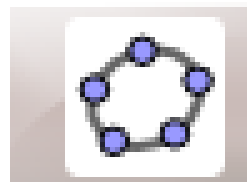
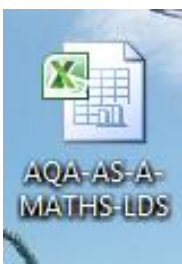




Using Large Data Sets Workbook Version C (AQA)



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 time series in Excel	Page 9
Drawing charts	Page 10
Drawing Scatterplots and curves of best fit	Page 11
Drawing charts side by side for comparison	Page 13
Testing for goodness of fit to a normal distribution	Page 14
Carrying out Hypothesis Tests	Page 15
Generating Random Samples	Page 16

Large Data Sets (AQA) 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 AQA dataset and to refer to the government website that the data is drawn from: <https://www.gov.uk/government/collections/family-food-statistics>

Key Skills

- Understand the dataset and its context
- Cleanse a dataset and know how to deal with outliers
- Sort and Filter the dataset where appropriate
- Produce a table of summary statistics
- Draw frequency charts, box plots, trend lines and 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 AQA-AS-A-MATHS-LDS excel file which contains the dataset and refer to <https://www.gov.uk/government/collections/family-food-statistics> . There are 10 sheets of food data for 9 different regions and for England as a whole. Students are required to understand the context of the data so that it is important that they read the glossary of terms whilst looking through the dataset. Some questions you might like to consider are:

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

Who collected it and how was it collected?

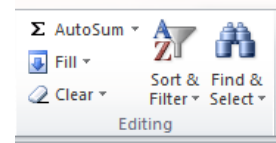
What are the units? Is there any missing data?

What does RSE mean?

Students should be encouraged to research further on the website so that they fully understand the context of the data.

2. Sorting and filtering the dataset

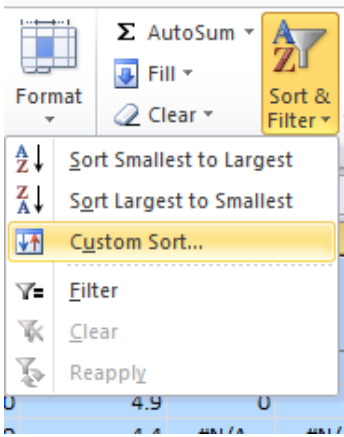
Further familiarity with the dataset can be gained by sorting and filtering the data within Excel. This can help identify any possible outliers or rogue values.



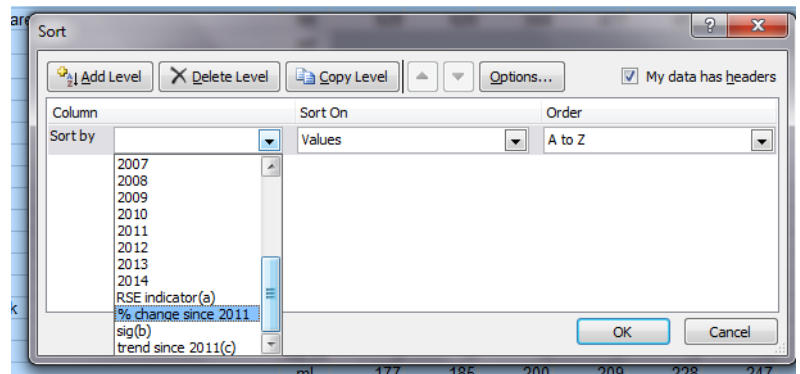
These functions can be found at the far end of the top toolbar:

Suppose we want to sort the data for the whole of England (last sheet) according to the % change since 2011. To make it easier to sort, first delete the first 7 rows so that now the first row contains the field headings. 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, say smallest to largest. Also make sure that the 'My data has headers' box is checked otherwise your column headings will get sorted as well.

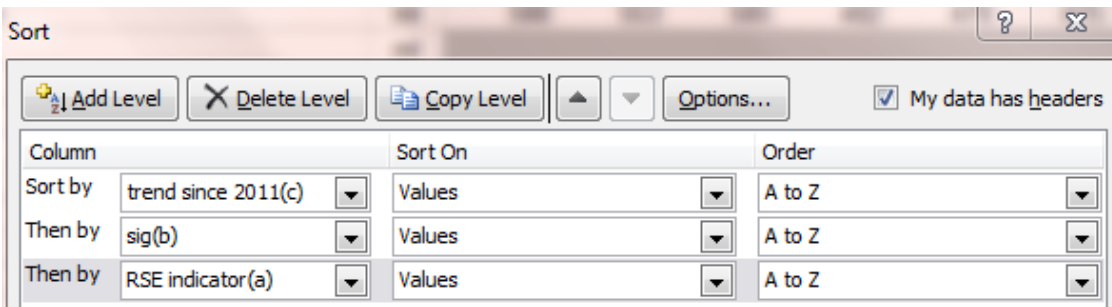


The data is now sorted in order of % change since 2011 with the largest decreases listed first (some of the earlier columns have been hidden).

Description	Units	2011	2012	2013	2014	RSE indicator ^(a)	% change since 2011	sig ^(b)	trend since 2011 ^(c)
Meals on wheels - items not specified	g	0	1	0	0	x	-88		
White bread, premium, sliced and unsliced	g	2	1	1	0	x	-84	yes	↘
Mints	g	3	2	2	2	✓	-51	yes	↘
Starch reduced bread and rolls	g	9	7	6	5		-43	yes	↘
Brown bread, sliced and unsliced	g	39	36	31	25	✓	-37	yes	↘
Dried milk products	ml	3	4	2	2	x	-36		
Soft drinks, concentrated, not low calorie^(h)	ml	381	340	315	255	✓	-33	yes	↘
Syrup, treacle	g	5	4	4	3	✓	-30	yes	↘
UHT milk	ml	5	4	3	3	x	-30		
Other margarine	g	2	3	3	2		-28		
Chewing gum	g	2	2	1	1	✓	-27	yes	↘
Other takeaway food brought home	g	0	0	0	0	x	-25		
Pasteurised/ homogenised	ml	348	288	276	260	✓✓	-25	yes	↘
Liquid wholemilk, including school and welfare	ml	356	293	280	267	✓✓	-25	yes	↘
Liquid wholemilk, full price	ml	356	293	280	267	✓✓	-25	yes	↘
Marmalade	g	8	8	6	6	✓	-25	yes	↘

The picture is a little confusing as some of the rows represent broad categories (in bold) and some sub-categories, so this needs to be borne in mind when making any conclusions. Which foods have shown the biggest changes? Are these genuine decreases or increases? Or has the way the data is classified changed?

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



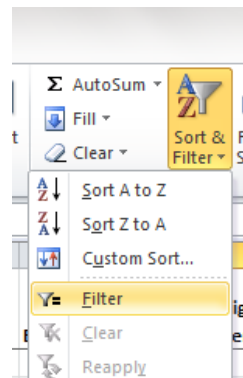
Try the above sort (remember to select all the data first using Ctrl A). Now the first rows are those food items which whose linear trend is statistically significant upwards, the change is greater than twice the standard error and the RSE has one or two ticks.

Description	Units	RSE indicator ^(a)	% change since 2011	sig ^(b)	trend since 2011 ^(c)
Mineral or spring waters	ml	✓	+36	yes	↗
Other food and drink	g	✓✓	+9	yes	↗
Soft drinks, concentrated, low calorie ^(h)	ml	✓✓	+23	yes	↗
Chocolate bars - solid	g	✓✓	+12	yes	↗

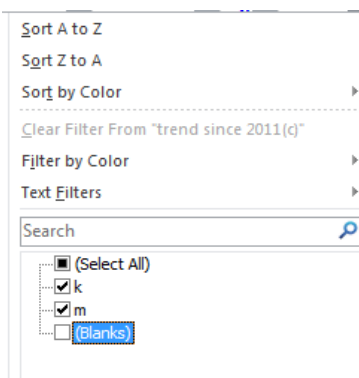
Ignoring 'Other food and drink' which is a broad heading, there are 3 food items: Mineral Water, Diet Drinks and Chocolate that have significant linear growth in consumption. These perhaps warrant further investigation.

Another way of focusing on part of the dataset is to use a filter:

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



Click on the arrow next to **sig** and then scroll down and uncheck the box next to blanks.



We now just have a list of those food items with a significant linear trend:

	Units	RSE indicator	% change since 2011	sig	trend since 2011 ^(c)
Liquid wholemilk, including school and welfare	ml	✓✓	-25	yes	↘
Liquid wholemilk, full price	ml	✓✓	-25	yes	↘
Pasteurised/ homogenised	ml	✓✓	-25	yes	↘
Fromage frais	ml	✓	-21	yes	↘
Dairy desserts - not frozen	ml	✓✓	-10	yes	↘
Non-dairy milk substitutes ^(e)	ml	✓	+59		↗

(Some columns are hidden here. To turn the filters off click on the filter button again)

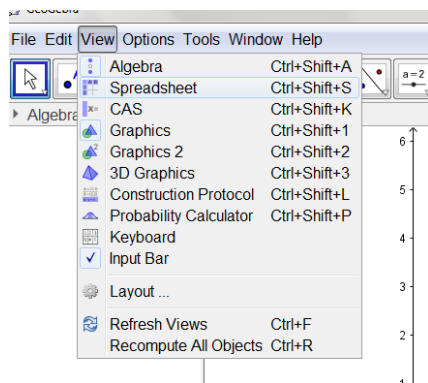
Exercise: Sort the whole England data into solids and liquids. Then sort by consumption in 2001-2 and filter out the bold items (the headings). After bread, what was the next most consumed foodstuff?

3. Producing a table of summary statistics in Geogebra

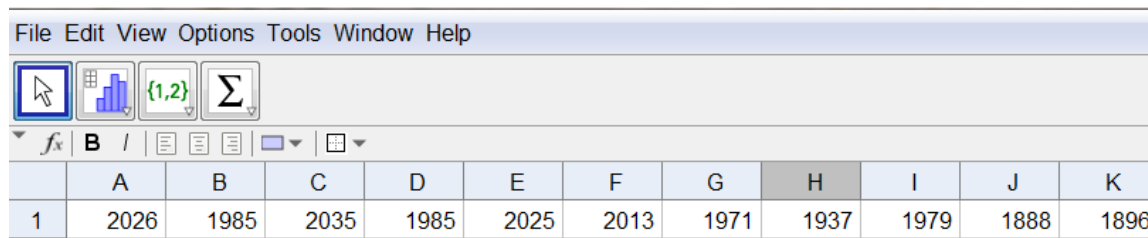
Select the England sheet of dataset and highlight the row corresponding to “Milk and milk products excluding cheese” and copy it (Ctrl C)

Description	Units	2001-02	2002-03	2003-04	2004-05	2005-06	2006	2007	2008	2009	2010	2011	2012	2013	2014
Milk and milk products excluding cheese	ml	2,026	1,985	2,035	1,985	2,025	2,013	1,971	1,937	1,979	1,888	1,896	1,885	1,843	1,845
Liquid wholemilk, including school and welfare	ml	588	553	585	482	470	475	420	412	411	346	356	293	280	267

Open GeoGebra in the spreadsheet view :

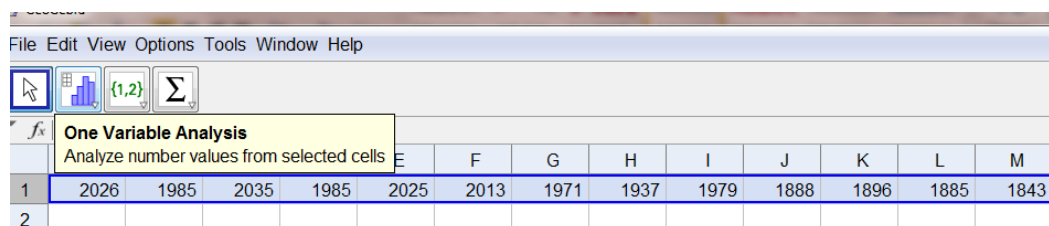


Select cell A1 and then paste (Ctrl V) the data into the first row of the GeoGebra spreadsheet:

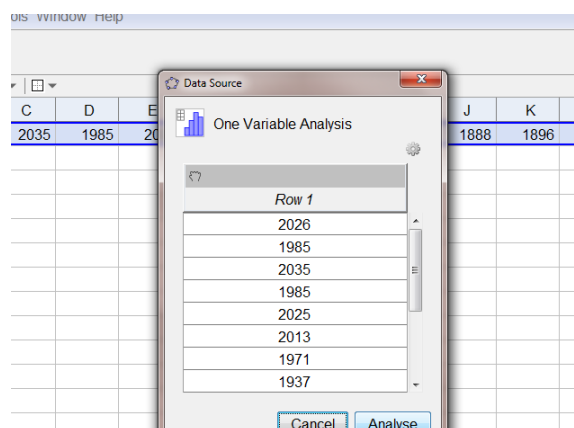


	A	B	C	D	E	F	G	H	I	J	K
1	2026	1985	2035	1985	2025	2013	1971	1937	1979	1888	1896

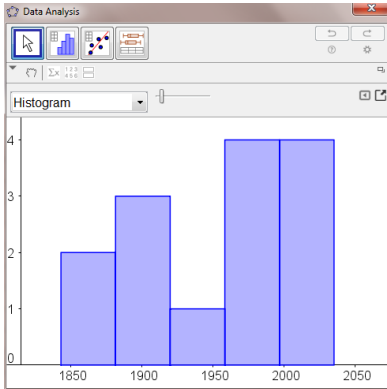
Highlight this row and then click on one variable analysis.



Confirm that you want to analyse this data:

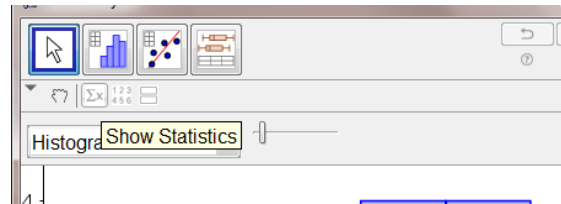


And a new dialogue box will appear:



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	14
Mean	1950.9286
σ	65.4342
s	67.9043
Σx	27313
Σx^2	53345655
Min	1843
Q1	1888
Median	1975
Q3	2013
Max	2035

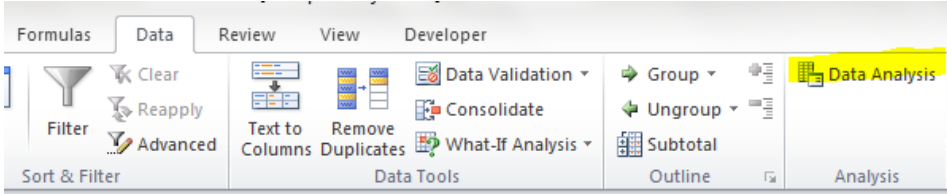
You might need to enlarge the window to see all the digits.

Exercise: Produce the Statistics Box for all Bread products, which is illustrated to the right.

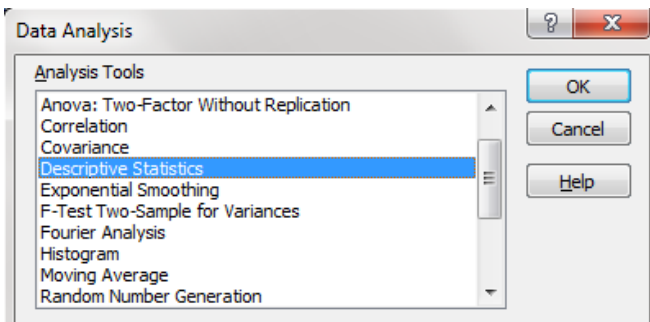
Statistics	
n	14
Mean	655.6429
σ	56.2019
s	58.3235
Σx	9179
Σx^2	6062367
Min	549
Q1	609
Median	654.5
Q3	688
Max	750

4. Producing a table of summary statistics in Excel

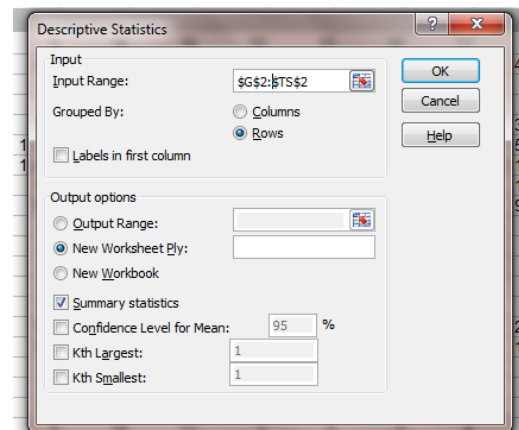
It is possible to produce a Statistics box in Excel but earlier versions require the data analysis add-in which has to be selected first. To select the add-in go to File>Options>Add-ins and select the Analysis Toolpak. Once the add-in is selected it will appear when the data tab is selected:



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



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



Then the statistics box will appear in a new sheet:

Mean	1950.99263
Standard Error	18.1622475
Median	1975.05082
Mode	#N/A
Standard Deviation	67.9569075
Sample Variance	4618.14127
Kurtosis	-1.3301482
Skewness	-0.3949288
Range	192.203216
Minimum	1843.0815
Maximum	2035.28472
Sum	27313.8968
Count	14

The Summary statistics box also needs to be checked.

5. Drawing time series in Excel

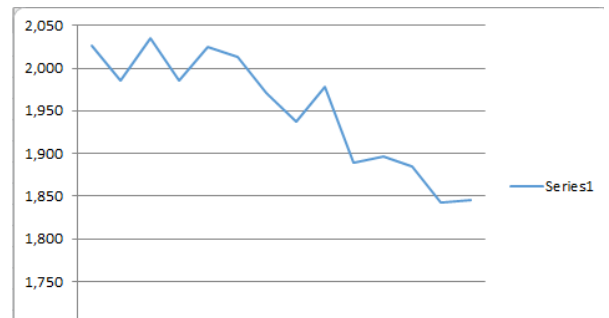
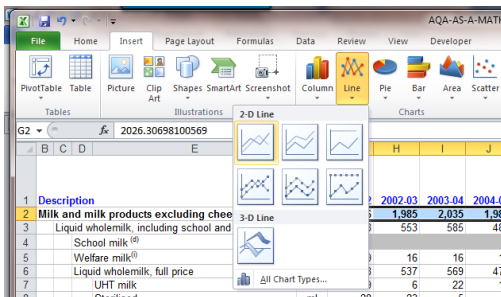
The AQA datasets are time series as they show how consumption changes over time. Excel can be used to plot these time series.

Consider the total milk consumption for all of England.

Highlight the data:

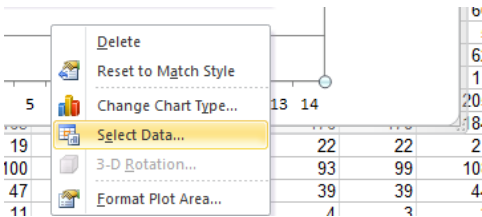
Units	2001-02	2002-03	2003-04	2004-05	2005-06	2006	2007	2008	2009	2010	2011	2012	2013	2014
ml	2,026	1,985	2,035	1,985	2,025	2,013	1,971	1,937	1,979	1,888	1,896	1,885	1,843	1,845
ml	588	553	585	482	470	475	420	412	411	346	356	293	280	267

Now click on Insert > Line > 2D-Line

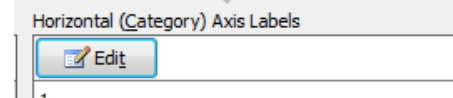


and select the first option. A basic graph will appear.

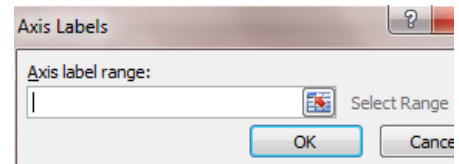
Right click on the graph so that we can assign the year to the horizontal axis:



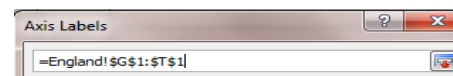
Choose 'select data'.



and edit the horizontal axis:

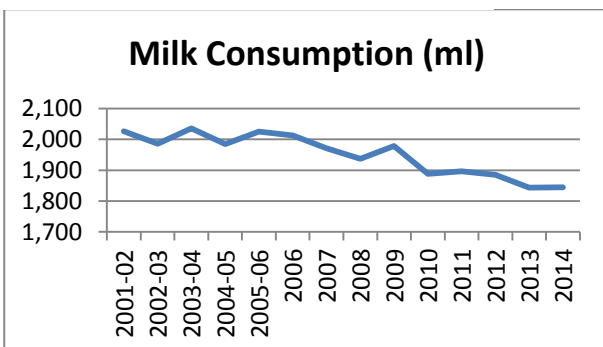
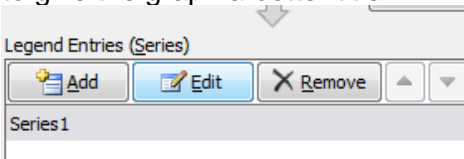


Select the row with the dates in.



Then click on OK to accept the dates.

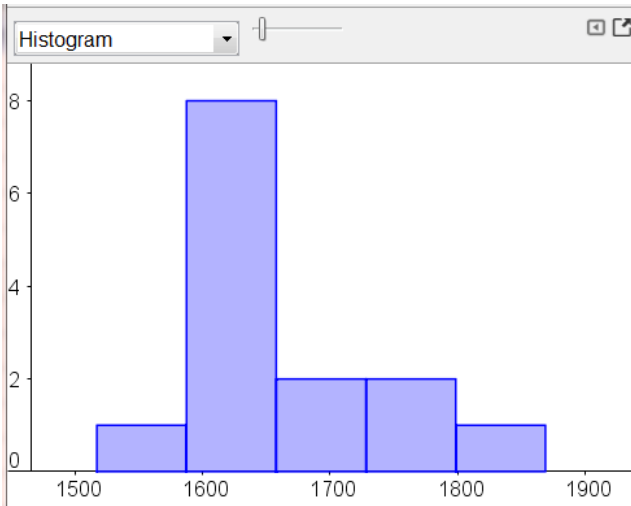
The legend edit option can be used to give the graph a better title



6. Drawing charts for a set of data

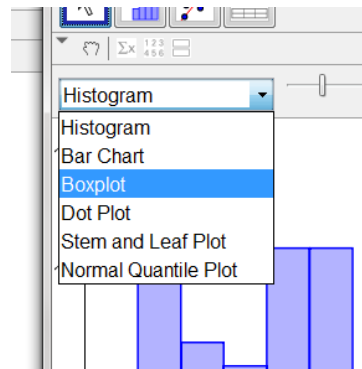
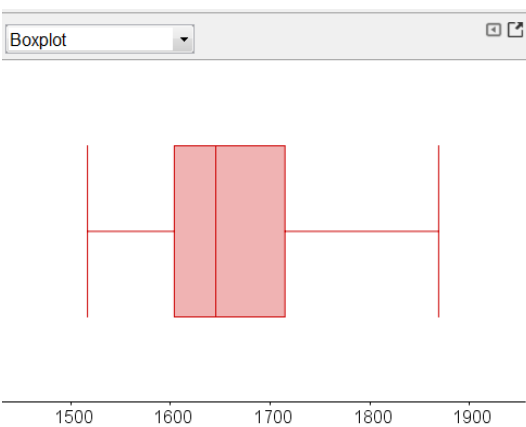
GeoGebra can display a range of graphs and charts. Using the soft drink data for all England, follow the previous steps for copying the data into the spreadsheet view and select one variable analysis again.

The default view is Histogram:

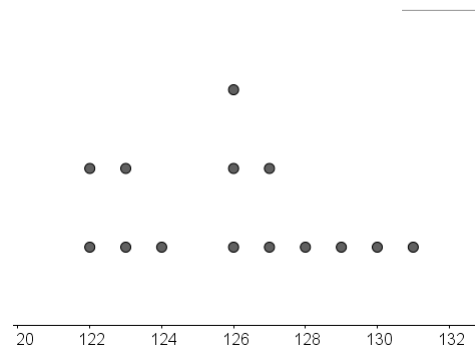
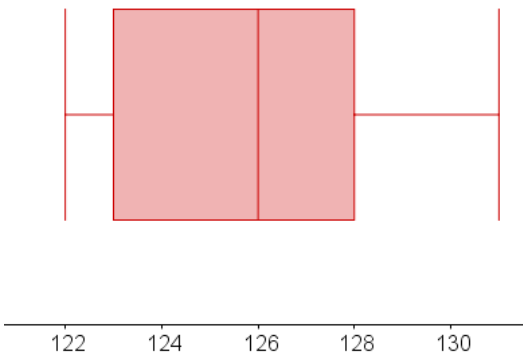


The slider can be used to alter the number of classes and it is interesting to note how the representation changes. This is not really a Histogram as it is that is frequency plotted on the vertical axis, rather than frequency density. All the classes are of equal length.

Different charts can be obtained by changing the option:



Exercise: Produce the diagrams below for confectionary consumption. Are these useful representations of this data?



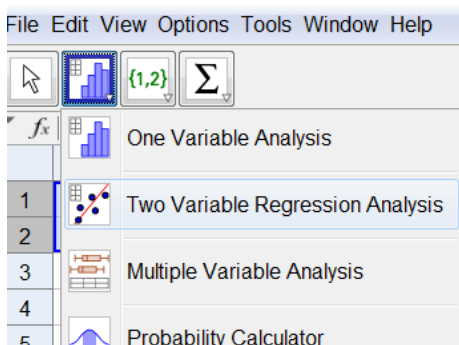
7. Drawing Scatterplots and curves of best fit

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.

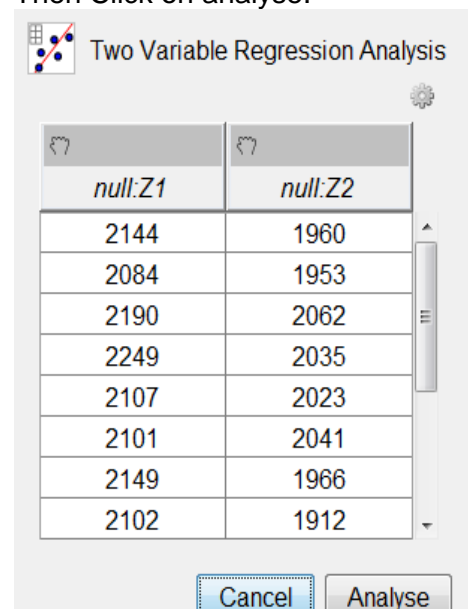
For example we might see if there is a relationship between milk consumption in the East Midlands and milk consumption in the South East.

	A	B	C	D	E	F	G	H	I
1	2144	2084	2190	2249	2107	2101	2149	2102	2250
2	1960	1953	2062	2035	2023	2041	1966	1912	1875

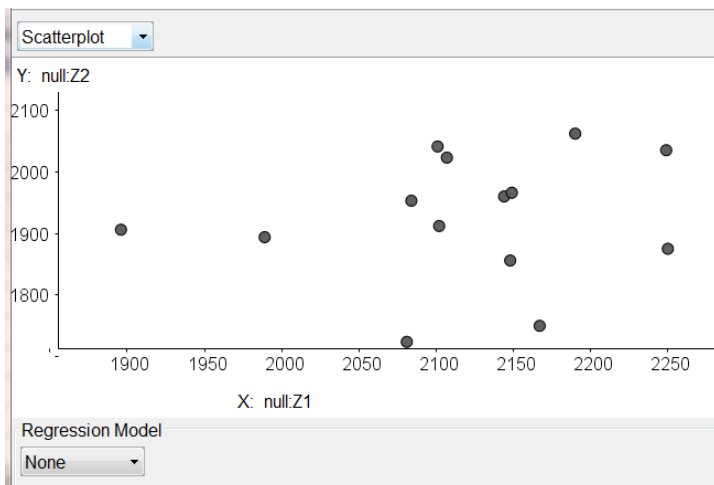
This time we need to highlight both rows and select two variable regression analysis:



Then Click on analyse:



And a Scatterplot is drawn:

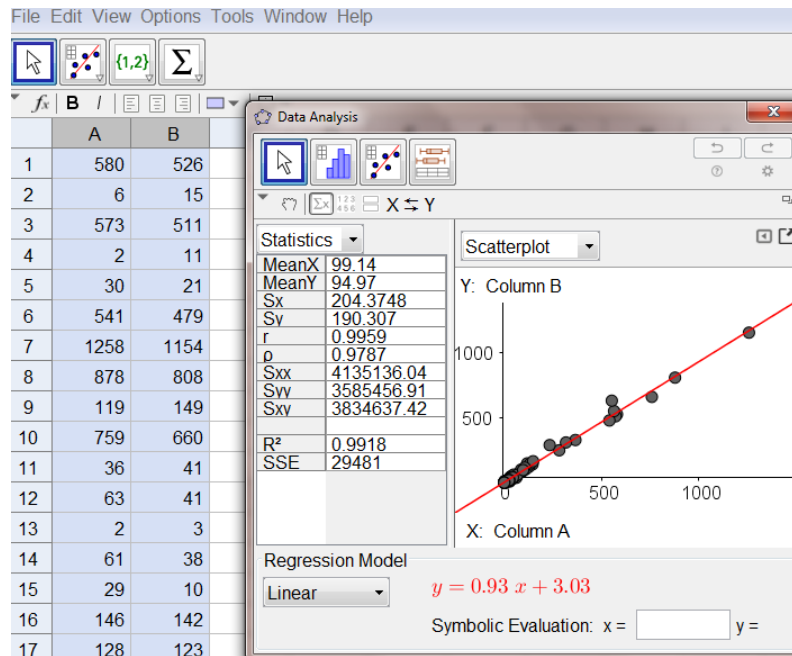


The Scatterplot shows little correlation between the two variables which can be confirmed by using the Statistics box $\sum x$:

Statistics	
MeanX	2118.3571
MeanY	1925.5
Sx	93.5707
Sy	102.8043
r	0.1685
p	0.2132
Sxx	113821.2143

Alternatively we might want to select two columns of data to compare. For example we might want to compare overall food consumption in two different years, say for London in 2001-2 and 2002-3. However before pasting into Geogebra delete the Bold items as these are totals (and we would be repeating the same data).

Units	2001-02	2002-03
ml	1,838	1,680
ml	580	526
ml		
ml	6	15
ml	573	511
ml	2	11
ml	30	21
ml	541	479
ml	1,258	1,154
ml	878	808
ml	119	149
ml	759	660
eq ml	36	41
eq ml	63	41
eq ml	2	3
eq ml	61	38



GeoGebra provides a selection of different types of regression models for this data. It will sometimes suggest one to start with (if the correlation is good enough). Try exploring the different models for this data.

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 food consumption in the South West in 2001-2 (x) and Yorkshire and Humber in 2001-2 (y). Suggest a y on x regression model for this data. Comment on the validity of your model.

8. Drawing Graphs side by side for comparison

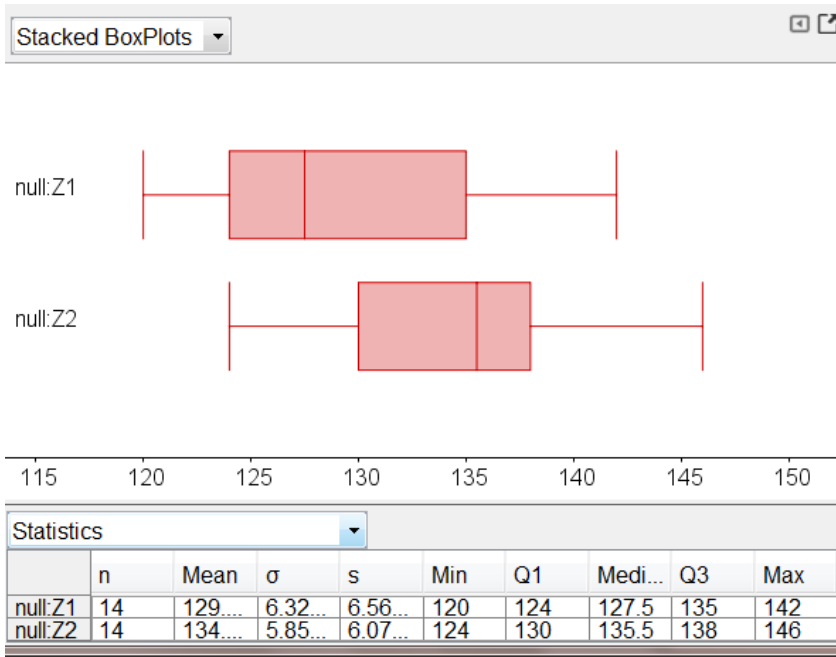
Let's compare cheese consumption in the SE to that in the SW. Copy and paste this data into Geogebra.

	A	B	C	D	E	F	G	H	I	J
	129	124	120	126	126	133	142	123	130	135
	126	129	134	130	137	136	132	136	146	143



Highlight both rows 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, the top one being the SE and the bottom SW?

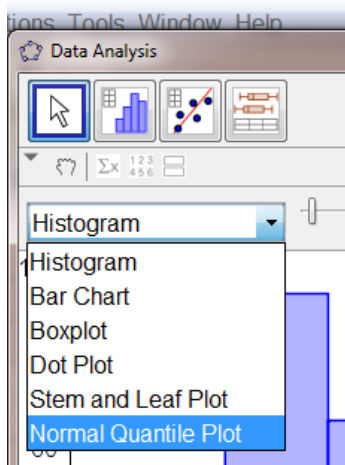
9. Testing for goodness of fit to a normal distribution using a quantile plot

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 a sample from a normal distribution.

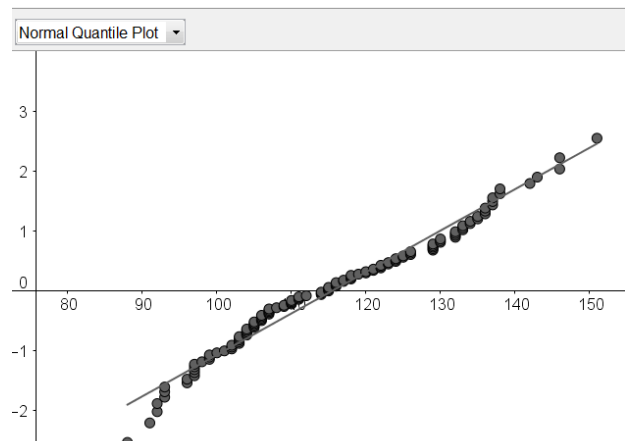
In order to get a good sample size, we copy and paste all the cheese consumption for all 9 regions in all years into GeoGebra:

A	B	C	D	E	F	G	H	I	J	K	L	M	N
97	97	96	114	96	104	106	99	97	98	97	103	91	88
106	104	118	106	106	107	115	107	103	105	116	107	111	116
99	106	110	97	110	107	110	103	109	111	115	108	103	103
132	121	120	132	125	123	118	116	134	130	129	119	146	116
104	114	112	101	109	121	115	100	114	104	110	117	107	102
124	118	133	117	151	122	122	118	129	137	132	129	133	115
104	105	93	92	93	105	105	92	93	102	105	102	111	99
129	124	120	126	126	133	142	123	130	135	136	122	137	125
126	129	134	130	137	136	132	136	146	143	138	135	124	138

Highlight all the data 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 but there are a few notable deviations, particularly at lower end. Looking at the histogram or box plot shows that the data is quite skewed as well.

Where do the data items that deviate the most come from?

Are they all from the same region?

Are they all from the same year?

10. 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.

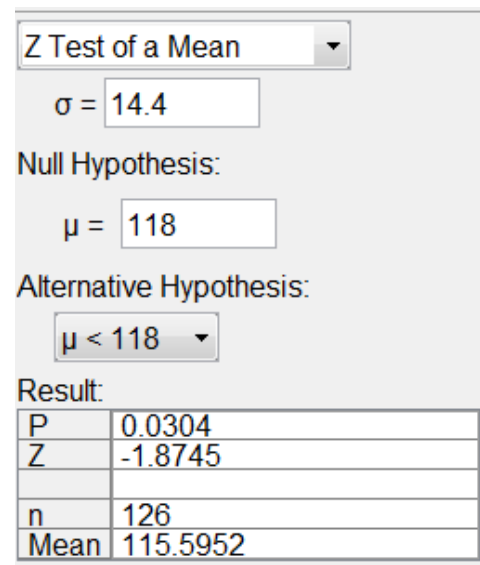
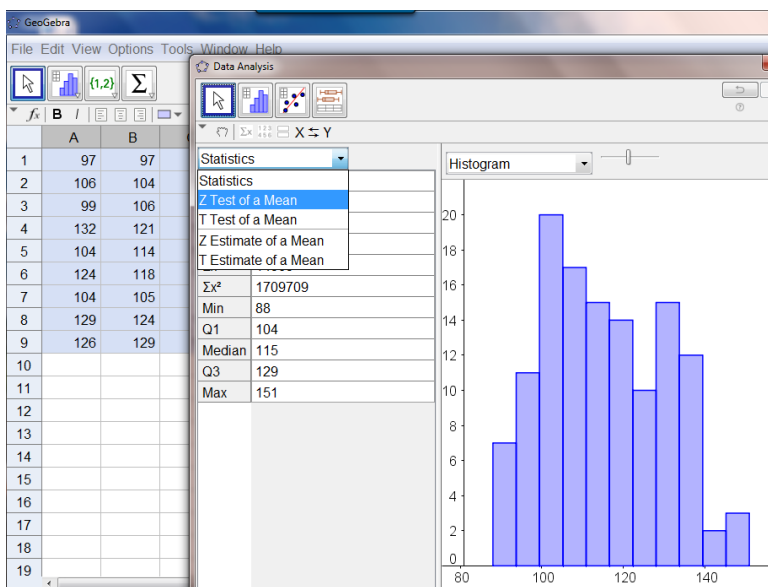
But it is straightforward to use GeoGebra to do this:

e.g. to test, at the 5% level,

$$H_0: \mu = 118 \quad \text{against} \quad H_1: \mu < 118$$

where μ is the mean cheese consumption from 2001-2014 in g/week and it is given that the population standard deviation is known and $\sigma = 14.4$

Using the pooled data from the last section, within the statistical box menu there is an option for a Z-test of a mean:



Select this and then enter the test parameters.

This shows that the probability that $\bar{X} = 115.60$ under H_0 is 3.0% and so I would reject the null hypothesis concluding that the population mean is in fact lower than 118.

A few health warnings here. First we have assumed the central limit theorem applies as n is large (126) 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 only a slight difference here giving $p = 0.032$.

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.

11. 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 .

To use this feature the data needs to be organised in a columns, so first organise the cheese data from the last section into a single column of 126 data items. Then insert a new column into column A and in the first blank cell in column A below the headers 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 the values 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
1	0.008023	118
2	0.009987	119
3	0.011798	103
4	0.013082	103
5	0.023339	114
6	0.043328	110
7	0.058671	123
8	0.066356	105
9	0.075456	97
10	0.076446	92
11	0.07761	135
12	0.086098	118
13	0.092033	106
14	0.099956	129
15	0.114579	123
16	0.1161	100
17	0.116142	103
18	0.122315	92
19	0.122812	107
20	0.126298	105

Exercise: Take a sample of size 40 from the cheese data and copy the data into GeoGebra. Perform the hypothesis test outlined in section 10 on this sample of 40 items. What is your result?