# Meeting Student Needs for Multivariate Data Analysis: A Case Study in Teaching an Undergraduate Multivariate Data Analysis Course

Amy Wagaman

Amherst College, Amherst, Massachusetts, 01002

Amy S. Wagaman is Associate Professor of Statistics, Amherst College, Amherst, Massachusetts, 01002 (e-mail: awagaman@amherst.edu)

## Abstract

Modern students encounter large, messy data sets long before setting foot in our classrooms. Many of these students need to develop skills in exploratory data analysis and multivariate analysis techniques for their jobs after college, but such topics are not covered in traditional introductory statistics courses. This case study describes my experience in designing and teaching an undergraduate course on multivariate data analysis with minimal pre-requisites, using real data, active learning, and other interactive activities to help students tackle the material. Multivariate topics covered include clustering and classification (among others) for exploratory data analysis and an introduction to algorithmic modeling.

**Key Words**: course development, statistics education, GAISE, algorithmic modeling

# 1. INTRODUCTION

Modern undergraduate students are encountering and using data in ways unlike their past counterparts. Their first experience with data is typically not in statistics courses and the type of data they encounter (large, messy, complex data sets) is not usually part of introductory statistics courses, which cover basic figures and descriptive statistics and work up to hypothesis testing and confidence intervals in one or two variable settings (Gould 2010). When our students graduate and enter the workforce, many of their jobs include data management (retrieving, filtering, or cleaning) and performing basic exploratory data analysis (Nolan and Temple Lang 2010). However, they are unlikely to see these topics in a traditional introductory statistics course. We need to adjust to our students' changing needs. There is good news on this front.

Statistics courses are undergoing significant change with rapid development in areas such as data science. We are moving away from the "tyranny of the computable" (Cobb 2007) and many instructors have shifted towards using randomization-based procedures to introduce inference in introductory courses (Tintle et al. 2011). Data science courses and programs are being introduced around the country, and coursework in statistics is becoming more data-focused, with guidance provided by the updated American Statistical Association (ASA) Curriculum Guidelines (ASA 2014). However, we still are not meeting the needs of our undergraduate students in terms of dealing with complex data sets, visualizing data, and performing exploratory data analysis.

**How can we introduce our undergraduate students to more complex data sets and basic techniques for multivariate data analysis**? Clearly, adding multivariate analysis topics to introductory statistics would strain an already taxed curriculum. Some statistics educators have suggested some multivariate topics (simple classification or decision trees, for example) might replace or become topics in introductory statistics (De Veaux 2015). However, instead of adding topics to introductory statistics or otherwise modifying that course, we consider a new course, that can be taught without any pre-requisites.

To meet the needs of our students, the course needs to have exploratory data analysis and cover techniques such as classification and clustering methods, taught at a level that undergraduates (including students without a previous statistics course) can handle. Many classification and clustering methods fall into an "algorithmic" modeling culture, as described by Breiman (2001). These methods tend not to focus on the underlying stochastic data model, and in particular for the classification methods, such as trees and random forests, focus on prediction. If we want to serve our undergraduate students well, we should find a place for these methods in

our curriculum, because these are methods they actually see employed in practice (Breiman 2001).

We designed a course to introduce students with a minimal background to multivariate analysis techniques. The overall course goals were for students to be introduced to more complex data sets than is typical for introductory statistics (say, at least six variables to start and most ending up with data sets with 15-25 variables for their projects), and gain an applied understanding of exploratory data analysis and several multivariate analysis techniques. That meant students should be able to recognize appropriate and inappropriate applications, be able to interpret results, including justifying choices they had to make during the analysis, and communicate their results. In order to accomplish this, it was clear that students would need to see many examples, have time to practice on their own, and receive feedback on their communication of ideas and results.

By making the course applied, instead of theoretical, we did not need to rely on student background knowledge, and initially developed the course to be taught with no pre-requisites. As our statistics program evolved, we also taught this course with an introductory statistics pre-requisite. The introductory statistics pre-requisite has been a boon to the course in terms of improved student computational knowledge (improved based on what they see in our introductory course). Ideally, this course should have minimal or no pre-requisites in order to have a broad student audience (O'Shea and Pollatsek 1997). Further discussion of the issue of pre-requisites is postponed to later sections, but we emphasize that it can be taught without any.

To develop the course, we researched teaching approaches that would help the students learn the techniques. There are great examples available from statistics education research and suggestions from many educational reports to pull from. For example, we used real data (often

from the web or from a repository) throughout the course, as suggested by the updated ASA Curriculum Guidelines, the Guidelines for Assessment and Instruction in Statistics Education (GAISE) college report and others (ASA 2014; Franklin and Garfield 2006; Singer and Willett 1990). Students learned to deal with some challenges with the data – working with multiple sources, creating new variables or subsetting, etc. In particular, using the framework of Grimshaw (2015) to describe an authentic data experience, the data skills required across all assignments for the class ranged from good to best for both the breadth and depth categories. More information on this framework and how it worked for one module of the course are detailed below.

We structured the course with a mixture of lectures and interactive activities ("labs"), including class discussions and group work opportunities, to provide students with a variety of learning methods to help them with each topic we were covering. As pointed out in Snee (1993) and suggested in the Curriculum Renewal Across the First Two Years (CRAFTY) statistics report (Moore, Peck, and Rossman 2000), students learn in different ways and we need to incorporate a variety of methods into how we teach. Additionally, we used assessment methods that let us give detailed feedback to students on their work. It is known that students' learning is enhanced when students receive good feedback on the expressions of their ideas (Garfield and Ben-Zvi 2007). For this particular course, we felt good feedback was essential as students wrapped their minds around the idea that in data analysis, there may be several viable ways to tackle a problem, and that justification of decisions during analysis is very important. In addition, these opportunities allowed students to practice communicating their statistical results and ideas to others, culminating in a group presentation of a course project. Providing opportunities to

practice communication skills (with refinement) is a key point in the updated ASA curriculum guidelines (ASA 2014).

We incorporated active learning activities in the course (in many of the lab activities). There has been general enthusiasm about these activities, where students take an active role in their learning in the classroom. Active learning encourages students to think about underlying concepts or processes while giving them a meaningful task. Numerous studies have found active learning to be effective in statistics, though there are some conflicting accounts. For details, as well as a more substantial review and discussion on active learning approaches, see Carlson and Winquist 2011.

Another decision motivated by the course goals and student needs was to spend a solid part of the course on exploratory data analysis and data visualization. Clustering (as an example of exploratory data analysis) is a useful topic that students have intuition about and is a natural topic to include so students see ways of uncovering patterns in data. Including data visualization was motivated in part because the GAISE college report states that appropriate figure use is a learning outcome in introductory statistics courses (Franklin and Garfield 2006). Learning and practicing appropriate figure use should be relevant for any statistics course, and in fact, data visualization can be taught as an entire course on its own (Nolan and Perrett 2015). For this course, we wanted to keep the focus on students learning what figures could show, and not become mired in the details about their creation. This is in line with research about how to get students to view figures as reasoning tools rather than just "pretty" pictures (Garfield and Ben-Zvi 2007). Figures can also be helpful teaching tools. For example, Valero-Mora and Ledesma (2011) discuss their experiences teaching some multivariate data analysis techniques using interactive figures.

In order to have the students generate figures and develop some computational skills, it was clear that statistical software was needed for the course. The use of statistical software is also recommended by GAISE (Franklin and Garfield 2006), and the development of computational skills is highlighted in the updated ASA curriculum guidelines (ASA 2014). In course design, it is important to consider what level of programming the students can handle. For example, Samsa et al. (2012) discussed student attributes as part of their course development considerations and noted that their students were not skilled programmers. We used R (R Core Development Team 2009) each time we have taught the course, with different interfaces, due to changes in the computing background of the students. A variety of other software packages could have been used; see Chance et al. (2007) for other options. When working with any software, educators should note that students may be overwhelmed by software commands. Chance et al. (2007) give guidance for how to avoid this, and we designed the student activities with this in mind.

We have taught the course three times, while adapting to changing needs in our curriculum and advances in technology. In the sections below, we present course design details, an example module (and activities), and outcomes/course evaluation for this applied undergraduate statistics course as a case study. We begin with course design in section two, offering further details on class structure, software, and assessment. In section three, we present the clustering module from the course as an example module. This was the fifth module for a multivariate analysis technique in the course, but it could have easily been an earlier module. This example module shows how the various course design components were integrated. Finally, in section four, we include some evaluation of the course to assess how well we met the course

goals. This case study ends with considerations for the future and advice on determining pre-requisites in section five.

## 2. COURSE DESIGN AND STRUCTURE

In this section, we describe basic course design considerations – available classrooms, course software, course topics, general module structure, and course assessment. The primary reflection deals with teaching the class twice with no pre-requisites.

### 2.1 Classroom Space

Class time each week was divided into four fifty minute blocks in a lecture-style classroom with large two-person desks and a computer projection system, suitable for 36 students, with plenty of power outlets. We required students to bring their laptops for computer labs (or work in pairs). Having the desks and laptops available meant it was easy to go from an example on the board to an example on the students' laptops. Over the entire course of the semester, we used computers in the class an average of three of the four class meetings each week. The classroom setup was also nice for group work – whether that involved the computers or not.

An alternative setup was used in previous course offerings. We had access to two classrooms – a lecture classroom and a computer lab classroom on fixed days. This schedule meant careful planning was necessary to optimize the time in the computer lab where students could try techniques on their own.

We found the flexibility in the current structure to be beneficial to planning for the class. In particular, if an activity on the computer was taking longer than planned, we could continue it the next day, rather than waiting until we were back in the computer classroom.

### 2.2 Software and Background Knowledge

Here, we consider teaching the course with and without an introductory statistics pre-requisite.

#### 2.2.1 Then – A Reflection on the Course with No Pre-requisites

The first two times we taught the course, students enrolling in it were not expected to have background statistical or linear algebra knowledge, or any programming experience. In other words, they were being introduced to these concepts without pre-requisites. The first two weeks of the course were devoted to covering background material to give students a common frame of reference and be introduced to the software.

With no pre-requisites required, we were aware that our students were not likely skilled programmers, and most would not have seen the R software previously. This motivated us to use the graphical interface The R Commander for R (Fox 2005). The interface for The R Commander is menu-driven and can be a comfortable starting point for students, while still allowing them to learn the script language. Eventually, all students in the course had to use R script for some of the multivariate techniques. We also used the additional free software Ggobi to aid in data visualization and exploration (Swayne et al. 2003). The first two weeks in the computer lab were devoted to learning ways of visualizing multivariate data, covering basic univariate and multivariate graphical displays, and using the software.

We covered other background material in the classroom in those two weeks of class. This included some statistics (types of variables, regression basics, ideas on variable selection and transformation), and some linear algebra (matrix notation, scaling effects, eigen decomposition, and singular value decomposition). Overall, we believe the course as taught with no pre-requisites was a useful and valuable experience for these students, and the course reached a

broader audience than would have been possible with pre-requisites. For example, the first time the course was taught, only half of the 19 enrolled students would have satisfied an introductory statistics pre-requisite.

### 2.2.2 Now – The Course with an Introductory Statistics Pre-requisite

In Fall 2014, major changes to the course were made. It was taught with a pre-requisite of introductory statistics, and was designed to fill the role of an elective course in our new statistics major. The change to the pre-requisites meant that many students came into the class familiar with R, as we use R with RStudio (RStudio 2011) as the interface in our introductory courses. Students were also already comfortable with basic statistical concepts, such as types of variables and univariate figures. We were able to capitalize on the student knowledge by going further with the material in several modules. For example, we used the lattice (Sarkar 2008) and ggplot2 (Wickham 2009) packages in R for visualization in that module, rather than the Ggobi software. We spent the first two days of class doing review activities via RMarkdown (RStudio 2014) to help students remember their R (or get a gentle introduction to it, if they were coming straight from AP statistics or had some other background in statistics). We still covered some linear algebra material as background information, but overall the first few weeks of class (introduction and visualization module) were radically different. Building on student's computing knowledge and basic statistics background made the visualization module much more enjoyable, and the students were able to tackle much more in that regard than in previous semesters.

The statistics pre-requisite knowledge was also useful when topics such as hypothesis testing (in factor analysis) or the multivariate normal distribution arose. We did not have to introduce students to the basic concepts of hypothesis testing, or lay out the specifics of the

univariate normal distribution before building up to the multivariate case, as they were assumed to have knowledge in these areas.

### 2.3 Course Topics, Principles, and Module Structure

The course topics (listed in Table 1) were covered in a module structure.

**Table 1**: Course Topics/Modules

| |
|---|
| I: Data Exploration and Visualization; Linear Algebra Background Material |
| II: Principal Components Analysis |
| III: Factor Analysis – primarily exploratory, some discussion on confirmatory |
| IV: Multidimensional Scaling – metric and non-metric |
| V: Clustering – hierarchical methods and K-means |
| VI: Classification – classification trees and random forests, nearest neighbor methods, linear and quadratic discriminant analysis, support vector machines |

The textbook for the class readings was Everitt and Hothorn's (2011) *An Introduction to Applied Multivariate Analysis with R*, which was sufficient for all topics except VI: Classification, which was supplemented with other notes. In previous semesters, students used Lattin, Carroll, and Green's (2003) *Analyzing Multivariate Data* text. The topics included are all frequently used multivariate data analysis techniques, covered in most multivariate texts. For some topics, theoretical depth was limited due to the level of linear algebra taught in the course. Students have natural intuition about some of these techniques (for example, spam filters classify emails as spam or not). Each module received 2-2.5 weeks of class time, and was followed by a homework assignment. The daily outline of a typical module is provided in Table 2.

We lectured to convey basics about the topic, at least one example (in R), considerations (such as tuning parameters) and relationships to other topics. Complete code for the examples

**Table 2**: Typical Module Outline

| | |
|---|---|
| Prep | Chapter/Sections Assigned from Text |
| Day 1 | Motivational Example(s) with New Data Set to Explore |
| Day 2 | Procedure Basics |
| Day 3 | Procedure Basics and Other Example(s) |
| Day 4 | Apply New Technique(s), Class Discussion |
| Day 5 | Settings and Connections to Other Topics (e.g. choosing tuning parameters, cross-validation procedures, bootstrap) |
| Day 6 | Further Applications of Technique(s) |
| Day 7 | Application Day(s) |
| Day 8 | More Applications, Discussion, Homework Assigned |

from class was provided via RMarkdown files. One day in each module was an "application day" which was devoted to examining applications of the technique in published work. This day took on several forms, with all three of the following formats used in Fall 2014. In the first format, we prepared slides showing key points from articles using the techniques. By showing real published articles and having students walk through the analysis as a group, students could show what they had learned about the topic so far, help each other, and learn to critique statistical work. In another format, we randomly divided the students into groups, each of which had a task to complete relative to the topic at hand and based on a published analysis article, with group presentations. In a final format, which the students seemed to prefer, we brought about 10 articles that included relevant methods to class, from a variety of subject areas, and students read through them to report to the class (in a short informal presentation) about how the methods were applied in the article, as well as any critiques they had about the application, etc. This could

either be done in one or two class periods, and the reports could be made formal as well. Appendix B in the Supplemental Materials contains a list of articles used for the Clustering Module.

For every topic, students explored the techniques using their laptops and examples in R/RMarkdown via RStudio. These "lab" activities were self/group-paced with instructor availability by request to help with any issues. We carefully designed the labs using RMarkdown, with some code provided and some not, and had flexibility for students to choose various options relevant to the techniques. The activities also provided a time to explore issues not covered in lecture and for students to learn about the challenges of real data analysis (Singer and Willett 1990). For example, we spent time discussing outliers and missing values, as well as how the computer software treated those values for each technique.

For each module, we wanted to keep an emphasis on data visualization, use real data for all examples, incorporate active learning activities, and use a variety of learning methods. Labs always began with data visualization and exploratory analysis, so the students were exploring the data before applying the new techniques they were learning. Active learning was incorporated in some lectures where students helped guide the discussion, as well as in most labs, where students worked through examples while addressing various questions. Finally, students had access to a variety of learning methods - readings from a text (Everitt and Hothorn 2011) and published articles, homework and a course project, class discussions, activities with exploratory components, with opportunities to share with classmates.

The data sets used for class and lab examples were a mixture of classic, tidy data sets and data sets requiring some data management (for example, requiring subsetting, having missing data coding problems, or needing to create new variables), from a variety of disciplines including

(but not limited to) biology, music, restaurant management, economics, medicine, chemistry, real estate, sociology, and anthropology. To summarize the data experiences of the students in the class based on one module, we constructed Table 3 based on Grimshaw's framework (Grimshaw 2015) and our clustering module. The framework divides the skills needed for analysis into "Data from Different Sources and Formats" and "Data Management". In each of these groups, three categories are defined to distinguish the depth of skill necessary - "good", "better", and "best". (For details on the breakdown, see Grimshaw (2015)). The coursework achieved the "better" and "best" categories for dealing with data in different courses and formats mostly during course projects which involved clustering, but the "better" category for management was obtained via the creation of new variables, requiring subsetting, and working with outliers in different examples and assignments (see the example homework assignment in Appendix C). Again, this table summarizes the experiences in a single module from the course. Clearly, we can work more on incorporating data in different sources and formats into the course.

**Table 3:** Clustering Module Data Experiences

| | Sources and Formats | | |
|---|---|---|---|
| Management | Good | Better | Best |
| Good | 7 | 1 | 0 |
| Better | 4 | 0 | 0 |
| Best | 0 | 1 | 1 |

As a final note about the data in the course, only one artificial (i.e. generated with context "added") data set was used throughout the entire course, so students were working with real data for every technique. Simulated data sets were used in some concept illustrations, but were clearly denoted and had no context added. (A quick note about that solitary artificial data set: we

designed it so that using scatterplots would result in letters being spelled out in the display, which could be properly arranged to spell a word. This data set was designed to reinforce the power of data visualization. For details, contact the author.)

**2.4 Assessment**

Designing assessment for this course was a challenge. Statistical data analyses can take many correct "paths", and we wanted that to be clear to the students. We also wanted to give the students good feedback on the expressions of their ideas, and offer several ways to convey their ideas. The breakdown of assessment tools that we decided on for the course is shown in Table 4. (Previous iterations had a take home final exam and less emphasis on the project).

**Table 4:** Assessment Tools

| Tool | # of Assignments/Duration | % of Final Grade |
|---|---|---|
| Homework | 7 – about 1 every 2 weeks, one per module, plus an R review | 20% |
| Midterms/Quiz | 2 Midterms – Concepts I-III, IV-V,1 Quiz - Concept VI | 20% /20%/10% |
| Final Project | Final four weeks of the semester; groups of 2-3 students | 25% |
| Participation | 4 major group activities assessed | 5% |

Each course concept had a related homework assignment with two portions, a theoretical portion on the technique and a data analysis portion, which was an RMarkdown submission of several pages of written analysis of a data set with supporting work and figures included which addressed a few questions we asked. The students grappled with homework with substantial decisions (which variables to include, for example), and they had to produce and provide appropriate supporting work for their analysis and solutions. For each assignment, we determined a basic rubric to grade on the statistical content, and while we commented on their writing and made corrections, we did not include a major writing assessment on homework. We

designed the rubrics so all data analysis assignments were a similar point value, giving points for preliminary analysis, appropriate technique choice and discussion of relevant option choices (for example, variable selection, rotation choice, or number of PCs/factors/clusters), as well as addressing any specific questions asked.

Midterms and the quiz were held in class and consisted of several problems each with multiple parts (with perhaps shared data sets). We supplied all results (figures and output), so the students were interpreting the results or picking between different options with explanations for their choices, and not applying techniques themselves. (Thus, they were not tested on their ability to code quickly). We did not write exam questions where students were given a problem and then had to suggest an appropriate analysis. Instead, midterms focused on students demonstrating understanding of each technique, reading and interpreting the output. We relied on the course project to give better insight into student abilities to apply appropriate techniques.

The course projects were group projects where students applied at least two of the main techniques from class on a data set (based on a data set they found or constructed) after brainstorming some questions of interest that they would try to answer. There are many websites that make data freely available and others that list collections of those sites. One of our favorite sites to find data is the University of California at Irvine's Machine Learning Repository (Frank and Asuncion 2010), and there are many other similar sites (too numerous to list here). Readers interested in additional sites used in the course should contact the author.

For their projects, students had to submit a proposed set of questions and receive approval before delving into their data analysis. They also had to introduce the class to a new statistical topic relevant to their analysis. This topic was tailored to each group with the difficulty level tuned to that group. For example, one group learned to code a bootstrap analysis in a new

setting, while another group learned about support vector machines and implemented one. The groups made class presentations to share their findings with the class during the last week of classes. For their final submission, groups summarized their results in a 10-12 page paper, with a technical appendix in RMarkdown.

This approach to the projects was a change from individual projects we had used in previous semesters. Overall, we felt the group projects were better quality than the individual projects from previous semesters. Course projects are a significant time investment on the part of the instructor - whether they are group or individual, but are very rewarding in our experience.

To demonstrate how all these course design aspects came together in the course, we present the clustering module in the next section.

# 3. AN EXAMPLE MODULE ON CLUSTERING

The clustering module was the fifth module in the course. At this point in the course, students were also brainstorming for their projects, and clustering gives them many possible techniques to explore that we do not cover in class. Briefly, we present the days in the module stating what happened in class and give some additional comments. We also describe the clustering related assessments used in the course. Many materials described here are provided in the Supplementary Material. Additional materials may be provided by request.

## 3.1 Clustering Module

### 3.1.0 Prep

What: Students were assigned to read the clustering chapter from the text with sections spread out over the course time on the module based on class progression.

Comments: If assigned, reading was usually 1-3 sections per day from the text. There were days with no reading assigned when the students were focused on an activity or application (where they might have to read an article).

### 3.1.1 Introduction to Clustering - Lecture 1

What: Introduction and motivation of the technique using interactive examples and data from our research, followed by simple examples for the students to see. The basics of linkage were discussed for hierarchical solutions.

Comments: Students worked in pairs to cluster 18 animals into two groups. Their results were shared with the class. This brought about a wonderful discussion of dissimilarities due to the wide variety in the solutions.

### 3.1.2 More on Clustering - Lecture 2

What: Continued developing understanding of hierarchical clustering and linkage with simple examples. Advanced linkages were discussed, as well as choice of dissimilarity. We began discussion of K-means.

Comments: The students reviewed a single linkage example and then tried a complete linkage example on their own, before the class moved on to K-means.

### 3.1.3 Clustering in R - Lab 1

What: Finish discussion on K-means and move into lab activity where students explored clustering solutions performed with R-code.

Comments: We designed this lab to familiarize the students with performing clustering in R based on all the code and examples in their text. Much of the R code is provided, though students have to make adjustments or code up their own solutions as they go – but example code is always close by. (See Supplemental Material - Appendix A for this lab activity.)

### 3.1.4 Clustering in R – Lab 2

What: Students continued working on the introductory lab activity. When completed, they had access to a second lab activity where clustering was being used to look for groups of crabs and decide if those groups matched sex or species of the crabs.

Comments: The second lab was designed to show students that the "natural groups" clustering finds in data may not match any categorical variables in the data set. The class had a discussion about stripe plots to be sure that everyone could understand them using those plots presented in Lab 1. Not all students made it through Lab 2, so they were encouraged to complete this on their own and check their responses versus our posted comments.

### 3.1.5 Clustering Challenges - Lecture 3

What: An RMarkdown (compiled) file was provided that showed various challenges in clustering. For example, a situation where K-means fails to find the "correct" three clusters expected based on two starting points falling in one large cluster in a simple two-dimensional setting was shown. The class walked through the different examples with guidance.

Comments: This series of examples was very useful to show students that different clustering techniques work well in different situations, but that there may be issues to be aware of. The K-means example in particular shows the importance of starting points of algorithms, and the students gained understanding about why it is necessary to randomize those starting points and consider multiple solutions.

### 3.1.6 Application Day – Lecture 4

What: Application day. Eleven (semi-short) articles were prepared for class discussion. The articles were applications of clustering in various disciplines and contexts. Students read/skimmed the articles in pairs or triples (eight groups) and presented some findings about

how clustering was used in the articles to the class in informal presentations. (See Supplemental Material – Appendix B for article references and class handout.)

Comments: Application days were discussed in Section 2.3 in some detail. The students seemed to prefer this setup – groups worked with an article and presented findings briefly to the class. Many variants on this activity are possible. Our idea was to bring current (year 2000+) examples of applications to the students so they could really appreciate the technique they were learning.

**3.2 Clustering Assessment**

As a main course topic, clustering had several related assessment materials - a homework and exam question (second midterm), and could also have been used for the course projects. The clustering homework (as with all the homeworks) had two associated components. First, there was a theoretical component where the students demonstrated their understanding of linkage and answered some questions. Then, they had an applied data analysis component. For this component, students had to do some preliminary data exploration, perform an appropriate clustering analysis, and explain what their analysis revealed. The clustering homework and the exam question on clustering are shown in the Supplemental Materials – Appendices C and D.

Clustering played a major role in several of the course projects – four of the nine groups planned to use it as one of their two required techniques in their original project proposals. Two groups used different aspects of cluster validation as their expository topics to share methods/ideas with the class that were not covered in the text. One group used clustering to identify natural groups of observations based on one subset of variables and then tried to follow up with classification techniques on a different subset of variables to see if the clusters could be recovered with different data (though this did not work unfortunately). With the course projects

and presentations, students got to see firsthand how different techniques worked well for some data sets and associated questions and not for others.

# 4. OUTCOMES AND COURSE EVALUATION

Recall that the goals we had outlined for the course were for students to be introduced to complex data sets (with more variables than usually presented in an introductory course), and gain an applied understanding of exploratory data analysis and several multivariate analysis techniques. This section provides some assessment of the course and whether the goals were achieved based on student course evaluations and short survey replies. Course evaluations were completed before the last day of classes in each semester the course was taught. The student evaluation included questions designed to generate feedback on the course and various course aspects, such as the application days, lab activities, and software. Over the three semesters the course was taught, the evaluation completion rate was around 80%. The short survey was sent to students who took the course the first time it was offered to obtain feedback nearly two years after that course offering (18 of 19 students in the first cohort were sent the survey, and six replied).

## 4.1 Comments on Course Aspects and Related Critical Comments

### 4.1.1 Class and Lab Activities

The class and lab activities were a crucial course aspect where students practiced techniques, worked with other students, and learned to code in R. Over the course of the semester, the activities got the students thinking about how data is stored and shared, the importance of appropriate visuals in data analysis, and how vital it is to understand the problem at hand, as well as the techniques being used. Most activities had an active-learning style and the students used RMarkdown via RStudio (RStudio 2011) for their analyses. Student comments

about the activities and software were overwhelmingly positive. Students enjoyed the activities, felt they were vital to their learning process about the techniques, and felt the software was not a deterrent to learning. (It was recommended that we create a master RMarkdown file of coded examples for reference - a good project for the future!). Prior to using R with RMarkdown, a few students had commented that they'd like to see other software used, such as Excel or Stata, because they were already familiar with those programs from other courses. However, we think the benefits from R (free, very nice interface in RStudio, access to packages, and reproducibility in RMarkdown) will continue to make it a valuable software choice.

### 4.1.2 Application Days

Another major feature of the course was the application day(s) in each module, and we asked the students what they thought of these days on their course evaluations. Each semester the course has been taught, these responses have been very positive. Recall that there were multiple formats where the students would be exposed to actual applications of the techniques in journal articles. A few students commented they did not prefer the format most often chosen by the class (majority vote), but still felt it had been useful to see the applications. The informal presentation aspect was new to the course this past fall, and several students wanted to formalize that, which likely would have meant spending at least two days on this component each module. With fewer modules, such an approach might work well, and it would give students more time to delve into the articles and practice their communication skills.

### 4.1.3 Course Accessibility

When the course is taught without a pre-requisite, perhaps the most important course aspect to consider is whether or not the material was accessible to all the students, and if they felt that they had gained good understanding of the techniques we had studied, no matter their

background. To address these questions, we can examine responses from the first two cohorts of students in the class, as well as the survey results. In particular, views on the level of theory in the course were requested on course evaluations. Many students were quite content with the applied nature of the course, and felt they got a lot out of the course. They voiced this in their final course evaluations as shown in the example replies that follow. One student stated that "It's nice to have another math class that is more hands on and has more application rather than theory based courses." Another stated: "It was good to take an applied class that I felt like had real life applications. I am very glad I took this class; it opened my eyes to ways I could use math in my future without just studying theory." (The course was listed as a Math course for two offerings).

A few students from the first two semesters commented that they would have liked to have seen more theory, and that requiring additional background knowledge might be necessary to do that. From the short survey, one student stated that he felt a deeper linear algebra background would have benefited him while learning about the techniques. Even in the third course offering, with the introductory statistics pre-requisite, two of the 21 enrolled students commented that they felt students who had linear algebra as well as intro stats would likely have gotten more out of the course than those without the linear algebra. These comments suggest that there is definitely a need to balance the desire for theory with student background, and that this is trickier when the course is taught without pre-requisites. For some recent discussion about "flattening" pre-requisites, see Cobb (2015).

## 4.2 What Did Students Do with the Knowledge They Gained in the Course?

One of our main motivations for sending the short survey out to course alumni from the first course offering was to find out how they had used the knowledge they gained in the course. We had limited feedback on this from course evaluations at the end of the semester. In this

section, we discuss relevant feedback from course evaluations from all three course offerings and short survey replies from the first cohort.

First, even at the end of the same semester the course was taught, several students had positive things to say about the course and how it had impacted their lives. From one student, we learned that "Seeing this course on my transcript has been a conversation starter during interviews" and this has actually happened several times since. One student listed us as a summer internship reference and we were able to state to the prospective employer that she had used R to perform a principal components analysis (PCA). Some seniors writing theses concurrently with the course used these techniques in their thesis work, and taking the course helped them to understand what they had been reading about and/or doing themselves, as well as seeing the techniques in a range of applications. From their short survey replies, two students reported greatly appreciating learning about data visualization and learning to work with large, real data sets for their current work. Another student reported seeing PCA applied often in her current field. Knowledge of R and some skill in coding has proven beneficial for several students (several have entered statistics graduate programs, and many reported this knowledge was helpful for job interviews and future work).

Finally, the course was important to several students due to its applied nature and course content. Students commented that the practical application was a nice contrast to theoretical courses (in our mathematics curriculum). Based on their evaluation comments and survey responses, students used the knowledge gained in the course in a variety of settings (other classes, projects, theses, internships, graduate school, and jobs), and appreciated their undergraduate applied introduction to the topics.

## 5. DISCUSSION

Briefly, we discuss some considerations for the course and potential research areas for the future. First, there are some considerations related to the teaching approaches used in the course. Each time we have taught the course, the students were amenable to group work, but we did not use this to its full advantages. We hope to incorporate more of the suggestions from Garfield's work on active learning with cooperative groups in future courses (Garfield 1993). Students might also benefit from more interaction with figures, though we feel the course in its third offering did much more on visualization than previous offerings. While some data skills were covered in the course, additional data wrangling skills would also be useful to incorporate before projects begin, in particular to assist students dealing with data sets from multiple sources requiring merging.

Second, there is the major consideration of teaching the course with or without a pre-requisite. We would like to emphasize that the course can be successful with or without pre-requisites. The relevant pre-requisites to consider are introductory statistics and linear algebra. The linear algebra pre-requisite would lock many students out of the course, so we will focus discussion here on the introductory statistics course as a pre-requisite. Our best advice is to consider the audience for the course. If the audience is statistics students (especially in a statistics program), it might make sense to offer the course as an elective after a pre-requisite of introductory statistics. If the audience is broader, then, in the spirit of Cobb (2015), perhaps offering the course with no pre-requisite is a better option. Another important issue to bear in mind is how much computation the students will be responsible for. The course can be taught with less emphasis on the computation, with more provided examples. Finally, if teaching the course with no pre-requisites and trying to cover as many methods as described herein seems daunting, cover fewer topics! We can envision a version of the course with no pre-requisites that

focuses on visualization, clustering, and classification, with a computational level that can be adjusted (or uses a slow pace to build up skills) that would still be beneficial for the students. In short, the core issue of the pre-requisite really depends on the student audience of the course.

A few areas for future pedagogical research are suggested by this case study. The use of real analyses from published articles as examples to help students learn techniques bears further investigation. Additionally, investigating active learning methods in statistics education in courses beyond introductory statistics will likely be enlightening, as we strive to help modern students learn the statistical concepts that they need to know in the real world.

**REFERENCES:**

American Statistical Association Undergraduate Guidelines Workgroup (2014), "2014 Curriculum Guidelines for Undergraduate Programs in Statistical Science." Alexandria, VA: American Statistical Association. http://www.amstat.org/education/curriculumguidelines.cfm, last accessed May 5, 2015.

Breiman, L. (2001), "Statistical Modeling: The Two Cultures," *Statistical Science*, 16 (3), 199-231.

Brown, E. and Kass, R. (2009), "What is Statistics?," *The American Statistician*, 63 (2), 105-110.

Carlson, K. and Winquist, J. (2011), "Evaluating an Active Learning Approach to Teaching Introductory Statistics: A Classroom Workbook Approach," *Journal of Statistics Education* [Online], 19 (1).

Chance, B., Ben-Zvi, D., Garfield, J., and Medina, E. (2007), "The Role of Technology in Improving Student Learning of Statistics," *Technology Innovations in Statistics Education* [Online], 1 (1). Available at: http://escholarship.org/uc/item/8sd2t4rr

Cobb, G. (2007), "The Introductory Statistics Course: A Ptolemaic Curriculum?," *Technology Innovations in Statistics Education* [Online], 1 (1). Available at: http://escholarship.org/uc/item/6hb3k0nz

Cobb, G. (2015), "Mere Renovation is Too Little Too Late: We Need to Rethink our Undergraduate Curriculum from the Ground Up," *The American Statistician*, 69 (4), 266-282.

De Veaux, R. (2015), Opening Presentation Slides from the USCOTS 2015 conference (United States Conference on Teaching Statistics).

Everitt, B., and Hothorn, T. (2011), An Introduction to Applied Multivariate Analysis with R. Springer, New York.

Fox, J. (2005), "The R Commander: A Basic-Statistics Graphical User Interface to R," *Journal of Statistical Software*, 14 (9).

Frank, A., and Asuncion, A. (2010), UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Available at: http://archive.ics.uci.edu/ml/index.html

Franklin, C., and Garfield, J. (2006), "The Guidelines for Assessment and Instruction in Statistics Education (GAISE) Project: Developing Statistics Education Guidelines for Pre K-12 and College Courses," In G.F. Burrill, (Ed.), Thinking and Reasoning about Data and Chance: Sixty-eighth NCTM Yearbook (pp. 345-375). Reston, VA: National Council of Teachers of Mathematics. Available at: http://www.amstat.org/Education/gaise/GAISECollege.htm

Garfield, J. (1993), "Teaching Statistics Using Small-Group Cooperative Learning," *Journal of Statistics Education* [Online], 1 (1).

Garfield, J., and Ben-Zvi, D. (2007) “How Students Learn Statistics Revisited: A Current Review of Research on Teaching and Learning Statistics,” *International Statistical Review*, 75 (3), 372-396.

Gould, R. (2010) “Statistics and the Modern Student,” *International Statistical Review*, 78 (2), 297-315.

Grimshaw, S. (2015) “A Framework for Infusing Authentic Data Experiences Within Statistics Courses,” *The American Statistician*, 69 (4), 307-314.

Lattin, J., Carroll, J., and Green, P. (2003), Analyzing Multivariate Data, Cengage Learning, Brooks/Cole.

Moore, T., Peck, R., and Rossman, A. (2000), “Statistics: CRAFTY Curriculum Foundations Project.” Chapter 14 in Curriculum Foundations Project: Voices of the Partner Disciplines from the Committee for the Undergraduate Program in Mathematics (CUPM). Available at: http://www.maa.org/cupm/crafty/Chapt14.pdf

Nolan, D. and Perrett, J. (2015), “Teaching and Learning Data Visualization: Ideas and Assignments,” *arXiv:1503.00781v1* [stat.OT].

Nolan, D. and Temple Lang, D. (2010), “Computing in the Statistics Curricula,” *The American Statistician*, 64 (2), 97-107.

O'Shea, D., and Pollatsek, H. (1997), "Do we need pre-requisites?," *Notices of the AMS*, 44 (5), 564-570.

R Development Core Team. (2009), "R: A language and environment for statistical computing," *R Foundation for Statistical Computing*, Vienna, Austria, ISBN 3-900051-07-0.

RStudio. (2011), "RStudio, new open-source IDE for R," RStudio Blog. http://blog.rstudio.org/2011/02/28/rstudio-new-open-source-ide-for-r/, last accessed July 15, 2015.

RStudio. (2014), "R Markdown v2," RStudio Blog. http://blog.rstudio.org/2014/06/18/r-markdown-v2/, last accessed July 15, 2015.

Samsa, G., Thomas, L., Lee, L., and Neal, E. (2012), "An Active Learning Approach to Teach Advanced Multi-predictor Modeling Concepts to Clinicians," *Journal of Statistics Education* [Online], 20 (1).

Sarkar, D. (2008), Lattice: Multivariate Data Visualization with R. Springer: New York, ISBN: 978-0-387-75968-5.

Singer, J., and Willett, J. (1990), "Improving the Teaching of Applied Statistics: Putting the Data Back Into Data Analysis," *The American Statistician*, 44 (3), 223-230.

Snee, R. (1993), "What's Missing in Statistical Education?," *The American Statistician*, 47 (2), 149-154.

Swayne, D., Temple Lang, D., Buja, A., and Cook, D. (2003), "Ggobi: Evolving from XGobi into an Extensible Framework for Interactive Data Visualization," *Computational Statistics and Data Analysis*, 43 (4), 423-444.

Tintle, N., VanderStoep, J., Holmes, V., Quisenberry, B., and Swanson, T. (2011), "Development and Assessment of a Preliminary Randomization-Based Introductory Statistics Curriculum," *Journal of Statistics Education* [Online], 19 (1).

Valero-Mora, P. and Ledesma, R. (2011), "Using Interactive Graphics to Teach Multivariate Data Analysis to Psychology Students," *Journal of Statistics Education* [Online], 19 (1).

Wickham, H. (2009). ggplot2: elegant graphics for data analysis. Springer: New York.

**Supplemental Materials**

The curriculum materials contained in this Supplemental Material are copyrighted by Amy Wagaman and distributed under a Creative Commons BY-NC-SA license**.**

**Appendix A – Lab Activity: Introduction to Clustering in R**

This lab was designed to show the students the code related to clustering covered in class and their textbook. It has questions for them throughout, but focuses on demonstrating code with an example. The material shown here is from a Word document formatted from an RMarkdown file for the lab. A solution was posted for the students.



---

# Intro to Clustering in R

The cluster library is the main one we need for clustering. Some others are used here to show you some other features.

```
require(mosaic)
require(cluster) #for access to daisy function
```

## Clustering Examples

The goal here is for you to run some pre-setup cluster analyses, in order to see how the code works and look at some examples in R. You will be requested to adjust some code as you run the examples.

## Cereals

```
cereals = read.table("http://www.amherst.edu/~awagaman/stat330/cereals.txt",h=T)
```

The data is cereal brands, manufacturers (also group - same info, but group is numeric, and manufacturer is categorical), and nutrition information (calories, protein, fat, sodium, fiber, carbs, sugar, potassium) per serving.

```
summary(cereals)
```

```
##         brand    manufacturer    calories      protein          fat
##  ACCheerios: 1   G:17         Min.   : 50   Min.   :1.00   Min.   :0.000
##  AllBran   : 1   K:20         1st Qu.:100   1st Qu.:2.00   1st Qu.:0.000
##  AppleJacks: 1   Q: 6         Median :110   Median :2.00   Median :1.000
##  CapNCrunch: 1                Mean   :108   Mean   :2.46   Mean   :0.977
##  Cheaties  : 1                3rd Qu.:110   3rd Qu.:3.00   3rd Qu.:1.500
```

```
##  Cheerios  : 1               Max.   :160   Max.   :6.00   Max.   :3.000
##  (Other)   :37
##      sodium        fiber          carbs          sugar
##  Min.   :  0   Min.   :0.00   Min.   : 1.0   Min.   : 0.00
##  1st Qu.:145   1st Qu.:0.50   1st Qu.:12.0   1st Qu.: 3.00
##  Median :190   Median :1.00   Median :14.0   Median : 8.00
##  Mean   :180   Mean   :1.71   Mean   :14.3   Mean   : 7.61
##  3rd Qu.:220   3rd Qu.:2.85   3rd Qu.:17.0   3rd Qu.:12.00
##  Max.   :320   Max.   :9.00   Max.   :22.0   Max.   :15.00
##
##    potassium         group
##  Min.   : 15.0   Min.   :1.00
##  1st Qu.: 37.5   1st Qu.:1.00
##  Median : 60.0   Median :2.00
##  Mean   : 84.4   Mean   :1.74
##  3rd Qu.:110.0   3rd Qu.:2.00
##  Max.   :320.0   Max.   :3.00
##
```

Be sure to explore the data before jumping into the analyses!

## Hierarchical Clustering

If we want to look for cereal groups via hierarchical clustering, we need to construct a distance matrix. Distances are constructed as in MDS/Isomap, and you again need to choose whether you compute them on scaled or unscaled variables (standardize or not).

```
cer.dist=dist(cereals[,-c(1:2,11)])
```

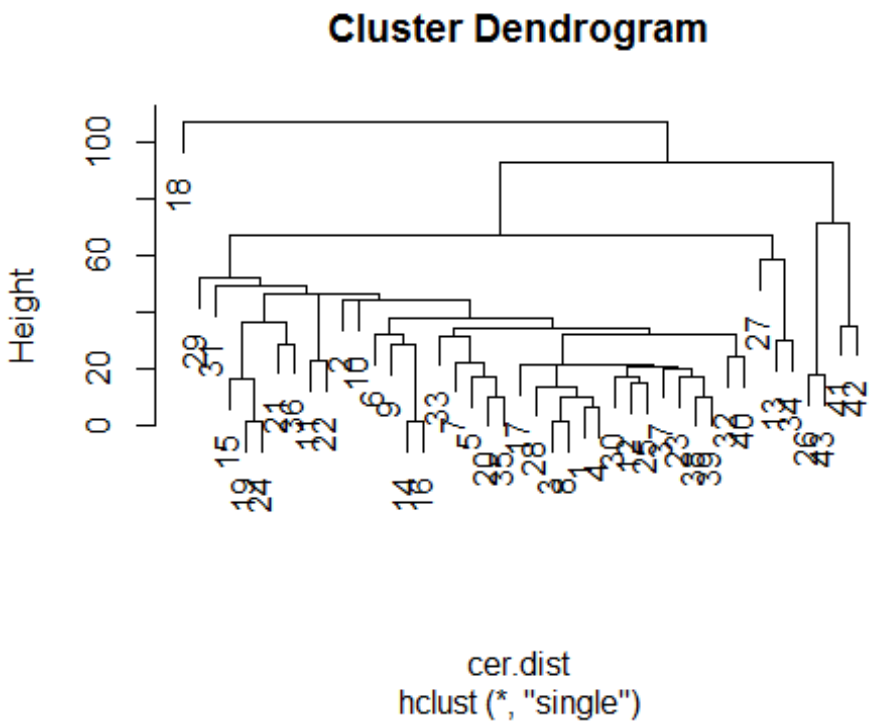
Now we look at how hierarchical clustering is applied. The relevant function is *hclust*.

```
hcsingle=hclust(cer.dist,method="single")
list(hcsingle) # reminds you of properties of the solution, if desired
```

```
## [[1]]
##
## Call:
## hclust(d = cer.dist, method = "single")
##
## Cluster method   : single
## Distance         : euclidean
## Number of objects: 43
```

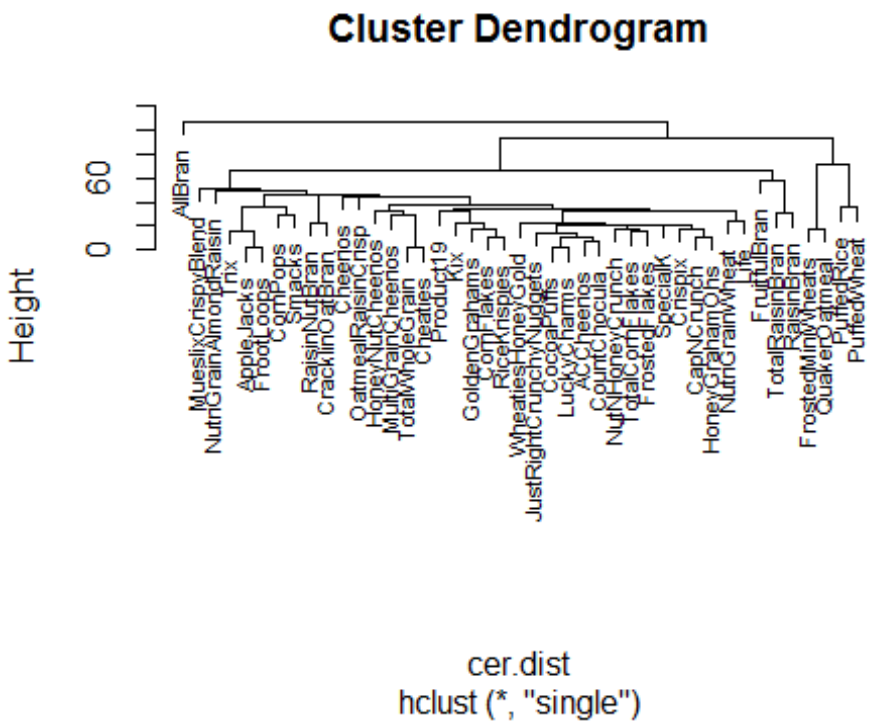
This creates the solution, and we can look at the dendrogram as:

```
plot(hcsingle)
```




The observation numbers show up as the labels. If you want those to be the cereals instead, try:

```
plot(hcsingle,labels=cereals$brand,cex=.7) #cex adjusts size of label
```




The options for hclust in terms of linkages are provided in the help under options for method. The following options are listed: "ward.D", "ward.D2", "single", "complete", "average", "mcquitty", "median" or "centroid".

Run a different hierarchical clustering solution, with a different name, and obtain a dendrogram with the cereal labels.

In order to obtain cluster labels, we need to *cut* our dendrograms.

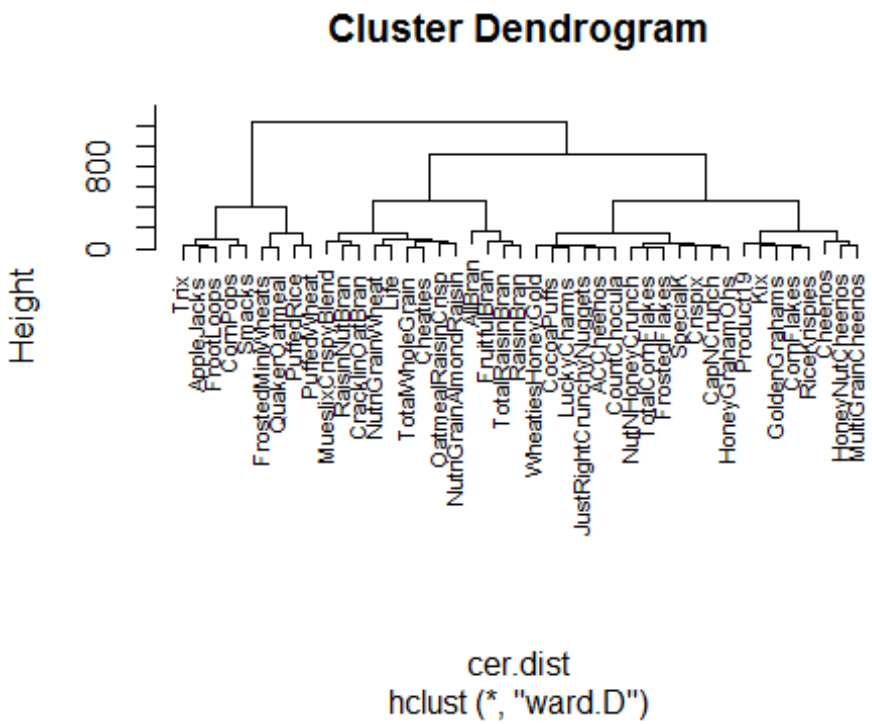
```
singleSol=(cutree(hcsingle, k = 5)) #cluster labels are numeric, k= # clusters
summary(as.factor(singleSol)) #as factor to get table

##  1  2  3  4  5
## 35  3  1  2  2
```

Try cutting your different solutions dendrogram.

To learn more details about the clusters you found, consider this solution and its details below:

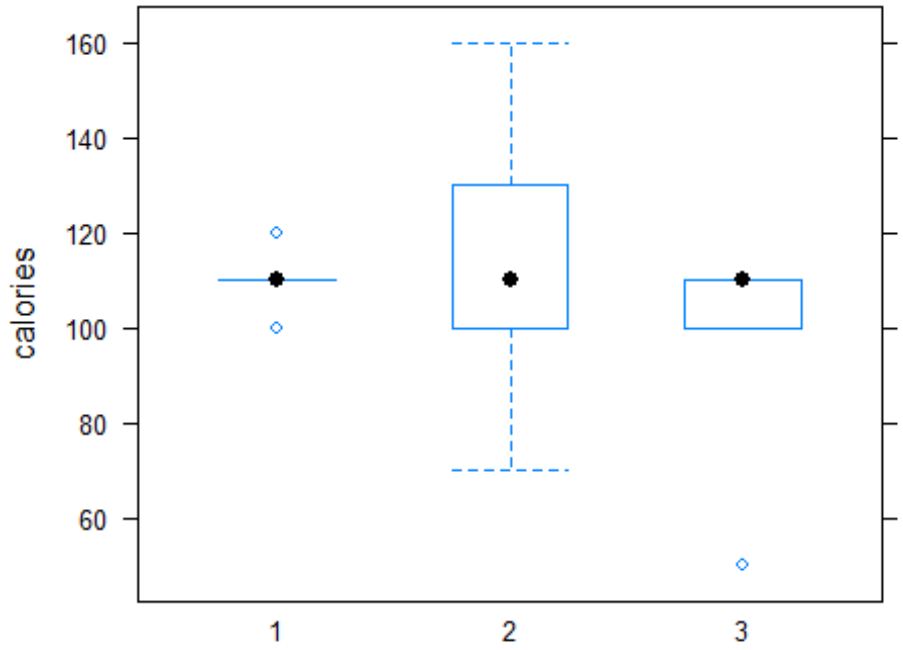
```
hcward=hclust(cer.dist,method="ward.D")
plot(hcward,labels=cereals$brand,cex=.7) #cex adjusts size of label
```




```
wardSol=(cutree(hcward, k = 3)) #cluster labels are numeric, k= # clusters
favstats(calories~wardSol,data=cereals) #can choose any variable

##   .group min  Q1 median  Q3 max   mean     sd  n missing
## 1      1 100 110    110 110 120 110.00  5.477 21       0
## 2      2  70 100    110 130 160 113.85 24.337 13       0
## 3      3  50 100    110 110 110  94.44 25.550  9       0

bwplot(calories~as.factor(wardSol),data=cereals)
```

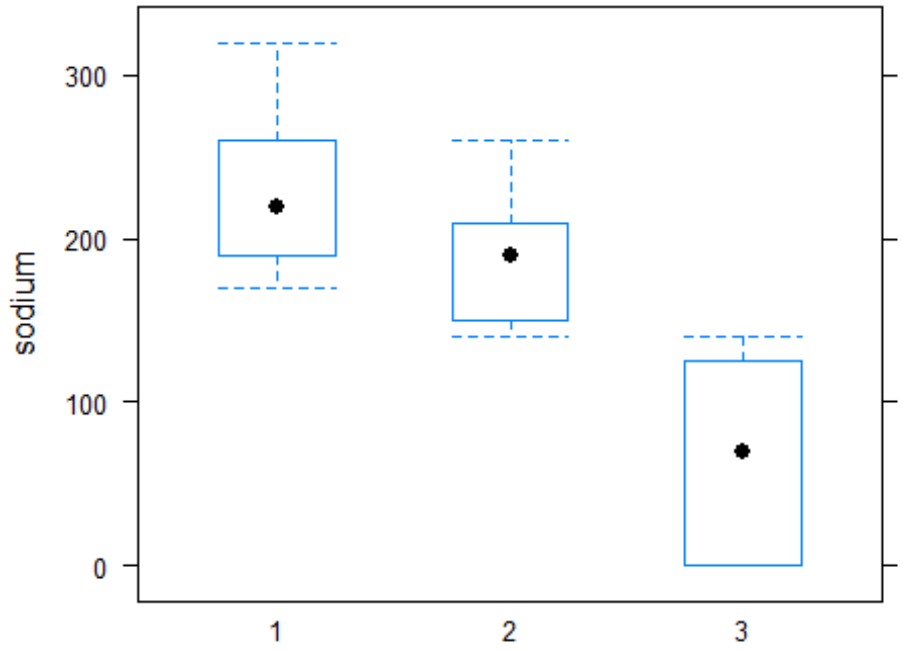



We see basically no difference in caloric content. What if we look at sodium?

```
favstats(sodium~wardSol,data=cereals)
```

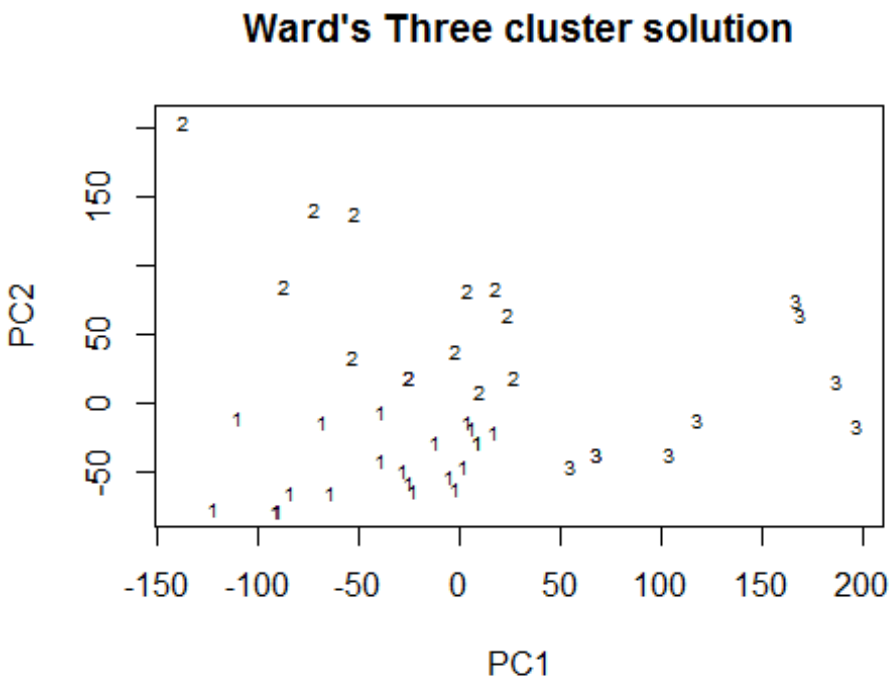
```
##   .group min  Q1 median  Q3 max   mean    sd  n missing
## 1      1 170 190    220 260 320 227.14 45.18 21       0
## 2      2 140 150    190 210 260 187.69 38.55 13       0
## 3      3   0   0     70 125 140  61.11 61.48  9       0
```

```
bwplot(sodium~as.factor(wardSol),data=cereals)
```




We can view the solution in the PC space (say 2-D) to see how well-separated the clusters are in that space. Because we used an unstandardized distance, I will run the PCA on the covariance matrix.

```
cerPCA=princomp(cereals[,-c(1,2,11)],cor=FALSE)
plot(cerPCA$scores[,1:2], type = "n", xlab = "PC1", ylab = "PC2",main="Ward's Three
cluster solution") #blank!
text(cerPCA$scores[,1:2], labels = wardSol, cex=0.6)
```




*Reminder* We used an unscaled distance matrix here. The clustering solution is largely driven by potassium and sodium values as a result. To really incorporate all the variables into the solution, we need to standardize the variables before computing the distance.

## K-means Methods

For k-means, you don't need to compute the distance matrix yourself. You should feed the function the data set to operate on. Here is an example k-means performed on the scaled cereals data:

```
Ksol1=kmeans(scale(cereals[,-c(1,2,11)]),centers=4) #centers is the # of clusters
list(Ksol1) #so you can see what it gives you
```

```
## [[1]]
## K-means clustering with 4 clusters of sizes 16, 10, 13, 4
##
## Cluster means:
##   calories protein      fat   sodium    fiber   carbs   sugar potassium
## 1   0.2092 -0.7388  0.02902 -0.04532 -0.60522 -0.4271  0.8587   -0.5868
## 2   0.5848  0.6015  0.77767  0.08250  1.21495 -0.2010  0.3958    1.4231
## 3  -0.1735  0.3748 -0.45089  0.69331 -0.18304  1.0241 -0.9812   -0.2996
## 4  -1.7348  0.2332 -0.59486 -2.27821 -0.02165 -1.1171 -1.2355   -0.2370
##
## Clustering vector:
##  [1] 1 3 1 1 1 1 3 1 3 2 2 3 2 3 1 3 1 2 1 3 1 2 3 1 1 4 2 3 2 1 2 3 3 2 3
## [36] 1 3 1 1 2 4 4 4
##
```

```
## Within cluster sum of squares by cluster:
## [1] 29.82 64.48 43.23 28.15
##  (between_SS / total_SS =  50.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

The list option provides us with lots of information. You can pull out the cluster means as:

```
Ksol1$centers
```

```
##   calories protein      fat   sodium    fiber   carbs   sugar potassium
## 1   0.2092 -0.7388  0.02902 -0.04532 -0.60522 -0.4271  0.8587   -0.5868
## 2   0.5848  0.6015  0.77767  0.08250  1.21495 -0.2010  0.3958    1.4231
## 3  -0.1735  0.3748 -0.45089  0.69331 -0.18304  1.0241 -0.9812   -0.2996
## 4  -1.7348  0.2332 -0.59486 -2.27821 -0.02165 -1.1171 -1.2355   -0.2370
```

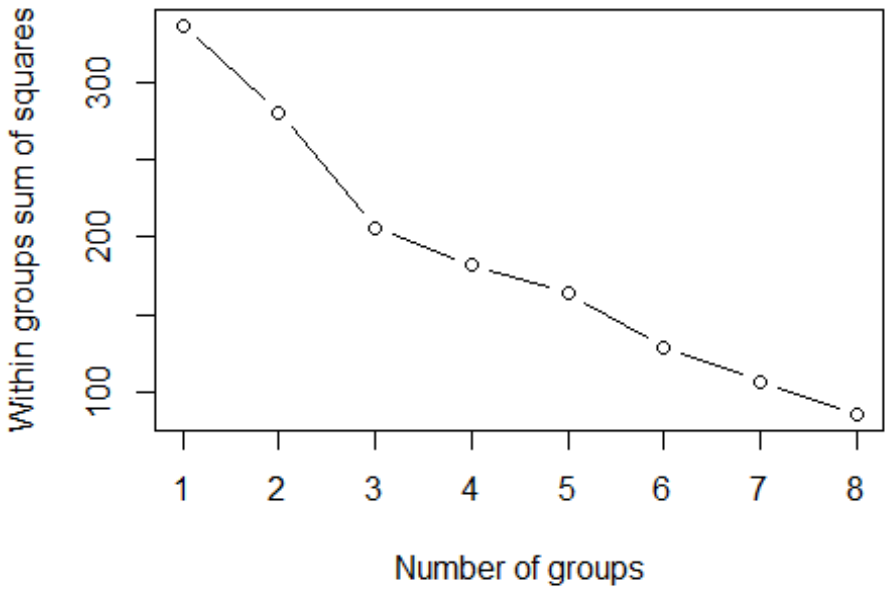
We can also get the clustering vector (with the cluster labels) as:

```
Ksol1$cluster
```

```
##  [1] 1 3 1 1 1 1 3 1 3 2 2 3 2 3 1 3 1 2 1 3 1 2 3 1 1 4 2 3 2 1 2 3 3 2 3
## [36] 1 3 1 1 2 4 4 4
```

In order to determine if we have chosen a "good" value of the number of clusters, we can look at the within cluster sum of squares for this solution and a few other options for k, the number of clusters. This runs the solution from 1 to 8 clusters and pulls the within group sum of squares from each.

```
n <- nrow(cereals) #number of observations
wss <- rep(0, 8) #creates 8 copies of 0 to create an empty vector
for(i in 1:8){wss[i] <- sum(kmeans(scale(cereals[,-c(1,2,11)]),centers= i)$withinss)}
plot(1:8, wss, type = "b", xlab = "Number of groups", ylab = "Within groups sum of
squares")
```




We look for elbows in the plot - here there are elbows at 4 and 6. So perhaps 4 is a decent value.

With four clusters, we should see if there is any relationship with cereal manufacturer.

```
tally(Ksol1$cluster~manufacturer,data=cereals,format="count")
```

```
##               manufacturer
## Ksol1$cluster G K Q
##             1 8 6 2
##             2 3 6 1
##             3 6 7 0
##             4 0 1 3
```

There does not appear to be a strong relationship here.

How do the K-means and Ward's solutions overlap?

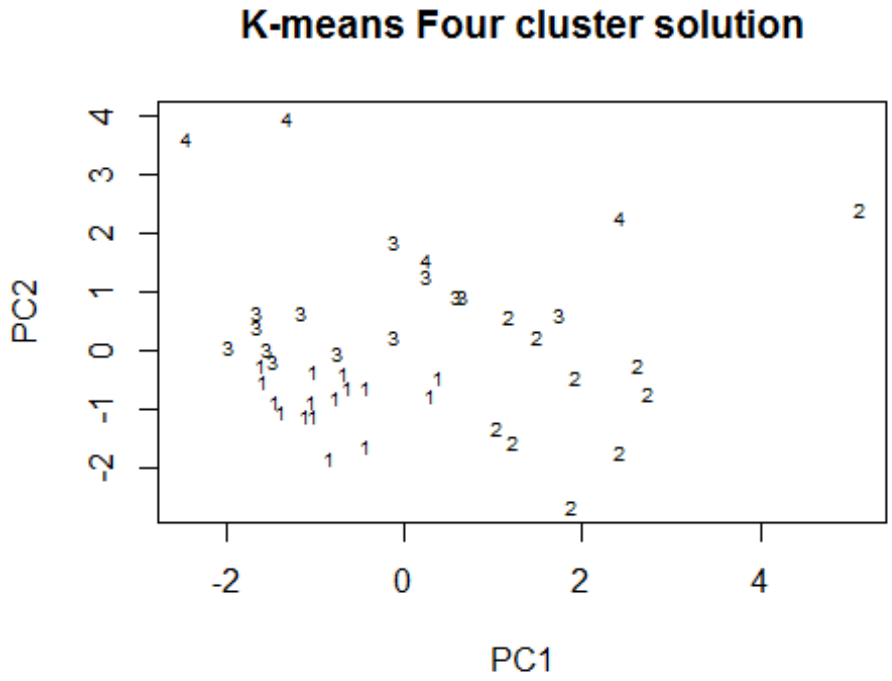
```
tally(Ksol1$cluster~wardSol,data=cereals,format="count")
```

```
##               wardSol
## Ksol1$cluster  1  2  3
##             1 11  0  5
##             2  0 10  0
##             3 10  3  0
##             4  0  0  4
```

What do you see? Bear in mind the Ward's linkage does incorporate some spread information, even though the original distances were not standardized, and this k-means is standardized.

We can plot the k-means solution in PC space as:

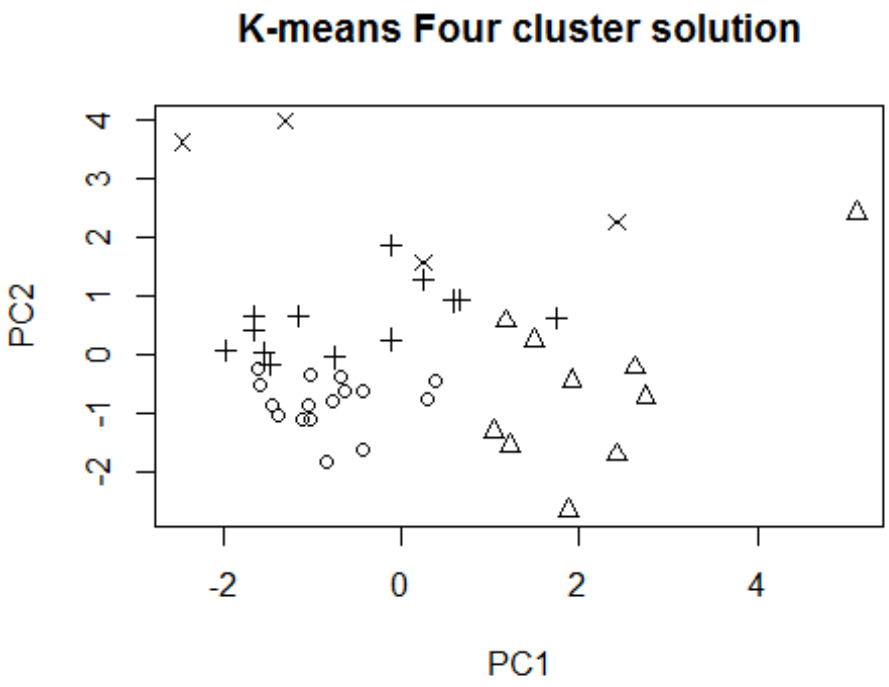
```
cerPCAs=princomp(cereals[,-c(1,2,11)],cor=TRUE)
plot(cerPCAs$scores[,1:2], type = "n", xlab = "PC1", ylab = "PC2",main="K-means Four
cluster solution") #blank!
text(cerPCAs$scores[,1:2], labels = Ksol1$cluster, cex=0.6)
```




Do the 4 clusters appear separated in 2-D PC space? They don't have to be. The separation can exist in higher dimensional space.

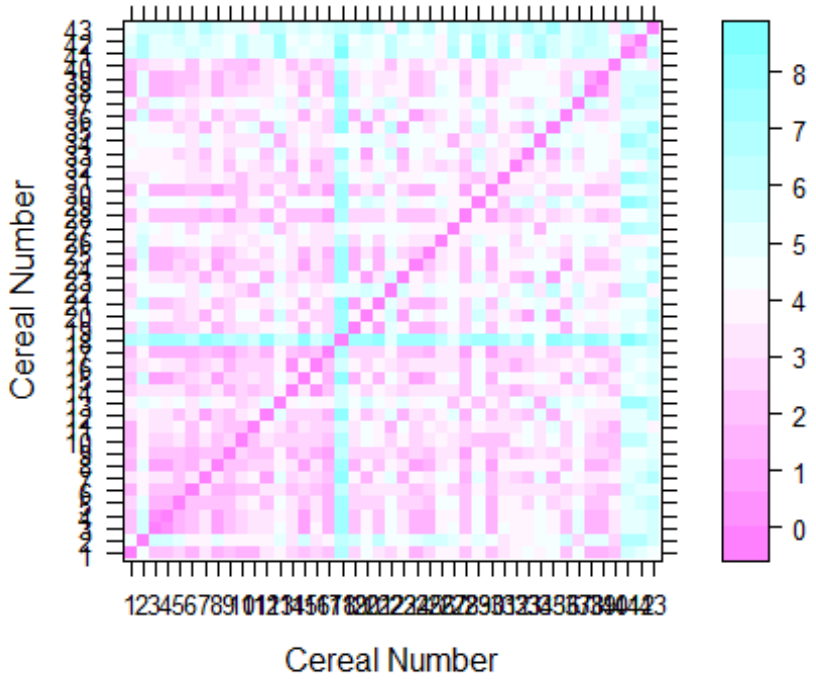
Here is the plot using symbols instead of text labels for the clusters.

```
plot(cerPCAs$scores[,1:2], type = "p", xlab = "PC1", ylab = "PC2",main="K-means Four
cluster solution", pch=Ksol1$cluster)
```




Your textbook shows you an image plot (or heatmap) of the dissimilarity matrix, so here we look at the one for the cereals data. The command requires the lattice library, which should be installed already, so we just require it.

```
require(lattice)
cer.dist.scale=dist(scale(cereals[,-c(1,2,11)]))
levelplot(as.matrix(cer.dist.scale), xlab = "Cereal Number",ylab = "Cereal Number")
```




Where are the dissimilarities 0? Why?

## Additional Material on Hierarchical Methods

You can run some alternative functions than hclust to obtain hierarchical solutions.

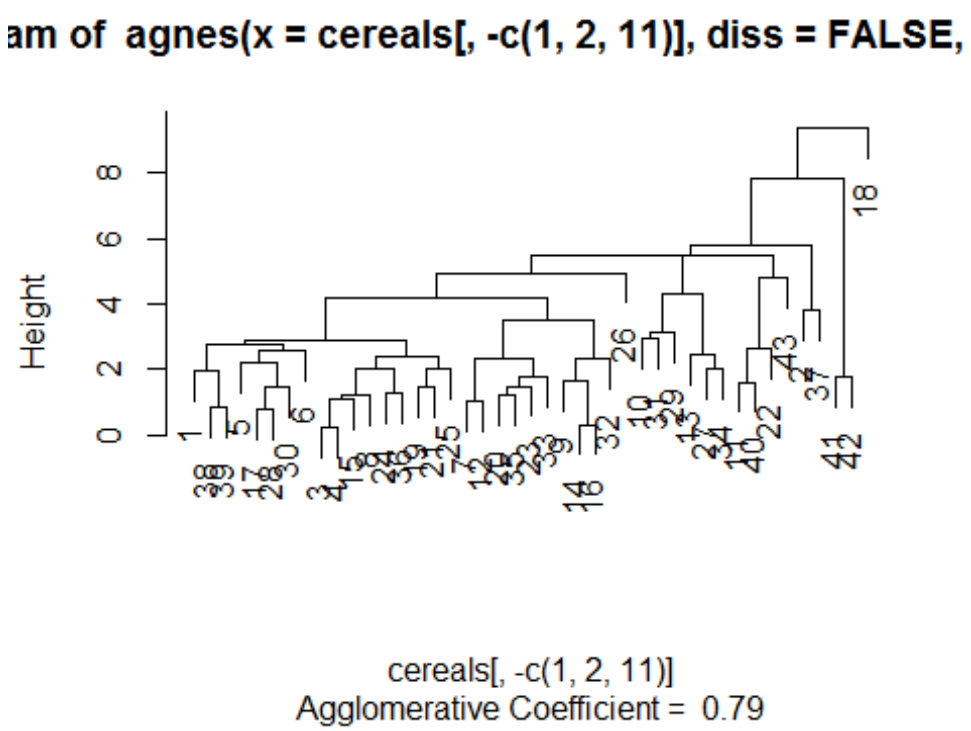
Pull up the help menu for *agnes* to investigate it. Agnes runs agglomerative clustering, can be fed either a data matrix or a dissimilarity matrix, has multiple linkages available, and can do the standardization FOR you. The default is euclidean distance using average linkage but without standardization. Here is an example application.

```
hagnes=agnes(cereals[,-c(1,2,11)],diss=FALSE,stand=TRUE)
```

What are the settings for this solution?

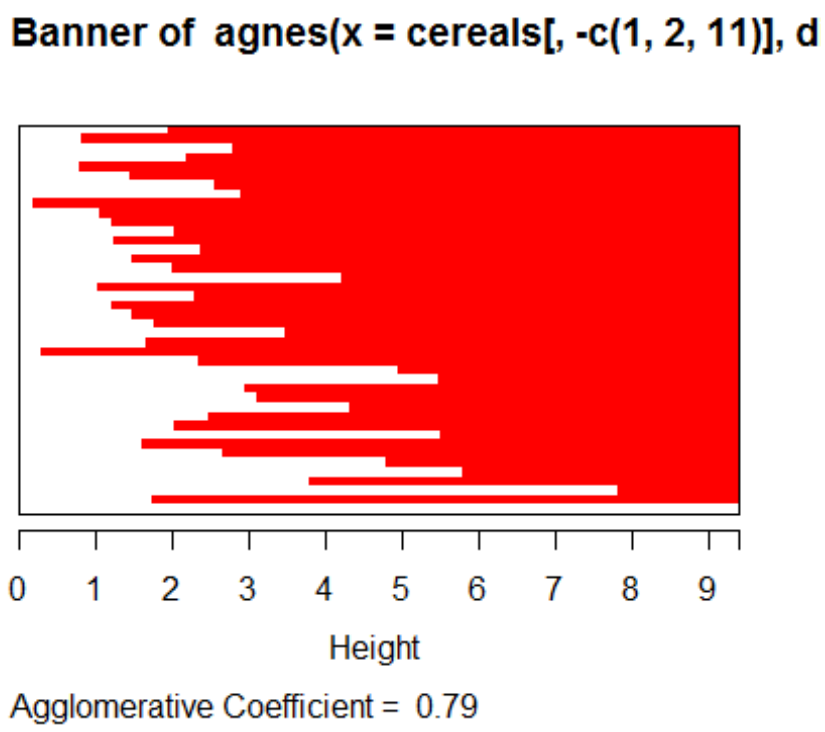
We can get the dendrogram as:

```
plot(hagnes,which=2)
```

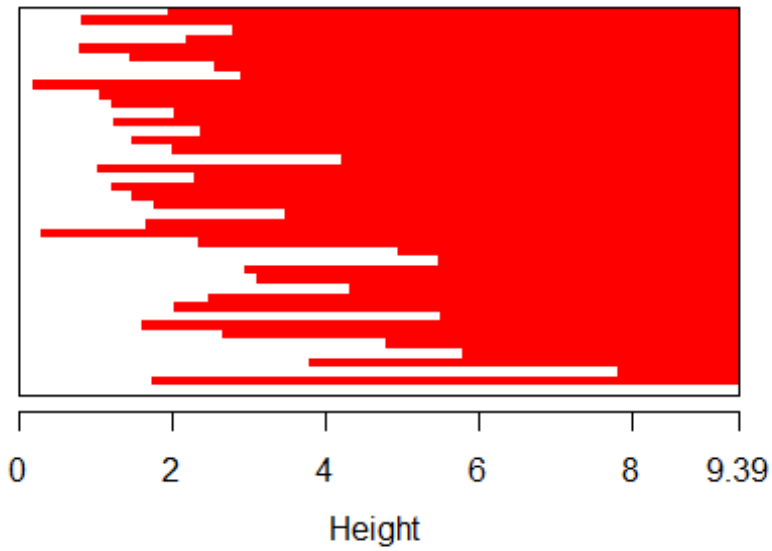



We can also get a banner plot from using this function. The banner plot is a different way of looking at a dendrogram.

```
plot(hagnes,which=1) #OR
```




```
bannerplot(hagnes)
```



Looking back at the dendrogram or at the bannerplot (first option), we see an agglomerative coefficient. Values close to 1 indicate a strong clustering solution has been found. .79 isn't terrible, but also isn't awesome.

You could look at other hierarchical clustering functions - *diana* does divisive clustering. Diana has an associated divisive coefficient. Again, values near 1 are optimal for strong clustering solutions.

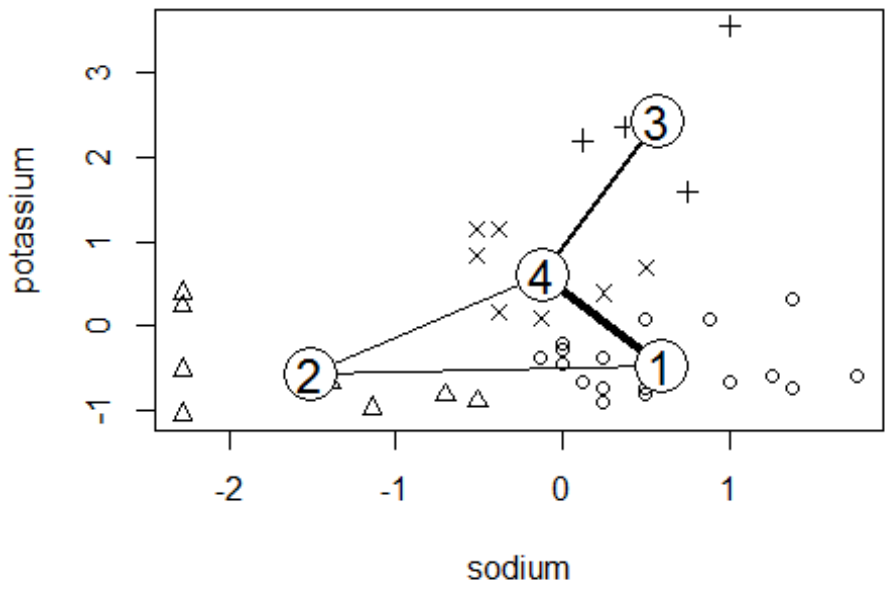
## Visualizations (From 6.6)

The convex clustering method/function is introduced in 6.6 in order to show you a neighborhood plot. Their example uses k-means clustering. Here, we use it to demonstrate the plot. It specifies this plot should only be used on a bivariate solution (or a project to PC axes or similar). So, we base our solution on potassium and sodium as an example. This requires the *flexclust* package, so you'll need to install it once. You could investigate the *cclust* function further.

```
require(flexclust)
```

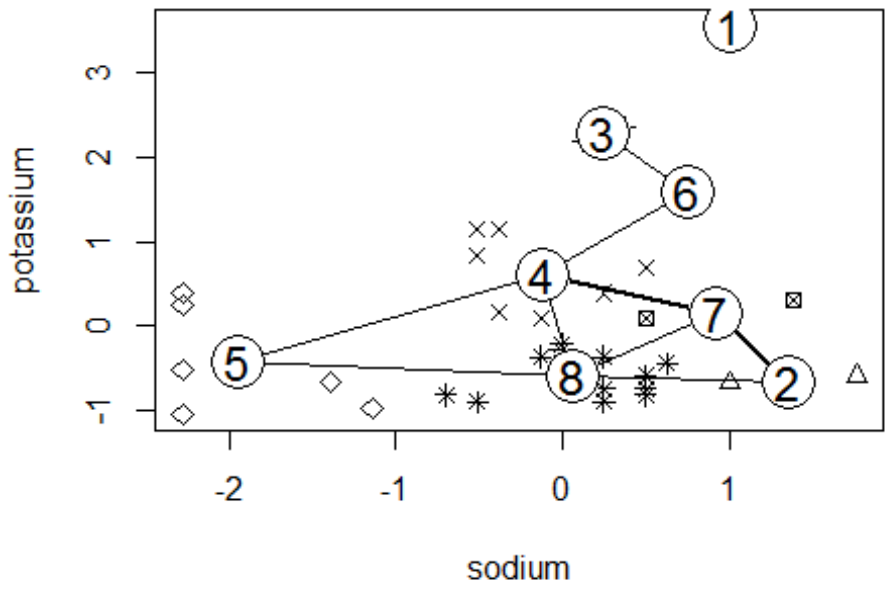
```
## Loading required package: flexclust
## Loading required package: grid
## Loading required package: modeltools
## Loading required package: stats4
```

```
km4=cclust(scale(cereals[,c(6,10)]),k=4,save.data=TRUE)
plot(km4,hull = FALSE, col = rep("black", 4))
```



I did not check here to see if 4 was a good number of clusters. The plot is designed to help assess this. Thicker lines indicate the clusters may not really be separate. Let's look at what happens when k=8.

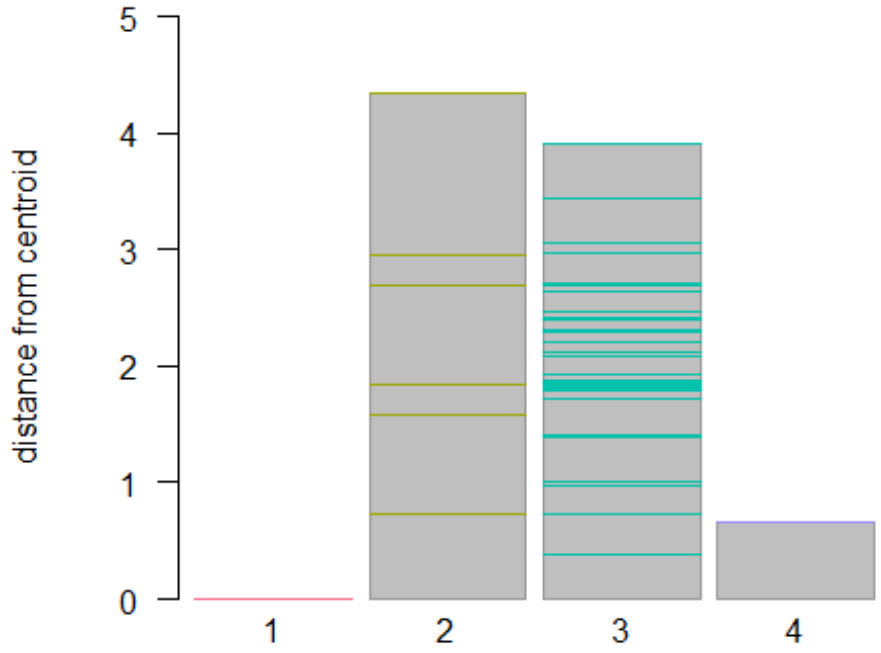
```
km8=cclust(scale(cereals[,c(6,10)]),k=8,save.data=TRUE)
plot(km8,hull = FALSE, col = rep("black", 8))
```



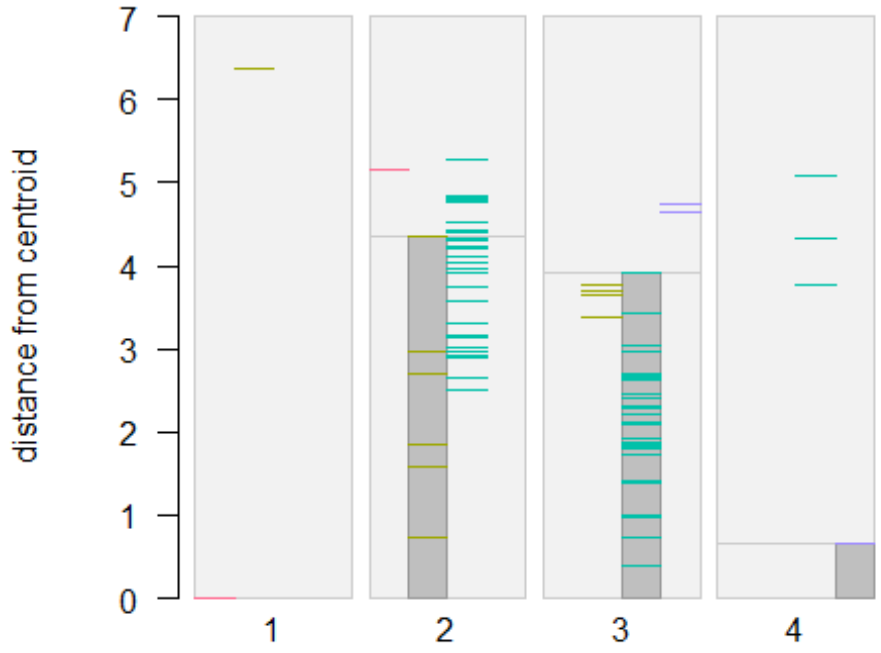

Which solution do you prefer? K=4 or k=8?

Another plot in 6.6 is the stripes plot. This is also in the flexclust package. Here, this shows you what the next closest cluster centroid for each observation is. Ideally, we want observations to fit well in their own clusters. Ask me if you have trouble understanding this plot after reading below. This is NOT restricted to 2-variable solutions, so we change that.

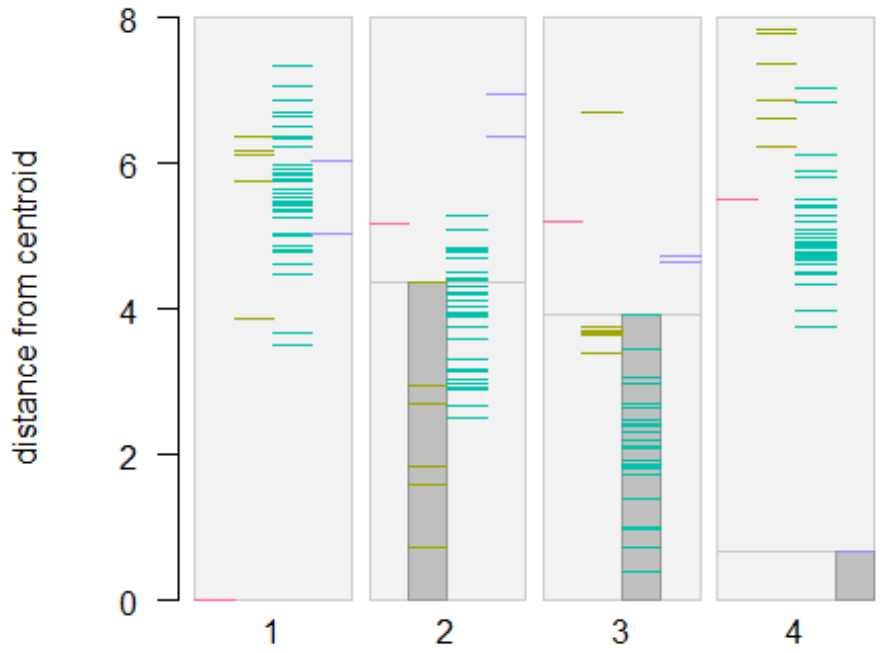
```
k4=cclust(scale(cereals[,-c(1,2,11)]),k=4,save.data=TRUE)
stripes(k4,type="first") #distance of each observation to its cluster centroid
```




```
stripes(k4,type="second") #to its centroid and second-closest centroid
```



```
stripes(k4,type="all")
```




Each cluster is shown. Let's look at the *type=second* plot. In the cluster 1 large rectangle, we see a smaller rectangle that has the distance each observation in cluster 1 is from the cluster one centroid. The OTHER lines in the large cluster 1 rectangle are from observations in clusters 2,3, or 4 whose second-closest cluster centroid is the cluster 1 centroid. A similar idea holds for the other cluster large rectangles. Ideally, you'd see a clusters own points at a low distance from that cluster centroid, and at higher distances from other cluster centroids. The type="all" plot shows how the points line up compared to the centroids for all other clusters.

**Appendix B** - Article Resources for Topic Exploration in the Clustering Application Day and Class Handout



**Articles:**

1. Olexa, Edward M., and Peter JP Gogan. "Spatial population structure of Yellowstone bison." *The Journal of wildlife management* 71.5 (2007): 1531-1538.

2. Mazzoleni, S., A. Lo Porto, and C. Blasi. "Multivariate analysis of climatic patterns of the Meditrranean basin." *Vegetatio* 98.1 (1992): 1-12.

3. Quigg, J. Michael, et al. "No bones about it: Using lipid analysis of burned rock and groundstone residues to examine Late Archaic subsistence practices in south Texas." *The Plains Anthropologist* (2001): 283-303.

4. Ridings, Rosanna, and C. Garth Sampson. "There's no percentage in it: Intersite spatial analysis of Bushman (San) pottery decorations." *American Antiquity* (1990): 766-780.

5. Reichmann, D., et al. "The modular architecture of protein–protein binding interfaces." *Proceedings of the National Academy of Sciences of the United States of America* 102.1 (2005): 57-62.

6. Hall, Kimberly R., and Susan L. Maruca. "Mapping a forest mosaic–A comparison of vegetation and bird distributions using geographic boundary analysis." *Plant Ecology* 156.1 (2001): 105-120.

7. Mannion, David, and Peter Dixon. "Authorship attribution: the case of Oliver Goldsmith." *Journal of the Royal Statistical Society: Series D (The Statistician)* 46.1 (1997): 1-18.

8. Lindsey, Delwin T., and Angela M. Brown. "Universality of color names." *Proceedings of the National Academy of Sciences* 103.44 (2006): 16608-16613.

9. Béjean, Sophie, Christine Peyron, and Renaud Urbinelli. "Variations in activity and practice patterns: a French study for GPs." *The European Journal of Health Economics* 8.3 (2007): 225-236.

10. Savage, Patrick E., and Steven Brown. "Mapping music: cluster analysis of song-type frequencies within and between cultures." *Ethnomusicology* 58.1 (2014): 133-155.

11. Perry, Raymond P., et al. "Attributional (explanatory) thinking about failure in new achievement settings." *European Journal of Psychology of Education* 23.4 (2008): 459-475.

**Class Handout:** (spread over 2 pages, question spacing removed)

This handout is designed to lead us through a discussion on clustering applications. Your group's goal is to skim/read through parts of your article that discuss the data and application of clustering analysis in order to answer the questions (designed to guide you for what to look for) and share your findings with the class in a brief (informal) presentation. Try to convey what the clustering was used for, describe the clustering method applied, and how well you feel the analysis was done in your informal presentation. Each group should have a hard copy of an article to look at. Links to online versions are available in the Word file posted on Moodle.

Clustering Applications

1. What is the stated purpose of the analysis? Why are the authors applying clustering methods?

2. What data is being analyzed in your article? (Variables, general area of application)

3. How is clustering described? What clustering method(s) are applied? Do they give a good description of the technique(s)?

4. What options were used to apply clustering in the article? They might mention their chosen distance measure or standardization choices.

5. What are the findings of the application of cluster analysis? How many clusters were found?

6. What, if any, visuals (graphs or tables) are used to convey the results of the analysis? Do the visuals work well?

7. Is the clustering the end of the analysis? What else is done? You may also list other techniques applied here that you are unfamiliar with.

8. Is the analysis performed well? What could be improved upon?

## **Appendix C**: Clustering Homework Assignment



**Theoretical Portion:**

**1.** Use agglomerative clustering with ***average*** linkage to construct a dendrogram based on the following dissimilarity matrix for 6 observations BY HAND (note: average linkage is harder than single/complete linkage because when you get to 3 points in a cluster, you **cannot** just average the dissimilarities in your current matrix):

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0 | | | | | |
| 2 | 4.7 | 0 | | | | |
| 3 | 6.6 | 11.1 | 0 | | | |
| 4 | 8.1 | 11.6 | 3.4 | 0 | | |
| 5 | 4.9 | 5.7 | 6.2 | 7.7 | 0 | |
| 6 | 2.9 | 2.5 | 7.0 | 6.8 | 10.4 | 0 |

Please show your work and your final dendrogram.

Take a look at the article cited below and then address the following questions. You can consider it supplemental reading (focused on k-means but dealing with general clustering concepts). There may be statistical terms you do not recognize, focus on understanding the clustering concepts.

Jain, Anil K. "Data clustering: 50 years beyond K-means." *Pattern Recognition Letters* 31.8 (2010): 651-666.

**2.** What year was k-means first published? (See abstract)

**3.** What is the point of Figure 2? What does it suggest about the power of human observation versus writing an algorithm to perform clustering?

**4.** In section 2.6, “Major approaches to clustering”, alternative clustering algorithms are discussed. Provide the name of an algorithm that is not covered in your textbook.

**5**. Figure 10 is interesting – it is a dendrogram of a clustering solution where cluster algorithms are the objects being clustered. In relation to this picture, **Sammon**’s mapping is mentioned. What version of MDS (classical/metric or non-metric) is Sammon’s mapping an application of?

**6**. In your own words, describe what a clustering ensemble is and what it is used for. (Section 4.1, and Figure 11)

**7.** In the summary, section 5, multiple issues are listed for clustering researchers to consider in the future. What three issues seem most important to you? Explain your reasoning in a short paragraph.

**Data Analysis Portion** – Return to Boston Housing

Previously, you explored the Boston housing data set to address some questions for a real estate agent. Now, we return to the data set in order to inform the agent about the "natural" types of housing available in Boston. Perform an appropriate analysis. How many housing groups do you find? What are the housing groups? How are the groups different? You should compare and contrast solutions from at least two different grouping methods (you may mention trying others that you don't show). Then pick a final model. For your final chosen housing model, determine (this is in your opinion based on your solution and it is okay to say that none of your groups match!):
What housing group best fits a couple with 2 young children who want a large home?
What housing group best fits an elderly couple with modest income but who do not want to live in an industrial area?
What housing group best fits a single adult whose job requires lots of local travel and who likes being near rivers?

Data set recap:
The data set *boston.txt* contains 14 variables on housing values in the suburbs of Boston. Variables include crime, zone (deals with residential percentage of housing), indust (deals with industry proportion), charles (on river or not), nox (amount of nitric oxides), rooms (avg. number per dwelling), age (proportion built before 1940), distance (weighted to 5 employment centers), radial (index of access to highways), tax (property tax per $10,000), ptratio (pupil-teacher ratio), minor (index of proportion of minorities), lstat (% lower status of population), and medv (median value of homes). Zone is always between 0 and 100. Charles is a binary variable. Radial is an index from 1- 24 (integers). For minor, values from 196 and below indicate a large proportion of minorities. Minor values around 396 indicate a completely non-minority area, and values below 396 down to 196 indicate areas that have decreasing proportions of minorities.

You do NOT need to use all the variables, but state which ones you remove and why. Remember that daisy can help you construct distances appropriate for mixtures of variable types, if you want to include those variables, or treat them differently than the default. Ask me for assistance if you are having trouble constructing the distance that you want to use!

(This question is modeled after an assignment from a graduate course in Multivariate Data Analysis at the University of Michigan using the same data set. That assignment was more involved than this, so it was modified to be more accessible to undergraduates and focus on clustering).

**Appendix D:** Clustering Question on Midterm 2



**1**. The data set primatedata.txt (seen in Homework 3) contains variables about shoulder blades from 105 primates. There are 5 indices (basically measured distances): AD.BD, AD.CD, EA.CD, Dx.CD, and SH.ACR, and 2 angles recorded (EAD and beta) for each shoulder blade. The primate genus (plural: genera) is also recorded. Gibbons (Hylobates), orangutangs (Pongo), chimpanzees (Pan), gorillas (Gorilla), and man (Homo) are all represented in the data set. A curious statistics student wants to know if there any natural groupings of shoulder blades based on the quantitative variables, and if so, if the groups correspond to the genera. Thus, the student undertakes a cluster analysis. Variables will be standardized using *scale* for the analysis.

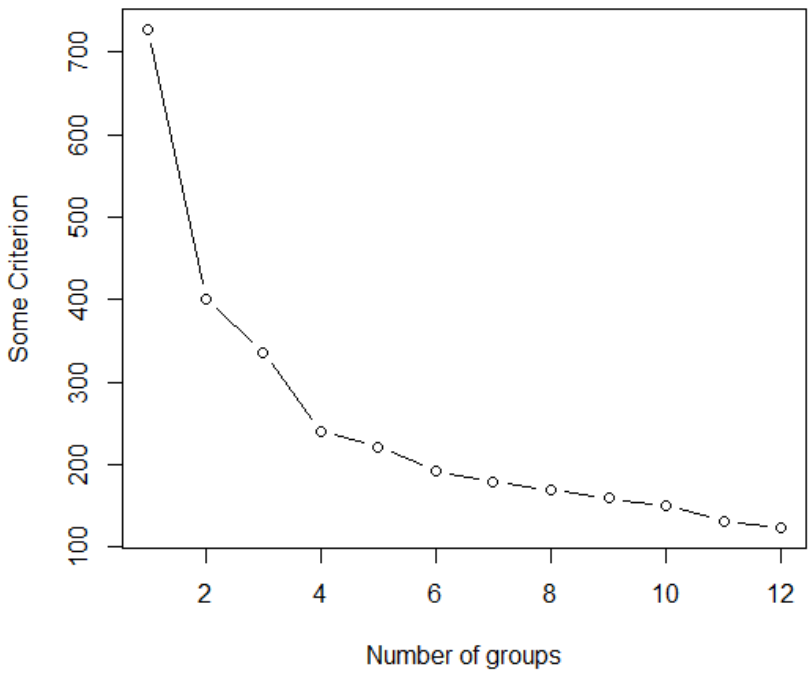
**a**. (1 point) We learned to implement partitioning and hierarchical clustering methods. One broad class of methods from your textbook that we didn't implement an example from is _____- based methods.

**b**. (1 point) If you wanted to implement agglomerative clustering with complete linkage, then, after merging the two closest clusters, you would update the distance between the new cluster and the other clusters with the _____ distance between any pair of points with one point in the new cluster and one point in the other cluster.

**c.** (2 points) The student begins with a k-means analysis, and refers to notes to generate the following plot with the code shown.

```
set.seed(58)
> criterion <- rep(0, 12)
> for(i in 1:12){criterion[i] <-
sum(kmeans(scale(primate[,-
c(8)]),centers= i)$withinss)}
> plot(1:12, criterion, type =
"b", xlab = "Number of groups",
ylab = "Some Criterion")
```

What is a more appropriate label for the Y-axis? What does this criterion measure?



**d**. (1 point) How many clusters would you pick for the k-means solution? Why?

**e.** (2 points) What is the significance of the set.seed(58) part of the code above? Why is that useful or necessary to include here?

**f.** (2 points) A hierarchical solution was obtained using average linkage. The solution was judged to be suboptimal, due to issues with average linkage being a compromise between single and

_____ linkage. In particular, average linkage can inherit _____ behavior from

single linkage and the tendency of the other linkage to find clusters of similar "diameter".

**g**. (2 points) Finish the sentence and fill-in-the-blanks.

Both _____coefficients and shadow coefficients measure....

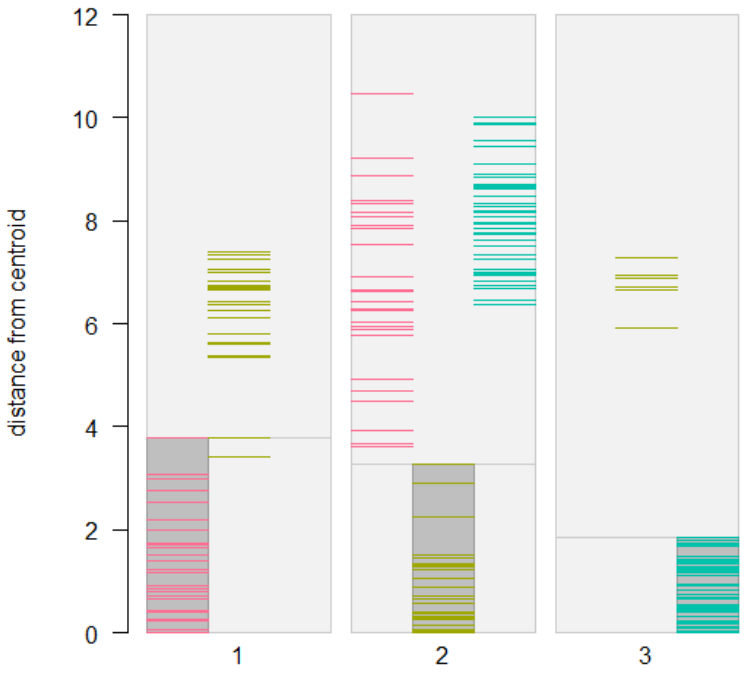
Both coefficients accomplish this by looking at the distance each observation is from its cluster centroid,

as well as its distance from its _____ closest cluster centroid, and combining the distances in some fashion.

**h**. (2 points) The curious student wants to compare the k-means 4 cluster solution to the genera, and generates the following table. Do the genera appear to correspond in any way to the clusters found? Explain.

```
tally(Ksol1$cluster~class,data=primate,format="count")
              class
Ksol1$cluster Gorilla Homo Hylobates Pan Pongo
            1      14    0         0  20     0
            2       0    0        16   0     0
            3       0   39         0   0     0
            4       0    1         0   0    15
```

**i**. (2 points) The plot at right is a _____ plot for a different clustering solution. Is this a "good" clustering solution for these observations? Explain using the plot.




**Bonus:** (1/2 point) Diana is an algorithm for hierarchical clustering, but it runs _____ rather than agglomerative clustering (agnes runs agglomerative).