percentage. The joint dataset was cleaned for the duplicated observations after the join as well.

Transaction Dataset	Location Dataset	Independent Variables	Dependent Variable
Customer ID	Order Number	Customer ID	Sales
Item Code	Customer Number	Source	
Source	City	Category	
Order Date	Prov	Order Date	
Item Description	Postal Code	Prov	
Category	Order First Time	Quantity	
Price	Source		
Quantity	Sales Amount		
We Calculated: **Sales = Quantity*Price**	Item Num		
	Item Description		
	Category Code		
	Quantity		

- Transaction had 34,288 rows
- Location had 1,284 rows
- Joined dataset 29,298 observations

Variables of Study –Definitions

The Variables used in the study can be summed up as below:

***Independent Variables*:**

Province – Location of Purchase/Purchaser Province, also location of Delivery might be the same

City – Same as above

Order Month – Month in which purchase was made. Generated from Order_Date

Order Year – Year in which purchase was made. Generated from Order_Date

Source – Includes all channels: Web, Regular, IVR and some Other categories

Dependent Variables (Calculated):

Total Sales Amount

Summary of Variables

The table below shows the summary of variables using the 'Proc Contents' procedure, after the creation of the 'Sales' Variable which was a result of the Price * Quantity, both Numeric Variables and Sales is also a Numeric Variable. 'Sales' has been used as the Dependent variable in the analysis.

- Note that the ‘Order_Date’ variable is in the ‘DATETIME’ format, which allowed the extraction of the year, quarter, month, week, day and day of the week using the year (), qtr (), month (), week (), day () and weekday () respectively.
- Also, note the reduction in the number of observations from 34,288 to 29,998 after removing the duplicate records

Contents of the joint Dataset

The CONTENTS Procedure

Data Set Name	MMEELAHI.INFERENTIAL	Observations	29298
Member Type	DATA	Variables	18
Engine	V9	Indexes	0
Created	07/28/2021 13:47:50	Observation Length	200
Last Modified	07/28/2021 13:47:50	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	YES
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	90
First Data Page	1
Max Obs per Page	327
Obs in First Data Page	313
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	C:\Users\ruzdomain\Desktop\SAS_Project\MMEELAHI\inferential.sas7bdat
Release Created	9.0401M6
Host Created	X64_10PRO
Owner Name	DINARTRIOSEDUCA\ruzdomain
File Size	6MB
File Size (bytes)	5963776

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
5	Category	Char	1	$1.	$1.
16	City	Char	30	$30.	$30.
1	Customer_ID	Num	8	BEST12.	BEST32.
13	Day	Num	8		
2	Item_Code	Char	8	$8.	$8.
4	Item_Description	Char	55	$55.	$55.
11	Month	Num	8		
8	Order_Date	Num	8	DATE9.	
18	Postal_Code	Char	6	$6.	$6.
7	Price	Num	8	DOLLAR10.2	
17	Prov	Char	2	$2.	$2.
6	Quantity	Num	8	BEST12.	BEST32.
10	Quarter	Num	8		
15	Sales	Num	8		
3	Source	Char	8	$8.	$8.
12	Week	Num	8		
14	Week_Day	Num	8		
9	Year	Num	8		

Sort Information	
Sortedby	Customer_ID Item_Code Source Item_Description Category Quantity Price Order_Date Year Quarter Month Week Day Week_Day Sales City Prov Postal_Code
Validated	YES
Character Set	ANSI
Sort Option	NODUPREC

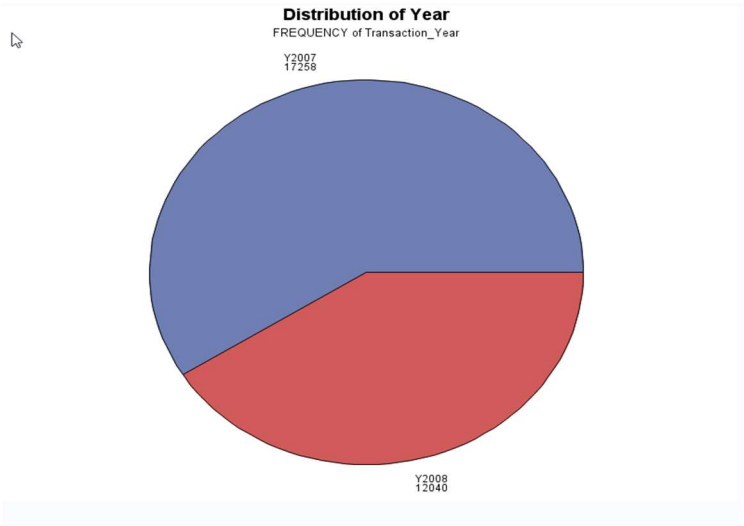
6. Descriptive Data Analysis

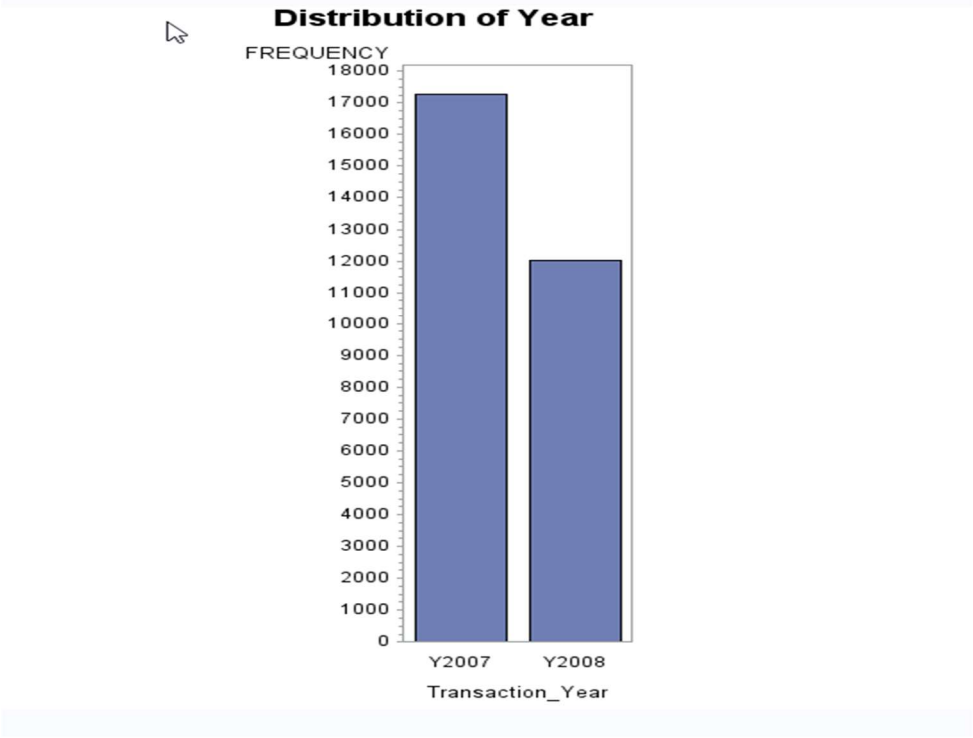
All the results of the descriptive analysis are described below.

Univariate Analysis:

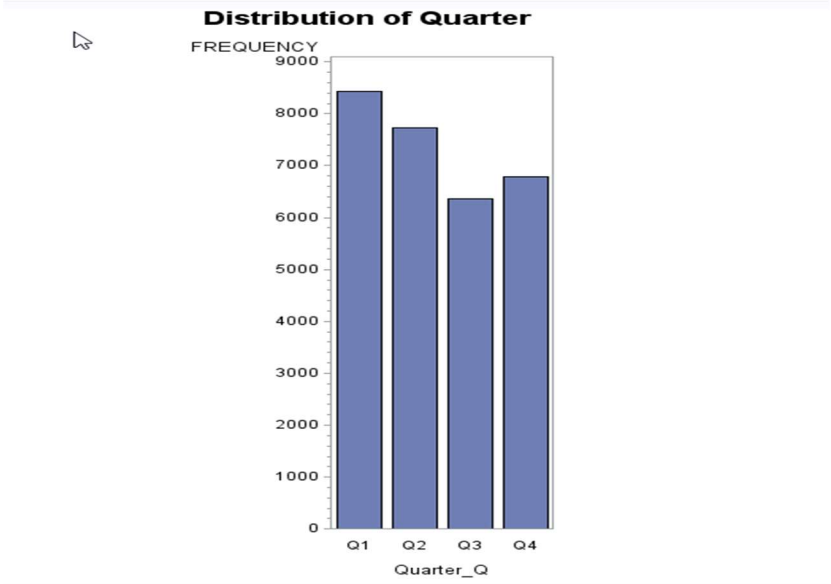
We performed descriptive analysis in graphical way for each of the columns of interest. For date we converted it to different categorical columns to have the analysis.

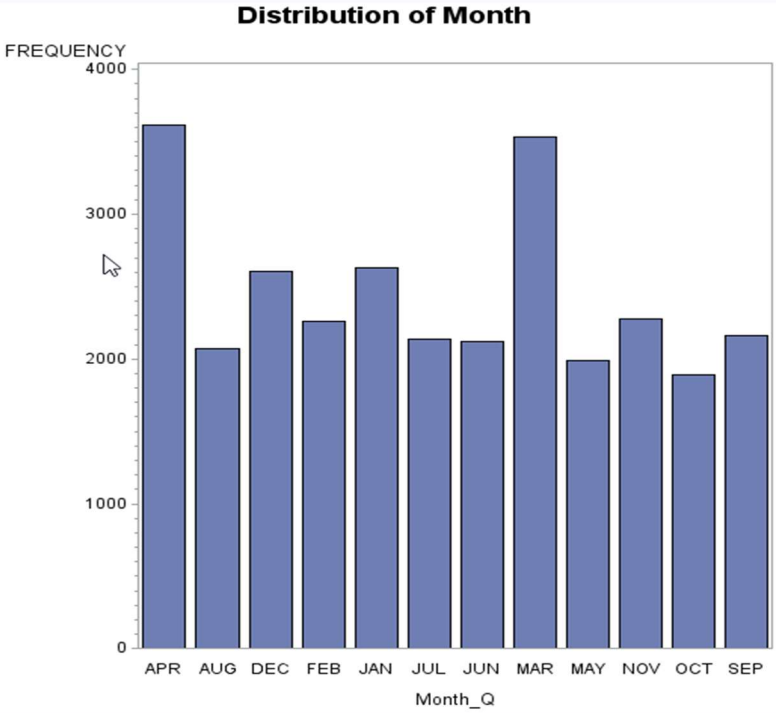
Order_Date:

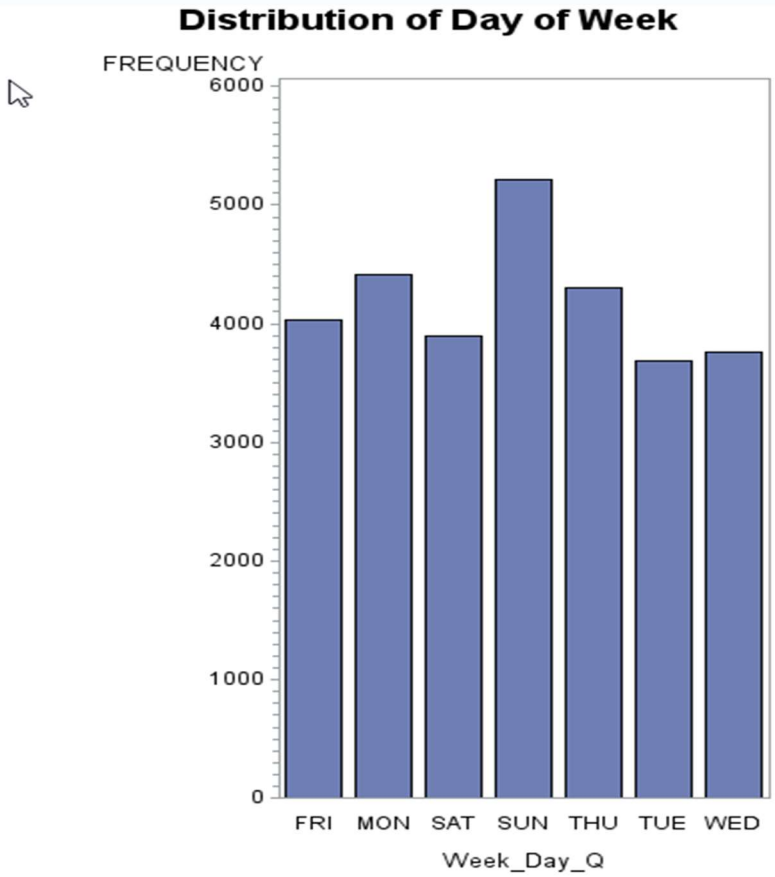


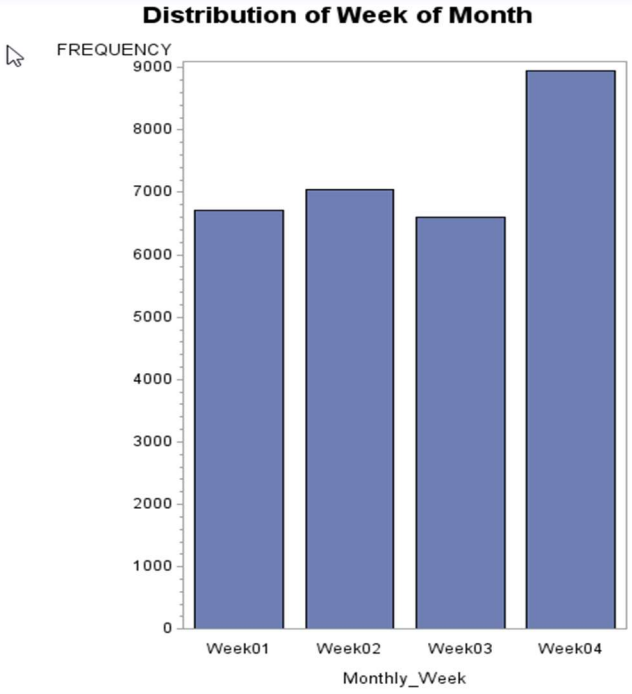


We find that the observation for the year 2007 is much higher compared to year 2008. But in fact, the data for 2007 covers only from May to December while the data for 2008 consists of for the duration of January to April.

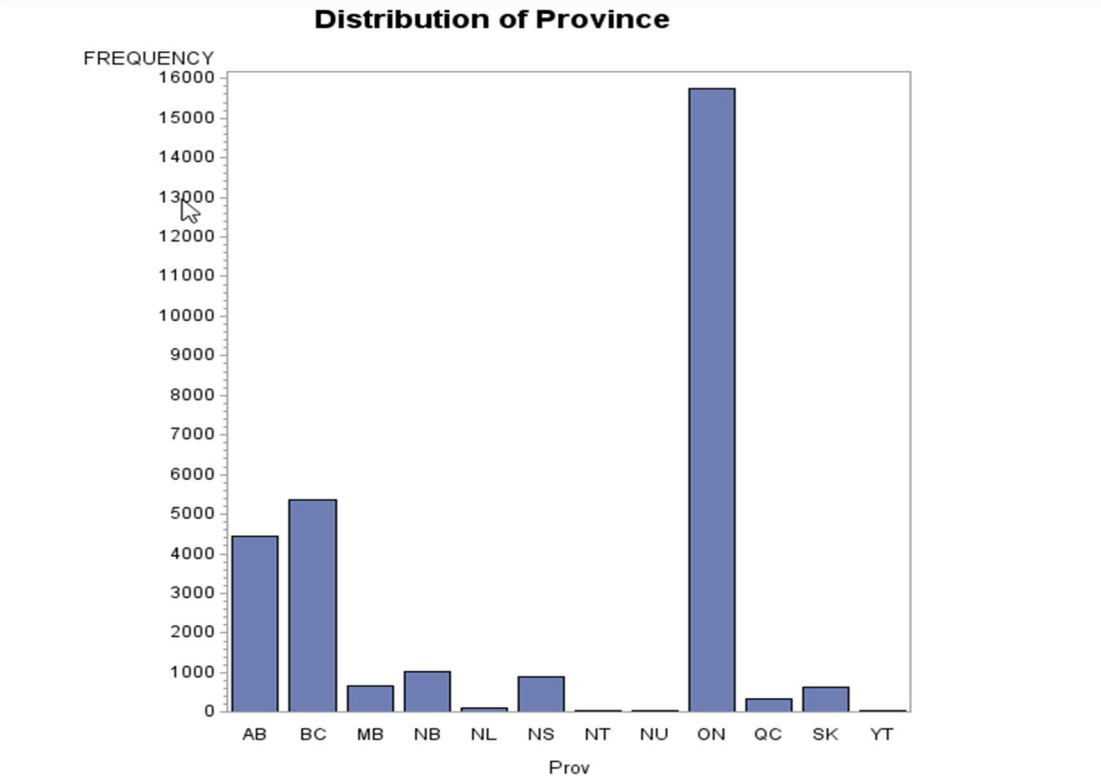

Distribution of Quarter
FREQUENCY
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Q1
Q2
Q3
Q4
Quarter_Q


Distribution of Month
FREQUENCY
4000
3000
2000
1000
0
APR
AUG
DEC
FEB
JAN
JUL
JUN
MAR
MAY
NOV
OCT
SEP
Month_Q

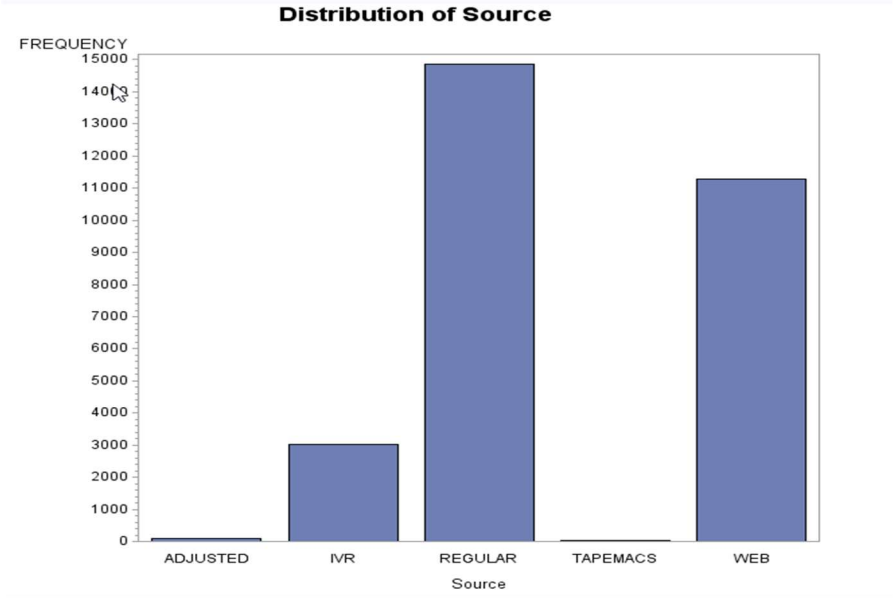

Distribution of Day of Week
FREQUENCY
6000
5000
4000
3000
2000
1000
0
FRI
MON
SAT
SUN
THU
TUE
WED
Week_Day_Q


Distribution of Week of Month
FREQUENCY
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Week01
Week02
Week03
Week04
Monthly_Week

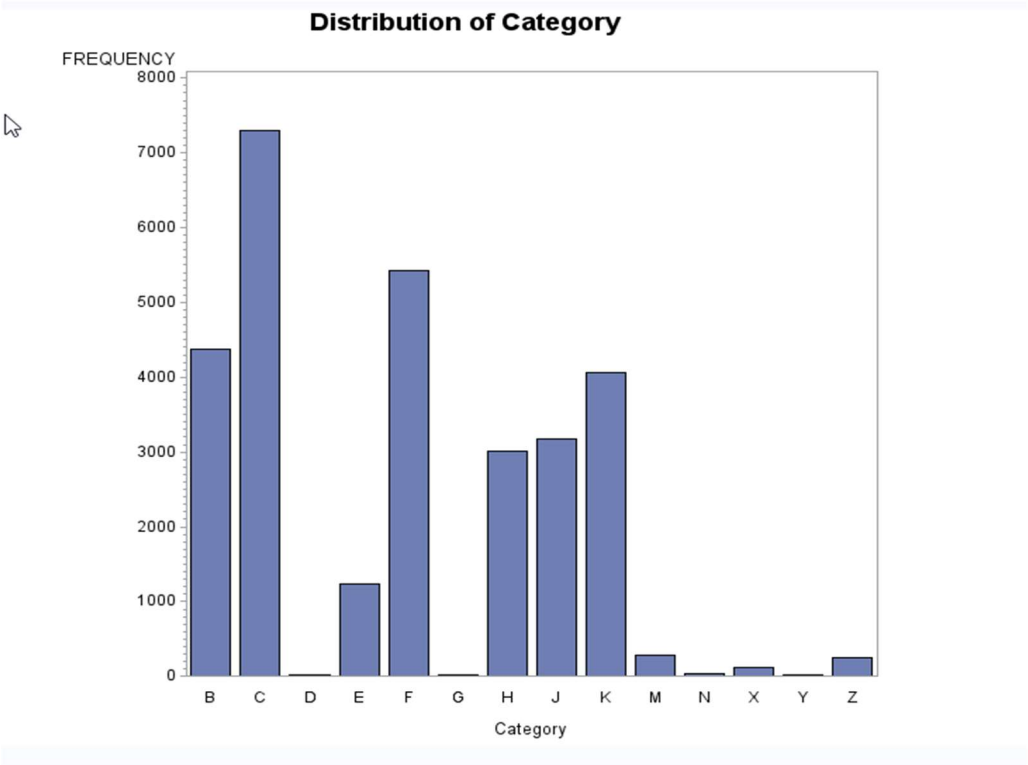
Province:



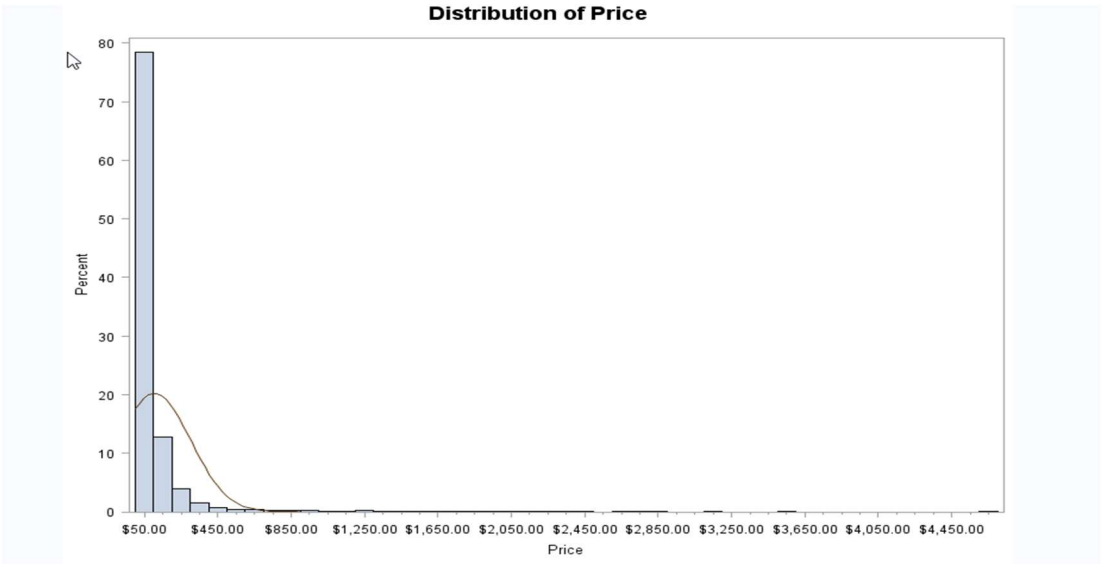
Source:



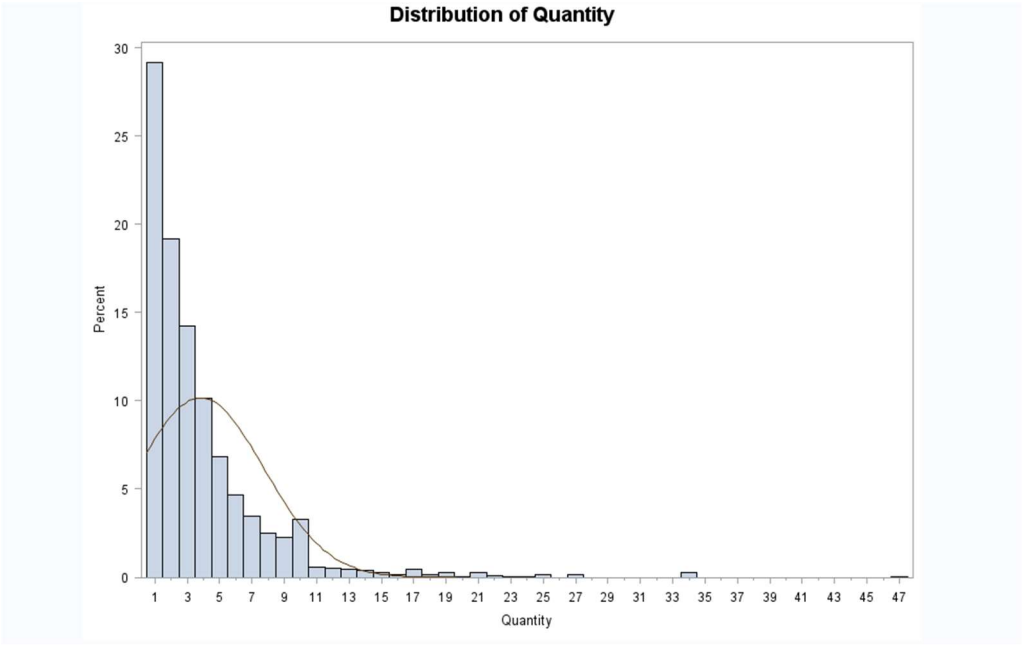
Category:



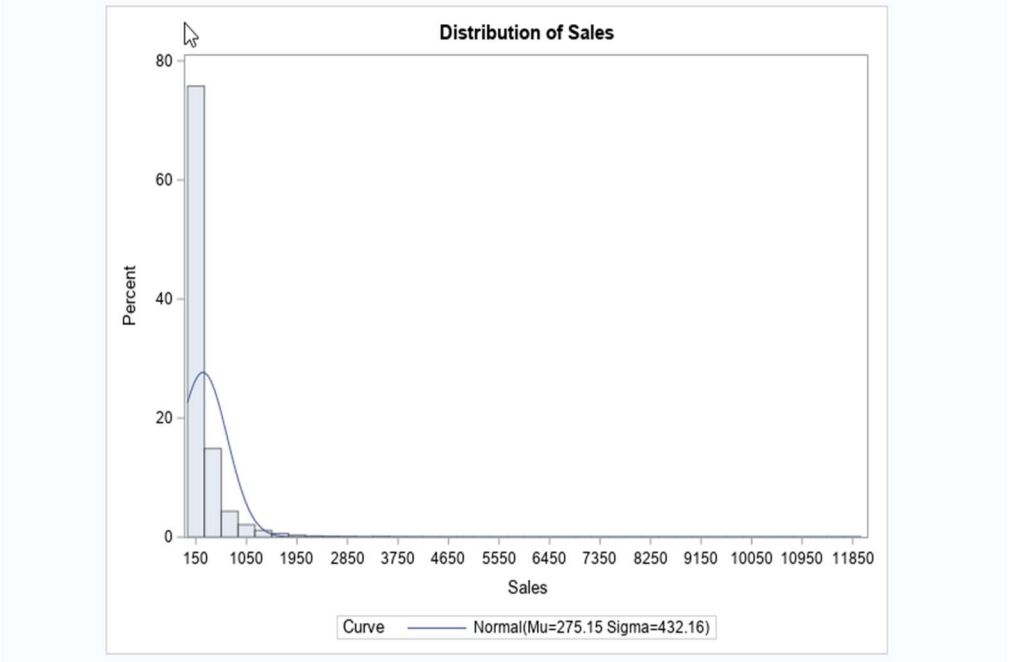
Price:



Quantity:



Sales:



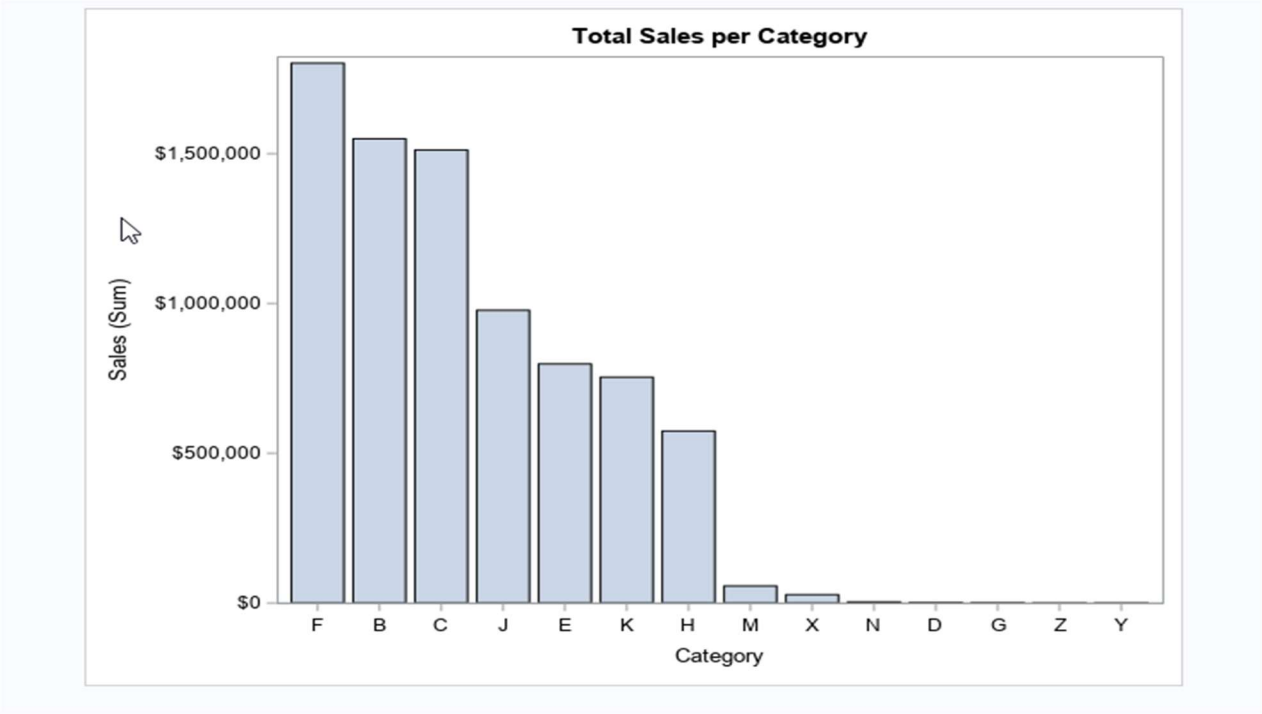
Summary of Univariate Analysis:

- Data is not balanced between the two years
- Data is there for last 8 months for 2007 and first four months for 2008.
- Monday and Sunday constitute the top two days for transactions
- Last week of the month experience higher transaction
- Ontario is way ahead in transaction
- Regular and Web are top two sources of transaction
- C, F, B, K are top four categories for transaction
- Price, Quantity and Sales are not normally distributed and skewed towards the right

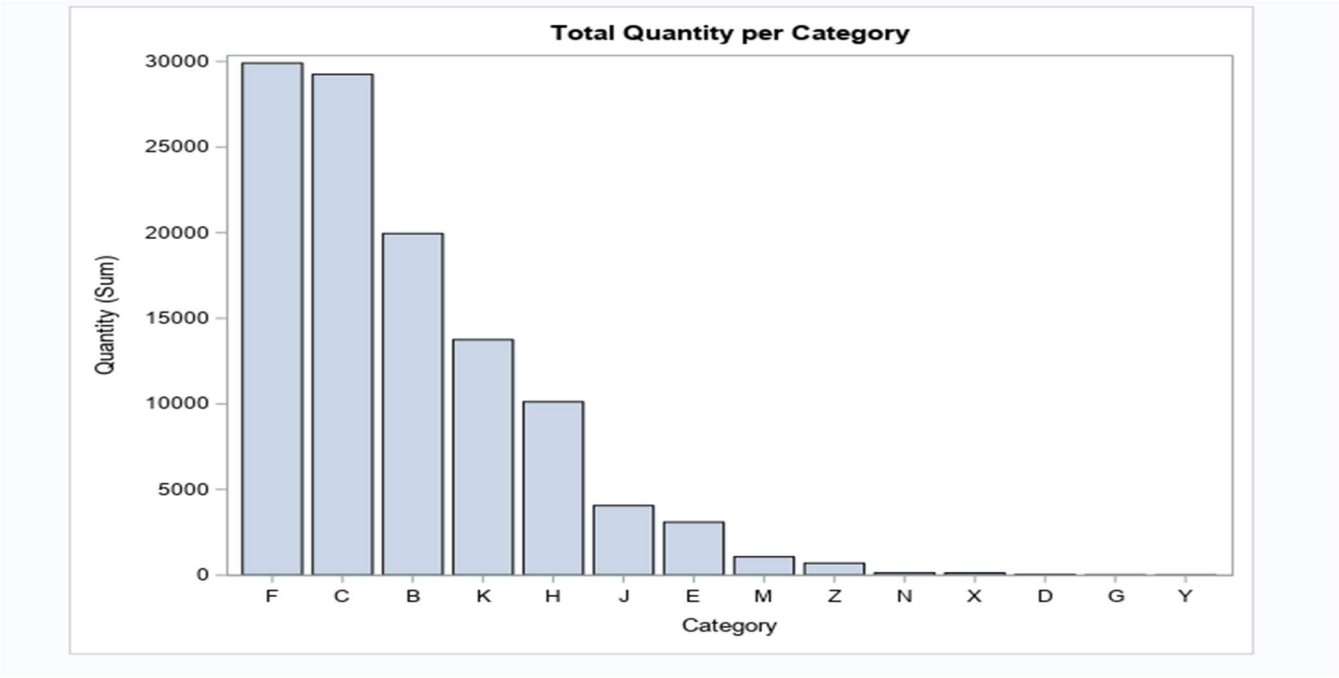
Bivariate Analysis:

We performed the analysis between all our features of interest with the target Sales column. Also, we performed bivariate analysis between our features of interest with the Quantity feature.

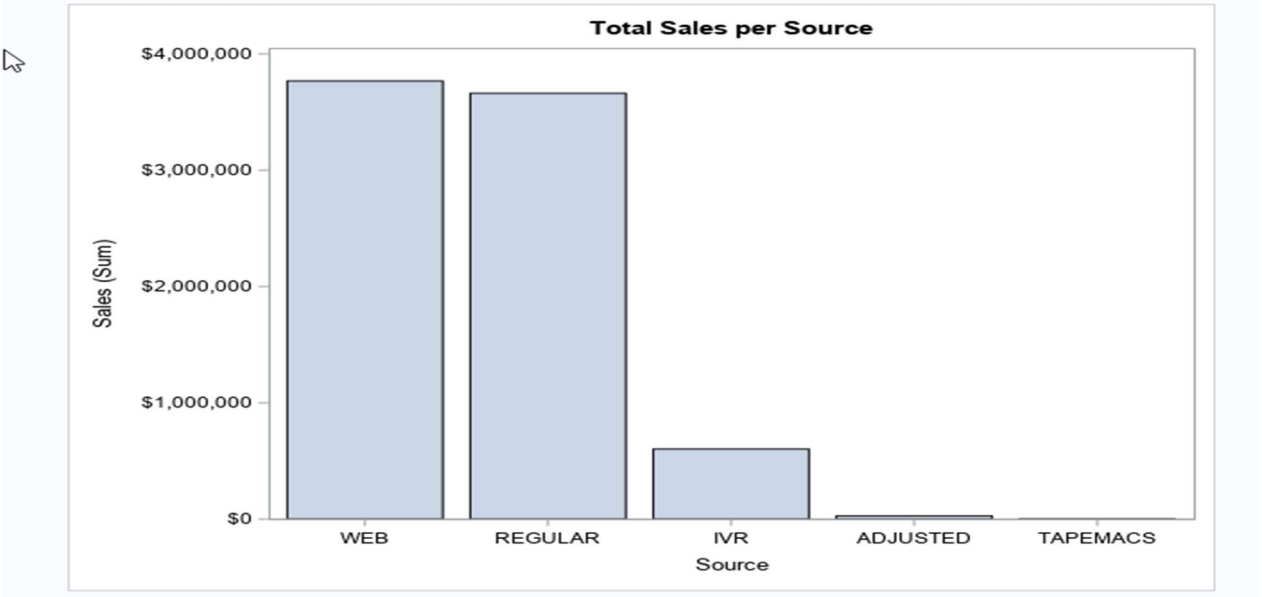
Sales versus Category:



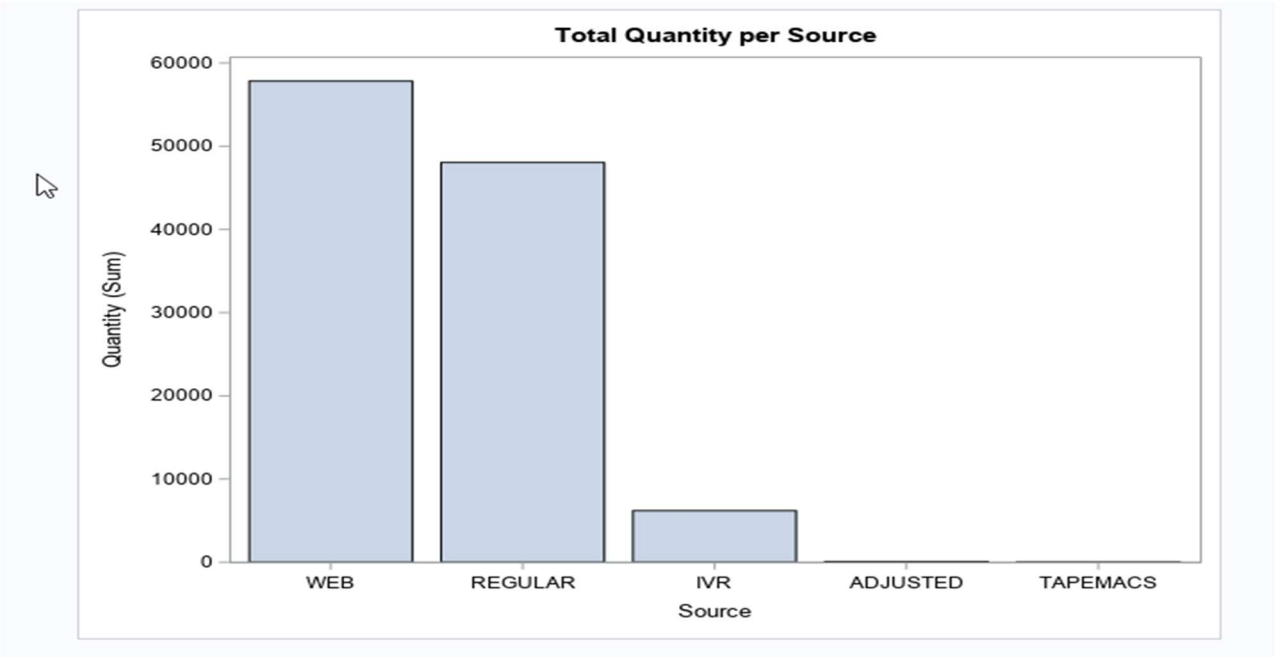
Quantity versus Category:



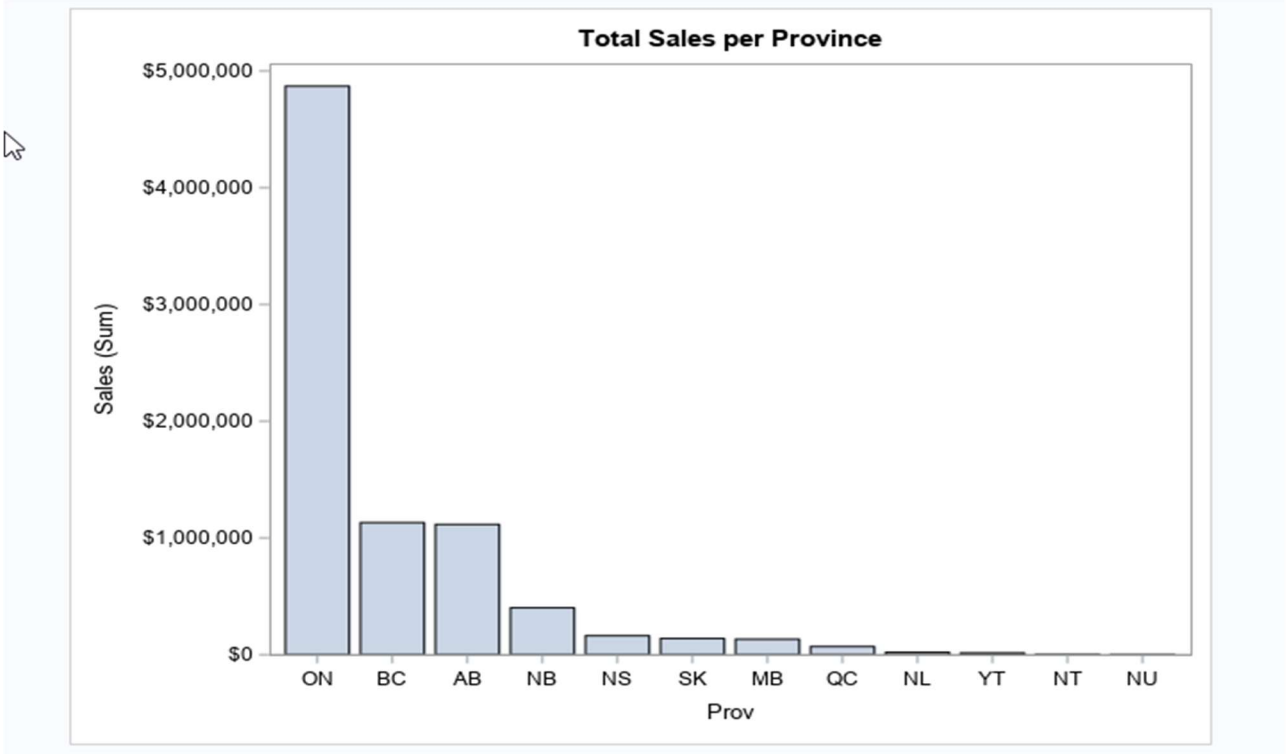
Sales versus Source:



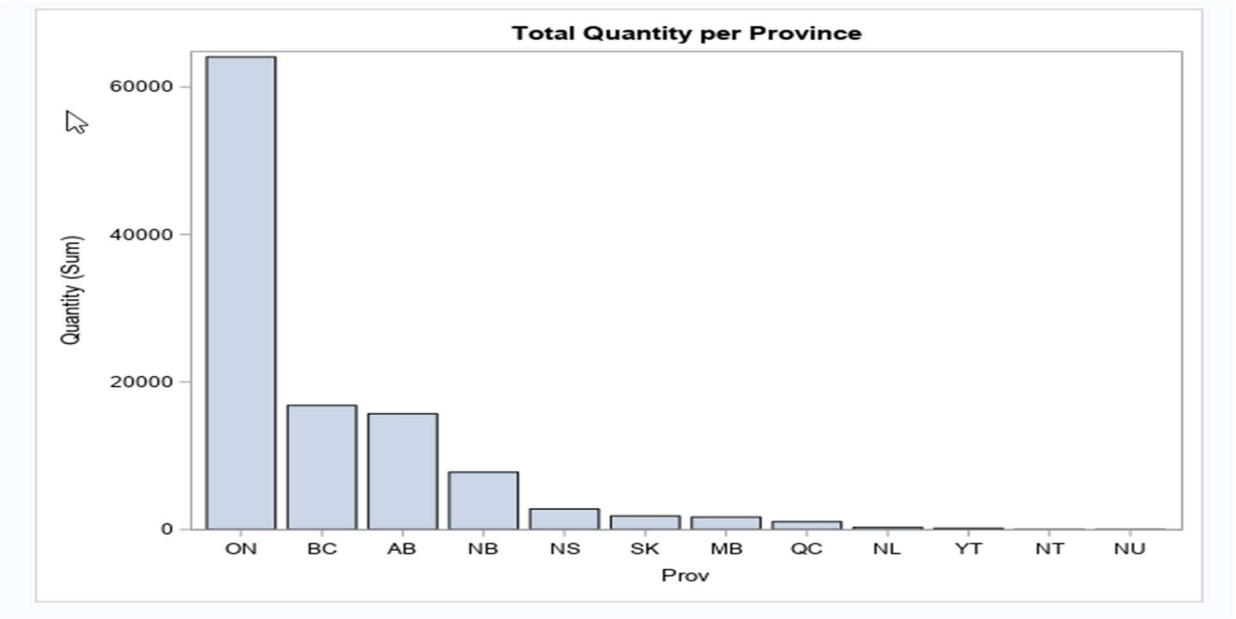
Quantity versus Source:



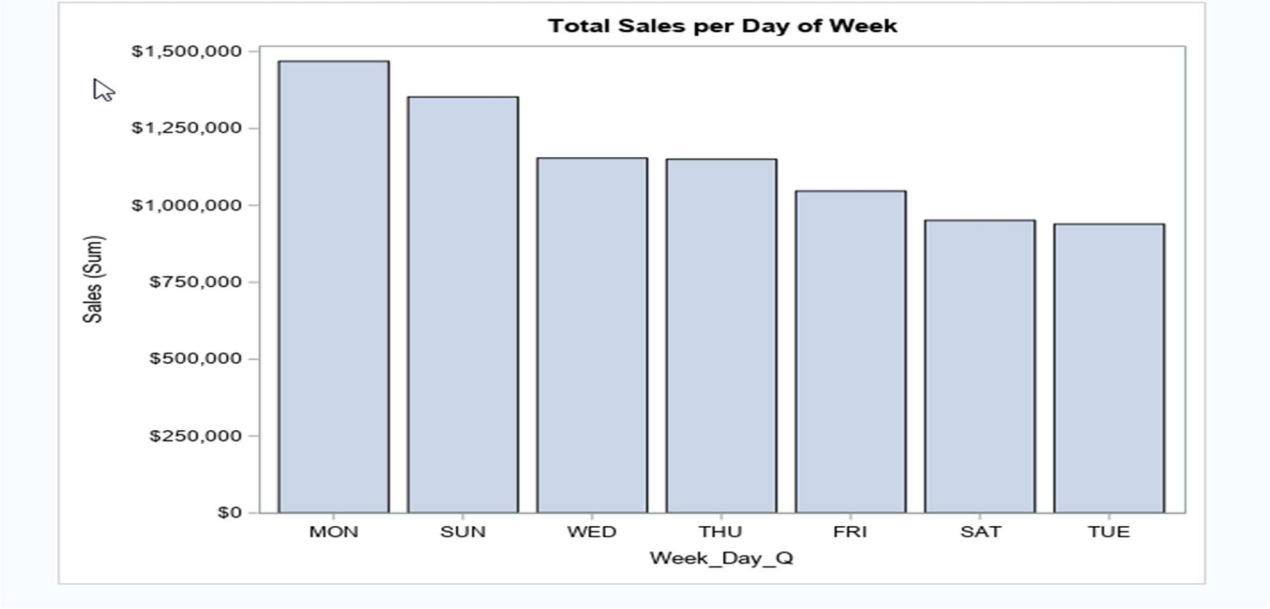
Sales versus Province:



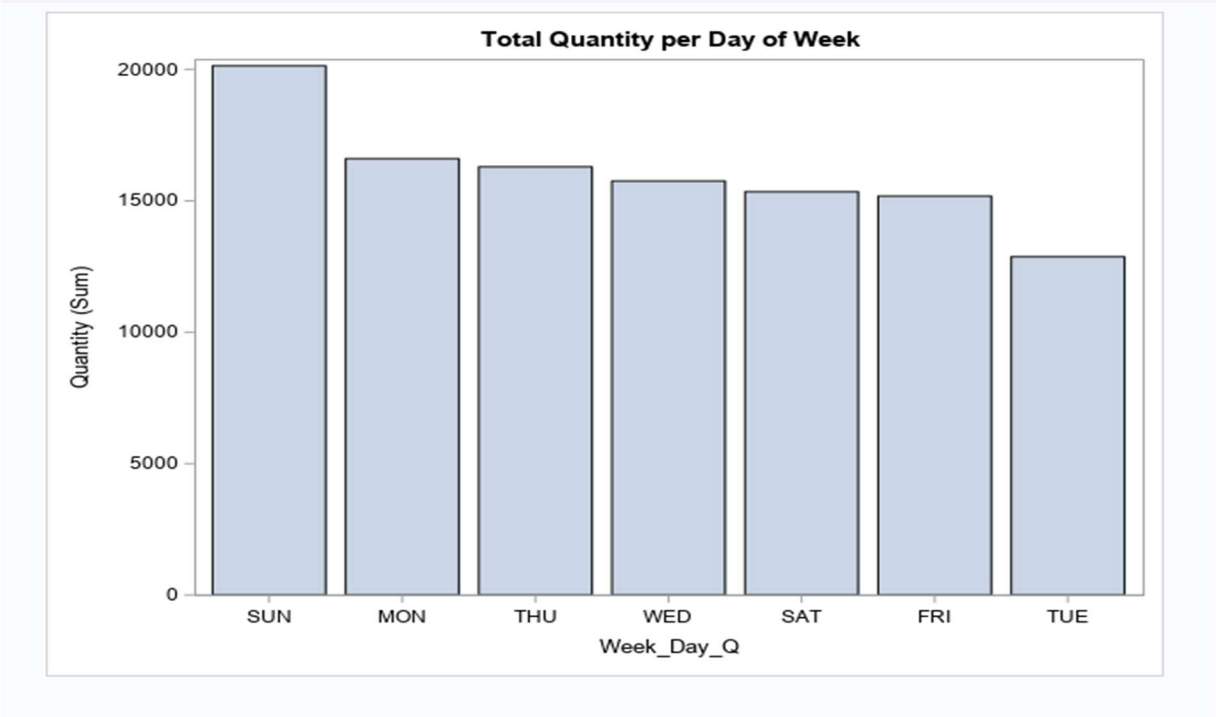
Quantity versus Province:



Sales versus Day of Week:



Quantity versus Day of Week:



Summary of Bivariate Analysis:

- The Analysis shows that ON has the highest value of Sales amount and Quantity
- The Analysis shows that the WEB was the channel with the highest Sales
- Category F, B, C are the most sold categories
- Monday and Sunday are the days when the sales are found to be high

7. Hypothesis Testing

ANOVA Analysis

We performed the ANOVA analysis between the numeric target or dependent column Sales and our independent variables of interest.

The Analysis shows that the Probability is below the significance level of 0.05, Therefore, we can reject the Null Hypothesis and conclude that there is a possibility of a relationship Between the independent variables of interest and Sales. Refer to the figures below for this analysis:

Target (Numeric)	Feature(Categorical)	P-Value	Statistical Significance
Sales	Category	<.0001	Significant
Sales	Prov	<.0001	Significant
Sales	Source	<.0001	Significant
Sales	Quarter	<.0001	Significant
Sales	Month	<.0001	Significant
Sales	Day Of Week	<.0001	Significant
Sales	Week of Month	0.0422	Significant

Spearman Correlation Analysis

We performed the Spearman Correlation Analysis (as the data of the concerned columns were not normally distributed) between the numeric target or dependent column Sales and our independent variables of interest which are Price and Quantity.

The Analysis shows that the Probability is below the significance level of 0.05, Therefore, we can reject the Null Hypothesis and conclude that there is a possibility of a relationship Between the independent variables of interest and Sales. Refer to the figures below for this analysis:

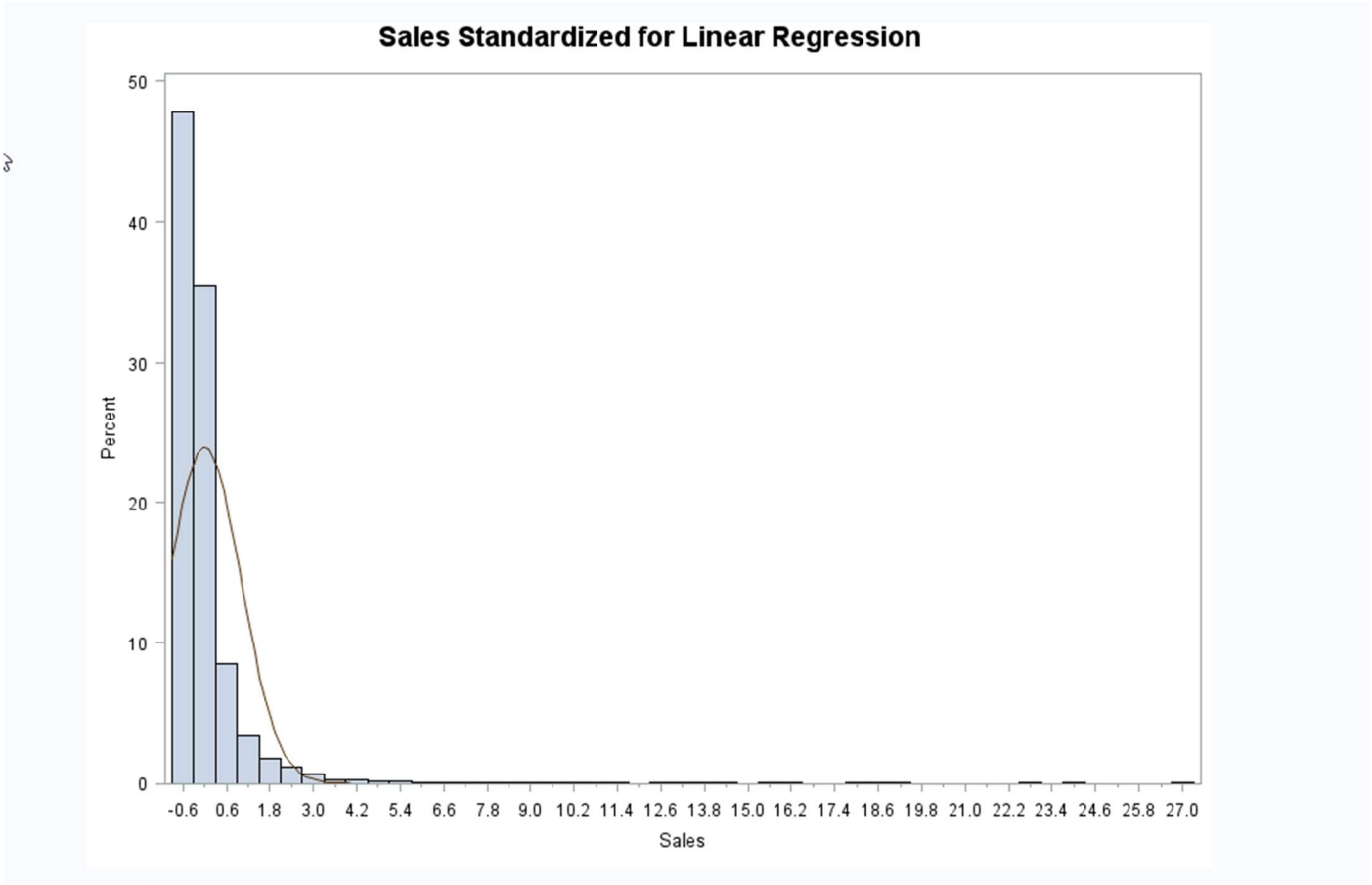
Target (Numeric)	Feature (Numeric)	P-Value	Statistical Significance
Sales	Quantity	<.0001	Significant
Sales	Price	<.0001	Significant

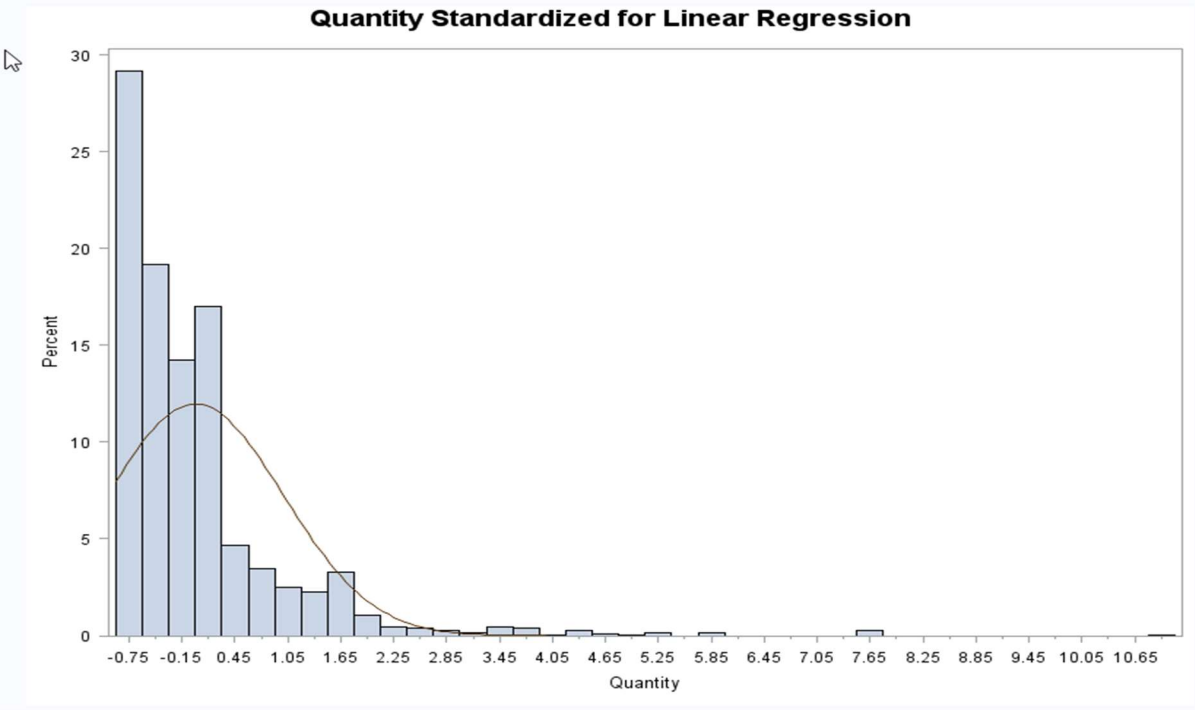
7. Predictive Modeling

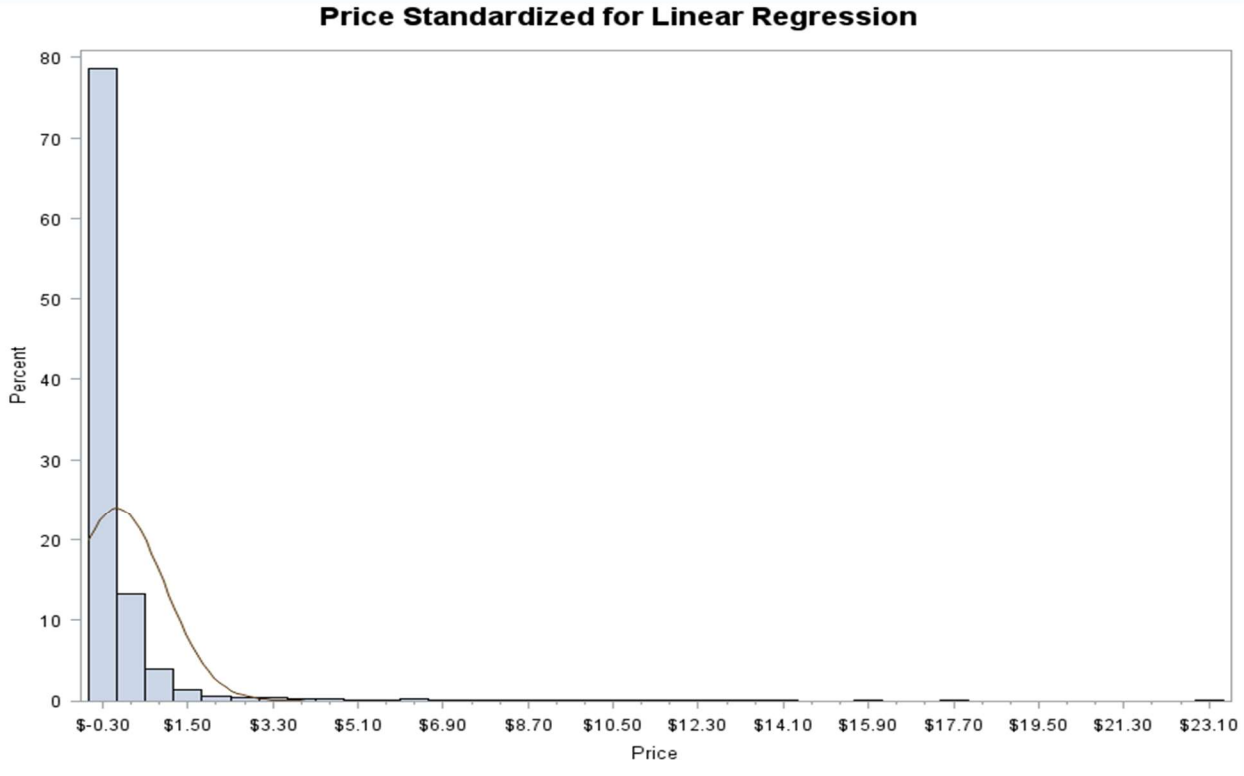
We performed predictive modeling using the linear regression model to predict the Sales value. Initially we left out the Price column but gave us very poor model R2 and Adjusted R2 score. Then we checked the correlation between the Price and Sales and being confirmed about the statistically significant relationship between them we included the Price column in to the model and we got far better model compared to the model that we built at the first time.

First, we had to standardize the numeric columns that were found to be not with normal distribution.

```
PROC UNIVARIATE DATA = MMEELAHI.INFERENTIAL_MODEL NORMAL;
var Sales;
histogram/normal;
title 'Sales Standardized for Linear Regression';
RUN;
```




Quantity Standardized for Linear Regression
Percent
30
25
20
15
10
5
0
-0.75 -0.15 0.45 1.05 1.65 2.25 2.85 3.45 4.05 4.65 5.25 5.85 6.45 7.05 7.65 8.25 8.85 9.45 10.05 10.65
Quantity


Price Standardized for Linear Regression
Percent
80
70
60
50
40
30
20
10
0
$-0.30 $1.50 $3.30 $5.10 $6.90 $8.70 $10.50 $12.30 $14.10 $15.90 $17.70 $19.50 $21.30 $23.10
Price

```
TITLE "Linear Regression Model";
ODS GRAPHICS ON;
PROC GLMSELECT DATA = MMEELAHI.INFERENTIAL_MODEL PLOTS = ALL;
    CLASS Prov Source Category;
    MODEL Sales = Quantity Price Prov Source Category / DETAILS = ALL STATS = ALL;
ODS GRAPHICS OFF;
RUN;
```

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 3).

Effects:	Intercept Quantity Price Category

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	15	18079	1205.23407	3145.85
Error	29282	11218	0.38312	
Corrected Total	29297	29297		

Root MSE	0.61897
Dependent Mean	-8.156E-14
R-Square	0.6171
Adj R-Sq	0.6169
AIC	1207.20787
AICC	1207.22877
BIC	-28091
C(p)	109.37522
PRESS	11252
SBC	-27960
ASE	0.38291

Adjusted R-Sq of 0.6169 suggests that the model is a moderate one.

8. Summary OF Findings

- Product of Category **F, B and C** are most sold
- **Web** is the largest Source for the Sales
- **Ontario** is by far the best revenue earning
- **Monday and Sunday** feature bigger Sales
- **Top 10 Customers** have been identified

9. Recommendations

Strengthen our promotion and other Sales and marketing strategies on the below aspects to increase revenue.

- ❑ **Province: Ontario**
- ❑ **Category: F, B & C**
- ❑ **Source: Web**
- ❑ **Days of Week: Mondays & Sundays**
- ❑ **Customers: As Listed in Top 10**

APPENDIX

A1. DATA SETS

Contents of transactionhistoryforcurrentcustomers.csv

The CONTENTS Procedure

Data Set Name	MMEELAHI.TRANSACTIONI	Observations	34288
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	07/27/2021 15:26:29	Observation Length	112
Last Modified	07/27/2021 15:26:29	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	59
First Data Page	1
Max Obs per Page	584
Obs in First Data Page	567
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	C:\Users\ruzdomain\Desktop\SAS_Project\MMEELAHI\transactioni.sas7bdat
Release Created	9.0401M6
Host Created	X64_10PRO
Owner Name	DINARTRIOSEDUCA\ruzdomain
File Size	4MB
File Size (bytes)	3932160

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
6	Category	Char	1	$1.	$1.
1	Customer_ID	Num	8	BEST12.	BEST32.
2	Item_Code	Char	8	$8.	$8.
5	Item_Description	Char	55	$55.	$55.
4	Order_Date	Num	8	DATETIME.	ANYDTDTM40.
8	Quantity	Num	8	BEST12.	BEST32.
3	Source	Char	8	$8.	$8.
7	price	Char	10	$10.	$10.

Contents of ec90_data.csv

The CONTENTS Procedure

Data Set Name	MMEELAHI.LOCATION	Observations	1427
Member Type	DATA	Variables	12
Engine	V9	Indexes	0
Created	07/27/2021 15:26:30	Observation Length	144
Last Modified	07/27/2021 15:26:30	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	4
First Data Page	1
Max Obs per Page	454
Obs in First Data Page	438
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	C:\Users\ruzdomain\Desktop\SAS_Project\MMEELAHI\location.sas7bdat
Release Created	9.0401M6
Host Created	X64_10PRO
Owner Name	DINARTRIOSEDUCA\ruzdomain
File Size	320KB
File Size (bytes)	327680

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
11	Category_code	Char	4	$4.	$4.
3	City	Char	30	$30.	$30.
2	Customer_Number	Num	8	BEST12.	BEST32.
10	Item_Description	Char	47	$47.	$47.
9	Item_Num	Num	8	BEST12.	BEST32.
6	Order_First_Time	Char	1	$1.	$1.
1	Order_Number	Char	8	$8.	$8.
5	Postal_Code	Char	6	$6.	$6.
4	Prov	Char	2	$2.	$2.
12	Quantity	Num	8	BEST12.	BEST32.
8	Sales_amount	Char	10	$10.	$10.
7	Source	Char	7	$7.	$7.

A2. Info on 2 Data Sets

Info: **Data Set 1** **Data Set 2**

Information	Dataset1	Dataset2
No. Of Records	34,288	1,297
Type of Products	Variety of different products	Electronics category EC90 only: Acer Aspire 16" Multimedia Notebook Computer
Demographics:Location Info	No Location information	City, Province and Postal Code
Different Order Details	Order Date	Order_Number, Order_First_Time (Y/N)
Sales information	Price & Quantity	Sales_Amount & Quantity
Common	Customer_ID, Item_Code, Source, Item_Description, Category, Quantity	Customer_Number, Source, Item_Num, Item_Description, Category_code, Quantity

A3. SAS SCRIPTS

```
libname MMEELAHI "C:\Users\ruzdomain\Desktop\SAS_Project\MMEELAHI";

/*STEP 1*/
*Import data to SAS;

PROC IMPORT OUT= MMEELAHI.TRANSACTIONI
DATAFILE="C:\Users\ruzdomain\Desktop\SAS_Project\MohammadMonjurEElahi_
SAS_Project\SAS_Project_Data_And_Script\transactionhistoryforcurrentcu
stomers.csv"
DBMS=CSV REPLACE;
GETNAMES=YES;
GUESSINGROWS=MAX;
DATAROW=2;
RUN;

PROC IMPORT OUT= MMEELAHI.LOCATION
DATAFILE="C:\Users\ruzdomain\Desktop\SAS_Project\MohammadMonjurEElahi_
SAS_Project\SAS_Project_Data_And_Script\ec90_data.csv"
DBMS=CSV REPLACE;
GETNAMES=YES;
GUESSINGROWS=MAX;
DATAROW=2;
RUN;


TITLE 'Contents of transactionhistoryforcurrentcustomers.csv';
PROC CONTENTS DATA=MMEELAHI.TRANSACTIONI;
RUN;


TITLE 'Contents of ec90_data.csv';
PROC CONTENTS DATA=MMEELAHI.LOCATION;
RUN;
```

```
*TRANSACTIONI - TRANSACTIONAL DATA SET WITH VARIOUS DIFFERENT
PRODUCTS;


PROC PRINT DATA = MMEELAHI.TRANSACTIONI (OBS = 30);
TITLE "TRANSACTIONI";
RUN;

*LOCATION - DATA WITH VARIUOUS RECORD ESPECIALLY LOCATION COLUMNS
LIKE Province, City, Postal Codes etc.;

PROC PRINT DATA = MMEELAHI.LOCATION (OBS = 30);
TITLE "LOCATION";
RUN;



*UNDERSTAND YOUR DATA AND ITS PROPERTIES;
TITLE "Contents TRANSACTIONI";
PROC CONTENTS DATA = MMEELAHI.TRANSACTIONI;
RUN;




TITLE "Contents LOCATION";
PROC CONTENTS DATA = MMEELAHI.LOCATION;
RUN;

* Converting the price column from categorical to numerical one;

DATA MMEELAHI.TRANSACTIONS;
SET MMEELAHI.TRANSACTIONI(RENAME = (price=old_price));
Price = input(old_price, COMMA10.2);
drop old_price;
format price DOLLAR10.2;
run;
```

```
* Convert the date to Year, Month, Quarter, Week, Day;

DATA MMEELAHI.TRANSACTIOND;
SET MMEELAHI.TRANSACTIONS(RENAME = (Order_Date=old_Order_Date));
Order_Date = datepart(old_Order_Date);
DROP old_Order_Date;
format Order_Date date9.;
RUN;



PROC PRINT DATA = MMEELAHI.TRANSACTIOND (OBS = 30);
TITLE "TRANSACTIOND";
RUN;



DATA MMEELAHI.TRANSACTION;
SET MMEELAHI.TRANSACTIOND;

Year = year(Order_Date);
Quarter = qtr(Order_Date);
Month = month(Order_Date);
Week = week(Order_Date);
Day = day(Order_Date);
Week_Day=weekday(Order_Date);

RUN;



PROC PRINT DATA = MMEELAHI.TRANSACTION (OBS = 10);
TITLE "TRANSACTION";
RUN;



TITLE "Contents TRANSACTION";
PROC CONTENTS DATA = MMEELAHI.TRANSACTION;
RUN;
```

```
*DUPLICATES;
*COUNT COLS;
TITLE "Count of Distinct Customer IDs in TRANSACTION";
PROC SQL;
SELECT COUNT(Customer_ID)AS TOTAL_COUNT, COUNT(DISTINCT
Customer_ID) AS
UNIQUE_COUNT
FROM MMEELAHI.TRANSACTION
;
QUIT;
*REMOVE DUPLICATE OBSERVATIONS and check count again;
PROC SORT DATA = MMEELAHI.TRANSACTION OUT = MMEELAHI.TRANSACTION_S
dupout=MMEELAHI.TRANSACTION_aDup NODUPRECS;
BY _ALL_;
RUN;


PROC SORT DATA = MMEELAHI.LOCATION OUT = MMEELAHI.LOCATION_S
dupout=MMEELAHI.LOCATION_aDup NODUPRECS;
BY _ALL_;
RUN;


*COUNT AGAIN;
TITLE 'Count of Customer IDs TRANSACTION_S';
PROC SQL;
SELECT COUNT(Customer_ID)AS TOTAL_COUNT, COUNT(DISTINCT
Customer_ID) AS
UNIQUE_COUNT
FROM MMEELAHI.TRANSACTION_S
;
QUIT;



TITLE 'Count of Customer Number LOCATION';
PROC SQL;
SELECT COUNT(Customer_Number)AS TOTAL_COUNT, COUNT(DISTINCT
Customer_Number) AS
UNIQUE_COUNT
FROM MMEELAHI.LOCATION_S
;
QUIT;
```

```
PROC format;
value $missfmt ' '='Missing' other='Not Missing';
value  missfmt  . ='Missing' other='Not Missing';
run;

PROC freq DATA=MMEELAHI.TRANSACTION_S;
format _CHAR_ $missfmt.; /* apply format for the duration of this
PROC */
tables _CHAR_ / missing missprint nocum nopercent;
format _NUMERIC_ missfmt.;
tables _NUMERIC_ / missing missprint nocum nopercent;
run;
```

```
PROC freq DATA=MMEELAHI.LOCATION_S;
format _CHAR_ $missfmt.; /* apply format for the duration of this
PROC */
tables _CHAR_ / missing missprint nocum nopercent;
format _NUMERIC_ missfmt.;
tables _NUMERIC_ / missing missprint nocum nopercent;
run;


*checking records for quantity = 0;

title 'checking records for quantity = 0';
PROC SQL;
SELECT * FROM MMEELAHI.TRANSACTION_S
WHERE Quantity = 0;
quit;


*checking records for price = .;
title 'Missing Price Record count';
PROC SQL;
SELECT count(*) FROM MMEELAHI.TRANSACTION_S
WHERE Price = .;
quit;
PROC SQL OUTOBS=10;
SELECT * FROM MMEELAHI.TRANSACTION_S
WHERE Price = .;
quit;
```

```
DATA MMEELAHI.TRANSACTION_T;
SET MMEELAHI.TRANSACTION_S;
IF quantity = 0 then delete;
run;

proc print data=MMEELAHI.TRANSACTION_T (OBS = 30);
title 'Omitting a Zero Quantity Observation';
run;



title 'checking records for quantity = 0';
PROC SQL;
SELECT * FROM MMEELAHI.TRANSACTION_T
WHERE Quantity = 0;
quit;


PROC freq DATA=MMEELAHI.TRANSACTION_T;
format _CHAR_ $missfmt.; /* apply format for the duration of this
PROC */
tables _CHAR_ / missing missprint nocum nopercent;
format _NUMERIC_ missfmt.;
tables _NUMERIC_ / missing missprint nocum nopercent;
run;



DATA MMEELAHI.TRANSACTION_U;
SET MMEELAHI.TRANSACTION_T;
IF Item_Description = " " then delete;
IF Category = " " then delete;
run;

proc print data=MMEELAHI.TRANSACTION_U (OBS = 30);
title 'Omitting Missing values';
run;



PROC freq DATA=MMEELAHI.TRANSACTION_U;
format _CHAR_ $missfmt.; /* apply format for the duration of this
PROC */
tables _CHAR_ / missing missprint nocum nopercent;
format _NUMERIC_ missfmt.;
tables _NUMERIC_ / missing missprint nocum nopercent;
run;
```

```
DATA MMEELAHI.TRANSACTION_V;
SET MMEELAHI.TRANSACTION_U;
Sales=Quantity*Price;
run;


proc print data=MMEELAHI.TRANSACTION_V (OBS = 50);
title 'Transaction Dataset with calculated Column Sales';
run;



TITLE 'Count of Top Sales';
PROC SQL  outobs = 20;
SELECT Sales
FROM MMEELAHI.TRANSACTION_V
ORDER BY Sales DESC

;
QUIT;


DATA MMEELAHI.LOCATION_T;
SET MMEELAHI.LOCATION_S;
IF Source  = " " then delete;
run;



PROC freq DATA=MMEELAHI.LOCATION_T;
format _CHAR_ $missfmt.; /* apply format for the duration of this
PROC */
tables _CHAR_ / missing missprint nocum nopercent;
format _NUMERIC_ missfmt.;
tables _NUMERIC_ / missing missprint nocum nopercent;
run;



proc print data=MMEELAHI.LOCATION_T (OBS = 20);
title 'Location Dataset';
run;
```

```
* Left joining transaction and location datasets;
proc sql;
create table MMEELAHI.INFERENTIAL_P as
select x.*, y.City, y.Prov, y.Postal_Code
from MMEELAHI.TRANSACTION_V x left join MMEELAHI.LOCATION_T y
on x.customer_id=y.customer_number
;
quit;


proc print data=MMEELAHI.INFERENTIAL_P (OBS = 20);
title 'Transaction Dataset';
run;



PROC freq DATA=MMEELAHI.INFERENTIAL_P ;
format _CHAR_ $missfmt.; /* apply format for the duration of this
PROC */
tables _CHAR_ / missing missprint nocum nopercent;
format _NUMERIC_ missfmt.;
tables _NUMERIC_ / missing missprint nocum nopercent;
run;


*REMOVE DUPLICATE OBSERVATIONS and check count again;
PROC SORT DATA = MMEELAHI.INFERENTIAL_P OUT = MMEELAHI.INFERENTIAL
dupout=MMEELAHI.INFERENTIAL_P_aDup NODUPRECS;
BY _ALL_;
RUN;



PROC freq DATA=MMEELAHI.INFERENTIAL ;
format _CHAR_ $missfmt.; /* apply format for the duration of this
PROC */
tables _CHAR_ / missing missprint nocum nopercent;
format _NUMERIC_ missfmt.;
tables _NUMERIC_ / missing missprint nocum nopercent;
run;

proc contents data=MMEELAHI.INFERENTIAL;
title 'Contents of the joint Dataset';
run;
```

```
DATA MMEELAHI.INFERENTIAL;
SET MMEELAHI.INFERENTIAL;
DROP Item_Code Item_Description City Postal_Code;
run;


DATA MMEELAHI.INFERENTIAL_Q;
SET MMEELAHI.INFERENTIAL;
run;



DATA MMEELAHI.INFERENTIALQ;
SET MMEELAHI.INFERENTIALQ;
IF Month = 1 THEN Month_Q = "JAN";
ELSE IF Month = 2 THEN Month_Q = "FEB";
ELSE IF Month = 3 THEN Month_Q = "MAR";
ELSE IF Month = 4 THEN Month_Q = "APR";
ELSE IF Month = 5 THEN Month_Q = "MAY";
ELSE IF Month = 6 THEN Month_Q = "JUN";
ELSE IF Month = 7 THEN Month_Q = "JUL";
ELSE IF Month = 8 THEN Month_Q = "AUG";
ELSE IF Month = 9 THEN Month_Q = "SEP";
ELSE IF Month = 10 THEN Month_Q = "OCT";
ELSE IF Month = 11 THEN Month_Q = "NOV";
ELSE IF Month = 12 THEN Month_Q = "DEC";
PROC PRINT DATA = MMEELAHI.INFERENTIALQ ( OBS = 20);
run;

* Converting Day of Week to String;
DATA MMEELAHI.INFERENTIALQ;
SET MMEELAHI.INFERENTIALQ;
IF Week_Day = 1 THEN Week_Day_Q = "MON";
ELSE IF Week_Day = 2 THEN Week_Day_Q = "TUE";
ELSE IF Week_Day = 3 THEN Week_Day_Q = "WED";
ELSE IF Week_Day = 4 THEN Week_Day_Q = "THU";
ELSE IF Week_Day = 5 THEN Week_Day_Q = "FRI";
ELSE IF Week_Day = 6 THEN Week_Day_Q = "SAT";
ELSE IF Week_Day = 7 THEN Week_Day_Q = "SUN";

PROC PRINT DATA = MMEELAHI.INFERENTIALQ ( OBS = 20);
run;
```

```
DATA MMEELAHI.INFERENTIALQ;
SET MMEELAHI.INFERENTIALQ;
DROP Week_Q;
IF 1 <= DAY <= 7  THEN Day_Q = "Week01";
ELSE IF 8 <= DAY <= 14  THEN Day_Q = "Week02";
ELSE IF 15 <= DAY <= 21  THEN Day_Q = "Week03";
ELSE IF 22 <= DAY <= 31  THEN Day_Q = "Week04";

PROC PRINT DATA = MMEELAHI.INFERENTIALQ ( OBS = 20);
run;

DATA MMEELAHI.INFERENTIALQ;
SET MMEELAHI.INFERENTIALQ(RENAME = (Day_Q=Monthly_Week));
run;

PROC PRINT DATA = MMEELAHI.INFERENTIALQ ( OBS = 20);
run;

DATA MMEELAHI.INFERENTIALQ;
SET MMEELAHI.INFERENTIALQ;
IF Year = 2007 THEN Transaction_Year = "Y2007";
ELSE IF Year = 2008 THEN Transaction_Year = "Y2008";

PROC PRINT DATA = MMEELAHI.INFERENTIALQ ( OBS = 20);
run;



DATA MMEELAHI.INFERENTIALQ;
SET MMEELAHI.INFERENTIALQ;
IF 1 <= Sales <= 150 THEN Sales_Q = "LOW";
ELSE IF 151 <= Sales <= 500 THEN Sales_Q = "MED";
ELSE IF Sales > 500 THEN Sales_Q = "HI";

PROC PRINT DATA = MMEELAHI.INFERENTIALQ ( OBS = 20);
run;



*UNIVARIATE ANALYSIS;

title 'Distribution of Sales Segments';
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
vbar Sales_Q;

RUN;
```

```
title 'Distribution of Source';
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
vbar Source;
RUN;
```

```
title 'Distribution of Year';
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
PIE Transaction_Year;
RUN;
```

```
title 'Distribution of Year';
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
vbar Transaction_Year;
RUN;
```

```
title 'Distribution of Quarter';
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
vbar Quarter_Q;
RUN;
```

```
title 'Distribution of Month';
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
vbar Month_Q;
RUN;
```

```
title 'Distribution of Day of Week';
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
vbar Week_Day_Q;
RUN;
```

```
title 'Distribution of Week of Month';
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
vbar Monthly_Week;
RUN;


title 'Distribution of Week of Year';
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
vbar Week_P;
RUN;



title 'Distribution of Week of Year';
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
vbar Week;
RUN;



title 'Distribution of Day of Month';
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
vbar Day;
RUN;









title 'Distribution of Province';
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
vbar Prov;
RUN;



title 'Distribution of Sales';
PROC UNIVARIATE DATA = MMEELAHI.INFERENTIAL;
var Sales;
histogram/normal;
RUN;
```

```
title 'Distribution of Category';
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
vbar Category;
RUN;



title 'Distribution of Price';
PROC UNIVARIATE DATA = MMEELAHI.INFERENTIAL;
var Price;
histogram/normal;
RUN;



title 'Distribution of Quantity';
PROC UNIVARIATE DATA = MMEELAHI.INFERENTIAL;
var Quantity;
histogram/normal;
RUN;



* BIVARIATE ANALYSIS;



TITLE "Total Sales per Category";
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
format Sales dollar20.;
pie3d Category / sumvar=Sales
VALUE = INSIDE
explode="F";
run;
quit;



TITLE "Total Sales per Category";
proc sgplot data=MMEELAHI.INFERENTIALQ;
format Sales dollar20.;
vbar Category/response=Sales stat=sum
       categoryorder=respdesc;
run;
```

```
TITLE "Total Quantity per Category";
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
pie3d Category / sumvar=Quantity
VALUE = INSIDE
explode="F";
run;
quit;
```

```
TITLE "Total Quantity per Category";
proc sgplot data=MMEELAHI.INFERENTIALQ;
vbar Category/response=Quantity stat=sum
       categoryorder=respdesc;
run;
```

```
TITLE "Total Sales per Source";
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
format Sales dollar20.;
pie3d Source / sumvar=Sales
VALUE = INSIDE
explode="F";
run;
quit;
```

```
TITLE "Total Sales per Source";
proc sgplot data=MMEELAHI.INFERENTIALQ;
format Sales dollar20.;
vbar Source/response=Sales stat=sum
       categoryorder=respdesc;
run;
```

```
TITLE "Total Quantity per Source";
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
pie3d Source / sumvar=Quantity
VALUE = INSIDE
explode="F";
run;
quit;
```

```
TITLE "Total Quantity per Source";
proc sgplot data=MMEELAHI.INFERENTIALQ;
vbar Source/response=Quantity stat=sum
       categoryorder=respdesc;
run;
```

```
TITLE "Total Sales per Province";
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
format Sales dollar20.;
pie3d Prov / sumvar=Sales
VALUE = INSIDE
explode="F";
run;
quit;
```

```
TITLE "Total Sales per Province";
proc sgplot data=MMEELAHI.INFERENTIALQ;
format Sales dollar20.;
vbar Prov/response=Sales stat=sum
       categoryorder=respdesc;
run;
```

```
TITLE "Total Quantity per Province";
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
pie3d Prov / sumvar=Quantity
VALUE = INSIDE
explode="F";
run;
quit;
```

```
TITLE "Total Quantity per Province";
proc sgplot data=MMEELAHI.INFERENTIALQ;
vbar Prov/response=Quantity stat=sum
       categoryorder=respdesc;
run;


TITLE "Total Sales per Quarter";
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
format Sales dollar20.;
pie3d Quarter_Q / sumvar=Sales
VALUE = INSIDE
explode="F";
run;
quit;



TITLE "Total Sales per Quarter";
proc sgplot data=MMEELAHI.INFERENTIALQ;
format Sales dollar20.;
vbar Quarter_Q/response=Sales stat=sum
       categoryorder=respdesc;
run;



TITLE "Total Quantity per Quarter";
PROC GCHART DATA = MMEELAHI.INFERENTIALQ;
pie3d Quarter_Q / sumvar=Quantity
VALUE = INSIDE
explode="F";
run;
quit;



TITLE "Total Quantity per Quarter";
proc sgplot data=MMEELAHI.INFERENTIALQ;
vbar Quarter_Q/response=Quantity stat=sum
       categoryorder=respdesc;
run;
```

```
TITLE "Total Sales per Month";
proc gchart data=MMEELAHI.INFERENTIALQ;
format Sales dollar20.;
vbar Month_Q / noframe type=SUM sumvar=Sales ;
format Month_Q ;
run; quit;
```

```
TITLE "Total Sales per Month";
proc sgplot data=MMEELAHI.INFERENTIALQ;
format Sales dollar20.;
vbar Month_Q/response=Sales stat=sum
       categoryorder=respdesc;
run;
```

```
TITLE "Total Quantity per Month";
proc gchart data=MMEELAHI.INFERENTIALQ;
vbar Month_Q / noframe type=SUM sumvar=Quantity ;
format Month_Q ;
run; quit;
```

```
TITLE "Total Quantity per Month";
proc sgplot data=MMEELAHI.INFERENTIALQ;
vbar Month_Q/response=Quantity stat=sum
       categoryorder=respdesc;
run;
```

```
TITLE "Total Sales per Day of Week";
proc gchart data=MMEELAHI.INFERENTIALQ;
format Sales dollar20.;
vbar Week_Day_Q / noframe type=SUM sumvar=Sales ;

run; quit;
```

```
TITLE "Total Sales per Day of Week";
proc sgplot data=MMEELAHI.INFERENTIALQ;
format Sales dollar20.;
vbar Week_Day_Q/response=Sales stat=sum
       categoryorder=respdesc;
run;



TITLE "Total Quantity per Day of Week";
proc gchart data=MMEELAHI.INFERENTIALQ;
 vbar Week_Day_Q / noframe type=SUM sumvar=Quantity ;

run; quit;



TITLE "Total Quantity per Day of Week";
proc sgplot data=MMEELAHI.INFERENTIALQ;
vbar Week_Day_Q/response=Quantity stat=sum
       categoryorder=respdesc;
run;




TITLE "Mean Sales per Week of Month";
proc gchart data=MMEELAHI.INFERENTIALQ;
 format Sales dollar20.;
 vbar Monthly_Week / noframe type=MEAN sumvar=Sales ;

run; quit;



TITLE "Mean Sales per Week of Month";
proc sgplot data=MMEELAHI.INFERENTIALQ;
format Sales dollar20.;
vbar Monthly_Week/response=Sales stat=sum
       categoryorder=respdesc;
run;

TITLE "Mean Quantity per Week of Month";
proc gchart data=MMEELAHI.INFERENTIALQ;
 vbar Monthly_Week / noframe type=MEAN sumvar=Quantity ;

run; quit;
```

```
TITLE "Mean Quantity per Week of Month";
proc sgplot data=MMEELAHI.INFERENTIALQ;
vbar Monthly_Week/response=Quantity stat=sum
       categoryorder=respdesc;
run;



TITLE "Top 10 Customers Based On Sales Volume";
PROC SQL outobs = 10;
SELECT Customer_ID, SUM(Sales) format dollar13.2 AS
Total_Per_Customer
FROM MMEELAHI.INFERENTIALQ
GROUP BY Customer_ID
ORDER BY Total_Per_Customer DESC
;
quit;



* Hypothesis testing;


*ANOVA : ANALYSIS OF VARIANCE;
title 'Anova Testing Between Sales and Source';
PROC ANOVA DATA = MMEELAHI.INFERENTIALQ;
 CLASS Source;
 MODEL Sales = Source;
 MEANS Source/SCHEFFE;
RUN;



PROC ANOVA DATA = MMEELAHI.INFERENTIALQ;
 CLASS Category;
 MODEL Sales = Category;
 MEANS Category/SCHEFFE;
RUN;
```

```
PROC ANOVA DATA = MMEELAHI.INFERENTIALQ;
 CLASS Prov;
 MODEL Sales = Prov;
 MEANS Prov/SCHEFFE;
RUN;



PROC ANOVA DATA = MMEELAHI.INFERENTIALQ;
 CLASS Quarter_Q;
 MODEL Sales = Quarter_Q;
 MEANS Quarter_Q/SCHEFFE;
RUN;



PROC ANOVA DATA = MMEELAHI.INFERENTIALQ;
 CLASS Month_Q;
 MODEL Sales = Month_Q;
 MEANS Month_Q/SCHEFFE;
RUN;


PROC ANOVA DATA = MMEELAHI.INFERENTIALQ;
 CLASS Week_Day_Q;
 MODEL Sales = Week_Day_Q;
 MEANS Week_Day_Q/SCHEFFE;
RUN;


PROC ANOVA DATA = MMEELAHI.INFERENTIALQ;
 CLASS Monthly_Week;
 MODEL Sales = Monthly_Week;
 MEANS Monthly_Week/SCHEFFE;
RUN;


* SPEARMAN CORRELATION;

*IF YOUR DATA IS NOT NL DISTRIBUTED : SPEARMAN CORRELATION;
title 'Correlation Testing Between Sales and Price & Quantity';
PROC CORR DATA = MMEELAHI.INFERENTIAL SPEARMAN;
VAR Price Quantity;
WITH Sales;
RUN;
```

```
* MODEL BUILDING;


PROC STANDARD DATA = MMEELAHI.INFERENTIAL MEAN = 0 STD = 1 OUT =
MMEELAHI.INFERENTIAL_MODEL;
     VAR Price Sales Quantity;
     title 'Price, Sales and Quantity are Standardized for Linear
Regression';
     RUN;



PROC UNIVARIATE DATA = MMEELAHI.INFERENTIAL_MODEL NORMAL;
var Sales;
histogram/normal;
title 'Sales Standardized for Linear Regression';
RUN;



PROC UNIVARIATE DATA = MMEELAHI.INFERENTIAL_MODEL NORMAL;
var Quantity;
histogram/normal;
title 'Quantity Standardized for Linear Regression';
RUN;



PROC UNIVARIATE DATA = MMEELAHI.INFERENTIAL_MODEL NORMAL;
var Price;
histogram/normal;
title 'Price Standardized for Linear Regression';
RUN;



TITLE "Linear Regression Model";
ODS GRAPHICS ON;
PROC GLMSELECT DATA = MMEELAHI.INFERENTIAL_MODEL PLOTS = ALL;
     CLASS Prov Source Category;
     MODEL Sales = Quantity Price Prov Source Category / DETAILS =
ALL STATS = ALL;
ODS GRAPHICS OFF;
RUN;
```