STAT 503 Homework 4: STAT 101 Grades

Andee Kaplan, Will Landau, Fangge Liu, Lindsay Rutter

March 27, 2015

Introduction

STAT 101 course instructors at Iowa State University usually claim their students are diverse. The undergrads who sign up have a wide variety of majors, backgrounds, perspectives, abilities, and levels of motivation. Visual and unsupervised analyses of homework grades may classify students in useful, insightful ways and even inform pedagogy.

We have three homework grade spreadsheets, each of which comes directly from Blackboard (Iowa State University 2015), Iowa State's system for managing course materials and grades. Each dataset corresponds to a single semester of STAT 101: either fall 2013, fall 2014, or spring 2014. Each semester has six or seven sections of roughly one hundred students each. Every spreadsheet has roughly twenty variables, each of which corresponds to a homework grade (either percentage of points earned or NA for a missing assignment) or the average homework score with missing assignments removed.

Missing values

A large fraction of homework scores appear as "NA", or missing. Since the spreadsheets come directly from Blackboard, we assume that almost all NA's correspond to homeworks that students failed to turn in, and only a small number, if any, are from bookkeeping errors. Before clustering, we need to impute these values, and for an imputation strategy, we look at patterns of missingness.

Visual patterns in missing values

Figure 1 plots the number of missing values per student for all students. Most of the students missed few to no assignments, and some students missed several. This pattern is consistent with each semester and section. Figure 2 shows the number of missing values for each assignment within each semester and section. For the fall semesters, notable spikes occur at chapter 9 ("Understanding Randomness") and topic 9 ("Sample Surveys"). Otherwise, for most sections, the number of missings increased steadily for each section over time.

Figures 3, 4, and 5 show each student's pattern of missingness. General patterns are consistent. Most students had few missing assignments scattered sporadically over the semester, some students dropped the class early, and a smaller students missed strings of around five or ten consecutive assignments in the beginning, middle, or end of the semester.

Grouping students by missingness

From inspection, we can partition the students in each semester into four groups.

Group 1 contains students who did not submit all of the last nine (about half) of all homework assignments. Group 2 contains students (who were not in Group 1) and missed at least nine (about half) of all homework assignments. Group 3 contains students (who were not in Group 1 or 2) and missed at least one homework assignment. Group 4 contains students (who were not in Group 1, 2, or 3) who did not miss any homework assignments.

As a result, we generated 12 main groups (the four groups across three semesters), as shown in Table 1.

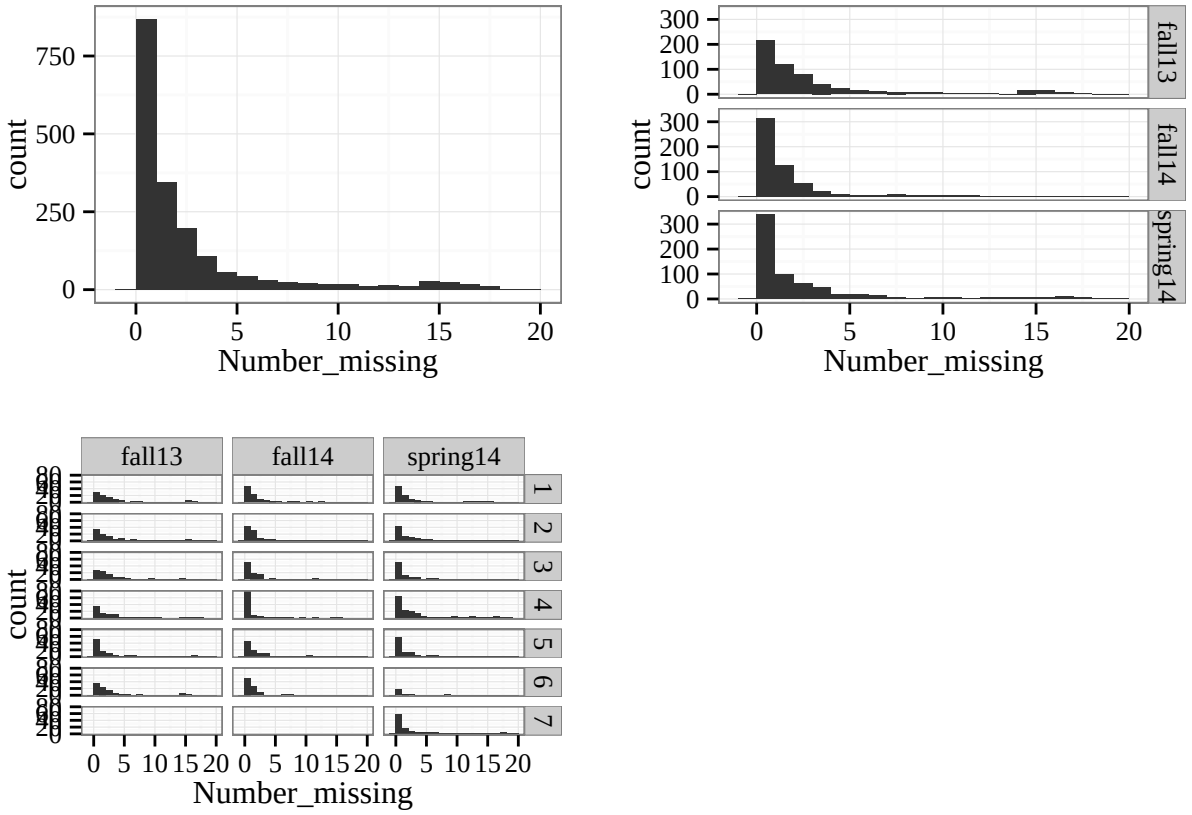

Figure 1: number of missing values per student for all students. The top right panel facets by semester, and the bottom left panel facets by semester and section number. Most of the students missed few to zero assignments, and some students missed several. This pattern is consistent with each semester and section.

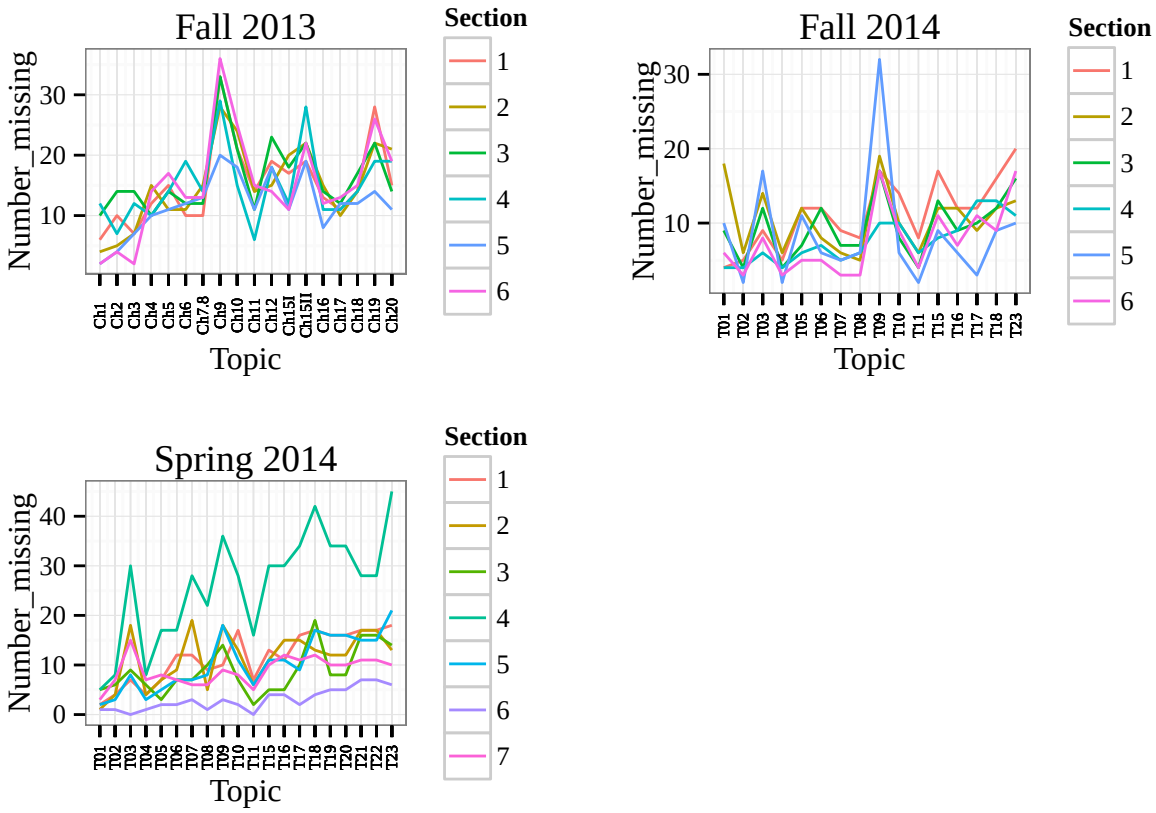

Figure 2: number of missing values per assignment for each semester and section. For the fall semesters, notable spikes occur at chapter 9 ("Understanding Randomness") and topic 9 ("Sample Surveys"). Otherwise, for most sections, the number of missings increased steadily for each section over time.

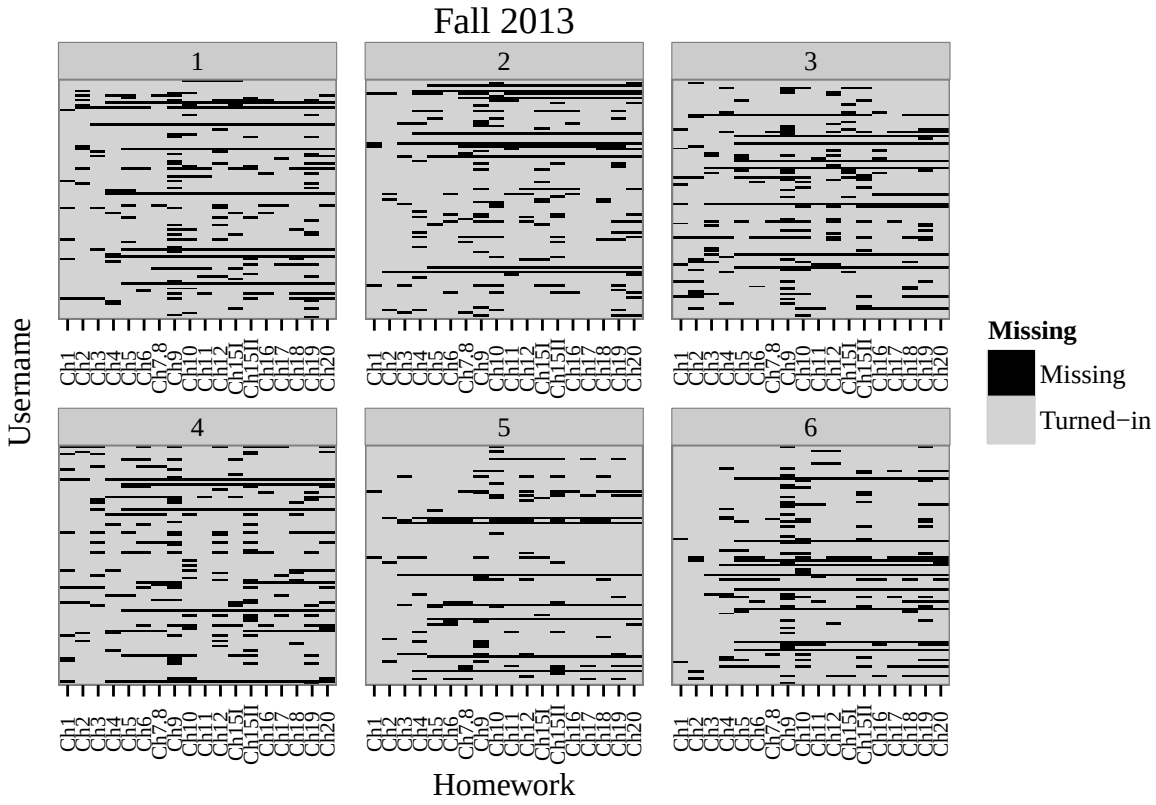

Figure 3: missing assignment records for fall 2013 students faceted by section number. Each row represents a student, each column is an assignment, and the tiles are colored according to the status of the corresponding assignment (missing or turned in). General patterns are consistent across section number. Most students missed few assignments, and the students with the most missings usually missed the last fifteen assignments. These students most likely dropped the class early.

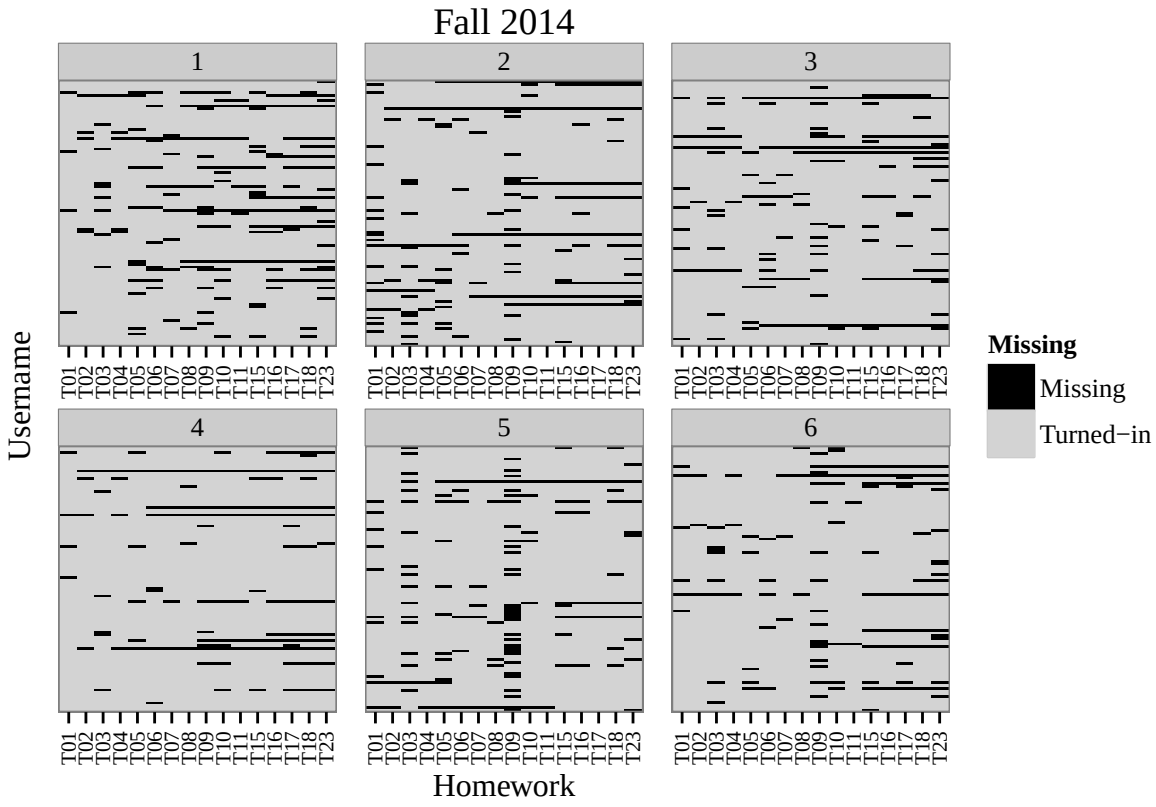

Figure 4: Same as Figure 3, but for fall 2014. We see some early drops, but also some students who failed to turn in either the first few or the middle few assignments.

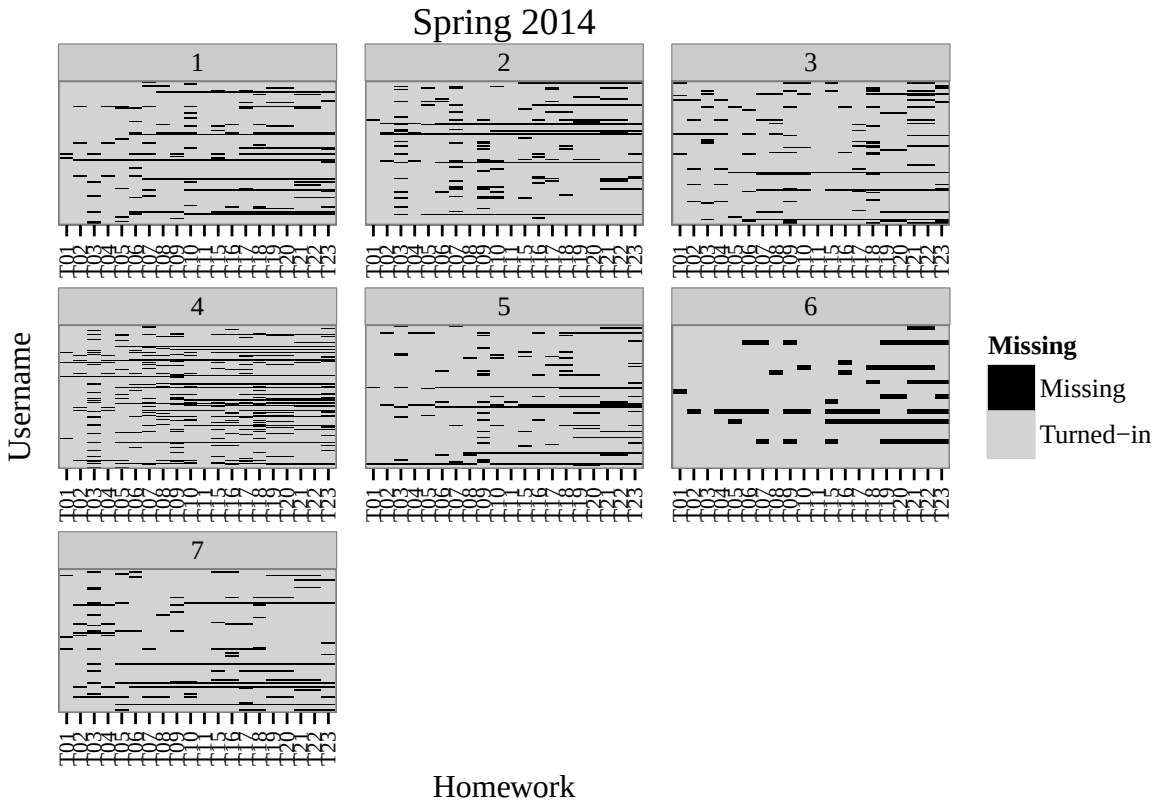

Figure 5: Same as Figure 3, but for spring 2014. Patterns are similar to those of Figure 3 (fall 2013).

Table 1: The number of students who were categorized into one of four mutually-exclusive groups, for the three semesters. We considered Groups 1 and 2 to be problematic, as they likely represented students who dropped the course or habitually missed assignments. However, even after removing students from Groups 1 and 2, we are still left with a very large dataset to cluster

	Fall 13	Spring 14	Fall 14
Group 1 - Drop outs	45	48	17
Group 2 - Common missings	24	23	20
Group 3 - Sporadic missings	298	268	236
Group 4 - No missings	216	338	314

Trim and impute

As we explain above, groups 1 and 2 are relatively small, and the students in these groups missed at least half of the homework assignments. With good reason, we give each of these groups its own cluster and concentrate the rest of our analysis on groups 3 and 4 only. These remaining students have some missing values left, and we impute them with the nearest neighbors imputation functionality in the DMwR (Torgo 2010) package. Dr. Cook wrote most of this code.

Number of clusters

With groups 3 and 4 put together and imputed as above, It may be useful to attempt to determine the appropriate number of clusters. First, we consider wb.ratio, the ratio of average within-cluster Euclidian distance to average between-cluster Euclidian distance. Figure 6 shows wb.ratio as a function of k, the number of clusters. We show results for kmeans (with the kmeans function in core R), along with hierarchical clustering with several linkage methods (with the hclust function in core R). It is alarming that wb.ratio does not monotonically decrease with k, so any clustering analysis should be cautious. It is especially important to note that single, average, and centroid linkages should not be trusted, because here, wb.ratio mostly increases with increasing k. (And yet, for small k, these linkages with increasing wb.ratio outperform other methods for small k.) The ward linkage clusterings are the most believable because wb.ratio behaves almost as it should: except for some erratic behavior for small k, wb.ratio dips sharply and then starts to level off. There is no clear choice of k, though $k = 6$ marks the point where most of this erratic behavior stops and a gentle downward slope begins.

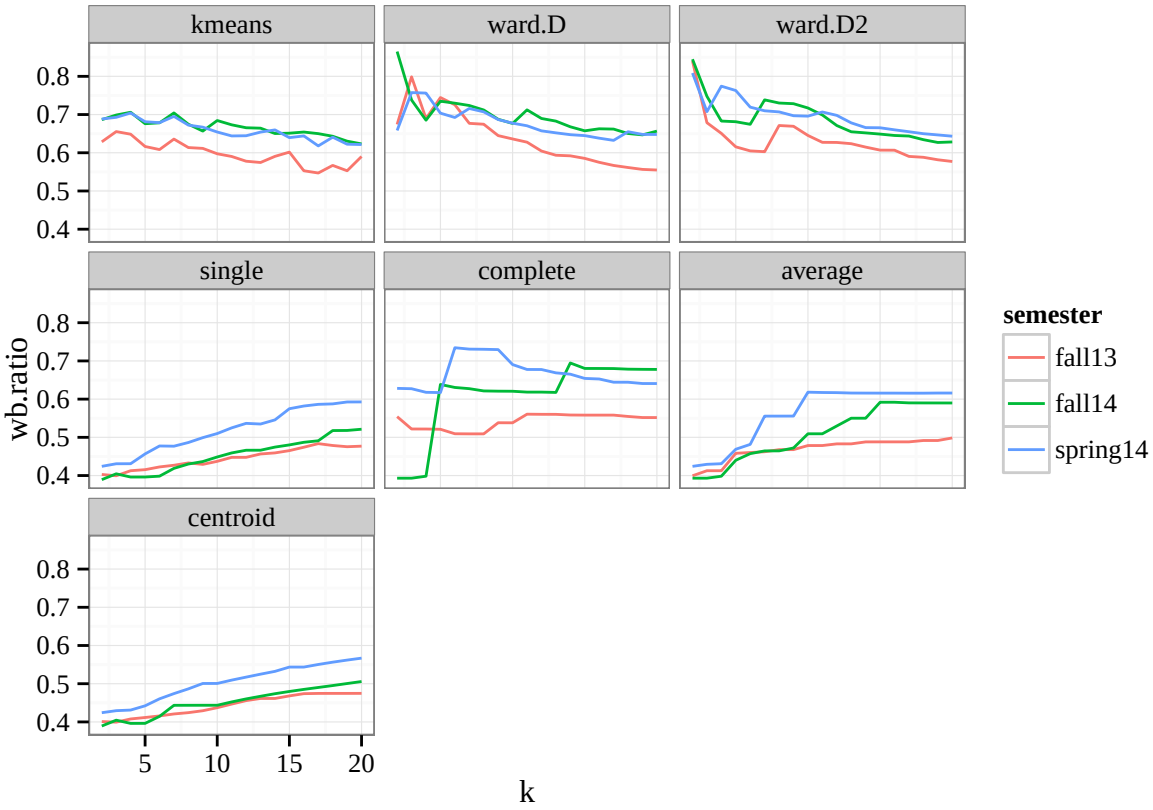

Figure 6: wb.ratio as a function of k, the number of clusters. We show results for kmeans (with the kmeans function in core R), along with hierarchical clustering with several linkage methods (with the hclust function in core R). It is alarming that wb.ratio does not monotonically decrease with k, so any clustering analysis should be cautious. It is especially important to note that single, average, and centroid linkages should not be trusted, since here, wb.ratio mostly with increasing k. (And yet, for small k, these linkages with increasing wb.ratio outperform other methods for small k.) The ward linkage clusterings are the most believable because wb.ratio behaves almost as it should: except for some erratic behavior for small k, wb.ratio dips sharply and then starts to level off. There is no clear choice of k, though $k = 6$ marks the point where most of this erratic behavior stops and a gentle downward slope begins.

We also look at dendrograms from hierarchical clustering to get a sense an optimal k, if we can determine k at all. Figure 7 shows dendrograms from hierarchical clustering using six linkage methods. Only spring 2014 dendrograms are shown, as results for the other two semesters are similar. If we use ward linkage,

arguably the most trustworthy choice in Figure 7 $k = 3$ is a reasonable choice. The other dendrograms are difficult to interpret. Single and centroid linkage dendrograms are nearly flat, and the complete and average dendrograms are likewise not definitive.

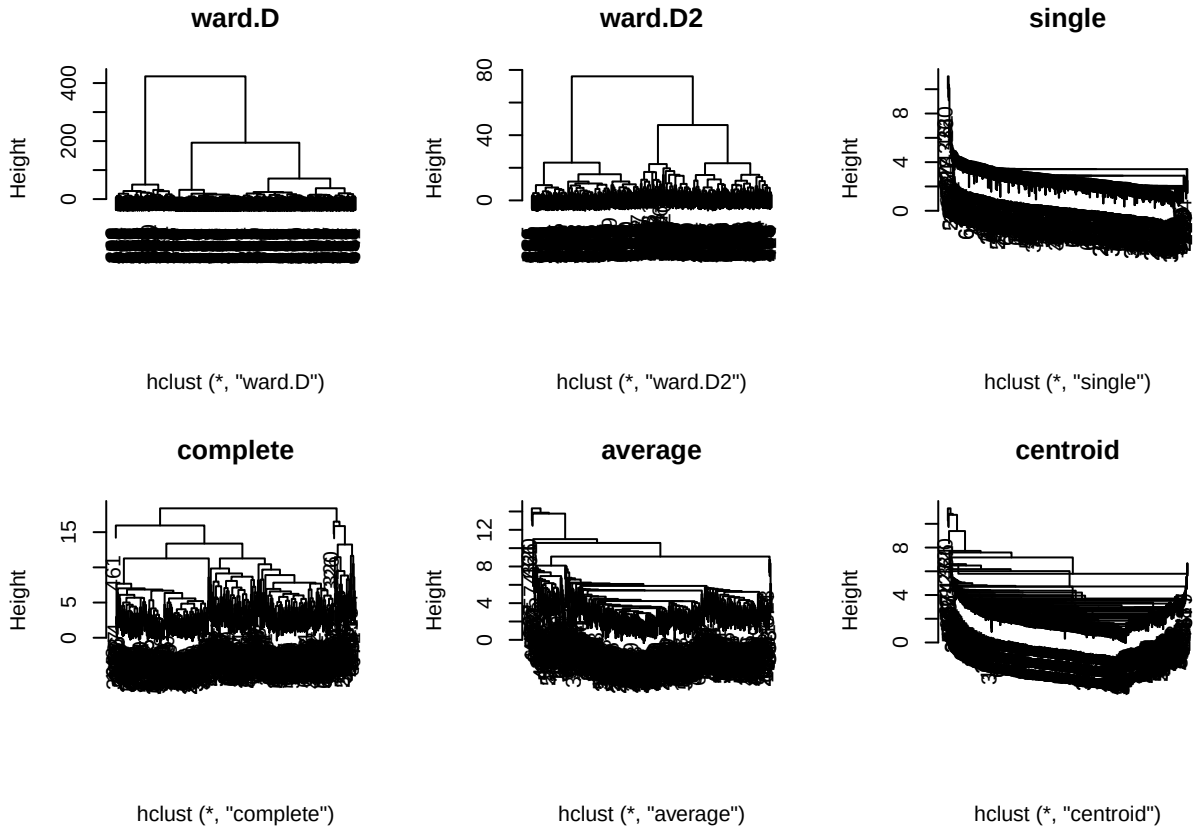

Figure 7: We also look at dendrograms from hierarchical clustering to get a sense an optimal k, if we can determine k at all. Figure ef{fig:dendros} shows dendrograms from hierarchical clustering using six linkage methods. Only spring 2014 dendrograms are shown, as results for the other two semesters are similar. If we use ward linkage, arguably the most trustworthy choice in Figure ef{fig:dendros} $k = 3$ is a reasonable choice. The other dendrograms are difficult to interpret. Single and centroid linkage dendrograms are nearly flat, and the complete and average dendrograms are likewise not definitive.

Acknowledgements

We would like to thank Dr. Cook for her advice on dealing with missing values and her imputation code. Also, we used the R packages DMwR (Torgo 2010), fpc (Hennig 2014), ggplot2 (Wickham 2009), gridExtra (Auguie 2012), and reshape2 (Wickham 2007).

References

Auguie, Baptiste. 2012. *GridExtra: Functions in Grid Graphics*. http://CRAN.R-project.org/package=gridExtra.

Hennig, Christian. 2014. *Fpc: Flexible Procedures for Clustering*. http://CRAN.R-project.org/package=fpc.

Iowa State University. 2015. "Blackboard Learn." https://bb.its.iastate.edu.

Torgo, L. 2010. *Data Mining with R, Learning with Case Studies*. Chapman; Hall/CRC. http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

———. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer New York. http://had.co.nz/ggplot2/book.