




Using Large Data Sets Workbook Version B (Edexcel)

 Edexcel dataset V2 Dec 2016



Updated with minor
corrections 28/4/17



Index

Key Skills	Page 3
Becoming familiar with the dataset	Page 3
Sorting and filtering the dataset	Page 4
Producing a table of summary statistics with GeoGebra	Page 6
Producing a table of summary statistics in Excel	Page 8
Drawing frequency charts, box plots, stem and leaf tables	Page 9
Drawing Scatterplots and curves of best fit	Page 10
Drawing Graphs side by side for comparison	Page 12
Testing for goodness of fit to a normal distribution	Page 13
Carrying out Hypothesis Tests	Page 14
Generating Random Samples	Page 15

Large Data Sets (Edexcel) Workbook

This workbook explores the different types of activities that students and teachers might undertake with a Large Data Set so that it can be used effectively to support the learning of statistical concepts. You will need the Edexcel Dataset

Key Skills

- Understand the dataset and its context
- Cleanse a dataset and know how to deal with outliers
- Sort and Filter the dataset
- Produce a table of summary statistics
- Draw frequency charts, box plots, stem and leaf tables for a set of data
- Draw scatterplots and plot lines and curves of best fit
- Calculate correlation coefficients and equations of regression lines
- Draw graphs of several datasets side by side for comparison
- Test data for goodness of fit to a normal distribution using a quantile plot
- Carry out hypothesis tests on data
- Take a random sample from a dataset

Software Used

- A spreadsheet (in this case excel)
- Graphing and statistical software (in this case GeoGebra).

Other spreadsheets such as Gnumeric, which has a wide range of statistical functions could be used. Likewise Autograph has similar functionality to GeoGebra.

1. Becoming familiar with the dataset

Open the Edexcel dataset V2 Dec 2016 excel file which contains the dataset. The first tab in the spreadsheet explains the source of the data and contains a glossary of terms. Students are required to understand the context of the data so that it is important that they read the glossary whilst looking through the dataset. Some questions you might like to consider are:

What are the sources of the data and how up to date is it?

Who collected it and how was it collected?

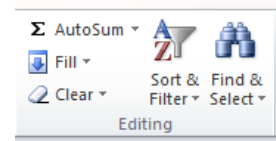
What are the differences in the data for the UK weather stations and those overseas?

What does N/A and tr mean and why are they used? How should we treat these items when analysing the data? Would we treat some fields differently?

Students need to understand each of the fields and how they are determined. Some of them warrant further discussion. Students should be encouraged to research further so that they fully understand the concepts. The [Met Office](#) website can be used to gather data from other years for comparison.

2. Sorting and filtering the dataset

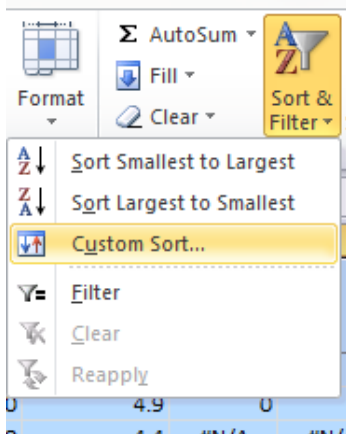
Further familiarity with the dataset can be gained by sorting and filtering the data within Excel. This can help identify any possible outliers or rogue values.



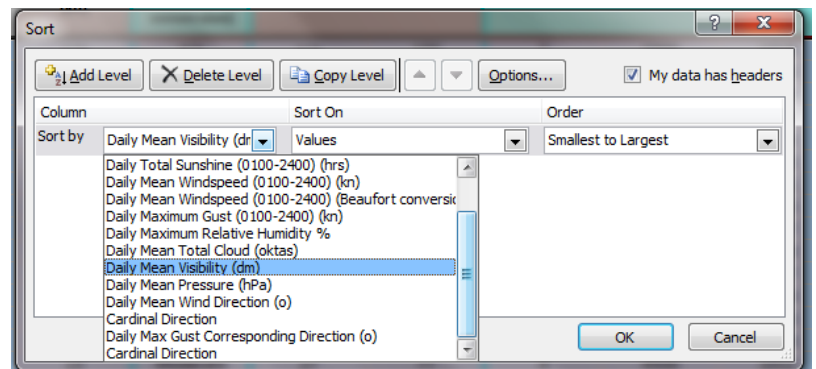
These functions can be found at the far end of the top toolbar:

Suppose we want to sort the Camborne 1987 data according to Daily Mean Visibility. Delete the first five rows of the data and then use Ctrl A to select all the data.

Select the custom sort option:



When the dialogue box appears select the field that you want to sort on and specify the order, smallest to largest. Also make sure that the 'My data has headers' box is checked otherwise your column headings will get sorted as well.

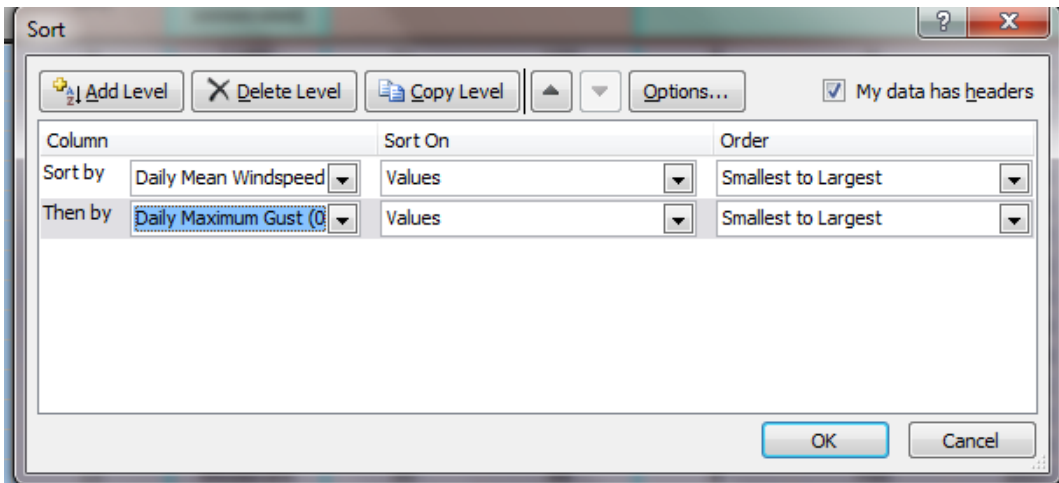


The data is now sorted in order of daily mean visibility.

A	B	C	D	E	F	G	H	I	J
Date	Daily Maximum Temperature (0900-0900) (°C)	Daily Total Rainfall (0900-0900) (mm)	Daily Total Sunshine (0100-2400) (hrs)	Daily Mean Windspeed (0100-2400) (kn)	Daily Mean Windspeed (0100-2400) (Beaufort conversion)	Daily Maximum Gust (0100-2400) (kn)	Daily Maximum Relative Humidity %	Daily Mean Total Cloud (oktas)	Daily Mean Visibility (dm)
28/06/1987	17	0.5	0	9	Light	21	100	8	0
27/06/1987	18	0.4	0	7	Light	15	100	8	200
20/09/1987	17.2	0.1	0	8	Light	21	100	6	200
03/10/1987	16.4	14	0	14	Moderate	37	99	8	200
24/06/1987	16.4	tr	2.5	7	Light	15	99	8	400
02/10/1987	16	12.2	0.3	13	Moderate	32	94	6	400
27/05/1987	12.6	0.1	0	9	Light	21	98	8	500
31/07/1987	18.7	tr	1	11	Moderate	24	100	8	500
31/08/1987	19.2	3.9	3.6	21	Fresh	39	91	5	500
01/09/1987	18.1	0.2	0	4	Light	n/a	100	8	500
17/09/1987	19.1	0.4	0.1	10	Light	23	100	8	500
07/05/1987	14.6	0	n/a	n/a	n/a	n/a	100	3	600
23/07/1987	17.9	0	5.6	7	Light	17	98	6	600
14/07/1987	18.8	3	3	3	Light	n/a	98	7	700
02/08/1987	18.4	2.2	4.3	12	Moderate	30	98	8	700
02/09/1987	19.5	3.5	2.5	6	Light	21	100	8	700
21/09/1987	18.4	3.8	0.5	16	Moderate	33	97	8	700

Looking at days of poorer visibility, we could ask when do these tend to occur? We could also ask students to find out about climatic conditions that lead to poor and good visibility.

It is possible to sort the data using several fields using the 'add level button



Try the above sort (remember to select all the data first using Ctrl A). It should give you the data in order of mean windspeed and then maximum gust

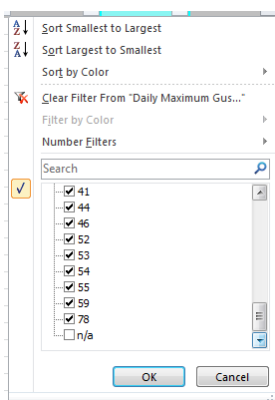
However there are a lot of annoying N/As in the gust field.

2	Light	n/a
2	Light	n/a
2	Light	n/a
3	Light	9
3	Light	10
3	Light	n/a

Click on filter and an arrow should appear next to each heading:

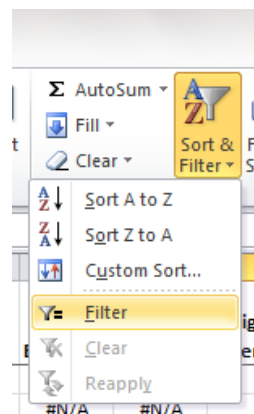


Click on the arrow next to Maximum Gust and then scroll down and uncheck the box next to N/A:



(To turn the filters off click on the filter button again)

However we can get rid of these by using a filter:



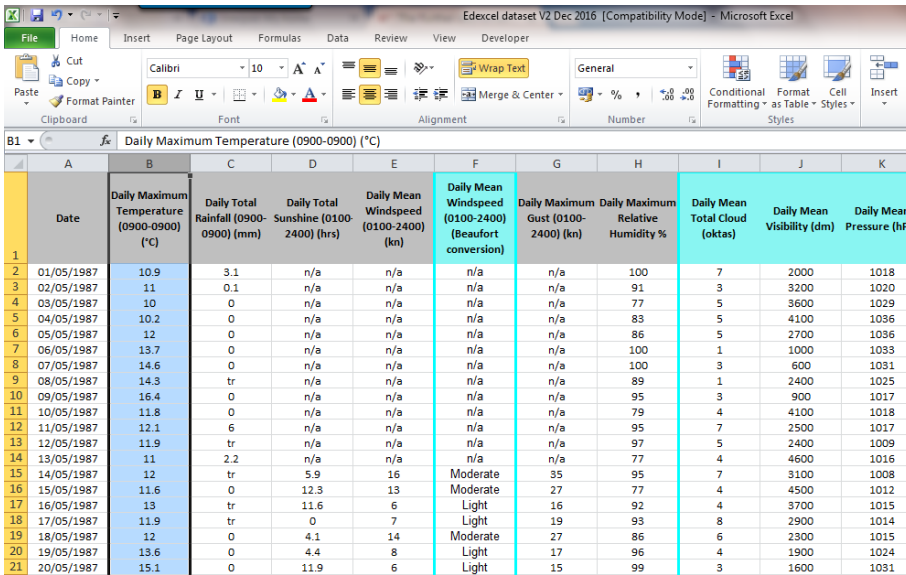
Now we have filtered out those records and we have only those with numerical entries:

3	Light	9
3	Light	10
4	Light	11
4	Light	12
4	Light	12
4	Light	13
4	Light	13
4	Light	13
4	Light	13
4	Light	14
4	Light	15

Exercise: Sort the data by Total Daily Rainfall and filter out the days that say 'tr'. Should these days just be ignored?

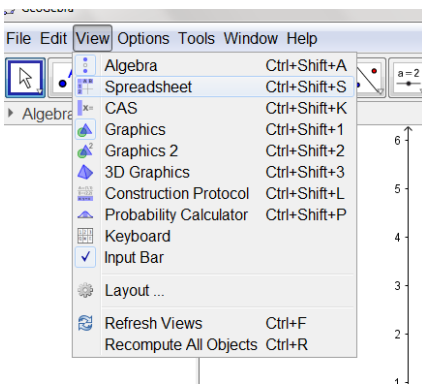
3. Producing a table of summary statistics in Geogebra

Load the excel file of the Edexcel dataset, select the 2nd sheet and highlight column B (Daily Maximum Temp) and copy it (Ctrl C).

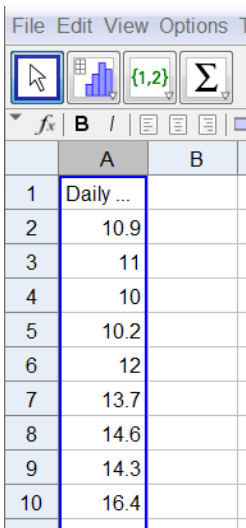


	A	B	C	D	E	F	G	H	I	J	K
	Date	Daily Maximum Temperature (0900-0900) (°C)	Daily Total Rainfall (0900-0900) (mm)	Daily Total Sunshine (0100-2400) (hrs)	Daily Mean Windspeed (0100-2400) (kn)	Daily Mean Windspeed (0100-2400) (Beaufort conversion)	Daily Maximum Gust (0100-2400) (kn)	Daily Maximum Relative Humidity %	Daily Mean Total Cloud (oktas)	Daily Mean Visibility (dm)	Daily Mean Pressure (hPa)
1											
2	01/05/1987	10.9	3.1	n/a	n/a	n/a	n/a	100	7	2000	1018
3	02/05/1987	11	0.1	n/a	n/a	n/a	n/a	91	3	3200	1020
4	03/05/1987	10	0	n/a	n/a	n/a	n/a	77	5	3600	1029
5	04/05/1987	10.2	0	n/a	n/a	n/a	n/a	83	5	4100	1036
6	05/05/1987	12	0	n/a	n/a	n/a	n/a	86	5	2700	1036
7	06/05/1987	13.7	0	n/a	n/a	n/a	n/a	100	1	1000	1033
8	07/05/1987	14.6	0	n/a	n/a	n/a	n/a	100	3	600	1031
9	08/05/1987	14.3	tr	n/a	n/a	n/a	n/a	89	1	2400	1025
10	09/05/1987	16.4	0	n/a	n/a	n/a	n/a	95	3	900	1017
11	10/05/1987	11.8	0	n/a	n/a	n/a	n/a	79	4	4100	1018
12	11/05/1987	12.1	6	n/a	n/a	n/a	n/a	95	7	2500	1017
13	12/05/1987	11.9	tr	n/a	n/a	n/a	n/a	97	5	2400	1009
14	13/05/1987	11	2.2	n/a	n/a	n/a	n/a	77	4	4600	1016
15	14/05/1987	12	tr	5.9	16	Moderate	35	95	7	3100	1008
16	15/05/1987	11.6	0	12.3	13	Moderate	27	77	4	4500	1012
17	16/05/1987	13	tr	11.6	6	Light	16	92	4	3700	1015
18	17/05/1987	11.9	tr	0	7	Light	19	93	8	2900	1014
19	18/05/1987	12	0	4.1	14	Moderate	27	86	6	2300	1015
20	19/05/1987	13.6	0	4.4	8	Light	17	96	4	1900	1024
21	20/05/1987	15.1	0	11.9	6	Light	15	99	3	1600	1031

Open GeoGebra in the spreadsheet view :

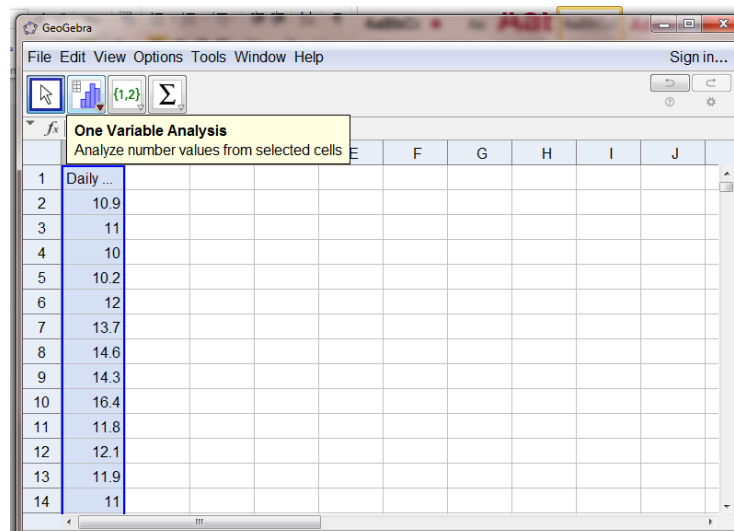


Then paste (Ctrl V) the data into the first column of the spreadsheet:

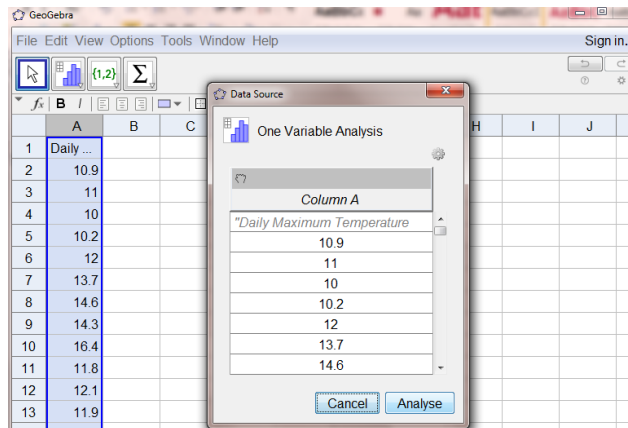


	A	B
1	Daily ...	
2	10.9	
3	11	
4	10	
5	10.2	
6	12	
7	13.7	
8	14.6	
9	14.3	
10	16.4	

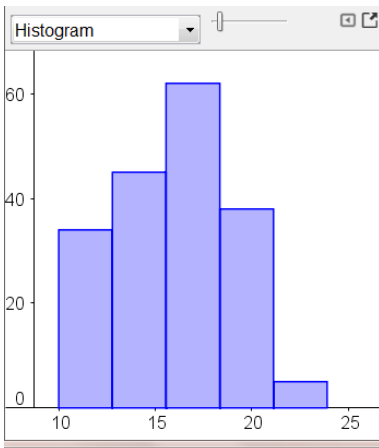
Highlight this column and then click on one variable analysis.



Confirm that you want to analyse this data:

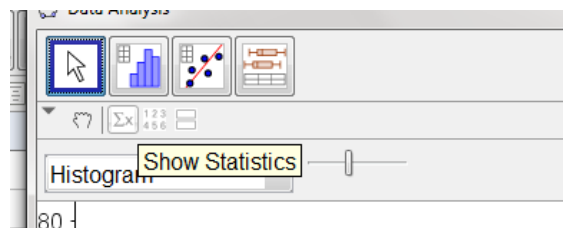


And a new dialogue box will appear:



This Data Analysis window provides a variety of different functions, some of which we consider later.

Click on the Σx icon to show Statistics:



The Statistics box will appear:

Statistics	
n	184
Mean	15.9212
σ	2.9343
s	2.9423
Σx	2929.5
Σx^2	48225.37
Min	10
Q1	13.5
Median	16.05
Q3	18.1
Max	23.9

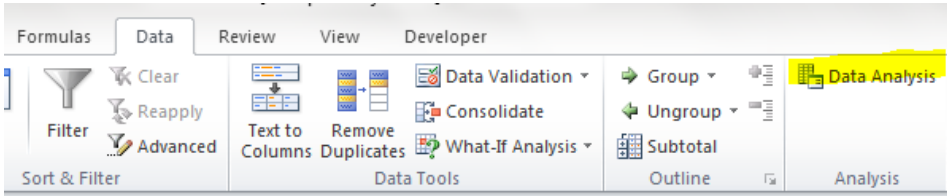
You might need to enlarge the window to see all the digits.

Statistics	
n	184
Mean	18.7837
σ	4.0811
s	4.0922
Σx	3456.2
Σx^2	67984.82
Min	9.4
Q1	15.5
Median	18.95
Q3	21.25
Max	29.2

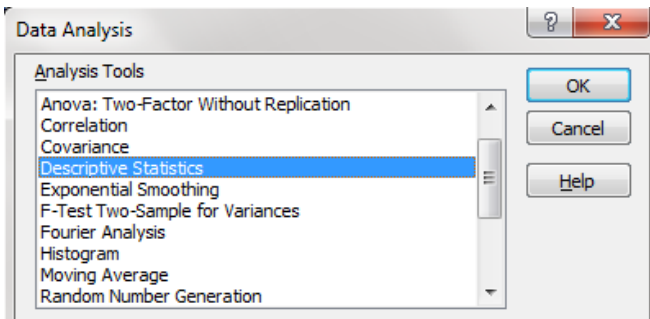
Exercise: Produce the Statistics Box for Daily Maximum Temp for Heathrow May-Oct 1987, which is illustrated to the right.

4. Producing a table of summary statistics in Excel

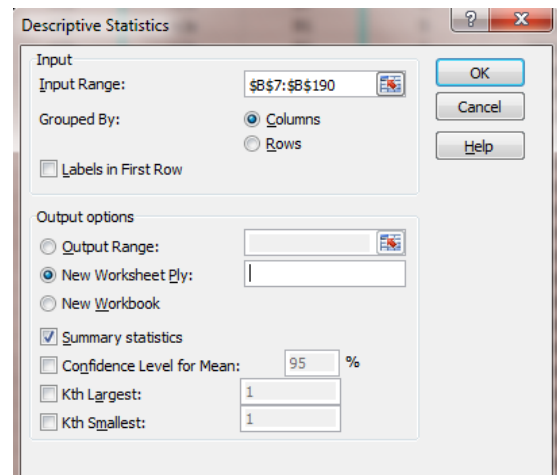
It is possible to produce a Statistics box in Excel but earlier versions require the data analysis add-in which has to be selected first. To select the add-in go to File>Options>Add-ins and select the Analysis Toolpak. Once the add-in is selected it will appear when the data tab is selected:



Click on Data Analysis and a box will appear from which descriptive statistics should be selected:



Then a further dialogue box requires the location of the data to be specified as well as giving the location of the output.



The Summary statistics box also needs to be checked.

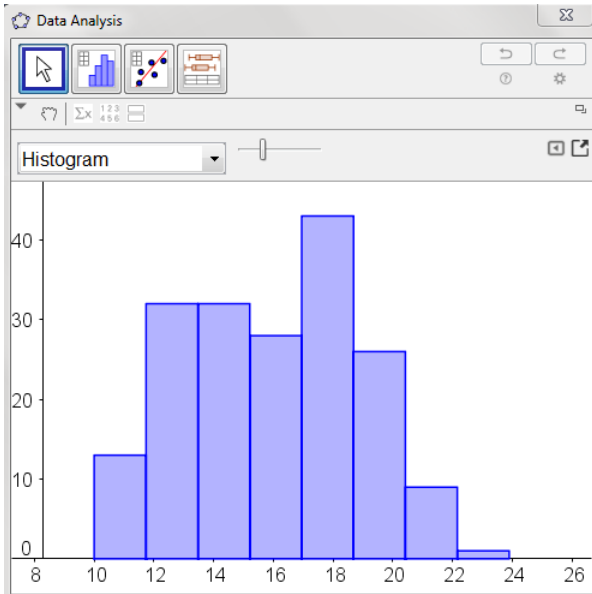
Unlike GeoGebra, Excel can't handle non-numeric data items. So whilst this method will work for 'Daily Maximum Temperature' because it has no N/A entries, Excel will give an error for a field such as 'Daily Total Sunshine' unless the N/A items are filtered out first and the remaining data items selected. This is the box for Daily Maximum Temp for Heathrow May-Oct 1987.

Column1	
Mean	18.7837
Standard Error	0.301685
Median	18.95
Mode	19.9
Standard Deviation	4.09225
Sample Variance	16.74651
Kurtosis	-0.29717
Skewness	0.300408
Range	19.8
Minimum	9.4
Maximum	29.2
Sum	3456.2
Count	184

5. Drawing frequency charts, box plots, stem and leaf tables for a set of data

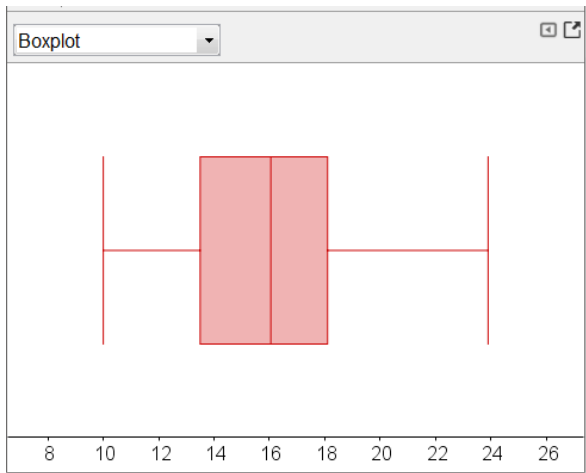
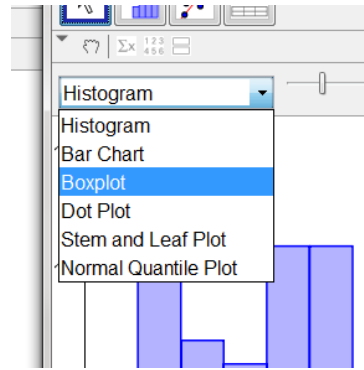
GeoGebra can display a range of graphs and charts. Using Daily Maximum Temp for Camborne May-Oct 1987, follow the previous steps for copying the data into the spreadsheet view and select one variable analysis again.

The default view is Histogram:



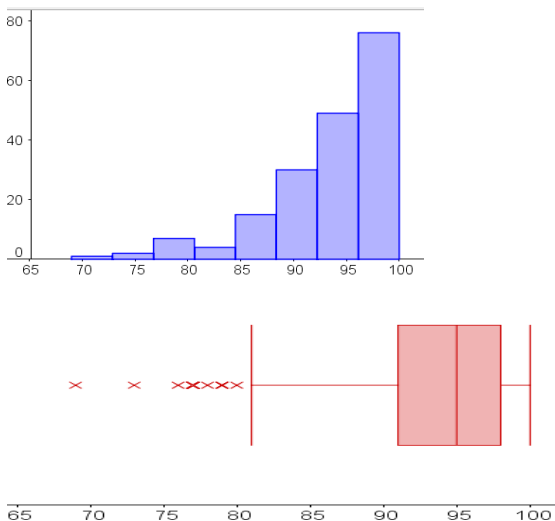
The slider can be used to alter the number of classes and it is interesting to note how the representation changes. This is not really a Histogram as it is that is frequency plotted on the vertical axis, rather than frequency density. All the classes are of equal length.

Different charts can be obtained by changing the option:



What does the box plot show about the daily maximum temperatures at this location?

Exercise: Produce the diagrams below for the relative humidity for Camborne in 1987 and state what they show about the data. How might the graphs for Heathrow differ? Why?



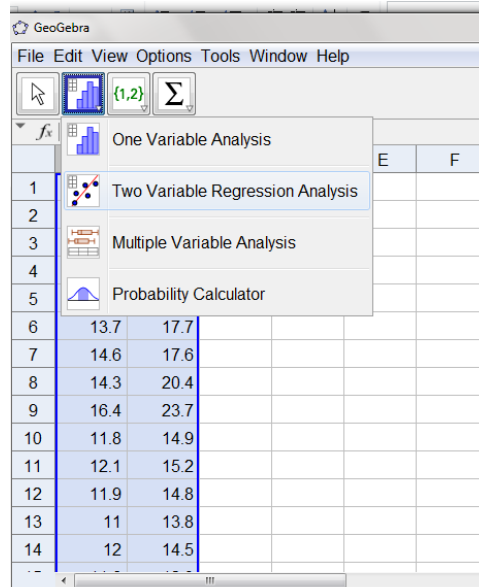
6. Drawing Scatterplots and curves of best fit

Here we will copy and paste two columns of data from Excel into Geogebra with a view to establishing if there is any relationship between the two variables, regarding the data as bivariate data.

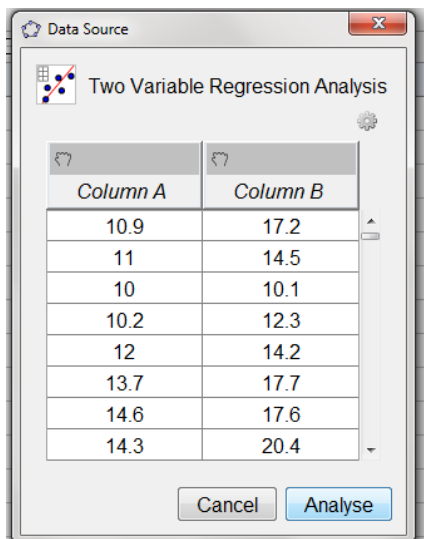
For example we might copy the maximum temperature from Camborne 1987 and Heathrow 1987:

	A	B
1	10.9	17.2
2	11	14.5
3	10	10.1
4	10.2	12.3
5	12	14.2
6	13.7	17.7
7	14.6	17.6
8	14.3	20.4
9	16.4	23.7
10	11.8	14.9
11	12.1	15.2
12	11.9	14.8

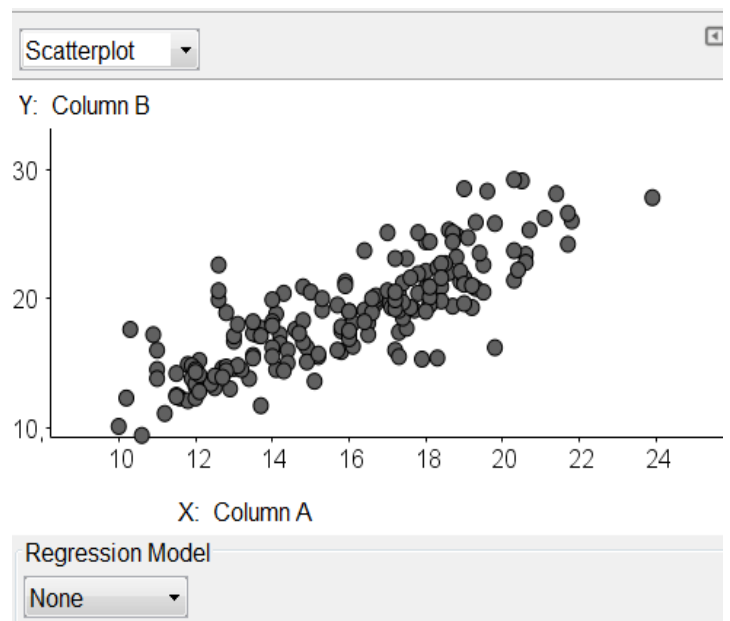
This time we need to highlight both columns and select two variable regression analysis:



Click on analyse:



And a Scatterplot is drawn:



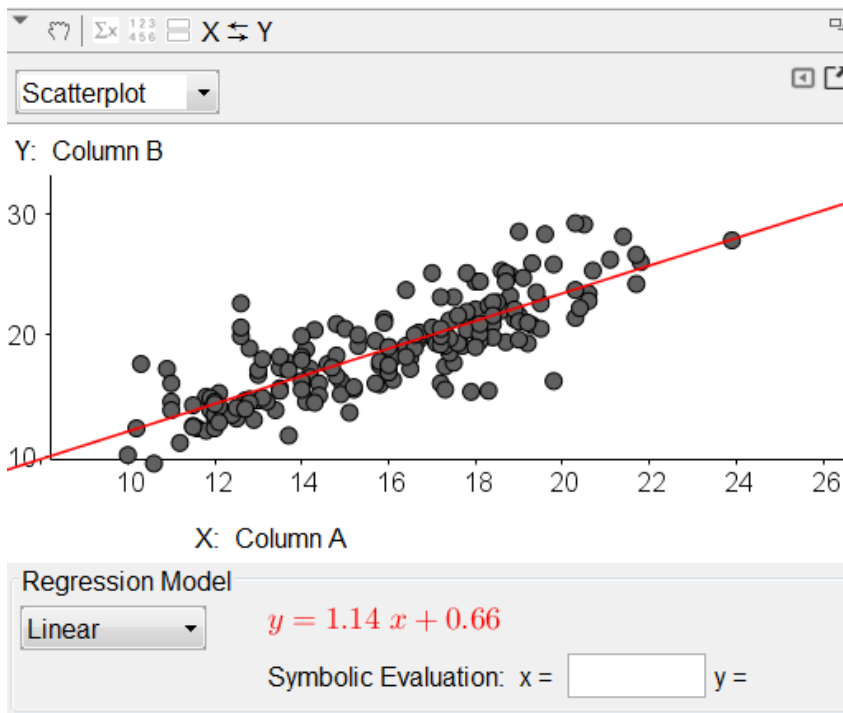
The Scatterplot shows good positive correlation between the two variables.

This can be confirmed by using the Statistics box Σx :

Statistics	
MeanX	15.9212
MeanY	18.7837
Sx	2.9423
Sy	4.0922
r	0.8185
ρ	0.8213
Sxx	1584.2273
Syy	3064.6111
Sxy	1803.4736

What is the difference between r and ρ here?

GeoGebra provides a selection of different types of regression models for this data. GeoGebra will often suggest one to start with but the model can be changed.



Often it is the best linear model that we require. Here the line of best fit is calculated (as the least squares regression line y on x). Values of x can be entered and values of y can be calculated.

Clicking on



will change the line to the x on y regression line.

Exercise: Find the amount of correlation between the rainfall in Camborne(x) and Heathrow(y) in May-Oct 1987. You will need to decide how to deal with trace rainfall. Suggest a regression model for this data. How good is your model?

7. Drawing Graphs side by side for comparison

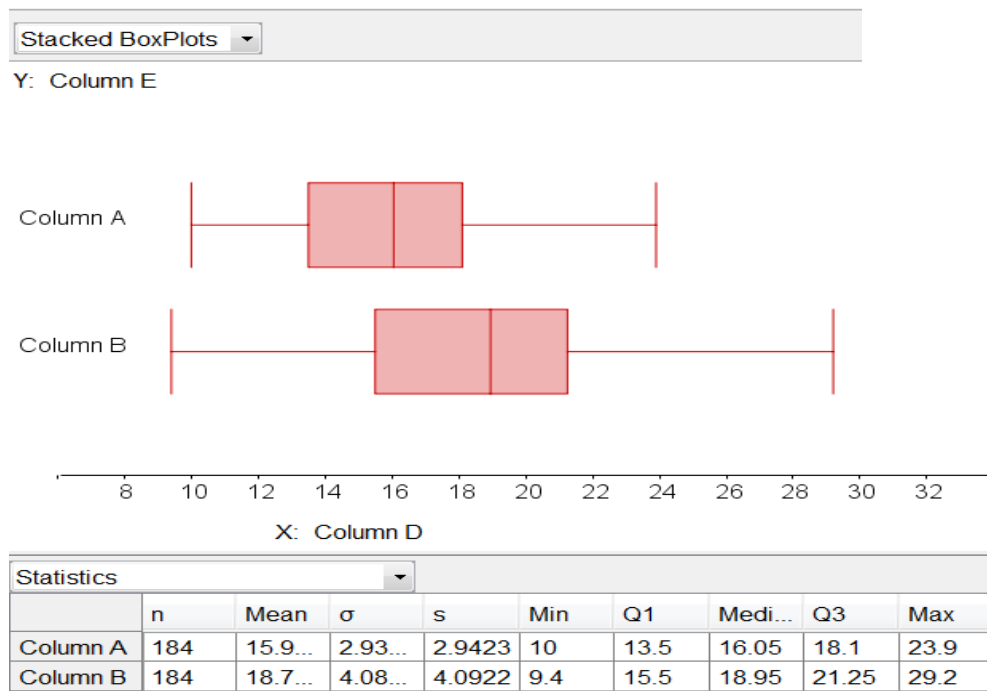
Let's compare the maximum temperatures in Camborne to those in Heathrow. Copy and paste the two sets of 1987 data into GeoGebra:

	A	B
1	10.9	17.2
2	11	14.5
3	10	10.1
4	10.2	12.3
5	12	14.2
6	13.7	17.7
7	14.6	17.6
8	14.3	20.4
9	16.4	23.7
10	11.8	14.9
11	12.1	15.2
12	11.9	14.8
13	11	13.8
14	12	14.5



Highlight both columns and select multi-variable analysis and then analyse.

Box plots are plotted and, if you click on the stats icon, summary stats are also calculated.



What conclusions can be reached by comparing these graphs?

8. Testing for goodness of fit to a normal distribution using a quantile plot

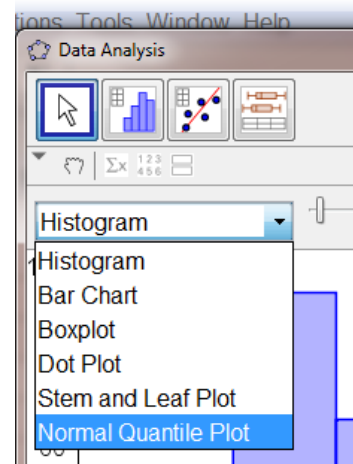
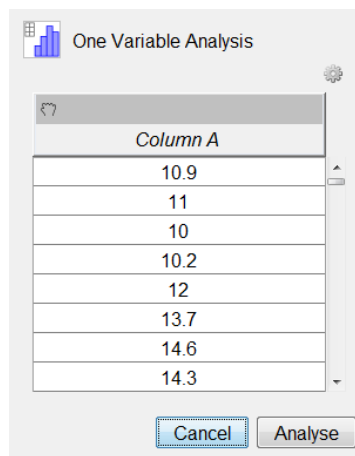
Whilst students are not expected to know any formal goodness of fit tests, to test informally whether data follows a normal distribution we can undertake a normal quantile plot or Q-Q plot. This plot compares the z-values of the data with the quantiles of the standard normal distribution to see how close they are. When plotted against each other, the closer they are to a straight line, the closer the data is to sample from a normal distribution.

GeoGebra has in-built functionality to produce a normal quantile plot

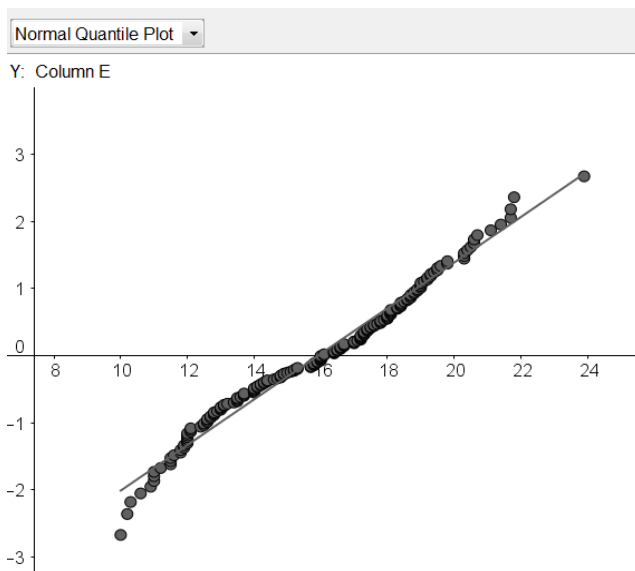
If we use the maximum temperature at Camborne:

	A	B
1	10.9	
2	11	
3	10	
4	10.2	
5	12	
6	13.7	
7	14.6	
8	14.3	
9	16.4	

Highlight the column and select One Variable Analysis and then click Analyse



Select Normal Quantile Plot from the menu and the plot will appear:



Here we see that most of the points lie close to a straight line, especially those in the middle of the data. However there are a few significant deviations either end.

Which days are these? Can they be considered outliers?

Exercise: Construct a Normal Quantile Plot for Daily Mean Pressure for Camborne 1987. What do you conclude?

9. Carrying out Hypothesis Tests

Excel has some in-built tests but they are for comparing 2 samples. The NORM.S.DIST function can be used to work out the probability of a z value and hence to test possible values of a population mean.

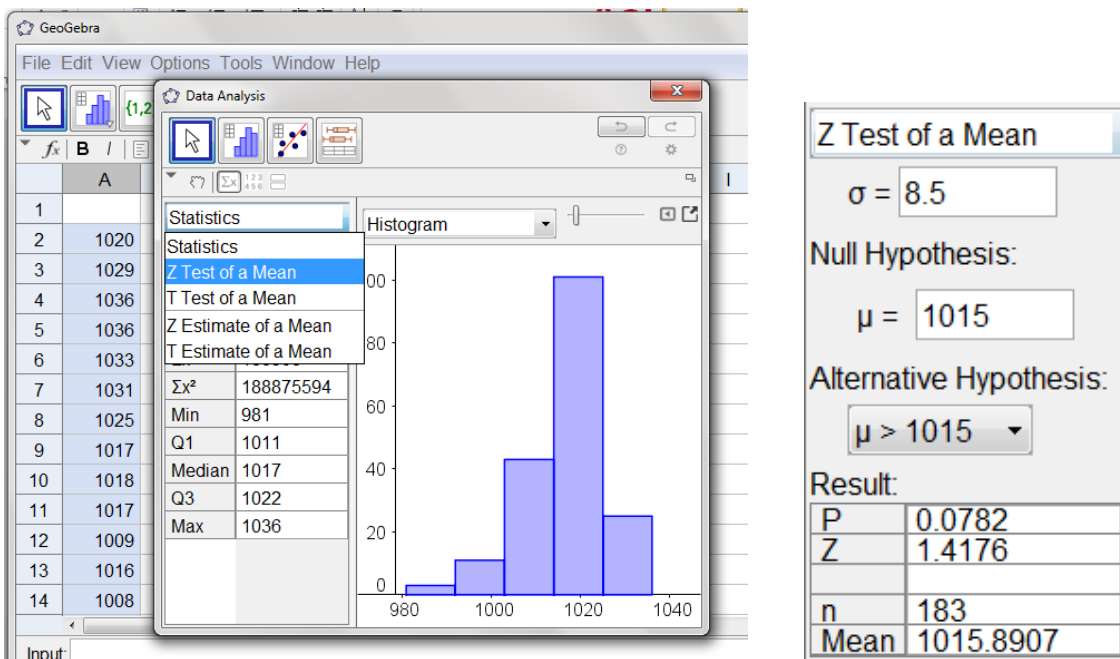
But it is straightforward to use Geogebra to do this:

e.g. to test, at the 5% level, for the daily mean pressure in Camborne 1987 (treating this is a sample) whether

$$H_0: \mu = 1015 \text{ against } H_1: \mu > 1015$$

where μ is the mean of the population from which it is drawn and it is given that the population standard deviation is known and $\sigma = 8.5$

Within the statistical box menu there is an option for a Z-test of a mean:



I select this and then enter the test parameters.

This shows me that the probability that $\bar{X} = 1015.89$ under H_0 is 7.8% and so I would not reject the null hypothesis concluding that there is insufficient evidence that the population mean is greater than 1015.

A few health warnings here. First we have assumed the central limit theorem applies as n is large (183) and so the sample means will be normally distributed. We sneakily used the sample standard deviation as an estimate for the population parameter. This is ok as n is very large but if it were smaller we should really use the t-test instead. It makes a slight difference here giving $p = 0.0808$

Secondly we have also regarded this dataset as a sample from a larger population, whereas in fact it could be argued this is a population and so we know μ . This is a problem with a dataset which constitutes a whole population if we wish to do work on inference. So it might be better to generate random samples from the data and use those to make inferences about the population and then you can see how often you make the correct decision as you will know the population parameters. The next section deals with generating random samples.

10. Generating Random Samples

Many of the models that we use at A level and beyond rely upon the fact that samples have been selected using simple random sampling. It is useful therefore to be able to generate a random sample. The easiest way in Excel is to generate random numbers and then use these to order the data set, selecting the first n items for a sample of size n .

Open the spreadsheet (say Camborne 1987) and insert a new column in Column A.

In the cell in column A below the headers (having deleted the title rows) type **= rand()**

This will generate a random number between 0 and 1.

Now copy this down the whole of column A to the bottom of the data by dragging the bottom right hand corner.

The problem with this facility in Excel is that it will refresh them every time an edit is made. So in order to keep these numbers we need to copy the values.

Select Column A and press Ctrl-C.



Now go to paste-values under the paste menu and paste the values on top of the originals in Column A. Now they will be numbers rather than functions.

We can now sort the data on this column and select the number of rows desired. For example for a sample of size 20:

	A	B	C	D	E	F
		Date	Daily Maximum Temperature (0900-0900) (°C)	Daily Total Rainfall (0900-0900) (mm)	Daily Total Sunshine (0100-2400) (hrs)	Daily Mean Windspeed (0100-2400) (kn)
1						
2	0.004917	04/10/1987	17.2	tr	6.5	13
3	0.007577	23/08/1987	18.9	1.1	12.3	8
4	0.012844	15/10/1987	10.3	30.2	0	11
5	0.026524	24/10/1987	11.5	0	9.2	3
6	0.033618	12/09/1987	17.6	6	0	12
7	0.03603	11/05/1987	12.1	6	n/a	n/a
8	0.040577	24/07/1987	17	tr	0.7	8
9	0.050923	06/09/1987	18.1	11.8	0	13
10	0.051051	12/06/1987	14	0	9.3	4
11	0.051713	18/10/1987	13.7	36.4	0	21
12	0.05364	26/08/1987	16.1	2.7	3.4	15
13	0.061202	30/06/1987	18	tr	3.1	6
14	0.065553	30/05/1987	12.6	7.1	0	11
15	0.070041	22/05/1987	12.5	0.3	7.8	11
16	0.076215	10/10/1987	10.6	5.2	6.3	13
17	0.07772	09/06/1987	12	1	9.9	6
18	0.086608	14/07/1987	18.8	3	3	3
19	0.088174	30/10/1987	12.9	9.3	0	13
20	0.106029	02/08/1987	18.4	2.2	4.3	12
21	0.113032	13/09/1987	16	0.3	1.7	2

The advantage of this method is that we can take a sample on all fields at the same time and we know which dates are in the sample.

Exercise: Take a sample of size 40 and copy the daily maximum temperature data into GeoGebra. Perform a z hypothesis test at the 5% level of $H_0: \mu = 15$ against $H_1: \mu \neq 15$ with $\sigma = 2.9$

What is your result? Repeat for a second random sample. Do you get the same result?