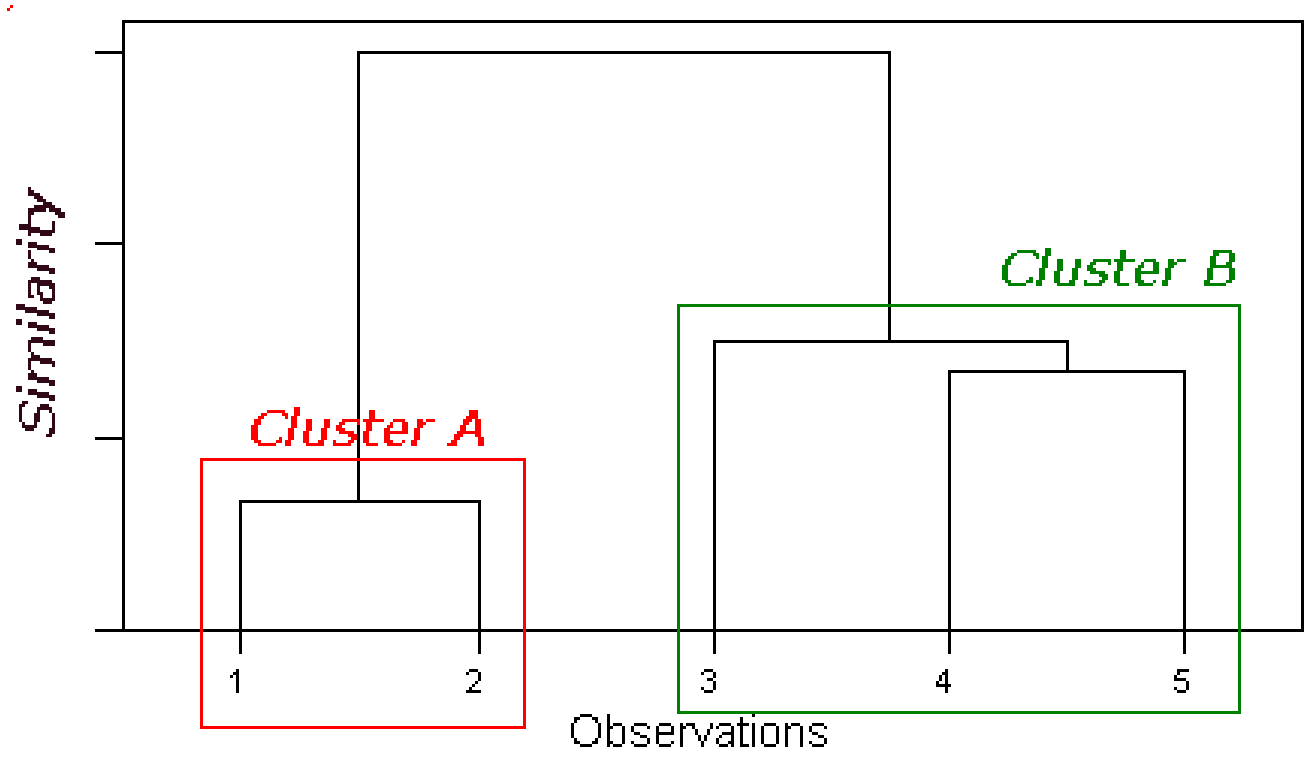
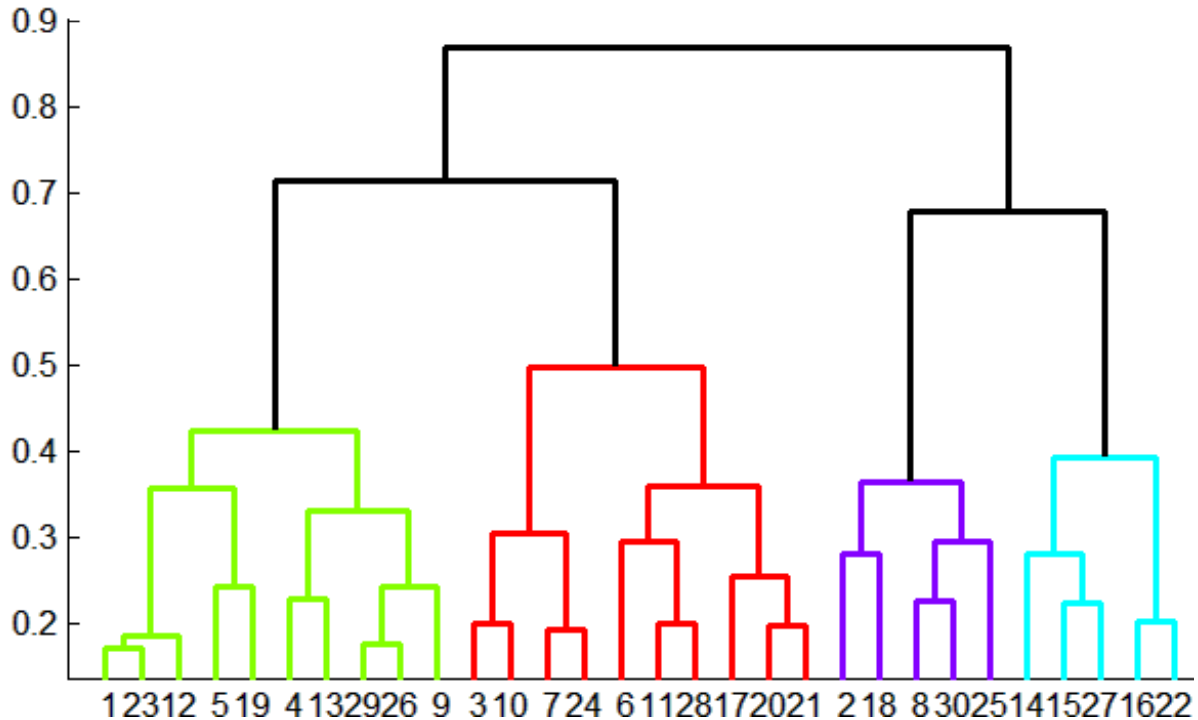
Hierarchical Clustering: The Dendrogram



Dendrograms: Display the Clustering Process

Tree-like diagram that summarize the clustering process

- Similar records joined by links
- Record location determined by similarity to other records



UG Business Programs: *Universities Clustering.xls*

Data for 25 undergraduate programs at business schools in US universities in 1995.

Student Quality

Program

Placement

Univ	SAT	Top10	Accept	SFRatio	Expenses	GradRate
Brown	1310	89	22	13	22,704	94
CalTech	1415	100	25	6	63,575	81
CMU	1260	62	59	9	25,026	72
Columbia	1310	76	24	12	31,510	88
Cornell	1280	83	33	13	21,864	90
Dartmouth	1340	89	23	10	32,162	95
Duke	1315	90	30	12	31,585	95
Georgetown	1255	74	24	12	20,126	92
Harvard	1400	91	14	11	39,525	97
JohnsHopkins	1305	75	44	7	58,691	87
MIT	1380	94	30	10	34,870	91
Northwestern	1260	85	39	11	28,052	89
NotreDame	1255	81	42	13	15,122	94
PennState	1081	38	54	18	10,185	80
Princeton	1375	91	14	8	30,220	95
Purdue	1005	28	90	19	9,066	69
Stanford	1360	90	20	12	36,450	93
TexasA&M	1075	49	67	25	8,704	67
UCBerkeley	1240	95	40	17	15,140	78
UChicago	1290	75	50	13	38,380	87
UMichigan	1180	65	68	16	15,470	85
UPenn	1285	80	36	11	27,553	90
UVA	1225	77	44	14	13,349	92
UWisconsin	1085	40	69	15	11,857	71
Yale	1375	95	19	11	43,514	96

This dataset excludes image variables (student satisfaction, employer satisfaction, deans' opinions)

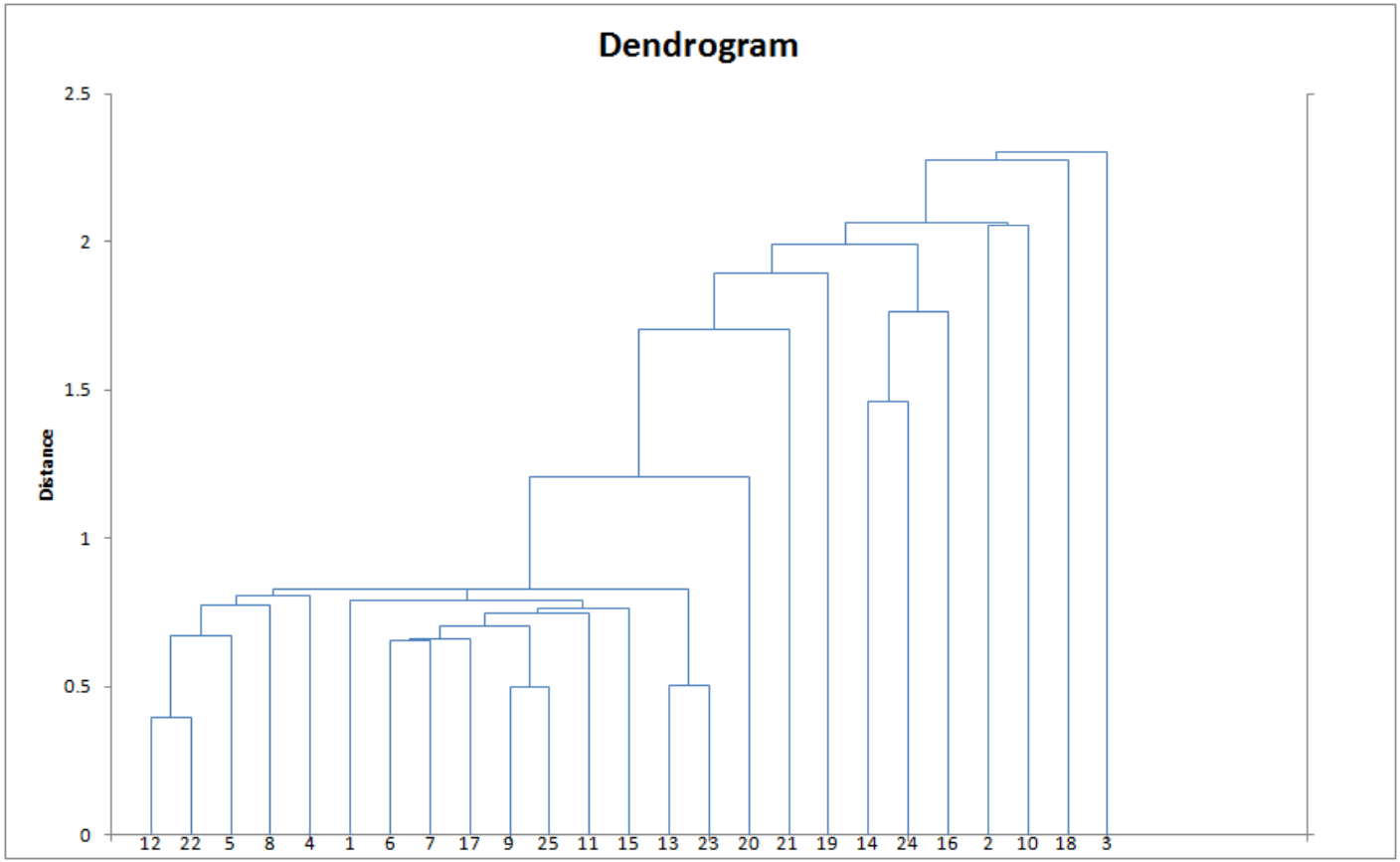
Dendrogram for Business School Example (XLMiner) with *Single Linkage*

- *XLMiner Platform > Cluster > Hierarchical Clustering*
- *Choose Data range (do not include 'Univ' column; why?)*
- *Choose input variables*
- *Next*
- *Normalize data, Single Linkage*
- *Draw dendrogram, Show cluster membership, #clusters 2*
- *Finish*
- *Click on **Help** if not clear at any step*

Dendrogram for Business School Example (XLMiner) with ***Single Linkage***

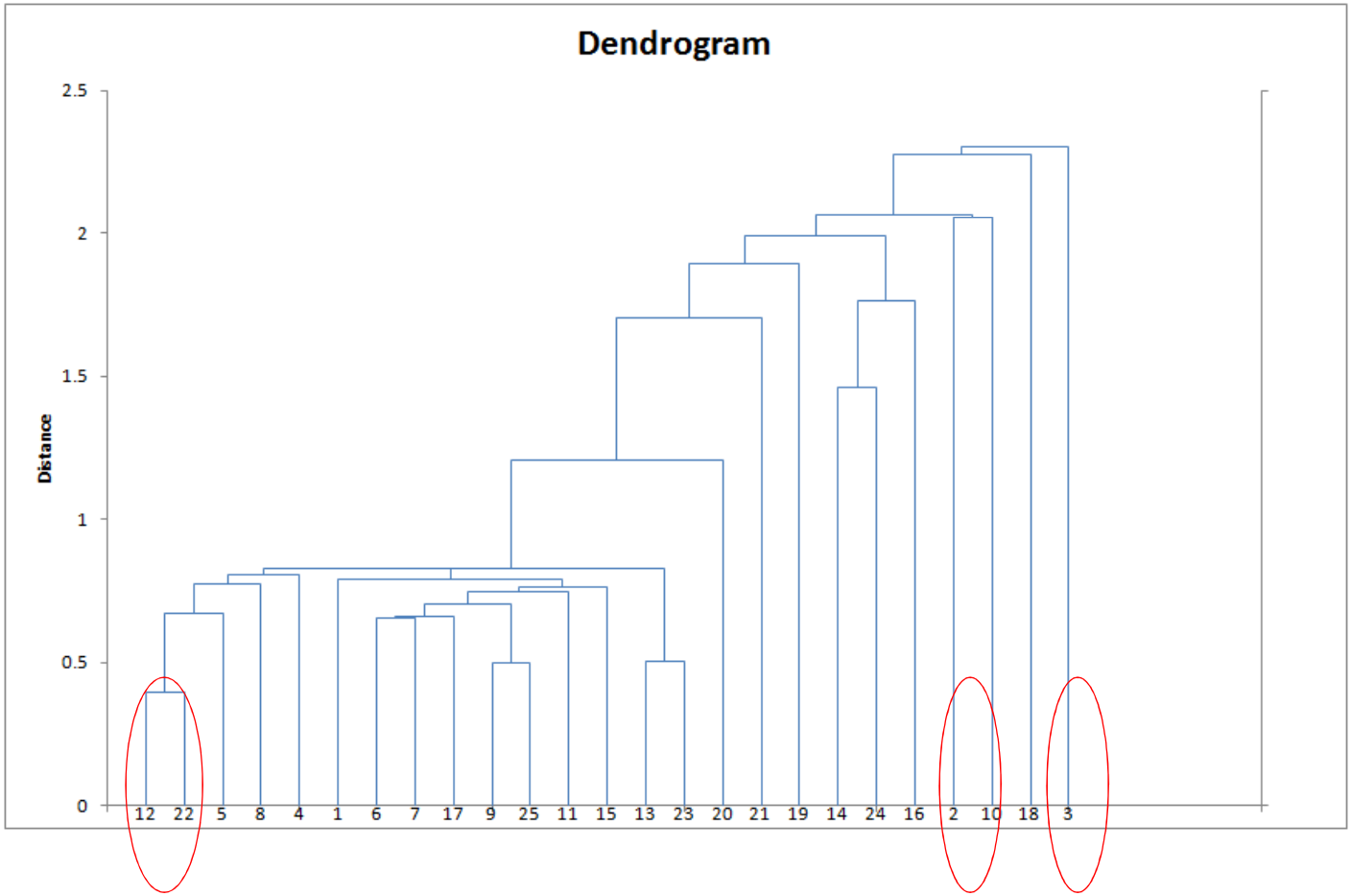
Worksheet:
HC_Dendrogram

Row Id.	University
1	Brown
2	CalTech
3	CMU
4	Columbia
5	Cornell
6	Dartmouth
7	Duke
8	Georgetown
9	Harvard
10	JohnsHopkins
11	MIT
12	Northwestern
13	NotreDame
14	PennState
15	Princeton
16	Purdue
17	Stanford
18	TexasA&M
19	UCBerkeley
20	UChicago
21	UMichigan
22	UPenn
23	UVA
24	UWisconsin
25	Yale



Insights? Anything Interesting?

Row Id.	University
1	Brown
2	CalTech
3	CMU
4	Columbia
5	Cornell
6	Dartmouth
7	Duke
8	Georgetown
9	Harvard
10	JohnsHopkins
11	MIT
12	Northwestern
13	NotreDame
14	PennState
15	Princeton
16	Purdue
17	Stanford
18	TexasA&M
19	UCBerkeley
20	UChicago
21	UMichigan
22	UPenn
23	UVA
24	UWisconsin
25	Yale



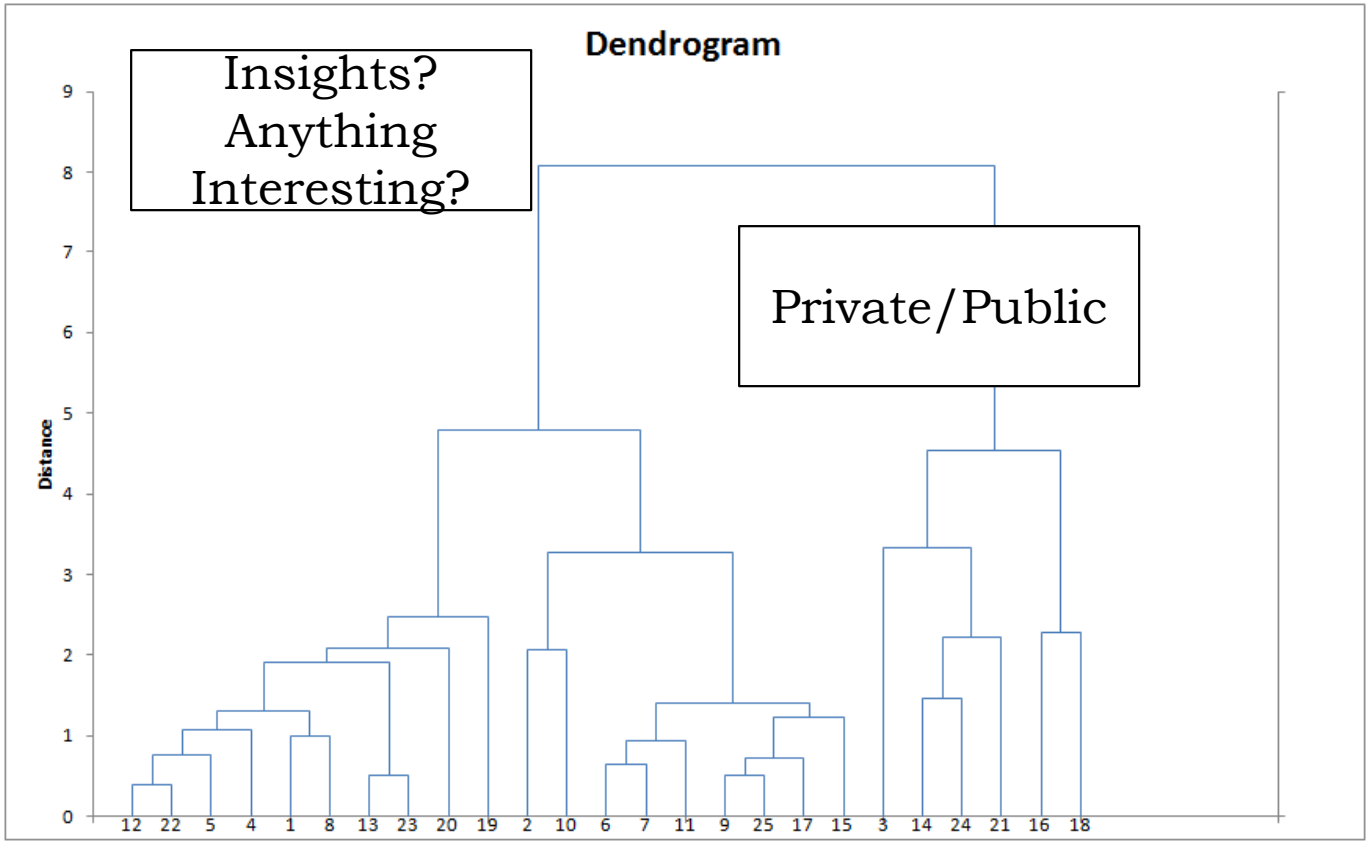
Dendrogram for Business School Example (XLMiner) with ***Complete Linkage***

- *XLMiner Platform > Cluster > Hierarchical Clustering*
- *Choose Data range (do not include 'Univ' column; why?)*
- *Choose input variables*
- *Next*
- *Normalize data,* ***Complete Linkage***
- *Draw dendrogram, Show cluster membership, #clusters 2*
- *Finish*
- *Click on* ***Help*** *if not clear at any step*

Dendrogram for Business School Example (XLMiner) with *Complete Linkage*

Row Id.	University	Cluster Id
1	Brown	1
2	CalTech	1
3	CMU	2
4	Columbia	1
5	Cornell	1
6	Dartmouth	1
7	Duke	1
8	Georgetown	1
9	Harvard	1
10	JohnsHopkins	1
11	MIT	1
12	Northwestern	1
13	NotreDame	1
14	PennState	2
15	Princeton	1
16	Purdue	2
17	Stanford	1
18	TexasA&M	2
19	UCBerkeley	1
20	UChicago	1
21	UMichigan	2
22	UPenn	1
23	UVA	1
24	UWisconsin	2
25	Yale	1

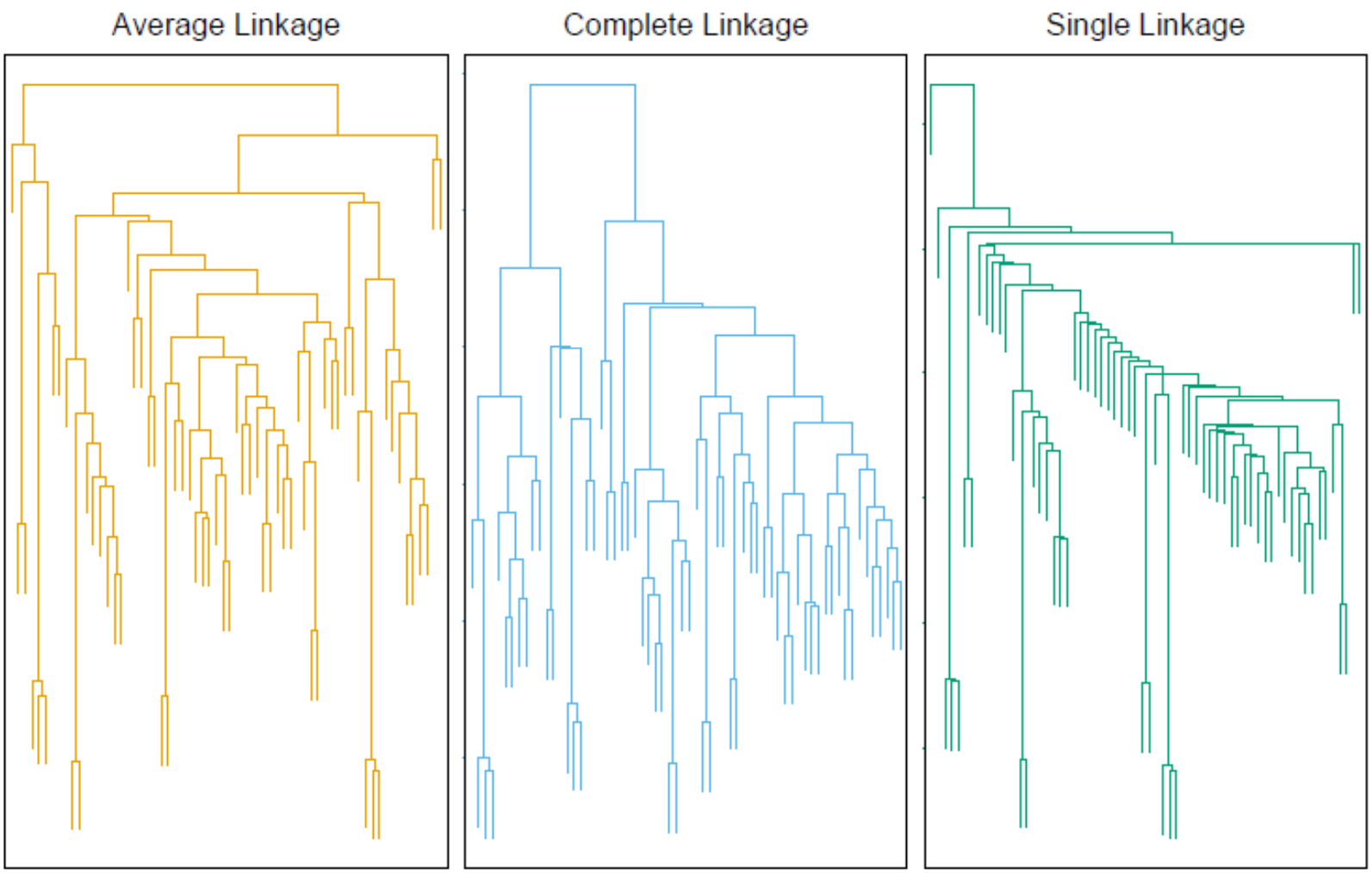
Worksheet: HC_Dendropgram1



Insights? Anything Interesting?

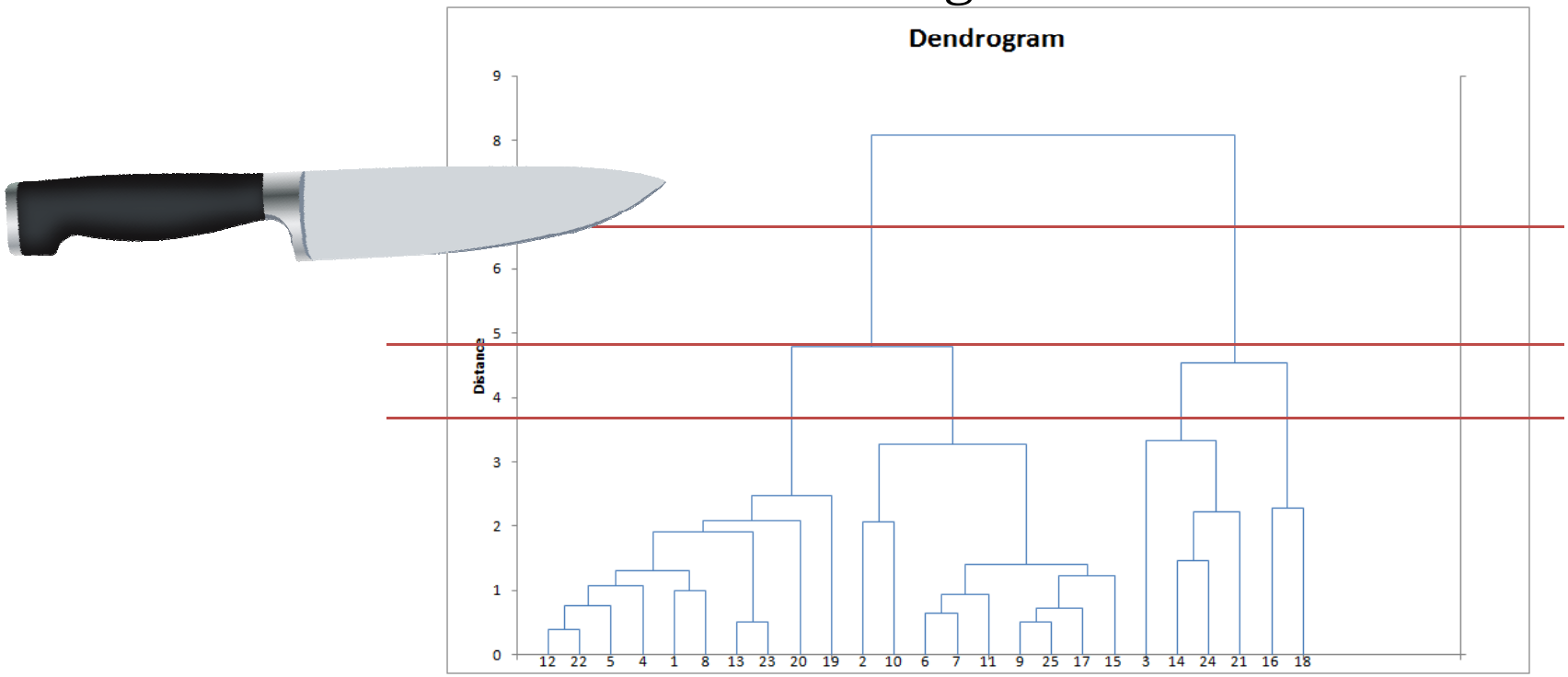
Private/Public

Comparing hierarchical algorithms



From Dendrograms to Clusters

- After dendrogram is obtained, **cut** it to create clusters. **How?**
- Examine *distance levels*
  - Cutpoint determines # clusters
  - Obtain statistics on resulting clusters



Run Hierarchical clustering again with ***Complete Linkage, and create 3 clusters***

- *XLMiner Platform > Cluster > Hierarchical Clustering*
- *Choose Data range (do not include 'Univ' column; why?)*
- *Choose input variables*
- *Next*
- *Normalize data,* ***Complete Linkage***
- *Draw dendrogram, Show cluster membership,* ***#clusters 3***
- *Finish*
- *Click on* ***Help*** *if not clear at any step*

Examine the clusters in Tableau

- *Copy-Paste the **cluster ID** column in HC_Clusters2 to Universities worksheet*
- *Save the file as "Universities_Clustering – with solution.xlsx"*
- *Open Tableau*
  - *Connect to data > Microsoft Excel > choose file "Universities_Clustering – with solution.xlsx" > choose worksheet "Universities" > Import all data*

http://kb.tableausoftware.com/articles/knowledgebase/measure-names-and-measure-values-explained

	A	B	C	D	E	F
1	**Date**	**Region**	**Sales**	**Profit**	**Discount**	
2	1/1/2009	East	$100	$50	0%	
3	1/2/2009	West	$300	$100	10%	
4	1/3/2009	Central	$500	$200	30%	
5	1/4/2009	East	$400	$160	40%	
6	1/5/2009	South	$600	$500	0%	
7	1/6/2009	West	$800	$750	0%	
8	1/7/2009	West	$400	$250	0%	
9	1/8/2009	Central	$100	$65	20%	
10	1/9/2009	East	$300	$254	50%	
11	1/10/2009	South	$200	$89	75%	
12	1/11/2009	South	$100	$40	30%	
13						

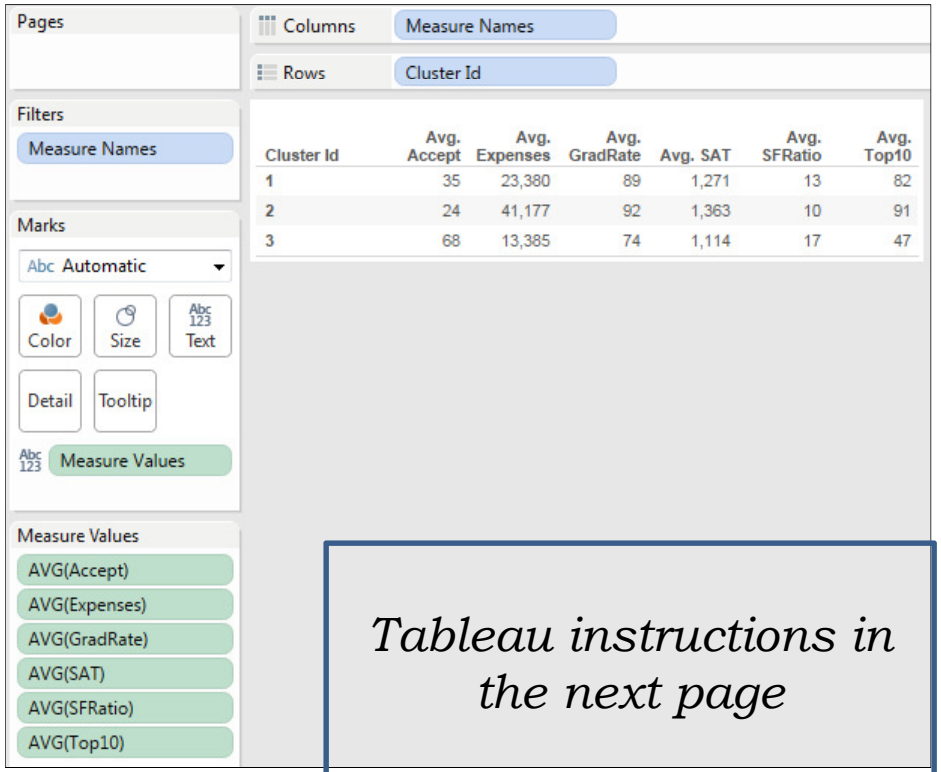
Measure Names

Sum of Sales
Sum of Profit
Sum of Discount

Measure Values

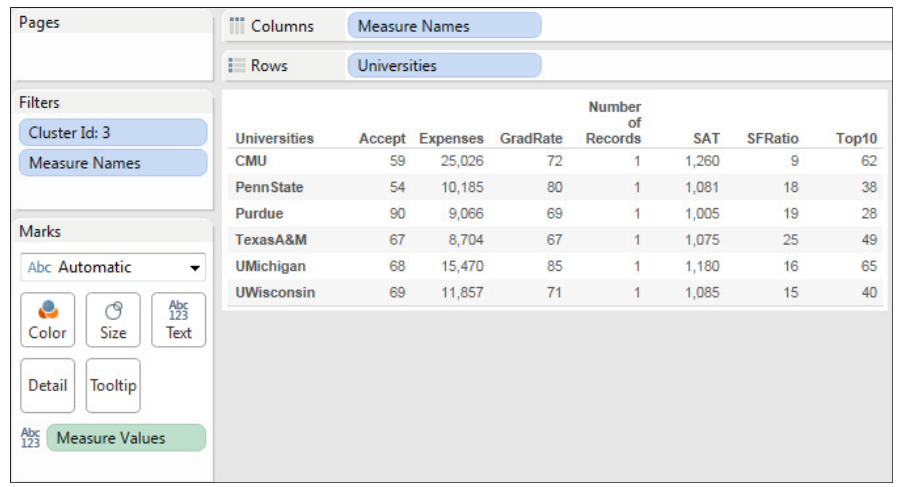
Sum of Sales	Sum of Profit	Sum of Discount
$100	$50	0%
$300	$100	10%
$500	$200	30%
$400	$160	40%
$600	$500	0%
$800	$750	0%
$400	$250	0%
$100	$65	20%
$300	$254	50%
$200	$89	75%
$100	$40	30%

1. What are the average acceptance rate, tuition fees, etc., across clusters?
2. What universities are there in those clusters?



Cluster Id	Avg. Accept	Avg. Expenses	Avg. GradRate	Avg. SAT	Avg. SFRatio	Avg. Top10
1	35	23,380	89	1,271	13	82
2	24	41,177	92	1,363	10	91
3	68	13,385	74	1,114	17	47

Tableau instructions in the next page



Universities	Accept	Expenses	GradRate	Number of Records	SAT	SFRatio	Top10
CMU	59	25,026	72	1	1,260	9	62
PennState	54	10,185	80	1	1,081	18	38
Purdue	90	9,066	69	1	1,005	19	28
TexasA&M	67	8,704	67	1	1,075	25	49
UMichigan	68	15,470	85	1	1,180	16	65
UWisconsin	69	11,857	71	1	1,085	15	40

Can you name/label those clusters?

Tableau Instructions

- *Move Cluster IDs to row area*
- *Move measure names to columns area.*
- *Move measure values to details area.*
- *Change the SUMs to AVGs.*

Tableau Instructions

- *Create a new worksheet*
- *Cluster IDs in Columns*
- *Universities in row*
- *Measure values in Detail*
- *Change Measure Values to Average*
- *Click on text tables*

- *You may filter on Cluster ID*

- You may also use Excel Data Filter combined with SUBTOTAL function in Excel

Row	Universities	Cluster Id	Sub Clus	SAT	Top	Acc	SFRa	Expens	GradR
1	Brown	1	1	1310	89	22	13	22704	94
4	Columbia	1	4	1310	76	24	12	31510	88
5	Cornell	1	5	1280	83	33	13	21864	90
8	Georgetown	1	8	1255	74	24	12	20126	92
12	Northwestern	1	12	1260	85	39	11	28052	89
13	NotreDame	1	13	1255	81	42	13	15122	94
19	UCBerkeley	1	19	1240	95	40	17	15140	78
20	UChicago	1	20	1290	75	50	13	38380	87
22	UPenn	1	22	1285	80	36	11	27553	90
23	UVA	1	23	1225	77	44	14	13349	92
			Min	1225	74	22	11	13349	78
			Average	1271	81.5	35.4	12.9	23380	89.4
			Max	1310	95	50	17	38380	94

Insights in terms of distribution, outliers?

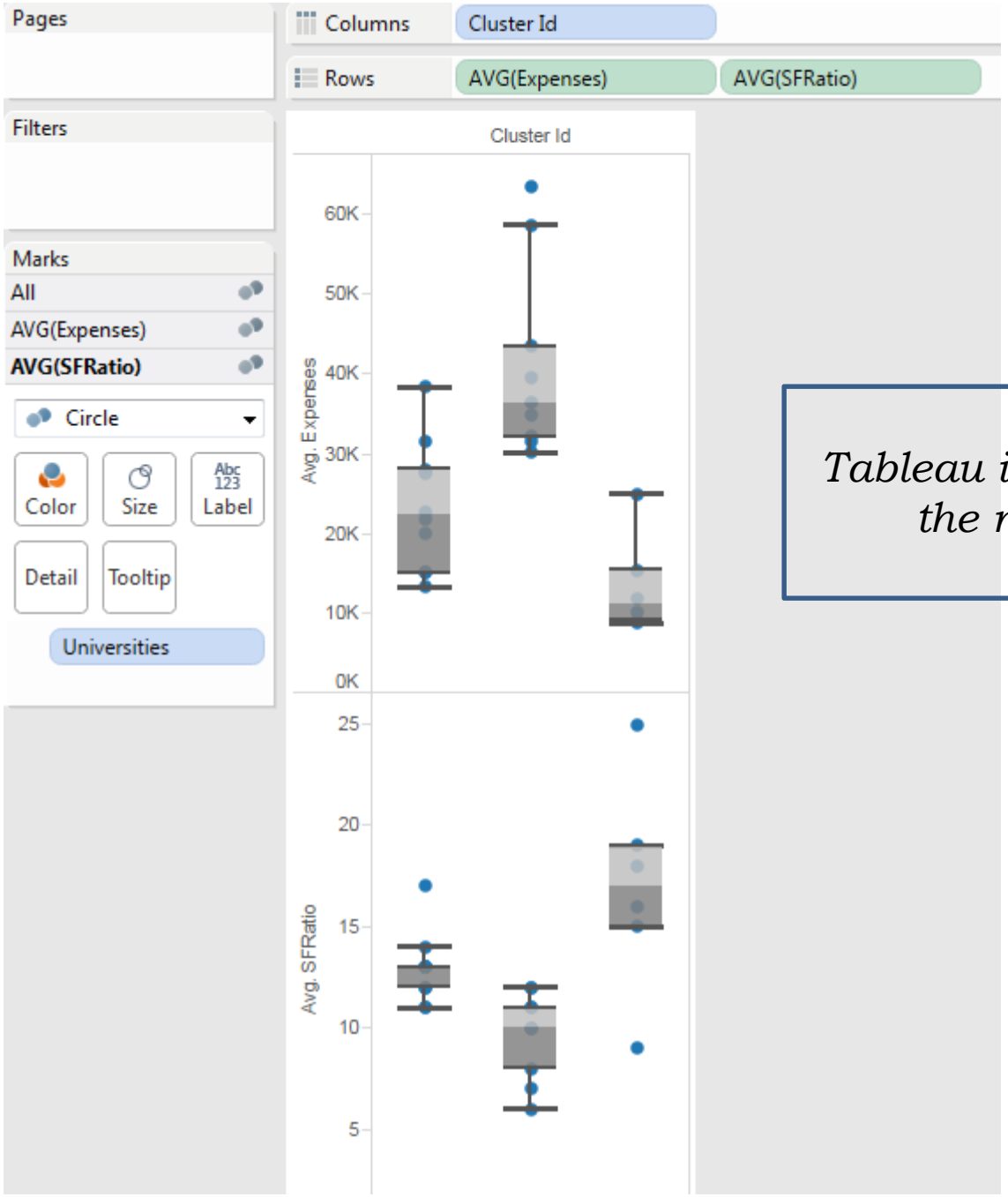


Tableau instructions in the next page

Tableau Instructions

- *Create a new worksheet*
- *Cluster IDs in the columns area*
- *Measure values in the row area*
- *Measure Values > Filter > select Expense and SFRatio*
- *Change the SUMs to AVGs.*
- *Universities in the Details area*
- *Click on Box plot from the right panel*
-
- *Lower Whisker* = $Q_1 - 1.5*(Q_3-Q_1)$
- *Upper Whisker* = $Q_3 + 1.5*(Q_3-Q_1)$

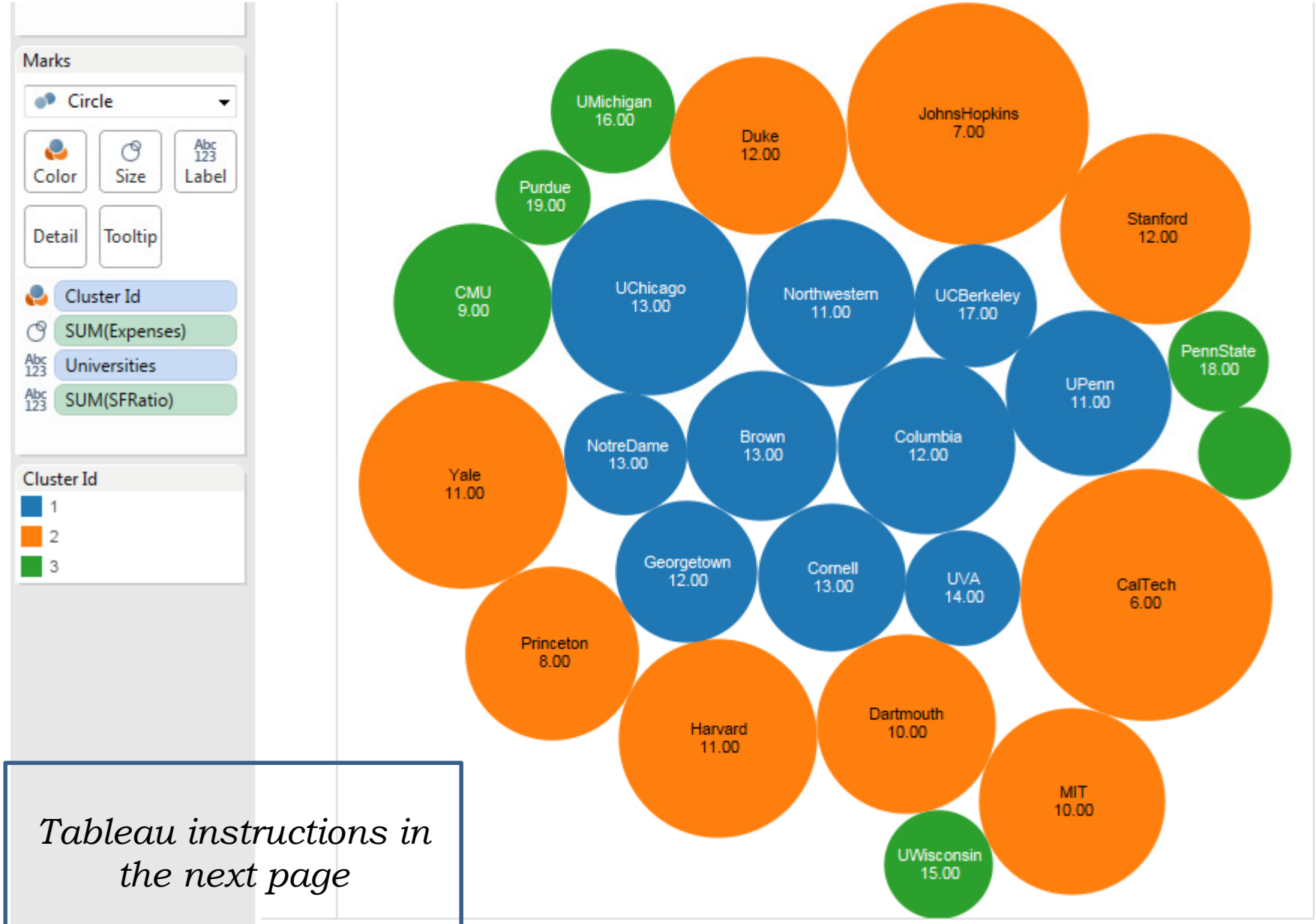


Tableau instructions in the next page

Tableau Instructions

- *Create a new worksheet*
- *Cluster IDs on color*
- *Universities in Detail*
- *Avg(Accept) on Size*
- *Click on Bubbles chat*

Is the clustering insightful?

Based on the interpretation, assign a ***name*** to each cluster

PRIZM NE Segmentation System Order B

01 Upper Crust	18 Kids & Cul-de-Sacs	35 Boomtown Singles
02 Blue Blood Estates	19 Home Sweet Home	36 Blue-Chip Blues
03 Movers & Shakers	20 Fast-Track Families	37 Mayberry-ville
04 Young Digerati	21 Gray Power	38 Simple Pleasures
05 Country Squires	22 Young Influentials	39 Domestic Duos
06 Winner's Circle	23 Greenbelt Sports	40 Close-In Couples
07 Money & Brains	24 Up-and-Comers	41 Sunset City Blues
08 Executive Suites	25 Country Casuals	42 Red, White & Blues

Final checks:

- Cluster **stability:** do cluster assignments change dramatically if some inputs are slightly altered?
- Cluster **separation:** compare between-cluster variation to within-cluster variation

Hierarchical Clustering: Advantages & Disadvantages

The Good

- Finds "natural" grouping – no need to specify number of clusters
- Dendrogram: transparency of process, good for presentation

The Bad

- Require computation & storage of *n x n distance matrix*
- Algorithm makes only one pass through the data. Records that are incorrectly allocated early on cannot be reallocated subsequently
- Low stability: Reordering data or dropping a few records can lead to different solution
- Single+complete linkage robust to distance metric as long as the relative ordering is kept. Average linkage is NOT.
- Most distances sensitive to outliers