

STAT 6348
Applied Multivariate Analysis
Spring 2024
Project 3

This project is individual work. So do not consult with anybody in or out of class. You can ask me questions.

Sign on this page below and attach with your project. Your project will not be graded without it.

This project is entirely my work. I have not discussed this project with anybody in or out of class. I understand and have complied with the academic integrity policies written in the Handbook of Operating Procedures of UT Dallas <https://policy.utdallas.edu/utdsp5003>.

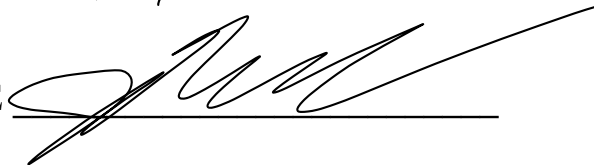
YOUR NAME

John Kenney

DATE

5/6/2024

YOUR SIGNATURE



Directions:

- Arrange your report in the following order: (1) Answers, (2) Outputs, and (3) Code. If using R markdown, code and output may be interspersed. Otherwise, attach your R script at the end by simply copying-pasting (just the commands – not the whole output) with comments included to indicate the parts.
- Type your answers including comments. These should not exceed more than 5 pages.
- Attach only the part of output that has been asked for in the question. Points may be deducted for attaching irrelevant output.
- At the end, attach your R script (just the commands – not the whole output) with comments included to indicate the parts and main steps.
- Your plots should be properly labeled and in a presentable format, e.g., use informative variable names (instead of V1, V2, etc), titles, axes labels, plotting symbols, colors, and legend. Do not present very small plots where it is difficult to see the pattern or plots that don't use the plotting region effectively (e.g., with large white space).

Question 1: Consider the Algerian forest fire dataset. The dataset includes 244 instances that are from two regions of Algeria, namely the Bejaia region located in the northeast of Algeria and the Sidi Bel-abbes region located in the northwest of Algeria. The 244 instances have been classified into fire (138 classes) and not fire (106 classes) classes.

Variable Information:

1. Date : (DD/MM/YYYY) Day, month ('june' to 'september'), year (2012)

Weather data observations

2. Temp : temperature noon (temperature max) in Celsius degrees: 22 to 42

3. RH : Relative Humidity in %: 21 to 90

4. Ws :Wind speed in km/h: 6 to 29

5. Rain: total day in mm: 0 to 16.8

FWI Components

6. Fine Fuel Moisture Code (FFMC) index from the FWI system: 28.6 to 92.5

7. Duff Moisture Code (DMC) index from the FWI system: 1.1 to 65.9

8. Drought Code (DC) index from the FWI system: 7 to 220.4

9. Initial Spread Index (ISI) index from the FWI system: 0 to 18.5

10. Buildup Index (BUI) index from the FWI system: 1.1 to 68

11. Fire Weather Index (FWI) Index: 0 to 31.1

12. Classes: two classes, namely fire and not fire.

Perform a principal components (PC) analysis of quantitative variables included in weather data observations and FWI components. How many PCs are needed to explain 80-90% of the total variation? Make a scree plot and based on it comment on how many PCs appear to be adequate. Interpret the selected PCs.

Question 2: Consider the Toothpaste data. This dataset includes consumer ratings on toothpaste about whether it prevents cavities, provides shiny teeth, strengthens gums, freshens breath, prevents decay, and gives attractive teeth. We would like to do factor analysis of these data.

- (a) Make a scatterplot matrix and comment on whether factor analysis is likely to be helpful for these data.
- (b) Conduct factor analysis using both PC and ML methods and compare them. How many factors are necessary to have an adequate fit of the model? Make this decision based on both residual matrix and test (for the test, state the hypotheses, test statistic, and p-value).
- (c) For the ML method, carry out factor rotation and compare the results (including interpretation) with and without rotation.
- (d) For the model with rotation, write down the approximate correlation/covariance matrix as per this model in terms of the factor loadings and specific variance matrices. Write all model assumptions needed for the ML analysis.
- (e) Perform a principal components analysis of these data and compare your results (including proportion of variance explained and interpretation) with the factor analysis results. Which one explains a higher proportion of variance and why?

Question 3: Consider a dataset about students' knowledge status on the subject of Electrical DC Machines.

Variable Information

STG: The degree of study time for goal object materials

SCG: The degree of repetition number of user for goal object materials

STR: The degree of study time of user for related objects with goal object

LPR: The exam performance of user for related objects with goal object

PEG: The exam performance of user for goal objects

UNS: The knowledge level of user

- a) Make a scatterplot matrix of variables by using different colors and symbols for different knowledge levels of user and include histogram of each variable in the diagonal of the matrix. Comment on any observed patterns/relationships.
- b) Perform linear discriminant analysis to classify a new observation into one of these three levels of knowledge and fully state the classification rule. Do this in R in two ways as discussed in class and compare the results. Use the rule to classify a new user whose STG and PEG values are at 1st quartile while the other three values are at 3rd quartiles.
- c) For the aid of understanding, make a scatterplot of the first two discriminant scores by labeling different levels of knowledge with different symbols and colors. Using this plot, which discriminant is more useful and why. Find the plug-in (APER) and leave-one-out (AER) estimates of misclassification rates. Which error rate is higher and why?
- d) Develop a classification rule using logistic regression (combine High with Medium) and repeat part (b). Find APER and AER estimates of misclassification rates. Compare the two methods in terms of misclassification rates and classification of the new user.
- e) Perform hierarchical clustering with single, complete, and average linkage. Decide on the number of clusters using a cutpoint (there is not necessarily a single best #, however, ease of interpretation can be a consideration) and report the means of clusters. Compare the clusters by commenting on how they differ from each other, and how they compare with the levels of knowledge of user (as used in parts b and d).
- f) Perform cluster analysis using kmeans approach. Repeat the parts mentioned in (e). Additionally use the plot of within groups sum of squares to decide the number of clusters. Compare the results with the ones from part (e).
- g) Perform model-based cluster analysis. Repeat the parts mentioned in (f). Compare the results with the ones from parts (e) and (f); include within groups sum of squares in this comparison.

STAT 6348 Project 2

John Kenney

2024-05-05

Answers

Question 1

You need 3 principal components to explain 82% of the total variation. With 4 principal components you can explain 90% of the total variation.

From the scree plot and using the elbow method it appears that it is adequate to use 3 and possibly just 2 principal components to summarize the data.

The first principal component seems to mostly explain the variables of the Fire Weather Index (FWI) with the Weather data observations either having lower coefficients, Wind speed is not included, and a slight contrast with the Relative humidity and total daily Rain fall.

The second principal component seems to mostly explain the different measurements for the true weather conditions.

The third principal component seems to mostly describe the measurements for the weather data observations along with the initial spread index from the FWI components.

Question 2

Question 2 (a)

There seems to be variables that are correlated with each other and other patterns between variables so Factor Analysis may be a good choice here. For example prevents cavities has linear positive correlation with shiny teeth and a negative correlation with decay prevention. There are also various variables that have a parabolic looking relationship between themselves.

Question 2 (b)

H_0 : 1 Factor is sufficient vs H_A : 1 Factor is not sufficient

$\chi^2_{14} = 92.13$ Test Statistic

$p - value = 1.5e - 13$

We reject the Null Hypothesis and conclude that 1 Factor is not sufficient.

H_0 : 2 Factors are sufficient vs H_A : 2 Factors are not sufficient

$\chi^2_8 = 15.8$ Test Statistic

$p - value = 0.0453$

We reject the Null Hypothesis and conclude that 2 Factors is not sufficient.

H_0 : 3 Factors are sufficient vs H_A : 3 Factors are not sufficient

$\chi^2_3 = 1.86$ Test Statistic

$p - value = 0.602$

We do not reject the Null Hypothesis and conclude that 3 Factors is sufficient.

Based on the Hypothesis tests with by comparing the residual matrices it seems that a 3 factor model is the best for our data. The ML method also seems to have a better residual matrix compared to the PC method's residual matrix. The $m = 2$ factor model based on the ML method only explains 67.1% of variation in the data, while $m = 3$ factor model based on the ML method explains 74.2% of the variation in the data.

ML Method Interpretation:

The first factor has high loadings on healthiness of mouth with a contrast between cavity prevention and gum strength with decay prevention.

The second factor has high loadings on cosmetic mouth features such as shiny teeth, fresh breath, and attractive teeth.

The third factor shows a contrast between age of consumer and fresh breath rating.

PC Method Interpretation:

The first factor has high loadings on healthiness of mouth with a contrast between cavity prevention and gum strength with decay prevention.

The second factor has high loadings on cosmetic mouth features such as shiny teeth, fresh breath, and attractive teeth.

The third factor has high loadings on age of the consumer.

The first 2 factors for both the ML and PC method are very similar with the third factor only being slightly different.

Question 2 (c)

ML Method Varimax Rotation Interpretation:

The first factor has high loadings on healthiness of mouth with a contrast between cavity prevention and gum strength with decay prevention also age has some weight but loading is smaller than others.

The second factor has high loadings on cosmetic mouth features such as shiny teeth, fresh breath, and attractive teeth.

The third factor shows a contrast between fresh breath rating and shiny teeth contrasted against age of consumer and gum strength.

The results of the rotation is that it makes the loadings in the ML Method no rotation more clear with little change in overall interpretation. With the same difference in the third factor to the PC Methods third factor that was discussed in part b.

Question 2 (d)

FA Model is $X - \mu = LF + \epsilon$

The Model Assumptions of the ML Method is that:

$F \perp \epsilon$

$F \sim N(0, I)$

$\epsilon \sim N(0, \Psi)$ where Ψ is a diagonal Matrix

$\hat{R} = \hat{L}T T^\top \hat{L}^\top + \hat{\Psi} = \hat{L}^* \hat{L}^{*\top} + \hat{\Psi}$

where \hat{L}^* is the Factor Loadings Matrix times the Rotation Matrix T and $\hat{\Psi}$ is the Specific Variance Matrix

$$\begin{aligned} \hat{R} &= \hat{L}T T^\top \hat{L}^\top + \hat{\Psi} = \hat{L}^* \hat{L}^{*\top} + \hat{\Psi} \\ &= \begin{bmatrix} 0.981 & -.035 & -.060 \\ -.021 & 0.642 & 0.292 \\ 0.868 & -.051 & -.341 \\ -.030 & 0.641 & 0.557 \\ -.874 & -.102 & 0.065 \\ 0.040 & 0.997 & 0.005 \\ 0.324 & -.197 & -.471 \end{bmatrix} \times \begin{bmatrix} 0.981 & -.021 & 0.868 & -.030 & -.874 & 0.040 & 0.324 \\ -.035 & 0.642 & -.051 & 0.641 & -.102 & 0.997 & -.197 \\ -.060 & 0.292 & -.341 & 0.557 & 0.065 & 0.005 & -.471 \end{bmatrix} \end{aligned}$$

$$\begin{aligned}
& + \begin{bmatrix} 0.034 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.502 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.128 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.278 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.221 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.005 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.635 \end{bmatrix} \\
& = \begin{bmatrix} 1.000 & -.060 & 0.873 & -.085 & -.858 & 0.004 & 0.353 \\ -.060 & 1.000 & -.151 & 0.575 & -.028 & 0.640 & -.271 \\ 0.873 & -.151 & 1.000 & -.249 & -.776 & -.018 & 0.452 \\ -.085 & 0.575 & -.249 & 1.000 & -.003 & 0.640 & -.398 \\ -.858 & -.028 & -.776 & -.003 & 1.000 & -.136 & -.294 \\ 0.004 & 0.640 & -.018 & 0.640 & -.136 & 1.000 & -.185 \\ 0.353 & -.271 & 0.452 & -.398 & -.294 & -.185 & 1.000 \end{bmatrix}
\end{aligned}$$

Question 2 (e)

The Cumulative Proportion of variance explained by 3 principal components is 85.8% for the Toothpaste data. The Cumulative Proportion of variance explained by the 3 Factor ML Method with no rotation is 74.2% for the Toothpaste data. The Cumulative Proportion of variance explained by the 3 Factor ML Method with varimax rotation is 74.2% for the Toothpaste data. The Cumulative Proportion of variance explained by the 3 Factor PC Method is 85.8% for the Toothpaste data the same as PC, because proportion of variance is $\frac{\lambda_i}{p}$ when using R .

Since 3 PC components explains 85% of the variation in the data will use 3 components also the screeplot shows a good elbow at 3 components. The FA with PC method and PC analysis explains the most about the total variation of the data, because the PC method maximizes the variance captured for each component. Therefore, it will always explain more of the total variation of the data in comparison with the ML method which tries to explain the covariance structure instead of focusing on capturing the most variance in each component.

Principal Components Interpretation:

The first principal component seems to mostly explain the overall healthiness of mouth with a contrast between cavity prevention and gum strength against decay prevention also the other variable also are included but have a smaller weight attached.

The second principal component seems to mostly explain cosmetic mouth features such as shiny teeth, fresh breath, and attractive teeth with a slight contrast with tooth decay prevention rating.

The third principal component is mostly explains the age of the consumer.

The first 2 principal components seem to roughly coincide with the Factor Analysis ML method with both rotation and varimax rotation, and the third principal component seems to be more similar to the Factor Analysis PC Method's third Factor.

Overall it seems that the Factor Analysis methods and Principal components Analysis performed similarly, so depending if you care more for the residual matrix you may want to go with the Factor Analysis ML Method results but if you are concerned about the amount of variation explained by the model it would be best to choose the PC Analysis or FA PC method.

Question 3

Question 3 (a)

It seems that for variables STG, SCG, STR, and LPR the data is pretty random with no clear observable user knowledge level groupings. With the inclusion the PEG it seems that we get a pretty clear separation

of the different knowledge levels of the users. It can also be seen where PEG is the response that we get an increasing group structure for the users knowledge level.

Question 3 (b)

$$x_0 = \{0.24075, 0.4975, 0.69, 0.6475, 0.25\}$$

The first classification Rule is:

Allocate x_0 to π_k if the linear discriminant score $\hat{d}_k(x_0)$ = the largest of $\hat{d}_1(x_0), \dots, \hat{d}_g(x_0)$

$$\hat{d}_i(x_0) = \bar{x}_i^\top S_{\text{pooled}}^{-1} x_0 - \frac{1}{2} \bar{x}_i^\top S_{\text{pooled}}^{-1} \bar{x}_i + \ln(p_i) \text{ for } i = 1, 2, \dots, g$$

p_i is the prior for the populations if unknown assume equal

$$S_{\text{pooled}} = \frac{1}{n_1 + \dots + n_g - g} ((n_1 - 1)S_1 + \dots + (n_g - 1)S_g)$$

I assumed equal prior so dropped $\ln(p_i)$ term.

What I calculated was that $\hat{d}_{\text{Low}}(x_0) = 20.376$, $\hat{d}_{\text{Middle}}(x_0) = 18.162$, and $\hat{d}_{\text{High}}(x_0) = 5.866$. Therefore based on our classification Rule we would conclude that x_0 belongs to the Low level group for the user's knowledge level.

The second classification rule is:

Allocate x_0 to π_k if the Squared distance of \hat{y}_j to \bar{y}_{kj} is smallest.

$$\text{ie } \min_{k \in \{1, \dots, g\}} D_i(x_0) = \sum_{j=1}^r (\hat{y}_j - \bar{y}_{kj})^2 = \sum_{j=1}^r (\hat{a}_j^\top x_0 - \hat{a}_j^\top \bar{x}_k)^2$$

What I calculated was that $D_{\text{Low}}(x_0) = 1.913$, $D_{\text{Middle}}(x_0) = 6.341$, and $D_{\text{High}}(x_0) = 30.934$. Therefore based on our classification Rule we would conclude that x_0 belongs to the Low level group for the user's knowledge level.

The two classification Rules coincide.

Question 3 (c)

$$APER = \frac{2+1+6}{258} = \frac{9}{258} = 3.5\%$$

$$AER = \frac{2+1+6}{258} = \frac{9}{258} = 3.5\%$$

The error rates are the same.

Question 3 (d)

The classification rule for logistic regression is defined as:

Assign x_0 to population High-Middle (value 0) if $\hat{p}(x_0) < 0.5$ and Assign x_0 to population Low (value 1) if $\hat{p}(x_0) > 0.5$

Our calculated that $\hat{p}(x_0) = 0.82$. Therefore based on our classification Rule we would conclude that x_0 (new user) belongs to the Low level group for the user's knowledge level (UNS).

$$APER = \frac{1+3}{258} = \frac{4}{258} = 1.6\%$$

$$AER = \frac{3+3}{258} = \frac{6}{258} = 2.3\%$$

AER has a higher misclassification rate of 2.3% which is 0.7% higher than the APER using the plugin method. This makes sense because APER is trained on what it is predicting so the estimates are biased whereas with AER you are using LOOCV.

Both methods LDA and Logistic both assign the new user to the Low level of UNS. In comparison with the LDA methods AER and APER they have an error rate of 3.5% each so it would seem that logistic regression in this instance is better, but we also collapsed two groups into 1 group so we lose some specification.

Question 3 (e)

The complete linkage cluster gets a better cluster of 3 groups whereas single and average linkage both have one cluster only having 1 observation. Also with single linkage it almost clusters all the observations to one cluster. The average linkage is a little better than the single linkage. Overall the results of the hierarchical clustering into 3 groups does not seem to be the best method. In comparison with the level of Knowledge groups the closest is complete linkage with cluster 2 being closest to the Middle group. For average linkage Cluster 3 is sorta similar to UNS's High Mean vector. Compared to UNS groups with 2 groups the complete

linkage is still the best in comparison to the group sizes and mean vectors. Overall complete linkage is the clustering method with the most comparable group sizes the the UNS grouping.

Question 3 (f)

Based on the Within Sum of Squares plot for k values from 1 to 5 it seems to me that $k = 3$ is the best choice with a WSS of 52.71, but I could have possibly used $k = 2$ as well. The Kmeans $k = 3$ is much better than the hierarchical clustering method with similar group sizes and Mean vectors to the UNS Group (3). Therefore out of the 2 methods I would definitely use Kmeans in this scenario if the class labels were not available.

Question 3 (g)

Only one of the Mean vectors seem to me to be somewhat close to the mean vector of UNS groupings, but we have here 4 clusters so this is somewhat to be expected. The group size for each cluster is overall well proportioned in comparison with the hierarchical clustering methods. The Cluster 1 from the model based clustering is somewhat close to the Low vector in the UNC 3 group's mean vector. By comparing the different clustering methods based on Within Cluster Sum of Squares we see that Model-Based Clustering with 4 groups has the smallest WSS equal to 42.057. Therefore, the Model-Based clustering method may be a better choice for further analysis of the different clusters based on the data.

Outputs

Question 1

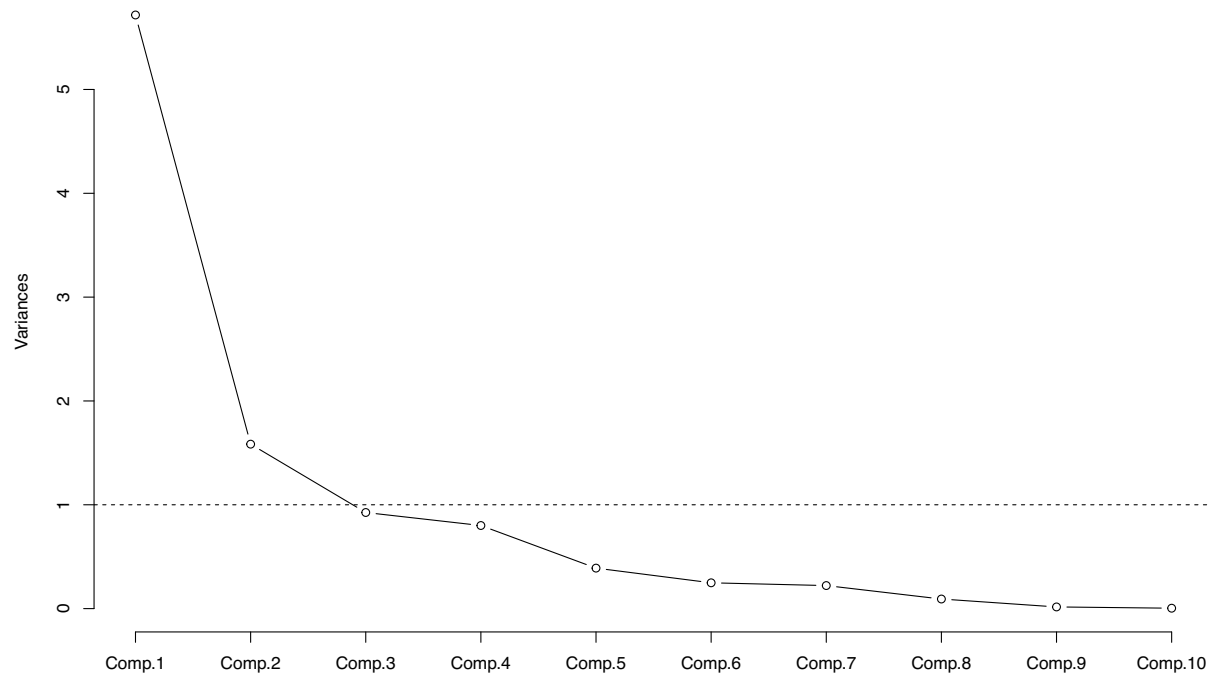


Figure 1: ScreePlot for Principal Components Analysis of Algerian Forest Fire Data

Question 2

Question 2 (a)

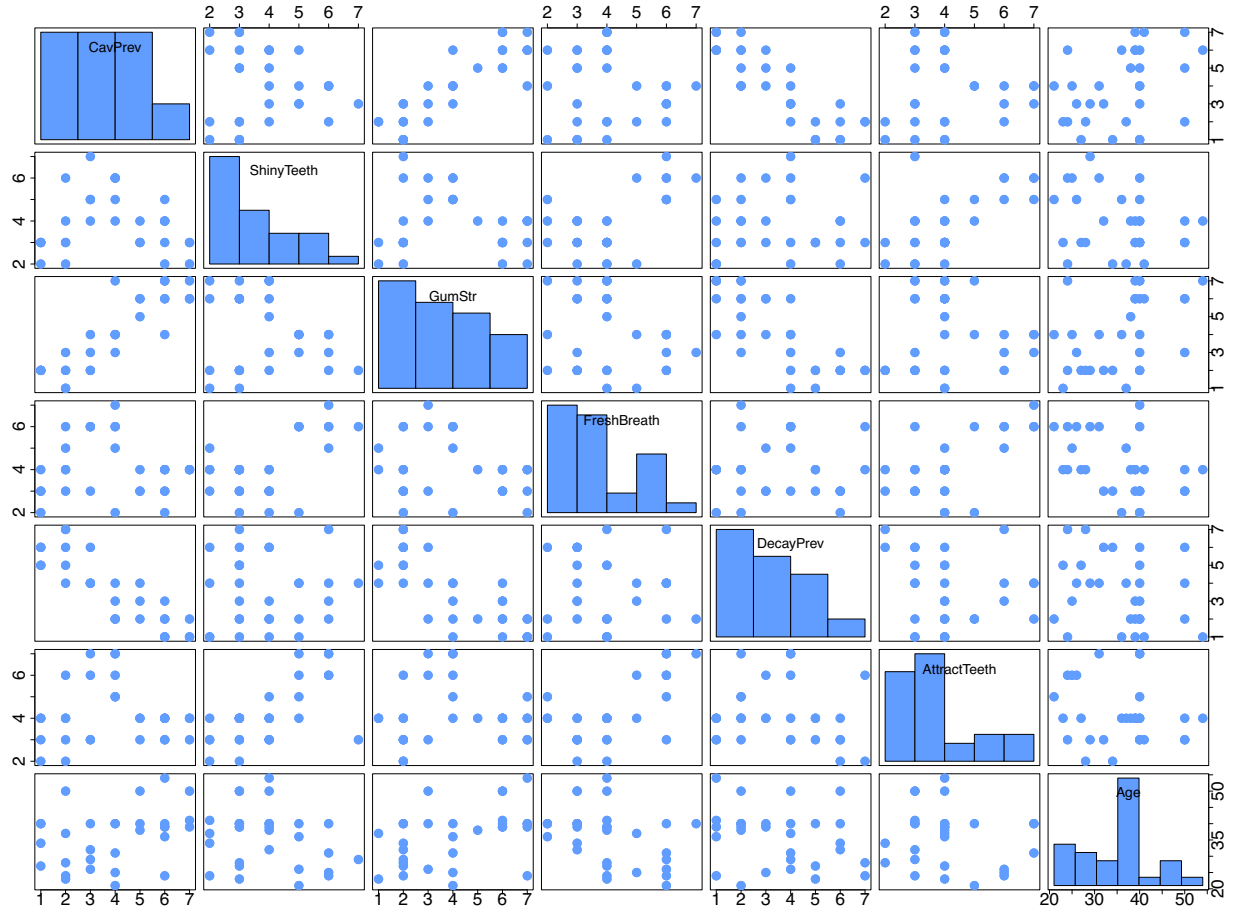


Figure 2: ScatterPlot Matrix for ToothPaste Data

Question 2 (b)

Table 1: Factor Loadings for 3 Factors

	ML Method			PC Method		
	Factor1	Factor2	Factor3	Factor1	Factor2	Factor3
CavPrev	0.981			0.859	-0.409	0.152
ShinyTeeth		0.644	0.274	-0.407	-0.726	-0.246
GumStr	0.906		-0.225	0.901	-0.284	
FreshBreath	-0.127	0.647	0.536	-0.477	-0.726	
DecayPrev	-0.871	-0.137		-0.774	0.500	-0.213
AttractTeeth		0.997		-0.292	-0.819	-0.246
Age	0.389	-0.19	-0.422	0.626	0.182	-0.727

Table 2: Residual Matrices

	CavPrev	ShinyTeeth	GumStr	FreshBreath	DecayPrev	AttractTeeth	Age
Residual Matrix ML Method with 1 Factor							
CavPrev	0.000	0.021	-0.002	0.026	-0.005	-0.007	-0.015
ShinyTeeth	0.021	0.000	-0.086	0.563	-0.047	0.641	-0.230
GumStr	-0.002	-0.086	0.000	-0.144	0.010	-0.029	0.106
FreshBreath	0.026	0.563	-0.144	0.000	-0.107	0.642	-0.359
DecayPrev	-0.005	-0.047	0.010	-0.107	0.000	-0.126	0.076
AttractTeeth	-0.007	0.641	-0.029	0.642	-0.126	0.000	-0.190
Age	-0.015	-0.230	0.106	-0.359	0.076	-0.190	0.000
Residual Matrix PC Method with 1 Factor							
CavPrev	0.000	0.297	0.099	0.324	-0.192	0.255	-0.181
ShinyTeeth	0.297	0.000	0.212	0.378	-0.296	0.522	-0.004
GumStr	0.099	0.212	0.000	0.182	-0.080	0.245	-0.114
FreshBreath	0.324	0.378	0.182	0.000	-0.376	0.501	-0.105
DecayPrev	-0.192	-0.296	-0.080	-0.376	0.000	-0.362	0.226
AttractTeeth	0.255	0.522	0.245	0.501	-0.362	0.000	-0.003
Age	-0.181	-0.004	-0.114	-0.105	0.226	-0.003	0.000
Residual Matrix ML Method with 2 Factors							
CavPrev	0.000	0.010	-0.001	0.016	-0.005	-0.031	-0.012
ShinyTeeth	0.010	0.000	0.005	-0.035	0.045	0.066	0.031
GumStr	-0.001	0.005	0.000	-0.038	0.008	0.058	0.053
FreshBreath	0.016	-0.035	-0.038	0.000	-0.006	-0.006	-0.064
DecayPrev	-0.005	0.045	0.008	-0.006	0.000	-0.013	0.038
AttractTeeth	-0.031	0.066	0.058	-0.006	-0.013	0.000	0.088
Age	-0.012	0.031	0.053	-0.064	0.038	0.088	0.000
Residual Matrix PC Method with 2 Factors							
CavPrev	0.000	-0.001	-0.018	0.027	0.012	-0.080	-0.106
ShinyTeeth	-0.001	0.000	0.006	-0.149	0.067	-0.073	0.128
GumStr	-0.018	0.006	0.000	-0.024	0.062	0.012	-0.062
FreshBreath	0.027	-0.149	-0.024	0.000	-0.014	-0.093	0.028
DecayPrev	0.012	0.067	0.062	-0.014	0.000	0.047	0.135
AttractTeeth	-0.080	-0.073	0.012	-0.093	0.047	0.000	0.147
Age	-0.106	0.128	-0.062	0.028	0.135	0.147	0.000
Residual Matrix ML Method with 3 Factors							
CavPrev	0.000	0.007	0.000	-0.001	0.000	0.000	0.005
ShinyTeeth	0.007	0.000	-0.004	-0.002	0.048	0.000	0.012
GumStr	0.000	-0.004	0.000	0.001	-0.002	0.000	-0.001
FreshBreath	-0.001	-0.002	0.001	0.000	-0.004	0.000	-0.005
DecayPrev	0.000	0.048	-0.002	-0.004	0.000	0.000	0.034
AttractTeeth	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Age	0.005	0.012	-0.001	-0.005	0.034	0.000	0.000
Residual Matrix PC Method with 3 Factors							
CavPrev	0.000	0.037	-0.023	0.013	0.045	-0.043	0.004
ShinyTeeth	0.037	0.000	0.015	-0.126	0.015	-0.133	-0.051
GumStr	-0.023	0.015	0.000	-0.027	0.070	0.021	-0.036
FreshBreath	0.013	-0.126	-0.027	0.000	0.006	-0.070	0.096
DecayPrev	0.045	0.015	0.070	0.006	0.000	-0.006	-0.020
AttractTeeth	-0.043	-0.133	0.021	-0.070	-0.006	0.000	-0.032
Age	0.004	-0.051	-0.036	0.096	-0.020	-0.032	0.000

Question 2 (c)

Table 3: Factor Loadings for 3 Factors

	ML Method Varimax Rotation			ML Method			PC Method		
	Factor1	Factor2	Factor3	Factor1	Factor2	Factor3	Factor1	Factor2	Factor3
CavPrev	0.981			0.981			0.859	-0.409	0.152
ShinyTeeth		0.642	0.292		0.644	0.274	-0.407	-0.726	-0.246
GumStr	0.868		-0.341	0.906		-0.225	0.901	-0.284	
FreshBreath		0.641	0.557	-0.127	0.647	0.536	-0.477	-0.726	
DecayPrev	-0.874	-0.102		-0.871	-0.137		-0.774	0.500	-0.213
AttractTeeth		0.997			0.997		-0.292	-0.819	-0.246
Age	0.324	-0.197	-0.471	0.389	-0.19	-0.422	0.626	0.182	-0.727

Question 2 (d)

Table 4: ML Method 3 factors Varimax Rotation

	CavPrev	ShinyTeeth	GumStr	FreshBreath	DecayPrev	AttractTeeth	Age
Estimated Correlation Matrix							
CavPrev	1.000	-0.060	0.873	-0.085	-0.858	0.004	0.353
ShinyTeeth	-0.060	1.000	-0.151	0.575	-0.028	0.640	-0.271
GumStr	0.873	-0.151	1.000	-0.249	-0.776	-0.018	0.452
FreshBreath	-0.085	0.575	-0.249	1.000	-0.003	0.640	-0.398
DecayPrev	-0.858	-0.028	-0.776	-0.003	1.000	-0.136	-0.294
AttractTeeth	0.004	0.640	-0.018	0.640	-0.136	1.000	-0.185
Age	0.353	-0.271	0.452	-0.398	-0.294	-0.185	1.000
Residual Matrix							
CavPrev	0.000	0.007	0.000	-0.001	0.000	0.000	0.005
ShinyTeeth	0.007	0.000	-0.004	-0.002	0.048	0.000	0.012
GumStr	0.000	-0.004	0.000	0.001	-0.002	0.000	-0.001
FreshBreath	-0.001	-0.002	0.001	0.000	-0.004	0.000	-0.005
DecayPrev	0.000	0.048	-0.002	-0.004	0.000	0.000	0.034
AttractTeeth	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Age	0.005	0.012	-0.001	-0.005	0.034	0.000	0.000

Question 2 (e)

Table 5: Loadings Matrices for 3 Components/Factors

	FA ML Method Varimax Rotation			FA ML Method			FA PC Method			PC Analysis		
	Factor1	Factor2	Factor3	Factor1	Factor2	Factor3	Factor1	Factor2	Factor3	Comp.1	Comp.2	Comp.3
CavPrev	0.981			0.981			0.859	-0.409	0.152	0.494	0.273	0.178
ShinyTeeth		0.642	0.292		0.644	0.274	-0.407	-0.726	-0.246	-0.234	0.483	-0.288
GumStr	0.868		-0.341	0.906		-0.225	0.901	-0.284		0.518	0.189	
FreshBreath		0.641	0.557	-0.127	0.647	0.536	-0.477	-0.726		-0.275	0.483	0.11
DecayPrev	-0.874	-0.102		-0.871	-0.137		-0.774	0.500	-0.213	-0.446	-0.333	-0.25
AttractTeeth		0.997			0.997		-0.292	-0.819	-0.246	-0.168	0.545	-0.288
Age	0.324	-0.197	-0.471	0.389	-0.19	-0.422	0.626	0.182	-0.727	0.360	-0.121	-0.852

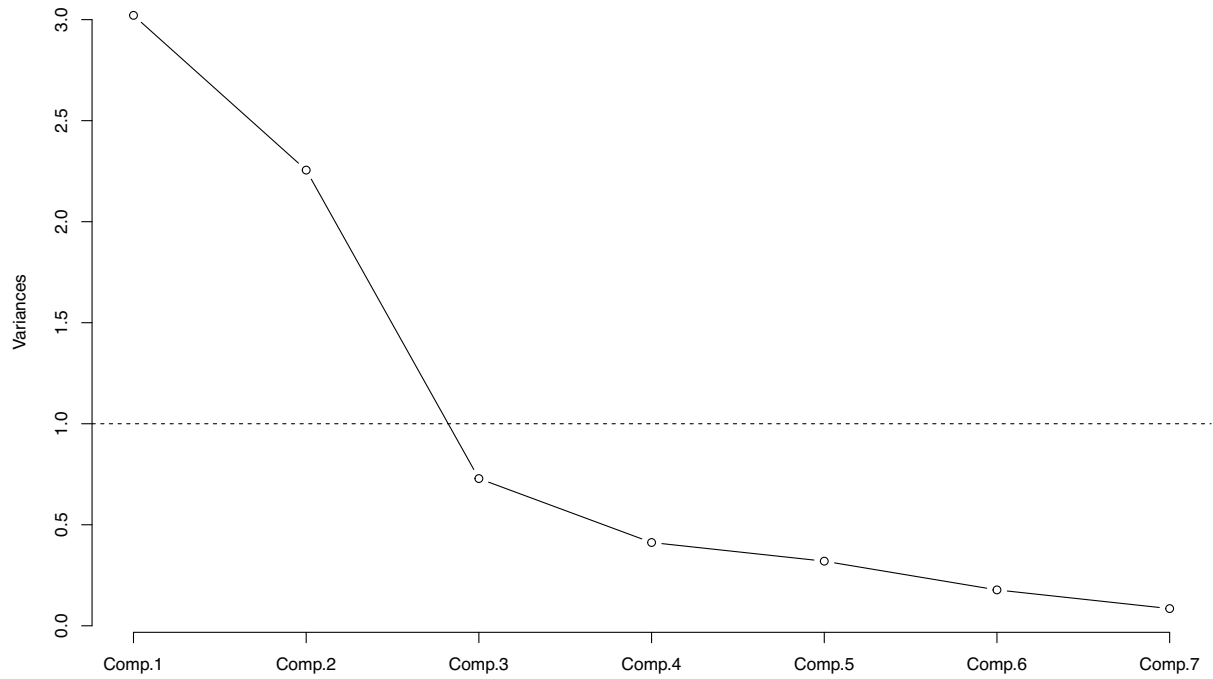


Figure 3: ScreePlot for Principal Component for ToothPaste Data

Question 3

Question 3 (a)

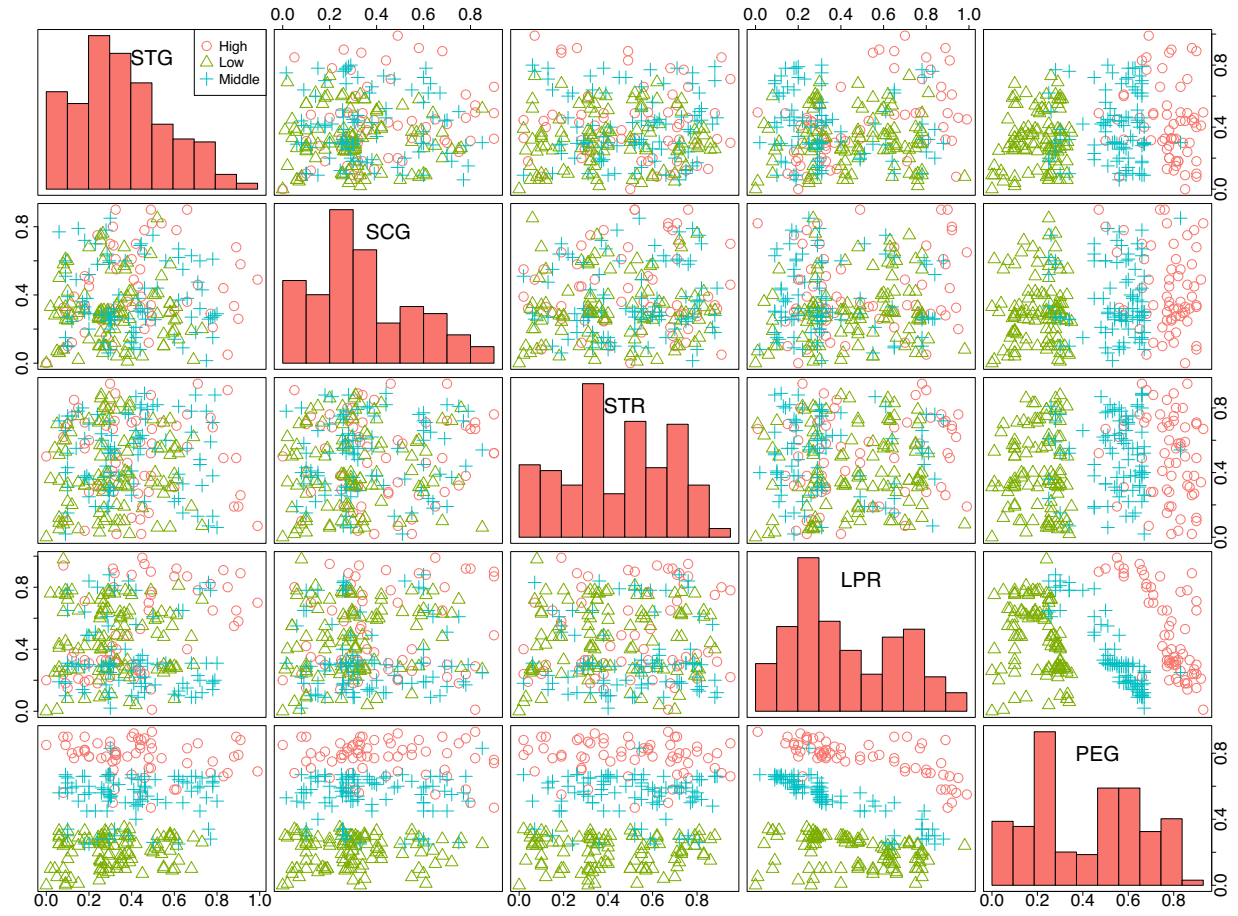


Figure 4: ScatterPlot Matrix given the knowledge level of the user

Question 3 (b)

Table 6: Classification Rule Calculation

	Low	Middle	High	Classification by Rule
Linear Discriminant Score	20.376	18.162	5.866	Low
Square Distance	1.913	6.341	30.934	Low

Question 3 (c)

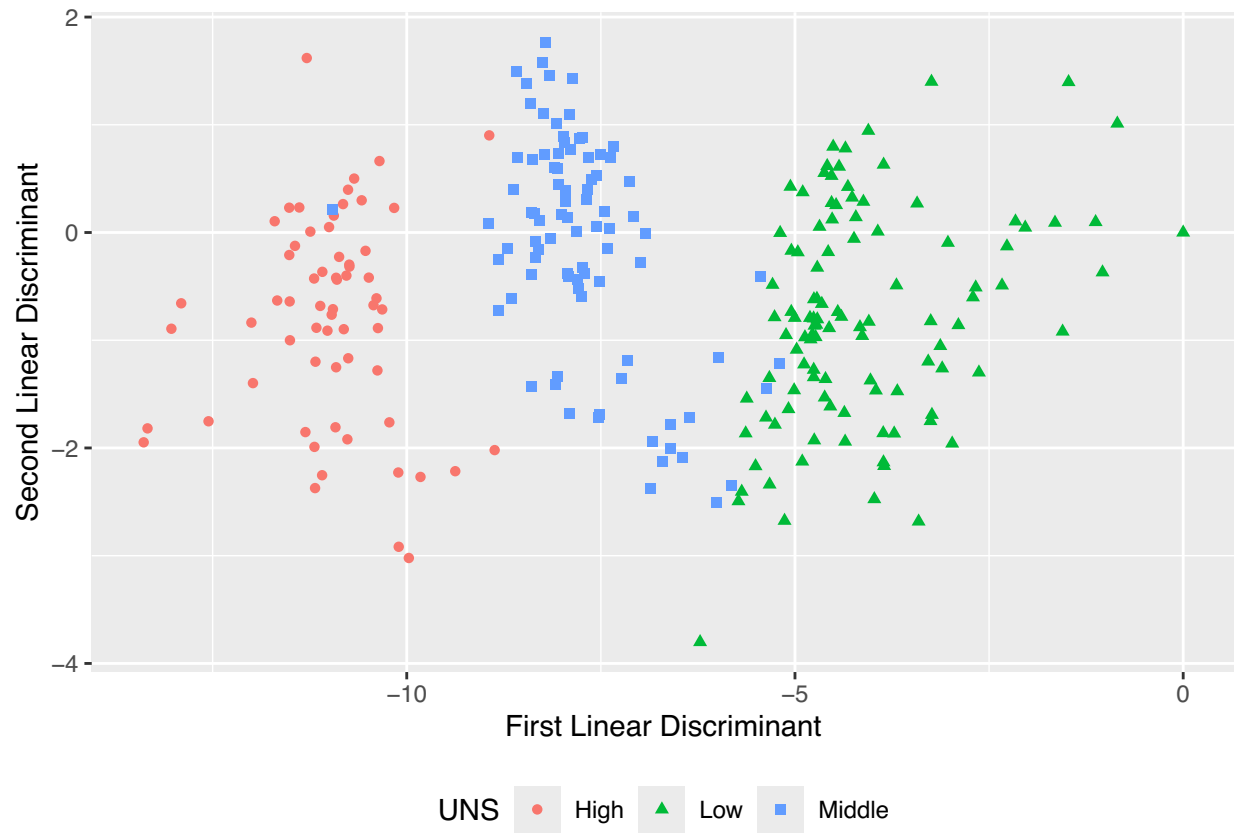


Figure 5: Discriminant ScorePlot

Table 7: Confusion Matrices

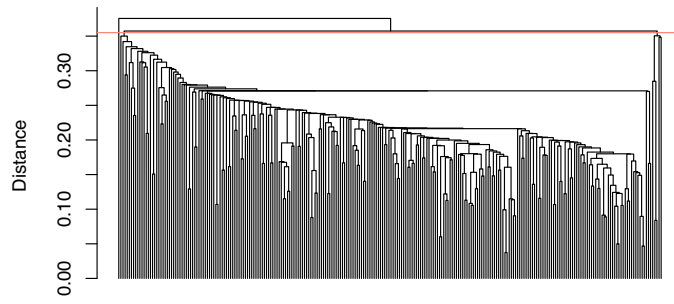
	Plugin Method			Cross-Validation Method		
	Pred.High	Pred.Low	Pred.Middle	Pred.High	Pred.Low	Pred.Middle
True.High	61	0	2	61	0	2
True.Low	0	107	0	0	107	0
True.Middle	1	6	81	1	6	81

Question 3 (d)

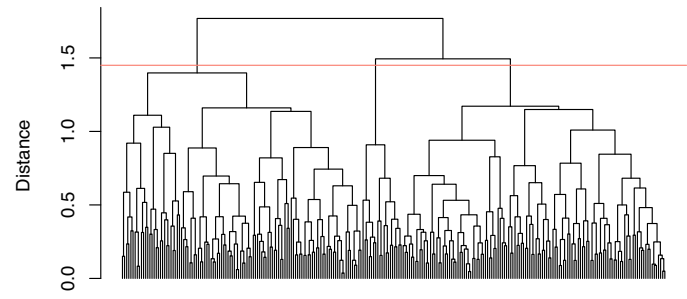
Table 8: Confusion Matrices

	Plugin Method		Cross-Validation Method	
	Pred.High-Middle	Pred.Low	Pred.High-Middle	Pred.Low
True.High-Middle	148	3	148	3
True.Low	1	106	3	104

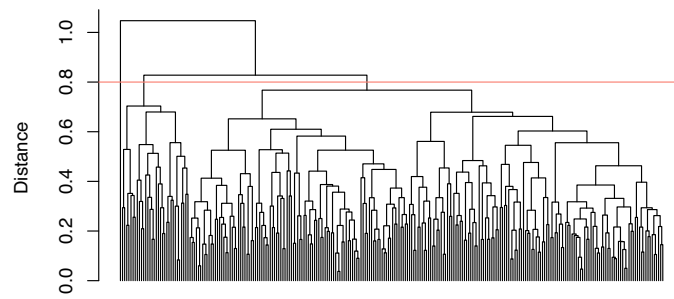
Question 3 (e)



(a) Single Linkage cut at $h = 0.355$



(b) Complete Linkage cut at $h = 1.45$



(c) Average Linkage cut at $h = 0.8$

Figure 6: Hierarchical Clustering Dendograms

Table 9: Mean Vectors of Hierarchical Clustering with 3 clusters chosen for ease of interpretation

	mean.STG	mean.SCG	mean.STR	mean.LPR	mean.PEG
Single Linkage					
Cluster 1 (Cluster size = 252)	0.360	0.351	0.474	0.429	0.452
Cluster 2 (Cluster size = 5)	0.914	0.469	0.240	0.674	0.808
Cluster 3 (Cluster size = 1)	0.520	0.850	0.060	0.270	0.250
Complete Linkage					
Cluster 1 (Cluster size = 19)	0.182	0.158	0.333	0.279	0.160
Cluster 2 (Cluster size = 125)	0.382	0.360	0.538	0.257	0.606
Cluster 3 (Cluster size = 114)	0.390	0.383	0.414	0.651	0.347
Average Linkage					
Cluster 1 (Cluster size = 1)	0.000	0.000	0.000	0.000	0.000
Cluster 2 (Cluster size = 225)	0.343	0.348	0.461	0.388	0.428
Cluster 3 (Cluster size = 32)	0.581	0.424	0.533	0.758	0.689
UNS Groups (3)					
Low (Group size = 107)	0.318	0.306	0.416	0.466	0.205
Middle (Group size = 88)	0.400	0.368	0.507	0.343	0.542
High (Group size = 63)	0.422	0.423	0.502	0.501	0.773
UNS Groups (2)					
Low (Group size = 107)	0.318	0.306	0.416	0.466	0.205
Middle-High (Group size = 151)	0.409	0.391	0.505	0.409	0.638

Question 3 (f)

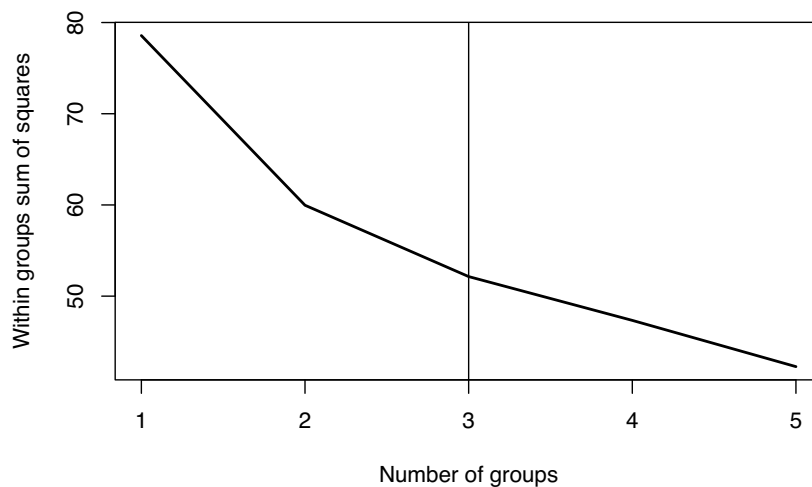


Figure 7: Within Sum of Squares for K-Means of $k = 1$ to 5

Table 10: Mean Vectors of K Means clustering with k = 3

	mean.STG	mean.SCG	mean.STR	mean.LPR	mean.PEG
K Means k = 3					
Cluster 1 (Cluster size = 110)	0.326	0.305	0.400	0.560	0.222
Cluster 2 (Cluster size = 65)	0.398	0.361	0.274	0.284	0.669
Cluster 3 (Cluster size = 83)	0.409	0.418	0.710	0.380	0.607
UNS Groups (3)					
Low (Group size = 107)	0.318	0.306	0.416	0.466	0.205
Middle (Group size = 88)	0.400	0.368	0.507	0.343	0.542
High (Group size = 63)	0.422	0.423	0.502	0.501	0.773
UNS Groups (2)					
Low (Group size = 107)1	0.318	0.306	0.416	0.466	0.205
Middle-High (Group size = 151)	0.409	0.391	0.505	0.409	0.638

Question 3 (g)

Table 11: Mean Vectors of Model-based clustering with k = 4 chosen by bic

	mean.STG	mean.SCG	mean.STR	mean.LPR	mean.PEG
Cluster 1 (Cluster size = 54)	0.313	0.273	0.393	0.282	0.245
Cluster 2 (Cluster size = 28)	0.593	0.441	0.526	0.793	0.686
Cluster 3 (Cluster size = 104)	0.369	0.382	0.504	0.252	0.679
Cluster 4 (Cluster size = 72)	0.333	0.348	0.452	0.673	0.216

Table 12: Within Cluster Sum of Squares for the different Clustering Methods

Hclust (Single Linkage)	Hclust (Complete Linkage)	Hclust (Average Linkage)	K-Means (k = 3)	Model-Based Clustering (k = 4)
67.751	52.574	62.484	47.91	42.057

Code

Question 1

```
alg <-  
  ↪ read.csv("https://utdallas.box.com/shared/static/e88slaf50oqp64svln7991j02ctidme9.csv")  
  
algff <- alg[,-c(1,2,3,15)]  
#head(algff)  
  
algff.pc <- princomp(algff[,-11],cor=TRUE) # used CORRELATION matrix  
#summary(algff.pc,loadings=TRUE) #loadings=TRUE reports the eigenvectors  
#names(algff.pc) # objects saved in usair.pc  
#algff.pc$loadings # eigenvectors of the correlation matrix  
#algff.pc$scores # values of all principal components  
  
#,fig.show='hide'}  
#Screeplot  
screeplot(algff.pc, type="lines",main = "")  
abline(h=1,lty=2)  
  
#algff.pc$loadings
```

Question 2

Question 2 (a)

```
toothpaste <-  
  ↪ read.csv("https://utdallas.box.com/shared/static/b84oi6a2oq625rgwr5huq167otl00v1r.csv")  
tooth <- toothpaste[,2:8]  
colnames(tooth) <- c("CavPrev", "ShinyTeeth", "GumStr", "FreshBreath", "DecayPrev",  
  ↪ "AttractTeeth", "Age")  
  
library(car)  
library(scales)  
hex <- hue_pal()(3)  
  
scatterplotMatrix(~CavPrev + ShinyTeeth + GumStr + FreshBreath + DecayPrev + AttractTeeth  
  ↪ + Age , diagonal = list(method = "histogram"),smooth = FALSE,col = hex[3],data =  
  ↪ tooth,regLine = FALSE,by.groups = FALSE,cex = 3,cex.labels = 2,cex.axis = 2.5,legend  
  ↪ = list(coords = c(-5,10),cex = 2),pch = 16)
```

Question 2 (b)

```
factanalpc <- function(data, factor = 2) {  
  R <- cor(data)  
  L <- matrix(unlist(lapply(1:length(colnames(data)),function(var)  
  ↪ sqrt(eigen(cor(data))$values[var])*eigen(cor(data))$vectors[,var] )),ncol =  
  ↪ length(colnames(data)),byrow = F)[,1:factor]
```

```

if(factor > 1){
  psi <- diag(R) - rowSums((L)^2)
}
else {
  psi <- diag(R) - ((L)^2)
}
resid <- R - (L %*% t(L) + diag(psi))
L <- as.data.frame(L)
rownames(L) <- colnames(data)
colnames(L) <- paste("Factor",1:factor,sep = "")
return(list(Loadings = L,uniqueness = psi,resid = round(resid,digits= 3)))
}

tooth.fa.ml<-lapply(1:3,function(nf) factanal(tooth,factors=nf,rotation="none"))
residual.fa.ml<-lapply(1:3,function(nf) round(cor(tooth) -
↪ (tooth.fa.ml[[nf]]$loadings%*%t(tooth.fa.ml[[nf]]$loadings) +
↪ diag(tooth.fa.ml[[nf]]$uniquenesses)) ,digits = 3))

tooth.fa.pc<-lapply(1:3,function(nf) factanalpc(tooth,factor=nf))

#for (i in 1:3) {
#print(as.numeric(tooth.fa.ml[[i]]$PVAL))
#}

#tooth.fa.ml[[3]]$loadings
#tooth.fa.ml[[3]]$uniqueness #Specific (unique) variance of each variable
#tooth.fa.pc[[3]]$Loadings

tmp <- round(as.data.frame(as.matrix(tooth.fa.ml[[3]]$loadings)[1:7,1:3]),3)
#apply(tmp, 2, function(col) col[which(abs(col) < 0.1)] = "")
tmp[which(abs(as.data.frame(as.matrix(tooth.fa.ml[[3]]$loadings)[1:7,1:3])[,1]) < 0.1),1]
↪ = ""
tmp[which(abs(as.data.frame(as.matrix(tooth.fa.ml[[3]]$loadings)[1:7,1:3])[,2]) < 0.1),2]
↪ = ""
tmp[which(abs(as.data.frame(as.matrix(tooth.fa.ml[[3]]$loadings)[1:7,1:3])[,3]) < 0.1),3]
↪ = ""
#tmp

tmp2 <- round(tooth.fa.pc[[3]]$Loadings,3)
#apply(tmp, 2, function(col) col[which(abs(col) < 0.1)] = "")
tmp2[which(abs(tmp2[,1]) < 0.1),1] = ""
tmp2[which(abs(tmp2[,2]) < 0.1),2] = ""
tmp2[which(abs(tmp2[,3]) < 0.1),3] = ""
#tmp2

library(kableExtra)
knitr::kable(cbind(tmp,tmp2),booktabs = T,caption = "Factor Loadings for 3 Factors",
↪ align = "ccccc") %>%
  kable_styling(latex_options = c("hold_position")) %>%

```

```
add_header_above(c(" " = 1,"ML Method" = 3, "PC Method" = 3))
```

```
library(kableExtra)
knitr::kable(rbind(residual.fa.ml[[1]], tooth.fa.pc[[1]]$resid, residual.fa.ml[[2]],
  tooth.fa.pc[[2]]$resid, residual.fa.ml[[3]], tooth.fa.pc[[3]]$resid),
  booktabs = T,caption = "Residual Matrices", align = "cccccc") %>%
  pack_rows("Residual Matrix ML Method with 1 Factor", start_row = 1, end_row = 7) %>%
  pack_rows("Residual Matrix PC Method with 1 Factor", start_row = 8, end_row = 14) %>%
  pack_rows("Residual Matrix ML Method with 2 Factors", start_row = 15, end_row = 21) %>%
  pack_rows("Residual Matrix PC Method with 2 Factors", start_row = 22, end_row = 28) %>%
  pack_rows("Residual Matrix ML Method with 3 Factors", start_row = 29, end_row = 35) %>%
  pack_rows("Residual Matrix PC Method with 3 Factors", start_row = 36, end_row = 42) %>%
  kable_styling(latex_options = c("hold_position",font_size = 2))
```

Question 2 (c)

```
tooth.ml.rot <- factanal(tooth,factors=3,rotation = "varimax")
#tooth.ml.rot

tmpprot <- round(as.data.frame(as.matrix(tooth.ml.rot$loadings)[1:7,1:3]),3)
tmpprot[which(abs(tmpprot[,1]) < 0.1),1] = ""
tmpprot[which(abs(tmpprot[,2]) < 0.1),2] = ""
tmpprot[which(abs(tmpprot[,3]) < 0.1),3] = ""
#tmpprot
library(kableExtra)
knitr::kable(cbind(tmpprot,tmp,tmp2),booktabs = T,caption = "Factor Loadings for 3
↪ Factors", align = "cccccc") %>%
  kable_styling(latex_options = c("hold_position")) %>%
  add_header_above(c(" " = 1,"ML Method Varimax Rotation" = 3,"ML Method" = 3, "PC
↪ Method" = 3))
```

Question 2 (d)

```
est.cor <- round((tooth.ml.rot$loadings%*%t(tooth.ml.rot$loadings) +
↪ diag(tooth.ml.rot$uniquenesses)) ,digits = 3)
residual <- round(cor(tooth) - (tooth.ml.rot$loadings%*%t(tooth.ml.rot$loadings) +
↪ diag(tooth.ml.rot$uniquenesses)) ,digits = 3)
library(kableExtra)
knitr::kable(rbind(est.cor,residual),booktabs = T,caption = "ML Method 3 factors Varimax
↪ Rotation", align = "cccccc") %>%
  pack_rows("Estimated Correlation Matrix", start_row = 1, end_row = 7) %>%
  pack_rows("Residual Matrix", start_row = 8, end_row = 14) %>%
  kable_styling(latex_options = c("hold_position",font_size = 2))
```

Question 2 (e)

```
tooth.pc <- princomp(tooth,cor=TRUE) # used CORRELATION matrix
#summary(tooth.pc,loadings=TRUE) #loadings=TRUE reports the eigenvectors
```

```

tmppc <- round(as.data.frame(as.matrix(tooth.pc$loadings)[1:7,1:3]),3)
tmppc[which(abs(tmppc[,1]) < 0.1),1] = ""
tmppc[which(abs(tmppc[,2]) < 0.1),2] = ""
tmppc[which(abs(tmppc[,3]) < 0.1),3] = ""
#tmppc

library(kableExtra)
knitr::kable(cbind(tmprot,tmp,tmp2,tmppc),format = "latex",booktabs = T,caption =
  ↳ "Loadings Matrices for 3 Components/Factors", align = "cccccccc") %>%
  kable_styling(latex_options = c("hold_position","scale_down")) %>%
  add_header_above(c(" " = 1,"FA ML Method Varimax Rotation" = 3,"FA ML Method" = 3, "FA
  ↳ PC Method" = 3,"PC Analysis" = 3))

```

```

screepplot(tooth.pc, type="lines",main = "")
abline(h=1,lty=2)

```

Question 3

Question 3 (a)

```

usermodeling <-
  ↳ read.csv("https://utdallas.box.com/shared/static/eyczzglabff2a37tpez8h2bwcmrggyek.csv")
usermodeling$UNS <- ifelse(usermodeling$UNS == "very_low","Low",usermodeling$UNS)
#head(usermodeling)

```

```

library(car)
library(scales)
hex <- hue_pal()(4)

```

```

scatterplotMatrix(~STG + SCG + STR + LPR + PEG| UNS, diagonal = list(method =
  ↳ "histogram"),smooth = FALSE,col = hex,data = usermodeling,regLine = FALSE,by.groups =
  ↳ FALSE,cex = 3,cex.labels = 3,cex.axis = 2.5,legend = list(cex = 2),pch = 1:4)

```

Question 3 (b)

```

#summary(factor(usermodeling$UNS))
library(MASS)
dis<-lda(UNS~STG + SCG + STR + LPR + PEG,data=usermodeling,prior = c(1/3,1/3,1/3))
#dis
#names(dis)
#dis$scaling #coefficients are saved here

# prediction on new observation
newdata <- rbind(c(quantile(usermodeling$STG,probs = .25)[[1]],
  quantile(usermodeling$SCG,probs = .75)[[1]],
  quantile(usermodeling$STR,probs = .75)[[1]],
  quantile(usermodeling$LPR,probs = .75)[[1]],

```

```

        quantile(usermodeling$PEG, probs = .25)[[1]])
colnames(newdata) <- colnames(usermodeling[, -6])
newdata <- as.matrix(c(newdata), ncols = 5)

# discriminants
a1 <- dis$scaling[,1]
a2 <- dis$scaling[,2]

# group means
m1 = dis$means[1,] # high
m2 = dis$means[2,] # low
m3 = dis$means[3,] # middle

# group Sample variances
S1 <- var(usermodeling[usermodeling$UNS == "High", -6])
n1 <- length(usermodeling[usermodeling$UNS == "High", 6])

S2 <- var(usermodeling[usermodeling$UNS == "Low", -6])
n2 <- length(usermodeling[usermodeling$UNS == "Low", 6])

S3 <- var(usermodeling[usermodeling$UNS == "Middle", -6])
n3 <- length(usermodeling[usermodeling$UNS == "Middle", 6])

# Sample variance pooled
g <- 3
S <- 1/(n1+n2+n3 - g) * ( (n1 - 1)*S1 + (n2 - 1)*S2 + (n3 - 1)*S3 )
Sinv <- solve(S)

# linear discriminant score
d1 <- t(m1) %*% Sinv %*% newdata - 0.5 * t(m1) %*% Sinv %*% m1
d2 <- t(m2) %*% Sinv %*% newdata - 0.5 * t(m2) %*% Sinv %*% m2
d3 <- t(m3) %*% Sinv %*% newdata - 0.5 * t(m3) %*% Sinv %*% m3

d <- round(c(d2,d3,d1),3)
#which(d == max(d))
d <- c(d,"Low")

# distance to group 1
D1 <- sum((a1%*%m1 - a1%*% newdata)^2, (a2%*%m1 - a2%*%newdata)^2)

# distance to group 2
D2 <- sum((a1%*%m2 - a1%*% newdata)^2, (a2%*%m2 - a2%*%newdata)^2)

# distance to group 3
D3 <- sum((a1%*%m3 - a1%*% newdata)^2, (a2%*%m3 - a2%*%newdata)^2)
D <- round(c(D2,D3,D1),3)
#which(D == min(D))
D <- c(D,"Low")

dfclass <- as.data.frame(rbind(d,D))

```



```

rownames(dfclass) <- c("Linear Discriminant Score","Square Distance")
colnames(dfclass) <- c("Low","Middle","High","Classification by Rule")
#dfclass
library(kableExtra)
knitr::kable(dfclass,booktabs = T,caption = "Classification Rule Calculation", align =
  ↪ "cccc") %>%
  kable_styling(latex_options = c("hold_position"))

```

Question 3 (c)

```

library(ggplot2)
df <- data.frame(UNS = usermodeling$UNS,firstdis = as.matrix(usermodeling[,-6]) %*%
  ↪ matrix(a1,ncol= 1),seconddis = as.matrix(usermodeling[,-6]) %*% matrix(a2,ncol= 1))
ggplot(df,aes(x = firstdis, y = seconddis,col = UNS,pch = UNS)) + geom_point() +
  ↪ xlab("First Linear Discriminant") + ylab("Second Linear Discriminant") +
  ↪ theme(legend.position = "bottom")

# Confusion Matrix - to get "plug-in" estimate of misclassification rate
# Prediction of classes for observations in the sample
pred.group<-predict(dis,method="plug-in")$class
#cbind(BULLS$Breed, pred.group)
t1 <- table(paste("True",usermodeling$UNS,sep = "."), paste("Pred",pred.group,sep = "."))

# Leave-one-out estimate of misclassification rate: use CV = TRUE option
dis.cv<-lda(UNS~STG + SCG + STR + LPR + PEG,data=usermodeling,prior = c(1/3,1/3,1/3),CV =
  ↪ TRUE)
#names(dis.cv)
#dis.cv$class
#cbind(BULLS$Breed, dis.cv$class)
t2 <- table(paste("True",usermodeling$UNS,sep = "."), paste("Pred",dis.cv$class,sep =
  ↪ "."))

library(kableExtra)
knitr::kable(cbind(t1,t2),booktabs = T,caption = "Confusion Matrices", align = "cccccc")
  ↪ %>%
  kable_styling(latex_options = c("hold_position")) %>%
  add_header_above(c(" " = 1,"Plugin Method" = 3, "Cross-Validation Method" = 3))

```

Question 3 (d)

```

userlogistic <- usermodeling
userlogistic$UNS <- ifelse(userlogistic$UNS == "High","Middle",userlogistic$UNS)
userlogistic$UNS <- ifelse(userlogistic$UNS == "Middle","High-Middle",userlogistic$UNS)

fit1 <- glm(as.factor(UNS)~STG + SCG + STR + LPR + PEG, family=binomial,
  ↪ data=userlogistic)
#summary(fit1)

#Plug-in estimate

```

```

t1log <- table(userlogistic$UNS,(predict(fit1, type="response")>0.5)) # predict
  ↪ probability
colnames(t1log) <- c("Pred.High-Middle","Pred.Low")
rownames(t1log) <- c("True.High-Middle","True.Low")

newdata <- rbind(c(quantile(usermodeling$STG,probs = .25)[[1]],
                    quantile(usermodeling$SCG,probs = .75)[[1]],
                    quantile(usermodeling$STR,probs = .75)[[1]],
                    quantile(usermodeling$LPR,probs = .75)[[1]],
                    quantile(usermodeling$PEG,probs = .25)[[1]]))
colnames(newdata) <- colnames(usermodeling[,-6])
newdata <- as.data.frame(newdata)
#predict(fit1,newdata=newdata,type="response")

#Cross-Validation (Leave-one-out method)
newpred <- numeric(length(userlogistic$UNS))

for (i in 1:length(userlogistic$UNS))
{
  newdat <- userlogistic[-i,]
  newfit <- glm(as.factor(UNS)~STG + SCG + STR + LPR + PEG, family=binomial, data=newdat)
  newpred[i] <- predict(newfit, newdat= data.frame(userlogistic[i,-6]), type="response")
}

t2log <- table(userlogistic$UNS,(newpred>0.5))
colnames(t2log) <- c("Pred.High-Middle","Pred.Low")
rownames(t2log) <- c("True.High-Middle","True.Low")

library(kableExtra)
knitr::kable(cbind(t1log,t2log),booktabs = T,caption = "Confusion Matrices", align =
  ↪ "ccc") %>%
  kable_styling(latex_options = c("hold_position")) %>%
  add_header_above(c(" " = 1,"Plugin Method" = 2, "Cross-Validation Method" = 2))

```

Question 3 (e)

```

#,out.width="58%", fig.ncol = 2,fig.subcap=c('Single Linkage cut at h = 0.355', 'Complete
  ↪ Linkage cut at h = 1.45', 'Average Linkage cut at h = 0.8'),fig.cap="Hierarchical
  ↪ Clustering Dendograms",fig.align='center'}

#dendrogram from hierarchical clustering
plot(hclust(dist(usermodeling[,-6]),method="single"), ylab="Distance", xlab = "",main =
  ↪ "", labels = FALSE, hang = -1,sub = "")
#cutting the dendrogram at height 0.355 to get various clusters formed at that point.
abline(h=0.355,col = "salmon")

cluster.single <-cutree(hclust(dist(usermodeling[,-6]),method="single"),h=0.355)
#max(cluster.single)

```

```

# To get components of each cluster separately
nc<-1
#row.names(usermodeling)[cluster.single==nc]
single.clus<-lapply(1:max(cluster.single),function(nc)
  ↪ {row.names(usermodeling)[cluster.single==nc]})
#single.clus

# Mean vectors for each cluster
#colMeans(usermodeling[cluster.single==nc,])
cluster.single.mean<-lapply(1:max(cluster.single),function(nc)
  ↪ {colMeans(usermodeling[cluster.single==nc,-6])})
#cluster.single.mean

plot(hclust(dist(usermodeling[,-6]),method="complete"),ylab="Distance", xlab = "",main =
  ↪ "", labels = FALSE, hang = -1,sub = "")
#cutting the dendrogram at height 1.45 to get various clusters formed at that point.
abline(h=1.45,col = "salmon")

```

```

cluster.comp <-cutree(hclust(dist(usermodeling[,-6]),method="complete"),h=1.45)
#max(cluster.comp)

# To get components of each cluster separately
nc<-1
#row.names(usermodeling)[cluster.single==nc]
comp.clus<-lapply(1:max(cluster.comp),function(nc)
  ↪ {row.names(usermodeling)[cluster.comp==nc]})
#comp.clus

# Mean vectors for each cluster
#colMeans(usermodeling[cluster.single==nc,])
cluster.comp.mean<-lapply(1:max(cluster.comp),function(nc)
  ↪ {colMeans(usermodeling[cluster.comp==nc,-6])})
#cluster.comp.mean

plot(hclust(dist(usermodeling[,-6]),method="average"), ylab="Distance", xlab = "",main =
  ↪ "", labels = FALSE, hang = -1,sub = "")
#cutting the dendrogram at height 0.8 to get various clusters formed at that point.
abline(h=0.8,col = "salmon")

```

```

cluster.avg<-cutree(hclust(dist(usermodeling[,-6]),method="average"),h=0.8)
#max(cluster.avg)

# To get components of each cluster separately
nc<-1
#row.names(usermodeling)[cluster.single==nc]
avg.clus<-lapply(1:max(cluster.avg),function(nc)
  ↪ {row.names(usermodeling)[cluster.avg==nc]})
#avg.clus

# Mean vectors for each cluster

```

```

#colMeans(usermodeling[cluster.single==nc,])
cluster.avg.mean<-lapply(1:max(cluster.avg),function(nc)
  ↳ {colMeans(usermodeling[cluster.avg==nc,-6])})
#cluster.avg.mean

tabler <- round(rbind(matrix(unlist(cluster.single.mean),ncol = 5, byrow =
  ↳ T),matrix(unlist(cluster.comp.mean),ncol = 5, byrow =
  ↳ T),matrix(unlist(cluster.avg.mean),ncol = 5, byrow = T)),3)
colnames(tabler) <- paste("mean",colnames(usermodeling[,-6]),sep = ".")
rownames(tabler) <- paste(rep(paste("Cluster",1:3),3), " (Cluster size = ",
  c(summary(factor(cluster.single)),
    summary(factor(cluster.comp)),
    summary(factor(cluster.avg))),"),",
  sep = "")

# add UNS 3 groups
truetbl <- rbind(Low = m2, Middle = m3, High = m1)
rownames(truetbl) <- c("Low (Group size = 107)", "Middle (Group size = 88)", "High (Group
  ↳ size = 63)")
colnames(truetbl) <- paste("mean",colnames(usermodeling[,-6]),sep = ".")

# add UNS 2 groups
# summary(factor(userlogistic[,6]))
tmp <- lda(UNS~STG + SCG + STR + LPR + PEG,data=userlogistic,prior = c(1/2,1/2))
m1tmp <- tmp$means[1,]
m2tmp <- tmp$means[2,]
truetbl2 <- rbind(Low = m2tmp, High = m1tmp)
rownames(truetbl2) <- c("Low (Group size = 107)", "Middle-High (Group size = 151)")
colnames(truetbl2) <- paste("mean",colnames(usermodeling[,-6]),sep = ".")

# putting it all together
tabler <- rbind(tabler,truetbl,truetbl2)
library(kableExtra)
knitr::kable(tabler,digits = 3,booktabs = T, caption = "Mean Vectors of Hierarchical
  ↳ Clustering with 3 clusters chosen for ease of interpretation", align = "cccc") %>%
  kable_styling(latex_options = c("hold_position")) %>%
  pack_rows("Single Linkage", start_row = 1, end_row = 3) %>%
  pack_rows("Complete Linkage", start_row = 4, end_row = 6) %>%
  pack_rows("Average Linkage", start_row = 7, end_row = 9) %>%
  pack_rows("UNS Groups (3)", start_row = 10, end_row = 12) %>%
  pack_rows("UNS Groups (2)", start_row = 13, end_row = 14)

```

Question 3 (f)

```

#Finding min & max of each column (option 2) and doing max-min to get range
rge<-apply(usermodeling[,-6],2,max)-apply(usermodeling[,-6],2,min)

# Dividing entries of each column (option 2) by range
usermodel.std<-sweep(usermodeling[,-6],2,rge,FUN="/")

```

```

#K-means with 2 clusters
#kmeans(usermodel.std,2)
#sum(kmeans(usermodel.std,2)$withinss)

#K-means with 3 clusters
#kmeans(usermodel.std,3)
K <- 5
wss <- numeric(K)
#set seed of random number generator - for randomly choosing an initial cluster
set.seed(1234)

#within-group ss for two to 10 cluster solutions
for(i in 1:K) {
  W<-sum(kmeans(usermodel.std,i)$withinss)
  wss[i]<-W
}

#Plotting the wss vs number of clusters
plot(1:K,wss,type="l",xlab="Number of groups",ylab="Within groups sum of squares",lwd=2)
abline(v = 3)

```

```

# K-means output for K=6 clusters
usermodel.kmean<-kmeans(usermodel.std,3)
#usermodel.kmean

#Cluster means of original (unstandardized) data saved in pottery.data
km3.means <- lapply(1:3,function(nc)
  ↳ {apply(usermodeling[usermodel.kmean$cluster==nc,-6],2,mean)})
tablef <- round(as.data.frame(matrix(unlist(km3.means),ncol = 5, byrow = T)),3)
colnames(tablef) <- paste("mean",colnames(usermodeling[,-6]),sep = ".")
rownames(tablef) <- paste(paste("Cluster",1:3)," (Cluster size = ",
  ↳ usermodel.kmean$size,")",sep = "")

rownames(truettbl2) <- c("Low (Group size = 107)", "Middle-High (Group size = 151)")
# putting it all together
tablef <- rbind(tablef,truettbl,truettbl2)

library(kableExtra)
knitr::kable(tablef,digits = 3,booktabs = T,caption = "Mean Vectors of K Means clustering
  ↳ with k = 3", align = "cccc") %>%
  kable_styling(latex_options = c("hold_position")) %>%
  pack_rows("K Means k = 3", start_row = 1, end_row = 3) %>%
  pack_rows("UNS Groups (3)", start_row = 4, end_row = 6) %>%
  pack_rows("UNS Groups (2)", start_row = 7, end_row = 8)

```

Question 3 (g)

```

library(mclust)
#For a specified number of clusters
mb3 = Mclust(usermodeling[,-6], 3)
mb3$classification
mb3$modelName

mb <- Mclust(usermodeling[,-6])

# optimal number of cluster
mb$G

# optimal selected model
mb$modelName

# probability for an observation to be in a given cluster
#head(mb$z)

# get probabilities, means, variances
#summary(mb, parameters = TRUE)

#table(usermodeling$UNS, mb$classification)
# vs
#table(usermodeling$UNS, mb3$classification)

#plot(mb, what=c("classification"))

#plot(mb, "density")

#plot(mb, "BIC")

## Useful for comparing clustering methods
library("fpc")
#cs = cluster.stats(dist(usermodeling[,-6]), mb$classification)
#cs$within.cluster.ss #within cluster sum of squares
#cs[c("within.cluster.ss", "avg.silwidth")] #average silhouette width - ranges from 0 to
↪ 1; value closer to 1 suggests the data are better clustered.

tableg <- t(mb$parameters$mean)
colnames(tableg) <- paste("mean", colnames(usermodeling[,-6]), sep = ".")
rownames(tableg) <- paste(paste("Cluster", 1:4), " (Cluster size = ",
↪ summary(factor(mb$classification)), ")", sep = "")
#tableg

library(kableExtra)
knitr::kable(tableg, digits = 3, booktabs = T, caption = "Mean Vectors of Model-based
↪ clustering with k = 4 chosen by bic", align = "ccccc") %>%
  kable_styling(latex_options = c("hold_position"))

```

```

tablewss <- cbind(cluster.stats(dist(usermodeling[,-6]),
  ↳ cluster.single)$within.cluster.ss,
                  cluster.stats(dist(usermodeling[,-6]), cluster.comp)$within.cluster.ss,
                  ↳
                  cluster.stats(dist(usermodeling[,-6]), cluster.avg)$within.cluster.ss,
                  cluster.stats(dist(usermodeling[,-6]),
                  ↳ usermodel.kmean$cluster)$within.cluster.ss,
                  cluster.stats(dist(usermodeling[,-6]),
                  ↳ mb$classification)$within.cluster.ss )
#colnames(tablewss) <- c("Hclust (Single Linkage)", "Hclust (Complete Linkage)", "Hclust
  ↳ (Average Linkage)",
# "K-Means (k = 3)", "Model-Based Clustering (k = 4)")

library(kableExtra)
knitr::kable(tablewss, digits = 3, booktabs = T, caption = "Within Cluster Sum of Squares
  ↳ for the different Clustering Methods", align = "ccccc", col.names = c("Hclust (Single
  ↳ Linkage)", "Hclust (Complete Linkage)", "Hclust (Average Linkage)", "K-Means (k = 3)",
  ↳ "Model-Based Clustering (k = 4)")) %>%
  kable_styling(latex_options = c("hold_position")) %>%
  column_spec(2:6, width = "3cm")

```