Introduction to
Data Science and Analytics

Stephan Sorger

www.StephanSorger.com

Unit 8. R Segmentation

Lecture: Introduction

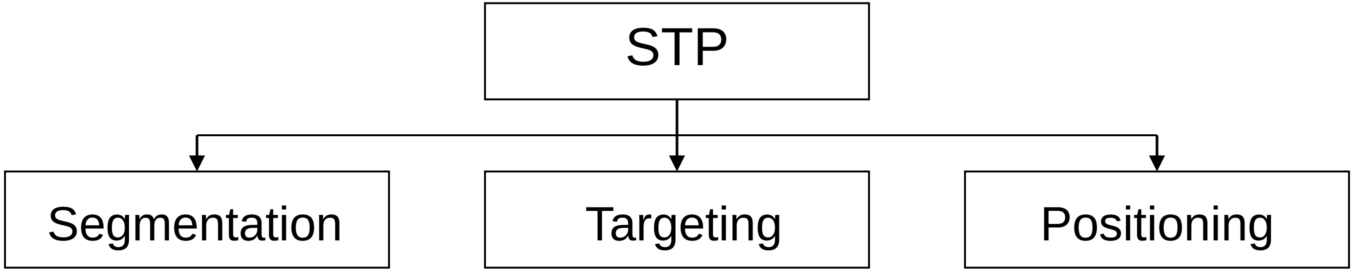
Disclaimer:

- All images such as logos, photos, etc. used in this presentation are the property of their respective copyright owners and are used here for educational purposes only
- Some material adapted from: Sorger, “Marketing Analytics: Strategic Models and Metrics”

Outline/ Learning Objectives

Topic	Description
Introduction	Overview of market segmentation, targeting, and positioning
A Priori	Comparison of A Priori and Post Hoc approaches
Techniques	Overview of different segmentation techniques
Naïve Bayes	Brief review of Naïve Bayes classification approach
Clusters	Discussion of cluster analysis for segmentation
R	Segmentation using R: K-means; Ward’s methods

STP: Segmentation, Targeting, Positioning

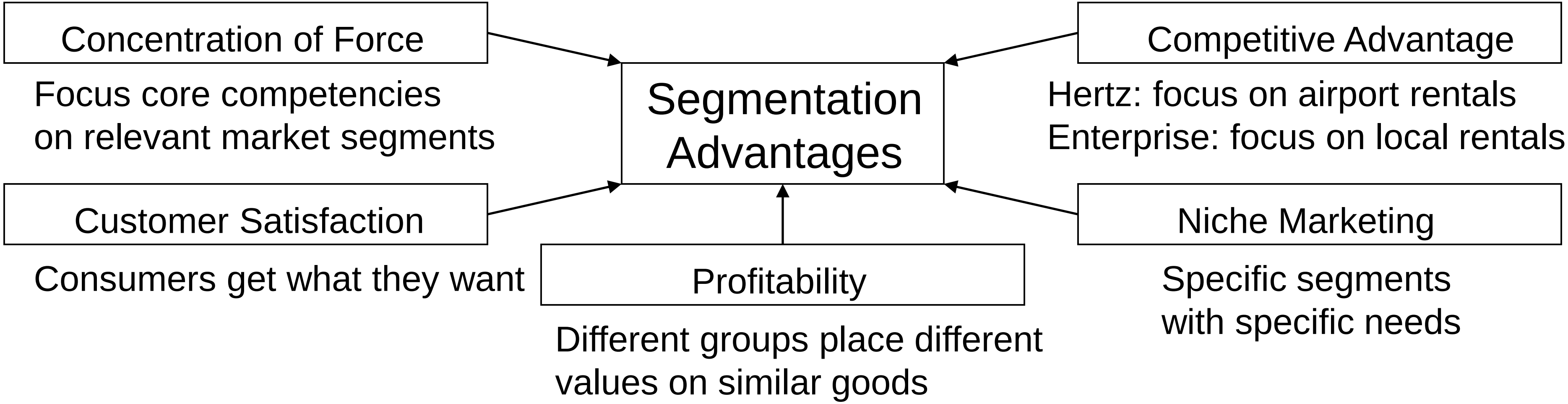


Segmentation:
Subdividing general markets into distinct segments with different needs, and which respond differently to marketing efforts.
-Increased customer satisfaction
-Increased marketing effectiveness

Targeting:
Selection of market segments. Cannot service every possible segment.

Positioning:
Activities to make consumers perceive that a brand occupies a distinct position relative to competing brands.

Segmentation Advantages



Segmentation Advantages

Concentration of Force

Focus core competencies
on relevant market segments

Customer Satisfaction

Consumers get what they want

Profitability

Different groups place different
values on similar goods

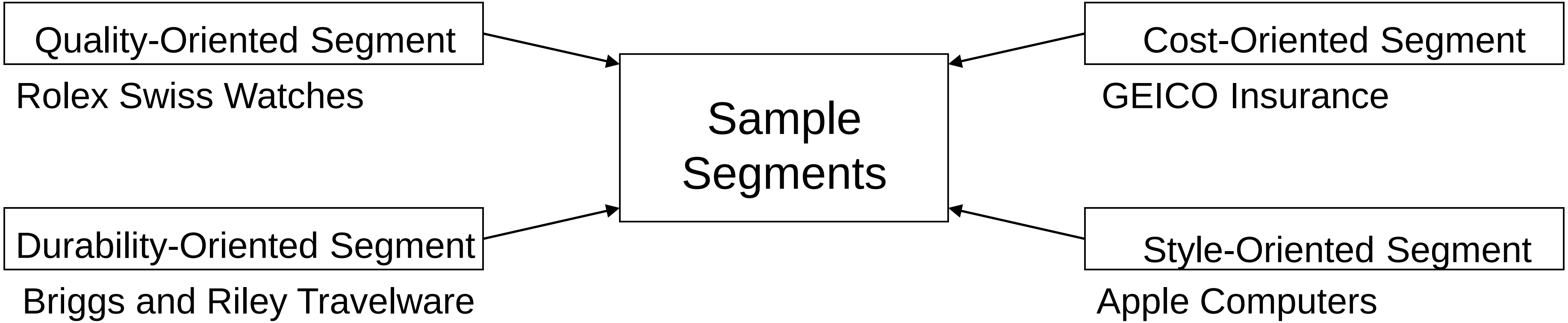
Competitive Advantage

Hertz: focus on airport rentals
Enterprise: focus on local rentals

Niche Marketing

Specific segments
with specific needs

Sample Segments



Quality-Oriented Segment

Rolex Swiss Watches

Durability-Oriented Segment

Briggs and Riley Travelware

Sample Segments

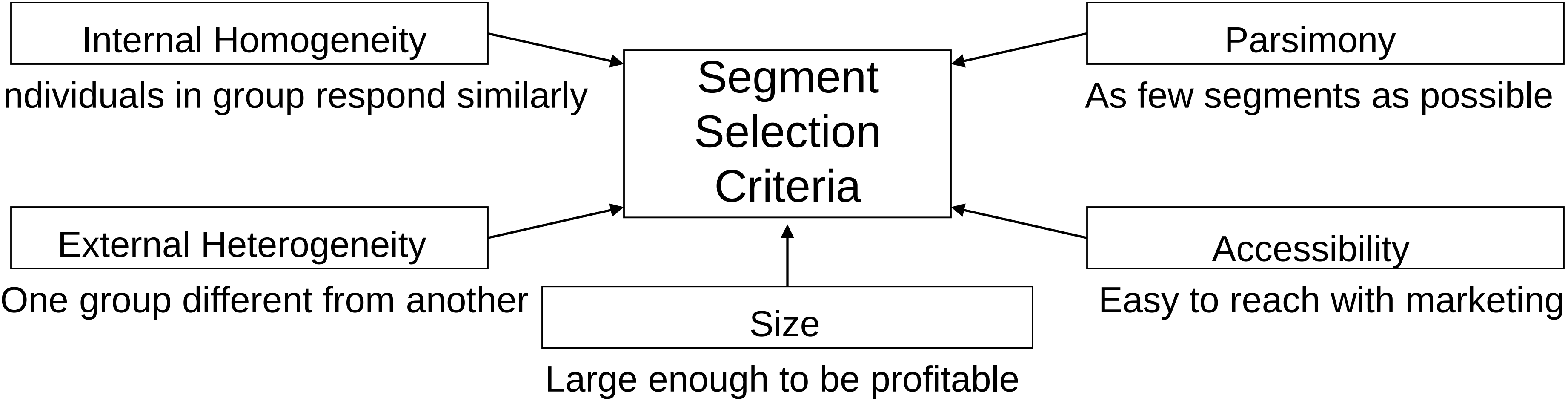
Cost-Oriented Segment

GEICO Insurance

Style-Oriented Segment

Apple Computers

Segment Selection Criteria



Internal Homogeneity

Individuals in group respond similarly

External Heterogeneity

One group different from another

Segment Selection Criteria

Size

Large enough to be profitable

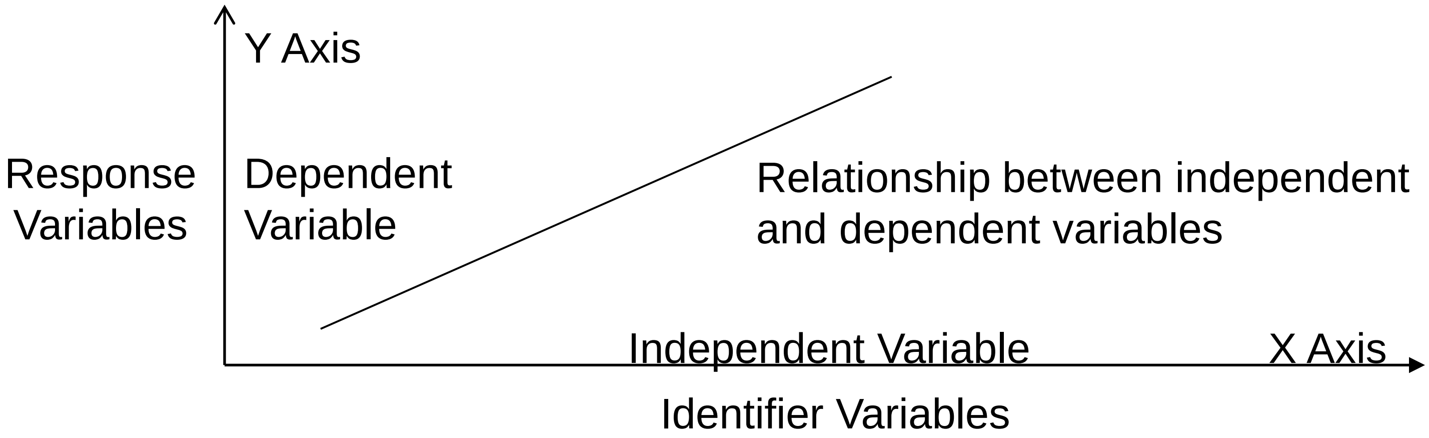
Parsimony

As few segments as possible

Accessibility

Easy to reach with marketing

Segmentation Variables



Y Axis

Response Variables

Dependent Variable

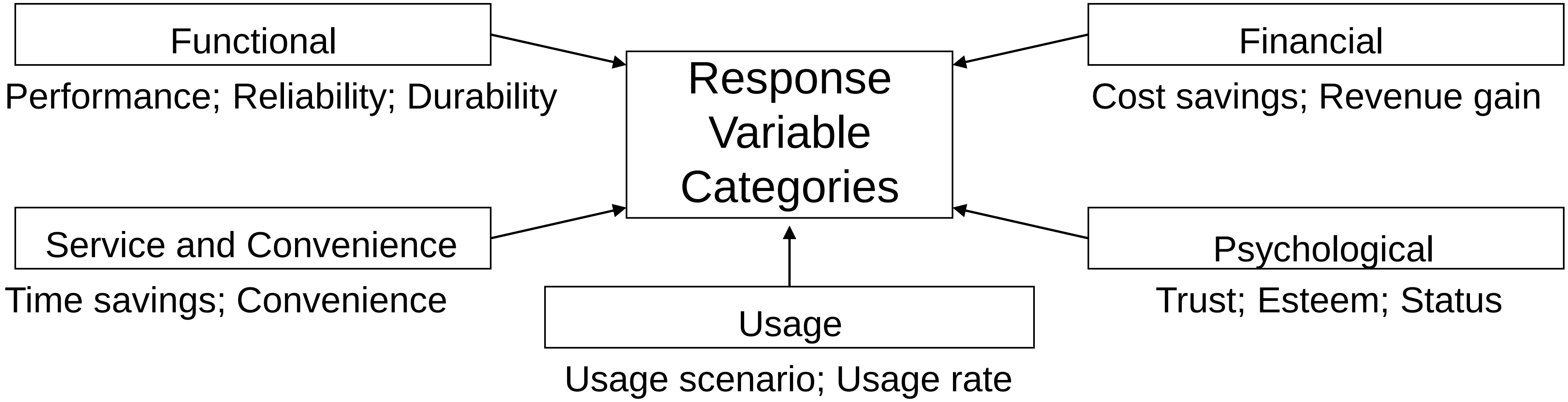
Relationship between independent and dependent variables

Independent Variable

X Axis

Identifier Variables

Response (Dependent) Variable Categories



Functional

Performance; Reliability; Durability

Service and Convenience

Time savings; Convenience

Response Variable Categories

Usage

Usage scenario; Usage rate

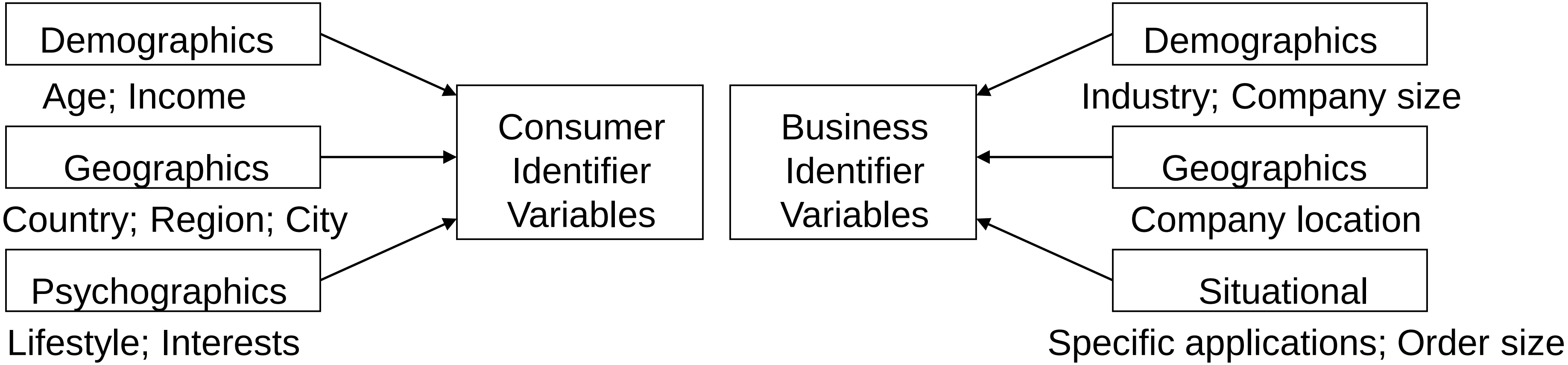
Financial

Cost savings; Revenue gain

Psychological

Trust; Esteem; Status

Segmentation Identifier (Independent) Variables



Demographics

Age; Income

Geographics

Country; Region; City

Psychographics

Lifestyle; Interests

Consumer Identifier Variables

Business Identifier Variables

Demographics

Industry; Company size

Geographics

Company location

Situational

Specific applications; Order size

Introduction to
Data Science and Analytics

Stephan Sorger

www.StephanSorger.com

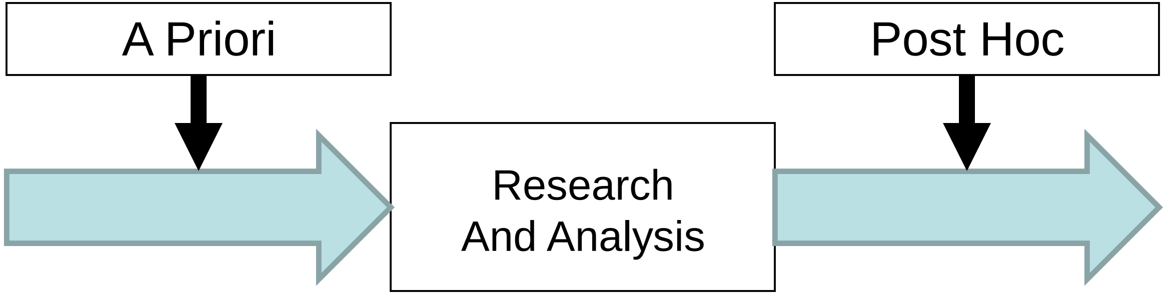
Unit 8. R Segmentation

Lecture: A Priori and Techniques Overview

Disclaimer:

- All images such as logos, photos, etc. used in this presentation are the property of their respective copyright owners and are used here for educational purposes only
- Some material adapted from: Sorger, “Marketing Analytics: Strategic Models and Metrics”

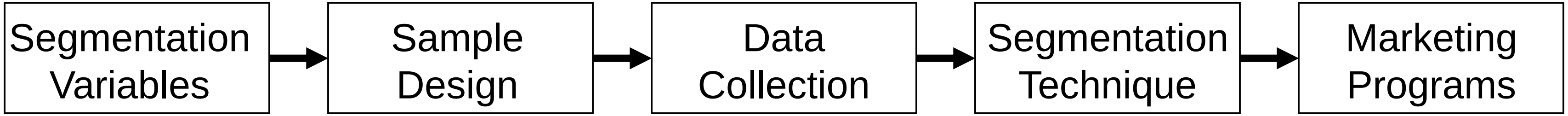
Segmentation Approaches: A Priori vs. Post Hoc



Latin: “From Before”
Segments defined before primary market research and analysis

Latin: “After This”
Segments defined after primary market research and analysis

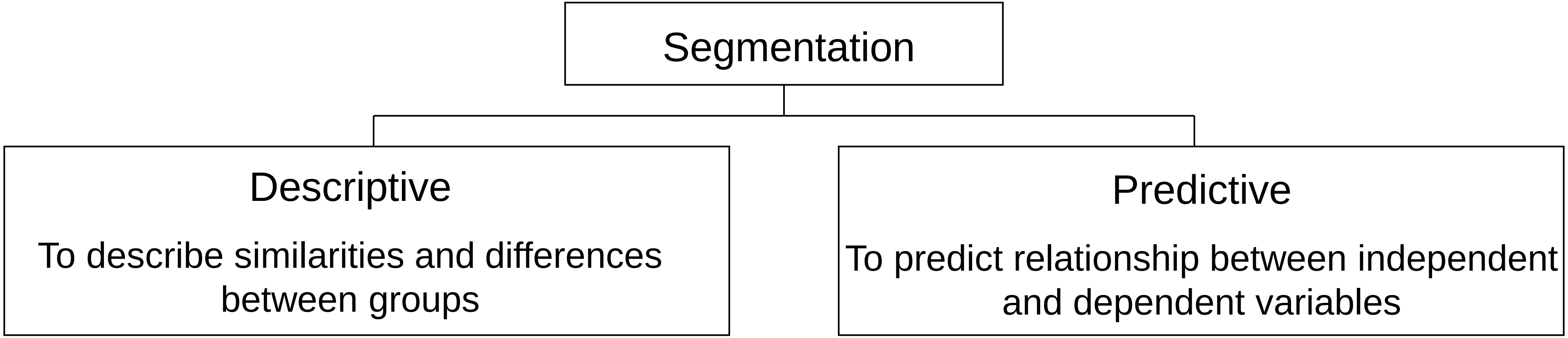
A Priori Market Segmentation Process



Segmentation Variables → Sample Design → Data Collection → Segmentation Technique → Marketing Programs

Step	Description
Segmentation Variables	Response Variable: Usage rate, etc. Identifier Variable: Age; Income; etc.
Sample Design	Large surveys: Often use random sample Small surveys: Often use non-random
Data Collection	Online survey tools: SurveyMonkey, etc.
Segmentation Technique	Cross-tab; Regression; etc.
Marketing Program	Leverage information known about segment

Segmentation: Descriptive vs. Predictive



Segmentation

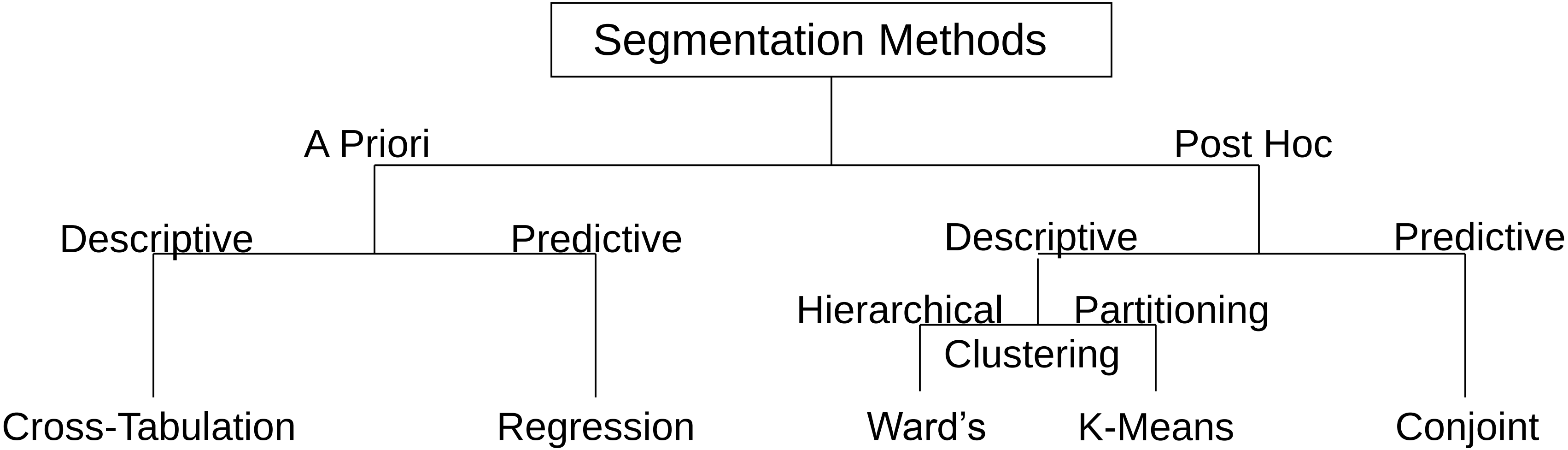
Descriptive

To describe similarities and differences between groups

Predictive

To predict relationship between independent and dependent variables

Segmentation: Analytic Techniques



Segmentation Methods

A Priori

Post Hoc

Descriptive

Predictive

Descriptive

Predictive

Hierarchical

Partitioning

Clustering

Cross-Tabulation

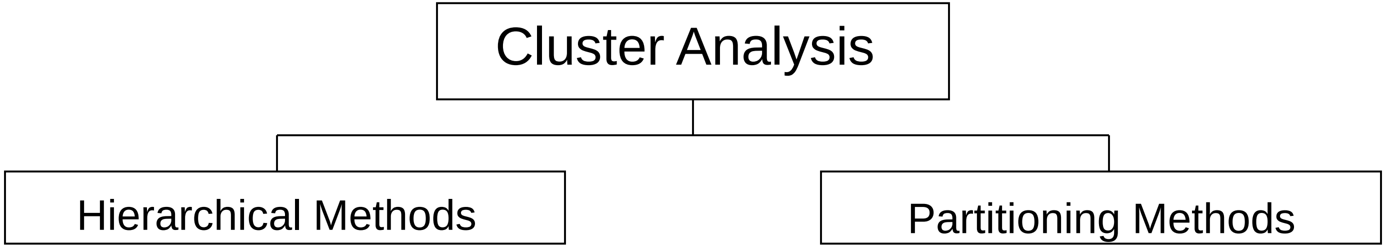
Regression

Ward’s

K-Means

Conjoint

Segmentation: Cluster Analysis



Cluster Analysis

Hierarchical Methods

Partitioning Methods

Example: Ward’s

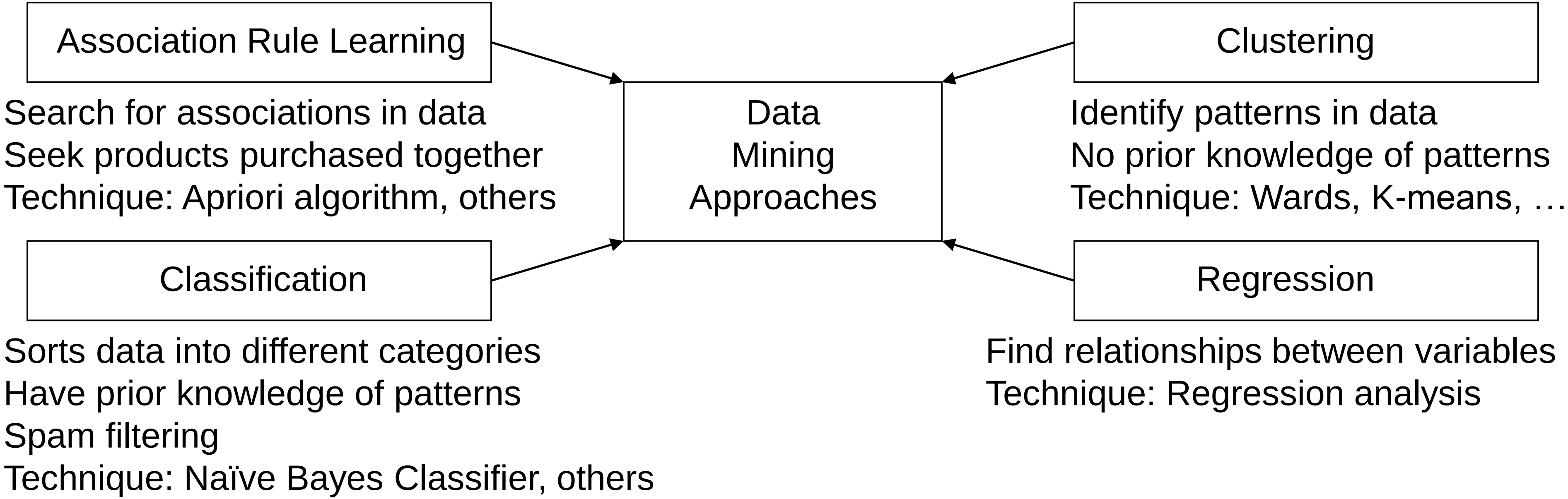
Example: K-Means

Ward’s Method:
Agglomerative hierarchical clustering
Groups clusters in hierarchy, from bottom up
Result is a tree-like diagram (dendogram)

K-Means:
Specify K, the number of final clusters to expect
Execute K-Means algorithm
Forms groups based on “distance” from “centroid”

Mathematics and algorithms of Cluster Analysis are complex;
Use cluster analysis built into R, SAS, SPSS, and other packages

Segmentation: Review of Data Mining Approaches



Association Rule Learning

Search for associations in data
Seek products purchased together
Technique: Apriori algorithm, others

Classification

Sorts data into different categories
Have prior knowledge of patterns
Spam filtering
Technique: Naïve Bayes Classifier, others

Data
Mining
Approaches

Clustering

Identify patterns in data
No prior knowledge of patterns
Technique: Wards, K-means, …

Regression

Find relationships between variables
Technique: Regression analysis

Introduction to
Data Science and Analytics

Stephan Sorger

www.StephanSorger.com

Unit 8. R Segmentation

Lecture: Naïve Bayes

Disclaimer:

- All images such as logos, photos, etc. used in this presentation are the property of their respective copyright owners and are used here for educational purposes only
- Some material adapted from: Sorger, “Marketing Analytics: Strategic Models and Metrics”

Classification: Naïve Bayes Classifier

Topic	Discussion
Naïve	Strong (naïve) independence assumptions between sets
Bayes	Thomas Bayes, b. 1701, English statistician and minister Developed Bayes theorem
Classifier	Sorts data based on probability
Applications	Spam filtering Text categorization: sports or politics? Medical diagnostics

T. Bayes.

Classification: Bayes Theorem

Topic	Discussion
Purpose	Converts results from tests into probability of events
Equation	True positive result, divided by chance of any positive result Pr(X)=chance of getting any positive result Chances of event A, given X, written as Pr(A|X)
Example	Next slide

$$\Pr(A|X) = \frac{\Pr(X|A) * \Pr(A)}{\Pr(X)}$$

Source: http://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/

Classification: Bayes Theorem

Topic	Discussion
Example	What is the probability it will rain during Alex’s wedding?
Given data	1. Alex getting married tomorrow outdoors in Palm Springs 2. Palm Springs: Rains 5 days/ year, on average 3. Weather app predicts rain for tomorrow 4. When it rains, weather app is correct 90% of the time 5. When it doesn’t rain, weather app is incorrect 10% of time

Event A1:	It does rain on Alex’s wedding
Event A2:	It does not rain on Alex’s wedding
Event B:	Weather app predicts rain
Problem:	P(A1|B): Probability of raining, given rain prediction

Source: http://stattrek.com/probability/bayes-theorem.aspx

Classification: Bayes Theorem

Topic	Discussion
Example	What is the probability it will rain during Alex’s wedding?
Given data	1. Alex getting married tomorrow outdoors in Palm Springs 2. Palm Springs: Rains 5 days/ year, on average 3. Weather app predicts rain for tomorrow 4. When it rains, weather app is correct 90% of the time 5. When it doesn’t rain, weather app is incorrect 10% of time

Event A1: It does rain on Alex’s wedding → P(A1) = 5/365 =0.014 (rains 5 days/year)
Event A2: It does not rain on Alex’s wedding → P(A2) = 360/365 = 0.986 (doesn’t rain)
Event B: Weather app predicts rain
P(B|A1) = 0.9 → When it does rain, weather app predicts rain 90% of the time
P(B|A2) = 0.1 → When it does not rain, weather app predicts rain 10% of the time

Source: http://stattrek.com/probability/bayes-theorem.aspx

Classification: Bayes Theorem

Topic	Discussion
Example	What is the probability it will rain during Alex's wedding?
Given data	1. Alex getting married tomorrow outdoors in Palm Springs 2. Palm Springs: Rains 5 days/ year, on average 3. Weather app predicts rain for tomorrow 4. When it rains, weather app is correct 90% of the time 5. When it doesn't rain, weather app is incorrect 10% of time

$P(A1\ldots\text{does rain}) = 5/365 = 0.014$ (rains 5 days/year)
$P(A2\ldots\text{does not rain}) = 360/365 = 0.986$ (doesn't rain)
$P(B|A1) = 0.9$; $P(B|A2) = 0.1$

$$\begin{aligned} P(A1|B) &= P(A1) * P(B|A1) / [\, P(A1) * P(B|A1) + P(A2) * P(B|A2) \,] \\ &= (0.014) * (0.9) / [\, (0.014) * (0.9) + (0.986) * (0.1) \,] \\ &= 0.111 \end{aligned}$$

← Even when weather app predicts rain, it only rains 11% of the time

Source: http://stattrek.com/probability/bayes-theorem.aspx

Classification: Naïve Bayes Classifier

Topic	Discussion
Spam Filtering	Event A: The message is spam Test X: The message contains certain words (free, Viagra)
Blacklist	Too restrictive: Many false positives Example: “Free introductory class on R techniques”
Bayes	Middle ground: Uses probabilities to compute chance of spam Rather than Yes/No decision 99.9% chance of spam → Classify “spam” Gets better over time with “training”

Source: http://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/

Introduction to
Data Science and Analytics

Stephan Sorger

www.StephanSorger.com

Unit 8. R Segmentation

Lecture: Cluster Analysis with R

Disclaimer:

- All images such as logos, photos, etc. used in this presentation are the property of their respective copyright owners and are used here for educational purposes only
- Some material adapted from: Sorger, “Marketing Analytics: Strategic Models and Metrics”

Segmentation and R

Topic	Discussion
R Power	Advanced market segmentation: Good application for R R features more specialized functions than Excel R features more advanced data handling than Excel
Demographic	Traditional segmentation: Demographic, Geographic, etc. Excel sufficient; Sort by age, Sort by ZIP code, etc.
Psychographic	Modern segmentation methods: Psychographic, etc. Need more powerful tools, such as R
Clusters	Given a general set of data, can we identify clusters? Groups of people in market who behave similarly

Cluster-Based Segmentation Example: Introduction

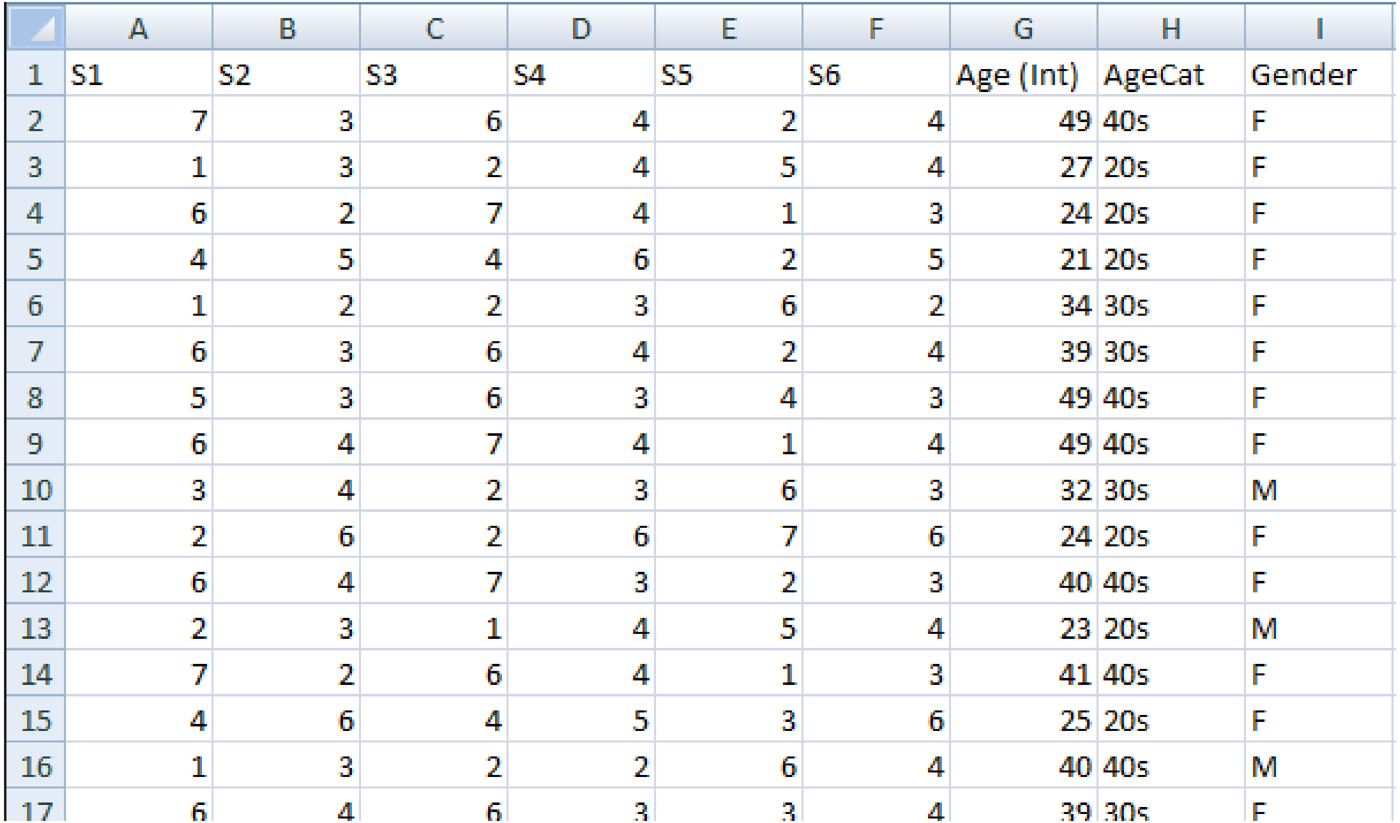
Topic	Discussion
Acme Dog	You are the marketing manager for Acme Dog Nutrition Organic, gluten-free food for active dogs
Groups	You seek to identify groups among dog owners
Market Survey	You conduct a market survey using a 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree)

Cluster-Based Segmentation Example: Survey

Topic	Discussion
Acme Dog	You are the marketing manager for Acme Dog Nutrition Organic, gluten-free food for active dogs
Groups	You seek to identify groups among dog owners
Market Survey	You conduct a market survey using a 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree)

S1: It is important for me to buy dog food that prevents canine cavities
S2: I like dog food that gives my dog a shiny coat
S3: Dog food should strengthen gums
S4: Dog food should make my dog's breath fresher
S5: It is not a priority for me that dog food prevent tooth decay or cavities (reverse coded)
S6: When I buy dog food, I look for food that gives my dog shiny teeth

Cluster-Based Segmentation Example: Dataset



	A	B	C	D	E	F	G	H	I
1	S1	S2	S3	S4	S5	S6	Age (Int)	AgeCat	Gender
2	7	3	6	4	2	4	49	40s	F
3	1	3	2	4	5	4	27	20s	F
4	6	2	7	4	1	3	24	20s	F
5	4	5	4	6	2	5	21	20s	F
6	1	2	2	3	6	2	34	30s	F
7	6	3	6	4	2	4	39	30s	F
8	5	3	6	3	4	3	49	40s	F
9	6	4	7	4	1	4	49	40s	F
10	3	4	2	3	6	3	32	30s	M
11	2	6	2	6	7	6	24	20s	F
12	6	4	7	3	2	3	40	40s	F
13	2	3	1	4	5	4	23	20s	M
14	7	2	6	4	1	3	41	40s	F
15	4	6	4	5	3	6	25	20s	F
16	1	3	2	2	6	4	40	40s	M
17	6	4	6	3	3	4	39	30s	F

Dataset: Survey results from 45 respondents, plus age and sex categories

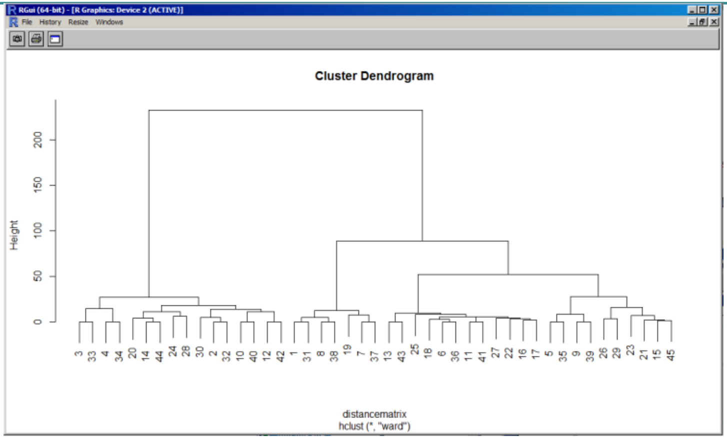
Cluster-Based Segmentation Example: Exercise

Topic	Discussion
1.	Using Wards Agglomerative Hierarchical Clustering, estimate the number of meaningful clusters present in the data
2.	Describe the resulting clusters so you can market to them State the messaging you would use for each segment
3.	Research actual segments used by dog food industry Compare those segments with segments you identified

Cluster-Based Segmentation Example: Exercise

Topic	Discussion
Wards	Apply Wards Agglomerative Hierarchical Clustering “Agglomerative” in that it gathers (agglomerates) data points “Hierarchical”: Smaller groups reporting to larger groups
Dendogram	Plot of data showing potential clusters Great visualization tool

Sample Dendogram



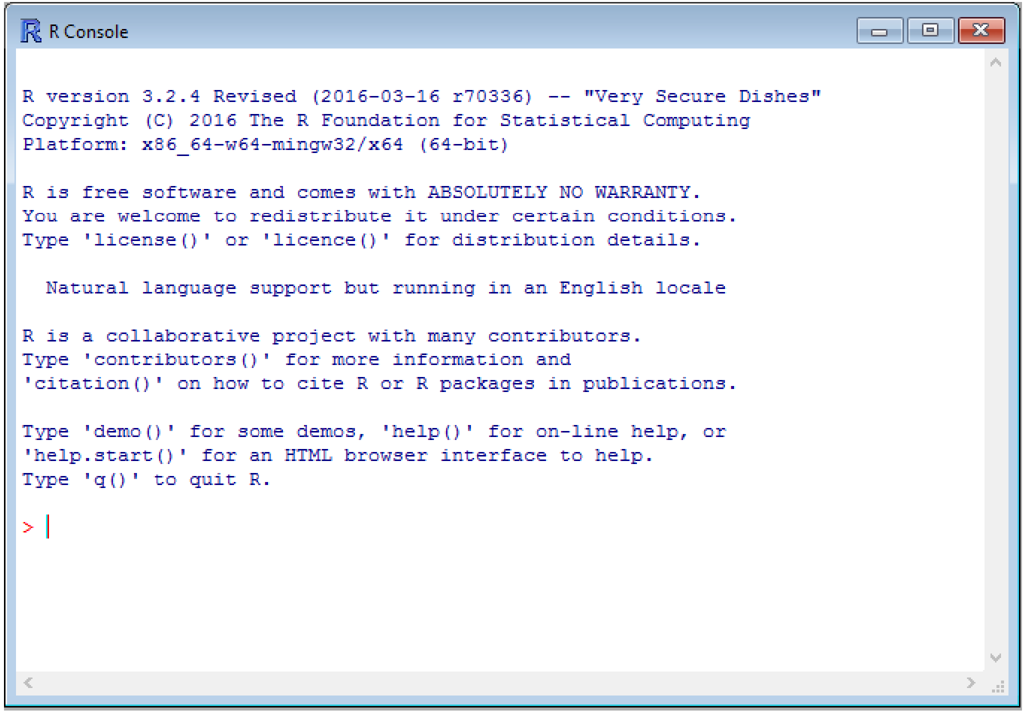
Cluster-Based Segmentation Example: Download R

Platform	Link
Windows	http://cran.r-project.org/bin/windows/base/
Mac	http://cran.r-project.org/bin/macosx/

Cluster-Based Segmentation Example: Launch R

Topic	Discussion
Prompt	You will see a “>” prompt in the “R Console”

You will be typing commands at the prompt: “ > “

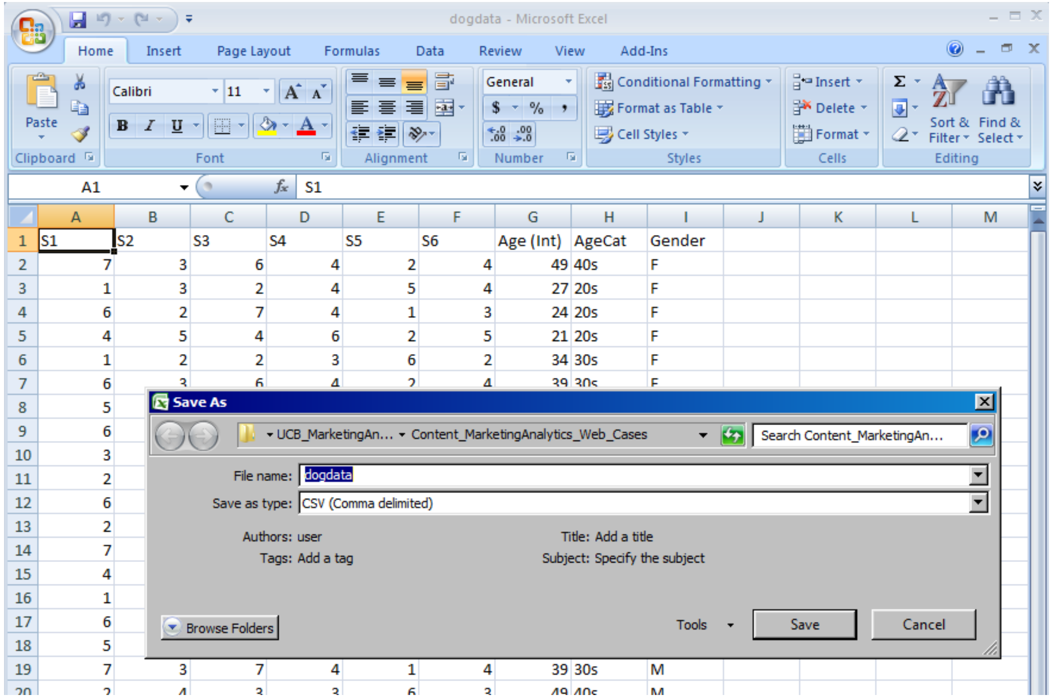


```
R version 3.2.4 Revised (2016-03-16 r70336) -- "Very Secure Dishes"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

Cluster-Based Segmentation Example: Prepare Data File

Topic	Discussion
Data File	Open data file, delete intro portion, save as CSV

Save as CSV



A1 fx S1

	A	B	C	D	E	F	G	H	I
1	S1	S2	S3	S4	S5	S6	Age (Int)	AgeCat	Gender
2	7	3	6	4	2	4	49	40s	F
3	1	3	2	4	5	4	27	20s	F
4	6	2	7	4	1	3	24	20s	F
5	4	5	4	6	2	5	21	20s	F
6	1	2	2	3	6	2	34	30s	F
7	6	3	6	4	2	4	39	30s	F
8	5								
9	6								
10	3								
11	2								
12	6								
13	2								
14	7								
15	4								
16	1								
17	6								
18	5								
19	7	3	7	4	1	4	39	30s	M
20	2	4	3	3	6	3	49	40s	M

Save As

UCB_MarketingAn... ▸ Content_MarketingAnalytics_Web_Cases

Search Content_MarketingAn...

File name: dogdata

Save as type: CSV (Comma delimited)

Authors: user

Title: Add a title

Tags: Add a tag

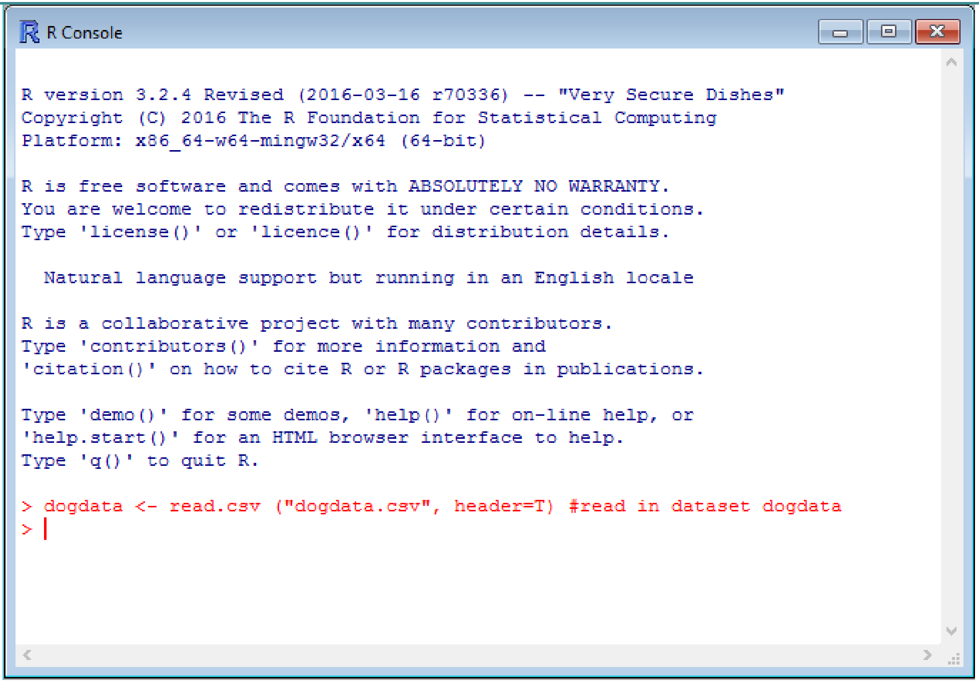
Subject: Specify the subject

Cluster-Based Segmentation Example: Read Data

Topic	Discussion

Read Data	dogdata<-read.csv(“C:\\Users\\user\\Desktop\\dogdata.csv”, header=T)
	dogdata<-read.csv(“dogdata.csv”, header=T) ← With working directory

Find out full filename, then insert filename into read.csv command

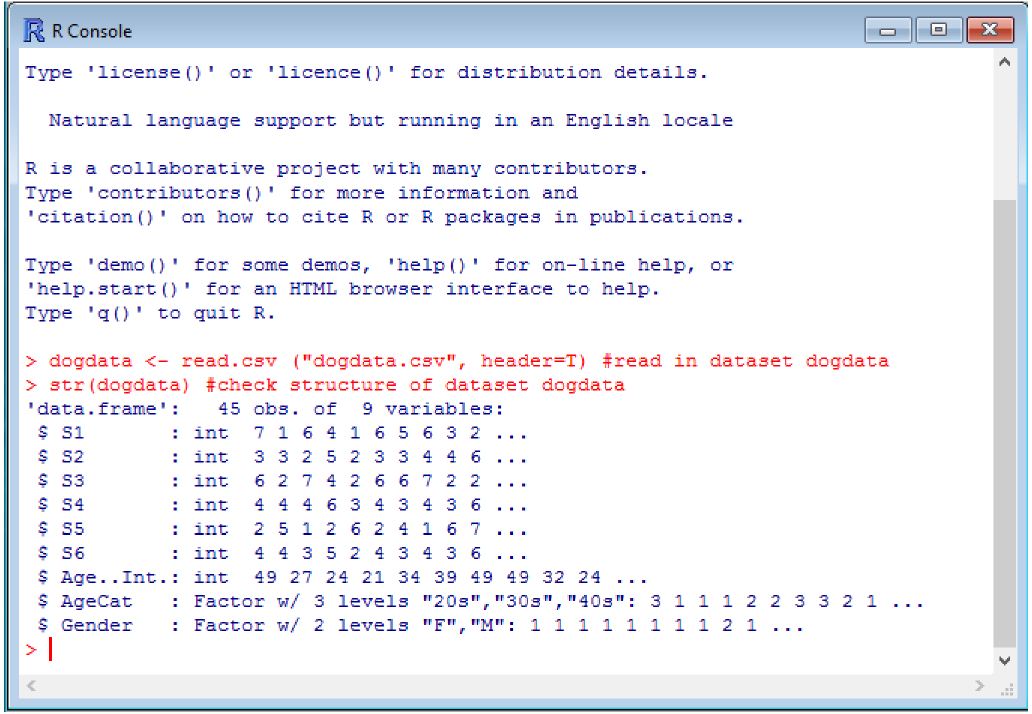


```
R version 3.2.4 Revised (2016-03-16 r70336) -- "Very Secure Dishes"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> dogdata <- read.csv ("dogdata.csv", header=T) #read in dataset dogdata
> |
```

Cluster-Based Segmentation Example: Confirm Reading Data

Topic	Discussion
Confirm Read	Ensure data was read in correctly

Confirm data was read in properly by asking R to tell you structure of dataset

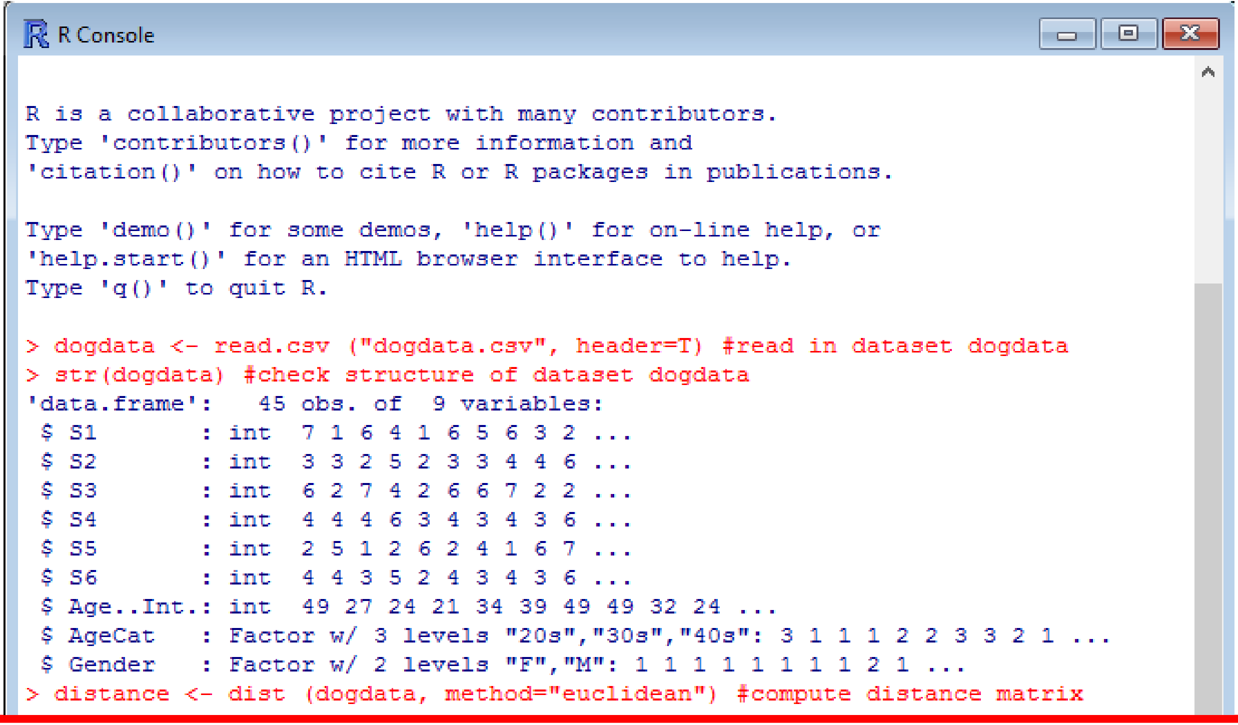


```
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> dogdata <- read.csv ("dogdata.csv", header=T) #read in dataset dogdata
> str(dogdata) #check structure of dataset dogdata
'data.frame':   45 obs. of  9 variables:
 $ S1       : int  7 1 6 4 1 6 5 6 3 2 ...
 $ S2       : int  3 3 2 5 2 3 3 4 4 6 ...
 $ S3       : int  6 2 7 4 2 6 6 7 2 2 ...
 $ S4       : int  4 4 4 6 3 4 3 4 3 6 ...
 $ S5       : int  2 5 1 2 6 2 4 1 6 7 ...
 $ S6       : int  4 4 3 5 2 4 3 4 3 6 ...
 $ Age..Int.: int  49 27 24 21 34 39 49 49 32 24 ...
 $ AgeCat   : Factor w/ 3 levels "20s","30s","40s": 3 1 1 1 2 2 3 3 2 1 ...
 $ Gender   : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 2 1 ...
> 
```

Segmentation Example: Distance Matrix for Wards

Topic	Discussion
Distance Matrix	distance <- dist (dogdata, method = “euclidean”)

First step of Wards:
Ask R to compute
the distances between
points in the dataset

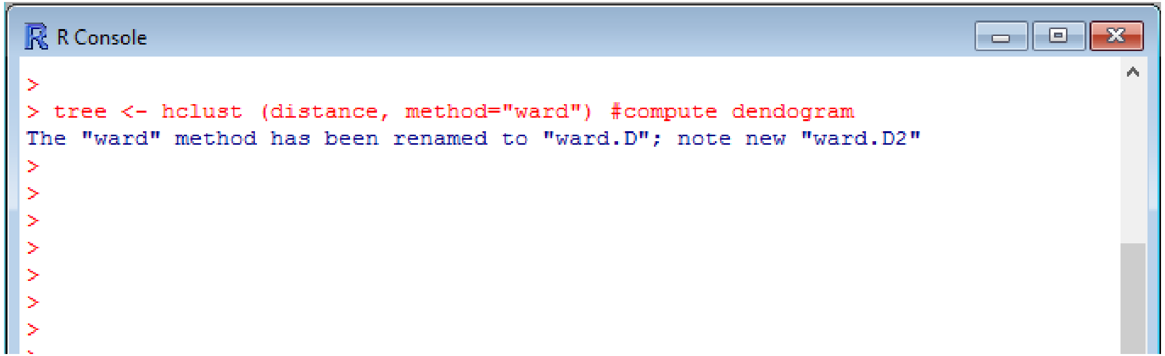


```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> dogdata <- read.csv ("dogdata.csv", header=T) #read in dataset dogdata
> str(dogdata) #check structure of dataset dogdata
'data.frame':   45 obs. of  9 variables:
 $ S1       : int  7 1 6 4 1 6 5 6 3 2 ...
 $ S2       : int  3 3 2 5 2 3 3 4 4 6 ...
 $ S3       : int  6 2 7 4 2 6 6 7 2 2 ...
 $ S4       : int  4 4 4 6 3 4 3 4 3 6 ...
 $ S5       : int  2 5 1 2 6 2 4 1 6 7 ...
 $ S6       : int  4 4 3 5 2 4 3 4 3 6 ...
 $ Age..Int.: int  49 27 24 21 34 39 49 49 32 24 ...
 $ AgeCat   : Factor w/ 3 levels "20s","30s","40s": 3 1 1 1 2 2 3 3 2 1 ...
 $ Gender   : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 2 1 ...
> distance <- dist (dogdata, method="euclidean") #compute distance matrix
```

Segmentation Example: Clusters for Wards

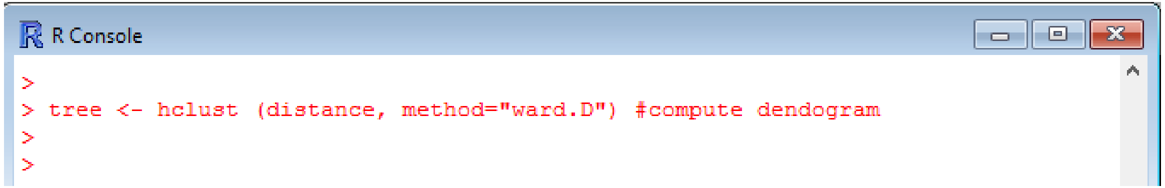
Topic	Discussion
Clusters	tree <- hclust (distance, method = “ward”)

Second step of Wards: Ask R to compute the hierarchical clusters (hclust), based on the distancematrix found in the previous step



```
>
> tree <- hclust (distance, method="ward") #compute dendogram
The "ward" method has been renamed to "ward.D"; note new "ward.D2"
>
>
>
>
>
>
>
```

R is open source code; Algorithms will change from time to time, such as “ward” changing to “ward.D”

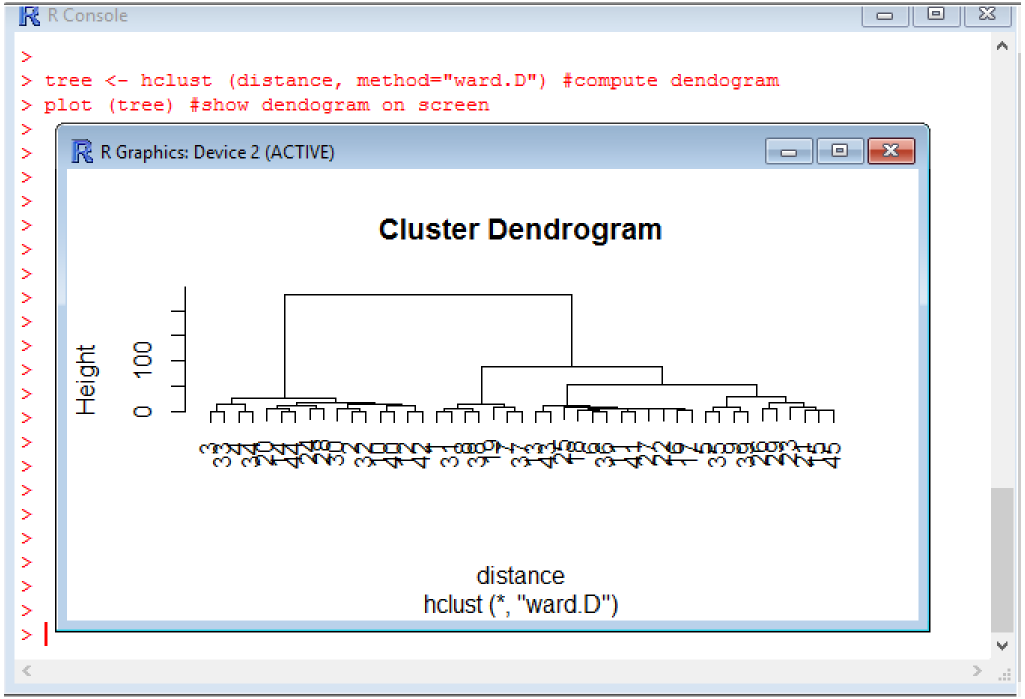


```
>
> tree <- hclust (distance, method="ward.D") #compute dendogram
>
>
```

Segmentation Example: Dendograms for Wards

Topic	Discussion
Dendograms	plot (tree)

Third step of Wards: Plot the “tree” dataset, which contains the cluster information

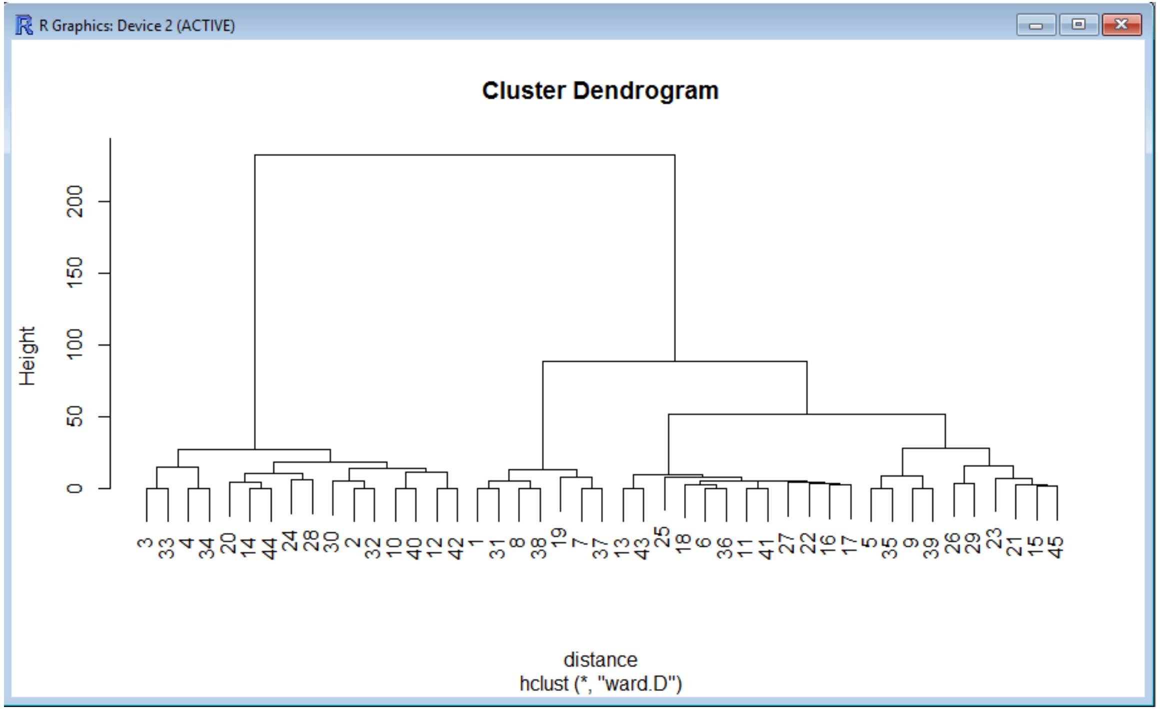


```
>
> tree <- hclust (distance, method="ward.D") #compute dendogram
> plot (tree) #show dendogram on screen
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
```

Segmentation Example: Dendograms for Wards

Topic	Discussion
Dendograms	plot (tree)

Third step of Wards: Plot the “tree” dataset, which contains the cluster information



Segmentation Example: Interpret Dendograms

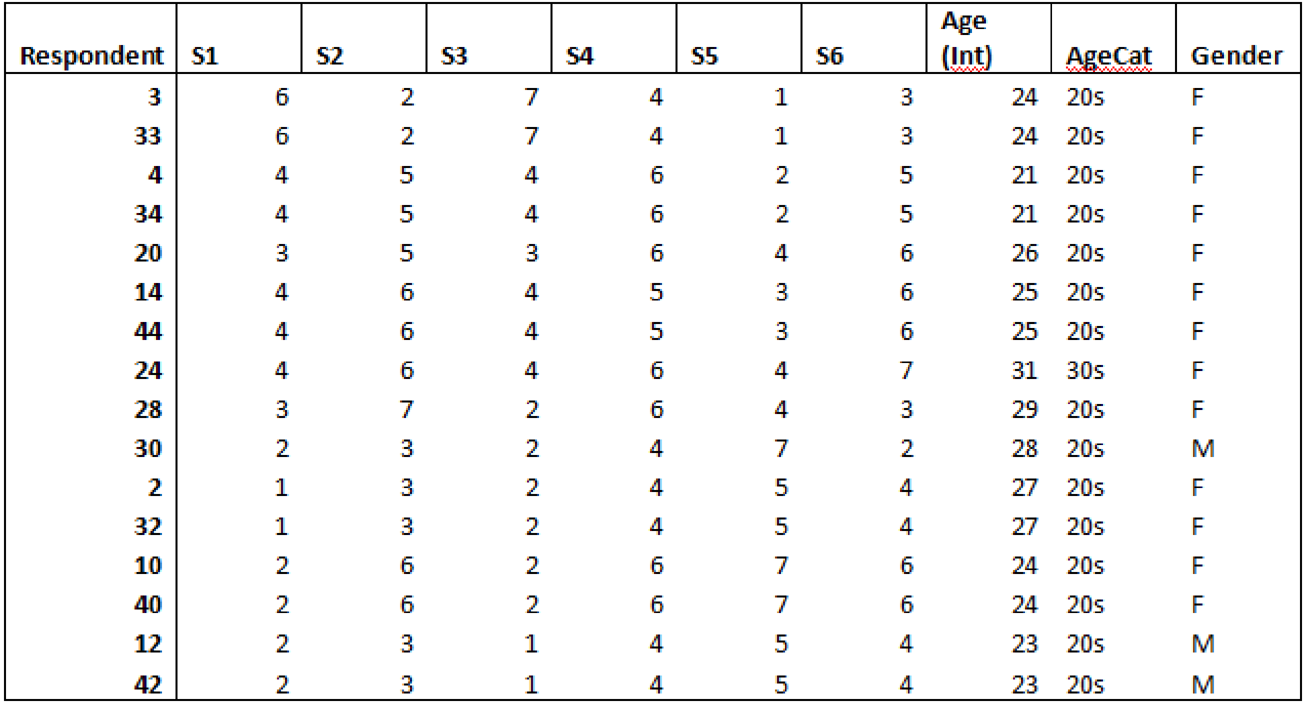
Topic	Discussion
Groupings	Data from respondents 3 and 33 are the same Wards plots the responses from “3” and “33” near each other Marketing to one would be like marketing to the other

Resp.	S1	S2	S3	S4	S5	S6	Age	AgeCat	Gender
3	6	2	7	4	1	3	24	20s	F
33	6	2	7	4	1	3	24	20s	F

Segmentation Example: Membership in Clusters

Topic	Discussion
Membership	Identify membership in each of the 3 clusters

Respondents (membership) in group 1 (cluster on left); 16 respondents total

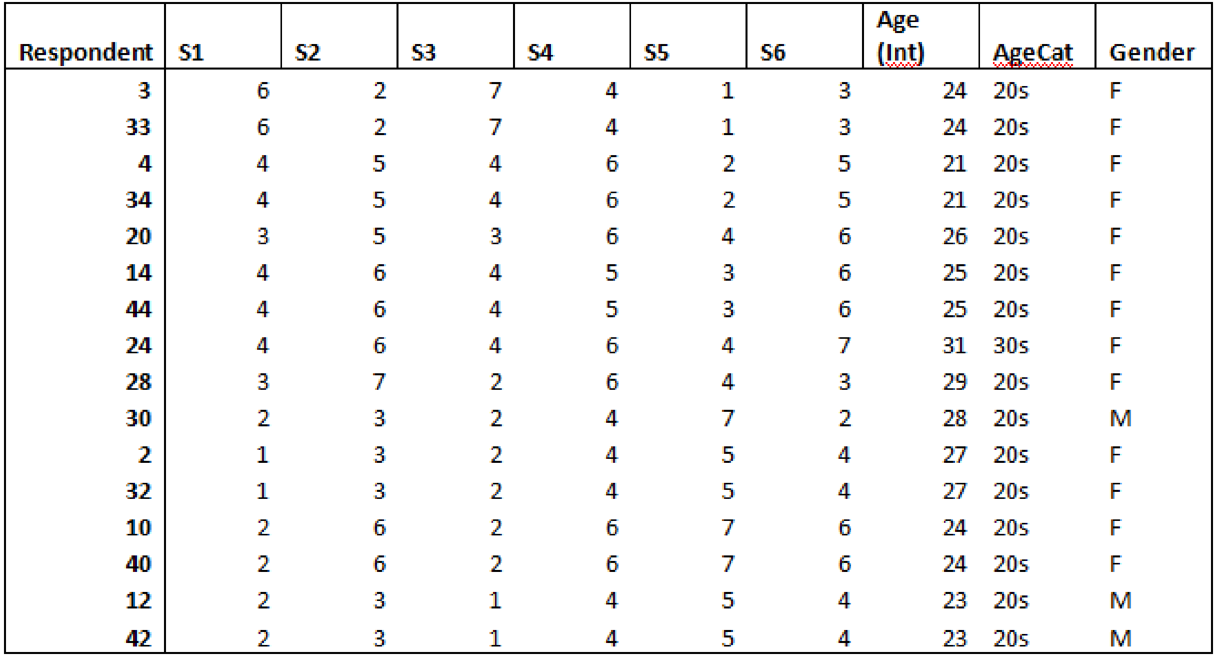


Respondent	S1	S2	S3	S4	S5	S6	Age (Int)	AgeCat	Gender
3	6	2	7	4	1	3	24	20s	F
33	6	2	7	4	1	3	24	20s	F
4	4	5	4	6	2	5	21	20s	F
34	4	5	4	6	2	5	21	20s	F
20	3	5	3	6	4	6	26	20s	F
14	4	6	4	5	3	6	25	20s	F
44	4	6	4	5	3	6	25	20s	F
24	4	6	4	6	4	7	31	30s	F
28	3	7	2	6	4	3	29	20s	F
30	2	3	2	4	7	2	28	20s	M
2	1	3	2	4	5	4	27	20s	F
32	1	3	2	4	5	4	27	20s	F
10	2	6	2	6	7	6	24	20s	F
40	2	6	2	6	7	6	24	20s	F
12	2	3	1	4	5	4	23	20s	M
42	2	3	1	4	5	4	23	20s	M

Segmentation Example: Cluster Mean: Group 1

Topic	Discussion
Means	Calculate the means (averages) for each of the 6 statements

Calculate the means (averages) for
S1, S2, S3, S4, S5, S6;
Add up and divide by 16

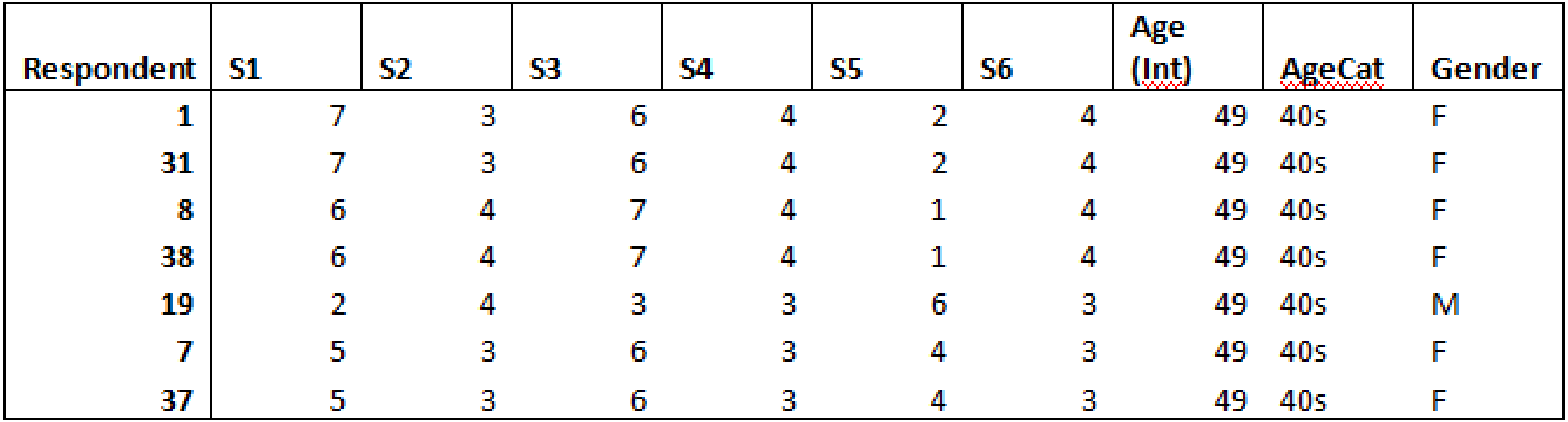


Respondent	S1	S2	S3	S4	S5	S6	Age (Int)	AgeCat	Gender
3	6	2	7	4	1	3	24	20s	F
33	6	2	7	4	1	3	24	20s	F
4	4	5	4	6	2	5	21	20s	F
34	4	5	4	6	2	5	21	20s	F
20	3	5	3	6	4	6	26	20s	F
14	4	6	4	5	3	6	25	20s	F
44	4	6	4	5	3	6	25	20s	F
24	4	6	4	6	4	7	31	30s	F
28	3	7	2	6	4	3	29	20s	F
30	2	3	2	4	7	2	28	20s	M
2	1	3	2	4	5	4	27	20s	F
32	1	3	2	4	5	4	27	20s	F
10	2	6	2	6	7	6	24	20s	F
40	2	6	2	6	7	6	24	20s	F
12	2	3	1	4	5	4	23	20s	M
42	2	3	1	4	5	4	23	20s	M

Means (Averages) → 3.13 4.44 3.19 5.00 4.06 4.63

Segmentation Example: Cluster Mean: Groups 2 & 3

Topic	Discussion
Means	Calculate the means (averages) for each of the 6 statements



Respondent	S1	S2	S3	S4	S5	S6	Age (Int)	AgeCat	Gender
1	7	3	6	4	2	4	49	40s	F
31	7	3	6	4	2	4	49	40s	F
8	6	4	7	4	1	4	49	40s	F
38	6	4	7	4	1	4	49	40s	F
19	2	4	3	3	6	3	49	40s	M
7	5	3	6	3	4	3	49	40s	F
37	5	3	6	3	4	3	49	40s	F

Means (Grp. 2)→	5.43	3.43	5.86	3.57	2.86	3.57
Means (Grp. 3)→	4.14	3.41	4.32	3.55	3.32	3.82

Segmentation Example: Cluster Mean: Summary

Topic	Discussion
Summary	Prepare table with means scores of each group

Group	S1	S2	S3	S4	S5	S6
1	3.13	4.44	3.19	5.00	4.06	4.63
2	5.43	3.43	5.86	3.57	2.86	3.57
3	4.14	3.41	4.32	3.55	3.32	3.82

Segmentation Example: Cluster Mean: Summary

Topic	Discussion
Summary	Prepare table with means scores of each group

Group	S1	S2	S3	S4	S5	S6
1	3.13	**4.44**	3.19	**5.00**	4.06	**4.63**
2	**5.43**	3.43	**5.86**	3.57	**2.86**	3.57
3	4.14	3.41	4.32	3.55	3.32	3.82

S1: It is important for me to buy dog food that prevents canine cavities
S2: I like dog food that gives my dog a shiny coat
S3: Dog food should strengthen gums
S4: Dog food should make my dog's breath fresher
S5: It is not a priority for me that dog food prevent tooth decay or cavities (reverse coded)
S6: When I buy dog food, I look for food that gives my dog shiny teeth

Cluster-Based Segmentation Example: Cluster Interpretation

Topic	Discussion
Interpretation	Establish the meaning for each group

Group	Description
1	“Beauty” segment: Buys dog food for the way it makes their dog beautiful
2	“Healthy” segment: Buys dog food for the health benefits the food provides
3	“Don’t Care” segment: No particular interest in how food helps dogs

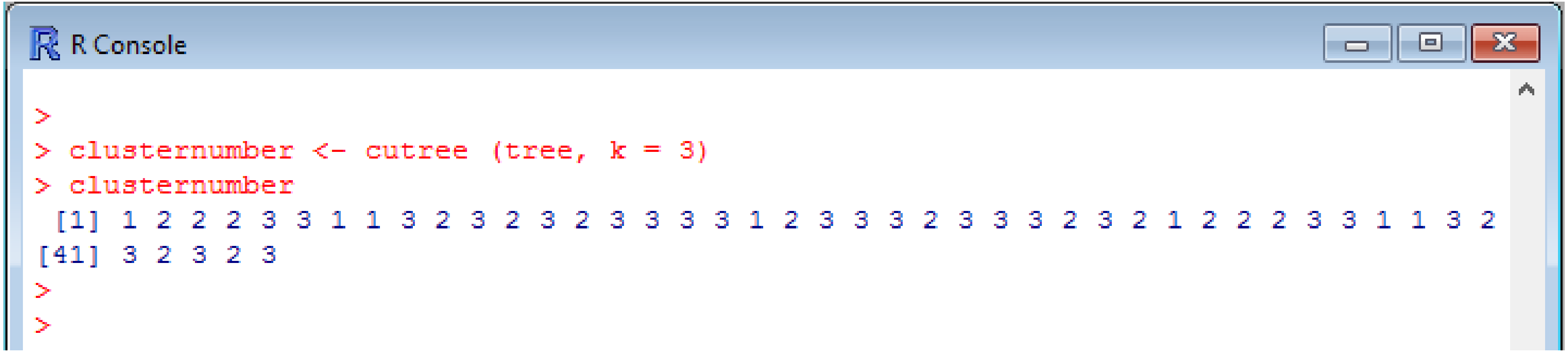
Cluster-Based Segmentation Example: Market Comparison

Topic	Discussion
Research	International Journal of Consumer Studies (Dec. 2014) *
Segments	“Strongly Attached Dog Owners”; “Price is no object” - Beauty emphasis - Healthy emphasis “Basic Dog Owner”; “Meet dogs’ basic needs”
Agrees	Research appears to agree well with our analysis

* Boya, Dotson, and Hyatt. “A Comparison of Dog Food Choice Criteria Across Dog Owner Segments: An Exploratory Study.” International Journal of Consumer Studies. December 2014. Pages 74-82. http://onlinelibrary.wiley.com/doi/10.1111/ijcs.12145/pdf

Market Segmentation Example: Advanced R

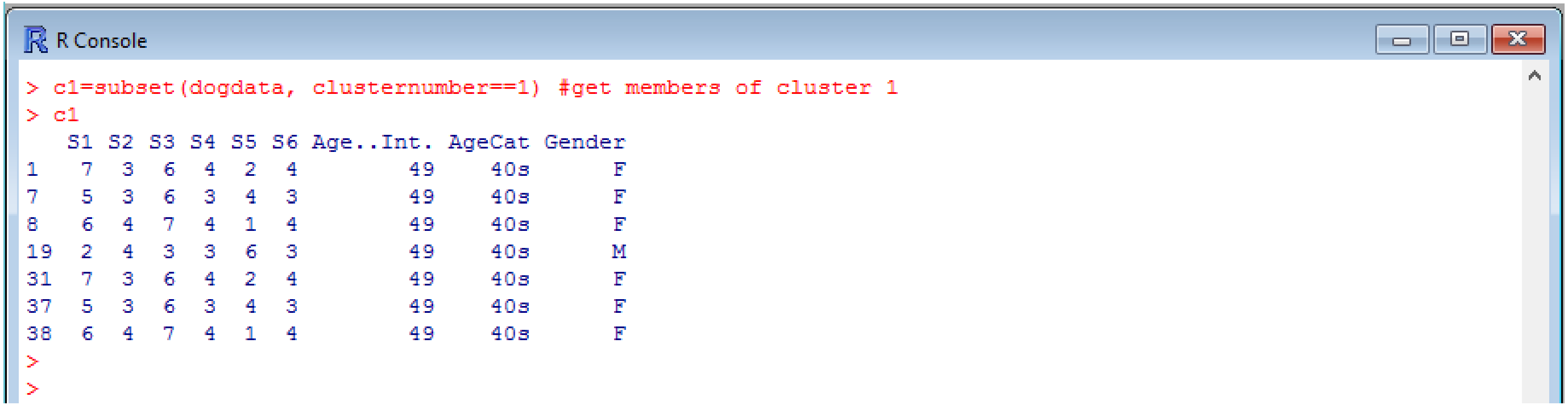
Topic	Discussion
Groups	cutree: Cut the dendogram tree into k segments/ clusters clusternumber <- cutree (tree, k = 3)
Members	Lists cluster number of each respondent Example: Respondent 1: 1; Respondent 2: 2; Resp 3: 2



```
>
> clusternumber <- cutree (tree, k = 3)
> clusternumber
 [1] 1 2 2 2 3 3 1 1 3 2 3 2 3 2 3 3 3 3 1 2 3 3 3 2 3 3 3 2 3 2 1 2 2 2 3 3 1 1 3 2
[41] 3 2 3 2 3
>
>
```

Market Segmentation Example: Interpret Dendograms

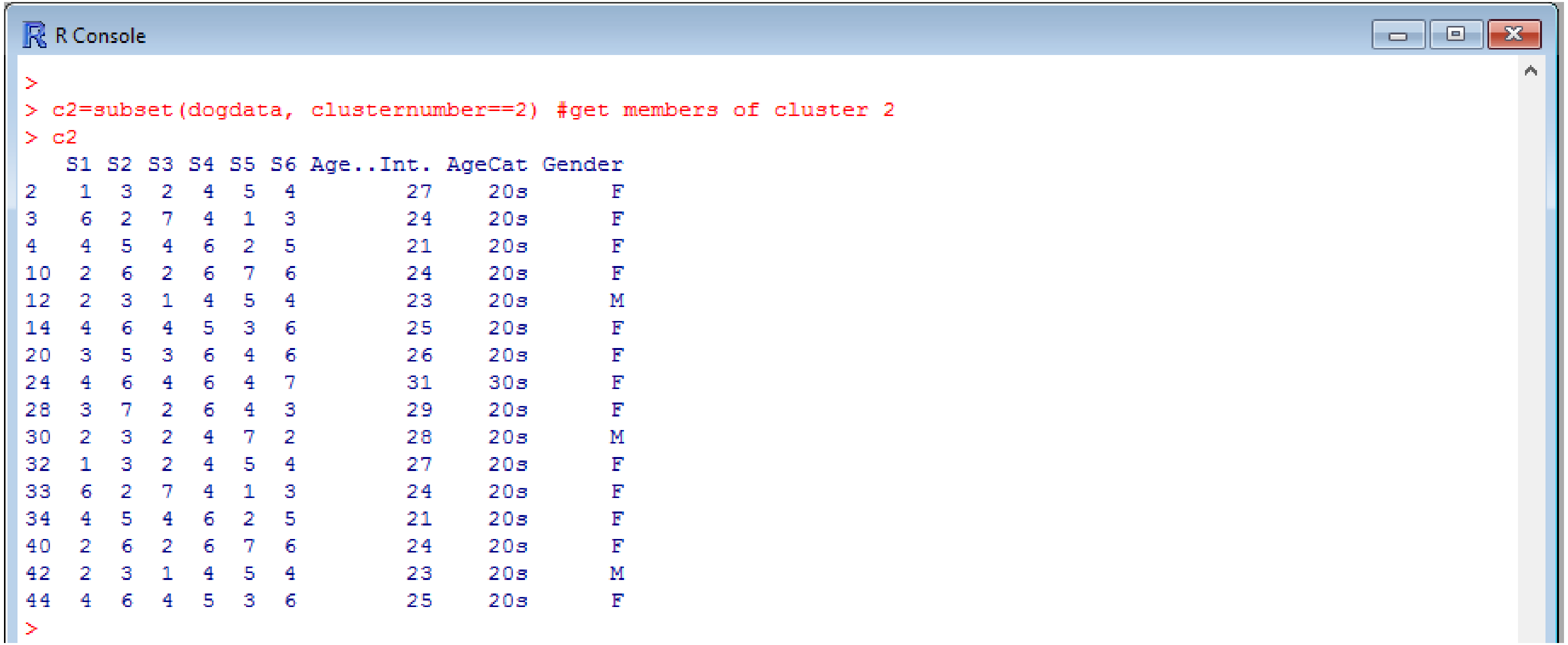
Topic	Discussion
Clusters	Subset: Get clusters of data based on clusternumber value c1 = subset (dogdata, clusternumber = 1) #cluster 1



```
> c1=subset(dogdata, clusternumber==1) #get members of cluster 1
> c1
   S1 S2 S3 S4 S5 S6 Age..Int. AgeCat Gender
1   7  3  6  4  2  4        49    40s      F
7   5  3  6  3  4  3        49    40s      F
8   6  4  7  4  1  4        49    40s      F
19  2  4  3  3  6  3        49    40s      M
31  7  3  6  4  2  4        49    40s      F
37  5  3  6  3  4  3        49    40s      F
38  6  4  7  4  1  4        49    40s      F
>
>
```

Market Segmentation Example: Interpret Dendograms

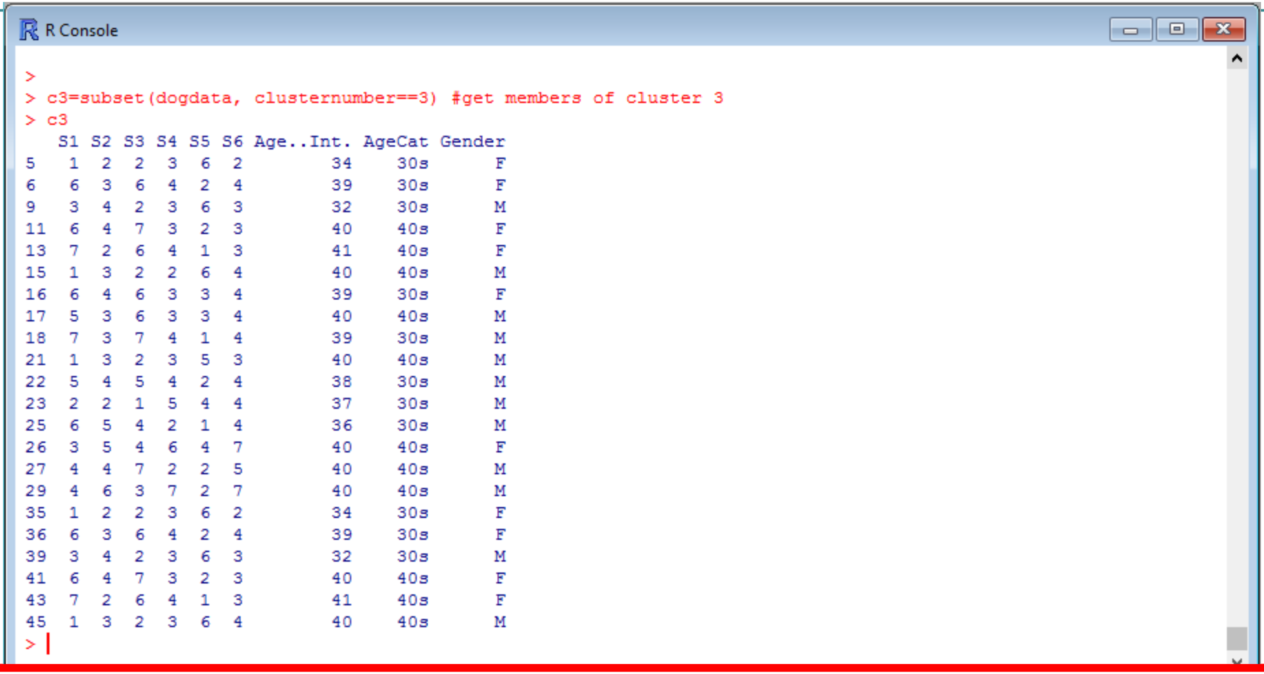
Topic	Discussion
Clusters	Subset: Get clusters of data based on clusternumber value c2 = subset (dogdata, clusternumber = 2) #cluster 2



```
>
> c2=subset(dogdata, clusternumber==2) #get members of cluster 2
> c2
   S1 S2 S3 S4 S5 S6 Age..Int. AgeCat Gender
2   1  3  2  4  5  4        27    20s      F
3   6  2  7  4  1  3        24    20s      F
4   4  5  4  6  2  5        21    20s      F
10  2  6  2  6  7  6        24    20s      F
12  2  3  1  4  5  4        23    20s      M
14  4  6  4  5  3  6        25    20s      F
20  3  5  3  6  4  6        26    20s      F
24  4  6  4  6  4  7        31    30s      F
28  3  7  2  6  4  3        29    20s      F
30  2  3  2  4  7  2        28    20s      M
32  1  3  2  4  5  4        27    20s      F
33  6  2  7  4  1  3        24    20s      F
34  4  5  4  6  2  5        21    20s      F
40  2  6  2  6  7  6        24    20s      F
42  2  3  1  4  5  4        23    20s      M
44  4  6  4  5  3  6        25    20s      F
>
```

Market Segmentation Example: Interpret Dendograms

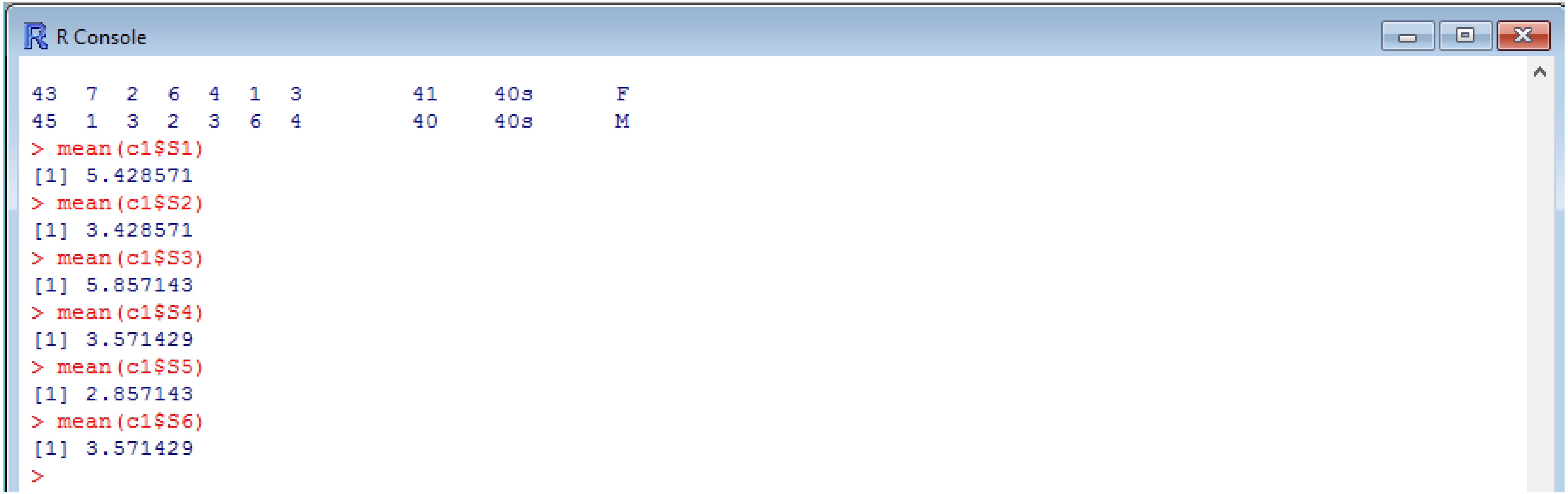
Topic	Discussion
Clusters	Subset: Get clusters of data based on clusternumber value c3 = subset (dogdata, clusternumber = 3) #cluster 3



```
>
> c3=subset(dogdata, clusternumber==3) #get members of cluster 3
> c3
   S1 S2 S3 S4 S5 S6 Age..Int. AgeCat Gender
5   1  2  2  3  6  2        34    30s      F
6   6  3  6  4  2  4        39    30s      F
9   3  4  2  3  6  3        32    30s      M
11  6  4  7  3  2  3        40    40s      F
13  7  2  6  4  1  3        41    40s      F
15  1  3  2  2  6  4        40    40s      M
16  6  4  6  3  3  4        39    30s      F
17  5  3  6  3  3  4        40    40s      M
18  7  3  7  4  1  4        39    30s      M
21  1  3  2  3  5  3        40    40s      M
22  5  4  5  4  2  4        38    30s      M
23  2  2  1  5  4  4        37    30s      M
25  6  5  4  2  1  4        36    30s      M
26  3  5  4  6  4  7        40    40s      F
27  4  4  7  2  2  5        40    40s      M
29  4  6  3  7  2  7        40    40s      M
35  1  2  2  3  6  2        34    30s      F
36  6  3  6  4  2  4        39    30s      F
39  3  4  2  3  6  3        32    30s      M
41  6  4  7  3  2  3        40    40s      F
43  7  2  6  4  1  3        41    40s      F
45  1  3  2  3  6  4        40    40s      M
> |
```

Market Segmentation Example: Interpret Dendograms

Topic	Discussion
Mean	Compute mean (average) for each column (S) in each cluster mean(c1$S1)

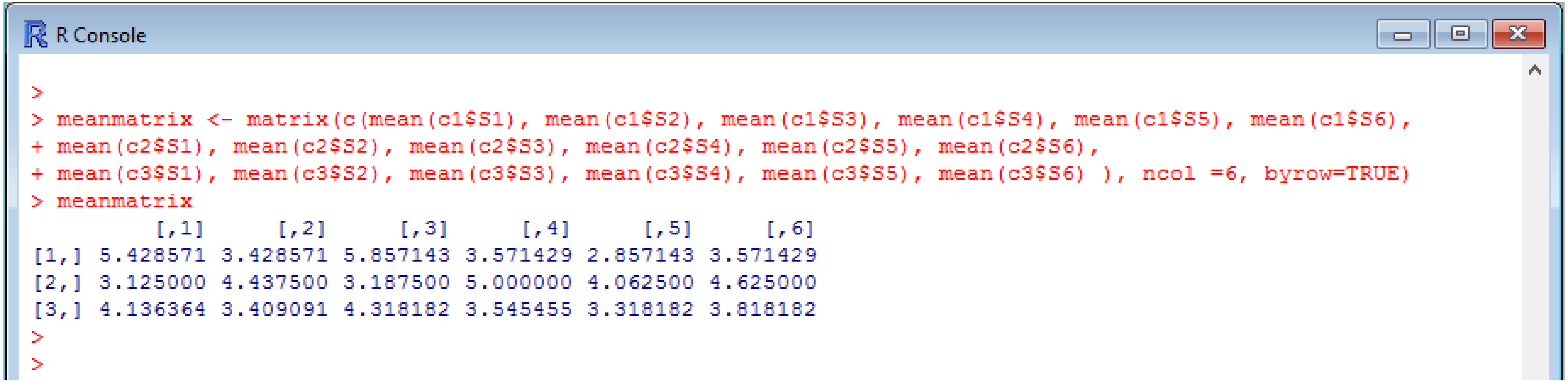


```
43   7   2   6   4   1   3          41     40s       F
45   1   3   2   3   6   4          40     40s       M
> mean(c1$S1)
[1] 5.428571
> mean(c1$S2)
[1] 3.428571
> mean(c1$S3)
[1] 5.857143
> mean(c1$S4)
[1] 3.571429
> mean(c1$S5)
[1] 2.857143
> mean(c1$S6)
[1] 3.571429
>
```

Market Segmentation Example: Interpret Dendograms

Topic	Discussion
Mean Matrix	matrix command; Build matrix of means for each cluster meanmatrix <- matrix(c(mean(c1$S1), mean(c1$S2), …

meanmatrix <- matrix(c(mean(c1$S1), mean(c1$S2), mean(c1$S3), mean(c1$S4), mean(c1$S5), mean(c1$S6),
mean(c2$S1), mean(c2$S2), mean(c2$S3), mean(c2$S4), mean(c2$S5), mean(c2$S6),
mean(c3$S1), mean(c3$S2), mean(c3$S3), mean(c3$S4), mean(c3$S5), mean(c3$S6)), ncol =6, byrow=TRUE)

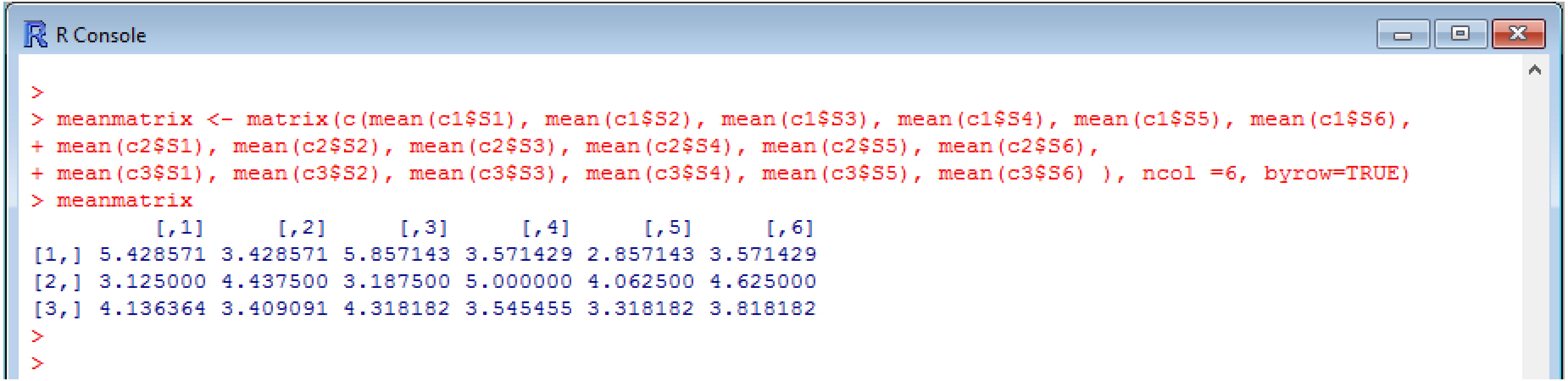


```
>
> meanmatrix <- matrix(c(mean(c1$S1), mean(c1$S2), mean(c1$S3), mean(c1$S4), mean(c1$S5), mean(c1$S6),
+ mean(c2$S1), mean(c2$S2), mean(c2$S3), mean(c2$S4), mean(c2$S5), mean(c2$S6),
+ mean(c3$S1), mean(c3$S2), mean(c3$S3), mean(c3$S4), mean(c3$S5), mean(c3$S6) ), ncol =6, byrow=TRUE)
> meanmatrix
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
[1,] 5.428571 3.428571 5.857143 3.571429 2.857143 3.571429
[2,] 3.125000 4.437500 3.187500 5.000000 4.062500 4.625000
[3,] 4.136364 3.409091 4.318182 3.545455 3.318182 3.818182
>
>
```

Market Segmentation Example: Compare Results

Note that R assigns a different group number than the number we arbitrarily assigned

Group	S1	S2	S3	S4	S5	S6
1	3.13	**4.44**	3.19	**5.00**	4.06	**4.63**
2	**5.43**	3.43	**5.86**	3.57	**2.86**	3.57
3	4.14	3.41	4.32	3.55	3.32	3.82



```
>
> meanmatrix <- matrix(c(mean(c1$S1), mean(c1$S2), mean(c1$S3), mean(c1$S4), mean(c1$S5), mean(c1$S6),
+ mean(c2$S1), mean(c2$S2), mean(c2$S3), mean(c2$S4), mean(c2$S5), mean(c2$S6),
+ mean(c3$S1), mean(c3$S2), mean(c3$S3), mean(c3$S4), mean(c3$S5), mean(c3$S6) ), ncol =6, byrow=TRUE)
> meanmatrix
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
[1,] 5.428571 3.428571 5.857143 3.571429 2.857143 3.571429
[2,] 3.125000 4.437500 3.187500 5.000000 4.062500 4.625000
[3,] 4.136364 3.409091 4.318182 3.545455 3.318182 3.818182
>
>
```

Introduction to
Data Science and Analytics

Stephan Sorger

www.StephanSorger.com

Unit 8. R Segmentation

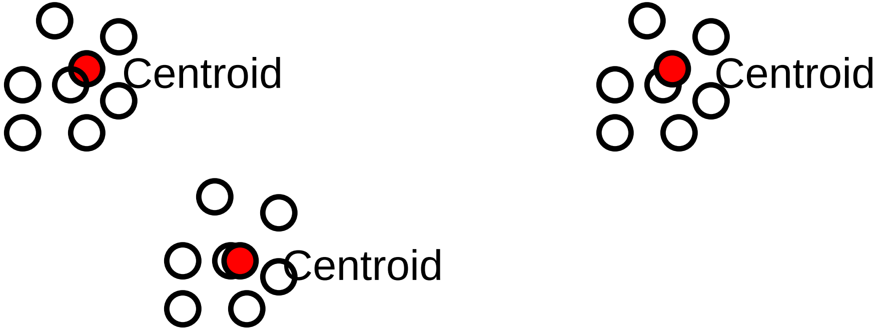
Lecture: K-Means Cluster Analysis

Disclaimer:

- All images such as logos, photos, etc. used in this presentation are the property of their respective copyright owners and are used here for educational purposes only
- Some material adapted from: Sorger, “Marketing Analytics: Strategic Models and Metrics”

Cluster-Based Segmentation: K-Means

Topic	Discussion
K-Means	Forms groups based on “distance” from “centroid”
Algorithm	Specify K, the number of final clusters to expect Execute K-Means algorithm Identify clusters; Change K as necessary
R	Standard function in R; No package install; Complex



K = 3; 3 Clusters

Cluster-Based Segmentation: K-Means

K-Means in R

Syntax:

$$\text{Kmeans } (\underbrace{\text{x , centers}}_{\text{Required}}, \underbrace{\text{iter.max, nstart, algorithm, trace}}_{\text{Optional}})$$

where

x = numeric matrix of data (your dataset)

centers = number of clusters (k)

iter.max = maximum number of iterations allowed (prevent computer running away); default=10

nstart = number of random sets to be chosen (default nstart=1)

algorithm = choice of different algorithms. Hartigan and Wong algorithm used by default
For more information, see help file

trace = integer number used to trace information on the progress of the algorithm
(to diagnose errors, or simply keep tabs on the process); default trace=FALSE

Kmeans Package Help File:
https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html

Cluster-Based Segmentation: K-Means

Sample K-Means Session

```
> #enable graphics
> require(graphics)
> #build 2-dimensional matrix for example purposes
> x <- rbind(matrix(rnorm(100, sd=0.3), ncol=2), matrix(rnorm(100, mean=1, sd=0.3), ncol=2))
> #name the columns of the matrix
> colnames(x) <- c("x", "y")
> (c1 <- kmeans(x,2))
> plot (x, col = c1$cluster)
```

Comments denoted with #hashtag

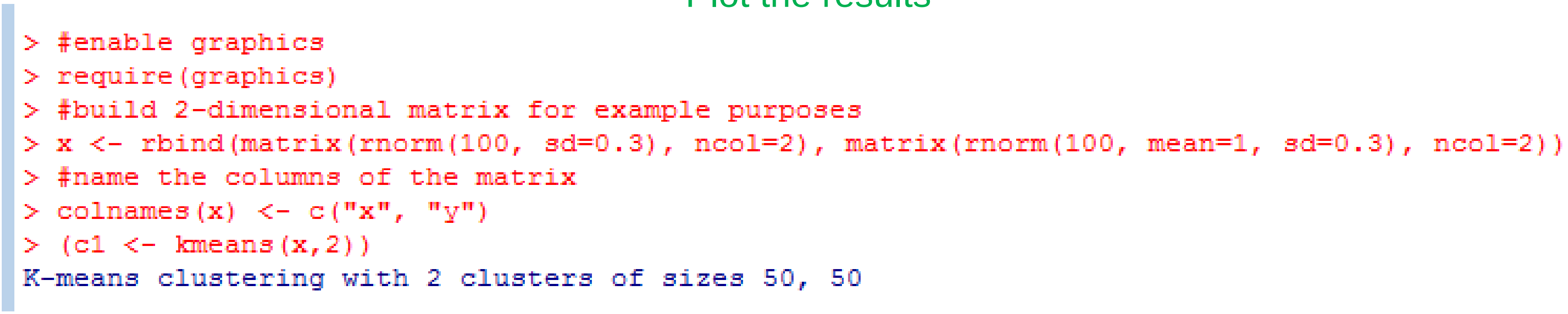
Invoke graphics capabilities

Arbitrary 2 x 2 matrix for example

Name columns so we can interpret plot

Invoke kmeans function

Plot the results

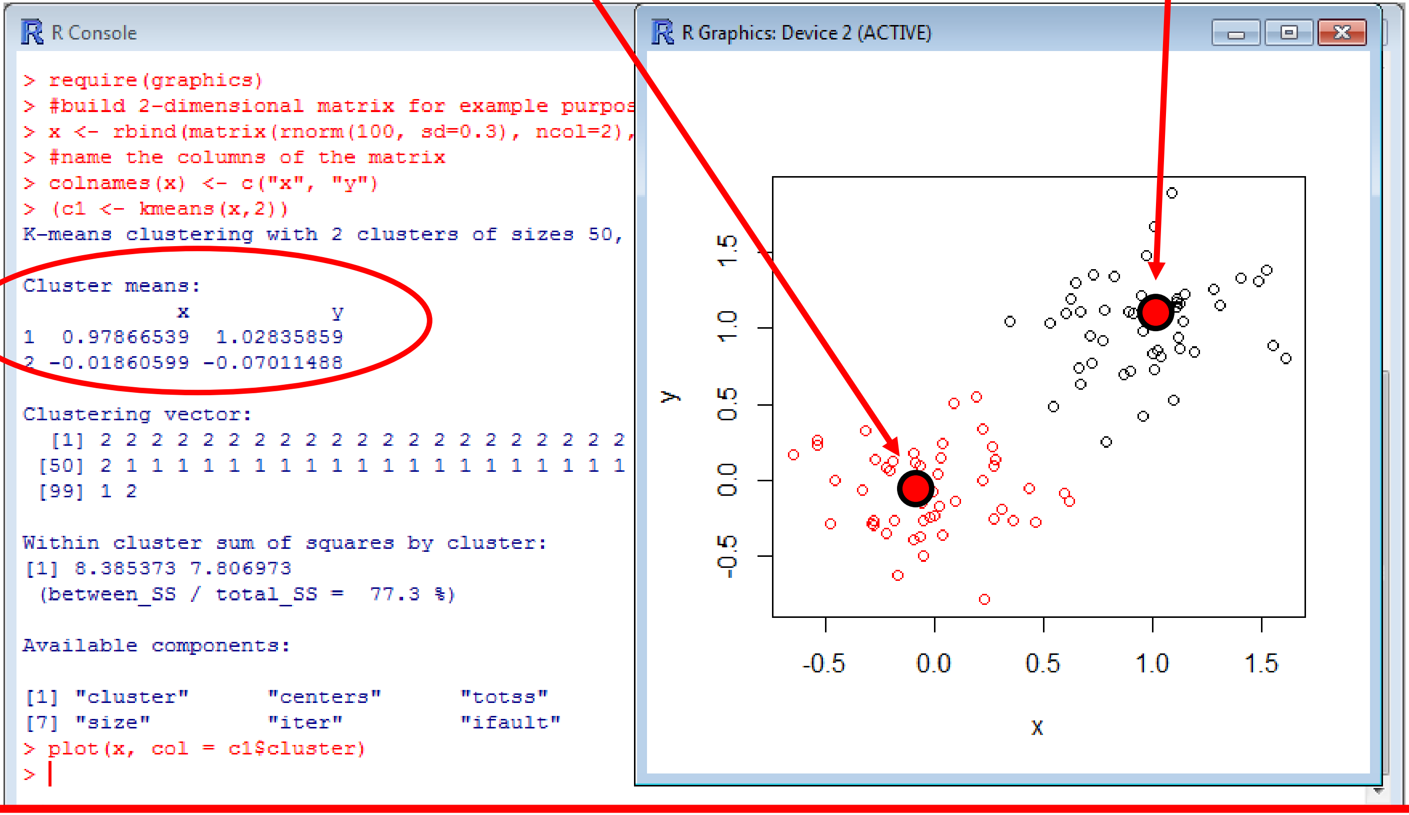


```
> #enable graphics
> require(graphics)
> #build 2-dimensional matrix for example purposes
> x <- rbind(matrix(rnorm(100, sd=0.3), ncol=2), matrix(rnorm(100, mean=1, sd=0.3), ncol=2))
> #name the columns of the matrix
> colnames(x) <- c("x", "y")
> (c1 <- kmeans(x,2))
K-means clustering with 2 clusters of sizes 50, 50
```

Cluster-Based Segmentation: K-Means

Cluster 1: x,y =(0.978, 1.028)

Cluster 2: x,y =(-0.0186, -0.070))



```
> require(graphics)
> #build 2-dimensional matrix for example purpos
> x <- rbind(matrix(rnorm(100, sd=0.3), ncol=2),
> #name the columns of the matrix
> colnames(x) <- c("x", "y")
> (c1 <- kmeans(x,2))
K-means clustering with 2 clusters of sizes 50,

Cluster means:
           x           y
1  0.97866539  1.02835859
2 -0.01860599 -0.07011488

Clustering vector:
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [50] 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [99] 1 2

Within cluster sum of squares by cluster:
[1] 8.385373 7.806973
 (between_SS / total_SS =  77.3 %)

Available components:

[1] "cluster"      "centers"      "totss"
[7] "size"         "iter"         "ifault"
> plot(x, col = c1$cluster)
> |
```

Outline/ Learning Objectives

Topic	Description
Introduction	Overview of market segmentation, targeting, and positioning
A Priori	Comparison of A Priori and Post Hoc approaches
Techniques	Overview of different segmentation techniques
Naïve Bayes	Brief review of Naïve Bayes classification approach
Clusters	Discussion of cluster analysis for segmentation
R	Segmentation using R: K-means; Ward’s methods