STRATEGIC ANALYSIS ON BLACK FRIDAY SALES

SAS Major Project Report

A sales data analysis report identifies critical factors affecting customer purchases and develops strategic ways to increase their revenue.

Presented to: Mr. Ar Kar Min
Presenter Name: Sajith Vellappillil
Date: 01-Aug-2021

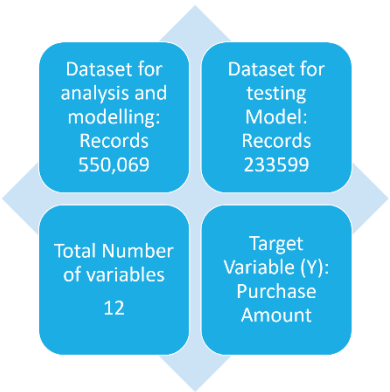
Table of Contents

Sr. No.	Particulars	Page No.
1	Introduction	1
2	Background	2
3	Objective	2
4	Methodology	2-4
5	Conceptual Framework	4
6	Descriptive Analysis	4-7
7	Inferential Statistics	7-8
8	Predictive Analysis	9-10
9	Conclusion	10-11
10	Recommendation	11
11	Appendix	12-25

Introduction

SAS, the most widely used statistical analysis software, was previously known as the 'Statistical Analysis system.' SAS was developed in the early 1970s at North Carolina State University originally intended for the management and analysis of agricultural field experiments. However, SAS is now used for data management, advanced analytics, multivariate analysis, business intelligence, criminal investigation, and predictive analytics. SAS is licensed software, hence preferred by large companies and government institutions.

For SAS major project, the dataset chosen is sales data of a retail company, "ABC Private Limited." They have shared purchase summaries of various customers for selected high-volume products from last month. The retailer records the customer's demographic details, product details and total purchases. The store, all together, generated total revenue of $5,095,812,742, and the total number of purchases made was 550,069. The dataset has information for their sales at 3 different cities, namely A, B and C, and wants to understand the customer purchase behaviour (precisely, purchase amount) against various products of different categories. The products have been categorized into 3 different groups.



Interestingly, we see that a product category does not remain consistent for a product for each purchase. Based on the given data analysis, the company would want to make strategic decisions and has shared a smaller dataset of 233599 records. Thus, we can test our predictive model to predict the purchase value.

Background

With all the shopping activity that takes place the Friday after Thanksgiving, the day became one of the most profitable days of the year for retailers and businesses. Because accountants use black to signify profit when recording each day's book entries (and red to indicate a loss), the day became known as Black Friday or when retailers see positive earnings and profits "in the black." So, Black Friday is now known as a profitable Friday in the retail industry and the rest of the economy.

Research Questions

1. WHICH product category must be focused on for promotion?
2. WHICH category for each below must be targeted for promotion?
    - Age Group
    - Male / Female
    - Married / Unmarried
    - City Category
    - No. of years in current city

Objectives

The primary objective is to increase the revenue by studying the data. The below sub-objective supports the primary objective:

1. Data Analysis: to find correlations / dependency of the purchases with customer's demographic factors.
2. Modeling: to create model to predict purchases for future business.
3. Develop Strategies: to focus on increasing revenue

Methodology

DATA COLLECTION: Secondary data from Metro College of Technology

DATA DEFINITION: There are 2 datasets Train.csv and Test.csv

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
4	Age	Char	5	$5.	$5.
6	City_Category	Char	1	$1.	$1.
3	Gender	Char	1	$1.	$1.
8	Marital_Status	Num	8	BEST12.	BEST32.
5	Occupation	Num	8	BEST12.	BEST32.
9	Product_Category_1	Num	8	BEST12.	BEST32.
10	Product_Category_2	Num	8	BEST12.	BEST32.
11	Product_Category_3	Num	8	BEST12.	BEST32.
2	Product_ID	Char	9	$9.	$9.
12	Purchase	Num	8	BEST12.	BEST32.
7	Stay_In_Current_City_Years	Char	2	$2.	$2.
1	User_ID	Num	8	BEST12.	BEST32.

Observations	550068
Variables	12
Indexes	0
Observation Length	80
Deleted Observations	0
Compressed	NO
Sorted	NO

The train dataset would be used to perform various analysis, study correlations / dependency and draw valuable conclusion. These insights serve basis for strategic

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
4	Age	Char	5	$5.	$5.
6	City_Category	Char	1	$1.	$1.
3	Gender	Char	1	$1.	$1.
8	Marital_Status	Num	8	BEST12.	BEST32.
5	Occupation	Num	8	BEST12.	BEST32.
9	Product_Category_1	Num	8	BEST12.	BEST32.
10	Product_Category_2	Num	8	BEST12.	BEST32.
11	Product_Category_3	Num	8	BEST12.	BEST32.
2	Product_ID	Char	9	$9.	$9.
7	Stay_In_Current_City_Years	Char	2	$2.	$2.
1	User_ID	Num	8	BEST12.	BEST32.

Observations	233599
Variables	11
Indexes	0
Observation Length	72
Deleted Observations	0
Compressed	NO
Sorted	NO

The test dataset serves the purpose of testing the model created using train dataset and predict the purchase value.

DATA VALIDATION: We combine the train and test dataset to find the missing values and the outliers for validating the data.

Gender	Frequency	Percent
Not Missing	783667	100.00

Age	Frequency	Percent
Not Missing	783667	100.00

City_Category	Frequency	Percent
Not Missing	783667	100.00

City_Years	Frequency	Percent
Not Missing	783667	100.00

source	Frequency	Percent
Not Missing	783667	100.00

Occupation	Frequency	Percent
Not Missing	783667	100.00

Marital_Status	Frequency	Percent
Not Missing	783667	100.00

PC1	Frequency	Percent
Not Missing	783667	100.00

PC2	Frequency	Percent
Missing	245982	31.39
Not Missing	537685	68.61

PC3	Frequency	Percent
Missing	545809	69.65
Not Missing	237858	30.35

On checking for the missing values for all the 12 variables, we get the below results:

- Variables Gender, Age, City_Category, City_Years, Occupation, Marital Status and Product Category 1 (PC1) have **no missing** values.
- Product Category 2 (**PC2**) has **missing values of 31.39%;** hence, it needs to be treated.
- Product Category 3 (**PC3**) has **missing values of 69.65%**; hence, the variable will be dropped.

EXCLUSION / INCLUSION: Exclude the below variable; only **9** variables remain for our analysis

- User ID
- Product ID
- Product Category 3 (PC3)

SOFTWARE: The data analysis is done using the statistical software SAS 9.4

STATISTICAL METHODS: The below SAS procedures were used for analysis:

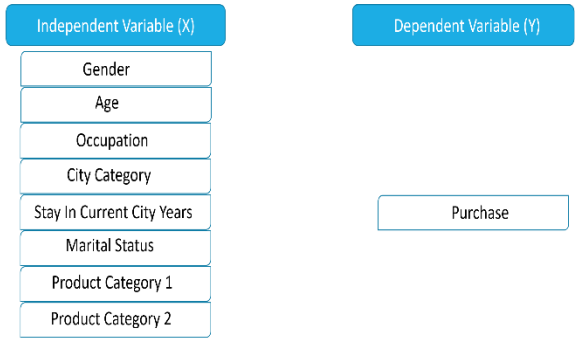
1. Proc Means
2. Proc Freq
3. Proc Sql
4. Proc Sgplot
5. Proc print
6. Proc Format
7. Proc Reg
8. Proc Standard

Conceptual Framework

The variables are divided into 2 – Independent Variable and Dependent Variable:

The dependent variable 'Y' is our target variable, which means that this variable is influenced by some or all other variables.

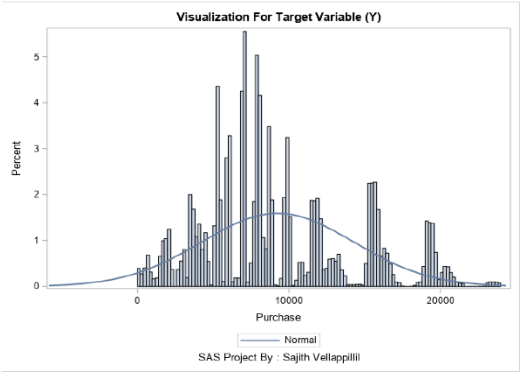
TARGET VARIABLE 'Y' – Purchase



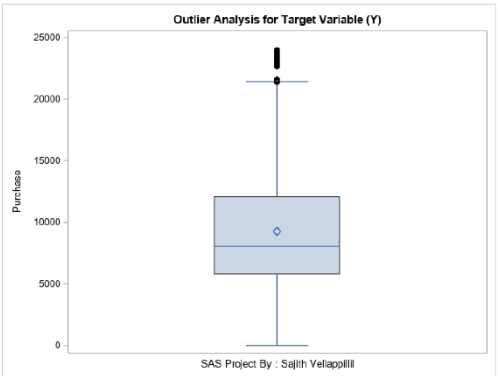
Descriptive Analysis

Univariate Analysis

THE TARGET VARIABLE - Purchase

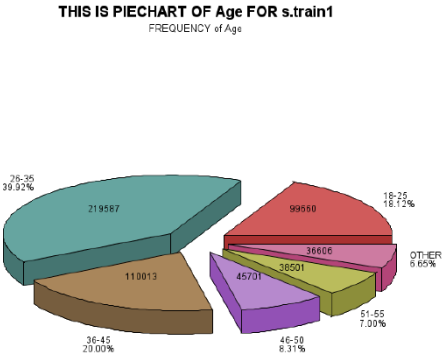


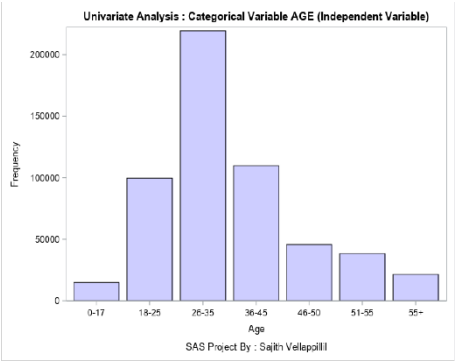
Basic Statistical Measures			
Location		Variability	
Mean	9263.969	Std Deviation	5023
Median	8047.000	Variance	25231186
Mode	7011.000	Range	23949
		Interquartile Range	6231



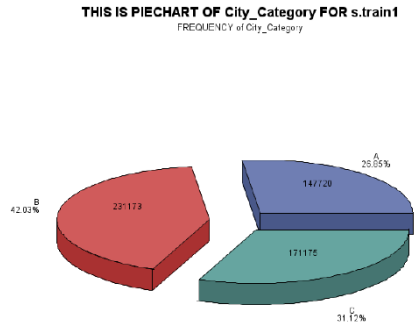
Based on the analysis, we see that that the distribution looks normal. There are few outliers, but taking into consideration the nature of the variable, treating any outliers would turn the data into biased data. Hence, we do not act on the outliers. From the basic statistical measure, we also interpret that the average purchase value is $9264.

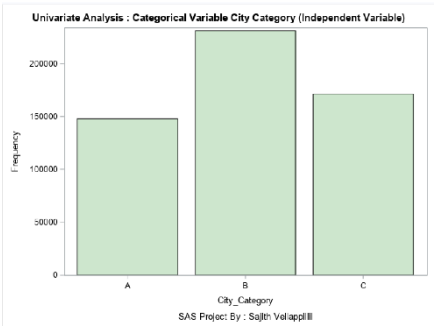
INDEPENDENT VARIABLES





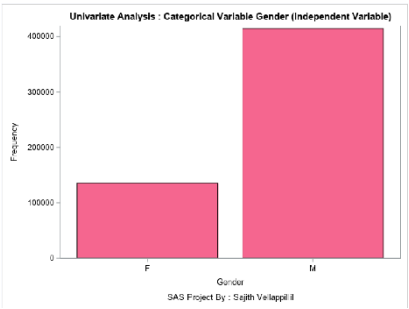
<u>Age</u>-The Pie-diagram and the bar diagram alongside show that the higher number of purchases are between age range 26-25 with 40% of the total followed by the age range of 36-45COMPRISING OF 20% and then 18-25 at 18%

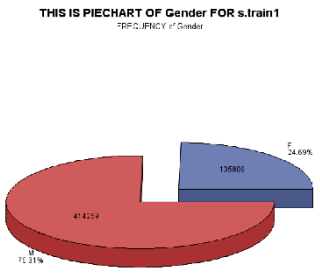




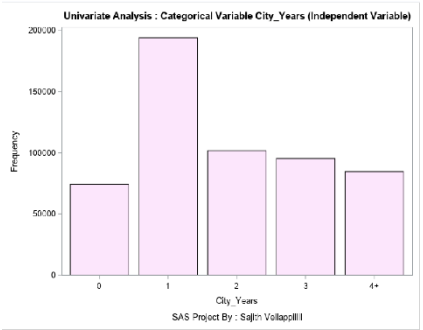
<u>City_Category</u> – From the chart, we understand that City B has the highest number of people making purchases at 42%.

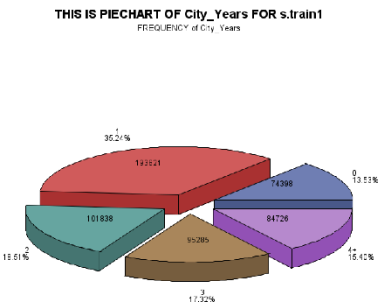
<u>Gender</u> – From the analysis on the right, we see that the male population are more active and involved in purchases at 75% than the female which is at 25%

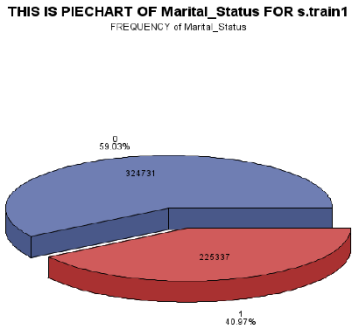


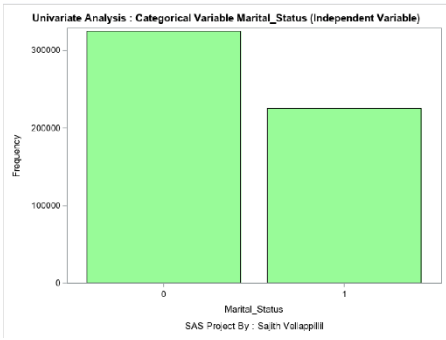


<u>City_Years</u> – The figures alongside represent that 35% of the customers are new in the city had stayed for only 1-year. It is followed by customers who lived 2-years in the city at 19%





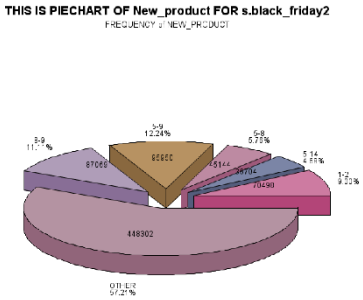




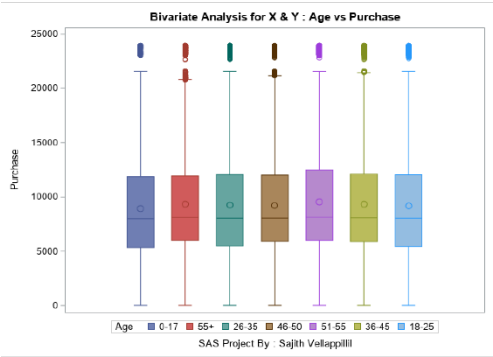
<u>Marital_Status</u> – We see that the singles are buying more than the married customers by 20%

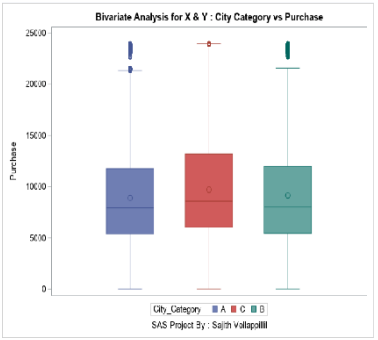
<u>New_Product</u> – We have combined PC1 and PC2 into a new column, 'New_Product.' The top 5 revenue-generating product categories are

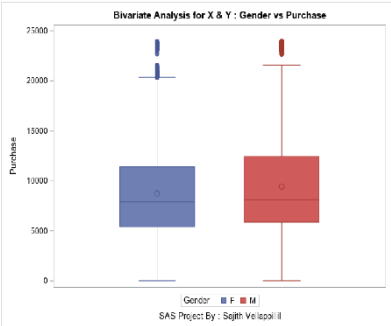
1. 5-9 at 12%
2. 8-9 at 11%
3. 1-2 at 9%
4. 5-8 at 6%
5. 5-14 at 5%



Bivariate Analysis (X vs Y)



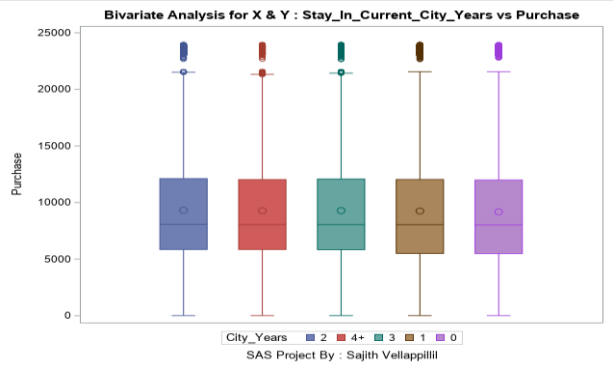


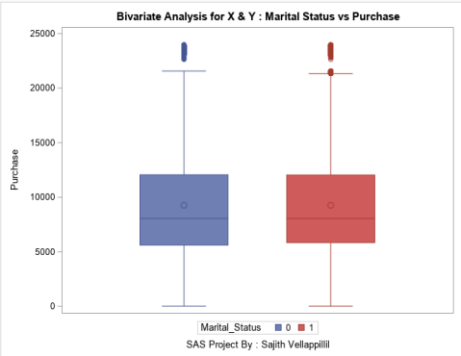


AGE vs PURCHASE – We see that the age group 51-55 has made the highest purchase value

CITY_CATEGORY vs PURCHASE – We see that the here city Cis leading in the total sales generated.

GENDER vs PURCHASE – We see that the male population shops more than female.





CITY_YEARS vs PURCHASE – We see that the irrespective of the number of years lived in the city, the purchase value for each group is similar.

MARITAL_STATUS vs PURCHASE – We see that there is no significant difference for total purchases between both the status.

Obs	NEW_PRODUCT	TOTAL_PURCHASE
1	1-2	670303864
2	8-9	459663440
3	5-9	408732166
4	1-15	250323220
5	1-8	223177440
6	5-8	207117667
7	1-16	203524781
8	6-8	191174483
9	3-4	167804291
10	5-14	160730868

NEW_PRODUCT Vs PURCHASE – From the fig on the left, the top 5 is the highest selling product categories are
1-2, 8-9, 5-9, 1-15 and 1-8.
The highest total sales value for the product category 1-2 is $670,303,864

Inferential Statistics

Product Category, Age, Gender, Marital status, City Category, and Stay in a city are factors that may influence total revenue earning.

Promotions have a massive impact on Black Friday Sales. It's a tool when utilized effectively, will generate considerable revenue. Analyzing the dependency, identifying yield prospects and strike with marketing gimmicks suitably is a proven way to achieve the primary business objective of 'PROFIT MAKING.'

H0: There is no statistically significant relationship between the given independent variable and purchase.

H1: There is a statistically significant relationship between the given independent variable and purchase.

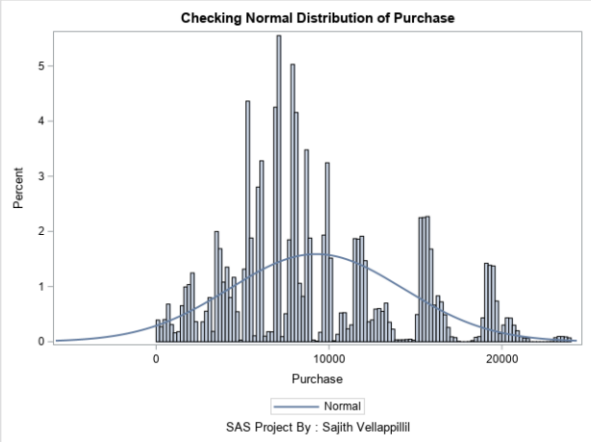
Significance level→5% -- below 5% --Reject the null hypothesis

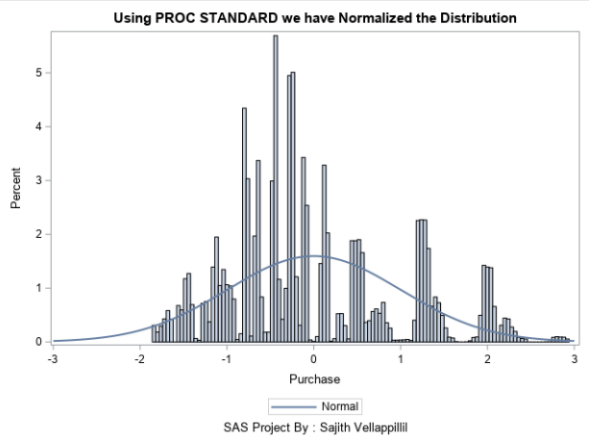
Decision: Reject/Fail to reject H0

Independent Variable (X) vs Purchase [Dependent Variable (Y)]			
AGE	<.001	ANOVA	Reject the null hypothesis
GENDER	<.001	T-TEST	Reject the null hypothesis
MARITAL STATUS	0.73	T-TEST	Fail to Reject the null hypothesis
OCCUPATION	<.001	ANOVA	Reject the null hypothesis
CITY CATEGORY	<.001	ANOVA	Reject the null hypothesis
STAY IN CURRENT CITY	<.001	ANOVA	Reject the null hypothesis
PRODUCT CATEGORY 1	<.001	ANOVA	Reject the null hypothesis

As the p-value for Marital_Status is higher than the significance level 5%, we fail to reject the null hypothesis, meaning, there is no correlation between Marital_Status and the Purchase.

Predictive Analysis





Here, the dataset is first checked to analyze if the distribution is normal. The bell-shaped density curve indicates that the distribution is normal. However, we can see that it is not a bell-shaped curve; we normalize the distribution using PROC STANDARD.

Even after standardizing, the result is the same, having no significant change.

The Test and Train datasets were merged to check and treat the missing values and outliers. Finally, the results are published under data validation.

Final dataset for Modeling

Gender	Age	Occupation	City_Category	City_Years	Purchase	NEW_PRODUCT
F	0-17	10	A	2	-0.17797274	3-9
F	0-17	10	A	2	1.1817547298	1-6
F	0-17	10	A	2	-1.561191842	12-9
F	0-17	10	A	2	-1.633856633	12-14
M	55+	16	C	4+	-0.25780447	8-9
M	26-35	15	A	3	1.1871299335	1-2
M	46-50	7	B	2	1.9810674373	1-8
M	46-50	7	B	2	1.3119541098	1-15
M	46-50	7	B	2	1.2785083975	1-16
M	26-35	20	A	1	-0.277314469	8-9

SAS Project By : Sajith Vellappillil

Using PROC GLMSELECT, we have created a prediction model. As PROC GLMSELECT has been used for modelling, no encoding has been done on the variables.

- P-Value less than significance level 5%
- R-Square at 0.647 – there is room for improvement.

Using PROC GLMSELECT Model

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 6).

Effects:	Intercept Age Gender City_Category Occupation City_Years NEW_PRODUCT

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	135	356370	2639.77960	7494.70	<.0001
Error	549932	193697	0.35222		
Corrected Total	550067	550067			

Root MSE	0.59348
Dependent Mean	-4.7168E-16
R-Square	0.6479
Adj R-Sq	0.6478
AIC	-23790
AICC	-23790
BIC	-573858
C(p)	136.00000
PRESS	193796
SBC	-572335
ASE	0.35213

Conclusions

- On analyzing **Product Category**, categories 1-2, 8-9, 5-9, 1-15,1-8 are top 5 for the high-value total purchases. Categories 5, 8, 1-2, 5-8 and 5-14 are in total times purchased.
- Regarding age, the highest number of people who purchased are from age group 26-35 comprising 40%, followed by group 36-45 having 20%. So, 60% of the total people count purchased are from the age group 26-45. However, the age group 51-55 spent the highest on purchases with an average of $9600 per person but comprised only 7% of the people count.
- Regarding **Gender** analysis, 75% of the total people purchased are male. The average spent by the male is $9438 whereas for the female is $8735.
- Based on **Marital Status** analysis, 60% of people purchased are single; however, the total purchase value is almost similar for each group. This means average spending is higher for people who are married. However, under the test of independence, marital status has no relation with the purchases.
- Our analysis on City Category shows that City B has a higher purchase count while total purchase value is higher for City C.
- Regarding **Years in City,** we see that many customers are newcomers whereas the total purchase value is higher for 2-years.

Recommendations

- Bundle offers must be offered combining the best-selling high revenue products like 1-2, 5-9 and 8-9 with the lower sale product, for instance, 6-8, 2-5, etc., with some margin. This is one of the best ways to boost the revenue from slow-moving items.
- Mobile-friendly and easily accessible websites and apps would positively impact increasing sales within the age group 51-55. In addition, multi-channel marketing is equally important, e.g., mail catalogues, as they may not be as tech-savvy as the age group 26-45.
- Even though the male purchase count and amounts are higher than the female, there is high earning potential from the female segment, which can be made real through marketing campaigns targeting women with a positive message and making them feel special.
- EMIs may be offered to attract married couples to make purchases as a trial and error.
- It is recommended to opt 'Segment Offers,' which means, for instance, bundle offer in one city would not be similar to the other as it must be based on which product is more prevalent in each city and bundle it with not so popular yet product that has potential to provide good margin.
- Referral-based discounts or coupons may be a solution to attract customers—be it new or old in the city. If executed effectively, it has a gamification aspect that can go viral.

Appendix

```
/*.............................********   Sajith Vellappillil    *********.............................*/
/*.........................***** SUBMITTED To : MR. AR KAR MIN ******.............................*/
/*.................................Black Friday Sales Data.....................................*/


libname S "C:\Users\sajit\OneDrive\Desktop\SAS Project\Personal DataSet\Black Friday Data";


*Setting The Header;

Title "SAS APPROACH TO BLACK FRIDAY SALES PREDICTION";

*Setting The Footer;

Footnote "SAS Project By : Sajith Vellappillil";


PROC IMPORT OUT= S.TRAIN
            DATAFILE= "C:\Users\sajit\OneDrive\Desktop\SAS Project\Personal DataSet\Black Friday Data\train.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
     GUESSINGROWS=5000;
RUN;


/*DATA PROFILING*/

/*Summarization of dataset contents*/
Title 'Summarization of Train DataSet';
PROC CONTENTS DATA=S.train;
RUN;

/*FOR PRINT ALL THE VARIABLE NAMES*/

PROC CONTENTS DATA = S.TRAIN VARNUM SHORT;
TITLE "VARIABLE NAME";
RUN;

* User_ID Product_ID Gender Age Occupation City_Category Stay_In_Current_City_Years
Marital_Status Product_Category_1 Product_Category_2 Product_Category_3 Purchase ;

TITLE "First 100 Train Dataset Obsevation";
PROC PRINT DATA =S.TRAIN (OBS=100) noobs;
run;

TITLE "First 100 Obsevation for specific column";
PROC PRINT DATA =S.TRAIN (OBS=100) ; *noobs;
var Product_ID Gender Age Occupation City_Category Stay_In_Current_City_Years
Marital_Status Product_Category_1 Product_Category_2 Product_Category_3 Purchase;
RUN;
```

```
 * Checking if there is any Duplicate key value in the Observation ;
proc sort data=S.train out = train nodupkey;
 by _All_;
run;

/*If We Want To Change Data Type - use put function

data s.train1;
set s.train;
Marital_Status_cat = put(Marital_Status, 12. -L); *character values are left-aligned for categorical column.;
Product_Category_1cat = put(Product_Category_1, 12. -L);
Product_Category_2cat = put(Product_Category_2, 12. -L);
Product_Category_3cat = put(Product_Category_3, 12. -L);
Occupation_cat = put(Occupation, 12. -L);
drop Product_Category_1 Product_Category_2
    Product_Category_3 Marital_Status Occupation
    user_ID Product_ID;
rename Product_Category_1cat=PC1 Product_Category_2cat=PC2 Product_Category_3cat=PC3
    Stay_In_Current_City_Years = City_Years;
options missing = ' '
;
run; */

data s.train1;
set s.train;
drop user_ID Product_ID;
rename Product_Category_1=PC1 Product_Category_2=PC2 Product_Category_3=PC3
    Stay_In_Current_City_Years = City_Years;
run;


TITLE "First 100 Obsevation";
PROC PRINT DATA =S.TRAIN1 (OBS=100) noobs;
run;

*************************************************************************************************************;
                                        *UNIVARIATE ANALYSIS for Numerical Variable;

Title 'Calculated Descriptive Statistics';
PROC MEANS DATA= s.train1 N NMISS MIN MEAN MEDIAN STD MAX MAXDEC=2;
var
Purchase
;
RUN;

/*Describe Numeric variable and Visualize for Continious Data*/
title 'Examining Data Distribution';
ods graphics on;
PROC UNIVARIATE DATA = s.train1;
 VAR
Purchase
;
HISTOGRAM;
RUN;
ods graphics off;

* Visualization for Numerical variable HISTOGRAM  AND DENISTY CRUVE;

Title 'Visualization For Target Variable (Y)';
ods graphics on;
PROC SGPLOT DATA =  s.train1;
 HISTOGRAM Purchase;
```

```
PROC SGPLOT DATA =  s.train1;
 HISTOGRAM Purchase;
 DENSITY Purchase;
RUN;
QUIT;
ods graphics off;

* BOXPLOT to Visually Check for Outliers - numeric variable ;

Title 'Outlier Analysis for Target Variable (Y)';
PROC SGPLOT DATA =  s.train1;
 VBOX Purchase;
RUN;
QUIT;

PROC SGPLOT DATA =  s.train1;
 HBOX Purchase;
RUN;
QUIT;

*************************************************************************************************************;
                                        *UNIVARIATE ANALYSIS for Categorical Variable;


/*Frequency for categorical Variable*/

Title 'Summarize Categorical Variable';
PROC FREQ DATA =s.train1;
 TABLE
Age
City_Category

/*Frequency for categorical Variable*/

Title 'Summarize Categorical Variable';
PROC FREQ DATA =s.train1;
 TABLE
Age
City_Category
Gender
Marital_Status
Occupation
PC1
PC2
PC3
City_Years
;
RUN;


PROC OPTIONS OPTION = MACRO;
RUN;


                            *MACRO FOR - PIE CHART AND FREQUENCY FOR CATEGORICAL VARIABLES ;
%MACRO UNI_ANALYSIS_CAT(DATA,VAR);
 TITLE "THIS IS FREQUENCY OF &VAR FOR &DATA";
  PROC FREQ DATA=&DATA;
  TABLE &VAR;
 RUN;

TITLE "THIS IS PIECHART OF &VAR FOR &DATA";
PROC GCHART DATA=&DATA;
```

```
TITLE "THIS IS PIECHART OF &VAR FOR &DATA";
PROC GCHART DATA=&DATA;
  PIE3D &VAR/discrete
             value=inside
             percent=outside
             EXPLODE=ALL
             SLICE=OUTSIDE
             RADIUS=20
;

RUN;
%MEND;
```

```
                                          *BAR CHART & PIE CHART FOR CATEGORICAL Variable;
Title 'Univariate Analysis : Categorical Variable AGE (Independent Variable)';
PROC SGPLOT DATA = s.trainl ;
 VBAR Age/FILLATTRS=(color=blue transparency=.8);
RUN;
QUIT;
```

```
TITLE "PIE CHART for Age";
%UNI_ANALYSIS_CAT(s.trainl,Age)
```

```
Title 'Univariate Analysis : Categorical Variable City Category (Independent Variable)';
PROC SGPLOT DATA = s.trainl ;
 VBAR City_Category/FILLATTRS=(color=green transparency=.8);
RUN;
QUIT;
```

```
TITLE "PIE CHART for City_Category";
%UNI_ANALYSIS_CAT(s.trainl,City_Category)
```

```
Title 'Univariate Analysis : Categorical Variable Gender (Independent Variable)';
PROC SGPLOT DATA = s.trainl ;
 VBAR Gender/FILLATTRS=(COLOR=CXEE0044 transparency=.4);
RUN;
QUIT;
```

```
TITLE "PIE CHART for Gender";
%UNI_ANALYSIS_CAT(s.trainl,Gender)
```

```
Title 'Univariate Analysis : Categorical Variable City_Years (Independent Variable)';
PROC SGPLOT DATA = s.trainl ;
 VBAR City_Years/FILLATTRS=(color=violet transparency=.8);
RUN;
QUIT;
```

```
TITLE "PIE CHART for City_Years";
%UNI_ANALYSIS_CAT(s.trainl,City_Years)
```

```
Title 'Univariate Analysis : Categorical Variable Marital_Status (Independent Variable)';
PROC SGPLOT DATA = s.trainl ;
 VBAR Marital_Status/FILLATTRS=(color=palegreen);
RUN;
QUIT;
```

```
TITLE "PIE CHART forMarital Status ";
%UNI_ANALYSIS_CAT(s.trainl,Marital_Status)


Title 'Univariate Analysis : Categorical Variable Occupation (Independent Variable)';
PROC SGPLOT DATA = s.trainl ;
 VBAR Occupation/FILLATTRS=(color=GRAY4F);
RUN;
QUIT;

TITLE "PIE CHART for Occupation";
%UNI_ANALYSIS_CAT(s.trainl,Occupation)


Title 'Univariate Analysis : Categorical Variable PC1 (Independent Variable)';
PROC SGPLOT DATA = s.trainl ;
 VBAR PC1/FILLATTRS=(color=V0F055FF);
RUN;
QUIT;

TITLE "PIE CHART for Product Category 1";
%UNI_ANALYSIS_CAT(s.trainl,PC1)


Title 'Univariate Analysis : Categorical Variable PC2 (Independent Variable)';
PROC SGPLOT DATA = s.trainl ;
 VBAR PC2/FILLATTRS=(color=H14055FF);
RUN;
QUIT;



TITLE "PIE CHART for Product Category 2";
%UNI_ANALYSIS_CAT(s.trainl,PC2)


Title 'Univariate Analysis : Categorical Variable PC3 (Independent Variable)';
PROC SGPLOT DATA = s.trainl ;
 VBAR PC3/FILLATTRS=(color=CX98FB98);
RUN;
QUIT;

TITLE "PIE CHART for Product Category 3";
%UNI_ANALYSIS_CAT(s.trainl,PC3)




*************************************************************************************************************;

                        *BIVARIATE ANALYSIS for (Continious) Numeric Variable (Y) and Categorical Variable (X);

*WHICH AGE CATEGORY SPENDS MORE ON A BLACKFRIDAY SALE;

ods graphics on;
TITLE "Bivariate Analysis for X & Y : Age vs Purchase";
PROC SGPLOT  DATA=S.trainl;
        VBOX Purchase / group = Age;
RUN;
ods graphics off;
```

```
/*Test Of Independence - Anova - if the Categorical Variable has more Thann 2 level*/
*HO: There is no statistically significant relation between AGE and PURCHASE
 H1: There is a statistically significant relation between AGE and PURCHASE;

PROC ANOVA DATA = S.train1;
 CLASS Age;
 MODEL PURCHASE = Age;
 MEANS Age/SCHEFFE;
RUN;
*<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<;

*WHICH CITY_CATEGORY CATEGORY SPENDS MORE ON A BLACKFRIDAY SALE;

TITLE "Bivariate Analysis for X & Y : City Category vs Purchase";
ods graphics on;
PROC SGPLOT  DATA=S.train1;
         VBOX Purchase / group = City_Category;
RUN;
ods graphics off;

/*Test Of Independence - Anova - if the Categorical Variable has more Thann 2 level*/

PROC ANOVA DATA = S.train1;
 CLASS City_Category;
 MODEL PURCHASE = City_Category;
 MEANS City_Category/SCHEFFE;
RUN;

*<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<;

*WHICH GENDER CATEGORY SPENDS MORE ON A BLACKFRIDAY SALE;

TITLE "Bivariate Analysis for X & Y : Gender vs Purchase";
ods graphics on;
PROC SGPLOT  DATA=S.train1;
         VBOX Purchase / group = Gender;
RUN;
ods graphics off;

/*if categorical column has only two levels :t-test*/
Title 'T-Test- Categorical Column Has Only 2 Levels';
proc ttest data=s.train1 ;
      class Gender;
      var Purchase;
   run;

*<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<;


* WHICH STAY IN CURRENT CITY YEARS CATEGORY SPENDS MORE ON A BLACKFRIDAY SALE;

TITLE "Bivariate Analysis for X & Y : Stay_In_Current_City_Years vs Purchase";
ods graphics on;
PROC SGPLOT  DATA=S.train1;
         VBOX Purchase / group = City_Years;
RUN;
ods graphics off;
```

```
/*Test Of Independence - Anova - if the Categorical Variable has more Thann 2 level*/

PROC ANOVA DATA = S.train1;
 CLASS City_Years;
 MODEL PURCHASE = City_Years;
 MEANS City_Years/SCHEFFE;
RUN;

*<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<;

*WHICH MARITAL_STATUS CATEGORY SPENDS MORE ON A BLACKFRIDAY SALE;

TITLE "Bivariate Analysis for X & Y : Marital Status vs Purchase";
PROC SGPLOT  DATA=S.train1;
         VBOX Purchase / group = Marital_Status;
RUN;
ods graphics off;


/*if categorical column has only two levels :t-test*/
Title 'T-Test Marital_Status vs Purchase';
proc ttest data=S.train1 ;
      class Marital_Status;
      var Purchase;
   run;

*<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<;

*WHICH OCCUPATION CATEGORY SPENDS MORE ON A BLACKFRIDAY SALE;


TITLE "Bivariate Analysis for X & Y : Occupation vs Purchase";
ods graphics on;
PROC SGPLOT  DATA=S.train1;
         VBOX Purchase / group = Occupation;
RUN;
ods graphics off;

/*Test Of Independence - Anova - if the Categorical Variable has more Thann 2 level*/

PROC ANOVA DATA = S.train1;
 CLASS Occupation;
 MODEL PURCHASE = Occupation;
 MEANS Occupation/SCHEFFE;
RUN;


*<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<;

*WHICH PRODUCT CATEGORY 1 CATEGORY SPENDS MORE ON A BLACKFRIDAY SALE;


TITLE "Bivariate Analysis for X & Y : PC1 vs Purchase";
ods graphics on;
PROC SGPLOT  DATA=S.black_friday;
         VBOX Purchase / group = NEW_PRODUCT;
RUN;
ods graphics off;
```

```
/*Test Of Independence - Anova - if the Categorical Variable has more Thann 2 level*/

title 'Anova testing for New Product Category';
PROC ANOVA DATA = S.black_friday;
 CLASS NEW_PRODUCT;
 MODEL PURCHASE = NEW_PRODUCT;
 MEANS NEW_PRODUCT/SCHEFFE;
RUN;

*<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<;

*WHICH PRODUCT CATEGORY 2 CATEGORY SPENDS MORE ON A BLACKFRIDAY SALE;

TITLE "Bivariate Analysis for X & Y : PC2 vs Purchase";
ods graphics on;
PROC SGPLOT  DATA=S.train1;
        VBOX Purchase / group = PC2;
RUN;

/*Test Of Independence - Anova - if the Categorical Variable has more Thann 2 level*/

PROC ANOVA DATA = S.train1;
 CLASS PC2;
 MODEL PURCHASE = PC2;
 MEANS PC2/SCHEFFE;
RUN;

*<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<;

*WHICH PRODUCT CATEGORY 3 CATEGORY SPENDS MORE ON A BLACKFRIDAY SALE;

*WHICH PRODUCT CATEGORY 3 CATEGORY SPENDS MORE ON A BLACKFRIDAY SALE;

TITLE "Bivariate Analysis for X & Y : PC3 vs Purchase";
ods graphics on;
PROC SGPLOT  DATA=S.train1;
        VBOX Purchase / group = PC3;
RUN;
ods graphics off;

/*Test Of Independence - Anova - if the Categorical Variable has more Thann 2 level*/

PROC ANOVA DATA = S.train1;
 CLASS PC3;
 MODEL PURCHASE = PC3;
 MEANS PC3/SCHEFFE;
RUN;


*************************************************************************************************************;
```

```
*************************************************************************************************************************;

/*                                      *If We Want To Treat The Outlier For Purchase;

proc univariate data = S.TRAIN1;
var purchase;
output out=boxStats median=median qrange = iqr;
run;

data null;
    set boxStats;
    call symput ('median',median);
    call symput ('iqr', iqr);
run;

%put &median;
%put &iqr;

data S.OUTLIER;
set S.TRAIN1;
    if (purchase le &median + 1.5 * &iqr) and (purchase ge &median - 1.5 * &iqr);
run;

proc print data = S.OUTLIER(obs=5);
run;

 PROC SGPLOT DATA = S.OUTLIER;
 VBOX PURCHASE;
 run;
 quit;
```

```
 /*Importing Test DataSet*/

PROC IMPORT OUT= S.Test
            DATAFILE= "C:\Users\sajit\OneDrive\Desktop\SAS Project\Personal DataSet\Black Friday Data\test.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
     GUESSINGROWS=5000;
 RUN;

proc contents data = S.Test;run;
 *************************************************************************************************************************;
 /*Removing Unwanted Column which removed from Test DataSet*/


data s.Test1;
 set s.Test;
 drop user_ID Product_ID;
 rename Product_Category_1=PC1 Product_Category_2=PC2 Product_Category_3=PC3
     Stay_In_Current_City_Years = City_Years;
 run;


                                /*MERGING TEST AND TRAIN DATASET WITH ADDED ONE COLUMN AS SOURCE*/

 Data S.T1;
  set S.Train1;
  source = 'train' ;
 run;
```

```
proc contents data=S.T1;
run;
Title 'Train dataset';
proc print data=S.T1 (obs=5) ; *noobs;
run;
Data S.T2;
 set S.Test1;
 source = 'test' ;
run;
Title 'Test dataset';
proc print data=S.T2 (obs=5) ; *noobs;
run;
*<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<;

*MERGING THE DATASETS TEST AND TRAIN FOR CLEANING PURPOSE;
data S.Black_Friday;
set S.T1 S.T2;
run;

Proc print data = S.black_friday (obs=78);run;

****************************************************************************************************************;

proc contents data = S.black_friday;run;


*<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<;
                                        * Identifying the Missing Value for Numerical Varibale;


                                        * Identifying the Missing Value for Numerical Varibale;

 Title "Count of Missing Value For Numerical Variables (Train Dataset)";
proc means data = S.Train1 n nmiss;
 var _numeric_;
 run;

 Title "Count of Missing Value For Numerical Variables (Test Dataset)";
proc means data = S.Test1 n nmiss;
 var _numeric_;
 run;


 Title "Count of Missing Value For Numerical Variables (Combined Dataset)";
proc means data = S.black_friday n nmiss;
 var _numeric_;
 run;

 *********************************************************************************************************;

                            * Identifying the Missing Value for Character & Numeric Varibale;

 Title "Count of Missing Value For Character Variables";

proc format ;
     value $ missfmt ' '="Missing" other="Not Missing";
     value  missfmt  . ='Missing' OTHER='Not Missing';
 run;
```

```
*Proc freq to count missing/non missing;
ods table onewayfreqs=temp;

proc freq data=S.black_friday;
    table _character_ / missing NOCUM;
    format  _character_ $missfmt.;
    FORMAT _Numeric_ missfmt.;
    TABLES _Numeric_ / Missing NOCUM;
run;

*************************************************************************************************

                                              /*TREAT THE MISSING VALUE*/


PROC STDIZE DATA = S.black_friday out =S.black_friday1 METHOD = MEDIAN REPONLY;
 VAR PC2;
RUN;

title 'Dataset after treating missing value';
PROC PRINT DATA= S.Black_Friday1(OBS=10);
RUN;

*<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<;

*<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<;

*concatenate  two product categories
 -----------------------------------;
DATA S.black_friday1;
SET S.black_friday1;
NEW_PRODUCT = CATX('-',PC1, PC2);
RUN;

proc print data = S.black_friday1 (obs=10);run;

************************************************************************************************;

/*REMOVING UNWANTED COLUMN*/


data S.Black_Friday1;
set S.Black_Friday1;
drop PC1 PC2 PC3 ;
run;

proc print data = S.Black_Friday1 (obs=10);run;
```

```
*************************************************************************************************************************;

* Checking our Dependent Variable (Y) is Normal Distribution or Not;
Title 'Checking Normal Distribution of Purchase';
PROC SGPLOT DATA = S.Black_Friday1;
HISTOGRAM Purchase;
DENSITY Purchase;
RUN;

*Dependent variable (Y) is not normally distributed. Using PROC STANDARD we have Normalized the Distribution.;
Title 'Using PROC STANDARD we have Normalized the Distribution';
PROC STANDARD DATA=S.Black_Friday1 MEAN=0 STD=1 OUT=S.Black_Friday2;
  VAR PURCHASE ;
RUN;

Title 'Checking Normal Distribution of Purchase';
PROC SGPLOT DATA = S.Black_Friday2;
HISTOGRAM Purchase;
DENSITY Purchase;
RUN;

/*Split TRAIN AND TEST*/

DATA S.Black_Friday_Train S.Black_Friday_Test;
SET  S.Black_Friday2;
IF  SOURCE='train' THEN OUTPUT S.Black_Friday_Train;
ELSE IF SOURCE='test' THEN OUTPUT S.Black_Friday_Test;
drop source Marital_Status;
RUN;


Title 'Final dataset for Modeling';
proc print data = S.Black_Friday_Train (obs = 10) noobs;
run;

proc print data = S.Black_Friday_Test (obs = 10) noobs;
run;

*Removing Dependent Variable Y from Test DataSet;

data S.Black_Friday_Test;
set S.Black_Friday_Test;
drop purchase;
run;


* Regression Analysis Using PROC GLMSELECT;

title 'Using PROC GLMSELECT Model';
PROC GLMSELECT data=S.Black_Friday_Train ;
class AGE GENDER City_Category City_Years Occupation new_product/ param=ref order=data;
model PURCHASE = AGE GENDER City_Category Occupation City_Years NEW_PRODUCT/ selection=stepwise select=SL
showpvalues stats=all STB;
QUIT;


*-------------------------------------------------------------------------------------------------------------------------
```

```
                                                    *BUSINESS QUESTIONS;

*1) Which Product category Must Be Focused For Promotion;

Title 'Business Question Answer';
PROC SQL;
CREATE TABLE S.Promotion AS SELECT DISTINCT New_product ,
SUM(PURCHASE) AS TOTAL_PURCHASE FROM S.black_friday1 GROUP BY new_product order by TOTAL_PURCHASE desc;
QUIT;

proc print data = S.Promotion (obs=5);run;

TITLE "PIE CHART for Product Category 3";
%UNI_ANALYSIS_CAT(s.black_friday2,New_product)

*-------------------------------------------------------------------------------------------------------------;

*2) Which Category For Each Below Must Be Targeted For Promotion

*Age group;

ods graphics on;
Title 'Business Question Answer';
PROC SGPLOT  DATA=S.train1;
        VBOX Purchase / group = Age;
RUN;
ods graphics off;


 /*Test Of Independence - Anova - if the Categorical Variable has more Thann 2 level*/
 *HO: There is no statistically significant relation between AGE and PURCHASE
  H1: There is a statistically significant relation between AGE and PURCHASE;

PROC ANOVA DATA = S.train1;
  CLASS Age;
  MODEL PURCHASE = Age;
  MEANS Age/SCHEFFE;
 RUN;

 *-------------------------------------------------------------------------------------------------------------

 *Gender;


 Title 'Business Question Answer';
 ods graphics on;
PROC SGPLOT  DATA=S.train1;
         VBOX Purchase / group = Gender;
RUN;
ods graphics off;

 /*if categorical column has only two levels :t-test*/

proc ttest data=s.train1 ;
       class Gender;
       var Purchase;
   run;
```

```
*----------------------------------------------------------------------------------------------------
*Marital Status;

Title 'Business Question Answer';
PROC SGPLOT  DATA=S.train1;
         VBOX Purchase / group = Marital_Status;
RUN;
ods graphics off;


/*if categorical column has only two levels :t-test*/

proc ttest data=S.train1 ;
      class Marital_Status;
      var Purchase;
   run;

*----------------------------------------------------------------------------------------------------

*City Category;

Title 'Business Question Answer';
ods graphics on;
PROC SGPLOT  DATA=S.train1;
         VBOX Purchase / group = City_Category;
RUN;
ods graphics off;

/*Test Of Independence - Anova - if the Categorical Variable has more Thann 2 level*/

PROC ANOVA DATA = S.train1;
 CLASS City_Category;
 MODEL PURCHASE = City_Category;
 MEANS City_Category/SCHEFFE;
RUN;

*----------------------------------------------------------------------------------------------------

*Stay In Current City Years;

Title 'Business Question Answer';
ods graphics on;
PROC SGPLOT  DATA=S.train1;
         VBOX Purchase / group = City_Years;
RUN;
ods graphics off;

/*Test Of Independence - Anova - if the Categorical Variable has more Thann 2 level*/

PROC ANOVA DATA = S.train1;
 CLASS City_Years;
 MODEL PURCHASE = City_Years;
 MEANS City_Years/SCHEFFE;
RUN;


****************************************************************************************************

**************************************         END          ****************************************
*****************************                                            ****************************
********************                                                             *******************


****************************************************************************************************
```