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

> setwd("H:/springer book/Case study/CaseStudy2")

> getwd()

[1] "H:/springer book/Case study/CaseStudy2"

# Import the data

market1 <- read.table("H:/springer book/Case study/CaseStudy2/MarketingData.csv", header=TRUE, sep=",", stringsAsFactors = FALSE)

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

# Understand the dimensions of the dataset - Obs and Vars

> dim(market1)

[1] 586 7

# type of data in the data frame

> str(market1)

'data.frame': 586 obs. of 7 variables:

$ Year : int 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...

$ Month : chr "January" "January" "January" "January" ...

$ Quarter : chr "Q1" "Q1" "Q1" "Q1" ...

$ CustomerType : chr "New Customer" "New Customer" "New Customer" "New Customer" ...

$ TypeOfCalling : chr "Inbound" "Inbound" "Outbound" "Inbound" ...

$ Vertical : chr "Media & Entertainment" "Education" "Manufacturing" "High Technology" ...

$ Monthly.Reccuring.Revenue.in.INR: int 0 1200 3838 8013 18618 1450 4500 8750 2700 4283 ...

# 5 number Summary of Recurring Revenue

> summary(market1$Monthly.Reccuring.Revenue.in.INR)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0 1546 4506 347800 14500 46130000

# Standard Deviation of Recurring Revenue

> sd(market1$Monthly.Reccuring.Revenue.in.INR)

[1] 2701403

> library("psych")

> describe(market1$Monthly.Reccuring.Revenue.in.INR, na.rm = TRUE, interp=FALSE,skew = TRUE, ranges = TRUE,trim=.1,type=3,check=TRUE)

vars n mean sd median trimmed mad min max range skew kurtosis

1 1 586 347768.8 2701403 4506 14927.69 5333.65 0 46132500 46132500 13.58 205.37

se

1 111593.9

Note :- The following are the Arguments in the describe function

x

A data frame or matrix

na.rm

The default is to delete missing data. na.rm=FALSE will delete the case.

interp

Should the median be standard or interpolated

skew

Should the skew and kurtosis be calculated?

ranges

Should the range be calculated?

trim

trim=.1 – trim means by dropping the top and bottom trim fraction

type

Which estimate of skew and kurtosis should be used? (See details.)

check

Should we check for non-numeric variables? Slower but helpful.

# Create Frequency Tables

> attach(market1)

> table(market1$CustomerType)

0 Existing Customer New Customer

88 191 307

> table(market1$TypeOfCalling)

Inbound Outbound

128 458

> table(market1$TypeOfCalling,market1$CustomerType)

0 Existing Customer New Customer

Inbound 26 10 92

Outbound 62 181 215

> table1<- table(market1$TypeOfCalling,market1$CustomerType)

> margin.table(table1,1)

Inbound Outbound

128 458

>

> margin.table(table1,2)

0 Existing Customer New Customer

88 191 307

# Tables with Ratios and percentages

> prop.table(table1) # cell percentages

0 Existing Customer New Customer

Inbound 0.04436860 0.01706485 0.15699659

Outbound 0.10580205 0.30887372 0.36689420

> prop.table(table1, 1)  # row percentages

0 Existing Customer New Customer

Inbound 0.2031250 0.0781250 0.7187500

Outbound 0.1353712 0.3951965 0.4694323

> prop.table(table1,2) # column percentages

0 Existing Customer New Customer

Inbound 0.29545455 0.05235602 0.29967427

Outbound 0.70454545 0.94764398 0.70032573

> table(market1$Vertical)

0 Automotive Business Services

3 4 48

Consumer Goods Consumer Services Education

19 2 6

Energy & Utilities Financial Services Foundation-Not for Profit

1 45 4

Gaming High Technology Hotel & Travel

55 66 18

Manufacturing Media & Entertainment Miscellaneous

23 138 4

Not Defined Pharma/Health Care Public Sector

1 8 30

Retail Software as a Service

97 14

# sort table to make sense

table2<-table(market1$Vertical)

> table4<-as.data.frame(table2) # convert to data frame

> View(table4)

> table5<- table4[order(-table4$Freq),]

sum(table4$Freq)

[1] 586

> table4$cfp<- table4$Freq/586

> table4

Var1 Freq cfp

1 0 3 0.005119454

2 Automotive 4 0.006825939

3 Business Services 48 0.081911263

4 Consumer Goods 19 0.032423208

5 Consumer Services 2 0.003412969

6 Education 6 0.010238908

> table5<- table4[order(-table4$cfp),]

> View(table5)

> table5

Var1 Freq cfp

14 Media & Entertainment 138 0.235494881

19 Retail 97 0.165529010

11 High Technology 66 0.112627986

10 Gaming 55 0.093856655

3 Business Services 48 0.081911263

8 Financial Services 45 0.076791809

18 Public Sector 30 0.051194539

13 Manufacturing 23 0.039249147

4 Consumer Goods 19 0.032423208

12 Hotel & Travel 18 0.030716724

20 Software as a Service 14 0.023890785

17 Pharma/Health Care 8 0.013651877

6 Education 6 0.010238908

2 Automotive 4 0.006825939

9 Foundation-Not for Profit 4 0.006825939

15 Miscellaneous 4 0.006825939

1 0 3 0.005119454

5 Consumer Services 2 0.003412969

7 Energy & Utilities 1 0.001706485

16 Not Defined 1 0.001706485

# 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 .

> market1$newvar[market1$Vertical=="Media & Entertainment"]<-1

> View(market1)

> market1$newvar[market1$Vertical=="Retail"]<-2

> market1$newvar[market1$Vertical=="High Technology"]<-3

> market1$newvar[market1$Vertical=="Gaming"]<-4

> market1[is.na(market1)]<-0 .

> table(market1$newvar)

0 1 2 3 4

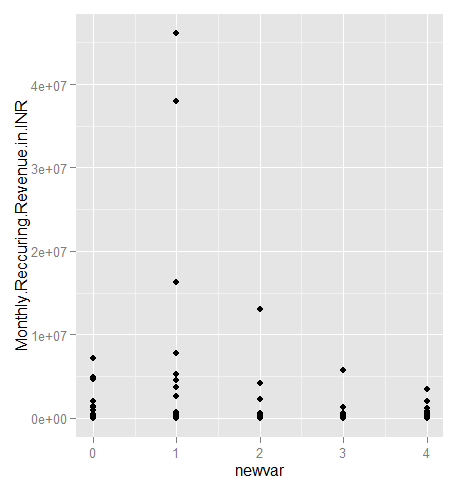
230 138 97 66 55

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

library("ggplot2")

> library("ggplot2")

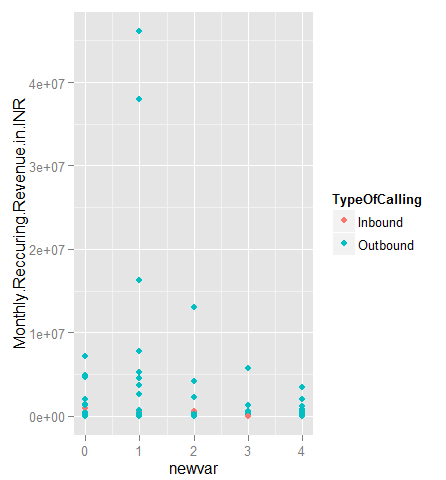
> qplot(newvar,Monthly.Reccuring.Revenue.in.INR, data = market1)



Note :- The Monthly Reccuring Revenues are in very large numbers and appear as exponential .

> market1$TypeOfCalling <- as.factor(market1$TypeOfCalling)

> qplot(newvar,Monthly.Reccuring.Revenue.in.INR, data = market1, color=TypeOfCalling)

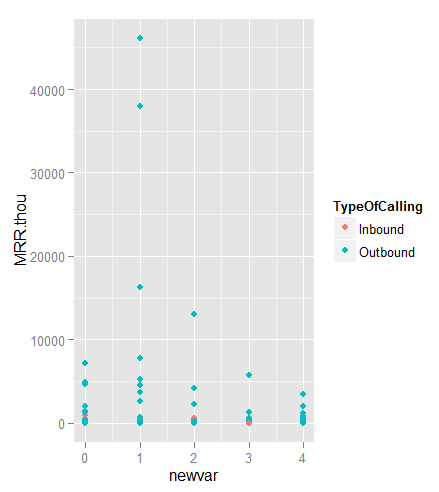


Note :- Include type of calling as a variable in the graph

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

> market1$MRR.thou<-market1$Monthly.Reccuring.Revenue.in.INR/1000

> qplot(newvar,MRR.thou, data = market1, color=TypeOfCalling)



# List of variables

> names(market1)

[1] "Year" "Month"

[3] "Quarter" "CustomerType"

[5] "TypeOfCalling" "Vertical"

[7] "Monthly.Reccuring.Revenue.in.INR" "newvar"

[9] "MRR.thou"

# Remove variables where Monthly Recurring Revenue is 0

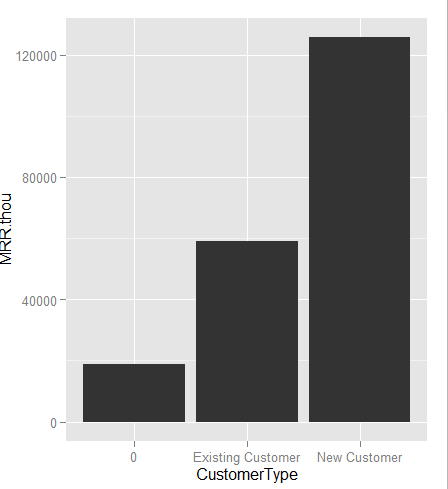
> market2<- market1[which( market1$Monthly.Reccuring.Revenue.in.INR>0),]

> dim(market2)

[1] 544 9

# Plots to explore sum of newvar (MRR in thousands) over the other variables

> ggplot(market2, aes(x=CustomerType, y=MRR.thou)) + geom\_bar(stat="identity")



> table1<- table(market2$CustomerType)

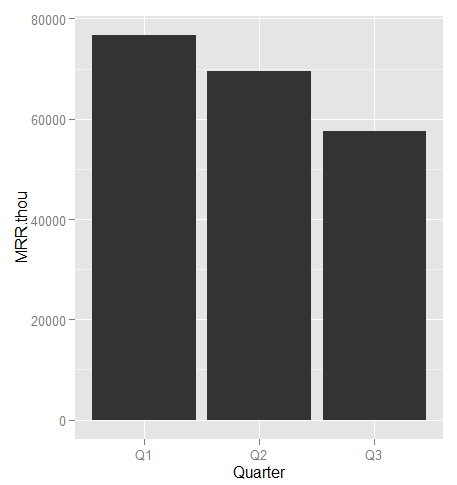
> prop.table(table1)

0 Existing Customer New Customer

0.1415441 0.3455882 0.5128676

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

> ggplot(market2, aes(x=Quarter, y=MRR.thou)) + geom\_bar(stat="identity")



> table1<- table(market2$CustomerType, market2$Quarter)

> table1

Q1 Q2 Q3

0 0 77 0

Existing Customer 56 98 34

New Customer 143 98 38

> prop.table(table1,2)

Q1 Q2 Q3

0 0.0000000 0.2820513 0.0000000

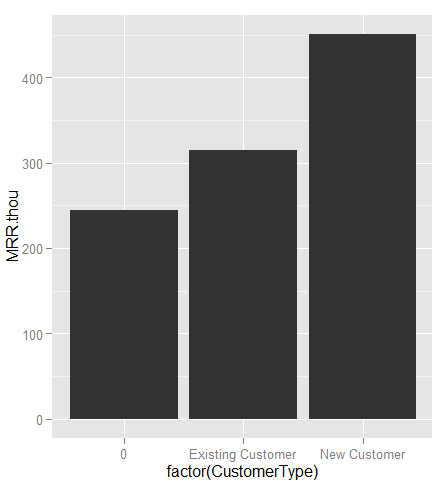
Existing Customer 0.2814070 0.3589744 0.4722222

New Customer 0.7185930 0.3589744 0.5277778

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.

# check average MRR across Customer Type

> ggplot(market2, aes(x=factor(CustomerType), y=MRR.thou)) + stat\_summary(fun.y="mean", geom="bar")



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 .**