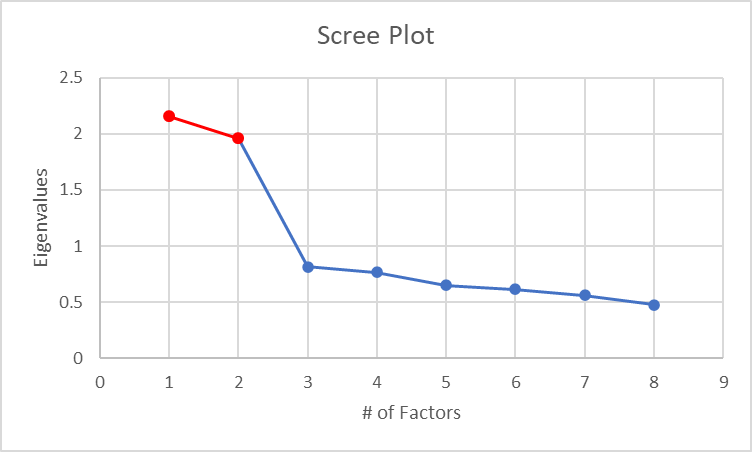
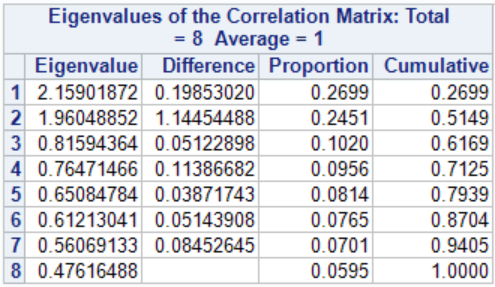
**Tesla Market Segmentation Project**

For this project, I am going to pretend I am the Tesla car company. My company has been developing a new family type vehicle with all the latest technologies and want to advertise it. My goal is to identify people who are family centered and who are into cars like Tesla. I want to see if I can identify the people who are likely to purchase this new vehicle. For this project, the target population is the U.S. adult population who are 18 years of age or older. To segment the market, I will use a combination of Principal Component Analysis and Cluster Analysis by way of K-Means Clustering and GAP Analysis.

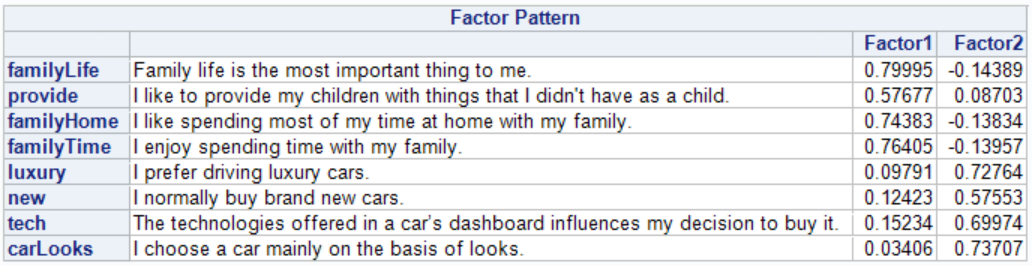
To develop the segmentation system, certain questions had to be identified to measure two different constructs for factor analysis. The two constructs for this project are: 1) People who are family centered, and 2) People who want nice cars. In addition to these variables, four other questions were identified to build clusters around. Finally, a set of non-driver descriptor variables were chosen to get a better picture of the individuals inside each cluster. The table below summarizes all the variables used in this analysis and frequency tables for each variable can be found in the Appendix.

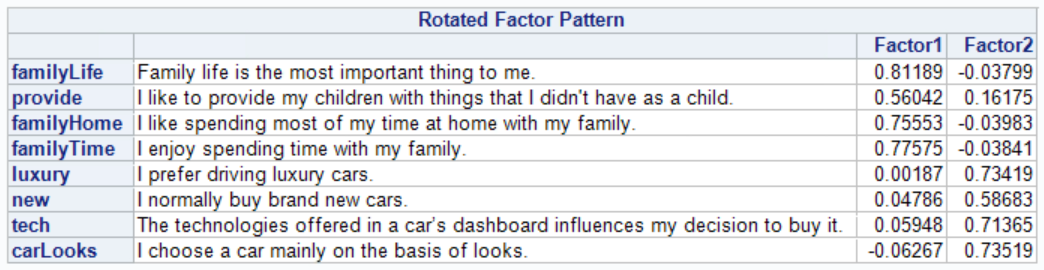
|  |  |  |
| --- | --- | --- |
|  | **Question** | **Variable Name(s)** |
| **Construct 1: Family Centered** | Family Life is the most important thing to me. | familyLife |
| I like to provide my children with things that I didn't have as a child. | provide |
| I like spending most of my time at home with my family. | familyHome |
| I enjoy spending time with my family. | familyTime |
|  |  |  |
| **Construct 2: Want Nice Cars** | I prefer driving luxury cars. | luxury |
| I normally buy brand new cars. | new |
| The technologies offered in a car’s dashboard influences my decision to buy it. | tech |
| I choose a car mainly on the basis of looks. | carLooks |
|  |  |  |
| **Cluster Variables** | Advertising helps me learn about the products companies have to offer. | advLearn |
| I love to buy new gadgets and appliances. | gadgets |
| I am worried about pollution and congestion caused by cars. | pollution |
| I am more likely to purchase products I see advertised on a social media/networking website. | advSocialMedia |
|  |  |  |
| **Descriptor Variables** | Race | Raceasian, Raceblack, Racewhite, Raceother |
| Watch ESPN in last 7 days. | espn |
| Next Vehicle: Used or New | NewVehicle |
| Chick-fil-a in last 30 days | chickfila |

In order to quantify the two constructs, factor analysis was utilized. To perform factor analysis, the method of Principal Component Analysis (PCA) with a Varimax Rotation was applied to the eight construct questions. Using PCA for this kind of extraction is traditionally used and it is computationally easier. PCA focuses on all of the variance in each of the variables. In addition, the factor scores produced by PCA have the advantage that the correlation of the factors exactly corresponds to the correlation of the produced factor scores. The advantage of using a Varimax Rotation is that it forces the factors to be orthogonal, so they are more likely to produce well separated clusters in cluster analysis. One disadvantage in doing this is that the rotation, to some extent, distorts reality by forcing the factors to be orthogonal.

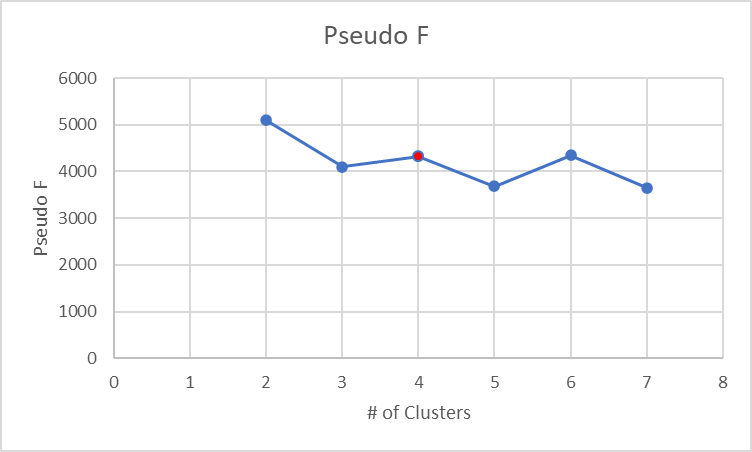
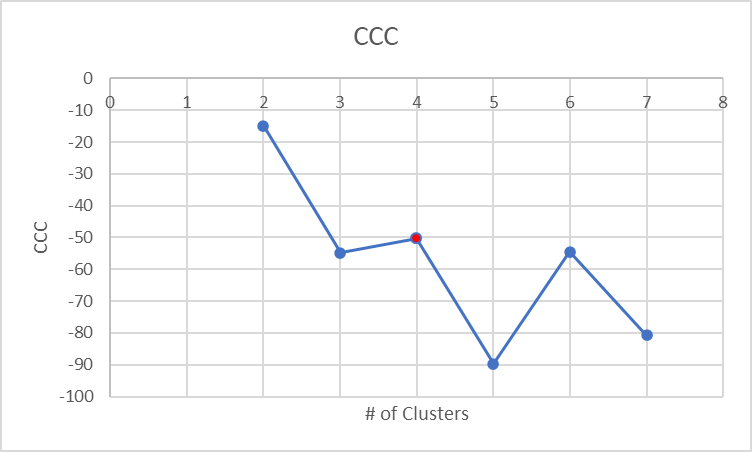
 Below is a summary table and a Scree plot of the eigenvalues calculated by PCA. Using the Kaiser criterion, stating that that the number of factors to keep is where its eigenvalue is greater than or equal to one, we can see that the optimal number of factors to keep is two. The Scree plot also tells the same story. At the factor number where the slope of the line suddenly changes, we chose the number of factors before it.

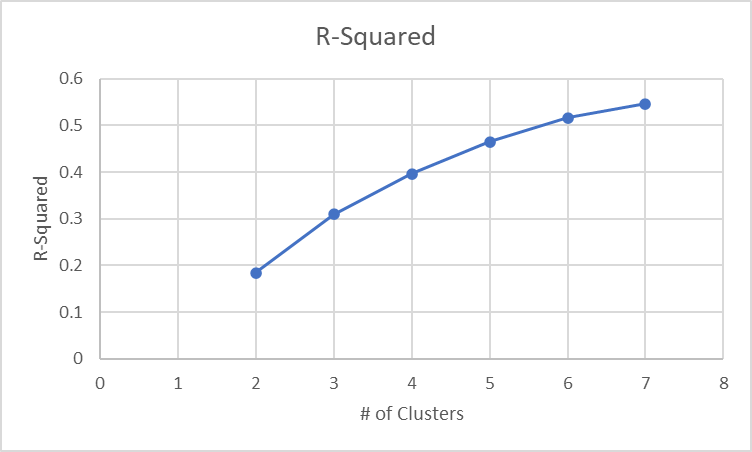
Once the optimal number of factors has been identified, the next step was to examine the factor pattern both before and after factor rotation. Looking at the two tables below, we can see that the Varimax rotation was able to split the constructs onto two separate factors, with the first construct being on factor 1 and the second construct being on factor 2.

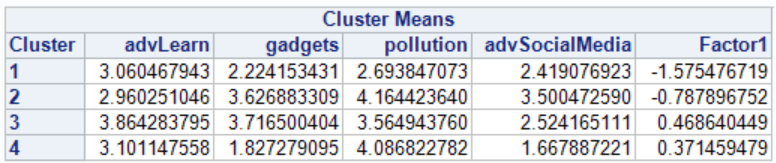


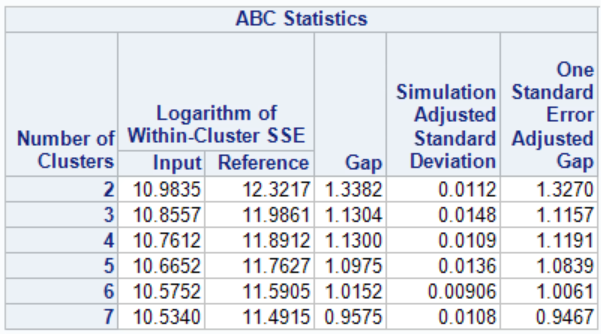


To determine the amount of variance explained by the factors, we look at the eigenvalues above. Each eigenvalue represents the amount of total variance explained by that factor. To get the amount of variance explained by the two factors we have, we just need to divide each of the eigenvalues by the sum of all eigenvalues, and then sum those together. In the table above, that is summarized in the “Cumulative” column. From that, we can see that the two factors we have explain 51.49% of the total variance.

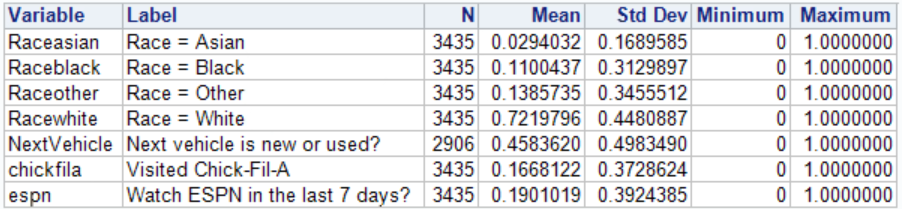
 The next step in the segmentation process is to identify the optimal number of clusters. To do this, using the four chosen clustering variables and factor 1 from PCA, the methods of K-Means Clustering and GAP Analysis were utilized. For K-Means Clustering, 2 – 7 clusters were tested and judged based on their Pseudo F and Cubic Clustering Criterion (CCC) statistics. Below are graphs summarizing these statistics, plus the R2 for the number of clusters. To find the optimal number of clusters based on the K-Means method, we look for the first local maxima of the Pseudo F and CCC statistics. If they agree on the number of clusters, then that is the optimal number. Looking at the graphs below, we see that the first local maxima for both statistics is at four clusters. Therefore, based on K-Means, four clusters are optimal.

 The table below summarizes the GAP Analysis statistics. To determine the optimal number of clusters, we first need to identify the maximum Gap. From the table below, we can see that it is 2. Next, we subtract the Simulation Adjusted Standard Deviation from the Gap number. This is the One Standard Error Adjusted Gap. Finally, we see if this number is farther away (greater than) one standard deviation from the next Gap number. Based on the table, 1.3270 > 1.1304. Therefore, based on the GAP Analysis, two clusters are optimal.

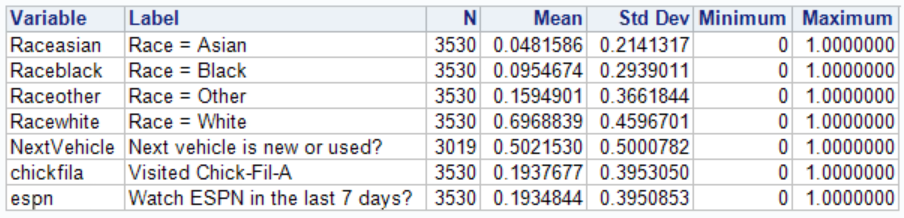
However, this does not match what K-Means Clustering provided. Because we don’t want a solution with only 2 clusters, the K-Means Clustering solution was chosen for the final number of clusters. The table below summarizes the cluster variable means across each cluster. From the table, we can see that each of the variables are somewhat decently separated between clusters. Based on the table, we can say the following:

* Cluster 1:
  + People don’t have an opinion about learning from advertising.
  + People tend to not acquire new gadgets.
  + People aren’t really concerned about pollution from cars
  + Advertising on social media discourages buying product.
* Cluster 2:
  + People don’t have an opinion about learning from advertising.
  + People tend to lean towards buying new gadgets.
  + People are concerned about pollution from cars.
  + Advertising on social media may have a slight positive effect on buying a product.
* Cluster 3:
  + People find advertising a little helpful for learning products
  + People tend to lean towards buying new gadgets.
  + People lean a little towards being concerned about pollution, but tend to not have an opinion.
  + Advertising on social media discourages buying product.
* Cluster 4:
  + People don’t have an opinion about learning from advertising.
  + People often do not acquire new gadgets.
  + People are concerned about pollution from cars.
  + Advertising on social media highly discourages buying product.

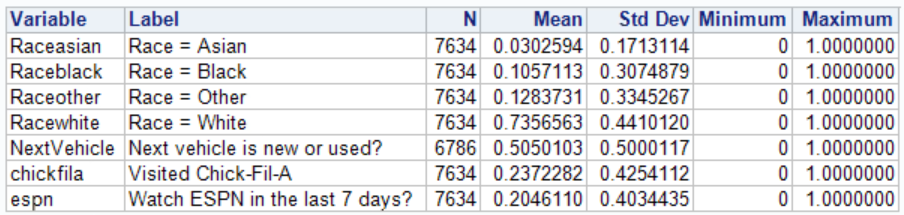
Finally, using these 4 clusters, the means of the descriptor variables were calculated for each cluster. These means are summarized in the tables below.

Cluster 1

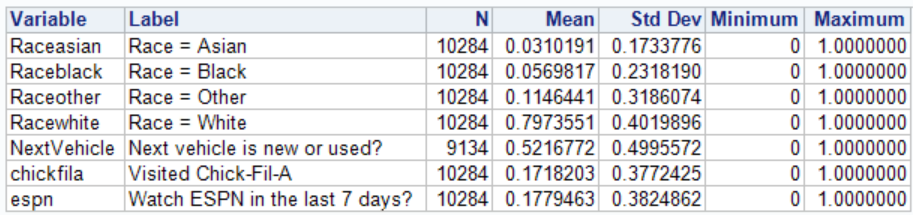
Cluster 2



Cluster 3



Cluster 4



Based on these tables, we can say the following:

* Cluster 1:
  + People are mostly Caucasian.
  + People tend to lean a little more to buying a used vehicle.
  + People tend to visit Chick-Fil-A the least compared to other clusters.
  + People watch ESPN the third most compared to other clusters.
* Cluster 2:
  + People are mostly Caucasian.
  + People buy new and used vehicles equally.
  + People tend to visit Chick-Fil-A the second most compared to other clusters.
  + People watch ESPN the second most compared to other clusters.
* Cluster 3:
  + People are mostly Caucasian.
  + People buy new and used vehicles equally.
  + People tend to visit Chick-Fil-A the most compared to other clusters.
  + People watch ESPN the most compared to other clusters.
* Cluster 4:
  + People are mostly Caucasian.
  + People buy new and used vehicles equally.
  + People tend to visit Chick-Fil-A the third most compared to other clusters.
  + People watch ESPN the least compared to other clusters.

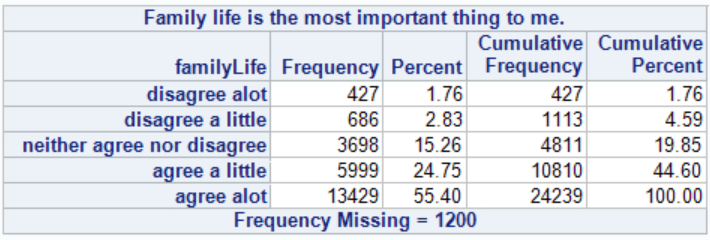
Bringing all of these cluster descriptions together, we can summarize them by name:

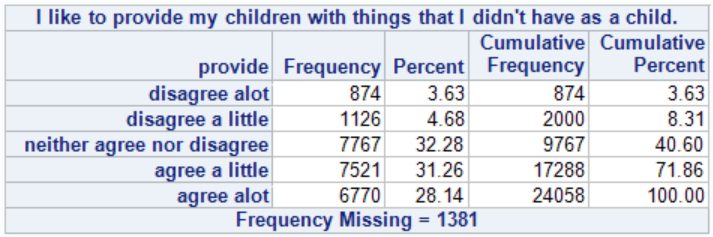
* Cluster 1 – Not Advertiser Friendly
* Cluster 2 – Environmental Techies
* Cluster 3 – Advertiser Friendly
* Cluster 4 – Environmental, Non-New-Tech

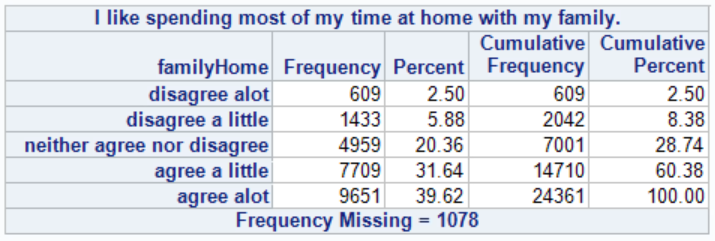
In conclusion, Tesla should focus its advertising campaigns on clusters 2 and 3 because 1) individuals in cluster 2 would be into technological, environmentally friendly cars like Teslas, and 2) individuals in cluster 3 are more receptive to advertisement. In addition, people in cluster 3, compared to other clusters, visit Chick-Fil-A and watch ESPN the most. Chick-Fil-A touts itself on being a family establishment and having family values, so Tesla might want to explore an advertising deal with Chick-Fil-A to feature its new family vehicle and target cluster 3. Tesla could also distribute adds on ESPN and target them at cluster 3. With this market segmentation, Tesla can focus its resources on specific people rather than potentially wasting them targeting everyone.

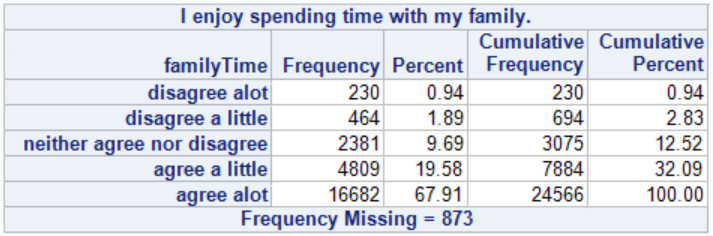
**Appendix**

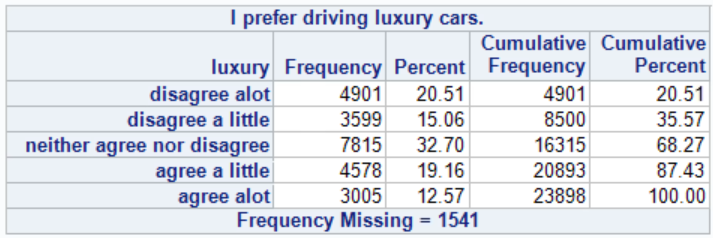
Frequency Tables:

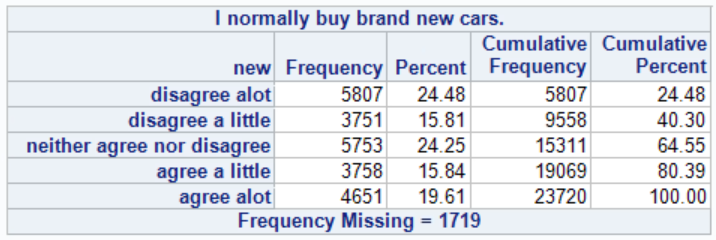
****

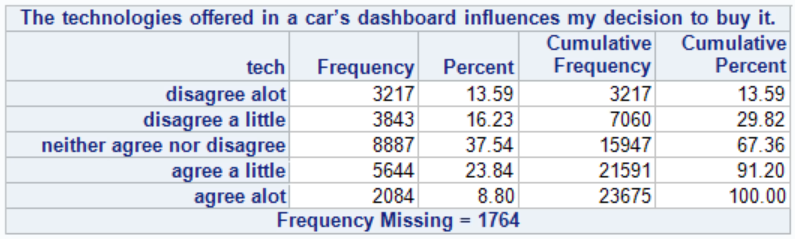
****

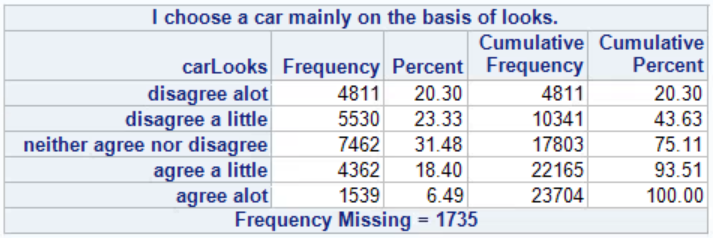
****

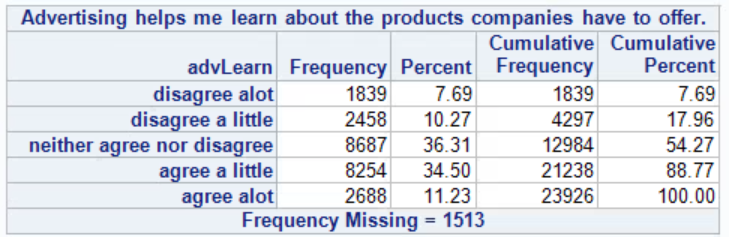
****

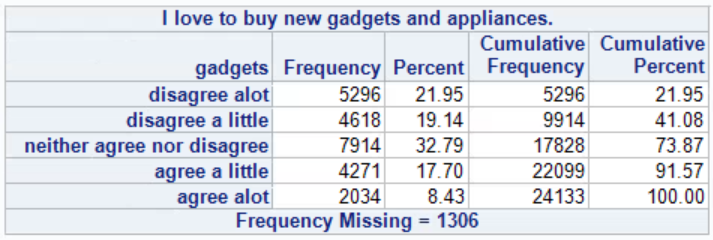


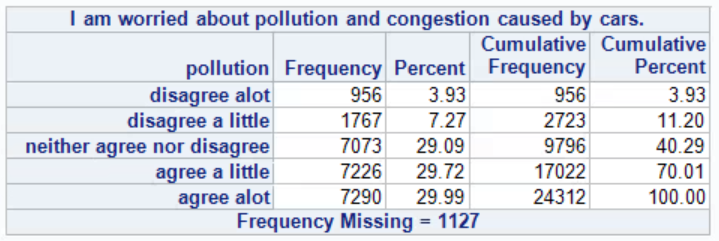


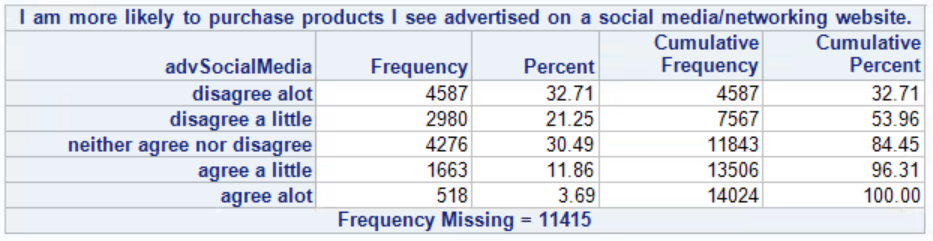


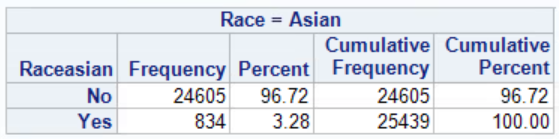


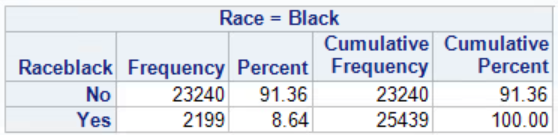


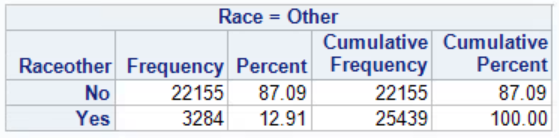


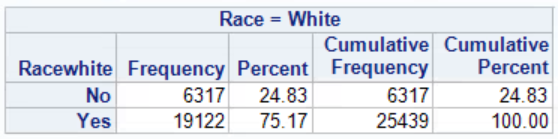


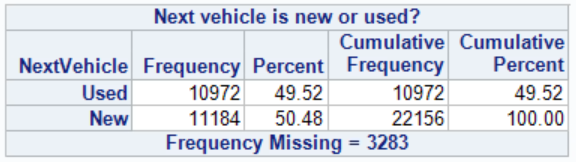


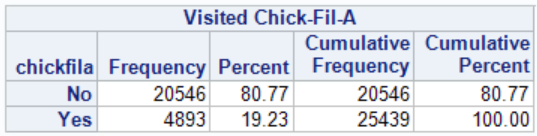


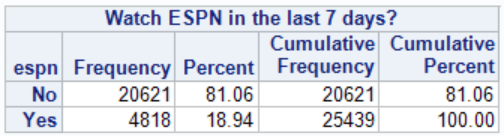












**SAS Code:**

/\*

Ian Scarff

iie728

Practicum I

Market Segmentation Project

/\*

/\*

I am going to imagine that I am Tesla. My company has been developing a new family car

with all the latest technologies and want to advertise it.

My goal is to identify people who are family centered and

who are into cars like Teslas.

\*/

/\* My target population is the U.S. adult population for 18 years of age or older. \*/

/\* Drivers

Construct 1: Family Centered

Question 1: Family life is the most important thing to me.

Question 2: I like to provide my children with things that I didn't have as a child.

Question 3: I like spending most of my time at home with my family.

Question 4: I enjoy spending time with my family.

Construct 2: People who want nice cars

Question 1: I prefer driving luxury cars.

Question 2: I normally buy brand new cars.

Question 3: The technologies offered in a car’s dashboard influences my decision to buy it.

Question 4: I choose a car mainly on the basis of looks

\*/

/\*

Cluster Variables:

Variable 1: Advertising helps me learn about the products companies have to offer.

Variable 2: I love to buy new gadgets and appliances.

Variable 3: I am worried about pollution and congestion caused by cars.

Variable 4: I am more likely to purchase products I see advertised on a social media/networking

website.

\*/

/\* Descriptor Variables:

Variable 1: Race

Variable 2: Watch ESPN in last 7 days.

Variable 3: Next Vehicle: Used or New

Variable 4: Chick-fil-a in last 30 days

\*/

/\* Go get variables \*/

filename rawData 'O:\Student$\MSDA2020\MKT6971\_003\Instructor\FA15\_Data.txt'

lrecl=**65378**;

/\* Grab the Driver variables \*/

**data** rawDrivers;

infile rawData;

input myid **1**-**7**

/\*Family life is the most important thing to me.\*/

familyLife\_agree\_alot **4680**

familyLife\_agree\_little **4757**

familyLife\_neither **4911**

familyLife\_disagree\_little **4988**

familyLife\_disagree\_alot **5065**

/\*I like to provide my children with things that I didn't have as a child.\*/

provide\_agree\_alot **4649**

provide\_agree\_little **4726**

provide\_neither **4880**

provide\_disagree\_little **4957**

provide\_disagree\_alot **5034**

/\*I like spending most of my time at home with my family.\*/

familyHome\_agree\_alot **4636**

familyHome\_agree\_little **4713**

familyHome\_neither **4867**

familyHome\_disagree\_little **4944**

familyHome\_disagree\_alot **5021**

/\*I enjoy spending time with my family.\*/

familyTime\_agree\_alot **4610**

familyTime\_agree\_little **4687**

familyTime\_neither **4841**

familyTime\_disagree\_little **4918**

familyTime\_disagree\_alot **4995**

/\*I prefer driving luxury cars.\*/

luxury\_agree\_alot **3617**

luxury\_agree\_little **3653**

luxury\_neither **3725**

luxury\_disagree\_little **3761**

luxury\_disagree\_alot **3797**

/\*I normally buy brand new cars.\*/

new\_agree\_alot **3620**

new\_agree\_little **3656**

new\_neither **3728**

new\_disagree\_little **3764**

new\_disagree\_alot **3800**

/\*The technologies offered in a car’s dashboard influences my decision to buy it.\*/

tech\_agree\_alot **3632**

tech\_agree\_little **3668**

tech\_neither **3740**

tech\_disagree\_little **3776**

tech\_disagree\_alot **3812**

/\*I choose a car mainly on the basis of looks.\*/

carLooks\_agree\_alot **3619**

carLooks\_agree\_little **3655**

carLooks\_neither **3727**

carLooks\_disagree\_little **3763**

carLooks\_disagree\_alot **3799**;

**run**;

/\* Now clean up the driver variables \*/

/\* use an array to convert missing values to zeros \*/

**data** DriverCleaning;

set rawDrivers;

array missing (**8**,**5**)

familyLife\_agree\_alot

familyLife\_agree\_little

familyLife\_neither

familyLife\_disagree\_little

familyLife\_disagree\_alot

provide\_agree\_alot

provide\_agree\_little

provide\_neither

provide\_disagree\_little

provide\_disagree\_alot

familyHome\_agree\_alot

familyHome\_agree\_little

familyHome\_neither

familyHome\_disagree\_little

familyHome\_disagree\_alot

familyTime\_agree\_alot

familyTime\_agree\_little

familyTime\_neither

familyTime\_disagree\_little

familyTime\_disagree\_alot

luxury\_agree\_alot

luxury\_agree\_little

luxury\_neither

luxury\_disagree\_little

luxury\_disagree\_alot

new\_agree\_alot

new\_agree\_little

new\_neither

new\_disagree\_little

new\_disagree\_alot

tech\_agree\_alot

tech\_agree\_little

tech\_neither

tech\_disagree\_little

tech\_disagree\_alot

carLooks\_agree\_alot

carLooks\_agree\_little

carLooks\_neither

carLooks\_disagree\_little

carLooks\_disagree\_alot;

/\* Convert missing values to zeros \*/

do i = **1** to **8**;

do j = **1** to **5**;

if missing(i,j) = **.** then missing(i,j) = **0**;

end;

end;

/\* Make an array for 8 variable sums \*/

array mysum (**8**);

/\*Make each variable, being sure to ignore zeroes and > 1\*/

do k = **1** to **8**;

mysum(k) = missing(k,**1**) + missing(k,**2**) + missing(k,**3**) + missing(k,**4**) + missing(k,**5**);

end;

/\*Now if the variable is not zero or > 1 create var \*/

array myvar(**8**);

do m = **1** to **8**;

if mysum(m) = **1** then

myvar(m) = (missing(m,**1**)\***5**) + (missing(m,**2**)\***4**) + (missing(m,**3**)\***3**) + (missing(m,**4**)\***2**) + (missing(m,**5**)\***1**);

else

myvar(m) = **.**;

end;

/\* Now make variable names pretty again\*/

familyLife = myvar(**1**);

provide = myvar(**2**);

familyHome = myvar(**3**);

familyTime = myvar(**4**);

luxury = myvar(**5**);

new = myvar(**6**);

tech = myvar(**7**);

carLooks = myvar(**8**);

**run**;

/\* Keep only the neccesary variables and give them labels \*/

**data** drivers;

set DriverCleaning;

keep myid familyLife provide familyHome familyTime luxury new tech carLooks;

label familyLife = "Family life is the most important thing to me.";

label provide = "I like to provide my children with things that I didn't have as a child.";

label familyHome = "I like spending most of my time at home with my family.";

label familyTime = "I enjoy spending time with my family.";

label luxury = "I prefer driving luxury cars.";

label new = "I normally buy brand new cars.";

label tech = "The technologies offered in a car’s dashboard influences my decision to buy it.";

label carLooks = "I choose a car mainly on the basis of looks.";

**run**;

/\* Now go grab the variables for clustering \*/

**data** rawClusters;

infile rawData;

input myid **1**-**7**

/\*Advertising helps me learn about the products companies have to offer.\*/

advLearn\_agree\_alot **5780**

advLearn\_agree\_little **5825**

advLearn\_neither **5915**

advLearn\_disagree\_little **5960**

advLearn\_disagree\_alot **6005**

/\* I love to buy new gadgets and appliances \*/

gadgets\_agree\_alot **6954**

gadgets\_agree\_little **6971**

gadgets\_neither **7005**

gadgets\_disagree\_little **7022**

gadgets\_disagree\_alot **7039**

/\* I am worried about pollution and congestion caused by cars. \*/

pollu\_agree\_alot **4193**

pollu\_agree\_little **4207**

pollu\_neither **4235**

pollu\_disagree\_little **4249**

pollu\_disagree\_alot **4263**

/\* I am more likely to purchase products I see advertised on a social media/networking website.\*/

advSocialMedia\_agree\_alot **6845**

advSocialMedia\_agree\_little **6860**

advSocialMedia\_neither **6890**

advSocialMedia\_disagree\_little **6905**

advSocialMedia\_disagree\_alot **6920**;

**run**;

/\* Now clean up the clustering variables \*/

/\* use an array to convert missing values to zeros \*/

**data** ClustersCleaning;

set rawClusters;

array missing (**4**,**5**)

advLearn\_agree\_alot

advLearn\_agree\_little

advLearn\_neither

advLearn\_disagree\_little

advLearn\_disagree\_alot

gadgets\_agree\_alot

gadgets\_agree\_little

gadgets\_neither

gadgets\_disagree\_little

gadgets\_disagree\_alot

pollu\_agree\_alot

pollu\_agree\_little

pollu\_neither

pollu\_disagree\_little

pollu\_disagree\_alot

advSocialMedia\_agree\_alot

advSocialMedia\_agree\_little

advSocialMedia\_neither

advSocialMedia\_disagree\_little

advSocialMedia\_disagree\_alot;

/\* Convert missing values to zeros \*/

do i = **1** to **4**;

do j = **1** to **5**;

if missing(i,j) = **.** then missing(i,j) = **0**;

end;

end;

/\* Make an array for 5 variable sums \*/

array mysum (**5**);

/\*Make each variable, being sure to ignore zeroes and > 1\*/

do k = **1** to **4**;

mysum(k) = missing(k,**1**) + missing(k,**2**) + missing(k,**3**) + missing(k,**4**) + missing(k,**5**);

end;

/\*Now if the variable is not zero or > 1 create var \*/

array myvar(**4**);

do m = **1** to **4**;

if mysum(m) = **1** then

myvar(m) = (missing(m,**1**)\***5**) + (missing(m,**2**)\***4**) + (missing(m,**3**)\***3**) + (missing(m,**4**)\***2**) + (missing(m,**5**)\***1**);

else

myvar(m) = **.**;

end;

/\* Now make variable names pretty again\*/

advLearn = myvar(**1**);

gadgets = myvar(**2**);

pollution = myvar(**3**);

advSocialMedia = myvar(**4**);

**run**;

**data** Clusters;

set ClustersCleaning;

keep myid advLearn gadgets pollution advSocialMedia;

label advLearn = "Advertising helps me learn about the products companies have to offer.";

label gadgets = "I love to buy new gadgets and appliances.";

label pollution = "I am worried about pollution and congestion caused by cars.";

label advSocialMedia = "I am more likely to purchase products I see advertised on a social media/networking website.";

**run**;

/\* Now grab Descripter variables \*/

**data** rawDescripters;

infile rawData;

input myid **1**-**7**

/\*Race \*/

Racewhite **2420**

Raceblack **2421**

Raceasian **2422**

Raceother **2423**

/\* Watch ESPN in last 7 days \*/

espn **9625**

/\* Next Vehicle\*/

nextNew **64921**

nextUsed **64922**

/\* Visited Chick-Fil-A \*/

chickfila **41753**;

**run**;

/\* Now clean up Descripter variables \*/

**data** DiscriptersCleaning;

set rawDescripters;

/\*Race \*/

if Racewhite = **.** then RaceWhite = **0**;

if Raceblack = **.** then Raceblack = **0**;

if Raceasian = **.** then Raceasian = **0**;

if Raceother = **.** then Raceother = **0**;

/\* Watch ESPN in last 7 days \*/

if espn = **.** then espn = **0**;

/\* Next Vehicle\*/

if nextNew = **1** then NextVehicle = **1**;

if nextUsed = **1** then NextVehicle = **0**;

/\* Visited Chick-Fil-A \*/

if chickfila = **.** then chickfila = **0**;

**run**;

**data** Descripters;

set DiscriptersCleaning;

keep myid Raceasian Raceblack Raceother Racewhite NextVehicle chickfila espn;

label Raceasian = "Race = Asian";

label Raceblack = "Race = Black";

label Racewhite = "Race = White";

label Raceother = "Race = Other";

label NextVehicle = "Next vehicle is new or used?";

label chickfila = "Visited Chick-Fil-A";

label espn = "Watch ESPN in the last 7 days?";

**run**;

/\* Join tables around my\_id. Sort them first \*/

**proc** **sort** data=Drivers;

by myid;

**proc** **sort** data=Clusters;

by myid;

**proc** **sort** data=Descripters;

by myid;

**data** DATA;

merge Drivers Clusters Descripters;

by myid;

**run**;

/\* Create various formats \*/

**proc** **format**;

value GenderScale **0** = "Male"

**1** = "Female";

**run**;

**proc** **format**;

value fivescale **5** = "agree alot"

**4** = "agree a little"

**3** = "neither agree nor disagree"

**2** = "disagree a little"

**1** = "disagree alot";

**run**;

**proc** **format**;

value Binaryscale **0** = "No"

**1** = "Yes";

**run**;

**proc** **format**;

value NewUsedScale **0** = "Used"

**1** = "New";

**run**;

/\* Create frequency tables for all variables \*/

**proc** **freq** data=DATA;

format Raceasian Raceblack Raceother Racewhite chickfila espn Binaryscale.;

format familyLife provide familyHome familyTime luxury new tech carLooks advLearn gadgets pollution advSocialMedia familyCar purchaseDec fivescale.;

format NextVehicle NewUsedScale.;

tables familyLife provide familyHome familyTime luxury new tech carLooks advLearn gadgets pollution advSocialMedia Raceasian Raceblack Raceother Racewhite NextVehicle chickfila espn;

**run**;

/\* Now run PCA on the two constructs \*/

**proc** **factor** data = DATA

maxiter = **100**

method = principal

mineigen = **1**

rotate = varimax

scree

score

print;

var familyLife provide familyHome familyTime luxury new tech carLooks;

**run**;

/\* Two factors are optimal \*/

/\* Output data \*/

**proc** **factor** data = DATA out=DATA2 nfactors=**2**

maxiter = **100**

method = principal

mineigen = **1**

rotate = varimax

scree

score

print;

var familyLife provide familyHome familyTime luxury new tech carLooks;

**run**;

/\* Now run cluster analysis \*/

**proc** **fastclus** data=DATA2 maxclusters=**2**;

var advLearn gadgets pollution advSocialMedia Factor1;

**proc** **fastclus** data=DATA2 maxclusters=**3**;

var advLearn gadgets pollution advSocialMedia Factor1;

**proc** **fastclus** data=DATA2 maxclusters=**4**;

var advLearn gadgets pollution advSocialMedia Factor1;

**proc** **fastclus** data=DATA2 maxclusters=**5**;

var advLearn gadgets pollution advSocialMedia Factor1;

**proc** **fastclus** data=DATA2 maxclusters=**6**;

var advLearn gadgets pollution advSocialMedia Factor1;

**proc** **fastclus** data=DATA2 maxclusters=**7**;

var advLearn gadgets pollution advSocialMedia Factor1;

**run**;

/\* K-means clustering says 4 clusters\*/

/\* Now run GAP analysis \*/

**proc** **hpclus** data=DATA2 maxclusters=**7**

noc=abc(b=**20** minclusters=**2** align=pca criterion=firstpeak);

Score out=mycluster;

input advLearn gadgets pollution advSocialMedia Factor1 / level=interval;

id myid;

**run**;

/\* GAP says 2. Go with 4 \*/

/\* 4 clusters are optimal \*/

**proc** **fastclus** data=DATA2 maxclusters=**4** out=myclus;

var advLearn gadgets pollution advSocialMedia Factor1;

**run**;

/\* Run means on descriptors \*/

**proc** **sort** data=myclus;

by cluster;

**proc** **means** data=myclus;

var Raceasian Raceblack Raceother Racewhite NextVehicle chickfila espn;

by cluster;

**run**;