Exercise 1

Moises Marin

Online CRN 23804
A20349918

mmarinm@okstate.edu

1/27/2022

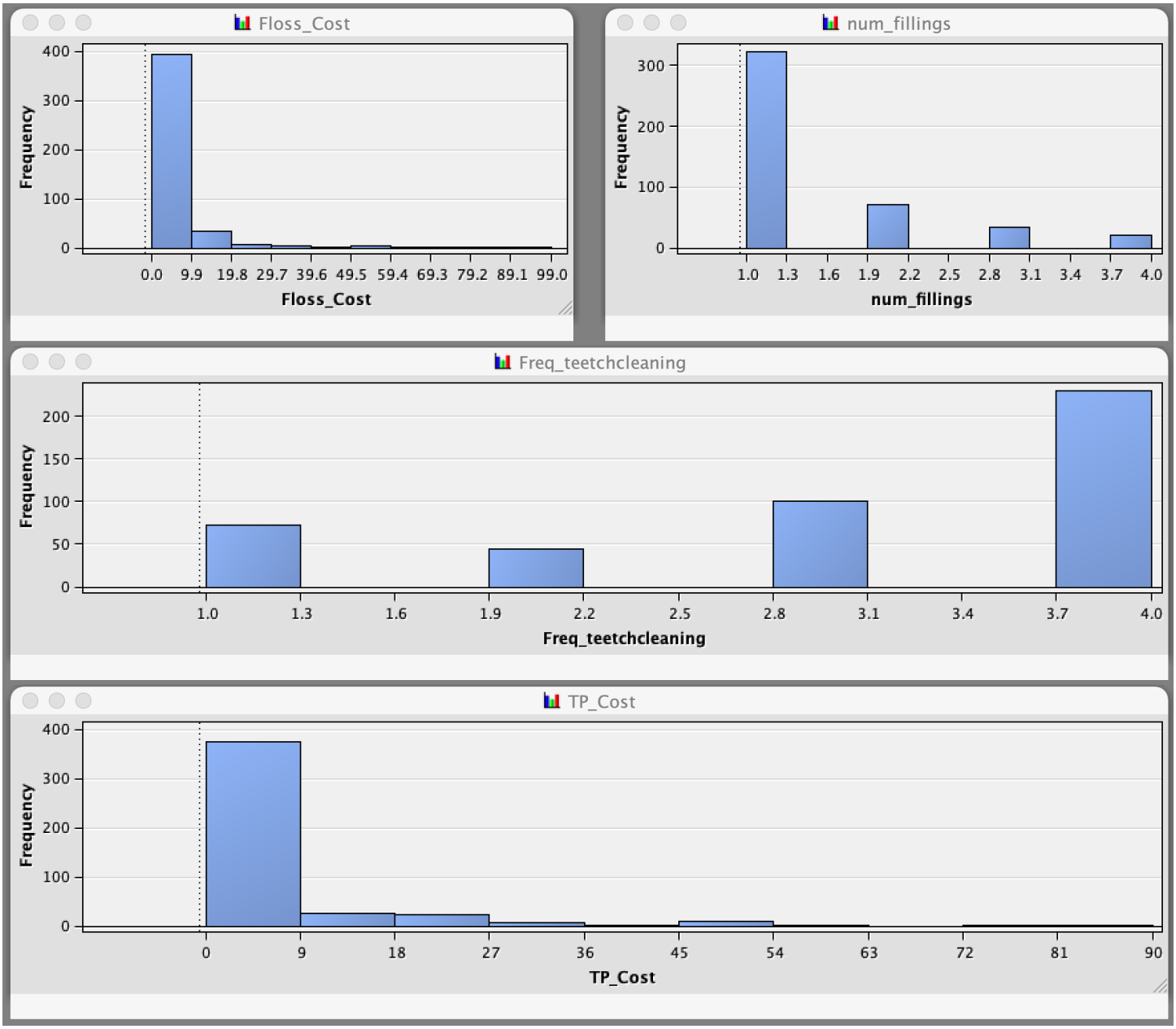
—

BAN 5743

Predictive Analytics

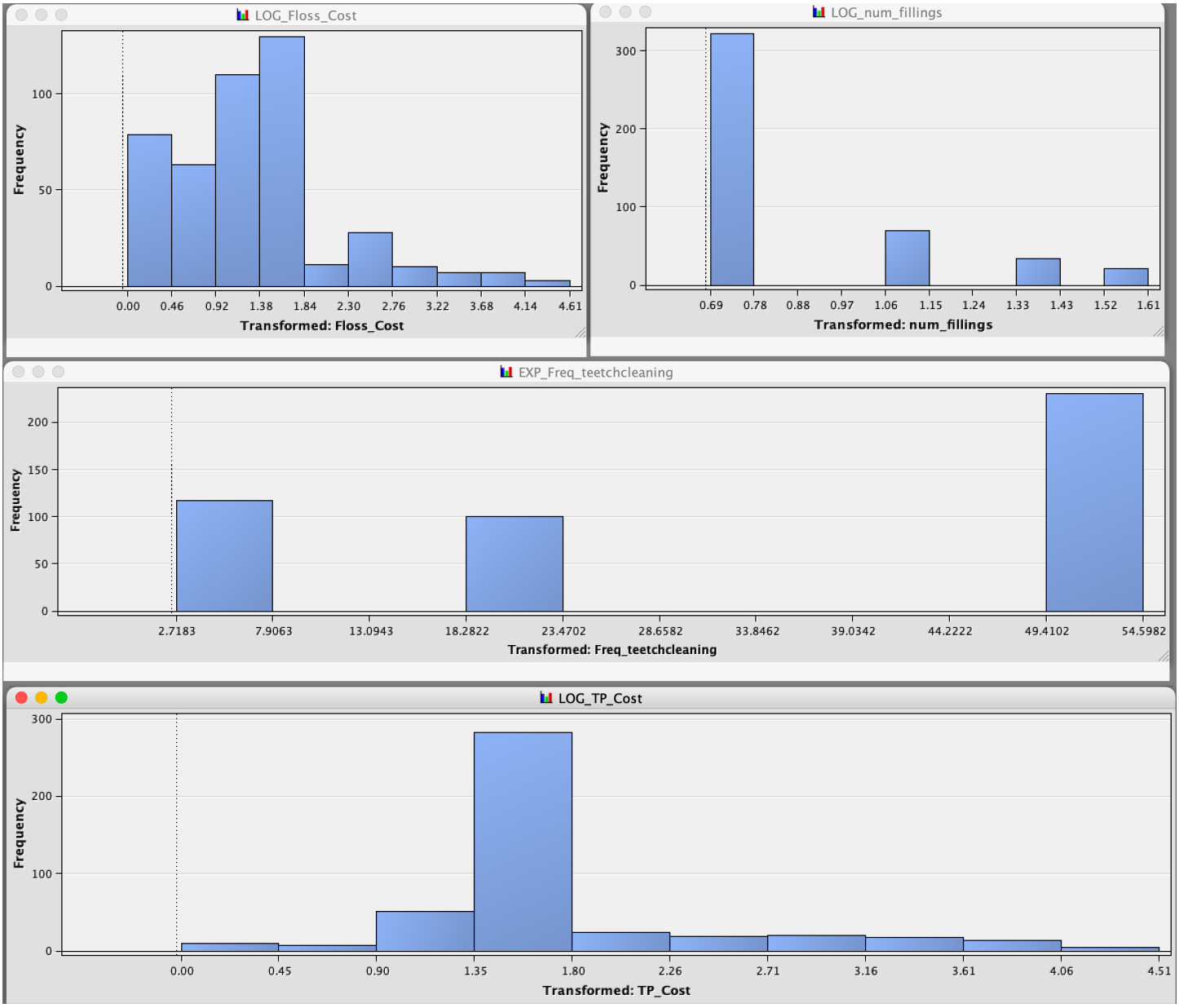
1. Compare and contrast the original variables versus the transformed variables. (1 point)

This is how the variables look before the transformation, i.e. the original variables



This is how the variables look after the transformation.

We can see that **Floss_cost** has a distribution that looks closer to a normal distribution than the original values, **num_fillings** and **Freq_teetchcleaning** don't look very different than the original values. TP_Cost has reduced the right skewness considerably.



The transformation statistics confirm the skewness change in the variables, for example TP_Cost changed from 3.8 (right skewed) to 1.2 (close to normal distribution) after log transformation.

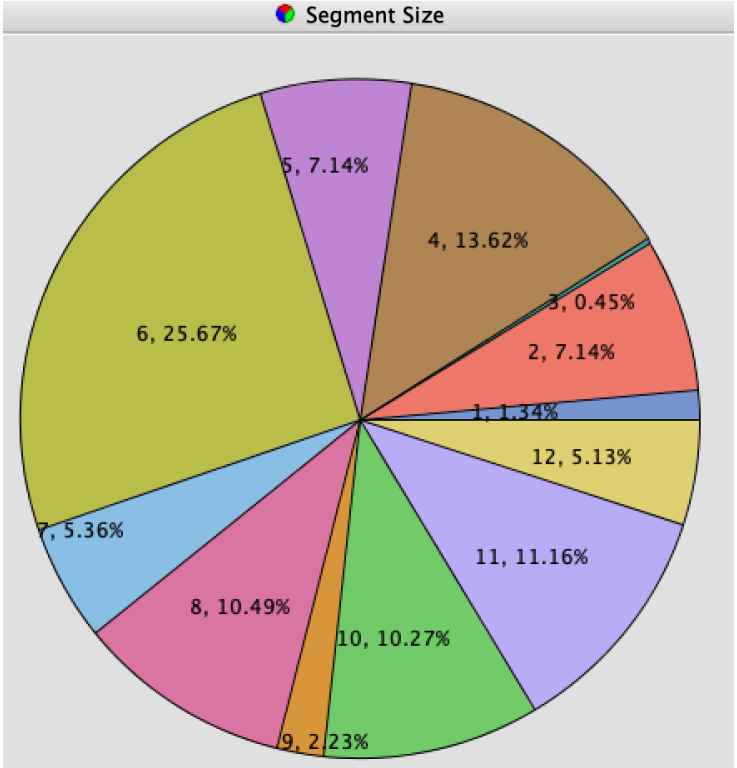
Transformations Statistics

Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	Floss_Cost		.	448	0	0	99	4.800312	9.903337	5.421592	35.7348	Floss_Cost
Input	Original	Freq_teetchcleaning		.	448	0	1	4	3.09375	1.119346	-0.87788	-0.70371	Freq_teetc...
Input	Original	TP_Cost		.	448	0	0	90	7.532879	11.35998	3.816792	16.86943	TP_Cost
Input	Original	num_fillings		.	448	0	1	4	1.448661	0.825786	1.828705	2.373855	num_fillings
Output	Computed	EXP_Freq_teetchcleaning	exp(Freq_teetchcleaning)	.	448	0	2.718282	54.59815	33.81462	22.16862	-0.24347	-1.74422	Transform...
Output	Computed	LOG_Floss_Cost	log(Floss_Cost + 1)	.	448	0	0	4.60517	1.25153	0.878245	0.807751	1.374649	Transform...
Output	Computed	LOG_TP_Cost	log(TP_Cost + 1)	.	448	0	0	4.51086	1.768165	0.737807	1.256471	2.486503	Transform...
Output	Computed	LOG_num_fillings	log(num_fillings + 1)	.	448	0	0.693147	1.609438	0.852057	0.27645	1.495521	0.917734	Transform...

The transformation statistics confirm the skewness change in the variables, for example TP_Cost changed from 3.8 (right skewed) to 1.2 (close to normal distribution) after log transformation.

2. How many clusters were selected by SAS Enterprise Miner? What are their relative sizes? (0.5 points)

There are 12 segments, the following image shows their relative sizes in percentage.

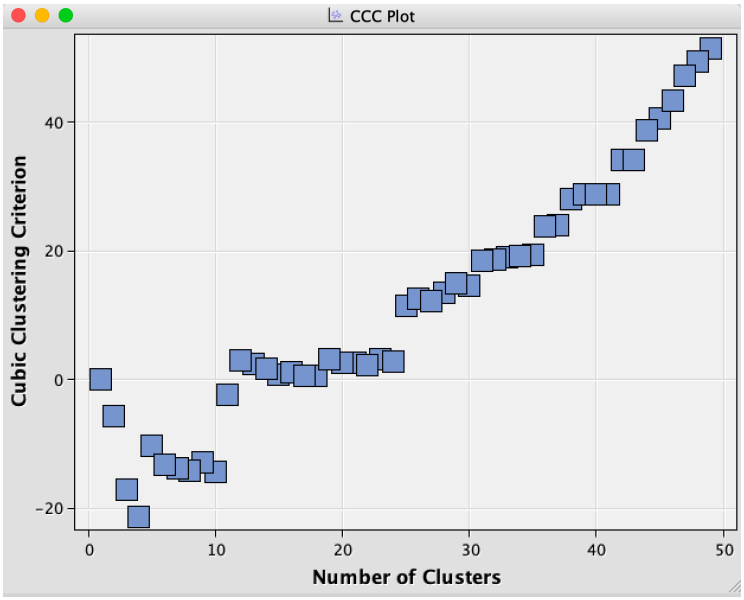


3. Look at the cluster history and the CCC plot and comment on the selection of the number of clusters by SAS Enterprise Miner. What other possible cluster solutions are being suggested by the cluster history or CCC plot? (1 point)

There were 50 clusters created before reaching to the final one, with the value of cubic clustering criterion going all the way to 51.57.

Cluster History

Number of Clusters	Clusters Joined		Freq	Pseudo F Statistic	Pseudo t-Squared	Norm RMS Distance	Tie
49	OB13	OB16	31	3273	33.8	0.1306	
48	CL49	OB35	36	2987	23.9	0.142	
47	OB17	OB26	16	2740	48.3	0.1491	
46	OB12	OB30	26	2316	96.5	0.1685	
45	OB23	OB38	80	2055	114	0.175	
44	OB15	OB36	14	1907	61.8	0.1807	
43	OB18	OB21	52	1552	205	0.1834	
42	OB11	OB29	2	1574	.	0.1877	
41	OB7	CL48	57	1237	119	0.1955	
40	OB20	OB28	5	1250	9.3	0.2027	
39	OB42	OB49	2	1271	.	0.2088	
38	OB4	OB24	22	1234	112	0.2125	
37	CL45	OB32	92	1039	169	0.2142	
36	OB25	OB34	6	1045	36.3	0.2171	
35	CL41	OB8	81	859	76.5	0.2219	
34	OB10	CL42	19	865	44.2	0.2276	
33	OB1	CL37	94	868	13.3	0.2407	
32	CL38	OB46	25	861	23.9	0.244	
31	OB22	OB44	11	867	123	0.2471	
30	CL35	CL46	107	736	54.9	0.2489	
29	OB41	OB45	3	756	286	0.2624	
28	CL44	OB33	26	716	72.6	0.271	
27	OB5	CL47	27	681	99.6	0.2732	
26	OB9	OB48	3	702	32.5	0.2886	
25	OB19	OB31	28	680	273	0.2958	
24	CL33	CL30	201	462	194	0.2959	
23	CL32	OB43	27	475	14.7	0.3137	
22	CL23	CL31	38	465	41.3	0.3237	
21	CL34	CL39	21	479	28.2	0.347	
20	CL40	OB27	14	485	95.6	0.3609	
19	OB39	OB47	2	509	.	0.3804	
18	CL28	CL25	54	457	91.1	0.4063	
17	OB6	CL21	25	464	37.4	0.4131	
16	OB2	CL18	56	485	5.7	0.4468	
15	CL22	OB14	43	489	27.0	0.4533	
14	CL20	OB37	15	520	5.3	0.4728	
13	CL43	OB50	54	544	68.9	0.5016	
12	CL26	CL36	9	574	26.7	0.5156	
11	CL27	CL13	81	457	292	0.5636	
10	CL24	CL11	282	261	354	0.6169	
9	OB3	CL15	45	287	15.1	0.6343	
8	CL16	CL17	81	276	116	0.6439	
7	CL10	OB40	293	288	40.7	0.6862	
6	CL9	CL14	60	305	85.5	0.7232	
5	CL8	CL29	84	365	15.6	0.873	
4	CL7	CL5	377	207	356	0.935	
3	CL6	CL12	69	279	46.7	0.9826	
2	CL4	CL19	379	518	19.0	1.5484	
1	CL2	CL3	448	.	518	1.5822	



4. Which are the four most important base variables in this cluster solution? (1 point)

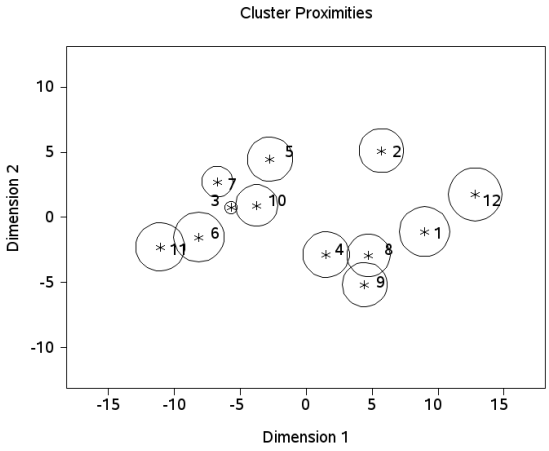
The following image shows that the four most important varaibles are F_aftermeal, Floss, F_wakingup and Brush.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
F_aftermeal	F_aftermeal	4	5	1.00000
Floss	Floss	0	7	0.88434
F_wakingup	F_wakingup	3	2	0.83581
Brush	Brush	0	5	0.77471

5. Do the cluster centers seem to be separated in the multidimensional space? (1 point)

This is how the centers of the cluster look like, they don't appear to be very separated in two dimensional space.



6. How do the important variables compare between the two cluster sets? (1 point)

The variable **F_aftermeal** remained the most important in both clusters, in the cluster 2ith 6 segments the second most important variable is **LOG_Floss_cost** which differs from the previous cluster that had **Floss** in that position. **Brush** remained in the 4th position in both clusters, the rest of variables have a different order.

Cluster with 6 segments

Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
F_aftermeal	F_aftermeal	2	2	1.00000
LOG_Floss_Cost	Transformed: Floss_Cost	0	4	0.92261
Floss	Floss	0	3	0.81997
Brush	Brush	0	2	0.69007
F_beforesleep	F_beforesleep	1	0	0.64900
Salary	Salary	0	2	0.61867
F_wakingup	F_wakingup	1	0	0.55431
B_beforesleep	B_beforesleep	0	1	0.54507
B_wakingup	B_wakingup	1	0	0.48698
F_anothertime	F_anothertime	0	1	0.48339
B_aftermeal	B_aftermeal	1	1	0.45916
EXP_Freq_teetchcleaning	Transformed: Freq_teetchcleaning	0	1	0.44845
LOG_TP_Cost	Transformed: TP_Cost	0	2	0.44315
YRS_Current_Position	YRS_Current_Position	0	1	0.44286
Birth_Year	Birth_Year	0	1	0.44200
Age_in_Years	Age_in_Years	0	1	0.44200
YRS_Current_Employer	YRS_Current_Employer	0	1	0.13859

Previous cluster with 12 segments

Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
F_aftermeal	F_aftermeal	4	5	1.00000
Floss	Floss	0	7	0.88434
F_wakingup	F_wakingup	3	2	0.83581
Brush	Brush	0	5	0.77471
LOG_Floss_Cost	Transformed: Floss_Cost	0	4	0.75105
LOG_TP_Cost	Transformed: TP_Cost	0	5	0.72954
B_beforesleep	B_beforesleep	1	1	0.64653
Age_in_Years	Age_in_Years	0	5	0.61218
B_wakingup	B_wakingup	1	1	0.54541
F_beforesleep	F_beforesleep	1	0	0.51802
B_aftermeal	B_aftermeal	1	1	0.51156
LOG_num_fillings	Transformed: num_fillings	0	1	0.42820
F_anothertime	F_anothertime	1	0	0.34090
EXP_Freq_teetchcleaning	Transformed: Freq_teetchcleaning	0	1	0.29333

7. Comparing the information from questions 1-5 above to the new cluster set. Which cluster set would you recommend to the dental association (3 clusters or 6 clusters)? (1 point)

I would recommend the first cluster because it does not have a fixed number of expected segments, imposing such a restriction affects the iterative process that looks for a better solution.

8. Which is the most important variable for segment 2? Which is the most important variable for segment 3? (0.5 point)

These are the most important variables for segment 2, there is no segment three.

```
Variable: _SEGMENT_ Segment: 2 Count: 32
Decision Tree Importance Profiles

Variable           Worth     Rank

B_aftermeal      0.025791      1
F_aftermeal      0.017335      2
F_beforesleep    0.008913      3
B_beforesleep    0.005025      4
B_wakingup       0.004078      5
F_wakingup       0.003802      6
Brush            0.003276      7
Floss            0.000907      8
F_anothertime    0.000654      9
B_anothertime    0.000186     10
```

9. Which are the top-2 important variables for segment 2? (0.5 point)

These are the most important variables for segment 2 using only descriptors.

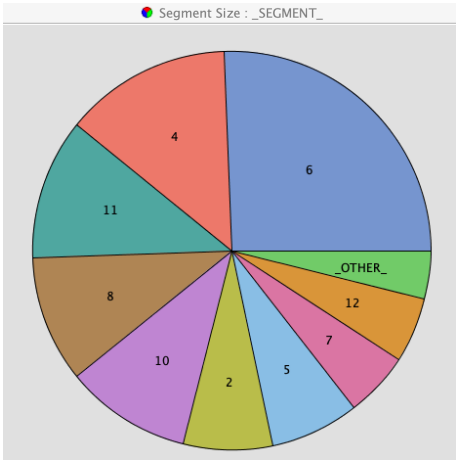
```
Variable: _SEGMENT_ Segment: 2 Count: 32
Decision Tree Importance Profiles

Variable                      Worth    Rank

Age_in_Years             .008553290       1
Birth_Year               .008553290       2
YRS_Current_Employer     .005476121       3
Salary                   .005262772       4
YRS_Current_Position     .004094721       5
debtundercontrol         .000482302       6
world                    .000408729       7
Gender                   .000255357       8
homeimprov               .000171801       9
School_Age_Children      .000131077      10
```

10. Provide a summary of the cluster profiles with the bases and descriptors and compare the results. (1.5 points)

The cluster with descriptors has 10 segments:

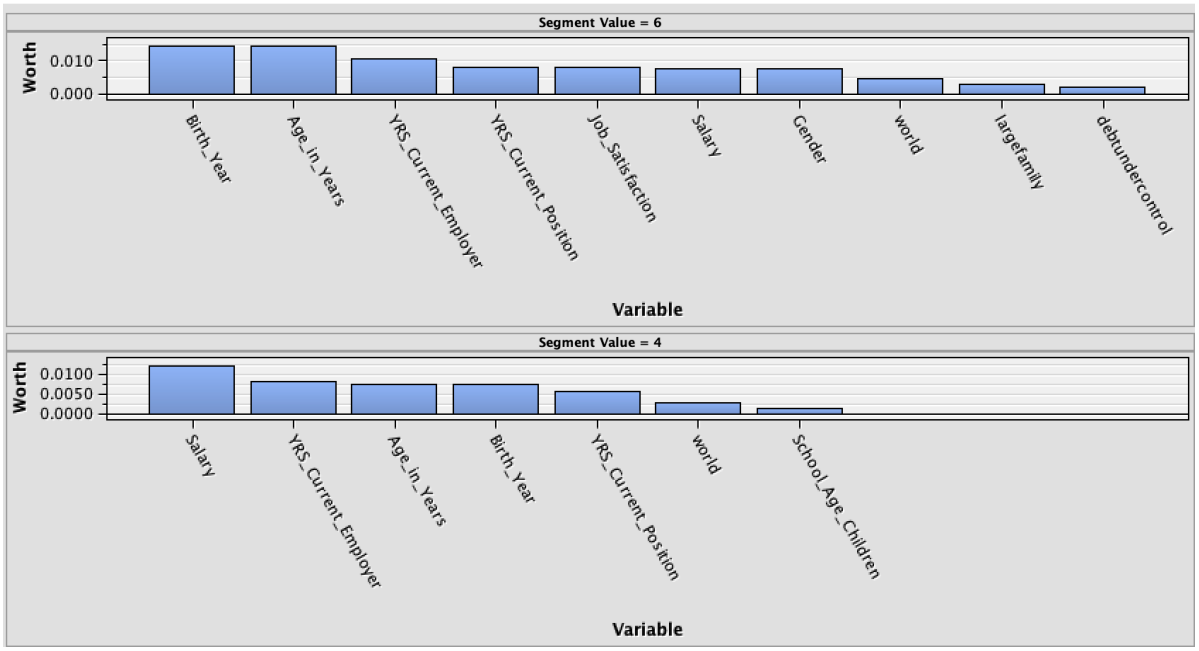


This is the description of its segments:

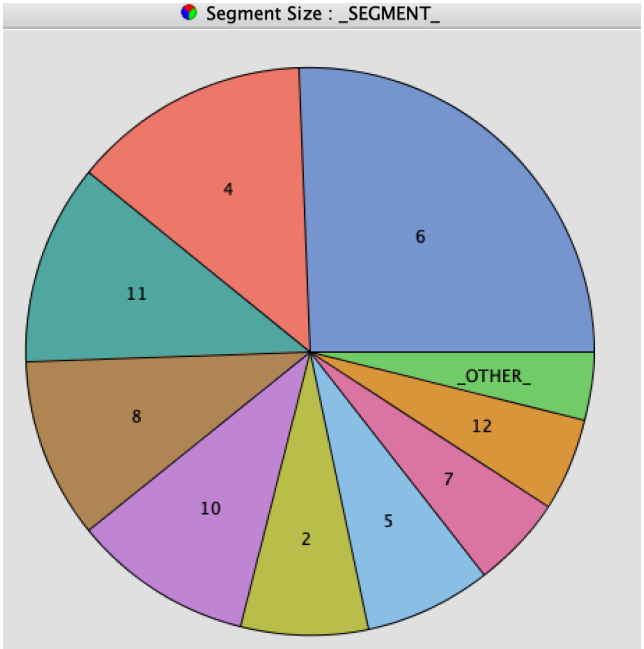
Frequencies: _SEGMENT_

Segment Variable	Segment Value	Frequency Count	Percent of Total Frequency
SEGMENT	6	115	25.6696
SEGMENT	4	61	13.6161
SEGMENT	11	50	11.1607
SEGMENT	8	47	10.4911
SEGMENT	10	46	10.2679
SEGMENT	2	32	7.1429
SEGMENT	5	32	7.1429
SEGMENT	7	24	5.3571
SEGMENT	12	23	5.1339
SEGMENT	_OTHER_	18	4.0179

The 2 largest segments have these distributions.



On the other hand, the cluster with base variables has 10 segments.

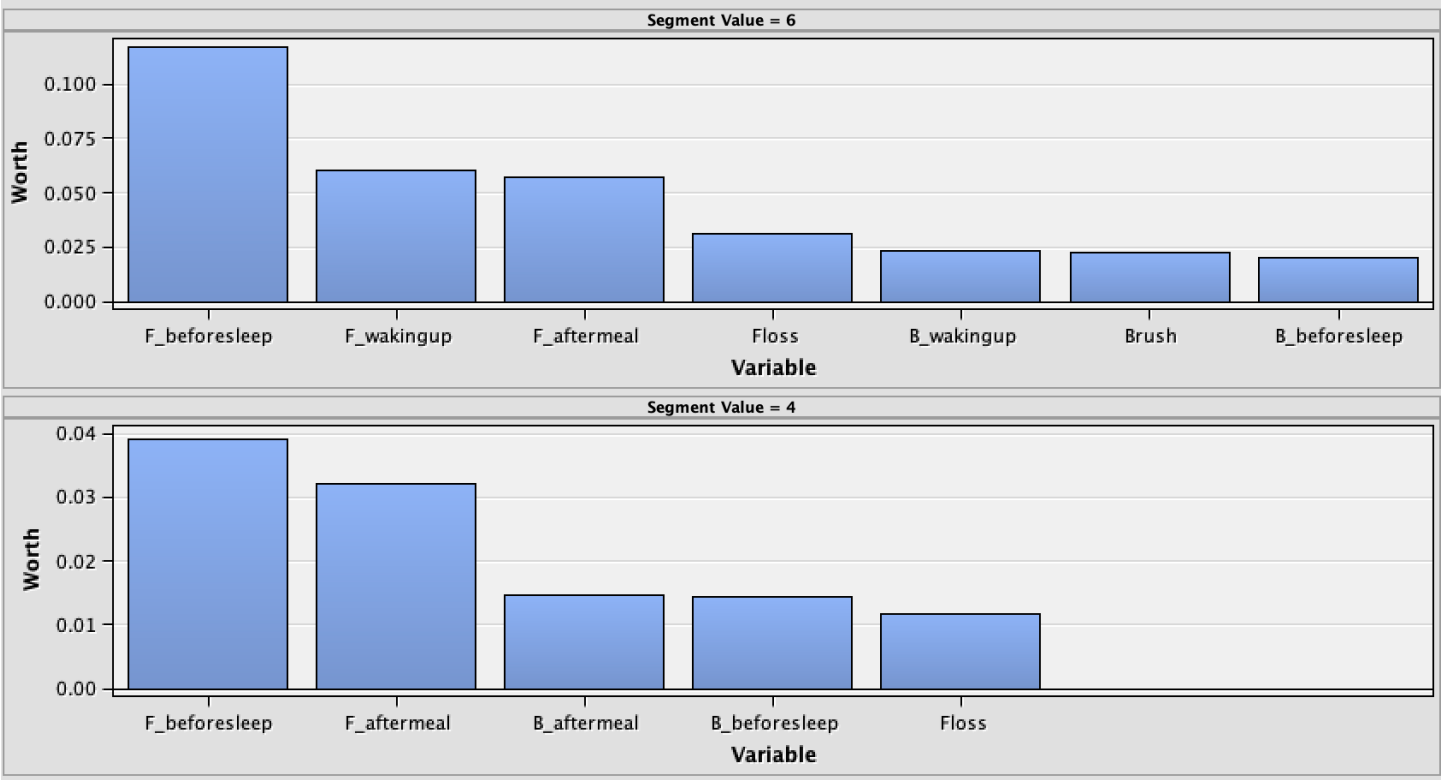


This is the description of its segments:

Frequencies: _SEGMENT_

Segment Variable	Segment Value	Frequency Count	Percent of Total Frequency
SEGMENT	6	115	25.6696
SEGMENT	4	61	13.6161
SEGMENT	11	50	11.1607
SEGMENT	8	47	10.4911
SEGMENT	10	46	10.2679
SEGMENT	2	32	7.1429
SEGMENT	5	32	7.1429
SEGMENT	7	24	5.3571
SEGMENT	12	23	5.1339
SEGMENT	_OTHER_	18	4.0179

The 2 largest segments have these distributions.



The clusters with descriptors and base variables are similar, they both have 10 segments and similar segment distribution. The variables change , as they are using a different set. The largest clusters with base variables identify **f_beforesleep** as the most important, on the other hand, the largest clusters with descriptor variables identify **birthyear** and **salary** as the most important.

11.What suggestions do you have for the dental non-profit group in relation to the marketing campaign?

Based on the segments of the cluster with base variables, I would suggest to considering flossing before sleep (F_beforesleep) as a key variable, as well as flossing after waking up (F_wakingup) and brusing (B_aftermeal) and flossing (F_aftermeal) after meals, because these where identified as being the most important in the largest segments. There are other segments but they gradually reduce in importance, focusing effort in the most important variables of the largest segments would have the biggest impact.