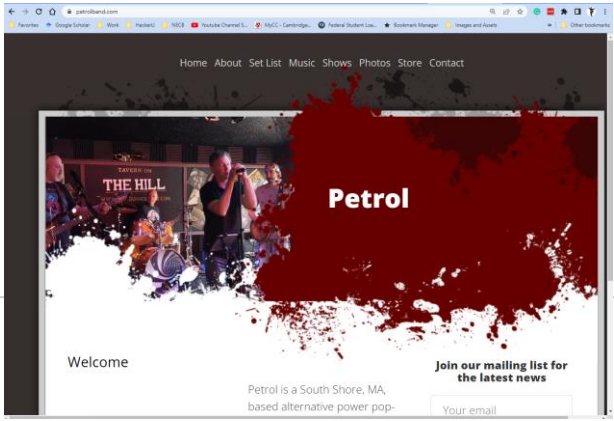
Introduction to Hierarchical Clustering Using College Scorecard Data

Dr. Rich Huebner

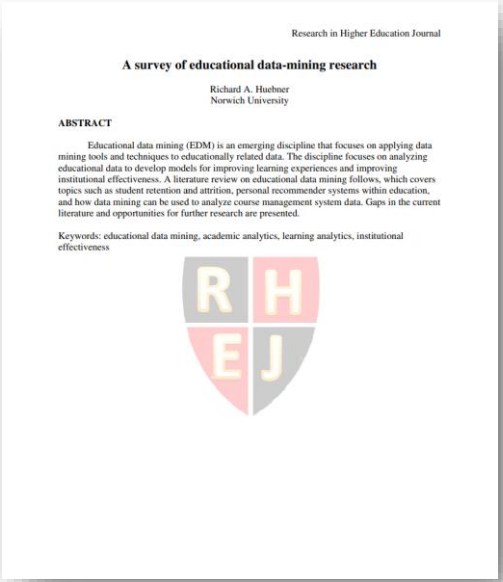
Dr. Rich Huebner



Senior Data Scientist, Nelnet

Professor, Cambridge College

Consultant, Author, Musician



Research in Higher Education Journal

A survey of educational data-mining research

Richard A. Huebner
Norwich University

ABSTRACT

Educational data mining (EDM) is an emerging discipline that focuses on applying data mining tools and techniques to educationally related data. The discipline focuses on analyzing educational data to develop models for improving learning experiences and improving institutional effectiveness. A literature review on educational data mining follows, which covers topics such as student retention and attrition, personal recommender systems within education, and how data mining can be used to analyze course management system data. Gaps in the current literature and opportunities for further research are presented.

Keywords: educational data mining, academic analytics, learning analytics, institutional effectiveness

Outline

- Background
- The Problem / Challenge
- Clustering Intuition
- Data Preparation
- Clustering Procedures
- Results
- Summary

TO PARTICIPATE IN POLLS, SCAN THE CODE OR GO TO:

https://pollev.com/richh199

- OR TEXT richh199 to 37607

Where are you from??

What is your familiarity with clustering algorithms, in general?

I can apply it and teach it to someone else

I can apply it as needed

I can explain the basic concepts of it to someone

I've heard of it but only in passing

No experience

Problem Background

Guiding Questions:

How can you compare schools when looking at college?

Can we deliver a solution that will aid students in finding colleges and programs that suit their interests and perhaps be a “good fit” for them?

Clustering Intuition

- Clustering approaches are fundamentally about grouping items together with similar characteristics.
- Marketing professionals look to clustering to group similar customers together based on characteristics like purchasing habits/sales, demographics, etc.
- There are a lot of applications of clustering.
  - Fraudulent sales, criminal activity, network traffic analysis, etc

Clustering Intuition (2)

- let's get "similar" things grouped together..... And at the same time, try to make sure dissimilar items are NOT grouped together.
- We'll put similar items in the same cluster, and dissimilar items in other clusters.
- Market segmentation, product segmentation, user segmentation, etc.
- A "cluster" will therefore be a collection of items which are similar.

Options for Clustering

- K-Means
- Hierarchical Clustering (Agglomerative and Divisive)
- DB-SCAN (Density-Based Spatial Clustering of Applications with Noise)
- Others (PAM – Partitioning Around Medoids)

Hierarchical Clustering

- Goal is to build a hierarchy of clusters.
- Agglomerative (Bottom-Up)
  - Make each point a single-point cluster (every item starts as its own cluster)
  - Take two closest points (using distance matrix) and make them one cluster (= n-1 clusters)
  - Take two closest clusters and make them 1 cluster (results in n-2 clusters)
  - Repeat the last step until one large cluster exists. (it is an iterative algorithm!)
- The **dendrogram** then can be used to divide the clusters after. Thus, we don't need to know K (# of clusters) beforehand.

Procedures

- Get data (College Scorecard: https://collegescorecard.ed.gov/data)
- Use data dictionary to determine viable attributes that we're interested in.
- Apply any filtering (region of country, type of school, etc)
- Ensure appropriate attributes/features have correct data types
- Ensure nulls are addressed (in this case I will avoid imputation)
- Apply transforms (OneHotEncoding, if needed), and scaling/normalizing
- Get distances / create distance matrix
- Apply clustering technique
- Visualize clusters

Filtering

- We can start by filtering down the number of potential schools we want to look at.
- The college scorecard data has well over 6,000 schools, so I recommend narrowing this down because the visualizations will NOT be helpful with a high number of schools.
- I am filtering on Massachusetts schools since a friend of mine is interested in this analysis, and they are ONLY looking at MA schools.
  - But this could be filtered on several other categorical attributes like Type of School, the primary type of degree they grant, and so on.

Variables of Interest

- State Abbreviation { STABBR } – used only for filtering down the data set.
- Type of programs { PREDDEG }
- Type of school { CONTROL } 1 = Public; 2 = Private Nonprofit; 3 = For-Profit
- Admissions Rate (aka Acceptance Rate) { ADM_RATE_ALL } Given as decimal
- Average SAT score for students admitted { SAT_AVG_ALL }
- Enrollment of undergraduate degree-seeking students { UGDS }
- Average net price { NPT4_PUB, NPT4_PRIV }
- Completion rate { C150_4 }
- Pct of all undergrads receiving a federal loan { PCTFLOAN }
- Pct of all undergrads receiving a PELL grant { PCTPELL }

Choosing Variables and NULL Handling

```
> View(subset)
> head(subset)
  UNITID   OPEID                              INSTNM PREDDEG CONTROL ADM_RATE_ALL SAT_AVG_ALL  UGDS NPT4_PUB NPT4_PRIV COSTT4_A TUITIONFEE_OUT PCTPELL PCTFLOAN C150_4
1 100654  100200               Alabama A & M University       3       1       0.8965         959  5090    15529      NULL    23445          18634  0.7095   0.7504 0.2866
2 100663  105200 University of Alabama at Birmingham       3       1        0.806        1245 13549    16530      NULL    25542          20400  0.3397   0.4688 0.6117
3 100690 2503400                    Amridge University       2       2         NULL        NULL   298     NULL     17618    20100           6950  0.7452   0.8493   0.25
4 100706  105500 University of Alabama in Huntsville       3       1       0.7711        1300  7825    17208      NULL    24861          23734  0.2403   0.3855 0.5714
5 100724  100500               Alabama State University       3       1       0.9888         938  3603    19534      NULL    21892          19396  0.7368   0.7805 0.3177
6 100751  105100              The University of Alabama       3       1       0.8039        1262 30610    20917      NULL    30016          31090  0.1718   0.3644 0.7214
>
```

- Because we filtered on PRIVATE schools, we can reasonably expect that NPT4_PUB (Average Net Price for Public Schools) to be NULL.

```
> nrow(subset)
[1] 1262
> str(subset)
'data.frame':   1262 obs. of  15 variables:
 $ UNITID        : int  100937 101189 101435 101453 101541 101675 101693 101912 102049 102234 ...
 $ OPEID         : int  101200 100300 101900 2199700 102300 102800 102900 103300 103600 104100 ...
 $ INSTNM        : chr  "Birmingham-Southern College" "Faulkner University" "Huntingdon College" "Heritage Christian University" ...
 $ PREDDEG       : int  3 3 3 3 3 3 3 3 3 3 ...
 $ CONTROL       : int  2 2 2 2 2 2 2 2 2 2 ...
 $ ADM_RATE_ALL  : chr  "0.6045" "0.7576" "0.5439" "0.6667" ...
 $ SAT_AVG_ALL   : chr  "1202" "1068" "1101" "NULL" ...
 $ UGDS          : chr  "1129" "1834" "917" "70" ...
 $ NPT4_PUB      : chr  "NULL" "NULL" "NULL" "NULL" ...
 $ NPT4_PRIV     : chr  "19808" "20500" "21632" "NULL" ...
 $ COSTT4_A      : chr  "32514" "34835" "37483" "NULL" ...
 $ TUITIONFEE_OUT: chr  "18900" "22990" "27900" "11532" ...
 $ PCTPELL       : chr  "0.2258" "0.5009" "0.4077" "0.4915" ...
 $ PCTFLOAN      : chr  "0.4615" "0.6384" "0.7252" "0.1017" ...
 $ C150_4        : chr  "0.7094" "0.2711" "0.4185" "0.1429" ...
```

Ensure Columns Have Correct Data Types

```
> str(complete_records)
'data.frame':   1262 obs. of  14 variables:
 $ UNITID         : int  100937 101189 101435 101453 101541 101675 101693 101912 102049 102234 ...
 $ OPEID          : int  101200 100300 101900 2199700 102300 102800 102900 103300 103600 104100 ...
 $ INSTNM         : chr  "Birmingham-Southern College" "Faulkner University" "Huntingdon College" "Heritage Christian University" ...
 $ PREDDEG        : int  3 3 3 3 3 3 3 3 3 3 ...
 $ CONTROL        : int  2 2 2 2 2 2 2 2 2 2 ...
 $ ADM_RATE_ALL   : chr  "0.6045" "0.7576" "0.5439" "0.6667" ...
 $ SAT_AVG_ALL    : chr  "1202" "1068" "1101" "NULL" ...
 $ UGDS           : chr  "1129" "1834" "917" "70" ...
 $ NPT4_PRIV      : chr  "19808" "20500" "21632" "NULL" ...
 $ COSTT4_A       : chr  "32514" "34835" "37483" "NULL" ...
 $ TUITIONFEE_OUT: chr  "18900" "22990" "27900" "11532" ...
 $ C150_4         : chr  "0.7094" "0.2711" "0.4185" "0.1429" ...
 $ PCTPELL        : chr  "0.2258" "0.5009" "0.4077" "0.4915" ...
 $ PCTFLOAN       : chr  "0.4615" "0.6384" "0.7252" "0.1017" ...
>
```

- Many attributes have been imported as strings (or character data type – chr).

```
# Ensure numeric
cols.num <- c("ADM_RATE_ALL", "SAT_AVG_ALL", "UGDS", "NPT4_PRIV", "COSTT4_A" ,"TUITIONFEE_OUT", "C150_4", "PCTPELL", "PCTFLOAN")
complete_records <- complete_records %>%
   mutate_at(cols.num, as.numeric)

str(complete_records)
```

What is the purpose of scaling or normalizing the data for a clustering task?

It converts the columns into a specific range

Improves the accuracy of clustering algorithms

It controls the variability of the data

It ensures attributes are on the same scale

All of the above

Scaling the Data

- This level-sets the important variables such that a single variable that may have large values would not “overpower” other variables with smaller values.
- You don’t want one variable to have an undue influence on the results of your clustering.
- Scaling the numeric data is necessary for clustering tasks.

Clustering Procedures

The Clustering Algorithm

After all cleanup...

- Steps: (we'll do this part in 4 lines of code in R)

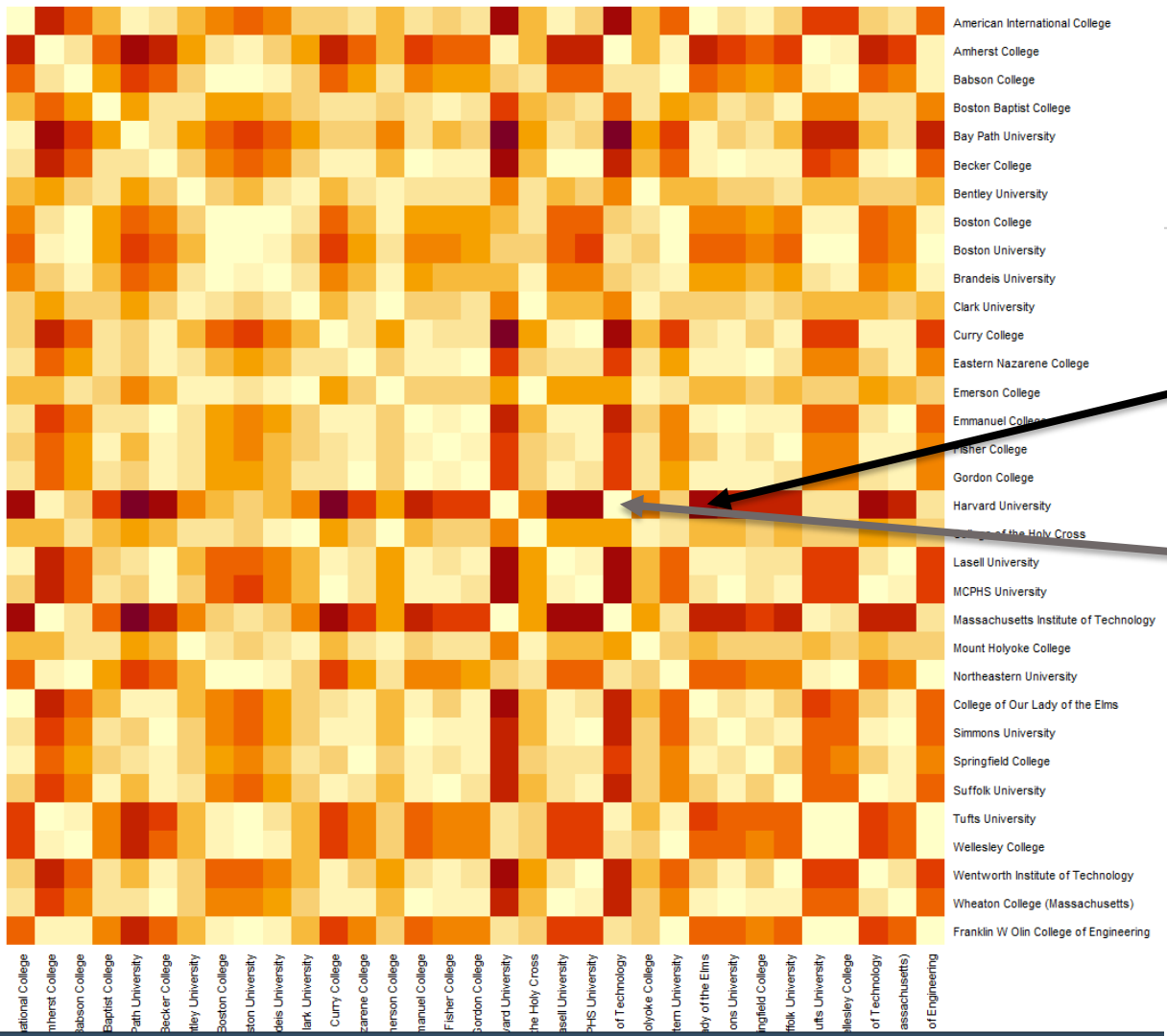
- Compute distances using the dist function in R.
- Pass the distance matrix to the HC algorithm (hclust)
  - A linkage method must be selected to determine how to separate the clusters based on centroids.
- Cut the tree at a selected number of clusters, *k*.
- Visualize the dendrogram.

```
d <- dist(df[, c(7,15)],method = "euclidian")
res.hc <- hclust(d, method = "complete" )
grp <- cutree(res.hc, k = 4)
ggdendrogram(res.hc, rotate = T, theme_dendro = FALSE, size = 2)
```

Distances and Distance Matrix

```
58  d <- dist(df[, c(7,15)],method = "euclidian")
```

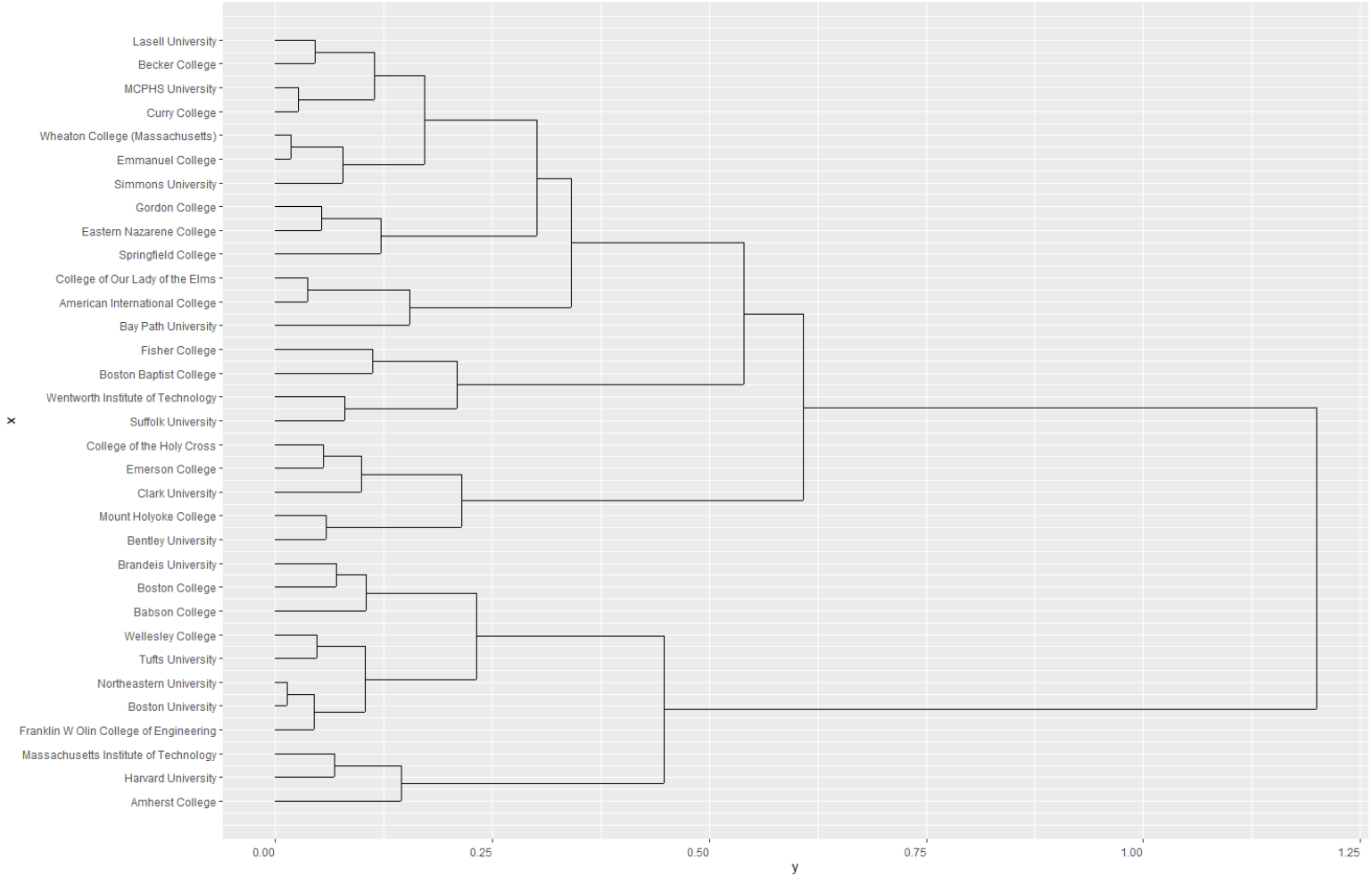
	American International College	Amherst College	Babson College	Boston Baptist College	Bay Path University	Becker College	Bentley University	Boston College	Boston University
American International College	0.00000000	2.69654127	1.93569145	0.19580694	0.33007455	0.397208362	1.488132737	2.33849430	2.26017153
Amherst College	2.69654127	0.00000000	0.76084982	2.50073434	2.36646672	2.299332913	1.208408538	0.35804697	0.43636975
Babson College	1.93569145	0.76084982	0.00000000	1.73988452	1.60561690	1.538483093	0.447558718	0.40280285	0.32448007
Boston Baptist College	0.19580694	2.50073434	1.73988452	0.00000000	0.13426762	0.201401423	1.292325798	2.14268736	2.06436459
Bay Path University	0.33007455	2.36646672	1.60561690	0.13426762	0.00000000	0.067133808	1.158058182	2.00841975	1.93009697
Becker College	0.39720836	2.29933291	1.53848309	0.20140142	0.06713381	0.000000000	1.090924375	1.94128594	1.86296316
Bentley University	1.48813274	1.20840854	0.44755872	1.29232580	1.15805818	1.090924375	0.000000000	0.85036156	0.77203879
Boston College	2.33849430	0.35804697	0.40280285	2.14268736	2.00841975	1.941285939	0.850361564	0.00000000	0.07832278
Boston University	2.26017153	0.43636975	0.32448007	2.06436459	1.93009697	1.862963163	0.772038788	0.07832278	0.00000000
Brandeis University	2.28814394	0.40839733	0.35245249	2.09233701	1.95806939	1.890935583	0.800011208	0.05035036	0.02797242
Clark University	1.43218790	1.26435338	0.50350356	1.23638096	1.10211334	1.034979535	0.055944840	0.90630640	0.82798363
Curry College	0.22377936	2.47276192	1.71191210	0.02797242	0.10629520	0.173429003	1.264353378	2.11471494	2.03639217
Eastern Nazarene College	0.08391726	2.61262402	1.85177420	0.11188968	0.24615729	0.313291102	1.404215477	2.25457704	2.17625427
Emerson College	1.60561690	1.09092437	0.33007455	1.40980996	1.27554235	1.208408538	0.117484163	0.73287740	0.65455462
Emmanuel College	1.04616850	1.65037277	0.88952295	0.85036156	0.71609395	0.648960141	0.441964234	1.29232580	1.21400302
Fisher College	0.43077527	3.12731654	2.36646672	0.62658220	0.76084982	0.827983628	1.918908003	2.76926957	2.69094679
Gordon College	0.80560569	1.89093558	1.13008576	0.60979875	0.47553114	0.408397330	0.682527045	1.53288861	1.45456583
Harvard University	2.82521441	0.12867313	0.88952295	2.62940747	2.49513985	2.428006044	1.337081670	0.48672011	0.56504288
College of the Holy Cross	1.99163629	0.70490498	0.05594484	1.79582936	1.66156174	1.594427932	0.503503558	0.34685801	0.26853523
Lasell University	0.38042491	2.31611636	1.55526654	0.18461797	0.05035036	0.016783452	1.107707827	1.95806939	1.87974661
MCPHS University	0.45315320	2.24338807	1.48253825	0.25734626	0.12307865	0.055944840	1.034979535	1.88534110	1.80701832
Massachusetts Institute of Technology	2.99304893	0.29650765	1.05735747	2.79724199	2.66297437	2.595840563	1.504916189	0.65455462	0.73287740
Mount Holyoke College	1.96925836	0.72728292	0.03356690	1.77345142	1.63918380	1.572049996	0.481125622	0.36923594	0.29091317
Northeastern University	2.66856886	0.02797242	0.73287740	2.47276192	2.33849430	2.271360493	1.180436118	0.33007455	0.40839733



Examples

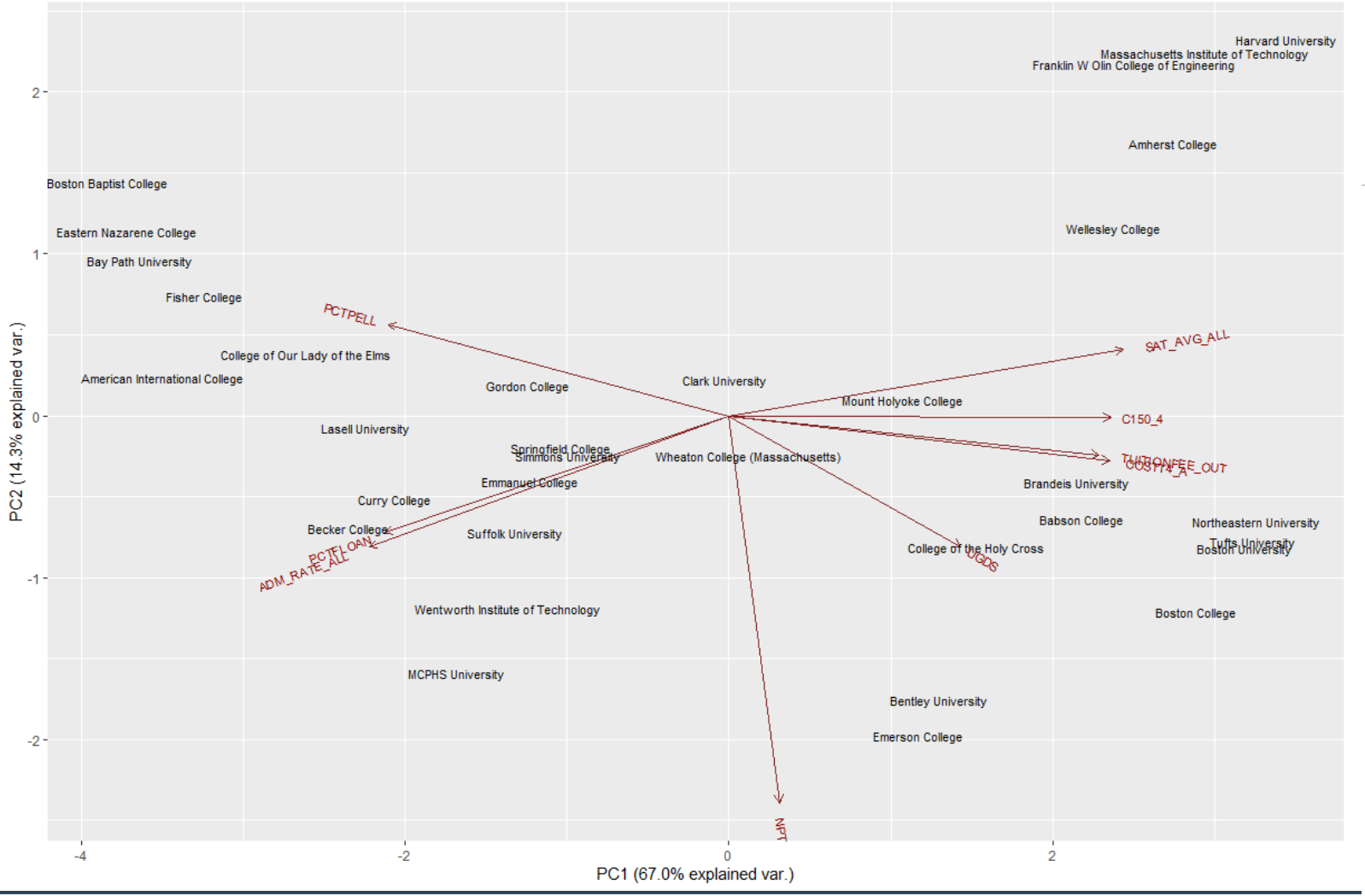
- **Highly dissimilar:**
  - Harvard University & College of Our Lady of the Elms
  - Bay Path & MIT
- **Highly similar:**
  - Harvard and MIT
  - Amherst College & Tufts U.
  - Becker College & Curry College
  - BC, BU
  - BC, Brandeis
- From visualizing the distances we can get a sense.

Dendrogram


Lasell University
Becker College
MCPHS University
Curry College
Wheaton College (Massachusetts)
Emmanuel College
Simmons University
Gordon College
Eastern Nazarene College
Springfield College
College of Our Lady of the Elms
American International College
Bay Path University
Fisher College
Boston Baptist College
Wentworth Institute of Technology
Suffolk University
College of the Holy Cross
Emerson College
Clark University
Mount Holyoke College
Bentley University
Brandeis University
Boston College
Babson College
Wellesley College
Tufts University
Northeastern University
Boston University
Franklin W Olin College of Engineering
Massachusetts Institute of Technology
Harvard University
Amherst College
x
0.00
0.25
0.50
0.75
1.00
1.25
y

Code (after all data cleanup!)

1. d <- dist(df[, c(7:13)], method = "euclidian")
2. d.mat <- as.matrix(d)
3. heatmap(d.mat, Rowv = NA, symm = T)
4. res.hc <- hclust(d, method = "complete")
5. grp <- cutree(res.hc, k = 4)
6. ggdendrogram(res.hc, rotate = T, theme_dendro = FALSE, size = 2)


Harvard University
Massachusetts Institute of Technology
Franklin W Olin College of Engineering
Amherst College
Boston Baptist College
Eastern Nazarene College
Wellesley College
Bay Path University
Fisher College
PCTPELL
College of Our Lady of the Elms
SAT_AVG_ALL
American International College
Gordon College
Clark University
Mount Holyoke College
C150_4
Lasell University
Springfield College
Simmons University
Wheaton College (Massachusetts)
TUITIONFEE_OUT
COSTT4_A_OUT
Emmanuel College
Brandeis University
Curry College
Babson College
Northeastern University
Becker College
Suffolk University
PCTFLOAN
College of the Holy Cross
Tufts University
Boston University
ADM_RATE_ALL
UGDS
Wentworth Institute of Technology
Boston College
MCPHS University
Bentley University
Emerson College
NPT4
PC2 (14.3% explained var.)
PC1 (67.0% explained var.)
2
1
0
-1
-2
-4
-2
0
2

- Code that produces the biplot.

```r
# PCA on the College Scorecard data.
library(ggbiplot)
collscore.pca <- prcomp(df[, c(7:15)], center = T, scale. = T )
summary(collscore.pca)
ggbiplot(collscore.pca, labels = rownames(df), ellipse = T, obs.scale = 1, choices = c(1,2) )
```

```
> grp
                American International College                          Amherst College                          Babson College 
                                             1                                        2                                        3 
                  Boston Baptist College                          Bay Path University                          Becker College 
                                             1                                        1                                        1 
                          Bentley University                          Boston College                          Boston University 
                                             3                                        3                                        4 
                      Brandeis University                          Clark University                          Curry College 
                                             3                                        1                                        1 
                 Eastern Nazarene College                          Emerson College                          Emmanuel College 
                                             1                                        3                                        1 
                          Fisher College                          Gordon College                          Harvard University 
                                             1                                        1                                        2 
          College of the Holy Cross                          Lasell University                          MCPHS University 
                                             3                                        1                                        1 
Massachusetts Institute of Technology                          Mount Holyoke College                  Northeastern University 
                                             2                                        3                                        4 
  College of Our Lady of the Elms                          Simmons University                       Springfield College 
                                             1                                        1                                        1 
                  Suffolk University                          Tufts University                       Wellesley College 
                                             1                                        3                                        2 
    Wentworth Institute of Technology        Wheaton College (Massachusetts)          Franklin W Olin College of Engineering 
                                             1                                        1                                        2 
>
```

Summary

- We can apply hierarchical clustering technique for clustering similar schools together based on a variety of characteristics.
- This helps students and families narrow down the MANY choices they have for schools.
- A lot of data preparation must be done prior to sending data to the hierarchical clustering algorithm. 70-80% of the work is data cleanup.
- HC is relatively easy to implement in R as you can see from the code snippets.

Thanks!

You can find me at

- LinkedIn: www.linkedin.com/in/RichHuebner
- Rich.Huebner@yahoo.com