Mike Mahoney

Take Home Exam #1

DA6823

Market Segmentation

Due 3/5/18

1. Select a problem or business issue to apply your market segmentation to. This may be a particular brand, a category of products or services, or a particular industry that you create the market segmentation for. You will likely want to take the position of a particular company that makes that product or service – let me know what company that is. Be sure to state your business problem.

The company I am taking a position in is Hyundai Motor Company, but specifically their subsidiary Hyundai Motor America. Hyundai Motor America is a 32-year-old company, that had a reputation for producing cheap, unreliable vehicles until recently. During the last 10 years, Hyundai has revamped their compact, intermediate and full-size cars classes to gain market share; even adding product lines such as small and medium SUVs classes, and the Genesis (brand) to appeal to the luxury consumers. As a marketing manager for Hyundai Motor America, I want to create an advertising strategy aimed to show the United States consumer that Hyundai can be reliable but still appeal to the cost-conscious vehicle consumers.

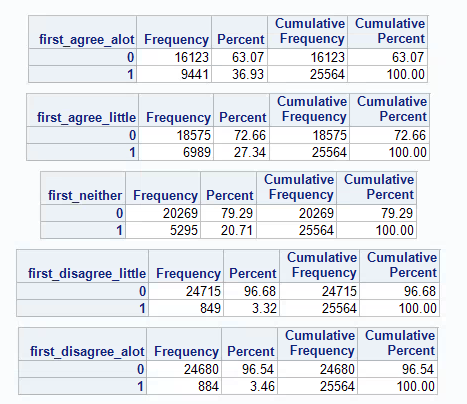
1. Select your target population that you are going to segment. Many times this just turns out to be the U.S. adult population 18 years of age or older (the entire NCS data set). Other times it might be something different such as adults aged 18-24 years or women 18-49 or people with digital tablets or frequent movie goers or tequila drinkers – the list is pretty endless. Be sure to tell me what your target population is and why you picked it.

I am targeting the US adult population 18 years and older because I want to capture as much of the market as I can. Realistically, I would be capturing all legal adults with valid driver’s licenses, anyone who drives in America and makes the decision on which vehicle to buy. We would be changing perceptions of the older adult population and creating brand recognition with the younger adult population, but our end goal is to find the best segment to aim advertisements to sell more vehicles in the United States.

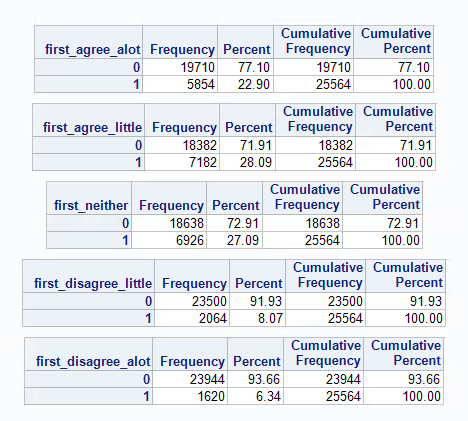
1. Clean and prepare the variables you are going to use in your segmentation system. This includes both driver variables in the statistical procedures as well profile variables. Please produce a frequency listing of all the variables you are using for the analysis so that I can see that the data are clean.

Output the factor scores for the next step:

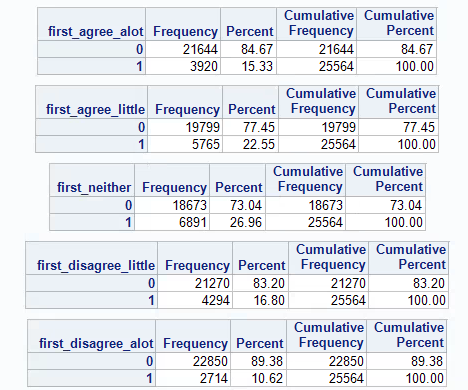
* true\_value frequency table



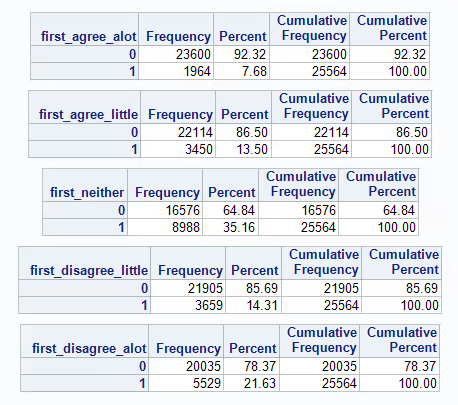
* before\_buy frequency table



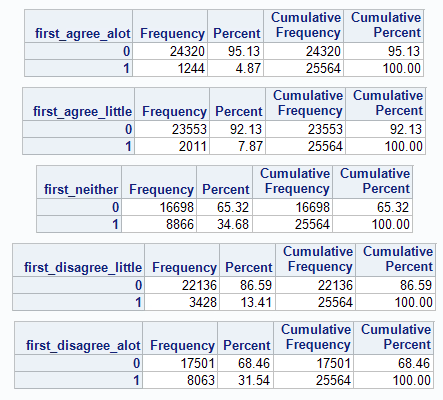
* car\_purpose frequency table



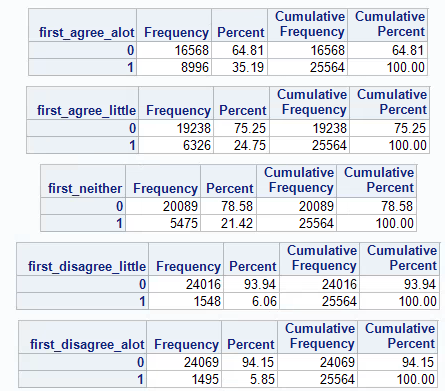
* foreign\_hq frequency table



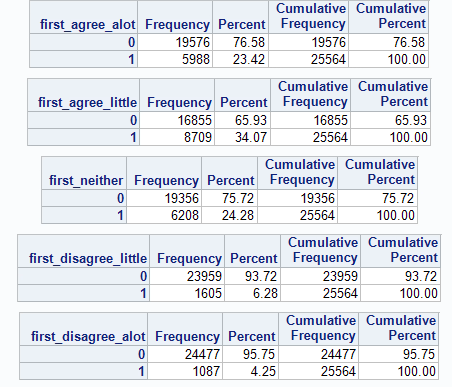
* foreign\_prestige frequency table



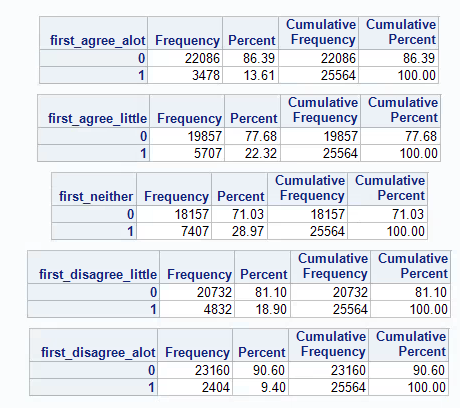
* plan\_ahead frequency table



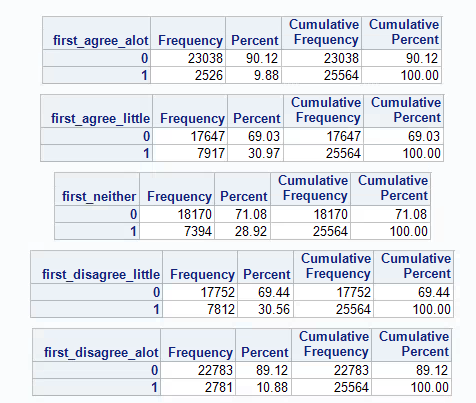
* special\_offers frequency table



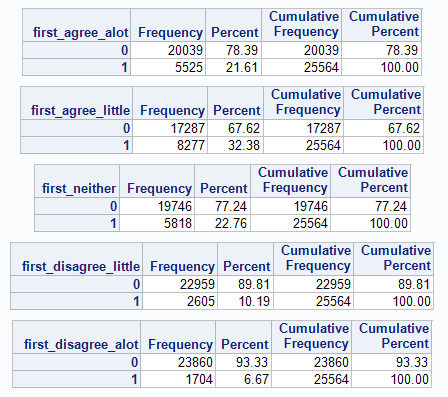
* unknown\_brands frequency table



* ask\_advice frequency table



* shop\_bargains frequency table



1. Be sure in your segmentation system to pick at least two sets of questions that measures two different constructs so you have to apply a factor analysis to the set of questions so that you can demonstrate your knowledge of factor analysis. Once that is done, the next step is to do a factor analysis of the two sets of questions that you selected that each measures a latent, unobserved construct.
   1. List out the questions that you selected to do the factor analysis on:

I chose two types of questions the first group is sentiment about vehicles: how the customer perceives a vehicle, if they prefer foreign vehicles, if they are looking for value. The second group focuses in on the value customer: do they look for bargains, do they ask advise for purchases, do they plan ahead for big purchases.

Vehicle Sentiment:

* + 1. The first variable I am using is labeled as “true\_value” in SAS, and reads: “The true value of a car is how long it lasts”. This variable is important because it implies our customers are looking for a vehicle that is reliable, dependable and are likely to keep this vehicle for many years.
    2. The second variable is “before\_buy” in SAS and reads “BEFORE BUYING CAR, I FIND OUT ABOUT THE CAR'S SAFETY RATING”. This variable implies our customer is seeking a safe vehicle
    3. The thirds is “car\_purpose” in SAS and reads “A CAR'S ONLY PURPOSE GET FROM POINT A TO POINT B”. This variable implies our customer only wants a vehicle to go from one place to another. They most likely do not care about the bells and whistles that are on the expensive vehicles, e.g. lane control/assists, back-up camera, and are more open for a less expensive vehicle.
    4. The fourth is “foreign\_hq” in SAS and reads “FOREIGN CARS ARE HIGHER QUALITY THAN AMERICAN”. This question is at the core of our business. If the customer perceives foreign manufactures over domestic, Hyundai would have an advantage (or not be eliminated from a customer) and could build brand loyalty.
    5. The fifth is “foreign in SAS and reads “OWNING A FOREIGN CAR IS MORE PRESTIGIOUS THAN OWNING AN AMERICAN CAR”. This implies the customer perceives value in a foreign car, Hyundai again would have an advantage. This could be an indicator of quality and value in a foreign car, an aspect Hyundai could capitalize on.

Value Customer:

* + 1. The first variable is ”plan\_ahead” and in SAS reads, “I GENERALLY PLAN FAR AHEAD TO BUY EXPENSIVE ITEMS SUCH AS AUTOMOBILES” . The variable implies the customer plans for their larger purposes, and often does research on the big-ticket items. This customer might be looking for a deal, the perfect product or just might be a value conscious customer.
    2. The second variable is “special\_offers” and in SAS reads “I ALWAYS LOOK OUT FOR SPECIAL OFFERS”. This variable specifically looks for the discount buyers. Hyundai does sell their vehicles at lower price points than other manufactures, so a customer looking for deals or cheaper products could be ideal.
    3. The thirds variable is “unknown\_brands” and in SAS reads, ““I DON'T BUY UNKNOWN BRANDS MERELY TO SAVE MONEY”. This variable is aiming at brand recognition/loyalty. Hyundai is a relatively new brand (or new rebrand) and has had a negative perception. Its possible to build a highly perceived brand for quality and affordability.
    4. The fourth variable is “ask\_advice” and in SAS reads, “I ASK PEOPLE ADVICE BEFORE BUYING NEW THINGS”. This vehicle implies someone does their research (on the internet maybe) or asks their network of acquaintances for perceptions on a certain vehicle or product. Building a high liked/perceived brand would spread by word of mouth, the most important uncontrolled factor in marketing.
    5. The fifth variable is “shop\_bargains” and in SAS read, “I SHOP AROUND ALOT TO TAKE ADVANTAGE OF SPECIALS OR BARGAINS”. This variables again is aiming for the value customer, someone looking for a good deal or bargain. This type of customer would be looking for a lower cost vehicle, at a price point Hyundai could service.
  1. Tell me what latent unobserved construct(s) you think they measure.
     1. I think the first latent unobserved construct is Foreign Car/ Reliability Fanatic
     2. I think the second is a Value Conscious Customer
  2. Decide which extraction technique to use and tell me why

I am going to use Principal Component Analysis (PCA) as an extraction technique to reduce the dimensions in the data. It uses the variance in the data, in which the most variance on relation to a straight line, is the first principal component. It allows us to group the data that has some correlations, so we can reduce the dimensions of the data into lower dimensions with still the same information. We will start out with 10 dimensions because we have 10 variables and reduce that down using the variance between closer points, therefore grouping the data (kind of).

* 1. Decide which rotation method you are going to use and why

I am going to use Promax rotational method because it is used for real world data (or data that does have correlation). Since we are using a real life survey/study, and I know there is some correlation in my variables, I think Promax is the better choice in this situation.

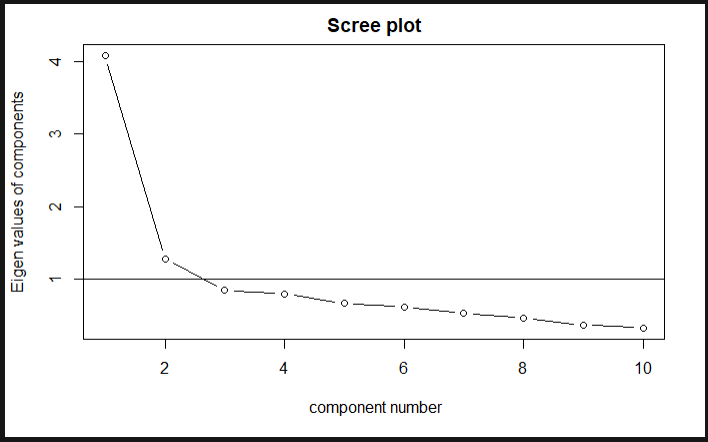
* 1. Run the factor analysis

It was run in R >> Practicum\_Exam1\_Final.Rmd

* 1. How many factors were extracted?

Two factors were extracted, because only two had Eigenvalues above 1

* 1. Make a pretty scree plot – how do you interpret it?



The horizontal line at 1 shows the cutoff of an Eigenvalue above 1. Here we have 2 components above 1, which indicate those two variables are explaining a large proportion of the variance. The remaining 8 components are not really of any value and are not explaining as much variance (they are not pulling their weight in the model).

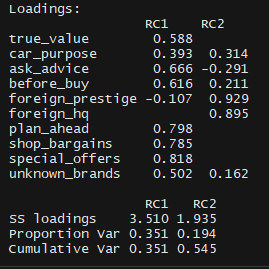
* 1. What criteria was used to determine number of factors? How does that work?

I used the Eigenvalues to determine which components were pulling their weight in the model. Only two components had eigenvalues pulling their weight, or explaining a large proportion of variance.

* 1. What percent of the variance is explained by the factors?

These two components explain about 54.5% of total variance, not great but these are the two most important components

* 1. Interpret the rotated factor matrix loadings and label the factor(s)



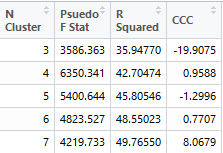
Our 10 dimensions of features was reduced to 2 dimensions or factors. The first factor explained the most variance and represents our bargain customers, or customers looking for a vehicle at a lower price point. These are our value customers and are seeking a vehicle without the bells and whistles. Our second factor explained the rest of the explainable variance and represents customers who believe a foreign car is superior over a domestic car.

1. The next step is the cluster analysis. Here I would like you to use at least one factor from the factor analysis in step 4 plus a few more relevant variables (say 3-4) to utilize in your cluster analysis. Here is how I would like you to proceed with the cluster analysis.
   1. You may wish to standardize the variables (other than the factor scores) to remove any influence due to scaling.

I standardized the variables >> Dataset : master\_ms\_stz (in R)

* 1. Pick a bracket of number of clusters (say from 3 to 7) and then run a k means cluster procedure to collect the following diagnostic statistics from the runs: R square, cubic clustering criterion (ccc) and pseudo F statistic (the proc FASTCLUS in SAS is the easiest way to obtain these but you can also get at least the ccc from R).

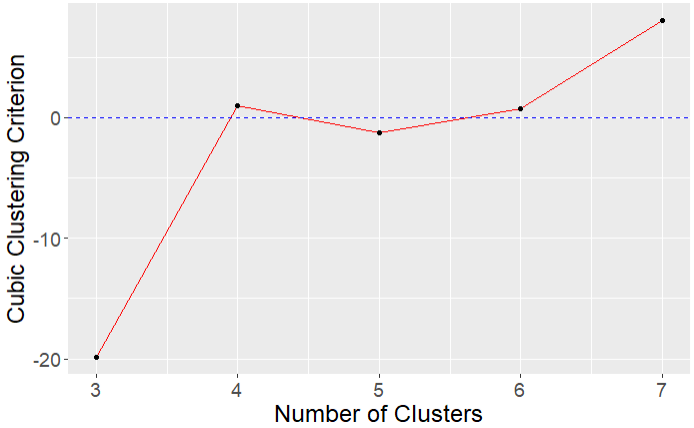
I chose a bracket from 3 to 7. I calculated the R-Squared, Pseudo F and Cubic Clustering Criterion >> See R file



* 1. Plot the diagnostic statistics against the number of clusters and apply the rules we discussed in class about the number of clusters that this graph might suggest to use (note that it may be the case that these diagnostics might not suggest any of the ones you ran. If so, note that).

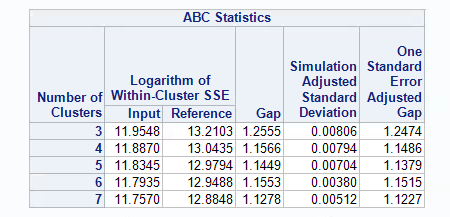






Our R squared is always going to go up if we add more clusters. The cubic clustering criterion (CCC) actually goes negative on Cluster 5, indicating the cost of the cluster outweigh the benefits of having the cluster. The Pseudo F statistic is highest at Cluster 4. We see the second highest peak in CCC at Cluster 4 (before going negative), our Voodoo chicken bones tell us Cluster 4 is where we begin.

* 1. Now perform a gap analysis on the data. Tell me how you know how many clusters you should select as your solution. Don’t point at the SAS output box that says here is the number of clusters – that’s too easy. Tell me the logic of how it works and then do it for your data.



Our desired cluster number = Gap(k) > Gap(k+1) – Simulated ADJ Standard Deviation

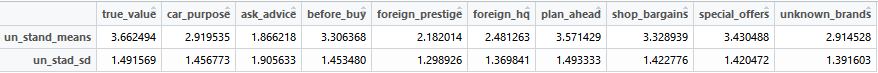
In our GAP analysis this occurs in cluster 6: Gap(5) > Gap(6) – Sim ADJ SD or

1.449 > 1.1553 – 0.0038 or 1.449 > 1.1515; which is false 1.449 is not greater than 1.1515, indicating cluster 6 is where we should stop. Our Gap analysis says that cluster 6 is further than 1 standard deviation from gap 5. Our intra-cluster variance will decrease as we add more clusters (it would be 0 with 10 clusters), but Cluster 6 is going to give us the lowest variance within a cluster (without every variable getting its own centroid).

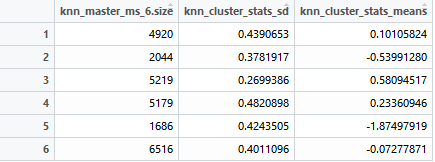
* 1. Run the cluster analysis with your choice of procedures.

See Practicum\_Exam1\_Final.Rmd >> Our final Cluster Analysis

* 1. Produce a table of means for the driver variables used in the cluster analysis. You should use the unstandardized variables for means wherever possible. Discuss how well the cluster analysis worked from your evaluation of this table.



Above is the non standadrized driver variables. Below is the standardized cluster means and standard deviations.



Our non standardized variables with the exception of ask\_advice are all relatively normal with standard deviations between 1.36 and 1.49 and means between 2.48 and 3.66. Ask advice has a low mean of 1.86 and higher standard deviation of 1.9. Our cluster analysis worked reasonably well, with the cluster means (centroids) relatively spaced away from each other. Our second largest cluster (3) with the lowest standard deviation of 0.26 and highest mean of 0.58. All of our clusters have standard deviations below 0.5, meaning each cluster has low intra-cluster variance. So I would say our clusters are adequately grouping our variables.

* 1. Select at least 3-4 more variables that are relevant to your business problem that are not drivers in the solution (be sure to include at least one demographic and one media variable) and construct a means table for each of these variables by cluster number. Note if you take the mean of a (0,1) variable it provides you with the proportion of individuals whose answered with the response coded 1.

The four non-drivers I chose:

1: Radio Formats-variety (cumulative Monday-Friday) >> If they are listening to the radio, they are probably driving (Mon-Fri, then they probably work or go to school)

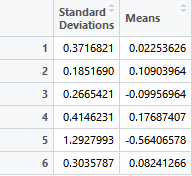
but I chose variety (like a Bob FM or Jack FM) because it captures a wide net of demographics

2: Driver's License-Currently have? - Yes >> We want to reach people with a driver’s license

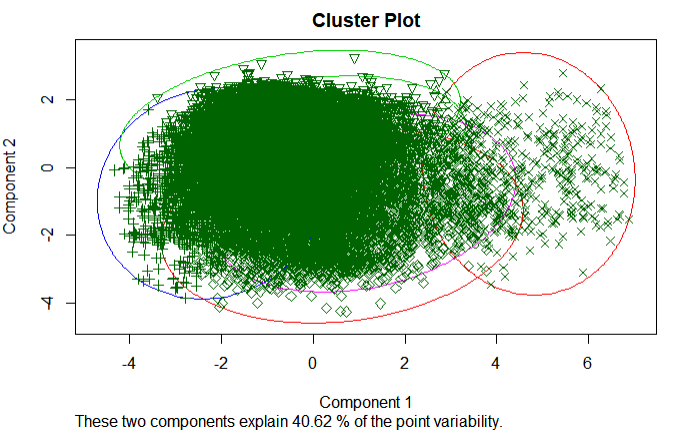
3: Demographics - Hispanic >> The Latino demographic is one of the largest and fastest growing across the US >> we need to capture this demographic

4: Most recently acquired- New/Used- Bought New >> We want individuals looking to buy a new car

This is a table of means and standard deviations of the non-driver variables, indexed from the original data set



* 1. Name the different clusters in your solution and provide a summary of the differences and similarities between the segments.



Above is our cluster plot of our 6 clusters, 5 clusters are overlapping each other almost completely overlapping. Only the most right cluster (cluster 4) is almost independent. This closely represents our factor analysis. I cannot adequately distinguish between the left clusters because they are so intertwined. If I were to name these two distinct groups, the left most group would be customers looking for a cheap reliable vehicles or our value/bargain customers (vars: true\_value, car\_purpose, ask\_advise, before\_buy, plan\_ahead, shop\_bargains, special\_offers and unknown\_brands). The distinct right group would be individuals who believe foreign cars are superior (vars: foreign\_prestiege, foreign\_hq)

* 1. Write a one paragraph executive summary of the segmentation system.

Our goal was to find distinct clusters that represent potential customers that would find the Hyundai brand favorable. The factor analysis reduced our 10 dimensions to 2 distinct factors, the first being the value or cost conscious customer and the second customers that think foreign cars are superior to domestic cars. Our cluster analysis gave mixed results to the optimal number of clusters. The Pseudo F statistic and CCC pointed to 4 clusters as optimal. The GAP analysis gave 6 clusters as optimal; I decided to use 6 clusters as a result because the GAP is usually the most accurate. When running the K means clustering with 6 clusters, I observed 5 clusters overlapping and 1 cluster was almost independent. The 6 observed clusters were synonymous with the 2 distinct factors. In conclusion, I believe I can provide Hyundai with insight with their customer segmentation, with two distinct groups. The first group is the core of our business, the value customer looking for a reliable vehicle at a lower price point. I believe Hyundai could fulfill those customer’s needs with its core products, the cars and SUV’s that are at a price point lower than current competitors, but are still reliable vehicles. The second group is customers that believe foreign vehicles are superior. Although these customers could be looking for luxury brands like BMW, Lexus and Mercedes, I believe Hyundai could fulfill these customer’s needs with the new Hyundai brand Genesis. The Genesis brand is Hyundai’s attempt to enter the luxury vehicle market, but still at a lower price point than the competitors (BMW, Lexus) and I believe that’s where this cluster overlaps with the larger group.

Appendix >> SAS Code

/\* Creating library called FA12 \*/

libname FA12 '\\Client\C$\Users\Mike\Documents\Practicum 1';

/\* Creating file called bigrec \*/

filename bigrec '\\Client\C$\Users\Mike\Documents\Practicum 1\FA12\_Data.txt' lrecl = **65576**;

/\*“TRUE VALUE OF A CAR IS HOW LONG A CAR LASTS” >> Reads 5 variables >>

Outputs a data frame with each my\_id's response to each of the 5 variables

With a "1" indicating a correct choice\*/

**data** true\_value;

infile bigrec;

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

first\_agree\_alot **3529**

first\_agree\_little **3565**

first\_neither **3637**

first\_disagree\_little **3673**

first\_disagree\_alot **3709**;

**run**;

/\* #true\_value Freq - Doing a Frequency for each of the 5 variables \*/

**proc** **freq** data=true\_value;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Change NA - Use an array to turn missing values to zeros \*/

**data** mycalcs\_1;

set true\_value;

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

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

/\* now make missy values 0 >> Loops through i (which is only one in our case because its one question) then

j (which is each question) >> turning the "." into "0" >> Using an if/then statement \*/

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

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

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

end;

end;

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

array mysum(**1**);

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

/\* now make each variable, meing sure to ignore zeros and larger than 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 greater than 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;

/\* Changing the variable name to first \*/

true\_value = myvar(**1**);

/\*Creates a sum of myvar \*/

mysum1 = mysum(**1**);

**run**;

/\* #Binary Freq Dist - Running a second freq distribution on to check if 1 - answered or 0- missing\*/

**proc** **freq** data=mycalcs\_1;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Multiple Box - Check indicies of multiple check marks >> To see if our surveyors followed instructions\*/

**proc** **freq** data=mycalcs\_1;

tables

mysum1;

**run**;

/\* #Sniffin' Vars - now sniff variables and compare - should be approx same as cell counts \*/

**proc** **freq** data=mycalcs\_1;

tables

true\_value;

**run**;

/\* Creating library called FA12 \*/

libname FA12 '\\Client\C$\Users\Mike\Documents\Practicum 1';

/\* Creating file called bigrec \*/

filename bigrec '\\Client\C$\Users\Mike\Documents\Practicum 1\FA12\_Data.txt' lrecl = **65576**;

/\*“BEFORE BUYING CAR, I FIND OUT ABOUT THE CAR'S SAFETY RATING” >> Reads 5 variables >>

Outputs a data frame with each my\_id's response to each of the 5 variables

With a "1" indicating a correct choice\*/

**data** before\_buy;

infile bigrec;

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

first\_agree\_alot **3509**

first\_agree\_little **3545**

first\_neither **3617**

first\_disagree\_little **3653**

first\_disagree\_alot **3689**;

**run**;

/\* #before\_buy Freq - Doing a Frequency for each of the 5 variables \*/

**proc** **freq** data=before\_buy;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Change NA - Use an array to turn missing values to zeros \*/

**data** mycalcs;

set before\_buy;

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

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

/\* now make missy values 0 >> Loops through i (which is only one in our case because its one question) then

j (which is each question) >> turning the "." into "0" >> Using an if/then statement \*/

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

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

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

end;

end;

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

array mysum(**1**);

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

/\* now make each variable, meing sure to ignore zeros and larger than 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 greater than 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;

/\* Changing the variable name to first \*/

before\_buy = myvar(**1**);

/\*Creates a sum of myvar \*/

mysum1 = mysum(**1**);

**run**;

/\* #Binary Freq Dist - Running a second freq distribution on to check if 1 - answered or 0- missing\*/

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

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Multiple Box - Check indicies of multiple check marks >> To see if our surveyors followed instructions\*/

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

tables

mysum1;

**run**;

/\* #Sniffin' Vars - now sniff variables and compare - should be approx same as cell counts \*/

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

tables

before\_buy;

**run**;

/\* Creating library called FA12 \*/

libname FA12 '\\Client\C$\Users\Mike\Documents\Practicum 1';

/\* Creating file called bigrec \*/

filename bigrec '\\Client\C$\Users\Mike\Documents\Practicum 1\FA12\_Data.txt' lrecl = **65576**;

/\*“A CAR'S ONLY PURPOSE GET FROM POINT A TO POINT B” >> Reads 5 variables >>

Outputs a data frame with each my\_id's response to each of the 5 variables

With a "1" indicating a correct choice\*/

**data** car\_purpose;

infile bigrec;

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

first\_agree\_alot **3510**

first\_agree\_little **3546**

first\_neither **3618**

first\_disagree\_little **3654**

first\_disagree\_alot **3690**;

**run**;

/\* #car\_purpose Freq - Doing a Frequency for each of the 5 variables \*/

**proc** **freq** data=car\_purpose;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Change NA - Use an array to turn missing values to zeros \*/

**data** mycalcs\_3;

set car\_purpose;

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

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

/\* now make missy values 0 >> Loops through i (which is only one in our case because its one question) then

j (which is each question) >> turning the "." into "0" >> Using an if/then statement \*/

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

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

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

end;

end;

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

array mysum(**1**);

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

/\* now make each variable, meing sure to ignore zeros and larger than 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 greater than 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;

/\* Changing the variable name to first \*/

car\_purpose = myvar(**1**);

/\*Creates a sum of myvar \*/

mysum1 = mysum(**1**);

**run**;

/\* #Binary Freq Dist - Running a second freq distribution on to check if 1 - answered or 0- missing\*/

**proc** **freq** data=mycalcs\_3;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Multiple Box - Check indicies of multiple check marks >> To see if our surveyors followed instructions\*/

**proc** **freq** data=mycalcs\_3;

tables

mysum1;

**run**;

/\* #Sniffin' Vars - now sniff variables and compare - should be approx same as cell counts \*/

**proc** **freq** data=mycalcs\_3;

tables

car\_purpose;

**run**;

/\* Creating library called FA12 \*/

libname FA12 '\\Client\C$\Users\Mike\Documents\Practicum 1';

/\* Creating file called bigrec \*/

filename bigrec '\\Client\C$\Users\Mike\Documents\Practicum 1\FA12\_Data.txt' lrecl = **65576**;

/\*“FOREIGN CARS ARE HIGHER QUALITY THAN AMERICAN” >> Reads 5 variables >>

Outputs a data frame with each my\_id's response to each of the 5 variables

With a "1" indicating a correct choice\*/

**data** foreign\_hq;

infile bigrec;

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

first\_agree\_alot **3513**

first\_agree\_little **3549**

first\_neither **3621**

first\_disagree\_little **3657**

first\_disagree\_alot **3693**;

**run**;

/\* #foreign\_hq Freq - Doing a Frequency for each of the 5 variables \*/

**proc** **freq** data=foreign\_hq;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Change NA - Use an array to turn missing values to zeros \*/

**data** mycalcs\_4;

set foreign\_hq;

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

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

/\* now make missy values 0 >> Loops through i (which is only one in our case because its one question) then

j (which is each question) >> turning the "." into "0" >> Using an if/then statement \*/

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

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

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

end;

end;

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

array mysum(**1**);

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

/\* now make each variable, meing sure to ignore zeros and larger than 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 greater than 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;

/\* Changing the variable name to first \*/

foreign\_hq = myvar(**1**);

/\*Creates a sum of myvar \*/

mysum1 = mysum(**1**);

**run**;

/\* #Binary Freq Dist - Running a second freq distribution on to check if 1 - answered or 0- missing\*/

**proc** **freq** data=mycalcs\_4;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Multiple Box - Check indicies of multiple check marks >> To see if our surveyors followed instructions\*/

**proc** **freq** data=mycalcs\_4;

tables

mysum1;

**run**;

/\* #Sniffin' Vars - now sniff variables and compare - should be approx same as cell counts \*/

**proc** **freq** data=mycalcs\_4;

tables

foreign\_hq;

**run**;

/\* Creating library called FA12 \*/

libname FA12 '\\Client\C$\Users\Mike\Documents\Practicum 1';

/\* Creating file called bigrec \*/

filename bigrec '\\Client\C$\Users\Mike\Documents\Practicum 1\FA12\_Data.txt' lrecl = **65576**;

/\*“OWNING A FOREIGN CAR IS MORE MORE PRESTIGIOUS THAN OWNING AN AMERICAN CAR” >> Reads 5 variables >>

Outputs a data frame with each my\_id's response to each of the 5 variables

With a "1" indicating a correct choice\*/

**data** foreign\_prestige;

infile bigrec;

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

first\_agree\_alot **3504**

first\_agree\_little **3540**

first\_neither **3612**

first\_disagree\_little **3648**

first\_disagree\_alot **3684**;

**run**;

/\* #foreign\_prestige Freq - Doing a Frequency for each of the 5 variables \*/

**proc** **freq** data=foreign\_prestige;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Change NA - Use an array to turn missing values to zeros \*/

**data** mycalcs\_5;

set foreign\_prestige;

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

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

/\* now make missy values 0 >> Loops through i (which is only one in our case because its one question) then

j (which is each question) >> turning the "." into "0" >> Using an if/then statement \*/

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

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

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

end;

end;

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

array mysum(**1**);

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

/\* now make each variable, meing sure to ignore zeros and larger than 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 greater than 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;

/\* Changing the variable name to first \*/

foreign\_prestige = myvar(**1**);

/\*Creates a sum of myvar \*/

mysum1 = mysum(**1**);

**run**;

/\* #Binary Freq Dist - Running a second freq distribution on to check if 1 - answered or 0- missing\*/

**proc** **freq** data=mycalcs\_5;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Multiple Box - Check indicies of multiple check marks >> To see if our surveyors followed instructions\*/

**proc** **freq** data=mycalcs\_5;

tables

mysum1;

**run**;

/\* #Sniffin' Vars - now sniff variables and compare - should be approx same as cell counts \*/

**proc** **freq** data=mycalcs\_5;

tables

foreign\_prestige;

**run**;

/\* Creating library called FA12 \*/

libname FA12 '\\Client\C$\Users\Mike\Documents\Practicum 1';

/\* Creating file called bigrec \*/

filename bigrec '\\Client\C$\Users\Mike\Documents\Practicum 1\FA12\_Data.txt' lrecl = **65576**;

/\*“I GENERALLY PLAN FAR AHEAD TO BUY EXPENSIVE ITEMS SUCH AS AUTOMOBILES >> Reads 5 variables >>

Outputs a data frame with each my\_id's response to each of the 5 variables

With a "1" indicating a correct choice\*/

**data** plan\_ahead;

infile bigrec;

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

first\_agree\_alot **7489**

first\_agree\_little **7513**

first\_neither **7561**

first\_disagree\_little **7585**

first\_disagree\_alot **7609**;

**run**;

/\* #plan\_ahead Freq - Doing a Frequency for each of the 5 variables \*/

**proc** **freq** data=plan\_ahead;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Change NA - Use an array to turn missing values to zeros \*/

**data** mycalcs\_6;

set plan\_ahead;

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

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

/\* now make missy values 0 >> Loops through i (which is only one in our case because its one question) then

j (which is each question) >> turning the "." into "0" >> Using an if/then statement \*/

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

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

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

end;

end;

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

array mysum(**1**);

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

/\* now make each variable, meing sure to ignore zeros and larger than 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 greater than 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;

/\* Changing the variable name to first \*/

plan\_ahead = myvar(**1**);

/\*Creates a sum of myvar \*/

mysum1 = mysum(**1**);

**run**;

/\* #Binary Freq Dist - Running a second freq distribution on to check if 1 - answered or 0- missing\*/

**proc** **freq** data=mycalcs\_6;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Multiple Box - Check indicies of multiple check marks >> To see if our surveyors followed instructions\*/

**proc** **freq** data=mycalcs\_6;

tables

mysum1;

**run**;

/\* #Sniffin' Vars - now sniff variables and compare - should be approx same as cell counts \*/

**proc** **freq** data=mycalcs\_6;

tables

plan\_ahead;

**run**;

/\* Creating library called FA12 \*/

libname FA12 '\\Client\C$\Users\Mike\Documents\Practicum 1';

/\* Creating file called bigrec \*/

filename bigrec '\\Client\C$\Users\Mike\Documents\Practicum 1\FA12\_Data.txt' lrecl = **65576**;

/\*“I ALWAYS LOOK OUT FOR SPECIAL OFFERS" >> Reads 5 variables >>

Outputs a data frame with each my\_id's response to each of the 5 variables

With a "1" indicating a correct choice\*/

**data** special\_offers;

infile bigrec;

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

first\_agree\_alot **7500**

first\_agree\_little **7524**

first\_neither **7572**

first\_disagree\_little **7596**

first\_disagree\_alot **7620**;

**run**;

/\* #special\_offers Freq - Doing a Frequency for each of the 5 variables \*/

**proc** **freq** data=special\_offers;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Change NA - Use an array to turn missing values to zeros \*/

**data** mycalcs\_7;

set special\_offers;

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

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

/\* now make missy values 0 >> Loops through i (which is only one in our case because its one question) then

j (which is each question) >> turning the "." into "0" >> Using an if/then statement \*/

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

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

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

end;

end;

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

array mysum(**1**);

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

/\* now make each variable, meing sure to ignore zeros and larger than 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 greater than 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;

/\* Changing the variable name to first \*/

special\_offers = myvar(**1**);

/\*Creates a sum of myvar \*/

mysum1 = mysum(**1**);

**run**;

/\* #Binary Freq Dist - Running a second freq distribution on to check if 1 - answered or 0- missing\*/

**proc** **freq** data=mycalcs\_7;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Multiple Box - Check indicies of multiple check marks >> To see if our surveyors followed instructions\*/

**proc** **freq** data=mycalcs\_7;

tables

mysum1;

**run**;

/\* #Sniffin' Vars - now sniff variables and compare - should be approx same as cell counts \*/

**proc** **freq** data=mycalcs\_7;

tables

special\_offers;

**run**;

/\* Creating library called FA12 \*/

libname FA12 '\\Client\C$\Users\Mike\Documents\Practicum 1';

/\* Creating file called bigrec \*/

filename bigrec '\\Client\C$\Users\Mike\Documents\Practicum 1\FA12\_Data.txt' lrecl = **65576**;

/\*“I DON'T BUY UNKNOWN BRANDS MERELY TO SAVE MONEY" >> Reads 5 variables >>

Outputs a data frame with each my\_id's response to each of the 5 variables

With a "1" indicating a correct choice\*/

**data** unknown\_brands;

infile bigrec;

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

first\_agree\_alot **7481**

first\_agree\_little **7505**

first\_neither **7553**

first\_disagree\_little **7577**

first\_disagree\_alot **7601**;

**run**;

/\* #unknown\_brands Freq - Doing a Frequency for each of the 5 variables \*/

**proc** **freq** data=unknown\_brands;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Change NA - Use an array to turn missing values to zeros \*/

**data** mycalcs\_8;

set unknown\_brands;

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

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

/\* now make missy values 0 >> Loops through i (which is only one in our case because its one question) then

j (which is each question) >> turning the "." into "0" >> Using an if/then statement \*/

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

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

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

end;

end;

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

array mysum(**1**);

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

/\* now make each variable, meing sure to ignore zeros and larger than 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 greater than 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;

/\* Changing the variable name to first \*/

unknown\_brands = myvar(**1**);

/\*Creates a sum of myvar \*/

mysum1 = mysum(**1**);

**run**;

/\* #Binary Freq Dist - Running a second freq distribution on to check if 1 - answered or 0- missing\*/

**proc** **freq** data=mycalcs\_8;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Multiple Box - Check indicies of multiple check marks >> To see if our surveyors followed instructions\*/

**proc** **freq** data=mycalcs\_8;

tables

mysum1;

**run**;

/\* #Sniffin' Vars - now sniff variables and compare - should be approx same as cell counts \*/

**proc** **freq** data=mycalcs\_8;

tables

unknown\_brands;

**run**;

/\* Creating library called FA12 \*/

libname FA12 '\\Client\C$\Users\Mike\Documents\Practicum 1';

/\* Creating file called bigrec \*/

filename bigrec '\\Client\C$\Users\Mike\Documents\Practicum 1\FA12\_Data.txt' lrecl = **65576**;

/\*“I ASK PEOPLE ADVICE BEFORE BUYING NEW THINGS" >> Reads 5 variables >>

Outputs a data frame with each my\_id's response to each of the 5 variables

With a "1" indicating a correct choice\*/

**data** ask\_advice;

infile bigrec;

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

first\_agree\_alot **7490**

first\_agree\_little **7514**

first\_neither **7562**

first\_disagree\_little **7556**

first\_disagree\_alot **7610**;

**run**;

/\* #ask\_advice Freq - Doing a Frequency for each of the 5 variables \*/

**proc** **freq** data=ask\_advice;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Change NA - Use an array to turn missing values to zeros \*/

**data** mycalcs\_9;

set ask\_advice;

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

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

/\* now make missy values 0 >> Loops through i (which is only one in our case because its one question) then

j (which is each question) >> turning the "." into "0" >> Using an if/then statement \*/

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

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

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

end;

end;

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

array mysum(**1**);

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

/\* now make each variable, meing sure to ignore zeros and larger than 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 greater than 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;

/\* Changing the variable name to first \*/

ask\_advice = myvar(**1**);

/\*Creates a sum of myvar \*/

mysum1 = mysum(**1**);

**run**;

/\* #Binary Freq Dist - Running a second freq distribution on to check if 1 - answered or 0- missing\*/

**proc** **freq** data=mycalcs\_9;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Multiple Box - Check indicies of multiple check marks >> To see if our surveyors followed instructions\*/

**proc** **freq** data=mycalcs\_9;

tables

mysum1;

**run**;

/\* #Sniffin' Vars - now sniff variables and compare - should be approx same as cell counts \*/

**proc** **freq** data=mycalcs\_9;

tables

ask\_advice;

**run**;

/\* Creating library called FA12 \*/

libname FA12 '\\Client\C$\Users\Mike\Documents\Practicum 1';

/\* Creating file called bigrec \*/

filename bigrec '\\Client\C$\Users\Mike\Documents\Practicum 1\FA12\_Data.txt' lrecl = **65576**;

/\*“I SHOP AROUND ALOT TO TAKE ADVANTAGE OF SPECIALS OR BARGAINS" >> Reads 5 variables >>

Outputs a data frame with each my\_id's response to each of the 5 variables

With a "1" indicating a correct choice\*/

**data** shop\_bargains;

infile bigrec;

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

first\_agree\_alot **7480**

first\_agree\_little **7504**

first\_neither **7552**

first\_disagree\_little **7576**

first\_disagree\_alot **7600**;

**run**;

/\* #shop\_bargains Freq - Doing a Frequency for each of the 5 variables \*/

**proc** **freq** data=shop\_bargains;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Change NA - Use an array to turn missing values to zeros \*/

**data** mycalcs\_10;

set shop\_bargains;

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

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

/\* now make missy values 0 >> Loops through i (which is only one in our case because its one question) then

j (which is each question) >> turning the "." into "0" >> Using an if/then statement \*/

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

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

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

end;

end;

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

array mysum(**1**);

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

/\* now make each variable, meing sure to ignore zeros and larger than 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 greater than 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;

/\* Changing the variable name to first \*/

shop\_bargains = myvar(**1**);

/\*Creates a sum of myvar \*/

mysum1 = mysum(**1**);

**run**;

/\* #Binary Freq Dist - Running a second freq distribution on to check if 1 - answered or 0- missing\*/

**proc** **freq** data=mycalcs\_10;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* #Multiple Box - Check indicies of multiple check marks >> To see if our surveyors followed instructions\*/

**proc** **freq** data=mycalcs\_10;

tables

mysum1;

**run**;

/\* #Sniffin' Vars - now sniff variables and compare - should be approx same as cell counts \*/

**proc** **freq** data=mycalcs\_10;

tables

shop\_bargains;

**run**;

**proc** **hpclus** data=work.master\_of\_tables maxclusters=**7**

noc=abc(b=**30** minclusters=**3** align=pca criterion=firstpeak);

input

ask\_advice before\_buy car\_purpose foreign\_prestige foreign\_hq plan\_ahead shop\_bargains special\_offers

true\_value unknown\_brands;

**run**;

/\* Creating library called FA12 \*/

libname FA12 '\\Client\C$\Users\Mike\Documents\Practicum 1';

/\* Creating file called bigrec \*/

filename bigrec '\\Client\C$\Users\Mike\Documents\Practicum 1\FA12\_Data.txt' lrecl = **65576**;

/\*“Adding 4 non driver variables >>

1:Radio Formats-variety (cumulative Monday-Friday) >> If they are listening to the radio, they are probably driving

(Mon-Fri, then they probably work or go to school)

but I chose variety (like a Bob FM or Jack FM) because it captures a wide net of demographics

2: Driver's Liscense-Curently have? - Yes >> Because we want people who can leagally drive

3: Demographics - Hispanic >> The Latino demographic is one of the largest and fastest growing across

the US >> we need to capture this demographic

4: Most recently aquaired- New/Used- Bought New >>> We want customers that buy new vehicles

Reads these 4 variables >>

Outputs a data frame with each my\_id's response to each of the 4 variables

With a "1" indicating a correct choice\*/

**data** non\_drivers;

infile bigrec;

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

variety\_radio **22724**

drivers\_license\_yes **64789**

hispanic\_yes **2271**

recent\_car\_new **64764**;

**run**;

/\* #car\_purpose Freq - Doing a Frequency for each of the 5 variables \*/

**proc** **freq** data=non\_drivers;

tables

variety\_radio

drivers\_license\_yes

hispanic\_yes

recent\_car\_new;

**run**;

R code >> Note: This is from a R markdown file

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

library(haven)

library(dplyr)

library(standardize)

library(NbClust)

library(cluster)

library(ggplot2)

library(psych)

library(matrixStats)

library(fpc)

```

## Loading In our Exported SAS Datasets >> For Practicum 1 Exam

```{r}

true\_value <- read\_sas("C:/Users/Mike/Documents/Practicum 1/Exam\_1/var\_data\_sets/true\_value\_export.sas7bdat")

car\_purpose <- read\_sas("C:/Users/Mike/Documents/Practicum 1/Exam\_1/var\_data\_sets/car\_purpose\_export.sas7bdat")

ask\_advice <- read\_sas("C:/Users/Mike/Documents/Practicum 1/Exam\_1/var\_data\_sets/ask\_advice\_export.sas7bdat")

before\_buy <- read\_sas("C:/Users/Mike/Documents/Practicum 1/Exam\_1/var\_data\_sets/before\_buy\_export.sas7bdat")

foreign\_prestige <- read\_sas("C:/Users/Mike/Documents/Practicum 1/Exam\_1/var\_data\_sets/foreign\_prestige\_export.sas7bdat")

foreign\_hq <- read\_sas("C:/Users/Mike/Documents/Practicum 1/Exam\_1/var\_data\_sets/foriegn\_hq\_export.sas7bdat")

plan\_ahead <- read\_sas("C:/Users/Mike/Documents/Practicum 1/Exam\_1/var\_data\_sets/plan\_ahead\_export.sas7bdat")

shop\_bargains <- read\_sas("C:/Users/Mike/Documents/Practicum 1/Exam\_1/var\_data\_sets/shop\_bargains\_export.sas7bdat")

special\_offers <- read\_sas("C:/Users/Mike/Documents/Practicum 1/Exam\_1/var\_data\_sets/special\_offers\_export.sas7bdat")

unknown\_brands <- read\_sas("C:/Users/Mike/Documents/Practicum 1/Exam\_1/var\_data\_sets/unknown\_brands\_export.sas7bdat")

```

## Subsetting our Data sets >> Getting rid of the Junk >> Only keeping "my\_id" and unique column

```{r}

true\_value <- true\_value[,c(1,13)]

car\_purpose <- car\_purpose[,c(1,13)]

ask\_advice <- ask\_advice[,c(1,13)]

before\_buy <- before\_buy[,c(1,13)]

foreign\_prestige <- foreign\_prestige[,c(1,13)]

foreign\_hq <- foreign\_hq[,c(1,13)]

plan\_ahead <- plan\_ahead[,c(1,13)]

shop\_bargains <- shop\_bargains[,c(1,13)]

special\_offers <- special\_offers[,c(1,13)]

unknown\_brands <- unknown\_brands[,c(1,13)]

```

## Merge our created data sets by "my\_id" >> to "master\_ms" >> Megre was not working >> So If found it easier to rbind >> THen remove the extra 9 "my\_id"s

>> Made sure the NAs were 0

```{r}

master\_ms <- NULL

master\_ms <- cbind(true\_value, car\_purpose, ask\_advice, before\_buy, foreign\_prestige, foreign\_hq, plan\_ahead, shop\_bargains, special\_offers, unknown\_brands)

master\_ms <- master\_ms[,c(1,2,4,6,8,10,12,14,16,18,20)] # removing the my\_id when all 10 vars we

master\_ms[is.na(master\_ms)] <- 0

pca\_master\_ms <- master\_ms[,2:11] # Removing "my\_id" just in case

```

## PCA >> Using master\_ms >> my\_id is indexed out >> To perform principal component analysis, to reduce our dimensions down from 10 to ?? >> Currently we have a 10 dimensional space which is ugly and extremely confusing (if you watch a Dr Kaku video) >> So we want to reduce this space to managable and understandable dimensions

```{r}

exam\_pca <- principal(master\_ms[,2:11], nfactors = 2, rotate= "promax", residuals = TRUE);exam\_pca

#exam\_pca <- prcomp(master\_ms[,2:11], scale. = TRUE)

names(exam\_pca)

exam\_pca$values

exam\_pca$n.obs

exam\_pca$loadings

exam\_pca$fit

```

## Using Scree plot in library(psych) >> It shows the Eigenvalues to determine the best number of factors to use

```{r}

scree\_plot <- VSS.scree(master\_ms[,2:11],main="Scree plot")

```

## #5 - Cluser Analysis >> Standarizing THe Predictor Variables >> Re adding "my\_id" and changing name from "V1"

```{r}

master\_ms\_stz <- master\_ms

master\_ms\_stz <- scale(as.matrix(master\_ms[,2:11]))

master\_ms\_stz <- cbind(master\_ms$my\_id, master\_ms\_stz)

colnames(master\_ms\_stz)[1] <- "my\_id"

#is.standardized(master\_ms\_stz[,2:11])

```

## K-means Clustering >> Going to try differenct "k"s >> knn\_master\_ms\_k

```{r}

knn\_master\_ms\_3 <- kmeans(master\_ms\_stz[,2:11], 3)#;knn\_master\_ms\_3 # 3 clusters

knn\_master\_ms\_4 <- kmeans(master\_ms\_stz[,2:11], 4)#;knn\_master\_ms\_4 # 4 clusters

knn\_master\_ms\_5 <- kmeans(master\_ms\_stz[,2:11], 5)#;knn\_master\_ms\_5 # 5 Clusters #### WINNER

knn\_master\_ms\_6 <- kmeans(master\_ms\_stz[,2:11], 6)#;knn\_master\_ms\_6 # 6 clusters

knn\_master\_ms\_7 <- kmeans(master\_ms\_stz[,2:11], 7)#;knn\_master\_ms\_7 # 7 Clusters, Ahhh haha haha

```

## Calculating thee Psudeo F >> Which is the Between Cluster Sum of Squared / Witin Cluster Sum of Squares >> KMeans inclues on the Value List >> betweenss: between cluster sum of squares, total.withinss: total within cluster sum of squares, length(knn\_master\_ms\_4$size): finds the number of clusters

```{r}

NBClust\_pseudo\_F\_3 <- (knn\_master\_ms\_3$betweenss/(length(knn\_master\_ms\_5$size)-1))/(knn\_master\_ms\_3$tot.withinss/(sum(knn\_master\_ms\_3$size)-length(knn\_master\_ms\_3$size)))

NBClust\_pseudo\_F\_4 <- (knn\_master\_ms\_4$betweenss/(length(knn\_master\_ms\_4$size)-1))/(knn\_master\_ms\_4$tot.withinss/(sum(knn\_master\_ms\_4$size)-length(knn\_master\_ms\_4$size)))

NBClust\_pseudo\_F\_5 <- (knn\_master\_ms\_5$betweenss/(length(knn\_master\_ms\_5$size)-1))/(knn\_master\_ms\_5$tot.withinss/(sum(knn\_master\_ms\_5$size)-length(knn\_master\_ms\_5$size)))

NBClust\_pseudo\_F\_6 <- (knn\_master\_ms\_6$betweenss/(length(knn\_master\_ms\_6$size)-1))/(knn\_master\_ms\_6$tot.withinss/(sum(knn\_master\_ms\_6$size)-length(knn\_master\_ms\_6$size)))

NBClust\_pseudo\_F\_7 <- (knn\_master\_ms\_7$betweenss/(length(knn\_master\_ms\_7$size)-1))/(knn\_master\_ms\_7$tot.withinss/(sum(knn\_master\_ms\_7$size)-length(knn\_master\_ms\_7$size)))

# Creating a list with our clusters Psuedo F

NBClust\_psudeo\_List <- c(NBClust\_pseudo\_F\_3, NBClust\_pseudo\_F\_4, NBClust\_pseudo\_F\_5, NBClust\_pseudo\_F\_6, NBClust\_pseudo\_F\_7)

Clusters\_List <- c(3:7)

Psudeo\_F\_DF <- data.frame(Clusters\_List, NBClust\_psudeo\_List)

```

## Calculating the R^2 >> In our clustering example >> Can be caluclated by Between cluster Sum of Squares / Total Sum of Squares >> for each cluster 3:7

```{r}

r\_sq\_3 <- knn\_master\_ms\_3$betweenss / knn\_master\_ms\_3$totss

r\_sq\_4 <- knn\_master\_ms\_4$betweenss / knn\_master\_ms\_4$totss

r\_sq\_5 <- knn\_master\_ms\_5$betweenss / knn\_master\_ms\_5$totss

r\_sq\_6 <- knn\_master\_ms\_6$betweenss / knn\_master\_ms\_6$totss

r\_sq\_7 <- knn\_master\_ms\_7$betweenss / knn\_master\_ms\_7$totss

r\_sq\_list <- c(r\_sq\_3, r\_sq\_4, r\_sq\_5, r\_sq\_6, r\_sq\_7)\*100

r\_sq\_DF <- data.frame(Clusters\_List, r\_sq\_list)

```

### K-means >> getting the cubic clustering criterion (ccc) >> it is the Value\_index >> and is $Best.nc >> Voodoo Chicken Bones

```{r}

knn\_master\_ms\_NBClust <- NbClust(master\_ms\_stz[,2:11], min.nc = 3, max.nc=7, method = "kmeans",

index="ccc")

knn\_master\_ms\_NBClust$Best.nc

knn\_master\_ms\_NBClust$All.index

knn\_master\_ms\_NBClust$Best.partition

CCC\_list <- knn\_master\_ms\_NBClust$All.index

```

## Create a dataframe with all of our diagnostic statistics >> Psuedo\_F, R-Squared, CCC and Cluster Numbers

```{r}

diag\_stats <- data.frame(Clusters\_List, NBClust\_psudeo\_List, r\_sq\_list, CCC\_list)#; diag\_stats

colnames(diag\_stats)[1] <- "N Cluster"

colnames(diag\_stats)[2] <- "Psuedo F Stat"

colnames(diag\_stats)[3] <- "R Squared"

colnames(diag\_stats)[4] <- "CCC"

diag\_stats

```

###### Using GGPLOT to plots some graphs

## R-Squared

```{r}

R\_Squared\_plot <- ggplot(data = diag\_stats, aes(x =diag\_stats[1], y = diag\_stats[3])) +

geom\_line(color = "purple") +

geom\_point() +

xlab("Number of Clusters") +

ylab("R Squared in Percentages") +

theme(text = element\_text(size=18))

R\_Squared\_plot

```

## Psuedo F Statistic Plot

```{r}

Psuedo\_F\_plot <- ggplot(data = diag\_stats, aes(x =diag\_stats[1], y = diag\_stats[2])) +

geom\_line(color = "blue") +

geom\_point() +

xlab("Number of Clusters") +

ylab("Psuedo F Statistic") +

theme(text = element\_text(size=18))

Psuedo\_F\_plot

```

## CCC\_plot

```{r}

CCC\_plot <- ggplot(data = diag\_stats, aes(x =diag\_stats[1], y = diag\_stats[4])) +

geom\_line(color = "red") +

geom\_point() +

geom\_hline(yintercept = 0, linetype="dashed", color = "blue") +

xlab("Number of Clusters") +

ylab("Cubic Clustering Criterion") +

theme(text = element\_text(size=18))

CCC\_plot

```

## Trying another GAP >> Trying to get 7 >> Takes 23 minutes

#```{r}

start\_clock <- proc.time()

gap\_cluster\_update <- clusGap(master\_ms\_stz[,2:11], FUN = kmeans, K.max = 6, B = 5)

end\_clock <- proc.time() - start\_clock

#```

## Getting values from >> gap\_cluster\_update >> Tab

```{r}

gap\_cluster\_update$Tab

gap\_cluster\_update$call

#gap\_cluster\_update$FUNcluster

gap\_cluster\_update$B

```

```{r}

K <- maxSE(gap\_cluster\_update$Tab[, "gap"], gap\_cluster\_update$Tab[, "SE.sim"], method = "Tibs2001SEmax")

gap\_results\_1 <- kmeans(master\_ms\_stz[,2:11], K)

```

###### Our final Cluster Analysis >> Using K means >> Utilizing 6 clusters obtained from our GAP analysis

```{r}

knn\_master\_ms\_6 <- kmeans(master\_ms\_stz[,2:11], 6)

knn\_master\_ms\_6$centers

knn\_master\_ms\_6$size

```

## Crating a dataframe: knn\_cluster\_stats\_means\_sd\_df >> With each row Standard Deviation and Means >> For the standardized variables

```{r}

knn\_cluster\_stats <- as.matrix(knn\_master\_ms\_6$centers)

knn\_cluster\_stats\_means<- rowMeans(knn\_cluster\_stats)

knn\_cluster\_stats\_sd <- rowSds(knn\_cluster\_stats)

knn\_cluster\_stats\_means\_sd\_df <- data.frame(knn\_master\_ms\_6$size, knn\_cluster\_stats\_sd, knn\_cluster\_stats\_means)

```

## Creating a dataframe: knn\_6\_means\_sd\_df >> With each column Standard Deviation and Means >> For the standardized variables

```{r}

knn\_master\_ms\_6$centers

knn\_6\_df <- as.matrix(knn\_master\_ms\_6$centers)

knn\_6\_means <- colMeans(knn\_6\_df)

knn\_6\_sd <- colSds(knn\_6\_df)

knn\_6\_means\_sd\_df <- data.frame(knn\_6\_sd, knn\_6\_means)

```

## Getting un standarized cluster means >> Putting in a dataframe: un\_stand\_means\_df

```{r}

un\_stand\_means <- colMeans(master\_ms[,2:11])

un\_stad\_sd <- colSds(as.matrix(master\_ms[,2:11]))

un\_stand\_means\_df <- data.frame(un\_stand\_means, un\_stad\_sd)

un\_stand\_means\_df <- t(un\_stand\_means\_df) # Transposing our dataframe to save space

```

## Importing our non-drivers >> Creating a master data set of our drivers and non\_drivers: master\_ms\_non\_drivers

```{r}

non\_drivers <- read\_sas("C:/Users/Mike/Documents/Practicum 1/Exam\_1/var\_data\_sets/non\_drivers\_1.sas7bdat")

non\_drivers[is.na(non\_drivers)] <- 0

#master\_ms\_non\_drivers <- cbind(master\_ms, non\_drivers[2:5])

```

## Finding mean on non-drivers

```{r}

non\_drivers\_col\_means <- colMeans(non\_drivers[2:5])

non\_drivers\_means\_df <- data.frame(non\_drivers\_col\_means)

```

## Standardizing the Data Frame: master\_ms\_non\_drivers >> Which contains our drivers and our non drivers

```{r}

non\_drivers\_means\_stz <- scale(as.matrix(non\_drivers\_means\_df))

master\_ms\_non\_drivers\_stz <- cbind(master\_ms\_stz, scale(non\_drivers[2:5]))

```

### Kmeans Clusering with our non\_drivers >> using: master\_ms\_non\_drivers\_stz >> Contains Standardized Drivers and Non-Driver variables

```{r}

knn\_master\_both\_6 <- kmeans(master\_ms\_non\_drivers\_stz[,2:15], 6)#;knn\_master\_both\_6

knn\_master\_both\_6$centers#[11:14]

knn\_master\_both\_6$size

colnames(Knn\_master\_both\_6\_means\_sd\_df)[1] <- "Standard Deviations"

colnames(Knn\_master\_both\_6\_means\_sd\_df)[2] <- "Means"

```

### Cluster Plot

```{r}

clust\_plot <- clusplot(master\_ms\_non\_drivers\_stz[,2:15], knn\_master\_both\_6$cluster, color = TRUE, lines = 0, main = "Cluster Plot")

```