University of Exeter

Clusters and Similarity

Week 04-BEM2031

Term2: 2024/25

University of Exeter

- Google Notebooks LM for summarising lecture material and content for audio revision.
- There is an optional homework from this week's material – see the ELE Assessments tab. NOTE: this is the same format as your first assignment.
- The website is updated with material from this week, including interactive code for clustering and a gently introduction to 'random'
- Apologies for timetabling making the switch to Monday this week. It happens again in Week 7.

University of Exeter

Today:

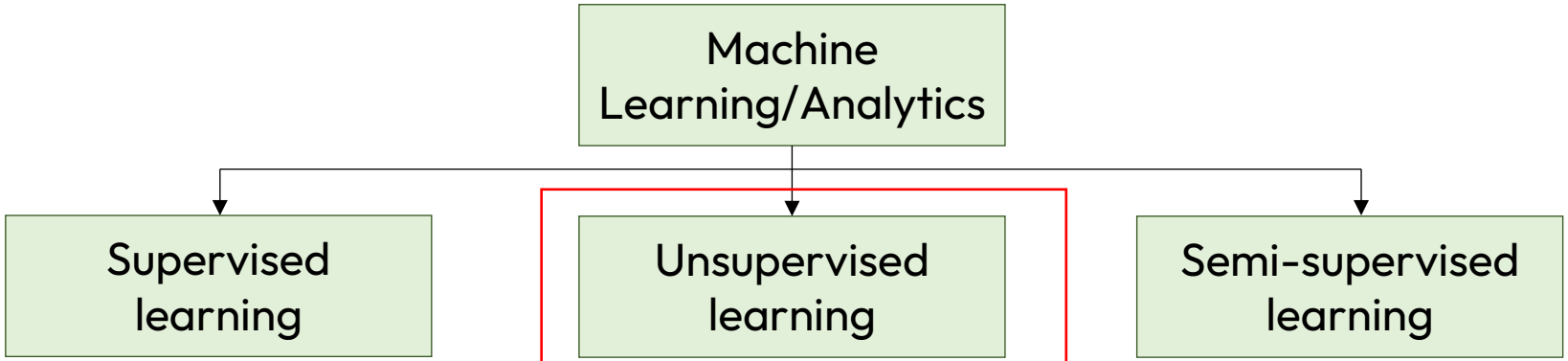
- Understand the spatial interpretation of data
  - Distances and Similarity
- Understand dimension reduction
  - Multi-dimensional scaling (MDS)
  - Principal Component Analysis (PCA)
- Experiment with clustering and interpreting results
  - Hierarchical Clustering
  - k-means Clustering

University of Exeter

Peer reviewed homework:

Clustering 14 February

Supervised vs Unsupervised methods



Supervised methods have a target, an objective.

"Can we find groups of customers who have particularly high likelihoods of cancelling their service soon after their contracts expire?"

Unsupervised methods have no specific target.

"Do our customers naturally fall into different groups?"

Exploratory Data Analysis

Exploratory data analysis

(EDA) largely uses unsupervised methods to examine the structure and patterns within the data.

The objectives of EDA:

Understand data structure:
- Identify variable types and their roles
- Understand the dimensions and structure of your dataset
- Assess completeness eg missing values, missing data, duplicates
- Outlier detection

Summarise the data

Visualise the data:
- Univariate (boxplots, histograms, density plots..)
- Bivariate (scatterplots, boxplots/violin plots..)
- Multivariate (correlation matrices, pair plots..)

Identify patterns/relationships:
Trends, clusters, correlations
Explore interactions between categorical and numerical variables

Prepare for modelling (feature selection, transformations, scaling etc)

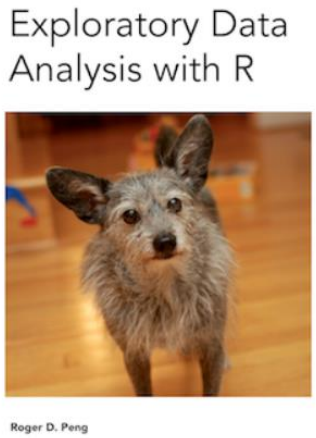
University of Exeter

Exploratory Data Analysis with R

Roger D. Peng

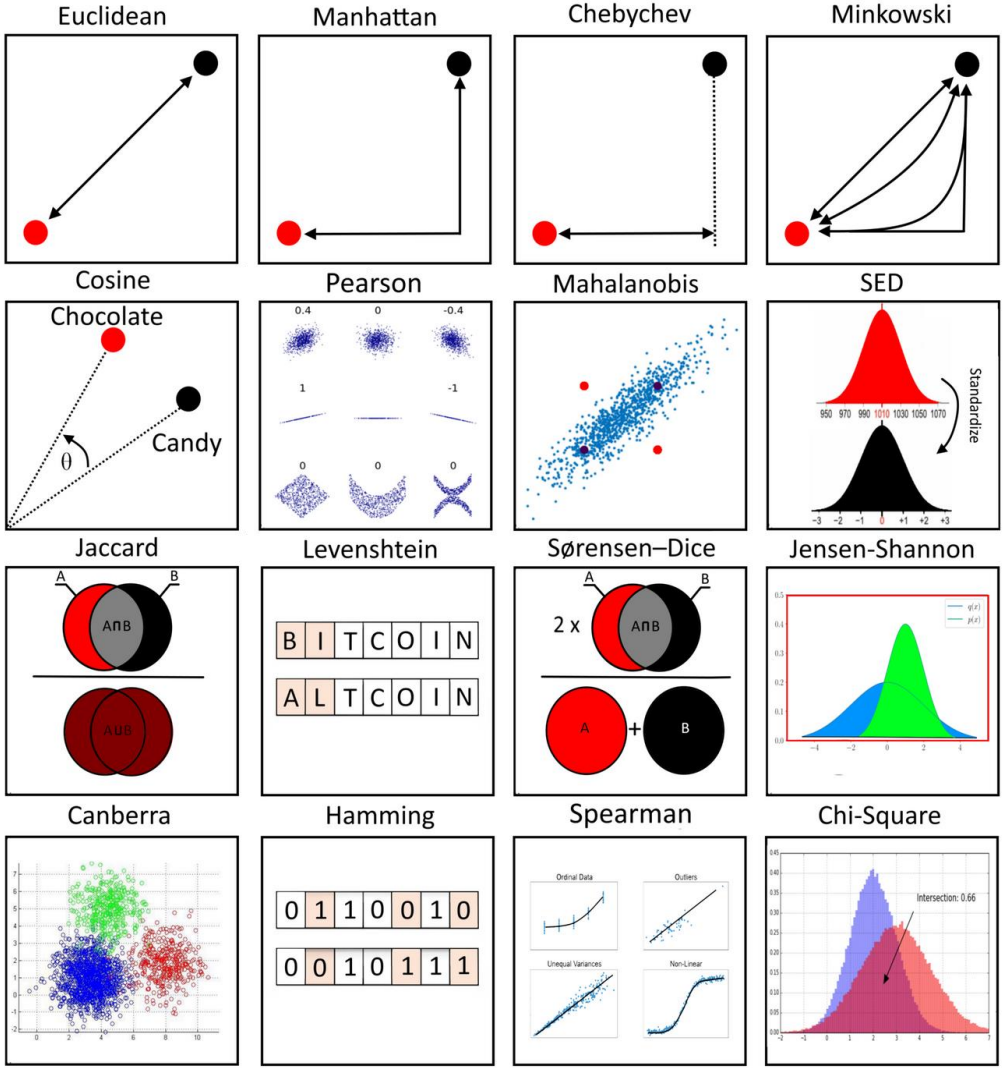
2020-05-01

Welcome



[Exploratory Data Analysis with R (bookdown.org)]

Distance and similarity



University of Exeter

Euclidean Distance

A distance function or metric $d(x, y)$ that tells us how far apart two data points are

- Euclidean Distance

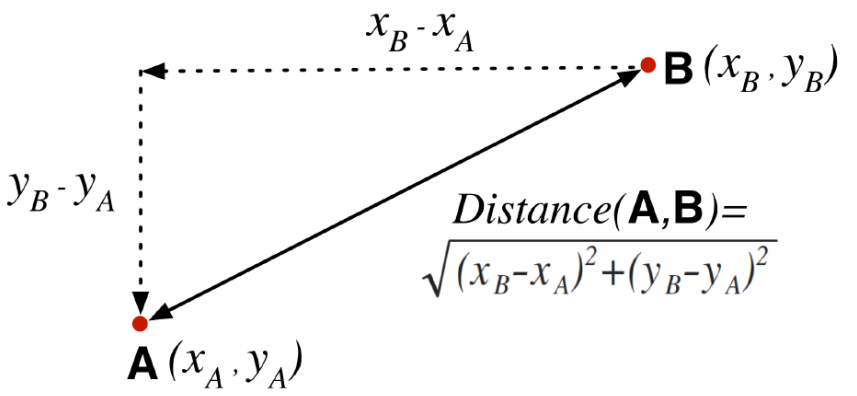


Table 6-1. Nearest neighbor example: Will David respond or not?

Customer	Age	Income (1000s)	Cards	Response (target)	Distance from David
David	*37*	*50*	*2*	*?*	0
John	35	35	3	Yes	$\sqrt{(35-37)^2 + (35-50)^2 + (3-2)^2} = 15.16$
Rachael	22	50	2	No	$\sqrt{(22-37)^2 + (50-50)^2 + (2-2)^2} = 15$
Ruth	63	200	1	No	$\sqrt{(63-37)^2 + (200-50)^2 + (1-2)^2} = 152.23$
Jefferson	59	170	1	No	$\sqrt{(59-37)^2 + (170-50)^2 + (1-2)^2} = 122$
Norah	25	40	4	Yes	$\sqrt{(25-37)^2 + (40-50)^2 + (4-2)^2} = 15.74$

National Student Survey - NSS - Office for Students

Euclidean Distance

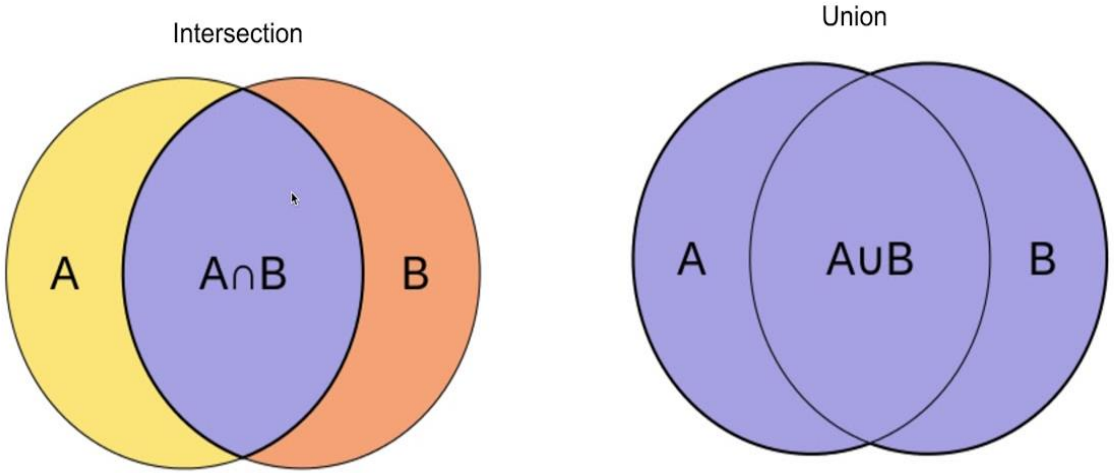
Level of study	Question	Responses	Population	Suppression rea	Option 1	Option 2	Option 3	Option 4	Option 5	This does not ap	Positivity meas
Other undergraduate	Q27: During your studies, how free did you feel to express your ideas, opinions, and beliefs?	15	18		7	7	1	0		2	93.3
All undergraduates	Q01: How good are teaching staff at explaining things?	92	111		37	50	5	0		0	94.6
All undergraduates	Q02: How often do teaching staff make the subject engaging?	92	111		28	43	21	0		0	77.2
All undergraduates	Q03: How often is the course intellectually stimulating?	92	111		23	49	19	1		0	78.3
All undergraduates	Q04: How often does your course challenge you to achieve your best work?	90	111		39	41	8	2		2	88.9
All undergraduates	Theme 1: Teaching on my course	92	111								84.8
All undergraduates	Q05: To what extent have you had the chance to explore ideas and concepts in depth?	90	111		24	44	19	3		2	75.6
All undergraduates	Q06: How well does your course introduce subjects and skills in a way that builds on what you have already learned?	90	111		39	40	8	3		2	87.8
All undergraduates	Q07: To what extent have you had the chance to bring together information and ideas from different topics?	90	111		24	50	14	2		2	82.2
All undergraduates	Q08: To what extent does your course have the right balance of directed and independent study?	90	111		38	35	15	2		2	81.1
All undergraduates	Q09: How well has your course developed your knowledge and skills that you think you will need for your future?	90	111		37	38	11	4		2	83.3
All undergraduates	Theme 2: Learning opportunities	91	111								81.2
All undergraduates	Q10: How clear were the marking criteria used to assess your work?	89	111		34	38	16	1		3	80.9
All undergraduates	Q11: How fair has the marking and assessment been on your course?	90	111		24	54	10	2		2	86.7
All undergraduates	Q12: How well have assessments allowed you to demonstrate what you have learned?	89	111		29	48	10	2		3	86.5
All undergraduates	Q13: How often have you received assessment feedback on time?	90	111		29	36	19	6		2	72.2
All undergraduates	Q14: How often does feedback help you to improve your work?	90	111		36	42	9	3		2	86.7
All undergraduates	Theme 3: Assessment and feedback	90	111								82.4
All undergraduates	Q15: How easy was it to contact teaching staff when you needed to?	90	111		30	38	15	7		2	75.6
All undergraduates	Q16: How well have teaching staff supported your learning?	90	111		38	36	15	1		2	82.2
All undergraduates	Theme 4: Academic support	90	111								78.9
All undergraduates	Q17: How well organised is your course?	90	111		27	33	20	10		2	66.7
All undergraduates	Q18: How well were any changes to teaching on your course communicated?	89	111		24	37	17	11		3	68.5
All undergraduates	Theme 5: Organisation and management	90	111								67.8
All undergraduates	Q19: How well have the IT resources and facilities supported your learning?	86	111		31	36	12	7		6	77.9
All undergraduates	Q20: How well have the library resources (e.g., books, online services and learning spaces) supported your learning?	86	111		32	44	8	2		6	88.4
All undergraduates	Q21: How easy is it to access subject specific resources (e.g., equipment, facilities, software) when you need them?	87	111		24	48	13	2		5	82.8
All undergraduates	Theme 6: Learning resources	89	111								83.1
All undergraduates	Q22: To what extent do you get the right opportunities to give feedback on your course?	90	111		24	44	19	3		2	75.6
All undergraduates	Q23: To what extent are students' opinions about the course valued by staff?	89	111		30	42	14	3		3	80.9
All undergraduates	Q24: How clear is it that students' feedback on the course is acted on?	89	111		28	31	24	6		3	66.3
All undergraduates	Theme 7: Student voice	90	111								74.3
All undergraduates	Q25: How well does the students' union (association or guild) represent students' academic interests?	76	111		19	40	10	7		16	77.6
All undergraduates	Q26: How well communicated was information about your university/college's mental wellbeing support services?	84	111		35	28	18	3		8	75

The single csv file will be too large to open in Excel, which is why we have also

Jaccard Distance

A distance function or metric $d(x, y)$ that tells us how far apart two data points are

- Jaccard Similarity / Distance
- Jaccard Distance = 1 – Jaccard similarity



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

e.g.:

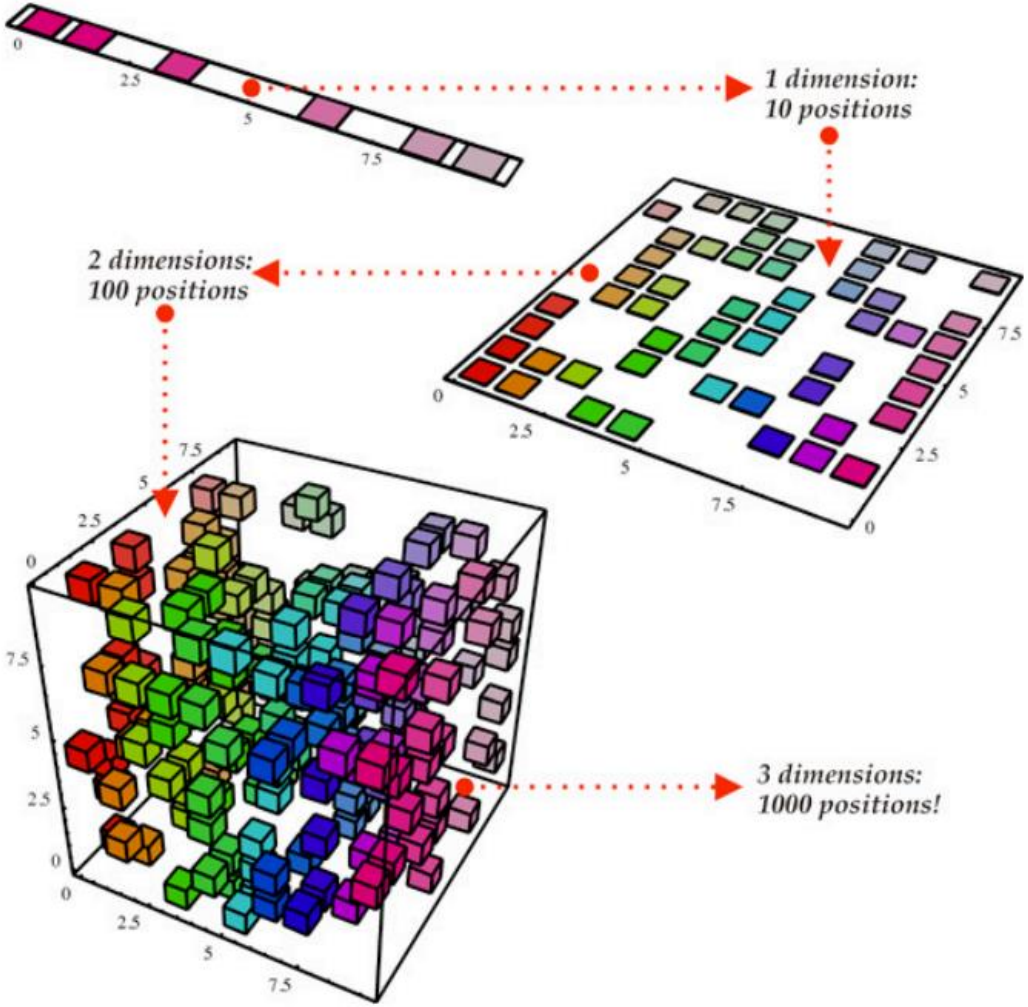
Recipe A: {salt, oil, mushrooms, bell peppers, cheese}

Recipe B: {pasta, oil, mushrooms}

J(Recipe A, Recipe B) = 2/6 = 1/3

https://youtu.be/Ah_4xqvS1WU

Dimensionality Reduction
#1: MDS

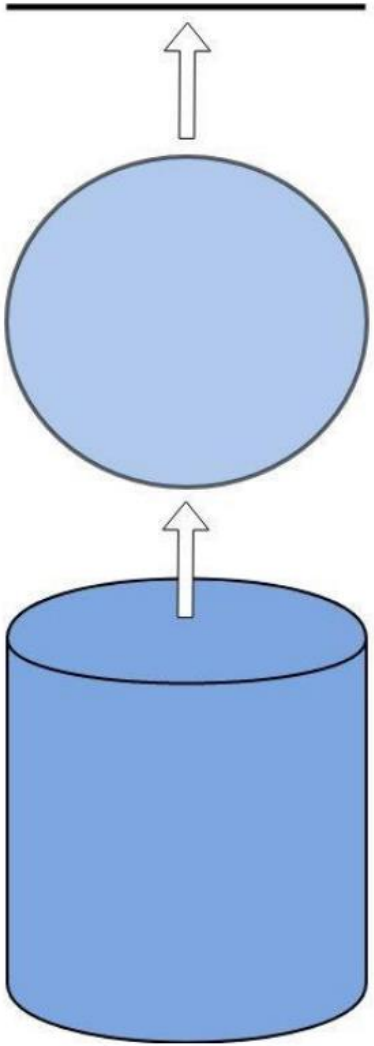


University of Exeter

Dimensionality Reduction

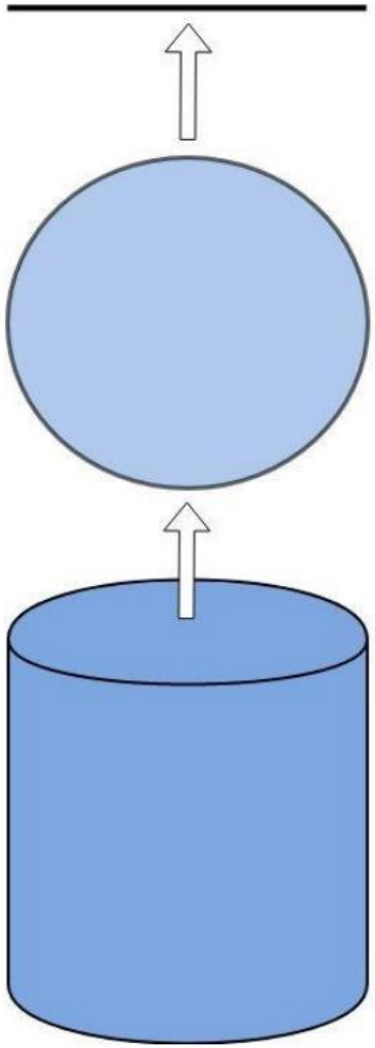
- Data can contain high level features
- Amazon books:
  - 1000s of customers
  - 1000s of books
  - Customer is a vector (0,1,0,0,0,…,0)
- What are some characteristics shared by many books/customers?

High-level features: language, genre, author etc..



Dimensionality Reduction

- Computational costs (time, memory, storage etc)
- Degradation of model performance
- Feature redundancy (correlations)
- Improve interpretability
- Reduce noise
- Data visualisation



Dimensionality Reduction

- **Feature Selection** – selecting a subset of relevant features (ignore redundant, irrelevant features)

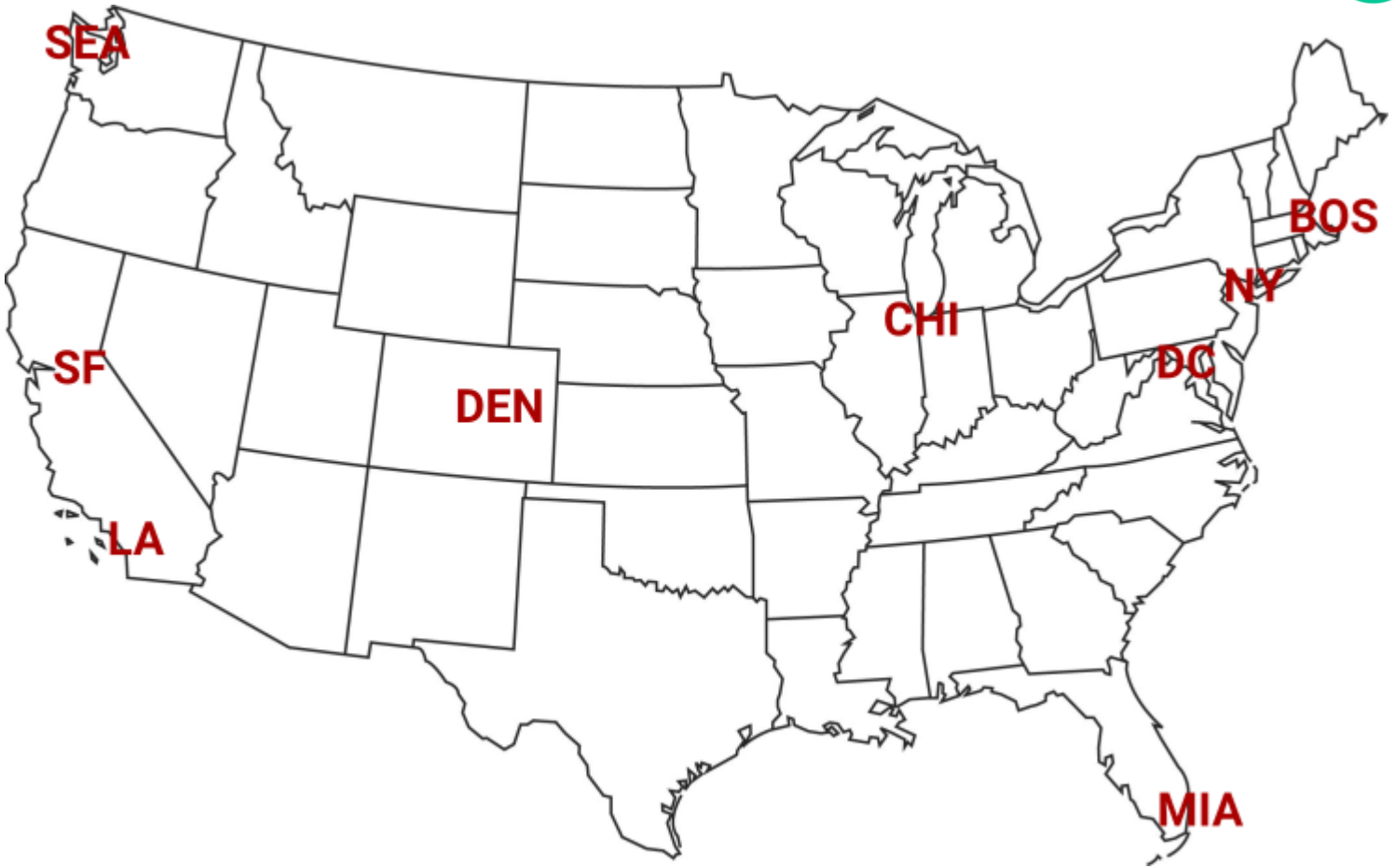
  - Original features are retained

- **Feature extraction** – finding a smaller set of features, in lower dimensional space, by extracting/deriving information from the original features in space.

  - Data is transformed by mapping it in the new lower dimensional feature space.

  - For example, multi-dimensional scaling (MDS), principal component analysis (PCA)

1. Multi-dimensional Scaling



Multi-dimensional Scaling (MDS)

```
library(tidyverse)

cities <- read_csv('city_distance.csv')
view(cities)
```

We don't know the actual locations of the cities – we only know the distance between them.

Each exists in high-dimensional space (9D), not just an x and a y.

	city	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
1	BOS	0	206	429	1504	963	2976	3095	2979	1949
2	NY	206	0	233	1308	802	2815	2934	2786	1771
3	DC	429	233	0	1075	671	2684	2799	2631	1616
4	MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
5	CHI	963	802	671	1329	0	2013	2142	2054	996
6	SEA	2976	2815	2684	3273	2013	0	808	1131	1307
7	SF	3095	2934	2799	3053	2142	808	0	379	1235
8	LA	2979	2786	2631	2687	2054	1131	379	0	1059
9	DEN	1949	1771	1616	2037	996	1307	1235	1059	0

Multi-dimensional Scaling (MDS)

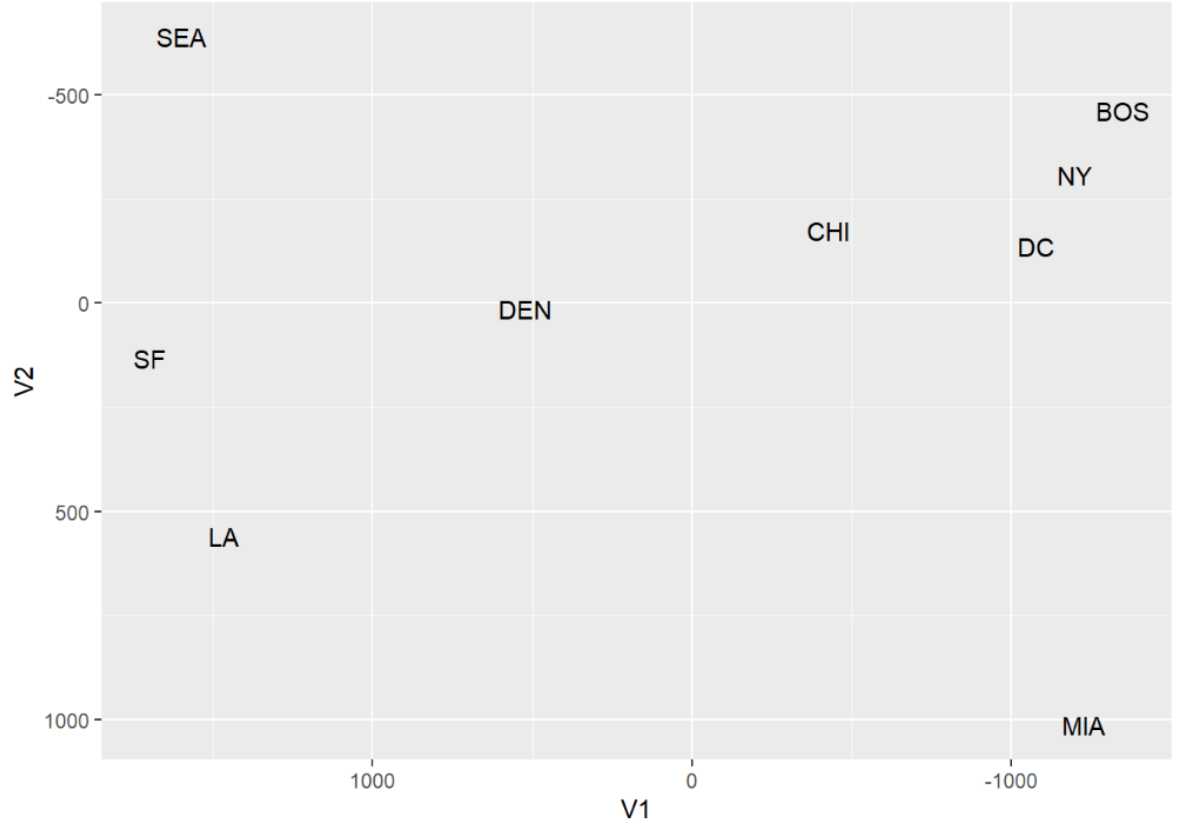
```
md_cities <- select(cities, -city) %>%
        cmdscale() %>%
        as.data.frame()
```

	V1	V2	city_name
1	-1348.6683	-462.40060	BOS
2	-1198.8741	-306.54690	NY
3	-1076.9855	-136.43204	DC
4	-1226.9390	1013.62838	MIA
5	-428.4548	-174.60316	CHI
6	1596.1594	-639.30777	SEA
7	1697.2283	131.68586	SF
8	1464.0470	560.58046	LA
9	522.4871	13.39576	DEN

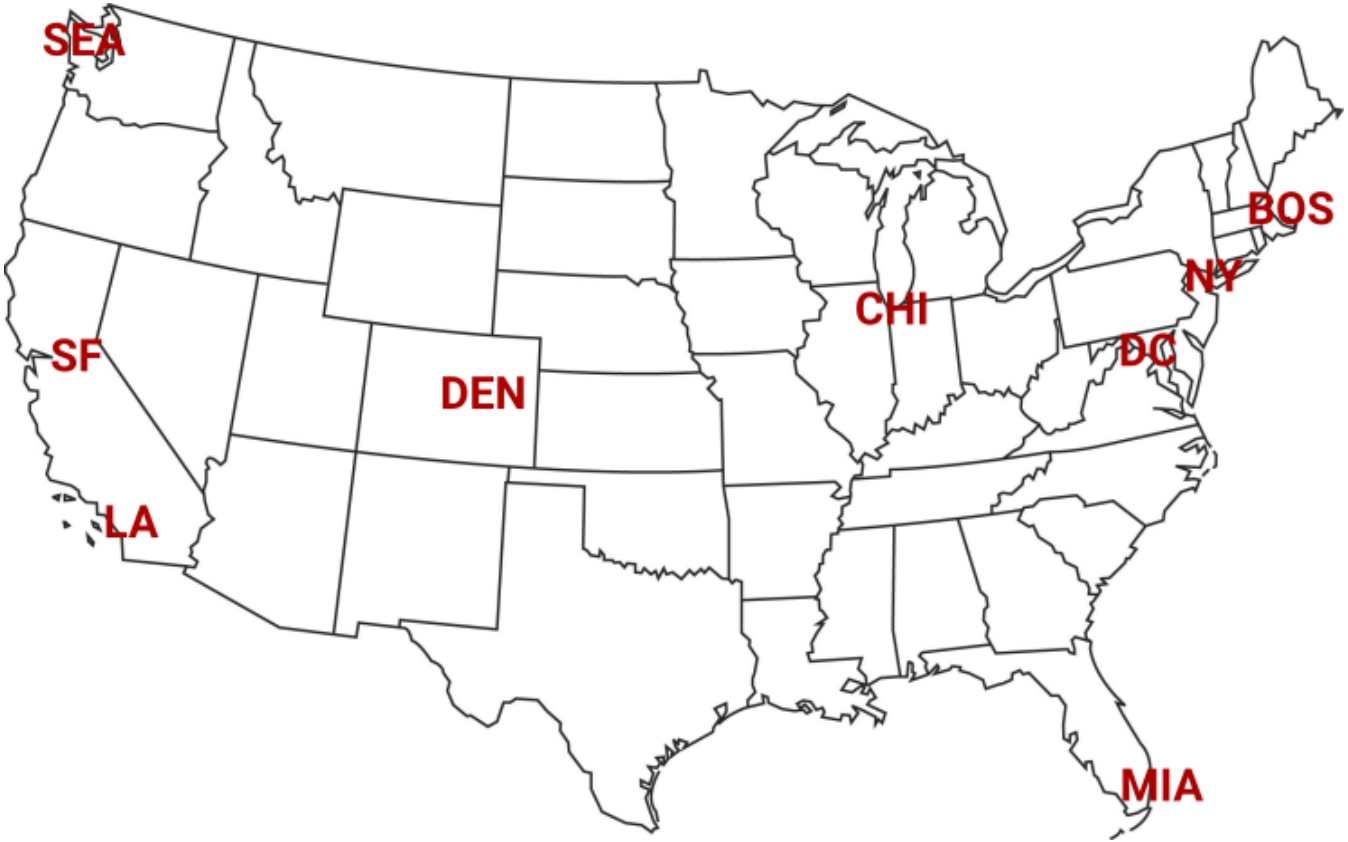
MDS creates a configuration of points in a lower-dimensional space, such that the distances between the points reflect the dissimilarities between the objects as closely as possible.

MDS can work with any type of data that can be expressed as distances or dissimilarities between objects.

```
ggplot(md_cities, aes(x = V1, y = V2)) +
       geom_text(aes(label = city_name)) +
       scale_x_reverse() +
       scale_y_reverse()
```

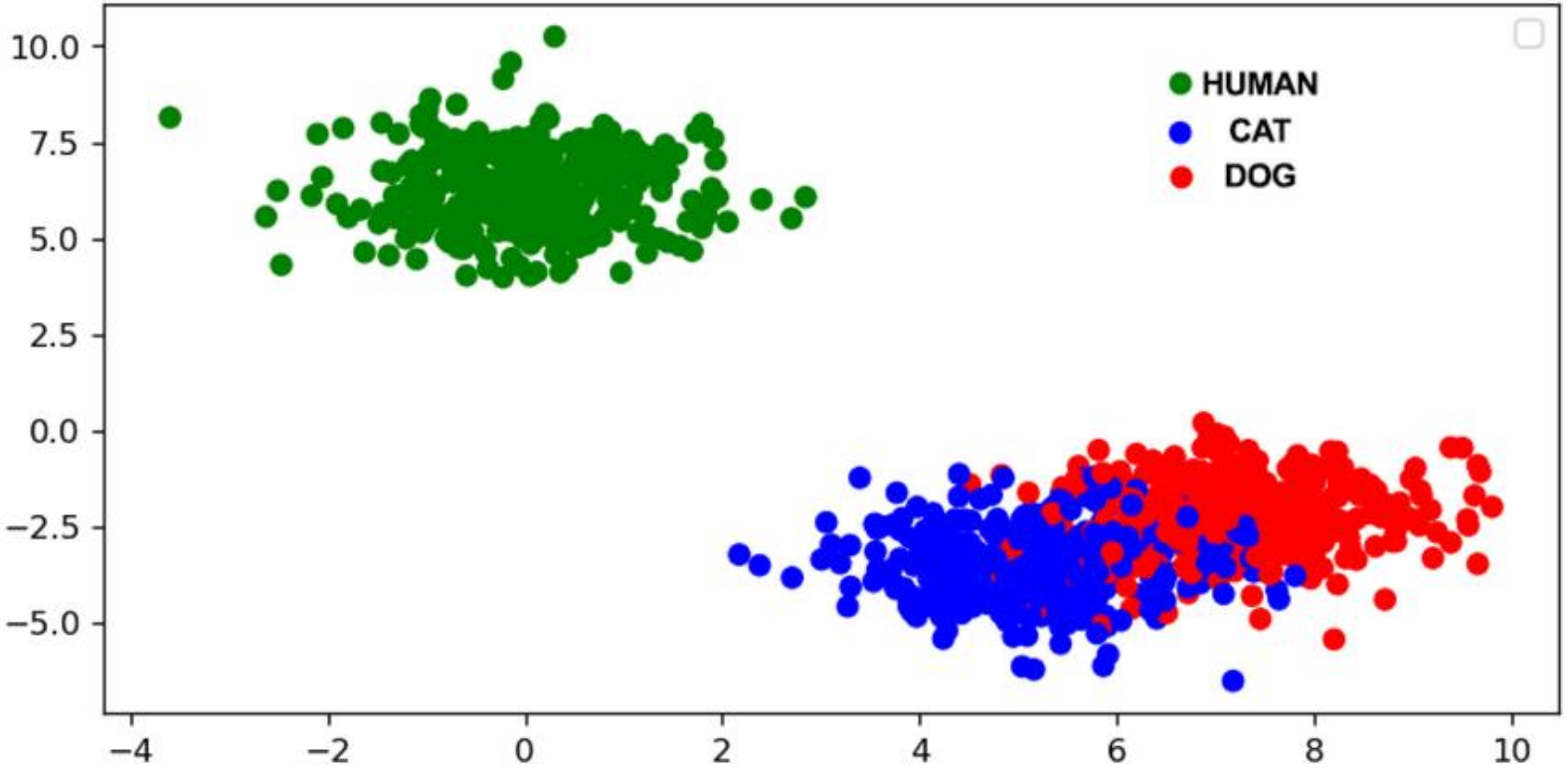


Multi-dimensional Scaling (MDS)

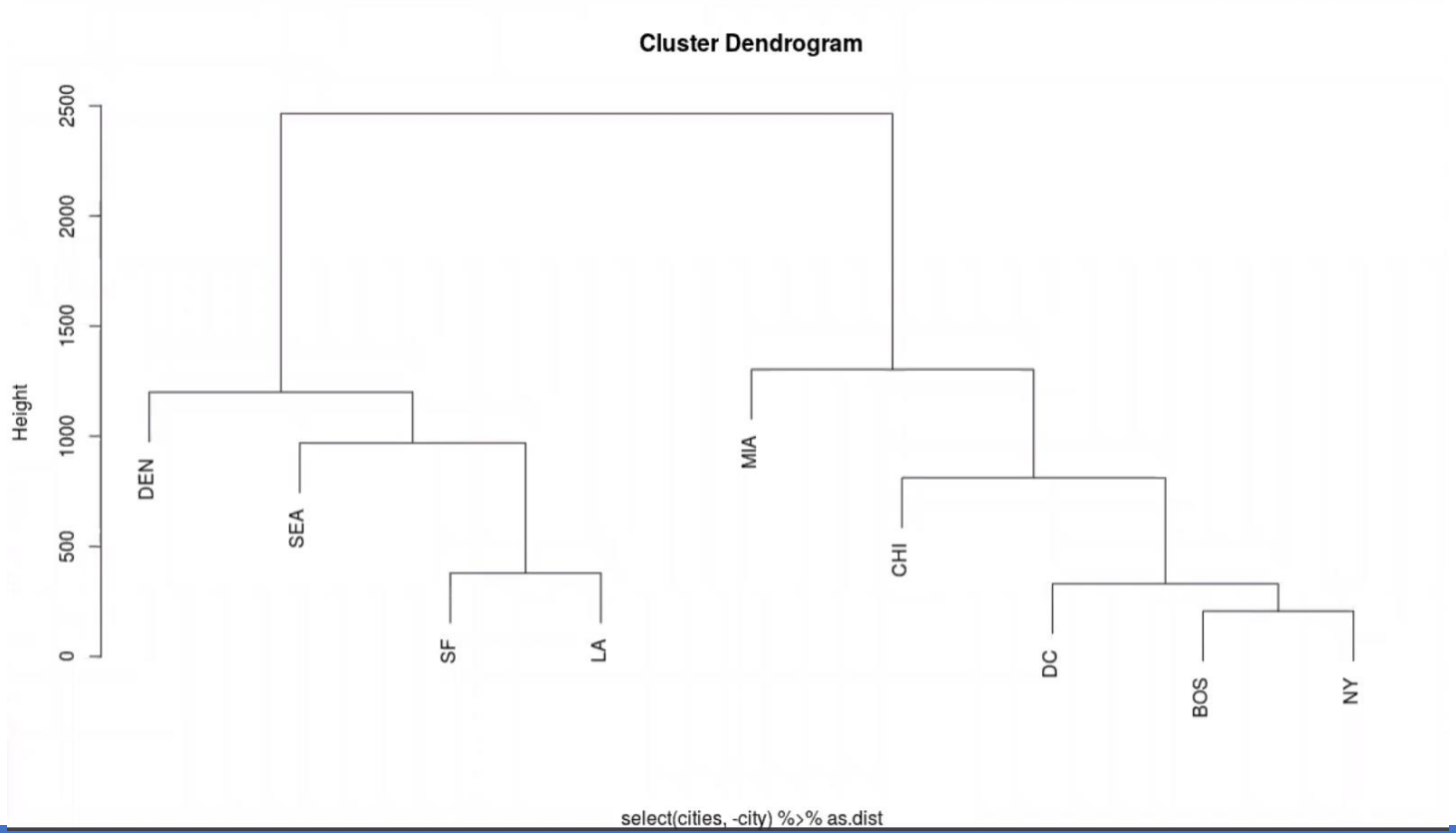


Cluster analysis

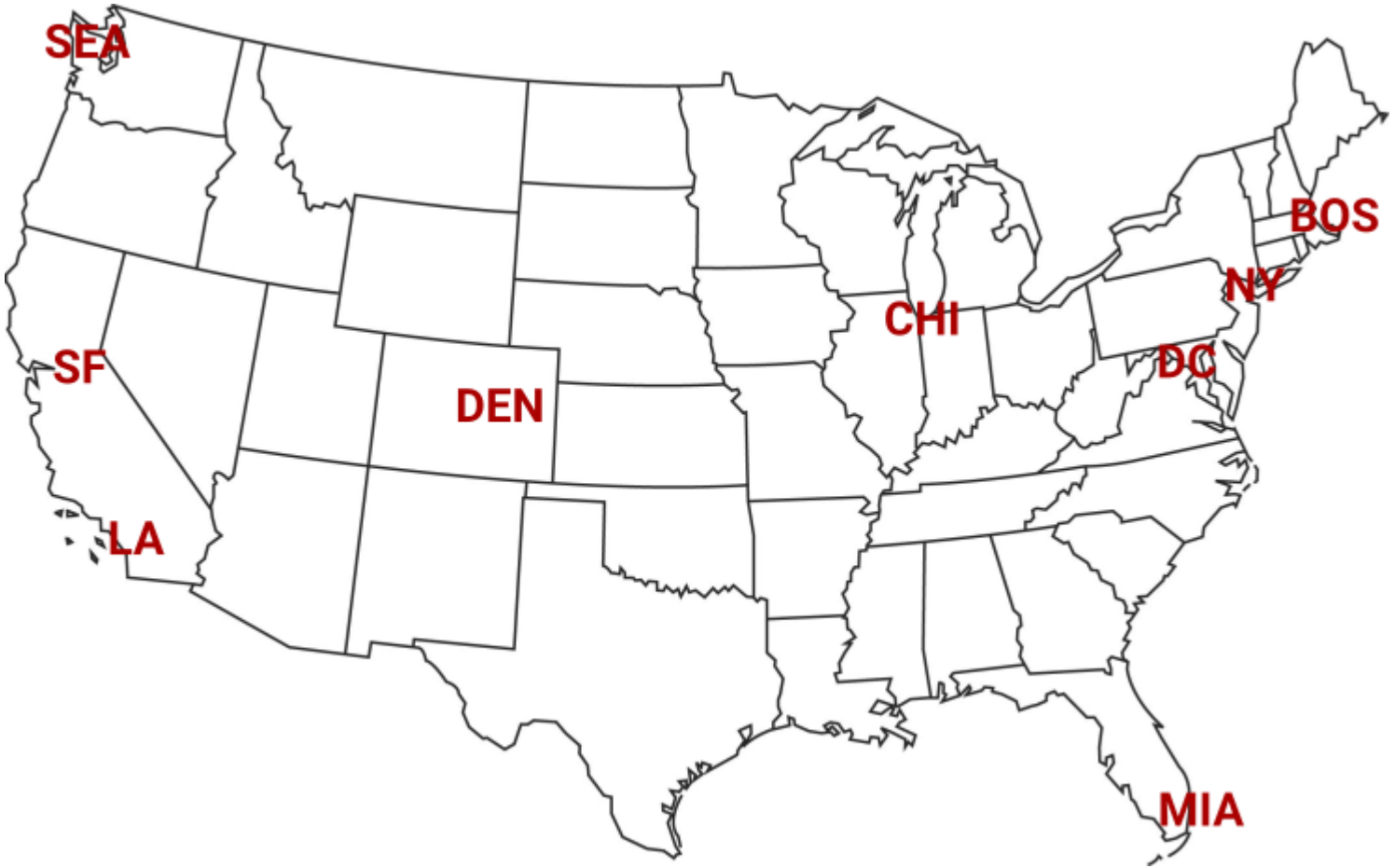
1. Heirarchical clustering
2. k-means clustering



1. Hierarchical Clustering

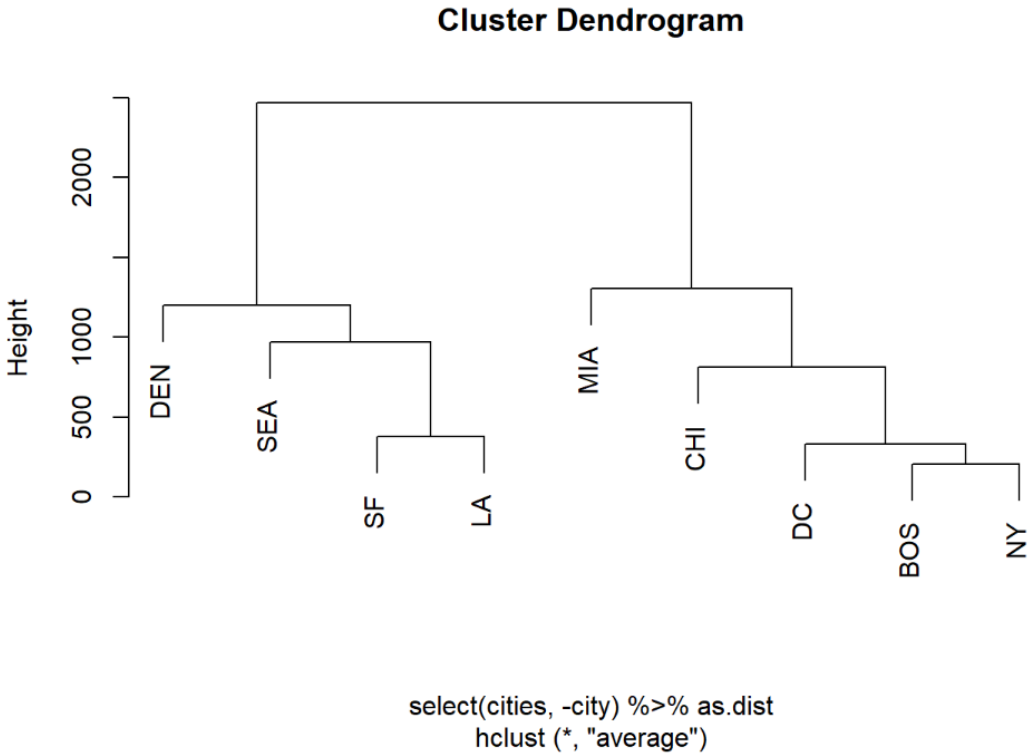


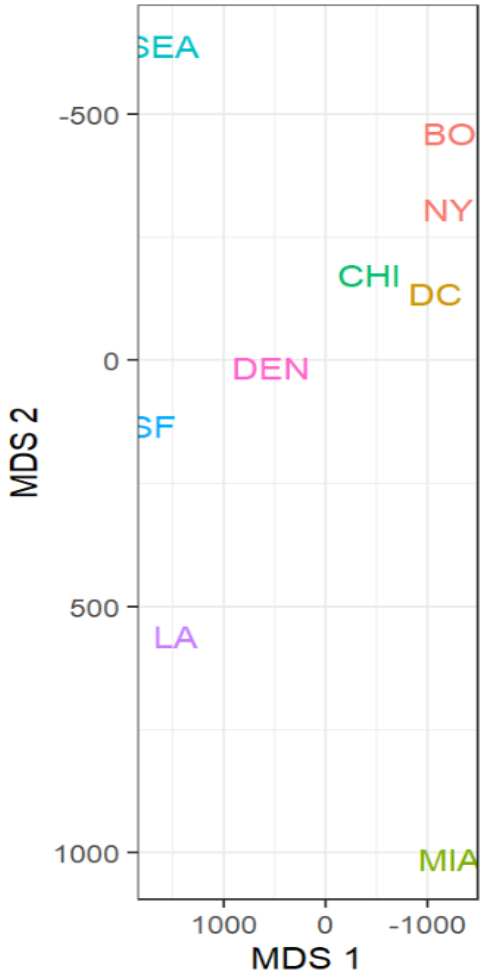
	city	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
1	BOS	0	206	429	1504	963	2976	3095	2979	1949
2	NY	206	0	233	1308	802	2815	2934	2786	1771
3	DC	429	233	0	1075	671	2684	2799	2631	1616
4	MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
5	CHI	963	802	671	1329	0	2013	2142	2054	996
6	SEA	2976	2815	2684	3273	2013	0	808	1131	1307
7	SF	3095	2934	2799	3053	2142	808	0	379	1235
8	LA	2979	2786	2631	2687	2054	1131	379	0	1059
9	DEN	1949	1771	1616	2037	996	1307	1235	1059	0

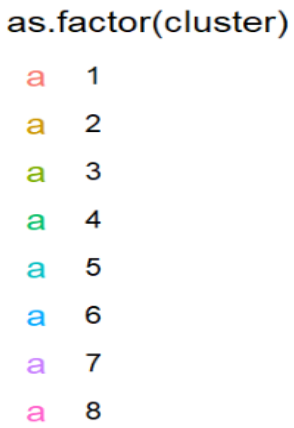


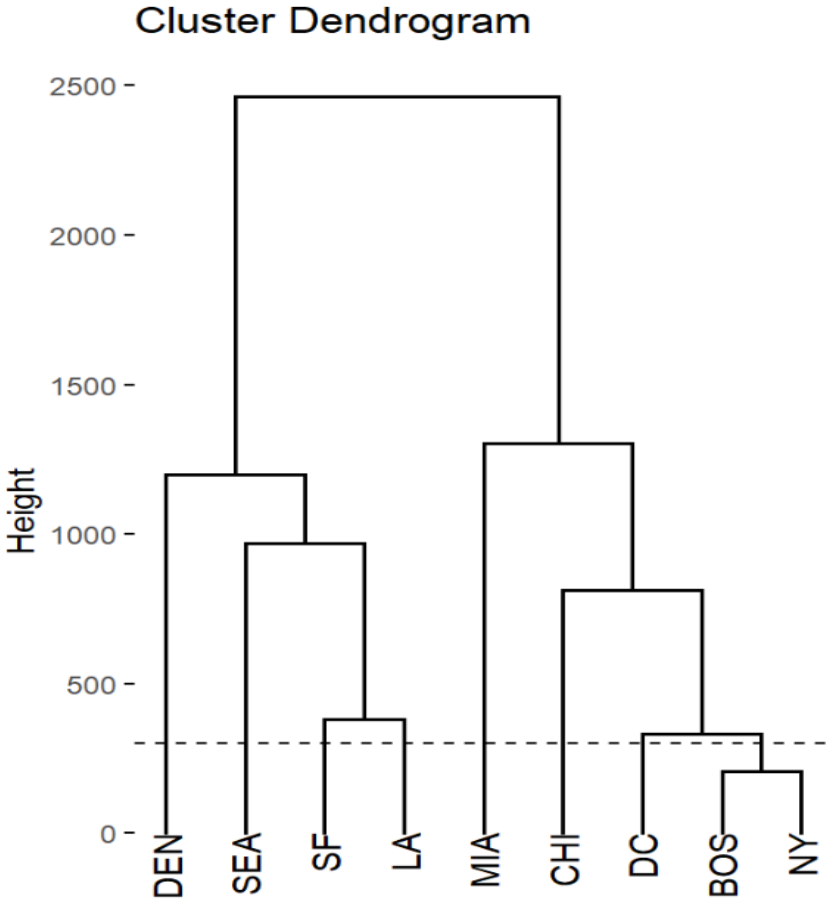
University of Exeter

```
cities_hc <- hclust(select(cities, -city) %>% as.dist,
method = 'ave')
plot(cities_hc)
```

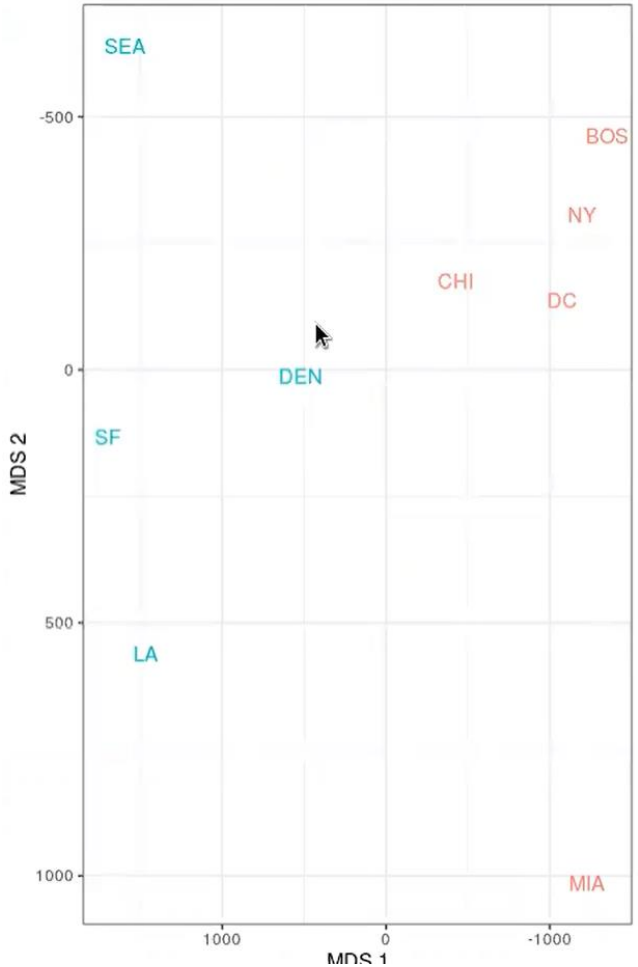


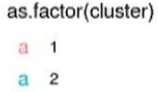


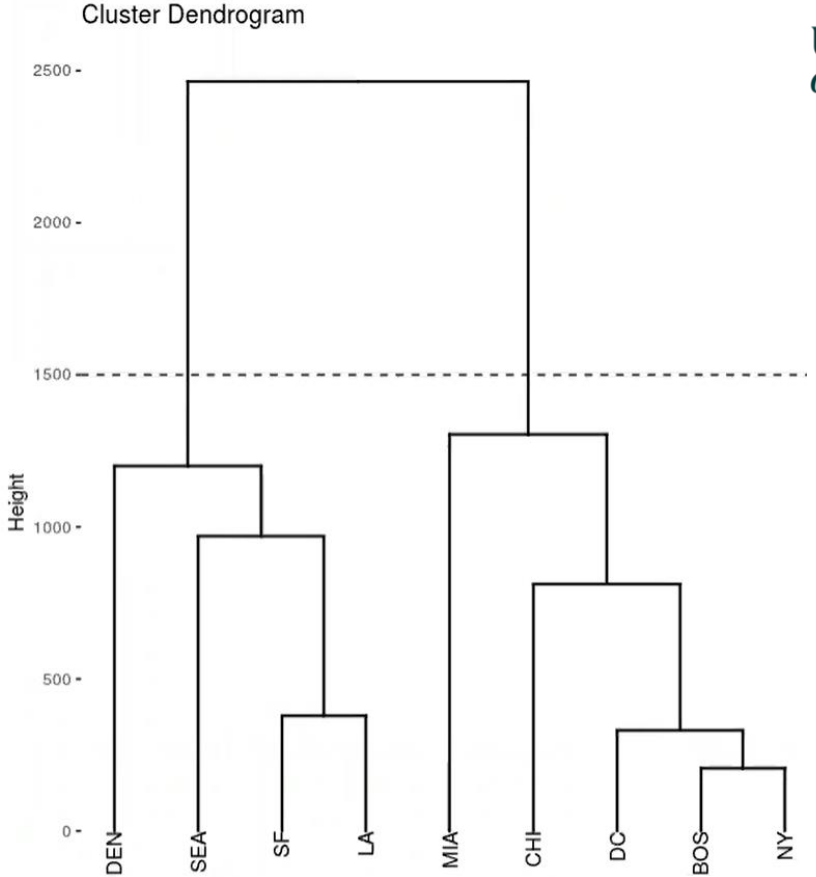




niversity
Exeter

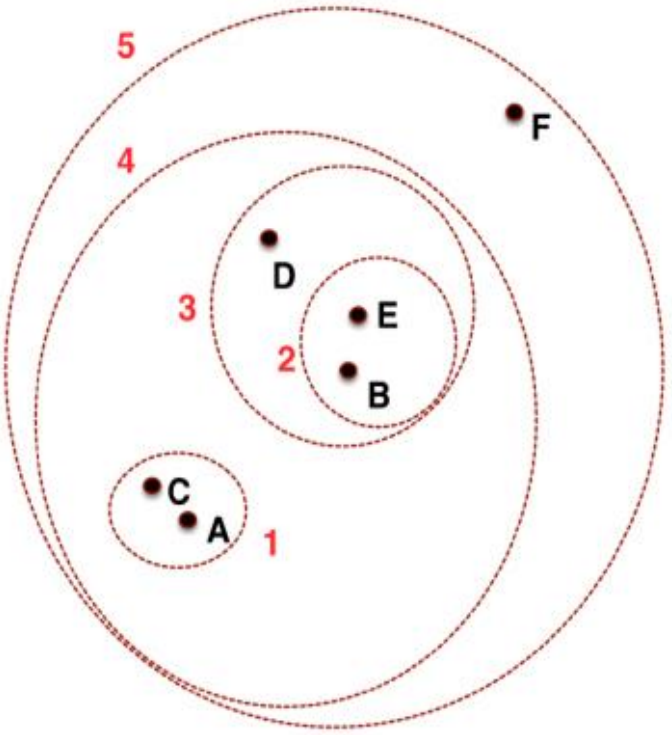


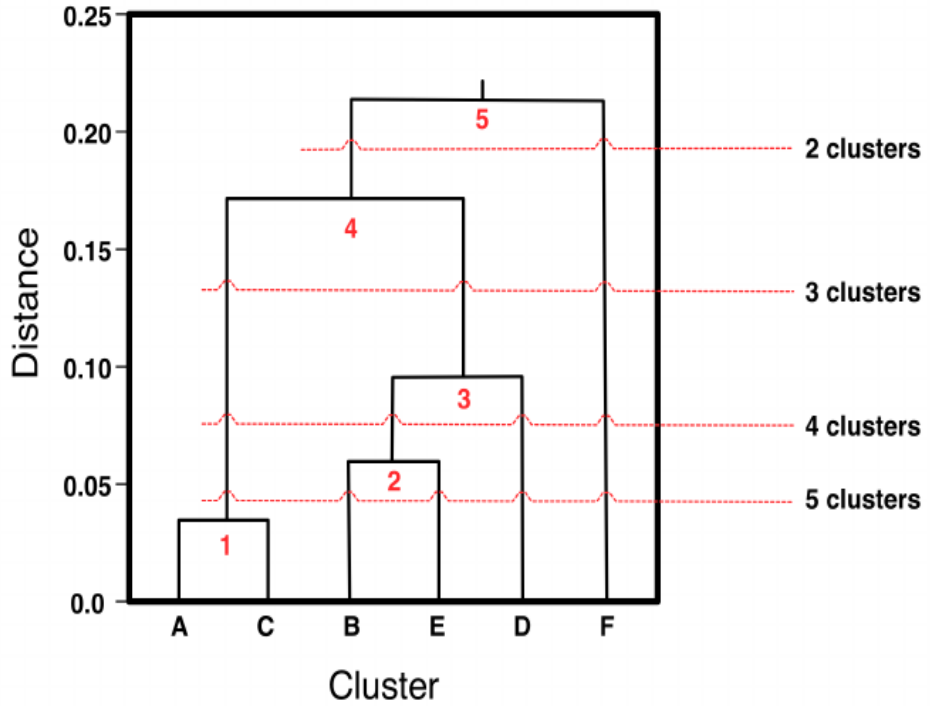


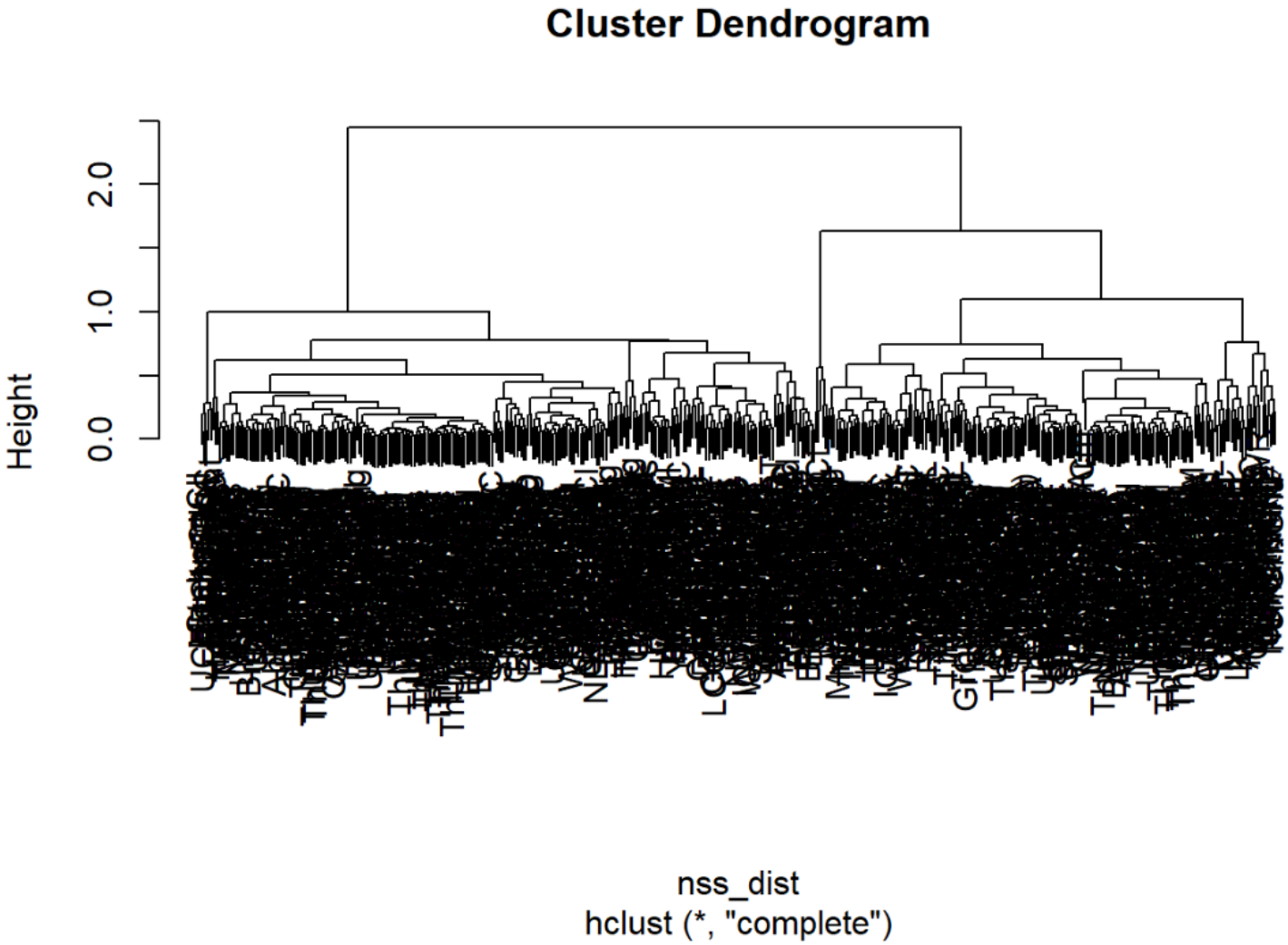


University of Exeter

Hierarchical Clustering






Cluster Dendrogram
Height
0.0
1.0
2.0
nss_dist
hclust (*, "complete")

2. K-means Clustering

Starting with N data points $\{x_1, x_2, \ldots, x_N\}$

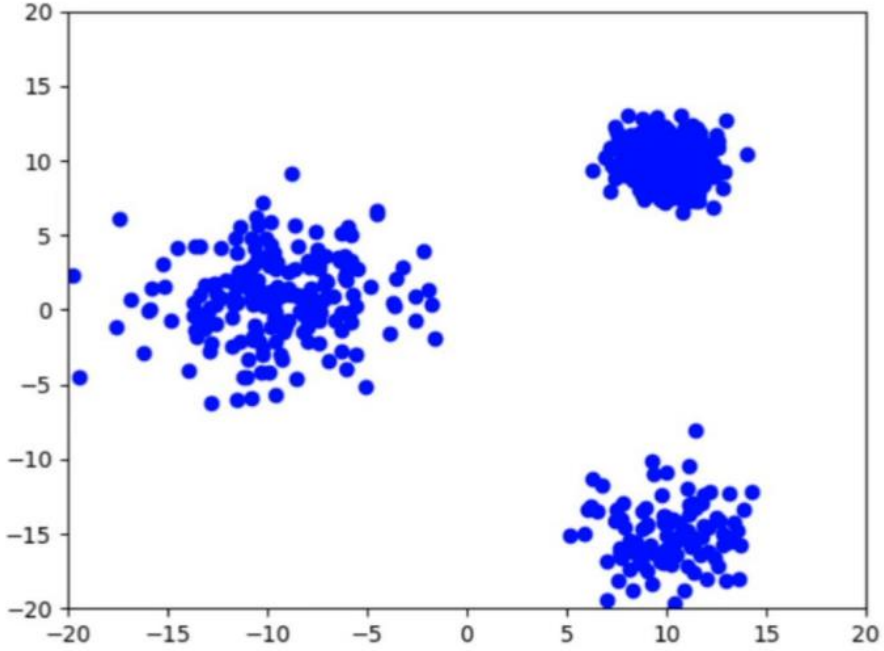
Choose *k*, i.e. the number of clusters

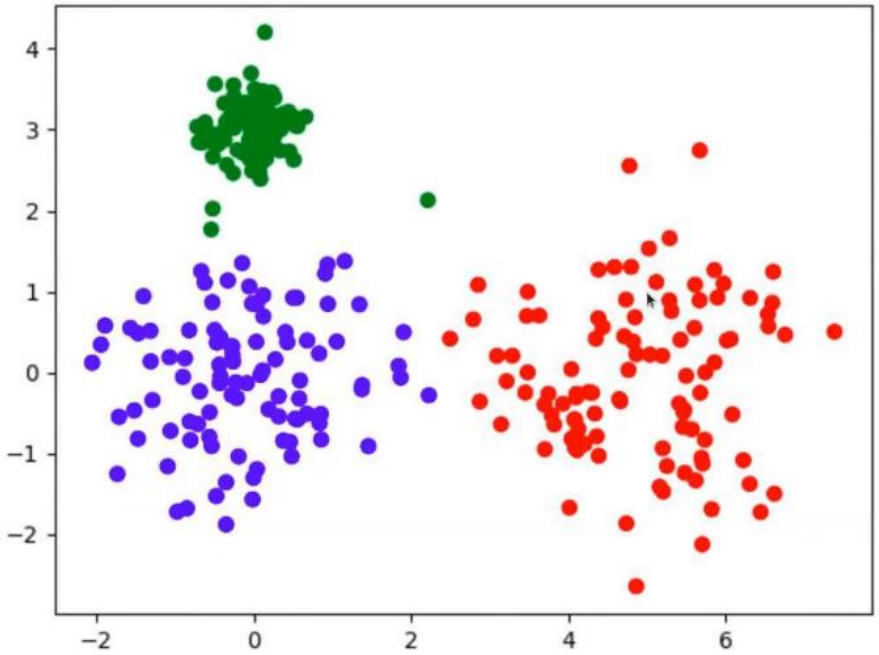
Choose *k* cluster centres $\{c_1, c_2, \ldots, c_N\}$

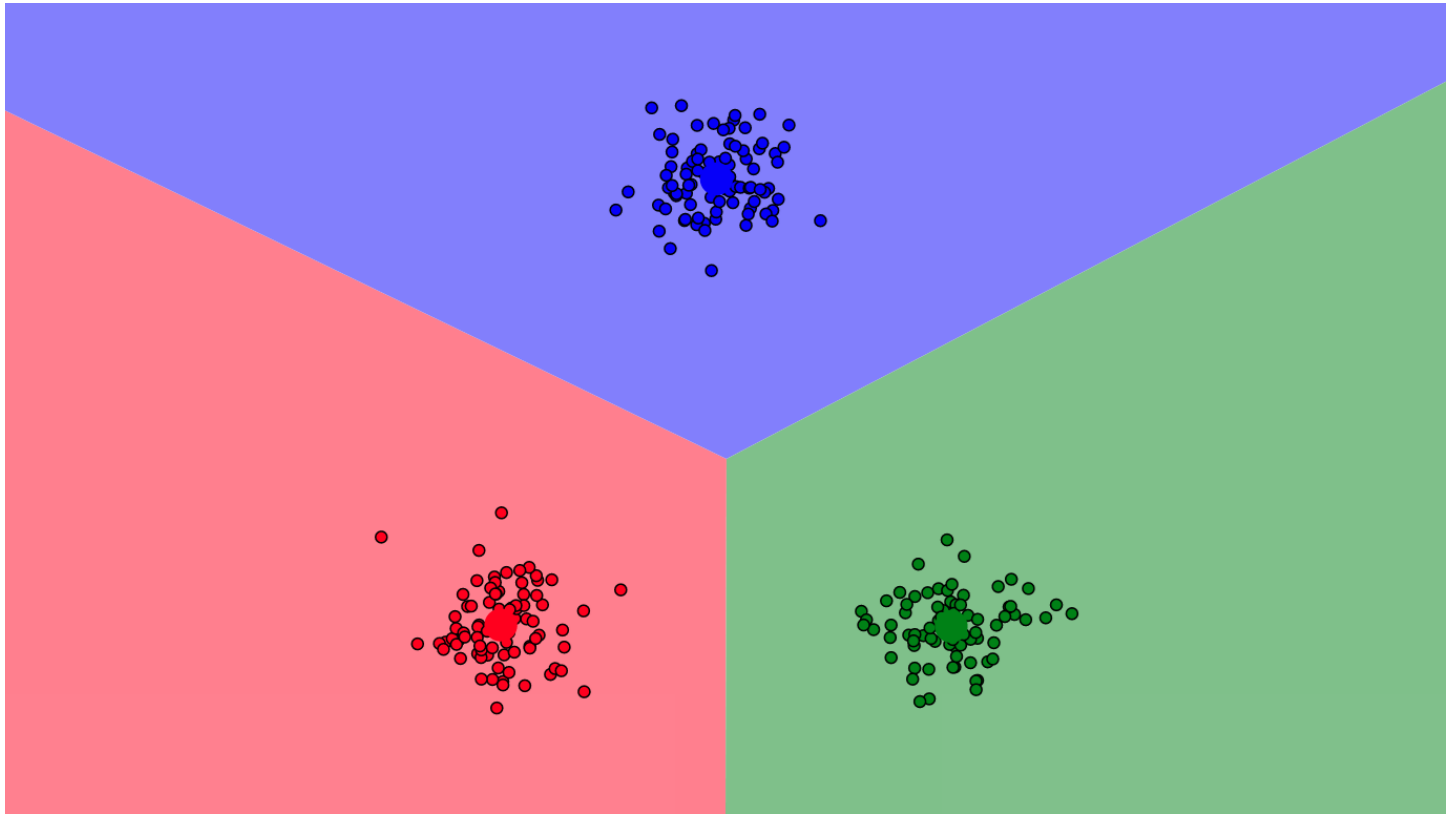
StatQuest: K-means clustering - YouTube

Clustering

University of Exeter

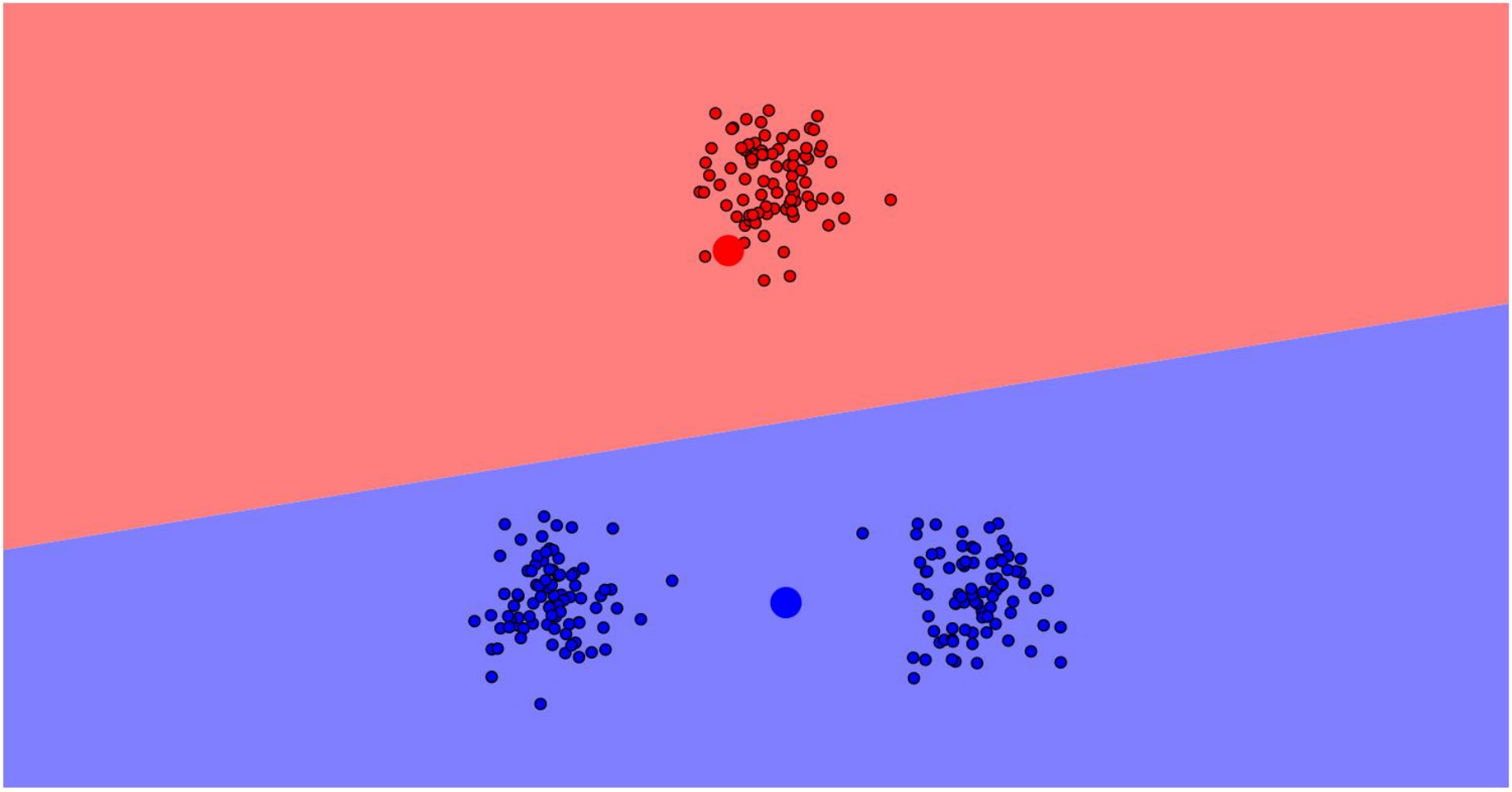




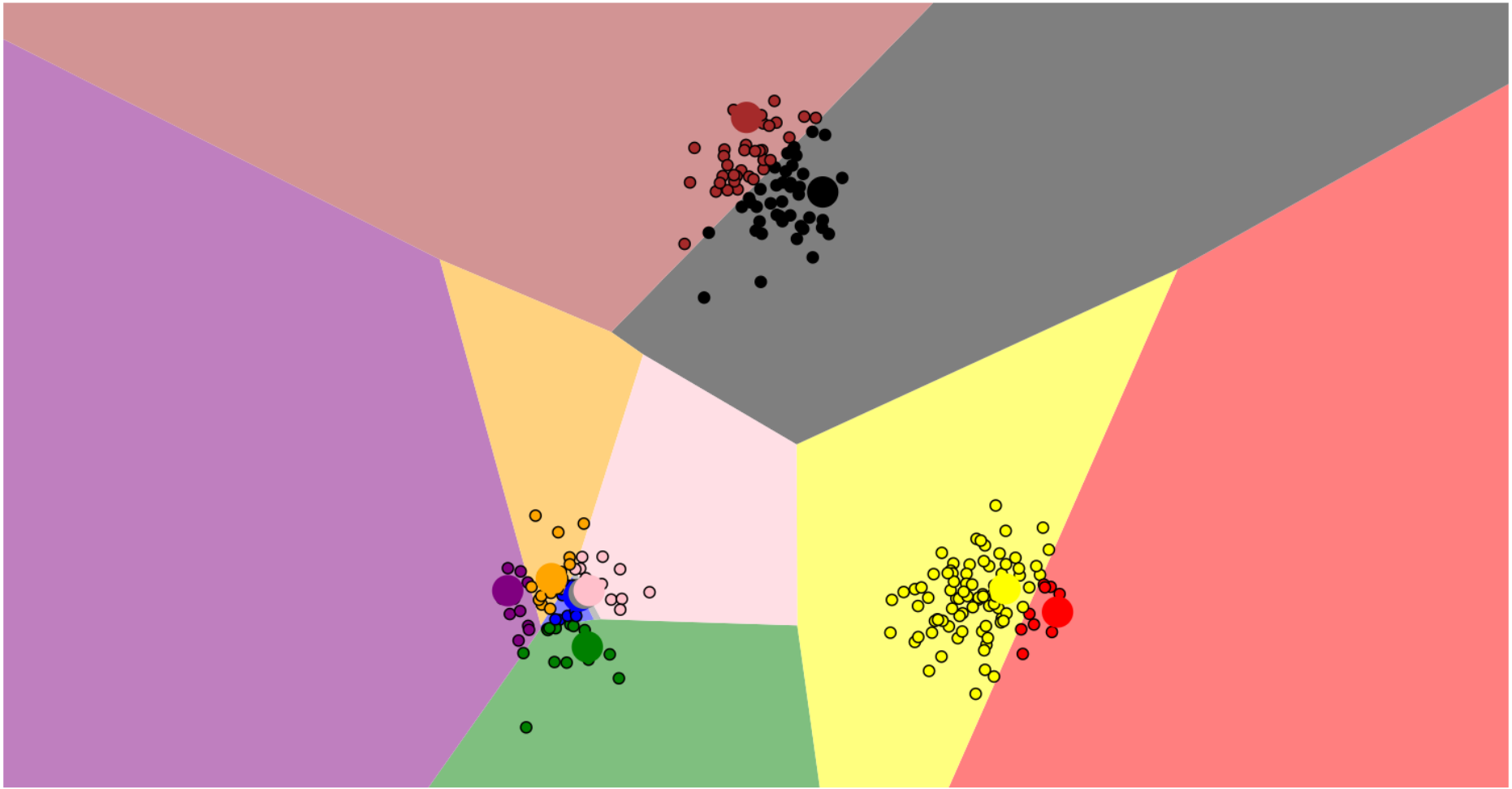


Visualizing K-Means Clustering (naftaliharris.com)

d3.js ~ Voronoi Diagram (strongriley.github.io)



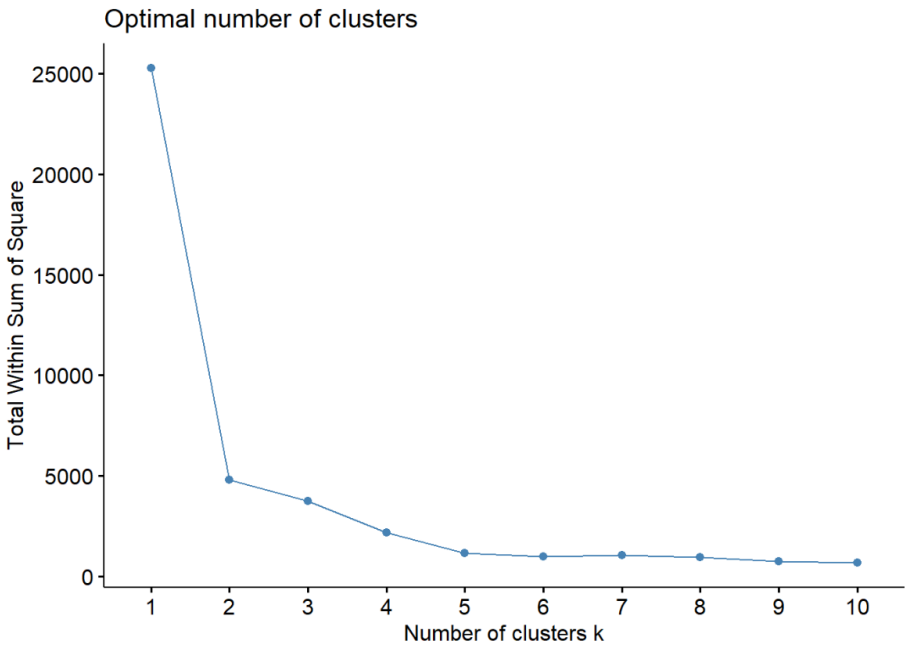
Restart | Add Centroid | Update Centroids

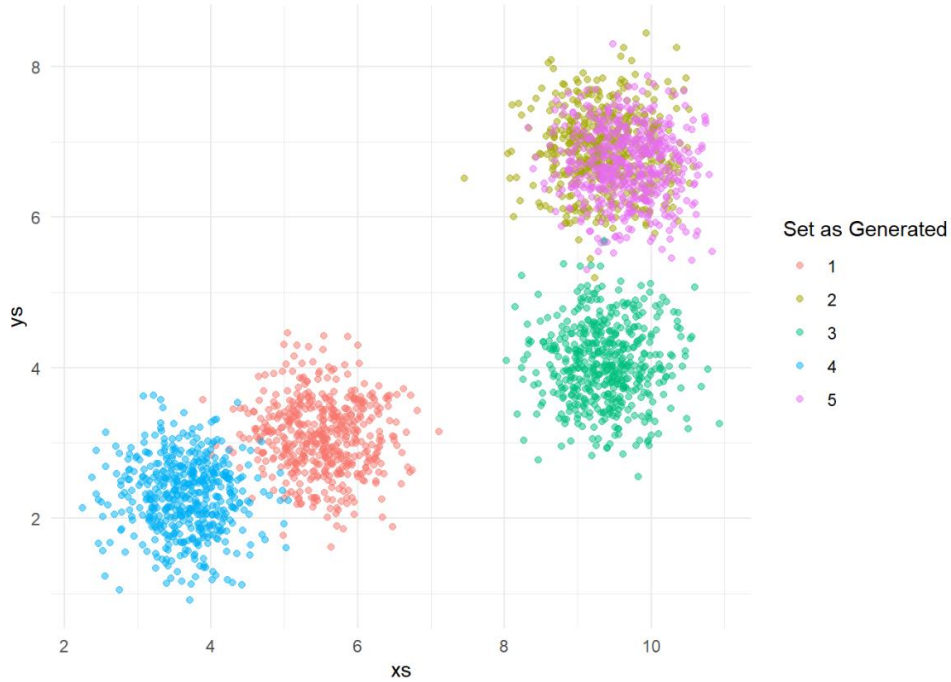


Restart | Update Centroids

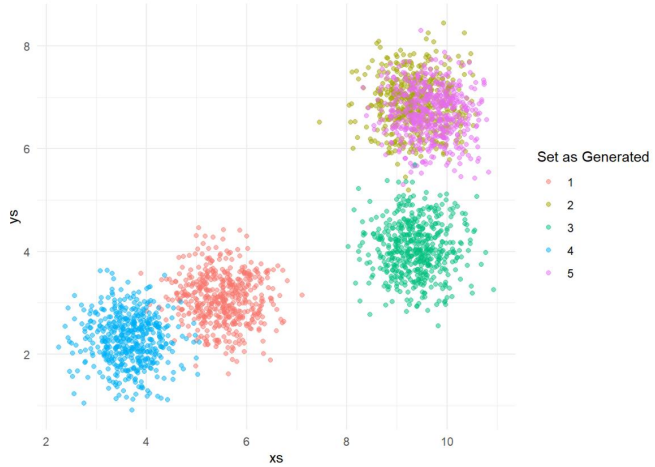
University of Exeter

```
fviz_nbclust(d,kmeans, method = 'wss')
```



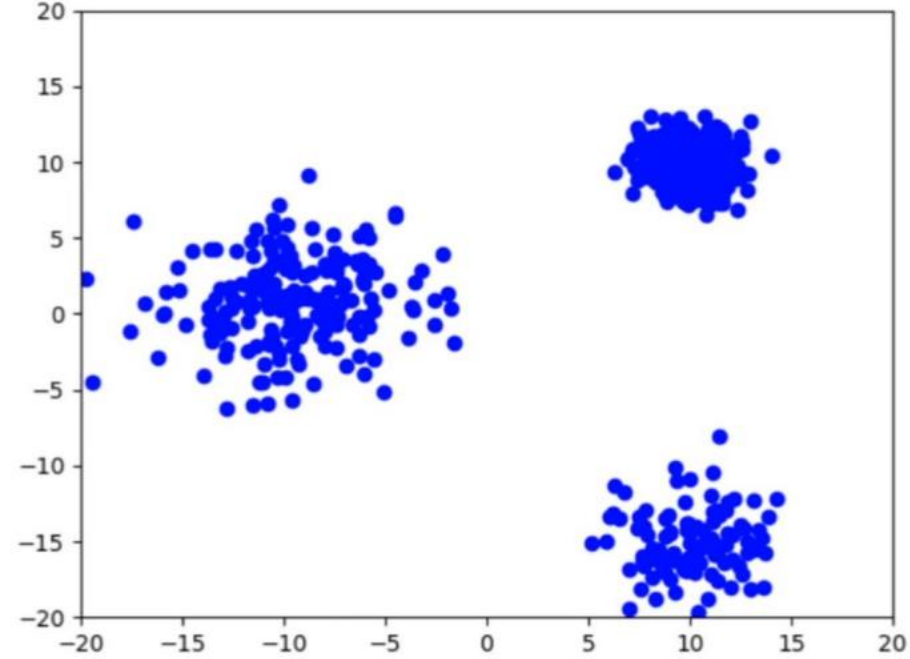


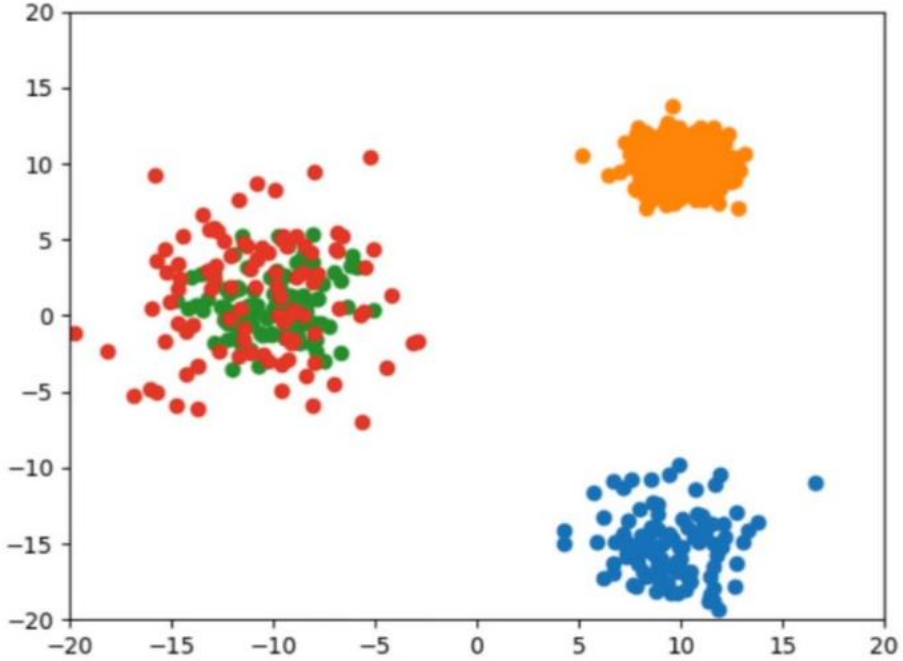
```r
set.seed(111222333)
k <- 5
N <- 500
xs <- runif(k, 0, 10)
ys <- runif(k, 0, 10)
d <- data.frame(xs = lapply(xs, function(x) rnorm(N, mean = x, sd = 0.5)) %>%
         unlist, ys = lapply(ys, function(x) rnorm(N, mean = x, sd = 0.5)) %>%
         unlist)

ggplot(d, aes(x = xs, y = ys, color = factor(rep(1:k, each = N)))) +
geom_point(alpha = 0.5) + theme_minimal() + scale_color_discrete('Set as
Generated')
```



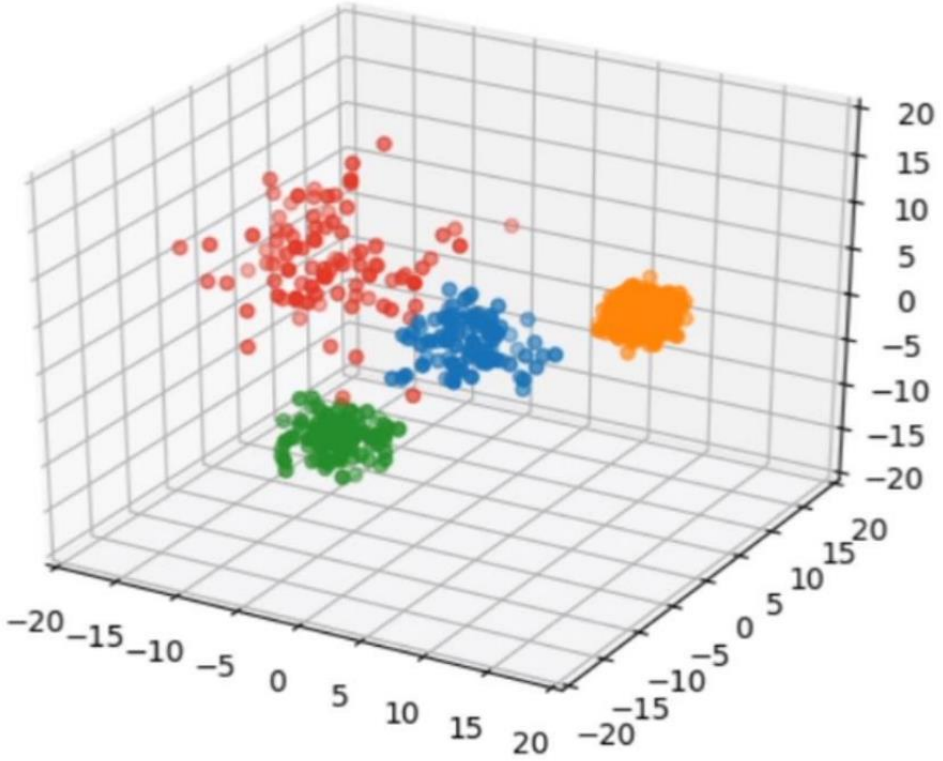
Clustering

University of Exeter

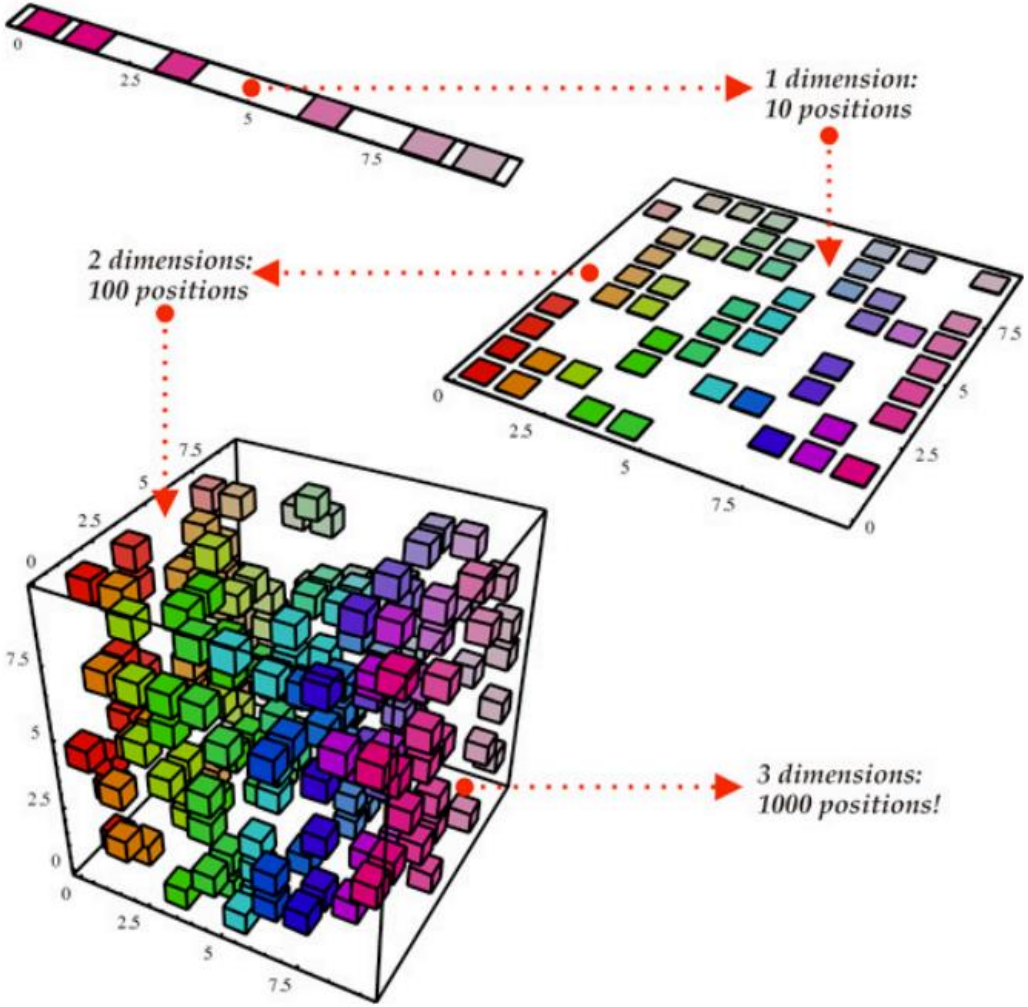




Clustering

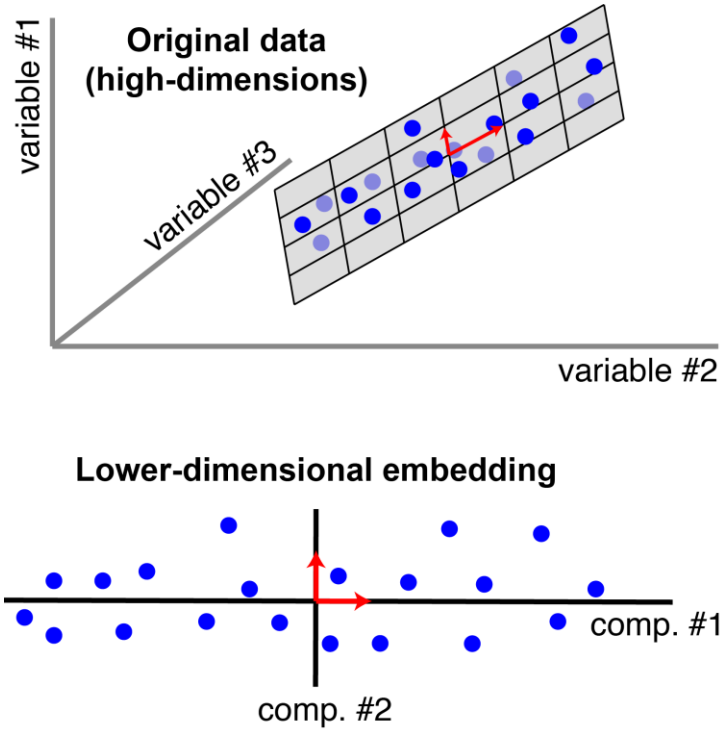


Dimensionality Reduction #2 PCA



University of Exeter

Principal Component Analysis (PCA)

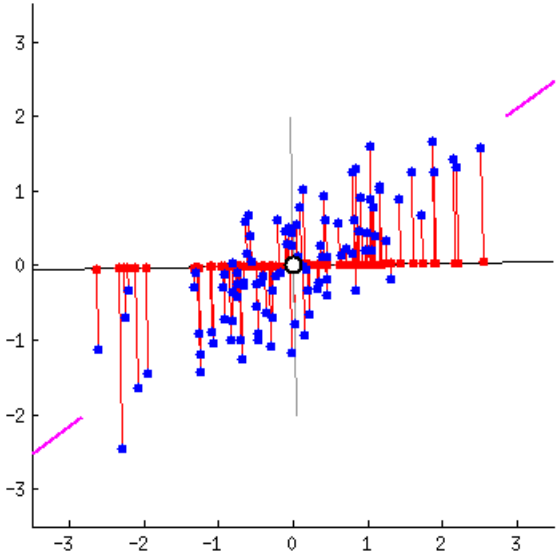


- A linear dimensionality reduction technique
- Set of new features called **principal components** are extracted from an existing set
- New features are expressed as **weighted linear combinations** of the original data

Everything you did and didn't know about PCA

http://setosa.io/ev/principal-component-analysis
Further explanation with pretty visualisations

Principal Component Analysis (PCA)



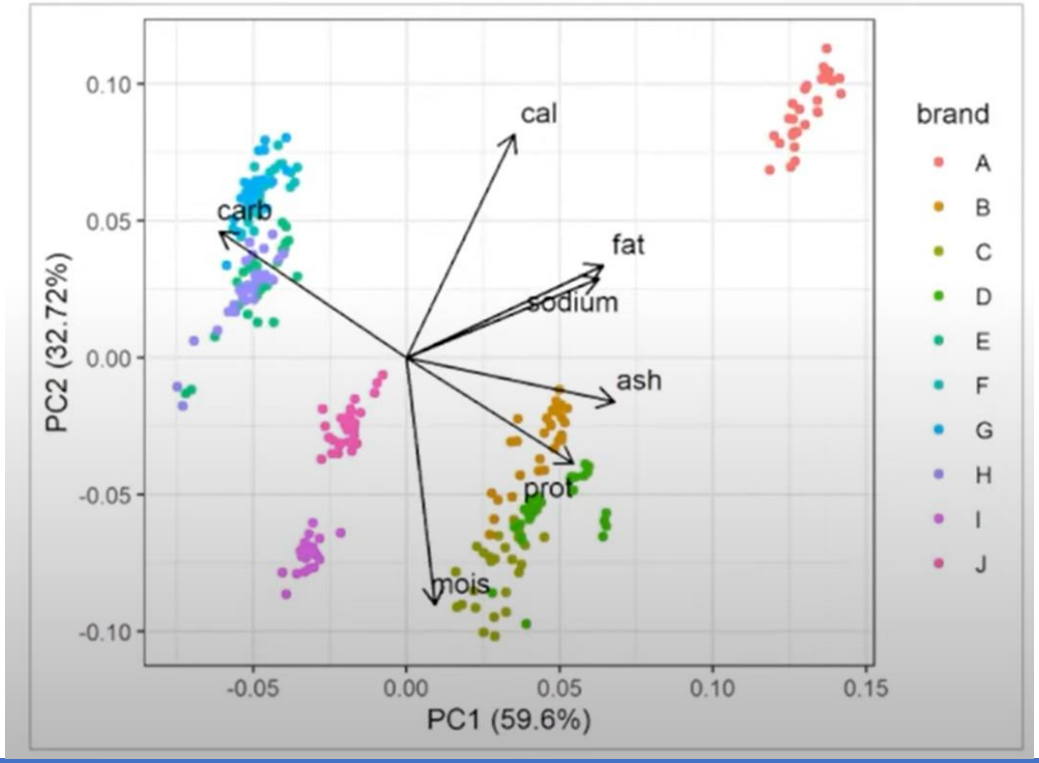
As there are as many principal components as there are variables in the data, principal components are constructed in such a manner that the first principal component accounts for the **largest possible variance** in the data set.

Everything you did and didn't know about PCA

http://setosa.io/ev/principal-component-analysis
Further explanation with pretty visualisations

Principal Component Analysis (PCA)

Biplot

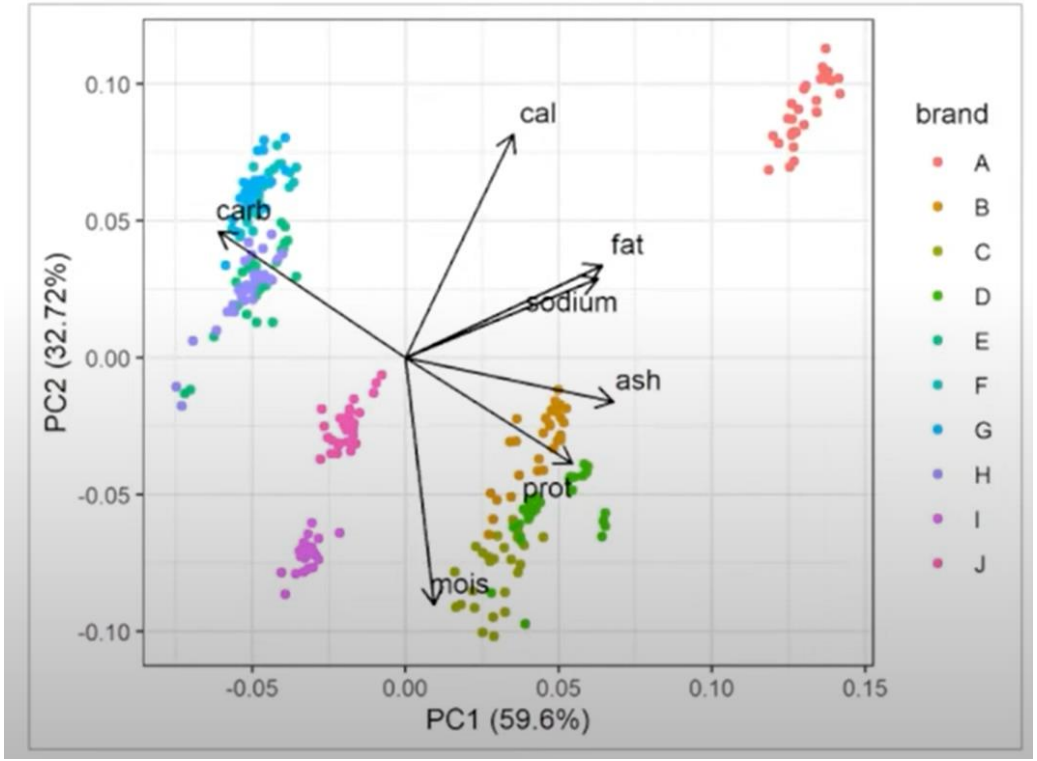


PCA is used to identify the directions (principal components) that maximise the variance in the data. It projects the data onto new axes which are linear combinations of the original variables, while preserving as much variation as possible.

vechttps://www.youtube.com/watch?v=DngS4LNNzc8&t=0s

Principal Component Analysis (PCA)

Biplot



9 dimensional data (attributes of pizzas) are reduced to 2 principal components (x and y).

Each pizza data point is plotted. Colour is added by pizza brand which shows some similarity within brands.

The vectors are calculated for each original feature. We can make some interpretations of the analysis.

vechttps://www.youtube.com/watch?v=DngS4LNNzc8&t=0s

Next Week: Predictive Modelling

- Read Data Science for Business, chapters 3 and 4

- Watch StatQuest: Decision Trees

- Watch StatQuest: Random Forests Part 1

- Watch StatQuest: Random Forests Part 2

- Play A Visual Introduction to Machine Learning

- Play Random Forest Playground

- Play Linear Regression (try clicking and dragging on points)

University of Exeter

Any questions?

?