Αναφορά εργασίας εξαμήνου στο μάθημα

Εφαρμοσμένα πληροφοριακά συστήματα ΙΙ

Εργαλείο:SAS

Ονόματα:	ΑΜ:
Δημητριάδης Μιλτιάδης(*)	5288
Πατρώνη Σωτηρία	5399
Τσαρνάς Βασίλειος	5444

Ερώτημα 1:

Για την ολοκλήρωση του καθαρισμού των δεδομένων του dataset χρειάστηκε αρχικά να ενωθούν τα δύο αρχεία δεδομένων Campaign1.csv και Campaign2.csv σε ένα dataset και έπειτα να ξεκινήσει ο καθαρισμός.

- Εισαγωγή στοιχείων του αρχείου Campaign1.csv στο dataset CAMP.campaign1

```
/*Import arxeiou Campaign1 */
/*Observations=1700, Variables=18*/

data CAMP.campaign1 ;
infile "&path/Campaign1.csv" dlm=';' dsd firstobs=2;
input id_ targetD:dollar10.2 GiftCnt36 GiftCntAll GiftCntCard36 GiftCntCardAll
        GiftAvgLast:dollar10.2 GiftAvg36:dollar10.2 GiftAvgAll:dollar10.2
        GiftAvgCard36:dollar10.2 GiftTimeLast GiftTimeFirst  PromCnt12
        PromCnt36 PromCntAll PromCntCard12 PromCntCard36 PromCntCardAll;
run;

proc print data=CAMP.campaign1;
run;

proc contents data=CAMP.campaign1;
run;
```

- Εισαγωγή στοιχείων του αρχείου Campaign2.csv στο dataset CAMP.campaign2

```
/*Import arxeiou Campaign2 */
/*Observations=1700, Variables=11*/
data CAMP.campaign2 ;
infile "&path/Campaign2.csv" dlm=';' dsd firstobs=2;
input id_ StatusCat96NK $ StatusCatStarAll DemCluster DemAge DeMGender $
        DemHomeOwner $ DemIncomeGroup DemMedHomeValue:dollar10.2
        DemPctVeterans TargetB;
run;

proc print data=CAMP.campaign2;
run;

proc contents data=CAMP.campaign2;
run;
```

- Ταξινόμηση των δύο dataset με βάση το id για την ένωση τους

```
/*Sorting */
proc sort data=CAMP.campaign1
out=CAMP.campaign1_sorted;
by id_;
run;

proc print data=CAMP.campaign1_sorted;
run;

proc sort data=CAMP.campaign2
out=CAMP.campaign2_sorted;
by id_;
run;

proc print data=CAMP.campaign2_sorted;
run;
```

- Ένωση των δύο dataset

```
/*Merging*/
data CAMP.campaign12 ;
merge CAMP.campaign1_sorted (in=a) CAMP.campaign2_sorted (in=b) ;
by id_;
if a=1 and b=1;
run;

proc print data=CAMP.campaign12;
run;

proc contents data=CAMP.campaign12;
run;

proc freq data=CAMP.campaign12;
tables _NUMERIC_ /missing;
tables _CHAR_ /missing;
run;
```

- Καθαρισμός του dataset μετά την ένωση

```
/*Cleaning the dataset*/
data CAMP.cleaned(drop= DemPctVeterans DemAge TargetB);
set CAMP.campaign12;

if DeMGender = 'Manan' then DeMGender = 'M';
if DeMGender = 'Man' then DeMGender = 'M';
if DemHomeOwner = 'Home' then  DemHomeOwner = 'H';
if DemHomeOwner = 'NoHome' then  DemHomeOwner = 'NH';
if DeMGender = 'Woman' then DeMGender = 'F';
if DeMGender = 'U' then delete;
if GiftAvgCard36 = '.' then delete;
if DemIncomeGroup = '.' then delete;
if TargetB = '0'  then delete;
if DemMedHomeValue = '0' then delete;
if GiftAvgLast = '0' then delete;
if PromCntCard12 = '0' then delete;
if GiftCnt36 = '0'  then delete;
if GiftCntCard36 = '0' then delete;
if GiftCntCardAll = '0' then delete;
if GiftAvg36 = '0' then delete;

run;
```

```
proc print data=CAMP.cleaned;
run;

proc contents data=CAMP.cleaned;
run;

proc freq data=CAMP.cleaned;
tables _NUMERIC_ /missing;
tables _CHAR_ /missing;
run;
```

Αρχικά, στην μεταβλητή DeMGender διορθώνονται όλες οι άτυπες τιμές και εκχωρούνται μόνο M (male) και F (female), έπειτα γίνεται το ίδιο και για τη μεταβλητή DemHomeOwner και τέλος διαγράφονται όλες οι λανθασμένες τιμές και τα missing values. Η τιμή U της μεταβλητής DeMGender και οι τιμές 0 των μεταβλητών DemMedHomeValue, GiftAvgLast, PromCntCard12, GiftCnt36, GiftCntCard36, GiftCntCardAll, GiftAvg36 είναι λανθασμένες τιμές, τα missing values βρίσκονται μόνο στις μεταβλητές GiftAvgCard36 και DemIncomeGroup και τέλος για την μεταβλητή TargetB η τιμή 0 είναι άχρηστη αφού δεν μας αφορούν οι πελάτες που δε θα πάρουν κάποιο δώρο. Επίσης, γίνεται διαγραφή των στηλών:

1. DemPctVeterans, αφού από την εκφώνηση είναι γνωστό ότι όλες οι μεταβλητές που υπολογίζουν κάποιο ποσοστό είναι άχρηστες και λανθασμένες,
2. DemAge, αφού το ποσοστό των λανθασμένων και των τιμών που απουσιάζουν είναι αρκετά μεγάλο για να αφαιρεθούν όλες αυτές οι εγγραφες οπότε γίνεται διαγραφή όλης της στήλης και
3. TargetB, αφού έχουν κρατηθεί όλες οι εγγραφές που είναι χρησιμες με βάση τη μεταβλητή αυτή (δηλαδή όλες οι εγγραφές των πελατών που θα λάβουν κάποιο δώρο).

Αποτελέσματα εκτέλεσης της διαδικασίας proc contents:

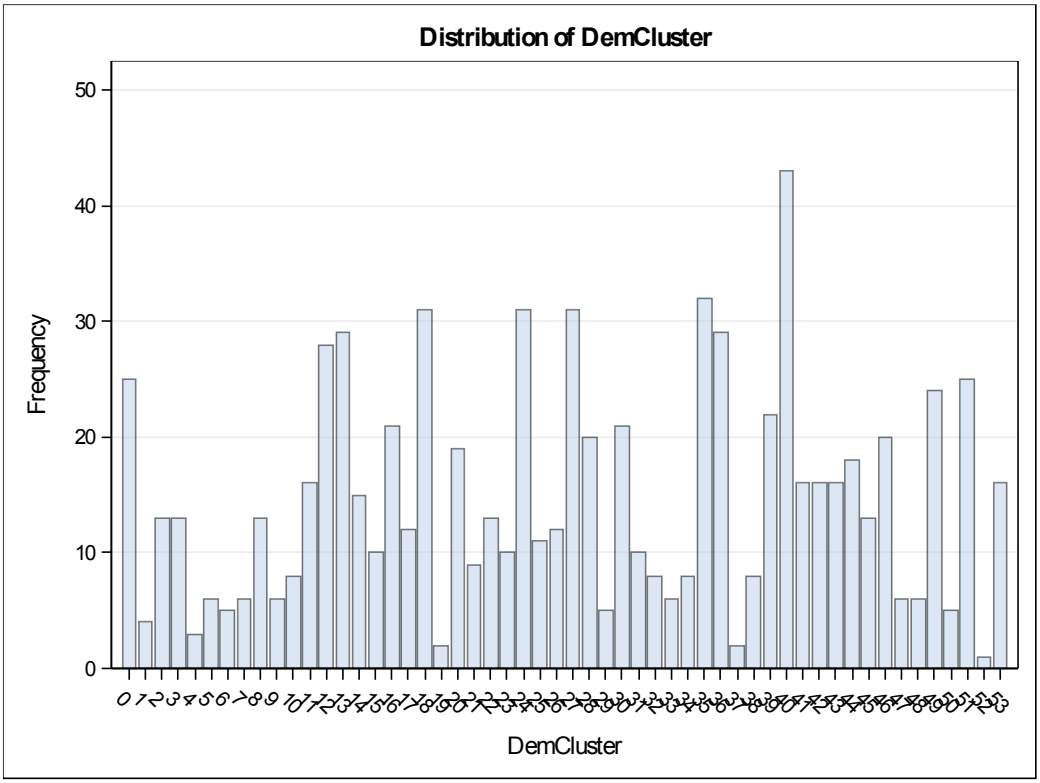
Data Set Name	CAMP.CLEANED	**Observations**	798
Member Type	DATA	**Variables**	25
Engine	V9	**Indexes**	0
Created	07/06/2015 09:59:51	**Observation Length**	200
Last Modified	07/06/2015 09:59:51	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

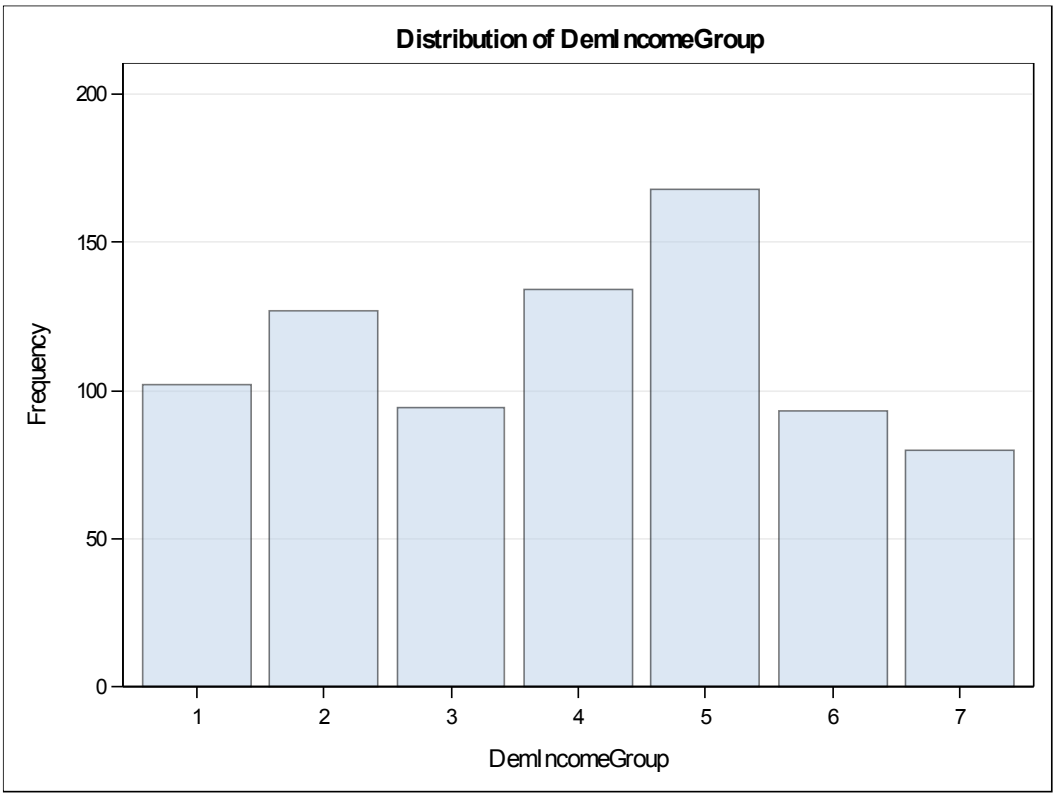
Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	3
First Data Page	1
Max Obs per Page	327
Obs in First Data Page	302
Number of Data Set Repairs	0
Filename	/folders/myfolders/myproject/cleaned.sas7bdat
Release Created	9.0401M2
Host Created	Linux
Inode Number	11554

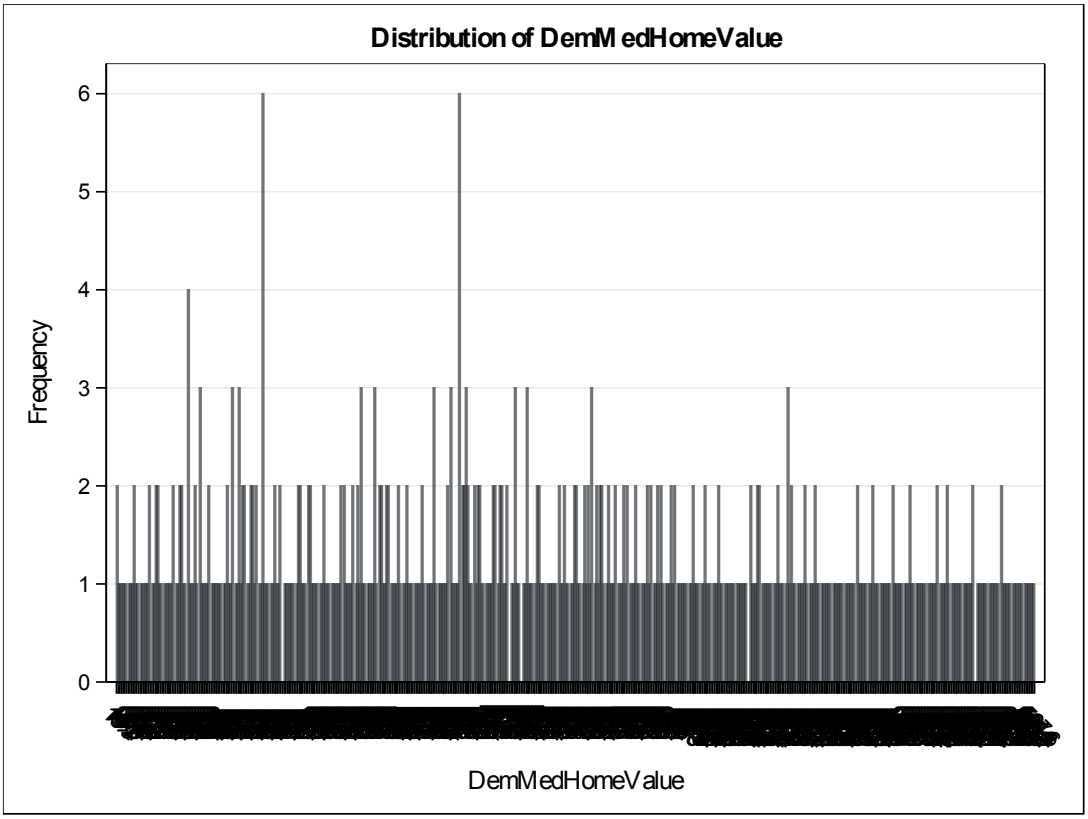
Engine/Host Dependent Information	
Access Permission	rwxrwxrwx
Owner Name	root
File Size (bytes)	262144

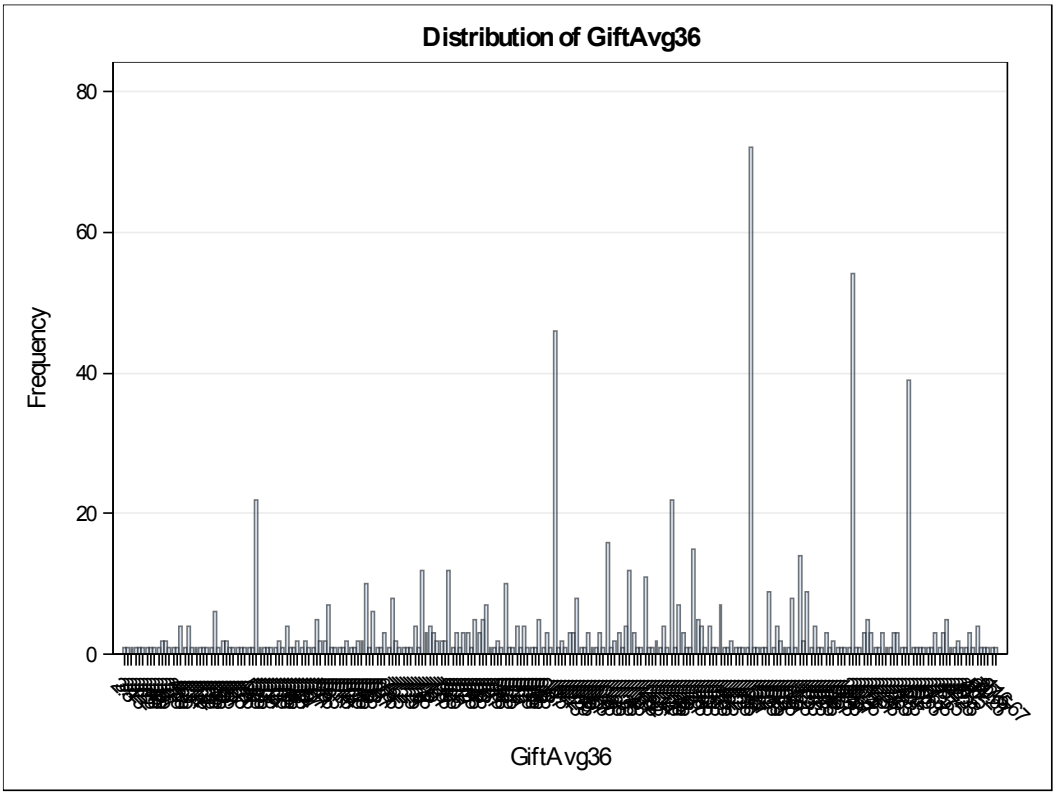
Alphabetic List of Variables and Attributes			
#	**Variable**	**Type**	**Len**
22	DeMGender	Char	8
21	DemCluster	Num	8
23	DemHomeOwner	Char	8
24	DemIncomeGroup	Num	8
25	DemMedHomeValue	Num	8
8	GiftAvg36	Num	8
9	GiftAvgAll	Num	8
10	GiftAvgCard36	Num	8
7	GiftAvgLast	Num	8
3	GiftCnt36	Num	8
4	GiftCntAll	Num	8
5	GiftCntCard36	Num	8
6	GiftCntCardAll	Num	8
12	GiftTimeFirst	Num	8
11	GiftTimeLast	Num	8
13	PromCnt12	Num	8
14	PromCnt36	Num	8
15	PromCntAll	Num	8
16	PromCntCard12	Num	8
17	PromCntCard36	Num	8
18	PromCntCardAll	Num	8
19	StatusCat96NK	Char	8
20	StatusCatStarAll	Num	8
1	id_	Num	8
2	targetD	Num	8

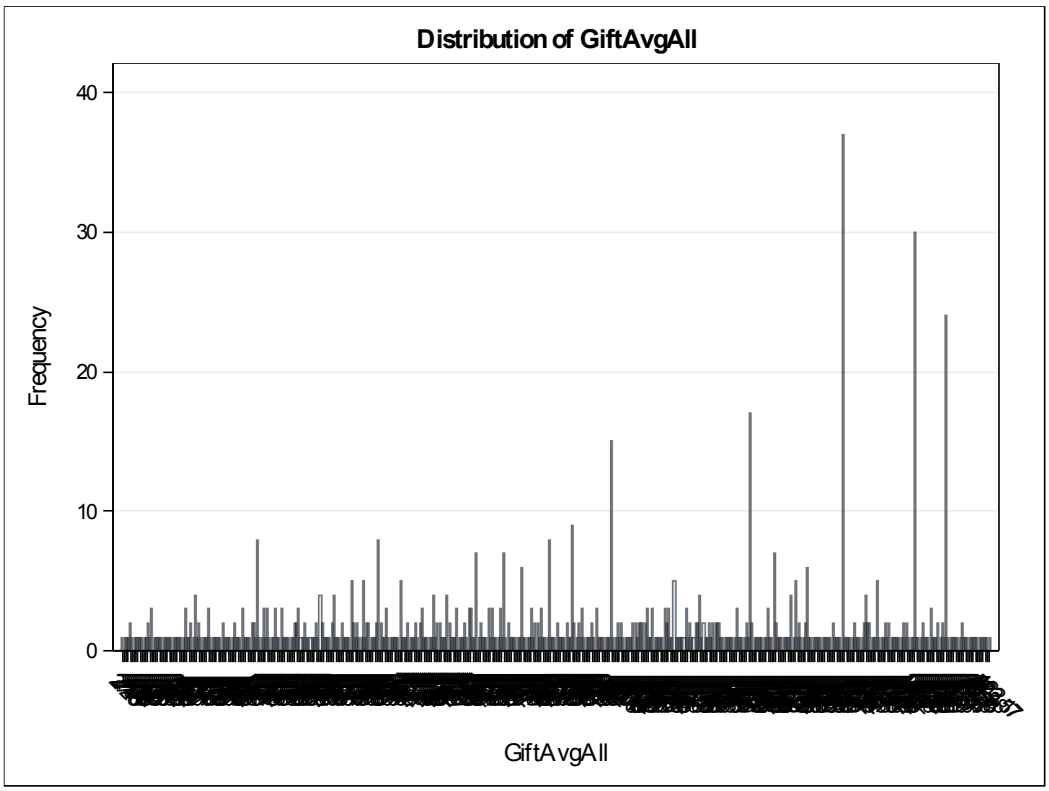
Ιστογράμματα:

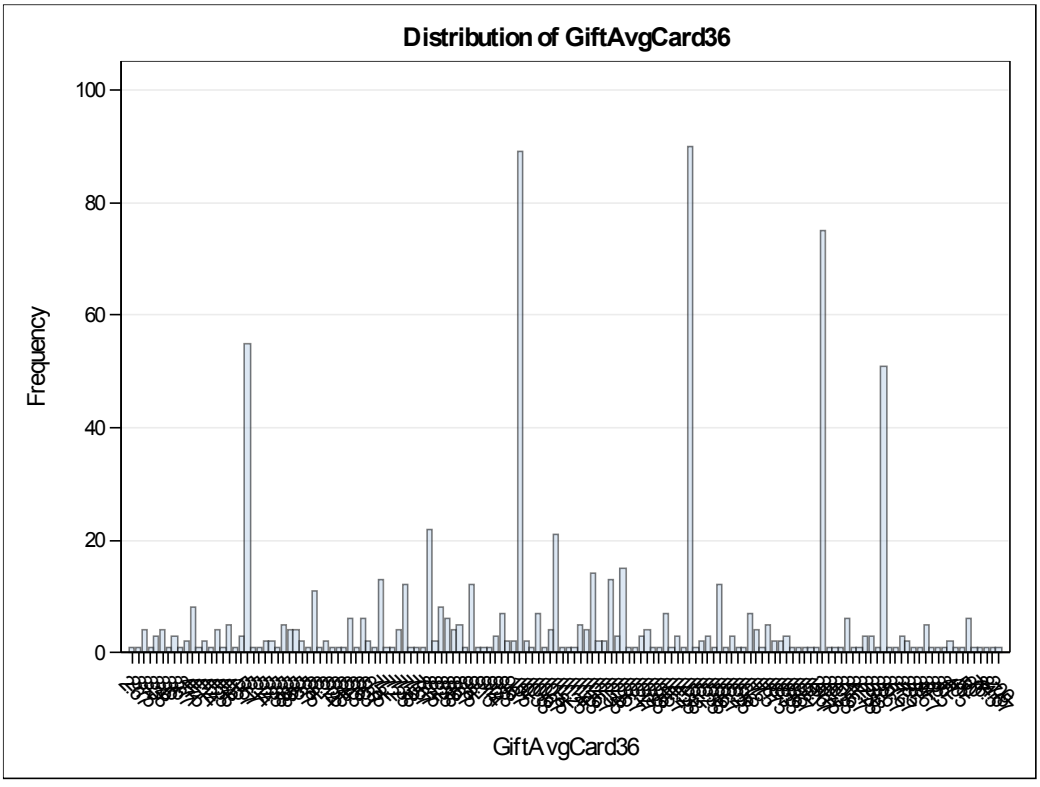


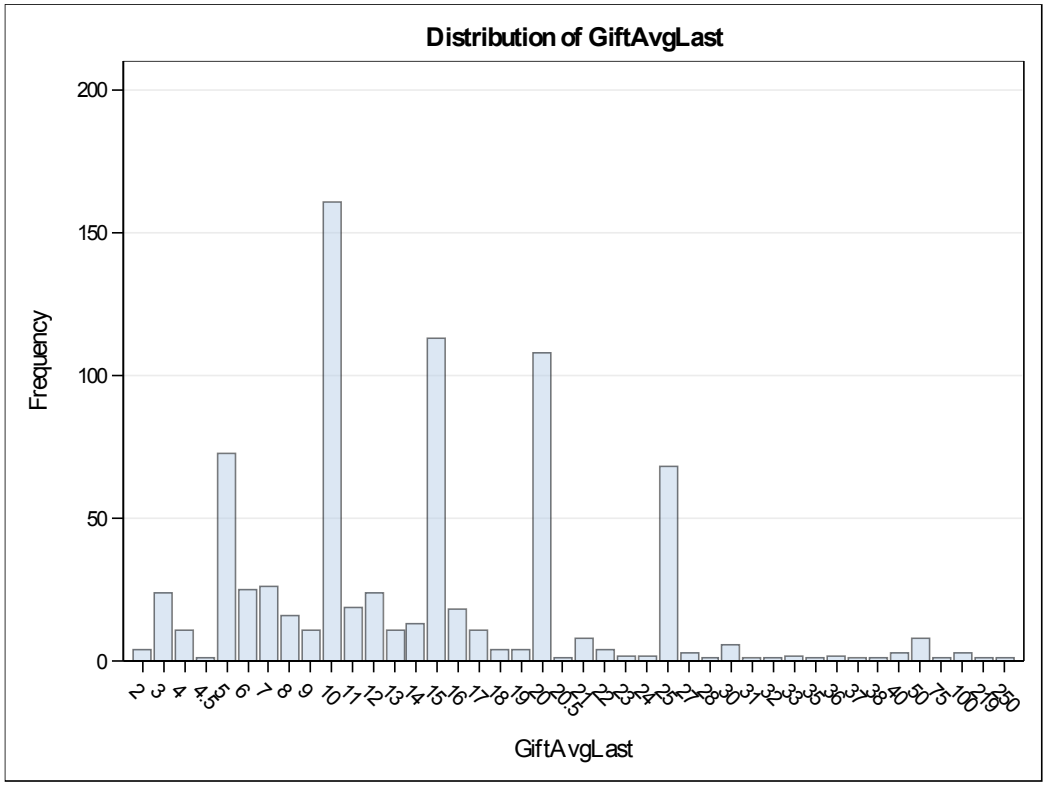


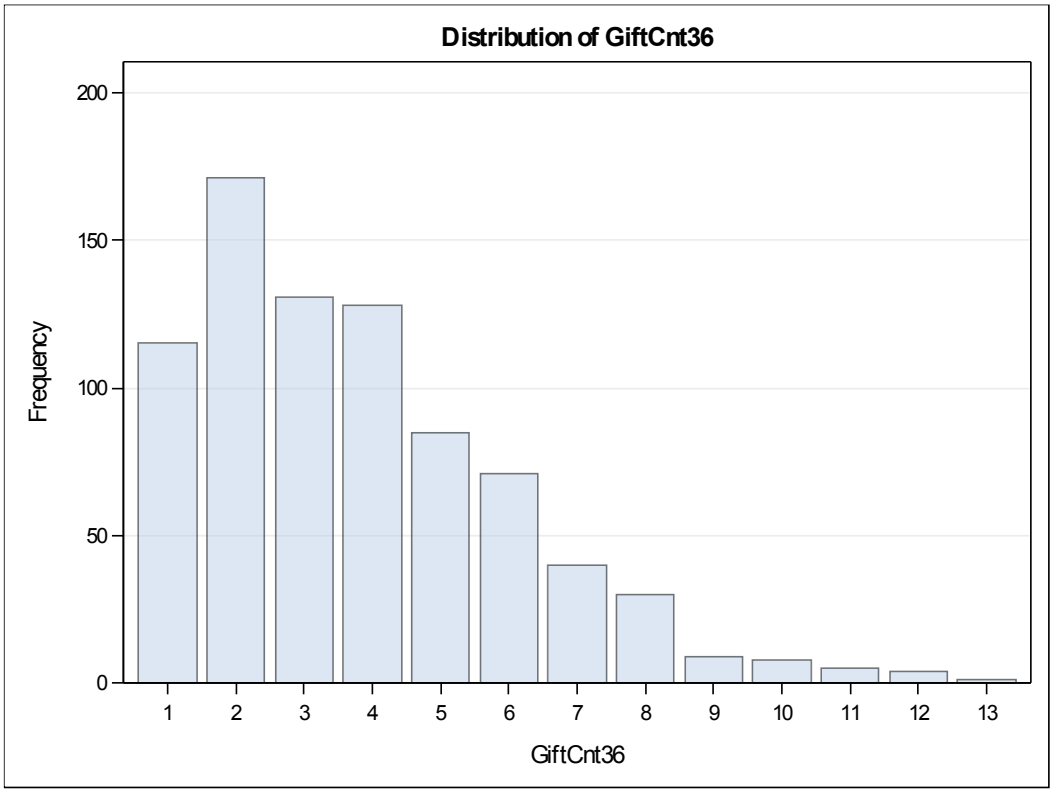

Distribution of DemMedHomeValue
Frequency
0
1
2
3
4
5
6
DemMedHomeValue

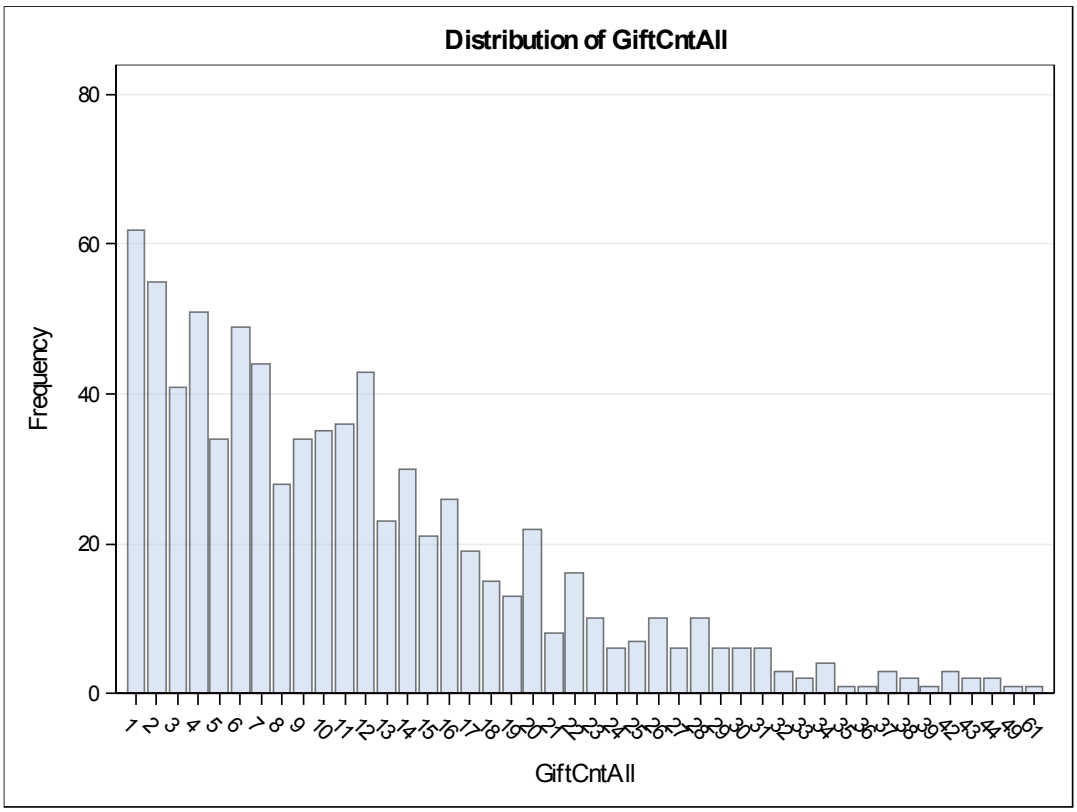

Distribution of GiftAvg36
Frequency
0
20
40
60
80
GiftAvg36

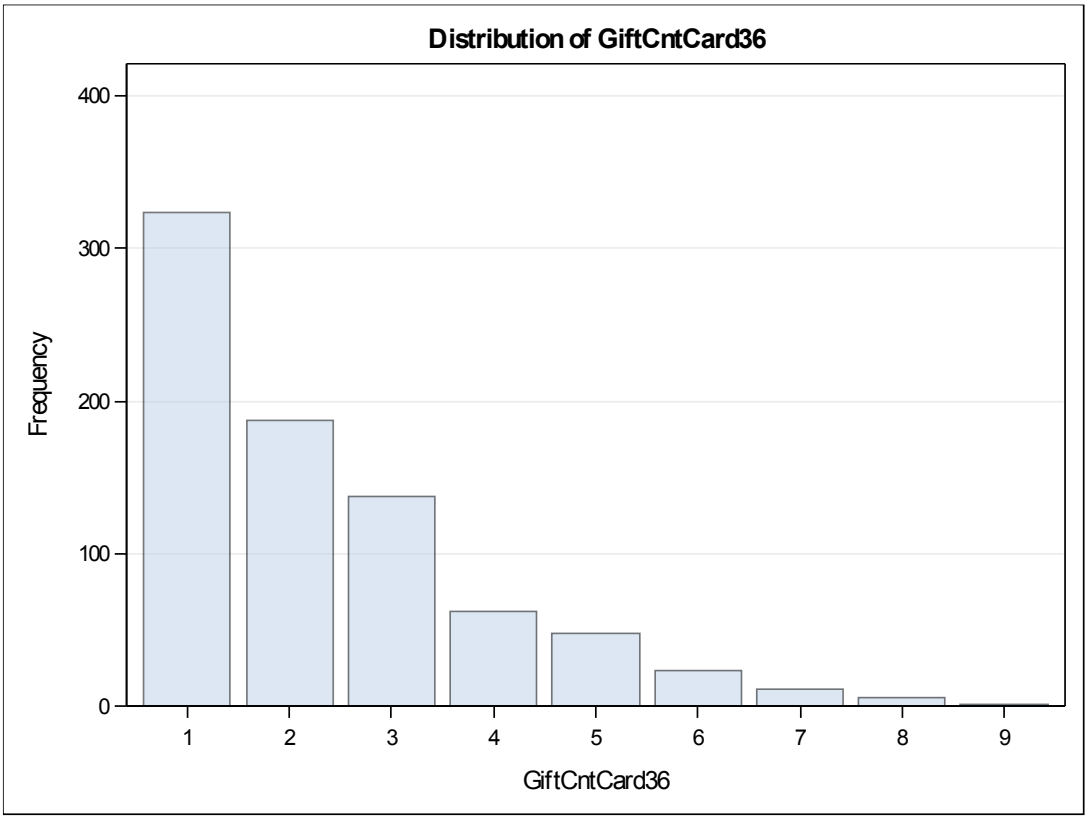

Distribution of GiftAvgAll
Frequency
40
30
20
10
0
GiftAvgAll

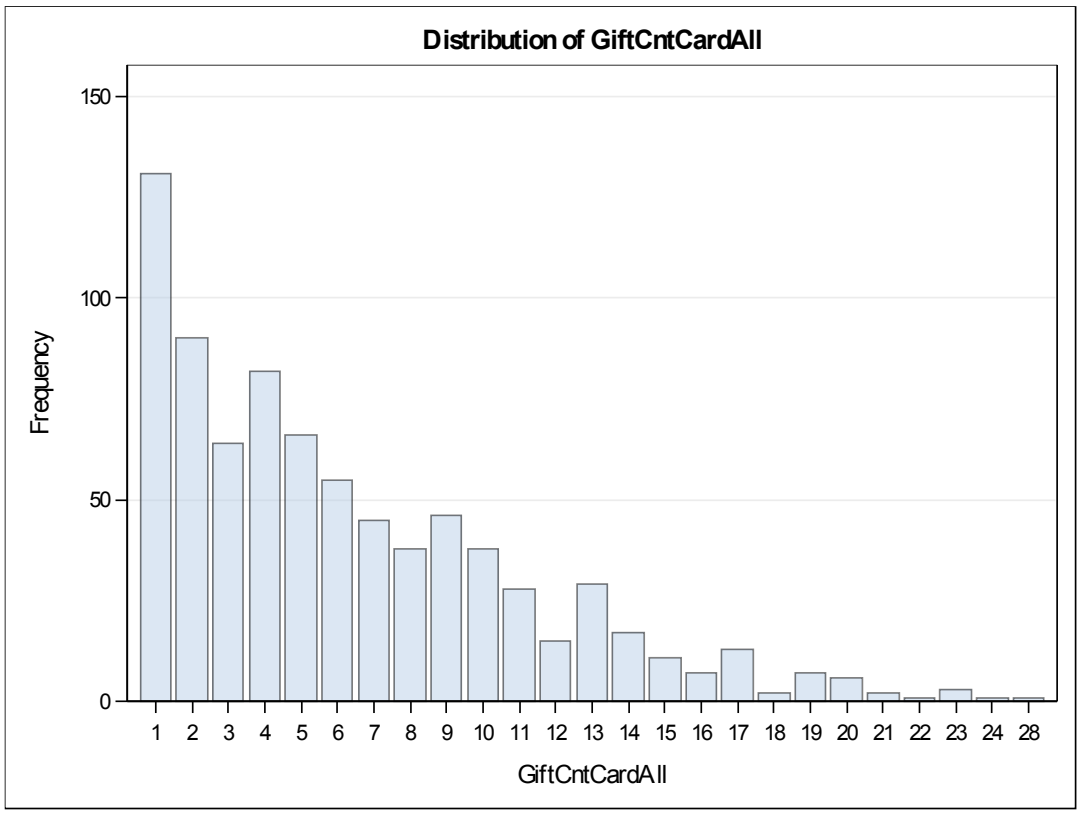

Distribution of GiftAvgCard36
Frequency
100
80
60
40
20
0
GiftAvgCard36

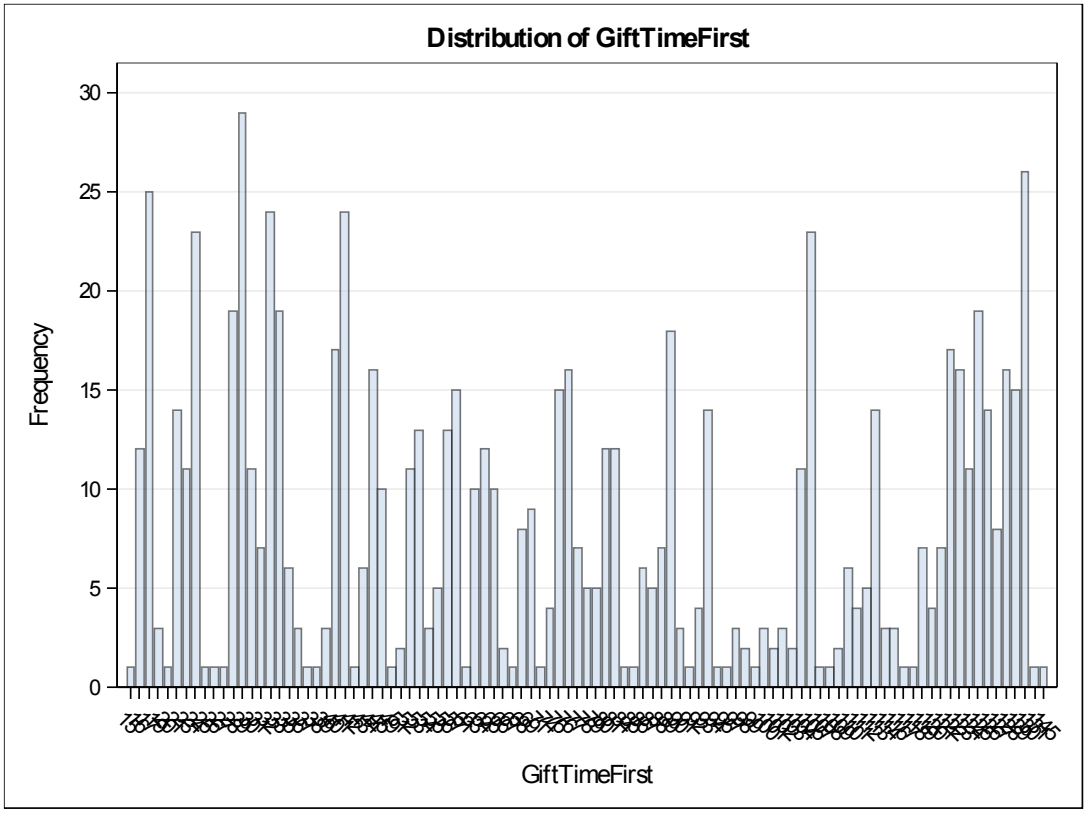

Distribution of GiftAvgLast
Frequency
200
150
100
50
0
GiftAvgLast

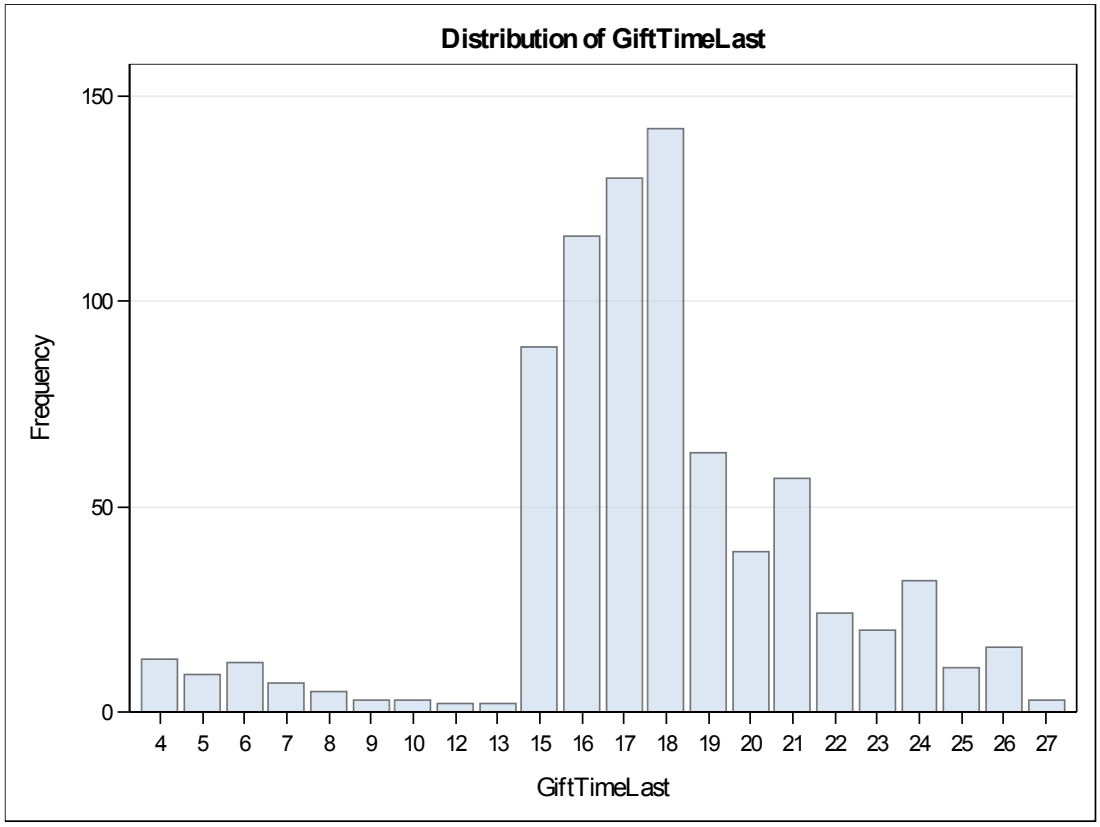

Distribution of GiftCnt36
Frequency
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
11
12
13
GiftCnt36

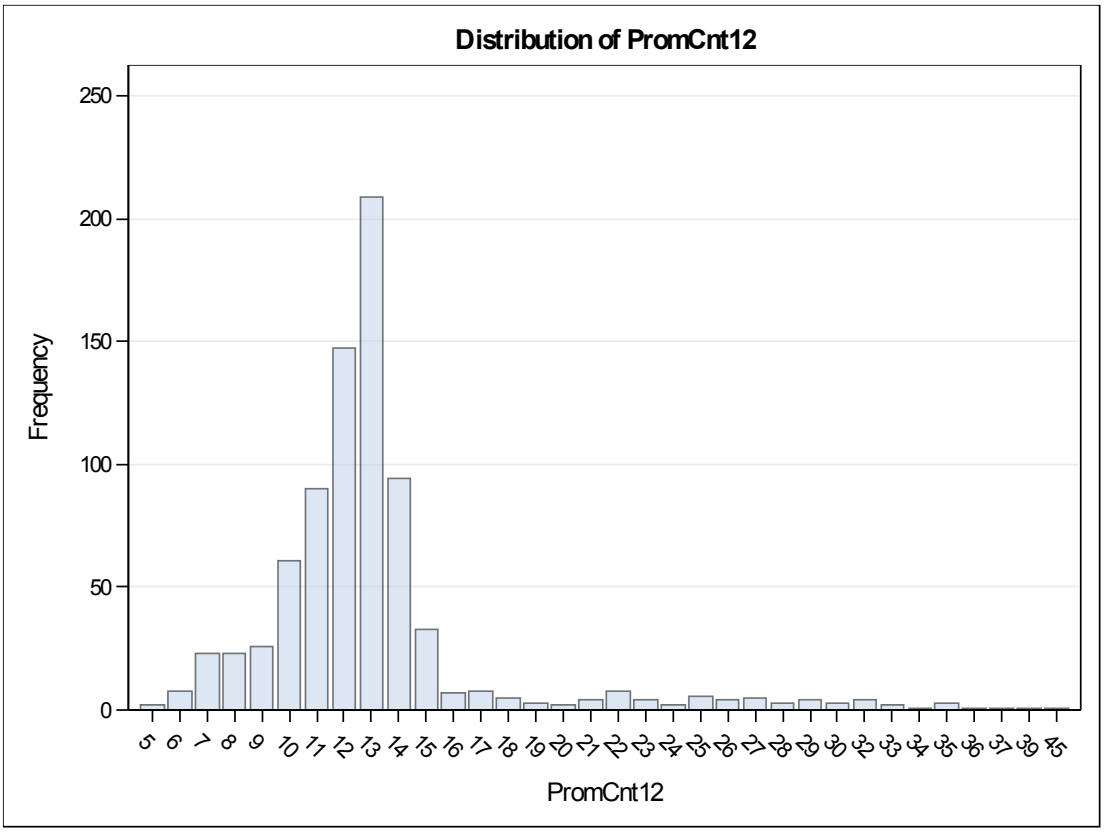

Distribution of GiftCntAll
Frequency
80
60
40
20
0
GiftCntAll

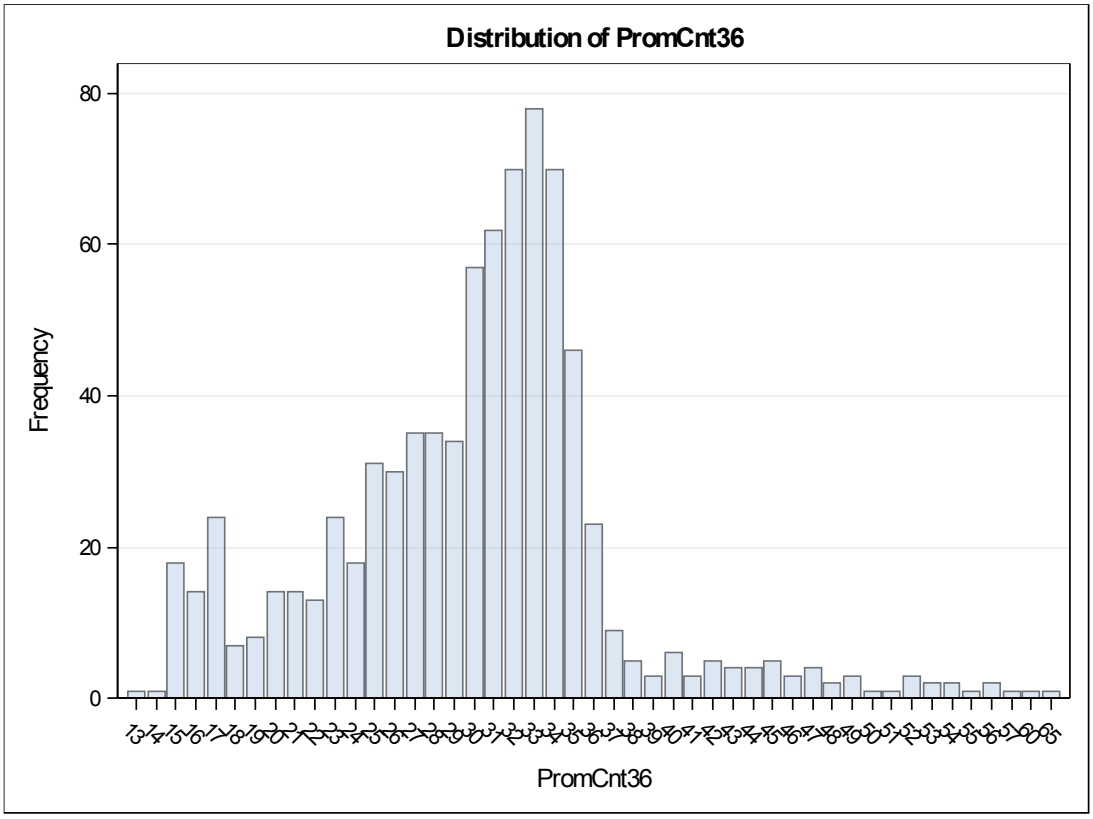

Distribution of GiftCntCard36
Frequency
400
300
200
100
0
1
2
3
4
5
6
7
8
9
GiftCntCard36

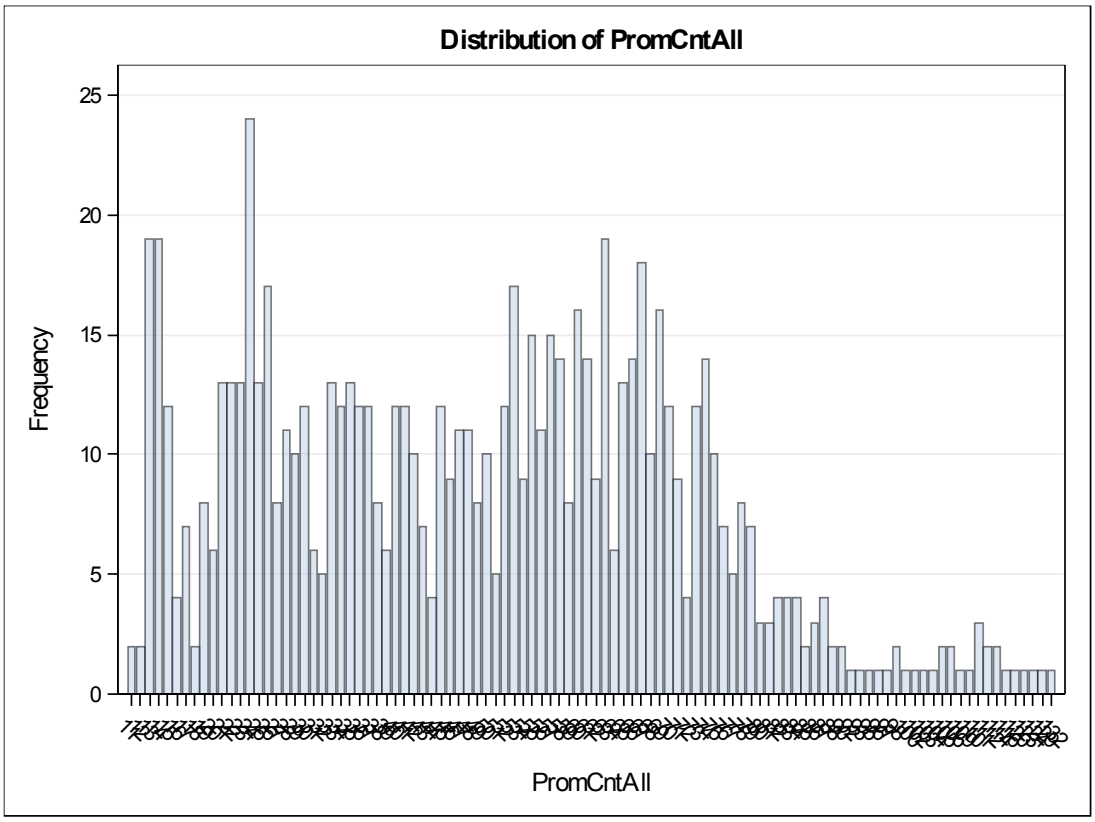

Distribution of GiftCntCardAll
Frequency
150
100
50
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 28
GiftCntCardAll

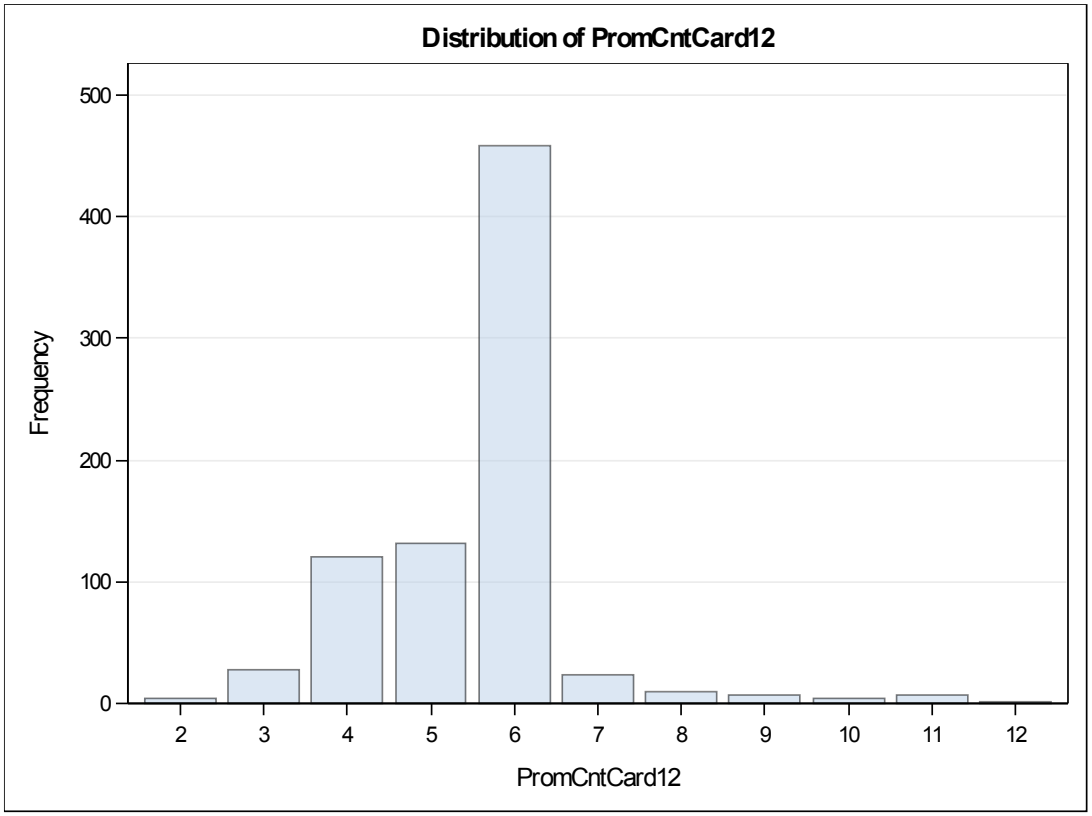

Distribution of GiftTimeFirst
Frequency
30
25
20
15
10
5
0
GiftTimeFirst

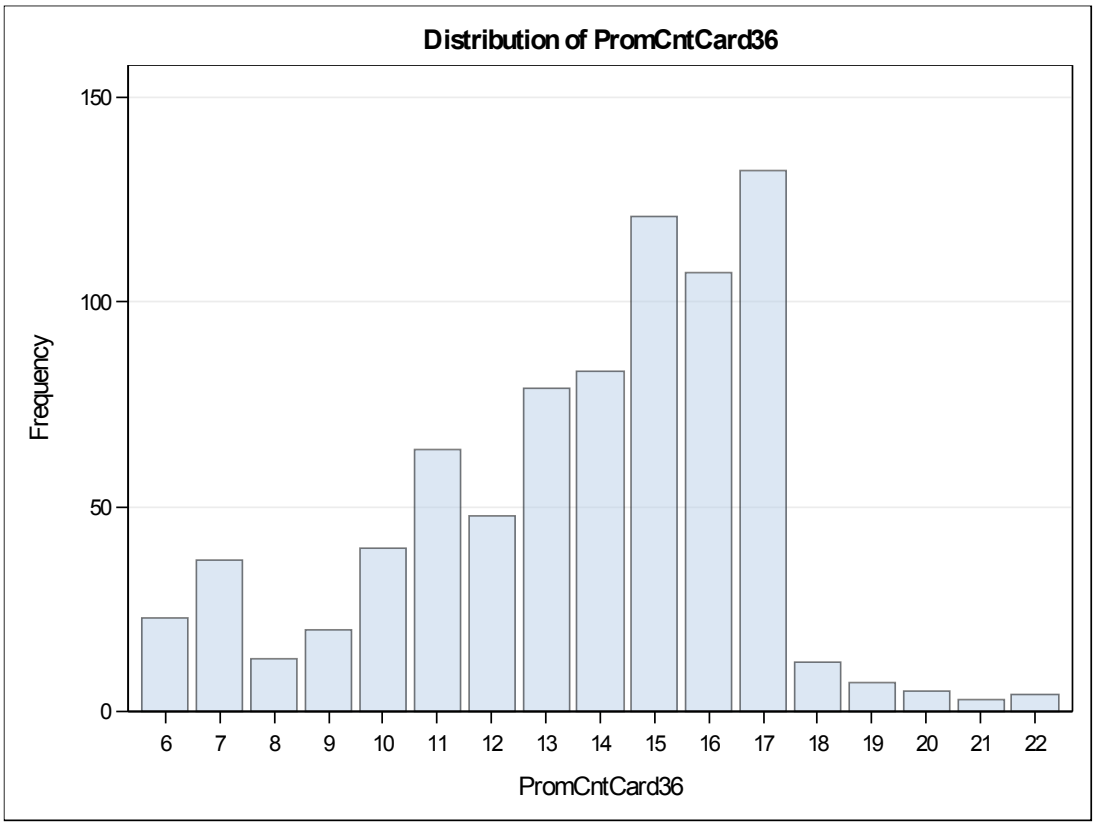

Distribution of GiftTimeLast
Frequency
150
100
50
0
4 5 6 7 8 9 10 12 13 15 16 17 18 19 20 21 22 23 24 25 26 27
GiftTimeLast

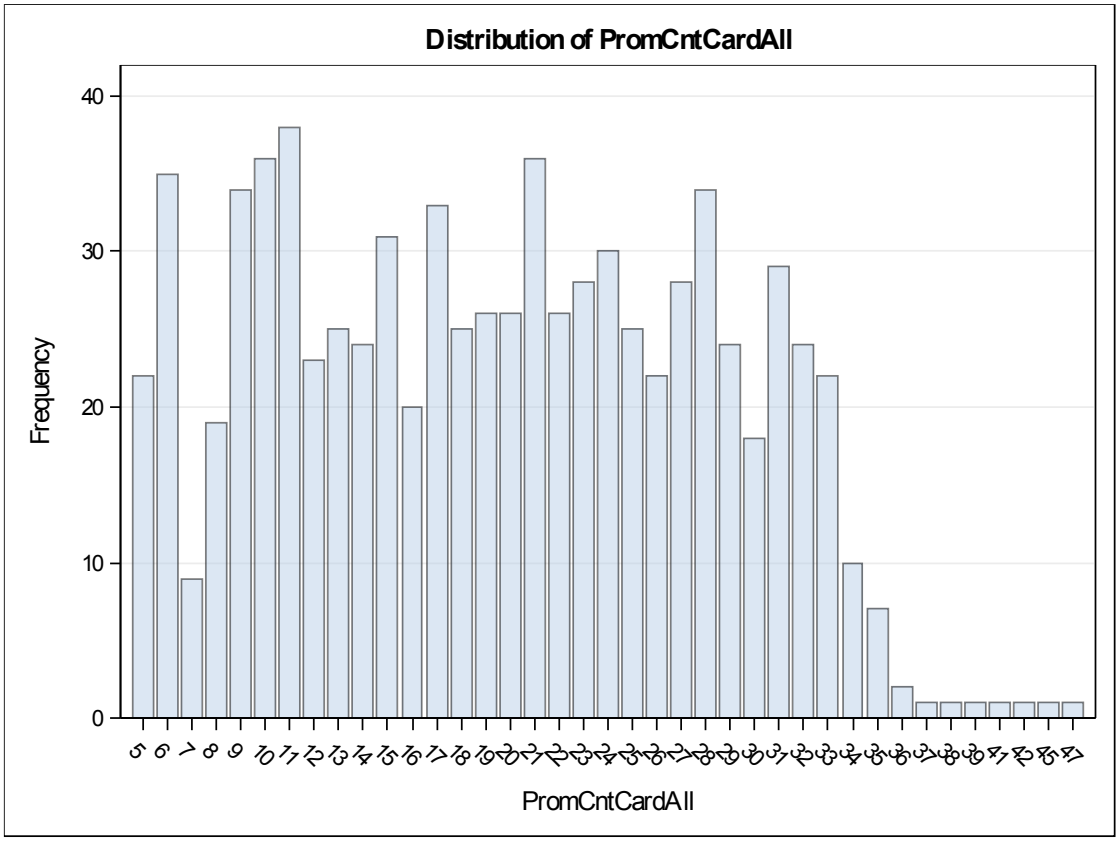

Distribution of PromCnt12
Frequency
250
200
150
100
50
0
PromCnt12


Distribution of PromCnt36
Frequency
0
20
40
60
80
PromCnt36


Distribution of PromCntAll
Frequency
0
5
10
15
20
25
PromCntAll


Distribution of PromCntCard12
Frequency
500
400
300
200
100
0
2
3
4
5
6
7
8
9
10
11
12
PromCntCard12


Distribution of PromCntCard36
Frequency
150
100
50
0
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
PromCntCard36



Ερώτημα 2:

Proc surveyselect και strata variables

Η surveyselect χρησιμοποιείται για την επιλογή δείγματος από ένα μεγάλο dataset, όπως αυτό της εργασίας. Το μέγεθος του δείγματος ισούται με το 70% του μεγέθους του καθαρισμένου dataset, άρα με 559 (από τα 798, στρογγυλοποίηση προς τα πάνω).

Για να κάνουμε stratified sampling στο εργαλείο SAS, πρέπει να χρησιμοποιηθεί το STRATA statement της surveyselect, το οποίο δέχεται μεταβλητές που εκφράζουν κατηγορία (όπως η DemHomeOwner στην δικιά μας περίπτωση). Με αυτό τον τρόπο, χωρίζουμε σε υποομάδες (**strata**) το δείγμα μας.

Το μέγεθος του κάθε stratum καθορίζεται από το ALLOC statement. Σε αυτήν την υλοποίηση επιλέχθηκε ALLOC=PROP, δηλαδή proportional allocation, που σημαίνει ότι αν (π.χ.) 60% του πληθυσμού έχει DemHomeOwner='H', τότε το 60% του δείγματος θα έχει DemHomeOwner='H'.

Ακολουθεί ο κώδικας για την δημιουργία του δείγματος.

```
/*sampling*/

title1 'Marketing campaign sampling (70% of population)';
title2 'Proportional Allocation';
proc surveyselect data=CAMP.cleaned_sorted
      n=559 out=CAMP.sample;
   strata  DemHomeOwner / alloc=prop;
run;
```

Έπειτα, γίνεται εκτέλεση της διαδικασίας proc corr για να εμφανιστούν οι συσχετίσεις όλων των μεταβλητών με τη μεταβλητή targetD, όπως φαίνεται παρακάτω:

```
proc corr data=CAMP.sample;
var DemCluster DemIncomeGroup DemMedHomeValue GiftAvg36 GiftAvgAll GiftAvgCard36
    GiftAvgLast GiftCnt36  GiftCntAll  GiftCntCard36  GiftCntCardAll  GiftTimeFirst
    GiftTimeLast PromCnt12  PromCnt36  PromCntAll  PromCntCard12  PromCntCard36
    PromCntCardAll;
    with targetD;
run;
```

Μετά την εκτέλεση του παραπάνω τμήματος προκύπτει ο παρακάτω πίνακας:

Pearson Correlation Coefficients, N = 559 Prob > |r| under H0: Rho=0						
	DemCluster	**DemIncomeGroup**	**DemMedHomeValue**	**GiftAvg36**	**GiftAvgAll**	**GiftAvgCard36**
targetD	-0.11019 0.0091	0.11320 0.0074	0.09884 0.0194	0.63304 <.0001	0.53800 <.0001	0.61434 <.0001

Pearson Correlation Coefficients, N = 559 Prob > |r| under H0: Rho=0							
	GiftAvgLast	**GiftCnt36**	**GiftCntAll**	**GiftCntCard36**	**GiftCntCardAll**	**GiftTimeFirst**	**GiftTimeLast**
targetD	0.76270 <.0001	-0.26357 <.0001	-0.18970 <.0001	-0.19870 <.0001	-0.16375 0.0001	-0.11745 0.0054	0.03967 0.3492

Pearson Correlation Coefficients, N = 559 Prob > |r| under H0: Rho=0						
	PromCnt12	**PromCnt36**	**PromCntAll**	**PromCntCard12**	**PromCntCard36**	**PromCntCardAll**
targetD	-0.02564 0.5452	-0.09256 0.0287	-0.10721 0.0112	-0.12500 0.0031	-0.14735 0.0005	-0.13456 0.0014

Από αυτόν τον πίνακα συμπεραίνουμε ότι οι μταβλητές με την μεγαλύτερη συσχέτιση με το targetD είναι όσες έχουν p-value<0.0001 και όσες είναι κοντά στο 1 κατ'απόλυτη τιμή. Αυτές οι μεταβλητές είναι οι GiftAvg36 GiftAvgAll GiftAvgCard36 GiftAvgLast GiftCnt36 GiftCntAll GiftCntCard36 και είναι αυτές που θα αποτελέσουν το μοντέλο πρόβλεψης που παρουσιάζεται παρακάτω:

```
127 proc reg data=CAMP.sample
128 outest=CAMP.estimates;
129
130 model targetD=GiftAvg36 GiftAvgAll GiftAvgCard36 GiftAvgLast
131  GiftCnt36  GiftCntAll GiftCntCard36/clm cli;
132
133 title "Regression of % targetD on sample";
134 run;
136 proc contents data=CAMP.estimates;
137 run;
```

Το αποτέλεσμα της παραπάνω εκτέλεσης φαίνεται στον παρακάτω πίνακα ANOVA:

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	7	55385	7912.07797	120.04	<.0001
Error	551	36316	65.91016		
Corrected Total	558	91701			

Root MSE	8.11851	**R-Square**	0.6040
Dependent Mean	14.47138	**Adj R-Sq**	0.5989
Coeff Var	56.10044		

Αποτελέσματα εκτέλεσης της διαδικασίας proc contents:

Data Set Name	CAMP.ESTIMATES	**Observations**	1
Member Type	DATA	**Variables**	13
Engine	V9	**Indexes**	0
Created	07/06/2015 13:33:48	**Observation Length**	136
Last Modified	07/06/2015 13:33:48	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type	EST	**Sorted**	NO
Label	Parameter Estimates and Statistics		
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	481
Obs in First Data Page	1
Number of Data Set Repairs	0
Filename	/folders/myfolders/myproject/estimates.sas7bdat
Release Created	9.0401M2
Host Created	Linux
Inode Number	21785
Access Permission	rwxrwxrwx
Owner Name	root
File Size (bytes)	131072

Alphabetic List of Variables and Attributes				
#	**Variable**	**Type**	**Len**	**Label**
6	GiftAvg36	Num	8	
7	GiftAvgAll	Num	8	
8	GiftAvgCard36	Num	8	
9	GiftAvgLast	Num	8	
10	GiftCnt36	Num	8	
11	GiftCntAll	Num	8	
12	GiftCntCard36	Num	8	
5	Intercept	Num	8	Intercept
3	_DEPVAR_	Char	13	Dependent variable
1	_MODEL_	Char	32	Label of model
4	_RMSE_	Num	8	Root mean squared error
2	_TYPE_	Char	8	Type of statistics
13	targetD	Num	8	

Ερώτημα 3:

Η ομαδοποίηση των δεδομένων με χρήση του k-means αλγόριθμου γίνεται με τη διαδικασία proc fastclus με βάση τις μεταβλητές που χρησιμοποιήθηκαν στην παραπάνω μοντελοποίηση, για k από 3 έως 7 (στον κώδικα το k ορίζεται ως maxclusters). Η εντολή maxiter=100 γίνεται για πιό ακριβές αποτέλεσμα εκτελώντας τον αλγόριθμο 100 φορές μέχρι να παραχθεί το αποτέλεσμα, ενώ γίνεται και μία ταξινόμιση του dataset με βάση το cluster για να μπορεί να τυπωθεί το αποτέλεσμα στον χρήστη.
Ο κώδικας για την ομαδοποίηση είναι ο εξής:

```
/*clustering*/
proc fastclus data=CAMP.cleaned out=CAMP.Clust
              maxclusters=7 maxiter=100;
   var GiftAvg36 GiftAvgAll GiftAvgCard36 GiftAvgLast GiftCnt36 GiftCntAll
       GiftCntCard36;
run;

proc sort data=CAMP.Clust;
by cluster;
run;
proc print;
by cluster;
run;
```

Μετά από πέντε εκτελέσεις του κώδικα αυτού για k από 3 έως 7 παρατηρήθηκε ότι καλύτερη ομαδοποίηση γίνεται για k=7, αφού γίνεται καλύτερη διαχείριση των outliers (απομόνωση σε ξεχωριστά clusters) με αποτέλεσμα να παράγεται ένα καλύτερο δείγμα δεδομένων.

Μετά από την επιλογή k=7 ομάδες επιλέγεται το cluster=7 λόγω περισσότερων παρατηρήσεων για το μοντέλο πρόβλεψης που θα δημιουργηθεί παρακάτω (Τα περισσότερα από τα υπόλοιπα clusters έχουν απομονώσει τα outliers και ως εκ τούτου δε μπορούν να χρησημοποιηθούν για μοντελοποίηση). Αυτό παρατηρείται απο τον παρακάτω πίνακα:

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	1	.	0		3	154.9
2	1	.	0		6	116.6
3	3	13.0043	33.3663		5	87.2145
4	302	3.4817	29.7666		7	18.6373
5	76	5.9202	51.5801		4	22.7041
6	1	.	0		2	116.6
7	414	4.3025	46.6018		4	18.6373

Ο κώδικας για τη συσχέτιση των μεταβλητών και τη μοντελοποίηση φαίνεται παρακάτω:

```
153 proc corr data=CAMP.Clust;
154 by cluster;
155 var DemCluster DemIncomeGroup DemMedHomeValue GiftAvg36 GiftAvgAll GiftAvgCard36
156     GiftAvgLast GiftCnt36   GiftCntAll  GiftCntCard36   GiftCntCardAll  GiftTimeFirst
157     GiftTimeLast PromCnt12  PromCnt36   PromCntAll  PromCntCard12   PromCntCard36
158     PromCntCardAll;
159     with targetD;
160 run;
```

Πίνακας συσχετίσεων:

Pearson Correlation Coefficients, N = 414 Prob > |r| under H0: Rho=0						
	DemCluster	**DemIncomeGroup**	**DemMedHomeValue**	**GiftAvg36**	**GiftAvgAll**	**GiftAvgCard36**
targetD	-0.01153 0.8151	0.02373 0.6302	0.03524 0.4746	0.40437 <.0001	0.32760 <.0001	0.39764 <.0001

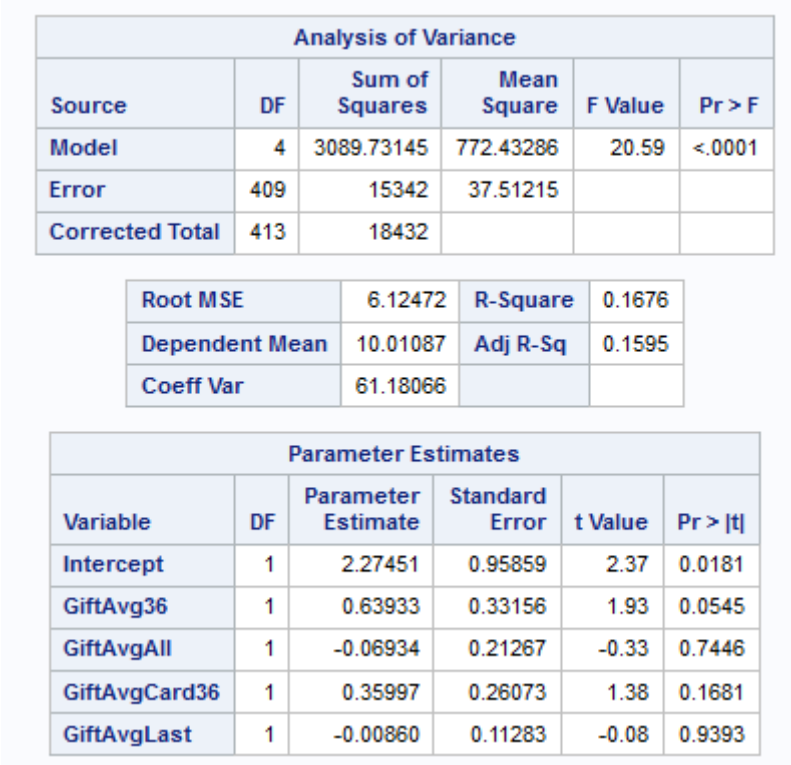
Pearson Correlation Coefficients, N = 414 Prob > |r| under H0: Rho=0							
	GiftAvgLast	**GiftCnt36**	**GiftCntAll**	**GiftCntCard36**	**GiftCntCardAll**	**GiftTimeFirst**	**GiftTimeLast**
targetD	0.27508 <.0001	-0.08317 0.0910	-0.03349 0.4967	-0.04636 0.3468	-0.03560 0.4701	0.00278 0.9550	0.04905 0.3194

Pearson Correlation Coefficients, N = 414 Prob > |r| under H0: Rho=0						
	PromCnt12	**PromCnt36**	**PromCntAll**	**PromCntCard12**	**PromCntCard36**	**PromCntCardAll**
targetD	0.04155 0.3991	0.05465 0.2672	0.05347 0.2777	0.04039 0.4124	0.06319 0.1995	0.03830 0.4370

Από τον πίνακα συμπεραίνουμε ότι οι μταβλητές με την μεγαλύτερη συσχέτιση με το targetD είναι για το λόγο που αναφέρθηκε παραπάνω οι GiftAvg36 GiftAvgAll GiftAvgCard36 GiftAvgLast και είναι αυτές που θα αποτελέσουν το μοντέλο πρόβλεψης για το cluster=7 που παρουσιάζεται παρακάτω:

```
162 proc reg data=CAMP.Clust
163 outest=CAMP.estimates2;
164 by cluster;
165 model targetD=GiftAvg36 GiftAvgAll GiftAvgCard36 GiftAvgLast /clm cli;
166     title "Regression of % targetD in clusters";
167 run;
168
169 proc contents data=CAMP.estimates2;
170 run;
```

Το αποτέλεσμα της παραπάνω εκτέλεσης φαίνεται στον παρακάτω πίνακα ANOVA:



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3089.73145	772.43286	20.59	<.0001
Error	409	15342	37.51215		
Corrected Total	413	18432			

Root MSE	6.12472	R-Square	0.1676
Dependent Mean	10.01087	Adj R-Sq	0.1595
Coeff Var	61.18066		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.27451	0.95859	2.37	0.0181
GiftAvg36	1	0.63933	0.33156	1.93	0.0545
GiftAvgAll	1	-0.06934	0.21267	-0.33	0.7446
GiftAvgCard36	1	0.35997	0.26073	1.38	0.1681
GiftAvgLast	1	-0.00860	0.11283	-0.08	0.9393

Αποτελέσματα εκτέλεσης της διαδικασίας proc contents:

Data Set Name	CAMP.ESTIMATES2	**Observations**	7
Member Type	DATA	**Variables**	11
Engine	V9	**Indexes**	0
Created	07/06/2015 13:48:17	**Observation Length**	120
Last Modified	07/06/2015 13:48:17	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type	EST	**Sorted**	NO
Label	Parameter Estimates and Statistics		
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	545
Obs in First Data Page	7
Number of Data Set Repairs	0
Filename	/folders/myfolders/myproject/estimates2.sas7bdat
Release Created	9.0401M2
Host Created	Linux
Inode Number	24591
Access Permission	rwxrwxrwx
Owner Name	root
File Size (bytes)	131072

Alphabetic List of Variables and Attributes				
#	**Variable**	**Type**	**Len**	**Label**
1	CLUSTER	Num	8	Cluster
7	GiftAvg36	Num	8	
8	GiftAvgAll	Num	8	
9	GiftAvgCard36	Num	8	
10	GiftAvgLast	Num	8	
6	Intercept	Num	8	Intercept
4	_DEPVAR_	Char	13	Dependent variable
2	_MODEL_	Char	32	Label of model
5	_RMSE_	Num	8	Root mean squared error
3	_TYPE_	Char	8	Type of statistics
11	targetD	Num	8	

Ερώτημα 4:

Η αναγωγή του παραπάνω μοντέλου στο dataset μετά τη δειγμοτοληψία γίνεται ως εξής:

```
174 proc reg data=CAMP.sample
175 outest=CAMP.estimates3;
176 model targetD=GiftAvg36 GiftAvgAll GiftAvgCard36 GiftAvgLast /clm cli;
177     title "Regression of % targetD on sample by model of cluster 6";
178 run;
```

Το αποτέλεσμα της παραπάνω εκτέλεσης φαίνεται στον παρακάτω πίνακα ANOVA:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	54451	13613	202.45	<.0001
Error	554	37250	67.23889		
Corrected Total	558	91701			

Root MSE	8.19993	R-Square	0.5938
Dependent Mean	14.47138	Adj R-Sq	0.5909
Coeff Var	56.66311		

Ερώτημα 5:

Παρατηρώντας του πίνακες ANOVA για τα δύο παραπάνω μοντέλα και ειδικότερα τιμές των πεδίων Sum of squares και R-Square τα οποία είναι μετρικές που πρέπει να λάβουμε υπόψη για να αποφασίσουμε ποιο από τα δύο μοντέλα είναι καλύτερο, συμπεραίνουμε ότι το αρχικό μοντέλο είναι καλύτερο από το δεύτερο. Το πεδίο R-square, το οποίο δείχνει πόσο καλά ταιριάζει ένα μοντέλο με την σχέση μεταβλητότητας ανεξάρτητης-εξαρτημένης μεταβλητής (το τετράγωνο της συσχέτισης), είναι ελαφρώς υψηλότερο στο πρώτο μοντέλο (r square υψηλότερο – μοντέλο καλύτερο). Το πεδίο Sum of squares όμως είναι μικρότερο στο δεύτερο μοντέλο (μικρότερο sum of squares σημαίνει καλύτερο μοντέλο). Κοιτώντας τα υπόλοιπα πεδία (γραμμή Error, το πρώτο μοντέλο έχει χαμηλότερες τιμές από το δεύτερο), αποφασίσαμε να χρησιμοποιήσουμε το πρώτο μοντέλο για την πρόβλεψη της δαπάνης.

Ερώτημα 6:

Η εφαρμογή του καλύτερου μοντέλου είναι η εφαρμογή του αρχικού μοντέλου και η πρόβλεψη γίνεται με τον παρακάτω τρόπο:

- Εισαγωγή στοιχείων του αρχείου New_Campaign.csv στο dataset CAMP. New_Campaign

```
/*Import arxeiou New_Campaign */
/*Observations=50, Variables=18*/

data CAMP.New_Campaign ;
infile "&path/New_Campaign.csv" dlm=';' dsd firstobs=2;
input targetB id_ targetD:dollar10.2 GiftCnt36 GiftCntAll GiftCntCard36
    GiftCntCardAll GiftAvgLast:dollar10.2 GiftAvg36:dollar10.2
    GiftAvgAll:dollar10.2 GiftAvgCard36:dollar10.2 GiftTimeLast
    GiftTimeFirst PromCnt12 PromCnt36 PromCntAll PromCntCard12
    PromCntCard36 PromCntCardAll StatusCat96NK $ StatusCatStarAll
    DemCluster DemAge DeMGender $ DemHomeOwner $ DemIncomeGroup
    DemMedHomeValue:dollar10.2 DemPctVeterans;
run;

proc print data=CAMP.New_Campaign;
run;
```

- Πρόβλεψη του τελικού συνολικού ποσού δαπάνης

```
/*telikh problepsh posou*/

proc score data=CAMP.New_Campaign score=CAMP.estimates out=CAMP.scored
    type=parms;
    var GiftAvg36 GiftAvgAll GiftAvgCard36 GiftAvgLast
        GiftCnt36   GiftCntAll GiftCntCard36;
    title "Problepsh telikou posou";
run;

proc print data=CAMP.estimates;
run;

data CAMP.scored1;
set CAMP.scored;
if MODEL1 = '.' then delete;
run;

proc print data=CAMP.scored1;
var id_ MODEL1;
sum MODEL1;
run;
```

Πίνακας αποτελεσμάτων μετά την εκτέλεση της διαδικασίας proc score:

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	GiftAvg36	GiftAvgAll
1	MODEL1	PARMS	targetD	8.11851	8.61576	-0.15486	-0.25077

Obs	GiftAvgCard36	GiftAvgLast	GiftCnt36	GiftCntAll	GiftCntCard36	targetD
1	0.31019	0.58109	-0.43353	-0.11673	0.38420	-1

Πίνακας αναλυτικού ποσού δαπάνης για κάθε πελάτη:

Obs	id_	MODEL1
1	11961	19.127
2	12122	22.903
3	19020	11.889
4	20522	15.734
5	23989	14.413
6	27137	7.054
7	43492	14.085
8	53576	20.591
9	63135	7.424
10	65073	11.166
11	65189	12.859
12	76165	11.111
13	77523	13.269
14	87706	13.552
15	92972	25.907
16	96548	12.539
17	97596	15.158
18	102163	20.591
19	109858	15.352
20	112172	12.828
21	115166	9.527
22	120128	14.969

Obs	id_	MODEL 1
23	124935	20.751
24	125164	39.136
25	126787	21.028
26	135265	12.778
27	137256	19.937
28	139628	18.253
29	140294	14.793
30	142838	11.154
31	151907	11.364
32	155653	18.163
33	157427	6.560
34	162494	16.874
35	168275	19.989
36	171265	20.145
37	172167	15.145
38	172750	17.958
39	173772	10.423
40	177190	11.914
41	179863	12.020
42	181337	14.882
43	182519	15.734
44	187274	15.734
45	189725	11.400
		698.185

Άρα τελικά το συνολικό ποσό δαπάνης της εταιρίας για τα δώρα των πελατών αναμένεται στα 698.185$.