DATA 630 – DATA MINING

SPRING 2017

Assignment 5-Classification Unsupervised

John Parsons

Professor Firdu Bati

University of Maryland University College

Introduction

The goal for Assignment five is to find patterns in the Abalone dataset that may lead to relevant correlations with the unsupervised Partitioning K-means or Hierarchical Clustering Algorithms in RStudio. The clustering algorithm is a goodness of fit measurement without external references. This algorithm will measure cluster cohesion (compactness) and cluster separation (isolation) which determines how well the clusters are separated from each other. The goal is to use these internal indices for unsupervised learning to find patterns in the data that can be used for future analysis in RStudio (Tan et. al., 2006).

The Abalone species has one genus in the family Halitidae, four to seven subgenera and around 100 species throughout the world. The Halitodae genus will cling to solid rocky structures and feed off red or brown algae (Toweraqua, 2017). The Blacklip Abalone (Haliotis rubra) is the species of Abalone measured in this dataset and is a large flat marine gastropod mollusk found in the southeastern Australian coastline (I & U NSW, ND). This species can live over 20 years, reach 22 cm in shell length and weigh over three kilograms for this time frame. This species matures around 9 to 10 cm in shell length (age is around 3 to 6 years) and this is the age needed to be able to reproduce and help sustain the population of the species.

The **Blacklip Albalone** species growth rates are highly variable along the Australian coast and these species have seen significant population declines due to commercial and recreational fishers and additional mortality rates due to parasites such as **Perkinsus**. Areas along the Port Stephens and Jervis Bay have banned Abalone fishing to correct this problem. This study has originally been conducted to determine what factors affect the age or number of rings to maintain a stable population for years to come (I & U NSW, ND).

The goal of this study is to find patterns or events in the Abalone dataset through unstructured K-means algorithm. These patterns can be used to understand the dataset and identify points within clusters that can be used for future research analysis in RStudio.

Analysis Method

The Abalone data set was chosen for this project and came from the UCI Learning Center and the KEEL Website (Lichman, 2013; KEEL, ND). The Abalone data from this site is being used for supervised learning algorithms to determine the age (number of rings) of this organism from the first eight variables listed in Table 1. The number of rings were measured through the physical cutting of the shell and counting the number of bands under a microscope. The Abalone dataset was in column delineated form and imported into RStudio for data preparation and analysis. Figure 1A (Appendix) has the R Commands and the list of the variables found in this dataset and Figure 2A is the summary of the dataset. Please note that all figures that end in an "**A**" will be in the Appendix and all others will be found in this document. The dataset contains a total of 9 attributes and 4,177 observations. These attributes are **Sex, length, diameter, height, whole_weight, viscera_weight, shell_weight and Rings**. The condensed version of this list can be seen in Table 1 below.

The first goal for the K-means and Hierarchical Clustering algorithm is to clean and prepare the data for analysis. The major preprocessing steps are to make sure all variables are numeric, the data is scaled or standardized and there are no missing values for the data set. The Abalone set was copied and then the sex attribute factor was changed to a numerical variable in R and saved as a **sex.num** attribute. The dependent or Rings variable needs to be removed and both the Sex and Rings attributes were removed from this study with a NULL command. The

commands and output can be seen in Figures 3A and 4A. Clustering algorithms do not handle missing variables and the summary () and apply() commands were used to validate and show this dataset did not have any missing variables and this can be seen in Figures 5A and 6A. The

Table 1: List of all 9 variables assigned to this study.

ID	Name of Variable	List of the top five variables for each category	Variable Type
1	Sex	"F","I","M": 3, 3, 1, 3, 2	Factor
2	Length	0.455, 0.35, 0.53, 0.44, 0.33	Number
3	Diameter	0.365, 0.265, 0.42, 0.365, 0.255	Number
4	Height	0.095, 0.09, 0.135, 0.125, 0.08	Number
5	Whole_weight	0.514, 0.226, 0.677, 0.516, 0.205	Number
6	Shucked_weigth	0.2245, 0.0995, 0.2565, 0.2155, 0.0895	Number
7	Viscera_weight	0.101, 0.0485, 0.1415, 0.114, 0.0395	Number
8	Shell_weight	0.15, 0.07, 0.21, 0.155, 0.055	Number
9	Rings	15, 7, 9, 10, 7	Integer

summary command was shown in Figure 2A, but was rerun to double check the data set before beginning the analysis.

The K-means algorithms needs to have a defined number of clusters to initiate the analysis and there are several ways to determine the cluster size. This study will use the "Elbow or Bend Method", Calinsky criterion and finally the NBClust package to determine the optimal cluster size for this study (STHDA, ND.: Stackoverflow, ND). The Elbow method will compare the sum of squared error for a number of cluster solutions. When the cluster size increases, the SSE should decrease and can produce an elbow in the data. This can show which cluster sizes do not have a major effect on the analysis (Figure 6A to 8A) (Peoples, M. 2017). The Calinski-Harabasz Criterion or variance ratio criterion is a ratio of overall between cluster variance over within-cluster variance times the number of clusters divided by the number of observations (Figure 9A to 10A) (MathWorks, ND). Finally, the last induce used for cluster analysis is the NbClust which computes 30 induces in a single function and lists the best cluster size to use by majority rules and the subsequent second and third best clusters for the analysis (Figure 11A to

12A) (STHDA, ND). The optimal cluster size for two of these indices was three and the third (Calinski) had two clusters as the optimal size. The first K-Means analysis will use three as the control cluster size and then two, seven and fifteen cluster sizes will be evaluated to determine if reducing or increasing cluster size will add additional information for the Abalone dataset.

The **cluster**, **caret**, **lattice** and **ggplot2** files were opened in RStudio to begin the K-means analysis using three clusters. The **set.seed(32)** command and K-means function was initiated with 3 clusters and the output from this model can be seen in Figures 13A and 14A . A total of four graphs will be generated for each K-means cluster and these are; Cluster plot, K-Means plot, Discriminant Projection Plot and a Parallel Coordinates Plot. These plots can be seen in Figure 15A to 18B. The size of the cluster, between sum of squares, total within sum of squares, within sum of squares, iteration number and cross validation evaluation was determined for each model and the Adjusted Rand Index was also calculated for each model. The output for these values in the three-cluster model can be seen in Figure 19A to 21A.

A total of five different K-mean cluster models were completed for this project and all of the previous commands were included in each analysis. The number of clusters that were tested are two (Figure 22A to 30A), seven (Figure 31A to 39A), fifteen (Figure 40A to 48A) and two clusters with different variations in the sex attributes for the Abalone species. These two variations for the sex variable are a K-means cluster with only male and female attributes (Figure 49A to 59A) (removed the infant category) and the K-means cluster with only the female and infant attributes (Figure 60A to 69A) (removed the male attribute). Finally, the PAM algorithm was run for this project and this can be seen in Figures 70A to 72A.

Results

The first goal after preprocessing the data for K-means is to determine the cluster size and the optimal cluster size of three was chosen by the **Elbow** and **NBClust Method**. The **Calinski** criterion had two as optimal cluster size but had several error messages due to convergence issues. The number of iterations was increased from 100,000 to 1 million and it still recommended two as the optimal cluster size, but had 50 warning messages on convergence. The **NBClust** Method also had two and fifteen clusters sizes chosen by four criteria in the model as the second-best choice for cluster selection. The number of clusters that will be used for this analysis is three as the base line and will be compared to algorithms with cluster sizes of two, seven and fifteen to determine the best model that will be used to look for patterns in the Abalone data set. The reason why cluster number seven was chosen is the within group sum of squares for Figure 7A had a good 200-point upswing in the data and this also divided the three to fifteen cluster range in half for the K-means analysis.

The cluster size, data instance per cluster, total within sum of squares, cross-tabulation percent (highest number of instances for each row divided by the total number of instances), ARI or Adjusted Rand Index and the number of iterations were summarized for all four K-means in Table 2. The K-means with two clusters (aba3b file) had the best ARI index value at 0.3965 and the cross-tabulation percent of 78.3%. The three-cluster model (**aba3**) had the second best cross tabulation percent and the fourth best ARI score of 0.0400. The seven-cluster model (aba7) had the second-best ARI score of 0.0563 and the third best cross tabulation percent with a value of 32.5%. Most of the API values are near zero which shows a complex data set with no clear separation of instances for the clusters in the model (Santos and Embrechts, ND). The total

Table 2

Cluster Size ()	Data instances for each cluster	Total Within Sum of Squares	Cross-Tabulation (%)	ARI	# of iterations
(3) aba3	1528, 1,304, 1345	1,000.968	2,186/4,177 or 52.3%	0.04001	2
(2) aba3b	2775, 1402	1,706.015	3,270/4,177 or 78.3%	0.3965	1
(7) aba7	715, 663, 545, 251, 813, 645, 545	364.25	1,359/4,177 or 32.5%	0.0563	2
(15) aba15	496, 603, 229, 142, 268, 664, 106, 188, 79, 33, 230, 232, 173, 493, 241	183.31	985/4,177 or 23.6%	0.0462	3

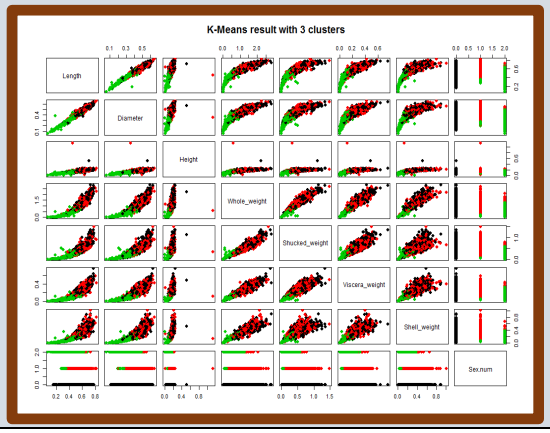
** ARI is Adjusted Rand Index

within sum of squares values does go down as cluster number increases and the total number of iterations used by these four models ranges from one to three. The best K-means models to use for this analysis is the two-cluster model compared to the control with three clusters in this study.

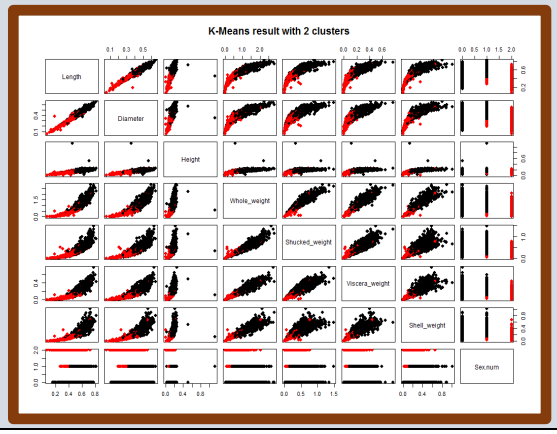
The K-means plot for three and two clusters is copied from the appendix and pasted in Figure 1 below. These plots had interesting results from this data set and this is what a biologist would expect to find from a population of species for a given area. The three cluster K-means plot shows the three colors used for separation of the data. The lime green for 1a and most of the red color for 1b represents the Infant population and is at the lower end of the physiological parameters measured in this sample. The black color for both 1a represent the males in this model and slightly more than half of the males for 1b in the two-cluster model. The black color in the 1b or two cluster model represent all the males and almost all the females. Figure 1 can be seen in the Appendix under Figure 15B and 25A.

The Cluster Plot and Discriminate Cluster Analysis was plotted to show how well these two algorithms did in separating the data and this can be seen in Figure 2. The plots for three

Figure 1: Comparing K-Means Plot comparing three cluster to two clusters model.



1a. K-means plot with three clusters

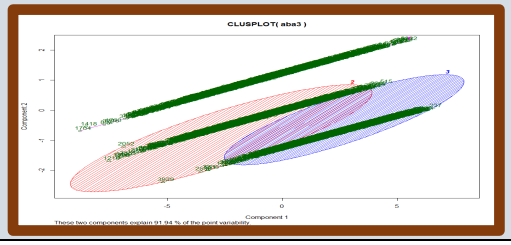


1b. K-means plot with two clusters

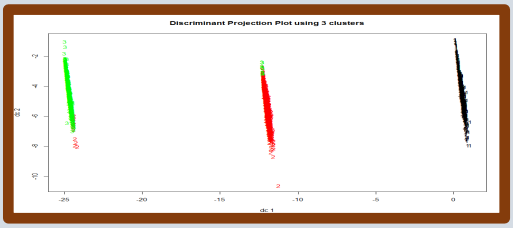
clusters did a good job in separating the sex attributes with some overlap between the females and infants. The plots with only two clusters did a good job in separating the infants from the adult population with again some mixing of females and infants. This algorithm did not separate the adult population by gender. When the K-mean algorisms were set to seven and fifteen clusters (Figure 35A and 44A), the infants were separated into four or more clusters (instead of one) and the males had two or more clusters for these models (instead of mostly one).

The kc$center means command was initiated in Rstudio for the three, two and seven cluster algorithms of the eight variables and the results can be seen in Figure 3. The differences between the means for the first two clusters is only by three hundreds of a point and more separation for the last cluster. There was slightly more separation between the means for both centers than the previous model for the algorithm with two clusters. The seven cluster center means was included for comparison purposes and again there is more variation between the center means and there seems to be two pairs of values that are closely related from the means within these clusters as seen in Figure 3.

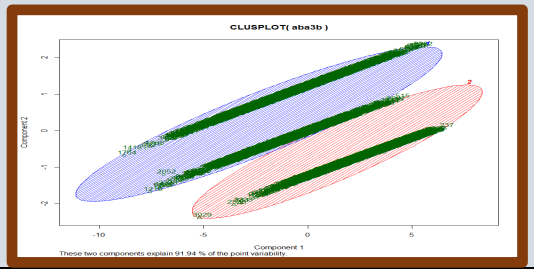
Figure 2: Cluster Plot and Discriminant Projection Plot Graphs



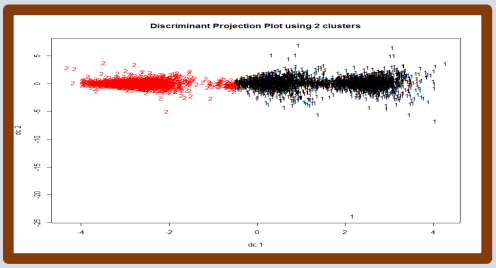
1a. Clusplot with three clusters



1b. Discriminant plot with three clusters



2a. Clusplot with two clusters



2b. Discriminant plot with two clusters

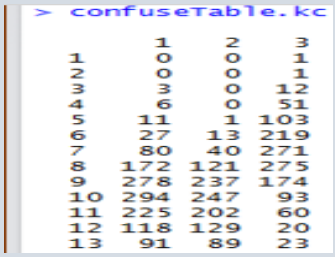
Figure 3

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ##################### Intiating K Center means for Three Cluster analysis
> kc$centers
     Length  Diameter     Height Whole_weight Shucked_weight Viscera_weight Shell_weight  Sex.num
1 0.5613907 0.4392866 0.1513809    0.9914594       0.4329460      0.21554450    0.2819692 0.000000
2 0.5825537 0.4575613 0.1591526    1.0606534       0.4520890      0.23370284    0.3060606 1.009202
3 0.4247286 0.3240372 0.1070000    0.4190439       0.1858829      0.08939703    0.1246428 1.988848
>
>
> ##################### Intiating K Center means for Two Cluster analysis
> kc1b$centers
     Length  Diameter     Height Whole_weight Shucked_weight Viscera_weight Shell_weight   Sex.num
1 0.5737477 0.4498072 0.1556018    1.0329150       0.4461133       0.2260555    0.2956755 0.4493694
2 0.4255100 0.3248966 0.1076783    0.4246198       0.1876701       0.0906102    0.1263174 1.9572040
>
>
> ################################ Initiating K Center Means for Seven Cluster Analysis
> kc7$centers
     Length  Diameter      Height Whole_weight Shucked_weight Viscera_weight Shell_weight Sex.num
1 0.6404406 0.5038531 0.17553846    1.3831399      0.61449930      0.29968042    0.38283077 0.00000
2 0.5151433 0.4022474 0.13970588    0.7085151      0.29612670      0.15716063    0.21150226 1.00000
3 0.4686881 0.3594404 0.11762385    0.4799890      0.21352477      0.10219908    0.14174679 2.00000
4 0.5677490 0.4416534 0.14924303    0.8826375      0.38914343      0.18742629    0.25859761 2.00000
5 0.4918696 0.3825031 0.13013530    0.6469926      0.27327737      0.14155043    0.19326568 0.00000
6 0.6450543 0.5088295 0.17692248    1.3955372      0.60094419      0.30659380    0.39559070 1.00155
7 0.3217798 0.2401009 0.07916514    0.1719321      0.07623761      0.03723761    0.05358807 2.00000
```

The last topic to discuss is the cross-tabulation table and this can be used to figure out how many instances the Abalone cluster assignment matches the actual Ring type for this algorithm and the percent of Abalone instances that were clustered in agreement with the Ring type. Most the ARI index values were very close to zero for these models which indicated a complex dataset and no clear separation of clusters for these models. The control for this analysis had the **confuseTable** copied and pasted into Figure 4 to show how the instances are

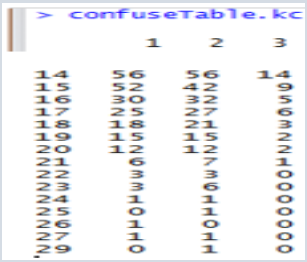
measured for all three clusters. The Ring attributes from one to four and 21 to 29 show some separation in the instances. However, the instances between five to 21 show a lot of mixing of data points and no clear separation of the values. The examples given in R with the Iris data set shows there was a clear separation of the three species of Iris' which this model does not have.

Figure 4



```
> confuseTable.kc
      1   2   3
1     0   0   1
2     0   0   1
3     3   0  12
4     6   0  51
5    11   1 103
6    27  13 219
7    80  40 271
8   172 121 275
9   278 237 174
10  294 247  93
11  225 202  60
12  118 129  20
13   91  89  23
```

1a. ConfuseTable 1 to 13 Rings



```
> confuseTable.kc
      1   2   3
14   56  56  14
15   52  42   9
16   30  32   5
17   25  27   6
18   18  21   3
19   15  15   2
20    1   2   2
21    6   7   1
22    3   3   0
23    3   6   0
24    1   1   0
25    0   1   0
26    1   0   0
27    1   1   0
29    0   1   0
```

1b. ConfuseTable 14 to 29 Rings

Interpretation of Results

The K-means algorithms were used to determine the structure of the Abalone data set from the UCI Learning Center. The data set was rather large with 4,177 samples and a total of nine variables. Selecting cluster size for K-means is a challenge and several indices were tested and the cluster size of three had the most votes. The cluster sizes of two, seven and fifteen were also selected based on these indices and were used to give a nice snapshot of this data set. These can be seen in 13A to 48A in the Appendix. The cross-tabulation percent and ARI index was also looked at and the K-means algorithms with three (control) and two clusters were selected to look for patterns and problems with the dataset.

The Abalone data set contained mostly weight values from the destructive sampling of these species. The article did not go into how these individuals were harvested from the Australian coastline. The population contained three variables for sex and these are Infants (sex is undetermined), males and females. The control or K-means with three clusters did a good job

in separating the infants from the individuals that could be sexed. These values were at the tail end of the axis near zero and then slightly overlapped with the female attributes. The male and female values were very hard to separate after the infant characteristics ended and the population could be sexed. The older population does not seem to be sexually dimorphic (one sex is larger than the other like a bull moose or male elephant seal) and this can be seen in Figure 16B. This shows the red color for females and tail end of the infants and the black color for males. When you look at the shell diameter that is greater than 0.3 cm, the data is evenly mixed with both sexes. The shell weight attribute does a good job in showing how the male and female points are evenly mixed as the values increases from 0.25 to 0.30 cm.

The data set did not contain any missing variables which helped in the preprocessing steps, but there were more outliers in the sample size than originally expected. The ramifications to the data with outliers will be discussed in the next section. A boxplot was run in RStudio and can be seen in Figure 5 below and the boxes or 50^{th} percentile show very little variation among each attribute except for **Whole_weight** and **Sex**. The other attributes values ranged between 0.25 to 0.60 with the data and each attribute between the Lower and Upper Quartile range was 0.20 for each category except for **Shucked_weight (**Figure 5). These values were very tight before and after the data had been scaled for the preprocessing step. The typical spherical clusters shown in other examples could be seen for the two and part of the three-cluster algorithm, but as cluster size increased the spheres were replaced with lines of highly concentrated data points. The K-means with seven or fifteen clusters might be the best approach to handle these data points because these lines are segmented into an equal number of clusters and this minimizes the chance of a data point from one cluster being closer to the mean of another cluster as shown in Figure 6 below.

The last two K-means analysis conducted on this data set can be seen in Figures 49A to 69A in the Appendix. The goal was to remove one of the sex attributes for each analysis to determine if the this produced any unusual patterns in the data. The first K-means algorithm had the infant attribute removed from the data set and the second K-means algorithm had the male attribute removed (only the female and infant were present). The same R commands were completed for these datasets and the output can be seen in the Appendix. The K-means for only the males and females had very good separation of points as shown from the clusterplot, but the cross validation still looked like an equal split of data points into both clusters. The ARI index was near zero and the para-coordinates plot show these variables being equally spread throughout the attributes. The last K-means analysis took out the male attribute and only had the infants and female attributes for this analysis. The ARI index was low and around 0.0567 but there was a little more separation of data points for the cross-tabulation values and this percent value was around 75.5.

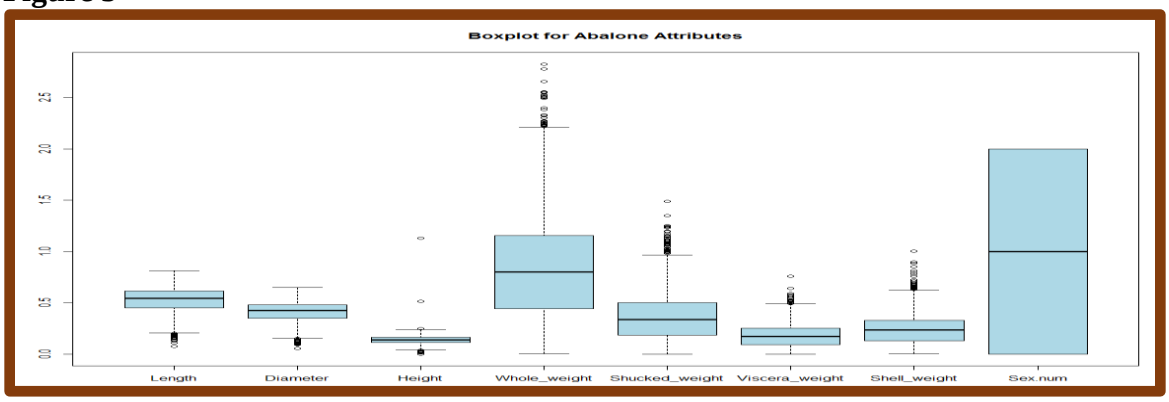
The PAM clustering was quickly run as an alternative approach to the K-means clustering. This stands for Partitioning around Medoids and is minimal distance between all observations and the center. The Results can be seen in Figures 70A to 72A. The recommended cluster size for the PAM algorithm was also three and the silhouette interpretation of these clusters based on the Statistics at Berkeley show Cluster 3 with a reasonable structure at 0.69 and two barely weak clusters for one and two.

Experimental Limitations

K means clustering is the most commonly used algorithm in R as discussed in the literature. There was a quote mentioned while looking for the possible limitations in dealing with mathematical theorems. The quote states there is “No Free Lunch Theorem” by Wolpert

and Macready in mathematical folklore (CrossValidated, ND). All models have their limitations and K-means is no exception. One experimental limitation is dealing with outliers in the dataset. The article discusses when squared error criteria is used for the algorithm, the presence of outliers can affect the cluster centroids and this may not be a representative sample for the data set. The article says the outliers need to be removed in the prepressing step before initiating the algorithm or they can be identified in the postprocessing step and be eliminated over multiple runs in R (Tan et al., 2006). This dataset had a few outliers which can be seen in Figure 5 below. The Abalone **height, Whole_weight, Shucked_weight**, **Viscera_weight** and **Shell Weight** all had several outliers away from the data set as shown in the boxplot () function in R for this dataset. These outliers can also be seen in the para-coordinates graph for Figure 18B. The outliers will need to be removed in the preprocessing or post processing steps to minimize their effects on of the cluster centroids.

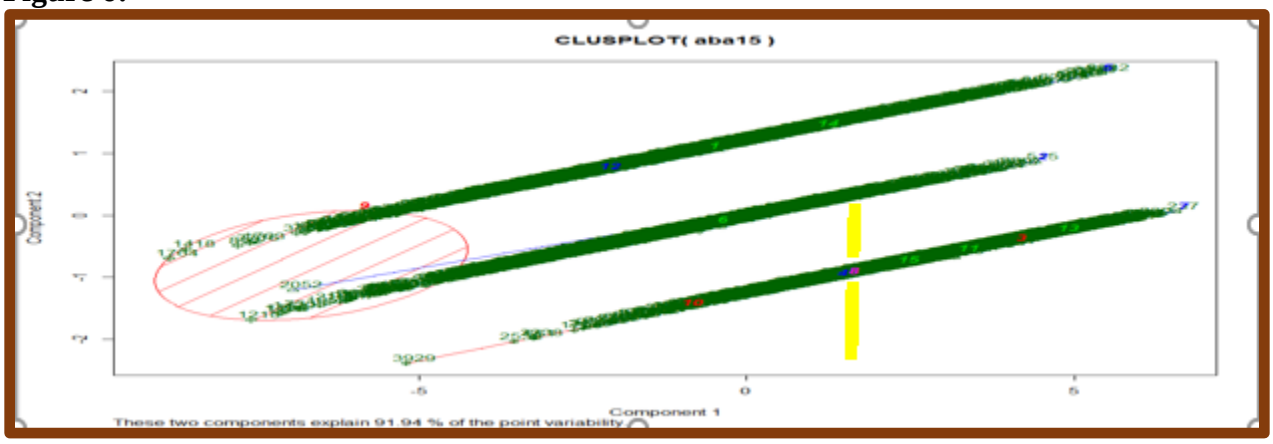
Figure 5



The K-means does well with generating a specific number of clusters that are globular in nature. However, it does not do well with non-globular clusters (chains) or clusters of different sizes. The K-means defines clusters based on the Euclidean distance to a point (center) and creates spheres separated from these points. When the data is non-globular, the k-means algorithm clusters are farther from the spheres affecting the algorithm because some points will

be closer to the wrong cluster center than the one it was placed in by the program (Biostars, ND). This data set was long lines of clustered data and the fifteen-cluster plot algorithm was copied and shown in Figure 6 below. This figure shows the chain or lines of data which these data points have been placed on for this analysis. The line closest to the x-axis shows if cluster size is minimal then cluster centers 4 and 8 could be placed in either cluster. The data for three lines are tightly packed and if the cluster centers are equally spaced apart then this should have minimal effect on the cluster centers.

Figure 6:



Conclusion

The Abalone data set contained a total of 8 attributes (Rings was the ninth one) that were used for unsupervised clustering for the K-means algorithm. The goal was to find interesting structure for this dataset that could be used for future analysis in a predictive model for age or number of Rings. Several cluster indices were used to determine appropriate cluster size and three clusters was the most recommended amount. The data set could be divided into two sections as shown in the two-cluster analysis, which did a good job of separating the infant population from the adult population. However, the adult population had equal values for each

attribute that could not be separated and this can be seen in the cross-tabulation tables in the appendix. The dataset did have a lot of outliers which need to be pre-or post-processed for the next analysis to minimize the errors for the K-means clustering.

The Abalone study needed additional attributes that could be divided into separate clusters such as biomass intake from the brown or red algae, predation rates for the specific area of interest, harvest rates by local or commercial fishing industries which could help determine the effects this has on the fecundity rates of these species.

References

Biostars.org. (ND). Question: K-means for non-spherical (non-globular) clusters. Retrieved from: https://www.biostars.org/p/137008/

CrossValidated. (ND). How to understand the drawbacks of K-means. Retrieved from: https://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means

Han, Kamber, and Pei. Data Mining: Concepts and Techniques, Third Edition (2011). Retrieved from: http://web.engr.illinois.edu/~hanj/cs412/bk3/08.pdf

I & I NSW. (ND). Blacklip Abalone (Haliotis rubra). Wildlife Fisheries Program. Retrieved from:http://www.dpi.nsw.gov.au/_data/assets/pdf_file/0009/375858/Blacklip Abalone.pdf

KEEL. (ND). Abalone data set. Retrieved from: http://sci2s.ugr.es/keel/dataset_ smja.php?cod=1346

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

MathWorks. (ND). Clustering.evaluation.CalinskiHarabaszEvaluation class. Retrieved from: https://www.mathworks.com/help/stats/clustering.evaluation.calinskiharabaszevaluation-class.html

Peoples, M. (2017). R Script for K-Means Cluster Analysis. Society for American Archeology Style. Retrieved from: http://www.mattpeeples.net/K-means.html

Santos, J. and Embrechts, M. **(ND).** On the use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. ISEP. Retrieved from: https://pdfs.semanticscholar.org/52d4/8b393f3f838f2370c50af03703eee0bbd669.pdf

Stackoverflow, (ND). Cluster analysis in R: determining the optimal number of clusters. Retrieved from: http://stackoverflow.com/questions/15376075/ cluster-analysis-in-r-determine-the-optimal-number-of-clusters/15376462

Stat.berkeley. (ND). Clustering Techniques. Retrieved from: https://www.stat.berkeley.edu /~s133/ Cluster2a.html

References (continued)

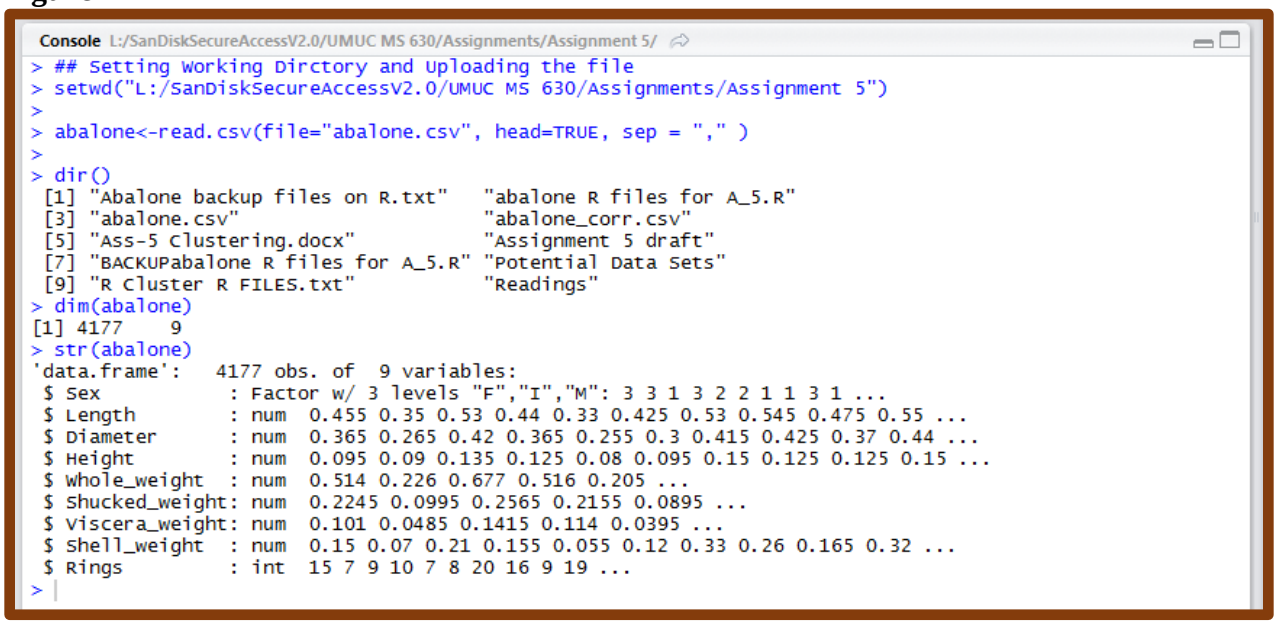
STHDA. (ND). Determining the optimal number of clusters: 3 must known methods-Unsupervised Machine Learning. Retrieved from: http://www.sthda.com/english/wiki/determining-the-optimal-number-of-clusters-3-must-known-methods-unsupervised-machine-learning

Tan, P., Steinback, M. and Kumar, V. (2006). Introduction to Data Mining: Cluster Analysis: basic concepts and algorithms. Retrieved from: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

Toweraqua, (2017). Abalone Facts. Tower Aqua Products, Europe's Abalone Specialists. Retrieved from: http://www.toweraqua.com/ourfamily.php

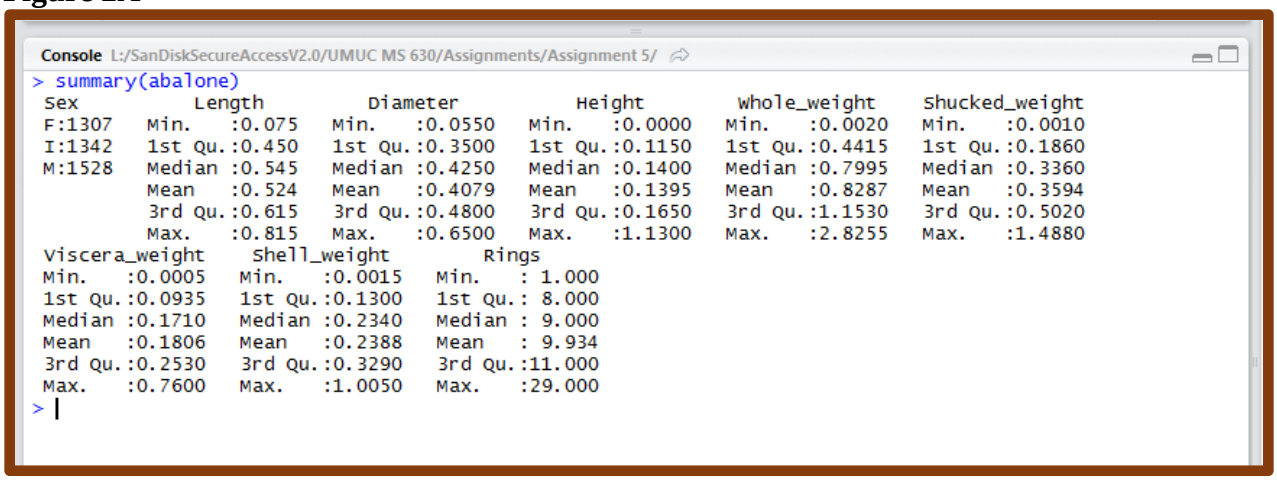
Appendix

Figure 1A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ## Setting Working Dirctory and Uploading the file
> setwd("L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5")
>
> abalone<-read.csv(file="abalone.csv", head=TRUE, sep = "," )
>
> dir()
 [1] "Abalone backup files on R.txt"   "abalone R files for A_5.R"
 [3] "abalone.csv"                     "abalone_corr.csv"
 [5] "Ass-5 Clustering.docx"           "Assignment 5 draft"
 [7] "BACKUPabalone R files for A_5.R" "Potential Data Sets"
 [9] "R Cluster R FILES.txt"           "Readings"
> dim(abalone)
[1] 4177    9
> str(abalone)
'data.frame':   4177 obs. of  9 variables:
 $ Sex           : Factor w/ 3 levels "F","I","M": 3 3 1 3 2 2 1 1 3 1 ...
 $ Length        : num  0.455 0.35 0.53 0.44 0.33 0.425 0.53 0.545 0.475 0.55 ...
 $ Diameter      : num  0.365 0.265 0.42 0.365 0.255 0.3 0.415 0.425 0.37 0.44 ...
 $ Height        : num  0.095 0.09 0.135 0.125 0.08 0.095 0.15 0.125 0.125 0.15 ...
 $ Whole_weight  : num  0.514 0.226 0.677 0.516 0.205 ...
 $ Shucked_weight: num  0.2245 0.0995 0.2565 0.2155 0.0895 ...
 $ Viscera_weight: num  0.101 0.0485 0.1415 0.114 0.0395 ...
 $ Shell_weight  : num  0.15 0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 ...
 $ Rings         : int  15 7 9 10 7 8 20 16 9 19 ...
>
```

Figure 2A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> summary(abalone)
 Sex          Length          Diameter          Height        Whole_weight    Shucked_weight
 F:1307   Min.   :0.075   Min.   :0.0550   Min.   :0.0000   Min.   :0.0020   Min.   :0.0010
 I:1342   1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150   1st Qu.:0.4415   1st Qu.:0.1860
 M:1528   Median :0.545   Median :0.4250   Median :0.1400   Median :0.7995   Median :0.3360
          Mean   :0.524   Mean   :0.4079   Mean   :0.1395   Mean   :0.8287   Mean   :0.3594
          3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650   3rd Qu.:1.1530   3rd Qu.:0.5020
          Max.   :0.815   Max.   :0.6500   Max.   :1.1300   Max.   :2.8255   Max.   :1.4880
 Viscera_weight    Shell_weight        Rings
 Min.   :0.0005   Min.   :0.0015   Min.   : 1.000
 1st Qu.:0.0935   1st Qu.:0.1300   1st Qu.: 8.000
 Median :0.1710   Median :0.2340   Median : 9.000
 Mean   :0.1806   Mean   :0.2388   Mean   : 9.934
 3rd Qu.:0.2530   3rd Qu.:0.3290   3rd Qu.:11.000
 Max.   :0.7600   Max.   :1.0050   Max.   :29.000
>
```

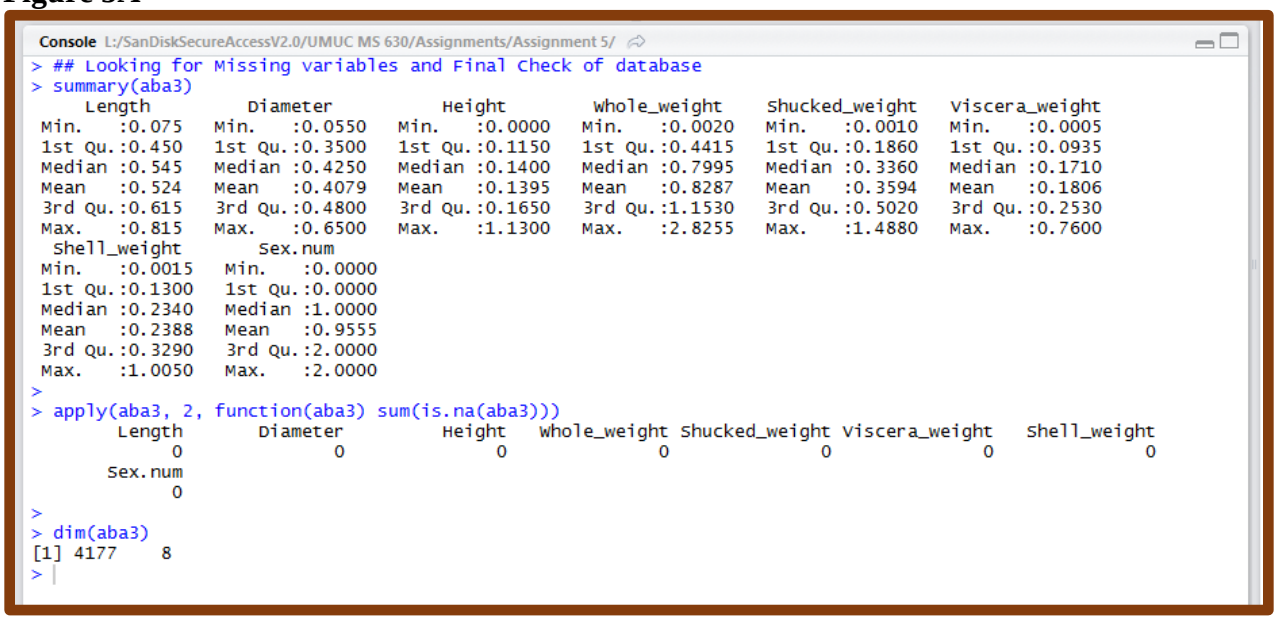
Figure 3A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> # Converting Factors in Gender to Numeric Variables
> aba2=abalone
> aba2$Sex.num[aba2$Sex=="M"]<-0
> aba2$Sex.num[aba2$Sex=="F"]<-1
> aba2$Sex.num[aba2$Sex=="I"]<-2
>
> str(aba2)
'data.frame':   4177 obs. of  10 variables:
 $ Sex           : Factor w/ 3 levels "F","I","M": 3 3 1 3 2 2 1 1 3 1 ...
 $ Length        : num  0.455 0.35 0.53 0.44 0.33 0.425 0.53 0.545 0.475 0.55 ...
 $ Diameter      : num  0.365 0.265 0.42 0.365 0.255 0.3 0.415 0.425 0.37 0.44 ...
 $ Height        : num  0.095 0.09 0.135 0.125 0.08 0.095 0.15 0.125 0.125 0.15 ...
 $ Whole_weight  : num  0.514 0.226 0.677 0.516 0.205 ...
 $ Shucked_weight: num  0.2245 0.0995 0.2565 0.2155 0.0895 ...
 $ Viscera_weight: num  0.101 0.0485 0.1415 0.114 0.0395 ...
 $ Shell_weight  : num  0.15 0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 ...
 $ Rings         : int  15 7 9 10 7 8 20 16 9 19 ...
 $ Sex.num       : num  0 0 1 0 2 2 1 1 0 1 ...
>
> ## Remove the Sex Factor and Ring Attribute
> aba3=aba2
> aba3$Sex=NULL
> aba3$Rings=NULL
>
>
> str(aba3)
'data.frame':   4177 obs. of  8 variables:
 $ Length        : num  0.455 0.35 0.53 0.44 0.33 0.425 0.53 0.545 0.475 0.55 ...
 $ Diameter      : num  0.365 0.265 0.42 0.365 0.255 0.3 0.415 0.425 0.37 0.44 ...
 $ Height        : num  0.095 0.09 0.135 0.125 0.08 0.095 0.15 0.125 0.125 0.15 ...
 $ Whole_weight  : num  0.514 0.226 0.677 0.516 0.205 ...
 $ Shucked_weight: num  0.2245 0.0995 0.2565 0.2155 0.0895 ...
 $ Viscera_weight: num  0.101 0.0485 0.1415 0.114 0.0395 ...
 $ Shell_weight  : num  0.15 0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 ...
 $ Sex.num       : num  0 0 1 0 2 2 1 1 0 1 ...
>
```

Figure 4A

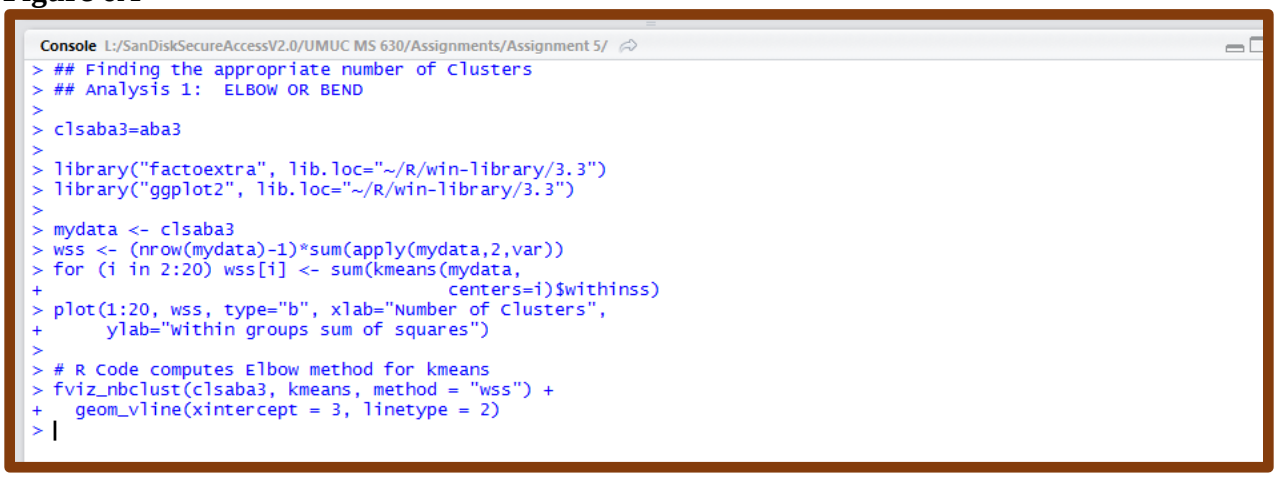
```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ## Scaling the attributes in the dataset
> scale(aba3)
            Length    Diameter      Height  Whole_weight Shucked_weight Viscera_weight Shell_weight     Sex.num
   [1,] -0.574489353 -0.43209706 -1.06429672 -0.6418213862  -0.6076126183  -0.7261246386 -0.638140488 -1.15420810
   [2,] -1.448812389 -1.43975662 -1.18383657 -1.2301298296  -1.1707696702  -1.2050769634 -1.212842110 -1.15420810
   [3,]  0.050027102  0.12211570 -0.10797794 -0.3094322136  -0.4634444130  -0.3566471308 -0.207114271  0.05379171
   [4,] -0.699392644 -0.43209706 -0.34705763 -0.6377429914  -0.6481599260  -0.6075269200 -0.602221637 -1.15420810
   [5,] -1.615350111 -1.54052258 -1.42291627 -1.2719333759  -1.2158222343  -1.2871830763 -1.320598665  1.26179152
   [6,] -0.824295935 -1.08707578 -1.06429672 -0.9731909600  -0.9838015289  -0.9405128221 -0.853653596  1.26179152
   [7,]  0.050027102  0.07173272  0.25064161 -0.1044928772  -0.5512969131  -0.3566471308  0.654938162  0.05379171
   [8,]  0.174930393  0.17249868 -0.34705763 -0.1238652523  -0.2944972975  -0.2836639194  0.152074243  0.05379171
   [9,] -0.407951631 -0.38171408 -0.34705763 -0.6509977744  -0.6436546696  -0.6212112722 -0.530383934 -1.15420810
  [10,]  0.216564823  0.32364761  0.25064161  0.1340932160  -0.2021395410  -0.2699795673  0.583100459  0.05379171
```

Figure 5A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ## Looking for Missing variables and Final Check of database
> summary(aba3)
     Length          Diameter          Height        Whole_weight    Shucked_weight   Viscera_weight
 Min.   :0.075   Min.   :0.0550   Min.   :0.0000   Min.   :0.0020   Min.   :0.0010   Min.   :0.0005
 1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150   1st Qu.:0.4415   1st Qu.:0.1860   1st Qu.:0.0935
 Median :0.545   Median :0.4250   Median :0.1400   Median :0.7995   Median :0.3360   Median :0.1710
 Mean   :0.524   Mean   :0.4079   Mean   :0.1395   Mean   :0.8287   Mean   :0.3594   Mean   :0.1806
 3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650   3rd Qu.:1.1530   3rd Qu.:0.5020   3rd Qu.:0.2530
 Max.   :0.815   Max.   :0.6500   Max.   :1.1300   Max.   :2.8255   Max.   :1.4880   Max.   :0.7600
  Shell_weight       Sex.num
 Min.   :0.0015   Min.   :0.0000
 1st Qu.:0.1300   1st Qu.:0.0000
 Median :0.2340   Median :1.0000
 Mean   :0.2388   Mean   :0.9555
 3rd Qu.:0.3290   3rd Qu.:2.0000
 Max.   :1.0050   Max.   :2.0000
>
> apply(aba3, 2, function(aba3) sum(is.na(aba3)))
        Length       Diameter         Height   Whole_weight Shucked_weight Viscera_weight   Shell_weight
             0              0              0              0              0              0              0
       Sex.num
             0
>
> dim(aba3)
[1] 4177    8
>
```

Figure 6A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ## Finding the appropriate number of Clusters
> ## Analysis 1:  ELBOW OR BEND
>
> clsaba3=aba3
>
> library("factoextra", lib.loc="~/R/win-library/3.3")
> library("ggplot2", lib.loc="~/R/win-library/3.3")
>
> mydata <- clsaba3
> wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
> for (i in 2:20) wss[i] <- sum(kmeans(mydata,
+                                      centers=i)$withinss)
> plot(1:20, wss, type="b", xlab="Number of Clusters",
+      ylab="Within groups sum of squares")
>
> # R Code computes Elbow method for kmeans
> fviz_nbclust(clsaba3, kmeans, method = "wss") +
+   geom_vline(xintercept = 3, linetype = 2)
>
```

Figure 7A

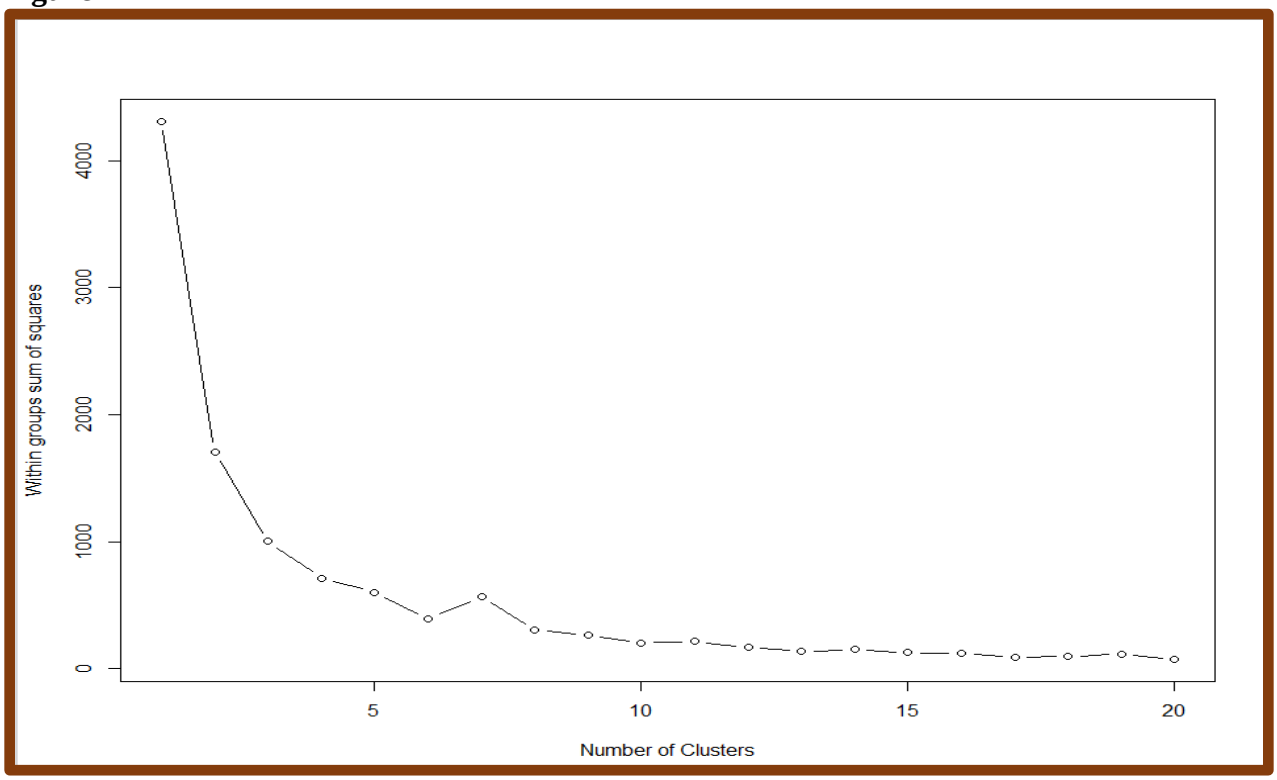


Figure 8A

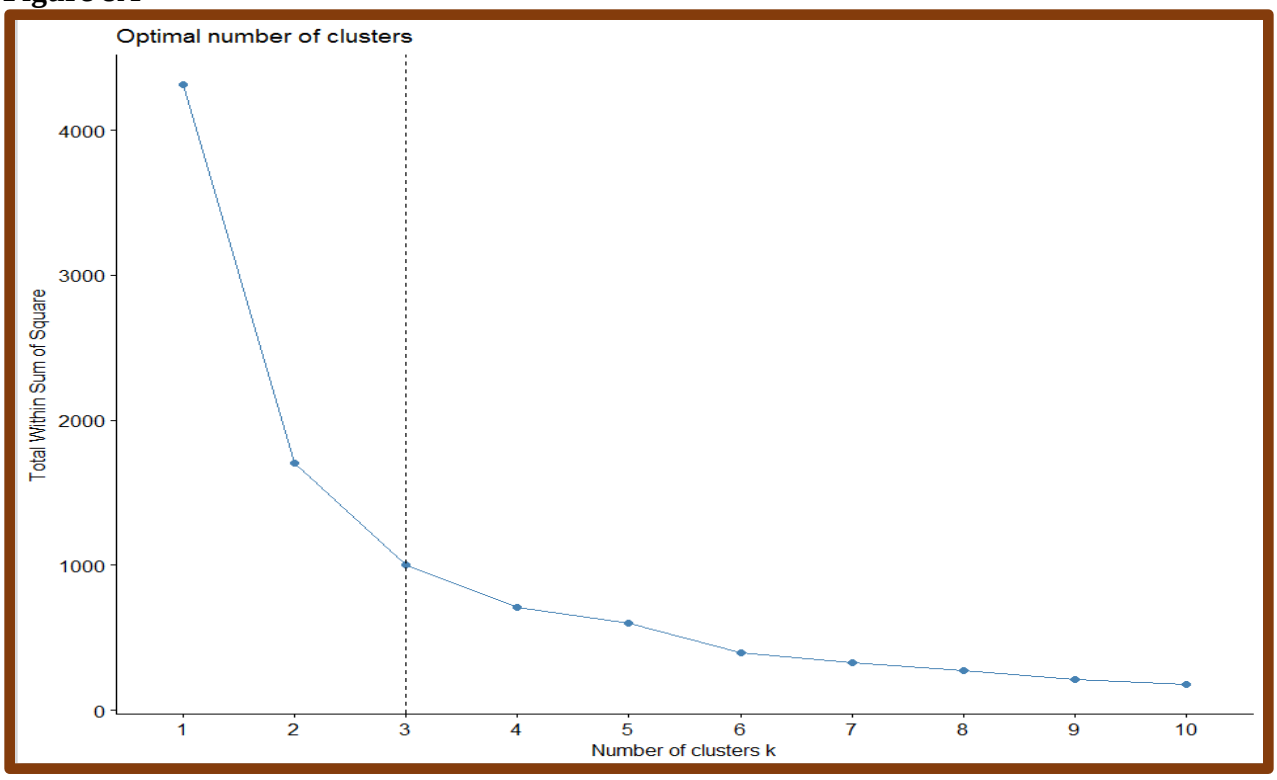


Figure 9A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ## Analysis 2 Calinsky Criterion from 1 to 10 groups
> library("permute", lib.loc="~/R/win-library/3.3")
> library("lattice", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("vegan", lib.loc="~/R/win-library/3.3")
> fit <- cascadeKM(scale(clsaba3, center = TRUE,  scale = TRUE), 1, 15, iter = 10000)
There were 24 warnings (use warnings() to see them)
> plot(fit, sortclsaba3 = TRUE, grpmts.plot = TRUE)
> calinski.best <- as.numeric(which.max(fit$results[2,]))
> cat("Calinski criterion  optimal number of clusters:", calinski.best, "\n")
Calinski criterion  optimal number of clusters: 2
>
```

Figure 10A

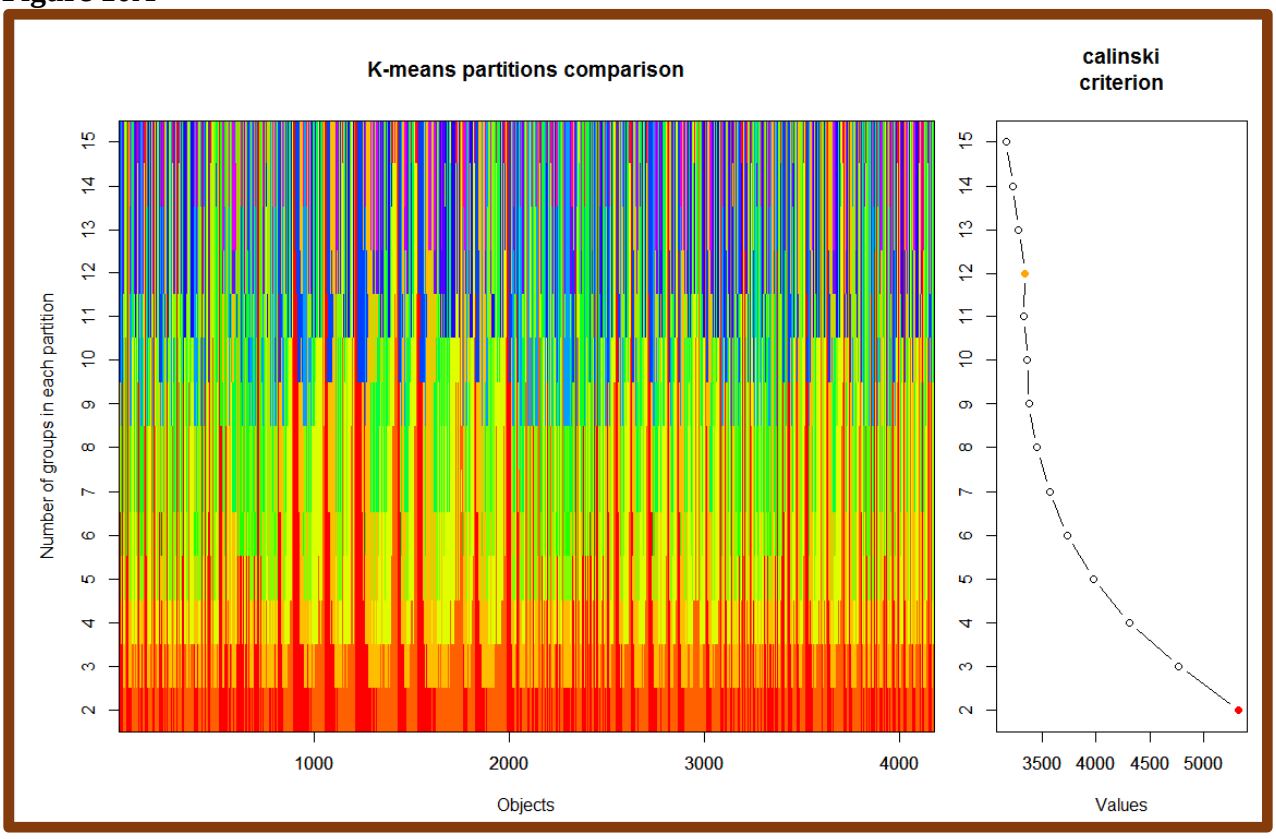
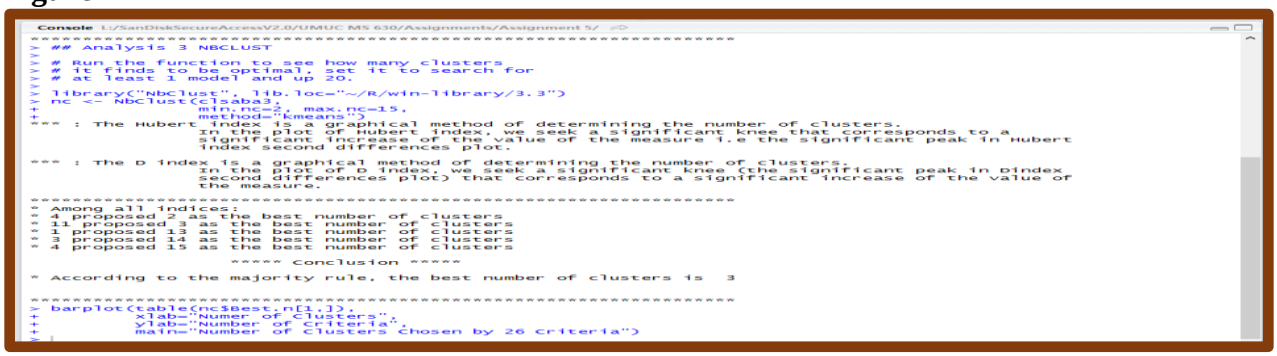


Figure 11A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MIS 630/Assignments/Assignment 5/
*******************************************************************
> ## Analysis 3 NBCLUST
>
> # Run the function to see how many clusters
> # it finds to be optimal, set it to search for
> # at least 1 model and up 20.
>
> library("NbClust", lib.loc="~/R/win-library/3.3")
> nc <- NbClust(clsaba3,
+               min.nc=2, max.nc=15,
+               method="kmeans")
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 4 proposed 2 as the best number of clusters
* 11 proposed 3 as the best number of clusters
* 1 proposed 13 as the best number of clusters
* 3 proposed 14 as the best number of clusters
* 4 proposed 15 as the best number of clusters

                ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3

*******************************************************************
> barplot(table(nc$Best.nc[1,]),
+         xlab="Numer of Clusters",
+         ylab="Number of Criteria",
+         main="Number of Clusters Chosen by 26 criteria")
>
```

Figure 12A

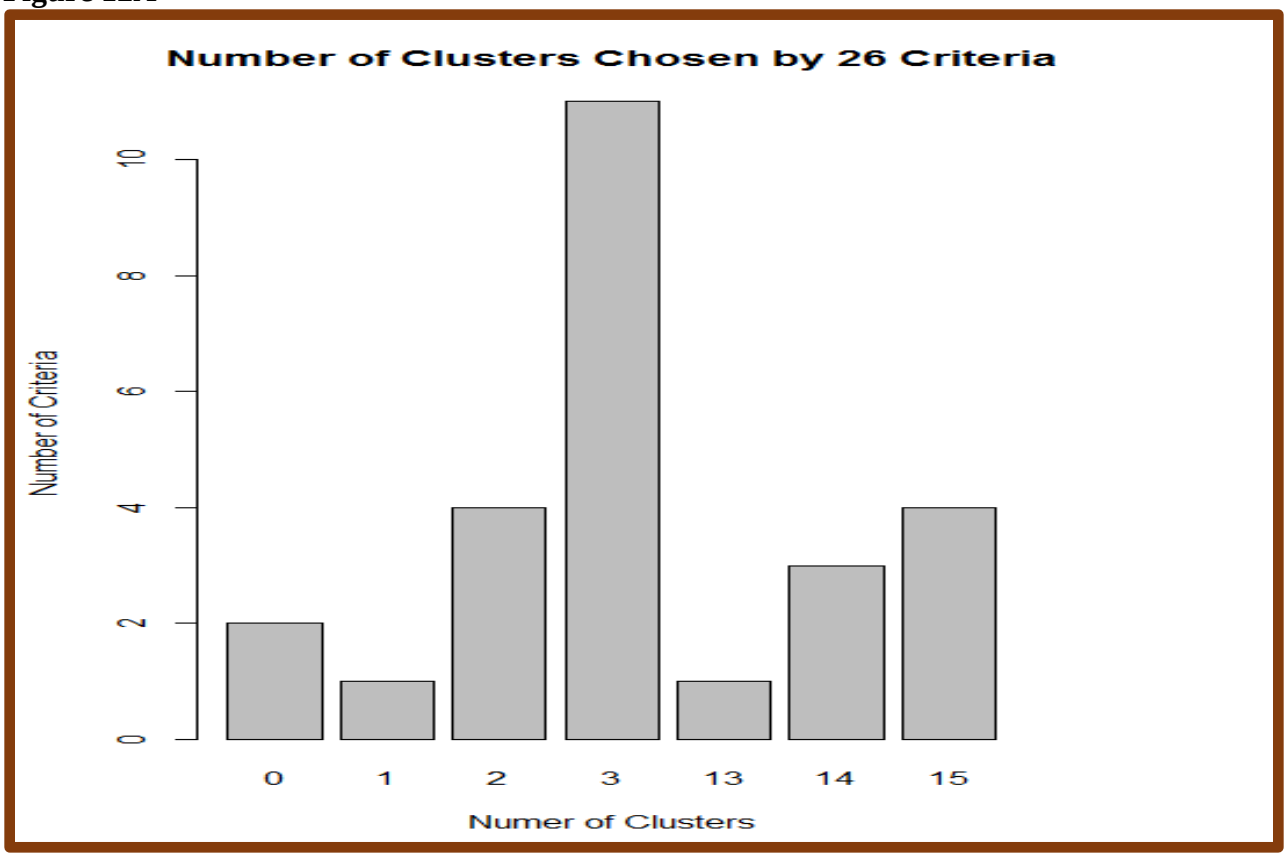
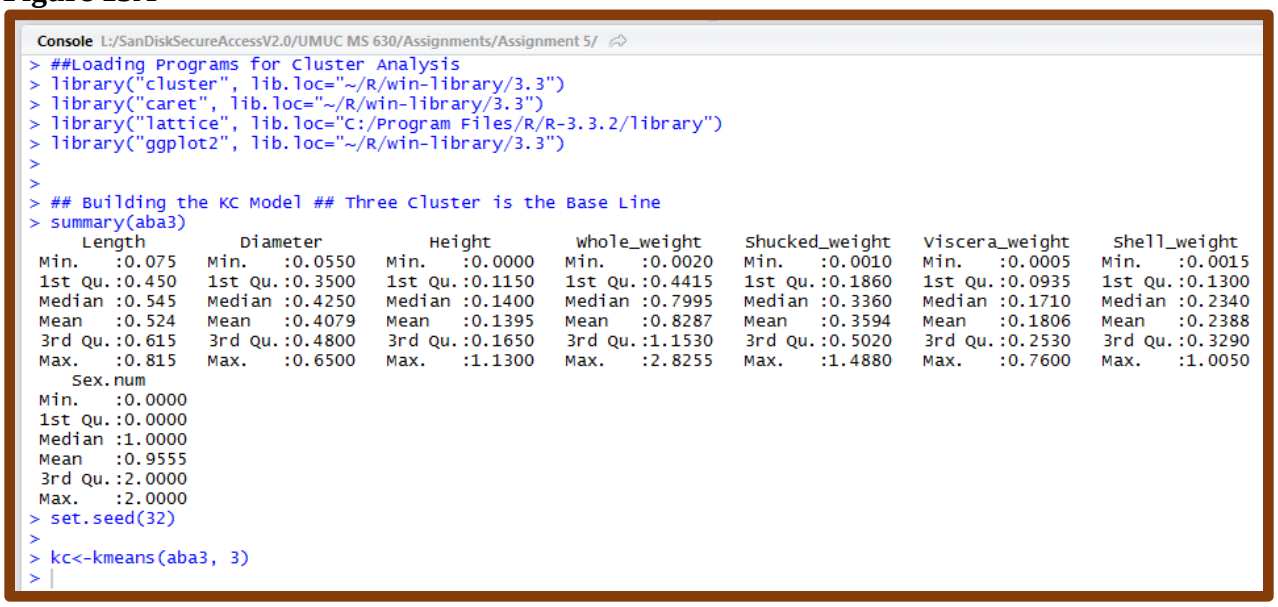
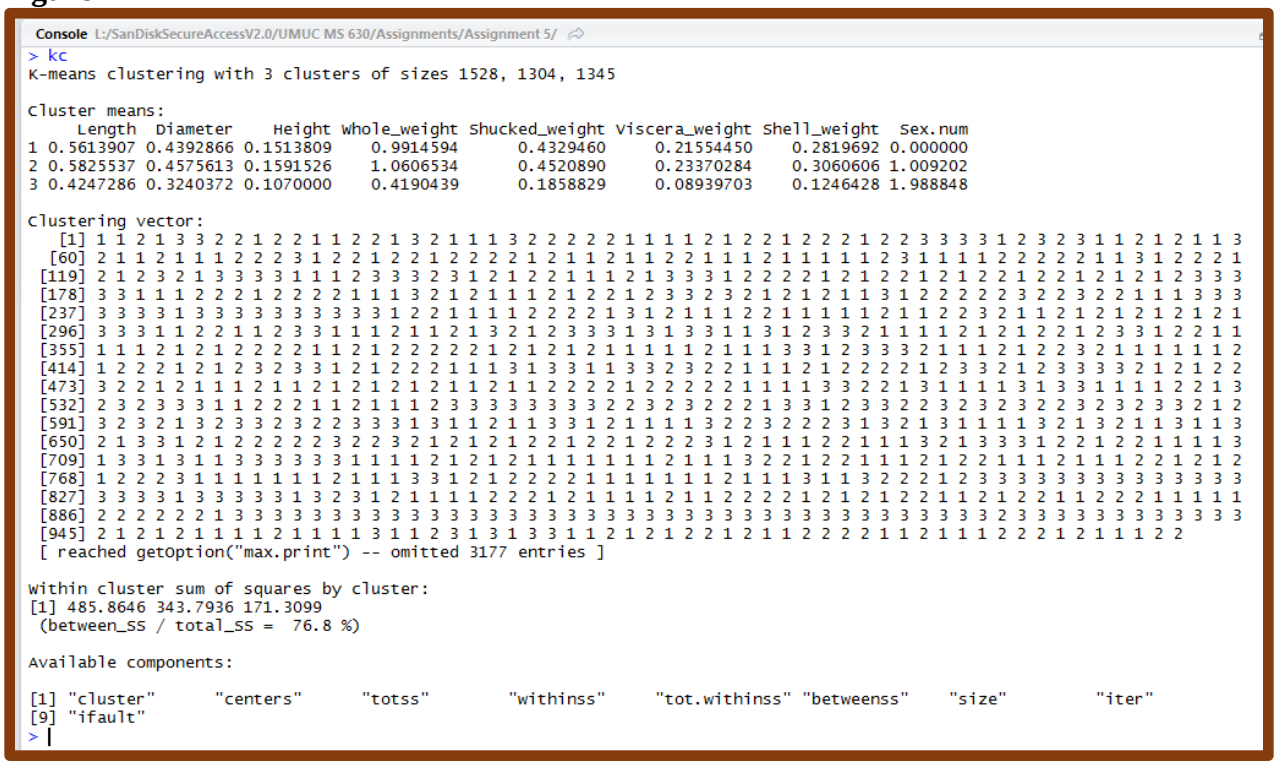


Figure 13A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ##Loading Programs for Cluster Analysis
> library("cluster", lib.loc="~/R/win-library/3.3")
> library("caret", lib.loc="~/R/win-library/3.3")
> library("lattice", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("ggplot2", lib.loc="~/R/win-library/3.3")
>
>
> ## Building the KC Model ## Three Cluster is the Base Line
> summary(aba3)
     Length          Diameter          Height        Whole_weight    Shucked_weight   Viscera_weight    Shell_weight
 Min.   :0.075   Min.   :0.0550   Min.   :0.0000   Min.   :0.0020   Min.   :0.0010   Min.   :0.0005   Min.   :0.0015
 1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150   1st Qu.:0.4415   1st Qu.:0.1860   1st Qu.:0.0935   1st Qu.:0.1300
 Median :0.545   Median :0.4250   Median :0.1400   Median :0.7995   Median :0.3360   Median :0.1710   Median :0.2340
 Mean   :0.524   Mean   :0.4079   Mean   :0.1395   Mean   :0.8287   Mean   :0.3594   Mean   :0.1806   Mean   :0.2388
 3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650   3rd Qu.:1.1530   3rd Qu.:0.5020   3rd Qu.:0.2530   3rd Qu.:0.3290
 Max.   :0.815   Max.   :0.6500   Max.   :1.1300   Max.   :2.8255   Max.   :1.4880   Max.   :0.7600   Max.   :1.0050
    Sex.num
 Min.   :0.0000
 1st Qu.:0.0000
 Median :1.0000
 Mean   :0.9555
 3rd Qu.:2.0000
 Max.   :2.0000
> set.seed(32)
>
> kc<-kmeans(aba3, 3)
>
```

Figure 14A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> kc
K-means clustering with 3 clusters of sizes 1528, 1304, 1345

Cluster means:
     Length  Diameter    Height Whole_weight Shucked_weight Viscera_weight Shell_weight  Sex.num
1 0.5613907 0.4392866 0.1513809    0.9914594      0.4329460     0.21554450    0.2819692 0.000000
2 0.5825537 0.4575613 0.1591526    1.0606534      0.4520890     0.23370284    0.3060606 1.009202
3 0.4247286 0.3240372 0.1070000    0.4190439      0.1858829     0.08939703    0.1246428 1.988848

Clustering vector:
   [1] 1 1 2 1 3 3 2 2 1 2 2 1 1 2 2 1 3 2 1 1 1 3 2 2 2 2 2 1 1 1 1 2 1 2 2 1 2 2 2 1 2 2 3 3 3 3 1 2 3 2 3 1 1 2 1 2 1 1 3
  [60] 2 1 1 2 1 1 1 2 2 2 3 1 2 2 1 2 2 1 2 2 2 2 1 2 1 1 2 1 1 2 2 1 1 1 2 1 1 1 1 1 2 3 1 1 1 1 2 2 2 2 2 1 1 3 1 2 2 2 1
 [119] 2 1 2 3 2 1 3 3 3 3 1 1 1 2 3 3 3 2 3 1 2 1 2 2 1 1 1 2 1 3 3 3 1 2 2 2 2 1 2 1 2 2 1 2 1 1 2 2 1 2 2 1 2 1 2 1 2 3 3 3
 [178] 3 3 1 1 1 2 2 2 1 2 2 2 2 1 1 1 3 2 1 2 1 1 1 2 1 2 2 1 2 3 3 2 3 2 1 2 1 2 1 1 3 1 2 2 2 2 2 3 2 2 3 2 2 1 1 1 3 3 3
 [237] 3 3 3 3 1 3 3 3 3 3 3 3 3 3 1 2 2 2 1 1 1 1 2 2 2 2 2 1 3 1 2 1 1 1 2 2 1 1 1 1 1 2 1 1 2 2 3 2 1 1 2 1 2 1 2 1 2 1 2 1
 [296] 3 3 3 1 1 2 2 1 1 2 3 3 1 1 1 2 1 1 2 1 3 2 1 2 3 3 3 1 3 1 3 3 1 1 3 1 2 3 3 2 1 1 1 1 2 1 2 1 2 2 1 2 3 3 1 2 2 1 1
 [355] 1 1 1 2 1 2 1 2 2 2 2 1 1 2 1 2 2 2 2 2 1 2 1 2 1 2 1 1 1 1 1 2 1 1 1 3 3 1 2 3 3 3 2 1 1 1 2 1 2 2 3 2 1 1 1 1 1 1 2
 [414] 1 2 2 2 1 2 1 2 3 2 3 3 1 2 1 2 2 2 1 1 1 3 1 3 3 1 1 3 3 2 3 2 2 1 1 1 2 1 2 2 2 2 1 2 3 3 2 1 2 3 3 3 3 2 1 2 1 2 2
 [473] 3 2 2 1 2 1 1 1 2 1 1 2 1 2 1 2 1 2 1 1 2 1 1 2 2 2 2 2 1 2 2 2 2 2 1 1 1 1 3 3 2 2 1 3 1 1 1 1 3 1 3 3 1 1 1 1 2 2 1 3
 [532] 2 3 2 3 3 3 1 1 2 2 2 1 1 2 1 1 1 2 3 3 3 3 3 3 3 3 3 2 2 3 2 3 2 2 2 1 1 3 3 1 2 3 3 3 2 2 3 2 3 2 3 2 2 3 2 3 2 3 3 2 1 2
 [591] 3 2 3 2 1 3 2 3 3 2 3 2 2 3 3 3 1 3 1 1 2 1 1 3 3 1 2 1 1 1 1 3 2 2 3 2 2 2 3 1 3 2 1 3 1 1 1 1 3 2 1 3 2 1 1 3 1 1 3
 [650] 2 1 3 3 1 2 1 2 2 2 2 2 3 2 2 3 2 1 2 1 2 1 2 2 1 2 2 1 2 2 2 3 1 2 1 1 1 2 2 1 1 1 3 2 1 3 3 3 1 2 2 1 2 2 1 1 1 1 3
 [709] 1 3 3 1 3 1 1 3 3 3 3 3 3 1 1 1 1 2 1 2 1 2 1 1 1 1 1 1 1 2 1 1 1 3 2 2 1 2 2 1 1 1 2 1 2 2 1 1 1 2 1 1 1 2 2 1 2 1 2 1 2
 [768] 1 2 2 2 3 1 1 1 1 1 1 1 2 1 1 1 3 3 1 2 1 2 2 2 2 2 1 1 1 1 1 1 1 2 1 1 1 3 1 1 3 2 2 2 1 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [827] 3 3 3 3 1 3 3 3 3 3 1 3 2 3 1 2 1 1 1 1 2 2 2 1 2 1 1 1 1 2 1 1 2 2 2 2 1 2 1 2 1 2 2 1 1 2 1 2 2 2 1 1 2 2 2 1 1 1 1 1
 [886] 2 2 2 2 2 2 2 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [945] 2 1 2 1 2 1 1 1 1 2 1 1 1 1 3 1 1 2 3 1 3 1 3 3 1 1 2 1 2 1 2 2 1 2 1 1 2 2 2 2 1 1 2 1 1 1 2 2 2 1 2 1 1 1 2 2
 [ reached getOption("max.print") -- omitted 3177 entries ]

Within cluster sum of squares by cluster:
[1] 485.8646 343.7936 171.3099
 (between_SS / total_SS =  76.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
```

Figure 15A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> clusplot(aba3, kc$cluster, color=TRUE, shade = TRUE, labels=2, line=0)
>
```

Figure 15B

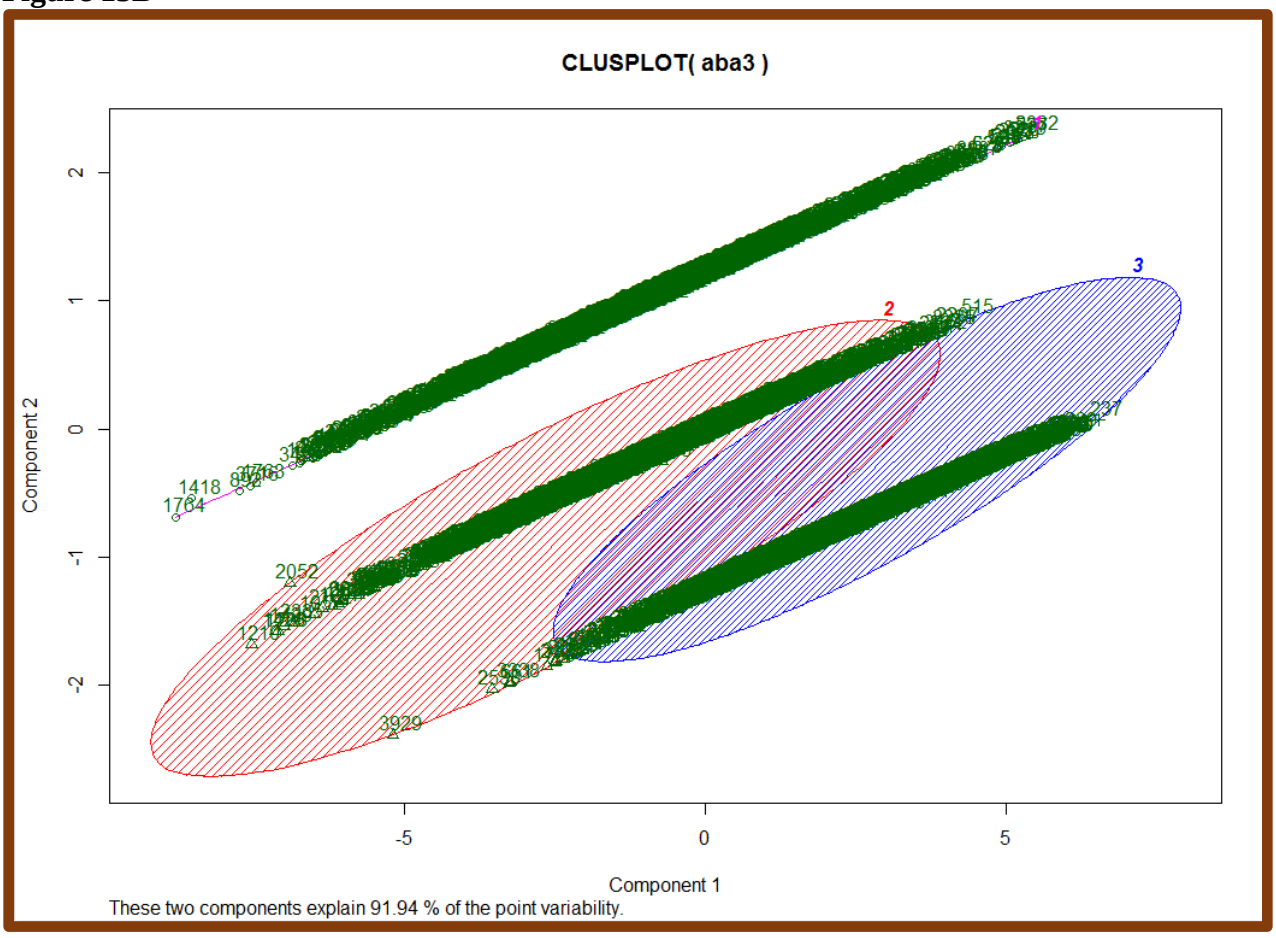


Figure 16A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> plot(aba3)
> plot(aba3, col =(kc$cluster) , main="K-Means result with 3 clusters", pch=20, cex=2)
>
```

Figure 16B

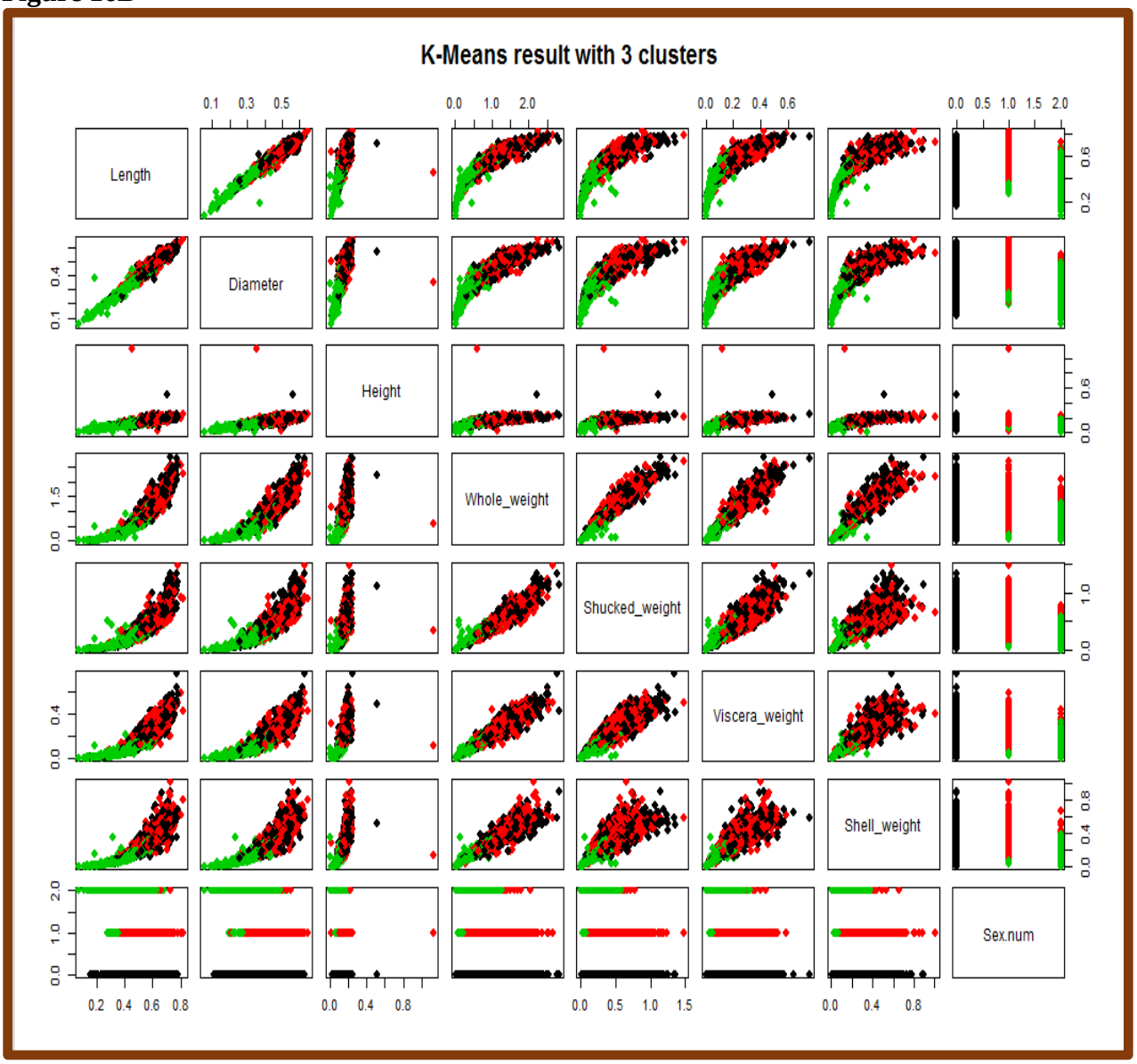


Figure 17A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ## plotcluster function from fpc package to draw discriminant projection plot
> library("fpc", lib.loc="~/R/win-library/3.3")
>
> plotcluster(aba3, kc$cluster, main="Discriminant Projection Plot using 3 clusters")
>
```

Figure 17B

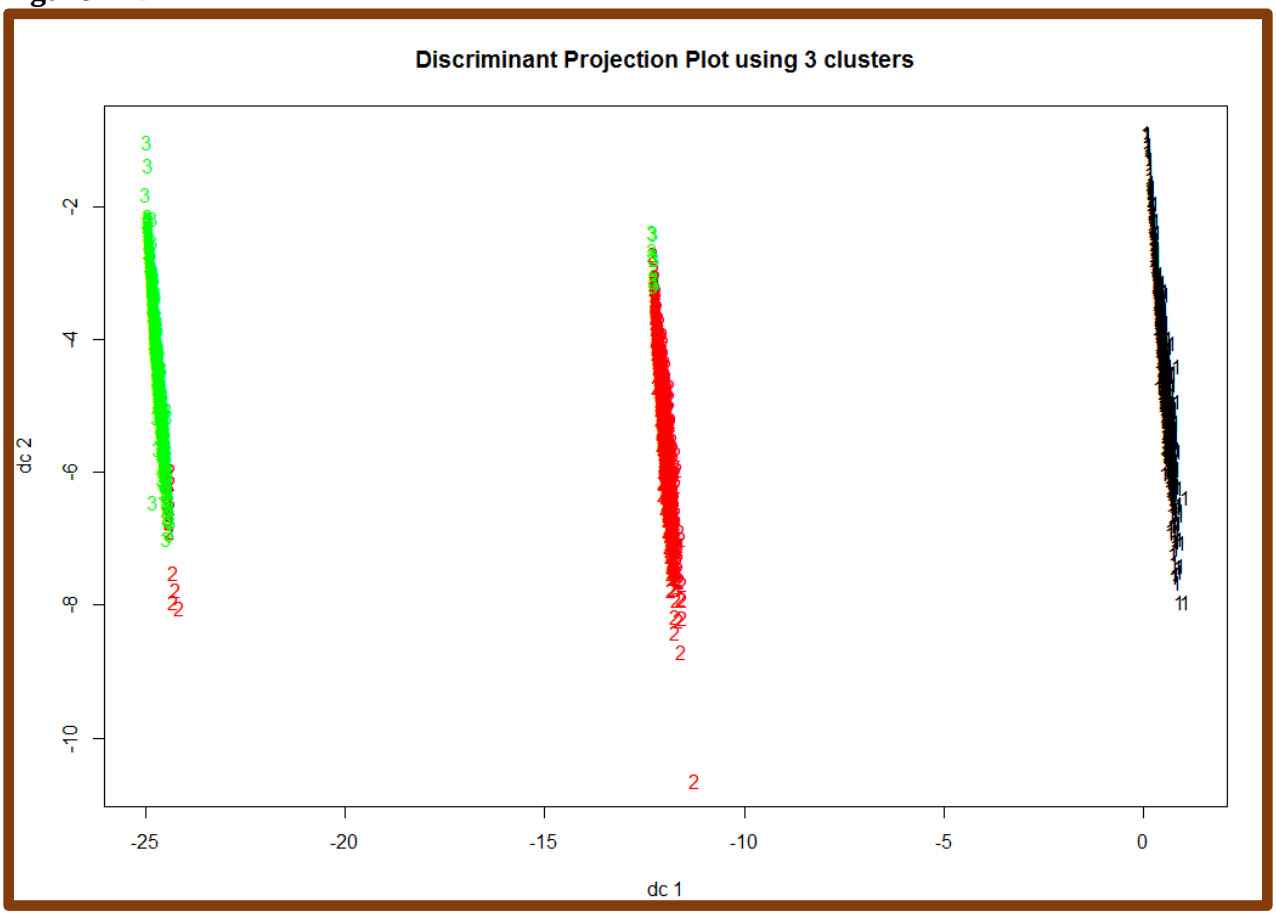


Figure 18A

```
>
> library("MASS", lib.loc="C:/Program Files/R/R-3.3.2/library")
> parcoord(aba3, kc$cluster)
>
```

Figure 18B

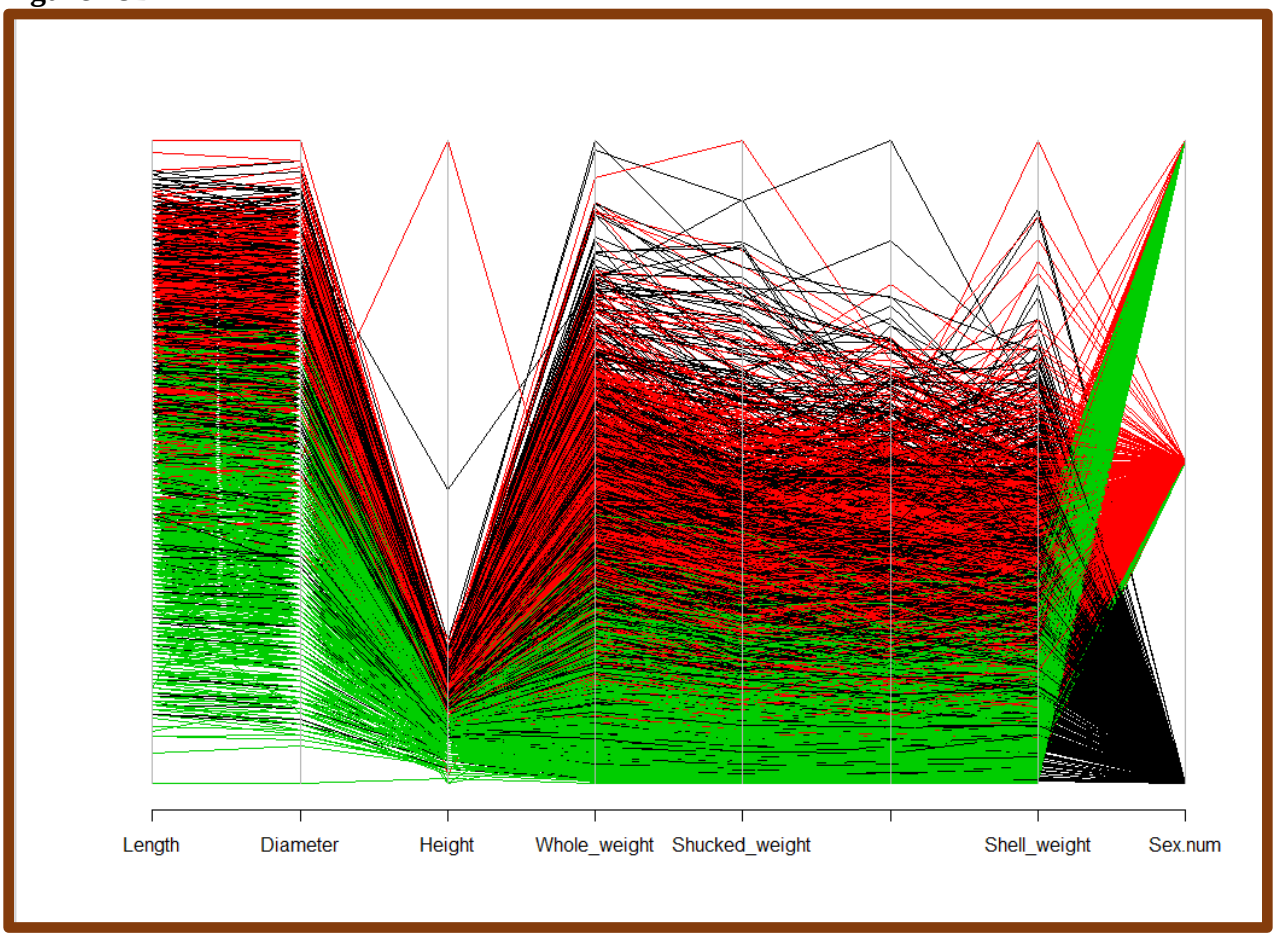


Figure 19A

```
> kc$size
[1] 1528 1304 1345
>
> kc$betweenss
[1] 3310.498
>
> kc$tot.withinss
[1] 1000.968
>
> kc$withinss
[1] 485.8646 343.7936 171.3099
>
> kc$iter
[1] 2
>
```

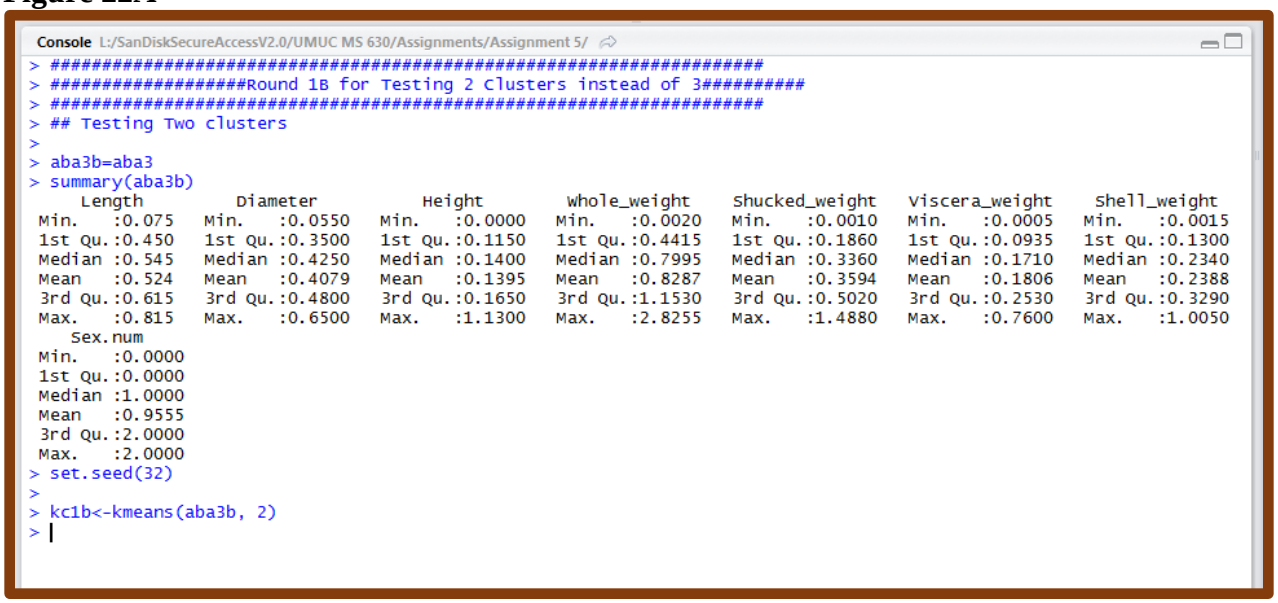
Figure 20A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ##Building the Clustering Evaluation Cross Validation
> table(abalone$Rings)

  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23
  1   1  15  57 115 259 391 568 689 634 487 267 203 126 103  67  58  42  32  26  14   6   9
 24  25  26  27  29
  2   1   1   2   1
>
> confuseTable.kc <- table(abalone$Rings, kc$cluster)
> confuseTable.kc

      1   2   3
1     0   0   1
2     0   0   1
3     3   0  12
4     6   0  51
5    11   1 103
6    27  13 219
7    80  40 271
8   172 121 275
9   278 237 174
10 294 247  93
11 225 202  60
12 118 129  20
13  91  89  23
14  56  56  14
15  52  42   9
16  30  32   5
17  25  27   6
18  18  21   3
19  15  15   2
20  12  12   2
21   6   7   1
22   3   3   0
23   3   6   0
24   1   1   0
25   0   1   0
26   1   0   0
27   1   1   0
29   0   1   0
>
```

Figure 21A

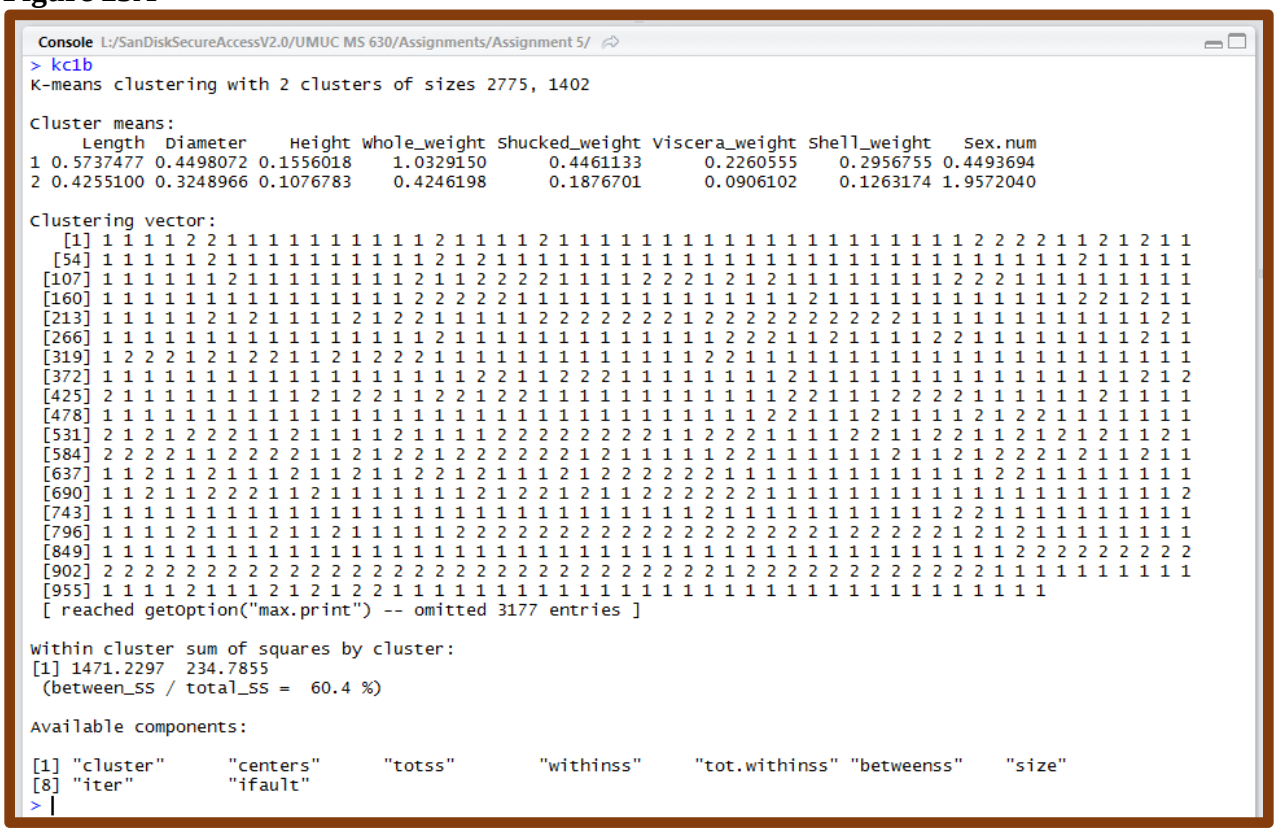
```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> library("stats4", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("modeltools", lib.loc="~/R/win-library/3.3")
> library("grid", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("flexclust", lib.loc="~/R/win-library/3.3")
> library("cluster", lib.loc="~/R/win-library/3.3")
>
> ## Adjusted Rand Index
> randIndex(confuseTable.kc)
       ARI
0.04001484
>
```

Figure 22A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ######################################################################
> ####################Round 1B for Testing 2 Clusters instead of 3##########
> ######################################################################
> ## Testing Two clusters
>
> aba3b=aba3
> summary(aba3b)
     Length          Diameter          Height        Whole_weight    Shucked_weight   Viscera_weight    Shell_weight
 Min.   :0.075   Min.   :0.0550   Min.   :0.0000   Min.   :0.0020   Min.   :0.0010   Min.   :0.0005   Min.   :0.0015
 1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150   1st Qu.:0.4415   1st Qu.:0.1860   1st Qu.:0.0935   1st Qu.:0.1300
 Median :0.545   Median :0.4250   Median :0.1400   Median :0.7995   Median :0.3360   Median :0.1710   Median :0.2340
 Mean   :0.524   Mean   :0.4079   Mean   :0.1395   Mean   :0.8287   Mean   :0.3594   Mean   :0.1806   Mean   :0.2388
 3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650   3rd Qu.:1.1530   3rd Qu.:0.5020   3rd Qu.:0.2530   3rd Qu.:0.3290
 Max.   :0.815   Max.   :0.6500   Max.   :1.1300   Max.   :2.8255   Max.   :1.4880   Max.   :0.7600   Max.   :1.0050
    Sex.num
 Min.   :0.0000
 1st Qu.:0.0000
 Median :1.0000
 Mean   :0.9555
 3rd Qu.:2.0000
 Max.   :2.0000
> set.seed(32)
>
> kc1b<-kmeans(aba3b, 2)
>
```

Figure 23A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> kc1b
K-means clustering with 2 clusters of sizes 2775, 1402

Cluster means:
     Length  Diameter    Height Whole_weight Shucked_weight Viscera_weight Shell_weight   Sex.num
1 0.5737477 0.4498072 0.1556018    1.0329150      0.4461133      0.2260555    0.2956755 0.4493694
2 0.4255100 0.3248966 0.1076783    0.4246198      0.1876701      0.0906102    0.1263174 1.9572040

Clustering vector:
   [1] 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 1 1 2 1 2 1 1
  [54] 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1
 [107] 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 2 2 2 2 1 1 1 1 2 2 2 1 2 1 2 1 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 1
 [160] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 2 1 2 1 1
 [213] 1 1 1 1 1 2 1 2 1 1 1 1 2 1 2 2 1 1 1 1 1 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 2 1
 [266] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 2 1 1 1 1 2 2 1 1 1 1 1 1 1 2 1 1
 [319] 1 2 2 2 1 2 1 2 2 1 1 2 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [372] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 2 2 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2
 [425] 2 1 1 1 1 1 1 1 1 1 2 1 2 2 1 1 2 2 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 2 2 2 1 1 1 1 1 2 1 1 1
 [478] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [531] 2 1 2 1 2 2 2 1 1 2 1 1 1 1 2 1 1 1 1 2 2 2 2 2 2 2 2 1 1 2 2 2 1 1 1 1 2 2 1 1 2 2 1 1 2 1 2 1 1 2 1
 [584] 2 2 2 2 1 1 2 2 2 2 1 1 2 1 2 2 1 2 2 2 2 2 2 2 1 2 1 1 1 1 1 2 2 1 1 1 1 2 1 1 2 1 2 2 2 1 2 1 1 2 1
 [637] 1 1 2 1 1 2 1 1 1 2 1 1 2 1 1 2 2 1 2 1 1 1 2 1 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1
 [690] 1 1 2 1 1 2 2 2 1 1 2 1 1 1 1 1 1 1 2 1 2 2 1 2 1 1 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2
 [743] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1
 [796] 1 1 1 1 2 1 1 1 2 1 1 2 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 1 2 1 2 1 1 1 1 1 1 1
 [849] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2
 [902] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1
 [955] 1 1 1 1 2 1 1 1 2 1 2 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [ reached getOption("max.print") -- omitted 3177 entries ]

Within cluster sum of squares by cluster:
[1] 1471.2297  234.7855
 (between_SS / total_SS =  60.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
>
```

Figure 24A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> #plotting or Graphing
> clusplot(aba3b, kc1b$cluster, color=TRUE, shade = TRUE, labels=2, line=0)
>
> plot(aba3b)
> plot(aba3b, col =(kc1b$cluster) , main="K-Means result with 2 clusters", pch=20, cex=2)
>
>
> ## plotcluster function from fpc package to draw discriminant projection plot
> library("fpc", lib.loc="~/R/win-library/3.3")
>
> plotcluster(aba3b, kc1b$cluster, main="Discriminant Projection Plot using 2 clusters")
>
> ##Next, we draw parallel coordinates plot to see how variables
> ##contributed in each cluster
>
> library("MASS", lib.loc="C:/Program Files/R/R-3.3.2/library")
> parcoord(aba3b, kc1b$cluster)
>
```

Figure 25A

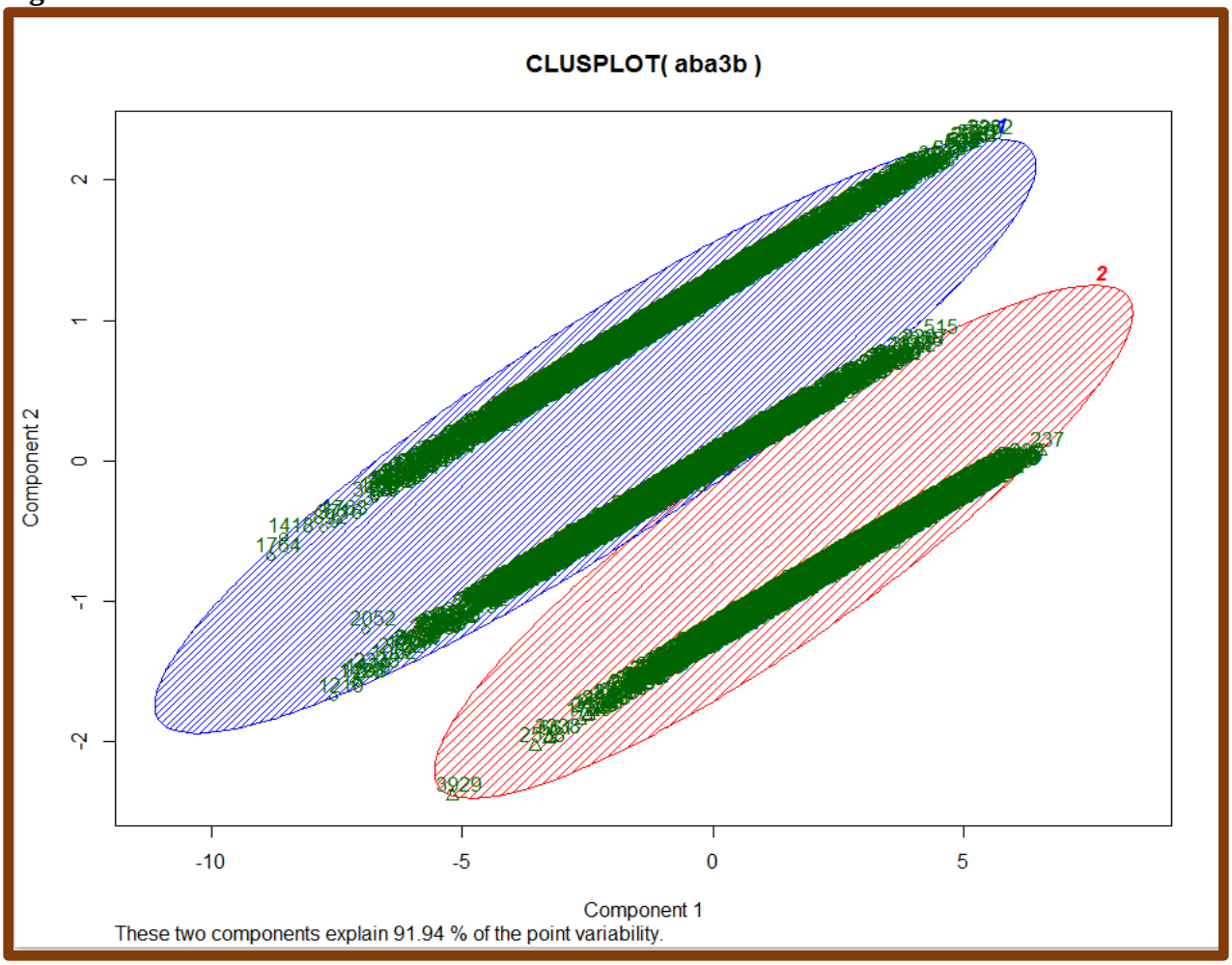


Figure 26A

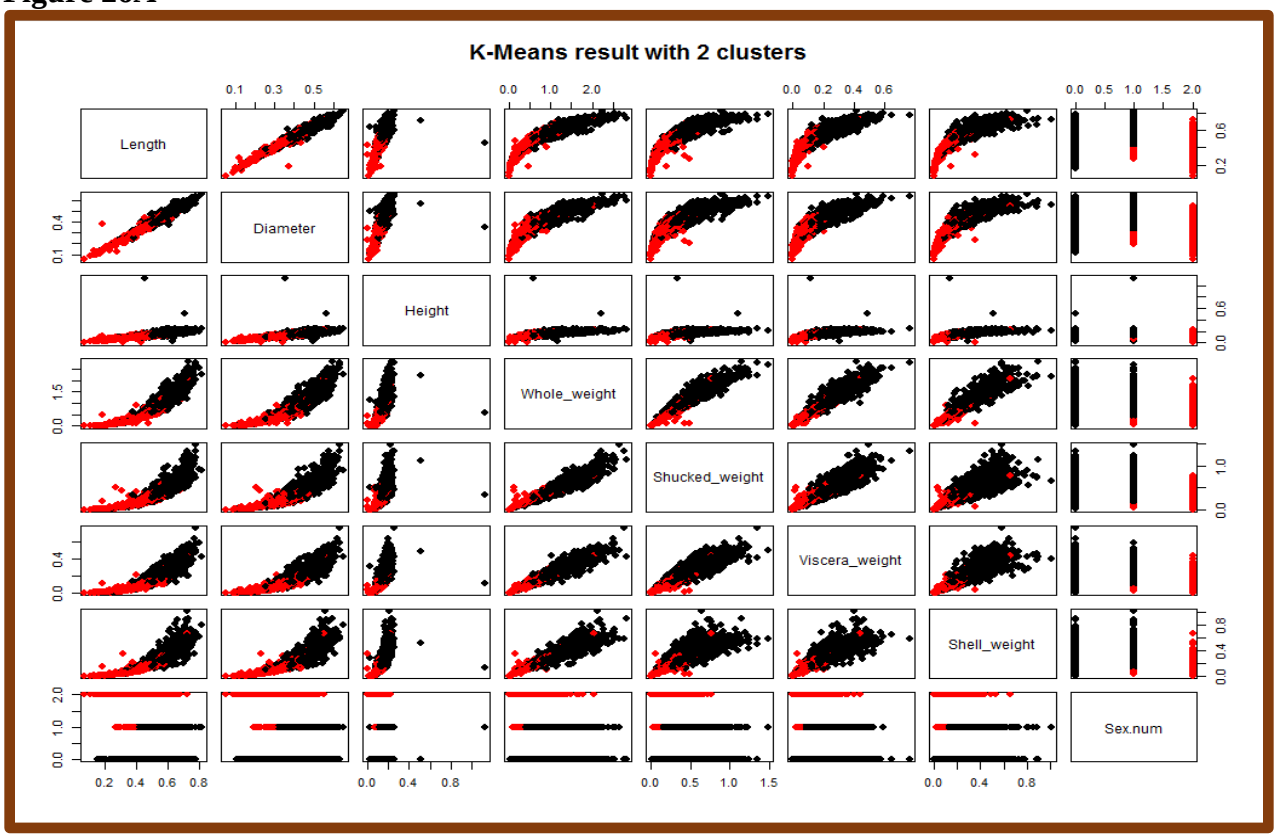


Figure 27A

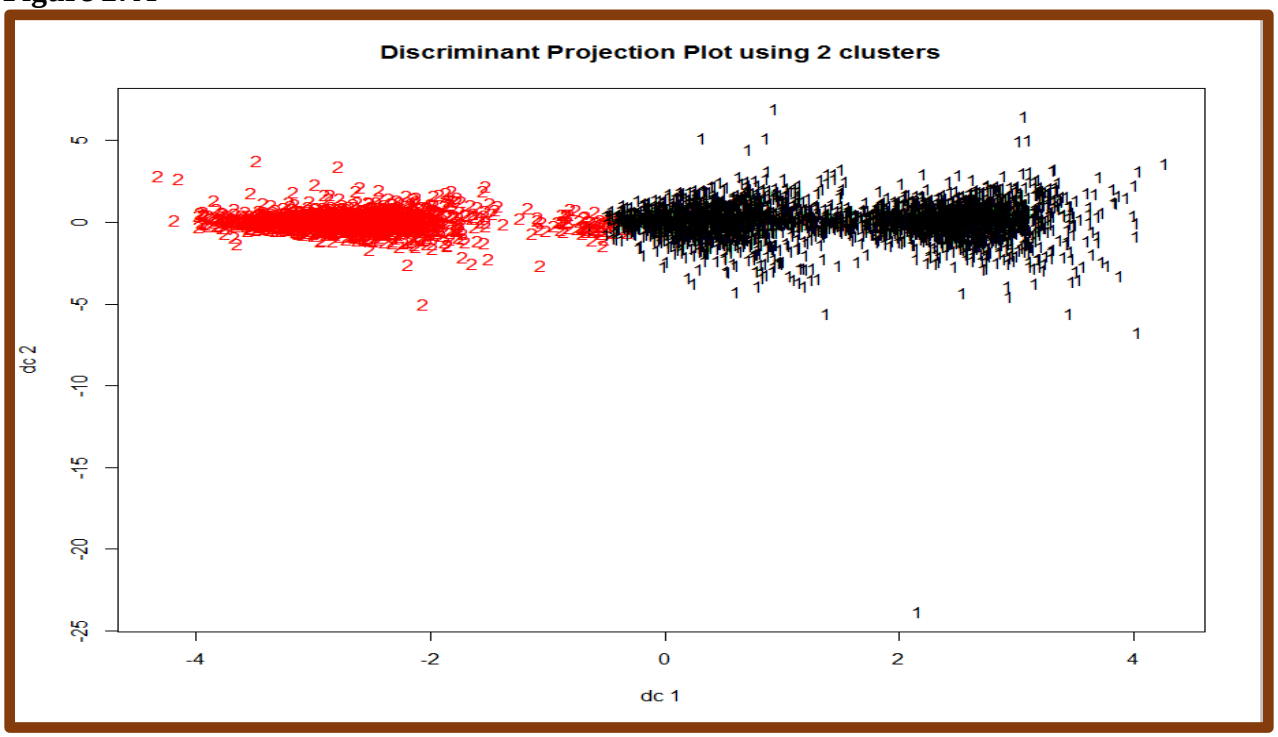


Figure 28A

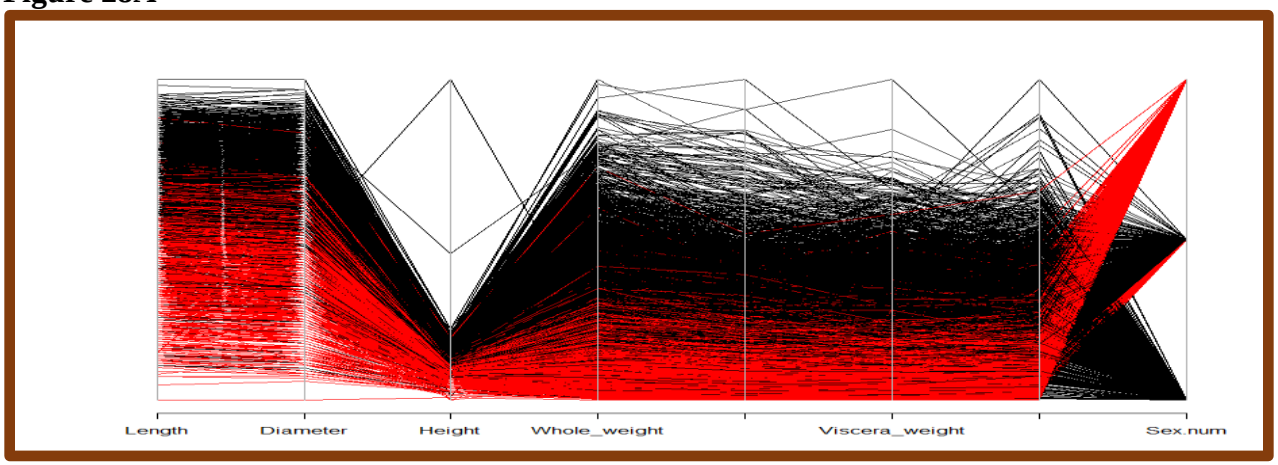


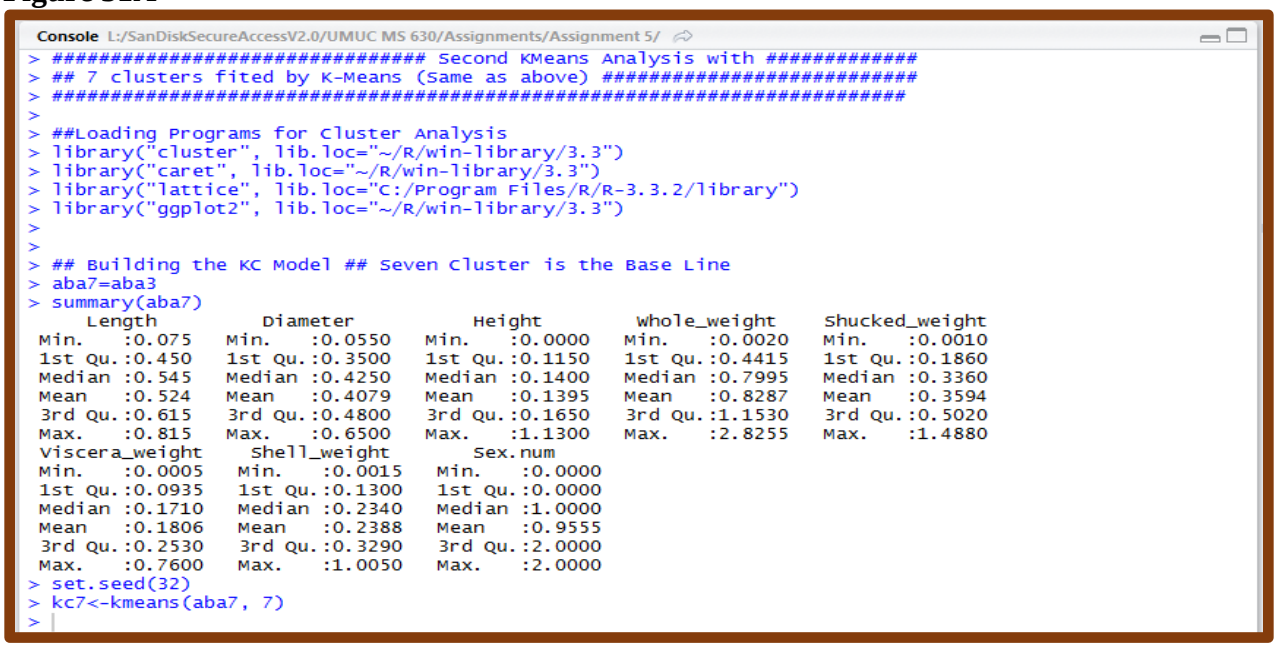
Figure 29A

```
> kc1b$size
[1] 2775 1402
>
> kc1b$betweenss
[1] 2605.451
>
> kc1b$tot.withinss
[1] 1706.015
>
> kc1b$withinss
[1] 1471.2297  234.7855
>
> kc1b$iter
[1] 1
>
>
>
> ##Building the Clustering Evaluation Cross Validation
> table(abalone$Rings, kc1b$cluster)

      1   2
  1   0   1
  2   0   1
  3   3  12
  4   6  51
  5  11 104
  6  38 221
  7 114 277
  8 285 283
  9 508 181
  10 526 108
  11 423  64
  12 243  24
  13 177  26
  14 112  14
  15  93  10
  16  59   8
  17  51   7
  18  37   5
  19  30   2
  20  24   2
  21  13   1
  22   6   0
  23   9   0
  24   2   0
  25   1   0
  26   1   0
  27   2   0
  29   1   0
>
> table(abalone$Rings)

  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  29
  1   1  15  57 115 259 391 568 689 634 487 267 203 126 103  67  58  42  32  26  14   6   9   2   1   1   2   1
>
```

Figure 30A

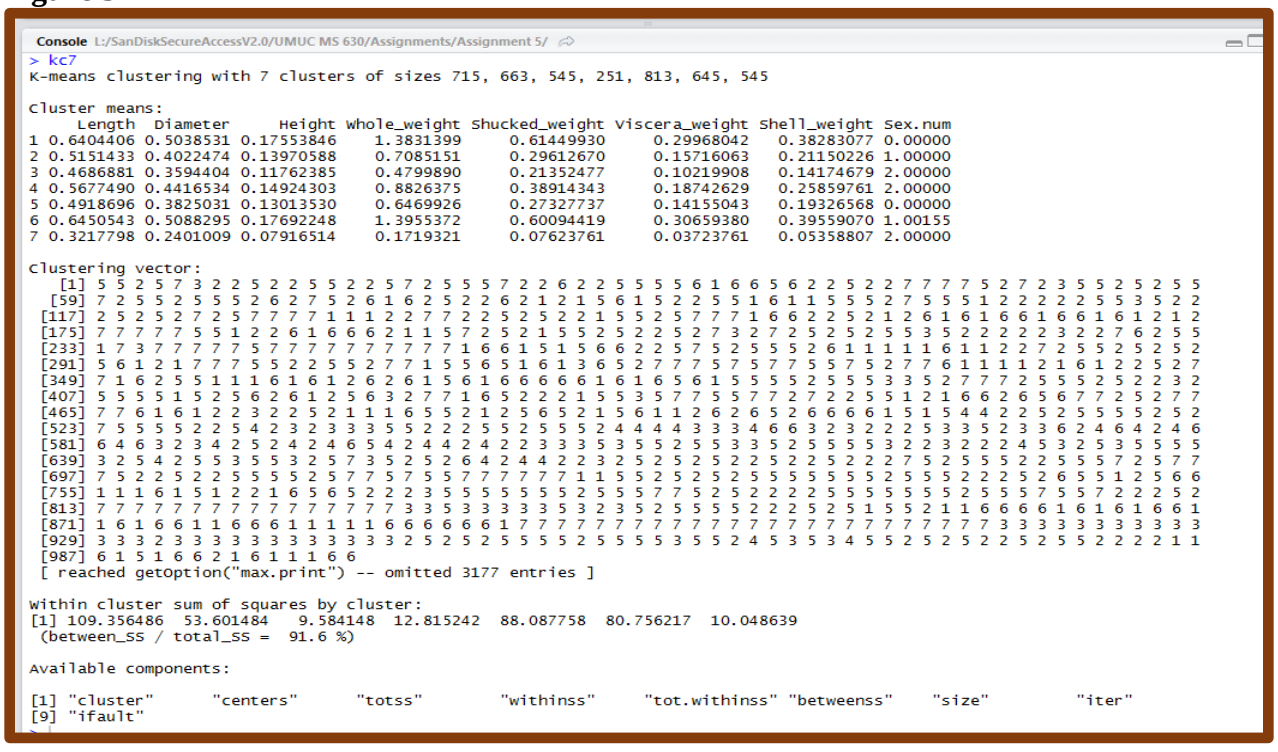
```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> library("stats4", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("modeltools", lib.loc="~/R/win-library/3.3")
> library("grid", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("flexclust", lib.loc="~/R/win-library/3.3")
> library("cluster", lib.loc="~/R/win-library/3.3")
>
> ## Adjusted Rand Index
> randIndex(confuseTable.kc1b)
        ARI
0.03965002
>
```

Figure 31A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ################################ Second KMeans Analysis with #############
> ## 7 clusters fited by K-Means (Same as above) ############################
> ########################################################################
>
> ##Loading Programs for Cluster Analysis
> library("cluster", lib.loc="~/R/win-library/3.3")
> library("caret", lib.loc="~/R/win-library/3.3")
> library("lattice", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("ggplot2", lib.loc="~/R/win-library/3.3")
>
>
> ## Building the KC Model ## Seven Cluster is the Base Line
> aba7=aba3
> summary(aba7)
     Length          Diameter          Height        Whole_weight    Shucked_weight
 Min.   :0.075   Min.   :0.0550   Min.   :0.0000   Min.   :0.0020   Min.   :0.0010
 1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150   1st Qu.:0.4415   1st Qu.:0.1860
 Median :0.545   Median :0.4250   Median :0.1400   Median :0.7995   Median :0.3360
 Mean   :0.524   Mean   :0.4079   Mean   :0.1395   Mean   :0.8287   Mean   :0.3594
 3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650   3rd Qu.:1.1530   3rd Qu.:0.5020
 Max.   :0.815   Max.   :0.6500   Max.   :1.1300   Max.   :2.8255   Max.   :1.4880
 Viscera_weight    Shell_weight       Sex.num
 Min.   :0.0005   Min.   :0.0015   Min.   :0.0000
 1st Qu.:0.0935   1st Qu.:0.1300   1st Qu.:0.0000
 Median :0.1710   Median :0.2340   Median :1.0000
 Mean   :0.1806   Mean   :0.2388   Mean   :0.9555
 3rd Qu.:0.2530   3rd Qu.:0.3290   3rd Qu.:2.0000
 Max.   :0.7600   Max.   :1.0050   Max.   :2.0000
> set.seed(32)
> kc7<-kmeans(aba7, 7)
>
```

Figure 32A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> kc7
K-means clustering with 7 clusters of sizes 715, 663, 545, 251, 813, 645, 545

Cluster means:
     Length  Diameter      Height Whole_weight Shucked_weight Viscera_weight Shell_weight Sex.num
1 0.6404406 0.5038531 0.17553846    1.3831399     0.61449930     0.29968042   0.38283077 0.00000
2 0.5151433 0.4022474 0.13970588    0.7085151     0.29612670     0.15716063   0.21150226 1.00000
3 0.4686881 0.3594404 0.11762385    0.4799890     0.21352477     0.10219908   0.14174679 2.00000
4 0.5677490 0.4416534 0.14924303    0.8826375     0.38914343     0.18742629   0.25859761 2.00000
5 0.4918696 0.3825031 0.13013530    0.6469926     0.27327737     0.14155043   0.19326568 0.00000
6 0.6450543 0.5088295 0.17692248    1.3955372     0.60094419     0.30659380   0.39559070 1.00155
7 0.3217798 0.2401009 0.07916514    0.1719321     0.07623761     0.03723761   0.05358807 2.00000

Clustering vector:
   [1] 5 5 2 5 7 3 2 2 5 2 2 5 5 2 2 5 7 2 5 5 5 7 2 2 6 2 2 5 5 5 5 6 1 6 6 5 6 2 2 5 2 2 7 7 7 7 5 2 7 2 3 5 5 2 5 2 5 5
  [59] 7 2 5 5 2 5 5 5 2 6 2 7 5 2 6 1 6 2 5 2 2 6 2 1 2 1 5 6 1 5 2 2 5 5 1 6 1 1 5 5 5 2 7 5 5 5 1 2 2 2 2 2 5 5 3 5 2 2
 [117] 2 5 2 5 2 7 2 5 7 7 7 7 1 1 1 2 2 7 7 2 2 5 2 5 2 2 1 5 5 2 5 7 7 7 1 6 6 2 2 5 2 1 2 6 1 6 1 6 6 1 6 6 1 6 1 2 1 2
 [175] 7 7 7 7 7 5 5 1 2 2 6 1 6 6 6 2 1 1 5 7 2 5 2 1 5 5 2 5 2 2 5 2 7 3 2 7 2 5 2 5 2 5 5 5 3 5 2 2 2 2 2 3 2 2 7 6 2 5 5
 [233] 1 7 3 7 7 7 7 7 5 7 7 7 7 7 7 7 7 7 7 1 6 6 1 5 1 5 6 6 2 2 5 7 5 2 5 5 5 2 6 1 1 1 1 1 6 1 1 2 2 7 2 5 5 2 5 2 5 2
 [291] 5 6 1 2 1 7 7 7 5 5 2 2 5 5 2 7 7 1 5 5 6 5 1 6 1 3 6 5 2 7 7 7 5 7 5 7 7 5 5 7 5 2 7 7 6 1 1 1 1 2 1 6 1 2 2 5 2 7
 [349] 7 1 6 2 5 5 1 1 1 6 1 6 1 2 6 2 6 1 5 6 1 6 6 6 6 6 1 6 1 6 5 6 1 5 5 5 5 2 5 5 5 3 3 5 2 7 7 7 2 5 5 5 2 5 2 2 3 2
 [407] 5 5 5 5 1 5 2 5 6 2 6 1 2 5 6 3 2 7 7 1 6 5 2 2 2 1 5 5 3 5 7 7 5 5 7 7 2 7 2 2 5 5 1 2 1 6 6 2 6 5 6 7 7 2 5 2 7 7
 [465] 7 7 6 1 6 1 2 2 3 2 2 5 2 1 1 1 6 5 5 2 1 2 5 6 5 2 1 5 6 1 1 2 6 2 6 5 2 6 6 6 6 1 5 1 5 4 4 2 2 5 2 5 5 5 5 2 5 2
 [523] 7 5 5 5 5 2 2 5 4 2 3 2 3 3 3 5 5 2 2 2 5 5 2 5 5 5 2 4 4 4 4 3 3 3 4 6 6 3 2 3 2 2 2 5 3 3 5 2 3 3 6 2 4 6 4 2 4 6
 [581] 6 4 6 3 2 3 4 2 5 2 4 2 4 6 5 4 2 4 4 2 4 2 2 3 3 3 5 3 5 5 2 5 5 5 3 3 5 2 5 5 5 5 5 3 2 2 3 2 2 2 4 5 3 2 5 3 5 5 5 5
 [639] 3 2 5 4 2 5 5 3 5 5 3 2 5 7 3 5 2 5 2 6 4 2 4 4 2 2 3 2 5 2 5 2 5 2 2 5 2 2 5 2 2 2 2 7 5 2 5 5 5 2 2 5 5 5 7 2 5 7 7
 [697] 7 5 2 2 5 2 2 5 5 5 5 2 5 7 7 5 7 5 5 5 7 7 7 7 7 7 1 1 5 5 2 5 2 5 2 5 5 5 5 5 5 5 5 2 5 5 5 2 2 2 5 2 6 5 5 1 2 5 6 6
 [755] 1 1 1 6 1 5 1 2 2 1 6 5 6 5 2 2 2 3 5 5 5 5 5 5 5 2 5 5 5 7 7 5 2 5 2 2 2 2 5 5 5 5 5 5 5 2 5 5 5 5 7 5 5 7 2 2 2 5 2
 [813] 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 3 3 5 3 3 3 3 3 3 5 3 2 3 5 2 5 5 5 5 2 2 2 5 2 5 1 5 5 2 1 1 6 6 6 6 1 6 1 6 1 6 6 1
 [871] 1 6 1 6 6 1 1 6 6 6 1 1 1 1 1 6 6 6 6 6 6 1 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 3 3 3 3 3 3 3 3 3 3 3 3 3
 [929] 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 2 5 2 5 2 5 5 5 5 2 5 5 5 5 5 3 5 5 2 4 5 3 5 3 4 5 5 2 5 2 5 2 2 5 2 5 5 2 2 2 2 1 1
 [987] 6 1 5 1 6 6 2 1 6 1 1 1 6 6
 [ reached getOption("max.print") -- omitted 3177 entries ]

Within cluster sum of squares by cluster:
[1] 109.356486  53.601484   9.584148  12.815242  88.087758  80.756217  10.048639
 (between_SS / total_SS =  91.6 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
```

Figure 33A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> #plotting or Graphing
> clusplot(aba7, kc7$cluster, color=TRUE, shade = TRUE, labels=2, line=0)
>
> plot(aba7)
> plot(aba7, col =(kc7$cluster) , main="K-Means result with 7 clusters", pch=20, cex=2)
>
>
> ## plotcluster function from fpc package to draw discriminant projection plot
> library("fpc", lib.loc="~/R/win-library/3.3")
>
> plotcluster(aba7, kc7$cluster, main="Discriminant Projection Plot using 7 clusters")
>
>
> library("MASS", lib.loc="C:/Program Files/R/R-3.3.2/library")
> parcoord(aba3, kc7$cluster)
>
```

Figure 34A

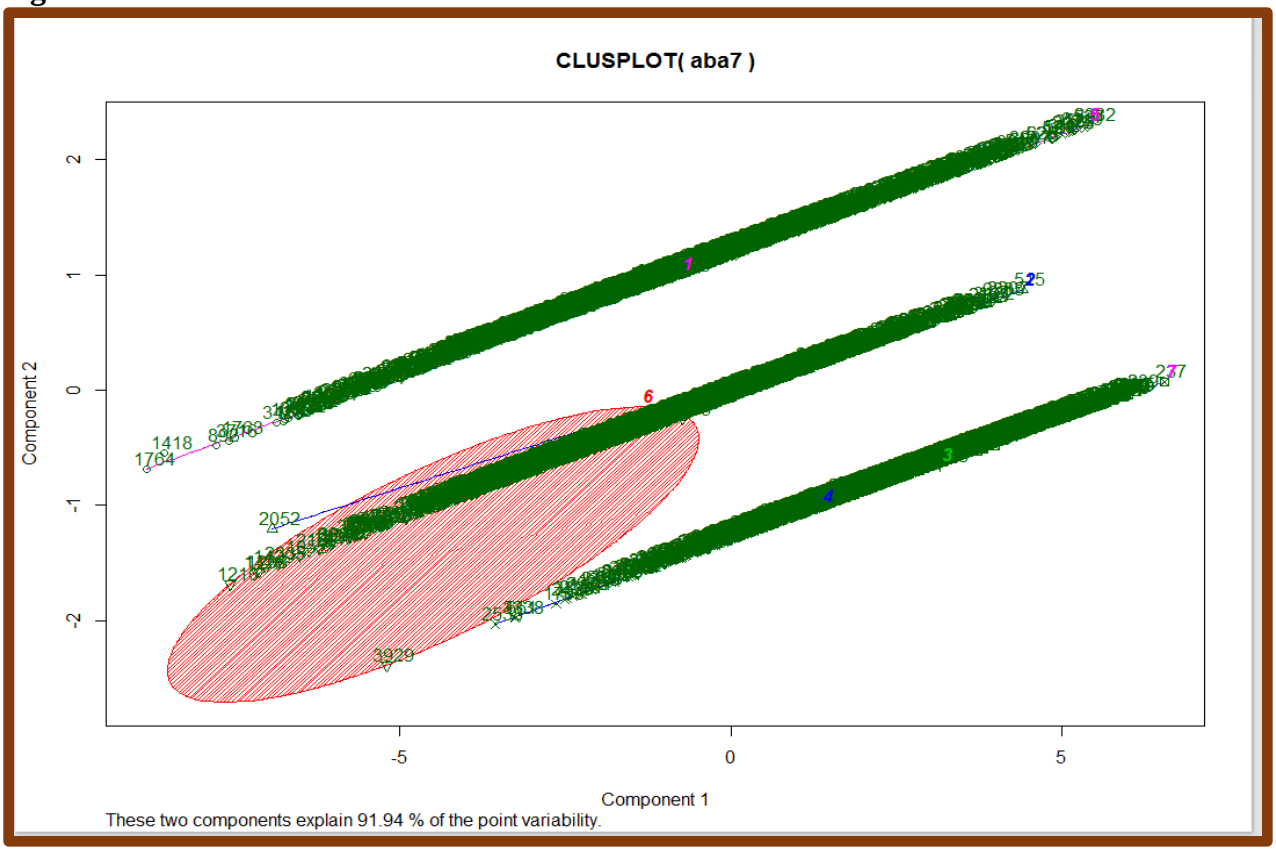


Figure 35A

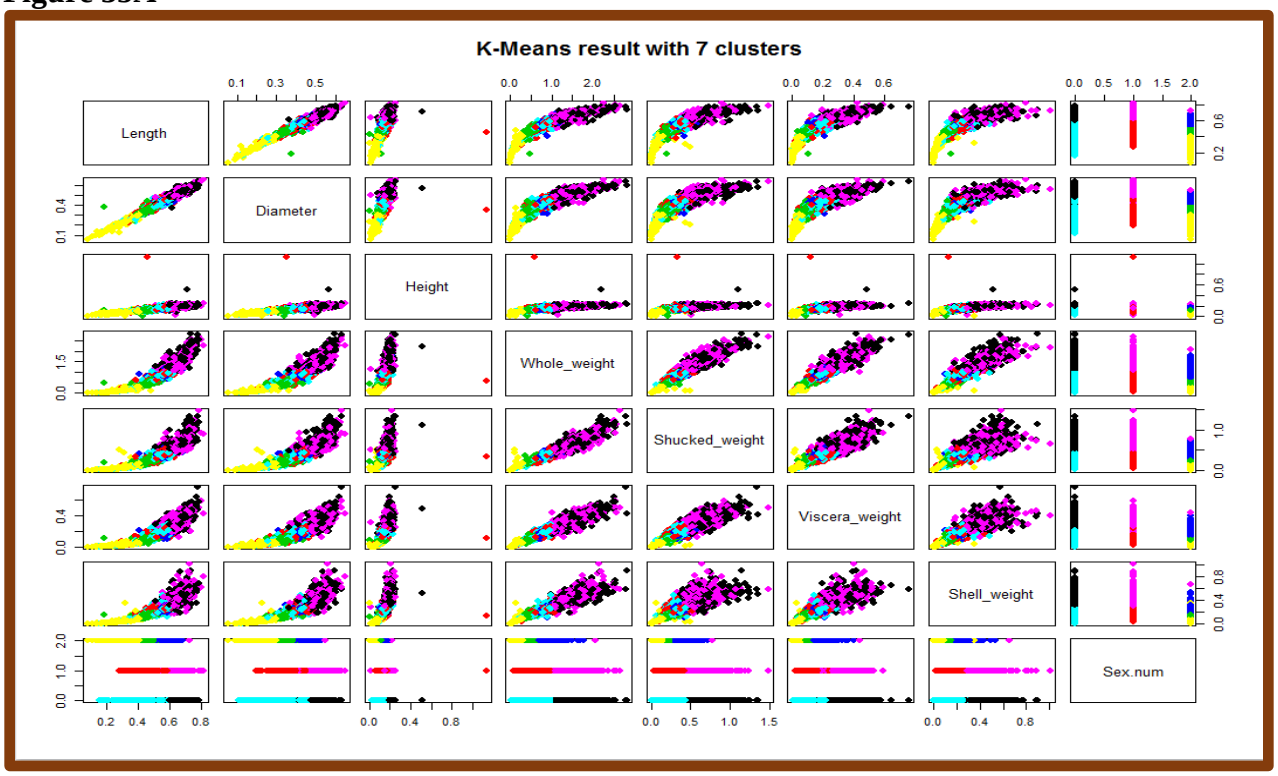


Figure 36A

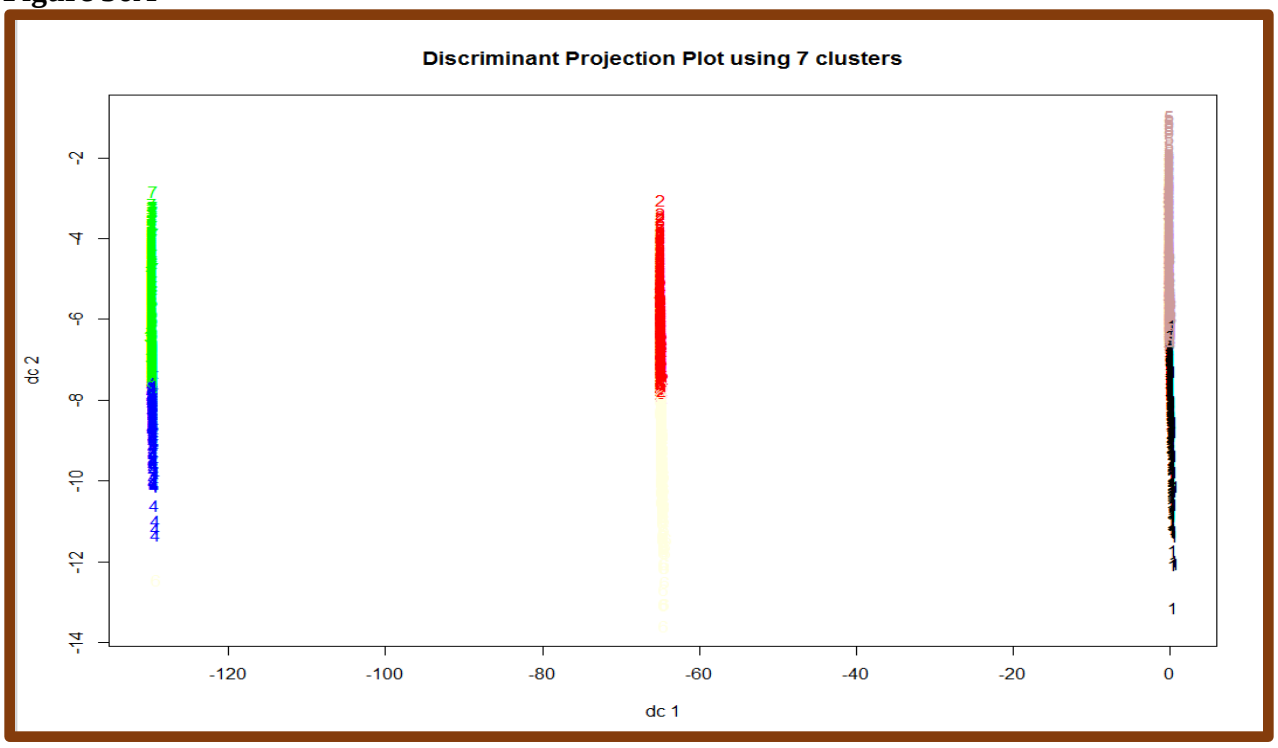


Figure37A

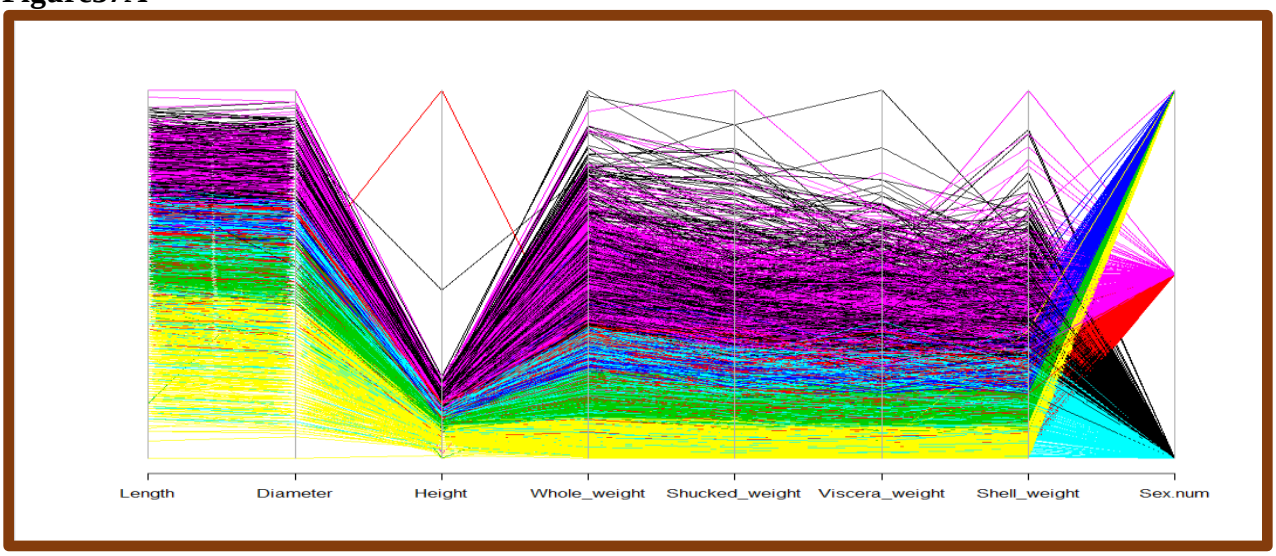


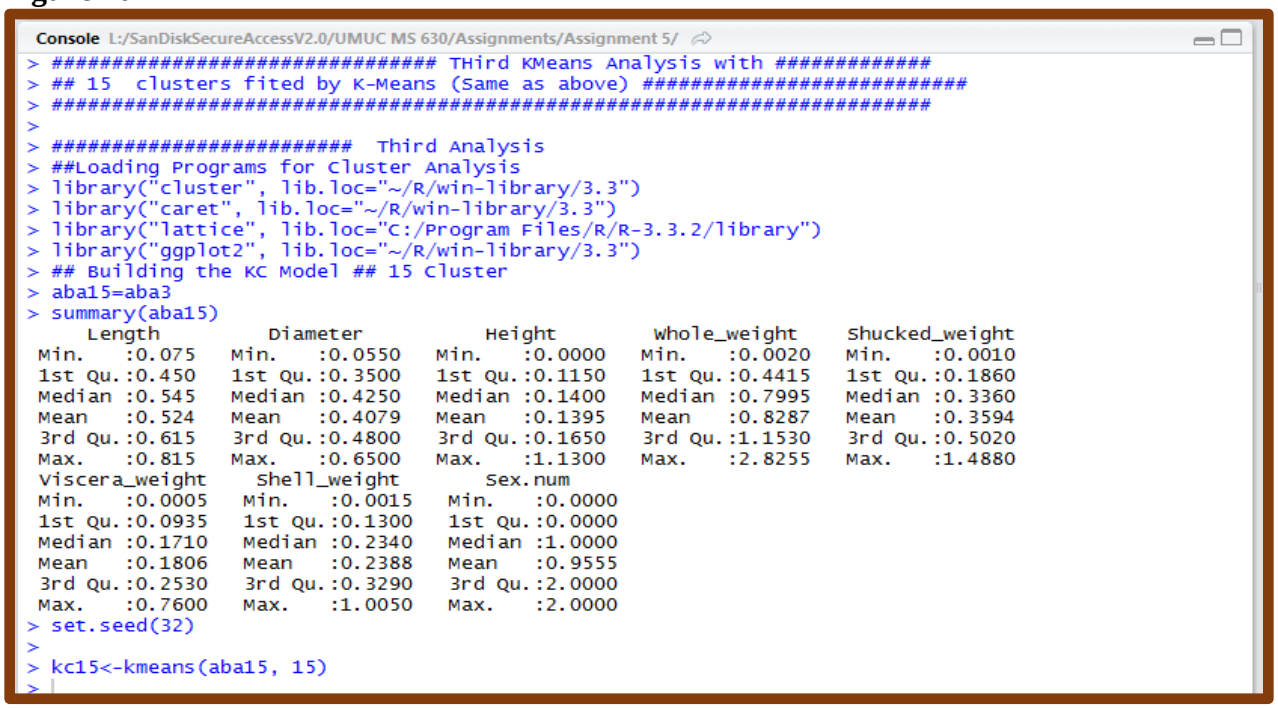
Figure 38A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ## Assignment 7 Additional parameters requested by Dr. Firdu Bati
> kc7$size
[1] 715 663 545 251 813 645 545
>
> kc7$betweenss
[1] 3947.216
>
> kc7$tot.withinss
[1] 364.25
>
> kc7$withinss
[1] 109.356486  53.601484   9.584148  12.815242  88.087758  80.756217  10.048639
>
> kc7$iter
[1] 2
>
>
>
> ##Building the Clustering Evaluation Cross Validation
> table(abalone$Rings, kc7$cluster)

      1   2   3   4   5   6   7
  1   0   0   0   0   0   0   1
  2   0   0   0   0   0   0   1
  3   0   0   0   0   3   0  12
  4   0   0   0   0   6   0  51
  5   0   4   2   0  11   0  98
  6   0  14  55   2  27   2 159
  7   4  41 130   2  76   3 135
  8  37  95 178  41 135  27  55
  9 117 134  94  62 161 104  17
 10 168 110  24  58 126 139   9
 11 148  67  24  34  77 133   4
 12  66  58  10  10  52  70   1
 13  51  41  12  11  40  47   1
 14  26  24   4   9  30  32   1
 15  26  22   6   4  26  19   0
 16  17  15   3   4  13  15   0
 17  16  13   0   7   9  13   0
 18  12   8   2   3   6  11   0
 19   7   7   1   1   8   8   0
 20   8   2   0   2   4  10   0
 21   5   3   0   1   1   4   0
 22   2   2   0   0   1   1   0
 23   2   3   0   0   1   3   0
 24   1   0   0   0   0   1   0
 25   0   0   0   0   0   1   0
 26   1   0   0   0   0   0   0
 27   1   0   0   0   0   1   0
 29   0   0   0   0   0   1   0
>
> table(abalone$Rings)

  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  29
  1   1  15  57 115 259 391 568 689 634 487 267 203 126 103  67  58  42  32  26  14   6   9   2   1   1   2   1
```

Figure 39A

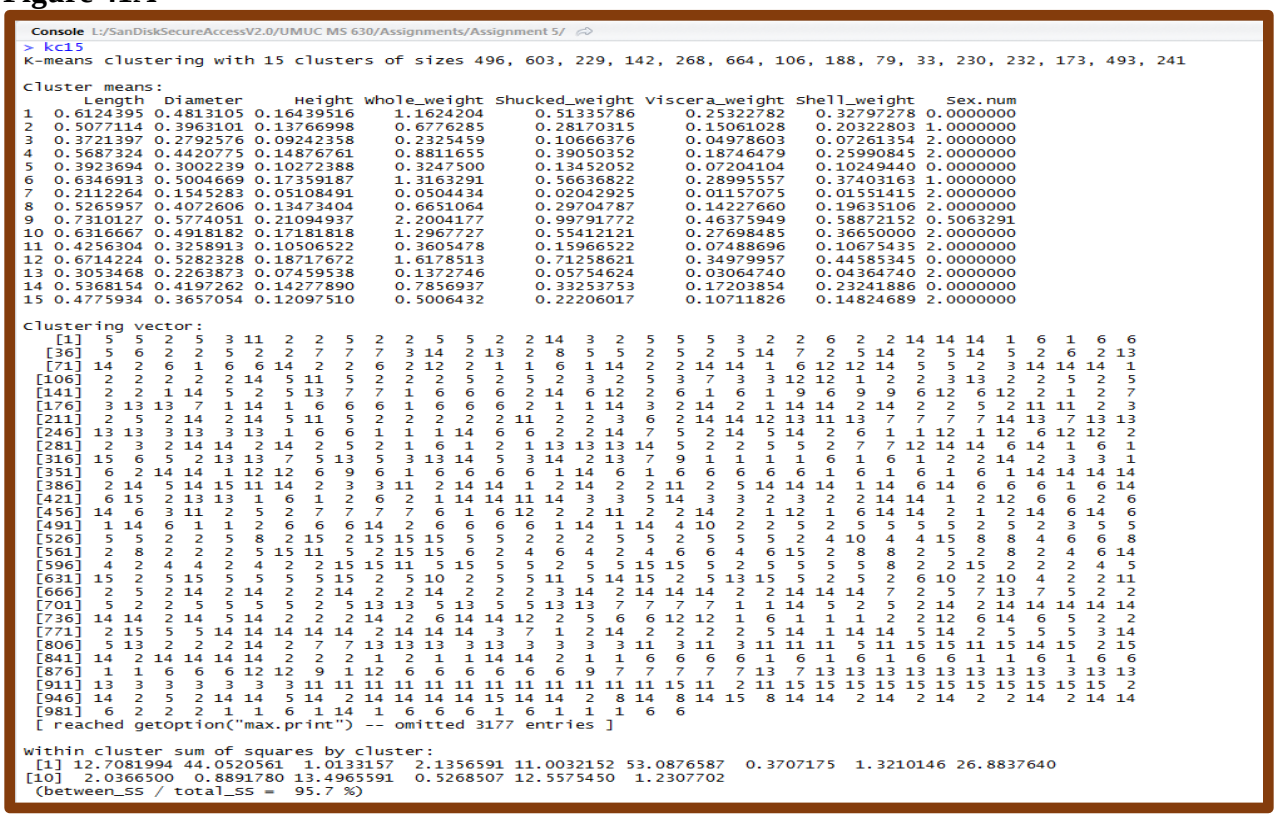
```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> library("stats4", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("modeltools", lib.loc="~/R/win-library/3.3")
> library("grid", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("flexclust", lib.loc="~/R/win-library/3.3")
>
> ## Adjusted Rand Index
> randIndex(confuseTable.kc7)
       ARI
0.05628814
>
```

Figure 40A



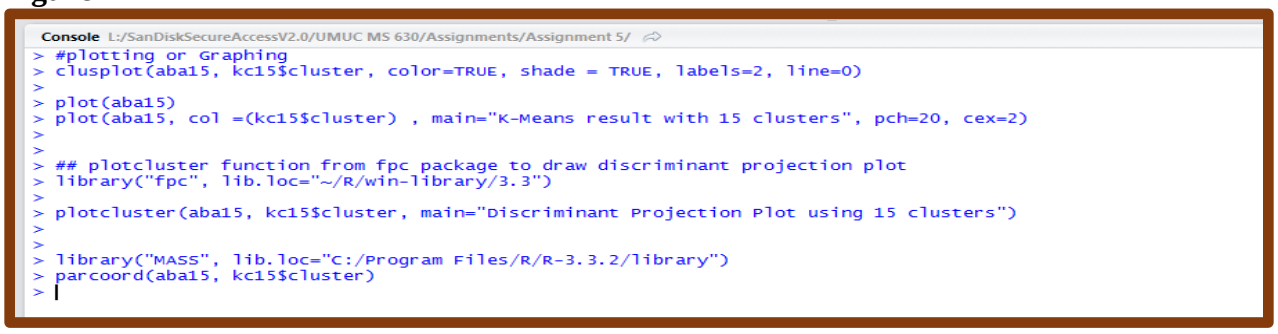
```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ################################ THird KMeans Analysis with #############
> ## 15  clusters fited by K-Means (Same as above) ##########################
> ########################################################################
>
> ########################  Third Analysis
> ##Loading Programs for Cluster Analysis
> library("cluster", lib.loc="~/R/win-library/3.3")
> library("caret", lib.loc="~/R/win-library/3.3")
> library("lattice", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("ggplot2", lib.loc="~/R/win-library/3.3")
> ## Building the KC Model ## 15 Cluster
> aba15=aba3
> summary(aba15)
     Length          Diameter          Height        Whole_weight    Shucked_weight
 Min.   :0.075   Min.   :0.0550   Min.   :0.0000   Min.   :0.0020   Min.   :0.0010
 1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150   1st Qu.:0.4415   1st Qu.:0.1860
 Median :0.545   Median :0.4250   Median :0.1400   Median :0.7995   Median :0.3360
 Mean   :0.524   Mean   :0.4079   Mean   :0.1395   Mean   :0.8287   Mean   :0.3594
 3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650   3rd Qu.:1.1530   3rd Qu.:0.5020
 Max.   :0.815   Max.   :0.6500   Max.   :1.1300   Max.   :2.8255   Max.   :1.4880
 Viscera_weight    Shell_weight       Sex.num
 Min.   :0.0005   Min.   :0.0015   Min.   :0.0000
 1st Qu.:0.0935   1st Qu.:0.1300   1st Qu.:0.0000
 Median :0.1710   Median :0.2340   Median :1.0000
 Mean   :0.1806   Mean   :0.2388   Mean   :0.9555
 3rd Qu.:0.2530   3rd Qu.:0.3290   3rd Qu.:2.0000
 Max.   :0.7600   Max.   :1.0050   Max.   :2.0000
> set.seed(32)
>
> kc15<-kmeans(aba15, 15)
>
```

Figure 41A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> kc15
K-means clustering with 15 clusters of sizes 496, 603, 229, 142, 268, 664, 106, 188, 79, 33, 230, 232, 173, 493, 241

Cluster means:
      Length  Diameter      Height Whole_weight Shucked_weight Viscera_weight Shell_weight    Sex.num
1  0.6124395 0.4813105 0.16439516    1.1624204      0.51335786     0.25322782   0.32797278 0.0000000
2  0.5077114 0.3963101 0.13766998    0.6776285      0.28170315     0.15061028   0.20322803 1.0000000
3  0.3721397 0.2792576 0.09242358    0.2325459      0.10666376     0.04978603   0.07261354 2.0000000
4  0.5687324 0.4420775 0.14876761    0.8811655      0.39050352     0.18746479   0.25990845 2.0000000
5  0.3923694 0.3002239 0.10272388    0.3247500      0.13452052     0.07204104   0.10249440 0.0000000
6  0.6346913 0.5004669 0.17359187    1.3163291      0.56636822     0.28995557   0.37403163 1.0000000
7  0.2112264 0.1545283 0.05108491    0.0504434      0.02042925     0.01157075   0.01551415 2.0000000
8  0.5265957 0.4072606 0.13473404    0.6651064      0.29704787     0.14227660   0.19635106 2.0000000
9  0.7310127 0.5774051 0.21094937    2.2004177      0.99791772     0.46375949   0.58872152 0.5063291
10 0.6316667 0.4918182 0.17181818    1.2967727      0.55412121     0.27698485   0.36650000 2.0000000
11 0.4256304 0.3258913 0.10506522    0.3605478      0.15966522     0.07488696   0.10675435 2.0000000
12 0.6714224 0.5282328 0.18717672    1.6178513      0.71258621     0.34979957   0.44585345 0.0000000
13 0.3053468 0.2263873 0.07459538    0.1372746      0.05754624     0.03064740   0.04364740 2.0000000
14 0.5368154 0.4197262 0.14277890    0.7856937      0.33253753     0.17203854   0.23241886 0.0000000
15 0.4775934 0.3657054 0.12097510    0.5006432      0.22206017     0.10711826   0.14824689 2.0000000

Clustering vector:
   [1]  5  5  2  5  3 11  2  2  5  2  2  5  5  2  2 14  3  2  5  5  5  3  2  2  6  2  2 14 14 14  1  6  1  6  6
  [36]  5  6  2  2  5  2  2  7  7  7  3 14  2 13  2  8  5  5  2  5  2  5 14  7  2  5 14  2  5 14  5  2  6  2 13
  [71] 14  2  6  1  6  6 14  2  2  6  2 12  2  1  1  6  1 14  2  2 14 14  1  6 12 12 14  5  5  2  3 14 14 14  1
 [106]  2  2  2  2  2 14  5 11  5  2  2  2  5  2  5  2  3  2  5  3  7  3  3 12 12  1  2  2  3 13  2  2  5  2  5
 [141]  2  2  1 14  5  2  5 13  7  7  1  6  6  6  2 14  6 12  2  6  1  6  1  9  6  9  9  6 12  6 12  2  1  2  7
 [176]  3 13 13  7  1 14  1  6  6  6  1  6  6  6  2  1  1 14  3  2 14  2  1 14 14  2 14  2  2  5  2 11 11  2  3
 [211]  2  5  2 14  2 14  5 11  5  2  2  2  2  2 11  2  2  3  6  2 14 14 12 13 11 13  7  7  7  7 14 13  7 13 13
 [246] 13 13  3 13  3 13  1  6  6  1  1  1 14  6  6  2  2 14  7  5  2 14  5 14  2  6  1  1 12  1 12  6 12 12  2
 [281]  2  3  2 14 14  2 14  2  5  2  1  6  1  2  1 13 13 13 14  5  2  2  5  5  2  7  7 12 14 14  6 14  1  6  1
 [316] 15  6  5  2 13 13  7  5 13  5  3 13 14  5  3 14  2 13  7  9  1  1  1  1  6  1  6  1  2  2 14  2  3  3  1
 [351]  6  2 14 14  1 12 12  6  9  6  1  6  6  6  6  1 14  6  1  6  6  6  6  6  1  6  1  6  1  6  1 14 14 14 14
 [386]  2 14  5 14 15 11 14  2  3  3 11  2 14 14  1  2 14  2  2 11  2  5 14 14 14  1 14  6 14  6  6  6  1  6 14
 [421]  6 15  2 13 13  1  6  1  2  6  2  1 14 14 11 14  3  3  5 14  3  3  2  3  2  2 14 14  1  2 12  6  6  2  6
 [456] 14  6  3 11  2  5  2  7  7  7  7  6  1  6 12  2  2 11  2  2 14  2  1 12  1  6 14 14  2  1  2 14  6 14  6
 [491]  1 14  6  1  1  2  6  6  6 14  2  6  6  6  6  1 14  1 14  4 10  2  2  5  2  5  5  5  5  2  5  2  3  5  5
 [526]  5  5  2  2  5  8  2 15  2 15 15 15  5  5  2  2  2  5  5  2  5  5  5  2  4 10  4  4 15  8  8  4  6  6  8
 [561]  2  8  2  2  2  5 15 11  5  2 15 15  6  2  4  6  4  2  4  6  6  4  6 15  2  8  8  2  5  2  8  2  4  6 14
 [596]  4  2  4  4  2  4  2  2 15 15 11  5 15  5  5  2  5  5 15 15  5  2  5  5  5  5  8  2  2 15  2  2  2  4  5
 [631] 15  2  5 15  5  5  5  5 15  2  5 10  2  5  5 11  5 14 15  2  5 13 15  5  2  5  2  6 10  2 10  4  2  2 11
 [666]  2  5  2 14  2 14  2  2 14  2  2  2  2  3 14  2 14 14 14  2  2 14 14 14  7  2  5  7 13  7  5  2  2
 [701]  5  2  2  5  5  5  5  2  5 13 13  5 13  5  5 13 13  7  7  7  7  1  1 14  5  2  5  2 14  2 14 14 14 14 14
 [736] 14 14  2 14  5 14  2  2  2 14  2  6 14 14 12  2  5  6  6 12 12  1  6  1  1  1  2  2 12  6 14  6  5  2  2
 [771]  2 15  5  5 14 14 14 14 14  2 14 14 14  3  7  1  2 14  2  2  2  2  5 14  1 14 14  5 14  2  5  5  5  3 14
 [806]  5 13  2  2  2 14  2  7  7 13 13 13  3 13  3  3  3  3 11  3 11  3 11 11 11  5 11 15 15 11 15 14 15  2 15
 [841] 14  2 14 14 14 14  2  2  2  1  2  1  1 14 14  2  1  1  6  6  6  6  1  6  1  6  1  6  6  1  1  6  1  6  6
 [876]  1  1  6  6  6 12 12  9  1 12  6  6  6  6  6  9  7  7  7  7  7 13  7 13 13 13 13 13 13 13  3 13 13
 [911] 13  3  3  3  3  3  3 11 11 11 11 11 11 11 11 11 11 11 11 11 15 11  2 11 15 15 15 15 15 15 15 15 15 15  2
 [946] 14  2  5  2 14 14  5 14  2 14 14 14 14 15 14 14  2  8 14  8 14 15  8 14 14  2 14  2 14  2  2 14  2 14 14
 [981]  6  2  2  2  1  1  6  1 14  1  6  6  6  1  6  1  1  1  6  6
 [ reached getOption("max.print") -- omitted 3177 entries ]

Within cluster sum of squares by cluster:
 [1] 12.7081994 44.0520561  1.0133157  2.1356591 11.0032152 53.0876587  0.3707175  1.3210146 26.8837640
[10]  2.0366500  0.8891780 13.4965591  0.5268507 12.5575450  1.2307702
 (between_SS / total_SS =  95.7 %)
```

Figure 42A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> #plotting or Graphing
> clusplot(aba15, kc15$cluster, color=TRUE, shade = TRUE, labels=2, line=0)
>
> plot(aba15)
> plot(aba15, col =(kc15$cluster) , main="K-Means result with 15 clusters", pch=20, cex=2)
>
>
> ## plotcluster function from fpc package to draw discriminant projection plot
> library("fpc", lib.loc="~/R/win-library/3.3")
>
> plotcluster(aba15, kc15$cluster, main="Discriminant Projection Plot using 15 clusters")
>
>
> library("MASS", lib.loc="C:/Program Files/R/R-3.3.2/library")
> parcoord(aba15, kc15$cluster)
>
```

Figure 43A

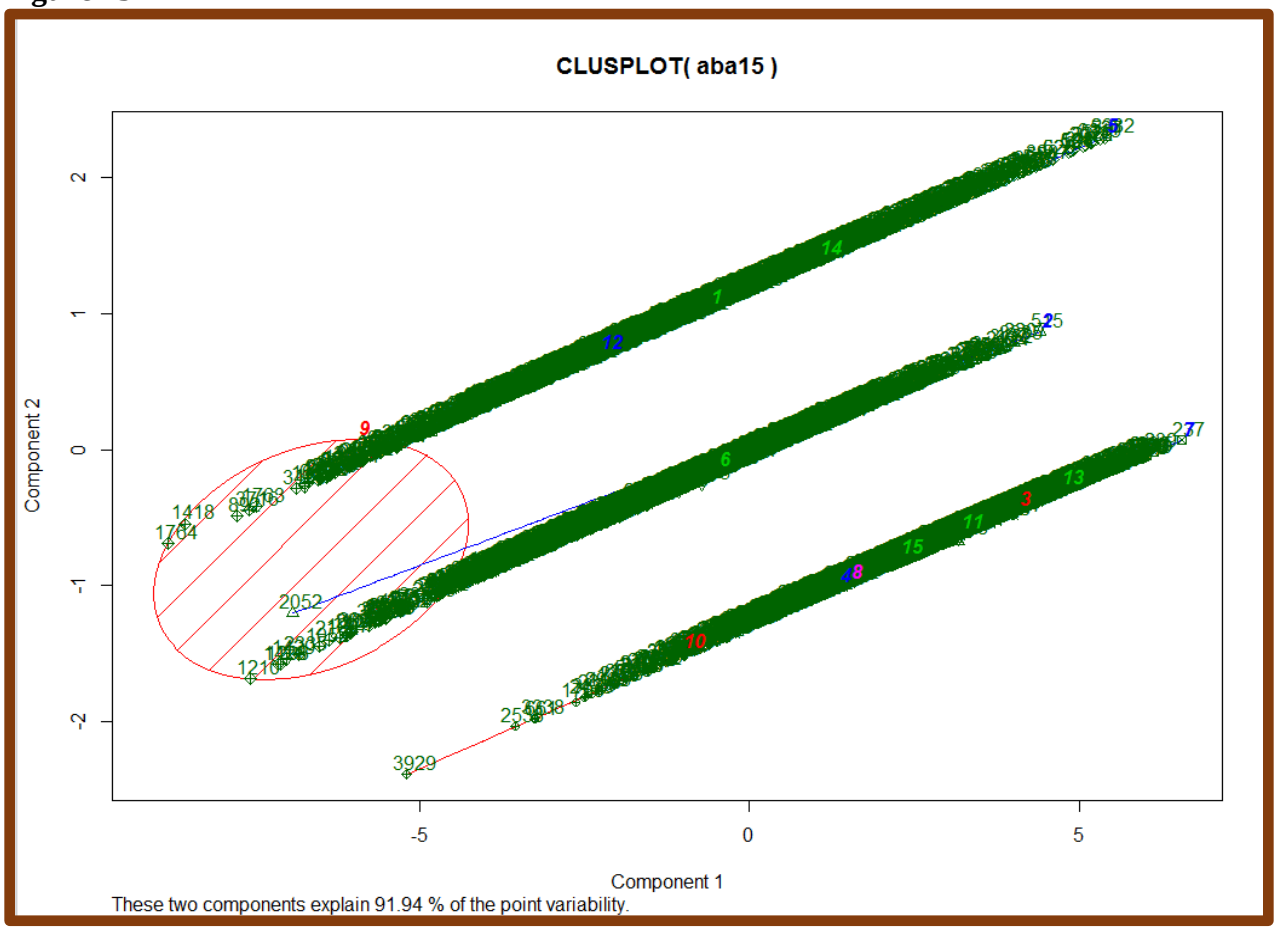

CLUSPLOT(aba15)
Component 2
Component 1
-5
0
5
-2
-1
0
1
2
These two components explain 91.94 % of the point variability.

Figure 44A

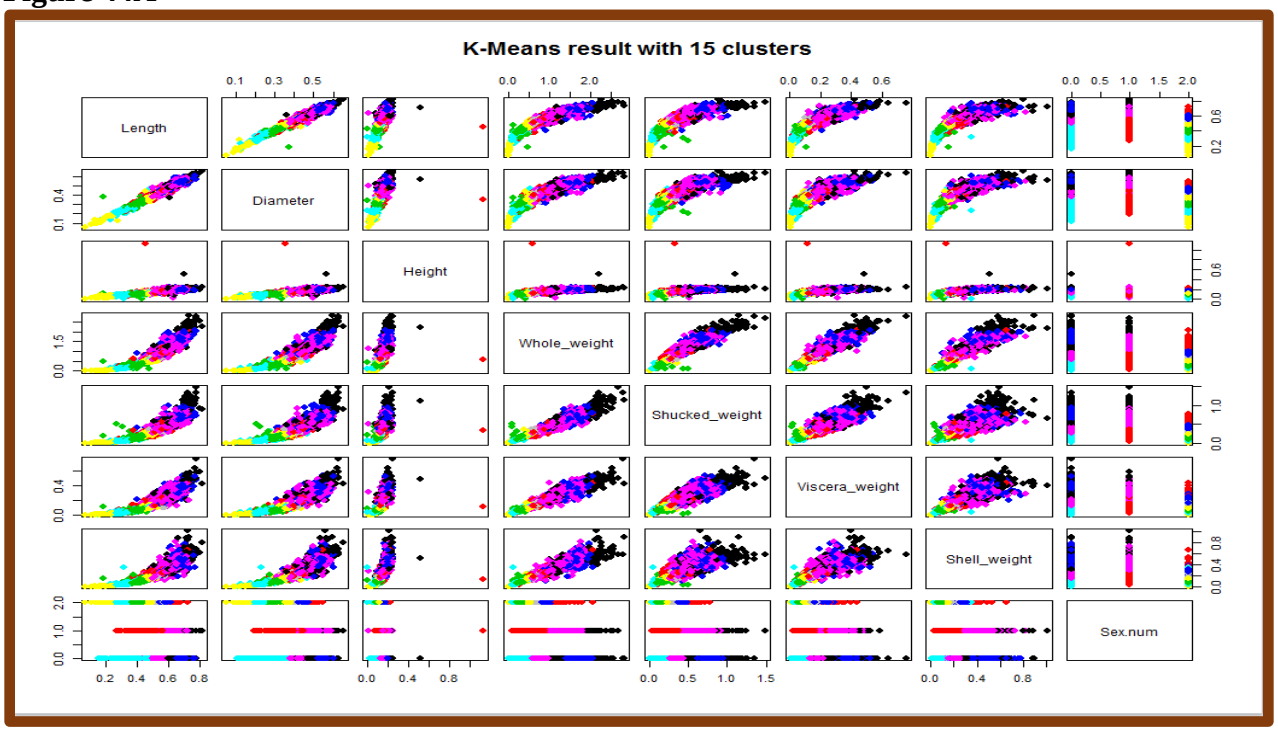


Figure 45A

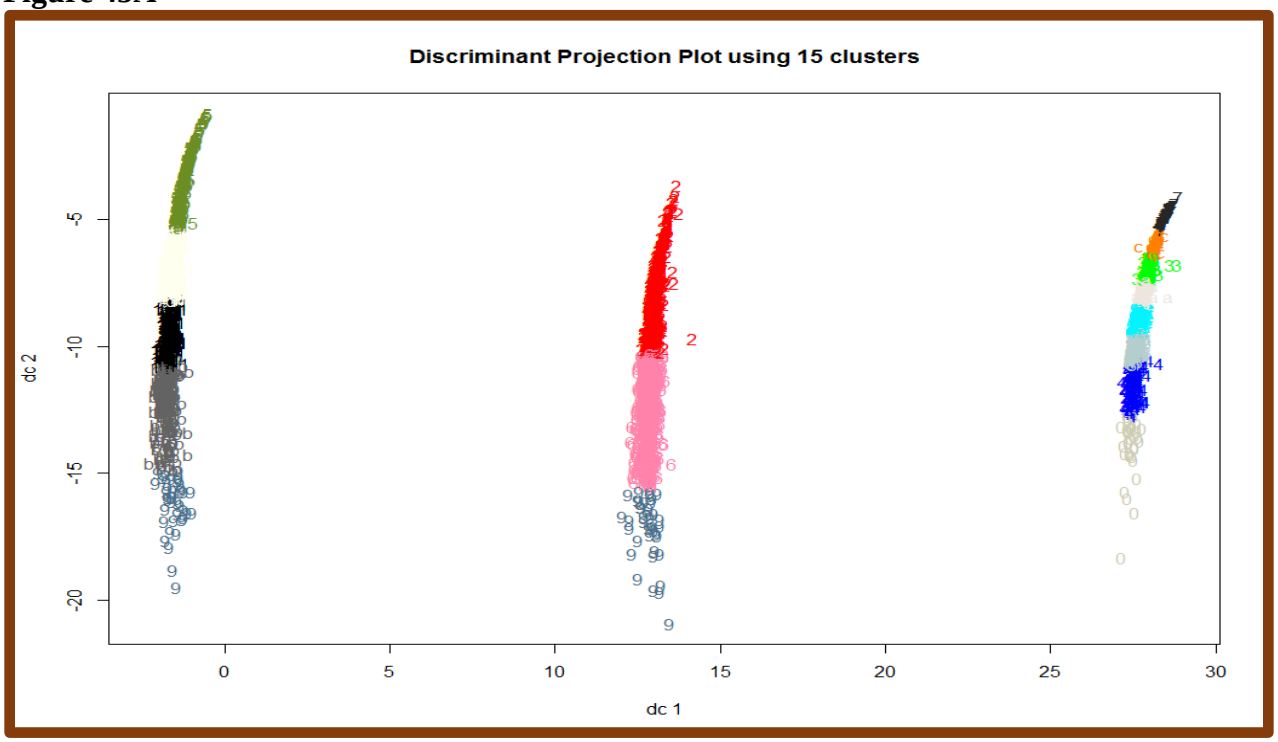


Figure 46A

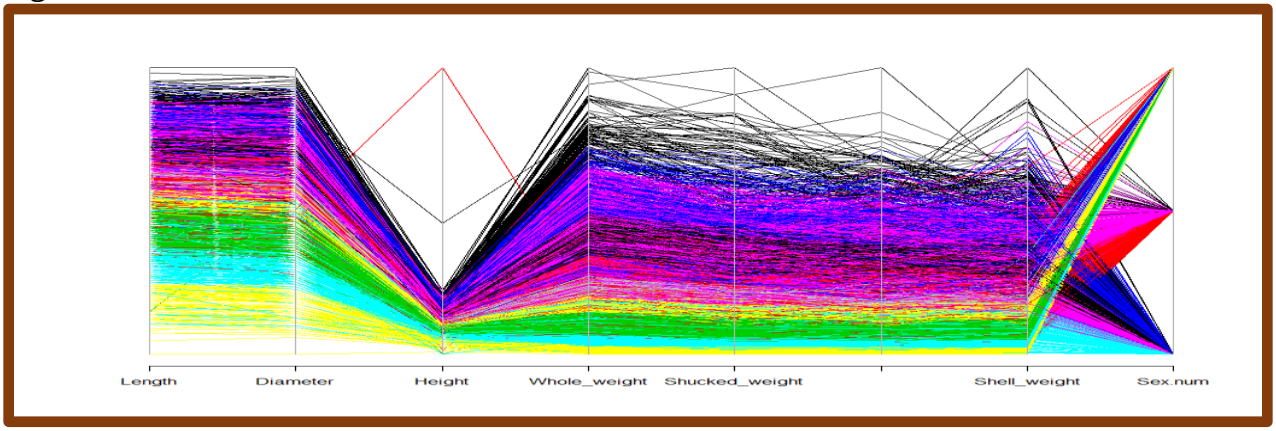
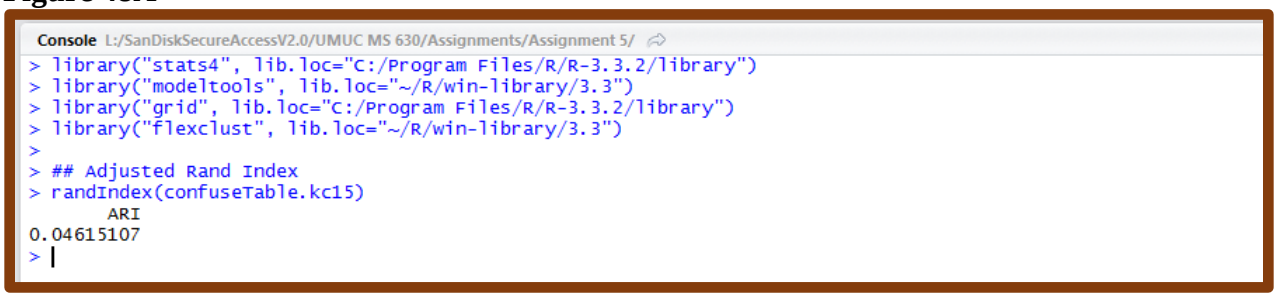


Figure 47A

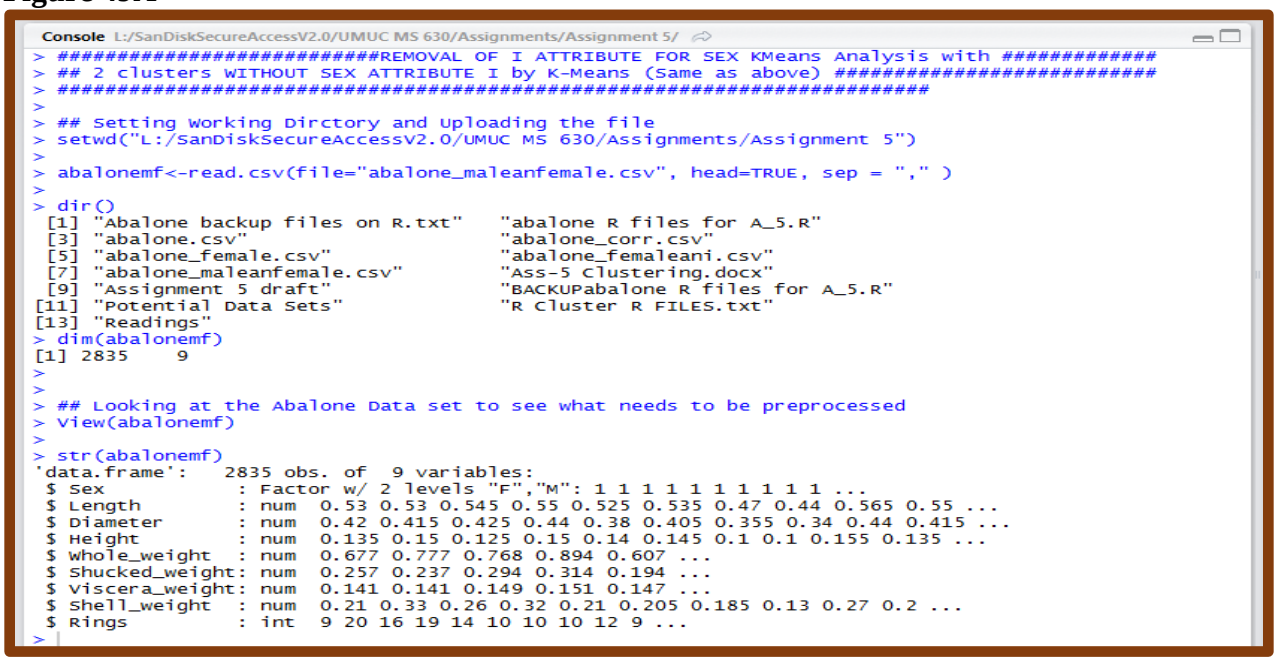
```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ## Assignment 7 Additional parameters requested by Dr. Firdu Bati
> kc15$size
 [1] 496 603 229 142 268 664 106 188  79  33 230 232 173 493 241
>
> kc15$betweenss
[1] 4128.153
>
> kc15$tot.withinss
[1] 183.3132
>
> kc15$withinss
 [1] 12.7081994 44.0520561  1.0133157  2.1356591 11.0032152 53.0876587  0.3707175  1.3210146 26.8837640  2.0366500  0.8891780 13.4965591
[13]  0.5268507 12.5575450  1.2307702
>
> kc15$iter
[1] 3
>
>
>
> ##Building the Clustering Evaluation Cross Validation
> table(abalone$Rings, kc15$cluster)

      1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
  1   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0
  2   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0
  3   0   0   0   0   3   0  12   0   0   0   0   0   0   0   0
  4   0   0   1   0   6   0  42   0   0   0   0   0   8   0   0
  5   0   4  17   0  11   0  34   0   0   0   4   0  45   0   0
  6   0  14  73   2  21   2  15   1   0   0  49   0  63   6  13
  7   5  39  85   1  43   5   1  12   0   0  84   0  34  32  50
  8  41  87  33  20  47  35   0  61   0   0  55   4  14  80  91
  9 108 122   9  34  53 114   0  56   3   3  17  19   6  97  48
  10 123 101   7  36  40 140   0  27  12   5   5  51   1  75  11
  11  79  59   2  23  21 133   0  14  17   6   6  66   1  50  10
  12  36  54   1   5  11  67   0   3  19   4   4  21   0  38   4
  13  27  39   0   6   4  45   0   6   9   3   4  21   1  34   4
  14  12  20   1   6   3  32   0   2   6   1   1  13   0  26   3
  15  19  17   0   1   3  24   0   3   0   3   0  11   0  19   3
  16  14  12   0   2   1  16   0   0   3   2   0   3   0  11   3
  17  10  13   0   4   0  12   0   0   4   3   0   5   0   7   0
  18   6   7   0   1   1  10   0   1   2   2   0   6   0   5   1
  19   4   6   0   0   0   9   0   1   0   0   1   4   0   7   0
  20   6   2   0   0   0   9   0   1   1   1   0   3   0   3   0
  21   3   2   0   1   0   5   0   0   0   0   0   2   0   1   0
  22   0   2   0   0   0   1   0   0   0   0   0   2   0   1   0
  23   2   3   0   0   0   2   0   0   1   0   0   0   0   1   0
  24   0   0   0   0   0   0   0   0   1   0   0   1   0   0   0
  25   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0
  26   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  27   0   0   0   0   0   1   0   0   1   0   0   0   0   0   0
  29   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0
>
> table(abalone$Rings)

  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  29
  1   1  15  57 115 259 391 568 689 634 487 267 203 126 103  67  58  42  32  26  14   6   9   2   1   1   2   1
```

Figure 48A



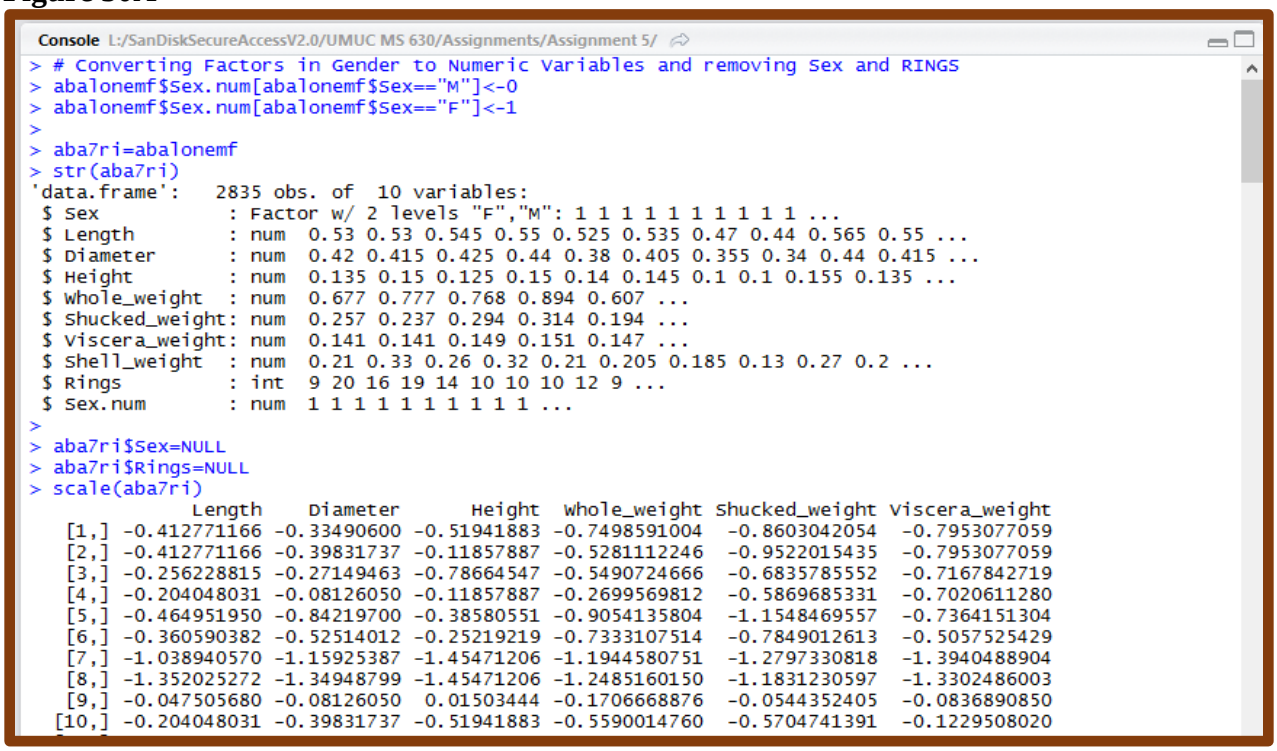
```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> library("stats4", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("modeltools", lib.loc="~/R/win-library/3.3")
> library("grid", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("flexclust", lib.loc="~/R/win-library/3.3")
>
> ## Adjusted Rand Index
> randIndex(confuseTable.kc15)
       ARI
0.04615107
>
```

Figure 49A



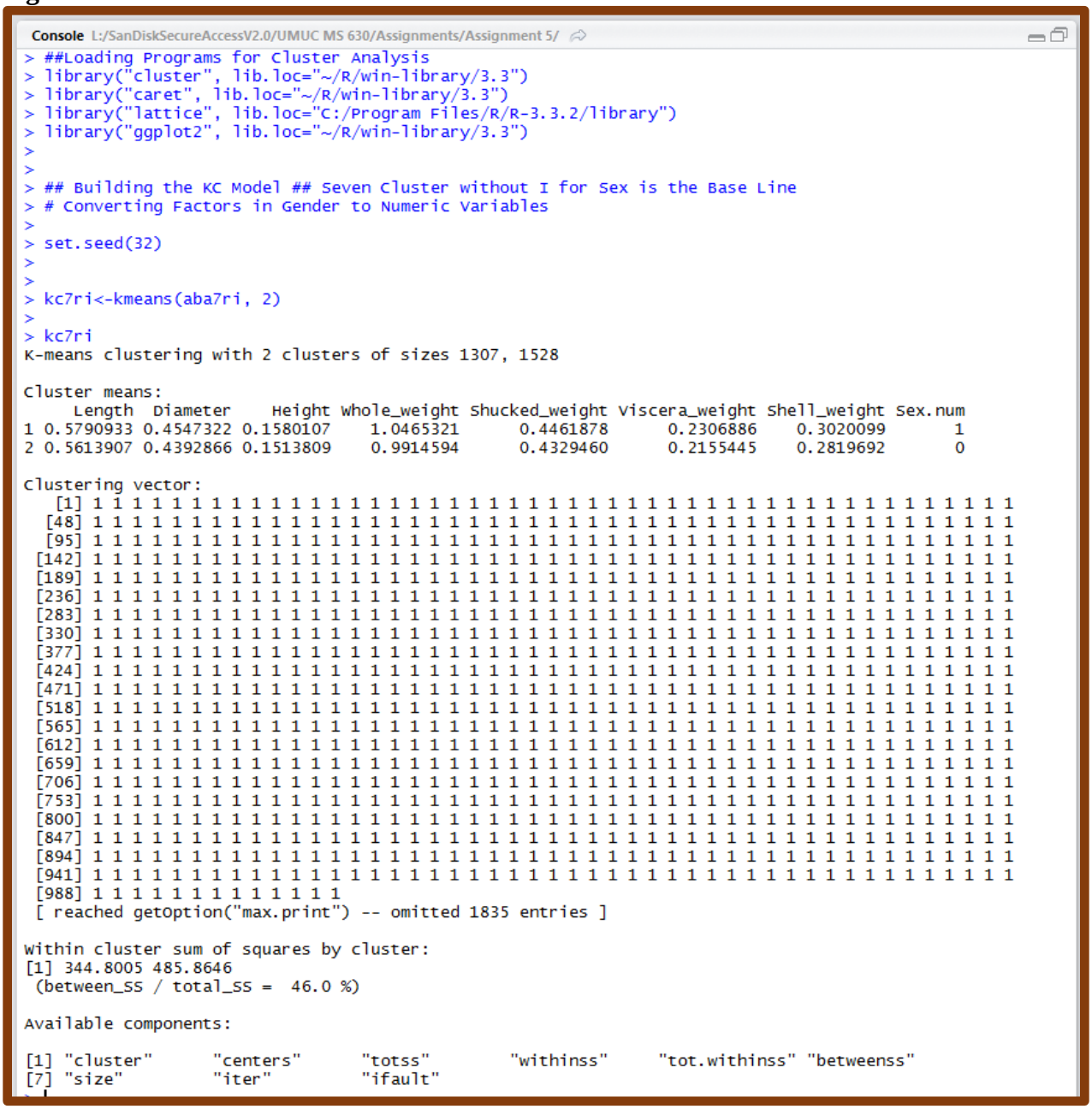
```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ##############################REMOVAL OF I ATTRIBUTE FOR SEX KMeans Analysis with ##############
> ## 2 clusters WITHOUT SEX ATTRIBUTE I by K-Means (Same as above) ############################
> ################################################################################
>
> ## Setting Working Dirctory and Uploading the file
> setwd("L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5")
>
> abalonemf<-read.csv(file="abalone_maleanfemale.csv", head=TRUE, sep = "," )
>
> dir()
 [1] "Abalone backup files on R.txt"   "abalone R files for A_5.R"
 [3] "abalone.csv"                     "abalone_corr.csv"
 [5] "abalone_female.csv"              "abalone_femaleani.csv"
 [7] "abalone_maleanfemale.csv"        "Ass-5 Clustering.docx"
 [9] "Assignment 5 draft"              "BACKUPabalone R files for A_5.R"
[11] "Potential Data Sets"             "R Cluster R FILES.txt"
[13] "Readings"
> dim(abalonemf)
[1] 2835    9
>
>
> ## Looking at the Abalone Data set to see what needs to be preprocessed
> View(abalonemf)
>
> str(abalonemf)
'data.frame':   2835 obs. of  9 variables:
 $ Sex           : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
 $ Length        : num  0.53 0.53 0.545 0.55 0.525 0.535 0.47 0.44 0.565 0.55 ...
 $ Diameter      : num  0.42 0.415 0.425 0.44 0.38 0.405 0.355 0.34 0.44 0.415 ...
 $ Height        : num  0.135 0.15 0.125 0.15 0.14 0.145 0.1 0.1 0.155 0.135 ...
 $ Whole_weight  : num  0.677 0.777 0.768 0.894 0.607 ...
 $ Shucked_weight: num  0.257 0.237 0.294 0.314 0.194 ...
 $ Viscera_weight: num  0.141 0.141 0.149 0.151 0.147 ...
 $ Shell_weight  : num  0.21 0.33 0.26 0.32 0.21 0.205 0.185 0.13 0.27 0.2 ...
 $ Rings         : int  9 20 16 19 14 10 10 10 12 9 ...
>
```

Figure 50A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> # Converting Factors in Gender to Numeric Variables and removing Sex and RINGS
> abalonemf$Sex.num[abalonemf$Sex=="M"]<-0
> abalonemf$Sex.num[abalonemf$Sex=="F"]<-1
>
> aba7ri=abalonemf
> str(aba7ri)
'data.frame':   2835 obs. of  10 variables:
 $ Sex           : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
 $ Length        : num  0.53 0.53 0.545 0.55 0.525 0.535 0.47 0.44 0.565 0.55 ...
 $ Diameter      : num  0.42 0.415 0.425 0.44 0.38 0.405 0.355 0.34 0.44 0.415 ...
 $ Height        : num  0.135 0.15 0.125 0.15 0.14 0.145 0.1 0.1 0.155 0.135 ...
 $ Whole_weight  : num  0.677 0.777 0.768 0.894 0.607 ...
 $ Shucked_weight: num  0.257 0.237 0.294 0.314 0.194 ...
 $ Viscera_weight: num  0.141 0.141 0.149 0.151 0.147 ...
 $ Shell_weight  : num  0.21 0.33 0.26 0.32 0.21 0.205 0.185 0.13 0.27 0.2 ...
 $ Rings         : int  9 20 16 19 14 10 10 10 12 9 ...
 $ Sex.num       : num  1 1 1 1 1 1 1 1 1 1 ...
>
> aba7ri$Sex=NULL
> aba7ri$Rings=NULL
> scale(aba7ri)
           Length    Diameter      Height  Whole_weight Shucked_weight Viscera_weight
   [1,] -0.412771166 -0.33490600 -0.51941883 -0.7498591004  -0.8603042054  -0.7953077059
   [2,] -0.412771166 -0.39831737 -0.11857887 -0.5281112246  -0.9522015435  -0.7953077059
   [3,] -0.256228815 -0.27149463 -0.78664547 -0.5490724666  -0.6835785552  -0.7167842719
   [4,] -0.204048031 -0.08126050 -0.11857887 -0.2699569812  -0.5869685331  -0.7020611280
   [5,] -0.464951950 -0.84219700 -0.38580551 -0.9054135804  -1.1548469557  -0.7364151304
   [6,] -0.360590382 -0.52514012 -0.25219219 -0.7333107514  -0.7849012613  -0.5057525429
   [7,] -1.038940570 -1.15925387 -1.45471206 -1.1944580751  -1.2797330818  -1.3940488904
   [8,] -1.352025272 -1.34948799 -1.45471206 -1.2485160150  -1.1831230597  -1.3302486003
   [9,] -0.047505680 -0.08126050  0.01503444 -0.1706668876  -0.0544352405  -0.0836890850
  [10,] -0.204048031 -0.39831737 -0.51941883 -0.5590014760  -0.5704741391  -0.1229508020
```

Figure 51A



```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ##Loading Programs for Cluster Analysis
> library("cluster", lib.loc="~/R/win-library/3.3")
> library("caret", lib.loc="~/R/win-library/3.3")
> library("lattice", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("ggplot2", lib.loc="~/R/win-library/3.3")
>
>
> ## Building the KC Model ## Seven Cluster without I for Sex is the Base Line
> # Converting Factors in Gender to Numeric Variables
>
> set.seed(32)
>
>
> kc7ri<-kmeans(aba7ri, 2)
>
> kc7ri
K-means clustering with 2 clusters of sizes 1307, 1528

Cluster means:
     Length  Diameter    Height Whole_weight Shucked_weight Viscera_weight Shell_weight Sex.num
1 0.5790933 0.4547322 0.1580107    1.0465321      0.4461878      0.2306886    0.3020099       1
2 0.5613907 0.4392866 0.1513809    0.9914594      0.4329460      0.2155445    0.2819692       0

Clustering vector:
   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  [48] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  [95] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [142] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [189] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [236] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [283] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [330] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [377] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [424] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [471] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [518] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [565] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [612] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [659] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [706] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [753] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [800] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [847] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [894] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [941] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [988] 1 1 1 1 1 1 1 1 1 1 1 1 1
 [ reached getOption("max.print") -- omitted 1835 entries ]

Within cluster sum of squares by cluster:
[1] 344.8005 485.8646
 (between_SS / total_SS =  46.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

Figure 52A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> #plotting or Graphing
> clusplot(aba7ri, kc7ri$cluster, color=TRUE, shade = TRUE, labels=2, line=0)
>
> plot(aba7ri)
> plot(aba7ri, col =(kc7ri$cluster) , main="K-Means result with 2 clusters only Male and Female", pch=20, cex=2)
>
>
> ## plotcluster function from fpc package to draw discriminant projection plot
> library("fpc", lib.loc="~/R/win-library/3.3")
>
> plotcluster(aba7ri, kc7ri$cluster, main="Discriminant Projection Plot using 2 clusters, only Male and Female")
>
>
> library("MASS", lib.loc="C:/Program Files/R/R-3.3.2/library")
> parcoord(aba7ri, kc7ri$cluster)
>
```

Figure 53A

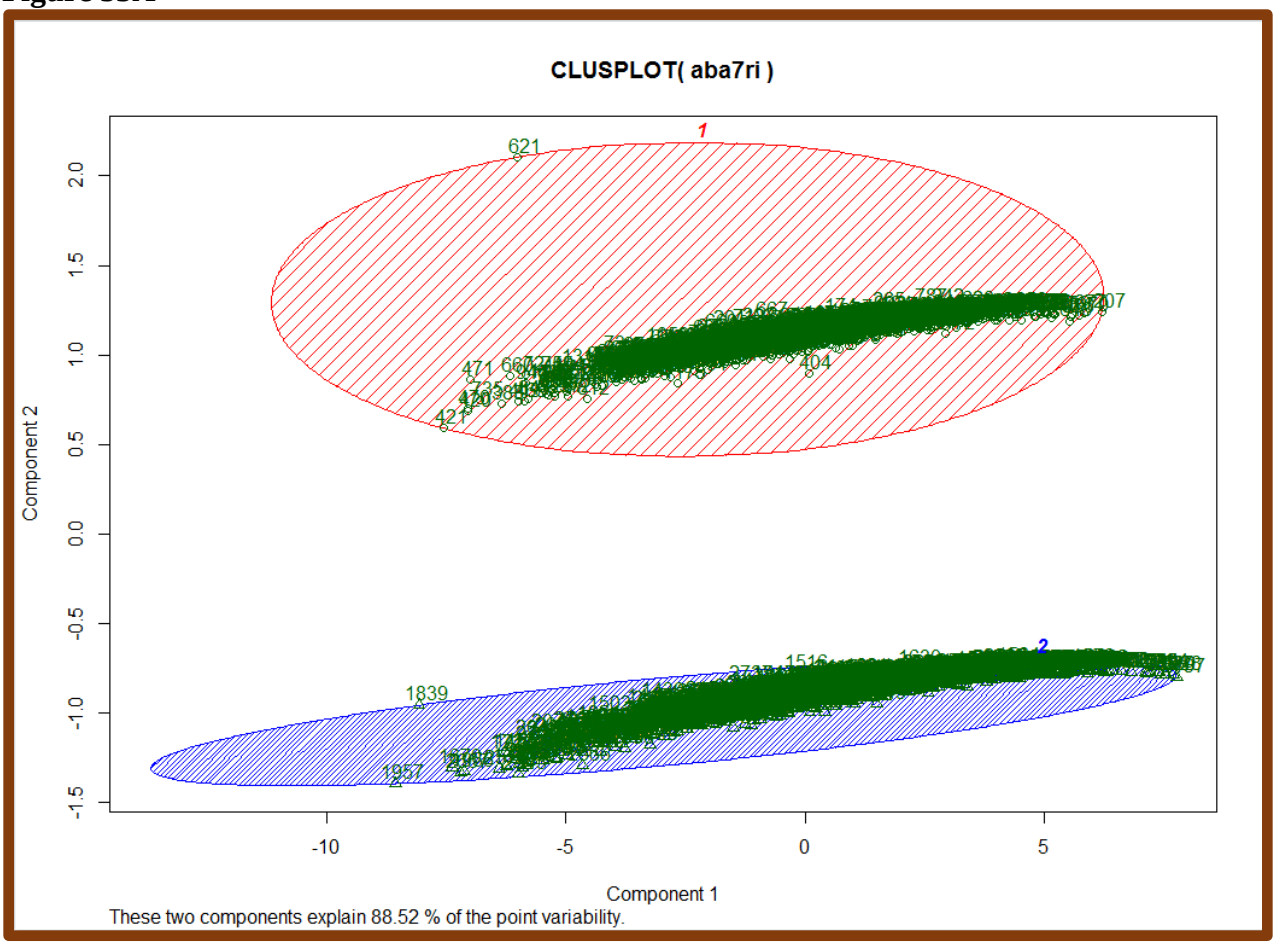


Figure 54A

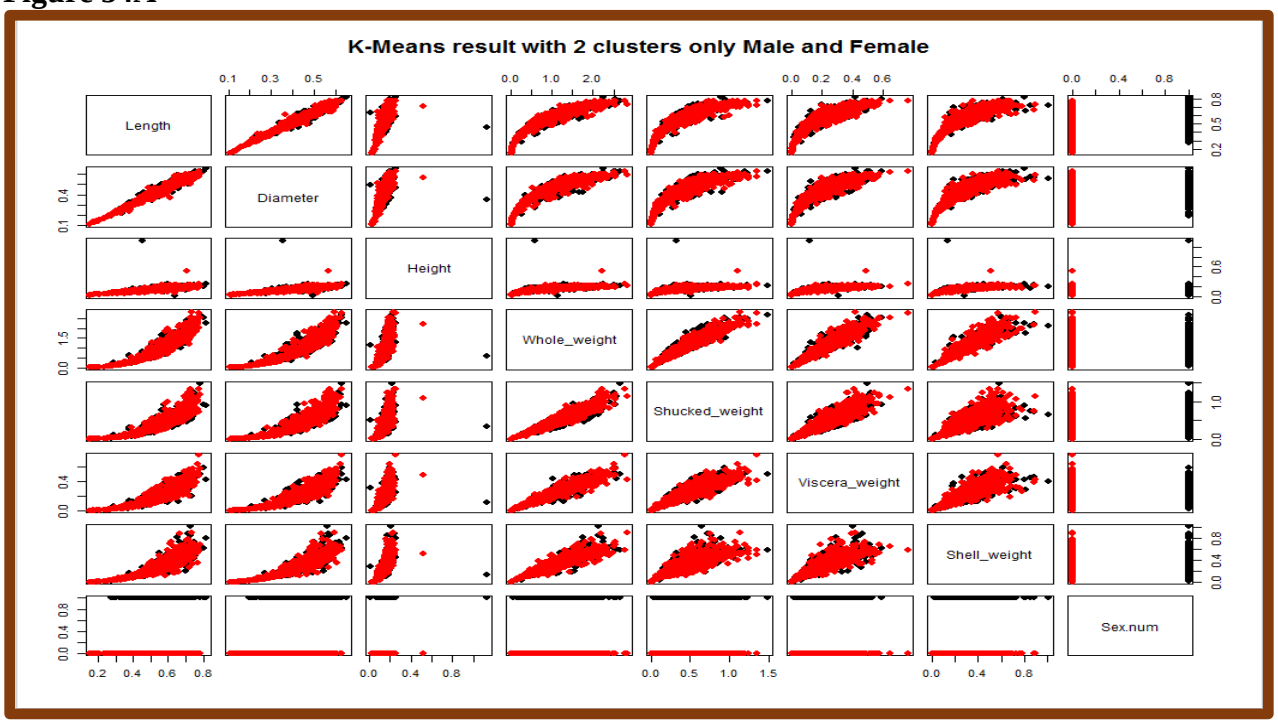


Figure 55A

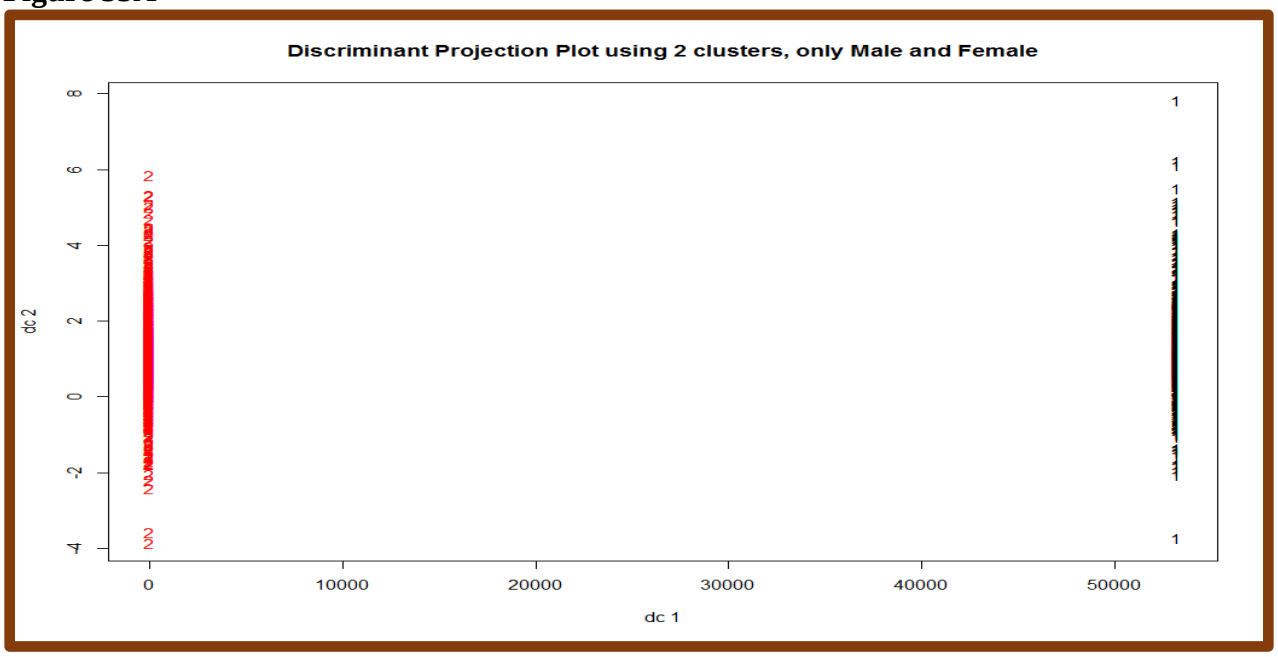


Figure 56A

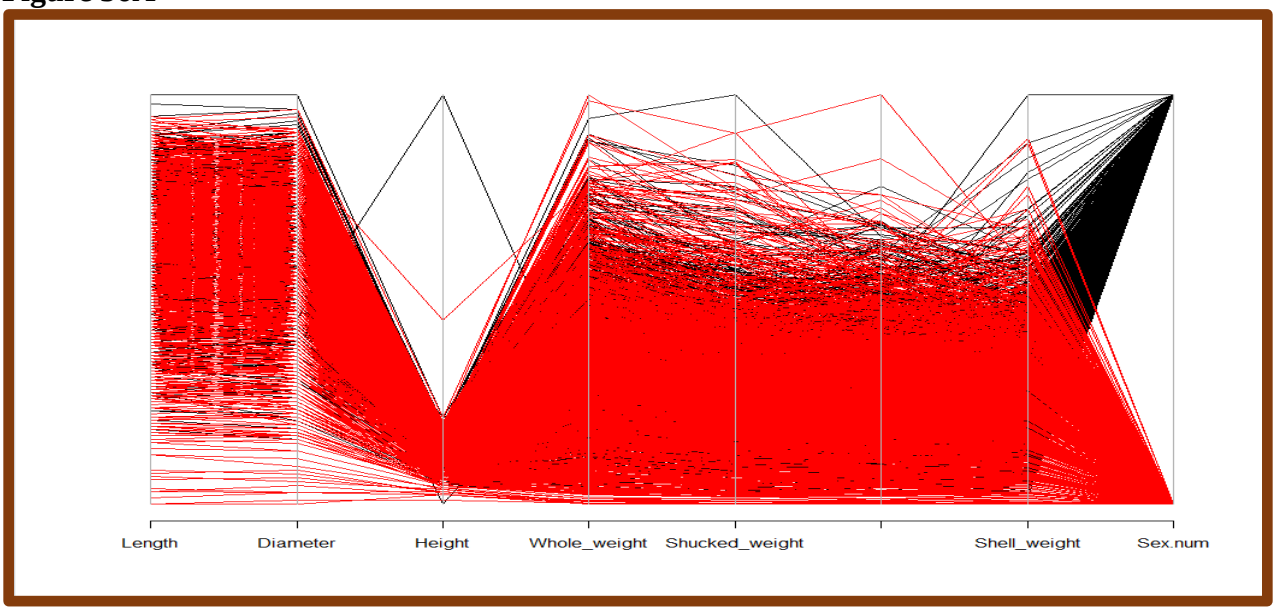


Figure 57A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ## Assignment 7 Additional parameters requested by Dr. Firdu Bati
> kc7ri$size
[1] 1307 1528
>
> kc7ri$betweenss
[1] 707.5674
>
> kc7ri$tot.withinss
[1] 830.6651
>
> kc7ri$withinss
[1] 344.8005 485.8646
>
> kc7ri$iter
[1] 1
```

Figure 58A

```
> ##Building the Clustering Evaluation Cross Validation
> table(abalonemf$Rings, kc7ri$cluster)

     1   2
3    0   3
4    0   6
5    4  11
6   16  27
7   44  80
8  122 172
9  238 278
10 248 294
11 200 225
12 128 118
13  88  91
14  56  56
15  41  52
16  30  30
17  26  25
18  19  18
19  15  15
20  12  12
21   7   6
22   3   3
23   6   3
24   1   1
25   1   0
26   0   1
27   1   1
29   1   0
> table(abalonemf$Rings)

  3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27
  3   6  15  43 124 294 516 542 425 246 179 112  93  60  51  37  30  24  13   6   9   2   1   1   2
 29
  1
>
```

Figure 59A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> library("stats4", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("modeltools", lib.loc="~/R/win-library/3.3")
> library("grid", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("flexclust", lib.loc="~/R/win-library/3.3")
>
> ## Adjusted Rand Index
> randIndex(confuseTable.kc7ri)
          ARI
5.879794e-05
>
```

Figure 60A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> ############################REMOVAL OF MALE ATTRIBUTE FOR SEX KMeans Analysis with ##############
> ## 2 clusters WITHOUT SEX ATTRIBUTE MALE by K-Means (Same as above) ##########################
> ######################################################################
>
> ## Setting Working Dirctory and Uploading the file
> setwd("L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5")
>
> abalonefi<-read.csv(file="abalone_femaleani.csv", head=TRUE, sep = "," )
>
> dir()
 [1] "Abalone backup files on R.txt"    "abalone R files for A_5.R"        "abalone.csv"
 [4] "abalone_corr.csv"                 "abalone_female.csv"               "abalone_femaleani.csv"
 [7] "abalone_maleanfemale.csv"         "Ass-5 Clustering.docx"            "Assignment 5 draft"
[10] "BACKUPabalone R files for A_5.R" "Potential Data Sets"              "R Cluster R FILES.txt"
[13] "Readings"
> dim(abalonefi)
[1] 2649    9
>
>
> ## Looking at the Abalone Data set to see what needs to be preprocessed
> View(abalonefi)
>
> str(abalonefi)
'data.frame':   2649 obs. of  9 variables:
 $ Sex           : Factor w/ 2 levels "F","I": 1 1 1 1 1 1 1 1 1 1 ...
 $ Length        : num  0.53 0.53 0.545 0.55 0.525 0.535 0.47 0.44 0.565 0.55 ...
 $ Diameter      : num  0.42 0.415 0.425 0.44 0.38 0.405 0.355 0.34 0.44 0.415 ...
 $ Height        : num  0.135 0.15 0.125 0.15 0.14 0.145 0.1 0.1 0.155 0.135 ...
 $ Whole_weight  : num  0.677 0.777 0.768 0.894 0.607 ...
 $ Shucked_weight: num  0.257 0.237 0.294 0.314 0.194 ...
 $ Viscera_weight: num  0.141 0.141 0.149 0.151 0.147 ...
 $ Shell_weight  : num  0.21 0.33 0.26 0.32 0.21 0.205 0.185 0.13 0.27 0.2 ...
 $ Rings         : int  9 20 16 19 14 10 10 10 12 9 ...
>
```

Figure 61A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> # Converting Factors in Gender to Numeric Variables and removing Sex and RINGS
> abalonefi$Sex.num[abalonefi$Sex=="F"]<-0
> abalonefi$Sex.num[abalonefi$Sex=="I"]<-1
>
> aba7fi=abalonefi
> str(aba7fi)
'data.frame':   2649 obs. of  10 variables:
 $ Sex           : Factor w/ 2 levels "F","I": 1 1 1 1 1 1 1 1 1 1 ...
 $ Length        : num  0.53 0.53 0.545 0.55 0.525 0.535 0.47 0.44 0.565 0.55 ...
 $ Diameter      : num  0.42 0.415 0.425 0.44 0.38 0.405 0.355 0.34 0.44 0.415 ...
 $ Height        : num  0.135 0.15 0.125 0.15 0.14 0.145 0.1 0.1 0.155 0.135 ...
 $ Whole_weight  : num  0.677 0.777 0.768 0.894 0.607 ...
 $ Shucked_weight: num  0.257 0.237 0.294 0.314 0.194 ...
 $ Viscera_weight: num  0.141 0.141 0.149 0.151 0.147 ...
 $ Shell_weight  : num  0.21 0.33 0.26 0.32 0.21 0.205 0.185 0.13 0.27 0.2 ...
 $ Rings         : int  9 20 16 19 14 10 10 10 12 9 ...
 $ Sex.num       : num  0 0 0 0 0 0 0 0 0 0 ...
> aba7fi$Sex=NULL
> aba7fi$Rings=NULL
> scale(aba7fi)
         Length     Diameter       Height Whole_weight Shucked_weight Viscera_weight  Shell_weight     Sex.num
   [1,]  0.22231974  0.294681792  0.05294211 -0.121367768   -0.287761211  -0.1766358013 -0.0286437839 -1.0131097
   [2,]  0.22231974  0.245948363  0.39416754  0.089357081   -0.380624552  -0.1766358013  0.8420505534 -1.0131097
   [3,]  0.34323235  0.343415222 -0.17454151  0.069437817   -0.109177863  -0.1020003604  0.3341455233 -1.0131097
   [4,]  0.38353655  0.489615509  0.39416754  0.334678548   -0.011552300  -0.0880062153  0.7694926920 -1.0131097
   [5,]  0.18201554 -0.095185641  0.16668392 -0.269189677   -0.585400123  -0.1206592206 -0.0286437839 -1.0131097
   [6,]  0.26262394  0.148481505  0.28042573 -0.105642033   -0.211565649   0.0985823869 -0.0649227146 -1.0131097
   [7,] -0.26133068 -0.338852786 -0.74325055 -0.543865849   -0.711599022  -0.7457310377 -0.2100384375 -1.0131097
   [8,] -0.50315589 -0.485053074 -0.74325055 -0.595236583   -0.613973459  -0.6850897420 -0.6091066755 -1.0131097
   [9,]  0.50444916  0.489615509  0.50790935  0.429032958    0.526578854   0.4997478814  0.4067033847 -1.0131097
  [10,]  0.38353655  0.245948363  0.05294211  0.060002376    0.005115479   0.4624301610 -0.1012016454 -1.0131097
```

Figure 62A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> summary(aba7fi)
     Length          Diameter          Height        Whole_weight    Shucked_weight   Viscera_weight  
 Min.   :0.0750   Min.   :0.0550   Min.   :0.0000   Min.   :0.0020   Min.   :0.0010   Min.   :0.0005  
 1st Qu.:0.4200   1st Qu.:0.3250   1st Qu.:0.1050   1st Qu.:0.3535   1st Qu.:0.1535   1st Qu.:0.0735  
 Median :0.5150   Median :0.4000   Median :0.1350   Median :0.6515   Median :0.2810   Median :0.1415  
 Mean   :0.5024   Mean   :0.3898   Mean   :0.1327   Mean   :0.7349   Mean   :0.3169   Mean   :0.1604  
 3rd Qu.:0.6000   3rd Qu.:0.4700   3rd Qu.:0.1600   3rd Qu.:1.0645   3rd Qu.:0.4550   3rd Qu.:0.2300  
 Max.   :0.8150   Max.   :0.6500   Max.   :1.1300   Max.   :2.6570   Max.   :1.4880   Max.   :0.5900  
  Shell_weight        Sex.num      
 Min.   :0.0015   Min.   :0.0000  
 1st Qu.:0.1050   1st Qu.:0.0000  
 Median :0.1945   Median :1.0000  
 Mean   :0.2139   Mean   :0.5066  
 3rd Qu.:0.3045   3rd Qu.:1.0000  
 Max.   :1.0050   Max.   :1.0000  
> ##Loading Programs for Cluster Analysis
> library("cluster", lib.loc="~/R/win-library/3.3")
> library("caret", lib.loc="~/R/win-library/3.3")
> library("lattice", lib.loc="C:/Program Files/R/R-3.3.2/library")
> library("ggplot2", lib.loc="~/R/win-library/3.3")
> 
> 
> ## Building the KC Model ## Seven Cluster without Male for Sex is the Base Line
> # Converting Factors in Gender to Numeric Variables
> 
> set.seed(32)
> 
> kc7fi<-kmeans(aba7fi, 2)
> 
```

Figure 63A

```
Console L:/SanDiskSecureAccessV2.0/UMUC MS 630/Assignments/Assignment 5/
> kc7fi
K-means clustering with 2 clusters of sizes 1304, 1345

Cluster means:
     Length  Diameter    Height Whole_weight Shucked_weight Viscera_weight Shell_weight     Sex.num
1 0.5825537 0.4575613 0.1591526    1.0606534      0.4520890     0.23370284    0.3060606 0.009202454
2 0.4247286 0.3240372 0.1070000    0.4190439      0.1858829     0.08939703    0.1246428 0.988847584

Clustering vector:
   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  [56] 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [111] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [166] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1
 [221] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [276] 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [331] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [386] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [441] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [496] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [551] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [606] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [661] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [716] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [771] 1 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [826] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [881] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [936] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1
 [991] 1 1 1 1 1 1 1 1 1 1
 [ reached getOption("max.print") -- omitted 1649 entries ]

Within cluster sum of squares by cluster:
[1] 343.7936 171.3099
 (between_SS / total_SS =  66.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"        
[8] "iter"         "ifault"      
> 
```

Figure 64A

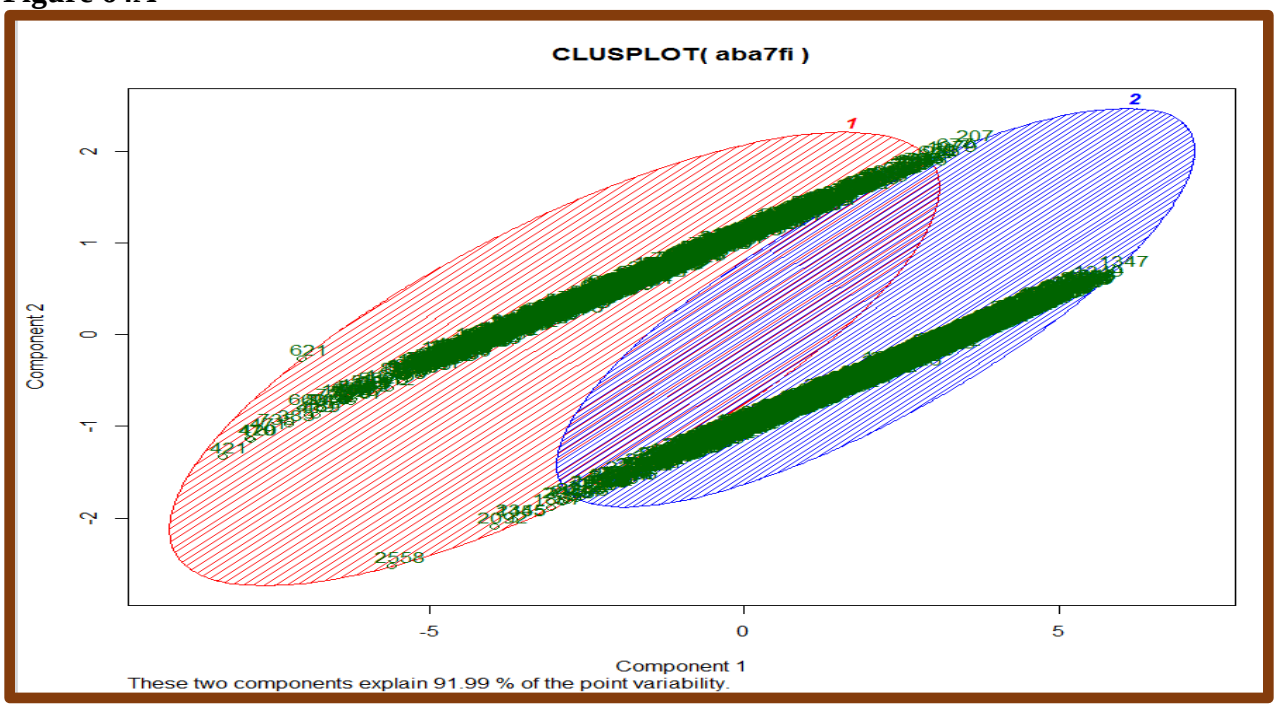


Figure 65A

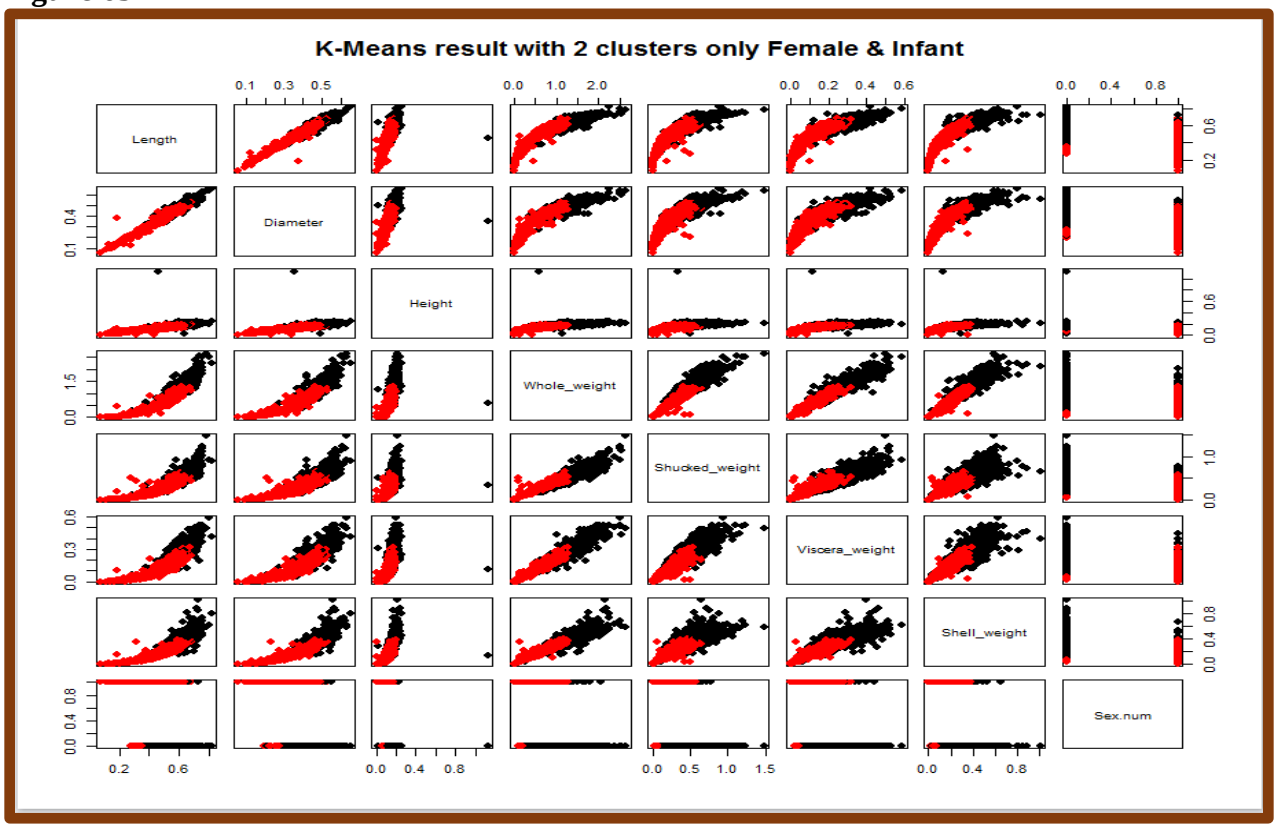


Figure 66A

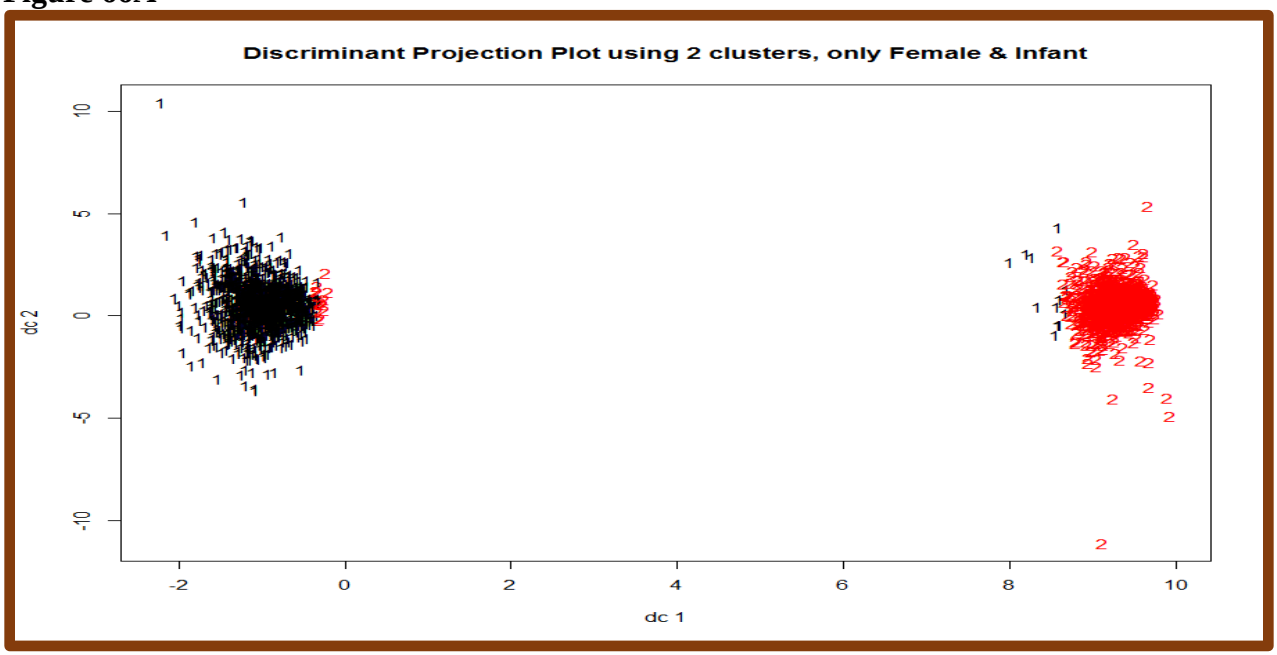


Figure 67A

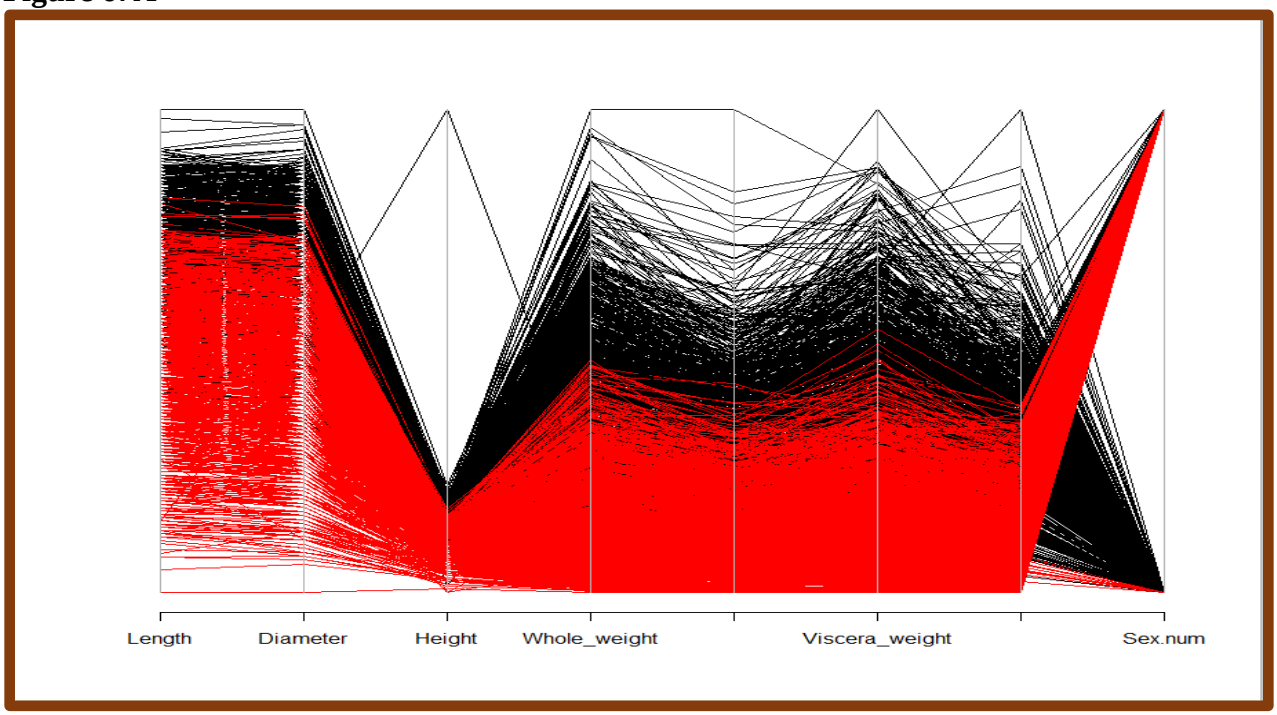


Figure 68A

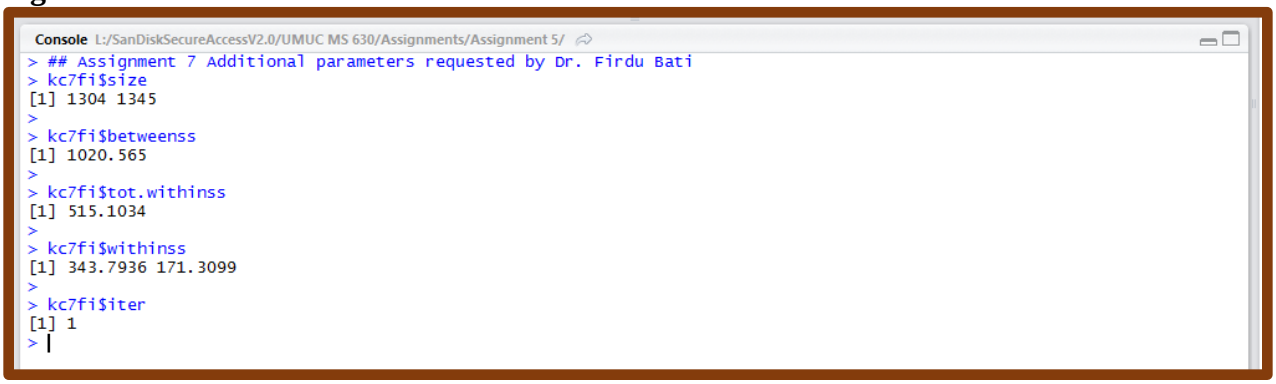


Figure 69A

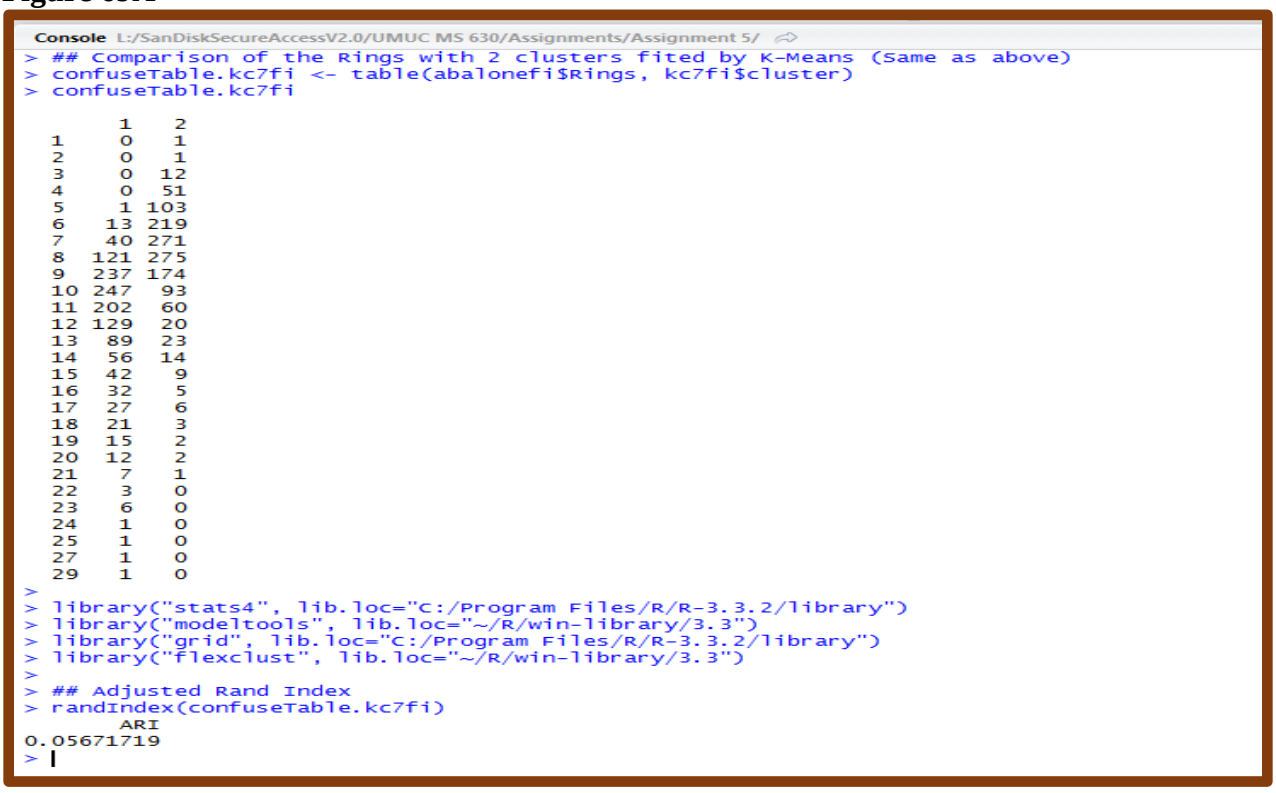


Figure 70A

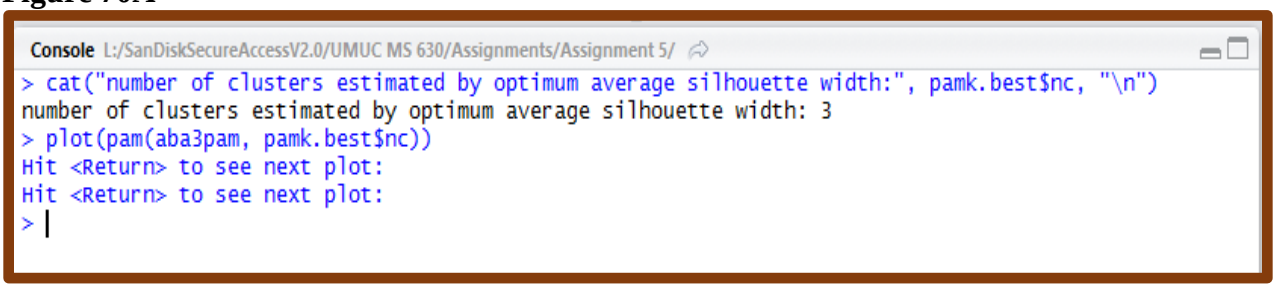


Figure 71A

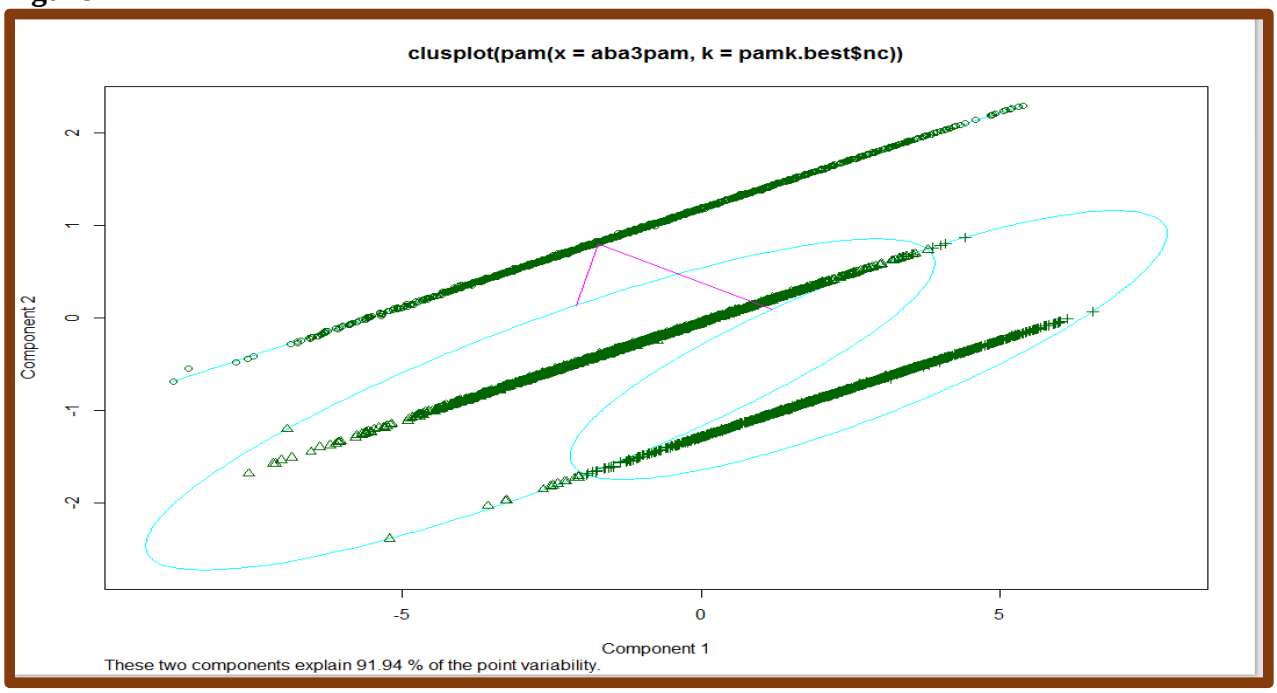
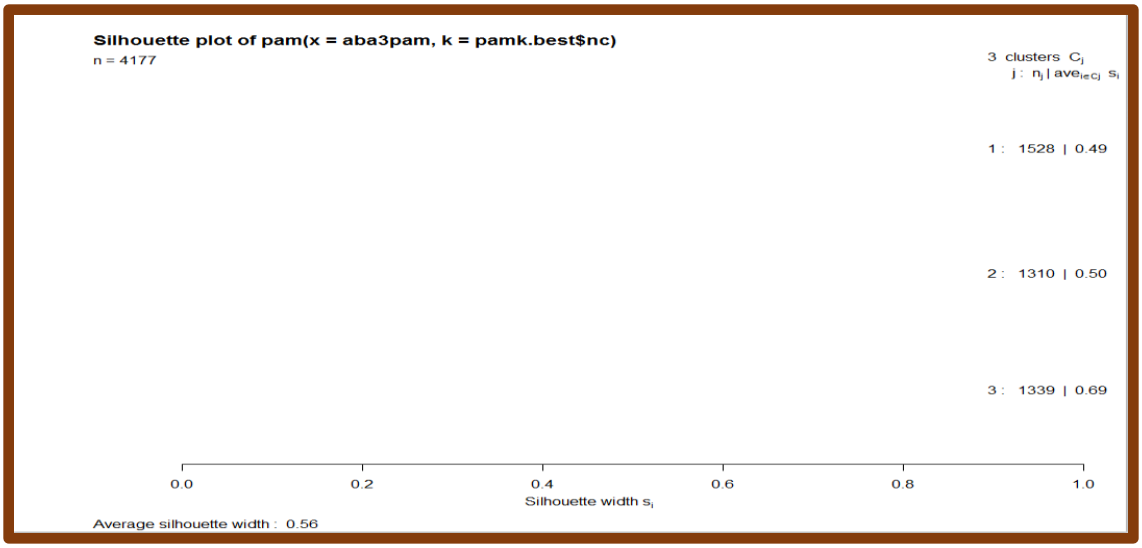


Figure 72A



Range of SC	Interpretation
0.71-1.0	A strong structure has been found
0.51-0.70	A reasonable structure has been found
0.26-0.50	The structure is weak and could be artificial
< 0.25	No substantial structure has been found

******* This table was copied from the Stat.berkeley web site and is used for this analysis.