Using Large Data Sets Workbook – Edexcel version

This booklet uses Excel and Desmos

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 which can be downloaded at: *https://qualifications.pearson.com/en/qualifications/edexcel-a-levels/mathematics-2017.coursematerials.html*

Key Skills

- Understand the dataset and its context
- Clean a dataset and know how to deal with outliers
- Sort and Filter the dataset
- Produce summary statistics
- Draw frequency charts and box plots for a set of data
- Draw graphs of several datasets side by side for comparison
- Draw scatterplots and plot lines and curves of best fit
- Use technology to calculate correlation coefficients and equations of regression lines
- Take a random sample from a dataset

Software Used

- A spreadsheet (in this case Excel)
- Graphing and statistical software (in this case Desmos).

Becoming familiar with the dataset

Open the "Pearson Excel GCE AS and AL Mathematics data set – Issue 1" 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 important that they read the glossary whilst looking through the dataset.

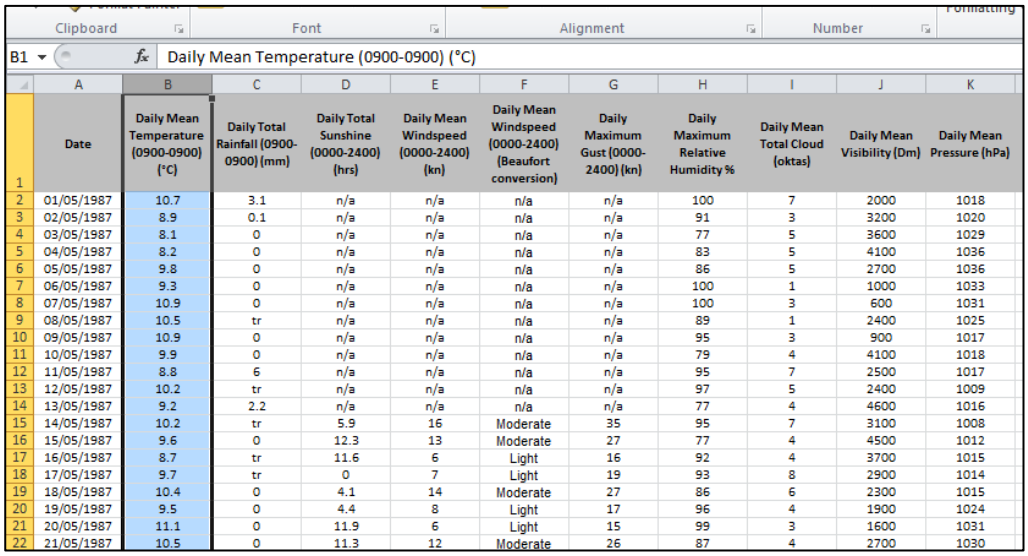
Some questions you might like to consider are:

- *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 (*http://www.metoffice.gov.uk/public/weather/climate-historic/*) can be used to gather data from other years for comparison.

1 Producing summary statistics

Load the excel file of the Edexcel dataset, select the 2nd sheet (Camborne May-Oct 1987) and highlight column B (Daily Mean Temp) and copy it (Ctrl-C).

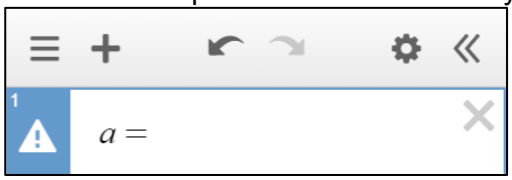


Clipboard | Font | Alignment | Number

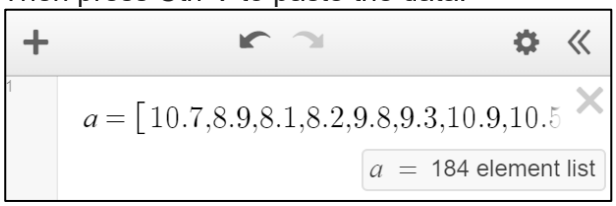
B1 — Daily Mean Temperature (0900-0900) (°C)

	A	B	C	D	E	F	G	H	I	J	K
1	Date	Daily Mean Temperature (0900-0900) (°C)	Daily Total Rainfall (0900-0900) (mm)	Daily Total Sunshine (0000-2400) (hrs)	Daily Mean Windspeed (0000-2400) (kn)	Daily Mean Windspeed (0000-2400) (Beaufort conversion)	Daily Maximum Gust (0000-2400) (kn)	Daily Maximum Relative Humidity %	Daily Mean Total Cloud (oktas)	Daily Mean Visibility (Dm)	Daily Mean Pressure (hPa)
2	01/05/1987	10.7	3.1	n/a	n/a	n/a	n/a	100	7	2000	1018
3	02/05/1987	8.9	0.1	n/a	n/a	n/a	n/a	91	3	3200	1020
4	03/05/1987	8.1	0	n/a	n/a	n/a	n/a	77	5	3600	1029
5	04/05/1987	8.2	0	n/a	n/a	n/a	n/a	83	5	4100	1036
6	05/05/1987	9.8	0	n/a	n/a	n/a	n/a	86	5	2700	1036
7	06/05/1987	9.3	0	n/a	n/a	n/a	n/a	100	1	1000	1033
8	07/05/1987	10.9	0	n/a	n/a	n/a	n/a	100	3	600	1031
9	08/05/1987	10.5	tr	n/a	n/a	n/a	n/a	89	1	2400	1025
10	09/05/1987	10.9	0	n/a	n/a	n/a	n/a	95	3	900	1017
11	10/05/1987	9.9	0	n/a	n/a	n/a	n/a	79	4	4100	1018
12	11/05/1987	8.8	6	n/a	n/a	n/a	n/a	95	7	2500	1017
13	12/05/1987	10.2	tr	n/a	n/a	n/a	n/a	97	5	2400	1009
14	13/05/1987	9.2	2.2	n/a	n/a	n/a	n/a	77	4	4600	1016
15	14/05/1987	10.2	tr	5.9	16	Moderate	35	95	7	3100	1008
16	15/05/1987	9.6	0	12.3	13	Moderate	27	77	4	4500	1012
17	16/05/1987	8.7	tr	11.6	6	Light	16	92	4	3700	1015
18	17/05/1987	9.7	tr	0	7	Light	19	93	8	2900	1014
19	18/05/1987	10.4	0	4.1	14	Moderate	27	86	6	2300	1015
20	19/05/1987	9.5	0	4.4	8	Light	17	96	4	1900	1024
21	20/05/1987	11.1	0	11.9	6	Light	15	99	3	1600	1031
22	21/05/1987	10.5	0	11.3	12	Moderate	26	87	4	2700	1030

Go to a new input bar in Desmos and type:

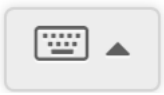


Then press Ctrl-V to paste the data:

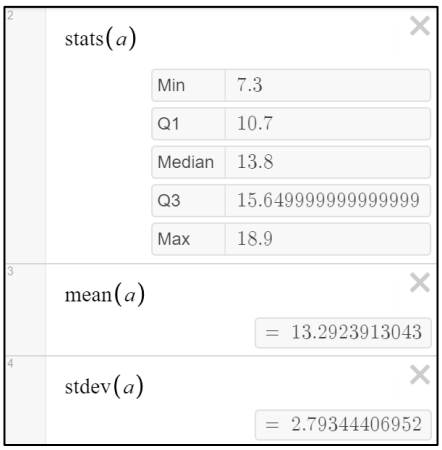


This should create the list (called a). You can now refer to the list in other commands.

The following commands can be found using functions > Stats from the onscreen keypad:



- $stats(a)$
- $mean(a)$
- $stdev(a)$



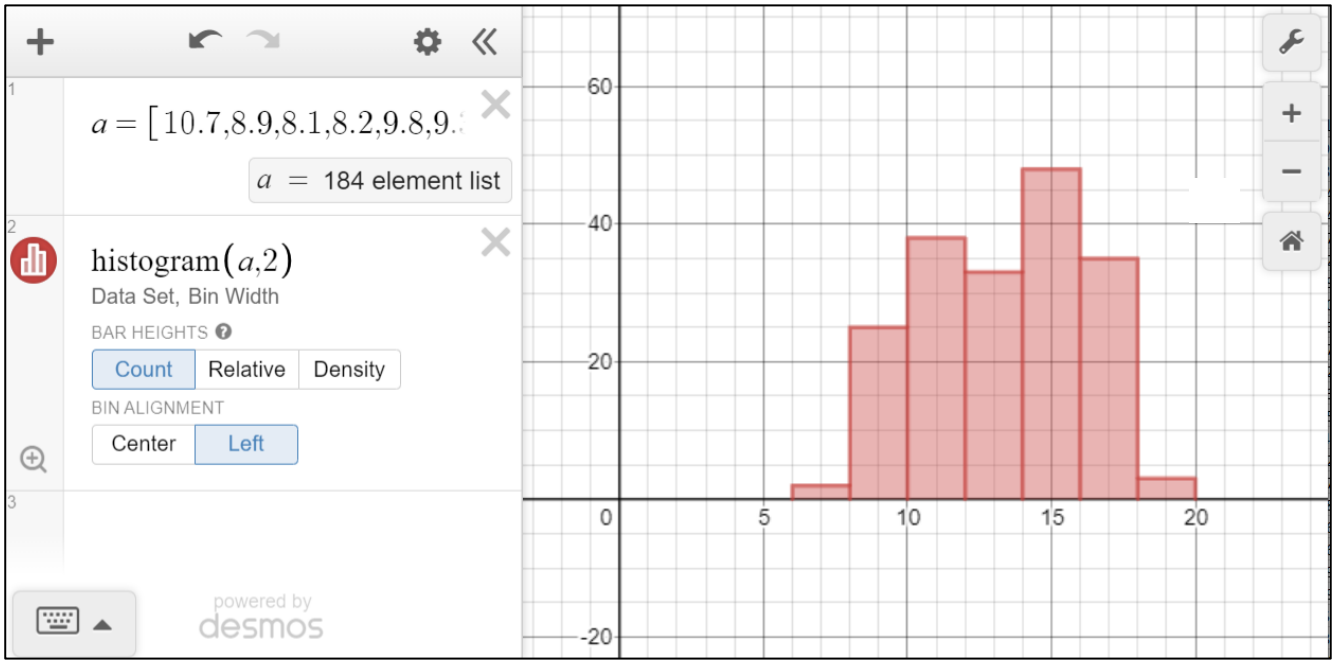
Exercise: Produce these statistics for the different columns from Cambourne May-Oct 1987 and Heathrow May-Oct 1987. What similarities or differences about the weather do these statistics show?

2 Drawing frequency charts and box plots for a set of data

Desmos can display a range of graphs and charts. You can use the previous steps for copying the data into Desmos and then select a visualization from: functions > Dist

Desmos includes both histograms and boxplots.

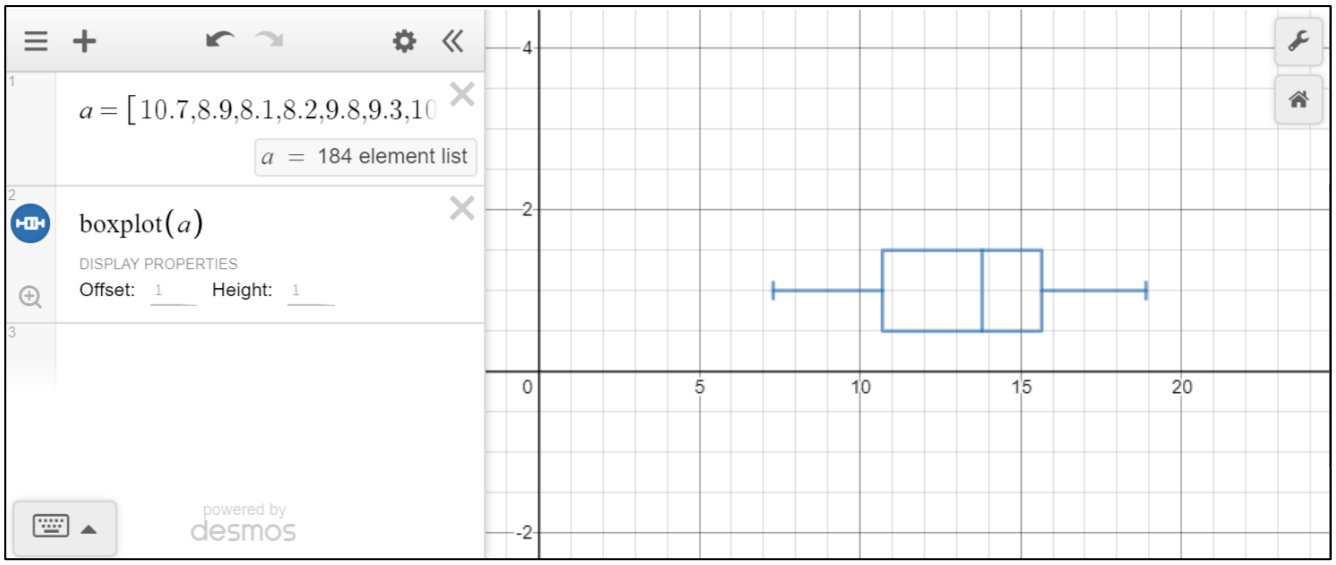
For a histogram you should enter the data set and the bin width.



The magnifying glass icon in the input row can be used to auto-scale. Setting the bin alignment to Left is often more useful. For bins of width 2 the first bin will contain values of x where $0 \leq x < 2$.

Desmos uses a definition of histogram that has frequency on the vertical axis and equal interval widths on the horizontal axis.

For a boxplot enter the data set.



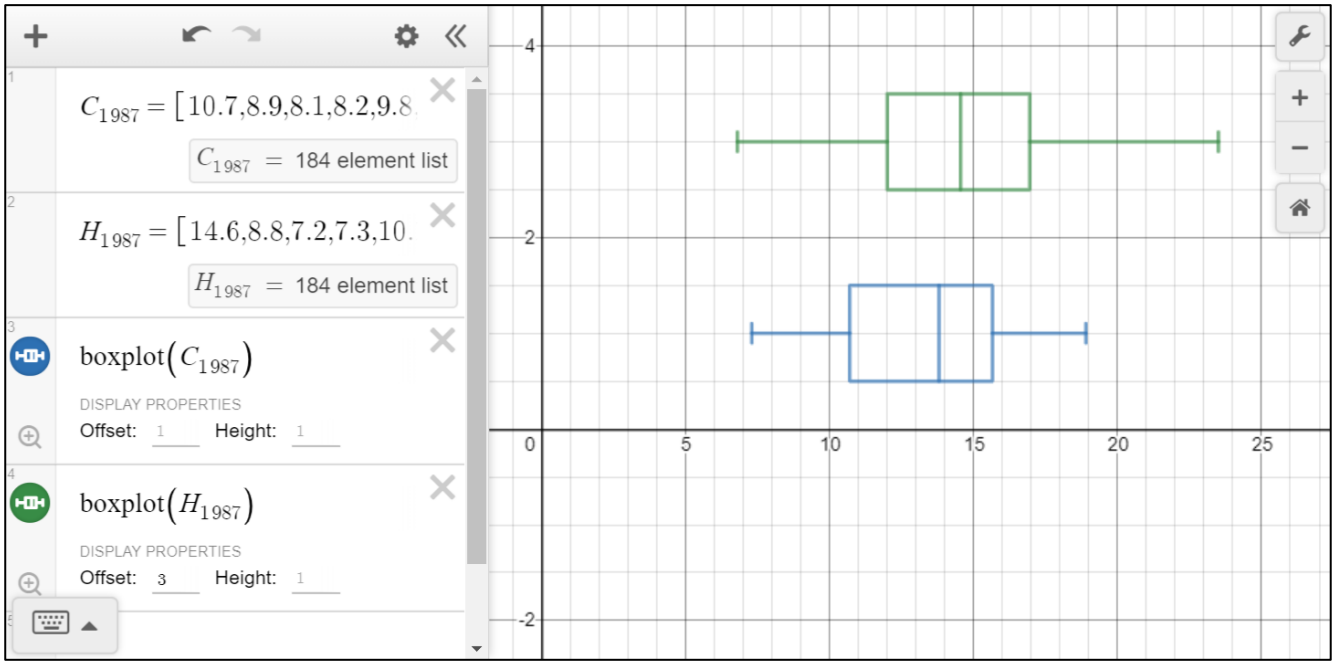
What picture of the data does the box plot give?

3 Drawing graphs side by side for comparison

The following example compares the daily temperature for Cambourne and Heathrow for May-Oct 1987.

Each set of numbers will need to be copied as a new list into Desmos.

Copying each set across into Desmos (NB just type C1987 to obtain the variable name with the subscript):

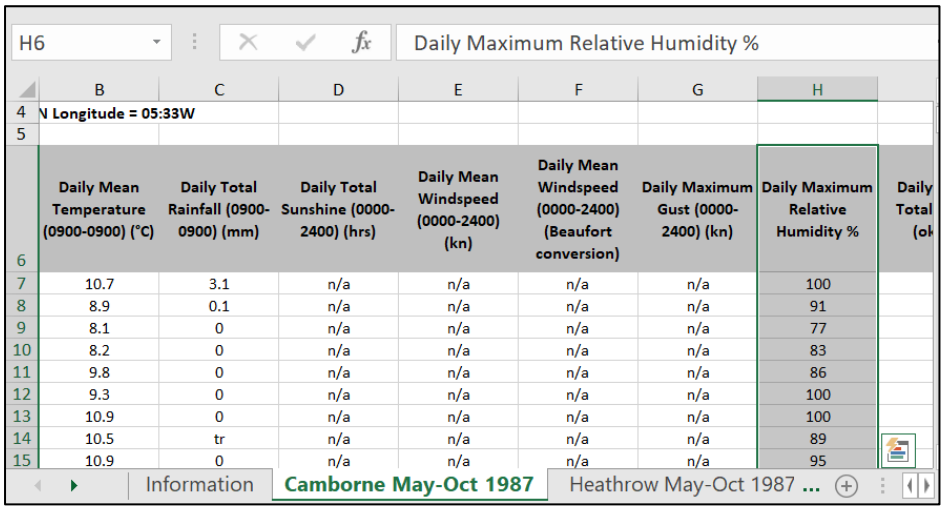


Exercise: Draw the boxplots for some the mean temperature in some other locations for 1987. What conclusions can be reached by comparing these plots?

4 Drawing Scatterplots

You can investigate whether there is a relationship between two variables, regarding the data as bivariate data, by copying and pasting two columns of data from Excel into Desmos.

In the following example the mean temperature will be compared against the relative humidity for Cambourne 1987. These two columns have been copied by selecting each one separately and copying into a new blank spreadsheet. The pair of adjacent columns in the new spreadsheet can then be copied and pasted into Desmos.

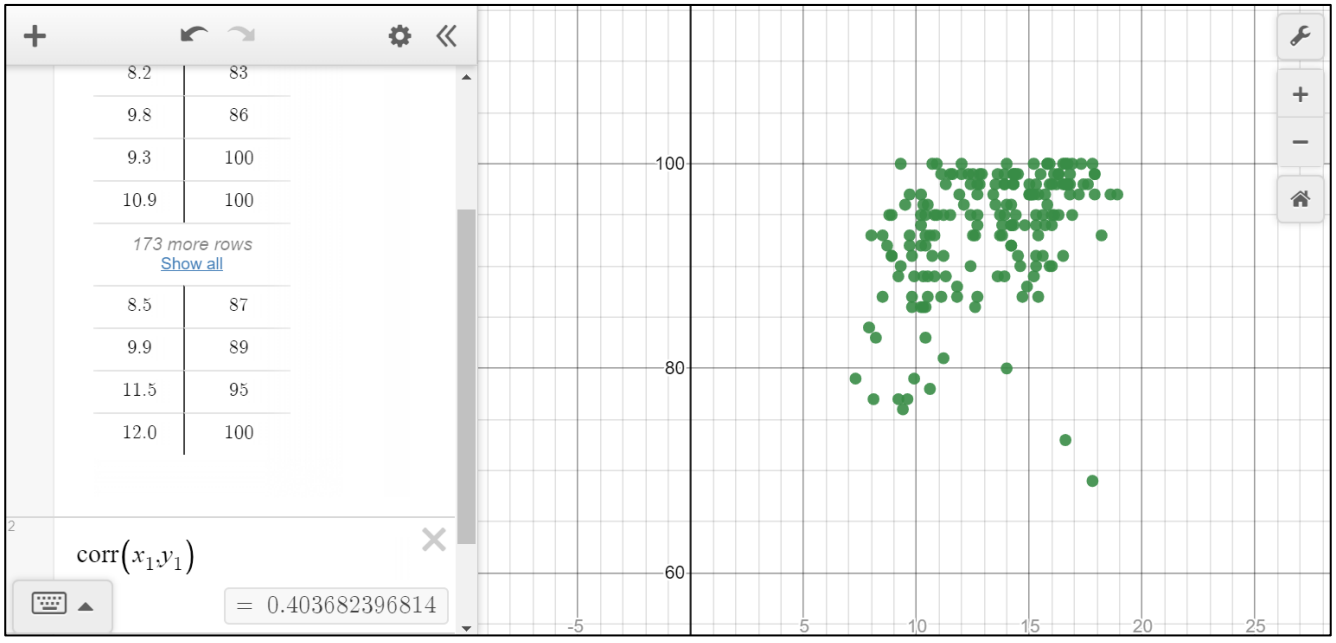


H6 — Daily Maximum Relative Humidity %

	B	C	D	E	F	G	H
4	N Longitude = 05:33W						
5							
6	Daily Mean Temperature (0900-0900) (°C)	Daily Total Rainfall (0900-0900) (mm)	Daily Total Sunshine (0000-2400) (hrs)	Daily Mean Windspeed (0000-2400) (kn)	Daily Mean Windspeed (0000-2400) (Beaufort conversion)	Daily Maximum Gust (0000-2400) (kn)	Daily Maximum Relative Humidity %
7	10.7	3.1	n/a	n/a	n/a	n/a	100
8	8.9	0.1	n/a	n/a	n/a	n/a	91
9	8.1	0	n/a	n/a	n/a	n/a	77
10	8.2	0	n/a	n/a	n/a	n/a	83
11	9.8	0	n/a	n/a	n/a	n/a	86
12	9.3	0	n/a	n/a	n/a	n/a	100
13	10.9	0	n/a	n/a	n/a	n/a	100
14	10.5	tr	n/a	n/a	n/a	n/a	89
15	10.9	0	n/a	n/a	n/a	n/a	95

Information | Camborne May-Oct 1987 | Heathrow May-Oct 1987 ...

	A	B
1	Daily Mean Temperature (0900-0900) (°C)	Daily Maximum Relative Humidity %
2	10.7	100
3	8.9	91
4	8.1	77
5	8.2	83
6	9.8	86
7	9.3	100
8	10.9	100
9	10.5	89



The Scatterplot shows some positive correlation between the two variables. This can be confirmed by calculating: $\text{corr}(x_1, y_1)$

NB for subscripts typing $x1$ will automatically be updated to x_1.

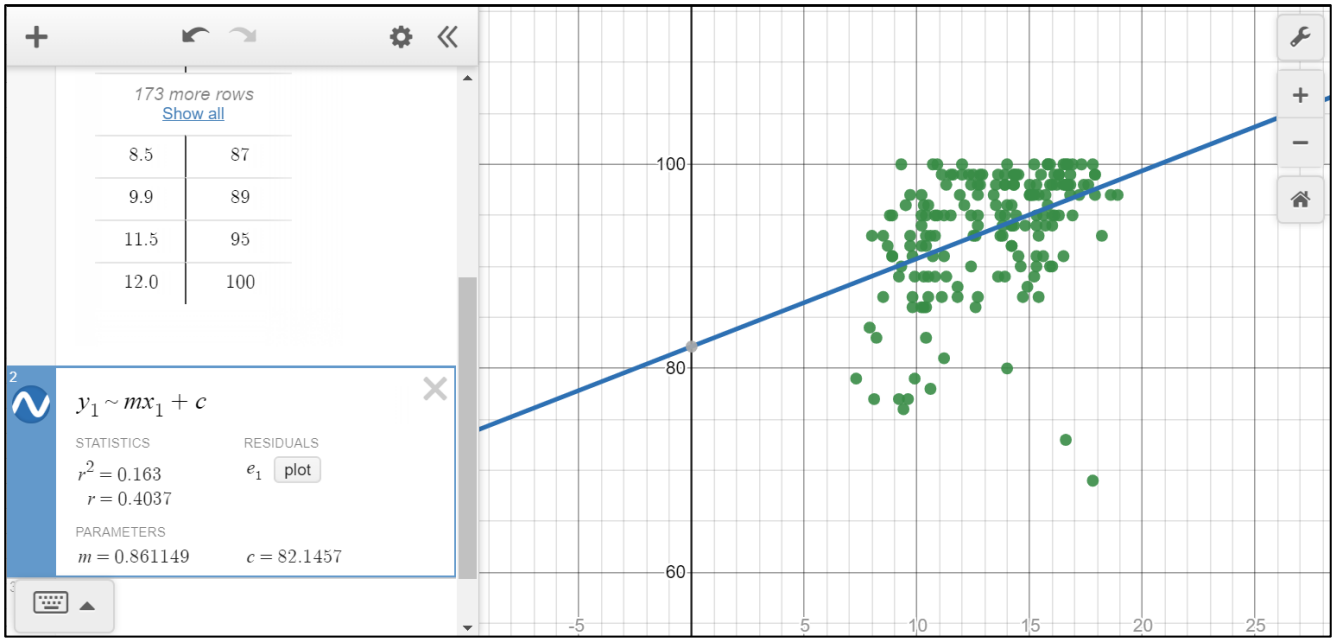
5 Regression models

Desmos allows for any regression model that can be defined as a generalised function. In practice students will mainly use a linear model: $y = mx + c$.

Bivariate data can be copied into Desmos as described in section 6.

The regression model is defined using the tilde symbol, ~, e.g. $y_1 \sim mx_1 + c$.

The example below shows this for the data used in the previous section: the mean temperature will be compared against the relative humidity for Cambourne 1987.



Students are expected to use (but not derive) non-linear models for data. For example you might consider a log model instead.

The residual sum of squares (RSS), also known as the sum of squared errors of prediction (SSE) gives a guide to how good a fit the model will be. The residuals can be plotted using the plot button.

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?

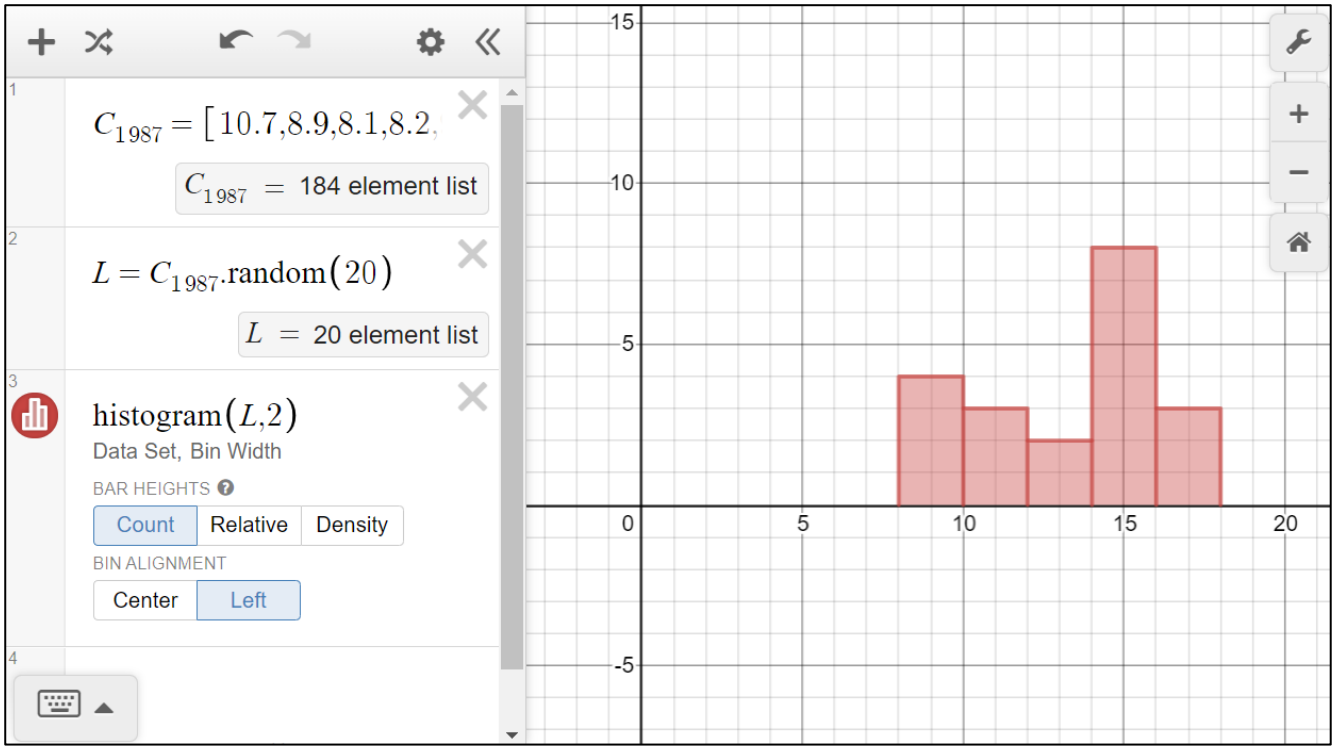
6 Random Sampling

You can use the random function to select a random sample from a list of values. In the following example a random sample of size 20 will be selected from the Daily mean temperature for Cambourne in 1987.

Enter a new variable of C_{1987} and copy the full list into Desmos.

To create a random sample of size 20 type: $L = C_{1987}.\text{random}(20)$

A histogram for this sample can be plotted using $\text{histogram}(L)$.



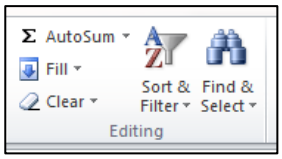
To generate different samples click on  (second icon top-left).

One application of this method is in selecting several samples of the same size and comparing statistics such as the mean or the standard deviation with their true values in the whole dataset. This illustrates the idea of statistical variation. The sample size can then be changed to see how that affects the variation.

Appendix 1: Sorting and filtering the dataset in Excel

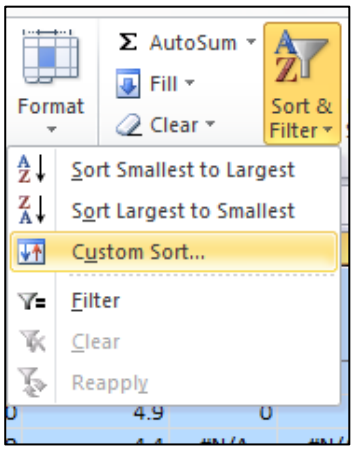
Further familiarity with the dataset can be gained by sorting and filtering the data within Excel. This can help identify any possible outliers or rogue values.

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

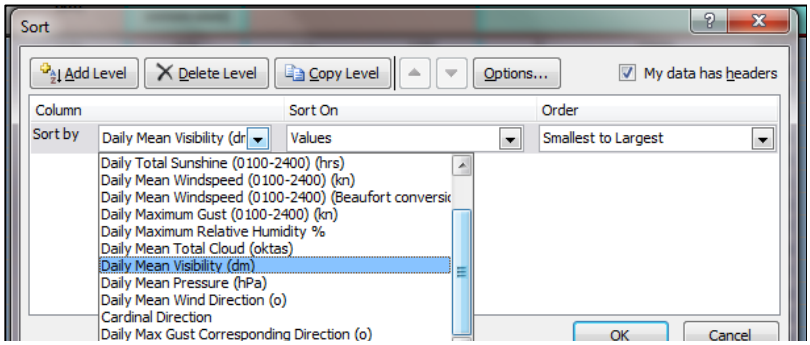


Suppose you want to sort the Camborne 1987 data according to Daily Mean Visibility. Delete the first five rows of the data and then use Crtl-A to select all the data.

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.

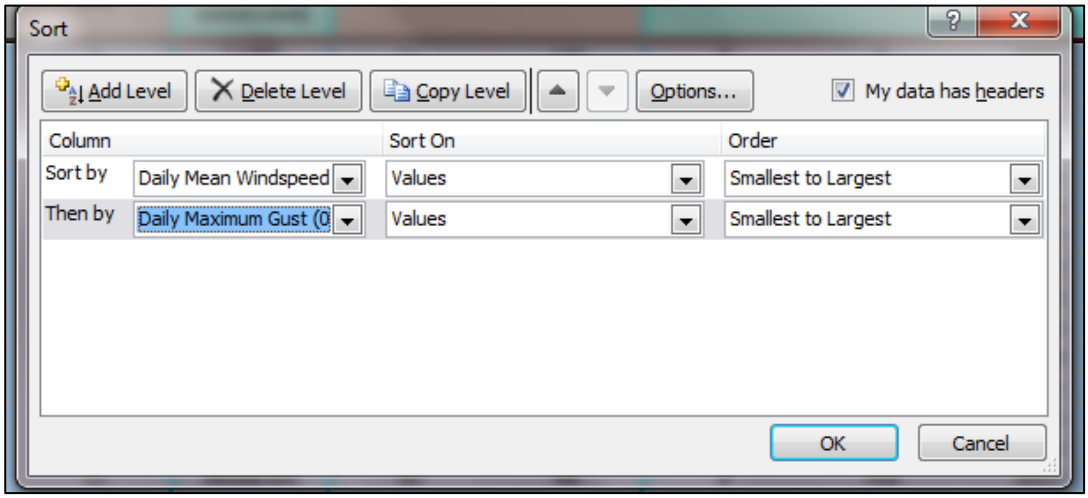


The data is now sorted in order of daily mean visibility.

D	E	F	G	H	I	J	K	L	M	N
Daily Total Sunshine (0000-2400) (hrs)	Daily Mean Windspeed (0000-2400) (kn)	Daily Mean Windspeed (0000-2400) (Beaufort conversion)	Daily Maximum Gust (0000-2400) (kn)	Daily Maximum Relative Humidity %	Daily Mean Total Cloud (oktas)	Daily Mean Visibility (Dm)	Daily Mean Pressure (hPa)	Daily Mean Wind Direction (o)	Cardinal Direction	Daily Max Gust Corresponding Direction (o)
0	9	Light	21	100	8	0	1021	190	S	190
0	7	Light	15	100	8	200	1016	230	SW	200
0	8	Light	21	100	6	200	1013	210	SSW	NA
0	14	Moderate	37	99	8	200	1015	80	E	90
2.5	7	Light	15	99	8	400	1023	290	WNW	280
0.3	13	Moderate	32	94	6	400	1015	90	E	90
0	9	Light	21	98	8	500	1017	320	NW	360
1	11	Moderate	24	100	8	500	1022	290	WNW	290
3.6	21	Fresh	39	91	5	500	1015	100	E	100
0	4	Light	n/a	100	8	500	1014	320	NW	110
0.1	10	Light	23	100	8	500	1018	210	SSW	250
n/a	n/a	n/a	n/a	100	3	600	1031	350	NNW	350
5.6	7	Light	17	98	6	600	1022	360	NNW	360
3	3	Light	n/a	98	7	700	1013	290	WNW	NA
4.3	12	Moderate	30	98	8	700	1017	270	W	270
2.5	6	Light	21	100	8	700	1017	190	S	NA
0.5	16	Moderate	33	97	8	700	1008	180	S	180
0.6	5	Light	17	100	7	800	1014	150	SSE	310
n/a	n/a	n/a	n/a	95	3	900	1017	360	NNW	NA
0	11	Moderate	28	99	7	900	1021	200	SSW	200
0	12	Moderate	31	99	8	900	1014	260	W	190
12.2	6	Light	15	95	3	900	1020	330	NNW	340

Looking at days of poorer visibility, you could ask when do these tend to occur? You 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:



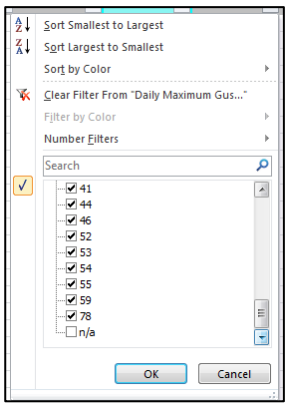
Try the above sort (remember to select all the data first using Crtl-A). It should give you the data in order of mean windspeed and then maximum gust

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

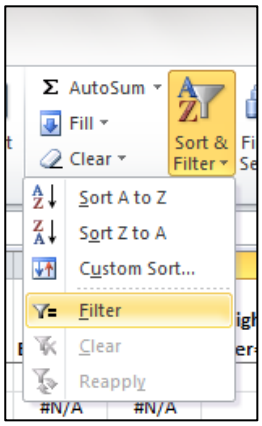
Click on filter and an arrow should appear next to each heading:

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.

You can get rid of these by using a filter:



Now you 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'. Compare the effect of ignoring these days, replacing the values with 0 or replacing these values with 0.025.

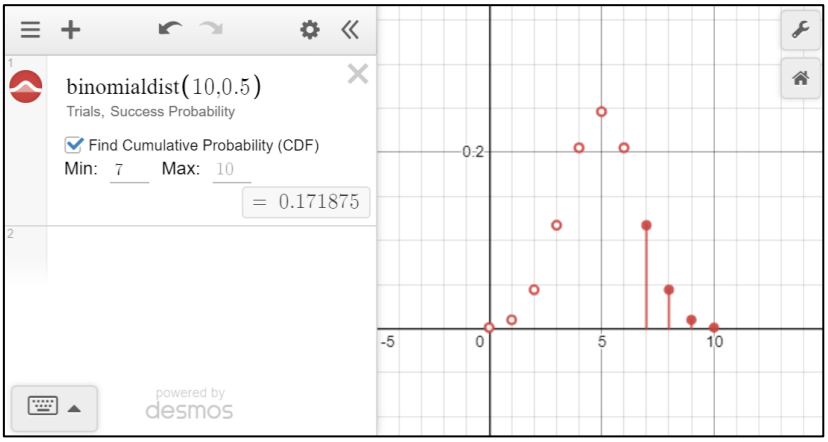
Appendix 2: Ideas for investigations

- How does the temperature vary with location?
- How does the wind speed vary with location?
- Are there any differences in weather from 1987 to 2015 in any of the locations?
- Is there a correlation between temperature and rainfall?
- Is there a correlation between temperature and sunshine?
- Is there a correlation between pressure and temperature?
- How likely is a week of warm, dry weather in the summer months at the UK locations?
- Which of the locations have the most variable weather?

Appendix 3: Using Desmos for distributions

Binomial distribution

- Select: functions > Dist > binomialdist
- Enter the number of trials and probability of success.
- To calculate the probability of a range select CDF and set the limits.



Normal distribution

- Select: functions > Dist > normaldist
- Enter the mean and standard deviation.
- To calculate the probability of a range select CDF and set the limits.