



Advanced Mathematics
Support Programme®

Using Large Data Sets Workbook – AQA (Cars) 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 AQA dataset for 2019-2020 which can be downloaded at

<https://www.aqa.org.uk/subjects/mathematics/as-and-a-level/mathematics-7357/assessment-resources>

Key Skills

- Understand the dataset and its context
- Cleanse 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).

1 Becoming familiar with the dataset

Open the “AQA-AS-A-MATHS-LDS-2019-2020” file which contains the dataset. The first sheet in the spreadsheet files explains the origin of the data and the third sheet contains a glossary of terms. The fourth sheet contains a key to the codes used. The data itself is in the second sheet.

Students are required to understand the context of the data so that it is important that they read these sheets whilst looking through the dataset. Some questions you might like to consider are:

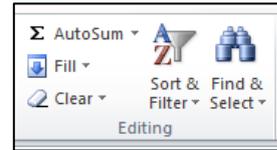
- What is the source of the data and how up to date is it?
- Who collected it and how was it collected?
- Why are there only 5 makes of car and how were these chosen? What other restrictions were applied to the dataset?
- There are some missing data items. How should you treat these items when analysing the data?

Students need to understand each of the fields. Some of them might warrant further discussion. Students should be encouraged to research further so that they fully understand the concepts. Further information can be found at <https://www.gov.uk/government/statistical-data-sets/all-vehicles-veh01> and <https://www.gov.uk/government/statistical-data-sets/all-vehicles-veh01>

2 Sorting and filtering the dataset in Excel

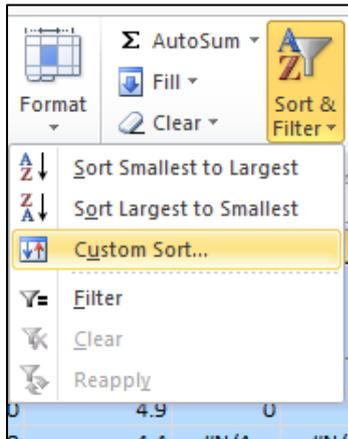
Further familiarity with the dataset can be gained by sorting and filtering the data within Excel. This can help identify any possible outliers or rogue values.

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

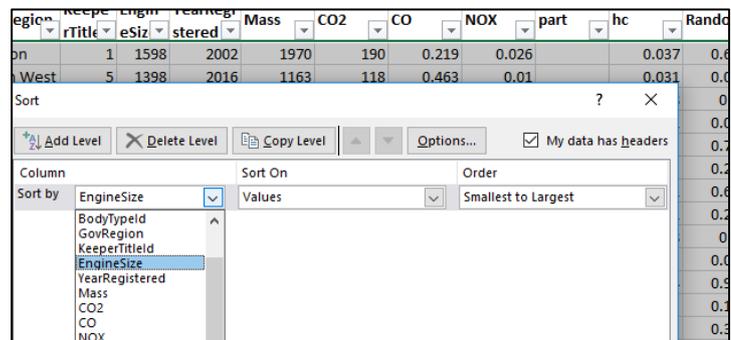


Suppose you want to sort the data according to Engine Size. Use Ctrl-A to select all the data.

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.

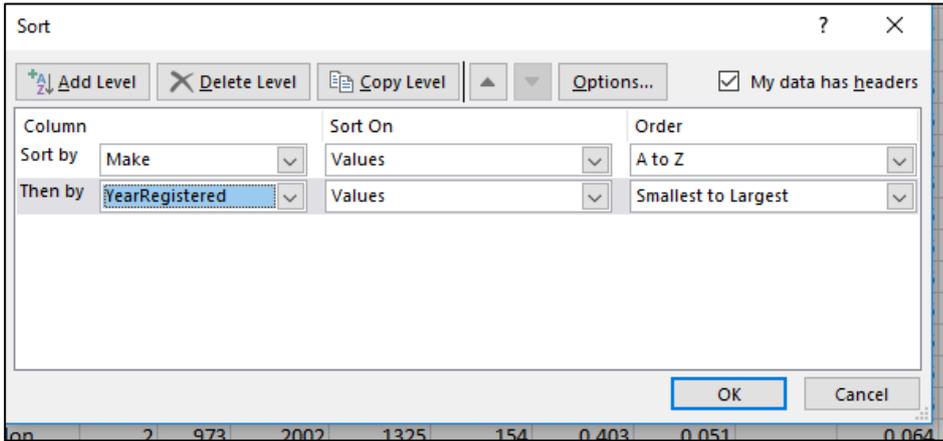


The data is now sorted in order of engine size:

| Reference | Make | PropulsionType | BodyType | GovRegion | KeeperTitle | EngineSize | YearRegistered | Mass | CO2 | CO | NOX |
|-----------|----------|----------------|----------|------------|-------------|------------|----------------|------|-----|-------|-------|
| 3645 | BMW | 3 | 14 | South West | 2 | 0 | 2016 | 1270 | 0 | | |
| 3530 | BMW | 8 | 14 | South West | 1 | 647 | 2016 | 1390 | 13 | 0.053 | 0.002 |
| 3495 | BMW | 8 | 14 | South West | 1 | 647 | 2016 | 1390 | 13 | 0.053 | 0.002 |
| 3025 | BMW | 8 | 14 | North West | 2 | 647 | 2016 | 1390 | 13 | 0.053 | 0.002 |
| 974 | VAUXHALL | 1 | 13 | North West | 2 | 973 | 2002 | 1430 | 135 | 0.342 | 0.048 |
| 505 | VAUXHALL | 1 | 13 | South West | 1 | 973 | 2002 | 1430 | 135 | 0.342 | 0.048 |
| 519 | VAUXHALL | 1 | 13 | North West | 1 | 973 | 2002 | 1430 | 135 | 0.342 | 0.048 |
| 337 | VAUXHALL | 1 | 13 | London | 2 | 973 | 2002 | 1430 | 135 | 0.342 | 0.048 |
| 503 | VAUXHALL | 1 | 14 | London | 1 | 973 | 2002 | 1430 | 135 | 0.342 | 0.048 |
| 1017 | VAUXHALL | 1 | 13 | London | 1 | 973 | 2002 | 1405 | 135 | 0.342 | 0.048 |

What problem has been identified here? How would you deal with this?

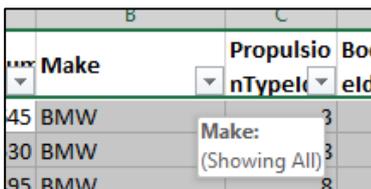
It is possible to sort the data using several fields using the 'add level button



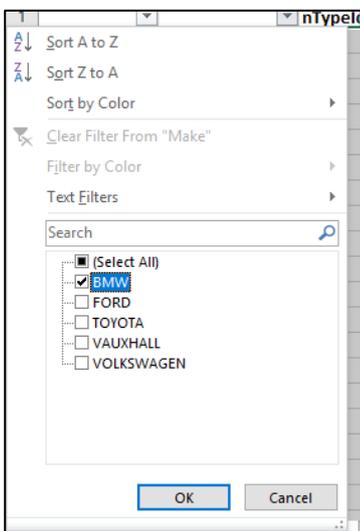
Try the above sort (remember to select all the data first using Ctrl-A). It should give you the data in order of make and the year.

An alternative approach if you just want to focus on a subset of the data is to use filters. The dataset comes with filters already set up. These are the arrows next to each heading.

Suppose you wish to just look at BMWs. Then click on the arrow next to make:



And then scroll down and uncheck the box for everything except BMW



Now you should have only BMWs:

| ReferenceNo | Make | PropulsionType | BodyType | GovRegion | KeepersTitle | EngineSize | YearRegistered | Mass | CO2 |
|-------------|------|----------------|----------|------------|--------------|------------|----------------|------|-----|
| 3645 | BMW | 3 | 14 | South West | 2 | 0 | 2016 | 1270 | 0 |
| 3530 | BMW | 8 | 14 | South West | 1 | 647 | 2016 | 1390 | 13 |
| 3495 | BMW | 8 | 14 | South West | 1 | 647 | 2016 | 1390 | 13 |
| 3025 | BMW | 8 | 14 | North West | 2 | 647 | 2016 | 1390 | 13 |
| 2335 | BMW | 2 | 14 | London | 5 | 1496 | 2016 | 1395 | 106 |
| 3606 | BMW | 2 | 14 | North West | 2 | 1496 | 2016 | 1395 | 103 |
| 2719 | BMW | 2 | 14 | South West | 5 | 1496 | 2016 | 1425 | 96 |
| 1947 | BMW | 2 | 13 | London | 5 | 1496 | 2016 | 1425 | 103 |
| 3309 | BMW | 2 | 14 | South West | 5 | 1496 | 2016 | 1395 | 89 |
| 2612 | BMW | 2 | 14 | London | 5 | 1496 | 2016 | 1395 | 94 |

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

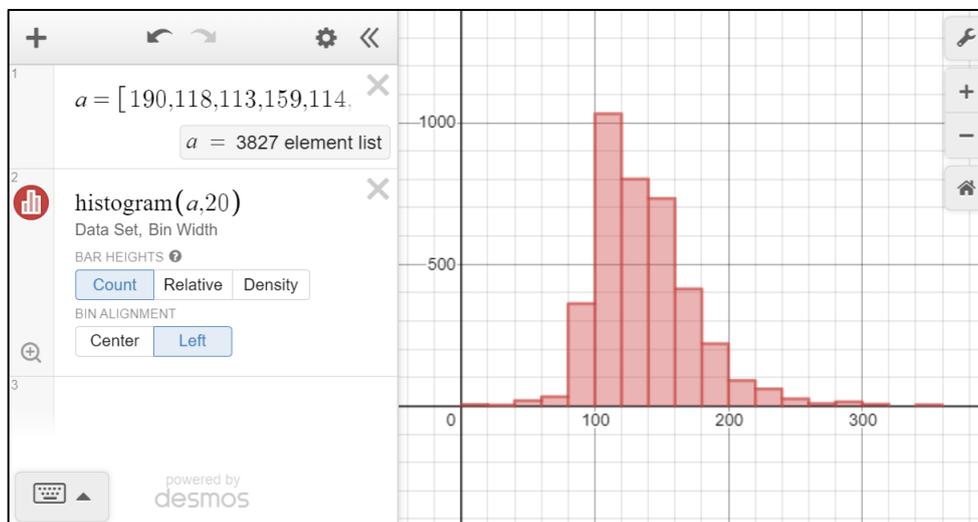
Exercise: Sort the data by Carbon Dioxide emissions. Do the newer cars generally have lower emissions? Are there any exceptions?

4 Drawing frequency charts and box plots

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.

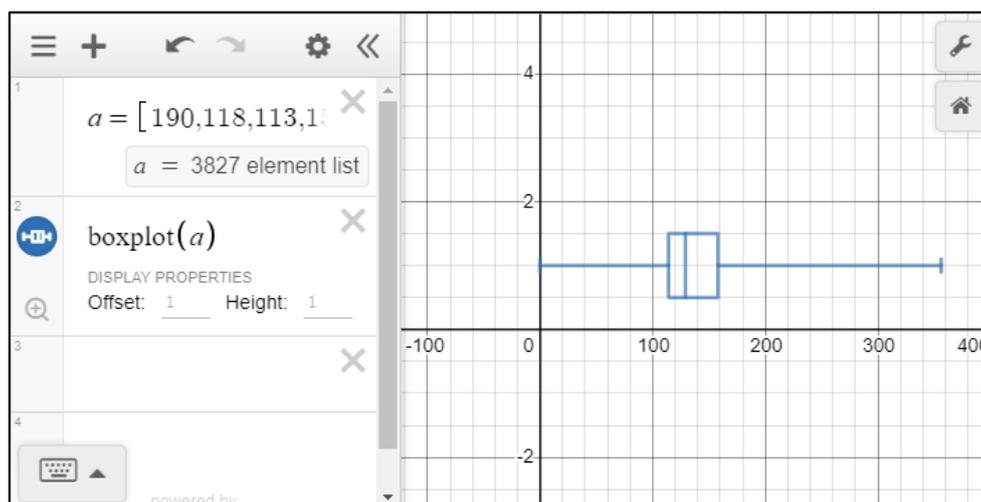
For a histogram you should enter the data set and the bin width.



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 20 the first bin will contain values of x where $0 \leq x < 20$.

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?

5 Drawing Graphs side by side for comparison

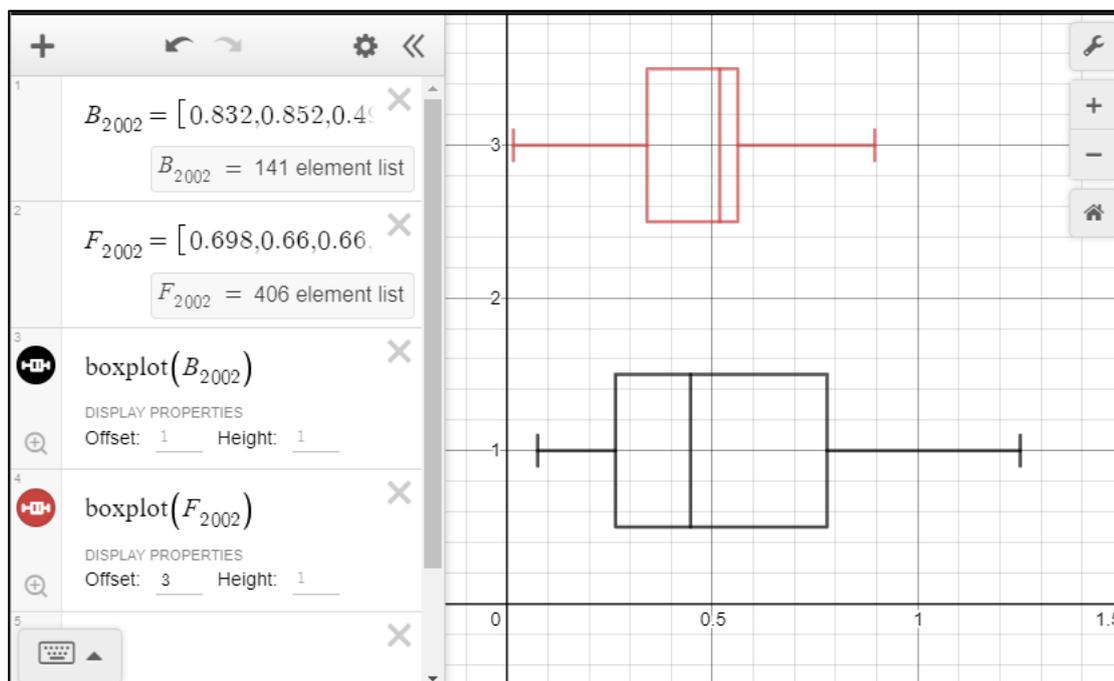
The following example compares the CO emissions for different makes and different years.

Going back to the original excel spreadsheet, **sort** the data by year and then by make:

| Reference | Make | PropulsionType | BodyType | GovRegion | KeeperTitle | EngineSize | YearRegistered | Mass | CO2 | CO | NOX | pa |
|-----------|------|----------------|----------|------------|-------------|------------|----------------|------|-----|-------|-------|----|
| 321 | BMW | 1 | 2 | South West | 1 | 2171 | 2002 | 1600 | 237 | 0.832 | 0.005 | |
| 886 | BMW | 1 | 2 | North West | 2 | 1995 | 2002 | 1395 | 175 | 0.852 | 0.05 | |
| 420 | BMW | 1 | 4 | South West | 1 | 2494 | 2002 | 1615 | 230 | 0.494 | 0.017 | |
| 588 | BMW | 1 | 3 | London | 1 | 2979 | 2002 | 0 | 230 | 0.447 | | |
| 737 | BMW | 2 | 13 | North West | 1 | 1995 | 2002 | 1505 | 185 | 0.203 | 0.375 | |
| 921 | BMW | 1 | 2 | London | 1 | 2171 | 2002 | 1600 | 237 | 0.832 | 0.005 | |
| 1021 | BMW | 1 | 6 | North West | 1 | 4619 | 2002 | 2180 | 356 | 1.231 | 0.068 | |
| 250 | BMW | 1 | 5 | South West | 1 | 2979 | 2002 | 1505 | 218 | 0.269 | 0.009 | |
| 484 | BMW | 1 | 6 | London | 1 | 2979 | 2002 | 2095 | 310 | 0.582 | 0.015 | |
| 920 | BMW | 2 | 2 | North West | 2 | 1995 | 2002 | 1490 | 148 | 0.101 | 0.377 | |
| 278 | BMW | 1 | 4 | South West | 2 | 2171 | 2002 | 1355 | 222 | 1.02 | 0.037 | |
| 805 | BMW | 1 | 5 | North West | 1 | 2171 | 2002 | 1500 | 226 | 0.498 | 0.014 | |

Each group will need to be copied as a new list into Desmos. Data items K2 to K142 correspond to CO emission from BMWs manufactured in 2002. Data items K143 to K549 to the CO emissions of 2002 Fords.

Copying each group across into Desmos (NB just type B2002 to obtain the variable name with the subscript):



Exercise: Draw the boxplots for the CO emissions for other makes. What conclusions can be reached by comparing these plots?

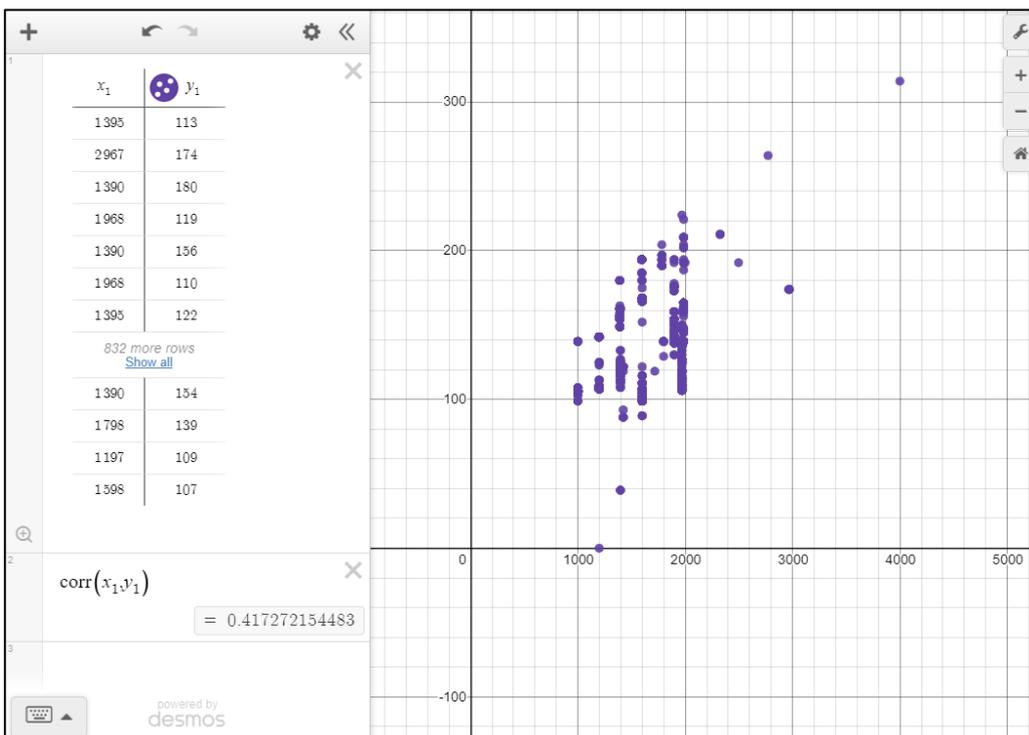
6 Drawing Scatterplots

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.

Desmos has a maximum of 1000 data pairs for a bivariate data so you will need to work with a subset of the full data set. In the following example the data has been filtered to show only Volkswagen cars and the columns for Engine Size and CO2 have been copied by selecting column G then pressing Ctrl and selecting column J. These have been copied into a new blank spreadsheet and then these two columns have been copied and pasted into Desmos.

| | A | B | C | D | E | F | G | H | I | J | K |
|-----|-----------|------------|----------------------|----------------|------------|-----------------|----------------|--------------------|------|-----|-------|
| 1 | Reference | Make | Propulsio nTypeId | BodyType Id | GovRegion | Keeper Title | Engine Size | YearRegis tered | Mass | CO2 | CO |
| 4 | 3434 | VOLKSWAGEN | 1 | 14 | South West | 2 | 1395 | 2016 | 1316 | 113 | 0.242 |
| 11 | 2971 | VOLKSWAGEN | 2 | 6 | South West | 5 | 2967 | 2016 | 2185 | 174 | 0.138 |
| 19 | 371 | VOLKSWAGEN | 1 | 14 | London | 1 | 1390 | 2002 | 1220 | 180 | 0.096 |
| 26 | 1929 | VOLKSWAGEN | 2 | 14 | London | 2 | 1968 | 2016 | 1394 | 119 | 0.079 |
| 37 | 1156 | VOLKSWAGEN | 1 | 14 | London | 2 | 1390 | 2002 | 1181 | 156 | 0.165 |
| 49 | 2274 | VOLKSWAGEN | 2 | 6 | South West | 5 | 1968 | 2016 | 1503 | 110 | 0.137 |
| 52 | 3446 | VOLKSWAGEN | 1 | 14 | South West | 2 | 1395 | 2016 | 1247 | 122 | 0.158 |
| 55 | 3331 | VOLKSWAGEN | 1 | 6 | London | 5 | 1984 | 2016 | 1574 | 162 | 0.609 |
| 59 | 3523 | VOLKSWAGEN | 1 | 14 | South West | 2 | 1197 | 2016 | 1107 | 107 | 0.157 |
| 86 | 2622 | VOLKSWAGEN | 1 | 13 | London | 1 | 999 | 2016 | 1055 | 106 | 0.407 |
| 99 | 3348 | VOLKSWAGEN | 1 | 14 | London | 2 | 1395 | 2016 | 1247 | 120 | 0.158 |
| 108 | 3500 | VOLKSWAGEN | 1 | 14 | South West | 2 | 1197 | 2016 | 1229 | 113 | 0.236 |
| 109 | 2196 | VOLKSWAGEN | 1 | 14 | South West | 2 | 1197 | 2016 | 1107 | 107 | 0.157 |
| 113 | 3647 | VOLKSWAGEN | 2 | 5 | South West | 1 | 1968 | 2016 | 1375 | 109 | 0.212 |
| 120 | 3305 | VOLKSWAGEN | 2 | 96 | London | 5 | 1968 | 2016 | 1843 | 136 | 0.052 |
| 122 | 53 | VOLKSWAGEN | 1 | 13 | London | 1 | 1596 | 2002 | 1285 | 185 | 0.53 |
| 123 | 2130 | VOLKSWAGEN | 1 | 13 | North West | 2 | 1984 | 2016 | 1382 | 139 | 0.341 |

| | A | B |
|----|------------|-----|
| 1 | EngineSize | CO2 |
| 2 | 1395 | 113 |
| 3 | 2967 | 174 |
| 4 | 1390 | 180 |
| 5 | 1968 | 119 |
| 6 | 1390 | 156 |
| 7 | 1968 | 110 |
| 8 | 1395 | 122 |
| 9 | 1984 | 162 |
| 10 | 1197 | 107 |
| 11 | 999 | 106 |
| 12 | 1395 | 120 |
| 13 | 1197 | 113 |
| 14 | 1197 | 107 |
| 15 | 1968 | 109 |
| 16 | 1968 | 136 |
| 17 | 1596 | 185 |
| 18 | 1984 | 139 |



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

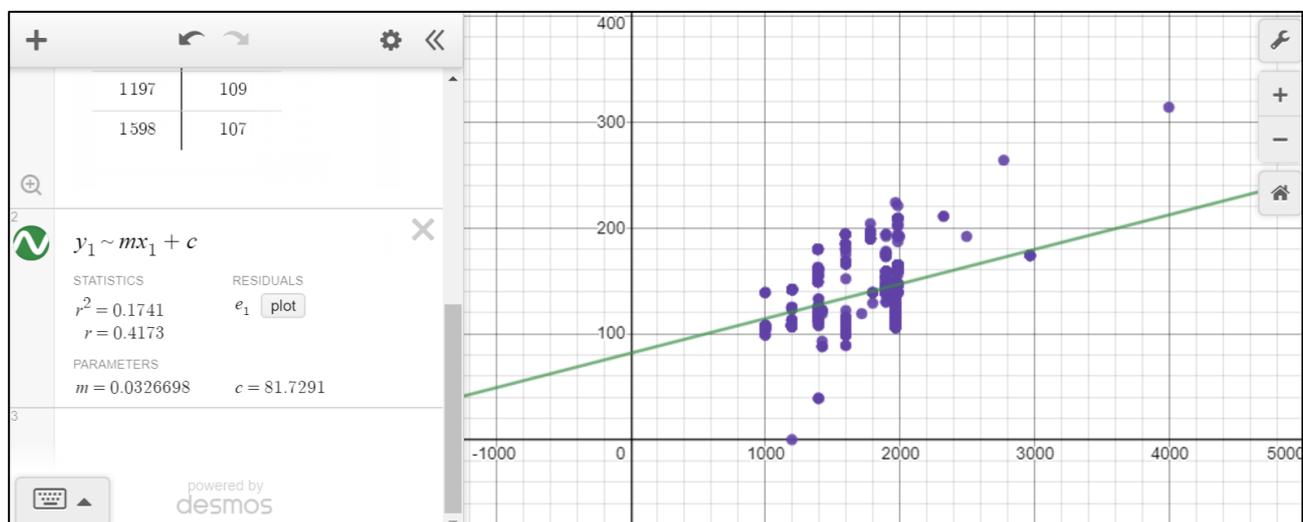
NB for subscripts typing x_1 will automatically be updated to x_1 .

7 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, \sim , e.g. $y_1 \sim mx_1 + c$.



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 engine size and mass. There are a number of cars whose mass is recorded as zero. Can you exclude these to get a better picture?

8 Random Sampling

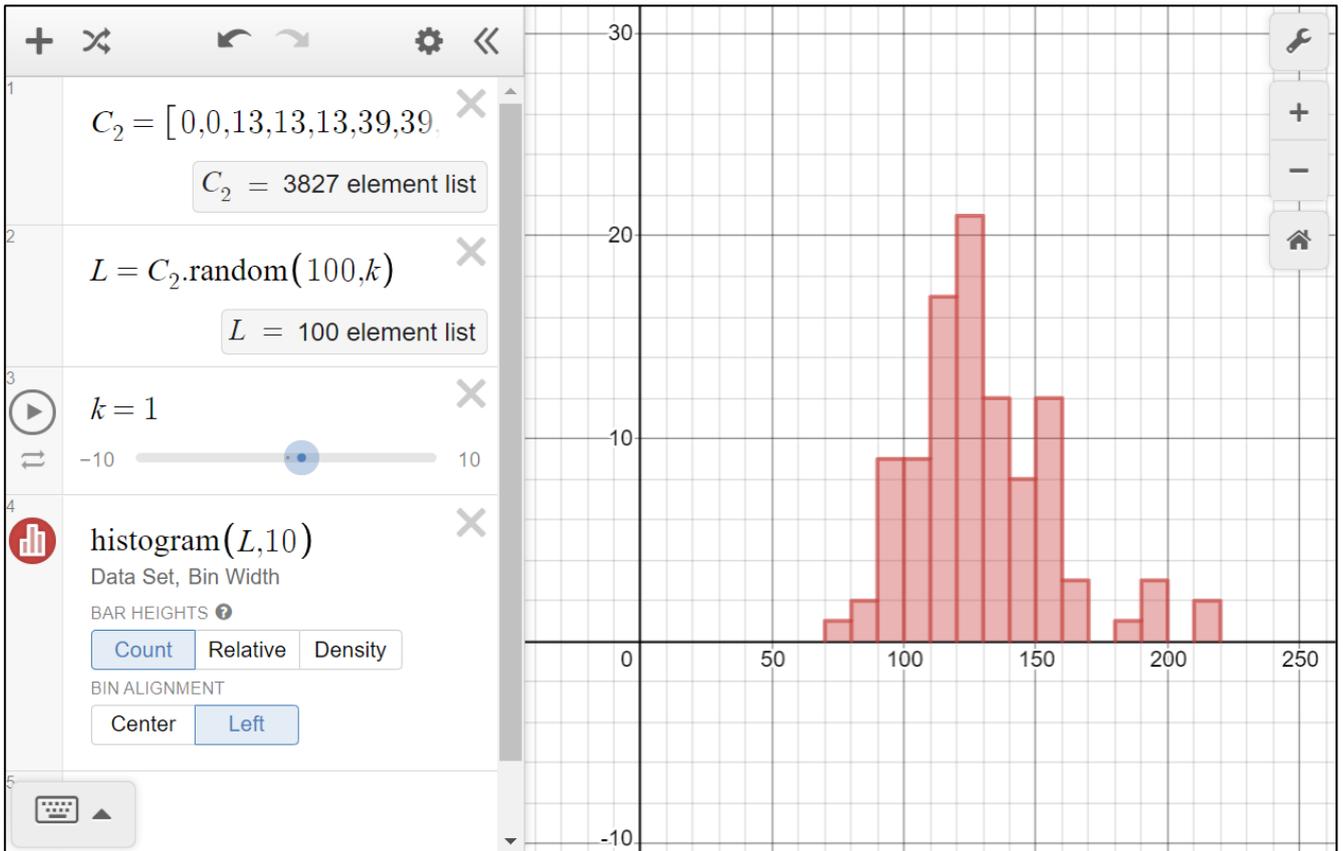
You can use the random function to select a random sample from a list of values. In the following example a random sample of size 20 will be selected from the CO₂ data.

Enter a new variable of C₂ and copy the full list into Desmos.

To create a random sample of size 100 type: $L = C_2.\text{random}(100)$

A histogram for this sample can be plotted using $\text{histogram}(L, 10)$.

To generate different samples change the list to $L = C_2.\text{random}(100, k)$. k is a “seed” for the random sample and changing this will generate a new sample.



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.

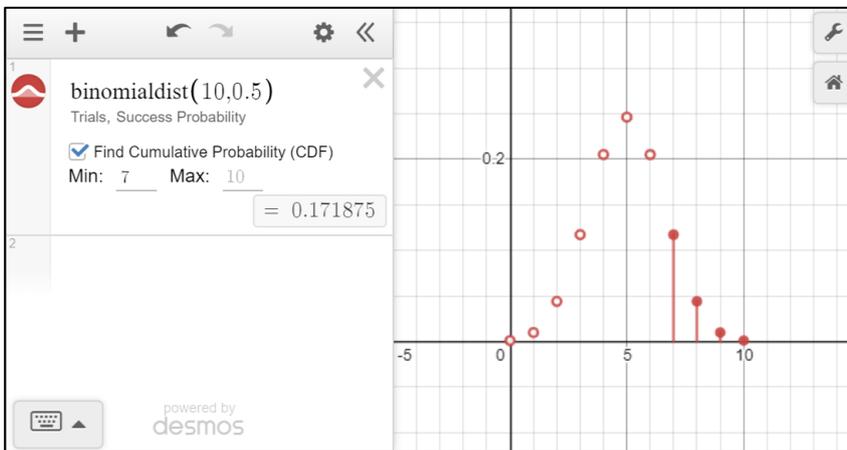
Appendix 1: Ideas for investigations

- Are there any differences in emissions between different makes of cars?
- Are the emissions for cars registered in 2002 worse than for the cars registered in 2016?
- Is there a correlation between mass and emissions?
- Is there a correlation between engine size and emissions?
- Is there a correlation between the CO₂ and NO_x emissions?
- Is there a link between the engine type and the emissions?
- Are any particular makes of cars more likely to be owned by companies?
- Do different genders have different sized cars?

Appendix 2: 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.

