

The Orange Lifestyle

Mohit Singhi and Sangar Rane, Oklahoma State University

ABSTRACT

Being a freshman at a large university, life can be fun as well as stressful. The choices a freshman makes while in college may impact his/her overall health. In order to examine the overall health and different behaviors of students at Oklahoma State University a survey was conducted among the freshman students. The survey focused on capturing the psychological, environmental, diet, exercise, alcohol and drug use among students. A total of 795 out of 1,036 freshman students completed the survey which included around 270 questions that covered the range of issues mentioned above. An exploratory factor analysis identified 26 factors. For example, two factors that relate to the behavior of students under stress are eating and relaxing. Further understanding the variables which contribute to alcohol and drug use may help the university in planning appropriate interventions and preventions. Factor analysis with Cronbach's alpha provided insight in to a more defined set of variables to help address these types of issues. We used SAS® to do factor analysis as well as to create different clusters of students with unique characteristics and profiled these clusters.

INTRODUCTION

College life molds the future of a person. Stress, Concentration, Anxiety, Drug abuse, Alcohol consumption are some factors that have a deep effect on the life of a university student. We have attempted to analyze all these factors using a survey among freshman students. Though this survey has only been done at Oklahoma State University, these findings are almost the same in other universities across the United States. We hope that our study helps in improving overall behavior of a student in the freshman year itself & control unwanted habits like drug & alcohol consumption.



Figure 1. Chart describing questions asked in survey

DATA PREPARATION

Our survey consisted of 270 questions asked to 1,036 students. We received responses from 795 students. The topics in the survey covered a variety of areas. Factor analysis process (Figure 2) identified 26 different factors/variables using Eigen value greater than 1 criterion.

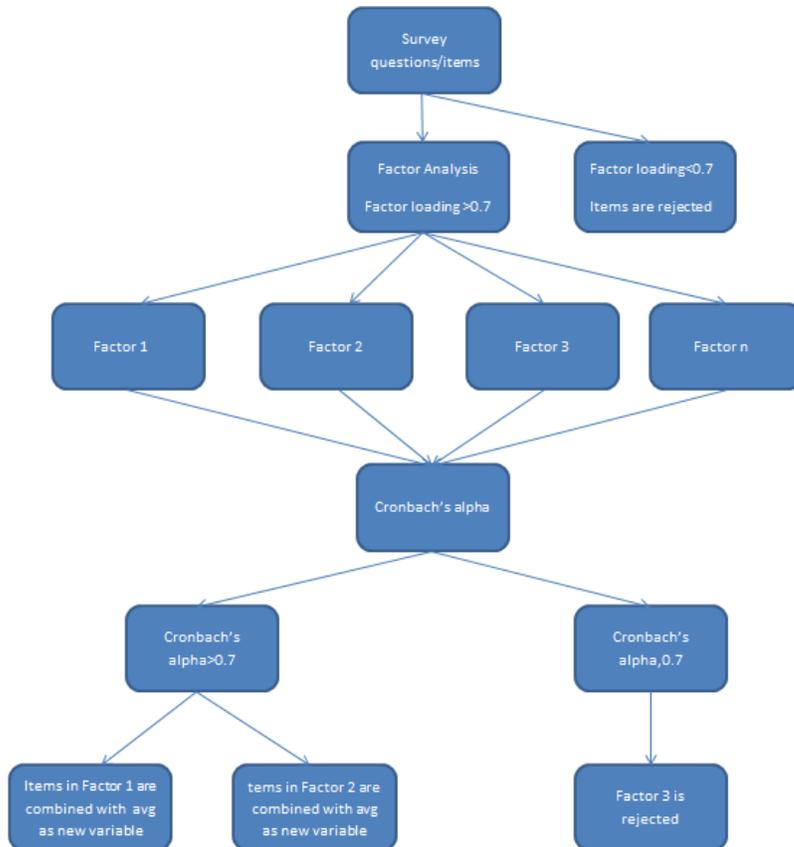


Figure 2. Factor Analysis Flow

These 26 factors were distributed in different areas like stress, confidence, concentration, physical activities, anxiety & addiction. Cronbach's alpha was used to measure the internal consistency of the factors. For the purpose of this research, we rejected any factors that had Cronbach's alpha value of less than 0.7. Our goal was to examine stress, confidence, anxiety, concentration, physical activities, drug & alcohol use and other such factors among freshman students. Our research also focused on how students deal with social anxiety, dropping confidence levels and increasing stress. Therefore, based on our focus area we decided to go ahead with the variables in Figure 3. Figure 4 shows the Pearson correlation amongst the selected variables.

Attribute Name	Description	scale	min of scale	max of scale
Imp_Pss_Avg	Perceived Stress	5 point	never	very often
Imp_Bad_Concentration_avg	Describes how bad the concentration of a student is.	10 point		0 100
Imp_Social_anxiety_avg	Social Anxiety Level of a student	5 point	not at all	extremely
Imp_conf_class_avg	Confidence level of a student	5 point	strongly disagree	strongly agree
IMP_Busy_hrs	workhours, volunteer work, study hours & campus	7 point	0 hours	more than 40 hours
IMP_EnoughSleep_NoOfDays	days in a week one gets adequate sleep	8 point	0 days	7 days
IMP_MinExercise_noOfDays	participates in minimum exercising(20-30mins)	8 point	0 days	7 days
Imp_describe_weight	How do you describe your weight?	5 point	very underweight	very over weight

Figure 3. Selected Variable Details

Pearson Correlation Coefficients, N = 795										
Prob > r under H0: Rho=0										
	IMP_Busy_hrs	IMP_EnoughSleep_NoOfDays	IMP_MinExercise_noOfDays	IMP_NCHA_max	IMP_PSS_Avg	IMP_Schly_Stress	IMP_Social_anxiety_Average	IMP_bad_concentration_Avg	IMP_conf_class	IMP_describe_weight
IMP_Busy_hrs	1	-0.04588	0.17061	0.00975	0.05925	0.14193	0.07554	0.0151	0.03893	0.01493
		0.1963	<.0001	0.7837	0.095	<.0001	0.0332	0.6708	0.2729	0.6742
IMP_EnoughSleep_NoOfDays	-0.04588	1	0.23238	-0.10371	-0.31906	-0.21922	-0.0394	-0.19717	0.17156	-0.04911
		0.1963	<.0001	0.0034	<.0001	<.0001	0.2672	<.0001	<.0001	0.1666
IMP_MinExercise_noOfDays	0.17061	0.23238	1	-0.04134	-0.15484	-0.10697	-0.00152	-0.07026	0.125	-0.1656
	<.0001	<.0001		0.2444	<.0001	0.0025	0.9659	0.0477	0.0004	<.0001
IMP_NCHA_max	0.00975	-0.10371	-0.04134	1	0.00397	0.00886	-0.01237	0.12178	-0.03439	0.02473
		0.7837	0.0034	0.2444		0.911	0.8031	0.7277	0.0006	0.3328
IMP_PSS_Avg	0.05925	-0.31906	-0.15484	0.00397	1	0.58836	0.33853	0.54617	-0.47216	0.1922
		0.095	<.0001	<.0001	0.911		<.0001	<.0001	<.0001	<.0001
IMP_Schly_Stress	0.14193	-0.21922	-0.10697	0.00886	0.58836	1	0.23835	0.35297	-0.33322	0.12974
	<.0001	<.0001	0.0025	0.8031	<.0001		<.0001	<.0001	<.0001	0.0002
IMP_Social_anxiety_Average	0.07554	-0.0394	-0.00152	-0.01237	0.33853	0.23835	1	0.39323	-0.35973	0.07951
		0.0332	0.2672	0.9659	0.7277	<.0001	<.0001		<.0001	0.025
IMP_bad_concentration_Avg	0.0151	-0.19717	-0.07026	0.12178	0.54617	0.35297	0.39323	1	-0.36517	0.14464
		0.6708	<.0001	0.0477	0.0006	<.0001	<.0001	<.0001		<.0001
IMP_conf_class	0.03893	0.17156	0.125	-0.03439	-0.47216	-0.33322	-0.35973	-0.36517	1	-0.08241
		0.2729	<.0001	0.0004	0.3328	<.0001	<.0001	<.0001	<.0001	
IMP_describe_weight	0.01493	-0.04911	-0.1656	0.02473	0.1922	0.12974	0.07951	0.14464	-0.08241	1
		0.6742	0.1666	<.0001	0.4863	<.0001	0.0002	0.025	<.0001	0.0201

Figure 4. Pearson Correlation

DATA TRANSFORMATION

DOUBLE STANDARDIZATION BEFORE CLUSTERING

A combination of row and column standardization is called double standardization. It is one of the recommended transformations of variables before clustering of survey data to overcome the response-style issues in surveys. Double standardization was carried out on all the above variables. Please refer Figure 5 & Figure 6 to see the effects of double standardization on the distribution of variables.

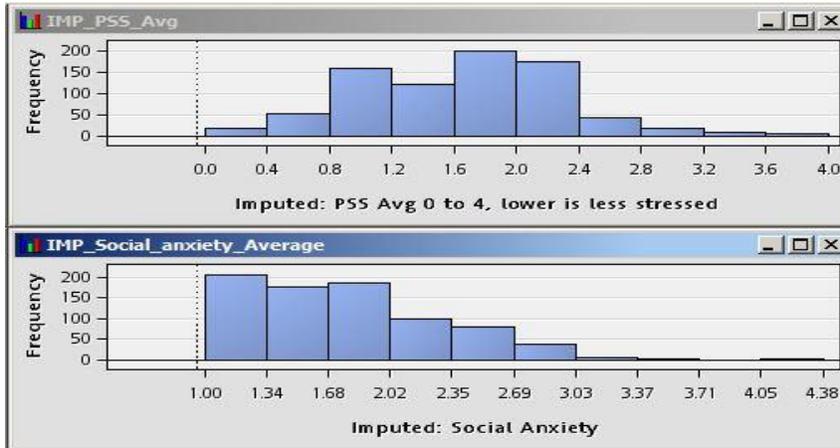


Figure 5. Distribution of Variables Before Double Standardization

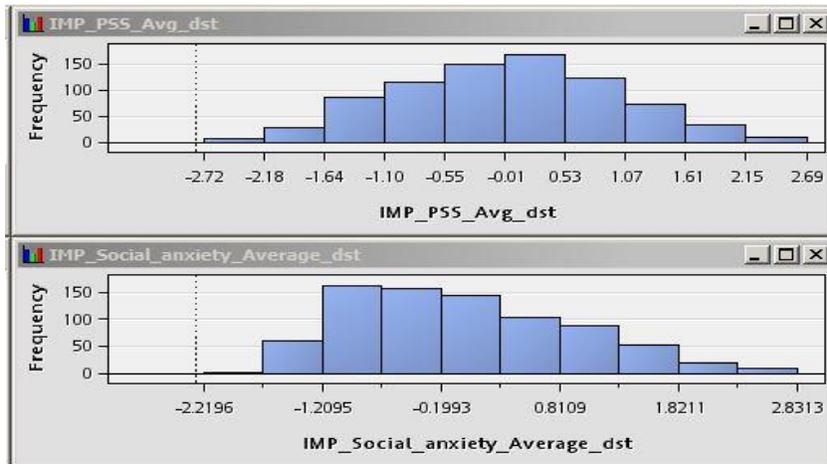


Figure 6. Distribution of Variables after Double Standardization

K-MEANS CLUSTERING

We used Ward's minimum variance as the clustering method and "Automatic" as the specification method for number of clusters. Please refer Figure 7. for process flow in SAS® Enterprise miner.

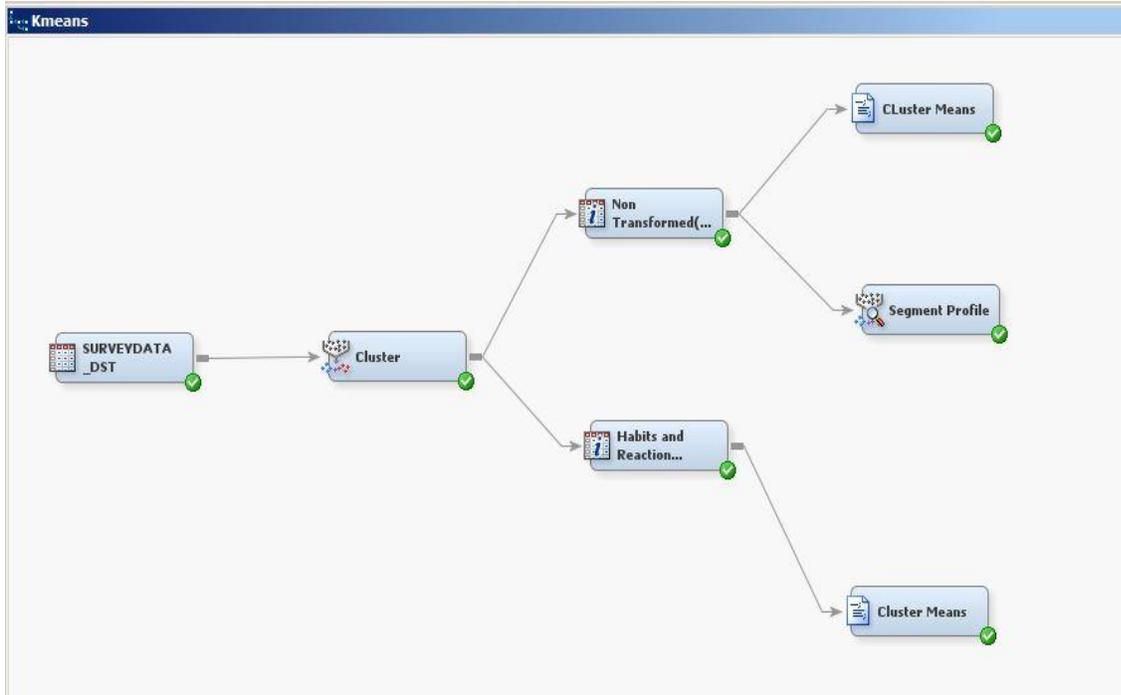


Figure 7. Process Flow in SAS Miner

10	CL13	OB44	144	0.0145	.403	58.8	19.5
9	CL22	CL21	93	0.0169	.386	61.7	21.3
8	CL31	CL29	92	0.0219	.364	64.4	36.0
7	CL10	CL18	200	0.0223	.342	68.2	27.5
6	CL12	CL19	169	0.0234	.318	73.7	29.6
5	CL8	CL9	185	0.0241	.294	82.3	26.6
4	CL11	CL5	330	0.0403	.254	89.7	44.0
3	CL6	CL14	265	0.0410	.213	107	48.4
2	CL4	CL7	530	0.0636	.149	139	64.1
1	CL3	CL2	795	0.1494	.000	.	139

Figure 8. Cluster History

Based on these settings it initially generated a 7 cluster solution. But as per the data and our requirement we opted to go for a 3 cluster solution based on the pseudo t-squared statistic.

Mean Statistics					
Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation
0.88249	0.001857	.	1	258	0.881712
0.88249	0.001857	.	2	259	0.876423
0.88249	0.001857	.	3	278	0.893517

Figure 9. Mean Statistics of Clusters

In Figure 9 we can see that the frequency of each of the clusters is reasonable. A good clustering solution is the one in which the clusters are roughly equal in terms of their compactness. This compactness can be defined by the Root-Mean-Square Standard deviation which in our case is nearly same for all the 3 clusters. Figure 13 gives the importance of each variable in predicting the segment to which a particular case (student) belongs. In our case PSS_Avg (Perceived stress) is the most important variable. The importance of the remaining variables is with reference to the importance of PSS_Avg. For example: Importance of EnoughSleep_NoOFDays and Busy_hrs is 0.945 and 0.912 times the importance of PSS_Avg respectively.

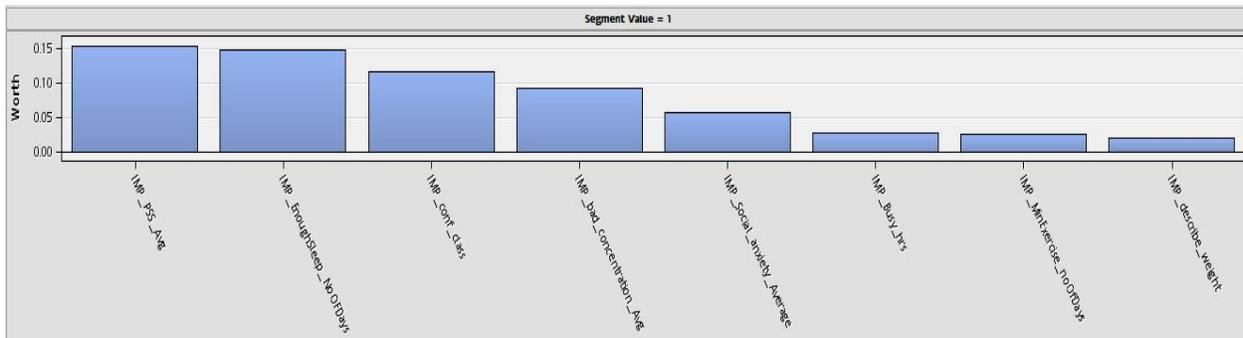


Figure 10. Variable Worth in Segment 1

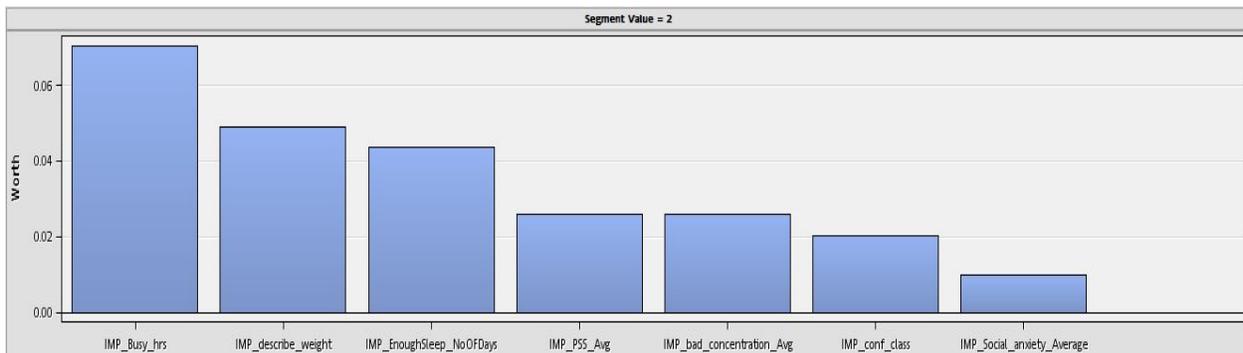


Figure 11. Variable Worth in Segment 2

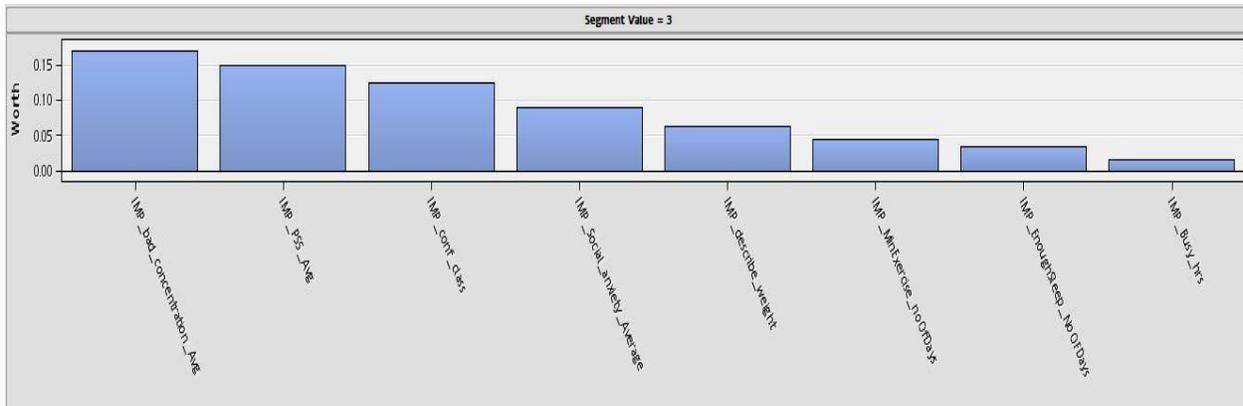


Figure 12. Variable Worth in Segment 3

Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
IMP_PSS_Avg_dst		2	12	1.00000
IMP_EnoughSleep_NoOfDays_dst		5	8	0.94543
IMP_Social_anxiety_Average_dst		0	12	0.91909
IMP_Busy_hrs_dst		3	11	0.91294
IMP_conf_class_dst		3	5	0.91047
IMP_bad_concentration_Avg_dst		3	9	0.87223
IMP_MinExercise_noOfDays_dst		2	7	0.73114
IMP_describe_weight_dst		2	8	0.69794
Yoga_Walk	Yogal_Walk2	0	1	0.23218

Figure 13. Variable Importance

SEGMENT PROFILING

- It is important to understand responses from each cluster behaves and this can be achieved via segment profiling. We have used cluster profiling node in SAS® enterprise miner to achieve this objective.
- Clustering was done on double standardized variables but we haven't used them for segment profiling because the double standardized values are difficult to interpret for profiling.
- To overcome this issue, we have used the actual variable values instead of the double standardized variable values in profiling.
- Segment profile node generates the mean statistics for the variables across each segment which makes it easy for us to compare it with the overall mean of the variables.
- For better understanding we have put the mean statistics for variables in an excel and highlighted the means with two colors shown in Figure 14 which identifies whether the mean of the variable for that particular segment is above or below the overall mean of that variable.

HIGH	
LOW	

Figure 14. Color Coding for High and Low Averages with respect to Overall Mean

PROFILING FOR FOCUS VARIABLES

- The focus variables of our study are perceived stress, social anxiety, confidence, concentration, busy hours, feeling about weight, number of days in week they do minimum exercise, no of days in a week they get enough sleep.
- Figure 15 gives the mean statistics for those variables. For easy understanding we have presented the information in a simpler manner in Figure 16.

Segment ID	IMP_busy_hrs_mean	Imp_enough Sleep_noOfDays	IMP_MinExercise_noOfDays_Avg	imp_Pss_Avg_mean	Imp_Social_Anxiety_Average_mean	Imp_bad_concentration_Avg_Mean	Imp_Conf_class	Imp_Describe_Weight
1	1.7781	4.8211	3.68217	1.10744	1.51162	17.9324	4.2184	3.1124
2	2.15541	2.40345	3.21493	1.64235	1.66369	25.2326	3.81924	2.82239
3	1.82822	2.48393	2.05036	2.09297	2.04865	44.5288	3.29132	3.50719

Figure 15. Mean Statistics for Variables Used for Clustering

Segment ID	Busy	Enough sleep Days	Min Exercise Days	Perceived Stress	Social Anxiety	Bad Concentration	Confidence	Feeling about weight
1	Less Busy	more days of enough sleep	More days of min exercise	Less stress	Low social anxiety level	Good concentration	high confidence level	Comfortable with weight
2	more busy	Less days of enough sleep	More days of min exercise	More Stress	slightly more social anxiety level then seg 1	slightly bad concentration as compared to segment 1	Low confidence level as compared to segment 1	Some believe they are underweight
3	slightly more busy then segment 1	slightly more days then segment 2	Less days of min exercise	most stress	High social anxiety level	Very bad concentration	Lower confidence than average of all students	some believe they are overweight

Figure 16. Profiling for Focus Variables

PROFILING FOR BEHAVIORAL VARIABLES

In a similar manner, profiling is done for variables such as activity to concentrate, coping with stress, activity under stress, activity to increase confidence, alcohol consumption, belief about smoking and drinking. Please refer Figure 17.

Segment	Activity to concentrate	coping with stress	Activity under stress	Activity to increase confidence	Alcohol Consequences	Belief about smoking and Drinking
1	Consciously relax or eat something and negligible use of drugs	cope actively with stress	Less eating	Think positive	very low concentration of all the things mentioned for segment 2 and 3	believe drinking doesn't help in any way and smoking is bad for mental health
2	Eat something	cope by planning/strategy	consciously relax	think positive	slight risky behaviour	Prefer smoke free environment
3	Consciously relax or eat something and high concentration of drug users	get into denial, more of substance use, Run away from the situation, use help from others	Eat more and consciously relax	do not think positive, eat more and resort to drug/alcohol use	Blacked out after drinking, physiologically dependent on alcohol, affected academics and risky behaviour	some believe that smoking is not bad for mental health and also believe that drinking helps

Figure 17. Profiling for Behavioral Variables

CONCLUSIONS

- Students who consider themselves as overweight tend to have high levels of stress, bad concentration and high social anxiety.
- Students with low anxiety and stress levels and good concentration are the ones who also report that they are getting enough sleep.
- High stress and anxiety among students leads to alcohol consumption and smoking and it also affects their ability to think positively.

REFERENCES

Eliminating Response Style Segments in Survey Data via Double Standardization before Clustering by Murali Krishna Pagolu and Goutam Chakraborty, Paper 165-2011, SAS Global Forum 2011.

ACKNOWLEDGEMENTS

We thank Dr. Josh Wiener, Professor & Head of Marketing Department, Dr. Goutam Chakraborty, Professor, Professor of Marketing and Director of MS in Business Analytics and Dr. Miriam McGaugh, Professor for their support throughout our research.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Mohit Singhi

Oklahoma State University

Stillwater Oklahoma

Email: <mailto:mohit.singhi@okstate.edu>

Sangar Rane

Oklahoma State University

Stillwater, Oklahoma

Email: <mailto:sangar.rane@okstate.edu>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.