INFS 762 Project

Name: Md Mominul Islam (101009250)

Case description

A supermarket is offering a new line of organic products. The supermarket's management wants to determine which customers are likely to purchase these products.

The supermarket has a customer loyalty program. As an initial buyer incentive plan, the supermarket provided coupons for the organic products to all of the loyalty program participants and collected data that includes whether these customers purchased any of the organic products.

The ORGANICS data set contains 13 variables and over 22,000 observations.

Step 1. You need to download the dataset “organics.csv” in D2l-> Content-> Project folder and import it to SAS.

The variables in the data set are shown below with the appropriate roles and levels:

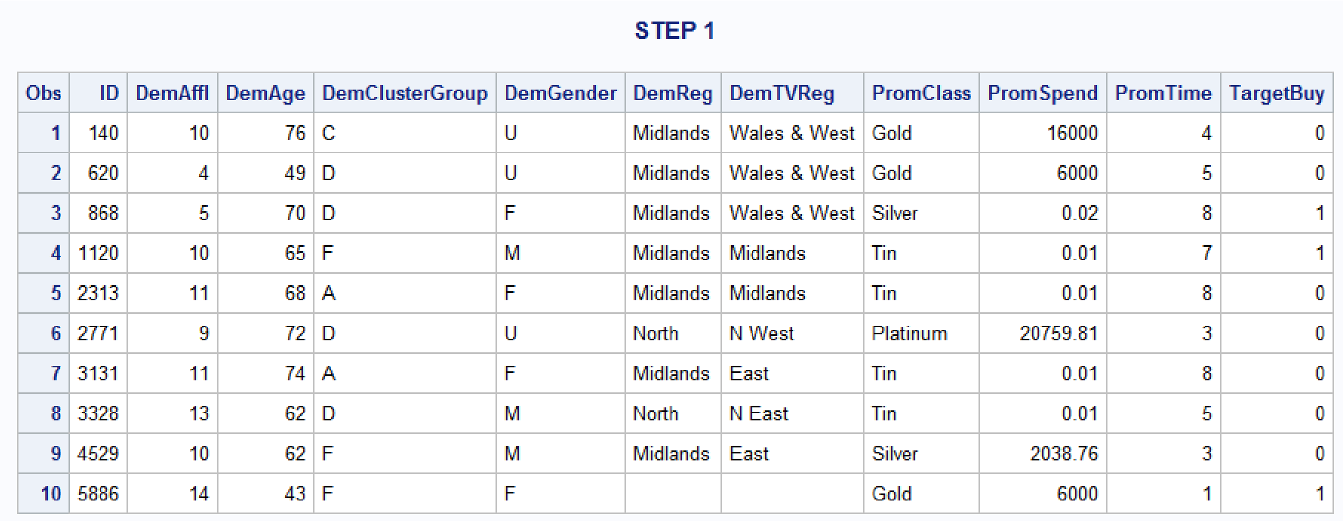
Now you want to remove two variables “DemCluster” and “TargetAmount”. The id of a customer shouldn’t be considered in data mining; however you don’t need to remove it. In the table below, the measurement level “nominal” means that the variable is categorical, and “Interval” means the variable is continuous. The target or dependent variable is “targetBuy”.

Variable	Type	Model role	Description
ID	Nominal	ID	Customer loyalty ID
DemAffl	Interval	Input	Affluence grade on a scale from 1 to 30
DemAge	Interval	Input	Age
DemCluster	Nominal	Rejected	Type of residential neighborhood
DemClusterGroup	Nominal	Input	Neighborhood group
DemGender	Nominal	Input	M, F, U
DemReg	Nominal	Input	Geo region
DemTVReg	Nominal	Input	TV region
PromClass	Nominal	Input	Loyalty status, tin, silver, gold or platinum
PromSpend	Interval	Input	Total amount spent
PromTime	Interval	Input	Time as a member
TargetBuy	Binary	Target	Organics purchased? 1 = yes 0 = no
TargetAmt	Interval	Rejected	Number of organic products purchased

Answer for Step 1:

```
TITLE 'STEP 1';

/*    Defiend Library first and then import data set “organics.csv” to SAS.*/

LIBNAME PROJECT1 'Z:\OneDrive - South Dakota State University - SDSU\INFS 762';
RUN;

/*Removing two variables named “DemCluster” and “TargetAmt” */

/* the code below illustrates how to drop a variable */
DATA PROJECT1.ORGANICS2;
SET PROJECT1.ORGANICS;
DROP DEMCLUSTER TARGETAMT;
RUN;
PROC PRINT DATA = PROJECT1.ORGANICS2(OBS = 10);

RUN;
```

First ten observations after dropping two variables named “DemCluster” and “TargetAmount”.



STEP 1

Obs	ID	DemAffl	DemAge	DemClusterGroup	DemGender	DemReg	DemTVReg	PromClass	PromSpend	PromTime	TargetBuy
1	140	10	76	C	U	Midlands	Wales & West	Gold	16000	4	0
2	620	4	49	D	U	Midlands	Wales & West	Gold	6000	5	0
3	868	5	70	D	F	Midlands	Wales & West	Silver	0.02	8	1
4	1120	10	65	F	M	Midlands	Midlands	Tin	0.01	7	1
5	2313	11	68	A	F	Midlands	Midlands	Tin	0.01	8	0
6	2771	9	72	D	U	North	N West	Platinum	20759.81	3	0
7	3131	11	74	A	F	Midlands	East	Tin	0.01	8	0
8	3328	13	62	D	M	North	N East	Tin	0.01	5	0
9	4529	10	62	F	M	Midlands	East	Silver	2038.76	3	0
10	5886	14	43	F	F			Gold	6000	1	1

In our data set, we have some missing values. The ‘ID’ column will not be used in data mining. Originally we had 13 variables. After removing two variables we ended up having 22,222 observations with 11 variables.

STEP 2: You to do a quality check. In this dataset, we do not have false/unreasonable values. You need to tell me: 1) which variables have a skewed distribution, and 2) which variables have missing values.

For the continuous variables, you need to print the histogram for each variable and also check the extreme values.

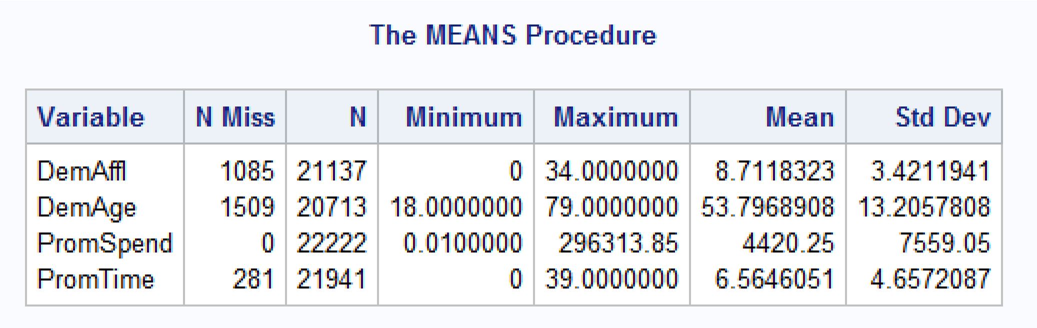
```
Answer:
```

```
/*Histogram for the continuous variables*/

TITLE 'STEP 2';

goptions reset=global
         gunit=pct
         hsize= 10.625 in
         vsize= 8.5 in
         htitle=4
         htext=3
         vorigin=0 in
         horigin= 0 in
         cback=white border
         ctext=black
         colors=(black blue green red yellow)
         ftext=swiss
         lfactor=3;

proc univariate data=PRJT.ORGANICS2 noprint;
   histogram DemAffl;
   title 'Histogram for Affluence grade on a scale from 1 to 30';
run;

proc univariate data=PRJT.ORGANICS2 noprint;
   histogram DemAge;
   title 'Histogram for Age';
run;

proc univariate data=PRJT.ORGANICS2 noprint;
   histogram PromSpend;
   title 'Histogram for Total amount spent';
run;

proc univariate data=PRJT.ORGANICS2 noprint;
   histogram PromTime;
   title 'Histogram for Time as a member';
run;
```

Summary of the continuous variables

Code:

```
proc means data = PRJT.ORGANICS2 nmiss N min max mean std;
var DemAffl DemAge PromSpend PromTime;
run;
```



The MEANS Procedure

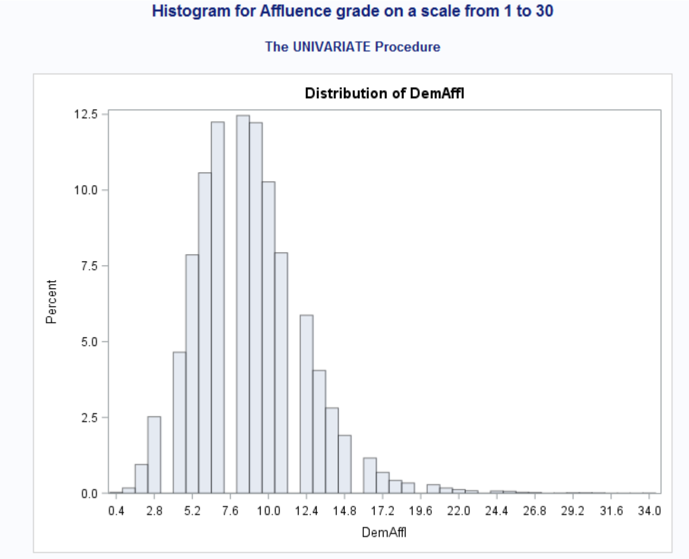
Variable	N Miss	N	Minimum	Maximum	Mean	Std Dev
DemAffl	1085	21137	0	34.0000000	8.7118323	3.4211941
DemAge	1509	20713	18.0000000	79.0000000	53.7968908	13.2057808
PromSpend	0	22222	0.0100000	296313.85	4420.25	7559.05
PromTime	281	21941	0	39.0000000	6.5646051	4.6572087

Note that although the dataset contains 22222 cases, three of the variables have fewer than given observations. We have 1085 missing information for ‘DemAffl’ variable, 1509 missing for DemAge and 281 missing information for PromTime.

Histogram, Extreme and Missing Values

Code:

```
/*The sas code for checking the extreme and missing values*/

ODS SELECT EXTREMEVALUES;
ODS select MissingValues;


PROC UNIVARIATE Data=PRJT.ORGANICS2 NEXTRVAL=10;
VAR DemAffl DemAge PromSpend PromTime;
title 'Extreme and Missing values for Continuous Variables';
RUN;
```



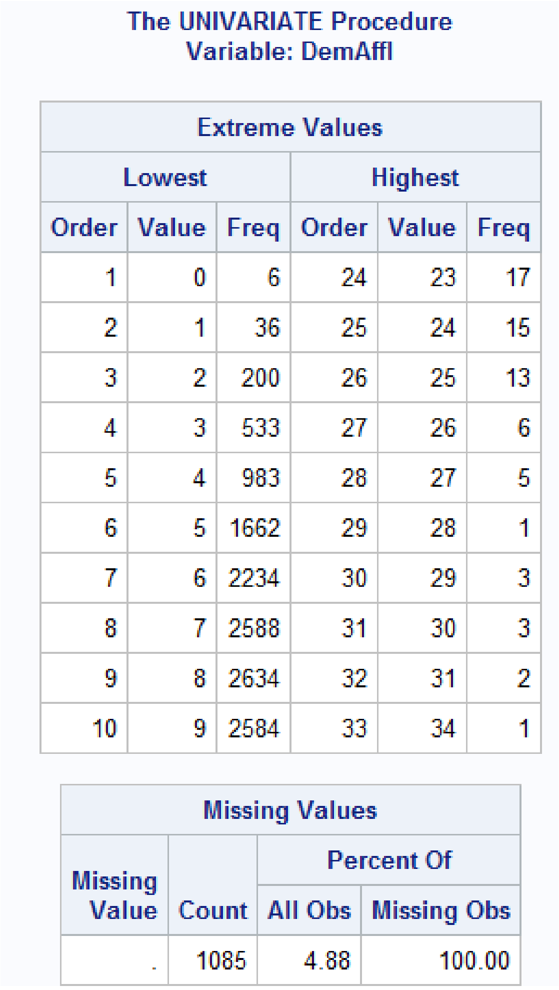
Histogram for Affluence grade on a scale from 1 to 30

The UNIVARIATE Procedure

The figure above is the histogram for affluence grade on a scale from 1 to 30. The histogram is right skewed with a peak left of the center. We have more than 12% data within a scale range of 6~10 scale.

There are approximately no percentage of data for scale greater than 26. Also the figure suggests that there are some missing values in the data set.

Also, we have gathered an idea about extreme and missing values for Affluence grade (scale from 1 to 30).



The UNIVARIATE Procedure
Variable: DemAffl

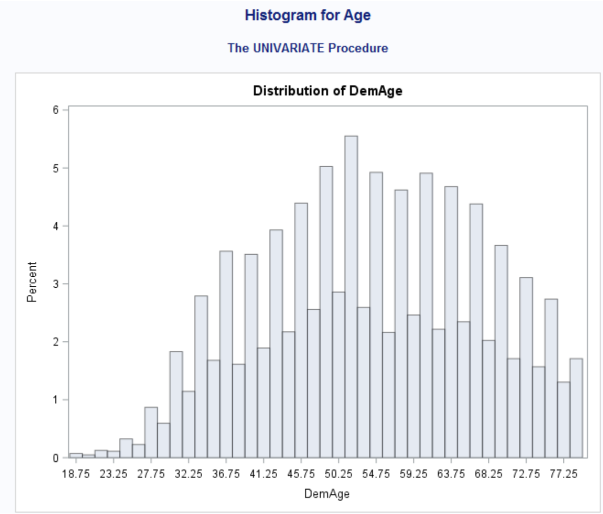
Extreme Values

Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	6	24	23	17
2	1	36	25	24	15
3	2	200	26	25	13
4	3	533	27	26	6
5	4	983	28	27	5
6	5	1662	29	28	1
7	6	2234	30	29	3
8	7	2588	31	30	3
9	8	2634	32	31	2
10	9	2584	33	34	1

Missing Values

Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	1085	4.88	100.00

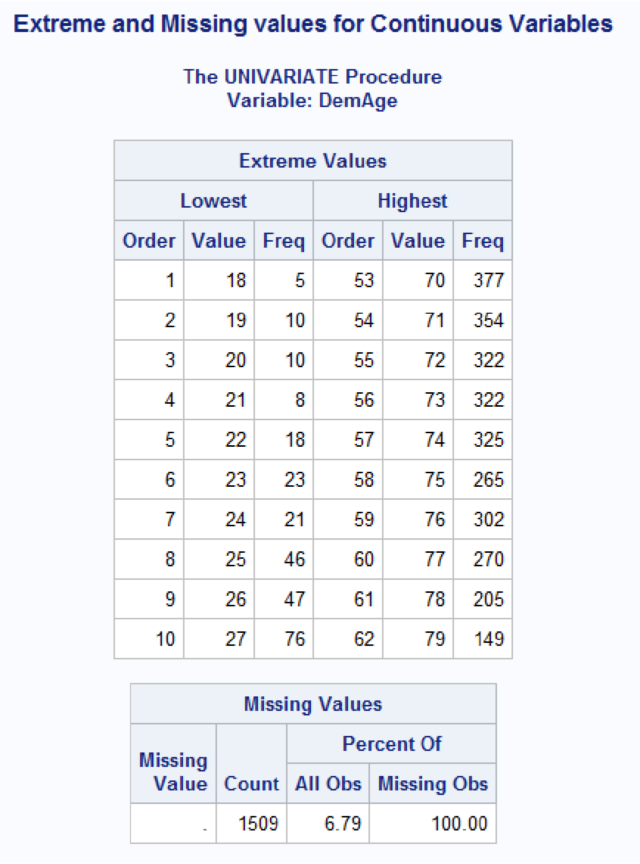
Based on our SAS code, we have also found some extreme values. We have 1085 missing values for the variable named ‘DemAffl’. With further analysis, we can also say that 4.88% is missing from our given data set for Affluence Grade.



Histogram for Age

The UNIVARIATE Procedure

In the following histogram of customer Age, about 5.5% percent of the customer are around 50 years old. Young customer with an age between 18 to 28 are less than 1%. We can tell that data is normally distributed if we can ignore 1% young customer within age of 18~28. Here, data shape is close enough to be called symmetric.



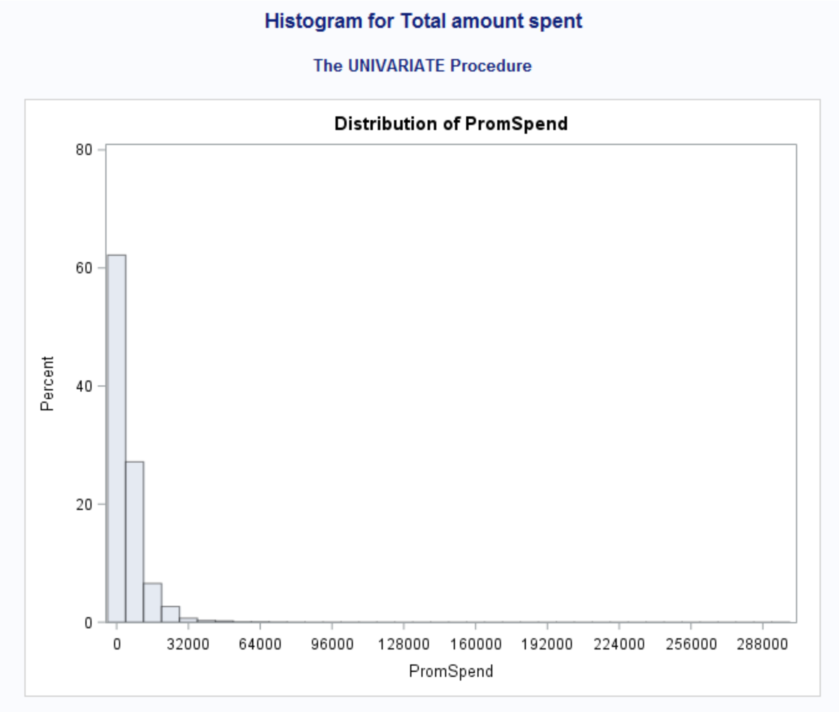
Extreme and Missing values for Continuous Variables

The UNIVARIATE Procedure
Variable: DemAge

Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	18	5	53	70	377
2	19	10	54	71	354
3	20	10	55	72	322
4	21	8	56	73	322
5	22	18	57	74	325
6	23	23	58	75	265
7	24	21	59	76	302
8	25	46	60	77	270
9	26	47	61	78	205
10	27	76	62	79	149

Missing Values			
		Percent Of	
Missing Value	Count	All Obs	Missing Obs
.	1509	6.79	100.00

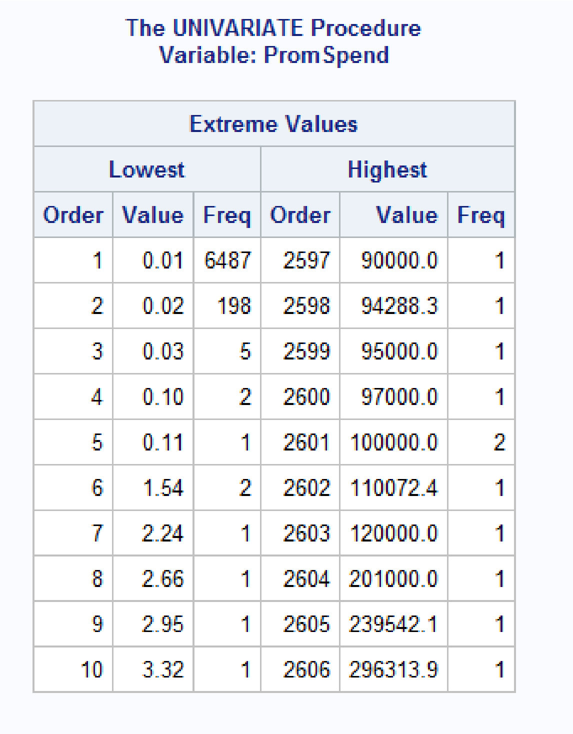
We can see for the above table that, we have more customers with a frequency greater than 300 from 70~74 age cycle. We have fewer customers with a range less than 10 for age group 18~21. Out of all observations we have 6.79% missing age.



Histogram for Total amount spent

The UNIVARIATE Procedure

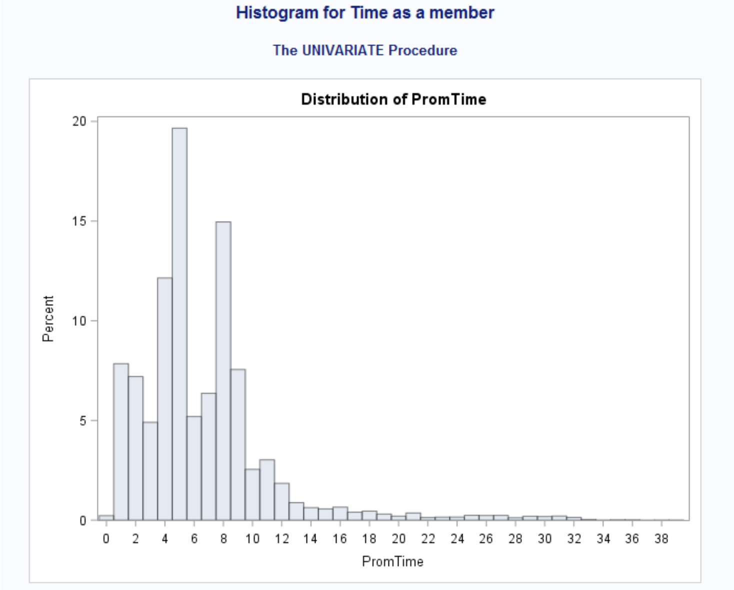
The data in the following graph are right-skewed. Most of the sample values are clustered on the right side of the histogram. The mean value was 4420.25 with minimum of 0.01 and maximum of 296313.85. We don’t have any missing values for this variable.



The UNIVARIATE Procedure
Variable: PromSpend

Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0.01	6487	2597	90000.0	1
2	0.02	198	2598	94288.3	1
3	0.03	5	2599	95000.0	1
4	0.10	2	2600	97000.0	1
5	0.11	1	2601	100000.0	2
6	1.54	2	2602	110072.4	1
7	2.24	1	2603	120000.0	1
8	2.66	1	2604	201000.0	1
9	2.95	1	2605	239542.1	1
10	3.32	1	2606	296313.9	1

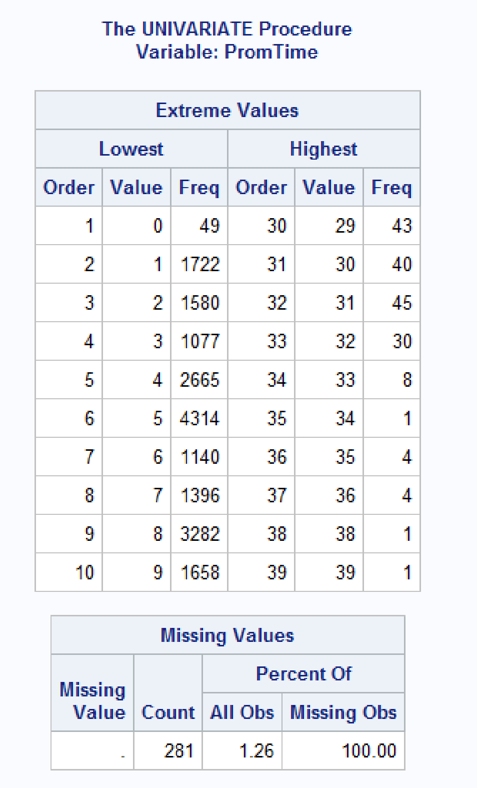
As histogram is left skewed we looked at the extreme values. We have found maximum frequency within a value range 0.01~0.02 which is not that much significant for our future analysis.



Histogram for Time as a member

The UNIVARIATE Procedure

The data in the following graph are right-skewed. Most of the sample values are clustered on the right side of the histogram.



The UNIVARIATE Procedure
Variable: PromTime

Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	49	30	29	43
2	1	1722	31	30	40
3	2	1580	32	31	45
4	3	1077	33	32	30
5	4	2665	34	33	8
6	5	4314	35	34	1
7	6	1140	36	35	4
8	7	1396	37	36	4
9	8	3282	38	38	1
10	9	1658	39	39	1

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	281	1.26	100.00

There are 281 missing values in this variable. We have seen from the histogram above that, we have more no of people within a time range between 1~9. We have extremely fewer number for a bigger time range.

Categorical Variable Analysis

For each categorical variable, you need to use proc freq to find out how many different categories and how many missing each categorical/nominal variable (except ID) has.

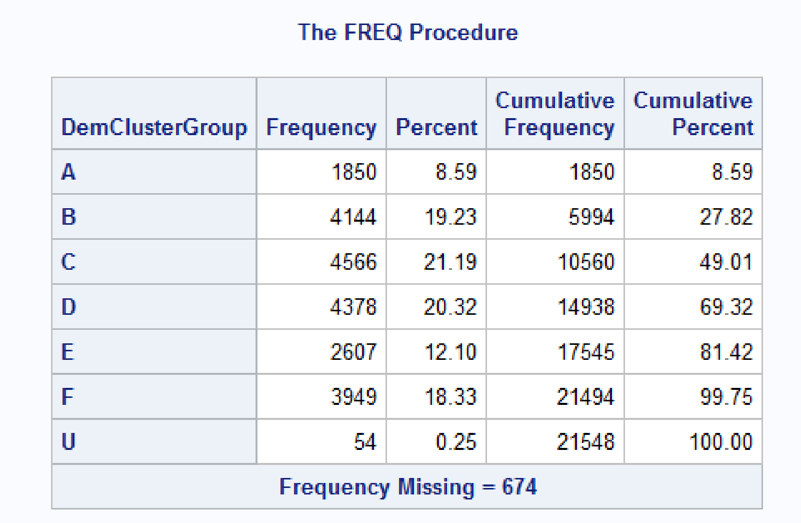
```
PROC freq Data=target.organics;
table demreg;

RUN;
```

Answer:

Code:

```
/*Categorical Variable Analysis*/

PROC freq Data=PRJT.ORGANICS2;
table DemClusterGroup DemGender DemReg DemTVReg PromClass;
title 'Categorical Variable Analysis';
RUN;
```

Frequency Table for Neighborhood group

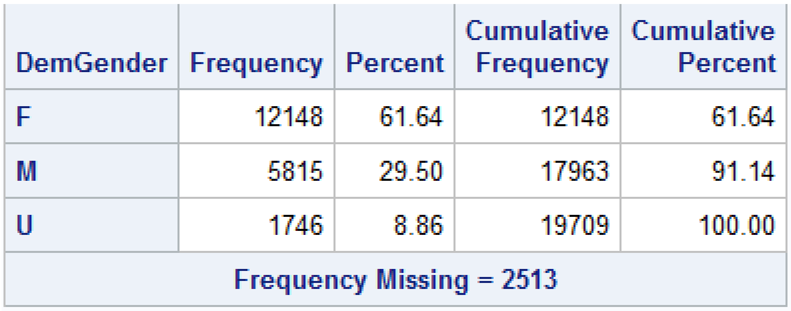


The FREQ Procedure

DemClusterGroup	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	1850	8.59	1850	8.59
B	4144	19.23	5994	27.82
C	4566	21.19	10560	49.01
D	4378	20.32	14938	69.32
E	2607	12.10	17545	81.42
F	3949	18.33	21494	99.75
U	54	0.25	21548	100.00
Frequency Missing = 674				

From this frequency table we can see that there are 674 missing values. There are 7 cluster groups in this data set. Group C has highest number of customers which is 21.19% of the total data set.

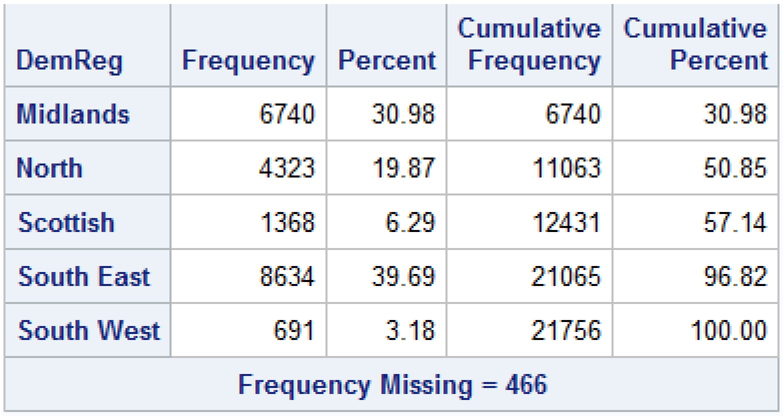
Frequency Table for Gender Description



DemGender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	12148	61.64	12148	61.64
M	5815	29.50	17963	91.14
U	1746	8.86	19709	100.00
Frequency Missing = 2513				

From this frequency table we can see that there are 2513 missing values. There are 3 groups in this data set. Among all, 61.64% are female, 29.50% are male and 8.86% are unknown.

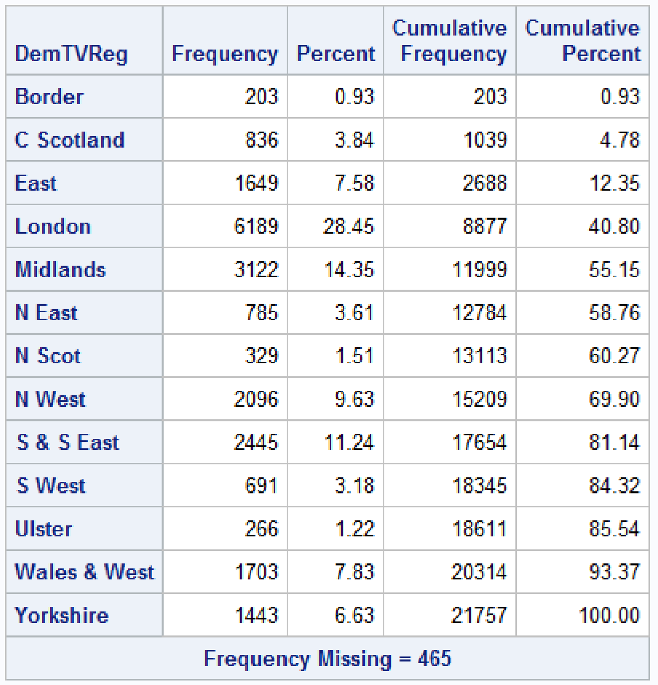
Frequency Table for Geographical Region



DemReg	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Midlands	6740	30.98	6740	30.98
North	4323	19.87	11063	50.85
Scottish	1368	6.29	12431	57.14
South East	8634	39.69	21065	96.82
South West	691	3.18	21756	100.00
Frequency Missing = 466				

From this frequency table we can see that there are 466 missing values. There are 5 groups in this data set. Based on our frequency table, we can say that 30.98% customers are from Midlands region, 19.87% from North region, 6.29% are Scottish, 39.69% are from South East region and 3.18% are from South West region.

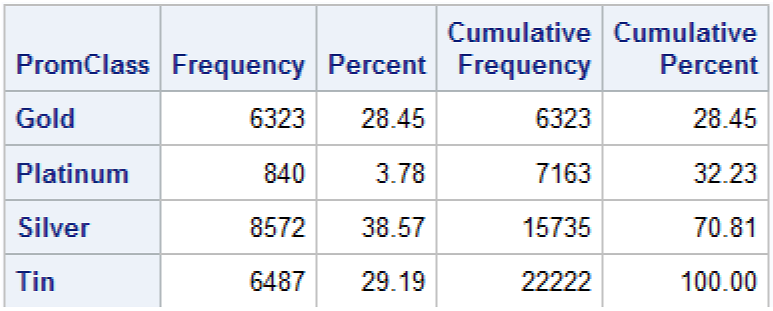
Frequency Table for TV Region



DemTVReg	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Border	203	0.93	203	0.93
C Scotland	836	3.84	1039	4.78
East	1649	7.58	2688	12.35
London	6189	28.45	8877	40.80
Midlands	3122	14.35	11999	55.15
N East	785	3.61	12784	58.76
N Scot	329	1.51	13113	60.27
N West	2096	9.63	15209	69.90
S & S East	2445	11.24	17654	81.14
S West	691	3.18	18345	84.32
Ulster	266	1.22	18611	85.54
Wales & West	1703	7.83	20314	93.37
Yorkshire	1443	6.63	21757	100.00
Frequency Missing = 465				

From this frequency table we can see that there are 465 missing values. There are 13 TV regions in this data set. Based on our frequency table, we can say that London TV region has the highest number (28.45%) of customers and Border TV region (0.93%) has the lowest number of customers.

Frequency Table for Loyalty Status



PromClass	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Gold	6323	28.45	6323	28.45
Platinum	840	3.78	7163	32.23
Silver	8572	38.57	15735	70.81
Tin	6487	29.19	22222	100.00

From our frequency table, we can clearly see that there are 4 groups based on loyalty status. Among all the customers based on loyalty status, 28.45% are Gold member, 3.78% are Platinum members, 38.57% are Silver Member and 29.19% are Tin member.

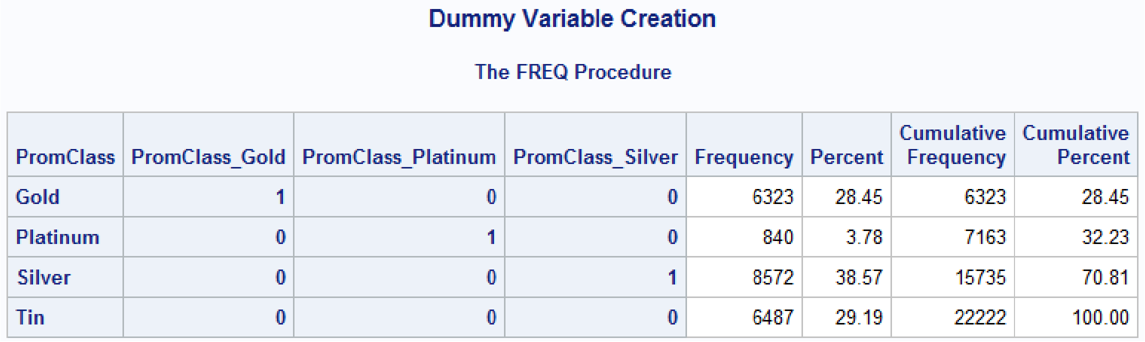
STEP 3: As I discussed in one of the lecture recordings, you are not required to do categorical/nominal variable dummy coding if you use SAS, but in this step, I will ask you to write SAS code for creating dummy variables. Please create dummy variables for the variable PromClass **(**Please remember you need to create k-1 dummies. PromClass include 4 categories; you hence need to create 3 dummy variables). You may choose to drop the variable "PromClass" after you create dummy variables for it. If you choose to keep the original variable, please remember not to include it when you fit your models. You don't need to do dummy coding for the other categorical/nominal variables.

Answer:

We have already seen that there are 4 groups based on loyalty status. They are called Gold member, Platinum members, Silver Member and Tin member. If a categorical variable has c levels/categories, the SAS output will include estimates of the regression coefficients for the (c-1) single level/category. Here there are 4 categories, we will create 3 dummy variables.

Code:

```
/* defining macro to create dummy variables */
%macro DummyVars(DSIn,    /* the name of the input data set */
                 VarList, /* the names of the categorical variables */
                 DSOut);  /* the name of the output data set */
   /* 1. add a fake response variable */
   data AddFakeY / view=AddFakeY;
      set &DSIn;
      _Y = 0;      /* add a fake response variable */
   run;
   /* 2. Create the design matrix. Include the original variables, if desired */
   proc glmselect data=AddFakeY NOPRINT outdesign(addinputvars)=&DSOut(drop=_Y);
      class      &VarList;
      model _Y = &VarList /  noint selection=none;
   run;
%mend;

/* Using macro to have desired dummy variables */

%DummyVars(PRJT.ORGANICS2, PROMCLASS, PROMCLASSDUMMY);


PROC FREQ DATA=PROMCLASSDUMMY;
TABLES PROMCLASS*PROMCLASS_GOLD*PROMCLASS_PLATINUM*PROMCLASS_SILVER / LIST;
RUN;
```



Dummy Variable Creation

The FREQ Procedure

PromClass	PromClass_Gold	PromClass_Platinum	PromClass_Silver	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Gold	1	0	0	6323	28.45	6323	28.45
Platinum	0	1	0	840	3.78	7163	32.23
Silver	0	0	1	8572	38.57	15735	70.81
Tin	0	0	0	6487	29.19	22222	100.00

Here, from the table depicted above we can clearly say that there are 3 dummy variables created. For Gold it is like 1 0 0 and for platinum 0 1 0 and for Silver 0 0 1. The rest can be defined as 0 0 0 which is classified as Tin.

STEP 4. You need to do missing value imputation. You want to replaced missing values for the interval (continuous) inputs with the input median (You need to use a SAS procedure called **proc means** to compute the median of the variable), and added unique imputation indicators for each input with missing values. For a categorical variable with missing values, you create a separate category for the missing values.

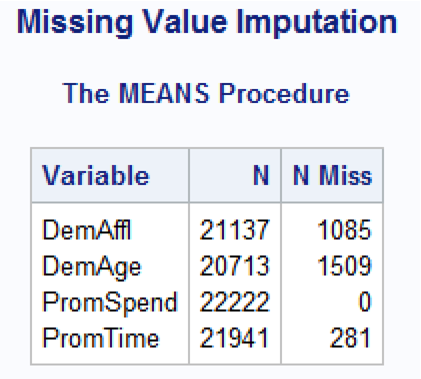
Answer:

To find the number of missing values for numeric variables we will use PROC MEANS procedure which creates a compact table that summarizes the number of missing values for each numerical variables for dataset "organics". The following statements use the N and NMISS options in the PROC MEANS statement to count the number of missing values in eight numerical variables in the PRJT.ORGANICS2 data set.

Code:

```
TITLE 'Missing Value Imputation';

/* Count missing values for numeric variables */
proc means data=PRJT.ORGANICS2 nolabels N NMISS;
var DemAffl DemAge PromSpend PromTime;
run;
```

Missing value Table for Continuous Variables



Missing Value Imputation

The MEANS Procedure

Variable	N	N Miss
DemAffl	21137	1085
DemAge	20713	1509
PromSpend	22222	0
PromTime	21941	281

The NMISS column in the table shows the number of missing values for each variable. There are 22,222 observations in the data set. Two variables have zero missing values, and another three have missing values. The DemAffl variable has 1085 missing values whereas the DemAge variable has 1509 missing values. PromTime has only 281 missing values.

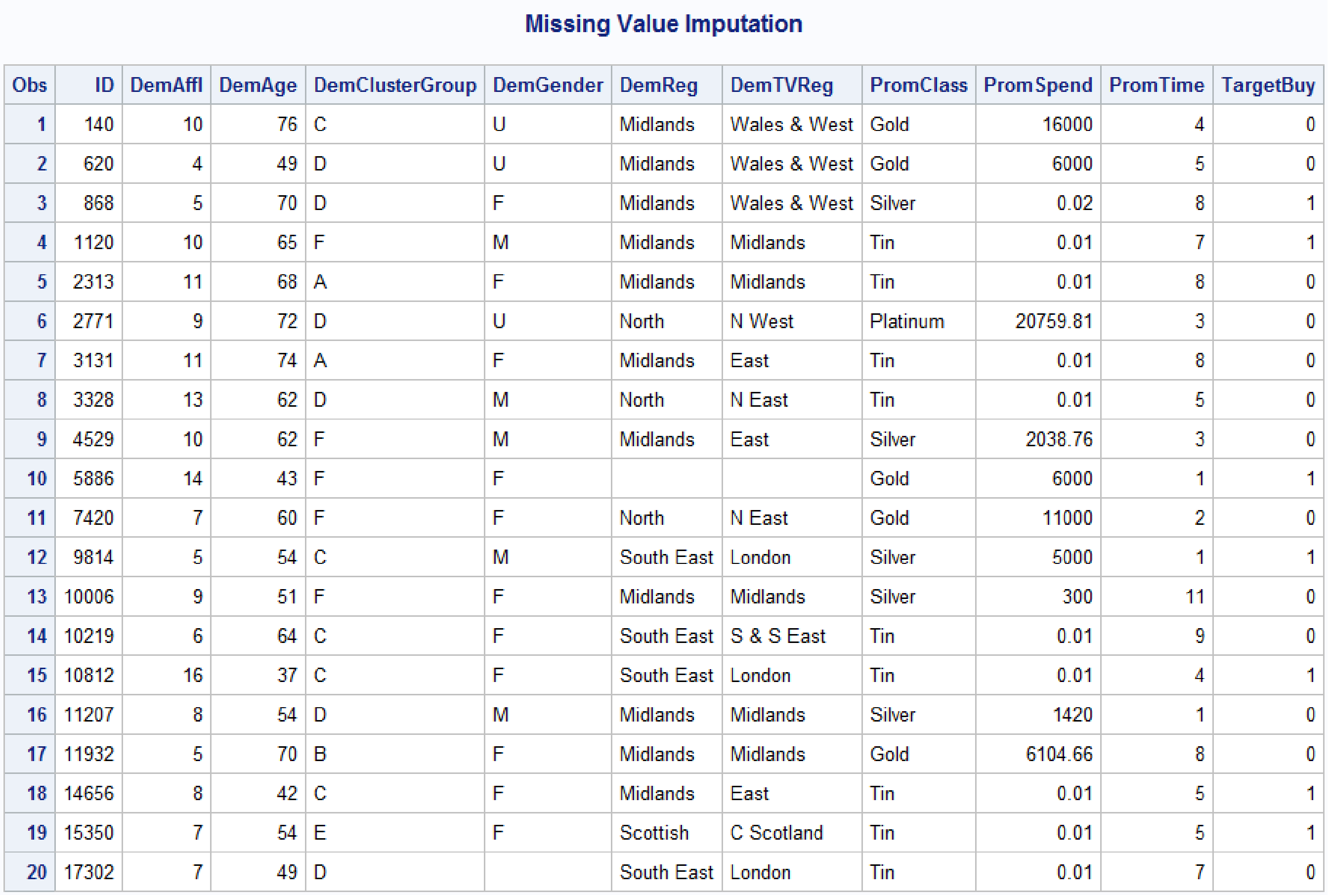
Median Imputation

Imputation of the missing value by either the mean, median or mode for the attribute are commonly used imputations. Both mean and median imputation can only be used on continuous attributes. For categorical data, the mode is often imputed whilst using either mean or median imputation.

Code:

```
/* Median imputation: Used PROC STDIZE to replace missing values with median */

proc stdize data= PRJT.ORGANICS2 out= PRJT.IMPUTED method= median reponly;
var DemAffl DemAge PromTime ;
run;

proc print data=PRJT.IMPUTED(obs = 20);  /*First 20 Observations with imputed
missing values */

run;
```

Output:



Missing Value Imputation

Obs	ID	DemAffl	DemAge	DemClusterGroup	DemGender	DemReg	DemTVReg	PromClass	PromSpend	PromTime	TargetBuy
1	140	10	76	C	U	Midlands	Wales & West	Gold	16000	4	0
2	620	4	49	D	U	Midlands	Wales & West	Gold	6000	5	0
3	868	5	70	D	F	Midlands	Wales & West	Silver	0.02	8	1
4	1120	10	65	F	M	Midlands	Midlands	Tin	0.01	7	1
5	2313	11	68	A	F	Midlands	Midlands	Tin	0.01	8	0
6	2771	9	72	D	U	North	N West	Platinum	20759.81	3	0
7	3131	11	74	A	F	Midlands	East	Tin	0.01	8	0
8	3328	13	62	D	M	North	N East	Tin	0.01	5	0
9	4529	10	62	F	M	Midlands	East	Silver	2038.76	3	0
10	5886	14	43	F	F			Gold	6000	1	1
11	7420	7	60	F	F	North	N East	Gold	11000	2	0
12	9814	5	54	C	M	South East	London	Silver	5000	1	1
13	10006	9	51	F	F	Midlands	Midlands	Silver	300	11	0
14	10219	6	64	C	F	South East	S & S East	Tin	0.01	9	0
15	10812	16	37	C	F	South East	London	Tin	0.01	4	1
16	11207	8	54	D	M	Midlands	Midlands	Silver	1420	1	0
17	11932	5	70	B	F	Midlands	Midlands	Gold	6104.66	8	0
18	14656	8	42	C	F	Midlands	East	Tin	0.01	5	1
19	15350	7	54	E	F	Scottish	C Scotland	Tin	0.01	5	1
20	17302	7	49	D		South East	London	Tin	0.01	7	0

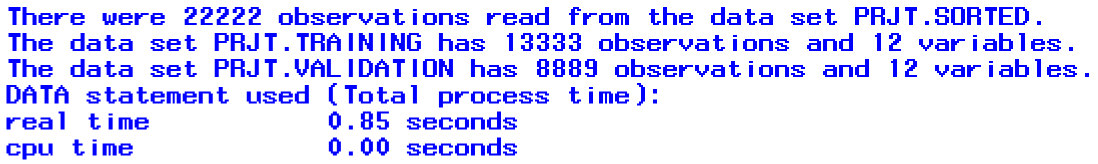
In the original data set, at 19th row, DemAffl data was missing and at 20th row, DemAge data was missing. With Median imputation I got rid of missing values. The first 20 observations were given above to justify my answer.

STEP 5: You need to randomly select 60% of the data for training and 40% for validation. Please refer to my SAS tutorial slides to find out how to create a random sample.

Answer:

```
TITLE 'Step 5: Random Training and Validation Data Set';

data PRJT.SORTED;
set PRJT.IMPUTED;
n=ranuni(20041206);

proc sort data=PRJT.SORTED; by n;

data PRJT.TRAINING PRJT.VALIDATION; /*Training and Validation Data*/
set PRJT.SORTED nobs=nobs;
if _n_<=.60*nobs then output PRJT.TRAINING;
else output PRJT.VALIDATION;

run;
```

Output:



```
There were 22222 observations read from the data set PRJT.SORTED.
The data set PRJT.TRAINING has 13333 observations and 12 variables.
The data set PRJT.VALIDATION has 8889 observations and 12 variables.
DATA statement used (Total process time):
real time           0.85 seconds
cpu time            0.00 seconds
```

Using the above code, I was able to successfully generate 60% Training Data and 40% Validation Data. Among, 22,222 observations 13,333 observations were assigned as Training data and 8,889 observations were assigned as Validation data.

STEP 6: You need to use the stepwise logistic regression method for variable selection and tell me what variables have been selected.

Answer:

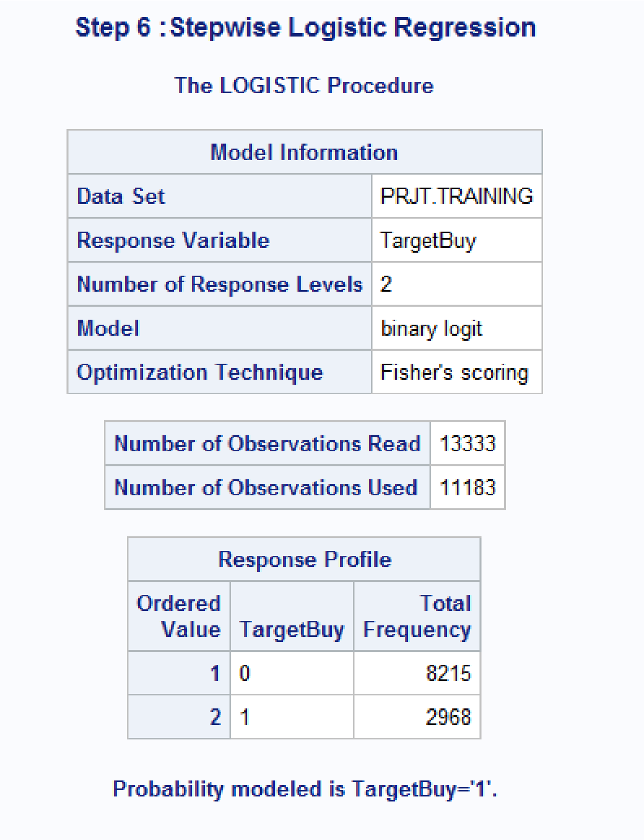
- The stepwise selection process consists of a series of alternating forward selection and backward elimination steps. The former adds variables to the model, while the latter removes variables from the model.
- The following invocation of PROC LOGISTIC illustrates the use of stepwise selection to identify the prognostic factors for **targetBuy** (Organics purchased? 1 = yes 0 = no). A significance level of 0.3 is required to allow a variable into the model (SLENTRY=0.3), and a significance level of 0.35 is required for a variable to stay in the model (SLSTAY=0.35).
- The OUTEST= and COVOUT options in the PROC LOGISTIC statement create a data set that contains parameter estimates and their covariances for the final selected model.
- The response variable option EVENT= chooses **targetBuy** =1 (Organics purchased) as the event so that the probability of targetBuy is modeled.

Code:

```
TITLE 'Step 6 :Stepwise Logistic Regression';

proc logistic data=PRJT.TRAINING outest=PRJT.TRAINING_REG covout;
class DemClusterGroup DemGender DemReg DemTVReg PromClass;
model targetBuy(event='1')=DemAffl DemAge DemClusterGroup DemGender DemReg DemTVReg
PromClass PromSpend PromTime
                / selection=stepwise  /*Stepwise selection process*/
                  slentry=0.3
                  slstay=0.35 details;

run;
```

Output:

Before adding a relevant variable to the model, stepwise selection attempts to remove any insignificant variables from the model. In the presented output, each addition or deletion of a variable to or from a model is indicated as a distinct step, and a new model is fitted at each step.



Step 6 : Stepwise Logistic Regression

The LOGISTIC Procedure

Model Information	
Data Set	PRJT.TRAINING
Response Variable	TargetBuy
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	13333
Number of Observations Used	11183

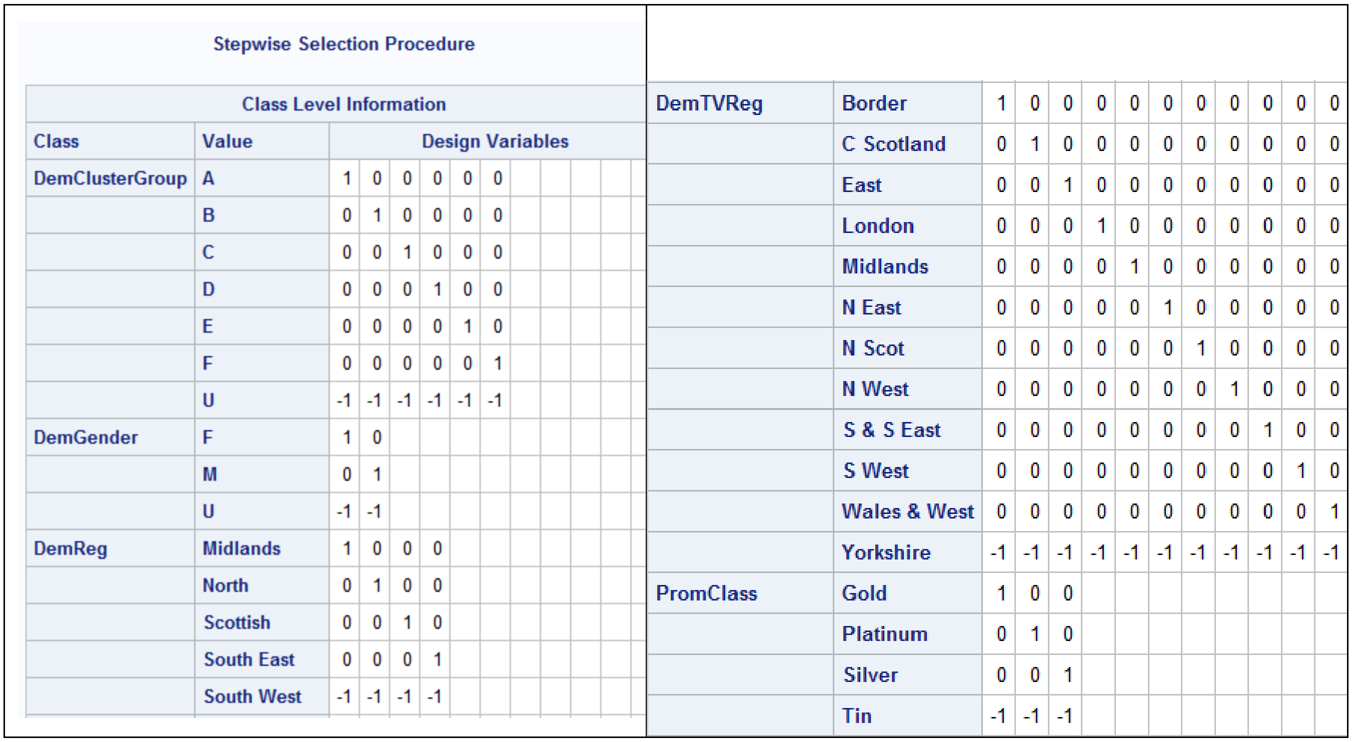
Response Profile

Ordered Value	TargetBuy	Total Frequency
1	0	8215
2	1	2968

Probability modeled is TargetBuy='1'.

From the above table, we can have an idea about our used training data set for our stepwise regression. Out of 13,333 observations, 11,183 observations were used in regression process. If we look at the response profile, we can see that 2968 customers used to buy organic products and 8215 didn't have any tendency to buy organic products. Other 2150 observations were deleted due to missing values for the response or explanatory variables.

Stepwise Selection Procedure:

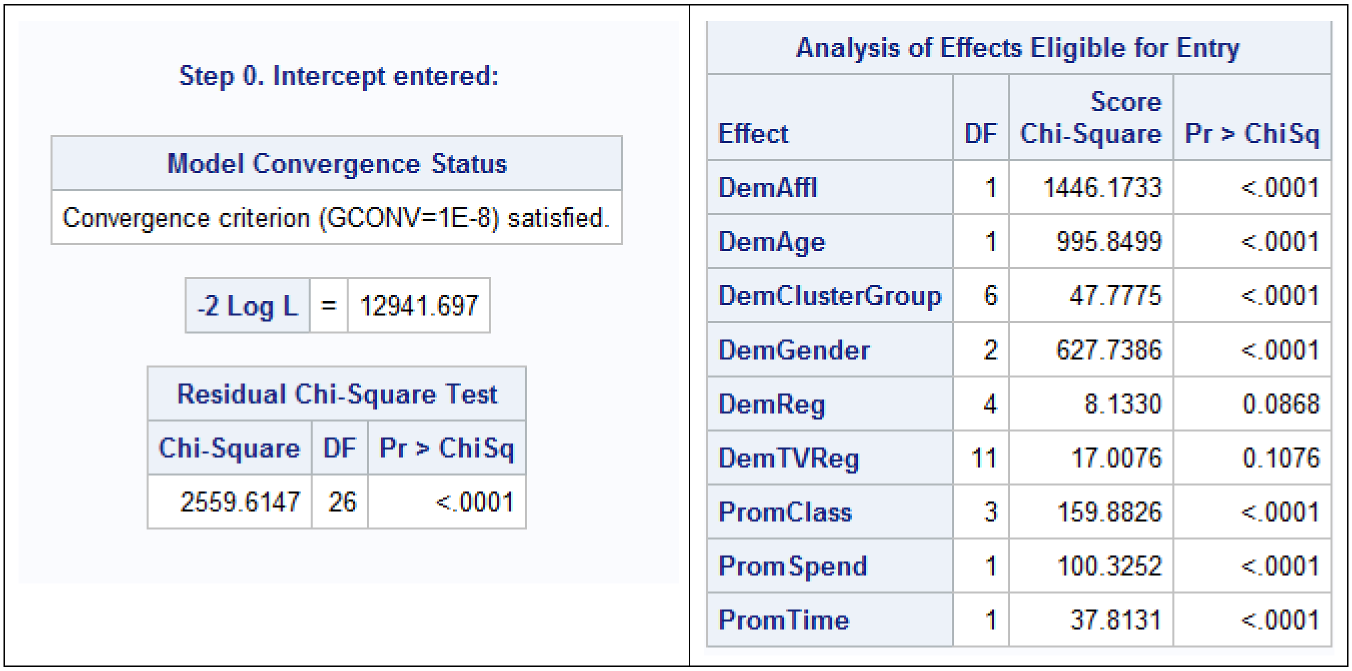


Stepwise Selection Procedure

Class Level Information

Class	Value	Design Variables										
DemClusterGroup	A	1	0	0	0	0	0					
	B	0	1	0	0	0	0					
	C	0	0	1	0	0	0					
	D	0	0	0	1	0	0					
	E	0	0	0	0	1	0					
	F	0	0	0	0	0	1					
	U	-1	-1	-1	-1	-1	-1					
DemGender	F	1	0									
	M	0	1									
	U	-1	-1									
DemReg	Midlands	1	0	0	0							
	North	0	1	0	0							
	Scottish	0	0	1	0							
	South East	0	0	0	1							
	South West	-1	-1	-1	-1							
DemTVReg	Border	1	0	0	0	0	0	0	0	0	0	0
	C Scotland	0	1	0	0	0	0	0	0	0	0	0
	East	0	0	1	0	0	0	0	0	0	0	0
	London	0	0	0	1	0	0	0	0	0	0	0
	Midlands	0	0	0	0	1	0	0	0	0	0	0
	N East	0	0	0	0	0	1	0	0	0	0	0
	N Scot	0	0	0	0	0	0	1	0	0	0	0
	N West	0	0	0	0	0	0	0	1	0	0	0
	S & S East	0	0	0	0	0	0	0	0	1	0	0
	S West	0	0	0	0	0	0	0	0	0	1	0
	Wales & West	0	0	0	0	0	0	0	0	0	0	1
	Yorkshire	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
PromClass	Gold	1	0	0								
	Platinum	0	1	0								
	Silver	0	0	1								
	Tin	-1	-1	-1								

Step 0:



Step 0. Intercept entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

-2 Log L	=	12941.697

Residual Chi-Square Test

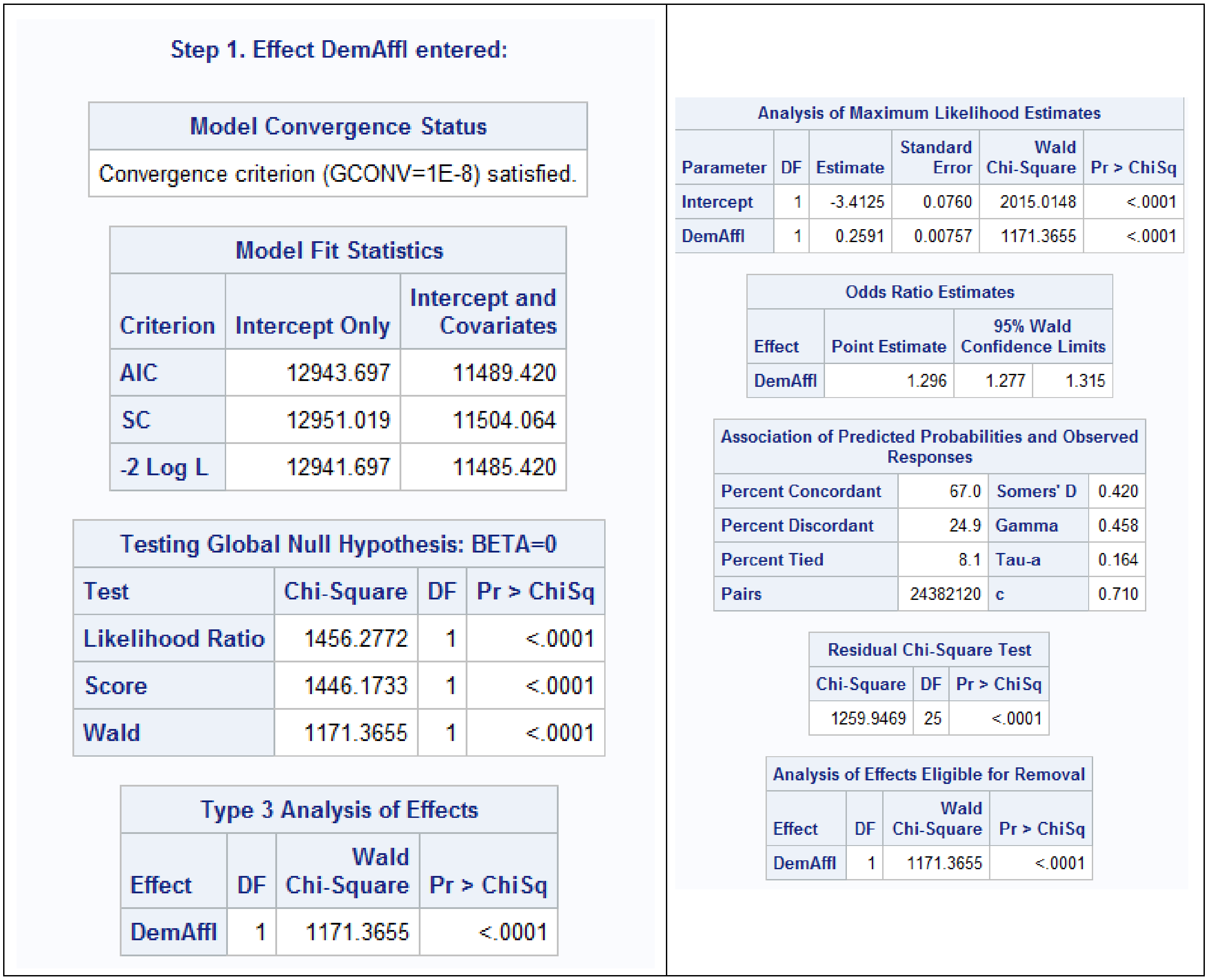
Chi-Square	DF	Pr > ChiSq
2559.6147	26	<.0001

Analysis of Effects Eligible for Entry

Effect	DF	Score Chi-Square	Pr > ChiSq
DemAffl	1	1446.1733	<.0001
DemAge	1	995.8499	<.0001
DemClusterGroup	6	47.7775	<.0001
DemGender	2	627.7386	<.0001
DemReg	4	8.1330	0.0868
DemTVReg	11	17.0076	0.1076
PromClass	3	159.8826	<.0001
PromSpend	1	100.3252	<.0001
PromTime	1	37.8131	<.0001

The intercept-only model is fit and individual score statistics for the potential variables are evaluated. Chi-square value is 2559.6147 with 26 degrees of freedom, which gives us p-value less than 0.05 which is considered to be significant.

Step 1:



Step 1. Effect DemAffl entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12943.697	11489.420
SC	12951.019	11504.064
-2 Log L	12941.697	11485.420

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1456.2772	1	<.0001
Score	1446.1733	1	<.0001
Wald	1171.3655	1	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
DemAffl	1	1171.3655	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.4125	0.0760	2015.0148	<.0001
DemAffl	1	0.2591	0.00757	1171.3655	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
DemAffl	1.296	1.277	1.315

Association of Predicted Probabilities and Observed Responses

Percent Concordant	67.0	Somers' D	0.420
Percent Discordant	24.9	Gamma	0.458
Percent Tied	8.1	Tau-a	0.164
Pairs	24382120	c	0.710

Residual Chi-Square Test

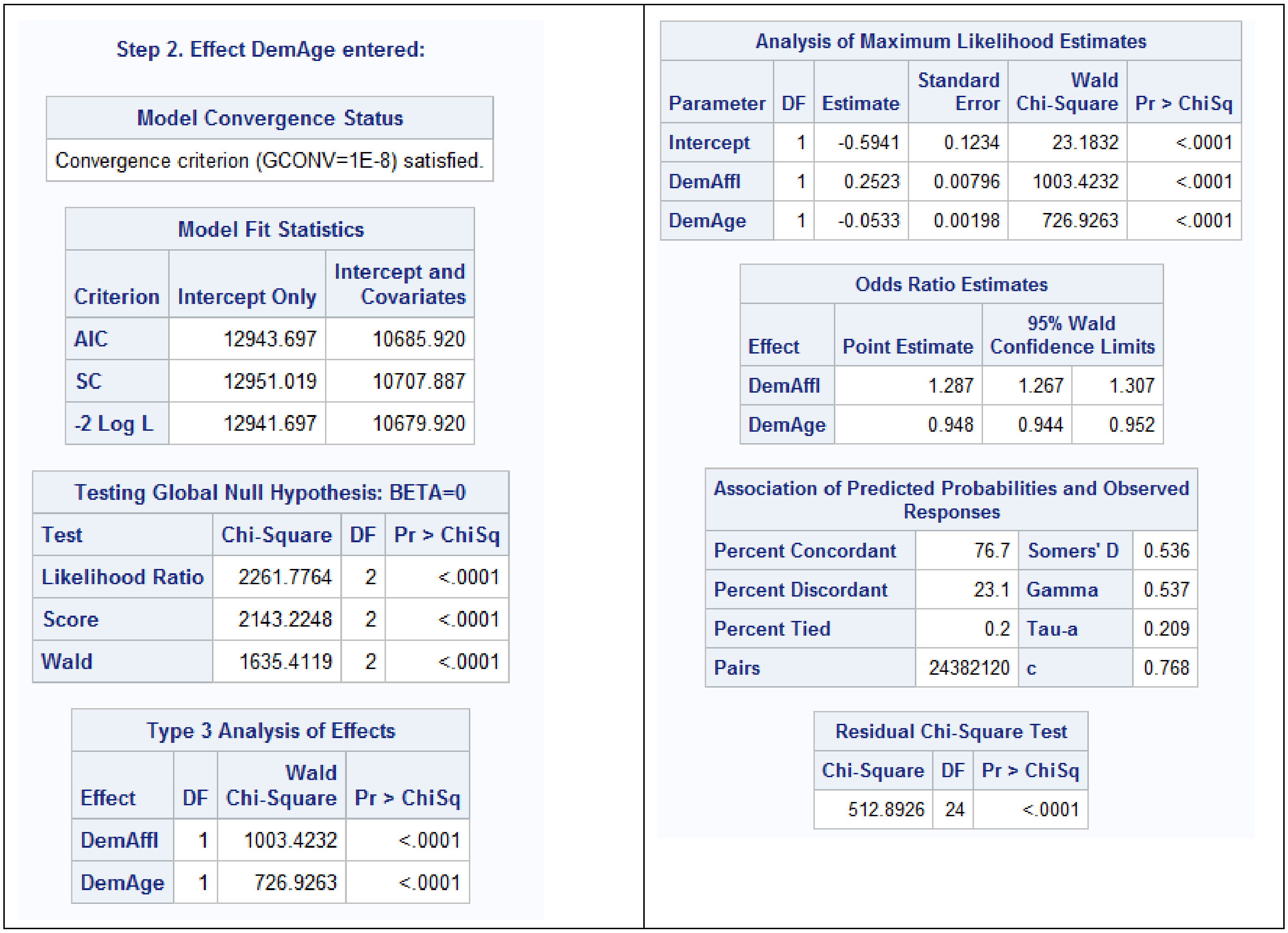
Chi-Square	DF	Pr > ChiSq
1259.9469	25	<.0001

Analysis of Effects Eligible for Removal

Effect	DF	Wald Chi-Square	Pr > ChiSq
DemAffl	1	1171.3655	<.0001

In Step 1, the variable DemAffl is selected along with the intercept into the model because it is the most significant variable. With chi-square value = 1259.9469 and 25 degrees of freedom, we ended up having a p value less than the significance level $\alpha = 0.05$. So we can reject our null hypothesis that $\beta = 0$. In our model, the variable DemAffl is significant.

Step 2:



Step 2. Effect DemAge entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12943.697	10685.920
SC	12951.019	10707.887
-2 Log L	12941.697	10679.920

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2261.7764	2	<.0001
Score	2143.2248	2	<.0001
Wald	1635.4119	2	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
DemAffl	1	1003.4232	<.0001
DemAge	1	726.9263	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.5941	0.1234	23.1832	<.0001
DemAffl	1	0.2523	0.00796	1003.4232	<.0001
DemAge	1	-0.0533	0.00198	726.9263	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
DemAffl	1.287	1.267	1.307
DemAge	0.948	0.944	0.952

Association of Predicted Probabilities and Observed Responses

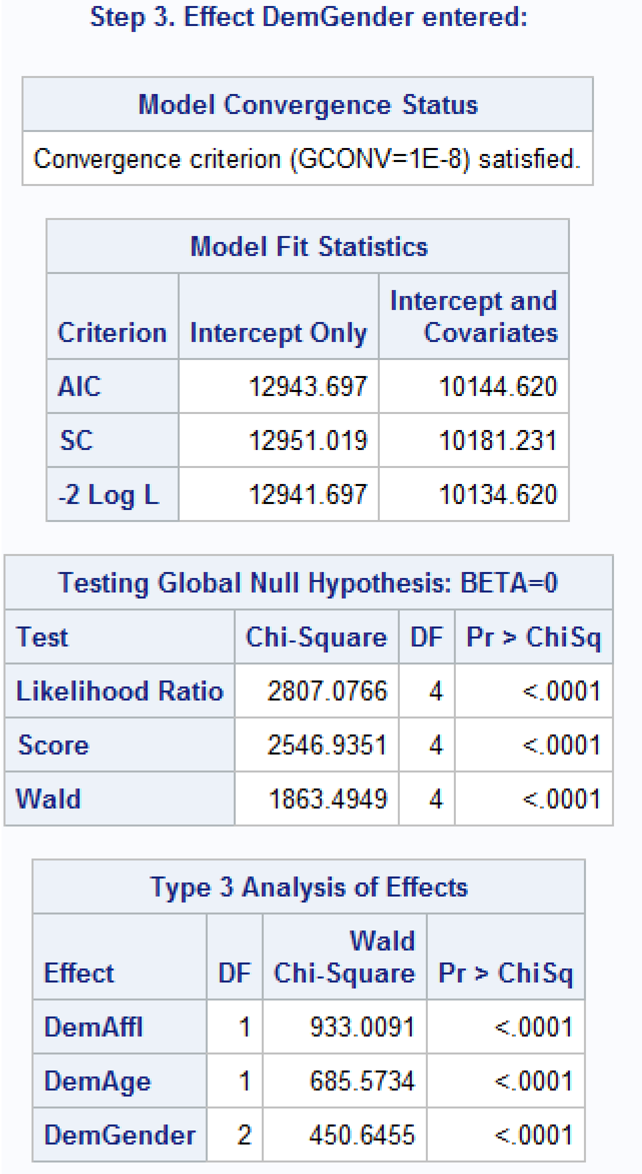
Percent Concordant	76.7	Somers' D	0.536
Percent Discordant	23.1	Gamma	0.537
Percent Tied	0.2	Tau-a	0.209
Pairs	24382120	c	0.768

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
512.8926	24	<.0001

In Step 3, the variable DemaAge is selected along with DemAffl and intercept into the model because it is the most significant variable. With chi-square value = 512.89 and 24 degrees of freedom, we ended up having a p value less than the significance level $\alpha = 0.05$. So we can reject our null hypothesis that $\beta_1 = 0$. In our model, the variable DemAge is significant.

Step 3:



Step 3. Effect DemGender entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12943.697	10144.620
SC	12951.019	10181.231
-2 Log L	12941.697	10134.620

Testing Global Null Hypothesis: BETA=0

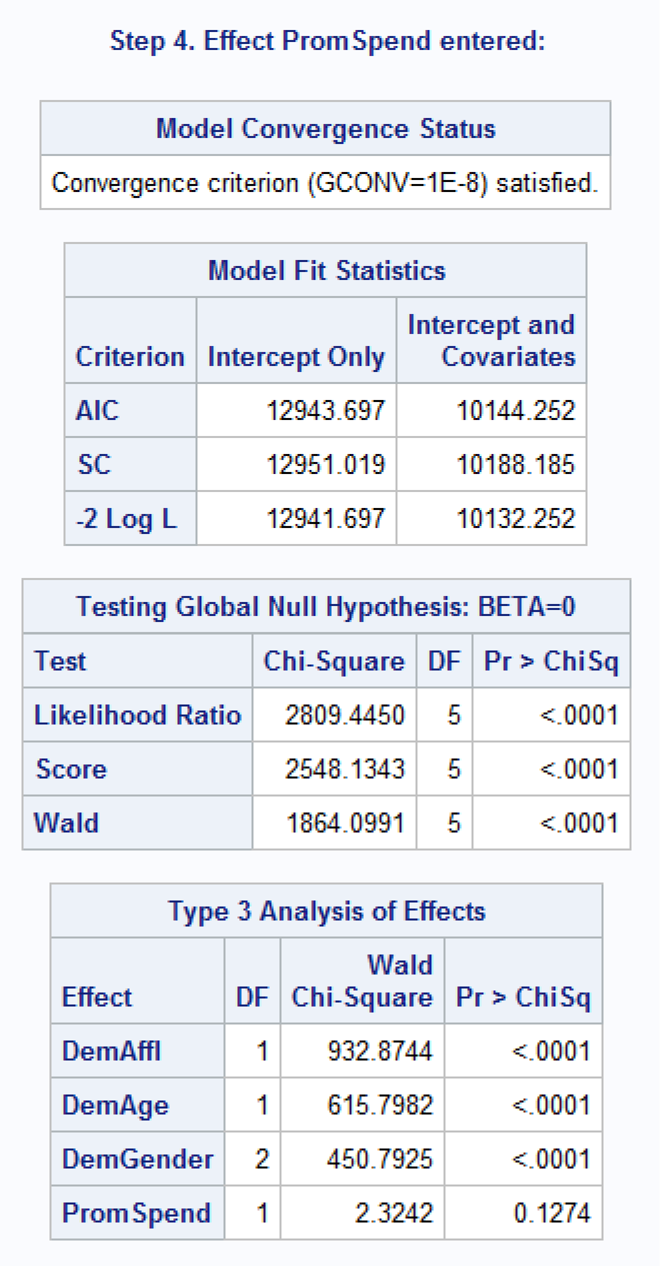
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2807.0766	4	<.0001
Score	2546.9351	4	<.0001
Wald	1863.4949	4	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
DemAffl	1	933.0091	<.0001
DemAge	1	685.5734	<.0001
DemGender	2	450.6455	<.0001

In Step 3, the variable DemGender is added along with other two variables. With chi-square value = 450.6455 and 2 degrees of freedom, we ended up having a p value less than the significance level $\alpha = 0.05$. So we can reject our null hypothesis that $\beta_2 = 0$. In our model, the variable DemGender is significant.

Step 4:



Step 4. Effect PromSpend entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12943.697	10144.252
SC	12951.019	10188.185
-2 Log L	12941.697	10132.252

Testing Global Null Hypothesis: BETA=0

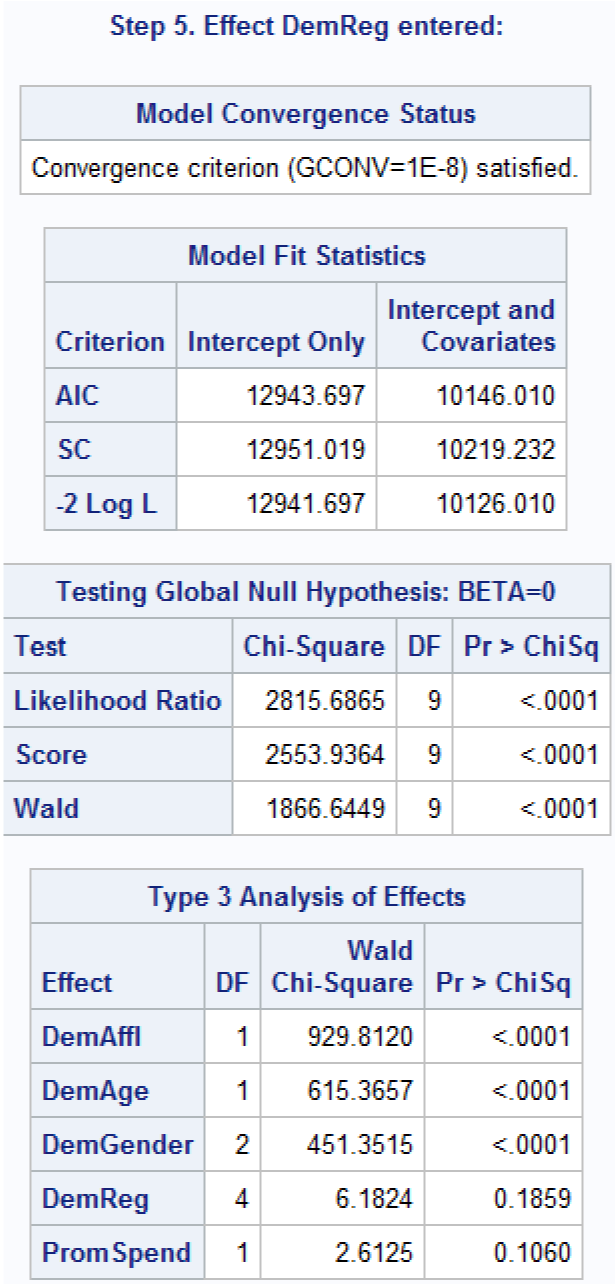
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2809.4450	5	<.0001
Score	2548.1343	5	<.0001
Wald	1864.0991	5	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
DemAffl	1	932.8744	<.0001
DemAge	1	615.7982	<.0001
DemGender	2	450.7925	<.0001
PromSpend	1	2.3242	0.1274

In Step 4, the variable PromSpend is added along with other three variables. With chi-square value = 2.32 and 1 degrees of freedom, we ended up having a p value greater than the significance level $\alpha = 0.05$. So we fail to reject our null hypothesis that $\beta_3 = 0$. In our model, the variable PromSpend is not significant.

Step 5:



Step 5. Effect DemReg entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	12943.697	10146.010
SC	12951.019	10219.232
-2 Log L	12941.697	10126.010

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2815.6865	9	<.0001
Score	2553.9364	9	<.0001
Wald	1866.6449	9	<.0001

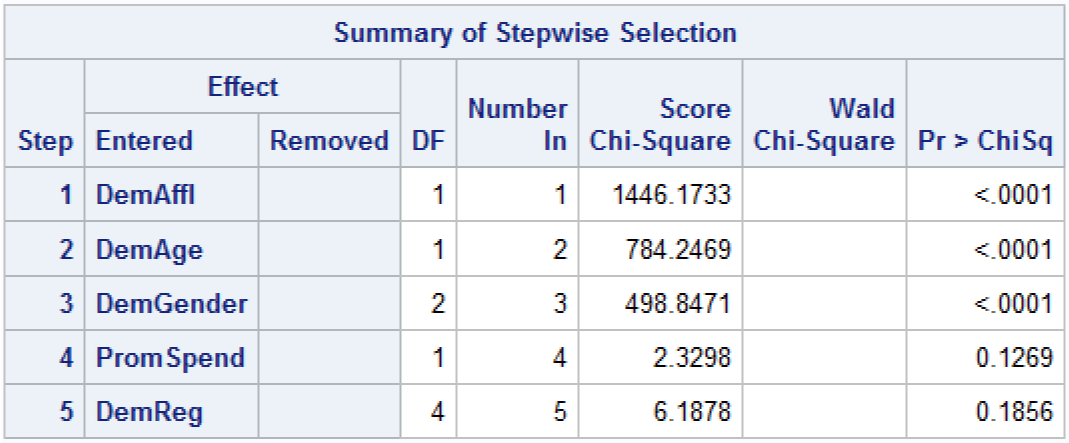
Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
DemAffl	1	929.8120	<.0001
DemAge	1	615.3657	<.0001
DemGender	2	451.3515	<.0001
DemReg	4	6.1824	0.1859
PromSpend	1	2.6125	0.1060

In Step 5, the variable DemReg is added along with other 4 variables. With chi-square value = 6.1824 and 4 degrees of freedom, we ended up having a p value greater than the significance level $\alpha = 0.05$. So we fail to reject our null hypothesis that $\beta_4 = 0$. In our model, the variable DemReg is not significant.

Summary of the Stepwise Selection

Finally, none of the remaining variables outside the model meet the entry criterion, and the stepwise selection is terminated. A summary of the stepwise selection is displayed below:



Summary of Stepwise Selection

Step	Effect Entered	 Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	DemAffl		1	1	1446.1733		<.0001
2	DemAge		1	2	784.2469		<.0001
3	DemGender		2	3	498.8471		<.0001
4	PromSpend		1	4	2.3298		0.1269
5	DemReg		4	5	6.1878		0.1856

From our stepwise summary, we ended up selecting **DemAffl, DemAge and DemGender** variables.

Exporting SAS Dataset

Task: Now you are done with variable processing, you can export your sas dataset using the following code (Now you have a training dataset and a validation dataset. Please remember to export both).

```
proc export data=target.organics
   outfile='C:\Users\jliu2188\organics.csv'//you need to specify the output file
name here.
   dbms=csv replace;

run;
```

Answer:

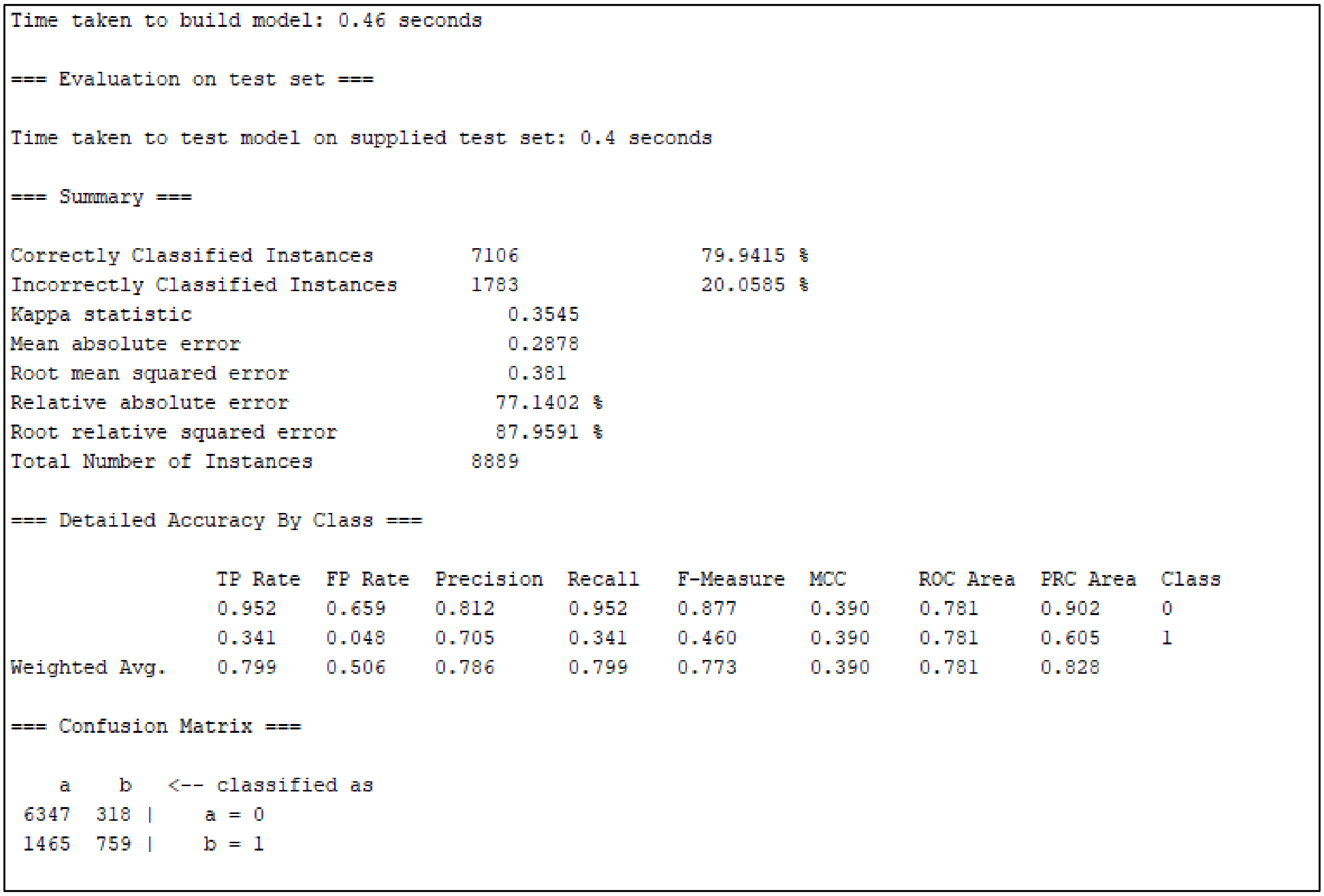
```
/*Exporting SAS training and validation data set*/
proc export data=PRJT.TRAINING
   outfile='Z:\OneDrive - South Dakota State University - SDSU\INFS 762\Project
1\organic_training.csv'
   dbms=csv replace;
run;

proc export data=PRJT.VALIDATION
   outfile='Z:\OneDrive - South Dakota State University - SDSU\INFS 762\Project
1\organic_validation.csv'
   dbms=csv replace;

run;
```

STEP 7: Based on the selected variables, you fit two additional models (e.g, neural network, random forest or SVM, but please not use a decision tree model). You need to use logistic regression with the selected variables as your baseline model and compare it with the two models you selected. Please create a table to show the precision/recall/accuracy of the three models including logistic regression and the two models you selected. You need to show the measures based on the validation dataset. When you fit the two additional models, remember to just use the variables that has been select in step 6. You can use Weka to fit your models (download: http://www.cs.waikato.ac.nz/ml/weka/downloading.html). You can watch my Weka tutorial and you can also find the Weka tutorial online such as https://www.youtube.com/watch?v=m7kpIBGEdkI. The dataset used in the tutorial is of the ARFF format. You can also load a cvs file to Weka. Please tell me which model you think is the best model for the problem.

Answer:

Model 1 : Logistic Regression Using Weka

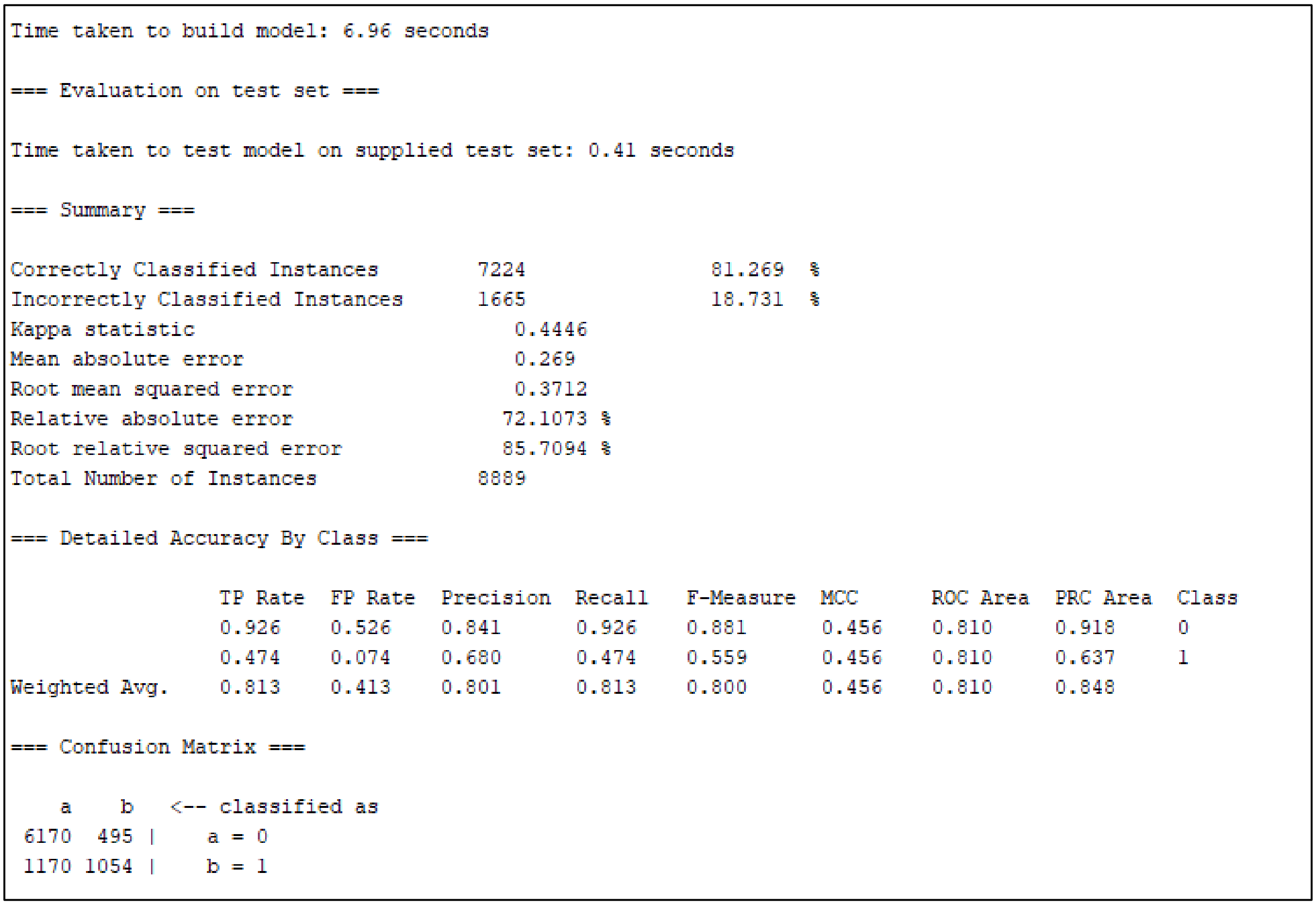
Logistic Regression Results: Using Training and Validation Data set



```
Time taken to build model: 0.46 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.4 seconds

=== Summary ===

Correctly Classified Instances        7106               79.9415 %
Incorrectly Classified Instances      1783               20.0585 %
Kappa statistic                          0.3545
Mean absolute error                      0.2878
Root mean squared error                  0.381
Relative absolute error                 77.1402 %
Root relative squared error             87.9591 %
Total Number of Instances             8889

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.952    0.659    0.812      0.952    0.877      0.390    0.781     0.902     0
                 0.341    0.048    0.705      0.341    0.460      0.390    0.781     0.605     1
Weighted Avg.    0.799    0.506    0.786      0.799    0.773      0.390    0.781     0.828

=== Confusion Matrix ===

    a    b   <-- classified as
 6347  318 |    a = 0
 1465  759 |    b = 1
```

Now with use of validation dataset, TargetBuy as Dependent variable and DemAffl, DemAge and DemGender as independent variables, we have 79.9415% overall efficiency. If we look at the precision, recall table, there are two rows defining class 0 and 1. This occurred because of the presence of the two different categories of dependent variable which are treated as positive. **For our convenience, we have selected class 0 as Positive.**

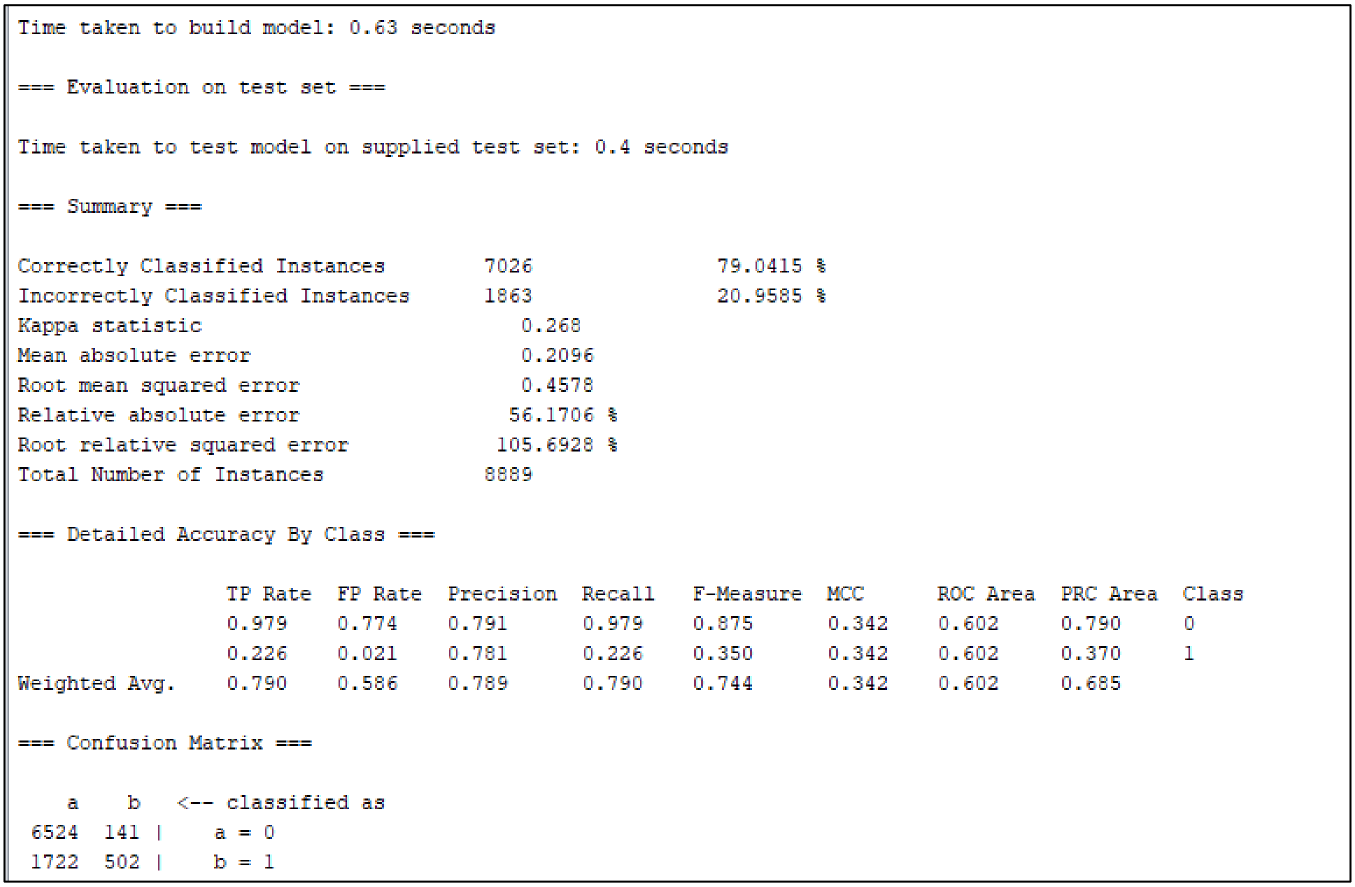
Model 2: Neural Network Using Weka



```
Time taken to build model: 6.96 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.41 seconds

=== Summary ===

Correctly Classified Instances        7224               81.269  %
Incorrectly Classified Instances      1665               18.731  %
Kappa statistic                          0.4446
Mean absolute error                      0.269
Root mean squared error                  0.3712
Relative absolute error                 72.1073 %
Root relative squared error             85.7094 %
Total Number of Instances             8889

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.926    0.526    0.841      0.926    0.881      0.456    0.810     0.918     0
                 0.474    0.074    0.680      0.474    0.559      0.456    0.810     0.637     1
Weighted Avg.    0.813    0.413    0.801      0.813    0.800      0.456    0.810     0.848

=== Confusion Matrix ===

    a    b   <-- classified as
 6170  495 |    a = 0
 1170 1054 |    b = 1
```

We have chosen MultilayerPerception for modeling a neural network. With TargetBuy as Dependent variable and DemAffl, DemAge and DemGender as independent variables, we have 81.269% overall efficiency which is a bit better than the logistic regression model.

Model 3: Stochastic Gradient Descent Using Weka



```
Time taken to build model: 0.63 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.4 seconds

=== Summary ===

Correctly Classified Instances        7026               79.0415 %
Incorrectly Classified Instances      1863               20.9585 %
Kappa statistic                          0.268
Mean absolute error                      0.2096
Root mean squared error                  0.4578
Relative absolute error                 56.1706 %
Root relative squared error            105.6928 %
Total Number of Instances             8889

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.979    0.774    0.791      0.979    0.875      0.342    0.602     0.790     0
                 0.226    0.021    0.781      0.226    0.350      0.342    0.602     0.370     1
Weighted Avg.    0.790    0.586    0.789      0.790    0.744      0.342    0.602     0.685

=== Confusion Matrix ===

    a    b   <-- classified as
 6524  141 |    a = 0
 1722  502 |    b = 1
```

We have chosen Stochastic Gradient Descent for modeling a neural network. With TargetBuy as Dependent variable and DemAffl, DemAge and DemGender as independent variables, we have 71.0415% overall efficiency.

By comparing three models based on precision, recall and accuracy we have created a table

Model	Precision	Recall	Accuracy (%)
Logistic Regression	0.812	0.952	79.94
Neural Network	0.841	0.926	81.27
Stochastic Gradient Descent	0.791	0.979	79.04

Based on accuracy we can say that Neural network gives us a good result with highest accuracy among those 3 models but recall value was less than the other two models. As we know that, recall is a good way to determine the fitness of a model. Based on recall and accuracy, I would choose logistic regression model would be a best model for this problem.

Step 8. In step 2, you have identified some variables with skewed distribution. Such distributions create high leverage points that can distort an input's association with the target. Let's now modify these variables before fitting the stepwise regression. You can take a natural log of the variables with the skewed distribution. However, logarithm of zero, log(0), is not defined. You can always do log(x+1), where x is the variable value you want to transform.

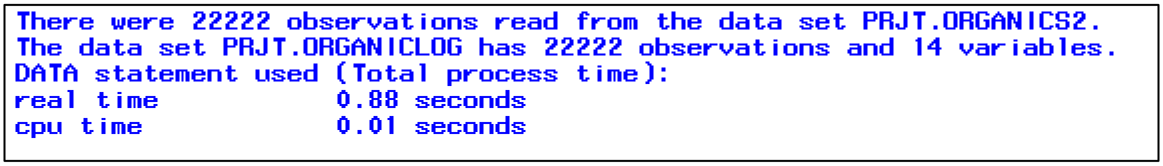
Answer:

From Step 2, we have found 3 variables which gave us right skewed histogram. We have applied log transformation for those 3 skewed variables.

Code:

```
TITLE 'Step 8 :Log Transformation of the Data';

data PRJT.ORGANICLOG;
   set PRJT.ORGANICS2;
   logDemAffl = log( DemAffl + 1 );
   logPromSpend = log( PromSpend + 1 );
   logPromTime = log( PromTime + 1 );
run;
```

Output:



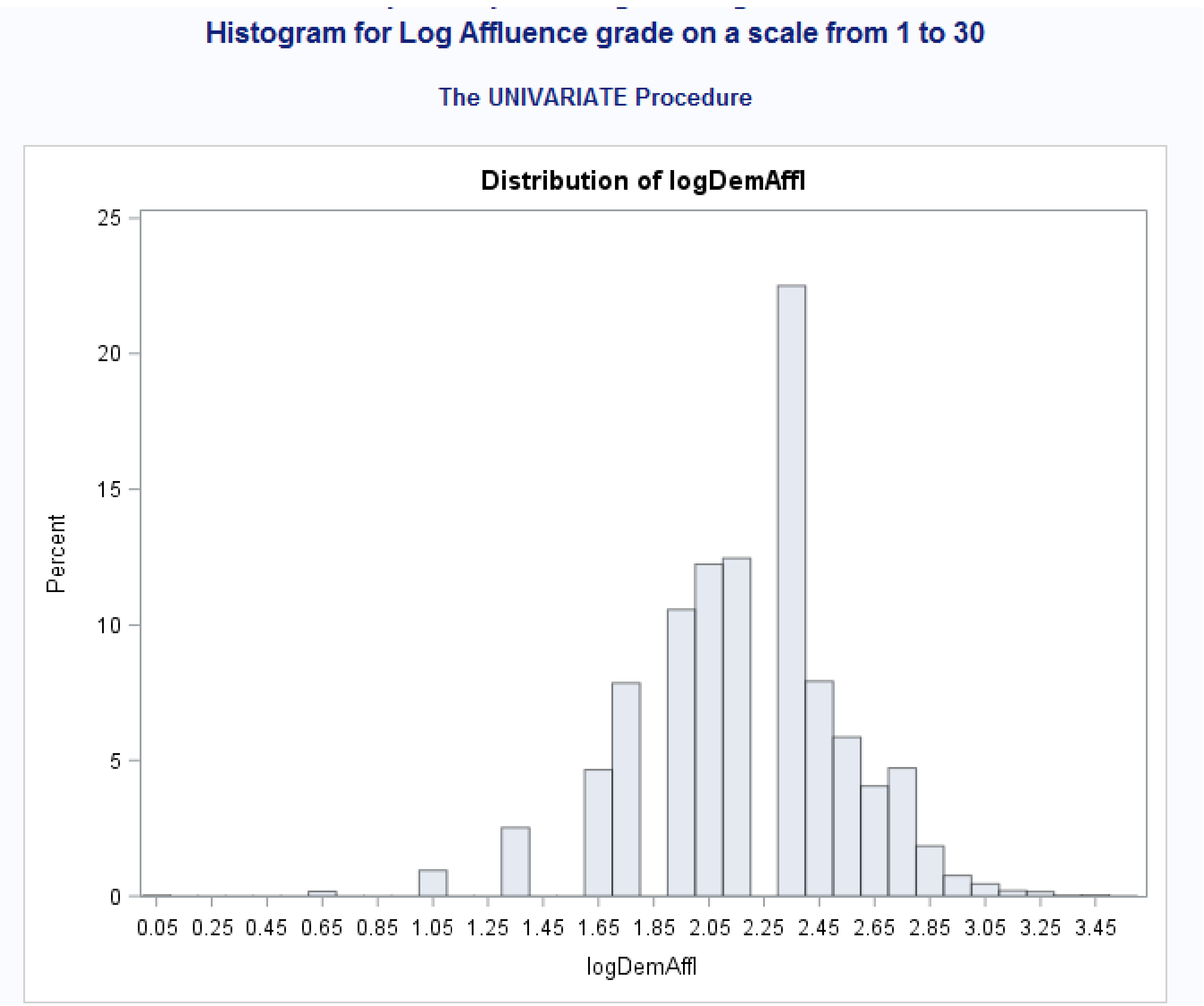
```
There were 22222 observations read from the data set PRJT.ORGANICS2.
The data set PRJT.ORGANICLOG has 22222 observations and 14 variables.
DATA statement used (Total process time):
real time           0.88 seconds
cpu time            0.01 seconds
```

Initially there were 11 observations, after log transformation we have successfully created 14 variables out of which 3 are log transformed.

Histogram after applying Log

Code:

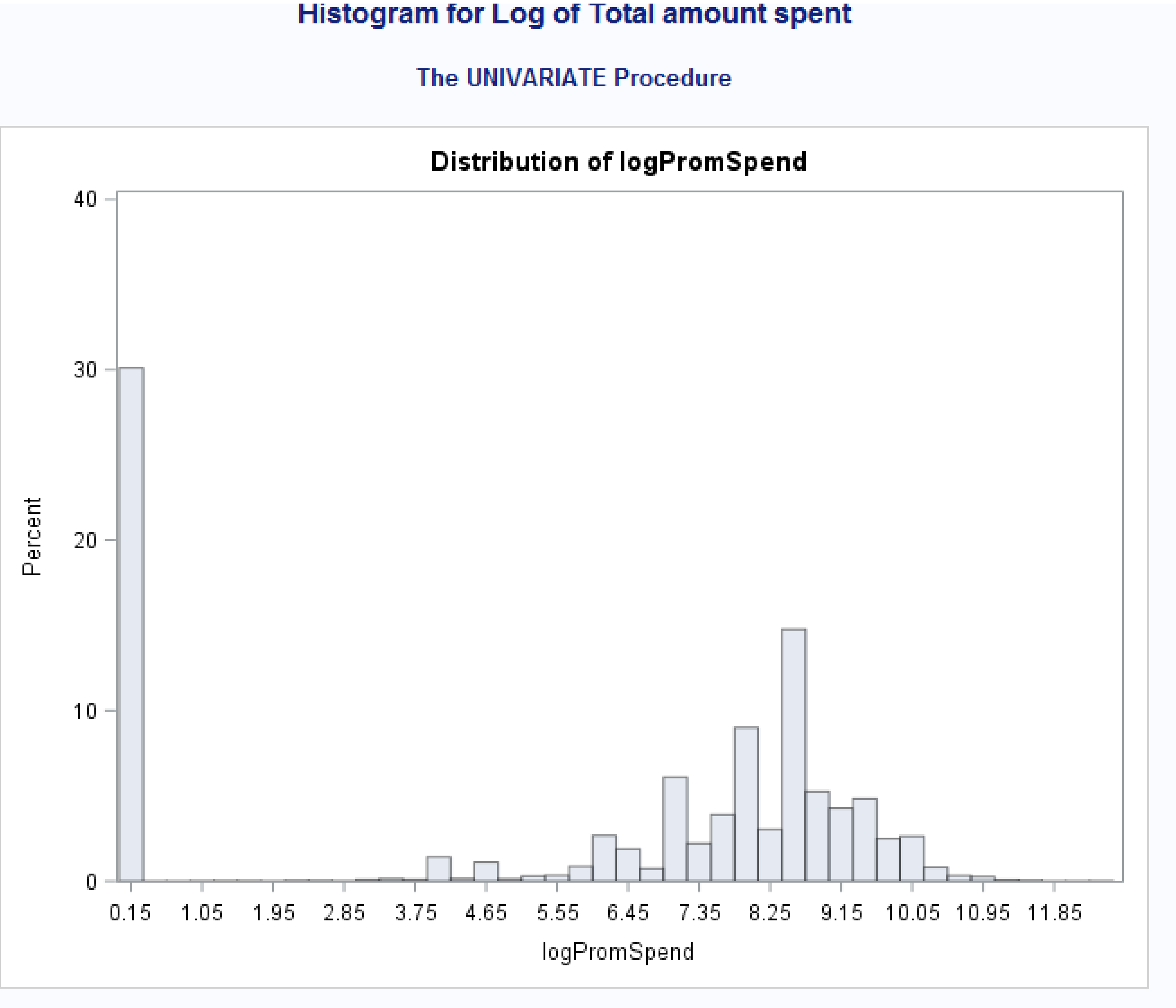
```
TITLE 'Step 8 :Log Transformation of the Data';

data PRJT.ORGANICLOG;
   set PRJT.ORGANICS2;
   logDemAffl = log( DemAffl + 1 );
   logPromSpend = log( PromSpend + 1 );
   logPromTime = log( PromTime + 1 );
run;

goptions reset=global
         gunit=pct
         hsize= 10.625 in
         vsize= 8.5 in
         htitle=4
         htext=3
         vorigin=0 in
         horigin= 0 in
         cback=white border
         ctext=black
         colors=(black blue green red yellow)
         ftext=swiss
         lfactor=3;

proc univariate data=PRJT.ORGANICLOG noprint;
   histogram logDemAffl;
   title2 'Histogram for Log Affluence grade on a scale from 1 to 30';
run;

proc univariate data=PRJT.ORGANICLOG noprint;
   histogram logPromSpend;
   title2 'Histogram for Log of Total amount spent';
run;

proc univariate data=PRJT.ORGANICLOG noprint;
   histogram logPromTime;
   title2 'Histogram for Log of Time as a member';
run;
```



Histogram for Log Affluence grade on a scale from 1 to 30

The UNIVARIATE Procedure

If we look at the histogram after applying log, we can clearly say that log transformation removed the skewness. Now we can come to a conclusion by saying that data is normally distributed as we can see a bell-shaped pattern in our data.

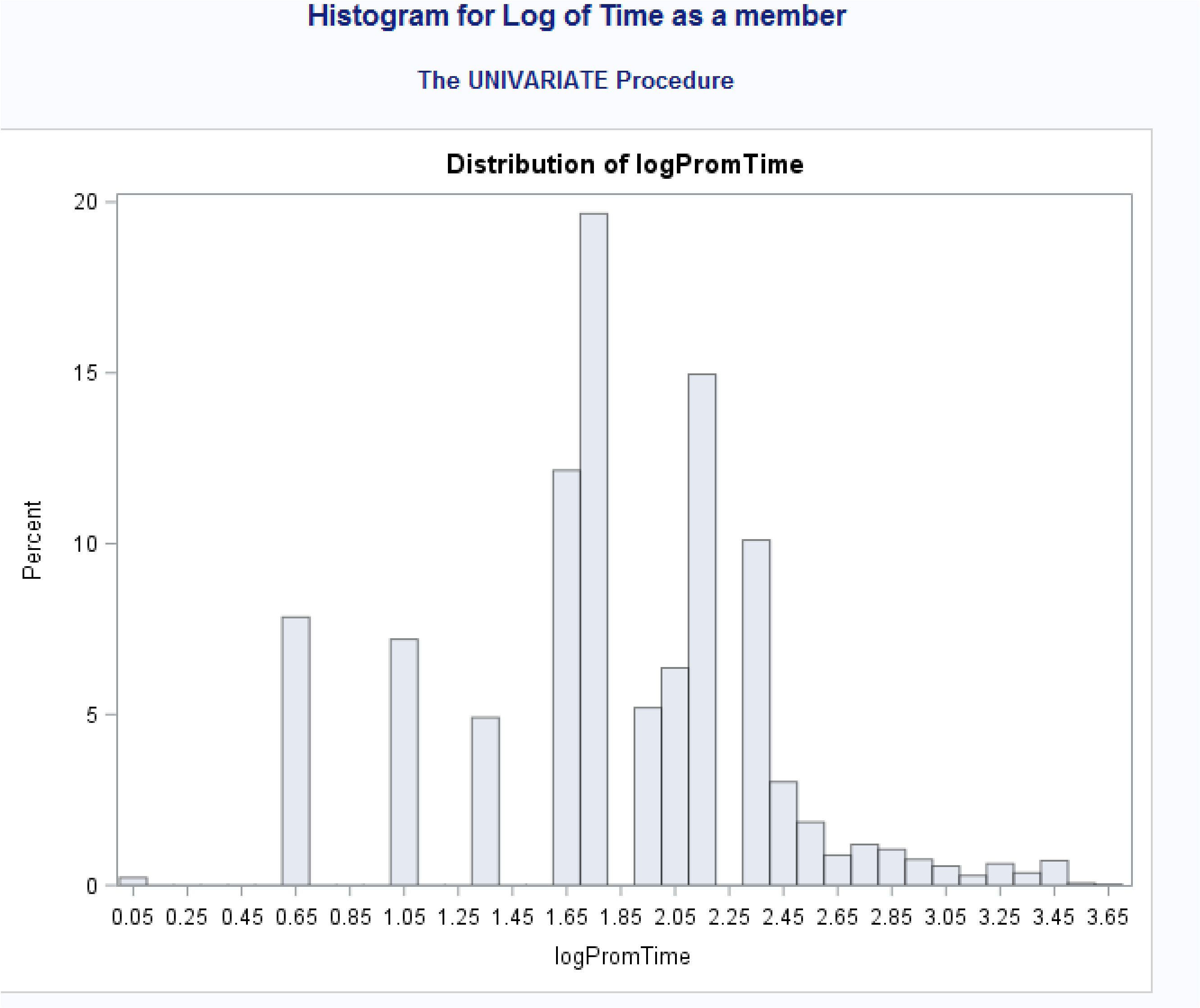


Histogram for Log of Total amount spent

The UNIVARIATE Procedure

Distribution of logPromSpend

If we ignore 30% outliers at 0.15, we can say that rest of the data is normally distribute dafter having a log transformation.



Histogram for Log of Time as a member

The UNIVARIATE Procedure

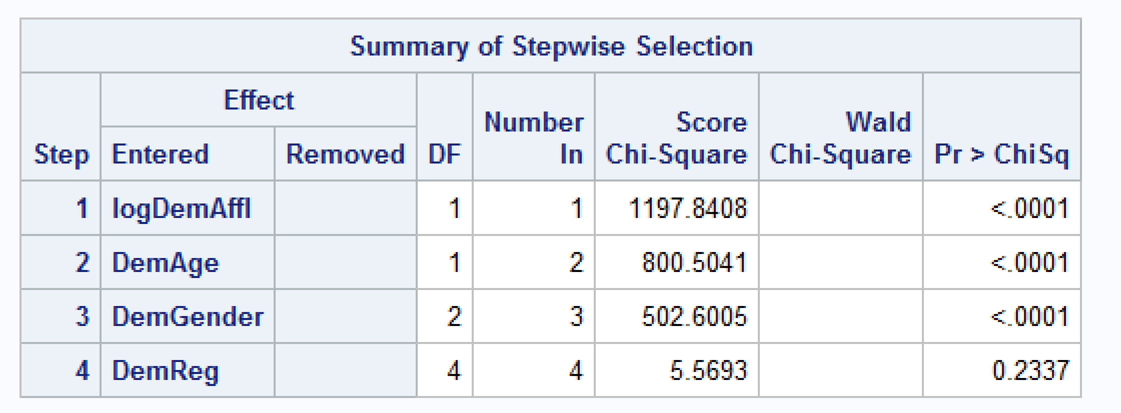
After log transformation, it gave us a better curve. But, we have missing values in our data set which affects our graph although we did log transformation.

Step 9: Please use the transformed variables and use the stepwise method for variable selection and tell me what variables have been selected. (If you have transformed a variable using log, you should consider just the log transformed variable, and the original one should be ignored in this run of variable selection)

Answer:

```
TITLE 'Step 9 :Stepwise Logistic Regression After Log Transformation';

/* Median imputation: Used PROC STDIZE to replace missing values with median */

proc stdize data= PRJT.ORGANICLOG out= PRJT.LOGIMPUTED method= median reponly;
var logDemAffl DemAge logPromTime ;
run;

data PRJT.LOGSORTED;
set PRJT.LOGIMPUTED;
n=ranuni(20041206);

proc sort data=PRJT.LOGSORTED; by n;

data PRJT.LOGTRAINING PRJT.LOGVALIDATION; /*Training and Validation Data*/
set PRJT.LOGSORTED nobs=nobs;
if _n_<=.60*nobs then output PRJT.LOGTRAINING;
else output PRJT.LOGVALIDATION;
run;


proc logistic data=PRJT.LOGTRAINING outest=PRJT.LOGTRAINING_REG covout;
class DemClusterGroup DemGender DemReg DemTVReg PromClass;
model targetBuy(event='1')=logDemAffl DemAge DemClusterGroup DemGender DemReg
DemTVReg PromClass logPromSpend logPromTime
                / selection=stepwise  /*Stepwise selection process*/
                  slentry=0.3
                  slstay=0.35 details;
                          TITLE 'Step 9 :Stepwise Logistic Regression After Log
Transformation';

run;
```

We used log transformed variables to do our stepwise logistic regression. The summary of the stepwise regression is given below



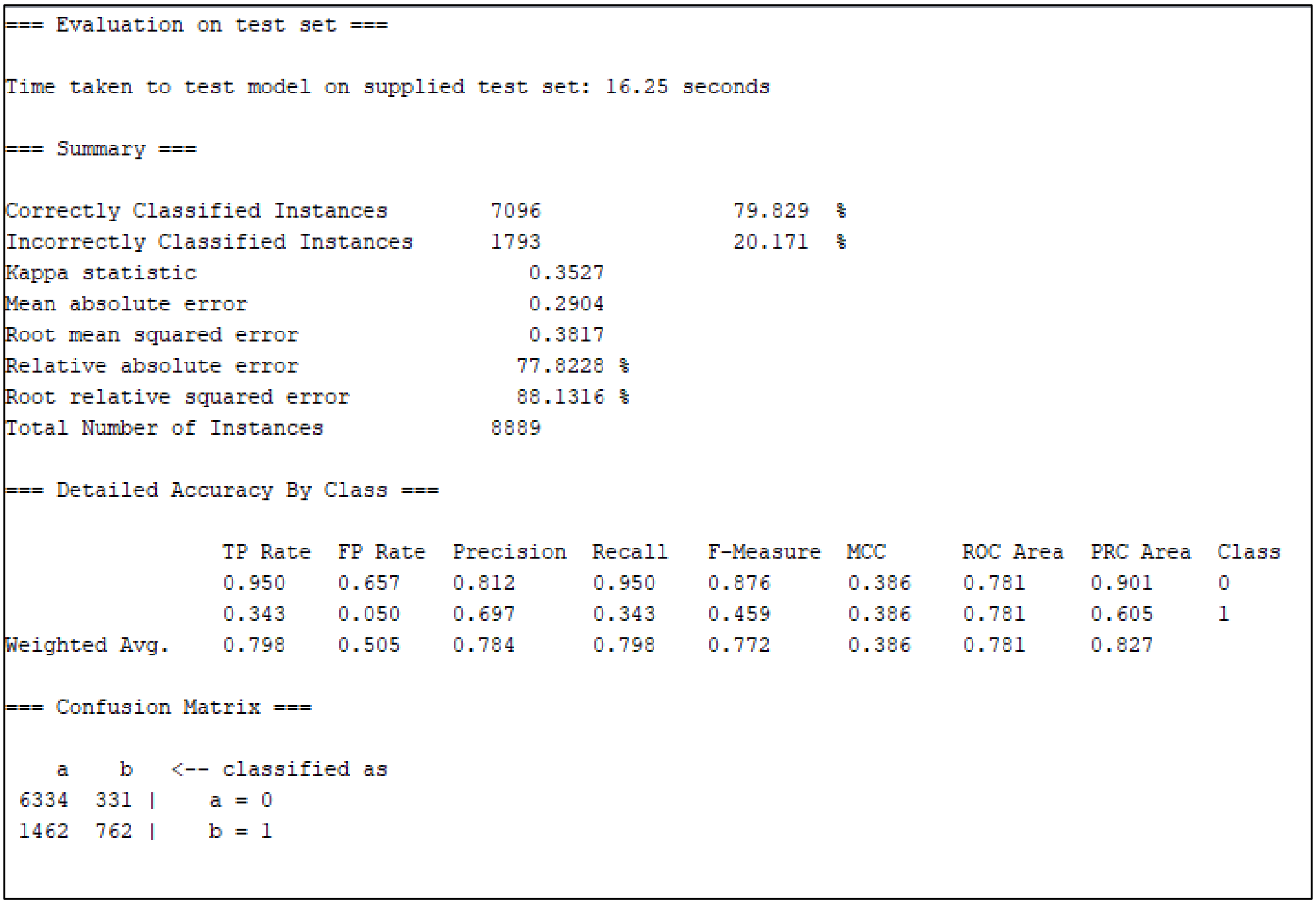
Summary of Stepwise Selection

Step	Effect Entered	 Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	logDemAffl		1	1	1197.8408		<.0001
2	DemAge		1	2	800.5041		<.0001
3	DemGender		2	3	502.6005		<.0001
4	DemReg		4	4	5.5693		0.2337

From our stepwise summary, we ended up selecting **logDemAffl, DemAge and DemGender** variables. As the p-values of the mentioned 3 variables are less than 0.05. We have considered them significant in our model.

Step 10. Please use the variable you selected in Step 9 to fit the same three models you used in step 7. Please create a table to show the precision/recall/accuracy of the three models. Please tell me which model works the best, and if transforming the variables helps achieve better results.

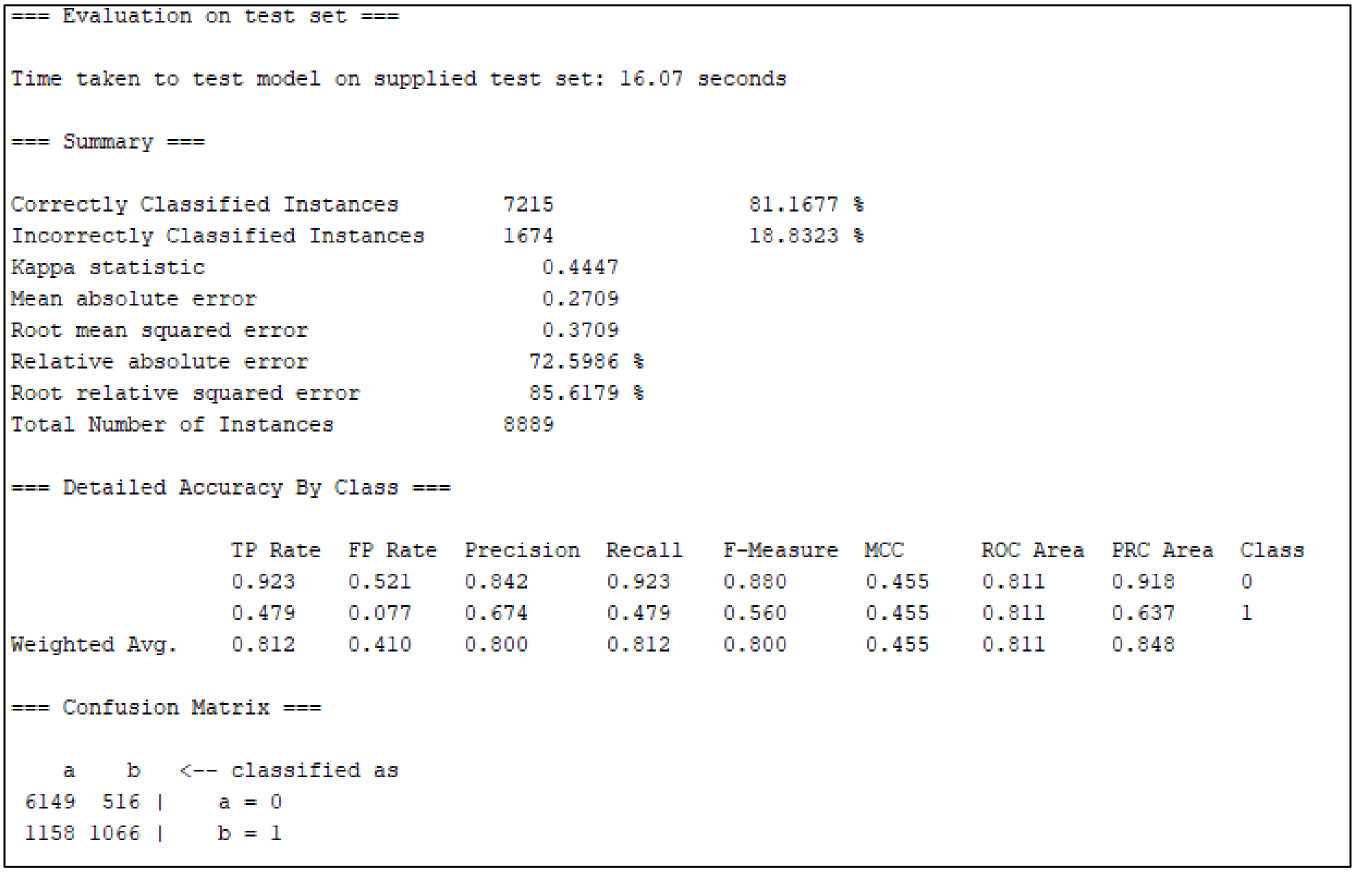
Model 1: Logistic Regression



```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 16.25 seconds

=== Summary ===

Correctly Classified Instances        7096               79.829  %
Incorrectly Classified Instances      1793               20.171  %
Kappa statistic                          0.3527
Mean absolute error                      0.2904
Root mean squared error                  0.3817
Relative absolute error                 77.8228 %
Root relative squared error             88.1316 %
Total Number of Instances             8889

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.950    0.657    0.812      0.950    0.876      0.386    0.781     0.901     0
                 0.343    0.050    0.697      0.343    0.459      0.386    0.781     0.605     1
Weighted Avg.    0.798    0.505    0.784      0.798    0.772      0.386    0.781     0.827

=== Confusion Matrix ===

    a    b   <-- classified as
 6334  331 |    a = 0
 1462  762 |    b = 1
```

Now with use of validation dataset, TargetBuy as Dependent variable and logDemAffl, DemAge and DemGender as independent variables, we have 79.83% overall efficiency. If we look at the precision, recall table, there are two rows defining class 0 and 1. This occurred because of the presence of the two different categories of dependent variable which are treated as positive. **For our convenience, we have selected class 0 as Positive.**

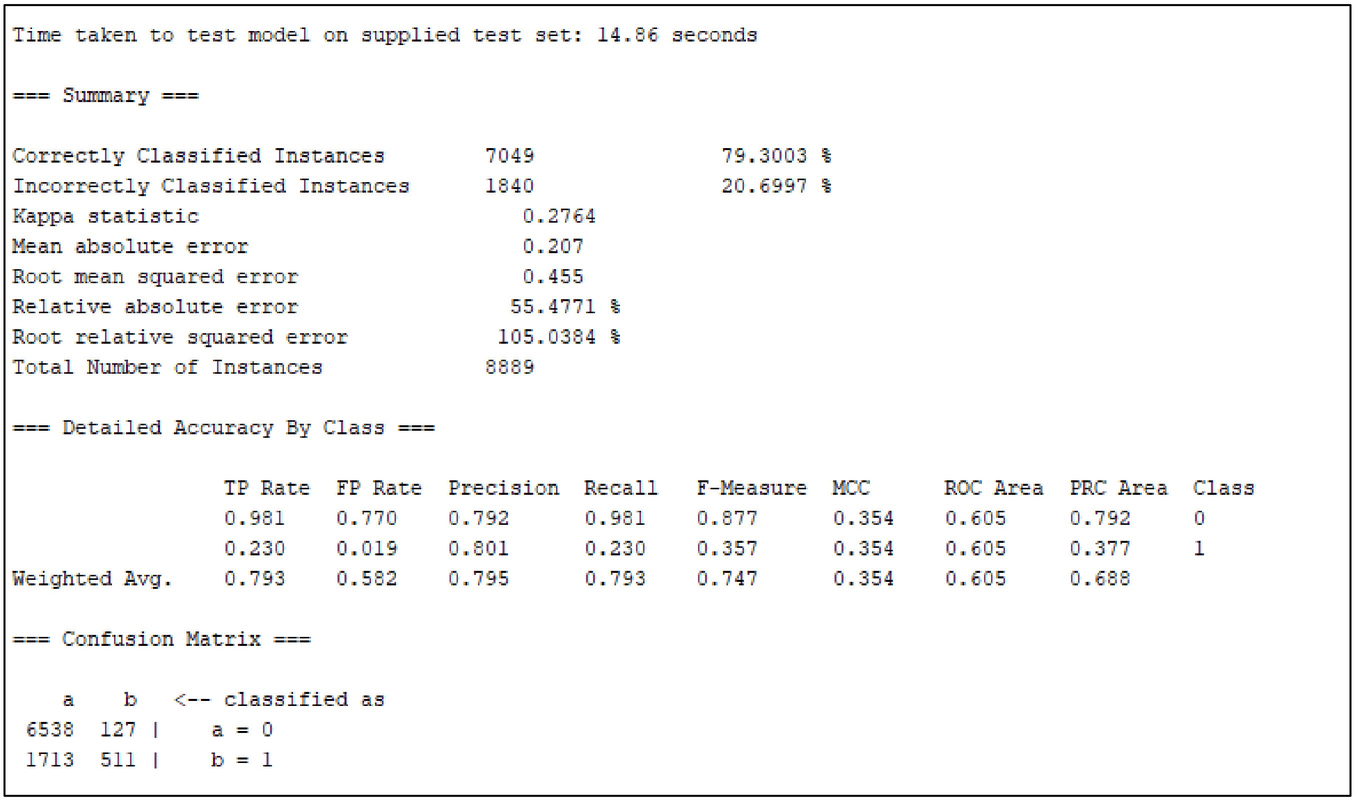
Model 2 : Neural Ntework



```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 16.07 seconds

=== Summary ===

Correctly Classified Instances        7215               81.1677 %
Incorrectly Classified Instances      1674               18.8323 %
Kappa statistic                          0.4447
Mean absolute error                      0.2709
Root mean squared error                  0.3709
Relative absolute error                 72.5986 %
Root relative squared error             85.6179 %
Total Number of Instances             8889

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.923    0.521    0.842      0.923    0.880      0.455    0.811     0.918     0
                 0.479    0.077    0.674      0.479    0.560      0.455    0.811     0.637     1
Weighted Avg.    0.812    0.410    0.800      0.812    0.800      0.455    0.811     0.848

=== Confusion Matrix ===

    a    b   <-- classified as
 6149  516 |    a = 0
 1158 1066 |    b = 1
```

We have chosen MultilayerPerception for modeling a neural network. With TargetBuy as Dependent variable and logDemAffl, DemAge and DemGender as independent variables, we have 81.17% overall efficiency which is a bit better than the logistic regression model.

Model 3



```
Time taken to test model on supplied test set: 14.86 seconds

=== Summary ===

Correctly Classified Instances        7049               79.3003 %
Incorrectly Classified Instances      1840               20.6997 %
Kappa statistic                          0.2764
Mean absolute error                      0.207
Root mean squared error                  0.455
Relative absolute error                 55.4771 %
Root relative squared error            105.0384 %
Total Number of Instances             8889

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.981    0.770    0.792      0.981    0.877      0.354    0.605     0.792     0
                 0.230    0.019    0.801      0.230    0.357      0.354    0.605     0.377     1
Weighted Avg.    0.793    0.582    0.795      0.793    0.747      0.354    0.605     0.688

=== Confusion Matrix ===

    a    b   <-- classified as
 6538  127 |    a = 0
 1713  511 |    b = 1
```

We have chosen Stochastic Gradient Descent for modeling a neural network. With TargetBuy as Dependent variable and logDemAffl, DemAge and DemGender as independent variables, we have 71.30% overall efficiency which is improved compared to the one without log transformation.

Model	Precision	Recall	Accuracy(%)
Logistic Regression	0.812	0.95	79.829
Neural Network	0.842	0.923	81.17
Stochastic Gradiant Descent	0.792	0.981	79.3

Based on accuracy we can say that Neural network gives us a good result with highest accuracy among those 3 models but recall value was less than the other two models. As we know that, recall is a good way to determine the fitness of a model. Based on recall and accuracy, I would choose logistic regression model would be a best model for this problem. With Log transformation, I got better result for the stochastic gradient descent model.

Below is a summary of things you need to submit:

1. All SAS code
2. Histograms
3. Tell me the variables that have missing values and variables that have skewed distribution.
4. Tell me which variables you selected in the first round of variable selection.
5. Use the variables you selected in the previous step to do model fitting. Tell me which three models (including logistic regression) you used.
6. Submit a table that shows the precision/recall/accuracy of the three models.
7. After you transformed the variables with skewed distribution, please do a variable selection again and tell me which variables you selected. Fit the three models and submit a table that shows the precision/recall/accuracy of the three models.
8. Please tell if log transformation results in better model performance.
9. Nothing else.