Visualizing data

Why it's important
How to do it well

Graphs: not just pretty pictures

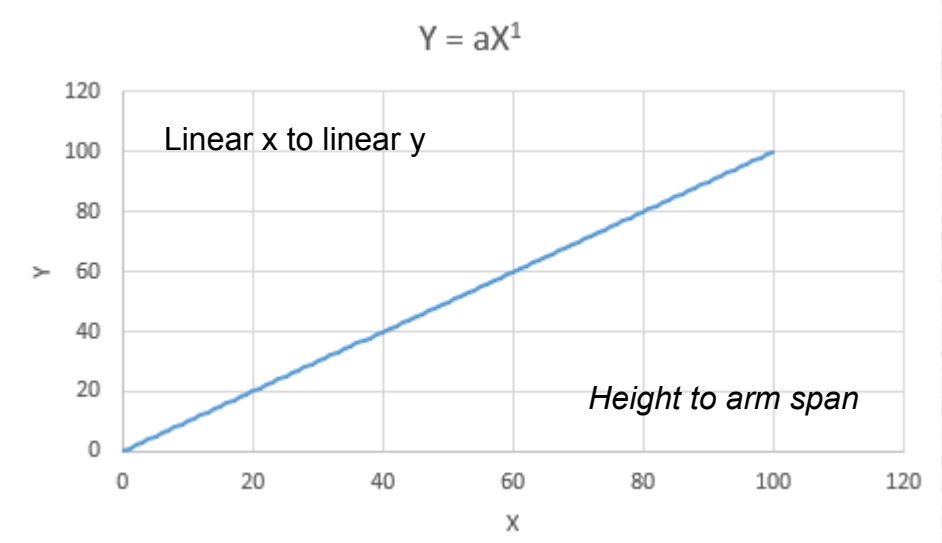
- Visualization is an important part of correct data analysis
  - Deriving meaning from data
    - Can support analysis
    - Can be a form of analysis
  - Especially important for large data sets
- Visualization is an important part of communicating results to others
  - We're visual creatures
  - A picture is worth a thousand words

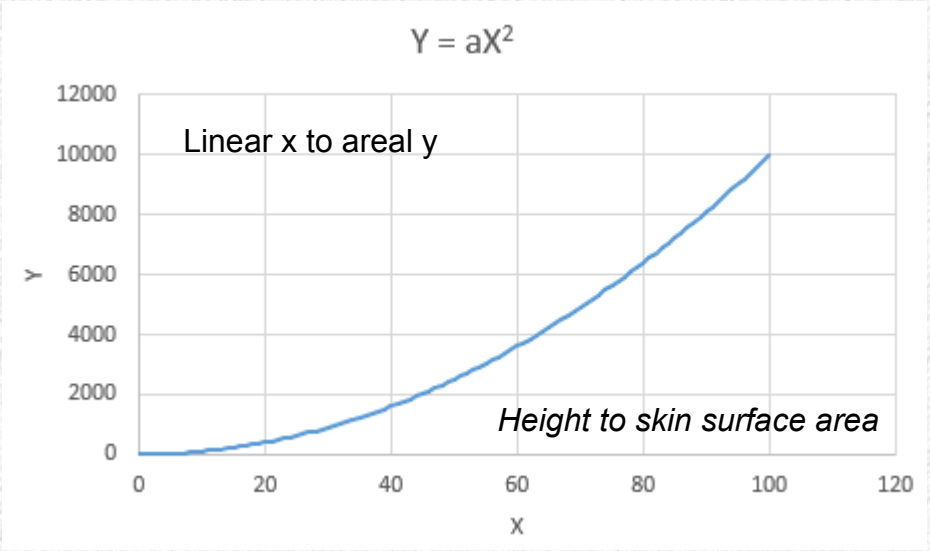
Graphs as a form of analysis: finding the relationship between variables

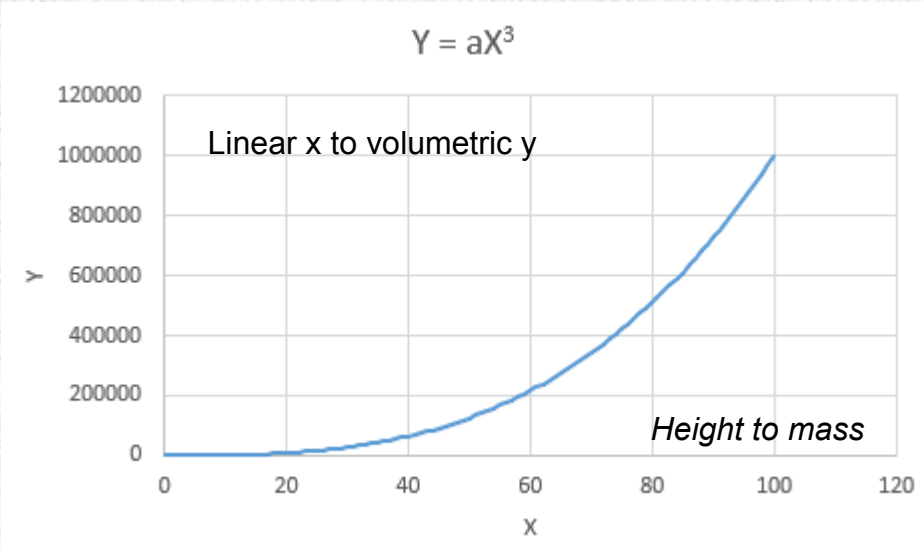
- If we want to know how one variable changes in response to another we need to know how they're related
- You can use graphical methods to find the functional relationship between variables
  - Different functions are straight lines on plots with logarithmic axes
  - Setting axes to log scale can help you identify the underlying functional relationship

Three power function relationships between x and y

$$Y = a\,X^{b}$$







Most body parts scale relative to one another by a power function

- Exponent depends on the variables

Power functions – straight lines on log-log plots

$$Y = a\,X^{b}$$

If the y-axis is log scale and the x-axis is log scale the relationship straightens out

$$\log(Y) = \log(a) + b\log(X)$$

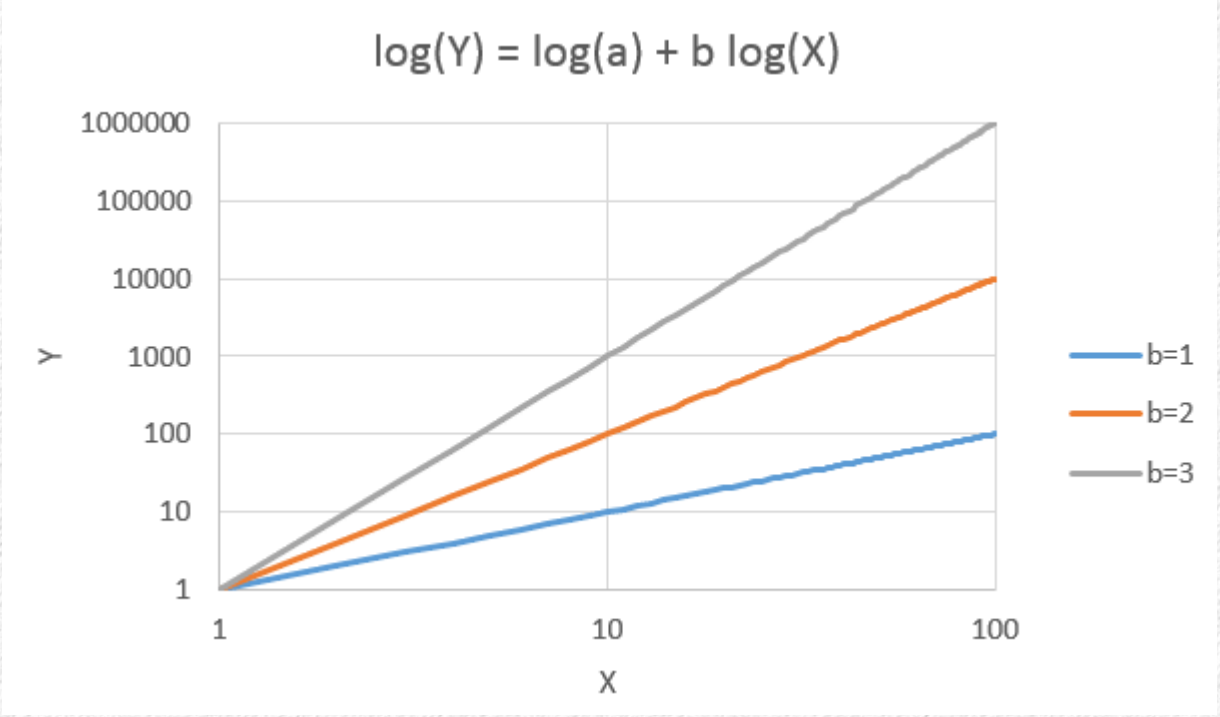
On a log scale has the form y = b + mx

Log-log plots

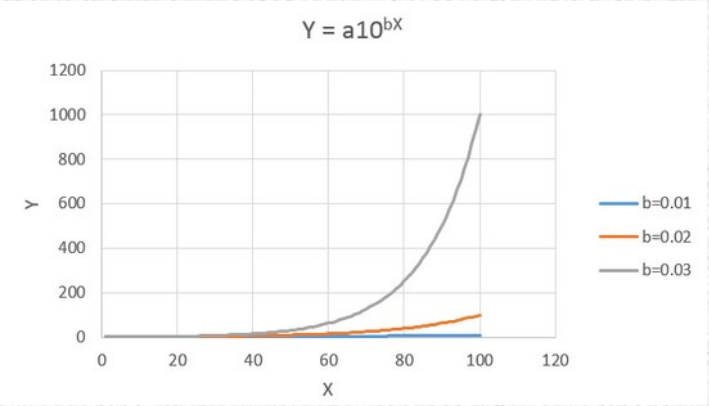
Both axes on a log base 10 scale

→ each tick is a 10-fold increase (order of magnitude)

If data are a straight line on a log-log plot, then the relationship is a power function

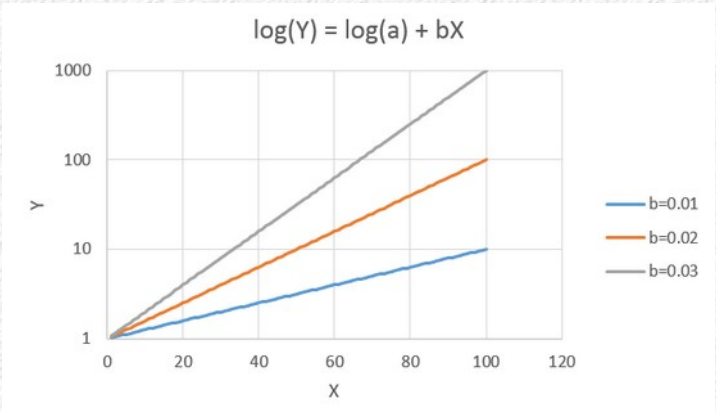


Exponential relationships



Constant multiplicative changes (population growth, radioactive decay)

$$Y = a\,10^{bX}$$



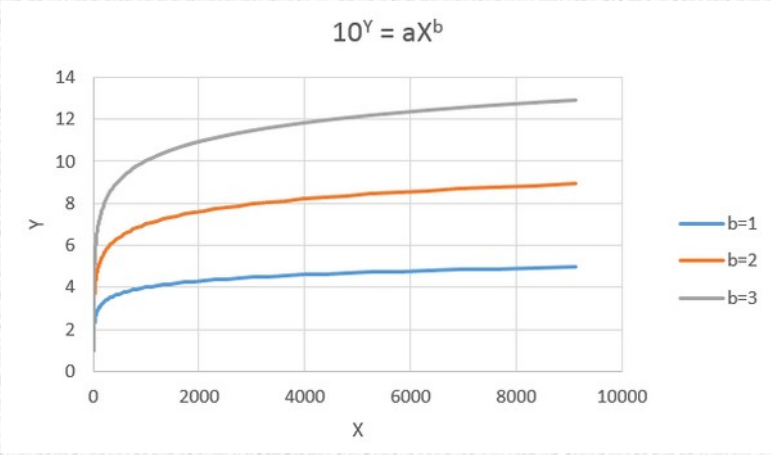
$$\log(Y) = \log(a) + bX$$

Straightens out when Y is on a log scale, X is linear

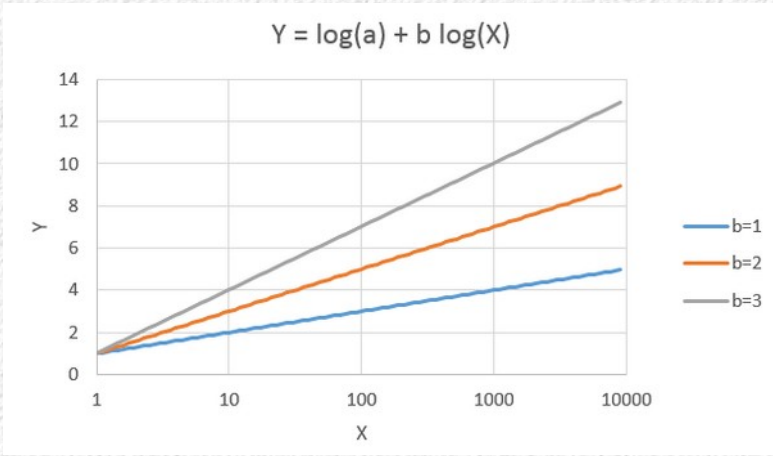
Logarithmic relationships

pH, fold changes for gene expression

$$10^{Y} = aX^{b}$$

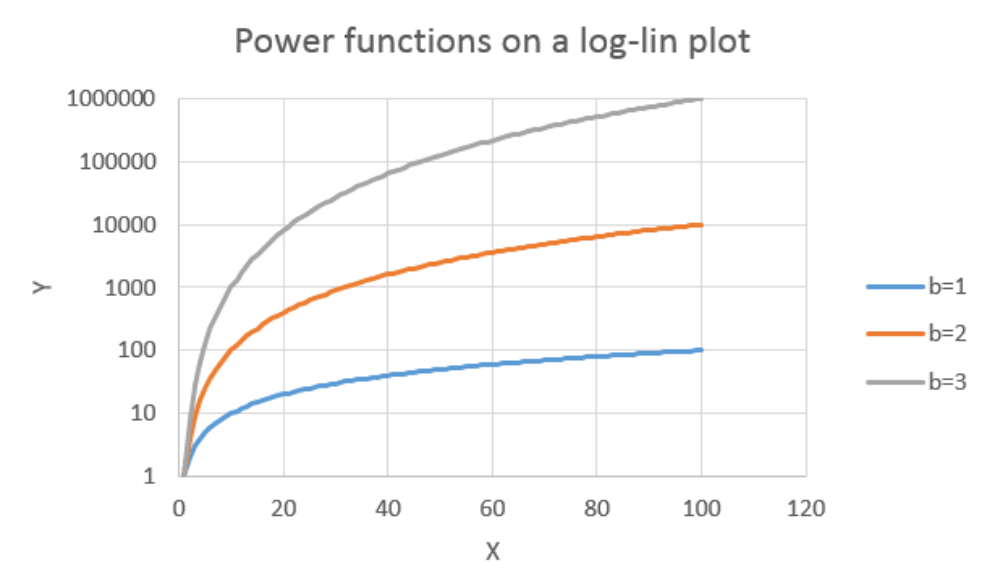


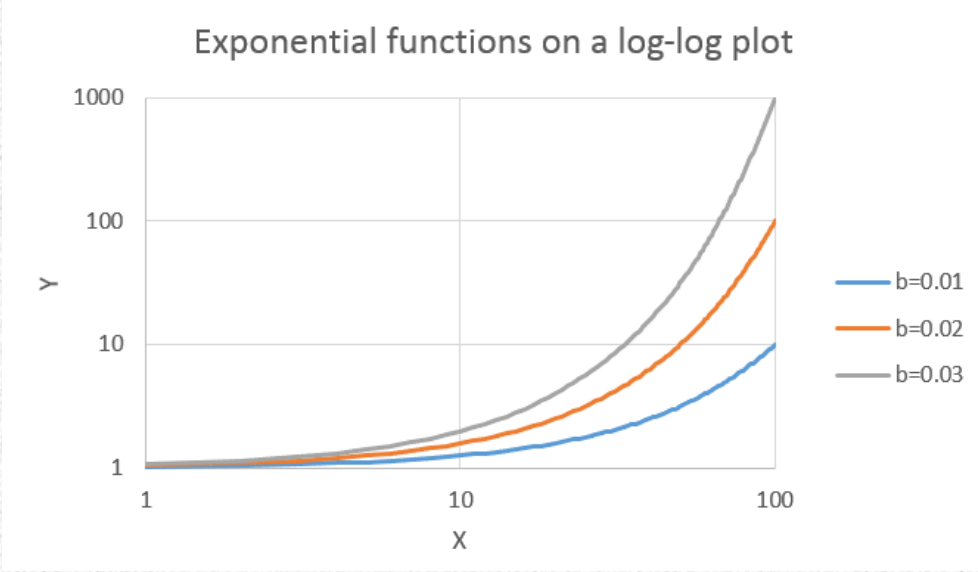
$$Y = \log(a) + b\log(X)$$



Becomes a straight line when Y is on a linear scale, X is logarithmic

Wrong choice of axis scales → curved lines





Thus, changing axis scale from linear to logarithmic can help you diagnose the functional relationship between variables

Once we know the relationship, we can have Excel give us the equation for the line

Plickers...

Graph types for data display

- Common types of graphs differ in:
  - Axis types – numeric, categorical, or a mix of the two
    - Choice of graph type depends on the variable types of the data
  - Axis scaling conventions – forcing axes to go to 0 or not
- Problem: the computer doesn't know what your variables are, only knows what you ask it to do
  - If you ask Excel to make the wrong graph type for your data, if it can it will
- Common graph types (i.e. those supported by Excel) cover most of the basic data display tasks

Excel's graph types

Graph type	Use
Column 	A numeric variable plotted at levels of one or more categorical variables
Bar 	A horizontal column chart
Line 	Values of a numeric variable displayed at the same levels of a categorical variable
Pie 	Composition data = frequencies, proportions, percentages
Area 	Line graph with the area below the lines shaded
Scatter 	Relationship between two numeric variables
Surface 	A three dimensional surface, with categorical x and y, and numeric z
Bubble 	A scatter plot with symbol size set to display a third variable
Radar 	Each numeric variable is a ray, each observation is plotted on each ray, with points connected
Histogram 	Frequency distribution of a binned numeric variable
Box plot 	Distribution statistics for a numeric variable

Column chart

Data in Excel

	A	B
1	Treatmen	Height
2	Control	11.5
3	Control	7.0
4	Control	10.9
5	Control	13.0
6	Control	7.2
7	Control	5.6
8	Control	9.0
9	Control	10.8
10	Control	9.7
11	Control	9.8
12	Fertilized	14.5
13	Fertilized	12.8
14	Fertilized	15.9
15	Fertilized	14.7
16	Fertilized	16.5
17	Fertilized	14.8
18	Fertilized	14.4
19	Fertilized	14.6
20	Fertilized	16.7
21	Fertilized	13.8
22		

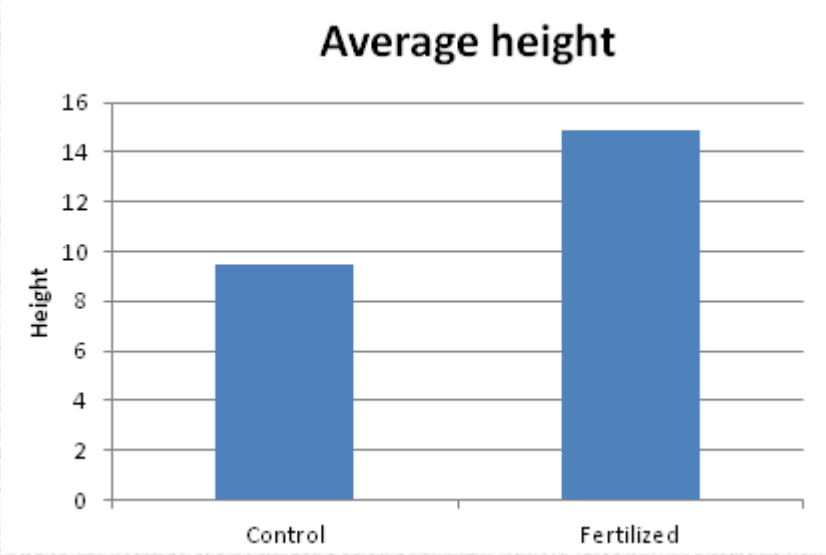
Pivot table for graphing

	D	E
	Average of Height	
	Treatment	Total
	Control	9.442980293
	Fertilized	14.88775842
	Grand Total	12.16536935

One categorical variable (grouping)

One numeric variable (response)

Graph of pivot table data



Bar height can be any numeric variable statistic (usually total or mean)

Grouped column chart

Data in Excel

	A	B	C
1	Treatmen	Plant	Height
2	Control	Corn	41.7
3	Control	Corn	37.4
4	Control	Corn	41.8
5	Control	Corn	44.6
6	Control	Corn	37.2
7	Control	Beans	24.2
8	Control	Beans	18.9
9	Control	Beans	25.5
10	Control	Beans	22.8
11	Control	Beans	26.8
12	Fertilized	Corn	53.1
13	Fertilized	Corn	60.5
14	Fertilized	Corn	71.9
15	Fertilized	Corn	64.9
16	Fertilized	Corn	63.3
17	Fertilized	Beans	32.7
18	Fertilized	Beans	32.3
19	Fertilized	Beans	39.3
20	Fertilized	Beans	34.1
21	Fertilized	Beans	32.9
22			

Pivot table for graphing

Average of Height	Plant		
Treatment	Beans	Corn	Grand Total
Control	23.64382879	40.56847979	32.10615429
Fertilized	34.2673524	62.7506732	48.5090128
Grand Total	28.9555906	51.6595765	40.30758355

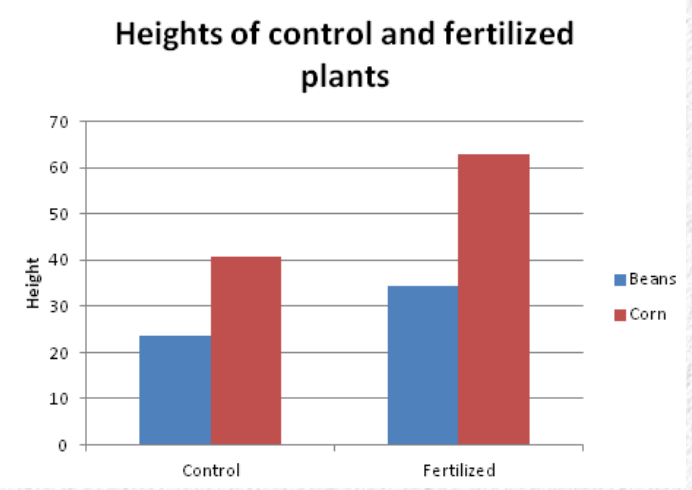
Two categorical variables (treatment group, plant type)

- Bars grouped/labeled on x-axis by first categorical variable

- Color coded by second categorical variable

One numeric variable (response)

Graph of pivot table data



Like simple column chart, bar height usually total or mean

Line charts

Data in Excel

	A	B	C
1	Date	Fertilized	Control
2	1/1/2012	6	7
3	1/4/2012	12	10
4	1/7/2012	14	12
5	1/10/2012	18	13
6	1/13/2012	20	14
7	1/16/2012	24	15
8	1/19/2012	26	18
9	1/22/2012	30	20
10			

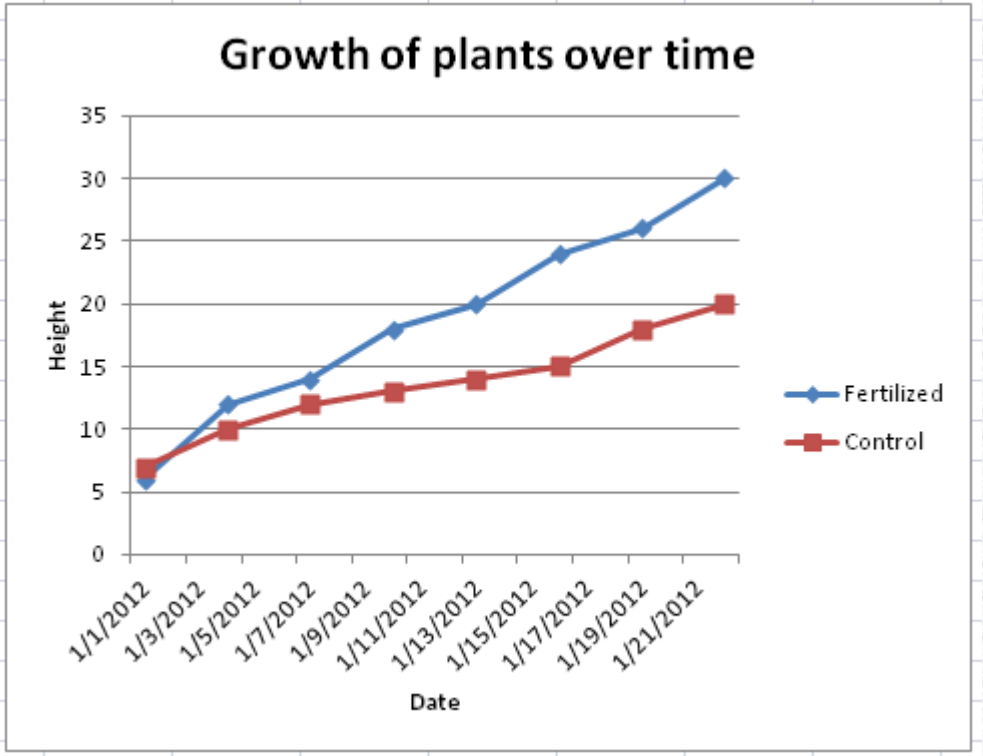
***Categorical** X axis, numeric Y*

Each line is a series (can select either row or column series – columns here)

Each series uses the same x-axis = first column selected

Order of values along the x-axis is the same as the order in the data table

Graph



WARNING: Line charts are a can of worms

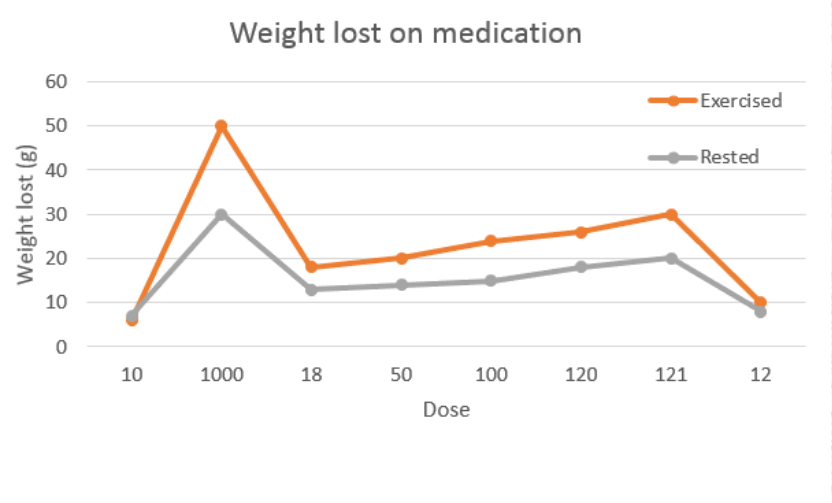
- Line charts plot the data as points and connect them with lines
  - Problem: the most conspicuous part of a line chart is the line, but the lines are not data
  - If the points are not plotted the data are not on the graph

- The x-axis is categorical, even if you use a numeric variable for it – the numbers are used as though they are text labels
  - Problems:
    - The ordering of the numbers along the x-axis can be wrong
    - The relative spacing of the numbers can be wrong
    - Excel lets you add trend lines, but the equations reported will (probably) be wrong

Data in Excel

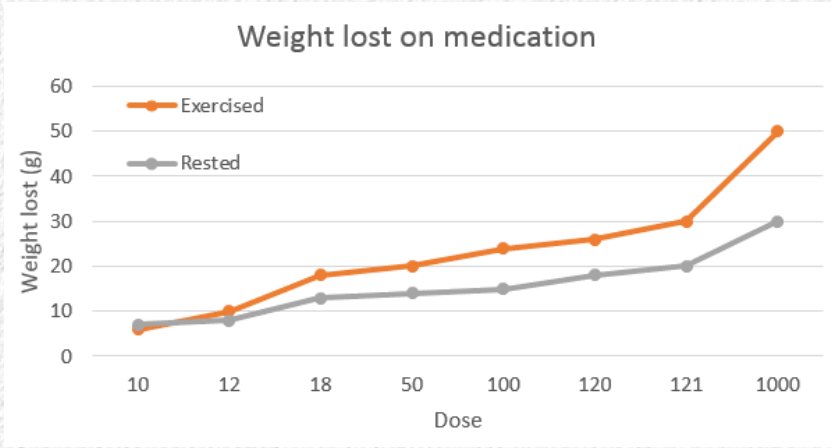
	A	B	C
1	Dose	Exercised	Rested
2	10	6	7
3	1000	50	30
4	18	18	13
5	50	20	14
6	100	24	15
7	120	26	18
8	121	30	20
9	12	10	8

Graph



X categories are not in order in the data sheet, so not in order on the graph

	A	B	C
1	Dose	Exercised	Rested
2	10	6	7
3	12	10	8
4	18	18	13
5	50	20	14
6	100	24	15
7	120	26	18
8	121	30	20
9	1000	50	30

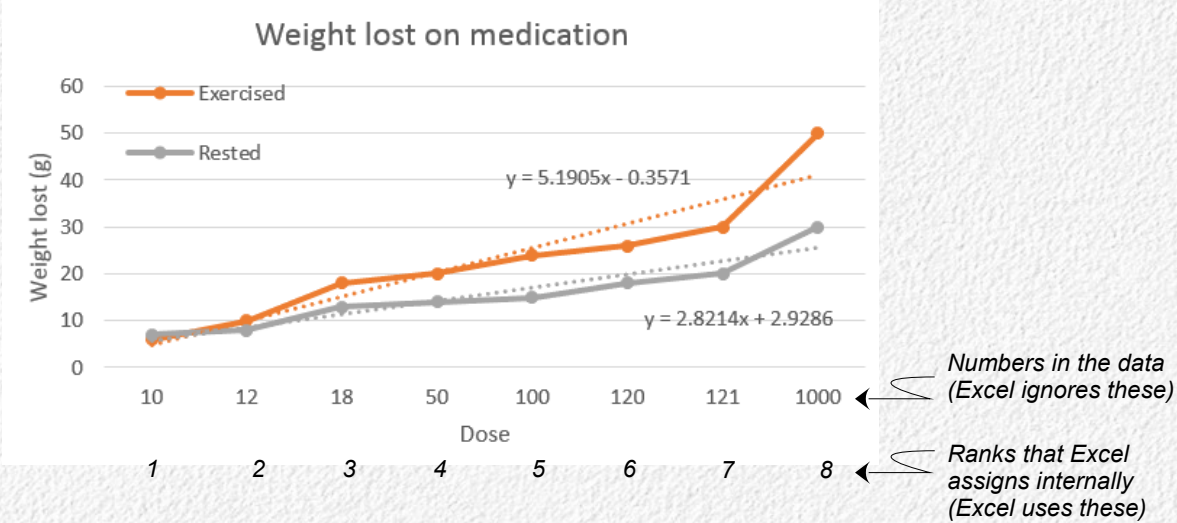


X categories in order now, but spacing between doses doesn't reflect numeric values

Excel allows you to fit a trend line on a line chart – but don't!

Exercise trend line slope and intercept

Based on numeric values		
Slope	0.03529	
Intercept	16.6879	
Based on rank order		
Slope	5.19048	
Intercept	-0.35714	

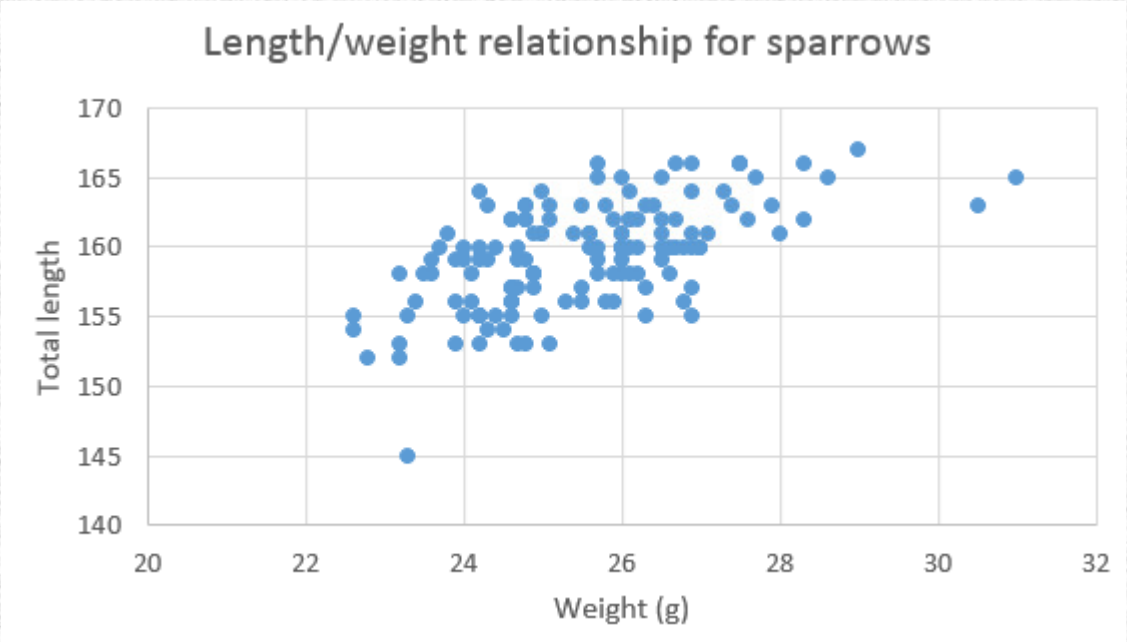


Since x-axis is categorical Excel ignores the x-axis numbers for the trend line

Instead, ranks assigned internally – 1 to the first, increasing by 1 to the last

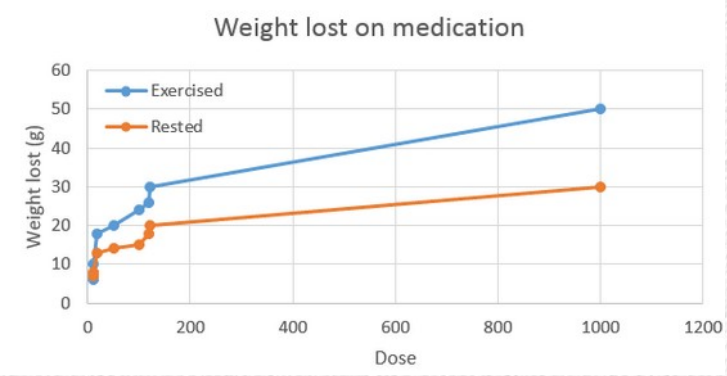
Slope and intercept from trend line are change in weight per unit change in rank, not per unit change in dose – in other words, they're wrong

Scatter plots – x and y both numeric



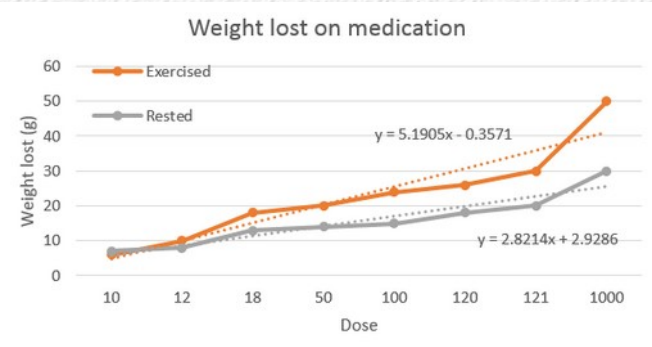
Best choice for displaying relationships between two numeric variables

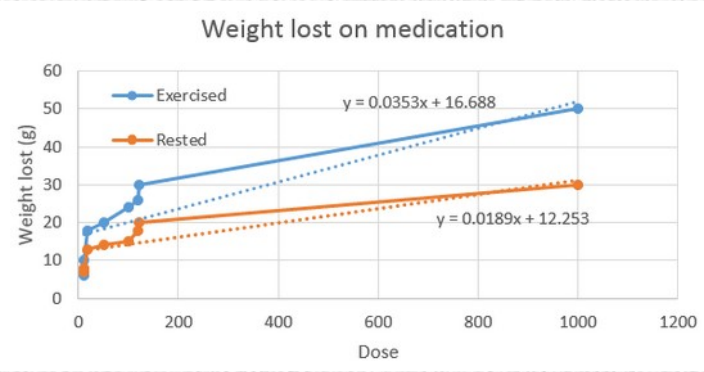
If x is numeric use scatter plot with connecting lines instead of line graph



With lines connecting data points

Same data as line graph





Trend line is based on dose, not rank of dose, so slope and intercept are correct

Based on numeric values	
Slope	0.03529
Intercept	16.6879

Plickers...

Composition data

- Composition data = data that represents values that are part of a whole
  - Yield of crops by type as part of total yield
  - Counts of observations by group as part of total number of observations
- Can be expressed in absolution number (kg, counts)
- Often expressed as proportions or percentages
- Good chart type choices:
  - Pie chart (meh)
  - Stacked column charts
  - Stacked area charts

Pie chart

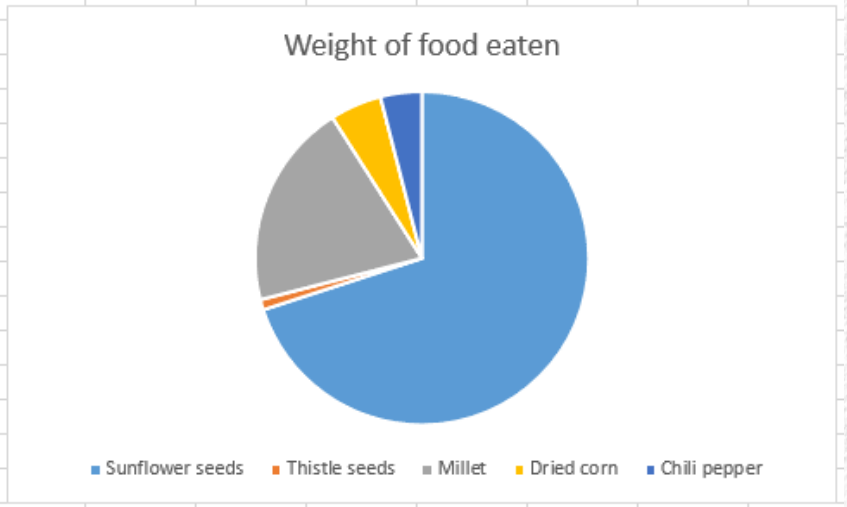
Data in Excel

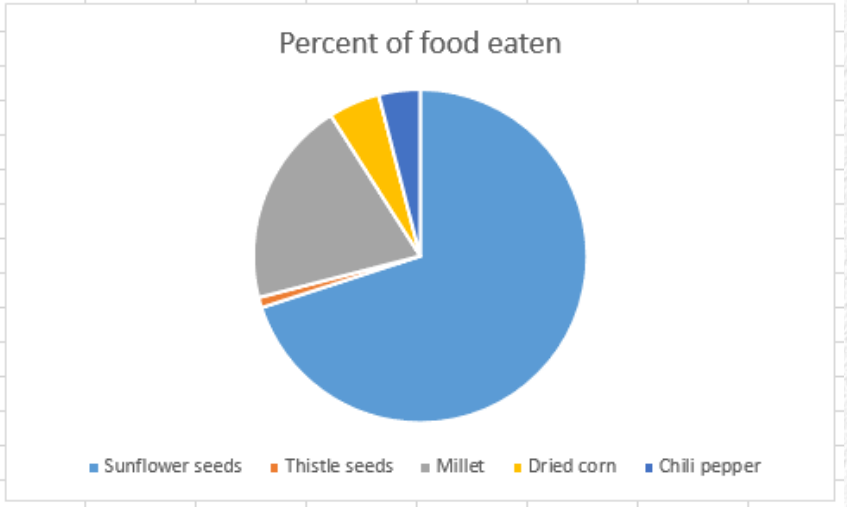
	A	B	C
1	Food	Weight of food eaten	Percent of food eaten
2	Sunflower seeds	105.0	70%
3	Thistle seeds	1.5	1%
4	Millet	30.0	20%
5	Dried corn	7.5	5%
6	Chili pepper	6.0	4%
7			
8	Total	150	100%

Any variable used will be converted to proportions of the total → set the size of the pie pieces

Only use pie charts for composition data!

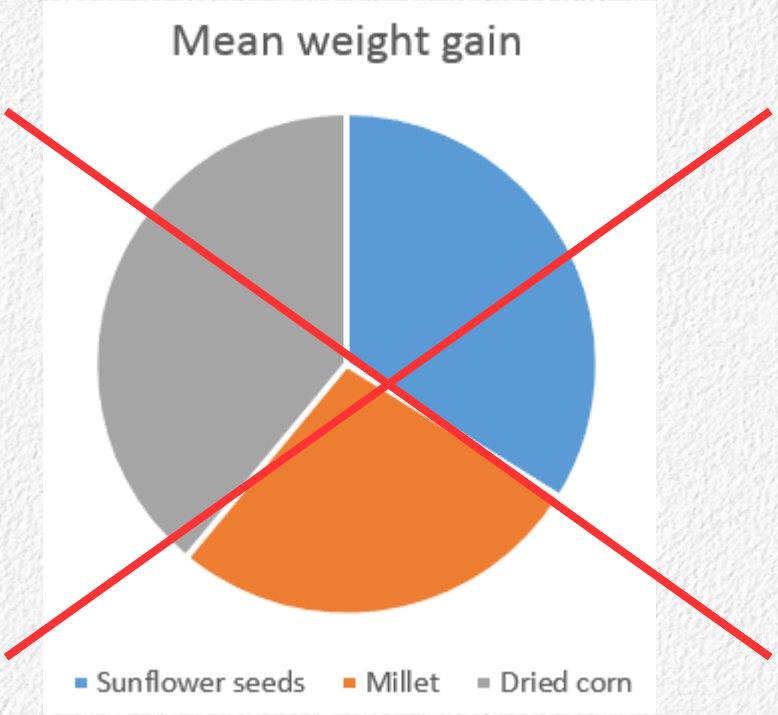
Graph of weight, percents





Pie charts treat any data as composition data

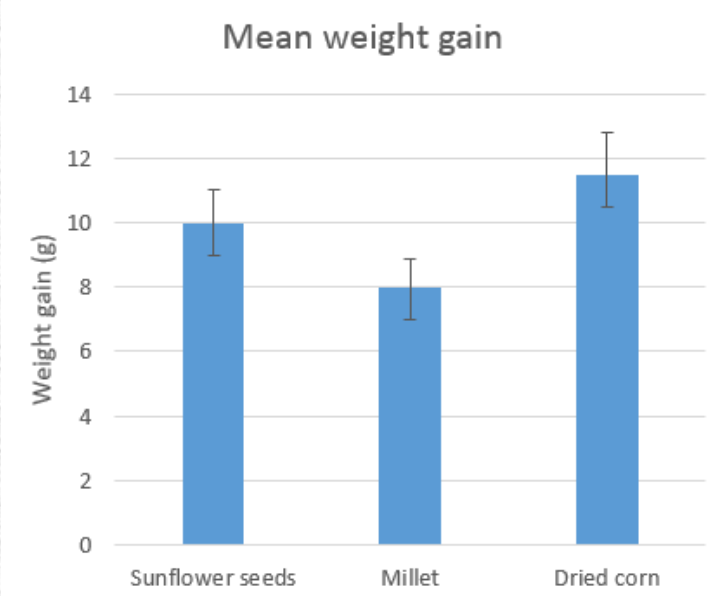
Mean weight gain
10
8
11.5



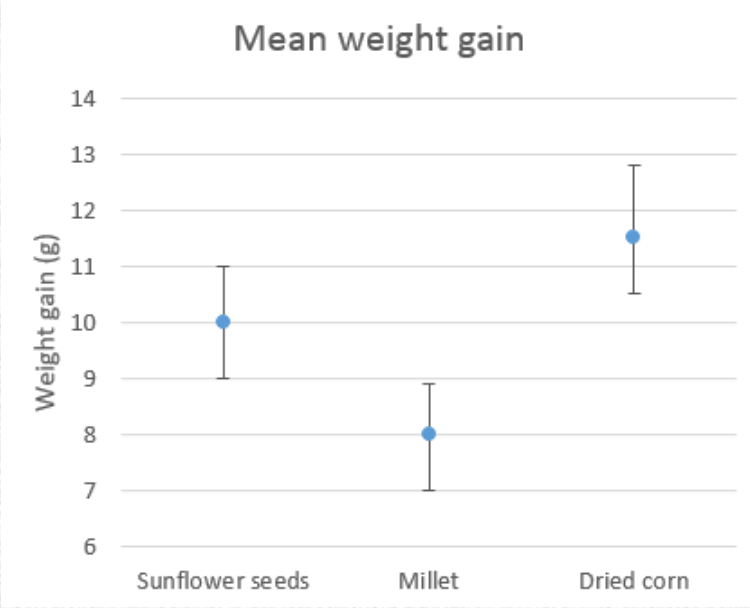
Don't use pie charts to display means – means are not parts of wholes

Better choice?

Better options for means by group



Bar chart with error bars



Mean symbols with error bars

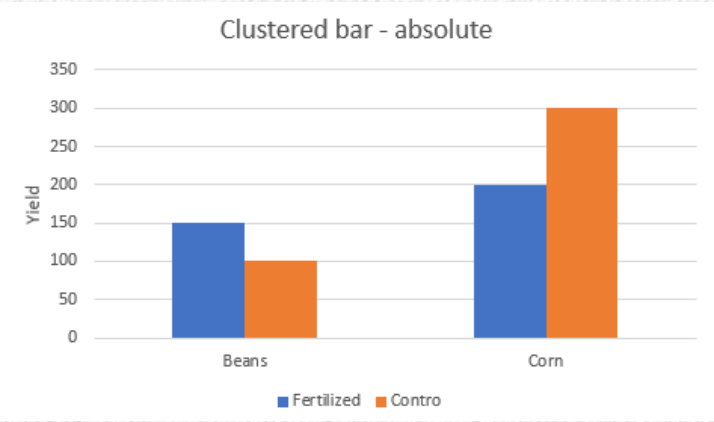
(in Excel done using line chart without lines)

Stacked column charts – absolute or relative

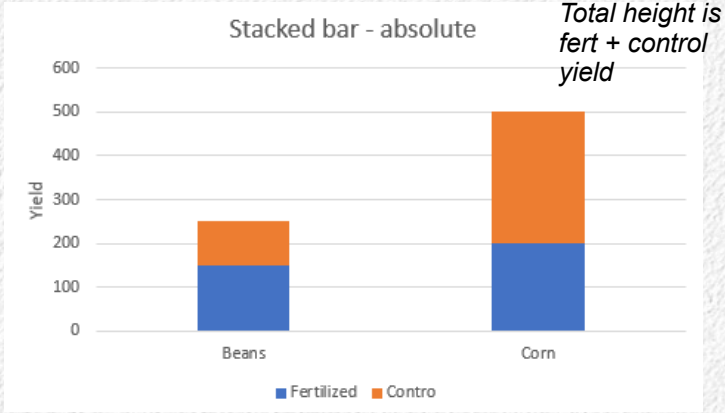
Total yield of beans and corn

Crop	Beans	Corn
Fertilized	150	200
Contro	100	300

Height of bar is yield on one of two treatments

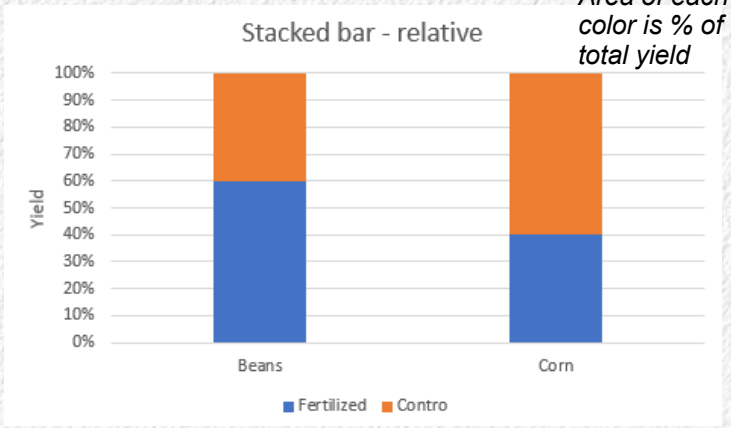


Stack 'em



Total height is fert + control yield

Alternative to two different pie charts



Area of each color is % of total yield

Stacked column charts can be used instead of pie charts

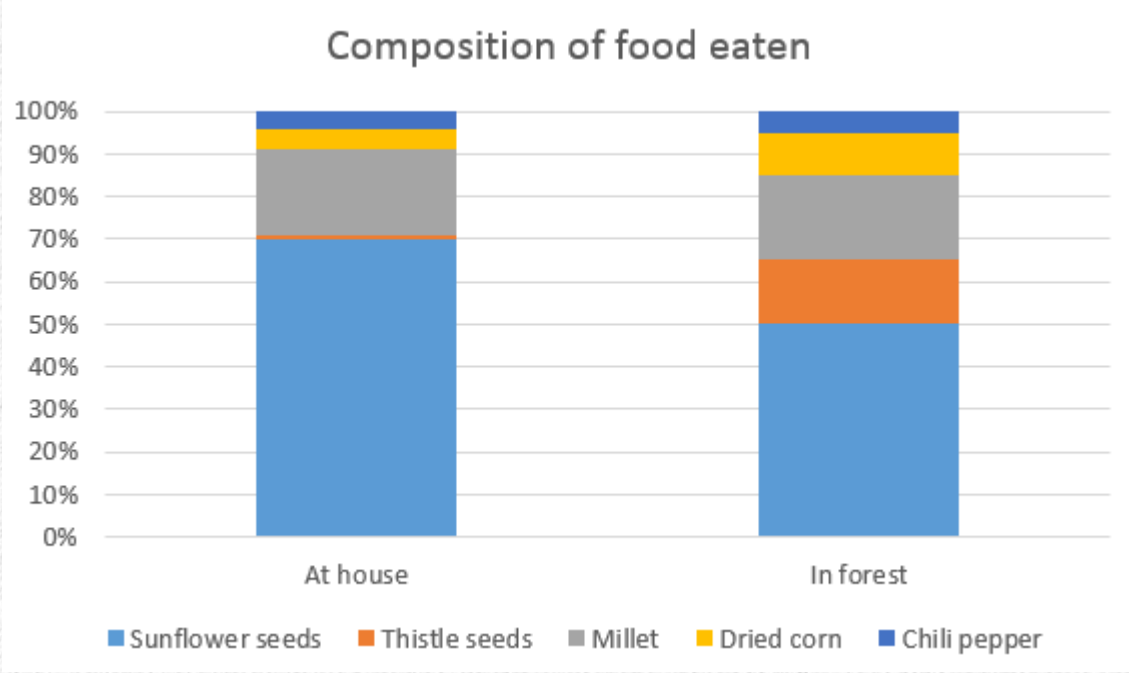
***Like a pie chart**:*

- Relative amount of each food type can still be seen

***Better than a pie chart**:*

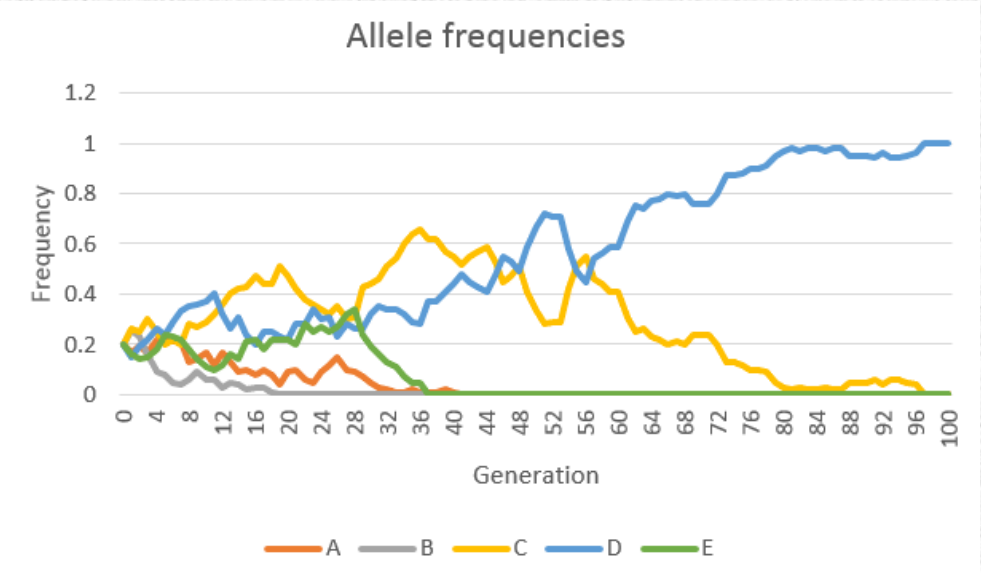
- Comparisons between groups are easier to make

- There is a numeric scale (y) that makes the amounts easier to judge

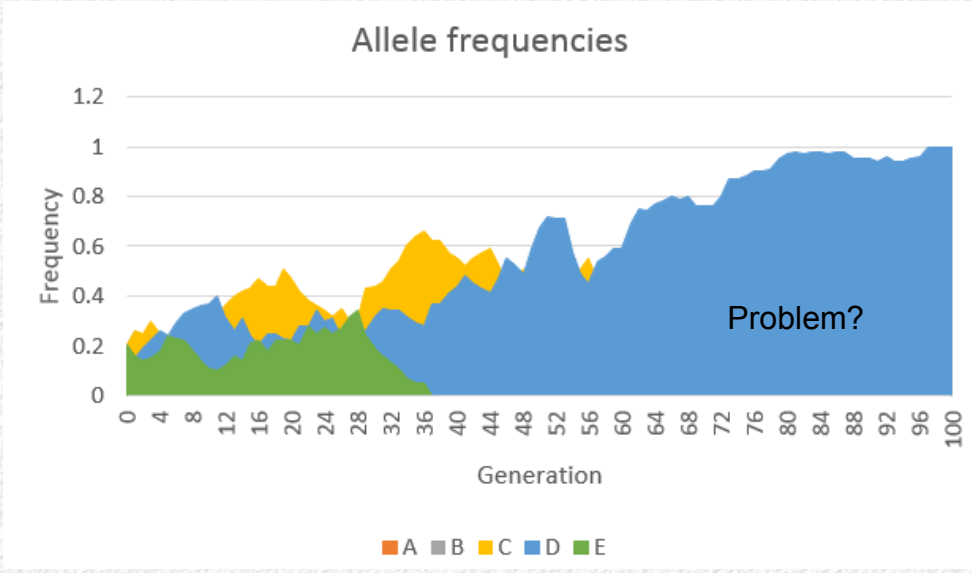


Stacked bar, percent scale

Area plots – space below lines filled

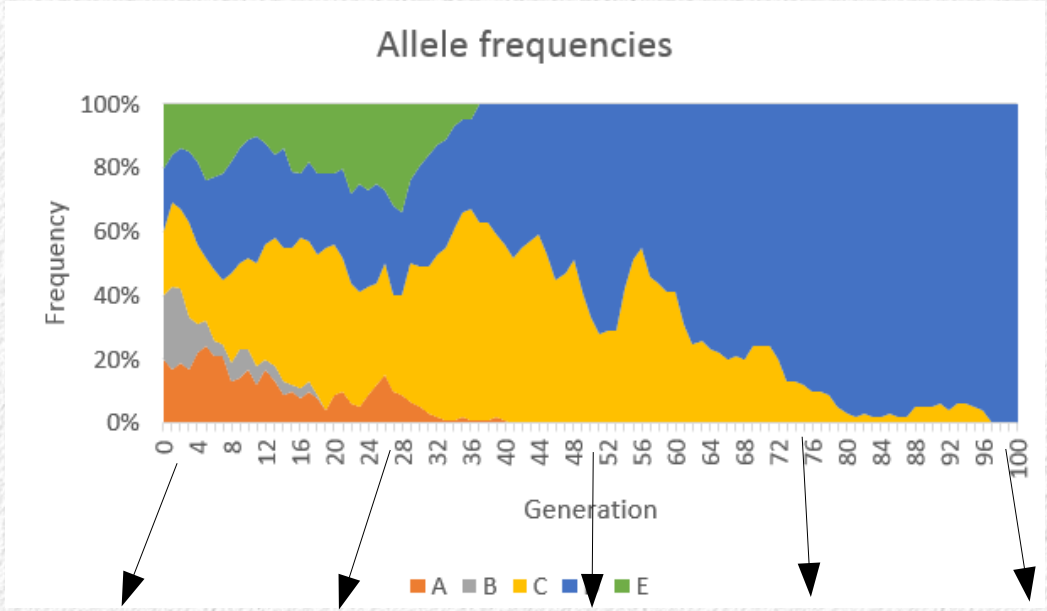


Line chart of
allele frequencies



Area chart of
allele frequencies

Composition data across many categories: stacked area plot



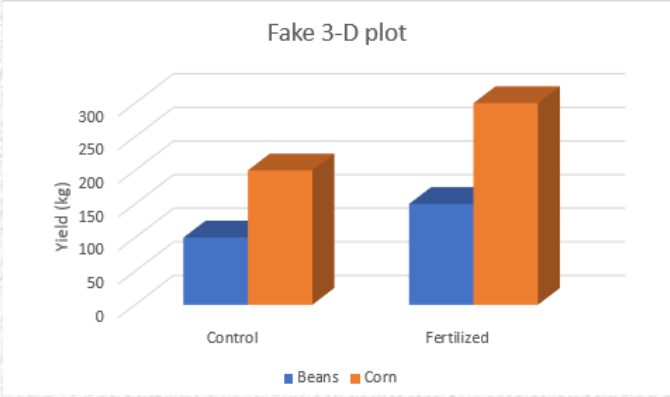
Alternative to stacked bar chart with many different bars

Note that x-axis is still categorical

Takes the place of 101 pie charts

High order data sets

- Some data sets have many different variables
- Flat screens only have 2 dimensions – two variables easy to display
- Adding variables means adding dimensions we don't have
  - Plot “slices” through the data
  - 3D graphs use depth cueing (perspective tricks)
  - Use symbols/lines



This nonsense is not a true 3D graph – not suitable for scientific work, avoid

Plotting slices across a third variable on a 2D graph

- Can group data based on levels of a third variable
- You can see the effect of the third (grouping) variable by comparing the groups
- Example: plotting the length, width, and area of rectangles as a series of lines

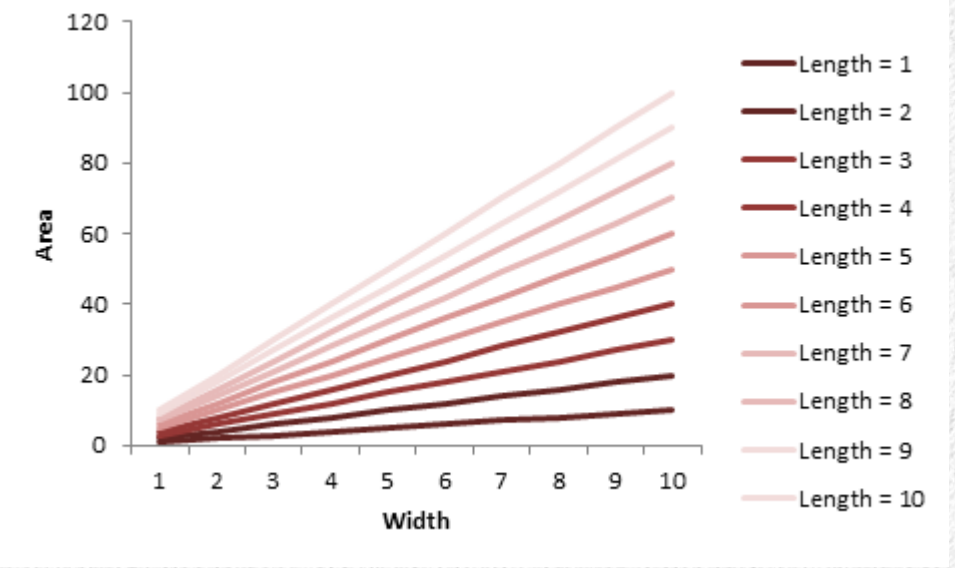
Slices through the data – lines are length columns

Data in Excel

	A	B	C	D	E	F	G	H	I	J	K	L
1			Length									
2	Width		1	2	3	4	5	6	7	8	9	10
3		1	1	2	3	4	5	6	7	8	9	10
4		2	2	4	6	8	10	12	14	16	18	20
5		3	3	6	9	12	15	18	21	24	27	30
6		4	4	8	12	16	20	24	28	32	36	40
7		5	5	10	15	20	25	30	35	40	45	50
8		6	6	12	18	24	30	36	42	48	54	60
9		7	7	14	21	28	35	42	49	56	63	70
10		8	8	16	24	32	40	48	56	64	72	80
11		9	9	18	27	36	45	54	63	72	81	90
12		10	10	20	30	40	50	60	70	80	90	100
13												

Length, width, and area of rectangles

Each length is a series, width on x-axis



Line colors selected to indicate lengths

Surface plot – plotting 3D on a 2D screen

Data in Excel

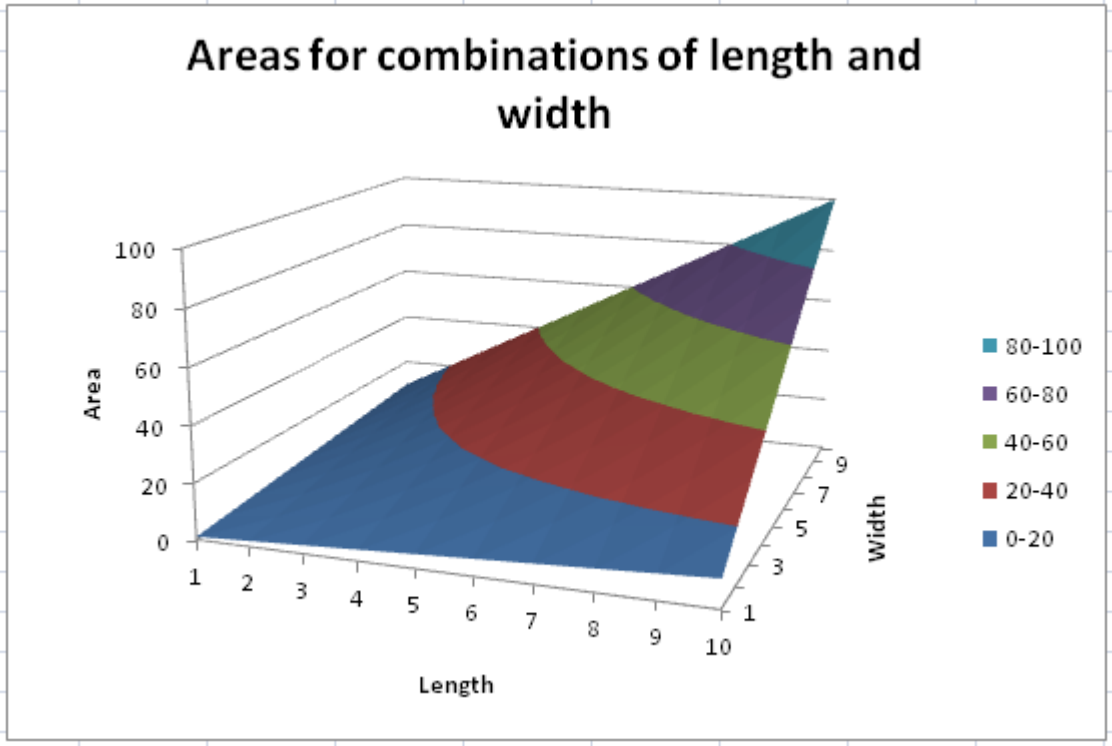
	A	B	C	D	E	F	G	H	I	J	K	L
1			Length									
2	Width		1	2	3	4	5	6	7	8	9	10
3		1	1	2	3	4	5	6	7	8	9	10
4		2	2	4	6	8	10	12	14	16	18	20
5		3	3	6	9	12	15	18	21	24	27	30
6		4	4	8	12	16	20	24	28	32	36	40
7		5	5	10	15	20	25	30	35	40	45	50
8		6	6	12	18	24	30	36	42	48	54	60
9		7	7	14	21	28	35	42	49	56	63	70
10		8	8	16	24	32	40	48	56	64	72	80
11		9	9	18	27	36	45	54	63	72	81	90
12		10	10	20	30	40	50	60	70	80	90	100
13												

Three dimensions, instead of 10 series

X and Y are row and column labels, treated as categorical

Z is numeric, in body of matrix

Graph of values in body of matrix



Symbol properties

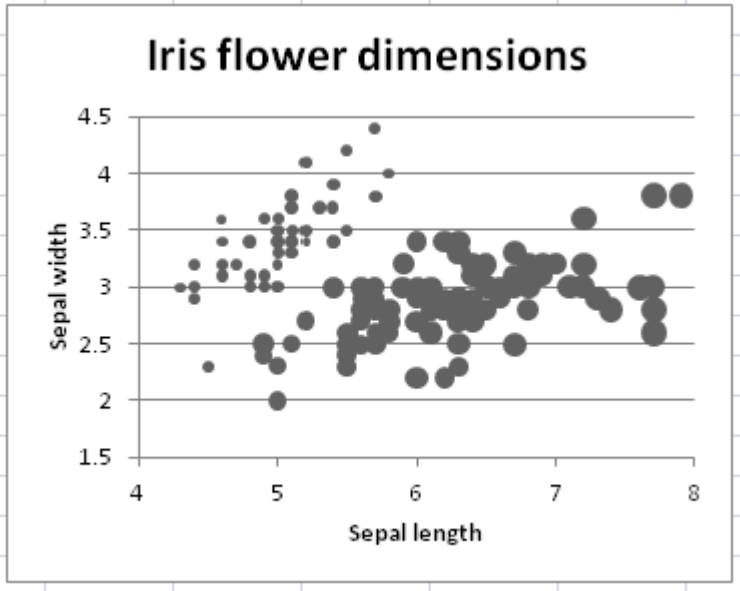
- Can subset the data and display with:
  - Symbol size
  - Symbol color
  - Symbol type

Data in Excel

	A	B	C
1	Sepal.Length	Sepal.Width	Petal.Length
2	5.1	3.5	1.4
3	4.9	3	1.4
4	4.7	3.2	1.3
5	4.6	3.1	1.5
6	5	3.6	1.4
7	5.4	3.9	1.7
8	4.6	3.4	1.4
9	5	3.4	1.5
10	4.4	2.9	1.4
11	4.9	3.1	1.5
12	5.4	3.7	1.5
13	4.8	3.4	1.6
14	4.8	3	1.4
15	4.3	3	1.1
16	5.8	4	1.2
17	5.7	4.4	1.5
18	5.4	3.9	1.3
19	5.1	3.5	1.4
20	5.7	3.8	1.7
21	5.1	3.8	1.5
22	5.4	3.4	1.7
23	5.1	3.7	1.5
24	4.6	3.6	1

Bubble chart

Chart – symbol size is proportional to petal length



Radar charts – multiple numeric axes

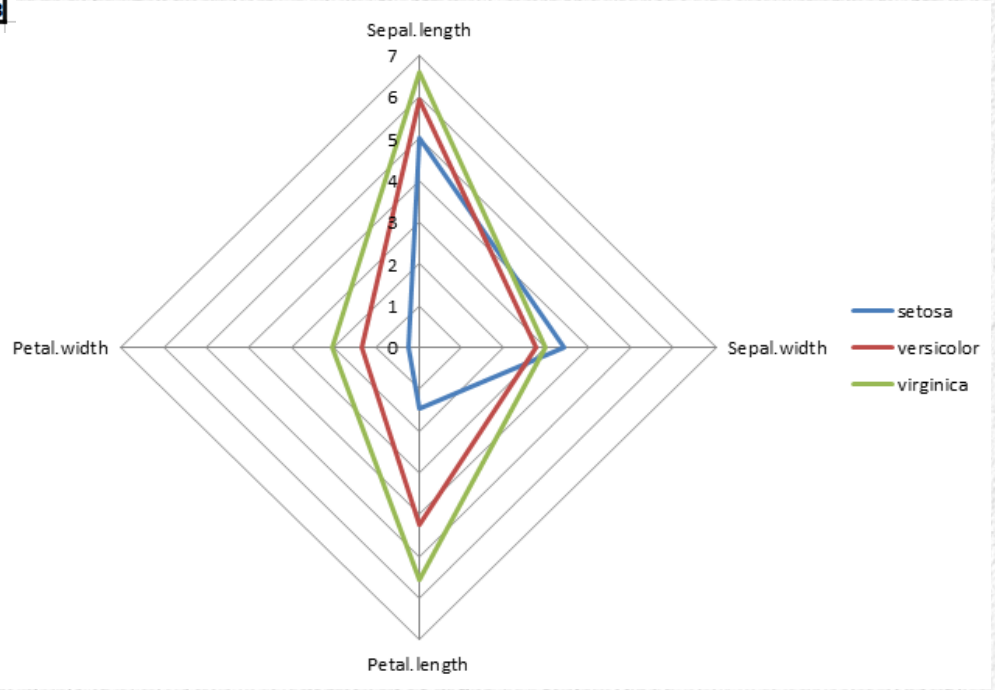
- Each ray is a different variable
- Each data point is plotted on each ray, with lines connecting

Radar plot with four axes

Data in Excel

	Data			
Species	Average - Sepal.Length	Average - Sepal.Width	Average - Petal.Length	Average - Petal.Width
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.77	4.26	1.326
virginica	6.588	2.974	5.552	2.026
Total Result	**5.84333333333333**	**3.05733333333333**	**3.758**	**1.19933333333333**

Chart



Histograms

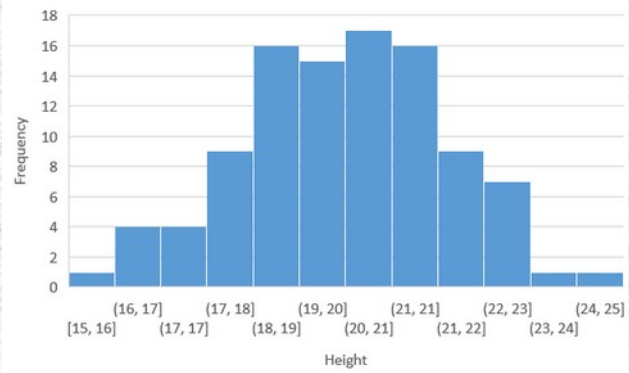
Data in Excel

	A	B
1	Height	
2	5.32	
3	4.53	
4	4.41	
5	5.76	
6	5.42	
7	4.68	
8	5.72	
9	5.23	
10	7.67	
11	4.89	
12	3.81	
13	5.81	
14	5.43	
15	5.54	
16	6.72	
17	5.49	
18	6.7	
19	4.95	
20	3.76	
21	6.84	
22	6.71	
23	5.17	

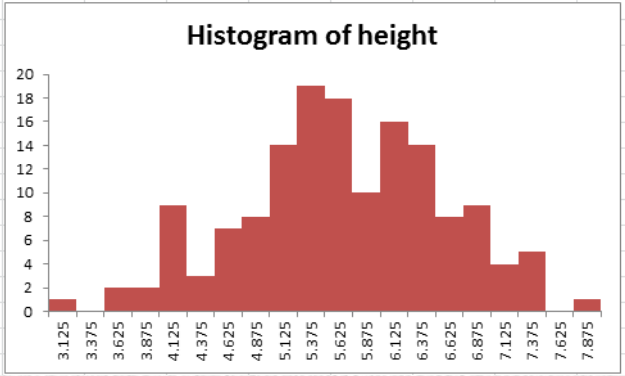
Bins, midpoints, and frequencies

Bins	Midpoints	Frequency
3	3.125	1
3.25	3.375	0
3.5	3.625	2
3.75	3.875	2
4	4.125	9
4.25	4.375	3
4.5	4.625	7
4.75	4.875	8
5	5.125	14
5.25	5.375	19
5.5	5.625	18
5.75	5.875	10
6	6.125	16
6.25	6.375	14
6.5	6.625	8
6.75	6.875	9
7	7.125	4
7.25	7.375	5
7.5	7.625	0
7.75	7.875	1

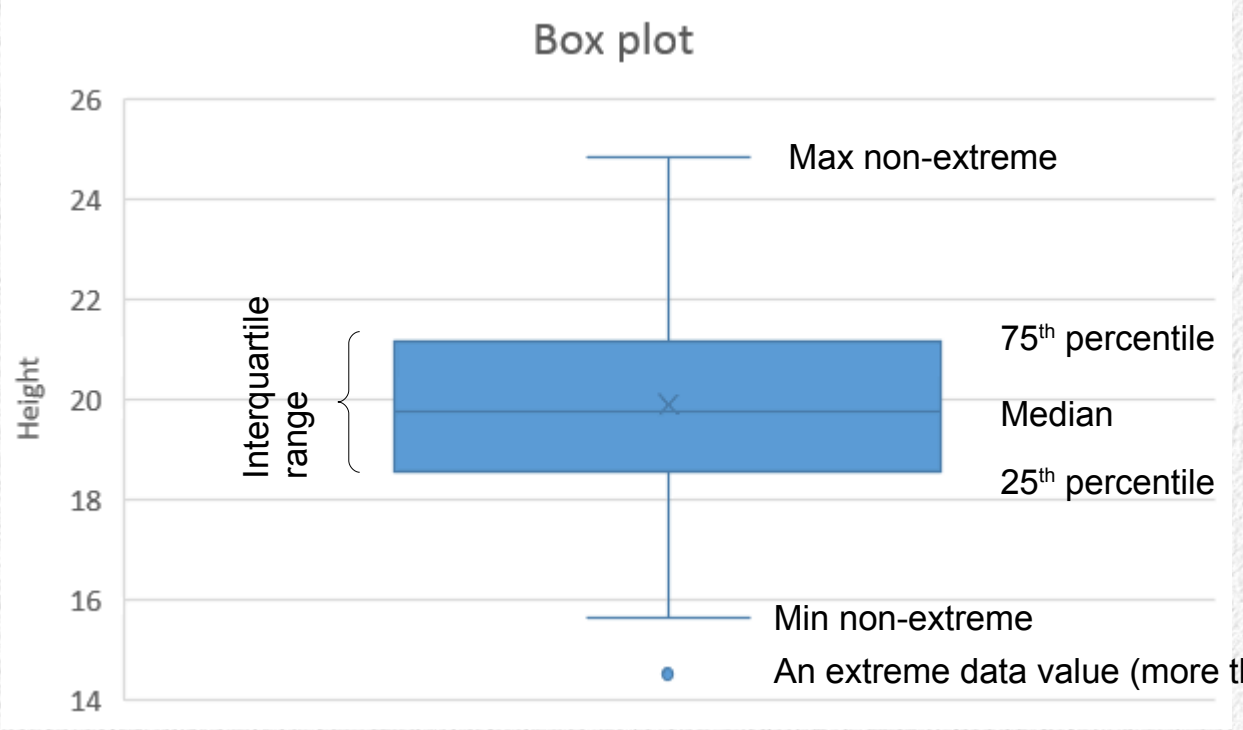
Built-in histogram chart type



Bar chart of freq's, no gap between bars



Box plots


Box plot
Height
26
24
22
20
18
16
14
Interquartile range
Max non-extreme
75th percentile
Median
25th percentile
Min non-extreme
An extreme data value (more than 1.5 IQR from bottom)

Specialized graph types Excel does not support

- Statistical graphs
  - Residual plots
  - Biplots
- Heat maps
- Various 3D visualizations (3D scatter plots)
- For visualizations like these, need to use another package