Charts (part 1)

Charts

Introduction

Excel can create a variety of charts however, for Health data only a subset of them are really useful and some important chart types are not available.

Useful Charts	Column, Bar, XY scatter and (Pie)
Less useful (unlikely to use)	Line (commonly mistaken for XY scatter with line), Area, Surface, Radar, Doughnut, Bubble, Stock, Cylinder, Cone and Pyramid

A chart is a graphic representation of worksheet data. When a chart is created based on a worksheet selection, Excel uses the values from the worksheet and presents them in the chart as data points, which are represented by bars, lines, columns, slices, dots and other shapes. These shapes are referred to as data markers.

Groups of data points, or data markers, originating from single worksheet rows or columns are grouped in data series. Each data series is then distinguished by a unique colour or pattern or both.

After creation, a chart can be enhanced to emphasise certain information by adding chart items, such as data labels, a legend (key), titles, text, trendlines, error bars and gridlines. Most chart items can be moved, sized and formatted.

When analysing data, the importance of plotting your data cannot be over-emphasised. Unless you are simply interested in frequency tables or crosstabulations, you should plot your data. Doing so will allow you to:

- study the distribution of your data (does it follow a normal distribution, for example?)
- check for any outliers or extreme values that need to be further investigated
- examine the relationships between two or more variables

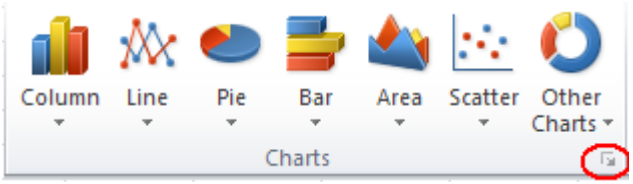
The computation of any statistics, whether summary statistics such as means or medians or performing statistical tests, such as t-tests or correlation coefficients, should only proceed after plotting the data.

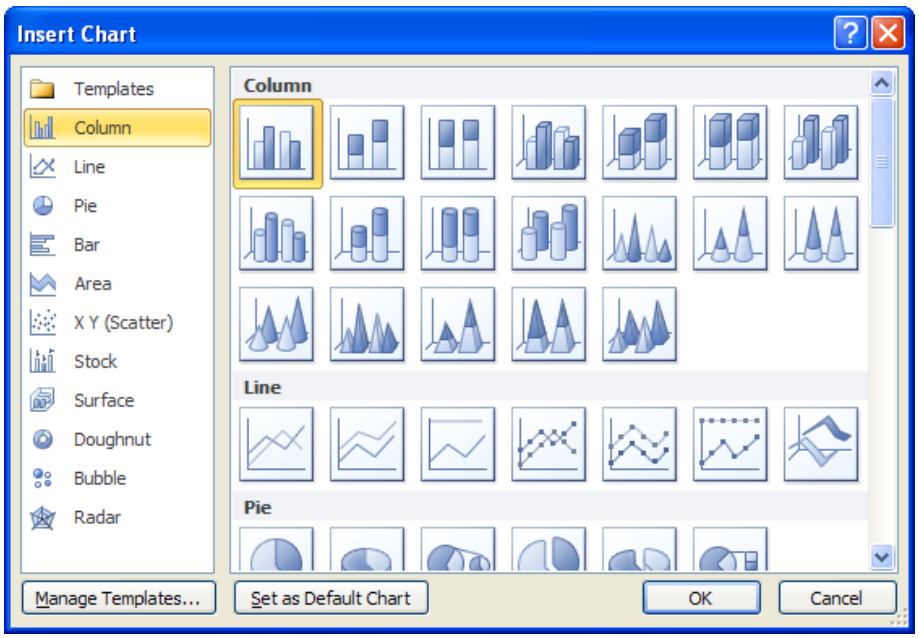
Creating a chart

The key to producing the type of chart you want is to organise your data correctly. Arrange the data in columns. You might wish to copy the data that you want to plot onto a new worksheet. Ensure that the labels for categories in a bar chart, or the X (horizontal) axis values for a scatter plot are in the first column. Use the first row in the worksheet to give the data a label.

	A	B
1	**Profession**	**Number**
2	Doctor	9
3	Nurse	13
4	Pharmacist	2
5	Physiotherapist	5
6	Radiographer	3

On the **Insert** tab you can see lots of icons in the ***Charts*** group. These are for creating different types of charts. To see all of the charts that are available click  to start the ***Insert Chart*** dialog box.

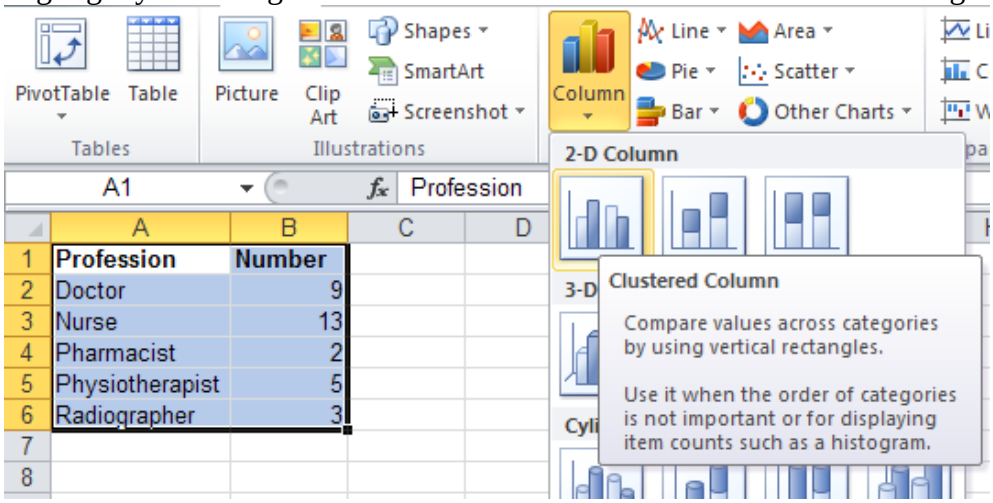




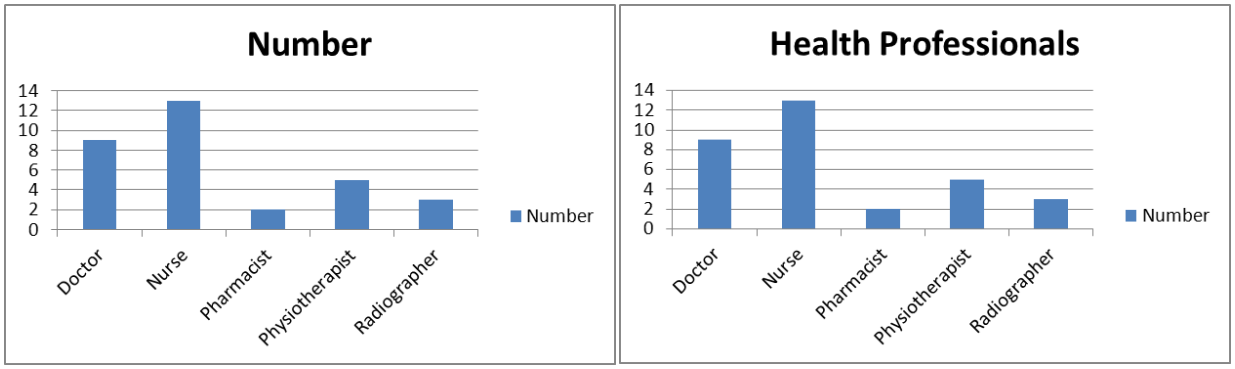
To get started, let's look at bar charts. In Excel, what we usually call a bar chart (a chart using parallel vertical bars of varying lengths, to illustrate comparative heights, weights, birth-rates, etc.) is called a column chart. What Excel calls a bar chart shows the same information with horizontal lines. This is clear from the icons that appear in the Charts group on the Insert tab.

In this course when I say bar chart I am referring to Excel's column chart.

Highlight your arranged data and click on the ***Column*** icon in the ***Charts*** group.

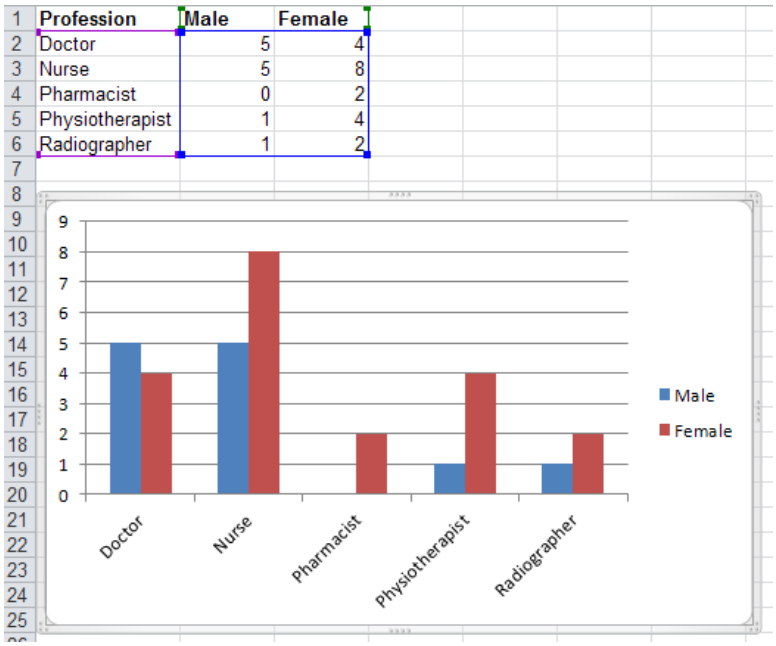


Select ***2-D Clustered Column***. A bar chart appears. By clicking twice (not double clicking) on the title we can edit it. Later we'll look at more ways to edit charts.

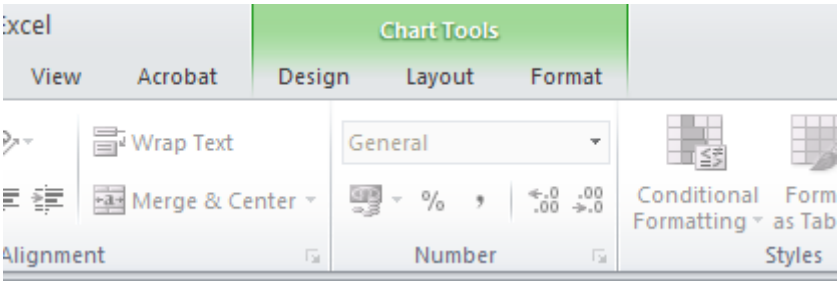


Producing a Clustered Bar Chart

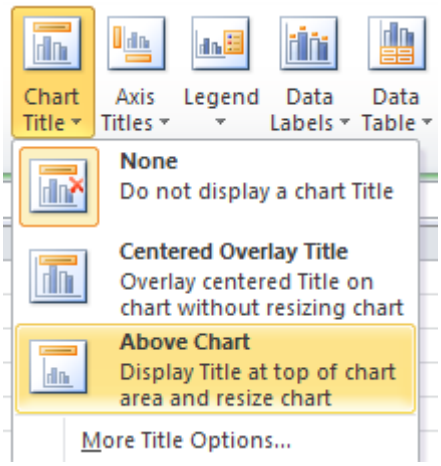
A clustered bar can be used to display categories grouped by an attribute such as gender. The chart opposite was produced in the same way as the previous bar chart. One of the cells in the range A1 to C6 was selected and the 2D clustered Column Chart was selected.

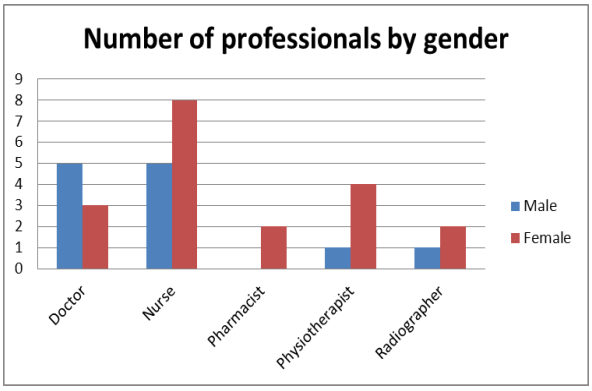


To add a title to this chart click anywhere on the chart to select it and the Chart Tools group will appear.



Click ***Layout*** then ***Chart Title*** and then ***Above Chart***. Then type in the title of the chart where it says Chart Title.



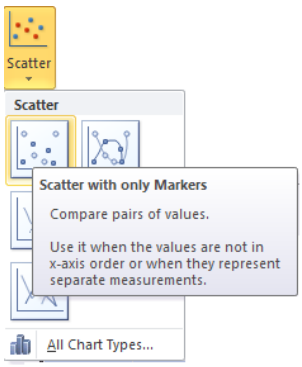


Producing a Scatterplot

You could use a scatter plot when plotting two continuous variables against each other. An example of this could be height vs. weight.

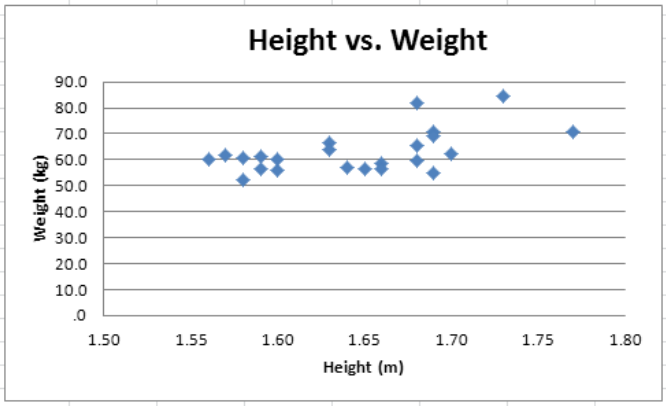
To arrange data on a worksheet for a scatter plot, you should place the x values in one row or column, and then enter the corresponding y values to the right of this in the adjacent rows or columns.

Highlight the data (or select a cell in the data and Excel will use all of the adjacent data to this cell) and click on the ***Scatter*** icon in the ***Charts*** group and select ***Scatter with only Markers***.



Click on ***Layout 1*** in the ***Chart Layouts*** group and change the axis labels and the title and delete the legend.

	A	B	C	D	E	F	G	H	I	J
1	Height (m)	Weight (kg)								
2	1.68	81.7								
3	1.73	84.6								
4	1.63	66.5								
5	1.60	56.0								
6	1.63	64.1								
7	1.66	58.6								
8	1.68	65.5								
9	1.57	62.0								
10	1.69	71.0								
11	1.59	61.5								
12	1.77	70.8								
13	1.66	56.5								
14	1.68	59.8								
15	1.58	60.5								
16	1.69	54.9								
17	1.70	62.4								
18	1.64	56.8								
19	1.59	56.7								
20	1.56	60.4								
21	1.60	60.3								
22	1.69	69.0								
23	1.58	52.2								
24	1.65	56.7								



Identify data points (XY scatter plot)

Move the cursor to a data point on the scatter plot and Excel will identify the data values.

Note *To select data that is not adjacent, use the <Ctrl> key to select non-consecutive columns or rows. Highlight the first column then hold down the <Ctrl> key whilst highlighting further column(s). Do not leave any empty rows or columns in the area of data to be plotted. Include cells containing labels for rows or columns.*

Line Charts

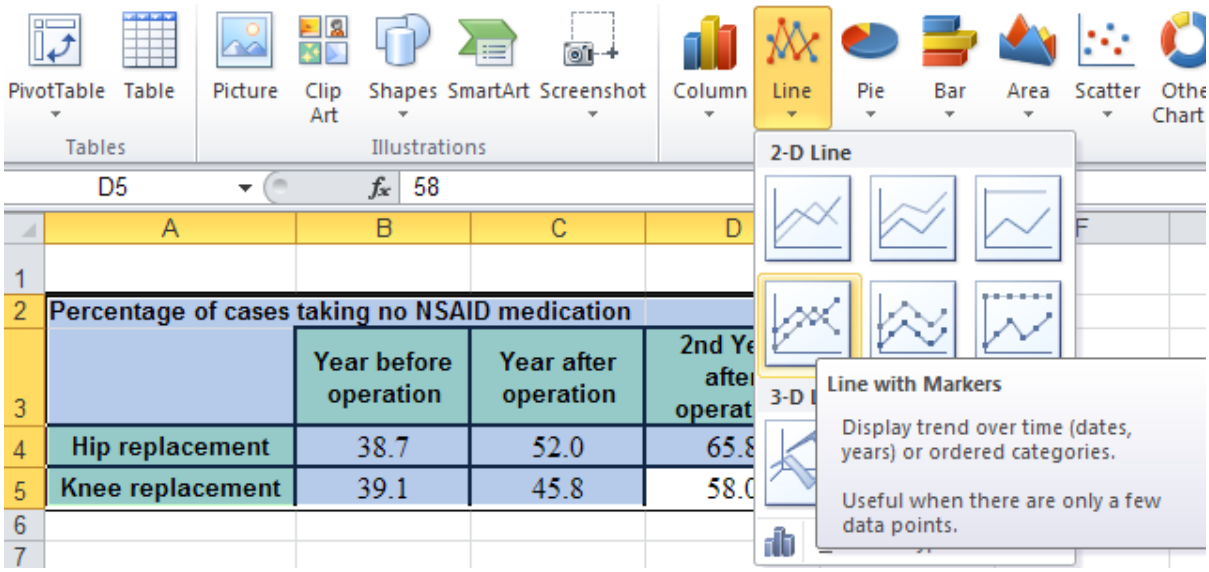
Line charts can be useful in graphing data that is continuous over time and is set against a common scale. It is useful for showing trends of data over time. An example might be the percentage of cases having knee and hip replacements who are taking no non-steroidal anti-inflammatory drugs in the year before and two years after their operations

The data could be stored as follows:

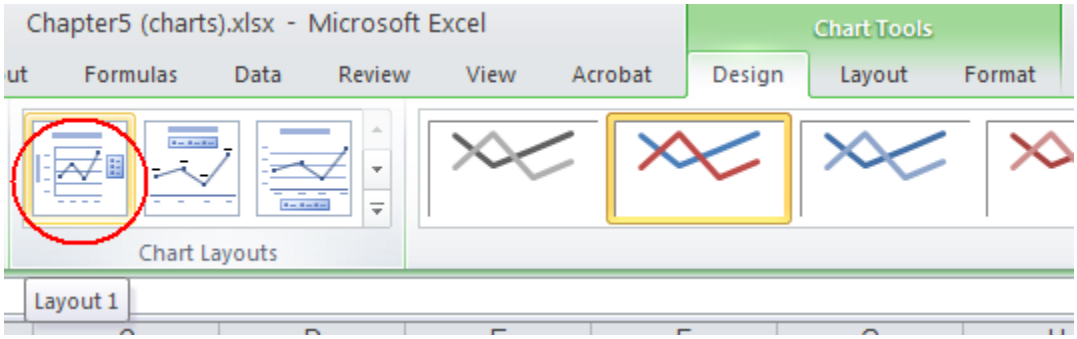
Percentage of cases taking no NSAID medication

	Year before operation	Year after operation	2nd Year after operation
Hip replacement	38.7	52.0	65.8
Knee replacement	39.1	45.8	58.0

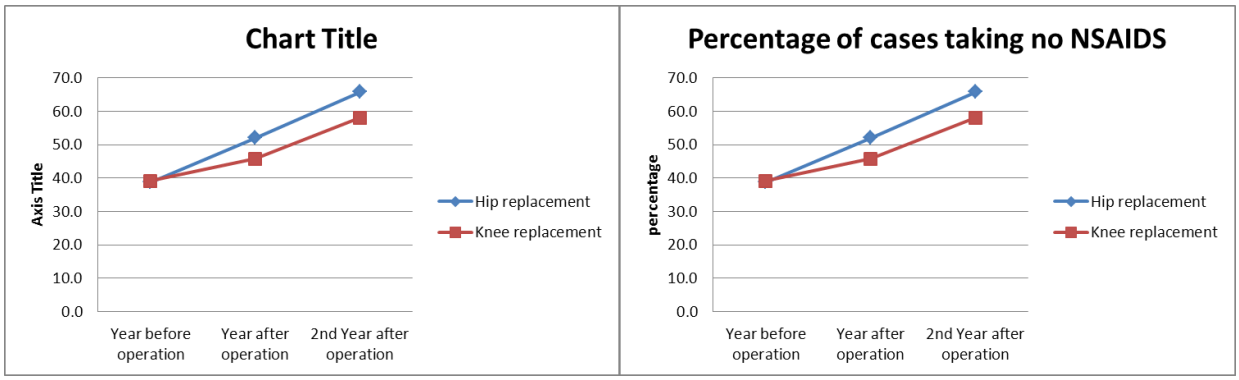
Highlighting this information and then clicking on the ***Line*** chart icon in the ***Charts*** group of the **Insert** tab and selecting ***2-D Line with Markers*** (shown below) will produce a line chart.



An easy way to add a title and a label for the y axis is to click on ***Chart Layout 1*** in the **Design** tab of the ***Chart Tools*** group. The ***Chart Tools*** group appears when you have a chart selected.



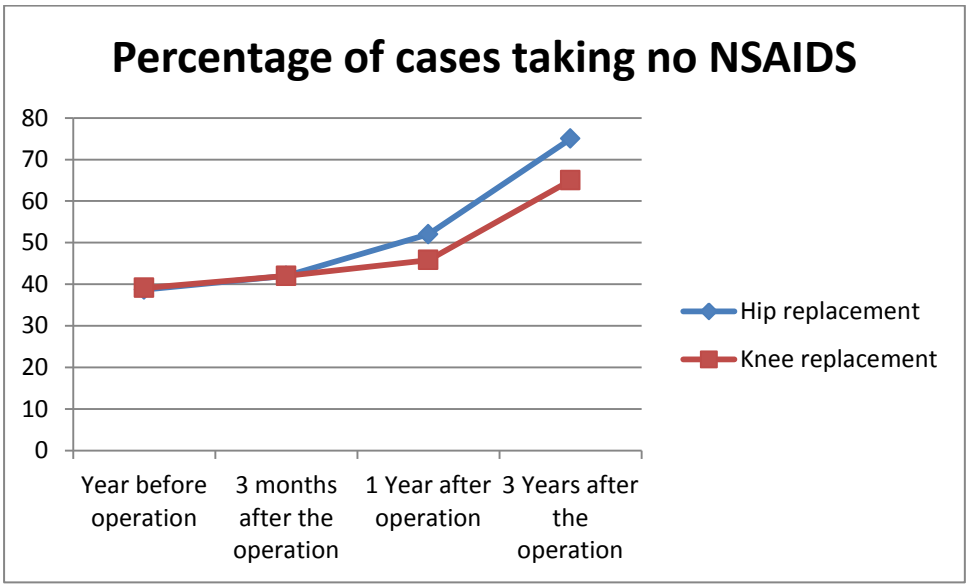
Then you can edit the title and axis label by clicking on them and typing in the text boxes.



Caution : If you use line charts then you must ensure that the time intervals on the horizontal axis have the same gap between them. Let's look at a problem that can occur if you aren't careful with line charts. Imagine your data is like this.

	Year before operation	3 months after the operation	1 Year after operation	3 Years after the operation
Hip replacement	38.7	42	52	75
Knee replacement	39.1	42	45.8	65

If we make a line chart like the previous one then we'll end up with something like this.

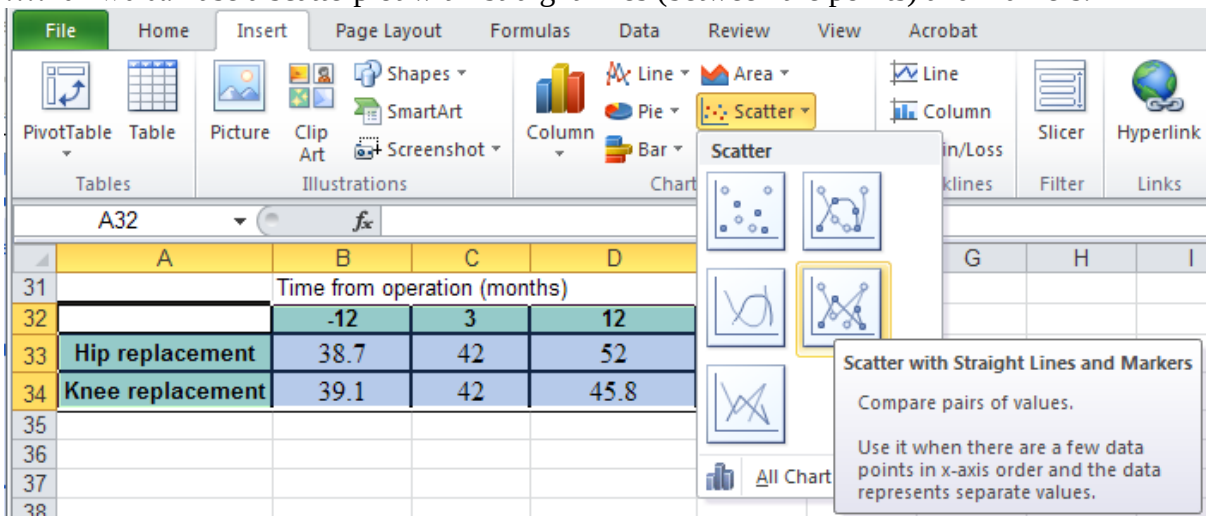


This looks fine at first glance but the distances between categories on the x-axis have no relation to the distance in time between these categories and this gives a distorted view of the data. I have heard reports of charts like this appearing in publications. You must be careful not to make this mistake. This mistake is corrected by using a scatter plot with lines between the points instead of a line chart.

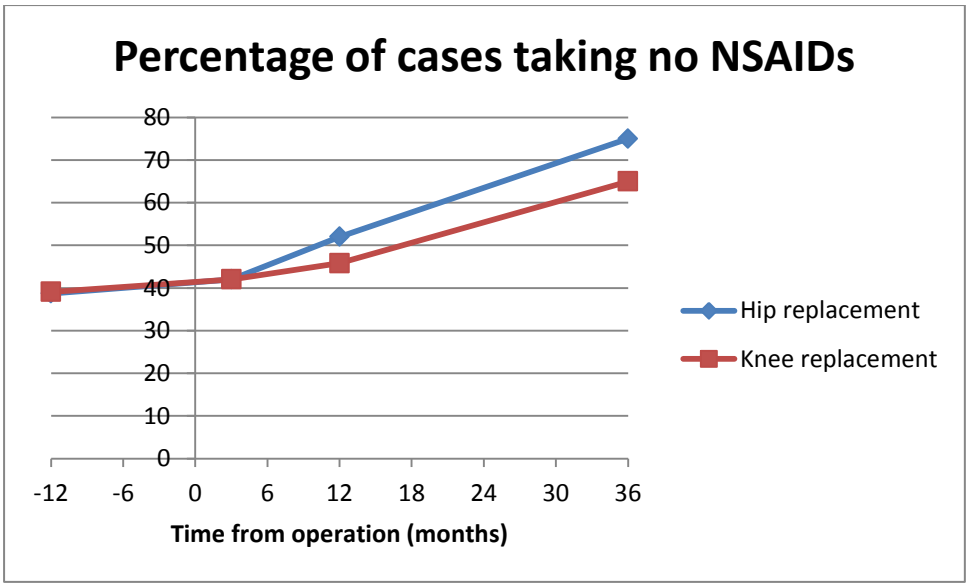
If we organise our data in a similar way to this…

	Time from operation (months)			
	-12	**3**	**12**	**36**
Hip replacement	38.7	42	52	75
Knee replacement	39.1	42	45.8	65

…then we can use a scatterplot with straight lines (between the points) and markers.



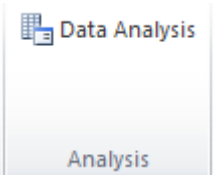
To produce a chart like this (after a little editing).



Now the distance between the points on the chart are representative of the time between the measurements.

Histograms

Histograms can be produced in Excel but it takes a little bit of extra work. To produce a histogram you need to have the ***Data Analysis*** button available in the ***Analysis*** group on the right of the **Data** ribbon.



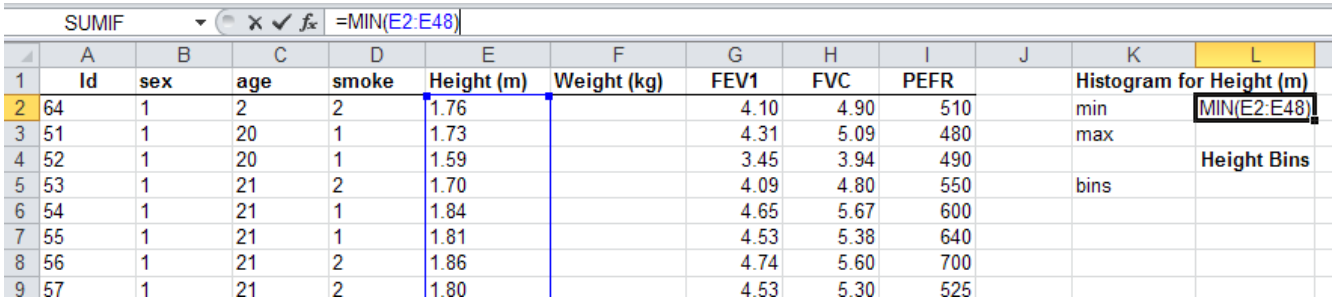
If this isn't there then click on the **File** tab and choose ***Options***. Then choose ***Add-Ins*** from the menu on the left. Down the bottom, where it says ***Manage:*** make sure the drop down menu has ***Excel Add-ins*** selected and then click ***Go…***Tick the box beside ***Analysis ToolPak*** and then click ***OK***.

Let's say that we want to do a histogram of the height data that we have just used in the scatterplot. Here's what to do.

First, because we have to define the widths of the rectangles in the histogram (these are called 'bins') it is helpful to know what the minimum and maximum values in our data are. When I make a histogram in Excel I type out the following information on the right to help me.

Histogram for Height (m)	
min	
max	
	Height Bins
bins	

I use formulas to find the min and max values.



	A	B	C	D	E	F	G	H	I	J	K	L
1	**Id**	**sex**	**age**	**smoke**	**Height (m)**	**Weight (kg)**	**FEV1**	**FVC**	**PEFR**		**Histogram for Height (m)**	
2	64	1	2	2	1.76		4.10	4.90	510		min	MIN(E2:E48)
3	51	1	20	1	1.73		4.31	5.09	480		max	
4	52	1	20	1	1.59		3.45	3.94	490			**Height Bins**
5	53	1	21	2	1.70		4.09	4.80	550		bins	
6	54	1	21	1	1.84		4.65	5.67	600			
7	55	1	21	1	1.81		4.53	5.38	640			
8	56	1	21	2	1.86		4.74	5.60	700			
9	57	1	21	2	1.80		4.53	5.30	525			

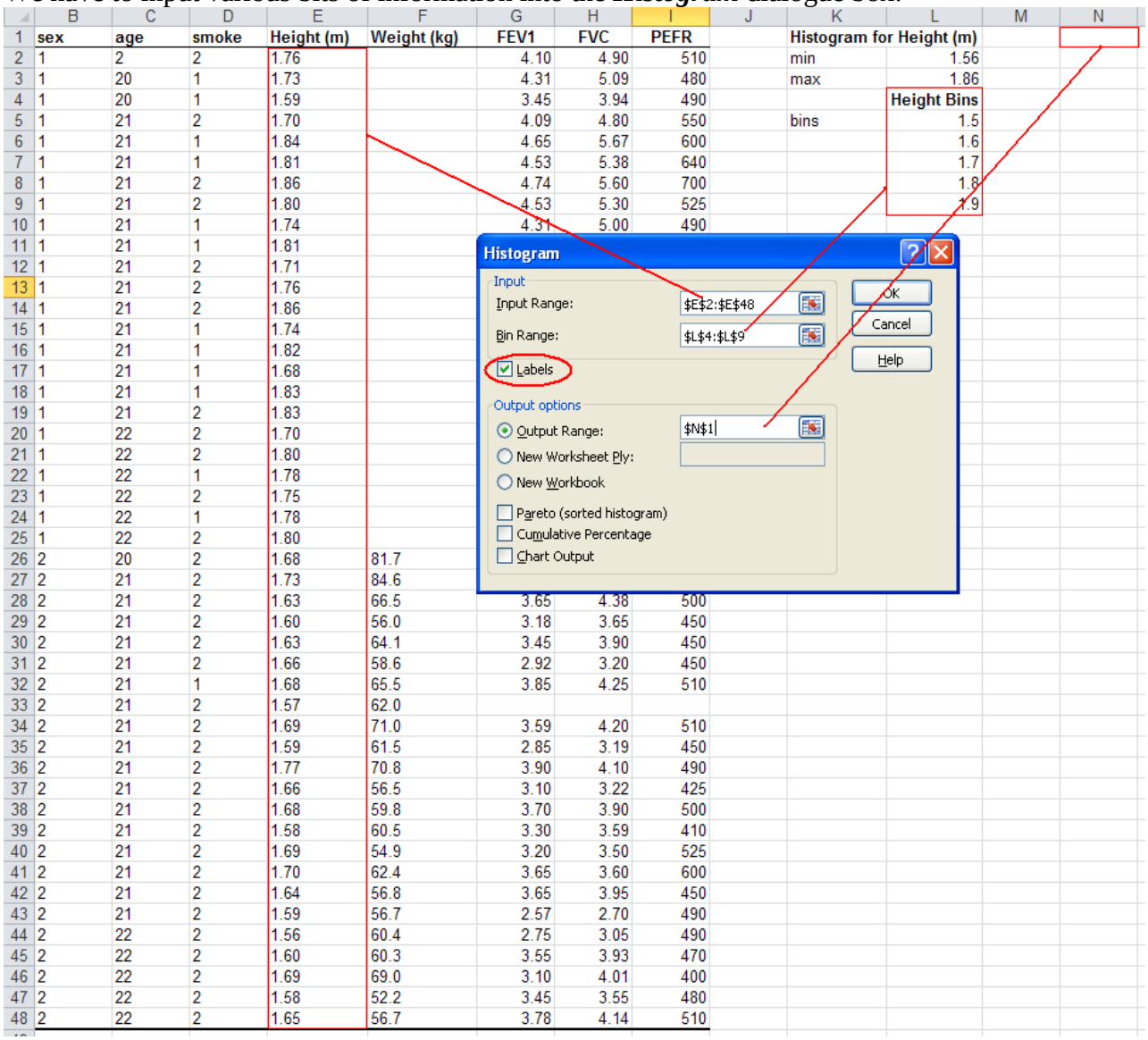
To do this you type "=min(" then you highlight the range of values that you want to find the minimum of and then you close the brackets by typing ")". To find the maximum value you do the same thing but you type "max" instead of "min".

From the minimum and maximum values type out the values for the bins with equal spacing starting from below the minimum to above the maximum. Excel will use this information to produce the information used to produce a histogram.

Histogram for Height (m)	
min	1.56
max	1.86
	Height Bins
bins	1.5
	1.6
	1.7
	1.8
	1.9

Next, select the **Data** tab and click on ***Data Analysis***. The ***Data Analysis*** dialogue box will appear. Select ***Histogram*** and click ***OK***.

We have to input various bits of information into the ***Histogram*** dialogue box.



	B	C	D	E	F	G	H	I	J	K	L	M	N
1	sex	age	smoke	Height (m)	Weight (kg)	FEV1	FVC	PEFR		Histogram for Height (m)			
2	1	2	2	1.76		4.10	4.90	510		min	1.56		
3	1	20	1	1.73		4.31	5.09	480		max	1.86		
4	1	20	1	1.59		3.45	3.94	490			Height Bins		
5	1	21	2	1.70		4.09	4.80	550		bins	1.5		
6	1	21	1	1.84		4.65	5.67	600			1.6		
7	1	21	1	1.81		4.53	5.38	640			1.7		
8	1	21	2	1.86		4.74	5.60	700			1.8		
9	1	21	2	1.80		4.53	5.30	525			1.9		
10	1	21	1	1.74		4.31	5.00	490					
11	1	21	1	1.81									
12	1	21	2	1.71									
13	1	21	2	1.76									
14	1	21	2	1.86									
15	1	21	1	1.74									
16	1	21	1	1.82									
17	1	21	1	1.68									
18	1	21	1	1.83									
19	1	21	2	1.83									
20	1	22	2	1.70									
21	1	22	2	1.80									
22	1	22	1	1.78									
23	1	22	2	1.75									
24	1	22	1	1.78									
25	1	22	2	1.80									
26	2	20	2	1.68	81.7								
27	2	21	2	1.73	84.6								
28	2	21	2	1.63	66.5	3.65	4.38	500					
29	2	21	2	1.60	56.0	3.18	3.65	450					
30	2	21	2	1.63	64.1	3.45	3.90	450					
31	2	21	2	1.66	58.6	2.92	3.20	450					
32	2	21	1	1.68	65.5	3.85	4.25	510					
33	2	21	2	1.57	62.0								
34	2	21	2	1.69	71.0	3.59	4.20	510					
35	2	21	2	1.59	61.5	2.85	3.19	450					
36	2	21	2	1.77	70.8	3.90	4.10	490					
37	2	21	2	1.66	56.5	3.10	3.22	425					
38	2	21	2	1.68	59.8	3.70	3.90	500					
39	2	21	2	1.58	60.5	3.30	3.59	410					
40	2	21	2	1.69	54.9	3.20	3.50	525					
41	2	21	2	1.70	62.4	3.65	3.60	600					
42	2	21	2	1.64	56.8	3.65	3.95	450					
43	2	21	2	1.59	56.7	2.57	2.70	490					
44	2	22	2	1.56	60.4	2.75	3.05	490					
45	2	22	2	1.60	60.3	3.55	3.93	470					
46	2	22	2	1.69	69.0	3.10	4.01	400					
47	2	22	2	1.58	52.2	3.45	3.55	480					
48	2	22	2	1.65	56.7	3.78	4.14	510					

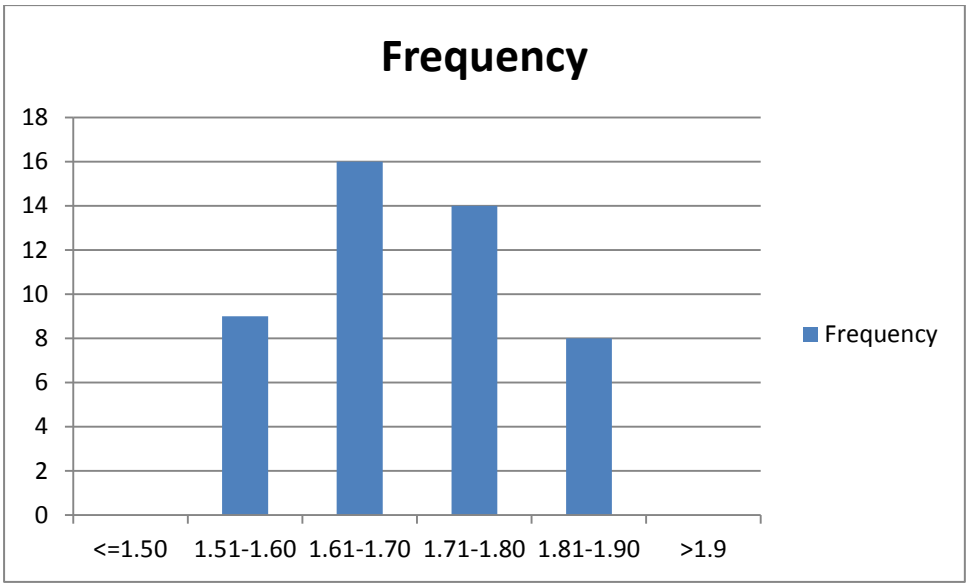
This will give us the following output…

Height Bins	*Frequency*
1.5	0
1.6	9
1.7	16
1.8	14
1.9	8
More	0

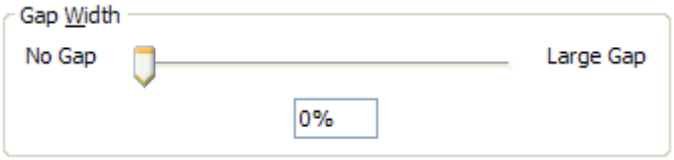
…which I edit so that we know the ranges of each bin. The information typed in here will appear on our histogram so we want it to be informative. Notice that I have used two decimal places because our source data has 2 decimal places.

Height Bins	*Frequency*
<=1.50	0
1.51-1.60	9
1.61-1.70	16
1.71-1.80	14
1.81-1.90	8
>1.9	0

This information is then turned into a chart. Highlight the information, go to the **Insert** tab and click on ***Column*** in the ***Charts*** group and click on ***2D Clustered Column*** to get this…



Double click on one of the columns and then when the ***Format Data Series*** dialogue box appears move the ***Gap Width*** to 0% to get a histogram.



We will edit this in the Editing Charts section later.