The Large Data Set OCR (A)

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

Overview

This resource establishes principles for working with the OCR (A) Large Data Set. The spreadsheet *Transport with headings.xlsx* contains a subset of the LDS and is available from the resource centre. The cleaned data *Transpt.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 the 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)
- Create and use new lists of data which are functions of the existing data (proportions)
- 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 OCR Large Data Set – transport use in 2011 for the North West and London

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

There are 10 columns of data from the Large Data Set showing the numbers of people in each authority.

List 1	List 2	List 3	List 4	List 5	List 6	List 7	List 8	List 9	List 10
NW people in employment	**NW train**	**NW driving car or van**	**NW passenger in car or van**	**NW Bicycle**	**London people in employment**	**London train**	**London driving car or van**	**London passenger in car or van**	**London Bicycle**

Look at a list on its own

1. Create a histogram for the number of people in employment in the London. Use Start 0 and Width 10 000. Comment on the shape of the graph. Does the data for the North West have the same shape?

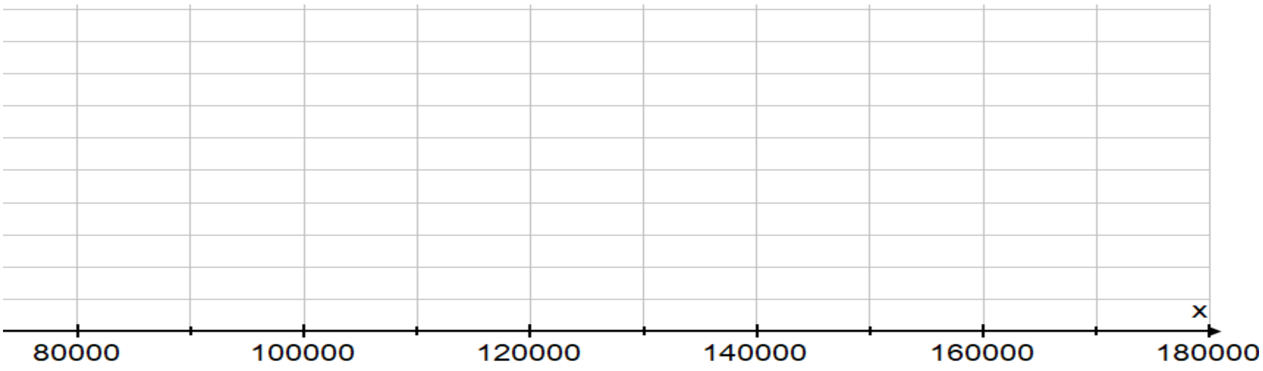
....

....

Compare one list with another

2. Calculate the following statistics for the numbers of people in employment in the North West and London. Draw a boxplot for each on the axes below.

	North West	London
Mean		
Standard deviation		
Boundary value for outliers using mean		
Boundary value for outliers using mean		
Minimum		
Q1 – lower quartile		
Median		
Q3 – upper quartile		
Maximum		
Boundary value for outliers using quartiles		
Boundary value for outliers using quartiles		



3. Comment on the calculations and graphs. ....

....

....

....

....

....

....

Look for correlation between two lists

4. Draw a scatter diagram of the number of people who cycle to work in the North West against the number who drive. Find the correlation coefficient and comment on its value. Is this the correlation you expected?

....

....

5. Investigate the correlation between other lists – why do you see so much positive correlation? ....
....

....

....

....

Create new data from the existing data and analyse it

6. Use the List functionality to create additional lists for the proportion of the number of people in employment who use each form of transport in the North West and London.

Scroll to the heading of List 11. Type List 2 ÷ List 1 and press EXE. Repeat for the other lists for the North West and London.

List 11	List 12	List 13	List 14	List 15	List 16	List 17	List 18
NW train proportion	**NW driving car or van proportion**	**NW passenger in car or van proportion**	**NW Bicycle proportion**	**London train proportion**	**London driving car or van proportion**	**London passenger in car or van proportion**	**London Bicycle proportion**

7. Investigate the correlation between the proportion of people in employment who cycle and the proportion who drive in the North West. What do you notice? ....

....

....

....

....

8. Repeat for other forms of transport. What do you notice? ....

....

....

9. Explain why this data is sometimes better considered as proportions rather than numbers of people.

....

....

10.Compare for the proportion of people in employment who cycle to work in the North West and London.

	North West	London
Mean		
Standard deviation		
Boundary value for outliers using mean		
Boundary value for outliers using mean		
Minimum		
Q1 – lower quartile		
Median		
Q3 – upper quartile		
Maximum		
Boundary value for outliers using quartiles		
Boundary value for outliers using quartiles		

....

....

....

Look at the regression line for an independent and dependent variable

11. Draw a scatter diagram for the **number** of people who are passengers in cars against the **number** of those who drive in London. (XList: List8, YList: List9). Explain why List 8 is a good choice for the independent variable. ....

....

....

12. Write down the value of the correlation coefficient and equation of the regression line. Interpret the values of these numbers. Comment on the appearance of the graph. ....

....

....

....

13. Repeat for the North West. ....

....

....

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 the OCR Large Data Set

To build familiarity with the OCR (A) 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 *Transpt.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.

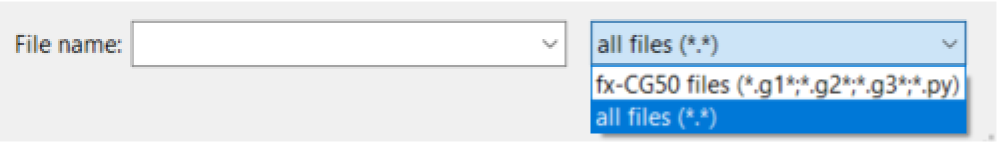
Opening a saved file on the fx-CG50

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.



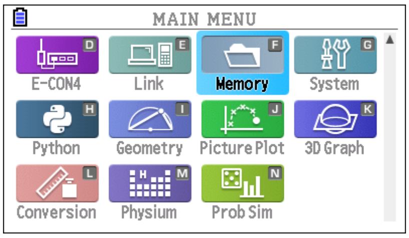
File name:

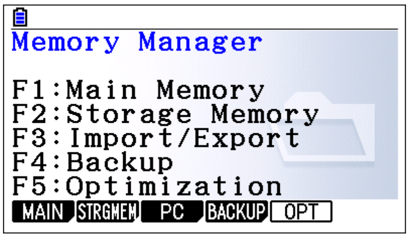
all files (*.*)

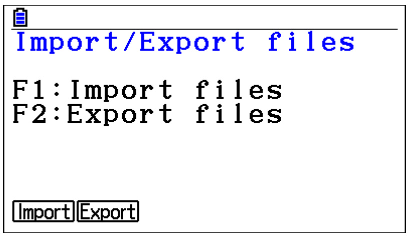
fx-CG50 files (*.g1*;*.g2*;*.g3*;*.py)

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.

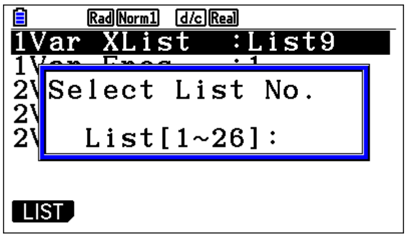




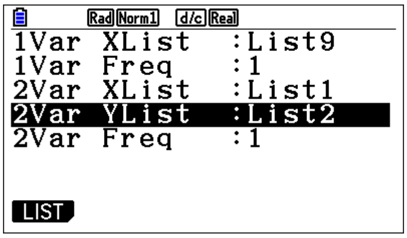


To calculate a set of statistics for a list of numbers

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
2\Select List No.
2\
2\  List[1~26]:
```



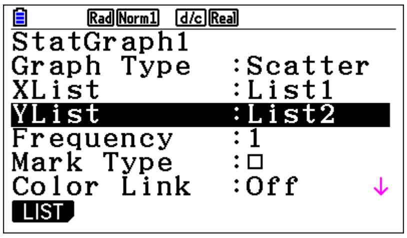
```
1Var XList   :List9
1Var Freq    :1
2Var XList   :List1
2Var YList   :List2
2Var Freq    :1
```

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.

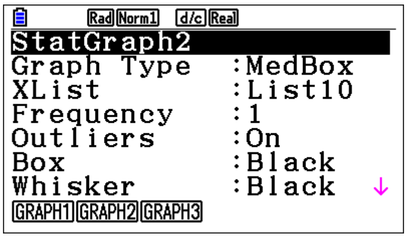


```
StatGraph1
Graph Type   :Scatter
XList        :List1
YList        :List2
Frequency    :1
Mark Type    :□
Color Link   :Off      ↓
```

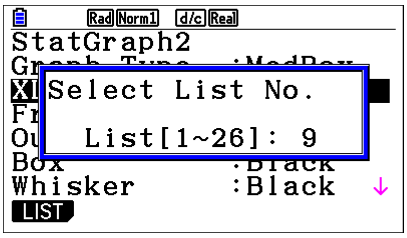
The How To videos listed above cover how to manage lists, calculate single variable statistics and draw graphs.

To create boxplots with different colours

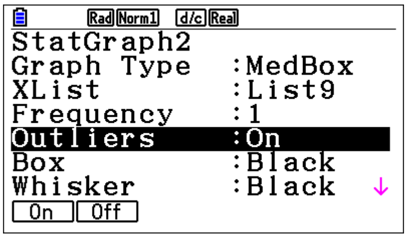
Go to **Statistics**, choose **GRAPH** and **SET**. Choose **GRAPH1** to be Graph Type: MedBox for one List and **GRAPH2** to be Graph Type: MedBox for the other. You can turn Outliers: On or Off and scroll down to change colours for the shading or the outliers.



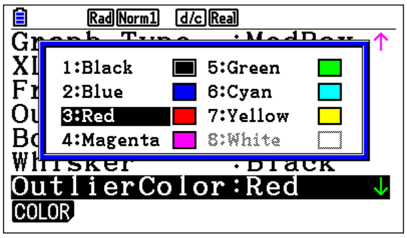
```
StatGraph2
Graph Type    :MedBox
XList         :List10
Frequency     :1
Outliers      :On
Box           :Black
Whisker       :Black
```



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



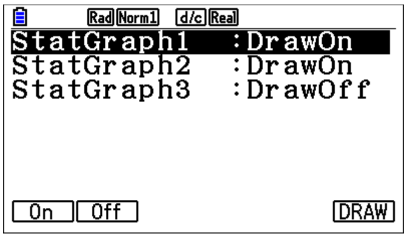
```
StatGraph2
Graph Type    :MedBox
XList         :List9
Frequency     :1
Outliers      :On
Box           :Black
Whisker       :Black
```



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

To show multiple graphs together

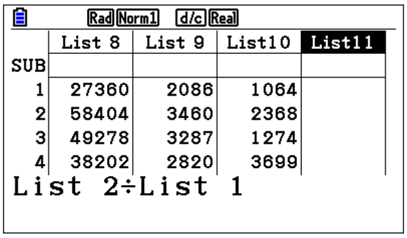
You can create two or three boxplots, or two scatter diagrams with similar ranges of data, as **GRAPH1** and **GRAPH2** for example. To display them on the same axes, go to **GRAPH**, choose **SELECT** (F4) and scroll to the graphs you wish to see together. Then switch them to DrawOn. Choose **DRAW**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



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

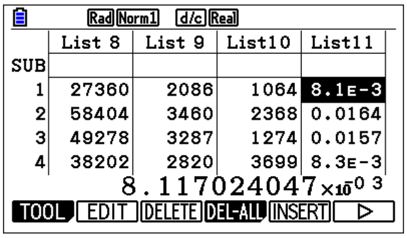
To create a new list in the data set from the existing data

In List 11, to show the result of dividing the value in list 2 by the value in list 1, scroll to the heading of List 11 and type List 2 ÷ List 1. To type List 2, press SHIFT - 1 - 2 . Type SHIFT - 1 - 1 for List 1.



	List 8	List 9	List10	List11
SUB				
1	27360	2086	1064	
2	58404	3460	2368	
3	49278	3287	1274	
4	38202	2820	3699	

List 2÷List 1



	List 8	List 9	List10	List11
SUB				
1	27360	2086	1064	8.1E-3
2	58404	3460	2368	0.0164
3	49278	3287	1274	0.0157
4	38202	2820	3699	8.3E-3

$8.117024047 \times 10^{-03}$

Dealing with the unexpected

A .csv file may not load to the fx-CG50 or emulator if the data it contains is not of the correct form. When creating a subset of their own to study, learners must make sure that the data is clean with no text anywhere, that the cells are formatted as General (no commas appear in the 4- or 5-digit numbers). It is wise to use Paste Values rather than simply Paste.

The List facility is not a spreadsheet, so if you use Lists 1 and 2 to create List 11, any 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 identifying outliers, prompt learners to look at the corresponding spreadsheet to find out what the outliers are and why. The outliers for number of people in employment are Manchester in the North West and the City of London for the London boroughs. The outliers for the proportion of cyclists are Barrow-in-Furness in the North West and Hackney in London. See the following for more information: https://visitbarrow.org.uk/cycling/ and https://hackney.gov.uk/movebybike

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.

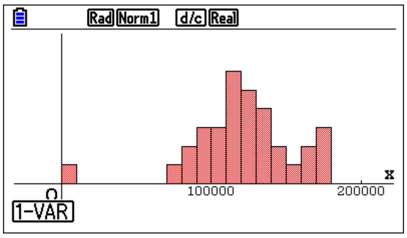
Extension questions

1. Do you think you will find similar things for other regions of the UK?
2. What other factors in the data set could be considered as independent variables for a new study?

Solutions

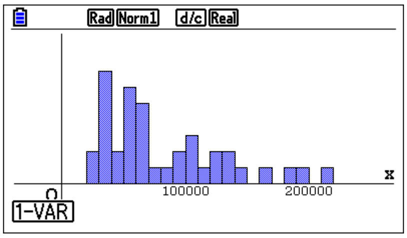
Look at a list on its own

1.



There is an outlier (City of London) and several large boroughs – it does not fit a normal distribution.

For the North West, the data has a positive skew.

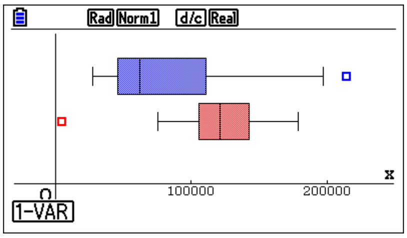


Compare one list with another

2.

	North West	London
Mean	82788	121179
Standard deviation	49013	33353
Boundary value for outliers using mean	-15238	54473
Boundary value for outliers using mean	180814	187885
Minimum	27461	4747
Q1 – lower quartile	46076	105793.5
Median	61691	120873
Q3 – upper quartile	110912	142505
Maximum	213705	178582
Boundary value for outliers using quartiles	-181945	50726
Boundary value for outliers using quartiles	338933	197572

Outliers: Manchester and City of London

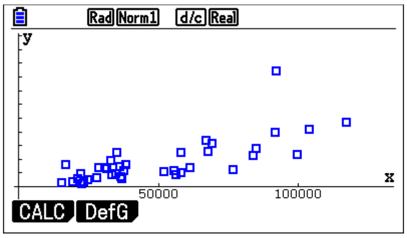


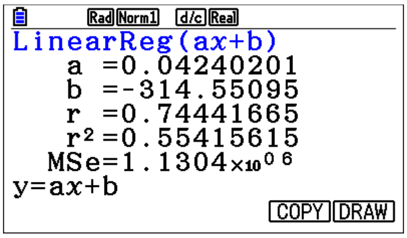
3. The average size of the boroughs is larger than the authorities in the North West as the mean and median are much bigger. There is more variation in the North West as the standard deviation is larger. London

has one outlier which is much smaller than the others, whereas the North West has one outlier at the top.

Look for correlation between two lists

4. The scatter diagram of the number of people who cycle to work in the North West against the number who drive.





```
LinearReg(ax+b)
   a  =0.04240201
   b  =-314.55095
   r  =0.74441665
   r² =0.55415615
  MSe=1.1304×10^06
y=ax+b
```

Correlation coefficient is 0.7444 which is strong positive correlation, where you might expect negative correlation as a lot of people driving might discourage people from cycling.

5. For the North West, the number using:

Train and driving r = 0.7529

Train and passenger r = 0.7852

Train and cycling r = 0.6091

Driving and passenger r = 0.9265 (Compare with the correlation between the proportion of people driving and the number of passengers.)

Driving and cycling r = 0.7444

Passenger and cycling r = 0.7085

For London, the number using:

Train and driving r = 0.3838

Train and passenger r = 0.3476

Train and cycling r = 0.0029

Driving and passenger r = 0.9810

Driving and cycling r = -0.5245

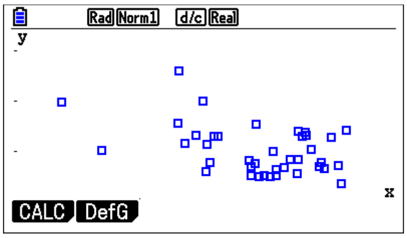
Passenger and cycling r = -0.5663

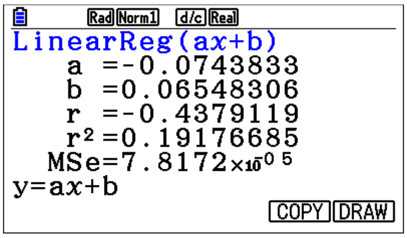
Correlation with the number in employment with train r = 0.8060, driving r = 0.9628, passenger r = 0.9514, bicycle r = 0.8468 in the North West. For London, correlation with the number in employment with train r = 0.4864, driving r = 0.4796, passenger r = 0.4599, bicycle r = 0.3034

Most of the numbers for each form of transport correlation strongly with the number of people in the authority, so other quantities correlate strongly with each other. It is necessary to compare proportions rather than numbers.

Create new data from the existing data and analyse it

6. Scatter diagram for the proportion who cycle against the proportion who drive.





```
LinearReg(ax+b)
   a  =-0.0743833
   b  =0.06548306
   r  =-0.4379119
   r² =0.19176685
  MSe=7.8172×10^-05
y=ax+b
```

7. The correlation coefficient is -0.4379 which is negative so the higher the proportion of people who drive, the lower the proportion of people who cycle. This is more what one might expect.
8. For the North West, the proportions using:

Train and driving *r* = -0.2228

The train and passenger *r* = -0.1294

The train and cycling *r* = 0.0212

Driving and passenger *r* = 0.0168 (almost no correlation – compare with the correlation between the numbers of people driving and the number of passengers.)

Driving and cycling *r* = -0.4379

Passenger and cycling *r* = -0.2491

For London, the proportions using:

Train and driving *r* = 0.3615

The train and passenger *r* = 0.3109

The train and cycling *r* = -0.2583

Driving and passenger *r* = 0.9729

Driving and cycling *r* = -0.7543

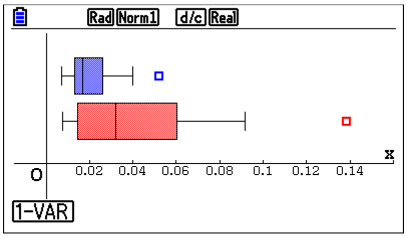
Passenger and cycling *r* = -0.7893

9. Proportions will be more useful when the absolute number varies a lot in magnitude between variables (for example train and cycling).

10. Proportion of people in employment who cycle to work in the North West and London.

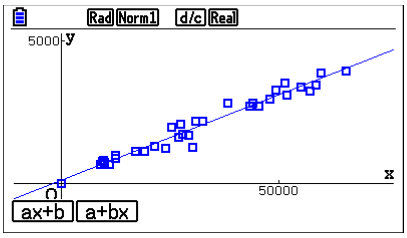
	North West	**London**
Mean	0.02044	0.03944
Standard deviation	0.00958	0.02911
Boundary value for outliers using mean	0.018524	-0.01878
Boundary value for outliers using mean	0.022356	0.09766
Minimum	0.00719	0.00763
Q1 – lower quartile	0.01323	0.01431
Median	0.01684	0.03188
Q3 – upper quartile	0.02623	0.06032
Maximum	0.05183	0.13824
Boundary value for outliers using quartiles	-0.00627	-0.05457
Boundary value for outliers using quartiles	0.04573	0.12911

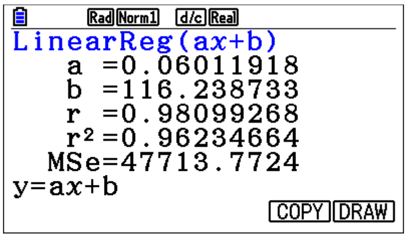
Proportion who use a bicycle: outliers are Barrow-in-Furness and Hackney.



Look at the regression line for an independent and dependent variable

11. List 8 is a good choice for the independent variable as changes in the number of drivers may cause a change in the number of passengers.

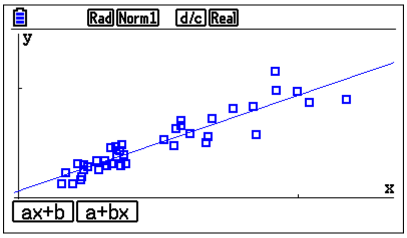


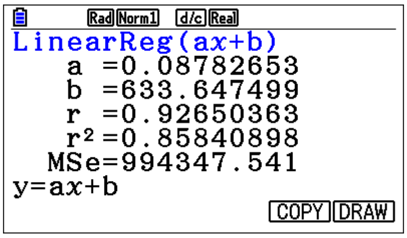


```
LinearReg(ax+b)
    a  =0.06011918
    b  =116.238733
    r  =0.98099268
    r² =0.96234664
   MSe=47713.7724
y=ax+b
```

12. The correlation coefficient *r* = 0.9810 (to 4sf) and equation of the regression line is $y = 0.06012x + 116.2$ This means that for every extra driver you would expect 0.06 extra passengers (about 1 car in 16 has a passenger) and that if there were no drivers, there would be 116 passengers (clearly nonsense to use the regression line for $x = 0$ as it is extrapolating too far).

13. For the North West





```
LinearReg(ax+b)
    a  =0.08782653
    b  =633.647499
    r  =0.92650363
    r² =0.85840898
   MSe=994347.541
y=ax+b
```

The correlation coefficient *r* = 0.9265 (to 4sf) and equation of the regression line is $y = 0.08783x + 633.6$ This means that for every extra driver you would expect 0.087 extra passengers (about 1 car in 11 or 12 has a passenger) and that if there were no drivers, there would be 634 passengers (clearly nonsense to use the regression line for $x = 0$ as it is extrapolating too far).