Using Large Data Sets Workbook – OCR 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 OCR Dataset which can be downloaded at *https://ocr.org.uk/qualifications/as-and-a-level/mathematics-a-h230-h240-from-2017/assessment/#as-level*

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 "OCR pre-release data 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 is a census, how is it conducted and how often?
- What does phrase "*dates of birth that imply an age over 110 are treated as invalid and the person's age is imputed*" mean?

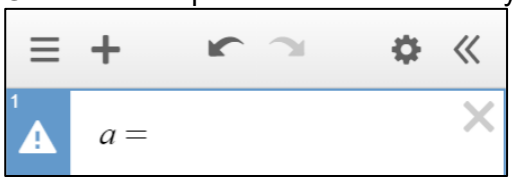
Students need to understand each of the fields and how they are determined. Some of them warrant further discussion. Students should be encouraged to research further so that they fully understand the concepts. The *Office for National Statistics* website can be used to find out more about the Census process.

1 Producing summary statistics

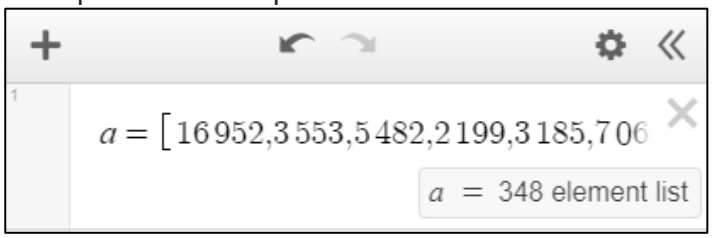
Select the 3rd sheet (Method of travel by LA 2001) and highlight column E (Work mainly at or from home) and copy it using Ctrl-C.

	A	B	C	D	E	F
1	geography code	Region	local authority: district / unitary	All Categories of people in employment	Work mainly at or from home	Underground, metro, light rail, tram
2	E06000047	North East	County Durham	203,614	16,952	190
3	E06000005	North East	Darlington	42,993	3,553	35
4	E08000020	North East	Gateshead	78,786	5,482	3,672
5	E06000001	North East	Hartlepool	33,762	2,199	44
6	E06000002	North East	Middlesbrough	49,317	3,185	46
7	E08000021	North East	Newcastle upon Tyne	101,498	7,066	5,591
8	E08000022	North East	North Tyneside	83,698	6,053	6,979
9	E06000048	North East	Northumberland	136,083	14,687	715
10	E06000003	North East	Redcar and Cleveland	54,295	3,898	48
11	E08000023	North East	South Tyneside	58,899	3,743	3,983
12	E06000004	North East	Stockton-on-Tees	75,904	5,449	61

Go to a new input bar in Desmos and type:

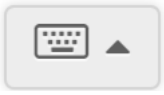


Then press Ctrl-V to paste the data:

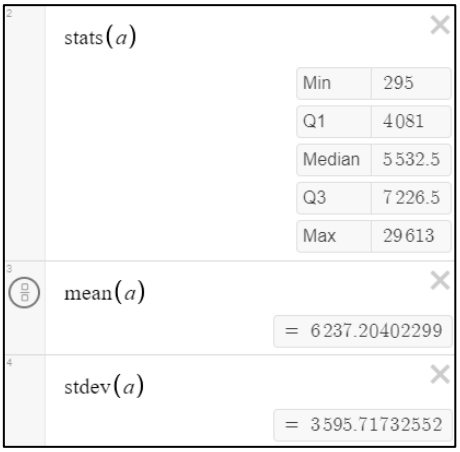


This should create the list (called a). You can now refer to the list in other commands.

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



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



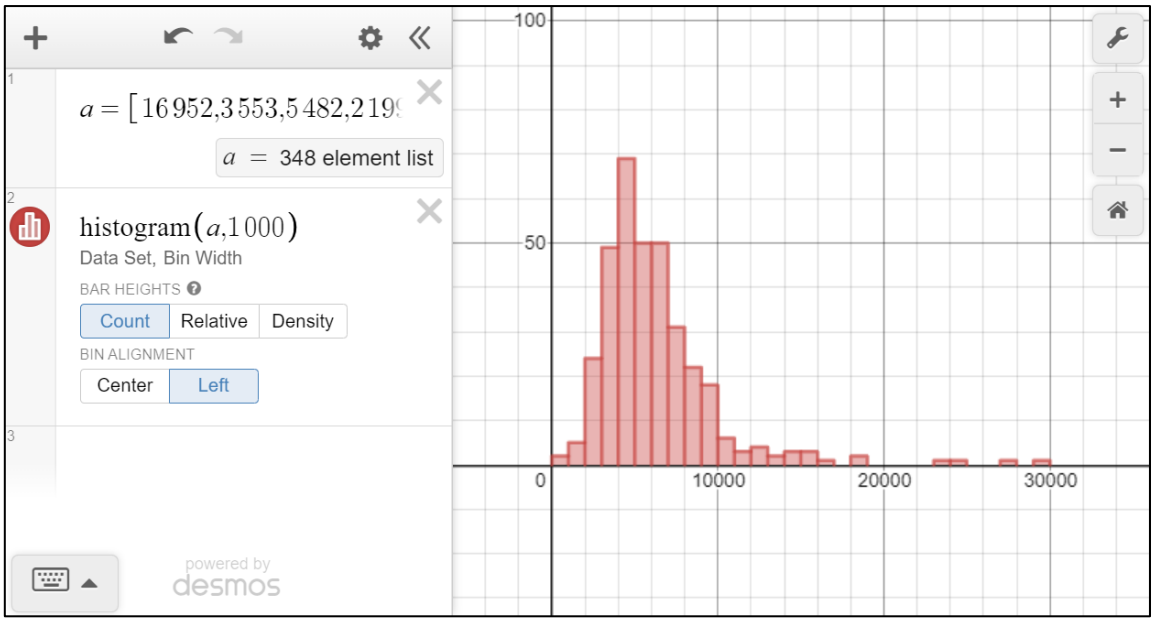
Exercise: Compare the number of people working from home in 2001 and 2011. How has the number changed? Are there any regional variations?

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.

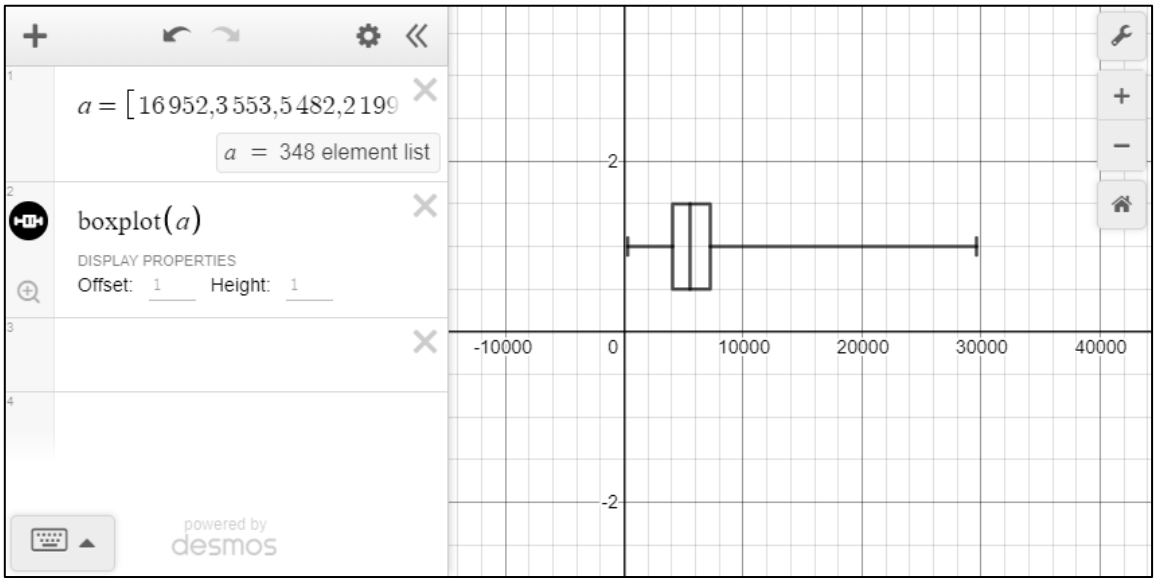
For a histogram you should enter the data set and the bin width.



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 5 the first bin will contain values of x where $0 \leq x < 5$.

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.

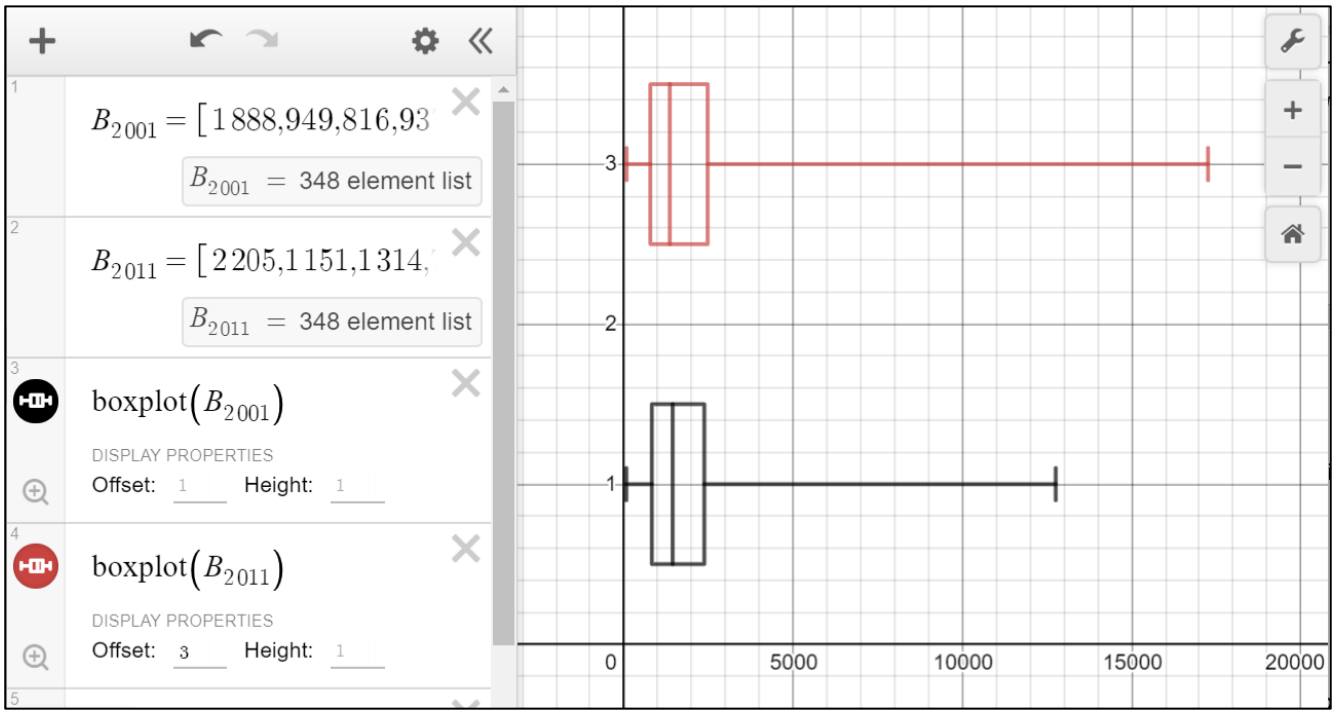


What information does each of these charts show you about the dataset? When might you use a histogram and when might you use a boxplot?

3 Drawing graphs side by side for comparison

The following example compares the number of people who use Bicycle as their main mode of transport in 2001 and 2011.

Each set of numbers will need to be copied as a new list into Desmos. (NB just type B2001 to obtain the variable name with the subscript). Use the *offset* to move the second boxplot.



Exercise: Draw the boxplots to compare some other modes of transport for 2001 and 2011. What differences can be observed?

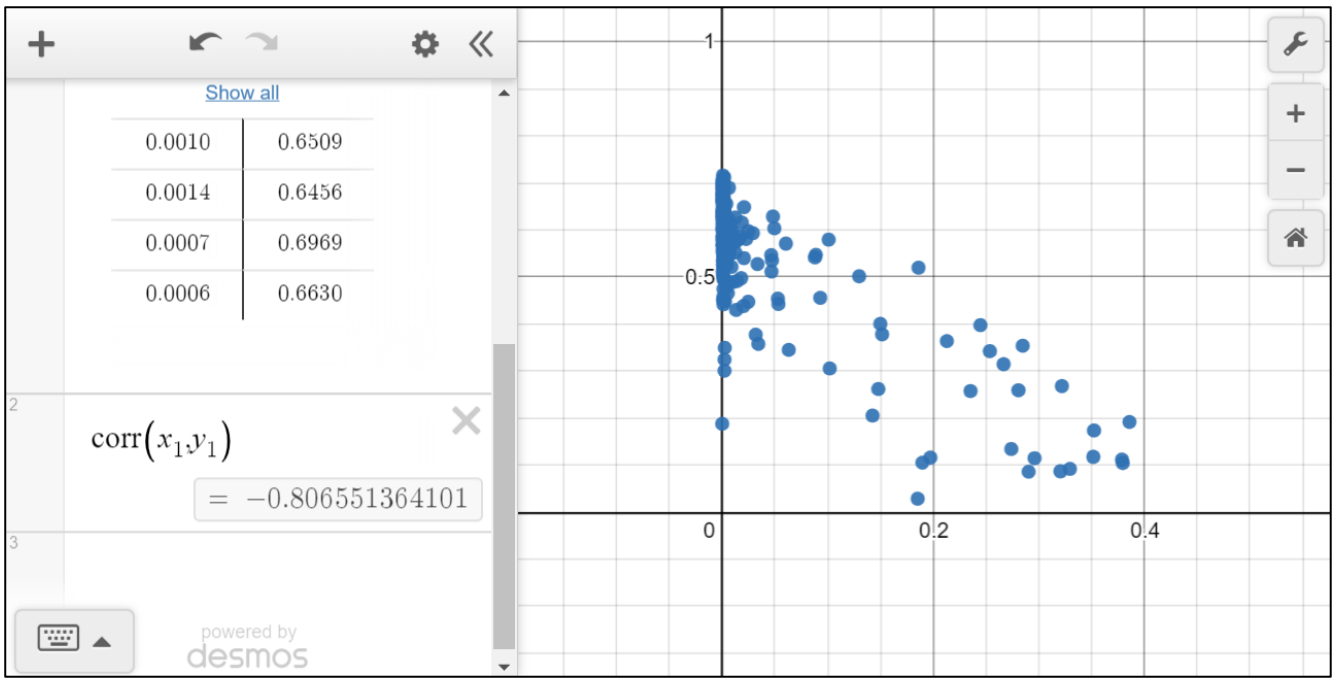
4 Drawing Scatterplots

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. For many of the columns in this dataset it will be more useful to compare the proportions of people using a particular mode of transport as opposed to the raw numbers.

In the following example the proportion of people travelling to work using underground, metro, light rail, tram will be compared to the proportion who travel to work as the driver of a car for 2011. These two columns have been calculated and then copied into a new blank spreadsheet. When copying into a new spreadsheet you should use the Paste Values option from the Paste menu. The pair of adjacent columns in the new spreadsheet can then be copied and pasted into Desmos.

	D	E	F	G	H
1	**All Categories of people in employment**	**Work mainly at or from home**	**Underground, metro, light rail, tram**	**Underground, metro, light rail, tram proportion**	**Train**
2	227,894	20,652	323	0.0014	1,865
3	49,014	4,180	33	0.0007	828
4	91,877	6,383	4,270	0.0465	705
5	37,767	2,473	33	0.0009	469
6	54,547	3,337	44	0.0008	698
7	119,335	8,430	6,304	0.0528	1,381
8	96,026	6,997	8,523	0.0888	1,385
9	146,901	17,894	659	0.0045	1,852

	A	B
1	Underground, metro, light rail, tram proportion	Driving a car or van propotion
2	0.0014	0.6435
3	0.0007	0.5913
4	0.0465	0.5468
5	0.0009	0.6054
6	0.0008	0.5712
7	0.0528	0.4539
8	0.0888	0.5472
9	0.0045	0.6165



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.

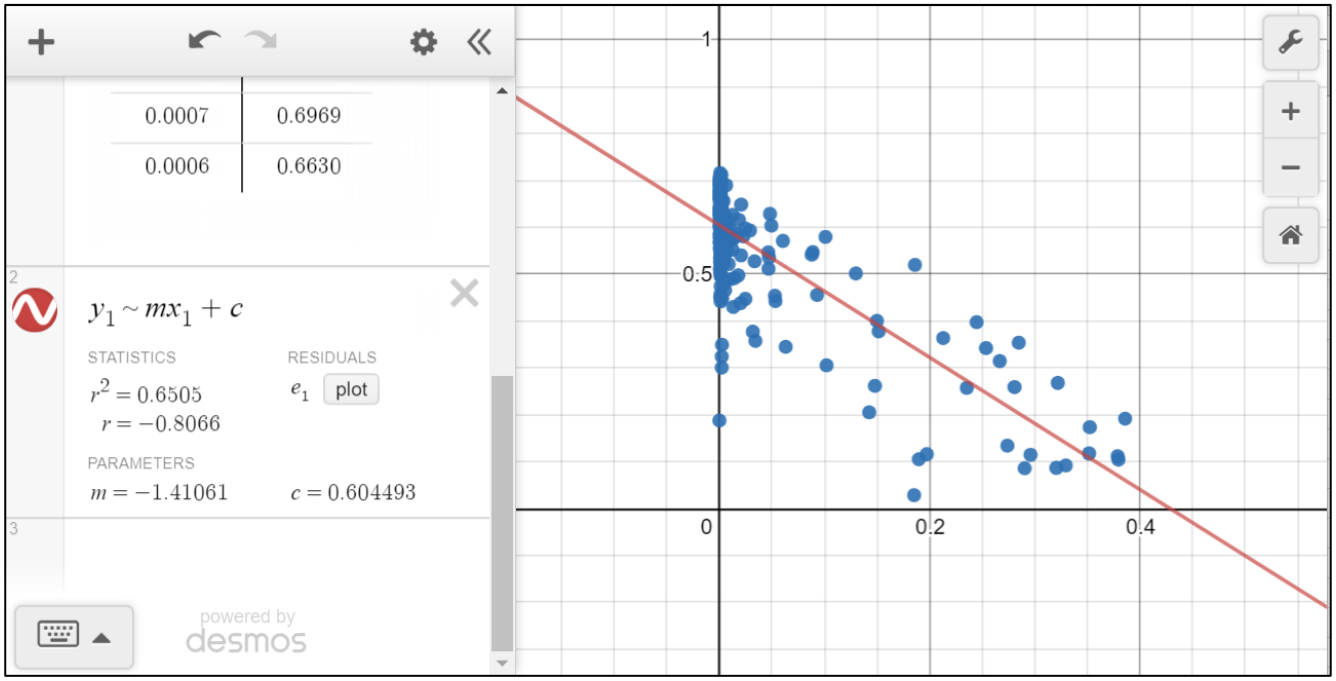
5 Regression models

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 proportion of people travelling to work using underground, metro, light rail, tram compared to the proportion who travel to work as the driver of a car for 2011.



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: Explore the correlation between other modes of transport for 2011. Which show positive and which show negative correlation? What might be the reasons for this?

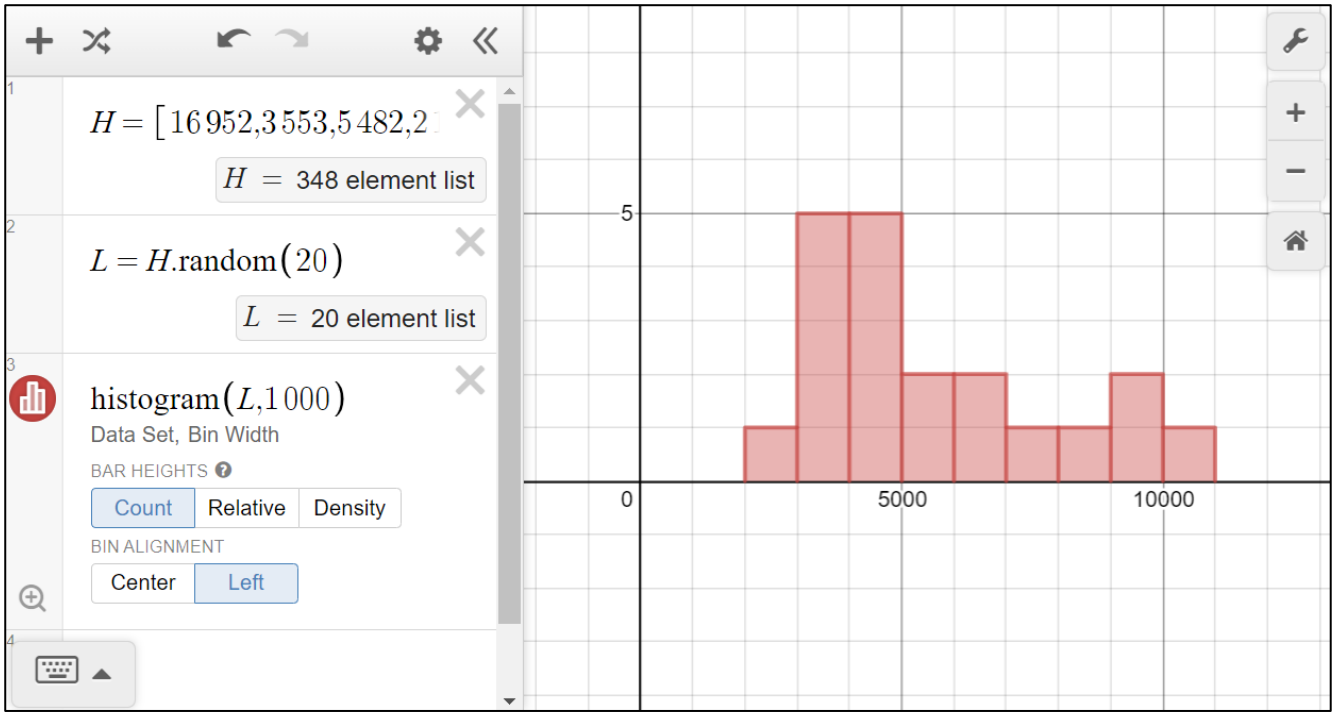
6 Random Sampling

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 the number of people who worked from home in 2001.

Enter a new variable of H and copy the full list into Desmos.

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

A histogram for this sample can be plotted using $\text{histogram}(L, 1000)$.



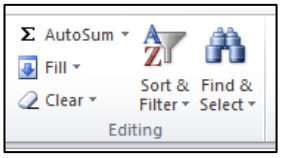
To generate different samples click on  (second icon top-left).

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

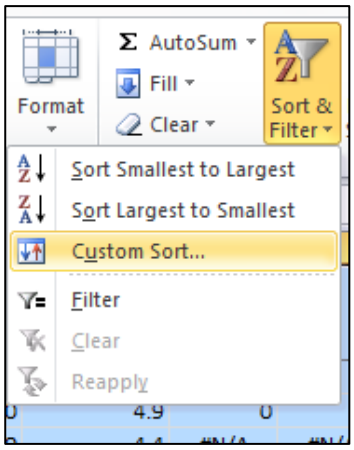
Further familiarity with the dataset can be gained by sorting and filtering the data within Excel. This can help identify any possible outliers or rogue values.

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

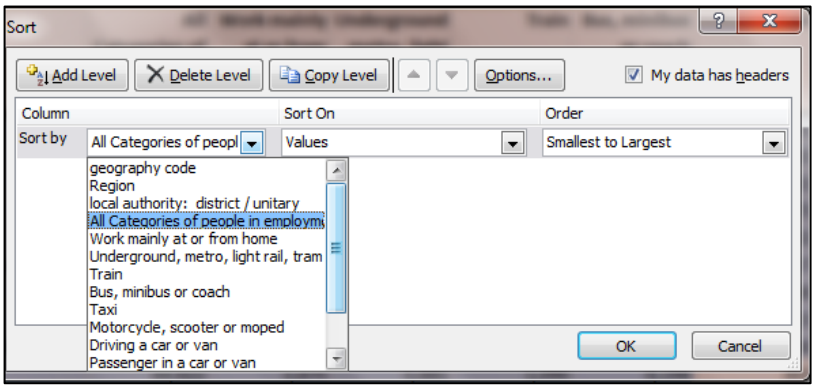


Suppose you want to sort the Method of Travel 2011 data (which is in the second sheet) according to number of people in employment. Use Crtl-A to select all the data (NB as column P is blank you might need to delete this before pressing Crtl-A to ensure all the data is selected).

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 number in employment.

B	C	D	E	F
Region	**local authority: district / unitary**	**All Categories of people in employment**	**Work mainly at or from home**	**Underground, metro, light rail, tram**
South West	**Isles of Scilly**	1,311	428	0
London	**City of London**	4,747	687	879
South West	**West Somerset**	15,355	3,885	17
East Midlands	**Rutland**	18,037	3,011	29
South West	**Christchurch**	20,301	2,530	19
South West	**Purbeck**	21,419	3,125	27
Wales	**Merthyr Tydfil**	25,099	1,680	24
South West	**West Devon**	25,241	5,245	31
Yorkshire and The Humber	**Ryedale**	25,504	5,114	24
East Midlands	**Melton**	26,184	3,489	17
East Midlands	**Oadby and Wigston**	26,399	2,187	43
North West	**Eden**	27,461	5,771	20
Yorkshire and The Humber	**Craven**	27,600	4,595	25
Yorkshire and The Humber	**Richmondshire**	27,795	4,910	29

For many of the columns in the dataset it is often more useful to compare the proportions of people in a region that use a particular mode of transport as opposed to using the raw number. To calculate this insert a new column and use a formula to divide the number by total of all people in employment.

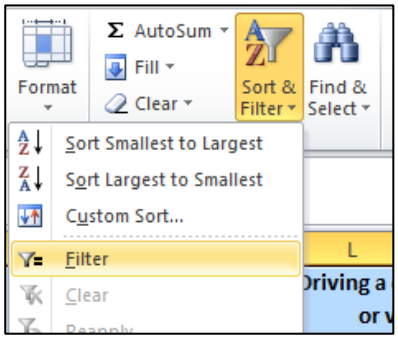
For example to calculate the proportion of people who travel by *Underground, metro, light rail, tram* for 2011 a new column G has been inserted and the formula =F2/E2 has been entered in cell G2. This can then be filled-down to complete the column and formatted to show 4 decimal places.

	D	E	F	G
1	All Categories of people in employment	Work mainly at or from home	Underground, metro, light rail, tram	Underground, metro, light rail, tram proportion
2	227,894	20,652	323	=F2/D2
3	49,014	4,180	33	
4	91,877	6,383	4,270	
5	37,767	2,473	33	
6	54,547	3,337	44	
7	119,335	8,430	6,304	
8	96,026	6,997	8,523	
9	146,901	17,894	659	

	D	E	F	G
1	All Categories of people in employment	Work mainly at or from home	Underground, metro, light rail, tram	Underground, metro, light rail, tram proportion
2	227,894	20,652	323	0.0014
3	49,014	4,180	33	0.0007
4	91,877	6,383	4,270	0.0465
5	37,767	2,473	33	0.0009
6	54,547	3,337	44	0.0008
7	119,335	8,430	6,304	0.0528
8	96,026	6,997	8,523	0.0888
9	146,901	17,894	659	0.0045

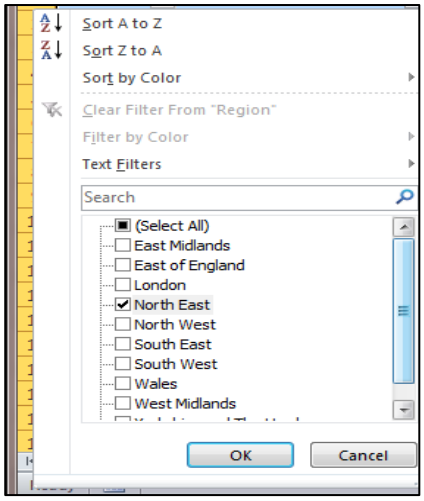
Why do some areas have such low employment numbers? Are they just small regions? Are their populations older and retired? Is unemployment high?

You can also use a filter to view the results just for one region. Click on filter and an arrow should appear next to each heading:



	A	B	C	D
1	geography code	Region	local authority: district / unitary	All Categories of people in employment
2	E06000053	South West	Isles of Scilly	1,311
3	E09000001	London	City of London	4,747
4	E07000191	South West	West Somerset	15,355
5	E06000017	East Midlands	Rutland	18,037
6	E07000048	South West	Christchurch	20,301

Click on the arrow next to Region and then scroll down and select the box next to North East only:



Now you should just see the data for the North East:

geography code	Region	local authority: district / unitary	All Categories of people in employment
E06000001	North East	Hartlepool	37,767
E06000005	North East	Darlington	49,014
E06000002	North East	Middlesbrough	54,547
E06000003	North East	Redcar and Cleveland	56,354
E08000023	North East	South Tyneside	64,622
E06000004	North East	Stockton-on-Tees	87,122
E08000020	North East	Gateshead	91,877

To turn the filters off click on the filter button again.

Exercise: Compare employment rates for different regions of the country.

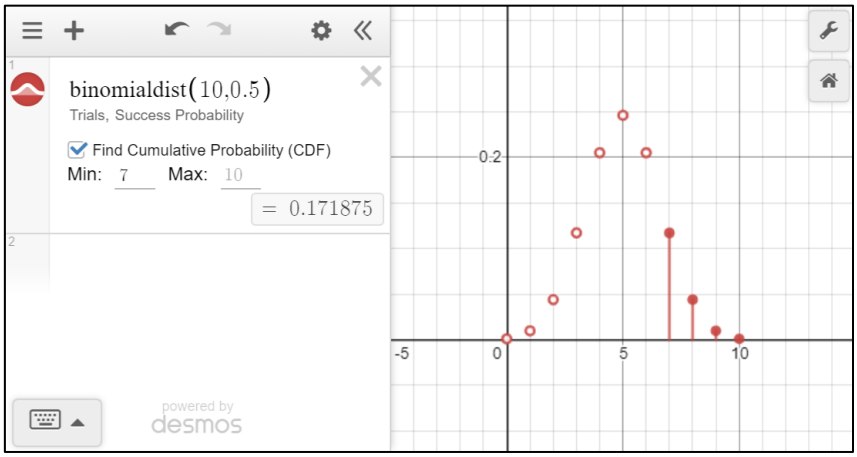
Appendix 2: Suggested Large Data Set Investigations

- Are people in London more likely to cycle to work than people in other parts of the country?
- Which areas of the country have the highest use of public transport?
- Have there been any changes in the modes of travel to work between 2001 and 2011?
- Do areas of the country with older populations drive more?
- Is there a correlation between cycling to work and walking to work?
- Is there a correlation between travelling by underground/metro/… and travelling by train?
- Are people in any parts of the country more likely to work from home?
- Are there any other regional differences in any of the categories?

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.