Customer Segmentation

Based on Factor Analysis, and K-means Clustering of Customer Satisfaction Survey for a cafe

Author: Saurabh Gupta (University of Auckland)
Date: 18th October 2017

Table of Contents

Executive Summary

Our analysis of the customer survey responses identified two segments based on customer priorities.

Segment 1 - **High Income Professionals**
It constitutes 34% of all respondents.
These are customers whose primary expectations from good customer service are

- Resolution of query
- Follow through
- Clarity
- Professional Manner

Most of high income post graduates and trades people belong to this segment. They are also focussed on a professional response. Hence, we will call this segment high income professionals.

Segment 2 - Convenience Driven
It constitutes the majority i.e. 66% of the respondents.
Customers whose primary expectations are:

- Speed of picking up the phone
- Ease of getting through to someone who can help
- Treating the customer as important

These attributes reflect **Convenience** as the priority; hence, we will call this segment by that name.

Demographic Profile

Segment	Preference by Education	Preference by Income
High Income Professionals	100% of Post Graduate respondents 80% of Trade Diploma holders Together, they constituted 36% of the segment	64% respondents, with household income of **$100,000+ p.a.**
Convenience Driven	It included majority of all other education groups. 72% were from education levels 2 & 3	It included majority of all other income groups. 67% were with household income of **less than $29,999 p.a.**

Creating the segments

I tried to understand the primary expectations of customers based on how the top score on specific attributes of customer service was related to an overall top score.

For this, I used factor analysis to segregate the responses into 2 primary factors influencing the top score. As an output of Factor analysis, factor scores for each observation were added as two columns to the dataset.

Based on the scores, we assigned the observations to segments using the following rule:

If Factor Score for Segment 1 is higher, assign to Segment 1 otherwise assign to Segment 2.

The segment sizes we got as a result were meaningful:

Segment	Frequency	Percent	Cumulative Frequency
1	335	34%	335
2	655	66%	990

Discrimination Rule

As the factor analysis results were meaningful, I could proceed further with creating a criteria for segmenting future customers / prospects based on observable demographics i.e. Income and Education levels.

We used linear discrimination analysis, we could come up with a reasonable criteria for segmentation

We could correctly classify 73% respondents into segments based on income and education.

The statistical model for it has been provided in **the technical appendix**.

Checking Accuracy: Cross Validation Results of Applying the Discrimination Rule ***Predicted Segment as % of Actual***			
	Predicted Segment 1	**Predicted Segment 2**	**Total Actual**
Actual Segment 1	**40%**	60%	335
Actual Segment 2	10%	**90%**	655
Total Predicted	199	791	990
Accuracy Rate	**Correctly Classified/ 990**		**73%**

Segment Profiles (Visual)

The following chart plots the proportion of customers that gave the top score on each attribute.

Blue is for segment 1 and Red is for segment 2.

Segment 2 has the highest proportion for the rest.

Technical Appendix

Correlation Matrix

Based on Continuous Scores

Observations:
A high degree of correlation between attributes, could indicates that a high score on one of them correlates to a high score on the remaining in that group.

It could imply that customers who value an attribute, are likely to similarly value the other attributes in its group.

Pearson Correlation Coefficients, N = 990

	Q32A1	Q37A1	Q37A2	Q37A3	Q37A 4	Q37A11	Q37A12	Q37A13	Q37A14	Q37A15	Q37A16	Q37A17
	Overall Score	**PickUp Speed**	**Ease**	**Listen**	**Resolution**	**Follow Up**	**Friendly**	**Clarity**	**Knowledge**	**Importance**	**Professional**	**Response Speed**
Q32A1 Overall Score	**1.00**	0.43	0.52	0.61	0.58	0.69	0.71	0.60	0.59	0.66	0.68	0.56
Q37A1 PickUp Speed	0.43	**1.00**	0.43	0.48	0.46	0.50	0.53	0.49	0.54	0.47	0.51	0.45
Q37A2 Ease	0.52	0.43	**1.00**	0.71	0.48	0.64	0.69	0.74	0.65	0.65	0.56	0.48
Q37A3 Listen	0.61	0.48	0.71	**1.00**	0.48	0.67	0.70	**0.78**	0.73	0.71	0.66	0.48
Q37A4 Resolution	0.58	0.46	0.48	0.48	**1.00**	0.59	0.66	0.53	0.55	0.51	0.62	0.68
Q37A11 FollowUp	0.69	0.50	0.64	0.67	0.59	**1.00**	**0.91**	**0.79**	0.72	0.71	**0.90**	0.71
Q37A12 Friendly	0.71	0.53	0.69	0.70	0.66	**0.91**	**1.00**	**0.77**	0.72	**0.75**	**0.88**	0.74
Q37A13 Clarity	0.60	0.49	0.74	**0.78**	0.53	**0.79**	**0.77**	**1.00**	0.75	0.72	0.74	0.54
Q37A14 Knowledge	0.59	0.54	0.65	0.73	0.55	0.72	0.72	0.75	**1.00**	0.68	0.69	0.54
Q37A15 Importance	0.66	0.47	0.65	0.71	0.51	0.71	**0.75**	0.72	0.68	**1.00**	0.71	0.57
Q37A16 Professional	0.68	0.51	0.56	0.66	0.62	**0.90**	**0.88**	0.74	0.69	0.71	**1.00**	0.74
Q37A17 Response Speed	0.56	0.45	0.48	0.48	0.68	0.71	0.74	0.54	0.54	0.57	0.74	**1.00**

Overall score is related to:

- Listening
- Follow up
- Friendly
- Importance
- Professional

In the matrix, we can identify that the following groups of attributes have a high correlation.

Group 1:

- Follow up
- Friendly
- Professionalism

Group2:

- Clarity
- Listen
- Knowledge

We will use Factor Analysis, to identify if these can form the basis for our segments.

Factor Analysis

Cluster analysis could not discriminate well between segments so I used Factor Analysis (for reference, cluster analysis results are given later in this appendix).

Factor analysis discriminated better.

The scores were converted to binary format with New Score = 1 if Score = 9 or 10 (otherwise new score = 0). It will indicate, customers who are most satisfied with an aspect of the service.

We will assume that these factors were the primary expectations of the respondents.

The table (Rotated Factor Pattern) on the next page indicates 2 customer segments as below.

Segment 1: Customers whose primary expectations from good customer service are

- Resolution
- Follow Up
- Clarity
- Professionalism

These attributes constitute **Professionalism**; hence, we will call this segment by that name.

Segment 2: Customers whose primary expectations are:

- Pick up Speed
- Ease of getting through
- Treating customer as important

These attributes reflect **Convenience**; hence, we will call this segment by that name.

Primary Customer Expectations by Segment:

Rotated Factor Pattern		Factor1 Professionalism	Factor2 Convenience
Q32A1	**Most Satisfied**	0.67	.
Q37A1	**Pick Up Speed**	.	0.80
Q37A2	**Ease**	.	0.76
Q37A3	**Listen**	0.69	0.54
Q37A4	**Resolution**	0.83	.
Q37A11	**Follow Up**	0.79	.
Q37A12	**Friendly**	0.75	0.54
Q37A13	**Clarity**	0.78	.
Q37A14	**Knowledge**	0.65	.
Q37A15	**Importance**	.	0.73
Q37A16	**Professional**	0.78	.
Q37A17	**Response Speed**	0.77	.
Values less than 0.5 are not printed.			

Creating Segments

As an output of Factor analysis, factor scores for each observation were added as two columns to the dataset.

Based on the scores, we assigned the observations to segments using the following rule:

If Factor Score for Segment 1 is higher, assign to Segment 1 otherwise assign to Segment 2.

The segment sizes we got as a result are meaningful as well:

Segment	**Frequency**	**Percent**	**Cumulative Frequency**
1	335	34%	335
2	655	66%	990

Demographics vs. Segments

Demographic data is more objective and easier to collect than customer expectations. Hence, we will see if demographic profiles can help us identify how to segment our customers.

Diagnostics

I will check if the 2 demographic variables Income and Education are statistically capable of discriminating and if yes, to what extent.

First, we will look at a simple frequency distribution of demographics by segments.

The discrimination seems moderate with 100% of those with a Post Graduate Degree (Education level 7) or 80% of those with a Trade Certificate (level 5) falling in segment 1.

It becomes clearer with Income Level. 64% of respondents with an income of $100k+ fall in Segment 1.

Segment by Education (q77)								
Segment 1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
Frequency	16.0	106.0	46.0	30.0	86.0	16.0	35.0	335.0
Percent	21%	30%	17%	27%	***80%***	43%	***100%***	
Segment 2								
Frequency	61.0	246.0	223.0	82.0	22.0	21.0	0.0	655.0
Percent	***79%***	***70%***	***83%***	***73%***	20%	57%	0%	
Total	77.0	352.0	269.0	112.0	108.0	37.0	35.0	990.0

Segment by Income (q84)								
Segment 1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
Frequency	83.0	65.0	39.0	51.0	14.0	4.0	79.0	335.0
Percent	23%	29%	37%	44%	37%	24%	***64%***	
Segment 2								
Frequency	283.0	159.0	66.0	66.0	24.0	13.0	44.0	655.0
Percent	***77%***	***71%***	***63%***	56%	***63%***	***76%***	36%	
Total	366.0	224.0	105.0	117.0	38.0	17.0	123.0	990.0

The correlation matrix between the 2 demographic variables Education and Income, indicate that they are not independent so many not be very good at discriminating. However, as correlation is less than 0.5, we will proceed with our analysis.

PEARSON CORRELATION COEFFICIENTS, N = 990

	Education (Q77)	Income (Q84)
EDUCATION (Q77)	1	0.39277
INCOME (Q84)	0.39277	1

Discriminant Analysis

To assess more accurately, we will use discriminant analysis on segments versus demographics.

74% of customers falling in segment 2 could be correctly classified compared to 57% for segment 1. This estimate seems reasonable given that we have only 2 demographic variables – income and education. While what we are trying to estimate is much more intangible and varied.

Higher the education level or income level, higher the chances that the customer lies in Segment 1. These customers are looking more for professionalism and speedy resolution.

The intercept for Segment 2 is much smaller. It indicates that those with a lower income or educational level are most likely to be in Segment 2. These customers are looking more for convenience and being paid importance to.

Discrimination Rule

To determine the likely segment of new customers based on income and education we can use the linear discrimination function below and apply the following rule.

Calculate the following expressions:

For Segment 1 use values in the first column:
Seg1 = w0 * + w1 * (Education Level) + w2 * (Income Level)

For Segment 2 use values in the second column:
Seg2 = w0 * + w1 * (Education Level) + w2 * (Income Level)

Classify the customer into the segment which has a higher value i.e.
If Seg1 > Seg2, classify as Segment 1, otherwise Segment 2.

Linear Discriminant Function for Segment			
Variable	**Label**	**Seg 1**	**Seg 2**
Constant (w0)		-5.28	-2.57
Q77 (w1)	Education	1.75	1.29
Q84 (w2)	Income	0.52	0.33

Accuracy: Cross Validation Results *Predicted Segment as % of Actual*			
	Predicted		
Actual	**1**	**2**	**Total**
1	**40%**	60%	335
2	10%	**90%**	655
Total	199	791	990
Priors	0.34	0.66	

Misclassification Rate			27%

Cluster Analysis of C Sat survey

K-means clustering – 2 clusters with continuous values

Summary – Classified groups into Most dissatisfied and Satisfied

K-means clustering was run using proc fastclus – once for 2 clusters and second time for 3 clusters.

- Logic for 2 clusters – Most Satisfied vs Rest
- Logic for 3 clusters – Most Satisfied vs Most Dissatisfied vs Rest

It was run for 10 iterations.

Convergence criterion was satisfied in both runs.

Observations from 2 clusters:

- Dissatisfied Customers were classified into Cluster 1 (86% - 100% of people who gave a score of 1 or 3)
- Satisfied Customers were classified into Cluster 2 (90-100% of people who gave scores of 7 to 10)
- However, those giving a score of 5 or 6 couldn't be classified accurately. In fact, a higher proportion (69%) of customers giving a score of 5 were classified as Satisfied compared to those who gave a score of 6 (50%)

SAS code – checking classification done by 2 clusters:

```
proc freq data = clusters;
tables cluster*q32a1;
run;
```

Output:

Overall Score (Q32A1) by Cluster

Cluster 1	**1**	**3**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Frequency	25	24	9	10	0	27	0	21	116
Percent	3	2	1	1	0	3	0	2	12
Row Pct	22	21	8	9	0	23	0	18	
Col Pct	86	100	31	50	0	9	0	5	
Cluster 2	**1**	**3**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Frequency	4	0	20	10	40	258	135	407	874
Percent	0	0	2	1	4	26	14	41	88
Row Pct	0	0	2	1	5	30	15	47	
Col Pct	14	0	69	50	100	91	100	95	
Total	29	24	29	20	40	285	135	428	990
	3	2	3	2	4	29	14	43	100

K-means clustering – 3 clusters with continuous values

Observations from 3 clusters:

- The middle were more accurately classified with 3 clusters.

SAS code – checking classification done by 3 clusters:

```
proc freq data = clusters;
tables cluster*q32a1;
run;
```

Output:

Overall Score (Q32A1) by Cluster

Cluster	1	3	5	6	7	8	9	10	Total
Cluster 1	**1**	**3**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Frequency	25	24	4	10	0	1	0	21	85
Percent	3	2	0	1	0	0	0	2	9
Row Pct	29	28	5	12	0	1	0	25	
Col Pct	***86***	***100***	14	50	0	0	0	5	
Cluster 2	**1**	**3**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Frequency	4	0	0	0	0	98	116	380	598
Percent	0	0	0	0	0	10	12	38	60
Row Pct	1	0	0	0	0	16	19	64	
Col Pct	14	0	0	0	0	34	***86***	***89***	
Cluster 3	**1**	**3**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Frequency	0	0	25	10	40	186	19	27	307
Percent	0	0	3	1	4	19	2	3	31
Row Pct	0	0	8	3	13	61	6	9	
Col Pct	0	0	***86***	***50***	***100***	***65***	14	6	
Total	**29**	**24**	**29**	**20**	**40**	**285**	**135**	**428**	**990**
	3	**2**	**3**	**2**	**4**	**29**	**14**	**43**	**100**

SAS Code used

```
** set working directory;

libname hold 'H:\STATS 747\Assignment4';

** import csat data;

PROC IMPORT OUT= work.csat
            DATAFILE= "BinaryData.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

** check data and distribution;

** import into work;

 data csat;
 set hold.csat;
 run;

proc contents data = csat;
run;

** check data;
proc freq data = csat;
      table
Q32: q77 q84 q37: ;
run;

** assign labels **;

data csat;
  set csat;
  label
      q32a1 = "MostSatisfied"
      q37a1 = "PickUpSpeed"
      q37a2 = "Ease"
      q37a3 =  "Listen"
      q37a4 =  "Resolution"
      q37a11 =  "FollowUp"
      q37a12 =  "Friendly"
      q37a13 =  "Clarity"
      q37a14 =  "Knowledge"
      q37a15 =  "Importance"
      q37a16 =  "Professional"
      q37a17 =  "Response Speed"

      q77 = "Education"
      q84 = "Income"
 ;
 run;


** check correlation;

ods graphics on;
```

```
 PROC CORR data= csat outp= csatcorr NOPROB;
 var q32a1 q37: ;
  run;
ods graphics off;


*** Factor Analysis because clustering wasn't discriminating well between
the data;

proc factor data = csat out = csat nfact = 2
rotate = varimax fuzz = 0.5;
var q32a1 q37: ;
run;


******* Segment based on Factor Scores;

data csat;
 set csat;
      maxfct=max(of factor1-factor2);
      seg=1;
      if factor2=maxfct then seg=2;
RUN;

***** Check Segment sizes;

proc freq data = csat;
      table seg ;
run;

********** data for spider chart;

proc tabulate data = hold.Csat;
class seg;
var q32a1 q37a1--q37a17 ;

table q32a1 q37a1--q37a17, mean * seg;

run;



******** check variation in education by segment;

proc freq data = csat;
tables seg * q77;
run;

******** check variation in income by segment;

proc freq data = csat;
tables seg * q84;
run;

******* check how education varies by income;

proc freq data = csat;
tables q77 * q84;
run;

**** check correlation between education and income;
```

```
ods graphics on;
 PROC CORR data= csat outp= incedu NOPROB;
 var q77 q84;
  run;
ods graphics off;


******* discrimanate - default priors 0.5, 0.5 ;

proc discrim data = csat outstat=outdisc method = normal pool=yes list
crossvalidate;

class seg;

var q77 q84;

run;

******* discrimanate - proportional priors ;

proc discrim data = csat outstat=outdisc method = normal pool=yes
crossvalidate;

class seg;

var q77 q84; priors prop;

run;

******* discrimanate - quadratic and proportional priors ;

proc discrim data = csat outstat=outdisc method = normal pool=no
crossvalidate;

class seg;

var q77 q84 ; priors prop;

run;


************ code not used for final result because clustering wasn't
useful;



** k means cluster - do this many times and find the most most
useful solution - change random seed (below it's 456);

 ** 3 clusters - Most Satisfied, Neutral and Least Satisfied; **
maxiter=10;

proc fastclus data=csat maxc=3 replace=random random=747 out=clusters
maxiter=10;
   var q37: ;
run;

** check how good the classification is;

proc freq data = clusters;
```

```
tables cluster*q32a1;
run;


** 2 clusters - Most Satisfied and Least Satisfied; ** maxiter=10;

proc fastclus data=csat maxc=2 replace=random random=747 out=clusters3
maxiter=10;
   var q37:;
run;

** check how good the classification is;

proc freq data = clusters3;
tables cluster*q32a1;
run;

** results are not encouraging because 9.5% of those who gave an overall
score of 8 were wrongly classified in Cluster 1;



* try ward's min var ;

proc cluster data=csat method=wards standard outtree=treedat pseudo;
      var q37:;
  run;

** build the tree ;

  proc tree  data=treedat;
run;

proc tree data = treedat nclusters=2 out=outclus;
run;

** sort the data by cluster;

proc sort data =outclus;
      by cluster;


proc means data =outclus mean;
      by cluster;
      var q37a1 ;
run;

proc freq data = csat;
      table q37a1 ;
run;
```