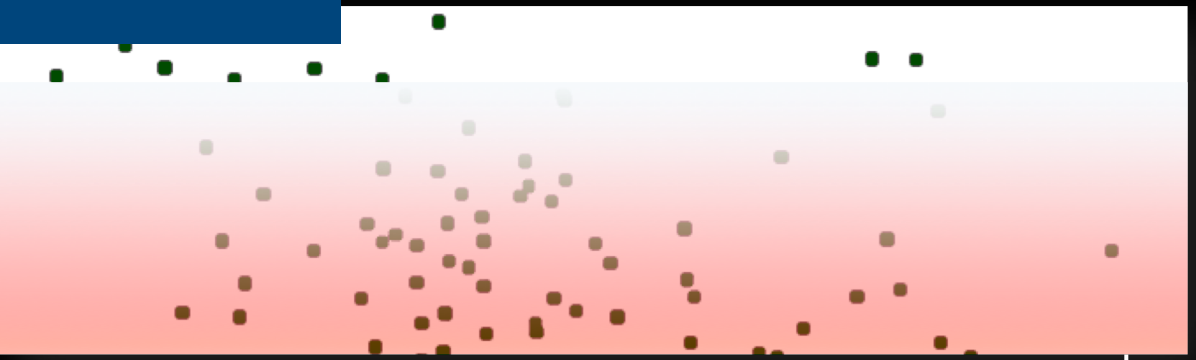
THE UNIVERSITY OF MELBOURNE



GEOM90007

SPATIAL VISUALISATION

LECTURE 4:
STATISTICAL FOUNDATION

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

http://www.bom.gov.au/vic/observations/melbournemap.shtml

OVERVIEW

A. Getting started: Terminology and data types

B. Methods for analysing numerical data – IGNORING LOCATION

C. Methods for analysing numerical data – CONSIDERING LOCATION

A. GETTING STARTED

Recall the first steps of the visualisation pipeline (lecture 1):

1. Data modelling:
   - User structures the data to facilitate visualisation

2. Data selection:
   - User or software identifies data to be visualised

Both are important steps in statistical analysis.

BASIC DATA CONCEPTS

A typical dataset for visualisation consists of:

- n records (,

with values for:

- m variables (,

A variable may be:

- *Independent*
  - ✓ date, time or location
- *Dependent:*
  - ✓ temperature

STATISTICAL DATA TYPES

Each observation represents a single piece of information.

Numerical

i. Discrete (integer values)
ii. Continuous (real values)

✓ Scatterplots and line charts

Categorical

i. Nominal (classes can not be ranked)
   - Count only (not order or measure) (e.g., gender)

ii. Ordinal (classes can be ranked) (e.g., house numbers)
   - Can count and order (not measure)

✓Bar charts

Australian Bureau of Statistics (2016)

DESCRIPTIVE VS INFERENTIAL STATISTICS

Descriptive

- Characteristics of a sample
- Data driven

Inferential

- Inference about a population from a sample
- Hypothesis testing
- Typically model driven

B. Methods for analysing numerical data – IGNORING LOCATION

1. Individual attribute
2. Two or more attributes
3. Exploratory data analysis

Example data set:
Australian Bureau of Statistics (ABS) Census 2011 – Employment
Download from: ABS Stat, http://stat.abs.gov.au/

PLEASE NOTE: DATA USED FOR ILLUSTRATIVE PURPOSES ONLY

1. INDIVIDUAL ATTRIBUTE

i. Tables

- **Raw tables**
  - Simplest form of tabular display
  - Data for a value of interest may be ordered low to high
  - Easy to find duplicates and potential outliers by visually scanning the columns
  - Range = Maximum(data) – Minimum(data)

1. INDIVIDUAL ATTRIBUTE

1. Clean data, format and sort using R

i.

	City_name	% Employed	% Unemployed	% Not_in_labour_force	% High_school	Population
1	Karratha	85.01	2.05	12.94	48.46	16474
2	Emerald	81.76	2.03	16.21	45.03	13218
3	Port Hedland	79.74	3.16	17.09	43.89	13771
4	Alice Springs	75.37	2.38	22.25	45.55	25186
5	Mount Isa	74.81	3.37	21.82	42.00	20570
6	Kalgoorlie - Boulder	74.75	3.35	21.90	38.11	30841
7	Darwin	74.17	2.75	23.09	48.10	106257
8	Broome	73.13	3.09	23.78	44.84	12762
9	Ellenbrook	72.56	2.87	24.57	45.51	28801
10	Gladstone - Tannum Sands	70.84	3.09	26.06	39.62	41967
11	Canberra - Queanbeyan	70.71	2.58	26.71	61.74	391644
12	Mackay	70.21	2.70	27.09	38.11	77293
13	Townsville					
14	Gisborne - Ma					
15	Singleton					
16	Highfields	67.18	2.52	30.30	40.87	16820
17	Torquay	67.16	2.68	30.16	52.65	15042
18	Cairns	66.66	4.83	28.51	46.05	133912
19	Perth	65.20	3.28	31.52	48.25	1670952
20	Bunbury	64.76	3.31	31.94	34.15	65607

igh
ally

2. Calculate data ranges using R

		% Employed	% Unemployed	% Not_in_labour_force	% High_school	Population
1	Minimum	40.68	2.03	12.95	18.14	10940
2	Maximum	85.01	5.49	56.12	61.74	4028524

1. INDIVIDUAL ATTRIBUTE

i. Tables

- **Grouped frequency table**
  1. Decide on how many groups (classes)
  2. Calculate class width using a basic formula

$$Width\ (basic) = \frac{Range}{Number\ of\ classes}$$

Determine lower and upper (-1) limits for each class

3. Count the number of observations within each class

1. INDIVIDUAL ATTRIBUTE

ii. Numerical summaries

- **Measures of central tendency**
  - Mode
    - The most frequently occurring value (for nominal data)
  - Median
    - The middle value in an ordered set of data (50^{th} percentile)
  - Mean
    - The 'average value'

Sample

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Population

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

1. INDIVIDUAL ATTRIBUTE

ii. Numerical summaries

- **Measures of dispersion**
  - Interquartile range
    - The absolute difference between the 25th and 75th percentiles
  - Standard deviation
    - Dispersion from the mean

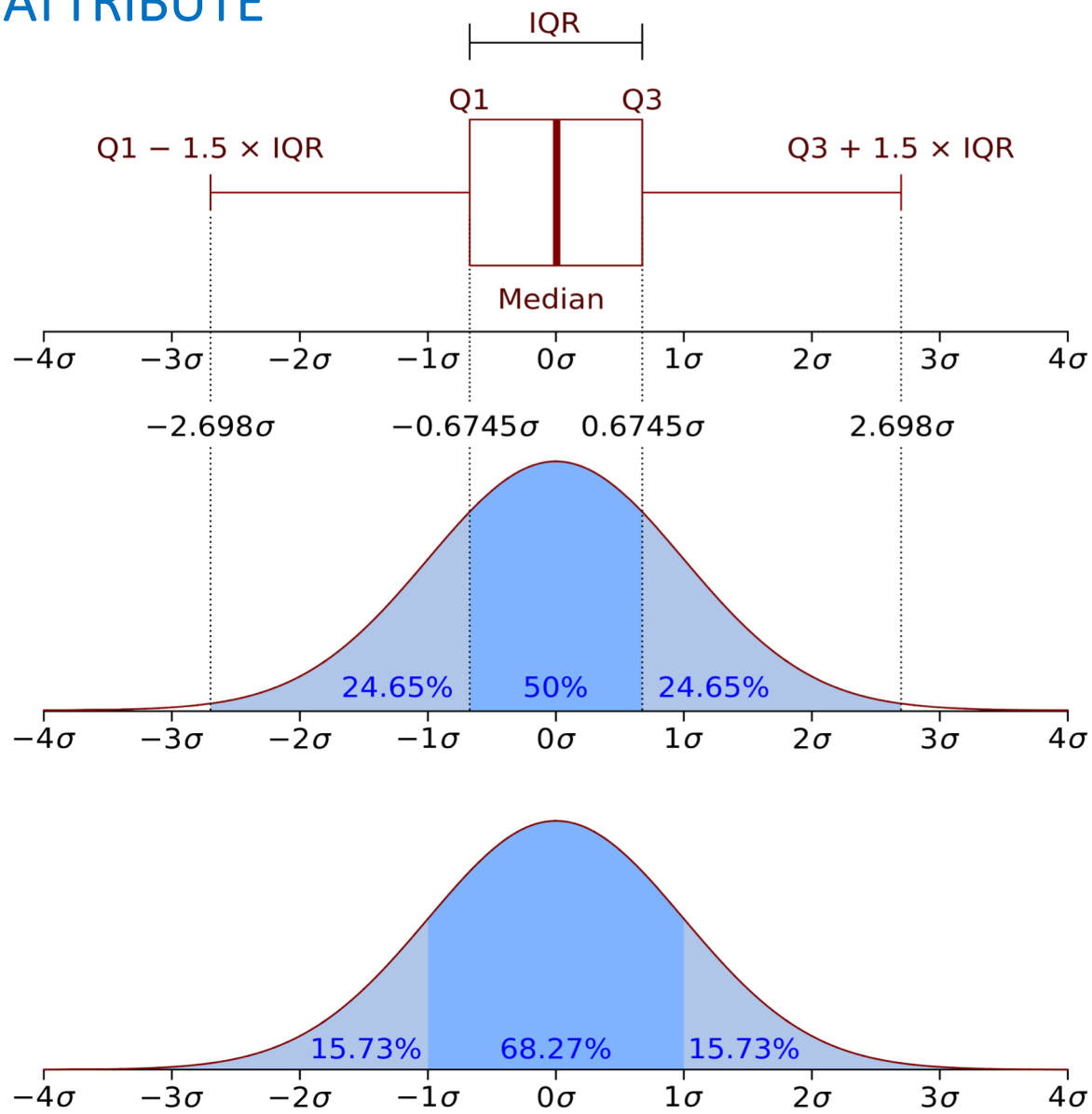
Sample

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Population

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

1. INDIVIDUAL ATTRIBUTE

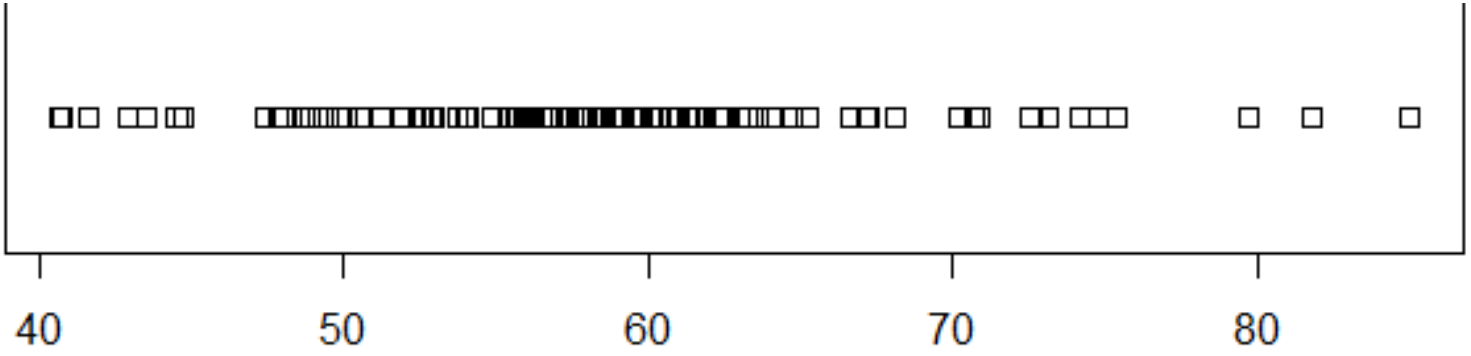


1. INDIVIDUAL ATTRIBUTE

iii. Graphs

- **Point graphs (1D scatterplot)**
  - Each data value is represented by a small point symbol plotted on a number line

1D scatterplot using R



1. INDIVIDUAL ATTRIBUTE

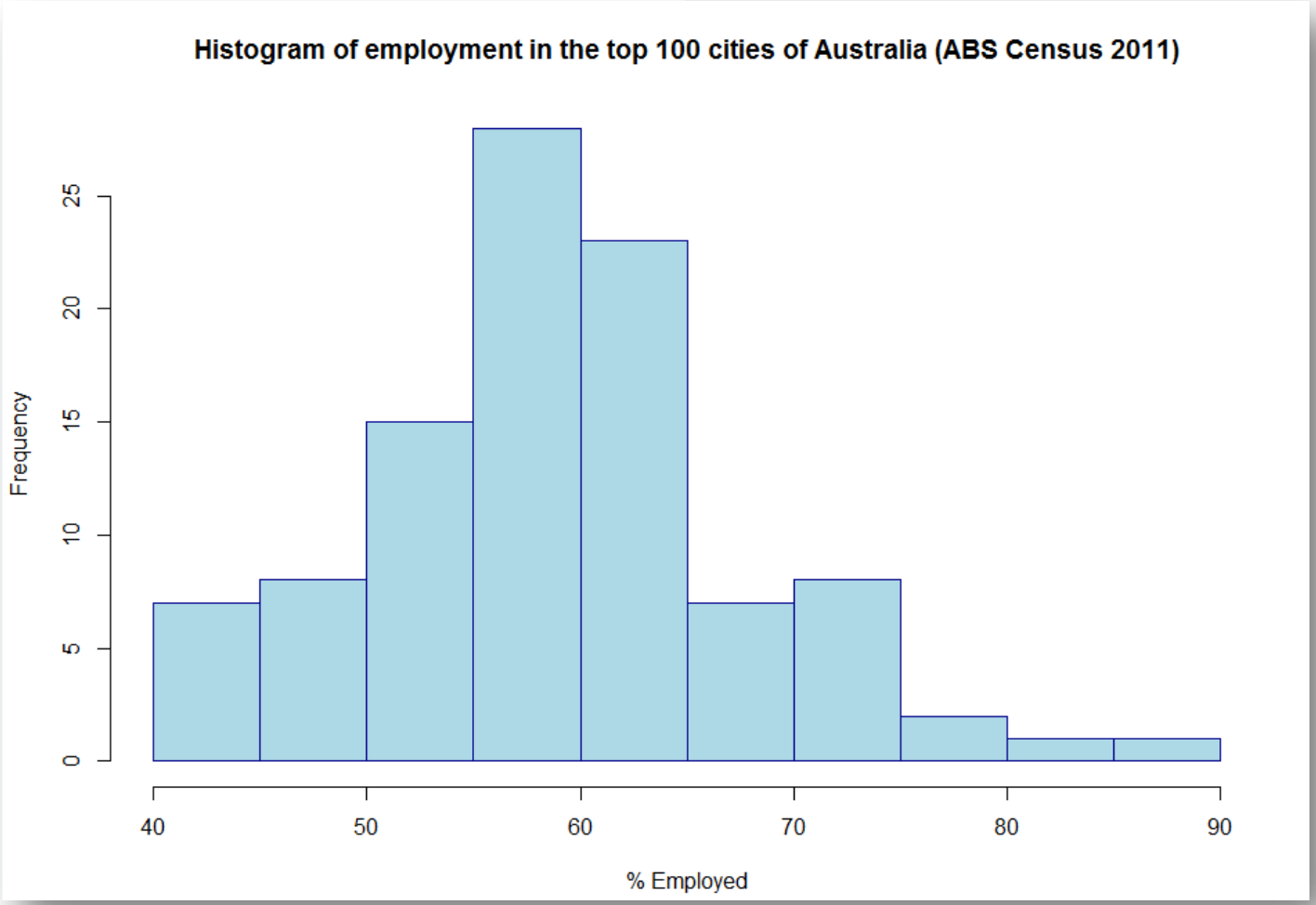
iii. Graphs (cont'd)

- **Histogram**
  - Data grouped into numerical classes
  - Bars of varying height are used to represent the counts (frequency) of observations

1. INDIVIDUAL ATTRIBUTE

Histogram using R

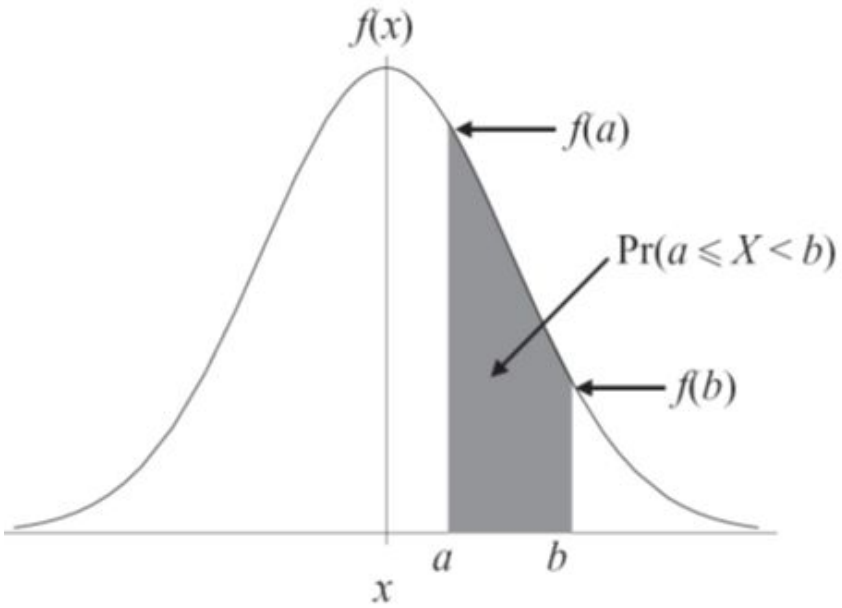
iii.



1. INDIVIDUAL ATTRIBUTE

iii. Graphs (cont'd)

- **Histogram with PDF**
  - Probability Distribution Function (PDF) of a continuous random variable (e.g., amount of rainfall, temperature, etc.)
  - Vertical axis change to 'density'
  - Compare against a hypothetical (normal) distribution

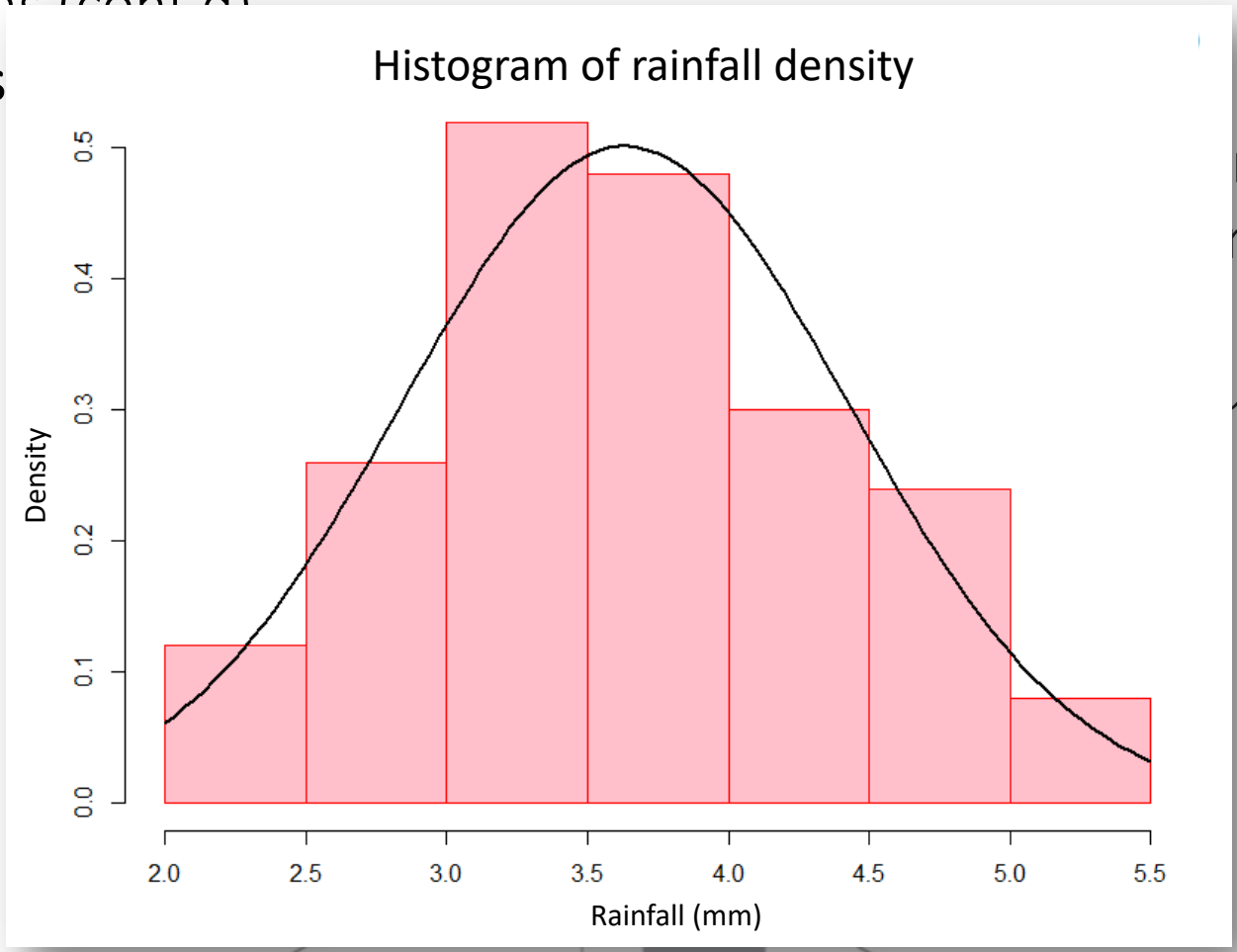


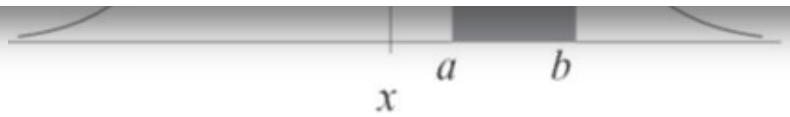
1. INDIVIDUAL ATTRIBUTE

Histogram with probability density using R

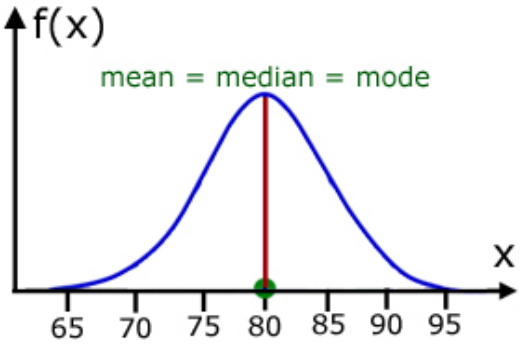
iii. Graphs (cont'd)

- His
  - ○ nuous
  - ○ rature, etc.)
  - ○ ution





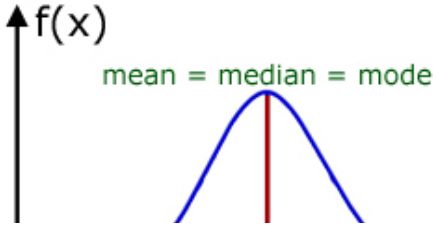
1. INDIVIDUAL ATTRIBUTE



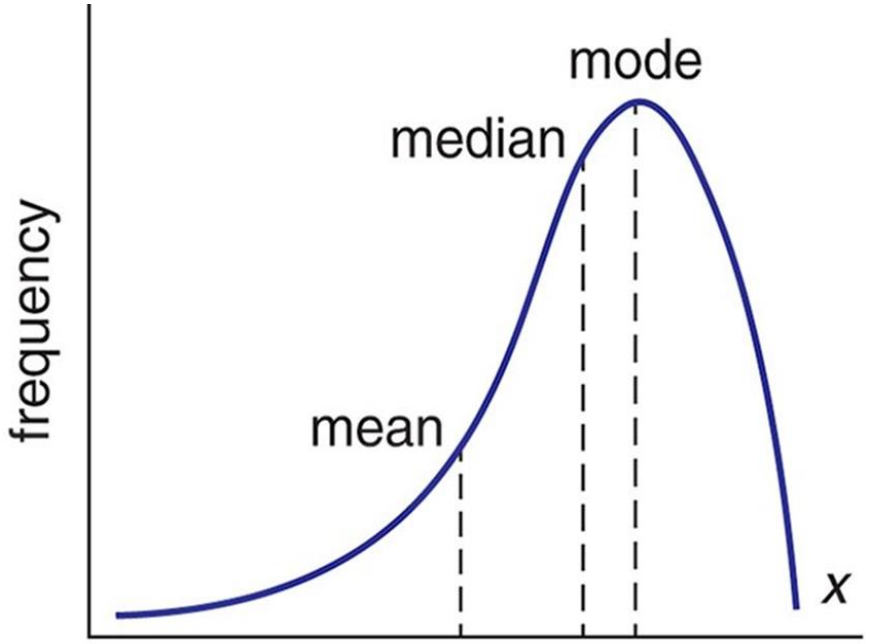
iii. Graphs (cont'd)

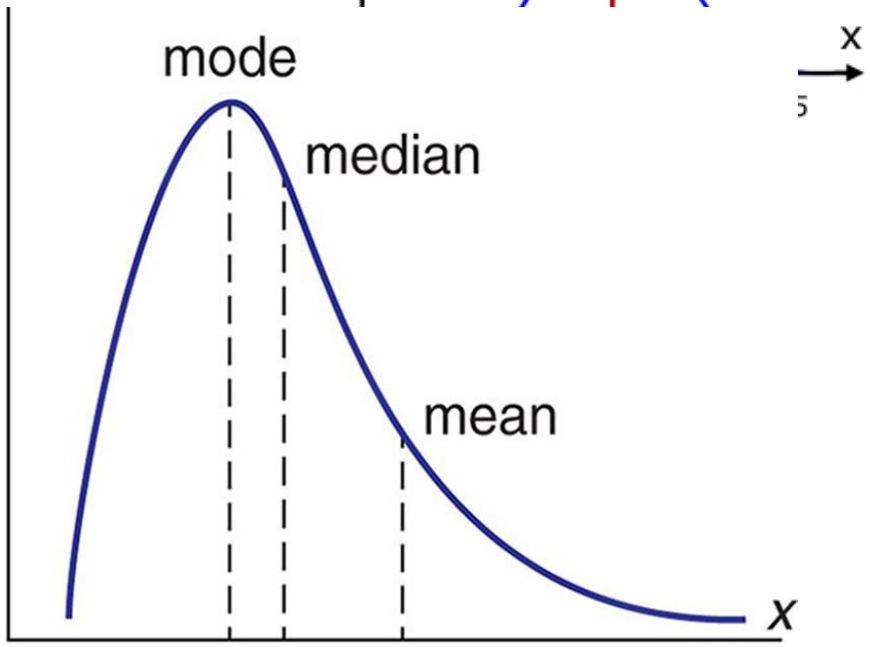
- **Distributions**
  - Normal (symmetric around the mean)
    - Required for testing
  - Skewed distributions (lack symmetry)
    - Positive skew
    - Negative skew
  - Transformations
    - Log
    - Square root

1. INDIVIDUAL ATTRIBUTE



iii. Graphs (cont'd)





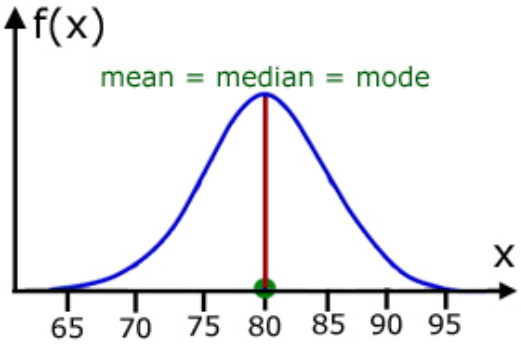
negative direction

positive direction

Negatively skewed

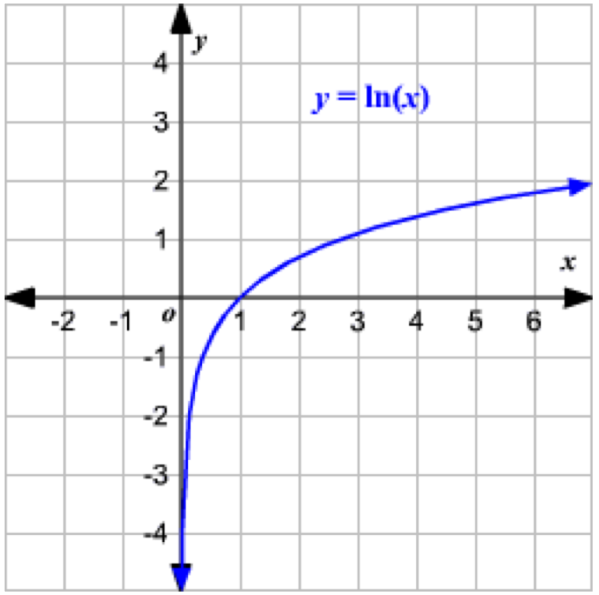
Positively skewed

1. INDIVIDUAL ATTRIBUTE



iii. Graphs (cont'd)

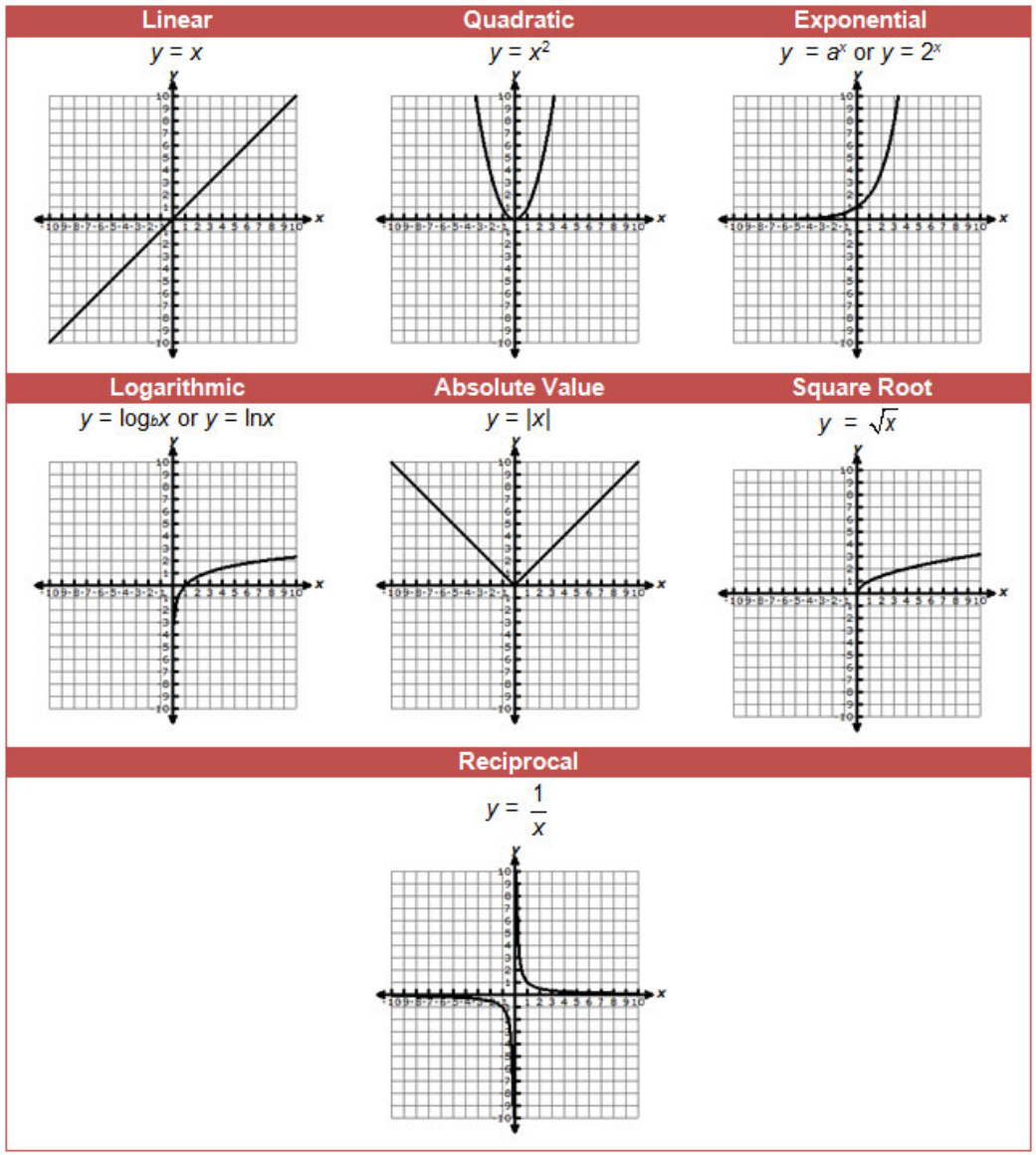
- **Distributions**
  - Normal (symmetric around the mean)
    - Required for testing
  - Skewed distributions (lack symmetry)
    - Positive skew
    - Negative skew
  - Transformations
    - Log
    - Square root

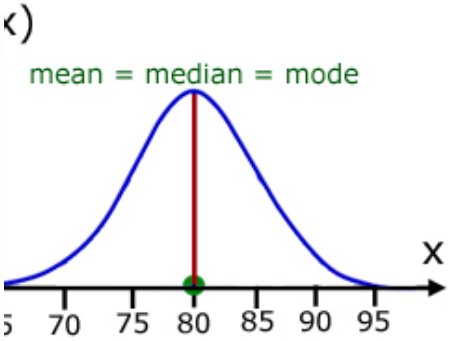


1. INDIVIDUA

iii. Graphs (

- Distrib
  - N
    -
  - Sk
    -
    -
  - Tr
    -
    -





2. RELATIONSHIPS BETWEEN TWO OR MORE ATTRIBUTES

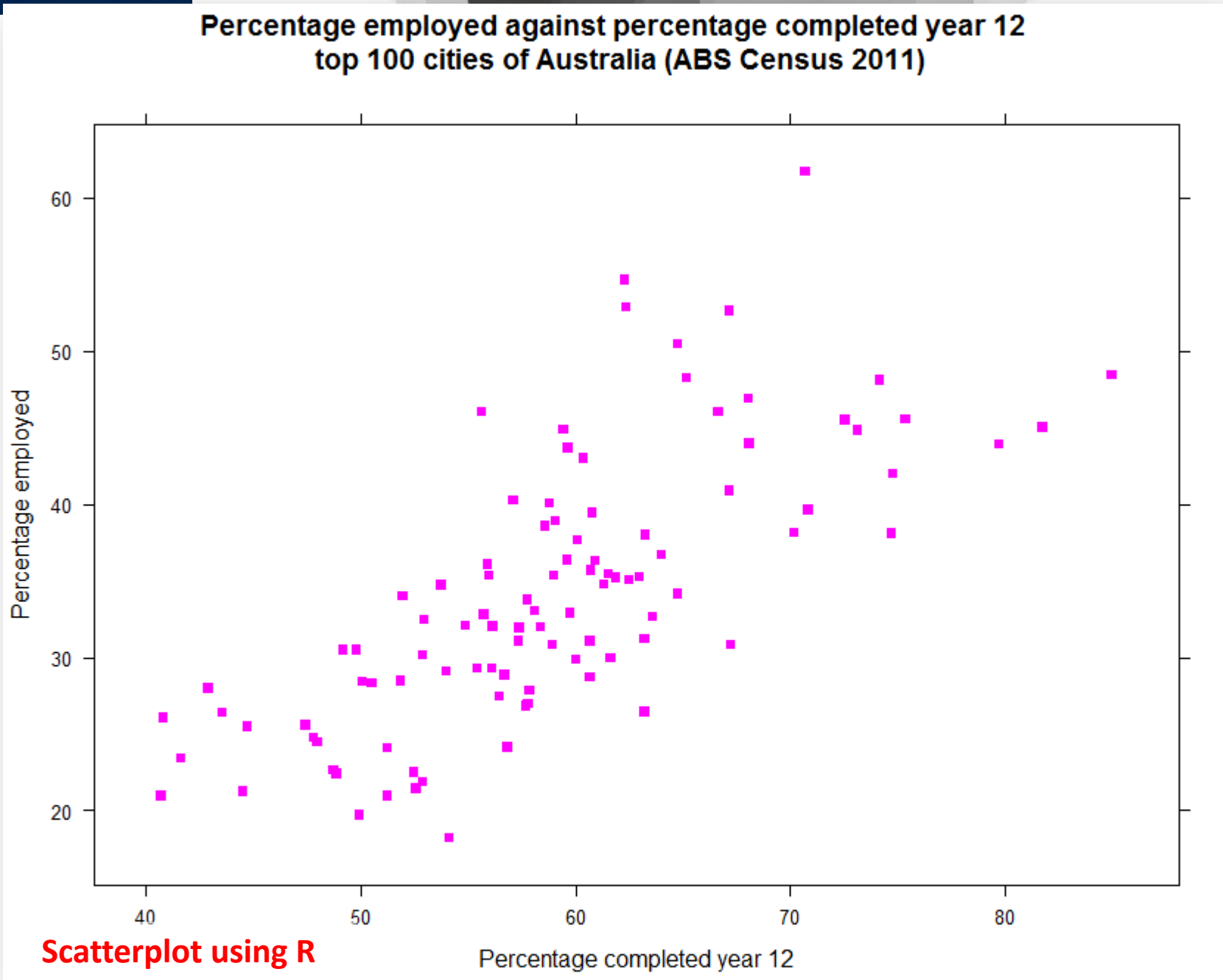
i. **Numerical summaries**
- Tables
  - Display dependent and independent variables in text form
  - Requires careful formatting for logic

ii. **Graphs**
- Scatterplot
  - Use to examine the relationship between two attributes in two-dimensional space
  - X-axis typically independent attribute
  - Y-axis typically dependent attribute
  - Higher dimension scatterplots (e.g., 3D) are difficult to interpret

2.

i.

ii.

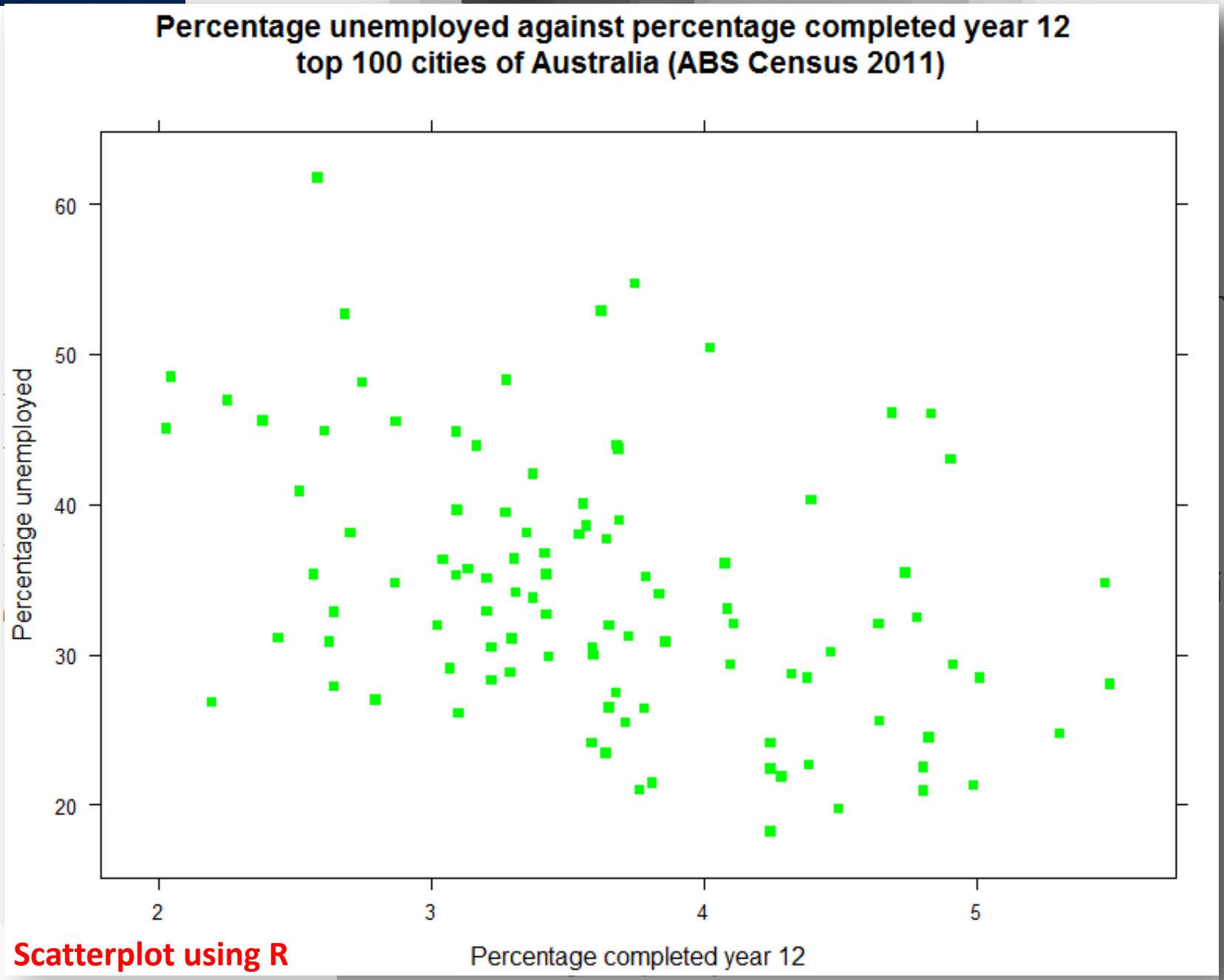


Scatterplot using R

2.

i.

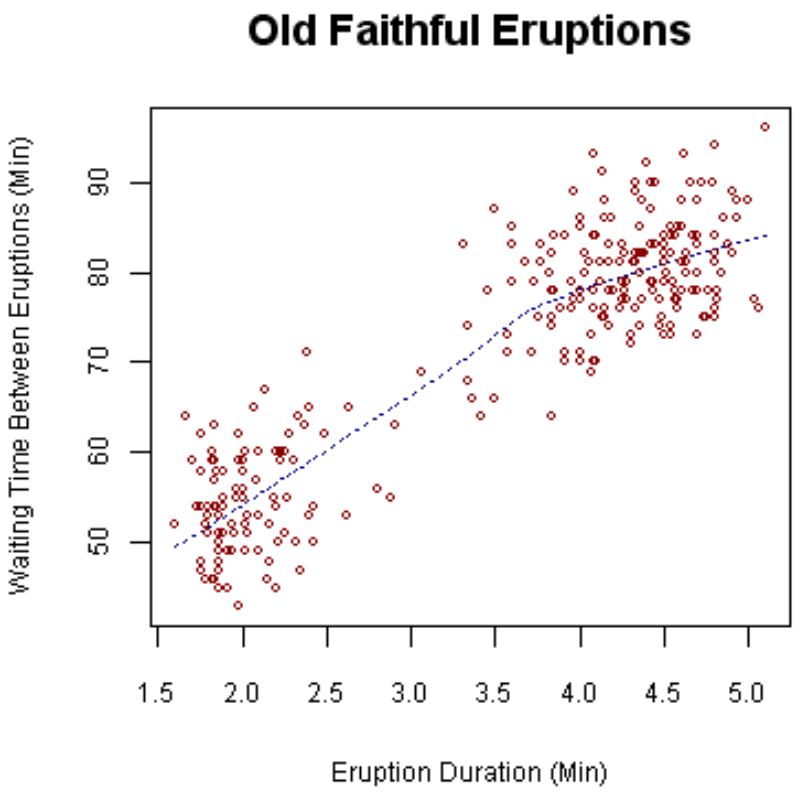
ii.



Scatterplot using R

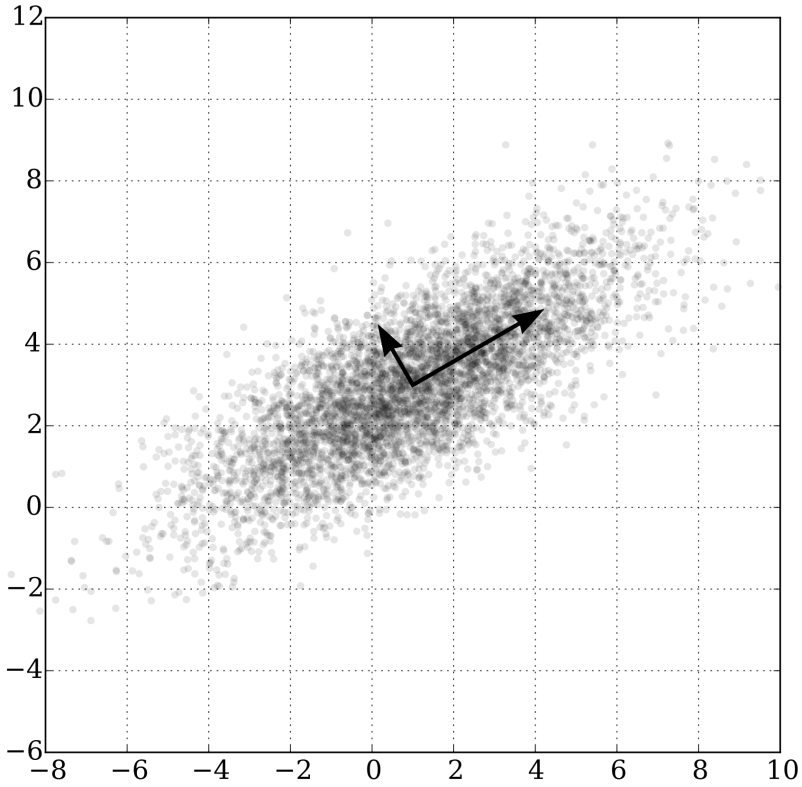
2. RELATIONSHIPS BETWEEN TWO OR MORE ATTRIBUTES

Waiting time between eruptions and the duration of the eruption for the Old Faithfull Geyser in Yellow Stone National Park, Wyoming, USA. This scatterplot suggests there are generally two "types" of eruptions: short-wait-short-duration, and long-wait-long-duration (Wikipedia Commons).



2. RELATIONSHIPS BETWEEN TWO OR MORE ATTRIBUTES

A scatter plot of samples that are distributed according to a multivariate (bivariate) Gaussian distribution centered at (1,3) with a standard deviation of 3 in the (0.866, 0.5) direction and of 1 in the orthogonal direction. The directions represent the Principal Components (PC) associated with the sample (Nicoguaro from Wikipedia Commons).



2. RELATIONSHIPS BETWEEN TWO OR MORE ATTRIBUTES

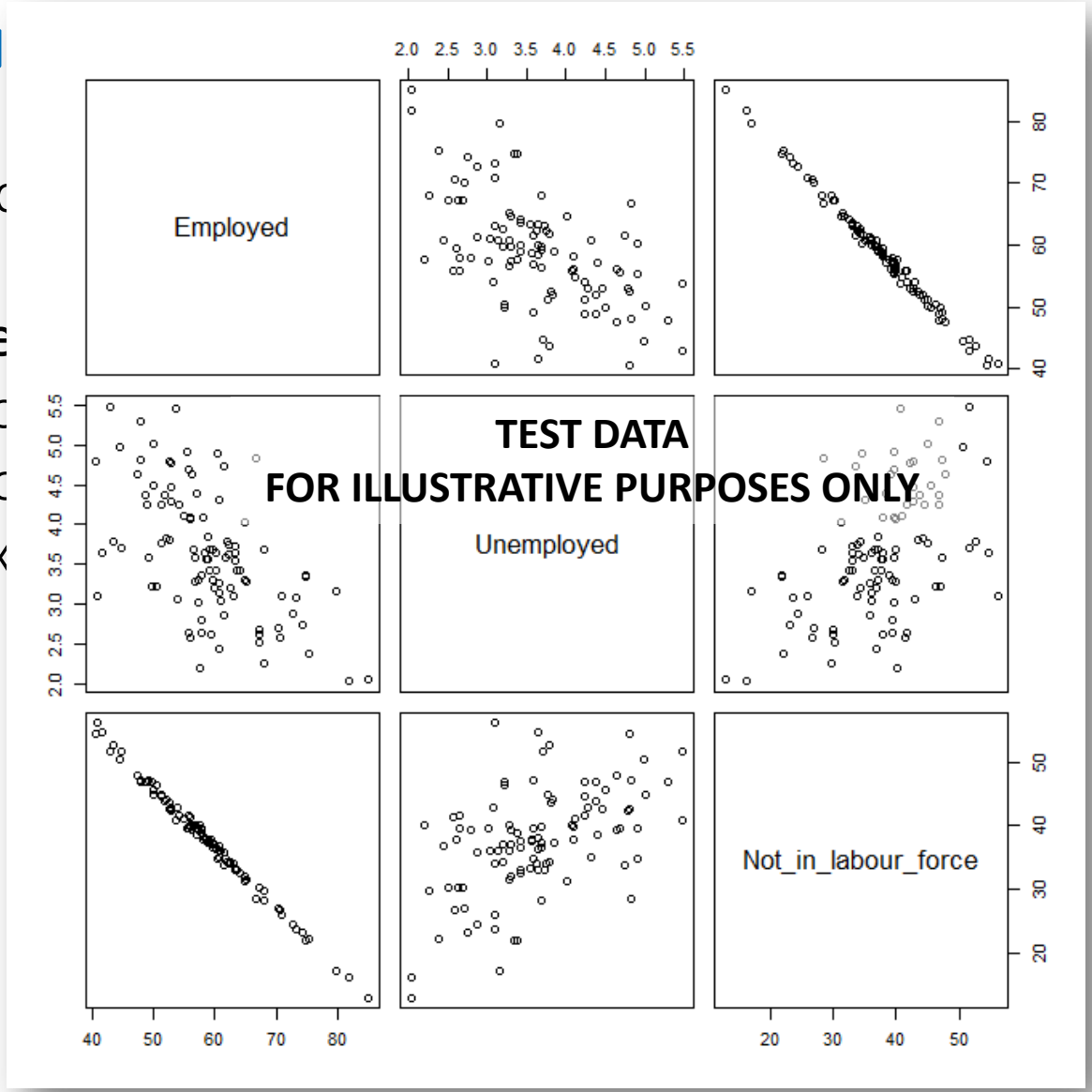
iii. Graphs (cont'd)

- **Scatterplot matrix**
  - Row considers attribute as dependent
  - Column considers attribute as independent
  - Explore and look for strong relationships

2. RELATION

iii. Graphs (

- Scatte
  - Rc
  - Cc
  - Ex



Scatterplot matrix using R

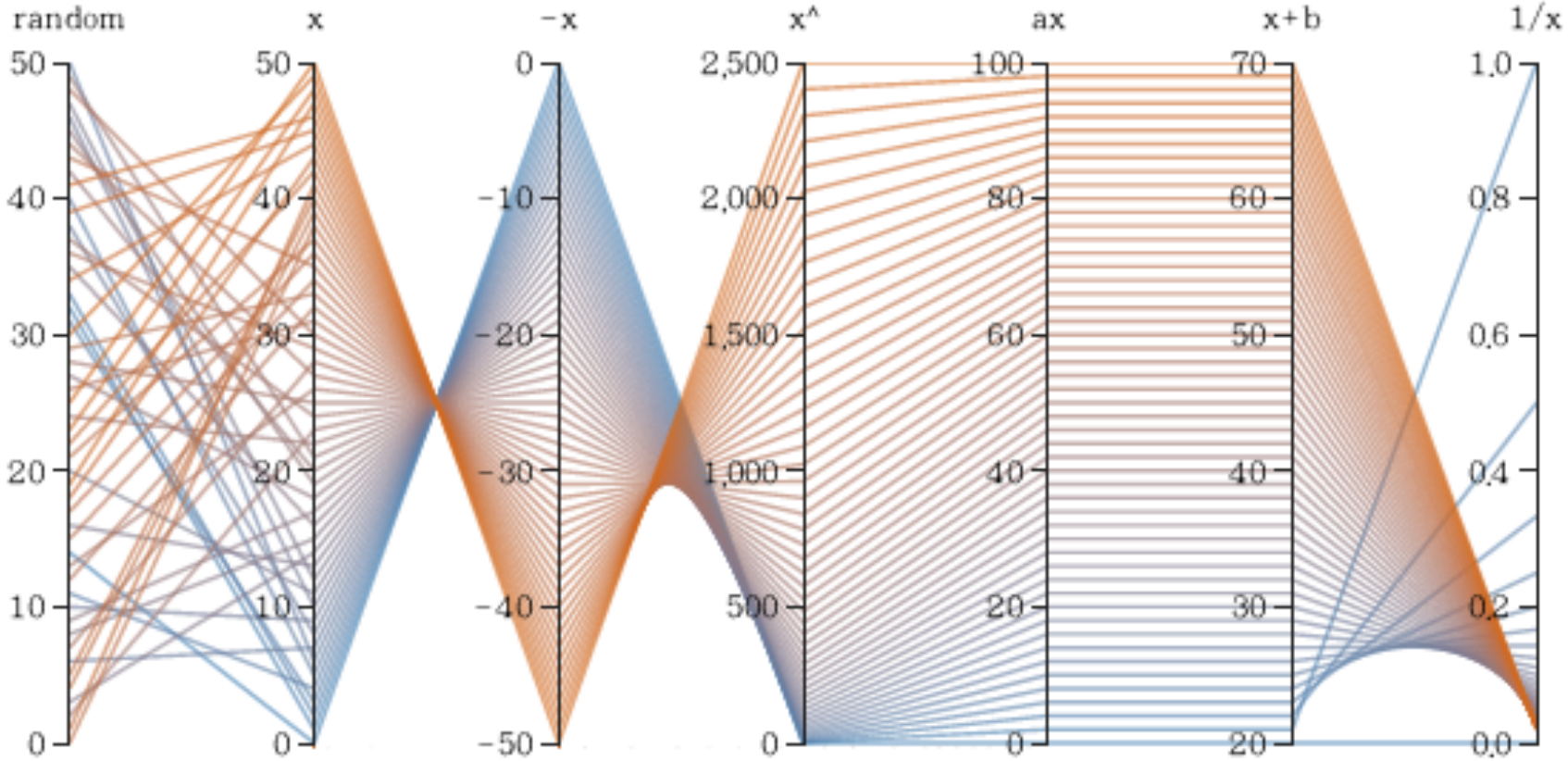
2. RELATIONSHIPS BETWEEN TWO OR MORE ATTRIBUTES

iii. Graphs (cont'd)

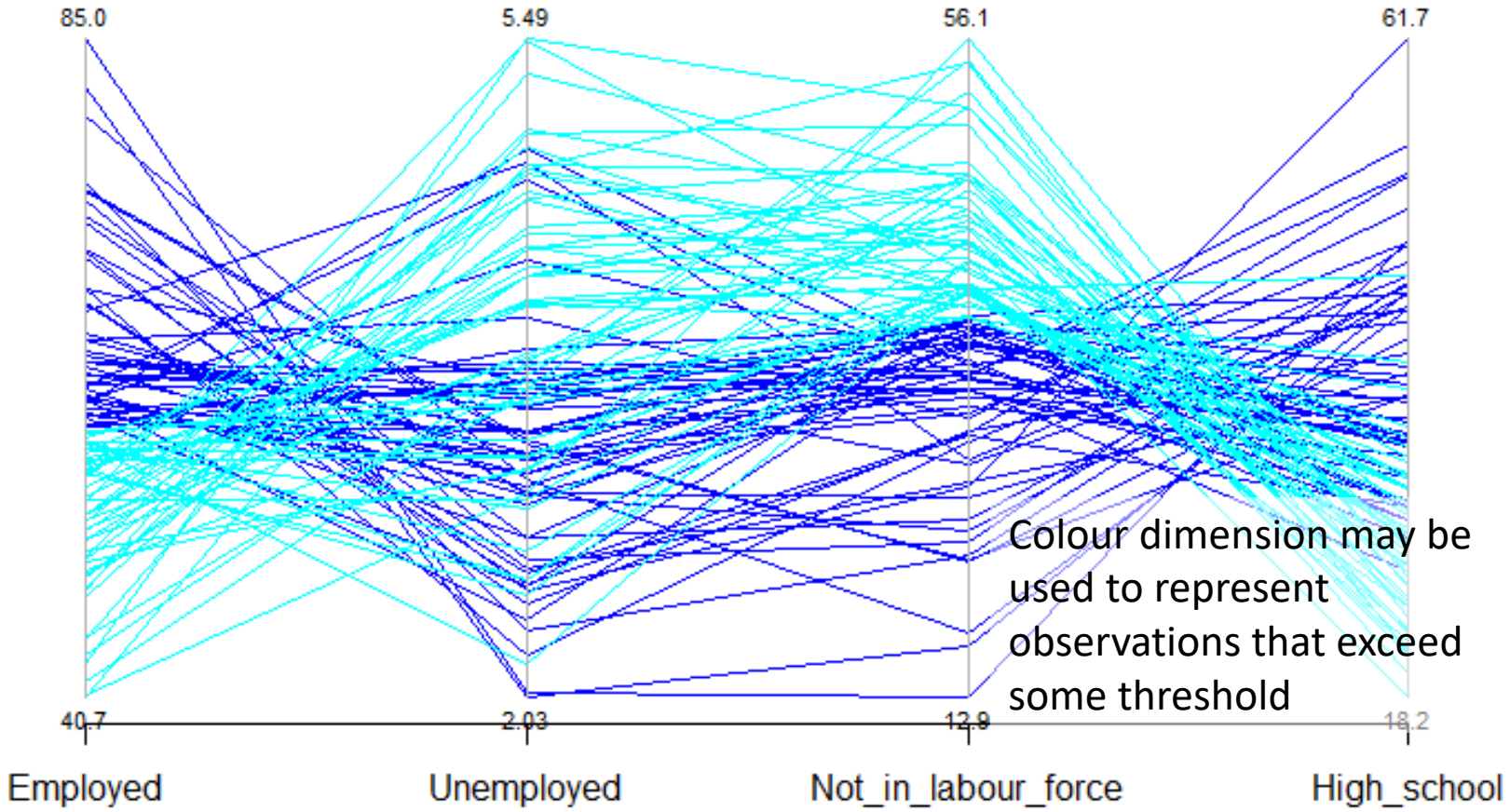
- **Parallel coordinate plot**
  - 3 or more attributes, each represented with a vertical line
  - Observations run horizontal using interconnected lines
  - Correlation coefficient, *r*
    - Different patterns for
      - Parallel lines *r*=1
      - Intersecting lines in the middle *r*=-1
      - Random lines *r*=0

2. RELATIONSHIPS BETWEEN TWO OR MORE ATTRIBUTES

iii. Graphs (cont'd)



Parallel coordinate plot using R



Colour dimension may be used to represent observations that exceed some threshold

3. EXPLORATORY DATA ANALYSIS (EDA)

- Many of the techniques presented can be used for EDA e.g., histogram
  - Tukey (1977) *Exploratory data analysis*. Pearson
- Other techniques include:
  - Stem and leaf plots
  - Box plots
- Information visualisation
  - Search for a hypothesis, suggest future work
- Exploratory Spatial Data Analysis (ESDA)
  - To be discussed in later lectures

C. Methods for analysing numerical data – CONSIDERING LOCATION

A brief introduction to spatial statistics

- Requires fundamental knowledge of
  - Spatial objects and data types
  - Random and other processes

1. Analysis of spatial location
   - Centroid
   - Shape

2. Analysing an attribute in association with a spatial location
   - Central tendency and dispersion of point data
   - Aggregation problem

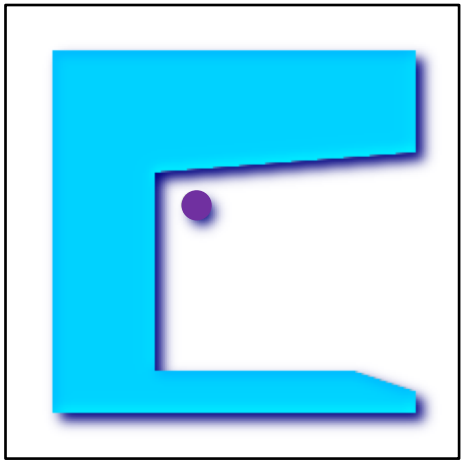
1. ANALYSIS OF SPATIAL LOCATION

i. Centroid

- The geometric centre (centre of mass or gravity) of a geometry

- Note: Centroid may be inside or outside of a feature!

Polygon



1. ANALYSIS OF SPATIAL LOCATION

ii. Shape

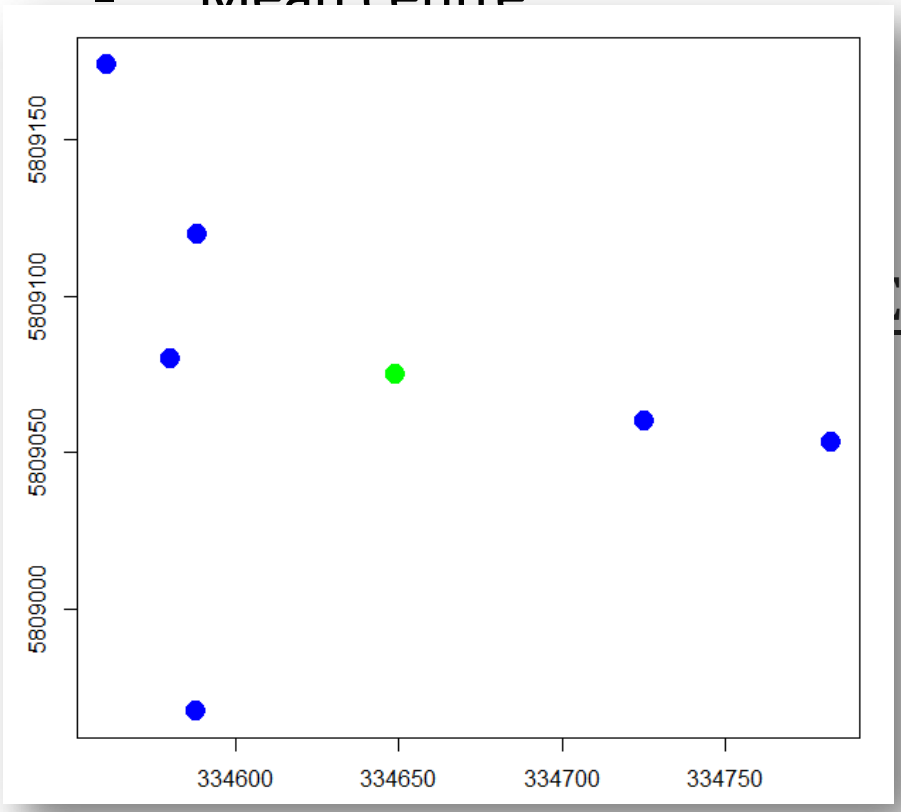
- Shape does matter (Angel et al., 2010)
  - Conceptually challenging

- Various indices
  - Compaction Index (CI)
    - Ratio of feature area to the circle area circumscribing it
  - Other of bounding geometry (e.g. rectangle)
    - Ratio

2. ANALYSIS OF AN ATTRIBUTE IN ASSOCIATION WITH SPATIAL LOCATION

i. Central tendency and dispersion of point data

- Mean centre



twice (x,y):

$$\frac{\sum_{i=1}^{n} Y_i}{n}$$

2. ANALYSIS OF AN ATTRIBUTE IN ASSOCIATION WITH SPATIAL LOCATION

i. Central tendency and dispersion of point data

- **Mean centre**
  - Apply non-spatial mean twice (x,y):

$$\overline{X_c} = \frac{\sum_{i=1}^{n} X_i}{n}, \overline{Y_c} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

  - Weighted mean centre (e.g., city population)

$$\overline{X_c} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}, \overline{Y_c} = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i}$$

2. ANALYSIS OF AN ATTRIBUTE IN ASSOCIATION WITH SPATIAL LOCATION

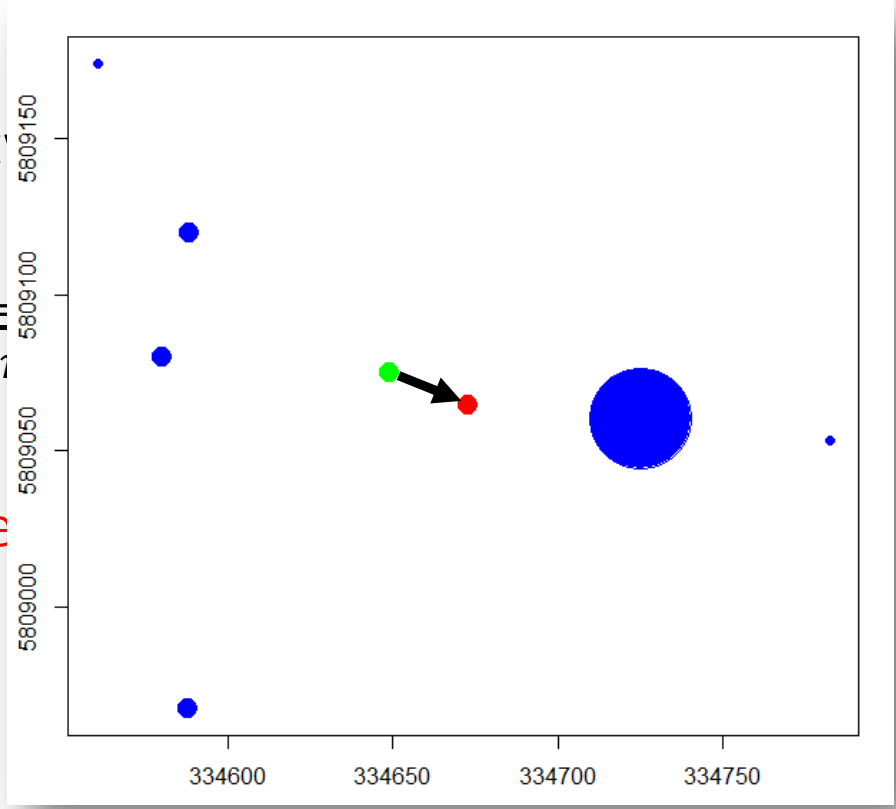
i. Central tendency and dispersion of point data

- **Mean centre**
  - Apply non-spatial mean t

$$\overline{X_c} = \frac{\sum_{i=1}^{n} X_i}{n}, \overline{Y_c} = \frac{\sum_{i=}^{n}}{}$$

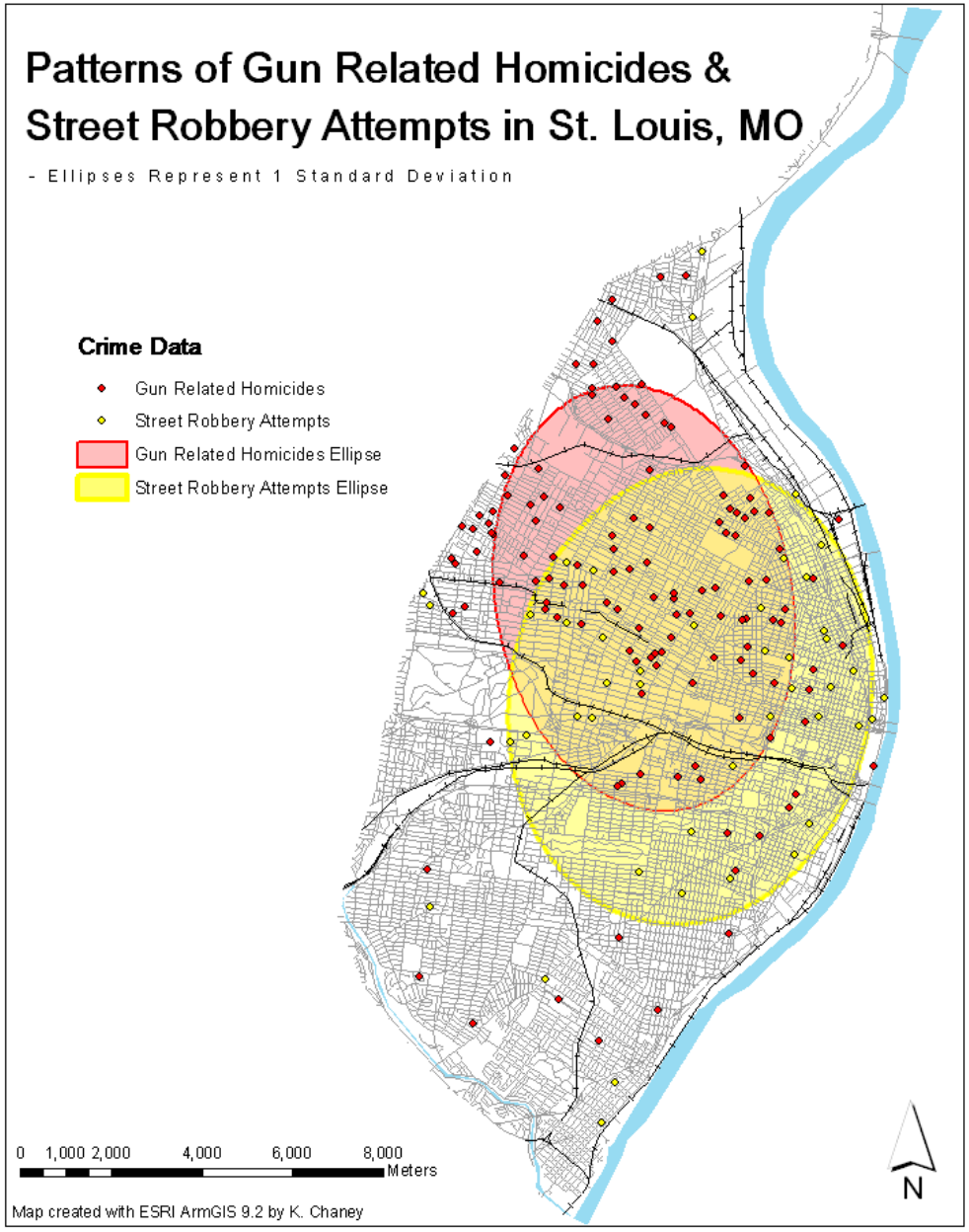
  - Weighted mean centre (e

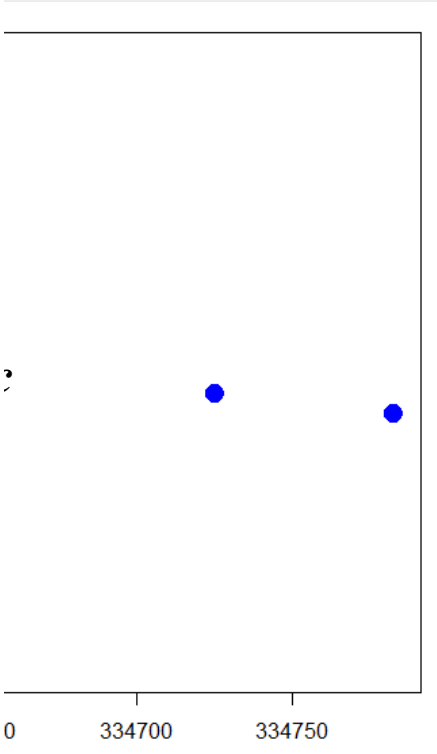
$$\overline{X_c} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}, \quad \overline{Y_c} =$$



2. ANALYSIS OF A … PATIAL LOCATION

i. Central tenc





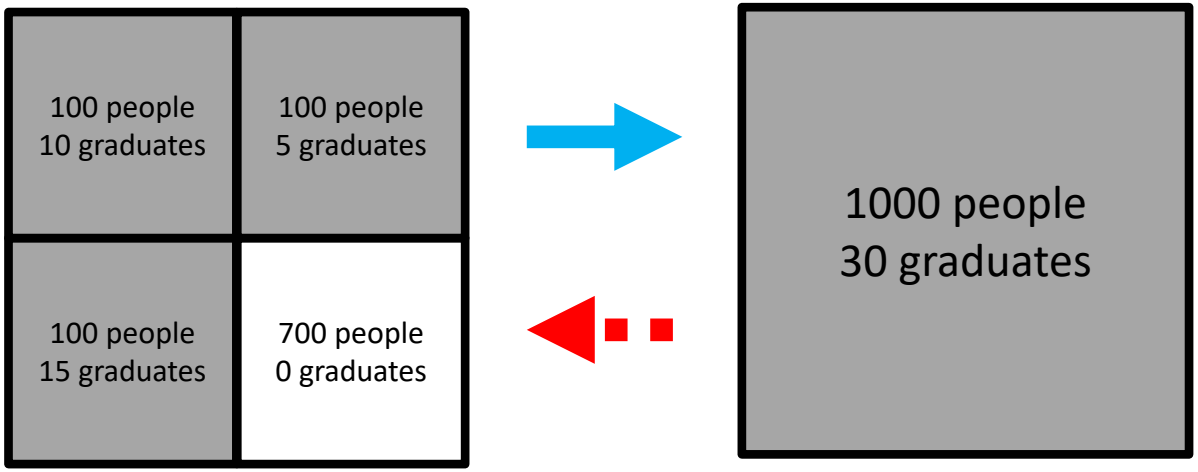
2. ANALYSIS OF AN ATTRIBUTE IN ASSOCIATION WITH SPATIAL LOCATION

i. Central tendency and dispersion of point data

- **Spatial (geometric) median**
  - Overcomes issue when mean centre does not coincide with where the data is clustered
  - Various techniques http://www.r-bloggers.com/multivariate-medians/
- **Direction**
  - Requires vector data

PROBLEMS WITH LOCATION DATA

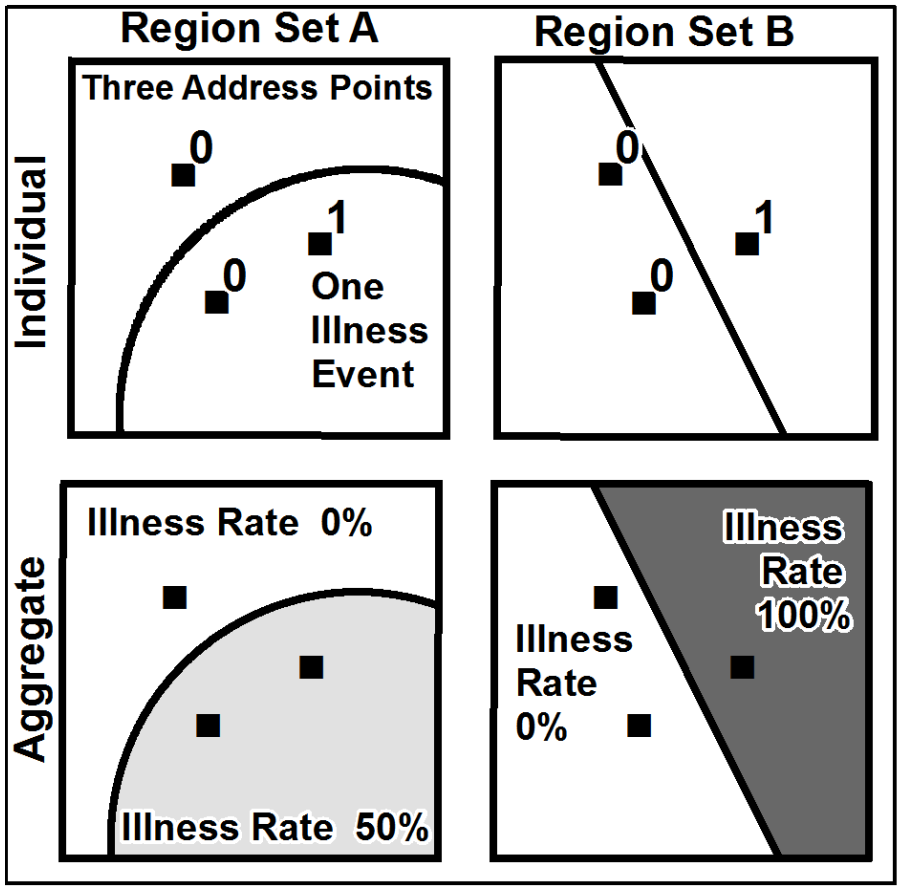
Average of the four regions:
7.5%



Average of region:
3%

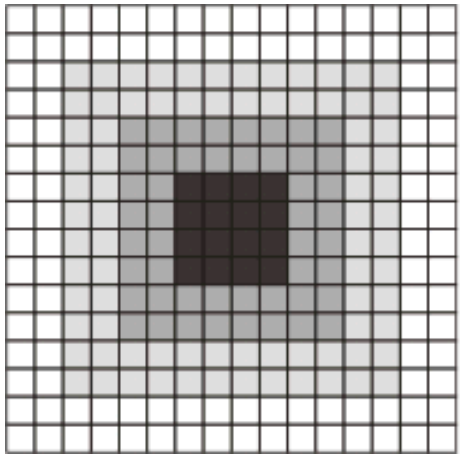
- Scaling up: data may be lost
- Scaling down: ecological fallacy (Anselin et al., 2000)
- Related to the Modifiable Areal Unit Problem (MAUP)
  e.g. http://giscollective.org/geographic-data-assumptions-maup-and-ecological-fallacies/

PROBLEMS WITH LOCATION DATA

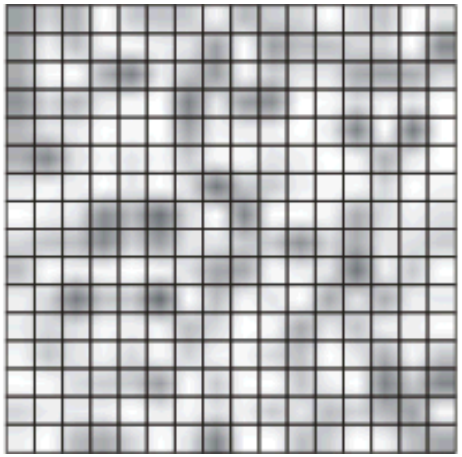


- Related to the Modifiable Areal Unit Problem (MAUP)
  e.g. http://giscollective.org/geographic-data-assumptions-maup-and-ecological-fallacies/

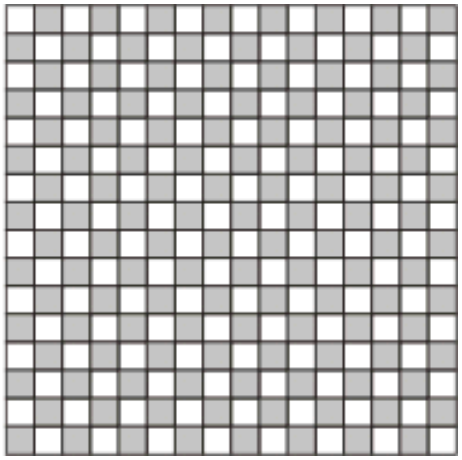
SPATIAL AUTOCORRELATION



If like values tend to cluster together, then the field exhibits high **positive spatial autocorrelation**



If there is no apparent relationship between attribute value and location then there is **zero spatial autocorrelation**



If like values tend to be located away from each other, then there is **negative spatial autocorrelation**

NEXT LECTURE

- Data graphics 1

READING

Refer LMS, Lecture materials: Week 3

- Section 3.4: Numerical summaries in which location is an integral component, Slocum et al., (2009) *Thematic cartography and geovisualisation*. Prentice Hall, Upper Saddle River, NJ.