Loading in Questions

/\*creating library for project and uploading the raw dataset\*/

libname mylib 'C:\Users\imo141\Dropbox\GitHub Code\Market Segmentation\Data';

filename o\_data 'C:\Users\imo141\Dropbox\GitHub Code\Market Segmentation\Data\FA12\_Data.txt' lrecl = **65576**;

/\*pulling in id number and 8 sample question with

4 being centered around the environment and 4 on fashion \*/

**data** temp1;

infile o\_data;

input my\_id **1**-**7**

Friendly\_Agree\_Alot **4075**

Friendly\_Agree\_Little **4084**

Friendly\_Neither **4102**

Friendly\_Disagree\_Little **4111**

Friendly\_Disagree\_Alot **4120**

Responsible\_Agree\_Alot **4077**

Responsible\_Agree\_Little **4086**

Responsible\_Neither **4104**

Responsible\_Disagree\_Little **4113**

Responsible\_Disagree\_Alot **4122**

Conscious\_Agree\_Alot **4079**

Conscious\_Agree\_Little **4088**

Conscious\_Neither **4106**

Conscious\_Disagree\_Little **4115**

Conscious\_Disagree\_Alot **4124**

EcoQuality\_Agree\_Alot **4081**

EcoQuality\_Agree\_Little **4090**

EcoQuality\_Neither **4108**

EcoQuality\_Disagree\_Little **4117**

EcoQuality\_Disagree\_Alot **4126**

HighQuality\_Agree\_Alot **3320**

HighQuality\_Agree\_Little **3347**

HighQuality\_Neither **3401**

HighQuality\_Disagree\_Little **3428**

HighQuality\_Disagree\_Alot **3455**

KeepUp\_Agree\_Alot **3317**

KeepUp\_Agree\_Little **3344**

KeepUp\_Neither **3398**

KeepUp\_Disagree\_Little **3425**

KeepUp\_Disagree\_Alot **3452**

Brand\_Agree\_Alot **3327**

Brand\_Agree\_Little **3354**

Brand\_Neither **3408**

Brand\_Disagree\_Little **3435**

Brand\_Disagree\_Alot **3462**

Styles\_Agree\_Alot **3325**

Styles\_Agree\_Little **3352**

Styles\_Neither **3406**

Styles\_Disagree\_Little **3433**

Styles\_Disagree\_Alot **3460**;

**run**;

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

**data** mycalcs;

set temp1;

array missy(**8**,**5**)

Friendly\_Agree\_Alot

Friendly\_Agree\_Little

Friendly\_Neither

Friendly\_Disagree\_Little

Friendly\_Disagree\_Alot

Responsible\_Agree\_Alot

Responsible\_Agree\_Little

Responsible\_Neither

Responsible\_Disagree\_Little

Responsible\_Disagree\_Alot

Conscious\_Agree\_Alot

Conscious\_Agree\_Little

Conscious\_Neither

Conscious\_Disagree\_Little

Conscious\_Disagree\_Alot

EcoQuality\_Agree\_Alot

EcoQuality\_Agree\_Little

EcoQuality\_Neither

EcoQuality\_Disagree\_Little

EcoQuality\_Disagree\_Alot

HighQuality\_Agree\_Alot

HighQuality\_Agree\_Little

HighQuality\_Neither

HighQuality\_Disagree\_Little

HighQuality\_Disagree\_Alot

KeepUp\_Agree\_Alot

KeepUp\_Agree\_Little

KeepUp\_Neither

KeepUp\_Disagree\_Little

KeepUp\_Disagree\_Alot

Brand\_Agree\_Alot

Brand\_Agree\_Little

Brand\_Neither

Brand\_Disagree\_Little

Brand\_Disagree\_Alot

Styles\_Agree\_Alot

Styles\_Agree\_Little

Styles\_Neither

Styles\_Disagree\_Little

Styles\_Disagree\_Alot

;

/\* now make missing values zeroes \*/

do i = **1** to **8**; /\*1 to 8 here since I am analyzing 8 questions\*/

do j = **1** to **5**; /\*1 to 5 here since each question has a likert scale of 5 levels\*/

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

end;

end;

/\* make array for 8 variable sums \*/

array mysum(**8**);

/\* sum up the vars and make no mark or > 1 mark missing\*/

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

mysum(k) = missy(k,**1**)+ missy(k,**2**)+missy(k,**3**) + missy(k,**4**) + missy(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) = (missy(m,**1**)\***5**) + (missy(m,**2**)\***4**)+ (missy(m,**3**)\***3**) + (missy(m,**4**)\***2**) + (missy(m,**5**)\***1**);

else

myvar(m) = **.**;

end;

/\* changing var names so they're easy to interpret\*/

friendly = myvar(**1**);

response = myvar(**2**);

conscious = myvar(**3**);

eco\_quality = myvar(**4**);

high\_quality = myvar(**5**);

keep\_up = myvar(**6**);

brand = myvar(**7**);

styles = myvar(**8**);

/\* make sum variables available \*/

mysum1=mysum(**1**);

mysum2=mysum(**2**);

mysum3=mysum(**3**);

mysum4=mysum(**4**);

mysum5=mysum(**5**);

mysum6=mysum(**6**);

mysum7=mysum(**7**);

mysum8=mysum(**8**);

**run**;

/\* check incidence of multiple check marks \*/

**proc** **freq** data=mycalcs;

tables

mysum1

mysum2

mysum3

mysum4

mysum5

mysum6

mysum7

mysum8;

**run**;

/\* do freqs for binary vars \*/

**proc** **freq** data=mycalcs;

tables

Friendly\_Agree\_Alot

Friendly\_Agree\_Little

Friendly\_Neither

Friendly\_Disagree\_Little

Friendly\_Disagree\_Alot

Responsible\_Agree\_Alot

Responsible\_Agree\_Little

Responsible\_Neither

Responsible\_Disagree\_Little

Responsible\_Disagree\_Alot

Conscious\_Agree\_Alot

Conscious\_Agree\_Little

Conscious\_Neither

Conscious\_Disagree\_Little

Conscious\_Disagree\_Alot

EcoQuality\_Agree\_Alot

EcoQuality\_Agree\_Little

EcoQuality\_Neither

EcoQuality\_Disagree\_Little

EcoQuality\_Disagree\_Alot

HighQuality\_Agree\_Alot

HighQuality\_Agree\_Little

HighQuality\_Neither

HighQuality\_Disagree\_Little

HighQuality\_Disagree\_Alot

KeepUp\_Agree\_Alot

KeepUp\_Agree\_Little

KeepUp\_Neither

KeepUp\_Disagree\_Little

KeepUp\_Disagree\_Alot

Brand\_Agree\_Alot

Brand\_Agree\_Little

Brand\_Neither

Brand\_Disagree\_Little

Brand\_Disagree\_Alot

Styles\_Agree\_Alot

Styles\_Agree\_Little

Styles\_Neither

Styles\_Disagree\_Little

Styles\_Disagree\_Alot;

**run**;

/\* sniffing new variables and comparing across questions\*/

**proc** **freq** data=mycalcs;

tables

friendly

response

conscious

eco\_quality

high\_quality

keep\_up

brand

styles;

**run**;

/\*keeping only the variables I need to move forward\*/

**data** mylib.questions;

set mycalcs;

keep my\_id friendly response conscious eco\_quality high\_quality keep\_up brand styles;

**run**;

/\*exporting new file into excel\*/

**proc** **export**

data=mylib.questions

dbms=xlsx

outfile= "C:\Users\imo141\Dropbox\GitHub Code\Market Segmentation\questions.xlsx"

replace;

**run**;

PCA on Questions

/\*Loading in questions questions data from previous chunk\*/

**data** mylib.pcadata;

set mylib.questions;

**run**;

ods graphics on;

/\* running the factor analysis \*/

**proc** **factor** data = mylib.pcadata

maxiter=**100**

method=principal

mineigen=**1**

rotate=varimax

scree

score

print

nfactors=**2**

plots= all

out= mylib.pcadata2;

var

friendly

response

conscious

eco\_quality

high\_quality

keep\_up

brand

styles;

**run**;

/\*renaming the factor variables \*/

**data** mylib.pcadata2;

set mylib.pcadata2;

rename factor1 = Environment;

rename factor2 = Fashion;

**run**;

/\*exporting new file into excel\*/

**proc** **export**

data=mylib.pcadata2

dbms=xlsx

outfile= "C:\Users\imo141\Dropbox\GitHub Code\Market Segmentationpcadata.xlsx"

replace;

**run**;

Loading in Driver Variables

/\*pulling in id number and 4 driver questions to help split the base data into clusters\*/

**data** temp2;

infile o\_data;

input my\_id **1**-**7**

Recycle\_Agree\_Alot **4509**

Recycle\_Agree\_Little **4590**

Recycle\_Neither **4752**

Recycle\_Disagree\_Little **4833**

Recycle\_Disagree\_Alot **4914**

Cell\_Agree\_Alot **3755**

Cell\_Agree\_Little **3779**

Cell\_Neither **3827**

Cell\_Disagree\_Little **3851**

Cell\_Disagree\_Alot **3875**

Change\_Agree\_Alot **4525**

Change\_Agree\_Little **4606**

Change\_Neither **4768**

Change\_Disagree\_Little **4849**

Change\_Disagree\_Alot **4930**

Copy\_Agree\_Alot **6555**

Copy\_Agree\_Little **6582**

Copy\_Neither **6636**

Copy\_Disagree\_Little **6663**

Copy\_Disagree\_Alot **6690**;

**run**;

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

**data** mycalcs2;

set temp2;

array missy(**4**,**5**)

Recycle\_Agree\_Alot

Recycle\_Agree\_Little

Recycle\_Neither

Recycle\_Disagree\_Little

Recycle\_Disagree\_Alot

Cell\_Agree\_Alot

Cell\_Agree\_Little

Cell\_Neither

Cell\_Disagree\_Little

Cell\_Disagree\_Alot

Change\_Agree\_Alot

Change\_Agree\_Little

Change\_Neither

Change\_Disagree\_Little

Change\_Disagree\_Alot

Copy\_Agree\_Alot

Copy\_Agree\_Little

Copy\_Neither

Copy\_Disagree\_Little

Copy\_Disagree\_Alot;

/\* now make missing values zeroes \*/

do i = **1** to **4**; /\*1 to 4 here since I am using 4 driver questions\*/

do j = **1** to **5**; /\*1 to 5 here since each question has a likert scale of 5 levels\*/

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

end;

end;

/\* make array for 4 variable sums \*/

array mysum(**4**);

/\* sum up the vars and make no mark or > 1 mark missing\*/

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

mysum(k) = missy(k,**1**)+ missy(k,**2**)+missy(k,**3**) + missy(k,**4**) + missy(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) = (missy(m,**1**)\***5**) + (missy(m,**2**)\***4**)+ (missy(m,**3**)\***3**) + (missy(m,**4**)\***2**) + (missy(m,**5**)\***1**);

else

myvar(m) = **.**;

end;

/\* changing var names so they're easy to interpret\*/

recycle = myvar(**1**);

cell = myvar(**2**);

change = myvar(**3**);

copy = myvar(**4**);

/\* make sum variables available \*/

mysum1=mysum(**1**);

mysum2=mysum(**2**);

mysum3=mysum(**3**);

mysum4=mysum(**4**);

**run**;

/\* check incidence of multiple check marks \*/

**proc** **freq** data=mycalcs2;

tables

mysum1

mysum2

mysum3

mysum4;

**run**;

/\* do freqs for binary vars \*/

**proc** **freq** data=mycalcs2;

tables

Recycle\_Agree\_Alot

Recycle\_Agree\_Little

Recycle\_Neither

Recycle\_Disagree\_Little

Recycle\_Disagree\_Alot

Cell\_Agree\_Alot

Cell\_Agree\_Little

Cell\_Neither

Cell\_Disagree\_Little

Cell\_Disagree\_Alot

Change\_Agree\_Alot

Change\_Agree\_Little

Change\_Neither

Change\_Disagree\_Little

Change\_Disagree\_Alot

Copy\_Agree\_Alot

Copy\_Agree\_Little

Copy\_Neither

Copy\_Disagree\_Little

Copy\_Disagree\_Alot;

**run**;

/\* sniffing new variables and comparing across questions \*/

**proc** **freq** data=mycalcs2;

tables

recycle

cell

change

copy;

**run**;

**data** mylib.drivers;

set mycalcs2;

keep my\_id recycle cell change copy;

**run**;

/\*exporting new file into excel so I can continue in R\*/

**proc** **export**

data=mylib.drivers

dbms=xlsx

outfile= "C:\Users\imo141\Dropbox\GitHub Code\Market Segmentation\drivers.xlsx"

replace;

**run**;

K-Means Clustering

/\*uploading the needed datasets\*/

**data** mylib.letscluster;

set mylib.pcadata2;

set mylib.drivers;

**run**;

/\*keeping only the variables used in analysis\*/

**data** mylib.letscluster;

set mylib.letscluster;

KEEP my\_id Environment Fashion recycle cell;

**run**;

/\*exporting new file into excel so I can continue in R\*/

**proc** **export**

data=mylib.letscluster

dbms=xlsx

outfile= "C:\Users\imo141\Dropbox\GitHub Code\Market Segmentation\letscluster.xlsx"

replace;

**run**;

/\*Running k-means using 2-7 clusters to see what is ideal number of clusters for this dataset\*/

ods graphics on;

**proc** **fastclus** data=mylib.letscluster maxclusters=**2**;

var Environment Fashion recycle cell;

**run**;

**proc** **fastclus** data=mylib.letscluster maxclusters=**3**;

var Environment Fashion recycle cell;

**run**;

**proc** **fastclus** data=mylib.letscluster maxclusters=**4**;

var Environment Fashion recycle cell;

**run**;

**proc** **fastclus** data=mylib.letscluster maxclusters=**5**;

var Environment Fashion recycle cell;

**run**;

**proc** **fastclus** data=mylib.letscluster maxclusters=**6**;

var Environment Fashion recycle cell;

**run**;

**proc** **fastclus** data=mylib.letscluster maxclusters=**7**;

var Environment Fashion recycle cell;

**run**;

GAP Clustering

/\*uploading the needed datasets\*/

**data** mylib.letscluster;

set mylib.pcadata2;

set mylib.drivers;

**run**;

/\*keeping only the variables used in analysis\*/

**data** mylib.letscluster;

set mylib.letscluster;

KEEP my\_id Environment Fashion change copy recycle cell;

**run**;

ods graphics on;

/\*Running GAP using 2-6 clusters to see

what is ideal number of clusters for this dataset\*/

**proc** **hpclus** data=mylib.letscluster maxclusters=**6**

noc=abc(b=**25** minclusters=**2** align=pca criterion=firstpeak);

score out=mylib.hpclusOut1;

input Environment Fashion recycle cell/ level=interval;

id Environment Fashion recycle cell;

**run**;

**proc** **candisc** data=mylib.hpclusOut1 anova out=mylib.can;

class \_CLUSTER\_ID\_;

var Environment Fashion recycle cell;

**run**;

**proc** **sgplot** data=mylib.can;

scatter y=Can2 x=Can1 / group=\_CLUSTER\_ID\_ ;

**run**;

Loading in Non-Driver Variables and Gender

/\*pulling in id number, 5 nondriver questions, and gender to help identify some demographics on the clusters\*/

**data** temp33;

infile o\_data;

input my\_id **1**-**7**

people\_6\_mo **11653** /\*people mag\*/

time\_6\_mo **11687** /\*time mag\*/

facebook\_30\_days **9457** /\*espn mag\*/

nike\_12\_mo **42556** /\*purchased nike shoe in past year\*/

male **2281** /\*is male\*/

female **2282** /\*is female\*/;

**run**;

**data** mycalcs33;

set temp33;

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

array missy(**6**,**1**)

people\_6\_mo

time\_6\_mo

facebook\_30\_days

nike\_12\_mo

male

female;

/\* now make missing values zeroes \*/

do i = **1** to **6**;

do j = **1**;

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

end;

end;

**run**;

/\* do freqs for binary vars \*/

**proc** **freq** data=mycalcs33;

tables

people\_6\_mo

time\_6\_mo

facebook\_30\_days

nike\_12\_mo

male

female

;

**run**;

**data** mylib.nondrivers;

set mycalcs33;

keep my\_id people\_6\_mo time\_6\_mo facebook\_30\_days nike\_12\_mo male female;

**run**;

/\*exporting new file into excel so I can continue in R\*/

**proc** **export**

data=mylib.nondrivers

dbms=xlsx

outfile= "C:\Users\imo141\Dropbox\GitHub Code\Market Segmentation\nondrivers.xlsx"

replace;

**run**;

Loading in Age

/\*loading in age groups to help add demographic context\*/

**data** temp10;

infile o\_data;

input my\_id **1**-**7**

Age\_18\_24\_Agree\_Alot **2299**

Age\_25\_34\_Agree\_Little **2304**

Age\_35\_44\_Neither **2308**

Age\_45\_54\_Disagree\_Little **2312**

Age\_55\_Plus\_Disagree\_Alot **2315**;

**run**;

**data** mycalcs22;

set temp10;

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

array missy(**1**,**5**)

Age\_18\_24\_Agree\_Alot

Age\_25\_34\_Agree\_Little

Age\_35\_44\_Neither

Age\_45\_54\_Disagree\_Little

Age\_55\_Plus\_Disagree\_Alot;

/\* now make missing values zeroes \*/

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

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

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

end;

end;

/\* make array for 1 variable sums \*/

array mysum(**1**);

/\* sum up the vars and make no mark or > 1 mark missing\*/

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

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

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

end;

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

array myvar(**1**);

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

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

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

else

myvar(m) = **.**;

end;

/\* now make the var names pretty again \*/

age = myvar(**1**);

/\* make sum variables available \*/

mysum1=mysum(**1**);

**run**;

/\* check incidence of multiple check marks \*/

**proc** **freq** data=mycalcs22;

tables

mysum1;

**run**;

/\* do freqs for binary vars \*/

**proc** **freq** data=mycalcs22;

tables

Age\_18\_24\_Agree\_Alot

Age\_25\_34\_Agree\_Little

Age\_35\_44\_Neither

Age\_45\_54\_Disagree\_Little

Age\_55\_Plus\_Disagree\_Alot;

**run**;

/\* now sniff new variables and compare - should be approx same as cell counts \*/

**proc** **freq** data=mycalcs22;

tables

age;

**run**;

**data** mylib.age;

set mycalcs22;

keep my\_id age;

**run**;

/\*exporting new file into excel so I can continue in R\*/

**proc** **export**

data=mylib.age

dbms=xlsx

outfile= "C:\Users\imo141\Dropbox\GitHub Code\Market Segmentation\age.xlsx"

replace;

**run**;

Loading in Race

/\*loading in race groups to help add demographic context\*/

**data** temp8;

infile o\_data;

input my\_id **1**-**7**

Race\_White **2318**

Race\_Black **2319**

Race\_Asian **2320**

Race\_Other **2321**

Race\_Not\_BW **2322**;

**run**;

**data** testrun;

set temp8;

**run**;

/\*combining race)other and race\_not\_BW since there was no desciption on the difference\*/

**proc** **sql**;

create table work.testrun2 as

SELECT

(CASE

WHEN Race\_Other = **1** or Race\_Not\_BW = **1**

THEN **1**

Else **.** END) as Race\_Other2, \*

FROM testrun;

**run**;

/\*keeping only the variables needed for analysis\*/

**data** work.mycalcs9;

set work.testrun2;

drop Race\_Not\_BW Race\_Other;

**run**;

/\*renaming variables\*/

**data** mycalcs11;

set mycalcs9;

rename Race\_Other2 = Race\_Other;

**run**;

**data** mycalcs11;

set temp8;

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

array missy(**1**,**4**)

Race\_White

Race\_Black

Race\_Asian

Race\_Other;

/\* now make missing values zeroes \*/

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

do j = **1** to **4**;

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

end;

end;

/\* make array for 1 variable sums \*/

array mysum(**1**);

/\* sum up the vars and make no mark or > 1 mark missing\*/

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

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

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

end;

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

array myvar(**1**);

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

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

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

else

myvar(m) = **.**;

end;

/\* now make the var names pretty again \*/

race = myvar(**1**);

/\* make sum variables available \*/

mysum1=mysum(**1**);

**run**;

/\* check incidence of multiple check marks \*/

**proc** **freq** data=mycalcs11;

tables

mysum1;

**run**;

/\* do freqs for binary vars \*/

**proc** **freq** data=mycalcs11;

tables

Race\_White

Race\_Black

Race\_Asian

Race\_Other;

**run**;

/\* now sniff new variables and compare - should be approx same as cell counts \*/

**proc** **freq** data=mycalcs11;

tables

race;

**run**;

**data** mylib.race;

set mycalcs11;

keep my\_id race;

**run**;

/\*exporting new file into excel so I can continue in R\*/

**proc** **export**

data=mylib.race

dbms=xlsx

outfile= "C:\Users\imo141\Dropbox\GitHub Code\Market Segmentation\race.xlsx"

replace;

**run**;

Final Dataset

/\*combining all of the datasets\*/

**DATA** mylib.final;

SET mylib.hpclusOut1;

SET mylib.AGE;

SET mylib.NONDRIVERS;

SET mylib.RACE;

**RUN**;

/\*exporting new file into excel so I can continue in R\*/

**proc** **export**

data=mylib.final

dbms=xlsx

outfile= "C:\Users\imo141\Dropbox\GitHub Code\Market Segmentation\final.xlsx"

replace;

**run**;

/\*analyzing a proc freq for each non-driver and demographic variable to better describe each output cluster\*/

**proc** **freq** data=mylib.final;

tables \_cluster\_id\_ \* age;

**RUN**;

**proc** **freq** data=mylib.final;

tables \_cluster\_id\_ \* race;

**RUN**;

**proc** **freq** data=mylib.final;

tables \_cluster\_id\_ \* facebook\_30\_days;

**RUN**;

**proc** **freq** data=mylib.final;

tables \_cluster\_id\_ \* female;

**RUN**;

**proc** **freq** data=mylib.final;

tables \_cluster\_id\_ \* nike\_12\_mo;

**RUN**;

**proc** **freq** data=mylib.final;

tables \_cluster\_id\_ \* people\_6\_mo;

**RUN**;

**proc** **freq** data=mylib.final;

tables \_cluster\_id\_ \* time\_6\_mo;

**RUN**;