MSCI152: Introduction to Business Intelligence and Analytics

Lecture 4: Qualitative Data

Lancaster University Management School

Overview

- Recoding and Aggregating
- Frequency Table
- Charts
  - Bar Charts
  - Pie Charts
  - Lots more charts . . .

Categorical Data Example

We want to analyze the sales of a car dealership to find out about

- mix of sales of different types of cars
- changes since last year

Date	Customer	Model	Version	Colour	Price
3rd Jan	Smith	Mondeo	Zetec	Grey	£18445
3rd Jan	Hendry	Galaxy	Titanium X	Blue	£28545
3rd Jan	Hudson	Focus	RS	Red	£27895
4th Jan	Thompson	C-Max	Style	White	£17245
4th Jan	Hastings	Kuga	Zetec	Blue	£20495
5th Jan	Lewis	Ka	Edge	Blue	£8995
5th Jan	Jones	Mondeo	Titanium	Black	£19945
5th Jan	Cole	Fiesta	Studio	Green	£9995

Price is numerical, other variables are categorical

Recoding and Aggregating

Recode:

- Combine models and versions into fewer categories for type of car
- Decide on the categories for type of car: small, family, 4x4, MPV, sports
- This is nominal data (no particular order)

Aggregate:

- Number of cars of each type sold in the year

Categorical Data Example

Recode: Create categorical variable **Type**, assign values based on the value of **Model**

Date	Customer	Model	Version	Colour	Price	Type
3rd Jan	Smith	Mondeo	Zetec	Grey	£18445	Family
3rd Jan	Hendry	Galaxy	Titanium X	Blue	£28545	MPV
3rd Jan	Hudson	Focus	RS	Red	£27895	Sports
4th Jan	Thompson	C-Max	Style	White	£17245	4x4
4th Jan	Hastings	Kuga	Zetec	Blue	£20495	MPV
5th Jan	Lewis	Ka	Edge	Blue	£8995	Small
5th Jan	Jones	Mondeo	Titanium	Black	£19945	Family
5th Jan	Cole	Fiesta	Studio	Green	£9995	Small

Aggregate: Count the number of different **Types** occuring within each year

Frequency Table

This table contains the data from the whole population

These data are the **counts** (i.e., **Frequency**) of each **Type** (i.e., **Class**) sold in each year

- **Year 1 total sales** = 278+425+159+231+114 = 1207
- **Year 2 total sales** = 252+364+104+172+54 = 946

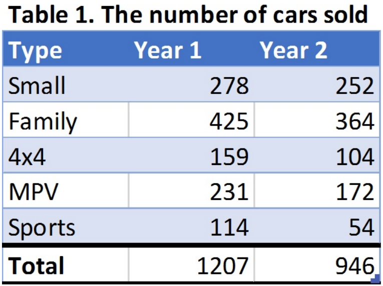


Table 1. The number of cars sold

Type	Year 1	Year 2
Small	278	252
Family	425	364
4x4	159	104
MPV	231	172
Sports	114	54
Total	1207	946

A table needs:

- Title/legend
- Column headings

Helpful:

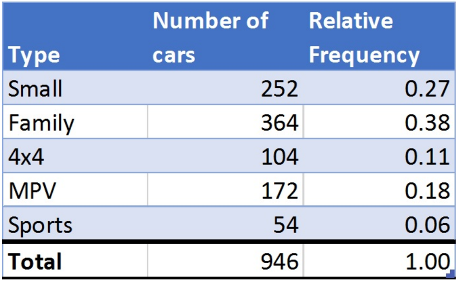
- Totals
- Percentages, Relative Frequency

Relative Frequency Table

Relative Frequency is the ratio between the frequency of a particular class and the count of all measurements (i.e., the total)

- Here, each cell is the **proportion** of all cars sold in Year 2 that were of that type

Table 2. The number of cars sold in Year 2



Type	Number of cars	Relative Frequency
Small	252	0.27
Family	364	0.38
4x4	104	0.11
MPV	172	0.18
Sports	54	0.06
Total	946	1.00

Relative Frequency

$$= \frac{\text{Count of class}}{\text{Total count}}$$

e.g. for 4x4's, the relative frequency

$$= \frac{104}{946} = 0.11$$

Percentage Relative Frequency Table

Percentage relative frequency reports the relative frequency as a percentage rather than a proportion

- It is the relative frequency multiplied by 100
- In Excel, just click the % symbol in the **Number** group in the **Home** tab

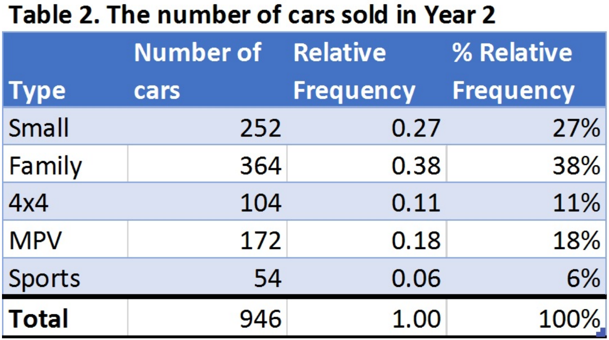


Table 2. The number of cars sold in Year 2

Type	Number of cars	Relative Frequency	% Relative Frequency
Small	252	0.27	27%
Family	364	0.38	38%
4x4	104	0.11	11%
MPV	172	0.18	18%
Sports	54	0.06	6%
Total	946	1.00	100%

Terminology Used

Class

- one of the categories into which qualitative data can be classified

Class frequency

- the number of observations in the data set falling into a particular class

Class relative frequency

- the class frequency divided by the total number of observations in the data set

Class percentage relative frequency

- the class relative frequency multiplied by 100

Frequency Tables

Rows

- categories (classes): every observation falls into **exactly one** category

Columns

- *frequencies*: sum to **total count** of data in class
- *relative frequencies*: sum to **1**
- *percentage relative frequencies*: sum to **100%**

Any table needs

- title/legend, column headings, etc.

Professional-looking tables

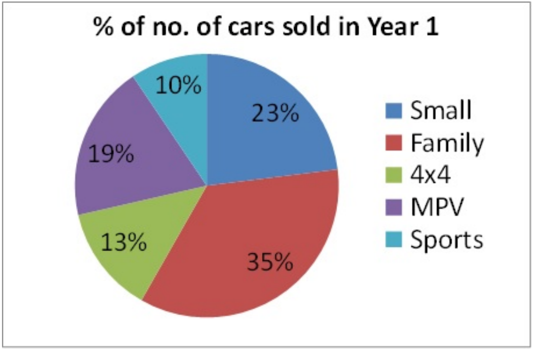
- align columns of text to the left
- align columns of numbers the right
- (e.g., same as Excel)

Charts: Pie Chart

Angle of each segment represents the proportion of the whole "Pie"

- **Angle** $= 360^\circ \times 11\% = 39.6^\circ$

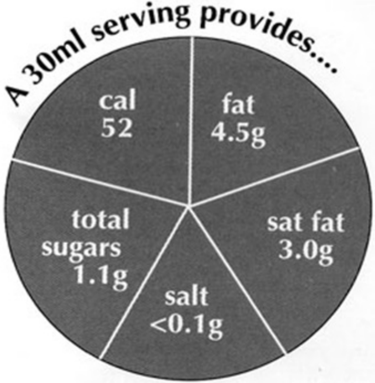
Categories must be separate and add up to a meaningful value



Pie chart needs:

- title
- legend
- values or %

Pie Chart: Spot the mistakes



from label for 'Crème fraiche' (Sainsbury's)
2006

Pie Charts

Segments/portions in pie charts:

- **represent categories:** every observation falls into exactly one category
- **follow the area principle:** angle of each portion represents the percentage of that category
- their **percentage relative frequencies sum up to 100%**

Any pie chart needs:

- title/caption, portion legends, % or numbers

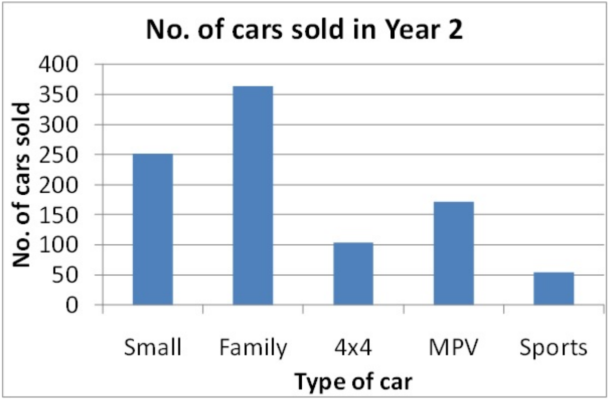
Professional looking pie charts:

- adjust portion colours
- ensure text and text colour is clear and readable

Charts: Bar Chart

Height of bar represents the frequency

- same width of bar
- gaps between bars (categorical data)

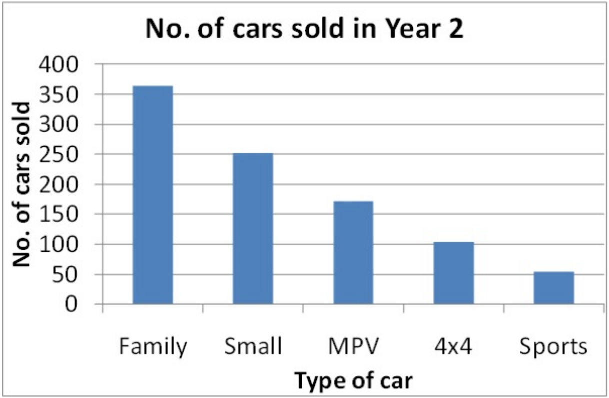


Bar chart needs:

- title
- axis title
- labels

Charts: Pareto Chart

Pareto chart is a bar chart with categories in decreasing order of frequency



Bar Charts

Are sometimes called **column charts**

Bars

- **represent categories:** every observation fits **exactly one** category
- **bar heights** are equal to the count or number of data falling into that category
    - heights of bars sum to the count/number of all data
    - e.g., percentages will sum up to 100%
- **bar width** has no meaning

Any bar chart needs:

- title/caption, axis titles , axis labels, legend

Professional looking bar charts:

- use clear, brief but understandable category labels

Pie chart vs bar chart

Pie Chart Advantages:

- Shows proportion of the whole
- e.g., can see whether category is about a quarter or a half of the total

Bar Chart Advantages:

- Shows relative values more clearly
- can see whether one category is more or less than another category
- Easy to read category labels even for categories with small frequencies

Table 3. The number of cars sold in Years 1 and 2

More complicated tables need good formatting

Complicated data may need more than one table

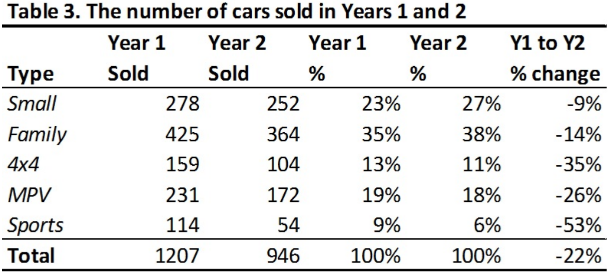


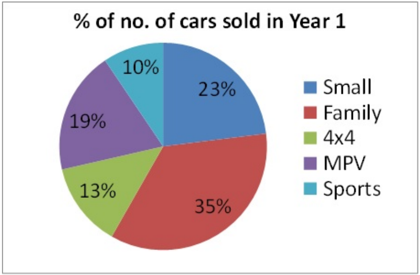
Table 3. The number of cars sold in Years 1 and 2

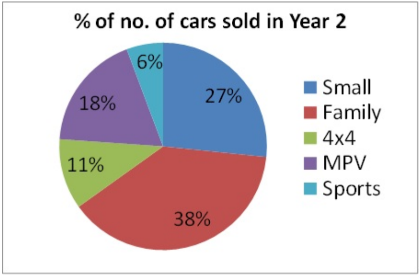
Type	Year 1 Sold	Year 2 Sold	Year 1 %	Year 2 %	Y1 to Y2 % change
Small	278	252	23%	27%	-9%
Family	425	364	35%	38%	-14%
4x4	159	104	13%	11%	-35%
MPV	231	172	19%	18%	-26%
Sports	114	54	9%	6%	-53%
Total	1207	946	100%	100%	-22%

Relative percentage change in sales from Year 1 to Year 2 for **Small** cars

$$\frac{(252 - 278)}{278} = -9\%$$

Comparing Years: Pie Chart





Challenging to **compare between the years**

- % shows the sales mix

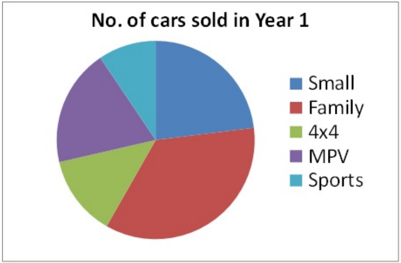
Not suitable if we showed number rather than %

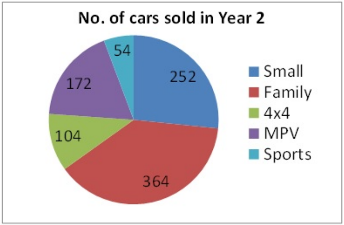
Pie Chart: Area Principle

The **area** should correspond to the value being represented

$$\text{Area} \;= \pi \times \;(\text{radius})^2; \quad \text{radius}_2 = \;\text{radius}_1 \times \sqrt{\frac{\text{total}_2}{\text{total}_1}}$$

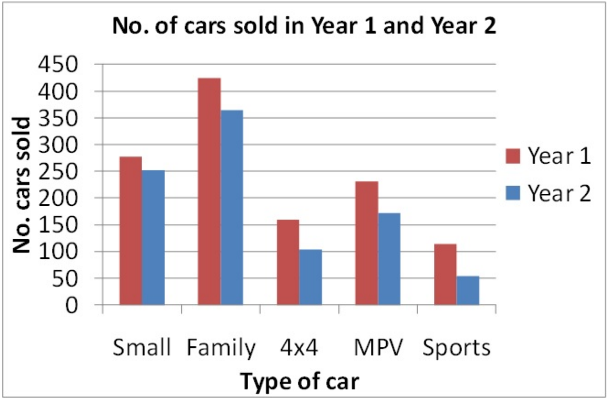
If this is not true then the chart can be **misleading**





How does the number of small cars change from Year 1 to Year 2?

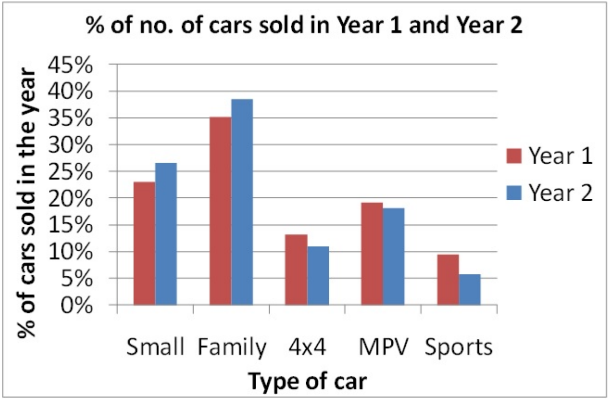
Comparing Years: Bar Charts



In comparative charts you need a legend to distinguish between the level in the data being compared.

- Here **Year 1** and **Year 2**

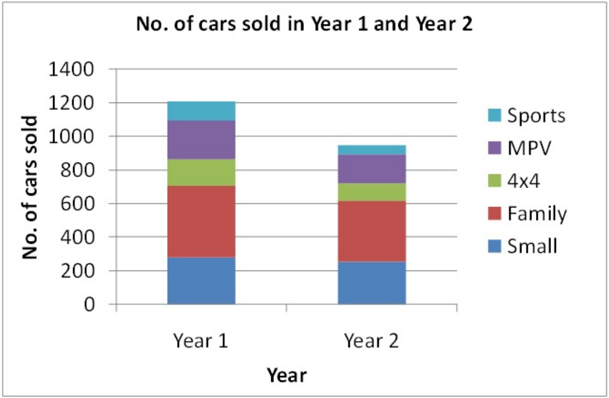
Comparing Years: Percentage Bar Charts



Percentage (%) values show the change in **sales mix**

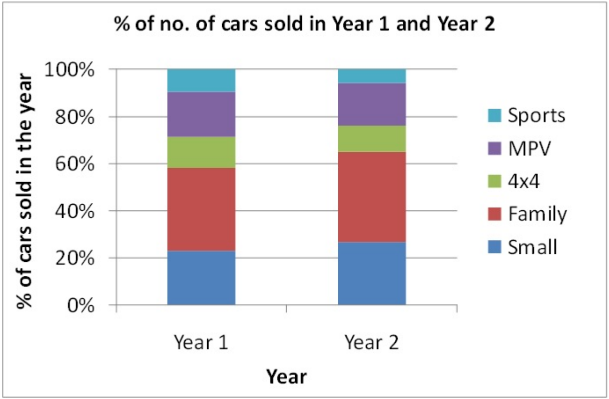
- Could be very misleading unless read carefully

Comparing Years: Stacked Bar Charts



Can be effective particularly if just 2 or 3 categories

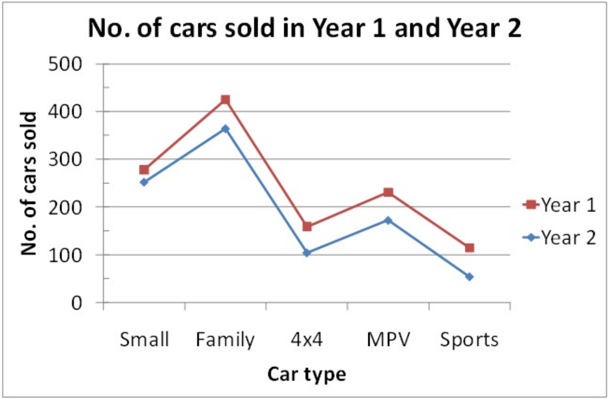
Comparing Years: Stacked (%) Bar Charts



Percentage (%) values show the change in **sales mix**

- Similar to comparative pie charts

Other Charts: Line Chart

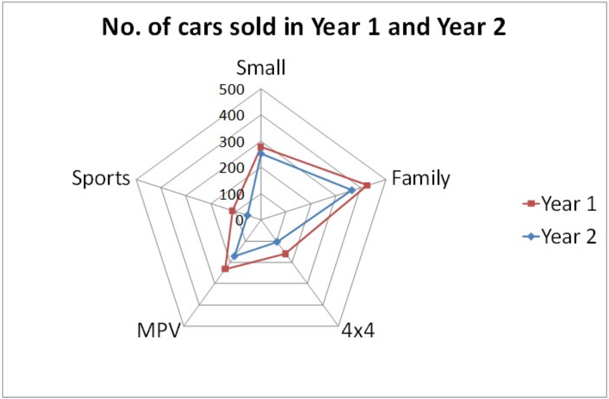


Like bar chart: draws line across tops of bars

Good for comparisons

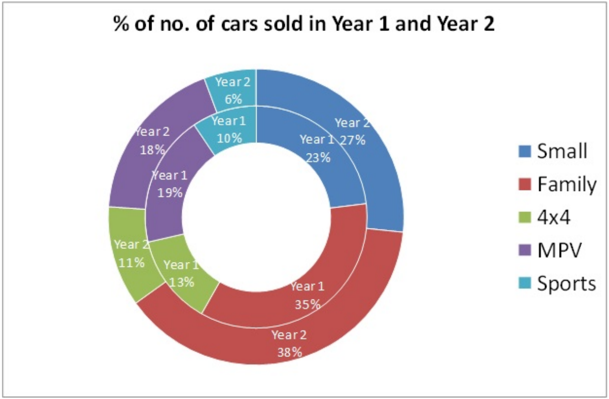
Careful: Must not be read as trend across the *x*-axis (categorical)

Other Charts: Radar or Spiderweb Chart



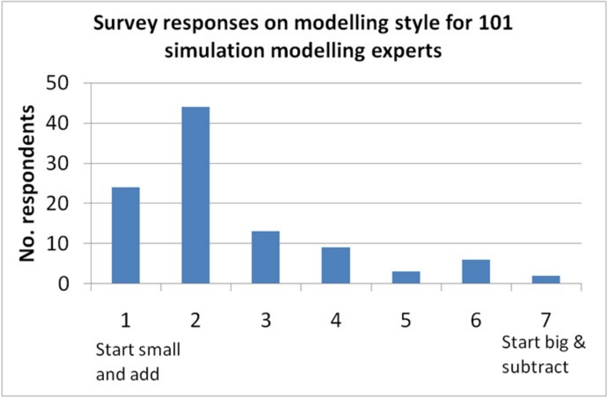
Gaining popularity

Other Charts: Doughnut Chart



Not recommended: potentially misleading as it violates the area principle

Other Charts: Bar Chart for Ordinal Data



Plot in **order** – the resulting pattern looking across the bars has meaning.

Bar chart better than pie chart for ordinal data.

Summary

Not difficult to draw a chart in Excel

But, high quality professional analysis and presentation requires:

- Careful thought in basic analysis of data
- Careful thought in choice of tables and charts
- Attention to detail
- titles, axis labels, data labels, legend, colours
- Ability to interpret results and identify key points
- Clear writing

Wrap up

Here we:

- Looked at a range of charts for **qualitative** data

Next time:

- Charts for **quantitative** data