Visualising Correlation

NANYANG THE INNOVATIVE POLYTECHNIC School of Information Technology

Learning Outcomes

By the end of this lesson, you should be able to

- Identify the patterns of correlation and distribution
- Use and explain the various data comparing approaches
- Explain and apply the techniques and best practices used

Introduction to Correlation

Correlation analysis involves comparing two quantitative variables to see if values in one vary systematically with the other, and if so,

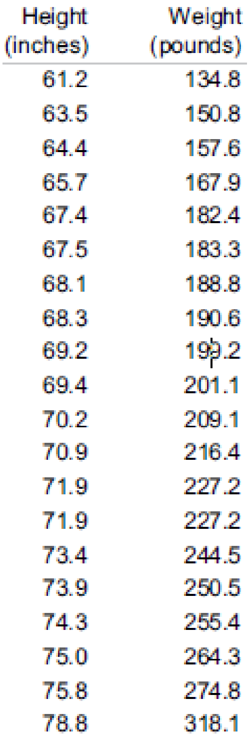
- in what manner,
- to what degree and
- why

In short, the relationships between/among different variables

Introduction to Correlation

- Sample list of 20 men, sorted in order of height from shortest tc tallest.

- Correlation can indicate the following:
  - One variable causes another's behavior
  - Neither causes the other's behavior
  - Apparent correlation is erroneous because of an insufficient or biased sample



Height (inches)	Weight (pounds)
61.2	134.8
63.5	150.8
64.4	157.6
65.7	167.9
67.4	182.4
67.5	183.3
68.1	188.8
68.3	190.6
69.2	199.2
69.4	201.1
70.2	209.1
70.9	216.4
71.9	227.2
71.9	227.2
73.4	244.5
73.9	250.5
74.3	255.4
75.0	264.3
75.8	274.8
78.8	318.1

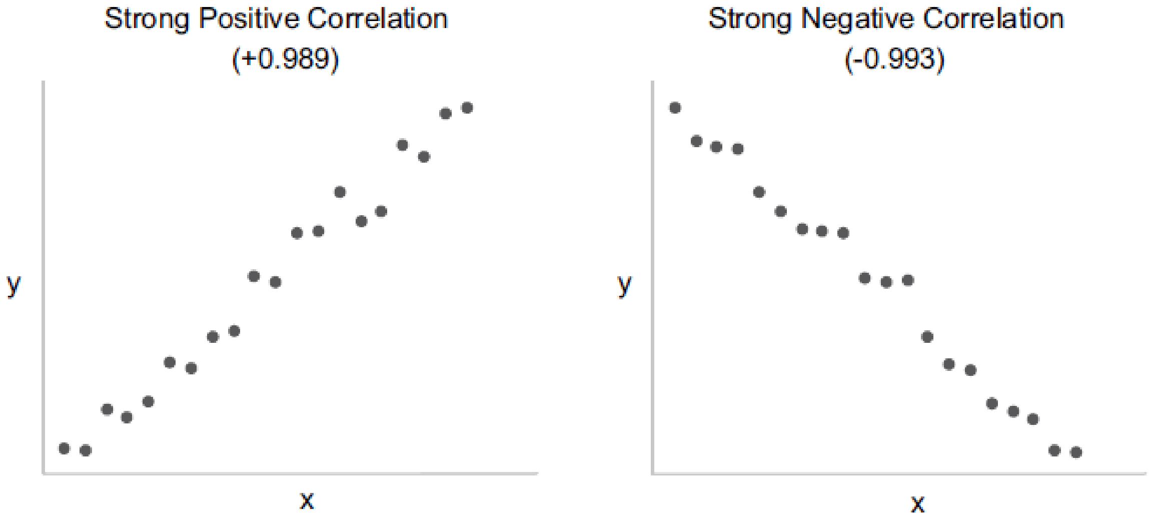
Describing Correlation

- Visual Characteristic of Correlations
  - Direction
    - Positive, Negative
  - Strength
    - Strong, Weak
  - Shape
    - Straight, Curved

- Statistical Summaries of Correlations
  - Linear correlation coefficient (r)
  - Coefficient of determination (r^2)

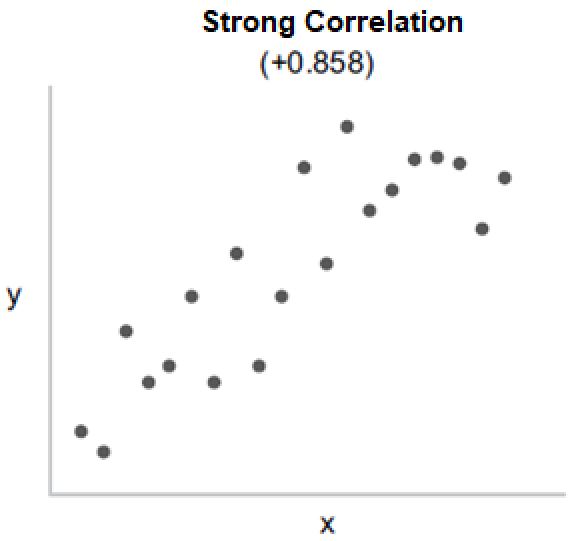
Visual Characteristic

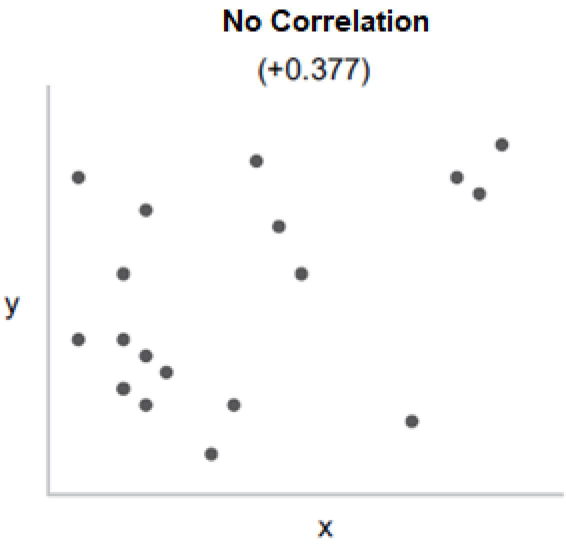
- Direction – positive, negative



Visual Characteristic

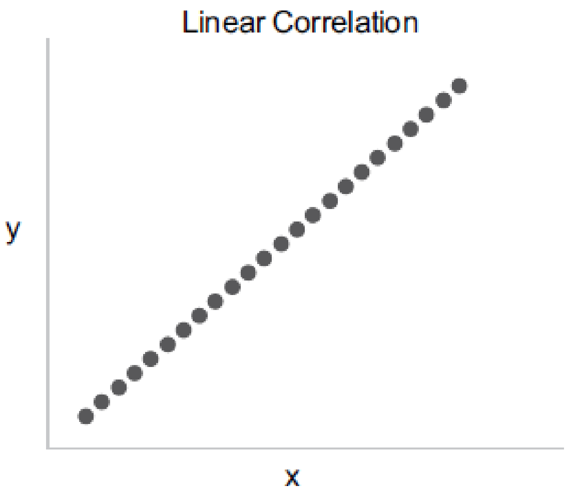
- Strength – strong, weak

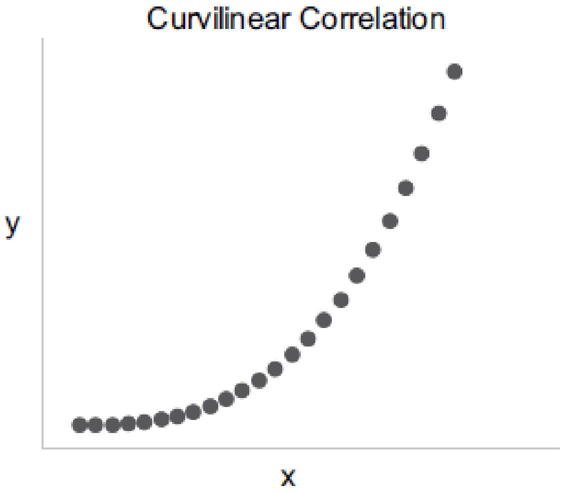




Visual Characteristic

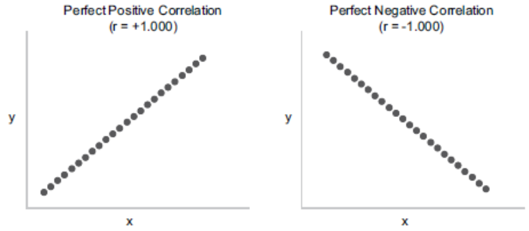
- Shape – straight (linear), curved (curvilinear)





Statistical Summaries of Correlations

- Linear correlation coefficient (r)
  - Describe both the direction and strength of a correlation (but only for linear)
  - Ranges from +1 to -1
- Coefficient of determination (r^2)
  - Describe strength of correlation but not direction
  - Can be expressed as a percentage. r^2 Of 0.986 indicates that 98.6% of the values (weight) can be determined by independent variable (height)
- r and r^2 can be used to test the strength but visual representation tell a richer story.



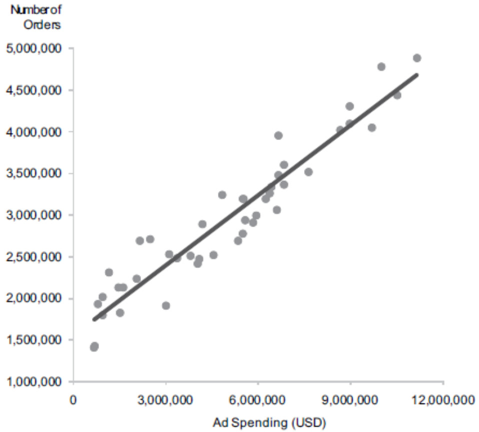
Correlation Patterns

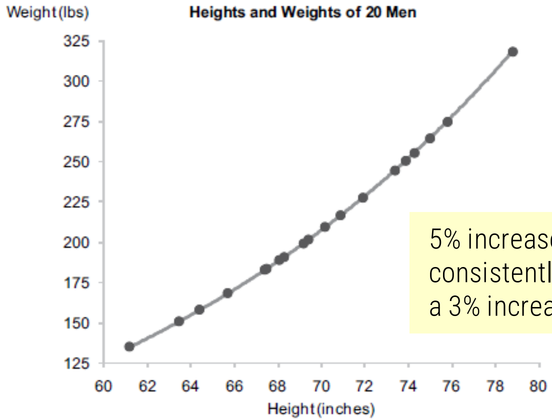
- Shape
  - Is it straight or curved?
  - If curved in one direction only, is it logarithmic, exponential, or some other shape?
  - If curved in both positive and negative directions, does it curve upward or downward?
  - Are there concentrations of values?
- Outliers

Correlation Patterns

- Straight or curved?

Amount of money spent on ads and number of orders a company receives in any given week, if linear correlation → linear trend line



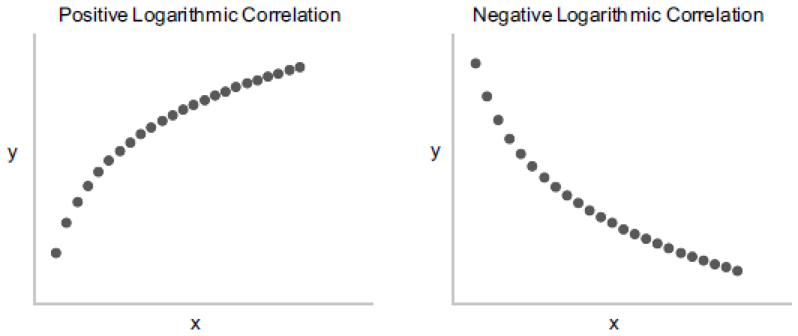


5% increase in height consistently corresponds to a 3% increase in weight

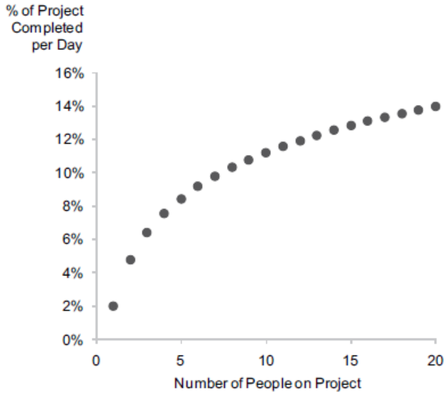
When a correlation is curvilinear, the relationship between values is not fixed to a consistent amount.

Correlation Patterns

Logarithmic or Exponential?

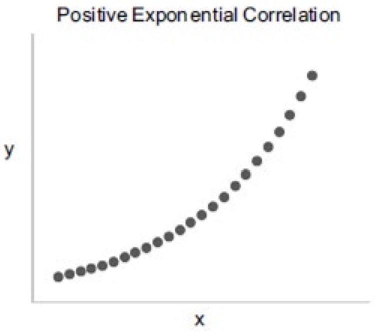


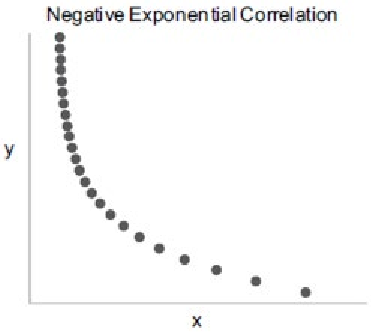
This logarithmic growth pattern is often seen when correlating the number of people who are assigned to a project with the amount of work they produce. The degree of change starts out great but then steadily decreases and eventually levels off.



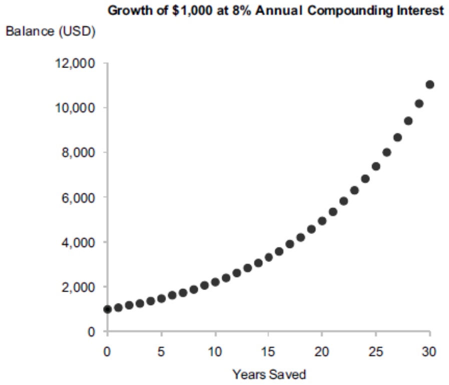
Correlation Patterns

Logarithmic or Exponential?



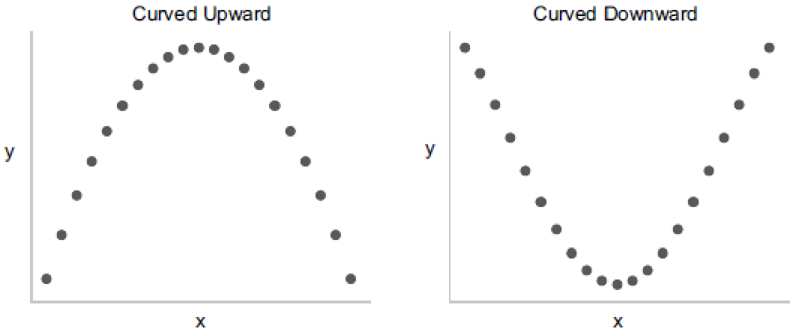


Compound interest that banks pay grows exponentially through time. Values go up by steadily increasing degree. Here’s what the pattern looks like when $1,000 is deposited into an account that pays 8% interest, compounded daily:

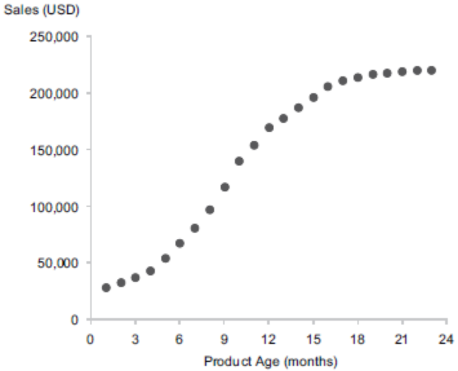


Correlation Patterns

- Curved upward or downward?



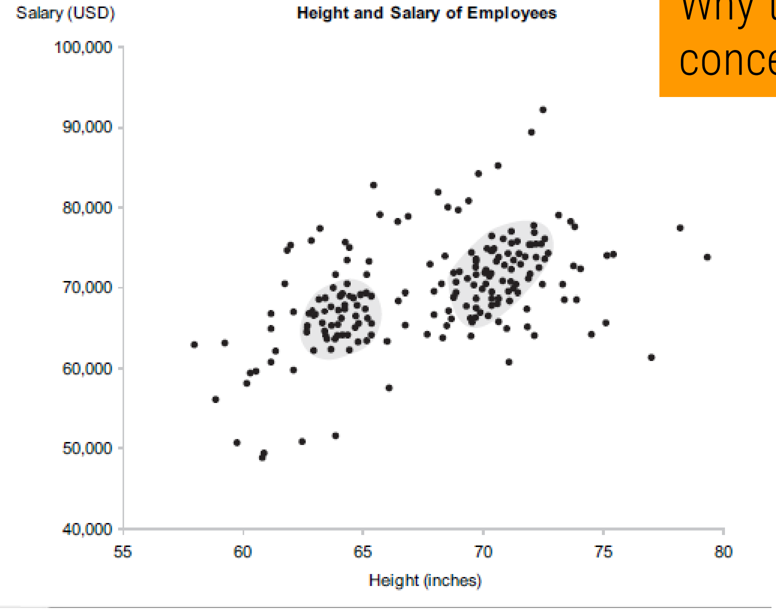
Curved upward – product that increases rapidly in sales during early life but gradually slows down.
S-curves – increasing sales until the product’s popularity reaches it peak and then levels off.



Correlation Patterns

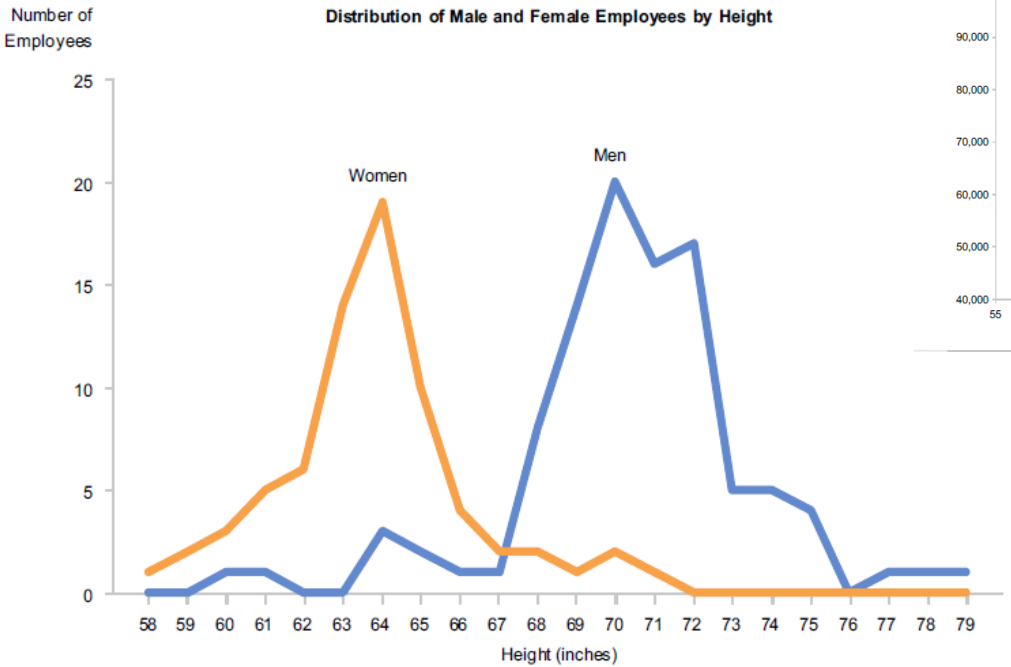
Concentrations?

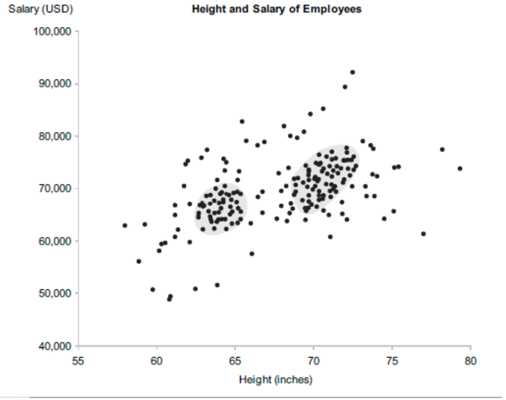
Correlations often exhibit groups of values that are close to one another, that is, concentrations of values, which are easy to spot in a scatterplot because the data points are clustered together.



Why two concentrations?

Concentration?



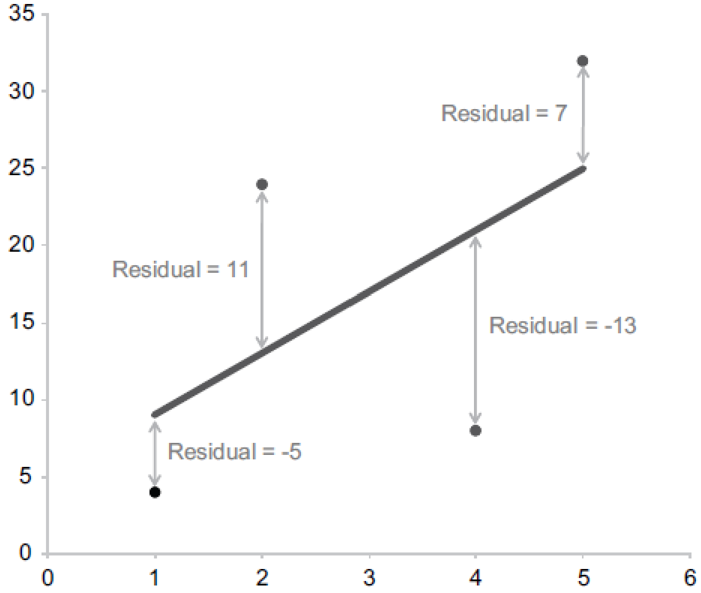


Correlation Patterns

Outliers

A few values that don’t fit the basic shape formed by majority are outliers and they are measured via residual.

It is important to understand under what circumstances values stray from the flock.

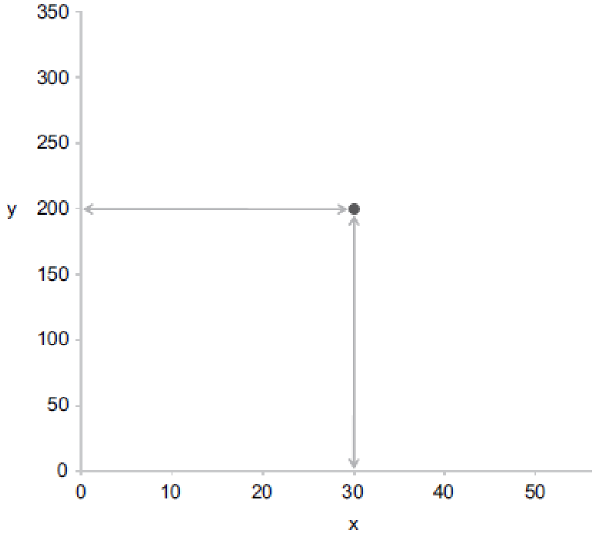


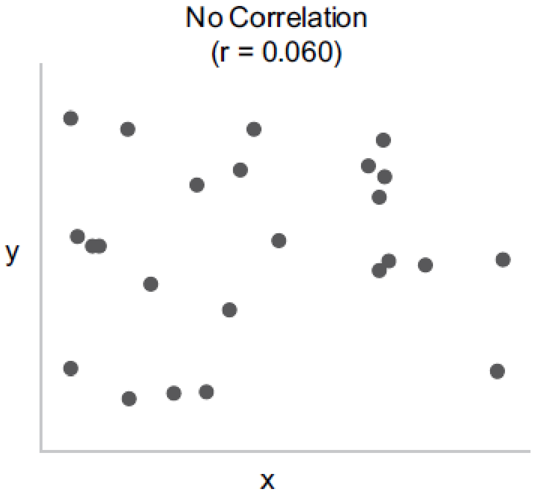
Correlation Displays

- Scatterplot
  - But it is designed to compare two quantitative variables only.
- Scatterplot matrix
  - Displays several scatterplots
- Table Lens
  - To detect possible correlations among many variables all at once in a single display

Scatterplots

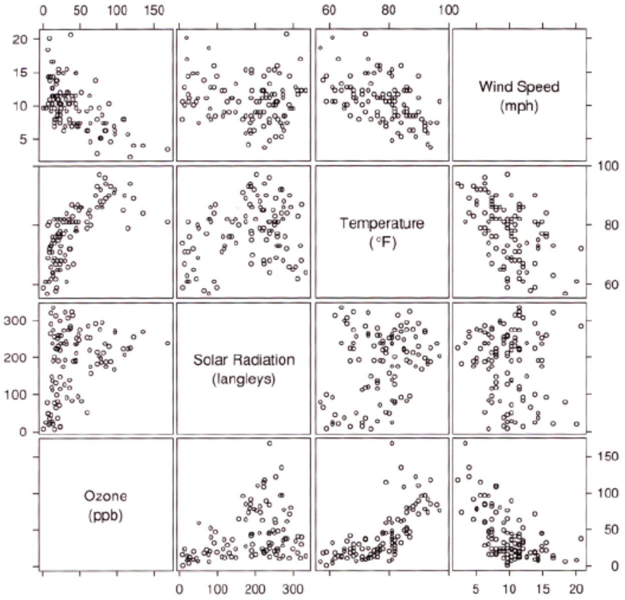
- Display two quantitative variables and how these variables are related.





Scatterplot Matrices

- Comparing multiple pairs of variables

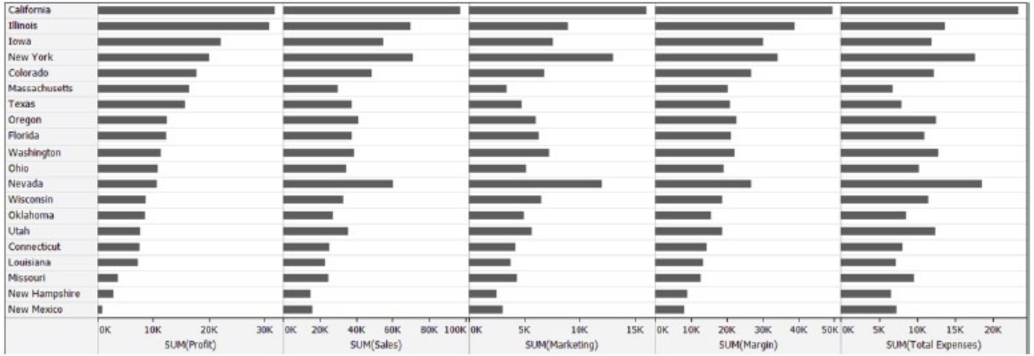


Ozone (an air pollutant when at ground level), solar radiation, air temperature, and wind speed. How do these four variables interact?

The Elements of Graphing Data,
William S. Cleveland,
Hobart Press, 1994, p. 257.

Table Lenses

- Comparing more than two variables simultaneously



Does not display correlations as richly as precisely as scatterplots but a good way to find out correlations among many variables.

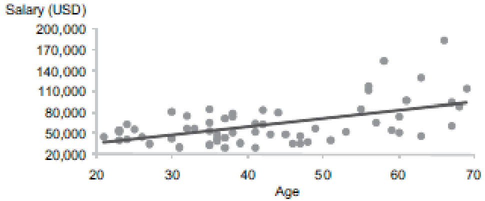
The table lens displays, for 20 states, five variables related to sales: profits, revenues, marketing expenses, profit margin, and total sales-related expenses. The quantitative scales for each of the variables are independent from one another.

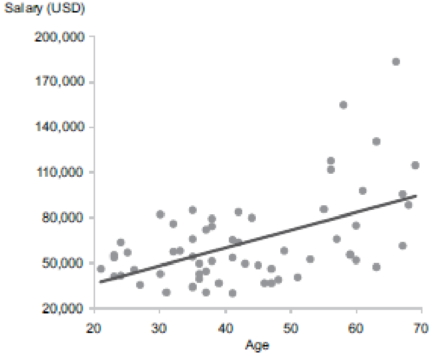
Correlation Analysis Techniques and Best Practices

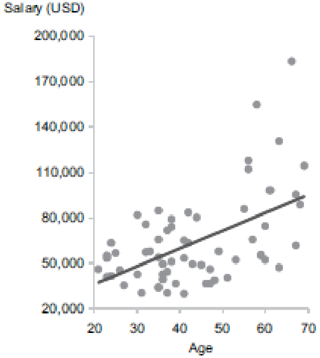
- Optimizing aspect ratio and quantitative scales
- Removing fill color to reduce over-plotting
- Comparing data to reference regions
- Visually Distinguishing Data Sets When Divided into Groups
- Using trend lines to enhance perception of the correlation’s shape, strength, and outliers
- Using multiple trend lines to see categorical differences
- Using trellis and crosstab displays to reduce complexity and over-plotting
- Using grid lines to enhance comparisons between scatterplots

Optimizing Aspect Ratio

- It is recommended to make the plot area roughly square in shape

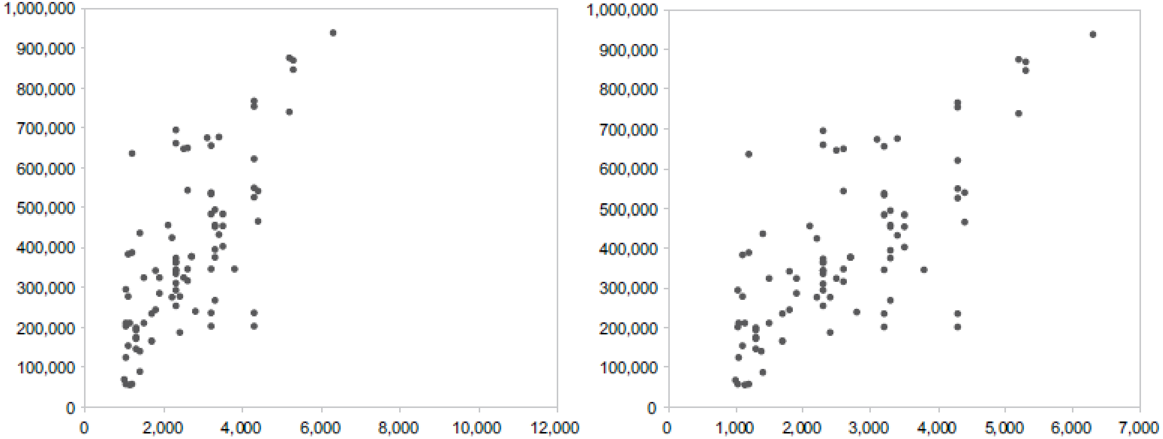






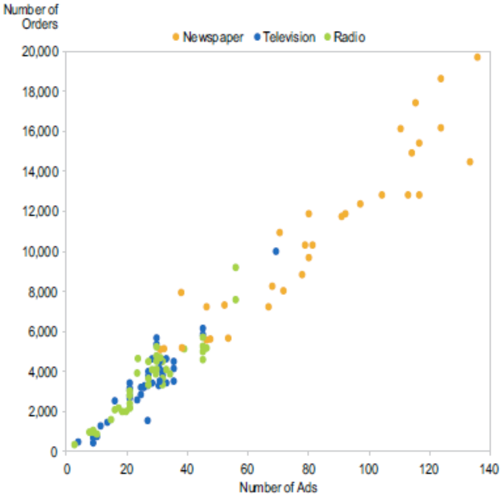
Optimizing Aspect Ratio

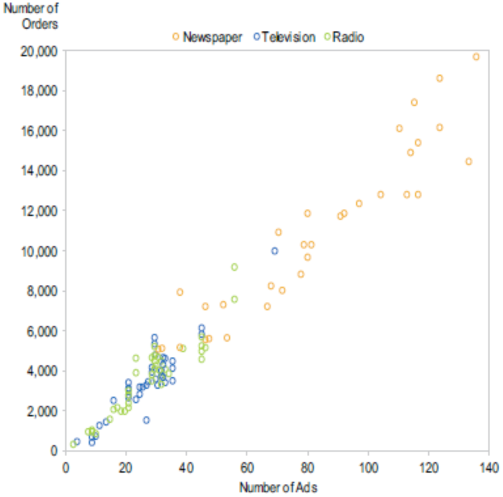
- It is beat to begin each scale a little below the lowest value and end it just a little above the highest.



Remove Fill Color

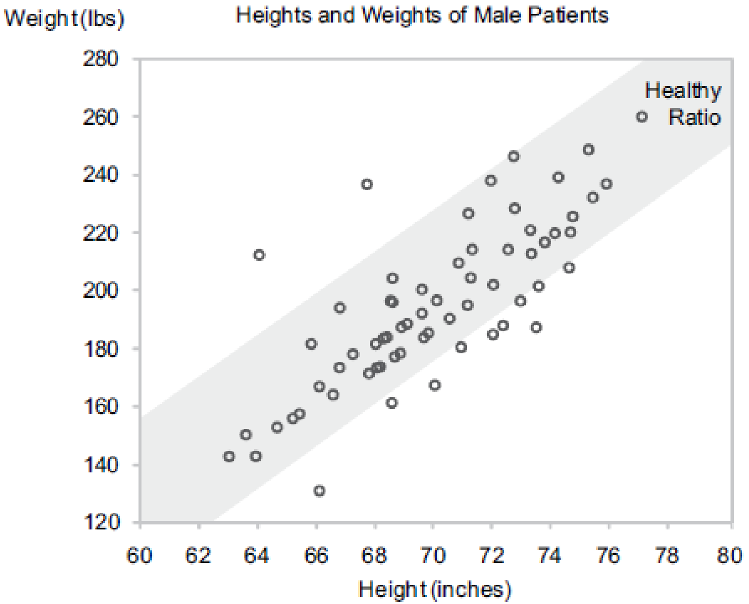
- Using only outlines can remove over-plotting and easier to see when data points overlap.



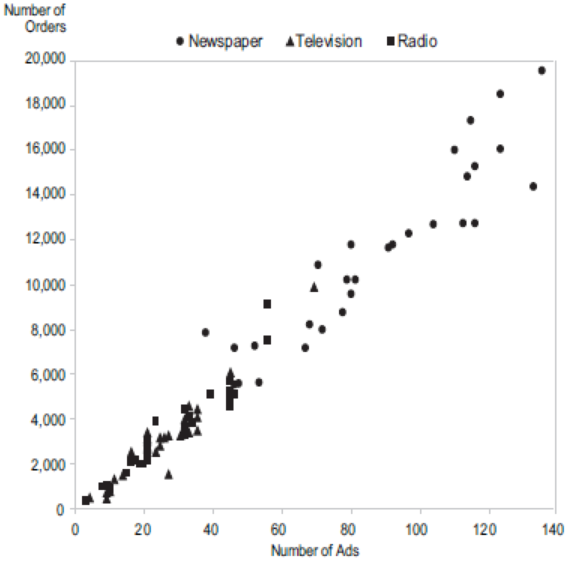


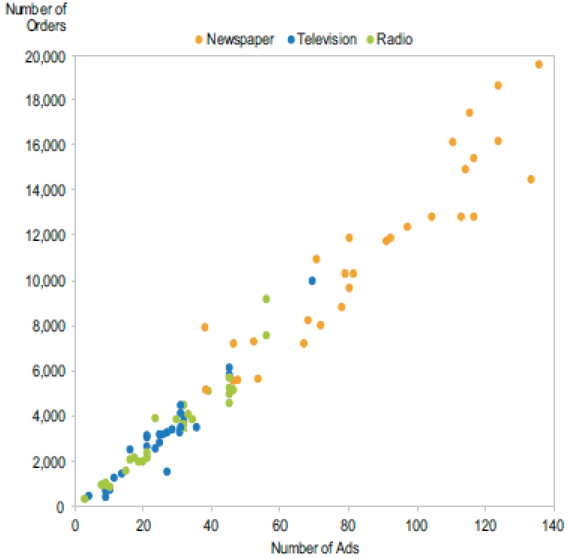
Reference Regions

- Easier to compare data to a reference set of values



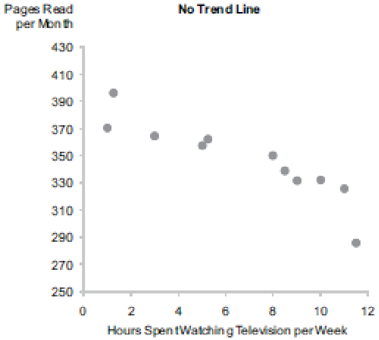
Visually Distinguishing Data Sets When Divided into Groups

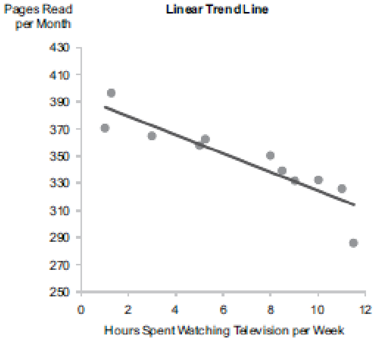


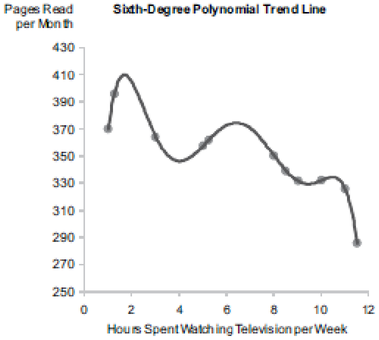


Trend lines

- Trend line traces basic shape of data from left to right.
- A line with least possible amount of residuals → line of best fit and can be used to predict value not in dataset.
- One way is to choose a line of highest r^2

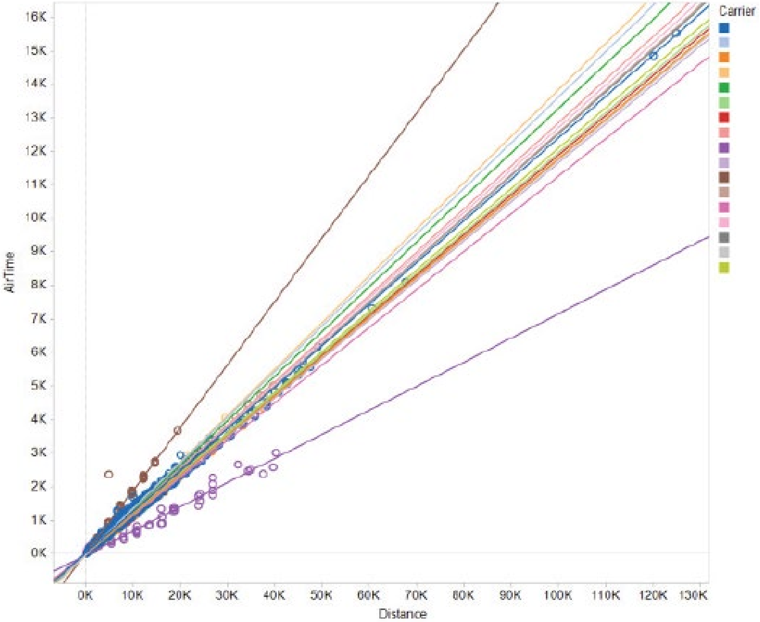






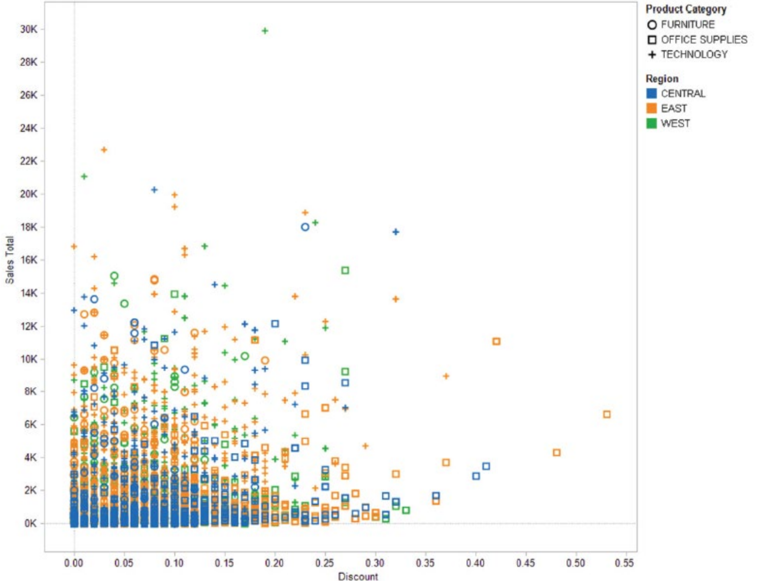
Multiple Trend Lines

All of the flights that took longer belong to a single airline (the brown dots) and all those that took less time belong to a single airline as well (the lavender dots)



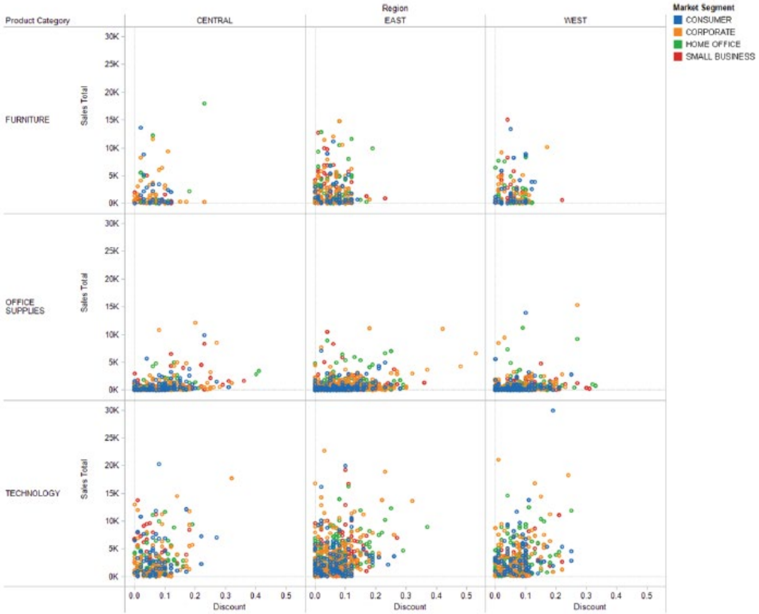
Crosstab Display

- To reduce complexity and Over-plotting



What is this chart trying to tell you?

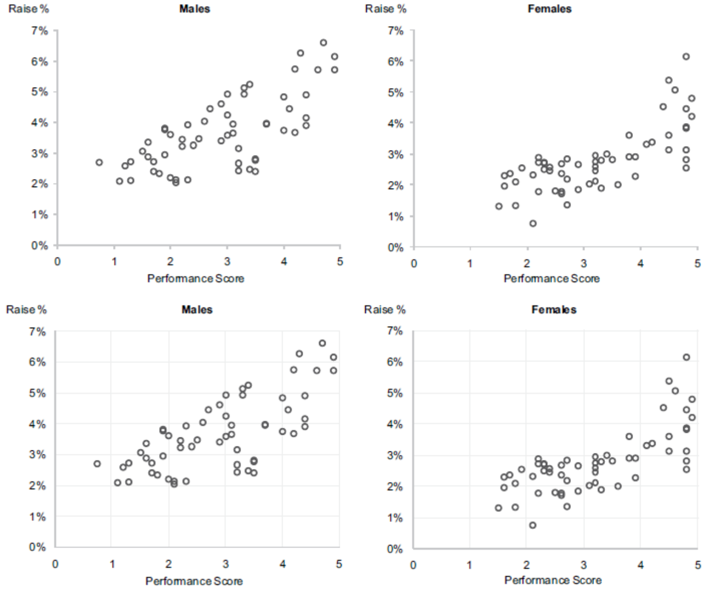
Crosstab Display



By separating each region and product category, we can now compare the correlation patterns formed by each group and more easily spot the differences.

Grid Lines

- To enhance comparisons between scatterplots



Significant difference around performance ratings of 3 and raises from 3% to 5%

Summary

- Visual Characteristics and Statistical Summaries of Correlations
- Correlation patterns
- How to display correlation data
- Techniques and best practices to consider for correlation analysis