Case Study 2

Xenon is an Internet marketing company. It has data related to the marketing revenues of its customers. Most of its customers plan their marketing budgets on a monthly basis. Xenon wants to understand its data better. It is looking at

creating some tables and charts to get some insight in its data.

The component of its data is as follows:-

1. Year- all the data belongs to the year 2014

2. Month - this is the month for which the data belongs

3. Quarter- Q1 to Q3 2014

4. Customer type- is it a new customer or an existing customer for the company

5. Type of calling- how did the customer engage with the company. Did the customer call in (inbound) and enquire/request for services? Or did the company call the customer (outbound) and engage with the customer?

6. Vertical- the industry to which the customer belongs

7. Monthly recurring revenue in INR- this is the amount of money the customer paid the company in a particular month. All amounts are in Indian rupees.

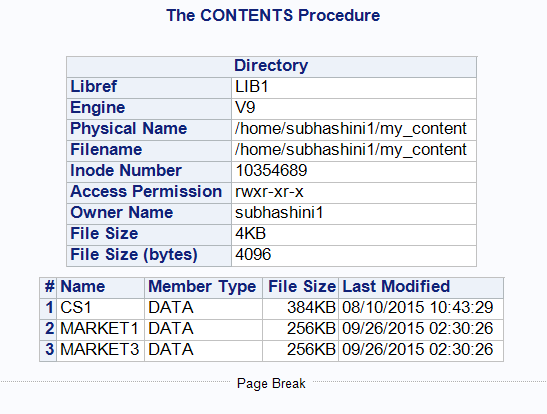
General Solution process:-

1. Define- the customer wishes to create tables and charts to understand the data better (descriptive statistics)
2. Collect-all the data required for this exercise has been given along with an understanding of the fields/variables that exist in the data
3. Organise-some basic checking for missing values and outliers is all that is required for this exercise
4. Visualise-tables and graphs have to be created to represent/summarise the data
5. Analyse-no particular statistical analysis is required
6. Insights-conclusions are to be drawn on the basis of the visualisations

# Prepare the Directory / space where you will work

libname lib1 " /home/subhashini1/my\_content "; **run**;

**PROC** **CONTENTS** data = lib1.\_ALL\_ NODS;**run**;



# Import the data

FILENAME REFFILE "/home/subhashini1/my\_content/MarketingData.csv" TERMSTR=CR;

**RUN**;

/\* Comment – I have renamed the file to REFILE so that in the Proc Import step I do not have to put the full path.\*/

**PROC** **IMPORT** DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.MARKET1;

GETNAMES=YES;

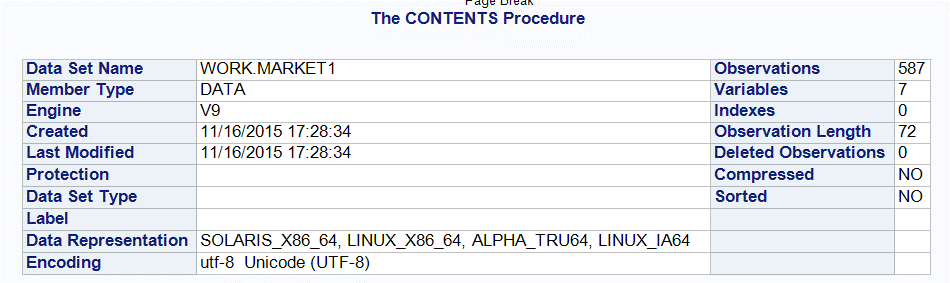
**RUN**;

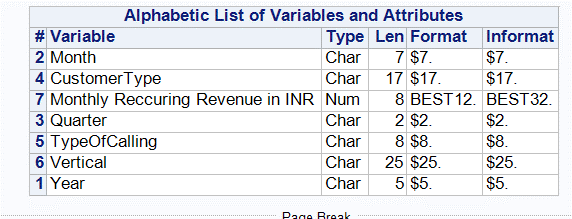
Note :- Types of delimiters can be :-

1. Csv – Comma Separator ,
2. Tab - text
3. Semi colon ;
4. Pipes |
5. Carets ^

# Understand the dimensions of the dataset - Obs and Vars – and type of data in the temporary SAS file in the work directory

**PROC** **CONTENTS** DATA=WORK.MARKET1; **RUN**;





Note :- Proc Contents procedure gives us details of the dataset in one go . If the data variables have labels / explanations / definitions , those will also be seen in the Proc Contents output .

# Save the SAS dataset back to the Permanent Directory (next time we can load in this dataset instead of the csv dataset)

**DATA** LIB1.MARKET1;

SET WORK.MARKET1; **RUN**;

Note :- For steps where no output is desired, the execution can be verified in the Log tab

47 DATA LIB1.MARKET1;

48 SET WORK.MARKET1; RUN;

NOTE: There were 587 observations read from the data set WORK.MARKET1.

NOTE: The data set LIB1.MARKET1 has 587 observations and 7 variables.

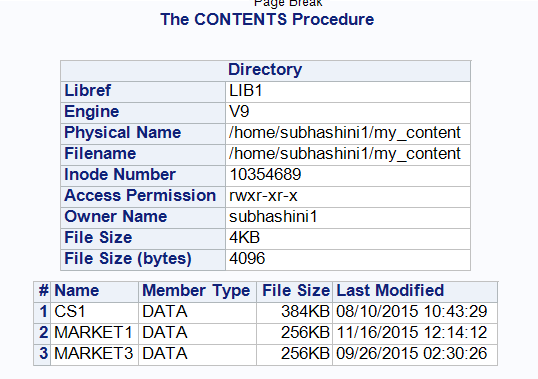
NOTE: DATA statement used (Total process time):

real time 0.05 seconds

user cpu time 0.01 seconds

# Check to confirm that the file has been saved

**PROC** **CONTENTS** data = lib1.\_ALL\_ NODS;**run**;



# Summary of Recurring Revenue

**PROC** **UNIVARIATE** DATA=WORK.MARKET1 ;

VAR 'Monthly Reccuring Revenue in INR'n; **RUN**;

Note :- the ‘n’ at the end of the variable name ensures that the blank spaces in the variable name are over-looked by the SAS system . As you will remember, SAS system does not understand blanks and any other special characters lie % : ; etc. It only understands the underscore \_ in place of blanks and other special characters.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **The UNIVARIATE Procedure** | | | | |
| **Variable:  Monthly Reccuring Revenue in INR** | | | | |
| **Moments** | | | |  |
| **N** | 586 | **Sum Weights** | 586 |  |
| **Mean** | 347768.848 | **Sum Observations** | 203792545 |  |
| **Std Deviation** | 2701402.67 | **Variance** | 7.30E+12 |  |
| **Skewness** | 13.6490561 | **Kurtosis** | 207.858855 |  |
| **Uncorrected SS** | 4.34E+15 | **Corrected SS** | 4.27E+15 |  |
| **Coeff Variation** | 776.781097 | **Std Error Mean** | 111593.916 |  |
| **Basic Statistical Measures** | | | |  |
| **Location** | | **Variability** | |  |
| **Mean** | 347768.8 | **Std Deviation** | 2701403 |  |
| **Median** | 4506 | **Variance** | 7.30E+12 |  |
| **Mode** | 0 | **Range** | 46132500 |  |
|  |  | **Interquartile Range** | 12955 |  |
| **Tests for Location: Mu0=0** | | | | |
| **Test** | **Statistic** | | **p Value** | |
| **Student's t** | **t** | 3.116378 | **Pr > |t|** | 0.0019 |
| **Sign** | **M** | 272 | **Pr >= |M|** | <.0001 |
| **Signed Rank** | **S** | 74120 | **Pr >= |S|** | <.0001 |
| **Quantiles (Definition 5)** | |  |  |  |
| **Level** | **Quantile** |  |  |  |
| **100% Max** | 46132500 |  |  |  |
| **99%** | 7152825 |  |  |  |
| **95%** | 520200 |  |  |  |
| **90%** | 188340 |  |  |  |
| **75% Q3** | 14500 |  |  |  |
| **50% Median** | 4506 |  |  |  |
| **25% Q1** | 1545 |  |  |  |
| **10%** | 500 |  |  |  |
| **5%** | 0 |  |  |  |
| **1%** | 0 |  |  |  |
| **0% Min** | 0 |  |  |  |
| **Extreme Observations** | | | |  |
| **Lowest** | | **Highest** | |  |
| **Value** | **Obs** | **Value** | **Obs** |  |
| 0 | 576 | 7789950 | 316 |  |
| 0 | 569 | 13000000 | 298 |  |
| 0 | 546 | 16275000 | 248 |  |
| 0 | 545 | 38000000 | 63 |  |
| 0 | 544 | 46132500 | 559 |  |
| **Missing Values** | | | |  |
| **Missing** | **Count** | **Percent Of** | |  |
| **Value** | **All Obs** | **Missing Obs** |  |
| . | 1 | 0.17 | 100 |  |

Inference: - at least 5% of the values are 0. These can be dropped from the analysis

Inference: - there is 1 missing value. This observation can be dropped from the dataset.

/\* Remove Missing value in Monthly Reccuring Reveue \*/

**DATA** WORK.MARKET2;

SET WORK.MARKET1;

IF 'Monthly Reccuring Revenue in INR'n = **.** THEN DELETE; **RUN**;

Log shows:-

NOTE: There were 587 observations read from the data set WORK.MARKET1.

NOTE: The data set WORK.MARKET2 has 586 observations and 7 variables.

/\*Remove 0 in Monthly Reccuring Revenue\*/

**DATA** WORK.MARKET2;

SET WORK.MARKET2;

IF 'Monthly Reccuring Revenue in INR'n = **0** THEN DELETE; **RUN**;

Log shows:-

NOTE: There were 544 observations read from the data set WORK.MARKET2.

NOTE: The data set WORK.MARKET2 has 544 observations and 7 variables.

# Create Frequency Tables

/\*frequency tables\*/

**PROC** **SORT** DATA = WORK.MARKET2;

BY CustomerType; **RUN**;

**PROC** **FREQ** DATA=WORK.MARKET2;

TABLES CustomerType; **RUN**;



Note :- Option user Proc Freq are

proc freq ;

by variables ;

exact statistic-options < / computation-options> ;

output options ;

tables requests < /options> ;

test options ;

weight variable ;

These options are described as follows:

BY calculates separate frequency or crosstabulation tables for each BY group

EXACT requests exact tests for specified statistics

OUTPUT creates an output dataset that contains specified statistics

TABLES specifies frequency or crosstabulation tables and requests tests and measures of association

TEST requests asymptotic tests for measures of association and agreement

WEIGHT identifies a variable with values that weight each observation

**PROC** **FREQ** DATA=WORK.MARKET2;

TABLES CustomerType\*Vertical; **RUN**;



/\* the table has too many outputs like Row Pct , Col Pct etc.\*/

**PROC** **FREQ** DATA=WORK.MARKET2;

TABLES Vertical\*CustomerType/norow nocol nopercent; **RUN**;

# Sort the out the table for easier understanding

**PROC** **FREQ** DATA=WORK.MARKET2 ORDER=FREQ;

TABLES Vertical\*CustomerType/norow nocol nopercent; **RUN**;



Note: - The order = option orders the values of the frequency and cross tabulation table variables according to the specified order, where:

data - orders values according to their order in the input dataset

formatted - orders values by their formatted values

freq - orders values by descending frequency count

internal - orders values by their unformatted values

**# Frequency table for Verticals**

**PROC** **FREQ** DATA=WORK.MARKET2 ORDER=FREQ;

TABLES Vertical/norow nocol; **RUN**;

| **Vertical** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| Media & Entertainment | 120 | 22.06 | 120 | 22.06 |
| Retail | 96 | 17.65 | 216 | 39.71 |
| High Technology | 65 | 11.95 | 281 | 51.65 |
| Business Services | 46 | 8.46 | 327 | 60.11 |
| Financial Services | 44 | 8.09 | 371 | 68.20 |
| Gaming | 39 | 7.17 | 410 | 75.37 |
| Public Sector | 29 | 5.33 | 439 | 80.70 |
| Manufacturing | 22 | 4.04 | 461 | 84.74 |
| Consumer Goods | 19 | 3.49 | 480 | 88.24 |
| Hotel & Travel | 18 | 3.31 | 498 | 91.54 |
| Software as a Service | 14 | 2.57 | 512 | 94.12 |
| Pharma/Health Care | 8 | 1.47 | 520 | 95.59 |
| Education | 6 | 1.10 | 526 | 96.69 |
| Automotive | 4 | 0.74 | 530 | 97.43 |
| Foundation-Not for Profit | 4 | 0.74 | 534 | 98.16 |
| Miscellaneous | 4 | 0.74 | 538 | 98.90 |
| 0 | 2 | 0.37 | 540 | 99.26 |
| Consumer Services | 2 | 0.37 | 542 | 99.63 |
| Energy & Utilities | 1 | 0.18 | 543 | 99.82 |
| Not Defined | 1 | 0.18 | 544 | 100.00 |

# There are far too many Verticals listed in the dataset. Most of them have a <5% contribution to the dataset (pls refer to the table above) . It will be desirable to club all the verticals except the top 4. Create a new variable to do so .

**DATA** WORK.MARKET3;

SET WORK.MARKET2;

IF VERTICAL = "Media & Entertainment" THEN NEWVAR=**1**;

ELSE IF VERTICAL = "Retail" THEN NEWVAR=**2**;

ELSE IF VERTICAL = "High Technology" THEN NEWVAR=**3**;

ELSE IF VERTICAL = "Business Services" THEN NEWVAR=**4**;

ELSE NEWVAR=**0**;

**RUN**;

**PROC** **FREQ** DATA=WORK.MARKET3;

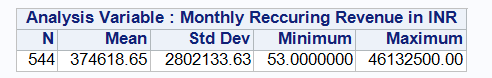
TABLES NEWVAR ; **RUN**;

| **NEWVAR** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| 0 | 217 | 39.89 | 217 | 39.89 |
| 1 | 120 | 22.06 | 337 | 61.95 |
| 2 | 96 | 17.65 | 433 | 79.60 |
| 3 | 65 | 11.95 | 498 | 91.54 |
| 4 | 46 | 8.46 | 544 | 100.00 |

# Create some tables and graphs to understand the data with respect to the continuous variable of Recurring Monthly Revenue

**PROC** **MEANS** DATA=WORK.MARKET3;

VAR 'Monthly Reccuring Revenue in INR'n; **RUN**;



Note :- The Monthly Recurring Revenues are in very large numbers

# Create a new variable to represent Monthly Recurring Revenue /1000 (in thousands)

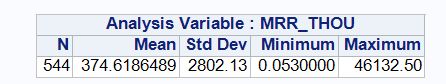
**DATA** WORK.MARKET3;

SET WORK.MARKET3;

MRR\_THOU= 'Monthly Reccuring Revenue in INR'n/**1000**; **RUN**;

**PROC** **MEANS** DATA=WORK.MARKET3;

VAR MRR\_THOU; **RUN**;

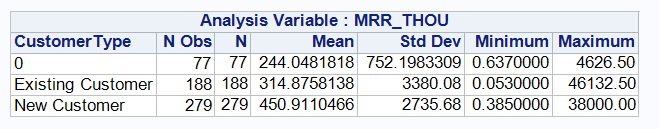


# Plots to explore sum of the new variable MRR\_THOU over the other variables

**PROC** **MEANS** DATA=WORK.MARKET3;

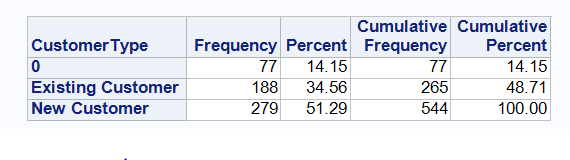
CLASS CustomerType;

VAR MRR\_THOU;**RUN**;



**PROC** **FREQ** DATA=WORK.MARKET3;

TABLES CustomerType /NOROW NOCOL;**RUN**;

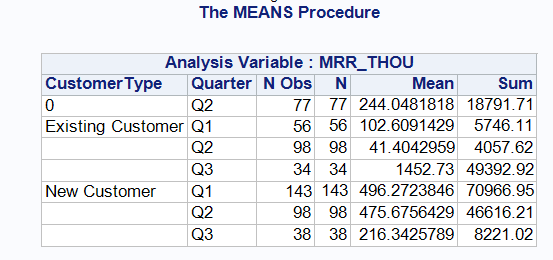
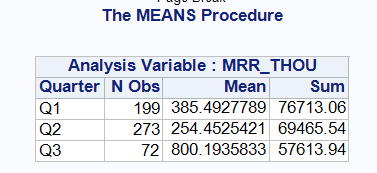


Take-away 1:- New customers are giving highest MRR and form the largest chunk of the business (51%)

**PROC** **MEANS** DATA=WORK.MARKET3 N MEAN SUM;

CLASS CustomerType Quarter ;

VAR MRR\_THOU;**RUN**;

Take-away 2:- Quarter wise MRR is showing decreasing trend. The count of New Customers is highest in Q1 and this is driving the trend.

Take-away 3:- We can see that the Average MRR for New Customer is highest.

**Assignment: - Do a similar exploration for Vertical (through the column newvar) and create some Take –away points .**