GSBA 524: Term 3

Professor Jacob Bien

Unsupervised learning

Unsupervised learning

- Just have X (no Y)
- Want to uncover insights without a specific prediction goal in mind

Example

- Data from Grosse Pointe Associates (consulting firm in Michigan) - consumer panel surveys

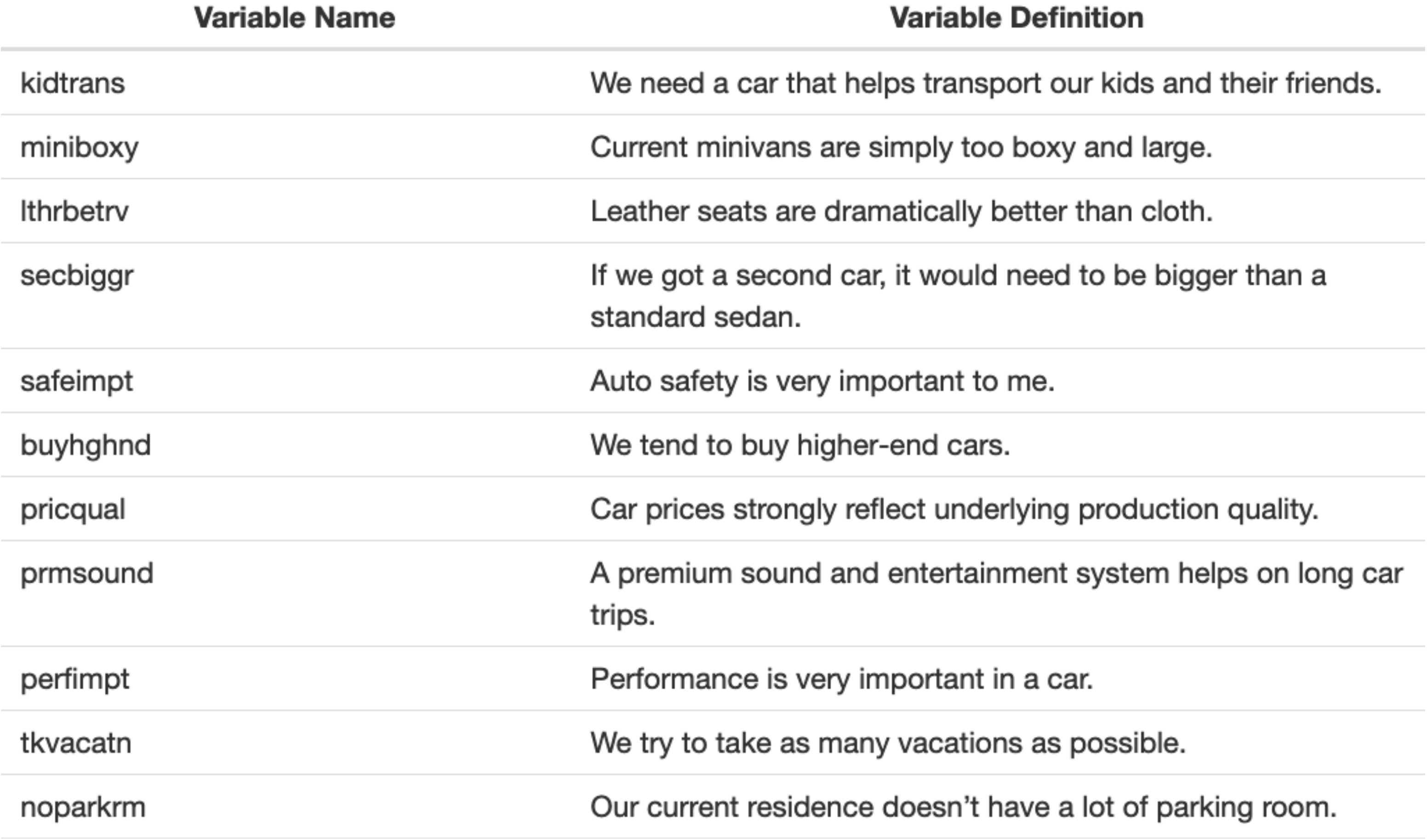
- Noticed increasing interest in minivans among affluent couples with children; yet reluctance due to their being too big.

- Wanted to investigate this trend… opportunity for a new category? A “microvan”?

- *Data set:* Random sample of $n = 400$ respondents answering 38 questions

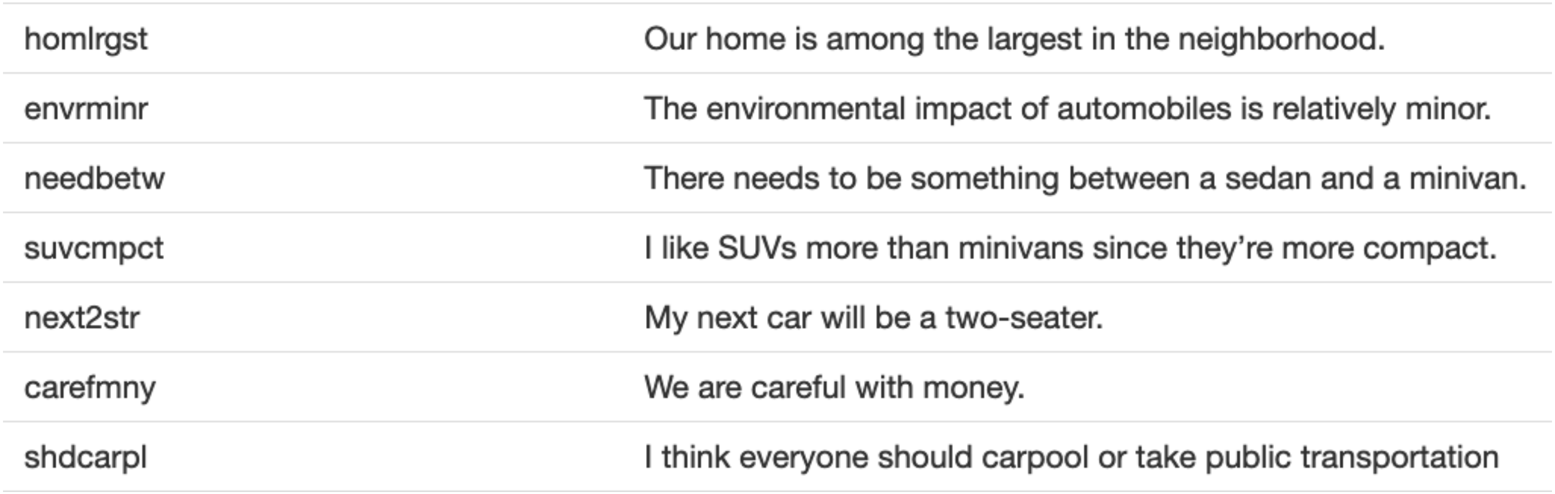
(From Feinberg, Kinnear, and Taylor, “Modern Marketing Research: Concepts, Methods and Cases”)

Questions asked



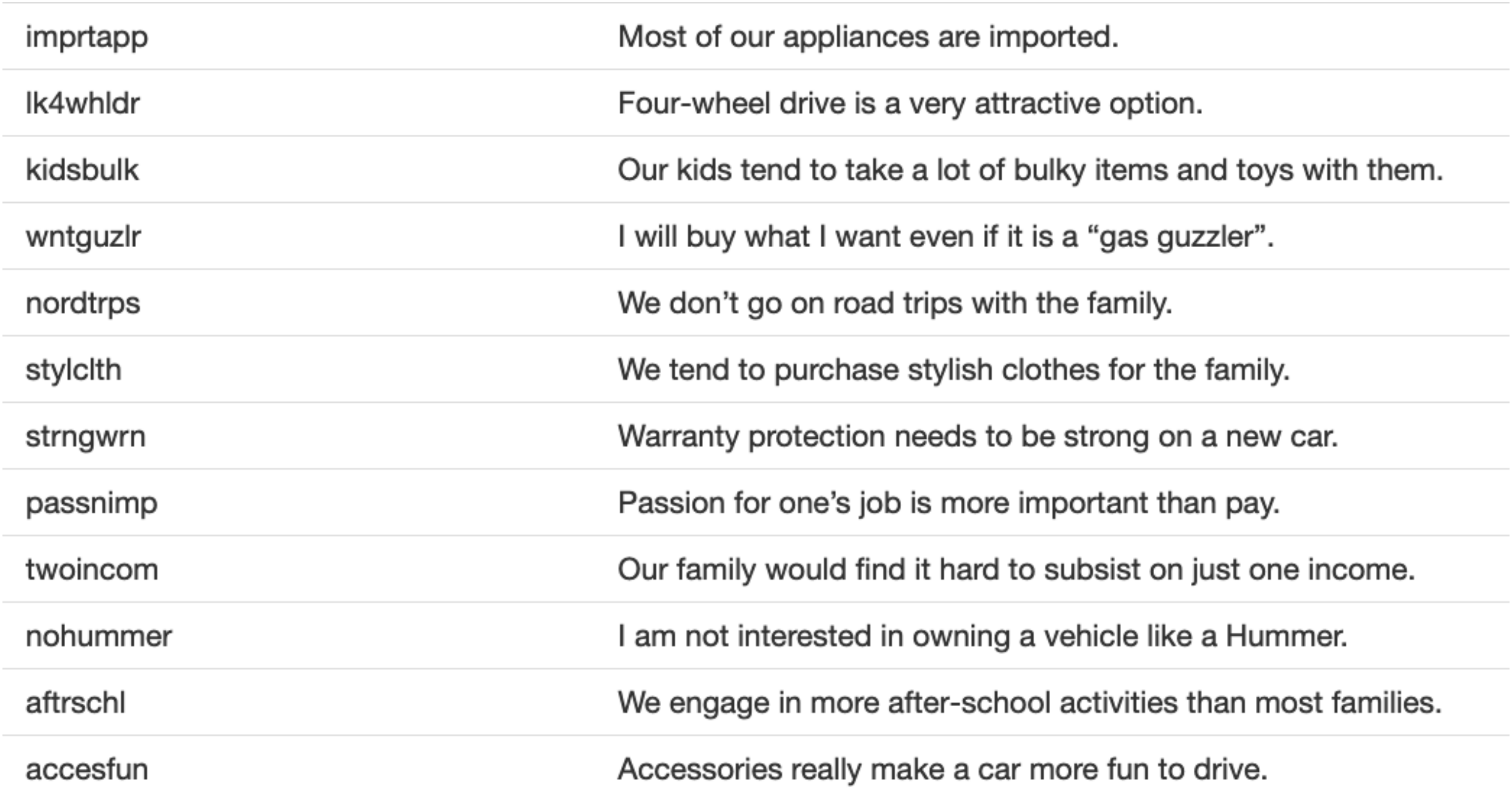
Variable Name	Variable Definition
kidtrans	We need a car that helps transport our kids and their friends.
miniboxy	Current minivans are simply too boxy and large.
lthrbetrv	Leather seats are dramatically better than cloth.
secbiggr	If we got a second car, it would need to be bigger than a standard sedan.
safeimpt	Auto safety is very important to me.
buyhghnd	We tend to buy higher-end cars.
pricqual	Car prices strongly reflect underlying production quality.
prmsound	A premium sound and entertainment system helps on long car trips.
perfimpt	Performance is very important in a car.
tkvacatn	We try to take as many vacations as possible.
noparkrm	Our current residence doesn’t have a lot of parking room.

Questions asked



homlrgst	Our home is among the largest in the neighborhood.
envrminr	The environmental impact of automobiles is relatively minor.
needbetw	There needs to be something between a sedan and a minivan.
suvcmpct	I like SUVs more than minivans since they’re more compact.
next2str	My next car will be a two-seater.
carefmny	We are careful with money.
shdcarpl	I think everyone should carpool or take public transportation

Questions asked



imprtapp	Most of our appliances are imported.
lk4whldr	Four-wheel drive is a very attractive option.
kidsbulk	Our kids tend to take a lot of bulky items and toys with them.
wntguzlr	I will buy what I want even if it is a “gas guzzler”.
nordtrps	We don’t go on road trips with the family.
stylclth	We tend to purchase stylish clothes for the family.
strngwrn	Warranty protection needs to be strong on a new car.
passnimp	Passion for one’s job is more important than pay.
twoincom	Our family would find it hard to subsist on just one income.
nohummer	I am not interested in owning a vehicle like a Hummer.
aftrschl	We engage in more after-school activities than most families.
accesfun	Accessories really make a car more fun to drive.

(Also some demographic variables)

Objectives

- *Dimension reduction:* Can we summarize/synthesize the 38 features into a small number of underlying factors? (**Principal components analysis**)

- *Market segmentation:* Do respondents fall naturally into groups? (**Clustering**)

Principal components analysis

$$\begin{pmatrix} x_{11}, x_{12}, \ldots, x_{1p} \\ x_{21}, x_{22}, \ldots, x_{2p} \\ x_{31}, x_{32}, \ldots, x_{3p} \\ \vdots, \vdots, \ldots, \vdots \\ x_{n1}, x_{n2}, \ldots, x_{np} \end{pmatrix} \xrightarrow{\text{dimension reduction}} \begin{pmatrix} z_{11}, z_{12}, \ldots, z_{1d} \\ z_{21}, z_{22}, \ldots, z_{2d} \\ z_{31}, z_{32}, \ldots, z_{3d} \\ \vdots, \vdots, \ldots, \vdots \\ z_{n1}, z_{n2}, \ldots, z_{nd} \end{pmatrix}$$

$$\boxed{p \gg d}$$

Can we form a small number of new variables that captures most of the interesting variability in respondents?

$$\begin{pmatrix} x_{11}, x_{12}, \ldots, x_{1p} \\ x_{21}, x_{22}, \ldots, x_{2p} \\ x_{31}, x_{32}, \ldots, x_{3p} \\ \vdots\,, \vdots\,, \ldots, \vdots \\ x_{n1}, x_{n2}, \ldots, x_{np} \end{pmatrix}$$

original features $\boxed{X_1, X_2, \ldots, X_p}$

$$\begin{pmatrix} z_{11}, z_{12}, \ldots, z_{1d} \\ z_{21}, z_{22}, \ldots, z_{2d} \\ z_{31}, z_{32}, \ldots, z_{3d} \\ \vdots\,, \vdots\,, \ldots, \vdots \\ z_{n1}, z_{n2}, \ldots, z_{nd} \end{pmatrix}$$

principal components $\boxed{Z_1, Z_2, \ldots, Z_d}$

Each of the principal components is a linear combination of the original features, as in

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

First principal component

$$Z_{\boxed{1}} = \phi_{1\boxed{1}} X_1 + \phi_{2\boxed{1}} X_2 + \cdots + \phi_{p\boxed{1}} X_p$$

1 **indicates the first PC**

$$Z_1 = \phi_{\boxed{1}1} X_{\boxed{1}} + \phi_{\boxed{2}1} X_{\boxed{2}} + \cdots + \phi_{\boxed{p}1} X_{\boxed{p}}$$

indices of the original features

“Scores” and “loadings”

$$\begin{pmatrix} x_{11}, x_{12}, \ldots, x_{1p} \\ x_{21}, x_{22}, \ldots, x_{2p} \\ x_{31}, x_{32}, \ldots, x_{3p} \\ \vdots, \vdots, \ldots, \vdots \\ x_{n1}, x_{n2}, \ldots, x_{np} \end{pmatrix} \qquad \begin{pmatrix} \boxed{\begin{matrix} z_{11} \\ z_{21} \\ z_{31} \\ \vdots \\ z_{n1} \end{matrix}} & \begin{matrix} , z_{12}, \ldots, z_{1d} \\ , z_{22}, \ldots, z_{2d} \\ , z_{32}, \ldots, z_{3d} \\ , \vdots, \ldots, \vdots \\ , z_{n2}, \ldots, z_{nd} \end{matrix} \end{pmatrix}$$

“scores” of the first PC

$$z_{i1} = \boxed{\phi_{11}} x_{i1} + \boxed{\phi_{21}} x_{i2} + \ldots + \boxed{\phi_{p1}} x_{ip}$$

Computing the first PC is equivalent to finding $\boxed{\phi_{11}, \phi_{21}, \ldots, \phi_{p1}}$

“Loadings” of the first PC

Second PC

$$\begin{pmatrix} x_{11}, x_{12}, \dots, x_{1p} \\ x_{21}, x_{22}, \dots, x_{2p} \\ x_{31}, x_{32}, \dots, x_{3p} \\ \vdots \;,\; \vdots \;,\dots,\; \vdots \\ x_{n1}, x_{n2}, \dots, x_{np} \end{pmatrix}$$

$$\begin{pmatrix} z_{11}, z_{12}, \dots, z_{1d} \\ z_{21}, z_{22}, \dots, z_{2d} \\ z_{31}, z_{32}, \dots, z_{3d} \\ \vdots \;,\; \vdots \;,\dots,\; \vdots \\ z_{n1}, z_{n2}, \dots, z_{nd} \end{pmatrix}$$

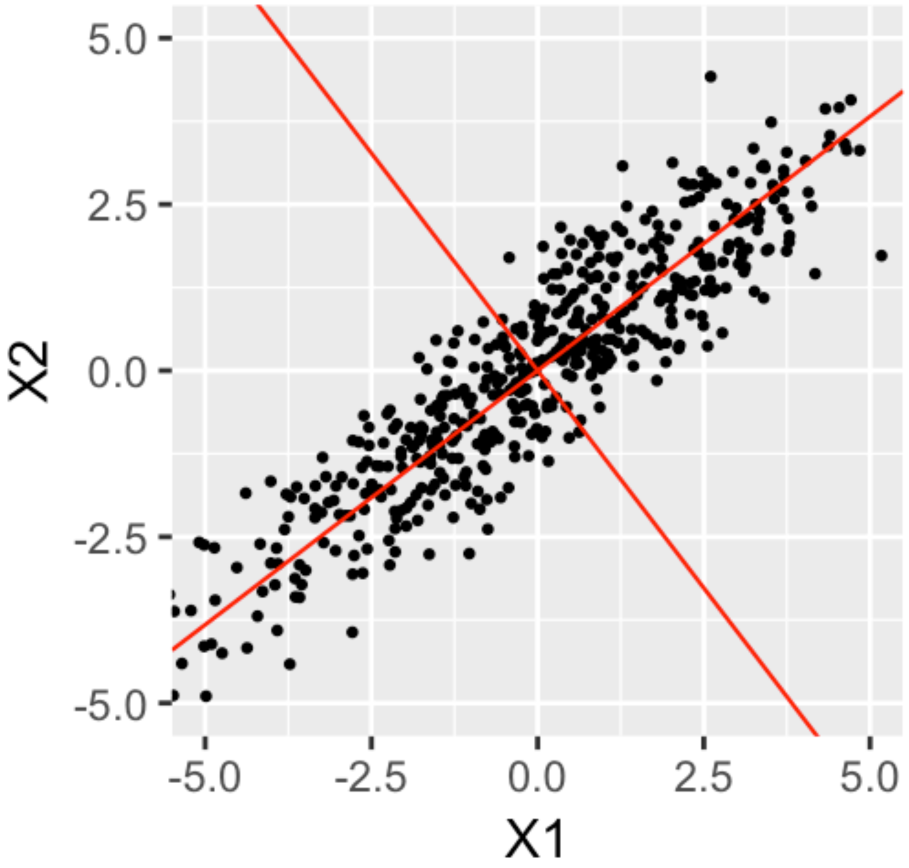
scores of the second PC

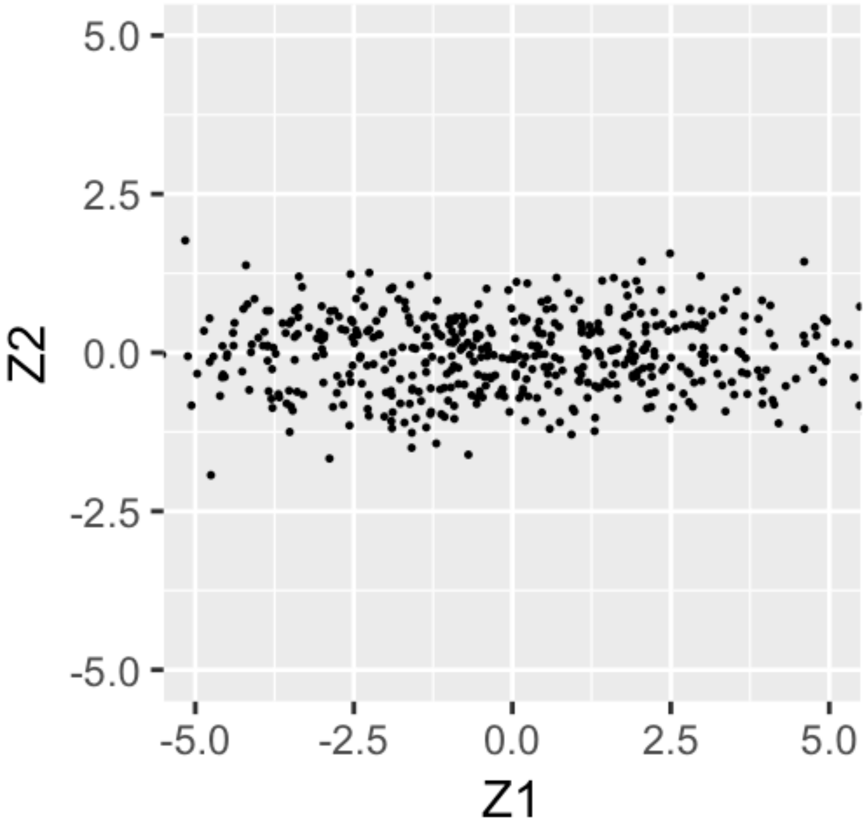
$$z_{i2} = \boxed{\phi_{12}} x_{i1} + \boxed{\phi_{22}} x_{i2} + \ldots + \boxed{\phi_{p2}} x_{ip}$$

Computing the 2nd PC is equivalent to finding $\boxed{\phi_{12}, \phi_{22}, \dots, \phi_{p2}}$

Loadings of the 2nd PC

Understanding PCs

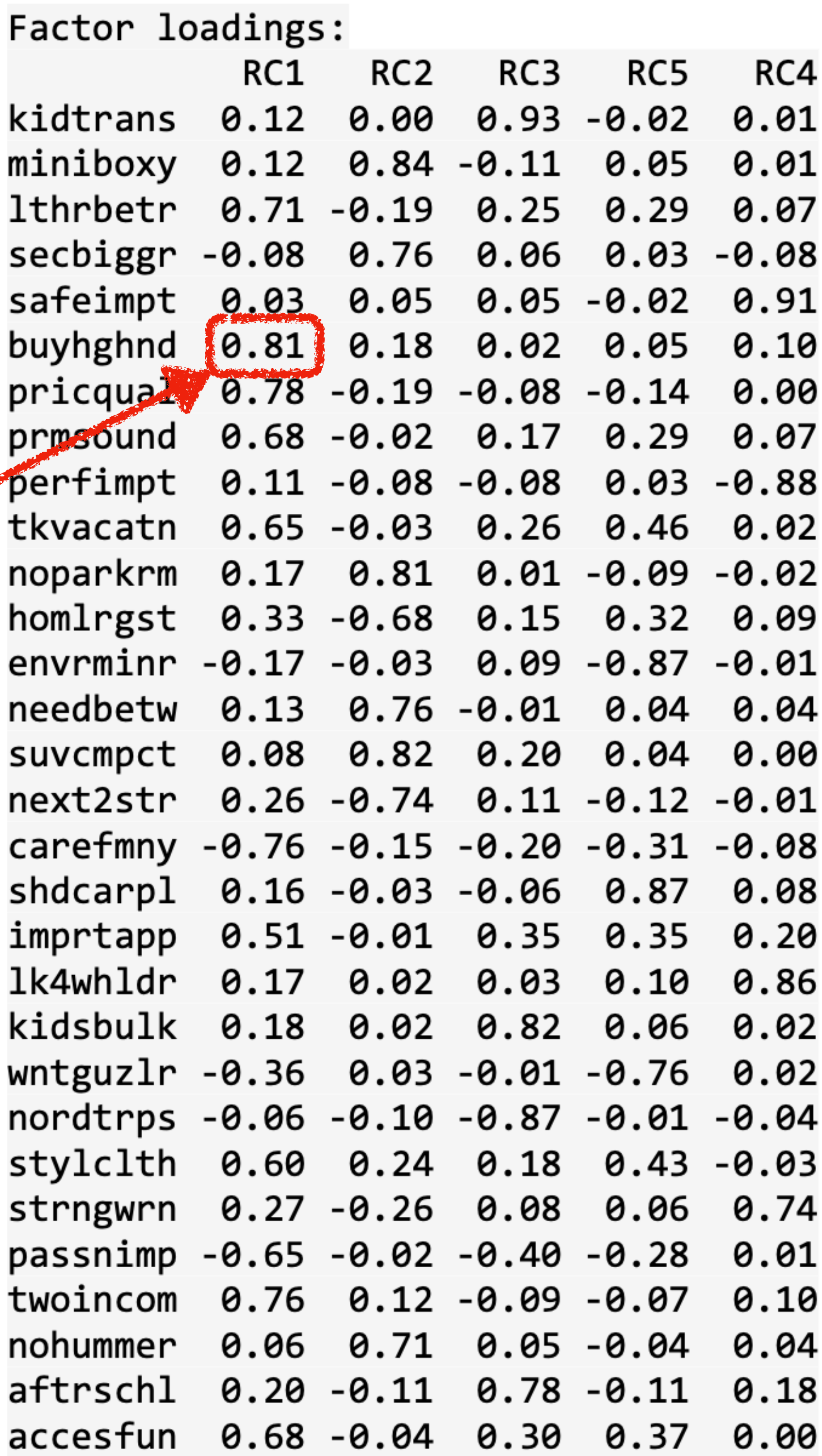




The two red lines represent the two PCs

Examining the factor loadings

This is ϕ_{61}. It’s large, meaning that the first PC has a lot to do with whether respondent tends to buy high-end cars.



```
Factor loadings:
              RC1   RC2   RC3   RC5   RC4
kidtrans     0.12  0.00  0.93 -0.02  0.01
miniboxy     0.12  0.84 -0.11  0.05  0.01
lthrbetr     0.71 -0.19  0.25  0.29  0.07
secbiggr    -0.08  0.76  0.06  0.03 -0.08
safeimpt     0.03  0.05  0.05 -0.02  0.91
buyhghnd     0.81  0.18  0.02  0.05  0.10
pricqual     0.78 -0.19 -0.08 -0.14  0.00
prmsound     0.68 -0.02  0.17  0.29  0.07
perfimpt     0.11 -0.08 -0.08  0.03 -0.88
tkvacatn     0.65 -0.03  0.26  0.46  0.02
noparkrm     0.17  0.81  0.01 -0.09 -0.02
homlrgst     0.33 -0.68  0.15  0.32  0.09
envrminr    -0.17 -0.03  0.09 -0.87 -0.01
needbetw     0.13  0.76 -0.01  0.04  0.04
suvcmpct     0.08  0.82  0.20  0.04  0.00
next2str     0.26 -0.74  0.11 -0.12 -0.01
carefmny    -0.76 -0.15 -0.20 -0.31 -0.08
shdcarpl     0.16 -0.03 -0.06  0.87  0.08
imprtapp     0.51 -0.01  0.35  0.35  0.20
lk4whldr     0.17  0.02  0.03  0.10  0.86
kidsbulk     0.18  0.02  0.82  0.06  0.02
wntguzlr    -0.36  0.03 -0.01 -0.76  0.02
nordtrps    -0.06 -0.10 -0.87 -0.01 -0.04
stylclth     0.60  0.24  0.18  0.43 -0.03
strngwrn     0.27 -0.26  0.08  0.06  0.74
passnimp    -0.65 -0.02 -0.40 -0.28  0.01
twoincom     0.76  0.12 -0.09 -0.07  0.10
nohummer     0.06  0.71  0.05 -0.04  0.04
aftrschl     0.20 -0.11  0.78 -0.11  0.18
accesfun     0.68 -0.04  0.30  0.37  0.00
```

Names	RC1	RC2	RC3	RC5	RC4
buyhghnd	0.81	0.18	0.02	0.05	0.1
pricqual	0.78	-0.19	-0.08	-0.14	0
carefmny	-0.76	-0.15	-0.2	-0.31	-0.08
twoincom	0.76	0.12	-0.09	-0.07	0.1
lthrbetr	0.71	-0.19	0.25	0.29	0.07
prmsound	0.68	-0.02	0.17	0.29	0.07
accesfun	0.68	-0.04	0.3	0.37	0
tkvacatn	0.65	-0.03	0.26	0.46	0.02
passnimp	-0.65	-0.02	-0.4	-0.28	0.01
stylclth	0.6	0.24	0.18	0.43	-0.03
imprtapp	0.51	-0.01	0.35	0.35	0.2
miniboxy	0.12	0.84	-0.11	0.05	0.01
suvcmpct	0.08	0.82	0.2	0.04	0
noparkrm	0.17	0.81	0.01	-0.09	-0.02
secbiggr	-0.08	0.76	0.06	0.03	-0.08
needbetw	0.13	0.76	-0.01	0.04	0.04
next2str	0.26	-0.74	0.11	-0.12	-0.01
nohummer	0.06	0.71	0.05	-0.04	0.04
homlrgst	0.33	-0.68	0.15	0.32	0.09
kidtrans	0.12	0	0.93	-0.02	0.01
nordtrps	-0.06	-0.1	-0.87	-0.01	-0.04
kidsbulk	0.18	0.02	0.82	0.06	0.02
aftrschl	0.2	-0.11	0.78	-0.11	0.18
envrminr	-0.17	-0.03	0.09	-0.87	-0.01
shdcarpl	0.16	-0.03	-0.06	0.87	0.08
wntguzlr	-0.36	0.03	-0.01	-0.76	0.02
safeimpt	0.03	0.05	0.05	-0.02	0.91
perfimpt	0.11	-0.08	-0.08	0.03	-0.88
lk4whldr	0.17	0.02	0.03	0.1	0.86
strngwrn	0.27	-0.26	0.08	0.06	0.74

Names	RC1	RC2	RC3	RC5	RC4
buyhghnd	0.81	0.18	0.02	0.05	0.1
pricqual	0.78	-0.19	-0.08	-0.14	0
carefmny	-0.76	-0.15	-0.2	-0.31	-0.08
twoincom	0.76	0.12	-0.09	-0.07	0.1
lthrbetr	0.71	-0.19	0.25	0.29	0.07
prmsound	0.68	-0.02	0.17	0.29	0.07
accesfun	0.68	-0.04	0.3	0.37	0
tkvacatn	0.65	-0.03	0.26	0.46	0.02
passnimp	-0.65	-0.02	-0.4	-0.28	0.01
stylclth	0.6	0.24	0.18	0.43	-0.03
imprtapp	0.51	-0.01	0.35	0.35	0.2
miniboxy	0.12	0.84	-0.11	0.05	0.01
suvcmpct	0.08	0.82	0.2	0.04	0
noparkrm	0.17	0.81	0.01	-0.09	-0.02
secbiggr	-0.08	0.76	0.06	0.03	-0.08
needbetw	0.13	0.76	-0.01	0.04	0.04
next2str	0.26	-0.74	0.11	-0.12	-0.01
nohummer	0.06	0.71	0.05	-0.04	0.04
homlrgst	0.33	-0.68	0.15	0.32	0.09
kidtrans	0.12	0	0.93	-0.02	0.01
nordtrps	-0.06	-0.1	-0.87	-0.01	-0.04
kidsbulk	0.18	0.02	0.82	0.06	0.02
aftrschl	0.2	-0.11	0.78	-0.11	0.18
envrminr	-0.17	-0.03	0.09	-0.87	-0.01
shdcarpl	0.16	-0.03	-0.06	0.87	0.08
wntguzlr	-0.36	0.03	-0.01	-0.76	0.02
safeimpt	0.03	0.05	0.05	-0.02	0.91
perfimpt	0.11	-0.08	-0.08	0.03	-0.88
lk4whldr	0.17	0.02	0.03	0.1	0.86
strngwrn	0.27	-0.26	0.08	0.06	0.74

Luxury

- Buy high end cars
- Price reflects quality
- Careful with money
- Hard to subsist on one income

Size

- Minivans too boxy and large
- SUVs better because more compact
- Don’t have a lot of parking room
- Need something between sedan and minivan

Kid Carrier

- Need car to transport kids
- No road trips with our family
- Kids have bulky items and toys with them

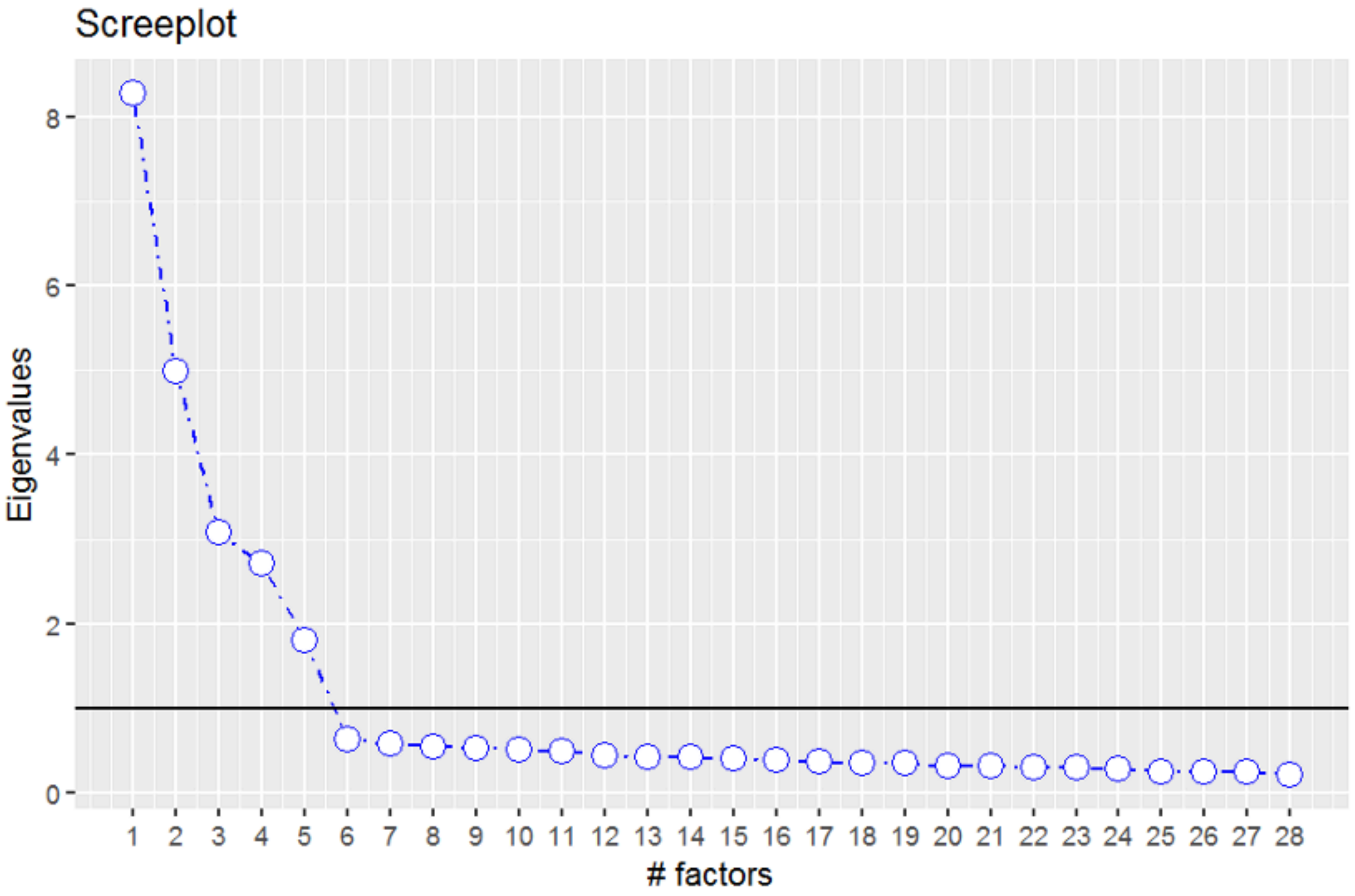
Eco Friendly

- Environmental impact of auto is small
- People should carpool

Safety Focused

- Auto safety is very important to me
- Performance is very important

How did we decide on 5 PCs?

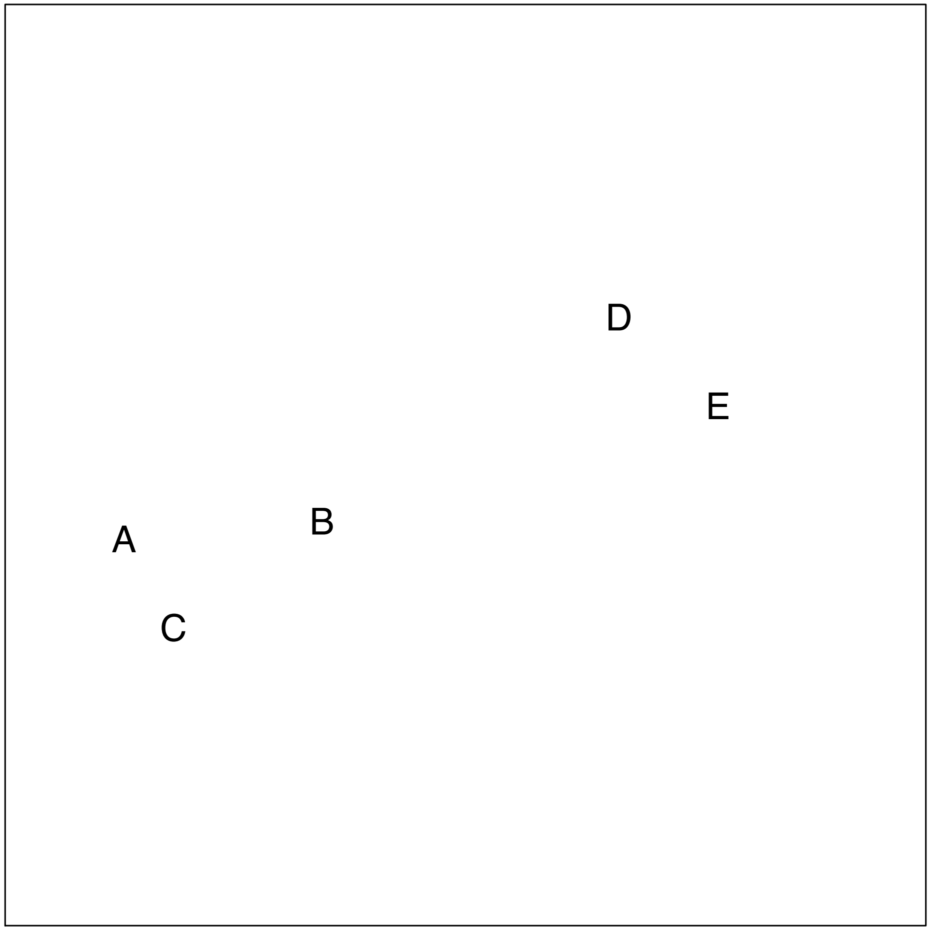


The y-axis tells us how much of the variability is explained. We look for an "elbow" in the plot.

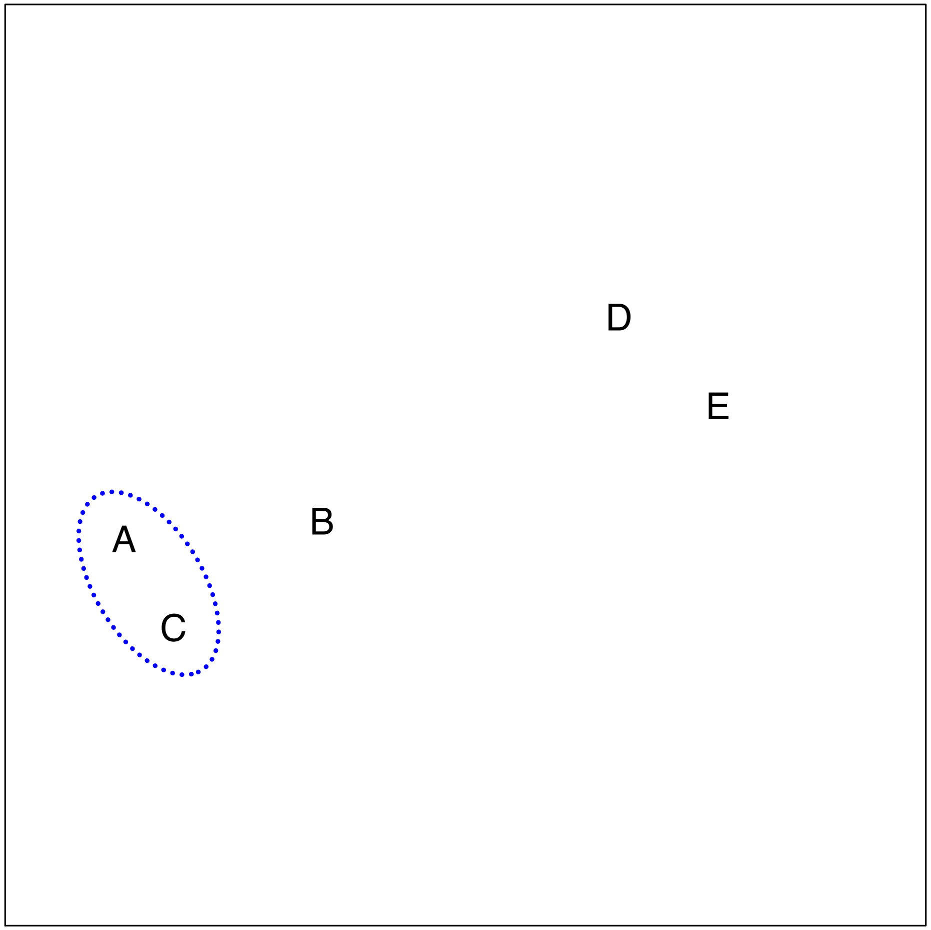
Clustering

- Partition observations into distinct subgroups (“clusters”)
- Observations within a cluster should be quite similar
- Observations in different clusters should be quite different
- Many methods for clustering.
- *Our focus:* Hierarchical clustering.

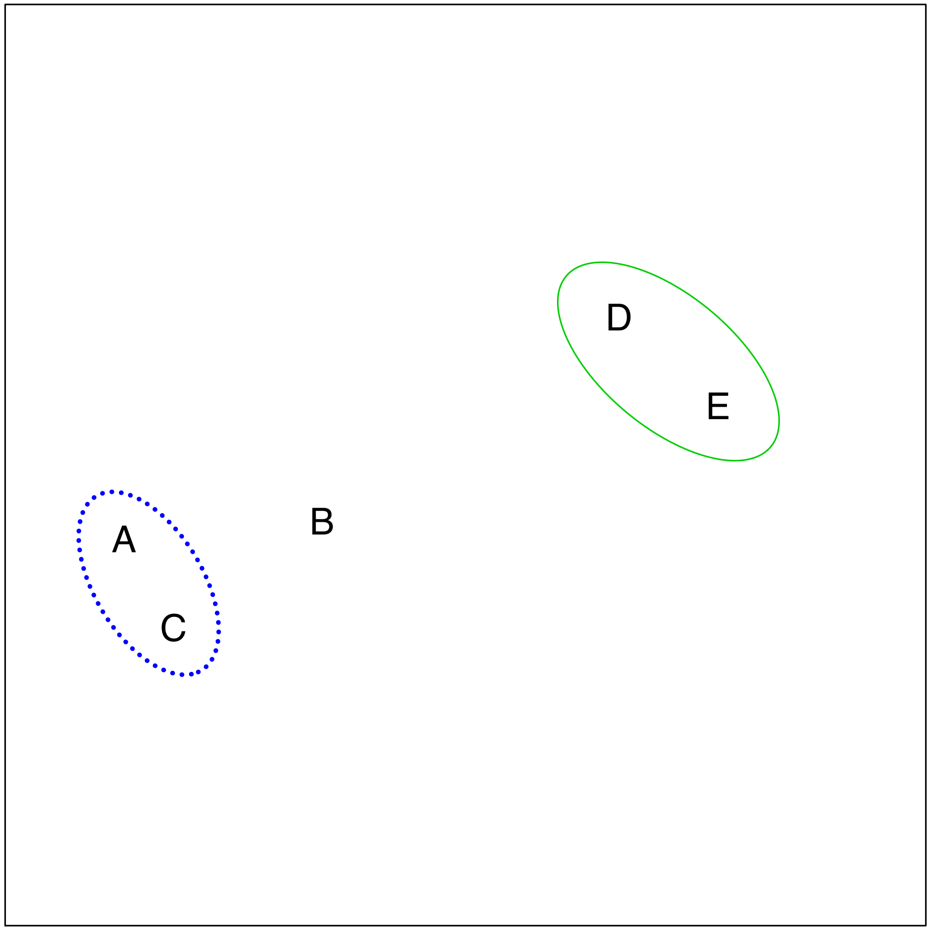
Hierarchical clustering



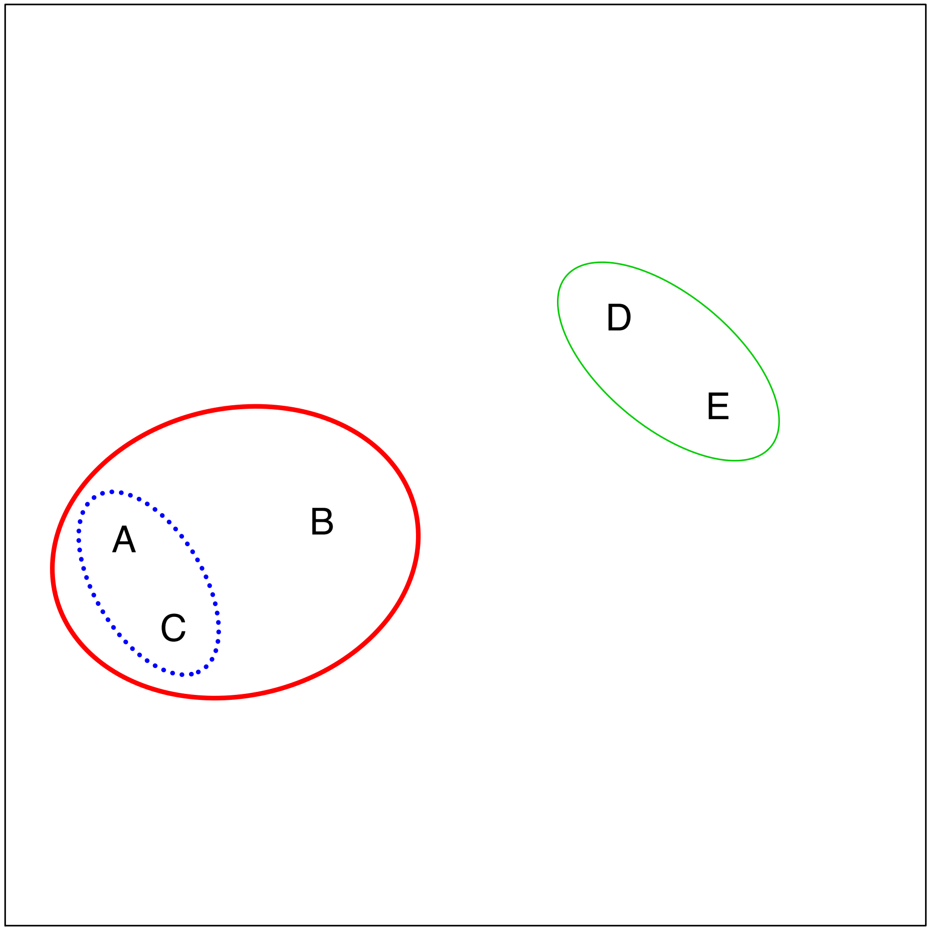
Hierarchical clustering



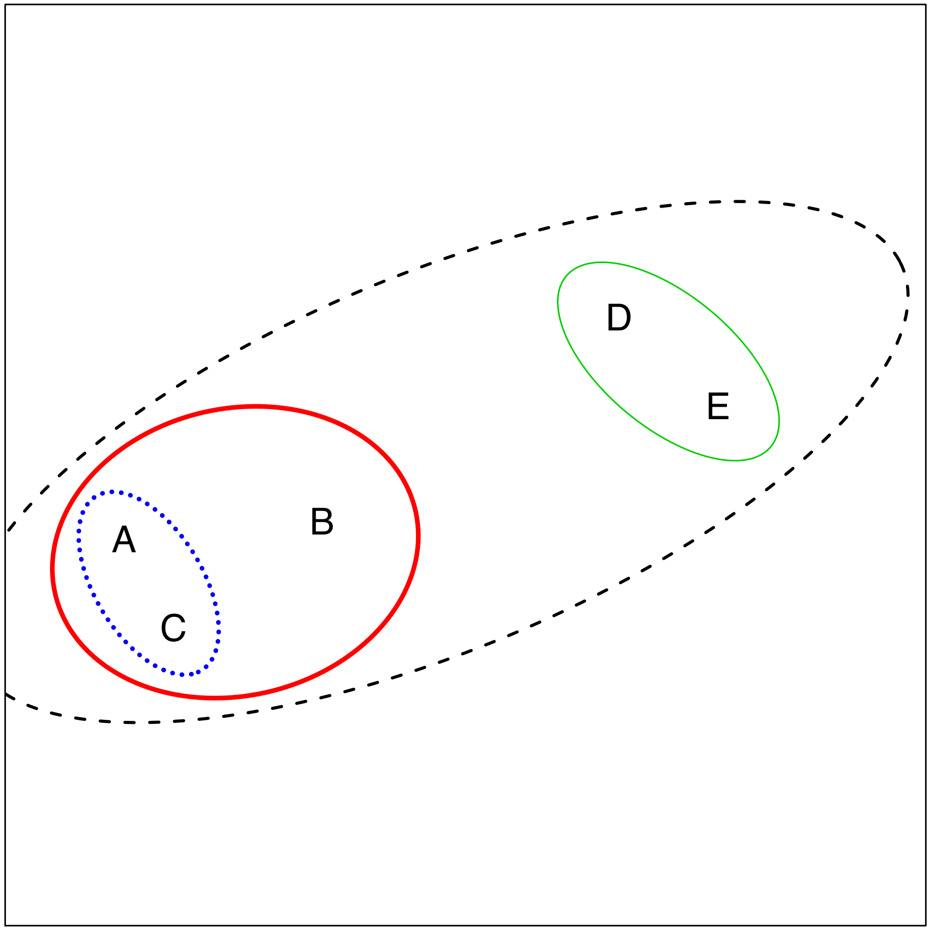
Hierarchical clustering



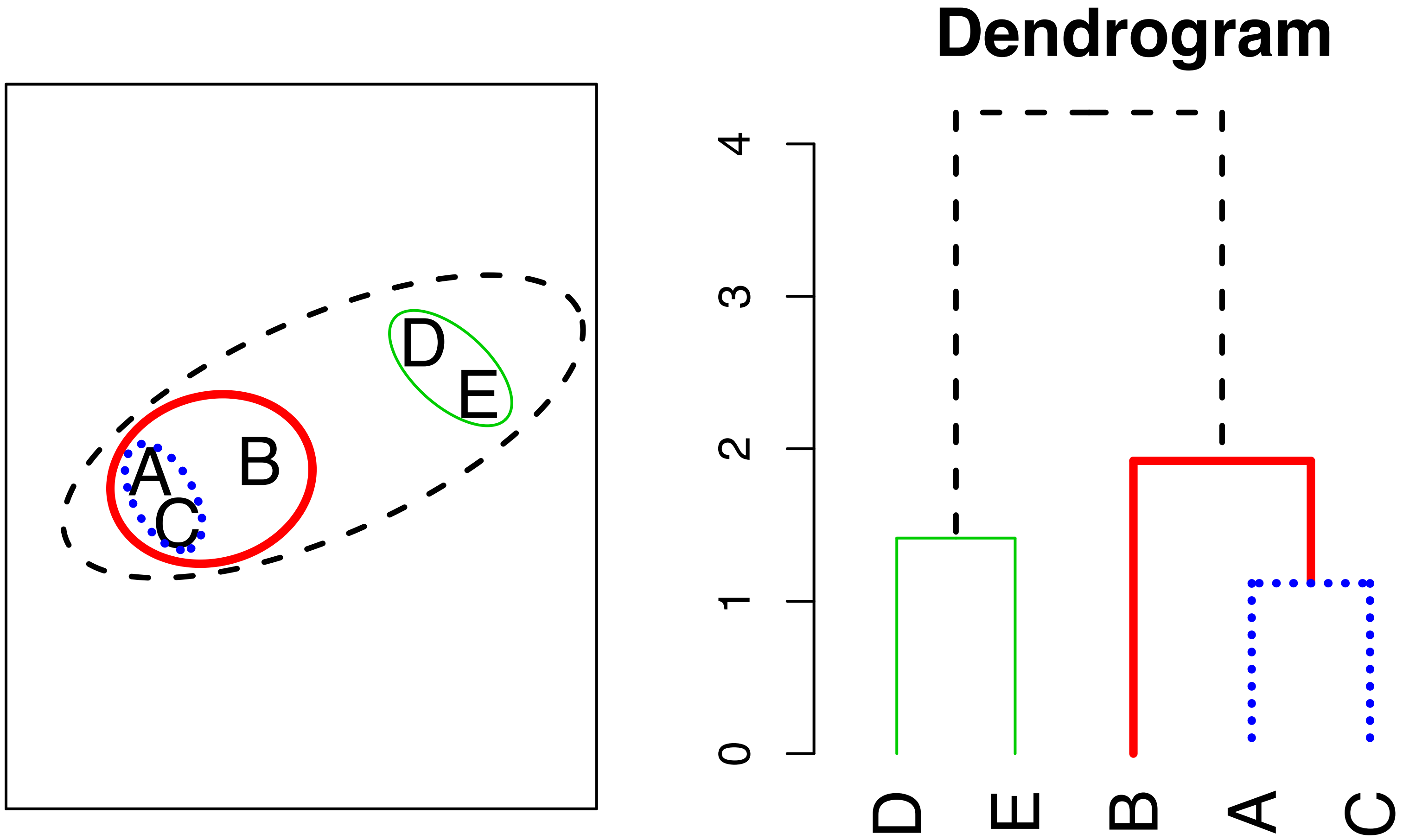
Hierarchical clustering



Hierarchical clustering

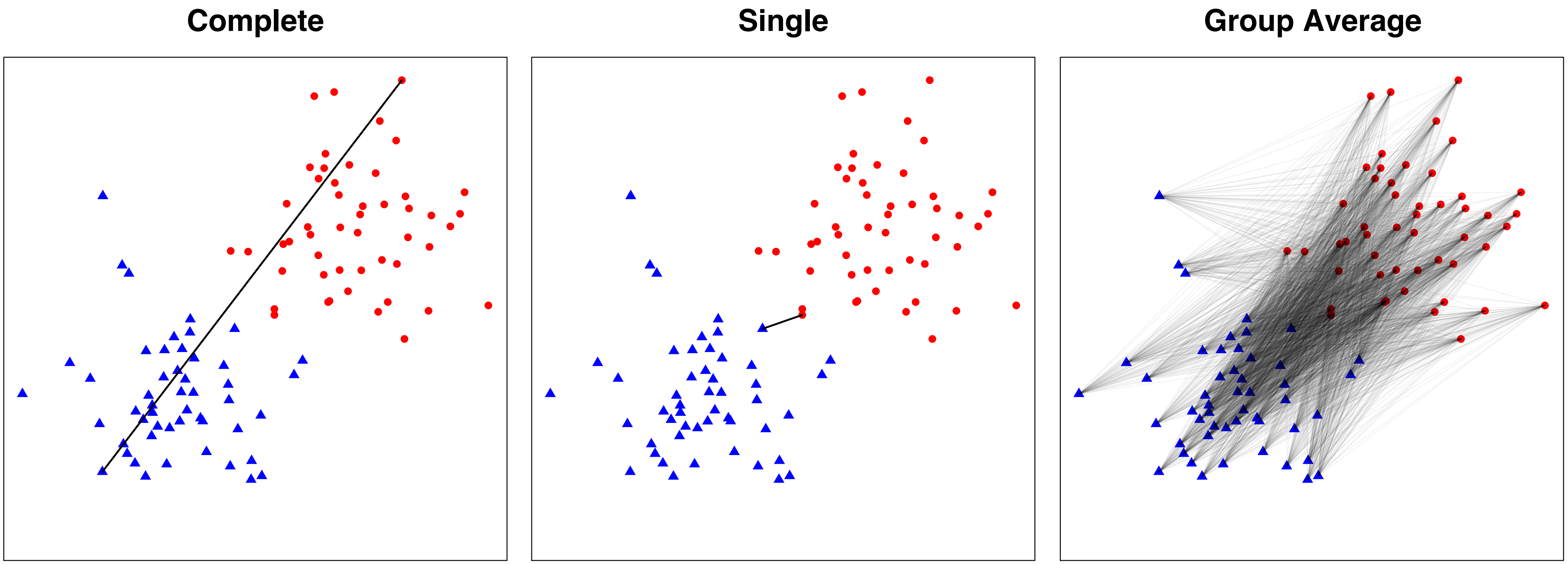


Hierarchical clustering



Choice of “linkage”

Specifies how to measure distance between two groups

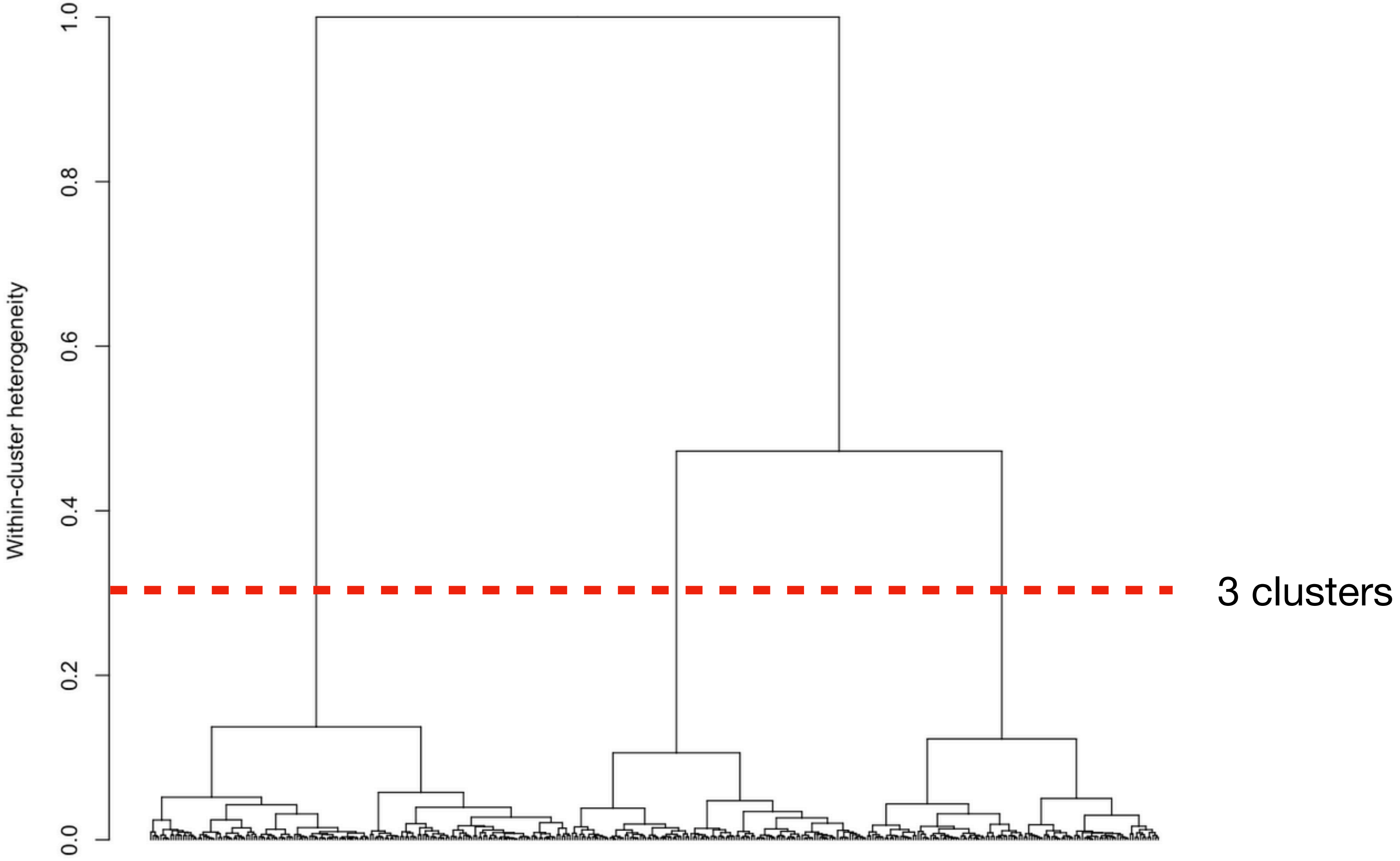


Complete

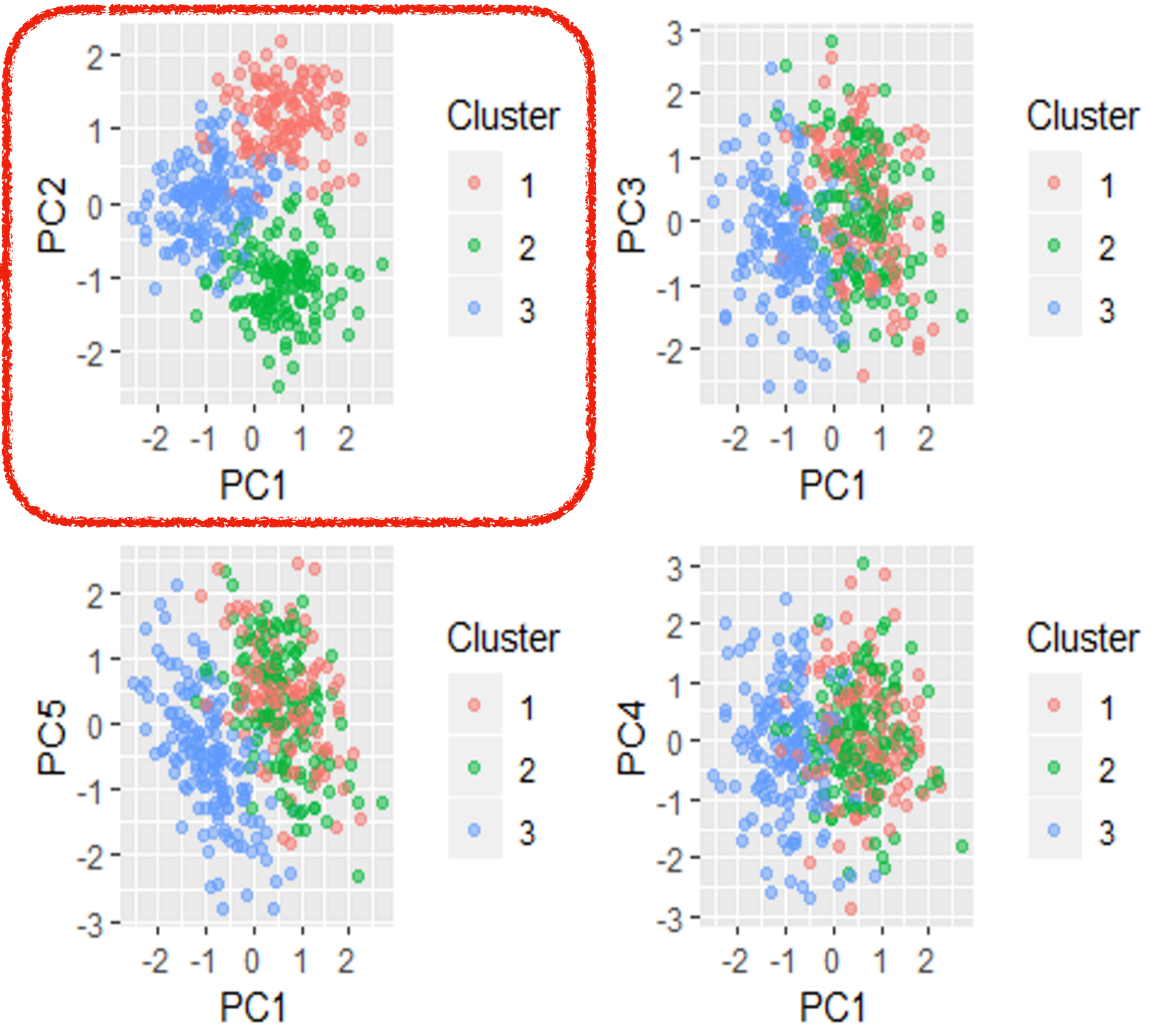
Single

Group Average

Microvans dendrogram



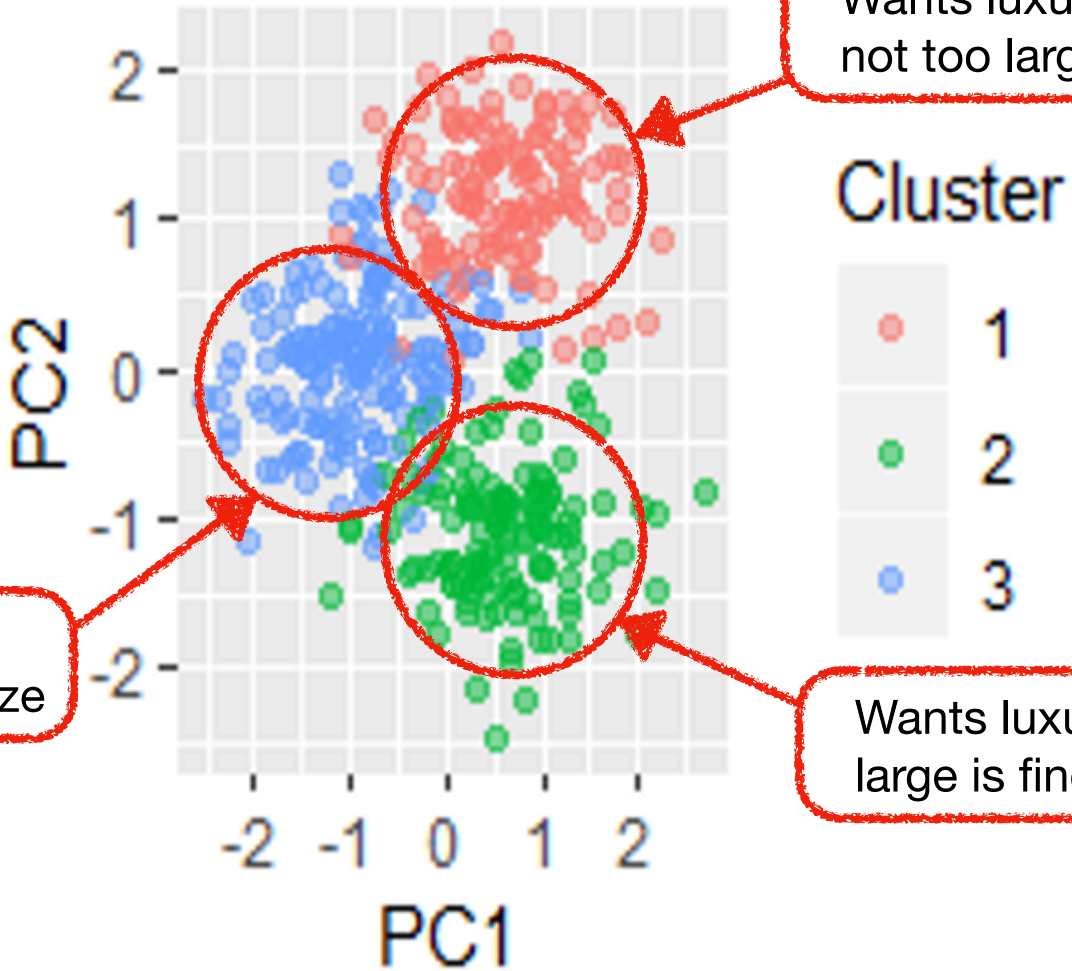
Examining clusters



Only PC1 and PC2 (Luxury and Size) seem to influence the clusters.

PC	Interpretation
PC1	Luxury
PC2	Size
PC3	Kid Carrier
PC4	Eco Friendly
PC5	Safety Focused

Examining clusters



Wants luxury vehicle that is not too large

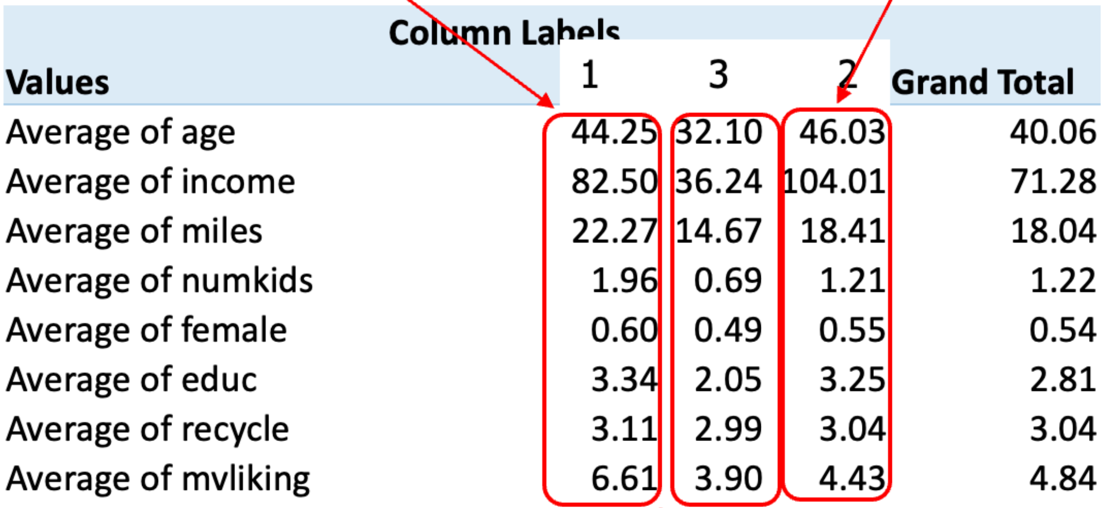
Prefer cheaper vehicle and don’t care about size

Wants luxury vehicle, but large is fine

Demographics of clusters

Older
Fairly high income
Drive further
More kids
Higher fraction female
Higher Education
High MVLiking

Older
Highest Income
Medium driving
Medium kids
High Education
Low MVLiking



	Column Labels			
Values	1	3	2	**Grand Total**
Average of age	44.25	32.10	46.03	40.06
Average of income	82.50	36.24	104.01	71.28
Average of miles	22.27	14.67	18.41	18.04
Average of numkids	1.96	0.69	1.21	1.22
Average of female	0.60	0.49	0.55	0.54
Average of educ	3.34	2.05	3.25	2.81
Average of recycle	3.11	2.99	3.04	3.04
Average of mvliking	6.61	3.90	4.43	4.84

Younger
Low income
Less driving
Few kids
Lower Education
Lowest MVLiking