The Large Data Set AQA

Learning objectives
To investigate a subset of data from the AQA Large Data Set To build familiarity with the AQA Large Data Set

Overview

This resource establishes principles for working with the AQA Large Data Set. The spreadsheet *Cars with headings.xlsx* contains a subset of the LDS and is available from the resource centre. The cleaned data *Cars.csv* also provided is ready to import to the fx-CG50. A learner worksheet leads learners to explore the data using calculations and graphs. Learners are then asked to choose different subsets of the data to work with, and to feed back their findings to the rest of the class. Some instructions on how to work with multivariate data on the fx-CG50 are included in teaching notes.

Context

This activity ideally comes immediately after learning the statistical techniques as the practice of calculating statistics and drawing graphs. It could be done in two sections – single variable statistics comparing two subset and bivariate statistics looking at correlation and regression between two lists from the data.

Activity

A .csv file of a subset of the data is provided to open using the fx-CG50. Guidance is then given to investigate information from the data. Learners are then asked to find their own subset from the LDS (ideally different from others in the class) and make their own investigation. Familiarity with the data set and the wording that is helpful when drawing inferences from the data are built as learners feed back to their peers what they have discovered.

Using this data set:

- Look at a list on its own (histograms)
- Compare one list with another list (measures of centre and spread, outliers, boxplots)
- Divide a list into subsets and compare the subsets
- Look for correlation between two lists (scatter diagram and correlation coefficient)
- Look at the regression line for an independent and dependent variable (interpret the coefficients from the equation).

A study of your own:

- Decide what would you like to investigate from the data (try to choose something different from your classmates)
- Create your own clean data subset and import to your calculator
- Follow the structure above and make a note of your key finds
- Share with the class.

Contents

Learning resource objective
Investigate part of the Large Data Set – 2016 petrol and 2016 diesel cars

Taking a sample

1. Explain how random numbers can be used to generate a sample of 100 cars from the large data set.

....

....

....

Download the data set *Cars.csv* and open in **Statistics**.

There are 10 columns of data from the large data set and a random sample of 100 cars of each type.

List 1	List 2	List 3	List 4	List 5	List 6	List 7	List 8	List 9	List 10
Petrol engine size	**Petrol mass**	**Petrol CO2**	**Petrol CO**	**Petrol NOX**	**Diesel engine size**	**Diesel mass**	**Diesel CO2**	**Diesel CO**	**Diesel NOX**

Look at a list on its own

2. Draw separate histograms for CO2 emissions for petrol and diesel cars. Comment..............................

....

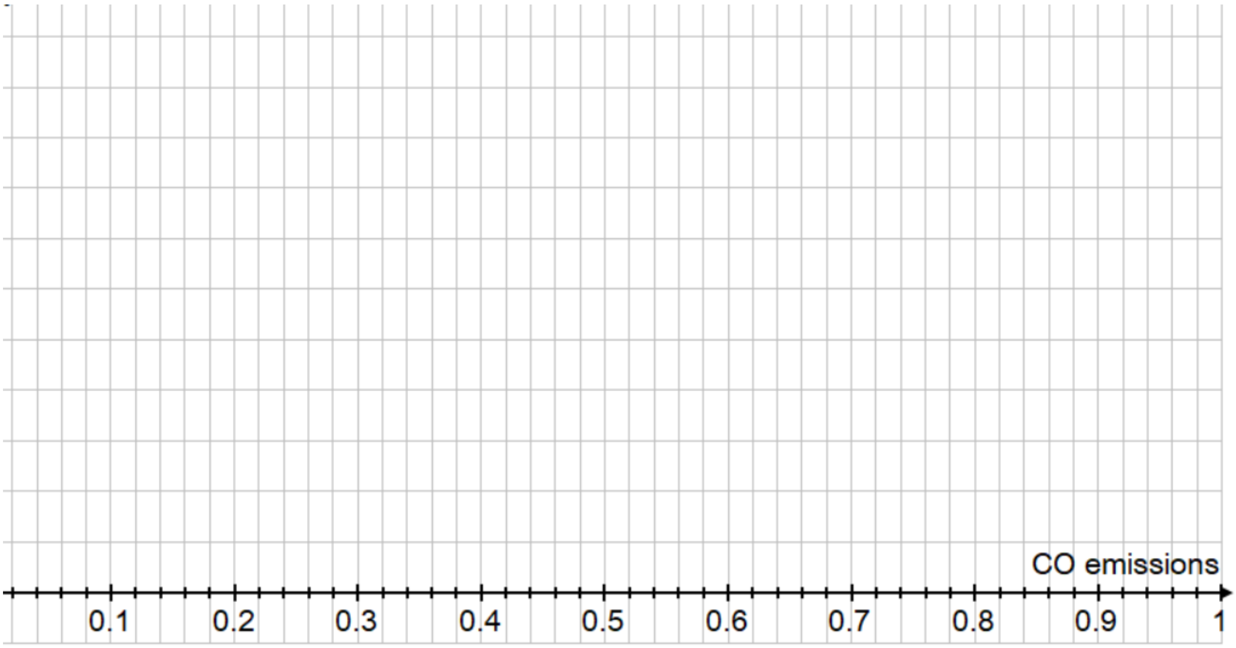
....

Compare one list with another

3. Calculate these statistics for petrol and diesel cars (Lists 4 and 9). The values for hybrid cars are also given.

	Petrol CO emissions	Diesel CO emissions	Hybrid CO emissions
Mean			0.1
Standard deviation			0.057242885
Boundary value for outliers using mean			0.02801423
Boundary value for outliers using mean			0.25698577
Minimum			0.053
Q1 – lower quartile			0.1
Median			0.131
Q3 - upper quartile			0.182
Maximum			0.267
Boundary value for outliers using quartiles			-0.023
Boundary value for outliers using quartiles			0.305

4. Draw three boxplots on these axes.



5. Compare the CO2 emissions for the three types of cars.

6. In 2016 there were government tax incentives to encourage buyers to choose diesel cars. Use the data to explain why.

7. In 2021 Birmingham introduced a clean air zone where some diesel cars had to pay to enter the zone. What can you find in this data subset to explain this decision?

CASIO

Look for correlation between two lists

8. Explain why the data in list 1 is a good choice to be an independent variable. ....

....

....

9. What correlation do you expect to see with engine size? Draw a scatter diagram of each other list against engine size and calculate the correlation coefficient. Comment on your findings.

Mass against engine size ....

....

CO2 emissions against engine size ....

....

CO emissions against engine size ....

....

NOX emissions against engine size ....

....

Is this the same for diesel cars? ....

....

Look at the regression line for an independent and dependent variable

10. Draw a scatter diagram for CO2 emissions against engine size. (**XList:** List1, **YList:** List3 or **XList:** List6, **YList:** List8). Write down the equation of the regression line. Interpret the values of the coefficients in the regression line. Comment on the appearance of the graph.

....

....

....

....

11. How well do CO2 and CO emissions correlate? Is that what you expected?

....

....

....

Extension

Create new data from the existing data and analyse that

12. Create an additional list for engine size divided by mass (possible indicator of acceleration). Compare petrol with diesel cars for this.

A study of your own

- Decide what would you like to investigate from the data. Try to choose something different from your classmates.
- Create your own clean data subset using a spreadsheet and save as a .csv file to transfer to your calculator.
- Follow the structure above and make a note of your key findings.
- Share with the class.

Aims

To investigate a subset of data from a Large Data Set

To build familiarity with the AQA Large Data Set

Resources

Supporting resources are available in our Resource Centre.

Before the lesson

How To Videos

- Statistics – Managing Lists
- Single Variable Statistics
- Boxplots – Plotting and Values
- Bivariate Data and Regression – Scatter Diagrams

During the lessons

- Learner worksheet

After the lesson

- Learners' own investigations into the Large Data set

Using the fx-CG50

Downloading a file to the fx-CG50

Download the *Cars.csv* file to your computer. Either watch the video on how to transfer files on our website or follow these instructions:

1) Use the USB lead to connect the fx-CG50 to your computer.
2) When prompted by the calculator choose **USB Flash** (F1).
3) The computer recognises the fx-CG50 as a USB drive and you can manage the files on the fx-CG50 in the same way as a flash drive. Drag and drop or copy and paste the file into a folder of your choosing.
4) Before unplugging the USB cable, terminate the USB connection on your computer.

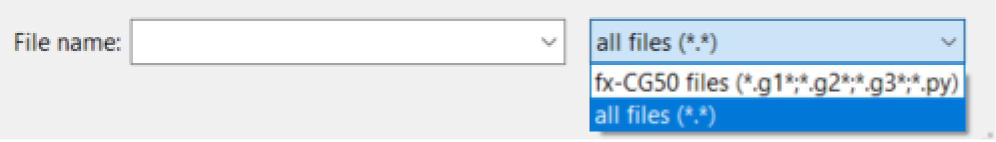
Opening a saved file on the fx-CG50

To open the file, from MENU choose **Statistics**. Then choose further menu ▷ (F6) twice followed by **CSV** (F1), **LOAD** (F1), **FILE** (F2) and scroll to the file.

Choose **OPEN** (F1). The 12 columns of data should now be visible. You can scroll to the row labelled **SUB** and type a heading if you wish. Use (A-LOCK) (SHIFT - ALPHA) to type several letters in a row.

Using the emulator

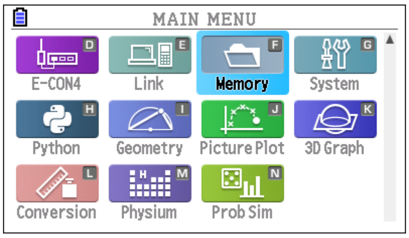
To open files with the emulator, from MENU choose **Memory** followed by **Import/Export** (F3). Then choose **Import** (F1) and search for the .csv file on your computer – you may need to change the file type to all files *.* to see files other than the Casio native files.

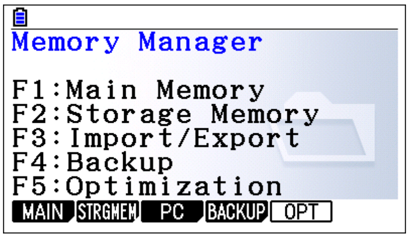


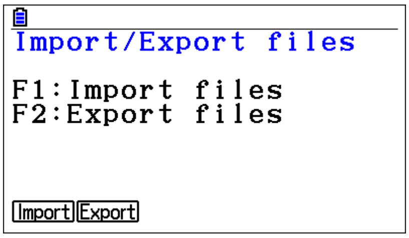
File name:

all files (*.*)

Double click the file you want to open, scroll to the location you want to save the file to and choose **SAVE** (F1). Press EXIT and go to **Statistics** to open the file as you would on the calculator itself.

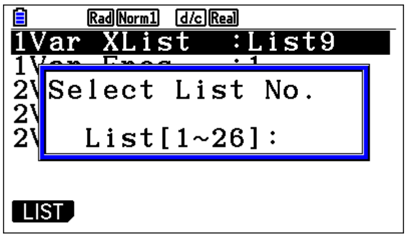




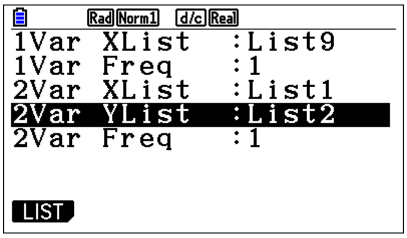


To calculate a set of statistics for a list of numbers

In **Statistics**, choose **CALC** (F2) and **SET** (F6). For one-variable calculations, choose the List using **LIST** (F1) and type the number of the List you wish to consider. To see 2 lists one after the other, or to see correlation and regression calculations, select the two lists in which you find the x and y variables. Notice they do not need to be together in the display.



```
1Var XList   :List9
1Var Freq    :1
Select List No.
  List[1~26]:
```



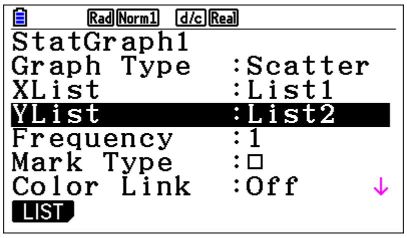
```
1Var XList   :List9
1Var Freq    :1
2Var XList   :List1
2Var YList   :List2
2Var Freq    :1
```

Press EXE to store the settings. Then choose **1-VAR** (F1) to see the statistics for the List you have selected and **2-VAR** F2 to see two lists of separate calculations. Choose **REG** (F3), **X** (F1) and either **ax+b** (F1) or **a+bx** (F2) to see the correlation coefficient and the coefficients in the equation of the regression line.

Note this can be used to create lists of statistics for several lists in quick succession if you need them.

To select lists to graphed

Use EXIT to return to the top menu in **Statistics**. Choose **GRAPH** (F1) and **SET** (F6). You can choose the settings for up to three graphs before returning to the previous menu. For **StatGraph**, choose **GRAPH1** (F1), **GRAPH2** (F2) or **GRAPH3** (F3). Scroll down to Graph Type and choose from the menus – note there is a further menu ▷ of graph types. Choose the List(s) for your graph and scroll down for more options for your graph. Repeat for the other graphs. Press EXE to save the settings and then choose the relevant graph (F1, F2 or F3) to display.

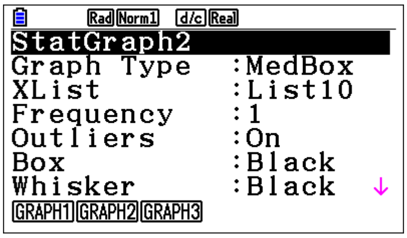


```
StatGraph1
Graph Type    :Scatter
XList         :List1
YList         :List2
Frequency     :1
Mark Type     :□
Color Link    :Off
```

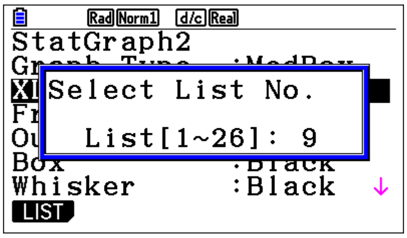
The How To videos listed above cover how to manage lists, calculate single variable statistics and draw graphs.

To create boxplots with different colours

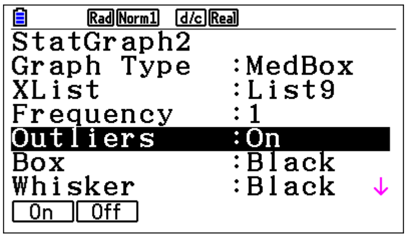
Go to **Statistics**, choose **GRAPH** and **SET**. Choose **GRAPH1** to be Graph Type: MedBox for one List and **GRAPH2** to be Graph Type: MedBox for the other. You can turn Outliers: On or Off and scroll down to change colours for the shading or the outliers.



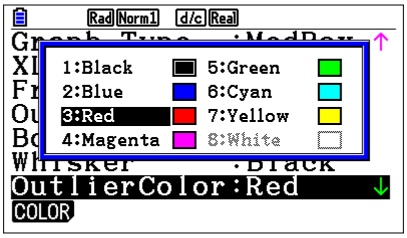
```
StatGraph2
Graph Type    :MedBox
XList         :List10
Frequency     :1
Outliers      :On
Box           :Black
Whisker       :Black
GRAPH1 GRAPH2 GRAPH3
```



```
StatGraph2
XL Select List No.
   List[1~26]: 9
Whisker       :Black
LIST
```



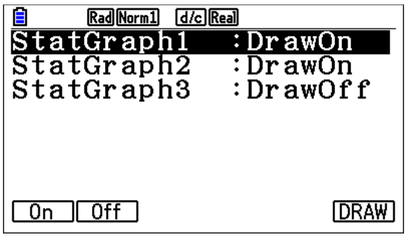
```
StatGraph2
Graph Type    :MedBox
XList         :List9
Frequency     :1
Outliers      :On
Box           :Black
Whisker       :Black
On  Off
```



```
1:Black   ■  5:Green  🟩
2:Blue    🟦  6:Cyan   ■
3:Red     🟥  7:Yellow 🟨
4:Magenta ■  8:White  □
OutlierColor:Red
COLOR
```

To show multiple graphs together

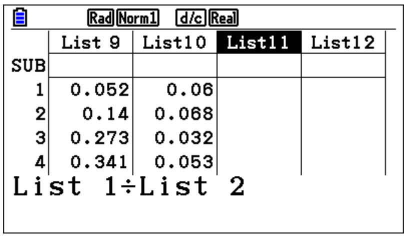
You can create two or three boxplots, or two scatter diagrams with similar ranges of data, as **GRAPH1** and **GRAPH2** for example. To display them on the same axes, go to **GRAPH**, choose **SELECT** (F4) and scroll to the graphs you wish to see together. Then switch them to DrawOn. Choose **DRAW** (F6) to see the selected graphs together.



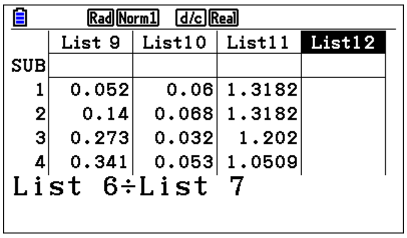
```
StatGraph1    :DrawOn
StatGraph2    :DrawOn
StatGraph3    :DrawOff

On  Off                DRAW
```

Creating a new list

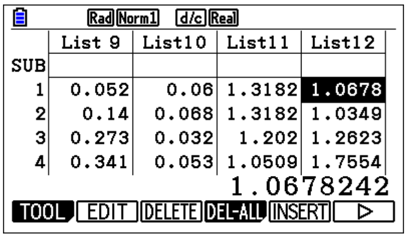
To create lists for engine size divided by mass (from Lists 1 and 2 respectively) scroll across to the heading of List 11. Press SHIFT - 1 - 1 to type List 1. Press ÷ , then SHIFT - 1 - 2 for List 2. Press EXE to see all the values in List 11. Similarly for List 12 using the data from Lists 6 and 7.



```
     List 9  List10  List11  List12
SUB
  1   0.052    0.06
  2    0.14   0.068
  3   0.273   0.032
  4   0.341   0.053
List 1÷List 2
```



```
     List 9  List10  List11  List12
SUB
  1   0.052    0.06  1.3182
  2    0.14   0.068  1.3182
  3   0.273   0.032   1.202
  4   0.341   0.053  1.0509
List 6÷List 7
```



```
     List 9  List10  List11  List12
SUB
  1   0.052    0.06  1.3182  1.0678
  2    0.14   0.068  1.3182  1.0349
  3   0.273   0.032   1.202  1.2623
  4   0.341   0.053  1.0509  1.7554
                          1.0678242
TOOL EDIT DELETE DEL-ALL INSERT ▷
```

Dealing with the unexpected

The scatter diagrams for many of the pairs of variables appear to be made up of vertical stripes – this is because there are relatively few distinct x-coordinates in the data set, each with a variety of y-coordinates.

The List facility is not a spreadsheet, so if you use Lists 1 and 2 to create List 11, any later changes to Lists 1 and 2 do not affect the values stored in List 11. If you anticipate making changes to the data, use the spreadsheet facility on the fx-CG50 or your PC instead.

Prompts

When looking at histograms, prompt learners to click through several lists by going to **GRAPH** then **SET** and change the list from which the data is gathered for the histogram. Explore the effect of changing the start and the width – use sensible values looking at the minimum and maximum for that list.

When comparing data from two lists, prompt learners to explore lots of pairs by comparing boxplots (MedBox). Go to **GRAPH** then **SET** and change the Lists from which the graphs are drawn. Choose **SELECT** and turn the relevant graphs **On**. You can click through lots of pairs to see those which will be interesting to compare in greater depth.

Similarly for correlation and regression, draw scatter diagrams for lots of pairs and look for those which are interesting – good correlation that you might not expect, or much less correlation than you might have thought.

Prompt learners when setting up their own study to choose something different from their classmates so that all aspects of the large data set have been examined by someone in the class. Encourage learners to share their findings with key graphs, interesting statistics and inferences that can be drawn from the data.

Extension questions

Do you think you will find similar things for the cars in 2002? Have cars got cleaner in that period?

What other factors in the data set could be considered as independent variables for a new study?

Solutions

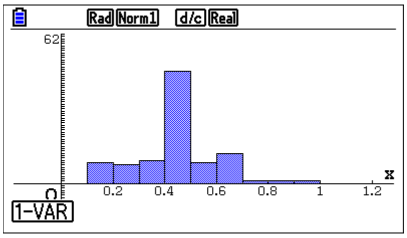
Taking a sample

1. In a spreadsheet, sort the data by the random number to mix the rows. Select the required number or rows from the top of the new lists.

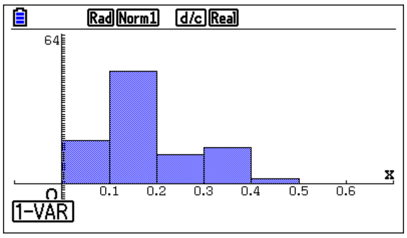
List 1	List 2	List 3	List 4	List 5	List 6	List 7	List 8	List 9	List 10
Petrol engine size	Petrol mass	Petrol CO2	Petrol CO	Petrol NOX	Diesel engine size	Diesel mass	Diesel CO2	Diesel CO	Diesel NOX

Look at a list on its own

Petrol CO2 emissions



Diesel CO2 emissions



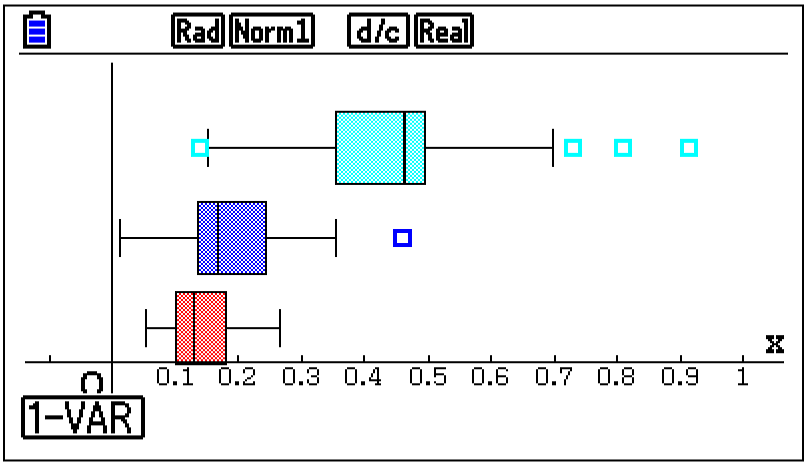
2. For each set one class is the modal class with a much higher frequency than the other classes – so you could say that typically petrol cars emit between 0.4 g/km and 0.5 g/km whereas a typical diesel car emits between 0.1 g/km and 0.2 g/km.

Compare one list with another

3.

	Petrol CO emissions	Diesel CO emissions	Hybrid CO emissions
Mean	0.4	0.2	0.1
Standard deviation	0.152023	0.09801	0.057242885
Boundary value for outliers using mean	0.134854	-0.01131	0.02801423
Boundary value for outliers using mean	0.742946	0.38073	0.25698577
Minimum	0.139	0.013	0.053
Q1 – lower quartile	0.355	0.136	0.1
Median	0.463	0.167	0.131
Q3 - upper quartile	0.48975	0.243	0.182
Maximum	0.913	0.459	0.267
Boundary value for outliers using quartiles	0.152875	-0.0245	-0.023
Boundary value for outliers using quartiles	0.691875	0.4035	0.305

4. :



5. Compare the CO2 emissions for the three types of cars.

   Measures of centre: The mean and median value for petrol cars is the highest and lowest for hybrid cars, so the hybrid cars emit least CO2.

   Measures of spread: The standard deviation and interquartile range is much bigger for petrol cars so the emissions vary a lot between petrol cars. These measures are smallest for hybrid cars, so the emissions are much more consistent between hybrid cars.

   Shape: There are several outliers for petrol cars – particular cars whose emission are much higher than the rest, and one unusually low one also. There is one outlier for diesel cars, but the emissions are only similar to the median for petrol cars. All the values for hybrid cars are less than almost all the petrol cars.

   The data for petrol cars has a slight negative skew, but the skew is positive for diesel and hybrid.

6. In 2016 there were government tax incentives to encourage buyers to choose diesel cars because their CO2 emissions are lower than petrol cars and this would contribute to the UK efforts to become carbon neutral.
7. In 2021, Birmingham introduced a clean air zone where some diesel cars had to pay to enter the zone. The average NOX emissions for diesel cars is significantly higher (mean and median both higher) with both data sets about equally spread. The oxides of nitrogen emissions are very bad for people with breathing problems such as asthma, so they are unwelcome in city areas where there is a lot of traffic. Also, older diesel cars emit particulates which are also bad for people with breathing problems – this is included in the original LDS and could be the focus of a different investigation.

	Petrol NOX emissions g/km	**Diesel NOX emissions g/km**
Mean	0.02079	0.04941
Standard deviation	0.01308	0.0151
Minimum	0.006	0.017
Q1 – lower quartile	0.01	0.035
Median	0.0145	0.051
Q3 - upper quartile	0.0315	0.061
Maximum	0.058	0.074

Look for correlation between two lists

8. Engine size in list 1 is a good choice to be an independent variable because it is likely to be at least partly an explanatory variable for emissions.
9.

Petrol engine size	**Correlation coefficient r**	**Correlation**	**Comment**
Mass	0.7482	Strong positive	Cars with bigger engine size tend to be heavier.
CO2 emissions	0.8040	Strong positive	Cars with bigger engine size tend to emit more CO2.
CO emissions	0.0192	None	No correlation between engine size and CO emissions
NOX emissions	0.1405	Weak positive	Very little correlation between engine size and CO emissions

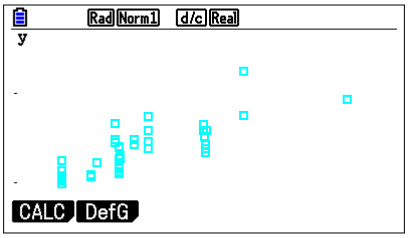
Diesel engine size	**Correlation coefficient r**	**Correlation**	**Comment**
Mass	0.7082	Strong positive	Cars with bigger engine size tend to be heavier.
CO2 emissions	0.7585	Strong positive	Cars with bigger engine size tend to emit more CO2.
CO emissions	0.00088	None	No correlation between engine size and CO emissions
NOX emissions	-0.0639	None	No correlation between engine size and CO emissions

The pattern is very similar for diesel cars.

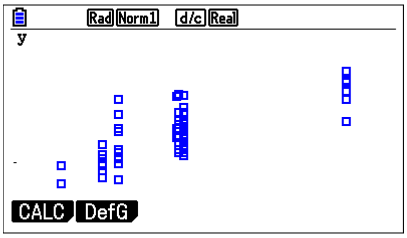
Look at the regression line for an independent and dependent variable

10. Draw a scatter diagram for CO2 emissions against engine size. (**XList:** List1, **YList:** List3 or **XList:** List6, **YList:** List8). Write down the equation of the regression line. Interpret the values of the coefficients in the regression line. Comment on the appearance of the graph.

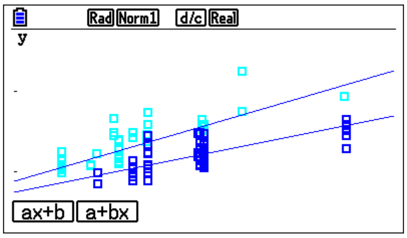
Petrol



Diesel



Both



For petrol cars the equation of the regression line is $y = 53.55 + 0.0518x$

The coefficient of *x* means that for every 1 addition cubic cm of engine size, the CO2 emission increase by 0.0518 g/km

The constant term means that for an engine size of 0 cubic cm, the emissions would be 53.55 g/km

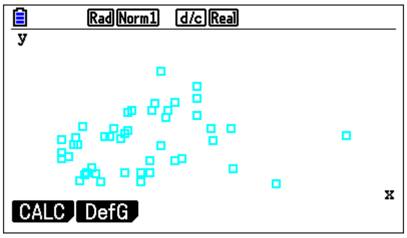
For diesel cars the equation of the regression line is $y = 50.60 + 0.0357x$

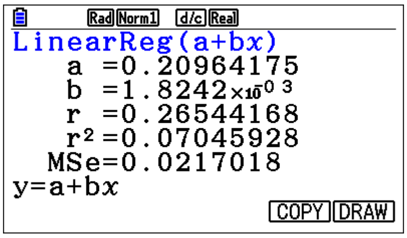
The coefficient of *x* means that for every 1 addition cubic cm of engine size, the CO2 emission increase by 0.0357 g/km

The constant term means that for an engine size of 0 cubic cm, the emissions would be 50.60 g/km

11. For petrol cars (cyan) there is weak positive correlation between CO2 and CO emissions.

For diesel cars (blue) there is weak negative correlation between CO2 and CO emissions.





LinearReg(a+bx)

a =0.20964175

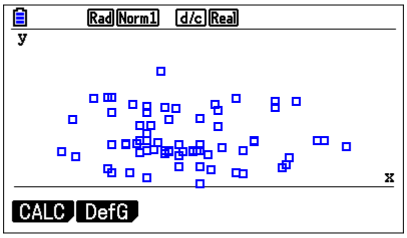
b =1.8242×10^{-03}

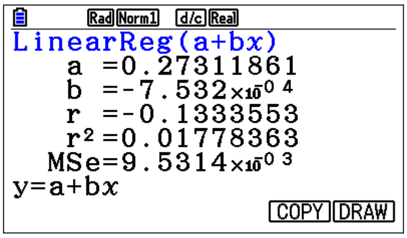
r =0.26544168

r^2 =0.07045928

MSe=0.0217018

y=a+bx





LinearReg(a+bx)

a =0.27311861

b =-7.532×10^{-04}

r =-0.1333553

r^2 =0.01778363

MSe=9.5314×10^{-03}

y=a+bx