Introduction

- Data Mining is the science of developing, identifying, and using suitable machine learning or statistical methods for extracting or uncovering useful information from the data.
- The main goal is to find important information from the data.
- Data Mining is multi-disciplinary, with contributions from diverse fields, *e.g.*
  - Computer science: artificial intelligence, machine learning, databases
  - Statistics
  - Engineering, Geographic Information Systems, and Commerce

Introduction

- There is no one-fits-all approach to data mining.
- It is important to first understand the property of the data and to understand the task at hand.
- Then, it is important to identify appropriate data mining method for the data and task at hand.

Introduction

Different terminology is used in different disciplines (computer science, statistics, machine learning, commerce, and engineering) for essentially the same concepts, *e.g.*

- Data mining ↔
  analytics, machine learning, big data analysis.
- Input ↔
  explanatory, predictor or independent variable, attribute, features
- Output ↔
  response, class, or dependent variable

Outline

Overview

- What is Visualisation?
- Why visualise?
- How to visualise?

Techniques

Outline

Overview

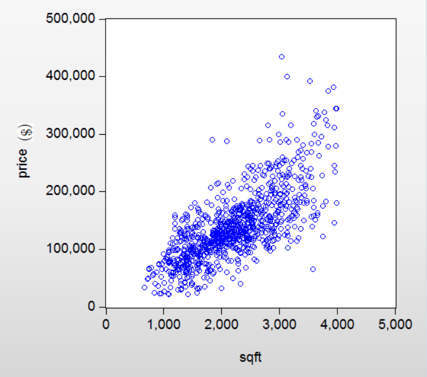
- What is Visualisation?
- Why visualise?
- How to visualise?

Techniques

Visualisation: What?

- Data visualisation is the display of information in a graphic or tabular format.
- The main motivation for using visualisation is that people can usually absorbs large amount of visual information and find patterns or outliers in it.
- Successful visualisation requires that the data (information) be converted into a visual format so that the characteristics of the data and the relationships among the data items or attributes can be analysed or reported.
- A key challenge in the data visualisation is to choose a technique or a method that makes the relationships of interest easily observable.

Simple Example



- Positive relationship between house price and house size.
- Variations of house price in big house is much bigger than in small house.

Visualisation: What?

In the context of data mining, what can be visualised?

- Raw or transformed data. For example: the log-transformation of the raw variable.
- Relationships between variables. For example, the relationship between house price and house size.
- Predicted output from the models. For example, forecast of the number of tourists from 2021-2025.
- Discrepancies (“errors”) between data and models.

Outline

Overview

- What is Visualisation?
- Why visualise?
- How to visualise?

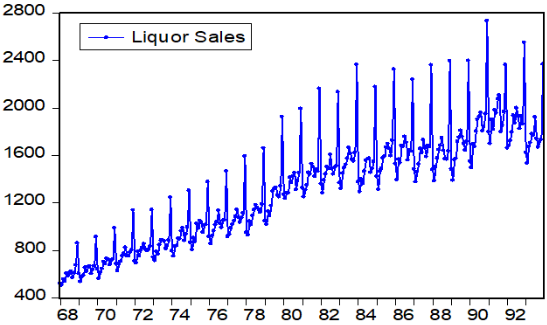
Techniques

Visualisation: Why?

Visualisation is important within data mining

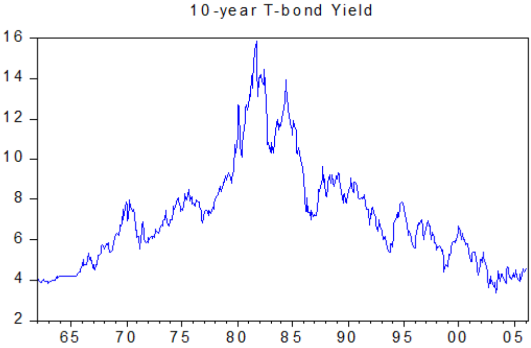
- to identify problems within raw data requiring “cleaning”, *e.g.* different units of measurment, extreme outliers
- to help with attribute selection
  - remove the input variable that does not have any relationship with the output variable.
- to help with model building because we may able to see the important patterns.
- to help with interpretation of results
- to communicate results to others.

Example: Monthly Liquor Sales, Time Series Plot



- trend (increase over time on average)
- seasonality (reoccuring pattern: Dec high and Feb low).

Visualisation, Time Series Plot



- Persistent (the yesterday value is quite close to today value).
- Random trend (subsamples are very different).

Quotes from John Tukey, a famous mathematician and statistician

- “Visualization is often used for evil - twisting insignificant data changes and making them look meaningful. Don’t do that if you want to be my friend. Present results clearly and honestly. If something isn’t working - those reviewing results need to know.”
- “Numerical quantities focus on expected values, graphical summaries on unexpected values.”

Outline

Overview

What is Visualisation?

Why visualise?

How to visualise?

Techniques

Visualisation: How?

Principles of graphical excellence

- Avoid distorting the meaning of data
- Display a lot of information within a small space
- Integrate graphics with verbal and statistical descriptions
- Compare the properties of multiple datasets

Visualisation: How?

Three categories of graphical properties which are perceived quickly without conscious intervention:

- Colour
- Form (shape, size)
- Movement (animation, spinning, zooming), some of this but not too many.

Outline

Overview

Techniques

- Quantitative Variables
- Categorical Variables
- Two

Example Dataset: Anderson's Iris Data

Anderson, Edgar (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, **59**, 2-5.

Measurements on 150 flowers from three species of iris (50 each):

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
5.8	4.0	1.2	0.2	setosa
5.2	3.5	1.5	0.2	setosa
4.5	2.3	1.3	0.3	setosa
6.5	2.8	4.6	1.5	versicolor
6.2	2.2	4.5	1.5	versicolor
5.5	2.4	3.7	1.0	versicolor
5.7	3.0	4.2	1.2	versicolor
6.7	2.5	5.8	1.8	virginica
7.7	2.8	6.7	2.0	virginica
7.7	3.0	6.1	2.3	virginica
5.9	3.0	5.1	1.8	virginica

Outline

Overview

Techniques

- Quantitative Variables
  - One
  - Two (or so)
  - Many
- Categorical Variables
- Two

Outline

Overview

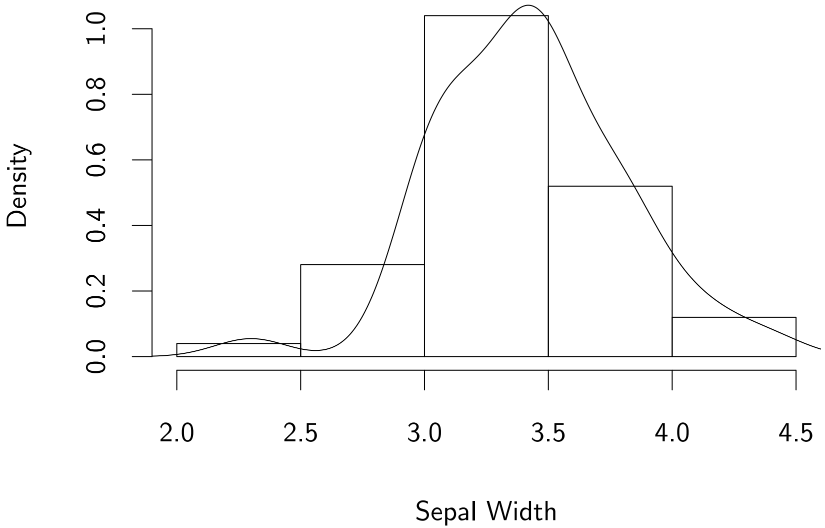
Techniques

- Quantitative Variables
  - One
  - Two (or so)
  - Many
- Categorical Variables
- Two

Histogram

- A histogram describes how data are distributed
- A histogram shows the frequency (vertical axis) of observations within each ‘bin’ along the horizontal axis (usually of constant width).
- A compromise must be made between too many bins (bumpy plot) and too few (interesting detail cannot be seen).
- Interpretation: allows detection of outliers (unusually high or low data values), and shows skewness (lack of symmetry) of distribution.
- R code: `hist`, or `density` to fit a curve to show the shape.

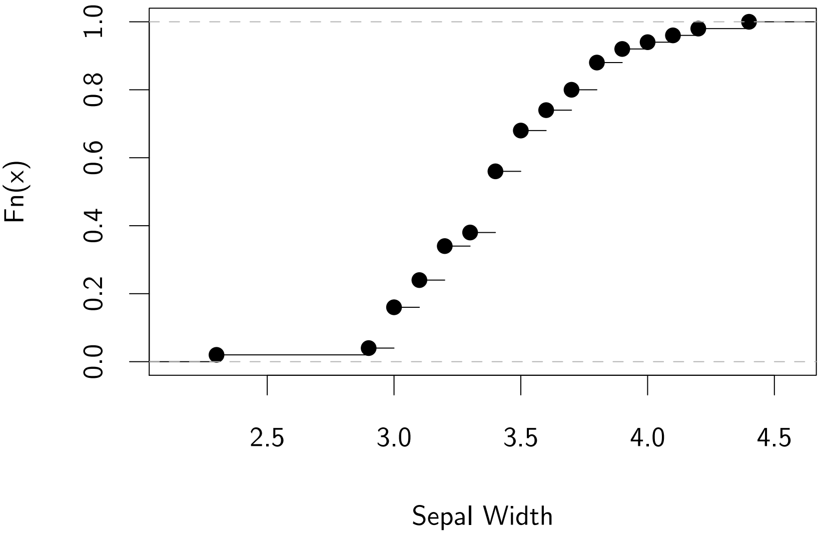
Histogram (overlaid density fit)



Empirical Cumulative Distribution Function

- A plot of the *empirical cumulative distribution function* (ecdf) displays the proportion of data values $\leq x$ for x (horizontal axis) within the range of the data.
- This takes the form of a *step* function which jumps by k/n at locations where there are k tied data values.
- The plot is non-decreasing, steeper in regions where data are concentrated, typically S-shaped.
- R code: `ecdf(x)`

Empirical Cumulative Distribution Function



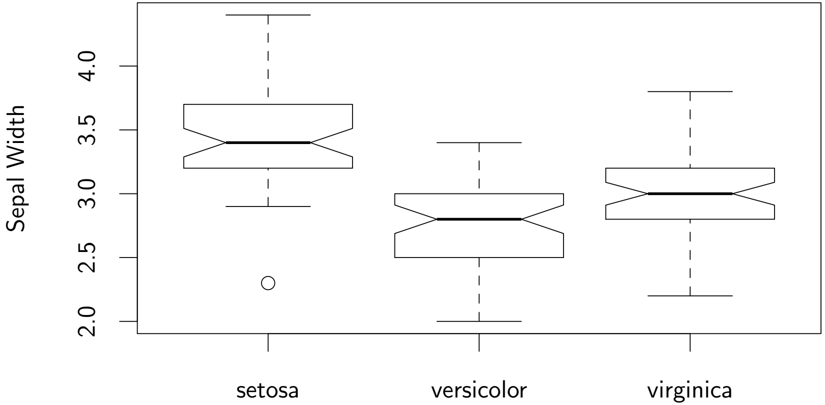
Quartiles

- The lower or first quartile of a single quantitative variable exceeds 25% of data values.
- The upper of third quartile exceeds 75% of data values.
- The interquartile range IQR is the difference between the quartiles.
- R code: `quantile(x,c(0.25,0.75))`

Box Plot

- A simple ‘box-and-whisker’ plot shows a box between the lower and upper quartiles, and whiskers extending to the extreme data values.
- Usually outliers lying more than 1.5× IQR above upper quartile or below lower quartile are displayed as separate points - then whiskers only extend to most extreme non-outliers.
- Parallel boxplots are useful for comparing different groups.
- R can add notches to show 95% confidence intervals for the median.

Box Plot



Outline

Overview

Techniques

- Quantitative Variables
  - One
  - Two (or so)
  - Many
- Categorical Variables
- Two

Scatterplots

- A relationship between two quantitative variables can be displayed as a *scatterplot*.
- R code: `plot(x,y)`
- Plotting symbols (R `plot()` optional argument: `pch`) and colours (argument: `col`).

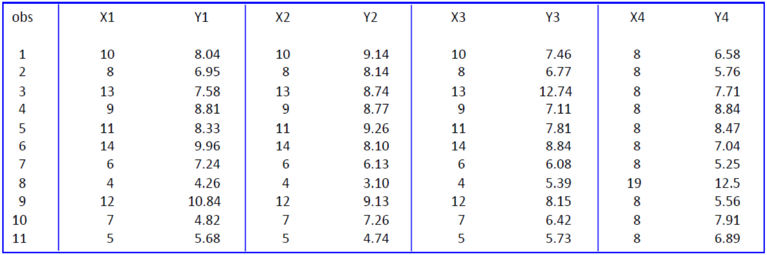
Scatterplots: What to look for?

- The direction of the relationship between two variables x and y. Positive? Negative? or No relationship?
- The variability of the y for different value of x. For example, variations of house price in big house is much bigger than in small house.
- Unusual data points/outliers

Classic Example: Anscombe's Quartet

Anscombe, Francis J. (1973) Graphs in statistical analysis. *American Statistician*, **27**, 17-21.

Four bivariate (x, y) datasets with 11 observations each.



obs	X1	Y1	X2	Y2	X3	Y3	X4	Y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.10	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.10	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

Four bivariate (x, y) datasets with 11 observations each, *all* having the following properties:

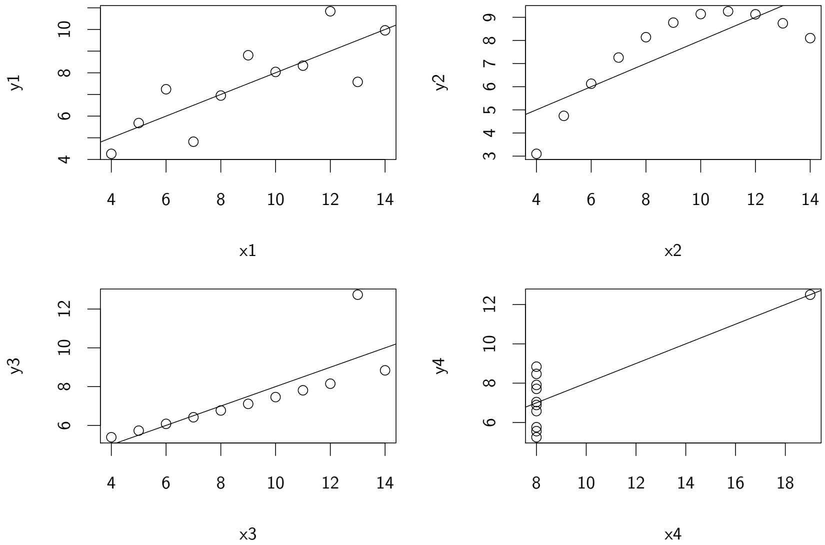
- $\bar{x} = 9$
- $s_x = 3.3166$
- $\bar{y} = 7.5009$
- $s_y = 2.0316$
- $r = 0.8164$

Four bivariate (x, y) datasets with 11 observations each, *all* having the following properties:

- $\bar{x} = 9$
- $s_x = 3.3166$
- $\bar{y} = 7.5009$
- $s_y = 2.0316$
- $r = 0.8164$

How different can they be?

Very different!



Outline

Overview

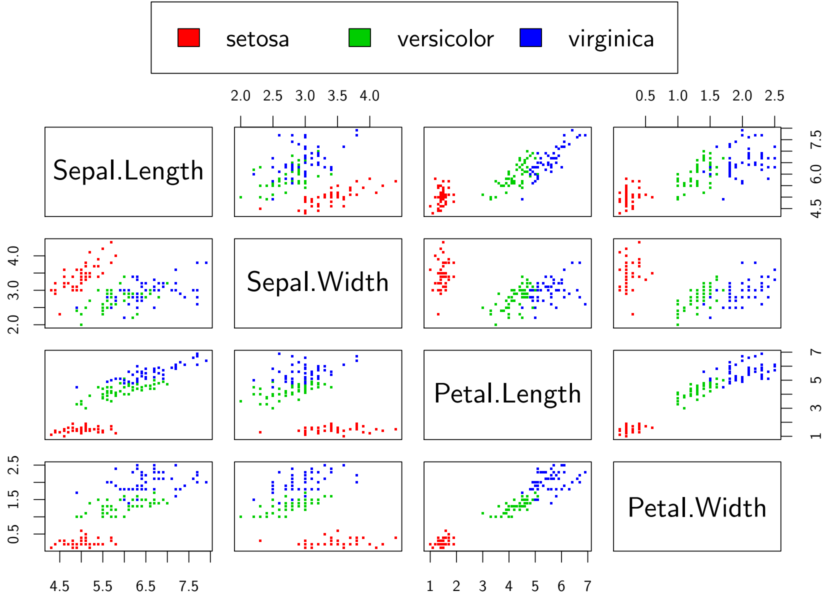
Techniques

- Quantitative Variables
  - One
  - Two (or so)
  - Many
- Categorical Variables
- Two

Scatterplot Matrix

- A *scatter plot matrix* is an array of scatterplots with aligned axes, with each attribute plotted against each other attribute (except itself).
- R code: `pairs(x)` (`x` has multiple columns)

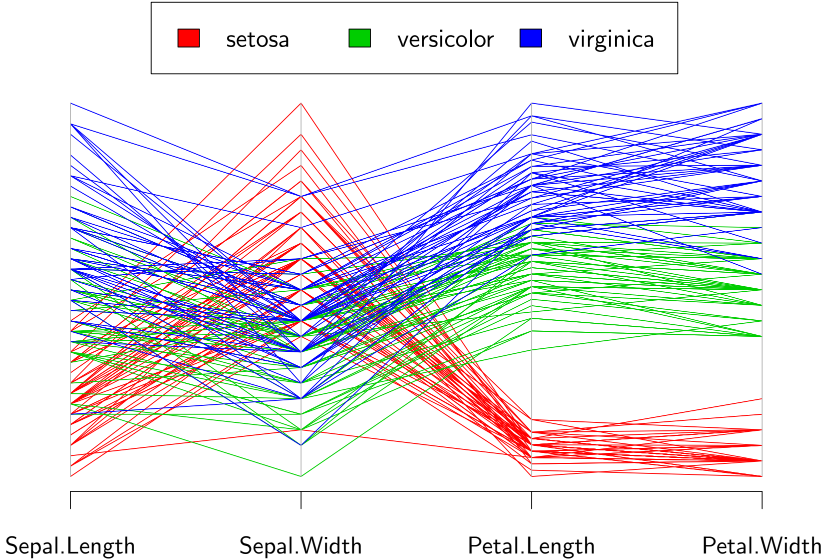
Scatterplot Matrix



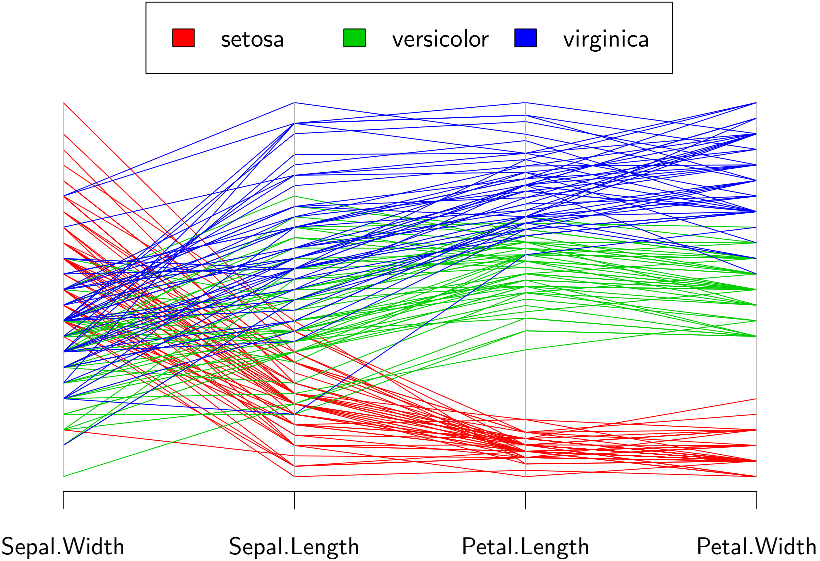
Parallel Coordinate Plot

- Visualise many variables, usually with a relatively small sample.
  1. Plot each observation from each variable, with variables on the horizontal axis and values (standardised) on the vertical axes.
  2. Connect points corresponding to the same observation with a line.
- R code: `parcoord(x)` (`x` has multiple columns)
- Ordering of the variables can be important.

Parallel Coordinate Plot (ordering 1)



Parallel Coordinate Plot (ordering 2)



Outline

Overview

Techniques

- Quantitative Variables
- Categorical Variables
  - One
- Two

Example dataset: Hair and eye colour of statistics students

Friendly, M. (1992a) Graphical methods for categorical data. *SAS User Group International Conference Proceedings*, **17**, 190-200. based on Snee, R. D. (1974) Graphical display of two-way contingency tables. *The American Statistician*, **28**, 9-12.

- Data from 592 students, across three categorical variables:
  - Hair colour: Black, Brown, Red, Blond
  - Eye colour: Brown, Blue, Hazel, Green
  - Sex: Male, Female

Male:

	Brown	Blue	Hazel	Green
Black	32	11	10	3
Brown	53	50	25	15
Red	10	10	7	7
Blond	3	30	5	8

Female:

	Brown	Blue	Hazel	Green
Black	36	9	5	2
Brown	66	34	29	14
Red	16	7	7	7
Blond	4	64	5	8

Outline

Overview

Techniques

Quantitative Variables

Categorical Variables

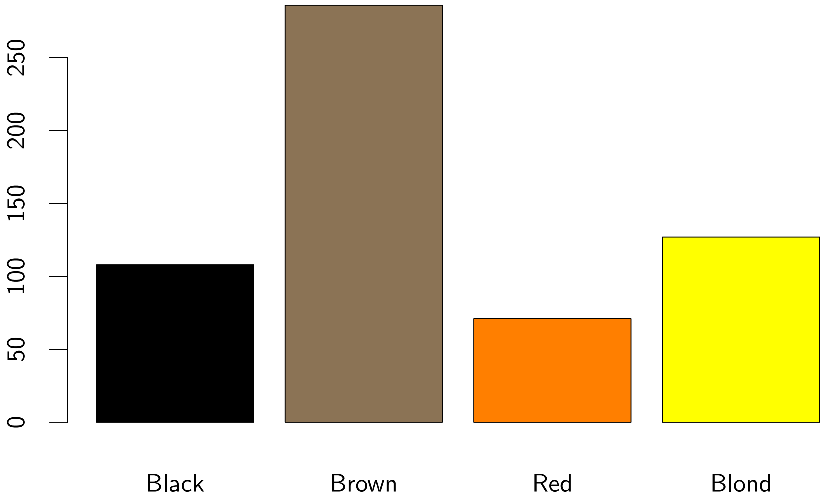
One

Two

Barplots

- Distinguished from histograms by having a categorical horizontal axis.
- Generally, a small gap between bars.
- Can be of frequencies or proportions.
- R code: `barplot(x)` (`x` a vector)

Barplot of hair colour



Outline

Overview

Techniques

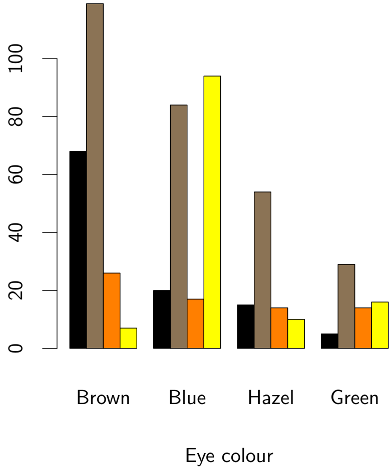
- Quantitative Variables
- Categorical Variables
- Two

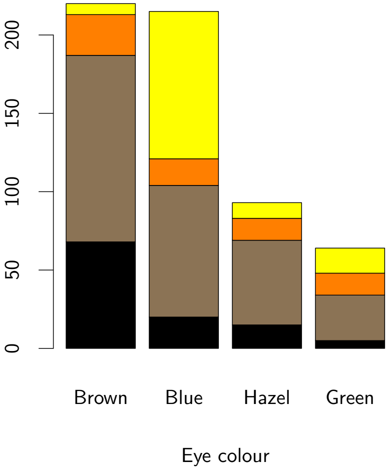
Bar plots for two variables

- Bars can be stacked or grouped. R code: `barplot(x, beside=TRUE/FALSE)` (`x` a matrix)

Bar plots for two variables

- Bars can be stacked or grouped. R code: `barplot(x, beside=TRUE/FALSE)` (`x` a matrix)
- Here, bar colour = hair colour.

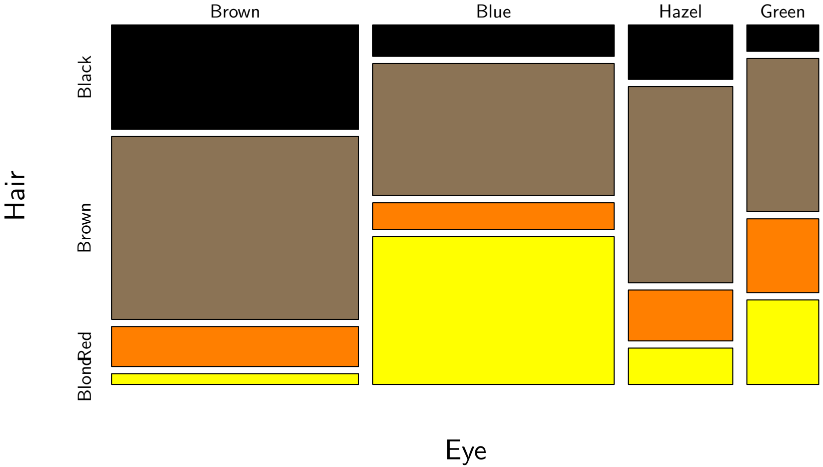




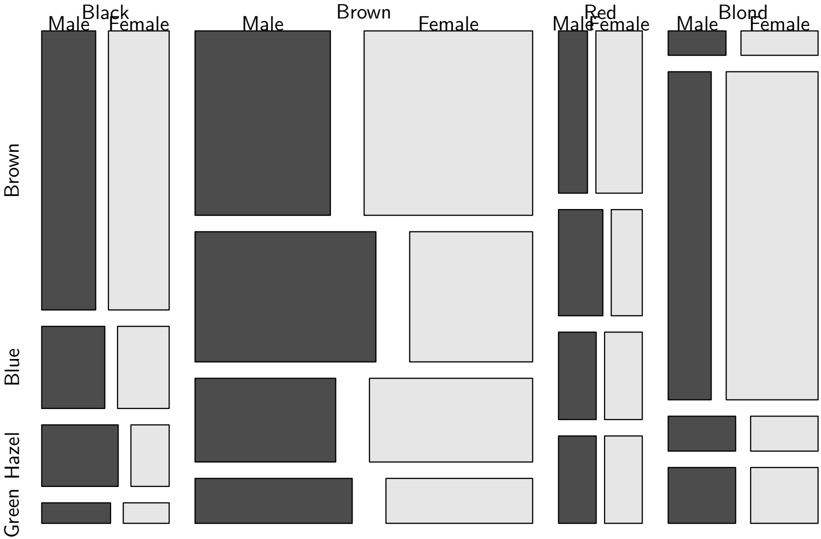
Mosaic Plots

- Like a stacked bar plot, except,
  - All bars are scaled to have the same total height.
  - All bars have *width* corresponding to the fraction of the data in the stacked bar.
  - Can be further subdivided.
- Display *conditional proportions* of one category within another.
- R code: `mosaicplot(x)` (`x` a matrix)
- Colours can be used to visually distinguish categories or to identify disproportionate ones.

Mosaic Plots (two variables)



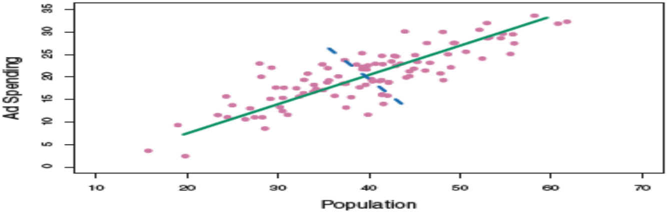
Mosaic Plots (three variables)



Principal Component Analysis (PCA, Dimensional Reduction Technique)

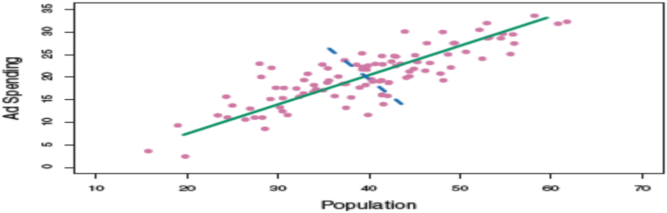
- Suppose that we wish to visualise n observations with measurements on a set of p features, $X_1, ..., X_p$, as part of an exploratory data analysis.
- We can examine two-dimensional scatterplots of the data, each of which contains the n observations measurements on two of the features.
- However, there are $p(p-1)/2$ such scatterplots. If p is large, then it will certainly not be possible to look at all of them.
- A better method is required to visualise the n observations when p is large.

Principal Components Analysis



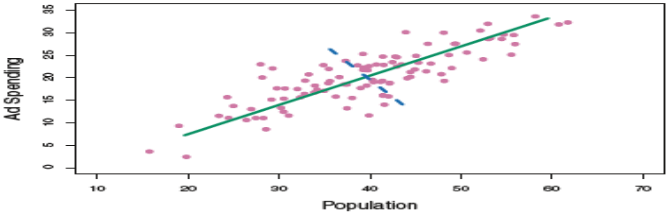
- Principal Component Analysis (PCA) is a popular approach for deriving a low-dimensional set of features from a large set of variables.
- PCA is especially valuable when we have subset of measurements that are measured on the same scale and are highly correlated.
- It provides a few variables (often as few as three) that are weighted linear combinations of the original variables, and that retains the majority of the information of the full original set.
- PCA is intended to use with numerical variables. For categorical variables, other methods such as correspondence analysis are more suitable (not cover here).

Principal Components Analysis



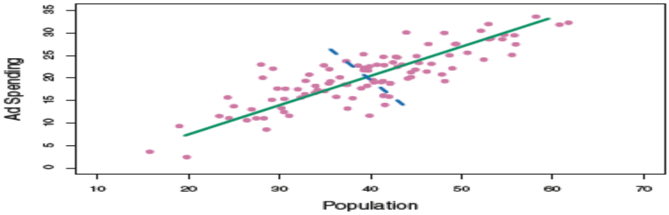
- The figure shows the population size in tens of thousands of people and ad spending for a company in thousands of dollars for 100 cities.
- They seem to have quite strong positive relationship.

Principal Components Analysis



- Can we use this fact, to reduce the number of variables, while retaining the majority of the information?
- Since there is redundancy in the information that the two variables contain, it might be possible to reduce the two variables to a single variable without losing too much information.

Principal Components Analysis



-
- The idea in PCA is to find a linear combination of the two variables that contains most of the information so that the new variable can replace the two original variables.
- The figure shows the population size in tens of thousands of people and ad spending for a company in thousands of dollars for 100 cities.
- The green solid line is the direction along which there is greatest variability in the data or in which the data vary the most. That is the first principal component.
- The blue dashed line indicates the second principal components.

Example: The advertising dataset

- The principal component is given by this formula

$$Z_1 = 0.83 \times (pop - \overline{pop}) + 0.54 \times \left(ad - \overline{ad}\right)$$

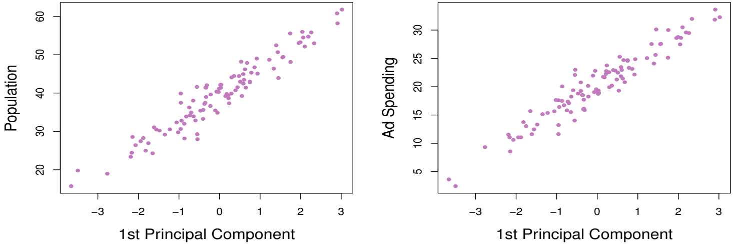
- The idea is that out of every possible linear combination of pop and ad such that $\phi_{11}^2 + \phi_{21}^2 = 1$, the above combination yields the highest variance.
- The linear combination in which

$$Var\left(\phi_{11} \times (pop - \overline{pop}) + \phi_{21} \times \left(ad - \overline{ad}\right)\right)$$

is maximised.

- The two loadings are both positive and have similar size, so Z_1 is almost an average of two variables.

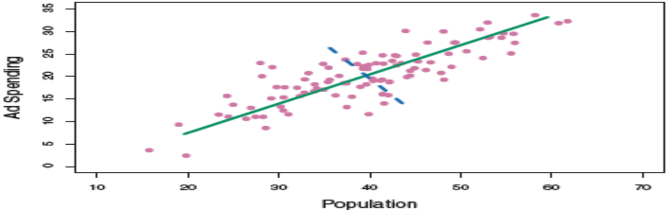
Example: The advertising dataset



- The plots show a strong relationship between the first principal component and the two predictors.
- In other words, the first principal component appears to capture most of the information contained in the pop and ad predictors.

Principal Components Analysis

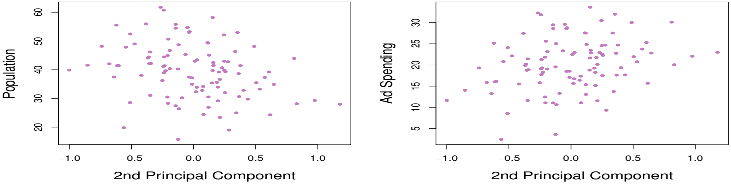
- In general, we can construct up to p distinct principal components.
- The second principal component is the linear combination of $X_1, ..., X_p$ that has maximal variance out of all linear combinations that are uncorrelated with the first principal component Z_1.
- The second principal component direction is given as a dashed blue line.



-

Example

- Since the advertising data has two predictors, the first two principal components contain all the information that is in pop and ad.
- The second principal component has little relationship with the two predictors.
- This indicates that the second principal component captures far less information compared to the first principal component.
-



Principal Component Analysis (PCA)

- When faced with a large set of correlated variables, principal components allow us to summarise this set with a smaller number of representative variables that collectively explain most of the variability in the original set.
- PCA extracts the main directions (Principal Components) across which the original data are highly variable.
  - First PC direction has the most variation across it.
  - Second PC direction has the leftover variation.
  - Etc.
- R code: `prcomp(x)` (`x` has multiple columns)

PCA Interpretation

- How much variability there is across a component `plot(prcomp(x))` tells how quickly the importance of PCs drops off after the first.
- Plotting data on the transformed scale via `biplot(prcomp(x))` tells you which dimensions are important, and which dimensions contain independent information.
  - Dimensions “pointing” the same way are “redundant”.

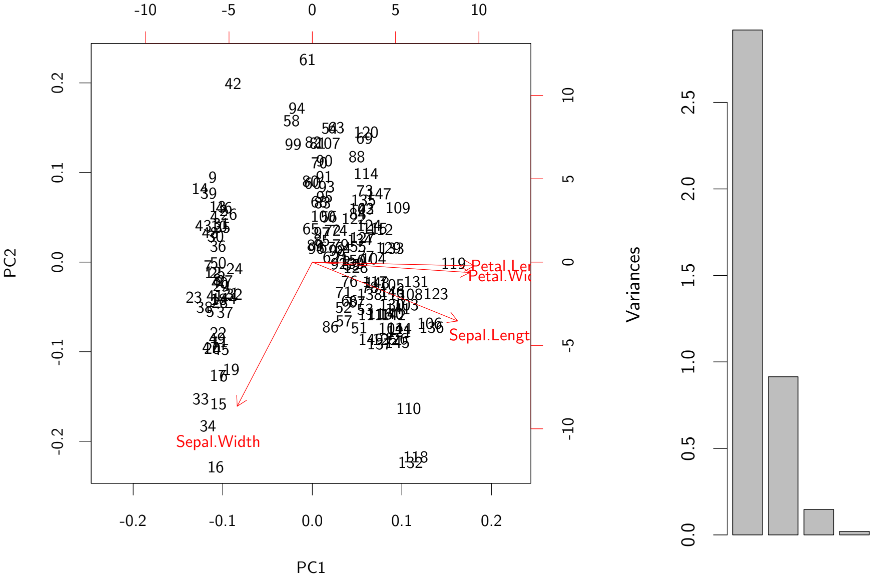
Example Dataset: Anderson's Iris Data

Anderson, Edgar (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, **59**, 2-5.

Measurements on 150 flowers from three species of iris (50 each):

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
5.8	4.0	1.2	0.2	setosa
5.2	3.5	1.5	0.2	setosa
4.5	2.3	1.3	0.3	setosa
6.5	2.8	4.6	1.5	versicolor
6.2	2.2	4.5	1.5	versicolor
5.5	2.4	3.7	1.0	versicolor
5.7	3.0	4.2	1.2	versicolor
6.7	2.5	5.8	1.8	virginica
7.7	2.8	6.7	2.0	virginica
7.7	3.0	6.1	2.3	virginica
5.9	3.0	5.1	1.8	virginica

PCA Example


PC1
PC2
-0.2
-0.1
0.0
0.1
0.2
-10
-5
0
5
10
Petal.Len
Petal.Wid
Sepal.Lengt
Sepal.Width
61
42
94
58
99
120
69
88
114
109
119
131
123
106
136
110
118
132
9
14
39
24
23
33
15
19
34
16
86
51
Variances
0.0
0.5
1.0
1.5
2.0
2.5