Multivariate Analysis (STU33011)

The topics we will cover include:

- Data Reduction
  - Principal Components Analysis
  - Multidimensional Scaling
- Cluster Analysis
  - Hierarchical Clustering
  - k-means Clustering
- Discriminant Analysis/Classification
  - k-Nearest Neighbours
  - Linear and Quadratic Discriminant Analysis
  - Logistic Regression

Multivariate Data

- Multivariate data arises when two or more attributes are recorded for each of a set of objects.
- **Example:** The heights and the weights of a set of students could be recorded.
- **Example:** A bank may record the current account balance, savings account balance and the credit card balance for its customers.
- In some cases we have only recorded numerical attributes (as above), but in other cases some of the attributes are numerical and some are categorical.
- **Example:** In the above banking example, the gender and geographical location of the customers may also be recorded.

2003 Decathlon Results

- The athletes' performance in the decathlon events in the 2003 world championships (only those who completed all events are included).

Name	100m	LongJ	ShotPutt	HighJ	400m	Hurdles	Discus	Pole	Javelin	1500m	Points
Qi	11.30	7.39	12.85	2.00	48.73	14.40	46.72	4.8	59.98	265.4	8126
Sebrle	11.00	7.64	15.47	2.06	47.90	14.25	47.47	4.8	69.79	274.5	8634
Dvorak	11.03	7.28	15.95	1.94	50.04	14.15	45.47	4.5	67.10	267.6	8242
Hernu	11.20	7.22	13.99	2.03	48.95	14.15	46.13	4.9	59.63	268.4	8218
Niklaus	11.19	7.21	13.87	1.97	49.95	14.50	42.68	5.1	57.55	268.8	8020
Bernard	10.91	7.22	15.39	2.03	49.31	14.76	43.47	4.3	59.47	274.5	8000
Karpov	10.72	7.75	15.51	2.12	47.33	13.95	47.38	4.4	47.53	277.7	8374
Warners	10.95	7.55	14.13	1.91	48.94	14.72	41.49	4.5	54.87	291.3	7753
Lobodin	10.99	7.08	15.43	1.97	49.54	14.36	48.36	5.0	56.50	274.6	8198
Pappas	10.80	7.62	16.11	2.09	47.58	13.99	46.94	5.1	65.90	284.3	8750
Smirnov	11.10	6.98	13.89	1.97	48.98	14.98	42.70	4.5	62.69	264.7	7897

2003 Heptathlon Results

	Name	100m	HighJump	ShotPutt	200m	LongJump	Javelin	800m	Points
1	Dufour	14.08	1.67	13.34	25.24	5.62	39.83	132.8	5723
2	Klueft	13.18	1.94	14.19	22.98	6.68	49.90	132.1	7001
3	Butor	13.92	1.79	12.51	24.67	5.86	46.43	136.7	6035
4	Sazanovich	13.67	1.76	16.81	24.25	6.47	44.93	136.5	6524
5	Netseporuk	13.91	1.76	13.97	24.96	6.05	50.05	139.8	6154
6	Barber	13.05	1.91	12.97	23.92	6.61	49.60	133.7	6755
7	Hollman	14.10	1.85	12.05	24.72	6.06	41.01	135.8	6018
8	Lewis	13.37	1.64	15.25	24.55	6.19	49.88	139.6	6254
9	Kesselschlaeger	13.34	1.76	13.77	24.94	6.17	41.57	137.2	6134
10	Strataki	13.93	1.79	13.34	24.69	6.03	44.27	138.1	6077
11	Bacher	14.01	1.76	13.32	24.91	5.99	47.11	129.8	6166
12	Kazanina	14.44	1.73	13.23	25.06	5.91	50.17	132.4	6047
13	Naumenko	14.20	1.79	12.88	24.98	6.00	42.15	135.4	5971
14	Klavina	13.92	1.76	14.24	24.37	6.07	41.17	149.6	5932
15	Skujyte	14.44	1.76	16.35	25.76	5.86	47.57	138.6	6077
16	Chernyavskaya	13.85	1.76	12.76	25.03	5.99	37.83	129.4	5969
17	Prokhorova	13.87	1.82	13.36	23.99	6.46	43.60	128.3	6452
18	Roshchupkina	14.32	1.70	14.53	24.21	5.85	43.22	133.0	6034

Italian Olive Oils

- The composition of 8 fatty acids found in the lipid fraction of 572 Italian olive oils were recorded; the oils were from three regions.

	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic	arachidic	eicosenoic
1	1075	75	226	7823	672	36	60	29
2	1088	73	224	7709	781	31	61	29
3	911	54	246	8113	549	31	63	29
4	966	57	240	7952	619	50	78	35
5	1051	67	259	7771	672	50	80	46
6	911	49	268	7924	678	51	70	44
7	922	66	264	7990	618	49	56	29
8	1100	61	235	7728	734	39	64	35
9	1082	60	239	7745	709	46	83	33
10	1037	55	213	7944	633	26	52	30
...	...	...	...	...	...	...	...	...
568	1280	110	290	7490	790	10	10	2
569	1060	100	270	7740	810	10	10	3
570	1010	90	210	7720	970	0	0	2
571	990	120	250	7750	870	10	10	2
572	960	80	240	7950	740	10	20	2

Italian Wine

- Forina *et al.* (1986) collected data recording the chemical and physical properties of Barbera, Barola and Grignolino wines.

Chemical Properties		
Alcohol	Malic Acid	Ash
Alkalinity of Ash	Magnesium	Total Phenols
Flavonoids	Nonflavonoid Phenols	Proanthocyanins
Color Intensity	Hue	OD280/OD315 Of Diluted Wines
Proline		

- The aim of the study was to develop a method of classifying wines into type from their chemical properties.

US Arrests

- The number of arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states were recorded.

	Murder	Assault	UrbanPop	Rape
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
Connecticut	3.3	110	77	11.1
.....				
Wisconsin	2.6	53	66	10.8
Wyoming	6.8	161	60	15.6

Fisher's Iris Data

- This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris.
- The species are Iris setosa, versicolor, and virginica.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
...					
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
...					
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
103	7.1	3.0	5.9	2.1	virginica

Old Faithful Geyser

- The waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA were recorded for 272 consecutive eruptions.

	eruptions	waiting
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85
6	2.883	55
7	4.700	88
8	3.600	85
.....		
271	1.817	46
272	4.467	74

Resting Pulse

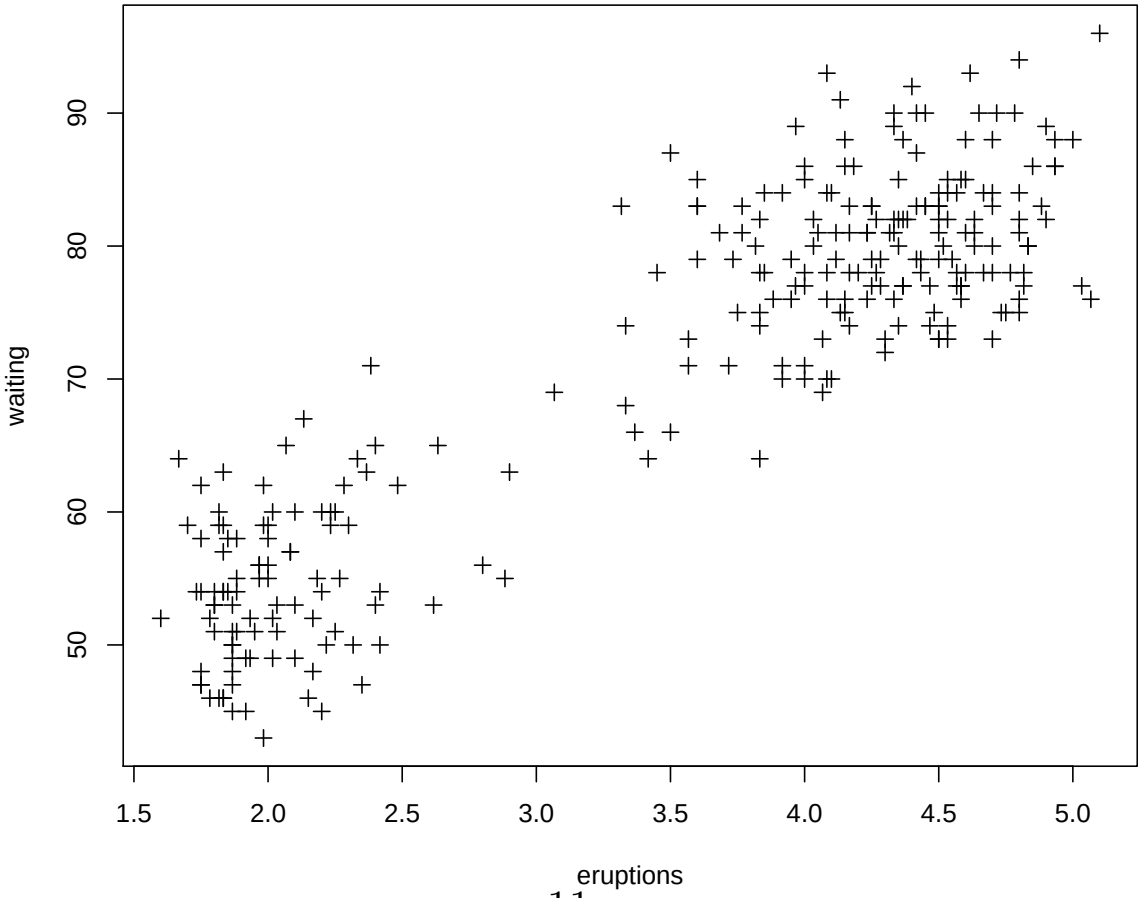
- A researcher is interested in understanding the effect of smoking and weight upon resting pulse rate (Low/High).

	PULSE	SMOKES	WEIGHT
1	Low	No	140
2	Low	No	145
3	Low	Yes	160
4	Low	Yes	190
5	Low	No	155
6	Low	No	165
7	High	No	150
8	Low	No	190
.....			
9	High	No	150
10	Low	No	108

- In this case, some of the variables are categorical.

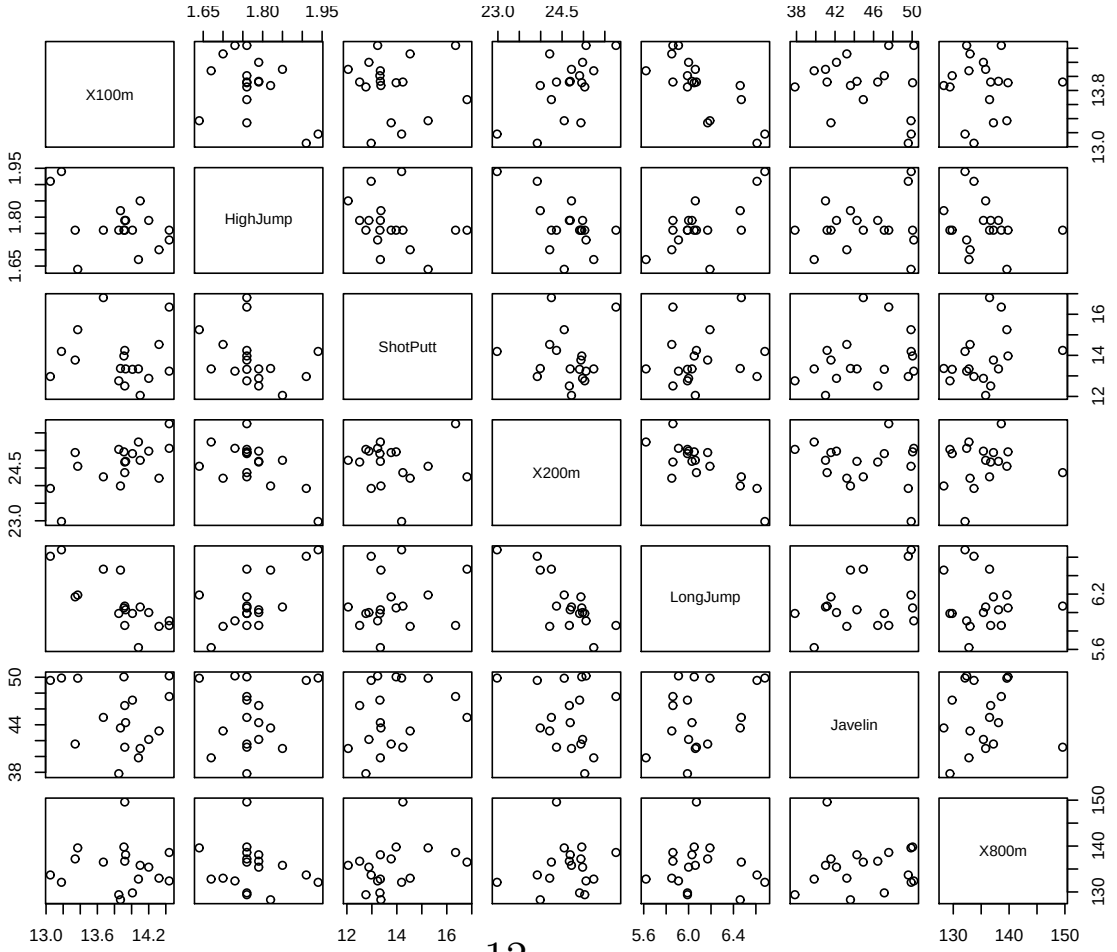
Graphical Summaries

- It is worth plotting your data to look for interesting structure.
- For two continuous variables, a scatter plot is a good choice.



Heptathlon Data: Scatter plot matrix

- A scatter plot matrix of the heptathlon data shows some relationships between the seven events.



Multivariate analysis

Our objective is to find structure in the data.

- **Dimension Reduction** – explain the data in a reduced number of dimensions, *e.g.*, principal component analysis and MDS;
- **Supervised methods** – assign labels to observations using some kind of **provided** structure within the data, *e.g.*, classification, or discriminant analysis.
- **Unsupervised methods** – using **only** the internal structure of the data to assign labels to observations *e.g.*, cluster analysis.

Data Reduction

- Multivariate data can have many variables recorded.
- Sometimes we can find a way of producing a reduced number of variables that contain most of the information of the original data.
- This is the aim of data reduction.
- **Example:** In the decathlon (heptathlon) data, the athletes are awarded points. The points are supposed to measure the overall ability of the athlete. Good athletes have high points and poor athletes have low points. The points score is a single number that is used to describe the athlete's performance in ten (seven) events.
- We will look at data-driven methods of reducing the dimensionality of the data that retains much of the information contained in the original data.
- The method of data reduction used depends on what information we wish to retain.

Principal Components Analysis

- Principal components analysis (PCA) finds linear combinations of the variables in the data which capture most of the variation in the data.
- The output from a PCA of the heptathlon data is as follows:

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
100m	0.460	0.095	0.264	-0.168	0.692	-0.286	-0.348
HighJump	-0.431	0.297	-0.076	-0.393	0.527	0.389	0.369
ShotPutt	-0.012	-0.697	0.141	0.462	0.392	0.208	0.288
200m	0.495	-0.026	0.097	-0.168	-0.161	0.810	-0.184
LongJump	-0.540	-0.088	-0.037	0.160	0.116	0.210	-0.785
Javelin	-0.238	-0.371	0.673	-0.542	-0.219	-0.103	0.004
800m	0.091	-0.520	-0.664	-0.506	0.046	-0.111	-0.100
Eigenv	3.099	1.5482	0.8667	0.6570	0.4533	0.2999	0.0754
Propor	0.443	0.221	0.124	0.094	0.065	0.043	0.011
Cumula	0.443	0.664	0.788	0.882	0.946	0.989	1.000

- What does this tell us?

Principal Components Analysis

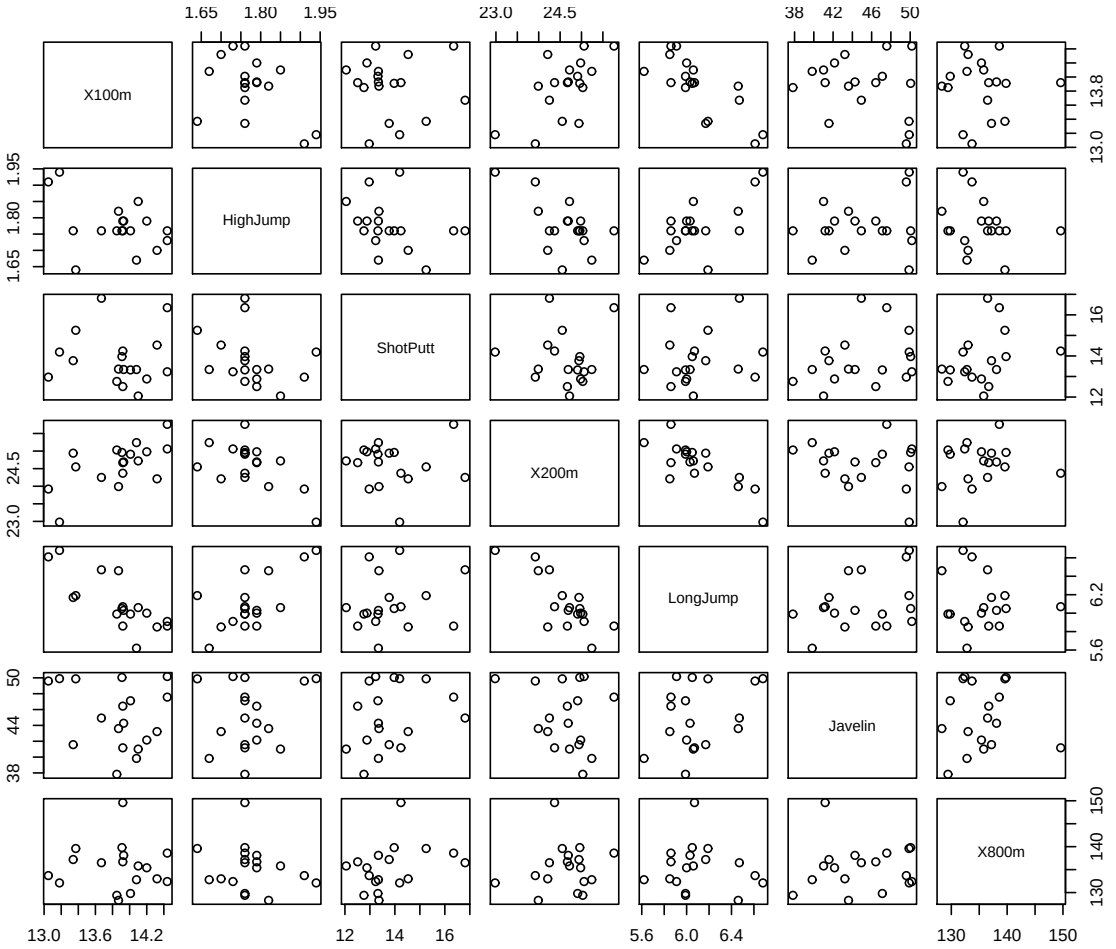
- The principal components are difficult to interpret because a small (quicker) running time is good, but a large (longer) throwing or jumping distance is also good.
- I replaced *time* by *-time* and re-ran the analysis.

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
100m	-0.460	-0.095	-0.264	0.168	-0.692	0.286	0.348
HighJump	-0.431	0.297	-0.076	-0.393	0.527	0.389	0.369
ShotPutt	-0.012	-0.697	0.141	0.462	0.392	0.208	0.288
200m	-0.495	0.026	-0.097	0.168	0.161	-0.810	0.184
LongJump	-0.540	-0.088	-0.037	0.160	0.116	0.210	-0.785
Javelin	-0.238	-0.371	0.673	-0.542	-0.219	-0.103	0.004
800m	-0.091	0.520	0.664	0.506	-0.046	0.111	0.100
Eigenv	3.099	1.5482	0.8667	0.6570	0.4533	0.2999	0.0754
Propor	0.443	0.221	0.124	0.094	0.065	0.043	0.011
Cumula	0.443	0.664	0.788	0.882	0.946	0.989	1.000

- What does this tell us?

Interpretation Of Principal Components

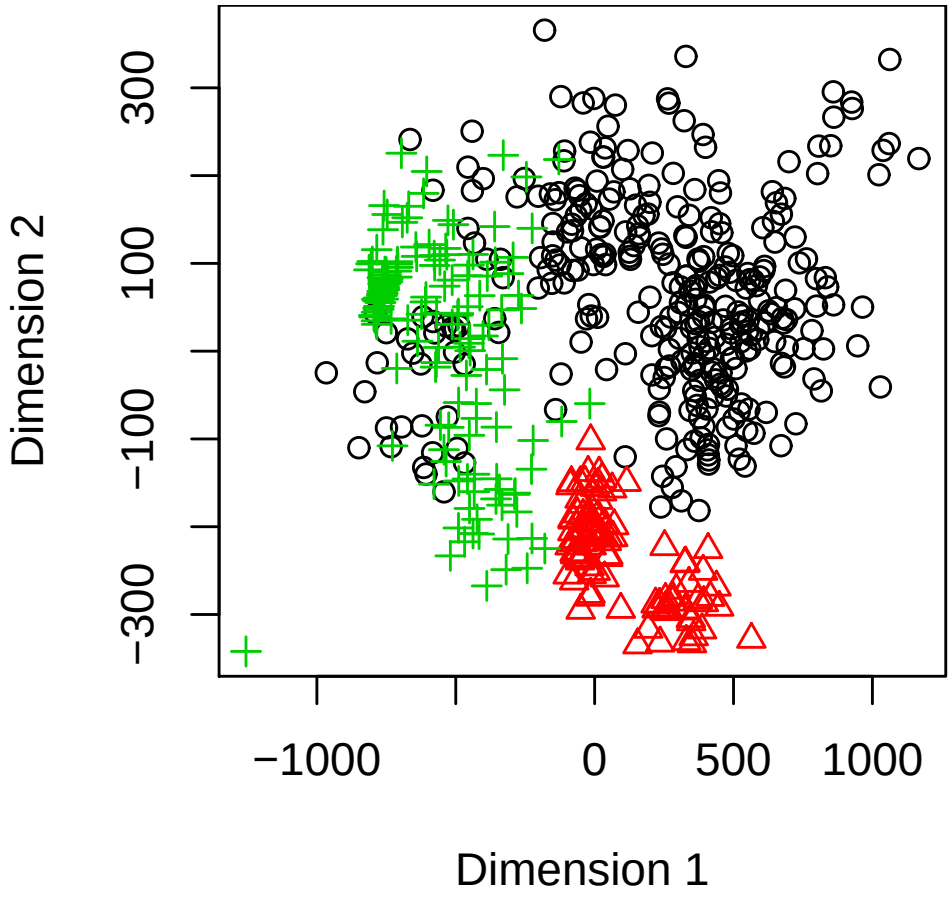
- The first principal component records some measure of overall ability.
- Note that the good athletes are on the left-hand side of the plot and the points tend to drop as we move right.

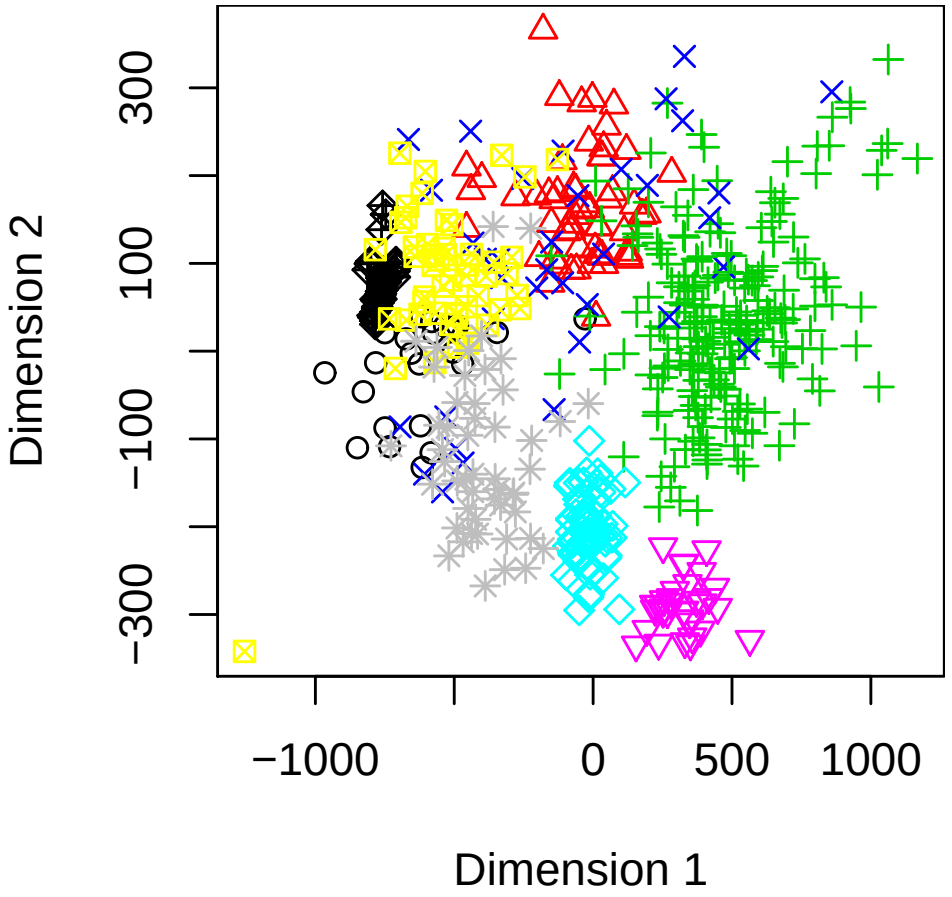


Discriminant Analysis/Classification

- The olive oil data is for oils collected from three regions (Southern Italy, Sicily, Northern Italy).
- Can we use the data to construct a rule that would help us to determine where new olive oil samples come from?
- This is the aim of discriminant analysis (or classification).
- In fact, we know the origin of the oils to an even greater accuracy (North-Apulia, Calabria, South-Apulia, Sicily, Inland-Sardinia, Coast-Sardinia, East-Liguria, West-Liguria, Umbria).
- Can we classify the oils to this fine an accuracy?

Olive Oil Data





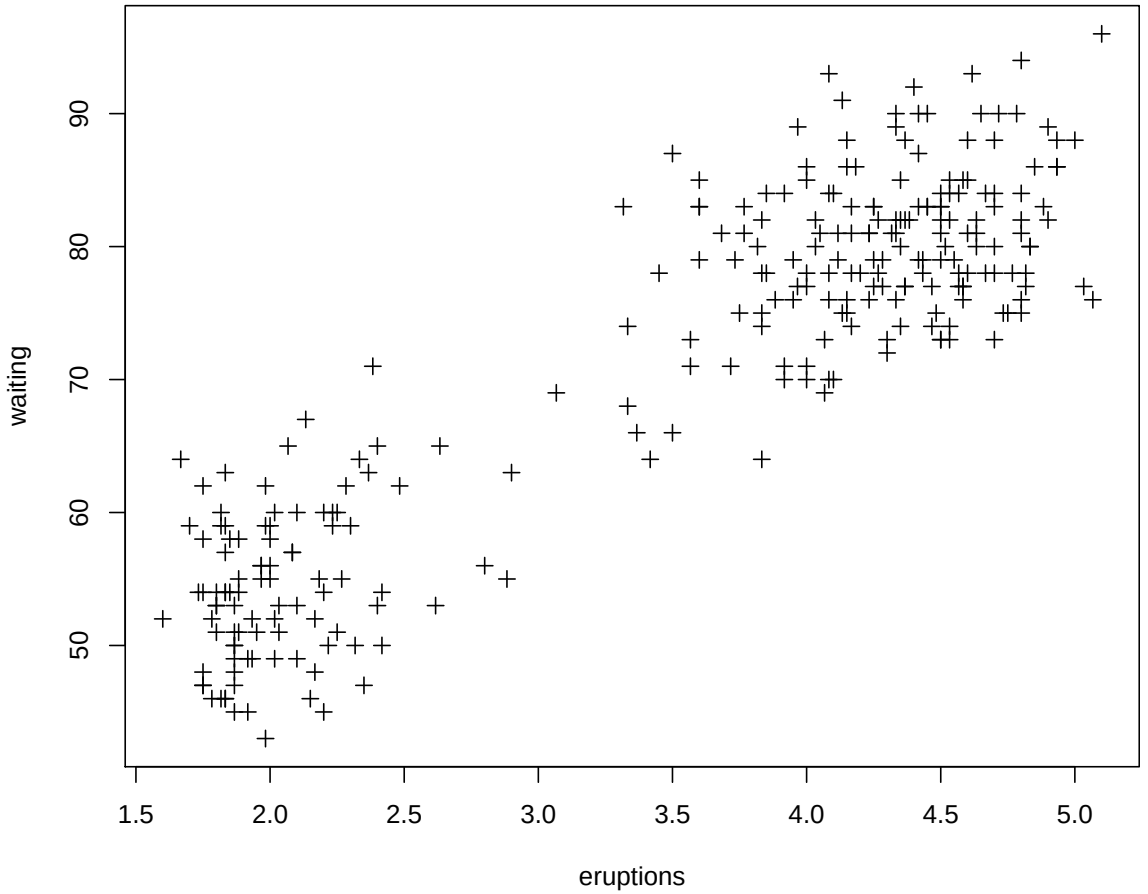
- We can classify to the three regions with almost 100% accuracy and to the nine areas with about 97% accuracy.

Cluster Analysis

- When doing discriminant analysis, we know that there are groups in the data. We want to be able to classify observations into the known groups.
- Cluster analysis is used to find unknown groups in the data. Once the existence and properties of groups have been established, we then want to classify observations into groups.

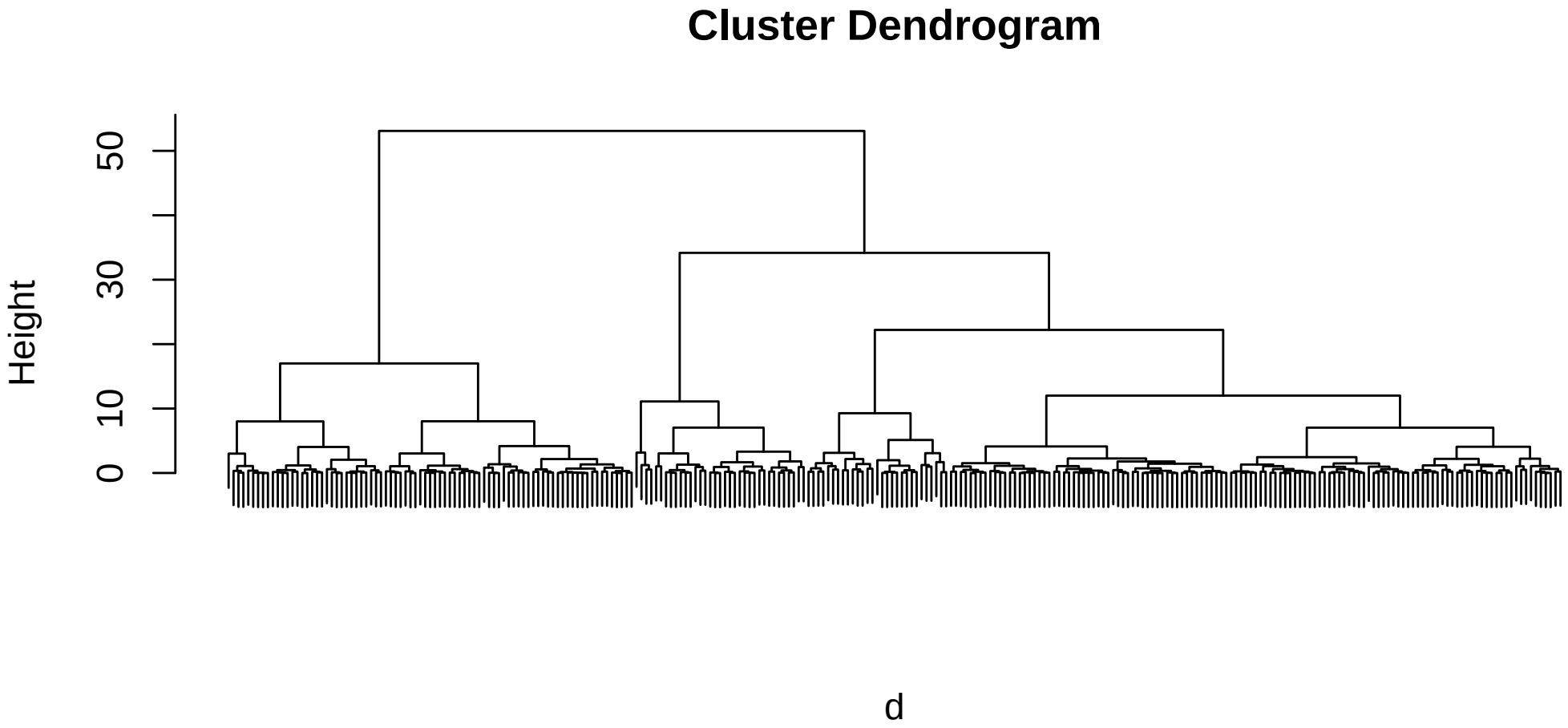
Old Faithful Data

- It looks like there are at least two different types of eruptions of the Old Faithful geyser.
- Let's see what cluster analysis tells us.



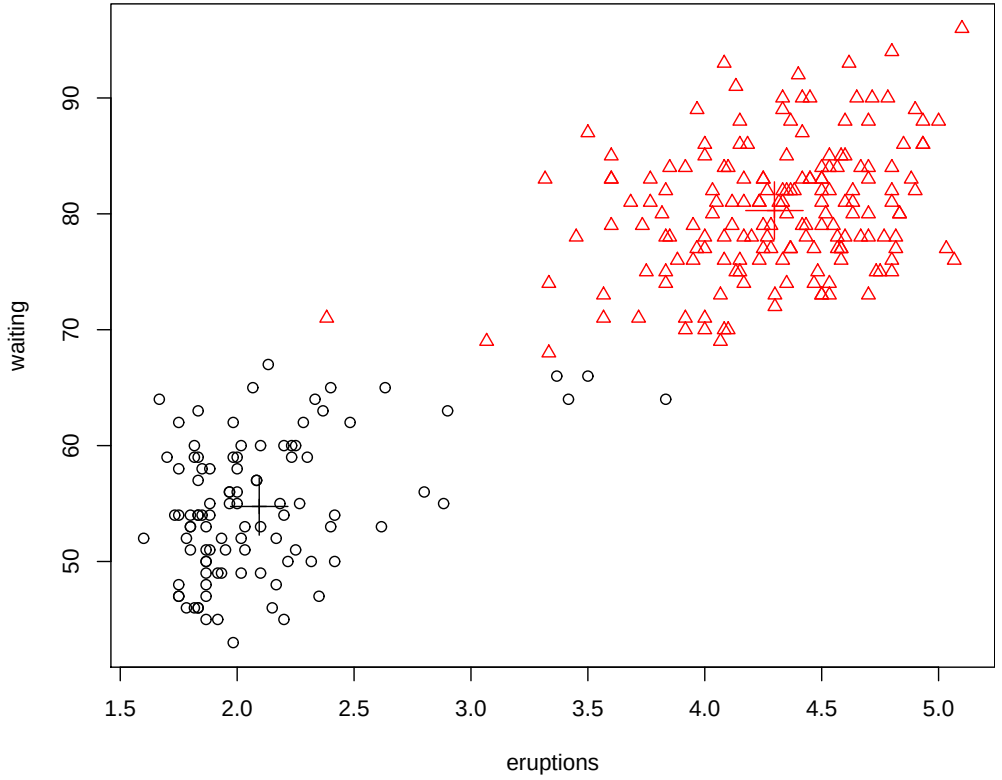
Hierarchical Clustering

- Hierarchical clustering builds a tree, where similar objects are joined low down and less similar objects are joined higher up.



k-Means Clustering

- k-means clustering assigns objects into k groups where the observations within a group are similar.
- A two group clustering of the Faithful data:

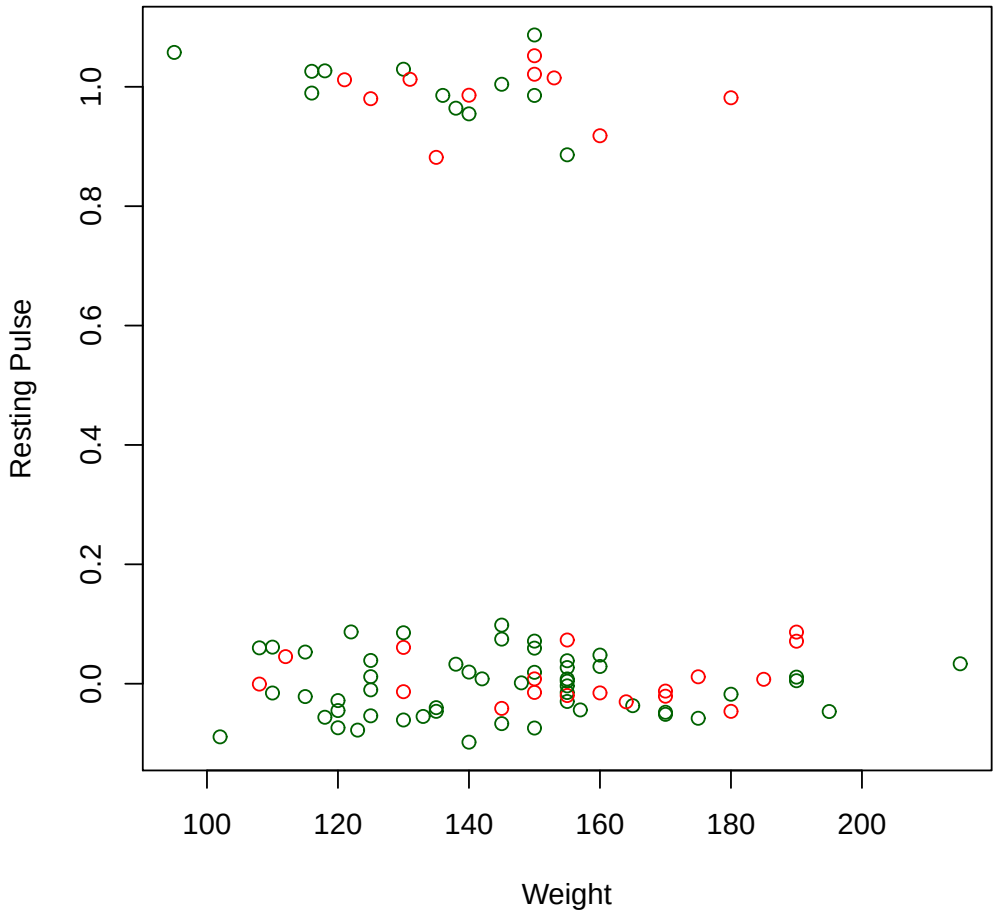


Logistic Regression

- Logistic regression is a special type of regression method where the response variable is a binary variable.
- Recall that in linear regression the response variable is continuous.
- **Examples:** Recall the resting pulse examples. In these cases, we can use a binary variable to indicate whether or not the response variable of interest (high resting pulse) occurs or not.
- Logistic regression can tell us which factors (smoking and weight) influence whether or not a patient has a high or low resting pulse, and by how much.

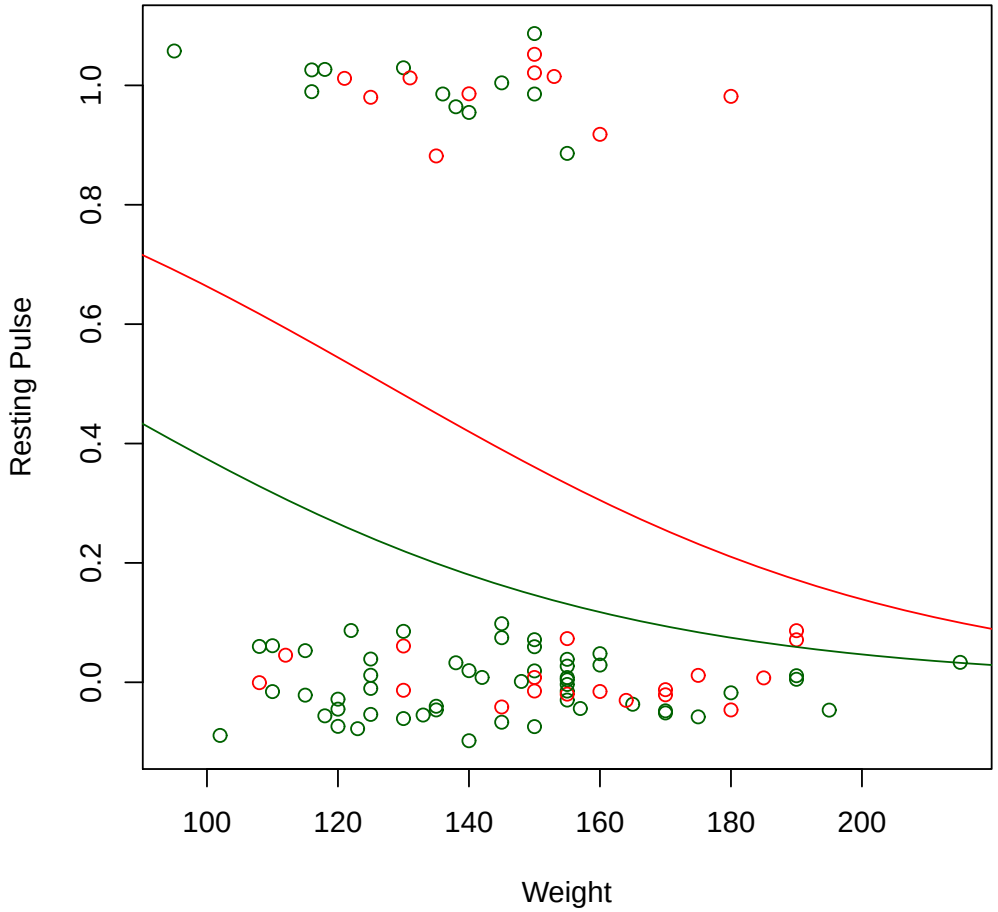
Logistic Regression

- 1 = High resting pulse, 0 = low; Green = smoker, Red = non-smoker; points are jittered to prevent overlap.



Logistic Regression

- Fitted lines – interpret these as the probability of the event occuring.



Conclusions and Questions

- Multivariate data are now easily available in very many fields of study. The examples we have seen should give you a feel for the material we will be covering in this module.
- Methods to analyse this kind of data can be organised into three topics:
  - Dimension reduction;
  - Classification;
  - Clustering.
- Can you think of an interesting data set you can apply the described methods to? What kind of questions could these methods help you answer?