ExamPA.net

Unsupervised Learning

Lesson

Learning Objectives

8. **Topic: Cluster and Principal Component Analyses**
Learning Objectives
The candidate will be able to apply cluster and principal components analysis to enhance supervised learning.
Learning Outcomes
The Candidate will be able to: a) Understand and apply *K*-means clustering. b) Understand and apply hierarchical clustering. c) Understand and apply principal component analysis.

+ Correlation analysis

What topics do you need to study?

Exam Date	Correlations	K-means	Hierarchical Clustering	Principal Component Analysis
6/19/2020	1			
6/18/2020	1			1
6/17/2020	1			1
6/16/2020	1			1
12/13/2019				
12/12/2019				
6/14/2019				1
6/13/2019				1

- Hospital Readmissions – K-means clustering
- Apartment Applicants & Health Costs - Hierarchical clustering

ExamPA.net

Unsupervised Learning

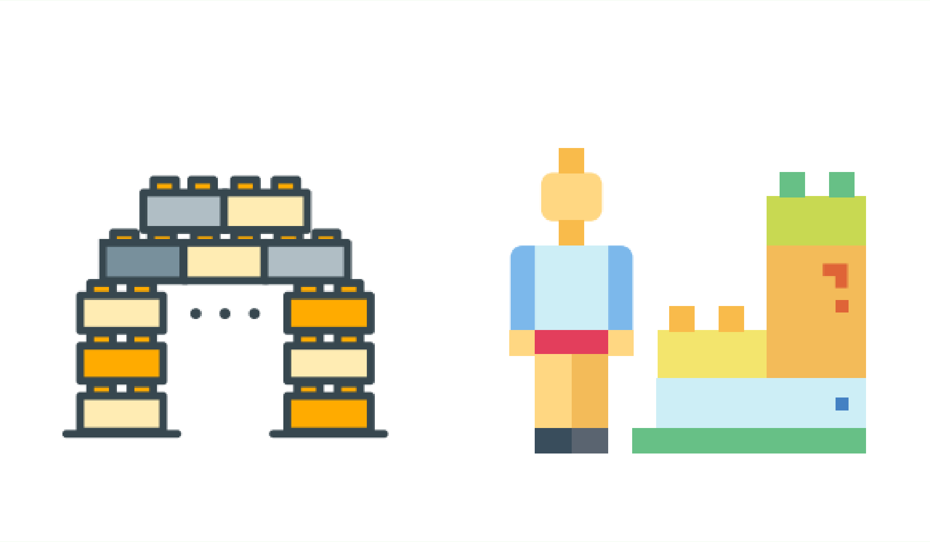
Unsupervised Learning

Supervised Learning

Unsupervised Learning

Supervised Learning

Unsupervised Learning



Unsupervised Learning

	Supervised Learning	Unsupervised Learning
Discrete		
Continuous		
	There **is** a target	There **is not** a target

Unsupervised Learning

	Supervised Learning	Unsupervised Learning
Discrete	Classification	Clustering
Continuous	Regression	Dimensionality Reduction
	There **is** a target	There **is not** a target

Unsupervised Learning

Supervised	Unsupervised
GLM	Correlation analysis
Lasso, Ridge, and Elastic Net	Principal component analysis (PCA)
Decision Tree	K-means clustering
Bagged Tree	Hierarchical clustering
Boosted Tree	

Semi-Supervised Learning: Using PCA or Clustering to create features that are used in a supervised model

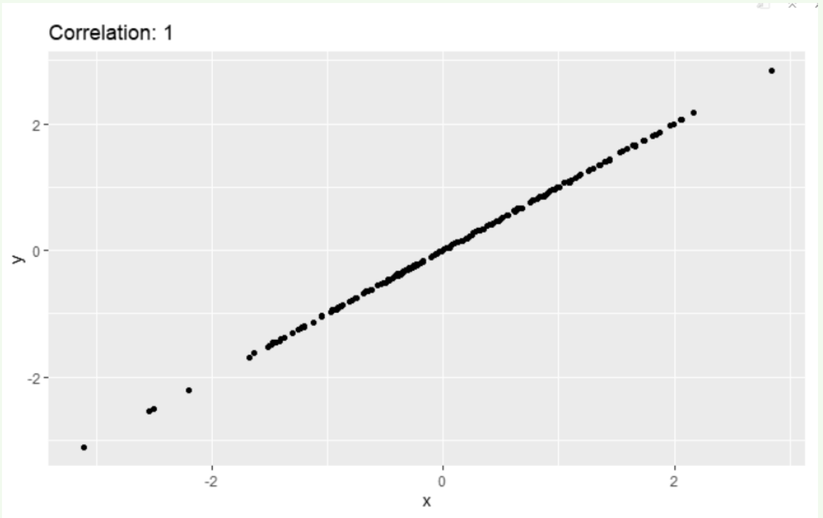
Pearson's Correlation

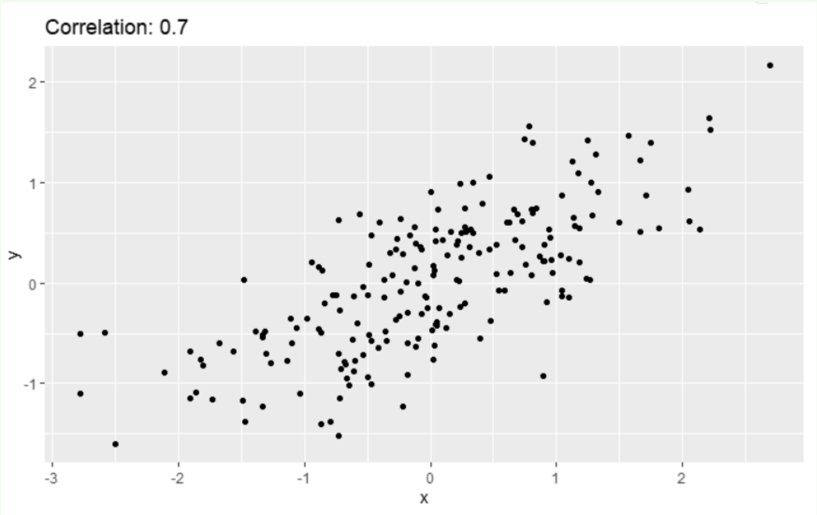
- Two variables are said to be **positively correlated** when increasing one tends to increase the other and **negatively correlated** when increasing one decreases the other
- Correlation is **unsupervised** because it does not depend on the target variable
- Only works for numeric variables
- The most common form: Pearson's Correlation

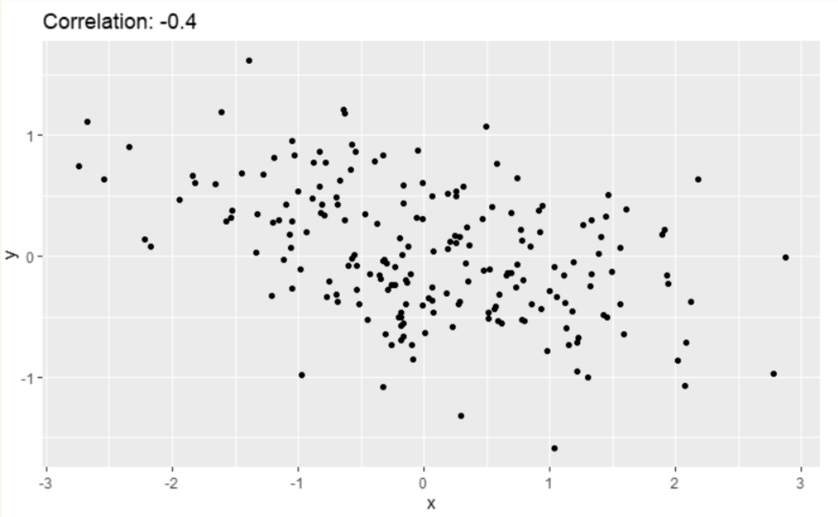
$$r = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2 \sum \left(y_i - \bar{y}\right)^2}}$$

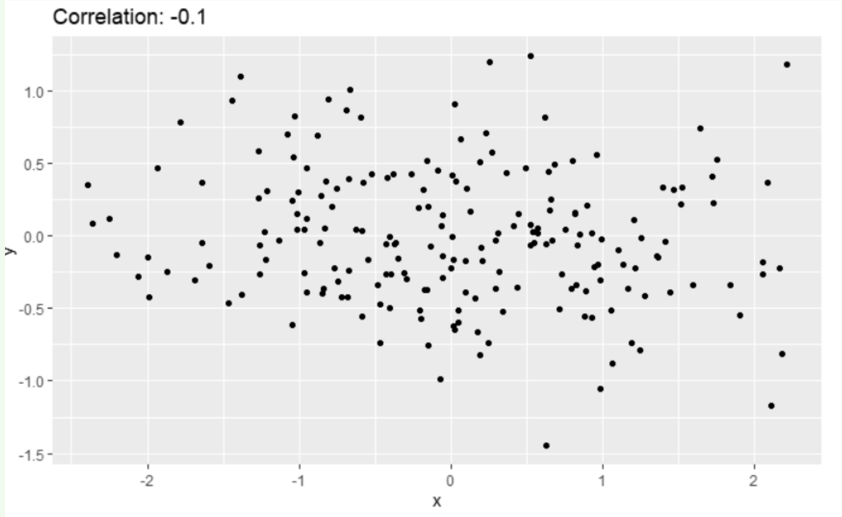
No need to memorize

What does it look like









Does not equal causation

Drownings rise when ice cream sales rise. It may seem that increased ice cream sales cause more drowning,

Does not equal causation

Drownings rise when ice cream sales rise. It may seem that increased ice cream sales cause more drowning,

Rising heat may cause more people to swim, as well as buy more ice cream.

Does not equal causation

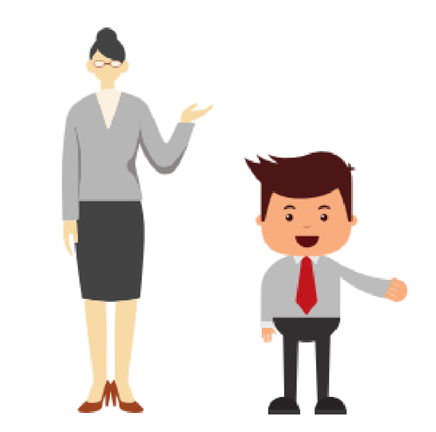
The U.S. murder rate from 2006-2011 dropped at the same rate as Microsoft Internet Explorer usage.

Does not equal causation

The U.S. murder rate from 2006-2011 dropped at the same rate as Microsoft Internet Explorer usage.

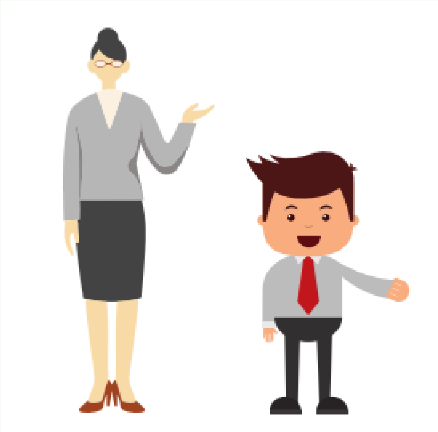
Google Chrome was launched in 2007 and the government increased funding to police departments in high-crime cities

Does not equal causation



Executives who say please and thank you more often enjoy better share performance.

Does not equal causation



Executives who say please and thank you more often enjoy better share performance.

People who take the extra effort to be polite also take the extra effort to do their job well

Example: SOA PA 6/16/20

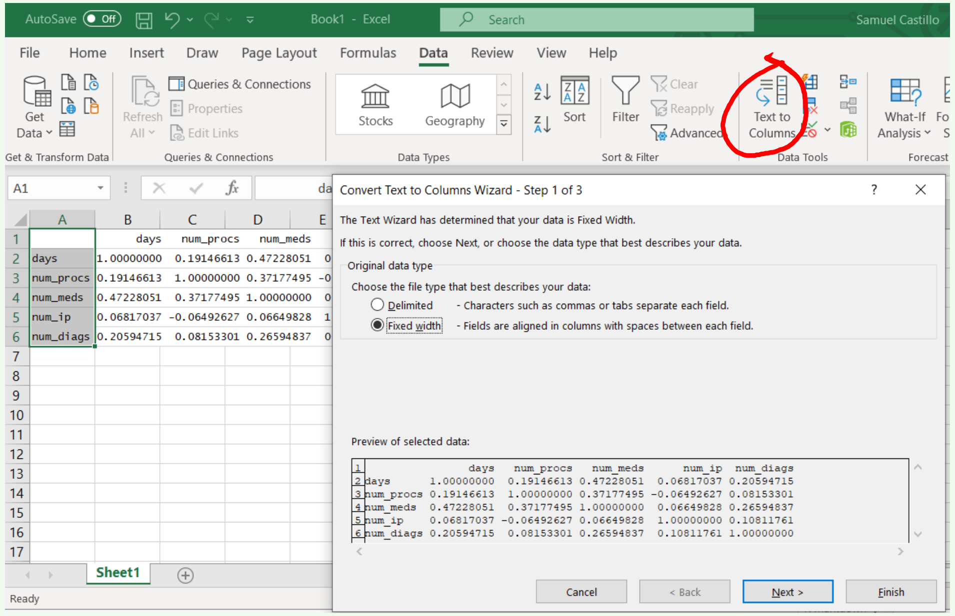
Exploratory Analysis

Target (Days spent in hospital)	Number of medical procedures	Number of prescriptions	Number of prior hospital visits

Example: SOA PA 6/16/20

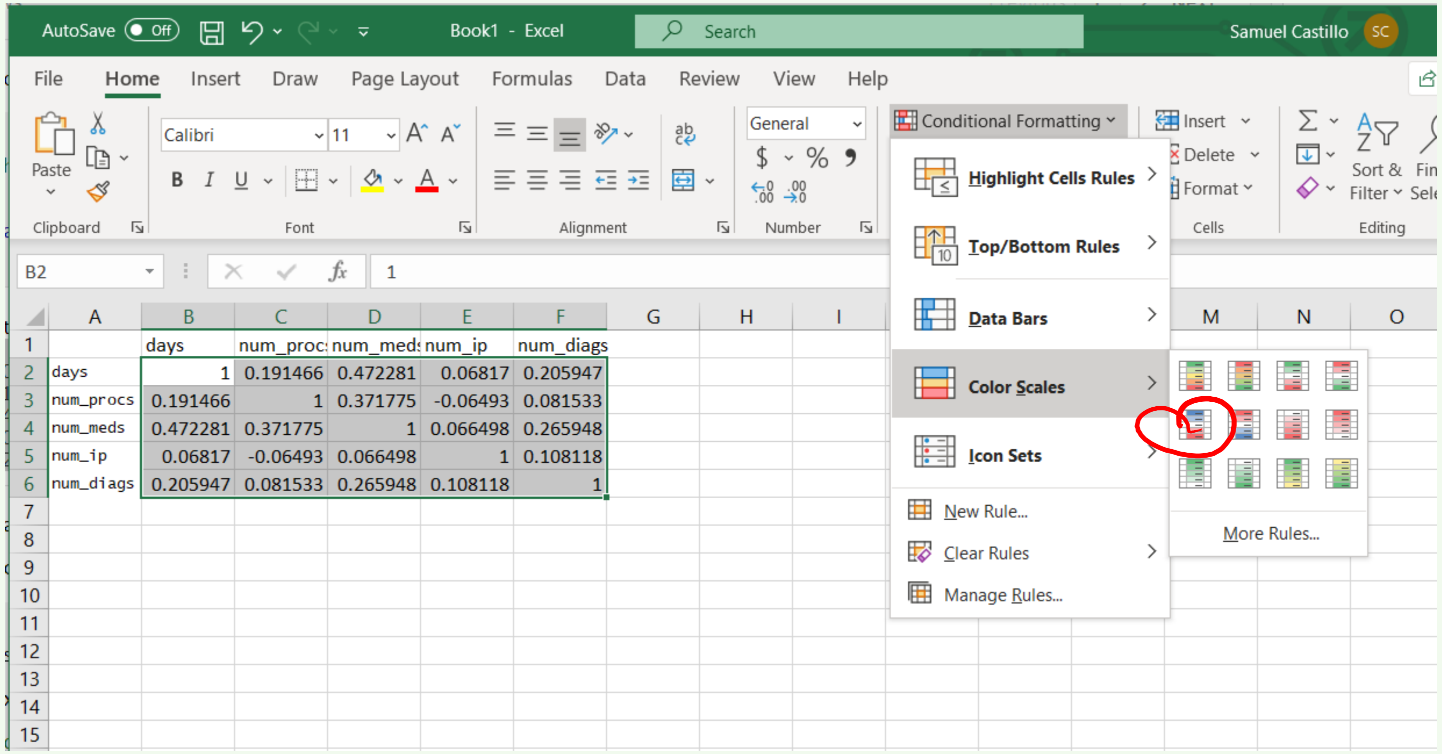
Exploratory Analysis

```
[1] "Correlation Matrix"
                days   num_procs   num_meds      num_ip  num_diags
days      1.00000000  0.19146613 0.47228051  0.06817037 0.20594715
num_procs 0.19146613  1.00000000 0.37177495 -0.06492627 0.08153301
num_meds  0.47228051  0.37177495 1.00000000  0.06649828 0.26594837
num_ip    0.06817037 -0.06492627 0.06649828  1.00000000 0.10811761
num_diags 0.20594715  0.08153301 0.26594837  0.10811761 1.00000000
```



Example: SOA PA 6/16/20

Exploratory Analysis



Example: SOA PA 6/16/20

Exploratory Analysis

	days	num_procs	num_meds	num_ip	num_diags
days	1.0	0.2	0.5	0.1	0.2
num_procs	0.2	1.0	0.4	-0.1	0.1
num_meds	0.5	0.4	1.0	0.1	0.3
num_ip	0.1	-0.1	0.1	1.0	0.1
num_diags	0.2	0.1	0.3	0.1	1.0

Multicollinearity in GLMs

Problem	Solutions
• Correlation among predictors or multicollinearity → Model instability → Extremely high or low coefficients → Standard errors which are very large • Not a problem for tree-based models	1. For any group of correlated predictors, remove all but one from the model 2. Pre-process the data using a dimensionality reduction technique such as PCA

How to find

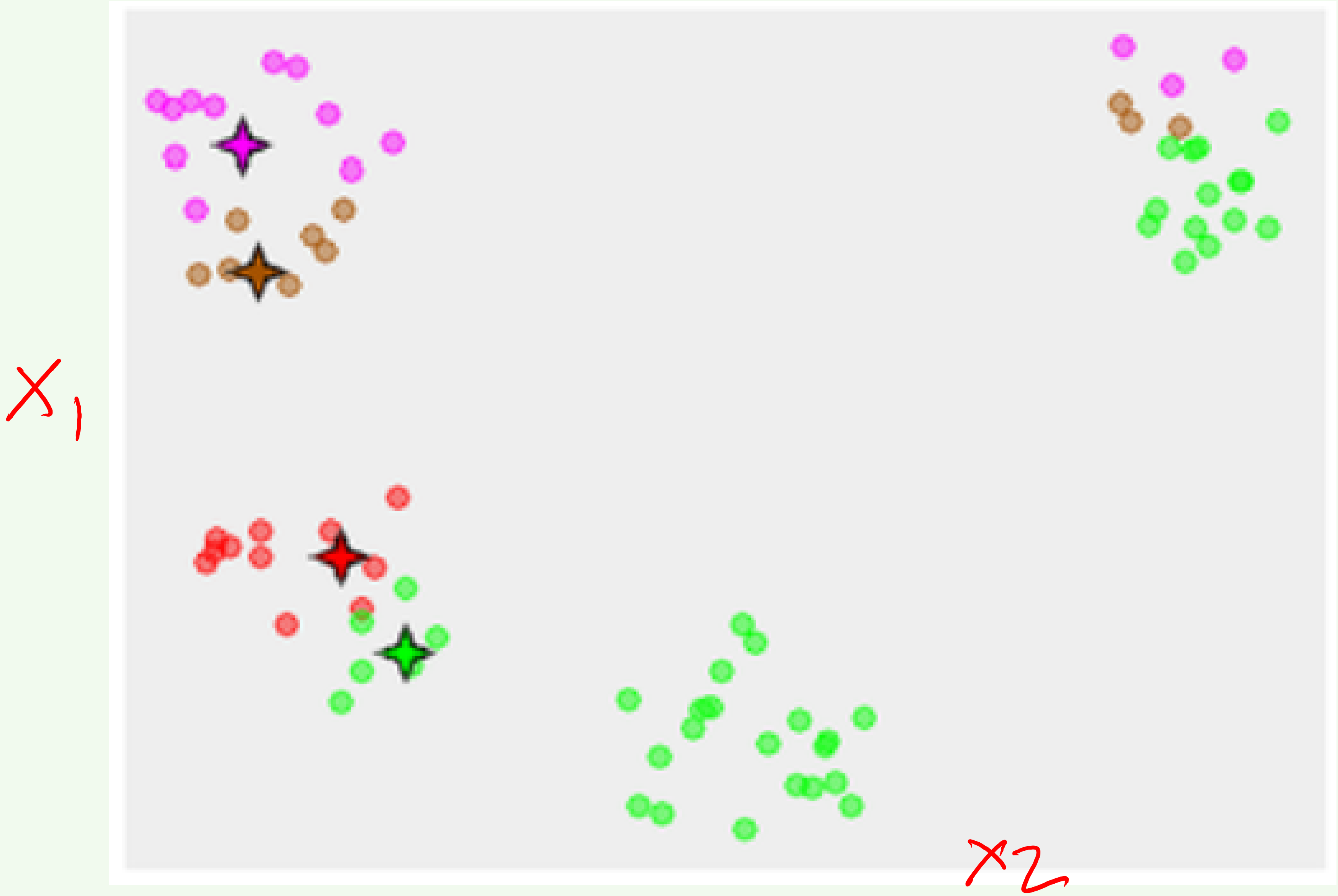
```
Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  0.7179703  0.0316229  22.704  < 2e-16 ***
genderMale  -0.0348400  0.0116236  -2.997 0.002723 ** 
age[60-70)  -0.0805106  0.0165719  -4.858 1.18e-06 ***
num_procs    0.0112080  0.0036594   3.063 0.002193 ** 
num_meds     0.0308537  0.0007074  43.618  < 2e-16 ***
num_ip       0.0140475  0.0044002   3.192 0.001411 ** 
num_diags    0.0273086  0.0035108   7.779 7.34e-15 ***
num_procs2          NA         NA      NA       NA    
```

Rank deficient

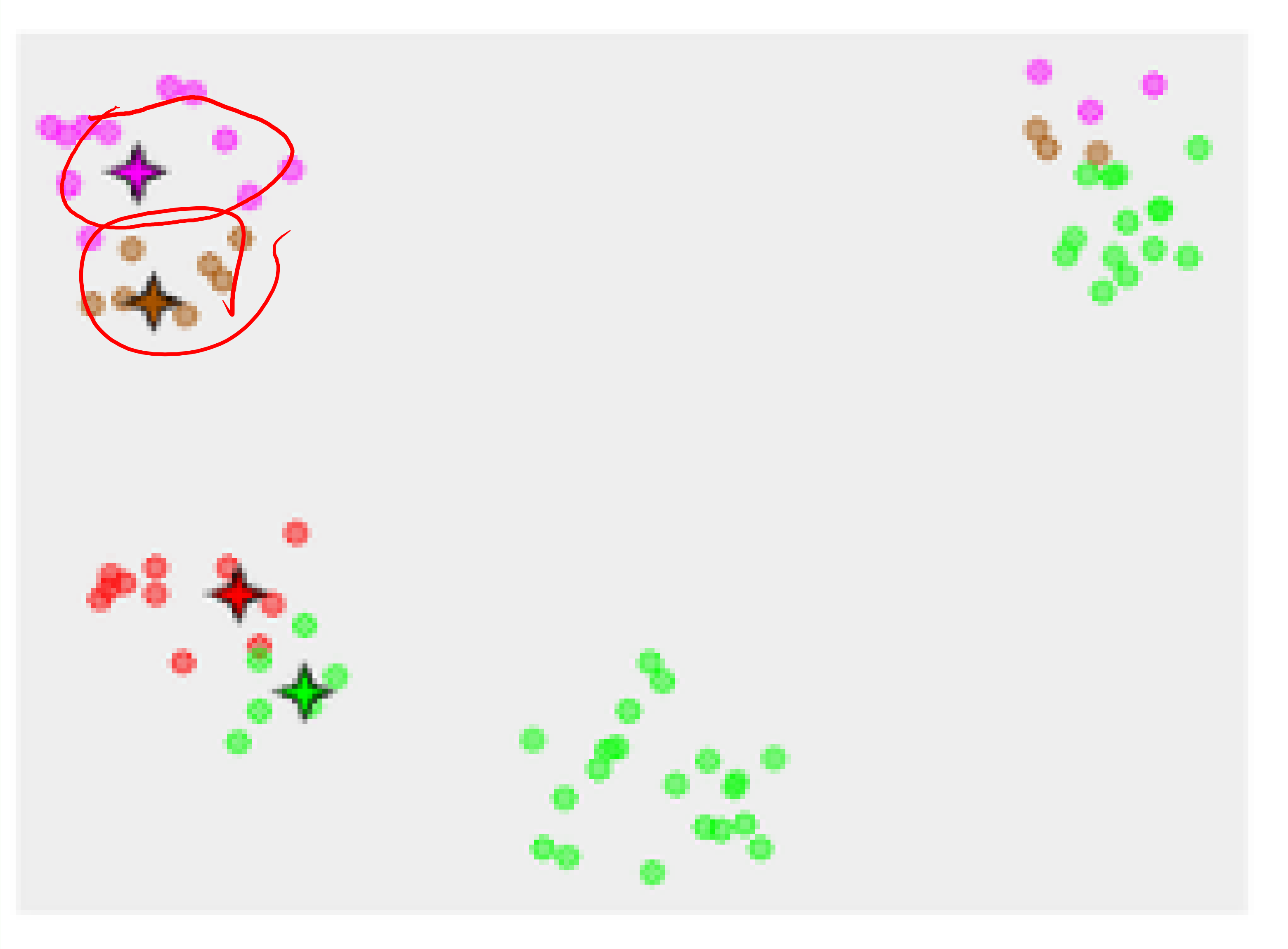
K-means clustering

Kmeans

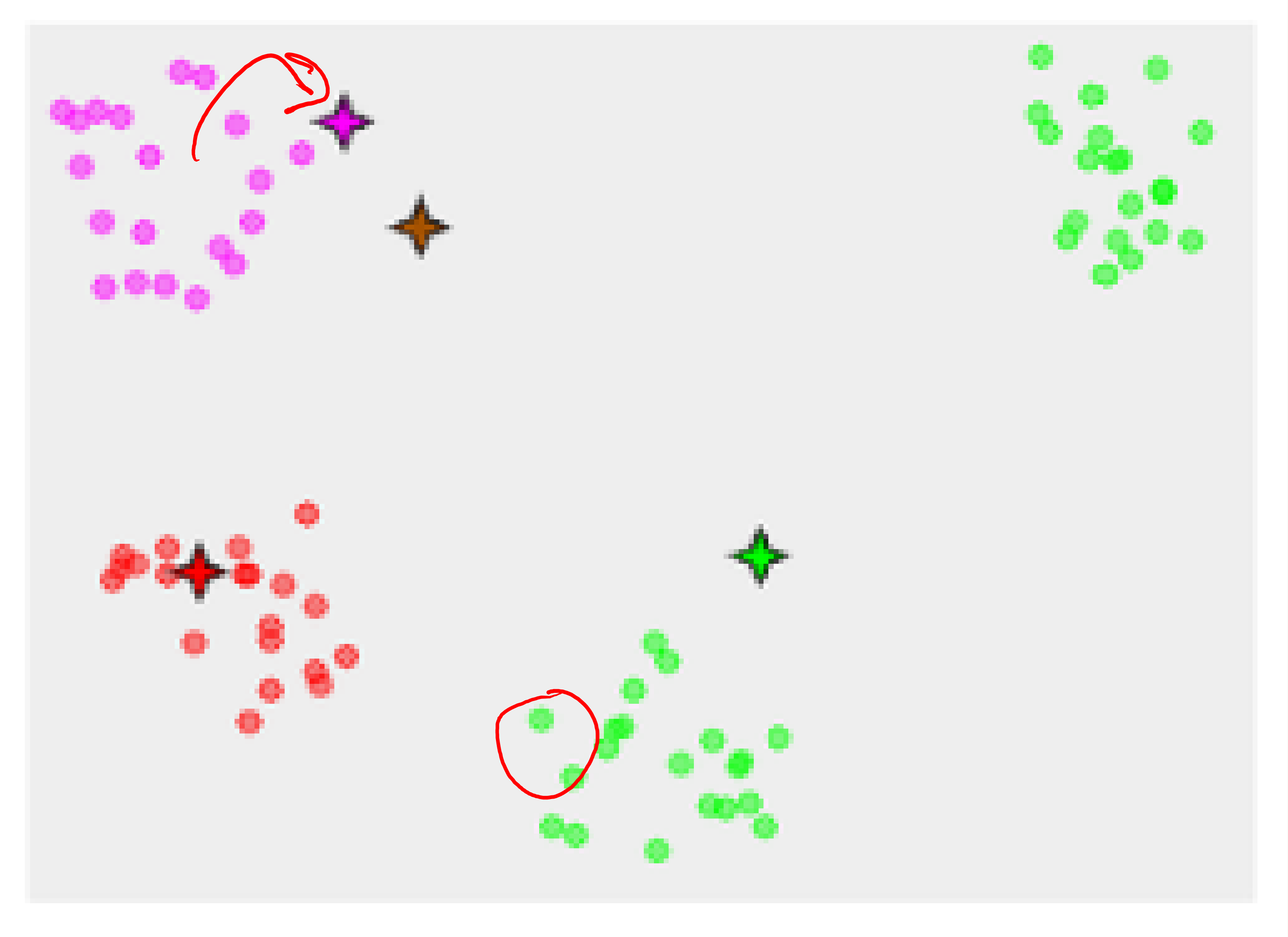
$k = 4$



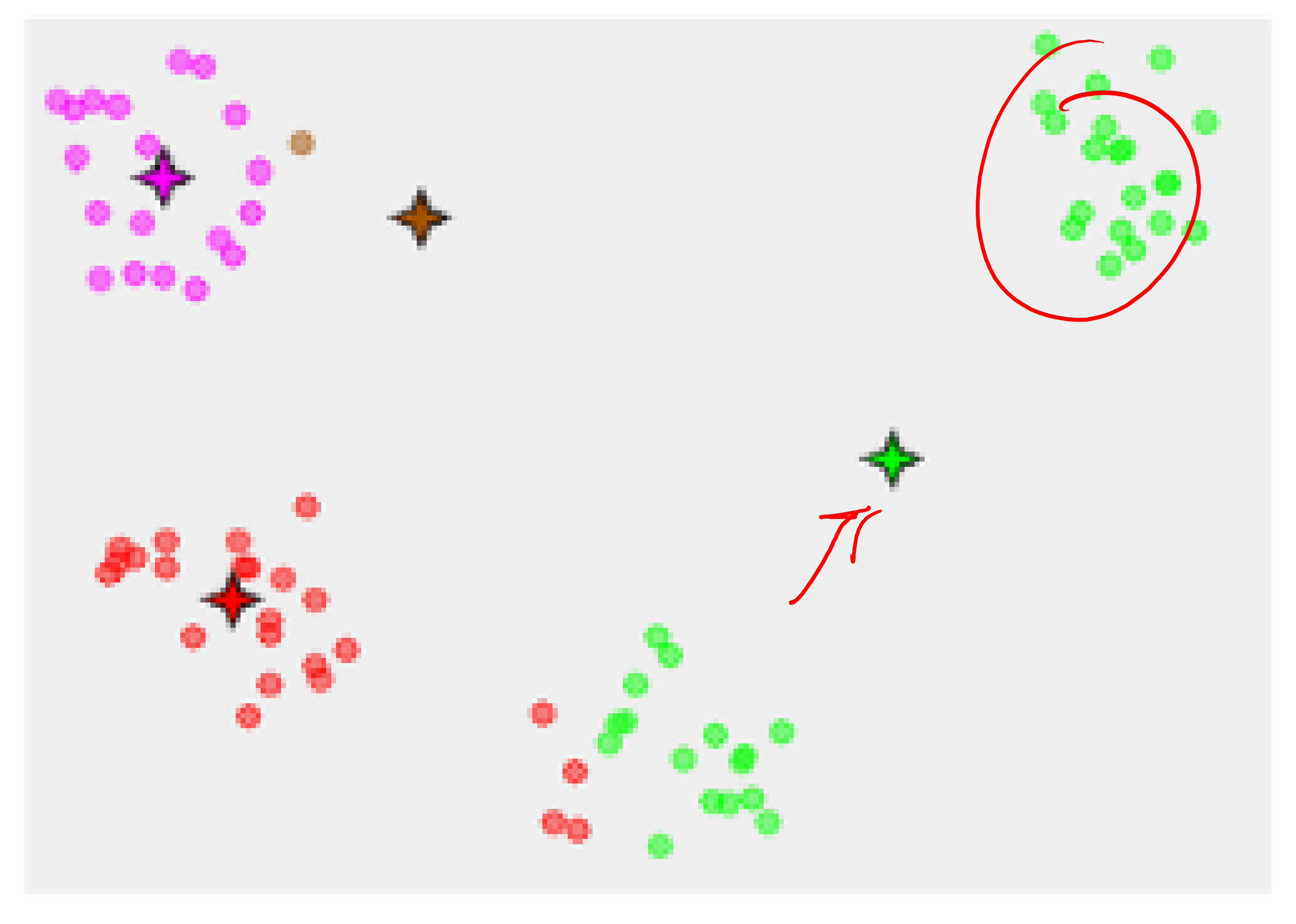
Kmeans



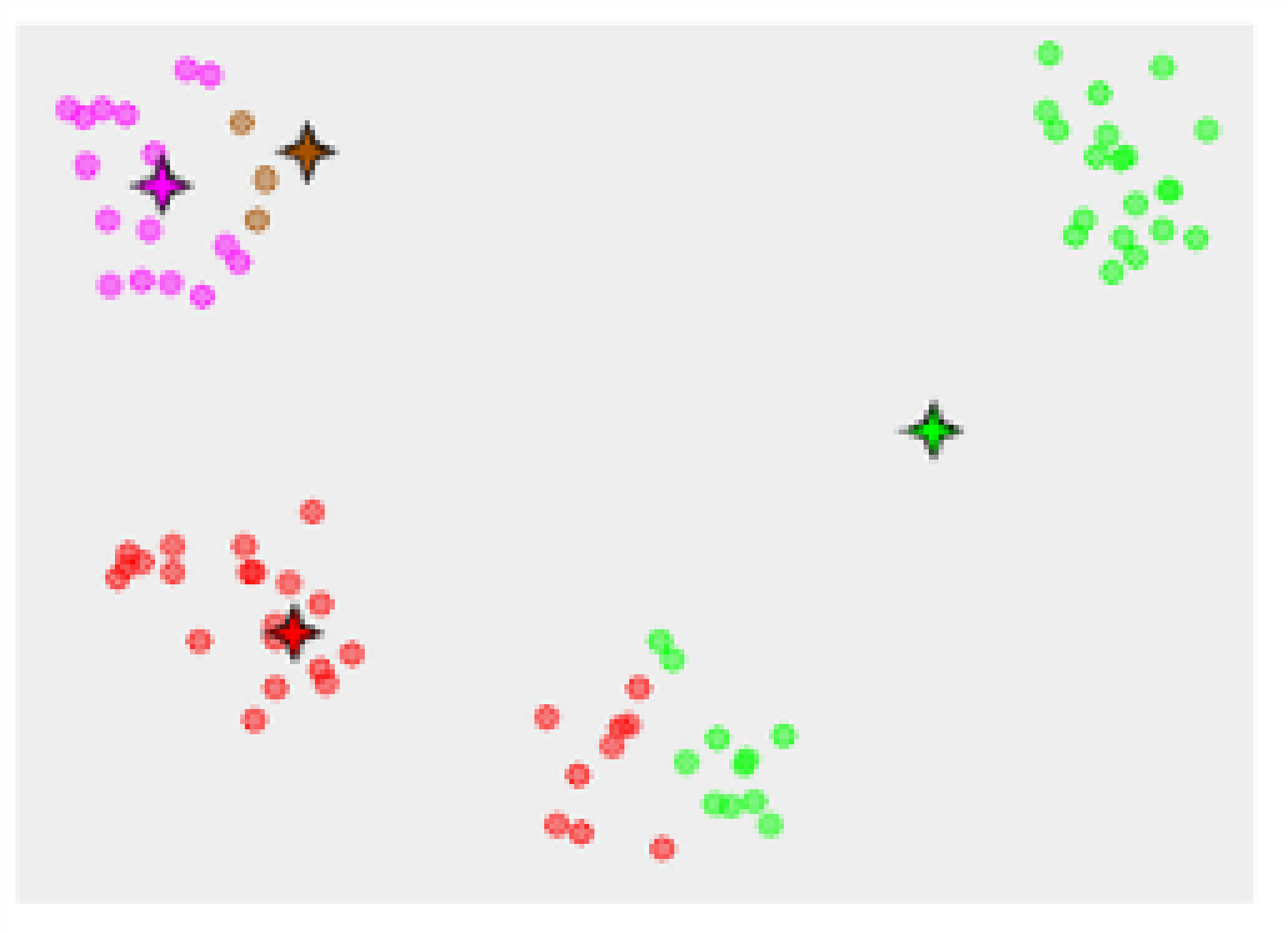
Kmeans



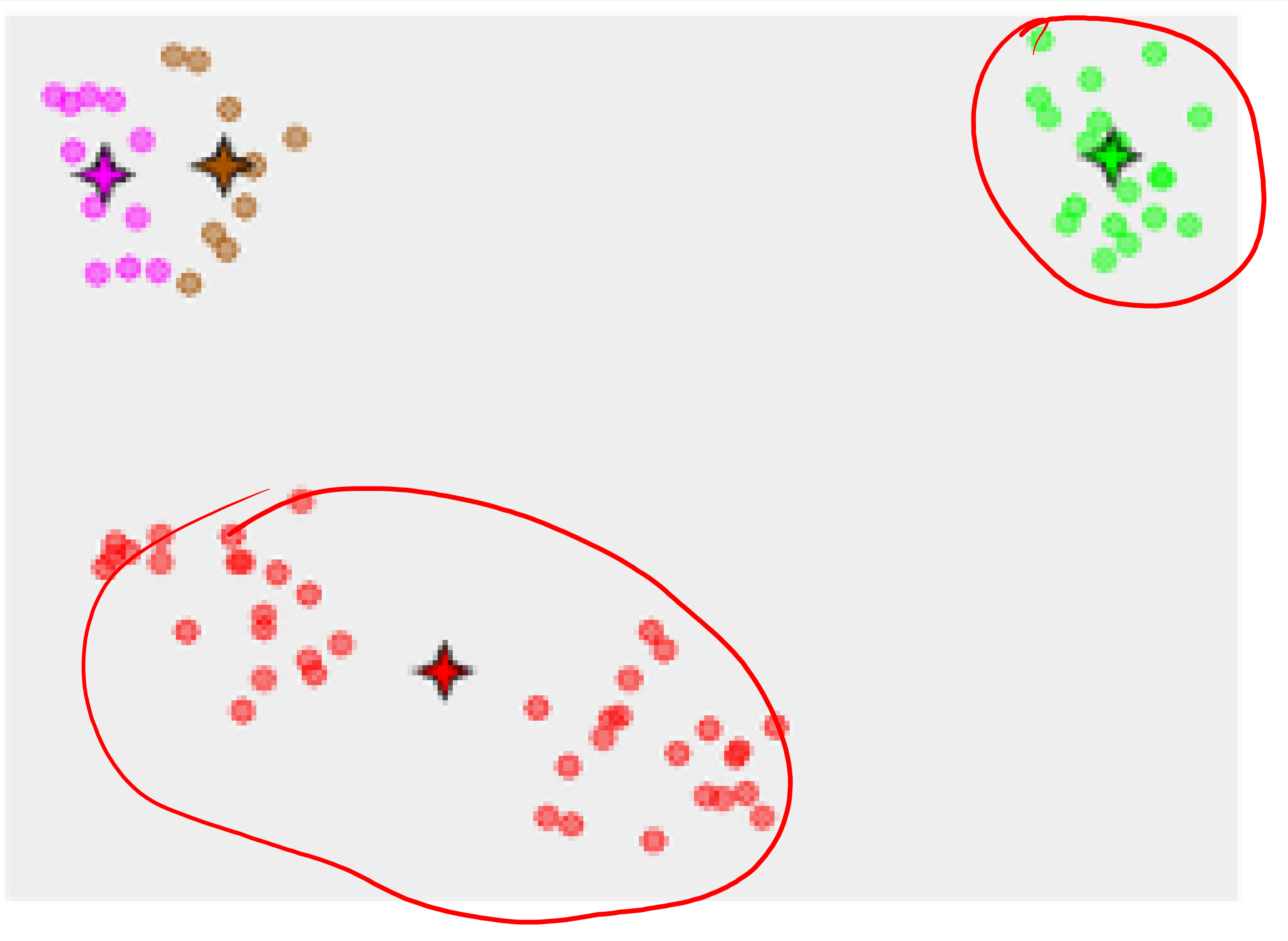
Kmeans



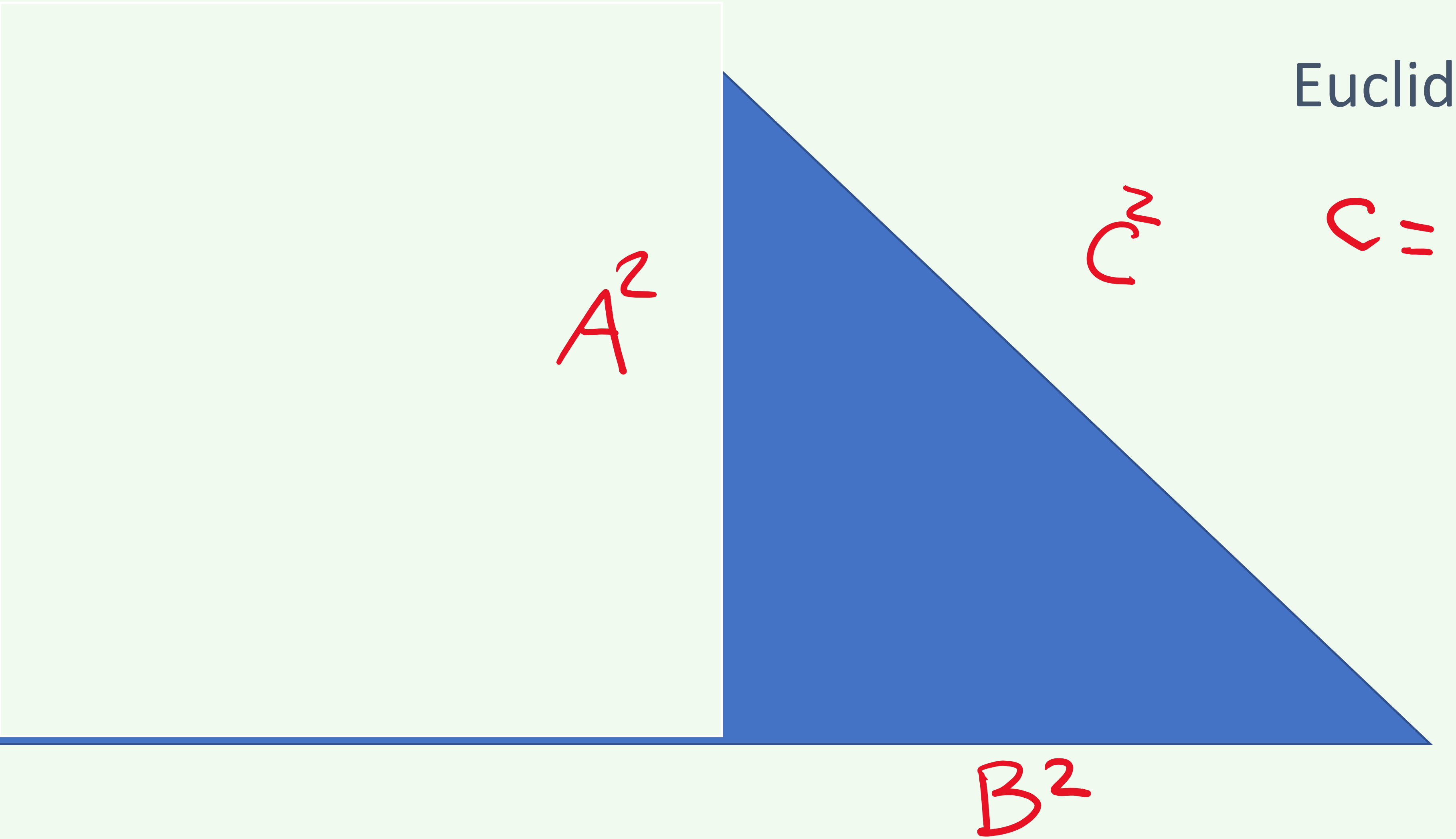
Kmeans



Kmeans



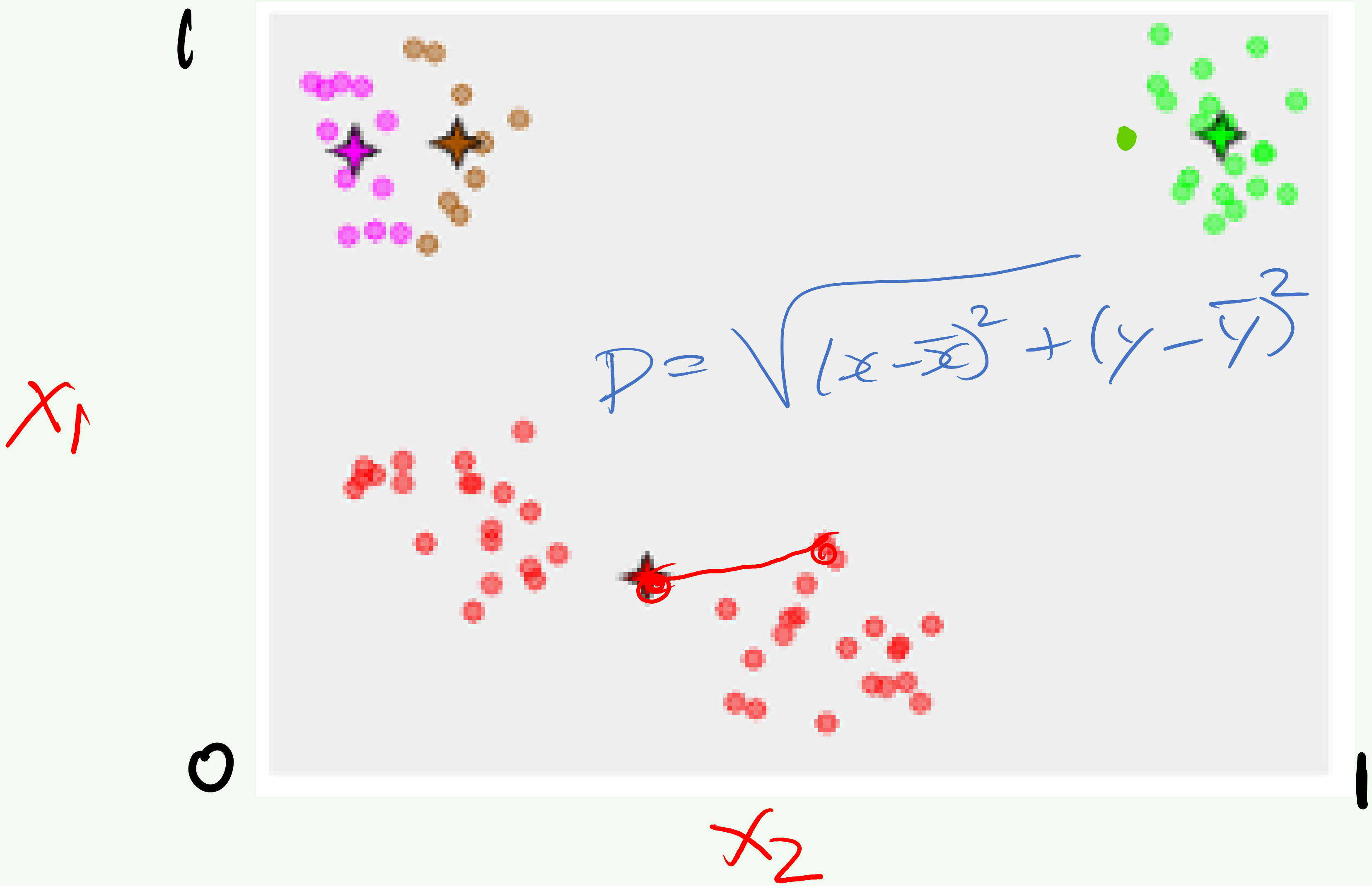
How is distance measured?



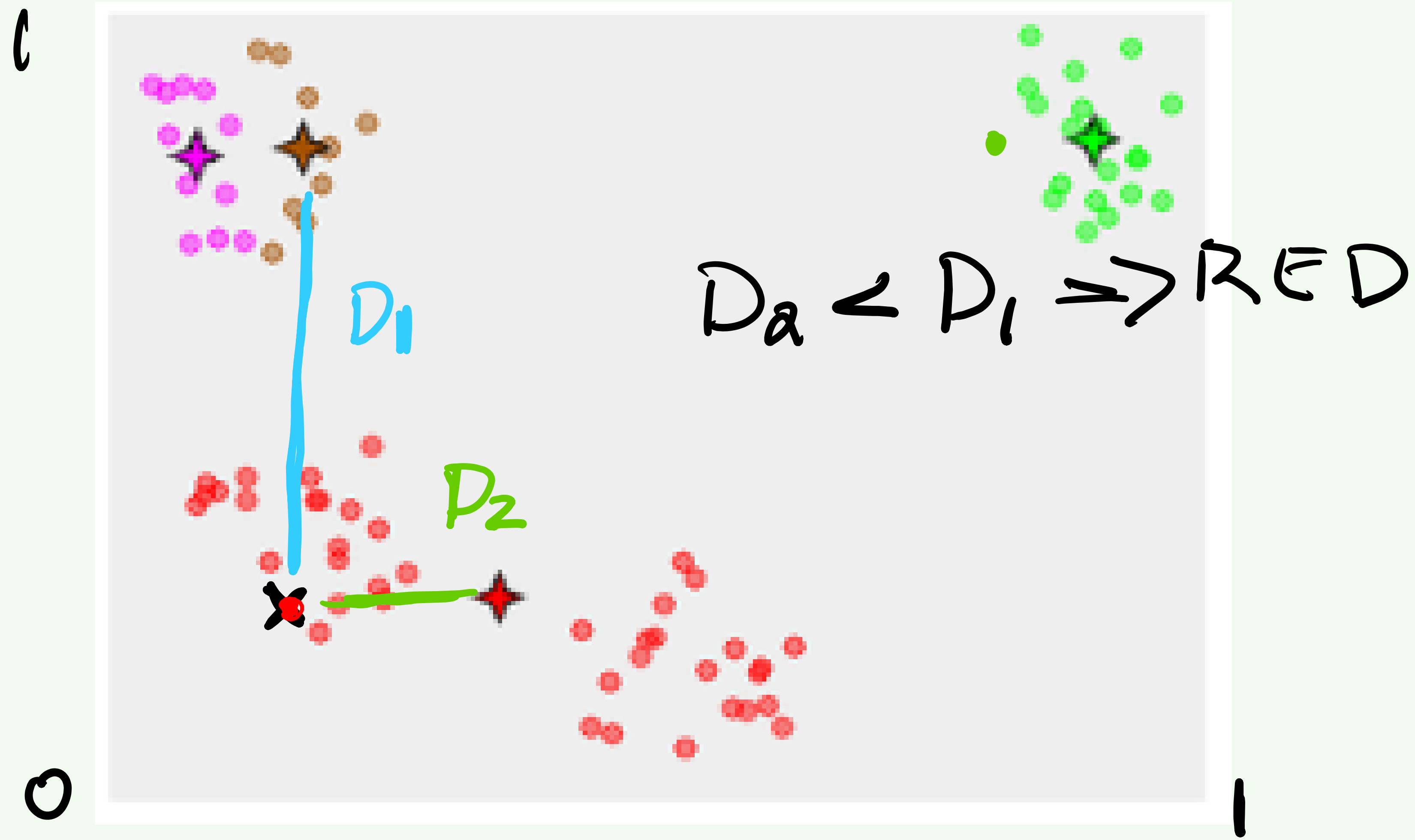
Euclidean distance

$$C = \sqrt{A^2 + B^2}$$

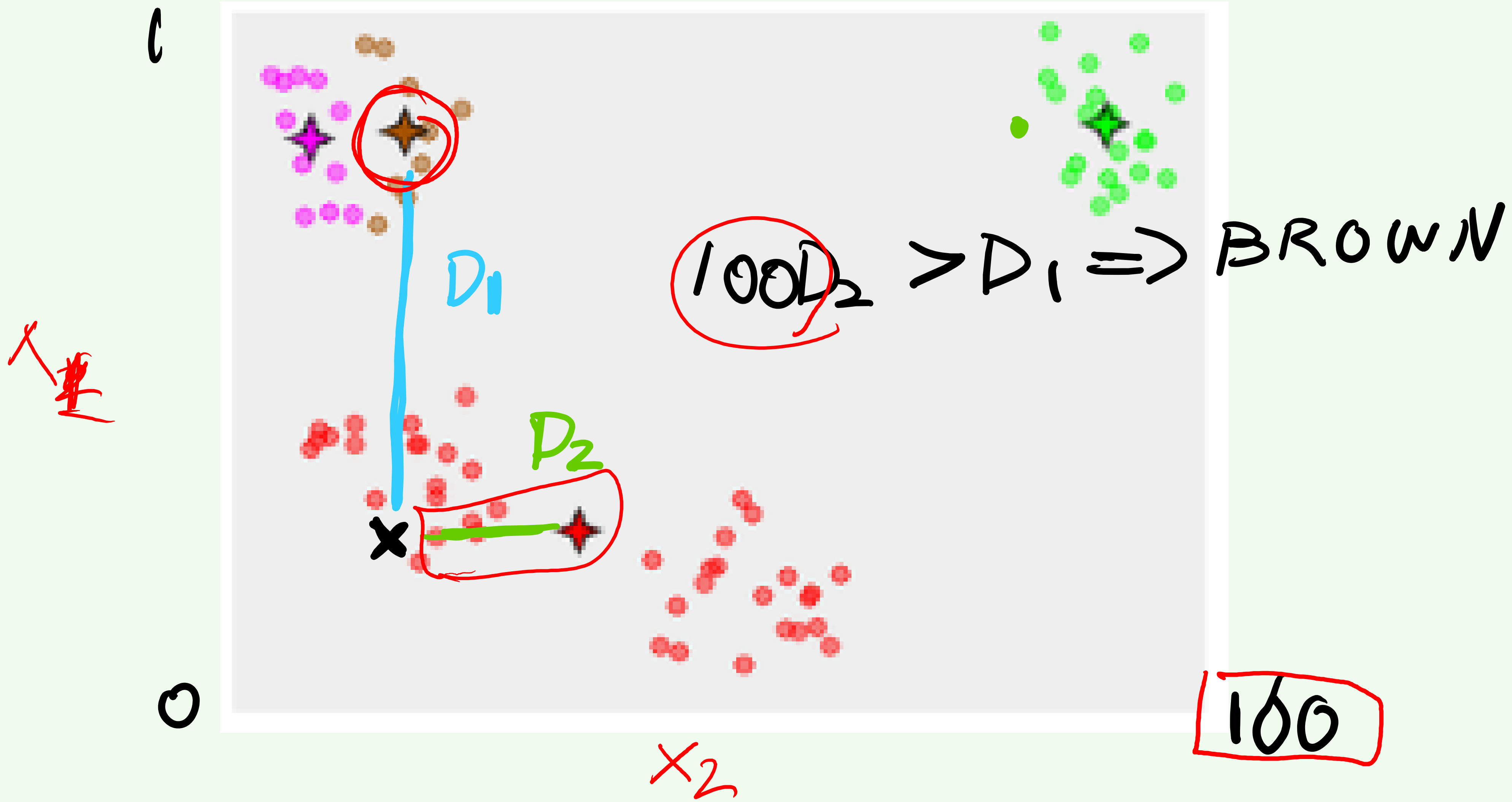
Why is scaling needed?



Why is scaling needed?

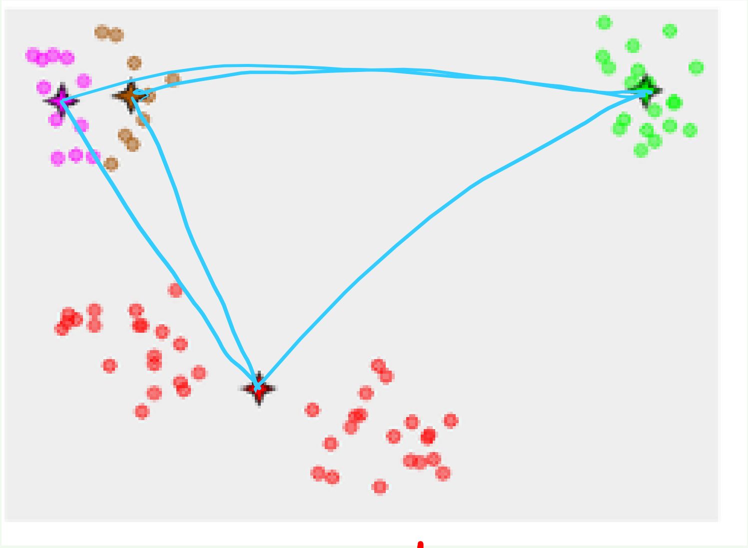


Kmeans



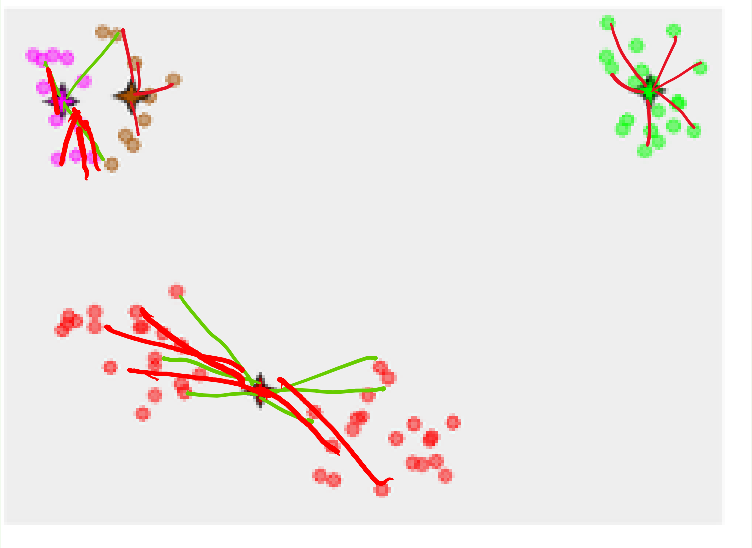
Between Cluster SS or *Within* Cluster SS

Between-Cluster Sum of Squares



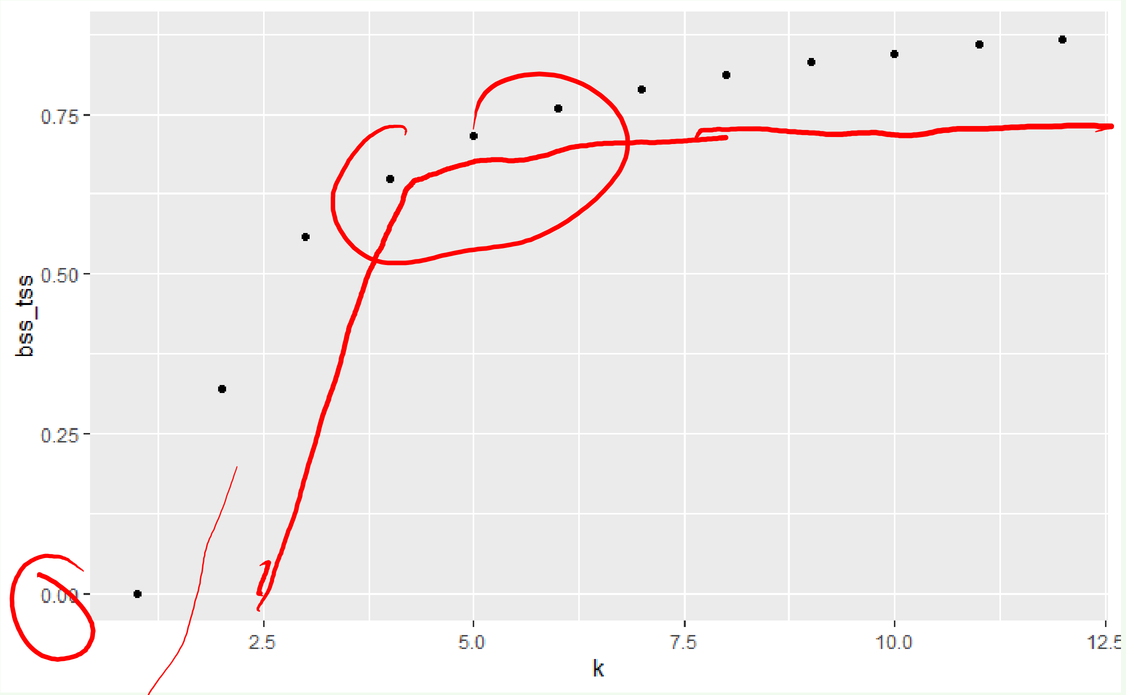
$d_1^2 + d_2^2 + d_3^2 + d_4^2$

Within-Cluster Sum of Squares



Selecting K – The Elbow Method

Between-Cluster Sum of Squares



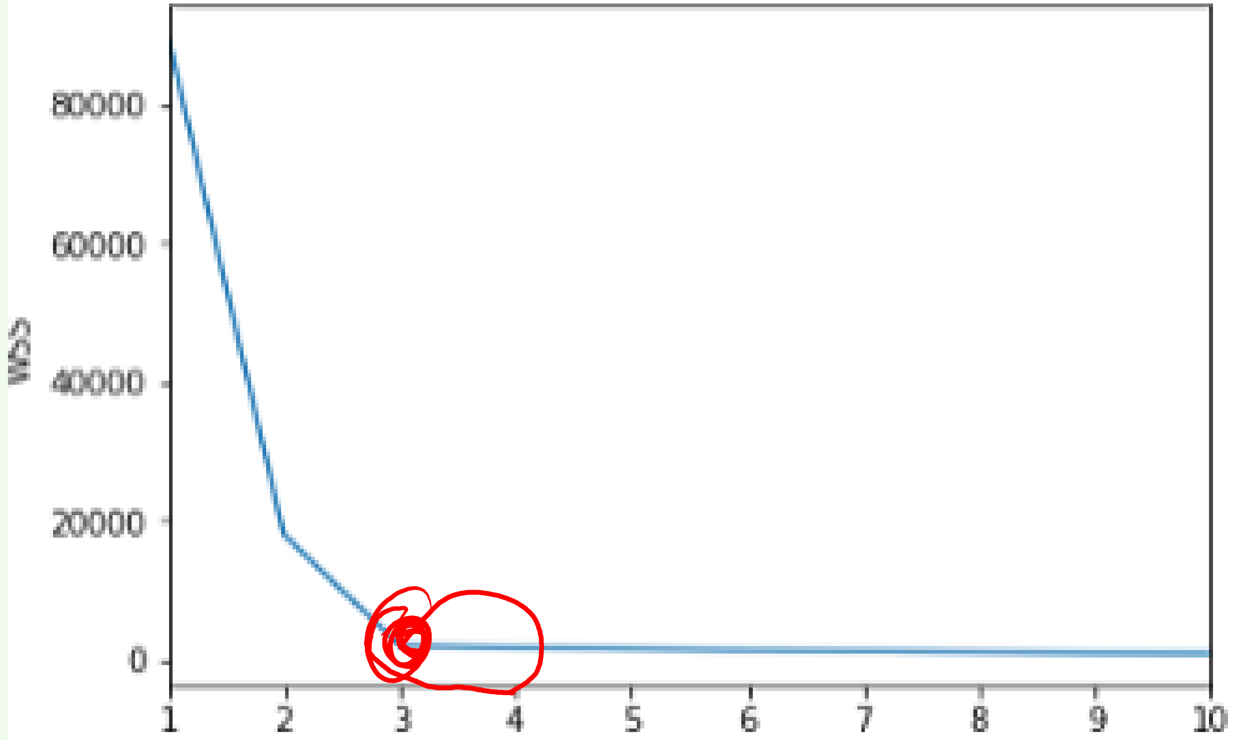
Number of Clusters (K)

K=1

K=n

Selecting K – The Elbow Method

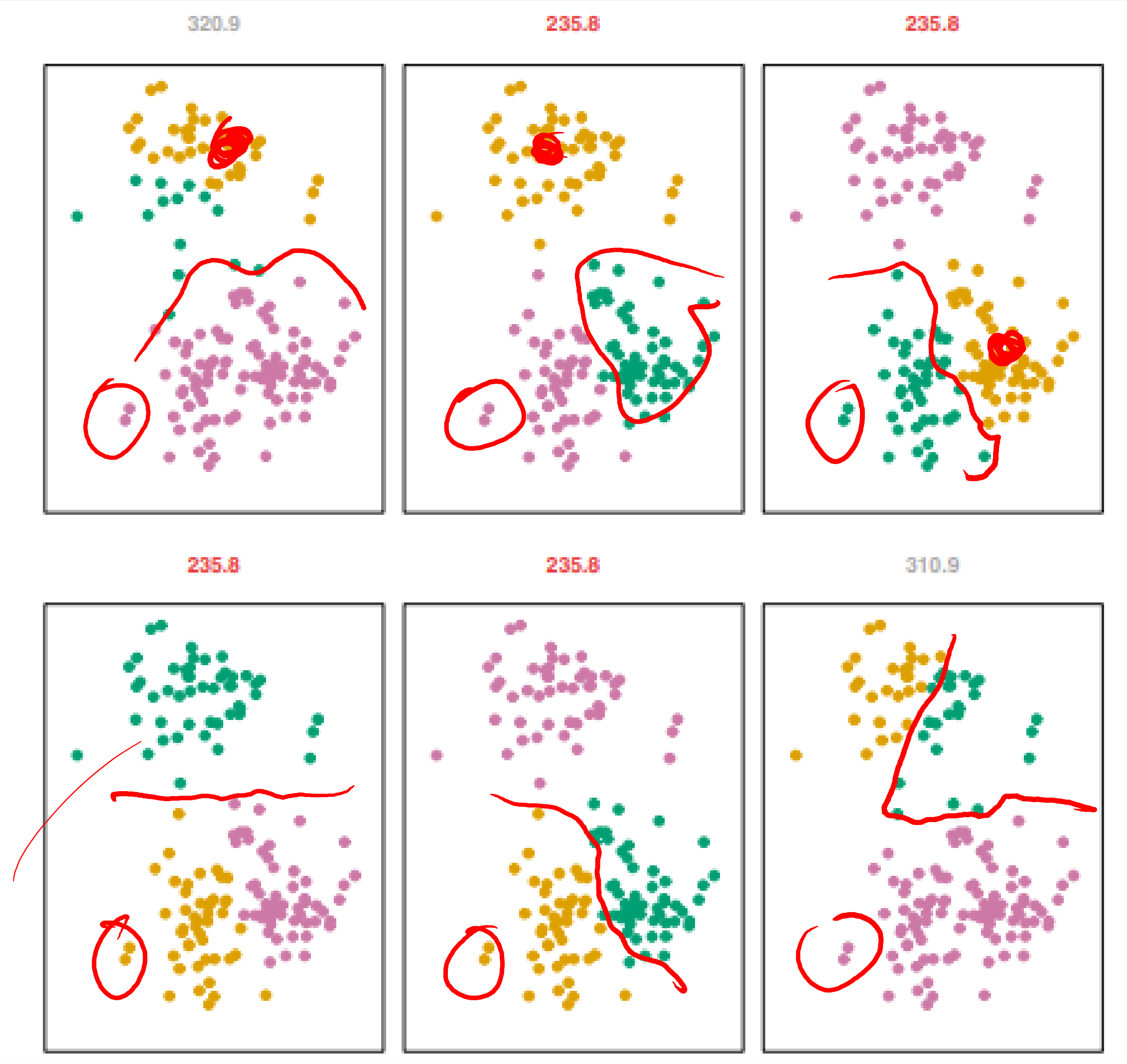
Within-Cluster
Sum of Squares



Number of Clusters (K)

k=n

Number of Starts (n.starts)



$(\bar{x}, \bar{y})$

Hierarchical clustering

Hierarchical Clustering

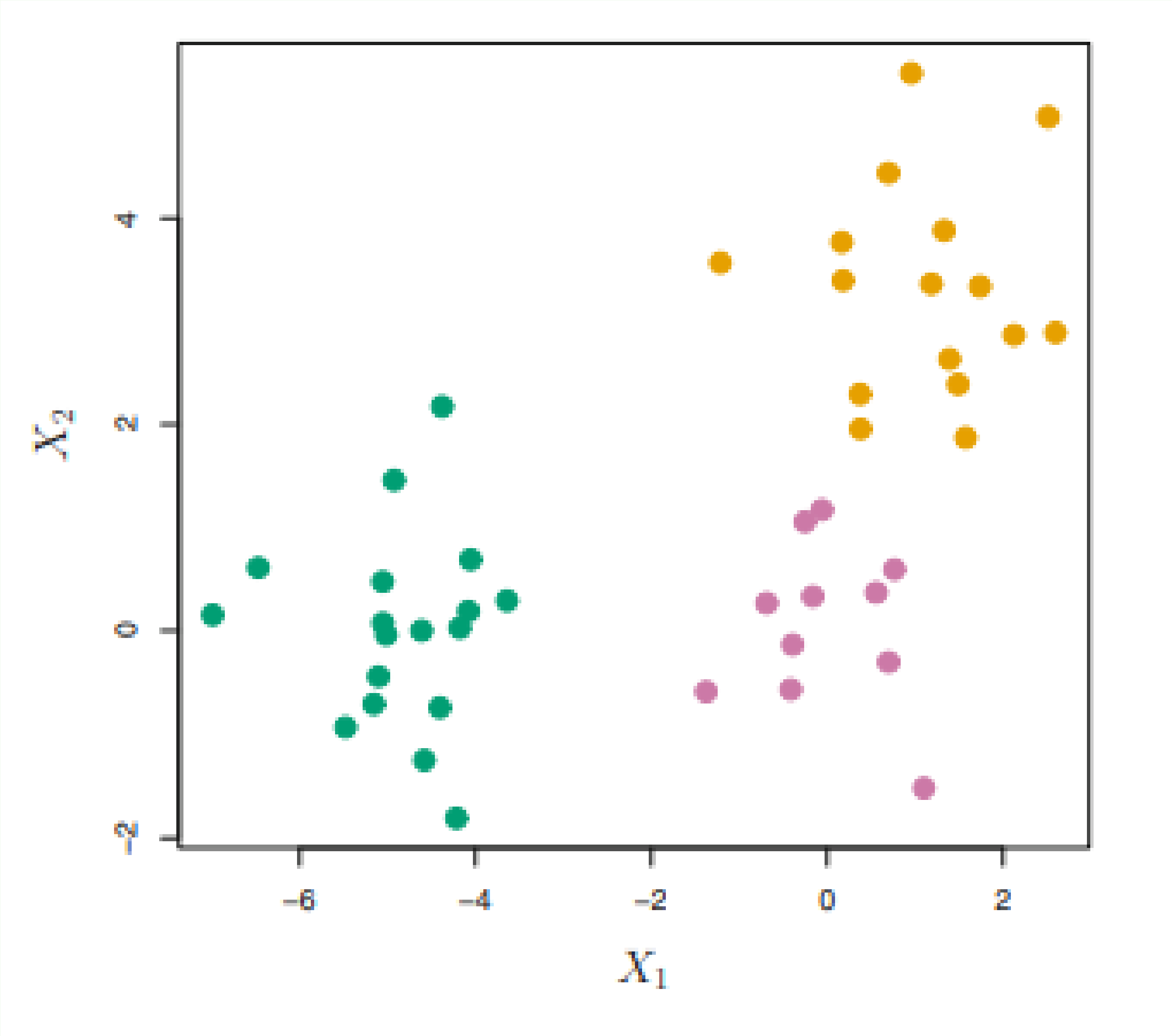
Why use instead of k-means?

- Does not require you to pre-specify the number of clusters
- Creates an easy-to-understand graph called a dendrogram

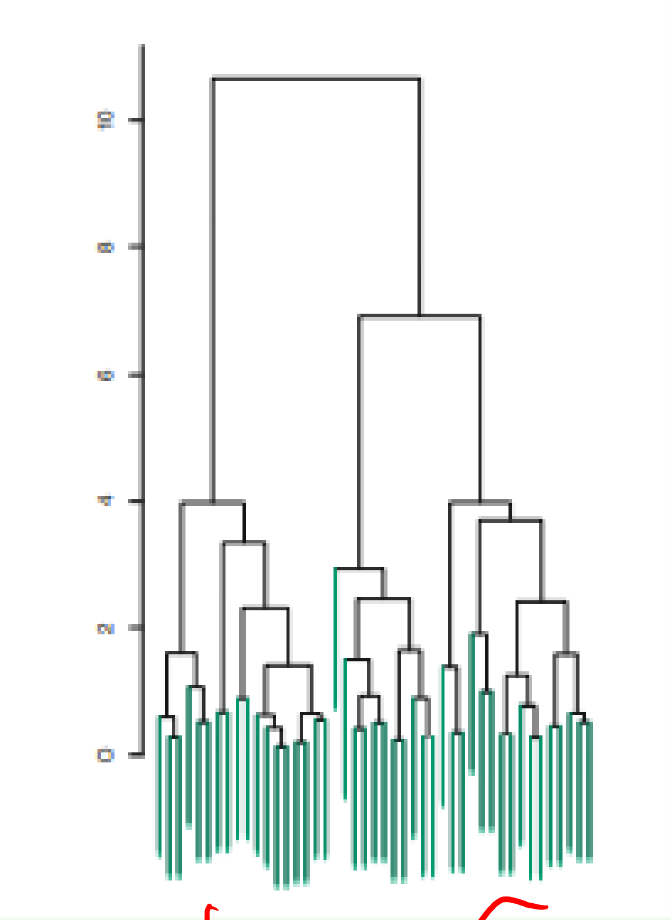
Dendrogram

Color = Target = Unknown / Not Used in Clustering

Y

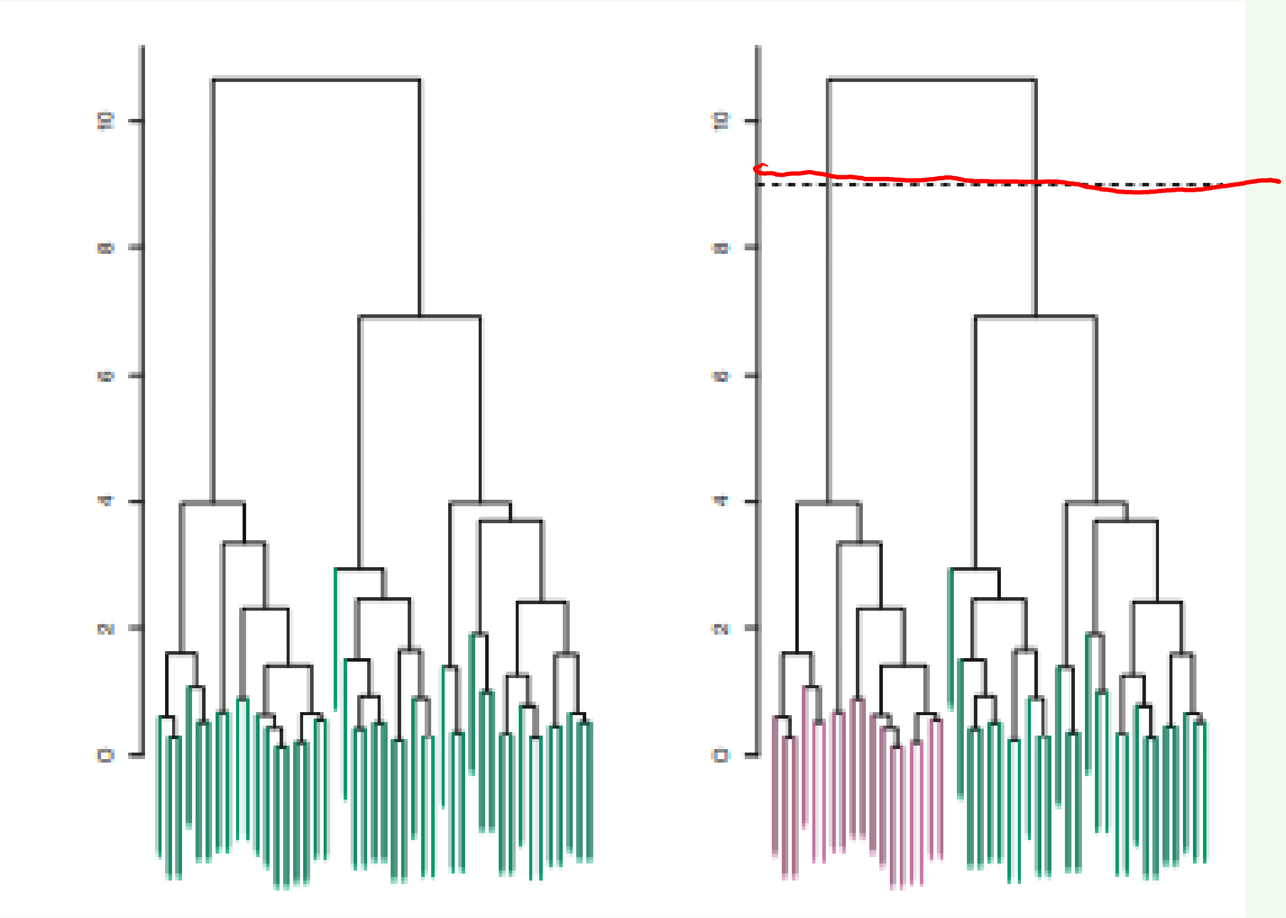


Dendrogram

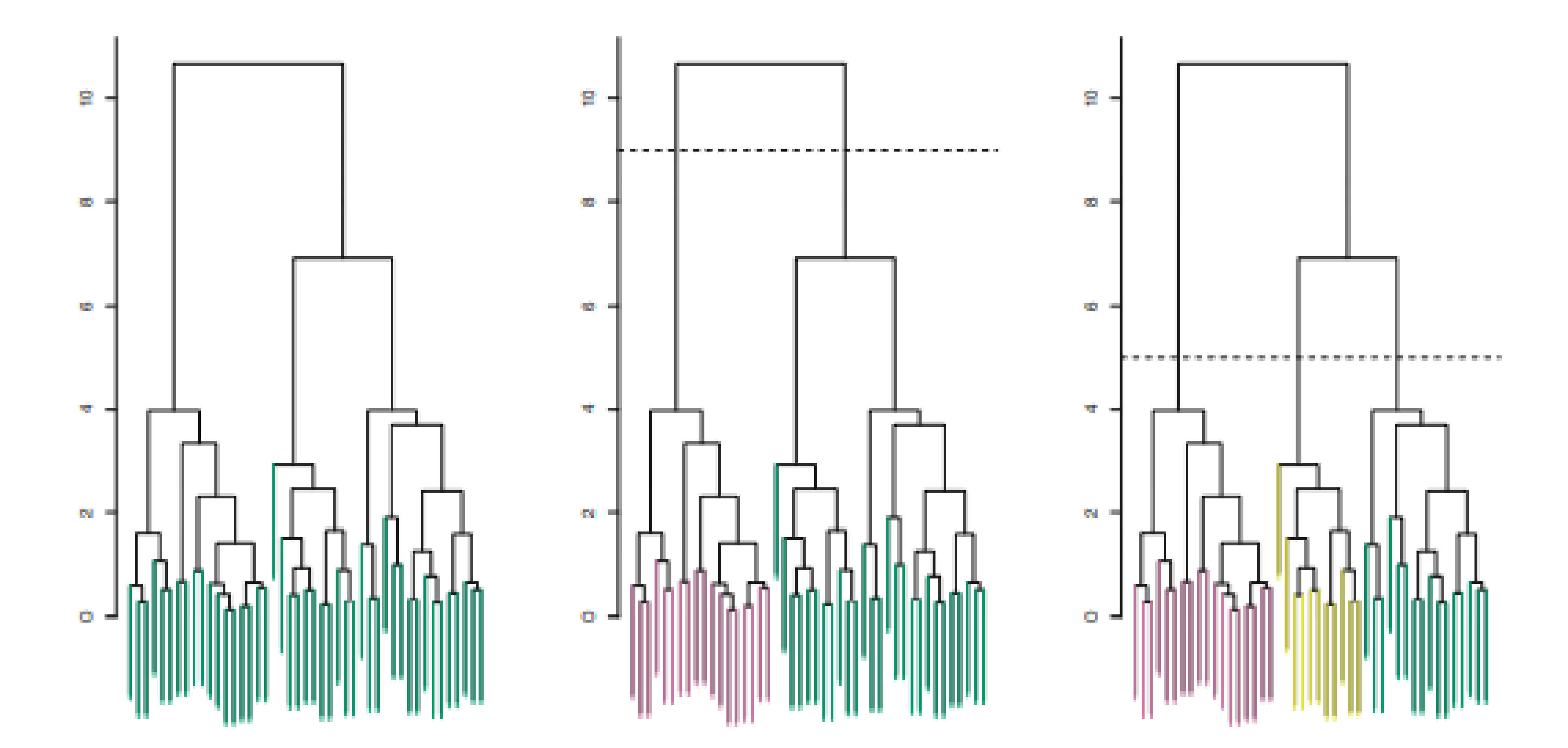


42

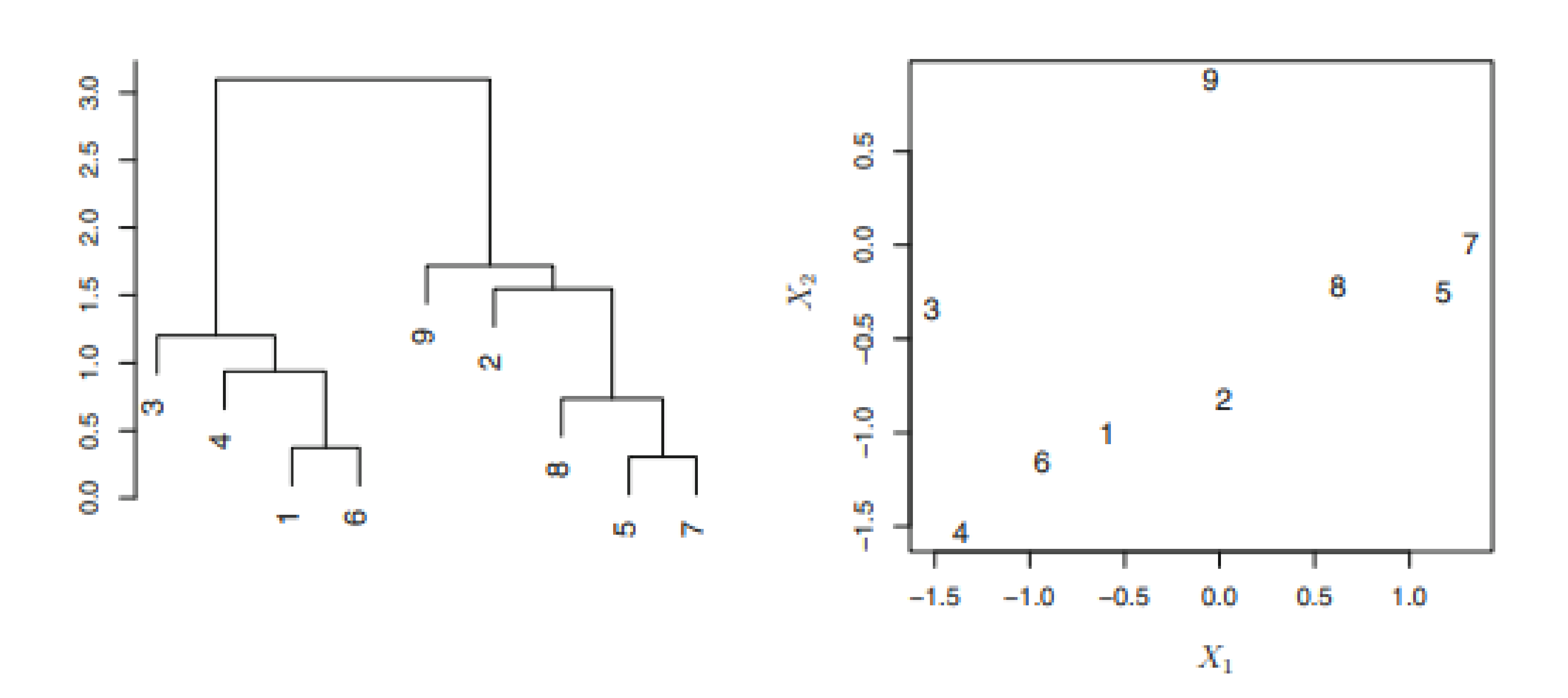
Dendrogram



Dendrogram

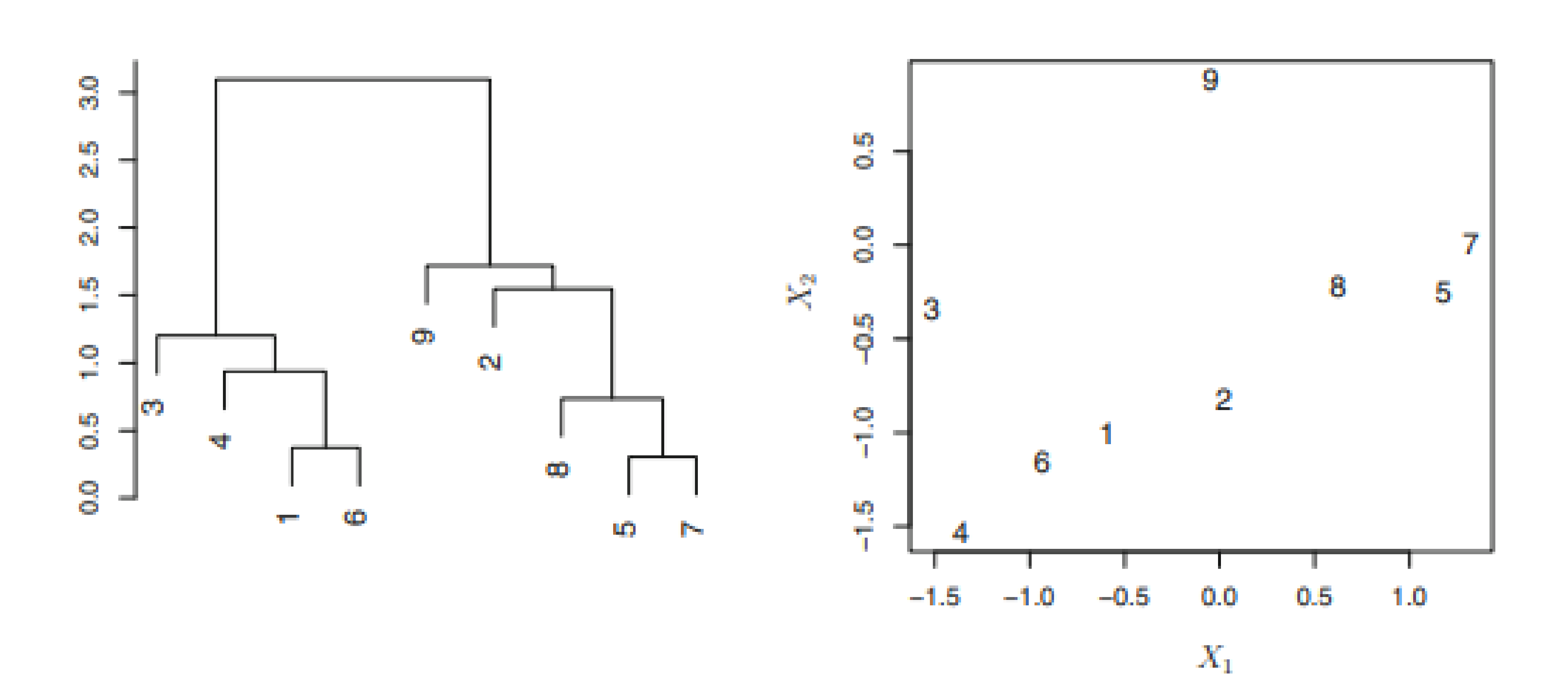


Dendrogram

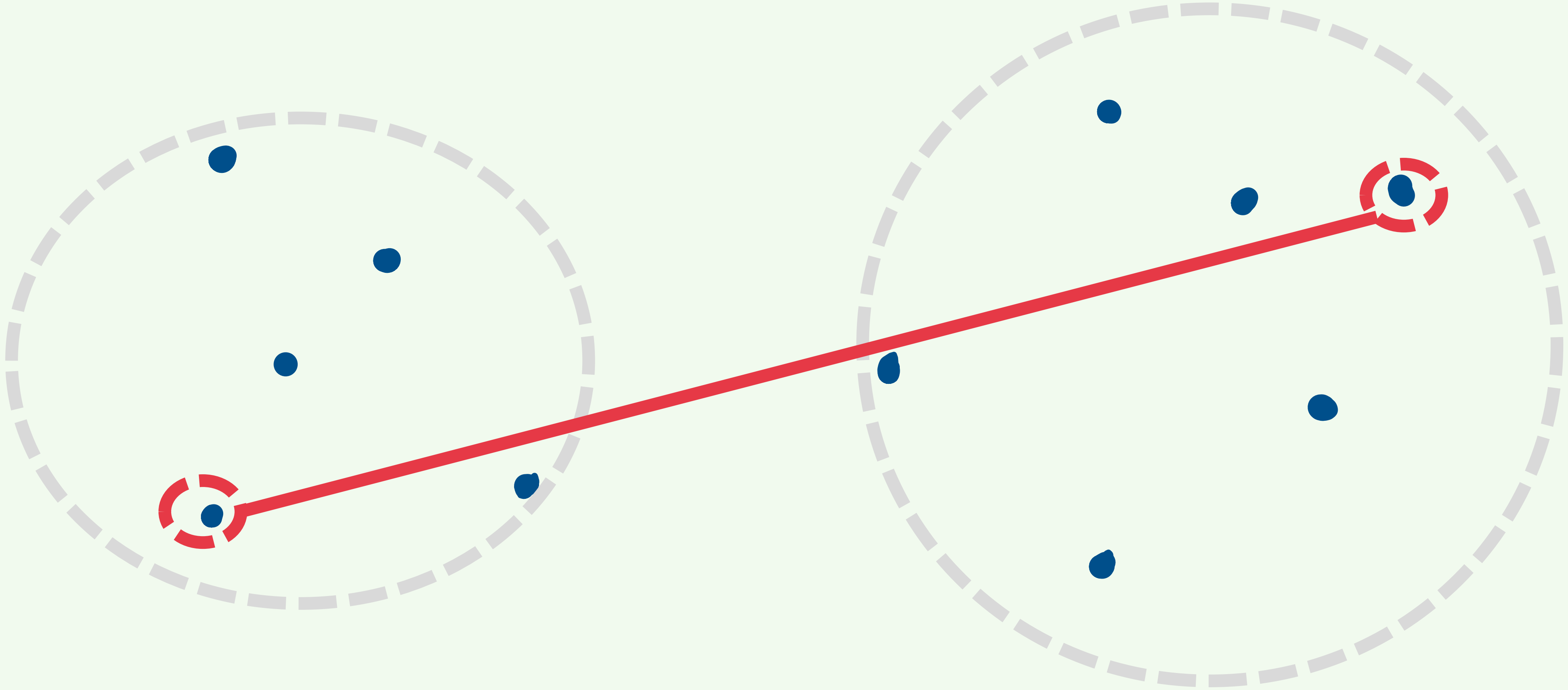


3.0
2.5
2.0
1.5
1.0
0.5
0.0
3
4
1
6
9
2
8
5
7
9
0.5
0.0
−0.5
−1.0
−1.5
X_2
3
8
5
7
2
1
6
4
−1.5
−1.0
−0.5
0.0
0.5
1.0
X_1

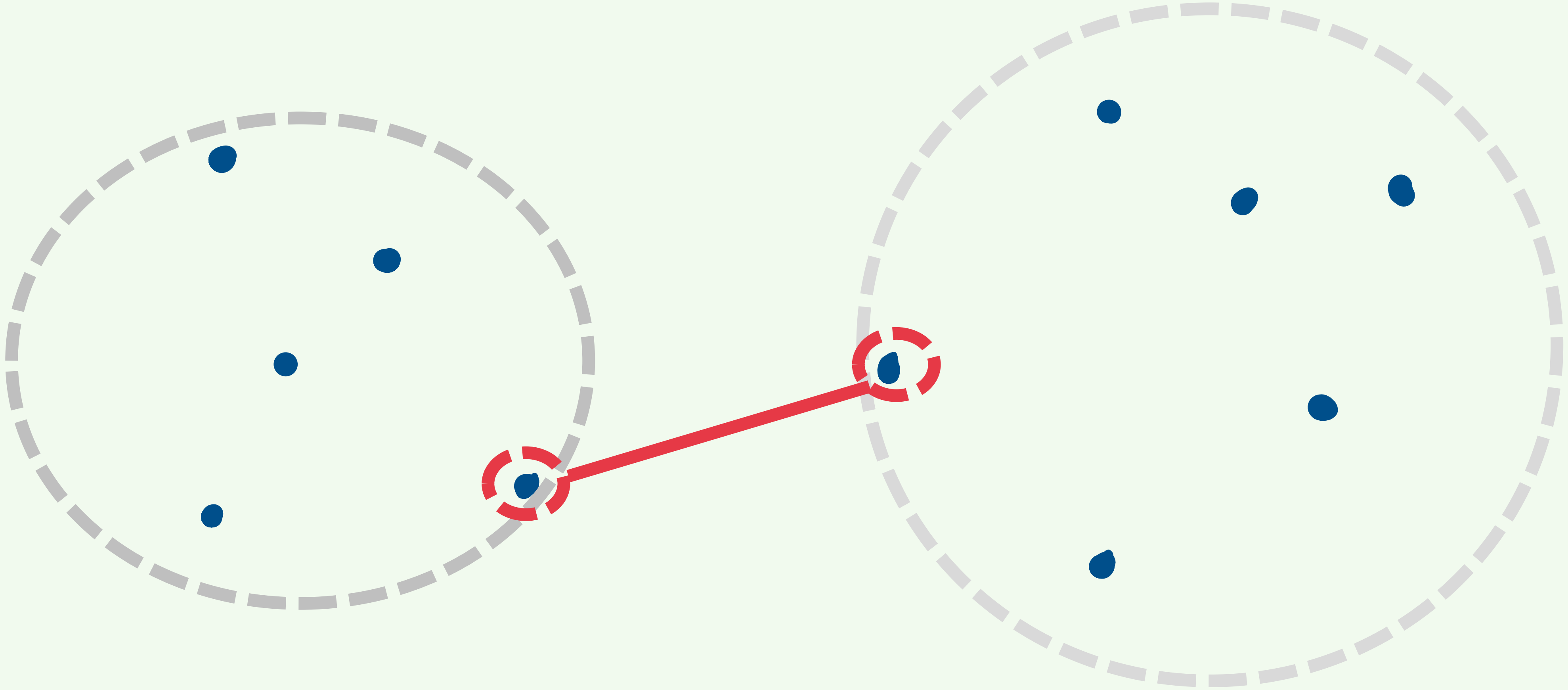
Dendrogram



Linkage types: Complete (Default) / Furthest Distance



Linkage types: Single / Shortest Distance



Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

X1	X2	X3	X4	X5

PCA

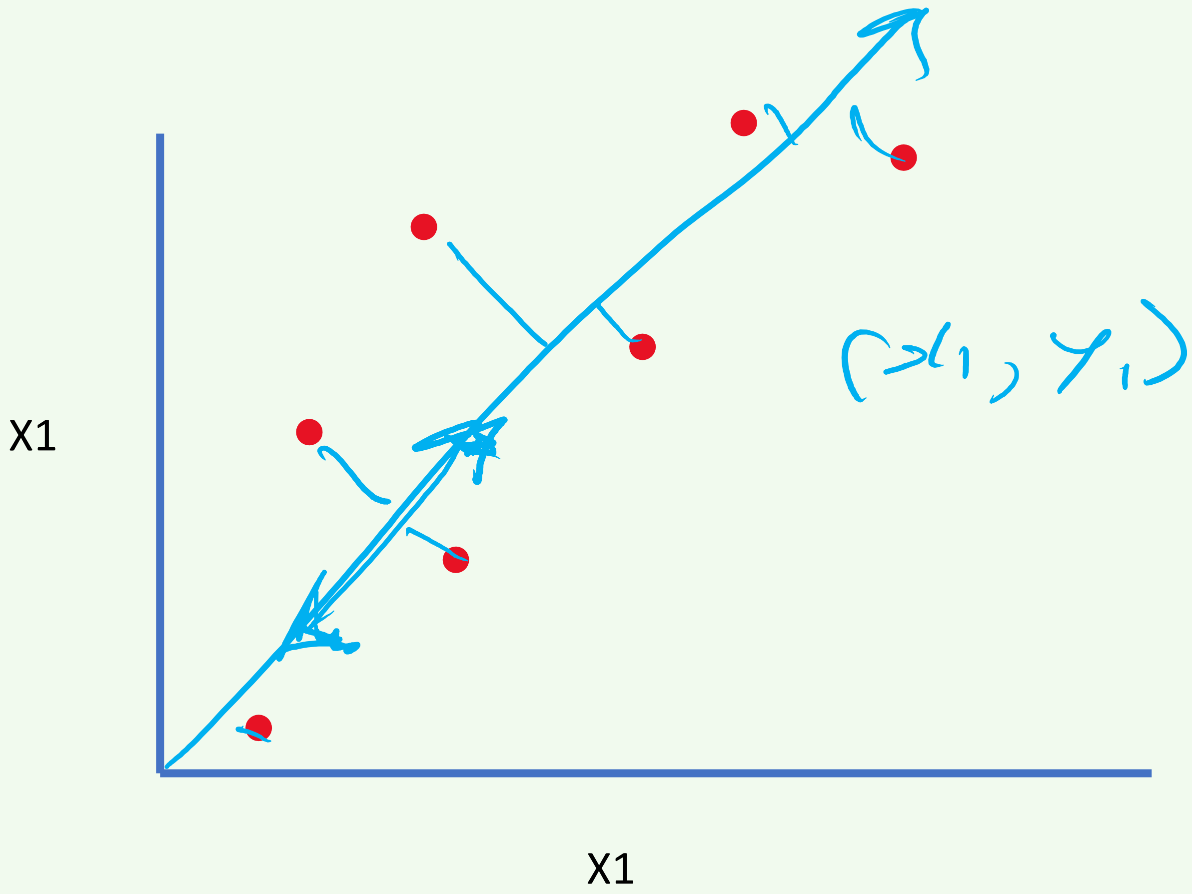
PC1	PC2	PC3	PC4	PC5

Variables have been centered

- Mean 0
- Variance 1

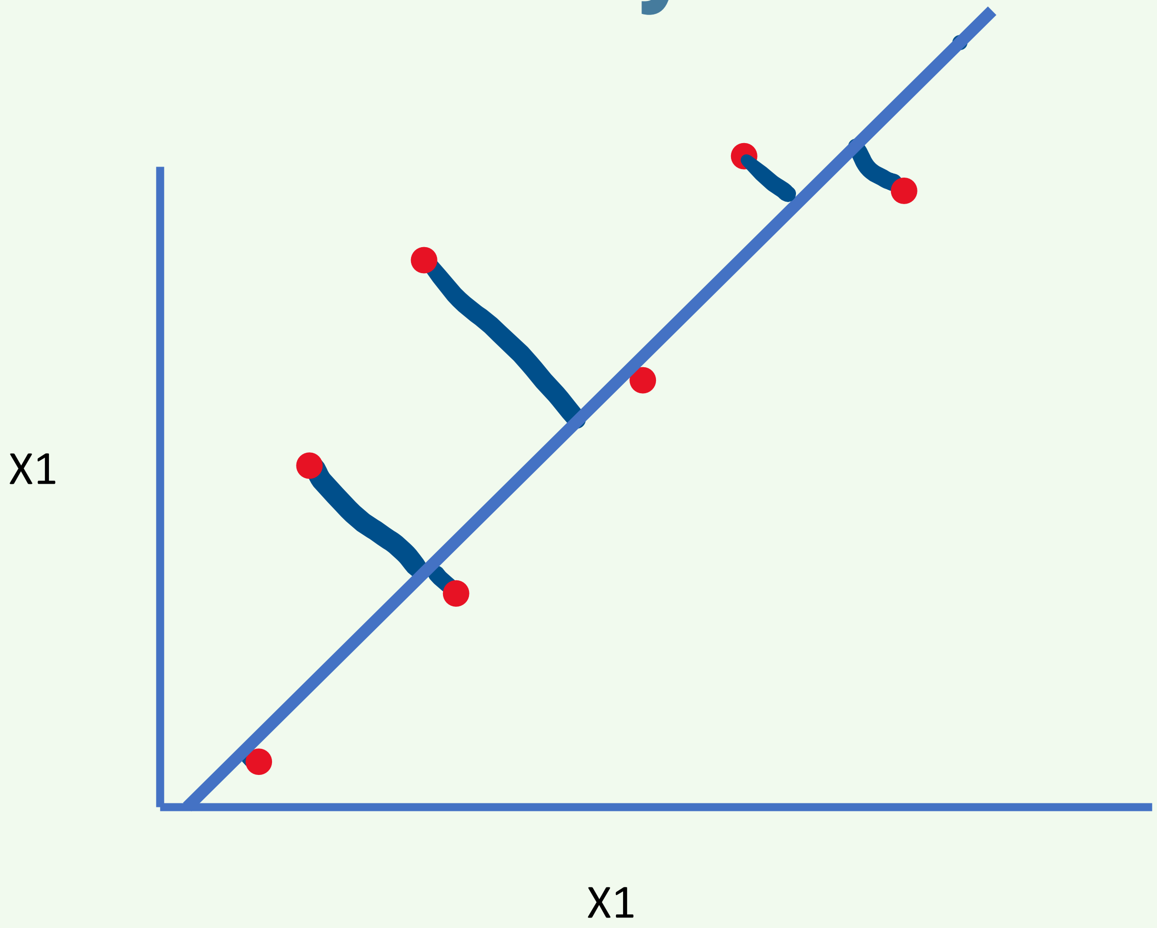
$$\frac{x-\mu}{\sigma}$$

Dimensionality Reduction



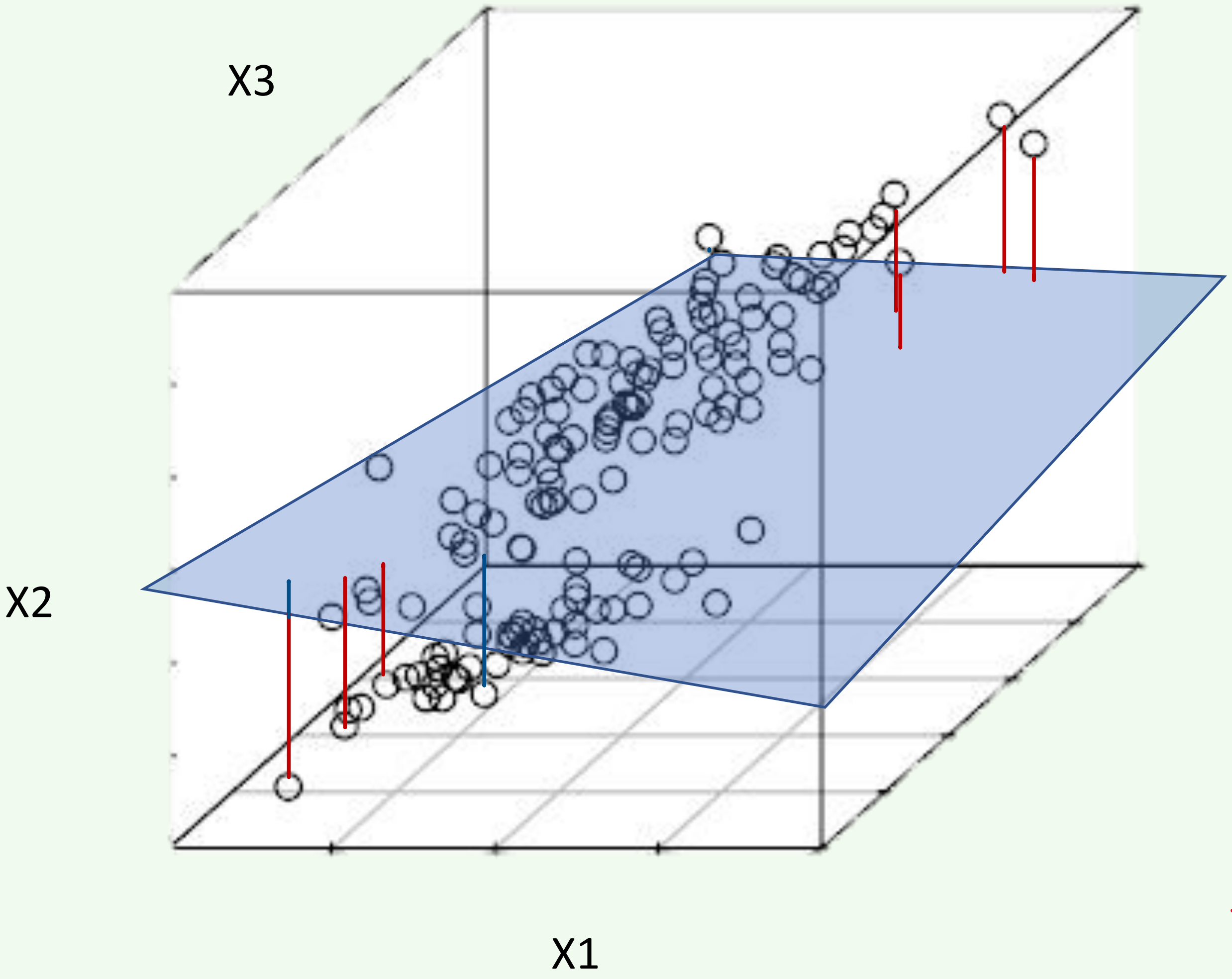
Source: Lecture 14.4 — Dimensionality Reduction | Principal Component Analysis Algorithm — [Andrew Ng]

Dimensionality Reduction



Source: Lecture 14.4 — Dimensionality Reduction | Principal Component Analysis Algorithm — [Andrew Ng]

Dimensionality Reduction



Principal Component Analysis (PCA)

$$PC_1 = 0.2\,X_1 + 0.1\,X_2 + 0.7\,X_3$$

$$\perp$$

$$PC_2 = 0.5X_1 + 0.4\,X_2 + 0.1\,X_3$$

$$\perp$$

$$PC_3 = 0.1X_1 + 0.8\,X_2 + 0.1\,X_3$$

Example: US Arrests

	Murder <dbl>	Assault <int>	UrbanPop <int>	Rape <dbl>
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7

50 states

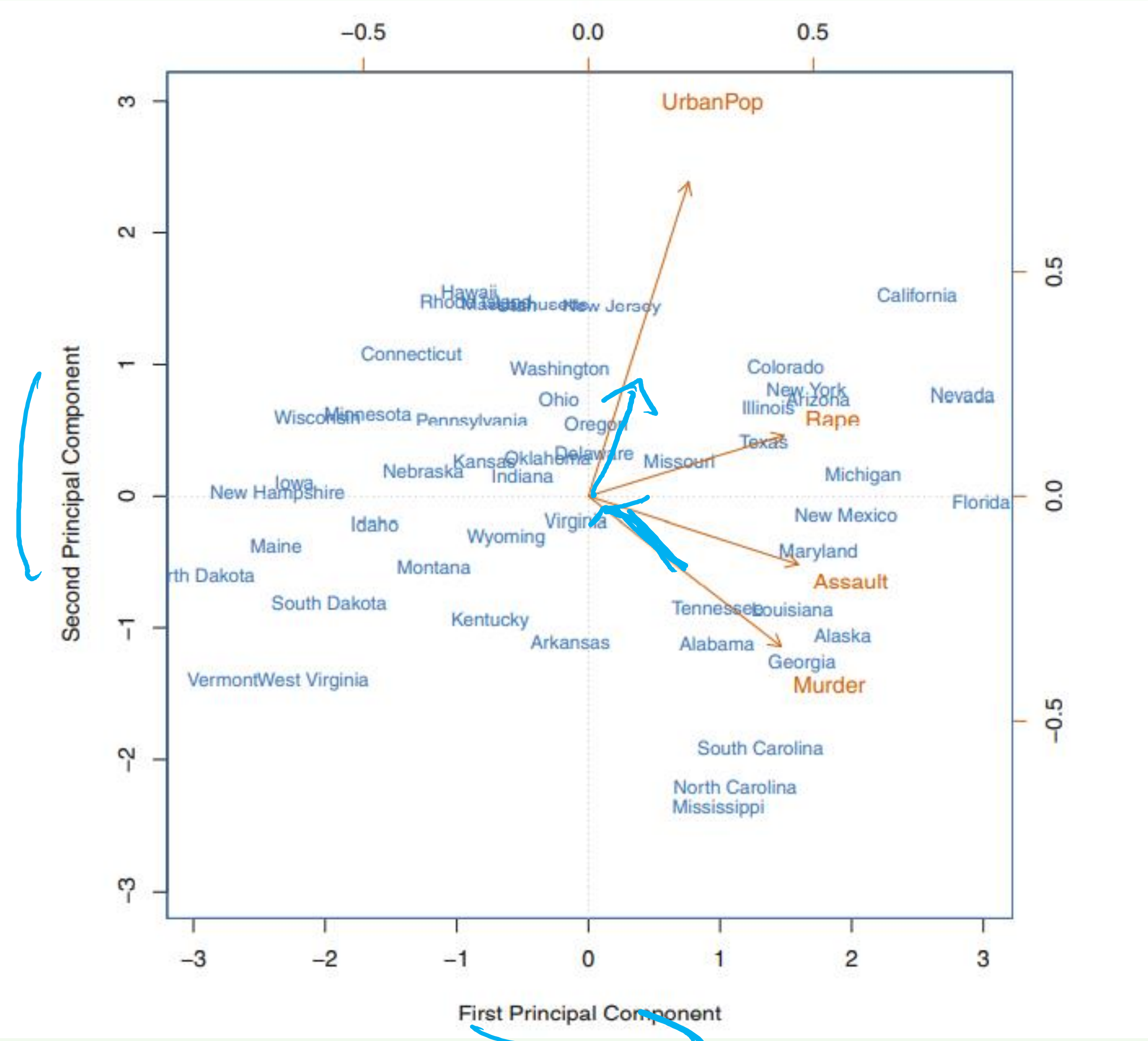
Example: US Arrests

Loadings = Rotations

	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

Biplot

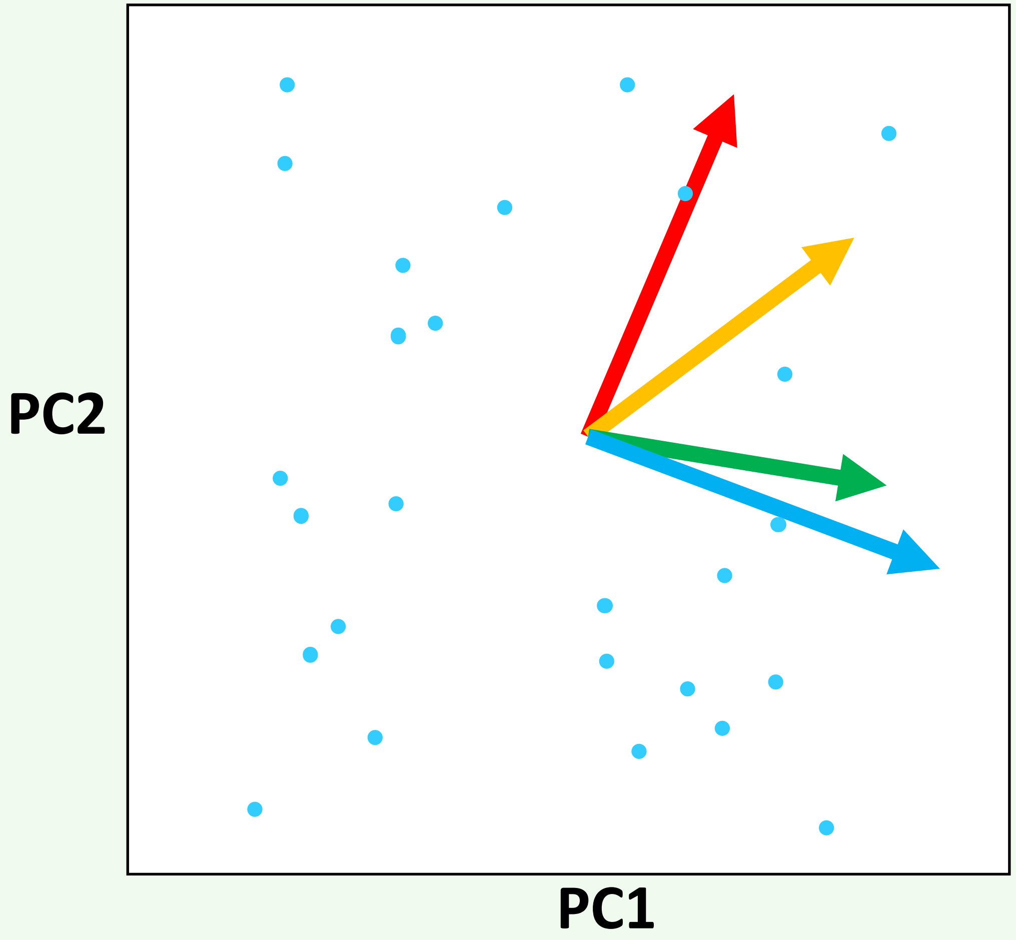
2 PC's



UrbanPop
Rape
Assault
Murder
First Principal Component
Second Principal Component
California
Nevada
Florida
Colorado
New York
Arizona
Illinois
Texas
Michigan
New Mexico
Maryland
Louisiana
Tennessee
Alaska
Alabama
Georgia
South Carolina
North Carolina
Mississippi
Hawaii
New Jersey
Connecticut
Washington
Ohio
Oregon
Pennsylvania
Minnesota
Delaware
Missouri
Oklahoma
Kansas
Indiana
Nebraska
Iowa
New Hampshire
Idaho
Virginia
Wyoming
Maine
Montana
South Dakota
Kentucky
Arkansas
VermontWest Virginia

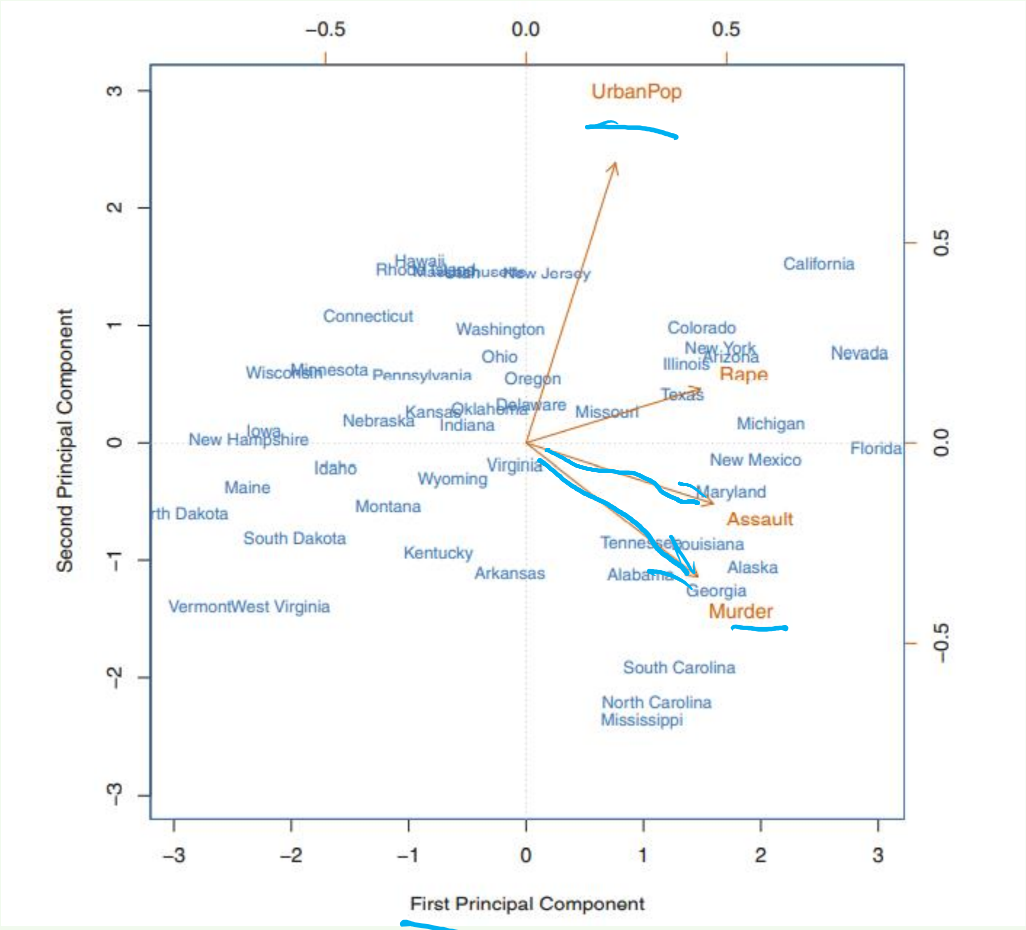
Biplot

	Loadings (Rotations)	
Variable	PC1	PC2
A	0.53	-0.42
B	0.58	-0.19
C	0.28	0.87
D	0.54	0.17



Biplot vs correlation

	Murder	Assault	UrbanPop	Rape
Murder	1.0	0.8	0.1	0.6
Assault	0.8	1.0	0.3	0.7
UrbanPop	0.1	0.3	1.0	0.4
Rape	0.6	0.7	0.4	1.0



Why you need scaling

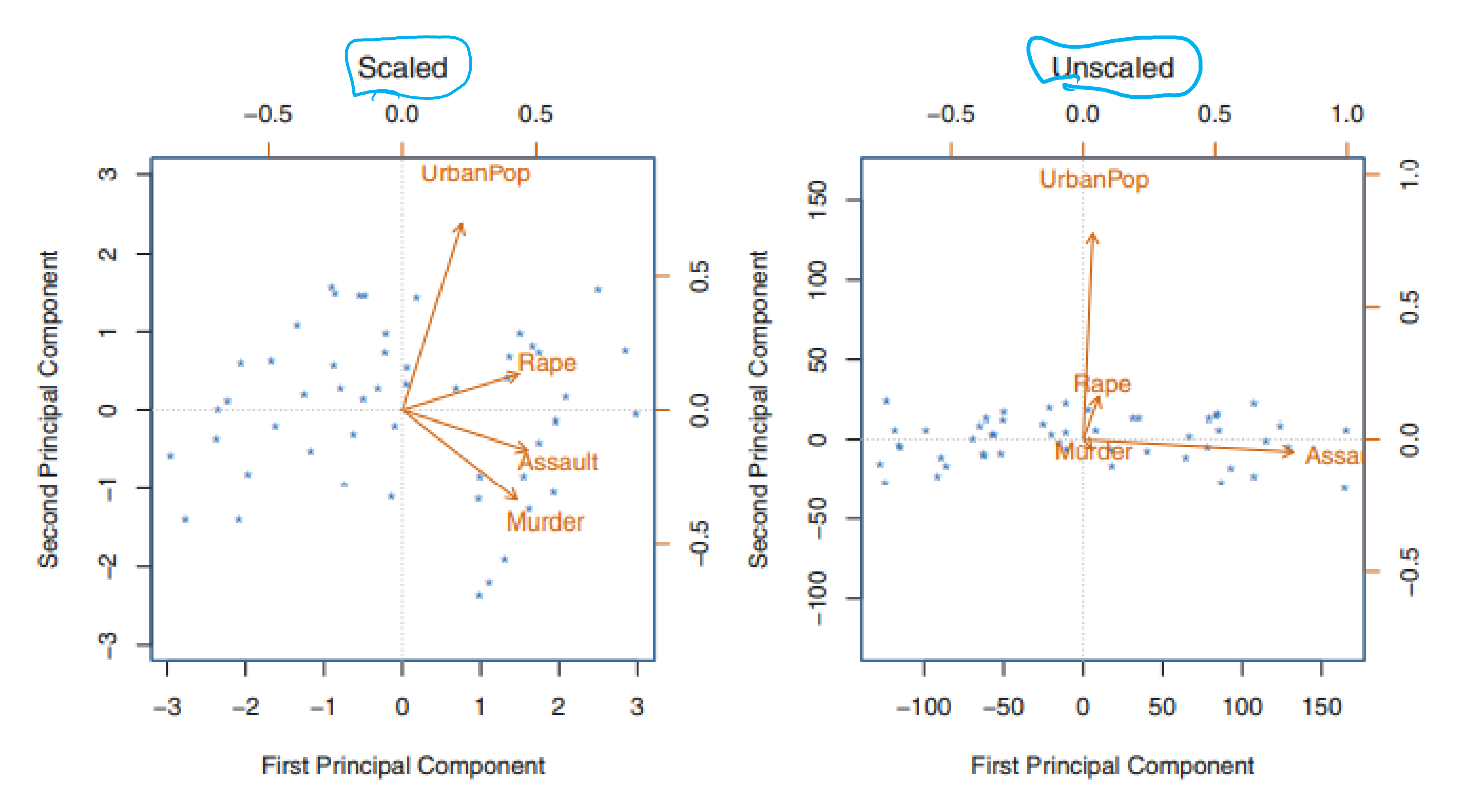
	Units
Murder	Occurrence Per 100,000 People
Assault	Occurrence Per 100,000 People
Rape	Occurrence Per 100,000 People
UrbanPop	% of Population that Lives in Urban Area

0% – 100%

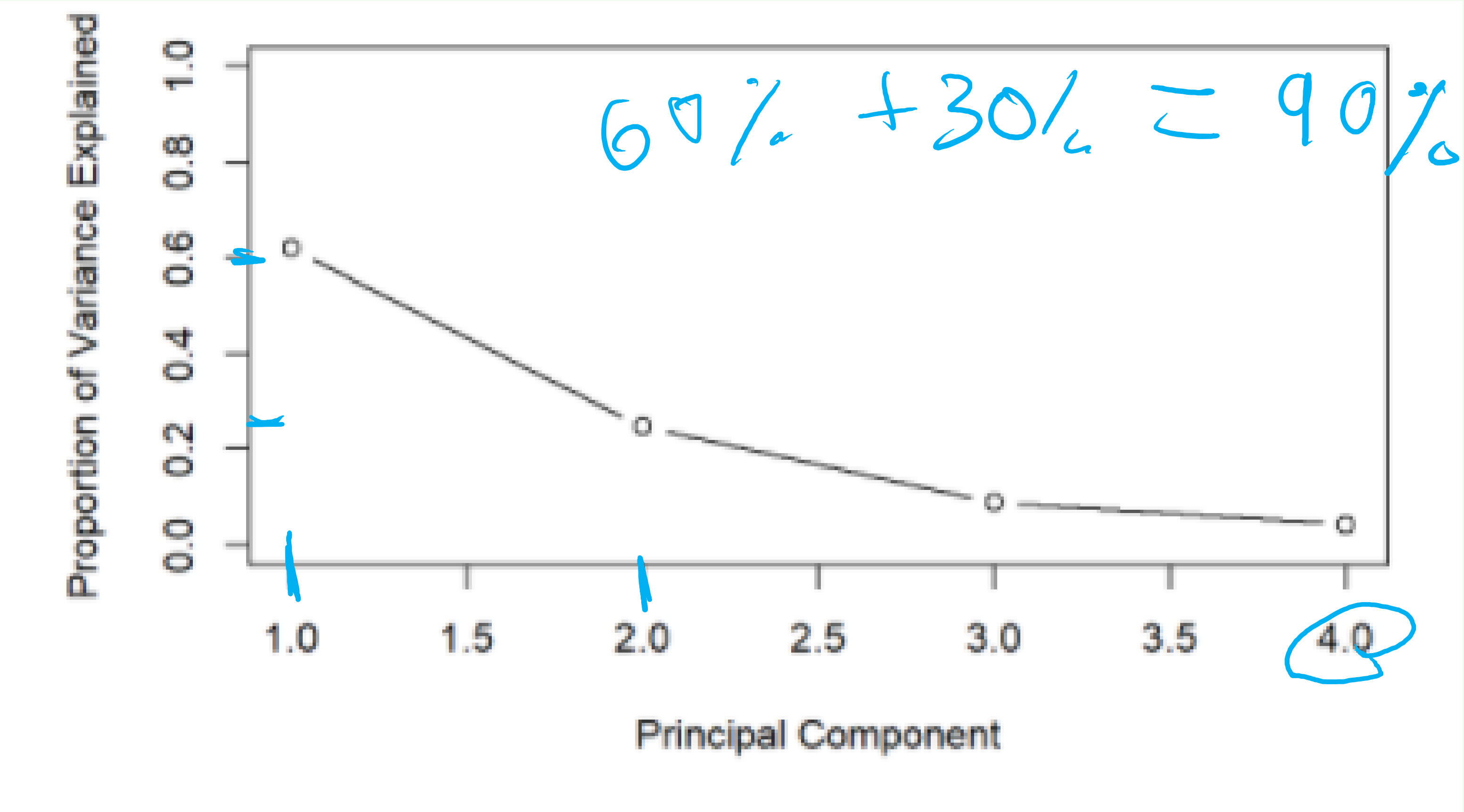
Why you need scaling

	Units	Variance
Murder	Occurrence Per 100,000 People	18.97
Assault	Occurrence Per 100,000 People	87.73
Rape	Occurrence Per 100,000 People	6,949.00
UrbanPop	% of Population that Lives in Urban Area	209.50

Why you need scaling



Why you need scaling



Example: SOA PA 6/13/19 (Traffic Safety), Task 3

$x_1, \cancel{x_2}, \cancel{x_3}$ PC_1

3. (*9 points*) Use observations from principal components analysis (PCA) to generate a new feature

Your assistant has provided code to run a PCA on three variables. Run the code on these three variables. Interpret the output, including the loadings on significant principal components. Generate one new feature based on your observations (which may also involve dropping some current variables). Your assistant has provided some notes on using PCA on factor variables in the Rmd file.

R output (summary)

```
Importance of components:
                         PC1    PC2    PC3    PC4     PC5     PC6     PC7    PC8    PC9    PC10    PC11    PC12
Standard deviation     1.829 1.3740 1.2796 1.2379 1.14429 1.03216 1.01236 1.0033 0.9174 0.79731 0.64583 0.54470
Proportion of Variance 0.223 0.1259 0.1092 0.1022 0.08729 0.07102 0.06833 0.0671 0.0561 0.04238 0.02781 0.01978
Cumulative Proportion  0.223 0.3489 0.4580 0.5602 0.64748 0.71851 0.78683 0.8539 0.9100 0.95241 0.98022 1.00000
                            PC13     PC14      PC15
Standard deviation     1.228e-13 8.07e-14 1.555e-14
Proportion of Variance 0.000e+00 0.00e+00 0.000e+00
Cumulative Proportion  1.000e+00 1.00e+00 1.000e+00
```

"Running PCA on these variables shows that 22% of the variation is explained by the first PC and 35% is explained by using the first two"

R output (rotation or weights)

	PC1	PC2	PC3
Rd_ConditionsDRY	-0.51165971	0.03279495	-0.074984796
Rd_ConditionsICE.SNOW.SLUSH	0.09037524	0.08506534	0.662448145
Rd_ConditionsOTHER	0.05610221	0.18320852	0.103092721
Rd_ConditionsWET	0.49654749	-0.10176327	-0.161823749
LightDARK.LIT	0.11584644	0.52794265	-0.134963861
LightDARK.NOT.LIT	0.05371675	0.19840327	-0.012771256
LightDAWN	0.03037488	0.07312351	0.008834873
LightDAYLIGHT	-0.14979749	-0.66027088	0.122825366
LightDUSK	0.04011811	0.17211754	-0.069299885
LightOTHER	0.03196240	0.20556965	0.097572239
WeatherCLEAR	-0.45856690	0.18940018	-0.043504511
WeatherCLOUDY	0.16796308	-0.22634633	0.028404961
WeatherOTHER	0.05593982	0.14313571	0.095611440
WeatherRAIN	0.43250589	-0.06252514	-0.190678603
WeatherSNOW	0.09667013	0.07063727	0.650123103

R output (rotation or weights)

	PC1	PC2	PC3
Rd_ConditionsDRY	-0.51165971	0.03279495	-0.074984796
Rd_ConditionsICE.SNOW.SLUSH	0.09037524	0.08506534	0.662448145
Rd_ConditionsOTHER	0.05610221	0.18320852	0.103092721
Rd_ConditionsWET	0.49654749	-0.10176327	-0.161823749
LightDARK.LIT	0.11584644	0.52794265	-0.134963861
LightDARK.NOT.LIT	0.05371675	0.19840327	-0.012771256
LightDAWN	0.03037488	0.07312351	0.008834873
LightDAYLIGHT	-0.14979749	-0.66027088	0.122825366
LightDUSK	0.04011811	0.17211754	-0.069299885

PC1 = -0.51(Rd_ConditionsDRY + 0.09(Rd_ConditionsICE.SNOW.SLUSH + 0.056(Rd_ConditionsOTHER) + 0.50(Rd_ConditionsWET) + …

Creating easy-to-interpret features

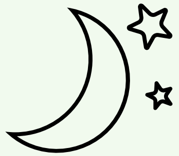
Rainy or Clear

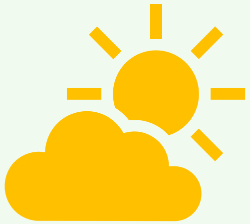
$$-0.51(Rd\_ConditionsDRY) + 0.5(Rd\_ConditionsWET) - 0.46(WeatherCLEAR) + 0.43(WeatherRAIN)$$

Applying these weights creates a variable that is strongly positive for rain/wet conditions and strongly negative for dry/clear conditions. It makes sense to pair up each of these as they would typically appear together, e.g. rain leads to wet roads.

Creating easy-to-interpret features

High or Low Visibility





-0.15(LightDAYLIGHT) + 0.11(LightDARK.LIT) + 0.05(LightDARK.LIT) - 0.46(WeatherCLEAR)

Applying these weights creates a variable that is strongly positive for dark conditions and strongly negative for daylight or lit conditions. It makes sense to pair up each of these as they would typicall y appear together, e.g. clear weather leads to brighter daylight