MSCI152: Introduction to Business Intelligence and Analytics

Lecture 5: Quantitative Data

Lancaster University Management School

Overview

- Graphs
- Outliers
- Presenting graphs

Sales Data

Sales data from 50 stores has been collected. It looks like this:

27.4, 85.2, 75.6, 54.6, 79.3, 76.9, 62.1, 28.1, 86.3, 86.9,
53.0, 87.1, 68.7, 72.4, 48.6, 62.4, 61.1, 103.6, 78.1, 64.0,
69.0, 55.7, 77.9, 54.2, 68.7, 80.2, 42.7, 73.8, 75.8, 84.2,
49.1, 51.9, 78.0, 57.4, 68.8, 57.6, 66.6, 100.1, 90.8, 46.3,
74.7, 88.7, 89.4, 78.9, 61.7, 61.4, 64.5, 50.3, 55.8, 50.6

- What can you tell me about this data?
- How can we make sense of it?

Visualising Quantitative Data

These graphs give an idea of the look, shape and distribution of the data

- histogram
- frequency polygon
- cumulative frequency chart* (in Measures of Spread)
- box plot* (in Measures of Spread)
- scatter plot
- time series

* These charts will be discussed in later in “Measures of Spread”

Store sales data approach

To see the pattern in the data we create a **frequency table** as for the car sales data:

- **Combine and Aggregate:** Create categories and count how many are in each category
- Here the categories will be (consecutive) numerical intervals

Next, we draw a chart of the categorised interval data:

- Histogram – similar to a bar chart but appropriate for numerical data

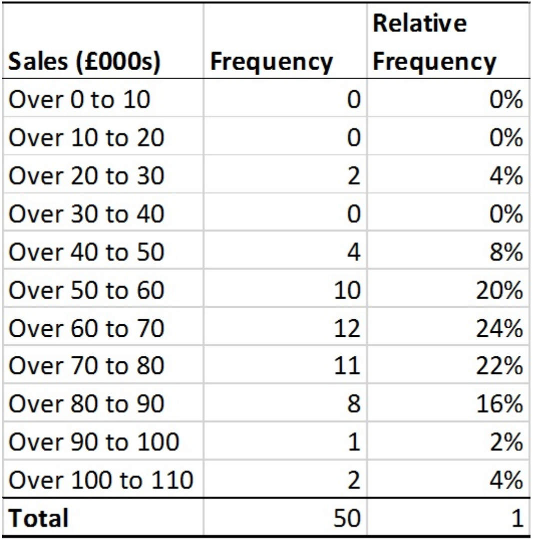
Frequency Table

Decide on the intervals:

- I have decided to have intervals of width £10,000
  - Over 0 to 10,
  - Over 10 to 20,
  - etc.
- Following Excel only **one boundary value** allowed in interval
  - e.g., a value of 10 can't be in both "0 to 10" and "10 to 20"!
- You should always use equal widths – see later
- Count the number of values in each interval

In Excel we can use the Histogram tool or COUNTIF function

Sales Frequency Table



		Relative
Sales (£000s)	**Frequency**	**Frequency**
Over 0 to 10	0	0%
Over 10 to 20	0	0%
Over 20 to 30	2	4%
Over 30 to 40	0	0%
Over 40 to 50	4	8%
Over 50 to 60	10	20%
Over 60 to 70	12	24%
Over 70 to 80	11	22%
Over 80 to 90	8	16%
Over 90 to 100	1	2%
Over 100 to 110	2	4%
Total	50	1

The table itself can help with understanding and presenting data

See Lecture 5 for *Relative Frequency*

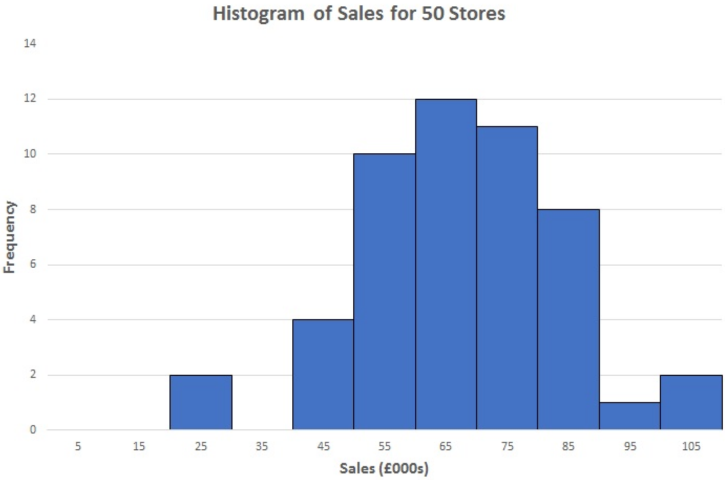
Histogram

Histogram is a chart that shows the distribution of the data.

Excel is a pain, as it draws Histograms as bar charts, so some key pointers:

- Numerical scale on x-axis
- Want to put values at “ticks” where bars join, but Excel labels “categories”
    - Can put mid-point of interval as category
    - Can put interval range also
- There should be no gaps between bars - contiguous interval scale

Histogram of Sales of 50 Stores



Histogram in Excel (pre-2016 version)

Column chart is closest

- Excel treats this as category data
- Can only label bars, not axis

Manipulate Excel:

- Format Data Series
- Series Options
- Change Gap width to 0

Histogram in Excel (2016 version)

Now has “Histogram” chart:

- Still a column chart
- Excel still treats this as category data
- Still can only label the bars, not the axis

Main advantage:

- Can easily try different intervals and see the results
- Format Axis – Axis Options

Limitations:

- intervals can only start at first data value, first interval not quite right

See next week’s workshop

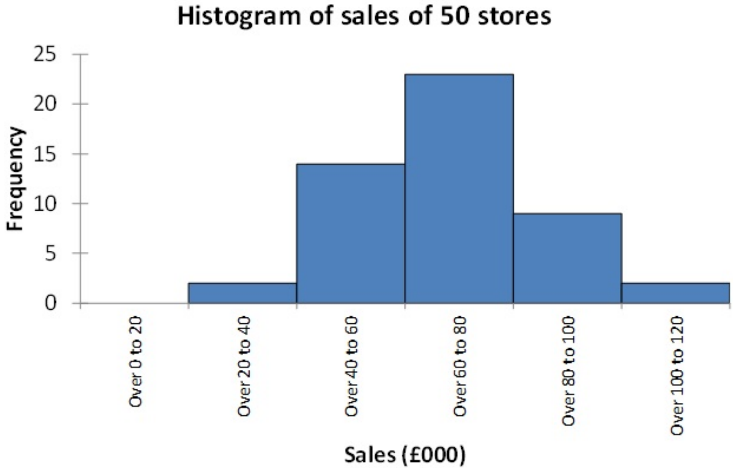
Bar Chart vs. Histogram

	Histogram	**Bar Chart**
Categories	Numerical Intervals	Qualitative Characteristics
x-axis	Numerical Scale	Description of Categories
Width of bars	Width of Interval	All the same
Gaps between Bars	**NO**	**YES**
Frequency	**AREA*** of Bar	**HEIGHT** of Bar

$*$ If the numerical intervals are of **equal width** the height also represents the frequency

Histogram issue 1

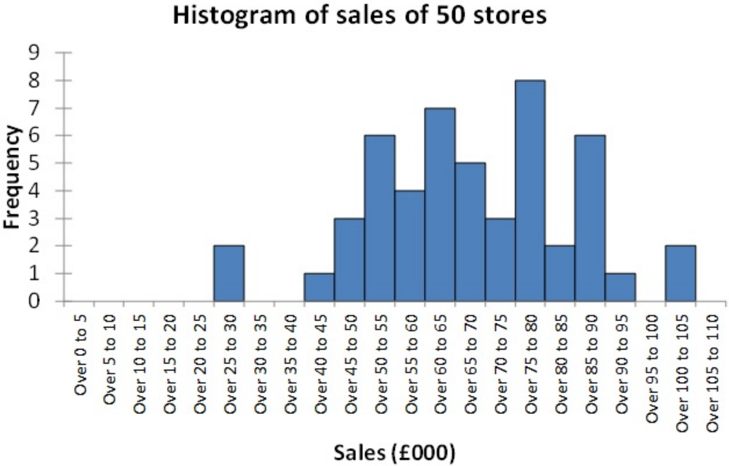
Pattern partly depends on number of intervals



Too few intervals and so not enough detail

Histogram issue 1

Pattern partly depends on number of intervals



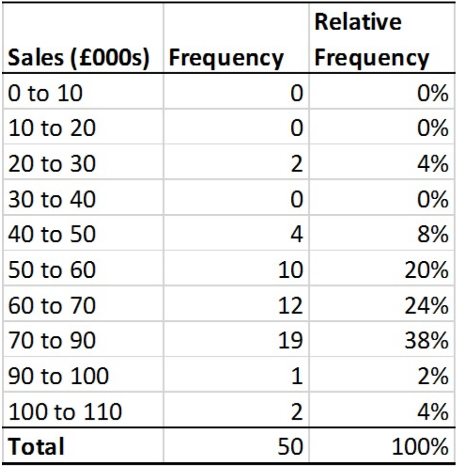
Too many intervals and so more difficult to see the shape

Histogram issue 2

Take care if intervals are of **different widths**

- Recall it is the **area of the bars** that measures the frequency
- Excel does not really draw a proper histogram and so stick to **equal width intervals**
- Note, most software packages use equal widths by default

Frequency table with unequal width intervals



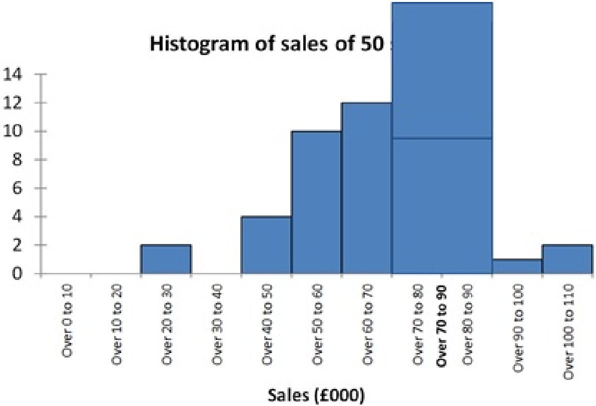
Sales (£000s)	Frequency	Relative Frequency
0 to 10	0	0%
10 to 20	0	0%
20 to 30	2	4%
30 to 40	0	0%
40 to 50	4	8%
50 to 60	10	20%
60 to 70	12	24%
70 to 90	19	38%
90 to 100	1	2%
100 to 110	2	4%
Total	50	100%

Misleading:

70–90 Combines

- 70–80: 11
- 80–90: 8

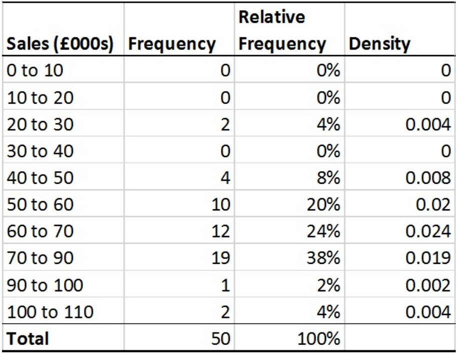
Histogram issue 2



If we have **unequal interval widths** we cannot use (Relative) Frequency

- Use **Density** (Frequency Density) to represent area of interval

Calculating the Density



Sales (£000s)	Frequency	Relative Frequency	Density
0 to 10	0	0%	0
10 to 20	0	0%	0
20 to 30	2	4%	0.004
30 to 40	0	0%	0
40 to 50	4	8%	0.008
50 to 60	10	20%	0.02
60 to 70	12	24%	0.024
70 to 90	19	38%	0.019
90 to 100	1	2%	0.002
100 to 110	2	4%	0.004
Total	50	100%	

Area Principle:

Divide Relative Frequency by interval width, e.g.

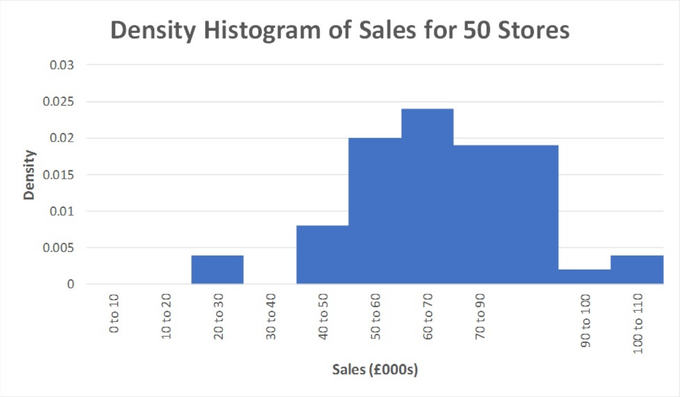
Over 20 to 30:

$$\frac{4\%}{10} = \frac{0.04}{10} = 0.004$$

Over 70 to 90:

$$\frac{38\%}{20} = \frac{0.38}{20} = 0.019$$

Histogram issue 2



The distribution of the data is now correct

But better to use **equal widths** wherever possible!

Issue: Outliers in data

What if the data we had originally been given looked like this?

27.4, 85.2, 75.6, 54.6, 79.3, 76.9, 62.1, 28.1, 86.3, 86.9,
53.0, 87.1, 68.7, 72.4, 48.6, 62.4, 61.1, 103.6, 78.1, 64.0,
69.0, 557, 77.9, 54.2, 68.7, 80.2, 42.7, 73.8, 75.8, 84.2,
49.1, 51.9, 78.0, 57.4, 68.8, 57.6, 66.6, 100.1, 90.8, 46.3,
74.7, 88.7, 89.4, 78.9, 61.7, 61.4, 64.5, 50.3, 55.8, 50.6

Is there anything unusual about these data?

Let's look at a histogram of the data...

Issue: Outliers in data

In general, an outlier means an unusual value

- In some cases there is a specific definition
- e.g., based on distance from some measure of location

In any data analysis we need to look out for such values and investigate:

- **Could be a correct value:** find out the reason for it
- **Could be incorrect:** replace by correct value or delete it
- **When reporting:** state any issue identified and any action taken

Comparing Distributions

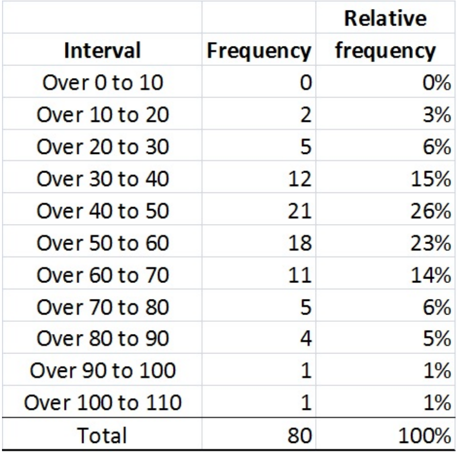
Sales of stores in region 2 in £000s:

50.6, 40.7, 71.1, 34.3, 53.6, 33.3, 34.7, 33.8, 49.7, 42.2,
48.1, 46.9, 58.5, 37.6, 88.1, 40.3, 54.5, 40.1, 46.7, 22.4,
54.2, 80.3, 56.2, 53.5, 46.1, 50.1, 18.5, 72.4, 66.0, 63.1,
56.7, 68.3, 53.0, 54.1, 39.5, 50.7, 69.7, 53.8, 18.5, 40.7,
35.5, 45.3, 44.3, 91.6, 68.9, 62.0, 61.2, 51.8, 44.3, 72.8,
54.2, 21.6, 39.9, 27.9, 42.5, 56.6, 66.4, 41.5, 45.1, 58.3,
62.9, 37.8, 107.6, 75.6, 23.0, 43.4, 42.0, 82.7, 31.3, 53.5,
60.1, 37.9, 39.5, 44.1, 65.6, 89.0, 72.8, 49.0, 45.2, 20.1

Compare using summary statistics and charts

Let's looks at some charts

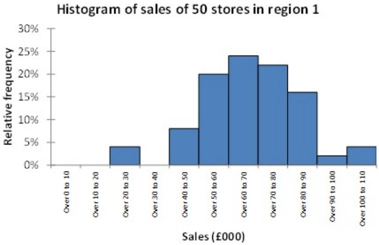
Comparing Distributions

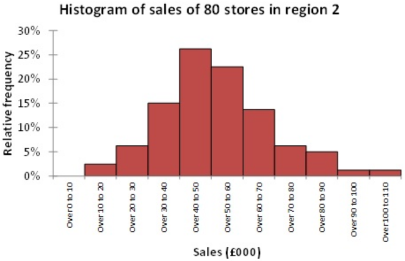


		Relative
Interval	Frequency	frequency
Over 0 to 10	0	0%
Over 10 to 20	2	3%
Over 20 to 30	5	6%
Over 30 to 40	12	15%
Over 40 to 50	21	26%
Over 50 to 60	18	23%
Over 60 to 70	11	14%
Over 70 to 80	5	6%
Over 80 to 90	4	5%
Over 90 to 100	1	1%
Over 100 to 110	1	1%
Total	80	100%

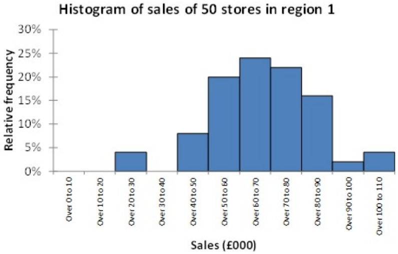
80 stores (compared to 50 stores in region 1) so use relative frequency to compare them

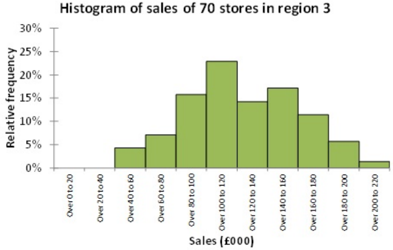
Comparing region 1 and region 2 (1)





Comparing axes scales (1)

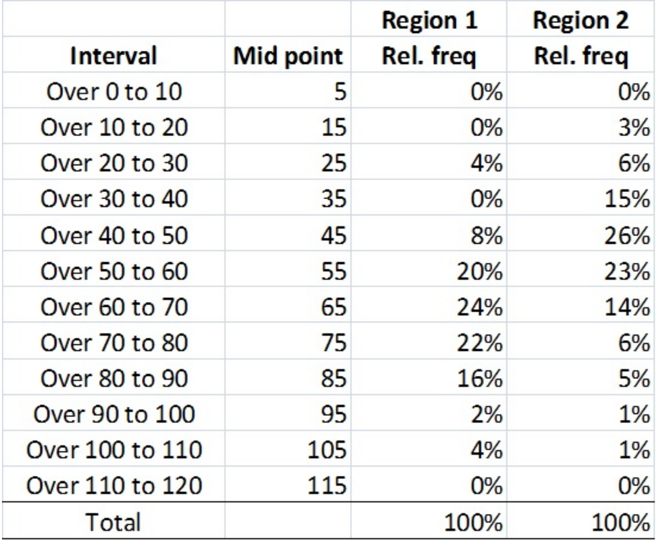




x-axis scales do not match

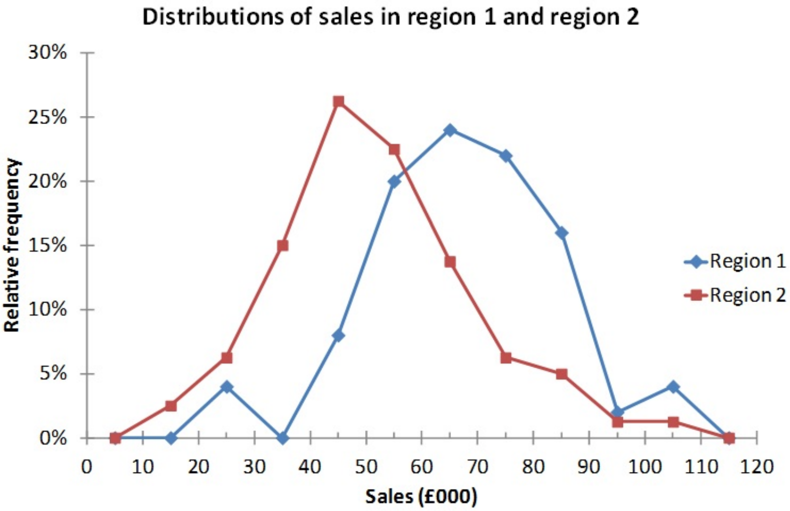
May reach incorrect conclusion

Comparing region 1 and region 2



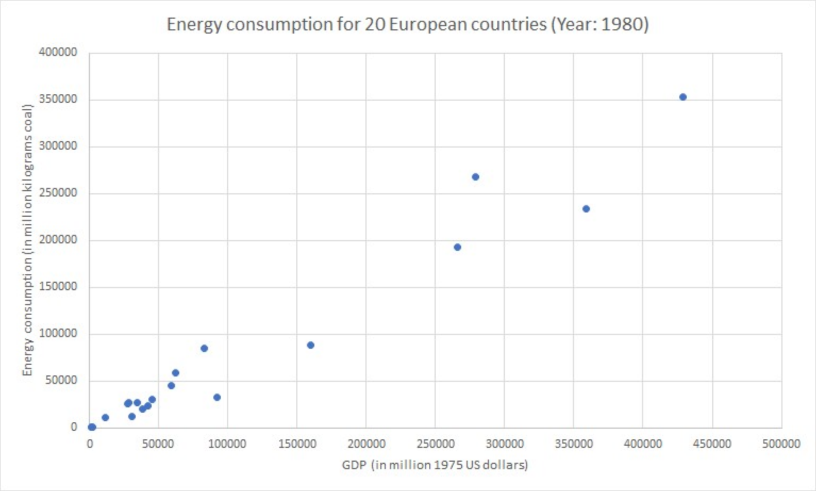
		Region 1	Region 2
Interval	Mid point	Rel. freq	Rel. freq
Over 0 to 10	5	0%	0%
Over 10 to 20	15	0%	3%
Over 20 to 30	25	4%	6%
Over 30 to 40	35	0%	15%
Over 40 to 50	45	8%	26%
Over 50 to 60	55	20%	23%
Over 60 to 70	65	24%	14%
Over 70 to 80	75	22%	6%
Over 80 to 90	85	16%	5%
Over 90 to 100	95	2%	1%
Over 100 to 110	105	4%	1%
Over 110 to 120	115	0%	0%
Total		100%	100%

Comparing: Frequency polygon



Excel: X-Y scatter chart – do not forget the legend!!

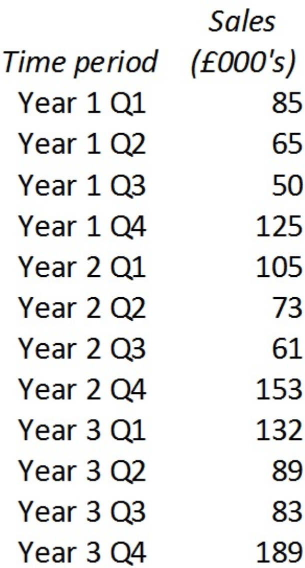
Relationships: Scatter plot



Excel: Scatter Chart

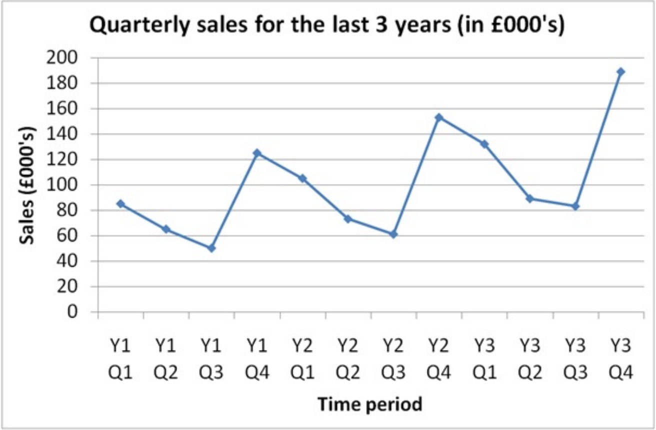
[Source: Baltagi, B.H. (2002). Econometrics, 3rd ed. Berlin, Springer]

Time Series Data



Time period	*Sales (£000's)*
Year 1 Q1	85
Year 1 Q2	65
Year 1 Q3	50
Year 1 Q4	125
Year 2 Q1	105
Year 2 Q2	73
Year 2 Q3	61
Year 2 Q4	153
Year 3 Q1	132
Year 3 Q2	89
Year 3 Q3	83
Year 3 Q4	189

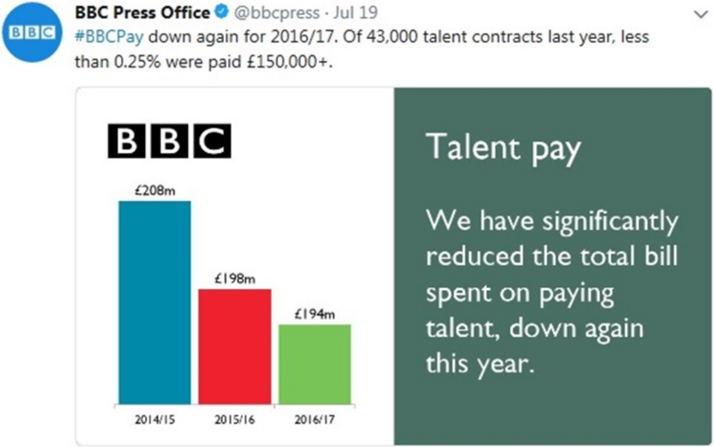
Time Series Chart



Excel: scatter or line chart

If *x*-axis data are numbers use a scatter chart; if labels then use a line chart

BBC Press Office Tweet



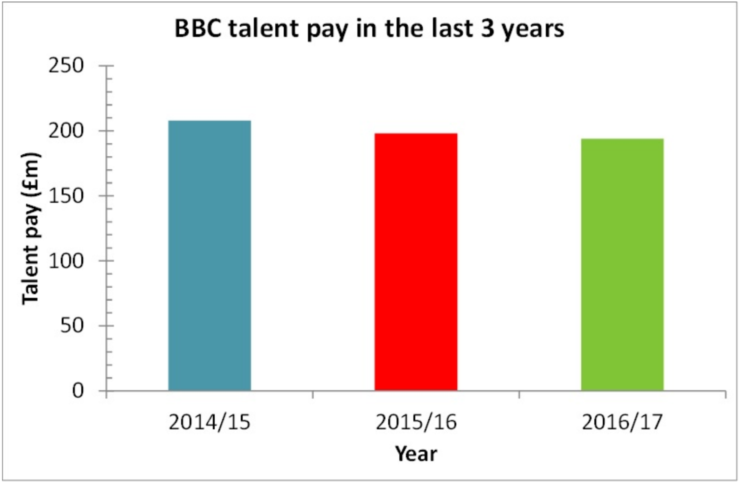
BBC Press Office @bbcpress · Jul 19

#BBCPay down again for 2016/17. Of 43,000 talent contracts last year, less than 0.25% were paid £150,000+.

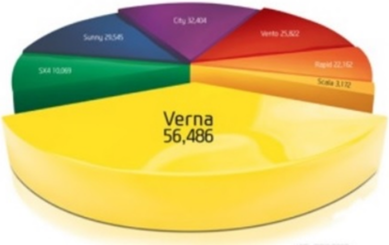
Talent pay

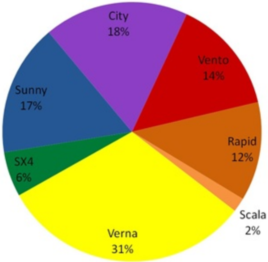
We have significantly reduced the total bill spent on paying talent, down again this year.

BBC pay Chart



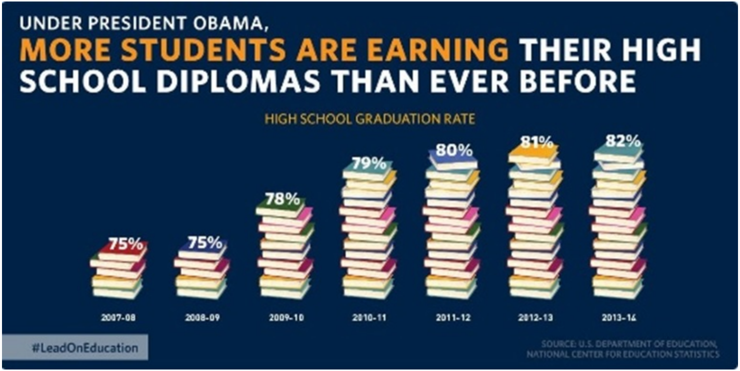
AREA PRINCIPLE





Important: The area in a chart must correspond to the value. Otherwise the chart is visually misleading

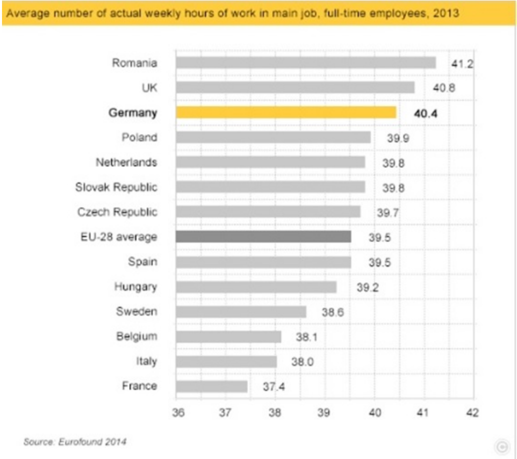
Truncation is a common problem!



Note also that 75% is 5 books, 78% 10 books, 79% 14 books

Source: White House tweet

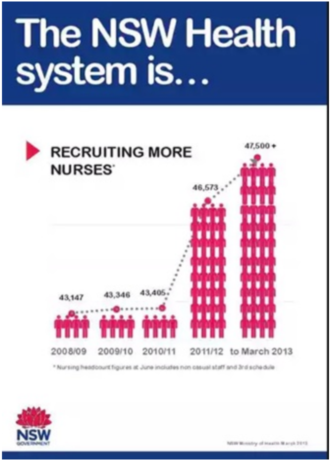
Truncation is a common problem!



Wanting to encourage companies to locate and invest in Germany

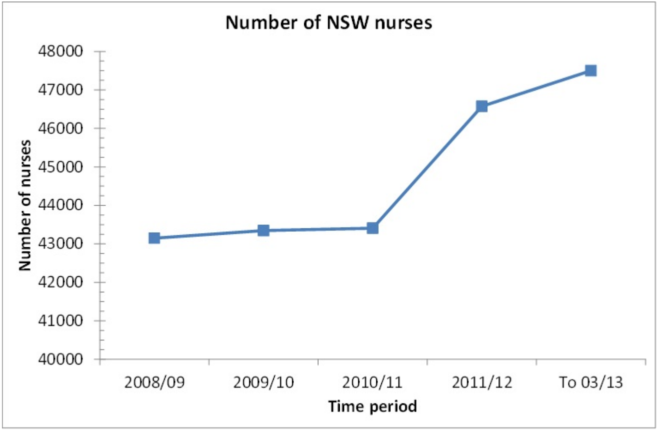
Source: Germany Trade & Invest (www.gtai.de)

Truncation is a common problem!



Source: theguardian.com

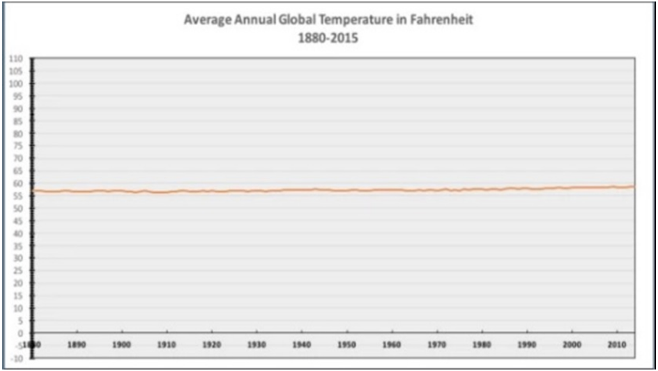
Using a line chart



Not too bad: Focus on change than actual value

Global Temperature

What is wrong with this?



Source: National Review tweet

Wrap up

Today we:

- Looked at charts for numerical data

Next time:

- we will discuss summarising data