Introduction

I am a creative designer at Nike. Nike is an American retail store that offers a wide range of products such as footwear, apparel, equipment, accessories, and services. With the growing concern in global warming I would like to conduct research to see if there is a market for Nike to produce a new streetwear shoe made from 100% recycled materials. In a recent study at MIT, a typical pair of running shoes was found to generate 30 pounds of carbon dioxide emissions, that is equivalent to keeping a 100-watt light bulb on for one week. While the majority of emissions are generated during the manufacturing process of the shoe lifecycle, by using recycled materials we are able to manufacture less new materials thus causing a decrease in carbon dioxide emissions.

The target population I am interested in is going to be the entire US adult population. While it is a bit broad, I picked this population because we would like to get an idea of what the ideal customer would look like for this type of product. Since this will be Nike's attempt to enter a new niche market, we do not want to start the analysis with a segment in mind and leave out potential buyers.

Questions for Analysis

The two major themes I will be examining are environmental enthusiasm and fashion forward. That being said, I selected the following questions for this analysis:

1) Are you more likely to purchase from an environmentally-friendly company.
2) Companies should help consumers become more environmentally responsible.
3) It is important to me that others see me as environmentally conscious.
4) Eco-friendly products are usually of higher quality.
5) Everything I wear is of the highest quality.
6) I like to keep up with the latest fashions.
7) I always look for my favorite brands first.
8) I am the first among my friends to try new styles.

I hypothesize that there are two potential factors within these selected questions.

F1 – Environmental Enthusiasts

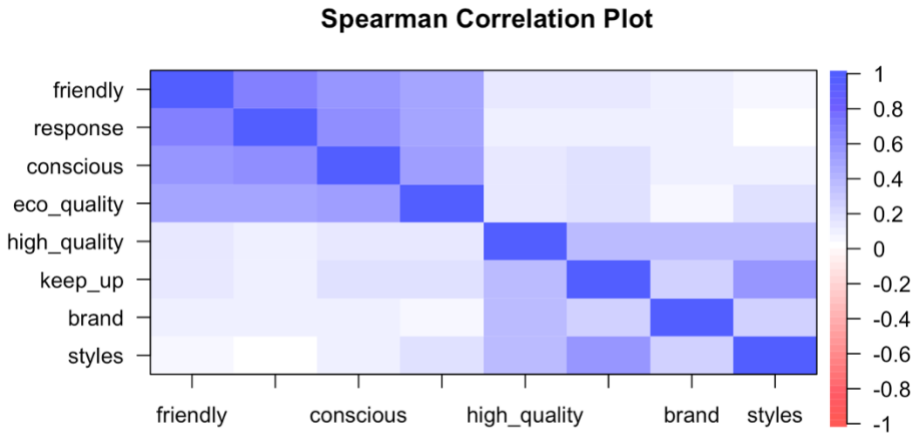
F2 – Fashion Forward

The first factor will measure how important the environment topic is to the population and how they think about eco-friendly or reused, recycled products while the second factor will measure how the population thinks about the fashion topic.

Accessing the Factorability of the Data

Before jumping into the dimensionality reduction portion of this analysis, I will need to do some due diligence on my question variables to ensure they are suitable for further analysis.

Since the survey uses the Likert scale, we can infer that the response data is ordinal and should use Spearman correlation to analyze the correlation between the selected questions.



From the correlation plot above, I can see that there are moderate to strong correlations between the questions I hypothesized would get grouped together in the PCA. This makes me hopeful that I selected appropriate questions. After running the analysis, both factors generated were located on the same side leading me to assume that I should proceed using varimax PCA.

Before running the factor analysis I need to do some due diligence and check to see if the data is even suitable using Kaiser-Meyer-Olkin (KMO) test and Bartlett 's test of sphericity.

Bartlett's Test of Sphericity

The Bartlett test will assess the intercorrelation amongst the selected variables, and should be found statistically significant in order to perform factor analysis.

```
cortest.bartlett(questions)

## $chisq
## [1] 48759.48
##
## $p.value
## [1] 0
##
## $df
## [1] 28
```

With a p-value of less than 1, we can conclude that the Bartlett test is statistically significant and can be more confident with our analysis moving forward

Kaiser-Meyer-Olkin test

The KMO test looks at partial correlations within the data and is essentially a better measure of factorability. The minimum acceptable value for this test is 0.5.

```
KMO(questions)

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = questions)
## Overall MSA =  0.77
## MSA for each item =
##     friendly     response    conscious  eco_quality high_quality
##         0.78         0.76         0.82         0.85         0.76
##      keep_up        brand       styles
##         0.70         0.75         0.67
```

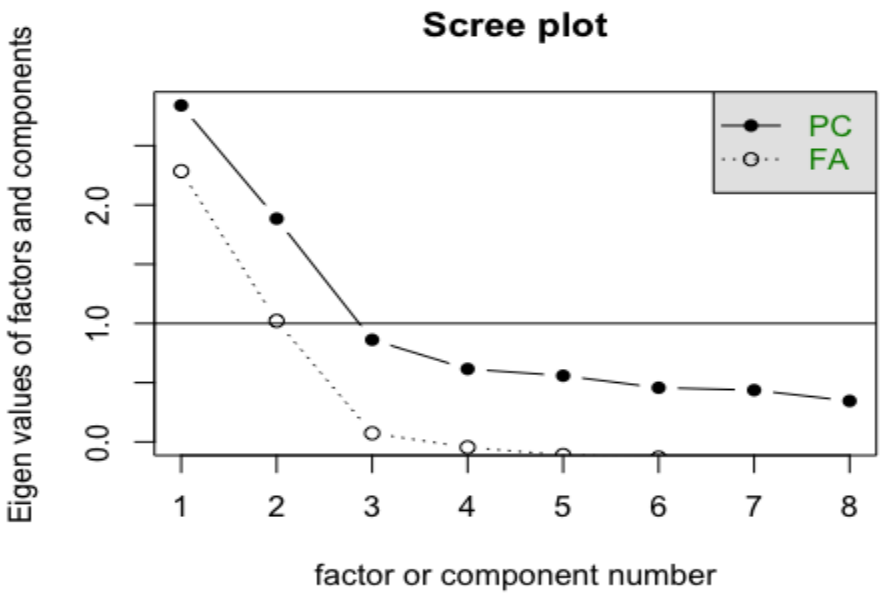
This test gives an overall MSA of 0.77 which is not fabulous, but definitely acceptable. Analyzing the individual MSA scores I noticed that the first four question, what I hope to be the Environment Enthusiast factor, has greater sampling adequacy than the second four questions, what I hope to be the Fashion Forward factor.

Determining the Number of Factors to Extract

I will be leveraging two different techniques to determine the appropriate number of factors parallel analysis and scree plot.

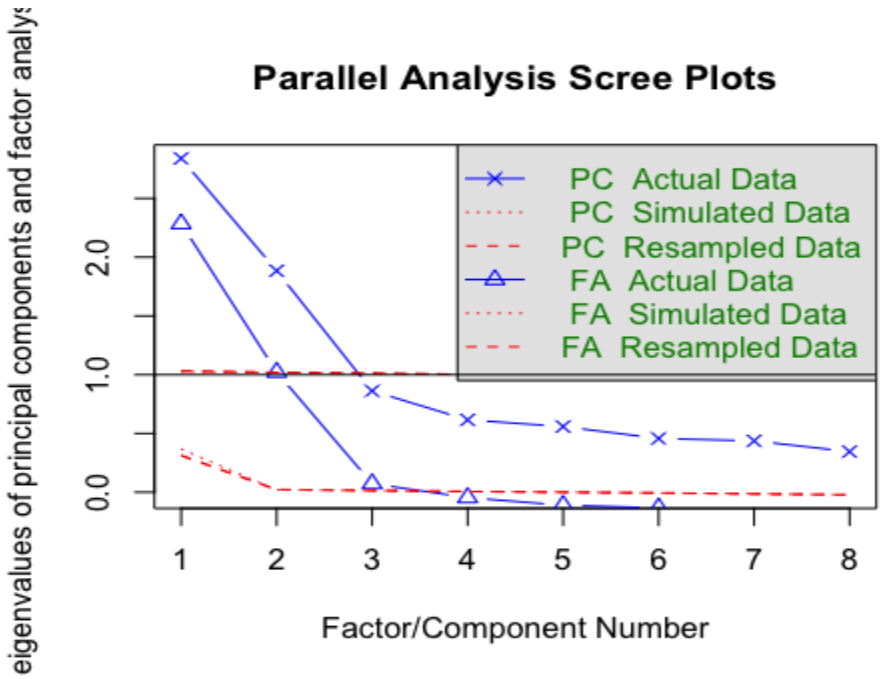
Scree Plot

The scree plot visually displays the eigenvalues per factor. After analyzing the plot below, I think two factors is the appropriate solution for this dataset. I settled on two factors as both the factors and principal components show two eigenvalues greater than one.



Parallel Analysis

In addition to plotting the eigenvalues from the factor analysis, a parallel analysis also generates random correlation matrices and compares the resulting eigenvalues to the eigenvalues of the observed data. This additional step makes the parallel analysis a bit better than an ordinary scree plot, as the observed eigenvalues that are higher than their corresponding random eigenvalues are more likely to be from meaningful factors. The plot below suggests that I should proceed with at most three factors and two principal components.



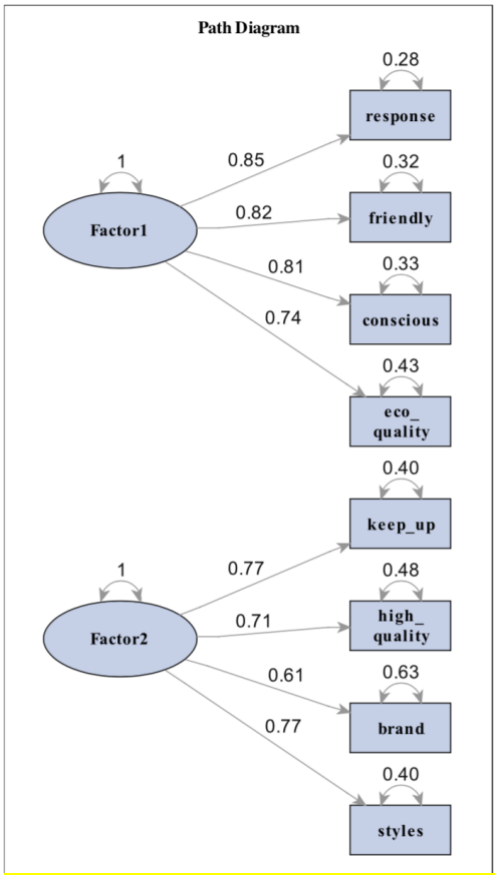
Extraction Technique

Now that I am confident in my data, I need to decide what type of extraction technique I am going to use before performing factor analysis. I will use be using principal component analysis (PCA) in an attempt to group these eight questions into two factors. As for the rotation, I will be using varimax since I do not want to have correlation between the two factors. Now that I have finished up all of the preparation, I can finally run the PCA.

I am able to conclude that my hypothesis is correct as two factors were retained with the Environmental Enthusiast and Fashion Forward questions being grouped together as predicted. I've included a table to display the rotated factor pattern as well as a visual representation of the factoring below. The percent

variance explained by the two factors is about 55%, which is acceptable but a little lower than what I was hoping for.

Rotated Factor Pattern		
	Factor1	Factor2
friendly	0.82464	0.05474
response	0.84709	0.01106
conscious	0.81410	0.08426
eco_quality	0.74494	0.11854
high_quality	0.06778	0.71484
keep_up	0.10391	0.76646
brand	0.02889	0.60677
styles	0.04665	0.77433

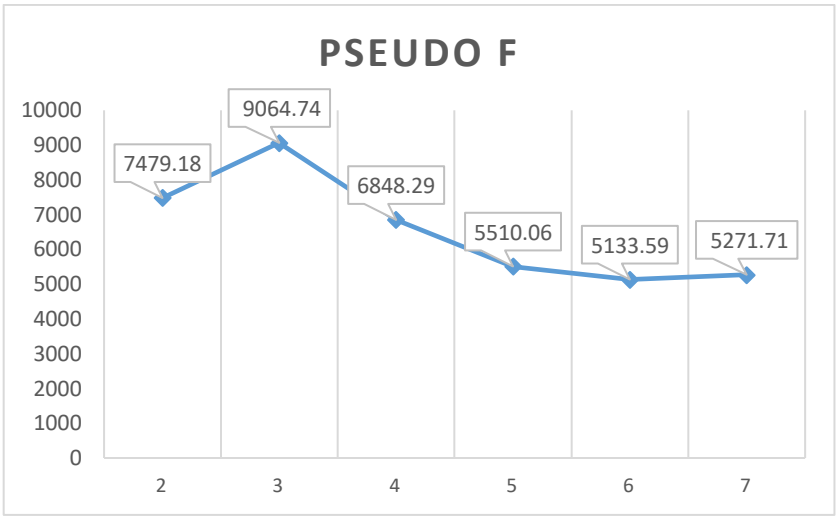


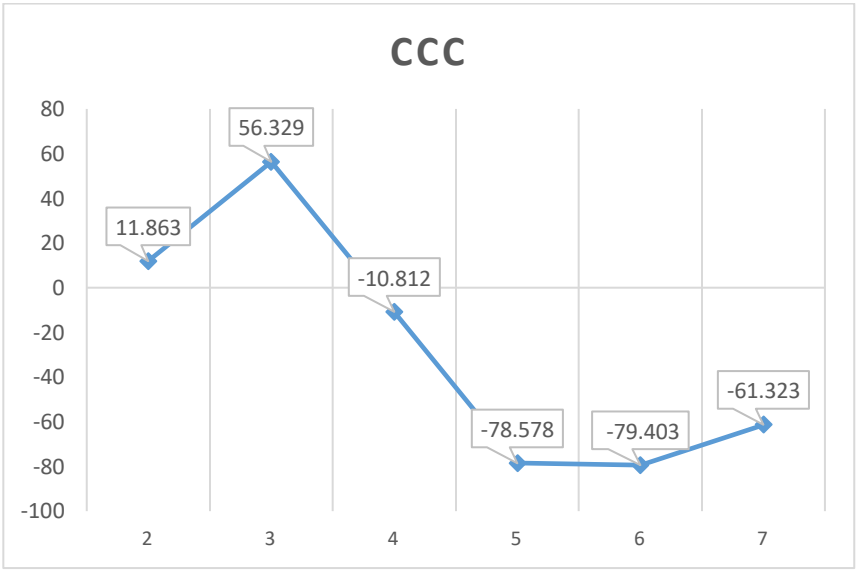
Cluster Analysis

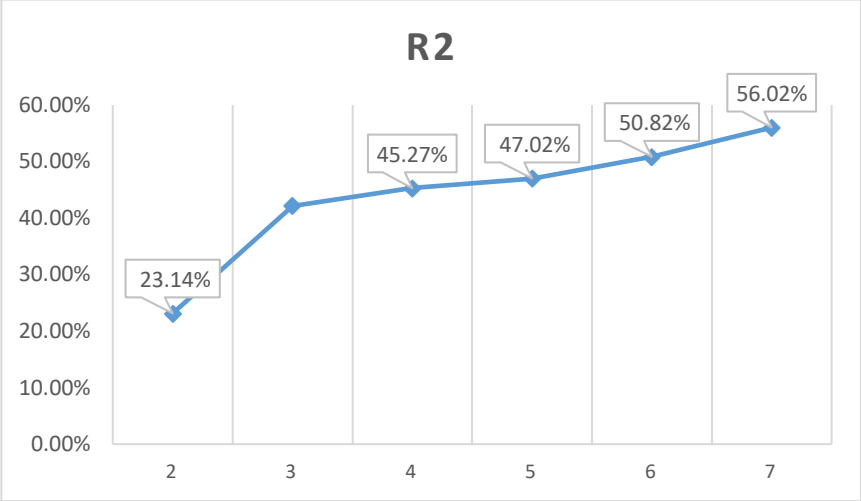
I chose to add “Likely to purchase products seen on cellphone advertisements” to find people who are easily persuaded by online marketing campaigns and “People have a responsibility to use recycled products” to narrow in on people who are going to buy this product. Additionally, since both of my driver variables were also on a Likert scale, I did not standardize the data.

K-Means Clustering

I see a peak in both the Pseudo F and CCC plots at 3 clusters, while the R2 plot shows that the hike starts at cluster 3.







Gap Analysis

For my dataset, the first peak was selected to be three clusters. I was able to make this inference from the ABC statistics table. I do this by first calculating the one standard error adjusted Gap statistic and comparing it to the previous gap number. If this number is greater, this is the first peak and the corresponding number of clusters is chosen. Given these criteria, if you look at the table below you will notice that this peak occurs at 3 clusters.

ABC Statistics					
Number of Clusters	Logarithm of Within-Cluster SSE		Gap	Simulation Adjusted Standard Deviation	One Standard Error Adjusted Gap
	Input	Reference			
2	11.0909	12.2915	1.2006	0.00868	1.1920
3	10.8455	12.1814	1.3360	0.0139	1.3220
4	10.6941	11.9409	1.2468	0.0184	1.2284
5	10.5704	11.7970	1.2267	0.0141	1.2126
6	10.5270	11.5569	1.0300	0.0115	1.0184

This estimated number of clusters table confirms my observations from the table above.

Estimated Number of Clusters	
Criterion	Number of Clusters
FIRSTPEAK	3

From the cluster summary table I notice that two out of the three clusters seem to distributed evenly, with cluster 1 being slightly smaller than clusters 2 and 3. While this could potentially be problematic I would like to evaluate the cluster means before concluding.

Cluster Summary								
		Distance from Cluster Centroid to Observation						
Cluster	Frequency	Maximum	Minimum	Average	SSE	Standard Deviation	Nearest Cluster	Distance to Nearest Cluster Centroid
1	4898	4.6171	0.3810	1.5551	13557.2	1.6637	3	2.3280
2	7425	5.2946	0.2732	1.5517	20833.1	1.6751	3	2.1719
3	8171	3.8900	0.3133	1.3484	16911.3	1.4386	2	2.1719

Looking at the within cluster statistics table I am able to evaluate the means across clusters by each variable used. The clusters look to be meaningful and offer a decent split on behavior.

Within Cluster Statistics			
Variable	Cluster	Mean	Standard Deviation
Environment	1	0.4776	2.1125
	2	-0.8592	2.2384
	3	0.4636	1.8465
Fashion	1	0.6149	2.5377
	2	-0.0268	2.1210
	3	-0.3298	1.7573
recycle	1	4.1676	8.6571
	2	2.7499	7.2095
	3	4.2902	5.6845
cell	1	3.3377	7.1786

Within Cluster Statistics			
Variable	Cluster	Mean	Standard Deviation
	2	1.9228	4.8120
	3	1.2136	2.7174

Non-Driver Variables and Demographics

To get a better understanding of what the average person looks like in one of these 3 clusters, I am going to add some descriptive variables to the analysis. For demographics I am going to include gender, age, and race. As for non-drivers, I am going to be looking at Facebook interaction in the last 30 days, purchase of Nike shoes within the last year, and magazine subscriptions to both Time and People.

Age:
1 = 18-24
2 = 25-34
3 = 35-44
4 = 45-54
5 = 55+

Table of _CLUSTER_ID_ by age

_CLUSTER_ID_(Cluster ID)	age					
Frequency Percent Row Pct Col Pct	1	2	3	4	5	Total
1	1959 9.56 40.00 23.84	1019 4.97 20.80 23.90	796 3.88 16.25 23.15	649 3.17 13.25 24.45	475 2.32 9.70 24.75	4898 23.90
2	2972 14.50 40.03 36.16	1522 7.43 20.50 35.70	1257 6.13 16.93 36.55	965 4.71 13.00 36.36	709 3.46 9.55 36.95	7425 36.23
3	3288 16.04 40.24 40.00	1722 8.40 21.07 40.39	1386 6.76 16.96 40.30	1040 5.07 12.73 39.19	735 3.59 9.00 38.30	8171 39.87
Total	8219 40.10	4263 20.80	3439 16.78	2654 12.95	1919 9.36	20494 100.00

Race:
1 = White
2 = Black or African American
3 = Asian
4=Other

Table of _CLUSTER_ID_ by race

_CLUSTER_ID_(Cluster ID)	race				
Frequency Percent Row Pct Col Pct	1	2	3	4	Total
1	567 2.77 11.58 24.11	158 0.77 3.23 23.65	419 2.04 8.55 25.50	3754 18.32 76.64 23.71	4898 23.90
2	904 4.41 12.18 38.44	247 1.21 3.33 36.98	596 2.91 8.03 36.28	5678 27.71 76.47 35.87	7425 36.23
3	881 4.30 10.78 37.46	263 1.28 3.22 39.37	628 3.06 7.69 38.22	6399 31.22 78.31 40.42	8171 39.87
Total	2352 11.48	668 3.26	1643 8.02	15831 77.25	20494 100.00

Gender:
0=Male
1=Female

Table of _CLUSTER_ID_ by female			
_CLUSTER_ID_(Cluster ID)	female		
Frequency Percent Row Pct Col Pct	0	1	Total
1	2098 10.24 42.83 23.57	2800 13.66 57.17 24.15	4898 23.90
2	3270 15.96 44.04 36.74	4155 20.27 55.96 35.84	7425 36.23
3	3532 17.23 43.23 39.69	4639 22.64 56.77 40.01	8171 39.87
Total	8900 43.43	11594 56.57	20494 100.00

Nike:
0=No
1=Yes

Table of _CLUSTER_ID_ by nike_12_mo			
_CLUSTER_ID_(Cluster ID)	nike_12_mo		
Frequency Percent Row Pct Col Pct	0	1	Total
1	3880 18.93 79.22 23.86	1018 4.97 20.78 24.07	4898 23.90
2	5896 28.77 79.41 36.25	1529 7.46 20.59 36.15	7425 36.23
3	6488 31.66 79.40 39.89	1683 8.21 20.60 39.79	8171 39.87
Total	16264 79.36	4230 20.64	20494 100.00

Facebook:
0=No
1=Yes

Table of _CLUSTER_ID_ by facebook_30_days			
_CLUSTER_ID_(Cluster ID)	facebook_30_days		
Frequency Percent Row Pct Col Pct	0	1	Total
1	2678 13.07 54.68 23.48	2220 10.83 45.32 24.43	4898 23.90
2	4160 20.30 56.03 36.48	3265 15.93 43.97 35.92	7425 36.23
3	4567 22.28 55.89 40.04	3604 17.59 44.11 39.65	8171 39.87
Total	11405 55.65	9089 44.35	20494 100.00

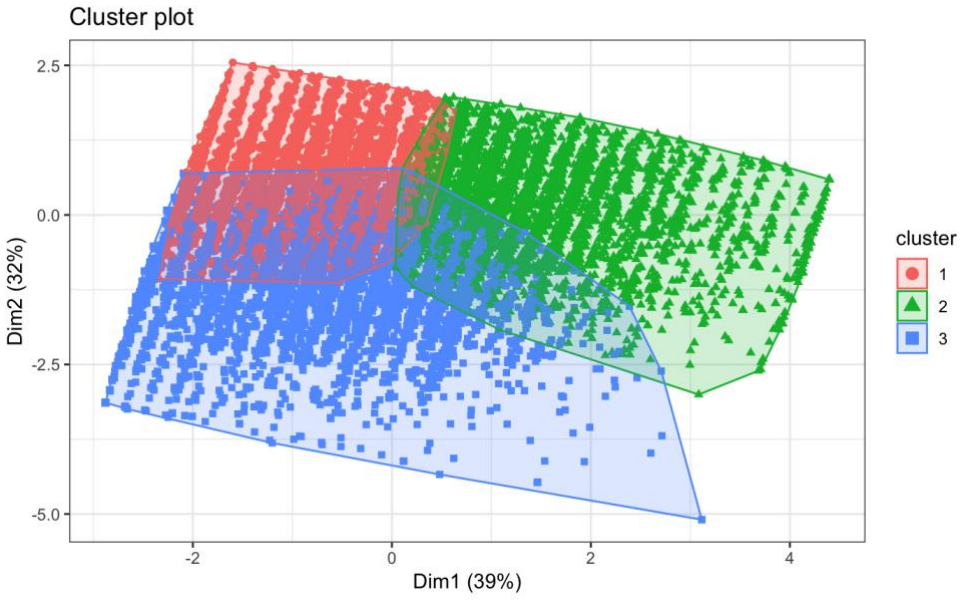
Time:
0=No
1=Yes

Table of _CLUSTER_ID_ by time_6_mo			
_CLUSTER_ID_(Cluster ID)	time_6_mo		
Frequency Percent Row Pct Col Pct	0	1	Total
1	4084 19.93 83.38 24.01	814 3.97 16.62 23.38	4898 23.90
2	6187 30.19 83.33 36.37	1238 6.04 16.67 35.56	7425 36.23
3	6742 32.90 82.51 39.63	1429 6.97 17.49 41.05	8171 39.87
Total	17013 83.01	3481 16.99	20494 100.00

People:
0=No
1=Yes

Table of _CLUSTER_ID_ by people_6_mo			
_CLUSTER_ID_(Cluster ID)	people_6_mo		
Frequency Percent Row Pct Col Pct	0	1	Total
1	3663 17.87 74.79 24.16	1235 6.03 25.21 23.15	4898 23.90
2	5496 26.82 74.02 36.25	1929 9.41 25.98 36.16	7425 36.23
3	6001 29.28 73.44 39.58	2170 10.59 26.56 40.68	8171 39.87
Total	15160 73.97	5334 26.03	20494 100.00

Cluster Summary



Cluster 1: Young, trendy, careless

The average person in this cluster is going to be extremely fashionable and somewhat ecofriendly. They will more than likely buy a product they see advertised on their phone, they are also the most likely to use Facebook compared to other groups. They are the smallest cluster of the three and also the youngest. The strategy here is going to be to pivot towards an online/digital marketing campaign. Since the new shoe will be part of a street wear line, we can expect this group to be interested if the shoe is fashionista approved.

Cluster 2: tree hugger, fashionable-ish

The average person in this cluster is going to be extremely ecofriendly and somewhat fashionable. They will probably not buy something advertised to them on their phone, and are the least likely to use social media like Facebook. It looks like they are more likely to read magazines so I would suggest, strategy wise, that while this customer does not seem very approachable there is still some hope. Their love for the environment and strong feelings towards the need to use recycled products makes me think if we advertise in a handful of magazines as well we might be able to pull some in.

Cluster 3: No sense of style, global warming who?

The average person in this cluster is not very interested in the environment and probably just throw on the first thing they see in the morning. While they are not likely to buy a product advertised to them on their phone, they are likely the most likely to use Facebook and read magazines. This cluster is also going to be the oldest. The strategy here is going to be to probably stay away from this group if possible. While they are the most likely to buy Nike shoes in the past year, given the answers from the other questions, this particular shoe would probably not spark interest.